

Wiley Encyclopedia of
Operations Research
and Management Science

James J. Cochran
Editor-in-Chief

Louis Anthony Cox, Jr., Pinar Keskinocak,
Jeffrey P. Kharoufeh, and J. Cole Smith
Area Editors

A (45)

A Concise Overview of Applications of Ant Colony Optimization	1
A Review of Equity in Nonprofit and Public Sector: A Vehicle Routing Perspective	17
A Review of Tools, Practices, and Approaches for Sustainable Supply Chain Management	29
A Societal Cost of Obesity in Terms of Automobile Fuel Consumption	48
A Structural Clustering Algorithm for Large Networks	62
Accelerated Life Models	79
Accident Precursors and Warning Systems Management: A Bayesian Approach to Mathematical Models	88
Advanced Branching Processes	102
Age Replacement Policies	113
Aggregate Planning	119
Aggregation and Lumping of DTMCs	129
Aging, Characterization, and Stochastic Ordering	137
Air Traffic Management	144
Airline Resource Scheduling	156
Allocation Games	175
Alternating Renewal Processes	186
American Football: Rules and Research	189
An Introduction to Linear Programming	198
An Introduction to Platelet Inventory and Ordering Problems	215
An Introduction to Probabilistic Risk Analysis for Engineered Systems	222
An Introduction to R&D Portfolio Decision Analysis	248
An Overview of Inquiry-Based Learning in Mathematics	262
An Overview of Operations Research in Tennis	273
Analysis of Pairwise Comparison Matrices	290
Analytic Modeling of Insurgencies	299
Analytics in Retail	308
Ant Colony Optimization	319
Antithetic Variates	330
Application of Operations Research in Amusement Park Industry	337
Applying Value of Information and Real Options in R&D and New Product Development	345
Approximate Dynamic Programming I: Modeling	366
Approximate Dynamic Programming II: Algorithms	377
Approximation Algorithms for Stochastic Optimization Problems in Operations Management	392
Assessing Probability Distributions from Data	412
Assessing Remaining Useful Lifetime of Products	420

Association of European Operational Research Societies–434

Asymptotic Behavior of Continuous-Time Markov Chains–439

Asymptotic Behavior of Discrete-Time Markov Chains–444

Austrian Society of Operations Research (Oesterreichische Gesellschaft für Operations Research, Oegor)–451

Availability Analysis: Concepts and Methods–455

Availability in Stochastic Models–466

Average Reward of a Given MDP Policy–479

Aviation Safety and Security–482

Axiomatic Measures of Risk and Risk-Value Models–492

Axiomatic Models of Perceived Risk–503

B (28)

Backtrack Search Techniques and Heuristics–514

Backward and Forward Equations for Diffusion Processes–524

Basic CP Theory: Consistency and Propagation (Advanced)–529

Basic CP Theory: Search–540

Basic Interdiction Models–561

Basic Polyhedral Theory–569

Basis Reduction Methods–582

Batch Arrivals and Service-Single Station Queues–601

Batch Markovian Arrival Processes (BMAP)–608

Bayesian Aggregation of Experts' Forecasts–616

Bayesian Network Classifiers–630

Behavioral Economics and Game Theory–640

Behavioral Operations: Applications in Supply Chain Management–651

Benders Decomposition–663

Biclustering: Algorithms and Application in Data Mining–671

Bilevel Network Interdiction Models: Formulations and Solutions–677

Bilinear Optimization–688

Biosurveillance: Detecting, Tracking, and Mitigating the Effects of Natural Disease and Bioterrorism–695

Birth-and-Death Processes–712

Block Replacement Policies–718

Branch and Cut–725

Branch-and-Bound Algorithms–734

Branch-Price-and-Cut Algorithms–742

Branch-Width and Tangles–755

Brazilian Society of Operational Research–763

Brownian Motion and Queueing Applications–766
 Business Process Outsourcing–773
 Byelorussian Operational Research Society (ByORS)–781

C (62)

Calculating Real Option Values–784
 Call Center Management–799
 Campaign Analysis: An Introductory Review–807
 Capacity Allocation in Supply Chain Scheduling–816
 Capacity Allocation–825
 Capacity Planning in Health Care–834
 Capacity Planning–842
 Category and Inventory Management–851
 Central Path and Barrier Algorithms for Linear Optimization–870
 Chilean Institute of Operations Research–886
 Chinese Postman Problem–892
 Classic Financial Risk Measures–902
 Clique Relaxations–912
 Closed-Loop Supply Chains: Environmental Impact–919
 Clustering–930
 Cognitive Mapping and Strategic Options Development and Analysis (SODA)–942
 Coherent Systems–952
 Collaborative Procurement–960
 Column Generation–968
 Combinatorial Auctions: Complexity and Algorithms–982
 Combinatorial Auctions–994
 Combinatorial Traveling Salesman Problem Algorithms–1004
 Combining Exact Methods and Heuristics–1013
 Combining Forecasts–1022
 Combining Scenario Planning with Multiattribute Decision Making–1030
 Common Failure Distributions–1039
 Common Random Numbers–1050
 Communicating Decision Information to the Lay Risk Manager: A Consultant's Perspective–1061
 Comparisons of Risk Attitudes Across Individuals–1072
 Competing Risks and Limited Failure–1085
 Complementarity Problems–1097
 Computation and Dynamic Programming–1107

Computational Biology and Bioinformatics: Applications in Operations Research–1124

Computational Methods for CTMCs–1134

Computational Methods for DTMCs–1141

Computational Pool: An OR-Optimization Point of View–1153

Concepts of Network Reliability–1165

Conceptual Modeling for Simulation–1176

Condition-Based Maintenance Under Markovian Deterioration–1188

Conic Optimization Software–1200

Conservation Laws and Related Applications–1210

Constraint Programming Links with Math Programming–1223

Constraint Qualifications–1238

Continuous Optimization by Variable Neighborhood Search–1247

Continuous-Time Control under Stochastic Uncertainty–1260

Continuous-Time Martingales–1287

Contributions to Software Reliability with OR Applications–1291

Control Variates–1310

Cooperative Game Theory with Nontransferable Utility–1319

Cooperative Games with Transferable Utility–1327

Coordination of Production and Delivery in Supply Chain Scheduling–1343

Cost-Effectiveness Analysis, Health-Care Policy, and Operations Research Models–1354

Cover Inequalities–1370

Credit Risk Assessment–1376

Credit Risk–1383

Croatian Operational Research Society–1393

Cross-Entropy Method–1396

CTMCs with Costs and Rewards–1403

Customer Relationship Management: Maximizing Customer Lifetime Value–1408

Customized Price Responses to Bid Opportunities in Competitive Markets–1419

Cusum Charts for Multivariate Monitoring and Forecasting–1428

Czech Society for Operations Research–1435

D (50)

Damage, Stress, Degradation, Shock–1437

Dantzig-Wolfe Decomposition–1444

Data Classification and Prediction–1456

Data Mining in Construction Bidding Policy–1462

Decision Analysis and Counterterrorism–1472

Decision Making Under Pressure and Constraints: Bounded Rationality–1480

Decision Making with Partial Probabilistic or Preference Information–1488

Decision Problems and Applications of Operations Research at Marine Container Terminals–1497

Decision Rule Preference Model–1517

Decision-Theoretic Foundations of Simulation Optimization–1533

Decomposition Algorithms for Two-Stage Recourse Problems–1543

Decomposition Methods for Integer Programming–1553

Defining Objectives and Criteria for Decision Problems–1564

Definition and Examples of Continuous-Time Markov Chains–1575

Definition and Examples of DTMCs–1579

Definition and Examples of Renewal Processes–1585

Degeneracy And Variable Entering/Exiting Rules–1589

Delayed Renewal Processes–1596

Demand Responsive Transportation–1599

Describing Decision Problems by Decision Trees–1608

Description of the French Operational Research and Decision-Aid Society: Société Française de Recherche Opérationnelle et d'aide À la Décision (ROADEF)–1625

Descriptive Models of Decision Making–1629

Descriptive Models of Perceived Risk–1645

Design and Control Principles of Flexible Workforce in Manufacturing Systems–1654

Design Considerations for Supply Chain Tracking Systems–1671

Design for Manufacturing and Assembly–1681

Design for Network Resiliency–1711

Deterministic Dynamic Programming (DP) Models–1728

Deterministic Global Optimization–1736

Different Formats for the Communication of Risks: Verbal, Numerical, and Graphical Formats–1756

Differential Games–1767

Direct Search Methods–1775

Discrete Optimization with Noisy Objective Function Measurements–1788

Discrete-Time Martingales–1802

Discretization Methods for Continuous Probability Distributions–1805

Disjunctive Inequalities: Applications and Extensions–1818

Disjunctive Programming–1828

Distributed Simulation in ORMS–1837

Domination Problems–1850

Drama Theory–1868

DTMCS with Costs and Rewards–1877

Dual Simplex–1882

Dynamic Auctions–1892

Dynamic Models for Robust Optimization–1903

Dynamic Pricing Strategies for Multiproduct Revenue Management Problems–1914

Dynamic Pricing Under Consumer Reference-Price Effects–1927

Dynamic Programming: Introductory Concepts–1944

Dynamic Programming Via Linear Programming–1955

Dynamic Programming, Control, and Computation–1961

Dynamic Vehicle Routing–1967

E (22)

Edgeworth Market Games: Price-Taking and Efficiency–1978

Effective Application of GRASP–1992

Effective Application of Guided Local Search–2001

Effective Application of Simulated Annealing–2012

Efficient Iterative Combinatorial Auctions–2022

Efficient Use of Materials and Energy–2036

Electronic Negotiation Systems–2042

Eliciting Subjective Probabilities from Individuals and Reducing Biases–2050

Eliciting Subjective Probability Distributions from Groups–2063

Ellipsoidal Algorithms–2070

Emergency Medical Service Systems that Improve Patient Survivability–2083

Estimating Failure Rates and Hazard Functions–2098

Estimating Intensity and Mean Value Function–2114

Estimating Survival Probability–2128

Estonian Operational Research Society–2144

Eulerian Path and Tour Problems–2147

Evacuation Planning–2154

Evaluating and Comparing Forecasting Models–2165

Evaluations of Single- and Repeated-Play Gambles–2176

Evolutionary Algorithms–2183

Evolutionary Game Theory and Evolutionary Stability–2196

Exact Solution of the Capacitated Vehicle Routing Problem–2207

F (18)

Fairness and Equity in Societal Decision Analysis–2219

Fast and Frugal Heuristics–2228

Feasible Direction Method–2236

Feature Extraction and Feature Selection: A Survey of Methods in Industrial Applications–2243

Fictitious Play Algorithm–2254

Finite Population Models-Single Station Queues–2262

Fire Department Deployment and Service Analysis–2268

Fluid Models of Queueing Networks–2277

Forecasting Approaches for the High-Tech Industry–2292

Forecasting for Inventory Planning under Correlated Demand–2298

Forecasting Nonstationary Processes–2309

Forecasting: State-Space Models and Kalman Filter Estimation–2320

Formulating Good MILP Models–2329

Foundations of Constrained Optimization–2343

Foundations of Decision Theory–2353

Foundations of Simulation Modeling–2364

Fritz-John and KKT Optimality Conditions for Constrained Optimization–2379

Fuzzy Measures and Integrals in Multicriteria Decision Analysis–2385

G (14)

Game-Theoretic Methods in Counterterrorism and Security–2393

Generating Homogeneous Poisson Processes–2399

Generating Nonhomogeneous Poisson Processes–2405

Generic Stochastic Gradient Methods–2409

Genetic Algorithms–2417

Geometric Programming–2441

German Operations Research Society (GOR) (Gesellschaft Für Operations Research)–2454

Gomory Cuts–2455

Gradient-Type Methods–2470

Graph Search Techniques–2478

Graphical Methods for Reliability Data–2485

Grasp: Greedy Randomized Adaptive Search Procedures–2496

Group Dynamics Processes for Improved Decision Making–2507

Guided Local Search–2515

H (12)

Hazard Rate Function–2528

Hazardous Materials Transportation–2535

Hellenic Operational Research Society–2543

Heuristics and Their Use in Military Modeling–2549

Heuristics for the Traveling Salesman Problem–2574

Heuristics in Mixed Integer Programming–2581
 History of Constraint Programming–2587
 History of LP Development–2598
 Holt-Winters Exponential Smoothing–2607
 Horse Racing–2616
 Housing and Community Development–2625
 Hyper-Heuristics–2638

I (35)

Ice Hockey–2646
 Icelandic Operations Research Society–2657
 IFORS: Bringing the World of OR Together for 50 Years–2661
 Implementing the Simplex Method–2666
 Impossibility Theorems and Voting Paradoxes in Collective Choice Theory–2682
 Improving Packaging Operations in the Plastics Industry–2691
 Improving Public Health in Developing Countries Through Operations Research–2702
 Inequalities from Group Relaxations–2717
 Infinite Horizon Problems–2730
 Infinite Linear Programs–2738
 Information Sharing in Supply Chains–2748
 Initial Transient Period in Steady-State Systems–2761
 Inspection Games–2773
 Instance Formats for Mathematical Optimization Models–2782
 Integer Programming Duality–2796
 Integrated Supply Chain Design Models–2805
 Interior Point Methods for Nonlinear Programs–2820
 Interior-Point Linear Programming Solvers–2828
 Introduction to Branching Processes–2837
 Introduction to Diffusion Processes–2842
 Introduction to Discrete-Event Simulation–2847
 Introduction to Facility Location–2860
 Introduction to Large-Scale Linear Programming and Applications–2878
 Introduction to Lévy Processes–2886
 Introduction to Multiattribute Utility Theory–2893
 Introduction to Point Processes–2906
 Introduction to Polynomial Time Algorithms for LP–2911
 Introduction to Rare-Event Simulation–2927

Introduction To Robust Optimization–2938

Introduction to Shop-Floor Control–2946

Introduction to Stochastic Approximation–2955

Introduction to the Use of Linear Programming in Strategic Health Human Resource Planning–2962

Inventory Inaccuracies In Supply Chains: How Can RFID Improve The Performance?–2972

Inventory Record Inaccuracy in Retail Supply Chains–2985

Iranian Operations Research Society (IORS)–3000

J (3)

Jackson Networks (Open and Closed)–3002

Job Shop Scheduling–3015

Just-In-Time/Lean Production Systems–3022

K (2)

Klimov's Model–3032

k-out-of-n Systems–3041

L (22)

Lagrangian Optimization for LP: Theory and Algorithms–3048

Lagrangian Optimization Methods for Nonlinear Programming–3060

Large Deviations in Queueing Systems–3068

Large Margin Rule-Based Classifiers–3075

Latin-Ibero-American Association for Operational Research–3087

Learning with Dynamic Programming–3089

Level-Dependent Quasi-Birth-and-Death Processes–3101

Level-Independent Quasi-Birth-and-Death Processes–3110

Lift-and-Project Inequalities–3120

Lifting Techniques For Mixed Integer Programming–3127

Limit Theorems for Branching Processes–3142

Limit Theorems for Markov Renewal Processes–3145

Limit Theorems for Renewal Processes–3148

Linear Programming and Two-Person Zero-Sum Games–3154

Linear Programming Projection Algorithms–3165

Lipschitz Global Optimization–3173

Little's Law and Related Results–3190

Load-Sharing Systems–3203

Location (Hotelling) Games and Applications–3215

Lot-Sizing–3226

Lovász-Schrijver Reformulation–3236

LP Duality and KKT Conditions for LP–3249

M (52)

MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique)–3257

Maintenance Management–3263

Management of Natural Gas Storage Assets–3272

Management Science/Operations Research Society of Malaysia–3281

Managing A Portfolio of Risks–3283

Managing Corporate Mobile Voice Expenses: Plan Choice, Pooling Optimization, and Chargeback–3307

Managing Perishable Inventory–3326

Managing Product Introductions and Transitions–3336

Managing R&D and Risky Projects–3348

Manufacturing Facility Design and Layout–3353

Markov and Hidden Markov Models–3364

Markov Chains of the M/G/1-Type–3374

Markov Regenerative Processes–3388

Markov Renewal Function and Markov Renewal-Type Equations–3392

Markov Renewal Processes–3394

Markovian Arrival Processes–3397

Mass Customization–3414

Materials Requirements Planning–3423

Mathematical Models for Perishable Inventory Control–3433

Mathematical Programming Approaches to the Traveling Salesman Problem–3450

Matrix Analytic Method: Overview and History–3457

Matrix-Geometric Distributions–3467

Maximum Clique, Maximum Independent Set, and Graph Coloring Problems–3476

Maximum Flow Algorithms–3489

MDP Basic Components–3505

Measures of Risk Equity–3511

Memetic Algorithms–3523

Metaheuristics for Stochastic Problems–3547

Methods For Large-Scale Unconstrained Optimization–3559

Military Operations Research Society (MORS)–3569

MILP Software–3581

Minimum Cost Flows–3591

Minimum Prediction Error Models and Causal Relations between Multiple Time Series–3603

Minimum Spanning Trees–3617

MINLP Solver Software–3629

Mixing Sets–3641

Model-Based Forecasting–3650

Modeling and Forecasting by Manifold Learning–3658

Modeling Uncertainty in Optimization Problems–3672

Models and Basic Properties–3681

Monte Carlo Simulation as an Aid for Deciding Among Treatment Options–3688

Monte Carlo Simulation for Quantitative Health Risk Analysis–3696

Multiarmed Bandits and Gittins Index–3705

Multiclass Queueing Network Models–3714

Multicommodity Flows–3722

Multiechelon Multiproduct Inventory Management–3729

Multimethodology–3737

Multistage (Stochastic) Games–3743

Multistate System Reliability–3759

Multivariate Elicitation: Association, Copulae, and Graphical Models–3766

Multivariate Input Modeling–3773

Music and Operations Research–3783

N (13)

Nash Equilibrium (Pure and Mixed)–3795

Network Reliability Performance Metrics–3812

Network Theory: Concepts and Applications–3817

Network-Based Data Mining: Operations Research Techniques and Applications–3836

Neuroeconomics and Game Theory–3846

Neuroeconomics Insights for Decision Analysis–3856

Newsvendor Models–3867

Newton-Type Methods–3877

Non-Expected Utility Theories–3891

Nonlinear Conjugate Gradient Methods–3903

Nonlinear Multiobjective Programming–3923

Nonstationary Input Processes–3953

Nurse Scheduling Models–3959

O (38)

Olympics–3969

Omega Rho International Honor Society for Operations Research and Management Science–3979

"On-The-Spot" Modeling and Analysis: The Facilitated Modeling Approach–3984

Operation Research in Golf–4001

Operational Research Society of India–4012

Operational Research Society of Nepal (ORSN): An Introduction–4016

Operational Research Society of Turkey–4023

Operational Risk–4025

Operations Research and Management Science in Fire Safety–4039

Operations Research Applications in Truckload Freight Transportation Networks–4050

Operations Research Approaches to Asset Management in Freight Rail–4062

Operations Research for Freight Train Routing and Scheduling–4071

Operations Research in Australia–4081

Operations Research in Data Mining–4084

Operations Research in Forestry and Forest Products Industry–4098

Operations Research in the Visual Arts–4117

Operations Research Models for Cancer Screening–4126

Operations Research Society of China–4140

Operations Research Society of New Zealand–4144

Operations Research Society of Taiwan–4146

Operations Research to Improve Disaster Supply Chain Management–4150

Operations Research Tools for Addressing Current Challenges in Emergency Medical Services–4159

Optimal Monitoring Strategies–4173

Optimal Reliability Allocation–4184

Optimal Replacement and Inspection Policies–4190

Optimal Risk Mitigation and Risk-Taking–4197

Optimization and Decision Sciences (Italy)–4212

Optimization for Dispatch and Short-Term Generation in Wholesale Electricity Markets–4217

Optimization Models for Cancer Treatment Planning–4225

Optimization of Public Transportation Systems–4239

Optimization Problems in Passenger Railway Systems–4249

Optimizing the Aviation Checkpoint Process to Enhance Security and Expedite Screening–4259

Option Pricing: Theory and Numerical Methods–4267

OR Models in Freight Railroad Industry–4280

OR/MS Applied to Cricket–4300

ORBEL, The Belgian Operations Research Society (SOGESCI-BVWB)–4308

ORSP: FORging Ahead in its Twenty-Fifth Year–4312

Overweighting of Small Probabilities–4314

P (39)

Paradoxes and Violations of Normative Decision Theory–4322

Parallel Configurations–4329

Parallel Discrete-Event Simulation–4333

Parallel Systems–4345

Parallel-Series and Series-Parallel Systems–4351

Parametric LP Analysis–4357

Partially Observable MDPs (POMDPs): Introduction and Examples–4364

PASTA and Related Results–4384

Penalty and Barrier Methods–4396

Percolation Theory–4406

Perfect Bayesian Equilibrium and Sequential Equilibrium–4415

Perfect Information and Backward Induction–4422

Performance Bounds in Queueing Networks–4429

Phase-Type (PH) Distributions–4438

Point and Interval Availability–4446

Poisson Process and its Generalizations–4457

Polynomial Time Primal Integer Programming via Graver Bases–4466

Portuguese Operational Research Society-APDIO–4475

Power Indices–4478

Presolving Mixed-Integer Linear Programs–4497

Pricing and Lead-Time Decisions–4506

Pricing and Replenishment Decisions–4514

Pricing and Scheduling Decisions–4521

Primal-Dual Methods for Nonlinear Constrained Optimization–4530

Probabilistic Distance Clustering–4543

Probability Weighting Functions–4562

Problem Structuring for Multicriteria Decision Analysis Interventions–4580

Procurement Contracts–4594

Product/Service Design Collaboration: Managing the Product Life Cycle–4605

Progressive Adaptive User Selection Environment (PAUSE) Auction Procedure–4615

Project-Based ORMS Education–4625

Prospect Theory–4640

Psychology of Risk Perception–4649

Public Health, Emergency Response, and Medical Preparedness I: Medical Surge–4657

Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts–4668

Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure–4690
Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation–4704
Pure Cutting-Plane Algorithms and their Convergence–4713
Push and Pull Production Systems–4724

Q (8)

Quality and Pricing Decisions–4731
Quality Design, Control, and Improvement–4743
Quality Management–4754
Quantum Command and Control Theory–4765
Quantum Game Theory–4775
Quasi-Newton Methods–4784
Queueing Disciplines–4799
Queueing Notation–4810

R (35)

R&D Risk Management–4813
Random Search Algorithms–4821
Random Variate Generation–4834
Randomized Simplex Algorithms–4843
Recycling–4848
Reduction of a POMDP to an MDP–4857
Reflected Brownian Motion–4866
Reformulation-Linearization Technique for MIPs–4873
Regenerative Processes–4880
Reinforcement Learning Algorithms for MDPs–4886
Relationship among Benders, Dantzig-Wolfe, and Lagrangian Optimization–4903
Reliability Indices–4909
Remanufacturing–4930
Rendezvous Search Games–4940
Renewal Function and Renewal-Type Equations–4952
Renewal Processes With Costs and Rewards–4955
Rent and Rent Loss in the Icelandic Cod Fishery–4961
Repairable Systems: Renewal and Nonrenewal–4969
Resource Modeling Association–4982
Retrial Queues–4985
Revenue Management in the Travel Industry–4992
Revenue Management with Incomplete Demand Information–5009

Revenue Management–5026

Reversibility in Queueing Models–5041

Reviews of Maintenance Literature and Models–5060

Risk Assessments and Black Swans–5075

Risk Averse Models–5083

Robust External Risk Measures–5091

Robust Offline Single-Machine Scheduling Problems–5106

Robust Ordinal Regression–5121

Robust Portfolio Selection–5131

Robustness Analysis–5143

Robustness for Operations Research and Decision Aiding–5148

Rule Developing Experimentation in Consumer-Driven Package Design–5158

Russian Scientific Operations Research Society–5169

S (79)

Saddle points and von Neumann Minimax Theorem–5173

Sales Optimization Models-Sales Force Territory Planning–5180

Sampling Methods–5189

Scenario Generation–5196

Scheduling Seaside Resources at Container Ports–5205

Scoring Rules–5227

Search Games–5238

Selective Support Vector Machines–5246

Self-Dual Embedding Technique for Linear Optimization–5257

Semidefinite Optimization Applications–5267

Semi-Infinite Programming–5279

Semi-Markov Decision Processes–5289

Semi-Markov Processes and Hidden Models–5298

Semi-Markov Processes–5307

Sensitivity Analysis and Dynamic Programming–5311

Sensitivity Analysis in Decision Making–5319

Sensitivity Analysis in Linear Programming–5330

Sensitivity Analysis of Simulation Models–5341

Sequential Decision Problems Under Uncertainty–5351

Sequential Quadratic Programming Methods–5357

Series Systems–5370

Service Outsourcing–5378

Shipper And Carrier Collaboration–5388
 Shortest Path Problem Algorithms–5401
 Simplex-Based LP Solvers–5414
 Simplifying and Solving Decision Problems by Stochastic Dominance Relations–5425
 Simulated Annealing–5435
 Simulation Optimization in Risk Management–5449
 Simultaneous Ascending Auctions–5458
 Simultaneous Perturbation and Finite Difference Methods–5466
 Single Machine Scheduling–5488
 Single-Dimensional Search Methods–5498
 Single-Search-Based Heuristics for Multiobjective Optimization–5513
 Slovak Society for Operations Research–5521
 Slovenian Society Informatika (SSI)-Section for Operations Research (SOR)–5523
 Soccer/World Football–5529
 Sociedad Peruana De Investigación Operativa Y De Sistemas (SOPIOS)–5543
 Software For Nonlinearly Constrained Optimization–5547
 Software for Solving Noncooperative Strategic form Games–5559
 Solving Influence Diagrams: Exact Algorithms–5567
 Solving Stochastic Programs–5580
 Some Optimization Models and Techniques for Electric Power System Short-term Operations–5592
 Spanish Society of Statistics and Operations Research–5609
 Split Cuts–5611
 Spreadsheet Modeling for Operations Research Practice–5621
 Stakeholder Participation–5629
 Standby Redundant Systems–5638
 Standby Systems–5650
 Statistical Analysis of Call-Center Operational Data: Forecasting Call Arrivals, and Analyzing Customer Patience and Agent Service–5660
 Statistical Methods for Optimization–5670
 Statistical Process Control–5679
 Stochastic Dynamic Programming Models and Applications–5687
 Stochastic Gradient Methods for Simulation Optimization–5697
 Stochastic Hazard Process–5712
 Stochastic Mixed-Integer Programming Algorithms: Beyond Benders' Decomposition–5725
 Stochastic Modeling and Optimization in Baseball–5736
 Stochastic Network Interdiction–5746
 Stochastic Optimal Control Formulations of Decision Problems–5761

Stochastic Orders for Stochastic Processes–5771

Stochastic Search Methods for Global Optimization–5778

Strategic and Operational Prepositioning in Case of Seasonal Natural Disasters: A Perspective–5788

Strategic Customer Behavior in a Single Server Queue–5801

Structural Results for POMDPs–5813

Structured Optimal Policies for Markov Decision Processes: Lattice Programming Techniques–5819

Subgradient Optimization–5844

Subjective Probability–5849

Supplier Selection–5863

Supply Chain Coordination–5876

Supply Chain Outsourcing–5886

Supply Chain Risk Management–5895

Supply Chain Scheduling: Origins and Application to Sequencing, Batching and Lot Sizing–5905

Support Vector Machines for Classification–5913

Supporting the Strategy Process: The Role of Operational Research/Management Science–5925

Surgery Planning and Scheduling–5940

Swiss Operations Research Society (Schweizerische Vereinigung Für Operations Research/Association Suisse Pour La Recherche Operationelle/Associazione Svizzera Di Ricerca Operativa)–5953

Symmetry Handling in Mixed-Integer Programming–5958

System Availability–5972

Systems in Series–5978

Systems Modeling to Inform Drug Policy–5984

T (59)

Tabu Search–5998

Take-Back Legislation and Its Impact on Closed-Loop Supply Chains–6008

Teaching ORMS/Analytics with Cases–6018

Teaching Soft OR/MS Methodologies: What, Why, and How–6031

The Analytics Society of Ireland–6040

The Association of Asia Pacific Operational Research Societies–6042

The Bulgarian Operational Research Society–6048

The Canadian Operational Research Society/Société Canadienne De Recherche Opérationnelle–6050

The Condition-Based Paradigm–6053

The Decision Sciences Institute–6061

The Exponentially Weighted Moving Average–6068

The Failure-Based Paradigm–6077

The G/G/1 Queue–6088

The G/G/s Queue–6094

The G/M/1 Queue–6105
 The Global Replenishment Problem–6115
 The Graph Model for Conflict Resolution–6132
 The Hungarian Operations Research Society–6144
 The Knowledge Gradient for Optimal Learning–6147
 The Law and Economics of Risk Regulation–6163
 The M/G/1 Queue–6175
 The M/G/s Queue–6182
 The M/G/s/s Queue–6190
 The M/M/1 Queue–6194
 The M/M/ ∞ Queue–6202
 The M/M/s Queue–6209
 The Manufacturing and Service Operations Management (MSOM) Society–6215
 The M/G/ ∞ Queue–6219
 The Nascent Industry of Electric Vehicles–6225
 The Naturalistic Decision Making Perspective–6233
 The North American Operations Research Societies–6242
 The Operational Research Society of Singapore–6245
 The Operational Research Society–6247
 The Operations Research Society of South Africa–6251
 The Scatter Search Methodology–6255
 The Search Allocation Game–6267
 The Shapley Value and Related Solution Concepts–6277
 The Simplex Method and Its Complexity–6290
 The Strategic Choice Approach–6299
 The Vehicle Routing Problem with Time Windows: State-of-the-Art Exact Solution Methods–6311
 The Weighted Moving Average Technique–6319
 Theory of Martingales–6327
 Total Expected Discounted Reward MDPS: Existence of Optimal Policies–6335
 Total Expected Discounted Reward MDPS: Policy Iteration Algorithm–6343
 Total Expected Discounted Reward MDPS: Value Iteration Algorithm–6349
 Tour Scheduling and Rostering–6354
 TPZS Applications: Blotto Games–6367
 Tracking Technologies in Supply Chains–6375
 Traffic Network Analysis and Design–6389
 Transient Behavior of CTMCs–6403
 Transient Behavior of DTMCs–6415

Transportation Algorithms–6419

Transportation Resource Management–6434

Travel Demand Modeling–6455

Treewidth, Tree Decompositions, and Brambles–6468

Triage in the Aftermath of Mass-Casualty Incidents–6478

Trust–6488

Two-Stage Stochastic Integer Programming: A Brief Introduction–6501

Two-Stage Stochastic Programs: Introduction and Basic Properties–6508

U (9)

Uncertainty in Forest Production Planning–6512

Understanding and Managing Variability–6522

Uniformization in Markov Decision Processes–6533

Use of a High-Fidelity UAS Simulation for Design, Testing, Training, and Mission Planning for Operation in Complex Environments–6540

Use of Lagrange Interpolating Polynomials in the RLT–6554

Using Holistic Multicriteria Assessments: The Convex Cones Approach–6565

Using Operations Research to Plan Natural Gas Production and Transportation on the Norwegian Continental Shelf–6579

Using or to Overcome Challenges in Implementing New Voting Technologies–6585

Using Queueing Theory to Alleviate Emergency Department Overcrowding–6593

V (5)

Value Functions Incorporating Disappointment and Regret–6602

Variants of Brownian Motion–6608

Variational Inequalities–6623

Vendor-Managed Inventory–6634

Very Large-Scale Neighborhood Search–6644

W (5)

Wardrop Equilibria–6655

Warranty Modeling–6667

Why Risk is Not Variance–6680

Why Traditional Kanban Calculations Fail in Volatile Environments–6683

Writing ORMS/Analytics Cases–6691

A CONCISE OVERVIEW OF APPLICATIONS OF ANT COLONY OPTIMIZATION

THOMAS STÜTZLE
MANUEL LÓPEZ-IBÁÑEZ
MARCO DORIGO
IRIDIA, CoDE, Université Libre de
Bruxelles (ULB), Brussels,
Belgium

Ant colony optimization (ACO) [1–3] is a metaheuristic for solving hard combinatorial optimization problems inspired by the indirect communication of real ants. In ACO algorithms, (artificial) ants construct candidate solutions to the problem being tackled, making decisions that are stochastically biased by numerical information based on (artificial) pheromone trails and available heuristic information. The pheromone trails are updated during algorithm execution to bias the ants search toward promising decisions previously found. The article titled *Ant Colony Optimization* gives a detailed overview of the main concepts of ACO.

Despite being one of the youngest metaheuristics, the number of applications of ACO algorithms is very large. In principle, ACO can be applied to any combinatorial optimization problem for which some iterative solution construction mechanism can be conceived. Most applications of ACO deal with $\mathcal{NP}$ -hard combinatorial optimization problems, that is, with problems for which no polynomial time algorithms are known. ACO algorithms have also been extended to handle problems with multiple objectives, stochastic data, and dynamically changing problem information. There are extensions of the ACO metaheuristic for dealing with problems with continuous decision variables, as well.

This article provides a concise overview of several noteworthy applications of ACO algorithms. This overview is necessarily incomplete because the number of currently available ACO applications goes into the hundreds. Our description of the applications

follows the classification used in the 2004 book on ACO by Dorigo and Stützle [3] but extending the list there with many recent examples. Tables 1 and 2 summarize these applications.

APPLICATIONS TO $\mathcal{NP}$ -HARD PROBLEMS

ACO was primarily intended for solving combinatorial optimization problems, among which $\mathcal{NP}$ -hard problems are the most challenging ones. In fact, no polynomial-time algorithms are known for such problems, and therefore heuristic techniques such as ACO are often used for generating high-quality solutions in reasonable computation times.

Routing Problems

Routing problems involve one or more agents visiting a predefined set of locations, and the objective function and constraints depend on the order in which the locations are visited. Perhaps the best-known example is the traveling salesman problem (TSP) [104,105]. In fact, the first ACO algorithm, ant system (AS) [4,5,106,107], was first tested using this problem. Although AS could not compete with state-of-the-art algorithms for the TSP, it was the starting point for the development of various high performing ACO algorithms. The application of AS to the TSP also stimulated the application of ACO to other routing and combinatorial problems.

For instance, ACO has obtained very good results for the sequential ordering problem, an extension of asymmetric TSP with precedence constraints among nodes. At the time it was proposed by Gambardella and Dorigo [18], the algorithm was the best available algorithm for this problem, improving upon many best-known solutions. Recently, stochastic sampling has been integrated into a Beam-ACO algorithm for the TSP with time windows [19], which is an extension of the classical TSP with time window constraints; Beam-ACO is a combination of ACO algorithms with beam-search [32].

Table 1. Applications of ACO Algorithms to $\mathcal{NP}$ -hard Problems

Problem Type	Problem Name	References
Routing	Traveling salesman	Dorigo <i>et al.</i> [4,5] Dorigo and Gambardella [6] Stützle and Hoos [7,8]
	Vehicle routing (VRP)	Bullnheimer <i>et al.</i> [9] Reimann <i>et al.</i> [10] Rizzoli <i>et al.</i> [11]
	VRP with time windows	Gambardella <i>et al.</i> [12]
	VRPMTWMV	Favoretto <i>et al.</i> [13]
	VRP with loading constraints	Doerner <i>et al.</i> [14] Fuellerer <i>et al.</i> [15,16]
	Team orienteering	Ke <i>et al.</i> [17]
	Sequential ordering	Gambardella and Dorigo [18]
	TSP with time windows	López-Ibáñez and Blum [19]
	Single machine	Den Besten <i>et al.</i> [20] Merkle and Middendorf [21,22] Meyer and Ernst [23] Liao and Juan [24] Meyer [25]
	Flow shop	Stützle [26] Rajendran and Ziegler [27]
Scheduling	Industrial scheduling	Gravel <i>et al.</i> [28]
	Project scheduling	Merkle <i>et al.</i> [29]
	Group shop	Blum [30]
	Job shop	Blum [30] Huang and Liao [31]
	Open shop	Blum [32]
	Car sequencing	Khichane <i>et al.</i> [33] Solnon [34] Morin <i>et al.</i> [35]
Subset	Multiple knapsack	Leguizamón and Michalewicz [36] Ke <i>et al.</i> [37]
	Maximum independent set	Leguizamón and Michalewicz [36]
	Redundancy allocation	Liang and Smith [38]
	Weight constraint graph tree partitioning	Cordone and Maffioli [39]
	Bin packing	Levine and Ducatelle [40]
	Set covering	Lessing <i>et al.</i> [41]
	Set packing	Gandibleux <i>et al.</i> [42]
	l -cardinality trees	Blum and Blesa [43]
	Capacitated minimum spanning tree	Reimann and Laumanns [44]
	Maximum clique	Solnon and Fenet [45]
Assignment and layout	Multilevel lot-sizing	Pitakaso <i>et al.</i> [46,47] Almeder [48]
	Edge-disjoint paths	Blesa and Blum [49]
	Feature selection	Sivagaminathan and Ramakrishnan [50]
	Multicasting ad-hoc networks	Hernández and Blum [51]
	Quadratic assignment	Maniezzo <i>et al.</i> [52,53] Stützle and Hoos [8]
	Graph coloring	Costa and Hertz [54]
	Generalized assignment	Lourenço and Serra [55]
	Frequency assignment	Maniezzo and Carbonaro [56]

Table 1. (Continued)

Problem Type	Problem Name	References
	Constraint satisfaction	Solnon [57,58]
	Course timetabling	Socha <i>et al.</i> [59,60]
	Ambulance location	Doerner <i>et al.</i> [61]
	MAX-SAT	Pinto <i>et al.</i> [62]
	Assembly line balancing	Bautista and Pereora [63]
	Simple assembly line balancing	Blum [64]
	Supply chain management	Silva <i>et al.</i> [65]
Machine learning	Bayesian networks	De Campos <i>et al.</i> [66,67]
	Classification rules	Pinto <i>et al.</i> [68] Parpinelli <i>et al.</i> [69] Martens <i>et al.</i> [70] Otero <i>et al.</i> [71]
Bioinformatics	Shortest common supersequence	Michel and Middendorf [72,73]
	Protein folding	Shmygelska and Hoos [74]
	Docking	Korb <i>et al.</i> [75,76]
	Peak selection in biomarker identification	Ressom <i>et al.</i> [77]
	DNA sequencing	Blum <i>et al.</i> [78]
	Haplotype inference	Benedettini <i>et al.</i> [79]

Table 2. Applications of ACO Algorithms to “Nonstandard” Problems

Problem Type	Problem Name	References
Multiobjective	Scheduling	Iredi <i>et al.</i> [80]
	Portfolio selection	Doerner <i>et al.</i> [81,82]
	Quadratic assignment	López-Ibáñez <i>et al.</i> [83,84]
	Knapsack	Alaya <i>et al.</i> [85]
	Traveling salesman	García-Martínez <i>et al.</i> [86]
	Activity crashing	Doerner <i>et al.</i> [87]
	Orienteering	Schilde <i>et al.</i> [88]
Continuous	Neural networks	Socha and Blum [89]
	Test problems	Socha and Dorigo [90]
Stochastic	Probabilistic TSP	Bianchi <i>et al.</i> [91]
		Bianchi and Gambardella [92]
		Balaprakash <i>et al.</i> [93]
Dynamic	Vehicle routing	Bianchi <i>et al.</i> [94]
	Screening policies	Brailsford <i>et al.</i> [95]
	Network routing	Di Caro and Dorigo [96]
		Di Caro <i>et al.</i> [97]
	Dynamic TSP	Guntsch and Middendorf [98,99]
		Eyckelhof and Snoek [100]
		Sammound <i>et al.</i> [101]
	Vehicle routing	Montemanni <i>et al.</i> [102]
		Donati <i>et al.</i> [103]

ACO algorithms have been successful in tackling various variants of the vehicle routing problem (VRP). The first application of ACO to the capacitated VRP (CVRP) was

due to Bullnheimer *et al.* [9]. More recently, Reimann *et al.* [10] proposed a particular ACO algorithm (D-Ants) for the capacitated VRP. Gambardella *et al.* [12] introduced

MACS-VRPTW, an ACO algorithm for the VRP with time window (VRPTW) constraints, which reached state-of-the-art results when it was proposed. Favaretto *et al.* [13] proposed an ACS algorithm for a variant of the VRP with multiple time windows and multiple visits (VRPMTWMV). Fullerer *et al.* [15] used an ACO algorithm for a problem that combines the two-dimensional packing and the capacitated vehicle routing problem, showing that it outperforms a tabu search (TS) algorithm. In this problem, items of different sizes and weights are loaded in vehicles with a limited weight capacity and limited two-dimensional loading surface, and then they are distributed to the customers. Other variants of VRP with different loading constraints have also been tackled by means of ACO [14,16].

Ke *et al.* [17] have recently proposed an ACO approach to the team orienteering problem (TOP), where the goal is to find the set of paths from a starting point to an ending point that maximizes the reward obtained by visiting certain locations taking into account that there are restrictions on the length of each path.

Scheduling Problems

Scheduling problems concern the assignment of jobs to one or various machines over time. Input data for these problems are processing times but also often additional setup times, release dates and due dates of jobs, measures for the jobs' importance, and precedence constraints among jobs. Scheduling problems have been an important application area of ACO algorithms, and the currently available ACO applications in scheduling deal with many different job and machine characteristics.

The single-machine total weighted tardiness problem (SMTWTP) has been tackled by both den Besten *et al.* [20] and Merkle and Middendorf [21,22] using variants of ACS (ACS-SMTWTP). In ACS-SMTWTP, a solution is determined by a sequence of jobs. The positions of the sequence are filled in their canonical order, that is, first a job is assigned to position 1, next a job to position 2, and so on, until position n . Pheromone trails

are defined as the desirability of scheduling job j at position i , a pheromone trail definition that is used in many ACO applications to scheduling problems [20,26,108,109]. Merkle and Middendorf [21] used sophisticated heuristic information and an algorithmic technique called *pheromone summation rule*, which has proven to be useful in many applications of ACO to scheduling problems. On the other hand, den Besten *et al.* [20] combined ACS-SMTWTP with a powerful local search algorithm, resulting in one of the best algorithms available for this problem in terms of solution quality. Another application of ACO to a variant of this problem with sequence-dependent setup times has recently been studied by Liao and Juan [24]. Meyer and Ernst [23] and Meyer [25] studied the integration of constraint programming techniques into ACO algorithms using a single-machine problem with sequence-dependent setup times, release dates, and deadlines for jobs, as a case study.

ACO algorithms have also been proposed for the permutation flow-shop problem (FSP). The first approach is due to Stützle [26], who proposed a hybrid between $\mathcal{MMAS}$ and ACS. Later, Rajendran and Ziegler [27] improved its performance by introducing the pheromone summation rule. For this problem, however, the results of existing ACO algorithms are behind the current state-of-the-art algorithms. This is also the case for the well-known job-shop problem [30], although recent results hybridizing ACO and TS seem promising [31]. Nevertheless, for various other scheduling problems ACO algorithms are among the best performing algorithms available nowadays. Beam-ACO, the hybrid between beam search and ACO, is a state-of-the-art algorithm for open shop scheduling [32]. In addition, a variant of $\mathcal{MMAS}$ obtained excellent results in the group shop problem [30].

Another scheduling problem where ACO obtained excellent results is the resource-constrained project scheduling problem, in which a set of activities must be scheduled, subject to resource constraints and precedence constraints among the activities, such that the last activity is completed as early as possible. At the time of its publication, the

ACO algorithm proposed by Merkle *et al.* [29] was the best available.

Finally, state-of-the-art results have been obtained in the car sequencing problem by the ACO algorithm proposed by Solnon [34], and these results have been further improved by Morin *et al.* [35] by means of a specialized pheromone model. The car sequencing problem has also been used as an example application by Khichane *et al.* [33] to explore the integration of constraint programming techniques into ACO algorithms.

Subset Problems

The goal in subset problems is, generally speaking, to find a subset of the available items that minimizes a cost function defined over the items and that satisfies a number of constraints. This is a wide definition that can include other classes of problems. There are, however, two characteristic properties of the solutions to subset problems: The order of the solution components is irrelevant, and the number of components of a solution may differ from solution to solution.

An important subset problem is the set covering problem (SCP). Lessing *et al.* [41] compared the performance of a number of ACO algorithms for the SCP, with and without the usage of a local search algorithm based on 3-flip neighborhoods [110]. The best performance results were obtained, as expected, when including local search. For a large number of instances, the computational results were competitive with state-of-the-art algorithms for the SCP.

Leguizamón and Michalewicz [36] proposed the first ACO applications to the multiple knapsack and to the maximum independent set problems, which were, however, not competitive with the state-of-the-art. Currently, the best performing ACO algorithm for the multiple knapsack problem is due to Ke *et al.* [37]. Levine and Ducatelle [40] adapted $\mathcal{MMAS}$ to the well-known bin-packing problem and compared its performance with the hybrid grouping genetic algorithm [111], and with Martello and Toth's reduction method [112]. The $\mathcal{MMAS}$ algorithm outperformed both, obtaining better solutions in a much shorter time. Solnon and Fener [45] carried out a comprehensive

study for the maximum clique problem. Their conclusion was that ACO combined with appropriate local search can match the quality of state-of-the-art heuristics. Blesa and Blum [49] applied ACO to the problem of finding edge-disjoint paths in networks, and found the performance of the proposed ACO superior in terms of both solution quality and computation time when compared with a multistart greedy algorithm. Another interesting application is the work of Sivagaminathan and Ramakrishnan [50], which discusses how ACO may be hybridized with neural networks for optimizing feature selection in multivariate analysis.

Cordone and Maffioli [39] introduced the weight constrained graph tree partition problem, and tested different variants of ACS with and without local search. Blum and Blesa [43] tackled the edge-weighted k -cardinality tree problem (or k -minimum spanning tree), where the goal is to find a tree over a graph with exactly k edges minimizing the sum of the weights. They compared a $\mathcal{MMAS}$ variant, TS, and an evolutionary algorithm. Their results showed that none of the approaches was superior to the others in all instance classes tested, and that $\mathcal{MMAS}$ was better suited for instances where the value of k was much smaller than the number of vertices.

A subset problem closely related to the CVRP is the capacitated minimum spanning tree problem, which has been effectively tackled by a hybrid ACO algorithm [44] based on a previous ACO algorithm for the CVRP [10]. More recently, Hernández and Blum [51] considered the minimization of power consumption when multicasting in static wireless ad-hoc networks. This problem can be stated as an $\mathcal{NP}$ -hard combinatorial problem, where the goal is to find a directed tree over the network of nodes. Their proposed ACO algorithm outperforms existing algorithms for several variants of this problem.

Finally, a class of problems for which ACO has recently shown competitive results is that of multilevel lot-sizing with [46,48] and without capacity constraints [47]. In these problems, a subset of items is scheduled for production at each time interval, and the goal is to minimize the cost of producing the items,

taking into account several constraints and relations between the items.

Assignment and Layout Problems

In assignment problems, a set of items has to be assigned to a given number of resources subject to some constraints. Probably, the most widely studied example is the quadratic assignment problem (QAP), which was among the first problems tackled by ACO algorithms [5,52,53]. Various high-performing ACO algorithms for the QAP have followed this initial work. Among them is the approximate nondeterministic tree search (ANTS) algorithm by Maniezzo [113] a combination of ACO with tree search techniques involving the usage of lower bounds to rate solution components and to prune extensions of partial solutions. The computational results of ANTS on the QAP were very promising. Another high-performing ACO algorithm is the *MAX-MIN* ant system (*MMAS*) proposed by Stützle and Hoos [8], which is among the best algorithms available for large, structured instances of the QAP.

The ANTS algorithm has also been applied to the frequency assignment problem (FAP), in which frequencies have to be assigned to links and there are constraints on the minimum distance between the frequencies assigned to each pair of links. ANTS showed good performance on some classes of FAP instances in comparison with other approaches [56]. Other applications of ACO to assignment problems include university course timetabling [59,60] and graph coloring [54]. The work of Solnon [57,58] applies ACO algorithms to the general class of constraint satisfaction problems (CSPs); in fact, decision variants of problems such as graph coloring and frequency assignment can be seen as cases of CSPs. Within this class, Pinto *et al.* [62] studied the application of ACO to regular and dynamic MAX-SAT problems.

Another notable example is the generalized assignment problem, where a set of tasks have to be assigned to a set of agents with a limited total capacity, minimizing the total assignment cost of tasks to agents. The *MMAS* algorithm proposed by Lourenço and

Serra [55] was, at the time of its publication, close to the state-of-the-art algorithm for this problem. More recently, Doerner *et al.* [61] tackled a real-world problem related to ambulance locations in Austria by means of an ACO algorithm; and Blum [64] has shown that the hybrid between beam search and ACO, Beam-ACO, is a state-of-the-art algorithm for simple assembly line balancing. In the section titled “Industrial Applications,” we mention an industrial application of ACO to assembly line balancing. Finally, Silva *et al.* [65] have used ACO for a complex supply chain management problem that combines aspects of the generalized assignment, scheduling, and vehicle routing problems.

Machine Learning Problems

Diverse problems in the field of machine learning have been tackled by means of ACO algorithms. Notable examples are the work of Parpinelli *et al.* [69] and Martens *et al.* [70] on applying ACO to the problem of learning classification rules. This work was later extended by Otero *et al.* [71] in order to handle continuous attributes. De Campos *et al.* [66,67] adapted Ant Colony System for the problem of learning the structure of Bayesian networks, and Pinto *et al.* [68] have recently extended this work. Finally, the work of Socha and Blum [89] for training neural networks by means of ACO is also an example of the application of ACO algorithms to continuous problems.

Bioinformatics Problems

Computer applications to molecular biology (bioinformatics) have originated many $\mathcal{NP}$ -hard combinatorial optimization problems. We include in this section general problems that have attracted considerable interest due to their applications to bioinformatics. This is the case of the shortest common super-sequence problem (SCSP), which is a well-known $\mathcal{NP}$ -hard problem with applications in DNA analysis. Michel and Middendorf [72,73] proposed an ACO algorithm for the SCSP, obtaining state-of-the-art results, in particular, for structured instances that are typically found in real-world applications.

An important problem in bioinformatics is protein folding, that is, the prediction of a protein's structure based on its sequence of amino acids. A simplified model for protein folding is the two-dimensional hydrophobic-polar protein folding problem [114]. Shmygelska and Hoos [74] have successfully applied ACO to this problem and its three-dimensional variant. The performance of the resulting ACO algorithm is comparable to the best existing specialized algorithms for these problems.

Interesting is also the work of Blum *et al.* [78], where they propose a *multilevel framework* based on ACO for the problem of DNA sequencing by hybridization. An earlier proposal of multilevel ACO frameworks is due to Korošec *et al.* [115]. Multilevel techniques [116,117] solve a hierarchy of successively smaller versions of the original problem instance. The solutions obtained at the lowest level of the hierarchy are transformed into solutions for the next higher level, and improved by an optimization algorithm, such as an ACO algorithm.

Other problems in bioinformatics have been successfully tackled by means of ACO algorithms: Korb *et al.* [75,76] considered the flexible protein–ligand docking problem, for which the proposed ACO algorithm reaches state-of-the-art performance, and Benedettini *et al.* [79] recently studied the problem of haplotype inference under pure parsimony. ACO algorithms are sometimes hybridized with Machine Learning techniques. An example is the recent work of Ressom *et al.* [77] on a selection problem in biomarker identification, which combines ACO with support vector machines.

APPLICATIONS TO PROBLEMS WITH NONSTANDARD FEATURES

We review in this section applications of ACO algorithms to problems having additional characteristics such as multiple objective functions, time-varying data, and stochastic information about objective values or constraints. In addition, we mention applications of ACO to network routing and continuous optimization problems.

Multiobjective Optimization

In many real-world problems, candidate solutions are evaluated according to multiple, often conflicting objectives. Sometimes the importance of each objective can be exactly weighted, and hence objectives can be combined into a single scalar value by using, for example, a weighted sum. This is the approach used by Doerner *et al.* [118] for a biobjective transportation problem. In other cases, objectives can be ordered by their relative importance in a lexicographical manner. Gambardella *et al.* [12] proposed a two-colony ACS algorithm for the vehicle routing problem with time windows, where the first colony improves the primary objective and the second colony tries to improve the secondary objective while not worsening the primary one.

When there is no *a priori* knowledge about the relative importance of objectives, the goal usually becomes to approximate the set of Pareto-optimal solutions—a solution is Pareto optimal if no other solution is better or equal for all objectives and strictly better in at least one objective. Iredi *et al.* [80] were among the first to discuss various alternatives for extending ACO to multiobjective problems in terms of Pareto-optimality. They also tested a few of the proposed variants on a biobjective scheduling problem. Another early work is the application of ACO to multiobjective portfolio problems by Doerner *et al.* [81,82]. Later studies have proposed and tested various combinations of alternative ACO algorithms for multiobjective variants of the QAP [83,84], the knapsack problem [85], activity crashing [87], and the biobjective orienteering problem [88]. García-Martínez *et al.* [86] reviewed existing multiobjective ACO algorithms and carried out an experimental evaluation of several ACO variants using the bicriteria TSP as a case study. Angus and Woodward [119] give another detailed overview of available multiobjective ACO algorithms.

Stochastic Optimization Problems

In stochastic optimization problems, data are not known exactly before generating a solution. Rather, because of uncertainty,

noise, approximation, or other factors, what is available is stochastic information on the objective function value(s), on the decision variable values, or on the constraint boundaries.

The first application of ACO algorithms to stochastic problems was to the probabilistic TSP (PTSP). In the PTSP, each city has associated a probability of requiring a visit, and the goal is to find an *a priori* tour of minimal expected length over all cities. Bianchi *et al.* [91] and Bianchi and Gambardella [92] proposed an adaptation of ACS for the PTSP. Very recently, this algorithm was improved by Balaprakash *et al.* [93], resulting in a state-of-the-art algorithm for the PTSP. Other applications of ACO to stochastic problems include vehicle routing problems with uncertain demands [94], and the selection of optimal screening policies for diabetic retinopathy [95]. The latter approach builds on the S-ACO algorithm proposed earlier by Gutjahr [120].

Dynamic Optimization Problems

Dynamic optimization problems are those whose characteristics change while being solved. ACO algorithms have been applied to such versions of classical $\mathcal{NP}$ -hard problems. Notable examples are applications to dynamic versions of the TSP, where the distances between cities may change or where cities may appear or disappear [98–101]. More recently, Montemanni *et al.* [102] and Donati *et al.* [103] discuss applications of ACS to dynamic vehicle routing problems, reporting good results on both artificial and real-world instances of the problem. Other notable examples of dynamic problems are routing problems in communication networks, which are discussed in the following section.

Communication Network Problems

Some system properties in telecommunication networks, such as the availability of links or the cost of traversing links, are time-varying. The application of ACO algorithms to routing problems in such networks is among the main success stories in ACO. One of the first applications by Schoonderwoerd *et al.* [121] concerned routing in

circuit-switched networks, such as classical telephone networks. The proposed algorithm, called *ABC*, was demonstrated on a simulated version of the British Telecom network.

A very successful application of ACO to dynamic network routing is the AntNet algorithm, proposed by Di Caro and Dorigo [96,122]. AntNet was applied to routing in packet-switched networks, such as the Internet. Experimental studies compared AntNet with many state-of-the-art algorithms on a large set of benchmark problems under a variety of traffic conditions [96]. AntNet proved to be very robust against varying traffic conditions and parameter settings, and it always outperformed competing approaches.

Several other routing algorithms based on ACO have been proposed for a variety of wired network scenarios [123,124]. More recent applications of these strategies deal with the challenging class of mobile ad hoc networks (MANETs). Because of the specific characteristics of MANETs (very high dynamics and link asymmetry), the straightforward application of the ACO algorithms developed for wired networks has proven unsuccessful [125]. Nonetheless, an extension of AntNet that is competitive with state-of-the-art routing algorithms for MANETs has been proposed by Ducatelle *et al.* [97, 126]. For recent, in-depth reviews of applications of ACO to dynamic network routing problems, we refer to Refs 127 and 128.

Continuous Optimization Problems

Continuous optimization problems arise in a large number of engineering applications. Their main difference from combinatorial problems, which were the exclusive application field of ACO in the early research efforts, is that decision variables in such problems have a continuous, real-valued domain. Recently, various proposals have been made on how to handle continuous decision variables within the ACO framework [129–131]. In the continuous ACO algorithm proposed by Socha and Dorigo [90], probability density functions, explicitly represented by Gaussian kernel functions, correspond to the pheromone models. Extensions of this approach also exist for mixed-variable—continuous

and discrete—problems [132]. A notable application of ACO algorithms for continuous optimization is the training of feed-forward neural networks [89]. Interestingly, there exist also successful applications of ACO to continuous problems that discretize the real-valued domain of the variables. An example is the PLANTS algorithm for the protein–ligand docking problem [76], which combines a discrete ACO algorithm with a local search that works on the continuous domain of the variables.

Industrial Applications

While most research is done on academic applications, commercial companies have started to use ACO algorithms for real-world applications [11]. The company AntOptima (www.antoptima.com) develops and markets ACO-based solution methods for tackling industrial vehicle routing problems. Features common to real-world applications are time-varying data, multiple objectives, or the availability of stochastic information about events or data. Moreover, engineering problems often do not have a mathematical formulation in the traditional sense. Rather, algorithms have to rely on an external *simulator* to evaluate the quality and feasibility of candidate solutions. Examples of applications of ACO relying on simulation are the design [133] and operation [134] of water distribution networks. Other interesting real-world applications are those of Gravel, Price and Gagné [28], who applied ACO to an industrial scheduling problem in an aluminum casting center, and those of Bautista and Pereira [63,135,136], who successfully applied ACO to solve an assembly line balancing problem for a bike line assembly.

CONCLUSIONS

Nowadays, ACO is a well-established metaheuristic applied to a wide range of optimization problems and with hundreds of successful implementations. Several of these implementations have shown to be, at least at the time of their publication, the state-of-the-art for the respective problems tackled,

including problems such as vehicle routing, sequential ordering, quadratic assignment, assembly line balancing, open-shop scheduling, and various others. Applications of ACO to dynamic routing problems in telecommunication networks have been particularly successful, probably because several algorithm characteristics match well with the features of the applications.

By analyzing the many available ACO implementations, one can identify ingredients necessary for the successful application of ACO. Firstly, an effective mechanism for iteratively constructing solutions must be available. Ideally, this construction mechanism exploits problem-specific knowledge by using appropriate heuristic information. Secondly, the best performing ACO algorithms have specialized features that allow to carefully control the balance between the exploration of new solutions and the intensification of the search around the best solutions. Such control mechanisms are offered by advanced ACO algorithms such as ACS or MMAS. In fact, the original AC has been abandoned by now in favor of better performing variants. Thirdly, the usage of local search algorithms for improving the solutions constructed by the ants is very successful in practice. Finally, the integration of other techniques such as constraint programming, tree search techniques, or multilevel frameworks often yields a further improvement in performance or increases the robustness of the algorithms.

Further information on ACO and related topics can be obtained by subscribing to the moderated mailing list *aco-list*, and by visiting the ACO web page (www.aco-metaheuristic.org).

Acknowledgments

This work was supported by the META-X project, an *Action de Recherche Concertée* funded by the Scientific Research Directorate of the French Community of Belgium. Marco Dorigo and Thomas Stützle acknowledge support from the Belgian F.R.S.-FNRS, of which they are Research Director and Research Associate, respectively.

REFERENCES

1. Dorigo M, Di Caro G. The Ant Colony optimization meta-heuristic. In: Corne D, Dorigo M, Glover F, editors. *New ideas in optimization*. London: McGraw Hill; 1999. pp. 11–32.
2. Dorigo M, Di Caro G, Gambardella LM. Ant algorithms for discrete optimization. *Artif Life* 1999;5(2):137–172.
3. Dorigo M, Stützle T. *Ant colony optimization*. Cambridge (MA): MIT Press; 2004. pp. 305.
4. Dorigo M, Maniezzo V, Colorni A. Positive feedback as a search strategy. Italy: Dipartimento di Elettronica, Politecnico di Milano; 1991. Report nr 91–016.
5. Dorigo M, Maniezzo V, Colorni A. Ant System: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern Part B* 1996;26(1):29–41.
6. Dorigo M, Gambardella LM. Ant Colony System: a cooperative learning approach to the traveling salesman problem. *IEEE Trans Evol Comput* 1997;1(1):53–66.
7. Stützle T, Hoos HH. The $MAX-MIN$ Ant System and local search for the traveling salesman problem. In: Bäck T, Michalewicz Z, Yao X, editors. *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*. Piscataway (NJ): IEEE Press; 1997. pp. 309–314.
8. Stützle T, Hoos HH. $MAX-MIN$ Ant System. *Future Gener Comput Syst* 2000;16(8):889–914.
9. Bullnheimer B, Hartl RF, Strauss C. An improved ant system algorithm for the vehicle routing problem. *Ann Oper Res* 1999;89:319–328.
10. Reimann M, Doerner KF, Hartl RF. D-ants: Savings based ants divide and conquer the vehicle routing problems. *Comput Oper Res* 2004; 31(4):563–591.
11. Rizzoli AE, Montemanni R, Lucibello E, *et al.* Ant colony optimization for real-world vehicle routing problems: from theory to applications. *Swarm Intell* 2007;1(2):135–151.
12. Gambardella LM, Taillard ED, Agazzi G. MACS-VRPTW: a multiple ant colony system for vehicle routing problems with time windows. In: Corne D, Dorigo M, Glover F, editors. *New ideas in optimization*. London: McGraw Hill; 1999. pp. 63–76.
13. Favaretto D, Moretti E, Pellegrini P. Ant colony system for a VRP with multiple time windows and multiple visits. *J Interdiscipl Math* 2007;10(2):263–284.
14. Doerner KF, Fuellerer G, Gronalt M, *et al.* Metaheuristics for the vehicle routing problem with loading constraints. *Networks* 2006;49(4):294–307.
15. Fuellerer G, Doerner KF, Hartl RF, *et al.* Ant colony optimization for the two-dimensional loading vehicle routing problem. *Comput Oper Res* 2009;36(3):655–673.
16. Fuellerer G, Doerner KF, Hartl RF, *et al.* Metaheuristics for vehicle routing problems with three-dimensional loading constraints. *Eur J Oper Res* 2009;201(3):751–759.
17. Ke L, Archetti C, Feng Z. Ants can solve the team orienteering problem. *Comput Ind Eng* 2008;54(3):648–665.
18. Gambardella LM, Dorigo M. Ant Colony System hybridized with a new local search for the sequential ordering problem. *INFORMS J Comput* 2000;12(3):237–255.
19. López-Ibáñez M, Blum C. Beam-ACO for the travelling salesman problem with time windows. *Comput Oper Res* 2010;37(9):1570–1583.
20. den Besten ML, Stützle T, Dorigo M. Ant colony optimization for the total weighted tardiness problem. In: Schoenauer M, *et al.*, editors. Volume 1917, *Proceedings of PPSN-VI, 6th International Conference on Parallel Problem Solving from Nature, Lecture Notes in Computer Science*. Heidelberg: Springer; 2000. pp. 611–620.
21. Merkle D, Middendorf M. An ant algorithm with a new pheromone evaluation rule for total tardiness problems. In: Cagnoni S, *et al.*, editors. Volume 1803, *Real-world applications of evolutionary computing, Lecture Notes in Computer Science*. Heidelberg: Springer; 2000. pp. 287–296.
22. Merkle D, Middendorf M. Ant colony optimization with global pheromone evaluation for scheduling a single machine. *Appl Intell* 2003;18(1):105–111.
23. Meyer B, Ernst AT. Integrating ACO and constraint propagation. In: Dorigo M, *et al.*, editors. Volume 3172, *Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science*. Heidelberg: Springer; 2004. pp. 166–177.
24. Liao CJ, Juan HC. An ant colony optimization for single-machine tardiness scheduling with sequence-dependent setups. *Comput Oper Res* 2007;34(7):1899–1909.
25. Meyer B. Hybrids of constructive metaheuristics and constraint programming. In:

- Blum C, Blesa MJ, Roli A, *et al.*, editors. Volume 117, Hybrid metaheuristics—an emergent approach to optimization: studies in computational intelligence. Berlin: Springer; 2008. pp. 85–116.
26. Stützle T. An ant approach to the flow shop problem. In: Volume 3, Proceedings of the 6th European Congress on Intelligent Techniques & Soft Computing (EUFIT'98). Aachen: Verlag Mainz; 1998. pp. 1560–1564.
 27. Rajendran C, Ziegler H. Ant-colony algorithms for permutation flowshop scheduling to minimize makespan/total flowtime of jobs. *Eur J Oper Res* 2004;155(2):426–438.
 28. Gravel M, Price WL, Gagné C. Scheduling continuous casting of aluminum using a multiple objective ant colony optimization metaheuristic. *Eur J Oper Res* 2002;143: 218–229.
 29. Merkle D, Middendorf M, Schneck H. Ant colony optimization for resource-constrained project scheduling. *IEEE Trans Evol Comput* 2002;6(4):333–346.
 30. Blum C. Theoretical and practical aspects of ant colony optimization [PhD Thesis]. Brussels, Belgium: IRIDIA, Université Libre de Bruxelles; 2004.
 31. Huang KL, Liao CJ. Ant colony optimization combined with taboo search for the job shop scheduling problem. *Comput Oper Res* 2008;35(4):1030–1046.
 32. Blum C. Beam-ACO—Hybridizing ant colony optimization with beam search: an application to open shop scheduling. *Comput Oper Res* 2005;32(6):1565–1591.
 33. Khichane M, Albert P, Solnon C. Integration of ACO in a constraint programming language. In: Dorigo M, *et al.*, editors. Volume 5217, Ant Colony Optimization and Swarm Intelligence: 6th International Conference, ANTS 2008, Lecture Notes in Computer Science. Heidelberg: Springer; 2008. pp. 84–95.
 34. Solnon C. Combining two pheromone structures for solving the car sequencing problem with ant colony optimization. *Eur J Oper Res* 2008;191(3):1043–1055.
 35. Morin S, Gagné C, Gravel M. Ant colony optimization with a specialized pheromone trail for the car-sequencing problem. *Eur J Oper Res* 2009;197(3):1185–1191.
 36. Leguizamón G, Michalewicz Z. A new version of Ant System for subset problems. In: Proceedings of the 1999 Congress on Evolutionary Computation (CEC'99). Piscataway (NJ): IEEE Press; 1999. pp. 1459–1464.
 37. Ke L, Feng Z, Ren Z, *et al.* An ant colony optimization approach for the multidimensional knapsack problem. *J Heuristics* 2010; 16(1):65–83.
 38. Liang YC, Smith AE. An Ant System approach to redundancy allocation. In: Proceedings of the 1999 Congress on Evolutionary Computation (CEC'99). Piscataway (NJ): IEEE Press; 1999. pp. 1478–1484.
 39. Cordone R, Maffioli F. Coloured Ant System and local search to design local telecommunication networks. In: Boers EJW, *et al.*, editors. Volume 2037, Applications of Evolutionary Computing, Proceedings of EvoWorkshops 2001, Lecture Notes in Computer Science. Heidelberg: Springer; 2001. pp. 60–69.
 40. Levine J, Ducatelle F. Ant colony optimisation and local search for bin packing and cutting stock problems. *J Oper Res Soc* 2003; 55(7):705–716.
 41. Lessing L, Dumitrescu I, Stützle T. A comparison between ACO algorithms for the set covering problem. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 1–12.
 42. Gandibleux X, Delorme X, T'Kindt V. An ant colony optimisation algorithm for the set packing problem. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 49–60.
 43. Blum C, Blesa MJ. New metaheuristic approaches for the edge-weighted k-cardinality tree problem. *Comput Oper Res* 2005;32(6):1355–1377.
 44. Reimann M, Laumanns M. Savings based ant colony optimization for the capacitated minimum spanning tree problem. *Comput Oper Res* 2006;33(6):1794–1822.
 45. Solnon C, Fenet S. A study of ACO capabilities for solving the maximum clique problem. *J Heuristics* 2006;12(3):155–180.
 46. Pitakaso R, Almeder C, Doerner KF, *et al.* Combining exact and population-based methods for the constrained multilevel lot sizing problem. *Int J Prod Res* 2006;44(22): 4755–4771.

47. Pitakaso R, Almeder C, Doerner KF, *et al.* A *MAX-MIN* Ant System for unconstrained multi-level lot-sizing problems. *Comput Oper Res* 2007;34(9):2533–2552.
48. Almeder C. A hybrid optimization approach for multi-level capacitated lot-sizing problems. *Eur J Oper Res* 2010;200(2):599–606.
49. Blesa MJ, Blum C. Finding edge-disjoint paths in networks by means of artificial ant colonies. *J Math Model Algorithms* 2007; 6(3):361–391.
50. Sivagaminathan RK, Ramakrishnan S. A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Syst Appl* 2007; 33(1):49–60.
51. Hernández H, Blum C. Ant colony optimization for multicasting in static wireless ad-hoc networks. *Swarm Intell* 2009;3(2):125–148.
52. Maniezzo V, Colorni A, Dorigo M. The Ant System applied to the quadratic assignment problem, Belgium: IRIDIA, Université Libre de Bruxelles; 1994.IRIDIA/94-28.
53. Maniezzo V, Colorni A. The Ant System applied to the quadratic assignment problem. *IEEE Trans Data Knowl Eng* 1999; 11(5):769–778.
54. Costa D, Hertz A. Ants can colour graphs. *J Oper Res Soc* 1997;48:295–305.
55. Lourenço H, Serra D. Adaptive approach heuristics for the generalized assignment problem, Economic Working Papers Series No. 304. Barcelona: Universitat Pompeu Fabra, Department of Economics and Management; 1998.
56. Maniezzo V, Carbonaro A. An ANTS heuristic for the frequency assignment problem. *Future Gener Comput Syst* 2000;16(8): 927–935.
57. Solnon C. Solving permutation constraint satisfaction problems with artificial ants. In: Horn W, editor. *Proceedings of the 14th European Conference on Artificial Intelligence*. Amsterdam, The Netherlands: IOS Press; 2000. pp. 118–122.
58. Solnon C. Ants can solve constraint satisfaction problems. *IEEE Trans Evol Comput* 2002;6(4):347–357.
59. Socha K, Knowles J, Sampels M. A *MAX-MIN* Ant System for the university course timetabling problem. In: Dorigo M, *et al.*, editors. Volume 2463, *Ant Algorithms: 3rd International Workshop, ANTS 2002, Lecture Notes in Computer Science*. Heidelberg: Springer; 2002. pp. 1–13.
60. Socha K, Sampels M, Manfrin M. Ant algorithms for the university course timetabling problem with regard to the state-of-the-art. In: Raidl GR, *et al.*, editors. Volume 2611, *Applications of Evolutionary Computing, Proceedings of EvoWorkshops 2003, Lecture Notes in Computer Science*. Heidelberg: Springer; 2003. pp. 334–345.
61. Doerner KF, Gutjahr WJ, Hartl RF, *et al.* Heuristic solution of an extended double-coverage ambulance location problem for Austria. *Cent Eur J Oper Res* 2005; 13(4):325–340.
62. Pinto P, Runkler T, Sousa J. Ant colony optimization and its application to regular and dynamic MAX-SAT problems. Volume 69, *Advances in biologically inspired information systems, Studies in Computational Intelligence*. Berlin: Springer; 2007. pp. 285–304.
63. Bautista J, Pereira J. Ant algorithms for a time and space constrained assembly line balancing problem. *Eur J Oper Res* 2007; 177(3):2016–2032.
64. Blum C. Beam-ACO for simple assembly line balancing. *INFORMS J Comput* 2008;20(4):618–627.
65. Silva CA, Sousa JMC, Runkler TA, *et al.* Distributed supply chain management using ant colony optimization. *Eur J Oper Res* 2009;199(2):349–358.
66. de Campos LM, Fernández-Luna JM, Gámez JA, *et al.* Ant colony optimization for learning Bayesian networks. *Int J Approx Reason* 2002;31(3):291–311.
67. de Campos LM, Gamez JA, Puerta JM. Learning Bayesian networks by ant colony optimisation: searching in the space of orderings. *Mathware Soft Comput* 2002; 9(2–3):251–268.
68. Pinto PC, Nägele A, DeJori M, *et al.* Using a local discovery ant algorithm for Bayesian network structure learning. *IEEE Trans Evol Comput* 2009;13(4):767–779.
69. Parpinelli RS, Lopes HS, Freitas AA. Data mining with an ant colony optimization algorithm. *IEEE Trans Evol Comput* 2002;6(4):321–332.
70. Martens D, De Backer M, Haesen R, *et al.* Classification with ant colony optimization. *IEEE Trans Evol Comput* 2007;11(5): 651–665.
71. Otero FEB, Freitas AA, Johnson CG. cAnt-Miner: an ant colony classification algorithm to cope with continuous attributes. In: Dorigo M, *et al.*, editors. Volume 5217, *Ant Colony*

- Optimization and Swarm Intelligence: 6th International Conference, ANTS 2008, Lecture Notes in Computer Science. Heidelberg: Springer; 2008. pp. 48–59.
72. Michel R, Middendorf M. An island model based Ant System with lookahead for the shortest supersequence problem. In: Eiben AE, *et al.*, editors. Volume 1498, Proceedings of PPSN-V, 5th International Conference on Parallel Problem Solving from Nature, Lecture Notes in Computer Science. Heidelberg: Springer; 1998. pp. 692–701.
 73. Michel R, Middendorf M. An ACO algorithm for the shortest supersequence problem. In: Corne D, Dorigo M, Glover F, editors. New ideas in optimization. London: McGraw Hill; 1999. pp. 51–61.
 74. Shmygelska A, Hoos HH. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics 2005;6:30.
 75. Korb O, Stützle T, Exner TE. PLANTS: application of ant colony optimization to structure-based drug design. In: Dorigo M, *et al.*, editors. Volume 4150, Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Lecture Notes in Computer Science. Heidelberg: Springer; 2006. pp. 247–258.
 76. Korb O, Stützle T, Exner TE. An ant colony optimization approach to flexible protein-ligand docking. Swarm Intell 2007; 1(2):115–134.
 77. Resson HW, Varghese RS, Drake SK, *et al.* Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics 2007;23(5):619–626.
 78. Blum C, Yábar Vallès M, Blesa MJ. An ant colony optimization algorithm for DNA sequencing by hybridization. Comput Oper Res 2008;35(11):3620–3635.
 79. Benedettini S, Roli A, Di Gaspero L. Two-level ACO for haplotype inference under pure parsimony. In: Dorigo M, *et al.*, editors. Volume 5217, Ant Colony Optimization and Swarm Intelligence: 6th International Conference, ANTS 2008, Lecture Notes in Computer Science. Heidelberg: Springer; 2008. pp. 179–190.
 80. Iredi S, Merkle D, Middendorf M. Bi-criterion optimization with multi colony ant algorithms. In: Zitzler E, Deb K, Thiele L, *et al.*, editors. Volume 1993, 1st International Conference on Evolutionary Multi-Criterion Optimization, (EMO'01), Lecture Notes in Computer Science. Heidelberg: Springer; 2001. pp. 359–372.
 81. Doerner KF, Gutjahr WJ, Hartl RF, *et al.* Ant colony optimization in multiobjective portfolio selection. In: Proceedings of the Fourth Metaheuristics International Conference; 2001. pp. 243–248.
 82. Doerner KF, Gutjahr WJ, Hartl RF, *et al.* Pareto ant colony optimization: a metaheuristic approach to multiobjective portfolio selection. Ann Oper Res 2004;131:79–99.
 83. López-Ibáñez M, Paquete L, Stützle T. On the design of ACO for the biobjective quadratic assignment problem. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 214–225.
 84. López-Ibáñez M, Paquete L, Stützle T. Hybrid population-based algorithms for the bi-objective quadratic assignment problem. J Math Model Algorithms 2006;5(1): 111–137.
 85. Alaya I, Solnon C, Ghédira K. Ant colony optimization for multi-objective optimization problems. Volume 1, 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). Los Alamitos (CA): IEEE Computer Society Press; 2007. pp. 450–457.
 86. García-Martínez C, Cordon O, Herrera F. A taxonomy and an empirical analysis of multiple objective ant colony optimization algorithms for the bi-criteria TSP. Eur J Oper Res 2007;180(1):116–148.
 87. Doerner KF, Gutjahr WJ, Hartl RF, *et al.* Nature-inspired metaheuristics in multi-objective activity crashing. Omega 2008; 36(6):1019–1037.
 88. Schilde M, Doerner KF, Hartl RF, *et al.* Metaheuristics for the bi-objective orienteering problem. Swarm Intell 2009; 3(3):179–201.
 89. Socha K, Blum C. An ant colony optimization algorithm for continuous optimization: An application to feed-forward neural network training. Neural Comput Appl 2007; 16(3):235–247.
 90. Socha K, Dorigo M. Ant colony optimization for continuous domains. Eur J Oper Res 2008;185(3):1155–1173.
 91. Bianchi L, Gambardella LM, Dorigo M. An ant colony optimization approach to the probabilistic traveling salesman problem.

- In: Merelo JJ, *et al.*, editors. Volume 2439, Proceedings of PPSN-VII, 7th International Conference on Parallel Problem Solving from Nature, Lecture Notes in Computer Science. Heidelberg: Springer; 2002. pp. 883–892.
92. Bianchi L, Gambardella LM. Ant colony optimization and local search based on exact and estimated objective values for the probabilistic traveling salesman problem, Manno: IDSIA; 2007. USI-SUPSI, IDSIA-06-07.
 93. Balaprakash P, Birattari M, Stützle T, *et al.* Estimation-based ant colony optimization algorithms for the probabilistic travelling salesman problem. *Swarm Intell* 2009;3(3):223–242.
 94. Bianchi L, Birattari M, Manfrin M, *et al.* Hybrid metaheuristics for the vehicle routing problem with stochastic demands. *J Math Model Algorithms* 2006;5(1):91–110.
 95. Brailsford SC, Gutjahr WJ, Rauner MS, *et al.* Combined discrete-event simulation and ant colony optimisation approach for selecting optimal screening policies for diabetic retinopathy. *Comput Manag Sci* 2006;4(1):59–83.
 96. Di Caro G, Dorigo M. AntNet: Distributed stigmergetic control for communications networks. *J Artif Intell Res* 1998;9:317–365.
 97. Di Caro G, Ducatelle F, Gambardella LM. AntHocNet: an adaptive nature-inspired algorithm for routing in mobile ad hoc networks. *Eur Trans Telecommun* 2005;16(5):443–455.
 98. Guntch M, Middendorf M. Pheromone modification strategies for ant algorithms applied to dynamic TSP. In: Boers EJW, *et al.*, editors. Volume 2037, Applications of Evolutionary Computing, Proceedings of EvoWorkshops 2001, Lecture Notes in Computer Science. Heidelberg: Springer; 2001. pp. 213–222.
 99. Guntch M, Middendorf M. A population based approach for ACO. In: Cagnoni S, *et al.*, editors. Volume 2279, Applications of Evolutionary Computing, Proceedings of EvoWorkshops 2002, Lecture Notes in Computer Science. Heidelberg: Springer; 2002. pp. 71–80.
 100. Eyckelhof CJ, Snoek M. Ant systems for a dynamic TSP: Ants caught in a traffic jam. In: Dorigo M, *et al.*, editors. Volume 2463, Ant Algorithms: 3rd International Workshop, ANTS 2002, Lecture Notes in Computer Science. Heidelberg: Springer; 2002. pp. 88–99.
 101. Sammoud O, Solnon C, Ghédira K. A new ACO approach for solving dynamic problems. In: 9th International Conference on Artificial Evolution (EA'09), Lecture Notes in Computer Science. Heidelberg: Springer, In press.
 102. Montemanni R, Gambardella LM, Rizzoli AE, *et al.* Ant colony system for a dynamic vehicle routing problem. *J Comb Optim* 2005;10:327–343.
 103. Donati AV, Montemanni R, Casagrande N, *et al.* Time dependent vehicle routing problem with a multi ant colony system. *Eur J Oper Res* 2008;185(3):1174–1191.
 104. Applegate D, Bixby RE, Chvátal V, *et al.* The traveling salesman problem: a computational study. Princeton (NJ): Princeton University Press; 2006.
 105. Lawler EL, Lenstra JK, Kan AHGR, *et al.* The travelling salesman problem. Chichester: John Wiley & Sons, Ltd.; 1985.
 106. Dorigo M. Optimization, Learning and Natural Algorithms [PhD thesis]. Italy: Dipartimento di Elettronica, Politecnico di Milano; 1992. In Italian.
 107. Dorigo M, Maniezzo V, Colnari A. The Ant System: an autocatalytic optimizing process. Italy: Dipartimento di Elettronica, Politecnico di Milano; 1991. 91-016 Revised.
 108. Bauer A, Bullnheimer B, Hartl RF, *et al.* An ant colony optimization approach for the single machine total tardiness problem. Proceedings of the 1999 Congress on Evolutionary Computation (CEC'99). Piscataway (NJ): IEEE Press; 1999. pp. 1445–1450.
 109. Merkle D, Middendorf M, Schneck H. Ant colony optimization for resource-constrained project scheduling. In: Whitley D, *et al.*, editors. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000). San Francisco (CA): Morgan Kaufmann Publishers; 2000. pp 893–900.
 110. Yagiura M, Kishida M, Ibaraki T. A 3-flip neighborhood local search for the set covering problem. *Eur J Oper Res* 2006;172:472–499.
 111. Falkenauer E. A hybrid grouping genetic algorithm for bin packing. *J Heuristics* 1996;2:5–30.
 112. Martello S, Toth P. Knapsack problems, algorithms and computer implementations. Chichester: John Wiley & Sons, Ltd.; 1990.
 113. Maniezzo V. Exact and approximate non-deterministic tree-search procedures for the quadratic assignment problem. *INFORMS J Comput* 1999;11(4):358–369.

114. Lau KF, Dill KA. A lattice statistical mechanics model of the conformation and sequence space of proteins. *Macromolecules* 1989;22:3986–3997.
115. Korošec P, Šilc J, Robič B. Solving the mesh-partitioning problem with an ant-colony algorithm. *Parallel Comput* 2004;30:785–801.
116. Brandt A. Multilevel computations: review and recent developments. In: McCormick SF, editor. Volume 110, Multigrid Methods: Theory, Applications, and Supercomputing, Proceedings of the 3rd Copper Mountain Conference on Multigrid Methods, Lecture Notes in Pure and Applied Mathematics. New York: Marcel Dekker; 1988. pp. 35–62.
117. Walshaw C, Cross M. Mesh partitioning: a multilevel balancing and refinement algorithm. *SIAM J Sci Comput* 2000;22(1): 63–80.
118. Doerner KF, Hartl RF, Reimann M. Are CompetAnts more competent for problem solving? The case of a multiple objective transportation problem. *Cent Eur J Oper Res Econ* 2003;11(2):115–141.
119. Angus D, Woodward C. Multiple objective ant colony optimization. *Swarm Intell* 2009; 3(1):69–85.
120. Gutjahr WJ. S-ACO: an ant-based approach to combinatorial optimization under uncertainty. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 238–249.
121. Schoonderwoerd R, Holland O, Bruten J, *et al.* Ant-based load balancing in telecommunications networks. *Adapt Behav* 1996; 5(2):169–207.
122. Di Caro G, Dorigo M. Mobile agents for adaptive routing. In: El-Rewini H, editor. Proceedings of the 31st International Conference on System Sciences (HICSS-31). Los Alamitos: IEEE Computer Society Press; 1998. pp. 74–83.
123. Di Caro G. Ant Colony Optimization and its application to adaptive routing in telecommunication networks [PhD Thesis]. Brussels, Belgium: IRIDIA, Université Libre de Bruxelles; 2004.
124. Sim KM, Sun WH. Ant colony optimization for routing and load-balancing: survey and new directions. *IEEE Trans Syst Man Cybern Part A: Syst Hum* 2003;33(5):560–572.
125. Zhang Y, Kuhn LD, Fromherz MPJ. Improvements on ant routing for sensor networks. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 154–165.
126. Ducatelle F, Di Caro G, Gambardella LM. Using ant agents to combine reactive and proactive strategies for routing in mobile ad hoc networks. *Int J Comput Intell Appl* 2005;5(2):169–184.
127. Farooq M, Di Caro G. Routing protocols for next-generation intelligent networks inspired by collective behaviors of insect societies. In: Blum C, Merkle D, editors. *Swarm intelligence: introduction and applications*, Natural Computing Series. Berlin: Springer; 2008. pp. 101–160.
128. Ducatelle F, Di Caro G, Gambardella LM. Principles and applications of swarm intelligence for adaptive routing in telecommunications networks. *Swarm Intell* 2010. In press.
129. Socha K. ACO for continuous and mixed-variable optimization. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, Lecture Notes in Computer Science. Heidelberg: Springer; 2004. pp. 25–36.
130. Tsutsui S. Ant colony optimisation for continuous domains with aggregation pheromones metaphor. In: Proceedings of the The 5th International Conference on Recent Advances in Soft Computing (RASC-04). Nottingham; 2004. pp. 207–212.
131. Tsutsui S. An enhanced aggregation pheromone system for real-parameter optimization in the ACO metaphor. In: Dorigo M, *et al.*, editors. Volume 4150, Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Lecture Notes in Computer Science. Heidelberg: Springer; 2006. pp. 60–71.
132. Socha K, Dorigo M. Ant colony optimization for mixed-variable optimization problems. Belgium: IRIDIA, Université Libre de Bruxelles; 2007. TR/IRIDIA/2007-019.
133. Maier HR, Simpson AR, Zecchin AC, *et al.* Ant colony optimization for design of water distribution systems. *J Water Resour Plann Manag ASCE* 2003;129(3):200–209.
134. López-Ibáñez M, Prasad TD, Paechter B. Ant colony optimisation for the optimal

- control of pumps in water distribution networks. *J Water Resour Plann Manag* ASCE 2008;134(4):337–346.
135. Bautista J, Pereira J. Ant algorithms for assembly line balancing. In: Dorigo M, Di Caro G, Sampels M, editors. Volume 2463, *Ant Algorithms: 3rd International Workshop, ANTS 2002, Lecture Notes in Computer Science*. Heidelberg: Springer; 2002. pp. 65–75.
136. Blum C, Bautista J, Pereira J. Beam-ACO applied to assembly line balancing. In: Dorigo M, *et al.*, editors. Volume 4150, *Ant colony optimization and swarm intelligence, Lecture Notes in Computer Science*. Heidelberg: Springer; 2006. pp. 96–107.

A REVIEW OF EQUITY IN NONPROFIT AND PUBLIC SECTOR: A VEHICLE ROUTING PERSPECTIVE

BURCU BALCIK

SEYED M. R. IRAVANI

KAREN SMILOWITZ

Department of Industrial Engineering
and Management Sciences,
Northwestern University,
Evanston, Illinois

Public, nonprofit, and commercial sector organizations differ from each other in many aspects including goals, activities, and stakeholders. The differences among these sectors have been the subject of research in various disciplines. In operations research (OR), public sector is mainly characterized by nonmonetary performance requirements, multiple constituencies, and public scrutiny on decisions [1].

In this article, *public service* refers to services provided to the society by public and/or nonprofit sector organizations. Although some public services such as library, emergency, and postal services are primarily provided by government agencies, there are numerous public services, including material and housing assistance, disaster relief, various health and social services that are also provided by nonprofit sector organizations, such as food banks, housing associations, and nursing homes. Indeed, nonprofit organizations play a large and increasing role in delivering services traditionally provided by governments. In some cases, nonprofit and government agencies work in partnership complementing each other in providing public services. While nonprofit organizations provide greater service flexibility and access to underserved regions and populations, governments can ensure larger coverage with more resources. In this article, we focus on a common concern of both public and nonprofit organizations: achieving high performance in delivering public service.

The fundamental differences in goals between public/nonprofit and private sector organizations lead to unique performance metrics for each sector. The most common types of metrics in OR are *effectiveness* and *efficiency*. Effectiveness assesses the extent to which organizations attain their goals, whereas efficiency measures the amount of resources used to achieve specified goals [2,3]. For instance, for a private sector organization, effectiveness can be measured in terms of increased sales and improved customer service quality, and efficiency can be assessed by measured inputs (costs, resources) used to achieve these goals. For public and nonprofit organizations, providing equal access is an important component of their strategic goals and missions. Indeed, *equity* is a distinguishing aspect of decision making in the nonprofit and public sectors in addition to traditional effectiveness and efficiency objectives.

Concepts such as equity, fairness, and justice are subject to broad interpretation and endless debate, as discussed in Ref. 4. These concepts have been interpreted and practiced in many context-dependent ways. As such, equity and fairness may have distinct meanings in different OR applications in the literature; for instance, airspace allocation in airline traffic management; bandwidth allocation in telecommunications; cost/benefit allocation in collaborative logistics; or organ, blood, and drug allocation in health care. While there are various OR applications for which equity is pertinent, this article focuses on equity issues in public and nonprofit sector applications. Since public and nonprofit organizations typically operate with limited resources, defining equity is closely related to determining equity principles for allocating those resources. The complexities associated with allocating resources in public and nonprofit sectors can be related to value judgments made in defining equity in a particular context; that is, *fairness based on what, who should benefit, and how is fairness measured?*

Equity-related issues in the public/non-profit sector have been widely studied using OR models primarily in few areas, such as *facility location*, while some other areas such as *vehicle routing* have received relatively less attention. There is a large body of literature in facility location both in terms of public sector applications (such as locating undesirable/controversial public facilities, emergency service facilities, and public housing) and the types of equity metrics used (see Refs 4, 5 for reviews). In locating such public service facilities, it is generally assumed that either users travel to the facility, or in some cases, such as in ambulance location, the resources positioned in the facilities travel to users. Therefore, providing equal access is usually defined based on the distance between facilities and the users. However, in some services provided by public and nonprofit organizations, demands for services and goods occur at spatially distributed locations and must be satisfied by visiting each user sequentially using the available logistical resources. Therefore, *vehicle routing* is another important problem faced by public and nonprofit decision makers. Although routing-related decisions such as sequence of visits and delivery amounts may affect equitable service provision, we observe that most of the vehicle routing literature focuses on the traditional efficiency (cost-based) objectives and equity is not well studied. In this article, we focus on equity issues within the context of vehicle routing and provide examples from public and nonprofit sector applications for which equity is a key performance indicator. Although we focus only on equity in the context of vehicle routing, this article also sheds light on how OR can be used to model equity in other contexts.

The remainder of this article is organized as follows. In section titled “Equity Overview,” we provide a broad overview of equity characterization and measurement. In section titled “Equity in Routing,” we focus on equity in vehicle routing and review the characteristics of various OR applications that consider equity. Section titled “Conclusion and Future Research” summarizes and identifies future research areas.

EQUITY OVERVIEW

There is no universal definition for equity. In general, equity is conceptually related to fairness, justice, and impartiality and considered in association with allocation of resources, benefits, or disbenefits. Many disciplines including economics, philosophy, political science, and mathematics have studied equity and equitable resource allocation. In this section, we provide a broad overview of how equity is approached in OR.

Equity is not new to OR; indeed, there is a diverse body of literature that addresses equity-related issues, from studies that describe the aspects of equity measurement through a discussion of different equity principles and metrics to papers that analyze the mathematical properties of existing or newly proposed equity objectives/metrics. Relatively few studies study the implications of using different equity metrics and the trade-offs between equity and other relevant objectives such as efficiency when incorporating equity in models.

Studying equity involves three interrelated steps: (i) definition of equity elements, (ii) policy design and implementation, and (iii) measurement of the outcomes of the policies, shown in Fig. 1.

Definition

Defining equity may not be straightforward since it often requires judgments regarding how individuals are affected by critical decisions; for example, allocating emergency relief resources, routing of hazardous waste. In general, a decision is deemed equitable when its effects are approximately equal across the affected parties. Therefore, defining equity requires characterizing three elements: *equity determinant*, *affected parties*, and their relevant *attributes*.

To characterize the effects, one must determine a basis for equity comparison (i.e., fairness based on what?), which Marsh and Schilling define as the *equity determinant*. Then, *the affected parties* addressed in the comparison (i.e., fairness among whom?) and their relevant *attributes* (i.e., population, demand, social needs) must be identified. As described in Ref. 4, an affected party can refer

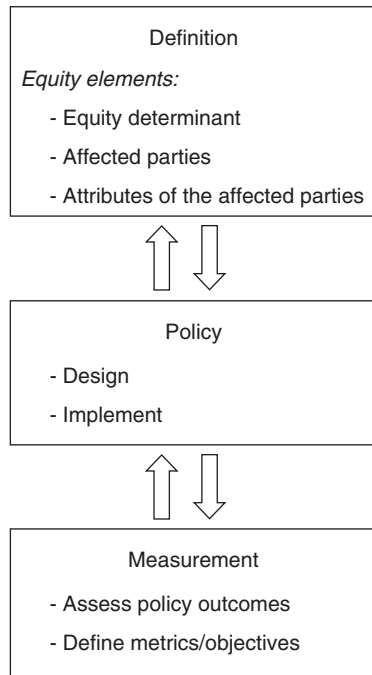


Figure 1. Equity characterization and measurement.

to a single individual or groups of individuals defined based on geographical, demographic, physical, and temporal aspects.

Equity determinants and affected parties capture the characteristics of the applications. For instance, in queueing systems, a typical equity determinant is service waiting time and can be compared among individuals. In resource allocation, an equity determinant can be the amount of commodities allocated among spatially distributed groups of individuals. In a workload allocation problem, an equity determinant can be the time required to complete tasks, measured across individual employees. In facility location, equity determinants are typically defined based on distance or time. For instance, in locating a desirable public facility such as a fire station, affected individuals can be grouped geographically and the effects can be characterized based on the distance or response time between the facility and the groups of beneficiaries.

Policy

In the next interrelated step, policies are designed and implemented with the above equity elements in mind; that is, equity determinants, affected parties, and critical attributes of the affected parties. Therefore, how one defines equity and fairness directly affects the types of policies to be designed and implemented. For instance, in queueing systems such as banks, policies based on first-come, first-served or first-in, first-out are traditionally considered the most fair. However, depending on how equity is characterized, policies that prioritize based on service times or urgency can also be considered to be fair. In public service provision, alternative policies may exist to allocate public resources equitably. In Ref. 3, the author discusses examples of an allocation policy based on equal inputs that may result in imbalanced outputs; for example, assigning an equal number of ambulances to each town may lead to higher casualties in large towns. Therefore, analyzing different equity policies is important in incorporating equity in decisions. As noted in Ref. 3, the implications of the options must be explored fully in the process.

Measurement

As discussed in Ref. 4, equity measurement involves comparing the effects of the decisions on the affected parties. Equity measurement is critical as it allows one to monitor and evaluate the degree of fairness achieved by the implemented policies and update policies if necessary.

An equity metric or a set of metrics can be used to evaluate the equity-based performance of a system. Each metric is formulated based on the specified equity elements: equity determinants, affected parties, and their attributes. Several equity metrics and objectives have been adopted in various problems and applications in the OR literature. For example, minmax type objectives that aim to improve the condition of the least advantaged have been widely used within the context of facility location and resource allocation problems. Twenty equity metrics relevant to facility location are reviewed in

Ref. 4. Other commonly used equity metrics include variance, range, mean deviation, and Gini coefficient. Given the number of alternatives in measuring equity, exploring the similarities and differences in solutions obtained from different metrics are important. Although there are several studies that compare different equity metrics, it is difficult to reach general conclusions regarding the use of a specific metric for different problems and settings.

Analyses that explore the trade-offs between equity and other objectives and evaluate the effects of various system aspects (e.g., regulations, budget) on equity can be insightful in understanding the use of different equity policies and metrics. Facility location and resource allocation literatures are relatively rich in this respect. For instance, in Ref. 6, the authors analyze the trade-offs between efficiency in terms of cost, and equity in terms of ambulance response time in providing equitable emergency medical service to people considering the inter- and intra-regional policies in Germany. Given the budget constraints in emergency care provision and the cost differences across regions, the policy makers face an important question: *whether to increase the total number of lives saved or provide equal access across regions*. Answering this question is difficult and open to debate. In Germany, policies for providing emergency medical services that emphasize efficiency over equity are followed across federal states; however, within each state equal access policies are mandated by law. Investigating the effects of various factors such as size of populations and regions, and regional income on the efficiency- and equity-based policies, the authors show that the actual implementation of the current equal access policy mandated by law values the lives saved in a rural area more than those in urban areas.

To summarize, there are many issues to be considered while incorporating equity in decisions. The literature indicates the importance of evaluating alternatives and analyzing implications while characterizing equity and selecting equity metrics. In the next section, we focus on equity as it pertains

to vehicle routing decisions in public service provision.

EQUITY IN ROUTING

In this section, we focus on equity issues within the context of vehicle routing. We review applications in which routing decisions affect equitable and fair allocation of resources, benefits, costs, or risks to the public. Table 1 summarizes the equity-related characteristics of the applications that integrate equity in routing decisions in the public/nonprofit sector. The table specifies the types of decisions addressed in different studies and provides examples of equity determinants and equity metrics/objectives for each application. In this section, we discuss several papers cited in Table 1 that highlight how equity is addressed in routing decisions.

The classical vehicle routing problem (VRP) determines a set of delivery routes for a fleet of vehicles dispatched from a depot to fulfill the demands of a set of customers, while minimizing total transportation costs. Since the introduction of the basic VRP over 50 years ago, the VRP and its extensions with different objectives and operational constraints have been widely studied. Extensions of the basic VRP include problems with multiple periods, customer and vehicle time windows, precedence constraints, heterogeneous vehicle fleets, and route length restrictions. Objectives such as minimizing the number of vehicles and routes or the length of the longest route have been used in various VRPs.

Many studies address vehicle routing for various commercial and public/nonprofit sector applications. Studies in commercial sector routing address delivery of a variety of goods and services. Several studies address problems of local city governments in providing public services such as street cleaning, electric meter reading, winter road maintenance, and waste collection. Other routing applications focusing on public services include public libraries, postal services, various health-care services such as blood collection, ambulance routing, home health care, disaster relief, material assistance, and mobility services.

Table 1. Routing Applications with Equity Considerations

Application (Example studies)	Decisions	Equity Determinants	Metrics/Objectives
Disaster relief distribution [11,12]	• Routing • Supply allocation	• Arrival time • Supplies	Equity metrics/objectives • [Minmax/Minsum] arrival time • Mean absolute upper semideviation of arrival times • [Minmax] unsatisfied demand percentage Other metrics/objectives • [Min] transportation cost (or time and distance)
Mobility services [10], [13–17]	• Routing • Location	• Travel time • Distance	Equity metrics/objectives • [Min] variation in route lengths • [Minmax/Limit] user travel time • [Limit] user walking time (or distance) to stops • [Limit] deviation from desired pickup/delivery time • Total absolute deviation from the mean travel time Other metrics/objectives • [Min] number of routes (or vehicles) • [Min] transportation cost • [Min] total route length (or time and distance) • [Min] total deviation from desired pickup/delivery times • Balance loads (number of users per route)
Food distribution [18–20]	• Routing • Supply allocation • Location	• Supplies • Distance	Equity metrics/objectives • [Max] expected minimum fill rate • [Min] total distance between unmet demands and facilities Other metrics/objectives • [Min] expected waste • [Min] transportation costs • [Min] facility costs
Hazardous material transportation [21–23]	• Routing • Location	• Risk • Distance	Equity metrics/objectives • [Minsum/Minmax/Limit] risk difference • [Min] average risk Other metrics/objectives • [Min] total risk • [Min] total cost (or time) • [Min] expected number of accidents

We observe that most of the routing studies in the literature address effectiveness and efficiency objectives; attention on equity-related issues has been recent and limited to few applications, mostly addressing public/nonprofit sector problems. In most commercial sector routing problems, equity is often not considered to be a factor affecting customer service. Equity in a commercial distribution system is studied in Ref. 7, which ensures equitable delivery times to customers in overnight delivery services by allowing a different set of routes on each day and limiting deviations from average delivery time through a set of constraints. Studies that explicitly consider equity as a major aspect in the public/nonprofit sector mainly address applications in the following areas: *disaster relief*, *food distribution*, *mobility services*, and *hazardous material transportation*, which are discussed in detail in the following sections.

We note that some studies that address routing decisions in commercial and public/nonprofit applications consider *route* and *load balancing* to ensure fairness for those who provide the service. More specifically, the problem seeks routes that yield balanced workloads and route distances/times to meet various regulations or for efficiency purposes [8–10]. Although balanced workloads for employees might have implications on the quality of service and fairness toward users, our focus is on applications that explicitly address equity and fairness for those served and/or affected by the operations, not the service providers.

Table 1 summarizes the equity-related characteristics of the applications that integrate equity in routing decisions in the public/nonprofit sector. The table specifies the types of decisions addressed in different studies and provides examples of equity determinants and equity metrics/objectives for each application. The following sections focus on each application area in detail. Specifically, section titled “Disaster Relief Distribution” discusses vehicle routing applications with equity considerations in disaster relief. The section titled “Mobility Services” presents several papers that consider equity in routing for mobility services such as school buses and

transportation services for the disabled and elderly. The section titled “Food Distribution” provides examples that consider equity in food distribution services. Finally, the section titled “Hazardous Material Transportation” discusses the equity issues in routing of hazardous materials.

Disaster Relief Distribution

Disaster relief services aim to save lives and reduce suffering after the onset of a disaster. The OR literature has widely addressed problems related to allocation and deployment of medical, police, and fire resources, in which equitable access is a major consideration. Most of the studies addressing these small-scale emergencies consider equity in terms of response time and coverage levels, and aim to position emergency resources to reduce the disparities in access to different regions. There has been an increasing attention in the literature for distribution problems in responding to emergencies caused by large-scale, high-impact disasters such as earthquakes, hurricanes, and floods. Large-scale disasters can be differentiated from smaller-scale emergencies by a number of aspects including their low frequencies and high-impacts. Once disasters occur, they are likely to create large demands for a variety of supplies at locations that may spread over a large area. The extent of demand created for relief supplies, staff, and services, combined with infrastructure difficulties, challenge the resource capacities of relief organizations in delivering aid. Owing to scarce logistic resources, it is typically very difficult, if not impossible, to immediately satisfy the entire relief demand. Therefore, allocating the available supplies fairly and equitably is a major concern in disaster relief. Moreover, response time equity is still critical in responding to large-scale disasters, especially during the initial days of response when search and rescue efforts continue, medical attention is vital, and time is the most valuable resource.

The majority of studies in the literature that address relief distribution problems use network flow type formulations [24–26]. Notably, in Ref. 26, the authors incorporate equity in their relief distribution model; a fill

rate-based equity objective that maximizes the minimum demand fraction is used in addition to efficiency objectives such as setup, operational, and transportation costs, and effectiveness objectives such as total travel time.

There are few studies that consider equity in classical vehicle routing type formulations for relief supply distribution. Equitable decisions are considered in local distribution of relief aid in Refs 11, 12. In both studies, the affected parties are defined as spatially distributed populations. Also, both papers incorporate equity in their models through defining equitable objectives. Specifically, in Ref. 12, the authors focus on evaluating two objective functions to ensure equity in routing decisions: minimizing the maximum arrival time of supplies and minimizing the average arrival time of supplies, which are analogous to minimizing makespan and sum of completion times in scheduling problems, respectively. They explore the potential impact of using these objective functions and compare solutions with those obtained with the traveling salesman problem (TSP) and VRP with the traditional objective of minimizing total travel time. Specifically, they analyze the relationship of the minmax and minsum objectives with the traditional minimizing total travel time objective and develop bounds on the performance of the new objectives for both TSP and VRP. It is shown that the new objective functions ensure better response times for the demand locations that are served later in a route. Also, average deviation of arrival times from the mean decreases, at the expense of reduced efficiency; that is, increased total route length.

In Ref. 11, routing and supply allocation decisions are considered jointly. The major problem is fair distribution of multiple types of supplies in a relief network. This study does not consider response time equity in terms of arrival times of supplies to the demand locations on each route; rather, delivery scheduling decisions are driven by supply allocation decisions and total transportation costs over a planning horizon. Given the supply, vehicle capacity, and delivery time restrictions, the authors address the problem of determining delivery

schedules, vehicle routes, and amounts of supplies to be delivered from a depot to demand locations during a relief planning horizon, whose length is unknown *a priori*. In characterizing equity in supply allocation, the authors capture criticality of supplies for different population groups by assigning penalty weights for unsatisfied demand at each location. Equitable allocation of relief supplies among demand locations is ensured by minimizing the maximum (weighted) unsatisfied demand percentage over all demand locations for each period. An efficiency objective based on total routing costs is also considered; however, decisions become mostly driven by supply allocation decisions as the importance of transportation cost is decreased through assigned weights. In their analysis, Balcik *et al.* [11] observe the effects of demand amounts, penalty weights, and remoteness of demand locations to the depot on supply allocation decisions.

Mobility Services

In this section, we focus on equity considerations in routing for mobility services. Examples of mobility services include school buses and transportation services for the disabled and elderly.

The school bus routing problem determines routes for public school buses to transport spatially distributed students from their residences to and from schools. The major considerations are efficiency and equity. Efficiency can be characterized by minimizing capital and transportation costs associated with operating buses, which alone might yield an inexpensive yet inequitable solution. Balanced bus route lengths and loads are the most frequently used equity indicators in school bus routing literature, which are incorporated in the models either in the objective function or the constraints; see Ref. 15 for a review of objectives and constraints used in school bus routing.

An urban school bus routing problem is considered in Ref. 13. The authors developed a multiobjective mathematical model that locates bus stops, assigns students to bus stops, and determines the vehicle routes

jointly. Equity is ensured via a set of objectives and constraints. In addition to balancing students travel times on buses, the walking distances of students to and from bus stops is also considered as an equity determinant. Each student is assigned to a bus stop such that each student walks less than the maximum walking distance, a constraint required by the school board. Variation in the route lengths (in terms of number of stops) is minimized. Finally, each bus route transports approximately equal number of students (i.e., balanced loads) to reduce the likelihood that some routes exceed capacity in the future. As such, load balancing is also an efficiency objective since it includes costs of reoptimization if new students are added to the system.

City characteristics can add further aspects to consider in modeling school bus routing problems. For example, a school bus routing problem in New York City is modeled in Ref. 10. The authors introduce mixed loads and multiple drop-off points in the model such that students from different schools may be allocated to the same bus. This flexibility can increase efficiency. The authors consider the routing problem in the mornings, which requires a higher number of buses due to more constraining time windows and traffic congestion. They note that a maximum distance constraint might be perceived as inequitable for students and parents if applied uniformly, since students whose bus stops very close to school might travel longer times than desired. Therefore, student-specific travel time constraints are used to ensure equity.

In Ref. 14, school bus routing in a sparse rural area is considered. Here students may experience longer travel times and greater variation in travel times, as compared to students in urban settings. Therefore, the maximum travel time is the major concern in rural school bus routing. As a result, in rural areas, buses reach an allowed route length in terms of time before they reach their maximal physical capacity. The model in Ref. 14 contains two objectives that are somewhat conflicting: (i) an equity objective of minimizing the maximum route length, and hence maximum time in the bus, and

(ii) an efficiency objective that minimizes the number of buses.

Similar equity objectives and constraints are relevant to other mobility services such as transportation for the disabled. These problems can be formulated as a dial-a-ride problem (DARP). In DARP, an origin and a destination is associated with each transportation request and each route starts and ends at a depot. DARP is a variation of pickup and delivery problem, in which people are transported door-to-door in groups or individually; see Ref. 27 for a review.

As discussed in Ref. 27, the efficiency objective of minimizing costs must be balanced against an objective of reducing user inconvenience in terms of long ride times and large deviations from desired departure and arrival times. Note that maximizing total effectiveness and quality of service does not necessarily lead to an equitable solution; either the worst-case service level over all users or service level differences among users or user groups must be explicitly considered for fairness. An example study that considers equity is Ref. 17, which considers the problem of meeting transportation requests of disabled people in an urban area. Transportation requests are known in advance and special types of vehicles may be needed to transport each user. The authors impose restrictions on maximum ride time for each user proportional to the direct travel time between the user's origin and destination.

Another study that focuses on free transportation services provided to disabled people by a nonprofit organization is Ref. 16. Vehicles transport people from their homes in the mornings to a training center, and return them home in the evenings. Working with the organization, Ref. 16 chooses the objective of minimizing total trip times by all vehicles as a proxy for cost subject to constraints on the minimum and maximum trip time for each vehicle and maximum and minimum number of seats occupied in each vehicle. An objective that aims to equalize trip times for each vehicle to prevent some trainees from excessive travel times is not found as a critical objective for the organization. As a result, while improvements are obtained in terms of travel times, distances and balanced loads, a

less equitable solution is obtained in terms of vehicle travel times.

Food Distribution

Similar to the studies discussed above in mobility services, OR methods can be used to improve food distribution services in communities. In Ref. 19, the authors consider collection and distribution of perishable food for a regional food bank. Each day, vehicles collect food from donor sites and deliver the food to recipient agencies on the scheduled routes. Donations and agency demands are observed upon the visit of the vehicles to the sites. This problem is formulated as a sequential resource allocation problem considering a single commodity. The authors develop allocation policies to provide equitable and sustainable service to agencies. Both throughput- and equity-based objectives are considered. It is shown that an objective that minimizes the expected waste and maximizes the total distribution results in inequitable solutions in terms of large discrepancies in fill rates for the agencies, whereas minimizing maximum expected fill rate tends to find an equitable solution with near-minimum waste. However, due to the mathematical structure of this equity-based objective, the existing methods from the commercial sequential allocation research are not directly applicable. Therefore, new solution methods are developed to solve the food distribution problem.

Another study that focuses on food assistance is Ref. 18, which considers distribution of hot meals to spatially dispersed homebound elderly people. This service is commonly known as “meals-on-wheels,” has also been studied in Ref. 20. In Ref. 18, the problem is formulated as an integrated location-routing problem, which involves determining kitchen locations and vehicle routes. Since the entire demand may not be satisfied using available kitchen capacity, an objective is to locate kitchens while unmet demand is fairly distributed over the region. This equity objective is modeled by minimizing total distance between unmet demand points and their closest kitchens. Other objectives are efficiency (minimization of fixed and variable facility

costs and transportation costs) and effectiveness (maximization of total throughput).

Hazardous Material Transportation

Hazardous materials such as radioactive materials, explosives, infectious substances pose a threat to the health or safety of people; therefore, minimizing population exposure when storing and transporting these materials is important. In hazardous material management, it is critical to obtain politically acceptable solutions that reduce safety risks and perceptions of injustice among local communities.

The problems related to locating sites for storing and disposing these materials and transporting the material from collection points to disposal sites have been addressed by a large body of literature; see review in Ref. 28. When optimizing the location of disposal sites, the general approach is to seek sites that are far from population centers to minimize risk. However, if facilities are located too far from urban areas, transportation costs tend to increase. Facility location decisions must incorporate these trade-offs, as well as fairness to all populations, recognizing that only equitable solutions would be accepted by communities.

In routing vehicles carrying hazardous materials, the critical issue is to limit and distribute equitably the risk over the geographical crossed regions [21]. While carriers focus on transportation costs, regulatory government agencies need to consider spatial risk equity to prevent perceptions of injustice that may result in public opposition to the use of nearby passageways and also excessive usage of some road segments, which may increase the chance of accidents [28]. Cost minimization approaches tend to ship large quantities over the inexpensive routes, whereas an equity objective may lead to transportation of smaller quantities over a large number of routes hence reducing the maximum exposure faced by each individual, which would increase transportation costs [22]. Although distributing risk over many routes reduces the risk to any one individual, the number of people exposed to risk may increase. Various multiobjective approaches are used in the literature to capture the trade-offs between

these objectives. Risk equity is modeled and assessed in a number of ways; for example, equity is ensured along the arcs, zones, or paths in a network. We provide a sample of equity objectives in Table 1.

Discussion

These applications summarize the main equity-related issues addressed in vehicle routing literature. Equity is defined, incorporated, and measured in models in different ways. As observed in Table 1, minmax and deviation type metrics and objectives are most frequently used. However, in general, types of equity metrics used in routing applications are relatively limited, for instance, compared to facility location applications.

We observe that equity is more frequently addressed in routing problems related to hazardous material transportation and various mobility services compared to other public services. Indeed, there are few examples that discuss equity in the distribution of supplies for disaster relief or other nonprofit operations, although we have seen a recent increase in these areas. Moreover, few studies discuss and compare the characteristics and implications of different equity policies and metrics or analyze the trade-offs between equity and other objectives such as effectiveness, efficiency, and flexibility.

Finally, most of the applications discussed in this article use mathematical modeling and heuristics. Since the VRP is NP-hard, incorporation of additional equity-related aspects usually brings additional challenges in solving problems. Moreover, resource allocation and facility location decisions are considered jointly with routing decisions in many applications further increasing the problem complexity.

CONCLUSION AND FUTURE RESEARCH

This article explores equity-related issues in the public and nonprofit sectors, focusing on vehicle routing applications. We discuss how equity is approached in OR in general and review studies from the literature that incorporate equity in vehicle routing applications.

Discussions and analyses regarding equity characterization and measurement are limited to few examples in vehicle routing applications. Therefore, future research can address various issues in studying equity in vehicle routing. For instance, implications of using different equity metrics can be explored for different problem settings in various application areas. Campbell *et al.* [12] show that modeling the delivery of relief supplies as a traditional TSP can double the latest arrival time to an aid recipient. Future work should continue to evaluate the trade-offs between efficiency and equity.

Additionally, methodological and practical implications of using different equity principles and policies can be explored. Solution approaches for nonprofit agencies often must focus on simplicity rather than on exploitation of technology. For example, the authors in Ref. 20 worked with meals-on-wheels to improve the distribution of lunches to the elderly. Meals-on-wheels needed a solution approach that could adapt to their changing client base without the use of computers. The project led to the development of a solution approach, based on the concept of space-filling curves, which could dynamically change routes, using only a map of Atlanta and two Rolodex files. The travel times of routes obtained with this method were generally within 25% of the shortest possible routes.

Studies addressing various facility location and resource allocation problems discuss the characteristics that are desirable when selecting equity metrics. For instance, criteria such as Pareto efficiency and principle of transfers are considered as minimum requirements for fair resource allocation. Future research may explore such criteria within the context of vehicle routing.

In providing public services, routing and resource allocation problems are often considered together. Future research can address integration of available resource allocation policies in the literature with routing decisions. For instance, in Ref. 29, the authors study a food allocation problem in a disaster relief setting and propose an equitable policy that considers two levels of food supply. Specifically, the proposed policy evaluates

the starvation level of each demand region before proportionally increasing the allocation amounts to reach a healthy existence level. The results and algorithms in such resource allocation studies can be used in developing solution methods for integrated routing and resource allocation problems.

In most routing applications, populations affected by decisions are assumed to be uniform while characterizing equity. This is often due to difficulties in capturing and quantifying differences among populations. For instance, identifying vulnerable groups and quantifying the criticality of emergency supplies at different locations are challenging in a disaster relief environment. Further research might explore how to incorporate the attributes of the affected parties in decisions.

Finally, future work may consider fairness in a multiple agency setting. Currently, most studies consider a single agency following a single equity principle. However, recent disasters such as Hurricane Katrina and the Asian tsunami highlighted the need for inter-agency coordination. It would be valuable to study the effects of multiple agencies implementing various policies in providing public service. Investigating the effects of coordination on equitable service provision might also have important implications in terms of efficiency and accountability of operations, particularly if agencies define their equity elements (equity determinants, affected parties, and relevant attributes) differently.

REFERENCES

1. Pollock S, Rothkopf M, Barnett A, editors. Operations research in the public sector. Volume 6, Handbooks in operations research and management science. Amsterdam: Elsevier Science, North-Holland; 1994.
2. Gass SI. Public sector analysis and operations research/management science. In: Pollock S, Rothkopf M, Barnett A, editors. Volume 6, Operations research in the public sector. Handbooks in operations research and management science. Amsterdam: North-Holland; 1994. pp. 23–46.
3. Savas ES. On inequality in providing public services. *Manage Sci* 1978;24(8):800–808.
4. Marsh MT, Schilling DA. Equity measurement in facility location analysis: a review and framework. *Euro J Oper Res* 1994;74(1):1–17.
5. Erkut E. Inequality measures for location problems. *Location Sci* 1993;1(3):199–217.
6. Felder S, Brinkmann H. Spatial allocation of emergency medical services: minimizing the death rate or providing equal access? *Reg Sci Urban Econ* 2002;32:27–45.
7. Dell RF, Batta R, Karwan MH. The multiple vehicle TSP with time windows and equity constraints over a multiple day horizon. *Transplant Sci* 1996;30(2):120–133.
8. Bertels S, Fahle T. A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. *Comput Oper Res* 2006;33(10):2866–2890.
9. Perrier N, Langevin A, Amaya CA. Vehicle routing for urban snow plowing operations. *Transplant Sci* 2008;42(1):44–56.
10. Simchi-Levi D, Chen X, Bramel J. A case study: School bus routing. In: *The logic of logistics: theory, algorithms, and applications for logistics and supply chain management*. New York: Springer; 2004. pp. 319–335.
11. Balcik B, Beamon BM, Smilowitz KR. Last mile distribution in humanitarian relief. *J Intell Transport Syst* 2008;12(2):51–63.
12. Campbell AM, Vandenbussche D, Hermann W. Routing for relief efforts. *Transplant Sci* 2008;42(2):127–145.
13. Bowerman R, Hall B, Calamai P. A multi-objective optimization approach to urban school bus routing: formulation and solution method. *Transport Res Part A* 1995;29(2):107–123.
14. Corberan A, Fernandez E, Laguna M, et al. Heuristic solutions to the problem of routing school buses with multiple objectives. *J Oper Res Soc* 2002;53(4):427–435.
15. Li L, Fu Z. The school bus routing problem: a case study. *J Oper Res Soc* 2002;53(5):552–558.
16. Sutcliffe C, Board J. Optimal solution of a vehicle-routing problem: transporting mentally handicapped adults to an adult training centre. *J Oper Res Soc* 1990;41(1):61–67.
17. Toth P, Vigo D. Heuristic algorithms for the handicapped persons transportation problem. *Transplant Sci* 1997;31(1):60–71.
18. Johnson M, Gorr WL, Roehrig S. Location/allocation/routing for home-delivered meals provision: model and solution approaches. *Int J Indus Eng, Special Issue on Facility Layout and Location* 2002;9(1):45–56.

19. Lien RW, Iravani SMR, Smilowitz KR. Sequential resource allocation for nonprofit operations. Department of Industrial Engineering and Management Sciences, Northwestern University, 2008. Working paper.
20. Bartholdi JJ, Platzman LK, Collins RL, et al. A minimal technology routing system for meals on wheels. *Interfaces* 1983;13(3):1–8.
21. Carotenuto P, Giordani S, Ricciardelli S. Finding minimum and equitable risk routes for hazmat shipments. *Comput Oper Res* 2007;34(5):1304–1327.
22. Current J, Ratick S. A model to assess risk, equity and efficiency in facility location and transportation of hazardous materials. *Location Sci* 1995;3(3):187–201.
23. Lindler-Dutton L, Batta R, Karwan MH. Equitable sequencing of a given set of hazardous material shipments. *Transplant Sci* 1990;25(2):124–137.
24. Barbarosoglu G, Arda Y. A two-stage stochastic programming framework for transportation planning in disaster response. *J Oper Res Soc* 2004;55(1):43–53.
25. Haghani A, Oh SC. Formulation and solution of a multi-commodity, multi-modal network flow model for disaster relief operations. *Transport Res Part A* 1996;30(3):231–250.
26. Tzeng GH, Cheng HJ, Huang TD. Multi-objective optimal planning for designing relief delivery systems. *Transport Res Part E* 2007;43(6):673–686.
27. Cordeau J-F, Laporte G. The dial-a-ride problem: models and algorithms. *Ann Oper Res* 2007;153(1):29–46.
28. Erkut E, Tjandra SA, Verter V. Hazardous materials transportation. In: Barnhart C, Laporte G, editors. Volume 14, *Handbooks in operations research and management science*. New York (NY): Elsevier; 2007. pp. 539–621.
29. Betts LM, Brown JR. Proportional equity flow problem for terminal arcs. *Oper Res* 1997;45(4):521–535.

A REVIEW OF TOOLS, PRACTICES, AND APPROACHES FOR SUSTAINABLE SUPPLY CHAIN MANAGEMENT

HENDRIK REEFKE

Cranfield University, Cranfield,
UK

JASON LO

University of Auckland Business
School, Auckland,
New Zealand

INTRODUCTION

From a business perspective, high economic growth cannot be achieved in the long run without protection of environmental and social needs but economic progress can also not be completely sacrificed for altruistic environmental or staff protection. The goal is rather to assure the long-term viability of a business model that can only be sustained if shareholders, suppliers, employees, and customers see a future in it [1]. Sustainability requirements can essentially be split into three interdependent dimensions encompassing environmental considerations, societal aspects, and economic development [2]. Sustainability as a concept provides a vision or roadmap for the future [3] but is difficult to grasp as related issues usually involve stakeholders with different interests and opinions regarding their responsibilities and sustainability requirements [4, 5]. Boundaries to an entity's impacts and associated responsibilities are difficult to assess and assign [6], often due to insufficient understanding of influences and interdependencies in a system environment [7, 8]. Many companies have started to implement sustainability into their internal operations, but it has frequently been emphasized that managers have to take their wider supply chains (SCs) into account as 50–70% of product value is actually derived through the SC [9, 10]. The

areas of sustainability and supply chain management (SCM) are characterized by inherent complexities, making it challenging to integrate sustainability considerations into SC decisions. However, examples have shown that making sustainability a strategic imperative helps overcome such challenges [11]. Many companies have recognized this pressing need and best-in-class companies take environmental initiatives and social responsibilities seriously [12].

Research Motivation and Objectives

Sustainable supply chain management (SSCM) has emerged as a relatively new field of research and practice in order to foster the integration of sustainability into SCs. While SSCM has become an enduring research area, there is a lack of conceptual theory development backed by rigorous research approaches [13–15]. SCs are complex structures that can span across multiple tiers of suppliers and customers [16, 17]. SC issues such as maintaining visibility, dependence on collaborative practices, or accurate performance measurement reach a new level of importance and difficulty when environmental and societal considerations are added to economic necessities. Theoretical understanding of sustainability requirements in SCs is limited and there is a lack of knowledge with regard to practices, methods, and prerequisites for SSCM. Accordingly, the currently available SC principles, frameworks, and models are not designed to meet these challenges and generally do not allow the transformation of existing SC processes toward a sustainable focus. Hence, theoretical and procedural support is required in the form of conceptual insights and practical approaches to address the challenges outlined.

On the basis of these motivational aspects, this article aims to identify practical tools and theoretical ideas that can help SC managers to implement sustainability into their operations and that furthermore hold the potential

to overcome the many systemic challenges. Furthermore, this article synthesizes these aspects within key elements of SCM. This not only provides context and tangibility to the review but also directly demonstrates the applicability of the aspects discussed.

DESCRIPTION OF THE REVIEW PROCESS

A literature review is “a systemic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners” [18]. Previous reviews have focussed on providing a descriptive analysis of related literature [13–15, 19–22] in order to provide structure to the field but did not necessarily concentrate on content. This article addresses the need to discover and summarize the conceptual content of aspects relevant to SSCM and thereby contributes to theory development and practical application.

Identification of Relevant Material

The review process relied on the use of multiple databases and appropriate keywords. The following databases that focus on the fields of business and economics were utilized:

- ABI/INFORM
- ACM Digital Library
- Business Source Premier

- Emerald
- SCOPUS
- Science Direct

The keywords shown in Table 1 were searched for within each of the databases. Firstly, the primary keywords were used in order to identify influential articles within the broader area of SSCM. The search then turned to more specific topics in SCM, that is, the secondary keywords. These were used in conjunction with the primary keywords in order to synthesize methods and tools specific to these SCM areas.

The primary keywords are essentially a deconstruction of the term *SSCM* and thus reflect how different publications may have referred to it. Using them in conjunction and separately allowed to identify a range of applicable publications. These articles provided the background knowledge for this review and allowed to delineate the field of study. The rationale for concentrating on certain SC issues, and hence the selection of secondary keywords, was motivated by the goal to deliver an extensive but nevertheless concise review. An overview of what SCM essentially entails can be derived from definitions. While some authors focus mainly on material flows and coordination efforts [23], others emphasize managerial and network perspectives [24]. A more holistic view of SCM is generally accepted now which includes the integration of business processes, management decisions, and activities

Table 1. Keywords and Synthesis

Primary Keywords	Secondary Keywords	Synthesis
• Supply chain • Sustainability • Sustainable • Sustainable supply chain management	• Contracts • Customer value • Distribution • Information technology (IT) • Inventory • Measurement • Network • Outsourcing • Partnership • Procurement • Product design • Strategy	• Approaches • Ideas • Methods • Practices • Results • Tools

across the entire SC [25]. On the basis of this realization, it was deemed best to focus on recognized key issues in SCM to guide the selection of secondary keywords. The chosen SC issues are derived from seminal sources in the field [26, 27] and range across strategic, tactical, and operational levels. They were thus considered to be a useful guide toward functions and influential factors that are likely to have a direct impact on the sustainability of an SC as well. They are reflected by the secondary keywords displayed in Table 1 and investigated through a sustainability lens in this review. In line with the aims of this article, relevant tools, methods, and ideas were synthesized. Owing to the choice of primary and secondary keywords, these are directly associated with sustainability considerations and the key SCM issues.

Content Analysis

A content analysis facilitated the systematic analysis of the identified material, akin to previous reviews in the field [13, 22, 28]. As outlined, the material was identified based on keywords. These superimposed categories were then further utilized to guide the analysis of the material and furthermore to structure this article. As indicated under 'synthesis' in Table 1, the identified material was closely examined in order to discover relevant applications and ideas for SSCM. The relevance of content was judged qualitatively, that is, by drawing a direct connection to sustainability in SCs, but more quantitative aspects were also considered including the currency and impact of reviewed publications. As a review, this article relies on well-recognized material for guidance and structure while the synthesis draws mainly on fairly current publications. This is also influenced by the development of SSCM as a field of research that has predominantly taken place from the mid-2000s onward [21, 22, 28].

Readers should be aware that there is some overlap between the secondary keywords and thus between the sections of this article. Categorized data can be interpreted in different ways [29], and hence, some of the reviewed material could have been included in multiple sections or additional sources

could have been considered based on the rationale employed by the researcher. Furthermore, additional keywords could have been considered. The use of superimposed keywords can be justified by the careful selection process based on acknowledged key issues in SCM and the aims of this article, that is, to provide a concise and timely review. While these potential issues need to be acknowledged, it can nevertheless be assumed that this article presents a wide-ranging and insightful overview of up-to-date methods and practices conducive to SSCM.

SUSTAINABLE SUPPLY CHAIN MANAGEMENT

Most research on corporate sustainability has focussed on manufacturing and management processes in single companies [30], neglecting SC issues as well as systemic linkages and relationships. It is the aim of this article to investigate strategies, methods, and tools that have the potential to support SSCM. As a starting point, it is therefore necessary to reach an understanding of what SSCM entails and identify influential factors that determine the sustainability of an SC.

SC activities add an increasing proportion of value to products and services [31], can provide competitive advantages [19, 22, 32], and may also account for adverse side effects [13, 33]. Dealing with sustainability-related risks and creating market opportunities [34] requires that sustainability considerations are embedded in SC operations and decisions [19, 22, 32]. A three-dimensional focus on economic, environmental, and social considerations is evident in most recent publications. Saying that, research on social SC aspects is lagging behind [21, 35, 36] while environmental considerations are predominant. Thus, there is a bias in SSCM research toward exploring environmental applications, which is consequently also reflected in this review article.

SCM as a research discipline has been approached from different directions, which resulted in a plethora of definitions [37, 38]. Similarly, various definitions of SSCM have been proposed over the years. Ahi and Searcy

[28] contrast these existing definitions, synthesize key elements, and define SSCM as:

the creation of coordinated SCs through the voluntary integration of economic, environmental, and social considerations with key inter-organizational business systems designed to efficiently and effectively manage the material, information, and capital flows associated with the procurement, production, and distribution of products or services in order to meet stakeholder requirements and improve the profitability, competitiveness, and resilience of the organization over the short- and long-term.

The definition shows that SSCM goes beyond traditional objectives of SCM as it involves the systemic coordination of SC resources and flows in accordance with sustainability considerations. Furthermore, the requirements of SC stakeholders have to be continuously assessed from short- and long-term perspectives. SC practice has to be informed by theoretical building blocks in order to deal with these complexities. Some useful higher level frameworks exist that outline supporting facets of sustainability in combination with SC considerations (Figure 1). As can be seen, application of SSCM demands that two or ideally all three sustainability dimensions are explicitly considered. This focus is also maintained

in this review article. In addition, more elaborate discussions of the theoretical underpinnings of SSCM or the development of this particular field of SC research can be found in the literature [13–15, 19–22, 28].

Network Structure

The network structure is a key element of SCM [16]. It is crucial to any SC and requires detailed information in order to make appropriate decisions including, for example, distribution strategies, the selection of warehouse locations and capacities, and production levels of products and plants. Furthermore, the transportation flows between different warehouses, production facilities, or SC members have to be determined. The general goal is to minimize the total costs related to production, inventory, and transportation while satisfying service-level requirements [26].

Winter and Knemeyer [14] elaborate on the importance of the network structure and interfirm relationships with regard to implementing sustainability. They summarize that a network needs to be flexible and adaptable based on partnerships in order to derive benefits for the SC. There is an emphasis on the need for coordinated and collaborative relationships between SC members. The relationships differ in that there may be ‘hard’

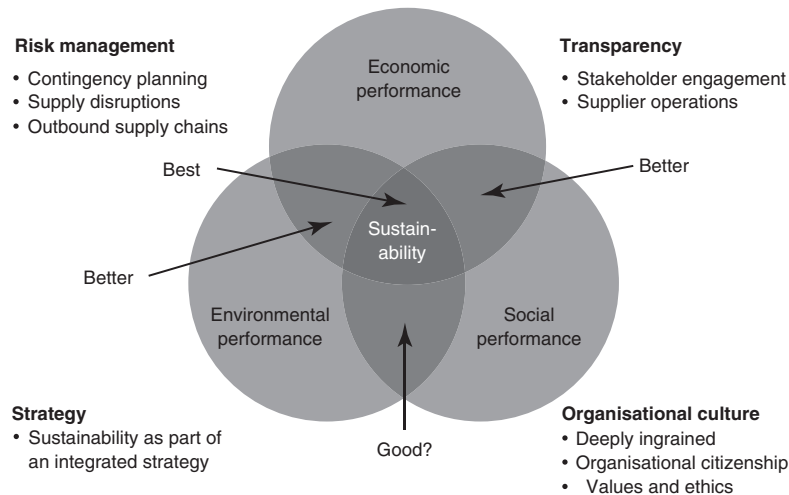


Figure 1. Facets of Sustainable Supply Chain Management [Source: Adapted from [19]].

relational ties such as material and monetary flows and ‘soft’ ties of friendship and information sharing. The traditional focus in SCM is on market and manufacturing issues and SC strategies are usually based on cheap transportation [39]. Poor design of transportation channels can lead to SC disruptions and higher environmental impacts. Requirements around reverse logistics, for example, returns and recycling, add further complexity [40]. Changes in market conditions, for example, rising fuel prices, can furthermore result in performance problems and risk of disruption demanding a sustainable orientation of the SC configuration [39]. Stringent controls of customer demand, manufacturing processes, and supplier performance have been associated with less uncertainty in the chain [31].

Schiller *et al.* [41] suggest to increase transport efficiencies through advanced transportation technologies, fleet differentiation, collaborative fleet management, load matching, and optimized scheduling. Potential outcomes are more appropriate vehicle selections depending on the transportation job while avoiding partial loadings and waiting times. Improved intermodal transitions can further minimize costs, congestion, noise, and pollution. More feasible supply sources and shortening of SCs present opportunities to mitigate the negative effects of distribution but may have to be promoted by policy efforts and weighed against purchasing costs, volatility risks, and infrastructural requirements [41, 42].

Reverse logistics is increasingly seen as an avenue to reduce SC waste and utilize resources from products at the end of their useful life [43]. Apart from utilizing obsolete products, reverse logistics can improve product designs as activities such as reuse, recycling, and remanufacturing come to the forefront [44]. Reverse logistics, however, also add to the complexities involved in network-related decisions [40]. Successful implementation of reverse logistics depends on whether monetary investments are warranted through economic benefits [30, 45]. Related research has revealed sustainability potential by reducing fluctuations of material costs, demand patterns, and repetitive

purchases [46]. The effectiveness of the idea has also been investigated by modeling the effectiveness of incentives such as rewards and advertising [47].

Another idea that influences network structures is industrial symbiosis or closed-loop SCs, which emphasizes reduction opportunities regarding system-level waste. It is based on utilizing the residual value of waste products from one operation as input material for other processes [30, 48]. Analyzing waste streams from an SC perspective can transform waste output from one company into value for the chain [48]. A high degree of cooperation and coordination, and hence exceptional relationship management, are required to facilitate such an improved flow of production by-products. In addition, any waste output has to be compatible with the input requirement in terms of quality, quantity, and logistics requirements. The high degree of interdependence also makes the system vulnerable to changing market requirements, for example, demand swings or price changes. Hence, industrial symbiosis holds potential in certain SC constellations but has limitations in terms of potential market risk and implementation difficulties [30].

Inventory Control

Inventories uncouple “the various phases of the process or service delivery system and thereby allow each to work independently of the other parts” [27]. Inventory control is vital for an SC as chain members have different requirements in regard to inventory size, reorder times, and demand of their process-related inventory and support inventory.

Inventory control is tightly linked with information technology (IT) and the availability of demand and supply information. These are crucial for reducing excessive inventory levels and associated costs caused by spoilage, insurance, theft, and obsolescence [49, 50]. Inventory control is therefore a priority for best-in-class companies and emphasis should be placed on accurate demand planning and forecasting to ensure optimized inventory levels [51, 52]. Demand amplification and the associated swings in inventories are also detrimental to sustainability goals and associated risks should

be decreased by better communication, inventory-related information exchange, and accepted mitigation strategies [53, 54].

The reduction of travel distances can support sustainability goals [41], suggesting a move from centralized inventories to several decentralized facilities, that is, increase proximity and accessibility to reduce transportation needs. Third-party logistics providers can help consolidate consignments, thus reducing costs, when dealing with resulting smaller shipping quantities. In addition, in case of supply disruptions, having larger amounts of inventory stored in various locations can reduce lead times for sourcing replacements and ensure general availability. Downsides are potentially higher fixed costs as more capital is locked in inventory while running additional storage facilities is likely to increase operational costs. Inventory holding costs are often higher than the generally assumed 20–25% but can add up to 60% of the cost of an item that stays in inventory for 12 months [51].

SC disruptions have negative economic impacts on the affected organisations [55] and are often due to shortages in supply and insufficient inventory levels. Single sourcing and lean supply strategies have resulted in increased dependence on suppliers. While such strategies often facilitate better cost control and help eliminate non-value-adding activities [17, 56], they can also make an SC more vulnerable to disruptions caused by, for example, natural disasters or political upheaval [57]. Risk management has become a major challenge in SCM [58] and possible risk mitigation options include a diversification of suppliers to provide backup sources for products and materials. Encouraging suppliers to establish additional supply locations and adding buffer stock for pivotal materials can provide further support to counteract SC disruptions [57].

Supply Contracts

Supply contracts generally specify the conditions of the relationship between suppliers and customers including pricing and volume discounts, lead times, or quality standards. Individual companies are traditionally mainly concerned with their own profitability

resulting in little attention paid to the effect of one's strategy and decisions on other SC members.

Building strong relationships between SC partners is crucial for SSCM as each SC member must meet sustainability criteria [59]. One way to ensure appropriate codes of conduct and supplier sustainability performance are audits that can include accreditation requirements, inspection of facilities, or reviewing documentations [60]. On the basis of such performance evaluations, supply contracts may be awarded, additional assessments may be scheduled, or certain conditions may be imposed on the supplier. Larger companies often require all their suppliers to obtain accreditation standards that force them to change their practices accordingly [61]. Management standards can contribute to sustainability performance [62, 63] and Pagell and Wu [64] emphasize that supplier certifications can specifically address social sustainability requirements, for example, child labor, worker safety, or working conditions.

Many companies tend to engage in collaboration only in order to facilitate assessments, whereas engagement in collaborative SC practices holds the potential to foster sustainability directly [65]. Instead of imposing sanctions, failure to meet supplier compliance standards can be addressed by joint efforts to develop necessary skills and procedures. Collaboration in supply relationships can be fostered through appropriate SC incentives and education across the tiers of an SC [66]. This can prevent issues such as concealing noncompliance or damaging a supplier's reputation and competitiveness, which in turn might lead to deterioration in sustainability performance. Dedication to partnerships can result in more supply options, higher transparency, more accurate performance assessment, and overall closer buyer–supplier relationships [67, 68]. An SC segmentation approach can help manage the complexities of collaborative arrangements by concentrating on critical SC partners [69].

Distribution Strategies

Distribution and transportation account for a high percentage of SC-related costs, making the distribution strategy a key issue in SCM. Logistics operations have also been identified as an area where most organizations can make sustainability improvements [70]. Distribution costs largely depend on the location of trading partners and the selected modes of transportation. The five commercially used transportation modes, that is, rail, road, water, air, and pipeline, differ significantly in terms of price, speed, and their share of goods moved [27]. Figure 2 does not include considerations regarding travel distances and only displays an approximated mix for developed economies. However, it becomes evident that selection of a ‘less appropriate’ mode can be costly in terms of money, time, and resources.

Low cost solutions that ensure the availability of goods and services are paramount for any SC. However, distribution strategies purely targeted at economic goals have begun to become obsolete as the environmental/social impacts and regulations are not considered [70]. Different distribution options and mode selections may be synergistic but can also create trade-offs due to incompatibilities. For example, lean and green strategies are often seen as compatible due to their focus on waste reductions, for example, the reduction of inventory and associated sourcing, producing, transporting, packaging, and handling activities. However, certain lean strategies, such as Just-in-time, are focussed on small lot sizes that require in fact more transportation, packaging, and handling [56]. Distribution strategies are often aimed at high product mobility to compensate for the prevalence of global, long-distance SCs. From a sustainability point of view, there are arguments against long SCs,

for example, vulnerability against world events, higher emissions, and difficulties to measure and assign externalities [41].

The optimization of distribution systems has long been a key goal for researchers and various techniques and solutions have been proposed. While not necessarily targeted at SSCM, the findings may also be beneficial from a holistic sustainability perspective. Overviews of such mainly quantitative optimization techniques are provided in more targeted articles [70]. Additionally, a variety of measures that can foster more sustainable distribution strategies can be identified [41, 56, 71, 72]:

- Increasing the efficiency of freight vehicles and use of alternative fuels
- Improved strategies for maintenance, disposal, and operator skill improvement
- Differentiation of the fleet, for example, smaller and nonmotorized vehicles
- More efficient scheduling and vehicle use, that is, increase in utilization
- Improved intermodal services and facilities through network redesign
- Policy efforts to promote shifting the truck-train breakeven point toward rail
- Ocean shipping reforms
- Slow steaming
- Shifting airfreight to high speed rail services
- Shipment consolidation
- Reverse logistics
- Redistribution of sourcing
- Carrier selection based on sustainability considerations
- Shortening of SCs
- Policies to promote shortening distances and reducing volumes of freight

Price per tonne/mile	Low	Pipeline/water	Rail	Road	Air	High
Speed, door to door	Slow	Water	Rail	Road	Air	Fast
% of goods moved	High	Rail	Road	Pipeline	Water	Low

Figure 2. Comparison of Modes of Transportation [Source: Adapted from [27]].

- Optimization of lot sizes
- Reusable packaging and containers

This list provides a series of distribution alternatives for SCs to explore and it has been shown that several of these options are already considered and applied in practice [73]. Maintaining decentralized local capacity and allowing for some slack resources also allow SCs to cope with uncertainties and risks of disruption [74]. Intermodal transport along with more appropriate, efficient, and cleaner forms of transportation can mitigate sustainability issues in distribution. As an example, slow steaming strategies are now widely employed in international shipping in order to cut emissions [72]. While high utilization remains important when selecting transportation modes, emission efficiencies have to be considered. For example, ocean shipping and rail have carbon emissions that are only a fraction of that caused by road haulage and airfreight [75, 76]. This is obviously of economic and environmental benefit and can furthermore foster logistics integration in SCs [77, 78]. A study among practitioners found that aspects of transportation, warehousing, and distribution have been neglected so far and emissions or energy consumption are not generally considered for supplier selection. Initiatives and sustainability objectives are also often not extended to second- or third-tier SC members [58].

Strategic Partnerships

Strategic partnerships and SC integration are necessary in order to stay competitive as companies increasingly compete on a SC level [79, 80]. Feasible SC partnerships are crucial considering the ongoing trends toward outsourcing, shifts to countries with low labor costs, reduction of physical infrastructure and inventories, and the vital role of SCs in providing delivery advantage [26, 27]. In such cases, consideration needs to be placed on indirect costs affiliated with intercontinental or long-haul SCs, including, loading, unloading, customs clearance, or transferring goods to warehouses [81]. Determining factors for choosing overseas suppliers

include, for example, supply structures, real-estate-related factors, and the characteristics of the actual industry [82]. An SC can be considered integrated when all product or service transfers are coordinated to manage costs and inventory levels while maintaining customers' delivery requirements. Planning activities and execution processes can be integrated through collaborative planning and vendor managed inventory [27].

One avenue toward sustainable competitive advantage is to create collaborative long-term relationships between SC partners [67]. SC collaboration was also found to foster internal company collaboration and improve service performance [83]. Harris *et al.* [40] suggest more horizontal and vertical collaboration between logistics operators in order to tackle sustainability challenges. Lee [84] points out the importance of long-term strategic relationships in order to support sustainability initiatives. A complete rethink of an SC structure may be necessary "including cutting out middlemen, changing suppliers and supply locations, taking a more holistic view of costs and benefits, and taking a longer term view of supplier relationships" [84]. It is emphasized that initiatives need to reinforce each another and that SC problems cannot be solved in isolation. More social and environmental SC control has been associated with financial benefits and long-term relationships that in turn support economic sustainability [1, 85].

Active management of strategic partnerships is crucial for sustained value and continuous improvement. Especially SC leaders or companies that are associated with the final product have to select their business partners with particular care to protect their reputation as they will be held responsible for unsustainable behavior within their SCs [86]. Relationship management programs have to ensure that relationships stay healthy by providing a platform for problem resolution, mutually beneficial continuous improvement goals, and control over performance objectives [51]. Two-way communication and mutual active engagement are pointed out as key to success. In order to manage related risks, implementation guidelines to implement

SSC practices can provide useful assistance, for example, Zhu, Sarkis, and Lai [87] develop and validate several measurements and underlying factors for the implementation of green SCM practices. Robinson [88] points out that sustainability requires the ongoing dynamic capacity to respond adaptively to changes. Hence, SC agility needs to be considered, that is, customer sensitivity, virtual integration, process integration, and network integration [89,90]. Customer sensitivity corresponds to the requirement to understand and quickly respond to market requirements through collaborative initiatives. Virtual integration is about accessing and visualizing information, knowledge, and competencies on an SC-level facilitated by information systems. Process integration highlights the interdependence of SC members, which is required to integrate key processes with their SC partners in order to manage change within their own organization and across the SC. Lastly, network integration demands a common identity, compatible information architectures and structures, and commitment to cooperation as well as performance measurement [89,90].

Outsourcing and Procurement

Outsourcing and procurement strategies determine whether a company decides to manufacture in-house or to buy from outside sources. Lakenan *et al.* [91] point out that outsourcing manufacturing functions allows companies to concentrate on their core capabilities while leaving production to suppliers with specialized knowledge. Other motivational factors include economies of scale at the suppliers' end as production volumes can be aggregated from multiple customers allowing to buy input material in bulk and at lower costs.

Similar to earlier business developments, sustainability may eventually become the norm, that is, an order qualifier [13]. Thus, sustainability needs to be actively encouraged by buyers so that sustainability performance can be marketed as an order winner [92]. Hence, outsourcing and procurement are crucial aspects in SSCM as "a company is no more sustainable than its SC—that is, a

company is no more sustainable than the suppliers that are selected and retained by the company" [92]. To support sustainable purchasing decisions, Pagell *et al.* [93] therefore propose a purchasing portfolio matrix taking into account transaction cost economics, the resource-based view, and stakeholder theory. The inclusion of other internal business functions in sourcing decisions can support the purchasing department by soliciting feedback and information on business objectives and strategies [51]. Krause *et al.* [92] suggest including sustainability as a competitive priority in purchasing and construct a purchasing model for SSCs based on the categorization of purchases into strategic, bottleneck, leverage, and noncritical as outlined by Kraljic [94]. Traditional competitive priorities in purchasing include quality, cost, delivery, and flexibility, which should be extended to include innovation. As a result, requirements for sustainable purchasing are derived for different product categories [95]:

- For strategic items, a focus on innovation is necessary to emphasize sustainability requirements for new product developments. Closer collaboration and transfer of know-how are decisive factors along with commitment to sustainability from suppliers.
- Bottleneck items put buyers in a dependent position making it difficult to assert pressure on suppliers. Industry standards for sustainable operations can be encouraged instead.
- For leverage items, emphasis should be put on recycling and waste prevention. Know-how transfer of improved practices across the SC is therefore important.
- For noncritical items, it may be easier to establish sustainability requirements due to a large supply base. Procedures regarding supplier selection and retention need to be adjusted accordingly, for example, certifications as selection criterion.

These purchasing requirements emphasize the need for SC collaboration and

long-term commitments between buyers and suppliers. Pursuing a sustainable sourcing strategy can be complicated by dynamic market requirements and globalization trends. Main driving factors for outsourcing are cost, flexibility, and concentration on core capabilities [91]. Global sourcing is often seen as the only option to gain a competitive advantage and is pursued especially by larger companies. However, it is often not supported by integrated global sourcing strategies [96]. Trade-offs between sustainability goals are often necessary, for example, maintaining control on working conditions or dealing with an increase in shipments along with potentially longer distances [67]. Cultural differences can present obstacles while effective supplier–buyer relationships are needed, which can be fostered through trust, commitment, and long-term orientation [96,97]. Thus, to support sustainable purchasing in increasingly global SCs, one needs to investigate the role of culture in creating and maintaining long-term relationships. Cannon *et al.* [97] point out that many uncertainties exist in this regard and suggest that an understanding of cultural values may help develop and improve relationships and increase the likelihood of long-term success.

Risk mitigation, for example, preventing SC disruptions is tightly coupled with sourcing decisions [51,57]. Sourcing decisions can be supported by risk mitigation methodologies including the identification of risk elements, determination of associated risk probabilities, assessment of potential financial impacts, and the prioritization of risks for monitoring and prevention [51]. Local production has been suggested in order to increase self-reliance and security of supply [41]. Diversification of the supply base and higher safety stock for critical items [57] along with compliance management, production standards, and real-time communication [98] can also help absorb the risks associated with sourcing. An associated approach is the ‘supply chain risk management process’ that classifies risk into nine separate categories along with feasible mitigation strategies [99]. These include phases such as risk identification,

monitoring, and controlling through risk analysis. Risk factors in SC sourcing cannot be viewed as disconnected and managers should, hence, pay attention to supplier relationships, information sharing, trust building, and collaboration [100].

Product Design

Decisions associated with product design can directly influence, for example, inventory holding or transportation costs by giving consideration to economic packaging and transportation, concurrent and parallel processing, and standardization [26]. Through SC design efforts, SC members can benefit from each other’s expertise while being able to concentrate on their own competencies. Suppliers can align their processes to new product requirements before the design is finalized and point out potential problems early on. Customers, on the other hand, can influence the product development so that the final design will specifically meet their requirements [101].

Developments in SC practice show that sustainability concerns are being addressed by modifying product design and packaging, incorporating sustainability initiatives into SC strategies, and establishing carbon management goals [58]. Product designs often reflect the prevalent market conditions during their creation, for example, products tend to be geared toward full replacement instead of repair when wages outweigh material costs. Extending product life helps reduce product obsolescence [102]. Modular product designs can be beneficial as they allow for easier repair, remanufacturing, and automated problem diagnosis. Furthermore, modular designs facilitate reuse, recycling, and disposal because of easy disassembly [103]. Reduction and reuse of product packaging also present opportunities to reduce SC impacts [76]. Factors that are often neglected in product design are by-products created during production, distribution, product use as well as disposal. SC integration can help mitigate associated negative impacts [67].

To support the aims of SSCM, product design along with product life cycle considerations can be of special importance [104]. Product life cycle management can help mitigate

sustainability risks by maintaining visibility, interoperability, and information exchange between SC partners and focussing on the impacts of a product throughout its entire life [42]. Life cycle assessment (LCA) can be useful to decrease adverse product impacts during preproduction, production, usage, and after disposal. It has been pointed out as one of the main methods to integrate environmental thinking into SCs since it emphasizes proactive behavior and demands careful supplier selection [42]. LCA thinking has also been associated with the creation of new business opportunities as previously unidentified problems may be recognized by SC partners [105]. However, there appears to be a lack of guidelines with regard to such integrations. SCs may require significant restructuring to integrate LCA and different types of SC structures demand different kinds of LCAs [106]. SC practices have impacts during each of the product life cycle stages and product design should reflect this accordingly, for example, ascertain that products can be manufactured, transported, used, and disposed of through sustainable practices [42]. It has, however, also been pointed out that LCA does not consider the entire SC and that a shift toward a holistic SC focus is required [71].

Design strategies targeting sustainability challenges include “Design for Environment” (DfE). DfE aims to reduce the amount of energy and material needed for the provision of goods and services by including environmental considerations in all design stages, for example, from project definition to concept and prototype development, field testing, and commercial launch [104]. Regulations already force companies to recycle or to be responsible for disposal of their products; for example, legislations in the European Union require manufacturers to design their cars so that 95% can be recycled [107, 108]. Thus, DfE and LCA are tightly connected and are useful concepts for developing and maintaining SSCs [104]. Borchardt *et al.* [109] point out that DfE is rarely considered by small and midsized companies due to, for example, lack of expertise, appropriate design tools, or missing know-how about how to change existing processes. Such difficulties have to be overcome in SCs as products are often

designed and built through an SC effort. Larger SC members with more resources and expertise may need to support their smaller SC partners in order to fully utilize the advantages of DfE and LCA.

Information Technology

Company success increasingly depends on SC coordination as much of the revenue is usually generated through the SC [31]. IT is one of the key aspects that drive SC efficiency nowadays. Information sharing within firms and across SCs is needed in order to enable and to fully exploit the power of SC integration [110]. To truly support the aims of the SC, it becomes increasingly important to store and communicate the “right kind” of information. Organizations and supply networks need to share information so that they are able to adapt quickly to changing requirements but are not overwhelmed by unnecessary data [26, 111]. IT systems should support easy retrieval and viewing of required data in order to make sound business decisions and avoid complicated workarounds [51]. Appropriate IT is of importance for exchanging inventory data and sales information. It can thereby support the reduction of cycle times and inventory costs while improving order fulfilment and customer service [50, 112]. Thus, information sharing impacts economic performance, for example, inventory levels and respective costs can decrease with an increase in the level of information sharing [49]. With smaller organizations also being able to afford and adopt IT, efficient information sharing and communication within the SC has been increasing [113, 114].

The control of sustainable practices is complicated in SCs [67] demanding compatible and user-friendly IT to support performance measurement. Considering the linkages and dynamics in a system like an SC is of special importance for accurate performance assessments [115]. Performance measurements in SCs have, however, been described as being beset by a lack of connection with strategy, focus on cost to the detriment of noncost indicators, lack of a balanced approach, insufficient focus on customers and competitors, focus

on local optimization, and a lack of system thinking [116]. In spite of such challenges, accurate performance measurements are essential for any improvement activity [117]. Indicators are commonly used in the context of sustainability in order to analyze progress and to communicate developments [118]. Thus, the effective communication and evaluation of an SSCM strategy requires a common language, facilitated through the use of KPIs to support informing, steering, and controlling [119]. Appropriate selection of performance measures as part of an SSCM strategy is essential to prevent functional silos and one-sided assessments. Standard measures to assess supplier performance include, for example, profitability, growth, service levels, technology use, or trade volumes [113]. However, SC performance cannot be holistically measured by financial ratios and logistics indicators alone but is affected by intraorganizational issues, the quality of relationships between SC members, as well as demands from customers and stakeholders [77]. Qualitative indicators are often neglected as it is sometimes neither financially viable nor practical to record qualitative data [119] but demand attention as some sustainability aspects may only be captured qualitatively. In support of SSCM, IT must capture the total cost of SC activities from the extended SC including usage of resources and creation of by-products [102]. A balanced approach has been prescribed to accurately evaluate SC performance including financial and nonfinancial measures classified according to strategic, tactical, and operational levels [120–122]. Researchers have made some advances in this area and have described approaches to develop and implement balanced scorecards specifically targeted at supporting SSCM [123, 124].

Customer Value

It is important for any company to assess the value offered by their services and products as perceived by the customer. This perception can be categorized into conformance to requirements, product selection, price and brand, value-added services, and relationships and experiences. SC success as a whole is also dependent on the value

provided to the end-customer, value being “the measure of desire for a product and its related services” [110]. Bowersox *et al.* [110] therefore emphasize that firms have to extend their management practices beyond suppliers and include suppliers’ suppliers so that their views on resource needs and constraints, threats, opportunities, and weaknesses can be considered.

Integrating sustainability considerations as a strategic priority can lead to improvements with regard to image and reputation [125]. Customer interest in SC sustainability is increasing, for example, the majority of customers who engage with third-party logistic providers are sensitive to sustainability issues [73]. Potential advantages of an SSC can go beyond reputation benefits including reliable long-term supply sources, increased visibility and control throughout the SC, reduction of price and volume volatility, improved quality, and increased efficiency [84]. The customer base may furthermore be extended toward environmental and socially conscious customers [67]. These aspects may help ensure the profitability of each company in the SC.

On the flip side, the current focus on sustainability and associated consumer preferences has also led to green-washing practices that in turn resulted in growing scepticism among consumers [126]. In order to reap the benefits of sustainable practices, SSCs have to distinguish themselves from SCs employing green-washing practices. Sustainability communication can provide customers with information about a SC’s practices and values. Suitable instruments include sustainability reports akin to financial reports, certifications through independent organizations, or the use of eco-labels for ones offerings [126, 127]. Success depends on such instruments to influence the acceptance level of customers, that is, customers have to see value in a SC’s sustainability efforts [126]. Especially in larger companies it has become common to report on SC sustainability as part of their corporate sustainability reports [12]. Such voluntary efforts have been described as crucial in order to effectively address industrially induced problems [128].

Table 2. Overview of SSC Characteristics

SCM Elements	Summary of SSC Characteristics	Tools and References
Network configuration	Elimination of distribution-related wastes and incorporation of sustainability costs. Developing partnerships and collaborating toward SC flexibility and a true SC perspective reaping the benefits of an advanced SC configuration	Marketing and manufacturing issues [31] Developing partnerships [14] Increasing transportation efficiencies [41] Reverse logistics [43] Industrial symbiosis [30] Closed-loop SCs and green SC innovation [48] Reduction of transportation distances [41] Control and capacity considerations [49] SC disruptions [55] Demand amplification [54]
Inventory control	Strategic placement of inventory to ensure accessibility. This may include the reduction of distances, altering distribution paths, and mitigating SC disruptions	Relationship building [59] Standards and certifications [62–64] Collaboration, incentives, and supplier education [65, 66] Intermodal transportation [75]
Supply contracts	Outlining multidimensional sustainability requirements and emphasizing collaborative relationships. Introduction of common standards and performance assessments	Sustainable distribution strategies [41, 56, 71, 72] Sustainability and other SC approaches [56, 70] Collaborative relationships [83, 84] Active management and continuous improvement [51] Agility and ability for change [89, 90]
Distribution strategies	Emphasis on intermodal connections and a high degree of accessibility. Use of suitable transportation modes, reduced freight distances, and efficient freight vehicles	
Strategic partnerships	Integrated planning and collaborative environment among strategic SC partners. Active SC management in order to develop mutual benefits with suppliers and engage in continuous improvement	
Outsourcing and procurement	Sourcing and procuring to support long-term SC goals with sustainability as a core competency. Availability of mitigation plans to deal with changing market conditions and supply disruptions	Purchasing priorities [92, 93] Supply risks [98–100] Buyer–supplier relationships [97]
Product design	Considers full life cycle of products along with reverse logistics to eliminate occurring waste and ensure nonharmful design. Utilize product design to improve sustainability of the SC	Product design and life stages [102, 103] Life cycle assessment [42, 106] Design for environment [104, 109]
Information technology	Integrated sustainability-oriented systems that allow for access to all relevant SC data. IT that supports balanced performance assessments, effective communication, and information sharing	Measurement in SCs [113, 115, 116] Key performance indicators [118, 119] Balanced performance measurement [120, 121] Balanced scorecard for SSCM [123, 124]
Customer value	The customer perception of value drives SC processes. Customer value is supported by improved SC performance, collaboration, long-term relationships, sustainability measure, and their communication	Sustainability and the customer [67, 73, 84, 125] Green washing practices [126] Sustainability reporting and compliance management [12, 128]

OVERVIEW OF KEY FINDINGS

The previous section provided an overview of important SC issues and sustainability concerns, outlining potential avenues for practitioners on how to move their SCs into a more sustainable direction. The interconnected nature of SC elements, sustainability concerns, associated requirements, and potential actions to take is also emphasized. It presents a starting point for decision makers in SCs that can guide a long-term SC strategy as well as operational decisions. As a key insight, this review has shown that companies should avoid disconnected ad hoc sustainability initiatives in their SCs. They should instead focus on a holistic SSC strategy with a long-term focus and align their practical choices accordingly.

On the basis of the review of key SCM elements through a sustainability lens, a summary (Table 2) was constructed that summarizes SSCM practices and identifies related reference material.

CONCLUDING COMMENTS

It can be expected that sustainability in general, and in extension the topic of SSCM, will continue to be of interest for researchers and practitioners alike due to the persisting nature of the underlying causal linkages [129]. The requirement to consider sustainability has been widely recognized by regulatory bodies and companies and is also increasingly demanded by consumers. As this review has shown, SCs have a crucial role in this endeavor and are well positioned to support sustainable development due to their wide-ranging impacts and influences. Decision makers in SCs are therefore tasked with strategic sustainability orientations and operational shifts. The thematic review of the literature has shown that traditional SC structures and operations may have to be redesigned in order to face current and future sustainability challenges. It became evident that the focus of many approaches is on isolated SC issues and it is usually not clear how these could be tailored and integrated as part of a more complete SSCM strategy. SCs and

sustainability requirements are both characterized by complex interactions that have to be understood in order to shape such a strategy. Customizable, prescriptive frameworks and models are required that can guide and facilitate strategic SSCM transformation and development while embedding more specific tools and methods at operational levels.

While this article can only provide a brief overview, many applicable methods, tools, and sustainable practices could be identified. The findings of this review can be of use for academic researchers and SC practitioners alike. Researchers are presented with a summary of methods and tools that hold the potential to support SSCM. It thereby informs academics of the current state of the field and points them toward new research directions. Practitioners can capitalize on this study by reviewing approaches that have been found feasible to support SSCM. While actual implementations in SCs depend on particular contexts, practitioners can use this review to identify potential avenues for sustainability development and understand the situations in which tools or ideas have been successfully applied in practice. Readers should be aware that the field of SSCM is characterized by rapid expansion. Research insights and SC practices continuously develop, making a truly comprehensive overview impossible. Hence, it needs to be acknowledged that ideas may have been overlooked in this review due to the dynamic nature of the field and the reliance on primarily academic sources. SSCM scholars are therefore advised to actively update their understanding as required based on new additions to the literature.

REFERENCES

1. McIntyre K. *Delivering sustainability through supply chain management*. In: Waters D, editor. Volume 245–260, *Global Logistics—New directions in supply chain management*. London, England: Kogan Page; 2007.
2. WCED. *Our common future*. New York, NY: The World Commission on Environment and Development (WCED)/Oxford University Press; 1987.

3. Viederman S. *Knowledge for sustainable development: What do we need to know?*. In: Trzyna TC, Osborn JK, editors. *A sustainable world: Defining and measuring sustainable development*. Sacramento, CA: IUCN—the World Conservation Union by the International Center for the Environment and Public Policy, California Institute of Public Affairs; 1995. pp. 37–43.
4. Mebratu D. *Sustainability and sustainable development: Historical and conceptual review*. *Environ Impact Assess Rev* 1998;18(6):493–520.
5. Nagpal T, Foltz C. *Choosing our future: Visions of a sustainable world*. Washington, DC: World Resources Institute; 1995. pp. 1–181.
6. Marshall JD, Toffel MW. *Framing the elusive concept of sustainability: A sustainability hierarchy*. *Environ Sci Technol* 2005;39(3):673–682.
7. Klein K, Kozlowski SW. *Multilevel theory, research, and methods in organizations*. San Francisco, CA: Jossey-Bass; 2000.
8. Starik M, Rands GP. *Weaving an integrated web: Multilevel and multisystem perspectives of ecologically sustainability organizations*. *Acad Manage Rev* 1995;20(4):908–935.
9. Mahler D. *The sustainable supply chain*. *Supply Chain Manag Rev* 2007;11(8):59–60.
10. Mahler D, Callieri C, Erhard A. *Chain reaction: Your firm cannot be “sustainable” unless your supply chain becomes sustainable first*. Chicago, IL: AT Kearney; 2007. pp. 1–8.
11. Beard A et al. *It's hard to be good*. *Harv Bus Rev* 2011;89(11):88–96.
12. KPMG. *The KPMG survey of corporate responsibility reporting 2013*. Amsterdam, The Netherlands: KPMG; 2013.
13. Carter CR, Easton PL. *Sustainable supply chain management: Evolution and future directions*. *Int J Phys Distrib Logist Manag* 2011;41(1):46–62.
14. Winter M, Knemeyer AM. *Exploring the integration of sustainability and supply chain management: Current state and opportunities for future inquiry*. *Int J Phys Distrib Logist Manag* 2013;43(1):18–38.
15. Ashby A, Leat M, Hudson-Smith M. *Making connections: a review of supply chain management and sustainability literature*. *Supply Chain Manag Int J* 2012;17(5):497–516.
16. Cooper MC et al. *Meshing multiple alliances*. *J Bus Logist* 1997;18(1):67–90.
17. Cooper R, Slagmulder R. *Supply chain development for the lean enterprise: Interorganizational cost management*. Strategies in confrontational cost management series. Portland, OR: Productivity Press xxxii; 1999. pp. 1–510.
18. Fink A. *Conducting research literature reviews: From the internet to paper*. 2nd ed. Thousand Oaks, CA: Sage Publications; 2005.
19. Carter CR, Rogers DS. *A framework of sustainable supply chain management: Moving toward new theory*. *Int J Phys Distrib Logist Manag* 2008;38(5):360–387.
20. Hassini E, Surti C, Searcy C. *A literature review and a case study of sustainable supply chains with a focus on metrics*. *Int J Prod Econ* 2012;140(1):69–82.
21. Seuring S. *A review of modeling approaches for sustainable supply chain management*. *Decis Support Syst* 2013;54(4):1513–1520.
22. Seuring S, Müller M. *From a literature review to a conceptual framework for sustainable supply chain management*. *J Clean Prod* 2008;16(15):1699–1710.
23. Monczka RM, Trent RJ, Handfield RB. *Purchasing and supply chain management*. Cincinnati, OH: South-Western College Publishing; 1998.
24. Christopher M. *Logistics and supply chain management: Creating value-adding networks*. 3rd ed. Harlow, England: Financial Times/Prentice Hall; 2005. pp. 1–305.
25. Mentzer JT et al. *Defining supply chain management*. *J Bus Logist* 2001;22(2):1–25.
26. Simchi-Levi D, Kaminsky P, Simchi-Levi E. *Designing & managing the supply chain: Concepts, strategies, and case studies*. 2nd ed. New York, NY: McGraw-Hill; 2003.
27. Hill T, editor. *Operations management: Strategic context and managerial analysis*. 1st ed. Macmillan Press: Basingstoke, England; 2000. pp. 1–704.
28. Ahi P, Searcy C. *A comparative literature analysis of definitions for green and sustainable supply chain management*. *J Clean Prod* 2013;52:329–341.
29. Ketokivi M, Mantere S. *Two strategies for inductive reasoning in organizational research*. *Acad Manage Rev* 2010;35(2):315–333.
30. Bansal P, McKnight B. *Looking forward, pushing back and peering sideways: Analyzing the sustainability of industrial symbiosis*. *J Supply Chain Manag* 2009;45(4):26–37.

31. Lambert DM, Cooper MC. *Issues in supply chain management*. Ind Mark Manag 2000;29(1):65–83.
32. Jayaraman V, Klassen R, Linton JD. *Supply chain management in a sustainable environment*. J Oper Manag 2007;25(6):1071–1074.
33. Dey A, LaGuardia P, Srinivasan M. *Building sustainability in logistics operations: A research agenda*. Manag Res Rev 2011;34(11):1237–1259.
34. Harms D, Hansen EG, Schaltegger S. *Strategies in sustainable supply chain management: An empirical investigation of large German companies*. Corp Soc Responsib Environ Manag 2012;20(4):205–218.
35. Sarkis J. *A boundaries and flows perspective of green supply chain management*. Supply Chain Manag Int J 2012;17(2):202–216.
36. Carbone V, Moatti V, Vinzi VE. *Mapping corporate responsibility and sustainable supply chains: An exploratory perspective*. Bus Strateg Environ 2012;21(7):475–494.
37. Croom S, Romano P, Giannakis M. *Supply chain management: An analytical framework for critical literature review*. Eur J Purch Supply Manag 2000;6(1):67–83.
38. Vaaland TI, Heide M. *Can the SME survive the supply chain challenges?* Supply Chain Manag Int Journal 2007;12(1):20–31.
39. Halldórsson Á, Kovács G. *The sustainable agenda and energy efficiency: Logistics solutions and supply chains in times of climate change*. Int J Phys Distr Log Manag 2010;40(1/2):5–13.
40. Harris I et al. *Restructuring of logistics systems and supply chains*. In: McKinnon A et al., editors. *Green logistics: Improving the environmental sustainability of logistics*. London, England: Kogan Page Limited; 2010. pp. 101–123.
41. Schiller PL, Bruun EC, Kenworthy JR. *An introduction to sustainable transportation: Policy, planning and implementation*. London, England: Earthscan; 2010.
42. Badurdeen F, Metta H, Gupta S. *Taxonomy of research directions for sustainable supply chain management*. IIE Annual Conference Proceedings. 2009. pp. 1–1256.
43. Kocabasoglu C, Prahinski C, Klassen RD. *Linking forward and reverse supply chain investments: The role of business uncertainty*. J Oper Manag 2007;25(6):1141–1160.
44. Swee Siong K, Sev Verl N, Yousef A. *Sustainable supply chain for collaborative manufacturing*. J Manuf Technol Manag 2011;22(8):984–1001.
45. Geyer R, Wassenhove LNV, Atasu A. *The economics of remanufacturing under limited component durability and finite product life cycles*. Manag Sci 2007;53(1):88–100.
46. Jayant A, Gupta P, Garg SK. *Reverse Supply Chain Management (R-SCM): Perspectives, Empirical Studies and Research Directions*. Int J Bus Insights Transform 2011;4(2):111–125.
47. Zeng AZ. *Coordination mechanisms for a three-stage reverse supply chain to increase profitable returns*. Nav Res Log 2013;60(1):31–45.
48. Jensen JK, Munksgaard KB, Arlbjørn JS. *Chasing value offerings through green supply chain innovation*. Eur Bus Rev 2013;25(2):124–146.
49. Wu YN, Edwin Cheng TC. *The impact of information sharing in a multiple-echelon supply chain*. Int J Prod Econ 2008;115(1):1–11.
50. Zhao X, Xie J, Zhang WJ. *The impact of information sharing and ordering co-ordination on supply chain performance*. Supply Chain Manag Int J 2002;7(1):24–40.
51. Engel B. *10 best practices you should be doing now*. CSCMP's Supply Chain Q 2011;5(1):48–53.
52. Chen, C.L., Yuan, T.Y., and Chang, C.Y. et al., *A multi-criteria optimization model for planning of a supply chain network under demand uncertainty*. Computer Aided Chemical Engineering, 2006, 21 (C), pp. 2075–2080. doi: 10.1016/S1570-7946(06)80354-8.
53. Lee HL, Padmanabhan V, Whang S. *Information distortion in a supply chain: The bullwhip effect*. Frontier Res Manuf Log 1997;43(4):546–558.
54. Lee HL, Padmanabhan V, Whang S. *The bullwhip effect in supply chains*. MIT Sloan Manag Rev 1997;38(3):93–102.
55. Hendricks KB, Singhal VR. *An empirical analysis of the effect of supply chain disruptions on long-run stock price performance and equity risk of the firm*. Prod Oper Manag 2005;14(1):35–52.
56. Mollenkopf D et al. *Green, lean, and global supply chains*. Int J Phys Distr Log Manag 2010;40(1/2):14–41.
57. Cooke, J.A. (2011) *Lessons from Japan's earthquake*. CSCMP's Supply Chain Quarterly. Quarter 2. <http://www.supplychain->

- quarterly.com/columns/20110525lessons_from_japans_earthquake/
58. IBM Global Services. Services IG, editor. *The smarter supply chain of the future - global chief supply chain officer study*. Somers, NY: IBM Corporation; 2009. p. 68.
 59. Hall J. *Environmental supply chain dynamics*. J Clean Prod 2000;8(6):455–471.
 60. Smith G, Feldman D. *Company codes of conduct and international standards: An analytical comparison*. In: T.W.B. Group, editor. In: *Corporate Social Responsibility Practice*. Washington, DC: The World Bank; 2003.
 61. Sroufe R, Curkovic S. *An examination of ISO 9000:2000 and supply chain quality assurance*. J Oper Manag 2008;26(4):503–520.
 62. Corbett CJ, Kirsch DA. *International diffusion of ISO 14000 certification*. Prod Oper Manag 2001;10(3):327–342.
 63. Zhu Q, Sarkis J, Geng Y. *Green supply chain management in China: Pressures, practices and performance*. Int J Oper Prod Manag 2005;25(5):449–468.
 64. Pagell M, Wu Z. *Building a more complete theory of sustainable supply chain management using case studies of 10 exemplars*. J Supply Chain Manag 2009;45(2):37–56.
 65. Gimenez C, Tachizawa EM. *Extending sustainability to suppliers: a systematic literature review*. Supply Chain Manag Int J 2012;17(5):531–543.
 66. Rao P, Holt D. *Do green supply chains lead to competitiveness and economic performance?* Int J Oper Prod Manag 2005;25(9):898–916.
 67. Faisal MN. *Sustainable supply chains: A study of interaction among the enablers*. Bus Process Manag J 2010;16(3):508–529.
 68. Jørgensen HB et al. *Strengthening implementation of corporate social responsibility in global supply chains*. In: T.W.B. Group, editor. In: *Corporate Social Responsibility Practice*. Washington, DC: The World Bank; 2003.
 69. Barratt M. *Understanding the meaning of collaboration in the supply chain*. Supply Chain Manag Int J 2004;9(1):30–42.
 70. Validi S, Bhattacharya A, Byrne PJ. *A case analysis of a sustainable food supply chain distribution system—A multi-objective approach*. Int J Prod Econ 2014;152:71–87.
 71. Colicchia C, Melacini M, Perotti S. *Benchmarking supply chain sustainability: Insights from a field study*. Benchmarking Int J 2011;18(5):705–732.
 72. Cariou P. *Is slow steaming a sustainable means of reducing CO₂ emissions from container shipping?* Transp Res Part D: Transp Environ 2011;16(3):260–264.
 73. Lieb KJ, Lieb RC. *Environmental sustainability in the third-party logistics (3PL) industry*. Int J Phys Distr Log Manag 2010;40(7):524–533.
 74. Jüttner U, Maklan S. *Supply chain resilience in the global financial crisis: An empirical study*. Supply Chain Manag Int J 2011;16(4):246–259.
 75. Winebrake JJ et al. *Assessing energy, environmental, and economic tradeoffs in intermodal freight transportation*. J Air Waste Manag Assoc 2008;58(8):1004–1013.
 76. Forum WE. In: Doherty S, Hoyle S, editors. *Supply chain decarbonization: The role of logistics and transport in reducing supply chain carbon emissions*. Geneva, Switzerland: World Economic Forum; 2009. pp. 1–41.
 77. de Brito MP, Carbone V, Blanquart CM. *Towards a sustainable fashion retail supply chain in Europe: Organisation and performance*. Int J Prod Econ 2008;114(2):534–553.
 78. Lambert DM. *The supply chain management and logistics controversy*. In: Brewer AM, Button KJ, Hensher DA, editors. *Handbook of logistics and supply chain management*. Amsterdam, The Netherlands: Pergamon Press; 2001. pp. 99–126.
 79. Chen IJ, Paulraj A. *Towards a theory of supply chain management: The constructs and measurements*. J Oper Manag 2004;22(2):119–150.
 80. Leenders MR. *Purchasing and supply management: With 50 supply chain cases*. 13th ed. New York, NY: McGraw-Hill; 2006. pp. 1–564.
 81. Co HC et al. *A continuous-review model for dual intercontinental and domestic outsourcing*. Int J Prod Res 2011;50(19):5460–5473.
 82. Brown RS. *Does institutional theory explain foreign location choices in fragmented industries?* J Int Bus Res 2011;10(1):59+.
 83. Stank TP, Keller SB, Daugherty PJ. *Supply chain collaboration and logistical service performance*. J Bus Logist 2001;22(1):29–48.
 84. Lee HL. *Embedding sustainability: Lessons from the front line*. Int Commerce Rev ECR J 2008;8(1):10–20.

85. Porter ME, van der Linde C. *Green and competitive: Ending the stalemate*. Harv Bus Rev 1995;73(5):120–134.
86. Handfield RB, Nichols EL. *Introduction to supply chain management*. Upper Saddle River, NJ: Prentice Hall; 1999.
87. Zhu Q, Sarkis J, Lai K-H. *Confirmation of a measurement model for green supply chain management practices implementation*. Int J Prod Econ 2008;111(2):261–273.
88. Robinson J. *Modelling the interactions between human and natural systems*. Int Soc Sci J 1991;43(4):629–647.
89. van Hoek RI, Harrison A, Christopher M. *Measuring agile capabilities in the supply chain*. Int J Oper Prod Manag 2001;21(1/2):126–147.
90. Yusuf YY *et al*. *Agile supply chain capabilities: Determinants of competitive objectives*. Eur J Oper Res 2004;159(2):379–392.
91. Lakenan B, Boyd D, Frey E. *Why Cisco fell: Outsourcing and its perils*. Strategy Bus 2001;3rd quarter(24):1–12.
92. Krause DR, Vachon S, Klassen RD. *Special topic forum on sustainable supply chain management: Introduction and reflections on the role of purchasing management*. J Supply Chain Manag 2009;45(4):18–25.
93. Pagell M, Wu Z, Wasserman ME. *Thinking differently about purchasing portfolios: An assessment of sustainable sourcing*. J Supply Chain Manag 2010;46(1):57–73.
94. Kraljic P. *Purchasing must become supply management*. Harv Bus Rev 1983;61(5):109–117.
95. Krause DR, Pagell M, Curkovic S. *Toward a measure of competitive priorities for purchasing*. J Oper Manag 2001;19(4):497–512.
96. Trent RJ, Monczka RM. *International purchasing and global sourcing - what are the differences?* J Supply Chain Manag 2003;39(4):26–36.
97. Cannon JP *et al*. *Building long-term orientation in buyer-supplier relationships: The moderating role of culture*. J Oper Manag 2010;28(6):506–521.
98. Olson DL, Wu D. *Risk management models for supply chain: A scenario analysis of outsourcing to China*. Supply Chain Manag Int J 2011;16(6):401–408.
99. Tummala R, Schoenherr T. *Assessing and managing risks using the Supply Chain Risk Management Process (SCRMP)*. Supply Chain Manag Int J 2011;16(6):474–483.
100. Faisal MN, Banwet DK, Shankar R. *Supply chain risk mitigation: modeling the enablers*. Bus Process Manag J 2006;12(4):535–552.
101. Hammer M. *The superefficient company*. Harv Bus Rev 2001;79(8):82–91.
102. Linton JD, Klassen R, Jayaraman V. *Sustainable supply chains: An introduction*. J Oper Manag 2007;25(6):1075–1082.
103. Kleindorfer PR, Singhal K, Van Wassenhove LN. *Sustainable operations management*. Prod Oper Manag 2005;14(4):482–492.
104. Bevilacqua M, Ciarapica FE, Giacchetta G. *Design for environment as a tool for the development of a sustainable supply chain*. Int J Sustain Eng 2008;1(3):188–201.
105. Birkin F, Polesie T, Lewis L. *A new business model for sustainable development: An exploratory study using the theory of constraints in nordic organizations*. Bus Strateg Environ 2009;18(5):277–290.
106. Hagelaar GJLF, van der Vorst JGAJ. *Environmental supply chain management: Using life cycle assessment to structure supply chains*. Int Food Agribus Manag Rev 2001;4(4):399–412.
107. European Union. *Directive 2000/53/EC of the European Parliament and of the Council of 18 September 2000 on end-of life vehicles*. Brussels: European Union; 2000.
108. European Union. *Directive 2002/96/EC of the European Parliament and of the Council of 27 January 2003 on waste electrical and electronic equipment*. Brussels: European Union; 2003.
109. Borchardt M *et al*. *Redesign of a component based on ecodesign practices: Environmental impact and cost reduction achievements*. J Clean Prod 2010;19(1):49–57.
110. Bowersox DJ, Closs DJ, Stank TP. *Ten megatrends that will revolutionize supply chain logistics*. J Bus Logist 2000;21(2):1–16.
111. Kolb DG, Collins PD, Lind EA. *Requisite connectivity: Finding flow in a not-so-flat world*. Organ Dyn 2008;37(2):181–189.
112. Stein, T. and Sweat, J., *Killer supply chains*. Information Week, 1998, 708(9), pp. 36–46.
113. Haug A, Pedersen A, Arlbjørn JS. *ERP system strategies in parent-subsidiary supply chains*. Int J Phys Distr Log Manag 2010;40(4):298–314.
114. Búrca SD, Fynes B, Marshall D. *Strategic technology adoption: extending ERP across the supply chain*. J Enterp Inf Manag 2005;18(4):427–440.

115. Singh R *et al.* *An overview of sustainability assessment methodologies.* *Ecol Indic* 2009;9(2):189–212.
116. Shepherd C, Günter H. *Measuring supply chain performance: Current research and future directions.* *Int J Product Perform Manag* 2006;55(3/4):242–258.
117. McCormack K, Ladeira MB, de Oliveira MPV. *Supply chain maturity and performance in Brazil.* *Supply Chain Manag Int J* 2008;13(4):272–282.
118. Segnestam L. *Indicators of environment and sustainable development - theories and practical experience.* Washington, DC: The World Bank; 2002. pp. 1–61.
119. Sürle C, Wagner M. *Supply chain analysis.* In: Stadler H, editor. *Supply chain management and advanced planning.* Berlin, Germany: Springer; 2008. pp. 37–64.
120. Gunasekaran A, Patel C, Tirtiroglu E. *Performance measures and metrics in a supply chain environment.* *Int J Oper Prod Manag* 2001;21(1/2):71–87.
121. Kaplan RS, Norton DP. *Putting the balanced scorecard to work.* *Harv Bus Rev* 1993;71(5):134–142.
122. Kaplan RS, Norton DP. *Using the balanced scorecard as a strategic management system.* *Harv Bus Rev* 1996;74(1):75–85.
123. Cetinkaya B. *Developing a sustainable supply chain strategy.* In: *Sustainable supply chain management.* Berlin, Heidelberg: Springer; 2011. pp. 17–55.
124. Reefke H, Trocchi M. *Balanced Scorecard for Sustainable Supply Chains: Design and Development Guidelines.* *Int J Product Perform Manag* 2013;62(8):805–826.
125. Chinander KR. *Aligning accountability and awareness for environmental performance in operations.* *Prod Oper Manag* 2001;10(3):276–291.
126. Blengini GA, Shields DJ. *Green labels and sustainability reporting: Overview of the building products supply chain in Italy.* *Manag Environ Qual* 2010;21(4):477–493.
127. Global Reporting Initiative. *Sustainability reporting guidelines, in Version 3.* Amsterdam, The Netherlands: Global Reporting Initiative; 2008.
128. Shrivastava P. *The role of corporations in achieving ecological sustainability.* *Acad Manag Rev* 1995;20(4):936–960.
129. Drake, D.F. and S. Spinler, *OM Forum—Sustainable Operations Management: An enduring stream or a passing fancy?* Manufacturing & Service Operations Management, 2013.

A SOCIETAL COST OF OBESITY IN TERMS OF AUTOMOBILE FUEL CONSUMPTION

SHELDON H. JACOBSON
Department of Computer Science,
University of Illinois at
Urbana-Champaign, Urbana

DOUGLAS M. KING
Department of Industrial and
Enterprise Systems Engineering,
Simulation and Optimization
Laboratory, University of Illinois,
Urbana, Illinois

The high levels of oil consumption and obesity in the United States have become important socioeconomic concerns. In June 2008, the national average price of regular unleaded gasoline exceeded US\$4 for the first time in the United States, more than tripling its price since June 2003 [1]. These record fuel prices, coupled with growing concerns over carbon emissions and their role in global warming, have piqued the national interest in reducing fuel consumption. At present, the United States is responsible for approximately 25% of the world's daily oil consumption, with two-thirds of this amount devoted to the transportation sector [2]; these facts suggest that changes in the transportation sector can lead to substantial reductions in oil consumption.

As oil consumption has become a national concern, so have the growing rates of obesity in the population, which have been rising for several decades. In 2004, more than 32% of US adults were estimated to be obese, up from 12% in 1991 [3,4]. This increase has occurred despite increasing evidence that individuals who are obese are more likely to suffer from health conditions such as coronary heart disease, type 2 diabetes, and high blood pressure [5], as well as the proliferation of government programs designed to promote healthier lifestyles.

These two issues may seem unrelated. However, several studies have demonstrated

that there is a link between the incidence of obesity and automobile use. A study of transportation and health trends in California revealed that the prevalence of obesity was higher for individuals who did more automobile travel [6]. The relationship between obesity and fuel prices was further investigated by Courtemanche [7], who estimated that a US\$1 increase in real gasoline prices will reduce obesity in the United States by 15% in five years, by increasing exercise due to discouraged automobile use, as well as by decreasing the number of meals eaten at restaurants.

One study has shown the effect of pedestrian-friendly urban planning on obesity rates and automobile use. This study, conducted in King County, Washington, found that both obesity and automobile use were reduced in areas where walking was an effective mode of transportation [8]. The effectiveness of walking is measured by a "walkability index" that depends on residential density, street connectivity, type of land use, and retail floor area ratio. While this study indicates that both obesity and automobile use can be reduced by creating a pedestrian-friendly environment, individuals living in the inner city, where residential density and street connectivity are high, also have a high prevalence of obesity [9], which suggests that the relationship between obesity and the built environment is not a simple one.

In addition to these studies, which establish a correlation between individual travel behavior and obesity, there is also a direct relationship between obesity and fuel consumption. Adding passenger weight to a vehicle tends to degrade fuel economy, thereby increasing fuel consumption. Therefore, increasing obesity among a vehicle's passengers will require additional fuel consumption during travel. This relationship has been explored in the literature in both airline and highway travel. Average weight gain in the United States during the 1990s accounted for 350 million gallons of jet fuel consumed during the year 2000, about

2.4% of the total volume of jet fuel used in domestic service [10], while average weight gain since the 1960s accounts for up to 938 million gallons of the gasoline consumed by cars and light trucks each year [11].

All historical weight gain, however, is not necessarily unhealthy. For example, if weight were only gained by individuals classified as underweight, average weight in the population would increase, but obesity rates would not. In fact, this weight gain would likely reflect an improvement in national health. Furthermore, an adult is considered overweight if their body mass index (BMI) is greater than 25, where BMI is computed by dividing height in meters by the square of weight in kilograms. Between 1960 and 2002, the height of the average adult in the United States increased by approximately one inch [12]; this increase in height should partially offset increasing obesity rates by reducing average BMI.

The goal of this study is to quantify the additional fuel consumption in the United States each year due to passenger overweight and obesity, as well as due to *historical weight gain* in the US population, which is estimated by the change in average weight over a specified time period. Fuel consumed by noncommercial passenger highway vehicles (i.e., cars and light trucks) was considered in the analysis. The analysis is based on the methodology presented by Jacobson and McLay [11], which estimates the fuel consumption that can be attributed to changes in passenger weight. There are four components to this methodology:

1. Average extra weight in the US population is estimated for six combinations of age (2–14 years, 15–19 years, 20–74 years) and gender (male and female).
2. Average extra weight per vehicle passenger is estimated by a probability model that considers the age and gender distribution of vehicle passengers. Different models are used for driving and nondriving passengers.
3. The average extra passenger weight per car and light truck is estimated by combining the average extra weight per passenger with the average number of passengers per vehicle.
4. The amount of fuel consumption attributable to this extra passenger weight is assessed by determining its impact on average fuel economy, which can be translated into the related impact on fuel consumption.

These components will be discussed individually in later sections. The last three components follow directly from the model proposed by Jacobson and McLay [11], while the average extra weight in the US population is assessed using two different methods. The first method estimates extra weight as the historical weight gain in the US population over several time periods, comparing national average weight estimates in 2005–2006 with those dating back as far as the 1960s. This is the method used by Jacobson and McLay [11]; the results of that study are made more current through the use of weight and travel statistics that have been released since its publication. The second method provides a direct estimate for the *average weight attributable to overweight and obesity* in the US population, using the model introduced by Jacobson and King [13]. The carbon emissions due to this additional fuel consumption are computed to quantify how this additional fuel consumption affects the environment.

This paper is organized as follows: The section titled “Data-Sets and Methodology” describes the data used in the analysis. The section titled “Estimating Weight Due to Overweight and Obesity” describes a model for estimating the average weight in the US population that can be attributed to overweight and obesity, which is the second method used to compute average additional weight in the US population in the methodology described above. The section titled “Estimating Fuel Consumption due to Overweight and Obesity” summarizes the Jacobson and McLay [11] model for estimating fuel consumption, which discusses the remaining three components of the methodology. The section titled “Analysis” summarizes the fuel consumption and carbon emission estimated by applying this methodology

to current travel and weight statistics in the United States, as originally reported by Jacobson and King [13], and the section titled “Conclusions” gives the key conclusions of the analysis and discusses further implications of this study and the results reported.

DATA-SETS AND METHODOLOGY

This section summarizes the data used within this study, which fall into three categories: passenger data that describe the age and gender of vehicle passengers, weight data that estimate the weights of these passengers based on those demographics, and vehicle data that estimate the effect of the added weight on national fuel consumption. These data are published by several government agencies, and represent a consistent level of data granularity, aggregating over all people and travel, rather than simulating the travel of each individual person or vehicle. As these data are published by the US government, they are assumed to be reliable. As this study presents fuel consumption estimates that update the results published by Jacobson and McLay [11], this section focuses on the differences between these data, to facilitate comparison between the two studies.

Passenger Data

Passengers are classified according to two demographics: age and gender. According to their ages, passengers are classified as children (0–14 years), teenagers (15–19 years), or adults (20–74 years). Passenger genders can be either male or female. For driving passengers, the distributions over these two demographics are determined by the distribution of licensed drivers in the United States, weighted by the expected daily miles traveled by each gender. Drivers are assumed to be teenagers or adults, with children excluded, leaving four age and gender combinations to consider. The number of licensed drivers in each combination in 2005, as reported by the US Department of Transportation, Federal Highway Administration, is $LD_{M,15-19\text{yrs}} = 4.78 \times 10^6$, $LD_{M,20+\text{yrs}} = 9.55 \times 10^7$, $LD_{F,15-19\text{yrs}} = 4.56 \times 10^6$, and

$LD_{F,20+\text{yrs}} = 9.57 \times 10^7$, where $LD_{G,A}$ is the number of licensed drivers of gender G with age A [14]. When comparing these statistics with the 2003 statistics used by Jacobson and McLay [11], a small increase in the number of teenage drivers is observed (less than 0.5% for each gender), while a larger increase in the number of adult drivers is observed (more than 2% for each gender). As such, the average driver is more likely to be an adult (as opposed to a teenager) in 2005 than in 2003. The US Department of Transportation, Bureau of Transportation Statistics [15], has estimated that males travel 37.6 miles per day in automobiles, while females travel 21.2 miles per day. These data are the same as those used by Jacobson and McLay [11], as new data have not been released.

The ages and genders of nondriving passengers are more difficult to estimate, as there is no reliable model to estimate their distribution. Therefore, following from Jacobson and McLay [11], several different cases are considered for these distributions:

- *Case 1.* The distribution of a nondriving passenger’s age and gender is identical to the distribution for the driver.
- *Case 2.* The distribution of a nondriving passenger’s age and gender follows the distribution defined by the US population demographics, as estimated for 2006 by the United States Census Bureau [16].
- *Case 3.* All nondriving passengers are children or teenagers (age 0–19 years).

In the first case, passengers cannot be children (who tend to weigh less than teenagers or adults) and are more likely to be males (who tend to weigh more than females); therefore, this case establishes an upper bound on passenger weight. In contrast, the third case prevents passengers from being adults, who tend to weigh more than children or teenagers; this case establishes a lower bound on passenger weight. In the second case, all ages are represented, according to the age and gender distribution established by the most recent US census, as estimated for the year 2006 [16]. The marginal distributions for

both age and gender are estimated by the proportion of individuals in each gender or age classification. The estimated age distribution is $P(0-14 \text{ yrs}) = 0.203$, $P(15-19 \text{ yrs}) = 0.071$, and $P(20-74 \text{ yrs}) = 0.726$. The gender distribution is $P(M) = 0.493$ and $P(F) = 0.507$. Passenger weight computations for each case are discussed in the section titled "Passenger Weight Computations," where they are scaled by the number of passengers per vehicle to estimate the total passenger weight in a vehicle.

This model of passenger demographics does not account for any trends other than age or gender. As such, this model assumes that no relationships between vehicle use and other demographics exist. For example, it must be assumed that an individual's travel behavior is independent of their weight; that is, two passengers with the same gender and age classifications will exhibit identical travel patterns, regardless of their weight classification. Similarly, it is assumed that the average weight of an individual passenger is the same, on average, for both cars and light trucks. As there are no robust national statistics that contradict these assumptions, they are assumed to be reasonable. However, at least one regional survey [6] has suggested a positive correlation between automobile use and obesity; if such a correlation does exist, then the fuel consumption estimates presented in this paper will tend to underestimate the additional passenger weight present in vehicles, making these estimates conservative lower bounds.

Weight Data

Weight data describe the average weight of individuals in the US population, given their age and gender demographics; when used in conjunction with the passenger data described in section titled "Passenger Data," these data can be used to compute the average weight of a vehicle passenger. Weight data are based on the results of the National Health and Nutrition Examination Survey (NHANES) for the years 2005–2006, which is conducted by the National Center for Health Statistics [17]. Every two years, the NHANES distributes survey data collected from approximately 10,000 participants;

roughly half of these participants being examined each year after being chosen from one of 15 designated counties [18]. To allow national statistics to be generated from the survey results, the NHANES uses a multistage probability model to choose participants [19]. A sample weight measuring the number of people in the US population that are represented by each participant is assigned based on the sampling process and nonresponse rates [20]. These sample weights should not be confused with the physical body weight associated with each participant; in this paper, the term *sample weight* will be used when referring to the sample weights assigned during the NHANES. All height and weight measurements in the NHANES are taken directly, rather than self-reported, to ensure that their values are accurate. Individuals tend to underestimate their weight and overestimate their height during self-report [21,22], which leads to inaccurate BMI computations and, consequently, inaccurate assessment of overweight and obesity; the direct measurement of these quantities in the NHANES avoids this bias.

While a total of 10,348 individuals participated in NHANES 2005–2006, some of these individuals are excluded from the analysis conducted for this study. A participant was excluded if their BMI could not be accurately assessed. Some participants did not have their height or weight measured; for example, no height measurements were taken for children younger than the age of two. In other cases, the height or weight measurements do not accurately reflect a person's true height or weight; an individual is excluded from analysis if their height was not measured at full standing, or if their weight measurement included a medical appliance. Pregnant women were not excluded from analysis. After excluding these individuals, 8678 participants remained and were used to compute weight estimates.

Average weight estimates for individuals in each age and gender combination are shown in Table 1. These average weights differ from those used by Jacobson and McLay [11], which were based on NHANES 1999–2002. During the time between these two studies, it is estimated that the average

Table 1. Average Weight (lbs) by Age and Gender, Including Reduced Weight Classifications, 2005–2006

Weight Classification Reduced	Age (years)	Weight ^a (Male)	Weight ^a (Female)
None	2–14	76.4	75.9
	15–19	167	142
	20–74	197	167
Overweight, obese, and extremely obese	2–14	73.6	72.9
	15–19	159	137
	20–74	168	139
Overweight only	20–74	190	163
Obese only	20–74	180	152
Extremely obese only	20–74	191	158

^a 1 lb = 0.4536 kg.

weight of an adult male has increased by six pounds (191–197 lbs), the average weight of male teenagers and female adults have each increased by three pounds (164–167 lbs), and the average weight of a female teenager has increased by two pounds (140–142 lbs). Based on these estimates, national weights have continued to increase between the completion of NHANES 1999–2002 and the completion of NHANES 2005–2006. As noted earlier, some of the increase in average weight may not be due to additional overweight and obesity, as these classifications are affected by changes in height.

To assess the amount of fuel consumption that can be attributed to extra passenger weight, the amount of such weight must be quantified. Table 1 includes average weight estimates for each age and gender combination if weight due to overweight, obesity, and/or extreme obesity were eliminated. The methodology used to generate these estimates is discussed in the section titled “Estimating Weight due to Overweight and Obesity.” Given these estimates, one method for assessing extra passenger weight is to determine the average weight that can be attributed to overweight and obesity in the US population; this value can be computed from the contents of Table 1. Alternatively, extra passenger weight can be found by comparing the current weight estimates (i.e., those with no weight classification reduced)

in Table 1 with average weight estimates from previous years to quantify historical weight gain over specific time periods; this methodology was used by Jacobson and McLay [11].

Vehicle Data

Once the extra passenger weight has been computed, vehicle data are used to assess the effect that this change in weight exerts on fuel consumption. To carry out this analysis, vehicle data are gathered. These data measure the use and performance of vehicles in each fleet being considered (i.e., cars and light trucks), as well as the sensitivity of vehicle performance to changes in weight. In this vein, four parameters are collected to reflect the current use and performance of each fleet of vehicles: annual vehicle miles traveled, annual passenger miles traveled, annual fuel consumption, and number of registered vehicles. This study collects these parameters for travel in 2005 [2], while Jacobson and McLay [11] consider travel in 2003 (Table 2). Between 2003 and 2005, fuel consumption has risen at a faster rate than vehicle miles traveled for light trucks, reflecting a decline in average fuel economy. In contrast, vehicle miles traveled for cars have increased, while fuel consumption has declined, reflecting an improvement in fuel economy.

Adding weight to a vehicle tends to degrade the vehicle’s fuel economy. A fifth parameter, reflecting this relationship, is reported by the US Environmental Protection Agency [23] for cars in model year (MY) 2007. This parameter measures the linear change in fuel consumption (measured as the number of gallons required to travel one hundred vehicle miles) caused by a one pound increase in vehicle weight. Hybrid and diesel vehicles are excluded from these estimates. The estimated slopes are 8.78×10^{-4} gal/100 miles/lb for cars and 8.95×10^{-4} gal/100 miles/lb for light trucks (1 gal/100 miles/lb = 5.186 L/100 km/kg). In contrast, Jacobson and McLay [11] use data based on the model year 2005, where these slopes are 7.27×10^{-4} gal/100 miles/lb for cars and 1.16×10^{-3} gal/100 miles/lb for light trucks. Between 2005 and 2007, fuel consumption in cars has become more sensitive

Table 2. Vehicle Use Statistics, Given Vehicle Type, 2003–2005

Quantity	Units	Cars		Trucks	
		2003	2005	2003	2005
Vehicle miles	miles ^a	1661 B	1690 B	998 B	1060 B
Passenger miles	miles ^a	2624 B	2670 B	1730 B	1837 B
Fuel consumption	gallons ^b	74.6 B	73.9 B	56.3 B	65.4 B
Number registered vehicles	—	136 M	137 M	87.0 M	95.3 M

^a 1 mile = 1.609 km.^b 1 gallon = 3.785 L.

to changes in weight, while fuel consumption in light trucks has become less sensitive.

ESTIMATING WEIGHT DUE TO OVERWEIGHT AND OBESITY

The first component of the fuel consumption model describes the estimation of the average extra weight in the US population. While this extra weight is evaluated as historical weight gain, for example, by comparing current national weight statistics to those from the 1960s [12] by Jacobson and McLay [11], it can also be quantified by considering the average weight loss required to eliminate overweight and obesity in the current US population. As discussed earlier, all historical weight gain is not necessarily unhealthy. By quantifying the passenger weight due to overweight and obesity, the resulting fuel consumption estimates are solely attributable to weight that is considered unhealthy. One method for computing such weight statistics, first introduced by Jacobson and King [13], is presented in this section. Rather than using the NHANES data directly, these data are altered such that each participant classified as overweight or obese is reduced to their *maximum normal weight*, or the weight at which the individual achieves the highest possible BMI to be classified as normal weight. Average weight statistics can be computed from these altered data and compared to weight statistics generated by the unaltered NHANES data to determine the average weight attributable to overweight and obesity in the United States, which can take the role of average extra weight in the model described by Jacobson and McLay [11].

Weight Classifications

To compute a person's maximum normal weight, a system of weight classifications must be established. These classifications are based on BMI, and differ by age. An adult can be considered underweight ($\text{BMI} < 18.5$), normal weight ($18.5 \leq \text{BMI} < 25$), overweight ($25 \leq \text{BMI} < 30$), obese ($30 \leq \text{BMI} < 40$), or extremely obese ($\text{BMI} \geq 40$); therefore, the maximum normal weight for an adult occurs when BMI is equal to 25 [3]. For children below 20 years of age, overweight status is determined by the BMI-for-age growth charts for boys and girls, as established by the Centers for Disease Control and Prevention [24]. A child is classified as overweight if their BMI is greater than the 95th percentile of BMI for their age. This cutoff is at its minimum at age four—when it is 17.8 for males and 18.0 for females—and at its maximum at age 19—when it is 30.6 for males and 31.8 for females.

Computing Weight Estimates

In order to eliminate overweight and obesity in the US population, all individuals must be reduced to their maximum normal weight. This will occur when each individual experiences a weight loss of

$$\text{WL} = \max\{W - \text{BMI}_N H^2, 0\}, \quad (1)$$

where BMI_N is the maximum BMI in the normal range (25 for adults, varies according to age for teenagers and children), and W and H are the individual's weight and height, respectively. No weight is lost if the individual is already of normal weight (or underweight); their weight loss is equal to zero.

The sample weights associated with the NHANES can be used to compute average weight statistics after these weight losses are applied. To compute the statistics for a particular age and gender classification, let J be the set of all survey participants that meet the age and gender requirements. For any particular individual, j , their sample weight, sw_j , determines the number of people in the US population that are represented by person j . Therefore, the average weight for individuals with this age and gender classification is computed as

$$W_A' = \frac{\sum_{j \in J} sw_j (W_j - WL_j)}{\sum_{j \in J} sw_j}, \quad (2)$$

where W_j is the weight of individual j , and WL_j is the weight loss for individual j , as computed in Equation (1).

Average weight estimates generated by this methodology are reported in Table 1. Four alternative cases are presented, each applying the weight loss computed in Equation (1) to a different set of surveyed individuals. In the first case, weight loss is applied to individuals classified as overweight, obese, or extremely obese. The remaining three cases apply weight loss to only one of the three classifications (e.g., weight loss is only applied to individuals classified as overweight). Since children and teenagers are not classified as obese or extremely obese, their average weights are omitted when weight loss is only applied to individuals with those classifications.

Table 1 shows that, for example, the average weight for an adult male falls from 197 to 168 lbs if weight loss is applied to overweight, obese, and extremely obese individuals, while it falls from 197 to 180 lbs if weight loss is only applied to obese individuals. Of the weight loss that results from eliminating overweight, obesity, and extreme obesity, some can be attributed to the weight loss in each classification.

In adult (age 20–74 years) females, prevalence of overweight is 25%, prevalence of obesity is 29%, and prevalence of extreme obesity is 7%. From Table 1, 13% of total

weight loss is due to individuals classified as overweight, 54% is due to individuals classified as obese, and 33% is due to individuals classified as extremely obese. Though a small fraction of adult females are classified as extremely obese, approximately one-third of the total weight loss is due to extreme obesity, reflecting the larger weight loss experienced by individuals classified as extremely obese. However, the majority of total weight loss is due to individuals who are classified as obese, reflecting the larger weight loss and higher prevalence of individuals classified as obese.

The prevalence of overweight in adult males is approximately 40%, while obesity has 30% prevalence and extreme obesity has 4% prevalence. However, from Table 1, 23% of total weight loss is due to individuals who are classified as overweight, 56% is due to individuals classified as obese, and 21% is due to individuals classified as extremely obese. Though obesity is less prevalent than overweight, an obese individual tends to experience more weight loss than an overweight individual, so more of the total weight loss is due to the reduction of obesity. Extreme obesity is much less prevalent, so less of the total weight loss is due to individuals who are extremely obese.

ESTIMATING FUEL CONSUMPTION DUE TO OVERWEIGHT AND OBESITY

Once average extra weight in the US population has been estimated, the impact of this weight change on fuel consumption can be assessed. First, the average extra weight of driving and nondriving vehicle passengers must be computed by determining the age and gender distributions of these passengers, in conjunction with the average extra weight of each age and gender combination. Second, the average extra passenger weight per vehicle is determined by scaling the average extra weight of each passenger type by the average number of passengers in each vehicle. These two tasks represent the second and third components of the fuel consumption model (discussed in the section titled “Passenger Weight Computations”). Finally, the fourth component of the fuel

consumption model assesses the effect of the average extra weight on fuel economy, which is then translated into a change in fuel consumption (discussed in the section titled “Fuel Economy Computations”).

Passenger Weight Computations

The age and gender demographics of each passenger are used to determine the weight of that passenger. Passengers can be either driving or nondriving passengers, whose weight and demographic data are given a subscript of D and ND, respectively. Clearly, the same demographic distributions cannot be used for both driving and nondriving passengers. For example, nondriving passengers can be children, while driving passengers can only be teenagers or adults. Therefore, the total passenger weight in a vehicle from fleet V is computed as

$$E[W_V] = E[W_D] + (N_V - 1)E[W_{ND}], \quad (3)$$

where $E[W_D]$ and $E[W_{ND}]$ are the average weights of driving and nondriving passengers, and N_V is the average number of passengers in vehicles from fleet V. An estimate for N_V is computed as passenger miles traveled by vehicles in the fleet, divided by vehicle miles traveled by those vehicles. Passenger weights are computed by the conditional expectations

$$E[W_D] = \sum_{A_D, G_D} E[W_D | A_D \cap G_D] \times P(A_D | G_D) P(G_D), \quad (4)$$

$$E[W_{ND}] = \sum_{A_{ND}, G_{ND}} E[W_{ND} | A_{ND} \cap G_{ND}] \times P(A_{ND} | G_{ND}) P(G_{ND}). \quad (5)$$

For both driving and nondriving passengers, average passenger weight given age and gender demographics is assumed to be equal to the average weight for a person in those same demographics in the entire US population, as reported in Table 1. Therefore,

$$\begin{aligned} E[W_D | A_D \cap G_D] &= E[W_{ND} | A_{ND} \cap G_{ND}] \\ &= E[W | A \cap G]. \end{aligned} \quad (6)$$

Given this assumption, the only difference between driving and nondriving passengers lies in their age and gender distributions. The data used to estimate these distributions were reported in the section titled “Passenger Data.”

For driving passengers, age and gender are determined by the distribution of licensed drivers in the United States, as well as the daily miles traveled by each gender. Children cannot be drivers; all drivers must be teenagers or adults. Given the driver’s gender (G_D), their age (A_D) follows a distribution that is defined by the proportion of licensed drivers in each age classification

$$P(A_D | G_D) = LD_{G_D, A_D} / \sum_A LD_{G_D, A}, \quad (7)$$

where $LD_{G,A}$ is the number of licensed drivers with gender G and age A . The gender distribution for drivers is defined by the proportion of miles driven by each gender

$$P(G_D) = \frac{E[DM|G_D] \sum_A LD_{G_D, A}}{\sum_G E[DM|G] \sum_A LD_{G, A}}, \quad (8)$$

where $E[DM|G]$ is the daily miles driven by individuals with gender G .

The distributions of age and gender for nondriving passengers are defined differently for each passenger case, as described in the section titled “Passenger Data.” For Case 1, these distributions are the same as those for driving passengers, so $P(A_{ND}|G_{ND}) = P(A_D|G_D)$ and $P(G_{ND}) = P(G_D)$. For Case 2, these distributions are defined by estimates from the US Census for the year 2006 [16]. In this case, age and gender are assumed to be independent variables, such that $P(A_{ND}|G_{ND}) = P(A_{ND})$. These probabilities are defined by the proportion of individuals in each age and gender classification in the census data, as described in the section titled “Passenger Data.”

For Case 3, passengers must be children and teenagers. Census data and daily miles

traveled by each gender define the age and gender distributions. Let $A' = \{0-14 \text{ years}, 15-19 \text{ years}\}$ be the set of age classifications for children and teenagers. The gender, G' , of a nondriving passenger for Case 3 is given by the weighted average

$$P(G') = \frac{E[DM | G'] \sum_{A \in A'} P(G', A)}{\sum_G \left(E[DM | G] \sum_{A \in A'} P(G, A) \right)}, \quad (9)$$

where $P(G, A) = P(G)P(A)$, and the proportion of individuals of each gender and age classification are computed using US Census estimates for 2006 [16]. The average weight of a nondriving passenger of gender G' is given by

$$\begin{aligned} & E[W_{ND} | G'] \\ &= \frac{\sum_{A \in A'} E[W_{ND} | A_{ND} \cap G_{ND}] P(G', A)}{\sum_{A \in A'} P(G', A)}. \end{aligned} \quad (10)$$

Therefore, the average weight of a non-driving passenger in Case 3 is computed as

$$E[W_{ND}] = E[W_{ND} | M] P(M) + E[W_{ND} | F] P(F), \quad (11)$$

rather than using Equation (5), where M and F reflect the case when the gender is male and female, respectively.

Fuel Economy Computations

In general, the fuel economy of a vehicle improves when the vehicle carries less weight. For a fleet of vehicles, V , let the average extra passenger weight per vehicle be given by ΔW_V . This extra weight is computed as historical weight gain or average weight attributable to overweight and obesity in the US population. Fuel economy is defined as the number of miles that a vehicle travels when consuming one gallon of gasoline; alternatively, it is the inverse of the number of gallons required to travel one

mile. Using this alternative formulation, the inverse of fuel economy falls by $\Delta W_V R_V / 100$ when passenger weight is decreased by ΔW_V (i.e., the average extra passenger weight per vehicle is removed), since $R_V / 100$ measures the number of additional gallons required to travel one vehicle mile caused by a one pound increase in vehicle weight. Therefore, if FE_V is the current fuel economy for vehicles in fleet V , this decrease in passenger weight causes fuel economy to increase to

$$FE_V' = [1/FE_V - \Delta W_V R_V / 100]^{-1}, \quad (12)$$

and fuel consumption decreases to

$$FC_V' = VM_V / FE_V', \quad (13)$$

where VM_V measures the vehicle miles traveled by fleet V . The fuel consumption estimates computed with this model can be compared with the fuel consumption for 2005, as reported in Table 2.

ANALYSIS

This paper estimates the annual additional fuel consumption for cars and light trucks in the United States due to the average extra passenger weight carried by cars and light trucks. This extra weight is estimated in two ways, both comparing current weights statistics computed from NHANES 2005–2006 with alternative weight cases. First, average extra weight is estimated by historical weight gain in the US population over several past time periods. Time periods considered in this study are 1960–1962, 1971–1974, 1976–1980, 1988–1994, and 1999–2002. Second, average extra weight is computed by assessing the weight loss that would be required to eliminate overweight, obesity, extreme obesity, or all three, from the US population. These weight estimates are computed using the methodology of section titled “Estimating Weight due to Overweight and Obesity.” Other than passenger weight, all data are held constant in the analysis.

Using the vehicle data described in the section titled “Vehicle Data,” several statistics can be computed. The average number of

passengers in a vehicle in fleet V is estimated as the passenger miles traveled divided by vehicle miles traveled. On average, each car holds 1.58 passengers, while each light truck carries 1.73 passengers. These values are indistinguishable from the values computed by Jacobson and McLay [11], which suggests that ridesharing patterns, on average, have not changed substantially between 2003 and 2005. Average fuel economy for vehicles in fleet V is estimated as vehicle miles divided by fuel consumed. Average fuel economy for cars is estimated as 22.9 miles per gallon (mpg), reflecting an improvement of 0.6 mpg from the estimate made by Jacobson and McLay [11]. Average fuel economy for light trucks is estimated as 16.2 mpg, which is 1.5 mpg less than the estimate made by Jacobson and McLay [11].

Additional Fuel Consumption Due to Historical Weight Gain

When average extra weight is estimated by historical weight gain, the estimated annual additional fuel consumption, as first reported by Jacobson and King [13], is reported in Table 3. To aid comparison, the results

reported by Jacobson and McLay [11] are also given. When comparisons are made in a single column, only changes in weight data between time periods are reflected. For example, the first column reports annual fuel consumption when weight estimates from 1999–2002 are compared to those in previous time periods (e.g., 1976–1980) using vehicle and passenger data from 2003. Since the two columns use vehicle and passenger data from different years, comparisons made between the two columns also reflect these differences. For example, consider the 1988–1994 row, for passenger Case 1. The first column reports that 357 million gallons of fuel are consumed each year according to historical weight gain between 1988–1994 and 1999–2002 in conjunction with passenger and vehicle data from 2003. The second column reports that 545 million gallons of fuel are consumed each year due to historical weight gain between 1988–1994 and 2005–2006 in conjunction with passenger vehicle data from 2005. Therefore, the 52.8% change observed between these two estimates incorporates not only historical weight gain between 1999–2002 and 2005–2006, but also changes in passenger and vehicle

Table 3. Additional Gallons of Fuel Consumed Annually Due to Historical Weight Gain in the US Population Over Several Time Periods

Time Period	Case	Additional Fuel (gallons ^a), When Compared to		% Change
		1999–2002 ^b	2005–2006 ^c	
1999–2002	1	*	182M	*
	2	*	182M	*
	3	*	199M	*
1988–1994	1	357M	545M	52.8
	2	335M	523M	56.3
	3	272M	473M	74.2
1976–1980	1	692M	886M	28.0
	2	655M	850M	29.7
	3	558M	758M	35.9
1971–1974	1	715M	909M	27.2
	2	678M	873M	28.8
	3	569M	770M	35.2
1960–1962	1	938M	1137M	21.2

^a 1 gallon = 3.785L.

^b Historical weight gain found by comparing weight statistics in 1999–2002 with those in the specified time period, using passenger and vehicle data from 2003 [11].

^c Historical weight gain found by comparing weight statistics in 2005–2006 with those in the specified time period, using passenger and vehicle data from 2005 [13].

Table 4. Estimated Annual Fuel Savings (Gallons) by Reducing Weight of Overweight, Obese, or Extremely Obese Individuals to Maximum Normal Weight BMI

Weight Classification Reduced	Additional Fuel (Gallons ^a)		
	Case 1	Case 2	Case 3
Overweight only	223M	209M	196M
Obese only	608M	549M	371M
Extremely obese only	273M	252M	166M
Overweight, obese, and extremely obese	1104M	1011M	734M

^a1 gallon = 3.785 L; sums may differ due to independent rounding.

data between 2003 and 2005. The largest annual additional fuel consumption is 1.137 billion gallons, according to historical weight gain since the 1960s when all passengers are adults and teenagers. This amount of fuel accounts for 0.8% of the gasoline consumed by cars and light trucks in 2005 [2].

According to this model, between 182 and 199 million additional gallons of fuel can be attributed to historical weight gain since 2002; by subtracting from 1.137 billion gallons, this computation indicates that 938–955 million additional gallons can be attributed to historical weight gain between 1960 and 2002, which is comparable to the 938 million gallon estimate that was reported by Jacobson and McLay [11]. If each passenger were to gain one pound, this model estimates that an additional 39.8 million gallons of fuel would be consumed each year, a small increase from the estimate of 39.2 million gallons reported by Jacobson and McLay [11].

Each gallon of gasoline consumption causes 19.4 pounds (8.80 kg) of carbon dioxide (CO₂) emissions [25]. Therefore, the 1.137 billion gallons of fuel consumed due to historical weight gain since the 1960s leads to 22.1 billion pounds (10.0 million metric tons) of CO₂ emissions, accounting for 0.5% of the total CO₂ emissions produced by combusting fossil fuels for use in the transportation sector in 2005 [26].

Projected Fuel Savings by Reducing Overweight and Obesity

When average extra weight is estimated by the average weight attributable to

overweight and obesity, the estimated annual additional fuel consumption is reported in Table 4. If overweight, obesity, and extreme obesity were eliminated from the US population, between 734 million and 1.104 billion gallons of fuel would be saved each year. This savings can be stratified by weight classification; regardless of which passenger case is used, more than half of the total savings is due to individuals classified as obese. This result is consistent with the earlier observation that more than half of the weight lost by eliminating overweight, obesity, and extreme obesity in adults would be due to the elimination of obesity. Furthermore, all passengers in Case 3 must be children and teenagers, for whom no classification of obesity or extreme obesity exists. Therefore, all fuel savings due to elimination of obesity and extreme obesity must be the average extra weight of the driver, which accounts for more than 73% of the total fuel saved.

Using these estimates of potential fuel savings, the 1.104 billion gallons of fuel that could be saved by eliminating overweight and obesity would also eliminate 21.4 billion pounds (9.71 million metric tons) of CO₂ emissions, or 0.5% of the total CO₂ emissions produced by combusting fossil fuels for use in the transportation sector in 2005 [26].

CONCLUSIONS

This paper quantifies the additional amount of fuel consumed each year due to average extra passenger weight in noncommercial passenger highway vehicles (i.e., cars and light trucks). Two scenarios are considered: in the first scenario, average extra passenger weight is based on historical weight gain in the US population over several past time periods, and in the second scenario, average extra passenger weight is computed as the average weight attributable to overweight and obesity in the US population. Regardless of which scenario is chosen, up to one billion gallons of fuel or more are found to be attributable to extra passenger weight. In the first scenario, if average extra weight is based on historical weight gain since the 1960s, estimates reach as high as 1.137 billion gallons,

while the second scenario generates estimates of up to 1.104 billion gallons. Both estimates represent approximately 0.8% of the fuel consumed by cars and light trucks each year, and consuming this additional fuel produces up to 0.5% of the CO₂ emissions generated by combusting fossil fuels in the transportation sector in 2005. Though these estimates represent a very small fraction of the total fuel consumption and CO₂ emissions produced annually in the United States, the estimates are large in absolute terms.

These estimates may change over time, particularly as travelers in the United States react to record gasoline prices. It has been estimated that a US\$1 increase in real gasoline prices will reduce obesity in the United States by 15% over five years by discouraging automobile use and decreasing the number of meals eaten at restaurants [7]. Such a result would not only decrease the vehicle miles traveled by cars and light trucks, but would also decrease the weight of passengers carried by those vehicles, both of which would reduce fuel consumption.

Fuel savings can also be realized in other ways. For example, the amount of fuel that could be saved by maintaining proper tire inflation was estimated to be 1.388 billion gallons each year, slightly more than the fuel consumption attributable to overweight and obesity [27]. Fuel can also be saved by ridesharing (i.e., carpooling). If one additional passenger were added to every hundred vehicles (cars and light trucks) without needing to be picked up, the annual fuel savings could be as high as 0.82 billion gallons [28]. Individually, these initiatives make up a small portion of the gasoline consumed in the United States, but each highlights an opportunity to reduce the high level of national oil consumption.

The results in this paper quantify the impact of obesity on fuel consumption. However, the reverse relationship may also be true. In particular, as fuel consumption has increased, so have obesity rates. This begs the question: If fuel consumption levels drop, will obesity rates follow? Lopez-Zetina *et al.* [6] report a positive relationship between time spent in a car and obesity rates in

California. It has been suggested that the recent surge in overweight and obesity rates in China correlates with an increased level of automobile ownership [29], indicating that the issues being faced by the United States will become international issues as nations become more affluent. From a practical point of view, every trip that replaces automobile use with walking, riding a bicycle, or using public transit requires a greater physical exertion of energy, and hence will lead to a weight reduction.

From a broad perspective, the model presented in this paper shows how mathematical modeling and operations research can quantify the direct relationship that exists between two seemingly unrelated issues. As this model is driven by national statistics, the use of data that control for the age and gender of driving and nondriving passengers provides a more accurate estimate of average extra passenger weight, and hence the additional fuel consumption due to this weight.

These national estimates are hindered by the dearth of data that describe how driving habits depend on factors such as income, race, and weight. Therefore, the model does not control for these factors and assumes that a passenger's income, race, and weight do not shape their driving habits. While these factors undoubtedly exert some level of influence, the degree of this influence is not clear. If estimates of how these demographics affect travel behavior could be quantified, they could be used to improve the estimation of passenger weight and further increase the accuracy of the fuel consumption estimates generated by this model. If a strong connection exists between these demographics and driving habits, then the fuel consumption estimates presented here may differ substantially from those generated by a more robust model that included these factors. If data that could support such a model become available in the future, the results presented in this paper should be reevaluated. In addition, the estimates presented in this paper estimate the fuel consumption due to overweight and obesity in the current national vehicle fleet. As the nature of the vehicle fleet changes (e.g., changes in the number of

larger vehicles such as SUVs, or the number of hybrid vehicles), the fleet's reaction to changes in passenger weight due to overweight and obesity will fluctuate. Therefore, the fuel consumption estimates presented here can be expected to change over time.

Quantifying the relationship between the socioeconomic issues of fuel consumption and obesity is valuable to both energy and public health policymakers, as it allows them to consider an additional benefit of public health initiatives that aim to reduce obesity in the US population. While the main economic impact of such initiatives would be a reduction in annual health-care expenditures associated with overweight and obesity, which were estimated to have been \$78.5 billion in 1998, accounting for 9.1% of all medical expenditures in the United States [30], the additional benefit of savings up to a billion gallons of fuel each year provides another significant and measurable benefit.

Acknowledgments

The material in this article is based upon work supported in part by the National Science Foundation under Grant No. 0 457 176. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The computational work was conducted with support from the Simulation and Optimization Laboratory at the University of Illinois.

REFERENCES

1. US Energy Information Administration. US regular weekly retail [report on the Internet]; Washington (DC). Available at http://www.eia.doe.gov/oil_gas/petroleum/data_publications/wrgp/mogas_history.html. Accessed 2008 Jul 2.
2. US Department of Transportation, Bureau of Transportation Statistics. National transportation statistics [report on the Internet]; Washington DC. Available at http://www.bts.gov/publications/national_transportation_statistics/. Accessed 2007 Sep 5.
3. Ogden CL, Carroll MD, Curtin LR. *et al.* Prevalence of overweight and obesity in the United States, 1999–2004. *J Am Med Assoc* 2006;295(13):1549–1555.
4. Mokdad AH, Serdula MK, Dietz WH. *et al.* The spread of the obesity epidemic in the United States, 1991–1998. *J Am Med Assoc* 1999;282(16):1519–1522.
5. Must A, Spadano J, Coakley EH, Field AE, *et al.* The disease burden associated with overweight and obesity. *J Am Med Assoc* 1999;282(16):1523–1529.
6. Lopez-Zetina J, Lee H, Friis R. The link between obesity and the built environment. Evidence from an ecological analysis of obesity and vehicles miles of travel in California. *Health Place* 2006;12:656–664.
7. Courtemanche C. A silver lining? The connection between gas prices and obesity [monograph on the Internet]. St. Louis (MO): Department of Economics, Washington University in St. Louis. Available at http://artsci.wustl.edu/~cjcourte/gas_obesity.pdf. Accessed 2007 Oct 22.
8. Frank LD, Sallis JF, Conway TL, Chapman JE, Saelens BE, *et al.* Many pathways from land use to health. *J Am Plann Assoc* 2006;72(1):75–87.
9. Lopez RP, Hynes HP. Obesity, physical activity, and the urban environment: public health research needs. *Environ Health* [serial on the Internet] 2006;5(25). Available at <http://www.ehjournal.net/content/5/1/25>. Accessed 2007 Oct 22.
10. Dannenberg AL, Burton DC, Jackson RJ. Economic and environmental costs of obesity: the impact on Airlines. *Am J Prev Med* 2004;27(3):264.
11. Jacobson SH, McLay LA. The economic impact of obesity on automobile fuel consumption. *Eng Econ* 2006;51:307–323.
12. Ogden CL, Fryar CD, Carroll MD, *et al.* Mean body weight, height, and body mass index, United States, 1960–2002, Advance data from vital health and statistics, No. 347. Hyattsville (MD): National Center for Health Statistics; 2004.
13. Jacobson SH, King DM. Measuring the potential for automobile fuel savings in the united states: the impact of obesity. *Transp Res D Transp Environ* 2009;14:6–13.
14. US Department of Transportation, Federal Highway Administration. Distribution of licensed drivers—2005 [report on the Internet]; Washington (DC). Available

- at <http://www.fhwa.dot.gov/policy/ohim/hs05/pdf/dl20.pdf>. Accessed 2007 Oct 22.
15. US Department of Transportation, Bureau of Transportation Statistics. National household travel survey 2001 highlights report [report on the Internet]; Washington (DC). Available at http://www.bts.gov/publications/highlights_of_the_2001_national_household_travel_survey/pdf/entire.pdf. Accessed 2007 Oct 22.
 16. US Census Bureau. Annual estimates of the population by five-year age groups and sex for the United States: April 1, 2000 to July 1, 2006. NC-EST2006-01 [report on the Internet]; Washington (DC). Available at <http://www.census.gov/popest/national/asrh/NC-EST2006/NC-EST2006-01.xls>. Accessed 2007 Oct 14.
 17. US Centers for Disease Control and Prevention, National Center for Health Statistics. National health and nutrition examination survey, NHANES 2005–2006 overview [report on the Internet]; Hyattsville (MD). Available at http://www.cdc.gov/nchs/about/major/nhanes/nhanes2005-2006/nhanes05_06.htm. Accessed 2007 Nov 5.
 18. US Centers for Disease Control and Prevention, National Center for Health Statistics. National health and nutrition examination survey, 2005–2006 overview [report on the Internet]; Hyattsville (MD). Available at http://www.cdc.gov/nchs/data/nhanes/nhanes_05_06/overviewbrochure_0506.pdf. Accessed 2007 Oct 6.
 19. US Centers for Disease Control and Prevention, National Center for Health Statistics. Analytic and reporting guidelines, the national health and nutrition examination survey (NHANES) [report on the Internet]; Hyattsville (MD). Available at http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf. Accessed 2007 Oct 6.
 20. US Centers for Disease Control and Prevention, National Center for Health Statistics. Continuous NHANES web tutorial, specifying weighting parameters [report on the Internet]; Hyattsville (MD). Available at <http://www.cdc.gov/nchs/tutorials/Nhanes/SurveyDesign/Weighting/intro.htm>. Accessed 2007 Oct 6.
 21. Gorber SC, Tremblay M, Moher D, *et al.* A comparison of direct vs. self-report measures for height, weight, and body mass index: a systematic review. *Obes Rev* 2007;8:307–326.
 22. Nawaz H, Chan W, Abdulrahman M, *et al.* Self-reported weight and height, implications for obesity research. *Am J Prev Med* 2002;20(4):294–298.
 23. US Environmental Protection Agency. Light-duty automotive technology and fuel economy trends: 1995–2007 EPA420-R-07-008 [report on the Internet]; Washington (DC). Available at <http://www.epa.gov/otaq/cert/mpg/fetrends/420r07008.pdf>. Accessed 2007 Oct 6.
 24. US Centers for Disease Control and Prevention, National Center for Health Statistics. United States clinical growth charts [report on the Internet]; Hyattsville (MD). Available at http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/clinical_charts.htm. Accessed 2007 Oct 5.
 25. US Environmental Protection Agency. Emission facts, average carbon dioxide emissions resulting from gasoline and diesel fuel EPA420-F-05-001 [report on the Internet]; Washington (DC). Available at <http://www.epa.gov/otaq/climate/420f05001.pdf>. Accessed 2007 Oct 18.
 26. US Environmental Protection Agency. Inventory of US Greenhouse gas emissions and sinks: 1990–2005, executive summary EPA430-R-07-002 [report on the Internet]; Washington (DC). Available at <http://www.epa.gov/climatechange/emissions/downloads/06/07ES.pdf>. Accessed 2007 Oct 18.
 27. Pearce JM, Hanlon JT. Energy conservation from systematic tire pressure regulation. *Energy Policy* 2007;35:2673–2677.
 28. Jacobson SH, King DM. Fuel saving and ridesharing in the US: motivations, limitations, and opportunities. *Transp Res D Transp Environ* 2009;14:14–21.
 29. Wu Y. Overweight and obesity in China. *Br Med J* 2006;333:362–363.
 30. Finkelstein EA, Fiebelkorn IC, Wang G. National medical spending attributable to overweight and obesity: how much, and who's paying? *Health Aff* 2003;W3:219–226.

A STRUCTURAL CLUSTERING ALGORITHM FOR LARGE NETWORKS

XIAOWEI XU

Department of Information
Science, University of Arkansas
at Little Rock, Little Rock,
Arkansas

INTRODUCTION

Networks are ubiquitous. Common networks include social networks, the World Wide Web, and computer networks. A network consists of a set of vertices interconnected by edges. A vertex represents some real entities such as a person, website, or piece of networking hardware. An edge connects two vertices if they have some relationship such as a friendship, hypertext link, or wired connection. In such networks, each vertex plays a role. Some vertices are members of clusters; a group of peers in a social network or a group of related websites in the WWW are examples. Some vertices are hubs that bridge many clusters but do not belong strongly to any one cluster; for example, politicians tend to play such a role in social networks, and websites like wikipedia.org are clearing-houses for all kinds of information. Some vertices represent outsiders that have only weak associations with any cluster; for example, a loner in a social network, or a parked domain on the internet.

To illustrate these points further, consider the network in Fig. 1. If one might confidently consider the vertices $\{0, 1, 2, 3, 4, 5\}$ and $\{7, 8, 9, 10, 11, 12\}$ to be clusters of peers, vertex 6 is difficult to classify. It could arguably belong to either cluster or to none. It is an example of a hub. Likewise, vertex 13 is weakly connected to a cluster. It is an example of an outsider.

Network clustering (or graph partitioning) is the detection of structures like those in Fig. 1, and it is drawing increased attention and application in computer science [1,2], physics [3], and bioinformatics [4]. Various such methods have been developed. They

tend to partition based on the principle that clusters should be sparsely connected, but the vertices within each cluster should be densely connected. Modularity-based algorithms [3–5] and normalized cut [1,2] are successful examples. However, they do not distinguish the roles of vertices.

The modularity-based algorithm [5] will cluster the network in Fig. 1 into two clusters: one consisting of vertices 0 to 6 and the other consisting of vertices 7 to 13. It does not isolate vertex 6 or vertex 13.

The identification of hubs gives valuable information. For example, hubs in the WWW are deemed authoritative information sources among web pages [6], and hubs in social networks are believed to play a crucial role in viral marketing [7] and epidemiology [8].

In this article, we propose a new method for network clustering. The goal of our method is to find clusters, hubs, and outsiders in large networks. To achieve this goal, we use the neighborhood of the vertices as clustering criteria instead of only their direct connections. Vertices are clustered by how they share neighbors. Doing so makes sense when you consider the detection of communities in large social networks. Two people who share many friends should be clustered in the same community.

Referring to Fig. 1, consider vertices 0 and 5, which are connected by an edge. Their neighborhoods are the vertex sets $\{0, 1, 4, 5, 6\}$ and $\{0, 1, 2, 3, 4, 5\}$, respectively. They share many neighbors and are thus reasonably clustered together. In contrast, consider the neighborhoods of vertex 13 and vertex 9. These two vertices are connected, but share only few common neighbors, that is, $\{9, 13\}$. Therefore, it is doubtful whether they should be grouped together. The situation for vertex 6 is a little different. It has many neighbors, but they are sparsely interconnected.

Our algorithm identifies two clusters, $\{0, 1, 2, 3, 4, 5\}$ and $\{7, 8, 9, 10, 11, 12\}$, and isolates vertex 13 as an outsider and vertex 6 as a hub.

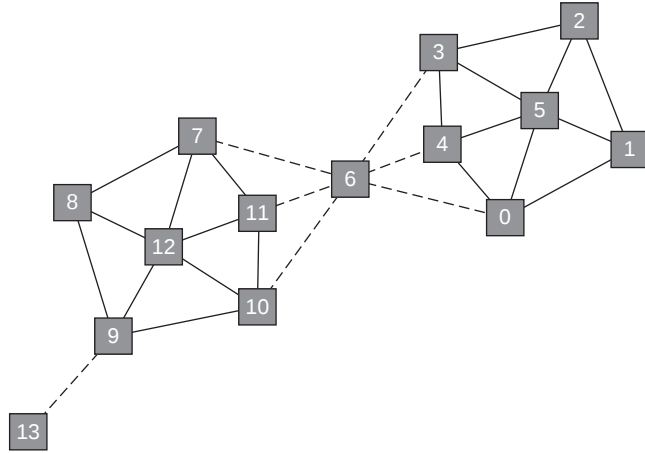


Figure 1. A network with two clusters, a hub, and an outsider.

Our algorithm has the following features:

- It detects clusters, hubs, and outsiders by using the structure and the connectivity of the vertices as clustering criteria.
- It is efficient. It clusters the given network by visiting each vertex exactly once.

Through theoretical analysis and experimental evaluation we demonstrate that our algorithm finds meaningful clusters and identifies hubs and outsiders in very large networks.

With respect to efficiency, our algorithm’s running time on a network with n vertices and m edges is $O(m)$. In contrast, the running time of the fast modularity-based algorithm [5] is $O(md \log n)$.

The article is organized as follows. We formalize the notion of structure-connected cluster (SCC) in the section titled “The Notion of Structure-Connected Cluster.” We devise an algorithm to find SCC in the section titled “Algorithm.” We give a computation complexity analysis of our algorithm in the section titled “Complexity Analysis.” We compare our algorithm to the fast modularity-based algorithm in the section titled “Evaluation.” We review the related work in the section titled “Related Work.” Finally, we present our conclusions and suggest future work in the final section.

THE NOTION OF STRUCTURE-CONNECTED CLUSTER

Our goal is both to cluster network optimally and to identify and isolate hubs and outsiders. Therefore, both connectivity and local structure are used in our definition of optimal clustering. In this section, we formalize the notion of a structure-connected cluster that extends that of a density-based cluster [9] and can distinguish clusters, hubs, and outsiders in networks.

The existing network clustering methods such as modularity-based algorithms are designed to find optimal clusters based on the number of edges between vertices or between clusters. Direct connections are important, but they represent only one aspect of the network structure. The neighborhood around two connected vertices is also important. The neighborhood of a vertex includes all the vertices connected to it by an edge. When you consider a pair of connected vertices, their combined neighborhood reveals neighbors common to both.

Our method is based on common neighbors. Two vertices are assigned to a cluster according to how they share neighbors. This makes sense when you consider social communities. People who share many friends create a community, and the more friends they have in common, the more intimate the community. But in social networks there are different kinds

of actors besides peers. There are also people who are outsiders (like hermits), and there are people who are friendly with many communities but belong to none (like politicians). The latter play a special role in small-world networks known as *hubs* [10]. An outsider is illustrated by vertex 13 in Fig. 1 and a hub is illustrated by vertex 6.

In this article, we focus on simple, undirected, and unweighted graph. Let $G = \{V, E\}$ be a graph representing a real network, where V is a set of vertices and E is a set of pairs (unordered) of distinct vertices, called *edges*.

The structure of a vertex can be described by its neighborhood. A formal definition of vertex structure is given as follows.

Definition 1 [Vertex Structure]. Let $v \in V$. The structure of v is defined by its neighborhood, denoted by $\Gamma(v)$:

$$\Gamma(v) = \{w \in V | (v, w) \in E\} \cup \{v\}$$

In Fig. 1, vertex 6 is a hub sharing neighbors with two clusters. If we only use the number of shared neighbors, vertex 6 will be clustered into either of the clusters or cause the two clusters to merge. Therefore, we normalize the number of common neighbors in different ways, which give us different similarity measures.

Definition 2 [Structural Similarity].

$$\sigma_{\cos}(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)| |\Gamma(w)|}} \quad (1)$$

$$\sigma_{\text{jaccard}}(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{|\Gamma(v) \cup \Gamma(w)|} \quad (2)$$

$$\sigma_{\min}(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\min(|\Gamma(v)|, |\Gamma(w)|)}. \quad (3)$$

The first similarity, called *cosine similarity*, normalizes the number of common neighbors by the geometric mean of the two neighborhoods' size and is commonly used for information retrieval. The second similarity, called *Jaccard similarity*, normalizes the number of common neighbors by the arithmetic mean

of the two neighborhoods' size. The third similarity, called *min similarity*, normalizes the number of common neighbors by the minimum of the two neighborhoods' size. In the section titled "Evaluation," we compare the similarities with respect to the clustering accuracy.

When a member of a cluster shares a similar structure with one of its neighbors, their computed structural similarity will be large. We apply a threshold ε to the computed structural similarity when assigning cluster membership, formulized in the following ε -neighborhood definition.

Definition 3 [ε -Neighborhood].

$$N_{\varepsilon}(v) = \{w \in \Gamma(v) | \sigma(v, w) \geq \varepsilon\}.$$

When a vertex shares structural similarity with enough neighbors, it becomes a nucleus or *seed* for a cluster. Such a vertex is called a *core vertex*. Core vertices are a special class of vertices that have a minimum of μ neighbors with a structural similarity that exceeds the threshold ε . From core vertices we grow the clusters. In this way the parameters μ and ε determine the clustering of the network. For a given ε , the minimal size of a cluster is determined by μ .

Definition 4 [Core]. Let $\varepsilon \in \mathbb{R}$ and $\mu \in \mathbb{N}$. A vertex $v \in V$ is called a core with respect to ε and μ , if its ε -neighborhood contains at least μ vertices, formally:

$$CORE_{\varepsilon, \mu}(v) \Leftrightarrow |N_{\varepsilon}(v)| \geq \mu.$$

We grow clusters from core vertices as follows. If a vertex is in ε -neighborhood of a core, it should also be in the same cluster because they share a similar structure and are connected. This idea is formulized in the following definition of direct structure reachability.

Definition 5 [Direct Structure Reachability].

$$\begin{aligned} DirREACH_{\varepsilon, \mu}(v, w) \\ \Leftrightarrow CORE_{\varepsilon, \mu}(v) \wedge w \in N_{\varepsilon}(v). \end{aligned}$$

Direct structure reachability is symmetric for any pair of cores. However, it is asymmetric if one of the vertices is not a core. The following definition is a canonical extension of direct structure reachability.

Definition 6 [Structure Reachability].

Let $\varepsilon \in \mathfrak{N}$ and $\mu \in \mathfrak{S}$. A vertex $w \in V$ is structure reachable from $v \in V$ w.r.t ε and μ , if there is a chain of vertices $v_1, \dots, v_n \in V$, $v_1 = v$, $v_n = w$ such that v_{i+1} is directly structure reachable from v_i , formally:

$$\begin{aligned} REACH_{\varepsilon, \mu}(v, w) \\ \Leftrightarrow \exists v_1, \dots, v_n \in V : v_1 = v \wedge v_n = w \wedge \\ \forall i \in \{1, \dots, n-1\} : DirREACH_{\varepsilon, \mu} \\ (v_i, v_{i+1}). \end{aligned}$$

The structure reachability is transitive, but it is asymmetric. It is only symmetric for a pair of cores. More specifically, the structure reachability is a transitive closure of direct structure reachability.

Two noncore vertices in the same cluster may not be structure-reachable because the core condition may not hold for them. But they still belong to the same cluster because they both are structure reachable from the same core. This idea is formulized in the following definition of structure connectivity.

Definition 7 [Structure Connectivity].

Let $\varepsilon \in \mathfrak{N}$ and $\mu \in \mathfrak{S}$. A vertex $v \in V$ is structure-connected to a vertex $w \in V$ w.r.t ε and μ , if there is a vertex $u \in V$ such that both v and w are structure reachable from u , formally:

$$\begin{aligned} CONNECT_{\varepsilon, \mu}(v, w) \\ \Leftrightarrow \exists u \in V : REACH_{\varepsilon, \mu}(u, v) \\ \wedge REACH_{\varepsilon, \mu}(u, w). \end{aligned}$$

The structure connectivity is a symmetric relation. It is also reflective for the structure reachable vertices.

Now, we are ready to define a cluster as structure-connected vertices, which is maximal w.r.t. structure reachability.

Definition 8 [Structure-Connected Cluster]. Let $\varepsilon \in \mathfrak{N}$ and $\mu \in \mathfrak{S}$. A nonempty subset $C \subseteq V$ is called a *structure-connected cluster* (SCC) w.r.t ε and μ , if all vertices in C are structure-connected and C is maximal w.r.t structure reachability, formally:

$$SCC_{\varepsilon, \mu}(C) \Leftrightarrow$$

1. Connectivity:

$$\forall v, w \in C : CONNECT_{\varepsilon, \mu}(v, w)$$

2. Maximality:

$$\begin{aligned} \forall v, w \in V : v \in C \wedge REACH_{\varepsilon, \mu}(v, w) \\ \Rightarrow w \in C. \end{aligned}$$

Now we can define a clustering of a network G w.r.t. the given parameters ε and μ as all structure-connected clusters in G .

Definition 9 [Clustering]. Let $\varepsilon \in \mathfrak{N}$ and $\mu \in \mathfrak{S}$. A clustering P of network $G = \langle V, E \rangle$ w.r.t. ε and μ consists of all structure-connected clusters w.r.t. ε and μ in G , formally:

$$\begin{aligned} CLUSTERING_{\varepsilon, \mu}(P) \Leftrightarrow P \\ = \{C \subseteq V \mid SCC_{\varepsilon, \mu}(C)\}. \end{aligned}$$

On the basis of the clustering definition above, some vertices may not belong to any clusters. They are outliers, in the sense that structurally they are not similar to their neighbors, formally:

Definition 10 [Outlier]. Let $\varepsilon \in \mathfrak{N}$ and $\mu \in \mathfrak{S}$. For a given clustering P , that is, $CLUSTERING_{\varepsilon, \mu}(P)$, if a vertex $v \in V$ does not belong to any clusters, it is an outlier w.r.t. ε and μ , formally,

$$OUTLIER_{\varepsilon, \mu}(v) \Leftrightarrow \forall C \in P : v \notin C.$$

The outliers may play different rolls. Some outliers, such as vertex 6 in Fig. 1, connect to many clusters and act as a hub. Others, such as vertex 13 in Fig. 1, connect to relatively few clusters and are potentially outsiders because they have only weak connections to

the network. In the following, we formalize the notion of an outlier and their classification as either hubs or outsiders.

Definition 11 [Hub]. Let $\varepsilon \in \mathbb{N}$ and $\mu \in \mathbb{N}$. For a given clustering P , that is, $CLUSTERING_{\varepsilon, \mu}(P)$, if an outlier $v \in V$ has neighbors belonging to two or more different clusters w.r.t. ε and μ , it is a hub (it bridges different clusters) w.r.t. ε and μ , formally,

$$HUB_{\varepsilon, \mu}(v) \Leftrightarrow$$

1. $OUTLIER_{\varepsilon, \mu}(v)$
2. v bridges different clusters:

$$\begin{aligned} \exists p, q \in \Gamma(v) : \exists X, Y \in P : X \\ \neq Y \wedge p \in X \wedge q \in Y. \end{aligned}$$

Definition 12 [Outsider]. Let $\varepsilon \in \mathbb{N}$ and $\mu \in \mathbb{N}$. For a given clustering P , that is, $CLUSTERING_{\varepsilon, \mu}(P)$, an outlier $v \in V$ is an outsider if and only if all its neighbors belong to a single cluster or other outliers, formally,

$$OUTSIDER_{\varepsilon, \mu}(v) \Leftrightarrow$$

1. $OUTLIER_{\varepsilon, \mu}(v)$
2. v does not bridge different clusters:

$$\begin{aligned} \neg \exists p, q \in \Gamma(v) : \exists X, Y \in P : X \\ \neq Y \wedge p \in X \wedge q \in Y. \end{aligned}$$

In practice, the definition of a hub and an outsider is flexible. The more clusters an outlier bridges, the more strongly that vertex is indicated to be a hub. This point is discussed further when actual networks are considered.

The following lemmas are important for validating the correctness of our proposed algorithm. Intuitively, the lemmas mean the following. Given a network $G = \langle V, E \rangle$ and two parameters ε and μ , we can find structure-connected clusters in a two-step approach. First, choose an arbitrary vertex from V satisfying the core condition as a seed. Second, retrieve all the vertices that are structure reachable from the seed to obtain the cluster grown from the seed.

Lemma 1. Let $v \in V$. If v is a core, then the set of vertices that are structure reachable from v is a structure-connected cluster, formally:

$$\begin{aligned} CORE_{\varepsilon, \mu}(v) \wedge C &= \{w \in V | REACH_{\varepsilon, \mu}(v, w)\} \\ &\Rightarrow SCC_{\varepsilon, \mu}(C). \end{aligned}$$

Proof.

1. $C \neq \emptyset$:
By assumption, $CORE_{\varepsilon, \mu}(v)$ and thus, $REACH_{\varepsilon, \mu}(v, v) \Rightarrow v \in C$.
2. Maximality:

Let $p \in C$ and $q \in V$ and

$$\begin{aligned} REACH_{\varepsilon, \mu}(p, q) \\ \Rightarrow REACH_{\varepsilon, \mu}(v, p) \\ \wedge REACH_{\varepsilon, \mu}(p, q) \\ \Rightarrow REACH_{\varepsilon, \mu}(v, q), \\ \text{since structure reachability} \\ \text{is transitive.} \\ \Rightarrow q \in C. \end{aligned}$$

3. Connectivity:

$$\begin{aligned} \forall p, q \in C : REACH_{\varepsilon, \mu}(v, p) \\ \wedge REACH_{\varepsilon, \mu}(v, q) \\ \Rightarrow CONNECT_{\varepsilon, \mu}(p, q), \text{ via } v. \end{aligned}$$

Furthermore, a structure-connected cluster C with respect to ε, μ is uniquely determined by any of its cores, that is, each vertex in C is structure reachable from any of the cores of C and, therefore, a structure-connected cluster C contains exactly the vertices that are structure reachable from an arbitrary core of C .

Lemma 2. Let $C \subseteq V$ be a structure-connected cluster. Let $p \in C$ be a core. Then, C equals the set of vertices that are structure reachable from p , formally:

$$\begin{aligned} SCC_{\varepsilon, \mu}(C) \wedge p \in C \wedge CORE_{\varepsilon, \mu}(p) \\ \Rightarrow C = \{v \in V | REACH_{\varepsilon, \mu}(p, v)\}. \end{aligned}$$

Proof. Let $\hat{C} = \{v \in V \mid REACH_{\varepsilon, \mu}(p, v)\}$. We have to show that $C = \hat{C}$:

1. $\hat{C} \subseteq C$: it is obvious from the definition of $\hat{C}$.
2. $C \subseteq \hat{C}$: Let $q \in C$. By assumption, $p \in C \wedge SCC_{\varepsilon, \mu}(C)$.

$\Rightarrow \exists u \in C : REACH_{\varepsilon, \mu}(u, p)$
 $\quad \wedge RECH_{\varepsilon, \mu}(u, q)$
 $\Rightarrow REACH_{\varepsilon, \mu}(p, u)$,
 since both u and p are cores,
 and structure reachability
 is symmetric for cores.

$\Rightarrow REACH_{\varepsilon, \mu}(p, q)$, since structure reachability is transitive.

$\Rightarrow q \in \hat{C}$.

ALGORITHM

In this section, we describe the algorithm that implements the search for clusters, hubs, and outsiders. As mentioned in the section titled “Structure-Connected Cluster,” the search visits each vertex once to find structure-connected clusters and the outliers, and then classifies each outlier as either a hub or an outsider based on their connectivity to the clusters.

```

ALGORITHM ( $G = \langle V, E \rangle, \varepsilon, \mu$ )
// all vertices in  $V$  are labeled as unclassified;

for each unclassified vertex  $v \in V$  do
// STEP 1. check whether  $v$  is a core;
  if  $CORE_{\varepsilon, \mu}(v)$  then
// STEP 2.1. if  $v$  is a core, a new cluster is expanded;
    generate new cluster ID;
    insert all  $x \in N_{\varepsilon}(v)$  into queue  $Q$ ;
    while  $Q \neq \emptyset$  do
       $y$  = first vertex in  $Q$ ;
       $R = \{x \in V \mid \text{DirRECH}_{\varepsilon, \mu}(y, x)\}$ ;
      for each  $x \in R$  do
        if  $x$  is unclassified or an outlier then
          assign current cluster ID to  $x$ ;
        if  $x$  is unclassified then
          insert  $x$  into queue  $Q$ ;
      remove  $y$  from  $Q$ ;
    else
// STEP 2.2. if  $v$  is not a core, it is labeled as an outlier
      label  $v$  as outlier;
    end for.
// STEP 3. further classifies outliers
for each outlier  $v$  do
  if  $(\exists x, y \in \Gamma(v) \mid x.\text{clusterID} \neq y.\text{clusterID})$  then
    label  $v$  as hub
  else
    label  $v$  as outsider;
  end for.
end.

```

Figure 2. The pseudocode of our algorithm.

The pseudocode of the algorithm is presented in Fig. 2. The algorithm performs one pass of a network and finds all structure-connected clusters for a given parameter setting. At the beginning, all vertices are labeled as unclassified. The algorithm either assigns a vertex to a cluster or labels it as an outlier. For each vertex that is not yet classified, it checks whether this vertex is a core (STEP 1 in Fig. 2). If the vertex is a core, a new cluster is expanded from this vertex (STEP 2.1 in Fig. 2). Otherwise, the vertex is labeled as an outlier (STEP 2.2 in Fig. 2). To find a new cluster, the algorithm starts with an arbitrary core v and search for all vertices that are structure-reachable from v in STEP 2.1. This is sufficient to find the complete cluster containing vertex v , due to Lemma 2. In STEP 2.1, a new cluster ID is generated that will be assigned to all vertices found in STEP 2.1. The algorithm begins by inserting all vertices in ε -neighborhood of vertex v into a queue. For each vertex in the queue, it computes all directly reachable vertices and inserts those vertices into the queue that are still unclassified. This is repeated until the queue is empty.

The outliers can be further classified as hubs or outsiders in STEP 3. If an outlier connects to two or more clusters, it is classified as a hub. Otherwise, it is an outsider. This final classification is done according to what is appropriate for the network. As mentioned earlier, the more the clusters in which an outlier has neighbors, the more strongly that vertex acts as a hub between those clusters. Likewise, a vertex might bridge only two clusters, but how strongly it is viewed as a hub may depend on how aggressively it bridges them.

As discussed in the section titled “The Notion of Structure-Connected Cluster,” the results of our algorithm do not depend on the order the vertices are processed. The partitioning (number of clusters and association of cores to clusters) is determinate.

COMPLEXITY ANALYSIS

In this section, we present an analysis of the computation complexity of the algorithm.

Given a network with m edges and n vertices, we first find all structure-connected clusters w.r.t. a given parameter setting by checking each vertex of the network (STEP 1 in Fig. 2). This entails retrieval of all the vertex’s neighbors. Using an adjacency list, a data structure where each vertex has a list of which vertices it is adjacent to, the cost of a neighborhood query is proportional to the number of neighbors, that is, the degree of the query vertex. Therefore, the total cost is $O(\deg(v_1) + \deg(v_2) + \dots + \deg(v_n))$, where $\deg(v_i)$, $i = 1, 2, \dots, n$ is the degree of vertex v_i . If we sum all the vertex degrees in G , we count each edge exactly twice: once from each end. Thus the running time is $O(m)$.

We also derive the running time in terms of the number of vertices, should the number of edges be unknown. In the worst case, each vertex connects to all the other vertices for a complete graph. The worst case total cost, in terms of the number of vertices, is $O(n(n-1))$, or $O(n^2)$. However, real networks generally have sparser degree distributions. In the following, we derive the complexity for an average case, for which we know the probability distribution of the degrees. One type of network is the random graph, studied by Erdős and Rényi [11]. Random graphs are generated by placing edges randomly between vertices. Random graphs have been employed extensively as models of real-world networks of various types, particularly in epidemiology. The degree of a random graph has a Poisson distribution:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{z^k e^{-z}}{k!},$$

which indicates that most nodes have approximately the same number of links (close to the average degree $E(k) = z$). In the case of random graphs the complexity of the algorithm is $O(n)$.

Many real networks, such as social networks, biological networks, and the WWW follow a power-law degree distribution. The probability that a node has k edges, $P(k)$, is on the order $k^{-\alpha}$, where α is the degree exponent. A value between 2 and 3 was observed for the degree exponent for most biological and non-biological networks studied by the Faloutsos

et al. [12] and Barabási and Oltvai [13]. The expected value of degree is $E(k) = \alpha/(\alpha - 1)$. In this case, the average cost of the algorithm is again $O(n)$.

We conclude that the complexity in terms of the number of edges in the network for our algorithm is, in general, linear. The complexity in terms of the number of vertices is quadratic in the worst case of a complete graph. For real networks like social networks, biological networks, and computer networks, we expect linear complexity with respect to the number of vertices. This is confirmed by our empirical study described in the next section.

EVALUATION

In this section we evaluate our algorithm using both synthetic and real datasets. We first compared the different structural similarities defined in the section titled “Structure-Connected Cluster” for the accuracy of the clustering. The performance of the algorithm is then compared with fast modularity-based network clustering algorithm proposed in Clauset *et al.* [5], which is faster than many competing algorithms: its running time on a network with n vertices and m edges is $O(md \log n)$ where d is the depth of the dendrogram describing the hierarchical cluster structure. We implemented our algorithm in C++. We used the original source code of fast modularity algorithm by Clauset *et al.* [14]. All the experiments were conducted on a PC with a 2.0 GHz Pentium 4 processor and 1 GB of RAM.

Efficiency

To evaluate the computational efficiency of the proposed algorithm, we generated 10 networks with the number of vertices ranging from 1000 to 1,000,000 and the number of edges ranging from 2182 to 2,000,190. We adapted the construction as used in Newman and Girvan [3] as follows: first we generate clusters such that each vertex connects to vertices within the same cluster with a probability P_i , and connects to vertices outside its cluster with a probability $P_o < P_i$. Next, we add a number of hubs and outsiders. An

example of a generated network is presented in Fig. 3.

The running time for fast modularity and our algorithm on the synthetic networks are plotted in Figs 4 and 5, respectively. The running time is plotted both as a function of the number of nodes and the number of edges. Figure 5 shows that our algorithm’s running time is in fact linear w.r.t. to the number of vertices and the number of edges, while fast modularity’s running time is basically quadratic and scales poorly for large networks. Note the difference in scale for the y-axis between the two figures.

Effectiveness

To evaluate the effectiveness of network clustering, we use real datasets whose clusters are known *a priori*. These real datasets include American College Football and Books about US politics. We also apply the clustering algorithm to customer data integration. We use adjusted Rand index (ARI) [15] as a measure of effectiveness of network clustering algorithms in addition to visually comparing the generated clusters to the actual.

Adjusted Rand Index. A measure of agreement is needed when comparing the results of a network clustering algorithm to the expected clustering. Rand index [16] serves this purpose. One problem with the Rand index is that the expected value is not constant when comparing two random clusters. An ARI was proposed by Hubert and Arabie [15] to fix this problem. The ARI is defined as follows:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2} \right] - \left[\sum_i \binom{n_{i,i}}{2} \sum_i \binom{n_{j,j}}{2} \right] / \binom{n}{2}},$$

where $n_{i,j}$ is the number of vertices in both clusters x_i and y_j ; and $n_{i,\cdot}$ and $n_{\cdot,j}$ are the number of vertices in cluster x_i and y_j , respectively.

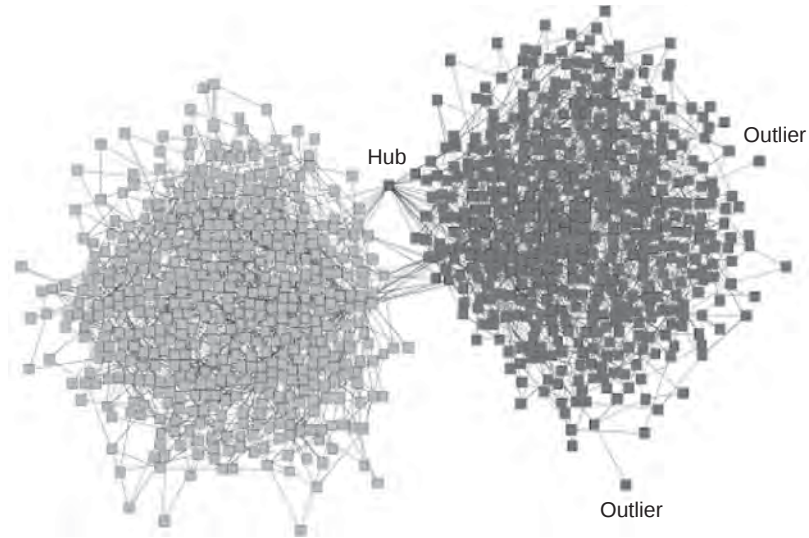


Figure 3. A Synthetic network with 1000 vertices.

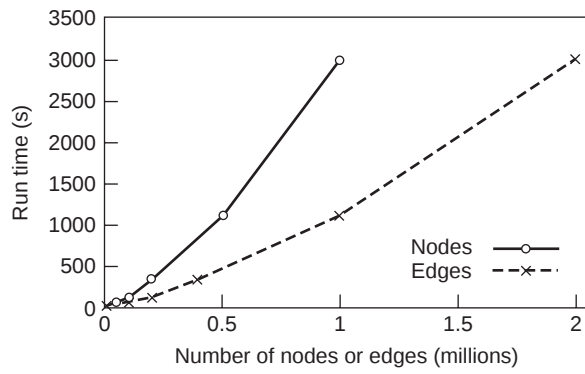


Figure 4. Running time for fast modularity algorithm.

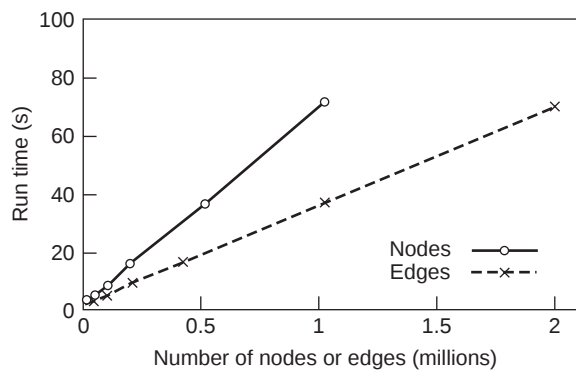


Figure 5. Running time for our algorithm.

Milligan and Cooper [17] evaluated many different indices for measuring agreement between two network clustering with different numbers of clusters and recommended the ARI as the measure of choice. We adopt the ARI as our measure of agreement between the network clustering result and the true clustering of the network. The ARI lies between 0 and 1. When the two clustering agree perfectly, the ARI is 1.

Performance of Various Structural Similarities. Our algorithm groups vertices based on their structural similarity. We compared the various structural similarities defined in the section titled “Structure-Connected Cluster” for the accuracy of the clusters they generate, measured by ARI. The results on real networks including college football, political books, and customers described in the sections titled “College Football,” “Books About US Politics,” and “Customer Data Integration,” respectively are listed in Table 1.

The cosine similarity (σ_{cos}) achieves overall the best accuracy in comparison with other similarity measures. In our following experiments, we use cosine similarity for our algorithm.

College Football. The first real dataset we examine is the 2006 National Collegiate Athletic Association (NCAA) Football Bowl Subdivision (formerly Division 1-A) football schedule. This example is inspired by the set studied by Newman and Girvan [3], who consider contests between Division 1-A teams in 2000. Our set is more complex; we consider all contests of the Bowl Subdivision schools including those against schools in lower divisions.

Table 1. Performance of Various Structural Similarities

	COSINE (σ_{cos})	MIN (σ_{min})	JACCARD (σ_{jaccard})
College football	1	0.255	0.983
Political books	0.708	0.661	0.574
CG1	1	1	1
CG2	1	0.942	1

The challenge is to discover the underlying structure of this network—the college conference system. The NCAA divides 115 schools into 11 conferences. In addition, there are four independent schools at this top level: Army, Navy, Temple, and Notre Dame. Each Bowl Subdivision school plays against schools within their own conference, against schools in other conferences, and against lower division schools. The network contains 180 vertices (119 Bowl Subdivision schools and 61 lower division schools) interconnected by 787 edges. Figure 6 shows this network with schools in the same conference identified by color.

This example illustrates the kinds of structures that our method seeks to address. Schools in the same conference are clusters. The four independent schools play teams in many conferences but belong to none; they are called *hubs*. The lower division schools are only connected weakly to the clusters in the network; they are called *outsiders*.

First we cluster this network by using the fast modularity algorithm. The results, for which the modularity is 0.599 is shown in Fig. 7. Maximizing Newman’s modularity gives a satisfying network clustering, identifying nine clusters. All schools in the same conference are clustered together. However, two of the conferences are merged (the Western Athletic and Mountain West conferences and the Mid-American and Big Ten conferences), the four independent schools are classified into various conferences despite their hub-like properties. All lower division teams are assigned to clusters.

Next, we cluster the network using our algorithm, using the parameters ($\varepsilon = 0.5$, $\mu = 2$). This clustering succeeds in capturing all the features of the network. The 11 clusters are identified, corresponding exactly to the 11 conferences. All schools in the same conference are clustered together. The independent schools and the lower division schools are unclassified; they stand apart from the clusters. The four independent schools show strong properties as hubs; they have inactive edges that connect them to a large number of clusters—at minimum five. In contrast, the lower division schools have only weak connections to clusters, one or

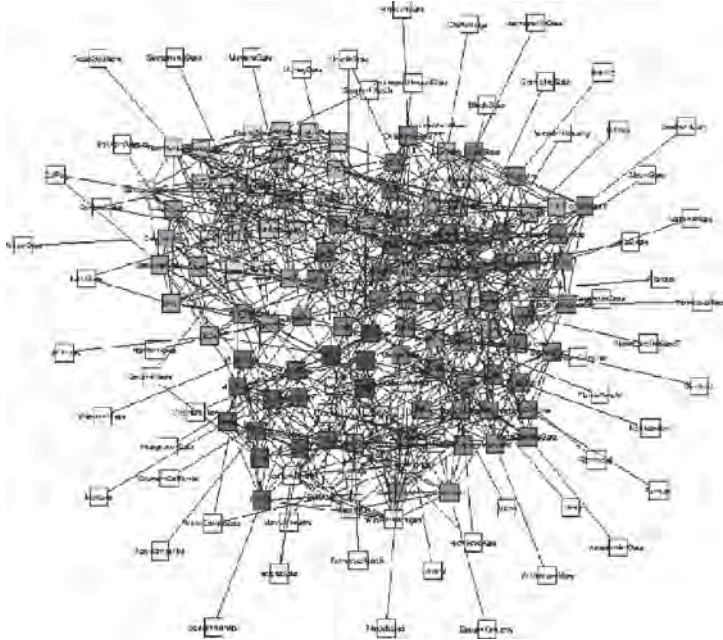


Figure 6. NCAA Football Bowl Subdivision schedule as a network, showing the 12 conferences in color, independent schools in black, and lower division schools in white.

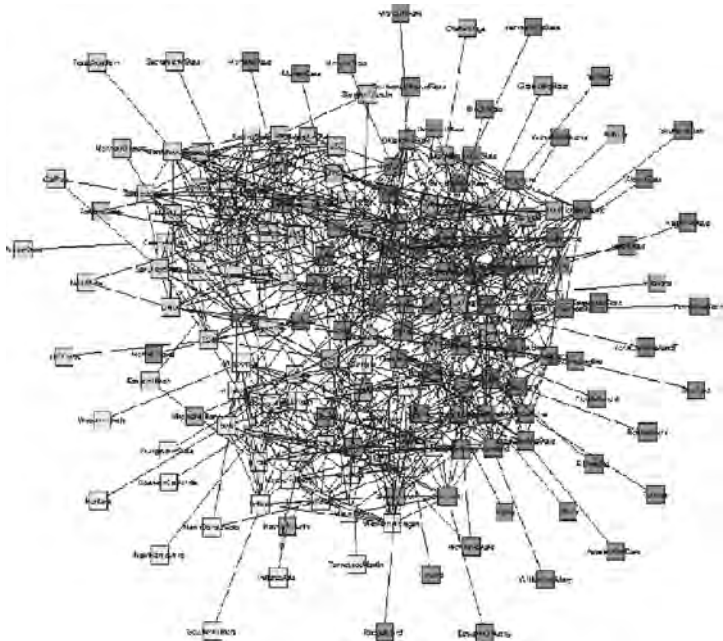


Figure 7. NCAA Football Bowl Subdivision schedule as clustered by fast modularity algorithm.

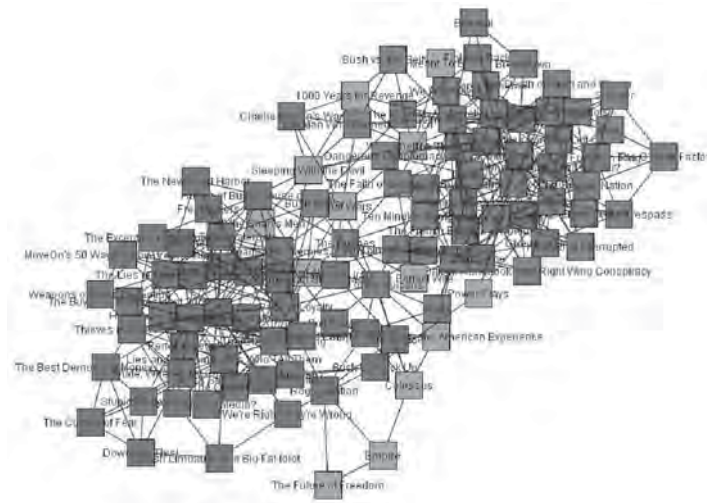


Figure 8. Political book network.

two, and in a single case three. They are true outsiders. This clustering matches perfectly with the underlying structure shown in Fig. 6.

Books about US Politics. The second example is the classification of books about US politics. We use the dataset of books about US politics compiled by Valdis Krebs [18]. The vertices represent books about US politics sold by the online bookseller Amazon.com. The edges represent frequent copurchasing of books by the same buyers, as indicated by the “customers who bought this book also bought these other books” feature on Amazon. The vertices have been given values “l,” “n,” or “c” to indicate whether they are “liberal,” “neutral,” or “conservative.” These alignments were assigned separately by Mark Newman [19] based on a reading of the descriptions and reviews of the books posted on Amazon. The political books network is illustrated in Fig. 8. The “conservative,” “neutral,” and “liberal” books are represented by red, gray, and blue, respectively.

First, we apply our algorithm to the political books network, using the parameters ($\varepsilon = 0.35$, $\mu = 2$). Our goal is to find clusters that represent the different political orientations of the books. The result is presented in Fig. 9. Our algorithm successfully finds

three clusters representing “conservative,” “neutral,” and “liberal” books, respectively. The obtained clusters are illustrated using three different shapes: squares for “conservative” books, triangles for “neutral” books, and circles for “liberal” books. In addition, each vertex is labeled with the book title.

The result for the fast modularity algorithm is presented in Fig. 10. The fast modularity algorithm found four clusters, presented using circles, triangles, squares, and hexagons. Although two dominant clusters, represented by circles and squares, align well with the “conservative” and “liberal” classes, the “neutral” class is mostly misclassified. This demonstrates again that fast modularity algorithm cannot handle vertices that bridge clusters.

Customer Data Integration. Finally we apply the network clustering algorithms to detect groups of records for the same individual, a problem called customer data integration (CDI). A large database of records consisting of names and addresses are matched against each other. The database contains multiple records for the same individual, but they manifest variations in the names and addresses whose causes range from the use of nicknames and abbreviations to data entry errors. If two records match, we

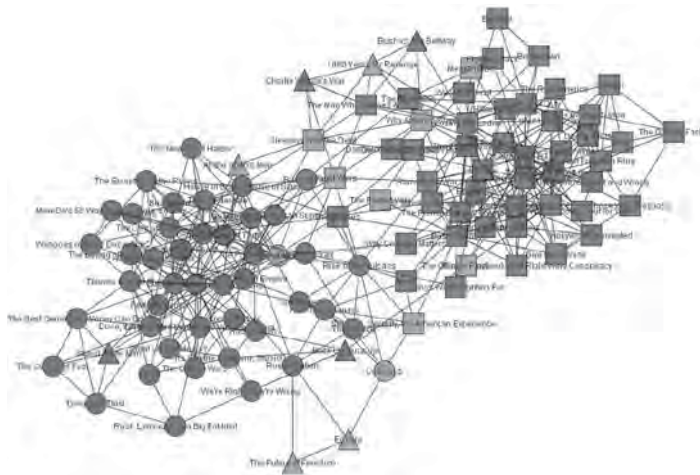


Figure 9. The result of our algorithm on political book network.

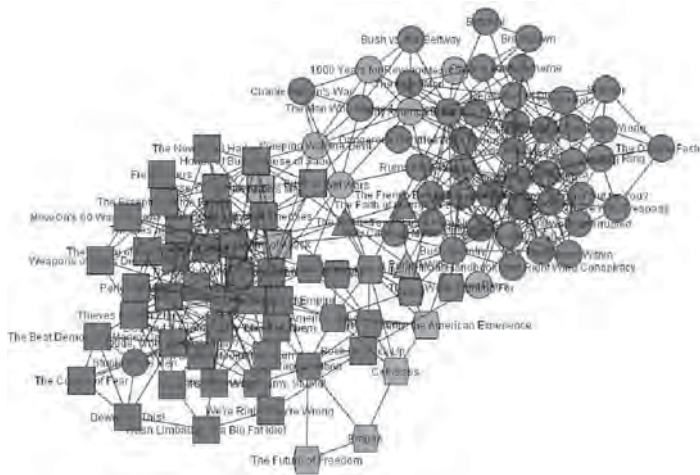


Figure 10. The result of fast modularity algorithm on political book network.

connect them with an edge. From a large file we extract sets of interconnected records for study. We test two networks, CG1 and CG2 (shown in Fig. 11). Network CG1 represents data for two individuals and two poor-quality records that represent no true individual. Network CG2 represents four individuals, one of whom is represented by a single instance.

The clustering results of our algorithm, using the parameters ($\varepsilon = 0.7, \mu = 2$), are presented in Fig. 12. The results demonstrate that it successfully found all the clusters and outliers. The results of fast modularity

algorithm are presented in Fig. 13. It is clear that it failed to identify any outliers.

Adjusted Rand Index Comparison. As mentioned in the section titled “Adjusted Rand Index,” the ARI is an effective measure of the similarity of a clustering result to the true clustering. The results for College Football, Political Books, CG1 and CG2 are presented in Table 2.

The ARI results clearly demonstrate that the proposed algorithm outperforms the fast modularity algorithm and produces clustering that resemble the true clustering for the real-world networks in our study.



Figure 11. Customer networks CG1 and CG2.

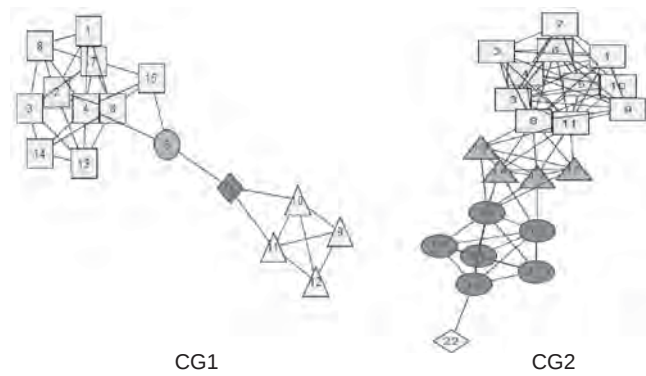


Figure 12. The result of our algorithm on CG1 and CG2.

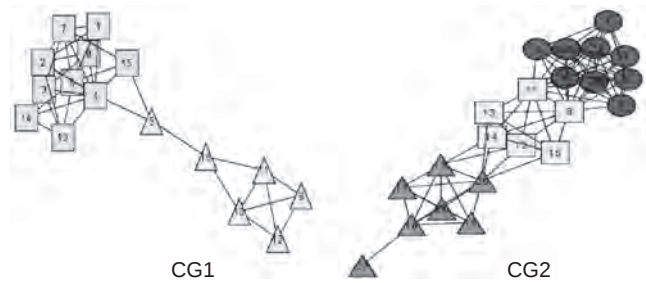


Figure 13. The result of fast modularity algorithm on CG1 and CG2.

Table 2. Adjust Rand Index Comparison

	Our Algorithm	Fast Modularity Algorithm
College football	1	0.24
Political books	0.71	0.64
CG1	1	0.85
CG2	1	0.68

Input Parameters. Our algorithm uses two parameters: ε and μ . To choose them we

adapted the heuristic suggested for DBSCAN in Ester *et al.* [9]. This involves making a k -nearest neighbor query for a sample of vertices and noting the nearest structural similarity as defined in the section titled “Structure-Connected Cluster.” The query vertices are then sorted in ascending order of nearest structural similarity. A typical k -nearest similarity plot is shown in Fig. 14. The knee indicated by a vertical line shows that an appropriate ε value for this network is 0.7. This knee represents a separation of

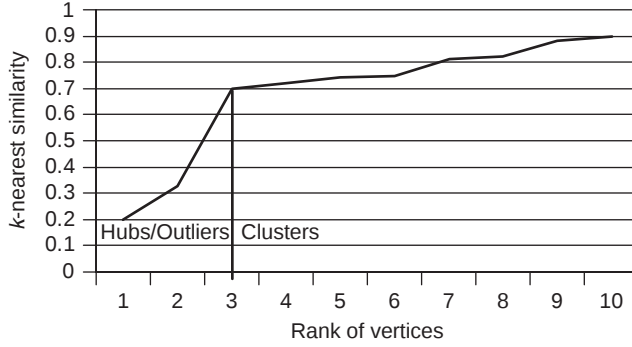


Figure 14. Sorted k -nearest structural similarity.

vertices belonging to clusters to the right from hubs and outliers to the left. Usually a sample of 10% of the vertices is sufficient to locate the knee. In the absence of such an analysis, an ε value between 0.5 and 0.8 is normally sufficient to achieve a good clustering result. We recommend a value of 2 for μ .

RELATED WORK

Network clustering is the division of a network into set of subnetworks, called *clusters*. More specifically, given a network $G = \{V, E\}$, where V is a set of vertices and E is a set of edges between vertices, the goal of network clustering is to divide G into k disjoint subnetworks $G_i = \{V_i, E_i\}$, in which $V_i \cap V_j = \emptyset$ for any $i \neq j$, and $V = \sum_{i=1}^k V_i$. The number of subnetworks, k , may or may not be known a priori. In this article, we focus on simple, undirected, and unweighted networks.

The problem of finding good clusters for networks has been studied for some decades in many fields, particularly computer science and physics. Here we review some of the more common methods.

The *min-max cut method* [1] seeks to cluster a network $G = \{V, E\}$ into two clusters A and B . The principle of min-max clustering is minimizing the number of connections between A and B and maximizing the number of connections within each. A *cut* is defined by the number of edges that would have to be removed to isolate the vertices in cluster A from those in cluster B . The min-max

cut algorithm searches for the clustering that creates two clusters whose *cut* is minimized while maximizing the number of remaining edges.

A pitfall of this method is that, if one cuts out a single vertex from the network, one will probably achieve the optimum. Therefore, in practice, the optimization must be accompanied with some constraint, such as A and B should be of equal or similar size, or $|A| \approx |B|$. Such constraints are not always appropriate; for example, in social networks some communities are much larger than the others.

To amend the issue, a *normalized cut* was proposed [2], which normalizes the cut by the total number of connections between each cluster to the rest of the network. Therefore, cutting out one vertex or some small part of the network will no longer always yield an optimum.

Both min-max cut and normalized cut methods cluster a network into two clusters. To divide a network into k clusters, one has to adopt a top-down approach, splitting the network into two clusters, and then further splitting these clusters, and so on, until k clusters have been detected. There is no guarantee of the optimality of recursive clustering. There is no measure of the number of clusters that should be produced when k is unknown. There is no indicator to stop the bisection procedure.

Recently, *modularity* was proposed as a quality measure of network clustering [3]. For a clustering of network with k clusters,

the modularity is defined as follows:

$$Q = \sum_{s=1}^k \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

L is the number of edges in the network, l_s is the number of edges between vertices within cluster s , and d_s is the sum of the degrees of the vertices in cluster s . The modularity of a clustering of a network is the fraction of all edges that lie within each cluster minus the fraction that would lie within each cluster if the network's vertices were randomly connected. Optimal clustering is achieved when the modularity is maximized. Modularity is defined such that it is 0 for two extreme cases: when all vertices clustered into a single cluster, and when the vertices are clustered at random. Note that the modularity measures the quality of any network clustering. Normalized and min-max cut measures only the quality of a clustering of two clusters.

Finding the maximum Q is NP-complete. Instead of performing an exhaustive search, various optimization approaches are proposed. For example, a greedy method based on a hierarchical agglomeration clustering algorithm is proposed in Clauset *et al.* [5], which is faster than many competing algorithms: its running time on a network with n vertices and m edges is $O(md \log n)$ where d is the depth of the dendrogram describing the hierarchical cluster structure. Also, Guimera and Amaral [4] optimize modularity using simulated annealing.

To summarize, the network clustering methods discussed in this section aim to find clusters such that there are many connections between vertices within the same clusters and few without. While all these network clustering methods successfully find clusters, they are generally unable to detect hubs and outsiders like those in the example in Fig. 1. Such vertices invariably are included in one cluster or another.

CONCLUSIONS

Organizing related data is a fundamental task in many fields of science and

engineering. Many algorithms for partitioning networks have been proposed by practitioners in different disciplines including computer science and physics. Successful examples are min-max cut [1] and normalized cut [2], as well as modularity-based network clustering algorithms [3–5]. While such algorithms can successfully detect clusters in networks, they tend to fail to identify and isolate two kinds of vertices that play special roles: vertices that bridge clusters (hubs) and vertices that are marginally connected to clusters (outsiders). Identifying hubs is particularly valuable for applications such as viral marketing and epidemiology. As vertices that bridge clusters, hubs are responsible for spreading ideas or disease. In contrast, outsiders have little or no influence, and may be isolated as noise in the data.

In this article, we proposed a new algorithm to detect clusters, hubs, and outsiders in networks. It clusters vertices based on their common neighbors. Two vertices are assigned to a cluster according to how they share neighbors. This makes sense when you consider social communities. People who share many friends create a community, and the more friends they have in common, the more intimate the community. But in social networks there are different kinds of actors. There are also people who are outsiders (like hermits), and there are people who are friendly with many communities but belong to none (like politicians). The latter play a special role in small-world networks [10].

We applied our algorithm to some real-world networks including finding conferences using only the NCCA College Football schedule, grouping political books based on copurchasing information, and customer data integration. In addition, we compared the new algorithm with the fast modularity-based algorithm in terms of both efficiency and effectiveness. The theoretical analysis and empirical evaluation demonstrate superior performance over the modularity-based network clustering algorithm.

In the future we plan to apply our algorithm to analyze biological networks such as metabolic networks and gene coexpression networks.

REFERENCES

1. Ding C, He X, Zha H, *et al.* A min-max cut algorithm for graph partitioning and data clustering. Proceedings of ICDM; 2001.
2. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 2000;22(8):888–905.
3. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004;69:026113.
4. Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature 2005;433:895–900.
5. Clauset A, Newman M, Moore C. Finding community in very large networks. Phys Rev E 2004;70:066111.
6. Kleinberg J. Authoritative sources in a hyperlinked environment. Proceeding of the 9th ACM-SIAM Symposium on Discrete Algorithms. San Francisco (CA); 1998.
7. Domingos P, Richardson M. Mining the network value of customers. Proceedings of the 7th ACM SIGKDD; San Francisco (CA); 2001. pp. 57–66.
8. Wang Y, Chakrabarti D, Wang C, *et al.* Epidemic spreading in real networks: an eigenvalue viewpoint. 22nd Symposium on Reliable Distributed Systems (SRDS'03). Florence: IEEE; 2003. pp. 25–34. ISBN: 0-7695-1955-5.
9. Ester M, Kriegel H-P, Sander J, *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96). Portland (OR): AAAI Press; 1996. pp. 291–316.
10. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998; 393:440–442.
11. Erdős P, Rényi A. On random graphs. Publ Math (Debrecen) 1959;6:290–297.
12. Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. Proceedings of SIGCOMM; Cambridge (MA); 1999.
13. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–113.
14. <http://cs.unm.edu/~aaron/research/fastmodularity.htm>.
15. Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.
16. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 1971;66:846–850.
17. Milligan GW, Cooper MC. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behav Res 1986;21:441–458.
18. <http://www.orgnet.com/>.
19. <http://www-personal.umich.edu/~mejn/netdata/>.

ACCELERATED LIFE MODELS

MIKHAIL NIKULIN
IMB, Université Victor Segalen,
Bordeaux, France

INTRODUCTION

We consider dynamic regression models which are well adapted to study the phenomena of longevity, aging, fatigue, and degradation of complex systems, and hence appropriate to be used in the organization of the efficient statistical process of quality control in dynamic environments.

It is well known that traditionally only the failure time data are usually used for the product reliability estimation or the estimation of survival characteristics. Failures of highly reliable units are rare, for example, the lifetime of semiconductors is very long, and to test devices under usual conditions would require too much test time and excessively large sample size. So other information should be used in addition to failure time data, which could be censored.

One way of obtaining a complementary reliability information is to use higher level of experimental factors, stresses or covariates (such as temperature, voltage or pressure) to increase the number of failures and hence to obtain reliability information quickly. This procedure provides the methods known today as the accelerated life testing (ALT). This method is described very well in statistical literature [1–6].

The ALT of technical or biotechnical systems is an important practical statistical method of estimation of the reliability and the quality improvement of new systems without waiting for the operating life of an item. The ALT has been recognized as a necessary activity to ensure the reliability of electronic products used in military, aerospace, automotive and mobile (cellular, laptop computers) applications, from which one can see

that accelerated testing of electronic products offers great potential for improving the reliability life testing, to obtain expected reliability results for the products more quickly than under normal operating conditions, given in terms of covariates (stresses), which could be time-varying. It is interesting to study the possibility of taking into account the cumulative effect of the applied stresses on aging, fatigue, and degradation of testing items or systems. It is evident that the extrapolating reliability or quality from the ALT always carries the risk that the accelerated stresses do not properly excite the failure mechanism, which dominates at operating (normal) stresses.

Another way of obtaining complementary reliability information is to measure some parameters, which characterize the aging or wear of the product in time. In analysis of longevity of highly reliable complex industrial or biological systems, the degradation processes provide an important additional information about the aging, degradation, and deterioration of systems, and from this point of view these degradation data are really a very rich source of reliability information and often offer many advantages over failure time data. Degradation is the natural response for some tests, and it is also natural that with degradation data it is possible to make useful reliability and statistical inference even with no observed failure. It is evident that sometimes it may be difficult and costly to collect degradation measures from some components or materials. Sometimes it is possible to apply the expert's estimation of the level of degradation.

Statistical inference from ALT is possible, if failure time regression models relating failure time distribution with external explanatory variables (covariates, stresses) influencing the reliability are well chosen. Statistical inference from failure time-degradation data with covariates needs even more complicated models relating failure time distribution not only with external but also with internal explanatory variables

(degradation, wear), which explain the state of units before the failures. In the last case, models for degradation process distribution are needed too.

In this paper we discuss the most-used failure time regression models used for analysis of failure time and failure time-degradation data with covariates.

NOTATIONS

Let us denote by T the random time-to-failure of a unit (or system). We also say that T is the time of *hard* or *traumatic failure*. Let S be the survival function and λ be the hazard rate, then

$$\begin{aligned} S(t) &= \mathbf{P}\{T > t\}, \\ \lambda(t) &= \lim_{h \rightarrow 0} \frac{1}{h} \mathbf{P}\{t \leq T < t+h | T \geq t\} \\ &= -\frac{d[\ln S(t)]}{dt} \end{aligned} \quad (1)$$

from where it follows that $S(\cdot)$ can be written as

$$S(t) = e^{-\Lambda(t)}, \quad \text{where} \quad \Lambda(t) = \int_0^t \lambda(s) ds$$

is the cumulative hazard function. $F(\cdot) = 1 - S(\cdot)$ is the cumulative distribution function of T . In survival analysis and reliability, the models are often formulated in terms of hazard rate functions. The most common shapes of hazard rates are monotone, $\cup$ -shaped or $\cap$ -shaped; for example, Ref. 7.

If the form of the survival or other equivalent function is supposed to be known and this function depends only on finite-dimensional parameter, the corresponding model is parametric. Classes of parametric models used in reliability are given in Refs 3, 5 and 6.

Suppose the explanatory variable is a deterministic time function

$$x = x(\cdot) = (x_1(\cdot), \dots, x_m(\cdot))^T : [0, \infty[\rightarrow B \in \mathbb{R}^m,$$

which is a vector of covariates itself or a realization of a stochastic process $X(\cdot)$, which is also called the covariate process, $X(\cdot) = (X_1(\cdot), \dots, X_m(\cdot))^T$. We denote by E , the set of

all possible (admissible) covariates and by E_1 , the set of all constants over time covariates, $E_1 \subset E$. We do not discuss here the questions of choice of X_i and m , but they are very important for the planning (design) of the experiments and for statistical inference. The covariates can be interpreted as the control [8], since we may consider models of aging in terms of differential equations and therefore use all theory and techniques from the optimal control theory. We may say that we consider statistical modeling with dynamic design or in dynamic environments.

Let $E_2, E_2 \subset E$, be a set of step-stresses of the form

$$x(t) = x_1 1_{\{0 \leq t < t_1\}} + x_2 1_{\{t_1 \leq t\}}, \quad x_1, x_2 \in E_1. \quad (2)$$

In accordance with Equation (1), the survival, the hazard rate, the cumulative hazard, and the distribution functions given x are

$$\begin{aligned} S(t|x) &= \mathbf{P}(T > t|x(s); 0 \leq s \leq t), \\ \lambda(t|x) &= -\frac{S'(t|x)}{S(t|x)}, \\ \Lambda(t|x) &= -\ln[S(t|x)], \\ F(t|x) &= \mathbf{P}(T \leq t|x(s); 0 \leq s \leq t) = 1 - S(t|x), \\ &x \in E, \end{aligned}$$

from where one can see their dependence on the life-history up to time t . On any family E of admissible stresses, we may consider a class $\{S(\cdot|x), x \in E\}$ of survival functions which could be very rich.

We say that the time $f(t|x)$ under the stress x_0 is *equivalent* to the time t under the stress x if the probability that a unit used under the stress x would survive till the moment t is *equal* to the probability that a unit used under the stress x_0 would survive till the moment $f(t|x)$:

$$\begin{aligned} S(t|x) &= \mathbf{P}\{T > t|x(s); 0 \leq s \leq t\} \\ &= \mathbf{P}\{T > f(t|x)|x_0(s); 0 \leq s \leq f(t|x)\} \\ &= S(f(t|x)|x_0). \end{aligned}$$

It implies that:

$$f(t|x) = S^{-1}[S(t|x)|x_0], \quad x \in E. \quad (3)$$

Let x and y be two admissible stresses: $x, y \in E$. We say that a stress y is *accelerated* with respect to x , if:

$$\begin{aligned} S(t|x) &\geq S(t|y), \quad \forall t \geq 0, \\ S(\cdot|x), S(\cdot|y) &\in \{S(\cdot|z), z \in E\}. \end{aligned}$$

ACCELERATED LIFE AND FAILURE TIME REGRESSION MODELS

Failure time regression models relating the lifetime distribution to possibly time dependent external explanatory variables are considered in this section. Failure time regression models relates failure time distribution not only with external but also with internal explanatory variables will be discussed in the next section. Now, such models are used not only in reliability but also in demography, dynamics of populations, gerontology, biology, survival analysis, genetics, radiobiology, biophysics; everywhere people study the problems of longevity, aging, and degradation using stochastic modeling.

In reliability, ALT in particular, the choice of a good regression model often is more important than in survival analysis. For example, in ALT units are tested under accelerated stresses which shorten the life. Using such experiments the life under the usual stress is estimated using some regression model. The values of the usual stress are often not in the range of the values of accelerated stresses, since the wide separation between experimental and usual stresses is possible; so if the model is mis-specified, the estimators of survival under the usual stress may be very bad.

Sedyakin's Model

The physical principle in reliability, proposed by Sedyakin [9], gives an interesting way to prolong any class of survival functions $\{S(\cdot|x), x \in E_1\}$ indexed by constant in time stresses to a class of survival functions indexed by step-stresses, for example, from E_2 , given by Equation (2).

According to Sedyakin we may consider the next model on E_2 : if two moments t_1 and

t_1^* are equivalent, that is, the probabilities of survival until these moments under stresses x_1 and x_2 , respectively, are equal, that is, $S(t_1|x_1) = S(t_1^*|x_2)$, then

$$\lambda(t_1 + s|x) = \lambda(t_1^* + s|x_2), \quad \forall s \geq 0, \quad (4)$$

with x defined by Equation (2). The meaning of this rule of time-shift for these step-stresses on E_2 one can also see in terms of the survival functions:

$$S(t|x) = \begin{cases} S(t|x_1), & 0 \leq t < t_1, \\ S(t - t_1 + t_1^*|x_2), & t \geq t_1. \end{cases} \quad (5)$$

The model given by Equation (4) (or Eq. 5) is called the *Sedyakin's model* on E_2 .

The general Sedyakin (GS) model Bagdonavičius and Nikulin (2000) generalizes this idea, by supposing that the hazard rate at any moment t depends on the stress at this moment and on the probability of survival until this moment:

$$\lambda(t|x) = g(x(t), S(t|x)), \quad x \in E.$$

The Sedyakin's model is too wide for ALT and failure time regression data analysis. It only states that units which did not fail until equivalent moments under different stresses have the same risk to fail after these moments under identical stresses. Nevertheless, this model is useful for construction of narrower models and for analysis of redundant systems. At the end of this section, we note that recently [10] studied an interesting application of Sedyakin's and the accelerated failure time (AFT) model for analysis of reliability of redundant systems with one main unit and $m - 1$ stand-by units operating in "warm" conditions, that is, under lower stress than the main one.

Accelerated Failure Time Model

AFT model is more adapted for failure time regression analysis [2,3,5,6,25,26]. In ALT, it is the most-used model.

We say that AFT model holds on E if there exists a positive continuous function $r : E \rightarrow \mathbf{R}^1$ such that for any $x \in E$, the survival and the cumulative hazard functions under a covariate realization x are given by formulas:

$$S(t|x) = G\left(\int_0^t r[x(s)] ds\right) \quad \text{and} \\ \Lambda(t|x) = \Lambda_0\left(\int_0^t r[x(s)] ds\right), \quad x \in E \quad (6)$$

respectively, where $G(t) = S(t|x_0)$, $\Lambda_0(t) = -\ln G(t)$, x_0 is a given (usual) stress, $x_0 \in E$. The function r changes locally with the timescale. From the definition of $f(t|x)$ (Eq. 3) it follows that for the AFT model on E

$$f(t|x) = \int_0^t r[x(s)] ds, \quad \text{hence} \\ \frac{\partial f(t|x)}{\partial t} = r(x(t)) \quad \text{at the continuity points} \\ \text{of } r[x(\cdot)]. \quad (7)$$

Note that the model can be considered as parametric (r and G belong to parametric families) semiparametric (one of these functions is unknown, other belongs to a parametric family) or nonparametric (both are unknown). The AFT model can also be given by the next formula:

$$\lambda(t|x) = r(x(t)) \quad q(\Lambda(t|x)), \quad x \in E. \quad (8)$$

This equality shows that different from the famous Cox model, the hazard rate $\lambda(t|x)$ is proportional not only to some functions of the stress at the moment t but also to a function of the cumulative hazard $\Lambda(t|x)$ at the moment t . It means that the hazard rate at the moment t depends not only on the stress applied at this moment but also on the stress applied in the past, that is, in the interval $[0, t]$.

In parametric modeling, a baseline survival function G is taken from some class of parametric distributions such as Weibull, log-normal, log-logistic etc., and the function r is taken in the form $r(x) = e^{\beta^T \varphi(x)}$, where $\varphi(x)$ is a specified, possibly multidimensional function of x [3,6]. A versatile model is obtained when G belongs to the *power generalized Weibull* (PGW) family of distributions [4]. In terms of the survival functions, the PGW family is given by the next formula

$$S(t, \sigma, \nu, \gamma) = \exp \left\{ 1 - \left[1 + \left(\frac{t}{\sigma} \right)^\nu \right]^{\frac{1}{\gamma}} \right\}, \\ t > 0, \gamma > 0, \nu > 0, \sigma > 0. \quad (9)$$

If $\gamma = 1$, we have the Weibull family of distributions. If $\gamma = 1$ and $\nu = 1$, we have the exponential family of distributions. This class of distributions has nice probability properties. For various values of the parameters, the hazard rate can be *constant*, *monotone* (increasing or decreasing), *unimodal* or $\cap$ -shaped, and *bathtub* or $\cup$ -shaped.

Semiparametric estimation procedures for the AFT model are given in Ref. 11 and developed by many authors. Nonparametric estimation procedures with special plans of experiments are given in Ref. 12. The AFT model is popular in reliability theory because of its *interpretability*, its convenient mathematical properties and its consistency with some engineering and physical principles. Nevertheless, the assumption that the survival distributions under different covariate values differ only in scale is rather *restrictive*.

Changing Shape and Scale (CHSS) Model

A natural generalization of the AFT model [3,6], the changing shape and scale (CHSS) model is obtained by supposing that different constant stresses $x \in E_1$ influence not only the scale but also the shape of the survival distribution: there exist positive functions on E_1 $\theta(x)$ and $\nu(x)$ such that for any $x \in E_1$

$$S(t|x) = S \left\{ \left(\frac{t}{\theta(x)} \right)^{\nu(x)} \middle| x_0 \right\}; \quad (10)$$

here x_0 is fixed stress, for example, design (usual) stress. Let us consider generalizations of the model (Eq. 10) to the case of time-varying stresses. We say that the CHSS model [13], holds on E if there exist two positive functions r and ν on E such that for all $x(\cdot) \in E$:

$$S(t|x) = S \left(\int_0^t r\{x(\tau)\} \tau^{\nu(x(\tau))-1} d\tau \middle| x_0 \right), \quad x \in E.$$

The variation of stress locally changes not only the scale but also the shape of distribution. In terms of the hazard rate, the CHSS

model can be written in the form:

$$\lambda(t|x) = r\{x(t)\}q(\Lambda(t|x))t^{v(x(t))-1}. \quad (11)$$

This model is not in the class of the GS models because the hazard rate $\lambda(t|x)$ depends not only on $x(t)$ and $\Lambda(t|x)$ but also on t . Generally it is recommended to choose $v(x) = e^{\gamma T x}$. Statistical analysis of the CHSS model is done in Ref. 14, and an interesting application is considered in Ref. 8.

FAILURE TIME-DEGRADATION MODELS WITH EXPLANATORY VARIABLES

Failures of highly reliable units are rare. As we saw, one way of obtaining complementary information is to use higher levels of experimental factors or covariates, that is, ALT. In previous sections, we described the most-applied ALT models. Other ways of obtaining this complementary information is to measure some parameters characterizing degradation or damage of the unit (the product) in time. Both methods can be combined. Here, we consider some approaches to model the relationship between failure time data and degradation data with external covariates.

Statistical modeling of observed degradation processes can help to understand the different real physical, chemical, medical, biological, physiological, or social degradation processes of aging, due to cumulated wear and tiredness. The information about real degradation processes helps us to construct degradation models, which permit us to predict the cumulative damage. Suppose that the following data may be available for reliability characteristics estimation: failure times (possibly censored), explanatory variables (covariates, stresses) and the values of some observable quantity characterizing the degradation of units. We call a failure nontraumatic when the degradation attains a critical level z_0 . Other failures are called traumatic.

The failure rate of traumatic failure may depend on covariates, degradation level and time.

Good reviews on failure time-degradation data modeling are given in Refs 13, 15–17.

We develop the models given in these excellent papers. Suppose that under fixed constant covariate the degradation is non-decreasing stochastic process $Z(t), t \geq 0$ with right continuous trajectories with finite left hand limits. Denoted by $T^{(tr)}$, the moment of the traumatic failure and by $\lambda^{(tr)}(t|Z) = \lambda^{(tr)}(t|Z(s), 0 \leq s \leq t)$ the conditional hazard rate of the traumatic failures given the degradation Z till time t .

Suppose that this conditional hazard rate has *two additive components*: one, λ , related to observed degradation values, other, μ , related to nonobservable degradation (aging) and to possible shocks causing sudden traumatic failures. For example, observable degradation of tires is the wear of the protector. The failure rate of tire explosion depends on thickness of the protector, on nonmeasured degradation level of other tire components and on intensity of possible shocks (hitting a kerb, nail). So the hazard rate of the traumatic failure is modeled as follows:

$$\lambda^{(tr)}(t|Z) = \lambda(Z(t)) + \mu(t).$$

The function $\lambda(z)$ characterizes the dependence of the rate of traumatic failures on degradation. Suppose that covariates influence degradation rate and traumatic event intensity. In such case the model needs to be modified.

Let $x = (x_1, \dots, x_m)^T$ be a vector of s possibly time dependent one-dimensional covariates. We assume in what follows that x_i are deterministic or realizations of bounded right continuous with finite left hand limits stochastic processes. Denote informally by $Z(t|x)$ the degradation level at the moment t for units functioning under the covariate x . We suppose that the covariates influence locally the scales of the traumatic failure time distribution component related to nonobservable degradation (aging) and to possible shocks, that is, the AFT model is true for this component.

Let us explain it in detail. Let us denote

$$S_1^{(tr)}(t|Z) = \exp \left\{ - \int_0^t \lambda[Z(u)] du \right\},$$

$$S_2^{(tr)}(t) = \exp \left\{ - \int_0^t \mu(u) du \right\}$$

as the survival functions correspond to the hazard rates $\lambda(Z(u))$ and $\mu(u)$. The first survival function is conditional given the degradation.

The AFT model defines the following relation of the second survival function and the covariates

$$S_2^{(tr)}(t|x) = S_2 \left(\int_0^t e^{\gamma T x(s)} ds \right);$$

the parameters γ have the same dimension as x . Set

$$f(t, x, \beta) = \int_0^t e^{\beta T x(u)} du,$$

and denote by $g(t, x, \beta)$ the inverse of $f(t, x, \beta)$ with respect to the first argument. The function $f(t, x, \beta)$ is time transformation in dependence on x . We consider the following model for degradation process under covariates:

$$Z(t|x) = Z(f(t, x, \beta)).$$

The covariates have double influence on the distribution of the first traumatic failure component: via degradation and directly. So we combine the AFT and the proportional hazards models:

$$S_1^{(tr)}(t|x, Z) = \exp \left\{ - \int_0^t e^{\beta T x(u)} \lambda(Z(u|x)) du \right\}.$$

Let us denote by

$$\begin{aligned} S^{(tr)}(t|x, Z) &= P(T^{(tr)} > t|x(u), \\ &\quad Z(u|x), 0 \leq u \leq t), \\ \lambda^{(tr)}(t|x, Z) &= - \frac{d}{dt} \ln S^{(tr)}(t|x, Z) \end{aligned}$$

the conditional distribution function and the failure rate of the traumatic failure given the covariates and the degradation.

So we consider the following *failure time-degradation covariate model*:

$$\begin{aligned} \lambda^{(tr)}(t|x, Z) &= e^{\beta T x(t)} \lambda(Z(f(t, x, \beta))) \\ &\quad + e^{\gamma T x(t)} \mu(f(t, x, \gamma)). \end{aligned}$$

Under this model,

$$\begin{aligned} S^{(tr)}(t|x, Z) &= \mathbf{P}(T^{(tr)} > t|x(u), \\ &\quad Z(u|x), 0 \leq u \leq t) \\ &= \exp \left\{ - \int_0^t e^{\beta T x(u)} \lambda(Z(u|x)) du - H(f(t, x, \gamma)) \right\}, \\ H(t) &= \int_0^t \mu(u) du. \end{aligned}$$

Let $T = \min(T^{(0)}, T^{(tr)})$ be the time of the unit failure. The survival function of the failure time T under the covariate x is

$$\begin{aligned} S(t|x) &= \mathbf{P}(T > t|x) = ES(t|x, Z), \\ S(t|x, Z) &= 1_{\{Z(t|x) < z_0\}} S^{(tr)}(t|x, Z). \end{aligned}$$

In particular, if $z_0 = \infty$, the survival functions $S(t|x, Z)$ and $S^{(tr)}(t|x, Z)$ coincide. Mean failure time under the covariate x is

$$\begin{aligned} e(x) &= \mathbf{E}(T|x) = \mathbf{E}(\mathbf{E}(T|x, Z)), \\ \mathbf{E}(T|x, Z) &= \int_0^{g(h(z_0), x, \beta)} S(t|x, Z) dt. \end{aligned}$$

The probability of nontraumatic failure under the covariate x in the interval $[0, t]$ is given by

$$\begin{aligned} P^{(0)}(t|x) &= \mathbf{E}P^{(0)}(t|x, Z), \\ P^{(0)}(t|x, Z) &= 1_{\{Z(t|x) \geq z_0\}} S(g(h(z_0), x, \beta)|x, Z). \end{aligned}$$

In particular, the probability of nontraumatic failure under the covariate x in the interval $[0, \infty)$ is obtained. The probability of traumatic failure under the covariate x in the interval $[0, t]$ is

$$\begin{aligned} P^{(tr)}(t|x) &= \mathbf{E}P^{(tr)}(t|x, Z), \\ P^{(tr)}(t|x, Z) &= 1 - S(t \wedge g(h(z_0), x, \beta)|x, Z). \end{aligned}$$

Analyzing failure time-degradation data requires not only the related probability of failures with degradation and covariates but also models for degradation process. The most-applied stochastic processes describing degradation are general path models and time-scaled stochastic processes with stationary and independent increments such as the γ -process, shock processes, and Wiener process with a drift:

General Path Models

$Z(t) = g(t, A, \theta)$, where g is a deterministic function and $A = (A_1, \dots, A_s)$ is a finite-dimensional random vector and θ is a finite-dimensional nonrandom parameter. The form of the function g may be suggested by the form of individual degradation curves. The degradation under the covariate x is modeled by

$$\begin{aligned} Z(t|x) &= g(f(t, x, \beta), A), \\ m(t|x) &= E g(f(t, x, \beta), A). \end{aligned}$$

Methods of estimation from degradation data are given in Ref. 3. Bagdonavičius *et al.* [18] considered estimation from failure time-degradation data.

Time-Scaled γ Process

$Z(t) = \sigma^2 \gamma(t)$, where $\gamma(t)$ is a process with independent increments such that for any fixed $t > 0$,

$$\gamma(t) \sim G(1, v(t)), v(t) = \frac{m(t)}{\sigma^2},$$

that is, $\gamma(t)$ has the γ distribution with the density

$$p_{\gamma(t)}(x) = \frac{x^{v(t)-1}}{\Gamma(v(t))} e^{-x}, \quad x \geq 0,$$

where $m(t)$ is an increasing function. Then,

$$Z(t|x) = \sigma^2 \gamma(f(t, x, \beta)).$$

The mean degradation and the covariances under the covariate x are

$$\begin{aligned} m(t|x) &= \mathbf{E}(Z(t|x)) = m(f(t, x, \beta)), \\ \mathbf{Cov}(Z(s|x), Z(t|x)) &= \sigma^2 m(f(s \wedge t, x, \beta)). \end{aligned}$$

Bagdonavičius and Nikulin [4,19] considered estimation from failure time-degradation data. Padgett and Tomlinson [20] considered estimation from degradation data.

Time-Scaled Wiener Process with a Drift

$Z(t) = m(t) + \sigma W(m(t))$, where W denotes the standard Wiener motion, that is, a process with independent increments such that $W(t) \sim N(0, t)$. Then,

$$Z(t|x) = m(f(t, x, \beta)) + \sigma W(m(f(t, x, \beta))).$$

The mean degradation and the covariance under the covariate x are

$$\begin{aligned} m(t|x) &= m(f(t, x, \beta)), \\ \mathbf{Cov}(Z(s|x), Z(t|x)) &= \sigma^2 m(f(s \wedge t, x, \beta)). \end{aligned}$$

Several authors [15,16,20,21] considered estimation from degradation data.

Shock Processes

Assume that degradation results from shocks, each of them leading to an increment of degradation. Let $T_n, (n \geq 1)$ be the time of the n th shock and X_n the n th increment of the degradation level. Let us denote by $N(t)$, the number of shocks in the interval $[0, t]$. Set $X_0 = 0$. The degradation process is given by

$$Z(t) = \sum_{n=1}^{\infty} 1\{T_n \leq t\} X_n = \sum_{n=0}^{N(t)} X_n.$$

Kahle and Wendt [17] model T_n as the moments of transition of the doubly stochastic Poisson process, that is, they suppose that the distribution of the number of shocks up to time t is given by

$$\mathbf{P}\{N(t) = k\} = \mathbf{E} \left\{ \frac{(Y \eta(t))^k}{k!} \exp \{-Y \eta(t)\} \right\},$$

where $\eta(t)$ is a deterministic function and Y is a nonnegative random variable with finite expectation. If Y is nonrandom, N is a nonhomogeneous Poisson process, in particular, when $\eta(t) = \lambda t$, N is a homogeneous Poisson process. If $\eta(t) = t$, then N is a mixed Poisson process. Other models for η may be used; for example, $\eta(t) = t^\alpha, \alpha > 0$. The random variable Y is taken from some parametric class of distributions.

Kahle and Lehmann [22] and Kahle and Wendt [27] considered parametric estimation from degradation data; Lehmann [16] considered estimation from failure time-degradation data; Harlamov [23] discusses inverse γ -process as a wear model; Zacks [24] gives failure distributions associated with general compound renewal damage processes. Bagdonavičius and Nikulin [4,19] considered the problem of estimation in accelerated degradation models, refer also to Nikulin *et al.* [25].

REFERENCES

1. Singpurwalla N. Inference from accelerated life tests when observations are obtained from censored data. *Technometrics* 1971;13: 161–170.
2. Viertl R. Statistical methods in accelerated life testing. Goettingen: Vandenhoeck and Ruprecht; 1988.
3. Meeker W, Escobar L. Statistical methods for reliability data. New York: Wiley; 1998.
4. Bagdonavičius V, Nikulin M. Accelerated life models. Boca Raton (FL): Chapman & Hall/CRC; 2002.
5. Lawless JF. Statistical models and methods for lifetime data. New York: Wiley; 2003.
6. Nelson W. Accelerated life testing: statistical models, test plans, and data analysis. New York: Wiley-Interscience; 2004.
7. Bagdonavičius V, Clerjaud L, Nikulin M. Accelerated life testing when the hazard rate function has cup shape. In: Huber C, Limnios N, Mesbah M, *et al.*, editors. Mathematical methods in survival analysis, reliability and quality of life. London: ISTE & J. Wiley; 2008. pp. 203–216.
8. Ceci C, Mazliak L. Optimal design in nonparametric life testing. *Stat Inference Stochastic Process* 2004;7:305–325.
9. Sedyakin NM. On a physical principle in reliability theory (in Russian). *Tech Cybern* 1966;3:80–87.
10. Bagdonavičius V, Masiulaityte I, Nikulin M. Statistical analysis of redundant systems with ‘warm’ stand-by units. *Stochastics Int J Prob Stochastic Process* 2000;80:115–128.
11. Lin DY, Ying Z. Semiparametric inference for accelerated life model with time dependent covariates. *J Stat Plan Inference* 1995;44:47–63.
12. Bagdonavičius V, Nikulin M. On nonparametric estimation in accelerated experiments with step stresses. *Statistics* 2000;33(2):349–365.
13. Singpurwalla N. Survival in dynamic environments. *Stat Sci* 1995;1:86–103.
14. Bagdonavičius V, Cheminade O, Nikulin M. Statistical planning and inference in accelerated life testing using the CHSS model. *J Stat Plan Inference* 2004;2:535–551.
15. Lehmann A. On degradation-failure models for repairable items. In: Nikulin M, Balakrishnan N, Mesbah M, *et al.*, editors. Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Boston: Birkhauser; 2004. pp. 65–80.
16. Lehmann A. Degradation-threshold-shock models. In: Nikulin M, Commenges D, Huber C, editors. Probability, statistics and modelling in public health. New York: Springer; 2006. pp. 286–298.
17. Kahle W, Wendt H. Statistical analysis of some parametric degradation models. In: Nikulin M, Commenges D, Huber C, editors. Probability, statistics and modelling in public health. New York: Springer; 2006. pp. 266–279.
18. Bagdonavičius V, Bikelis A, Kazakevičius V, *et al.* Non-parametric estimation from simultaneous renewal-failure-degradation data with competing risks. *J Stat Plan Inference* 2007;137:2191–2207.
19. Bagdonavičius V, Nikulin M. Estimation in degradation models with explanatory variables. *Lifetime Data Anal* 2001;7(1):85–103.
20. Padgett WJ, Tomlinson MA. Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. *Lifetime Data Anal* 2005;11:511–527.
21. Doksum KA, Normand SLT. Gaussian models for degradation processes - part I: methods for the analysis of biomarker data. *Lifetime Data Anal* 1995;1:131–144.
22. Kahle W, Lehmann A. Parameter estimation in damage processes: dependent observations of damage increments and first passage time. In: Kahle W, von Collani E, Franz F, *et al.*, editors. Advances in stochastic models for reliability, quality and safety. Boston (MA): Birkhauser; 1998. pp. 139–152.
23. Harlamov B. Inverse gamma-process as a model of wear. In: Antonov V, Huber C, Nikulin M, *et al.*, editors. Volume 2, Longevity, aging and degradation models in reliability, health, medicine and biology. Saint Petersburg: St. Petersburg State

- Polytechnical University Press; 2004. pp. 180–190.
24. Zacks S. Failure distributions associated with general compound renewal damage processes. In: Antonov V, Huber C, Nikulin M, *et al.*, editors. Volume 2, Longevity, aging and degradation models in reliability, public health, medicine and biology. St.Petersburg: St.Petersburg State Polytechnical University Press; 2004. pp. 336–344.
 25. Nikulin M, Limnios N, Balakrishnan N, *et al.*, editors. Advances on degradation models: applications to industry, medicine and finance. Boston (MA): Birkhauser; 2009.
 26. Martinussen T, Scheike TH. Dynamic regression models for survival data. New York: Springer; 2006.
 27. Wendt H, Kahle W. On parametric estimation for a position-dependent marking of a doubly stochastic Poisson process. In: Nikulin M, Balakrishnan N, Mesbah M, *et al.*, editors. Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Boston (MA): Birkhauser; 2004. pp. 473–486.

ACCIDENT PRECURSORS AND WARNING SYSTEMS MANAGEMENT: A BAYESIAN APPROACH TO MATHEMATICAL MODELS

ELISABETH PATÉ-CORNELL

Department of Management Science and
Engineering, Stanford University,
Stanford, California

THE IMPORTANCE OF PRECURSORS AND THE DIFFICULTY TO MONITOR THEM

Many accidents are preceded by near misses and/or precursors. Of course, in hindsight, they often could have been predicted and something should have been done in time to prevent a failure or reduce the risk. It is indeed a rare disaster that in retrospect did not involve a signal that is easy to identify after the fact and should have been detected earlier. One problem is that many such signals and near misses occur without significant consequences, and that organizations cannot always stop operations because someone has observed an unusual and troublesome event. Another problem is that in retrospect, it is easy to see why and how the accident and the signal were correlated but estimating the risk before the fact is much more problematic [1,2]. One key issue is to structure the hazard warnings to permit proper response [3]. The questions are thus first, how to monitor what most needs to be, and second how to assess the probability of an event (accident, disaster, of given severity) given the observation of such a signal. The next question is how to respond to a signal, sometimes with some lead time, sometimes immediately, given the costs and the benefits of different risk-reduction measures.

Identification and reporting of precursors through monitoring systems, and interpretation of the signal has been implemented in several industries, for example, the aviation reporting system [4,5]; in the nuclear power

industry, the sequence of status reports on precursors to potential severe core damage accidents [6–8] and the USNRC accident sequence precursor program [9,10]; and in the chemical industry, near miss management systems [11]. Also in the chemical industry, it was shown that accidents reoccur for similar reasons, and that observable disruptions can provide useful warnings [12]. Finally, in the context of storage of flammable gas, Zhou *et al.* [13] studied the dynamics of warnings based on the analysis of real-time risk data.

Precursors come in several forms. Signals can be observed through systematic reporting of near misses [14]. They can also be picked up by chance and indeed, by surprise. For example, an unknown terrorist group may appear to try to stage an attack, or a known technical system may threaten to fail in ways that were not anticipated. This is similar to finding by chance a needle in haystack. The problem is then to determine the significance of the signal, and the implications of that new piece of information in order to follow up if it is justified. This may entail collecting additional information that may either confirm or negate various hypotheses.

Third, one can observe near misses under the form of incomplete accident sequences; for example, an employee may have come close to derailing a train by excessive speed in a dangerous area. The repetition of such events can confirm the existence of a problem. The Concord supersonic aircraft, for instance, experienced more than 50 hits of its fuselage by pieces of tires that had blown up at take off, before such an event caused a tank rupture and a plane crash in 2000 [15]. In that case also, one can learn from a near miss by asking what was the probability of a full-fledged accident given what happened? And what should have been done to prevent it?

Fourth, through the direct monitoring of a continuous dynamic process, one can observe clearly, for example, how close one is from an event such as a flood by watching the level of water in a rising river. Given how fast the

process evolves and how much time one has to react when the hazard level exceeds different thresholds of threat, one can identify an alert threshold that would strike a balance between the lead time for protective actions and the probability of a false alert.

Precursors thus provide important information for risk management [16–19]. They are mostly useful when looking ahead to decide what to do to mitigate the risk, although observations of signals prior to an accident are often used in retrospect to find the root cause of a problem and/or assign the blame.

Relying on signals to take safety measures generally requires managing a trade-off between the probabilities of false positives and false negatives [20]. Classical statistics can provide an approach to the analysis of precursors [21,22] but they require data samples of sufficient size and a stable underlying condition that may not exist. Bayesian methods, by contrast rely on all existing information, which is updated given new data [23]. The use of these methods has been explored in the context of signals and precursors as an improvement over classical statistics [24]. They have also been described in the context of intelligence analysis [25], and counterterrorism [26,27].

What is described in this article relies in general on a probabilistic Bayesian method of precursor signal analysis in the context of a decision, which permits best observation and response [28–30]. The data then become input in a rational decision analysis [31] for risk management. This implies that the organization itself must be equipped to identify, communicate, and act upon warning signals [32–36]. The culture of the organization is thus a key factor in its ability to process such signals [37]. In designing a warning system, the main decisions are therefore which systems or process to monitor, what specific signals to observe, how to interpret the results, how to communicate the warnings, and how to best use the lead time that they provide. The fundamental issue is thus to set up an organizational warning system to observe and communicate appropriate signals. Confidentiality in reporting system is often an important success factor [38]. On

the one hand, one needs to filter out signals that do not require immediate attention so as not to swamp potential decision makers under lots of less relevant facts. On the other hand, filtering out relevant signals can be catastrophic as was, for example, the decision of FBI managers to ignore pre 9/11 signals that a number of people were taking suspect training in airplane operations.

In the end, the value of a warning or alert lies in the possibility that it may affect risk management decisions for the better. The concept of value of information [39] is thus central to the analysis of uncertainties about a signal, and the benefits of observation, analysis and response. That concept itself has its foundations in decision analysis and requires both a probabilistic assessment and an evaluation of the different possible outcomes given each option. This article describes a general probabilistic approach to the interpretation of signals in the context of risk management decisions, and presents examples based on Bayesian treatment of that information.

In the first part of the article, the importance of signals of a specific subsystem's malfunction is captured using a probabilistic risk analysis (PRA) of the overall system [40,41]. The analysis is extended to include human errors and thus signals that specific mistakes have taken place, along with their potential implications. A general model of optimization of a warning system is then presented, based on an explicit formulation of the trade-off between type I and type II errors. Finally, the principles of design and implementation of an organizational warning system are described.

PROBABILISTIC ANALYSIS OF WARNING SIGNALS

The principle behind the probabilistic analysis of a piece of information, such as a signal that something may be going wrong, is to separate the prior probability or base rate of the event (accident, catastrophe, disaster) and the quality of the signal. The prior probabilities of such events are provided by the information gathered so far, including performance statistics, test results, and expert

opinions. The quality of a precursor (signal, test, monitoring results, etc.) is described by the probabilities of false positives and false negatives, or put another way, the sensitivity (probability of true positive) and the specificity (probability of a true negative) of the test.

Noting E the adverse event that one is trying to avoid, S the signal that is observed, and NE or NS the negation of the event of the signal, the probability of an event *given* the signal ($p(E|S)$)¹ can be written using Bayes theorem and the total probability theorem:

$$\begin{aligned} p(E|S) &= p(E \text{ and } S)/p(S) \\ &= p(E)p(S|E)/[p(E \text{ and } S) + p(NE \text{ and } S)] \\ &= p(E)p(S|E)/[p(E)p(S|E) + p(NE) \\ &\quad \times p(S|NE)]. \end{aligned}$$

To describe the probabilities of errors, one can write that $p(S|E)$ is $1 - p(NS|E)$, where $p(NS|E)$ is the probability of a false negative and $p(S|NE)$ is the probability of a false positive defined per time unit or operation.

This formulation again allows separation of the characteristics of the event ($p(E)$) and of the quality of signals and of the performance of their sources (captors, sensors, human information, etc.).

The denominator $p(S)$, which is sometimes considered a mere “normalization factor,” is actually important because it requires that the analyst envision the possible cases in which the signal could be observed with or without the considered event. In the case of an intelligence system, for example, it requires constructing a structured set of scenarios that would yield the same information under different hypotheses.

¹As in classic probability notations, the vertical bar in $p(X|Y)$ means: probability of X *conditional* on Y or *given* Y, the comma in $p(X,Y)$ means joint probability of X and Y. The negation of X is noted here NX and $p(NX)$ means probability of NOT X, which is $1 - p(X)$.

Numerical Illustration

Assume a case in which the prior probability (base rate) of event E is 1/100 per year, where a signal S of E can be observed with a probability of false positive of 0.05 per year, and where the probability of a false negative is 0.01 given that E is going to occur. The formula above yields the probability of the event given the signal, which is 0.165 (16.5% chance of occurrence). Even though the signal occurred, it is still low because the prior of 1/100 is low.

Several Signals

The same type of analysis applies to the case where several signals can be observed. These signals may be conditionally dependent on the event, for example, because they come from the same source or related sources of information. The signals that are most significant are conditionally independent, that is, independent given the event. In general, however, even these signals are marginally dependent given an occurrence of the event to which they are both correlated. In addition, there can be further dependences if they are correlated not only by the occurrence of the event of interest but also by the source(s) of information. The probabilistic analysis of several signals thus encompasses all these possible dependence aspects. It relies on an extension of Bayes theorem, and in the general case can be expressed as

$$\begin{aligned} p(E|S_1 \text{ and } S_2) &= p(E \text{ and } S_1 \text{ and } S_2)/p(S_1 \text{ and } S_2) \\ &= p(E)p(S_1 \text{ and } S_2|E)/[p(E \text{ and } S_1 \text{ and } S_2) \\ &\quad + p(NE \text{ and } S_1 \text{ and } S_2)] \\ &= p(E)p(S_1|E)p(S_2|S_1 \text{ and } E)/ \\ &\quad [p(E)p(S_1|E)p(S_2|S_1 \text{ and } E) \\ &\quad + p(NE)p(S_1|NE)p(S_2|S_1 \text{ and } NE)]. \end{aligned}$$

In the case where S_1 and S_2 are conditionally independent, this formula can be simplified because $p(S_2|S_1, E)$ is equal to $p(S_2|E)$ and $p(S_2|S_1, NE)$ is equal to $p(S_2|NE)$. Therefore,

$$\begin{aligned}
p(E|S_1, S_2) &= p(E \text{ and } S_1 \text{ and } S_2) / p(S_1 \text{ and } S_2) \\
&= p(E) p(S_1 \text{ and } S_2 | E) / [p(E \text{ and } S_1 \text{ and } S_2) \\
&\quad + p(NE \text{ and } S_1 \text{ and } S_2)] \\
&= p(E) p(S_1 | E) p(S_2 | E) / \\
&\quad [p(E) p(S_1 | E) p(S_2 | E) \\
&\quad + p(NE) p(S_1 | NE) p(S_2 | NE)].
\end{aligned}$$

For an example of conditional dependence or independence of two signals, consider two weather predictions for the same place and the same day. If one is based on a satellite picture and the other on the arthritic pains of an individual who has linked them to the weather patterns, the signals seem to be independent given the weather (but generally correlated by the weather). If both predictions are generated from two photographs from the same satellite, they may be correlated not only by the weather itself but also by characteristics of the satellite's performance.

USING A PROBABILISTIC RISK ANALYSIS (PRA) TO IDENTIFY POTENTIAL PRECURSORS

One way to identify signals of possible malfunctions in large complex systems (including accidents that have not happened yet) is to use a PRA to identify the elements of accident scenarios. PRA [40,41] is based on a systematic identification of the failure modes of a system, structured into a set of exhaustive, mutually exclusive scenarios. If a minimum set of events that lead to a failure ("min-cut set") is not completed, the system does not fail but the subset of a failure mode event(s) that have occurred can be regarded as a near miss. For instance, if an aircraft operates on four jet engines but can work with a single one, failure of the four is a failure mode (leading to a crash) and failures of one, two, or three jet engines are near misses, the last one being the closest call.

More importantly perhaps, PRA allows computing the probability of a system failure given the occurrence of some basic component failures. One can thus use that analytical

framework to compute the probability of a failure given the unfolding of a partial accident scenario (failure mode, "min-cut set"), or given the deterioration of components.

Failure of a Redundancy in a Subsystem of Parallel Components

Consider an electric system with two generators in parallel, or a cooling system with several tanks in parallel. If one fails, the system still works but the probability of failure has become higher with the loss of one of the redundant elements. The failure probability may become very high, especially if the failures of the redundant elements are dependent. One can use a PRA to compute how close one then is to system failure.

Assume, for instance, that the failure modes (sets of events leading to system failure) have been identified as $M_1, M_2, M_3, \dots$. According to the total probability theorem, the probability of system failure as a function of that of the failure modes is

$$\begin{aligned}
p(F) &= \sum_i p(M_i) - \sum_{ij} p(M_i, M_j) \\
&\quad + \sum_{ijk} p(M_i, M_j, M_k) - \dots
\end{aligned}$$

Given that the failure mode M_i , for example, has not been completed, but that a subset $M_{i'}$ of the events that constitute M_i has been observed (e.g., failure of one out of two redundant elements), one can write the $p(F)$ given $M_{i'}$ conditioning the above equation on the occurrence of $M_{i'}$:

$$\begin{aligned}
p(F|M_{i'}) &= \sum_i p(M_i|M_{i'}) - \sum_{ij} p(M_i, M_j|M_{i'}) \\
&\quad + \sum_{ijk} p(M_i, M_j, M_k|M_{i'}) - \dots
\end{aligned}$$

$M_{i'}$ is a "near miss" but how close to an actual failure $M_{i'}$ has taken the system can thus be computed using this formula.

Numerical Illustration

Consider a system (Fig. 1) that has three subsystems in series, one of which is made of two components in parallel.

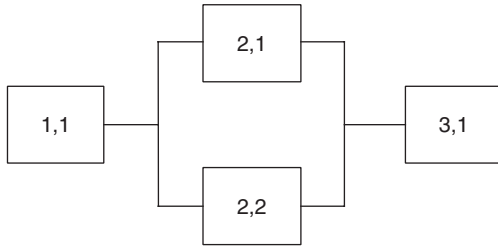


Figure 1. Simple system of three subsystems in series, one of which has two components in parallel.

Assume that the failures of 1.1 and 3.1 are independent of all others, that there is a dependence only between the failures of 2.1 and 2.2, and that the failure probabilities of the components per time unit or operation are

$$p(1.1) = 10^{-3}$$

$$p(2.1) = 10^{-2}$$

$$p(2.2|2.1) = 0.5$$

$$p(3.1) = 10^{-2}$$

$$\begin{aligned}
 p(F) &= p(1.1) + p(2.1)p(2.2|2.1) + p(3.1) \\
 &\quad - \Sigma p(\text{two failure modes}) \\
 &\quad + p(\text{three failure modes}).
 \end{aligned}$$

The failure of 2.1 can be considered a close call (near miss) because the probability of system failure is then

$$p(F|2.1) = p(1.1) + p(3.1) + p(2.2|2.1) \approx 0.511.$$

At that point, the safety of the system relies in large part on the performance of the redundant component. The risk manager has to strengthen it immediately and/or replace 2.1. If possible, dependences among failures of redundant systems should be reduced or eliminated. In the future, lessons should be drawn, in particular, from near misses involving human errors, in which case a critical mistake that affect several components can bring the system close to failure.

Increase of System Failure Probability by Weakening of a Component

Another precursor to system failure is the weakening of one of its components. PRA allows computing the probability of system failure given the increase of component failure probability. It can also guide the schedule, extent, and priorities of inspection and maintenance of the different subsystems and components according to their contribution to the system failure risk. An example is that of the heat shield of the space system. A weakening of the tiles or panels that protect the edges of the wings is an accident precursor. It can happen at take off, for example, under the impact of a piece of debris. It can happen in orbit, for instance, if the tiles are damaged during extravehicular activities. It can also happen if the tiles are poorly maintained and some bonds are weak. The following example shows the vulnerability of a mission to such weakening of the tiles.

Illustration: Effect of Errors in the Maintenance of the Space Shuttle Tiles

In 1986, the space shuttle Challenger exploded at liftoff due to a failure of the O-rings of the solid rocket boosters. Following that accident, NASA decided to do a number of PRAs to identify potential weak points of the system, and to try to address these problems before they cause another accident. In a study conducted at Stanford [42,43], a PRA was performed to assess the risks of shuttle failure due to a failure of the tiles of the orbiters' heat shield. Twenty-five thousand tiles protect the orbiter, each different from the other. The objective was to determine potential improvements of the tile maintenance given that weak bonds could become accident precursors. There are two main failure modes. The first one is a debris hit that causes tiles to debond. The second one involves debonding of a tile under regular loads (e.g., vibrations) because the tile capacity has been decreased. This can be caused, for example, by poor bonding of the tile during maintenance, which, in turn, can be the result of organizational factors, for example, excessive time constraints. After

a tile debonds, the adjacent ones at reentry are subjected to additional heat loads due to turbulences in the cavity. Hot gases can then enter the structure and damage critical subsystems under the orbiter's skin causing a catastrophic accident. The model was structured using four parameters that are characterized for each tile as a function of its location on the surface: the density of debris hits at that place that had been observed in previous missions, the aerodynamic forces (that contribute to the loss of adjacent tiles), the heat load, and the criticality level of the subsystems under the aluminum skin.

Figure 2 shows the influence diagram that was used in the study to compute the resulting probability of losing a shuttle because of a failure of the tiles.

The first results of the study were presented as a view of the orbiters' underside in which each zone was shaded as a function of the contribution of each tile to the overall failure probability (Fig. 3). The contribution of tile failure to the overall probability of a shuttle accident was found to be in the order of 10% of the overall mission failure probability

(10^{-3} per flight for an overall probability of accident in the order of 10^{-2}).

Warnings and Signals

Some weak bonds could be detected by manual inspection of the bond. Because this task was long and delicate, the prioritization of the tiles in terms of system vulnerability was essential.

But the 1990 study provided a warning of things to come, that is, the 2003 Columbia accident, where a piece of debris at take-off opened a gap in the heat shield [44]. An important part of the study results was the fact that debris hits contributed about half of the risks of losing an orbiter due to tile failure. Some possible debris trajectories had been computed at Johnson Space Center. Backtracking these trajectories from the more risk critical parts of the orbiter to their potential sources revealed that a critical part of the risk was attributable to the insulation of the external tank. The bonding of that insulation was weakened, in particular, by the attachment of a fuel line along the tank surface. Pieces of the tank insulation could

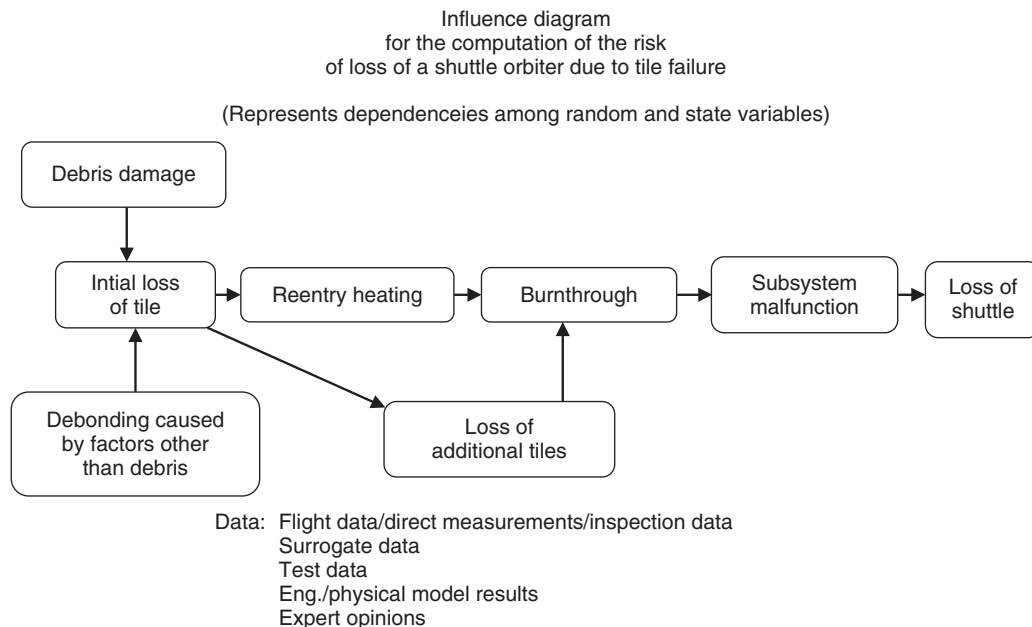


Figure 2. Structure of the risk assessment model for the loss of a shuttle mission due to loss of tiles. (Source: From Ref. 43).

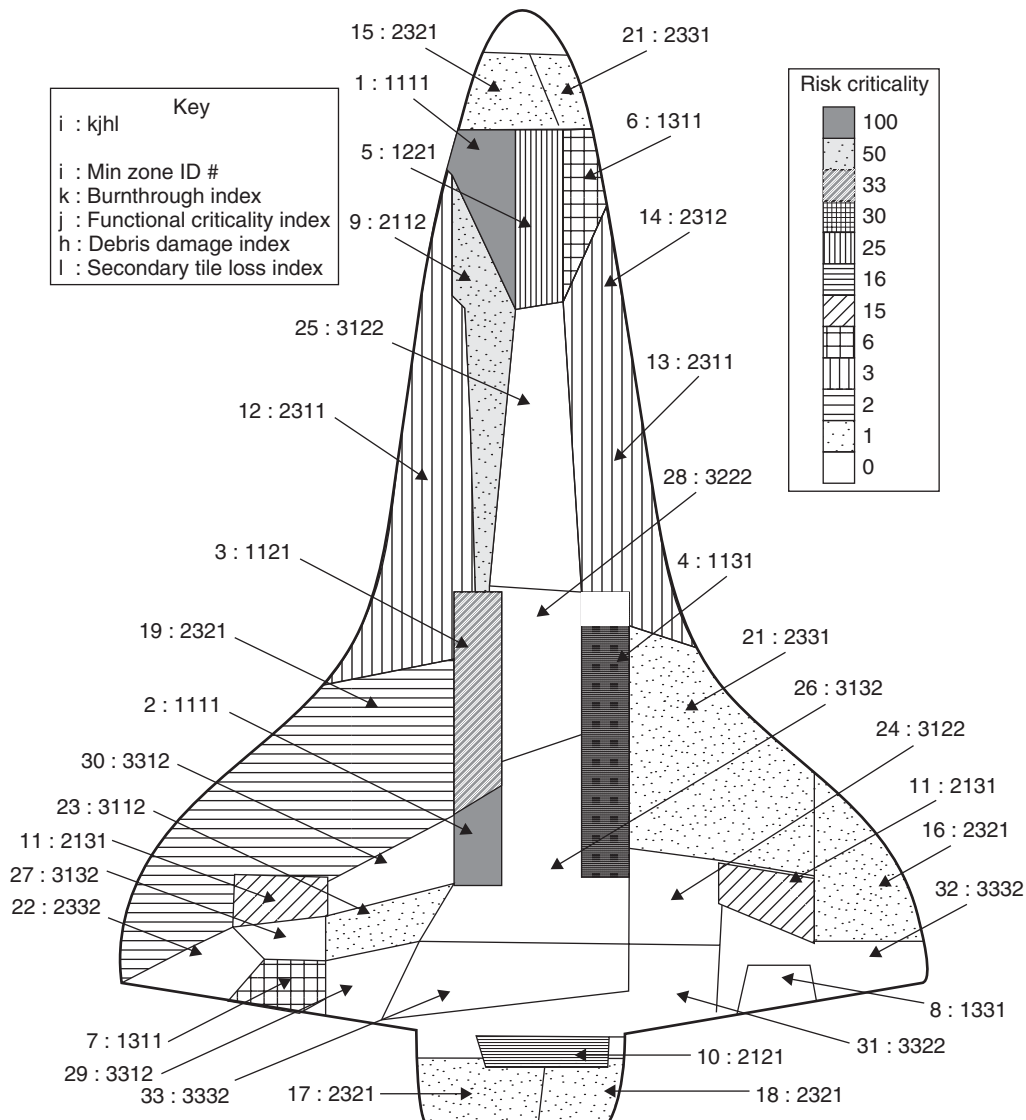


Figure 3. Results of the study of the space shuttle tile failure risks. (Source: From Ref. 43).

then hit the wings of the orbiter (and small ones had hit the tiles before). This was a signal that more attention needed to be given to the insulation of the external tank. This failure mode was identified in as a priority in risk management in spite of the fact that failure of the heat shield had not yet caused an accident (only two tiles had been lost in flight prior to the 1990 study). There was some attempt by NASA to address the problem, but, obviously, the response was

not satisfactory. In 2003, the Columbia space shuttle exploded in flight because a piece of debris from the insulation of the external tank had hit the reinforced carbon-carbon panels that protected the leading edge of a wing.

Human Errors as Accident Precursors: Observation in the PRA Context

Human errors that have been observed without causing an accident can be considered

useful precursors. In operating rooms, they include wrong intubation of the patient (ventilating the stomach instead of the lungs) or wrong dosage of an anesthetic [48]. The question is how close a call it was, what should be the response, and how urgent it is. The problem is to link these human errors to the probabilities of accident sequences. But to derive some risk management conclusions, one then needs to relate human errors to the alertness and competence of the anesthesiologists and to the relevant management factors such as hiring and training. One can then decide what action to take to monitor the performance of the physicians in charge.

Organizational factors are often part of the root causes of industrial accidents and system failures. For example, Grabowski *et al.* [45] propose an approach to developing leading risk indicators in virtual organizations. Identifying accident sequences, precursors, and the effectiveness of risk management measures can be achieved through a probabilistic analysis of the whole system such as the systems–action–management (SAM) model presented in Fig. 4 [46]. The bottom part represents the PRA for the system, for instance the influence diagram of Fig. 2. The level above represents the decisions and actions of the people directly involved, for example the

maintenance technicians in charge of replacing damaged tiles. One error that can appear here is that a technician may decide to take a shortcut in cleaning the cavity before bonding a new tile, thus weakening the bond. In turn, that decision may be influenced by the decision of the management to impose strict deadlines and a daily quota of tiles to be maintained, when a bit more flexibility (e.g., a weekly quota) might avoid the temptation of a shortcut. In another illustration of the SAM model, one can focus on the probability that a ship (e.g., an oil tanker) lose propulsion, hits the ground causing a breach in the hull, and spills oil in the sea. The loss of propulsion itself may be attributable to a problem of design or maintenance, and the grounding to control of the drift by the crew, location, weather, and speed at the time of the incident, and the source term (amount of oil released) to the design of the hull (e.g., single vs double hull). All these factors depend in turn on management decisions, involving resource allocation and personnel management. The SAM model allows connecting these decisions to the risk of an oil spill knowing the ship's routes. A breach in the hull is thus an immediate precursor to the oil spill but the training and the coherence of the crew are thus early signal of their ability to cope with an incident.

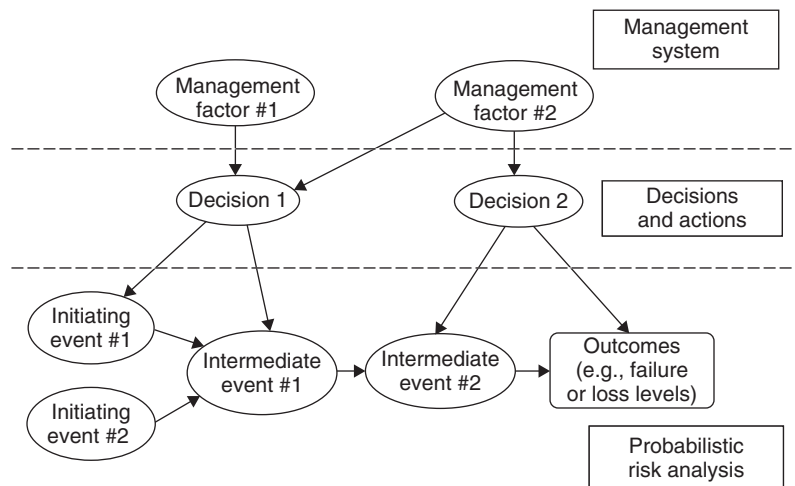


Figure 4. The structure of the systems–actions–management model that links technical PRA to human decisions (and errors) to management decisions.

The SAM model thus allows assessing the probability of a full-fledged accident given a close call from a human error. Furthermore, it allows identifying the management factors that can be improved to decrease the probability that it happens, and to make those choices based on the costs and the risk-reduction benefits of these measures.

Anesthesia Patient Risks

Human errors have been recognized for a longtime as a risk factor in medicine [47] where observing, diagnosing, and acting upon precursors of accident sequences is critical. An example of precursor involving human and organizational factors is found in the case of an analysis of anesthesia patient risk (Fig. 5). That study [48] was based first on statistical data about the precursors, that is, the initiating events that start an accident sequence. We also had statistics about mortality in anesthesia accidents. We needed to model the dynamics of accident scenarios in between to use this model in risk assessment and risk management.

From the probability of the precursors (e.g., of a tube disconnect), one could then compute the probability that the corresponding signal is observed, that the diagnosis of the problem is made and a proper solution is found, and that the patient recover given that oxygen deprivation can cause a serious injury or death within minutes depending on the characteristics of the patient.

This is another case in which the immediate observation of and reaction to precursors is essential in risk management, but also where precursors regarding the incompetence or the alertness problems of a practitioner constitute real warnings that something needs to be addressed before an accident occurs.

Challenges of Linking Occurrences of Signals and a Probabilistic Risk Analysis Model

All signals may not fit neatly and obviously in the framework of the PRA for a given system [see, for example, the chapter on ASP by Sattison [10]]. What is clear is that for the system to fail, one or more of the failure modes must take place. So the general equation linking occurrences of signals to the probabilities of the failure modes and of system failure can be written as

$$p(F|S) = \sum_i p(M_i|S) - \sum_{ij} p(M_i \text{ and } M_j|S) + \sum_{ijk} p(M_i \text{ and } M_j \text{ and } M_k|S) - \dots$$

What may be less clear are the links between the failure modes (and their factors) and the occurrence of a signal ($p(M_i|S)$, $p(M_i, M_j|S)$, ...). For instance, a human error such as the deterioration of the performance of a maintenance crew, which may seem removed from the system's safety, may affect

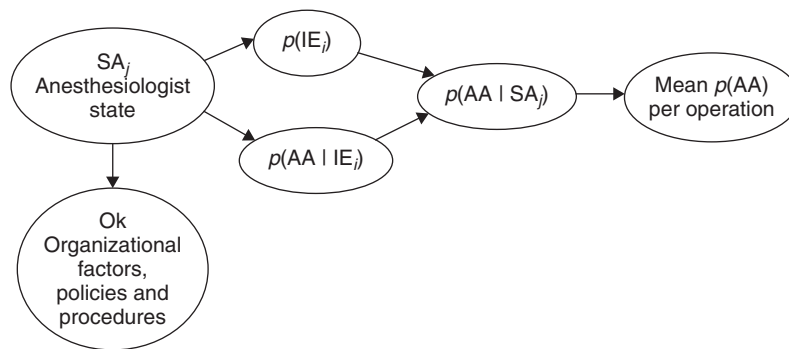


Figure 5. Structure of a dynamic model for the analysis of anesthesia patient risks: SA_j , state of the anesthesiologist (level of alertness and competence); $p(IE_i)$, probability of the initiating event IE_i ; $p(AA|IE_i)$ probability of an anesthesia accident given an initiating event; Ok, organizational factors and risk management measures that influence the state of the anesthesiologist.

a chain of events, resulting in the increase of the probability of failure more than intuition suggests. The importance of a change in external loads, such as a change in the climate that increases the probability of severe storms and may affect several components, may not be immediately recognized. In other cases, a minor deterioration of an external system—for example, a small change in the reliability of the electric grid—may increase the failure risk. One can find many other instances in which the PRA model does not yield immediately the variation of the failure probability given the signal. The problem is generally case-specific and the value of the result depends on the ability of the analyst to identify the existence of the link between the signal and the system's performance, and to gather and use relevant information.

A STRING OF MISSED PRECURSORS AND SIGNALS

Neglecting Repeated Identical Problems because They Have Not Caused a Disaster Yet

One problem when no accident happens in a string of similar near misses is that these events can be considered neglected (and underestimated as survivable) when, in fact, the probability of catastrophic failure may be high and randomness and luck are deceitful.

An example of that phenomenon is the sequence of missed precursors that preceded the crash of the supersonic plane Concorde in 2000. After about 75,000 flights, 57 tires had burst at takeoff without critical damage to the aircraft. The problem was the need for long runways given the shape of the plane, and the danger of heating and bursting given the technology of the tires. Their rubber surface was divided into lozenges. These tires had split with bursts, hitting the under surface of the aircraft, but for a while missed the fuel tanks, until they pierced the fuselage and the tank in an accident that occurred at takeoff in July 2000. By then it should have been clear that the probability of such a disaster was high by aviation safety standards but the fear was that fixing that problem might create another one. It seems that the balance of risks was not properly considered.

This is typical of cases where the evidence is clear and the fact that no accident has happened causes the complacency of the organization. The likelihood of a hit causing a fuel tank rupture could have been computed based on observations of trajectories and patterns, and a crash could have avoided. The argument that is sometimes heard when such luck allows avoiding a disaster for a while can be called the *zillion-mile argument*: we are safe because we have survived these incidents in the past. In cases like this, one can sometimes point to the gap between the safety-first discourse and the reaction to precursors. Because it is true that organizations cannot react to all remotely possible events, an analysis of the probabilities of a disaster can be very useful in interpreting the precursors and their predictive values in probabilistic terms.

Failure to Monitor a System and to Detect a Serious Problem

Some clear signals are not always properly monitored by the organizations responsible for a system. This was the case, for example, in the Ford Firestone fiasco in which the failure of Firestone tires (by tread separation) on Ford vehicles caused many rollovers resulting in more than 250 deaths and 3000 serious injuries. The problem was apparently detected in the end by insurance companies but after much time had been lost. Ford Motor Company then decided to set up a report system that would bring all incidents to the attention of the company's executive. This is one example of what is described below as an "organizational warning system." Such a system, however, has to provide channels to communicate warnings to the appropriate decision maker(s). It also has to include some filtering so that the system is not clogged with false alerts, which are costly not only because of the time and production lost but also because people start losing confidence in the system itself and may stop responding to signals.

Another case where such an organizational warning system is essential is the monitoring of medical devices once they are released on the market—given the nature of these devices (e.g., cardiac), the uncertainty

that remains once they are approved (offlabel uses, unusual patients, vulnerability to the skills of the physicians). Postmarket monitoring needs to be set up so that the Food and Drug Administration is aware as soon as possible of problems of design, manufacturing, or use that require immediate adjustment. These signals were missed for a while in the case of the Guidant drug-eluting stents, which caused multiple accidents before being withdrawn from the market.

FILTERING SIGNALS: MANAGING THE TRADE-OFF BETWEEN TYPE I AND TYPE II ERRORS

One key problem in setting an early-warning system is to manage the trade-off between the probabilities and the consequences of false positives and false negatives, both linked to the sensitivity of the signal when it is observed. Consider here a continuous stochastic process representing the level at a given time of a potentially hazardous phenomenon (for instance, the density of smoke in a room or the level of a river that can overflow). Figure 6 illustrates the case of the risk of a fire with lethal smoke density or a flood from an overflow of a river.

The problem here has multiple dimensions. The question is to know where to set the alert level considering the trade-off between the lead time that the signal provides (the higher the threshold, the shorter the lead time) and the probability of a false alert that has costs in itself (e.g., evacuation) but also decreases the rate of response to the signal in the future. For each possible threshold level (here: $d(\Phi)$), one can compute for the given stochastic model, the upcrossing rate (the rate of alerts) some being false alerts and some true ones that permit reducing the losses given the lead time.

Implementing this method thus requires three types of models. The first one is a model of the underlying phenomenon (how often does it occur, how serious it is, and how fast does it evolve when it is a true alert). The second is a model of the response rate given the past history and the system's performance. It involves the human memory and the "cry-wolf" effect on the response, which obviously depend on the potential severity of the event and the costs of a response. The third model represents the effectiveness of the use of the lead time once the alert threshold has been exceeded, and therefore, the risk-reduction benefits (in probabilistic terms) of the warning system defined by that

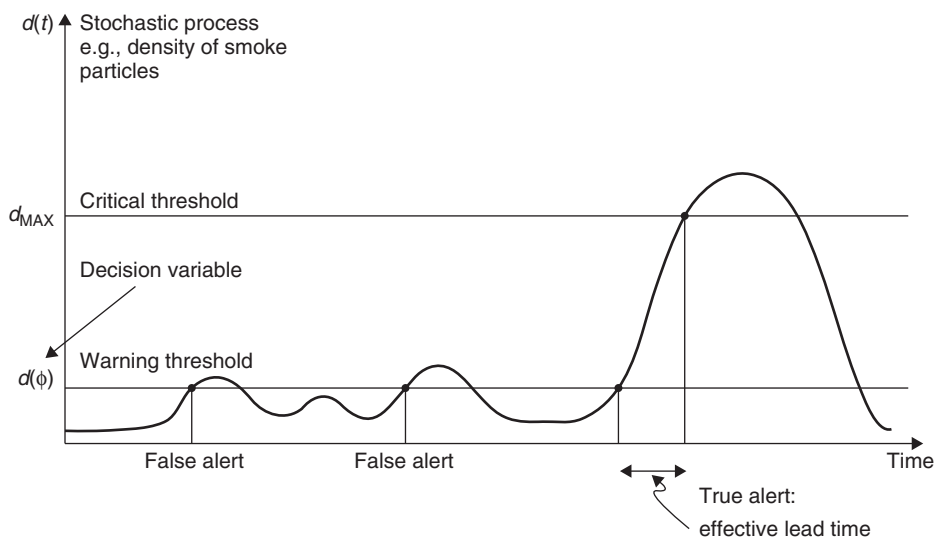


Figure 6. Stochastic process and possible alert thresholds. (Source: From Ref. 20).

alert threshold. Given the benefits of the use of the lead time and the costs of the alert system, one can then compute an optimal threshold that maximizes either a benefit-cost ratio or a utility function, possibly including several attributes (human safety and property damage) and risk attitudes. Note that in this case, another way to proceed is simply to figure out the lead time that is needed (for instance, for the evacuation of a threatened region) and to back-figure the appropriate threshold of alert.

ORGANIZATIONAL WARNING SYSTEMS

Observing a worrisome signal somewhere in an organization is not always sufficient to ensure proper response. This was the case, for instance, of the FBI shortly before 9/11/01, when an agent had observed that potential terrorists were learning how to fly under suspect circumstances and tried to warn the hierarchy above her head, but her message was ignored.

The design of an organizational warning system requires addressing several problems. Communication channels first have to exist or to be set up. The message may get distorted along the communication lines, or simply ignored and blocked at a particular level of the hierarchy. The careful design of an organizational warning system thus requires first deciding what can possibly be the weak points of a system, and to ensure adequate monitoring of the corresponding signals. Second, transmission lines must be set up (sometimes with redundancies) so that the message reaches the appropriate decision maker. But sometimes, filtering has to occur to prevent clogging of the system, and sorting the signals requires a clear examination of the risks involved. The probabilities of false positives and false negatives are essential to a rational interpretation of the information and an assessment of the posterior probability of an event given what has been observed [36].

The organization thus has to ensure that someone will monitor potential problems and set up a communication system with incentives such that the message is communicated to the appropriate level. From a theoretical

point of view, these incentives have to be aligned (e.g., on the basis of a “principal-agent” model), so that the decision maker gets the story straight and that there be no attempt to punish the messenger.

The dynamics of system deterioration and the speed at which events unfold are critical: the organization has to respond on the spot to immediate threats, which sometimes requires a radical switch to a crisis mode. In any case, managing the trade-off between false positives and false negatives requires an implicit or explicit valuation of their probabilities and outcomes. Communicating the message effectively may involve a candid and accurate description of uncertainties, even if it seems to weaken the message.

CONCLUSIONS

Identifying, observing, and communicating precursors and their risk implications are essential parts of risk management. In retrospect, it is often easy after an accident to find that precursors could and should have been observed. In reality, and looking forward, the question is first to make sure that appropriate warning systems are in place and that the message will be adequately communicated. An unavoidable challenge when uncertainties exist is to assess and communicate the chances of possible outcomes. The value of the information provided by any warning lies in its ability to make timely and adequate decisions. There is often an unavoidable trade-off between the potential for false positives and false negatives. Managing this trade-off requires a value system based on risk attitudes, whether in private or public decisions.

REFERENCES

1. Fischhoff B. Hindsight/foresight: the effect of outcome knowledge on judgement under uncertainty. *J Exp Psychol Hum Percept Perform* 1975;1:288–299.
2. Hawkins SA, Hastie R. Hindsight: biased judgments of past events after the outcomes are known. *Psychol Bull* 1990;107:311–327.
3. Lees FP. Hazard warning structure: some illustrative examples based on actual cases. *Reliab Eng Syst Saf* 1985;10(2):65–81.

4. ASRS (Aviation Safety Reporting System). 2001. Available at http://asrs.arc.nasa.gov/overview_nf.htm.
5. FAA (Federal Aviation Administration). The Global Aviation Information Network (GAIN): using information proactively to improve aviation safety. Washington (DC): FAA Office of System Safety; 2002. Available at <http://www.gainweb.org/>.
6. Cottrell WB, Minarick JW, Austin PN, Hagen EW, Harris JD. Precursors to potential severe core damage accidents: 1980–1981, A Status Report. NUREG/CR-3591, July 1984. Washington (DC): U.S. Nuclear Regulatory Commission; 1984.
7. Minarick JW, Kukielka CA. Precursors to potential severe core damage accidents: 1969–1979, A Status Report. NUREG/CR-2497, June 1982. Washington (DC): U.S. Nuclear Regulatory Commission; 1982.
8. USNRC. Precursors to potential severe core damage accidents, A Status Report. NUREG/CR-4674. Washington (DC): U.S. Nuclear Regulatory Commission; 1986 to 1992.
9. Johnson JW, Rasmuson DM. The US NRC's accident sequence precursor program: an overview and development of a Bayesian approach to estimate core damage frequency using precursor information. *Reliab Eng Syst Saf* 1996;53:205–216.
10. Sattison M. Nuclear precursor assessment: the accident sequence precursor program: the accident precursor analysis and management. Proceedings of the National Academy of Engineering Workshop on Precursors. Washington (DC): National Academy Press; 2004. pp. 45–59.
11. Phimister JR, Oktem U, Kleindorfer PR, Kunreuther H. Near miss management systems in the chemical process industry. *Risk Anal* 2003;23(3):445–453.
12. Sonnemans PJ, Körvers PMW. Accidents in the chemical industry: are they foreseeable? *J Loss Prev Process Ind* 2006;19(1):1–12.
13. Zhou J, Chen G, Chen Q. Real-time data-based risk assessment for hazard installations storing flammable gas. *Process Saf Prog* 2008;27(3):205–211.
14. Van der Schaaf TW, Lucas DA, Hale AR. Near miss reporting as a safety tool. Oxford: Butterworth-Heinemann; 1991.
15. BEA. Accident on 25 July 2000 at La Patte d'Oie in Gonesse (95) to the Concorde registered F-BTSC operated by Air France. Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile, Ministère de l'équipement, des transports et du logement, Paris, France; 2002.
16. Bier V, editor. Proceedings of Workshop on accident sequence precursors and probabilistic risk analysis. Center for Reliability Engineering, College Park (MD): University of Maryland; 1998.
17. National Academy of Engineering. Accident precursor analysis and management. In: Bier V, Kunreuther H, Phimister J, editors. Proceedings of the National Academy of Engineering Workshop on Precursors. Washington (DC): National Academy Press; 2004. pp. 45–59.
18. Tamuz M. Understanding accident precursors: the accident precursor analysis and management. Proceedings of the National Academy of Engineering Workshop on Precursors. Washington (DC): National Academy Press; 2004. pp. 45–59.
19. Oktem U, Meel A. Near-Miss management: a participative approach to improving system reliability. In: Melnick E, Everitt B, editors. Encyclopedia of quantitative risk assessment and analysis. Chichester: John Wiley & Sons, Ltd.; 2008. pp. 1154–1163.
20. Paté-Cornell ME. Warning systems in risk management. *Risk Anal* 1986;5(2):223–234.
21. Bier VM, Yi W. The performance of Precursor-based estimators for rare event frequencies. *Reliab Eng Syst Saf* 1995;50:241–251.
22. Cooke R, Bier VM. Simulation results for precursor estimates. In: Bier V, editor. Accident sequence precursors and probabilistic risk analysis. College Park (MD): University of Maryland; 1998. pp. 61–76.
23. de Finetti B. Theory of probability. New York: Wiley; 1974.
24. Yi W, Bier VM. An application of copulas to accident precursor analysis. *Manage Sci* 1998;44:S257–S270.
25. Paté-Cornell E. Fusion of intelligence information: a Bayesian approach. *Risk Anal* 2002;22(3):445–454.
26. Garrick BJ. Perspectives on the use of risk assessment to address terrorism. 2002;22(3): 421–424é.
27. Paté-Cornell ME, Guikema SD. Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures. *Mil Oper Res* 2002;7(4): 5–23.

28. Kaplan S. On the inclusion of precursor and near miss events in quantitative risk assessments: a Bayesian point of view and a space shuttle example. *Reliab Eng Syst Saf* 1990; 27:103–115.
29. Paté-Cornell ME. On signals, response, and risk mitigation: a probabilistic approach to the detection and analysis of precursors: the Accident precursor analysis and management. Proceedings of the National Academy of Engineering Workshop on Precursors. Washington (DC): National Academy Press; 2004. pp. 45–59.
30. Meel A, Seider WD. Plant-specific dynamic failure assessment using Bayesian theory. *Chem Eng Sci* 2006;61:7036–7056.
31. Raiffa H. Decision analysis. Cambridge (MA): Addison Wesley; 1968.
32. Reason J. Managing the risks of organizational accidents. Hants: Ashgate Publishing England; 1997.
33. Carroll JS. Organizational learning activities in high-hazard industries: the logics underlying self-analysis. *J Manag Stud* 1998; 35:699–717.
34. Marcus A, Nichols M. On the Edge: Heading the Warning of Unusual Events. *Organization Science* 1999;10(4):482–499.
35. Weick KE, Sutcliffe KM. Volume 1, Managing the unexpected: assuring high performance in an age of complexity. New York: John Wiley & Sons, Inc.; 2001.
36. Lakats LM, Paté-Cornell ME. Organizational warnings and system safety: a probabilistic analysis. *IEEE Trans Eng Manag* 2004;51(2): 183–196.
37. Vaughan D. The challenger launch decision: risky technology, culture, and deviance at NASA. Chicago (IL): University of Chicago Press; 1997.
38. CIRAS. Confidential incident reporting and analysis system. Glasgow: CIRAS Executive Report; 2003.
39. Howard RA. Value of information lotteries. Volume 2, Readings in the principles and practice of decision analysis. Menlo Park (CA): Strategic Decisions Group; 1984. p. 785.
40. Henley E, Kumamoto H. Probabilistic risk assessment: reliability engineering, design, and analysis. New York: IEEE Press; 1992.
41. Paté-Cornell ME. The engineering risk analysis method and some applications. In: Edwards W, Miles R, von Winterfeldt D, editors. *Advances in decision analysis*. Cambridge (UK): Cambridge University Press; 2007.
42. Paté-Cornell ME, Fischbeck PS. “Safety of the Thermal Protection System of the STS orbiter: quantitative analysis and organizational factors,” Phase 1: “the probabilistic risk analysis model and preliminary observations,” Research Report to NASA, Kennedy Space Center, Dec 1990.
43. Paté-Cornell E, Fischbeck PS. Probabilistic risk analysis and risk-based priority scale for the tiles of the space shuttle. *Reliab Eng Syst Saf* 1993;40:221–238.
44. CAIB. Columbia accident investigation board, Columbia Accident Investigation Board releases working scenario. Houston (TX): Press Release; 2003.
45. Grabowski M, Ayyalasomayajula P, Merrick J, Harrald JR, Roberts K. Leading indicators of safety in virtual organizations. *Saf Sci* 2007;45(10):1013–1043.
46. Murphy DM, Paté-Cornell ME. The SAM framework: a systems analysis approach to modeling the effects of management on human behavior in risk analysis. *Risk Anal* 1996;16(4):501–515.
47. IOM (Institute of Medicine). To Err is human: building a safer health system. In: Kohn LT, Corrigan JM, Donaldson MS, editors. Washington (DC): National Academy Press; 2000.
48. Paté-Cornell ME, Lakats LM, Murphy DM, Gaba DM. Anesthesia patient risk: a quantitative approach to organizational factors and risk management options. *Risk Anal* 1996;17(4):511–523.

ADVANCED BRANCHING PROCESSES

FLORIAN SIMATOS
Eindhoven University of
Technology, Eindhoven,
Netherlands

INTRODUCTION

Branching processes have their roots in the study of the so-called family name extinction problem (see *Introduction to Branching Processes*) and find their most natural and important applications in biology, especially in the study of population dynamics. They were also motivated by the study of nuclear fission reactions and underwent rapid development during the Manhattan project under the impulse of Szilárd and Ulam. To date, they continue to be very important in reactor physics. They also play a major role in (applied) probability at large, and appear in a wide variety of problems in queuing theory, percolation theory, random graphs, statistical mechanics, the analysis of algorithms, and bins and balls problems, to name a few.

The appearance of branching processes in so many contexts has triggered the need for extensions and variations around the classical Galton—Watson branching process. For instance, their application in particle physics provided an impetus to study them in continuous time. The possible extensions are almost endless, and indeed new models of processes exhibiting a branching structure are frequently proposed and studied. Such models allow for instance time and/or space to be continuous, individuals to have one of several types, immigration to take place, catastrophes to happen, individuals to move in space, each individual's dynamic to depend on time, space, the state of the process itself or some exogenous resources, a combination of all these ingredients, and many more.

In this article, we focus more specifically on three advanced models of branching

processes: (i) branching processes in random environment, which are examples of branching processes, where the dynamic evolves (randomly) over time; (ii) branching random walks that exhibit a spatial feature; and (iii) continuous state branching processes (CSBPs) that can be seen as continuous approximations of Galton—Watson processes where both time and space are continuous. The presentation of CSBPs will also be a good place to briefly discuss superprocesses. We each time focus on the most basic properties of these processes, such as the extinction probability or the behavior of extremal particles.

This choice of topics does not aim to be exhaustive and reflects a personal selection of exciting and recent research on branching processes. It leaves out certain important classes of models, which include multitype branching processes, branching processes with immigration, and population-size-dependent branching processes. For multitype branching processes, the book by Mode [1] offers a good starting point. Branching processes with immigration were initially proposed by Heathcote [2,3] as branching models that could have a non-trivial stationary distribution. Lyons *et al.* [4] showed, via change of measure arguments, that they played a crucial role in the study of Galton—Watson processes. Finally, population-size-dependent branching processes, originally proposed by Labkovskii [5], find their motivations in population dynamics: they are elegant models that introduce dependency between individuals and can account for the important biological notion of carrying capacity, see for instance [6–8]. The interested reader can find more results in the extensive survey by Vatutin and Zubkov [9,10] that gathers results up to 1993 as well as in the recent books by Haccou *et al.* [11] and by Kimmel and Axelrod [12].

Before going on, recall (see *Introduction to Branching Processes*) that a Galton—Watson branching process $(Z_n, n \geq 0)$ is an $\mathbb{N}$ -valued Markov chain obeying to the following recursion:

$$Z_{n+1} = \sum_{k=1}^{Z_n} X_{ni}, \quad n = 0, 1, 2, \dots, \quad (1)$$

where the $(X_{ni}, n, i = 0, 1, 2, \dots)$ are independent and identically distributed (i.i.d.) random variables following the so-called *offspring distribution*. A Galton—Watson process is classified according to the value of the mean $m = \mathbb{E}(X_{ni})$ of its offspring distribution. If $m < 1$, the process is subcritical: it dies out almost surely, the survival probability $\mathbb{P}(Z_n > 0)$ decays exponentially fast at speed m^n , and Z_n conditioned on being non-zero converges weakly. If $m = 1$, the process is critical: it dies out almost surely, the survival probability $\mathbb{P}(Z_n > 0)$ decays polynomially fast, and Z_n conditioned on being non-zero grows polynomially fast. Finally, if $m > 1$, the process is supercritical: it may survive forever, and grows exponentially fast in the event $\{\forall n \geq 0 : Z_n > 0\}$ of survival.

BRANCHING PROCESSES IN RANDOM ENVIRONMENT

A first possible generalization of the Galton—Watson model allows for the offspring distribution to vary over time: then, the recursion (1) still holds, the X_{ni} s are still independent but the law of X_{ni} may depend on n . If π_{n+1} is the offspring distribution in generation n and $\Pi = (\pi_n)$, that is, π_{n+1} is the common law of the $(X_{ni}, i = 0, 1, 2, \dots)$ and Π is the environmental process, then this model defines a branching process in varying environment Π . We talk about branching process in random environment when the sequence Π is itself random and independent from Z_0 . Note that in this case, π_n is a random probability distribution on $\mathbb{N}$.

As always in the case of stochastic processes in random environment, one may follow two approaches for their study: (i) the quenched approach, which fixes a realization of the environment and studies the process in it; it is most natural from the point of view of the applications and (ii) the annealed approach, where the various characteristics of interest are calculated by averaging over the environment. For instance, the

extinction probability is a random variable in the quenched approach, and a deterministic number in the annealed approach.

When the environmental process is assumed to be stationary and ergodic, which includes for instance the case of i.i.d. environment or the case where the environment is a stationary Markov chain, it is known since the pioneering works of Smith [13], Smith and Wilkinson [14], and Athreya and Karlin [15,16] that the extinction problem and the description of the asymptotic growth have fairly general solutions. Although in the classical Galton—Watson case, the classification of Z_n is in terms of the mean of the offspring distribution, it is not difficult to see that in the case of random (stationary and ergodic) environment the mean of the logarithm of the mean is the meaningful quantity to look at. More precisely, if π is a probability distribution on $\mathbb{N}$, let $m(\pi) = \sum_y y\pi(\{y\})$ be its mean. Then, by definition (1), we have

$$\mathbb{E}(Z_n | \Pi) = Z_0 m(\pi_1) \cdots m(\pi_n) = Z_0 e^{S_n},$$

where we have defined

$$S_n = \log m(\pi_1) + \cdots + \log m(\pi_n).$$

By the ergodic theorem, we have $S_n/n \rightarrow \mathbb{E}(\log m(\pi_1))$ as $n \rightarrow +\infty$, which implies that $[\mathbb{E}(Z_n | \Pi)]^{1/n} \rightarrow \exp[\mathbb{E}(\log m(\pi_1))]$. In particular, conditionally on the environment, the mean of Z_n goes to 0 if $\mathbb{E}(\log m(\pi_1)) < 0$ and to $+\infty$ if $\mathbb{E}(\log m(\pi_1)) > 0$. This suggests to classify the behavior of Z_n in terms of $\mathbb{E}(\log m(\pi_1))$, and under some mild technical assumptions it holds indeed that Z_n dies out almost surely if $\mathbb{E}(\log m(\pi_1)) \leq 0$ (subcritical and critical cases) and has a positive chance of surviving if $\mathbb{E}(\log m(\pi_1)) > 0$ (supercritical case). More precisely, we have the following quenched result: if $q(\Pi)$ is the (random) extinction probability of Z_n given Π , then $\mathbb{P}(q(\Pi) = 1) = 1$ in the former case and $\mathbb{P}(q(\Pi) < 1) = 1$ in the latter case.

In the supercritical case $\mathbb{E}(\log m(\pi_1)) > 0$, there is an interesting technical condition that is both necessary and sufficient to allow the process to survive with positive probability: namely, in addition to $\mathbb{E}(\log m(\pi_1)) > 0$ one also needs to assume

$\mathbb{E}(-\log(1 - \pi_1(\{0\}))) < +\infty$. This condition shows the interesting interplay that arises between Z_n and the environment: even though $\mathbb{E}(\log m(\pi_1)) > 0$ is sufficient to make the conditional mean of Z_n diverge, if $\mathbb{E}(-\log(1 - \pi_1(\{0\}))) = +\infty$ then the process almost surely dies out because the probability of having an unfavorable environment is large, where by unfavorable environment we mean an environment π where the (random) probability $\pi(\{0\})$ of having no offspring is close to 1. In other words, if $\mathbb{E}(-\log(1 - \pi_1(\{0\}))) = +\infty$ then the process gets almost surely extinct because of the wide fluctuation of the environment.

The classification of Z_n into the subcritical, critical, and supercritical cases also corresponds to different asymptotic behaviors of Z_n conditioned on non-extinction (here again, we have the quenched results of Athreya and Karlin [15] in mind). In that respect, Z_n shares many similarities with a Galton—Watson process, although there are some subtle differences as we see at the end of this section. In the supercritical case, Z_n grows exponentially fast in the event of non-extinction, whereas in the subcritical case, Z_n conditioned on being non-zero converges weakly to a non-degenerate random variable. In the critical case, Z_n conditioned on being non-zero converges weakly to $+\infty$, a result that can be refined in the case of i.i.d. environment.

Indeed, the case where the (π_i) are i.i.d. has been extensively studied. In this case, S_n is a random walk and recent works have highlighted the intimate relation between Z_n and S_n . In particular, the classification of Z_n can be generalized as follows. It is known from random walk theory that, when one excludes the trivial case where $S_n = S_0$ for every n , there are only three possibilities concerning the almost sure behavior of (S_n) : either it drifts to $-\infty$, or it oscillates with $\liminf_n S_n = -\infty$ and $\limsup_n S_n = +\infty$, or it drifts to $+\infty$. Then, without assuming that the mean $\mathbb{E}(\log m(\pi_1))$ exists, Z_n can be said to be subcritical if $S_n \rightarrow -\infty$, critical if S_n oscillates, and supercritical if $S_n \rightarrow +\infty$. Within this terminology, Afanasyev *et al.* [17] studied the critical case and were able to obtain striking results linking the behavior of Z_n to

the behavior of its associated random walk. In particular, this work emphasized the major role played by fluctuation theory of random walks in the study of branching processes in random (i.i.d.) environment, a line of thought that has been very active since then.

Let us illustrate this idea with some of the results of Afanasyev *et al.* [17], so consider Z_n a critical branching process in random environment. As Z_n is absorbed at 0, we have $\mathbb{P}(Z_n > 0 \mid \Pi) \leq \mathbb{P}(Z_m > 0 \mid \Pi)$ for any $m \leq n$ and as Z_n is integer-valued, we obtain $\mathbb{P}(Z_n > 0 \mid \Pi) \leq \mathbb{E}(Z_m \mid \Pi)$. It follows that

$$\mathbb{P}(Z_n > 0 \mid \Pi) \leq Z_0 \exp \left(\min_{0 \leq m \leq n} S_m \right),$$

which gives an upper bound, in term of the infimum process of the random walk S_n , on the decay rate of the extinction probability in the quenched approach. It turns out that this upper bound is essentially correct, and that the infimum also leads to the correct decay rate of the extinction probability in the annealed approach, although in a different form. Indeed, it can be shown under fairly general assumptions that

$$\mathbb{P}(Z_n > 0) \sim \theta \mathbb{P}(\min(S_1, \dots, S_n) > 0) \quad (2)$$

for some $\theta \in (0, \infty)$. Moreover, conditionally on $\{Z_n > 0\}$, Z_n/e^{S_n} converges weakly to a random variable W , almost surely finite and strictly positive, showing that S_n essentially governs the growth rate of Z_n . Finally, although it is natural to consider the growth rate and extinction probability of the process Z_n , one can also reverse the viewpoint and study the kind of environment that makes the process survive for a long time. And actually, the conditioning $\{Z_n > 0\}$ has a strong impact on the environment: although S_n oscillates, conditionally on $\{Z_n > 0\}$ the process $(S_k, 0 \leq k \leq n)$ suitably rescaled can be shown to converge to the meander of a Lévy process, informally, a Lévy process conditioned on staying positive. This provides another illustration of the richness of this class of models, where the interplay between the environment and the process leads to very interesting behavior.

These various results concern the annealed approach: Equation (2) is for

instance obtained by averaging over the environment. However, the connection between Z_n and S_n continues to hold in the quenched approach. In Ref. 18, it is for instance shown that Z_n passes through a number of bottlenecks at the moments close to the sequential points of minima in the associated random walk. More precisely, if $\tau(n) = \min\{k \geq 0 : S_j \geq S_k, j = 0, \dots, n\}$ is the leftmost point of the interval $[0, n]$ at which the minimal value of $(S_j, j = 0, \dots, n)$ is attained, $Z_{\tau(n)}$ conditionally on the environment and on $\{Z_n > 0\}$ converges weakly to a finite random variable. For further reading on this topic, the reader is referred to Refs 19 and 20.

Let us conclude this section by completing the classification of branching processes in random environment. We have mentioned that similarly as Galton—Watson processes, branching processes in random environment could be classified as subcritical, critical, or supercritical according to whether $\mathbb{E}(Y) < 0$, $\mathbb{E}(Y) = 0$, or $\mathbb{E}(Y) > 0$, with $Y = \log m(\pi_1)$ (in the “simple” case where Y is indeed integrable). Interestingly, assuming that $\mathbb{E}(e^Y)$ is finite for every $t \geq 0$, the subcritical phase can be further subdivided, according to whether $\mathbb{E}(Ye^Y) > 0$, $\mathbb{E}(Ye^Y) = 0$, or $\mathbb{E}(Ye^Y) < 0$ corresponding respectively, in the terminology of Birkner *et al.* [21], to the weakly subcritical, intermediate subcritical, and strongly subcritical cases. These three cases correspond to different speeds of extinction: in the weakly subcritical case, there exists $\beta \in (0, 1)$ such that $\mathbb{E}(Ye^{\beta Y}) = 0$ and $\mathbb{P}(Z_n > 0)$ decays like $n^{-3/2}[\mathbb{E}(e^{\beta Y})]^n$; in the intermediate subcritical case, $\mathbb{P}(Z_n > 0)$ decays like $n^{-1/2}[\mathbb{E}(e^Y)]^n$; finally, in the strongly subcritical case, $\mathbb{P}(Z_n > 0)$ decays like $[\mathbb{E}(e^Y)]^n$. These decay rates are to be compared to the classical Galton—Watson process, wherein the subcritical case $\mathbb{P}(Z_n > 0)$ decays like m^n , corresponding to the strongly subcritical case, because when Y is determinist we have the relation $m = \mathbb{E}e^Y$.

Further reading on branching processes in random environment includes, for example, Refs 22 and 23 for the study of the subcritical case using the annealed approach, while the trajectories of Z_n under various conditionings, namely dying at a distant given moment

and attaining a high level, have been studied in Refs 24,25 and 26,27, respectively. In Ref. 28, the survival probability of the critical multitype branching process in a Markovian random environment is investigated.

BRANCHING RANDOM WALKS

Branching random walks are extension of Galton—Watson processes that, in addition to the genealogical structure, add a spatial component to the model. Each individual has a location, say for simplicity on the real line. Typically, each individual begets a random number of offspring, as in a regular branching process, and the positions of these children form a point process centered around the location of the parent. For instance, if $B = \sum_{x \in B} \delta_x$ is the law of the point process governing the locations of the offspring of a given individual, with δ_x the Dirac measure at $x \in \mathbb{R}$, the locations of the children of an individual located at y are given by the atoms of the measure $\sum_{x \in B} \delta_{y+x}$. Branching random walks can therefore naturally be seen as measure-valued Markov processes, which will turn out to be the right point of view when discussing superprocesses. Another viewpoint is to see branching random walks as random labeled trees: the tree itself represents the genealogical structure, whereas the label on an edge represents the displacement of a child with respect to its parent. Nodes of the tree then naturally inherit labels recursively, where the root is assigned any label, and the label of a node that is not the root is given by the label of its parent plus the label on the corresponding edge.

There is an interesting connection between branching random walks and (general) branching processes. In the case where the atoms of B are in $(0, \infty)$, particles of the branching random walk live on the positive half-line and their positions can therefore be interpreted as the time at which the corresponding particle is born. Keeping track of the filiation between particles, we see that within this interpretation, each particle gives birth at times given by the atoms of the random measure B . This is exactly the model of general, or Crump—Mode—Jagers,

branching processes (see *Introduction to Branching Processes*).

One of the most studied questions related to branching random walks concerns the long-term behavior of extremal particles. Of course, as the branching random walk is absorbed when there are no more particles, this question only makes sense when the underlying Galton—Watson process is supercritical and conditioned on surviving. Let for instance M_n be the location of the leftmost particle in the n th generation, that is, the smallest label among the labels of all nodes at depth n in the tree. Assume for simplicity that each individual has two children with i.i.d. displacements, say with distribution D . Then by construction, a typical line of descent (i.e., the labels on the successive nodes on an infinite path from the root) is equal in distribution to a random walk with step distribution D , thus drifting to $+\infty$ if $\mathbb{E}D > 0$. However, M_n is then equal in distribution to the minimum between 2^n random walks, and although a typical line of descent goes to $+\infty$, the exponential explosion in the number of particles makes it possible for the minimal displacement M_n to follow an atypical trajectory and, say, diverge to $-\infty$. Finer results are even available, and the speed at which M_n diverges has been initiated in a classical work by Hammersley [29] (who was interested in general branching processes), and later extended by Kingman [30] and Biggins [31] leading to what is now commonly referred to as the *Hammersley—Kingman—Biggins theorem*. For instance, in the simple case with binary offspring and i.i.d. displacements, it can be shown that $M_n \rightarrow -\infty$ if $\inf_{\theta \geq 0} \mathbb{E}(e^{-\theta D}) > 1/2$ and simple computations even give a precise idea of the speed at which this happens. Indeed, if $S_n^{(k)}$ for $k = 1, \dots, 2^n$ are the 2^n labels of the nodes at depth n in the tree, we have by definition for any $a \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}(M_n \leq an) &= \mathbb{P}\left(S_n^{(k)} \leq an \text{ for some } k \leq 2^n\right) \\ &\leq \sum_{1 \leq k \leq 2^n} \mathbb{P}\left(S_n^{(k)} \leq an\right) \end{aligned}$$

using the union bound for the last inequality. As the $S_n^{(k)}$'s are identically distributed, say

with common distribution S_n equal to the value at time n of a random walk with step distribution D , we obtain for any $\theta \geq 0$

$$\begin{aligned} \mathbb{P}(M_n \leq an) &\leq 2^n \mathbb{P}(S_n \leq an) \\ &\leq 2^n \left[e^{\theta an} \mathbb{E}(e^{-\theta S_n}) \right] \leq [2\mu(a)]^n \end{aligned}$$

using Markov inequality for the second inequality, and defining $\mu(a)$ as $\mu(a) = \inf_{\theta \geq 0} (e^{\theta a} \mathbb{E}(e^{-\theta D}))$ in the last term. In particular, $\mathbb{P}(M_n/n \leq a) \rightarrow 0$ if a is such that $\mu(a) < 1/2$, which makes $M_n/n \rightarrow \gamma$, with $\gamma = \inf\{a : \mu(a) > 1/2\}$, the best we could hope for. It is quite surprising that these simple computations lead to the right answer, but the almost sure convergence $M_n/n \rightarrow \gamma$ is indeed the content of the aforementioned Hammersley—Kingman—Biggins theorem.

The case $\gamma = 0$ can be seen as a critical case, where the speed is sublinear; for a large class of branching random walks, this case also corresponds, after some renormalization (typically, centering the branching random walk), to study the second-order asymptotic behavior of M_n for a general γ . In the case $\gamma = 0$, several asymptotic behaviors are possible and the reader can for instance consult Addario-Berry and Reed [32] for more details. Bramson [33] proved that if every particle gives rise to exactly two particles and the displacement takes the value 0 or 1 with equal probability, then $M_n - \log \log n / \log 2$ converges almost surely. Recently, Aïdékon [34] proved in the so-called boundary case that $M_n - (3/2) \log n$ converges weakly.

These results concern the behavior of the extremal particle, and there has recently been an intense activity to describe the asymptotic behavior of all extremal particles, that is, the largest one, together with the second largest one, and third largest one. Informally, one is interested in the behavior of the branching random walk “seen from its tip,” which technically amounts to consider the point process recording the distances from every particle to the extremal one. This question was recently solved by Madaule [35], building on previous results by different authors, in particular the aforementioned work by Aïdékon [34]. The limiting point process can be seen as a “colored” Poisson process, informally obtained by attaching to each

atom of some Poisson process independent copies of some other point process.

Initially, one of the main motivations for studying the extremal particle of a branching random walk comes from a connection with the theory of partial differential equations. Namely, one can consider a variation of the branching random walk model, called the *branching Brownian motion*. In this model, time and space are continuous; each particle lives for a random duration, exponentially distributed, during which it performs a Brownian motion, and is replaced on death by k particles with probability p_k . Then, McKean [36] and later Bramson [37] observed that if $\sum_k k p_k = 2$ and $\sum_k k^2 p_k < +\infty$, then the function $u(t, x) = \mathbb{P}(M(t) > x)$, with $M(t)$ now the maximal displacement of the branching Brownian motion at time t , that is, the location of the rightmost particle, is a solution to the so-called Kolmogorov—Petrovskii—Piskunov (KPP) equation, which reads

$$\frac{\partial u}{\partial t} = \frac{1}{2} \frac{\partial^2 u}{\partial x^2} + \sum_{k \geq 1} p_k u^k - u$$

with the initial condition $u(0, x) = 1$ if $x \geq 0$ and $u(0, x) = 0$ if $x < 0$. One of the key properties of the KPP equation is that it admits traveling waves: there exists a unique solution satisfying

$$u(t, m(t) + x) \rightarrow w(x) \text{ uniformly in } x \text{ as } t \rightarrow +\infty.$$

Using the connection with the branching Brownian motion, Bramson [37] was able to derive extremely precise results on the position of the traveling wave, and essentially proved that $m(t) = \sqrt{2}t - (3/2^{3/2})\log t$. In probabilistic terms, this means that $M(t) - \sqrt{2}t + (3/2^{3/2})\log t$ converges weakly. Similarly as for the branching random walk, there has recently been an intense activity to describe the branching Brownian motion seen from its tip, which culminated in the recent works by Arguin *et al.* [38] and Aïdékon [34].

Beyond the behavior of extremal particles, the dependency of the branching random walk on the space dimension has been

investigated in Refs 39–45. There are also a number of articles for the so-called catalytic random walk when the particles performing random walk on $\mathbb{Z}^d$ reproduce at the origin only [42, 46, 47] for which a wide range of phenomena has been investigated. For this and more, the reader can for instance consult the recent survey by Bertacchi and Zuccha [48].

CONTINUOUS STATE BRANCHING PROCESSES

From a modeling standpoint, it is natural in the context of large populations to wonder about branching processes in continuous time and with a continuous state space. In the same vein, Brownian motion (or, more generally, a Lévy process) approximates a random walk evolving on a long time scale. The definition (1) does not easily lend itself to such a generalization.

An alternative and, from this perspective, more suitable characterization of Galton—Watson processes is through the branching property. If Z^y denotes a Galton—Watson process started with $y \in \mathbb{N}$ individuals, the family of processes $(Z^y, y \in \mathbb{N})$ is such that

$$Z^{y+z} \stackrel{(d)}{=} Z^y + \tilde{Z}^z, y, z \in \mathbb{N}, \quad (3)$$

where $\stackrel{(d)}{=}$ means equality in distribution and $\tilde{Z}^z$ is a copy of Z^z , independent from Z^y . In words, a Galton—Watson process with offspring distribution X started with $y + z$ individuals is stochastically equivalent to the sum of two independent Galton—Watson processes, both with offspring distribution X , and where one starts with y individuals and the other with z . It can actually be shown that this property uniquely characterizes Galton—Watson processes, and Lamperti [49] uses this characterization as the definition of a continuous state branching process (CSBP). Formally, CSBPs are the only time-homogeneous Markov processes (in particular, in continuous time) with state space $[0, \infty]$ that satisfy the branching property, see also Ikeda *et al.* [50] for more general state spaces. Note that even in the case of real-valued branching processes, the

state space includes $+\infty$: in contrast with Galton—Watson branching processes, CSBP can in principle explode in finite time.

One of the achievements of the theory is the complete classification of CSBPs, the main result being a one-to-one correspondence between CSBPs and Lévy processes with no negative jumps and, in full generality, possibly killed after an exponential time. This result has a long history dating back from Lamperti [49], for which the reader can find more details in the introduction of Caballero *et al.* [51] (note that CSBPs were first considered by Jiřina [52]). There are two classical ways to see this result. Until further notice, let $Z = (Z(t), t \geq 0)$ be a CSBP started at $Z(0) = 1$.

The first one is through random time-change manipulations, and more specifically through the Lamperti transformation $\mathcal{L}$ that acts on positive functions as follows. If $f : [0, \infty) \rightarrow [0, \infty)$, then $\mathcal{L}(f)$ is defined implicitly by $\mathcal{L}(f)(\int_0^t f) = f(t)$ for $t \geq 0$, and explicitly by $\mathcal{L}(f) = f \circ \kappa$ with $\kappa(t) = \inf\{u \geq 0 : \int_0^u f > t\}$. Then, it can be proved that $\mathcal{L}(Z)$ is a Lévy process with no negative jumps, stopped at 0 and possibly killed after an exponential time; conversely, if Y is such a Lévy process, then $\mathcal{L}^{-1}(Y)$ is (well defined and is) a CSBP.

The second way is more analytical. Let $u(t, \lambda) = -\log \mathbb{E}(e^{-\lambda Z(t)})$: then u satisfies the semigroup property $u(s + t, \lambda) = u(s, u(t, \lambda))$, which leads, for small $h > 0$, to the approximation

$$\begin{aligned} u(t + h, \lambda) - u(t, \lambda) &= u(h, u(t, \lambda)) - u(0, u(t, \lambda)) \\ &\approx h \frac{\partial u}{\partial t}(0, u(t, \lambda)) = -h \Psi(u(t, \lambda)) \end{aligned}$$

once one defines $\Psi(\lambda) = -\frac{\partial u}{\partial t}(0, \lambda)$. It can indeed be shown that u satisfies the so-called branching equation

$$\frac{\partial u}{\partial t} = -\Psi(u), \quad (4)$$

with boundary condition $u(0, \lambda) = \lambda$. In particular, Ψ uniquely characterizes u , and thus Z , and it is called the *branching mechanism* of Z . Moreover, it can be proved that Ψ is a Lévy exponent, that is, there exists a Lévy process Y with no negative jumps such

that $\Psi(\lambda) = -\log \mathbb{E}(e^{-\lambda Y(1)})$ or equivalently, in view of the Lévy—Khintchine formula, Ψ is of the form

$$\begin{aligned} \Psi(\lambda) &= \varepsilon + \alpha\lambda - \frac{1}{2}\beta\lambda^2 \\ &\quad + \int_{(0, \infty)} (1 - e^{-\lambda x} - \lambda x \mathbf{1}_{\{x < 1\}}) \pi(dx) \end{aligned}$$

for some measure π satisfying $\int_{(0, \infty)} (1 \wedge x^2) \pi(dx) < +\infty$ and some parameters $\varepsilon, \beta \geq 0$, and $\alpha \in \mathbb{R}$. Conversely, any Ψ of this form is the branching mechanism of a CSBP, and as Lévy processes are characterized by their Lévy exponent, this provides an alternative view on the one-to-one correspondence between CSBPs and Lévy processes.

The long-term behavior of CSBPs was one of the earliest questions studied by Grey [53], and they exhibit a richer range of behavior than Galton—Watson processes. First, Z is conservative, that is, $\mathbb{P}(Z(t) < +\infty) = 1$ for every $t \geq 0$, if and only if $1/\Psi$ is integrable at $0+$, a sufficient condition for this being that $\Psi(0) = 0$ and $\Psi'(0) > -\infty$. Further, we observe by differentiating the branching equation (4) with respect to λ and using $\mathbb{E}(Z(t)) = \frac{\partial u}{\partial \lambda}(t, 0)$ that $\mathbb{E}(Z(t)) = \exp(-\Psi'(0)t)$, which suggests to classify a (conservative) CSBP as subcritical, critical, or supercritical according to whether $\Psi'(0) > 0$, $\Psi'(0) = 0$, or $\Psi'(0) < 0$, respectively. This classification turns out to be essentially correct for conservative processes, under the additional requirement $\beta > 0$: in this case, supercritical processes may survive forever, with probability $e^{-\lambda_0}$ where λ_0 is the largest root of the equation $\Psi(\lambda) = 0$, while critical and subcritical processes die out almost surely, that is, the time $\inf\{t \geq 0 : Z(t) = 0\}$ is almost surely finite.

When $\beta = 0$, the situation may be slightly different. indeed, for any CSBP, the extinction probability $\mathbb{P}(\exists t \geq 0 : Z(t) = 0)$ is strictly positive if and only if $1/\Psi$ is integrable at $+\infty$ and $\Psi(\lambda) > 0$ for λ large enough; in this case, the extinction probability is equal to $e^{-\lambda_0}$ with λ_0 as discussed earlier. In particular, we may have a subcritical CSBP (with $\Psi'(0) > 0$) satisfying both $\beta = 0$ and $\int^\infty (1/\Psi) = +\infty$, in which case $Z(t) \rightarrow 0$ but $Z(t) > 0$ for every $t \geq 0$. In other words, although Z vanishes,

in the absence of the stochastic fluctuations induced by β , it never hits 0. This behavior is to some extent quite natural, because the α term corresponds to a deterministic exponential decay (for $\Psi(\lambda) = \alpha\lambda$ we have $Z(t) = e^{-\alpha t}$) and the jumps of Z are only positive, so one needs stochastic fluctuations in order to make Z hit 0.

We have mentioned in the beginning of this section the motivation for studying CSBPs as continuous approximations of Galton—Watson processes. This line of thought is actually present in one of the earliest articles by Lamperti [54] on the subject. In particular, CSBPs are the only possible scaling limits of Galton—Watson processes, that is, if $(Z^{(n)}, n \geq 1)$ is a sequence of Galton—Watson processes with $Z_0^{(n)} = n$ such that the sequence of rescaled processes $(\bar{Z}^{(n)}, n \geq 1)$, where $\bar{Z}^{(n)}(t) = Z_{[ant]}^{(n)}/n$ for some normalizing sequence a_n , converges weakly to some limiting process $\bar{Z}$, then $\bar{Z}$ must be a CSBP. And conversely, any CSBP can be realized in this way.

There is, at least informally, an easy way to see this result, by extending the Lamperti transformation at the discrete level of Galton—Watson processes. Indeed, consider $(S(k), k \geq 0)$ a random walk with step distribution $X' = X - 1$ for some integer-valued random variable X , and define recursively $Z_0 = S(0)$ and $Z_{n+1} = S(0) + S(Z_1 + \dots + Z_n)$ for $n \geq 0$. Then, writing $S(k) = S(0) + X'_1 + \dots + X'_k$ with (X'_k) i.i.d. copies of X' , we have

$$\begin{aligned} Z_{n+1} - Z_n &= X'_{Z_1 + \dots + Z_{n-1} + 1} \\ &\quad + \dots + X'_{Z_1 + \dots + Z_{n-1} + Z_n} \end{aligned}$$

and so Z_{n+1} is the sum of Z_n i.i.d. copies of $X' + 1$, that is, Z_n is a branching process with offspring distribution X . This realizes Z as the time-change of a random walk, and leveraging on classical results on the convergence of random walks toward Lévy processes and continuity properties of the time-change involved [55], one can prove that the limit of any sequence of suitably renormalized Galton—Watson processes must be the time-change of a Lévy process, that is, a CSBP. This approach is for instance carried on by Ethier and Kurtz [56].

If a CSBP can be viewed as a continuous approximation of a Galton—Watson process, it is natural to ask about the existence of a corresponding genealogical structure. This question was answered by Duquesne and Le Gall [57], who for each CSBP Z exhibited a process H , which they call height process, such that Z is the local time process of H . This question is intrinsically linked to the study of continuum random trees initiated by Aldous [58,59]. As a side remark, note that this genealogical construction plays a key role in the construction of the Brownian snake [60]. There has also been considerable interest in CSBP allowing immigration of new individuals: these processes were defined by Kawazu and Watanabe [61], their genealogy studied by Lambert [62] and the corresponding continuum random trees by Duquesne [63].

Finally, let us conclude this section on CSBPs by mentioning superprocesses. Superprocesses are the continuous approximations of branching random walks, in the same vein as CSBPs are continuous approximations of Galton—Watson processes. They were constructed by Watanabe [64], and can technically be described as measure-valued Markov processes. Similarly as for the branching Brownian motion, Dynkin [65] showed that superprocesses are deeply connected to partial differential equations. The recent book by Li [66] offers a nice account on this topic.

REFERENCES

1. Mode CJ. Multitype branching processes. Theory and applications, Modern Analytic and Computational Methods in Science and Mathematics, No. 34. New York: American Elsevier Publishing Co., Inc.; 1971.
2. Heathcote CR. A branching process allowing immigration. *J R Stat Soc Ser B* 1965;27:138–143.
3. Heathcote CR. Corrections and comments on the paper “A branching process allowing immigration”. *J R Stat Soc Ser B* 1966;28:213–217.
4. Lyons R, Pemantle R, Peres Y. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann Probab* 1995;23(3):1125–1138.

5. Labkovskii V. A limit theorem for generalized random branching processes depending on the size of the population. *Theory Probab Appl* 1972;17(1): 72–85.
6. Jagers P. Population-size-dependent branching processes. *J Appl Math Stochast Anal* 1996;9(4): 449–457.
7. Jagers P, Klebaner FC. Population-size-dependent, age-structured branching processes linger around their carrying capacity. *J Appl Probab* 2011;48A (New frontiers in applied probability: a Festschrift for Søren Asmussen): 249–260.
8. Klebaner FC. On population-size-dependent branching processes. *Adv Appl Probab* 1984;16(1): 30–55.
9. Vatutin VA, Zubkov AM. Branching processes. I. *J Math Sci* 1987;39:2431–2475. DOI: 10.1007/BF01086176.
10. Vatutin VA, Zubkov AM. Branching processes. II. *J Math Sci* 1993;67:3407–3485. DOI: 10.1007/BF01096272.
11. Haccou P, Jagers P, Vatutin VA. Branching processes: variation, growth, and extinction of populations, Cambridge Studies in Adaptive Dynamics. Cambridge: Cambridge University Press; 2007.
12. Kimmel M, Axelrod DE. Branching processes in biology. Volume 19, Interdisciplinary Applied Mathematics. New York: Springer-Verlag; 2002.
13. Smith WL. Necessary conditions for almost sure extinction of a branching process with random environment. *Ann Math Stat* 1968;39:2136–2140.
14. Smith WL, Wilkinson WE. On branching processes in random environments. *Ann Math Stat* 1969;40:814–827.
15. Athreya KB, Karlin S. Branching processes with random environments. II. Limit theorems. *Ann Math Stat* 1971;42: 1843–1858.
16. Athreya KB, Karlin S. On branching processes with random environments. I. Extinction probabilities. *Ann Math Stat* 1971;42:1499–1520.
17. Afanasyev VI, Geiger J, Kersting G, *et al.* Criticality for branching processes in random environment. *Ann Probab* 2005;33(2): 645–673.
18. Vatutin V, Dyakonova E. Branching processes in a random environment and bottlenecks in the evolution of populations. *Theory Probab Appl* 2007;51(1): 189–210.
19. Vatutin V, Dyakonova E. Galton-Watson branching processes in a random environment I: limit theorems. *Theory Probab Appl* 2004;48(2): 314–336.
20. Vatutin V, Dyakonova E. Galton-Watson branching processes in a random environment. II: Finite-dimensional distributions. *Theory Probab Appl* 2005;49(2): 275–309.
21. Birkner M, Geiger J, Kersting G. Branching processes in random environment—a view on critical and subcritical cases. *Interacting stochastic systems*. Berlin: Springer; 2005. p 269–291.
22. Afanasyev VI, Böinghoff C, Kersting G, *et al.* Limit theorems for weakly subcritical branching processes in random environment. *J Theor Probab* 2012;25(3): 703–732.
23. Afanasyev V, Böinghoff C, Kersting G, *et al.* Conditional limit theorems for intermediately subcritical branching processes in random environment. To appear in *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*.
24. Böinghoff C, Dyakonova EE, Kersting G, *et al.* Branching processes in random environment which extinct at a given moment. *Markov Process Relat Fields* 2010;16(2): 329–350.
25. Vatutin V, Wachtel V. Sudden extinction of a critical branching process in a random environment. *Theory Probab Appl* 2010;54(3): 466–484.
26. Afanasyev V. Brownian high jump. *Theory Probab Appl* 2011;55(2): 183–197.
27. Afanasyev V. Invariance principle for a critical Galton-Watson process attaining a high level. *Theory Probab Appl* 2011;55(4): 559–574.
28. Dyakonova E. Multitype Galton-Watson branching processes in Markovian random environment. *Theory Probab Appl* 2012;56(3): 508–517.
29. Hammersley JM. Postulates for subadditive processes. *Ann Probab* 1974;2:652–680.
30. Kingman JFC. The first birth problem for an age-dependent branching process. *Ann Probab* 1975;3(5): 790–801.
31. Biggins JD. The first- and last-birth problems for a multitype age-dependent branching process. *Adv Appl Probab* 1976;8(3): 446–459.
32. Addario-Berry L, Reed B. Minima in branching random walks. *Ann Probab* 2009;37(3): 1044–1079.

33. Bramson MD. Minimal displacement of branching random walk. *Z Wahrsch Verw Gebiete* 1978;45(2): 89–108.
34. Aïdékon E. Convergence in law of the minimum of a branching random walk. *Annals of Probability*. 2013;41(3A):1362–1426.
35. Madaule T. Convergence in law for the branching random walk seen from its tip 2011. arXiv:1107.2543.
36. McKean HP. Application of Brownian motion to the equation of Kolmogorov–Petrovskii–Piskunov. *Commun Pure Appl Math* 1975;28(3): 323–331.
37. Bramson MD. Maximal displacement of branching Brownian motion. *Commun Pure Appl Math* 1978;31(5): 531–581.
38. Arguin L-P, Bovier A, Kistler N. The extremal process of branching Brownian motion. *Probab Theory Relat Fields* 2012. DOI: 10.1007/s00440-012-0464-x.
39. Bramson M, Cox JT, Greven A. Ergodicity of critical spatial branching processes in low dimensions. *Ann Probab* 1993;21(4): 1946–1957.
40. Bramson M, Cox JT, Greven A. Invariant measures of critical spatial branching processes in high dimensions. *Ann Probab* 1997;25(1): 56–70.
41. Cox JT, Greven A. On the long term behavior of some finite particle systems. *Probab Theory Relat Fields* 1990;85(2): 195–237.
42. Fleischmann K, Vatutin V, Wakolbinger A. Branching systems with long-living particles at the critical dimension. *Theory Probab Appl* 2003;47(3): 429–454.
43. Fleischmann K, Vatutin VA. An integral test for a critical multitype spatially homogeneous branching particle process and a related reaction-diffusion system. *Probab Theory Relat Fields* 2000;116(4): 545–572.
44. Klenke A. Different clustering regimes in systems of hierarchically interacting diffusions. *Ann Probab* 1996;24(2): 660–697.
45. Klenke A. Clustering and invariant measures for spatial branching models with infinite variance. *Ann Probab* 1998;26(3): 1057–1087.
46. Alberverio S, Bogachev LV. Branching random walk in a catalytic medium. I. Basic equations. *Positivity* 2000;4(1): 41–100.
47. Vatutin V, Topchii V. Limit theorem for critical catalytic branching random walks. *Theory Probab Appl* 2005;49(3): 498–518.
48. Bertacchi D, Zucca F. Statistical mechanics and random walks: principles, processes and applications. Recent results on branching random walks. Nova Science Publishers Inc.; 2012.
49. Lamperti J. Continuous state branching processes. *Bull Am Math Soc* 1967;73:382–386.
50. Ikeda N, Nagasawa M, Watanabe S. Branching Markov processes. I. *J Math Kyoto Univ* 1968;8:233–278.
51. Caballero ME, Lambert A, Uribe Bravo G. Proof(s) of the Lamperti representation of continuous-state branching processes. *Probab Surv* 2009;6:62–89.
52. Jiřina M. Stochastic branching processes with continuous state space. *Czech Math J* 1958;8(83):292–313.
53. Grey DR. Asymptotic behaviour of continuous time, continuous state-space branching processes. *J Appl Probab* 1974;11(4): 669–677.
54. Lamperti J. The limit of a sequence of branching processes. *Z Wahrsch Verw Gebiete* 1967;7:271–288.
55. Helland IS. Continuity of a class of random time transformations. *Stoch Process Appl* 1978;7(1): 79–99.
56. Ethier SN, Kurtz TG. *Markov Processes, Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. New-York: John Wiley & Sons Inc.; 1986.
57. Duquesne T, Le Gall J-F. Random trees, Lévy processes and spatial branching processes. *Astérisque* 2002;281:vi+147.
58. Aldous D. The continuum random tree. I. *Ann Probab* 1991;19(1): 1–28.
59. Aldous D. The continuum random tree. III. *Ann Probab* 1993;21(1): 248–289.
60. Le Gall J-F. Spatial branching processes, random snakes and partial differential equations. *Lectures in Mathematics ETH Zürich*. Basel: Birkhäuser Verlag; 1999.
61. Kawazu K, Watanabe S. Branching processes with immigration and related limit theorems. *Theory Probab Appl* 1971;16(1): 36–54.
62. Lambert A. The genealogy of continuous-state branching processes with immigration. *Probab Theory Relat Fields* 2002;122(1): 42–70.
63. Duquesne T. Continuum random trees and branching processes with immigration. *Stoch Process Appl* 2009;119(1): 99–129.

- 64. Watanabe S. A limit theorem of branching processes and continuous state branching processes. *J Math Kyoto Univ* 1968;8: 141–167.
- 65. Dynkin EB. Superprocesses and partial differential equations. *Ann Probab* 1993; 21(3):1185–1262.
- 66. Li Z. Measure-valued branching Markov processes. *Probability and its Applications*. Springer: Berlin Heidelberg; 2011.

AGE REPLACEMENT POLICIES

NADER EBRAHIMI
Division of Statistics, Northern
Illinois University, DeKalb,
Illinois

INTRODUCTION

Maintenance-management activities have become part of the overall quality improvement in many companies. Preventive maintenance is a schedule of planned maintenance activities aimed at the prevention of breakdown and failures. The primary goal of preventive maintenance is to prevent the failure of a unit before it actually occurs. We assume an understanding of a unit which when placed in a socket performs some operational function, see Ascher and Feingold [1] for more details. Most preventive maintenance strategies are based on the use of planned replacements made before the failure and service rectification made after failure. Planned replacements are generally less expensive than service replacements. For different planned replacement policies we refer you to Barlow and Proschan [2], Sahin and Polatoglu [3], Yeh *et al.* [4], Chien [5], and many references cited there.

One of the basic and simple replacement policies is the age replacement (AR) policy. Under an AR policy we replace a unit at failure or at the end of a specified time interval, whichever occurs first. Of course, this makes sense if a failure replacement costs more than a planned replacement. If a unit is replaced upon its failure only, then we refer to this policy as a renewal replacement (RR) policy.

Sometimes, in practice, the age of a unit is measured in more than one timescale, for example, cars age in the “parallel” scales of calendar time since purchase and the number of miles driven. In such situations, a maintenance policy should take into account the parallel scales in which a unit operates. AR

policies can also be developed based on several timescales. For more details see Frickenstein and Whitaker [6] and many references cited there. In this article, however, we concentrate on the AR policy based on a single timescale.

The article is organized as follows. In the section titled “Properties of Age Replacement Policy”, we give several properties of AR policy. In the section titled “Optimal Age Replacement Policy”, we obtain an optimal AR policy. Finally, in the section titled “Multivariate Age Replacement”, we define a multivariate version of AR policy and give its properties.

PROPERTIES OF AGE REPLACEMENT POLICY

Let X be a nonnegative continuous random variable representing the time to failure of a unit. Let $\bar{F}(x) = P(X > x)$ be the survival function of X and $F(x) = 1 - \bar{F}(x)$ be the cumulative distribution function of X . The probability density and the hazard functions of X are $f(x) = -\frac{d}{dx}(\bar{F}(x))$ and $h(x) = \frac{f(x)}{\bar{F}(x)}$, respectively. It is assumed that there are many identical items in stock for this unit.

Let $N(t)$ be the number of failures in $[0, t]$ for an RR policy and $N_{AR}(t, T)$ be the number of failures in $[0, t]$ under an AR policy with replacement interval T . Suppose W_i and $W_i(AR)$ are the intervals between $(i - 1)$ th and i -failure under RR and AR with replacement interval T policies respectively, $i = 1, 2, \dots$. Then,

$$P(W_i > x) = \bar{F}(x), i = 1, 2, \dots,$$

and

$$P(W_i(AR) \geq x) = ((\bar{F}(T))^j \bar{F}(x - jT))$$

$$\text{for } jT \leq x \leq (j + 1)T,$$

$$j = 0, 1, 2, \dots, i = 1, 2, \dots$$

Using the above equations the following result holds.

Result 1. For all $t \geq 0, T \geq 0, P(N(t) \geq k) \geq P(N_{AR}(t, T) \geq k)$ for $k = 0, 1, 2, \dots$ if and only if $\bar{F}$ is new better than used (NBU). $\bar{F}$ is said to be NBU if $\bar{F}(x+y) \leq \bar{F}(x)\bar{F}(y)$ for all $x, y \geq 0$.

The above result states that NBU class of life distributions is the largest class for which AR diminishes stochastically the number of failures experienced in any particular time interval $[0, t], 0 < t < \infty$. In this sense, the class of NBU distributions is a natural class to consider in studying AR. As an application of the above result consider the Weibull distribution $\bar{F}(x) = \exp(-\alpha x^\beta), x > 0, \beta \geq 1, \alpha > 0$. For this distribution, since $\bar{F}$ is NBU, one can conclude that in any interval the chance of having a smaller number of failures under AR policy is higher compared to RR policy.

It is not always possible to carry out preventive replacement at any moment in time. In an opportunity-based age replacement (OAR) preventive replacements are possible at randomly occurring opportunities. Therefore, in the OAR strategy, replacement of a unit occurs at the first opportunity after a specified age, say T . See Coolen-Schrijner *et al.* [7] and many references cited there.

If $W_i(\text{OAR})$ be the interval between the $(i-1)$ st and i th failure under OAR, $i = 1, 2, \dots$, then,

$$\begin{aligned} P(W_i(\text{OAR}) \geq x) &= \int \cdots \int \bar{F}(T+y_1)\bar{F}(T+(y_2-y_1)) \cdots \\ &\quad \bar{F}(T+(y_j-y_{j-1}))\bar{F}(x-(jT+y_j)) \quad (1) \\ &\quad \times \prod_{l=1}^j f(y_l) dy_1 dy_2 \cdots dy_j, \\ &\quad j = 1, 2, \dots, i = 1, 2, \dots, \end{aligned}$$

where the integral is over the set $A = \{(y_1, \dots, y_j) : 0 \leq y_l \leq T \text{ and } jT + y_j < x < (j+1)T\}$.

Using Equation (1), we get the following result.

Result 2. For all $t \geq 0, T \geq 0, P(N(t) \geq k) \geq P(N_{OAR}(t, T) \geq k), k = 0, 1, 2, \dots$ if and only if $\bar{F}$ is NBU. Here, $N_{OAR}(t, T)$ is the number of failures in $[0, t]$ under the OAR strategy with replacement interval T .

The above result is very interesting in the sense that under OAR, which is administratively easier to implement than AR, the NBU class of life distributions is still the largest class where the chance of having a smaller number of failures under OAR policy is higher compared to RR policy. Also, like AR policy, the class of NBU distributions is a natural class to consider in studying OAR.

OPTIMAL AGE REPLACEMENT POLICY

An AR cycle starts immediately after the installation of a new unit, with survival function $\bar{F}$, and ends with its failure, or after an operating time T , whichever comes first. If a cycle ends with a failure at age $x \leq T$, an unplanned failure replacement is performed, the cost $c_s + c_q = c_f$ is incurred and a new cycle starts. Here c_s represents the service replacement cost, c_q the failure cost and c_f is the total cost. If the unit does not fail before time T , then the preventive maintenance is carried out, the planned replacement cost, c_p , is incurred and a new cycle starts. It should be noted that the assumption of independent identically distributed (i.i.d.) failure times implies i.i.d. cycles. Throughout this section, we assume that $c_f \geq c_p$ which makes sense in many practical applications.

The average length of an AR cycle is $\int_0^T \bar{F}(x) dx$. Also, the average planned replacement, service replacement, and failure costs during the cycle are $c_p \bar{F}(T), c_s \bar{F}(T)$, and $c_q \bar{F}(T)$ respectively. Now, letting $C(T)$ denote the long-run average cost (cost rate) of an AR strategy for the unit, we then have

$$\begin{aligned} C(T) &= \frac{E(C_0(T))}{E(T_0(T))} \\ &= \frac{c_p \bar{F}(T) + c_f \bar{F}(T)}{\int_0^T x f(x) dx + \int_0^T T f(x) dx} \\ &= \frac{c_p \bar{F}(T) + c_f \bar{F}(T)}{\int_0^T \bar{F}(x) dx}, \quad (2) \end{aligned}$$

where $T_0(T) = \begin{cases} x, & \text{if } x \leq T \\ T, & \text{if } x > T \end{cases}$ and $C_0(T) = \begin{cases} c_f, & \text{if } x \leq T \\ c_p, & \text{if } x > T \end{cases}$ are cycle time and cycle

cost for replacement age T . In Equation (2), if we let $T \rightarrow 0$, then $C(T) \rightarrow \infty$. Also, if $T \rightarrow \infty$ (service replacement only policy), then $C(T) = \frac{c_f}{\mu}$, where $\mu = E(X)$ is the expected lifetime of the unit. That is, an AR policy $T(T < \infty)$ is justified if $C(T) \leq \frac{c_f}{\mu}$.

Using Equation (2), the optimal rectification period, T^* , can be obtained by minimizing the $C(T)$. More specifically,

$$C'(T) = \frac{\bar{F}(T)}{\left(\int_0^T \bar{F}(x) dx\right)^2} [(c_f - c_p)L(T) - c_f], \quad (3)$$

where $L(u) = h(u) \int_0^u \bar{F}(x) dx + \bar{F}(u)$. Now $L'(u) = h'(u) \int_0^u \bar{F}(x) dx$ and therefore if $h(t)$ is nondecreasing (nonincreasing), then $L(u)$ is nondecreasing (nonincreasing). Thus, if X has a nonincreasing failure rate (DFR) distribution, then $T^* = \infty$. If X has a nondecreasing failure rate (IFR) distribution, then T^* can be obtained by solving the following equation

$$L(T^*) = \frac{c_f}{c_f - c_p}, \quad (4)$$

provided that $\mu h(\infty) \geq \frac{c_f}{c_f - c_p}$. This means if $h(\infty) \cdot \mu < \frac{c_f}{c_f - c_p}$, then $T^* = \infty$.

As an application, consider the following example.

Example 1. Let X be the lifetime of a unit. The failure distribution of X is described by a 2-parameter Weibull distribution, with $\beta = 2.5$ and $\alpha = 1000$ h. That is, $\bar{F}(x) = \exp(-1000x^{2.5})$. Suppose the cost of corrective maintenance is $c_f = \$5$ and the cost for a preventive replacement is $c_p = \$1$. Solving Equation (4), $T^* \approx 493.047$. That is, the optimum replacement age is about 493.047 h.

Suppose the hazard function of X , $h(x)$ contains IFR regions. Examples of such hazard functions include bathtub hazard function and inverse bathtub hazard function. For such hazard functions $L(u)$ has local maximum at the end of each IFR region, and the global maximum value will exist at the end of one of the IFR regions. Also, one can obtain

T^* by still solving Equation (4) provided that $h(t_m)\mu > \frac{c_f}{c_f - c_p}$, where t_m corresponds to maximum $h(t)$. For more details see, Amari and Fulton [8] and references cited there.

Under OAR, the random length $T_{\text{OAR}}(T)$ and cost $L_{\text{OAR}}(T)$ of the cycle under consideration are

$$T_{\text{OAR}}(T) = \min(X, T + Y), \quad (5)$$

and

$$L_{\text{OAR}}(T) = c_p I(X \geq T + Y) + c_f I(X < T + Y), \quad (6)$$

where Y is the residual time to the next opportunity. Using Equations (5) and (6), the expected cost function is

$$\begin{aligned} C_{\text{OAR}}(T) &= E\left(\frac{L_{\text{OAR}}(T)}{T_{\text{OAR}}(T)}\right) \\ &= \frac{c_f}{E(X)} P(X < T + Y) \\ &\quad + \frac{c_p}{T + E(Y)} P(X \geq T + Y). \end{aligned} \quad (7)$$

Under a one-cycle criterion, the optimal threshold age T_{OAR}^* can be obtained by minimizing $C_{\text{OAR}}(T)$ in Equation (6). For more details, we refer you to Coolen-Schrijner *et al.* [7].

As an application consider the following example.

Example 2. Continuing from Example 1, keeping $c_p = 1, c_f = 5$, and $\bar{F}(x) = \exp(-1000x^{2.5}), x \geq 0$. Assume Y has the uniform distribution over the interval $(0-20)$. Solving Equation (7), $T_{\text{OAR}}^* \approx 505.28$. That is, under OAR policy, the optimum replacement age is about 505.28 h.

MULTIVARIATE AGE REPLACEMENT

Frequently, in reliability theory, we deal with a system that consists of two or more sockets and their associated units or components, interconnected to perform one or more functions. Consider a system with k components. In this section, we describe the following two replacement policies, which are extensions of

AR and RR to the multivariate case. For more details see Ebrahimi [9].

1. Under a multivariate age replacement (MAR) policy, component i of a system is replaced either at age T_i , or upon its failure, $i = 1, \dots, k$. We refer to this as the MAR at $(T_1, T_2, \dots, T_k)$.
2. Under a multivariate renewal replacement (MRR), components are replaced upon their failures.

For simplicity we confine our study to the case where $k = 2$ and define the following counting processes:

- $N_i(t)$ = the number of component i failures in $(0, t]$ under MRR, $i = 1, 2$;
- $N_{AR}(t; i, T)$ = the number of component i failures in $(0, t]$ under MAR (T, T) , $i = 1, 2$.

Let X_1 and X_2 be two nonnegative continuous random variables representing the times to failure of components 1 and 2, respectively. Let $\bar{F}(x_1, x_2) = P(X_1 > x_1, X_2 > x_2)$ be the joint survival function of X_1 and X_2 and $\bar{F}_{X_i}(x_i) = P(X_i > x_i)$ be the survival function of X_i , $i = 1, 2$. It is assumed that $\bar{F}(0, 0) = \bar{F}_{X_1}(0) = \bar{F}_{X_2}(0) = 1$ and $\bar{F}(\infty, \infty) = \bar{F}_{X_1}(\infty) = \bar{F}_{X_2}(\infty) = 0$. We denote the conditional survival function of X_1 given $X_2 > y$ and the conditional survival function of X_2 given $X_1 > x$ by

$$\bar{H}_y^{(1)}(x) = P(X_1 > x | X_2 > y), \quad (8)$$

and

$$\bar{H}_x^{(2)}(y) = P(X_2 > y | X_1 > x), \quad (9)$$

respectively.

Before we present our main result in this section, we need to make the following assumptions.

1. As soon as a component of the system fails or it reaches the age T , it will be replaced by a new and an identical component whose lifetime is independent of the replaced component but is dependent on the lifetimes of the components that are currently in service.

2. The replacement time (the time it takes to replace either component) is negligible.

The following result which is an extension of Result 1 to the bivariate case provides a stochastic comparison of failures experienced under MAR (T, T) and MRR.

Result 3. If (a) $\bar{F}(x + t_1, x + t_2) \leq \bar{F}(x, x)$ for all $x, t_1, t_2 \geq 0$ and (b) $\bar{H}_x^{(i)}(t_1 + t_2) \leq (\geq) \bar{H}_x^{(i)}(t_1) \bar{H}_x^{(i)}(t_2)$, $i = 1, 2$, for all $x, t_1, t_2 \geq 0$, then

$$\begin{aligned} P(N_1(t_1) \leq n_1, N_2(t_2) \leq n_2) & (\leq) \\ & \geq P(N_{AR}(t_1; 1, T) \leq n_1, N_{AR}(t_2; 2, T) \leq n_2) \end{aligned} \quad (10)$$

for all $n_1, n_2, t_1, t_2 \geq 0$.

The above result says that under conditions (a) and (b), MAR (T, T) diminishes stochastically the number of failures experienced in any particular time intervals $[0, t_1]$ and $[0, t_2]$, $0 < t_1, t_2 < \infty$ by components 1 and 2 respectively. As an application, suppose the joint failure distribution of components 1 and 2 is described by a bivariate Gumbel distribution,

$$\begin{aligned} \bar{F}(x, y) &= \exp(-\lambda_1 x - \lambda_2 y - \lambda_3 xy), \quad x, y > 0, \\ \lambda_1, \lambda_2, \lambda_3 &> 0. \end{aligned}$$

Since the assumptions of Result 3 holds for this distribution, we can claim that the chance of having less numbers of components 1 and 2 failures under MAR (T, T) policy is higher compared to MRR policy.

Optimal Replacement Policy under MAR (T, T)

In this section, we determine an optimal replacement policy T^* under MAR (T, T) such that

$$\begin{aligned} C(T, T) &= \frac{\text{the expected cost incurred in a cycle}}{\text{the expected length of a cycle}} \end{aligned} \quad (11)$$

is minimized. Here a cycle is the time between two consecutive system failures. We note that $C(T, T)$ is being used as the criterion for evaluating the replacement policies.

To evaluate $C(T, T)$ in Equation (11) we have to describe system failure. For that we need to know how the components are connected. In this section we consider the following cases.

Case I. The system works if both components work. Let c_f be the constant cost of replacement of one or both components at the system's failure and let c_p be the constant cost of a preventive replacement of both components at age T with $0 < c_p < c_f < \infty$. Since the system fails if at least one component fails, therefore the lifetime of the system in a given cycle, say j , is $\min(Y_1(j), Y_2(j))$ and the number of preventive replacements of both components at age T in the cycle is $[\min(Y_1(j), Y_2(j))/T]$. Here $Y_i(j)$ is the interval between $(j-1)$ th and j th arrivals of the process $N_A(t; i, T)$, $i = 1, 2$, $j = 1, 2, \dots$. Consequently, the total cost is

$$c_p[\min(Y_1(j), Y_2(j))/T] + c_f, \quad j = 1, 2, \dots, \quad (12)$$

where $[a]$ is the largest integer less than or equal to a . Now,

$$E(\min(Y_1(j), Y_2(j))) = \frac{1}{1 - \bar{F}(T, T)} \int_0^T \bar{F}(u, u) du, \quad j = 1, 2, \dots, \quad (13)$$

and

$$E\left[\frac{\min(Y_1(j), Y_2(j))}{T}\right] = \frac{\bar{F}(T, T)}{1 - \bar{F}(T, T)}. \quad (14)$$

Using Equations (14) and (15), Equation (12) reduces to

$$C_1(T, T) = \frac{1}{\int_0^T \bar{F}(u, u) du} (c_p \bar{F}(T, T) + c_f(1 - \bar{F}(T, T)))$$

$$= \frac{1}{\int_0^T \bar{F}(u, u) du} (c_f - \beta \bar{F}(T, T)), \quad (15)$$

where $\beta = (c_f - c_p) > 0$. We can determine the optimal replacement policy T^* by minimizing $C_1(T, T)$. The minimization procedure can be done by analytical or numerical methods. See Ebrahimi [9] for more details.

As an application consider the following example.

Example 3. Let X_1 and X_2 be lifetimes of components 1 and 2 respectively. The joint distribution of X_1 and X_2 is described by a 3-parameter bivariate Gumble distribution, with $\lambda_1 = \lambda_2 = \lambda_3 = 1$ year. That is, $\bar{F}(x, y) = \exp(-x - y - xy)$. Suppose the cost of corrective maintenance is $c_f = \$2$ and the cost for a preventative replacement is $c_p = \$1$. For this case, Equation (15) can be written as

$$\exp[-2T - T^2] + (2 + 2T) \int_0^T \exp[-2u - u^2] du = 2. \quad (16)$$

Solving Equation (16) we get T^* approximately equal to 0.27. That is, the optimum replacement age is about 3.12 months.

Case II. The system works if at least one component works. Under this case the system fails if both components fail, therefore the lifetime of the system in a given cycle, say j is $\max(Y_1(j), Y_2(j))$. Using arguments similar to those in Case I, Equation (9) reduces to

$$C_2(T, T) = \frac{c_p E[\max(Y_1(j), Y_2(j))/T] + c_f}{E(\max(Y_1(j), Y_2(j)))} = \frac{1}{\sum_{i=1}^2 \frac{\int_0^T \bar{F}_i(u) du}{1 - \bar{F}_i(T)} - \frac{\int_0^T \bar{F}(u, u) du}{1 - \bar{F}(T, T)}}$$

$$\times \left(c_f + c_p \sum_{i=1}^2 \frac{\bar{F}_i(T)}{1 - \bar{F}_i(T)} - \frac{c_p \bar{F}(T, T)}{x1 - \bar{F}(T, T)} \right). \quad (17)$$

We may proceed as in Case I and determine the optimal replacement policy T^* by minimizing $C_2(T, T)$ in Equation (17). However, unlike Case I, it is hard to obtain a general necessary condition for a unique and finite T^* (T^* is the time that minimizes Equation 17). Each parametric family of distributions must be studied separately in order to obtain such a condition.

REFERENCES

1. Ascher H, Feingold H. Repairable systems reliability: modeling, inference, misconceptions and their causes, Volume 7, Lecture notes in Statistics. New York: Marcel Dekker Inc.; 1984.
2. Barlow R, Proschan F. Mathematical theory of reliability. Volume 17, Classics in applied mathematics. Philadelphia (PA): SIAM 1996.
3. Sahin I, Polatoglu H. Quality, warranty, and preventive maintenance. Kluwer Academic Publishers; Boston. 1998.
4. Yeh RH, Chen GC, Chen MY. Optimal replacement policy for nonrepairable product under renewing free replacement warranty. IEEE Trans Reliab 2005;54:92–97.
5. Chien YH. Optimal age replacement policy under an imperfect renewing free-replacement Warranty. IEEE Trans Reliab 2008; 57:125–133.
6. Frickenstein SG, Whitaker LR. Age replacement policies in two time scales. Nav Res Logist 2003;50:592–613.
7. Coolen-Schrijner P, Shaw SC, Coolen FPA. Opportunity-based age replacement with a cycle criterion. J Oper Res 2009;60:1428–1438.
8. Amari S, Fulton W. Bounds on optimal replacement time of age replacement policy. IEEE Trans Reliab 2003;52:717–723.
9. Ebrahimi N. Multivariate age replacement. J Appl Probab 1997;34:1032–1040.

AGGREGATE PLANNING

MIKE SWIHART

ArrowStream in Logistics
Engineering, Chicago,
Illinois

Aggregate planning creates a high level plan that has been simplified by combining a number of entities into a single entity. This article focuses on the application of aggregate planning to strategic situations for supply chains. A full supply chain starts with raw materials and ends with a final customer who uses finished goods. As such it can include facilities that supply raw materials, manufacture finished goods, and store products. It also includes the transportation links that connect these facilities. A supply chain will typically include multiple companies, since a single company normally does not encompass all of these functions [1]. Since the breadth and depth of a supply chain can be overwhelming, it is a good candidate for aggregate planning.

In strategic situations aggregate planning is used for developing models that inform strategic decisions. Some examples of these types of decisions include adding, removing, or changing suppliers, manufacturers, or storage locations. Other examples are changing the capabilities of a facility, or changing the transportation links between facilities. The reason why aggregate planning is used is because all the details are not required for the solution or where details cannot be included either because the construction would be too complex, the solve time would be too long, or a final solution would not be useful at that level of detail [2].

- *Strategic Applications of an Aggregate Plan.* In strategy the objective is to make decisions that are difficult to reverse, are expensive, and/or have an impact on a large number of other decisions. For example, in the context of optimizing a supply chain an aggregate

plan might combine a large number of UPCs. (A UPC is the Universal Product Code that identifies what a product is and who made it. UPCs are used on virtually all retail products. [1]) The aggregate planning model would be used for making decisions such as where to place a new production plant or where to add a new production line. For strategic situations, the typical implementation is to construct a number of aggregate plans, each one representing a scenario. After a set of scenarios has been made and optimized they are compared and analyzed. This is used to inform decisions.

- *Mathematical Approach for Aggregate Planning.* The mathematical approach behind aggregate planning is typically a linear program or LP (see the section titled “Linear Programming” in this encyclopedia). Sometimes some binary variables are required turning the problem into a mixed-integer program or MIP (see the section titled “Models and Algorithms” in this encyclopedia), but the majority of the structure will still be linear. The mathematical approaches for this type of problem are well developed, and software is employed that addresses the basic mathematical needs.
- *Types of Software Available for Aggregate Planning.* It is not uncommon to find software that is stand alone, and it is often PC based. While it is possible to solve aggregate planning models with a general LP solver, most often specialized software is used especially in the context of supply chain models. This is because this type of problem is well known and commonly desired structures and rules relating to facilities, production, shipping, inventory, and so on, can be anticipated, so it is worthwhile to have features built in to allow quick modeling.

- *Key Areas that are Challenges in Aggregate Planning.* There are several areas where issues can arise. One area is with respect to the mathematics itself. Other areas are more likely to be the source of challenges, including the business processes, the data processes, and the decision processes.

THE USE OF AGGREGATE PLANNING IN STRATEGY

What Is Strategic Aggregate Planning

Strategy addresses questions that can be characterized as being long term, being difficult to reverse, costing capital, having a significant expense, and/or affecting many other decisions. Further, to justify doing a true aggregate model there has to be a significant amount of complexity. If not, then aggregate plan can simply be a spreadsheet exercise. Some ways that it might be complicated could be a supply chain with a large number of products, a large number of production sites, or having many production lines capable of running different types of products with interactions between lines. There can also be complexity within the demand profile, such as a high seasonality. Another example is a situation where there are a significant number of vendors that give rebates when order quantities hit negotiated thresholds minimums.

When to Use Strategic Aggregate Planning

A model is most valuable when the system is in a significant state of change. New situations require new solutions. If the real-world situation is not changing very quickly, then the current supply chain setup can simply be maintained. If there is a reason to believe that the existing supply chain is suboptimal, then modeling can still be useful even if the current system is not changing. Obviously, it is preferred to optimize a supply chain immediately after a significant change rather than run suboptimally for some period of time.

One type of change is when there is a high level of growth in product demand,

either organic growth or growth thru acquisitions. There could be a large flux in the mix of products that are offered, or (hopefully not) a significant decrease in demand where the size of the supply chain needs to be reduced.

What Strategic Aggregate Planning Is NOT

Some of the things that aggregate planning is NOT intended to address are detailed production plans and it is definitely not intended to produce a production schedule. It follows that modeling at a UPC level is not appropriate. Production decisions on a short term basis are not an appropriate application. Low level decisions are not appropriate, such as trying to site a facility within a predefined metro area.

Aggregate planning is also not designed for optimizing inventory levels under uncertainty (see the section titled “Inventory Management and Control” in this encyclopedia). The most typical inventory question—“How much inventory should be carried to protect against probabilistic fluctuations in demand or supply?” is stochastic based. Aggregate planning is intended for problems that can be addressed with deterministic mathematical models. If the stochastic elements outweigh the deterministic parts of the problem then the problem is not well suited for aggregate planning. Of course it is true that with the use of many scenarios a variety of inputs values can be addressed, but aggregate planning does not take variability into account in the model itself (see the section titled “Stochastic Models” in this encyclopedia).

Questions Answered by a Strategic Aggregate Planning Model

There are a number of common strategic questions that aggregate planning is used to address in the context of a supply chain. First, it is used to determine when more production capacity is required. The capacity can be over all, by product, by region, or by time period. If the demand has been split into different sales channels then this can also be by channel. When more capacity is needed, a number of scenarios can be done

to investigate options for new plants, new individual production lines, and enhancements of existing production lines by adding capacity or new capabilities. These scenarios delve into the many options for the location, capability, and capacity for these capacity changes. Similarly, when there are large reductions in demand these analyses can be repeated but for a shrinking supply chain.

Another common set of questions center around the use of distribution centers (DCs). Aggregate planning is used to determine when a change in DC space is required, including options for expanding or contracting existing DCs, adding DCs, and closing DCs. Aggregate planning can help in determining which products to flow through which DCs, and to set service territories for DCs.

What Is Aggregated

At an appropriate level of detail this type of problem can have a very broad scope, and in fact the breadth of scope is often a driver requiring the detail to be aggregated. The scope can accommodate production, logistics, and inventory, with the key being how each of these is aggregated. All three of these are affected in the way that demand is aggregated. The demand will not be at a UPC level but instead at a product level. Demand would typically be at a monthly or even annual level as opposed to by shift or by day, and often not even by week. This means that production, logistics, and inventory will be aggregated at a product level for the same time buckets. The demand may not be at the customer level since a company can have thousands of ship to points for customers. Instead the demand may be by metro area or even by DC, the warehouse that faces the customer, which determines how the logistics flow will be aggregated. This implies that individual loads are aggregated, so that the aggregate plan only has overall product flow. Further, facilities are often included for storing inventory, but some may be aggregated. If there are several outside warehouses in one metro area, they may be modeled as one large storage space. While these are common types of aggregation, other

entities may require aggregation depending on the details of the system.

What Is Included in a Strategic Aggregate Planning Model

Production. Production is commonly included. Production can be by line and by product. It can take into account external and internal production along with the differences in variable production costs by site, for example, labor, materials, ingredients, and utilities. Tiered pricing and rebates can also be included for external production sources. Production can take into account the capacity limits and can take into account the labor constraints at an aggregate level. Even though there is some level of aggregation in the products themselves, individual product lines typically do not need to be aggregated. Despite the fact that a broad scope of production information is included in a strategic aggregate plan, the output is not intended to be a production schedule.

Logistics. Logistics is normally included, with interfacility transportation, customer shipping, and shuttles. The cost of lane rates, fuel, and handling can be included. Multiple modes can be included, such as over the road truck versus rail but this is an aspect that may also require aggregation. However the aggregation requires that individual loads are not considered. What is considered is the flow between facilities or from a facility to a customer.

Inventory. Finally it is not uncommon to include inventory with safety stock and pre-building limits. Costs can be included for storage costs, cost of capital, and related items. The inventory would be constrained between a preset minimum and maximum. These limits are inputs to the plan, and are not being calculated by it to take into account the variability of demand or supply (see the section titled "Inventory Management and Control" in this encyclopedia). Other forms of modeling would be used to set the appropriate level of inventory based on a target level of product availability or stock out.

MATHEMATICAL ISSUES

What Are the Mathematics

The fundamental mathematical approach behind aggregate planning is an LP (see the section titled “Linear Programming” in this encyclopedia) or a mixed integer program (see the section titled “Models and Algorithms” in this encyclopedia). The vast majority of the structure is linear. The reasons why binary variables are introduced are due to cost structure or the need to capture particular system behavior. An example of a cost structure that requires binary variables is a third party supplier that has tiered pricing or provides a rebate for production above a certain threshold. An example of a system behavior that requires binary variables is when there is plant or third party supplier that is only operationally viable when a minimum amount of production is reached. In that case the model has the choice between no production at that site or production at least equal to the minimum.

Strategic models also use binary variables as an efficient means to sort through many options. For example, a manufacturer might need to add the capability to produce a new product, and have to choose from 20 existing manufacturing lines located in six existing plants, with a capital cost for each line where the capability is added. Rather than run a separate model for each option, a single model that is permitted to choose between the options is sometimes more efficient. This introduces binary variables. This approach has the drawback that the results given only cover the optimal solution. Very little information is gained on good solutions that are suboptimal, that is, the runner-up solutions, so this approach is not as useful as it first appears.

What Are the Challenges in the Mathematics

Usually the mathematics are taken care of in the software. The areas where the math becomes important are in developing the structure and determining how to approach the scenarios. Constructing the model itself requires some knowledge of the impact of adding linear variables and binary variables,

especially around the impact on the combinatorics. This is important because of the impact on the time to solve or optimize a scenario. In a business context, a good solution in a reasonable amount of time is more valuable than a perfect solution that comes too late. Some knowledge of the mathematics is needed to know when to make trade-offs between the accuracy of having a set of binary variables to get the right behavior or cost structure versus using an approximation to reduce the solve time.

Some amount of mathematical knowledge is also useful because most programs will allow some level of control within the LP, or branch and bound for MIPs. Knowledge of the mathematics that is used in these approaches can be useful in knowing what settings can be helpful when a scenario is taking too long to optimize. It is also good to have some knowledge of the combinatorics so that when there are equally valid choices in structure, the one that minimizes the time to optimize can be chosen.

Further, when using the output of a strategy aggregate model financial concepts such as the time value of money, internal rate of return (IRR), and net present value (NPV) will all come into play in the final decision. These may or may not be incorporated into the model itself. The proper mathematical background is helpful to know how to use the output of a model that may span a single year for a decision that affects many years.

What Can be done about the Challenges in the Mathematics

As far as the mathematical approach is concerned, the basic math is largely well solved. When and if the time to optimize is inordinately long the business is typically sufficiently well understood to know what trade-offs in model construction are reasonable to make. In other words, when the modeling is being done well, the modeler will know whether or not different elements of the cost or structure are critical and likely to drive the results. A good modeler will know where large amounts of detail are required and what areas can be approximated. The implication of this is that to address the mathematical challenges that may arise

what is normally required is not just an understanding of the mathematics but also a strong understanding of the business.

BUSINESS PROCESS

What Is the Business Process

Strategic aggregate planning is project oriented and not a continuous process like normal operational production planning. There is a specific set of questions that are to be answered in a reasonable amount of time to make a long-term decision in a timely manner. The business process tends to be more difficult than the mathematics. For example, if a supply chain requires an additional capacity, then the decision to be made using aggregate planning may be to determine a good location for a new plant that will be cost effective for many years. However, the decisions need to be made such that the plant can be built soon enough to avoid cutting sales or at least to avoid expensive manufacturing alternatives for an extended period.

This means that the project output required is not just a single optimal answer but a complete analysis to inform the decision. In most studies this requires not just the optimal result but also a number of good, suboptimal alternatives, with enough detail for financial analysis. Internal customers or external customers who are going to use the information to make decisions need to see and understand the alternatives and see how close they are to the optimal solution. They need to weigh the calculated financial impact of choosing one option over another versus the difficulty in measuring intangible differences between scenarios. Further, a number of sensitivities are required to understand under what conditions a solution is good or even optimal and when it becomes a poor solution. Almost always scenarios will be needed that consider higher or lower sales, higher or lower costs for production or transportation, as well as scenarios that take into account other potential decisions that are in process in the organization.

A model can choose a solution that is operationally very different for a very small

amount of money from a small change in assumption values. This can lead to a solution that is unstable or brittle from an operational perspective. Since strategic decisions cannot be quickly changed, there is a benefit in choosing a suboptimal solution that is robust. For these reasons many explicit scenarios will be run. For reference, the author has seen many projects that required over 100 scenarios and a number of projects that used 500 scenarios. An individual modeler may do 750–1000 scenarios in a given year over several projects.

What Is the Workflow

The ideal work flow for an aggregate planning process would be as shown in Fig. 1. Note that it is expected even in the ideal case to be an iterative process where one result leads to more probing questions that require more scenarios. This is good thing because modeling should lead to a greater depth of understanding with new insights triggering more questions.

The actual workflow is often as shown in Fig. 2. Note that the more realistic workflow has more iterative aspects than the ideal. Models are crafted and difficult to construct in a single shot. When multiple projects are done that analyze the same supply chain the workflow will actually look like Fig. 3.

What Are the Challenges in the Business Process

Because so many scenarios can be required due to the options available, to test many inputs, and do sensitivities, it is a challenge to track and organize the information going into and coming out of the various scenarios. Therefore having sound organizational skills is critical. Further, in this type of project, there are always new options and sensitivities that can be devised. This means timeline management is critical, as is the ability to limit scope creep. These types of issues are not unique to aggregate planning projects, and can be addressed with solid project management.

Who Is Needed for the Project Team

In order for the project to be successful, the right project team needs to be assembled

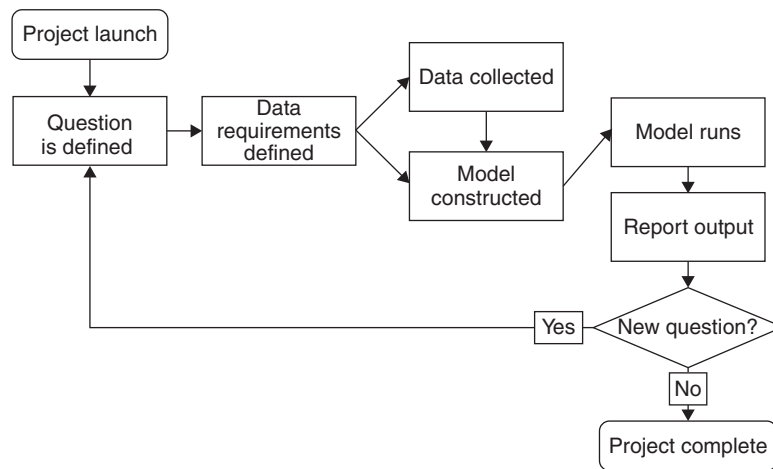


Figure 1. Ideal Work Flow.

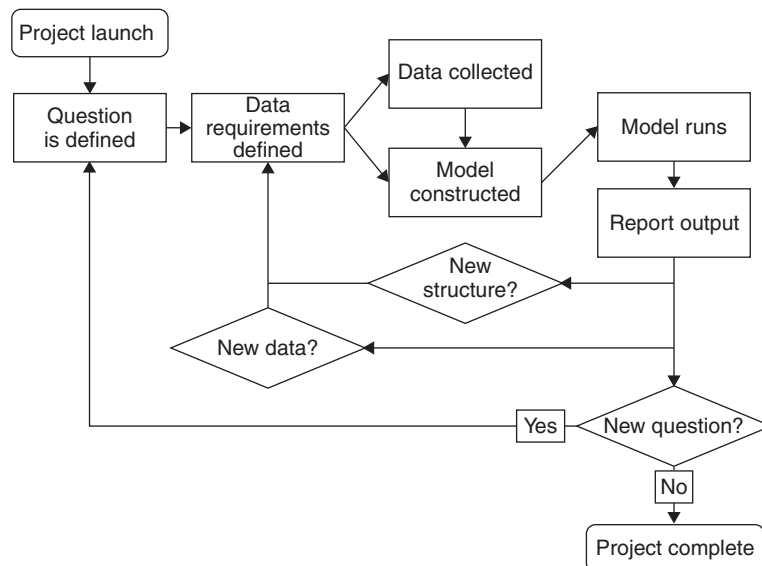


Figure 2. Realistic Work Flow.

that covers the right skill sets. Obviously, one member of the team needs to be a modeler. The modeler requires technical skills, business skills, and communication skills. The technical skills include first an understanding of the mathematics that the software is using. The modeler also requires skills in using spreadsheets and data bases at an above average level. The information that is normally supplied in aggregate planning projects does not come from a

single unified source. It is typically found in many fragments in different data bases and spreadsheets. The modeler has to be able to effectively and efficiently rearrange and process the data into a form that the software can use. It is strongly preferred that the modeler should also be able to develop automation, for example, macros and shell-scripts, to ensure accuracy and repeatability in the data processing, in using the modeling software, and in the report

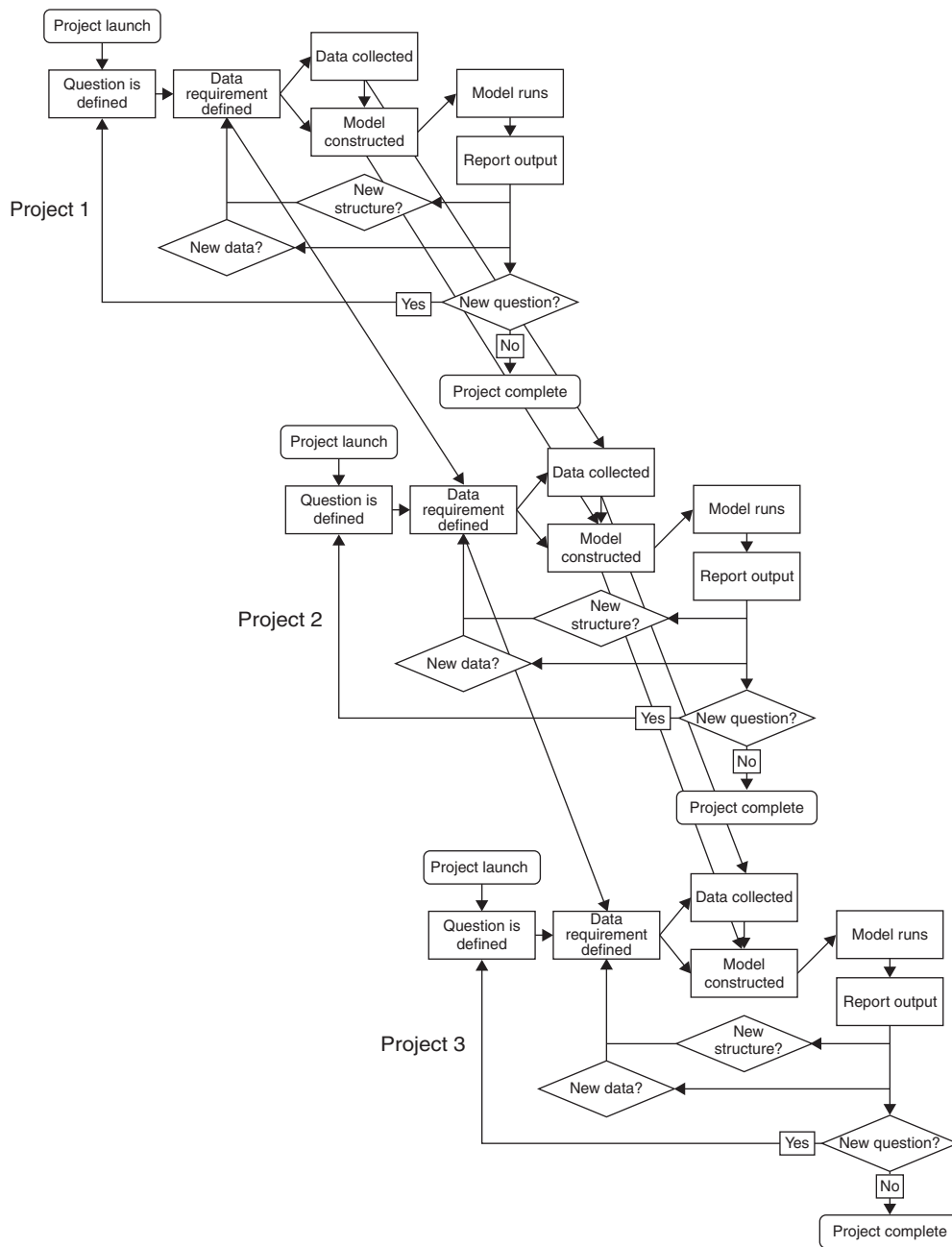


Figure 3. Realistic Multi-Project Work Flow.

generation. Further the modeler needs to have some understanding of the business. This is required so that the model structure properly captures the key elements of the

real life system that will drive the decisions that the project is required to make. This is also needed so that the modeler can apply the proper judgment of when and how to

approximate different elements of the model. The modeler needs to be able to translate the raw model output into actionable information that explains what is happening in the model and why. The modeler needs to be able to generate the types of reports that quickly and cleanly communicate the results to a wider, and frequently less technical and audience.

A number of area specialists are required. In a supply chain aggregate planning, project specialists in areas such as manufacturing, operations, logistics, and engineering are required. Their participation can be on an as-needed basis. The project will also require a financial specialist because ultimately the decisions are financially driven, so financial analysis will be a key part of the final decisions.

Finally, the project team requires a project manager. This person does not require as much technical skills as the modeler, but they are vital in the role of maintaining timelines, information flow, contacts, and team coordination. In this manner an aggregate planning project is not different from any other project where sound project management is required for success. The project manager may also supply some amount of the analysis on the project. The exact split between the analysis done by the modeler and the project manager largely depends on the technical skills and business knowledge of both. The best results are arrived when the final analysis is the product of a good set of interaction among the team members. Finally, someone in the team needs to have presentation skills to communicate out the results of the overall analysis. This would often be the project manager, though it is not required that they be the only presenter. If any of these skill sets are lacking, then no matter how sound the underlying mathematics, the project is in danger of not being successful.

DATA PROCESS

What Is the Data Process

The data process is the process to get the information needed for the aggregate plan, for example, obtaining the data for demand, production, logistics, and so on. Similar to

the business process, the data process is an area that tends to cause more issues than the mathematics. Getting good data that is clean in a format that can be used is key in constructing an aggregate plan but difficult in practice. Related to this is the ability to track the data that is going into multiple scenarios to maintain the traceability of assumptions.

What Are the Challenges in the Data Process

There are several challenges in the data process. One is that readily available data is typically from the financial tracking system. The issue is that cost accounting standards are often not designed for purposes of an aggregate plan. For example, an aggregate plan usually requires variable costs of production by product by site. Cost accounting standards that are developed for production by product by site may have allocated fixed costs or allocated over time labor costs. This means that making an incremental unit of product would not actually increase the total spend of the supply chain by the amount that is indicated in the cost standard.

Another challenge is simply having errors buried in the data. Optimization models are very good at finding the bad data. Unfortunately this is not a beneficial feature. Any information that is incorrectly cheap will be exploited by the model and then produce a false positive with respect to financial benefits. Any information that is incorrectly expensive will be avoided by the model so that good and reasonable solutions will not be used due to an error in the data.

What Can be done about the Challenges in the Data Process

The first step in dealing with the issues in the data process is simply to be prepared to invest the time to understanding what is in the data, what assumptions went into creating the data, and in identifying and correcting errors. This is best done up front in the project. Another good practice is to avoid inheritance, where one project inherits data or structure from a previous project back several generations. If a model is used on multiple projects, there should be a clear

well documented schema for the input in the current version of the model. What should be avoided is to have legacy numbers whose origin is currently unknown that were added to the model in a previous project where traceability has been lost.

Ideally the customers (internal or external) for the results of an aggregate planning exercise will be the same group that is supplying the most important inputs. This gives the group that understands the data the best the motivation for cleaning the data. This also sets up a feedback loop on the data, providing a natural incentive to balance the demands of accuracy versus the effort that can be saved in obtaining data by using approximations.

A significant amount of time is needed to analyze the results that come out of the aggregate plan. Errors that were not detected in the input can be found in understanding the behavior of the output. Therefore, it is helpful to the data process to trace the drivers that explain the behavior of particular scenarios. Again, being able to generate reports quickly and accurately that illuminate model behavior is useful since it aids in identifying the drivers. This needs to be done not just to understand what the model is doing but also what the model is NOT doing. Understanding the business to understand the reasonableness of the results is critical.

Finally a number of types of automation are incredibly useful. Automation of the data processing, of creation of scenarios, and of the output generation is profoundly useful in maintaining the consistency and quality of the results. As assumptions change or different option combinations are needed they can be handled in a nonmanual manner. A manual process can be risky, reckless, and error prone. They are also the most difficult to debug. This is the same as the difference between trying to add a long column of numbers with a calculator versus using a spreadsheet.

DECISION PROCESS

What Is the Decision Process

The decision process is simple the way that the aggregate planning analysis is used in

making a decision or a set of decisions. This can involve two types of questions—absolute and relative. Relative questions are where there is a base case scenario and an alternative scenario and the question is which one is better and by how much. This is the preferred type of question to answer because the model will not match real life, but in comparing two scenarios the biases of the model will tend to cancel. Absolute questions are where no comparison is made and the absolute values from a single scenario are used. An example of a relative question is “What is the decrease in costs due to adding a production line in plant X?” An example of an absolute question is “How many months during the year would that production line have utilization > 80%?”

What Are the Challenges in the Decision Process

The first difficulty in the decision process is to persuade the decision makers that an aggregate model should be part of the decision process. If it is not then frequently the decisions are based on the rules of thumb, which put a very real limit on the number and types of options that are explored. This also limits the thoroughness and due diligence that can be applied. The thoroughness of a model means that it can be used to uncover approaches that a person may overlook.

Once aggregate planning is part of the decision process, it is important to be clear on the purpose of the model. The point of an aggregate plan for strategic purposes is to determine what is the optimal course of action GIVEN a set of assumptions, not to determine how likely those assumptions are to be true. It is important to use an aggregate model for the purpose for which it was designed. More than that, each individual model and scenario is designed for a specific question, so it is important to craft the model for that question. If this is neglected then the risk is that modeling may be done for the sake of modeling and not to add value to the decision process.

On the other hand, once aggregate planning has gained acceptance, there is the danger of having an over dependence on it. It can become a substitute for judgment and not a means to augment judgment in the decision

making process. There is also the danger that because a specific aggregate plan covers a wide scope of aspects of a supply chain that it will be used to answer questions about everything in that scope. For example, a model may be intended to answer production questions, so it is designed with production data with significant detail and logistics data at only high level approximations. The tendency is to try to use the same model for both production and logistics questions despite the design intent.

In other words:

1. Without strategy, there is only drift.
2. You cannot predict the future. You can only prepare for it.

3. Models are only a means to answering a question. What is the question?
4. Models inform strategic decisions. They do not make them.
5. The only universal model is reality.

REFERENCES

1. Council of Supply Chain Management Professionals, CSCPM. See glossary at <http://cscmp.org/digital/glossary/document.pdf>, Accessed May 16, 2010.
2. Simchi-Levi D, Kaminsky P, Simchi-Levi E. Designing and managing the supply chain. New York: Irwin McGraw-Hill; 2000.

AGGREGATION AND LUMPING OF DTMCs¹

MARLIN U. THOMAS
Department of Electrical and
Computer Engineering, Air Force
Institute of Technology,
Wright-Patterson AFB, Ohio

Markov chains are fundamental in operations research modeling, ranking among the most applied methodologies in the field. Markov decision processes, queueing networks, and inventory systems are examples of OR/MS modeling platforms that are based on Markov chains. This is due largely to the appeal and practicality of the embedded Markov property along with the computational simplicity provided by the Chapman–Kolmogorov equations. Another attraction is the robustness and flexibility that can sometimes be gained by working with functions of Markov chains that can simplify the structure and analysis procedures.

This article provides an overview of the methods and conditions for transforming discrete time Markov chains (DTMCs) into a smaller process with fewer states by partitioning the state space into subsets each of which can be treated as a single state. Burke and Rosenblatt [1] discovered, through their early more general investigations of functions of Markov chains, conditions for which an ergodic Markov chain can be partitioned to form a smaller chain with transition probabilities that satisfy the Chapman–Kolmogorov equations for arbitrary choices of initial probability vectors. Their seminal finding initiated the development and established the foundation for lumpability theory.

¹The views in this paper are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government.

MARKOV CHAIN LUMPABILITY

Lumpability is the process of partitioning the state space of a Markov chain into subsets each of which can be treated as a single state of a smaller chain that retains the Markov property. Consider a DTMC $\{X_t : t = 0, 1, \dots\}$ with finite state space $\Omega = \{1, 2, \dots, n\}$, stationary transition probability matrix $P = [p_{ij}]$, where $i, j = 1, \dots, n$, and initial state probability vector $\mathbf{p}^0 = (p_1^0, \dots, p_n^0)$.

Let $\tilde{\Omega} = \{L_1, L_2, \dots, L_m\}$ be a nontrivial partition of Ω into $m < n$ subsets.

Strong Lumpability

Definition 1. $\{X_t\}$ is strongly lumpable with respect to $\tilde{\Omega}$ if for every initial state probability vector $\mathbf{p}^0$, the resulting chain, $\{\tilde{X}_t\}$ is Markov with transition probability matrix $\tilde{P} = [\tilde{p}_{ij}]$, $i, j = 1, 2, \dots, m$, that is invariant under choices of $\mathbf{p}^0$.

The criterion for establishing that a given Markov chain is lumpable is based on whether the particular partition $\tilde{\Omega}$ of the state space will result in $\tilde{P}$ that satisfies the Chapman–Kolmogorov equations. This forms the basis for the following conditions for strong lumpability

Theorem 1. [Ref. 2, Theorem 6.3.2]. A necessary and sufficient condition for a DTMC $\{X_t : t = 0, 1, \dots\}$ on $\Omega = \{1, 2, \dots, n\}$ with transition probability matrix $P = [p_{ij}]$ to be strongly lumpable with respect to $\tilde{\Omega} = \{L_1, \dots, L_m\}$, $m < n$, for each pair (L_i, L_j) is

$$\tilde{p}_{ij} = \sum_{r \in L_j} p_{kr}, \text{ for any } k \in L_i \quad (1)$$

Corollary 1. If $\{X_t\}$ is ergodic and $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ is the limiting probability distribution, then

$$\tilde{\pi}_j = \sum_{r \in L_j} \pi_r, j = 1, \dots, m \quad (2)$$

is the corresponding limiting probability distribution for the lumped chain.

Example 1. Consider the DTMC with the transition probability matrix

$$P = \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0.2 & 0.1 \\ 0.1 & 0.4 & 0.2 & 0.2 & 0.1 \\ 0 & 0.2 & 0.5 & 0 & 0.3 \\ 0 & 0.3 & 0.1 & 0.2 & 0.4 \\ 0.3 & 0 & 0.1 & 0.3 & 0.3 \end{bmatrix}.$$

It follows by inspection that this chain is lumpable with respect to $\tilde{\Omega} = \{(1, 2), 3, (4, 5)\} = \{L_1, L_2, L_3\}$. Thus from Equation (1)

$$\begin{aligned} p_{11} + p_{12} &= p_{21} + p_{22} & p_{13} &= p_{23} \\ p_{14} + p_{15} &= p_{24} + p_{25} \\ p_{41} + p_{42} &= p_{51} + p_{52} & p_{43} &= p_{53} \\ p_{44} + p_{45} &= p_{54} + p_{55} \end{aligned}$$

therefore,

$$\begin{aligned} \tilde{P} &= \begin{bmatrix} p_{11} + p_{12} & p_{13} & p_{14} + p_{15} \\ p_{31} + p_{32} & p_{33} & p_{34} + p_{35} \\ p_{41} + p_{42} & p_{43} & p_{44} + p_{45} \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}. \end{aligned}$$

It can be shown that the limiting probability distribution for the original chain is $\pi = (0.14, 0.21, 0.22, 0.18, 0.25)$ and for the lumped chain from Equation (2), $\tilde{\pi} = (0.35, 0.22, 0.43)$.

An equivalent characterization of strong lumpability that can be useful in examining lumpability options is given by the following.

Corollary 2. $\{X_t\}$ is strongly lumpable to $\{\tilde{X}_t\}$ if and only if there exist matrices A and B such that

$$BAPB = PB, \quad (3)$$

where $B = [b_{ij}]$ with

$$b_{ij} = \begin{cases} 1, & i \in L_j \\ 0, & i \notin L_j, \end{cases} \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m,$$

and $A = (B'B)^{-1}B'$.

The position of the 1s in each column of B correspond to the subset of states in Ω that form a lumped state in $\tilde{\Omega}$. It follows that if Equation (3) is satisfied then the transition probability matrix for the lumped chain is given by

$$\tilde{P} = APB. \quad (4)$$

Matrices A and B are the distributor and aggregator matrices, and it follows that they are useful in deriving properties of interest for the lumped process, which include the following:

1. Since $AB = I$, the s -step state transition probability matrix for the lumped chain is

$$\tilde{P}^s = AP^s B. \quad (5)$$

2. Given $\{X_t\}$ has the limiting state probability distribution π , the limiting state probability distribution for $\{\tilde{X}_t\}$ is given by

$$\tilde{\pi} = \pi B. \quad (6)$$

3. If $\mathbf{p}^0$ is the initial state probability vector for the original chain, then for the lumped chain

$$\tilde{\mathbf{p}}^0 = \mathbf{p}^0 B. \quad (7)$$

The proofs for these results follow directly from the fact that $AB = I$ and Equation (4).

Example 1 (continued). To apply the matrix operations for the lumped chain in Example 1, from Equation (3)

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$A = (B'B)^{-1}B'$$

$$= \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix}.$$

From Equation (4), it follows that

$$\tilde{P} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

$$\times P \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{bmatrix},$$

and from Equation (6), $\tilde{\pi} = \pi B = (0.35, 0.22, 0.43)$.

While the necessary and sufficient conditions of Theorem 1 and Corollary 2 for Markov chain lumpability provide useful means of establishing or confirming that a particular choice of $\tilde{\Omega}$ qualifies for $\{X_t\}$ being a lumped chain, there is a major challenge in efficiently finding these choices. Moreover, for parent DTMCs with large state spaces the related combinational problem can be insurmountable. The following approach provides useful guidance for finding alternative lumpings.

Approach for Generating Alternative Lumpings

Operationally, lumping involves taking linear functions of the parent transition probability matrix P that will satisfy the necessary and sufficient conditions (1) and (4) for suitable choices of $\tilde{\Omega}$ for which the associated transition probability matrix $\tilde{P}$ will satisfy the Chapman–Kolmogorov equation. To this end, the following results for the geometric properties of DTMCs provide useful guidance in developing lumping options [3].

We consider a process of systematically prescribing alternative lumpings by generating matrices B for a given transition probability matrix P on Ω . Consider the left

eigenvectors of P such that for λ an eigenvalue, $xP = x\lambda$, $x \neq 0$. So for $\lambda = 1$, the corresponding positive unit-eigenvector is the limiting state probability distribution π of $\{X_t\}$. The remaining eigenvectors corresponding to eigenvalues other than 1, are orthogonal to the vector $\mathbf{1} = (1, \dots, 1)$.

The procedure incorporates results from [3] on the eigenvector structure of P . Let $\{X_t\}$ be a DTMC with transition probability matrix P that is lumpable to $\{\tilde{X}_t\}$ with transition probability matrix $\tilde{P}$.

1. The eigenvalues of $\tilde{P}$ are eigenvalues of P .
2. For left eigenvector x associated with eigenvalue λ of P , $(xB)\tilde{P} = (xB)\lambda$.
3. If λ is *not* an eigenvalue of $\tilde{P}$, then $xB = 0$.

The problem of searching for alternative lumpings from Corollary 2 can be treated as that of finding B matrices that satisfy Equation (3).

Example 2. Consider the simple Markov chain with the transition probability matrix

$$P = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.5 & 0.1 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}.$$

The eigenvalues λ^j and associated eigenvectors, x^j , $j = 1, 2, 3$, for P are

$$\lambda^1 = -0.1 : x^1 = (-0.5, -0.5, 1)$$

$$\lambda^2 = 0.1 : x^2 = (0, -1, 1)$$

$$\lambda^3 = 1 : x^3 = (-1, -1, -0.75)$$

Applying Result 3 and generating the matrix

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

$x^1B = (-0.5, 0.5)$, $x^2B = (0, 0)$, and $x^3B = (-1, 0)$. Therefore, $\tilde{\Omega} = \{1, (2, 3)\}$ is a “candidate” lumping of $\Omega = \{1, 2, 3\}$. Since Equation (3) is satisfied,

$$\tilde{P} = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

with eigenvalues and vectors $\tilde{\lambda}^1 = -0.1$, $\tilde{x}^1 = (-0.5, 0.5)$; $\tilde{\lambda}^2 = 1$, $\tilde{x}^2 = (-0.1, -1.75)$.

Results 1–3 are incorporated into the following procedure for generating alternative lumpings for a DTMC $\{X_t\}$.

Barr–Thomas Algorithm

1. Compute eigenvalues of P : $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = 1$.
2. For each λ^j compute left eigenvectors $x^j = (x_1^j, x_2^j, \dots, x_n^j)$.
3. Generate B such that $x^j B = 0$, $j = 1, \dots, n-1$.
4. Compute A from $AB = I$ and $\tilde{P} = APB$.
5. Check the eigenvalues of $\tilde{P}$.

The ordering of the eigenvalues in Step 1 is not necessary but is generally convenient. Since P and $\tilde{P}$ are stochastic matrices, $\lambda = 1$ will always be an eigenvalue and the associated eigenvector can be normalized to give the steady state probability vector π and $\tilde{\pi}$. So the value $\lambda = 1$ can be eliminated from consideration in generating B matrices. It should also be noted that in Step 3, it is possible to have multiple B matrices that satisfy $x^j B = 0$. Thus, since the B matrices satisfying the condition are not unique Step 5 is necessary. If one of the eigenvalues $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \dots \leq \tilde{\lambda}_m = 1$, $m < n$ is not found among the λ_j s in Step 1, then a different choice of B should be examined [4].

Example 3. As a final illustration of the algorithm consider the Markov chain in Example 1 with

$$P = \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0.2 & 0.1 \\ 0.1 & 0.4 & 0.2 & 0.2 & 0.1 \\ 0 & 0.2 & 0.5 & 0 & 0.3 \\ 0 & 0.3 & 0.1 & 0.2 & 0.4 \\ 0.3 & 0 & 0.1 & 0.3 & 0.3 \end{bmatrix}$$

Starting with Steps 1 and 2, all of the eigenvalues and associated eigenvectors for

Table 1. Eigenvalues and Eigenvectors for P in Example 1

j	λ_j	Eigenvectors, x^j				
1	-0.1	1	-1	0	1	-1
2	0.2	1	-1	0	0	0
3	0.3	-0.1	0.3	-0.2	0.1	-0.1
4	0.3	-0.1	0.3	-0.2	0.1	-0.1
5	1	0.14	0.21	0.22	0.18	0.25

P are given in Table 1. Continuing in Step 3 with eigenvalue $\lambda_1 = -0.1$ and corresponding vector $x^1 = (1, -1, 0, 1, -1)$,

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

In Step 4, computing A and then from Equation (4),

$$\tilde{P} = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}.$$

Checking the eigenvalues of $\tilde{P}$, $\tilde{\lambda}_1 = -0.1$, $\tilde{\lambda}_2 = 0.3$, and $\tilde{\lambda}_3 = 1$, from which it follows from Corollary 2 that this choice of B is an alternative lumping with the partition $\tilde{\Omega} = \{(1, 2), 3, (4, 5)\}$.

Weak Lumpability

Strong lumpability is a useful theoretical construct for Markov chain modeling but it is generally too restrictive for practical applications. A less restrictive form of lumpability, apparently conceived by Burke and Rosenblatt [1] and Kemeny and Snell [2] requires that the transition probability matrix $\tilde{P}$ on the partitioned state space $\tilde{\Omega}$ satisfy the Chapman–Kolmogorov equation for only selected initial probability vectors. One of these vectors will be the steady state probability vector. Formally, as with strong lumpability we consider a DTMC $\{X_t : t = 0, 1, \dots\}$ with finite state space $\Omega = \{1, 2, \dots, n\}$, stationary transition probability matrix $P = [p_{ij}]$, $i, j = 1, \dots, n$, and $\tilde{\Omega} = \{L_1, L_2, \dots, L_m\}$ is a nontrivial partition

of Ω into $m < n$ subsets. Let $\mathbf{A}$ denote the set of all initial probability vectors.

Definition 2. $\{X_t\}$ is weakly lumpable with respect to $\tilde{\Omega}$ if for selected initial state probability vectors $\alpha^{(k)} \in \mathbf{A}$ the resulting chain, $\{\tilde{X}_t\}$, is Markov with transition probability matrix $\tilde{P} = [\tilde{p}_{ij}]$, $i, j = 1, 2, \dots, m$.

The major problem in establishing strong lumpability is in finding alternative lumpings for a given DTMC. While weak lumpability is a less restrictive condition, it is further complicated by the need for finding suitable initial probability vectors for which the Chapman–Kolmogorov equation holds.

The basic framework for examining weak lumpability conditions originated with the early work of Kemeny and Snell [2]. They showed that some but not necessarily all initial probability vectors $\alpha \in \mathbf{A}$ will permit the aggregated chain $\tilde{X}_t = \text{agg}(\alpha, P, \tilde{\Omega})$ to satisfy the Chapman–Kolmogorov equation. Moreover, the transition probability matrix for the lumped chain, $\tilde{P}$, will be the same for all such $\alpha \in \mathbf{A}$ that do result in a Markov chain. The challenge is in finding such a set of initial probability vectors. Abdel-Moneim and Leysieffer [5] introduced an approach for characterizing weak lumpability that was perfected and extended by Rubino and Sericola [6,7] to provide a procedure for computing the set of initial probability vectors that will lead to lumpable Markov chains.

Denoting the aggregated chain from (α, P) over a partition $\tilde{\Omega} = \{L_1, \dots, L_m\}$, $m < N$ by $\text{agg}(\alpha, P, \tilde{\Omega})$, and the cardinality of L_i by $\kappa(i)$, the partitioned subsets can be written as

$$\begin{aligned} L_1 &= \{1, \dots, \kappa(1)\} \\ &\dots \\ L_j &= \{\kappa(1) + \dots + \kappa(j-1) \\ &\quad + 1, \dots, \kappa(1) + \dots + \kappa(j)\}. \\ &\dots \\ L_m &= \{\kappa(1) + \dots + \kappa(m-1) + 1, \dots, N\} \end{aligned}$$

For $L \subseteq \Omega$ and $\sum_{j \in L} \alpha(j) \neq 0$ the vector α^L of $\mathbf{A}$ is defined by

$$\alpha^L(i) = \begin{cases} \alpha(i) / \sum_{j \in L} \alpha(j), & i \in L \\ 0, & i \notin L \end{cases}. \quad (8)$$

Further denote for each $\ell \in \Omega$ and $\alpha \in \mathbf{A}$ the vector $T_\ell \cdot \alpha$ having $\kappa(\ell)$ elements as the restriction of α corresponding to the subset L_ℓ . To illustrate this notation, consider the DTMC in Example 1 for which $N = 5$ and $L = \{L_1, \dots, L_3\}$ with $L_1 = \{1, 2\}$, $L_2 = \{3\}$, and $L_3 = \{4, 5\}$. So for $\alpha = \pi = (0.14, 0.21, 0.22, 0.18, 0.25)$:

$$\begin{aligned} \alpha^{L_1} &= (.4, .6, 0, 0, 0), \quad T_1 \cdot \alpha^{L_1} = (.4, .6) \\ \alpha^{L_2} &= (0, 0, 1, 0, 0), \quad T_2 \cdot \alpha^{L_2} = (1) \\ \alpha^{L_3} &= (0, 0, 0, .42, .58), \\ T_3 \cdot \alpha^{L_3} &= (0.42, 0.58) \end{aligned}$$

The Rubino–Sericola procedure for finding the set of initial probability vectors,

$$\mathbf{A}_M = \{\alpha \in \mathbf{A} | \tilde{X}_t = \text{agg}(\alpha, P, \tilde{\Omega}) \text{ is a Markov chain}\}$$

is based on the following conditions. Let

$$\mathbf{A}^1 = \{\alpha \in \mathbf{A} | (T_\ell \cdot \alpha^{L_\ell}) \tilde{P}_1 = \hat{P}_\ell, \forall \ell \in \tilde{\Omega}\}$$

and

$$\mathbf{A}^j = \{\alpha \in \mathbf{A} | \beta = f(\alpha; L_1, \dots, L_k), \\ k \leq j, \beta \in \mathbf{A}^1, j \geq 2\}.$$

1. If $\tilde{P}$ is the transition probability matrix for the aggregated Markov chain $\tilde{X} = \text{agg}(\alpha, P, \tilde{\Omega})$ then $\tilde{P}$ is the same for every α leading to an aggregated Markov chain [2].
2. $\mathbf{A}_M \neq \emptyset$ iff $\mathbf{A}^1$ is stable by right product by P (i.e., subset U is stable by right product by P iff $\forall u \in U$ vector uP is an element of U), thus $\mathbf{A}_M = \mathbf{A}^1$ [6].
3. $\mathbf{A}_M$ is a convex closed set [6].

Now for $\ell \in \Omega$ denote by $\hat{P}_\ell$ the probability of transitions in one step from state j in L_ℓ to L_m which is the $\kappa(\ell) \times m$ matrix defined by

$$\begin{aligned} \hat{P}_\ell(j, m) &= P\{\kappa(1) + \dots + \kappa(\ell-1) \\ &\quad + j, L_m\}, \quad 1 \leq j \leq \kappa(\ell), m \in \tilde{\Omega}. \end{aligned} \quad (9)$$

It follows that $\hat{P}_\ell$ is the ℓ th row vector of the transition probability matrix for the lumped chain $\hat{P}$. When $\mathbf{A}_M \neq \emptyset$, $\hat{P}$ is the same for all $\alpha \in \mathbf{A}_M$ and can be computed by $\tilde{P}_\ell = (T_\ell \cdot \alpha^{L_\ell}) \hat{P}_\ell$.

Example 1 (continued). Applying this procedure to the DTMC in Example 1,

$$\begin{aligned}\hat{P}_1 &= \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.2 & 0.3 \end{pmatrix}, \\ \hat{P}_2 &= (0.2 \quad 0.5 \quad 0.3), \\ \hat{P}_3 &= \begin{pmatrix} 0.3 & 0.1 & 0.6 \\ 0.3 & 0.1 & 0.6 \end{pmatrix},\end{aligned}$$

from which it follows that

$$\begin{aligned}\tilde{P}_1 &= (T_1 \cdot \alpha^{L_1}) \hat{P}_1 \\ &= (0.4, 0.6) \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.5 & 0.2 & 0.3 \end{pmatrix} \\ &= (0.5, 0.2, 0.3).\end{aligned}$$

and similarly, $\tilde{P}_2 = (0.2, 0.5, 0.3)$, $\tilde{P}_3 = (0.3, 0.1, 0.6)$, and

$$\tilde{P} = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{pmatrix}.$$

The set of initial probability vectors that lead to a time homogeneous Markov chain, $\mathbf{A}^1$ and $\mathbf{A}^j$, $j \geq 2$, is derived through the following recursive operations. Let $\beta_j = P\{X_t = j | X_t \in L_t, \dots, X_0 \in L_0\}$ and for $\alpha \in \mathbf{A}$ define the vector $\beta_k = f(\alpha; L_1, \dots, L_k) \in \mathbf{A}$ and the recursive relationship:

$$\begin{aligned}f(\alpha; L_1) &= \alpha^{L_1} \\ f(\alpha; L_1, L_2) &= (\alpha^{L_1} P)^{L_2} \\ &\dots \\ f(\alpha; L_1, \dots, L_k) &= f(\alpha; L_1, \dots, L_{k-1}) P^{L_k}\end{aligned}\tag{10}$$

Example 4. [from Ref. 6]. Consider the Markov chain $\{X_t\}$ with one-step transition matrix

$$P = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/2 \\ 1/8 & 3/8 & 1/4 & 1/4 \\ 3/8 & 1/8 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix},$$

with the alternative aggregated partition $\tilde{\Omega} = \{L_1, L_2\}$, $L_1 = (1, 2, 3)$, $L_2 = (4)$. Note that the chain is not strongly lumpable since $p_{14} = 1/2 \neq p_{24} = 1/4$. Applying Equation (6) for the limiting probability distribution $\pi = (3/13, 3/13, 3/13, 4/13)$,

$$\begin{aligned}\pi_1^{L_1} &= (1/3, 1/3, 1/3, 0), \\ \pi_1^{L_1} P &= (2/9, 2/9, 2/9, 1/3) \\ \pi_2^{L_1} &= (0, 0, 0, 1), \\ \pi_2^{L_1} P &= (0, 0, 0, 1)\end{aligned}$$

It follows that

$$\hat{P}_1 = \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \\ 3/4 & 1/4 \end{pmatrix}, \quad \hat{P}_2 = (3/4, 1/4)$$

from which it follows that

$$\begin{aligned}\tilde{P}_1 &= (T_1 \cdot \alpha^{L_1}) \hat{P}_1 \\ &= (1/3, 1/3, 1/3) \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \\ 3/4 & 1/4 \end{pmatrix} \\ &= (2/3, 1/3).\end{aligned}$$

and similarly, $\tilde{P}_2 = (3/4, 1/4)$, and

$$\tilde{P} = \begin{pmatrix} 2/3 & 1/3 \\ 3/4 & 1/4 \end{pmatrix}.$$

Now to construct $\mathbf{A}^1$, applying Equation (10)

$$\begin{aligned}f(\alpha; L_1) &= \alpha^{L_1} = (1/3, 1/3, 1/3, 0) \\ f(\alpha; L_1, L_2) &= (f(\alpha; L_1) P)^{L_2} \\ &= (2/9, 2/9, 2/9, 1/3)^{L_2} = (0, 0, 0, 1).\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{A}^1 &= \{\alpha \in \mathbf{A} | (T_1 \cdot \alpha^{L_1}) \tilde{P}_1 = \hat{P}_1, (T_2 \cdot \alpha^{L_2}) \\ &\quad \times \tilde{P}_2 = \hat{P}_2\},\end{aligned}$$

$$(\sigma_1, \sigma_2, \sigma_3) \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \\ 3/4 & 1/4 \end{pmatrix} = (2/3, 1/3),$$

thus leading to $\mathbf{A}^1 = \{\lambda (\frac{1}{3}, t, \frac{2}{3} - t) + (1 - \lambda) (0, 0, 0, 1), 0 \leq t \leq 2/3, 0 \leq \lambda \leq 1\}$.

Further details on weak lumpability can be found in Refs 2, 6–8. As with the case for strong lumpability, the current

theory does not provide a simple procedure for identifying and generating potential lumping which restricts the applications in operations research. Ledoux [9,10] has developed some interesting and promising results based on the geometric properties associated with weak lumpability. He established the equivalence of weak lumpability with the existence of a direct sum of polyhedral cones that is positively invariant by the transition probability matrix of the parent chain. This will hopefully lead to further results and insights for finding efficient algorithms for identifying lumping options.

STATE AGGREGATION

When the state space of a DTMC is very large even though numerous possible lumping can exist, the restrictive theory of lumpability does not provide easy methods for efficiently identifying and examining lumpability. Therefore, state space aggregation generally produces smaller chains that lack the Markov property and are therefore only approximations to the parent Markov Chain. In this context, lumpability can be thought of as a special case of state space aggregation where the aggregated chain is Markov. Still, lumpability theory provides the basis for developing approximation methods for computing the various measures of interest for the parent Markov chain from the aggregated chain.

The objective of state space aggregation methods is to determine the various stationary and transient results such as state probabilities, sojourn times, and mean passage times from a decomposed or aggregated chain. Takahashi [11] developed an iterative aggregation–disaggregation (IAD) procedure for Markov chain analysis by state space aggregation. This method involves computing the ergodic probabilities through an iterative process of allocating lumping of states to individual states, alternately solving an aggregated system and a disaggregated system. Schweitzer [12] extended the algorithm with a process that would ensure geometric convergence of the state probability vector computation, and further showed that for the special case of the DTMC being

weakly lumpable (called “exactly lumpable”), the IAD algorithm would converge in one step. Sumita and Rieders [13] further showed that for cases where a DTMC is lumpable, the aggregation element of the IAD procedure can be eliminated. For more details on state space aggregation methods see Refs 14–19.

SUMMARY REMARKS

State space aggregation is the process of partitioning the state space of a Markov chain and combining groups of states to form a smaller chain. Lumpability is the special case whereby the resulting smaller chain retains the Markov property as evidenced by the transition probability matrix of the lumped chain satisfying the Chapman–Kolmogorov equations. The degree of lumpability, that is, strong versus weak will depend on whether the Markov condition holds for all or selected initial state probability vectors. There are three motivations for lumpability. First is in structuring Markov models. The Markov dependence and geometric distributed sojourn times that are inherent with Markov chains make them appealing in modeling operational systems. It is quite often reasonable and conservative to assume these properties in the absence of a lot of information. So in constructing such models lumpability can provide additional features for selected applications. For example, the state spaces for manpower planning models typically have gradual growth among neighboring states with few if any backward transitions. The transition probability matrices, P , will then have entries about the diagonal with zeros in the off-diagonal positions. Lumpability options are accordingly more prevalent and easy to develop from Equation (3).

A second motivation for lumpability is the computational advantages that can be gained in working with smaller chains. This is particularly beneficial for lumpable chains with very large state spaces. The difficulty, however, is in identifying valid lumpability alternatives. The current theory, albeit complicated, provides necessary and sufficient conditions for examining lumpability for a

given partitioning of the state space. Efficient methods for finding these partitions are lacking. What is needed is a process that, given a Markov chain $\{X_t\}$ on Ω with transition probability matrix P , will generate all exhaustive partitions $\tilde{\Omega}_1, \tilde{\Omega}_2, \dots, \tilde{\Omega}_K$ to which viable lumpings exist.

The third motivation for lumpability is the value in advancing the knowledge and understanding of Markov chains. The focus of this article has been on finite time homogeneous DTMCs which have received most attention in the literature. Other lumpability efforts include Markov chains with denumerable state spaces [20], stationary reversible DTMCs [21,22], pseudo-stationarity [23], nonirreducible chains [24], and continuous time Markov chains [25].

REFERENCES

- Burke CJ, Rosenblatt MA. Markovian function of a Markov chain. *Ann Math Stat* 1958; 29:112–122.
- Kemeny JG, Snell JL. *Finite Markov chains*. Berlin: Springer; 1976.
- Barr DR, Thomas MU. An eigenvector condition for Markov chain lumpability. *Oper Res* 1977;25:1028–1031.
- Thomas MU. Computational methods for lumping Markov chains. *Proc Am Stat Assoc* 1977;364:367.
- Abdel-Moneim AM, Leysieffer FW. Weak lumpability in finite Markov chains. *J Appl Probab* 1982;19:685–691.
- Rubino G, Sericola B. On weak lumpability in Markov chains. *J Appl Probab* 1989; 26:446–457.
- Rubino G, Sericola B. A finite characterization of weak lumpable Markov processes Part I: the discrete time case. *Stoch Proc Appl* 1991;38:195–204.
- Peng N. On weak lumpability of finite Markov chains. *Stat Probab Lett* 1996;27:313–318.
- Ledoux J. A necessary condition for weak lumpability in finite Markov processes. *Oper Res Lett* 1993;13:165–168.
- Ledoux J. A geometric invariant in weak lumpability of finite Markov chains. *J Appl Probab* 1997;34:847–858.
- Takahashi Y. A lumping method for numerical calculations of stationary distributions of Markov chains. Research Report B-18. Department of Information Sciences, Tokyo Institute of Technology; 1975.
- Schweitzer P. In: Iazeolla G, Courtois PJ, *et al.* editors. *Aggregation methods for large Markov chains*. Amsterdam: Elsevier North Holland; 1984. pp. 275–286.
- Sumita U, Rieders M. Lumpability and time reversibility in the aggregation-disaggregation method for large Markov chains. *Commun Stat Stoch Models* 1989;5(1):63–81.
- Weilu CA, Stewart WJ. Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains. *J Assoc Comput Mach* 1985;32(3):702–719.
- Haviv M. An aggregation/disaggregation algorithm for computing the stationary distribution of a large Markov chain. *Stoch Models* 1992;8(3):565–575.
- Schweitzer PJ. A survey of aggregation-disaggregation in large Markov chains. In: Stewart WJ, editor. *Introduction to numerical solutions of Markov chains*. Princeton (NJ): Princeton Press; 1994. pp. 63–87.
- Kim DS, Smith RL. An exact aggregation-disaggregation algorithm for mandatory set decomposable Markov chains. In: Stewart WJ, editor. *Introduction to numerical solutions of Markov chains*. Princeton (NJ): Princeton Press; 1994. pp. 89–103.
- Heyman DP, Goldsmith MJ. Comparisons between aggregation/disaggregation and a direct algorithm for computing the stationary probabilities of a Markov chain. *ORSA J Comput* 1995;7(1):101–108.
- Marek I. Iterative aggregation/disaggregation methods for computing some characteristics of Markov chains. *Appl Numer Math* 2003; 45(1):11–28.
- Hachigian J. Collapsed Markov chains and the Chapman-Kolmogorov equation. *Math Stat* 1963;34:233–237.
- Rosenblatt M. Functions of a Markov process that are Markovian. *J Math Mech* 1959; 8:585–596.
- Hachigian J, Rosenblatt M. Functions of reversible Markov processes that are Markovian. *J Math Mech* 1962;11:951–960.
- Ledoux J, Leguesdron P. Weak lumpability and pseudo-stationarity of finite Markov chains. *Stoch Models* 2000;16:46–67.
- Abdel-Moneim AM, Leysieffer FW. Lumpability for non-irreducible finite Markov chains. *J Appl Probab* 1984;21:567–574.
- Leysieffer FW. Functions of finite Markov chains. *Anal Math Stat* 1967;38:206–212.

AGING, CHARACTERIZATION, AND STOCHASTIC ORDERING

FÉLIX BELZUNCE

Departamento Estadística e
Investigación Operativa,
Universidad de Murcia, Murcia,
Spain

MOSHE SHAKED

Department of Mathematics,
University of Arizona, Tucson,
Arizona

Many reliability systems or components of such systems have random lifetimes that indicate aging properties of these systems or components. Most of these aging properties can be characterized by various stochastic orders. Such characterizations are useful for the identification of the corresponding aging properties and for a better understanding of the meaning of these properties. In this article, we describe such characterizations. The characterizations can be used in practice to develop various inequalities that yield bounds on various probabilistic quantities of interest such as the survival function and moments. Some of these characterizations have been used to provide statistics for testing exponentiality against aging properties [1].

We also often encounter aging notions in the recent operations research and management science literature, in areas other than reliability theory. The common definitions of these aging notions, and the terminology that is associated with them, are usually stated in the language of reliability theory. This is because concepts such as monotone failure rate, monotone residual life, and new better than used (NBU) are meaningful and useful in the context of reliability theory. However, these reliability theoretic concepts are easily grasped by researchers whose expertise lies in areas other than reliability theory. Thus, although reliability theory has “monopolized” the language that is associated with aging notions, it should be

pointed out that, from a mathematical point of view, these notions abound in many other areas of operations research and management science.

Consider, for example, the failure rate notion. Mathematically speaking, this is just the intensity of an occurrence of a particular event (death of an item). Other events that occur randomly in time, in other areas of operations research and management science, also have occurrence intensity. Thus, any statement about failure rates in reliability theory has an analog in other areas. For instance, consider a G/G/1 queue. Arrivals to the queue are occurring randomly in time, and the “failure rates” that correspond to the associated interarrival times are just the arrival intensities. Similarly, the “failure rates” that correspond to the service times are just the end-of-service intensities.

As another example, consider an insured risky asset with a deductible d . The insurer may then be interested in the behavior of the expected claim size as a function of the deductible. For instance, monotonicity of the expected claim amount as a function of d , may be a useful fact. But, in the language of reliability theory, the claim size is just the residual life given that a failure has not occurred before time d . Thus, monotonicity of the expected claim amount in risk management is mathematically the same as the monotonicity of the mean residual life in reliability theory.

Similarly, the residual life at time d has a nice interpretation in reinsurance. It represents the amount paid by the reinsurer in a stop-loss agreement, given that the retention d has been reached (2, p. 124).

An interesting appearance of aging notions occurs in the theory of auctions. In a buyer's auction, the bidder with the highest bid is awarded the goods. Suppose that the valuations of the bidders are independent random variables with some common distribution function. The rent of the winner is the difference between his valuation and the price that he pays. It turns out that if the

common distribution function of the bidders valuations has some aging properties then the expected rent of the winner in a buyer's auction is monotonically decreasing in the number of bidders [3].

Many other instances of aging notions in operations research and management science areas, other than reliability theory, can be listed. We will not do it here.

In this article “increasing” and “decreasing” stand for “nondecreasing” and “non-increasing”, respectively. Expectations are assumed to exist whenever they are mentioned.

SOME STOCHASTIC ORDERS

In this section, we give the definitions of some stochastic orders that will be used in the sequel. Useful references that cover the area of stochastic orders are Refs 4 and 5. Let X and Y be two absolutely continuous nonnegative random variables with distribution functions F and G , survival functions $\bar{F} \equiv 1 - F$ and $\bar{G} \equiv 1 - G$, and density functions f and g . Let F^{-1} and G^{-1} denote the quantile functions of X and Y . Below, the definitions that follow the symbol $\bullet$ are of stochastic orders that can compare any two random variables, whereas the definitions that follow the symbol $\star$ are of stochastic orders that compare nonnegative random variables. The random variable X is said to be smaller than the random variable Y in the

- *ordinary stochastic order* (denoted as $X \leq_{st} Y$) if $\bar{F}(x) \leq \bar{G}(x)$ for all x ;
- *hazard rate order* (denoted as $X \leq_{hr} Y$) if $\bar{F}(x)\bar{G}(y) \geq \bar{F}(y)\bar{G}(x)$ for all $x \leq y$;
- *likelihood ratio order* (denoted as $X \leq_{lr} Y$) if $f(x)g(y) \geq f(y)g(x)$ for all $x \leq y$;
- *mean residual life order* (denoted as $X \leq_{mrl} Y$) if $\bar{G}(x) \int_x^\infty \bar{F}(y) dy \leq \bar{F}(x) \int_x^\infty \bar{G}(y) dy$ for all x ;
- ★ *harmonic mean residual life order* (denoted as $X \leq_{hmrl} Y$) if $(\int_x^\infty \bar{F}(y) dy) / EX \leq (\int_x^\infty \bar{G}(y) dy) / EY$ for all $x \geq 0$;
- *dispersive order* (denoted as $X \leq_{disp} Y$) if $G^{-1}(\alpha) - F^{-1}(\alpha)$ is increasing in $\alpha \in (0, 1)$;

- *excess wealth order* (denoted as $X \leq_{ew} Y$) if $\int_{F^{-1}(\alpha)}^\infty \bar{F}(x) dx \leq \int_{G^{-1}(\alpha)}^\infty \bar{G}(x) dx$ for all $\alpha \in (0, 1)$;
- *location-independent riskier order* (denoted as $X \leq_{lir} Y$) if $\int_{-\infty}^{F^{-1}(\alpha)} F(x) dx \leq \int_{-\infty}^{G^{-1}(\alpha)} G(x) dx$ for all $\alpha \in (0, 1)$;
- *increasing convex order* (denoted as $X \leq_{icx} Y$) if $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing convex functions ϕ for which the expectations are defined;
- *increasing concave order* (denoted as $X \leq_{icv} Y$) if $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing concave functions ϕ for which the expectations are defined;
- *dilation order* (denoted as $X \leq_{dil} Y$) if $X - EX \leq_{icx} Y - EY$;
- ★ *convex transform order* (denoted as $X \leq_c Y$) if $G^{-1}F(x)$ is convex in $x \geq 0$;
- ★ *star order* (denoted as $X \leq_* Y$) if $G^{-1}F(x)$ is starshaped in x ; that is, if $G^{-1}F(x)/x$ increases in $x \geq 0$;
- ★ *superadditive order* (denoted as $X \leq_{su} Y$) if $G^{-1}F(x)$ is superadditive in x ; that is, if $G^{-1}F(x+y) \geq G^{-1}F(x) + G^{-1}F(y)$ for all $x \geq 0$ and $y \geq 0$;
- ★ *Laplace transform order* (denoted as $X \leq_{lt} Y$) if $E[\exp\{-sX\}] \geq E[\exp\{-sY\}]$ for all $s > 0$.

CHARACTERIZATIONS OF AGING PROPERTIES

Throughout this section, X denotes a random variable that may have any of the aging properties that are defined and discussed below. For simplicity of exposition, we assume that X is absolutely continuous with distribution function F , survival function $\bar{F} \equiv 1 - F$, and density function f . For any event A , we use the notation $[X|A]$ to denote any random variable that is distributed according to the conditional distribution of X given A . The results below, as well as additional references, can be found in Ref. 5, unless stated otherwise.

Most of the characterizations in this section are based on stochastic comparisons of residual lives at different times. Another kind of characterizations that we describe

are based on comparisons of the “aging” random variable with an exponential random variable, which in the context of reliability theory, has the “nonaging” property.

For a nonnegative random variable X with a finite mean, let A_X denote the corresponding asymptotic equilibrium age. That is, if the distribution function of X is F then the distribution function F_e of A_X is defined by

$$F_e(x) = \frac{1}{EX} \int_0^x \bar{F}(y) dy, \quad x \geq 0.$$

The asymptotic equilibrium age is of importance in renewal theory. Suppose that we start observing a renewal process, with inter-renewal distribution F , at some time t . Then the process that we end up actually observing is a delayed renewal process, and if t is large, then the distribution of the delay is approximately the distribution of A_X given above. It is also worthwhile to mention that a delayed renewal process, with interrenewal distribution F , with the delay distribution F_e , is a stationary process—this points out another aspect of the importance of the asymptotic equilibrium age. The asymptotic equilibrium age will be used in some characterizations of aging notions below.

The relationships among the aging notions that are characterized below are given in the following chart [1]. The exact definitions of the notions in the chart will be given in the sequel.

$$\begin{array}{ccccc} \text{IFR} & \Rightarrow & \text{IFRA} & \Rightarrow & \text{NBU} \\ \downarrow & & & & \downarrow \\ \text{DMRL} & \Rightarrow & & & \text{NBUE} \Rightarrow \text{HNBUE} \end{array}$$

Increasing and Decreasing Failure Rate (IFR and DFR)

The random variable X is said to have the aging property of increasing failure rate (IFR) if $\bar{F}$ is logconcave. It has the property of decreasing failure rate (DFR) if $\bar{F}$ is logconvex on its support. The ordinary stochastic order can be used to characterize the IFR and DFR notions as follows.

Theorem 1. *The random variable X is IFR [DFR] if, and only if, $[X - t|X > t] \geq_{\text{st}} [\leq_{\text{st}}] [X - t'|X > t']$ whenever $t \leq t'$.*

The hazard rate order can also be used to characterize the IFR and DFR notions:

Theorem 2. *The random variable X is IFR [DFR] if, and only if, one of the following equivalent conditions holds (when the support of X is bounded, condition (iii) does not have a simple DFR analog):*

- (i) $[X - t|X > t] \geq_{\text{hr}} [\leq_{\text{hr}}] [X - t'|X > t']$ whenever $t \leq t'$.
- (ii) $X \geq_{\text{hr}} [\leq_{\text{hr}}] [X - t|X > t]$ for all $t \geq 0$ (when X is a nonnegative random variable).
- (iii) $X + t \leq_{\text{hr}} X + t'$ whenever $t \leq t'$.

Some similar characterizations of IFR and DFR random variables are given in Ref. 6. Other characterizations of IFR and DFR random variables, by means of ordering delayed record value spacings with respect to the order $\leq_{\text{hr}}$, are given in Ref. 7.

Here is how the dispersive order can characterize the IFR and DFR notions [8,9]:

Theorem 3. *The nonnegative random variable X is IFR [DFR] if, and only if, one of the following equivalent conditions holds:*

- (i) $[X - t|X > t] \geq_{\text{disp}} [\leq_{\text{disp}}] [X - t'|X > t']$ whenever $t \leq t'$.
- (ii) $X \geq_{\text{disp}} [\leq_{\text{disp}}] [X - t|X > t]$ for all $t \geq 0$.

Some similar characterizations of IFR and DFR random variables are given in Ref. 6.

Next we describe a result of Ref. 10, which characterizes the IFR and DFR notions by means of the location-independent riskier order:

Theorem 4. *Let X be a random variable with support of the form (a, ∞) , where $a \geq -\infty$ [respectively, $a > -\infty$]. Then X is IFR [DFR] if, and only if, $[X - t|X > t] \geq_{\text{lir}} [\leq_{\text{lir}}] [X - t'|X > t']$ for all $t' > t > a$.*

Another characterization of the IFR and DFR notions is the following [11]:

Theorem 5. *The random variable X is IFR [DFR] if, and only if, $[X - t|X > t] \geq_{\text{icv}} [\leq_{\text{icv}}] [X - t'|X > t']$ whenever $t \leq t'$.*

In some places in the literature, if the condition of Theorem 5 holds, that is, if $[X - t|X > t] \geq_{\text{icv}} [\leq_{\text{icv}}][X - t'|X > t']$ whenever $t \leq t'$, then X is said to have the property of IFR [DFR] in the second-order stochastic dominance (IFR(2) [DFR(2)]). Theorem 5 shows that the IFR(2) [DFR(2)] property is equivalent to the IFR [DFR] property.

Still another characterization, similar to the characterizations in Theorems 1–5, is given in the next theorem [12]:

Theorem 6. *The random variable X is IFR if, and only if, $[X - t|X > t] \geq_{\text{Lt}} [X - t'|X > t']$ whenever $t \leq t'$.*

The convex transform order characterizes the IFR notion as follows. We denote any exponential random variable by Exp (no matter what the mean is).

Theorem 7. *The nonnegative random variable X is IFR if, and only if, $X \leq_c \text{Exp}$.*

Using the notion of the asymptotic equilibrium age, we have the following further characterization of the IFR and DFR concepts:

Theorem 8. *The nonnegative random variable X with a finite mean is IFR [DFR] if, and only if, $X \geq_{\text{lr}} [\leq_{\text{lr}}]A_X$.*

Increasing Failure Rate Average (IFRA)

The nonnegative random variable X is said to have the aging property of increasing failure rate average (IFRA) if $-\log \bar{F}$ is star shaped; that is, if $-\log \bar{F}(t)/t$ is increasing in $t \geq 0$. The main reason for interest in this particular aging notion comes from reliability theory—this notion defines the largest class of distribution functions that are closed with respect to construction of coherent systems. The star order can be used to characterize the IFRA notion as follows. As in Theorem 7, Exp denotes any exponential random variable.

Theorem 9. *The nonnegative random variable X is IFRA if, and only if, $X \leq_* \text{Exp}$.*

New Better (Worse) than Used (NBU and NWU)

The nonnegative random variable X is said to have the aging property of new better than used (NBU) if $\bar{F}(s)\bar{F}(t) \geq \bar{F}(s+t)$ for all $s \geq 0$ and $t \geq 0$. It has the property of new worse than used (NWU) if $\bar{F}(s)\bar{F}(t) \leq \bar{F}(s+t)$ for all $s \geq 0$ and $t \geq 0$. The ordinary stochastic order can be used to characterize the NBU and NWU notions as follows:

Theorem 10. *The random variable X is NBU [NWU] if, and only if, $X \geq_{\text{st}} [\leq_{\text{st}}][X - t|X > t]$ for all $t > 0$.*

Similar to Theorem 7 and 9, the superadditive order can be used to characterize the NBU notion as follows:

Theorem 11. *The random variable X is NBU if, and only if, $X \leq_{\text{su}} \text{Exp}$.*

The order $\leq_{\text{icx}}$ was used in Ref. 13 to define the property of new better than used in convex order (NBUC) as follows. The nonnegative random variable X is said to have the aging property of NBUC if $X \geq_{\text{icx}} [X - t|X > t]$ for all $t > 0$. In fact, earlier, [14] encountered this aging notion, and called it *new better than used in mean (NBUM)*. Using the asymptotic equilibrium age the NBUC aging property can be characterized as follows: X is NBUC if, and only if,

$$X \geq_{\text{st}} [A_X - t|A_X > t] \quad \text{for all } t \geq 0.$$

The order $\leq_{\text{icv}}$ was used in Ref. 15 to define the property of NBU in second-degree stochastic dominance [NBU(2)] as follows. The random variable X is said to have the aging property of NBU(2) if $X \geq_{\text{icv}} [X - t|X > t]$ for all $t > 0$.

Decreasing and Increasing Mean Residual Life (DMRL and IMRL)

The random variable X with a finite mean is said to have the aging property of decreasing mean residual life (DMRL) if $E[X - t|X > t]$ is decreasing in t . It has the property of increasing mean residual life (IMRL) if $E[X -$

$t|X > t]$ is increasing in t . The mean residual life order can be used to characterize the DMRL notion as follows:

Theorem 12. *The random variable X is DMRL if, and only if, one of the following equivalent conditions holds:*

- (i) $[X - t|X > t] \geq_{\text{mrl}} [X - t'|X > t']$ whenever $t \leq t'$.
- (ii) $X \geq_{\text{mrl}} [X - t|X > t]$ for all $t \geq 0$ (when X is a nonnegative random variable).
- (iii) $X + t \leq_{\text{mrl}} X + t'$ whenever $t \leq t'$.

In a similar manner, the order $\leq_{\text{hmrl}}$ can be used to characterize the DMRL notion:

Theorem 13. *The random variable X is DMRL if, and only if, $[X - t|X > t] \geq_{\text{hmrl}} [X - t'|X > t']$ whenever $t \leq t'$.*

Next we describe a characterization of the DMRL and IMRL aging notions by means of the dilation order [8]:

Theorem 14. *The random variable X is DMRL [IMRL] if, and only if, $[X - t|X > t] \geq_{\text{dil}} [\leq_{\text{dil}}] [X - t'|X > t']$ whenever $t \leq t'$.*

Some similar characterizations of DMRL and IMRL random variables are given in Ref. 6.

When the support of X is bounded from below, the excess wealth order can characterize the DMRL and IMRL aging notions as follows [16]:

Theorem 15. *Let X be a continuous random variable with a finite left end point of support $a > -\infty$. Then X is DMRL [IMRL] if, and only if, any one of the following equivalent conditions holds:*

- (i) $[X - t|X > t] \geq_{\text{ew}} [\leq_{\text{ew}}] [X - t'|X > t']$ whenever $t' \geq t \geq a$.
- (ii) $X \geq_{\text{ew}} [\leq_{\text{ew}}] [X - t|X > t]$ for all $t \geq 0$ (when $a = 0$).

Again, some similar characterizations of DMRL and IMRL random variables are given in Ref. 6.

A final result of the type above that we give here uses the order $\leq_{\text{icx}}$ for characterizing the DMRL and IMRL notions [13]:

Theorem 16. *The nonnegative random variable X is DMRL [IMRL] if, and only if, $[X - t|X > t] \geq_{\text{icx}} [\leq_{\text{icx}}] [X - t'|X > t']$ whenever $t \leq t'$.*

There exists an analog of Theorems 7, 9, and 11 for the DMRL aging notion. In order to describe it, we first need to introduce the so-called DMRL stochastic order. The random variable X is said to be *smaller than Y in the DMRL order* (denoted by $X \leq_{\text{dmrl}} Y$) if

$$\frac{\int_{G^{-1}(\alpha)}^{\infty} \bar{G}(x) dx}{\int_{F^{-1}(\alpha)}^{\infty} \bar{F}(x) dx} \text{ is increasing in } \alpha \in [0, 1].$$

Then

Theorem 17. *The nonnegative random variable X is DMRL if, and only if, $X \leq_{\text{dmrl}} \text{Exp}$.*

Li and Li [17] introduced another order, denoted by $\leq_{\text{drlc}}$, such that X is DMRL if, and only if, $X \leq_{\text{drlc}} \text{Exp}$.

Using the notion of the asymptotic equilibrium age, we have the following further characterization of the DMRL and IMRL concepts:

Theorem 18. *The nonnegative random variable X with a finite mean is DMRL [IMRL] if, and only if, $X \geq_{\text{hr}} [\leq_{\text{hr}}] A_X$.*

New Better (Worse) than Used in Expectation (NBUE and NWUE)

The nonnegative random variable X with finite mean is said to have the aging property of new better than used in expectation (NBUE) if $E[X] \geq E[X - t|X > t]$ for all $t \geq 0$. It has the property of new worse than used in expectation (NWUE) if $E[X] \leq E[X - t|X > t]$ for all $t \geq 0$. The harmonic mean residual life order characterizes the NBUE notion as follows:

Theorem 19. *Let X be a nonnegative random variable with positive mean. Then X is NBUE if, and only if, any one of the following equivalent conditions holds:*

- (i) $X \leq_{\text{hmrl}} X + Y$ for any nonnegative random variable Y with a finite positive mean, which is independent of X .
- (ii) $X + Y_1 \leq_{\text{hmrl}} X + Y_2$ whenever Y_1 and Y_2 are almost surely positive random variables with finite means, which are independent of X , such that $Y_1 \leq_{\text{hmrl}} Y_2$.

There exists an analog of Theorems 7, 9, 11, and 17 for the NBUE aging notion. In order to describe it, we first need to introduce the so-called NBUE stochastic order of nonnegative random variables. Let X and Y be nonnegative random variables. Then X is said to be *smaller than Y in the NBUE order* (denoted by $X \leq_{\text{nbue}} Y$) if

$$\frac{1}{EX} \int_{F^{-1}(\alpha)}^{\infty} \bar{F}(x) dx \leq \frac{1}{EY} \int_{G^{-1}(\alpha)}^{\infty} \bar{G}(x) dx$$

for all $\alpha \in [0, 1]$.

Then

Theorem 20. *The nonnegative random variable X is NBUE if, and only if, $X \leq_{\text{nbue}} \text{Exp}$.*

Note that for any two nonnegative random variables Z and W with positive expectations we have $Z \leq_{\text{nbue}} W \iff (Z/EZ) \leq_{\text{ew}} (W/EW)$. Thus, Theorem 20 can be rewritten as follows.

Theorem 21. *The random variable X is NBUE if, and only if, $(X/EX) \leq_{\text{ew}} \text{Exp}(1)$, where $\text{Exp}(1)$ denotes an exponential random variable with mean 1.*

Theorem 1.A.32 in Ref. 5 describes still another similar characterization of the NBUE aging notion using the ordinary stochastic order $\leq_{\text{st}}$.

The asymptotic equilibrium age still gives a further characterization of the NBUE and NWUE concepts:

Theorem 22. *The nonnegative random variable X with a finite mean is NBUE [NWUE] if, and only if, $X \geq_{\text{st}} [\leq_{\text{st}}] A_X$.*

An aging notion that is closely related to the NBUE notion is the notion of the harmonic new better than used in expectation (HNBUE). Formally, the nonnegative random variable X with mean $\mu > 0$ is said to have the aging property of HNBUE if $X \leq_{\text{icx}} \text{Exp}(\mu)$, where $\text{Exp}(\mu)$ denotes an exponential random variable with mean μ . The random variable X with mean $\mu > 0$ is said to have the property of harmonic new worse than used in expectation (HNWUE) if $X \geq_{\text{icx}} \text{Exp}(\mu)$. Note that by known basic properties of the orders $\leq_{\text{icx}}$ and $\leq_{\text{dil}}$, it follows that the random variable X with mean $\mu > 0$ is HNBUE (HNWUE) if, and only if, $X \leq_{\text{dil}} [\geq_{\text{dil}}] \text{Exp}(\mu)$. The harmonic mean residual life order can be used to characterize the HNBUE and HNWUE notions as follows:

Theorem 23. *The random variable X with mean $\mu > 0$ is HNBUE [HNWUE] if, and only if, $X \leq_{\text{hmrl}} [\geq_{\text{hmrl}}] \text{Exp}(\mu)$.*

It is easy to verify that a nonnegative random variable X with mean $\mu > 0$ satisfies $X \leq_{\text{icx}} [\geq_{\text{icx}}] \text{Exp}(\mu)$ if, and only if, $A_X \leq_{\text{st}} [\geq_{\text{st}}] \text{Exp}(\mu)$. This observation yields the following characterization of the HNBUE and HNWUE concepts:

Theorem 24. *The nonnegative random variable X with a finite mean $\mu > 0$ is HNBUE [HNWUE] if, and only if, $A_X \leq_{\text{st}} [\geq_{\text{st}}] \text{Exp}(\mu)$.*

REFERENCES

1. Lai C-D, Xie M. Stochastic ageing and dependence for reliability. New York: Springer; 2006.
2. Denuit M, Dhaene J, Goovaerts M, *et al.* Actuarial theory for dependent risks: measures, orders and models. West Sussex: Wiley; 2005.
3. Li B, Li X. New partial orderings of life distributions with respect to the residual life function. J Lanzhou Univ (Natural Sciences) 2005;41:134–138.

4. Müller A, Stoyan D. Comparison methods for stochastic models and risks. New York: Wiley; 2002.
5. Shaked M, Shanthikumar JG. Stochastic orders. New York: Springer; 2007.
6. Hu T, He F, Khaledi B-E. Characterizations of some aging notions by means of the dispersion-type or dilation-type variability orders. Chin J Appl Probab Stat 2004;20(1):66–76.
7. Wei G, Hu T. Characterizations of aging classes in terms of spacings between record values. Stoch Models 2007;23:575–591.
8. Belzunce F, Candel J, Ruiz JM. Dispersive orderings and characterizations of ageing classes. Stat Probab Lett 1996;28:321–327.
9. Pellerey F, Shaked M. Characterizations of the IFR and DFR aging notions by means of the dispersive order. Stat Probab Lett 1997;33:389–393.
10. Sordo MA. On the relationship of location-independent riskier order to the usual stochastic order. Stat Probab Lett 2009;79:155–157.
11. Belzunce F, Hu T, Khaledi B-E. Dispersion-type variability orders. Probab Eng Inform Sci 2003;17:305–334.
12. Belzunce F, Gao X, Hu T, *et al.* Characterizations of the hazard rate order and IFR aging notion. Stat Probab Lett 2004;70:235–242.
13. Cao J, Wang Y. The NBUC and NWUC classes of life distributions. J Appl Probab 1991;28:473–479.
14. Bergmann R. Some classes of distributions and their application in queueing. Math Operationsforschung Stat Ser Stat 1979;10:583–600.
15. Deshpande JV, Kocher SC, Singh H. Aspects of positive ageing. J Appl Probab 1986; 23:748–758.
16. Belzunce F. On a characterization of right spread order by the increasing convex order. Stat Probab Lett 1999;45:103–110.
17. Li X. A note on expected rent in auction theory. Oper Res Lett 2005;33:531–534.

AIR TRAFFIC MANAGEMENT

ROBERT HOFFMAN
Metron Aviation, Inc.,
Dulles, Virginia

AVIJIT MUKHERJEE
University-Affiliated
Research Center,
University of California,
Santa Cruz, California

THOMAS W. M. VOSSEN
Leeds School of Business,
University of Colorado,
Boulder, Colorado

INTRODUCTION

Air traffic management (ATM) can be defined as the broad collection of services that support safe, efficient, and expeditious aircraft movement. It is common to distinguish two basic ATM components: air traffic control (ATC) and air traffic flow management (ATFM).

Air Traffic Control refers to processes that provide tactical separation services, that is, real-time separation procedures for collision detection and avoidance. ATC is usually performed by human controllers who watch over three-dimensional regions of airspace, called *sectors*, and dictate local movements of aircraft. Their aim is to maintain separation between aircraft while moving traffic as expeditiously as possible and presenting the traffic in an orderly and useful manner to the next sector. Each sector can only be occupied by a limited number of aircraft; the limit is determined by a controller's ability as well as the complexity of traffic patterns. As such, ATC actions are of a more tactical nature and primarily address immediate safety concerns of airborne flights.

Air Traffic Flow Management, on the other hand, refers to processes of a more strategic nature. ATFM procedures detect and resolve demand-capacity imbalances that

jeopardize safe separation. By keeping the workload of air traffic controllers to a manageable level, traffic flow management can be viewed as the first line of defense in maintaining system safety. Whereas ATC generally controls individual aircraft, ATFM usually adjusts aggregate traffic flows to match scarce capacity resources. Accordingly, ATFM actions have a greater potential to address system efficiency.

Air Traffic Flow Management Objectives and Challenges

The objective of ATFM is to mitigate imbalances between the demand for air traffic services and the capacity of the air transportation system, so as to ensure that aircraft can flow through the airspace *safely* and *efficiently*. In the long term, this implies efforts to prevent structural imbalances by reducing demand (by, for example, congestion pricing or auctioning off landing slots; see Ball *et al.* [1]) or increasing capacity (i.e., building new runways). In the short term, however, ATFM aims to limit—as much as possible—the impact of the congestion and delays that arise when the system is disrupted.

Fluctuating weather conditions, equipment outages, and demand surges all cause significant capacity–demand imbalances. Adverse weather conditions, in particular, frequently cause substantial reductions in airspace and airport capacity. Because these disruptions are highly unpredictable, ATFM will need to resolve the resulting capacity–demand imbalances in a dynamic manner. This is further complicated by the fact that airlines' flight schedules are highly interconnected. The aircraft, crews, and passengers that compose the flight schedule might all follow different itineraries, thus creating a complex interaction between an airline's flight legs. As a result, delays of a single flight leg can propagate throughout the network and local disruptions might have a global impact.

At the heart of the objectives and challenges of ATFM is the fact that

decision-making responsibilities are shared between a number of stakeholders. The actions performed by these stakeholders are highly interdependent, and therefore necessitate a significant degree of coordination. It is therefore no surprise that the coordination and cooperation between air traffic service providers and the airspace users have become increasingly important. In the United States, for instance, nearly all efforts to improve ATFM, nowadays, are guided by the so-called collaborative decision-making (CDM) philosophy.

The CDM philosophy recognizes that to implement appropriate ATFM actions, the service provider needs an accurate assessment of flight status and intent. Airspace users, on the other hand, need the flexibility to adjust their schedules, and can only provide accurate information if they know the actions planned by the provider. Given the relatively short response times, the real-time exchange of information between the service provider and users is therefore a critical component of ATFM functionality.

In addition, it has become increasingly clear that the service provider should not be solely responsible for determining the delays, reroutes, and so on, required to resolve congestion. While both the service provider and users can possibly delay or reroute flights, certain actions that can alleviate congestion are only available to airlines. In the United States, for example, only an airline can decide to cancel flights or to reassign passengers, crew, and aircraft. Consequently, the notion of CDM emphasizes that any successful attempt at flow management requires significant involvement from airlines and other users. Such decisions involve economic trade-offs that the air traffic service provider is not in a position to make.

Air Traffic Flow Management Initiatives

For the major portion of the previous century, the coordination of air traffic proceeded largely through tactical ATC procedures. In the United States, it was not until the aftermath of the air traffic controllers' strike of

1981 that the Federal Aviation Administration (FAA) first implemented a systematic form of flow management known as *ground holding*. Under ground holding, aircraft departures are restricted until it is determined that sufficient airspace is available for the aircraft. Initially, the use of ground holding was primarily instituted to reduce workload for the inexperienced controllers that were hired in the wake of the mass firings that accompanied the strike. However, the continued growth in air traffic that followed the airline deregulation act of 1978, together with changes in traffic patterns such as the "hub and spoke" scheduling practices used by airlines, have gradually increased the scope of ATFM initiatives.

To implement these initiatives, traffic management has a variety of control techniques at their disposal. These control techniques can be organized as follows:

- ground holding controls, that is, the selective assignment of delays to flights prior to their departure;
- rerouting controls, which impose constraints on the flight paths that an aircraft can fly; and
- airborne holding controls, which result in flight delays after take-off. Airborne delays can be applied using a variety of methods, ranging from *spacing* to *speed controls* and *vectoring*. Spacing, between aircraft traveling in the same direction, specifies and controls the separation between successive aircraft. Speed control aims to ensure safe and efficient flow of aircraft by selectively increasing or decreasing their speed, while vectoring, corresponds to minor spatial deviations from flight path.

Generally speaking, ground holding and rerouting techniques are used to support strategic activities, in that they are applied proactively hours in advance. Airborne holding controls, with the exception of spacing, are commonly used for tactical flow management and are initiated reactively. It is important to note, however, that the use of these proactive controls is perhaps uniquely relevant to air transportation: in contrast to

most other forms of transportation, aircraft cannot be stopped en route and therefore ATFM cannot allow traffic jams to develop.

MODELS FOR AIR TRAFFIC FLOW MANAGEMENT

The explosive growth and ensuing congestion in air traffic has motivated a considerable amount of research that considers the application of operations research models to ATFM. The use of decision models to support ATFM received relatively little attention prior to the 1980s, and most of the literature dates to after the emergence of formal flow management procedures that followed the air traffic controllers' strike in 1981.

This section provides an overview of the principal operations research models that have been proposed in support of ATFM. We note that, due to space limitations, our emphasis is on optimization models; for a review that also includes the more descriptive simulation or queueing models that have been used in ATFM, we refer the reader to the surveys [2,3]. Here, we distinguish between airport allocation models, where an airport represents the single constrained resource, and airspace allocation models, where congestion occurs throughout a network of airports and/or sectors of airspace.

Airport Allocation

Without a doubt, the prototypical application of optimization models for ATFM is in the so-called ground holding problem (GHP). GHP was first introduced by Odoni [4] and by Andreatta and Romanin-Jacur [5], and assumes that only a single airport in the system faces a reduction in capacity for some period of time. As a result, the flights that are scheduled to arrive during this time period will have to be delayed: due to both safety and economic concerns, this is typically done by delaying flights prior to their departure. A central concern in this setting is that the problem of assigning ground delays is both stochastic, because capacity forecasts have a significant degree of uncertainty, and dynamic, because the forecasts are updated frequently and provide new information on

how the weather conditions at an airport are changing. Thus, the overall goal of the GHP is to balance the risk of excessive ground delays (which can lead to underutilization of the airport) with the risk of excessive airborne delays (which can lead to dangerous levels of airborne holding).

Richetta and Odoni [6] were the first to propose an integer programming model to solve a multiperiod stochastic GHP. In their model, uncertainty in airport capacity is represented by a finite set of scenarios, each of which represents a time-varying profile of the airport capacity that is likely to occur. The goal is to assign ground delays to flights, given uncertainty in airport capacity, in order to minimize the total expected delay cost. The model formulation is given below.

As in most of the discrete optimization models for ATFM, the planning horizon is divided into equal time periods. Let there be Q capacity scenarios, each scenario depicting a possible evolution of airport arrival capacity over the planning period with the scenario $q \in \{1, \dots, Q\}$ having a probability of occurrence equal to p_q . Let M_t^q denote the capacity at time period t under the scenario q . In order to ensure that all flights that are scheduled to land get assigned a landing slot during a time period, let there be a time period $T + 1$ with unlimited capacity under all scenarios. In their model, Richetta and Odoni classified the flights that are scheduled to arrive during each time period into K cost classes. Let N_{kt} denote the number of flights, belonging to cost category k , that are scheduled to arrive at the airport during the time period $t \in 1, \dots, T$. The cost of ground holding a flight of class k for i time units is denoted by the cost function $G(k, i)$. As illustrated below, this cost function allows us to capture nonlinear ground delay costs for flights, while keeping the objective function linear in decision variables. Let C_a denote the unit cost of airborne holding for all flights.

The decision variables are denoted by X_{ktj}^q , which indicate the number of flights in class k scheduled to arrive during time period t that are reassigned to arrive during time period j under capacity scenario q . Let W_t^q denote the number of aircraft that are unable to land during time period t under scenario q ,

and hence face airborne holding during that time period. The objective function minimizes the total expected cost of ground and airborne delays. The integer program is given as follows:

$$\begin{aligned} \text{Min} \quad & \sum_{k=1}^K \sum_{t=1}^T \sum_{j=t}^{T+1} C_g(k, j-t) X_{ktj} \\ & + C_a \left\{ \sum_{q=1}^Q p_q \sum_{t=1}^T W_t^q \right\} \end{aligned} \quad (1)$$

subject to

$$\begin{aligned} \sum_{j=t}^{T+1} X_{ktj} &= N_{kt}, \quad k = 1, \dots, K; \\ t &= 1, \dots, T \end{aligned} \quad (2)$$

$$\begin{aligned} \sum_{k=1}^K \sum_{j=1}^t X_{kjt} + W_{t-1}^q - W_t^q &\leq M_t^q, \\ q &= 1, \dots, Q; \quad t = 1, \dots, T+1 \end{aligned} \quad (3)$$

$$W_0^q = W_{T+1}^q = 0, \quad \forall q = 1, \dots, Q \quad (4)$$

$$\begin{aligned} X_{ktj} &\geq 0 \quad \text{and integer}, \quad k = 1, \dots, K; \\ t &= 1, \dots, T; j = 1, \dots, T+1 \end{aligned} \quad (5)$$

$$\begin{aligned} W_t^q &\geq 0 \quad \text{and integer}, \quad t = 1, \dots, T; \\ q &= 1, \dots, Q. \end{aligned} \quad (6)$$

Constraint set (2) ensures that all flights scheduled to arrive during any time period get rescheduled to land before the end of planning horizon. Constraint set (3) imposes an upper bound on the number of aircraft that can land during a time period under different scenarios. Kotnyek and Richetta [7] showed that the constraint matrix of the above formulation is not totally unimodular, and in some cases the LP relaxation to the above IP will not yield integer solutions. However, if all flights belong to only one cost category, and if the ground delay cost function is monotonically increasing, the Richetta–Odoni model guarantees integer solutions (see Kotnyek and Richetta [7] for details).

It is important to note that formulation described above presents a *static* model, in that the model does not incorporate the recourse options that may be available at the start of each decision epoch. While this

model can be implemented in a dynamic manner, by repeatedly solving the model in each decision epoch (and executing the first stage), other approaches have aimed to explicitly incorporate the dynamic nature of the problem. In particular, Richetta and Odoni [8] were also the first to develop a multistage stochastic integer program with recourse for the GHP. Such a model not only accounts for uncertainty, but also utilizes updated information on capacity changes in the decision-making process, thus explicitly accounting for GHP dynamics. As in the static models, uncertainty in airport capacity is represented by a finite set of scenarios. These scenarios are arranged in a scenario tree, which reveals the availability of information on airport operating conditions (see Fig. 1 for an example). The information is based on forecasts, so that capacity changes are anticipated before they occur. If branching points in the scenario tree occur only when the operating conditions change physically (for example, at possible fog burn-off times) the active branch of the scenario tree will reflect the actual capacity at any instant. In the Richetta–Odoni model, flights are assigned ground delays as their departure time approaches, so that the decisions can be made with the most up-to-date information. Once assigned, however, ground delays cannot be revised, even though this is technically possible so long as the flight has not yet departed. On the one hand, this results in less efficient solutions as there might be unnecessary ground delays that can be recovered by appropriate revision, while on the other hand it results in a higher degree of predictability of flight departure times. Mukherjee and Hansen [9,10] extended the multistage stochastic IP formulation of Richetta and Odoni, and proposed a dynamic model for the GHP that can revise ground delays of flights in response to updated information on airport capacity. Their formulation is discussed below (wherever possible, we use the same notation as in the static model presented above).

Let Φ be a set of flights that are scheduled to fly to an airport for which a ground-holding program is necessary. As in the static model, the time of day is divided into a finite set

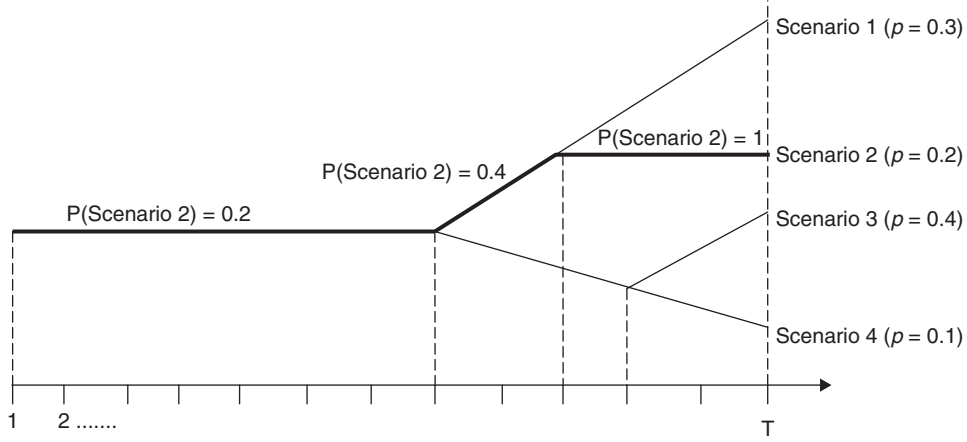


Figure 1. Scenario tree of evolving airport arrival capacity.

of time periods of equal duration. The scheduled departure and arrival times of a flight f are denoted by d_f and a_f respectively. The scenario tree is input to the model through the following variables. Let B ($B \geq Q$) be the total number of branches in the scenario tree. Each branch corresponds to a set of scenarios. The scenarios corresponding to branch $b \in \{1, \dots, B\}$ are given by the set $\Omega_b = \{S_1^b, \dots, S_k^b, \dots, S_{\pi_b}^b\}; S_k^b \in \{1, \dots, Q\}$. We assume that each branch has start and end nodes corresponding to the beginning of time periods. The time periods corresponding to the start and end nodes of branch b are denoted by o_b and μ_b , $b \in \{1, \dots, B\}$. The decision variables in the model are binary variables defined as follows:

$$X_{ft}^q = \begin{cases} 1, & \text{if flight } f \text{ is planned to arrive} \\ & \text{during time period } t \text{ under} \\ & \text{scenario } q; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

Corresponding to the X_{ft}^q is a set of corresponding auxiliary variables for the departure time period. Specifically we define

$$Y_{ft}^q = \begin{cases} 1, & \text{if flight } f \text{ is released for} \\ & \text{departure during time period} \\ & t \text{ under scenario } q; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

The departure release variables track the planned arrival times but are displaced earlier in time by the amount $a_f - d_f$. Hence the variables Y_{ft}^q are related to X_{ft}^q as follows:

$$Y_{ft}^q = X_{f, t+a_f-d_f}^q \quad \in \{d_f..T\}, q \in \{1, \dots, Q\}. \quad (7)$$

The objective function and the set of constraints are given as follows:

$$\text{Min } \sum_{q=1}^Q p_q \left\{ \sum_{f \in \Phi} \sum_{t=a_f}^{T+1} (t - a_f) X_{ft}^q + \sum_{t=1}^T \lambda W_t^q \right\} \quad (8)$$

subject to

$$\sum_{t=a_f}^{T+1} X_{ft}^q = 1, \quad f \in \Phi, q \in \{1, \dots, Q\} \quad (9)$$

$$W_{t-1}^q - W_t^q + \sum_{f \in \Phi: a_f \leq t} X_{ft}^q \leq M_t^q, \quad t \in \{1, \dots, T+1\};$$

$$q \in \{1, \dots, Q\}; \quad (W_0^q = W_{T+1}^q = 0) \quad (10)$$

$$Y_{ft}^{S_1^b} = \dots = Y_{ft}^{S_k^b} = \dots = Y_{ft}^{S_{\pi_b}^b}; f \in \Phi, \\ t \in \{1, \dots, T\}; S_k^b \in \Omega_b: \pi_b \geq 2, o_b \leq t \leq \mu_b \quad (11)$$

$$X_{ft}^q, Y_{ft}^q \in \{0, 1\}; W_t^q \geq 0 \quad \text{and integer} \\ f \in \Phi, q \in \{1, \dots, Q\}, t \in \{1, \dots, T\}. \quad (12)$$

Constraint set (9) implies that all flights are planned to arrive by the end of planning horizon $T + 1$. Constraint set (10), which is similar to Equation (3) in the static model presented above, ensures that the number of arrivals during any time period is limited by the scenario-specific airport arrival capacity for that time period. The number of arrivals in a time period t is the sum of the reduction in the size of the arrival queue between the end of t and the end of the previous time period $t - 1$, and the number of flights whose planned arrival time is in t . If the number of planned arrivals during a time period exceeds the arrival capacity, then the excess flights are subject to airborne delay and added to the arrival demand for the next time period. Constraint set (11) is a set of coupling constraints, sometimes known as *nonanticipatory constraints* in the literature [11], on the ground holding decision variables. These constraints force ground delay decisions to be made solely on information available at time t . For a given time period t , it is required that the ground holding decisions are the same for all scenarios associated with the same scenario tree branch b (in other words the scenarios belonging to the set Ω_b) in that time period.

Variants and Extensions. The static and dynamic stochastic models outlined above illustrate the fundamental decisions and trade-offs that arise in the GHP. Nevertheless, there are also a number of additional concerns, which have been addressed in several variants of the GHP.

Most of the optimization models for the GHP, for example, address only the arrival capacity shortfall at an airport, and decide on the amount of ground delay to impose on various incoming flights. Gilbo [12,13], however, presented optimization models to assign ground delays to both arrival and departure traffic at an airport. His model not only solves for ground delays of aircraft, but also optimal allocation of airport capacity to arrival and departure operations. Another stream of research [14–16] has a more tactical perspective, and considers the optimal sequencing of aircraft landings while taking into account the (flight-pair dependent)

separation standards that are required in the wake turbulence that is generated.

With the emergence of the CDM paradigm, fairness issues related to the distribution of delays in the GHP have also received attention. Under CDM, a three-step process is used to allocate airport capacities during ground delay programs (GDPs). First, airport arrival slots are assigned to airlines in a first-scheduled, first-served manner, using the ration-by-schedule (RBS) procedure. Subsequently, a substitution process allows airlines to adjust their part of the schedules according to their internal objectives. Finally, a reallocation process called *compression* aims to ensure maximum utilization when flights have been canceled or delayed [17]. In such a setting, the GHP has to determine the aggregate number of slots that are made available in each period; subsequently, these slots can be allocated using the RBS procedure. Ball *et al.* [18] address this in a model for the static stochastic GHP, and show that the resulting integer programming formulation has a dual network structure and thus can be solved efficiently. In practice, flow managers often address risk by exempting long-haul flights. This, however, raises equity issues. Vossen *et al.* [19] show that exemptions can result in a systematic bias toward airlines operating long-haul flights, and present an optimization model to mitigate those biases that can be used within a CDM framework. Hoffman *et al.* [20] also proposed an algorithm, which they termed as *equitable ration-by-distance*, which performs better, with respect to both equity and efficiency, than the distance-based flight exemptions currently used in practice by the FAA when implementing a GDP.

Finally, scenario and scenario tree generation present a vital issue in application of multistage stochastic optimization problems. While scenario generation has mostly been studied in the context of stochastic programming problems in finance, Liu [21] has recently applied statistical clustering techniques to develop capacity scenarios and scenario trees for the stochastic single airport ground holding problem (SAGHP) from historical airport capacity data. More generally, one of the primary shortcomings

of the scenario-based models for SAGHP is that they assume that a limited number of capacity profiles can occur. However, in reality the set of possible scenarios can change with time. Furthermore, scenario-based models impose a decision tree structure when in reality improved information about future capacity can be obtained continually rather than at discrete branching points. In light of these shortcomings, Liu and Hansen [22] proposed a “scenario-free” sequential decision making model, based on dynamic programming techniques, for the stochastic SAGHP. To reduce the computational complexity associated with large-scale problems, they proposed several prioritization-based heuristics.

Airspace Allocation

Optimization models and algorithms that address en route capacity constraints treat the airspace system as a multiple origin–destination network along which traffic flow must be assigned. As such, these models often incorporate the network effects (or delay propagation) of arrival delays of flights that are not considered in the GHP.

Deterministic optimization models addressing en route capacity constraints were first formulated as a multicommodity network-flow problem by Helme [23]. These models deal with aggregate flows instead of individual flights, and aim to determine an optimal assignment of ground and airborne delays in an air transportation network; that is, rerouting decisions are not considered. Disaggregate deterministic 0-1 integer programming models for deciding ground and airborne holding of individual flights when faced with airport and airspace capacity constraints were formulated by Lindsay *et al.* [24]. The proposed model, which was named the *Time Assignment Model* (TAM), decides on the temporal and spatial location of each aircraft, given a set of capacity constraints on national airspace system (NAS) resources. The input parameters are the origin and destination airports, a set of en route fixes each aircraft must fly over, as well as the time-varying capacity profile of each of these airspace elements. More recently, Bertsimas and Stock Patterson [25]

presented a deterministic 0-1 IP model to solve a similar problem. For each aircraft, a predetermined set of en route sectors is specified as the route between its origin and destination. The model decides on the departure time and sector occupancy time of each aircraft. Bertsimas and Stock-Patterson showed that their formulation is NP-hard. In many practical cases however the LP relaxation of their IP yields integer optimal solutions, and hence their model is considered to be computationally efficient in practice. The Bertsimas–Patterson model can be summarized as follows.

Let K denote a set of airports and F be the set of flights scheduled between those airports, and let J denote the set of en route sectors. Let Ψ denote the set of pairs of flights that are continued, that is, $\Psi := \{(f, f') : f' \text{ is continued by } f\}$. Let the planning horizon be divided into T time intervals of equal duration. For a given flight f , let N_f denote the number of resources (i.e., sectors and airports), and $P(f, i)$, $1 \leq i \leq N_f$, denote the i th resource along flight f 's path. Note that $P(f, 1)$ and $P(f, N_f)$ represent the departure and arrival airports respectively. Depending on the trajectory, each flight is required to spend a minimum number of time units, l_{fj} , in a sector j that lies along its flight path. Let the capacity of resources during a time interval t be denoted as follows: $D_k(t)$ equals the departure capacity of airport $k \in K$, $A_k(t)$ the arrival capacity of k , and $S_j(t)$ the sector capacity (i.e., the number of aircraft allowed to be present) in sector $j \in J$. The flight-specific scheduled times and delay costs are denoted as follows: d_f , a_f , and s_f are the scheduled departure, arrival, and turnaround times (the minimum ground time for an aircraft between flights) respectively, while c_f^g and c_f^a denote the unit costs of delaying a flight on the ground and in the air.

The binary decision variables, which are nondecreasing, are defined as follows:

$$X_{ft}^j = \begin{cases} 1, & \text{if flight } f \text{ arrives at sector } j \text{ by} \\ & \text{time } t; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

To reduce the size of the formulation, Bertsimas and Patterson define a feasible time

window for each flight that establishes when that flight can occupy a resource along its flight path. The feasible time periods, for a flight f to be present in sector j , are represented by a set $T_f^j, j \in P(f), 1 \leq i \leq N_f$. On the basis the decision variables, the total ground and airborne delays of a flight are given by the following expressions:

$$g_f = \sum_{t \in T_f^k, k=P(f,1)} t(w_{ft}^k - w_{f,t-1}^k) - d_f; \text{ and}$$

$$r_f = \sum_{t \in T_f^k, k=P(f,N_f)} t(w_{ft}^k - w_{f,t-1}^k) - a_f - g_f.$$

The objective function and the set of constraints are defined as follows:

$$\text{Min} \sum_{f \in F} (c_f^g g_f + c_f^a r_f)$$

subject to

$$\sum_{f: P(f,1)=k} (w_{ft}^k - w_{f,t-1}^k) \leq D_k(t) \quad \forall k \in K, t \in \{1, \dots, T\} \quad (13)$$

$$\sum_{f: P(f,N_f)=k} (w_{ft}^k - w_{f,t-1}^k) \leq A_k(t) \quad \forall k \in K, t \in \{1, \dots, T\} \quad (14)$$

$$\sum_{f: P(f,i)=j, P(f,i+1)=j'} (w_{ft}^j - w_{f,t-1}^{j'}) \leq S_j(t) \quad \forall j \in J, t \in \{1, \dots, T\} \quad (15)$$

$$w_{f,t+l_{fj}}^{j'} - w_{ft}^j \leq 0 \quad \begin{cases} f \in F, t \in T_f^j, j = P(f,i), \\ j' = P(f,i+1), i < N_f \end{cases} \quad (16)$$

$$w_{ft}^k - w_{f',t-s_{f'}}^k \leq 0 \quad \begin{cases} (f',f) \in \Psi, t \in T_{f'}^k, \\ k = P(f,1) = P(f',N_{f'}) \end{cases} \quad (17)$$

$$w_{ft}^j - w_{f,t-1}^j \geq 0 \quad \begin{cases} \forall f \in F, j \in P(f,i), \\ 1 \leq i \leq N_f, t \in T_f^j \end{cases} \quad (18)$$

$$w_{ft}^j \in \{0, 1\} \quad \forall f \in F, j \in J, t \in \{1, \dots, T\}. \quad (19)$$

The objective function minimizes the total cost of flight delays. The set of constraints are classified into two categories: capacity constraints (13–15) and connectivity constraints

(16–18). The capacity constraints ensure that the flow is bounded by the capacities of each resource in the system—airports and sectors. For example, constraint set (15) ensures that the total number of flights within a sector during any time interval does not exceed the sector capacity during that time period. Within the connectivity constraints, there are two subcategories: sector and flight connectivity. The sector connectivity constraints (16) ensure that each flight passes through the proper sequence of sectors in its route between origin and destination airports. The flight connectivity constraints (17) ensure that an aircraft must spend a minimum “turnaround” time at an airport before it can depart on its subsequent leg. Constraint set (18) ensures that the decision variables are nondecreasing, while Equation (19) ensures that they are binary.

Variants and Extensions. The Bertsimas–Patterson formulation allows several important variants and extensions. If sector capacity constraints are removed, for example, the formulation corresponds to a multiairport ground holding problem (MAGHP). MAGHPs [26–28] consider a network of airports and optimize the ground delay assignment to various flights, so that delay on a given flight segment can propagate to downstream segments flown by the same aircraft. At the same time, multiple connections at a hub airport can also be addressed in the Bertsimas–Stock formulation by modifying flight connectivity constraints, while interdependence between arrival and departure capacity constraints can be captured using a notion of capacity envelopes. Another interesting alternative is proposed by Lulli and Odoni [29], who introduce a more macroscopic version of the Bertsimas–Stock formulation that omits some of its details (i.e., speed control, en route airborne holding). Lulli and Odoni argue that the resulting model is particularly appropriate for ATFM in Europe, where congestion on en route sectors is common and much more prevalent than in the United States. Using their model, the authors show that fundamental conflicts may arise between efficiency and equity, and

illustrate the potential benefits of selective airborne delay assignments.

In addition, the incorporation of rerouting decisions presents another important research direction. Bertsimas and Patterson present a more aggregate model (similar to the approach by Helme [23]) that addresses routing as well as ground and airborne holding decisions. Their model is formulated as a dynamic, multicommodity, integer network-flow problem with certain side constraints. Aggregate flows are generated by solving a Lagrangian relaxation of the LP, in which the capacity constraints are relaxed into the objective function. Subsequently, a randomized rounding heuristic is applied to decompose the aggregate flows into a collection of individual flight paths. Finally an integer packing problem is solved to obtain feasible, and near-optimal, flight routes. Another important alternative is proposed by Bertsimas *et al.* [30], who extend the Bertsimas–Patterson formulation by allowing for rerouting decisions. Specifically, they propose a formulation where reroute options are represented using a compact set of additional constraints. Their experiments indicate that the resulting models can be solved efficiently for realistic large-scale problem instances. At a more microscopic level, routing decisions have also been considered by Sherali *et al.* [31,32]; their models consider smaller regions of airspace, but incorporate more detailed representation of flight trajectories and sector capacities.

Applying the Bertsimas–Patterson model for large-scale problems in ATM can pose serious computational challenges. Several variants of the model that address these computational issues have also been proposed in the literature. Rios [33] applied the Dantzig–Wolfe decomposition technique to solve the Bertsimas model. Such decomposition techniques are particularly beneficial when the subproblems are solved simultaneously in multiple processors. In another study [34], the decision variables were limited to assigning flight departure times only. No airborne holding was allowed. Airborne holding, and subsequent rerouting was invoked when aircraft approached weather-impacted regions. Thus the model proposed

by Grabbe *et al.* [34] accounts for uncertainty in weather forecast by delaying the rerouting and airborne holding decisions until further in time when precise information becomes available. Several other recent papers [35,36], also discuss stochastic models for airspace capacity allocation, and we expect this to be a fertile area for future research.

Finally, we also note the so-called Eulerian models [37–39] that have been proposed for ATFM. These models spatially aggregate air traffic count to generate models for air traffic flow in one-dimensional control volumes. The complexity of aggregate models depends on the number of such control volumes, which is typically much less than the total aircraft count. A linear, dynamic systems model, proposed in Sridhar *et al.* [39], was developed using the number of departures from various centers, and by estimating the flows between adjacent centers using historical air traffic data. Subsequently, an Eulerian–Lagrangian model based on multicommodity network flow was developed based on historical traffic data [40]. The Eulerian flow model is similar to that in Sridhar *et al.* [39]. The problem of controlling flow was posed as an IP in which the dynamic flow model represents the set of constraints.

CHALLENGES AND RESEARCH DIRECTIONS

While the current body of research in ATFM is certainly large and varied, it is nevertheless important to note that operations research models have generally not seen a widespread adoption within ATFM practice. Whereas the field of operations research has had a critical impact within other areas of the airline industry (e.g., revenue management, airline schedule planning, and crew rostering), its applications to flow management have been more isolated and, generally speaking, of a more limited scope. It is important to note, however, that the context in which ATFM operates makes it difficult to adopt operation research models on a large scale. Because ATFM is concerned with safety and day-to-day operations, the environment will naturally be more conservative and slower to adopt new models than a

for-profit business operating in a competitive environment. In a system this complex, an incremental and evolutionary approach to adopting new technologies has long been preferred. At the same time, it should also be noted that the models' intensive data requirements together with sometimes restrictive assumptions also complicates the implementation of optimization models in practice. The large body of work on stochastic GHPs, for instance, generally uses a scenario tree as an input. The development of appropriate decision trees, however, presents a formidable task in and by itself, and has received attention only recently [41,42]. Moreover, the model's intended users might find it hard to determine appropriate values for models parameters that—while mathematically convenient—do not correspond well with the manner in which they make their decisions. Models for the GHP have traditionally used parameters to represent the relative “cost” of ground and airborne holding. It can be difficult for the service provider to determine values for such parameters, or understand how varying these parameters will impact results. Thus, even though we believe the research on ATFM has made significant progress and yielded important new ideas, it is safe to say that the field as a whole is still in its early stages.

As such, we believe that there is a clear need and opportunity for further ATFM research. Over the next two decades, the demand for air traffic in the United States is expected to grow to two or three times its current level [43]. Given that the air transportation system can barely manage current demand levels, stakeholders in the system are actively pursuing ways to accommodate this future growth. These developments are wide-ranging, and—in the United States—are organized in an integrated plan that is known as the *Next Generation Air Transportation System (NextGen)*. The NextGen concept envisions a fundamental departure from the current approach to air transportation operations that provides a common framework for safety, efficiency, security, and environmental concerns. And, while the NextGen vision will undoubtedly undergo changes and revisions in the years to

come, we believe that its key capabilities and fundamental characteristics offer numerous opportunities for operations research analyses and modeling.

One important aspect of these models will be the ability to account for user behavior and response. A potential area of research in this area is the application of market-based mechanisms, both in the medium and short term, for managing the air transportation system's resources. In the medium-term, for instance, airlines could bid for and own a certain proportion of system resources. Subsequently, they might be able to trade the resources they own with other users on a daily basis. Such an approach would require users to develop models that value their resources, and support decisions related to resource trading on a secondary market. The service provider, on the other hand, would have to design this “marketplace,” and provide a platform for resource trading. Examples of initial research in this area include the use of auctions to assign airport arrival slots [1] and the use of slot trading during GDPs [44].

In addition, the use of new models for disruption management also offers several possibilities. Given that most disruptions are due to bad weather conditions, further development related to the above-mentioned decision models that integrate uncertainty forms one important research direction in this area. Another promising area is the development of decision support models that will facilitate contingency planning, to increase responsiveness under changing conditions. Such a framework also needs corresponding models at the user side, to allow for fully integrated airline recovery methods that can evaluate and establish user preferences under the various potential scenarios.

REFERENCES

1. Ball M, Donohue G, Hoffman K. Auctions for the safe, efficient, and equitable allocation of airspace system resources. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge: MIT Press; 2005. pp. 507–538.
2. Sridhar B, Grabbe SR, Mukherjee A. Modeling and optimization in traffic flow management. *Proc IEEE* 2008;96(12):2060–2080.

3. Hoffman R, Mukherjee A, Vossen T. Air traffic flow management. Working Paper, Leeds School of Business. Available at http://leeds.colorado.edu/Faculty_and_Research/interior.aspx?id=5448, 2009.
4. Odoni A. The flow management problem in air traffic control. In: Odoni AR, Bianco L, Szego G, editors. Flow control of congested networks. Berlin: Springer-Verlag; 1987. pp. 269–288.
5. Andreatta G, Romanin-Jacur G. Aircraft flow management under congestion. *Transport Sci* 1987;21:249–253.
6. Richetta O, Odoni A. Solving optimally the static ground holding policy problem in air traffic control. *Transport Sci* 1993;24: 228–238.
7. Kotnyek B, Richetta O. Equitable models for the stochastic ground-holding problem under collaborative decision making. *Transport Sci* 2006;40:133–146.
8. Richetta O, Odoni A. Dynamic solution to the ground-holding problem in air traffic control. *Transport Res Part A, Policy Practice* 1994;28:167–185.
9. Mukherjee A, Hansen M. A dynamic stochastic model for the single airport ground holding problem. *Transport Sci* 2007;41:444–456.
10. Mukherjee A. Dynamic stochastic optimization models for air traffic flow management. PhD thesis, University of California at Berkeley, 2004.
11. Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
12. Gilbo E. Airport capacity: representation, estimation, optimization. *IEEE Trans Control Syst Technol* 1993;1:144–154.
13. Gilbo E. Optimizing airport capacity utilization in air traffic flow management subject to constraints at arrival and departure fixes. *IEEE Trans Control Syst Technol* 1997;5:490–503.
14. Dear RG, Sherif YS. An algorithm for computer assisted sequencing and scheduling of terminal area operations. *Transport Res* 1991;25(2):129–139.
15. Beasley JE, Krishnamoorthy M, Sharaiha YM, *et al.* Scheduling aircraft landings—the static case. *Transport Sci* 2000;34(2): 180–197.
16. Bianco L, Dellà-Olmo P, Giordani S. Coordination of traffic flows in the TMA. In: Bianco L, Dellà-Olmo P, editors. New concepts and methods in air traffic management Berlin: Springer; 2001; pp.95–124.
17. Vossen T, Ball M. Optimization and mediated bartering models for ground delay programs. *Naval Res Logistics* 2006; 53(1):75–90.
18. Ball M, Hoffman R, Odoni A, *et al.* A stochastic integer program with dual network structure and its application to the ground-holding problem. *Oper Res* 2003;51:167–171.
19. Vossen T, Ball M, Hoffman R, *et al.* A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. *Air Traffic Control Q* 2003;11(4):277–292.
20. Hoffman R, Ball MO, Mukherjee A. Ration-by-distance with equity guarantees: A new approach to ground delay program planning and control. Proceedings of the 7th USA/Europe Air Traffic Management R&D Seminar; Barcelona, Spain. 2007.
21. Liu B. Managing uncertainty in the single airport ground holding problem using scenario-based and scenario-free approaches. PhD thesis, University of California, Berkeley, 2006.
22. Liu B, Hansen M. Scenario-free sequential decision model for the single airport ground holding problem. Proceedings of the 7th USA/Europe Air Traffic Management R&D Seminar; Barcelona, Spain. 2007.
23. Helme MP. Reducing air traffic delay in a space-time network. *IEEE Int Conf Syst, Man, Cybernetics* 1992;1:236–242.
24. Lindsay K, Boyd E, Burlingame R. Traffic flow management modeling with the time assignment model. *Air Traffic Control Q* 1994; 1:255–276.
25. Bertsimas D, Stock Patterson S. The air traffic flow management problem with enroute capacities. *Oper Res* 1998;46:406–422.
26. Vranas P, Bertsimas D, Odoni A. The multi-airport ground holding problem in air traffic control. *Oper Res* 1994;42:249–261.
27. Vranas P, Bertsimas D, Odoni A. Dynamic ground-holding policies for a network of airports. *Transport Sci* 1994;28:275–291.
28. Navazio L, Romanin-Jacur G. Multiple connections multi-airport ground holding problem: Models and algorithms. *Transport Sci* 1998;32:268–276.
29. Lulli G, Odoni A. The european air traffic management problem. *Transport Sci* 2007; 41:431–443.
30. Bertsimas D, Lulli G, Odoni A. An integer optimization approach to large-scale air traffic flow management. Proceedings of the 13th International Conference on Integer

- Programming and Combinatorial Optimization; Bertinoro, Italy. 2008; pp.34–46.
31. Sherali HD, Staats RW, Trani AA. An airspace planning and collaborative decision-making model: part i-probabilistic conflicts, workload, and equity considerations. *Transport Sci* 2003; 37:434–456.
 32. Sherali HD, Staats RW, Trani AA. An airspace-planning and collaborative decision-making model: Part ii-cost model, data considerations, and computations. *Transport Sci* 2006;40:147–164.
 33. Rios J, Ross K. Massively parallel Dantzig-Wolfe decomposition technique applied to traffic flow scheduling. *Proceedings of the AIAA Guidance, Navigation and Control Conference*, Chicago, IL; 2009.
 34. Grabbe S, Sridhar B, Mukherjee A. Sequential traffic flow optimization with tactical flight control heuristics. *AIAA J Guidance, Control, Dynamics* 2009;32(3):810–820.
 35. Ganji M, Lovell D, Ball MO. Resource allocation in flow-constrained areas with stochastic termination times considering both optimistic and pessimistic reroutes. *Proceedings of the 8th USA/Europe Air Traffic Management R&D Seminar*; Napa (CA). 2009.
 36. Clarke JP, Solak S, Chang Y-H, *et al.* Air traffic flow management in the presence of uncertainty. *Proceedings of the 8th USA/Europe Air Traffic Management R&D Seminar*; Napa (CA). 2009.
 37. Menon PK, Sweriduk GD, Bilimoria KD. New approach for modeling, analysis, and control of air traffic flow. *AIAA J Guidance, Control, Dynamics* 2008;27(5):737–744.
 38. Bayen AM, Raffard RL, Tomlin CJ. Adjoint-based control of a new Eulerian network model for air traffic flow. *IEEE Trans Control Syst Technol* 2006;15(5).
 39. Sridhar B, Soni T, Sheth K, *et al.* An aggregate flow model for air traffic management. *AIAA J Guidance, Control, Dynamics* 2006; 29(4).
 40. Sun D, Bayen AM. Multi-commodity Eulerian-Lagrangian large-capacity cell transmission model for en route traffic. *AIAA J Guidance, Control, Dynamics* 2004;31(3):616–628.
 41. Innis T, Ball MO. Estimating one-parameter airport arrival capacity distributions for air traffic flow management. *Air Traffic Control Q* 2004;12:223–252.
 42. Hansen M, Liu B, Mukherjee A. Scenario-based air traffic flow management: From theory to practice. Technical report, Technical report. Berkeley: University of California; 2006.
 43. <http://www.jpdo.gov>. Accessed in 2010.
 44. Vossen TWM, Ball MO. Slot trading opportunities in collaborative ground delay programs. *Transport Sci* 2006;40(1):29–54.

AIRLINE RESOURCE SCHEDULING

AMY COHN
MARCIAL LAPP
Industrial and Operations
Engineering, University of
Michigan, Ann Arbor,
Michigan

INTRODUCTION

Passenger aviation is critical to today's society, with passengers relying on airlines (carriers) to provide safe, reliable, and affordable travel for both business and leisure. In 2009, more than 9.9 million flights originated and/or terminated in the United States alone, carrying more than 764 million passengers [1]. Every one of these flights required the coordinated utilization of many shared resources including aircraft, crews (cockpit, cabin, and ground), taxiways and runways, the airspace, and more. In some cases, resources are shared across multiple flights within a single company (e.g., aircraft, crews) while other resources (such as runways and the airspace) must be shared across airlines, adding further complexity. This sharing of resources, along with the associated underlying network structure of an airline, results in significant coordination challenges.

The operations research (OR) community has long played an active role in virtually all aspects of the airline industry, helping to plan, schedule, coordinate, and operate it. In the past decade, this role has been particularly important. Major challenges such as SARS, the US terrorist attacks of 9/11, the 2008 spikes in fuel prices, and a global economic downturn have made it increasingly important that airlines utilize resources efficiently. To accommodate, the OR community has expanded its focus to include topics such as robust planning, integrated planning, and enhanced recovery techniques.

In the process of solving these challenging airline problems, the OR community has also made broader contributions. Specifically, several of the modeling and algorithmic techniques developed to solve airline planning problems have applicability to a broad class of other application areas as well.

This article therefore has two purposes. The first is to introduce readers new to the field of airline OR to the problems that have been solved and the problems currently under investigation, and to provide initial references to some of the key literature. The second is to review some of the key modeling and algorithmic contributions, which have relevance in many other fields beyond aviation.

OPERATIONS RESEARCH PROBLEMS IN THE AIRLINE INDUSTRY

Within passenger aviation, there is a vast array of complex decisions to be made, ranging from aircraft design and airport construction to the control of the airspace to airline planning and scheduling. Similarly, there is a wide range of decision makers, including carriers, government regulators, airport authorities, and passengers. We focus here on resource scheduling from the perspective of the carriers. Resource scheduling problems range in timescale from years, such as the decision to purchase an aircraft, to minutes, such as deciding how to reaccommodate passengers who have missed their connections. These problems cover many different resources including aircraft, pilots, flight attendants, gates, baggage, maintenance workers and facilities, and, of course, passengers. In addition, they must all address underlying uncertainty, including variability in demand, inclement weather, and unexpected maintenance issues. Furthermore, these problems must all address the system complexity associated with the underlying network structure of airline systems and the sharing of a finite set of resources.

Within the set of carrier resource scheduling problems, we focus primarily on the established literature in *fleet assignment*, *crew scheduling*, and *aircraft routing*, as well as the emerging literature on integrated planning and robust planning. We also briefly touch upon important future areas of research. Before doing so, we first briefly highlight some important areas of airline OR. These references are not intended to be an exhaustive survey, but rather to give a sense of the wide range of work that has been done and some initial sources for the interested reader.

Schedule generation

Airlines work on the *schedule generation* process a year or more in advance of the day of operation, predicting the demand for flights between origin–destination (O–D) pairs and subsequently deciding which flights to offer and with what frequency, as well as how to partner with other airlines through code-sharing and partnering agreements. Gopalan and Talluri [2] provide a survey paper identifying various schedule generation strategies. Early work by Daskin and Panayotopoulos [3] provides a Lagrangian relaxation to solve the assignment of aircraft to routes. Warburg *et al.* [4] and Jiang and Barnhart [5] provide a dynamic scheduling model that uses changes in the fleet assignment and minor flight retimings to update the schedule as booking data becomes available.

Revenue management, pricing, and passenger flow models

Revenue management and *pricing* problems focus on strategies to maximize profits from ticket sales. This is an evolving field of study, especially as new purchasing channels and new information systems become available. Related work can be found in McGill van Ryzin [6], Talluri and van Ryzin [7], Belobaba *et al.* [8], Fiig *et al.* [9], Lardeux [10]. Acting as a bridge between revenue management and fleet assignment, passenger flow

modeling [11] finds an optimal (i.e., revenue-maximizing) selection from a set of candidate itineraries given a fixed set of flight capacities. This solution, idealized in the sense that it assumes that an airline has complete control over which itineraries are purchased, provides a bound on the revenue that can be generated from a given fleet schedule. More recently, Dumas and Soumis [12] have incorporated uncertain demand and spill estimates within passenger flow modeling.

Demand driven dispatch and dynamic fleet

Even though airline plans are set far in the future, they are subject to uncertainty until the day of operations. Early work by Berge and Hopperstad [13] suggests that airlines can benefit by dynamically adjusting their fleet assignment to better match aircraft capacity to passenger demand when updated information about passenger bookings is obtained. For a more recent discussion of this topic, see Shebalov [14] and Tekiner *et al.* [15]. As an extension to demand driven dispatch, two recent articles by Warburg *et al.* [4] and Jiang and Barnhart [5] explore the idea of not only modifying the fleet assignment but also slightly altering the flight schedule itself to increase the revenue in response to evolving information about passenger demand.

Recovery

Airline plans are rarely, if ever, executed as designed. Unexpected disruptions such as inclement weather and unplanned maintenance issues often lead to flight delays. The inherent underlying network structure is such that these delays can further propagate to cause other delays (e.g., a down-stream flight delayed owing to the delay of its incoming aircraft). The *recovery* problem focuses on how to quickly return to the original plan, re-accommodating passengers, crews, aircraft, and more, often through the use

of heuristics and rules-of-thumb. On the other hand, resource planning problems focus more heavily on profit optimization and have greater flexibility in their computational solution time. Recent research in this important and challenging area includes the works of Eggenberg *et al.* [16], Abdelghany *et al.* [17–19], and Kohl *et al.* [20].

Airport operations

- *Gate assignment* The *gate assignment* problem determines which terminal gates are assigned to which inbound/outbound flights. The objectives that have been considered in the literature include minimizing walking distance for connecting passengers or minimizing the total number of missed passenger connections. Related work can be found in Mangoubi and Mathaisel [21], Bihr [22], Haghani and Chen [23], and Bolat [24] and a recent survey paper by Dorndorf *et al.* [25].
- *Boarding strategies* Boarding strategies vary among airlines. In most cases, the objective of a successful boarding strategy is to minimize the overall boarding time on a full aircraft. For an example of how OR is used in developing and analyzing boarding strategies, see van den Briel *et al.* [26].
- *Baggage handling* Although not as visible as passengers, baggage handling also presents many challenges for airline operations. Ensuring that baggage is transported from origin to destination, often with connections in between, presents an opportunity for OR contributions. For an example on this research, see Abdelghany *et al.* [27].
- *Check-in staffing* Scheduling of ground staff has also been of recent interest to the OR community. Stollatz [28] examines the operational workforce plan at check-in counters.

On-demand air transportation

In recent years, on-demand air transportation has begun to evolve as a new

business model for air travel. Constructing and evaluating the networks and operating practices of such companies yields many interesting OR problems. See Espinoza *et al.* [29,30] for two recent papers in this area.

Congestion pricing and slot auctions

Certain airports exhibit very high levels of congestion, often because of both very high demand for travel into and out of that area and also limited geographical opportunity for airport expansion. The volume of traffic at these airports can lead to significant congestion-based delays, which can in turn propagate throughout the aviation system. Congestion pricing [31] and slot auctions [32] are two examples of external influences on how airlines choose to generate their flight schedules. Both of these have benefited from OR tools for analysis and assessment of the impact of such approaches. For some recent work on this topic, we refer the reader to the thesis of Harsha [33].

Analysis of delays

The OR community has also conducted significant empirical and quantitative analysis on passenger airline performance. Examples of this include AhmadBeygi *et al.* [34], Stollatz [35], Baik and Trani [36], and Balakrishna *et al.* [37].

We close this section by noting some valuable textbooks focusing on the airline industry and airline decision making: [38–40].

RESOURCE SCHEDULING PROBLEMS IN PASSENGER AVIATION

Within passenger aviation, the three resource planning problems that have received the greatest attention from the OR community (and achieved the greatest successes) are *fleet assignment*, *aircraft routing*, and *crew scheduling*. As such, we focus our primary attention here.

Note that most airlines typically offer between one and four flight schedules per year. For example, they may offer a winter

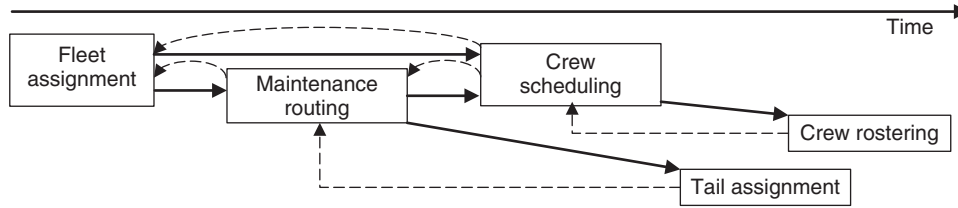


Figure 1. Resource allocation solution approach.

schedule and a summer schedule. Within each schedule, there is usually a consistent pattern that repeats weekly, with many domestic flights repeating daily. Approximately six months to a year before a new schedule begins operations (varying by carrier), the solving of fleet assignment, aircraft routing, and crew scheduling typically begins, with the three problems solved in sequence. The general time-line of solving these airline resource allocation problems can be seen in Fig. 1. In this figure, the solid lines show initial flow of information, with output from one problem providing input for the next. The dashed lines illustrate a feedback loop, where information from a later problem is used to revise the solution to an earlier problem.

First, the fleet assignment problem is solved, in which each flight is assigned a specific aircraft type. The goal is to maximize the projected revenue minus the operating cost associated with the assignments, subject to *cover constraints* (every flight must have exactly one fleet type), *balance constraints* (for each fleet type, the flow into an airport must equal the flow out), and *count constraints* (you cannot use more aircraft of a given type than you have in your fleet). Note that this problem results in a partitioning of the flights by subfleet. We can then solve a separate (and independent) aircraft routing and crew scheduling problem for each fleet type.

The goal of the aircraft routing problem is to build *lines-of-flight*, that is, sequences of flights to be flown by individual aircraft (these lines-of-flight will subsequently be assigned to specific aircraft in the tail assignment problem). There are two primary concerns when establishing lines-of-flight.

The first is to meet strict *maintenance requirements* as required by the Federal Aviation Administration (FAA) (in the United States) or other governing bodies. In order to ensure that it is possible to meet these scheduled maintenance requirements, lines-of-flight are created that start and end at *maintenance stations* (airports that have the capability to perform routine maintenance) without exceeding maintenance limits. The second goal in building lines-of-flight is to establish *flight connections*, that is, to identify pairs of sequential flights that will share a common aircraft. This has benefits from both the revenue side (charging a premium on desirable itineraries for flight pairs that do not require passengers to change aircraft) and the crew scheduling side (identifying good opportunities to allow crew to remain with an aircraft over multiple flights, which reduces the propagation of delays).

Once the aircraft routing problem has been solved, the crew scheduling problem can be addressed. Like aircraft routing, a separate crew scheduling problem is solved for each aircraft type, since pilots are trained to fly a specific aircraft type. Aircraft routing and crew scheduling can largely be solved independently of one another, with the exception of *short connects*. These are pairs of flights with a *tight turn*, that is, very little time between the arrival of the first flight and the departure of the second. Thus, it is only possible for a crew to be assigned to both of these flights if the flights were assigned to a common aircraft in the aircraft routing solution. In addition, because it is desirable to keep a crew with the same aircraft, the aircraft routing problem acts as a key input to the crew scheduling problem.

Similar to the aircraft routing problem, in which sequences of flights are constructed to be flown by a common aircraft, the main component of crew scheduling, the *crew pairing* problem, builds sequences of flights (*pairings*) to be flown by an individual crew (the specific crews are then matched to the pairings in a *crew rostering* or *bidline* problem). A pairing is a multiday sequence of flights that are not only sequential in space and time but also comply with all federally mandated rest requirements and duty limitations. The goal of the crew pairing problem is to construct the least-cost set of pairings such that all flights are covered exactly once.

Observe that there is significant interdependence between all of these problems. In particular, the fleet assignment substantially impacts the feasible regions for the aircraft routing and crew scheduling problems by partitioning the flights into independent sets. As such, this raises the question of whether higher quality solutions could be found by solving the three problems simultaneously. In fact, there is such benefit, but it comes at the cost of substantially increased computational challenges. A sequential approach has been used in the past, primarily for reasons of tractability. In contrast, a sequential approach may not only lead to suboptimal solutions, but can in fact result in infeasibility. As a result, carriers typically perform multiple sequential iterations with feedback between the three problems.

Once these three problems have been solved to satisfaction, typically months before the schedule's start date, the shorter term problems of tail assignment and crew rostering are conducted—typically on a repeated, rolling horizon throughout the duration of the schedule.

Finally, fleeting, routing, and crew scheduling decisions all continue to be made even at the operational level, when recovery decisions must be made in response to disruptions. In these cases, the problem constraints are largely the same, but the goals are often quite different (for example, instead of focusing on optimizing profits, the goal may be to return to the planned schedule

as quickly as possible) and the permissible run time to find solutions is much tighter, leading to a focus on fast-running heuristics and rules-of-thumb over optimization-based approaches.

FLEET ASSIGNMENT

Time-Space Networks

In the fleet assignment problem (FAM), we want to assign an aircraft fleet type to each flight in the schedule. The goal is to maximize profits subject to the cover, balance, and count constraints described in the section titled “Basic Fleet Assignment Model (FAM).” Before presenting formulations for this problem, we introduce the notion of *time-space networks* [41]. In a time-space network, each node in the network represents a physical location along with a specific moment in time. The arc connecting two nodes in such a network then represents a transition in both space and time. Such networks can be very powerful in a variety of applications, including but not limited to passenger aviation resource planning [42].

In airline planning problems, we often make use of a time-space network in which we have a *time-line* for each station (i.e., airport). A node on this time-line indicates a *flight event* at that station, that is, either an arrival or departure. Each flight then has two nodes, one on the time-line of its origin airport, at the time of its departure, and one at its destination airport, at the time of its arrival. We refer to the arc connecting these two nodes as a *flight arc*. In addition, we create a *ground arc* from each node on a time-line to the next node in time on that same time-line. These arcs represent aircraft remaining at the station in between the flight events. Figure 2 represents such a time-space network for two stations and two flights.

Basic Fleet Assignment Model (FAM)

Given the concept of a time-space flight network, we now present an integer programming (*IP*) formulation for FAM, as originally formulated by Hane *et al.* [41] and Jacobs *et al.* [43].

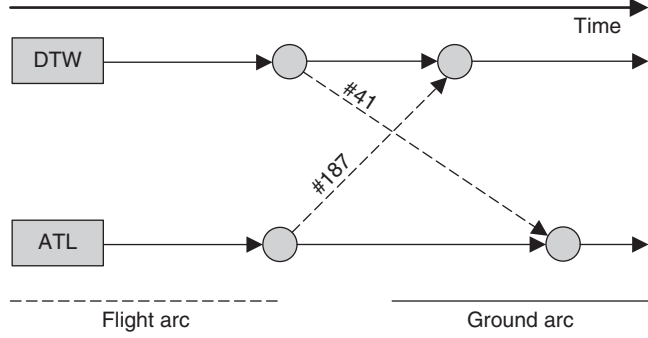


Figure 2. Example of a time-space network.

Sets

- F the set fleet types.
- L the set of flight legs.
- T the set of nodes, that is, flight events.
- S the set of stations.

Parameters

- c_{lf} the profit gained when leg l is assigned to fleet type f , $\forall l \in L, \forall f \in F$.
- N_f the number of aircraft of type f in the fleet, $\forall f \in F$.
- $C \subset L$ the set of flight legs that cross a count-line (e.g., 3:00 am)
- $I(f, s, t)$ the set of flight legs that are *inbound* to (f, s, t) , $\forall f \in F, \forall s \in S, \forall t \in T$.
- $O(f, s, t)$ the set of flight legs that are *outbound* from (f, s, t) , $\forall f \in F, \forall s \in S, \forall t \in T$.

Decision Variables

- x_{lf} a binary variable that is 1 if aircraft type f is assigned to leg l , $\forall l \in L, \forall f \in F$.
- y_{fst+} the number of aircraft of type f on the ground at station s just *after* flight event t , $\forall f \in F, \forall s \in S, \forall t \in T$.
- y_{fst-} the number of aircraft of type f on the ground at station s just *before* flight event t , $\forall f \in F, \forall s \in S, \forall t \in T$.

Objective:

$$\max \sum_{l \in L} \sum_{f \in F} c_{lf} x_{lf} \quad (1)$$

Subject to:

$$\sum_{f \in F} x_{lf} = 1 \quad \forall l \in L \quad (2)$$

$$y_{fst-} + \sum_{l \in I(f, s, t)} x_{lf} - \sum_{l \in O(f, s, t)} x_{lf} - y_{fst+} = 0 \quad \forall f \in F, \forall s \in S, \forall t \in T \quad (3)$$

$$\sum_{l \in C} x_{lf} + \sum_{s \in S} y_{is0-} \leq N_f \quad \forall f \in F \quad (4)$$

$$x_{lf} \in \{0, 1\} \quad \forall l \in L, \forall f \in F \quad (5)$$

$$y_{fst} \geq 0 \quad \forall f \in F, \forall s \in S, \forall t \in T. \quad (6)$$

Constraint (2) enforces the *cover* constraints, that is, each flight must be covered by exactly one aircraft type. Constraint (3) enforces aircraft *balance*: the total number of aircraft of a given type on the ground at a given station immediately prior to a flight event plus the number of aircraft of that type that land at that event must equal the number of aircraft of that type that leave that station at that event plus the number of aircraft that remain on the ground. These balance constraints, in conjunction with Equation (4), enforce *count*. Specifically, we use the concept of a *count-line* that represents a single point in time (typically a time of low activity, such as 3:00 a.m.). For each aircraft type, we force the number of aircraft of that type assigned to a flight spanning the count time plus the number of aircraft of that type on the ground at that station at the count time to not exceed the number of aircraft of that fleet type available. Because network balance is enforced, if we do not

exceed the fleet count at the count time, then we will not exceed it at any time.

Arc Copies

When assigning fleet types to flights, we would ideally like to match each flight to its optimal aircraft, that is, the one that best trades off between operating cost and capacity (and hence ability to capture revenue). However, such a match is not necessarily feasible for all flights because of the balance and count constraints. In practice, it has been observed that small shifts in the timings of the flight schedule can increase the feasible region of FAM, leading to solutions with reduced cost. In the simplistic example in Fig. 3, we consider two different departure timings for the flight from station DTW to station ATL. By choosing the earlier of the two departures in set $\{A\}$ out of DTW, we achieve coverage of two flights using a single aircraft of a given fleet type. That is, it is possible to cover both the flights from DTW to ATL and then from ATL to DTW with a single aircraft.

To take advantage of these potential benefits, Rexing *et al.* [44] introduced the *fleet assignment problem with time windows*. The idea is to allow small, discrete shifts in time to enable a better fleet assignment. FAM with time windows, as described here, is an example of the integration between fleet assignment and schedule design. Further integration approaches are illustrated in the section titled “Integrated Planning.” To formulate this problem, the time-space network is first modified to contain *arc copies*. Specifically, for each flight, we create

one arc for each possible time when the flight might depart. This is typically limited to a small window (e.g., 15 to 30 min before/after the originally scheduled departure time) so that dramatic changes in potential passenger demand will not be observed. Given this modified network, the basic FAM formulation must also be slightly modified. Specifically, the decision variables now represent choosing not only a fleet type for each flight but also a specific departure time. The objective is then given by Equation (7).

$$\min \sum_{f \in F} \sum_{l \in L} \sum_{n \in N_{lf}} c_{lf} x_{lfn}. \quad (7)$$

We replace constraint (2) with constraint (8), so that the cover constraint now selects not only a fleet type but a departure time as well. We also update the variable definition accordingly.

$$\sum_{f \in F} \sum_{n \in N_{lf}} x_{lfn} = 1 \quad \forall l \in L \quad (8)$$

$$x_{lfn} \in \{0, 1\} \quad \forall l \in L, \forall f \in F, \forall n \in N_{lf}. \quad (9)$$

Both the basic FAM and time windows version depend on the input parameters c_{lf} . In practice, to estimate these cost parameters can be difficult for many reasons, which motivates the need for more advanced FAMs.

Itinerary-Based Fleet Assignment (IFAM)

The objective function of FAM depends on the objective coefficients, c_{lf} , which capture both the cost and revenue component of a fleet assignment. The cost component can be

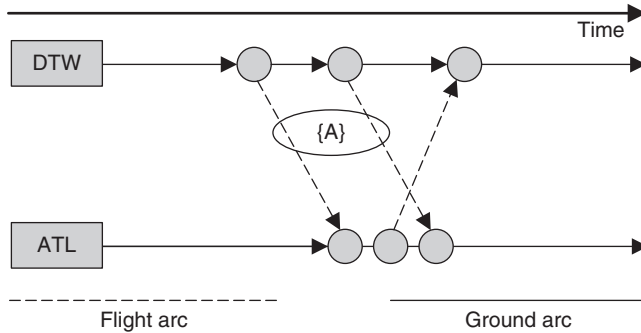


Figure 3. Example of the multiple arc formulation.

fairly straightforward to estimate, but the revenue component is much more difficult. It is usually thought of not as revenue captured but rather potential revenue that is lost, or *spilled*, due to insufficient capacity. For example, if 200 passengers want to buy tickets for a particular flight and that flight is assigned to a fleet type with only 170 seats, then the revenue from 30 passengers is lost. This spill cost is added to the operating cost to determine the coefficients c_{lf} .

There are several challenges associated with calculating spill. The first is the fact that demand is dynamic. Only one fleet type will be chosen for the entire schedule period, but demand will vary daily over this period. An even bigger challenge is the fact that passengers do not just fly individual flights, but often fly multileg itineraries. By only looking at individual legs in FAM, we miss the interdependencies that stem from these itineraries. For example, suppose a passenger wants to fly from Boston to Los Angeles via Chicago. If the basic FAM model is solved, a large aircraft may be assigned to the Boston to Chicago flight with more than enough capacity to meet all demand. If the flight from Chicago to Los Angeles is assigned to an aircraft with inadequate capacity, however, the passenger may be spilled from this flight. In reality, we would lose the revenue of this passenger from both flights; in the model, we would still capture their revenue on the first leg, even though they were spilled from the second.

To address this, Barnhart *et al.* [11] developed the following extended version of FAM, known as *itinerary-based fleet assignment* (IFAM). In this approach, the fleet assignment decisions are augmented with passenger “spill variables” that take into account the demand for each itinerary rather than each flight leg. Specifically, a fleet assignment also implicitly defines the capacity on each flight leg. Given this capacity, IFAM can simultaneously determine the number of passengers spilled (i.e., potential passengers whose revenue is lost due to inadequate capacity) and corresponding revenue lost across the entire itinerary, not on an individual flight leg. Although there are still many challenges with this approach (e.g., it ignores the fact

that passenger purchases occur over a rolling time horizon and cannot be fully controlled by the airline), it nonetheless is a substantial step toward overcoming the limitations of a leg-based approach.

Parameters

$SEATS_f$	the number of seats available on aircraft of fleet type f , $\forall f \in F$.
$\widetilde{fare}_p$	the fare for itinerary p , $\forall p \in P$.
b_p^r	recapture rate from p to r , that is, the fraction of passengers spilled from itinerary p that the airline succeeds in redirecting to itinerary r , $\forall p \in P, \forall r \in P$.
CAP_l	is the capacity of the aircraft assigned to leg l , $\forall l \in L$.
δ_l^p	1 if itinerary p includes flight leg l and is 0 otherwise, $\forall l \in L, \forall p \in P$.

Decision Variables

t_r^p	the number of passengers requesting itinerary p that are redirected by the model to itinerary r , $\forall p \in P, \forall r \in P$.
---------	--

We augment the FAM formulation by replacing the objective with Equation (10) and adding constraints (11)–(13).

Modified objective:

$$\min \sum_{l \in L} \sum_{f \in F} c_{lf} x_{lf} + \sum_{p \in P} \sum_{r \in P} (\widetilde{fare}_p - b_p^r \widetilde{fare}_r) t_r^p \quad (10)$$

Additional constraints:

$$\begin{aligned} \sum_{f \in F} SEATS_f x_{lf} + \sum_{p \in P} \sum_{r \in P} \sigma_l^p t_r^p \\ - \sum_{r \in P} \sum_{p \in P} \sigma_l^p b_p^r t_r^p \geq Q_l - CAP_l \quad \forall l \in L \end{aligned} \quad (11)$$

$$\sum_{r \in P} t_r^p \leq D_p \quad \forall p \in P \quad (12)$$

$$t_r^p \geq 0 \quad \forall r \in P. \quad (13)$$

In the augmented model, we not only assign fleet types to flights (thereby determining the capacity on each leg) but also choose the number of passengers to assign

to each itinerary through constraint set (11). We note here that the parameters (SEATS_f) indicates the number of seats available on aircraft type f , which is followed by our decision variable, x_{if} . On the right side of this equation, we represent the demand, Q_i , as defined in Equation (14) and subtract from it the available capacity for the particular leg, CAP_l . Finally, Equation (12) ensures that we do not assign more passengers to an itinerary than there is demand, and Equation (13) ensures that we do not assign negative numbers of passengers.

$$Q_i = \sum_{p \in P} \delta_i^p D_p. \quad (14)$$

Most recently, Dumas and Soumis [12] and Dumas *et al.* [45] have provided additional research on how to estimate and model itinerary-based passenger demand and its effect on the quality of FAM solutions.

AIRCRAFT ROUTING

Once flights have been assigned to specific fleet types, the subsequent planning problems can be partitioned into independent sets. For each fleet type, we next solve an *aircraft routing* (AR) problem, as described in Gopalan and Talluri [46] and Clarke *et al.* [47]. The primary goal of AR is to ensure that every aircraft has adequate opportunity to undergo the required routine maintenance. For example, in the United States, an A check must be completed after every 65 flight hours; any aircraft exceeding this limit will be grounded by the FAA. Several months before a new schedule begins, carriers therefore build *lines-of-flight* (LOF). These sequences dictate a consequence series of flights to be flown by a single aircraft. Aircraft routes can then be constructed by connecting LOF that start and end with maintenance events, while ensuring maintenance feasibility over the flights in between.

In constructing OFs, *flight connections* are established as well. When two consecutive flights are flown by a common aircraft, this provides opportunities for improved passenger itineraries and crew schedules (as there is no need to change planes between flights),

and also has implications for gate scheduling and similar operational activities.

Note that at this stage, specific aircraft (also known as *tails* because they are identified by the unique number painted on the tail of the aircraft) are not assigned to the maintenance routes. Although the intent is to repeatedly fly the same routes over the course of the schedule, these routes will not always be flown by the same aircraft. This is in part to balance the utilization of aircraft over the system, but more importantly it is a reflection of the operational deviations that often occur in practice. For example, as illustrated in Fig. 4, suppose flights A and B are scheduled to arrive at the same station at roughly the same time, and then their aircraft will be used for departures C and D, respectively. If A is delayed in arrival, an operational decision may be made to use the aircraft from B for C instead of D, as was originally scheduled (for example, to ensure that passengers on C can make international connections that they would otherwise miss). In doing so, the aircraft routes have been swapped, with the aircraft from A now flying B's route and vice versa. In the process, because the different aircraft have different histories (e.g., one may have already flown more hours since its last maintenance than the other), the new routings may be maintenance infeasible.

Although processes vary substantially by carrier, it is not uncommon for the assignment of specific tails to routes to occur on a rolling horizon five to seven days before the day of operation. Each day these assignments are modified both to add a new day to the end of the horizon and also to modify the existing routes to take into account any changes such as aircraft swaps and unplanned maintenance needs.

For more information, we refer the reader to Gabteni and Grönkvist [48].

Aircraft Routing Models

There are several different ways to model and solve the aircraft routing problem, each with different benefits and challenges and each appropriate for different carriers and different contexts. For example, Gopalan and Talluri [46], Kabbani and Patty [49],

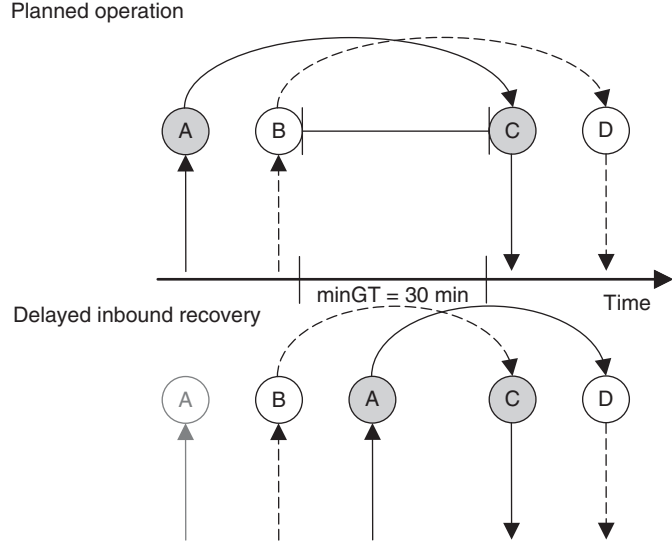


Figure 4. Recovery swaps under disruption.

and Talluri [50] present some of the seminal early work in this area that focuses on challenges such as building aircraft routings that spend every third or fourth night in a maintenance station (i.e., an airport with maintenance capabilities) so that short-term maintenance checks can be completed on a regular basis. This work draws largely from graph theory and the development of Euler tours. Clarke *et al.* [47] pose the problem as being similar to an asymmetric traveling salesman problem (TSP) [51], which they solve using a Lagrangian relaxation [52] and subgradient optimization techniques.

We focus here on two particular modeling approaches, one based on *multicommodity flow (MCF)* formulation techniques, and the other on *string-based* models.

Multicommodity Flow Formulations

More recent formulations of the aircraft routing problem have focused on traditional, linear programming formulations. As detailed in Grönkvist [53], the maintenance-routing problem can be formulated as a variation of the *MCF* problem. In the general *MCF* problem [54,55], we are given a set of commodities, a set of nodes (each with a supply or demand for each commodity), and a set of arcs. The objective is to find the least-cost way

to move commodities across the network from supply to demand while satisfying capacity constraints on the individual arcs.

Sets

F the set of all flights (recall that AR is solved separately for each fleet type).

Parameters

c_{ij} the cost of assigning the connection i to j . This cost often represents the (negative of) potential additional revenue that can be gained by offering this flight connection as a direct flight with no change of planes. In actuality, carriers are usually more concerned with feasibility than optimality when solving the aircraft routing problem.

Decision Variables

x_{ij} a binary variable that indicates if flight i is followed by flight j and 0 otherwise.

Objective:

$$\min \sum_{i \in F} \sum_{j \in F} c_{ij} x_{ij} \quad (15)$$

Subject to:

$$\sum_{j \in F} x_{ij} = 1 \quad \forall i \in F \quad (16)$$

$$\sum_{j \in F} x_{ij} - \sum_{j \in F} x_{ji} = 0 \quad \forall i \in F \quad (17)$$

$$x_{ij} \in \{0, 1\} \quad \forall i \in F, \forall j \in F. \quad (18)$$

To formulate aircraft routing as a variation of MCF, we define a network in which each commodity is an aircraft route, each flight is represented by a node, and an arc exists between each pair of nodes corresponding to flights that form a feasible connection. Constraints (16) require each flight to be covered exactly once (i.e., to be included in exactly one route). Constraints (17) enforce balance. Additional constraints are used to enforce maintenance feasibility as implemented by Grönkvist [53].

String-Based Models

The challenge of the MCF formulation lies in capturing all maintenance requirements. Therefore, as an alternative, several researchers have taken a string-based modeling approach to solve aircraft routing, often in the context of integrating AR with other planning problems [56,57].

In the string-based approach, a variable corresponds to the assignment of a particular route to a complete “string,” that is, a complete LOF. Each string has an associated cost that spans the entire set of assignments in that string. Constraints primarily focus on building continuous aircraft routes out of strings; the maintenance constraints by definition are enforced through the variable definition, with a variable not being included in the model unless the corresponding string is maintenance feasible. The following is an example of a string-based model.

Sets

- S the set of all possible strings.
- F the set of flights.
- N the set of nodes that represent time and space points in the flight network.
- $S(s, n)$ the set of strings s that start at node n , $\forall n \in N, \forall s \in S$.
- $E(s, n)$ the set of strings s that end at node n , $\forall n \in N, \forall s \in S$.

Parameters

- K a parameter that indicates the number of available aircraft.
- α_{fs} a binary parameter that indicates if string s contains flight f , $\forall s \in S, \forall f \in F$

Decision Variables

- g_n variables that represents ground arcs, which indicate the number of aircraft on the ground at node point n , $\forall n \in N$.
- d_s a binary variable that indicates if string s is chosen in the final solution, $\forall s \in S$.

$$\min \sum_{s \in S} c_s d_s \quad (19)$$

Subject to:

$$\sum_{s \in S} \alpha_{fs} d_s = 1 \quad \forall f \in F \quad (20)$$

$$\sum_{s \in E(s, n)} d_s + g_n^- - \sum_{s \in S(s, n)} d_s - g_n^+ = 0 \quad \forall n \in N \quad (21)$$

$$\sum_{s \in S} d_s + \sum_{n \in Z^T} g_n^+ \leq K \quad (22)$$

$$d_s \in \{0, 1\} \quad \forall s \in S \quad (23)$$

$$g_n^+, g_n^- \geq 0 \quad \forall n \in N. \quad (24)$$

The objective in Equation (19) minimizes the cost of the chosen strings. Constraint set (20) ensures that each flight is included in exactly one chosen string, with α_{fs} indicating whether string s covers flight f . Constraint set (21) ensures that continuous aircraft routes can be formed from the given strings (note that the mapping from strings to routes may not be unique). Finally, constraint set (22) provides a count constraint. As in the FAMs, the total number of available strings that are assigned at a given time cannot exceed the number of aircraft available.

The challenge in solving a string-based formulation is the exponentially large number of variables. One approach to overcome this challenge is to solve the LP relaxation of the problem via *column generation* [58]. In column generation, a *restricted master problem* is provided in which a limited subset of

the variables (here, the strings) are included. This restricted master is solved to optimality. The dual information is then passed on to a *subproblem*, a secondary optimization problem that is used to identify the string with the most negative reduced cost. If this yields a string with strictly negative reduced cost, then this can be passed back to the restricted master that will then continue with pivoting. Otherwise, if there are no negative reduced cost strings then the optimality of the problem has been established.

The key challenge in this approach is to find an efficient way to solve the subproblem. This can be done, for example, by formulating and solving a network flow problem (similar to the MCF approach above), where the arc costs now include the dual information as well as the true costs.

CREW SCHEDULING

Just as aircraft are required to follow a complex set of maintenance requirements, there are also many rules that restrict how crews can be assigned to flights. For example, on a given work day (known as a *duty*), crew members are limited in both the total number of hours that they can fly and also the total *elapsed time* from the start of the duty's first flight to the end of the duty's last flight. There are also limits on the minimum and maximum time between any two consecutive flights. In addition, crew members often are on duty for multiple sequential days. Their multiday schedule is known as a *pairing*. A pairing is simply a string of consecutive duties, where the first duty starts and the last duty ends at the airport where the crew is *based*, and nights in between the duties are spent at a hotel. A pairing also has many restrictions on the amount of flying and on-duty time permitted, as well as on the amount of rest required between duties. More information on crew restrictions can be found in Barnhart *et al.* [59].

In addition to these complex feasibility rules, the cost structure for paying crews is quite complex as well. For example, the cost of a duty is defined by equation (25).

$$b_d = \max\{mg_1, f_1 \times \text{elapsed}, \text{fly}\} \quad (25)$$

Here mg_1 represents a guaranteed minimum number of hours. For example, if a crew flies a short flight, followed by a substantial wait time on the ground, followed by another short flight, the number of compensated flying hours could be minimal. To prevent such a situation from occurring, each crew is paid at least mg_1 , which represents a lower threshold on the number of hours. In addition, f_1 represents a contractual fraction that is multiplied by the total elapsed duty period, to ensure adequate compensation for a duty period of very limited duration. Finally, fly represents the total number of flying hours in a duty period.

$$c_p = \max \left\{ \text{NDP} \times mg_2, f_2 \times \text{TAFB}, \sum_{d \in P} b_d \right\}. \quad (26)$$

In the first term, NDP represents the number of duties in a pairing. This is multiplied by mg_2 , which is the minimum guarantee per duty. Next, f_2 is again a contractual fraction that is then multiplied by the time-away-from-base (TAFB). The final term represents the total of all of the individual duty periods as computed in Equation (25).

Given the cost functions stated above, the crew scheduling problem thus becomes one of finding a minimum-cost assignment of crews to flights while satisfying all of these feasibility requirements. See Barnhart *et al.* [59] for a more detailed discussion of this complex problem.

Crew scheduling has many parallels to aircraft routing in the sense of assigning resources to sequences of flights while ensuring a number of complex constraints. And like aircraft routing, crew scheduling is solved in two stages. In the first stage, several months in advance of the start of the upcoming schedule, a set of *pairings* is constructed that collectively covers all of the flights. These pairings (analogous to the LOF in aircraft routing) will be repeated throughout the schedule period, but not always by the same crew members (just as LOFs are flown by different tails on different days). It is not until the second stage (typically solved on a monthly basis) that specific crew members are actually assigned to these pairings (analogous

to the tail assignment problem in aircraft routing). This assignment of crews to pairings is typically solved through either a *crew rostering problem* or a *bidline problem*. For more information on these problems we refer the reader to Caprara *et al.* [60]. We focus for the remainder of this section on the crew pairing problem.

Crew Pairing Formulation

A crew pairing is a fully self-contained assignment for an individual crew member. That is, given the set of pairings that make up a crew member's monthly schedule, if each pairing is feasible then the full schedule will be feasible. Furthermore, the cost of a schedule is simply the sum of the costs of the pairings.

Thus, we can formulate the crew pairing problem quite simply as a *set partitioning problem* [61] in which each variable represents a feasible pairing. The sole constraint is to choose a set of pairings such that each flight is included in exactly one pairing.

Sets

- P the set of all possible strings.
- F the set of flights.

Parameters

- δ_f^p a binary parameter that indicates whether flight f is included in pairing p , $\forall f \in F, \forall p \in P$.

Decision Variables

- x_p a binary variable that indicates if pairing p is included in the solution, $\forall p \in P$.

Objective:

$$\sum_{p \in P} c_p x_p \quad (27)$$

Subject to:

$$\sum_{p \in P} \delta_f^p x_p = 1 \quad \forall f \in F \quad (28)$$

$$x_p \in \{0, 1\} \quad \forall p \in P. \quad (29)$$

In constraint set (28), the parameter δ_f^p is a binary parameter that indicates whether flight f is included in pairing.

Note that although this problem is very concise to formulate, it can be quite difficult to solve for a moderately sized airline; the number of feasible pairings (and thus binary variables in the model) can easily reach billions. Again, similar to the aircraft routing, this problem is often solved using column generation to solve the linear programs (for an alternative solution technique, see Vance *et al.* [62]).

When solving the LP relaxation of the crew pairing problem via column generation, it is necessary to pose a subproblem in which we generate the pairing with the most negative reduced cost. This is more challenging than the aircraft routing subproblem because of the parameters of the complex feasibility rules as well as the non-linear cost function, both of which cannot simply be summed across the flights in a pairing. To overcome these challenges, one of the most successful approaches has been through *multilabel shortest path* algorithms [63]. These approaches are similar to Dijkstra or other label setting algorithms, with the key distinction being that at each node, rather than just keeping one cost label and pruning any path to that node with a higher cost, we must keep multiple labels (e.g., one for cost, one for elapsed time in the duty accrued so far, and one for flying time in the duty accrued so far), and we can only prune when *all* labels are dominated.

Finally, we conclude this section with a discussion of branching strategies. When using column generation to solve string-based models such as aircraft routing or crew scheduling, a subproblem is used to generate candidate pivot variables, rather than explicitly enumerating all of the variables and computing their respective reduced costs. Note that this approach only solves the LP relaxation of the problem. When column generation is embedded within *branch and bound* to solve an integer program, it is often referred to as *branch and price* [58]. Using column generation within *branch and bound* introduces its own new set of challenges, as we need to be able to enforce

a branching strategy that is consistent with the subproblem formulation. This is not necessarily a trivial task. For example, in traditional branching strategies, we often pick some variable that has a fractional value and impose additional constraints to rule out this value. If x is a binary variable with value 0.5 in the current solution to the LP relaxation, we might enforce $x = 0$ on one half of the tree and $x = 1$ on the other. These new constraints impose an additional dual value in the reduced cost calculation for variable x however, which is not easily captured without modifying the subproblem structure to treat that variable as a special case.

As an alternative, we can use a strategy known as *branching-on-follow-ons* [64]. We explain this strategy via a simple example. Suppose we have a pairing composed of four sequential flights, $A \rightarrow B \rightarrow C \rightarrow D$, and that this pairing is assigned to a fractional value. To prevent this fractional solution in subsequent nodes of the tree, we do not branch on the fractional pairing variable, but rather on a fractional flight connection. For example, we can force A to be followed by B in one half of the tree and A *not* to be followed by B in the other half. Forcing A to be followed by B can be imposed by simply combining the nodes representing the two flights into one single node, while forcing A to be followed by B can be imposed by simply deleting the arc connecting the nodes corresponding to these flights. The structure of the subproblem remains unchanged, and in fact each progressive subproblem becomes easier to solve, as more connections are forced a priori.

We conclude by noting that this branching strategy extends to set partitioning problems in the broader sense, where we can branch on items A and B *are* in the same set on one side of the tree and A and B *are not* in the same set on the other.

BEYOND BASIC PLANNING

Integrated Planning

There is clearly a strong link between each of the three planning problems described above (as well as with the schedule design problem, in which the set of flights itself is

determined). Significant benefits can therefore be achieved through an integrated rather than sequential approach to solving them. On the other hand, given that each problem is itself challenging to solve individually, solving them simultaneously requires significant advances in modeling and algorithms in order to ensure tractability.

Over the past 15 years, many advances have been made in this area, with two primary focuses. One is in developing heuristics to quickly find high-quality solutions to large-scale integer programs. The other is in partial integration, trying to identify the most critical relationships between different problems and focusing on capturing these relationships in an integrated approach. The following is merely a sampling of this rich literature.

In Clarke *et al.* [65], the basic fleet assignment problem is extended to include maintenance and crew considerations. That is, although FAM is still solved prior to aircraft routing and crew scheduling, extra constraints are added to increase the likelihood that the FAM solution will be maintenance- and crew-feasible.

In Sherali *et al.* [66] fleet assignment is augmented with additional schedule design decisions, and the focus is on computational techniques to solve the resulting large-scale integer program. Schedule design and fleet assignment are also integrated in Lohatepanont and Barnhart [67] and Barnhart *et al.* [68].

Finally, we note that there have been several papers on the integration of aircraft routing and crew scheduling, exploiting the fact that the link between these two problems is fairly narrow—the key connection is that when two flights with a very tight connection time are assigned to a common aircraft, then these two flights become a viable connection for a crew as well. (If the flights were assigned to two different aircraft, there would not be time for the crew to move through the terminal and cover both flights.) Thus decisions made in the aircraft routing problem impact the feasible region of decisions to be made in the crew scheduling problem. This problem structure naturally lends itself to a variety of decomposition approaches. These are explored by Cohn and Barnhart [56], Mercier

et al. [57], Cordeau *et al.* [69], Mercier and Soumis [70], and Weide *et al.* [71].

Robustness and Recovery

It is important to note that airline planning problems are often modeled as static and deterministic, although the real-world problems are both dynamic and stochastic. For example, the same fleet assignment is flown repeatedly over the course of a schedule period, even though demand varies daily over this horizon. Furthermore, the flight times needed to define the time-space network are taken as fixed, whereas actual flight times can vary quite substantially in practice.

The reasons for static and deterministic modeling are twofold. The first is that repeating schedules have operational benefits, reducing the number of decisions that must be made and communicated, and allowing workers to develop familiarity with a plan over time. The second is that even when solved statically and deterministically, these problems are computationally very challenging.

The cost of these simplifying assumptions is that it is quite common that a carrier will be unable to fully operate a schedule as planned on any given day. Maintenance problems, weather delays, and many other sources of disruption will require modifications to the original plan to recover from these disruptions. There are two ways to reduce the impact of disruptions.

One is through sophisticated recovery tools that allow the user to quickly modify the current schedule in response to a disruption. There is a vast literature on this topic. The aircraft recovery problem is formalized in Rosenberger *et al.* [72] and is further studied in Eggenberg *et al.* [16]. For more information on disruption management, Clausen *et al.* [73] provide a survey paper that covers aircraft recovery, crew recovery, and integrated crew aircraft and passenger recovery. Sarac *et al.*, in their work [74], use column generation approach to deal with day-of-operations disruption and subsequent recovery.

The second approach is to incorporate *robustness* into the planning process. This can take the form of reducing the impact of

delays (e.g., adding buffer between flights decreases the likelihood that one flight delay will propagate to a subsequent down-stream flight) [34]. It can also take the form of creating greater opportunities for recovery when disruptions do occur (e.g., building crew pairings that provide extra swap opportunities when an inbound crew is delayed and enabling another crew to take over the delayed crew's outbound flight) [75,76].

We conclude by highlighting a few recent approaches to improving robustness in airline planning:

- Rosenberger *et al.* [77] note that when airlines cancel flights, they tend to cancel entire cycles, that is, sequences of flights that begin and end at the same airport. As a result, a fleet assignment and aircraft rotation with many short cycles is frequently less sensitive to a flight cancellation than one with only a few short cycles.
- Ahmadbeygi *et al.* [34] quantify the prevalence of delay propagation in modern airline schedules. This provided motivation for the work of Ahmadbeygi *et al.* [78], in which minor modifications are made to flight departure times to redistribute the network's existing slack, moving extra slack to turns that are historically prone to delay propagation and away from turns that are historically reliable. This work builds on that in Lan *et al.* [79] and most recently on that in ISMP [80].
- The motivation behind the work in Lapp and Cohn [81] is the fact that tail assignments are frequently swapped over the course of the day to adjust for disruptions. This can make longer term maintenance plans infeasible. This paper takes an existing set of LOFs and modifies them so as to maintain important crew and passenger connections, while maximizing the number of opportunities for overnight recovery of the maintenance plan.
- In Schaefer *et al.* [82], simulation (in the form of a tool called SimAir) is used

to evaluate the quality of crew schedules when implemented under stochastic conditions.

- Ehrgott Ryan [83], Tekiner *et al.* [15], and Yen and Birge [84] explore the construction of robust crew schedules under uncertainty.
- The authors use the idea of a *station purity* measure to improve the robustness of fleet assignments in Smith and Johnson [85].
- Weide *et al.* [71] provide an iterative approach to generating integrated aircraft routing and crew scheduling. By first solving one problem to optimality, the authors are able to obtain trade-off points between cost and robustness.

FUTURE WORK

Despite all of the efforts and accomplishments that the OR community has had in passenger aviation, many challenges still remain. These include further advances in robust and integrated planning, as well as many other problems including those briefly introduced in this article.

In addition, the industry faces many new challenges, which the OR community can help to address by playing a pivotal role. These include the following:

- escalating and volatile fuel costs;
- increasing congestion, both at airports and in the airspace;
- environmental concerns; and
- security concerns.

We expect to see many advances in addressing these challenges in the coming years.

REFERENCES

1. Bureau of Transportation Statistics. Transtats, flights. Bureau of Transportation Statistics; 2010 Apr.
2. Gopalan R, Talluri K. Mathematical models in airline schedule planning: a survey. *Ann Oper Res* 1998;76:155–185.
3. Daskin MS, Panayotopoulos ND. A Lagrangian relaxation approach to assigning aircraft to routes in hub and spoke networks. *Transp Sci* 1989;23(2):91–99.
4. Warburg V, Hansen TG, Larsen A, *et al.* Dynamic airline scheduling: an analysis of the potentials of refueling and retiming. *J Air Transp Manage* 2008;14(4):163–167.
5. Jiang H, Barnhart C. Dynamic airline scheduling. *Transp Sci* 2009;43(3):336–354.
6. McGill JI, van Ryzin GJ. Revenue management: research overview and prospects. *Transp Sci* 1999;33(2):233–256.
7. Talluri K, van Ryzin G. The theory and practice of revenue management. Boston (MA): Kluwer Academic Publishers; 2004.
8. Belobaba P, Odoni A, Barnhart C. Fundamentals of pricing and revenue management. West Sussex: John Wiley & Sons, Ltd; 2009.
9. Fiig T, Isler K, Hopperstad C, *et al.* Optimization of mixed fare structures: theory and applications. *J Revenue Pricing Manage* 2010;9:152–170.
10. Lardeux B, Goyons O, Robelin C-A. Availability calculation based on robust optimization. *J Pricing Revenue Manage* 2010;9(4):313–325.
11. Barnhart C, Kniker TS, Lohatepanont M. Itinerary-based airline fleet assignment. *Transp Sci* 2002;36(2):199–217.
12. Dumas J, Soumis F. Passenger flow model for airline networks. *Transp Sci* 2008;42(2):197–207.
13. Berge ME, Hopperstad CA. Demand driven dispatch: a method for dynamic aircraft capacity assignment, models and algorithms. *Oper Res* 1993;41(1):153–168.
14. Shebalov S. Practical overview of demand-driven dispatch. *J Pricing Revenue Manage* 2009;8:166–173.
15. Tekiner H, Birbil SI, Bülbül K. Robust crew pairing for managing extra flights. *Comput Oper Res* 2009;36(6):2031–2048.
16. Eggenberg N, Salani M, Bierlaire M. Constraint-specific recovery network for solving airline recovery problems. *Comput Oper Res* 2010;37(6):1014–1026.
17. Abdelghany KF, Shah SS, Raina S, *et al.* A model for projecting flight delays during irregular operation conditions. *J Air Transp Manage* 2004;10(6):385–394.
18. Abdelghany A, Ekollu G, Narasimhan R, *et al.* A proactive crew recovery decision support tool for commercial airlines during irregular operations. *Ann Oper Res* 2004;127:309–331.

19. Abdelghany KF, Abdelghany AF, Ekollu G. An integrated decision support tool for airlines schedule recovery during irregular operations. *Eur J Oper Res* 2008;185(2):825–848.
20. Kohl N, Larsen A, Larsen J, *et al.* Airline disruption management-perspectives, experiences and outlook. *J Air Transp Manage* 2007;13(3):149–162.
21. Mangoubi RS, Mathaisel DFX. Optimizing gate assignments at airport terminals. *Transp Sci* 1985;19(2):173–188.
22. Bihr RA. A conceptual solution to the aircraft gate assignment problem using 0, 1 linear programming. *Comput Ind Eng* 1990;19(1–4):280–284.
23. Haghani A, Chen M-C. Optimizing gate assignments at airport terminals. *Transp Res Part A: Policy Pract* 1998;32(6):437–454.
24. Bolat A. Assigning arriving flights at an airport to the available gates. *J Oper Res Soc* 1999;50(1):23–34.
25. Dorndorf U, Drexel A, Nikulin Y, *et al.* Flight gate scheduling: state-of-the-art and recent developments. *Omega* 2007;35(3):326–334.
26. van den Briel MHL, Villalobos JR, Hogg GL, *et al.* America West Airlines develops efficient boarding strategies. *Interfaces* 2005;35(3):191–201.
27. Abdelghany A, Abdelghany K, Narasimhan R. Scheduling baggage-handling facilities in congested airports. *J Air Transp Manage* 2006;12(2):76–81.
28. Stolletz R. Operational workforce planning for check-in counters at airports. *Transp Res Part E: Logist Transp Rev* 2010;46(3):414–425.
29. Espinoza D, Garcia R, Goycoolea M, *et al.* Per-seat, on-demand air transportation Part I: problem description and an integer multicommodity flow model. *Transp Sci* 2009;42(3):263–278.
30. Espinoza D, Garcia R, Goycoolea M, *et al.* Per-seat, on-demand air transportation Part II: parallel local search. *Transp Sci* 2009;42(3):279–291.
31. Ball MO, Ausubel LM, Berardino F, *et al.* Market-based alternatives for managing congestion at New York's LaGuardia airport. Papers of Peter Cramton 07mbac. Department of Economics, University of Maryland; 2007.
32. Rassenti SJ, Smith VL, Bulfin RL. A combinatorial auction mechanism for airport time slot allocation. *Bell J Econ* 1982;13(2):402–417.
33. Harsha P. Mitigating airport congestion : market mechanisms and airline response models [PhD thesis]. Cambridge (MA): MIT; 2009.
34. AhmadBeygi S, Cohn A, Guan Y, *et al.* Analysis of the potential for delay propagation in passenger airline networks. *J Air Transp Manage* 2008;14(5):221–236.
35. Stolletz R. Non-stationary delay analysis of runway systems. *OR Spectr* 2008;30(1):191–213.
36. Baik H, Trani AA. Framework of a time-based simulation model for the analysis of airfield operations. *J Transp Eng* 2008;134(10):397–413.
37. Balakrishna P, Ganesan R, Sherry L. Airport taxi-out prediction using approximate dynamic programming: intelligence-based paradigm. *Transp Res Rec: J Transp Res Board* 2008;2052(1):53–61.
38. Abdelghany A, Abdelghany K. Modeling applications in the airline industry. Farnham: Ashgate; 2010.
39. Doganis R. Flying off course, the economics of international airlines. New York (NY): Routledge; 2002.
40. Belobaba P, Odoni A, Barnhart C. The global airline industry. West Sussex: John Wiley & Sons, Ltd; 2009.
41. Hane CA, Barnhart C, Johnson EL, *et al.* The fleet assignment problem: solving a large-scale integer program. *Math Program* 1995;70(2):211–232.
42. Klier N, Mellouli T, Suhl L. A time-space network based exact optimization model for multi-depot bus scheduling. *Eur J Oper Res* 2006;175(3):1616–1627.
43. Jacobs TL, Smith BC, Johnson EL. Incorporating network flow effects into the airline fleet assignment process. *Transp Sci* 2008;42(4):514–529.
44. Rexing B, Barnhart C, Kniker T, *et al.* Airline fleet assignment with time windows. *Transp Sci* 2000;34(1):1–20.
45. Dumas J, Aithnard F, Soumis F. Improving the objective function of the fleet assignment problem. *Transp Res Part B: Methodol* 2009;43(4):466–475.
46. Gopalan R, Talluri KT. The aircraft maintenance routing problem. *Oper Res* 1998;46(2):260–271.
47. Clarke L, Johnson E, Nemhauser G, *et al.* The aircraft rotation problem. *Ann Oper Res* 1997;69(1):33–46.
48. Gabteni S, Grönkvist M. Combining column generation and constraint programming to solve the tail assignment problem. *Ann Oper Res* 2009;171(1):61–76.

49. Kabbani NM, Patty BW. Aircraft routing at american airlines. Proceedings of The Thirty-Second Annual Symposium of AGIFORS. Budapest, Hungary; 1992.
50. Talluri KT. The four-day aircraft maintenance routing problem. *Transp Sci* 1998;32(1):43–53.
51. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc Am Math Soc* 1956;7(1):48–50.
52. Fisher ML. The Lagrangian relaxation method for solving integer programming problems. *Manage Sci* 2004;50(12):1861–1871.
53. Grönkvist M. The tail assignment problem [PhD thesis]. Chalmers University of Technology; 2005.
54. Ford LR, Fulkerson DR. A suggested computation for maximal multi-commodity network flows. *Manage Sci* 2004;50(12):1778–1780.
55. Barnhart C, Hane CA, Vance PH. Volume 3011/2004, Integer multicommodity flow problems. Berlin/Heidelberg: Springer; 1996.
56. Cohn AM, Barnhart C. Improving crew scheduling by incorporating key maintenance routing decisions. *Oper Res* 2003;51(3):387–396.
57. Mercier A, Cordeau J-F, Soumis F. A computational study of benders decomposition for the integrated aircraft routing and crew scheduling problem. *Comput Oper Res* 2005;32(6):1451–1476.
58. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. *Oper Res* 1998;46(3):316–329.
59. Barnhart C, Cohn A, Johnson EL, *et al.* Airline crew scheduling. New York: Kluwer Scientific Publishers; 2003.
60. Caprara A, Toth P, Vigo D, *et al.* Modeling and solving the crew rostering problem. *Oper Res* 1998;46(6):820–830.
61. Ryan DM. The solution of massive generalized set partitioning problems in aircrew rostering. *J Oper Res Soc* 1992;43(5):459–467.
62. Vance PH, Atamturk A, Barnhart C, *et al.* A heuristic approach for the airline crew pairing problem. 1997.
63. Lavoie S, Minoux M, Odier E. A new approach for crew pairing problems by column generation with an application to air transportation. *Eur J Oper Res* 1988;35(1):45–58.
64. Anbil R, Tanga R, Johnson EL. A global approach to crew pairing. *IBM Syst J* 1992;31:71–78.
65. Clarke LW, Hane CA, Johnson EL, *et al.* Maintenance and crew considerations in fleet assignment. *Transp Sci* 1996;30(3):249–260.
66. Sherali HD, Bae K-H, Haouari M. Integrated airline schedule design and fleet assignment: polyhedral analysis and Benders' decomposition approach. *INFORMS J Comput* 2009. DOI: [ijoc.1090.0368](https://doi.org/10.1090.0368).
67. Lohatepanont M, Barnhart C. Airline schedule planning: integrated models and algorithms for schedule design and fleet assignment. *Transp Sci* 2004;38(1):19–32.
68. Barnhart C, Farahat A, Lohatepanont M. Airline fleet assignment with enhanced revenue modeling. *Oper Res* 2009;57(1):231–244.
69. Cordeau J-F, Stojkovic G, Soumis F, *et al.* Benders decomposition for simultaneous aircraft routing and crew scheduling. *Transp Sci* 2001;35(4):375–388.
70. Mercier A, Soumis F. An integrated aircraft routing, crew scheduling and flight retiming model. *Comput Oper Res* 2007;34(8):2251–2265.
71. Weide O, Ryan D, Ehrgott M. An iterative approach to robust and integrated aircraft routing and crew scheduling. *Comput Oper Res* 2010;37(5):833–844.
72. Rosenberger JM, Johnson EL, Nemhauser GL. Rerouting aircraft for airline recovery. *Transp Sci* 2003;37(4):408–421.
73. Clausen J, Larsen A, Larsen J, *et al.* Disruption management in the airline industry—concepts, models and methods. *Comput Oper Res* 2010;37(5):809–821.
74. Sarac A, Batta R, Rump CM. A branch-and-price approach for operational aircraft maintenance routing. *Eur J Oper Res* 2006;175(3):1850–1869.
75. Shebalov S, Klabjan D. Robust airline crew pairing: move-up crews. *Transp Sci* 2006;40(3):300–312.
76. Gao C, Johnson E, Smith B. Integrated airline fleet and crew robust planning. *Transp Sci* 2009;43(1):2–16.
77. Rosenberger JM, Johnson EL, Nemhauser GL. A robust fleet-assignment model with hub isolation and short cycles. *Transp Sci* 2004;38(3):357–368.
78. Ahmadbeygi S, Cohn A, Lapp M. Decreasing airline delay propagation by re-allocating scheduled slack. *IIE Trans* 2009.

79. Lan S, Clarke J-P, Barnhart C. Planning for robust airline operations: optimizing aircraft routings and flight departure times to minimize passenger disruptions. *Transp Sci* 2006;40(1):15–28.
80. ISMP. Robust airline scheduling: improving schedule robustness with flight re-timing and aircraft swapping. ISMP: 20th International Symposium on Mathematical Programming; 2009 Aug 23–28; Chicago. 2009.
81. Lapp M, Cohn A. Modifying lines-of-flight in the planning process for improved maintenance robustness. *Comput Oper Res* 2010.
82. Schaefer AJ, Johnson EL, Kleywegt AJ, *et al.* Airline crew scheduling under uncertainty. *Transp Sci* 2005;39(3):340–348.
83. Ehrgott M, Ryan DM. Constructing robust crew schedules with bicriteria optimization. *J Multi-Criteria Decis Anal* 2002;11(3):139–150.
84. Yen JW, Birge JR. A stochastic programming approach to the airline crew scheduling problem. *Transp Sci* 2006;40(1):3–14.
85. Smith BC, Johnson EL. Robust airline fleet assignment: imposing station purity using station decomposition. *Transp Sci* 2006;40(4):497–516.

ALLOCATION GAMES

BRIAN ROBERSON
Department of Economics,
Krannert School of
Management, Purdue
University, West
Lafayette, Indiana

INTRODUCTION

This article examines a class of resource allocation games in which two budget-constrained players strategically allocate resources across multiple simultaneous contests (see also the article titled ***TPZS Applications: Blotto Games*** in this encyclopedia). In each of the individual component contests the players allocate scarce resources (such as time, money, or effort) in order to affect the probability of winning the contest, where this probability is increasing in a player's own sunk resource expenditure and decreasing in their opponent's resource expenditure in that contest (for an introduction to contest theory, see Konrad [2]). The outcome of the allocation game is a function of the outcomes in the individual component contests. Two of the most common objectives in such environments are (i) to maximize the sum of the winnings across the individual contests, henceforth the *plurality objective*, and (ii) to maximize the probability of winning a majority of the component contests, henceforth the *majority objective*.¹

There are a number of important applications that may be characterized as

budget-constrained players strategically allocating resources across multiple simultaneous contests. Blackett [4], in a 1953 panel that was chaired by Oskar Morgenstern and highlighted this particular class of resource allocation games, lists several military applications including (i) the allocation of bombers and interceptors across a set of target areas, (ii) the routing of convoys and the choice of submarine locations across disjoint routes, (iii) the location of amphibious landings and the allocation of defensive forces, and (iv) the allocation of forces across a set of distinct battlefields. Clearly, military applications of the model are plentiful. However, the problem of allocating resources across multiple component contests also arises in economic and political applications. For example, in a presidential campaign there are a number of combinations of states that result in an Electoral College victory, and each candidate allocates his or her campaign resources across the states in an attempt to win a majority of the votes within each of the states in any one of the winning combinations. Other applications of this theoretical framework include political competition over taxation and redistribution, political competition over vote-buying, multidimensional research and development competition, and multimarket advertising resource allocation.

Partly because it is a foundational problem that is well suited for abstract theoretical modeling, multidimensional strategic resource allocation was one of the first problems examined in modern game theory. Borel [5] formulates this problem as a constant-sum game involving two symmetric players, A and B, who must each allocate a fixed amount of resources, normalized to one unit of resources, over three contests (or battlefields). Each player must distribute their resources without knowing their opponent's distribution of resources. In each of the component contests, the player who allocates the higher level of resources wins the contest, and

¹For a survey of multidimensional contests in which there exist linkages in how battlefield outcomes and costs aggregate in determining performance in the overall conflict, see Kovenock and Roberson [3].

each player maximizes the expected number of component contest wins.²

Borel's game, which came to be known as the *Colonel Blotto game*, was a focal point in the early game theory literature [6–10]. To some degree, this was due to the fact that the Colonel Blotto game is, as Golman and Page [11] describe, an “elaborate version of Rock, Paper, and Scissors.” For example, in the symmetric Colonel Blotto game with three contests, each player i chooses three numbers $(x_{i,1}, x_{i,2}, x_{i,3})$ such that $\sum_{j=1}^3 x_{i,j} = 1$. If player i wishes to increase his allocation of force to contest j , then he must decrease his allocation of force to either or both of the other two contests. Furthermore, if player A knew player B's allocation of resources $(x_{B,1}, x_{B,2}, x_{B,3})$, then, because within each component contest the player who allocates the higher level of resources wins that contest, player A could win two of the three contests. As a result, the Colonel Blotto game (as with the Rock, Paper, and Scissors game) has no pure-strategy equilibria. In this game, a mixed strategy is a joint distribution function that specifies both the randomization in the allocations to the individual component contests (provided by a univariate marginal distribution function for each of the component contests) and a correlation structure that ensures that across the set of resource allocations the budget constraint is satisfied with probability one.

From a theoretical point of view, the Colonel Blotto game also provides an important benchmark for multidimensional strategic resource allocation games.³ In particular, let $p_{i,j}(x_{i,j}, x_{-i,j})$, henceforth the

contest success function (CSF), denote the probability that player i wins component contest j when player i allocates $x_{i,j}$ resources and player $-i$ allocates $x_{-i,j}$ resources to component contest j , and consider the general ratio-form contest success function $p_{i,j}(x_{i,j}, x_{-i,j}) = x_{i,j}^m / (x_{i,j}^m + x_{-i,j}^m)$, where the parameter $m \geq 0$ specifies the level of randomness or noise in the component contests. When $m = 0$, each player wins each component contest with equal probability regardless of the players' resource allocations. For low values of m the outcome of the component contest is largely random, or noisy (for further information on how much noise is implied, see Konrad and Kovenock [14]). As m increases, the amount of noise in the contest decreases. Now, consider a Blotto-type resource allocation game in which (i) each player maximizes the expected number of component contest wins, (ii) in each of the n component contests, the probability that player i ($i = A, B$) wins component contest j ($j = 1, \dots, n$) is given by the general ratio-form CSF, and (iii) each player has one unit of resource to distribute among the n contests. For this resource allocation game, the Lagrangian for player i 's optimization problem may be written as,

$$\pi_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \sum_{j=1}^n \left[\frac{x_{i,j}^m}{x_{i,j}^m + x_{-i,j}^m} - \lambda_i x_{i,j} \right] + \lambda_i, \quad (1)$$

where $\mathbf{x}_i$ and $\mathbf{x}_{-i}$ are elements of the standard simplex in $\mathbb{R}^n$. Recalling that the Colonel Blotto game features deterministic, or no noise, component CSFs in which the player who allocates the higher level of resources wins that contest,⁴ it is clear that the Colonel Blotto game corresponds to the limiting case of this game in which m is set to infinity. This resource allocation game has also been examined by Friedman [17] for the case that $m = 1$,

²Although Borel actually assumes that the players maximize the probability of winning a majority of the component contests, in the case of symmetric players and three contests the majority objective is strategically equivalent to the plurality objective. Furthermore, with four or more contests, the solution to the majority game is still an open question. Following Gross and Wagner [6], the literature on multidimensional contests has primarily focused on the plurality objective.

³The discussion given here focuses on the general ratio-form CSF. Note, though, that a similar argument can be made for the difference-form CSF as

in Lazear and Rosen [12]. For further details, see Che and Gale [13].

⁴This type of CSF is commonly known as the *auction CSF*. See, for example, the closely related literature on all-pay auctions [15,16].

and by Robson [18] for the case that $m \in (0, 1]$. Observe that Equation (1) is concave with respect to $\mathbf{x}_i$ only for m less than or equal to 1. Furthermore, for m greater than 2, as in the $m = \infty$ case, no pure-strategy equilibria exist. Even in the case of a single contest with linear costs, the equilibrium set for the $m > 2$ case has not yet been characterized, except in the case of $m = \infty$ [16]. Note also that in the single contest game with linear costs, it is known that for $m > 2$ there exist equilibria that are payoff equivalents to the $m = \infty$ case [19,20]. To summarize, the equilibrium characterization of the Colonel Blotto game (i.e., the $m = \infty$ case) provides an important theoretical benchmark that sheds light on all specifications of the general ratio-form CSF game in which $m > 2$.

The remainder of this article is outlined as follows. The section titled “The Model” presents the formal specification of the Colonel Blotto game. The section titled “The Colonel Blotto Game (Plurality Objective)” provides an introduction to the Colonel Blotto game with a focus on the intuition for the key results. Partly for the reasons mentioned above and partly because of several theoretical breakthroughs in the area of multidimensional contests, there has been a resurgence of interest in the Colonel Blotto game [21–32]. The section titled “Variations and Extensions” briefly summarizes several of these recent developments including nonconstant-sum formulations of the Colonel Blotto game, additional restrictions on the strategy space, a simplified form of the Colonel Blotto game, and majoritarian and other objectives. The final section concludes the article.

THE MODEL

In this section we present a framework for examining various multidimensional strategic resource allocation games of the Blotto type. Two players, A and B , simultaneously allocate resources across a finite number, $n \geq 3$, of independent component contests. Each player has a fixed level of available resources (or budget). The stronger player A has a normalized budget of 1 unit of resources, and the

weaker player B has a normalized budget of $\beta \leq 1$. Each component contest $j \in \{1, \dots, n\}$ has a value of v_j for each player. We will focus primarily on the case that $v_j = v_k$ for each $j, k \in \{1, \dots, n\}$, a condition henceforth referred to as *homogenous contests*, but will also examine the case in which this condition does not hold, henceforth referred to as the case of *heterogenous contests*.

Let $\mathbf{x}_i$ denote the n -tuple $(x_{i,1}, \dots, x_{i,n})$ of player i 's allocation of resources across the component contests. The level of resources allocated to each contest must be nonnegative. For player A , the set of feasible allocations of resources across the n component contests is denoted by

$$S_A = \left\{ \mathbf{x} \in \mathbb{R}_+^n \mid \sum_{j=1}^n x_j = 1 \right\}.$$

Player B 's set of feasible allocations of resources, denoted S_B , is similarly delineated by $\mathbf{x} \in \mathbb{R}_+^n$ such that $\sum_{j=1}^n x_j = \beta$. We will focus primarily on the case that the resource is continuous, but will also briefly examine additional restrictions on the strategy space.

Objectives

Let $\pi_{i,j}(x_{i,j}, x_{-i,j})$ denote the payoff to player i in contest j when player i allocates $x_{i,j}$ resources and player $-i$ allocates $x_{-i,j}$ resources to contest j . Within each component contest the player that allocates the higher level of resources wins, and in the case of a tie, each player wins the component contest with equal probability. That is,

$$\pi_{i,j}(x_{i,j}, x_{-i,j}) = \begin{cases} v_j & \text{if } x_{i,j} > x_{-i,j} \\ \frac{v_j}{2} & \text{if } x_{i,j} = x_{-i,j} \\ 0 & \text{if } x_{i,j} < x_{-i,j}. \end{cases}$$

We will focus primarily on the general form of the *plurality objective* in which each player i 's payoff across the set of component contests, denoted π_i , is equal to the sum of the winnings in the individual component contests:

$$\pi_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \sum_{j=1}^n \pi_{i,j}(x_{i,j}, x_{-i,j}),$$

in this case the game is constant-sum with a total value of $\sum_{j=1}^n v_j$. Note that when $v_j = 1$ for all j , $\pi_i(\cdot, \cdot)$ is equal to the number of contests that player i wins.

We will also address several alternative objectives such as the *majority objective* in which each player i 's payoff across the set of component contests, denoted $\hat{\pi}_i$, is given by

$$\hat{\pi}_i(\mathbf{x}_i, \mathbf{x}_{-i}) = \begin{cases} 1, & \text{if } \sum_{j=1}^n \pi_{i,j}(x_{i,j}, x_{-i,j}) > \frac{\sum_{j=1}^n v_j}{2} \\ \frac{1}{2}, & \text{if } \sum_{j=1}^n \pi_{i,j}(x_{i,j}, x_{-i,j}) = \frac{\sum_{j=1}^n v_j}{2} \\ 0, & \text{otherwise} \end{cases}$$

in this case the game is constant-sum with a total value of 1.

Strategies

For each player i , a pure strategy is a budget-balancing n -tuple consisting of a nonnegative allocation of resources to each of the n component contests. A mixed strategy, which we term a *distribution of resources*, for player i is an n -variate distribution function $P_i : \mathbb{R}_+^n \rightarrow [0, 1]$ with support (denoted $\text{Supp}(P_i)$) contained in player i 's set of feasible resource allocations $\mathcal{S}_i$ and with the set of one-dimensional marginal distribution functions $\{F_{i,j}\}_{j=1}^n$, one univariate marginal distribution function for each component contest j . In a mixed strategy, the n -tuple of player i 's allocation of resources to each of the n contests is a random n -tuple drawn from the n -variate distribution function P_i .

THE COLONEL BLOTTO GAME (PLURALITY OBJECTIVE)

Throughout this section we will focus on the plurality objective. We begin with the symmetric Colonel Blotto game (i.e., $\beta = 1$) with homogeneous contests ($v_j = v_k \ \forall j, k \in \{1, \dots, n\}$). Let $v(\equiv v_j)$ denote the common value for each of the n component contests and let $\mathcal{S}(\equiv \mathcal{S}_A = \mathcal{S}_B)$ denote the symmetric strategy space.

Symmetric Players and Homogeneous Contests

In the case of symmetric players and homogeneous contests, it is straightforward to show that there is no pure-strategy equilibrium. Borel and Ville [33] provide an equilibrium in the case of three-component contests. Gross and Wagner [6] extend the analysis of the symmetric game with homogeneous contests to allow for $n \geq 3$ and $n = 2$. Theorem 1 provides Gross and Wagner's [6] sufficient conditions for equilibrium in the symmetric Colonel Blotto game with homogeneous contests.

Theorem 1 [Gross and Wagner [6]]. *The pair of n -variate distribution functions P_A^* and P_B^* constitute a Nash equilibrium of the symmetric Colonel Blotto game with homogeneous contests if they satisfy the following two conditions: (i) For each player i , $\text{Supp}(P_i^*) \subset \mathcal{S}$ and (ii) P_i^* , $i = A, B$, provides the corresponding set of univariate marginal distribution functions $\{F_{i,j}^*\}_{j=1}^n$ outlined below.*

$$\forall j \in \{1, \dots, n\}, \quad F_{i,j}^*(x) = \frac{x}{\frac{2}{n}}, \quad \text{for } x \in [0, \frac{2}{n}].$$

There exist such strategies, and, in equilibrium, each player has an expected payoff of $(nv/2)$.

The following discussion provides a brief sketch of the proof of Theorem 1 with an eye on intuition. We will then use this as a baseline case with which to compare other specifications of the Colonel Blotto game.

Let P^* denote a feasible n -variate distribution function (i.e., $\text{Supp}(P^*) \in \mathcal{S}$) with the set of univariate marginal distribution functions $\{F_j^*\}_{j=1}^n$ specified in Theorem 1. If player A is using P^* , then player A 's expected payoff under the plurality objective, π_A , when player B chooses any pure strategy $\mathbf{x}_B \in \mathcal{S}$ is

$$\pi_A(P^*, \mathbf{x}_B) = nv - v \sum_{j=1}^n F_j^*(x_{B,j}) \quad (2)$$

recalling that for all j , $F_j^*(x) = \frac{nx}{2}$ for $x \in [0, \frac{2}{n}]$

$$\pi_A(P^*, \mathbf{x}_B) \geq nv - nv \sum_{j=1}^n \frac{x_{B,j}}{2} = \frac{nv}{2}.$$

In a symmetric and constant-sum game that sums to nv , this is sufficient to prove that uniform univariate marginal distributions are optimal.

Before moving on to the case of heterogeneous contests, it is instructive to provide an n -variate distribution function P^* that satisfies the condition that $\text{Supp}(P^*) \in \mathcal{S}$ and has the set of univariate marginal distribution functions $\{F_j^*\}_{j=1}^n$ specified in Theorem 1. The following section provides the construction for such a distribution function. The construction given below can also be extended to cover the case of the symmetric players and heterogeneous contests as in Theorem 2. However, the case of asymmetric players (with either homogeneous or heterogeneous contests), as in Theorem 3, requires a different approach (see Roberson [29] for further details).

Equilibrium Joint Distributions. We now examine the construction of a joint distribution function, which satisfies the properties listed in Theorem 1. Gross and Wagner [6] examine several different constructions of sufficient joint distribution functions (including a fractal construction). The solution that we focus on here extends Borel and Ville's [33] disk solution for the case of three homogeneous contests to allow for any arbitrary finite number n of component contests. This generalized-disk solution exploits the following two properties of regular n -gons: (i) the sum of the perpendiculars from any point in a regular n -gon to the sides of the regular n -gon is equal to n times the inradius, and (ii) letting s be the side length and r be the inradius,

$$s = 2r \tan \frac{\pi}{n}$$

for all regular n -gons.

This outline of the generalized-disk solution follows along the lines of Laslier [26]. Let Ω be the incenter of a regular n -gon with sides of length $(2/n) \tan(\pi/n)$. The inradius is, thus, $(1/n)$. Let S be the sphere of radius $(1/n)$ centered at Ω . Let M be a point randomly chosen from the surface of S , according to the uniform distribution on the surface of

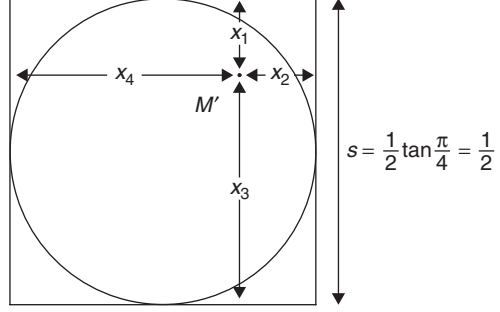


Figure 1. Generalized disk solution for $n = 4$.

S . Let M' be the projection of M on the plane that contains the regular n -gon. Let x_i be the perpendicular from side i to the point M' . For $n = 4$, the generalized disk is shown in Fig. 1.

The surface area of a spherical cap of S with height h is $2\pi rh$. Let $\text{Cap}(i, h)$ denote the spherical cap on S with height h as measured from side i of the regular n -gon. Thus,

$$\begin{aligned} \Pr[x_i \leq h] &= \Pr[M \in \text{Cap}(i, h)] \\ &= \frac{\text{Area}[\text{Cap}(i, h)]}{\text{Area}[S]} = \frac{2\pi rh}{4\pi r^2} = \frac{h}{2r}. \end{aligned}$$

Because the radius $r = (1/n)$, each x_i is uniformly distributed on the interval 0 to $\frac{2}{n}$.

Note that Gross and Wagner's [6] use of regular n -gons places severe restrictions on the supports of the resulting n -variate distributions, and the supports of the resulting joint distributions are contained in a strict subset of the intersection of the n -box $[0, \frac{2}{n}]^n$ and $\mathcal{S}$. For example with $n = 4$, each univariate marginal distribution randomizes continuously on $[0, \frac{1}{2}]$ such that at each point in the support, $x_1 + x_3 = \frac{1}{2}$ and $x_2 + x_4 = \frac{1}{2}$. More generally, for all n , the generalized-disk solution possesses the property that each n -tuple in its support is entirely determined by any two x_i that are not opposite to each other.

Symmetric Players and Heterogeneous Contests

Consider now the case of symmetric players' and heterogeneous contests. If any component contest j satisfies the property that $v_j \geq \sum_{k \neq j} v_k$, then the unique pure-strategy equilibrium trivially involves each player allocating all of their resources to the contest

with the maximal value. But if $v_j < \sum_{k \neq j} v_k$ for all j , then there is no pure-strategy equilibrium. Theorem 2 provides the modification of Theorem 1 that applies in this case. This result is due to Gross [34] (see also Laslier [26]).

Theorem 2 [Gross [34]]. *If $v_j < \sum_{k \neq j} v_k$ for all j , then the pair of n -variate distribution functions P_A^* and P_B^* constitute a Nash equilibrium of the symmetric Colonel Blotto game with heterogeneous contests if they satisfy the following two conditions: (i) For each player i , $\text{Supp}(P_i^*) \subset S$ and (ii) P_i^* , $i = A, B$, provides the corresponding set of univariate marginal distribution functions $\{F_{i,j}^*\}_{j=1}^n$ outlined below.*

$$\forall j \in \{1, \dots, n\}, \quad F_{i,j}^*(x) = \frac{\frac{x}{2v_j}}{\sum_{k=1}^n v_k},$$

$$\text{for } x \in \left[0, \frac{2v_j}{\sum_{k=1}^n v_k} \right].$$

There exist such strategies, and in equilibrium each player's expected payoff is $(1/2) \sum_{j=1}^n v_j$.

For the proof of the existence of a joint distribution which satisfies the conditions of Theorem 2, see Laslier [26] or Gross [34], both of which extend the generalized-disk solution to allow for the differing values of the heterogeneous contests. To see that a pair of distributions of resources, which satisfy the conditions of Theorem 2, form an equilibrium, observe that if player A is using such a P_A^* , then player A 's expected payoff, π_A , when player B chooses any pure strategy $\mathbf{x}_B \in S$ is

$$\pi_A(P_A^*, \mathbf{x}_B) = \sum_{j=1}^n v_j - \sum_{j=1}^n v_j F_{A,j}^*(x_{B,j}) \quad (3)$$

inserting $F_{A,j}^*(x_{B,j})$ from the statement of Theorem 2

$$\begin{aligned} \pi_A(P^*, \mathbf{x}') &\geq \sum_{j=1}^n v_j - \left(\sum_{j=1}^n v_j \right) \sum_{k=1}^n \frac{x_{B,j}}{2} \\ &= \frac{\sum_{j=1}^n v_j}{2}. \end{aligned}$$

Thus, such uniform univariate marginal distributions are optimal.

Asymmetric Players

We now examine the case of asymmetric players ($\beta < 1$). Note that if the strong player (A) has sufficient resources to outbid the weaker player's (B 's) maximal resource allocation β on all n contests (i.e., if $1 \geq n\beta$) then there, trivially, exists a pure-strategy equilibrium, and the strong player (A) wins all of the contests. It is well known that for the remaining parameter configurations, $(1/n) < \beta \leq 1$, there is no pure-strategy equilibrium for this class of games.

The earliest attempt at the asymmetric game is by Friedman [17], who simplifies the game by assuming that for each player i a strategy is a set of one-dimensional marginal distribution functions $\{F_{i,j}\}_{j=1}^n$ that satisfies the condition that the budget holds in expectation ($\sum_{j=1}^n \int_0^\infty x dF_{A,j} = 1$ and $\sum_{j=1}^n \int_0^\infty x dF_{B,j} = \beta$ for players A and B respectively). However, because Friedman [17] focuses on only sets of univariate marginal distributions that satisfy the budget in expectation, this analysis leaves open the question of whether—in the original specification of the asymmetric game—the constraint on the support of the joint distribution function (i.e., that the budget is satisfied with probability one) imposes restrictions on the feasible sets of univariate marginal distributions. These issues are resolved in Roberson [29], which solves the asymmetric game at the level of the n -variate distribution functions. For a large range of parameter configurations, Roberson [29] shows that the constraint on the support of the joint distribution function is, in fact, binding. In such cases the univariate marginal distributions given by Friedman [17] do not arise in equilibrium.

For the sake of brevity we will only examine the case of $n \geq 3$ and $\beta \in ((2/n), 1]$. This corresponds to the portion of the parameter space in which the univariate marginal distributions given by Friedman [17] do arise in equilibrium. See Roberson [29] for the remaining case, $n \geq 3$ and $\beta \leq (2/n)$, in which this relationship breaks down. For the case of $n = 2$, Gross and Wagner [6] provide an equilibrium and Macdonnel and Mastronardi [28] provide the complete characterization of the equilibrium joint distributions. Moving from $n = 2$ to $n \geq 3$ greatly enlarges the set of feasible distributions of resources, and for $n = 2$, the equilibrium strategies are qualitatively different from the $n \geq 3$ case. Theorem 3 summarizes Roberson's [29] characterization of equilibrium in the Colonel Blotto game for $n \geq 3$ and $\beta \in ((2/n), 1]$.

Theorem 3 [Roberson 29]]. *If $n \geq 3$ and $\beta \in ((2/n), 1]$, then the pair of n -variate distribution functions P_A^* and P_B^* constitute a Nash equilibrium of the Colonel Blotto game if and only if they satisfy the following two conditions: (1) For each player i , $\text{Supp}(P_i^*) \subset S_i$ and (2) P_i^* , $i = A, B$, provides the corresponding unique set of univariate marginal distribution functions $\{F_{i,j}^*\}_{j=1}^n$ outlined below.*

$$\begin{aligned} \forall j \in \{1, \dots, n\}, \quad F_{B,j}^*(x) &= (1 - \beta) + \frac{x\beta}{n}, \\ &\text{for } x \in [0, \frac{2}{n}]. \\ \forall j \in \{1, \dots, n\}, \quad F_{A,j}^*(x) &= \frac{x}{2}, \\ &\text{for } x \in [0, \frac{2}{n}]. \end{aligned}$$

Moreover, such strategies exist, and in any Nash equilibrium the expected payoff of the weak player (B) is $\beta(nv/2)$ and the expected payoff of the strong player (A) is $nv - \beta(nv/2)$.

A major part of the proof of Theorem 3 involves showing that there exist strategies that satisfy the two conditions specified above. To do this, Roberson [29] makes a break from the classical constructions, which exploit properties of n -gons, and

provide an entirely new approach.⁵ Since the appearance of the classical solutions to the symmetric case, it had been an open question whether uniform univariate marginal distributions were a necessary condition for equilibrium.⁶ Roberson [29] shows that—as long as the level of symmetry (as measured by the ratio of the players' resource constraints) is above a threshold—there exists a unique set of univariate marginal distribution functions for each player and these involve uniform marginals. Note also that when the players have asymmetric resource constraints, the disadvantaged player optimally uses a “guerrilla warfare” strategy, which involves the stochastic allocation of zero resources to a subset of the component contests. Conversely, the player with the larger budget plays a “stochastic complete coverage” strategy that with probability one, stochastically allocates a strictly positive level of resources to each of the component contests.

For all configurations of the asymmetric Colonel Blotto game, Roberson [29] provides the characterization of the unique equilibrium expected payoffs.⁷ These expected payoffs are illustrated in Fig. 2, as a function of the ratio of the players' resource constraints (β). These payoffs are for the case of at least three-component contests ($n \geq 3$) with a value of $v = (1/n)$ each and player B being the disadvantaged player ($\beta \leq 1$).

Although Roberson [29] focuses on the case of homogenous contests, it is straightforward to extend that analysis to allow for heterogeneous contests, as long as for each distinct contest valuation there are at least three contests with that valuation.

VARIATIONS AND EXTENSIONS

In what follows, we provide a brief overview of several of the directions in which the Colonel

⁵See also Weinstein [32], which provides a related construction in the case of the symmetric Colonel Blotto game.

⁶See, for example, Gross and Wagner [6] and Laslier and Picard [27], which discuss this issue.

⁷Note that uniqueness of the equilibrium expected payoffs follows immediately from the fact that the Colonel Blotto game is constant sum.

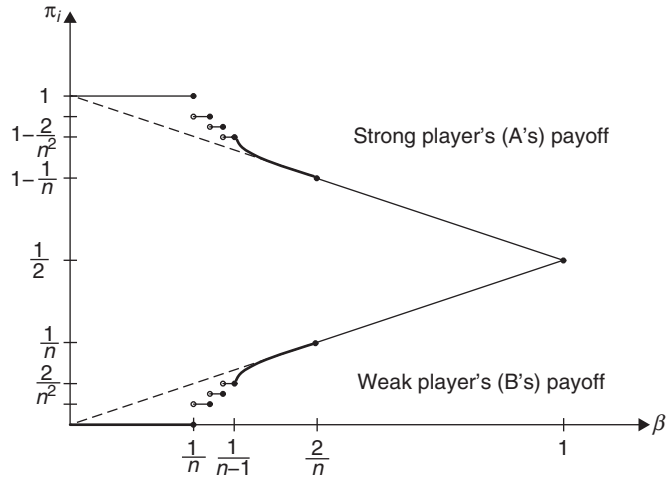


Figure 2. Colonel Blotto game payoffs.

Blotto game has been extended and provide directions for future work in this area.

The Majoritarian Objective

The majoritarian Colonel Blotto game is largely an open question. In the case of three contests, symmetric players, and homogeneous contests, the majority objective game is strategically equivalent to the plurality objective game. Therefore, the case of Theorem 1 with $n = 3$ applies directly (see Kovenock and Roberson [3] for further details). The case of three contests, symmetric players, and heterogeneous contests is addressed by Laslier [35], who shows that (as long as $v_j < \sum_{k \neq j} v_k$) the case of Theorem 1 with $n = 3$ applies directly. Observe that this differs from Theorem 2, the corresponding heterogeneous contest game with the plurality objective. That is, in the case of symmetric players and three heterogeneous contests, the equilibrium in the majoritarian game involves each contest receiving the same average resource allocation, while the equilibrium in the plurality game involves the contests with higher valuations receiving higher average resource allocations.

In the case of three contests and asymmetric resource constraints, Weinstein [32] provides bounds on the equilibrium payoffs. But beyond the case of three contests, little is known about the majoritarian Colonel Blotto game.

A Nonconstant-Sum Formulation

In the Colonel Blotto game, each player has a resource constraint and resources are “use-it-or-lose-it” in the sense that any unused resources have no value. There are a number of applications in which unused resources may have positive value. This issue was first examined by Szentes and Rosenthal [36] who examine, among other things, a nonconstant-sum formulation of the majoritarian Colonel Blotto game with three contests and symmetric players.⁸ The equilibrium in this extension is quite different from the constant-sum majority game, and we refer the interested reader to Szentes and Rosenthal [36] for further details. As with the case of the constant-sum formulation of the majoritarian Colonel Blotto game with symmetric players, the case of $n > 3$ is unresolved.

Kvasov [25] examines a nonconstant-sum formulation of the plurality Colonel Blotto game with symmetric players, and Roberson and Kvasov [31] examine the corresponding asymmetric game. The key insight from this variation of the Colonel Blotto game is that—as long as the level of asymmetry between the players’ budgets is below a threshold—there exists a one-to-one mapping from the unique set of equilibrium

⁸In this game the players do not face a budget constraint.

univariate marginal distribution functions in the constant-sum game to those in the nonconstant-sum game. That is, the key features of the equilibrium in the Colonel Blotto game are robust to the relaxation of the “use-it-or-lose-it” feature.

A Continuum of Contests

Myerson [37] introduces a Blotto-type game with a continuum of homogeneous contests, and symmetric players each with the plurality objective. In this game, a feasible distribution of resources is a univariate probability distribution that exhausts the budget in expectation, and the support of which is contained in $\mathbb{R}_+$. Instead of drawing an n -tuple from a joint distribution, the distribution of resources specifies a measure, over any interval of $\mathbb{R}_+$, that corresponds to the proportion of the component contests receiving an allocation of resources in that interval. Myerson [37] applies this Blotto-type game to political parties competing for vote share by simultaneously announcing binding commitments as to how they will allocate a fixed budget across a continuum of voters. Each voter votes for the party offering the highest level of utility, and each party’s payoff is the fraction of votes received by that party. This particular formulation of the redistributive politics model⁹ has been used to study the inequality created by political competition [37], incentives for generating budget deficits [42], inefficiency of public good provision [43,44], campaign spending regulation [45], redistributive competition in an electorate with heterogeneous party preferences [23], inefficient redistributive politics [24], and distortionary taxation [46].

Restrictions on the Strategy Space

Restrictions to the strategy space are an important area for future work on the Colonel Blotto game. Hart [22] provides a complete

characterization of the discrete version of the plurality Colonel Blotto game (with both symmetric and asymmetric players and homogeneous contests). In the case that the players are restricted to choosing only integer amounts of the resource to allocate to each contest, the main features of the equilibria are similar in spirit to those arising in the continuous game. For example, the weaker player utilizes a stochastic guerrilla warfare strategy and the stronger player utilizes a stochastic complete coverage strategy. In addition, the upper bounds of the supports of the univariate marginal distributions are the same in both the discrete and the continuous formulations.

Closely related to the discrete version of the Colonel Blotto game is the restriction to the strategy space examined by Arad [47], which places a further restriction on the set of integer allocations. In particular, that paper considers the case of four component contests and symmetric players with a budget of 10 units of resources who must choose a permutation of the numbers (1, 2, 3, 4).

Alternative Objectives

Although most of the work on multidimensional strategic resource allocation has focused on only the plurality and majoritarian objectives, there are a number of other relevant objectives. Golman and Page [21] examine a variation of the plurality game, which allows for the players to value not only isolated fronts, but also pairs of fronts and even all sets of fronts. In several of these games, pure-strategy equilibria are found to exist. Szentes and Rosenthal [48] examine a variation of the majoritarian game in which winning the overall game requires winning the supermajority of all but one of the component contests. That paper provides an equilibrium in the case that there are a sufficient number of players (strictly greater than two). Except for the case of three-component contests, the two-player case is unresolved.

CONCLUSION

As an exercise in both abstract and applied game-theoretic analysis, the Colonel Blotto

⁹There are several other variations of the redistributive politics model. See, for example, the literature in Cox and McCubbins [38], Lindbeck and Weibull [39], Dixit and Londregan [40], and Dixit and Londregan [41].

game provides a unified theoretical framework that facilitates new ways of thinking about the foundations of optimal resource allocation. This article (i) provides a brief outline of some of the main results on the classic Colonel Blotto game, (ii) surveys several of the theoretical extensions of this framework, and (iii) provides direction for future work in this area.

Acknowledgments

I wish to thank two anonymous referees for valuable suggestions.

REFERENCES

1. Washburn A, TPZS applications: Blotto Games. In: Cochran JJ, editor. *Encyclopedia of Operations Research and Management Science*. Hoboken(NJ): John Wiley&Sons. In press.
2. Konrad KA. *Strategy and dynamics in contests*. Oxford: Oxford University Press; 2009.
3. Kovenock D, Roberson B. Modeling multiple battlefields. In: Garfinkel M, Skaperdas S, editors. *Handbook of the Economics of Peace and Conflict*. Oxford: Oxford University Press. In press.
4. Blackett DW. Blotto-type games. Presented at the 4th Annual Logistics Conference (Part II- Restricted Session); 1953; Washington (DC).
5. Borel E. La theorie du jeu les equations integrales a noyau symetrique. *C R Acad*. 1921;173:1304–1308. English translation by Savage L. The theory of play and integral equations with skew symmetric kernels. *Econometrica* 1953;21:97–100.
6. Gross O, Wagner R. A continuous colonel blotto game RM-408. Santa Monica (CA): RAND Corporation; 1950.
7. Bellman R. On Colonel Blotto and analogous games. *Siam Rev* 1969;11:66–68.
8. Blackett DW. Some blotto games. *Nav Res Log Q* 1954;1:55–60.
9. Blackett DW. Pure strategy solutions to Blotto games. *Nav Res Log Q* 1958;5:107–109.
10. Tukey JW. A problem of strategy. *Econometrica* 1949;17:73.
11. Golman R, Page SE. General blotto: games of strategic allocative mismatch. University of Michigan, mimeo; 2006.
12. Lazear EP, Rosen S. Rank-order tournaments as optimum labor contracts. *J Polit Econ* 1981;89:841–864.
13. Che Y-K, Gale I. Difference-form contests and the robustness of all-pay auctions. *Games Econ Behav* 2000;30:22–43.
14. Konrad KA, Kovenock D. Multi-battle contests. *Games Econ Behav* 2009;66:256–274.
15. Hillman AL, Riley JG. Politically contestable rents and transfers. *Econ Polit* 1989;1:17–39.
16. Baye MR, Kovenock D, de Vries CG. The all-pay auction with incomplete information. *Econ Theory* 1996;8:291–305.
17. Friedman L. Game-theory models in the allocation of advertising expenditures. *Oper Res* 1996;6:699–709.
18. Robson ARW. Multi-item contests. Australian National University, Working Paper No. 446; 2005.
19. Baye MR, Kovenock D, de Vries CG. The solution to the Tullock rent-seeking game when $R > 2$: mixed-strategy equilibria and mean dissipation rates. *Public Choice* 1994;81:363–380.
20. Alcalde J, Dahm M. Rent seeking and rent dissipation: a neutrality result. *J Public Econ* 2010;94:1–7.
21. Golman R, Page SE. General blotto: games of strategic allocative mismatch. *Public Choice* 2009;138:279–299.
22. Hart S. Discrete colonel blotto and general lotto games. *Int J Game Theory* 2008;36:441–460.
23. Kovenock D, Roberson B. Electoral poaching and party identification. *J Theor Polit* 2008;20:275–302.
24. Kovenock D, Roberson B. Inefficient redistribution and inefficient redistributive politics. *Public Choice* 2009;139:263–272.
25. Kvasov D. Contests with limited resources. *J Econ Theory* 2007;127:738–748.
26. Laslier JF. How two-party competition treats minorities. *Rev Econ Des* 2002;7:297–307.
27. Laslier JF, Picard N. Distributive politics and electoral competition. *J Econ Theory* 2002;103:106–130.
28. Macdonell S, Mastronardi N. Colonel Blotto equilibria: a complete characterization in the two battlefield case. University of Texas, mimeo; 2010.
29. Roberson B. The colonel blotto game. *Econ Theory* 2006;29:1–24.
30. Roberson B. Pork-barrel politics, targetable policies, and fiscal federalism. *J Eur Econ Assoc* 2008;6:819–844.

31. Roberson B, Kvasov D. The non-constant-sum colonel blotto game. CESifo Working Paper No. 2378; 2008.
32. Weinstein J. Two notes on the blotto game. Northwestern University, mimeo; 2005.
33. Borel E, Ville J. Application de la théorie des probabilités aux jeux de hasard. Paris: Gauthier-Villars; 1938. reprinted in Borel E, Chéron A. Théorie mathématique du bridge à la portée de tous. Paris: 1991. Editions Jacques Gabay.
34. Gross O. The symmetric blotto game RM-718. Santa Monica (CA): RAND Corporation; 1951.
35. Laslier JF. Party objectives in the “Divide a dollar” electoral competition. In: Austen-Smith D, Duggan J, editors. Social choice and strategic decisions: essays in honor of Jeffrey S. Banks. New York: Springer; 2003.
36. Szentes B, Rosenthal RW. Three-object two-bidder simultaneous auctions: chopsticks and tetrahedra. Games Econ Behav 2003;44:114–133.
37. Myerson RB. Incentives to cultivate favored minorities under alternative electoral systems. Am Polit Sci Rev 1993;87:856–869.
38. Cox GW, McCubbins MD. Electoral politics as a redistributive game. J Polit 1986; 48:370–389.
39. Lindbeck A, Weibull J. Balanced-budget redistribution as the outcome of political competition. Public Choice 1987;52: 272–297.
40. Dixit A, Londregan J. Redistributive politics and economic efficiency. Am Pol Sci Rev 1995;89:856–866.
41. Dixit A, Londregan J. The determinants of success of special interests in redistributive politics. J Polit 1996;58:1132–1155.
42. Lizzeri A. Budget deficits and redistributive politics. Rev Econ Stud 1999;66:909–928.
43. Lizzeri A, Persico N. The provision of public goods under alternative electoral incentives. Am Econ Rev 2001;91:225–239.
44. Lizzeri A, Persico N. A drawback of electoral competition. J Eur Econ Assoc 2005; 3:1318–1348.
45. Sahuguet N, Persico N. Campaign spending regulation in a model of redistributive politics. Econ Theory 2006;28:95–124.
46. Crutzen BSY, Sahuguet N. Redistributive politics with distortionary taxation. J Econ Theory 2009;144:264–279.
47. Arad A. The tennis coach problem: a game-theoretic and experimental study. Tel Aviv University, mimeo; 2009.
48. Szentes B, Rosenthal RW. Beyond chopsticks: symmetric equilibria in majority auction games. Games Econ Behav 2003;45: 278–295.

ALTERNATING RENEWAL PROCESSES

NILAY TANIK ARGON
Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

Consider a stochastic process $\{Z(t), t \geq 0\}$ that alternates between two states—"up" and "down." Assume that at time zero, it is in state "up" and it stays there for a random amount of time (denoted by U_1) before it moves to state "down." After a sojourn of random duration (denoted by D_1) in the "down" state, it moves back to state "up." This cycle repeats indefinitely. Let U_n and D_n denote the time spent in states "up" and "down," respectively, during the n th cycle. If $\{(U_n, D_n), n \geq 1\}$ is a sequence of independent and identically distributed (i.i.d.) bivariate random variables, then $\{Z(t), t \geq 0\}$ is called an *alternating renewal process*. Note that for an alternating renewal process, the "up times" and "down times" of different cycles should be independent from one another. However, within the same cycle, the time spent in the "up" and "down" states could be dependent.

In the following, we assume that $U_n + D_n$ is an aperiodic random variable with cumulative distribution function $F(x)$. (A random variable X is called a *periodic random variable*, if it only takes on integral multiples of some nonnegative number d ; it is called an *aperiodic random variable* otherwise.) We first present the renewal-type equations for alternating renewal processes that arise by conditioning on $U_1 + D_1$. These equations have the following form:

$$H(t) = D(t) + \int_0^t H(t-x) dF(x), \quad t \geq 0, \quad (1)$$

where $D(t)$ is a known function and $H(t)$ is to be determined. These equations are

particularly useful in obtaining $\lim_{t \rightarrow \infty} H(t)$ by applying the key renewal theorem.

We next present, perhaps, the most important result for alternating renewal processes. This result provides the limiting distribution of $\{Z(t), t \geq 0\}$ and can be proved by applying the key renewal theorem to a renewal-type equation in the form of Equation (1).

Theorem 1. *If $\mathbb{E}[U_1 + D_1] < \infty$, then we have*

$$\lim_{t \rightarrow \infty} \Pr\{Z(t) \text{ is in state "up"}\} = \frac{\mathbb{E}[U_1]}{\mathbb{E}[U_1] + \mathbb{E}[D_1]}.$$

For a proof of this result and the case where $U_n + D_n$ is periodic, the reader is referred to Section 8.8 in Kulkarni [1]. In the remainder of this article, we illustrate the various uses of alternating renewal processes and Theorem 1 with two examples.

Example 1. Suppose that $\{N(t), t \geq 0\}$ is a renewal process with an i.i.d. sequence of aperiodic interrenewal times $\{X_n, n \geq 1\}$ having mean $\tau > 0$ and distribution $G(\cdot)$. Let $C(t) = X_{N(t)+1}$ be the length of the interrenewal time that contains t for some $t \geq 0$. ($C(t)$ is sometimes called the *total life* at time t .) We will obtain the limiting distribution of $C(t)$ (as $t \rightarrow \infty$) by defining an embedded alternating renewal process. For fixed $x > 0$, let

$$Z(t) = \begin{cases} \text{up;} & \text{if } C(t) > x, \\ \text{down;} & \text{if } C(t) \leq x. \end{cases}$$

Then $\{Z(t), t \geq 0\}$ is an alternating renewal process with

$$U_n = \begin{cases} X_n; & \text{if } X_n > x, \\ 0; & \text{otherwise;} \end{cases} \quad (2)$$

and

$$D_n = X_n - U_n. \quad (3)$$

Note that $\{(U_n, D_n), n \geq 1\}$ is a sequence of i.i.d. bivariate random variables, where U_n and D_n are dependent for a given n . Using Equations (2) and (3), we get

$$\mathbb{E}[U_n] = \int_x^\infty u \, dG(u)$$

and $\mathbb{E}[U_n + D_n] = \tau$ for all $n \geq 1$. Now applying Theorem 1, we obtain the limiting distribution of $C(t)$ as

$$\begin{aligned} \lim_{t \rightarrow \infty} \Pr\{C(t) > x\} &= \lim_{t \rightarrow \infty} \Pr\{Z(t) = \text{"up"}\} \\ &= \frac{1}{\tau} \int_x^\infty u \, dG(u). \end{aligned}$$

Note that for any $x > 0$, we have

$$\begin{aligned} \frac{1}{\tau} \int_x^\infty u \, dG(u) &= \frac{\mathbb{E}[X_1 | X_1 > x]}{\mathbb{E}[X_1]} \\ &\times \Pr\{X_1 > x\} \geq \Pr\{X_1 > x\}. \end{aligned}$$

This means that for large t , the interrenewal time containing t is stochastically larger than an arbitrarily picked interrenewal time. This result, which may sound counterintuitive at first, is known as the *inspection paradox*. To understand this paradox, suppose that we choose a t at random, where any t is equally likely to be picked. Then, the probability that the selected t lands in a particular interrenewal time must be proportional to the length of that interval. Hence, the interrenewal time where t falls into is expected to be larger than a generic interrenewal time in some stochastic sense.

Example 2. Consider a queueing system with a single server and Poisson arrivals with rate λ . The queue capacity is $K < \infty$, that is, an arriving customer finding K customers in the queue will be lost. In this queueing system, the service is given in bulk to a group of exactly K customers. More specifically, the server does not start service if there are fewer than K customers waiting; when the number waiting reaches K , all K customers are taken into service and administered service collectively. The service time for the group taken into service at the n th service cycle is denoted by X_n , which is an i.i.d. random variable with mean τ and distribution $G(\cdot)$. We will obtain the limiting probability that the server is busy for this queueing system. We first identify a suitable alternating renewal process. For $t \geq 0$, define

$$Z(t) = \begin{cases} \text{up}; & \text{if the server is busy at time } t, \\ \text{down}; & \text{otherwise.} \end{cases}$$

Assume that at time zero, the server has just started a new service. Then $\{Z(t), t \geq 0\}$ is an alternating renewal process with $U_n = X_n$ and

$$D_n = \begin{cases} \sum_{i=1}^{K-k} Y_i; & \text{if there are } k \text{ arrivals} \\ & \text{during the } n\text{th service cycle for} \\ & k = 0, 1, \dots, K-1, \\ 0; & \text{if there are } K \text{ or more arrivals} \\ & \text{during the } n\text{th service cycle,} \end{cases}$$

where Y_i 's are i.i.d. exponential random variables with rate λ . It is clear from above that the n th down time D_n depends on the n th up time U_n and that $\{(U_n, D_n), n \geq 1\}$ is a sequence of i.i.d. bivariate random variables. We next obtain the mean down time. Let B_n be the random variable denoting the number of arrivals during the n th service cycle (or up time). Then, by conditioning on the service time, for all $k \geq 0$ and $n \geq 1$, we obtain

$$\Pr\{B_n = k\} = \int_0^\infty \frac{e^{-\lambda u} (\lambda u)^k}{k!} dG(u).$$

We next condition on the number of arrivals during an up time and obtain the mean down time as

$$\begin{aligned} \mathbb{E}[D_1] &= \sum_{k=0}^\infty \mathbb{E}[D_1 | B_1 = k] \Pr\{B_1 = k\} \\ &= \frac{1}{\lambda} \sum_{k=0}^{K-1} (K-k) \Pr\{B_1 = k\} \\ &= \frac{1}{\lambda} \sum_{k=0}^{K-1} \frac{K-k}{k!} \int_0^\infty e^{-\lambda u} (\lambda u)^k dG(u). \end{aligned}$$

Now, using the fact that $\mathbb{E}[U_1] = \tau$ and applying Theorem 1, we get

$$\begin{aligned} \lim_{t \rightarrow \infty} \Pr\{\text{server is busy at time } t\} &= \frac{\lambda \tau}{\lambda \tau + \sum_{k=0}^{K-1} \frac{K-k}{k!} \int_0^\infty e^{-\lambda u} (\lambda u)^k dG(u)}. \end{aligned}$$

For further examples on alternating renewal processes, the interested reader is referred to textbooks on stochastic processes such as those by Kulkarni [1], Ross [2], and Tijms [3]. For further applications in reliability systems, see Aven and Jensen [4].

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman and Hall; 1995.
2. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1996.
3. Tijms HC. A first course in stochastic models. New York: John Wiley & Sons, Inc.; 2003.
4. Aven T, Jensen U. Stochastic models in reliability. New York: Springer; 1999.

AMERICAN FOOTBALL: RULES AND RESEARCH

RICK L. WILSON
Oklahoma State University,
Oklahoma City, Oklahoma

AMERICAN FOOTBALL—RULES AND BACKGROUND

The objective of American football is to outscore your opponent, an obviously common feature of many competitive sports. It is a team sport, and often thought as one of the most violent and physical sports. Nonetheless, because of how play has repetitive starts and stops, it has developed into a very strategic game, one in which scientific analysis can be helpful to decision makers, whether it be coaches, players, owners, and so on.

Explaining the game of American football to a novice is a challenge due to the more complex set of rules and the “start and stop” nature of the game, which allows a variety of strategic decisions throughout. For instance, in the game known as *soccer* in the United States, you score a goal by advancing the ball past the opponent’s goalkeeper into the net. There are rules that dictate advancement and play of the game (cannot use your hands and arms to advance the ball, corner kick rules, etc.), but arguably, play is quite straightforward.

The same cannot be said for American football. In general, the objective is to advance the ball into an opponent’s end zone. The ball can be advanced either by running of the football by a player (a running play) or by passing the ball to a teammate (passing play). Points can be scored in a variety of ways. For instance, if the ball is advanced successfully into the opponent’s end zone, a touchdown occurs, and the team scoring the touchdown is awarded six points. There are other ways in which points can be scored (field goals, safeties, extra points, etc.), and they are outlined later in the article.

American football is a timed sport, and the winner is the team with the most points when game time expires. As with other sports, most levels of play now allow for “overtime” in case the game ends in a tie, akin to extra time in “soccer.” This “overtime” process has not always been in place, and as pointed out later, the process differs depending upon what “level” of American football one is playing.

While primarily a US sport, the game is played outside the United States as well. A similar game is played in Canada (Canadian Rules football), while professional and collegiate leagues have existed in Europe and elsewhere. Of note is the attempt by the professional National Football Leagues (NFLs) to globalize interest in the sport by playing select games in Mexico, London, other points in Europe, and so forth.

Both the United States and Canadian version of football have their origins from rugby. Football is played at all age levels, as kids from the United States grow up playing in backyard games and in empty lots in neighborhoods. This tends to become more serious at decidedly younger ages with each passing year, and there are many competitive youth leagues starting from age 5, culminating with the usual teams fielded by the junior high and high school age teams. American football has become immensely popular at the collegiate and professional level, perhaps even surpassing the sport of baseball as “America’s pastime.”

College football tends to be the major revenue producing sport for most major US University athletic departments, and millions of dollars are spent (and earned) on its care and feeding. The NFL is the major professional league in the United States, and the Super Bowl, the championship game of the league played every February, draws millions of viewers from many nations. The Super Bowl is typically the most watched TV event each year. The money spent on attending the games, the popularity of the sport for TV, and how people live their lives around the sport illustrates that American Football has just

as “rabid” fans of any sport around the world. Even college football, where players are not paid to play (they are amateurs by definition), has become a big money operation.

BASIC RULES OF AMERICAN FOOTBALL

Fundamentals

Each team has 11 players on the field at any one time. Because plays start and stop, free substitutions are allowed in between plays. The offense is the team in possession of the ball, and they will try to advance the ball toward its opponent’s end zone. The defense attempts to stop the offense from moving the ball down the field.

At the major levels (college and professional) the game is played on a field that is 100 yd long, not counting the two end zones (each of which is 10 yd long). (Youth leagues may play on slightly smaller dimensions, but even this is getting rarer.) The field is approximately 53.3 yd wide. At the back of the end zones, goal posts are placed in the middle of the field. The goal posts have a crossbar 10 ft high, and have “uprights” on each end of the crossbar. This provides a target for successful field goals and extra points that are attempted via kicking. The goalposts are typically about 18–23 ft apart (depending on the level of play, wider at the younger levels, closer together for the NFL).

Starting the Game

To start the game, a coin toss is held to determine which team gets to go on offense first, and which side of the field a team “defends.” Because the football is thrown and kicked (and it is not particularly aerodynamic), the outside elements (such as the wind) are important aspects to consider in the strategy of the game. The team that wins the coin toss can elect to receive the ball, or choose which side of the field to defend, or defer the choice to the second half. The second team chooses accordingly based upon the first team’s choice. Often this choice is dictated by weather conditions, emotional thoughts of the coaching staff (“Let’s take the ball and ram it down their throats and score right

away!!”), or in the case of deferring, the concept of trying to get the last choice much like the alleged advantage possessed by the home team hitting last in a baseball game.

At the collegiate and professional level, the game is split into four quarters of 15 min each. The first two quarters are referred to as the *first half*, and the second two the *second half*. The team that has the ball at the end of the first quarter retains possession for the start of the second quarter, and likewise for the end of the third quarter, going into the fourth quarter. However, at the end of the first half, possession is not retained, as the start of the third quarter begins with a choice by the team that either deferred at the coin toss or that did not pick its “option” at the beginning of the game.

The game clock stops after certain plays, so it is unlike soccer, which has a continuously running clock. The clock stops when an “incomplete pass” is thrown (described below), when there is a change in who has “possession” of the ball, when a player with the ball runs or is hit out of bounds during the final 2 min of a half, among other reasons. If each team has scored the same number of points at the end of the four quarters (“regulation time”), most levels of play have incorporated an overtime feature that allow additional playing by the teams in attempt to break the tie. The process employed varies by level (e.g., the collegiate overtime process is considerably different than the NFL).

At the start of the game and the start of the second half, and after a team scores, the team “kicks off” to the other team, and the receiving team tries to advance the kickoff by running with the football. When a runner with the ball is tackled, the play is over, the spot where the runner was tackled is marked, and the team with the ball is now on offense.

Note that even when kicking off after scoring, a team can try to recover its own kickoff. This is called an *on-side kick*. A kickoff must travel 10 yd before a member of the kicking team can touch it and try to recover it for themselves. A team might employ this strategy on the kickoff when they are behind late in the game and are trying to quickly make up the deficit. Alternatively, coaches can sometimes use an on-side kick when the

other team is not expecting it for the “surprise” effect.

Running “Plays” from Scrimmage

The location of the ball is called the *line of scrimmage*. Each team must stay on its side of the line of scrimmage before a play is started. The offensive team (the team in possession of the football) has four downs (plays) to advance the ball 10 yd. If they advance the ball at least 10 yd during these plays, they are deemed to have received a “first down,” and then are awarded four more plays, with another “goal” of 10 yd for another first down. If the defense stops the offense from advancing 10 yd in four plays, they (the defensive team) get possession of the ball at the point where the ball was last “downed.” Then, due to the specialization of today’s teams, the offensive specialists for the team that just stopped the other would come out to play, and the team that gave up possession of the ball would bring out its defensive specialists.

A play starts when the offense player called the *center* gives the ball (typically between his legs called a *snap*) to the “quarterback.” The quarterback is somewhat like the field general, and typically handles the ball on every play. The quarterback might hand or lateral (a backward pass) the ball to a running back who will try to advance the ball forward before being tackled (or driven out of bounds). Alternatively, the quarterback might retreat further behind the line of scrimmage and try to throw the ball to a receiver down the field to advance the ball.

If a running back, or anyone who is running the ball, has his knee go down and hit the ground, he is considered tackled, and the next play will start from that point. If a person running the ball loses possession of it, it is a live ball and may be advanced freely by either the offense or the defense . . . so it is possible for a defense to score points as well (say on a fumble return for a “touchdown”).

Similarly, if a pass is thrown and it hits the ground, it is called an *incomplete pass*, and the ball returns to the previous line of scrimmage for the next play. If the offense team catches a pass, then the receiver can

advance the ball until he is tackled. If the defensive team catches the pass, it is called an *interception*. As with a fumble, the defense can advance the intercepted ball toward the other team’s goal line until they are tackled, and they will have possession of the ball at the end of the play.

There is much strategy involved in play calling, both offensively and defensively. Many assistant coaches analyze other team’s tendencies for play calling in various situations, and there exists sophisticated video systems and software to assist in the analysis [e.g., see the Hudl system [1]]. The game has become very specialized and scientific even at this micro level.

Typically, if a team is faced with fourth down and has more than just a yard or two to make a first down, or if it is close to its own goal line (or not close to their opponent’s goal line), it will “punt” the ball away to the other team. Today, punters can routinely kick the ball 50 yd from scrimmage, so this conservative strategy is used to minimize the risk in trying to make a first down and failing; giving the opponent the ball close to its goal line, they might be able to score easily. Of course, the receiving team can advance a punt as well, much like a kickoff. This is yet another case of “conventional coaching wisdom,” where a coach would be considered a high risk taker if they tried to get a first down on fourth down on its “own” side of the field. Given the high pressure of being successful, and the high salary made by most coaches, few are willing to do something that would be viewed as risky or unusual, especially where a single poorly timed failure, counter to conventional wisdom, could lead to the firing of a high profile coach.

Scoring Points in American Football

When a team, by a running play, a passing play, a kickoff return, a punt return, a fumble return, or an interception return crosses its opponent’s goal line in possession of the ball, this is called a *touchdown* and the team earns six points.

After a touchdown, the scoring team then attempts an “extra point.” The ball is placed at the 3-yd line (college) or the 2-yd line

(NFL). Typically, a team will try to score a 1-point extra point by attempting to kick the ball through the aforementioned goalposts. If they are successful, one additional point is awarded, for a total of seven points.

If the team tries to run or pass the ball into the end zone from there, they are awarded two points if they are successful. If they are not successful, they of course score no additional points. This is a more risky choice (typically) than a kicked extra point. Conventional wisdom holds that two-point extra points are successful between 40% and 45% of the time [2]. This number varies in this range each year, and also differs throughout the levels of competitive football.

Note that in both cases, should the defensive team block a kick and advance it all the way down the field into the offensive team's end zone, or likewise return a fumble or interception, the defensive team is awarded two points in both cases. This is a rare occurrence but is worth noting.

Sometimes, when an offensive team decides that they do not want to try to achieve a first down on fourth down, and they are close enough to its opponents end zone, it lines up to attempt a field goal. A field goal is similar to the previously mentioned extra point, except it occurs from a variable (typically further) distance away. If a kicker successfully kicks the ball through the goalposts, the team is awarded three points. Field goal kickers have become very accurate and strong of leg over the last 20 years, and can routinely make a majority of their kicks from 50 yd and less.

Note that if a kicker misses, the other team gains possession of the ball at either the 20-yd line or the previous line of scrimmage, whichever is most advantageous. If an opposing team blocks an attempted field goal, the ball becomes "live" and can be advanced as if it was a fumble.

Finally, a defensive team can score two points for a safety by tackling the offensive team in its own end zone (behind the goal line). This too is a relatively rare occurrence. If a team scores a safety, the other team must execute a "free kick" from the 20-yd line, as the defensive team gets a bonus whereby they receive possession of the ball even after

being awarded the two points for the safety. The same rules apply for the free kick that applies for the kickoff.

Final Comments on American Football Rules

There are many minor differences between college football and professional football (NFL) rules, but many are small issues of such things as how the clock is stopped and started, where "hashmarks" are placed on the field (this impacts the starting point for plays), the width of the goalposts, the definition of when a player is declared tackled, and how penalties are assessed when players are caught in rule infractions, among other "subtle" items. The basic premises, rules, and scoring remain basically the same.

One of the biggest differences between NFL and college football rules are the means for how they deal with games that are tied at the end of the fourth quarter. Both employ overtime, but the NFL uses "sudden death" where the first team that scores wins, while college football uses an overtime process that allows each team an equal number of times where they have the ball ("possessions"). Each approach has been criticized and/or analyzed, and this is one of the areas discussed in this article where research has examined some strategic decisions or processes in American Football.

The next section presents a brief summary of some representative analyses that have been undertaken applying research studies to American Football. This discussion is not meant to be exhaustive nor complete, but provides the reader with an idea on how the rich strategic nature of American Football can be studied using operations research techniques.

EXAMPLES OF RESEARCH IN AMERICAN FOOTBALL—ANALYZING THE CONVENTIONAL COACHING WISDOM

American College Football Overtime

Beginning in 1996, the National Collegiate Athletic Association (NCAA; governing body of major college football) adopted new rules for college football overtime games. Prior to

this time, games were allowed to end in ties. Under these new rules, each team is given one offensive possession starting at the opponent's 25-yd line. The team with the most points after the first overtime period is declared the winner. If the game is tied after the first overtime period, a second period will be played. If the game goes to a third overtime period, and for all subsequent periods, teams are required to attempt a two-point extra point after any touchdown [3].

Before the first overtime period, a coin toss is held. The winner of the coin toss chooses to start on either offense or defense in overtime (they could also opt for end of the field, but this choice is very rare and generally occurs because of weather and/or field conditions). Interestingly, this coin toss determines the possession for all overtime periods, as the team's alternate offense and defense first in each subsequent period that is necessary. Thus, a team that starts on offense in Period 1 will start on offense in every odd-numbered overtime period. This same team would start on defense in Period 2 and all subsequent even-numbered overtime periods.

The overtime method used in college football is designed to be fair to both teams and minimizes the importance of the coin toss. This is in stark contrast to the NFL's sudden death overtime system. In sudden death, the first team that scores wins the game, regardless of whether or not the other team had the opportunity to be on offense (i.e., "has possession of the ball"). This system has long been criticized for favoring the team that gets possession of the ball first, implying that the team that wins the coin toss has an unfair advantage and usually wins the game.

Between the years of 1974 and 2003, 28% of all NFL overtime games ended with the team winning the coin toss getting to go on offense first, scoring, and winning the game without allowing the other team to have the ball. Critics charge that this high percentage demonstrates that the system is unfair to the team that has the misfortune of losing the coin toss [4]. In the most recent NFL season, this number was even more extreme, as Clayton [5] reports that 43.4% of NFL's overtime games during the 2008–2009 season were won in the first possession by the

team that won the coin toss, and that overall, 63% of the overtime winners won the coin toss.

Even though both teams get an offensive possession in college football overtime rules, conventional coaching wisdom dictates that the team that wins the coin toss should start on defense (somewhat like being the home team in baseball, getting the "last at-bat"). Since this conventional wisdom has been practiced in all but four occasions (out of more than 390 games to date), there seems to be an implied advantage to the winner of the coin toss.

Research showed that in college football, there was not a large advantage to being on defense first [3]. Coaches were surveyed, and they thought that teams who started on defense won as much as 75% of the time. The actual figures through 2005 were 56%, and in fact the result was nearly 50% for the years from 2001 to the present. From 2001 through 2003, the team that went first actually won nearly 60% of the time!

The studies found that one interesting factor is the "pressure" of being on defense first and having to match the opponent if they score a touchdown on its first possession (A team that scores a touchdown on its first possession wins approximately 70% of the time). There is evidence that if you are the coach of a good offensive team, or fear that your team is overmatched the longer the overtime periods continue, you might be better-off by breaking tradition and choosing to take the ball first in overtime.

At least college football overtime rules seem fair to both teams. Interestingly, the NFL has again opted not to change its overtime process for the 2009 season, even with the evidence that its overtime process gives an unfair advantage to the coin toss winner. The reason—not enough support from teams, and players who expressed "safety" concerns if current rules were to be changed. Funny, what making hundreds of thousands of dollars per game will do to your competitive spirit.

The Fourth Down Punting Strategy

In the previous description of American Football rules and the typical process of a game,

the concept of punting away the ball on fourth down was discussed. Oftentimes, coaches opt to punt the ball on fourth down and inches rather than to go for a first down due to their cautious nature (why take a chance if it might cause me to lose my million-dollar job?).

Romer [6] studied the specific fourth down punting decision in the NFL and found that conservative decisions made were counter to a maximize wins objective, and estimated that this cost teams at least one win every three years.

Taking this research study to an extreme, Pulaski Academy (Arkansas) High School coach Kevin Kelley [7] has adopted an offensive philosophy where his team never punts—they have not punted for 20 straight games. This past season, they won the 5A State Championship in Arkansas by applying this philosophy. Coach Kelley basically uses a form of expected value calculations in justifying his very nontraditional approach to fourth down, and his results (83% winning percentage during his time there) speak for themselves. Coach Kelley also uses onside kicks as a regular strategy. He also attributes some of his strategy to the output of ZEUS (pigskinrevolution.com), an analytic computer-based tool that studies strategic decision making in the NFL.

Two-Point Conversions and Decision Analysis

In the days before overtime in college football, coaches were oftentimes faced with tough decisions late in the game. When scoring a touchdown late, and trailing by 1, should a coach go for the win, or settle for a tie?

An even more interesting scenario was raised in the classic Janssen and Daniel article [8] on using decision analysis in deciding when to go for a two-point conversion. The game under study was the 1967 Harvard-Cornell game, but it very well could have been the 1969 Arkansas-Texas game (a game of the two top-rated teams at the time), or one of the most famous bowl games of all time, the 1984 Orange Bowl game between Miami (Florida) and Nebraska.

In all of these cases, one team was faced with a 14-point deficit, and scored a touchdown late in the game to bring them within

eight points. Not much time remained, so the best scenario that the team trailing could hope for would be to hold the other team, get the ball back, and then score another touchdown with little or no time remaining in the game. Before overtime, a coach would be interested (one would assume) in trying to maximize the likelihood that his team would win.

Janssen and Daniel discussed the implications and circumstances surrounding Cornell's decision to go for two points after its first touchdown that made the score 14-6. The team opted to go for two points and failed, and when they scored later in the game and failed again at the two-point conversion; it was left with a 14-12 loss, and was criticized for not following conventional wisdom (which would have been to go for the kick after the first touchdown, then do the all-or-nothing two-point conversion after the second touchdown). Janssen and Daniel explain that the conventional wisdom strategy is oftentimes suboptimal and devalues a tie.

Interestingly, coach Darryl Royal of Texas used the optimal strategy in his team's 15-14 victory over Arkansas. Most people recall the gambling fourth down pass that was completed in that game—few recall coach Royal's gutsiness in choosing the arguably "correct" way of executing extra points when Texas scored at the start of the fourth quarter to make it 14-6. Texas went for a two-point extra point, made it, and the score stood at 14-8 when Texas scored late in the game. A simple one-point kick was all Texas needed to win 15-14.

Nebraska's coach Tom Osborne was roundly praised for his decision to go for two late in the 1984 Orange Bowl, when a tie almost assuredly would have given Nebraska the so-called National Championship. Nebraska trailed 31-17 late in the fourth quarter, when they scored to make it 31-23; they then kicked the one-point extra point to make it 31-24. Then, scoring with 47s left, at 31-30, Osborne elected to go for two points; it failed, and Miami won the game, and Osborne won the respect of the nation. Unfortunately, decision analysis indicated that perhaps the better decision was to go for two when it was 31-23!!

THE BOWL CHAMPIONSHIP SERIES (BCS), COMPUTER RATINGS, AND VOTER POLL RESEARCH

College Football—No Play-off?

NCAA Division I-A college football remains the only major US college sport that does not utilize a play-off to determine its champion. There are a variety of reasons for this, not the least of which is money [9].

For years, voter polls determined the so-called “Mythical National Champion.” A quick look through college football history will show many seasons where voter polls disagreed with the “champion.” This phenomenon has seemed to trouble football fans much more during the last 20 years, perhaps due to society’s increased focus on “Who’s No 1?” It is also worth noting for those not familiar with Division I-A football that with nearly 120 teams playing a normal schedule of 12 games each, many teams may have the same record or same number of wins, and their win–loss record has never been solely utilized as a way of awarding the mythical national championship.

A series of controversial endings to the college football season in the late 1980s and early 1990s led influential people in college football to attempt to develop a process in which a single champion could be named. This led to the creation of the Bowl Coalition process, which started in 1992. Unfortunately, not all major collegiate conferences participated in this process, and so split or controversial national champions still occurred in 1993, 1994, and 1997. The continued controversy and the hope for a more robust method for selecting the participants led to the creation of the Bowl Championship Series (BCS) in 1998, which included “objective” measures of team strength by utilizing computer rankings (in addition to voter polls). However, controversy has been the rule rather than the exception since the BCS’s inception, and it has seemingly changed its approach in determining the top two teams each year it has existed. The controversy has been so pervasive that congressional hearings were held during the last few years to discuss the “BCS Mess.”

Team Rankings

The root of the problem is the BCS leadership’s failure to explicitly state criteria for determining the final two teams. Additionally, the mathematical models used in the BCS approach have often come under criticism and scrutiny. There has been a fair amount of research in the academic literature on college football ranking methods that date back many years ago [10]. Most of the published research has focused on how a particular approach would rank teams in a given season, and then argue, using some form of face validity, that one approach is superior to another.

Research continues in a variety of areas related to ranking—a quick scan of the academic literature turns up a number of different recent studies [11–14]. As the BCS and academic researchers continue looking for the holy grail of ranking methodologies, one of the challenges faced is how to objectively determine which ranking method performs best. There appears to be no satisfactory solution to the problem at present, but academic and college football fans keep searching.

Voter Poll Findings

One of the key components of the formula that makes up the BCS calculations is the use of voter polls. Human “experts” (at least expert in theory) rank the football teams. For many years, these rankings have been controversial, and the voters have claimed to possess many biases and inaccuracies. There has been a fair amount of research seeking to validate or invalidate these claims.

A study by Goff [15], validated in part by Lebovic and Sigelman [16], found that there was “path dependence” in voting in the AP voter poll. Thus, a team’s standing at the beginning of the season, when very little actual performance data is known about team strength to drive voter decisions, has a significant impact on the final ranking of the team. This has led some to call for no opinion polls prior to the midway point of the season.

Campbell *et al.* [17] observed the 2003 and 2004 voting polls and found that a team that appeared more on television had bigger adjustments made in its poll standings, all

other things being equal. Logan [18] studied AP voter polls and found some interesting results that countered “conventional wisdom” of voter polls—that it is better to lose later in the season than early, that voters do not pay attention to the strength of the opponent, and that the benefit of winning by a large margin is negligible. Stone [19] also studied the voters and found systematic and statistically significant errors from a Bayesian standpoint. Findings included evidence that voters overreact to losses by higher ranked teams, improve rankings excessively after wins at home, and worsen rankings excessively after large margin losses on the road, especially for higher ranked teams. Decisive wins against unranked teams are unappreciated by voters, as well.

As major college football in the United States seeks to try to sort out the “BCS Mess,” even involving the strategic focus list of new US President Barack Obama, the quantitative research undertaken by operations researchers will continue to play a role in trying to “optimize” the decisions.

MARKET EFFICIENCY AND OTHER ISSUES IN THE NFL BETTING MARKET

Legal or not, betting on sporting events has long held the interest of the general population, and has served as an easily accessible data source for academicians seeking to test financial theories of market efficiency. Research has also been undertaken in public to try to find strategies that can be used to do more than break even in betting. There is a rich history of published research in this area as it is related to American Football.

Stern’s 1991 study [20] that deals with the probability of winning an NFL football game is a great starting point in this area. Stern’s empirically developed premise was that the probability of winning a football game is a random normal variable with a mean equal to the Las Vegas determined betting “point spread” and a standard deviation of 14. The point spread is a surrogate indicator of the difference in team strength but, in practice, it also considers additional

factors. It is well known that Las Vegas casinos set the odds, or point spreads, for football games such that the amount of money bet on each team is approximately equal. As those who accept bets receive a certain percentage of all winning bets (and keep all losing bets), this approach ensures steady Casino profits.

As an example of this premise, consider two teams—Baltimore and New York—and the Vegas odds have chosen Baltimore as a seven-point favorite. Stern’s premise would indicate that Baltimore would have a $Z(7/14) = 0.6914$ or 69.14% chance of winning the game, while New York would have a $1 - 0.6914 = 0.3086$ or 30.86% chance.

Stern’s work has oftentimes been used in trying to determine strength of schedule for teams, even in a ranking context. Other studies have focused more on the wagering issue.

Numerous studies have appeared in the marketing literature that analyze the professional and college football betting markets [21], using an analogy with securities markets. Past research included statistical approaches to test for market efficiencies [22], while others have explored specific betting strategies to determine if they lead to unusual profits. A comprehensive paper by Sapra [23] has recently looked at intraseason efficient and interseason overreaction to the NFL betting market. They also reference a number of the most recent studies in this area. They conclude that point spread by itself indicates the likelihood of victory, and thus the market is efficient. They point out the need for additional research on variations from year to year.

CONCLUSION

This brief article has reviewed the rules and playing process of American football. In the second half of the paper, some relevant research related to college and profession American football has been reviewed. This review is not meant to be comprehensive, or necessarily in depth, but gives the reader an idea of some of the interesting and practical research questions that exist in this domain for fellow operations researchers.

REFERENCES

1. Dreaming of fields. Available at http://www.economist.com/business/businesseducation/displaystory.cfm?story_id=12451400. Accessed 2009 May 28.
2. Johnson G. Two point conversion turns 50, NCAA News. Available at <http://www.ncaa.org/wps/ncaa?ContentID=35763>. Accessed 2009 May 28.
3. Rosen P, Wilson R. An analysis of the defense first strategy in college football overtime games. *J Quant Anal Sports* 2007;3(2):1–17.
4. Peterson I. Football's overtime bias. Available at https://www.maa.org/mathland/math_trek_11_08_04.html. Accessed 2009 May 20.
5. Clayton J. NFL's overtime rules won't change. Available at <http://sports.espn.go.com/nfl/news/story?id=3993657>. Accessed 2009 May 15.
6. Romer D. Do firms maximize? Evidence from professional football. *J Polit Econ* 2006; 114:340–365.
7. Matheson L. Pulaski academy coach kelly explain no punting philosophy. Available at <http://footballrecruiting.rivals.com/content.asp?cid=888058>. Accessed 2009 May 26.
8. Janssen CTL, Daniel TE. A decision theory example in football. *Decis Sci* 1984;15: 253–259.
9. Wilson R. Validating a Division I-A college football season simulation system. Proceedings of the Winter Simulation Conference. Orlando (FL); Dec 2005.
10. Wilson R. Ranking college football teams: a neural network approach. *Interfaces* 1995; 25(4):44–59.
11. Beard TR. Who's number one? - ranking college football teams for the 2003 season. *Appl Econ* 2009;41(3):307–310.
12. Annis DH, Craig BA. Hybrid paired comparison analysis, with applications to the ranking of college football teams. *J Quant Anal Sports* 2005;1:3–33.
13. Wang T, Keller JM. Iterative ordering using fuzzy logic and application to ranking college football teams. Annual Conference of the North American Fuzzy Information Processing Society—NAFIPS; Banff, Alberta Canada. Volume 2; 2004. pp. 729–733.
14. Mease D. A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins. *Am Stat* 2003;57(4):241–248.
15. Goff B. An assessment of path dependence in collective decisions: evidence from football polls. *Appl Econ* 1996;28:291–297.
16. Lebovic JH, Sigleman Lee. The forecasting accuracy and determinants of football rankings. *Int J Forecast* 2001;17:105–120.
17. Campbell N *et al*. Evidence of television exposure effects in AP top 25 college football rankings. *J Sports Econ* 2007;8:425–434.
18. Logan TD. Whoa Nellie! empirical tests of college football's conventional wisdom. Working paper 13956, NBER Working paper series. 2007.
19. Stone DT. Testing Bayesian updating with the AP top 25. Working paper, Johns Hopkins University, 2008.
20. Stern H. On the probability of winning a football game. *Am Stat* 1991;45:116–123.
21. Golec J, Tamarkin M. The degree of inefficiency in the football betting market. *J Financ Econ* 1991;31:311–323.
22. Gandar J *et al*. Testing market rationality in the point spread betting market. *J Finance* 1988;43:995–1007.
23. Sapra SG. Evidence of betting market intraseason efficiency and interseason overreaction to unexpected NFL team performance 1988–2006. *J Sports Econ* 2008;9(5): 488–503.

AN INTRODUCTION TO LINEAR PROGRAMMING

JAMES J. COCHRAN
Department of Marketing and
Analysis, College of Business,
Louisiana Tech University,
Ruston, Louisiana

A PRELIMINARY OVERVIEW OF OPTIMIZATION

At its most elemental level, optimization refers to the identification of the best element(s) from a domain (or allowable set of available alternatives) with respect to the application of some function to these elements. The unknown values of these elements are referred to as *decision variables*, the set of available alternatives (i.e., collection of values of the decision variables that may be considered) is referred to as the *feasible set*, the function used to evaluate the relative performance of these elements is referred to as the *objective function*, and the coefficients and exponents that operate on the decision variables in the objective function are referred to as *parameters*. An element from the feasible set that generates the optimal value with respect to the objective function is called an *optimal solution*.

In the simplest case of optimization, one seeks to minimize or maximize a real objective function through systematic evaluation of this function for values of real variables that belong to the feasible set. The systematic approach used to identify various elements of the feasible set, evaluate the objective function for these various elements of the feasible set, and find an optimal solution is referred to as a *solution algorithm*.

Depending on the characteristics of the optimization problem, a solution algorithm for a class of optimization problems may identify either a global or local optimal solution. A *global optimal solution* is optimal over the

entire feasible set, while a *local optimal solution* is optimal over a subset (or local neighborhood) of the feasible set. Some optimization problems have more than one global optimal solution; in such cases, these solutions are referred to as *alternative optima*. Ideally, a solution algorithm will identify and confirm a global optimal solution(s) for a problem in a reasonably short period of time.

These concepts combine to form the basis of optimization theory, which is an important area of application and research in operations research. Optimization plays an important role in many areas of application including manufacturing and production; finance and investing; engineering; marketing, logistics, transportation, distribution, and supply-chain management; network design; and telecommunications.

UNCONSTRAINED VERSUS CONSTRAINED OPTIMIZATION

Optimization problems may be classified as *unconstrained* or *constrained*. An unconstrained maximization problem is defined as

find $\mathbf{X}^* = \operatorname{argmax}_{\mathbf{X}} f(\mathbf{X})$ where $f: \mathbb{R}_n \rightarrow \mathbb{R}$,

where

$\mathbf{X}$ is an element of n -dimensional real space;

$f(\cdot)$ is the objective function that maps an n -dimensional element onto the set of real numbers.

An unconstrained minimization problem is defined similarly through the use of the *argmin* function. In unconstrained optimization problems, no restrictions are placed on the potential solutions $\mathbf{X}$ (see ***Methods for Large-Scale Unconstrained Optimization*** for a technical discussion of unconstrained optimization).

In constrained optimization, one is again attempting to identify a best element from the feasible set with respect to the objective

function. However, instead of searching over all elements in n -dimensional real space (as is done in unconstrained optimization), the search is now limited to a predefined subset of the n -dimensional real space.

Constrained optimization problems can be represented in the following manner:

$$\begin{aligned} \text{find } \mathbf{X}^* \in \operatorname{argmax}_{\mathbf{X} \in \mathbf{A}_n} f(\mathbf{X}) \\ \text{where } f : \mathbf{A}_n \rightarrow \mathbb{R} \text{ and } \mathbf{A}_n \subset \mathbb{R}_n, \end{aligned}$$

where $\mathbf{A}_n$ is the predefined feasible space containing all values of $\mathbf{X}$ that satisfy all limitations that have been imposed on values of the decision variables.

The limitations that are imposed on values of the decision variables are referred to as *constraints*. If the objective function and all constraints are linear with respect to the decision variables, the problem is a *linear programming* problem.

A SIMPLE LINEAR PROGRAMMING EXAMPLE

The process of taking information from a real situation and representing it symbolically is referred to as *mathematical modeling* or *problem formulation*. Consider the following simple example:

Suppose a producer of single serve packets of powdered fruit drink mixes wants to decide how much of two products, *Zesty Lemon* and *Super Zesty Lemon*, to produce on a given day. Each of these products consists of only two ingredients (sugar and powdered lemon juice); a single serve packet of *Zesty Lemon* contains two ounces of sugar and one ounce of powdered lemon juice, while a single serve packet of *Super Zesty Lemon* contains two ounces of sugar and two ounces of powdered lemon juice. Eight hundred ounces of sugar and six hundred ounces of powdered lemon juice are available for daily production of *Zesty Lemon* and *Super Zesty Lemon*. An ounce of sugar costs the producer 3¢ and an ounce of powdered lemon juice costs the producer 5¢, and the producer can sell packets of *Zesty Lemon* and *Super Zesty Lemon* for 21¢ and 32¢, respectively. How many packets of *Zesty Lemon* and *Super Zesty Lemon* powdered drink mix should the producer make on a daily basis in order to maximize profit?

In this problem the goal is to maximize profit, and the decision variables (unknown values that will ultimately determine how well this goal is met) are the number of packets of *Zesty Lemon* and *Super Zesty Lemon* powdered drink mix to produce. To facilitate the formulation of this problem, arbitrarily designate the decision variables as X_1 (the number of packets of *Zesty Lemon* to produce) and X_2 (the number of packets of *Super Zesty Lemon* to produce).

A packet of *Zesty Lemon* powdered drink mix sells for 21¢ and uses 2 ounces of sugar (each of which costs 3¢) and 1 ounce of powdered lemon juice (which costs 5¢). This information can be used to determine that the profit per packet of *Zesty Lemon* powdered drink mix is

$$21¢ - [2(3¢) + 1(5¢)] = 10¢.$$

Similarly, since a packet of *Super Zesty Lemon* powdered drink mix sells for 32¢ and uses 2 ounces of sugar (each of which costs 3¢) and 2 ounces of powdered lemon juice (each of which costs 5¢), the profit per packet of *Super Zesty Lemon* powdered drink mix is

$$32¢ - [2(3¢) + 2(5¢)] = 16¢.$$

Thus, the profit the producer will earn can be stated functionally (in cents) as

$$10X_1 + 16X_2,$$

and the objective function of this problem (in cents) is

$$\text{Maximize } 10X_1 + 16X_2.$$

The coefficients applied to the decision variables in the objective function are called the *objective function coefficients*. In this problem, 10 is the objective function coefficient for *Zesty Lemon* powdered drink mix and 16 is the objective function coefficient for *Super Zesty Lemon* powdered drink mix.

If this were a complete statement of the problem, the manufacturer could produce an unlimited amount of *Zesty Lemon* and *Super Zesty Lemon* powdered drink mixes and generate an infinite profit. However, limitations

on the amount of sugar and powdered lemon juice available for production prevent the producer from pursuing this strategy.

The producer has 800 ounces of sugar available, and it takes 2 ounces of sugar to make a single serve packet of *Zesty Lemon* and 2 ounces of sugar to make single serve packet of *Super Zesty Lemon*. Since the producer cannot use more than the 800 ounces of sugar available in the daily production, this implies the following limitation:

$$2X_1 + 2X_2 \leq 800.$$

The coefficients applied to the decision variables in a constraint are called *constraint coefficients*, and the value on the right side of the inequality is called the *right hand side value* of the constraint. In this constraint, the expression on the left hand side of the inequality represents the total ounces of sugar used in the daily production and the value on the right hand side of the inequality represents the ounces of sugar available.

The producer also has 600 ounces of powdered lemon juice available, and it takes 1 ounce of powdered lemon juice to make a single serve packet of *Zesty Lemon* and 2 ounces of powdered lemon juice to make a single serve packet of *Super Zesty Lemon*. The producer cannot use more than 600 ounces of powdered lemon juice in the daily production, and this implies the following limitation:

$$X_1 + 2X_2 \leq 600.$$

In this constraint, the expression on the left hand side of the inequality represents the total ounces of powdered lemon juice used in the daily production and the value on the right hand side of the inequality represents the ounces of powdered lemon juice available.

These two constraints represent limits on values of the decision variables (the number of packets of *Zesty Lemon* and *Super Zesty Lemon* powdered drink mix to produce). Note also that it is impossible to produce a negative number of packets of either *Zesty Lemon* or *Super Zesty Lemon* powdered drink mix. These limitations, which are referred to as *nonnegativity constraints*, can be stated in

the following manner

$$X_1 \geq 0.$$

$$X_2 \geq 0.$$

The nonnegativity constraints combine with the constraints on available ounces of sugar and powdered lemon juice to define the feasible region A_n , and the complete problem formulation is

$$\text{Maximize } 10X_1 + 16X_2.$$

$$\text{subject to } 2X_1 + 2X_2 \leq 800$$

$$X_1 + 2X_2 \leq 600$$

$$X_1 \geq 0$$

$$X_2 \geq 0.$$

This simple example is provided for illustrative purposes; real applications of linear programming can (and often do) encompass millions of decision variables and constraints [1]. Also note that the relationship between the left and right hand sides of a constraint may be less than or equal to ($\leq$), greater than or equal to ($\geq$), or equal ($=$); constraints that feature strict inequality relationships ($<$ or $>$) between the left and right hand sides are typically avoided. Finally note that in the expressions that represent the constraints, all decision variables have been placed on the left side and all constants have been placed on the right side; although not necessary, this format generally facilitates understanding of the model and will be utilized throughout this discussion.

LINEAR PROGRAMMING ASSUMPTIONS

Four conditions must be met in order for a linear programming formulation to provide an appropriate representation of a problem scenario. These conditions are as follows:

- *Additivity.* *Nonlinear interactions* between decision variables cannot occur. The contribution made by any decision variable to the objective function must be independent of the values of all other decision variables.

Furthermore, the contribution made by any decision variable to the left hand side of each constraint must be independent of the values of all other decision variables.

- *Proportionality.* Relationships must be *linear*. The contribution made by a decision variable to the objective function must be proportional to the value of the decision variable. Furthermore, the contribution made by a decision variable to the left hand side of each constraint must be proportional to the value of the decision variable.
- *Divisibility.* Decision variables must be permitted to take on any value in a continuous range.
- *Deterministic Nature.* The value of each parameter (i.e., each objective function coefficient, constraint coefficient, and right hand side value) must be known with certainty.

Alternative approaches for modeling problems that do not meet these conditions have been developed. For example, problems that violate the additivity and/or proportionality assumptions may be modeled using *nonlinear programming*.

Problems that violate the divisibility assumption may be modeled with *integer programming*. Such problems may be designated as *integer* (all decision variables are restricted to integer values) or *mixed integer* (a subset of the decision variables are restricted to integer values). Combinatorial optimization problems generally fall into the category of integer programming problems.

Problems that violate the deterministic nature assumption can be modeled with *stochastic programming* (when the probability distributions associated with the values of the unknown parameters are known) or *sample-based programming* (when sample data is used to estimate the values of the unknown parameters). *Robust programming* and *fuzzy programming* are other approaches to dealing with uncertainty in the values of the parameters.

In addition, problems that have several inherent goals can be modeled with *goal*

programming, and problems for which only a subset of constraints must be satisfied can be modeled with *disjunctive programming*. For detailed discussions of several of these concepts, see ***Nonlinear Multiobjective Programming; Sampling Methods; Solving Stochastic Programs; and Stochastic Mixed-Integer Programming Algorithms: Beyond Benders' Decomposition.***

SOLVING LINEAR PROGRAMMING PROBLEMS GRAPHICALLY

Solving a linear programming problem to optimality initially appears daunting. Consider the powdered fruit drink mix problem discussed in an earlier section; how can one determine the number of packets of *Zesty Lemon* and *Super Zesty Lemon* powdered drink mix to produce in order to maximize profit? An infinite number of solutions are feasible (i.e., satisfy all constraints). One could utilize a *greedy approach* and first produce the maximum amount of the product that is most profitable on a per unit basis (*Super Zesty Lemon*), then make as much of the least profitable product (*Zesty Lemon*) as possible with the resources that remain.

Given the amount of sugar available (800 ounces), daily production of *Super Zesty Lemon* cannot exceed 400 single serve packets (recall that each single serve packet of *Super Zesty Lemon* includes 2 ounces of sugar). The available powdered lemon juice (600 ounces) limits daily production of *Super Zesty Lemon* to 300 single serve packets (each single serve packet of *Super Zesty Lemon* includes 2 ounces of powdered lemon juice). Maximum daily production of *Super Zesty Lemon* can therefore not exceed 300 single serve packets, which would generate a profit of 4800¢ (or \$48.00). This production strategy would use all available powdered lemon juice, and 200 ounces of sugar would remain unused. Since production of *Zesty Lemon* requires powdered lemon juice, no packets of this product can be produced with the remaining resources. Perhaps a solution that would make better use of the sugar would produce a superior profit.

Since a single serve packet of *Zesty Lemon* requires only a single ounce of powdered

lemon juice, one can consider producing the maximum quantity of *Zesty Lemon* possible and then producing as much *Super Zesty Lemon* as possible with the remaining resources. Given the amount of sugar available (800 ounces), production of *Zesty Lemon* could not exceed 400 single serve packets (each single serve packet of *Zesty Lemon* includes 2 ounces of sugar). The available powdered lemon juice (600 ounces) limits production of *Zesty Lemon* to 600 single serve packets (each single serve packet of *Zesty Lemon* includes 1 ounce of powdered lemon juice). Maximum daily production of *Zesty Lemon* therefore cannot exceed 400 single serve packets, which would generate a profit of 4000¢ (or \$40.00). This production strategy would use all available sugar, and 200 ounces of powdered lemon juice would remain unused. At this point it is not possible to produce packets of *Super Zesty Lemon* (which require sugar). Obviously, this solution is inferior (by \$8.00) to the greedy solution identified in the previous paragraph.

One may also wish to consider solutions that involve production of some amount of each product; such a solution may make better use of the available resources (sugar and powdered lemon juice) and produce a superior profit. Note that for every packet of *Zesty Lemon* that is removed from the solution identified in the previous paragraph, one sacrifices 10¢ of profit and makes available 2 ounces of sugar and 1 ounce of powdered lemon juice. These resources could be combined with one of the remaining ounces of the powdered lemon juice to produce one packet of *Super Zesty Lemon*, which will generate 16¢ of profit (for a net gain of 6¢ of profit). If this were done 200 times, the number of packets of *Zesty Lemon* produced would decrease by 200 and the corresponding contribution to profit would fall by 2000¢, while the number of packets of *Super Zesty Lemon* produced would simultaneously increase by 200 and the corresponding contribution to profit would be 3200¢. The net result would be an increase of 1200¢ (or \$12.00) of profit over the strategy of producing only packets of *Zesty Lemon*. The new solution is to produce 200 packets of *Zesty Lemon*

and 200 packets of *Super Zesty Lemon* and generate a profit of 5200¢ (or \$52.00). This result suggests that if one can effectively manage the trade-offs in the marginal profits generated by the two products, one can potentially identify superior solutions to the problem (and perhaps even identify an optimal solution to the problem).

So how does one manage these trade-offs in a systematic manner? This task is surprisingly straightforward and simple. In the case of a linear programming problem involving only two decision variables, an optimal solution can be found using basic geometry and algebra in tandem. One can

- plot the areas represented by the constraints (i.e., identify the area that is feasible for each constraint);
- find the feasible region for the problem (the intersection of the areas that are feasible with respect to the constraints);
- set the objective function equal to some arbitrary value and plot the resulting equality; (this is referred to as an *isoprofit line* or *objective contour*);
- move this isoprofit line in a parallel fashion through the feasible region in the direction of improvement with respect to the objective function, stopping when the parallel-shifted isoprofit line is tangent to the feasible region;
- convert the constraints that intersect at this optimal solution to equalities, and then solve these equalities with respect to the decision variables; this will yield the optimal values of the decision variables associated with this optimal solution; and
- substitute the optimal values of the decision variables into the objective function and solve; this will yield the optimal objective function value.

Again consider the powdered fruit drink mix problem from the previous section; because this problem has only two decision variables, it can be solved graphically using the steps outlined above.

- Plot the areas represented by the constraints (i.e., identify the area that is feasible for each constraint).

Let the x -axis represent X_1 (the number of packets of *Zesty Lemon* powdered drink mix produced) and the y -axis represent X_2 (the number of packets of *Super Zesty Lemon* powdered drink mix produced). The regions that satisfy each of the constraints and their intersection (the feasible region) are shown on the graph in Fig. 1:

- Set the objective function equal to some arbitrary value and plot the resulting isoprofit line. Set the objective function equal to 3200¢, that is,

$$10X_1 + 16X_2 = 3200,$$

and plot the resulting isoprofit line (Fig. 2).

- Move this isoprofit line in a parallel fashion through the feasible region in the direction of improvement with respect to the objective function, and stop when the parallel-shifted isoprofit line is tangent with the feasible

region; this is an optimal solution (Fig. 3).

- Convert the constraints that intersect at this optimal solution to equalities, and then solve these equalities with respect to the decision variables; this will yield the optimal values of the decision variables associated with this optimal solution.

After converting the sugar and powdered lemon juice constraints to equalities, solving for X_1 (packets of *Zesty Lemon*) and X_2 (packets of *Super Zesty Lemon*) at this point yields

$$2X_1 + 2X_2 = 800$$

$$\begin{array}{r} -(X_1 + 2X_2 = 600) \\ \hline X_1 = 200, \end{array}$$

and

$$\begin{aligned} X_1 + 2X_2 &= 600 \rightarrow 200 + 2X_2 \\ &= 600 \rightarrow X_2 = 200. \end{aligned}$$

The optimal values of the decision variables X_1 (packets of *Zesty Lemon*) and X_2 (packets of *Super Zesty Lemon*) are 200 and 200, respectively.

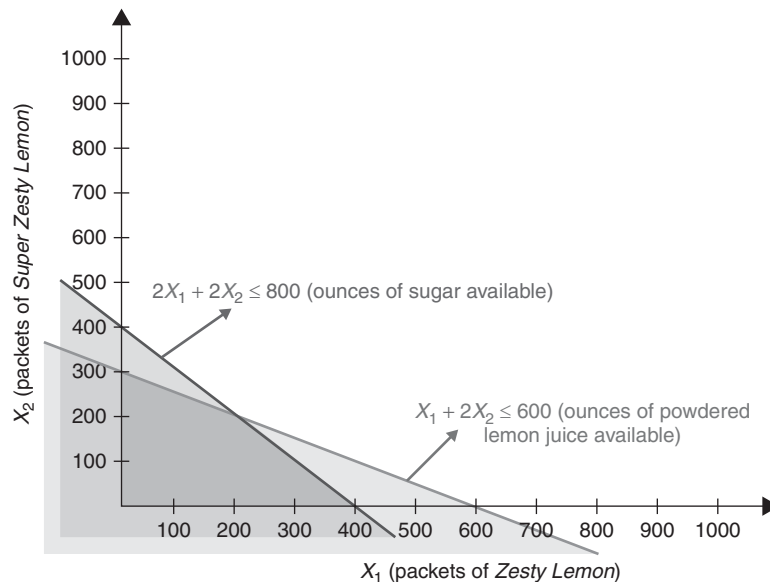


Figure 1. Feasible Region for the Powdered Drink Mix Problem.

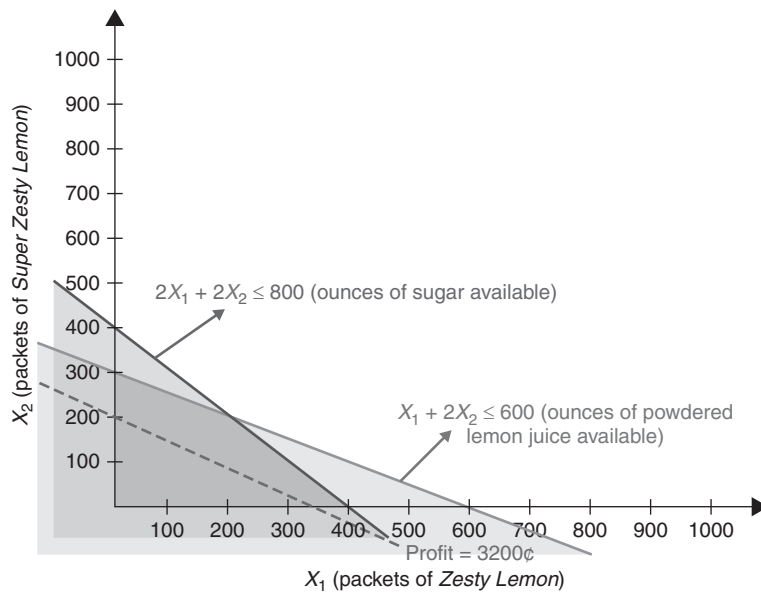


Figure 2. The Iso-Profit Line Corresponding to a 3200¢ Profit.

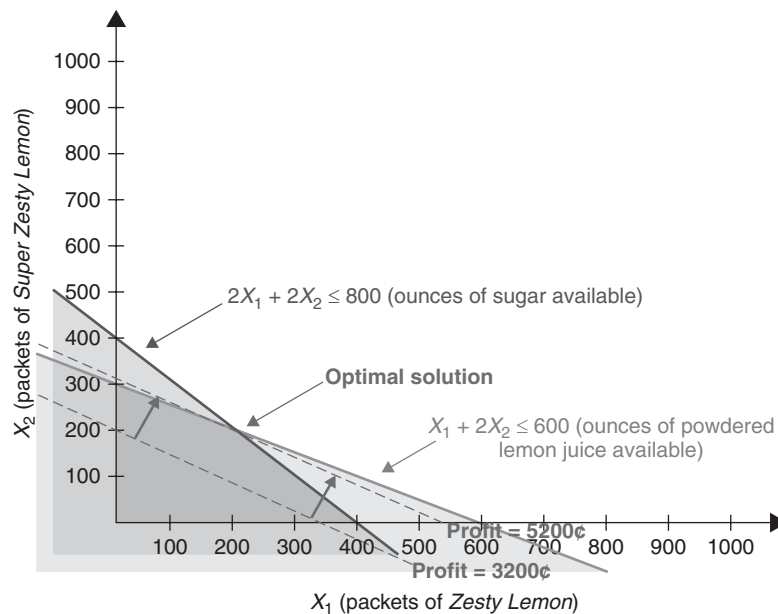


Figure 3. Parallel Shifting the Iso-Profit Line in the Direction of Improvement until it is Tangent to the Feasible Region.

- Substitute the optimal values of the decision variables into the objective function and solve; this will yield the optimal objective function value.

Substitution of the optimal values of the decision variables X_1 (packets of *Zesty Lemon*) and X_2 (packets of *Super Zesty Lemon*) yields

$$\begin{aligned}
 10X_1 + 16X_2 &= 10(200) + 16(200) \\
 &= 5200,
 \end{aligned}$$

and so the optimal objective function value is 5200¢ (\$52.00).

SOLVING LINEAR PROGRAMMING PROBLEMS THROUGH ENUMERATION

Each constraint has a perimeter that can be found by setting the left hand side of the constraint equal to the right hand side of the constraint. If feasible, the points where perimeters of constraints (including nonnegativity) intersect are called *extreme points* and are extremely important. By the *Fundamental Theorem of Linear Programming*, an optimal solution to the linear programming problem (if one exists) can be found at an extreme point due to the linearity of the objective function and the convexity of feasible region (see Dantzig [2] for a detailed explanation and Martin [3] for a succinct proof). This discussion now considers linear programs for which an optimal solution exists; discussion of linear programs for which no

optimal solution exists is provided in the section titled “Special Cases” that follows.

In linear programming, the enumeration method refers to the determination of an optimal solution (if one exists) through the evaluation of the solution at each point of intersection of perimeters of constraints. Note that this set of points includes all extreme points and so, by the fundamental theorem of linear programming, will include an optimal solution (if one exists). This means that for the powdered fruit drink mix problem, an optimal solution may be found at one or more of points A, B, C, D, E, and F identified in Fig. 4. Solving for the decision variables X_1 and X_2 at each of these points yields the solutions provided in Table 1.

Of these solutions, points C and F have the largest objective function values (6400 and 6000, respectively); however, neither of these solutions is feasible (point C violates the powdered lemon juice constraint and point F violates the sugar constraint). Of the remaining points (each of which is feasible and so is an extreme point), point D provides the largest objective function value. Thus, through the enumeration method one would again identify the optimal solution as

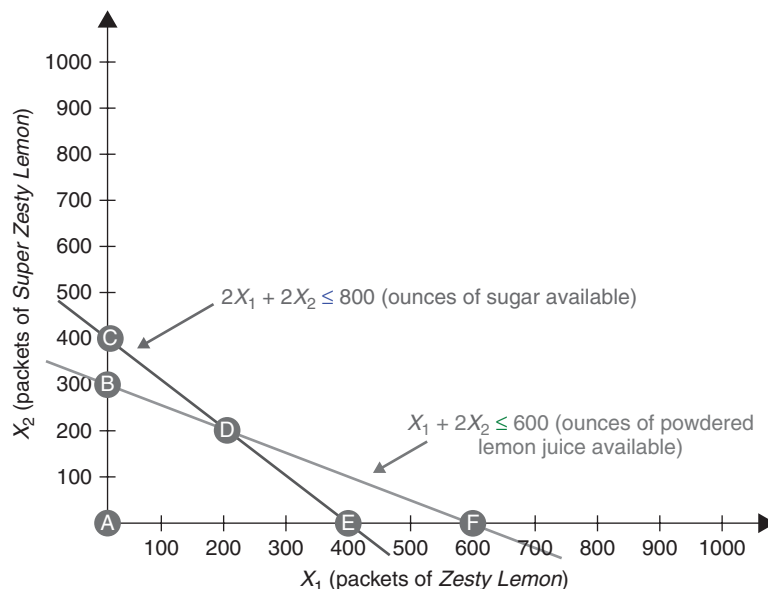


Figure 4. Points to be considered in solving the powdered drink mix problem by enumeration.

Table 1. Solutions at the Intersections of Perimeters of Constraints

Extreme Point	X_1 (packets of <i>Zesty Lemon</i>)	X_2 (packets of <i>Super Zesty Lemon</i>)	Objective Function
A	0	0	0
B	0	300	4800
C	0	400	6400
D	200	200	5200
E	400	0	4000
F	600	0	6000

point D, where $X_1 = 200, X_2 = 200$, and the objective function is 5200.

A linear program with n decision variables and $m + n$ constraints (which includes n nonnegativity constraints) will require evaluation of as many as $\binom{m+n}{n}$ solutions when using the enumeration method. This implies that the enumeration method requires evaluation of (i) potentially several infeasible solutions and (ii) all extreme points—these are the inherent weaknesses of the enumeration method. Enumeration of a relatively small linear program with 100 decision variables and 150 constraints (including nonnegativity) may require the evaluation of as many as $2.01287\text{E} + 40$ potential solutions. In order for an algorithm to be practical, it must be able to quickly find an optimal solution for linear programs with thousands or even millions of constraints and decision variables, and this can only be accomplished if the algorithm avoids these inherent weaknesses of the enumeration method.

SOLVING LINEAR PROGRAMMING PROBLEMS WITH THE SIMPLEX ALGORITHM

Real applications of linear programming frequently feature far more decision variables and constraints than can be solved for efficiently through graphing or the enumeration method. For such problems, the *simplex algorithm* is often used to find an optimal solution. This algorithm was created by George Dantzig, who is also credited with the first use of the term *linear programming* in reference to problems in logistics programs that he worked on for the Allied Forces during

World War II [4]. The algorithm relies on a corollary of the fundamental theorem of linear programming and successfully addresses the inherent weaknesses of the enumeration method discussed in the proceeding discussion; the simplex algorithm generally (i) will require the evaluation of only a subset of all extreme points and (ii) will not require the evaluation of infeasible solutions.

One can think of the simplex algorithm as a systematic approach to solving a series of simultaneous equations for a set of variables with unknown values in a manner that best satisfies the objective function. For example, again consider the powdered fruit drink mix problem; for the first constraint (sugar), it is obvious that values of the decision variables X_1 and X_2 must be chosen so that the resulting value on the left side of the inequality (ounces of sugar used to produce *Zesty Lemon* and *Super Zesty Lemon*) does not exceed the value on the right side of the inequality (ounces of sugar available). This inequality can be converted into an equality by adding a nonnegative variable (designated S_1) to the left side of this constraint, where S_1 is defined to be equal to the number of ounces of unused sugar, that is,

$$S_1 = 800 - (2X_1 + 2X_2).$$

The resulting equality is

$$2X_1 + 2X_2 + S_1 = 800,$$

and S_1 is referred to as the *slack variable* associated with the first constraint. The second constraint (powdered lemon juice) can be converted to an equality in a similar manner

$$X_1 + 2X_2 + S_2 = 600,$$

where

$$S_2 = 600 - (X_1 + 2X_2).$$

Here, S_2 is referred to as the *slack variable* associated with the second constraint and represents the ounces of unused powdered lemon juice sugar that remain if one produces X_1 packets of *Zesty Lemon* and X_2 packets of *Super Zesty Lemon*. Since neither S_1 nor S_2 contributes to the objective function, these variables both have objective function coefficients of 0. The resulting revised formulation, called the *standard form* of the original formulation, is

$$\begin{aligned} \text{Maximize } & 10X_1 + 16X_2 + 0S_1 + 0S_2 \\ & 2X_1 + 2X_2 + S_1 = 800 \\ & X_1 + 2X_2 + S_2 = 600 \\ & X_1 \geq 0 \\ & X_2 \geq 0 \\ & S_1 \geq 0 \\ & S_2 \geq 0. \end{aligned}$$

The result is a system of $m = 2$ equalities and $n + m = 4$ variables

$$\begin{aligned} 2X_1 + 2X_2 + S_1 &= 800, \\ X_1 + 2X_2 + S_2 &= 600, \end{aligned}$$

for which the nonnegativity constraints provide the conditions under which one can ignore a solution to this set of equations due to infeasibility. Note that a constraint for which the left hand side must be at least as large as the right hand side can be converted to an equality in a similar manner by subtracting a surplus variable (defined as the amount by which the left hand side exceeds the right hand side) from the left hand side of the constraint.

With m equalities one can find unique values for no more than m variables; thus solving a system of equations such as this will generally involve setting n of the decision variables, slack variables, and surplus variables equal to 0 (these are referred to as *nonbasic variables*) and solving for the remaining m variables (these are referred

to as *basic variables*). Thus, at each iteration for the powdered drink mix problem, the simplex algorithm systematically solves for two of the variables in this set of simultaneous equations while setting the remaining two variables equal to zero, and then evaluates the resulting solution with respect to the objective function.

One begins by identifying a feasible solution at an extreme point and evaluating the objective function at that point; this is called the *current solution*. Movement of the current solution toward adjacent extreme points is then considered. Note here that moving from an extreme point that represents the current solution to an adjacent extreme point is equivalent to allowing one of the nonbasic variables (i.e., variables with values of zero in the current solution) to become basic (i.e., potentially take on a positive value). If none of these adjacent extreme points produces a superior objective function value, then the current solution is an optimal solution. On the other hand, if any of these adjacent extreme points produces an objective function value that is superior to the value produced by the current solution, the adjacent extreme point that produces the best marginal increase per unit of the new basic variable becomes the current solution. These steps constitute one *iteration* of the algorithm, and the process continues until no adjacent extreme point produces a value of the objective function that is superior to the current solution.

If one initializes the simplex algorithm on the powdered drink mix problem by arbitrarily selecting X_1 and X_2 to be nonbasic variables (i.e., setting $X_1 = 0$ and $X_2 = 0$), the slack variables become basic variables and take on the values $S_1 = 800$ and $S_2 = 600$; this yields an objective function value of 0. This is reasonable—if one produces no packets of *Zesty Lemon* (X_1) or *Super Zesty Lemon* (X_2), 800 ounces of sugar (S_1) and 600 ounces of powdered lemon juice (S_2) will remain and no profit will be earned. This solution occurs at the origin, which is extreme point A in Fig. 4, and is the initial current solution.

Now consider swapping the roles of one basic variable and one nonbasic variable in this solution. Which of the basic variables (S_1

or S_2) should be set to 0 (i.e., made nonbasic), and which of the nonbasic variables (X_1 or X_2) should be allowed to potentially take a nonzero value (i.e., become basic)? In other words, does increasing the value of X_1 or X_2 by one unit increase the objective function value, and if so, for which of these variables will a one unit increase result in the greatest marginal improvement in the objective function value?

A one unit increase in X_1 will contribute 10¢ directly to the objective function while decreasing the slack (leftover) ounces of sugar by 2 ounces and the slack (leftover) ounces of powdered lemon juice by 1 ounce. Since neither the slack (leftover) ounces of sugar nor the slack (leftover) ounces of powdered lemon juice contributes to the objective function, there is no cost associated with using either resource, and so the marginal contribution to the objective function that is made by increasing X_1 by one unit at this iteration is 10¢.

Similarly, a one unit increase in X_2 will contribute 16¢ directly to the objective function while decreasing the slack (leftover) ounces of sugar by 2 ounces and the slack (leftover) ounces of powdered lemon juice by 2 ounces. Again, since neither the slack (leftover) ounces of sugar nor the slack (leftover) ounces of powdered lemon juice contributes to the objective function, there is no cost associated with using either resource, and so the marginal contribution to the objective function that is made by increasing X_2 by one unit at this iteration is 16¢.

Since the marginal increase in the objective function associated with a one unit increase in X_2 is 16¢ and the marginal increase in the objective function associated with a one unit increase in X_1 is 10¢ at this iteration, X_2 is allowed to potentially take some positive value (i.e., become basic). Note that these marginal values are often referred to as *reduced costs*. The algorithm continues to increase the value of X_2 until some variable that is basic in the current solution (S_1 or S_2) becomes nonbasic, which at this iteration means that the supply of either sugar or powdered lemon juice is exhausted. This occurs at $X_2 = 300$, which results in $S_1 = 200$

and $S_2 = 0$ (the supply of powdered lemon juice that remained in the previous iteration is exhausted) and yields an objective function value of 4800. If one produces 0 packets of *Zesty Lemon* (X_1) and 300 packets of *Super Zesty Lemon* (X_2), 200 ounces of sugar (S_1), and 0 ounces of powdered lemon juice (S_2) will remain, and 4800¢ profit will be earned; this solution occurs at extreme point B in Fig. 4 (which is adjacent to the extreme point A).

Note that if one had allowed X_1 to become basic (potentially take some positive value) and continued to set $X_2 = 0$ in the first iteration, the value of X_1 would have become 400 and as a result $S_1 = 0$ and $S_2 = 200$, which would have yielded an objective function value of 4000. If one produces 400 packets of *Zesty Lemon* (X_1) and 0 packets of *Super Zesty Lemon* (X_2), 0 ounces of sugar (S_1) and 200 ounces of powdered lemon juice (S_2) will remain and 4000¢ profit will be earned; this solution occurs at extreme point D in Fig. 4 (which is the other extreme point that is adjacent to the extreme point A).

All extreme points that are adjacent to the current solution have now been systematically examined, and both result in objective function values that are superior to the objective function value produced by the current solution. The algorithm therefore selects extreme point B (the extreme point at which X_2 , the nonbasic variable that makes the greatest per unit improvement to the objective function of all nonbasic variables at this iteration, becomes basic) as the new current solution. $X_1 = 0, X_2 = 300, S_1 = 200, S_2 = 0$, and profit = 4800. Evaluation of all extreme points that are adjacent to a current solution constitutes one iteration of the simplex algorithm, and the algorithm will continue to iterate until no improvement to the objective function value that is associated with the current solution can be found; at that point an optimal solution has been identified.

Now, again consider allowing one of the nonbasic variables (i.e., a variable with a value of 0 in this new current solution) to become basic (take a positive value). A one-unit increase in X_1 will directly contribute

10¢ to the objective function. However, note that production of one unit of X_1 requires 2 ounces of sugar and 1 ounce of powdered lemon juice, and the supply of powdered lemon juice is exhausted by the current solution. The only way to increase the number of units of X_1 by one is to use some powdered lemon juice that is currently devoted to production of X_2 . Since it takes 2 ounces of powdered lemon juice to produce one unit of X_2 and only 1 ounce of powdered lemon juice to produce one unit of X_1 , one must sacrifice $\frac{1}{2}$ unit of X_2 for every unit of X_1 added at this iteration. Since one unit of X_2 makes a direct contribution of 16¢ to the objective function, the loss of $\frac{1}{2}$ unit of X_2 decreases the objective function by 8¢. Because the slack (leftover) sugar at this iteration is positive, it can be used in the production of X_1 at no cost. Thus, the marginal increase in the objective function at this iteration that corresponds to a one unit increase in X_1 is $10¢ - 8¢ = 2¢$. The algorithm will continue to add units of X_1 until one of the basic variables for this iteration (X_2 or S_2) becomes nonbasic (i.e., becomes 0). This occurs at $X_1 = 200$ and $X_2 = 200$, with $S_1 = 0$ (the slack sugar that remained in the previous iteration is now exhausted) and $S_2 = 0$, which yields an objective function value of 5200¢. This becomes the current solution, and the second iteration is complete.

At this point, one could consider allowing one of the slack variables (each of which equals 0 in this new current solution) to take a positive value. However, a one unit increase in either slack variable S_1 or S_2 will not improve the objective function, so the algorithm stops at this iteration. The optimal solution to this problem is $X_1 = 200, X_2 = 200, S_1 = 0$, and $S_2 = 0$, which yields an objective function value of 5200¢.

In reviewing these iterations, it is apparent that in each iteration the simplex algorithm started at an extreme point (the current solution at that point), considered all adjacent extreme points, and moved to the extreme feasible point that yielded the greatest marginal contribution to the objective function per unit increase of the associated new basic variable.

Because the simplex algorithm considers only a subset of extreme points and completely ignores infeasible solutions, it is referred to as an *intelligent enumeration algorithm*. It is for these reasons that the simplex algorithm is remarkably efficient and capable of quickly solving very large problems to optimality in a relatively short period of time [5,6]. For a more detailed discussion, see **The Simplex Method and Its Complexity** and **Simplex-Based LP Solvers**.

SPECIAL CASES

Note that it is possible for multiple extreme points to be optimal; under these circumstances, these multiple extreme points and all points that lie on the perimeter of the constraint that connects them have the same value of the objective function and are referred to as the *alternate optimal solutions* (or *alternate optima*). This phenomenon occurs when the objective function and the perimeter of a constraint that combines with other constraints to form an optimal extreme point are parallel (for example, if the objective function in the powdered fruit drink problem was Maximize $4X_1 + 8X_2$ as shown in Fig. 5).

For this problem, the objective function and the perimeter of the powdered lemon juice constraint have the same slope, and the powdered lemon juice constraint is one of the constraints that intersect to form an optimal extreme point. Thus, extreme points B and D and all points that lie on the line segment that connects these points are alternate optima.

It is also possible for a linear programming problem to have no optimal solution. This can happen in two distinct ways. Consider the following formulation of a linear programming problem with two decision variables:

$$\begin{aligned} &\text{Maximize } 10X_1 + 16X_2 \\ &\text{subject to } 2X_1 + 2X_2 \geq 800 \\ &\quad X_1 + 2X_2 \geq 600 \\ &\quad X_1 \geq 0 \\ &\quad X_2 \geq 0. \end{aligned}$$

If one graphs the feasible region for this problem and then plots and parallel shifts an

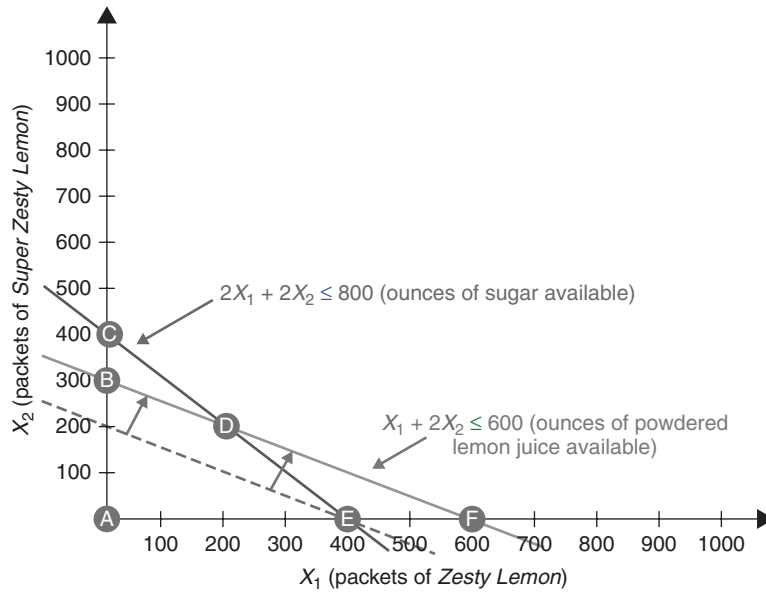


Figure 5. An example of a linear program with alternate optimal solutions.

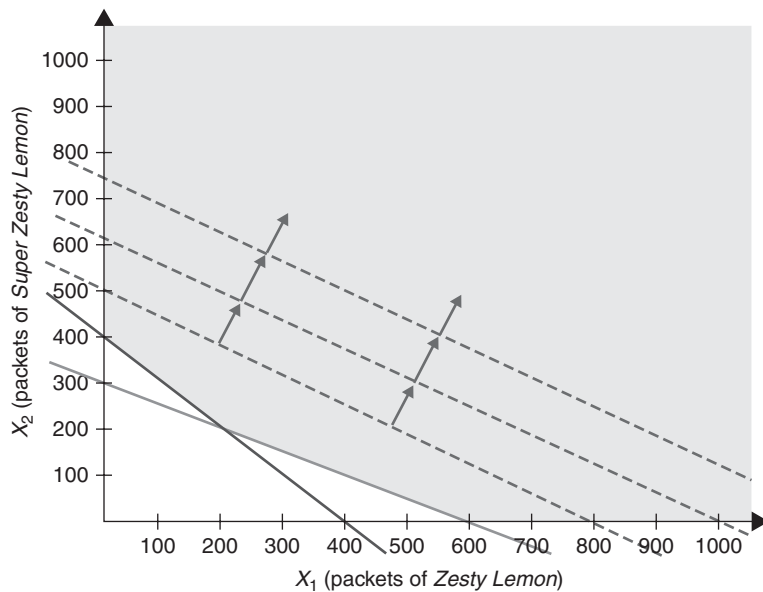


Figure 6. An example of an unbounded linear program.

isoprofit line in its direction of improvement, one never reaches the perimeter of the feasible region. The feasible region for problems such as this is said to be *unbounded* (Fig. 6).

A linear programming problem also has no optimal solution if there is no feasible region;

this happens when there is no intersection of the areas that are feasible for the individual constraints and is called *infeasibility*. Consider the following formulation of a linear programming problem with two decision variables:

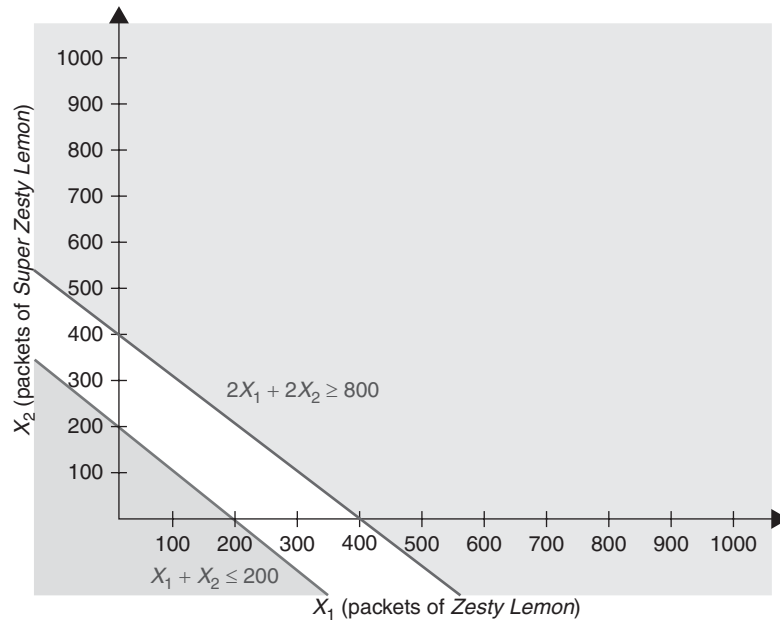


Figure 7. An example of an infeasible linear program.

$$\begin{aligned}
 &\text{Maximize } 10X_1 + 16X_2 \\
 &\text{subject to } 2X_1 + 2X_2 \geq 800 \\
 &\quad X_1 + X_2 \leq 200 \\
 &\quad X_1 \geq 0 \\
 &\quad X_2 \geq 0.
 \end{aligned}$$

A graph of the feasible spaces that correspond to the first two constraints for this problem is provided in Fig. 7. The feasible regions for these two constraints obviously have no intersection (i.e., no solution lies in an area that satisfies both constraints), and so there is no feasible region from which to select a potential solution.

OTHER CONSIDERATIONS

Note that several methods have been developed for assessing the changes in the solution that result from changes in the values of parameters. These methods are collectively referred to as *sensitivity analysis*, *postoptimality analysis*, or more colloquially as *what-if analysis* (see ***Sensitivity Analysis in Linear Programming***).

It is also important to note that this entire discussion pertains to problems for which the objective is to minimize some function of the decision variables. In fact, multiplication of a maximization objective function by -1 creates an equivalent minimization objective function (the converse is also true).

Duality

Associated with every linear programming problem is a unique corresponding formulation that is referred to as the problem's *dual* formulation. In this context, the original problem is referred to as the *primal* formulation.

Conventionally, the primal constraints are indexed by $i = 1, \dots, m$ and the primal decision variables by $j = 1, \dots, n$. Using this indexing scheme, we can denote the objective function coefficient for the j th primal decision variable x_j as c_j , the coefficient associated with the j th primal decision variable in the i th primal constraint as a_{ij} , and the right hand of the i th primal constraint as b_i . If the objective of the primal formulation is maximization of the objective function, all primal decision variables are on the left hand sides and all constants are on the right hand

sides for each constraint, and each primal constraint is expressed such that the left hand side is less than or equal to the right hand side, the generic formulation using this notation is

$$\begin{aligned} &\text{Maximize} && \sum_{j=1}^n c_j x_j \\ &\text{subject to} && \sum_{j=1}^n a_{ij} x_j \leq b_i \quad i = 1, 2, \dots, m \\ &&& x_j \geq 0 \quad j = 1, 2, \dots, n. \end{aligned}$$

This is referred to as the *standard form* of the primal problem. Note that any linear program can be put into this form through basic algebra.

If the primal formulation is in the standard form, then the goal of the associated dual formulation is minimization of its objective function. The associated dual formulation will have a decision variable corresponding to each primal constraint and a constraint corresponding to each primal decision variable. Thus the dual constraints are indexed by $j = 1, \dots, n$ and the dual decision variables by $i = 1, \dots, m$. The dual decision variables may be denoted by y_i .

The objective function coefficient for the i th dual decision variable y_i is the right hand of the i th primal constraint b_i , the coefficient associated with the i th dual decision variable in the j th dual constraint is a_{ji} , and the right hand of the j th dual constraint is the objective coefficient of the j th primal variable c_j . Finally, if primal problem is in standard form (i.e., the left hand side of each primal constraint is restricted to be less than or equal to its corresponding right hand side), then the left hand side of each constraint in the corresponding dual problem is restricted to be greater than or equal to its corresponding right hand side. Given these relationships and the primal formulation in standard form, the corresponding generic dual formulation is

$$\text{Minimize} \quad \sum_{i=1}^m b_i y_i$$

$$\begin{aligned} &\text{subject to} && \sum_{i=1}^m a_{ji} y_i \geq c_j \quad j = 1, 2, \dots, n \\ &&& y_i \geq 0 \quad i = 1, 2, \dots, m. \end{aligned}$$

For example, the dual formulation associated with the powdered fruit drink mix problem is

$$\begin{aligned} &\text{Minimize} && 800Y_1 + 600Y_2 \\ &\text{subject to} && 2Y_1 + Y_2 \geq 10 \\ &&& 2Y_1 + 2Y_2 \geq 16 \\ &&& Y_1 \geq 0 \\ &&& Y_2 \geq 0, \end{aligned}$$

with an optimal solution of $Y_1 = 2$ and $Y_2 = 6$, and an associated objective function value of 4800. Note that the optimal objective function values associated with the primal and dual formulations for this problem are equal; this will always be so, and this result is referred to as the *strong duality theorem* (which is discussed shortly).

While the dual formulation can yield additional insightful information and can sometimes be solved to optimality in less time than the corresponding primal formulation, the greatest value of the dual formulation lies in three results on the relationship between a primal formulation and its corresponding dual formulation.

The first of these results is commonly referred to as the *weak duality theorem*. This theorem states that for a primal formulation in standard form, the objective function value of the associated dual problem at any of its feasible solutions will be greater than or equal to the objective function value of the primal problem at any of its feasible solutions. Thus, if a primal problem is in standard form, given any feasible primal solution $x_1, x_2, \dots, x_n$ and any feasible solution $y_1, y_2, \dots, y_m$ for the associated dual,

$$\sum_{j=1}^n c_j x_j \leq \sum_{i=1}^m b_i y_i.$$

This relationship allows for the identification of an upper bound for the primal problem objective value through the identification of

a feasible solution of the corresponding dual problem. Similarly, if the goal of the primal formulation is minimization of its objective function, the objective function value of any feasible solution of the associated dual problem will be less than or equal to the objective function value of any feasible solution of the primal problem (and so constitutes a lower bound for the primal problem objective value).

The second of these three results is referred to as the *complementary slackness theorem*. This theorem states that for any feasible primal solution $x_1, x_2, \dots, x_n$ and any feasible solution $y_1, y_2, \dots, y_m$ for the associated dual

$$\left(b_i - \sum_{j=1}^n a_{ij}x_j\right)y_i = 0 \quad i = 1, 2, \dots, m,$$

which is referred to as *primal complementary slackness*, or conversely

$$\left(c_j - \sum_{i=1}^m a_{ji}y_i\right)x_j = 0 \quad j = 1, 2, \dots, n,$$

which is referred to as *dual complementary slackness*. Either of these sets of conditions is also commonly referred to as the *optimality conditions* as they constitute necessary and sufficient conditions for optimality of the primal or dual problem.

Complementary slackness, in combination with requirements for feasibility of the primal and dual solutions, comprise what are known as the *Karush–Kuhn–Tucker* (or *KKT*) *optimality conditions* [7].

The third of these results is commonly referred to as the *strong duality theorem*. This theorem, which is often cited as a lemma of complementary slackness, states that if the primal problem has an optimal solution $x_1^*, x_2^*, \dots, x_n^*$, then the associated dual problem also has an optimal solution $y_1^*, y_2^*, \dots, y_m^*$ such that

$$\sum_{j=1}^n c_j x_j^* = \sum_{i=1}^m b_i y_i^*,$$

that is, if the primal formulation has an optimal solution, the associated dual problem

also has an optimal solution with an objective function value equal to the optimal primal objective value. This relationship allows for the solution of a primal problem through the formulation and solution of the corresponding dual problem.

An important corollary of the strong duality theorem states that if a primal linear program is unbounded, then the associated dual linear program is infeasible (and the converse is also true).

For proof of these theorems and the corollary associated with the strong duality theorem see [3].

Interior Point Methods

While the solution algorithms discussed in this article search the perimeter of the feasible region (and sometimes beyond) for an optimal solution, other solution algorithms search for an optimal solution by moving directly through the interior of the feasible region. These algorithms are referred to as *interior point methods* [8,9]. While early work in this area by Khachian [10] on what is referred to as the *ellipsoid method* ultimately showed initial promise, it was Karmarkar's efforts [11] that first demonstrated the true potential of interior point algorithms for solving large linear programs efficiently. Modern solution algorithms that are based on interior point methods often have better worst-case running times than the simplex algorithm, and are commonly used in practice to solve large linear programs [12,13]. See **Interior-Point Linear Programming Solvers** for a more detailed discussion of this concept.

Finally, it is important to reiterate that (i) large linear programming problems can generally be solved to optimality rapidly and (ii) many problems faced by small businesses and scientists across disciplines can be solved using linear programming.

REFERENCES

1. Wagner M, Meller J, Elber R. Large-scale linear programming techniques for the design of protein folding potentials. *Math Program* 2004;101(2):301–318.

2. Dantzig GB. Linear programming and extensions. Princeton, NJ: Princeton University Press; 1963.
3. Martin K. Large scale linear and integer optimization. Norwell, MA: Kluwer; 1999.
4. Cottle R, Johnson E, Wets R, George B, Dantzig. Notices of the AMS 2007;54(3): 344–362.
5. Forrest JJH, Tomlin JA. Implementing the simplex method for the optimization subroutine library. IBM Syst J 1992;31:11–25.
6. Dantzig GB, Thapa MN. Linear Programming 1: Introduction. New York, NY: Springer; 1997.
7. Bazaraa M, Jarvis J, Sherali H. Linear programming and network flows. 4th ed. New York, NY: John Wiley and Sons; 2009.
8. Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. SIAM Rev 2002;44:525–597.
9. Potra F, Wright S. Interior-point methods. J Comput Appl Math 2000;124:281–302.
10. Khachian LG. A polynomial algorithm in linear programming. Dokl Akad Nauk SSSR 1979;244:1093–1096 (English translation in Sov Math Doklady 1979;20:191–194.).
11. Karmarkar N. A new polynomial-time algorithm for linear programming. Combinatorica 1984;4:373–395.
12. Zhang Y, Tapia R, Dennis J. On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms. SIAM J Optim 1992; 2(2): 304–324.
13. Zhang Y, Tapia R, Porta F. On the superlinear convergence of interior point algorithms for a general class of problems. SIAM J Optim 1993;3(2):413–422.

AN INTRODUCTION TO PLATELET INVENTORY AND ORDERING PROBLEMS

JOHN T. BLAKE
Department of Industrial Engineering,
Dalhousie University,
Halifax, Nova Scotia, Canada

BACKGROUND

Platelets are blood cells that initiate the hemostatic plug that causes blood clot formation. Patients receiving intense chemotherapy or suffering massive bleeding complications require platelet transfusions for the prevention of a potentially fatal hemorrhage. A stable, readily available inventory of platelets is required for the safe and effective delivery of health care [1].

Platelets are typically produced from whole blood through a process that separates blood into three main products: red cells, plasma, and platelets. Platelets can also be collected directly through apheresis, a donation process that removes platelets and some plasma from a donor's blood, and returns the remaining red cells and plasma back to the donor's system.

Since platelets must be kept warm to remain viable, they are subject to bacterial contamination and thus have a shelf life of 5–7 days. The exact shelf life of platelets depends on the method used to collect platelets from donors and the use of systems to detect bacterial contamination. The short shelf life poses a number of inventory management problems. Since transmissible disease testing, component processing, nucleic acid testing, and blood bank testing require 12–48 h to complete, platelets may be available for transfusion for as little as 78 h before they must be discarded. By comparison, red blood cells have a shelf life of 42 days at present; though there is some

indication that the older red cells may be less effective than the newer cells [2].

PLATELET INVENTORY AND ORDERING PROBLEM

Managing platelet inventory is a difficult problem because of the necessity of minimizing both stock-outs and outdates. Since they are vital for medical care, an adequate supply of platelets must be available on demand. Since supply and demand are stochastic, hospitals and blood product suppliers must maintain stock to minimize stock-outs. However, because platelets are perishable, excess inventory leads to outdates. Both stock-outs and outdates are considered to be expensive. Stock-outs are expensive in an absolute sense because patients may suffer ill-health (or worse) without a timely transfusion of platelets. In a practical sense, a shortage of platelets requires a hospital or blood supplier to expedite platelets from another source (another distribution center or hospital), often with a large resultant cost penalty. Outdates are also considered to be expensive. There is, of course, the tangible cost of the materials and labor required by the blood supplier to collect, produce, test, and distribute the unit, as well as the cost to the hospital of receiving and storing a unit that ultimately was never needed. There is also an intangible cost to the donor of having lost personal time and suffering discomfort to provide a unit that ends up being incinerated. Blood agencies are sensitive to the intangible costs to donors, since they feel that individuals may be less inclined to donate if they think that their gift will just be scrapped. Because donor participation rates in western countries are typically less than 5% of the eligible adult population [3], large outdatedness rates are believed to be a threat to the underlying stability of the blood supply chain and must be avoided.

DEFINITION OF PRODUCER'S AND CONSUMER'S PROBLEM

There are two subproblems in platelet inventory and ordering: the producer's problem and the consumer's problem.

Producer's Problem

The producer's problem follows a classical inventory management structure. At the beginning of each decision epoch (typically a day), the inventory state, comprising the total number of units available and their age distribution, is observed. The decision maker places an order for platelets to be collected during the day, before demand is observed. Collections are usually assumed to be deterministic. A nonzero cost for placing an order is assumed ($c_o = f_o + v_o o$). Orders are assumed to take a minimum of one period to be filled. Thus, orders placed today are not available for distribution until tomorrow.

Demand is observed over the course of the day. Demand is assumed to be stochastic and unknown by the producer prior to its realization, but to follow a known distribution. The producer fills the demand from the available stock, starting with the oldest stock on hand (i.e., a first-in-first-out or FIFO policy). If orders cannot be filled from the available stock, a shortage is registered and the unmet demand is assumed to have been lost. A penalty is assumed for any shortage. It is generally assumed that any instance of a shortage is to be avoided and thus, a step-type penalty cost with a large fixed component and a small per unit cost is often assumed ($c_S = f_S + v_S S$).

At the end of each day, all stock remaining in the inventory is "aged" by 1 day. For example, the stock that will expire in 2 days becomes the stock that will be outdated in 1 day. Any remaining stock with 1 day to outdate becomes stock with 0 days to outdate and is thus "outdated" and removed from the inventory. A nonzero outdate cost is assumed, generally with both a fixed and variable component ($c_W = f_W + v_W W$). It is generally assumed that $c_S \gg c_W$.

Finally, the stock that was ordered yesterday becomes available for use and enters

inventory as m day to outdatedness stock. The cycle then repeats.

It should be noted that stock on hand is typically represented as a vector $\mathbf{x} = (x_1, x_2, \dots, x_m)$, where x_i is the inventory with i days to outdate. Thus, the stock vector is typically "backwards" with x_1 representing the oldest stock on hand (i.e., the stock with 1 day to outdate) and x_m representing the newest stock (i.e., the stock with m days to outdate).

Consumer's Problem

The consumer's problem is similar to the producer's problem with two important differences. As in the producer's problem, the inventory state, comprising the number of units available and their age distribution, is observed at the beginning of the decision epoch. A decision is then made about the number of units to order (if any). However, unlike the producer's problem, a consumer's order may be assumed to arrive before demand is observed. This assumption is reasonable for most medium to large hospitals in which it might be expected that an order could be received within a few hours from a nearby distribution center. In instances where deliveries cannot be made within a few hours, such as rural or isolated facilities, orders must be made 1 or more days in advance and the problem more closely resembles that of the producer's problem.

Unlike the producer's problem, the units arriving to the consumer may or may not be of a consistent age. For instance, the consumer may receive some units with 5 days to outdate, some with 4 days to outdate, and so on. Since distribution and testing often take 1 or more days to complete, it is rare for consumers to receive "new" products; more commonly, consumers receive products that have lost 1 or more days of shelf life prior to receipt. Furthermore, because of the variations in collection and production schedules at the producer, (day-to-day fluctuations in collections as well as a systematic fluctuation occurring because demand is experienced 7 days a week while platelet collection and testing may occur only Monday to Friday), the age of arriving units varies from day to day. For instance,

platelets collected on Fridays, are typically not available for distribution until Monday and thus, consumers receive platelets that have lost at least 2 days of shelf life.

LITERATURE REVIEW

There is an extensive literature on perishable inventory as related to blood supply. However, much of the literature deals with red blood cells that have a much longer shelf life than platelets.

Veinott [4] describes a periodic review policy under the assumption of stationary demand. Results show that optimal ordering policies for perishable inventory correspond closely to the nonperishable case. Optimal order quantities are set as

$$\min(Q^*, \lambda m) \quad (1)$$

where Q^* is the economic order quantity (EOQ), λ the demand rate, and m the lifetime. Under this set of assumptions, no units expire.

Pierskalla and Roach [5] use a dynamic programming (DP) formulation to show that FIFO policies are optimal in perishable inventory problems.

Fries [6] describes a DP approach to perishable inventory policies under the assumption of no backordering. Ordering policies in this case depend on the stock on hand, its age distribution, and the length of the planning horizon. Fries considers three specific cases. In instances where the shelf life (m) of the product is one period, the problem reduces to that of the well-known newspaper vendor problem. When the shelf life is two or more periods and the planning horizon (n) is 1 day, the optimal policy is an (s, S) type policy where a quantity of product is ordered to bring the inventory on hand up to a critical value x^* . When the planning horizon is $1 < n < m$ the ordering policy is an (s, S) policy, where an order is placed to bring the total usable inventory on hand in the next period up to y^* , where y^* depends on the age of the stock on hand. When the planning horizon is greater than the life span of the product, the optimal ordering policy is based on the

expected number of units to be consumed in the next period and the age distribution of the remaining stock.

Nahmias [7] adopts a similar approach to Fries, but notes the extreme difficulty of computing optimal policies when m is greater than 2 days. In place of exact solutions, Nahmias argues for the use of heuristic solutions [8,9] and [10] provides an excellent summary of the relevant perishable inventory literature.

Cohen *et al.* [11], acknowledging the problem of *a priori* shortage rates, suggest a simple decision rule that obviates the need for managers to set explicit rates in the case of red blood cell supply. Using regression techniques in combination with simulation methods, they develop a target inventory level S^* that depends on daily demand, average transfusion-to-cross match ratio and cross match release period. Brodheim *et al.* [12] adopt a similar approach, but suggest an equation for setting target inventory that depends on the mean daily demand in conjunction with an explicit management decision regarding acceptable shortage rates. Brodheim and Prastacos [13] describe a model for setting hospital inventory policies under the assumption of a fixed delivery schedule from the regional blood bank.

Kendall and Lee [14] employ a goal-programming approach to develop policies for red cell rotation in blood provision networks. Their model is novel in that it does not take a cost minimization approach, but rather focuses on obtaining a set of objectives that includes minimizing shortages and avoiding outdatedness, both on the local and regional levels. Kendall and Lee note that blood networks vary greatly in their composition and usage patterns and thus suggest that what constitutes a good inventory policy depends on local demand patterns, transportation links, and donor availability and participation rates. Results from tests on two blood networks suggest that outdates are minimized when stock is able to freely rotate between hospitals. Stock age and outdate rates were shown to improve with greater rotation without any increase in shortages. However, Kendall and Lee do not explicitly consider transportation costs

and thus note that hospital mix, geography, and population affect the generalizability of their results.

Freidman *et al.* [15] describe the use of simulation to set inventory levels for red blood cells under the assumption of a 35-day shelf life. Describing blood management policies from a clinician's standpoint, they argue against the setting of *a priori* shortage rates. Instead, they suggest an empirical approach to inventory policy in which safety stocks are gradually reduced.

Hesse *et al.* [16] describe an application of inventory management techniques to platelets, in a system in which a centralized blood bank supplies 35 client hospitals. Hesse *et al.* adopt a periodic review model and develop (s, S, t) policies for each of the client institutions, using a simulation model as a test platform. Noting the complexity of a DP approach, the authors aggregate institutions into risk pools and develop, via an enumerative process, an (s, S, t) policy for each pool.

Sirelson and Brodheim [17] use simulation to test platelet ordering policies for a blood bank, based on the average demand and a fixed base stock level. They show that a base stock level based on a mean demand plus a multiple of standard deviation, can be used to reduce current outdatedness and shortage rates. They also show that, on a regional level, low shortage and outdatedness rates can be readily obtained; within individual hospitals, low outdatedness and shortage rates are more difficult to achieve. Katz *et al.* [18] report similar results.

Blake *et al.* [1] present a DP formulation for solving an instance of the platelet inventory problem for an environment in which there is a single producer of platelets and a single consumer. They implement a DP model for both the producer and the consumer, to identify optimal local ordering policies. These policies are then tested via a simulation model to identify good practical policies that minimize the overall wastage and outdatedness rates. Blake *et al.* note the potential for developing a DP approach for developing optimal joint producer/consumer policies, but found that the so-called *curse of dimensionality* limited the scale of their model, despite efforts to minimize the state space by

aggregating units and demand into standard adult doses.

Katsaliaki and Brailsford [19] describe the use of a large-scale simulation model to evaluate the function of a blood supply chain in southern England. Their model includes multiple products, including platelets. A number of operational policies are tested via simulation. For the system under study (in the simplest case a single producer and a single consumer), it is shown that the amount of inventory stored can be reduced if improved ordering and cross-matching policies are implemented. The single consumer–single supplier model is extended to cover a longer run period and a larger number of consumers through a distributed simulation environment [20].

van Dijk *et al.* [21] suggest a multistep procedure for identifying a heuristic solution to the platelet ordering problem. They formulate the solution to the platelet ordering problem as a DP problem using standard conventions. However, to make the problem tractable, they scale the problem by a factor of 4, since platelet units produced from red cells are typically given to adults in doses of four. van Dijk *et al.* then solve exactly the downsized problem, using a scaled demand function. The solution to the problem is recorded for all instances of day, time in the planning horizon, and amount and age distribution of stock. A simulation model then records 1M weeks of ordering behavior under the assumption of the downsized problem. For each day, the amount of stock on hand and the order size suggested by the DP model are recorded. The resulting order sizes tend to follow a classical “order-up-to” policy seen in the nonperishable inventory problem. Based on this observation, van Dijk *et al.* suggest that solutions to the platelet ordering problem could ignore the age distribution of stock and condition orders only on the total stock available. The order-up-to rule selected for each day is the solution most frequently seen in the simulated history. The solution is then rescaled back to the original problem size and the full-size problem is again simulated to verify that the solution remains feasible for the actual problem. Blake [22], however, argues that age cannot always be ignored.

Erhun *et al.* [23] take a systemic view of the platelet supply chain to recommend a series of practical policy improvements for a university-based blood bank. They note the importance of agility to respond to sudden changes in demand level and suggest a collection and testing regime that extends over 7 days of the week and that is explicitly tied to expected demand. They also suggest shortening the platelet rotation horizon, such that small hospitals hold platelets for only a single day before rotating them to larger demand institutions, and improving issuing policies to ensure strict adherence to FIFO policies.

DISCUSSION

For the deterministic inventory problem, it is simple to show, using the EOQ model, how a fixed order point and quantity can be determined if lead time is constant and known. If lead time or demand varies, safety stock is necessary to ensure that enough stock is on hand to guarantee a given level of service that can be realized between order placement and order receipt. In the case where demand varies, it can be shown that a “two-bin” inventory system is optimal. A two-bin system assumes that inventory is reviewed, either continuously or at fixed intervals. If the inventory is below a certain level (called a *trigger level*), then an order is placed. The order size is not fixed, but depends on the actual amount of stock on hand. The policy suggests an “order-up-to rule” that states that, if the inventory level (i) is less than the trigger level (s) an order of size o is placed that will bring the total amount of inventory on hand up to a target level (S) where $o = S - i$. The methods for finding good two-bin policies, while not always simple, are well known for nonperishable products.

Perishable inventory models are much more difficult to solve. While it is known that an exact solution can be found for perishable inventory problems via DP, it is difficult to actually solve a DP model for problems of realistic size. DP is essentially a search across the solution space for a particular problem. All combinatorial problems (of which the platelet inventory ordering problem is an

example), can be solved by naïve enumeration. That is, if we simply check all of the possible combinations of decisions for every possible combination of inventory state, we will eventually find the solution to any problem. The difficulty, in practice, is that the solution space for a realistic size problem may be so large that even the fastest computers, working for hundreds or thousands of years, cannot generate all the possible combinations. Thus, while it is easy to define the platelet ordering problem in a DP format, actually solving such a model is usually impractical.

In the absence of exact solution methods, most research in the area of platelet ordering problem has focused on the solution of either approximate models or the development of heuristic techniques to find good, if not optimal, solutions to the problem quickly. Some researchers have attempted to solve the platelet ordering problem for restricted problem sizes (see Refs 1 and 21). However, even if the size of the problem is restricted and demand is aggregated from units to doses, exact solutions to the platelet ordering problem remain very difficult to obtain. Other researchers have attempted to set platelet inventory policy through a heuristic search of possible trigger points and order-up-to levels. While such models can be used to develop policies that are good, their solution cannot be proven to be optimal.

In addition to being hard to obtain, DP solutions are difficult to characterize and thus, to implement. Solution values depend on the exact number of units on hand, their age distribution, the day of the week, and the particular day in a planning horizon. This makes it difficult to implement a DP solution in practice, since the number of combinations of day, date, inventory status, and expected demand may be in the range of hundreds of millions.

Given that the exact solutions to the platelet problem are hard to calculate and difficult to implement, most research in the area relies on some form of problem simplification or a heuristic approach to produce a usable solution. Heuristics may involve simple rules of thumb, for example ordering up to the expected demand over the expected life span of the platelets, or they

may involve more complicated procedures, such as the one outlined in Ref. 21. There is a trade-off between the quality of a solution obtained and the time required to obtain that solution. More effort implies a better result, but the law of diminishing returns applies to combinatorial problems and, hence, doubling the computational time does not necessarily improve the quality of the solution by a factor of 2.

CONCLUSION

The safe and effective delivery of health care requires that a sufficient supply of platelets be available when and where required. However, because platelets have a very limited shelf life and are expensive to collect and produce, outdates are an important practical concern for hospitals and blood system operators. The platelet inventory problem revolves around identifying policies for ordering and holding platelets such that unit availability is maximized, while ensuring that outdates are minimized. Cost minimization is also a concern, but it is unclear as to how the intangible effects of shortages and outdatedness can be priced relative to one another and also against concrete operational costs.

The platelet inventory and ordering problem can be formulated as a DP problem. Exact solutions can be obtained for small problems. However, the computational complexity associated with DP makes this method intractable for realistically sized problems. Hence, heuristics are generally employed to solve practical problems.

While there is extensive literature on perishable inventory, with much focus on red blood cells, the OR literature on platelet inventory and ordering is surprisingly sparse. Despite the fact that platelet ordering is a very common practical problem with significant impact on human health, it is not well represented in the literature. The platelet inventory and ordering problem therefore, remains a rich area for theoretical development and an avenue for application of operational research with significant impact.

REFERENCES

1. Blake J, Smith S, Arellano R, Anderson D, Bernard D. Optimizing the platelet supply chain in Nova Scotia. In: Proceedings of the 29th Meeting of the European Working Group on Operational Research Applied to Health Services. Prague: ORAHS; 2006. pp.47–66.
2. Tinmouth A, Fergusson D, Yee I, Hebert P. Clinical consequences of red cell storage in the critically ill. *Transfusion* 2006;46(11):2014–2027.
3. van der Poel C, Janssen M. The collection, testing, and use of blood and blood products in Europe in 2003. Strasbourg: Council of Europe; 2005.
4. Veinott A. Optimal policy for a multi-product, dynamic, non-stationary inventory problem. *Manag Sci* 1965;12(3):206–222.
5. Pierskalla W, Roach C. Optimal issuing policies for perishable inventory. *Manag Sci* 1972;18(11):603–614.
6. Fries B. Optimal ordering policy for a perishable commodity with fixed lifetime. *Oper Res* 1975;23(1):46–61.
7. Nahmias S. Optimal ordering policies for perishable inventory. *Oper Res* 1975;23(4):735–749.
8. Nahmias S. On ordering perishable inventory when both demand and lifetime are random. *Manag Sci* 1977;24(1):82–90.
9. Nahmias S. The fixed charge perishable inventory problem. *Oper Res* 1978;26(3):464–481.
10. Nahmias S. Perishable inventory theory: a review. *Oper Res* 1982;30(4):680–708.
11. Cohen M, Pierskalla W, Sassetti S, Consolo J. An overview of a hierarchy of planning models for regional blood bank management. *Transfusion* 1979;19(5):526–534.
12. Brodheim E, Hirsch R, Prastacos G. Setting inventory levels for hospital blood banks. *Transfusion* 1976;16(1):63–70.
13. Brodheim E, Prastacos G. A regional blood management system with prescheduled deliveries. *Transfusion* 1979;19(4):455–462.
14. Kendall K, Lee S. Formulating blood rotation policies with multiple objectives. *Manag Sci* 1980;26(11):1145–1157.
15. Freidman B, Abbott R, Williams G. A blood ordering strategy for hospital blood banks derived from a computer simulation. *Am J Clin Pathol* 1982;78(2):154–160.
16. Hesse S, Coullard C, Daskin M, Hurter A. A case study in platelet inventory management. In: Curry G, Bidanda B, Jagdale S,

- editors. Sixth industrial engineering research conference proceedings. Norcross, GA: IIE; 1997. pp. 801–806.
17. Sirelson V, Brodheim E. A computer planning model for blood platelet production and distribution. *Comput Meth Programs Biomed* 1991;35(4):279–291.
 18. Katz A, Carter C, Saxton P, Blutt J, Kakaiya R. Simulation analysis of platelet production and inventory management. *Vox Sang* 1983;44(1):31–36.
 19. Katsaliaki K, Brailsford S. Using simulation to improve the blood supply chain. *J Oper Res Soc* 2007;58(2):219–227.
 20. Brailsford S, Katsiliaki K, Mustafee N, Taylor S. Modelling very large complex systems using distributed simulation: a pilot study in a healthcare setting. In: Robinson S, Taylor S, Brailsford S, Garnett J, editors. 2006 OR society simulation workshop. Leamington Spa: OR Society; 2006. pp. 257–262.
 21. van Dijk N, Haijema R, van der Wal J, Sibinga J. Blood platelet production: a novel approach for practical optimization. *Transfusion* 2009;49(3):411–420.
 22. Blake J. On the use of Operational Research for managing platelet inventory and ordering. *Transfusion* 2009;49(3):396–401.
 23. Erhun F, Chung Y, Fontaine M, Galel S, Rogers W, Sussmann H. Publications. Retrieved July 4, 2009, from Feryal Erhun: 2008. <http://www.stanford.edu/~ferhun/>

AN INTRODUCTION TO PROBABILISTIC RISK ANALYSIS FOR ENGINEERED SYSTEMS

ELISABETH PATÉ-CORNELL
Department of Management
Science and Engineering,
Stanford University, Stanford,
California

FAILURE RISK ASSESSMENT

The risk of failure of a system includes both, the probabilities and the consequences of the different failure scenarios.¹ It can be described by the probability distribution of the damage per time unit or operation.² For a complex engineered system, one may not have a sufficient statistical database to assess failure probability at the system level using classical frequentist definitions [1]. That may be true because there is not enough experience with the system, because it is not in a stable state, or because failures are too rare to have been observed systematically in the past. In that case, one has to rely on systems analysis and on the Bayesian definition of probability as a rational degree of belief about the chances of occurrence of an event in a specified reference frame [2–4].

Probabilistic risk analysis³ (PRA) has been developed in engineering, in particular

in the nuclear power industry [5] to permit computing the risk in cases where there is insufficient statistical information at the system level, but where some information is available at the level of subsystems or components [6–12]. Events and random variables in the possible scenarios are combined systematically, accounting for dependencies and rare occurrences which are often difficult for the human mind to grasp without analytical support.

A Brief History of PRA

PRA has a rich history. In the nuclear power industry, as mentioned above, it was particularly helpful in providing safety information in the early years of the civilian nuclear power program, when there was limited experience at the system level, but considerable amounts of additional information at the subsystem level, for example from the US nuclear Navy. PRA has been used since then in many other settings, such as chemical plants [13] and space systems [14]. In electrical engineering, the reliability of a circuit can be assessed as a function of the reliability of its components using fault trees to provide a logical (Boolean⁴) relationship between the failure of the whole system and that of its components, in parallel or in series. For instance, Haasl [15] describes the early use of fault trees in the analysis of aviation safety. In civil engineering, one problem is to compute the probability of failure of a structure, given its capacity and the loads to which it may be subjected [16]. When both are uncertain, they can be described by probability distributions. The probability of failure is then computed as the probability

¹In some cases, the problem is to compute only the failure probability per time unit or operation for instance, because the consequences are well known.

²Note that the expected value of the losses is generally not a sufficient description of the risks, especially in cases where rare failures can cause large damage.

³Probabilistic risk analysis (PRA) is also called probabilistic risk assessment, or quantitative risk analysis (QRA) or probabilistic safety assessment or analysis (PSA).

⁴Boolean algebra provides logical relations (e.g., AND and OR functions) among variables that can take values of 1 or 0—true or false—for example, to represent the state—failure or no failure—of an element.

that the loads (for example, the seismic loads) exceed the capacity in a specified time frame.

In mechanical engineering, aeronautics, and astronautics, the same methods have been applied to new and/or complex systems, ranging from automobiles involving new components to space systems.⁵ A probabilistic analysis of the risk of a potential failure of the heat shield of the US space shuttle [17,18] showed that 15% of the tiles that protect the orbiter at reentry represented about 80% of the risk.⁶ More recently, the same probabilistic approach has been used for medical devices, which may need to be tested in patients before they are approved by the Food and Drug Administration [19]. In new systems, with which there is little experience *in situ* such as a new type of satellite or medical device, the failure probability can be computed by a PRA-type of model, based on an analysis of the functions to be performed and on marginal and conditional probabilities of component failures.

PRA allows using all available information to represent uncertainties about a system's performance. The data may include direct observations, surrogate data (same elements in another setting), engineering models, test results, and expert opinions. PRA relies on an analysis of the functions to be performed by the whole system and on the probabilities of failure of its basic components and subsystems. As described by Garrick and Kaplan, the analysis is guided by the questions: What can go wrong? With what probability? And with what consequences? [6]. Similarly, the risk management problem can be described as finding and fixing a system's weaknesses [21] by asking the questions: How does it work? How can it fail? And what can be done about it, given that we are not infinitely rich and that days have only 24 h? The risk assessment results then become inputs into

a problem of risk management and optimization of resource allocation.

As described further, the PRA method relies in part on event trees and fault trees, or similarly, on Bayesian networks (or influence diagrams). It includes the effects of external events that may affect the performance of the components, possibly several at the same time, thus creating dependencies among basic failures. It also includes a consequence model (e.g., economic) to assess the outcome of each scenario. The PRA method, which was originally designed on the basis of the technical performance of components, can be extended as shown further to include human and organizational factors. In a different form, the same approach can be used to address environmental problems [22] and to examine other types of systems involving human networks and organizations. Risk analysis can then be combined with game theory to assess for instance, the risk of an attack when intelligent actors are involved [23,24].

Objectives of PRA

PRA has two main purposes: to optimize resource allocation (e.g., to minimize the probability of system failure given budget and schedule constraints) and to check that the failure risk is tolerable. Proactive risk management requires recognizing, anticipating, and correcting problems before an accident occurs. PRA allows improvement of both system design and operations by setting priorities among risk reduction measures, while accounting for costs, benefits, and uncertainties about them [25]. Clearly, such decisions also require a value judgment; for instance, what costs are justified by a given decrease of the failure risks, or what level of failure probability is tolerable given that few risks can be reduced to zero unless the hazard is eliminated altogether [26–28].

Therefore, one of the major functions of a risk analysis is to represent uncertainties in risk management decisions, especially for systems that are poorly known or when failures are rare. These uncertainties can be treated at different levels of complexity [29]. What is described here is a PRA method that generally yields a distribution of the

⁵For example, a complete risk analysis has been performed for the International Space Station ([20]).

⁶It is such a heat shield failure that eventually caused the accident of the Columbia orbiter in 2003.

outcomes by their complementary cumulative distribution.⁷ Such a risk curve can then be used as input in a single-attribute decision analysis [30] in the framework of rationality defined by the von Neumann axioms [31]. The result can also involve several dimensions (monetary losses, human casualties, and environmental damage). The result of the PRA can then be represented by a surface, which becomes an input in a multiattribute decision analysis [32].

Risk assessment is not a static exercise. It represents the state of knowledge about a system state at a given time, but the result may change, either with additional information or with changes in the system (improvements or deterioration). Therefore, to be useful, a PRA must be a “living document,” updated as more information becomes available.⁸ In that respect, warnings and precursors play an essential role. Observing and interpreting precursors and signals of defects or malfunctions are essential in guiding both risk assessment and risk management [34]. This involves integrating occurrences and observations of these precursors, updating the probabilities of the corresponding failure modes or accident sequences, and taking timely corrective measures.

Organization of This Article

This article describes and illustrates the PRA method and some of its extensions. First, the notion of probability and the fundamental rules of probability computation are presented. The next section describes the basic tools of PRA: events trees, fault trees, and functional block diagrams. The PRA process is then described and illustrated by several examples including the risks of subsystem failures in nuclear reactors and oil spills caused by loss of propulsion of oil tankers.

⁷The complementary cumulative distribution of a random variable shows the probability of exceeding different loss levels. These risk curves are sometimes called “Farmer risk curves” [33].

⁸Note that more information does not always imply a reduction of uncertainty. This occurs for example, when experience shows that the prior probability (assessment *a priori*) was way off base.

The dynamic aspect of accident sequences is illustrated by the case of patient risks in anesthesia, starting from the initiating event (e.g., a disconnection of the oxygen tubes) and ending with the recovery—or not—of a patient who cannot live when deprived of oxygen for more than a few minutes [35]. In many cases, the main challenge was in the formulation of the problem; that is, which variables and relationships to consider, dynamic representation, economic analysis, and so on.

PROBABILITY AND SOURCES OF DATA

As mentioned earlier, failure probabilities can be based on two different definitions: the classical statistics interpretation based on the frequency of an event in a sufficiently large sample, and the Bayesian approach, which relies on the degree of belief of a decision maker or a risk analyst. Bayesian probabilities are obtained by considering first, the prior probability of an event, a state or a hypothesis before additional information is acquired, then by updating this prior with new data. This is done, as shown further, by combining the prior probability with the likelihood of a new piece of information (probability of obtaining these data given the state of interest) to obtain a posterior (updated) probability of that state.

It is important to note that the Bayesian definition of probability allows using all relevant information, not only statistical data when they exist, but also expert opinions. The Bayesian method has been used, for instance, in the nuclear power industry for parameter estimation [36].

Fundamental Rules of Bayesian Probability Computation

Two laws of probability are at the basis of probability computation and in particular, in a PRA context: the total probability theorem and Bayes theorem.

Notations:

$P(A)$ = marginal probability of event or property A ,

$P(A, B) = P(A \text{ AND } B)$: joint probability of A and B ,

$P(A | B) = P(A \text{ GIVEN } B)$: conditional probability of A given B ,
 $P(\text{NOT } A) = \text{probability of NOT } A = 1 - P(A)$.

The *Total Probability Theorem* links the probability of an event A (e.g., system failure) to the probability of the intersection of the set of scenarios containing A with a set of mutually exclusive and collectively exhaustive scenarios B_i that may or may not contain A .⁹ For the simple case where there are only two B_i cases, scenarios B or NOT B , the total probability theorem can be written as

$$P(A) = P(A \text{ AND } B) + P(A \text{ AND } (\text{NOT } B))$$

More generally, for a set of appropriately structured scenarios B_i , the theorem can be written as

$$P(A) = \sum_i P(A \text{ AND } B_i)$$

The *Bayes theorem* links the prior probability of A and the likelihood of B to the obtain the posterior probability of A once it was established that B is true. Consider the case where A is a particular hypothesis and B a signal or some type of observation.¹⁰ $P(A)$ is the prior probability of A , $P(A | B)$ is the posterior probability of A once B has been observed. $P(B | A)$ represents the likelihood of observing that signal if A is true (or about to happen). The Bayes theorem can be written as

$$P(A | B) = P(A, B) / P(B) = P(A) \times P(B | A) / [P(B, A) + P(B, \text{NOT } A)]$$

This formula permits computing, by expansion, the probability of a scenario involving A and B as the joint probability of these

two components.¹¹ It allows, in particular, accounting for dependencies among A and B represented by $P(A | B)$ or by $P(B | A)$. Only in the case where A and B are independent, can one write that $P(A | B) = P(A)$ and therefore, $P(A, B) = P(A) \times P(B)$.

The same formula applies to a scenario involving more than two variables or events, for example, A and B and C and D . The Bayes theorem permits computation of this scenario as the joint probability of A , B , C and D expanded as

$$P(A, B, C, D) = P(A) \times P(B | A) \times P(C | A, B) \times P(D | A, B, C).$$

Two Kinds of Uncertainties

It is sometimes assumed that probability simply represents the randomness that one may observe through a statistical sample. In reality, uncertainty encompasses aleatory uncertainty (randomness) and epistemic uncertainty that reflects one's lack of knowledge about fundamental mechanisms or phenomena [37]. PRA may involve both, in which case the results may be represented not only by a single risk curve (combining both types of uncertainties) but by a family of risk curves separating both in the display of results [38]. In any case, both types of uncertainties can be described by Bayesian probability, which is the basis of the methods of risk assessment presented here.

One main problem is to assess the different probabilities that are the input of a PRA [39]. In what follows, it is understood that one can use classical statistics to generate data about the different parts of the problem and rely on observed failure frequencies of various components when that information is available. This requires samples of sufficient size and failure mechanisms that are stable enough over time so

⁹For instance, if A represents a system's failure, the probability of A is computed as the sum of the probabilities of all scenarios containing A , provided these scenarios do not overlap and cover all possibilities.

¹⁰ A can be a system's failure and one may want to compute the probability of A given that a crack has been observed in one of its critical subsystems (B).

¹¹Note that the order in which A and B are introduced does not matter because $p(A \text{ AND } B)$ is equal to $p(B \text{ AND } A)$. Therefore, the Bayesian formula as written here is equivalent to $P(A, B) = P(A) \times P(B | A) = P(B) \times P(A | B)$.

that past statistics provide adequate probability estimates. Often however, it is not the case by, for example, for new systems. The Bayesian approach then allows using all relevant information including test data and surrogate data that are collected in similar but not identical settings, and have to be updated to accurately represent the phenomenon of interest in actual operations. Relevant information also includes expert opinions, that is, subjective degrees of belief from specialists of the different parts of the system, which the analyst adopts based on his or her confidence in the expert.

Expert Opinions

In the Bayesian framework, expert opinions play a critical role, among other things because they allow assessing epistemic uncertainties since by definition, one does not have statistical frequencies to represent them. The opinions of experts reflect their experience but may be distorted by cognitive biases [40,41] as well as a desire to influence the risk management decision. The likelihood function that is used to include an expert's opinion in the updating of a prior probability thus reflects the confidence of the analyst in the judgment of the expert.

Problems arise when using the opinions of several experts who disagree. The analyst then faces the challenge of aggregating these opinions to generate PRA inputs. This can be done by simply weighting the expert opinions in a linear aggregation function to obtain the probability of interest. A more sophisticated approach is to use a Bayesian analysis, which allows accounting for possible dependencies among the opinions of different experts who may share for example, the same fundamental model about the variable of interest [42,43]. This approach requires the use of probabilities (likelihoods) that represent not only the confidence of the analyst in each expert, but also the analyst's assessment of dependencies among the opinions of the experts.

The aggregation of expert opinions can also be done by an iterative process such as the Delphi method, in which each expert is asked to provide an estimate of a probability, the results are then aggregated by the

analysts and sent back to the experts who are given an opportunity to revise their individual judgments. These adjustments can be reiterated until the process converges [44]. Another approach is based on direct interactions among experts, which permits explicit elicitation of mental models, exchange of experience, and comparison of information bases, in order to come up with a collective estimation. This was done in the case of seismic risk analysis by Budnitz *et al.* [45].

The use of expert opinions introduces an element of subjectivity in a Bayesian analysis. The question is—what are the alternatives? In the absence of a systematic analysis, the same element of subjectivity would exist, but at a higher level of analysis, where the information base and the level of experience may be much thinner. Instead, decomposing the problem into different parts allows using the experience of the most qualified experts in the different fields involved. For example, when assessing seismic risk for a given structure, the problem can be decomposed between seismic hazard, which is the speciality of seismologists, and buildings' vulnerability which is in the domain of structural engineers. In general, the result is thus better informed than a direct subjective assessment of the failure risk.

Using all these sources of information and data, the role of the analyst is to support a decision by representing uncertainties about all variables and events (including experts' dissents) as accurately as possible and as free of value judgment as can be achieved. It is then the role of the decision maker to include his or her risk attitude in the risk management decision to be made.

BASIC TOOLS OF ENGINEERING RISK ANALYSIS: EVENT TREES, FAULT TREES, AND FUNCTIONAL BLOCK DIAGRAMS

The PRA process involves identification of the failure scenarios (conjunctions of events leading to failure) and computation of the probabilities and consequences of these scenarios. Two fundamental tools used in that process are event trees and fault trees. Event trees permit systematic identification and

structuring of the set of possible scenarios so that they are collectively exhaustive (all possibilities are included) and mutually exclusive (each scenario appears only once in the tree). Fault trees permit identification of the scenarios (conjunctions of component failures) that lead to system failure. In complex systems, the first question is—what functions must be performed for the system to work? This is described by functional block diagrams.

Event Trees

Event trees represent a systematic identification of possible scenarios, which are conjunctions of uncertain events and/or random variables. These may include component failures as well as external events such as waves, winds, or earthquakes that affect the probability of failure of components or subsystems. The structure of a generic event tree is shown in Fig. 1.

The tree is read from left to right. The circles represent chance nodes corresponding to an event or a random variable. Each chance node is followed by branches representing

the possible realizations of that chance event or variable (e.g., the subsystem succeeds or fails). On each branch, one places the probability of that realization of the chance node given everything else that precedes it in the tree. A “path” (or scenario) is a set of branches that link the original event to a possible outcome (e.g., overall system failure or success). Event trees thus provide a simple way of representing both the scenarios and the dependencies among events. Figure 1 includes three chance nodes, A , B , and C . The probability of each scenario (hence the probability of the corresponding outcome) is simply the product of the conditional probabilities along the corresponding path. For example, the probability of the path or scenario (A , NOT B , C) is

$$P(A, \text{NOT } B, C) = P(A) \times P(\text{NOT } B | A) \times P(C | \text{NOT } B, A).$$

The structure of an event tree is thus determined by conditional probability. Its elements are not necessarily displayed in chronological order, even though it is often

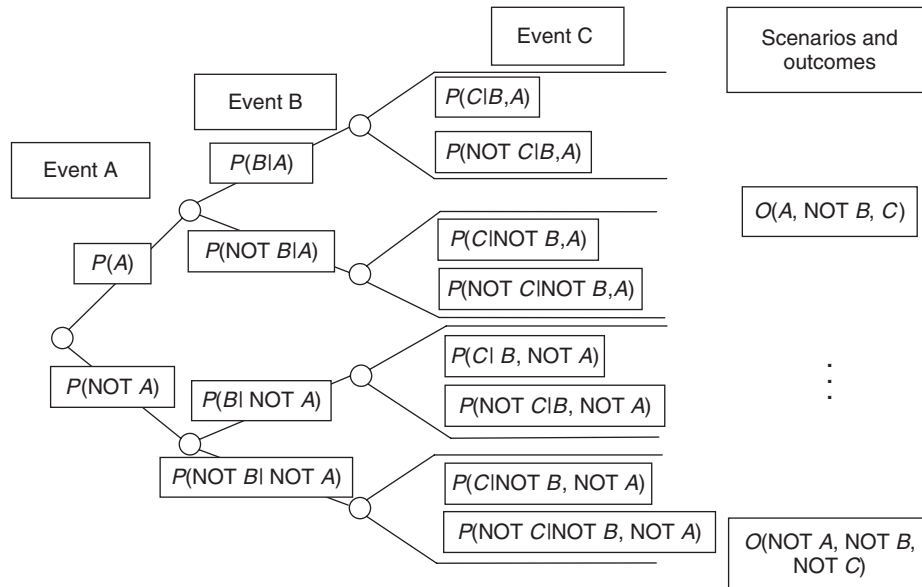


Figure 1. Representation of an event tree including three binary events A , B , and C (for example, A : initiating event; B : intermediate development; C : final system state, and O : outcome such as the amount of a product spilled in the environment as a result of system rupture).

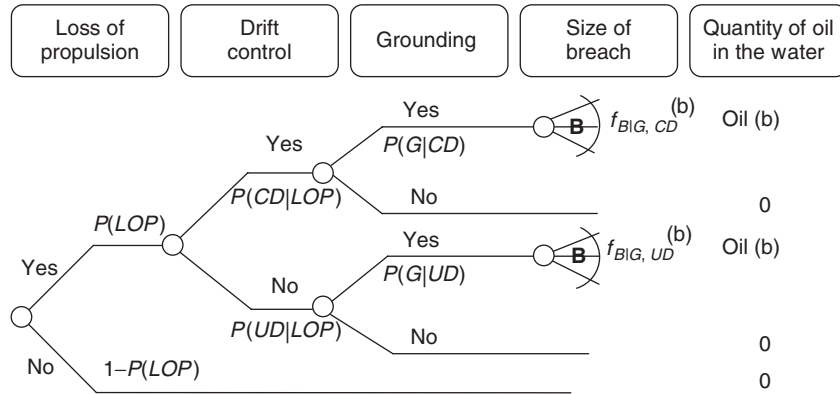


Figure 2. Event tree to compute the risk of ship grounding due to loss of propulsion. (Source: Paté-Cornell [11].)

a convenient way to build the tree. Instead, the order of the variables may be adapted to fit that of the available information.¹² At the end of each path, the scenario's outcome is displayed, whether it represents a system's failure or decreased performance. The consequences can be financial, health-related, or environmental.

An example of an event tree in a risk analysis is shown in Fig. 2. That simplified event tree can be used to compute the risk of an oil spill due to loss of propulsion, LOP, of an oil tanker. The simplified scenarios are the following. If a tanker loses propulsion, it may start drifting depending on external factors (currents, tides, winds etc.) and on the skill of the crew to control the drift. If it is close to the sea floor or obstacles (e.g., rocks), the risk may be grounding, which can cause a breach (of different possible sizes) in the hull. At that point, various amounts of oil can be released into the sea.

Figure 2 thus shows both discrete events (LOP; drift control, CD or not, UD, and grounding of the ship G), and continuous random variables (the size of the breach in the hull, B , and the amount of oil spilled in the sea, O). The discrete random variables

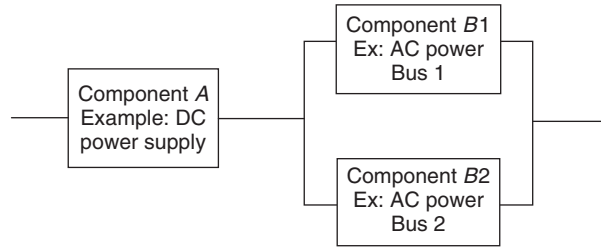
are characterized by their marginal and conditional probabilities. The size of the breach in the ship hull is characterized by its probability distribution function, conditional on drift control, and grounding (realizations of B are noted b). The amount of oil spilled in the seawater is the result of that event tree. It is characterized by its probability distribution conditional on the breach size. Its overall (marginal) probability distribution is obtained by summing the probabilities of oil spilled for all breach sizes.¹³ Further analysis of the consequences (not shown here) allows quantifying the probability distribution of associated losses.

In the case of ship grounding, the sequence of events leading to failure is relatively straightforward. It is not necessarily the case for complex systems involving multiple redundancies. What an event tree does not always show is whether or not a conjunction of component failures causes the whole system to fail. That information can be provided by a description of the functions that the system must perform to work and by fault trees, which use that information to derive the sets of component failures leading to failure of the whole system.

¹²For instance, it may be convenient to place A followed by B (in the tree) if one can assess directly $P(A)$ and $P(B|A)$. But if $P(B)$ and $P(B|A)$ are easier to assess, the order may be B followed by A .

¹³The amount of oil spilled in the sea water (in this case) can also be referred to as the "source term."

Figure 3. Functional block diagram for a system of two subsystems in series *A* and *B*, the second one (*B*) composed of two components in parallel, *B1* and *B2*.



Functional Block Diagrams

Sometimes called “reliability block diagram”, functional block diagrams represent the main functions to be performed for the system to work and the corresponding subsystems in series. For each of these main functions, the diagram then includes the components or subsystems that can perform the function. Each function can then be analyzed according to its structure (components in series or in parallel); for example, whether or not each function involves redundancies. Figure 3 represents as an illustration, the functional block diagram for a power supply system that requires both DC and AC power for the system to work. That system is composed of one DC power supply and two redundant AC power buses. The system shown here fails either if component *A* (DC power) fails or if subsystem *B* (AC power) fails, which requires that both redundant components *B1* and *B2* fail.

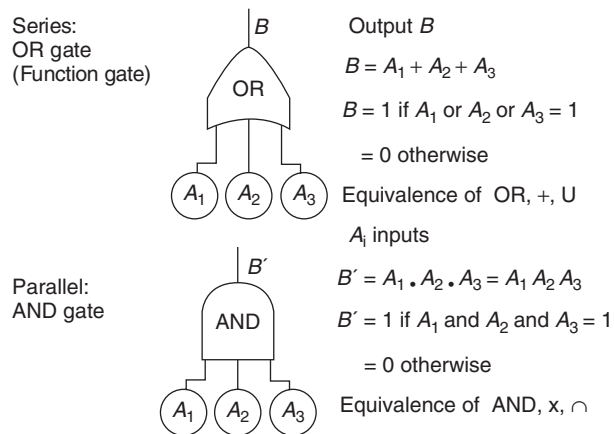
Fault Trees

A fault tree represents the logical relationship between the failure of a system and the failures of its subsystems and basic components. It is a snapshot showing how the state of the system depends on the state of its components. It is composed of a set of logical Boolean functions (for instance, OR gates and AND gates as shown below in Fig. 3) whose inputs and outputs are Boolean (0 or 1) variables.¹⁴

Figure 4 shows two examples of fault tree functions (“gates”). The top one (“OR gate”) represents the logical link between the failure of a system ($B = 1$) and that of its components in the case where $B = 1$ if any of the three inputs A_1, A_2 , or A_3 are equal to 1. This function represents the case of a system (*B*) of three components in series (A_1, A_2 , and

¹⁴ $X = 1$ means here that component *X* is in a failed state.

Figure 4. Logical (OR and AND) gates in a fault tree.



A3) so that the whole system B fails if any of the A's fail. On the right-hand side of the tree, the figure shows several notations used to represent that OR function¹⁵:

$$B = A1 \text{ OR } A2 \text{ OR } A3;$$

$$B = A1 + A2 + A3;$$

$$B = A1 \cup A2 \cup A3,$$

which means that B represents the unions of all sets of scenarios including A1, A2 or A3.

The bottom part of Fig. 4 represents the logical link between the failure of a system ($B = 1$) that occurs if all of the three inputs A1, A2, or A3 are equal to 1; that is, B fails if all of the three components A1, A2 and A3 have failed. This function represents the case of a system (B) composed of three components in parallel (A1, A2, and A3) such that the whole system B fails if all three redundant elements A1 and A2 and A3 fail. Again, on the right-hand side of the tree are several notations used to represent the function¹⁶:

$$B = A1 \text{ AND } A2 \text{ AND } A3;$$

$$B = A1 \cdot A2 \cdot A3$$

$$B = A1 \cap A2 \cap A3,$$

which means the intersection of A1, A2, and A3.

The fault tree corresponding to a particular system is derived from its functional structure as described by its functional block diagram. For a system of elements in series to work, all the elements must work. For a system of elements in parallel to work, at least one of the (redundant) components must work. The system or subsystem thus fails if any of its basic functions in series are not performed, or if all of its elements in parallel have failed. Boolean algebra allows

writing and simplifying the logical equations represented by a fault tree. A Boolean polynomial made of the Boolean functions shown in Fig. 3 can be written based on the fault tree to represent the “top event” (output of the fault tree) as a function of the basic events (inputs of the tree).

A fault tree is generally constructed top-down, in a deductive manner, starting with the failure of the whole system. The failure event is then decomposed into failures of subsystems and components in parallel (AND gate) and/or in series (OR gate) according to the structure of its functional block diagram. Figure 5 shows a fault tree corresponding to the power system represented in Fig. 3 whose role is to provide power to the emergency safety features (ESF) of a reactor. Figure 5 (which can be read from top to bottom) indicates that the system fails if there is no AC power or no DC power (event A), and that for AC power to fail, there must be loss of power from both on-site (event B1) and off-site (event B2) buses. The top event (main failure) is the loss of power to ESF.

Fault trees and events trees are often used together in technical risk analyses. The structures of the model and its submodels, involving the identification of all possible classes of scenarios, showing their probabilities and their consequences, are represented by event trees. Fault trees, based on functional block diagrams, are used to identify the scenarios that lead to system failures and to compute the probability of failure of the whole system. That information can be direct input into an event tree, which can include in addition, external events, human errors and other factors that affect component failure probabilities.

THE PRA PROCESS AND RESULTS

The PRA process is a systematic way to identify a system's failure modes and quantify its failure risks. The first step is to identify the functions involved in the system's operations, and the set of scenarios that may follow an “initiating event” triggering an

¹⁵The OR function, represented by the + sign in Boolean algebra, is such that $1 + 1 = 1$, $1 + 0 = 1$, and $0 + 0 = 0$.

¹⁶The AND function in Boolean algebra represented by the $\times$ sign, is such that $1 \times 1 = 1$, $1 \times 0 = 0$, and $0 \times 0 = 0$.

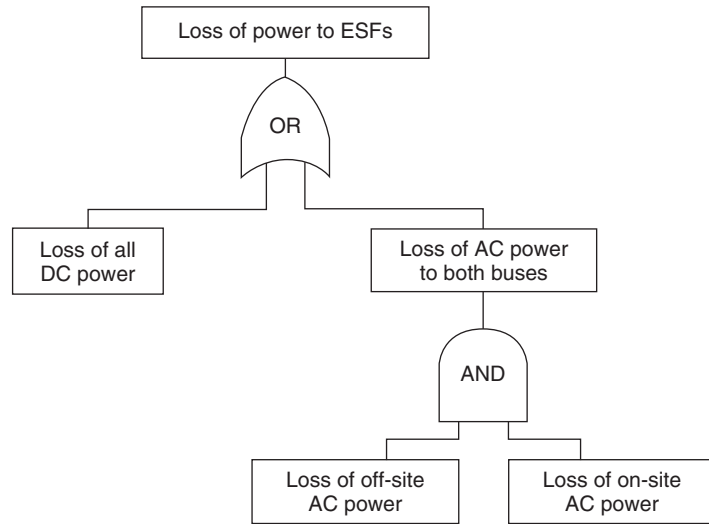


Figure 5. Fault tree for the failure of the power supply of an emergency system (ESF: Emergency Safety Feature). (Source: USNRC [5].)

accident sequence.¹⁷ As mentioned earlier, to be amenable to probabilistic analysis, these classes of scenarios have to be structured logically (beyond simple “what if?” questions) so that they are mutually exclusive and collectively exhaustive. The second step involves computation of the scenarios’ probabilities, including external events that can affect the probabilities of basic failures. The third step is to assess the scenarios’ consequences, including the different attributes relevant to the decision; for example, the effects of system failures on human safety, monetary results, and environmental damage. The results can be presented as a single risk curve if the inputs of the PRA are the probabilities of events and distributions of random variables. The analysis can be performed at a second level of uncertainty analysis, which involves uncertainties in

the future frequencies of component failures and the results of these uncertainties on the overall failure probability.

First Level of Uncertainty Analysis: A Single Risk Curve

Given a specified engineered system, the PRA process often starts with a functional block diagram. The functional block diagram can then support the construction of a fault tree. One can derive from that fault tree the failure modes or “minimal cut sets”, that is the minimum sets of component failures leading to failure of the whole system.¹⁸ To compute the probability of each type of scenario, one can use event trees (or their equivalent) especially in the case where external events and loads are involved in failure probabilities. As described earlier, these scenarios represent the conjunctions of events defining

¹⁷The choice of a level of detail is critical to the feasibility of the analysis. Very detailed scenarios add complexity that is not always useful. Instead, the issue is to decompose the problem into classes of scenarios that are manageable, and for which one can identify clearly the relevant information. The appropriate depth of analysis may vary across the subsystems; for instance, more details may be required in places where there is little operational experience.

¹⁸Note that fault trees involve Boolean variables (1 or 0) that represent functional failure (or not) of each component. In their classic form they do not include partial failures in which the component may not work perfectly but still well enough for the system to function.

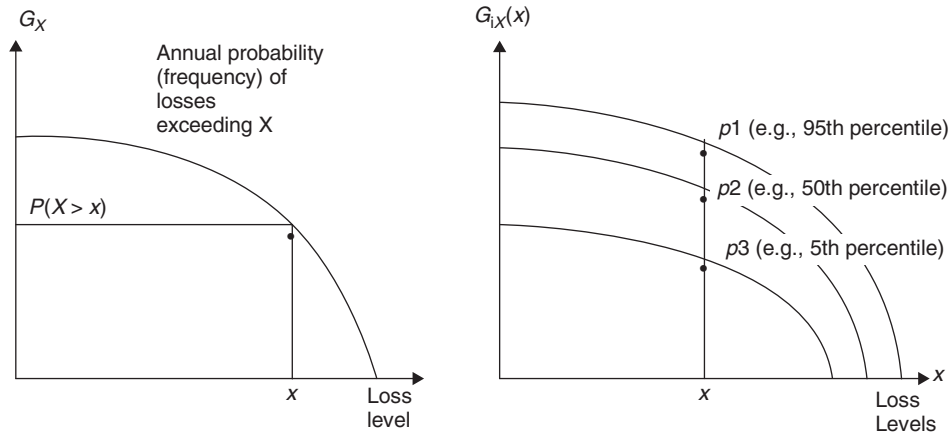


Figure 6. Risk curve and family of risk curves representing the results of first and second levels of uncertainty analysis in a PRA.

each scenario and allow computing the corresponding joint probability.¹⁹

Fault trees and event trees, as illustrated further for a nuclear power system, are thus fundamental tools of the PRA method; the former relying on logic functions and the latter on conditional probability [46]. Finally, specific consequence models (e.g., dispersion analyses, dose-response analyses, economic analyses) are needed to assess the outcomes of system failures (partial or total). When the consequences of the different scenarios have been assessed, the results can be displayed as a probability distribution (density function) of the losses per time unit or operation. Alternatively, they can be represented as a cumulative distribution function or its complement; that is, a risk curve that shows the probability of exceeding different loss levels in a given time unit (see left-hand side of Fig. 6).

¹⁹Fault trees and event trees can be considered roughly equivalent, although fault trees represent logical functions and event trees represent exhaustive sets of scenarios (including failures or not) and are based on marginal and conditional probabilities. The main difference arises when external events are involved. They are not part of failure modes but affect the probability of failure of components. They are thus best handled by event trees.

The main characteristic of this process is to be systematic and rational according to the classic criteria of von Neumann [31]. It also allows accounting not only for uncertainties, but also for dependencies among failures. Dependencies can arise for example, when failure of one component puts a higher load on a redundant one, or when an external event (e.g., an earthquake) applies simultaneous loads on several components. Dependencies can also be caused by a common manufacturing process or an operator error that affect several components.

Second Level of Uncertainty Analysis: A Family of Risk Curves

A second level of analysis may be required to represent uncertainties about the input; that is, about the future failure rates of basic components. This is needed if the decision maker is sensitive to uncertainties about the failure probabilities [47]. Also, in many cases, the analysis is done for one system per time unit. The second level of uncertainty is needed to extend that analysis to several identical systems for several time units [48]. This requires propagation of uncertainties through the PRA and representation of these uncertainties in the result [38].

Uncertainties about future failure frequencies can sometimes be characterized by lognormal distributions. If these variables

are multiplied in the analysis, the product is also lognormal yielding a closed-form solution. Also, uncertainties about the probabilities of basic events can be represented, for instance, by beta distributions. In most cases, one has to simulate the propagation of uncertainties about basic event probabilities throughout the PRA model to represent their effects on a distribution of the overall system failure probability, or on the probability of exceeding different loss levels per time unit. The risk analysis result is then a family of risk curves (right-hand side of Fig. 6). Each curve represents a fractile (e.g., 10%, or 95%) of the probability distribution of the chances of exceeding various levels of losses in a given time period or operation [38]. In Fig. 6, the meaning of $P_1(x)$ is that there is a probability 0.95 that $P(X > x)$ is less than P_1 , or, $P[P(X > x) < P_1] = 0.95$. The meaning of P_3 is that there is a probability 0.05 that $P(X > x)$ is less than P_3 . What follows in this article is limited to the first level of uncertainty analysis, resulting in a single risk curve.

EXAMPLES AND ILLUSTRATIONS FROM THE NUCLEAR POWER INDUSTRY

Different Models That Need to be Combined

The general model of the risk of failure of nuclear reactors (and of many other systems that can release toxic or polluting material) can be described as a combination of submodels as shown in Fig. 7.

Note that this model structure applies with only minor modifications to many other cases. For example, submodel #3 as shown in Fig. 7 yields the amount of fission product released in the atmosphere in the case of a nuclear reactor or toxic substance released in the case of a chemical plant. This amount released (“source term”) often depends on the final state of the system. In nuclear reactors, it can be characterized (among other factors) by the damage to the containment structure. In the case of grounding of an oil tanker, this final system state can be described by the size of the breach in the hull, the source term is the quantity of oil released in the sea, and the “dose-response” relationship describes the health and environmental damage associated with the oil spill.

The two following examples (both from a risk analysis for nuclear reactors) illustrate the tools and the process described in the previous sections. First, the risk posed by the possibility of a steam-generator pipe rupture is used to show the structure of a risk analysis model involving both event trees and fault trees. Second, the failure risk of an auxiliary feed water system (AFWS) illustrates the use of functional block diagrams to identify failure modes and compute their probabilities given the possibility of external events such as earthquakes.

Example 1. Steam-generator pipe rupture. In this example (from the Reactor Safety Study [5]), the initiating event that can start an accident sequence is a feed

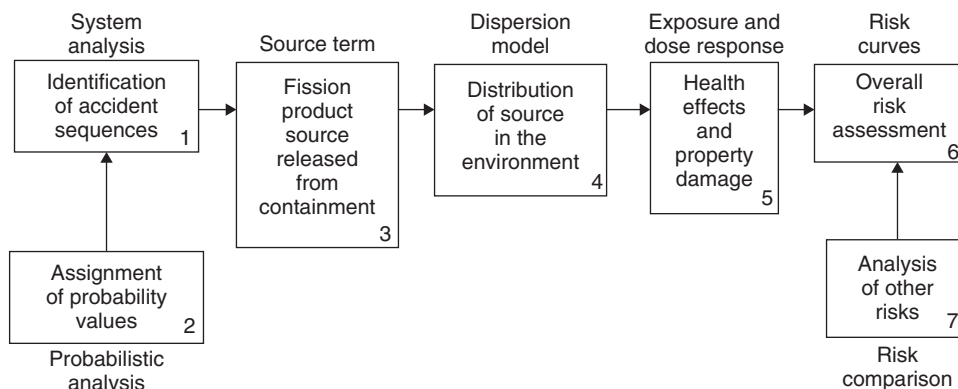


Figure 7. The different models of a nuclear PRA. (Source: USNRC [5].)

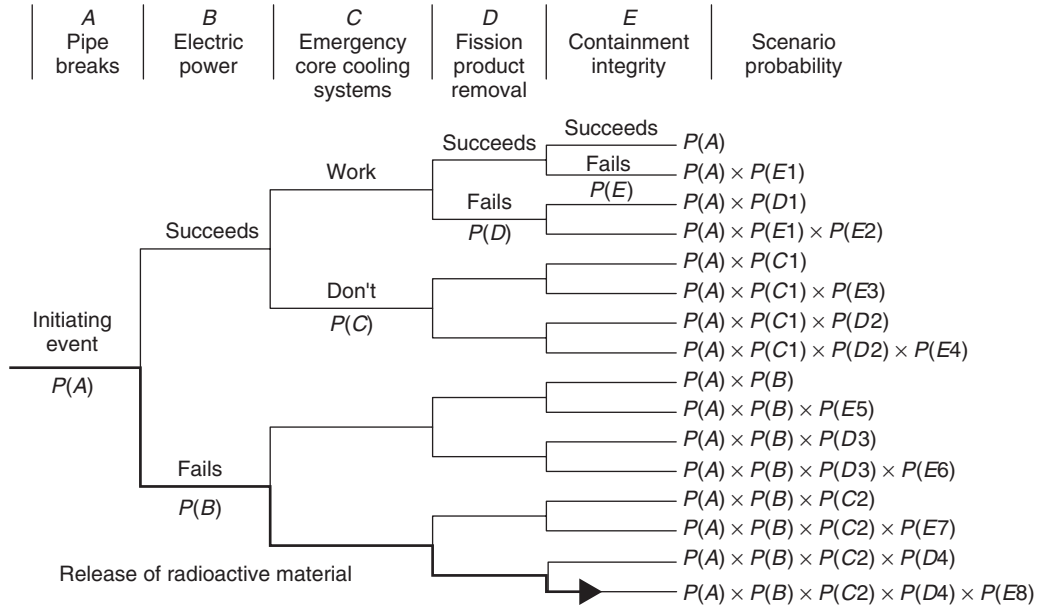


Figure 8. Event tree for steam generator pipe rupture. (Source: USNRC [5].)

water pipe rupture in the steam generator. Steam generators are heat exchangers used in pressurized water reactors to generate electricity. They are at the interface between the primary (radioactive) coolant loop that goes through the nuclear reactor core, and the secondary (nonradioactive) coolant loop that delivers steam to the turbines. The integrity of these pipes is critical to prevent radioactive steam from contaminating the rest of the system. To be able to face such a loss-of-coolant accident, nuclear reactors are equipped with an emergency core cooling system, which requires electricity. If radioactive material is released in the containment structure, an emergency safety feature is in place to remove fission products. Finally, a (secondary) containment structure sits over the whole reactor as a barrier to the release of radioactive material in the atmosphere.

As discussed earlier, the general structure of this model relies first on event trees that collectively represent an exhaustive, mutually exclusive set of failure scenarios following (in this case) an initiating event, then on fault trees, to compute the probabilities of the scenarios that lead to system failure.

Figure 8 represents the event tree that can be used to compute the probabilities of the different scenarios (i.e., the paths from the initiating event to the different outcomes) that can follow a steam-generator pipe rupture (event A). If one of the steam-generator pipes breaks, the regular water flow does not cool the core anymore. The emergency core cooling system then kicks in, provided that electric power is available (electric power failure: event B) and the emergency cooling system itself works (failure: event C). If it does not, radioactive material may be released inside the containment structure, and removed by an emergency fission product removal (failure: event D). If that does not work, the containment structure itself may prevent the radioactive material from contaminating the atmosphere. If that structure does not hold (event E), the radioactive material is released into the environment (failure F). The probability of each scenario is the product of the marginal probability of the initiating event and of the conditional probabilities along the path.

The lowest branch in the tree shown in bold in Fig. 8 involves the loss of containment

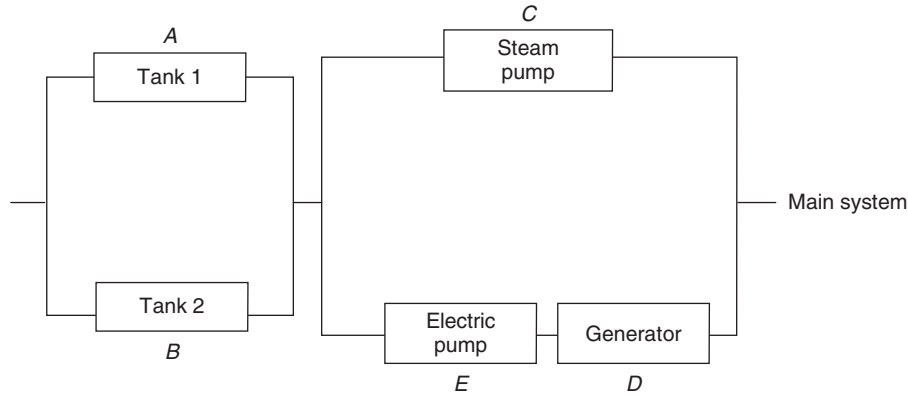


Figure 9. Functional block diagram for an auxiliary feed water system in a nuclear reactor. (Source: Cornell and Newmark [49].)²⁰

integrity. It represents the only scenario that leads to system failure F (release of radioactive material in the atmosphere) following the initiating event “steam-generator pipe rupture” (A). That failure scenario includes the loss of containment after all other emergency systems failed (B , C , and D) following the initiating event A . The failure probability associated with that particular initiating event is that of this last scenario (using Bayes theorem as shown earlier); that is,

$$P(F) = P(A) \times P(B | A) \times P(C | A, B) \\ \times P(D | A, B, C) \times P(E | A, B, C, D).$$

The probability of failure of each subsystem can, in turn, be calculated separately. For example, the probability of failure of the power system can be estimated through a fault tree such as the simplified one that was shown in Fig. 5.

The Case of the Auxiliary Feed Water System (AFWS) of Nuclear Reactors

The AFWS of a nuclear reactor is called upon when the main cooling system is not functioning. The structure of that system is represented by a simplified functional block diagram shown in Fig. 9 as comprising

two subsystems in series: water storage and pumps. In this illustration, the AFWS involves two storage tanks in parallel, and two pumps in parallel, a turbine pump, and an electric pump, which requires energy from a generator.

The fault tree for this simplified AFWS can be represented as shown in Fig. 10 (although this representation is not unique). From that fault tree, one can identify the failure modes (or minimal cut sets); that is, the minimum sets of component failures that would lead to the failure of the AFWS. To do that, one generally uses a Boolean representation of the fault tree. Calling A , B , C , D , E , the Boolean variables that represent failure (or not) of the different components shown in Fig. 9, the

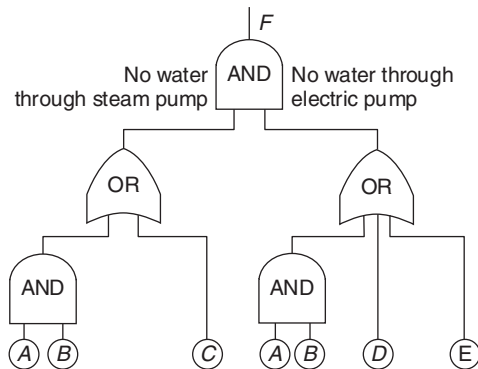


Figure 10. Fault tree for the AFWS shown in Fig. 9. (F : failure of the system).

²⁰Note that in such a diagram, the lines do not represent physical elements such as pipes or wires but the structure of the systems and subsystems.

Boolean expression that links the failure of the system ("top event" F) to those of the basic events is as follows (after application of some Boolean laws of logic):

$$F = (C \text{ AND } D) \text{ OR } (C \text{ AND } E) \\ \text{OR } (A \text{ AND } B).$$

The three failure modes (M_i) are thus C AND D (M_1 : loss of turbine pump and generator), C AND E (M_2 : loss of turbine pump and electric pump), and A AND B (M_3 : loss of both water tanks). The first two failure modes (M_1 AND M_2) imply that both pumps are off, and the last one (M_3) that no water is available.

One can then assess the probabilities of the different failure modes and combine them to compute the overall probability of failure of the AFWS accounting for failure dependencies, such as a failure of the two water tanks.²¹ For example, the probability of M_3 (failure of both water tanks) is

$$P(M_3) = P(A) \times P(B | A) = P(B) \times P(A | B).$$

Only in the case where A and B are independent events, can the probability of M_3 be written as $P(A) \times P(B)$.

Following the total probability theorem described in the section titled "Probability and Sources of Data," the probability of failure of the whole system is

$$P(F) = P(M_1) + P(M_2) + P(M_3) \\ - P(M_1 \text{ AND } M_2) \\ - P(M_1 \text{ AND } M_3) - P(M_2 \text{ AND } M_3) \\ + P(M_1 \text{ AND } M_2 \text{ AND } M_3).$$

External Events

The example of the AFWS has been used in the literature to illustrate the effect of external events such as earthquakes on the performance of a system [49]. External events

must be included in PRA, separating on the one hand the probability that they occur and the severity of an event if it occurs (hazard analysis) and on the other hand, the probability of system failure (fragility) given an event's severity. Hazard analysis thus represents the uncertainty on the loads and fragility analysis, the uncertainty about the capacity, which is the maximum load that the structure can withstand.

Loads and capacities are often measured on a continuous scale (e.g., the peak ground acceleration of an earthquake). Note their density distribution functions $f_L(x)$ and $f_C(x)$ respectively, and F_L and F_C the corresponding cumulative distribution functions, in which the probability of the load is estimated for a given time period (e.g., a year). The probability of failure during that time unit is the probability that the load exceeds the capacity or that capacity is less than the load:

$$\left\{ \begin{array}{l} P_F = \int_X f_L(X) F_C(X) dX \\ \quad = P(L = X \text{ and } C < X) \\ \text{or} \\ P_F = \int_X f_C(X) G_L(X) dX \\ \quad \text{with } G_L(X) = 1 - F_L(X) \\ \quad = P(C = X \text{ and } L > X). \end{array} \right.$$

Knowing the probability distributions of loads and capacities through experience in operations, tests, engineering models or expert opinions, one can compute the overall probability of a system's failure accounting for occurrences of external events, their effects on the components, and the resulting failure dependencies.

An Overarching Analytical Framework for Risk Analysis Based on "Pinch Points"

The different models that yield the probabilities and consequences of different accident scenarios need to be combined to obtain an overall assessment of the failure risk. One way of doing so is to structure the analysis as a product of state vectors and transition matrices in an overarching model as represented in Fig. 11 for the particular case of a nuclear reactor [50].

²¹This means that knowing that one of the tanks has failed, changes the probability that the other one has too.

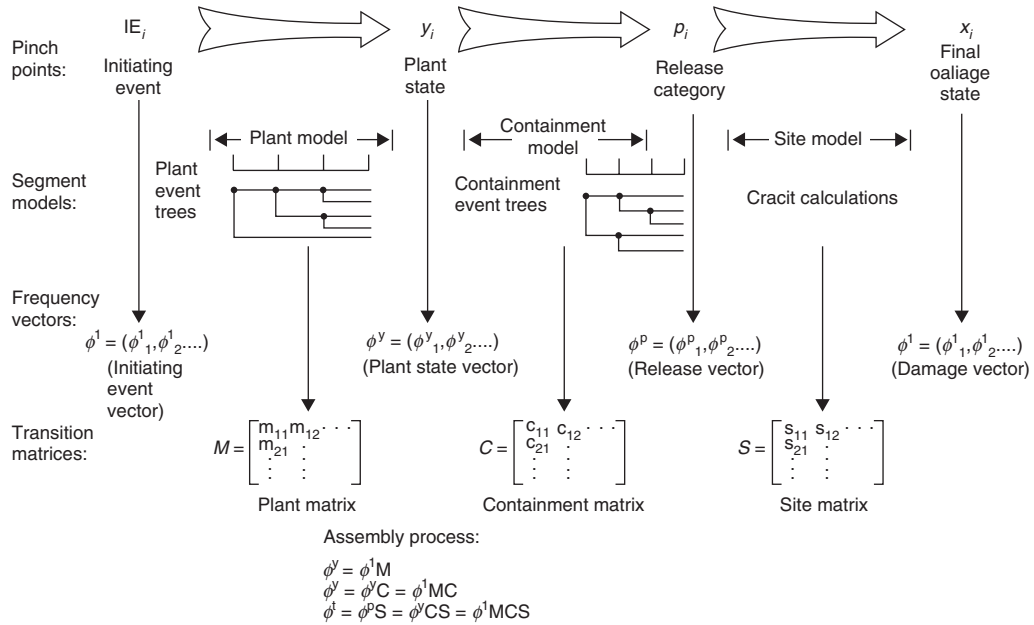


Figure 11. Assembly model for nuclear power plants. (Source: Garrick [50].)

This model can be generalized to most systems that present the risk of releasing a pollutant or toxic substance in the environment (e.g., an oil tanker, or a chemical plant). The computation starts with a vector representing the probability of the different initiating events. The risk is then estimated using the models described earlier. The different steps include a sequence of computations: of the probability that the system/plant/reactor ends in different states given the possibility of different initiating events; of the probability that given the plant state and its containment system is compromised to various degrees; of the resulting distribution of quantities released; and of the probabilities of the final states of the site based on environmental models.

This overarching model can be represented by a sequence of products of vectors and matrices that simply reflects the fact that the probability of reaching a given state at any stage is the sum of the probabilities of starting from any possible state in the previous stage and reaching the state of interest in the next stage. The sequence of transition computations is thus the following.

Notations

- *Initiating Events.* Vector $\{IE_i\}$ is the probability of initiating events indexed in i per time unit.
- *Plant (or Final System) State.* Vector $\{PS_j\}$ represents the probabilities of the final system states; that is, of the different end points of accident sequences indexed in j .
- *Plant Model.* Matrix $[PM_{ij}]$ is the set of probabilities of transition from IE_i to PS_j .
- *Source Term.* Vector $\{R_k\}$ is the probability of the release category (released quantity) indexed in k .
- *Containment Model (for Nuclear Reactors).* Matrix $[CM_{jk}]$ is the probability of transition from plant state j , PS_j , to release level R_k .
- *Damage Vector.* $\{D_l\}$ is the probability of loss (or damage) level l .
- *Site Model.* Matrix $[SM_{kl}]$ is the probability of transition from source term k to damage level l .

Assembly Model

1. From initiating events to final plant states: The nuclear power plant risk analysis model (or PRA model of another system of interest, using fault and event trees or similar techniques) yields the probability that an accident that started with initiating event IE_i ends in the final state PS_j . The probabilities of transition from IE_i to PS_j are the elements PM_{ij} of matrix PM , which represents the plant model. They are computed through combinations of event trees and fault trees as described earlier.

$$\begin{aligned}\{PS\} &= \{\tilde{IE}\}[PM] \Leftrightarrow PS_j \\ &= \sum_i IE_i \times PM_{ij}.\end{aligned}$$

2. From plant states to release category: Next, the containment model allows computation of the probability that a plant state PS_j leads to a release category R_k . The probabilities of transition from PS_j to R_k are the elements CM_{jk} of the matrix CM , which represents the results of the containment model.

$$\begin{aligned}\{R\} &= \{PS\}[CM] \Leftrightarrow R_k \\ &= \sum_j PS_j \times CM_{jk}.\end{aligned}$$

3. From release category to site damage: In turn, the site model allows computation of the probability that a release category R_k leads to a level of damage D_l . The probabilities of transition from R_k to D_l are the elements SM_{kl} of the matrix SM , which represents the site model.

$$\{D\} = \{R\}[SM] \Leftrightarrow D_l = \sum_k R_k \times SM_{kl}$$

4. From initiating event to site damage: These three equations can then be combined to link the distribution of the final damage to the probabilities of the initiating events through the different

submodels: plant (system) model, containment model, and site model.

$$\begin{aligned}\{D\} &= \{\tilde{IE}\} \times [PM] \times [CM] \times [SM] \\ \Leftrightarrow D_l &= \sum_k SM_{kl} \\ &\times \left(\sum_j CM_{jk} \left(\sum_i IE_i \times PM_{ij} \right) \right).\end{aligned}$$

Therefore the risk, that is the probability distribution of the damage per time unit or operation, can be computed starting with the probabilities of the initiating events, using the plant model, the containment model, and the site model.

Pinch Point Concept

An important element of the structure of this assembly model is the notion of “pinch point” [50]. This means that what follows in the model only depends on the state reached so far, not on the scenario by which it was reached (e.g., a particular state of the system, which is all that matters to compute the consequences). Therefore, the downstream computations following a pinch point depend only on the state at that particular stage, which allows multiplication of the transition matrices PM , CM , and SM as independent transition models.

This model has wide applications, for example, to assess the risk of an oil spill caused by loss of ship propulsion followed for instance, by an uncontrolled drift in a shallow zone, so that the hull is breached and a quantity of oil is released into the sea. In the end, it is in large part the size of the breach that determines the amount of oil in the sea, not the events that led to that breach.

INFLUENCE DIAGRAMS AND BAYESIAN NETWORKS AS TOOLS TO STRUCTURE A RISK ANALYSIS MODEL

Influence diagrams can represent on one figure the elements of risk analysis and risk management models, and their relationships. Therefore, they are very convenient

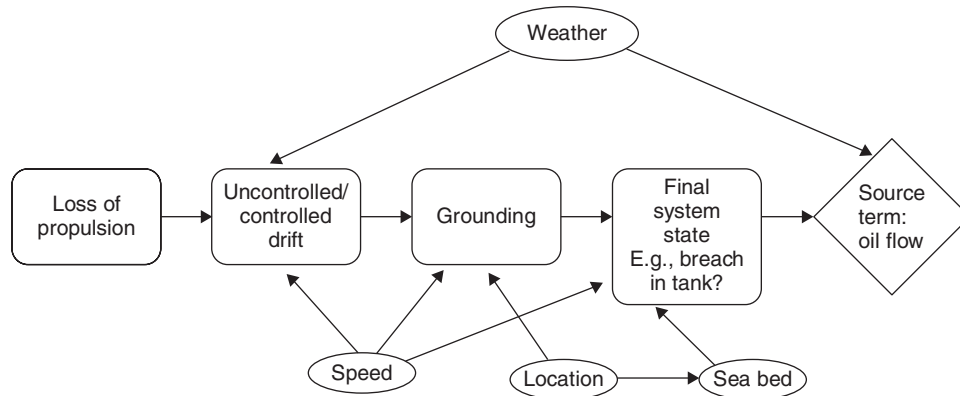


Figure 12. Influence diagram: risk of ship grounding and oil release in the sea. (Source: Paté-Cornell [11].)

tools, especially in the formulation phase of such models. Influence diagrams are directed graphs that are essentially homomorphic to decision trees. They explicitly display the probabilistic links among the different variables (decisions, random events, state variables, and scenario outcomes) of a decision analysis model [51]. These variables are represented by nodes (respectively, rectangles, ovals, and lozenges) and the arrows among them represent the conditional dependencies between the variables that they connect. Influence diagrams include not only a graph, but also the numerical tables that represent the realizations of the different variables, their probability distributions conditional on what precedes them in the diagram, and the values of the possible outcome realizations.

Bayesian networks used in risk analysis can be viewed as influence diagrams that do not include decision nodes but only state and outcome variables. These networks/diagrams are resolved by a computer's "inference engine" based on the laws of Bayesian probability. They provide in the end, a probability distribution of the losses that can be incurred given the probabilities of adverse (initiating) events and the models of the system's evolution thereafter.

Ship Grounding

An example of such diagrams represents the risk, as described earlier, of an oil spill as

a result of ship grounding caused by loss of propulsion, LOP.²² The Bayesian network of Fig. 12 represents the same information as the event tree of Fig. 2 (but includes a few more variables). In the same way, it starts with one initiating event: the LOP of an oil tanker. Given that this event has occurred, the weather, and the speed of the boat, the next question (state variable) is whether or not the crew is able to control the drift. Given the drift control, the speed, and the location, the next question is whether grounding occurs and at what energy level. Given that grounding occurs, the nature of the seabed, and the speed, the next variable is the final system state, measured by the size of the breach in the hull. Finally, the breach size and possibly the weather, determine the amount of oil released into the sea [11].

This model, like the event tree presented earlier, can then be used to assess the effectiveness of risk management measures such as requiring double hulls for some types of ships or improving the response to an incident. The next problem is an optimal allocation of risk management resources, which can be formulated either as the minimization of failure probability given a budget, or minimization of a budget to reach a certain level

²²A subsequent site model could be added to the diagram shown here to assess the damage given the total amount of oil released and the local conditions at the site location.

of safety. These computations require data such as risk/cost functions, linking the reduction of the probability of component failure to the costs involved.

Influence diagrams and Bayesian networks are useful as computational tools, but also an effective means of communication, first with the experts at the onset of the analysis to structure the problem and in the end, with the decision makers, to communicate the results and explain how they were obtained.

The Space Shuttle Heat Shield

In the space shuttle study, an influence diagram was also key to the problem formulation. The objective was to compute the contribution of the black tiles that protect the underside of the orbiters against heat loads at reentry [17]. There are about 25,000 tiles on each orbiter, each different and subjected to different loads according to its location. Each tile is bonded to a felt pad, itself glued to the aluminum surface. The concern is that a tile can debond in flight, creating a gap in the heat shield, and that the turbulence in the tile cavity at reentry could cause adjacent tiles to debond, melting the aluminum and causing loss of critical system functions. After each flight, the tiles are inspected and replaced if necessary; this requires careful cleaning of the cavity and positioning of the tile.

There are two main causes of debonding of the tiles: under normal loads (vibrations, heat, aerodynamic forces, etc.) because

of a weak bond, or under the impact of debris either external (e.g., micrometeorites) or internal to the system (e.g., pieces of insulation of the external tank).

Figure 13 represents the influence diagram showing the structure of the PRA model that was developed for this case. Following one of two possible initiating events (debonding under the impact of debris, or for other reasons such as a weak bond under normal loads), a first tile may be lost. After that, depending on the aerodynamic forces and the heat at reentry at its location, additional tiles may be lost. Under the heat loads at reentry in the atmosphere, a hole may appear in the aluminum structure ("burnthrough"). Hot gases then penetrate the structure causing a subsystem malfunction, which in turn, can cause the loss of the shuttle (orbiter and crew).

The probabilistic computations involved the values of four parameters that varied depending on the location on the surface: heat load, aerodynamic forces, density of debris hits, and criticality of the subsystems under the orbiter skin. The surface was thus partitioned into 33 zones that were characterized by these parameter values as a function of the location. The results were that the tiles contributed to about 10% of shuttle mission failure risks. They also involved a map of the orbiter showing the criticality of the tiles in different locations. It turned out that 15% of the tiles were the source of about 80% of the risk. The challenge was in the formulation of the model using both, the influence

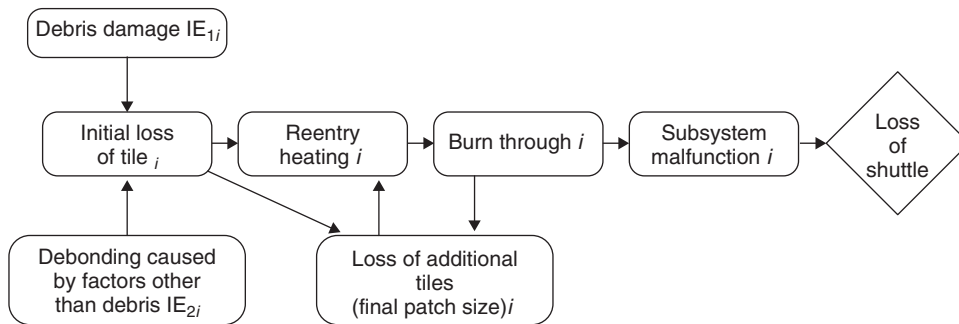


Figure 13. Influence diagram for a risk analysis of the loss of a space shuttle due to tile failure. i : index of min-zones, that is, tile location on the orbiter surface. (Source: Paté-Cornell and Fischbeck [17].)

diagram shown in Fig. 13 and the partition of the orbiter surface into zones of similar characteristics.

EXTENSIONS OF THE CLASSIC PRA MODEL

The PRA models as described above can be extended in several ways to better represent the realities of a system's behaviors for example, the dynamics of accident sequences and system deterioration, or the effects of human and organizational factors. The method can also be extended to other non-technical problems (e.g., an analysis of the risk posed by some human groups) that can be addressed through systems analysis and probability. Again, this systems analysis approach is most useful when relevant statistics do not exist because they involve events or conjunctions of events that are rare or have not occurred before, and when the problem is too complex to be processed directly by the human mind.

Evolution and Deterioration

The evolution (generally deterioration) of the components of a technical system can be represented by different stochastic processes, a simple one being a Markov model, which assumes that there is no memory in state transition and that the time to transition to another state is exponentially distributed. Consider for example, the case of imperfect inspections, after which there is a nonzero probability that the system is in an undetected deteriorated state. The inspection can be described by the distribution of the possible states after such an inspection. One can represent the deterioration of the system by the probability of transition from one state of deterioration to the next, then to failure. The results can then be used to assess the distribution of the time to failure following an inspection, and to decide on an optimal maintenance schedule.

The Dynamics of Accident Sequences

Assessing the dynamics of an accident sequence following an initiating event involves computation of the probability (per

time unit) of transition from each event to the next in each accident sequence. An example of such a dynamic risk analysis model was a study of patient risk under anesthesia [52]. Time is a critical element because the final state of the patient depends on the duration of oxygen deprivation. Following an initiating event, the subsequent steps involve observation of a signal that the event has occurred (e.g., patient reaction); detection, diagnosis, and remediation of the problem; and outcome to the patient (recovery, brain damage or death). For example, following a disconnection of the breathing tube, the oxygen no longer flows and the first signal observed might be that the patient turns blue. The attendants have to observe the signals and the disconnection, then reconnect the tube. The outcome of such an incident is determined by the total time that elapses between the initiating event and the end of the episode.

In this study, a Markov model was used to represent the dynamics of accidents. Partial statistics (e.g., about initiating events and overall mortality rate) allowed calibration of the model. The results included first, the risk of an anesthesia accident per initiating event; second, how this risk could be linked to the "state of the anesthesiologist" characterized in terms of competence and alertness; and third, how a change of management practice (e.g., mandating periodic recertification) affecting the performance factors could reduce the patient risk. It was found for instance, that periodic recertification of practitioners and improved supervision of residents could significantly reduce the risk to the patient.

This dynamic model can be generalized to represent different evolution paths of an accident sequence, the deterioration of various components, and the corresponding changes of the failure probability of the whole system. The results can be used to decide for example, an inspection and maintenance schedule.

Human and Organizational Factors

Human errors are a common cause of system failure, some rooted in personal behavioral problems such as distraction, some in management issues such as an

inappropriate incentive system [53]. Human reliability analysis has been the subject of a whole field of research focusing on both, the causes of human errors and on best practices to eliminate the risks that they pose. In the nuclear power industry for example, Swain and Guttman [54] estimate the rates of human errors that are part of failure modes in the Handbook of Human Reliability using the THERP method (Technique for Human Error Rate Prediction). Kolaczowski *et al.* [55], show for example, the influence of organizations on performance-shaping factors, error mechanisms, actions, human failures, and “unacceptable outcomes.” They use as an example, the 1982 Air Florida crash in Washington DC. In the nuclear power industry, although the two events were of a very different nature, they show common cognitive problems between the Three Mile Island and Chernobyl nuclear accidents. They point in particular, to the fact that in both cases, the lack of operators’ understanding of the basic physics involved led them to ignore instrument readings and field reports. Gertman [56] focuses on performance-shaping factors in the nuclear power industry based on the standardized plant analysis risk human reliability analysis (SPARH), which is designed to encompass all factors that may influence human performance. Bley [57] links the occurrence of human errors to the context of human operations, and the challenges that it sometimes presents, as critical to the probability of an error.

The goal of introducing human and management factors in PRA is to identify and assess possible risk mitigation measures that are human in nature—such as a change in the incentive structure for the operators—as opposed to technical modifications such as adding redundancies to the system. The link between human performance and system reliability can be modeled in several ways. One is to relate the system’s failure risk to the management factors that influence operators’ behaviors. This requires linking the technical elements of the system’s risk analysis (submodel S) to the immediate decisions and actions of the operators (or pilots, anesthesiologists etc.) (submodel

A), and in turn, to link these actions to management decisions (submodel M) in what we call the SAM model [58].

Figure 14 represents the integration of these three levels into a SAM model for the simplified model of ship grounding presented earlier. The lower part (Level 1: starting point of the model) is the technical risk analysis shown in Fig. 12. The intermediate level represents the actions of the maintenance crews as well as those of the ship’s captain and crew in case of a propulsion problem. They include human errors but also, regular and exceptional decisions that may save the system in difficult circumstances. This second level is represented here as a model of random variables, whose realizations and distributions are estimated from the point of view of the decision makers, that is the managers.²³ The third level represents the management decisions that affect the actions of operators at the intermediate level. Organizations provide the structure, procedures, and culture within which employees operate and make decisions. Managers decide how to allocate their resources (e.g., how much should be dedicated to maintenance). In general, they select and train employees, set incentives, and provide information to operators.²⁴

The SAM model has been used for instance, to examine the effect of some organizational factors on the risk of failure of jacket-type offshore oil platforms [60]. One question was whether or not to require an external design review, given the structure of the risk analysis model, the nature of the current design review process, and the nature of the human errors that can lead to failures (e.g., gross factual errors versus errors of judgment, and their severity level). A computation of the benefits of doing so showed that

²³Note that the realizations of these random variables can be represented in a simple way as “low, medium or high,” or by more sophisticated measures involving for example, the individual experience level of the captain and the crew, and the time they have worked together.

²⁴To link the decisions to the response of the operators, one can use a variety of behavioral models, or an economic model such as a “principal-agent” model [59].

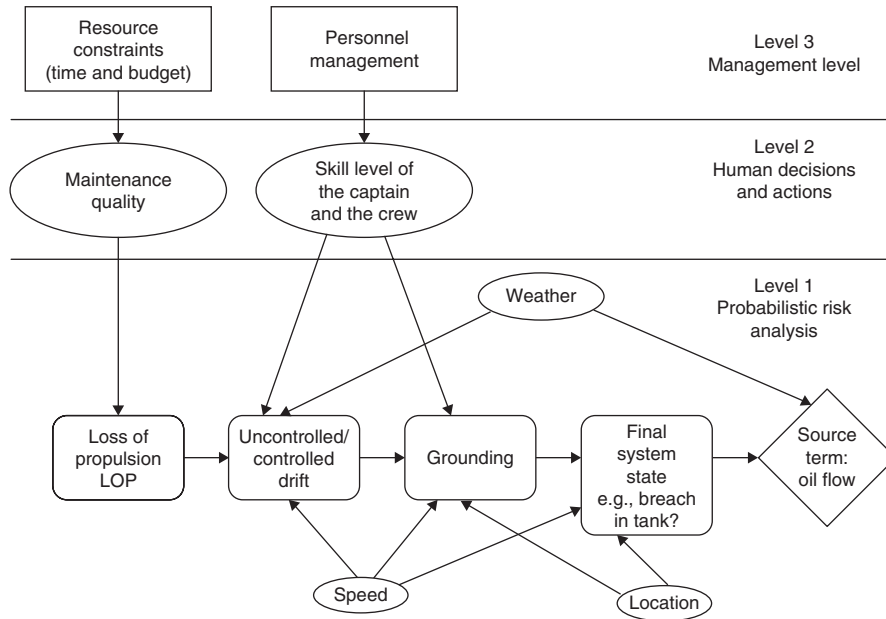


Figure 14. Influence diagram showing the structure of the SAM model (system, action, management) for the case of ship grounding presented in Fig. 12.

their costs were inferior to those of adding steel to the structure to achieve the same risk reduction benefits.

Another form of the SAM model was used in conjunction with the risk analysis model described earlier in Fig. 13 to link potential failures of the space shuttle heat shield to decisions and actions of the technicians during tile maintenance. These actions were linked in turn, to management decisions such as time pressures, the space center structure (thus the flow of information), and the incentives to remain with one's job which is provided by NASA's compensation structure [18]. The study showed, for example, the benefits of improving the inspection process by focusing first on the most risk-critical tiles.

Another form of such models designed to link technical failures and human factors is to start with the work process (instead of the system). For instance, the work process analysis model (WPAM) is based directly on a detailed study of the work process and links its elements to the performance of the system's components [61]. Similarly, Galan *et al.* [62] use Bayesian networks to

incorporate organizational factors in PRAs for nuclear power plants.

Extensions to Human Groups, Game Theory, and Counterterrorism

The structure of technical PRA models can also be useful to address problems that require system's analysis and quantification of uncertainties when the system's elements are human groups such as terrorist organizations. The methods can then be used in a game analysis setting, with explicit representation of the adversary's objectives, information, and means [23]—all factors that may not be fully known to the main player.

In one study, systems analysis, probability, and bounded-rationality concepts were used to rank the risks of different scenarios of terrorist attack on the United States [63]. In another application, the simulation of a game between a government and a group of insurgents permitted an assessment of various government strategies [24,64]. The application in that case was an insurrection in the Philippines. It required another way

to represent the failure risk attached to a specified policy as the probability that after a specified number of time periods, the political and economic stability of the country remained below a given threshold. That model allowed assessing various approaches to balancing immediate resource spending to avoid attacks, and long-term expenditures to address the fundamental problems that fuel the insurrection in the first place. In both cases—counterterrorism and counterinsurrection models—the problem was represented by influence diagrams showing the decision process for the two sides involved, and linked by the results of their alternate moves.

The structure of these models and the probabilistic analysis involved are similar to those used to assess the performance of technical systems: identification of the key state variables, of their realizations and distributions, of the failure scenarios, and of the probability distribution of the spectrum of possible outcomes, positive or negative.

Obviously, the PRA methods have limitations [65]. For instance, one can never be sure that all possible failure scenarios have been considered. Indeed, some are difficult to imagine; that is a problem of implementation, specific to each case. More fundamentally, explicit representation of uncertainty is not part of the engineering tradition—and to a large extent, engineering education—in which a deterministic description of risks combined with safety factors is often preferred to probability.²⁵ The problem of course, is that safety factors can be costly, economically inefficient, and inadequate if one does not consider the chances that even with such factors, extreme loads can be greater than the capacities provided, given the safety margins. Another argument that is sometimes advanced is that a risk analysis cannot be done in the absence of sufficient statistical data. In reality, one often performs a PRA of the type described above precisely because there is not sufficient statistical information

at the system level. In that case, a system analysis combined with probability allows addressing problems such as rare events or events that have not yet been experienced. It is true in engineering and in medicine, where new medical devices must be tested in patients in order to gather the required statistics [19]. It is also true in finance where the chances of “perfect storms” cannot be computed directly from the performances of the last decades because circumstances have changed. In many of these cases, the probabilities used in the analysis have to come from expert opinions. These represent the best available information, even though they may reflect some biases that have been described in the literature [67].

CONCLUSION

PRA methods have been developed mostly in engineering and in particular, in the nuclear power industry, to compute the risks of failure of systems for which sufficient statistical data sets do not exist at the system level. They permit prioritization of risk management measures and allocation of scarce risk management resources. They also allow checking of whether the final risk is tolerable. The methods can be extended to involve human and organizational factors and to represent other types of systems including those in which human groups play a role. The main challenge of these methods is in the problem formulation and in the gathering of data. The approaches presented here allow addressing both to the best of the analyst’s ability. The most important feature of these PRA models is that they explicitly represent uncertainties in a systematic way and can support rational decisions, both in government and industry.

REFERENCES

1. von Mises R. Probability, statistics and truth. 2nd rev. English ed. New York: Dover; 1981.
2. Savage LJ. The foundations of statistics. New York: John Wiley & Sons; 1954.
3. de Finetti B. Theory of probability. New York: John Wiley & Sons; 1974.

²⁵Quasi-deterministic descriptions of risks include for example, maximum probable floods and maximum credible earthquakes.

4. Press SJ. Bayesian statistics: principles, models, and applications. New York: John Wiley & Sons; 1989.
5. US Nuclear Regulatory Commission (USNRC). Reactor safety study: assessment of accident risk in US Commercial Nuclear Plants. WASH 1400 (NUREG-75/014). Washington (DC): USNRC; 1975.
6. Kaplan S, Garrick BJ. On the quantitative definition of risk. *Risk Anal* 1981;1(1):11–27.
7. Lave LB. Risk assessment and management. New York: Plenum Press; 1987.
8. Henley E, Kumamoto H. Probabilistic risk assessment: reliability engineering, design, and analysis. New York: IEEE Press; 1992.
9. Bedford T, Cooke R. Probabilistic risk analysis: foundations and methods. Cambridge (UK): Cambridge University Press; 2001.
10. Paté-Cornell ME. Greed and ignorance: motivations and illustrations of the quantification of major risks. Proceedings of the study week on “Science for Survival and Sustainable Development”: Pontificiae Academiae Scientiarum Scripta Varia (Report of the Pontifical Academy of Sciences); The Vatican. 2000. pp. 231–270.
11. Paté-Cornell ME. The engineering risk analysis method and some applications. In: Edwards M, von Winterfeldt D, editors. *Advances in decision analysis*. Cambridge (UK): Cambridge University Press; 2007.
12. Haimes YY. Risk modeling, assessment, and management. 3rd ed. New York: John Wiley & Sons; 2008.
13. Fullwood RR. Probabilistic safety assessment in the chemical and nuclear industries. Boston (MA): Butterworth and Heinemann Publishers; 2000.
14. Garrick BJ. The approach to risk analysis in three industries: nuclear power, space systems, and chemical processes. *Reliab Eng Syst Saf* 1988;23(3):195–205.
15. Haasl DF. Advanced concepts in fault tree analysis. Proceedings of System Safety Symposium; Seattle. 1965.
16. Cornell CA, Newmark NM. On the seismic reliability of nuclear power plants. Invited paper, Proceedings of ANS Topical Meeting on Probabilistic Reactor Safety; Newport beach, California: 1974. 10–14.
17. Paté-Cornell ME, Fischbeck PS. Probabilistic risk analysis and risk-based priority scale for the tiles of the space shuttle. *Reliab Eng Syst Saf* 1993;40(3):221–238.
18. Paté-Cornell ME, Fischbeck PS. PRA as a management tool: organizational factors and risk-based priorities for the maintenance of the tiles of the space shuttle orbiter. *Reliab Eng Syst Saf* 1993;40(3):239–257.
19. Pietzsch JB, Paté-Cornell ME. Early technology assessment of new medical devices. *Int J Technol Assess Health Care* 2008; 24(1):37–45.
20. Frank M. Choosing safety: a guide to using probabilistic risk assessment and decision analysis in complex high-consequence systems. London (UK): Earthscan; 2008.
21. Paté-Cornell ME. Finding and fixing systems weaknesses: probabilistic methods and applications of engineering risk analysis. *Risk Anal* 2002;22(2):319–334.
22. Cox LA. Risk analysis: foundations, models and methods. New York: Springer; 2002.
23. Bier VM. Game-theoretic and reliability methods in counter-terrorism and security. In: Wilson A. *Modern statistical and mathematical methods in reliability*. Series on quality, reliability and engineering statistics. London (UK): World Scientific Publishing Co.; 2005.
24. Paté-Cornell ME. Games and risk analysis: three examples of single and alternate moves. In: Bier V, Azaiez M, editors. *Game theoretic risk analysis of security threats*. New York: Springer; 2008.
25. Starr C. Social benefit versus technological risk. *Science* 1969;165:1232–1239.
26. Fischhoff B, Lichtenstein S, Slovic P, *et al.* *Acceptable risk*. New York: Cambridge University Press; 1981.
27. Paté-Cornell ME. Acceptable decision processes and acceptable risks in public sector regulations. *IEEE Trans Syst Man Cybern* 1983;SMC-13(3):113–124.
28. Paté-Cornell ME. Quantitative safety goals for risk management of industrial facilities. *Struc Saf* 1994;13(3):145–157.
29. Paté-Cornell ME. Uncertainties in risk analysis: six levels of treatment. *Reliab Eng Syst Saf* 1996;54:95–111.
30. Raiffa H. *Decision analysis*. Cambridge (MA): Addison Wesley; 1968.
31. von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton (NJ): Princeton University Press; 1947.
32. Keeney RL, Raiffa H. *Decision analysis with multiple objectives: preferences and value trade-offs*. New York: John Wiley & Sons; 1976.

33. Farmer FR. Reactor safety and sitting: a proposed risk criterion. *Nucl Saf* 1967; 8(6):539–548.
34. Phimister JR, Bier VM, Kunreuther HC, editors. Accident precursor analysis and management: reducing technological risk through diligence. Washington (DC): National Academies Press; 2004.
35. Paté-Cornell ME, Lakats LM, Murphy DM, *et al.* Anesthesia patient risk: a quantitative approach to organizational factors and risk management options. *Risk Anal* 1996; 17(4):511–523.
36. Atwood CL, LaChance JL, Martz HF, *et al.* Handbook of parameter estimation for probabilistic risk assessment. NUREG/CR-6823. Washington (DC): US Nuclear Regulatory Commission; 2003.
37. Apostolakis G. The concept of probability in safety assessments of technological systems. *Science* 1990;250:1359–1364.
38. Helton JC. Treatment of uncertainty in performance assessments for complex systems. *Risk Anal* 1994;14:483–511.
39. Benjamin JR, Cornell CA. Probability, statistics and decision for civil engineers. New York: McGraw Hill; 1970.
40. Hogarth RM. Cognitive processes and the assessment of subjective probability distributions. *J Am Stat Assoc* 1975;70(350):271–289.
41. Kahneman D, Slovic P, Tversky A, editors. Judgment under uncertainty: heuristics and biases. New York: Cambridge University Press; 1982.
42. Morris PA. Combining expert judgment: a Bayesian approach. *Manage Sci* 1977;23(7): 679–693.
43. Winkler RL. Expert resolution *Manage Sci* 1986;32(3):298–303.
44. Dalkey NC. Delphi, The RAND Corporation, P-30704, Santa Monica (CA). 1967.
45. Budnitz RJ, Apostolakis G, Boore DM, *et al.* Use of technical expert panels: applications to probabilistic seismic hazard analysis. *Risk Anal* 1998;18(4):463–469.
46. Paté-Cornell ME. Fault trees versus event trees in reliability analysis. *Risk Anal* 1984;4(3):177–186.
47. Ellsberg D. Risk, ambiguity, and the Savage axioms. *Q J Econ* 1961;75(4):643–669.
48. Paté-Cornell ME. Conditional uncertainty analysis and implications for decision making: the case of the waste isolation pilot plant. *Risk Anal* 1999;19(5):995–1002.
49. Cornell CA, Newmark NM. On the seismic reliability of nuclear power plants. Proceedings of ANS Topical Meeting on Probabilistic Reactor Safety; 1978 May 8–10; Newport Beach, CA. 1978.
50. Garrick BJ. Recent case studies and advancements in probabilistic risk assessment. *Risk Anal* 1984;4(4):267–279.
51. Shachter RD. Probabilistic inference and influence diagrams. *Oper Res* 1988;36:871–882.
52. Paté-Cornell ME. Medical application of engineering risk analysis and anesthesia patient risk illustration. *Am J Ther* 1999; 6(5):245–255.
53. Reason JT. Human error. Cambridge, UK: Cambridge University Press; 1990.
54. Swain AD, Guttman HE. Handbook of human reliability analysis with emphasis on nuclear power applications. NUREG/CR-1278. US Nuclear Regulatory Commission; 1983.
55. Kolaczowski A, *et al.* Good practices for implementing human reliability analysis (HRA). NUREG-1792. US Nuclear Regulatory Commission; 2005.
56. Gertman D. The SPAR-H human reliability analysis method. NUREG/CR-6883. US Nuclear Regulatory Commission; 2005.
57. Bley D. New methods for human reliability analysis. *Environ Manage Health* 2002;13(3):277–289.
58. Murphy DM, Paté-Cornell ME. The SAM framework: a systems analysis approach to modeling the effects of management on human behavior in risk analysis. *Risk Anal* 1996;16(4):501–515.
59. Milgrom PR, Roberts J. Economics, organization and management. New Jersey: Prentice-Hall; 1992.
60. Paté-Cornell ME. Organizational aspects of engineering system safety: the case of offshore platforms. *Science* 1990;250:1210–1217.
61. Davoudian K, Wu J, Apostolakis G. Incorporating organizational factors into risk assessment through the analysis of work processes. *Reliab Eng Syst Saf* 1994;45:85–105.
62. Galan SF, Mosleh A, Izquierdo JM. Incorporating organizational factors into probabilistic safety assessment of nuclear power plants through canonical probabilistic models. *Reliab Eng Syst Saf* 2007;92:1131–1138.
63. Paté-Cornell ME, Guikema SD. Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures. *Mil Oper Res* 2002;7(4):5–23.

64. Kucik P. Probabilistic modeling of insurgency. Doctoral thesis, Department of management Science and Engineering. Stanford University; 2007.
65. Bier VM. Challenges to the acceptance of probabilistic risk analysis. *Risk Anal* 1999;19:703–710.
66. Pate-Cornell ME, Dillon R. Probabilistic risk analysis for the NASA space shuttle: a brief history and current work. *Reliab Eng Syst Saf* 2001;74(3):345–352.
67. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science* 1974;185:1124–1131.

AN INTRODUCTION TO R&D PORTFOLIO DECISION ANALYSIS

KELLY J. DUNCAN
Altria, Richmond, Virginia

JASON R.W. MERRICK
Statistical Sciences and Operations
Research, Virginia Commonwealth
University, Richmond, Virginia

A portfolio is a purposeful combination of items. For research and development (R&D) portfolios, these items are technologies, projects, or products. Companies have widely varying practices for portfolio selection. This article examines existing literature to determine the key characteristics of a good portfolio and how decision analysis is used to find one. The approach needs to handle multiple objectives, account for project interactions, and address the social aspect of decision making. The resulting portfolio should be aligned with business strategy, balanced, and maximize value. The article introduces general concepts that have been used to select portfolios and reviews specific applications.

HOW DO COMPANIES DECIDE ON R&D PORTFOLIOS?

A survey of 205 businesses shows that the techniques for portfolio management are inconsistent even within industries or groups of successful companies [1]. The survey asked company executives to identify all methods they used as part of their portfolio management strategy. The executives then identified the dominant strategy among the ones they used. Financial methods ranked as the most popular primary technique. These methods frequently include net present value (NPV) analysis for selecting projects. Project selection based on business strategy was also popular. This method allocates a percentage of the available budget to different strategies or divisions. Projects are then added into

the pipeline in these areas until all funding is allocated. Scoring models were next in popularity and establish weights and metrics for various attributes of a project. Scoring methods align expenditures with business strategy but are more cumbersome and less user-friendly than the graphical methods of bubble diagrams and portfolio mapping. The graphical methods are next in popularity. These methods typically plot potential projects on a graph of risk versus reward, although other measures can be used on the axes. The graphical methods are easy to read and tend to produce portfolios that are well-balanced but not necessarily strategically aligned with business objectives. The bubble chart is a popular graphical method [1]. Bubble charts allow executives and decision makers to visualize the entire portfolio from a number of perspectives. The visual representation could look at projects based on the decision maker's preference. Two examples of representations are distribution of projects based either on risk or on launch horizon (near or long term). A simple checklist was the least popular and least effective method identified. In this technique, projects that satisfied a given number of questions made the cut into the portfolio.

When evaluating the characteristics of companies that ranked near the top in R&D, Cooper *et al.* [1] found that the most successful companies relied least on financial methods. The top companies used methods that were understood by the senior management, perceived to be effective, and used in making Go/Kill decisions. The top firms in *Smart Organizations* use metrics to ensure that projects aligned with corporate strategy [2]. They are also able to show what creates value for the company and encourage development of projects that increase the value. Human judgment tops the list of the many techniques for portfolio management reported in a survey of pharmaceutical companies [3]. In the survey, 60% of companies report satisfaction with their current portfolio management strategy. In many of

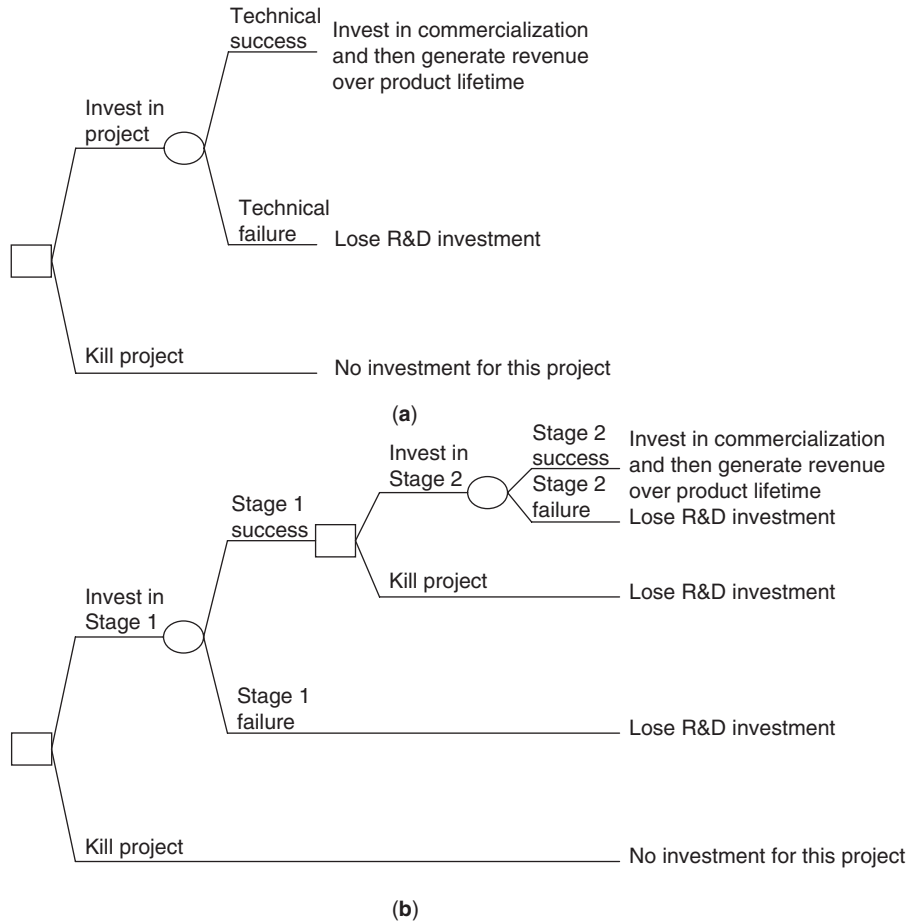


Figure 1. A decision tree for (a) a simple R&D investment decision (adapted from Matheson and Matheson [2]), and (b) a two-stage R&D investment decision.

these companies, transparency of information for decision making is contributing to their satisfaction. Companies that were dissatisfied based their views on inability to gain consensus and focus on individual projects instead of overall portfolio [3].

Decision analysis is not the most common portfolio selection method. However, it does offer significant advantages over the methods commonly used in practice. In the simplest form, Fig. 1a shows a decision tree for evaluating the value of a given R&D project (adapted from Matheson and Matheson [2]). At an individual project level, the decision is whether to continue investing or to kill the project. If the investment is made, the

decision maker must assess the probability of technical success and then the NPV of revenue generated, if successful. Figure 1b shows a two-stage investment decision with an investment or kill decision and a probability of technical success at each stage. However, each of these simple R&D decisions is made at the project level. The broader question is which set of projects to invest in given a budget constraint. In essence, this becomes a problem of maximizing the value of a set of R&D projects, or an R&D portfolio, given resource constraints. This form of decision was used as early as Balthasar *et al.* [4] and more recently in Sharpe and Keelin [5]. Clearly, a key part of such an

approach is the assessment of probabilities, a topic covered in detail in article titled *Eliciting Subjective Probabilities from Individuals and Reducing Biases* in this encyclopedia.

In this article, we first discuss the complexities of portfolio selection decisions and define five criteria for a good method for making such decisions. We then discuss portfolio selection applications and methods that do use decision analysis and discuss which best meet the five criteria.

THE PROBLEM AND CHALLENGES

The problem of portfolio selection poses a set of unique challenges. Undertaking new projects or products requires accepting some level of risk and addressing the uncertainty of both the technical and market success of the project. Decision makers frequently face the task of balancing benefits against costs and risk of realizing the benefits. Phillips and Bana e Costa [3] identified five challenges specific to the R&D portfolio problem.

1. Benefits are typically characterized by multiple and possibly conflicting objectives.
2. When a large number of alternatives are presented, the decision maker cannot know the details of each well enough to make an informed decision.
3. If resources are allocated to several organizational units based on individual needs, the result is rarely an optimal allocation for the overall organization. This problem is a situation that illustrates the "Commons Dilemma".
4. Many people are generally involved. People providing advice or expert opinions can end up competing against each other. Besides that, it is difficult to identify all the people with the power to interfere with or influence the decision.
5. Implementation by people who do not agree with the resource allocation can lead to small groups of people working on unapproved projects.

Chien and Sainfort [6] described two additional complications associated with portfolio selection. First, decision makers face the challenge of measuring preference for the portfolio as a whole against the preference for specific items in a portfolio. The objectives of a portfolio could include measures such as achieving optimal balance among projects, whereas objectives for an individual project could include different types of measures such as maximizing technical merit. Second, items in the portfolio often have interrelations. According to Phillips and Bana e Costa [3], these problems demonstrate the need for an approach that balances the costs, benefits, and risks and takes into account differing perspectives of the people involved. This objective cannot be accomplished solely with a technical solution. A social process to engage the involved parties is also required. Top performing companies maintain portfolios that are aligned with their strategies and objectives, of high value, and balanced.

This article reviews various decision analysis approaches to addressing some of the areas key to successful portfolio strategy. Applications are evaluated against the following criteria for a good portfolio selection method:

- alignment with strategy
- balance within the portfolio
- interrelationship between items in a portfolio
- maximizing value of the portfolio
- social acceptance (including transparency and gaining consensus), and
- handling of multiple and conflicting objectives.

SELECTING A PORTFOLIO USING DECISION TREES

Jackson *et al.* [7] designed a portfolio of remediation measure at nuclear waste storage sites. The problem consists of a complex set of sequential decisions involving interdependent technologies and uncertainties in cost and time. Over a 75-year period, the Department of Energy (DOE) plans to

remediate landfills throughout the United States and Puerto Rico at significant expense. There are seven technology process steps associated with stabilizing a landfill: (i) characterization and assessment, (ii) stabilization, (iii) retrieval, (iv) treatment, (v) containment, (vi) disposal, and (vii) monitoring. Several technology options exist to address each of these processes. Technologies under consideration range from proven technologies to prototypes still under laboratory investigation. Risk comes from the maturity of a given technology, the ability to characterize and assess a waste site with accuracy, and applying the correct technologies to a given site. To incorporate risk into the tool described, one must clearly define the risk. Jackson *et al.* [7] described the development of a formal decision analysis tool to support the decision maker when selecting remediation technologies. Known life cycle cost (LCC) simulation models within the DOE provide inputs to the tool. Decision analysis techniques combine output from LCC tools with information about technology risk and uncertainty in cost and time.

A senior DOE official defined the appropriate criteria for this model using value-focused thinking. As a result, the decisions focus on risks for cost, time, and safety. The uncertainty and trade-offs between cost and time make utility functions a good fit for this application. The decision analysis tool proposed by the authors uses sequential remediation decisions to determine the total time required for a project. A distribution of the present value of the portfolio cost is produced. Constraints are added to ensure compatibility of projects and adherence to timelines and budgetary requirements. An additive utility function describes the decision maker's preference and utility for time and cost.

Jackson *et al.* created an influence diagram for each process where a technology selection is required. The uncertain events in this model are R&D costs, operations and maintenance (O&M) costs, R&D time, and O&M time. Parameters for the probability distributions in the uncertainty nodes come from estimation of distribution parameters from the LCC model. A selected technology

has a chance of failure that would lead to additional time and costs. The probability of failure of a project contributes additional penalty time and cost to the expected values for a technology. The decision makers use the diagrams to visualize and validate the process. A complete model of the decision combines the seven processes described by the influence diagrams into a decision tree. A partial decision tree is shown in Fig. 2. The decision tree shows the sequential nature of the remediation process. In addition to choosing whether to stabilize and whether to treat or contain, the decision maker selects from several available technologies for each process step.

The model accounts for the attributes of cost and time and the category of each technology. Categories ensure technologies in a portfolio are compatible. These constraints can model several types of technology relationships based on Boolean logic and are similar to the approach employed in multicriteria programming approaches. The total cost and time values constrain the model. A user can use a constraint to penalize any portfolio that exceeds allowed time or budget by assigning. Assessing a high penalty could completely exclude an undesirable portfolio from consideration. Since time and cost uncertainties exist within a portfolio, a portfolio could have a nonzero probability of exceeding either time or budget constraints. The user can penalize a portfolio more as the probability of the portfolio exceeding the limits increases. For example, in a portfolio with a 0.10 chance of exceeding the limits, the user could assign a utility of -0.5 to the leaves of the decision tree that exceed a constraint. The expected utility function would then account for the possibility of exceeding the limits. The decision tree model produces time and expected NPV for cost for each leaf on the decision tree. A utility function for these attributes takes into account the decision maker's preferences as a basis for selecting technologies. Jackson *et al.* developed a general utility function based on information from the DOE. The DOE has a high utility for costs and time that are below the target plus a 10% error and a very low utility for costs and time that exceed the target values. Using lotteries, decision makers

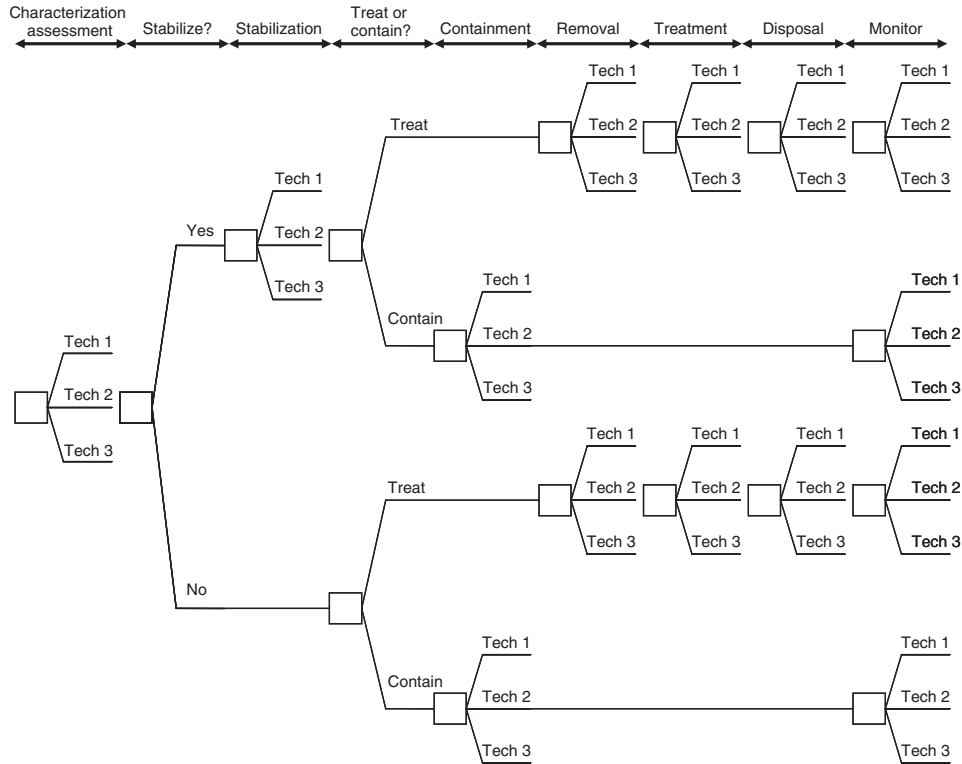


Figure 2. A partial decision tree of the landfill remediation decision (adapted from Jackson *et al.* [7]).

determined the midpoint utilities. From the known points, two exponential utility curves were created. One curve for cost and time is less than 10% above target and the other for cost and another for cost and time that exceeds the target by more than 10%. The user can incorporate the known utility function into the model and choose a portfolio based on highest utility. In this example, the decision makers examined best and worst case values from the portfolio options and a target option. These values determined the starting point for a utility function. The decision maker then adjusted the shape of the function until content with the shape.

Jackson *et al.* used lotteries to establish utility independence for cost and time attributes. To confirm the stronger additive independence condition, the authors presented each of the decision makers a choice between lottery X, which compares low cost, long time with high cost, short time, and

lottery Y, which compares low cost, short time with high cost, long time. All decision makers were indifferent as long as the cost and time were within the established limits. Cost and time satisfy the additive independence constraints if their values are less than the maximum allowed by constraints. If the additive independence conditions are true for both attributes, an additive utility function can represent the decision maker's objective function. The additive utility function is relatively straightforward and relies on a weighting parameter to represent the decision maker's preference between the attributes. Jackson *et al.* are able to calculate multiattribute utility for a portfolio once all weights are assigned and select the appropriate technology for each stage of the process.

Having drawn from multiple methods from the toolbox of project selection techniques, Jackson *et al.* successfully address

the issues of uncertainty as they relate to timing and cost concerns using decision trees and multiattribute utility theory. They also lay out a transparent method for project selection due to the government funding of the process. They address issues of compatibility and balance within this portfolio by requiring the selection of one technology per stage. The approach works well for the specific application but does not address dependence among projects or balance in a portfolio in an application where such specific constraints did not exist.

SELECTING A PORTFOLIO USING MULTIATTRIBUTE UTILITY THEORY

Golabi *et al.* [8] take a portfolio view of selecting solar energy projects and expand on popular techniques for use in government procurement. They address several areas where they identify shortcomings in earlier R&D project selection procedures including: treatment of multiple criteria, handling of project interactions, approach toward nonmonetary aspects of the problem, and the perception of difficulty understanding models. The project they tackle focuses on the selection of solar energy projects for funding. Since the projects focus on increasing the knowledge in this area of study, minimal risk or uncertainty exists. All projects funded will increase the knowledge base.

Golabi *et al.* utilize multiattribute utility theory in order to address the issues identified. For this application, they determine that there are limited interactions between proposed projects. Redundancy in project selection was not required, but diversity in technologies was. In order to use a simple multilinear utility function, the selection of a project must be based on utility independent of its complement. Golabi *et al.* express a concern preference for a project of medium quality versus one with equal chances of high quality or low quality. This preference could depend on the overall quality of projects already included in the portfolio. Since the condition of utility independence is not met, Golabi *et al.* decided to decouple the evaluation of technical merit from the portfolio

problem to avoid the complexity of addressing dependence. The technical evaluators determined that budget and diversity concerns were the primary consideration for the portfolio. Upon reviewing a list of cost and diversity issues, the technical evaluators determined that a portfolio would need to achieve a minimum level of diversity. Below the minimum level, the portfolio would be unacceptable. However, no additional value was gained by increasing diversity beyond this level. Thus, a trade-off could not be made between budget and diversity. Constraints were added to assure that the desired level of diversity was achieved. One example was determining the allocation of funding to small, medium, and large sized projects. In many cases, it was difficult for the technical evaluators to identify the best level of diversity. The portfolio selection algorithm was first run with only a budgetary constraint. The technical evaluators then reviewed the portfolio of maximum technical utility. If they did not think the identified portfolio demonstrated sufficient diversity, then they added diversity constraints and reran the model.

To assess the technical utility of the entire portfolio, the technical evaluators identified 22 attributes of interest (Table 1), the utility function associated with the attribute, and the weights given to each attribute. Projects that did not meet a minimum threshold for technical quality were eliminated from consideration. Computer support was used to calculate the utilities once the technical evaluators had input values for each attribute. Once all the attributes had been evaluated, the model was turned over to a panel to experiment with different levels of funding and diversity and make final project selections. Golabi *et al.* report that this procedure allowed 77 projects to be evaluated over a period of two weeks and the selection of 17 projects to be completed in three days. They report a successful implementation of their procedure to this application.

While being successful in this application, the procedure does not provide a method for addressing interactions between projects that would occur in an industrial R&D setting. It also fails to address risk and uncertainty as the issue was not deemed relevant

Table 1. Technical Worth Attributes for the Solar Energy Project Portfolio (adapted from Golabi *et al.* [8])

1. Concept design and system analysis
a. Application identification
• X_1 = Load characterization
• X_2 = Matchup of array output load
b. System conceptual design
• X_3 = System conceptual design description
c. Analysis, optimization, and trade-off studies
• X_4 = Consideration of system design options
• X_5 = Parameter identification and optimization approach
2. Technical performance and integration
a. Component specification plan
• X_6 = Specification of array
• X_7 = Specification of other major components
b. System control of interfacing
• X_8 = Understanding of system control and operation
• X_9 = Interface with other energy sources
c. Evaluation of potential performance
• X_{10} = Evaluation of potential performance
d. Development of major components
• X_{11} = Development of major components
3. Implementation plan
• X_{12} = Definition of work tasks for phase 1
• X_{13} = Identification of team members for phase 1
• X_{14} = General phase 2 and 3 plans
• X_{15} = Program management for phase 1
4. Proposer's capabilities
• X_{16} = Experience of firm(s)
• X_{17} = Experience of personnel assigned to project
• X_{18} = Disciplines of personnel
5. Other characteristics
• X_{19} = Accessibility to technical community and visibility to public
• X_{20} = Potential for low cost
• X_{21} = Percent of load met by photovoltaic system
• X_{22} = Institutional considerations

to the specific decision process described. Golabi *et al.* do describe a more rigorous check of the conditions required for a linear-additive utility function than commonly considered.

DEVELOPING PORTFOLIOS USING STRATEGIC THEMES

Poland [9] and Skaf [10] described the evaluation of portfolios for a variety of industries including pharmaceutical, plastic and packaging, oil and gas, and entertainment. Both authors draw on their experience in consulting with Strategic Decisions Group. Poland focuses on addressing the uncertainty inherent in portfolio problems. He also proposes a unique approach for grouping a portfolio that aligns with the business strategy. The approach Poland describes for setting of portfolio themes is also utilized in Skaf's application in an upstream oil and gas organization.

The assessment of uncertainty by calculating probability distributions on key value measures such as NPV is computationally complex. Poland proposes a simplified method for assessment of the portfolio distribution. The method attempts to balance the communication challenge of presenting a large number of probability distributions for multiple businesses in a meaningful way. A presentation with too little detail could mask important insights. A presentation with too much detail could lead to undue focus on certain details and detract from the high level approach to the analysis [9]. The computational requirements for this type of work are high. For example, describing portfolios for a plastics and packaging company with 20 businesses would produce a probability tree with approximately 3.5 billion branches. Poland limits the expansion by focusing on uncertainties with the most impact on the outcome as determined by a tornado chart and fixing the value at the mean for all low-impact items. In many long-term business models, the top five uncertainties can often account for nearly 90% of the total variance, but in portfolio evaluations, many more uncertainties could be required.

Poland [9] uses decision trees to calculate the distributions for various strategies for each business, analytically combining the moments of the distributions for a given portfolio, and fitting a distribution for overall risk and return. Initially, the consultants evaluate the distributions of business value for various business strategies. Then the senior management sets an overall portfolio strategy theme that would guide the strategy for each business. The theme allows management to account for constraints not explicitly modeled and to some extent could address interactions between items within the portfolio. For example, an overall “Aggressive” strategy could lead to an “Expansion” strategy for Business 1 and an “Acquisition” strategy for Business 2.

The portfolio strategy drives the business-level strategy and thereby portfolio value. Both global uncertainties and business-level uncertainties affect portfolio value. The consultants needed to determine how to approximate distributions of portfolio value quickly, given the distribution of value for each value measure, strategy, business, and global scenario. The solution has four steps: summarize business value distribution with the first three cumulants (mean, variance, and skewness); sum cumulants across businesses to get portfolio values (based on the assumption that the values from each business are independent for a global scenario); convert the portfolio cumulants for each global scenario to raw moments and find the overall raw moments for the portfolio; and, fit a smooth distribution to the moments.

In a workshop setting, the consultants used a spreadsheet implementation allowing for quick and interactive use and summarizing the results in a user-friendly flying-bar chart. During the workshops, many strategy themes are explored to account for constraints such as resources not accounted for with this value model. The consultants have used these techniques in the areas of drug development, oil and gas fields, telecommunications, agricultural products, and potential TV pilot shows. If some subsets are highly correlated (such as two drugs that could cannibalize each other’s markets), they are

preevaluated as a single combined asset. Other scenarios could also lead to evaluation of subset groupings. Another challenge occurs when the probability distribution is not accurately represented by the first three cumulants [9].

While the strategy method does take into account alignment, a key item in successful portfolios, it neglects to address how one would evaluate an individual business-level strategy. For example, there is no explanation for how to choose which “Expansion” plan to apply to Business 1 in the “aggressive” portfolio strategy. Poland’s strategy also mentions the issue of interaction in the form of cannibalization but ignores how one evaluates interrelated products as a single asset.

A HIERARCHICAL DECISION PROCESS

Peerenboom *et al.* [11] takes a hierarchical approach in allocating funds to a supplemental environmental program (SEP) based on synthetic fuels. The funding was tied to a multibillion-dollar loan agreement between the DOE and the Great Plains coal gasification facility. The funding requirements for the proposed projects exceeded the available funds by more than a factor of two. Also, national attention was focused on the Great Plains facility [11]. Thus, DOE chose decision analysis to produce a well-documented and traceable record of the decision process.

The DOE established a steering committee and five technical subcommittees to develop the SEP. Each subcommittee proposed a number of detailed studies for health or environmental concerns. The subcommittees did not have individual budget constraints, but the overall budget was \$12 million. The budget allocation decision was complex: the organizational structure included five independent subcommittees; there were uncertainties around research needs, data availability, and costs; value trade-offs were required at both the committee and subcommittee levels; there were numerous strategies of more than 100 projects to evaluate.

The procedure used builds on previous applications of decision analysis techniques to rank projects and evaluate portfolios. It

uses a hierarchical structure to integrate lower level and portfolio level decision analysis. The procedure was tailored to the structure of the committee and subcommittees. Each subcommittee was responsible for ranking its proposed studies. The subcommittee then quantified the degree to which a portfolio met a set of portfolio objectives as a function of funding level. The subcommittees used this information to produce a standardized set of performance curves [11].

The analysts followed a four-step procedure. Step 1 defined the portfolio objectives and attributes. Committee members developed a hierarchy of objectives. Specific objectives were used to build up to broader, more general objectives. The committee then developed scales and attributes for each objective to indicate how well each portfolio objective was met by subcommittee plans. Step 2 ranked the subcommittee studies and developed performance curves. Each subcommittee developed objectives that were more specific than the overall portfolio objectives. This step required quantifying a multiattribute utility function that represented the subcommittee chairperson's preferences over the subprogram objectives. The process involved determining: (i) the trade-offs the chairperson was willing to make between competing subprogram objectives, and (ii) the chairperson's attitude toward risk. The subcommittee evaluated each proposed study in terms of the utility function developed previously, used probability distributions to represent uncertainty, ranked studies on the basis of expected utility, and performed sensitivity analysis. This step links lower and higher levels in the hierarchy. Each subcommittee quantified how well its proposed studies met the portfolio objectives for given levels of funding. As funding levels were reduced, lower ranked studies were cut first in most cases. Subcommittees reviewed the proposed plan to assure that the selections made sense together. Step 3 quantified preferences for portfolio objectives defined in Step 1. In this step, the committee quantified a multiattribute utility function to represent the committee chairperson's preferences over portfolio objectives. In addition to determining chairperson's value trade-offs and attitude

toward risk, the committee addressed utility trade-offs between the five subcommittee plans. This evaluation produced a set of subprogram scaling constants. Step 4 evaluated and compared feasible funding strategies to finalize SEP portfolio. A model using a backward dynamic programming algorithm to maximize utility from the funding of studies in the subprogram areas was used to identify and evaluate the large number of feasible funding strategies. The hierarchical approach came in at Step 2 when the subcommittee sets priorities for its set of subprogram studies. This feature is a major contribution to this procedure but represents only one input into the portfolio level decision making.

At the portfolio level the chairperson identified comprehensiveness, relevance, and cost effectiveness as the broad areas of concern. The committee established objectives, attributes, and scoring criteria for each area of broad concern. Performance curves were created for each attribute to show how well the subcommittee portfolio would do based on a given percentage of requested funding. The performance curves show that for the attribute of coverage, the value is 100% at full funding of the toxicology subprogram. If funding drops by 20%, the coverage of the toxicology subprogram decreases by nearly 50%. The performance curves allow the portfolio to be assessed as a whole unit.

The steering committee allocated a reduced amount of funding of \$9 million across the five subcommittees. Prior to final allocation sensitivity analysis was completed on changes in levels of (i) subprogram scaling constants, (ii) portfolio level utility function scaling constants, and (iii) subprogram performance curves. The chairperson adjusted the funding priorities from the model following extensive reviews and discussions with stakeholders. This adjustment impacted only three of the 88 proposed studies [11].

The method described by Perrenboom *et al.* addresses alignment with strategy and handling of multiple criteria. It also creates a transparent process for the decision and allows for adjustment to build consensus among committee members. Some of the attributes, such as coverage, defined at the

portfolio level address the balance in the portfolio. One consideration not explicitly modeled was interactions between funded projects.

DECISION CONFERENCING TO DEVELOP PORTFOLIOS

Phillips and Bana e Costa [3] describe a multicriteria decision analysis approach to portfolios that they have utilized in numerous consulting applications in various industries. They use multiattribute utility theory, but place a greater emphasis on the social aspects of the decision than, for instance, Golabi *et al.* [8]. Much of their focus is on transparency and consensus building. The primary metric used in their evaluation of projects is value for money determined by the ratio of risk-adjusted benefit to cost. They noted that, much literature recommends this approach. However, in practice, most companies without formal decision analysis support rely on just the expected benefit. Figure 3 shows that picking elements of the portfolio with the highest expected benefit is not the optimal use of the budget. The benefit only curve is always under the cost adjusted benefit curve.

Similar to the approach previously described by Golabi *et al.*, Phillips and Bana

e Costa collapse multiple dimensions of benefit into a single risk-adjusted benefit. The benefit criteria must be mutually preference independent in order to use an additive value function, where the overall value of option i is described by the equation below

$$V_i = \sum_j w_j v_{i,j},$$

where, $v_{i,j}$ represents the value associated with consequence i on criterion j and w_j represents the weight assigned to criterion j .

Several software programs exist for portfolio analysis. Phillips and Bana e Costa describe the approach taken in the software package EQUITY. The basic structure mimics an organization of K areas whose options are appraised against J benefit and risk criteria, producing $K \times J$ scales. The options for each area are appraised against each criterion separately, resulting in a value score $v_{i,j}$ for each option i on criterion j , such that for each scale 100 represents the most preferred option and 0 the least. Then each of the scales for criterion j will be assigned a within-criterion weight, w_j , using swing weighting. The scale associated with the largest difference in value between two reference points is assigned a weight of 100, and others are given a weight relative to 100. The scales assigned within-criterion weights of 100 for each criterion are compared for their swings, producing a set of across-criteria weights w_j . Value scores, within-criterion weights, and across criterion weights are required inputs for EQUITY to calculate the overall value. EQUITY then calculates the benefit-to-cost ratios by dividing each option's overall value by its total cost.

This process results in a single value-for-money triangle associated with each option. The triangles are stacked in declining order of value-for-money priority to create an efficient frontier of projects as seen in Fig. 4. The portfolios of projects up to a given budget are examined and projects that fall outside of the portfolio are examined to make sure exclusion is realistic. The shaded area under the efficient frontier includes all possible portfolios.

At this stage in the decision process constraints are introduced. The decision maker

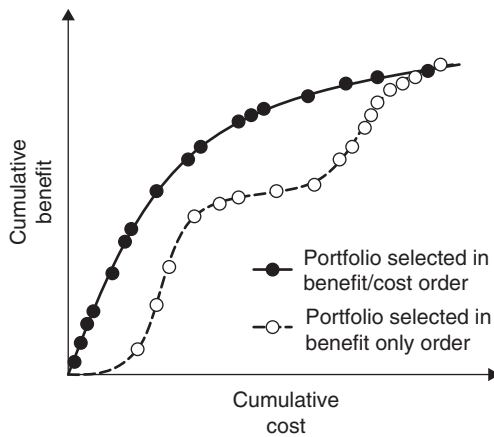


Figure 3. Portfolio value with increasing budget when chosen in order of highest benefit/cost ratio versus just highest benefit (adapted from Philips and Bana e Costa [3]).

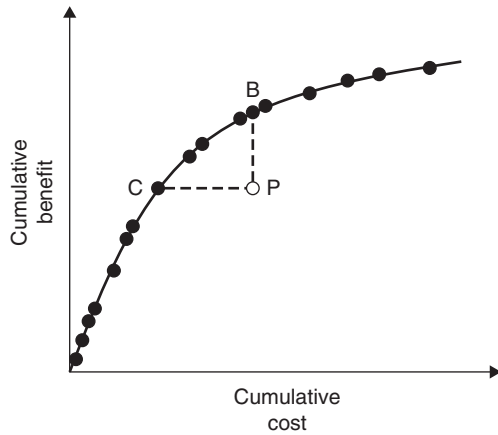


Figure 4. Illustration of the efficient frontier (adapted from Phillips and Bana e Costa [3]).

can determine that an excluded project is too far along to stop or that new projects are infeasible due to other current conditions. The decision maker can propose a portfolio of current projects only. This proposed portfolio, P, is below the efficient frontier. Figure 4 shows that an improvement could be made by moving to portfolio C (same benefit lower at a lower cost) or portfolio B (same cost increased benefit). In 20 applications of Equity, benefit of moving from P to B averaged 30%. This approach helps decision makers make difficult decisions to close down projects that do not look promising. Participants also gain an understanding that what is best for an individual area is not always best for the whole organization.

Portfolio item dependencies are handled visually, following an ad hoc procedure. If the proposed portfolio includes or excludes two dependent projects, no further action is required. Otherwise, the omitted project is manually included and the resulting portfolio analyzed. If two projects are truly dependent on each other, it may be more effective to model them as a single option. Phillips and Bana e Costa pointed out that it is most efficient in practice to focus only on a few important dependencies. A major challenge facing consultants is managing the trade-off between sophisticated modeling and social acceptance of the process. In opting for an approach that favors social acceptance,

Phillips and Bana e Costa must necessarily simplify the complex issue of project dependence. Alignment to objectives is considered in the benefit assessment. The model itself does not account for a balance in selected projects but balance is considered in the decision conference by visually imposing additional constraints as requested to explore balance across various dimensions. Phillips and Bana e Costa place the most emphasis on transparency and acceptance of any of the studies surveyed in this article.

A CONTINGENT PORTFOLIO PROGRAMMING APPROACH

Gustafsson and Salo [12] point out the limited acceptance of the decision analytic method in industrial portfolio selection. They suggest that the slow industrial uptake is due in part to the inability of existing methods to address all areas relevant to the problem. They build on the existing work from decision analysis, R&D management, and financial portfolios to develop the contingent portfolio programming (CPP) method. In addition to drawing on multiattribute methods, Gustafsson and Salo identify optimization models and dynamic programming models as the most relevant to the portfolio problem. Optimization models can capture project interactions and resource constraints, but traditionally fail to address uncertainty. Gustafsson and Salo consider decision trees and real options analysis to be in the category of dynamic programming. This type of modeling captures the sequential nature of decision making, but fails to address project interactions or resource constraints. They point to the options literature that addresses risk preferences but fails to mimic a continuous range of options not a discrete set such as in project selection. Smith and Nau [13] and Perdue *et al.* [14] also integrate decision trees and real options.

CPP provides a methodology for a decision maker to select risky projects over multiple time periods. The CPP approach incorporates decision trees to mimic flexibility of the decision maker to make ongoing go/kill decisions based on available information. CPP

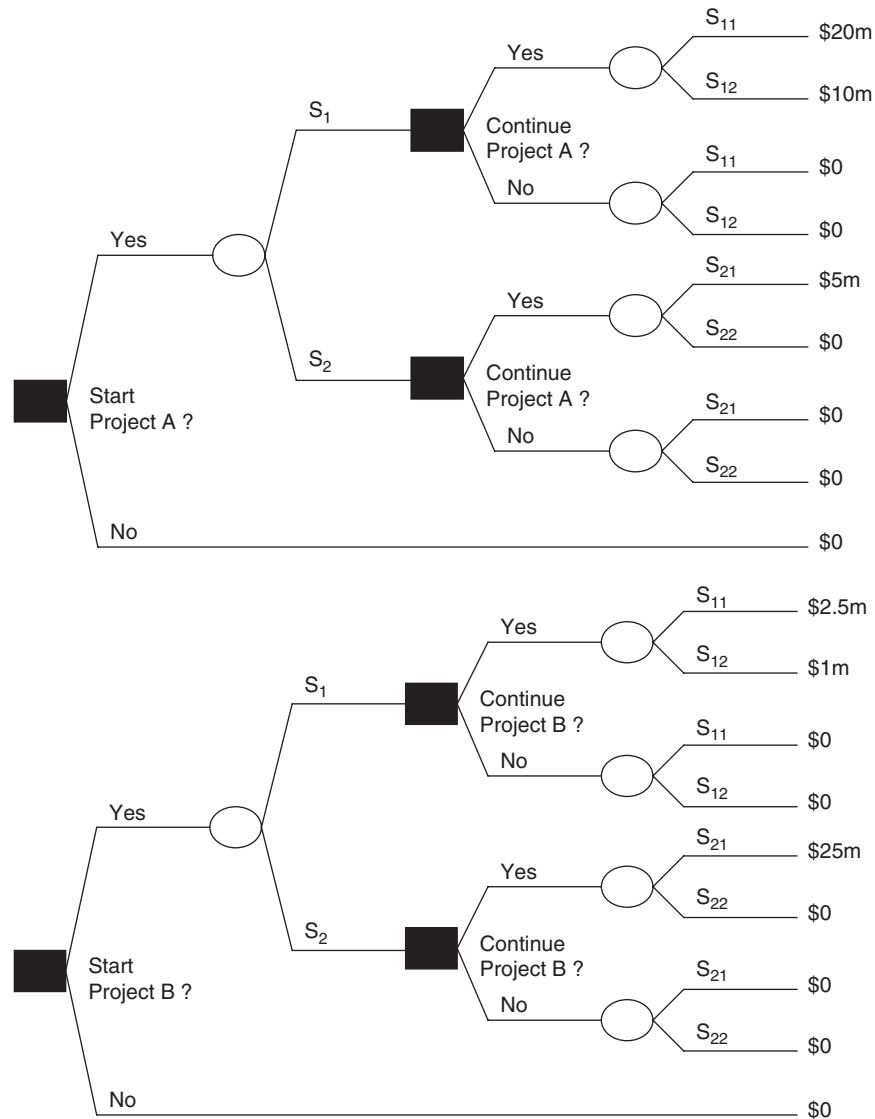


Figure 5. Decision trees for two projects with common state uncertainties (adapted from Gustafsson and Salo [12]).

offers flexibility to accommodate a range of risk attitudes. The CPP model is defined by resource types, a state tree, and decision trees by project. The method accommodates many types of resources both tangible (e.g., capital or equipment) and intangible (e.g., skill sets). Future states of nature are represented by a state tree. A decision maker has choices at a number of decision points for each project. At each decision-point, the decision maker

chooses the action taken. Decision variables are defined for each action. The variable is binary with a value of 1 if the action is made and 0 otherwise. Sample decision trees are defined for two projects A and B and shown in Fig. 5. The same states from the state tree are included in each decision tree, allowing for project dependencies.

Resource flows are defined at each state. Resources can either be gained or consumed

at each point depending on actions chosen by the decision maker. In evaluating this decision, the decision maker's objective is to maximize the utility of the initial position. Gustafsson and Salo focus on a special case that has a reasonable model of risk aversion and is appropriate for linear programming (LP). In addition to the objective function, they define several classes of constraints including decision consistency constraints, resource constraints, and a number of optional constraints. The simple two-project example demonstrates the benefits to considering the projects together instead of each individually. Project A and B succeed inversely. Project A fares better if state S_1 occurs in period two and Project B fares better if state S_2 occurs. Either project selected individually would have a negative NPV. If the decision maker invests in both projects in the first stage and then makes a decision about which project to fund for the second phase depending on the current state of nature, the expected NPV is positive. The diversification of the portfolio mitigates some of the risk. As the number of projects, resources, and constraints increase the problem becomes more computationally complex. Gustafsson and Salo test a number of scenarios using C++ and an LP solver. They find that LP models can be solved in a reasonable period, but the time to solve mixed integer programming (MIP) formulation increased exponentially with the number of integer variables.

Gustafsson and Salo recommend theoretical extension of the model to include more complex resource dynamics. They identify situations where decision trees can be defined for each project and correlated projects. The more complex theoretical approach that Gustafsson and Salo embrace stands in stark contrast to the beliefs of Phillips and Bana e Costa [3] that lean to a simplified model and rely on social process to guide decision making.

CONCLUSIONS

The topic of R&D portfolios is complicated and demands adequate tools to address all relevant concerns. Each method described herein has its own advantages suited to

the specifics of each application. However, Phillips and Bana e Costa [3] and Gustafsson and Salo [12] performed well, generally on our five criteria of a good portfolio selection method. The former approach ensures alignment with strategy, balance across the portfolio, value maximization, consideration of conflicting objectives, and an exemplary social process. Interrelationships within the portfolio are handled in an ad hoc manner. The latter approach is more technical with less social emphasis, but also offers alignment with strategy, balance across the portfolio, value maximization, consideration of conflicting objectives, and explicitly models portfolio interrelationships with the state space diagrams.

Thus, the question turns to the level of detail needed in a model or technical solution versus the social process, traceability, and simplicity of the approach. The industry in question and the corporate environment must affect the choice of tool for a specific application. Keeney and von Winterfeld [15] discuss "practical" value models, noting that it is not always necessary or desirable to construct a complex value model even though it might be theoretically justifiable. They also acknowledge that in some cases theoretically valid assessment procedures are not required. The appropriate level of complexity is driven by the decision scenario, the resources available to gather data or implement a model, and the time allowed for making the decision. Phillips [16] discusses a similar concept of requisite modeling. A requisite model is one that is sufficient to resolve the issues under consideration. The iterative process between consultants and decision makers to define the model increases the understanding of the situation and resolves decision makers' concerns on validity of output from a model. A model is then requisite when no additional insight is evolving. In the end, the specific application for portfolios and the industry in question will drive method selection and the implementation plan.

REFERENCES

1. Cooper RG, Edgett SJ, Kleinschmidt EJ. Best practices for managing R&D portfolios. *Res Technol Manage* 1998;41(4):20–33.

2. Matheson D, Matheson J. The smart organization: creating value through strategic R&D. Boston (MA): Harvard Business School Press; 1998.
3. Phillips LA, Bana e Costa CA. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Ann Oper Res* 2007;154(1):51–68.
4. Balthasar HU, Boschi RA, Menke MM. Calling the shots in R&D. *Harv Bus Rev* 1978;56(3): 151–160.
5. Sharpe P, Keelin T. How Smithkline Beecham makes better resource-allocation decisions. *Harv Bus Rev* 1998;76(2):45–57.
6. Chien C-F, Sainfort F. Evaluating the desirability of meals: an illustrative multiattribute decision analysis procedure to assess portfolios with interdependent items. *J Multi-Criteria Decis Anal* 1998;7:230–238.
7. Jackson JA, Kloeber JM Jr, Ralston BE, *et al.* Selecting a portfolio of technologies: an application of decision analysis. *Decis Sci* 1999;30(1):217–238.
8. Golabi K, Kirkwood CG, Sicherman A. Selecting a portfolio of solar energy projects using multiattribute preference theory. *Manage Sci* 1981;27(2):174–189.
9. Poland WB. Simple probabilistic evaluation of portfolio strategies. *Interfaces* 1999;29(6): 75–83.
10. Skaf MA. Portfolio management in an upstream oil and gas organization. *Interfaces* 1999;29(6):84–104.
11. Peerenboom JP, Buehring WA, Joseph TW. Selecting a portfolio of environmental programs for a synthetic fuels facility. *Oper Res* 1989;37(5):689–699.
12. Gustafsson J, Salo A. Contingent portfolio programming for the management of risky projects. *Oper Res* 2005;53(6):946–956.
13. Smith JE, Nau RF. Valuing risky projects: option pricing theory and decision analysis. *Manag Sci* 1995;41(5):795–816.
14. Perdue RK, McAllister WJ, King PV, *et al.* Valuation of R and D projects using options pricing and decision analysis models. *Interfaces* 1999;29(6):57–74.
15. Keeney RL, von Winterfeldt D. Practical value models. In: Edwards W, Miles RE Jr, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. pp. 232–252.
16. Phillips LA. Decision conferencing. In: Edwards W, Miles RE Jr, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. pp. 375–399.

AN OVERVIEW OF INQUIRY-BASED LEARNING IN MATHEMATICS

STAN YOSHINOBU

Department of Mathematics,
California Polytechnic State
University, San Luis Obispo,
California

MATTHEW JONES

Mathematics Department, California
State University, Dominguez
Hills, Carson, California

"If education is always to be conceived along the same antiquated lines of mere transmission of knowledge, there is little to be hoped from it in the bettering of man's future."

— Maria Montessori

INTRODUCTION

A number of studies have found that undergraduates may not be learning what instructors would hope in undergraduate mathematics courses. Traditionally, instructors of mathematics have rationalized the process of students turning away from mathematics as a process of weeding out students unfit to study mathematics. However, evidence from "Talking About Leaving" [1] casts doubt on this perspective. This study found that students who switched out of mathematics and other sciences were not necessarily failing out of courses; on the basis of grades, even grades in the major, it was difficult to predict who would switch and who would remain. Instead, the primary reasons for switching were based on the way the courses were taught. Indeed, the top four factors reported as contributing to students' decision for switching majors are:

1. lack of/loss of interest in science, mathematics, and engineering (SME),
2. non-SME major offers better education/more interest,

3. poor teaching by SME faculty,
4. overwhelming curriculum overload and fast pace.

Rather than the common notion that students leave because they cannot perform in the discipline, these reasons, although more broad than teaching alone, highlight that the way courses are commonly taught is a primary problem.

In addition to the problem of capable students leaving mathematics for other disciplines, there is the issue of the skills and beliefs of students who complete courses in mathematics. Muis, in her review of the literature [2], found that across 33 studies of students from first grade through college mathematics, there is a pervasive belief that only geniuses are capable of understanding or doing mathematics, that good students can solve mathematics problems in 5 min or less, and that one succeeds in mathematics by following procedures described by the instructor to the letter. Muis terms these nonavailing beliefs as an indication that such beliefs are correlated with poorer learning outcomes. These beliefs are held even in a tenth-grade geometry class, as observed by Schoenfeld [3], in which there were high-achieving students, as measured by standardized tests.

Studies also indicate that students are unable to apply what they know. Research by Selden and colleagues [4–6] are illustrative of students' difficulties. In their work, the researchers demonstrated that students with C's, as well as those with A or B grades, had very little success in solving nonroutine problems, even when associated tests showed that the students possessed the requisite algebra and calculus skills. Let us take the study by Carlson [7] as another example. Carlson [7] studied students who had just received A's in college algebra, in second-semester honors calculus, or in first-year graduate mathematics, and, from examinations and interviews, found that not even the top students completely understood the concepts taught in a course, and when faced with an unfamiliar

problem, they had difficulty in utilizing material recently taught.

Another study illustrating the issues faced by collegiate mathematics instructors is work on proof schemes [8]. Harel and Sowder, through data amassed from 128 students in six teaching experiments, described three broad categories of proof schemes, which they defined as consisting of “what constitutes ascertaining and persuading for that person.” Many students exhibited external conviction proof schemes, meaning that what the students found to be sufficient justification depended on things such as an authoritative instructor or an authoritative book, on the appearance of the argument, or on not-necessarily-meaningful manipulations of symbols. Other students exhibited empirical proof schemes, in which evidence from examples is considered sufficient justification. Only the third category, analytical proof schemes, contained those students whose views of proof were consistent with mathematical proof.

In contrast to this evidence, in the remainder of this article, we will describe inquiry-based learning (IBL) as a means to actively engage students in learning mathematics, and we will review evidence suggesting that IBL may improve student outcomes on measures such as problem solving and beliefs about mathematics, while not sacrificing outcomes on basic skills.

IBL encompasses a fairly broad range of approaches to course design and delivery. Historically, many practitioners of IBL in mathematics trace an influence to the mathematician RL Moore. Moore taught mathematics with an approach that we can consider an example of IBL. What is often called the *Moore method* [9] has many variants, and taken as a whole, these are indistinguishable from the approach we call IBL. At its core, IBL is a paradigm in which (i) *students are deeply engaged with mathematical tasks* and (ii) *there are opportunities for substantive collaboration with peers*. These are the foundational principles on which IBL is built.

WHAT DOES RESEARCH INDICATE ABOUT INQUIRY-BASED LEARNING COURSES?

In mathematics, the research base supporting IBL is still developing, but the studies that do exist support IBL as a means to achieve outcomes better than lecture-based courses on measures such as problem solving and beliefs about mathematics, without sacrificing outcomes on procedural knowledge. This finding has held across studies large and small, in calculus and upper division mathematics, and is consistent with the larger research base on IBL.

Smith [18] compared students in two inquiry-oriented number theory courses to students in a number theory course taught via the traditional lecture approach, by interviewing a small number of students from each course. The students from the IBL section exhibited quite different views of proof and approaches to proof when compared with the students from the traditional course section. In particular, students from the IBL sections approached proofs by trying to make sense of the statement and would work through examples to generate insight into the idea of the proofs, while students from the traditional section would search for proof techniques and were reluctant to work on examples, as examples were not going to be part of the proof. Students from the traditional course section tried to relate a current proof task to other proofs, based on the surface features (for instance, “prime” is mentioned, whether or not primality is actually an essential condition for the proof), rather than based on understanding the concept. The IBL students also sought to understand their proofs and try to write them in ways that were meaningful, whereas the traditional section students were more interested in figuring out what they were supposed to do. Although these findings are limited by a small sample, they are suggestive of results that arise from larger projects, which we review next.

We turn our attention to studies of other IBL courses, not necessarily proof-oriented courses. The Inquiry-Oriented Differential Equations Project investigated student learning of a specially designed curriculum

in which the course goals were described as follows:

We wanted students to essentially reinvent many of the key mathematical ideas and methods. . . We wanted challenging tasks, often situated in realistic situations, to serve as the starting point for students' mathematical inquiry. . . We wanted a balanced treatment of analytic, numerical, and graphical approaches, but we wanted these various approaches to emerge more or less simultaneously for learners [19].

Students in the course developed, used, and compared multiple approaches to solving differential equations. This contrasts with the traditional model of the instructor demonstrating solutions to well-defined classes of differential equations and then assigning the students similar problems to practice solving. Courses using this approach were compared with courses using the traditional approach. At the end of the semester, students from the inquiry-oriented courses performed similarly to the students compared on procedural items, but outperformed students from traditional courses on modeling and conceptual tasks. In a follow-up study one year later, students from one of the inquiry-oriented courses and one of the traditional courses took a retention test, again composed of procedural, modeling, and conceptual tasks. In this case, the results were the same: students from the inquiry-oriented course performed similarly to the students from the traditional course on procedural items, but outperformed on modeling and conceptual items. This finding held even though 75% of the traditional students and none of the inquiry-oriented students had taken a course in numerical differential equations in the interim year [20]. This finding, that students in IBL courses tend to perform about as well as students from traditional sections on procedural items, but do better on conceptual items, was confirmed by six of the seven studies that reported statistical significance, as reported in a review of studies of reform calculus [21]. Thus, across eight studies of calculus and differential equations, students from

IBL courses tend to perform better than comparison students on conceptual items, with no cost to performance on procedural items.

Laursen and her colleagues [22] engaged in a multiyear, multisite study of the impact of IBL on students. Mathematics departments at four research universities (three public and one private) agreed to offer IBL courses and to participate in data collection. Across a variety of courses at the four sites, the main features of students solving problems from a carefully sequenced problem set, with class time primarily spent on student-centered activities, and with the instructor participating but not significantly in the role of lecturer, were all present to varying degrees. Overall, 60% of IBL class time was spent on student-centered activity, whereas 87% of comparison non-IBL course time was spent listening to the instructor talk. Surveys asked students to report gains in several areas. Students who took IBL math-track courses (i.e., excluding courses for future elementary teachers) had higher gains in areas such as understanding of concepts, improved thinking and problem-solving skills, gains in confidence and persistence, collaboration, and comfort in teaching mathematical ideas to others. Preservice teachers did not have a comparison group of non-IBL preservice teachers. However, preservice teachers had gains higher than math-track non-IBL courses in areas such as applying mathematical knowledge, collaboration, and comfort in teaching mathematical ideas to others. In addition, students with low prior achievement who took IBL courses and then took math courses after their IBL experience also tended to earn higher math GPAs in comparison to low-achieving students who took non-IBL courses. In other words, students with low prior achievement in mathematics made stronger gains in mathematics achievement by taking IBL courses. These findings add depth to the picture of student performance painted by the earlier studies, by highlighting particular ways in which students perceived their own gains both on content learning and on beliefs and attitudes toward mathematics.

These findings in collegiate mathematics are similar to the findings elsewhere, where there is more evidence investigating various forms of IBL in college classrooms. The evidence from collegiate science courses indicates that on higher level cognitive outcomes such as thinking, problem solving, and laboratory skills, IBL is superior than traditional science instruction [23].

Support for an IBL approach can also be found in the literature on mathematics education at the precollegiate level. While it is beyond the scope of this review to summarize this literature, we review two studies that take a thorough look at the scope of the differences in student experiences under IBL versus a traditional approach. Boaler [24] conducted a three-year study of two high schools in the United Kingdom, one with a traditional approach and one that used an IBL approach, emphasizing projects throughout the curriculum. In spite of a de-emphasis on textbook-like basic skill problems, students at the IBL-oriented Phoenix Park performed similarly to the traditionally taught Amber Hill students. Simultaneously, Phoenix Park students outperformed Amber Hill students on contextualized problems. Moreover, while Phoenix Park students were learning to apply the mathematics they were learning, Amber Hill students exhibited cue-based behavior, and possessed a kind of “inert procedural knowledge” [24, p. 59] that they could only use in a narrow range of textbook-like situations. In a separate four-year study in the United States [25], a comparison of 700 students across three high schools yielded similar results in favor of the school with the inquiry-based approach.

Summarizing, we find that comparative studies indicate that students in IBL courses perform better than students in traditional courses on measures of conceptual understanding, perform about as well on measures of procedural competence, and report higher gains in areas such as confidence and persistence, and the ability to teach or explain mathematics to others.

BASIC COMPONENTS OF INQUIRY-BASED LEARNING

First, we present a typical first couple of days in an IBL class to give a sense for the overall structure of an IBL class, and then later we describe principles and structures typical of IBL. What is presented in this section represents the main features of IBL courses. Individual instructors have developed various adaptations of IBL, and describing the entire portfolio of IBL implementation is beyond the scope of this article. What we strive to do is to convey the main principles of IBL teaching in mathematics, and leave open to the reader how it might vary depending on the course, students, and other factors. Moreover, we describe specifics about IBL from a mathematician’s perspective, and emphasize that the core ideas presented in this article have broad applicability across all fields of study. Indeed, as subjects become more applied in nature, it is perhaps easier to motivate students with relevant problems presented in context.

Typical day one: After a brief introduction, students are given problems from a well-crafted problem sequence. They start working on the problems in class. Basic definitions and a handful of starter problems are given to the students. Starter problems are first and foremost accessible questions that ask students to unpack the meaning of definitions, develop specific skills, or work with fundamental concepts.

Definition 1. An integer N is even if and only if $N = 2k$ for some integer k .

Definition 2. An integer N is odd if and only if $N = 2k + 1$ for some integer k .

Problem 1. Prove or disprove the following statement. The sum of two even numbers is even.

Problem 2. Prove or disprove the following statement. The sum of two odd numbers is odd.

After giving the two definitions and a list of problems (including problems 1 and 2), the instructor asks the students to work on the first few problems on their own. After students have had a chance to develop their

own ideas, they may be asked to share their ideas with one or two partners. Next, the class is instructed to produce, as best they can, a proof of problems 1 and 2. As students are working and completing their proofs, the instructor's role is to visit with several (possibly all) groups, and find two groups capable of presenting proofs to the first two problems. After the class has had enough time on these problems, two individuals or groups are selected to present a proof to the class for problems 1 and 2. The students in the class would be given the role of making sure that the proofs are correct. The instructor's role is to facilitate the discussion of the proofs, specifically managing the question-and-answer component of the discussion. All of this should be carefully explained to students.

During and after a student (or group) presents a proof, the class is allowed to ask questions for clarification, to point out gaps in arguments, and to offer suggestions. Once all questions are fielded, then the class is asked to approve the solution, request clarification from the presenter, or disagree with the solution. If the solution is approved, the presenters receive credit for producing a correct proof. If the class cannot arrive at a consensus, the outstanding issues are typically written on the board as questions for students to work on. A choice is now made by the instructor to continue working on the questions that were raised or to ask the presenter(s) to return the next class period with an updated solution. Then the class moves on to the next problem. Normally, with well-chosen starter problems, the class as a whole makes it through the first few problems without difficulty.

One of the goals of day 1 is to get some students to successfully solve a problem and present a proof in class. This establishes how the class will transpire for the rest of the term. At this point of the course, it is vital for an IBL instructor to very clearly outline (i) students' roles as young mathematicians and problem solvers; (ii) the instructor's role as facilitator, mentor, and coach; and (iii) how students will be stuck and that being stuck is ok, provided these situations are turned into opportunities to search for

new ideas and strategies. Instructors often start the first day by informing the students of the above, and revisiting these topics as necessary throughout the term. When the first period is about to end, students are asked to continue to work on the next five to seven problems from the list. Their standing instructions are to work on problems outside the class and return to the next class with their own solutions. All outside resources are strictly forbidden, and their goal should be to get stuck and enjoy the growth that comes from being stuck. One of the most important messages students should take to heart and instructors should develop is, "Being stuck is okay!" In fact, helping students deal with being stuck is one of the most important skills needed for a successful IBL course and is discussed further under Students' Roles and Some Pitfalls to Avoid.

On day 2, class starts with the instructor reminding students of their roles and how class will proceed. Since only a few students have presented thus far, volunteers are requested among those with a proof of one of the next four to five problems (e.g., problem 3 and on). Students are selected and asked to go to the board and write their proofs simultaneously. When all students are done writing their proofs on the board, the student with a proof for number 3 is asked to present his/her proof to the rest of the class, line by line. The class is asked to check that the presented proof is correct. If errors are found, then the class is obligated to ask questions and offer suggestions. If a presenter is unable to fix a problem at the board, they are given the option of returning the next class period with a fix. Otherwise, if all the steps are fine and the entire class is satisfied with the solution, then the solution is deemed correct, and the next presenter will then share his/her proof of the next problem. The course proceeds in this way.

After a few days, some of the students will have presented, and some will not have had a turn at the board. One way to keep students motivated and provide ample opportunities for all students is to keep track of who has presented, and to pass out a sign-up sheet at the beginning of each class meeting. The sign-up sheet will have the students

listed alphabetically along the left column, and the next 10 or so problems listed across the top. Students can indicate the problems they believe they are capable of presenting by marking appropriate columns with an X. Students who have presented the fewest number of times get first priority. Instructors also have the option of seeking out at the beginning of class those students with the fewest presentations to ask them whether they have a proof to present to the class. Occasionally, it becomes apparent that a student will not present unless asked, in which case the instructor may choose to respond by assigning the student a specific problem ahead of time and asking the student to prepare a proof for presentation. Additionally, working in pairs on accessible tasks in class is another strategy to generate opportunities for reticent students to present in class. (Extremely shy students can sometimes be accommodated through written work in lieu of presentations and/or giving presentations in office hours. It is emphasized that such instances are rare and special, and are handled outside the standard operation of the course.) Therefore, mechanisms exist to provide for students a way to fairly distribute problems and ensure that every student who puts forth a reasonable effort will have chances to present their solutions.

In summary, the basic components of a typical IBL course are:

- a well-crafted problem set,
- the students' role defined to include problem solving and presenting solutions in class,
- class discussion and peer review of all solutions,
- the instructor's role to primarily guide, coach, and mentor.

BIG IDEAS IN IBL TEACHING

In addition to the basic structure of an IBL course, it is valuable to consider some of the "big ideas" in IBL teaching. These ideas are how to construct appropriate course materials, deal with common pitfalls, expand on the

instructor's role, and build a community of learners.

IBL Course Materials

Course materials are generally constructed so that a standard textbook is not used, although there exist books specifically designed for IBL courses. IBL course materials consist of a well-crafted list of problems, theorem statements, axioms (or basic assumptions), and definitions. The list of problems is normally given to students as one section or one unit at a time. An important feature of these IBL course materials is that they do not contain proofs, solutions, or carefully worked examples that students can mimic. Instead of presenting answers to students, students are asked to make sense of the questions and to construct their own solutions. Students must read and understand definitions, construct examples, explore strategies to solve problems, examine other students' proofs, and find their own proofs/solutions to problems that they do not already know how to solve. The nature of the questions tends to be different than standard textbooks, because they must include opportunities for students to learn the big ideas and strategies in the subject. In this way, students are deeply engaged in rich mathematical tasks, and are doing mathematics like mathematicians.

One of the major reasons why problems are handed out one section at a time is that students may decide to take a path that is not the instructor's intended route through the material. The learning experience is akin to exploring a landscape. As students progress through the material, unworked examples and lemmas may need to be added (or subtracted). These alterations tend to be minor, but classes vary from term to term, and there are occasions when students create unexpected paths of discovery. Moreover, such alterations cannot be planned for in advance, because they depend on the specific nature of the mathematics generated by the students in the class. With that said, it is unlikely that major overhauls are required or that the entire course framework comes apart. Additional course material is sometimes needed to help students overcome specific learning

challenges or to honor their ingenuity by finding another path to a solution. Then, the course problem set is a living document.

How does one find IBL course materials? Several problem sets can be found at *The Journal of Inquiry-Based Learning in Mathematics* (www.jiblm.org). This journal is a collection of peer-reviewed and field-tested IBL course materials. In addition to the peer-reviewed materials, the Journal also has links to a separate collection of IBL course materials, which have not been peer-reviewed, but however may be useful to instructors. In addition to the Journal, several IBL textbooks have been published:

- *Chapter Zero* [10],
- *Closer and Closer* [11],
- *Essentials of Mathematics: Introduction to Theory, Proof, and the Professional Culture* [12],
- *The Heart of Mathematics: An Invitation to Effective Thinking* [13],
- *Number Theory Through Inquiry* [14], and
- *Linear Optimization: A Simplex Workbook* [15].

How does one construct IBL course materials (or a module/unit)? In proof-based courses (topology, number theory, algebra, analysis, intro to proofs, etc.), course materials are generally constructed by reverse engineering from goal theorems. The first step is to identify the goal theorems in a course. Goal theorems encapsulate the major learning objectives of the course and from these goal theorems the problem sequences can be reverse engineered via finding all the assumptions, definitions, lemmas, and techniques needed to prove the theorems. These problems are then sequenced and made suitable for one's students. Suitability depends on the level of the course, student body, and the prior student experiences with IBL or problem-solving approaches. One of the advantages of IBL is that course materials can be tailored to specific needs that can vary across institutions and cohorts.

When all is said and done, the problem sequence starts with basic definitions and a

good supply of starter problems to engage students in the topic being studied. The starter problems should be at a level that is easily accessible to all students and leads students toward understanding definitions, core ideas, and fundamentally important techniques. These starter problems lead to lemmas, theorems, and eventually to the first goal theorem. Then the class is on to the next topic or challenge, and students begin again with new definitions, starter problems, lemmas, theorems, and goal theorems.

The following is a list of key components of IBL course materials:

- Units start with first principles and definitions.
- Problem sets include a good supply of “starter problems” so that all students, especially those who struggle the most, have successes early and often.
- The sequence of problems builds up to lemmas and goal theorems.
- The entire course is organized around the goal theorems.

Students' Roles and Some Pitfalls to Avoid

The first and foremost student role is to be a problem solver. Students should do math, be stuck, and think hard about mathematics. This poses great opportunities as well as some pitfalls. The opportunities are clear—students can learn to become better problem solvers and proof writers. Students are to be instructed not to use outside resources of any kind (e.g., web, books, students not enrolled in the class). Instead of looking up answers, their job is to create their own solutions. Although more challenging, it is better for students, as they will develop the capacities that society values, such as problem solving, creative thinking, skillful communication, and experimentation. Despite the benefits, IBL courses often go against students' prior experiences in math classes. Thus, it is very easy for students to fall back on old habits.

Another of the core principles in IBL courses is that students bear the responsibility of evaluating whether a proof or

solution is correct. Answers are deemed correct because the *logic and reasoning* are convincing to everyone in the class. If a student presents a proof to the class, then the students in the class peer-review the proof, just as a journal article is peer-reviewed.

Consequently, the pitfalls are related to the fact that most undergraduates have not had many experiences solving problems on their own and thus they can be uncomfortable and not used to being stuck or confused for relatively long stretches of time. It is vital to understand that in most cases students are *in the process of becoming good problem solvers, but they are not yet fully developed*. Students may utter, “I need to be shown how to solve the problem...” or “I don’t learn this way.” Perhaps students will assume that getting stuck means they do not have the capacity to solve problems. The IBL instructor must first diagnose these issues and then address them quickly to avoid a situation where students shut down and give up. Instructors can

- ensure that there exists a wide range of problems so that all students can successfully solve some problems.
- slow down as needed to provide time for students to work on difficult problems. It is common for IBL instructors to augment a problem set with additional problems that provide more scaffolding of a problem.
- utilize small group work to provide a needed problem-solving activity that can generate new ideas and provide students a chance to ask questions and experiment.
- help students by providing sample strategies that can be employed, for instance reminding students that they should look at examples and try to understand and generalize them.
- coach students through the process of building a sense of proof by making sure students know that being stuck is normal and that help is available to them at all times. It can be helpful to share personal stories or stories from famous mathematicians, for example, Wiles’ work on Fermat’s Last Theorem, to remind students that mathematicians often spend very long periods working to understand a single problem.
- ask students who are frustrated, “How can I help you?” and “What have you tried?” Such prompts guide students to refocus attention on the specific math questions they have. Rather than providing a solution, one could outline some ideas that could be helpful, for example, “Has anyone tried looking at these examples or cases?”

If an instructor does not diagnose and manage student expectations and the natural frustration that arises from the problem-solving process, then this leaves open the possibility that students shut down or give up. Instead, an effective IBL instructor provides the support necessary to keep students engaged in the journey through the mathematical landscape. At the same time, the IBL format provides tremendous opportunities for advanced students to grow, as such students discover their own power to do mathematics, and, with a well-crafted problem set, there are sufficient opportunities for advanced students to solve problems that are at a level that engages and challenges them. Indeed, advanced students should be given problems to match their wit, and in mathematics there exists a vast wealth of problems for such students to tackle.

Community of Learners

An essential feature of an IBL course is the community of learners. The specific arrangement and use of the community differs across instructors. Nevertheless, reliance on some form of collaboration between students is one of the principal features of an IBL course. Some instructors do not allow students to work with one another, except during in-class presentations and discussion. Students are otherwise required to work alone on problems outside class. Others encourage collaboration in and out of class. In either of the cases, a core feature of IBL is the sense that a community of learners is working together to investigate and establish facts, and that the differences are relatively minor in the larger

scheme, in which the main idea is the focus on student-generated mathematics.

SOME DETAILS OF IBL COURSE STRUCTURE

Specific use of examinations, written work, presentations, and other assessments differs by instructors. We present some “canonical forms” that instructors can use as a foundation for their own course. One important concept in choosing a course structure is to provide enough structure for the student body at one’s institution. Some students are more independent, while others are still developing their sense of independence. For courses with students who are less independent as learners, it is recommended that the course be more structured.

- Presentations and final,
- presentations, midterm(s), final,
- presentations, midterm(s), final, homework, and
- presentations, midterm(s), final, homework, and portfolio.

Presentations, midterms, and final exams are generally equally weighted, with a tendency for IBL instructors to put the most weight on presentations. Presentations provide the best opportunities for students to demonstrate their understanding of the material, and should be highly valued as a form of assessment of learning. Note that, depending on the course and the instructor, IBL can be a highly collaborative environment, yet individual accountability is the central feature of a successful IBL classroom, and so individual examinations and/or individual presentations are still used in this setting.

Portfolios can be highly beneficial to learning and can take on several different forms. The main idea behind the portfolio is to ensure that students keep a well-organized, comprehensive set of solutions from the entire term. Portfolios are normally checked two to three times during the term. Other tasks can be included in the portfolio, such as requiring students to select their top eight to ten proofs from the quarter, and/or

a summary essay on the main ideas in the course. The portfolio typically is weighted at about 5–10% of the course grade.

Grading presentations is often a task that new IBL instructors have not done before. Sample rubrics are presented here as models for how to grade presentations and provide feedback.

Sample #1

10	Perfect
8–9	Correct with minor technical issues
7	Incorrect with reasoning why incorrect
0–6	Incorrect

Sample #2

4	I know you understand
3	I think you understand
2	I don’t think you understand
1	I know you don’t understand
0	Nothing turned in

Sample #3 (Translated sample #2 for approximate 90-80-70-60 grading scale)

8	I know you understand
7	I think you understand
6	I don’t think you understand
5	I know you don’t understand
0–4	Nothing turned in

Sample #4

1	Correct
0	Incorrect

Class size affects instruction. Most IBL instructors indicate that small class sizes are ideal, and a class size of about 35 students is roughly the upper limit for a full IBL course. As class size increases beyond 35 students, instructors have fewer options for class activities. In this case, instructors can choose a hybrid IBL structure, which diminishes or eliminates presentations from the grading

system. The designation hybrid IBL is meant to connote a classroom in which students are regularly engaged in sense-making activities, but instructors may still rely on lectures for a significant portion of the class. For instance, one day a week may be reserved for IBL activities. Or, group activities may be interspersed at regular intervals within a lecture. This is a natural consequence of the fact that in large classes, students would present only once or twice a term at the most. Thus, it is recommended that instructors who wish to teach a full IBL course do so in courses with enrollments of 35 or fewer. Hybrid IBL courses can be implemented in larger (and smaller) classes, with the understanding that some of the most effective learning environments are being sacrificed.

WHEN IS IBL APPROPRIATE AND HOW DOES IT DIFFER BY COURSE OR AUDIENCE?

We believe that IBL or hybrid IBL methods are appropriate across the curriculum. The specific methods and course structure are largely dependent on the environmental factors and the students in the course. IBL is most easily implemented in upper division or graduate-level courses that are proof-based. Such courses are self-contained, and students in those courses are more mathematically mature. These courses are generally considered to be the best places to implement IBL, particularly for instructors new to IBL.

In freshmen and sophomore-level courses, student backgrounds, experiences, majors, and intended careers vary more widely. Additionally, students have had fewer mathematics courses and thus need to learn more ideas, skills, and habits of mind compared to an upper division student. In courses such as calculus, instructors may not have the option of choosing textbooks or the topics that must be covered. Further, instructors usually do not have as much leeway or flexibility in time as they would in upper level courses. Thus, a hybrid approach is often employed, where the instructor spends less than half of the time introducing a topic, and students work in smaller groups on problems or well-crafted handouts, based on one section of the textbook.

A particularly important group that can benefit from IBL is preservice elementary and secondary teachers. According to The Principles and Standards for School Mathematics [16] and the upcoming Common Core State Standards [17], teachers will increasingly be asked to teach students critical reasoning and higher-level thinking, such as problem solving. It is very difficult to teach problem solving, exploration, how to handle being stuck, and so on, if one has not experienced them firsthand. Thus, we also advocate using IBL in courses for preservice teachers. The flavor of IBL differs in these courses as the course content is not proof-based and axiomatic for all topics. Nevertheless, problems can still be developed from first principles. Further, students can present their findings and work in groups on tasks that require them to explain or justify why. For instance, most students in such courses have never been asked to consider why “invert and multiply” works, why we do not define division by zero, or why subtracting a negative is equivalent to adding.

CONCLUDING REMARKS

In this article, we have discussed what IBL is, what it looks like in a typical course, general features of an IBL approach, and the research findings that support its use. We end this article with some insights from the field. While the data is compelling, what is equally compelling is how IBL has generated a strong, enthusiastic group of practitioners. The reason lies at the heart of education. It is our belief that instructors are ultimately in the business of transforming lives. Despite the extra work needed to learn how to use and continue to implement IBL, instructors who use IBL see growth in students that goes far beyond just learning more facts. We have witnessed, for many students, a transformation of the students’ relationship with, and approach to, mathematics. It is our hope that readers of this article will be inspired to look beyond mere surface features, and invest the necessary time required to become an effective IBL instructor. Assistance for new IBL instructors is available

at The Academy of Inquiry-Based Learning (www.inquirybasedlearning.org).

REFERENCES

1. Seymour E, Hewitt NM. Talking about leaving. Boulder (CO): Westview Press; 2007.
2. Muis KR. Personal epistemology and mathematics: a critical review and synthesis of research. *Rev Edu Res* 2004;74(3):317–377.
3. Schoenfeld AH. When good teaching leads to bad results: the disasters of “Well-taught” mathematics courses. *Edu Psychol* 1988;23:145–166.
4. Selden J, Mason A, Selden A. Can average calculus students solve nonroutine problems? *J Math Behav* 1989;8:45–50.
5. Selden A, Selden J, Hauk S, Mason A. Why can’t calculus students access their knowledge to solve non-routine problems? Research in Collegiate Mathematics Education IV. CBMS Issues Math Edu 2000;8:128–153.
6. Selden J, Selden A, Mason A. Even good calculus students can’t solve nonroutine problems. In: Kaput JJ, Dubinsky E, editors. Research issues in undergraduate mathematics learning (MAA Notes No. 33). Washington (DC): Mathematical Association of America; 1994. p 19–26.
7. Carlson MP. A cross-sectional investigation of the development of the function concept. Research in Collegiate Mathematics Education III. CBMS Issues Math Edu 1988;7:114–162.
8. Harel G, Sowder L. Students’ proof schemes: results from exploratory studies. Research in Collegiate Mathematics Education III. CBMS Issues Math Edu 1998;7:234–283.
9. Coppin CA, Mahavier WT, May EL, Parker GE. The Moore method—a pathway to learner-centered instruction. Washington (DC): Mathematical Association of America; 2009.
10. Smith JC. A sense-making approach to proof: strategies of students in traditional and problem-based number theory courses. *J Math Behav* 2005;25:73–90.
11. Rasmussen C, Kwon ON. An inquiry-oriented approach to undergraduate mathematics. *J Math Behav* 2007;26:189–194.
12. Kwon ON, Rasmussen C, Allen K. Students’ retention of mathematical knowledge and skills in differential equations. *School Sci Math* 2005;105:227–239.
13. Darken B, Wynegar R, Kuhn S. Evaluating calculus reform: a review and a longitudinal study. Research in Collegiate Mathematics Education IV. CBMS Issues Math Edu 2000;8:16–41.
14. Laursen S, Hassi M-L, Kogan M, Hunter A-B, Weston T. Evaluation of the IBL mathematics project: student and instructor outcomes of inquiry-based learning in college mathematics. Boulder (CO): Ethnography & Evaluation Research, University of Colorado Boulder; 2007. [Report prepared for the Educational Advancement Foundation and the IBL Mathematics Centers].
15. Prince M, Felder R. The many faces of inductive teaching and learning. *J College Sci Teach* 2007;36(5):14–20.
16. Boaler J. Open and closed mathematics: student experiences and understandings. *J Res Math Edu* 1998;29:41–62.
17. Boaler J, Staples M. Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Rec* 2008;110:608–645.
18. Schumacher C. Chapter zero: fundamental notions of abstract mathematics. 2nd ed. New York: Addison Wesley; 2000.
19. Schumacher C. Closer and closer. Burlington (MA): Jones & Bartlett Publishers; 2007.
20. Hale M. Essentials of mathematics: introduction to theory, proof, and the professional culture. Washington (DC): Mathematical Association of America; 2003.
21. Burger E, Starbird M. The heart of mathematics: an invitation to effective thinking. 3rd ed. Emeryville (CA): Key Curriculum Press; 2009.
22. Marshall D, Odell E, Starbird M. Number theory through inquiry. Washington (DC): Mathematical Association of America; 2007.
23. Hurlbert G. Linear optimization: the simplex workbook. New York: Springer; 2009.
24. National Council of Teachers of Mathematics. Principles and standards of school mathematics. Reston (VA): National Council of Teachers of Mathematics; 2000.
25. Common Core State Standards, 2013. Available from <http://www.corestandards.org/>. Accessed March 15, 2013.

AN OVERVIEW OF OPERATIONS RESEARCH IN TENNIS

GEOFF POLLARD
Faculty of Life and Social Science,
Swinburne University of
Technology, Melbourne, Victoria,
Australia
Tennis Australia, Melbourne, Victoria,
Australia

DENNY MEYER
Tennis Australia, Melbourne, Victoria,
Australia

The modern game of “tennis” began in 1874 as “lawn tennis,” as an outdoor version of the ancient game now known as *royal tennis*. In nearly 100 years, the equipment and the rules hardly changed. Originally designed as a portable game to be played on grassed areas using balls and stringed rackets made of wood, the courts soon became permanent and the surface extended from grass to clay and then, to various hardcourt surfaces. Rackets and balls continued to improve in quality, but remained essentially the same. The three-tiered scoring system of points, games, and sets remained consistent throughout, along with the principle of winning by at least two points (in a game) and at least two games (in a set).

In 1968, the long-standing division between amateurs and professionals ended and tournaments were opened to all players. The game expanded rapidly. The equipment and the scoring system that had served the amateur game for so long were repeatedly thrown into question by technical developments and by the demands of television coverage.

Scientific and technological developments both created problems and opportunities for the game. This article considers some of these developments, highlighting the current and potential use of operations research (OR) and management science (MS) methods in this regard.

CURRENT AND POTENTIAL USE OF OR/MS METHODS IN TENNIS

The following table illustrates how OR/MS methods are currently used in tennis. However, modern technological developments mean that there are now ways to measure many more of the relevant variables, opening up many new uses for these methods. Under a range of topics, Table 1 illustrates various decision problems which can be addressed by OR/MS methods using appropriate objective functions and constraints. These topics are discussed in detail in this article.

DEVELOPING ALTERNATIVE AND OPTIMAL SCORING SYSTEMS

The scoring system used in the modern game of (lawn) tennis is a much simplified version of the ancient game of royal tennis given the absence of the “chase,” the walls, galleries, and other hazards of royal tennis. When Major Walter Wingfield patented the new game (initially called “sphairistike”) in 1874, he patented both the rules and the equipment (a patent which he subsequently allowed to drop). Soon there were many variants to the rules and in the absence of any ruling body, the Marylebone Cricket Club (MCC) was called upon to adjudicate on the rules of tennis. However, the MCC only set guidelines with some variations allowed. The Rules Committee of the first Wimbledon Championships in 1877 adopted a set of rules which then became the standard and have remained fundamentally the same to this day.

Tennis has a unique scoring system, in that it is triple nested (points, games, and sets). One player serves and their opponent receives and the server has two chances to put the ball into play. The same player continues serving until a game is completed, which is won by the player winning the “best of six” points (i.e., first to four points) provided that player leads by at least two points. In the following game, the receiver becomes the server and the players continue to swap roles each

Table 1. Current and Potential Use of OR/MS Methods in Tennis				
Topic	Objective Function	Constraints	Decisions	Appropriate OR/MS Methods
Developing optimal scoring systems	Minimize the mean and variation in length of matches	Integrity of tennis' unique scoring system, better player should win	Alternative procedures and scoring systems	Deterministic optimization using mathematical models
Developing optimal match strategies	Maximize the probability of player success	Surface properties (e.g., court speed), current scoring system	Serving strategy, When to lift	Deterministic optimization, Markov chain and dynamic programming models
Surface optimization	Minimize capital and maintenance costs Maximize income from court hire and the chance of winning	Weather, player preferences in terms of speed and comfort	Choice of surface	Deterministic optimization, Multicriteria optimization
2 Optimizing ball characteristics	Optimize ball speed, spin, trajectory, durability, bounce, etc. Minimize cost effects and maximize ease of play	Surface and racket properties; player ability, style and age; cloth texture	Internal pressure or pressureless dynamic stiffness	Stochastic optimization using mathematical modeling and computer simulation
Optimizing racket characteristics	Maximize ease of play and minimize cost effects Optimize ball trajectory and power for any stroke	Preservation of the nature of tennis, ball, racket and surface properties; racket material, atmospheric conditions	Head size, weight, and grip size; frame stiffness, balance point, sweet spots, swingweight, stability, shock, power, vibration, feel	Stochastic optimization using mathematical modeling and computer simulation
Optimization from a coaching perspective	Optimize performance, minimize injury risk	Physical and mental attributes of player; ball, racket, and surface characteristics	Technical advice and short/long-term training (periodization) plans	Pattern recognition using video footage; deterministic optimization

ω	Optimizing playing technique	Optimize power and control combination	Equipment, surface, athletic, and physical attributes of player, injury risk	Choice of style and technique for individual players	Multicriteria programming
	Minimizing medical risk	Minimize injury risk, doping detection Optimize rehabilitation and training	Player characteristics, opponent's strength; ambient conditions, other risk factors	Tournament rules (e.g., minimum age of players, heat policies, racket design)	Deterministic optimization using mathematical models
	Efficient tournament management	Maximize the number of feasible matches	Availability of courts and players, Maximum time for matches	Match schedule	Constraint programming with heuristic search
	Strengthening the mental side of the game	Minimize variation in playing ability through the control of heart rhythms and emotion	Physical and mental attributes of player; ambient conditions, surface, other risk factors	Choice of techniques for maintaining concentration (e.g., breathing, stalling, grunting, variation in play)	Deterministic optimization using mathematical models
	Accurate player rankings	Maximize accuracy of rankings in terms of performance	Different levels of tennis, singles/doubles, surface, effects of injury	Weightings for importance and closeness of a match, strength of opponent, protective rankings, average or "best of x " results rules	Kalman filter, Monte Carlo simulation, nonlinear optimization
	Improving officiating	Optimized accuracy of line calls and foot faults	Technology used, surface, human error	Choice of method, model, training of officials	Pattern recognition using video footage, deterministic optimization

game. The set is “best of ten” games (i.e., first to six games) provided the player leads by at least two games (called an *advantage set*). A match consists of the “best of three” sets (i.e., first to win two sets) or in Grand Slam men’s tournaments and Davis Cup, the “best of five” sets (i.e., first to win three sets).

The first significant change of the open tennis era intended to help reduce the mean and variation in the length of matches, was to introduce the option of a “tie-break set” instead of an “advantage set.” The tie-break system was introduced by James Van Allen at the Philadelphia Indoor Tournament in 1970 and was the best of nine points. The fact that both players had a match point at four points all, but one player had the advantage of serving soon lead to a second tie-break system, the 12- point method described below. This tie-break system soon became the more popular system and was used at Wimbledon in 1971 (at eight games all, whereas most other tournaments applied it at six all).

The Rules of Tennis [1] published by the International Tennis Federation (ITF) now provide that, if the score reaches six games all, the set can be decided by playing a tie-break game where the first player to win seven points wins the game and the set, provided there is a margin of two points within the tie-break game. If necessary the tie-break game continues until this margin is achieved. Thus, although the duration of a tie-break set is still unbounded, it clearly has a lower mean and variance than the advantage set. The tie-break set is now used in all sets in all tournaments except the Australian Open, Roland Garros, Wimbledon, Davis Cup, and Fed Cup which maintain that the fifth and deciding set (third set for women) should be decided by an advantage set (i.e., requiring a break of service).

Tennis, pushed by the demands of television, searched (and still searches) for a solution to the reasonably unpredictable length of tennis matches, while maintaining the integrity of its unique scoring system. Short matches were accepted, but extremely long matches, primarily brought about by the dominance of the service, became increasingly difficult to handle. Appendix IV, of the Rules of Tennis [1] offers alternative

procedures and scoring methods such as no ad scoring, short sets, match tie-break to replace the deciding final set, and eliminating the let during service. These alternatives have long been used by social players and in selected competitions such as World Team Tennis and Intercollegiate tennis, but were only formally recognized in the ITF Rules of Tennis from 2002. In the forward of Rules of Tennis, the ITF [1] invites interested parties to submit applications to officially trial other scoring systems.

There is a growing collection of computer simulation, probability, and statistical analyses of the official scoring system and alternative more efficient scoring systems. The simplest stochastic models (unipoints) assume all points are independent and identically distributed. A more realistic but more complicated model for elite tennis recognizes there are two types of points depending on which player is serving (bipoints).

Kemeny and Snell [2] modeled a single game of tennis using a Markov chain. Schultz [3] examined the feasibility of using a Markov chain with stationary transition probabilities to evaluate scoring systems. Hsi and Burych [4], taking a bipoints approach, evaluated the probability that a player wins an advantage set. Carter and Crews [5], taking a unipoints approach, found the expected number of points in a game, games in a set and sets in a match. Miles [6], using the bipoints approach, found the expected duration of a set in tennis. Pollard [7] showed mathematically that the expected duration and variance of a tie-break set and match are always smaller than using advantage sets, but the probability that the better player wins is slightly reduced.

Morris [8] introduced the concept of importance for each point played and the more complex notion of “time-importance” where the importance is weighted by the expected number of times the point is played. Miles [9] noted the link between sports scoring systems and symmetric sequential statistical hypothesis testing and examined the efficiencies of sports scoring systems such as tennis. Pollard [10,11] linked Miles’ and Morris’ work and found an important relationship between importance of points within a system and the

efficiency of the system itself. He uses this relationship to find the optimally efficient tennis scoring system. Pollard [12] showed how, in many cases, the relative efficiency of two different scoring systems could be found without calculating the efficiency of each system separately. Pollard and Noble [13] suggested the 50–40 game and showed that it leads to good scoring systems in terms of variance and efficiency, particularly for doubles. Pollard and Pollard [14] generalized the Miles formula from singles (two parameters) to doubles (four parameters) and identified very efficient doubles scoring systems. They also found the moment-generating function for the present three set tennis scoring system [15].

DEVELOPING OPTIMAL MATCH STRATEGIES

Croucher [16] summarizes much of the literature on developing strategies within current scoring in tennis and separates his analysis into firstly, serving strategies (given you have a first and second serve), secondly, the probability of winning a game, set or match, and thirdly, the analysis of match data. All involve mathematical and OR techniques.

Gale [17] used a simple mathematical model to identify the best first serve and the best second serve from a set of serves. Redington [18] concluded that in the 1971 Wimbledon final, Newcombe would have done better if he had served two first serves rather than a first and second serve. George [19] used a simple probabilistic model to show that a “strong” first serve followed by a weaker second serve was generally, but not always, the best strategy. Using considerably more data, King and Baker [20] came to the same conclusion. Hannan [21] proposed a game theory approach that factored in the opponents return while Norman [22] used dynamic programming to decide when a first serve should be used. With even more data, McMahon and de Mestre [23] also confirmed that the “strong weak” serve pattern was usually the best but were surprised at the number of matches where an alternative was better. Barnett and Clarke [24] used server and receiver statistics to predict strategies

and outcomes in any match between two players. Barnett and Pollard [25] adjusted this analysis for the effect of court surface. Pollard and Pollard [26] and Pollard [27] looked at the mathematical relationship between the probability a serve goes in and the probability a player wins the point and thus determined the optimal strategy for first and second serves. Barnett *et al.* [28] used the large OnCourt database (www.oncourt.info) to calculate match statistics for each court surface and thus determine service strategies to improve performance.

There has been much research over the years, including work by Fischer [29], Pollard [7], and Croucher [30] on the probability of winning a game, set (advantage and tie-break), and match where the probability of winning a point is constant for each player and each point is independent. For example, Croucher [30] calculated the probability that the server held serve from each of the 16 possible score lines. Morris [8] introduced the concept of the importance of a point defined as the difference between the two conditional probabilities that the server wins the game if he wins that point and the probability that he wins the game after losing that point. Morris showed that increasing the effort on the more important points and decreasing effort on the least important points increased the probability of winning. Pollard [10] expanded and generalized this analysis of importance and Barnett [31] showed numerically when a player should lift (increase) his/her probability of winning a point. Using 10 years of Grand Slam men’s singles data, Pollard, Cross and Meyer [32] showed that the better player had the ability to lift his play. Pollard and Pollard [33] used differential calculus to identify the points, games or sets when the player gets the most reward for lifting. All these mathematical analyses assume each point is independent of the previous points. However, using 4 years data from Wimbledon 1992–1995, Klaassen and Magnus [34] showed that points in tennis are neither independent nor identically distributed. Nevertheless, they claimed that these assumptions are “sufficient as a first order approximation,” which is similar to

Croucher's [16] comment that these assumptions "provided some interesting, if debatable, conclusions."

Brimberg *et al.* [35] addressed the problem of how to allocate your limited energy over (say) a first-to-three (best of five) set match. Using both probability and dynamic planning, they showed that where there are only two choices (high and base energy) it does not matter when the high energy is expended (although it does affect the expected length of the match). However, where there are three or more energy choices the player who is behind should divide his energy evenly over the remaining games but for the player who is ahead, it is best to divide his remaining energy between high and low rather than evenly.

Selected statistics (aces, double faults, winners, errors, total points) were often collected for selected important matches such as Davis Cup Challenge Rounds but were primarily for media use. For these and some of the earlier analyses above, data was laboriously collected by pencil and paper by someone watching the match or a video replay. In 1982, Bill Jacobsen began to record his son's matches on a micro computer. This soon developed into a system called *Compu Tennis* which became a useful tool for coaches. Today, the umpire records each point live into the computer and organizers record additional statistics at all Grand Slam matches and many matches on the Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) Tours, although only limited data is published and access to the data is restricted. By accessing Wimbledon data 1992–1995, Magnus and Klaassen (summarized in Ref. 36) were able to test and mostly refute some of the common hypotheses in tennis such as: any advantage in serving first, the first use of new balls, who has the advantage in the final set in a five set match, and even forecasting the winner before and at various stages during a match. The fact that their data is still being utilized demonstrates the limited access for researchers to officially collected data, although individual matches are provided to the media and selected match

statistics are available on the OnCourt database.

SURFACE OPTIMIZATION

Originally designed to be played on flat, manicured English lawn surfaces, British travelers soon took the game to Europe and the colonies and it spread quickly. Where suitable grass was not available, enthusiasts simply played on gravel, sand, asphalt, concrete, brick dust, and other surfaces. As technology developed, an increasing range of synthetic hard and cushioned acrylic surfaces as well as artificial grass and artificial clay surfaces appeared. The oldest and most prestigious tournament, Wimbledon (1877) has always been played on grass. The US Open (1881) was played on grass until 1974 when it was briefly played on clay, but was changed to hard court in 1977 at the request of the US players. The Australian Open (1905, although the Victorian Championships date back to 1880) was also played on grass until 1988 when it changed to cushioned acrylic (Rebound Ace) and since 2007 to cushioned hardcourt (Plexicushion). The French Open (1891) has always been played on clay (crushed brick) but was restricted to French players until 1925. The importance of these four Grand Slam tournaments ensures most other tournaments and venues use the same or similar surfaces.

Further, court surface has a very significant effect on the nature of the way the game is played, so that, for example, selection of court surface is a prized home ground advantage in the Davis Cup. Selection of court surface for a tennis facility is a multicriteria optimization problem where the capital cost of construction, maintenance costs, and player preferences in terms of court speed and comfort are some of the factors to be considered. Tennis Australia has introduced a Court Rebate Program to encourage clubs to lay the surfaces used at the Grand Slams rather than the popular synthetic grass surface.

The rules of tennis were effectively silent on court surfaces, but to allow quantification

of the speed of a court surface, the ITF has developed a Court Pace Classification Program [37]. This system divides court pace as measured by the Court Pace Rating (CPR) into five categories:

Category 1 (slow pace)	CPR 0–29
Category 2 (medium slow)	CPR 30–34
Category 3 (medium)	CPR 35–39
Category 4 (medium fast)	CPR 40–44
Category 5 (fast)	CPR 45 and over

and nine court surface types:

A	Acrylic
B	Artificial clay
C	Artificial grass
D	Asphalt
E	Carpet
F	Clay
G	Concrete
H	Grass
J	Other (e.g., modular tiles, wood, canvas)

Currently, 57 different court surface suppliers are recognized by the ITF and their products cover 11 (Category 1), 15 (Category 2), 43 (Category 3), 38 (Category 4), and 26 (Category 5) surfaces. However, few companies laying clay or grass courts register as their products are inherently variable, although in general they are recognized as slow and fast respectively.

Barnett and Pollard [25] analyzed the performance of players on grass, hard, and clay courts used on the ATP and WTA Tours, and how the changing number of tournaments on each surface, especially the decline in grass courts, affects player rankings and also injuries. Only Wimbledon (grass) modifies official rankings when doing seedings

for its Championships. Rankings and medical risks are discussed later, but both would benefit from further OR analysis factoring in the effect of court surface.

The development of a court pace classification program has been an interesting technical, scientific, and statistical exercise [38], especially since 2008 when the ITF introduced rules to limit the ability of home nations to select extreme surfaces for the Davis Cup Competition. However, player comfort has not been subject to any attempt at classification.

OPTIMIZATION OF BALL CHARACTERISTICS

Compared to the technological development of the golf ball over the past 100 years, the tennis ball has remained almost unchanged except for the development of the pressurized can, to ensure longer shelf life, and the change in color from white to yellow. Pressureless and high altitude balls are also available which differ from the “standard ball” by virtue of their internal pressure and bounce height.

Over the last decade, the ITF has tried to address the issues of court speed and style of play by encouraging manufacturers to make three types of balls, where ball type one (fast speed ball) is designed for use on slow paced court surface; ball type three (slow speed ball) is designed for use on fast paced court surfaces, while the standard ball type two (medium) is designed to be used on medium-slow, medium and medium-fast surfaces (i.e., the majority of courts). While there is clear evidence that using the appropriate ball on the relevant surface makes the game easier to play (for both players), the use of balls other than ball type two is limited.

Brody *et al.* [39] have a number of chapters devoted to the tennis ball and the testing of ball properties for perfect bounce, ball spin, trajectories, and similar technical characteristics. All ITF ball testing measures the properties of new balls only. The ITF Technical Department has now revolutionized ball approval by introducing the testing of ball durability into the process. A key challenge has been to accurately simulate in the laboratory the effects of play on the ball [40].

Another development is the comparison between the static stiffness of tennis balls as measured in ITF ball testing and the dynamic stiffness in real play [41], the potential outcome of which is the introduction of approval tests that are more representative of player perceptions of ball properties.

Special tennis balls have been developed to make it easier for starter players, especially 10 and under, to play the game and to position tennis as easy and fun. These introductory balls are color coded into red (stage 3), orange (stage 2), and green (stage 1) and are officially tested for approval by the ITF.

OPTIMIZATION OF RACKET CHARACTERISTICS

For nearly 100 years, tennis rackets were made of wood and generally strung with strings made from sheep's gut. Aluminum frames appeared in the 1960s leading to the larger head size in the 1970s. Aluminum was replaced with graphite and then by composite rackets made mainly from carbon fiber reinforced with graphite. Titanium and more recently ceramic fibers, boron, and other products, in many cases, outcomes of the research and development in the aerospace industry, have also been used.

Comparing modern carbon fiber composite rackets to the best wooden rackets at the beginning of the open era shows they are much lighter (180–350 g compared to 370–425 g), longer (68–73 cm compared to 66–68 cm) with a larger head size (580–870 sq. cm compared to 420–450 sq. cm). Amongst the top 20 male players, Warinka uses a 271-g racket while the average racket weight is 320 g. Amongst the top 20 women players, the Williams sisters use 270-g rackets while the average racket weight is 305 g.

The modern composite rackets have a greater “sweet spot” and are considerably more “powerful.” The term *sweet spot* is the commonly accepted term for the area on the racket face where the impact feels best, generally the node point since there are no vibrations. In fact, the “sweet spot” can be any of three spots; the center of percussion

where shock (i.e., acceleration) of the racket is minimized, the node where vibration is minimized and, finally, the location of maximum ball rebound speed, although the impact at this point does not necessarily feel “sweet.” On service, top players serve from near the top of the racket due to the added height advantage and the faster racket speed at the top.

Inherently, the material of which modern rackets are made are more powerful than wood; in that, less energy is lost (due to bending) during impact. On the assumption that a player puts constant energy into a stroke, a modern (lighter) racket will have a greater impact speed and will therefore generate a higher ball speed, which is used as the definition of “power.” In theory, lighter rackets should be less powerful than the older heavier rackets. But swing speed (inversely related to mass) is more important, so you can generate more ball speed with a lighter racket. Stiffness of racket material has an effect, but not to the same extent as mass and swing speed.

At the social level, larger rackets have made the game easier to play by being more “forgiving” to off-centre impacts. At the elite level, they have changed the way the game is played (technique) and professional tennis is now a considerably faster and different game as compared to that of the amateur era.

Strings and stringing have also contributed to the development of the modern racket. Until the 1940s, all rackets were strung with gut strings but gradually nylon strings appeared, although nylon was inferior to gut in the wood rackets. The development of the modern composite rackets and the considerable improvement in strings with nylon, polyester, kevlar, and other substances means that, today, there is little difference between different string types and all are used by the top professional players, with polyester being the most preferred string followed by gut and then nylon. Polyester strings do generate more spin [42].

But it was the method of stringing that led the ITF to finally act in 1978 on the definition of the racket when “double stringing” or “spaghetti stringing” allowed some players to

generate so much spin that the ball was virtually unplayable. Such stringing was banned. In the 1980s, the increasing size of the racket led the ITF to finally limit the size of a tennis racket in the Rules of Tennis and to define the parameters for stringing patterns.

In the 1990s, the ITF established a research laboratory at its headquarters in London with the analysis of rackets and strings and the increasing power they can generate being a major issue for ongoing research and for the possible introduction into the rules of tennis [43].

Brody *et al.* [39] have a number of chapters devoted to the tennis racket and the technical measurement of relatively simple characteristics such as head size, weight, and grip size, through to the more complicated characteristics of frame stiffness, balance point, sweet spot, swing weight, stability, shock, vibration, feel, and power.

As part of its role to preserve the nature of tennis, the ITF Science and Technical Department has developed a software program which simulates the effects of ball, racket and surface properties, and atmospheric conditions on ball trajectory for any stroke. Not only is the software (known as *Tennis GUT*) able to establish the effects of any combination of equipment on the nature of the game, but it also allows the user to modify the properties of that equipment. Trends in equipment design and properties can be used, for example, to predict future properties and, as a consequence, to predict how these trends may affect the way the game is played.

OPTIMIZATION FROM A COACHING PERSPECTIVE

Coaching is the method by which one generation passes on its considerable knowledge, experiences, and expertise to the next generation, thus enabling the game to grow and develop. In tennis, coaches play a significant role at every level from introducing youngsters to the game and teaching them the basic rudiments of how to play, through to helping the world's best player to become even better and retain his/her number one

ranking. This includes analyzing each opponent to detect any potential weakness that might be exploited.

Traditionally, tennis coaching involved either past champions describing their experience and how they played the game or other teachers, who may not have had the same on-court success, but who appreciated and understood the style and technique of the champions and had the ability to describe and teach it. While there is still a role for the traditional coach, the modern coach also makes considerable use of all the sports science and technology that have been introduced into the game over recent decades.

Coaching courses today include a significant component of sports science and technology and coaches are encouraged to access the ITF publications and the ITF coaching website www.tenniscoach.com to see the extent of modern technology available and to engage in on-line learning or upgrading. The ITF has recommended syllabi for level one, two, and three coaches that are available in a number of languages and which are now used in over 80 countries.

Many coaches use video players to record their pupil in practice or match conditions and then use software such as Dartfish to analyze their game. Other match and player analysis programs commercially available include the Swinger video analysis, Chart-Mate Pro, and Stats Master video analysis. Player match statistics are available on the OnCourt website www.oncourt.info allowing players and coaches to analyze their performance and that of their next opponent.

For running their businesses, coaches use specifically designed commercial programs such as Tennis Biz, Tennis Logic or Software Management Solutions. For marketing and communications purposes, there is increasing use of SMS, e-mail, and in some cases, their own website.

National associations use products such as Athlete Management System to track all players in their academies. Coaches record their analysis following matches, practice sessions, training regimes, technical advice, injury, periodization plans, communication

with other coaches, and even video footage and analysis.

Using Wimbledon singles data 1992–1995, Klaassen and Magnus [44] converted player rankings into the probability a particular player won a match between two players. They then recalculated that probability as the match progressed. They suggested such information was most useful for television commentators rather than coaches, but obviously also for punters and betting agencies. For further probability predictions, see Bedford and Clarke [45] and Barnett and Clarke [46].

OPTIMIZING PLAYER TECHNIQUE

Biomechanics is the study of human movement. In tennis, this involves the search for the optimum playing technique, the most efficient and effective combination of power and control in both stroke and movement while minimizing the risk of injury.

Traditionally, coaches performed this task simply by observing the world's best players and then trying to teach these observed techniques to their pupils. But tennis is such a multidimensional sport. It is played by men and women, young and old, clumsy and talented, serve/volley players and baseliners and with ever changing equipment on a wide range of surfaces from clay (slow) to hard courts (medium) to grass (fast). There is no single optimum way to play and champions have displayed a wide range of styles and techniques. However, modern technology has made it easier to describe body movement, to quantify movement, and to compare players' style.

Elliott and Reid [47] give a good summary of the wide range of descriptive and objective technologies now available to the coach or the tennis biomechanical researcher. Generally they involve one or more high speed video cameras linked to a computer loaded with one of the increasing range of sport analysis software packages. However, they repeat the warning of Elliott and Knudson [48] that any recommendation by a coach to change a player's game first needs a comprehensive assessment of a wide range of other player characteristics.

For a complete study of the biomechanics of advanced tennis with respect to the effectiveness and efficiency of a player's on-court movement and stroke production, and the relationship with performance enhancement and injury prevention, the reader is referred to the ITF publication, *Biomechanics of Advanced Tennis* [49] and its wealth of references.

MINIMIZING MEDICAL RISK

Sports medicine has become a legitimate and recognized discipline within medicine. Basically, it involves the prevention of athletic injuries or their detection, treatment and rehabilitation. It is closely linked to the other sports sciences, although in most countries only medical practitioners are licensed to approve drugs as part of the treatment. There are now doctors who specialize in injuries specific to tennis. The WTA and ATP are in the process of introducing an injury registration database which will provide accurate data on tennis injuries in the future, leading to opportunities for statistical and OR analysis not previously available.

The best known injury in tennis is the so-called tennis elbow. Many people acquiring tendonitis of the elbow from other causes unrelated to tennis will still call the injury "tennis elbow." The different injuries causing elbow pain and their diagnosis, treatment and rehabilitation are described by Renstrom [50].

The International Olympic Committee and the International Tennis Federation, together with the Society for Tennis Medicine and Science and the ATP and WTA Tours, have combined to produce a *Handbook of Sports Medicine and Science in Tennis* [51]. The nature of the game means that injuries occur throughout the whole body with shoulder, back, ankle, and knee being the most common sites of injury, but these are generally minor injuries that respond well to treatment. The majority of injuries are due to microtrauma repetitive overwork. Macrotrauma, such as sprains or fractures, does occur in a minority of cases and are difficult to prevent with training activities.

Although injuries do occur, Pluim *et al.* [52] have conducted an extensive literature search of the health benefits of tennis and concluded that “people who choose to play tennis appear to have significant health benefits, including improved aerobic fitness, a lower body fat percentage, a more favorable lipid profile, a reduced risk for developing cardiovascular disease, and improved bone health.”

Other aspects of sports science and medicine currently undergoing research, analysis, progressive introduction or modification, include minimum age limits for boys and girls for international tournament play, continuous registration of injuries and their treatment for international players, measurement of heat and humidity and appropriate rules affecting play in extreme conditions, and, finally, antidoping. Data now being collected will provide considerable opportunities for research on each of these aspects.

Tennis has an Antidoping Program [53] to maintain the integrity of the sport and to protect the health and rights of all tennis players. The ITF is a signatory to the World Antidoping Code. Any player who participates in an event organized, sanctioned, or recognized by the ITF, ATP or WTA Tours is subject to In-competition and Out-of-competition testing under the program.

Statistics is commonly used in “Tennis Medicine.” For example, Hatch *et al.* [54] report an experiment in which 16 asymptomatic Division I and II collegiate tennis players performed single-handed backhand ground strokes with rackets of three different grip sizes (recommended measurement, undersized 1/4 in., and oversized 1/4 in.).

However, Pluim *et al.* [55] have concluded that there have been few longitudinal cohort studies that have investigated the relationship between risk factors and injuries in tennis. This, together with the complete absence of randomized control trials concerning injury prevention, means that this is an area that is desperately in need of more rigorous analysis in the future.

Other useful references to sports medicine in tennis include Pluim and Safran [56] and Petersen and Nittinger [57] while coaches

are referred to the ITF Manual on Tennis Medicine for Tennis Coaches [58].

EFFICIENT TOURNAMENT MANAGEMENT

Most components of the work required for the running of a successful tournament have been gradually converted from long hours and substantial paperwork to computer assisted administrative systems such as CAT (Computer Assisted Tournaments), TMS (Tournament Management Systems), and TP (Tournament Planner). As explained by Della Croce *et al.* [59] the goal is to maximize the number of feasible matches, given the constraints in terms of court and player availability.

The world’s first tournament to be managed with the aid of computer software (CAT) was in Hay, Australia in October 1980. This first version of CAT was designed to produce the daily schedule for each player competing in up to five or six events in a 3-day country tournament. The current version handles virtually all aspects of tournament management.

In the 1990s, the Americans introduced a similar product called TMS for which the US Tennis Association purchased a nationwide license. Many other countries followed the United States making TMS the most popular program for tournament management for some time.

The most recent tournament management software, TP, was introduced by the Dutch and adopted in 2007 by the ITF for management of all its tournaments worldwide. Its wide range of features and flexibility means that it has been adopted for small tournaments through to Grand Slams (e.g., Australian Open 2009).

The ITF has recently introduced a compulsory International Player Identification Number (IPIN) and an associated on-line player entry system that, in 2009, covers tens of thousands of players in well over one thousand tournaments on the ITF Junior Circuit, ITF Pro Circuit, and ITF Seniors Circuit. This allows the on-line scheduling of a worldwide calendar of tournaments, permitting players

to easily select their most appropriate tournament at any time according to their ranking and the entry decisions of other players.

STRENGTHENING THE MENTAL SIDE OF THE GAME

Sports psychology deals with the mental side of tennis and covers the study of human behavior within the context of the game of tennis. Many matches are won or lost because of the mental differences between the two players rather than any technical or physical difference.

The role of psychology in tennis is summarized in the comprehensive book *Tennis Psychology* by Crespo *et al.* [60] including over 200 references on playing with confidence, playing in “the zone,” “choking,” lack of concentration, burnout, motivation, fear of losing or winning, or simply how to play matches the same way as you play in practice. This book published by the ITF gives over 200 on-and off-court psychological training techniques, drills, and activities to strengthen the psychological approach of players to match play.

But as noted by Forzoni [61] and Calstedt [62], because it deals with peoples’ minds, sport psychology does not have the same degree of objective measurement as the physical and technical side of the game. But both these authors also recognize that with technological advances, decreasing costs, and decreasing size of equipment, along with the ease of use of biofeedback and neurofeedback devices, it is becoming easier to measure and monitor over time the effectiveness of specific mental skills strategies. This will lead to opportunities for statistical and OR studies to minimize fluctuations in performance during a match and to optimize mental skills.

Emotions being experienced during a game of tennis have an effect on the human body and some aspects can be measured. Forzoni [61] strongly recommends the Heart Math Freeze Framer which measures an athlete’s heart coherence under various situations, so that with various relaxation techniques (controlled breathing, visualizing positive images) you can learn to control

your heart coherence and consequently your emotions. Calstedt [62] suggests that the most reliable indication of psychological states in athletes is heart rate variability, or heart rhythms, which can be measured by electrocardiograms or pulse wave recordings. He recommends that players should learn techniques to control their heart rhythms and consequently their emotions. He has developed the Calstedt Protocol as one possible approach for athlete psychological assessment, mental training, and interaction efficiency.

ACCURATE RANKING OF PLAYERS

In 1973, the leading male tennis players formed their own union, the ATP, and one of their first acts was to introduce a 12-month weighted moving average computing rankings system to determine fairly which players gained entry into tournaments worldwide and to determine which players were seeded. A separate doubles ranking was introduced in 1976. The WTA was also founded in 1973 and introduced its computer rankings system in 1975.

Prior to this time, entry into tournaments and seeding was done by the tournament director or tournament committee. National rankings were compiled annually by national tennis associations. World rankings were generally announced by various journalists so there were many unofficial world rankings and no official world ranking.

Under the original ranking systems, tournament importance or “strength” was determined by prize money and player performance was measured by the round reached. A schedule of points was agreed based on the above and a player’s ranking was calculated as the average points earned for tournaments played in the previous 12 months. Utilizing the capacity of a computer to create a more rigorous ranking, Musante and Yellin [63] were able to rank the strength of tournaments by the ranking of all players entered in that event and, also, to measure performance using the ranking of defeated opponents, not just the round reached. The concept of bonus points for defeating a higher

ranked player was used by the WTA for some years, but was subsequently discontinued. The ATP publishes a calendar year race to the ATP finals as well as the normal ranking on performances over the past 12 months.

Subsequently, Blackman and Casey [64] recognized that previous rankings were useful for determining tournament entry for players and for tournament seedings, but not for determining match result probabilities, betting odds, and equitable handicapping methods. Using the actual scores in all matches between the players being ranked, they developed a rating in numerical units similar to a golf handicap. The difference in these rating units for any two players was shown to be a very good indication of match result probabilities and could also be used to determine what handicap should be given to the weaker player to make the match more even. See also Klaassen and Magnus [44], Bedford and Clarke [45], and Barnett and Clarke [46] discussed earlier under the section on coaching.

In 1989, the ATP players union combined with the tournaments (excluding ITF and Grand Slams) to form the ATP Tour. Likewise in 1995, the WTA Players Association combined with the Women's Tennis Council (including ITF and Grand Slam) to form the WTA Tour. Since then, both computer rankings have rewarded quantity as well as quality by selecting a player's best results from a minimum number of tournaments, with the ATP Tour and WTA Tour, and the ITF for junior and senior rankings all using different formulae. Only the ITF uses one ranking for each player based on both singles and doubles matches.

Mathematicians could clearly develop a ranking or rating system superior to any of those currently used by tennis governing bodies but the players and tournaments prefer simplicity and are reluctant to make changes to their current systems.

IMPROVING OFFICIATING

A game of singles in tennis involving two players can involve no umpire, one central chair umpire or up to 12 officials (center

chair plus up to 11 lines persons) making tennis one of the most highly officiated sports. The largest tournaments, the Grand Slams, require a team of 350 officials, gradually reducing in numbers as players are eliminated.

Clay courts have always had the advantage of the ball leaving a mark which can be examined if there is a dispute over a line call. Ball mark inspection procedures are included as an Appendix to the Rules of Tennis [1].

The first technical development to assist line calling was the "Cyclops" machine which served tennis for 20 years. It was only used on the service line and involved sending beams of light either side of the service line to detect whether the service was in or out of play.

The first attempt to cover the whole court was the TEL machine which involved an undetectable modification to the ball (iron particles in the cover) and placing wires in the court on either side of the line to detect whether a ball that bounced close to the line was either in or out. It was successfully used for one year at the Hopman Cup with a center chair and no linesmen, but the cost, standby arrangements (still needed to keep linemen on call in case of failure), and the need to dig up the court for installation, limited its use.

The Auto-Ref and Hawkeye optical tracking system takes a different approach to line calling. These systems track the flight of a tennis ball with software being used to map the point of impact relative to the lines. As explained by Szimák *et al.* [65] the software uses pattern recognition and other algorithms to find the ball when it is hit over the net, to convert 2D to 3D images, estimating the ball trajectory and predicting where the ball will intersect with the court surface.

Originally introduced as an aid to television coverage, by increasing the number of cameras and other technical improvements, the ITF Technical Department gave approval for Hawkeye's use for officiating in tournament play. Electronic review procedures are also included as an appendix to the Rule of Tennis [1].

Research continues into other technical systems to assist in-line calling including

heat sensors (infrared cameras recording the heat generated when a ball bounces).

Mathematical modeling and statistics have been used extensively in the development of the above and other line-calling systems. For example, Jonkhoff [66] used experimental data to show that electronic line-calling using the TEL system is consistent and accurate and also cheaper than line umpires. Marshall [67] showed that the multibeam LineHawk provided superior accuracy than a single master beam. However, Collins *et al.* [68] have expressed concerns at the way Hawkeye decisions are naively accepted although the technology is certainly not perfect. They suggest that measurement error needs to be made salient and that confidence levels should be attached whenever ball impacts are reconstructed. The average accuracy is about 5 mm but players accept the machine-made decision over a human decision. Mather [69] has used Monte Carlo simulation to show that, for the best-fitting space constants, the condition for a challenge (player and umpire disagree) was met in 20.8% of all trials. In 39.6% of the simulated challenges, a line judge error was recorded, which is very close to the 39.3% of errors found in the actual challenge records.

At www.freepatentsonline.com/5908361.html details of a new line-calling invention (Automated tennis line-calling system: US Patent 5908361) were reported on March 20, 2009. This invention uses mathematical models to merge information from loudspeakers, ball impact sensors close to boundary lines, and pressure sensing devices to make line calls and foot fault calls. Foot faults are detected by comparing (i) the time of occurrence obtained from signals induced by contact of the serving player's foot on a pressure sensor at the baseline with (ii) the computed time of occurrence of racket contact with the ball as derived from the racket sounds received by three or more microphones. This system can also be used to monitor the progress of play and to keep score using the sequence of locations of ball bounces, net-cord hits, service foot faults, and racket hits. In particular, player statistics such as ball speed on service,

winners, errors, and other measures of player accuracy and effectiveness can be compiled from this data.

This is clearly an area of great potential for further research using management science and operations research methods along with technical development.

CONCLUSION

Technical development has already been made and will continue to make a significant contribution to the development of the game and the way it is played, managed, and viewed. The challenge is to maintain the integrity of the sport while keeping it relevant in the twenty-first century. In this introductory article, it has been shown that in this context there is plenty of scope for OR and MS methodologies because the technical developments are allowing more scientific measurement of the important decision-making variables and constraints.

For a more extensive list of scientific tennis articles than included here, the ITF web site www.itftennis.com/coaching/publications contains hundreds of articles in categories including sports science, biomechanics, medicine and conditioning, psychology, and technique and methodology.

Acknowledgment

We would like to express our appreciation to Stephen Clarke, Dave Miley, Stuart Miller, and Graham Pollard for their advice and comments on sections of this article and the referees for their constructive suggestions.

REFERENCES

1. International Tennis Federation (ITF). Rules of Tennis. London: International Tennis Federation Ltd; 2009.
2. Kemeny JG, Snell JL. Finite Markov chains. Princeton (NJ): Van Nostrand; 1960.
3. Schultz RW. A mathematical model for evaluating scoring systems with specific reference to tennis. *Res Q Exerc Sport* 1970;41:552–561.
4. Hsi BP, Burych DM. Games of two players. *J R Stat Soc [Ser C]* 1971;20:86–92.

5. Carter WH Jr., Crews SL. An analysis of the game of tennis. *Am Stat* 1974;28:130–134.
6. Miles RE. Scoring systems in sport. In: Kotz S, editor. Volume 8, *Encyclopedia of statistical science*. London: John Wiley & Sons, Inc.; 1988. pp. 607–610.
7. Pollard G. An analysis of classical and tie-breaker tennis. *Aust J Stat* 1983;25:496–505.
8. Morris C. The most important points in tennis. In: Ladany SP, Machol RE, editors. *Optimal strategies in sport*. New York: North Holland Publishing Company; 1977. pp. 131–140.
9. Miles R. Symmetric sequential analysis: the efficiencies of sports scoring system (with particular reference to those of tennis). *J R Stat Soc [Ser B]* 1984;46(1):93–108.
10. Pollard G. A stochastic analysis of scoring systems [PhD thesis]. Canberra: Australian National University; 1986.
11. Pollard G. The optimal test for selecting the greater of two binomial probabilities. *Aust J Stat* 1992;34(2):273–284.
12. Pollard G. A method for determining the asymptotic efficiency of some sequential probabilities. *Aust J Stat* 1990;32(1):191–204.
13. Pollard G, Noble A. The benefits of a new game scoring system in tennis: the 50–40 game. *Proceedings of the 7th Conference on Mathematics and Computers in Sport*; Palmerston North, New Zealand. 2004. pp. 262–265.
14. Pollard G, Pollard G. The efficiency of doubles scoring systems. *Proceedings of the 9th Conference on Mathematics and Computers in Sport*; Coolangatta, Australia. 2008. pp. 45–51.
15. Pollard G, Pollard G. Moment generating function for a tennis match. *Proceedings of the 9th Conference on Mathematics and Computers in Sport*; Coolangatta, Australia. 2008. pp. 204–207.
16. Croucher JS. Developing strategies in tennis. In: Bennett J, editor. *Statistics in sport*. London: Arnold; 1998. pp. 157–171.
17. Gale D. Optimal strategy for serving in tennis. *Math Mag* 1971;5:197–199.
18. Redington F. Usurpers. *J Inst Actuar Stud Soc* 1972;20(3):353–354.
19. George SL. Optimal strategy in tennis: a simplistic probabilistic model. *Appl Stat* 1973;22:97–104.
20. King HA, Baker JAW. Statistical analysis of service and match-play strategies in tennis. *Can J Appl Sports Sci* 1979;4(4):298–301.
21. Hannan EL. An analysis of different serving strategies in tennis. In: Machol RE, Ladany SP, Morrison DJ, editors. *Management science in sports*. Amsterdam: North Holland Publishing Company; 1976. pp. 125–135.
22. Norman JM. Dynamic programming in tennis—when to use a fast serve. *J Oper Res Soc* 1985;36:75–77.
23. McMahon G, de Mestre N. Tennis serving strategies. *Proceedings of the 6th Australasian Conference on Mathematics and Computers in Sport*; Gold Coast, Australia. 2002. pp. 177–182.
24. Barnett T, Clarke S. Combining player statistics to predict a long tennis match at the 2003 Australian Open. *Int J Manage Math* 2005;16:113–120.
25. Barnett T, Pollard G. How the tennis court surface affects player performance and injuries. *Med Sci Tennis* 2007;12(1):34–37.
26. Pollard G, Pollard G. Optimal risk taking on first and second serve. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 273–280.
27. Pollard G. Balancing the use of first and second serves. *Med Sci Tennis* 2008;13(1):30–33.
28. Barnett T, Meyer D, Pollard G. Applying match statistics to increase serving performance. *Med Sci Tennis* 2008;13(2):24–27.
29. Fischer G. Exercise in probability and statistics, or the probability of winning at tennis. *Am J Phys* 1980;48(1):14–19.
30. Croucher JS. The conditional probability of winning games of tennis. *Res Q Exerc Sport* 1986;57(1):23–26.
31. Barnett T. Mathematical modelling in hierarchical games with specific reference to tennis [PhD thesis]. Melbourne: Swinburne University; 2006.
32. Pollard G, Cross R, Meyer D. An analysis of ten years of the four Grand Slam men's singles data for lack of independence of set outcomes. *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport*; Coolangatta, Australia. 2006. pp. 239–246.
33. Pollard G, Pollard G. Importances 1: the most important sets in a match, and the most important points in a game of tennis. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 281–291.

34. Klaassen FJGM, Magnus JR. Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *J Am Stat Assoc* 2001;96:500–509.
35. Brimberg J, Hurley WJ, Lior DU. Allocating energy in a first-to- n match. *J Manage Math* 2004;15(1):25–27.
36. Magnus J, Klaassen FJGM. Myths in tennis. In: Albert J, Koning RH, editors. *Statistical thinking in sports*. Boca Raton (FL): Chapman and Hall; 2008. pp. 217–240.
37. International Tennis Federation. ITF approved tennis balls and classified court surfaces. London: International Tennis Federation Ltd; 2009.
38. Spurr J, Capel-Davies J, Miller S. Player perception of surface pace rating in tennis. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 73–80.
39. Brody H, Cross R, Lindsay C. *The physics and technology of tennis*. Chicago: Independent Publishers Group; 2002.
40. Spurr J, Capel-Davies J. Tennis ball durability: simulation of real play in the laboratory. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 41–48.
41. Downing M. A comparison of static and dynamic ball stiffness. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 31–40.
42. Goodwill S, Douglas J, Miller S, Haake S. Measuring ball spin off a tennis racket. *Proceedings of the 6th International Conference on the Engineering of Sport*; Munich, Germany. 2006. pp. 379–384.
43. Goodwill S, Haake S, Miller S. Validation of the ITF racket power machine. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 113–120.
44. Klaassen FJGM, Magnus JR. Forecasting the winner of a tennis match. *Eur J Oper Res* 2003;148:257–267.
45. Bedford AB, Clarke SR. A comparison of the ATP rating with a smoothing method for match prediction. *Proceedings of the 5th Australasian Conference on Maths and Computers in Sport*; Sydney, Australia. 2000. pp. 43–51.
46. Barnett TJ, Clarke SR. Using Microsoft Excel to model a tennis match. *Proceedings of the 6th Australasian Conference on Maths and Computers in Sport*; Gold Coast, Australia. 2002. pp. 63–68.
47. Elliott B, Reid M. The use of technology in tennis biomechanics. *ITF Coach Sports Sci Rev* 2008;15(45):2–4.
48. Elliott B, Knudson D. Analysis of advanced stroke production. In: Elliott B, Reid M, Crespo M, editors. *Biomechanics of advanced tennis*. London: International Tennis Federation Ltd; 2003. pp. 137–154.
49. Elliott B, Reid M, Crespo M, editors. *Biomechanics of advanced tennis*. London: International Tennis Federation Ltd; 2003.
50. Renstrom AFH. Elbow injuries in tennis. In: Renstrom AFH, editor. *Handbook of sports medicine and science in tennis*. London: International Tennis Federation Ltd; 2002.
51. IOC Medical Commission. In: Renstrom AFH, editor. *Handbook of sports medicine and science: tennis*. Oxford: Blackwell Publishing; 2002.
52. Pluim B, Staal JB, Marks BL, *et al.* Health benefits of tennis. *Br J Sports Med* 2007;41:760–768.
53. International Tennis Federation. *Tennis anti-doping programme*. London: International Tennis Federation Ltd; 2009.
54. Hatch GF, Pink MM, Mohr KJ, *et al.* The effect of tennis racket grip size on forearm muscle firing patterns. *Am J Sports Med* 2006;34:1977.
55. Pluim BM, Staal JB, Windler GE, *et al.* Tennis injuries: occurrence, aetiology and prevention. *Br J Sports Med* 2006;40:415–423.
56. Pluim B, Safran M. *From breakpoint to advantage: a practical guide to optimal tennis health and performance*. Vista (CA): Racquet Tech Publishing; 2006.
57. Petersen C, Nittinger N. *Fit to play tennis - practical tips to optimize training & performance*. Vista (CA): Racquet Tech Publishing; 2006.
58. International Tennis Federation. In: Crespo M, Pluim B, Reid M, editors. *ITF manual on tennis medicine*. London: International Tennis Federation Ltd; 2001.
59. Della Croce F, Tadei R, Asioli PS. Scheduling a round robin tennis tournament under courts and players availability constraints. *Ann Oper Res* 1999;92:349–361.
60. Crespo M, Reid M, Quinn A. *Tennis Psychology: 200+ practical drills and the latest*

- research. London: International Tennis Federation Ltd; 2006.
61. Forzoni R. Psychology and the use of technology. *ITF Coach Sports Sci Rev* 2008;15(45):26–27.
 62. Calstedt RA. Integrative evidence-based tennis psychology: perspectives, practices and findings from a ten year validation investigation of the Calstedt protocol. In: Miller S, Capel-Davies J, editors. *Tennis science and technology 3*. London: International Tennis Federation Ltd; 2007. pp. 245–254.
 63. Musante TM, Yellin BA. The USTA/Equitable computerized tennis ranking system. *Interfaces* 1979;9(4):33–37.
 64. Blackman SS, Casey JW. Development of a rating system for all tennis players. *Oper Res* 1980;28(3):489–502.
 65. Szimák P, Harmath M. Automated tennis line judging. In: Miller S, editor. *Tennis science and technology 2*. London: International Tennis Federation Ltd; 2003. pp. 401–410.
 66. Jonkhoff HC. Electronic line-calling using the TEL system. In: Miller S, editor. *Tennis science and technology 2*. London: International Tennis Federation Ltd; 2003. pp. 377–384.
 67. Marshall J. Theoretical comparison between single and multi light beam semi-automatic line judging systems. In: Miller S, editor. *Tennis science and technology 2*. London: International Tennis Federation Ltd; 2003. pp. 393–400.
 68. Collins H, Evans R. You cannot be serious! Public understanding of technology with special reference to Hawk-Eye. *Public Underst Sci* 2008;17:283–308.
 69. Mather G. Perceptual uncertainty and line-call challenges in professional tennis. *Proc R Soc B Biol Sci* 2008;275:1645–1651.

ANALYSIS OF PAIRWISE COMPARISON MATRICES

JACINTO GONZÁLEZ-PACHÓN
Department of Artificial
Intelligence, Computer
Science School, Technical
University of Madrid, Madrid,
Spain

CARLOS ROMERO
Department of Forest Economics
and Management, Forestry
School, Technical University
of Madrid, Madrid, Spain

INTRODUCTION

Following [1], the pairwise comparison method was introduced in a rudimentary form by G.T. Fechner [2] in 1860, and it was formalized first by L.L. Thurstone [3] in 1927. The method has been a basic ingredient of several interactive multicriteria methods with many applications reported in several areas such as forecasting, investment decision, projects of national importance, and other socioeconomic planning issues.

The basic idea underlying the pairwise comparison method is simple and very close to common sense. Thus, let us assume a scenario where there is a set A of n objects $\{a_1, a_2, \dots, a_n\}$, such as criteria and alternatives, to be evaluated by a single decision maker (DM). If the cardinality of set A or the number n of objects is relatively large, it is very difficult for the DM to provide global preferences for the n objects; that is, to compare simultaneously the characteristics of n objects. However, if we establish a partition of the set A into subsets of cardinality two, such as:

$$\{\{a_j, a_k\}\}_{j,k=1,\dots,n}$$

then, we are demanding from the DM a local information much easier to provide and likely much more robust than the global information. In short, the main virtue of the pairwise

comparison procedure is its simplicity, which is close to the limited human cognitive capacities: **“take two parts at a time when the whole is too difficult to handle simultaneously, especially when the whole is a large one.”**

The final purpose of the pairwise comparison method is to deduce a set of numerical values $(w_1, w_2, \dots, w_n)$ representing the relative importance of each object for the DM. This fact implies that a generic element of the pairwise comparison matrix, m_{ij} , is expected to be equal to the ratio w_i/w_j . For this reason, an ideal pairwise comparison matrix is expected to verify the following properties:

- Reflexivity: $m_{ii} = 1 \forall i$
- Reciprocity: $m_{ij} \times m_{ji} = 1 \forall i, j$
- Consistency: $m_{ij} \times m_{jk} = m_{ik} \forall i, j, k$

These three properties define the concept of rational DM during a pairwise comparison process. However, it is an empirical issue that, in practice imperfect judgments due to human cognitive limitations, leads to matrices without these ideal properties. The challenge faced is how to deduce the priority weights when different scenarios of rationality are verified. This challenge has been addressed by following these two directions: when we state “*a priori*” hypothesis about the rationality of the DM, or when the rationality study is made “*a posteriori*”; that is, when a complete pairwise comparison matrix has already been obtained. These two directions define sections titled “Deriving Priority Weights with Hypotheses about the DM Rationality” and “Deriving Priority Weights Without Making any Hypothesis about the DM Rationality” of the article, respectively.

Finally, in order to illustrate how the pairwise comparison method works and how the corresponding pairwise comparison matrices are obtained, we shall resort to a simple example. Thus, let us assume that in a hypothetical civil engineering project, a DM must provide preferential information with respect to three relevant criteria in the

decision-making process: cost (C), environmental impact (EI), and energy consumption (EC) associated with the undertaking of the project. For this particular example, a pairwise comparison procedure requires for this particular example six judgment values (in general $n(n-1)$) from the DM, which lead to the following square matrix:

$$\begin{array}{c} \text{C} \quad \text{EI} \quad \text{EC} \\ \text{C} \quad \begin{pmatrix} 1 & m_{12} & m_{13} \\ m_{21} & 1 & m_{23} \\ m_{31} & m_{32} & 1 \end{pmatrix} \\ \text{EI} \\ \text{EC} \end{array}$$

The interpretation of the elements of the above matrix is straightforward. For instance, the element m_{12} represents the relative importance attached by the DM to the criterion cost when it is compared with the criterion environmental impact. It is now clear that the pairwise comparison matrix will be worthy only if a cardinal information/numerical value is associated with its elements and obviously, for that, we shall need a numerical scale. Without loss of generality in this article, we shall resort to Saaty's scale, a widely used scale in the applied literature [4].

DERIVING PRIORITY WEIGHTS WITH HYPOTHESES ABOUT THE DM RATIONALITY

This issue will be addressed by considering three different possible scenarios according to the level of rationality underlying the information provided by the DM.

Case I. Rational Decision Maker

In this scenario of perfect rationality, we assume that the information provided by the DM holds the three properties: reflexivity, reciprocity, and consistency. Except for small problems, this scenario is unrealistic, but it has theoretical interest as a starting point for the presented analysis.

In this context, the DM only needs to provide $(n-1)$ comparisons. Moreover, it is mathematically quite simple to derive the vector of preferential weights w from matrix M . Thus, for the above simple example in this

scenario, the vector w would be obtained by solving the following system of homogeneous equations:

$$\begin{aligned} w_1 - m_{12}w_2 &= 0 \\ w_1 - m_{13}w_3 &= 0 \\ w_2 - m_{23}w_3 &= 0 \end{aligned}$$

The consistency condition guarantees that the rank of the matrix of coefficients is less than the number of unknowns, which implies the existence of a nontrivial solution for the above system of equations. Moreover, the reciprocity condition avoids the inclusion of the three equations corresponding to the lower part of the matrix.

Case II. Partially Rational Decision Maker

In this new scenario, the rationality conditions are relaxed by eliminating the consistency condition. Now, the DM holds the tautological condition of reflexivity (i.e., one object is of the same importance as itself) and the reciprocity condition. In this new scenario, the number of pairwise comparisons increases from $(n-1)$ to $n(n-1)/2$. Thus, for our simple example with three objects, we will now require three pieces of information instead of two. This small difference is due to the small cardinality of the set A of objects. If, for example, the DM needs to compare 10 objects, the number of pairwise comparisons will move from 9 to 45, which is a significant difference.

As we are now assuming that the consistency matrix does not hold, the system of homogeneous equations only has as a solution the trivial one (i.e., $w_1 = w_2 = w_3 = 0$). There are several procedures for approximating a solution to the corresponding system of equations. The theory attempts to provide a consistent matrix "as close as possible to" M . One of these procedures was proposed by Saaty [4] in 1977. With this method, the original matrix M is replaced by a consistent matrix $C = (c_{ij} = v_i/v_j)_{ij}$, where $v = (v_1, v_2, \dots, v_n)$ is the eigenvector associated with the largest eigenvalue of M . By the theorem of Perron–Fröbenius,

$\lambda_{\max}$ is unique, positive, and simple. Furthermore, v can be chosen as having all its positive coordinates. Now, the consistent matrix C shares the eigenvector v with the inconsistent matrix M . In this sense, it can be considered that C is “as close as possible to” M .

Another family of procedures is based on the idea of approximating a solution to the above system of homogeneous equations by transforming the systems into a goal programming (GP) formulation. There are many variants in this direction; see [5] or [6]. The most straightforward GP formulation of our problem leads to the following weighted GP formulation:

Achievement function:

$$\text{Min } n_1 + p_1 + n_2 + p_2 + n_3 + p_3$$

Goals:

$$w_1 - m_{12}w_2 + n_1 - p_1 = 0$$

$$w_1 - m_{13}w_3 + n_2 - p_2 = 0$$

$$w_2 - m_{23}w_3 + n_3 - p_3 = 0$$

Auxiliary constraint:

$$w_1 + w_2 + w_3 = 1$$

Nonnegativity constraints:

$$n_i, p_i, w_i \geq 0 \quad i = 1, 2, 3$$

The last equation of the above GP model has an auxiliary character, making the interpretation of the weights as percentages of importance easier, and also preventing the optimal solution from being arbitrarily scaled. It is interesting to note that the value of the achievement function for the optimum solution will indicate the level of consistency associated with matrix M . This matter will be clarified later with the help of a numerical example.

Case III. Irrational Decision Maker

In this new scenario, the only rationality condition is the tautological condition of reflexivity. Therefore, the number of pairwise comparisons requested is now $n(n-1)$; that is, the removal of the reciprocity condition implies doubling the number of questions raised from the DM. Moreover, in a scenario of nonreciprocity, the maximum eigenvalue method does not work. However, the family of procedures based on GP is straightforwardly applicable to this scenario. In this case, the GP model will now have double the number of equations (goals) with respect to the previous case. Thus, the general weighted GP model to derive the vector of weights for a general pairwise comparison matrix of dimension 3×3 , without enjoying properties of consistency and reciprocity will be given by:

Achievement function:

$$\begin{aligned} \text{Min } n_1 + p_1 + n_2 + p_2 + n_3 + p_3 \\ + n_4 + p_4 + n_5 + p_5 + n_6 + p_6 \end{aligned}$$

Goals:

$$w_1 - m_{12}w_2 + n_1 - p_1 = 0$$

$$w_1 - m_{13}w_3 + n_2 - p_2 = 0$$

$$w_2 - m_{21}w_1 + n_3 - p_3 = 0$$

$$w_2 - m_{23}w_3 + n_4 - p_4 = 0$$

$$w_3 - m_{31}w_1 + n_5 - p_5 = 0$$

$$w_3 - m_{32}w_2 + n_6 - p_6 = 0$$

Auxiliary constraint:

$$w_1 + w_2 + w_3 = 1$$

Nonnegativity constraints:

$$n_i, p_i, w_i \geq 0 \quad i = 1, 2, 3$$

If we represent $\text{Min} \sum_{i=1}^6 (n_i + p_i)$ by θ and $\text{Max} \sum_{i=1}^6 (n_i + p_i)$ by θ' , the index of consistency (IC) associated with the solution of the above GP model will be equal to:

$$\text{IC} = [1 - (\theta/\theta')] \times 100$$

as θ' measures the maximum possible level of inconsistency (see the numerical example below).

Numerical Example

The above ideas will be illustrated with the help of the following example. For an exercise with four objects, Saaty's scale, without the reciprocity property, has been used for valuations, yielding the following pairwise comparison matrix, which is nonreciprocal and inconsistent:

$$\begin{pmatrix} 1.00 & 5.00 & 0.50 & 3.00 \\ 0.25 & 1.00 & 3.00 & 0.33 \\ 2.00 & 0.50 & 1.00 & 0.50 \\ 0.33 & 3.00 & 2.00 & 1.00 \end{pmatrix}$$

The vector of preferential weights corresponding to the above pairwise comparison matrix can be obtained by solving the following weighted GP model:

Achievement function:

$$\begin{aligned} \text{Min} \quad & n_1 + p_1 + n_2 + p_2 + n_3 + p_3 + n_4 \\ & + p_4 + n_5 + p_5 + n_6 + p_6 + n_7 \\ & + p_7 + n_8 + p_8 \end{aligned}$$

Goals:

$$\begin{aligned} w_1 - 5w_2 + n_1 - p_1 &= 0 \\ 2w_1 - w_3 + n_2 - p_2 &= 0 \\ w_1 - 3w_4 + n_3 - p_3 &= 0 \\ 4w_2 - w_1 + n_4 - p_4 &= 0 \\ w_2 - 3w_3 + n_5 - p_5 &= 0 \\ 3w_2 - w_4 + n_6 - p_6 &= 0 \\ 2w_3 - w_2 + n_7 - p_7 &= 0 \\ 2w_3 - w_4 + n_8 - p_8 &= 0 \end{aligned}$$

Auxiliary constraint:

$$w_1 + w_2 + w_3 + w_4 = 1$$

Nonnegativity constraints:

$$n_i, p_i, w_i \geq 0 \quad i = 1, 2, 3, 4$$

It should be noted that, in order to avoid redundancies, in the above model, the equations corresponding to reciprocal elements have not been replicated. By solving the model, the following solution is obtained:

$$\begin{aligned} w_1 &= 0.612, w_2 = 0.122, w_3 = 0.062, w_4 \\ &= 0.204 \end{aligned}$$

For this solution, $\theta = 1.59$ and $\theta' = 8.01$. Therefore, the consistency of the solution obtained, according to the index IC previously defined, is equal to:

$$\text{IC} = [1 - (1.59/8.01)] \times 100 = 80.14\%$$

DERIVING PRIORITY WEIGHTS WITHOUT MAKING ANY HYPOTHESIS ABOUT THE DM RATIONALITY

In this scenario, we have had to obtain the complete pairwise comparison matrix;

that is, the number of pairwise comparisons requested is $n(n-1)$, similar to case III of section titled “Deriving Priority Weights with Hypotheses about the DM Rationality”.

We consider that, in practice, imperfect judgments, due to the human cognitive limitations, lead to matrices with a high level of violation of the rationality properties defined earlier. This challenge has been addressed in the literature by following one of the following three directions:

- (a) A threshold of inconsistency is defined in one way or another. Matrices surpassing this threshold, as they are derived from highly inconsistent information, are eliminated from the analysis. This is a clear normative approach.
- (b) The inconsistency is considered to be a fact. As the information is provided by a real DM, it must be considered in the analysis. This approach can be considered to be a positive or descriptive one. In this case, the scenario is similar to the one stated in case III of section titled “Deriving Priority Weights with Hypotheses about the DM Rationality”.
- (c) An approximation perspective, which can be considered as a prescriptive orientation; that is, a compromise between the above normative and descriptive orientations. With this perspective, when there are inconsistencies, a new matrix M' that differs from M “as little as possible” and that holds the rationality conditions “as much as possible” is sought, see [7, 8], or [9]. This is the scenario in this section.

According to the third direction, and in order to obtain the new matrix $M' = (m'_{ij})_{ij}$, it is interesting to review a general procedure based on a distance function framework [10]. For this purpose, the following objective functions are introduced for a general metric $p \in [1, \infty)$.

- (a) *Imposing similarity between matrices M and M' .* Imposing similarity between

matrices M and M' implies the minimization of the following nonlinear function.

$$\sum_{\substack{i,j \\ i \neq j}} |m_{ij} - m'_{ij}|^p$$

- (b) *Imposing reciprocity to matrix M' .* Imposing reciprocity in matrix M' implies the minimization of the following nonlinear function.

$$\sum_{\substack{i,j \\ i \neq j}} |m'_{ij} \times m'_{ji} - 1|^p$$

- (c) *Imposing consistency to matrix M' .* Imposing consistency in matrix M' implies the minimization of the following nonlinear function.

$$\sum_{\substack{i,j,k \\ i \neq j \neq k}} |m'_{ij} \times m'_{jk} - m'_{ik}|^p$$

The optimization of the above objective functions is subjected to the fulfillment of certain scale conditions used in the pairwise comparison procedure. This consideration leads to the following constraints set:

$$L \leq m'_{ij} \leq U \quad \forall i, j$$

The nonsmooth character of the above objective functions makes their optimization extremely difficult. However, this type of optimization problems can be reduced to GP formulations considering the relationship between distance function models and mathematical programming [11]. Thus, the above objective functions can be simultaneously optimized in an aggregated way by formulating and solving the following Archimedean GP model:

Achievement function:

$$\begin{aligned} \text{Min} \sum_l (n_l^{(1)} + p_l^{(1)})^p + \sum_s (n_s^{(2)} + p_s^{(2)})^p \\ + \sum_t (n_t^{(3)} + p_t^{(3)})^p \end{aligned}$$

Goals:

$$m_{ij} - m'_{ij} + n_l^{(1)} - p_l^{(1)} = 0$$

$$l = 1, 2, \dots, n(n-1)$$

$$m'_{ij} \times m'_{ji} - 1 + n_s^{(2)} - p_s^{(2)} = 0$$

$$l = 1, 2, \dots, \frac{n(n-1)}{2}$$

$$m'_{ij} \times m'_{jk} - m'_{ik} + n_t^{(3)} - p_t^{(3)} = 0$$

$$l = 1, 2, \dots, n(n-1)(n-2)$$

Auxiliary constraints:

$$L \leq m'_{ij} \leq U \quad \forall i, j$$

Nonnegativity constraints:

$$\mathbf{n} \geq \mathbf{0}, \mathbf{p} \geq \mathbf{0}$$

It can be observed that there are three blocks of goals corresponding to conditions of similarity, reciprocity, and consistency, respectively. Deviation variables in the first block have a very different range of values with respect to the deviation values associated with the other two blocks. Hence, it might be suitable to implement a normalization process in the achievement function.

The reader should note that for small values of metric p , the Archimedean GP model may provide solutions, which are too biased toward the achievement of one of the three blocks of conditions formulated earlier. As p increases, more importance is attached to the largest deviation. Thus, for $p = \infty$, the maximum disagreement is minimized. This type of analysis can be generalized, within a context of linearity, by resorting to the formulation of an extended GP model (see for details [10] and [12]).

Numerical Example

By applying the above procedure for metric 1 to the pairwise comparison matrix of the previous section, the following improved matrix was obtained:

$$\begin{pmatrix} 1.00 & 5.00 & 3.33 & 1.67 \\ 0.20 & 1.00 & 0.67 & 0.33 \\ 0.30 & 1.50 & 1.00 & 0.50 \\ 0.60 & 3.00 & 2.00 & 1.00 \end{pmatrix}$$

From this matrix, by applying the GP procedure proposed earlier, the following vector $\mathbf{W}$ of preferential weights was obtained:

$$w_1 = 0.476, w_2 = 0.095, w_3 = 0.145, w_4 = 0.286$$

For this solution, $\theta = 0.0022$ and $\theta' = 5.01$; therefore,

$$\text{IC} = [1 - (0.002/5.01)] \times 100 = 99.95\%$$

which improves the consistency of matrix M' with respect to the initial matrix M .

THE GROUP DECISION-MAKING PROBLEM

The purpose of this section is to provide some ideas on how to extend the above analysis to a context where there is a finite number r of DM, each one providing a pairwise comparison matrix. Depending on where the group wants to reach the consensus, in the final priority weights or in the initial pairwise comparison information, we can foresee two different aggregation problems.

- (a) Individual priority weights are computed and combined in a consensus vector of preferential weights [13, 14].
- (b) A consensus pairwise comparison is computed and a priority weight is obtained [15]. This is the scenario that we have adopted in this section.

In this context, we are concerned with the computation of a consensus pairwise comparison matrix M^C that differs from individual

matrices $M^1, M^2, \dots, M^r$ “as little as possible.”

In order to make operational the assertion “as little as possible,” we have again resorted to a family of distance functions based on the generic metric p , which leads to the following optimization problem:

$$\begin{aligned} \text{Min } & \sum_{t=1}^r \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |m_{ij}^{(t)} - m_{ij}^C|^p \\ & \text{s.t.} \\ & L \leq m_{ij}^C \leq U \quad i, j \in \{1, \dots, n\} \end{aligned}$$

where $m_{ij}^{(t)}$ are the elements of matrix M^t ($t = 1, \dots, r$) and m_{ij}^C are the elements of the consensus matrix M^C sought. The constraint set must establish, as usual, some scale conditions.

Again the nonsmooth character of the above objective function makes its optimization extremely difficult. However, it is once more quite simple to transform the above model into the following Archimedean GP formulation [12]:

Achievement function:

$$\text{Min } \sum_{t=1}^r \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (n_{ij}^t + p_{ij}^t)^p$$

Goals:

$$m_{ij}^C - m_{ij}^{(t)} + n_{ij}^t - p_{ij}^t = 0$$

$$i, j \in \{1, \dots, n\} \quad t \in \{1, \dots, r\}$$

Constraints:

$$L \leq m_{ij}^C \leq U \quad i, j \in \{1, \dots, n\}$$

$$\mathbf{n} \geq \mathbf{0}, \mathbf{p} \geq \mathbf{0}$$

Again the value of metric p attached to the objective function might have a strong influence in obtaining a good solution from the point of view of the average (small values of p) or a good solution from the point of view of the DM with views that are more displaced with respect to the consensus obtained (large values of p). Once more, the analysis can be generalized in a linear context by the formulation of an extended GP model (see for details [11, 15]).

Numerical Example

Let us illustrate how the theory presented in this section works with the help of a simple example. Thus, let us assume that we have four DMs with the following pairwise comparison matrices over four objects, using Saaty's scale:

$$M^1 \equiv \begin{bmatrix} 1 & 1/5 & 5 & 3 \\ 3 & 1 & 1/7 & 1/3 \\ 1/5 & 7 & 1 & 1/3 \\ 1/3 & 3 & 3 & 1 \end{bmatrix}$$

$$M^2 \equiv \begin{bmatrix} 1 & 3 & 1/3 & 1/3 \\ 1/3 & 1 & 1 & 5 \\ 3 & 1 & 1 & 7 \\ 5 & 1/5 & 1/5 & 1 \end{bmatrix}$$

$$M^3 \equiv \begin{bmatrix} 1 & 1 & 1/2 & 7 \\ 1 & 1 & 1/4 & 5 \\ 2 & 4 & 1 & 8 \\ 1/5 & 1/5 & 1/8 & 1 \end{bmatrix}$$

$$M^4 \equiv \begin{bmatrix} 1 & 7 & 7 & 3 \\ 5 & 1 & 1 & 1/5 \\ 1/7 & 1 & 1 & 1/5 \\ 1/3 & 5 & 5 & 1 \end{bmatrix}$$

By applying the above model to the four matrices, and by resorting to metric $p = 1$, the resulting consensus pairwise comparison matrix M^C is the following one:

$$M^C \equiv \begin{pmatrix} 1 & 2.11 & 4.43 & 3 \\ 1 & 1 & 1 & 0.33 \\ 2 & 1 & 1 & 7 \\ 0.33 & 3 & 0.20 & 1 \end{pmatrix}$$

It should be noted that the problem of inconsistencies analyzed in the previous section can be transferred to a group decision-making context in two different ways:

- (a) Analyzing individual pc matrices, M^1 , M^2 , . . . , M^r .
- (b) Analyzing the final consensus matrix M^C .

CONCLUSIONS

Pairwise comparisons leading to pairwise comparison matrices has become, in the last few years, a powerful multicriteria procedure, in order to obtain robust information from a single DM or a group of DMs. In this way, this procedure is a basic ingredient of several interactive multicriteria methods with many applications reported in the literature.

The success of this approach lies in its simplicity, which is close to common sense: “when it is difficult to provide preferential information about a large set of objects, then it is advisable to take only two parts at a time.”

Moreover, the derivation of a vector of preferential weights from a pairwise comparison matrix is relatively easy from a computational point of view. Thus, according to the rationality properties underlying the pairwise comparison matrix, there are different computational procedures (maximum eigenvalue, different GP formulations, etc.) that can be applied.

It is interesting to note that the gap between descriptive and normative models, which is a controversial topic in the decision-making literature, can be compromised within a pairwise comparison context without excessive computational difficulties. Finally, the extension of the pairwise comparison analysis from a single DM to a group decision-making context is a relatively straightforward task, leading to models easy to be computed.

Acknowledgments

This research has been funded by the Government of Madrid Autonomous Region

under project QM100705026. The authors are indebted to Diana Badder for English language editing.

RELATED ARTICLES

Decision Analysis Insights from Behavioral Economics; Solving Multicriterion (MCDM) Problems; Analytic Hierarchy Process and Critique

REFERENCES

1. Janicki R, Koczkodaj WW. A weak order solution to a group ranking and consistency-driven comparisons. *Appl Math Comput* 1998 Aug; 94(2–3):227–241.
2. Fechner, GT Elements of Psychophysics Vol 1. 1965. Holt, Rinehart & Winston, New York, translation of H.E. Adler of Elemente der Psychophysik. 1860. Breitkopf und Härtel, Leipzig.
3. Thurstone LL. A law of comparative judgments. *Psychol Rev* 1927Jul; 34(4):273–286.
4. Saaty TL. Scaling method for priorities in hierarchical structures. *J Math Psychol* 1977; 15(3):234–281.
5. Bryson N. A goal programming method for generating priority vectors. *J Oper Res Soc* 1995May; 46(5):641–648.
6. Jones DF, Mardle SJ. A distance-metric methodology for the derivation of weights from a pairwise comparison matrix. *J Oper Res Soc* 2004Aug; 55(8):869–875.
7. Chu MT. On the optimal consistent approximation to pairwise comparison matrices. *Linear Algebra Appl* 1998Mar; 272:155–168.
8. Koczkodaj W, Orłowski M. Computing a consistent approximation to a generalized pairwise comparisons matrix. *Comput Math Appl* 1999Feb; 37(3):79–85.
9. Choo EU, Wedley WC. A common framework for deriving preference values from pairwise comparison matrices. *Comput Oper Research* 2004May; 31(6):893–908.
10. González-Pachón J, Romero C. A method for dealing with inconsistencies in pairwise comparisons. *Eur J Oper Res* 2004Oct; 158(2):351–361.
11. Romero C. Extended lexicographic goal programming a unifying approach. *Omega-Int. J. Manage. Sci.* 2001Feb; 29(1):63–71.

12. Romero C. A general structure of achievement function for a goal programming model. *Eur J Oper Res* 2004Mar; 153(3):675–686.
13. Ramanathan R, Ganesh LS. Group preference aggregation methods employed in AHP: an evaluation and an intrinsic process for deriving members' weightages. *Eur J Oper Res* 1994Dec; 79(2):249–265.
14. Saaty TL, Sodenkamp M. The analytic hierarchy and analytic network measurement processes: the measurement of intangibles decision making under benefits, opportunities, costs and risks. In: Constantin Z, Pardalos PM, editors. *Handbook of multicriteria analysis*. Berlin: Springer; 2010. pp. 91–166 (Chapter 4).
15. González-Pachón J, Romero C. Inferring consensus weights from pairwise comparison matrices without suitable properties. *Ann Oper Res* 2007Oct; 154(1):123–132.

ANALYTIC MODELING OF INSURGENCIES

MOSHE KRESS
Naval Postgraduate School,
Monterey, California, USA

INTRODUCTION

Combat modeling is one of the oldest areas of operations research, dating back to the Kriegsspiels—board war-games developed in the early nineteenth century for training, planning, and testing military operations in the Prussian Army. The ground-breaking work of Lanchester in 1916 [1] marks the beginning of formal models of conflicts, where mathematical formulas and, later on, computer simulations replace board games and sand tables. From WWII until a decade ago combat models have mostly been focused on the physical aspect of conflicts. These models have been typically applied to regular force-on-force engagements, where adversary armies, navies and air forces engage in, so-called, kinetic actions that comprise fire, maneuver, and attrition.

Insurgencies are not new either; they date as far back as the Jewish revolt against the Seleucid Empire in the second century BC. Since then, insurgencies have been prevalent throughout history. While fighting the Japanese Imperial Army occupying China, Mao Tse-tung was one of the first to offer a conceptual “model” of insurgencies in 1937 [2]. He described the insurgents as “fish” that must swim in the population’s “sea” to survive and prevail. More recent and formal models of insurgencies include Deitchman’s Guerrilla Warfare model [3], which is a Lanchester-based mathematical model (see more details about this model later on), and McCormick’s Magic Diamond model [4], which is a conceptual model that identifies the players in an insurgency and specifies the interrelationships among them. However, it is only since the early 2000s,

following major insurgencies in Colombia, Iraq, Afghanistan, Libya, and elsewhere in the Middle East, that operations-oriented modeling of these armed conflicts has become prevalent. In particular, during the Iraq and Afghanistan wars operations-research analysts were deployed with combat units to collect and analyze data. They also utilized reach-back support from operations research analysts in the United States for developing decision support models, for both combat tactics against the insurgents and information operations aimed at gaining the support of the local population [5].

Insurgencies are different than regular force-on-force engagements. The two adversaries in an insurgency are *government* organized forces on one side, and, on the other side, loosely organized nonstate actors, henceforth called *insurgents*, such as violent demonstrators, armed rebels, and terrorists. While there are some differences between *counterinsurgency* and *counterterrorism* operations—mostly in terms of scale of operations and countermeasures (see, e.g., [6])—in some places in the following, the two terms may be used interchangeably.

In one-on-one situations, the nonstate actors—the insurgents—are no match to the state-controlled government forces, who are significantly larger, better equipped, and better trained. To avoid eradication, the insurgents must reduce their signature as targets, and this elusiveness is attained by blending in with the civilian population among which the insurgents operate. The insurgents use relatively simple, yet lethal, weapons such as small arms, improvised explosive devices, and suicide bombs.

The asymmetry described earlier is one significant characteristic of insurgencies. The second characteristic is the active role played by civilians who provide insurgents, either willingly or as a result of coercion, hiding places, shelters, logistical support, and most importantly—information and recruits. Civilians play other roles too: They may provide information (intelligence) to

the government forces regarding insurgents' activities and whereabouts, they consume social and economic resources that are provided either by the insurgents or the government, and they are possible targets to violent actions by both sides (see e.g., the conflict in Syria that started in 2011). All these characteristics make civilians a key component in insurgency modeling, which is absent in legacy armed-conflict models. In particular, models from behavioral science, sociology, political science, and economics play a major role in insurgency modeling.

There are several ways to model insurgencies. Probably, the most popular models are detailed (e.g., agent-based) simulations that represent both physical and cognitive interrelations among stakeholders in such conflicts. Some examples of simulation models are mentioned later on. Another possible modeling approach is system dynamics [7]. A "softer" side of insurgency analysis—by means of political science, sociology, and behavioral science—is embodied in the works of Killcullen [8], which has been influential in planning counterinsurgency operations in Iraq and Afghanistan. Another important political and economic analysis of an insurgency is reported in [9].

In this article, however, we focus on OR analytic modeling based on formal methodologies such as differential equations, utility theory, game theory, and probability models. Besides their elegance, the conceptual simplicity of these models provides transparency, facilitates a clear description of cause-and-effect relations, and thus offers strategic insights regarding insurgency situations. Because of these features, some of the models presented later have been briefed to top leadership in the US Army.

Following a general discussion on analytic modeling of insurgencies presented in the next section, we describe three families of models addressing three major issues: (i) *Eyes and Ears*: the impact of information, intelligence, and situational awareness on the conduct of counterinsurgency operations, (ii) *Hearts and Minds*: the effect of civilians' behavior and attitude toward the insurgents and the government, and (iii) *Bullets and*

Fire: the "kinetics" of insurgencies that represents the physical attrition in this type of conflicts. The three families are obviously related—for example, informational models may have some attrition and behavioral components. Thus, the classification mentioned—information, behavior, and attrition—only serves for highlighting the main thrust of the models included in each corresponding section.

THE BIG PICTURE

The three major players in an insurgency are the two adversaries—the government forces and the insurgents—and the civilian population, which is caught in the middle. A fourth player is the international community [4, 10] who may (e.g., Libya, 2011) or may not (e.g., Syria, 2011–2014) be actively involved in the conflict. Because international forces, if they get actively involved in the insurgency, typically side with either the government (e.g., Afghanistan) or the insurgents (e.g., Libya), we will focus only on the three main players, as shown in Fig. 1.

The government and the insurgents are engaged in an armed conflict (the "Combat" link in Fig. 1) governed by violent actions and mutual attrition. The civilian population is divided into three groups: *supporters* of the government, *supporters* of the insurgents henceforth called *contrarians* and those who maintain neutrality. The government and the insurgents, while trying to gain the support of the population, provide social benefits such as healthcare and education; however, they may also impose some requirements such as dress codes, taxes, and draft. The insurgents also use aimed violence against civilians to coerce them to collaborate and deter them from supporting the government. These actions, combined with observations on the state of the armed-conflict between the insurgents and the government forces, lead to shifts in civilians' behavior regarding their support to the two sides. Contrarians support the insurgents by providing hiding places, recruits, logistics support, and information, while government supporters mostly provide information, in the form of human intelligence (HUMINT), to government forces.

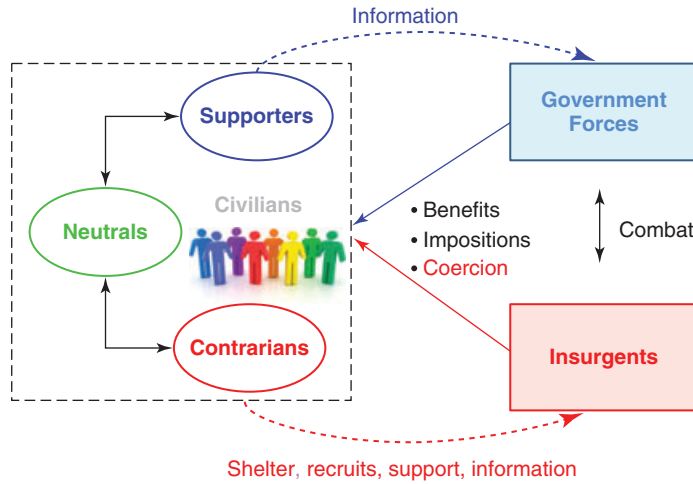


Figure 1. The stakeholders in an insurgency.

As mentioned earlier, the blue block corresponding to government forces in Fig. 1 may contain several stakeholders such as regular government army, police forces, and international forces supporting the government. The red block corresponding to the insurgents may also comprise nonhomogeneous forces in the form of a number of competing tribal or sectarian militias who may also fight each other to gain control of the insurgency (e.g., Syria 2013–2014).

There were several attempts to develop analytic insights regarding the observed evolution of insurgencies. Bohorquez *et al.* [11] examine the patterns of violence in insurgencies and terror events and identify a common pattern regarding the size distribution of such events and their timing. Their dynamic model explaining this pattern is based on the notion of coalescence and fragmentation of insurgents or terror organizations, thus producing ecology of groups. A more recent paper [12] reveals a dynamical pattern of fatal insurgency attacks. This pattern, which is manifested in a power law, identifies possible escalation scenarios of such attacks. The authors establish a new metric for understanding the momentum of these attacks and the effectiveness of COIN actions—a metric that appears to be stable across multiple conflicts and at different scales.

EYES AND EARS

Situational awareness regarding the battle-field, generated by a steady, relevant, and reliable stream of intelligence reports, is an important factor in any armed conflict. In order to effectively deploy and operate its weapons and other combat assets, a military force needs to know the deployment, capabilities, plans, and intentions of its adversary, as well as details regarding the combat environment. Situational awareness becomes even more critical in counterinsurgency operations because of the two aforementioned characteristics: asymmetry and the significant active role of civilians. First, the asymmetry is manifested in elusive, well-hidden insurgents, diffused among civilians and thus reducing their signature as targets. This elusiveness requires extra effort by the government forces—mostly by acquiring intelligence from human sources—to locate and effectively engage the insurgents. Second, the impact of civilians on the evolution of an insurgency necessitates social, cultural, and behavioral intelligence about the population’s attitude, mood, and sentiments. This type of intelligence is typically absent in legacy force-on-force conflicts (see also Section HEARTS AND MINDS).

Generalized Deitchman Model

The first to capture the asymmetry feature in a Lanchesterian setting was Deitchman [3] in his Guerrilla Warfare model, which is a pair of differential equations. If $G(t)$ and $I(t)$ are the sizes of the government forces and the insurgents at time t , respectively, and P is the size of the civilian population, then Deitchman's model states that

$$\begin{aligned} G'(t) &= -\alpha I(t) \\ I'(t) &= -\gamma G(t) \frac{I(t)}{P} \end{aligned}$$

where α, γ are attrition coefficients and the signature of the insurgents—their proportion in the population—is represented by $\frac{I(t)}{P}$. The larger the population, among which the insurgents are diffused, the smaller is the insurgents' signature and thus the smaller the effectiveness of the government forces in fighting the insurgents. Deitchman's model was extended in [13–15]. Absent accurate situational awareness, that is, information regarding the whereabouts of the insurgency targets, not only might the guerrillas be able to continue their insurgency actions unhindered, but also the collateral damage caused to civilians from poor targeting by the government forces may generate an adverse response against the government, thus creating popular support for the insurgents [16]. This popular support may translate into new cadres of recruits to the insurgency ranks [17]. This dynamics of two-sided attrition, collateral casualties among civilians, recruitment to the insurgency, and reinforcement to the government forces, is captured by a pair of differential equations in [14]. The imperfect situational awareness, which results in collateral casualties, turns out to be crucial. Analyzing the differential equations model results in some operational insights. While perfect targeting would eradicate the insurgents in no time, it is shown that, under reasonable assumptions, the government forces can never do that if the situational awareness is imperfect. The best the government forces can hope for is containing the insurgency at some manageable size, a result that has been confirmed empirically in recent insurgencies

(e.g., Afghanistan and Syria). Moreover, it is shown that there may be two steady-state containment situations. One stalemate scenario involves relatively low level of violent activity by the government forces. However, this stalemate is fragile from the government point of view; small changes in the balance of forces could lead to quick government demise. The other steady-state containment scenario involves high level of violent activity and is more stable. In any case, the only alternative, which is the worse-case scenario for the government, is when the latter actually loses, as was the case in Libya, Tunisia, and Yemen in 2011 (aka “Arab Spring”).

Searching for Insurgents

Optimizing the search for hiding insurgents could be modeled as a *whereabouts search* problem [18]. The objective here is to detect, as fast as possible, an insurgents' cell that is hidden in one out of n possible locations observed by an imperfect sensor. The cost of searching a location is c , and the intelligence obtained from this search may be false negative (the searcher failed to detect an insurgents' cell) with probability $1 - p$, false positive (the searcher erroneously identifies a location as an insurgents' cell), with probability q . An indication by the sensor that a certain location is an insurgents' cell, may be *correct*—insurgents are indeed hiding in that location—or *wrong*—the location is void of insurgents. Following such an indication (correct or wrong), a combat team is sent out with the objective to capture the insurgents in the indicated location. The operational costs of such actions are C_P and C_W , for *correct* and *wrong* scenarios, respectively. In case the indication is wrong, the cost may be collateral damage and lost of time and effort. It is shown in [19] that the optimal search sequence of the sensor follows a greedy policy that only depends on the prior location probabilities and the values of $\frac{p}{C_P + qC_W}$ in the various locations.

Balancing Between Intelligence Collection and Analysis

Collecting, processing, and analyzing intelligence for obtaining a current and reliable

situational awareness is a complex task that requires considerable resources. In particular, it is crucially important to balance off between the collection effort—gathering information and data from an assortment of sources—and the analysis effort—analyzing the collected information and producing useful operational knowledge. In a model that addresses this issue [20], using game-theoretic arguments in a queuing setting, assuming two types of traffic—light and heavy, the authors identify equilibria under various operational scenarios. They conclude that contrary to current practices in the counterterrorism intelligence community, more should be invested in analysis so as to reduce the size of the queue of intelligence input.

HEARTS AND MINDS

First coined by the British in Malaya during the early 1950s of the last century, and later on used in Vietnam (1960s) and more recently in Iraq and Afghanistan, “winning hearts and minds” is an operational concept that aims at swaying public support toward one side in the conflict. Because public opinion plays a major role in shaping the way an insurgency evolves, both sides—the government and the insurgents—take actions to win the support of the people. Which actions are effective? How do these actions interact? What is the end-effect of these actions? These types of questions have been addressed mostly in simulations [21], such as agent-based simulations [22, 23]. Various empirical approaches for addressing these questions have been reported in the political science literature, for example, [9, 24]. Modeling public behavior takes into account the difference between the *attitude* of an individual (his “heart”) and his *behavior* (his “mind”). For example, a person who fundamentally opposes the insurgents in his heart (attitude) may express latent or even active support to the insurgents in his mind (behavior) based on pragmatic considerations. The terms attitude and behavior used here are related, respectively, to the terms “private preference” and “public preference” used in [25]. While it may be hard to change the attitude of people—their fundamental

beliefs and values that have been shaped over centuries of cultural evolution—it may be possible to affect their manifested behavior, which is influenced by interests and utilities.

Carrots and Sticks

An analytic model that captures the aforementioned utilitarian aspect is presented in [26]. The situation modeled therein is as follows: a certain region is under the control of the insurgents who act there quite freely. Civilians in that region react to these actions based on their perception regarding the capabilities and intentions of the government, had it been in control instead of the insurgents, to improve or worsen their welfare. The insurgents execute two types of actions: (i) violent actions, aimed to coerce potential supporters of the government, which are only targeted at such suspected civilians (y), and (ii) nonviolent supportive economic and social actions aimed at gaining the population support (x). The dilemma of the insurgents is how to balance these two types of actions—the “sticks” y and the “carrots” x . A dynamic utility-based model is developed in [26] in which the state variables are the fractions of contrarians (supporters of the insurgency) (C), latent supporters of the government (L) and active supporters of the government (A) in the population. The model is parameterized by the activity levels of the insurgents (x, y) for both nonviolent and violent actions, which are assumed to be bounded by the available resources to the insurgents, as depicted in Fig. 3. It is shown that the model converges into equilibrium, and the main insight is that tipping points may occur, where small changes in the aforementioned balance between insurgents’ violent (y) and nonviolent (x) actions can drastically change the size of active supporters who help the government. See Fig. 2.

The government should be aware of potential tipping points that lead to the elimination of most active supporters and attempt to avoid situations that lead to them.

Civilians’ Reaction to Violence

How civilians remember and react to violent actions by both sides—the insurgents

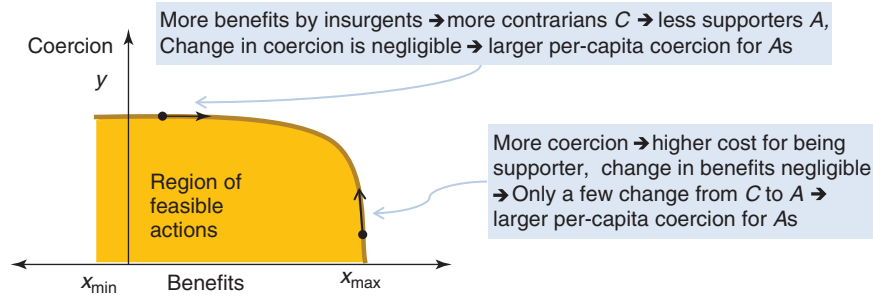


Figure 2. Cascading effects of insurgency's actions.

and the government—is modeled in [27]. Arguably, the exposure to violence affects the people's sense of security and is a major factor in shaping their allegiance to one side or another. *Ceteris Paribus*, people will support the side that provides a better sense of security [17, 28]. A key question is, how do people remember violence; is it the first exposure to violence that shapes their behavior or the last such encounter? The dynamic differential equations model in [27] takes into account the violence intensity ratio of the government and the insurgents, the effectiveness of coercion by the insurgents, the targeting accuracy (see Section “Eyes and Ears”), and the way civilians remember and respond to violence. The main conclusions of the analysis are (i) excess violence and poor targeting accuracy may lead to situations where civilians' support for a certain side will vanish; (ii) the government should not be discouraged by an initial small level of popular support, because there are situations where this would actually play to its advantage if the insurgents are very violent and have poor situational awareness; (iii) the effect of the initial distribution of opinions (support or opposition to a certain side) among civilians on the outcome of the insurgency depends on the way people remember and respond to violent experiences. For some responses the outcome is insensitive to this initial distribution.

BULLETS AND FIRE

As observed by Deitchman [3] and others (see Section “Eyes and Ears”) counterinsurgency

is essentially a contest of attrition that lends itself to classical combat modeling [29, 30].

Confronting Entrenched Insurgents

Similarly to Deitchman [3], Kaplan *et al.* [31] used modified Lanchester models to study the force allocation problem of both the government and the insurgents and, using a sequential force allocation game between the two sides, obtain an equilibrium. It is shown that the insurgents' optimal strategy depends on the government level of situational awareness; when the government has perfect intelligence in equilibrium, the insurgents concentrate their force in a single stronghold that the government either attacks or leaves unengaged, depending on the resulting casualty count. Otherwise, under reasonable assumptions regarding the government's behavior and intelligence capabilities, it is optimal for the insurgents to “spread out” in a way that maximizes the number of soldiers required to win all battles. This type of behavior was observed during the 2006 war against the Hezbollah in Lebanon, and the 2014 war against the Hamas in the Gaza Strip. On the other hand, for a given allocation of insurgents across strongholds, it is shown that an optimal selection of insurgents' strongholds to attack can be (approximately) accomplished with a simple knapsack rule that depends on the force size of the insurgents in a certain stronghold, the one-on-one fire-exchange relative strength, and the level of situational awareness.

Controlling Territories

Also based on a Lanchester setting, a model that accounts for the split between territorial regions loyal to the government and regions favoring the insurgents is described in [32]. Let B denote the government forces and R the insurgents. Also, the fraction of regions inhabited by government supporters is S and the fraction of insurgents' supporters country is C , $S + C = 1$ —see Fig. 3. SB and CR indicate the fraction of regions that are liberated. That is, SB is the proportion of the S regions approvingly controlled by the government, and CR is the proportion of the C regions approvingly controlled by the insurgents. Similarly, SR and CB indicate subjugated regions—regions coercively controlled by the other side. Solid lines in Fig. indicate changes in control due to liberation, while dashed lines indicate subjugation. Regions do not change allegiances even under occupation, and each side can exert military force, to liberate or subjugate, only out of regions that support that side. It is shown that contrary to classical Lanchesterian insights regarding traditional force-on-force engagements, the outcome of an insurgency is independent of the initial force sizes; it will only depend on the fraction of regions supporting each side and the combat effectiveness of each side. Moreover, unlike legacy force-on-force situations, counterinsurgency allows for stalemates (see also Section “Eyes and Ears”). Very often it is hard to avoid stalemate without foreign support. The foreign support in Libya in 2011 led to a victory by the insurgents, while lack of such support in Syria has resulted in an ongoing stalemate where some regions are controlled by the Assad forces and others are controlled by the insurgents. The model's predictions have been consistent with the recently observed situations (2010–2014) in Afghanistan, Libya and Syria.

Tactical Issues

A common threat in insurgencies is roadside attacks by improvised explosive devices [33]. A model addressing this problem is given in [34]. On the basis of collected data on attacks against coalition forces in Iraq, the paper presents a stochastic, game-theoretic, model

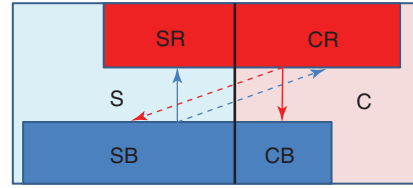


Figure 3. Schematic dynamics of the territorial model.

for optimally allocating clearing devices by the Coalition forces on a network of roads. The insurgents are strategic; they observe the Coalition forces' actions and they react to them. A related problem is the optimal interaction between a single military convoy of government forces and a single route-clearing team operating on a single roadway [35].

Another common threat by insurgencies—mostly used as terrorizing and coercing means against civilians—is suicide bombing. A person acts as a live bomb, killing and harming other people while killing himself or herself. The physics of this type of attack is studied in [36], where the effect of crowd blocking emerges as a significant factor in estimating the number of casualties. Mitigation tactics for this type of attack are examined, using probability models, in [37]. It is shown that even under best-case assumptions regarding the effectiveness and timeliness of widespread deployment of detection sensors against potential attackers, the expected number of casualties will not be significantly reduced compared to no detection. Thus, suicide-bomber-detector schemes will not likely to prove effective in protecting civilian populations from random pedestrian suicide-bomber attacks. The effort should be focused in prevention, rather than mitigation. In other words—invest in intelligence.

SUMMARY

Insurgencies are different than regular armed conflicts in the way, they are physically manifested and the role civilians play in them. These differences are reflected in mathematical models that attempt to study,

analyze, and understand this type of confrontations. From purely attritional focused models that characterize legacy force-on-force abstractions, insurgency models move to behavioral, social, political, and economic domains. The main challenges in insurgency modeling lie in these domains and are mostly associated with data collection and interpretation. Insurgency modeling will experience a huge leap in application and relevance once effective and reliable data-collection methods will be in place for monitoring public mood, sentiments, and behavior doing such armed conflicts. The advent and proliferation of observable social networks may give this capability a significant boost.

REFERENCES

1. Lanchester FW. *Aircraft in warfare: the dawn of the fourth arm*. London: Constable; 1916.
2. Tse-tung M. *On Guerrilla warfare* (Translated by S. B. Griffith II). Chicago: University of Illinois Press; 2000.
3. Deitchman SJ. A Lanchester model of Guerrilla warfare. *Operations Research* 1962;10:818–827.
4. McCormick G. *The Shining Path and Peruvian terrorism*. Santa Monica, CA: RAND; 1987.
5. Connable B, Perry WL, Doll A, *et al.* Modeling, simulation, and operations analysis in Afghanistan and Iraq. Santa Monica, CA: RAND; 2014.
6. V. M. Bier, *Game-theoretic methods in counterterrorism and security*, *Wiley Encyclopedia of Operations Research and Management Science*, Wiley, 2011, 10.1002/9780470400531.eorms1035.
7. Anderson EG. A dynamic model of counterinsurgency policy including the effects of intelligence, public security, popular support, and insurgent experience. *System Dynamics Review* 2011;27(2):111–141.
8. Kilcullen D. *Counterinsurgency*. New York: Oxford University Press; 2010.
9. Berman E, Shapiro JN, Felter JH. Can hearts and mind be bought? The economics of counterinsurgency in Iraq. *J. Political Economy* 2011;119(4):766–819.
10. Schaffer MB. A model of 21st century counterinsurgency warfare. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 2007;4:252–261.
11. Bohorquez JC, Gourley S, Dixon AR, *et al.* Common ecology quantifies human insurgency. *Nature* 2009;462(7275):911–914.
12. Johnson N, Carran S, Botner J, *et al.* Pattern in escalations in insurgent and terrorist activity. *Science* 2011;333:81–84.
13. Schaffer MB. Lanchester models of guerrilla engagements. *Operations Research* 1968;16:457–488.
14. Kress M, Szechtman R. Why defeating insurgencies is hard: the effect of intelligence in counterinsurgency operations—a best case scenario. *Operations Research* 2009;57(3):578–585.
15. Kress M, MacKay NJ. Bits or shots in combat? The generalized Deitchman model of guerrilla warfare. *Operations Research Letters* 2014;42:102–108.
16. Condra LN, Shapiro JN. Who takes the blame? The strategic effects of collateral damage. *American Journal of Political Science* 2012;56(1):167–187.
17. Hammes TX. Countering evolved insurgent networks. *Military Review*, July–August, 18–26 2006.
18. Kadane JB. Optimal whereabouts search. *Operations Research* 1971;19:894–904.
19. Kress M, Lin K, Szechtman R. Optimal discrete search with imperfect specificity. *Mathematical Methods of Operations Research* 2008;68:539–549.
20. Feinstein JS, Kaplan EH. Counterterrorism intelligence operations and terror attacks. *Public Choice* 2011;149:281–295.
21. Farley J. Evolutionary dynamics of the insurgency in Iraq: a mathematical model of the battle for hearts and minds. *Studies in Conflict and Terrorism* 2007;30:947–962.
22. Epstein J. Modeling civil violence: an agent-based computational approach. *Proceedings of the National Academy of Sciences* 2002;99:7243–7250.
23. Cioff-Revillai C, Rouleau M. MASON Rebelland: an agent-based model of politics, environment, and insurgency. *International Studies Review* 2010;12:31–52.
24. Blair G, Fair CC, Malhotra N, *et al.* Poverty and support for militant politics: evidence from Pakistan. *American Journal of Political Science* 2013;57(1):30–48.
25. Kuran T. Sparks and prairie fire: a theory of unanticipated political revolution. *Public Choice* 1989;61:41–74.

26. Atkinson MP, Kress M, Szechtman R. Carrots, sticks and fog during insurgencies. *Mathematical Social Sciences* 2012;64:203–213.
27. Atkinson MP, Kress M. On popular response to violence during insurgencies. *Operations Research Letters* 2012;40(4):223–229.
28. Lynn JA. Patterns of insurgency and counterinsurgency. *Military Review*, July–August 2005;22–27.
29. Washburn A, Kress M. *Combat modeling*. New York: Springer; 2009.
30. Kress M. Modeling armed conflicts. *Science* 2012;336(6083):865–869.
31. Kaplan EH, Kress M, Szechtman R. Confronting entrenched insurgents. *Operations Research* 2010;58(2):329–341.
32. Atkinson MP, Gutfraind A, Kress M. When do armed revolts succeed: lessons from Lanchester Theory. *Journal of the Operational Research Society* 2011;63:1363–1373.
33. Wilson C. *Improvised Explosive Devices (IEDs) in Iraq and Afghanistan: effects and countermeasures*. Washington DC: Congressional Research Service, The Library of Congress; 2006.
34. Washburn A, Ewing PL. Allocation of IED assets in IED warfare. *Naval Research Logistics* 2011;58(3):180–187.
35. Kolesar P, Leister K, Stimpson D, *et al.* A simple model of optimal clearance of improvised explosive devices. *Annals of Operations Research* 2013;208:451–468.
36. Kress M. The effect of crowd density on the expected number of casualties in a suicide attack. *Naval Research Logistics* 2005;52(1):22–29.
37. Kaplan EH, Kress M. Operational effectiveness of suicide bomber detector schemes: a best-case analysis. *Proceedings of the National Academy of Science* 2005;102(29):10399–10404.

ANALYTICS IN RETAIL

DIEGO KLABJAN

YAN JIANG

Department of Industrial Engineering and
Management Sciences, Northwestern
University, Evanston, Illinois

LOREN WILLIAMS

EVP and Chief Scientist Predictix, LLC,
Atlanta, Georgia

Thanks to the abundant consumer transaction data collected and ample cheap computing power, retailers have begun to develop and employ analytical methods as decision support tools for their various operation and management tasks. They hire either analytically trained individuals or specialized retailing consulting firms to develop their proprietary software for that purpose. In either case, they turn to the wealth of academic models for ideas of developing their own practical models. Generally, however, there are differences between the academic and practical models. Sometimes, academic models are overly simplified such that a lot of challenging issues are left unsolved; sometimes, academic models are too complicated such that it would be too hard or costly to implement them in practice. In this article, we focus on two types of commonly used analytics in retail practice, forecasting, and optimization, and present two models, motivated by academic research, that are in production use or development in retail business settings.

FORECASTING

Forecasting is one of the earliest and most common analytics carried out in retail practice. It provides vital input to almost all of the other analytics in retail functions such as marketing, merchandising, and operation, and for management tasks such as planning, budgeting, and controlling. Its methods can be roughly divided into judgmental and statistical methods. Judgmental methods

prevailed in the early days of retailing, when there were only limited products as well as scarce recorded data. These methods are subjective and are nearly pure art with little science. They are still useful tools for forecasting sales in some cases such as innovative products (e.g., the iPad) or fashion goods, which have no direct relevant historical sales. And, expert judgment can convey valuable knowledge and experience, which can be used to improve the performance of statistical methods [1]. But, statistical methods are the basic tools for modern forecasting system in retail. They make use of the large-scale data collected through an IT system. They are objective, scientific, and can be automated. Statistical forecasting methods can be further divided into extrapolation and causal methods. Techniques of extrapolation methods include the simpler moving average and exponential smoothing family, and the more sophisticated Box–Jenkins approach [2]. They use only the time series data of the forecasting subject. Franses [3] discusses the application of extrapolation methods for business and economic forecasting. Causal methods, on the other hand, build statistical models using both the data of the forecasting subject and potential causal factors. Some causal factors, such as prices, promotions, advertising, are under the control of management, and the others, such as competitor prices, weather, changes in competitive landscape, and market demographics are not. Despite the simplicity of judgmental and extrapolation methods in retail practice, the more complicated causal forecasting methods play an important role in the retail industry and are the focus of this article.

In what follows, we use a promotion forecasting example to illustrate these causal forecasting methods. This example employs a multiple linear regression model for forecasting, which is most often used in causal methods. We highlight some of the issues commonly encountered in retail forecasting using causal statistical methods, and an innovative solution for one of them.

Promotion Forecasting

An important class of decisions made by retailers is related to planning a promotional strategy. These decisions include which products to promote; what promotional prices to offer; how to communicate the promotional offers to customers, either via various media channels or in-store; and how to execute the promotions. Promotion decisions may be made on a chain-wide basis, or may be tailored to specific markets, store types, and stores. These decisions can be informed and supported by promotion forecasting, which aims to forecast the magnitude of the impact of a particular promotion on both the promoted products, and products complementary or substitute to the promoted ones. Promotions forecasting methods embedded in decision support have been shown to provide substantial increments in revenue and gross profit. The benefits accrue in three areas. First, retailers can make more informed decisions about promotional plans, including items, timing, and targets, subject to a given level of support from vendors. By basing such decisions on better understanding of the profitability of a promotion, the retailer can lessen the risk of unprofitable promotions, and make the most effective use of promotional “budgets.” Second, the retailer is able to provide higher service levels on the promoted items by improving the forecast accuracy, which translates into fewer lost sales and greater revenue. Finally, better forecasts for promotional items can also improve inventory management, and hence reduce the ordering and holding costs.

The forecasting method described in this section features an integrated demand model, which captures the sales effects of various promotional features, both in the item being promoted and other related items. Ordinary least square estimates of these effects, calibrated on historical sales and ancillary data, are used first in a promotional planning tool to support choices on promotional activities, and, once plans are finalized, in the inventory management tools to plan inventory for items influenced by the planned promotions.

Forecast Constituents

Secular Effects. Secular effects are predictable effects of phenomena that are time based. Although predicting these effects is not the chief goal of the promotion forecasting model, they are important for two reasons. First, in order to estimate the effects of promotional features, it is necessary to “untangle” the effects of the promotion on sales from the effects of these other phenomena. Second, when the retailer is considering the specific details of a promotion under consideration, the ultimate effects of the promotion on sales often depend on the level of sales when the promotion is absent, which in turn depends on these secular effects. Our model incorporates three secular effects: demand level, seasonal effects, and holiday effects; it could be expanded to include a trend effect as well. The demand level is interpreted as the average (or deseasonalized) sales under a regular pricing regime. The seasonal effects indicate how sales vary over a year. They are captured at the weekly level in the promotion forecasting model, which works well in practice. We include holiday effects in addition to the weekly seasonal effects because the same holiday may not happen at the same week each year.

Own Effects. Own promotional effects are the influence of the promotional features on the sales of the item being promoted. Any particular promotion is characterized by a vector of promotional features. These features can be categorized into four classes: the discount or temporary price reduction; the mechanism by which the offer is extended (e.g., straight discount, “buy one, get one free”); the way in which the offer is communicated or promoted in the store (e.g., on shelf and aisle end cap); and the way in which the offer is communicated outside the store (e.g., circular front page, brand advertising). For example, a particular promotion may be a 50% discount on shelf promotion with “buy one, get one free” mechanism and circular front page advertising. To forecast the own effects for such a promotion, we just need to combine

the effects of its feature vector. These effects are combined multiplicatively.

Own effects of a promotion can either be on sales that occur contemporaneously with the promotion or on sales before or after the promotion. The effect on sales before or after the promotion is sometimes referred to as *retiming*, *self-cannibalization*, or *pantry loading*. It occurs when consumers either defer purchases until the promotion is in effect, if it can be anticipated, or when consumers stock up on the item during the promotion and purchase less afterwards. Our model incorporates both the contemporaneous own effects and the retiming effects.

Cross-Effects. Cross-effects are the influence of the promotion on items other than the one being promoted. Cross-effects can happen either for complement or substitute items. Items that are substitutes for the promoted item may experience a decrease in sales. These are sometimes referred to as promotional *victims* (e.g., different pack sizes of the same item and other brands of the same item). Other items, which are complements to the promoted item, may experience an increase in sales. These are sometimes referred to as *halo* items (e.g., snacks, when beer is promoted and accessories, when furniture is promoted).

Another type of cross-effects arises when a promotion is undertaken to stimulate traffic in the whole store, from which sales of many items may benefit. Some of these items may be neither complement nor substitute to the promoted items. The promoted item is sometimes referred to as a *loss leader* or a *traffic-driver*. A typical example of a loss leader in grocery is milk. Retailers sometimes reduce the price of milk less than its cost to attract shoppers to their stores. Our model incorporates both types of cross-effects.

Full Model. We model the effects in the promotion forecast multiplicatively. For seasonal and holiday effects, this is merely a choice of convenience, and is consistent with most other work in the area of retail forecasting. For the own- and cross-effects, the multiplicative formulation is motivated by the fact that the own- and cross-effects clearly

depend on the baseline sales of the promoted item and related items. Multiplicative models automatically capture this kind of dependence while additive models cannot.

The full forecast model for item i is given by

$$\begin{aligned} \text{Sales}_{i,t} = & \text{Level}_i \times \text{SeasonalEffects}_{i,t} \\ & \times \text{HolidayEffects}_{i,t} \\ & \times \prod_{k=1}^K \beta_{i,k}^{\text{own}} \times \text{PromoFactor}_{k,t} \\ & \times \prod_{l=1}^L \prod_{j=1}^J \beta_{i,l,j}^{\text{cross}} \text{RelatedItemPromoFactor}_{l,j,t}, \end{aligned}$$

where k indexes the factors representing the promotional features for item i , which are being offered at time t , and β^{own} are the own effects for each of the K factors; l indexes the items being promoted with cross-effects on item i ; j indexes the attributes of each of the L other items being promoted at time t and β^{cross} are the cross-effects for each of the J factors.

Implementing such a model requires making choices about inclusion of effects and the way in which effects are modeled. We discuss some of these choices with an example.

In this promotion planning example, the promotion under consideration is for a branded grocery product, peanut butter. The features of the promotion are a “buy one, get a second for half price” offer and a large display at an end of an aisle. For simplicity we consider only two related items, a larger package of the same brand and a similar size package of a house brand. The promotion is planned to run for two weeks, corresponding to week indices 13 and 14. The estimated effects are presented in Table 1.

The first column displays the own effects. The demand level and seasonal effects provide the baseline (i.e., nonpromoted) forecasts. This promotion offers an effective discount of 25%; we have an estimated effect (demand lift) of 1.32 for discounts in the range of 20–25%. An alternative way to model this effect is to treat the discount as a continuous variable and the effect as price elasticity. An advantage of the discrete discount buckets

Table 1. Promotions Forecasting Effects

Promotion Forecast Effects	Premium Brand 14 oz	Premium Brand 32 oz	House Brand 15 oz
Demand level	66	27.5	83.7
Seasonal effect, week 13	1.16	1.16	1.16
Seasonal effect, week 14	1.16	1.16	1.16
Seasonal effect, week 15	1.14	1.14	1.14
20–25% discount	1.32		
Buy one, get one half off	1.04		
Major end aisle display	1.23		
Retiming effect	0.78		
Cross-effects of 14 oz premium brand promo with display		0.41	0.68

is that it is nonparametric, offering a more flexible way to model the discount effects. The promotional mechanism “buy one, get one half off” is estimated to have a lift of 1.04. The fact that this is greater than 1.0 indicates that this form of the offer is more effective than simply offering a 25% discount. The way by which the promotion is communicated in the store, using a large end cap display, is estimated to have a lift effect of 1.23. Finally, we have a retiming effect of 0.78 on the sales of the promoted peanut butter the following week.

For many retailers, there can be a large set of promotional mechanisms and in-store communications, as well as promotional advertising possibilities. For the purpose of modeling and estimating the effects, these can be consolidated into a smaller number of alternatives of each of the three main categories.

The second and third columns in Table 1 display the baseline forecast effects and the cross promotional effects for the other two items in the subcategory. The cross-effects predict the sales impact of a promotion on the 14 oz Premium brand, on each of these other two items, given that the promotion includes a display. In the present case, the 0.41 effect for the 32 oz Premium brand indicates a large reduction in sales volume, due to a high degree of substitutability across sizes of identical items. The effect on the sales of the like-sized House brand is less, indicating a lesser degree of substitutability with the promoted item.

As with the own effects, there are a variety of ways to model the cross-effects. In this

example, we strike a compromise between the simplest model, which reflects only the presence or absence of a promotion of a related item, and a much more complex model, which attempts to estimate the effects of all the various promotional attributes on the sales of the related item. Note also that, in this promotions planning example, there are no complementarity cross-effects.

Challenges

Data Preparation. The presented forecast model is based on information from the point-of-sales (POS) and the promotion planning and execution systems. Neither the POS nor the promotion execution systems were designed with a view toward supplying data to support a forecasting model and process, and thus the source data require inspection and cleansing prior to being used. The primary hygiene tasks are to ensure the plausibility of data values and consistency of data across systems. Once these checks are accomplished, the sales data are summed to weekly values, whose start and end dates correspond to the promotions calendar. These values are then further cleansed of outliers and are matched up to the historical promotion information, providing the essential ingredients to estimate the effects.

Typical Problem Sizes. In 2008, the average number of Stock Keeping Units (SKUs) carried in a typical US supermarket was 46,852 according to the Food Marketing Institute. The number of stores a large retail chain may have is in the order of thousands.

For example, Wal-Mart has more than 8000 retail stores under 53 different banners in 15 countries [4].

Promotion planning systems will typically support dozens to hundreds of different promotion features. Localization of promotional planning and execution of some large retailers can increase the large-size retail forecasting problem even further. For example, a grocery chain with approximately 1000 stores may have 25–30 different price zones in which there are different promotional effects to forecast. Although costs of computing are not much of a constraint, especially with the advent of cloud computing, the large size of the problem requires careful design of the automation of the estimation process, which can take many hours to execute. Further, with so many combinations of item, location, and period, the majority of historical data used to calibrate the model is sparse. Aggregation and pooling are strategies for facilitating estimate of the effects in the presence of sparsity.

Aggregation and Pooling. In forecasting problems amenable to the extrapolation or time series methods, sparsity is often addressed through aggregation and disaggregation methods [5,6]. In promotional forecasting models (and causal models, generally), aggregation is problematic, since SKUs which might be aggregated are likely to have histories with different promotion timing and features.

In causal forecasting models, pooling data for different items and different locations can improve the estimability and reliability of effect estimates. The challenge is to devise a means of defining pools or clusters of items that are homogeneous in the effects.

One approach to find item-level clusters is a two-stage method. We first assign each item an attribute vector. After that, we can group items together according to the assigned attribute vector by some clustering algorithm like K-means or hierarchical clustering. For example, Zotter *et al.* [5] used normalized store sales to group stores with similar seasonal effects. Products are often grouped according to certain physical or usage attributes, such as product category

or performance grade. The problem with this approach is that it is hard, *a priori*, to align the attribute vector with the regression effects across which we want to pool.

KMeans—GA Algorithm. An alternative approach not relying on these artificially assigned attribute vectors is cluster-wise linear regression. The basic idea of this modeling approach is to consider clustering and regression concurrently. In this way, the difficulty of aligning the attribute vector and the regression effects is bypassed.

We have developed a heuristic algorithm, termed *the KMeans-GA algorithm*, which is able to find clusters of similar items efficiently and effectively. This KMeans-GA algorithm hinges on the work by Maulik and Bandyopadhyay [7], where the authors embed a K-means procedure into the genetic algorithm framework for the standard clustering problem. Along the same line, our KMeans-GA algorithm also embeds a structure exploring K-means-like procedure into the genetic operations. For our cluster-wise linear regression problem, each cluster of items is encoded by a vector that stores the regression effects. These regression effects define similarity among items, according to which the K-means part assigns items to different clusters. The vector of regression effects is updated by running regression over new clusters generated through the K-means part. The genetic algorithm part performs genetic operations on the population of vectors of regression effects. The genetic operations include the standard selection, crossover, mutation, and elitism. We refer to Fig. 1 for a simple example that illustrates the KMeans-GA algorithm. In this example, each cluster i has the same regression model $y = \sum_{j=1}^3 x_j \beta_j^i$ with different estimates for the regression effects β_j^i 's, where the superscript i indicates cluster i . We divide the items into two clusters.

We tested our KMeans-GA algorithm on the promotion forecasting problem to find SKUs with similar seasonality patterns. We found that the heuristic approach performs very well; the difference in total sum of squared regression errors between our heuristic and optimal solutions is below

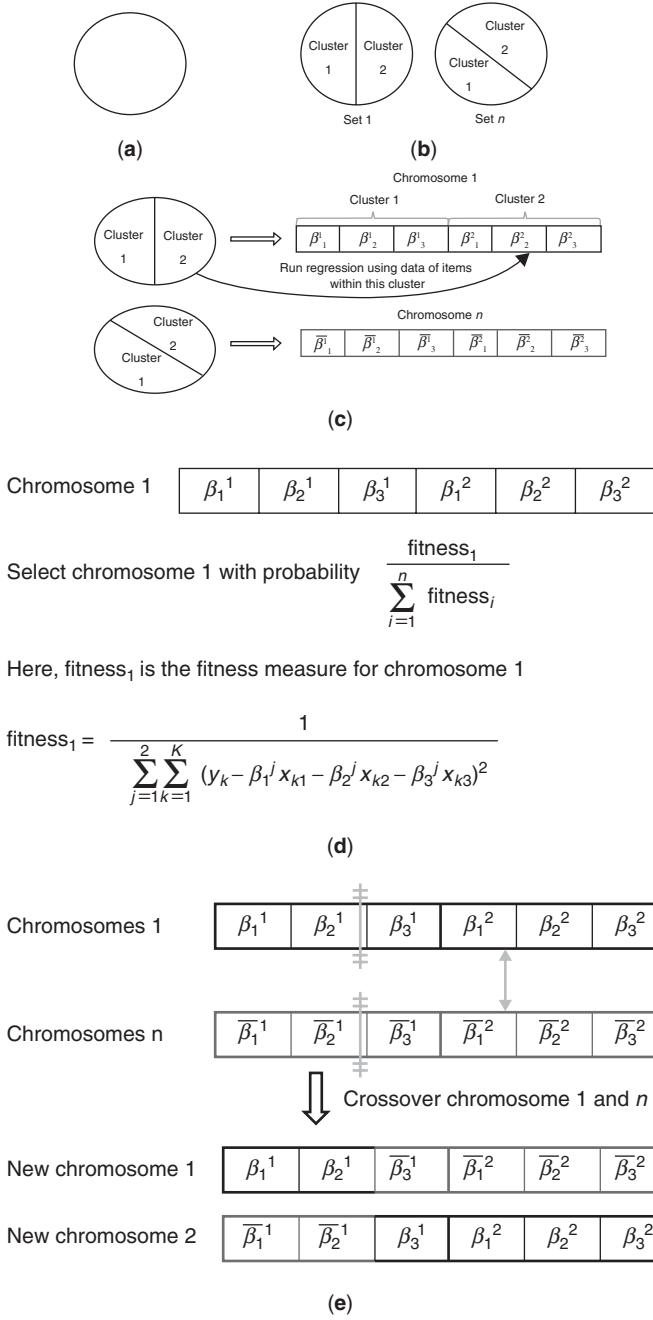
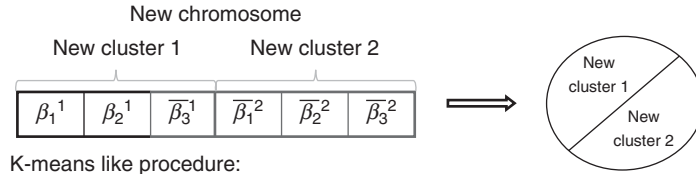


Figure 1. Illustration of the KMeans-GA algorithm. (a) Set of items to cluster; (b) random generation of n sets of clusters; (c) encoding of each set of clusters using corresponding regression effects; (d) selection of two chromosomes for genetic operations; (e) crossover in genetic operation; (f) generation of new set of clusters through K-means-like procedure; (g) update of coding for new chromosome.

2% in the test cases. In addition, the improvement of our KMeans-GA algorithm over the attribute-based two-stage method is significant; reductions in Sum-of-Squares Error (SSE) of greater than 30%. Finally, in a number of test cases, the procedure

revealed evidently better and more granular predictions about seasonal effects than the two-stage method. Figure 2 shows the seasonal effects estimated for a group of SKUs pooled on the basis of an attribute, the subcategory of the product. The only



K-means like procedure:

Item A is assigned to new cluster 1 if and only if

$$\sum_{k=1}^K (y_k - \beta_1^1 x_{k1} - \beta_2^1 x_{k2} - \beta_3^1 x_{k3})^2 \leq \sum_{k=1}^K (y_k - \beta_1^2 x_{k1} - \beta_2^2 x_{k2} - \beta_3^2 x_{k3})^2$$

Otherwise, Item A is assigned to new cluster 2.

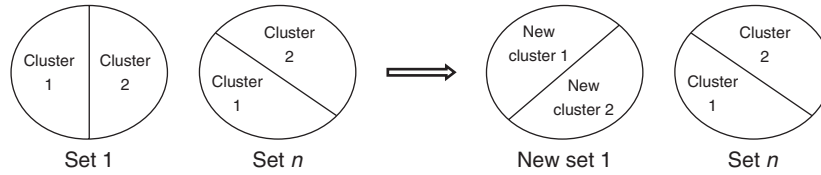
Here, $(y_k, x_{k1}, x_{k2}, x_{k3})_{k=1}^K$ is the data for item A

(f)



(g)

Replace chromosome 1 with new chromosome 1 if the latter has larger fitness.



(h)

Figure 1. (Continued)

significant seasonal pattern was observed in weeks 51 and 52. For the same set of SKUs, the method revealed pools with appreciably different and more pronounced seasonal effects, as shown in Fig. 3. These seasonal patterns were confirmed by retail experts for the SKUs under study.

OPTIMIZATION

Optimization is not a commonly used tool in retailing as forecasting. Retailers are more comfortable making decisions by asking “what-if” questions than resorting to optimization models. The reason is the lack of the underlying intuition of an optimization model in the absence of technical background

and confidence. But thanks to recent collaborations between researchers in academia and some leading retailers, selected optimization models have made their headway into retail practice. Among them, assortment planning and price optimization models are the two prominent classes. Nevertheless, there are gaps between academic and industry models. Please refer to the section titled “Retail Applications” in this encyclopedia for academic models for assortment planning and price optimization. In what follows, we introduce a novel optimization model jointly optimizing price, assortment, and presentation (i.e shelf space) for a group of substitutable products. The model strikes a balance between complexity, relevance, and adoptability.

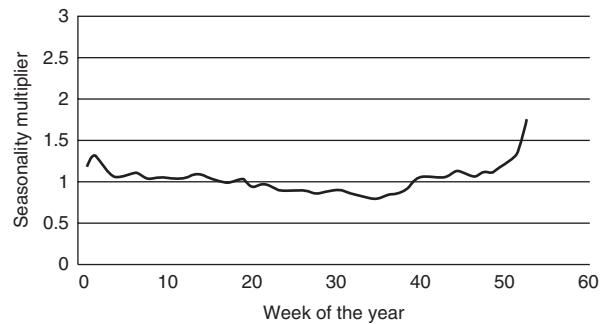


Figure 2. Seasonal pattern for an attribute-based cluster of approximately 350 SKUs.

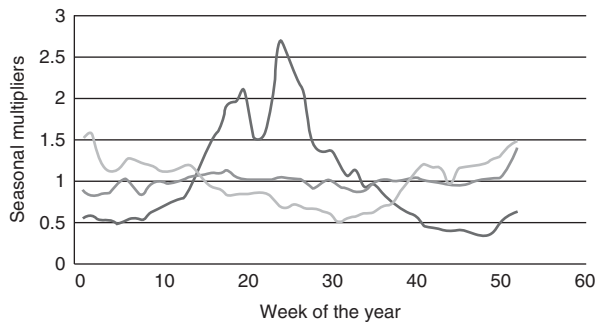


Figure 3. Three distinct seasonal patterns revealed for same SKUs.

Optimizing Price, Assortment, and Presentation

In a typical retailer, macrolevel assortment decisions, for example, how much space to allocate to each category or department, are strategic decisions driven by considerations of the image or position of the chain, an appreciation of what items have been winners and losers, relationships with vendors, and the size and layout of the store. As a matter of operational practice, these macro-level assortment decisions either ignore heterogeneity in tastes and preferences across markets, or accommodate it coarsely by differentiating assortment plans by geographic store groups. More micro-level decisions are typically made by buyers or category managers and presentation decisions are typically undertaken by space planners, especially in grocery retailing. They have to decide, in each subcategory or finer categorization of items, precisely what size package of which brand and how many facings of each to display. These decisions are made on the basis of the particular display space available, visual appeal, negotiations with vendors, and received wisdom about

what has “worked” in the past. Finally, item-level pricing decisions are the responsibility, in some cases, of a pricing department, and, in other cases, by the merchants or category managers. For most retailers, these decisions are made based on judgment, informed of high-level policy rules governing markups, price image, and promotional positioning. In some more sophisticated retailers, analytic-based tools are used to support these decisions, sometimes including price optimization tools. The latter rely on models of demand which predict sales changes due to price changes, and may even include substitution and complementarity effects. But even these sophisticated price optimization models are insensitive to assortment and presentation decisions.

Studies of these assortment, pricing, and presentation decisions suggest that category profits may be improved by upward of 50% by independently optimizing these decisions, [8–12].

Demand Effects

Substitutable Item Groups and Price Effects. A “substitutable item group” (SIG), is a set

of candidate items to offer and present which are strong, but imperfect substitutes for each other. They would typically comprise items of the same brand in different packages and flavors, as well as comparable items in different brands. A typical grocery retailer, for example, will sell 30,000–50,000 items in 50–150 categories. Each category would comprise 15–100 SIGs. Thus, an SIG may comprise from as few as 2 to 20 or more items. In practice, on an average, there are 7–10 items per SIG.

We first model demand for the SIG as a whole, and then use a demand share model to obtain the demand for each individual item. Let M be the set of all items in the SIG and M^* be the subset of M comprising the items actually assorted. Quantity Q_{SIG} , the total demand of the entire SIG, is assumed to be responsive to $\mathbf{p} = (p_i)_{i \in M^*}$, the selling prices of the items offered:

$$Q_{\text{SIG}}(\mathbf{p}, M^*) = Q_{\text{SIG}}^r \left(\frac{p_{\text{SIG}}(\mathbf{p}, M^*)}{p_{\text{SIG}}^r(M^*)} \right)^{e_{\text{SIG}}}, \quad (1)$$

where Q_{SIG}^r is the reference sales for the SIG under the reference prices, e_{SIG} is the own-price elasticity of demand for the SIG, and p_{SIG} and p_{SIG}^r are the reference share-weighted selling prices and reference prices, respectively, of the items that are offered in the SIG. More specifically,

$$p_{\text{SIG}}(\mathbf{p}, M^*) = \sum_{i \in M^*} S_i^r p_i \Big/ \sum_{i \in M^*} S_i^r,$$

$$p_{\text{SIG}}^r(M^*) = \sum_{i \in M^*} S_i^r p_i^r \Big/ \sum_{i \in M^*} S_i^r,$$

where S_i^r is the demand share for item i under the reference prices p_i^r .

To model the demand shares assorted for individual items, we include an “outside” good to account for no-purchase decisions. Both the own items’ shares and the outside good’s shares respond to their prices:

$$ST(\mathbf{p}, M^*) = \sum_{i \in M^*} \left(S_i^r \left(\frac{p_i}{p_i^r} \right)^{e_i} \right)$$

$$+ S_{\text{OG}}^r \left(\frac{p_{\text{OG}}}{p_{\text{SIG}}} \right)^{e_{\text{OG}}},$$

$$S_i(\mathbf{p}, M^*) = S_i^r \left(\frac{p_i}{p_i^r} \right)^{e_i} / ST(\mathbf{p}, M^*) \quad i \in M^*,$$

$$S_i(\mathbf{p}, M^*) = 0 \quad i \notin M^*,$$

$$S_{\text{OG}}(\mathbf{p}, M^*) = S_{\text{OG}}^r \left(\frac{p_{\text{OG}}}{p_{\text{SIG}}} \right)^{e_{\text{OG}}} / ST(\mathbf{p}, M^*),$$

where e_i is the own-price share elasticity of item i , and e_{OG} analogously for the outside good. S_i^r and S_{OG}^r are the reference shares of item i and the outside good, respectively; they can be thought of as the intrinsic preference weights for the item. Quantity $ST(\mathbf{p}, M^*)$ normalizes the shares to ensure that they sum to unity. We take p_{OG} as fixed, since it is outside the control of the decision maker; note, however, that S_{OG}^r does depend on $\mathbf{p}$. This share model is equivalent to the widely used multiplicative competitive interaction (MCI) model due to Nakanishi and Cooper [13–16].

In its most basic form, demand Q_i of item i and demand Q_{OG} for the “outside” good is given by

$$Q_i(\mathbf{p}, M^*) = Q_{\text{SIG}}(\mathbf{p}, M^*) S_i(\mathbf{p}, M^*) \quad (2)$$

and

$$Q_{\text{OG}}(\mathbf{p}, M^*) = Q_{\text{SIG}}(\mathbf{p}, M^*) S_{\text{OG}}(\mathbf{p}, M^*).$$

Note that by definition we have $\sum_{i \in M^*} S_i(\mathbf{p}, M^*) + S_{\text{OG}}(\mathbf{p}, M^*) = 1$.

Assortment or Variety Effects. Retailer experience suggests that having a larger assortment, with greater variety, can lead to larger sales. This can be understood either as accommodating consumer heterogeneity, or as a consumer preference for variety [17]. We use a simple extension of the basic SIG demand model in Equation (1) to accommodate these effects:

$$\bar{Q}_{\text{SIG}}(\mathbf{p}, M^*) = \left[\left(Q_{\text{SIG}}^r \left(\frac{p_{\text{SIG}}(\mathbf{p}, M^*)}{p_{\text{SIG}}^r(M^*)} \right)^{e_{\text{SIG}}} \right) - Q_{\text{OG}}(\mathbf{p}, M^*) \right] |M^*|^{\delta_{\text{SIG}}},$$

where $0 < \delta_{\text{SIG}} < 1$ is the “assortment elasticity.” While this is not strictly elasticity, we adopt that label for expository convenience and restrict parameter values to model diminishing returns to assortment breadth. Note that here the effects of assortment on demand of an SIG depend only on the size of the assortment.

Presentation Effects. As is the case with the assortment or variety effects, retailers acknowledge that the extent of a presentation of an item is also correlated with sales for that item. A greater number of facings of an item will make it more prominent and attract more demand, and the more facings allocated to an item, the less likely that the display quantity will be depleted prior to restocking. We employ a simple extension of the basic item quantity model in Equation (2) to accommodate these effects:

$$Q_i(\mathbf{p}, M^*, f_i) = (\bar{Q}_{\text{SIG}}(\mathbf{p}, M^*) S_i(\mathbf{p}, M^*)) f_i^{\gamma_i},$$

where f_i is the number of “facings” of item i and $0 < \gamma_i < 1$ is the “presentation elasticity” of item i . Again, this is not strictly elasticity, but we adopt that label for expository convenience and restrict parameter values to model diminishing returns to presentation intensity. Finally, we note that if $f_i = 0$, there are no sales, which corresponds to the case of i being excluded from the assortment. The set of assorted items, M^* , is determined by $\mathbf{f} = (f_i)_{i \in M^*}$ as follows:

$$M^*(\mathbf{f}) = \{i \mid f_i > 0\}.$$

Full Model and Parameters. Combining all these elements, we have an item-level sales model that reflects the effects of prices on all items in the SIG, as well as the items that are assorted and the presentation, reflected in the number of facings of each item:

$$Q_i(\mathbf{p}, \mathbf{f}) = (Q_{\text{SIG}}(\mathbf{p}, M^*(\mathbf{f}))(1 - S_{\text{OG}}(\mathbf{p}, M^*(\mathbf{f}))) |M^*(\mathbf{f})|^{\delta_{\text{SIG}}} S_i(\mathbf{p}, M^*(\mathbf{f})) f_i^{\gamma_i}.$$

We are interested in the choices that confront the decision maker to set values for p_i and f_i

(and thus for M^* , as well). We assume that p_{OG} and p_i^r are fixed and given, and that $Q_{\text{SIG}}, S_i^r, S_{\text{OG}}^r, \delta_{\text{SIG}}, \gamma_i, e_i, e_{\text{SIG}}$, and e_{OG} are parameters to be set by expert judgment or to be calibrated from historical data.

Optimization Model

The decision problem is to choose the prices p_i and the number of facings f_i for each item in the SIG (which subsumes the problem of choosing the set of items to include in the assortment $M^*(\mathbf{f})$). In practice, a retailer may wish to maximize revenue, but, in concept, the optimization problem is to maximize some measure of gross profit. Thus, the model discussed in the previous section yields the following optimization problem:

$$\max_{\mathbf{p}, \mathbf{f}} \sum_{i \in M} (Q_{\text{SIG}}(\mathbf{p}, M^*(\mathbf{f}))(1 - S_{\text{OG}}(\mathbf{p}, M^*(\mathbf{f}))) |M^*(\mathbf{f})|^{\delta_{\text{SIG}}} S_i(\mathbf{p}, M^*(\mathbf{f})) f_i^{\gamma_i} (p_i - c_i),$$

where $p_i - c_i$ is the gross profit for item i .

The dominant constraint in this problem is the space constraint. We assume that each item occupies a certain amount of linear shelf space, per facing. If we have K_{SIG} linear space allocated to the SIG, then we have

$$\sum_{i \in M} w_i f_i \leq K_{\text{SIG}},$$

where w_i is the width of item i . Typically, there are other business constraints, such as an item being required to be assorted, or the maximum space that can be allocated to an item. These can be condensed into a set of range constraints for each item:

$$l_i \leq f_i \leq u_i,$$

where l_i and u_i are minimum and maximum allowable number of facings for item i . It is also typical that there are limits on the magnitude of price changes. For simplicity, we assume that the current prices are the reference prices, which yields for each item constraints

$$p_i^r(1 - \theta_i) \leq p_i \leq p_i^r(1 + \theta_i),$$

where θ_i is the maximum allowable price change of item i .

This model is hard to solve; however, we resort to the fact that the number of items in the SIG is relatively low. To this end, the algorithm enumerates all possible assortments M , and for each of them we solve the resulting nonlinear continuous optimization problem.

CONCLUSION

Models much like the promotion forecasting model in the section titled “Forecasting” are presently used by managers in some retail chains, both for the purposes of planning promotions and improving supply chain decisions. The assortment, space, and price optimization model in the section titled “Optimization” is the subject of active research, but, to the authors’ knowledge, it is not employed at present in a production system. K  k and Fisher [18] present estimation results and management implications from implementing a similar model at a European grocery retailer.

Retailing has been imagined to be “a paradise for operations researchers” [12]. It is rich with interesting problems and an increasing appetite for adapting to some of the methods and tools of OR. We have described a few areas where the fertile interplay between academic research, analytic solution vendors and consultants, and retail decision makers is yielding new and useful OR applications.

REFERENCES

1. Bunn D, Wright G. Interaction of judgemental and statistical forecasting methods: issues and analysis. *Manage Sci* 1991;37(5):501–518.
2. Box GE, Jenkins GM. *Time series analysis: forecasting and control*. Hoboken (NJ): John Wiley & Sons, Inc.; 2008.
3. Franses PH. *Time series models for business and economic forecasting*. Cambridge: Cambridge University Press; 1998.
4. Wal-Mart Stores, Inc. (n.d.). Available at <http://investors.walmartstores.com/phoenix.zhtml?c=112761&p=irol-irhome>. Accessed 2010 Feb 4.
5. Zotter G, Kalchschmidt M, Caniato F. The impact of aggregation level on forecasting performance. *Int J Prod Econ* 2005; 93–94 (8):479–491.
6. Kahn KB. Revisiting top-down versus bottom-up forecasting. *J Bus Forecast* 1998;17(2): 14–19.
7. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. *Pattern Recognit* 2000;33(9):1455–1465.
8. McIntyre SH, Miller C. The selection and pricing of retailing assortments: an empirical approach. *J Retail* 1999;75(3):588–604.
9. Green PE, Savitz J. Applying conjoint analysis to product assortment and pricing in retailing research. *Pricing Strat Pract* 1994;2(3):4–19.
10. Phillips RL. *Pricing and revenue optimization*. Stanford (CA): Stanford University Press; 2005.
11. Talluri K. *The theory and practice of revenue management*. Boston (MA): Kluwer Academic Publishers; 2004.
12. Fisher M. Rocket science retailing: the 2006 Philip McCord Morse lecture. *Oper Res* 2009; 57(3):527–540.
13. Nakanishi M, Cooper LG. Parameter estimation for a multiplicative competitive interaction model: least square approach. *J Mark Res* 1974;11(3):303–311.
14. Nakanishi M, Cooper LG. Simplified estimation procedures for MCI models. *Mark Sci* 1982;1(3):313–322.
15. Cooper LG, Nakanishi M. *Market-share analysis*. Norwell (MA): Kluwer Academic Publishers; 1988.
16. Cooper LG. Market share models. In: Eliashberg J, Lilien GL, editors. *Volume 5, Handbooks in operations research and management science: marketing*. Amsterdam, The Netherlands: Elsevier Science Publishers; 1993. pp. 259–354.
17. Kim J, Allenby GM, Rossi PE. Modeling consumer demand for variety. *Mark Sci* 2002; 21(3):229–250.
18. K  k AG, Fisher ML. *Demand estimation and assortment optimization under substitution: methodology and application*. *Oper Res* 2007;55(6):1001–1021.

ANT COLONY OPTIMIZATION

MARCO DORIGO
MARCO A. MONTES DE OCA
SABRINA OLIVEIRA
THOMAS STÜTZLE
IRIDIA, CoDE, Université Libre de
Bruxelles (ULB), Brussels,
Belgium

Ant colony optimization (ACO) [1–8] is a class of algorithms for tackling optimization problems that is inspired by the pheromone trail laying and following behavior of some ant species. While foraging, ants leave on the ground a chemical substance, called *pheromone*, that attracts other fellow nest-mates [9]. The pheromone trail laying and following behavior of the ants induces a positive feedback process whereby trails with high concentration of pheromones become more and more attractive as more ants follow them [10–12]. As a result, whenever two paths to the same food source are discovered, the colony is more likely to select the shortest one because ants will traverse it faster and thus it will have a higher pheromone concentration than the longer one.

ACO algorithms exploit a mechanism analogous to the one that allows colonies of real ants to find shortest paths. In ACO, (artificial) ants construct candidate solutions to the problem instance under consideration. Their solution construction is stochastically biased by (artificial) pheromone trails, which are represented in the form of numerical information that is associated with appropriately defined solution components, and possibly by heuristic information based on the input data of the instance being solved. A key aspect of ACO algorithms is the use of a positive feedback loop implemented by iterative modifications of the artificial pheromone trails that are a function of the ants' search experience; the goal of this feedback loop is to bias the colony toward the most promising solutions.

The ACO metaheuristic is a high-level algorithmic framework for applying the above ideas to the approximate solution of optimization problems. When applied to a specific optimization problem, this ACO framework needs to be concretized by taking into account the specifics of the problem under consideration and possibly by adding additional techniques such as problem-specific solution improvement procedures. The development of effective ACO algorithm variants has been one of the most active research directions in ACO: this article gives an overview of the most important of these developments. For more information about successful applications of ACO, we refer to the article titled ***A Concise Overview of Applications of Ant Colony Optimization***, in this encyclopedia.

ACO EXAMPLE APPLICATIONS

Perhaps the easiest way to understand how ACO algorithms work is through examples. Here, we present two examples where the ACO algorithms use different solution representations. The first example shows how ACO is applied to solve the traveling salesman problem (TSP), which was the first optimization problem to which an ACO algorithm was applied. The second example concerns the set covering problem (SCP): it shows how ACO algorithms can be used to solve problems using a binary representation.

Example 1. Ant Colony Optimization for the Traveling Salesman Problem

The TSP is one of the most widely studied combinatorial optimization problems. It is also a problem to which the application of ACO algorithms is rather intuitive and straightforward. An instance of the TSP is determined by a set of locations (cities) and by the distances between them. The goal is to find a closed tour of minimal length that visits each city exactly once. A TSP instance

can be represented by a fully connected graph $G = (V, E, d)$, V being the set of $n = |V|$ vertices (representing the cities), E being the set of edges that fully connects the vertices, and d being a distance function that assigns to each edge (i, j) a distance d_{ij} . Here, we assume that the distance function is symmetric, that is, we have $d_{ij} = d_{ji}$, meaning that the distance is the same whether one goes from i to j or in the opposite direction.

To tackle a TSP instance with an ACO algorithm, each edge $(i, j) \in E$ needs a pheromone value τ_{ij} associated with it. The pheromone values are represented by real numbers that are modified while running the algorithm; they reflect the *learned* desirability of choosing an edge: the higher the pheromone value τ_{ij} , the higher is the desirability of choosing edge (i, j) as a solution component. Additionally, each edge has an associated heuristic value $\eta_{ij} = 1/d_{ij}$. For the TSP, the value η_{ij} is a measure of the *heuristic* desirability of having edge (i, j) as a component of a tour: the shorter the distance, the higher is the heuristic desirability.

An intuitive approach for constructing a tour is to first choose a vertex randomly and then, at each step, to go from the current vertex to the closest one that has not yet been visited. This solution construction ends when all vertices have been visited and the round trip is closed by returning to the initial vertex. In all ACO algorithms that have been implemented so far, the ants follow a randomized version of this construction rule. In fact, at each construction step, they choose randomly a next vertex based on the pheromone trail information and the heuristic information. The probabilistic choice is biased by pheromone and heuristic values: the higher the pheromone and the heuristic values associated with an edge, the higher the probability that an ant will choose it. Once all ants have completed their tours, the pheromone on the edges is updated. First, all pheromone values are decreased by a constant factor, simulating the phenomenon of pheromone evaporation. Then, each edge receives an amount of pheromone proportional to the quality of the solutions to which it belongs (there is one solution per ant); that is, the shorter the associated tour, the more

pheromone is deposited on the edges, making them more attractive in future iterations.

Example 2. Ant Colony Optimization for the Set Covering Problem

The SCP is a problem in which a candidate solution is represented by a subset of elements from some other set subject to some feasibility constraints. In the SCP, one is given two sets A and B . Each element B_i of B is a subset of A and it has associated a cost c_i . The goal of the SCP is to find a subset of the set B of minimal cost such that A is covered, that is, every element of the set A occurs in at least one of the elements chosen from set B . To guarantee that such a solution exists, one necessary assumption to make is that the elements of B cover the set A , that is, $\bigcup_{i=1}^n B_i = A$. A candidate solution for the SCP can be represented by an n -dimensional binary vector $X = [x_i]$, where n is the cardinality of set B , $x_i = 1$ if B_i is selected to be part of the solution, otherwise $x_i = 0$.

For solving an instance of the SCP with an ACO algorithm, we define the solution components as the elements of the set B . Each set B_i has associated a pheromone trail τ_i . The pheromone trails represent, analogous to the TSP, the ants' cumulated experience in solving the problem. For the SCP, the pheromone τ_i gives the desirability for an ant to choose element B_i , that is, to set the decision variable $x_i = 1$. The heuristic information η_i can be defined in various ways. One possibility is to use $\eta_i = k_i/c_i$, where $k_i = |B_i|$ is the total number of elements covered by the subset B_i . Hence, the heuristic function gives the average cost for B_i of covering elements of A . In this case, the heuristic information makes use only of *a priori* available information. It is therefore possible to compute the heuristic information before running the algorithm and therefore to compute the values of $\tau_i \cdot \eta_i$ before each algorithm iteration, saving in this way computation time. However, it may be advantageous to make the heuristic information more accurate (but slower to compute) by taking into account an ant's partial solution. In the SCP case, η_i could then measure the unit cost of covering one additional, still uncovered element of set A . This can be done by using $\eta_i = e_i/c_i$, where e_i is the number of

additional elements of set A covered when B_i is added to an ant's partial solution. Which of the two options—using the faster but less accurate precomputed heuristic information or adapting the heuristic information based on the ants' partial solutions—is preferable typically depends on the particular problem to which ACO is applied.

Ants construct solutions taking into account both the pheromone value and the heuristic information associated with each solution component. In the SCP case, an ant starts with an empty solution and chooses, at each construction step, one element of B until all elements of A are covered. In other words, an ant starts with all decision variables set to zero and at each construction step it sets one decision variable to one until all elements of A occur in at least one of the chosen elements of B . Note that in the SCP application, the number of construction steps to complete a solution may differ among the ants. Once each ant has terminated the construction of a candidate solution, it can remove subsets B_i that may have become redundant while constructing a solution before the pheromone trails are updated.

ACO METAHEURISTIC

ACO can be applied to any combinatorial optimization problem for which it is possible to devise an incremental solution construction procedure. Let us consider a general description of a combinatorial optimization problem that is modeled by the tuple (S, f, Ω) , where

- S is the set of candidate solutions defined over a finite set of discrete decision variables $\mathbb{X}$. S is referred to as the *search space* of the problem being tackled;
- $f: S \rightarrow \mathbb{R}$ is an objective function to be minimized;
- Ω is a (possibly empty) set of constraints among the decision variables.

A decision variable $X_i \in \mathbb{X}$, with $i = 1, \dots, n$, is said to be instantiated when a value v_i^j that belongs to its domain

```

procedure ACOMetaheuristic
  ScheduleActivities
    ConstructSolutions
    DaemonActions //optional
    UpdatePheromones
  end-ScheduleActivities
end-procedure

```

Figure 1. ACO metaheuristic in pseudocode. It works by intertwining three high-level procedures: ConstructSolutions, DaemonActions, and UpdatePheromones.

$D_i = \{v_i^1, \dots, v_i^{|D_i|}\}$ is assigned to it. A solution $s \in S$ is called *feasible* if each decision variable has been instantiated satisfying all constraints in the set Ω . Solving the optimization problem requires finding a solution s^* such that $f(s^*) \leq f(s) \forall s \in S$. Note that maximizing the value of an objective function f is the same as minimizing the value of $-f$; hence, every model of a combinatorial optimization problem can be described as a minimization problem.

ACO works by intertwining three high-level procedures: ConstructSolutions, DaemonActions, and UpdatePheromones as shown in Fig. 1. The ScheduleActivities construct does not specify how the three algorithmic components are scheduled and synchronized. However, in most applications, these procedures are executed in the depicted order.

- *ConstructSolutions.* This procedure implements the artificial ants' incremental construction of candidate solutions. In ACO, an instantiated decision variable $X_i \leftarrow v_i^j$ is called a *solution component* $c_{ij} \in C$, where C denotes the set of solution components. A pheromone trail value τ_{ij} is associated with each component $c_{ij} \in C$. (More formally, each solution component has an associated pheromone variable that can take a value, the pheromone trail value, in a specific range.) A solution construction starts from an initially empty partial solution s^p . At each construction step, s^p is extended by appending to it a feasible solution component from the set of its feasible neighbors $N(s^p) \subseteq C$ that satisfies the constraints in Ω . The

choice of a solution component is guided by a stochastic decision policy, which is biased by both the pheromone trail and the heuristic values associated with c_{ij} . The exact rules for the probabilistic choice of solution components vary across different variants of ACO. The best known rule is the one used first in the ant system algorithm [4]

$$p_{c_{ij}|s^p} = \frac{[\tau_{ij}]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{c_{il} \in N(s^p)} [\tau_{il}]^\alpha \cdot [\eta_{il}]^\beta}, \quad (1)$$

where τ_{ij} and η_{ij} are, respectively, the pheromone trail value and the heuristic value associated with the component c_{ij} . The parameters $\alpha > 0$ and $\beta > 0$ determine the relative importance of pheromone versus heuristic information.

- *DeamonActions*. This procedure, although optional, is important when state-of-the-art results are sought [7]. It allows the execution of problem-specific operations, such as the use of local search procedures, or of centralized actions that cannot be performed by artificial ants. It is usually executed before the update of pheromone values in order to bias the ants' search toward high quality solutions.
- *UpdatePheromones*. This procedure updates the pheromone trail values associated with the solution components in the set C . The modification of the pheromone trail values is performed in two stages: (i) *pheromone evaporation*, which decreases the pheromone values of all components by a constant factor ρ (called *evaporation rate*) in order to avoid premature convergence, and (ii) *pheromone deposit*, which increases the pheromone trail values associated with components of a set of promising solutions S_{upd} . The general form of the pheromone update rule is as follows:

$$\tau_{ij} \leftarrow (1 - \rho) \cdot \tau_{ij} + \rho \cdot \sum_{s \in S_{upd} | c_{ij} \in s} F(s), \quad (2)$$

where $\rho \in (0, 1]$ is the evaporation rate, and $F: S \rightarrow \mathbb{R}^+$ is a function such that $f(s) < f(s') \Rightarrow F(s) \geq F(s')$, $\forall s \neq s' \in S$. $F(\cdot)$ is called the *fitness function*. Different definitions for the set S_{upd} exist. Two common choices are $S_{upd} = s_{bsf}$, and $S_{upd} = s_{ib}$, where s_{bsf} is the best-so-far solution, that is, the best solution found since the start of the algorithm, and s_{ib} is the best solution of the current iteration. The specific implementation of the pheromone update mechanism differs across ACO variants [1,4,13–15].

When applying the ACO metaheuristic to a specific problem, the definition of solution components and, hence, the definition of the interpretation of the pheromone trails is decisive for the final performance of the ACO algorithm. In fact, even when restricting to problems where candidate solutions can be represented by a same representation (e.g., permutations), different interpretations for solution components and pheromone trails may be useful. For example, while in the TSP case (see Example 1 in the previous section), the successor relationship is important, that is, τ_{ij} should refer to the desirability of visiting city j directly after city i , in scheduling applications, it is often preferable to interpret a pheromone trail τ_{ij} as the desirability of assigning a job j to position i . When facing problems for which several alternative definitions of pheromone are reasonable, which one would be the best choice has to be determined experimentally.

ACO ALGORITHMS

The ACO metaheuristic is a general algorithmic framework. Various specific ACO algorithms, which all follow the high-level rules of the ACO metaheuristic, have been proposed in the literature. In fact, these variants are obtained by various instantiations of the three main procedures that build the ACO metaheuristic. Some of the most noteworthy variants are described below.

Ant System

Ant System (AS) was the first ACO algorithm reported in the literature [2–4]. In AS, the

pheromone values are updated at each iteration by all m ants of the colony. All pheromone trail values τ_{ij} are updated as follows:

$$\tau_{ij} \leftarrow (1 - \rho) \cdot \tau_{ij} + \sum_{k=1}^m \Delta\tau_{ij}^k, \quad (3)$$

where ρ is the evaporation rate, and $\Delta\tau_{ij}^k$ is the quantity of pheromone laid on c_{ij} by ant k . $\Delta\tau_{ij}^k$ is defined as follows:

$$\Delta\tau_{ij}^k = \begin{cases} F(s_k) & \text{if component } (i,j) \text{ is in the} \\ & \text{solution constructed by ant } k, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where the value of $F(s_k)$ is a function of the quality of the solution constructed by ant k . Normally, the better the solution, the higher is the amount of pheromone deposited.

In the solution construction, ants select solution components according to a stochastic mechanism, following Equation (1).

AS is mainly of historical interest because it was the first ACO algorithm proposed in the literature. Initial computational results have been interesting in the sense that they showed that the underlying mechanism works and allows to find good quality solutions. However, the performance of AS was still quite far from state-of-the-art methods. The main importance of AS is that it has seeded follow-up work by various researchers on better performing algorithmic variants, such as the two presented next.

$\mathcal{MAX}\text{-}\mathcal{MIN}$ Ant System

$\mathcal{MAX}\text{-}\mathcal{MIN}$ Ant System ($\mathcal{MMAS}$) [15] is an improvement over the original Ant System. Its main features are (i) only one of the best ants deposits pheromone, and (ii) the range of the allowed pheromone trail values is bounded. The pheromone update is implemented as follows:

$$\tau_{ij} \leftarrow \begin{cases} \tau_{\min} & \text{if } \tau_{ij} < \tau_{\min}, \\ (1 - \rho) \cdot \tau_{ij} + \Delta\tau_{ij}^{\text{best}} & \text{if } \tau_{\min} \leq \tau_{ij} \leq \tau_{\max}, \\ \tau_{\max} & \text{if } \tau_{ij} > \tau_{\max}, \end{cases} \quad (5)$$

where $\tau_{\max}$ and $\tau_{\min}$ are, respectively, the upper and lower bounds imposed on the pheromone, and $\Delta\tau_{ij}^{\text{best}}$ is defined as

$$\Delta\tau_{ij}^{\text{best}} = \begin{cases} F(s_{\text{best}}) & \text{if solution component } (i,j) \\ & \text{is part of } s_{\text{best}}, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where the value of $F(s_{\text{best}})$ is a function of the quality of the best solution found. This solution can be s_{ib} , s_{bsf} , a combination of them, or possibly some other high-quality solution.

Concerning the lower and upper bounds on the pheromone values, a bound on the maximum value may be calculated analytically as $\tau_{\max}^b = F(s^*)/\rho$, if the optimal solution s^* is known [16]. If s^* is not known, it can be approximated by s_{bsf} . Usually, setting $\tau_{\max} = \tau_{\max}^b$ (or to its approximation) results in good behavior of $\mathcal{MMAS}$. The initial value of the trails is set to $\tau_{\max}$ to increase the diversification of the search at the start of the algorithm. Some heuristic considerations for defining the setting of $\tau_{\min}$ have been proposed [15,17]. Finally, $\mathcal{MMAS}$ was the first ACO algorithm to use additional mechanisms for increasing the diversification of the search such as a reinitialization of the pheromone trails or a smoothing of the pheromone trail values when no improvement is observed for a given number of iterations. For a detailed description of these mechanisms, a general overview of $\mathcal{MMAS}$, and some variants of $\mathcal{MMAS}$, we refer the reader to [17].

Ant Colony System

Ant Colony System (ACS) [13] differs in some key aspects from other ACO algorithms. The first is that it uses a different decision rule in the ants' solution construction, which is known as the *pseudorandom proportional rule*. In this rule, with probability q_0 the next solution component j is the one that maximizes the product of the pheromone and heuristic values, that is, $j = \arg \max_{c_{il} \in N(s^p)} \{\tau_{il} \eta_{il}^\beta\}$. With probability $1 - q_0$ the probabilistic choice is made using Equation (1).

Similar to $\mathcal{MMAS}$, a pheromone update is applied at the end of each iteration by only one ant. The ACS pheromone update formula is as follows:

$$\tau_{ij} \leftarrow \begin{cases} (1 - \rho) \cdot \tau_{ij} + \rho \cdot \Delta\tau_{ij}^{\text{best}} & \text{if solution} \\ & \text{component } (i,j) \text{ is part of } s_{\text{best}}, \\ \tau_{ij} & \text{otherwise.} \end{cases} \quad (7)$$

As in $\mathcal{MMAS}$, $\Delta\tau_{ij}^{\text{best}} = F(s_{\text{best}})$, where s_{best} can be either s_{ib} or s_{bsf} . It is noteworthy that in ACS only the pheromone values of solution components associated with the best solution are updated.

To avoid search stagnation, in ACS a local pheromone update is performed by each ant after each construction step. This update decreases the pheromone trail value of the solution component that has been chosen in the previous step. The goal is to diversify the search performed by subsequent ants during an iteration: by decreasing the pheromone concentration for chosen components, these get less desirable for subsequent ants, thus increasing the chances of producing different solutions. Each ant applies the local pheromone update only to the pheromone trail of the last solution component added

$$\tau_{ij} \leftarrow (1 - \varphi) \cdot \tau_{ij} + \varphi \cdot \tau_0, \quad (8)$$

where $\varphi \in (0, 1)$ is a parameter called *pheromone decay coefficient*, and τ_0 is a parameter that determines the initial value for the pheromone trails. A good value for τ_0 was found to be $F(s^h)/n$, where n is the size of the instance and s^h is a solution constructed using a problem-specific heuristic [7].

Other Variants

In addition to the variants described above, there are others that have been reported in the literature. Table 1 summarizes the main ACO variants, including those discussed in the previous sections, which have been proposed in the literature for the approximate solution of NP-hard problems. For each of these variants, we give the main references and the year in which they were proposed.

The main characteristics of the ACO algorithms that were not discussed so far are the following. Elitist AS is a direct variant of AS that gives a strong additional feedback to the best solution constructed since the start of the algorithm. Ant-Q is a predecessor of ACS that is inspired by the well-known Q-learning method from reinforcement learning. Rank-based AS extends Elitist AS by allowing not only the best-so-far ant, but also the r best ranked ants of the current iteration to deposit pheromone; the weight given to each ant in the pheromone update is inversely proportional to its rank, the highest weight being given to the best-so-far ant. ANTS is an ACO algorithm that exploits the connection with tree-search procedures by including elements from branch-and-bound techniques, such as lower bound information, into an ACO algorithm. Best-worst AS is an AS variant where the worst ant of the current iteration is used to subtract pheromone from solution components that are part of this worst ant but that do not occur in the best-so-far solution. Population-based ACO uses a set of elite solutions to define, at each iteration, the pheromone trail matrix. The set of elite solutions is managed by a population management mechanism that is responsible for updating the pheromone matrix each time a solution is added or removed from the elite set. Finally, Beam-ACO incorporates a heuristic derived from branch-and-bound algorithms called *beam search*.

Table 1. Overview of the main ACO algorithms for NP-hard problems that have been proposed in the literature

ACO Algorithm	Main References	Year
Ant System (AS)	[2–4]	1991
Elitist AS	[2–4]	1992
Ant-Q	[18]	1995
Ant Colony System	[13,14]	1996
$\mathcal{MAX-MIN}$ AS	[15,19,20]	1996
Rank-based AS	[21,22]	1997
ANTS	[23,24]	1998
Best-worst AS	[25,26]	2000
Population-based ACO	[27]	2002
Beam-ACO	[28,29]	2004

Given are the ACO algorithm names, the main references where these algorithms are described, and the year in which they were first published.

Hybrid ACO Algorithms

Currently, it is a well-established fact that ACO algorithms reach best performance for most combinatorial optimization problems when they are combined with either iterative improvement algorithms or more complex local search methods such as tabu search or simulated annealing [7]. In these hybrid algorithms, the local improvement methods are used to improve the solutions constructed by one or more ants after each iteration. The usage of local search algorithms is also one example of the daemon actions that have been mentioned in the description of the ACO metaheuristic (see Fig. 1).

Various alternative ways of hybridizing ACO algorithms with other techniques have been studied. The ANTS and Beam-ACO algorithms, mentioned in the previous section, were explicitly designed as hybrid algorithms that integrate features from branch-and-bound techniques into ACO algorithms. Another active area is the integration of constraint programming techniques into ACO algorithms [30,31], which is particularly attractive for problems where, due to the problem constraints, it is difficult for the ants to generate feasible candidate solutions. Other hybrid techniques exploit the idea of using partial candidate solutions to seed an ant's solution construction. Examples of these hybrid methods are the use of external memory in ACO algorithms [32] or the extensions called *iterated ants* [33] and *cunning ants* [34].

The investigation of hybrid ACO algorithms is currently one of the most active areas in the research on ACO.

ACO APPLICATIONS

ACO algorithms have been successfully applied to a large variety of important problems from both the academic and industrial worlds (see the article ***A Concise Overview of Applications of Ant Colony Optimization*** for more information). The main application areas are the following:

NP-Hard Problems. The best known algorithms that are guaranteed to

find an optimal solution to this kind of problems have exponential time complexity in the worst case [35]. However, heuristic methods such as ACO can be used to find high-quality solutions in a reasonable amount of time. Some examples of NP-hard problems for which ACO algorithms have been successful are *routing problems* [36–38], in which the goal is to find the shortest route that visits a set of locations; *assignment problems* [15], where a set of items (objects, activities, etc.) has to be assigned to a given number of resources (locations, agents, etc.) subject to some constraints; *subset problems* [39], where a solution to a problem is considered to be a selection of a subset of available items; and *scheduling problems* [29], in which the main concern is to optimally allocate scarce resources to tasks over time.

Rich Academic and Industrial Problems.

After initial encouraging results on classic academic problems, ACO started to be applied to real industrial problems such as those arising in the food or in manufacturing industry [38,40]. As a result, richer versions of the academic problems started to be studied. Among the features of these problems are time-varying data [41], stochasticity [42,43], the presence of multiple objectives [44,45], continuous variables [46], mixed variables [47], and so on. Practically relevant dynamic problems are those found in the domain of telecommunication networks because some important properties, such as the cost of using links or the availability of nodes, vary over time. Some ACO algorithms have been shown to be very effective at solving these types of problems [48–50].

ACO THEORY

Most of the research results on metaheuristics, in general, and on ACO, in particular, are of experimental nature. However,

there is also a significant interest in more fundamental properties of ACO algorithms.

A first question that is usually asked is whether, given enough time, the algorithm will eventually find an optimal solution. An initial answer to this question was given by Gutjahr, who proved for an ACO algorithm called *graph based ant system* (GBAS) the convergence with probability $1 - \epsilon$ to the optimal solution [51]. In a later paper [52], convergence with probability 1 was proven for two variants of GBAS. While GBAS has not been studied in practical applications, it is remarkable that convergence proofs for two of the practically most successful ACO algorithms, ACS and $\mathcal{MMAS}$, have also been obtained [53]. More recently, the focus of research has shifted to studies of the expected runtime to find optimal solutions in ACO applications to specific problems. An overview of proof techniques and some results are given by Gutjahr [54]; recent publications in this direction can be found in [55–57].

Other contributions on theoretical aspects of ACO have focused on establishing connections to other methods. Zlochin *et al.* [58] have defined the framework of model-based search algorithms, of which ACO is one representative. Connections of ACO to stochastic gradient descent, an algorithm used, for example, for learning weights in neural networks, have been studied in [59]. Of more practical interest are studies on the behavior of ACO algorithms. Merkle and Middendorf were the first to analyze the dynamic behavior of the pheromone model in ACO algorithms [60]. Search bias in ACO algorithms is studied in [61], where the authors show that ACO algorithms may suffer from the same type of deception as evolutionary algorithms do. In addition, they show that ACO algorithms may suffer from what they call a second order deceptive behavior, where, due to an interaction between the pheromone update and the pheromone model chosen, the quality of the solutions generated by an ACO algorithm can decrease over time.

A more detailed discussion of theoretical results about ACO algorithms is given in [7,62].

CONCLUSIONS

ACO is now one of the main metaheuristics and an active area of research. Early research on ACO focused mainly on the development of effective ACO algorithm variants and a common framework for these algorithmic developments is given by the ACO metaheuristic. Currently, the main active research directions in ACO concern applications to computationally challenging problems, the hybridization of ACO algorithms with other search techniques, and the theoretical study of the behavior of specific ACO algorithms.

Evidence of the success of ACO algorithms is the number of specialized meetings, where researchers can discuss their research results on ACO algorithms and their applications. ACO is one of the main subjects of the biannual conference ANTS (International Conference on Swarm Intelligence; <http://iridia.ulb.ac.be/ants/>) and of the IEEE Swarm Intelligence Symposium series. In addition, ACO is a central topic at various conferences on metaheuristics and evolutionary algorithms. Finally, research on ACO has frequently been featured in journal special issues [63–66] and is a fundamental subject of the journal *Swarm Intelligence*. Information on ACO and related topics can be obtained through the moderated mailing list `aco-list`, and the ACO web page (www.aco-metaheuristic.org).

Acknowledgments

This work was supported by the META-X project, an *Action de Recherche Concertée* funded by the Scientific Research Directorate of the French Community of Belgium and the E-SWARM ERC Advanced Grant. Marco Dorigo and Thomas Stützle acknowledge support from the Belgian F.R.S.-FNRS, of which they are a Research Director and a Research Associate, respectively.

REFERENCES

1. Dorigo M, Maniezzo V, Colorni A. Positive feedback as a search strategy. Italy: Dipartimento di Elettronica, Politecnico di Milano; 1991. pp. 91–016.

2. Dorigo M, Maniezzo V, Colorni A. The ant system: an autocatalytic optimizing process. Italy: Dipartimento di Elettronica, Politecnico di Milano; 1991. pp. 91–016. Revised.
3. Dorigo M. Optimization, learning and natural algorithms (in Italian). Italy: Dipartimento di Elettronica, Politecnico de Milano; 1992.
4. Dorigo M, Maniezzo V, Colorni A. Ant System: optimization by a colony of cooperating agents. *IEEE Trans Syst Man and Cybern-Part B* 1996;26(1):29–41.
5. Dorigo M, Di Caro G. The ant colony optimization meta-heuristic. In: Corne D, editor. *New ideas in optimization*. London: McGraw Hill; 1999. pp. 11–32.
6. Bonabeau E, Dorigo M, Theraulaz G. Inspiration for optimization from social insect behaviour. *Nature* 2000;406(6791):39–42.
7. Dorigo M, Stützle T. *Ant Colony Optimization*. Cambridge (MA): MIT Press; 2004.
8. Dorigo M. Ant colony optimization. *Scholarpedia* 2007;2(3):1461.
9. Grassé PP. La reconstruction du nid et les coordinations interindividuelles chez *Bellicositermes natalensis* et *Cubitermes sp.* La théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs. *Insectes Sociaux* 1959;6:41–81.
10. Pasteels JM, Deneubourg JL, Goss S. Self-organization mechanisms in ant societies (I): trail recruitment to newly discovered food sources. *Experientia Suppl* 1987; 54:155–175.
11. Goss S, Aron S, Deneubourg JL, *et al.* Self-organized shortcuts in the Argentine ant. *Naturwissenschaften* 1989;76:579–581.
12. Deneubourg JL, Aron S, Goss S, *et al.* The self-organizing exploratory pattern of the Argentine ant. *J Insect Behav* 1990;3(2):159–168.
13. Dorigo M, Gambardella LM. Ant Colony System: a cooperative learning approach to the traveling salesman problem. *IEEE Trans Evol Comput* 1997;1(1):53–66.
14. Gambardella LM, Dorigo M. Solving symmetric and asymmetric TSPs by ant colonies. In: *ICEC'96 Proceedings of the IEEE Conference on Evolutionary Computation*. Piscataway (NJ): IEEE Press; 1996. pp. 622–627.
15. Stützle T, Hoos HH. *MAX-MIN* ant system. *Future Gener Comput Syst* 2000;16(8): 889–914.
16. Stützle T, Dorigo M. A short convergence proof for a class of ant colony optimization algorithms. *IEEE Trans Evol Comput* 2002; 6(4):358–365.
17. Stützle T. Volume 220, *Local search algorithms for combinatorial problems: analysis, improvements, and new applications*, DISKI. Germany: Infix, Sankt Augustin; 1999.
18. Gambardella LM, Dorigo M. Ant-Q: a reinforcement learning approach to the traveling salesman problem. In: Frieditis A, Russell S, editors. *Proceedings of the 12th International Conference on Machine Learning (ML-95)*. Palo Alto (CA): Morgan Kaufmann Publishers; 1995. pp. 252–260.
19. Stützle T, Hoos HH. Improving the Ant System: a detailed report on the *MAX-MIN* Ant System. Germany: FG Intellektik, FB Informatik, TU Darmstadt; 1996. AIDA-96-12.
20. Stützle T, Hoos HH. The *MAX-MIN* Ant System and local search for the traveling salesman problem. In: Bäck T, Michalewicz Z, Yao X, editors. *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*. Piscataway (NJ): IEEE Press; 1997. pp. 309–314.
21. Bullnheimer B, Hartl RF, Strauss C. A new rank based version of the Ant System—a computational study. Vienna: Institute of Management Science, University of Vienna; 1997.
22. Bullnheimer B, Hartl RF, Strauss C. A new rank-based version of the Ant System: a computational study. *Cent Eur J Oper Res Econ* 1999;7(1):25–38.
23. Maniezzo V. Exact and approximate non-deterministic tree-search procedures for the quadratic assignment problem. Italy: Scienze dell'Informazione, Università di Bologna, Sede di Cesena; 1998. CSR 98-1.
24. Maniezzo V. Exact and approximate non-deterministic tree-search procedures for the quadratic assignment problem. *INFORMS J Comput* 1999;11(4):358–369.
25. Cordon O, de Viana IF, Herrera F. Analysis of the best-worst Ant System and its variants on the TSP. *Mathware Soft Comput* 2002;9(2–3):177–192.
26. Cordon O, de Viana IF, Herrera F, *et al.* A new ACO model integrating evolutionary computation concepts: the best-worst Ant System. In: Dorigo M, Middendorf M, Stützle T, editors. *Abstract proceedings of ANTS 2000-From Ant Colonies to Artificial Ants: Second International Workshop on Ant Algorithms*. Brussels: IRIDIA, Université Libre de Bruxelles; 2000. pp. 22–29.

27. Guntsch M, Middendorf M. A population based approach for ACO. In: Cagnoni S, *et al.*, editors. Volume 2279, Applications of Evolutionary Computing, Proceedings of EvoWorkshops2002, LNCS. Berlin: Springer; 2002. pp. 71–80.
28. Blum C. Theoretical and practical aspects of ant colony optimization. Brussels: IRIDIA, Université Libre de Bruxelles; 2004.
29. Blum C. Beam-ACO—hybridizing ant colony optimization with beam search: an application to open shop scheduling. *Comput Oper Res* 2005;32(6):1565–1591.
30. Meyer B, Ernst A. Integrating ACO and constraint propagation. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence, 4th International Workshop, ANTS 2004, LNCS. Berlin: Springer; 2004. pp. 166–177.
31. Khichane M, Albert P, Solnon C. Integration of ACO in a constraint programming language. In: Dorigo M, *et al.*, editors. Volume 5217, Ant Colony Optimization and Swarm Intelligence, 6th International Conference, ANTS 2008, LNCS. Berlin: Springer; 2008. pp. 84–95.
32. Acan A. An external memory implementation in ant colony optimization. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, LNCS. Berlin: Springer; 2004. pp. 73–84.
33. Wiesemann W, Stützle T. Iterated ants: an experimental study for the quadratic assignment problem. In: Dorigo M, *et al.*, editors. Volume 4150, Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, LNCS. Berlin: Springer; 2006. pp. 179–190.
34. Tsutsui S. cAS: ant colony optimization with cunning ants. In: Runarsson TP, *et al.*, editors. Volume 4193, Parallel Problem Solving from Nature-PPSN IX, 9th International Conference, LNCS. Berlin: Springer; 2006. pp. 162–171.
35. Garey MR, Johnson DS. Computers and Intractability: a Guide to the Theory of NP-Completeness. New York: W.H. Freeman & Co.; 1979.
36. Gambardella LM, Dorigo M. Ant Colony System hybridized with a new local search for the sequential ordering problem. *INFORMS J Comput* 2000;12(3):237–255.
37. Reimann M, Doerner K, Hartl RF. D-Ants: savings based ants divide and conquer the vehicle routing problem. *Comput Oper Res* 2004;31(4):563–591.
38. Rizzoli AE, Montemanni R, Lucibello E, Gambardella LM. Ant colony optimization for real-world vehicle routing problems: from theory to applications. *Swarm Intell* 2007; 1(2):135–151.
39. Blum C, Blesa MJ. New metaheuristic approaches for the edge-weighted k-cardinality tree problem. *Comput Oper Res* 2005; 32(6):1355–1377.
40. Dorigo M, Birattari M, Stützle T. Ant colony optimization: artificial ants as a computational intelligence technique. *IEEE Comput Intell Mag* 2006;1(4):28–39.
41. Montemanni R, Gambardella LM, Rizzoli AE, *et al.* Ant Colony System for a dynamic vehicle routing problem. *J Comb Optim* 2005; 10(4):327–343.
42. Bianchi L, Gambardella LM, Dorigo M. An ant colony optimization approach to the probabilistic traveling salesman problem. In: Merelo JJ, *et al.*, editors. Volume 2439, Parallel Problem Solving from Nature-PPSN VII, 7th International Conference, LNCS. Berlin: Springer; 2002. pp. 883–892.
43. Balaprakash P, Birattari M, Stützle T, *et al.* Estimation-based ant colony optimization algorithms for the probabilistic traveling salesman problem. *Sist Intell* 2009;3(3): 223–242.
44. Doerner K, Gutjahr WJ, Hartl RF, *et al.* Pareto ant colony optimization: a metaheuristic approach to multiobjective portfolio selection. *Ann Oper Res* 2004;131(1–4):79–99.
45. Angus D, Woodward C. Multiple objective ant colony optimisation. *Swarm Intell* 2007; 3(1):69–85.
46. Socha K, Dorigo M. Ant colony optimization for continuous domains. *Eur J Oper Res* 2008;185(3):1155–1173.
47. Socha K. ACO for continuous and mixed-variable optimization. In: Dorigo M, *et al.*, editors. Volume 3172, Ant Colony Optimization and Swarm Intelligence: 4th International Workshop, ANTS 2004, LNCS. Berlin: Springer; 2004. pp. 25–36.
48. Di Caro G, Dorigo M. AntNet: distributed stigmergetic control for communications networks. *J Artif Intell Res* 1998;9:317–365.
49. Mong Sim K, Hong Sun W. Ant colony optimization for routing and load-balancing: survey and new directions. *IEEE Trans Syst Man Cybern Part A: Syst Humans* 2003; 33(5):560–572.

50. Di Caro G, Ducatelle F, Gambardella LM. AntHocNet: an adaptive nature-inspired algorithm for routing in mobile ad hoc networks. *Eur Trans Telecommun* 2005;16(5):443–455.
51. Gutjahr WJ. A Graph-based Ant System and its convergence. *Future Gener Comput Syst* 2000;16(8):873–888.
52. Gutjahr WJ. ACO algorithms with guaranteed convergence to the optimal solution. *Inf Process Lett* 2002;82(3):145–153.
53. Stützle T, Dorigo M. A short convergence proof for a class of ACO algorithms. *IEEE Trans Evol Comput* 2002;6(4):358–365.
54. Gutjahr WJ. Mathematical runtime analysis of ACO algorithms: survey on an emerging issue. *Swarm Intell* 2007;1(1):59–79.
55. Gutjahr WJ, Sebastiani G. Runtime analysis of ant colony optimization with best-so-far reinforcement. *Methodol Comput Appl Probab* 2008;10:409–433.
56. Neumann F, Sudholt D, Witt C. Analysis of different MMAS ACO algorithms on unimodal functions and plateaus. *Swarm Intell* 2009;3(1):35–68.
57. Neumann F, Witt C. Runtime analysis of a simple ant colony optimization algorithm. *Algorithmica* 2009;54(2):243–255.
58. Zlochin M, Birattari M, Meuleau N, *et al.* Model-based search for combinatorial optimization: a critical survey. *Ann Oper Res* 2004;131(1–4):373–395.
59. Meuleau N, Dorigo M. Ant colony optimization and stochastic gradient descent. *Artif Life* 2002;8(2):103–121.
60. Merkle D, Middendorf M. Modeling the dynamics of ant colony optimization. *Evol Comput* 2002;10(3):235–262.
61. Blum C, Dorigo M. Search bias in ant colony optimization: On the role of competition-balanced systems. *IEEE Trans Evol Comput* 2005;9(2):159–174.
62. Dorigo M, Blum C. Ant colony optimization theory: a survey. *Theor Comput Sci* 2005;344(2–3):243–278.
63. Cordon O, Herrera F, Stützle T. Special issue on Ant Colony Optimization: Models and applications. *Mathware Soft Comput* 2003;9(2–3):137–268.
64. Doerner KF, Merkle D, Stützle T. Special issue on ant colony optimization. *Swarm Intell* 2009;3(1):1–85.
65. Dorigo M, Di Caro G, Stützle T. Special issue on “Ant Algorithms”. *Future Gener Comput Syst* 2000;16(8):851–946.
66. Dorigo M, Gambardella LM, Middendorf M, *et al.* Special issue on “Ant Algorithms and Swarm Intelligence”. *IEEE Trans Evol Comput* 2002;6(4):317–365.

ANTITHETIC VARIATES

DONGHAI HE

CHUN-HUNG CHEN

Department of Systems Engineering and
Operations Research, George Mason
University, Fairfax, Virginia

One of the most recommended variance reduction techniques for Monte Carlo simulation is antithetic variates, first introduced by Hammersley and Morton [1]. It was motivated by the desire to achieve a smaller variance than simple Monte Carlo random sampling could obtain from simulation samples of the same size. The basic idea is to induce negative correlation into random samples. This survey article gives an overview of antithetic variates and discusses their recent development. Numerical results are presented to illustrate their effectiveness.

Originally, the antithetic variates method was developed to allow for the evaluation of integrals by Monte Carlo sampling methods with better accuracy (smaller variance) than one could achieve for a given sample size with pure random sampling, when the integral function is bounded. Wilson [2] shows that unbounded functions are also suitable for antithetic variates method. Instead of focusing on integrals, Page [3] extends antithetic variates to estimate the statistics of queueing system simulation. Furthermore, Mitchell [4] applies antithetic variates to $GI/G/1$ queueing models and extends certain results obtained by Page [3]. George [5] uses antithetic variates to reduce the variance of prediction of replacement requirements in a multiple-component multiple-period replacement process. Nelson [6] applies antithetic variates to obtain better estimates of steady-state simulations. Avramidis and Wilson [7] extend antithetic variates to estimate quantiles instead of expected values. Further, antithetic variates have been successfully applied to Monte Carlo radiation transport problems [8]

and particle transport problems [9]. Cheng [10], L'Ecuyer [11], and Law and Kelton [12] present a nice overview of antithetic variates. More discussions on antithetic variates can be found in Refs 13–21.

The rest of this article is organized as follows: we first present fundamental statistics and then explain the basic ideas of antithetic variates. We show how they work to reduce variance in Monte Carlo simulation. Limitations of current antithetic variates methods are then discussed, and new developments presented. Finally, we conduct two numerical experiments to demonstrate the effectiveness of antithetic variates. The numerical results illustrate what can be expected from this variance reduction method.

FUNDAMENTAL AND BASIC IDEAS

The basic idea of antithetic variates is to construct pairs of simulation samples such that a small observation in one of the samples tends to be offset by a bigger observation in the other. As a result, the average of these two observations will tend to be closer to the mean value we are trying to estimate. To achieve that, the observations must be negatively correlated. We start with a simple case.

Suppose X_1 and X_2 are two simulation samples. The mean is estimated using a sample mean estimator

$$\bar{X} = \frac{1}{2}(X_1 + X_2).$$

When X_1 and X_2 are independent, the variance of the sample mean estimator is

$$\text{Var}(\bar{X}) = \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2)].$$

If we introduce negative correlation between these two samples, then

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)] \\ &< \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2)]. \end{aligned} \quad (1)$$

The higher the negative correlation between X_1 and X_2 , the more we can reduce the variance of the sample mean estimator.

Now consider a general case. Suppose

$$X = f(W) \quad (2)$$

is an unbiased estimator of interest, where $f(\cdot)$ is either a bounded or an unbounded function with finite variance [2], and W is the random input. The goal is to estimate the expected value of X . In the setting of Monte Carlo simulation, we take n samples of W and so obtain n samples of X . Then the expected value of X can be estimated as

$$\bar{X}(n) = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n f(W_i), \quad (3)$$

where W_i is the i th random input, X_i is the i th response (or observation), $\bar{X}(n)$ denotes the sample mean and n is the sample size. The variance of $\bar{X}(n)$ is

$$\begin{aligned} \text{Var}(\bar{X}(n)) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n X_i \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j) \right). \end{aligned} \quad (4)$$

If X_i and X_j , for $i \neq j$, $i, j \in \{1, \dots, n\}$, are negatively correlated, the variance of the sample mean estimator is reduced. The question is how to induce negative correlation among the samples.

INDUCING NEGATIVE CORRELATION

In Monte Carlo simulation, random variates are normally generated using uniform random numbers. Hammersley and Morton [1] induce negative correlation in these random numbers while ensuring that the estimator of interest remains unbiased. In the simplest form of antithetic variates, the negative correlation is induced by using *complementary* random numbers. Define

$$W_i^{(1)} = U_i$$

$$\text{and } W_i^{(2)} = 1 - U_i,$$

where $W_i^{(1)}$ and $W_i^{(2)}$ constitute the i th pair of antithetic variates, and the random number U_i is uniformly distributed between 0 and 1. Note that

$$\text{Cov}(W_i^{(1)}, W_i^{(2)}) = -\text{Var}(U_i) < 0.$$

By introducing antithetic variates, Equation (3) can be rewritten as

$$\bar{X}(n) = \frac{1}{n} \sum_{i=1}^{n/2} [f(W_i^{(1)}) + f(W_i^{(2)})]. \quad (5)$$

Thus, the variance of $\bar{X}(n)$ in Equation (4) can be rewritten as

$$\begin{aligned} \text{Var}(\bar{X}(n)) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^{n/2} [f(W_i^{(1)}) + f(W_i^{(2)})] \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^{n/2} \text{Var}(f(W_i^{(1)})) \right. \\ &\quad \left. + \sum_{i=1}^{n/2} \text{Var}(f(W_i^{(2)})) \right. \\ &\quad \left. + 2 \sum_{i=1}^{n/2} \text{Cov}(f(W_i^{(1)}), f(W_i^{(2)})) \right). \end{aligned} \quad (6)$$

When $f(W_i^{(1)})$ and $f(W_i^{(2)})$ are simulated independently, the covariance terms are zero. If we induce negative correlation between them, $\text{Cov}(f(W_i^{(1)}), f(W_i^{(2)})) < 0$ and the variance of $\bar{X}(n)$ is reduced.

Note that the simulation output responses, $f(W_i^{(1)})$ and $f(W_i^{(2)})$, are not always negatively correlated, even when the random inputs, $W_i^{(1)}$ and $W_i^{(2)}$, are negatively correlated. Hammersley and Morton [1] show that if the response function $f(\cdot)$ is a monotone, increasing or decreasing, then $\text{Cov}(f(W_i^{(1)}), f(W_i^{(2)})) \leq 0$. On the other hand, Cheng [22,23] demonstrates that $\text{Cov}(f(W_i^{(1)}), f(W_i^{(2)}))$ might be positive if the response function is not a monotone, in which case the variance increases.

Nonuniform Random Input

In most cases, the random inputs are not uniformly distributed. For example, the interarrival and service times in an $M/M/1$ queueing model follow exponential distributions. To ensure negative correlation, we generate random inputs using the inverse transform method. Thus,

$$W^{(1)} = g(U) \\ \text{and } W^{(2)} = g(1 - U),$$

where $g(\cdot)$ is an inverse cumulative distribution function (ICDF). The correlation coefficient between U and $1 - U$ is

$$\begin{aligned} \rho_{U,1-U} &= \frac{\text{Cov}(U, 1 - U)}{\sqrt{\text{Var}(U) \cdot \text{Var}(1 - U)}} \\ &= \frac{-\text{Var}(U)}{\text{Var}(U)} = -1. \end{aligned}$$

However, the correlation coefficient between $W^{(1)}$ and $W^{(2)}$, $\rho_{W^{(1)}, W^{(2)}}$, may be much greater than -1 due to the transformation of $g(\cdot)$.

Page [3] and Franta [24] show that if $W^{(1)}$ and $W^{(2)}$ are a pair of normally distributed antithetic variates generated by the Box Muller method (polar method), $\rho_{W^{(1)}, W^{(2)}} \approx -0.582$. If they are generated by the Marsaglia method, $\rho_{W^{(1)}, W^{(2)}} = -1$. If $W^{(1)}$ and $W^{(2)}$ are a pair of exponentially distributed (with parameter 1) antithetic variates, in which case the ICDF $g(u) = -\log(u)$, then $\rho_{W^{(1)}, W^{(2)}} = -0.645$.

In addition to the inverse transform method, random variates can be generated using other approaches. Schmeiser and Kachitvichyanukul [25] develop fast approaches to generate correlated random inputs using composition and acceptance/rejection methods. However, these approaches produce less correlation than that obtained by the inverse transform method.

Extensions

Several extensions and variations of the antithetic variates method have been developed.

Multivariate Inputs. Wilson [26] extends antithetic variates to the case of two or more

random variables. Consider an unbiased estimator which is a summation of m responses of random inputs:

$$\sum_{j=1}^m f_j(W_j). \quad (7)$$

Instead of using complementary random numbers between samples (e.g., U and $1 - U$), a key step of this extension consists of inducing appropriate dependencies among these input random variables while preserving their distributions. Intuitively, this estimator should be composed of mutually counteracting random variables so that the sum has a small net variance. Wilson [26] also offers a general framework, with antithetic variate theorems, in attempting to minimize the variance of the estimator given in Equation (7).

Fishman and Huang [27] extend antithetic variates to multivariate inputs in an estimator similar to Equation (7). Like Wilson [26], instead of using U and $1 - U$ between simulation samples, they intend to induce the desired negative correlation among these m responses in a sample using a rotation sampling scheme. One example of rotation sampling is

$$W_j(U) = \begin{cases} U + \theta_j & 0 \leq U < 1 - \theta_j \\ U + \theta_j - 1 & 1 - \theta_j \leq U < 1 \end{cases}, \quad (8)$$

for $0 \leq \theta_j \leq 1, j = 1, \dots, m$. A theoretical foundation for minimizing variance is provided.

Queueing Simulation. In addition to Monte Carlo simulation, the application of antithetic variates has been extended to more complex system simulations, such as discrete-event simulation. As we discuss below, a major challenge in these extensions is how to control the generated random variates in a way that ensures negative correlation.

Page [3] applies antithetic variates to $M/M/1$ queueing simulation. Like Monte Carlo simulation, the basic idea is to induce negative correlation by using complementary random numbers in two consecutive runs of the simulation. Specifically, if U_i is a particular random number used for a

particular purpose (e.g., to generate the i th interarrival time or i th service time) in the first run, we use $1 - U_i$ for this same purpose in the second run. By doing so, negative correlation between the two runs is obtained and the variance of the estimator is reduced. In addition, Mitchell [4] uses antithetic variates to reduce the variance of estimates of interests in a $GI/G/1$ queueing simulation, such as waiting time and queue length. Tuffin [28] presents the application of antithetic variates to closed product-form multiclass Jackson networks and achieves improvement in simulation analysis.

Nelson [29] investigates the consequence of incorporating antithetic variates into steady-state simulation when initial-condition bias is present. He shows that antithetic variates can lead to improvement in all standard criteria under typical assumptions.

Cheng [23] gives two examples to demonstrate the use of antithetic variates for transient simulation. In the example of gas demand, the input contains an autocorrelated error sequence and the target is to calculate the average peak demand. Another example deals with average peak flow in water resource management. Both examples have a time series output. The numerical results show the significant benefit of antithetic variates in variance reduction.

Other Strategies. Instead of estimating a mean of interest, Avramidis and Wilson [7] extend antithetic variates to estimate quantile of a large-scale finite-horizon stochastic simulation. This work is based on their previous attempt [30], which integrates antithetic variates and Latin hypercube sampling into finite-horizon stochastic simulation to estimate a mean response. They provide a framework for order statistics based on quantile estimation. Antithetic variates and Latin hypercube sampling are introduced into the general framework to induce correlation. The variance reduction improvement is observed in both, single-sample quantile estimators and multiple-sample quantile estimators.

Yang and Liou [31] combine antithetic variates with another popular variance reduction technique, “control variates.” They

apply antithetic variates to generate control variates across paired replications and show that the integrated control-variate estimator is unbiased and yields a smaller variance than the conventional control-variate estimator without using antithetic variates.

LIMITATIONS

In practice, antithetic variates have two major limitations.

Efficacy is Sensitive to Problems

A basic requirement for antithetic variates is monotone response functions. This requirement can generally be satisfied in simple simulations, particularly when the inverse transform method is applied. However, it may become difficult to ensure that the response is monotone when the simulated system is complex. Cheng [22,23] shows that the method can backfire when non-monotone functions are used.

In addition, the efficiency of antithetic variates depends on how much negative correlation we can induce in the sampling. That is usually unknown beforehand. Franta [24] shows that negative correlation is highly dependent on the response functions. Cheng [22] demonstrates that a skewed distribution, such as the exponential distribution, may significantly reduce the efficacy of negative correlation.

Synchronization

When applying antithetic variates to discrete-event simulation, we should carefully synchronize the random number sequences between the two replications (runs) in a pair, in order to ensure that the outputs are negatively correlated. As presented earlier, if U_i is a particular random number used to generate an interarrival time in the first replication, $1 - U_i$ should be used to generate the interarrival time for the same customer in the second replication. The erroneous use of U and $1 - U$ can undermine the basic intent of antithetic variates, which attempt to complement a small response of U with a large response of $1 - U$, and

Table 1. Variances of the Estimators With and Without Antithetic Variates

	$f(x) = x^4$	$f(x) = \sin(0.5\pi x)$	$f(x) = \frac{e^x - 1}{e - 1}$
Without AV	7.0×10^{-4}	9.5×10^{-4}	8.3×10^{-4}
With AV	3.3×10^{-4}	7.8×10^{-5}	2.6×10^{-5}

vice versa. For simple queueing simulation, synchronization of random number sequences is not too difficult. However, it can be quite challenging when the simulated systems are complicated. Law and Kelton [12] present extensive discussions on this synchronization issue and offer several useful suggestions.

NUMERICAL EXAMPLES

We apply antithetic variates to two numerical experiments to demonstrate their effectiveness. The numerical results illustrate what can be expected from this variance reduction method.

Monte Carlo Integration

The first experiment is to evaluate a standard one-dimensional integral using Monte Carlo simulation. We want to evaluate

$$I = \int_0^1 f(x) dx.$$

Note that $I = E_U[f(x)]$, where U is uniformly distributed between 0 and 1. In Monte Carlo simulation, I is estimated using a sample mean estimator

$$\frac{1}{n} \sum_{i=1}^n f(U_i).$$

In this experiment, three functions are tested.

- (1) $f(x) = x^4$,
- (2) $f(x) = \sin(0.5\pi x)$,
- and (3) $f(x) = \frac{e^x - 1}{e - 1}$.

All of the above functions are monotone over the range we are considering (i.e., $[0, 1]$). In our numerical testing, we set $n = 100$. To estimate the variance of the estimator, we repeat 10,000 independent macroreplications.

Table 1 gives the estimated variances and shows that the variance is reduced when antithetic variates are applied.

M/G/1 Queueing Simulation

The second experiment is a single server M/G/1 queueing system, which has been discussed in many queueing theory textbooks. The interarrival time is exponentially distributed with mean 0.5 and the service time is uniformly distributed between 0 and 1. We are interested in the average waiting time in the queue and the average total time in the system for the first 100 customers.

Using antithetic variates, we conduct 10,000 independent pairs of simulation replications, making a total of 20,000 replications. In each pair, the scheme of U and $1 - U$ is used to generate the random variates. Interarrival times and service times are generated separately to ensure the synchronization. For the cases without using antithetic variates, we conduct 20,000 independent simulation replications.

Table 2. Variance of Estimators With and Without Antithetic Variates

	Var(Average Wait Time)	Var(Average System Time)
Without AV	1.34	1.36
With AV	0.57	0.57

Table 2 compares the variances obtained from these simulations. Antithetic variates significantly reduce the variances of both measurements.

REFERENCES

1. Hammersley JM, Morton KW. A new Monte Carlo technique: antithetic variates. *Proc Camb Philol Soc* 1956;52:449–475.
2. Wilson JR. Proof of the antithetic variates theorem for unbounded functions. *Math Proc Camb Philos Soc* 1979;86:477–479.
3. Page E. Simulation of queuing systems. *Oper Res* 1965;13:300–305.
4. Mitchell B. Variance reduction by antithetic variates in GI/G/1 queueing simulations. *Oper Res* 1973;21:988–977.
5. George LL. Variance reduction for a replacement process. *Simulation* 1977;29:65–74.
6. Nelson BL. Antithetic-variate splitting for steady-state simulations. *Eur J Oper Res* 1988; 36(3):360–370.
7. Avramidis AN, Wilson JR. Correlation-induction techniques for estimating quantiles in simulation experiments. *Oper Res* 1998;46: 574–591.
8. Sarkar PK, Prasad MA. Variance reduction in Monte Carlo radiation transport using antithetic variates. *Ann Nucl Energy* 1992; 19(5):253–265.
9. Milgram MS. On the use of antithetic variates in particle transport problems. *Ann Nucl Energy* 2001;28(4):297–332.
10. Cheng RCH. Variance reduction methods. In: Wilson JR, Henriksen JO, Roberts SD, editors. *Proceedings of the 18th Conference on Winter Simulation*; 1986 Dec 08–10; Washington (DC). New York: WSC '86 ACM; 1986. pp. 60–68.
11. L'Ecuyer P. Efficiency improvement and variance reduction. In: Manivannan MS, Tew JD, editors. *Proceedings of the 26th Conference on Winter Simulation*; Winter Simulation Conference; 1994 Dec 11–14; Orlando (FL). San Diego (CA): Society for Computer Simulation International; 1994. pp. 122–132.
12. Law AM, Kelton DM. *Simulation modeling and analysis*. 3rd ed. Boston (MA): McGraw-Hill Higher Education; 2000.
13. Halton JH, Handscomb DC. A method for increasing the efficiency of monte carlo integration. *J ACM* 1957;4(3):329–340.
14. Hammersley JM, Handscomb DC. *Monte Carlo methods*. London: Methuen; 1964.
15. Kleijnen JPC. *Statistical techniques in simulation, Part I*. New York: Marcel Dekker; 1974.
16. Rubinstein RY, Samorodnitsky G, Shaked M. Antithetic variates, multivariate dependence, and simulation of complex stochastic systems. *Manag Sci* 1985;31:66–77.
17. Brantley P, Fox BL, Schrage LE. *A guide to simulation*. 2nd ed. New York: Springer; 1987.
18. McGeoch C. Analyzing algorithms by simulation: variance reduction techniques and simulation speedups. *ACM Comput Surv* 1992; 24(2):195–212.
19. Halton JH, Sarkar PK. Increasing the efficiency of radiation shielding calculations by using antithetic variates. *Math Comput Simul* 1998;47(2–5):309–318.
20. Staum J. State of the art tutorial II: simulations for financial engineering: efficient simulations for option pricing. *Proceedings of the 35th Conference on Winter Simulation: Driving innovation*, Winter Simulation Conference; 2003 Dec; New Orleans (LA). 2003. pp. 258–266.
21. Cole GP, Johnson AW, Miller JO. Feasibility study of variance reduction in the logistics composite model. In: *Proceedings of the 39th Conference on Winter Simulation: 40 Years! the Best Is Yet To Come*; Winter Simulation Conference; 2007 Dec 09–12; Washington (DC). Piscataway (NJ): IEEE Press; 2007. pp. 1410–1416.
22. Cheng RCH. The use of antithetic variates in computer simulations. *J Oper Res Soc* 1982; 33:229–237.
23. Cheng RCH. Antithetic variate methods for simulation of processes with peaks and troughs. *Eur J Oper Res* 1984;15:227–236.
24. Franta WR. A note on random variate generators and antithetic sampling. *INFOR* 1975;13:112–117.
25. Schmeiser BW, Kachitvichyanukul V. Correlation induction without the inverse transformation. *Proceedings 1986 Winter Simulation Conference*. Washington (DC); 1986. pp. 266–274.
26. Wilson JR. Antithetic sampling with multivariate inputs. *Am J Math Manag Sci* 1983; 3:121–144.
27. Fishman GS, Huang BD. Antithetic variates revisited. *Commun Assoc Comput Mach* 1983;26:964–971.

28. Tuffin B. Variance reduction applied to product form multiclass queueing networks. *ACM Trans Model Comput Simul* 1997; 7(4):478–500.
29. Nelson BL. Variance reduction in the presence of initial-condition bias. *IIE Trans* 1990;22:340–350.
30. Avramidis AN, Wilson JR. Integrated variance reduction strategies for simulation. *Oper Res* 1996;44:327–346.
31. Yang W, Liou W. Combining antithetic variates and control variates in simulation experiments. *ACM Trans Model Comput Simul* 1996;6(4):243–260.

APPLICATION OF OPERATIONS RESEARCH IN AMUSEMENT PARK INDUSTRY

UTKU YILDIRIM
Prorize, LLC, Atlanta,
Georgia

The amusement park is a well-known concept that describes a highly capitalized recreational environment consisting of rides, attractions, and facilities aimed toward family-oriented entertainment. Although the concept of amusement park is regarded as an American invention, the world's oldest operating amusement park, Bakken, dates back to 1583 and is located in Klampenborg, Denmark [1]. The first amusement park built in the United States, which is still in operation, is Lake Compounce in Bristol, Connecticut, which was opened in 1846 [2]. In 1955, Disneyland was introduced as a new amusement park concept that puts emphasis on several themes by combining architectural elements with costumed personnel, restaurants, and retail shops.

In the latter part of the twentieth century, theme parks have turned into one of most visited tourist destinations. Currently, there are more than 400 amusement and theme parks in the United States. According to International Association of Amusement Parks and Attractions (IAAPA), 341 million people visited amusement and theme parks in the United States in 2007 and enjoyed more than 1.5 billion rides [2]. Number of visitors and revenue increased 14% and 42%, respectively, from 1997 to 2007. In the same time frame, amusement parks and theme parks in the United States generated \$12 billion in revenues [2].

Amusement parks face increasing competition as the number of amusement parks increases and new entertainment technologies emerge. In order to attract new customers and increase repeat visits of loyal customers, theme park management needs to understand and manage customer expectations.

To that end, theme parks are collecting more information about their customers, including but not limited to ride selection, food and beverage choice, and shopping habits through the use of cutting edge technology.

Operations research is uniquely positioned to take advantage of this explosion of data. Customer behavior inside the park can be tracked and modeled with the goal of improving customer experience by reducing the wait time at attractions or supporting facilities (such as food and beverage locations). Optimizing the throughput of each facility based on the level of demand can also serve to improve operational efficiency. Lastly, pricing and revenue management techniques can be used to match customer demand segments with revenue-maximizing ticket prices.

The main purpose of this article is to provide a review of operations research models applied to the amusement park industry. Although there is a long history of research on the marketing and strategic management of theme parks, literature on applications of operations research to the amusement park industry is limited. The remainder of the article is organized as follows: First, we give an overview of simulation techniques used in the design of rides and management of the wait time associated with rides. Then we concern ourselves with the intelligent management of park resources to increase visitor satisfaction via utilization of optimization techniques. The next two sections provide an overview of revenue management tactics that can be applied to pricing tickets effectively to prevent revenue leakage and to ancillary revenue sources. Since theme park and amusement park possess similar characteristics except the concept of theming used in theme parks, these terms are used interchangeably in the article.

RIDE DESIGN AND MANAGEMENT

In a survey completed by IAAPA, rides are listed as the number one factor that attracts

visitors to theme parks, and 46% of people put the roller coasters as their favorite ride [1]. Visitors frequently attempt to take as many rides as possible to get the most enjoyment out of their park visits. Unpleasant consequences of high participation, such as long lines and high wait time, can be avoided by increasing the park capacity.

On the other hand, construction of new rides and renovation of existing rides make up the biggest portion of the theme park's capital investment. O'Brien reported that investment required for a new roller coaster is between \$3 and \$26 million dollar per ride [3]. Signature attractions such as Walt Disney World's "Mission: Space" can cost up to \$150 million [4]. The cost of increasing an existing ride's capacity can also cost millions. For example, the renovation of the "Finding Nemo" ride cost Disney nearly \$100 million [4]. Hence, a ride designed to handle maximum expected demand and avoid long lines requires a massive initial investment that may lead to wasted/idle capacity.

Theme park management can utilize Operations research models to address this problem. The first step in this process is to estimate the number of customers who will attend the park each day of the year. Since these numbers are highly variable and seasonal, a "design day" concept is utilized to summarize this information. A design day is a hypothetical day that possesses a specific characteristic defined by the targets of a theme park: i.e., forecasted demand for the design day has to be greater than forecasted demand for 95% of days across year. Finally, design day is employed to identify the optimal capacity of rides, facilities, and other supporting elements that balance the initial cost and customer experience.

Estimation of the capacity of a ride due to complex safety regulations, theme-specific requirements, and customer mix arrival pattern is challenging. The realized ride capacity is a combination of theoretical ride capacity and throughput rate of the ride. Ahmadi reports high variability of the throughput rate due to individual visitors and families who wish to ride together [5,6]. For instance, take a ride that consists of operating units, such as cars or trains, with four seats. When a

family of three is seated in an operating unit and there are no individuals in the queue to fill the last spot, throughput is 75% of the theoretical ride capacity. Ahmadi used a neural network approach to reveal the relationship between group size mixture, queue length, and throughput of the ride [7].

Desai and Hunsucker [8] proposed a simulation-based tool to estimate the capacity of a ride when one or more operational variables is altered during the design or renovation of a ride. Three existing rides are selected to develop this tool. Common operational characteristics for these rides consist of fixed ride time and capacity, loading/unloading method, and restraint checking process.

Common activities, and hence operational variables, are defined and calculated on the basis of collected data and common operational characteristics of these rides. Real-life data is organized in three groups: (i) operational variables such as loading time (time required to load customers), storage time (time required for customers to store their personal items), restraint checking time, signal ok time (time required to complete safety checks), and unloading time (time required to unload customers); (ii) ride configuration-related variables such as number of cars, number of seats per car, and ride time; (iii) customer-related variables such as arrival pattern and group sizes. In order to generalize the model, ride speed is adjusted while track length is fixed to accommodate different ride times, and triangular distribution is used to fit all random variables.

Desai and Hunsucker compared the throughput of the final model to empirical observations to validate their model. They reported that the difference between the theoretical model and the empirical results is not statistically significant. Furthermore, they designed an experiment in which they changed the parameters of each random and fixed variable to see the changes in the hourly and daily ride capacity. The results of the experiment illustrated that reduction in time spent during storage, restraint checking, and signal ok activities appeared to generate the greatest improvement in capacity. Other variables such as unloading time and number of parallel queues showed

no significant impact on capacity. Although the experiment does not consider whether the parameters are achievable in real life, it demonstrates how a decision maker can utilize this tool: (i) to estimate the increase in throughput if number of employees is increased, restraining time is reduced or extra cars are added, and (ii) to identify the optimal operating environment when demand is less than maximum capacity.

This tool can also be used in the design and analysis of new rides by using design parameters for the new ride (when available) along with operational data associated with existing rides that are similar in characteristics. Also see Tibben-Lembke [6] where closed queueing models are used rather than simulation under similar settings.

Despite all the efforts, waiting lines are unavoidable. One way to boost the customer satisfaction is to improve the quality of the time spent in the line. Theme parks often attempt to ensure that their visitors are occupied with theming, shows, or TV monitors while they are waiting. Parks also manage customers' expectation by posting estimated wait times. However, Dickson *et al.* report that wait time is still a major contributor to customer dissatisfaction [9].

Disney introduced the concept of a virtual queue, called Fastpass, at Walt Disney World. The main promise behind the concept is that visitors are assigned to a virtual location in the line and their physical attendance is not required until it is their turn in the line. As a consequence, average time spent waiting is reduced while average spending per person is increased [9].

Currently, several paid and unpaid versions of Disney's Fastpass are used in the market. The success of such an implementation depends on the number and timing of the Fastpass distribution during the day. Operations research models can be used to identify the optimal distribution that minimizes the overall wait time of visitors while considering the decision-making process of visitors (i.e., when the waiting time at the physical line is 30 min, would a visitor immediately join the line or return in an hour later?). Simulation, Kalman filtering, and queueing theory based models are proposed solve the problem

[7,10,11]. Note that virtual queues and their implementation should also be incorporated in the simulation models described earlier in this section.

PARK CAPACITY MANAGEMENT

As mentioned in the previous section, customers generally strive to maximize the number of rides they take during their visit to a theme park. One of the main factors that impacts customer flows in the park is the wait time for these rides. As the wait time for a ride increases, the customers are tempted to try other rides with shorter wait times to benefit from their limited time in the park; else they leave the park early and unsatisfied. Hence, it is crucial to understand and direct the flow of visitors in the park in order to improve service quality and customer satisfaction.

Ahmadi [5] used real data provided by a major theme park and analyzed how visitors' flow can be managed. The real data used in the analysis consists of survey data that tracks visitors' transition from one ride to another, operational data such as park opening and closing hours, ride capacities, and historical daily demand. Conceptualization of a visitor flow model consists of the destination of the visitors and the time that they spent at each destination. Once the transition probabilities are estimated via the survey answers, the remaining issue to be addressed is determining how much time the visitors are spending on rides. Intending to produce the desired result, one of the models described in the previous section can be utilized to estimate the throughput of each ride.

Ahmadi provides two different optimization models that maximize minimum-weighted number of rides given any period of time during the day [5]. The author uses a weighted average rather than the average number of rides per person per day (which is a standard metric in the industry due to fact that all rides are not equally desired). The first optimization model utilizes the actual transition probabilities and determines optimal ride capacities but ignores the flow management problem. Second optimization

model imposes actual transition probabilities when they are zero and determines the minimum ride capacity as well as guidelines for transition probabilities. The author reports substantial gains when visitor transition probabilities are influenced by the theme park management along with the ride capacity optimization.

An emerging trend in the research is the utilization of multiagent systems to model complex social structures. Kawamura *et al.* [12] implemented this concept to the theme park industry. The main requirement behind the model is utilization of some type of communication device that provides information on rides to visitors. The model is designed to control the flow of the visitors via the communication device in order to reduce congestion and increase visitor satisfaction using their preferences. Although it seems straightforward to conceptualize the problem, it is very difficult to find the optimal solution. Hence, authors developed several approximations that maximize the social welfare of the all visitors.

At this point the question is: What is required to implement a successful park capacity management solution? One key aspect is the workforce scheduling. Workforce has to be managed carefully to achieve optimal ride capacity while reducing the cost as well as considering demand variability and business requirements such as minimum customer satisfaction level. Work force of a theme park consists of part-time, seasonal, and full-time employees. To better utilize the flexibility provided by the part-time employees, number of visitors arriving every hour and how those visitors flow in the park have to be tracked. Operations research techniques, such as resource allocation models, can then be utilized to determine the number and mix of employees as well as their schedule during the day.

TICKET PRICING

As mentioned in Formica and Olsen [13], the theme park industry requires substantial capital investment not only to facilitate the design and development of the initial

establishment but also to update existing attractions and build new attractions. Compared to the upfront high fixed cost, the impact of additional customers on the variable operating cost is minimal [14]. On the other hand, the unused capacity (or unsold tickets) at the end of each day represents wasted opportunity cost. Characteristics of the theme park business, such as high fixed cost, low variable cost, seasonal demand, and the perishability of the inventory, are similar to those faced by other resemble industries (e.g., airline) in which revenue management tactics are implemented effectively. Similar to those industries, the theme park industry also seeks high utilization via price differentiation and increased repeat customers.

Differential pricing and promotional strategies are used to increase the number of visitors during the low season and to control the surplus demand during the peak season. To that end, several ticket types and bundles are introduced to accommodate different customer segments such as locals, frequent visitors, corporate affiliates, and free travelers. Introduction of annual passes is one example of how demand is stimulated during the low season by providing visitors full year access at a deeply discounted price. As one would expect, this deep discount comes with restrictions. Restrictions such as number and timing of block-out days that an annual pass holder cannot access the park determines the price tag and purchase likelihood of each annual pass.

Several annual pass packages are introduced on the basis of add-on features such as free parking and restrictions in order to segment locals and domestic visitors. On local bases, short-term marketing campaigns such as three day resident packages are frequently used to stimulate local demand during periods when demand is expected to be lower than usual. Additionally, several discount levels are made available to corporate affiliates and multiday visitors.

Although the main portion of a theme park's revenue comes from admission fees, a number of theme parks generate additional revenue through hotel reservations, merchandise, and food and beverage sales. Additional services provide an advantage and

flexibility in pricing by allowing park tickets to be bundled with hotel rooms and/or food and beverage credit to attract families.

The key problem that the decision maker faces is determining the mix of the customer segments in order to maximize the admission revenue of the theme park. Heo and Lee [14] reviewed the current ticket pricing practice in the theme park industry. Based on their findings, they proposed implementing the traditional revenue management tactics used in hospitality and travel industry such as pricing based on season, day of week, demand level, and ticket purchase time. Next, an example is provided to illustrate how pricing based on purchase time is used in the hospitality industry to separate customer segments. Airlines as well as hotels provide discounted fares if customers make reservations 14 days ahead of their actual travel dates. Unlike most business trips, leisure trips can be planned in advance; hence leisure customers readily take advantage of the low fares by reserving hotel rooms and/or plane tickets early.

Heo and Lee tested the perceived fairness of the proposed strategies and compared these with the hospitality industry where the customers are well aware of the implementation of revenue management tactics. As a result of their survey, they concluded that customers' perceived pricing based on season and day of week in the theme park industry to be fairer than a similar strategy implemented in the hospitality industry. Although most of the US theme parks currently implement flat-rate admission fees and try to achieve these authors' suggestions through seasonal promotions and discounting programs, the authors propose a more dynamic pricing strategy based on the season and the day of week. One should also note that implementation of such a strategy in the theme park industry is costly and not straightforward. The name of the customer has to be printed on the tickets and validated while accessing the parks to prevent revenue leakages that can occur due to secondary markets.

While Heo and Lee do not provide details on how the suggested strategies might be operationalized, one might surmise that a successful implementation of a disciplined

pricing strategy requires understanding and anticipation; in other words, forecasts of daily demand for each customer segment and estimates of seasonal demand–price relationship. The forecasts then have to be fed into some type of optimization or heuristic algorithm to develop ticket prices while considering business constraints.

In a manner similar to travel and hospitality industries, one solution may be to rely on an algorithm that generates a minimum ticket price (bid price) for each day and then blocks all discount levels below the minimum ticket price. At first, this solution might appear easily implementable, but the problem with implementation of this strategy is twofold: First, it is very difficult to associate a daily ticket value with an annual pass. Second, annual pass demand blocked during the peak season will impact the demand for the low season. As the number of dates that are blocked during peak season increases, the volume of the annual pass sales decreases as well as the business that annual pass holders bring during the low season.

An alternative to this approach is building a simulation model that gives the decision maker flexibility to test different parameters and strategies easily. Advantages and disadvantages of each approach can then be determined before making the final decision.

OTHER REVENUE SOURCES

Although the main portion of a theme park's revenue comes from admission fees, other sources, such as hotel reservations, merchandise, and food and beverage sales, contribute up to 70% of the revenue. Six Flags and Disney generated, respectively, 47.6% and 68% [15,16] in the same order of their revenue from admissions in 2007.

Price Waterhouse Coopers completed a research on the on-site accommodations at 29 top European theme parks [17]. Of these 29 theme parks, 14 possess on-site accommodations and 5 planned to develop on-site accommodations by the end of 2010. A former group of theme parks reported that on-site accommodations contributed to the satisfaction of customers by providing a complete

experience, and contributed to the theme park's revenue by tapping into the corporate market.

Although the degree of the theming varies across hotels, it is reported as the main element that contributed to the overall experience of the leisure visitors. Consequently, this led to an increased mean length of stay of 2.8 days for theme parks with hotels compared to 1.7 days without hotels. Theme parks also command higher room rates with the introduction of the themed rooms. On the other hand, facilities such as conference rooms and spas allowed theme parks to reach different markets and increase utilization during the low season. In addition, park tickets that are bundled with hotel rooms increased the marketing power and revenue of the theme parks.

During the high season, the real challenge is to match excess demand with the limited supply on hand. Hence, theme parks have to be clever about how to price the hotel rooms in order to maximize revenue. Rooms priced too high keep the customers away and the unused capacity will perish; rooms priced too low cause a theme park to miss revenue opportunity. This problem is a specific case of the capacity control problem that is well known in operations research. The solution of the problem is inventory assigned to each price level: in other words, price for each remaining room in the inventory. While it is known that the capacity control problem can be formulated as a dynamic programming model, this formulation is intractable in practice due to its size and complexity. As a result, various approximation methods are proposed in the literature. Decomposition and deterministic linear programming approximations are formulated and have been successfully used in practice. Lately, several stochastic programming approaches that consider demand uncertainty have been published. For further discussion see Buke *et al.* [18].

Food and beverage facilities in a theme park consist of quick-service and sit-down restaurants. Sit-down restaurants share several characteristics with the hotels, such as reservation lead time, fixed capacity, perishable inventory, and seasonal demand. Hence, the associated revenue optimization problem

can be conceptualized in a similar manner. In the case of a restaurant business, inventory is defined as the time that is required to complete a meal and output is the number of tables available for reservation in addition to price differentiation. For further discussion of this topic, see Kimes [19].

In addition to hotels and restaurants, Rajaram and Ahmadi [20] estimated that merchandise sales contribute up to 40% of theme park profits. Unlike the hotel and restaurant businesses, merchandise inventory is not perishable and location of inventory in the store is adjustable. Hence, a different revenue optimization approach must be taken. Pricing with the consideration of inventory is a well-studied topic not just in general but also specifically in the field of revenue management. Please see the literature for a general review of this topic [21]. For such problems, the kernel of the solution is in the estimation of the price–demand relation. In real-world applications, it is difficult to find a case for which demand for an item is related only to its price. Ke [22] took a more holistic approach and considered shelf space and location of each product as well. Assortment changes in a store are incorporated in to the price–demand model along with seasonality. The model aimed to identify the assortment changes (i.e., shelf space and/or location of one or more products are changed) that made positive impact on the store revenue. He utilized data from a major resort destination and reported 5% revenue improvement over historical assortment changes. Other factors that cannot be controlled by the store management such as number of visitors, weather, and competing products that contribute to the variability in demand must be identified and incorporated into the formulation in order to isolate their impact on demand.

Rajaram and Ahmadi [20] developed a survey to assess the relationship between visitor flow and store sales. In the survey, visitors were expected to enter the rides in which they participated, time that they entered the queue, time that they left the ride, and amount of money they spent after each ride. A regression in which the dependent variable is store revenue and the independent

variable is the number of visitors is developed, and the adjusted R^2 is for the resulting model is 0.96. Contrary to the aforementioned approach, authors developed a model that utilizes this relationship to maximize revenue rather than price. They built an optimization model in which revenue is modeled as weighted visitor flow where weights are calculated for each location using per-person spending at that location. As a result of the analysis, they estimated an average profit lift of 9%.

CONCLUSION

Although marketing and strategic management issues in the theme park industry have been studied extensively, documented applications of operations research in the theme park industry are limited. Simulation techniques have been used to design rides and manage the wait time associated with rides as well as controlling the flow of the visitors via multiagent systems. Optimization techniques have been used to manage park resources intelligently and increase visitor satisfaction. Finally, revenue management tactics have been applied to pricing tickets effectively to prevent revenue leakage.

Because theme parks share several characteristics with other service industries, the research on the latter one can be a good starting point to augment the literature. Although there is detailed research in the field of revenue management on hotels, merchandise, and food and beverage, it is still an open question how existing models can be improved by incorporating the characteristics of the theme park industry. Also, there is no published research completed on theme park work force scheduling problem.

REFERENCES

1. Adams JA. The american amusement park industry: a history of technology and thrills. Boston (MA): Twayne Publishers; 1991. ISBN: 0805798218.
2. IAAPA. Available at <http://www.iaapa.org/pressroom/AmusementParkIndustryIndex.asp>.
3. O'Brien T. New coasters highlight capital park improvements for 2000. *Amusement Bus* 1999;111(37):32–33.
4. Kirsner S. Rebuilding tomorrowland. *Wired* 2002;10:12.
5. Ahmadi R. Managing capacity and flow at theme parks. *Oper Res* 1997;45(1):1–13.
6. Tibben-Lembke R. Maximum happiness: amusement park rides as closed queueing networks. Working Paper.
7. Guo Q, Liu J, Chen X. Optimization model and simulation of the queueing system with QuickPass. *Proceedings of the 6th World Congress on Intelligent Control and Automation*. Dalian, China; 2006. pp. 1401–1404.
8. Desai SS, Hunsucker JL. A sensitivity analysis tool for improving the capacity of amusement rides. *J Simul* 2008;2(2):117–126.
9. Dickson D, Ford RC, Laval B. Managing real and virtual waits in hospitality and service organizations. *Cornell Hotel Restaur Adm Q* 2005;46(1):52–68.
10. Lovejoy TC, Aravkin AY, Schneider-Mizell C. Kalman queue: an adaptive approach to virtual queueing. *UMAP J* 2004;25(3):337–352.
11. Brega ML, Cantarero AL, Lee CL. Developing improved algorithms for quickpass systems. *UMAP J* 2004;25(3):319–336.
12. Kawamura H, Kurumatani K, Ohuchi A. Modeling of theme park problem with multiagent for mass user support. *Multi-Agent for Mass User Support, International Workshop, MAMUS 2003*. Volume 3012, *Lecturer Notes in Computer Science*. Berlin: Springer; 2004. pp. 48–69.
13. Formica S, Olsen MD. Trends in the amusement park industry. *Int J Contemp Hospitality Manage* 1998;10(7):297–308.
14. Heo CY, Lee S. Application of revenue management practices to the theme park industry. *Int J Hospit Manage* 2009;28(3):446–453.
15. Disney 10-K. Annual Report. 2008.
16. Six Flags 10-K. Annual Report. 2008.
17. Clark J, Hall L. European theme park wars: hotels help refresh park revenues. *Hospitality Directions: Europe Edition*. PriceWaterHouseCoopers; 2004.
18. Buke B, Kuyumcu A, Yildirim U. New stochastic linear programming approximations for network capacity control problem with buyups. *J Pricing Revenue Manage* 2008;7(1):61–84.

19. Kimes S. Restaurant revenue management. *Cornell Hotel Restaur Adm Q* 1998;39(3): 32–39.
20. Rajaram K, Ahmadi R. Flow management to optimize retail profits at theme parks. *Oper Res* 2003;51(2):175–184.
21. Elmaghraby W, Keskinocak P. Dynamic pricing in the presence of inventory considerations: research overview, current practices, and future directions. *Manage Sci* 2003;49(10):1287–1309.
22. Ke W. The marketing operations interface in consumer retail: theory and practical approach [PhD dissertation]. Columbia University; 2009. Chapter 4.

APPLYING VALUE OF INFORMATION AND REAL OPTIONS IN R&D AND NEW PRODUCT DEVELOPMENT

JEFFREY M. KEISLER

Department of Management Science
and Information Systems, Boston
College of Management, University
of Massachusetts, Boston,
Massachusetts

PAUL Y. MANG

McKinsey & Company, Chicago,
Illinois

INTRODUCTION

A typical research and development (R&D) or new product development () project involves up-front investment in research, which, if technically successful, will lead to a viable product, which will hopefully succeed in the market and result in a net profit. The way the investment decisions and uncertainties about their outcomes unfold in stages has made such problems fruitful areas for application of decision analysis (DA). DA methods result in improved valuation of project business plans.

Rather than evaluating plans using a risk-adjusted required rate of return to determine the value of a proposed project (i.e., the discounted cash flow approach, which tends to penalize long-term R&D due to its inherent, though diversifiable, risk), the positive cash flows in the market phase of the plan are decreased by the probability that those revenues will not be realized. In addition, DA methods take into account the fact that, if the product never reaches the market, costs associated with commercializing the product do not have to be incurred. A project that will be stopped if the product performs poorly has higher expected value (EV) than a project that will be pursued and launch its product in the market no matter what its performance is, that is, resolving

uncertainty about performance before making the product launch decision leads to an increase in the project EV. This increase in value is the expected value of information (VoI) regarding product performance.

Real options analysis (ROA) represents a concept closely related to VoI. They are based on an analogy—sometimes literal and sometimes just suggestive—to options on financial instruments, for example, the most basic financial options—put and call options on stock shares. Options are called *options* because the owner of the option actually has an option to do something or not. With put options on stocks, the owner of the option has the right to sell a stock for a fixed exercise price by some future expiration date (according to the terms of the option contract), as well as the right not to sell. If the stock is worth less than the exercise price at the expiration date, the owner of the option sells the stock for more than it is worth and gains \$1 for each \$1 by which the exercise price exceeds the stock's market price at that time. The owner of the option typically does not actually sell the stock at the exercise price, but rather settles up with the seller of the option for the gain, if it is positive. A call option allows the holder to buy at the exercise price and is of value if the actual stock is worth more than the exercise price and to do nothing otherwise. *Real options*, a term originated in the 1970s [1], came to prominence around the late 1990s. There are by now many high quality articles and books on the topic, and the references in this article are a tremendous resource for the reader interested in learning more. Real options are real situations where there are choices decision makers can make that typically limit their downside and preserve the upside by waiting before taking final actions. Practitioner-oriented articles [2,3] have described how business decisions can be mapped to financial options and to parameters for value calculation. One situation isomorphic to a financial call option, for example, is where a decision maker waits to launch a product and does

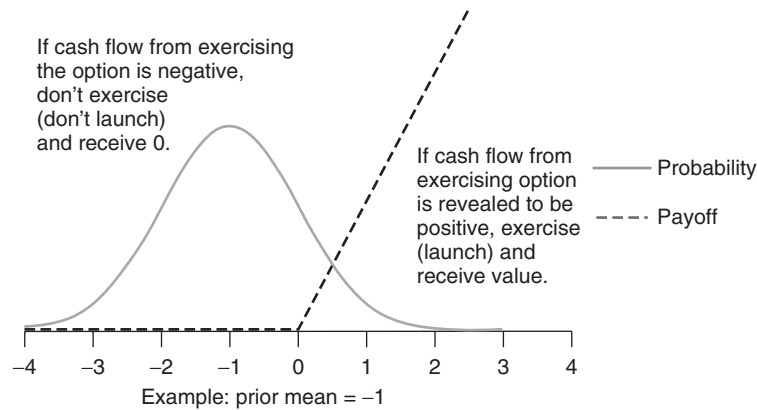


Figure 1. Option value or VoI is the integral of payoff $\times$ probability density.

so only if its performance ultimately appears promising.

Other than accounting for discount rates and for the premium paid to purchase the option, the value of the option is simply the integral of the product of probability density and value gained at different actual prices of the asset that may be obtained at the time the exercise decision must be made (Fig. 1). This is similar or equivalent (in the case of log-normal distributions) [4], to VoI in the two-act linear loss problem [5,6]. In the options theory, instantaneous returns on the underlying asset are usually assumed to be Brownian, resulting in a log-normal distribution on the asset's final price. This distribution (with a parameter for the variance of instantaneous changes) is combined with a linear loss integral and discount rate in the form of the Black–Scholes [7] option pricing formula. In DA, the probability distribution over uncertain values are typically assessed subjectively (e.g., by fitting to assessments for discrete sections of the cumulative probability curve; in particular, Gaussian distributions are sometimes used and fitted to match the first and second moments). The uncertain values assessed in DA could be the project value itself, or parameters that give information relevant to project value, and either perfect or imperfect information about these values may be available at the time of later decisions.

The situation described above is one kind of real option, which we call an *innovation*

option, where the company retains the choice of whether to launch a product until after learning the results of technical development. In the field of real options, a number of standard business decisions under uncertainty have been characterized with simple underlying microeconomic cash flow models each with its specific assumptions about which variables are random. Trigeorgis [8] classified a variety of real options, including deferring investment, defaulting during staged construction, expanding production, contracting production, shutting down and restarting operations (much of this is analyzed in rigorous economic detail in Dixit and Pindyck [9]), abandoning assets or projects, switching uses of assets, and growth of business lines.

Over time, a variety of phenomena such as corporate capabilities [10], regulations [11], and many others have been considered as real options. In DA, there has not been the same move to formally classify business application-specific decision models, and technique focuses on structuring models in any situation.

Under the right assumptions, ROA and DA frames have been shown to be equivalent [12]. Valuing a project that contains a real option should be the same as valuing it using DA, and valuing the option itself should be the same as valuing some obtainable information. The applied methods are compatible [13]. As practical tools, they have relative strengths. DA is more flexible with respect

to problem type and distribution, ROA has ready-made valuation formulas for specific problems and specific assumptions about distributions on prices (which is a plus if those assumptions reflect reality, but a limitation if not [14]). Combined approaches have been proposed [15,16] in which real options are used for parameters where public market prices are available and for risks where trading is possible, while DA assessment techniques are used to incorporate internal information for other uncertain parameters and for private risks. Furthermore, a combined approach may prove to be more transparent and in some cases allow simpler encoding of models to calculate the option value [17]. In addition, the ROA frame is more information focused rather than decision focused. With financial options, the focus is on what drives option price, for example, volatility, duration, and exercise price. Likewise, ROA focuses on understanding and, in some cases, influencing these parameters, thereby making a strategic problem out of maximizing the value added by information. DA most often identifies VoI, but tends to focus attention on understanding the drivers and distributions of the value for alternative strategies, especially those involving a more complex sequence of events and choices. ROA can include some of the structures associated with multistage decision trees in the form of compound options [18], that is, options on options, but, because these require more complex numerical calculations, their adoption has been limited.

Overall, real options and value-of-information/DA approaches can improve planning for innovation-oriented/R&D projects [19–22]. By definition, in innovative projects, new information is obtained over the course of the project, and the information is useful in setting that course. The key to realizing the benefits of these approaches is having a “smart organization” [23] oriented toward them leveraging them. Organizations may view real options and corresponding DA methods as a way of thinking (and a language), an analytical tool, or an organizational process [24]. The following sections detail, for one common type of option, how organizations can best add value to the

opportunities they obtain by understanding how their own capabilities combine with the project-based drivers of real option value and VoI.

Example

Consider a company developing a new product based on a promising technology for which it has just received a patent. The company funds the development of the product based on the hope that, after seven years with reasonable additional investment, the new product will grab a large market share and be highly profitable for a long period of time. That is the optimistic business case. But it is uncertain what it will take to make the technology commercially viable. It must be engineered to improve performance along several criteria (mass, speed of response, strength, quality, and durability). At this level, factors might be similar for all sorts of development projects ranging from a new pharmaceutical compound to clean-energy technology. Depending on the difficulty of development, the stream of investment costs, the time to market (and therefore the patent life), and the ultimate performance levels are all uncertain. Related to these factors and also uncertain are production cost and the market share the product will attain, and ultimately the profit (net present value, or NPV, of the stream of cash flows) that the entire effort will generate.

If the uncertainty is high enough, there is a good chance that even an initially attractive product could lose money—or an initially questionable product could become profitable. If at some point prospects for the product appear negative enough, efforts can be terminated to avoid incurring additional costs.

ROA and VoI can both be used capture the impact of this uncertainty, its anticipated resolution, and the impact on the probability distribution on the project value from flexibility to change in plans as the situation changes.

In a classic real options approach (explained in more detail elsewhere in this volume), we construct a binomial lattice (Fig. 2) showing the project’s projected value going up (u) or down (d) by a fixed amount in each

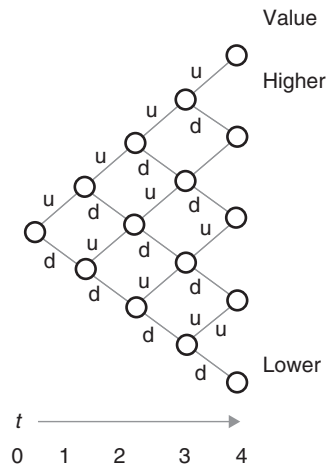


Figure 2. Binomial lattice for real options analysis.

period, for example, each year through the point at which the product launch decision is made. The value goes up if engineering results or market changes have been positive compared to the initial expectation, and it goes down if they are negative. At any point in the lattice, the nominal project value at time t is the result of all the up and down moves, with an adjustment for the time value of money. If the up and down moves each multiply project value by 100% plus or minus some percentage, the nominal project value is the initial value multiplied by the product of the changes. If the changes are additive, then nominal project value is the initial value plus the sum of the changes. The value at any point in time follows a binomial distribution. As the number of time periods increases, this distribution approaches log-normal (if changes are multiplicative) or normal (if changes are additive). In fact, the size of the up and down steps would typically be estimated based on a holistic estimate of the uncertainty in a single later period, and fitted to be of the right magnitude to make this work for the number of discrete time increments assumed. At any point, the project can be stopped with known costs. Using dynamic programming, we work back from the end of the lattice and calculate for each point in the lattice whether it would be best to stop at that point or to continue,

given the probability distribution on the value associated with continuing.

A DA approach to the same problem would use a decision tree with each uncertainty (time, cost, performance, market share, etc.) represented as a chance node with two or more possible outcomes and probabilities assigned for each outcome. End point values represent the NPV of the project under each path or scenario. Depending on the timing of when those uncertainties are resolved, the tree can contain decision nodes following some of the chance nodes, representing choice points at which the project may be abandoned if this leads to greater EV than continuing.

If the identified uncertainties are associated with sequential investments and are resolved essentially in sequence and at evenly spaced intervals, the decision tree closely resembles the binomial lattice. If the real option model is structured to allow for different volatilities at different times (or with respect to different drivers of uncertainty), the lattice may closely resemble a standard decision tree. Finally, if it is practical to assume that there is a single decision point in the future at which (if a flexible approach is taken) it is possible to terminate the project, then all the uncertainties prior to that point can be combined to generate a single distribution over project value, and the problem can be represented as a simple two-stage decision tree (Fig. 3). We can compare the EV of the project if it is simply continued with uncertain consequences with the EV at the first stage of the decision tree if there is a later decision point after some or all of the uncertainty is resolved. The difference in these two values can be thought of as either the value of the real option or as the VoI.

ORGANIZATIONAL CAPABILITIES AND REAL OPTIONS

Organizational capabilities are recognized as being important sources of competitive advantage for firms. Proponents of the resource-based view (RBV) perspective [25,26] argue that capabilities, whether in product development, efficient production, or market access, provide firms with

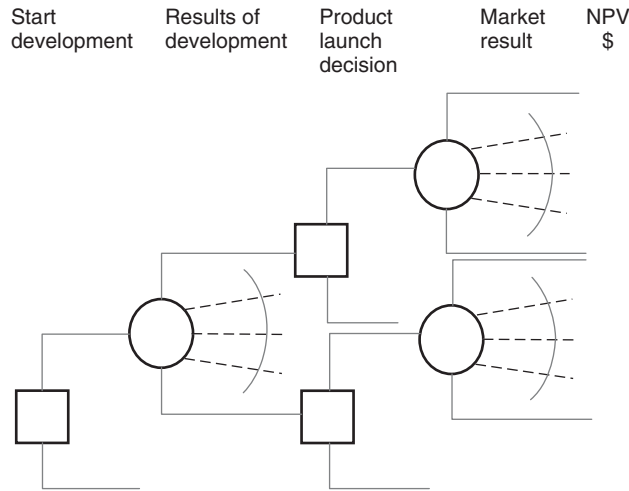


Figure 3. Two-stage decision tree.

opportunities to achieve profitable strategic positions in product markets. Over time, firms accumulate idiosyncratic combinations of capabilities that, in turn, provide unique opportunities for profitable business growth or expansion [27–31].

The options framework described above has been applied to address this concept of the firm's capabilities. Myers [1] proposed that some portion of a firm's value is based on *growth options*, yet to be realized opportunities for profitable investments. Existing research has focused on capital investment decisions [9,32] and firm or technology acquisition decisions [33–35]. Given the importance of product development as a source of new business opportunities, the application of the options framework for the management of R&D activities merits attention. While it is understood that decisions concerning R&D that create and preserve options are of value to the firm [10,36,37] and that uncertainty can actually increase a project's value [38], the relationships among the project and the firm's characteristics must also be explored as ROA seeks broader acceptance [39]. What drives option value for an organization?

During the course of translating a new technology idea into a novel commercial product or service, information regarding the technical and market feasibility of a project is revealed. The firm has the opportunity to exploit the accumulated information

about the potential net benefits of a project before committing partially or completely irreversible resources to commercialize the technology.¹ In its most simple form, an innovation project should be viewed as a sequence of two decisions: a decision to gather information about the project's prospects (experimentation stage), followed by a decision to commercialize the project (implementation stage). With this sequence of events, investment decisions for these innovation options should differ from the standard investment criteria associated with other productive assets [12,40].

The flexibility to delay making investment decisions involving commercialization activities, such as investments in specialized plant and equipment or introductory marketing campaigns, limits the appropriateness of standard investment criteria. For example, a computer software firm might initiate numerous seemingly negative NPV projects to develop new application programs given that it maintains the flexibility to abandon any individual project if prospects appear unfavorable after early development steps

¹Thus, *exploiting* an innovation option is defined as the realization of the additional EV by developing the project to a later stage. The *commercialization* of an innovation option would involve the actual production of goods or delivery of services.

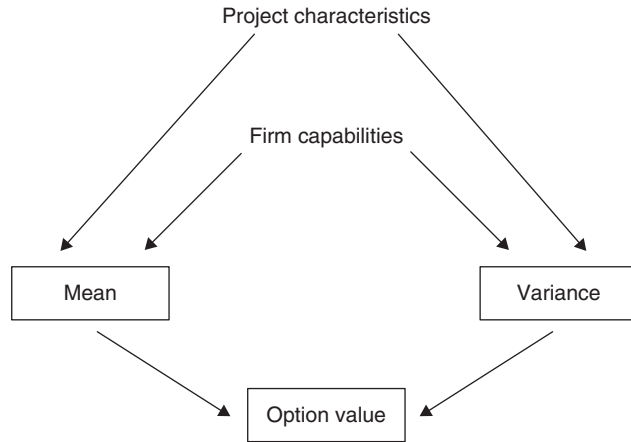


Figure 4. Option frame decision model.

are conducted. Under these circumstances, the firm will find it worthwhile to finance the first steps to gather preliminary information even on those projects that initially seem least attractive.

We can combine the statistical decision theory and VoI with the concept of the firm's capabilities. While a number of authors have prescribed that R&D projects should be treated as options [37], it is common to consider only the initial project variance as the primary contributor to option value [41].² Additional project and firm-level variables are also critical in understanding how firms can best manage their investments in innovation opportunities.

A basic dyadic model illustrates how a firm can maximize the value of its innovation efforts by investing in projects that match its organizational capabilities. First, we identify characteristics that determine the potential option value of an innovation project. We

then consider the firm's capabilities that permit the firm to exploit high option value opportunities. We specifically consider the firm's *learning capability* and *abandonment capability*. An appropriate fit between project and the firm's characteristics permits the firm to appropriate the option value associated with its innovation opportunities. In addition to tactical management of individual projects, organizations should think about strategically balancing their capabilities (which themselves represent a cost necessary to exploit options) with their flow of projects.

ONE-SHOT DECISION FRAME VERSUS OPTION DECISION FRAME

In our model, we translate qualitative characteristics of the firm and the project into quantitative descriptions of the determinants of project value, from which the option value is computed (Fig. 4). In determining how the option value changes as the firm's and project's characteristics change, we identify simple, testable explanations for differences in firm's performance. This modeling approach is flexible since future extensions can quantify the impact on the firm option value of any number of different capabilities and characteristics. Since the same units are used to value projects and value organizational capabilities, our framework provides a set of tools for R&D management decisions.

²For example, Sykes and Dunham [41] suggest identifying *critical assumptions* (those factors whose uncertainty can lead to a significant negative project NPV) as a risk-management technique. They advocate setting tasks and milestones for uncertain ventures so as to achieve maximum learning (reduction in the range of uncertainty) per dollar. Our model captures this concept of maximum learning per dollar more explicitly, in a way that is consistent with Bayes' rule and with the actual loss function faced by the investor.

We consider two possible cases of a firm evaluating its decision regarding an innovation project.

Case A. First, consider a firm endowed with a project, but with no option to delay decisions about implementation until a later date. The firm can decide to go ahead with the project or to abandon the project at time t_0 only. We call this case the *one-shot decision frame*, and refer to the EV of the project to the firm here as the *intrinsic value* (V_1). In practice, such a firm would use a rule that depends on *a priori* information; the firm might, for example, approve projects that exceed a preset hurdle rate adjusted for risk. If proceeding with the project has positive EV, then V_1 is the project EV, otherwise the firm would reject the project and $V_1 = 0$.

Case B. The firm can decide to go ahead with the project or to abandon the project either at time t_0 or later at time t_f , when more information is available. We call this case the *option decision frame*, and refer to the EV of the project to the firm here as the *potential project value* (V^*_2). The firm is again assumed to maximize its EV.³ We often refer to the value of the option implicit in Case B. We call this the *potential option value*, $V^*_{(2-1)} = V^*_2 - V_1$.

In practice, we may observe that a firm treats certain investments as one-shot decisions, and others as option-type decisions. What might determine the type of behavior observed? In CASE A, the firm's observed behavior could be due to the fact that the project has no potential option value or the fact that the firm is not in a position to realize the potential option value of the project.⁴

In Case B, we can infer that there is both an option and the firm is in a position to exploit it.

Project Factors that Contribute to Option Value

Before we analyze the factors that determine the magnitude of a project's option value, we define the following terms.

x	the ultimate value of the project if it is pursued
$\mu_0 = E(x)$ at t_0	EV of the project if the firm proceeds at t_0
$\mu_f = E(x)$ at t_f	EV of the project at t_f

We assume that, at time t_0 , the ultimate value of the project, x , is normally distributed with mean μ_0 and variance U , where $U = \sigma_0^2$. At time t_f , x is normally distributed with mean μ_f and variance σ_f^2 ; this implies that some uncertainty may remain at t_f . Finally, at time t_0 , μ_f is normally distributed with mean μ_0 and variance σ_{0f}^2 (the *preposterior variance* of μ_f).

We assume normally distributed variables and a rate of information revelation (resolvability, denoted by R) for a project:

$$\sigma_0^2 = \sigma_{0f}^2 + \sigma_f^2$$

$$R = (\sigma_f^{-2} - \sigma_0^{-2}) / (t_f - t_0).$$

In order to simplify exposition, let us now assume that

$$(t_f - t_0) = 1, \text{ so } R = (\sigma_f^{-2} - \sigma_0^{-2}),$$

in other words, R is the percentage of uncertainty that can be resolved prior to the final decision point (t_f). If it is reasonable to consider R to be constant, that is, signals about true project value arrive at a constant rate,

³The firm is assumed to maximize the following: $\text{Max}\{0, E(\text{Max}[0 - \text{abandonment costs at } t_f, E(\text{project value at } t_f)])\}$. For simplicity, we have no discounting and use EV instead of the expected NPV.

⁴The firm would not be able to *exploit* the innovation option because it lacked the capability to reveal

information about the project through development efforts. For our basic model, we assume that projects or innovation options are nontransferable. The transferability of a project (project characteristic) and the capability to engage in interfirm R&D transactions (firm capability) are factors that can be added to future extensions of this model.

R can be assumed to be the upper bound on the rate of increase in precision (inverse of variance) in the estimate of the project's value (i.e., how fast can the uncertainty be resolved). It could be a project characteristic that does not change when t_0 and t_f are varied, and it depends on the characteristics of the project but not the firm. If all of the uncertainty is not resolved before the decision, then all the volatility is not actionable [42] and is captured as the option value.

We define

$$z = \mu_0 / \sigma_{0f},$$

$$V_1 = \text{Max}(\mu_0, 0).$$

The term z is the standardized value of the mean of the distribution; it serves as a useful measure of whether, in practical terms, the decision is more or less of a long-shot. Now, the upper limit on the option value (potential option value) only depends on certain project characteristics and is determined by the equation:

$$V_{(2-1)}^*(\mu_0, U, R)$$

$$= V_2^* - V_1 = E[\max(x, 0)] - \max[0, E(x)],$$

where x is normally distributed with mean μ_0 and variance $= U - [U/(1 - R)]^{-1}$, and $H(z)$ refers to the standard normal hazard function⁵ evaluated at z ; then we get

$$V_{(2-1)}^* = \sigma_{0f} H(z). \quad (1)$$

This value is increasing in σ_{0f} and decreasing in the absolute value of μ_0 , as shown in Charts 2 and 3. These relationships restate well-known results [43].

By definition, when $V_{(2-1)}^* = 0$, it makes no difference whether a firm uses the option decision frame or the one-shot decision frame. However, when $V_{(2-1)}^* > 0$, viewing the project as a one-shot decision could mean a forfeit in the EV; the potential option value component of the project would

go unrecognized. In fact, the higher the potential option value, the greater is the opportunity that may be lost.

We find that there are relationships among the different project characteristics as they relate to the potential option value. For example, increasing preposterior variance increases the option value more for projects with mean near 0 than projects with mean far from 0; and, similarly, shifting a project's mean value toward 0 increases the potential option value more for high variance projects than for low variance projects.

Equation (1) demonstrates that in our model only three project characteristics are needed to determine the potential option value, and therefore the appropriate decision frame. These are *mean* (μ_0 ranging from $-\infty$ to ∞) and *resolvable uncertainty* (σ_{0f}^2), which can be determined from the initial uncertainty ($U = \sigma_0^2$, ranging from 0 to ∞) and the *resolvability* of that uncertainty (R , ranging from 0 to 1). Mean, initial uncertainty, and resolvability are sufficient to determine the option value because R and U determine σ_f , while U and σ_f determine σ_{0f} . Specifically, the potential option value component of an innovation project is high when

- mean is close enough to 0 so that there is a reasonable likelihood that the firm would change its decision on the basis of what it learns during the experimentation stage;
- there is uncertainty to resolve and prior standard deviation must be large enough so that, under some circumstances, a firm would want to change its decision after experimentation; and
- the uncertainty must be resolvable and posterior standard deviation must be lower than prior standard deviation, or else the presence of uncertainty merely implies risk that cannot be avoided.⁶

⁵The hazard function for a probability distribution evaluated at the point z is the ratio of the probability density at z divided by the cumulative probability at z .

⁶It is possible to have two projects with identical resolvable uncertainty values, σ_{0f} , one of which arises from high values for both σ_0 and σ_f , and the second from low values for both σ_0 and σ_f . Since resolvable uncertainty rather than initial uncertainty determines option value, both projects, all

Firm's Characteristics and Option-Exploiting Capabilities

Firms in our basic model have two characteristics: *Abandonment capability* and *learning capability*. Abandonment capability is incorporated into the model simply as the cost the firm pays to abandon the project at t_f .⁷ The lower the cost to drop out of projects, the greater is the firm's abandonment capability. We denote this capability as A , and define it as the incremental cash flow⁸ to the firm when it drops out of a project. A is assumed to be negative, ranging from $-\infty$ to 0.

For learning capability from experimentation, we define L , with values that can range from 0 to 1. A value of 1 corresponds to a perfect firm that resolves all the uncertainty that can be resolved between t_o and t_f . The actual effectiveness of experimentation for a given firm on a given project is simply (RL) , that is, the product of the resolvability of the uncertainty for the project and the firm's learning capability about (resolvable) uncertainty. We shall denote the values that depend on L by using it as a superscript to distinguish those values from what they would be for

the perfect firm (potential values), which are denoted with the superscript “*”.

The firm's learning capability is a reflection of its technical expertise as well as its organizational skills at absorbing and applying new information. Making an analogy between information revelation and sampling,⁹ we define the increase in precision (σ^{-2}) between t_0 and t_f for the firm with learning capability, L , as

$$\Delta(\sigma_{of}^{-2})^L = L\Delta(\sigma_{of}^{-2})^*.$$

Since precision is additive, we can paraphrase this as

$$(\sigma_f^{-2})^L = \sigma_0^{-2} + \Delta(\sigma_{of}^{-2})^L. \quad (2)$$

Because the preposterior variance and the posterior variance must sum to the prior variance, we also have

$$(\sigma_{of}^2)^L = \sigma_0^2 - (\sigma_f^2)^L.$$

If we modify our definition of z such that

$$z(L, A) = (\mu_0 - A)/\sigma_{of}^L,$$

we find that the firm's realized project value is

$$V_2(L, A) = \sigma_{of}^L H(z(L, A)) + V_1 + A. \quad (3)$$

And therefore, the firm's *realized option value*, $V_{(2-1)}(L, A) = E[\max(x, A)] - \max[0, E(x)]$. Under our assumptions, x is normally distributed with mean μ and variance σ^2 ,

else being equal, have the same option value component. Thus, it is not enough to merely state that projects with high uncertainty should be viewed through the option frame; it is the resolvability of uncertainty, as well as the uncertainty itself that drives potential option value.

⁷In a more general model, we could also include the cost of experimentation, which would allow L to be treated directly as a choice variable that could be changed for specific projects (e.g., working staff overtime). For the purpose of discussing permanent capabilities, we assume instead that there is no incremental cost to experimentation, but rather that the effectiveness of that experimentation is a firm's characteristic. The abandonment capability parameter is specifically the additional cost incurred by abandoning the project at t_f rather than any sunk costs incurred prior to t_f .

⁸It is conceivable that a project would have a positive salvage value, but it is unlikely that this would exceed the investment in the experimental stage, which is considered lost if the project is abandoned, and capitalized if the project is pursued. A is the incremental amount written off.

⁹In sampling, adding a single sample increases precision by the inverse of the sample standard deviation squared; the latter is analogous to R in our model. The increase in precision per dollar is this value divided by the cost of a sample. A more efficient sampler would simply be one with a lower cost of sampling. This way of quantifying the real-world characteristics corresponding to R and L allows the use of statistical decision theory to describe them, thus creating a rich framework for theoretical investigation.

where $\sigma^2 = U - [(1-L)/U + LU/(1-R)]^{-1}$. Analogous to Equation (1), this is

$$V_{(2-1)}(L, A) = \sigma_{0f}^L H(z(L, A)) + A. \quad (4)$$

The realized project value and option value both depend directly on μ_0, σ_{0f}^L , and A . As a check, when the firm has perfect capabilities, the realized value is the same as the potential value (i.e., $V_{(2-1)}^* = V_{(2-1)}(1, 0)$). The comparative statics of the model, with results derived from Keisler [5], provide insights regarding the relationships among project and firm's characteristics. First, we observe that $\sigma_{0f} = (RUL)^{0.5}$ and that this is the only impact of R, U , and L on realized option value. This implies that a given percentage change in R, U , or L would have the same impact on the option value, and that the impact on σ_{0f} of an increase in L is larger when R and U are larger.

Where $V_{(2-1)}(L, A) > 0$, several facts are known about the value of sample information (analogous to option value) in a two-act linear loss problem with normal prior distributions (for details, see Raiffa and Schlaifer [43]). The impact on the option value of an increase in σ_{0f} starts low and approaches a constant slope as σ_{0f} increases. Also, the impact of a change in σ_{0f} on the realized option value is highest when z is nearest 0. The impact on z of a change in A is the same as the impact on z of the same absolute change in μ_0 , while there is an additional one to one increase in the realized option value for increases in A .

More formally (see the section titled "Appendix" for proofs):

T1: A firm's realized option value $V_{(2-1)}$ is positively related to its abandonment capability (A).

$$dV(A, \mu, z)/dA > 0.$$

T1a: The magnitude of the relationship in T1 is positively related to the project uncertainty (U) when abandonment capability is low, and negatively related when abandonment capability is high.

$$\partial^2 V(A, U)/\partial A \partial U > 0 \text{ when } A < \mu_0,$$

$$\text{and } \partial^2 V(A, U)/\partial A \partial U < 0$$

$$\text{when } \mu_0 < A.$$

T1b: The magnitude of the relationship in T1 is positively related to project resolvability (R) when abandonment capability is low, and negatively related when abandonment capability is high.

$$\partial^2 V(A, R)/\partial A \partial R > 0.$$

T1c: The magnitude of the relationship in T1 is negatively related to project mean (μ)

$$\partial^2 V(A, \mu_0)/\partial A \partial \mu_0 < 0.$$

T1d: The magnitude of the relationship in T1 is positively related to learning capability (L).

$$\partial^2 V(L, A, \sigma, \mu)/\partial A \partial L > 0$$

$$\text{when } A < \mu_0.$$

T2: A firm's realized option value is positively related to its learning capability (L).

$$dV(L, \sigma)/dL > 0.$$

T2a: The magnitude of the relationship in T2 is positively related to project uncertainty (U).

$$\partial^2 V(L, U, \sigma)/\partial L \partial U > 0.$$

T2b: The magnitude of the relationship in T2 is negatively related to project resolvability (R), if uncertainty (U) is varied to maintain a constant value for the potentially resolvable uncertainty, $U(1-R) = k$.¹⁰

$$\partial^2 V(R, U, L, \sigma)/\partial A \partial R < 0$$

$$\text{when } U(1-R) \text{ is held fixed.}$$

¹⁰As R increases, $UR = R(1-R)/k = (R-R^2)/k$, the coefficient on L in determining σ_{0f} , increases more slowly for greater values of R , diminishing to 0 as R approaches 1.

T2c: The magnitude of the relationship in T2 is negatively related to absolute value of project mean (μ).

$$\partial^2 V(L, \sigma, \mu) / \partial L \partial \mu < 0.$$

T2d: The magnitude of the relationship in T2 is positively related to the firm's abandonment capability (A).

$$\partial^2 V(L, A, \sigma, \mu) / \partial L \partial A > 0 \text{ when } A < \mu_0.$$

It is important to note that the option pricing approach for financial securities can be considered to be a special case of our model. The term μ is comparable to the strike price less the current price; however, the resolvability (R) is trivial in the case of financial assets. In the case of financial call options, the quantity about which uncertainty is resolved is the price of the underlying asset at a given point in time. Since the market updates such information over time, all parties are equally capable of assessing the market value of a share price at the appropriate date. Even though there will be uncertainty about the future value of the asset, the uncertainty of the asset at the exercise date is completely resolved.

Furthermore, the literature on option pricing for financial securities does not generally treat abandonment as a variable quantity as it is in this innovation option model.

The costs of letting a financial call option expire are trivial when compared to abandoning innovation projects, and do not vary significantly among financial market traders. A simulation of the model provided a means to examine numerical examples with varying the firm's capabilities and project characteristics. The more interesting results are described below.

Learning Capability. As Chart 1 and Chart 2 show, increasing the variance of a project does not necessarily add to its EV in an option-oriented firm. Only increasing the uncertainty that can be resolved by the firm adds value. Firms can maximize the EV by (i) selecting projects to increase the initial uncertainty while holding the final uncertainty relatively constant, and

(ii) improving their learning capability to decrease the final uncertainty.

The resolvable uncertainty facing the firm is best described by our term σ_{0f}^L , and this is what should be increased. Learning capability is particularly interesting from a strategic perspective. It is possible to have two projects with identical *theoretically resolvable uncertainty* (σ_{0f}^*) one of which arises from higher values for both σ_0 and σ_f , and the second from lower values for both σ_0 and σ_f . In the first case, the uncertainty is of a more difficult nature; the incremental value of learning capability is seen in Chart 2 as the difference between the option value at learning capability $L = 1$ and $L < 1$. Since both curves have the same value for $L = 1$ (when the firm resolves all resolvable uncertainty, regardless of the difficulty of resolution), the fact that the second curve is lower implies that learning capability is more valuable when resolvability is relatively low. In the extreme, $L = 0$ implies that the firm has no learning capability, and therefore receives no value added from the option. In fact, a firm with abandonment cost greater than 0 and for which $L = 0$ would always abandon projects with negative EV at time t_0 rather than wait to do so at t_f .

Abandonment Capability. Abandonment capability improves a firm's ability to realize option value, and it increases it the most when (i) the likelihood that the project will be abandoned after the experimentation stage is fairly high, and (ii) the project is attractive enough that there is value in going through with experimentation. The value of abandonment capability is seen in Chart 3 as the vertical gap between the two curves. In particular, as the cost of abandonment becomes arbitrarily large, no project will be pursued which has negative prior EV, since none of these projects will be abandoned at t_f . This is the case of the firm that has no abandonment capability and therefore views all project investments as one-shot decisions. As the cost of abandonment goes to 0 (perfect abandonment capability), all projects are pursued at least until t_f and abandoned then only if the implementation stage EV (μ_f) is negative. This corresponds to the firm at

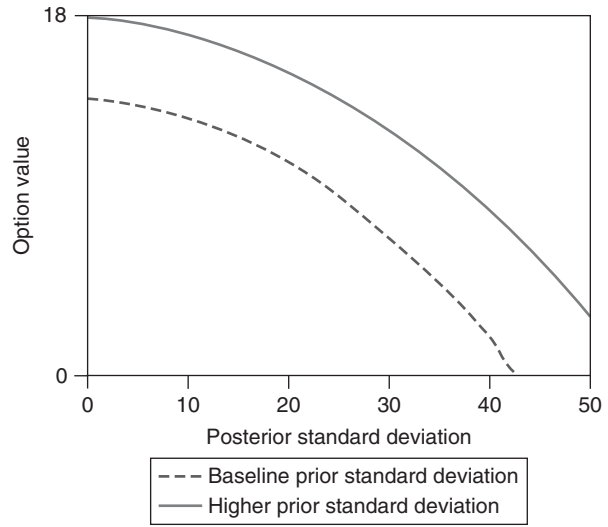


Chart 1. Option value is increasing in prior standard deviation and decreasing in posterior standard deviation of project value.

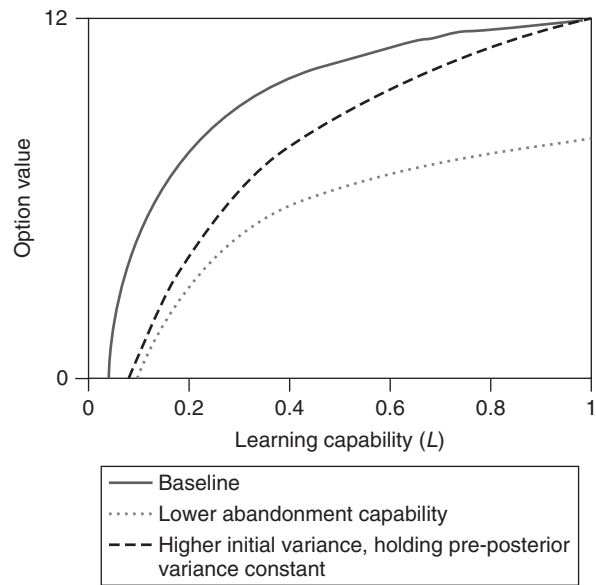


Chart 2. Option value is increasing in learning capability.

the other extreme that views all projects as options.

Interestingly, the value of abandonment capability is asymmetric in terms of the project mean value. Projects with a mean value near 0 are fairly likely to be abandoned. Projects with a high mean value are unlikely to be abandoned regardless of abandonment cost, and so abandonment capability is of less value.

It is also clear from our model that abandonment capability is of higher marginal value for projects with high preposterior uncertainty. This implies that the marginal value of abandonment capability is greater when firms also have greater learning capability. For this reason, it is conceptually convenient to combine both capabilities as the *option-exploiting capability* of the firm. Figure 5 illustrates the complementary

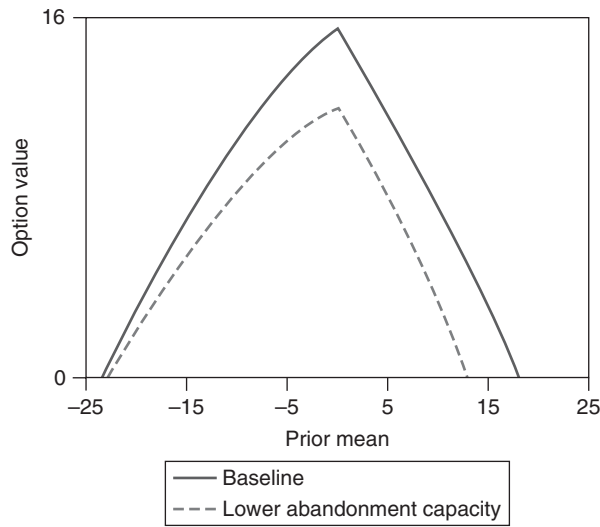


Chart 3. Option value is asymmetric due to abandonment costs.

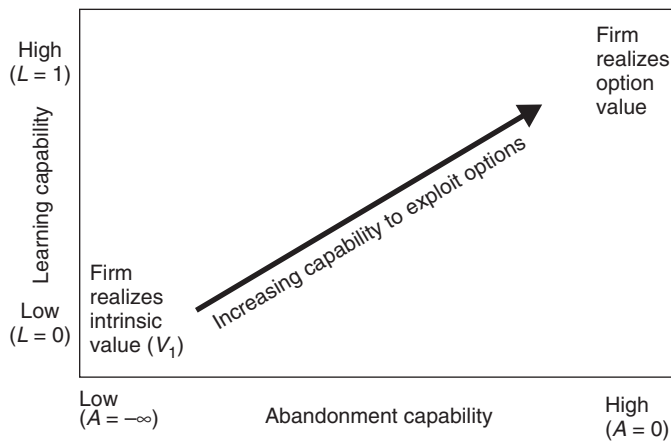


Figure 5. Option-exploiting capability.

nature of abandonment and learning capabilities.

Making Judgments about the Firm's Capabilities and Project Characteristics

The model described here obtains illustrative results using three parameters for project characteristics and two parameters for the firm's capabilities. In practice, a firm would use more detailed models of the decision problems before determining which projects to seek or develop and which capabilities to build. But it is useful to think at least qualitatively about how to gauge a firm at a general

level. This is easier with some representative projects in mind. For a given project, if it is possible to identify a single major downstream decision point, for example, launch or do not launch the product, then standard DA methods can be used estimate a prior probability distributions (summarized by mean and standard deviation) for ultimate NPV (after all relevant cash flows have actually occurred) as well as posterior probability distributions of NPV for each branch of the tree starting at the point at which the decision must be made, and, finally, given the probability for each branch up to the point of decision, a prior distribution on the mean of the posterior distribution.

This thought experiment could be done with the firm as it is, and then an ideal firm would know everything that is possible to know at the time of the decision—and would integrate this knowledge—but not more. An ideal learning firm could discover amounts of inputs per unit needed prior to its product launch decision, but would not know future market prices of inputs; it could know the physical characteristics of the product and how it functions in many different physical environments, but it would not know exactly how competitors and customers would react when it is introduced in a free market. Thus, much uncertainty is resolved but there is residual uncertainty. The resolvability parameter R depends on the prior standard deviation on NPV and the posterior standard deviation faced by this ideal firm at the time of its decision. An example of a project with completely resolvable uncertainty is the construction and sale of a building—at the time it is sold at a determined price, all relevant cash flows are known.

To gauge the firm's learning capability, L , results from the thought experiment above would be analyzed. The firm could consider how much its own estimate might vary from the ideal firm's estimate. If the firm could freeze time at the downstream decision point and process all the information it might have, how much could its estimate of expected net present value (ENPV) change? What is the variance of this change and how does that compare with the other sources of variance? A simpler approach might be to just pick a real firm to treat as a reference point for the ideal well-forecasting firm ($L = 1$), say the best in any closely related industry, and to pick another reference point for a poor learning firm, for example, a bulky bureaucratic firm that does not use any new information ($L = 0$), and to ask how far along the spectrum from the former to the latter one's own firm falls.

Assessment of abandonment capability would be, we believe (based on experience with DA assessments), fairly straightforward once one asks the question about a specific project and defines the costs of abandonment. The cost of abandonment is the sum of nonrecoverable project costs incurred up to

the point of abandonment and the additional costs that are incurred as a result of abandonment (e.g., disposal of equipment, severance of personnel, payments to terminate contracts). These figures would be available if the planning process requires pro forma cases for scenarios involving abandonment. In fact, firms that exercise options to abandon with any frequency ought to consider such cases simply in order to make good up-front decisions (although this is not always the case).

MATCHING PROJECT CHARACTERISTICS AND THE FIRM'S CAPABILITIES

The organizational capability to take advantage of the flexibility afforded by certain investment opportunities is without doubt of value [10,29]. It is clear from the present model, however, that the value of a firm's *option-exploiting capability* depends on the availability of innovation projects with a suitable option value. Firms operating in a technical and market environment with numerous innovation opportunities with high option value components can more easily recoup investments in abandonment capability and learning capability than firms facing lower option value projects. Figure 6 illustrates that the option decision frame is most appropriate for high option value projects owned by firms with high option-exploiting capability (quadrant 2). The one-shot decision frame is appropriate in the three remaining quadrants.

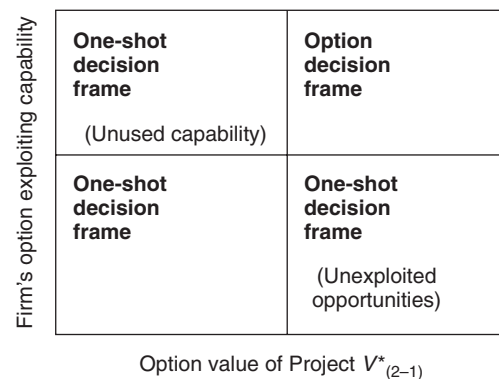


Figure 6. One-shot versus option decision frames.

It is critical to note that the need to match capabilities and opportunities is based on an assumption that projects cannot be traded in efficient markets for innovation options. A relevant market could exist if there are other firms whose abandonment capabilities, learning capabilities, or even commercialization capabilities are superior to the original owner of the project. The use of such markets would, of course, depend on the transaction costs involved with transferring innovation-related knowledge through contractual arrangements. Our model could be extended to allow for such markets by incorporating the expected price received for transferring the innovation option into our abandonment capability variable (A).

Firms in quadrants 1 and 4 have a mismatch between their capabilities and their opportunities. In quadrant 1, the firm's option-exploiting capabilities are underutilized. Such firms could seek higher option value projects or could reduce their investments in option-exploiting capabilities and move into quadrant 3. In quadrant 4, the option value of innovation projects goes unrecognized. These firms would miss opportunities for profitable investments and might benefit from investing in their option-exploiting capabilities. Firms in quadrant 4 may also have an opportunity to sell their high option value projects to other organizations with the appropriate capabilities to recognize the value of such opportunities. Thus, firms can employ both types of levers—internal capability management and project selection—to achieving an efficient resource balance that does not lead to resources wasted on projects where they are not needed, and also does not lead to losses on projects that another firm may have avoided.

Our analysis also demonstrates that the drivers of option value (e.g., mean, uncertainty, resolvability) and the drivers of option-exploiting capability (e.g., abandonment capability and learning capability) should also match each other. If a firm faces large and difficult-to-resolve uncertainty, then its investment in option-exploiting capability should be concentrated in learning capability. If it faces projects with mean

generally near 0, its investment in option-exploiting capability should be concentrated in abandonment capability. Because of the *complementarity* between abandonment capability and learning capability, firms should tend toward high levels or low levels of both capabilities.

Our simple model offers insights about the complex interrelationships that affect how successful a firm might be in extracting option value from its innovation opportunities. The primary implication of this modeling exercise is that a research manager's approach to evaluating projects should depend on both the character of the opportunity and the character of the firm in question. In particular, in high option value environments, firms that can effectively manage market exploration, technical innovation, and other organizational learning activities have a decided advantage in the race to accumulate resources and skills for competitive advantage [25,26]. Therefore, it is important for firms to know what type of option environment they operate in, and what capabilities they need to bring to exploit their particular circumstances.

An Example

The emergence of new biotechnology firms (NBFs) in the pharmaceutical industry provides a useful context to examine the implications of our analysis. Recent advances in biotechnology research have created opportunities for significant investments with significant uncertainty about prospects for success at different stages. For this reason, we might expect that pharmaceutical industry participants are endowed with numerous high option value innovation opportunities. In that case, firms can avoid significant costs associated with new drug introductions by abandoning projects if early stage research and preliminary clinical trials reveal unprofitable prospects.

This would imply that a high portion of the value of biopharmaceutical R&D opportunities resides in their option value.¹¹

¹¹Depending on the project, outcomes might be better modeled not with a normally distributed

These projects have a relatively low expected mean (especially if they must be pursued to completion) and high uncertainty that is in large measure resolved during preliminary research procedures. Do the numerous NBFs that have recently entered the industry have an advantage over incumbent pharmaceutical firms in this high option value environment?¹²

Firm's Option-exploiting Capabilities. Small entrant biotechnology firms, most started through the entrepreneurial efforts of academic researchers, have the technical strength to efficiently conduct the experimentation stage of innovation (which we believe translates to high *learning capability* because managers would quickly recognize whether the project is achieving its technical goals). NBFs would also likely exhibit high *abandonment capability*. Since researchers at these entrepreneurial firms typically have several projects "on the back burner," research attention can be quickly shifted to the most promising projects as information is gained about each project. The research function at biotechnology firms thus resembles a constantly shifting network of projects; a project that appears promising would attract researchers while a failing project would quickly lose support from internal scientists. Furthermore, an NBF's need to attract external sources of capital on an almost continual basis¹³ improves

their abandonment capability [44]; many biopharmaceutical projects are abandoned when the involved NBF fails to convince its venture capital backers or public equity market participants of the logic of additional investments in a project with disappointing preliminary results.

Large established pharmaceutical firms, with their strengths in traditional chemical-based disciplines, may be at a disadvantage in the new biology-based biopharmaceutical field [45]; this can be interpreted as a lower *learning capability* relative to NBFs. The large size of the R&D function at most pharmaceutical firms requires a formal resource allocation system that may hinder their *abandonment capabilities* [46].¹⁴ In some corporate cultures, a failed experiment is viewed as a failure of the experimenter; in such environments, abandonment can in practice be more difficult and costly. Ironically, large internal sources of capital to fund research may serve to limit a pharmaceutical firm's *abandonment capability*; unlike NBFs, which are at the mercy of the external financial markets, a large firm's internal resource allocation procedure may not respond as quickly to information revealed during the experimentation stage of development.

value, but rather as a project that can take either a high or low value. The uncertainty in this case would be expressed as the probability of receiving the high value. Whatever the firm's initial estimate of the probability—that it will receive the high value—this estimate would typically move closer to either 0 or 1 at time t_f . The fundamental arguments about the firm's capabilities and project characteristics would remain unchanged from our conclusions using the normal distribution.

¹²NBFs are very active in biopharmaceutical development. In 1991, the Pharmaceutical Manufacturers' Association reported that NBFs conducted early stage development work for approximately two-thirds of the 132 biopharmaceutical projects under development for the US drug market.

¹³For example, Sahlman [44] notes that venture capitalists, an important source of funds for NBFs,

commonly stage their investments. He observes that "By staging capital the venture capitalists preserve the right to abandon a project whose prospects look dim. The right to abandon is essential because an entrepreneur will almost never stop investing in a failing project as long as others are providing capital" (pp. 506–507).

¹⁴In fact, finance executives at Merck & Co., an established pharmaceutical firm, recognize the inadequacies of their traditional NPV analysis and have adopted option analysis for research projects [46]. It is unclear, however, how easily Merck will be able to build their option-exploiting capabilities; abandonment capability may be very difficult to develop in a large organization that spends over \$2 billion in R&D and capital expenditures per year. It is interesting to note that Merck and other pharmaceutical companies continue to enter into numerous interorganizational arrangements with NBFs to gain access to early stage research projects.

The high option-exploiting capabilities of NBFs may help explain their intensive activity in high option value biopharmaceutical projects. We hypothesize that the efficient production and marketing capabilities of established pharmaceutical firms are more appropriate for implementation (drug commercialization) activities. Thus, incumbent pharmaceutical firms might be better suited to buying the rights to drugs in latter stages of development from NBFs (when the option value component of the project diminishes), rather than developing early stage projects internally.

SUMMARY

Real options and the closely related concept of VoI in DA can improve R&D and NPD. We have considered some of the basic characteristics that determine the option value in an investment opportunity. The firm's capabilities and project characteristics whose fit drives value have precise definitions that can be quantified by borrowing established techniques from DA. While an uncertain outcome is a necessary condition for the creation of the option value, the *resolvability* of that uncertainty is also important. To capture the inherent option value of a given innovation project, however, a firm must be able to reveal information about its prospects through development efforts and have the flexibility to act in a meaningful way to acquire information. Learning capability and abandonment capability represent valuable organizational resources that can be the basis of strategic advantage.

The framework here is a starting point for addressing innovation-related capabilities,¹⁵ and it could also be applied to questions such as: How long and how intensely should

firms pursue the experimental stage before committing to implementation? Under what conditions would firms specialize in either experimentation or implementation activities? How should a firm manage interrelated projects within a portfolio? How much should firms invest in option-exploiting capabilities?

The globalization of markets and the quickening pace of technological advance have increased the volatility of many industries. In this environment, a firm's ability to add value may be determined by its recognizing real options and then utilizing the information revealed during the innovation process.

APPENDIX

This appendix provides the analytical support for the hypotheses in the main text. We use the following notation to simplify the equations in this section:

$$\sigma = (RLU)^{1/2} \quad \mu = \mu_0 - A \quad z = \mu/\sigma$$

To remove ambiguity from the partial derivatives in this section, we also use the notation

$$V(\bullet) = V_{(2-1)}(R, L, U, A, \mu, \mu_0, \sigma),$$

to define V as a function of the parameters $(\bullet)$ that are allowed to vary, with all other parameters held fixed. Finally, $f(z)$ denotes the standard unit normal probability density function evaluated at z [i.e., $f(z) = \exp(z^2/2)/(2\pi)^{1/2}$] and $G(z)$ denotes the standard unit normal right tail cumulative density function evaluated at z .

Preliminary Facts

Recalling that option value is equal to $\max(0, G(z) + f(z) - zG(z))$, we note that for $A = 0$,

$$V = \sigma \int_z^\infty x f(x) dx = \sigma(f(z) - zG(z)) \quad [43]$$

For $A \neq 0$, z is changed, and the payoff is shifted so that,

$$V = \max(0, \sigma(f(z) - zG(z)) + A).$$

¹⁵In a similar vein, Copeland and Tufano [3] essentially considers what might be called *timing* capability and calculates for a specific example the loss of option value associated with exercising at too low a market price, and the loss associated with waiting too long to even consider exercising the option, as well as how the percentage of value lost varies with the volatility of the option.

We also note for later use that $d\sigma/dR = R/2\sigma$, $d\sigma/dU = U/2\sigma$, $d\sigma/dL = L/2\sigma$, $dz/d\sigma = 1/\mu$,

$$dz/d\mu = 1/\sigma, \quad \text{and,} \quad df/dz = -zf(z).$$

The following analyses assume $V > 0$ so that the option frame is appropriate.

Proofs

T1: $dV(A, \mu, z)/dA > 0$

Proof.

$$\begin{aligned} dV(A, \mu, z)/dA &= \partial V(A, \mu, z)/\partial A \\ &\quad + \partial V(\mu, z)/\partial \mu \partial \mu / \partial A, \\ \partial V(A, \mu, z)/\partial A &= 1, \\ \partial \mu / \partial A &= -1, \\ \partial V(\mu, z)/\partial \mu &= \sigma df(z)/d\mu - \sigma dz/d\mu (G(z)) \\ &\quad + \sigma z dG(z)/d\mu \\ &= df(z)/dz - dG(z)/dz + z dG(z)/dz. \end{aligned}$$

Because $dG(z)/dz = -f(z)$, this reduces to

$$df(z)/dz + f(z) - zf(z) = -f(z),$$

and the whole derivative becomes $1 + f(z)$, which is greater than 0.

Note, for $\mu_0 > A$, the G term must be replaced by F , the left tail cumulative probability, but the first derivative of option value with respect to A is still positive.

T1a: $\partial^2 V(A, U)/\partial A \partial U > 0$ when $A < \mu_0$, and $\partial^2 V(A, U)/\partial A \partial U < 0$ when $\mu_0 < A$.

Proof.

$\partial V(A, U)/\partial A = G(z)$, that is, the probability that the project will be abandoned.

$$\begin{aligned} \partial^2 V(A, U)/\partial A \partial U &= \partial/\partial U (\partial V(A, U)/\partial A) \\ &= \partial G(z)/\partial U \\ &= (U/2\sigma) \partial G(z)/\partial \sigma \\ &= (U/2\sigma) (-f(z)) \partial z/\partial \sigma \\ &= (U/2\sigma) (-f(z)) (-(\mu_0 - A))/\sigma^2. \end{aligned}$$

Noting that $f(z) > 0$, $U > 0$ and $\sigma > 0$, we can see that the derivative is positive only when $\mu_0 - A > 0$. Of course, if the one-shot frame is appropriate, then $\partial V(A, U)/\partial A = 0$.

T1b: $\partial^2 V(A, R)/\partial A \partial R > 0$

Proof. Same as T1a, only replace U with R .

T1c: $\partial^2 V(A, \mu_0)/\partial A \partial \mu_0 < 0$

Proof.

$$\partial^2 V(A, \mu_0)/\partial A \partial \mu_0 = \partial/\partial \mu_0 (\partial V(A, \mu_0)/\partial A)$$

We know that

$$\partial V(A, \mu_0)/\partial A = G(z)$$

because the marginal value from an incremental increase in A is simply the probability that the project will be abandoned. Knowing also that an increase in μ_0 makes abandonment less likely, we get

$$\partial/\partial \mu_0 (G(z)) = \partial G(z)/\partial z \partial z/\partial \mu_0 = -f(z)/\sigma < 0$$

T1d: $\partial^2 V(L, A, \sigma, \mu)/\partial A \partial L > 0$ when $A < \mu_0$

Proof. Same as T1a only replace R with L .

T2: $dV(L, \sigma)/dL > 0$

Proof.

$$\begin{aligned} dV(L, \sigma)/dL &= \partial V(L, \sigma)/\partial L \\ &\quad + \partial V(L, \sigma)/\partial \sigma \partial \sigma / \partial L \end{aligned}$$

and

$$\partial V(L, \sigma)/\partial L = 0 \quad (\text{when } \sigma \text{ is held constant}),$$

$$\partial V(L, \sigma)/\partial \sigma > 0, \quad \text{and} \quad \partial \sigma / \partial L = RU > 0.$$

T2a: $\partial^2 V(L, U, \sigma)/\partial L \partial U > 0$

Proof.

$$\begin{aligned} \partial^2 V(L, U, \sigma)/\partial L \partial U &= \partial/\partial U (\partial V(L, U, \sigma)/\partial L) \end{aligned}$$

$$\begin{aligned}
&= \partial/\partial U \partial V(L, U, \sigma)/\partial \sigma \partial \sigma/\partial L, \\
&= \partial/\partial U (\partial V(L, U, \sigma)/\partial \sigma) \partial \sigma/\partial L \\
&\quad + \partial/\partial U (\partial \sigma/\partial L) \partial V(L, U, \sigma)/\partial \sigma \\
&= 0 + \partial/\partial \sigma \{[(f(z) - zG(z)) \\
&\quad + z^2 f(z)]L/2\sigma\} \partial \sigma/\partial U \\
&\partial \sigma/\partial U = U/2\sigma > 0, \text{ and} \\
&\partial/\partial \sigma (\partial V(L, U, \sigma)/\partial L) > 0.
\end{aligned}$$

Intuitively, increasing σ scales up the effect of L on σ , as well as decreasing the absolute value of z , which makes the option frame decision more likely to differ from the one-shot frame decision.

T2b: $\partial^2 V(R, U, L, \sigma)/\partial L \partial R < 0$ when $U(1 - R)$ is held fixed

Proof.

$$\begin{aligned}
&\partial^2 V(R, U, L, \sigma)/\partial L \partial R \\
&= \partial/\partial \sigma [\partial V(R, U, L, \sigma)/\partial L] \partial \sigma/\partial R
\end{aligned}$$

If we add the constraint $U(1 - R) = k$, that is, that total resolvable uncertainty is held constant, we find that

$$\partial \sigma/\partial R = -L^{1/2}k/(1 - R^2) < 0, \text{ and because}$$

$$\partial/\partial \sigma [\partial V(R, U, L, \sigma)/\partial L] > 0 \text{ (shown in T2a),}$$

the product of the two partial derivatives is less than 0.

T2c: $\partial^2 V(L, \sigma, \mu)/\partial L \partial \mu < 0$

Proof.

$$\begin{aligned}
&\partial^2 V(L, \sigma, \mu)/\partial L \partial \mu \\
&= \partial/\partial \mu (\partial V(L, \sigma, \mu)/\partial L) \partial \sigma/\partial \mu
\end{aligned}$$

We know that

$$\partial/\partial \mu (\partial V(L, \sigma, \mu)/\partial \sigma) < 0$$

(because a greater value of μ implies a lower probability of exercising the option

to abandon and smaller loss avoided by exercising the option), and that

$$\partial \sigma(L)/\partial L > 0,$$

so their product is also less than 0.

T2d: $\partial^2 V(L, A, \sigma, \mu)/\partial L \partial A > 0$ when $A < \mu_0$

Proof.

$$\partial^2 V(L, A, \sigma, \mu)/\partial L \partial A = \partial^2 V(L, A, \sigma, \mu)/\partial A \partial L$$

Then the proof is the same as T1a, only we replace R with L .

Observations

When $V = 0$ (even the experimentation stage is not worth pursuing), the marginal value of changes in parameters are 0 until the changes are large enough to change the decision maker from the one-shot frame to the option frame, at which point the results hold. Again, when $V = 0$, it is interesting to consider the dual question of how much of a change in project characteristic or the firm's capability does it take to shift the decision maker to the option frame. For example, the abandonment capability a firm must have in order to pursue the option frame is decreasing in the amount of uncertainty and decreasing in the project mean, while the uncertainty required to shift the decision maker to the option frame is decreasing in A . We also observe that $\partial^2 V(L)/\partial L^2$ may go from positive to negative, depending on UR ; in other words, the value of learning capability saturates at some point.

Acknowledgments

We appreciate comments and suggestions from Charles Rosa, Elizabeth Teisberg, and George Wu and others. The authors contributed equally to this article. Dr. Mang's contribution was largely completed while he was at the McCombs School of Business, University of Texas at Austin.

REFERENCES

1. Myers S. Determinants of corporate borrowing. *J Financ Econ* 1977;5(2):147–175.
2. Luehrman T. Investment opportunities as real options: getting started on the numbers. *Harv Bus Rev* 1998;7(4):51–60.
3. Copeland T, Tufano P. A real-world way to manage real options. *Harv Bus Rev* 2004;82(3):90–99.
4. Herath H, Park C. Real options valuation and its relation to Bayesian decision making methods. *Eng Econ* 2001;46(1):1–32.
5. Keisler J. Comparative static analysis of information value in a canonical decision problem. *Eng Econ* 2004;49(4):339–349.
6. Bickel E. The relationship between perfect and imperfect information in a two-action risk-sensitive problem. *Decis Anal* 2008;5(3):116–128.
7. Black F, Scholes M. The pricing of options and corporate liabilities. *J Polit Econ* 1973; 81(3):637–654.
8. Trigeorgis L. Real options: managerial flexibility and strategy in resource allocation. Cambridge (MA): MIT Press; 1996.
9. Dixit A, Pindyck R. Investment under uncertainty. Princeton (NJ): Princeton University Press; 1994.
10. Kogut B, Kulatilaka N. Options thinking and platform investments: investing in opportunity. *Calif Manage Rev* 1994;36(2):52–71.
11. Alleman J, Rappoport P. Modelling regulatory distortions with real options. *Eng Econ* 2002;47(4):389–417.
12. Smith J, Nau R. Valuing risky projects: option pricing theory and decision analysis. *Manage Sci* 1995;41(5):795–816.
13. Miller L, Park C. Decision making under uncertainty: real options to the rescue. *Eng Econ* 2002;47(2):105–150.
14. Bowman E, Moskowitz G. Real options analysis and strategic decision making. *Organ Sci* 2001;12(6):772–777.
15. Smith J, McCardle K. Valuing oil properties: integrating option pricing and decision analysis approaches. *Oper Res* 1998;46(2):198–217.
16. Borison A. Real options analysis: where are the emperor's clothes? *J Appl Corp Finance* 2005;17(2):17–31.
17. Brandão L, Dyer J, Hahn W. Using binomial trees to solve real-option valuation problems. *Decis Anal* 2005;2(2):69–88.
18. Herath H, Park C. Multi-stage capital investment opportunities as compound real options. *Eng Econ* 2002;47(1):1–27.
19. Perdue R, McAllister W, King P, *et al.* Valuation of R and D Projects using options pricing and decision analysis models. *Interfaces* 1999;29(6):57–74.
20. Neely J, de Neufville R. Hybrid real options valuation of risky product development projects. *Int J Technol Policy Manage* 2001; 1(1):29–46.
21. Mun J. Real options analysis: tools and techniques for valuing strategic investment and decisions. 2nd ed. New York: Wiley Finance; 2005.
22. Huchzermeier A, Loch CH. Project management under risk: using the real options approach to evaluate flexibility in R&D. *Manage Sci* 2001;47(1):85–101.
23. Triantis A, Borison A. Real options: state of the practice. *J Appl Corp Finance* 2001; 14(2):8–24.
24. Matheson D, Matheson J. The smart organization. Boston (MA): Harvard Business School Press; 1998.
25. Barney J. Firm resources and sustained competitive advantage. *J Manage* 1991; 17(1):99–120.
26. Peteraf M. The cornerstones of competitive advantage: a resource-based view. *Strateg Manage Rev* 1993;14(3):179–191.
27. Baldwin C, Clark K. Capabilities and capital investment: new perspectives on capital budgeting. *J Appl Corp Finance* 1992;5:67–82.
28. Dierickx I, Cool K. Asset stock accumulation and sustainability of competitive advantage. *Manage Sci* 1989;35(12):1504–1511.
29. Henderson R, Cockburn I. Measuring competence? Exploring firm effects in pharmaceutical research. *Strateg Manage J* 1994;15:63–84.
30. McGrath R, MacMillan I, Venkataraman S. Defining and developing competence: a strategic process paradigm. *Strateg Manage J* 1995;16(4):251–275.
31. Teece D, Pisano G, Shuen A. Dynamic capabilities and strategic management. *Strateg Manage J* 1997;18(7):509–533.
32. Pindyck R. Irreversibility, uncertainty, and investment. *J Econ Lit* 1991;29:1110–1152.
33. Folta T, Leiblein M. Technology acquisition and the choice of governance by established firms: insights from option theory in a multinomial logit model. *Academy*

- of Management Proceedings. Dallas, Texas; 1994. pp. 27–31.
34. Hurry D, Miller A, Bowman E. Calls on high-technology: Japanese exploration of venture capital investment in the United States. *Strateg Manage J* 1992;13(2):85–101.
35. Kogut B. Joint ventures and the option to expand and acquire. *Manage Sci* 1991;37(1):19–33.
36. Bowman E, Hurry D. Strategy through the options lens: an integrated view of resource investments and the incremental-choice process. *Acad Manage Rev* 1993;18:760–782.
37. Dixit A, Pindyck R. The options approach to capital investment. *Harv Bus Rev* 1995;73(3):105–115.
38. Morris P, Teisberg E, Kolbe A. When choosing R&D projects, go with long shots. *Res Technol Manage* 1991;34(1):35–40.
39. Hartman M, Hasan A. Application of real options analysis for pharmaceutical R&D project valuation—Empirical results from a survey. *Res Policy* 2006;35(3):343–354.
40. Roberts K, Weitzman M. Funding criteria for research, development, and exploration projects. *Econometrica* 1981;49(5):1261–1288.
41. Sykes H, Dunham D. Critical assumption planning: a practical tool for managing business development risk. *J Bus Venturing* 1995;10(6):413–424.
42. Lewis N, Eschenbach T, Hartman J. Can we capture the value of option volatility? *Eng Econ* 2008;53(3):230–258.
43. Raiffa H, Schlaifer R. *Applied statistical decision theory*. Boston (MA): HBS Division of Research; 1961.
44. Sahlman W. The structure and governance of venture-capital organizations. *J Financ Econ* 1990;27(2):473–521.
45. Pisano G. The R&D boundaries of the firm: an empirical analysis. *Adm Sci Q* 1991;35:153–176.
46. Nichols N. Scientific management at Merck. *Harv Bus Rev* 1994;72(1):89–99.

APPROXIMATE DYNAMIC PROGRAMMING I: MODELING

WARREN B. POWELL
Department of Operations Research
and Financial Engineering,
Princeton University, Princeton,
New Jersey

INTRODUCTION

Stochastic optimization problems pose unique challenges in how they are represented mathematically. These problems arise in a number of different communities, often in the context of problems that introduce specific computational characteristics. As a result, a number of contrasting notational styles have evolved, which complicate our ability to communicate research across communities. This is particularly problematic in the general area of multistage, stochastic optimization problems, where different communities have made significant algorithmic contributions, which have applications to a wide variety of problems.

The range of problems that can be modeled as stochastic, dynamic optimization problems is vast. Examples of major problem classes include:

- *Optimization over Stochastic Graphs.* This is a fundamental problem class that addresses the problem of managing a single entity in the presence of different forms of uncertainty with finite actions.
- *Dynamic Resource Allocation Problems.* These include scheduling people and machines, routing vehicles, managing inventories, and investing in new facilities and technologies. These problems arise in supply chain management, personnel management, health care, military operations, agriculture, and energy.

- *Demand Management.* These problems include booking strategies for airlines, hotels, hospitals, vendor-managed inventories, and incentives to control the demand for energy.
- *Management of Financial Portfolios.* These include how a portfolio should be spread over different investments to strike a balance between risk and return
- *R&D Portfolio Problems.* These include how research and development portfolios should be managed to reach specific technological goals; what investment strategy to pursue to ensure that we will meet government targets for renewable energy in 30 years. These decisions need to be made in the presence of uncertainty about prices, climate, technology, and government policy.
- *Pricing Problems.* These include how products and services should be priced to maximize total revenue.
- *Engineering Control Problems.* These include how much CO₂ should we release into the atmosphere; what time window should one commit to for providing service; at what speed should you fly your aircraft and so on.
- *Sensor Management Problems.* These include how to manage a team of technicians collecting information about the presence of disease in the population, the concentration of pollution or radiation in the atmosphere, or the concentration of pollutants in the water.

These problems are hardly exhaustive, but hint at the range of applications and types of complexities that we might encounter. In all of these problems, we face the challenge of making decisions sequentially, in that we make a decision, and then observe information that we did not know when we made the first decision. We then get to make another decision, after which we see more information. The goal is to make decisions over time that achieve some objective.

There are several ways to model these problems, and different communities have evolved modeling and algorithmic strategies to deal with specific problem classes. For example, the simulation community typically uses myopic policies (rules that do not directly consider the impact of decisions now on the future), which might depend on one or more tunable parameters. For example, a (q, Q) inventory policy orders new product if the inventory falls below q , and places an order to bring the total inventory up to Q . In this case, q and Q are tunable parameters, which can be optimized to find the best policy, indirectly taking into account the impact of decisions now on the future.

The Markov decision process (MDP) community assumes that we can represent our system as being in a state s at time t . If we choose action a , then we let $p(s'|s, a)$ be the probability that we then land in state s' . If $C(s, a)$ is the contribution (reward) we earn if we choose action a when in state s , then we can find the best action by solving Bellman's optimality equation given by

$$V(s) = \max_a (C(s, a) + \gamma \sum_{s'} p(s'|s, a) V(s')), \quad (1)$$

where γ is a discount factor. We are assuming that we are maximizing total discounted contributions over an infinite horizon. The challenge is computing the value $V(s)$ for each (discrete) state s . There are powerful algorithms for solving this problem, but they require enumerating the set of potential states. While there are many problems that can be solved with this strategy, the method breaks down when s consists of a vector of elements. This produces the well-known curse of dimensionality of dynamic programming.

Approximate dynamic programming (ADP) is both a modeling and algorithmic framework for solving stochastic optimization problems. Most of the literature has focused on the problem of approximating $V(s)$ to overcome the problem of multidimensional state variables. In addition to the problem of multidimensional state variables, there are many problems with multidimensional random variables, and multidimensional

decision variables (most commonly referred to as *actions* in the dynamic programming community, or *controls* in the engineering literature). These three challenges make up what have been called the *three curses of dimensionality*.

It is important in any presentation on dynamic programming to acknowledge the different communities that have contributed to the field. The challenge of making good decisions over time in the presence of uncertainty arises in a number of fields, and as a result it is not surprising to see similar ideas being developed under different notational systems and different vocabularies. These communities include:

- *Discrete MDPs*. This covers research in computer science as well as the MDP community in operations research. These problems are typically characterized by discrete states (with possibly many states), and discrete actions, but typically not very many actions.
- *Control Theory*. These communities include engineering in the physical sciences and economics. Problems are often modeled in continuous time, with decision variables (controls) that are typically continuous and low-dimensional (e.g., one to a dozen dimensions). Randomness often arises in the form of measurement noise.
- *Stochastic Programming*. This community deals with vector-valued (and often high-dimensional) decision vectors and general forms of uncertainty, which are represented using scenario trees. This community typically does not use Bellman's optimality equation as an algorithmic device.
- *Simulation Optimization*. The simulation community generally uses myopic policies to make decisions over time, but these policies may be governed by a vector of tunable parameters that can be optimized. This community also does not use Bellman's equation to guide decisions, but there are close parallels between the problem of optimizing policies in simulation, and policy optimization in dynamic programming.

Although the roots of ADP can be traced to early work by Bellman [1], the ideas evolved independently within different fields, notably the early work on training computers to play games [2,3] and the work in control theory [4–6]. The work in computer science evolved under the name “reinforcement learning,” where the first published use of this term is in Ref. 7 (the roots of this work can be found in Minsky’s Ph.D. dissertation [8]; see also Ref. 9). Reinforcement learning as a field did not really emerge until the 1980s with Barto *et al.* [10], followed by numerous contributions by Sutton and Barto through the 1980s, eventually leading to their ground breaking book [11]. Work in control theory took place under a variety of names, including reinforcement learning, adaptive dynamic programming and (later) ADP, with important early contributions by Paul Werbos [4–6,12]. The seminal book by Bertsekas and Tsitsiklis [13] introduced the term *neurodynamic programming*, but it appears that this term is being replaced with ADP (see for example, Ref. 14, Chapter 6).

While ADP in its various forms really accelerated in the 1990s in computer science and control theory, there was relatively little attention given to ADP in the operations research community until after 2000. One of the earliest papers in the operations research literature to explicitly use the term *approximate dynamic programming* is Ref. 15, although others have done similar work under different names such as *adaptive dynamic programming* [16–18]. Methods for handling vector-valued decision variables in a formal way using the language of dynamic programming appear to have emerged quite late (see in particular, Ref. 19), although other authors have used specialized techniques from math programming to solve multistage stochastic optimization problems. Pereira and Pinto [20] in particular, introduce the idea of using Benders cuts to overcome the curse of dimensionality in dynamic programming (see also Ref. 21, Chapter 11). This idea has enjoyed a substantial literature (see Birge and Louveaux [22], Higle and Sen [23]), but these have evolved independently of the ADP community.

From this discussion, we feel that any discussion of ADP has to acknowledge the fundamental contributions made within computer science (under the umbrella of reinforcement learning) and control theory. The one dimension that these communities largely ignored was problems that involved high-dimensional decision variables, which are common in operations research. The first book to bridge the gap with mainstream operations research in a thorough way did not appear until Powell [21].

MODELING A STOCHASTIC OPTIMIZATION PROBLEM

Before we can solve a problem, we have to model it. In this section, we review the five fundamental dimensions of a stochastic, dynamic systems, which include: (i) states, (ii) actions/decisions/controls, (iii) exogenous information/random variables, (iv) transition function, and (v) objective function. We use this presentation to review the different notational systems used by different communities.

States

It is with some surprise that we have found that few authors attempt to actually define a state variable. Powell [21] offers the following definition:

Definition 2.1. A **state variable** is the minimally dimensioned function of history that is necessary and sufficient to compute the decision function, the transition function, and the contribution function.

This definition is familiar to researchers in control theory (in particular, in electrical engineering). The modifier “minimally dimensioned” is intended to restrict the state variable to all the information needed, but only the information needed, so that it is as compact as possible.

The MDP framework, used in both computer science and operations research, uses s for state, or S_t for the random variable describing the state at time t . Control theorists use x . Most applications involve

modeling a physical state (the status of a piece of equipment, the amount of products in different inventories), but many problems require modeling an information state (information used to make a decision), and for some applications, a belief state (when we are unsure about the actual state of our system).

Decisions/Actions/Controls

We might refer to action a (MDPs), control u or decision x . While it is easy to view these as different variable names for the same quantity, it is also the case that these communities tend to work on different types of problems. Although exceptions abound, in most cases a refers to a relatively small (that is, easy to enumerate) set of discrete actions; the control u is typically a low-dimensional (e.g., 1–10) continuous vector (density, pressure, velocity, acceleration, price); and the decision vector x in operations research is often very high-dimensional, with hundreds to tens of thousands of dimensions.

We defer until later the problem of determining how to make a decision, other than to say that we will ultimately look for a *policy* π , which is a rule (or function) for determining a decision using the information in the state variable S_t . We might use π to represent this rule, or a function $A^\pi(S_t)$ to determine the action a_t or $X^\pi(S_t)$ to determine the decision x_t . We let Π be the set of all possible policies (or functions), which takes on different meanings as we give a policy structure in a specific setting.

Exogenous Information

We are typically interested in problems that are driven by some sort of exogenous information process. This might come from a physical system (observed prices or rainfall) or a probability distribution. These are modeled as random variables, but there is an important distinction between whether the underlying probability distribution is known or not.

While communities have standard notation for states and actions, we are unaware of any standard notation for exogenous information. The MDP community typically does not explicitly model exogenous information,

preferring the more compact representation of the one-step transition function $p(s'|s, a)$. In control theory, we often see w_t for random information. Given the preference of the applied probability community to use capital letters for random variables, we use W_t as our generic notation for random information.

A separate modeling issue for discrete-time models is the modeling of time. In a continuous-time model, W_t would represent information arriving between t and $t + dt$. For discrete-time problems, there are many authors who would let W_t be new information (about rainfall, changes in prices, new demands) arriving between t and $t + 1$, but this means that at time t , W_t is random. We prefer the convention, widely used in the applied probability community, that W_t represents information arriving between $t - 1$ and t . With this convention, any variable indexed by t is known at time t .

We let ω represent a sample path $(W_1, W_2, \dots)$ where $\omega \in \Omega$. To finish the formalism, we let $\mathcal{F}$ be the sigma-algebra on Ω (the set of events), and let $\mathcal{P}$ be a probability measure on $(\Omega, \mathcal{F})$. Finally, because information evolves, over time, we let $\mathcal{F}_t$ be the sigma-algebra generated by the variables $(W_1, \dots, W_t)$, which implies that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ is a sequence of filtrations. We can use this notation to express the dependence of state variables and decisions on information that has arrived prior to time t .

There are instances where it is useful to provide an explicit model of the history. For this purpose, we can define

$$\begin{aligned} H_t &= \text{The history of the process,} \\ &\quad \text{consisting of all the information} \\ &\quad \text{known through time} \\ &\quad t, = (W_1, W_2, \dots, W_t), \\ \mathcal{H}_t &= \text{The set of all possible histories} \\ &\quad \text{through time } t, \{H_t(\omega) | \omega \in \Omega\}, \\ h_t &= \text{A sample realization of a history,} \\ &\quad H_t(\omega), \\ \Omega(h_t) &= \{\omega \in \Omega | H_t(\omega) = h_t\}. \end{aligned}$$

In the stochastic programming community, it is common to model uncertainty through scenario trees, which represent the branching of outcomes as new information becomes available. Imagine all the outcomes

$\omega \in \Omega(h_t)$, which correspond to a common history h_t at time t . We can model this juncture as a node $n \in \mathcal{N}$, where each node n captures a particular history at a point in time. All the outcomes that meet at node n follow a common path (corresponding to the history h_t). Rather than use an explicit model of time, the modeling of scenario trees typically refers to predecessor nodes and successor nodes. A decision made at node n of the tree has to depend on the information up to that juncture. This representation does not use the concept of a state variable, where outcomes with different histories can lead to the same state.

It is interesting to contrast the use of scenario trees in stochastic programming with state variables in dynamic programming. A node in a scenario tree corresponds to an entire history. Not surprisingly, scenario trees grow exponentially in size as the number of time periods increases. For this reason, there is a literature addressing the problem of generating scenario trees, which capture desirable properties with a minimum number of outcomes [24–26]. It is perhaps interesting that the stochastic programming community, which focuses primarily on problems with multidimensional decisions x_t , views dynamic programming as a method that is limited to small problems (due to the “curse of dimensionality”) when in fact scenario trees suffer from a similar curse of dimensionality when representing exogenous information processes.

In practice, scenario trees are used most often when the history of a process plays an important role. Although the history can be easily added to a state variable, the result is an extremely large state space where there may be a unique state for each sample path (which shares a common history). We note that scenario trees are generated prior to solving the problem, which means that this method of modeling uncertainty is unable to handle problems where the exogenous outcome depends on a prior decision. For example, we may be modeling random prices to determine the value of an asset. If we sell more, the prices may drop. For such problems, we cannot generate scenario trees in advance.

The Transition Function

There are different styles for modeling how the system evolves over time. The convention in operations research is to use systems of equations, such as

$$A_t x_t + B_{t-1} x_{t-1} = b_t. \quad (2)$$

In the MDP community, the evolution of the system is described using the one-step transition matrix $p(s'|s, a)$.

In the control theory community, it is common to define a function that maps state, action, new information to new state, as in

$$S_{t+1} = S^M(S_t, a_t, W_{t+1}).$$

The function $S^M(\cdot)$ goes under many names such as *plant model* (literally, the model of a physical production plant), *plant equation*, *law of motion*, *transfer function*, *system dynamics*, *system model*, *transition law*, and *transition function*. We use “transition function,” but adopt notation that captures the widely used term *system model*. Transition functions are typically straightforward to specify (with exceptions as we note below), although in many engineering applications, they can be quite complex. For this reason, it is common practice to specify the existence of a transition function when modeling a problem without actually writing out the details. This contrasts sharply with writing out systems of linear equations such as Equation (2), where all the details of the transition are fully specified.

It is often overlooked that the one-step transition matrix is actually an expectation, since it can be derived directly from the transition function as follows:

$$\begin{aligned} \mathbb{P}(s'|s, a) &= \mathbb{E}\{\mathbf{1}_{\{s'=S^M(s, a, W_{t+1})\}} | S_t = s\} \\ &= \sum_{\omega_{t+1} \in \Omega_{t+1}} \mathbb{P}(W_{t+1} = \omega_{t+1}) \\ &\quad \times \mathbf{1}_{\{s'=S^M(s, a, W_{t+1})\}}. \end{aligned}$$

The problem with one-step transition matrices is that they tend to be extremely large, measuring the number of states times the number of states times the number of

actions. One-step transition matrices have enjoyed a rich history in the literature for MDPs, where they have facilitated an elegant theory. But in practice, it is only an extremely narrow class of problems where they can actually be computed.

Objective Function

The final dimension of our model is the objective function. We might minimize a cost or maximize a contribution (or reward). Assuming that we are maximizing, we define

$C(S_t, a_t)$ = Contribution received for being in state S_t and taking action a_t .

The contribution might be a random variable, which we would then write as $C(S_t, a_t, W_{t+1})$. In many applications, we observe the next state S_{t+1} but do not explicitly observe W_{t+1} , in which case we may write the contribution as $C(S_t, a_t, S_{t+1})$. In either case, $C(S_t, a_t)$ would be the expected contribution. The most common assumption is that we are maximizing total discounted rewards over a finite or infinite horizon, which we would write as

$$F^\pi(S_0) = \mathbb{E} \left\{ \sum_{t=0}^T \gamma^t C(S_t, A^\pi(S_t)) \right\}, \quad (3)$$

where γ is a discount factor. It is very common to let $T \rightarrow \infty$ and solve for a steady-state policy, but for other problems, solving undiscounted, finite-horizon problems is the standard model. We note that we are assuming that our objective function can be written in the form of additive rewards.

We let $A^\pi(S_t)$ be a function, parameterized in some way by $\pi \in \Pi$, that determines the action a_t given the information in the state S_t . Our goal is to find the best policy, which means solving

$$\max_{\pi \in \Pi} \mathbb{E} \left\{ \sum_{t=0}^T \gamma^t C(S_t, A^\pi(S_t)) \right\}. \quad (4)$$

Solving this optimization problem directly, even for very simple problems, is computationally intractable. The breakthrough of dynamic programming is realizing

that this problem can be solved using Bellman's optimality equation (see Ref. 27 for a modern and thorough discussion of this field), which can be written as

$$V_t(S_t) = \max_a (C(S_t, a) + \gamma \sum_{s'} P(s'|S_t, a) V_{t+1}(s')), \quad (5)$$

$$= \max_a (C(S_t, a) + \gamma \mathbb{E}\{V_{t+1}(S_{t+1})|S_t\}), \quad (6)$$

where $S_{t+1} = S^M(S_t, a, W_{t+1})$. For steady-state problems, we let $V_t(S) = V_{t+1}(S) = V(S)$. If states are discrete (and there are not too many of them), the number of actions is not too large, and the expectation is easy to compute, then Equation (5) or (6) can be used to find the best action a for each state S , which gives us a lookup table representation of a policy. Unfortunately, only a small number of toy problems meet these criteria, which is what leads us to the field of ADP.

MAJOR PROBLEM CLASSES

ADP arises because of the computational difficulties in solving Bellman's equation. Now that we have our modeling framework in place, we can discuss more precisely about the nature of these complexities.

There are three computational challenges that arise in the solution of Bellman's equation:

1. Finding the value function $V(S)$ (or $V_t(S_t)$ for finite-horizon problems).
2. Computing the expectation.
3. Finding the best action.

The nature of these challenges arises from the characteristics of three variables: the state variable S_t , the exogenous information variable W_t , and the action a_t /control u_t /decision x_t .

State variables can typically be divided between whether the state space is (i) discrete and easy to enumerate (up to thousands of states, but not millions), (ii) scalar and continuous, or (iii) a vector (whether it is discrete

or continuous). Often, cases (ii) and (iii) are equivalent from a computational perspective, but scalar, continuous variables tend to offer special structure. The third case spans discrete vectors (how many cars of each model are in inventory), continuous vectors (how much money is invested in each investment choice), vectors of categorical attributes, and of course a mixture of all of these. There are many applications where the state variable is complicated by the need to retain some portion of the history of the process. When this happens, a common modeling strategy is to use scenario trees.

For the random vector W_t , we are primarily interested in whether we can compute the expectation in Bellman's equation. Vectors of random variables may be easy if they are independent, and of course the problem is easiest when there are no autocorrelations linking observations over time. It is possible that the random information is simple, but we do not know the probability distribution. For example, the random variable may simply capture whether a person accepts a bid in an auction or not; the random variable is Bernoulli, but we do not know the probability distribution describing the person's behavior. It is also important to separate problems where W_t is independent of all prior history; problems where W_t depends on the state S_t , and problems where W_t depends on both the state S_t and action x_t .

For decisions, there may be a small number of discrete actions a , a single scalar decision, low-dimensional continuous vectors u , or high-dimensional continuous or discrete vectors x . For vector-valued actions, we typically need some sort of search algorithm such as a linear program or genetic algorithm to find x . The choice of algorithmic search strategy can have an impact on how we represent future events when making a decision.

The default way to handle any dynamic program is to discretize all the states and actions, and assume that we can compute expectations. When the number of states and actions is small enough to enumerate when discretized, and when we can compute expectations, we typically can solve Bellman's equation (Eq. 1) exactly using classical techniques [27]. If the state variable is a vector,

discretizing it can produce an exponentially large number of states, a problem that is routinely referred to as the *curse of dimensionality* (a term coined by Bellman). The same problem arises with the information variable W and the action $a/u/x$. However, there are problems where W may be continuous, but where the expectation is still easy. There are other problems where the information may be fairly simple (e.g., the behavior of an opponent or the price of a stock), but where we do not know the distribution, and therefore, we cannot compute the expectation. Finally, there are problems where the decision variable is a vector, which makes it impossible to enumerate all the actions. When all three of these problems arise (vector-valued states, uncomputable expectation, and vector-valued decisions), we say that we have three curses of dimensionality.

TYPES OF POLICIES

The challenge of dynamic programming is finding a good rule for making decisions given a state. We refer to this rule as a *policy*. Policies come in a number of forms, and the precise form can play a major role in the design of an algorithm for finding a good policy. A policy is often denoted as π , which is a generic mapping from a state to an action. It is convenient to emphasize that this is really a function. If the action is a , we might designate the function as $A^\pi(S)$ for the action that we would take if we are in state S . If your action is x , we could use $X^\pi(S)$. We design our space of possible policies using $\pi \in \Pi$, but what this means computationally is very dependent on the nature of the policy (or decision function). Examples of policies include:

1. *Lookup Tables.* For a discrete state s , $A^\pi(s)$ is the discrete action we should take. If there are 100 states and 10 actions per state, our policy space would have 1000 parameters.
2. *Parameterized Myopic Policies.* Let S_t be the amount of inventory on hand at time t . A reorder policy might be to order $X^\pi(S_t) = Q - S_t$ if $S_t < q$ and 0

otherwise. This policy is parameterized by q and Q , so we might say $\pi = (q, Q)$. The set of all possible policies is the set of potential values for q and Q .

3. *Statistical Models.* When controlling energy commitments from a wind farm, let W_t be the wind speed. We have to decide how much energy to commit to the grid, which we might model using $x_t = \theta_0 + \theta_1 W_t + \theta_2 W_t^2 + \theta_3 W_t^3$. This regression function is a policy parameterized by $(\theta_0, \theta_1, \theta_2, \theta_3)$. Within this category, we would include training a neural network to represent a policy, a strategy that is common in the neural network community (see Ref. 28, Chapter 12).
4. *Myopic Optimization Models.* Let $C(S_t, x_t)$ be the contribution earned by using decision x when we are in state S . An example might be a resource allocation problem where we are allocating people, products or machinery to different tasks or demands. S_t captures the current status of our resources and tasks, and x_t is our vector of decisions of who gets assigned to what. We could solve our problem using

$$X^\pi(S_t) = \arg \max_{x \in \mathcal{X}} C(S_t, x).$$

This means finding the best assignment of resources now, without regard to the impact of these decisions on the future. We note that solving this problem may involve using a solver for linear, nonlinear, or integer programs, or using a heuristic search algorithm such as tabu search or genetic algorithms.

5. *Tree Search.* For problems with typically small action spaces, it is possible to estimate the value of a particular state by enumerating all the actions and subsequent states that result from reaching a particular state. These methods are widely used in the design of algorithms for playing games (see Ref. 29).
6. *Roll-Out Heuristics.* When it is not possible to enumerate all the actions out of a state (as in tree search), it may

be the case that we have access to a reasonable (but suboptimal) policy. A roll-out heuristic evaluates the value of a particular state s' (that we might reach from state s using a potential action a) by following this policy starting from s' for a specified number of iterations. The value of this simulated sample path can be used to approximate the value of reaching state s' . See Ref. 30 for a more in-depth discussion of this strategy.

7. *Rolling Horizon Policies (Deterministic).* We could also optimize over a planning horizon T using

$$U^\pi(S_t) = \arg \max_{u \in \mathcal{U}} \sum_{t'=t}^{t+T} C(S_{t'}, u_{t'}),$$

where point forecasts are used to make decisions over the horizon $t, \dots, t+T$. We only use u_t as our decision to implement right now. This strategy is also known as a *receding horizon policy*, or, in the control theory community, *model-predictive control*. Rolling horizon policies are mathematically equivalent to tree search, but are normally written in the context of multidimensional decision/control problems where a solver of some sort is used to solve the optimization problem.

8. *Rolling Horizon Policies (Stochastic).* Classical rolling horizon policies use a point forecast of the future, but it is possible to use a stochastic model, which captures uncertainty in potential future events. This strategy is most popular in the stochastic programming community, which represents possible outcomes of future events as scenarios in a set Ω_t . An outcome $\omega \in \Omega_t$ would be viewed as a set of potential events over time periods $t, t+1, \dots, t+T$.

$$\begin{aligned} X^\pi(S_t) = \arg \max_{x_t \in \mathcal{X}} & C(S_t, x_t) \\ & + \sum_{\omega \in \Omega_t} p(\omega) \\ & \times \sum_{t'=t+1}^{t+T} C(S_{t'}(\omega), x_{t'}(\omega)). \end{aligned}$$

These problems have to be solved subject to constraints (known as *nonanticipativity constraints* in the stochastic programming community), which ensure that a decision x'_t does not see information that becomes available at time periods later than t' .

9. *Value Function Approximations.* In this strategy, we use an approximation of the value function in Bellman's equation, where we would make a decision by solving

$$A^\pi(S_t) = \arg \max_a (C(S_t, a) + \gamma \mathbb{E}[\bar{V}_{t+1}(S^M_t(S_t, a, W_{t+1})) | S_t]), \quad (7)$$

where $\bar{V}(S)$ is an approximation of the value of being in state S . The space of policies is the space of potential value function approximations.

Lookup tables are easy to visualize but require enumerating states and actions, so this is precisely the type of policy that is sensitive to curse of dimensionality issues. Finding the best parameters in a parameterized policy is a topic that has been widely studied in the literature known as simulation optimization [31–34] where the techniques of stochastic search are used [35]. Rolling horizon policies have been viewed as a form of ADP (see Ref. 36 for a discussion of the relationship between ADP and model-predictive control), since they are in the same mathematical class as tree search algorithms and roll-out policies [30].

ADP arises primarily when we are looking for a value function approximation that determines a decision using equations such as Equation (7). However, finding the best statistical model (policies of type 3 above) is also an important strategy in the ADP literature, where it is referred to as *approximate policy optimization*. This is particularly common in the control theory community where a policy is a neural network (literally, a statistical model). It can be argued that finding the best regression function (type 3) and finding the best value function (type 9) are mathematically equivalent (both produce functions

that are determined by regression methods), but the computational issues are different. The problem of finding the best approximation of a policy is closest to policy iteration of dynamic programming, while finding the best value function approximation is closest to value iteration.

MODEL-FREE DYNAMIC PROGRAMMING

A topic that is very popular in computer science (in the reinforcement learning community) and engineering is a problem class that is referred to as *model free*. These applications arise in the context of more complex applications, but the term *model free* can take on different meanings in different settings. In a nutshell, model-free dynamic programming arises when we cannot compute

$$a_t^n = \max_a (C(S_t^n, a) + \gamma \mathbb{E} V_{t+1}(S^M_t(S_t^n, a_t, W_{t+1}(\omega^n))). \quad (8)$$

There are three calculations implied in the solution of Equation (8):

1. computing $S_{t+1}^n = S^M(S_t^n, a_t, W_{t+1}(\omega^n))$ using the transition function,
2. computing the expectation, and
3. computing the contribution function $C(S_t^n, a)$.

There are many applications where we cannot compute some combination of these calculations. The most common are problems where we do not have an explicit transition function. Given the fact that many refer to this as the model, the lack of a model (transition function) resulted in algorithmic strategies, which address this dimension as model-free dynamic programming. These techniques apply equally to problems where we cannot compute the expectation, which can easily arise because observations are from an exogenous process where we do not know the underlying probability distribution. There are also problems where we do not have an explicit contribution (or reward or utility) function. These can arise when we are trying to mimic a human making

decisions, where we do not know the precise utility function that guides the human.

The reinforcement learning community often requires a model-free framework since this community is frequently working on problems that involve mimicking human behavior. Model-free dynamic programming is so common, in fact, that authors feel that they have to explicitly state when an algorithm is model-based. The control theory community often encounters model-free applications when the physics of a particular problem (e.g., modeling a chemical plant) is simply too complex to represent as a mathematical model.

CLOSING REMARKS

The goal of this article was to outline a general strategy for modeling stochastic, dynamic problems. Designing effective policies is a difficult challenge, which requires taking advantage of the nature of a particular problem. ADP offers very general algorithmic framework for solving these problems. An introduction to this approach is given in Ref. 37.

REFERENCES

1. Bellman R, Kalaba R. On adaptive control processes. *IRE Trans Automat Control* 1959;4:1–9.
2. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev* 1959;3:211–229.
3. Samuel AL. Some studies in machine learning using the game of checkers II – recent progress. *IBM J Res Dev* 1967;11:601–617.
4. Werbos PJ. Beyond regression: new tools for prediction and analysis in the behavioral sciences [PhD thesis]; 1974.
5. Werbos PJ. Backpropagation and neurocontrol: a review and prospectus. *Proceedings of the International Joint Conference on Neural Networks* New York: IEEE; 1989. pp. 209–216.
6. White DA, Sofge DA. *Handbook of intelligent control: neural, fuzzy, and adaptive approaches*. New York: Van Nostrand Reinhold Company; 1992.
7. Minsky ML. Steps toward artificial intelligence. *Proc Inst Radio Eng* 1961;49:8–30.
8. Minsky ML. *Theory of neural-analog reinforcement systems and its application to the brain-model problem* [PhD thesis]; 1954.
9. Mendel JM, McLaren RW. Volume 66, Reinforcement learning control and pattern recognition systems. New York: Academic Press; 1970. pp. 287–318.
10. Barto A, Sutton R, Brouwer P. Associative search network: a reinforcement learning associative memory. *Biol Cybern* 1981;40:201–211.
11. Sutton R, Barto A. Volume 35, Reinforcement learning. Cambridge (MA): MIT Press; 1998.
12. Si J, Barto AG, Powell WB, *et al.* *Handbook of learning and approximate dynamic programming*. Hoboken (NJ): Wiley-IEEE Press; 2004.
13. Bertsekas D, Tsitsiklis J. *Neuro-dynamic programming*. Belmont (MA): Athena Scientific; 1996.
14. Bertsekas DP. Volume II, *Dynamic programming and optimal control*. Belmont (MA): Athena Scientific; 2007.
15. Bertsimas D, Demir R. An approximate dynamic programming approach to multidimensional knapsack problems. *Manag Sci* 2002;48:550–565.
16. Powell WB, Shapiro JA, Simao HP. *A representational paradigm for dynamic resource transformation problems*. Basel, Switzerland: J.C. Baltzer AG; 2001. pp. 231–279.
17. Godfrey G, Powell W. An adaptive dynamic programming algorithm for dynamic fleet management, I: Single period travel times. *Transport Sci* 2002;36:21–239.
18. Papadaki KP, Powell WB. An adaptive dynamic programming algorithm for a stochastic multiproduct batch dispatch problem. *Nav Res Logist* 2003;50:742–769.
19. Powell WB, Van Roy B. Approximate dynamic programming for high dimensional resource allocation problems. In: Si J, Barto AG, Powell WB, *et al.*, editors. *Handbook of learning and approximate dynamic programming*. New York: IEEE Press; 2004.
20. Pereira MVF, Pinto LMVG. Multistage stochastic optimization applied to energy planning. *Math Progr* 1991;52:359–375.
21. Powell WB. *Approximate dynamic programming: solving the curses of dimensionality*. Hoboken (NJ): John Wiley & Sons Inc.; 2007.
22. Birge JR, Louveaux F. *Introduction to stochastic programming*. New York: Springer; 1997.
23. Higle JL, Sen S. *Stochastic decomposition: a statistical method for large scale stochastic*

- linear programming. Boston (MA): Kluwer Academic Publishers; 1996.
24. Dupačová J, Consigli G, Wallace SW. Scenarios for multistage stochastic programs. *Ann Oper Res* 2000;100(1):25–53.
 25. Høyland K, Wallace SW. Generating scenario trees for multistage decision problems. *Manag Sci* 2001;295–307.
 26. Heitsch H, Romisch W. Scenario tree modeling for multistage stochastic programs. *Math Progr* 2009;118(2):371–406.
 27. Puterman ML. Markov decision processes. Hoboken (NJ): John Wiley & Sons Inc.; 1994.
 28. Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River (NJ): Prentice Hall; 1999.
 29. Pearl J. Heuristics: intelligent search strategies for computer problem solving. Reading (MA): Addison-Wesley; 1984.
 30. Bertsekas DP, Castanon DA. Rollout algorithms for stochastic scheduling problems. *J Heuristics* 1999;5:89–108.
 31. Nelson BL, Swann J, Goldsman D, *et al.* Simple procedures for selecting the best simulated system when the number of alternatives is large. *Oper Res* 2001;49:950–963.
 32. Fu M, Glover F, April J. Simulation optimization: a review, new developments, and applications. Proceedings of the 37th conference on Winter simulation; Orlando (FL): 2005. pp. 83–95.
 33. Kim SH, Nelson BL. Selecting the best system. Chapter 17. Amsterdam: Elsevier; 2006.
 34. Chang HS, Fu MC, Hu J, *et al.* Simulation-based algorithms for Markov decision processes. Berlin: Springer; 2007.
 35. Spall JC. Introduction to stochastic search and optimization: estimation, simulation and control. Hoboken (NJ): John Wiley & Sons, Inc.; 2003.
 36. Bertsekas D. Dynamic programming and suboptimal control: a survey from ADP to MPC. *Eur J Contr* 2005;11:310–334.
 37. Powell WB. Approximate dynamic programming II: algorithms. Encyclopedia of Operations Research Management and Science. Hoboken (NJ): John Wiley & Sons, Inc.; 2010.

APPROXIMATE DYNAMIC PROGRAMMING—II: ALGORITHMS

WARREN B. POWELL
Department of Operations Research
and Financial Engineering,
Princeton University, Princeton,
New Jersey

INTRODUCTION

Approximate dynamic programming (ADP) represents a powerful modeling and algorithmic strategy that can address a wide range of optimization problems that involve making decisions sequentially in the presence of different types of uncertainty. A short list of applications, which illustrate different problem classes, include the following:

- *Option Pricing.* An American option allows us to buy or sell an asset at any time up to a specified time, where we make money when the price goes under or over (respectively) a set strike price. Valuing the option requires finding an optimal policy for determining when to exercise the option.
- *Playing Games.* Computer algorithms have been designed to play backgammon, bridge, chess, and recently, the Chinese game of Go.
- *Controlling a Device.* This might be a robot or unmanned aerial vehicle, but there is a need for autonomous devices to manage themselves for tasks ranging from vacuuming the floor to collecting information about terrorists.
- *Storage of Continuous Resources.* Managing the cash balance for a mutual fund or the amount of water in a reservoir used for a hydroelectric dam requires managing a continuous resource over time in the presence of stochastic information on parameters such as prices and rainfall.

- *Asset Acquisition.* We often have to acquire physical and financial assets over time under different sources of uncertainty about availability, demands, and prices.
- *Resource Allocation.* Whether we are managing blood inventories, financial portfolios, or fleets of vehicles, we often have to move, transform, clean, and repair resources to meet various needs under uncertainty about demands and prices.
- *R&D Portfolio Optimization.* The Department of Energy has to determine how to allocate government funds to advance the science of energy generation, transmission, and storage. These decisions have to be made over time in the presence of uncertain changes in technology and commodity prices.

These problems range from relatively low-dimensional applications to very high-dimensional industrial problems, but all share the property of making decisions over time under different types of uncertainty. In Ref. 1, a modeling framework is described, which breaks a problem into five components.

- *State.* S_t (or x_t in the control theory community), capturing all the information we need at time t to make a decision and model the evolution of the system in future.
- *Action/Decision/Control.* Depending on the community, these will be modeled as a , x , or u . Decisions are made using a decision function or policy. If we are using action a , we represent the decision function using $A^\pi(S_t)$, where $\pi \in \Pi$ is a family of policies (or functions). If our action is x , we use $X^\pi(S_t)$. We use a if our problem has a small number of discrete actions. We use x when the decision might be a vector (of discrete or continuous elements).
- *Exogenous Information.* Lacking standard notation, we let W_t be the family

of random variables that represent new information that first becomes known by time t .

- *Transition Function.* Also known as the *system model* (or just *model*), this function is denoted by $S^M(\cdot)$, and is used to express the evolution of the state variable, as in $S_{t+1} = S^M(S_t, x_t, W_{t+1})$.
- *Objective Function.* Let $C(S_t, x_t)$ be the contribution (if we are maximizing) received when in state S_t if we take action x_t . Our objective is to find a decision function (policy) that solves

$$\max_{\pi \in \Pi} \mathbb{E} \left\{ \sum_{t=0}^T \gamma^t C(S_t, X^\pi(S_t)) \right\}. \quad (1)$$

We encourage readers to review Ref. 1 (in this volume) or Chapter 5 in Ref. 2, available at <http://www.castlelab.princeton.edu/adp.htm>, before attempting to design an algorithmic strategy. For the remainder of this article, we assume that the reader is familiar with the modeling behind the objective function in Equation (1), and in particular the range of policies that can be used to provide solutions. In this article, we assume that we would like to find a good policy by using Bellman's equation as a starting point, which we may write in either of two ways:

$$\begin{aligned} V_t(S_t) &= \max_a (C(S_t, a) \\ &\quad + \gamma \sum_{s'} p(s'|S_t, a) V_{t+1}(s')), \quad (2) \\ &= \max_a (C(S_t, a) + \gamma \mathbb{E}\{V_{t+1}(S_{t+1})|S_t\}), \quad (3) \end{aligned}$$

where $V_t(S_t)$ is the value of being in state S_t and following an optimal policy from t until the end of the planning horizon (which may be infinite). The control theory community replaces $V_t(S_t)$ with $J_t(S_t)$, which is referred to as the cost-to-go function. If we are solving a problem in steady state, $V_t(S_t)$ would be replaced with $V(S)$.

If we have a small number of discrete states and actions, we can find the value function $V_t(S_t)$ using the classical techniques of value iteration and policy iteration [3]. Many

practical problems, however, suffer from one or more of the three curses of dimensionality: (i) vector-valued (and possibly continuous) state variables, (ii) random variables W_t for which we may not be able to compute the expectation, and (iii) decisions (typically denoted by x_t or u_t), which may be discrete or continuous vectors, requiring some sort of solver (linear, nonlinear, or integer programming) or specialized algorithmic strategy.

The field of ADP has historically focused on the problem of multidimensional state variables, which prevent us from calculating $V_t(s)$ for each discrete state s . In our presentation, we make an effort to cover this literature, but we also show how we can overcome the other two curses using a device known as the *postdecision state variable*. However, a short article such as this is simply unable to present the vast range of algorithmic strategies in any detail. For this, we recommend for additional reading, the following: Ref. 4, especially for students interested in obtaining thorough theoretical foundations; Ref. 5 for a presentation from the perspective of the reinforcement learning (RL) community; Ref. 2 for a presentation that puts more emphasis on modeling, and more from the perspective of the operations research community; and Chapter 6 of Ref. 6, which can be downloaded from <http://web.mit.edu/dimitrib/www/dpchapter.html>.

A GENERIC ADP ALGORITHM

Equation (2) (or Eq. 3) is typically solved by stepping backward in time, where it is necessary to loop over all the potential states to compute $V_t(S_t)$ for each state S_t . For this reason, classical dynamic programming is often referred to as *backward dynamic programming*. The requirement of looping over all the states is the first computational step that cannot be performed when the state variable is a vector, or even a scalar continuous variable.

ADP takes a very different approach. In most ADP algorithms, we step forward in time following a single sample path. Assume we are modeling a finite horizon problem. We are going to iteratively simulate this problem.

At iteration n , assume that at time t we are in state S_t^n . Now assume that we have a policy of some sort that produces a decision x_t^n . In ADP, we are typically solving a problem that can be written as

$$\begin{aligned} \hat{v}_t^n &= \max_{x_t} (C(S_t^n, x_t) \\ &\quad + \gamma \mathbb{E}[\bar{V}_{t+1}^{n-1}(S^M(S_t^n, x_t, W_{t+1})) | S_t]). \end{aligned} \quad (4)$$

Here, $\bar{V}_{t+1}^{n-1}(S_{t+1})$ is an approximation of the value of being in state $S_{t+1} = S^M(S_t^n, x_t, W_{t+1})$ at time $t+1$. If we are modeling a problem in steady state, we drop the subscript t everywhere, recognizing that W is a random variable. We note that x_t here can be a vector, where the maximization problem might require the use of a linear, nonlinear, or integer programming package. Let x_t^n be the value of x_t that solves this optimization problem. We can use $\hat{v}_t^n$ to update our approximation of the value function. For example, if we are using a lookup table representation, we might use

$$\bar{V}_t^n(S_t^n) = (1 - \alpha_{n-1})\bar{V}_t^{n-1}(S_t^n) + \alpha_{n-1}\hat{v}_t^n, \quad (5)$$

where α_{n-1} is a step size between 0 and 1 (more on this later).

Now we are going to use Monte Carlo methods to sample our vector of random variables, which we denote by W_{t+1}^n , representing the new information that would have first been learned between time t and $t+1$. The next state would be given by

$$S_{t+1}^n = S^M(S_t^n, x_t^n, W_{t+1}^n).$$

The overall algorithm is given in Fig. 1. This is a very basic version of an ADP algorithm, one which would generally not work in practice. But it illustrates some of the basic elements of an ADP algorithm. First, it steps forward in time, using a randomly sampled set of outcomes, visiting one state at a time. Secondly, it makes decisions using some sort of statistical approximation of a value function, although it is unlikely that we would use a lookup table representation. Thirdly, we use information gathered as we step forward in time to update our value function approximation, almost always using some sort of recursive statistics.

Step 0. Initialization:

Step 0a. Initialize $\bar{V}_t^0(S_t)$ for all states S_t .

Step 0b. Choose an initial state S_0^1 .

Step 0c. Set $n = 1$.

Step 1. Choose a sample path ω^n .

Step 2. For $t = 0, 1, 2, \dots, T$ do:

Step 2a. Solve

$$\hat{v}_t^n = \max_{x_t \in \mathcal{X}_t^n} \left(C_t(S_t^n, x_t) + \gamma \mathbb{E} \bar{V}_{t+1}^{n-1}(S^M(S_t^n, x_t, W_{t+1}^n)) \right)$$

and let x_t^n be the value of x_t that solves the maximization problem.

Step 2b. Update $\bar{V}_t^{n-1}(S_t)$ using

$$\bar{V}_t^n(S_t) = \begin{cases} (1 - \alpha_{n-1})\bar{V}_t^{n-1}(S_t^n) + \alpha_{n-1}\hat{v}_t^n & S_t = S_t^n \\ \bar{V}_t^{n-1}(S_t) & \text{otherwise.} \end{cases}$$

Step 2c. Compute the next state to visit. This may be chosen at random (according to some distribution), or from

$$S_{t+1}^n = S^M(S_t^n, x_t^n, W_{t+1}^n).$$

Step 3. Let $n = n + 1$. If $n < N$, go to step 1.

Figure 1. An approximate dynamic programming algorithm using expectations.

The algorithm in Fig. 1 goes under different names. It is a basic “forward pass” algorithm, where we step forward in time, updating value functions as we progress. Another variation involves simulating forward through the horizon without updating the value function. Then, after the simulation is done, we step backward through time, using information about the entire future trajectory.

We have written the algorithm in the context of a finite horizon problem. For infinite horizon problems, simply drop the subscript t everywhere, remembering that you have to keep track of the “current” and “future” state.

There are different ways of presenting this basic algorithm. To illustrate, we can rewrite Equation (5) as

$$\bar{V}^n(S_t^n) = \bar{V}^{n-1}(S_t^n) + \alpha_{n-1}(\hat{v}_t^n - \bar{V}^{n-1}(S_t^n)). \quad (6)$$

The quantity

$$\begin{aligned} \hat{v}_t^n - \bar{V}^{n-1}(S_t^n) &= C_t(S_t^n, x_t) + \gamma \bar{V}_{t+1}^{n-1} \\ &\quad \times (S_{t+1}) - \bar{V}^{n-1}(S_t^n) \end{aligned}$$

where

$$S_{t+1} = S^M(S_t, x_t, W_{t+1})$$

is known as the *Bellman error*, since it is a sampled observation of the difference between what we think is the value of being in state S_t^n and an estimate of the value of being in state S_t^n . In the RL community, it has long been called the *temporal difference (TD)*, since it is the difference between estimates at two different iterations, which can have the interpretation of time, especially for steady-state problems, where n indexes transitions forward in time.

The RL community would refer to this algorithm as TD learning (first introduced in Ref. 7). If we use a backward pass, we can let $\hat{v}_t^n$ be the value of the entire future trajectory, but it is often useful to introduce a discount factor (usually represented as λ) to discount rewards received in the future. This discount factor is introduced purely for algorithmic reasons, and should not be confused with the discount γ , which is intended

to capture the time value of money. If we let $\lambda = 1$, then we are adding up the entire future trajectory. But if we use $\lambda = 0$, then we obtain the same updating as our forward pass algorithm in Fig. 1. For this reason, the RL community refers to this family of updating strategies as $TD(\lambda)$.

$TD(0)$ (TD learning with $\lambda = 0$) can be viewed as an approximate version of classical value iteration. It is important to recognize that after each update, we are not just changing the value function approximation, we are also changing our behavior (i.e., our policy for making decisions), which in turn changes the distribution of $\hat{v}_t^n$ given a particular state S_t^n . $TD(1)$, on the other hand, computes $\hat{v}_t^n$ by simulating a policy into the future. This policy depends on the value function approximation, but otherwise $\hat{v}_t^n$ does not directly use value function approximations.

An important dimension of most ADP algorithms is Step 2c, where we have to choose which state to visit next. For many complex problems, it is most natural to choose the state determined by the action x_t^n , and then simulate our way to the next state using $S_{t+1}^n = S^M(S_t^n, x_t^n, W_{t+1}(\omega^n))$. Many authors refer to this as a form of “real-time dynamic programming” (RTDP) [8], although a better term is trajectory following [9], since RTDP involves other algorithmic assumptions. The alternative to trajectory following is to randomly generate a state to visit next. Trajectory following often seems more natural since it means you are simulating a system, and it visits the states that arise naturally, rather than from what might be an unrealistic probability model. But readers need to understand that there are few convergence results for trajectory following models, due to the complex interaction between random observations of the value of being in a state and the policy that results from a specific value function approximation, which impacts the probability of visiting a state.

The power and flexibility of ADP has to be tempered with the reality that simple algorithms generally do not work, even on simple problems. There are several issues we have to address if we want to develop an effective ADP strategy:

- We avoid the problem of looping over all the states, but replace it with the problem of developing a good statistical approximation of the value of being in each state that we might visit. The challenge of designing a good value function approximation is at the heart of most ADP algorithms.
- We have to determine how to update our approximation using new information.
- We assume that we can compute the expectation, which is often not the case, and which especially causes problems when our decision/action variable is a vector.
- We can easily get caught in a circle of choosing actions that take us to states that we have already visited, simply because we may have pessimistic estimates of states that we have not yet visited. We need some mechanism to force us to visit states just to learn the value of these states.

The remainder of this article is a brief tour through a rich set of algorithmic strategies for solving these problems.

Q-LEARNING AND THE POSTDECISION STATE VARIABLE

In our classical ADP strategy, we are solving optimization problems of the form

$$\begin{aligned} \hat{v}_t^n = \max_{x_t \in \mathcal{X}_t^n} & (C(S_t^n, x_t) \\ & + \gamma \mathbb{E}[\bar{V}_{t+1}(S^M(S_t^n, x_t, W_{t+1})) | S_t]). \end{aligned} \quad (7)$$

There are many applications of ADP, which introduce additional complications: (i) we may not be able to compute the expectation, and (ii) the decision x may be a vector, and possibly a very big vector (thousands of dimensions). There are two related strategies that have evolved to address these issues: *Q*-learning, a technique widely used in the RL community, and ADP using the post-decision state variable, a method that has evolved primarily in the operations research community. These methods are actually closely related, a topic we revisit at the end.

Q-Factors

To introduce the idea of *Q*-factors, we need to return to the notation where a is a discrete action (with a small action space). The *Q* factor (so named because this was the original notation used in Ref. 10) is the value of being in a state *and* taking a specific action,

$$Q(s, a) = C(s, a) + \gamma \mathbb{E} \bar{V}(S^M(S^n, a, W)).$$

Value functions and *Q*-factors are related using

$$V(s) = \max_a Q(s, a),$$

so, at first glance it might seem as if we are not actually gaining anything. In fact, estimating a *Q*-factor actually seems harder than estimating the value of being in a state, since there are more state-action pairs than there are states.

The power of *Q*-factors arises when we have problems where we either do not have an explicit transition function $S^M(\cdot)$, or do not know the probability distribution of W . This means we cannot compute the expectation, but not because it is computationally difficult. When we do not know the transition function, and/or cannot compute the expectation (because we do not know the probability distribution of the random variable), we would like to draw on a subfield of ADP known as *model-free* dynamic programming. In this setting, we are in state S^n , we choose an action a^n , and then observe the next state, which we will call s' . We can then record a value of being in state s and taking action a using

$$\hat{q}^n = C(S^n, a^n) + \bar{V}^{n-1}(s')$$

where $\bar{V}(s') = \max_{a'} Q^{n-1}(s', a')$. We then use this value to update our estimate of the *Q*-factor using

$$Q^n(S^n, a^n) = (1 - \alpha) Q^{n-1}(S^n, a^n) + \alpha \hat{q}^n.$$

Whenever we are in state S^n , we choose our action using

$$a^n = \arg \max_a Q^{n-1}(S^n, a). \quad (8)$$

Note that we now have a method that does not require a transition function or the need to compute an expectation, we only need access to some exogenous process that tells us what state we transition to, given a starting state and an action.

The Postdecision State Variable

A closely related strategy is to break the transition function into two steps: the pure effect of the decision x_t , and the effect of the new, random information W_{t+1} . In this section, we use the notation of operations research, because what we are going to do is enable the solution of problems where x_t may be a very large vector. We also return to time-dependent notation since it clarifies when we are measuring a particular variable.

We can illustrate this idea using a simple inventory example. Let S_t be the amount of inventory on hand at time t . Let x_t be an order for a new product, which we assume arrives immediately. The typical inventory equation would be written as

$$S_{t+1} = \max\{0, S_t + x_t - \hat{D}_{t+1}\}$$

where $\hat{D}_{t+1}$ is the random demand that arises between t and $t + 1$. We denote the pure effect of this order using

$$S_t^x = S_t + x_t.$$

The state S_t^x is referred to as the *postdecision state* [11], which is the state at time t , immediately after a decision has been made, but before any new information has arrived. The transition from S_t^x to S_{t+1} in this example would then be given by

$$S_{t+1} = \max\{0, S_t^x - \hat{D}_{t+1}\}.$$

Another example of a postdecision state is a tic-tac-toe board after you have made your move, but before your opponent has moved. Reference 5 refers to this as the after-state variable, but we prefer the notation that emphasizes the information content (which is why S_t^x is indexed by t).

A different way of representing a postdecision state is to assume that we have a forecast $\bar{W}_{t,t+1} = \mathbb{E}\{W_{t+1}|S_t\}$, which is a point

estimate of what we think W_{t+1} will be given what we know at time t . A postdecision state can be thought of as a forecast of the state S_{t+1} given what we know at time t , which is to say

$$S_t^x = S^M(S_t, x_t, \bar{W}_{t,t+1}).$$

This approach will seem more natural in certain applications, but it would not be useful when random variables are discrete. Thus, it makes sense to use the expected demand, but not the expected action of your tic-tac-toe opponent.

In a typical decision tree, S_t would represent the information available at a decision node, while S_t^x would be the information available at an outcome node. Using these two steps, Bellman's equations become

$$V_t(S_t) = \max_{x_t} (C(S_t, x_t) + \gamma V_t^x(S_t^x)), \quad (9)$$

$$V_t^x(S_t^x) = \mathbb{E}\{V_{t+1}(S_{t+1})|S_t\}. \quad (10)$$

Note that Equation (9) is now a deterministic optimization problem, while Equation (10) involves only an expectation. Of course, we still do not know $V_t^x(S_t^x)$, but we can approximate it just as we would approximate $V_t(S_t)$. However, now we have to pay attention to the nature of the optimization problem. If we are just searching over a small number of discrete actions, we do not really have to worry about the structure of our approximate value function around S_t^x . But if x is a vector, which might have hundreds or thousands of discrete or continuous variables, then we have to recognize that we are going to need some sort of solver to handle the maximization problem.

To illustrate the basic idea, assume that we have discrete states and are using a lookup table representation for a value function. We would solve Equation (9) to obtain a decision x_t^n , and we would then simulate our way to S_{t+1}^n using our transition function. Let $\hat{v}_t^n$ be the objective function returned by solving Equation (9) at time t , when we are in state S_t^n . We would then use this value to update the value at the *previous postdecision state*. That is,

$$\bar{V}_{t-1}^n(S_{t-1}^n) = (1 - \alpha_{n-1})\bar{V}_{t-1}^{n-1}(S_{t-1}^n) + \alpha_{n-1}\hat{v}_t^n. \quad (11)$$

Thus, the updating is virtually the same as we did before; it is just that we are now using our estimate of the value of being in a state to update the value of the previous postdecision state. This is a small change, but it has a huge impact by eliminating the expectation in the decision problem.

Now imagine that we are managing different types of products, where S_{ti} is the inventory of product of type i , and S_{ti}^x is the inventory after placing a new order. We might propose a value function approximation that looks like

$$\bar{V}_t^x(S_t^x) = \sum_i \left(\theta_{1i} S_{ti}^x - \theta_{2i} (S_{ti}^x)^2 \right).$$

This value function is separable and concave in the inventory (assuming $\theta_{2i} > 0$). Using this approximation, we can probably solve our maximization problem using a nonlinear programming algorithm, which means we can handle problems with hundreds or thousands of products. Critical to this step is that the objective function in Equation (9) is deterministic.

Comments

On first glance, Q -factors and ADP using the postdecision state variable seem to be very different methods to address very different problems. However, there is a mathematical commonality to these two algorithmic strategies. First, we note that if S is our current state, (S, a) is a kind of postdecision state, in that it is a deterministic function of the state and action. Second, if we take a close look at Equation (9), we can write

$$Q(S_t, x_t) = C(S_t, x_t) + \gamma V_t^x(S_t^x).$$

We assume that we have a known function S_t^x that depends only on S_t and x_t , but we do not require the full transition function (which takes us to time $t + 1$), and we do not have to take an expectation. However, rather than develop an approximation of the value of a state and an action, we only require the value of a (postdecision) state. The task of finding the best action from a set of Q -factors (Eq. 8) is identical to the maximization problem in Equation (9). What has been overlooked in the ADP/RL community is that this

step makes it possible to search over large, multidimensional (and potentially continuous) action spaces. Multidimensional action spaces are a relatively overlooked problem class in the ADP/RL communities.

APPROXIMATE POLICY ITERATION

An important variant of ADP is approximate policy iteration, which is illustrated in Fig. 2 using the postdecision state variables. In this strategy, we simulate a policy, say, M times over the planning horizon (or M steps into the future). During these inner iterations, we fix the policy (i.e., we fix the value function approximation used to make a decision) to obtain a better estimate of the value of being in a state. The general process of updating the value function approximation is handled using the update function $U^V(\cdot)$. As $M \rightarrow \infty$, we obtain an exact estimate of the value of being in a particular state, while following a fixed policy (see Ref. 12 for the convergence theory of this technique).

If we randomize on the starting state and repeat this simulation for an infinite number of times, we can obtain the best possible value function given a particular approximation. It is possible to prove convergence of classical stochastic approximation methods when the policy is fixed. Unfortunately, there are very few convergence results when we combine sampled observations of the value of being in a state with changes in the policy.

If we are using low-dimensional representations of the value function (as occurs with basis functions), it is important that our update be of sufficient accuracy. If we use $M = 1$, which means we are updating after a single forward traversal, the result may mean that the value function approximation (and therefore the policy) is changing too rapidly from one iteration to another. The resulting algorithm may actually diverge [13,14].

TYPES OF VALUE FUNCTION APPROXIMATIONS

At the heart of approximate dynamic programming is approximating the value of

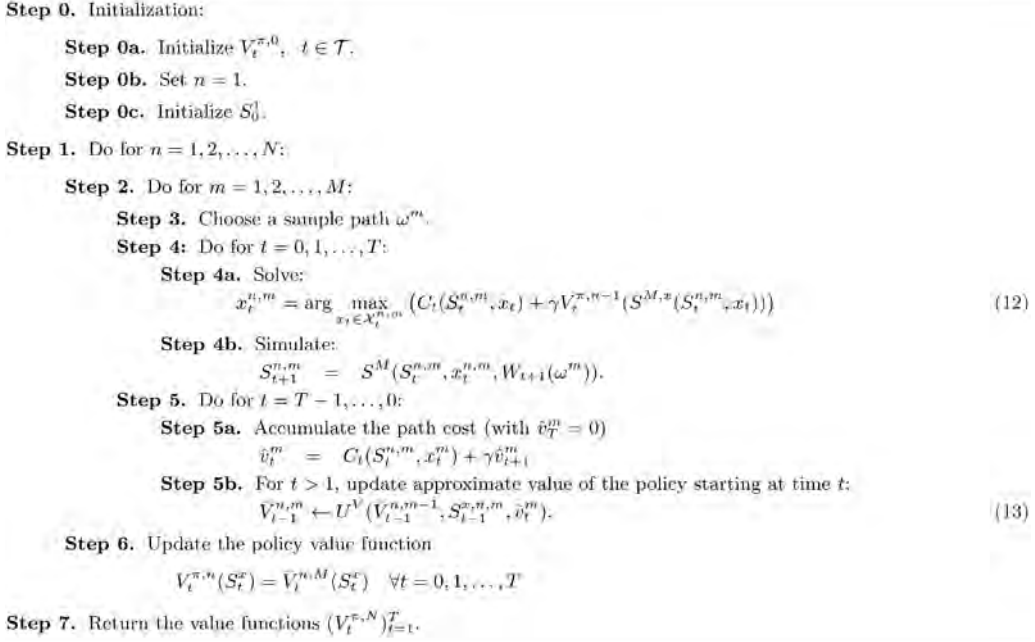


Figure 2. Approximate policy iteration using value function-based policies.

being in a state. It is important to emphasize that the problem of designing and estimating a value function approximation draws on the entire field of statistics and machine learning. Below, we discuss two very general approximation methods. You will generally find that ADP is opening a doorway into the field of machine learning, and that this is the place where you will spend most of your time. You should pick up a good reference book such as Ref. 15 on statistical learning methods, and keep an open mind regarding the best method for approximating a value function (or the policy directly).

Lookup Tables with Aggregation

The simplest form of statistical learning uses the lookup table representation that we first illustrated in Equation (5). The problem with this method is that it does not apply to continuous states, and requires an exponentially large number of parameters (one per state) when we encounter vector-valued states.

A popular way to overcome large state spaces is through aggregation. Let $\mathcal{S}$ be the

set of states. We can represent aggregation using

$$G^g : \mathcal{S} \rightarrow \mathcal{S}^{(g)}.$$

$\mathcal{S}^{(g)}$ represents the g^{th} level of aggregation of the state space $\mathcal{S}$, where we assume that $\mathcal{S}^{(0)} = \mathcal{S}$. Let

- $s^{(g)} = G^g(s)$, the g^{th} level aggregation of state s .
- $\mathcal{G}$ = The set of indices corresponding to the levels of aggregation.

We then define $\bar{V}^g(s)$ to be the value function approximation for state $s \in \mathcal{S}^{(g)}$.

We assume that the family of aggregation functions $G^g, g \in \mathcal{G}$ is given. A common strategy is to pick a single level of aggregation that seems to work well, but this introduces several complications. First, the best level of aggregation changes as we acquire more data, creating the challenge of working out how to transition from approximating the value function in the early iterations versus later iterations. Second, the right level of

aggregation depends on how often we visit a region of the state space.

A more natural way is to use a weighted average. At iteration n , we can approximate the value of being in state s using

$$\bar{v}^n(s) = \sum_{g \in \mathcal{G}} w^{(g,n)}(s) \bar{v}^{(g,n)}(s).$$

where $w^{(g,n)}(s)$ is the weight given to the g^{th} level of aggregation when measuring state s . This weight is set to zero, if there are no observations of states $s \in \mathcal{S}^{(g)}$. Thus, while there may be an extremely large number of states, the number of positive weights at the most disaggregate level $w^{(0,n)}$ is never greater than the number of observations. Typically, the number of positive weights at more aggregate levels will be substantially smaller than this.

There are different ways to determine these weights, but an effective method is to use weights that are inversely proportional to the total squared variation of each estimator. Let

$\bar{\sigma}^2(s)^{(g,n)}$	the variance of the estimate of the value of state s at the g^{th} level of aggregation after collecting n observations,
$\bar{\mu}^{(g,n)}(s)$	an estimate of the bias due to aggregation.

We can then use weights that satisfy

$$w^{(g)}(s) \propto \left((\bar{\sigma}^2(s))^{(g,n)} + (\bar{\mu}^{(g,n)}(s))^2 \right)^{-1}. \quad (14)$$

The weights are then normalized so that they sum to one. The quantities $\bar{\sigma}^2(s)^{(g,n)}$ and $\bar{\mu}^{(g,n)}(s)$ are computed using the following:

$$\begin{aligned} (\bar{\sigma}^2(s))^{(g,n)} &= \text{Var}[\bar{v}^{(g,n)}(s)] \\ &= \lambda^{(g,n)}(s) (s^2(s))^{(g,n)}, \end{aligned} \quad (15)$$

$$(s^2(s))^{(g,n)} = \frac{\bar{v}^{(g,n)}(s) - (\bar{\beta}^{(g,n)}(s))^2}{1 + \lambda^{n-1}}, \quad (16)$$

$$\lambda^{(g,n)}(s) = \begin{cases} (\alpha^{(g,n-1)}(s))^2, & n = 1, \\ (1 - \alpha^{(g,n-1)}(s))^2 \\ \quad \times \lambda^{(g,n-1)}(s) \\ \quad + (\alpha^{(g,n-1)}(s))^2, & n > 1, \end{cases} \quad (17)$$

$$\begin{aligned} \bar{v}^{(g,n)}(s) &= (1 - \alpha_{n-1}) \bar{v}^{(g,n-1)}(s) \\ &\quad + \alpha_{n-1} (\bar{v}^{(g,n-1)}(s) - \hat{v}^n(s))^2, \end{aligned} \quad (18)$$

$$\begin{aligned} \bar{\beta}^{(g,n)}(s) &= (1 - \alpha_{n-1}) \bar{\beta}^{(g,n-1)}(s) \\ &\quad + \alpha_{n-1} (\hat{v}^n - \bar{v}^{(g,n-1)}(s)), \end{aligned} \quad (19)$$

$$\bar{\mu}^{(g,n)}(s) = \bar{v}^{(g,n)}(s) - \bar{v}^{(0,n)}(s). \quad (20)$$

The term $\bar{v}^{(g,n)}(s)$ is the total squared variation in the estimate, which includes variation due to pure noise, $(s^2(s))^{(g,n)}$, and two sources of bias. The term $\bar{\beta}^{(g,n)}(s)$ is the bias introduced when smoothing a nonstationary data series (if the series is steadily rising, our estimates are biased downward), and $\bar{\mu}^{(g,n)}(s)$ is the bias due to aggregation error.

This system for computing weights scales easily to large state spaces. As we collect more observations of states in a particular region, the weights gravitate toward putting more weight on disaggregate estimates, while keeping higher weights on more aggregate estimates for regions that are only lightly sampled. For more on this strategy, see Ref. 16 or Chapter 7 in Ref. 2. For other important references on aggregation see Refs 4, 17, 18 and 19.

Basis Functions

Perhaps the most widely cited method for approximating value functions uses the concept of basis functions. Let

$\mathcal{F}$	A set of features drawn from a state vector,
$\phi_f(S)$	A scalar function that draws what is felt to be a useful piece of information from the state variable, for $f \in \mathcal{F}$.

The function $\phi_f(S)$ is referred to as a *basis function*, since in an ideal setting we would have a family of functions that spans the state space, allowing us to write

$$V(S|\theta) = \sum_{f \in \mathcal{F}} \theta_f \phi_f(S).$$

for an appropriately chosen vector of parameters θ . In practice, we cannot guarantee that we can find such a set of basis functions, so we write

$$\bar{V}(S|\theta) \approx \sum_{f \in \mathcal{F}} \theta_f \phi_f(S).$$

Basis functions are typically referred to as *independent variables* or *covariates* in the statistics community. If S is a scalar variable, we might write

$$\bar{V}(S|\theta) = \theta_0 + \theta_1 S + \theta_2 S^2 + \theta_3 \sin(S) + \theta_4 \ln(S).$$

This type of approximation is referred to as *linear*, since it is linear in the parameters. However, it is clear that we are capturing nonlinear relationships between the value function and the state variable.

There are several ways to estimate the parameter vector. The easiest, which is often referred to as *least squares temporal differencing* (LSTD), uses $\hat{v}_t^n$ as it is calculated in Equation (4), which means that it depends on the value function approximation. An alternative is least squares policy estimation (LSPE), where we use repeated samples of rewards from simulating a fixed policy. The basic LSPE algorithm estimates the parameter vector θ by solving

$$\theta^n = \arg \max_{\theta} \sum_{m=1}^n \left(\hat{v}_t^m - \sum_{f \in \mathcal{F}} \theta_f \phi_f(S_t^m) \right)^2.$$

This algorithm provides nice convergence guarantees, but the updating step becomes more and more expensive as the algorithm progresses. Also, it is putting equal weight on all iterations, a property that is not desirable in practice. We refer the reader to Ref. 6 for a more thorough discussion of LSTD and LSPE.

A simple alternative uses a stochastic gradient algorithm. Assuming we have an initial estimate θ^0 , we can update the estimate of theta at iteration n using

$$\begin{aligned} \theta^n &= \theta^{n-1} - \alpha_{n-1} (\bar{V}_t(S_t^n | \theta^{n-1}) - \hat{v}_t^n) \nabla_{\theta} \bar{V}_t \\ &\quad \times (S_t^n | \theta^{n-1}) \end{aligned}$$

$$\begin{aligned} &= \theta^{n-1} - \alpha_{n-1} (\bar{V}(S_t^n | \theta^{n-1}) - \hat{v}^n(S_t^n)) \\ &\quad \times \begin{pmatrix} \phi_1(S_t^n) \\ \phi_2(S_t^n) \\ \vdots \\ \phi_F(S_t^n) \end{pmatrix}. \end{aligned} \quad (21)$$

Here, α_{n-1} is a step size, but it has to perform a scaling function, so it is not necessarily less than 1. Depending on the problem, we might have a step size that starts around 10^6 or 0.00001.

A more powerful strategy uses recursive least squares. Let ϕ^n be the vector created by computing $\phi(S^n)$ for each feature $f \in \mathcal{F}$. The parameter vector θ can be updated using

$$\theta^n = \theta^{n-1} - H^n \phi^n \hat{\varepsilon}^n, \quad (22)$$

where ϕ^n is the vector of basis functions evaluated at S^n , and $\hat{\varepsilon}^n = \bar{V}(S_t^n | \theta^{n-1}) - \hat{v}^n(S_t^n)$ is the error in our prediction of the value function. The matrix H^n is computed using

$$H^n = \frac{1}{\gamma^n} B^{n-1}. \quad (23)$$

B^{n-1} is an $F+1$ by $F+1$ matrix (where $F = |\mathcal{F}|$), which is updated recursively using

$$B^n = B^{n-1} - \frac{1}{\gamma^n} (B^{n-1} \phi^n (\phi^n)^T B^{n-1}). \quad (24)$$

γ^n is a scalar computed using

$$\gamma^n = 1 + (\phi^n)^T B^{n-1} \phi^n. \quad (25)$$

Note that there is no step size in these equations. These equations avoid the scaling issues of the stochastic gradient update in Equation (21), but hide the fact that they are implicitly using a step size of $1/n$. This can be very effective, but can work very poorly. We can overcome this by introducing a factor λ and revising the updating of γ^n and B^n using

$$\gamma^n = \lambda + (x^n)^T B^{n-1} x^n, \quad (26)$$

and the updating formula for B^n , which is now given by

$$B^n = \frac{1}{\lambda} \left(B^{n-1} - \frac{1}{\gamma^n} (B^{n-1} x^n (x^n)^T B^{n-1}) \right).$$

This method puts a weight of λ^{n-m} on an observation from iteration $n - m$. If $\lambda = 1$, then we are weighting all observations equally. Smaller values of λ put a lower weight on earlier observations. See Chapter 7 of Ref. 2 for a more thorough discussion of this strategy. For a thorough presentation of recursive estimation methods for basis functions in ADP, we encourage readers to see Ref. 19.

Values versus Marginal Values

There are many problems, where it is more effective to use the derivative of the value function with respect to a state, rather than the value of being in the state. This is particularly powerful in the context of resource allocation problems, where we are solving a vector-valued decision problem using linear, nonlinear, or integer programming. In these problem classes, the derivative of the value function is much more important than the value function itself. This idea has long been recognized in the control theory community, which has used the term *heuristic dynamic programming* to refer to ADP using the value of being in a state, and “*dual heuristic*” dynamic programming when using the derivative (see Ref. 20 for a nice discussion of these topics from a control-theoretic perspective). Chapters 11 and 12 in Ref. 2 discuss the use of derivatives for applications that arise in operations research.

Discussion

Approximating value functions can be viewed as just an application of a vast array of statistical methods to estimate the value of being in a state. Perhaps the biggest challenge is model specification, which can include designing the basis functions. See Refs 14 and 21 for a thorough discussion of feature selection. There has also been considerable interest in the use of nonparametric methods [22]. An excellent reference for statistical learning techniques that can be used in this setting is Ref. 15.

The estimation of value function approximations, however, involves much more than just a simple application of statistical methods. There are several issues that have to be addressed that are unique to the ADP setting:

1. Value function approximations have to be estimated recursively. There are many statistical methods that depend on estimating a model from a fixed batch of observations. In ADP, we have to update value function approximations after each observation.
2. We have to get through the early iterations. In the first iteration, we have no observations. After 10 iterations, we have to work with value functions that have been approximated using 10 data points (even though we may have hundreds or thousands of parameters to estimate). We depend on the value function approximations in these early iterations to guide decisions. Poor decisions in the early iterations can lead to poor value function approximations.
3. The value $\hat{v}_t^n$ depends on the value function approximation $\bar{V}_{t+1}(S)$. Errors in this approximation produce biases in $\hat{v}_t^n$, which then distorts our ability to estimate $\bar{V}_t(S)$. As a result, we are recursively estimating value functions using nonstationary data.
4. We can choose which state to visit next. We do not have to use $S_{t+1}^n = S^M(S_t^n, x_t^n, W_{t+1}^n)$, which is a strategy known as *trajectory following*. This is known in the ADP community as the “*exploration vs. exploitation*” Problem—do we explore a state just to learn about it, or do we visit a state that appears to be the best? Although considerable attention has been given to this problem, it is a largely unresolved issue, especially for high-dimensional problems.
5. Everyone wants to avoid the art of specifying value function approximations. Here, the ADP community shares the broader goal of the statistical learning community, which is a method that will approximate a value function with no human intervention.

THE LINEAR PROGRAMMING METHOD

The best-known methods for solving dynamic programs (in steady state) are value iteration and policy iteration. Less well known is that we can solve for the value of being in each (discrete) state by solving the linear program

$$\min_v \sum_{s \in S} \beta_s v(s) \quad (27)$$

subject to

$$\begin{aligned} v(s) &\geq C(s, a) \\ &+ \gamma \sum_{s' \in S} p(s'|s, a) v(s') \text{ for all } s \text{ and } a, \end{aligned} \quad (28)$$

where β is simply a vector of positive coefficients. In this problem, the decision variable is the value $v(s)$ of being in state s , which has to satisfy the constraints (Eq. 26). The problem with this method is that there is a decision variable for each state, and a constraint for each state-action pair. Not surprisingly, this can produce very large linear programs very quickly.

The linear programming method received a new lease on life with the work of de Farias and Van Roy [23], who introduced ADP concepts to this method. The first step to reduce complexity involves introducing basis functions to simplify the representation of the value function, giving us

$$\min_{\theta} \sum_{s \in S} \beta_s \sum_{f \in \mathcal{F}} \theta_f \phi_f(s)$$

subject to

$$\begin{aligned} \sum_{f \in \mathcal{F}} \theta_f \phi_f(s) &\geq C(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) \\ &\sum_{f \in \mathcal{F}} \theta_f \phi_f(s') \text{ for all } s \text{ and } a. \end{aligned}$$

Now, we have reduced a problem with $|S|$ variables (one value per state) to a vector $\theta_f, f \in \mathcal{F}$. Much easier, but we still have the problem of too many constraints. de Farias and Van Roy [23] then introduced the novel idea of using a Monte Carlo sample of the constraints. As of this writing, this method

is receiving considerable attention in the research community, but as with all ADP algorithms, considerable work is needed to produce robust algorithms that work in an industrial setting.

STEP SIZES

While designing a good value function, approximations are the most central challenge in the design of an ADP algorithm; a critical and often overlooked choice is the design of a step size rule. This is not to say that step sizes have been ignored in the research literature, but it is frequently the case that developers do not realize the dangers arising from an incorrect choice of step size rule.

We can illustrate the basic challenge in the design of a step size rule by considering a very simple dynamic program. In fact, our dynamic program has only one state and no actions. We can write it as computing the following expectation

$$F = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t \hat{C}_t.$$

Assume that the contributions $\hat{C}_t$ are identically distributed as a random variable $\hat{C}$ and that $\mathbb{E}\hat{C} = c$. We know that the answer to this problem is $c/(1 - \gamma)$, but assume that we do not know c , and we have to depend on random observations of $\hat{C}$. We can solve this problem using the following two-step algorithm:

$$\hat{v}^n = \hat{C}^n + \gamma \bar{v}^{n-1}, \quad (29)$$

$$\bar{v}^n = (1 - \alpha_{n-1}) \bar{v}^{n-1} + \alpha_{n-1} \hat{v}^n. \quad (30)$$

We are using an iterative algorithm, where $\bar{v}^n$ is our estimate of F after n iterations. We assume that $\hat{C}^n$ is the n^{th} sample realization of $\hat{C}$.

We note that Equation (29) handles the problem of summing over multiple contributions. Since we depend on Monte Carlo observations of $\hat{C}$, we use Equation (30) to then perform smoothing. If $\hat{C}$ were deterministic, we would use a step size $\alpha_{n-1} = 1$ to achieve the fastest convergence. On the other hand,

if $\hat{C}$ is random, but if $\gamma = 0$ (which means we do not really have to deal with the infinite sum), then the best possible step size is $1/n$ (we are just trying to estimate the mean of $\hat{C}$). In fact, $1/n$ satisfies a common set of theoretical conditions for convergence which are typically written

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \\ \sum_{n=1}^{\infty} (\alpha_n)^2 < \infty.$$

The first of these conditions prevents stalling (a step size of 0.8^n would fail this condition), while the second ensures that the variance of the resulting estimate goes to zero (statistical convergence).

Not surprisingly, many step size formulas have been proposed in the literature. A review is provided in Ref. 24 (see also Chapter 6 in Ref. 2). Two popular step size rules that satisfy the conditions above are

$$\alpha_n = \frac{a}{a + n - 1}, \quad \text{and} \\ \alpha_n = \frac{1}{n^\beta}.$$

In the first formula, $a = 1$ gives you $1/n$, whereas larger values of a produce a step size that declines more slowly. In the second rule, β has to satisfy $0.5 < \beta \leq 1$. Of course, these rules can be combined, but both a and β have to be tuned, which can be particularly annoying.

Reference 24 introduces the following step size rule (called the *bias adjusted Kalman filter* (BAKF) step size rule):

$$\alpha_n = 1 - \frac{\sigma^2}{(1 + \lambda^n)\sigma^2 + (\beta^n)}, \quad (31)$$

where

$$\lambda^n = \begin{cases} (\alpha_{n-1})^2, & n = 1 \\ (1 - \alpha_{n-1})^2 \lambda^{n-1} + (\alpha_{n-1})^2, & n > 1. \end{cases}$$

Here, σ^2 is the variance of the observation noise, and β^n is the bias measuring the difference between the current estimate of the

value function $\bar{V}^n(S^n)$ and the true value function $V(S^n)$. Since these quantities are not generally known they have to be estimated from data (see Refs 2 and 24, Chapter 6 for details).

This step size enjoys some useful properties. If there is no noise, then $\alpha_{n-1} = 1$. If $\beta^n = 0$, then $\alpha_{n-1} = 1/n$. It can also be shown that $\alpha_{n-1} \geq 1/n$ at all times. The primary difficulty is that the bias term has to be estimated from noisy data, which can cause problems. If the level of observation noise is very high, we recommend a deterministic step size formula. But if the observation noise is not too high, the BAKF rule can work quite well, because it adapts to the data.

PRACTICAL ISSUES

ADP is a flexible modeling and algorithmic framework that requires that a developer make a number of choices in the design of an effective algorithm. Once you have developed a model, designed a value function approximation, and chosen a learning strategy, you still face additional steps before you can conclude that you have an effective strategy.

Usually the first challenge you will face is debugging a model that does not seem to be working. For example, as the algorithm learns, the solution will tend to get better, although it is hardly guaranteed to improve from one iteration to the next. But what if the algorithm does not seem to produce any improvement at all?

Start by making sure that you are solving the decision problem correctly, given the value function approximation (even if it may be incorrect). This is really only an issue for more complex problems, such as where x is an integer-valued vector, requiring the use of more advanced algorithmic tools.

Then, most importantly, verify the accuracy of the information (such as $\hat{v}$) that you are deriving to update the value function. Are you using an estimate of the value of being in a state, or perhaps the marginal value of an additional resource? Is this value correct? If you are using a forward pass algorithm (where $\hat{v}$ depends on a value function approximation of the future), assume the

approximation is correct and validate $\hat{v}$. This is more difficult if you are using a backward pass (as with TD(1)), since $\hat{v}$ has to track the value (or marginal value) of an entire trajectory.

When you have confidence that $\hat{v}$ is correct, are you updating the value function approximation correctly? For example, if you compare $\hat{v}_t^n$ with $\bar{V}^{n-1}(S_t^n)$, the observations of $\hat{v}^n$ (over iterations) should look like the scattered noise of observations around an estimated mean. Do you see a persistent bias?

Finally, think about whether your value function approximations are really contributing better decisions. What if you set the value function to zero? What intelligence do you feel your value function approximations are really adding? Keep in mind that there are some problems where a myopic policy can work reasonably well.

Now, assume that you are convinced that your value function approximations are improving your solution. One of the most difficult challenges is evaluating the quality of an ADP solution. Perhaps you feel you are getting a good solution, but how far from optimal are you?

There is no single answer to the question of how to evaluate an ADP algorithm. Typically, you want to ask the question: If I simplify my problem in some way (without making it trivial), do I obtain an interesting problem that I can solve optimally? If so, you should be able to apply your ADP algorithm to this simplified problem, producing a benchmark against which you can compare.

How a problem might be simplified is situation specific. One strategy is to eliminate uncertainty. If the deterministic problem is still interesting (for example, this might be the case in a transportation application, but not a financial application), then this can be a powerful benchmark. Alternatively, it might be possible to reduce the problem to one that can be solved exactly as a discrete Markov decision process.

If these approaches do not produce anything, the only remaining alternative is to compare an ADP-based strategy against a policy that is being used (or most likely to be used) in practice. If it is not the best policy, is it at least an improvement? Of course, there

are many settings, where ADP is being used to develop a model to perform policy studies. In such situations, it is typically more important that the model behave reasonably. For example, do the results change in response to changes in inputs as expected?

ADP is more like a toolbox than a recipe. A successful model requires some creativity and perseverance. There are many instances of people who try ADP, only to conclude that it “does not work.” It is very easy to miss one ingredient to produce a model that does not work. But a successful model not only solves a problem, but it also enhances your ability to understand and solve complex decision problems.

REFERENCES

1. Powell WB. Approximate dynamic programming — I: Modeling. Encyclopedia for operations research and management science. New York: John Wiley & Sons, Inc.; 2010.
2. Powell WB. Approximate dynamic programming: solving the curses of dimensionality. New York: John Wiley & Sons, Inc.; 2007.
3. Puterman ML. Markov decision processes. New York: John Wiley & Sons, Inc.; 1994.
4. Bertsekas D, Tsitsiklis J. Neuro-dynamic programming. Belmont (MA): Athena Scientific; 1996.
5. Sutton R, Barto A. Reinforcement learning. Cambridge (MA): The MIT Press; 1998.
6. Bertsekas D. Volume II, Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2007.
7. Sutton R. Learning to predict by the methods of temporal differences. *Mach Learn* 1988;3(1):9–44.
8. Barto AG, Bradtke SJ, Singh SP. Learning to act using real-time dynamic programming. *Artif Intell Spec Vol Comput Res Interact Agency* 1995;72:81–138.
9. Sutton R. On the virtues of linear learning and trajectory distributions. In: *Proceedings of the Workshop on Value Function Approximation, Machine Learning Conference*; 1995. pp. 95–206.
10. Watkins C. Learning from delayed rewards [PhD thesis]. Cambridge: Cambridge University; 1989.
11. Van Roy B, Bertsekas DP, Lee Y, *et al.* A neuro-dynamic programming approach to

- retailer inventory management. In: Proceedings of the IEEE Conference on Decision and Control, Volume 4, 1997. pp. 4052–4057.
12. Tsitsiklis J, Van Roy B. An analysis of temporal-difference learning with function approximation. *IEEE Trans Autom Control* 1997;42:674–690.
 13. Bell DE. Risk, return, and utility. *Manage Sci* 1995;41:23–30.
 14. Tsitsiklis JN, Van Roy B. Feature-based methods for large-scale dynamic programming. *Mach Learn* 1996;22:59–94.
 15. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, Springer series in Statistics. New York: Springer; 2001.
 16. George A, Powell WB, Kulkarni S. Value function approximation using multiple aggregation for multiattribute resource management. *J Mach Learn Res* 2008;2079–2111.
 17. Bertsekas D, Castanon D. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE Trans Automat Control* 1989;34(6):589–598.
 18. Luus R. Iterative dynamic programming. New York: Chapman & Hall/CRC; 2000.
 19. Choi DP, Van Roy B. A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning. *Discrete Event Dyn Syst* 2006;16:207–239.
 20. Ferrari S, Stengel RF. Model-based adaptive critic designs. In: Si J, Barto AG, Powell WB, *et al.*, editors. Handbook of learning and approximate dynamic programming. New York: IEEE Press; 2004. pp. 64–94.
 21. Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In: Sanz-Solé M, Soria J, Varona JL, *et al.* editors. Volume III, Proceedings of international congress of mathematicians. Zurich: European Mathematical Society Publishing House; 2006. pp. 595–622.
 22. Fan J, Gijbels I. Local polynomial modeling and its applications. London: Chapman and Hall; 1996.
 23. de Farias D, Van Roy B. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math Oper Res* 2004;29(3):462–478.
 24. George A, Powell WB. Adaptive step sizes for recursive estimation with applications in approximate dynamic programming. *Mach Learn* 2006;65(1):167–198.

APPROXIMATION ALGORITHMS FOR STOCHASTIC OPTIMIZATION PROBLEMS IN OPERATIONS MANAGEMENT

CONG SHI
University of Michigan,
Ann Arbor, MI, USA

INTRODUCTION

The difficulty of sifting through large amount of data in order to make an informed choice is ubiquitous nowadays, thanks to the advances in information technologies and high speed networking. One of the promises of the information technology era is that many decisions can now be made rapidly by computers, such as deciding inventory levels, routing vehicles, planning facility locations, managing revenue, and so on. Many of these applications can be modeled as discrete optimization problems. Unfortunately, most interesting discrete optimization problems and their stochastic variants are NP-hard. Thus, we cannot simultaneously have algorithms (i) that find optimal solution (ii) in polynomial time, and (iii) for any instance. In order to deal with such optimization problems, we need to relax at least one of the three requirements: relaxing the “for any instance” requirement, the requirement of polynomial-time solvability, or the requirement of finding an optimal solution. The third relaxation is the most common approach, where we only need to find a “good-enough” solution instead of the best one. Next, we formally define the notion of *approximation algorithms* for discrete optimization problems.

Definition 1 An α -approximation algorithm for an optimization problem is a polynomial-time algorithm that for all instances of the problem produces a solution whose value is within a factor of α of the value of an optimal solution.

For an α -approximation algorithm, we always call α the *performance guarantee*, *approximation ratio*, or *approximation factor* of the algorithm. We will follow the convention that $\alpha > 1$ for minimization problems, while $\alpha < 1$ for maximization problems.

For a particular class of problems of interest (e.g., knapsack problem and Euclidean traveling salesman problem), we are able to obtain extremely good approximation algorithms; in fact, these problems have *polynomial-time approximation schemes* (PTASs).

Definition 2 A PTAS is a family of algorithms $\{P(\epsilon)\}$, where there is an algorithm for each $\epsilon > 0$, such that $P(\epsilon)$ is a $(1 + \epsilon)$ -approximation algorithm for minimization problems or a $(1 - \epsilon)$ -approximation algorithm for maximization problems.

However, there exists a large class of interesting (but not so easy) problems called MAX SNP (e.g., max satisfiability problem and max cut problem), which fails to have a PTAS, unless $P = NP$.

There is a vast body of literature in both computer science and operations research devoted to designing approximation algorithms for deterministic discrete optimization problems. For an excellent and a detailed exposition on deterministic models, we refer interested readers to the following books: Ausiello *et al.* [1], Vazirani [2], and Williamson and Shmoys [3]. This article mainly focuses on approximation algorithms for stochastic optimization models arising in operations management, which have gained much momentum recently. Before giving an overview or a survey on the recent development, we would like to first provide readers some basic ideas of discrete optimization models *under uncertainty* through the following example.

A Stochastic Vertex Cover Problem

We first use a stochastic version of the vertex cover problem to motivate our discus-

sion of how approximation algorithms can be designed for stochastic optimization problems. We review the 2-approximation algorithm and its worst-case analysis by Ravi and Sinha [4].

A vertex cover of a graph is a set of vertices such that each edge of the graph is incident to at least one vertex of the set. Formally, a vertex cover of an undirected graph $G = (V, E)$ (vertices and edges) is a subset S of V such that, if an edge $(u, v) \in E$, then either $u \in S$ or $v \in S$ (or both). The set S is said to *cover* the edges of G . As picking each vertex $i \in V$ incurs a cost c_i , a *minimum vertex cover* is a vertex cover of smallest possible cost. A simple deterministic LP rounding algorithm yields a 2-approximation. The best-known approximation algorithm has performance ratio of $2 - \log \log |V| / (2 \log |V|)$, because of Monien and Speckenmeyer [5]. A lower bound of 1.16 on the hardness of approximating the problem was shown by Håstad [6]. We refer readers to the studies by Williamson and Shmoys [3] and Vazirani [2] for the extensive discussion of the deterministic vertex cover problem.

Now, we describe a two-stage stochastic version of the vertex cover problem. In the first stage, we are given a (undirected) graph $G = (V, E_0)$. In the stochastic version, the edges E that will be present in the second stage are uncertain *a priori*, and we only know its distributional information, that is, there are m possible scenarios, each consisting of a set of *realized* edges E_k with probability of occurrence p_k . Note that E_k ($k = 1, \dots, m$) may or may not be subsets of the first-stage edge set E_0 . The first-stage cost of vertex v is c_v^0 , and its cost in scenario k at the second-stage is c_v^k . The edges in $E_k \cap E_0$ may be covered in either the first or second stage, while edges in E_k/E_0 must be covered in the second stage. The objective is to identify a set of vertices to be selected in the first stage, so that the expected cost of extending this set to a vertex cover of the edges of the realized second-stage scenario is minimized.

Ravi and Sinha [4] provided a primal-dual algorithm that rounds the integer programming (IP) formulation of stochastic vertex cover. Variable x_v^k indicates whether or not vertex v is picked in scenario k (where $k = 0$

corresponds to the first stage). The IP formulation is given by

$$\begin{aligned} \min \quad & \sum_v c_v^0 x_v^0 + \sum_{k=1}^m p_k c_v^k x_v^k, \quad (\text{IP-SVC}) \\ \text{s.t.} \quad & x_u^0 + x_v^0 + x_u^k + x_v^k \geq 1, \quad \forall (u, v) \in E_k \cap E_0, \forall k, \\ & x_u^k + x_v^k \geq 1, \quad \forall (u, v) \in E_k/E_0, \forall k, \\ & x \in \mathbb{Z}_0^+. \end{aligned}$$

The LP relaxation of IP-SVC, called *LP-SVC*, replaces the constraint $x \in \mathbb{Z}_0^+$ by $x \geq 0$. Then, we write the dual of LP-SVC further. The dual variable y_e^k packs edge e in E_k if $e \in E_k$, and it packs e in E_0 if $e \in E_k \cap E_0$.

$$\begin{aligned} \max \quad & \sum_{k=1}^m \sum_{e \in E_0 \cup E_k} y_e^k, \quad (\text{DLP-SVC}) \\ \text{s.t.} \quad & \sum_{e \in E_k: v \in e} y_e^k \leq p_k c_v^k, \quad \forall v, \forall k, \\ & \sum_{k=1}^m \sum_{e \in E_0 \cap E_k: v \in e} y_e^k \leq c_v^0, \quad \forall v, \\ & y \geq 0. \end{aligned}$$

The algorithm π is a greedy dual-ascent type of primal-dual algorithm with two phases.

- (a) In the first phase, we raise the dual variable y_e^k uniformly for all edges in E_k/E_0 , separately for each k . All vertices that become tight (the first dual constraint packed to $p_k c_v^k$) have x_v^k set to 1 and deleted along with adjacent edges. We proceed this way until all edges in E_k/E_0 are covered and deleted.
- (b) In the second phase, we raise the dual variable y_e^k uniformly for all *uncovered* edges in E_k . Note that these uncovered edges are contained in $E_k \cap E_0$. If a vertex is tight for x_v^0 (i.e., second dual constraint packed to c_v^0), then we pick it in the first stage solution by setting $x_v^0 = 1$, and if it is not tight for x_v^0 (the first dual constraint packed to $p_k c_v^k$), then we pick it in the second stage by setting $x_v^k = 1$ as a recourse decision.

Theorem 1 (Ravi and Sinha [4]) *The integer program IP-SVC can be bounded by the primal-dual algorithm described earlier within a factor of 2 in polynomial time.*

Proof. Let the cost of DLP-SVC generated by π be C^π (DLP-SVC). In addition, let the optimal costs of LP-SVC and IP-SVC be C^* (LP-SVC) and C^* (IP-SVC), respectively. By linear programming (LP) duality,

$$C^\pi(\text{DLP-SVC}) \leq C^*(\text{LP-SVC}) \leq C^*(\text{IP-SVC}).$$

Now if π generates a feasible solution for IP-SVC and a cost C^π (IP-SVC), then it suffices to show that

$$\begin{aligned} C^\pi(\text{IP-SVC}) &\leq 2 \cdot C^\pi(\text{DLP-SVC}) \\ &= 2 \sum_{k=1}^m \sum_{e \in E_0 \cup E_k} y_e^k. \end{aligned}$$

The feasibility is rather obvious. Consider an edge $e = (u, v) \in E_k$ in scenario k . We must have picked one of its endpoints in either the first phase or the second phase (or both) by the construction of algorithm π .

To complete the proof, we shall show that, each time we set $x_v^k = 1$, we assign some dual variables to it such that (i) the sum of dual variables assigned to each such x_v^k variable has to equal $p_k c_v^k$ (where $p_0 = 1$) and (ii) each dual variable or each edge is assigned at most twice.

Consider a vertex v that was selected in scenario k in either the first or second phase. We assign all dual variables y_e^k such that v is incident to e . By construction of π , v is only chosen when the first constraint becomes tight, that is,

$$\sum_{e \in E_k: v \in e} y_e^k = p_k c_v^k, \quad \forall k, \quad (1)$$

thereby guaranteeing (i). An edge $e \in E_k/E_0$ is assigned to vertex v only if x_v^k is set to 1 for $k \neq 0$. Thus, (ii) for $e \in E_k/E_0$ is ensured because each edge has at most two vertices.

Then, we consider a vertex v for which x_v^0 is set to 1. By our construction of π , the second constraint has to be tight, that is,

$$\sum_{k=1}^m \sum_{e \in E_0 \cap E_k: v \in e} y_e^k = c_v^0,$$

and all edges in the sum are assigned to the variable x_v^0 , ensuring (i). In addition, as these edges in the sum are not assigned to any other variable x_v^k for $k \neq 0$ and each such edge has only two vertices, (ii) is maintained. $\square$

As shown in this example, instead of actually solving the dual LP (DLP-SVC), we can construct a feasible dual solution maintaining some desired properties. In this case, constructing the dual solution is much faster than solving the dual LP and, hence, leads to a much faster algorithm. Other common techniques for designing approximation algorithms include various rounding techniques, greedy algorithms, or randomized algorithms.

Literature Review

Traditionally, approximation algorithm techniques have been applied primarily to deterministic combinatorial optimization problems, for instance, the set cover problem, the knapsack problem, the bin-packing problem, the traveling salesman problem, scheduling problems, and so on. We refer interested readers to Ausiello *et al.* [1], Vazirani [2], and Williamson and Shmoys [3] for more details on various deterministic models. Our literature review will mainly focus on the design of approximation algorithms for stochastic optimization models, with emphasis on the recent development in various operations management models.

Stochastic optimization is a vast field, beginning with the works of Dantzig [7] and Beale [8] in the 1950s and seeing much activity until this day; we refer interested readers to the following books that survey the field: Birge and Louveaux [9], Kall and Wallace [10], Stougie and Van Der Vlerk [11], and Ruszczyński and Shapiro [12]. Stochastic optimization problems are often computationally quite difficult and often more difficult than their deterministic counterparts, both from the viewpoint of complexity theory and from a practical perspective. In many settings, the computational difficulty stems from the fact that the distribution might assign a

nonzero probability to an exponential number of scenarios, leading to a considerable increase in the problem complexity, a phenomenon often called the *curse of dimensionality*.

The work on approximation algorithms for stochastic combinatorial problems goes back to the work on stochastic scheduling problem of Möhring *et al.* [13, 14] and the more recent work of Möhring *et al.* [15]. By its very nature, scheduling algorithms often have to account for uncertainty in the sizes and arrival times of future jobs. Next, we will survey some recent development on approximation algorithms for these stochastic optimization models related to operations management. We attempt to divide the relevant work into three groups. The first group surveys the growing stream of approximation results for two-stage or multistage stochastic optimization models. These results are usually derived for general purposes and can be potentially applied to specific problems in operations management. The second group discusses the recent advances in designing approximation algorithms for stochastic inventory systems, with a summary of many key results. The third group presents relevant work in other core operations management models, such as scheduling, facility location, vehicle routing, and revenue management.

General Stochastic Optimization Models. The first worst-case analysis of approximation algorithms for two-stage stochastic programming problems was the work on service provisioning in a telecommunication network by Dye *et al.* [16]. Ravi and Sinha [4] studied two-stage stochastic versions of several combinatorial optimization problem (e.g., shortest path, vertex cover, facility location, set cover, and bin-packing problems) and provided nearly tight approximation algorithms for them when the number of scenarios is polynomial. Independently, Immorlica *et al.* [17] provided approximation algorithms for several covering and packing problems. Their model allowed for exponential scenarios (through an independent activation model) but the costs in the two stages have to be proportional.

Gupta *et al.* [18] were the first to consider the *black-box* model and provided sampling-based approximation algorithms for various two-stage problems with a proportional cost structure. In the black-box model, the underlying distribution is not specified exactly, but the algorithm has access to an oracle that can be used to sample from the underlying distribution, and the running time is measured in terms of the number of calls to this oracle. Gupta *et al.* [19] extended this boosted sampling framework for multistage stochastic optimization problems with recourse. Shmoys and Swamy [20–22] showed that one could derive a PTAS using sample average approximation and adapted ellipsoid methods for the exponentially large LP relaxations and, then, use a simple rounding approach to derive approximation algorithms for the original black-box model without any proportional cost assumption. Swamy and Shmoys [23] extended their results to multistage stochastic optimization problems and showed that the LP solution for each stage can be rounded to an integer solution independent of other stages. Charikar *et al.* [24] provided a general technique based on a sample average approximation that reduced the problem of obtaining a good approximation algorithm for the two-stage black-box model, to the problem of obtaining the analogous results in the polynomial scenario setting. Srinivasan [25] improved on the work of Swamy and Shmoys [23] by showing an approximability, which does not depend multiplicatively on the number of stages.

Dhamdhere *et al.* [26] introduced the robust version of two-stage combinatorial covering problems under uncertainty and provided approximation algorithms for them. There are also some recent works related to robust or risk-averse version under uncertainty (see Gupta *et al.* [27] on covering problems using a guess and prune idea, Golovin *et al.* [28] on two-stage min-cut and shortest path problems, and So *et al.* [29] on problems with controllable risk-aversion level).

Stochastic Inventory Systems. The concept of approximation algorithms has been

applied to several deterministic problems in inventory management, see Silver and Meal [30], Roundy [31], Levi *et al.* [32–34], Shen *et al.* [35], Cheung *et al.* [36]. We then focus on stochastic inventory systems.

The recent stream of research on designing approximation algorithms for the multiperiod stochastic inventory control problems was initiated by Levi *et al.* [37], who proposed a 2-approximation algorithm for the basic uncapacitated backlogged model. Subsequently, Levi *et al.* [38] proposed a 2-approximation algorithm for the capacitated backlogged model. Levi *et al.* [39] designed a 2-approximation algorithms for uncapacitated models with lost sales. Levi and Shi [40] provided a 3-approximation algorithm for the uncapacitated backlogged model with setup costs. Subsequently, Shi *et al.* [41] provided a 4-approximation algorithm for the capacitated backlogged problem with setup costs.

More recently, Truong [42] provided a 2-approximation algorithm for the stochastic inventory problem via a look-ahead (myopic) optimization approach. Tao and Zhou [43] designed a 2-approximation algorithm for stochastic inventory systems with remanufacturing. Chao *et al.* [44] proposed an approximation algorithm with a worst-case guarantee between 2 and 3 for perishable inventory systems. Subsequently, Chao *et al.* [45, 46] studied perishable inventory systems with capacity and setup cost, respectively. There are also recent works on designing approximation algorithms for multi-echelon inventory systems (e.g., Chu and Shen [47] for one-warehouse-multiretailer problems, Levi *et al.* [48] for serial inventory systems, and Levi and Shi [49] for joint replenishment problems). For the distribution-free or the black-box model, Levi *et al.* [50] also proposed an approximation scheme based on a sample average approximation approach for multiperiod stochastic inventory problems.

Other Core Stochastic Operations Management Models. The following list is by no means exhaustive.

- (a) *Scheduling.* Dean [51] in his doctoral thesis provided approximation algorithms for a broad class of stochastic

scheduling problems. He also conducted an exhaustive survey of this topic before his thesis. Shmoys and Sozio [52] designed approximation algorithms based on the approach of Charikar *et al.* [24] for two-stage stochastic scheduling problems, extending the results of Bar-Noy *et al.* [53].

- (b) *Vehicle Routing.* Gupta *et al.* [54] gave randomized approximation algorithms with factor $1 + \alpha$ for split-delivery vehicle routing problem with stochastic demands and $2 + \alpha$ for unsplit-delivery counterpart, where α is the best approximation guarantee for the traveling salesman problem. They also showed that the cyclic heuristic for split-delivery achieves a constant approximation ratio, thereby confirming the conjecture in Bertsimas [55]. More recently, Gørtz *et al.* [56] formulated the stochastic vehicle routing problem via a two-stage stochastic optimization with recourse and gave approximation results.
- (c) *Facility Location.* For the two-stage recourse model, Shmoys and Swamy [20] gave a $(3.225 + \epsilon)$ -approximation algorithm, whereas Gupta *et al.* [18] gave an 8.45-approximation algorithm in the black-box model with proportional cost, and Srinivasan [25] presented a 3.25-approximation algorithm for facility location in the polynomial scenario setting. So *et al.* [29] gave an 8-approximation algorithm for the risk-adjusted two-stage stochastic facility location problem. Shen *et al.* [57] studied a reliable facility location problem wherein some facilities are subject to failure from time to time and proposed a 4-approximation algorithm.
- (d) *Stochastic Network Design.* Gupta *et al.* [58, 59] gave LP rounding approximation algorithms for stochastic Steiner tree and stochastic network design problems. Krishnaswamy *et al.* [60] gave approximation algorithms for node-capacitated network design to minimize the energy consumption.

- (e) *Resource Allocation and Revenue Management*. Dean *et al.* [61] presented approximation algorithms for stochastic knapsack problems. Chan and Farias [62] gave a 2-approximation algorithm for multiperiod stochastic depletion problems. Geunes *et al.* [63] proposed a 1.582-approximation algorithm via LP rounding scheme for supply chain planning and logistics problems with market choice under demand uncertainties. Goyal *et al.* [64] devised a PTAS for the assortment planning problem under dynamic substitution and stochastic demand. Levi and Radovanoic [65] gave a 2-approximation algorithm for revenue management models with reusable resources. Subsequently, Levi and Shi [66] then extended the same results allowing for advanced reservations.

Many if not most of the core problems in operations management fall into the category of multistage stochastic optimization models. In particular, one has to make multiple, typically dependent decisions over time to optimize a certain objective function under uncertainty on how the system will evolve over the future time horizon. Unfortunately, it is often computationally intractable to find the exact optimal solutions for these fundamental and important models. Approximation algorithms seem to be the natural remedy for overcoming this prohibitive computing resource requirement.

Indeed, most of the heuristics and algorithms that have been proposed for operations management models were evaluated merely through computational experiments on randomly generated instances. This does not necessarily provide strong indications that the proposed heuristics are good in general, beyond the instances that were actually tested. In contrast, approximation algorithms have the advantage that they provide *a priori* and *a posteriori* guarantees on the quality of the solution produced by the algorithm. Moreover, the performance analysis provides insights on how to design algorithms that have good empirical performance, which

is in most cases significantly better than the worst-case performance guarantees.

Organization of the Article

The remainder of the article is organized as follows. In the section titled “Approximation Algorithms on Stochastic Inventory Models”, we discuss recent results on designing approximation algorithms for stochastic inventory systems. In the section titled “Approximation Algorithms on Revenue Management Models”, we present recent results on designing approximation algorithms for revenue management problems. The section titled “Conclusion and Future Directions” concludes our article and points out plausible avenues for future research.

APPROXIMATION ALGORITHMS ON STOCHASTIC INVENTORY MODELS

In this section, we focus our attention on designing approximation algorithms for stochastic inventory models that allow for generally correlated demand structures capturing demand seasonality and forecast updates. Table 1 shows the current available results on various stochastic inventory systems.

When decision makers want to incorporate any forecast update mechanisms, the future demands can be represented as a function of the current information set consisting of past realized demand information and some other possible exogenous information that is available to them. This unavoidably introduces correlations between the future demands, which make the dynamic programming formulation computationally intractable, as the state space grows exponentially fast. This is usually referred to as *the curse of dimensionality*. In this article, we present a single-echelon problem called the *stochastic lot-sizing problem*. The term lot-sizing captures the setup cost or interchangeably the fixed ordering cost whenever a strictly positive order is placed. We demonstrate how randomized approximation algorithms can be designed for this problem.

Table 1. Summary of Current Results on Stochastic Inventory Control Systems

Stochastic Inventory Control Systems	Approximation Ratio	References
Backlogged	2	[37, 42]
Lost sales	2	[39]
Backlogged, capacity	2	[38]
Backlogged, setup cost	3	[40]
Backlogged, setup cost, capacity	4	[41]
Backlogged, perishable	2–3	[44]
Backlogged, perishable, capacity	2–3	[45]
Backlogged, perishable, setup cost	3–4	[46]
Backlogged, remanufacturing	2	[43]
Backlogged, serial system	2	[48]
Service-level, one-warehouse multiretailer	1.26	[47]

Model

We consider a finite planning horizon of T periods indexed $t = 1, \dots, T$. The demands over these periods are random variables, denoted by $D_1, \dots, D_T$, and the goal is to coordinate a sequence of orders over the planning horizon to satisfy these demands with minimum expected cost. As a general convention, from now on, we will refer to a random variable and its realization using capital- and lower-case letters.

In each period $t = 1, \dots, T$, four types of costs are incurred, a per-unit ordering cost c_t for ordering any number of units at the beginning of period t , a per-unit holding cost h_t for holding excess inventory from period t to $t + 1$, a per-unit backlogging penalty b_t that is incurred for each unsatisfied unit of demand at the end of period t , and a *fixed ordering cost* K that is incurred in each period with strictly positive ordering quantity. (It should be noted that our analysis remains valid for nonstationary K_t satisfying $\alpha K_{t+1} \leq K_t$, which is commonly assumed in the literature. Without loss of generality, we can assume that the discount factor $\alpha = 1$.) Unsatisfied units of demand are usually called *backorders*. Each unit of unsatisfied demand incurs a per-unit backlogging penalty cost b_t in each period t until it is satisfied. In addition, we consider a model with a lead time of L periods between the time an order is placed and the time at which it actually arrives. We assume that the lead time is a known integer L . We assume without loss of generality that the

discount factor is equal to 1 and that $c_t = 0$ and $h_t, b_t \geq 0$, for each t .

At the beginning of each period s , we observe what is called an *information set* denoted by f_s . The information set f_s contains all of the information that is available at the beginning of time period s . More specifically, the information set f_s consists of the realized demands $d_1, \dots, d_{s-1}$ over the interval $[1, s)$ and possibly some exogenous information. The information set f_s in period s is one specific realization in the set of all possible realizations of the random vector $F_s = (D_1, \dots, D_{s-1})$. The set of all possible realizations is denoted by $\mathcal{F}_s$. The observed information set f_s induces a given conditional joint distribution of the future demands $(D_s, \dots, D_T)$. For ease of notation, D_t will always denote the random demand in period t according to the conditional joint distribution in some period $s \leq t$, where it will be clear from the context to which period s it refers. The index t will be used to denote a general time period, and s will always refer to the current period. The only assumption on the demands is that for each $s = 1, \dots, T$, and each $f_s \in \mathcal{F}_s$, the conditional expectation $\mathbb{E}[D_t | f_s]$ is well defined and finite for each period $t \geq s$. In particular, we allow for nonstationary and correlation between the demands in different periods.

The traditional approach to study these models has been dynamic programming. Using the dynamic programming approach, it can be shown that *state-dependent* (s, S)

policies are optimal (see Zipkin [68]). However, the computational complexity of the resulting dynamic programs is very sensitive to the dimension of the sets $\mathcal{F}_s$. In particular, in many practical scenarios, these sets are of high dimension, which leads to dynamic programming formulations that are computationally intractable. In fact, it has been shown in Halman *et al.* [69] that this model is NP-hard, even for the relatively simple special case of independent discrete, finite support demands. In fact, for this special case, it is possible to construct a PTAS; that is, the problem can be approximated to an arbitrary degree of accuracy with a running time that depends on the degree of accuracy.

In addition, Guan and Miller [70, 71], Huang and Küçükyavuz [72], and Jiang and Guan [73] proposed exact and polynomial-time algorithms for the stochastic lot-sizing problem if the stochastic programming scenario tree is polynomially representable. These models allow for stochastic and correlated demands. However, the scenario tree in our model is exponentially large. Besides the exact (dynamic programming) approaches to stochastic lot-sizing problems, Guan *et al.* [74] and Zhang *et al.* [75] also proposed branch-and-cut methods to this class of problems.

Randomized Cost-Balancing Policies

In this section, we shall describe the *randomized cost-balancing policy* following Levi and Shi [40]. This policy is based on two major ideas: (i) *Marginal cost accounting scheme*. The standard dynamic programming approach directly assigns to the decision of how many units to order in each period only the expected holding and backlogging costs incurred in that period, although this decision might affect the costs in future periods. Instead, marginal cost accounting scheme assigns to the decision in each period *all* the expected costs that, once this decision is made, become unaffected by any decision made in future periods. These costs may still depend on future demands. (ii) *Randomizing cost balancing*. The idea of cost balancing was used in the past to construct heuristics with constant performance guarantees for deterministic inventory problems (Silver and

Meal [30]). The key observation in this model is that any policy in any period incurs potential expected costs because of *overordering* (namely, expected holding costs of carrying excess inventory) and *underordering* (namely, expected backlogging costs incurred when demand is not met on time). To address the nonlinearity induced by the fixed costs, a randomized decision rule is employed to balance the expected fixed ordering costs, holding costs, and backlogging costs, in each period. In particular, the order quantity in each period is decided based on a carefully designed randomized rule that chooses among various possible order quantities with carefully chosen probabilities.

Marginal Cost-Accounting Scheme. We first introduce a *marginal holding cost-accounting approach*. Without loss of generality, assume that the ordered supply units are consumed on *first-ordered, first-consumed basis*. The key observation under this assumption is that once an order is placed in some period, then the expected holding cost that the units just ordered will incur over the rest of the planning horizon is a function only of the realized demands over the rest of the horizon and not of any future orders. Hence, within each period, we can associate the overall expected holding cost that is incurred by the units ordered in this period over the entire horizon. We note that similar ideas of holding cost accounting were used previously in the context of models with continuous time, infinite horizon, and stationary (Poisson-distributed) demand (see the work of Axsäter and Lundell [76] and Axsäter [77]). More specifically, let x_s be the *inventory position* at the beginning of period s that captures the total sum of the physical *on-hand inventory* and the *outstanding orders* (placed in past periods but still on the way) minus the pending backlogged demand. Say now that q_s units were ordered in period s , and consider a future period $t \geq s + L$. Then, the holding cost incurred by the q_s units ordered in period s at the end of period t is $h_t(q_s - (D_{[s,t]} - x_s)^+)$, where $x^+ = \max(x, 0)$ and $D_{[s,t]} = \sum_{j=s}^t D_j$ is the cumulative demand over the interval $[s, t]$. Observe that if $D_{[s,t]} \leq x_s$, then none of the q_s units has been yet consumed. When

$D_{[s,t]}$ exceeds x_s , the q_s units are used to satisfy the demand until all of them are consumed. It follows that the total holding cost incurred by the q_s units ordered in period s over the entire horizon is equal to

$$H_s = H_s(Q_s) \triangleq \sum_{t=s+L}^T h_t(Q_s - (D_{[s,t]} - X_s)^+)^+. \quad (2)$$

Because X_s and Q_s are realized at the beginning of period s (whereas, x_s and q_s are the realizations of X_s and Q_s , respectively), then, as seen from the beginning of period s , this quantity depends only on future demands and not on any future decision.

In addition, in an uncapacitated model, the decision of how many units to order in each period affects the expected backlogging cost in only a single future period, namely, a lead time ahead. Now, let Π_s be backlogging cost incurred in period $s+L$, for each $s = 1-L, \dots, T-L$. In particular, it is straightforward to verify that

$$\Pi_s \triangleq b_{s+L}(D_{[s,s+L]} - (X_s + Q_s))^+, \quad (3)$$

where $D_j \triangleq 0$ with probability 1 for each $j \leq 0$. (Observe that the supply units captured by $X_s + Q_s$ will become available by time period $s+L$ and that no order placed after period s will arrive by time period $s+L$.)

Now, let $\mathcal{C}(P)$ be the cost of a feasible policy P and use the superscript P to relate the respective quantities to that policy. Clearly,

$$\begin{aligned} \mathcal{C}(P) \triangleq & \sum_{t=1-L}^0 \Pi_t^P + H_{(-\infty,0]} \\ & + \sum_{t=1}^{T-L} (K \cdot \mathbf{1}(Q_t^P > 0) + H_t^P + \Pi_t^P), \end{aligned} \quad (4)$$

where $H_{(-\infty,0]}$ denotes the total holding cost incurred by units ordered before period 1 (given as an input). We note that the first two expressions $\sum_{t=1-L}^0 \Pi_t^P$ and $H_{(-\infty,0]}$ are the same for any feasible policy and each realization of the demand, and, therefore, we will omit them. Because they are nonnegative, this will not affect our approximation results. In addition, observe that, without loss of generality, it can be assumed that $Q_t^P = H_t^P = 0$ for any policy P and each period

$t = T-L+1, \dots, T$, because nothing that is ordered in these periods can be used within the given planning horizon. We now can write

$$\mathcal{C}(P) \triangleq \sum_{t=1}^{T-L} (K \cdot \mathbf{1}(Q_t^P > 0) + H_t^P + \Pi_t^P). \quad (5)$$

The cost-accounting scheme in Equation (5) is marginal; that is, in each period, we account for all the expected costs that become unaffected by any future decision.

Randomized Cost-Balancing Policy. To describe the policy, we modify the definition of the information set f_t to also include the randomized decisions of the randomized balancing policy up to period $t-1$. Thus, given the information set f_t , the inventory position at the beginning of period t is known. However, the order quantity in period t is still *unknown* because the policy randomizes among various order quantities. We denote the randomized cost-balancing policy by RB. The decision in each period, whether to order and how much to order, is based on the following quantities.

- Compute the *balancing quantity* $\hat{q}_t$, which balances the expected marginal holding cost incurred by the units ordered against the expected backlogging cost in period $t+L$. That is, $\hat{q}_t$ uniquely solves

$$\mathbb{E}[H_t^{RB}(\hat{q}_t)|f_t] = \mathbb{E}[\Pi_t^{RB}(\hat{q}_t)|f_t] \triangleq \theta_t. \quad (6)$$

- Compute the *holding-cost- K quantity* $\tilde{q}_t$ that solves $\mathbb{E}[H_t^{RB}(\tilde{q}_t)|f_t] = K$, that is, $\tilde{q}_t$ is the order quantity that brings the expected marginal holding cost to K .
- Compute $\mathbb{E}[\Pi_t^{RB}(\tilde{q}_t)|f_t]$, that is., the expected backlogging cost if one orders $\tilde{q}_t$ units in period t .
- Compute $\mathbb{E}[\Pi_t^{RB}(0)|f_t]$, that is, the expected backlogging cost resulting from not ordering in period t .

On the basis of the above-mentioned quantities computed, the following randomized rule is used in each period t . Let P_t denote our ordering probability, which is *a priori*

random. With the observed information set f_t , the ordering probability $p_t = P_t|f_t$ in period t is defined differently in the two cases below.

Case I. If the balancing cost exceeds K , that is, $\theta_t \geq K$, the RB policy orders the balancing quantity $q_t^{RB} = \hat{q}_t$ with probability $p_t = 1$. The intuition is that when $\theta_t \geq K$, the fixed ordering cost K is less dominant compared to marginal holding and backlogging costs. Moreover, if the RB policy does not place an order, the conditional expected backlogging cost is potentially large. Thus, it is worthwhile to order the balancing quantity $q_t^{RB} = \hat{q}_t$ with probability $p_t = 1$.

Case II. If the balancing cost is less than K , that is, $\theta_t < K$, the RB policy orders the holding-cost- K quantity (i.e., $q_t^{RB} = \tilde{q}_t$) with probability p_t and nothing with probability $1 - p_t$. That is,

$$q_t^{RB} = \begin{cases} \tilde{q}_t, & \text{with probability } p_t \\ 0, & \text{with probability } 1 - p_t \end{cases}. \quad (7)$$

The probability p_t is computed by solving the following equation:

$$p_t K = p_t \cdot \mathbb{E}[\Pi_t^{RB}(\tilde{q}_t)|f_t] + (1 - p_t) \cdot \mathbb{E}[\Pi_t^{RB}(0)|f_t]. \quad (8)$$

The underlying reason behind the choice of this particular randomization in Equation (8) is that the policy perfectly balances the three types of costs, namely, the marginal holding cost, the marginal backlogging cost, and the fixed ordering cost associated with the period t . In particular, as we order the holding-cost- K quantity with probability p_t and nothing with probability $1 - p_t$, the conditional expected marginal holding cost in this case is

$$\begin{aligned} \mathbb{E}[H_t^{RB}(q_t^{RB})|f_t] &= p_t \mathbb{E}[H_t^{RB}(\tilde{q}_t)|f_t] \\ &+ (1 - p_t) \mathbb{E}[H_t^{RB}(0)|f_t] = p_t K. \end{aligned} \quad (9)$$

By the construction of p_t in Equation (8), the conditional expected backlogging cost is

$$\begin{aligned} \mathbb{E}[\Pi_t^{RB}(q_t^{RB})|f_t] &= p_t \mathbb{E}[\Pi_t^{RB}(\tilde{q}_t)|f_t] \\ &+ (1 - p_t) \mathbb{E}[\Pi_t^{RB}(0)|f_t] = p_t K. \end{aligned} \quad (10)$$

As p_t is the ordering probability in case II, the expected fixed ordering cost is $p_t K$. It can be shown that (8) has the following solution:

$$0 \leq p_t = \frac{\mathbb{E}[\Pi_t^{RB}(0)|f_t]}{K - \mathbb{E}[\Pi_t^{RB}(\tilde{q}_t)|f_t] + \mathbb{E}[\Pi_t^{RB}(0)|f_t]} < 1. \quad (11)$$

The inequalities in Equation (11) follows from the fact that $\theta_t < K$ and $\tilde{q}_t > \hat{q}_t$, which implies that $\mathbb{E}[\Pi_t^{RB}(\tilde{q}_t)|f_t] < \mathbb{E}[\Pi_t^{RB}(\hat{q}_t)|f_t] = \theta_t < K$.

Worst-Case Performance Analysis. To obtain a 3-approximation algorithm, one wishes to show that, on expectation, the cost of an optimal policy can “pay” for at least one-third of the expected cost of the randomized cost-balancing policy. The periods are decomposed into subsets in which we will define explicitly. For certain well-behaved subsets, we want to show that the holding and backlogging costs incurred by an optimal policy can pay for one-third of the cost incurred by the RB policy. The difficulty arises in analyzing the remaining subset of *problematic periods*, for which it is not a priori clear how to pay for their cost. These problematic periods are further partitioned into intervals defined by each pair of two consecutive orders placed by the optimal policy. It can be shown that the total expected cost incurred by the RB policy in problematic periods within each interval does not exceed $3K$. This implies that the fixed ordering cost incurred by an optimal policy can pay on expectation one-third of the cost incurred by the randomized cost-balancing policy in problematic periods. Let Z_t^{RB} be a random variable defined as

$$Z_t^{RB} \triangleq \mathbb{E}[H_t^{RB}(Q_t^{RB})|F_t] = \mathbb{E}[\Pi_t^{RB}(Q_t^{RB})|F_t]. \quad (12)$$

Note that Z_t^{RB} is a random variable that is realized with the information set in period t . Observe that by the construction of the RB policy, the random variable Z_t^{RB} is well defined because the expected marginal holding costs and the expected marginal backlogging costs are always balanced. That is, the conditional expected marginal holding cost is always equal to the conditional expected backlogging cost. In addition, the expected fixed ordering cost in period t is also Z_t^{RB}

by the construction of the algorithm, and, therefore, we have the following lemma:

Lemma 1 (Levi and Shi [40]) *Let $C(RB)$ be the total cost incurred by the RB policy. Then, we have*

$$\mathbb{E}[C(RB)] \leq 3 \cdot \sum_{t=1}^{T-L} \mathbb{E}[Z_t^{RB}]. \quad (13)$$

To complete the worst-case analysis, we would like to show that the expected cost of an optimal policy denoted by OPT is at least $\sum_{t=1}^{T-L} \mathbb{E}[Z_t^{RB}]$. This will be carried out by amortizing the cost of OPT against the cost of the RB policy. In particular, we shall show that, on expectation, OPT pays for a large fraction of the cost of the RB policy. In the subsequent analysis, we will use a random partition of periods $t = \{1, 2, \dots, T-L\}$ to the following sets: The set $\mathcal{T}_{1H} \triangleq \{t : \Theta_t \geq K \text{ and } Y_t^{OPT} > Y_t^{RB}\}$ consists of periods in which the balancing cost Θ_t exceeds K and the optimal policy had higher inventory position than that of the RB policy after ordering [recall that, if $\Theta_t \geq K$, then the RB policy orders the balancing quantity with probability 1 and the value Y_t^{RB} is known deterministically (i.e., realized) with F_t]. The set $\mathcal{T}_{1\pi} \triangleq \{t : \Theta_t \geq K \text{ and } Y_t^{OPT} \leq Y_t^{RB}\}$ consists of periods in which the balancing cost exceeds K and the inventory position of the optimal policy does not exceed that of the RB policy after ordering (see the above-mentioned comment regarding $\mathcal{T}_{1H}$). The set $\mathcal{T}_{2H} \triangleq \{t : \Theta_t < K \text{ and } Y_t^{OPT} \geq X_t^{RB} + \tilde{Q}_t^{RB}\}$ consists of periods in which the balancing cost is less than K , and in such periods, the inventory position of the RB policy after ordering would be either X_t^{RB} if no order was placed or $X_t^{RB} + \tilde{Q}_t^{RB}$ if the *holding-cost- K quantity* is ordered, depending on the randomized decision of the RB policy. However, the inventory position of OPT after ordering exceeds even $X_t^{RB} + \tilde{Q}_t^{RB}$. [Note again that the quantity $\tilde{Q}_t^{RB}$ is known deterministically (i.e., realized) with F_t .] Analogous to $\mathcal{T}_{2H}$, the set $\mathcal{T}_{2\pi} \triangleq \{t : \Theta_t < K \text{ and } X_t^{RB} \geq Y_t^{OPT}\}$ consists of periods in which the inventory position of OPT after ordering is below X_t^{RB} . The set $\mathcal{T}_{2M} \triangleq \{t : \Theta_t < K \text{ and } X_t^{RB} < Y_t^{OPT} < X_t^{RB} + \tilde{Q}_t^{RB}\}$ consists of

periods in which the balancing cost is less than K and the inventory position of OPT after ordering is within $(X_t^{RB}, X_t^{RB} + \tilde{Q}_t^{RB})$. Thus, whether the RB policy or OPT has more inventory depends on whether the RB policy placed an order. Note that the sets $(\mathcal{T}_{1H} - \mathcal{T}_{2M})$ are disjoint, and the union makes a complete set. Conditioning on f_t , it is already known which part of the partition period t belongs.

Next, we will explain that the total holding cost incurred by OPT is higher than the marginal holding cost incurred by the RB policy in periods that belong to $\mathcal{T}_{1H} \cup \mathcal{T}_{2H}$ and that the total backlogging cost incurred by OPT is higher than the backlogging cost incurred by the RB policy associated with periods within $\mathcal{T}_{1\pi} \cup \mathcal{T}_{2\pi}$.

Lemma 2 (Levi and Shi [40]) *The overall holding cost and backlogging cost incurred by OPT are denoted by H^{OPT} and Π^{OPT} , respectively. Then, we have, with probability 1,*

$$\begin{aligned} H^{OPT} &\geq \sum_t H_t^{RB} \cdot \mathbf{1}(t \in \mathcal{T}_{1H} \cup \mathcal{T}_{2H}), \Pi^{OPT} \\ &\geq \sum_t \Pi_t^{RB} \cdot \mathbf{1}(t \in \mathcal{T}_{1\pi} \cup \mathcal{T}_{2\pi}). \end{aligned} \quad (14)$$

The idea of Lemma 2 is as follows: In each period $t \in \mathcal{T}_{1H} \cup \mathcal{T}_{2H}$, the inequality $Y_t^{RB} < Y_t^{OPT}$ holds and implies that the Q_t^{RB} units ordered by the RB policy in period t have been ordered by OPT either in period t or even earlier. Thus, the holding cost they incur under OPT are higher than those incurred under the RB policy. On the other hand, in each period $t \in \mathcal{T}_{1\pi} \cup \mathcal{T}_{2\pi}$, the inequality $Y_t^{RB} \geq Y_t^{OPT}$ holds and implies that the backlogging incurred by OPT at the end of period $t+L$ will be higher than that of the RB in that period. We are still left with the problematic set $\mathcal{T}_{2M}$. Note that, in this particular set, whether the RB policy or OPT has more inventory depends on whether the RB policy placed an order. Fortunately, Lemma 3 shows that the fixed ordering costs incurred by OPT can cover the randomized balancing costs in $\mathcal{T}_{2M}$.

Lemma 3 (Levi and Shi [40]) *The expected randomized cost in set $\mathcal{T}_{2M}$ is less than the*

total expected fixed ordering cost incurred by OPT , that is,

$$\begin{aligned} & \mathbb{E} \left[\sum_t Z_t^{RB} \cdot \mathbf{1}(t \in \mathcal{T}_{2M}) \right] \\ & \leq \mathbb{E} \left[\sum_{t=1}^{T-L} K \cdot \mathbf{1}(Q_t^{OPT} > 0) \right]. \end{aligned} \quad (15)$$

As an immediate consequence of Lemmas 2 and 3, we obtain the following lemma and theorem.

Lemma 4 (Levi and Shi [40]) *Let $C(OPT)$ be the total cost incurred by the cost-balancing policy RB . Then, we have,*

$$\mathbb{E}[C(OPT)] \geq \sum_{t=1}^{T-L} \mathbb{E}[Z_t^{RB}]. \quad (16)$$

Theorem 2 (Levi and Shi [40]) *For each instance of the stochastic lot-sizing problem, the expected cost of the randomized cost-balancing policy RB is at most three times the expected cost of an optimal policy OPT , that is,*

$$\mathbb{E}[C(RB)] \leq 3 \cdot \mathbb{E}[C(OPT)]. \quad (17)$$

Some Remarks and Future Directions

The worst-case analysis suggests that the theoretical worst-case performance bound is three. However, extensive computational experiments (Levi and Shi [40]) show that the randomized cost-balancing policy performs significantly better than the worst-case guarantee, in most cases within 5% of optimum.

It has been shown (see, Levi and Shi [40]) that one can consider parametric versions of these policies and use known lower and upper bounds on the optimal ordering levels to devise more sophisticated policies with better empirical (typical) performance, some of which still admit worst-case analysis. The parameterization of these policies is motivated by the fact that, in the context of worst-case analysis, one wishes to choose policies

that “protect” against all possible instances, whereas per a given (known) instance, it is possible to choose the parameters of the policy optimally with respect to that instance. This naturally leads to improved empirical performance.

Besides the summary of current results shown in Table 1, designing approximation algorithms to stochastic assemble-to-order systems, stochastic one-warehouse multiretailer problems, stochastic joint inventory, and pricing models, the stochastic perishable inventory with depletion decisions remains an open challenge and will definitely require novel ideas and techniques. Another important future research direction is to study the performance of cost-balancing policies under various assumptions on the underlying demand distributions. As much as it is powerful to establish general worst-case analysis, it is equally important to refine this analysis to various parametric regimes of the underlying demand distributions and other key parameters of the problem. We call this *parametric worst-case analysis*.

APPROXIMATION ALGORITHMS ON REVENUE MANAGEMENT MODELS

In this section, we consider a class of revenue management problems that arise in systems with *reusable resources* and *advanced reservations*. This work is motivated by both traditional and emerging application domains, such as hotel room management, car rental management, and workforce management. For instance, in hotel industries, customers make requests to book a room in the future for a specified number of days. This is called *advanced reservation*. Rooms are allocated to customers based on their requests, and after one customer used a room, it becomes available to serve other customers. One of the major issues in these systems is how to manage capacitated pool of reusable resources over time in a dynamic environment with many uncertainties.

Model

We consider revenue management problems of a single pool of reusable resources used to

serve multiple classes of customers through advanced reservations. There is a single pool of resources of integer capacity $C < \infty$ that is used to satisfy the demands of M different classes of customers. The customers of each class $k = 1, \dots, M$, arrive according to an independent Poisson process with respective rate λ_k . Each class- k customer requests to reserve one unit of the capacity for a specified *service time interval* in the future.

Let D_k be the reservation distribution of a class- k customer and S_k be the respective service distribution with mean μ_k . We assume that they are nonnegative, discrete, and bounded. In particular, upon an arrival of a class- k customer at some random time t , the customer requests to reserve the service time interval $[t + d, t + d + s]$, where d is distributed according to D_k and s according to S_k . Note that D_k and S_k are independent of the arrival process and between customers; however, per customer, D_k and S_k can be correlated. (We assume that both D_k and S_k are finite discrete distributions.) During the time a customer is served (i.e., $[t + d, t + d + s]$), the requested unit cannot be used by any other customer; after the service is over, the unit becomes available again to serve other customers. If the resource is reserved, the customer pays a class-specific rate of r_k dollars per unit of service time. The resource can be reserved for an arriving customer only if on arrival there is at least one unit of capacity that is available (i.e., not reserved) throughout the entire requested interval $[t + d, t + d + s]$. Specifically, a customer's request can be satisfied if the maximum number of already reserved resources throughout the requested service interval is smaller than the capacity C . However, customers can be rejected even if there is available capacity. Rejecting a customer now possibly enables serving more profitable customers in the future. Customers whose requests are not reserved upon arrival are *lost* and leave the system. The goal is to find a feasible admission policy that maximizes the expected long-run average revenue. Specifically, if $\mathcal{R}_\pi(T)$ denotes the revenue achieved by policy π over the interval $[0, T]$, then the expected long-run average revenue of π is defined

as $\mathcal{R}(\pi) \triangleq \liminf_{T \rightarrow \infty} (\mathbb{E}[\mathcal{R}_\pi(T)]/T)$, where the expectation is taken with respect to the probability measure induced by π .

Like many stochastic optimization models, one can formulate this problem using dynamic programming approach. However, even in special cases (e.g., no advanced reservations allowed and with exponentially distributed service times), the resulting dynamic programs are computationally intractable because of the *curse of dimensionality*.

An LP-Based Approach

In this section, we describe a simple LP that provides an upper bound on the achievable expected long-run average revenue. The LP conceptually resembles the one used by Levi and Radovanovic [65], Key [78], and Iyengar and Sigman [79], who studied models without advanced reservations. It is also similar in spirit to the one used by Adelman [80] in the queueing networks framework with unit resource requirements again without advanced reservations. We shall show how to use the optimal solution of the LP to construct a simple admission control policy that is called *class selection policy* (CSP).

At any point of time t , the state of the system is specified by the entire booking profile consisting of the class, reservation, and service information of each customer in the booking system as well as the customers currently served. Without loss of generality, we restrict attention to *state-dependent policies*. Note that each state-dependent policy induces a Markov process over the state space, and one can show that the induced Markov process has a unique stationary distribution, which is ergodic. The detailed technical proof can be found in Levi and Shi [66] following the arguments in Sevastyanov [81] and Lu and Radovanovic [82]. The key idea is to find a *fixed* auxiliary probability distribution that can be scaled (by positive constants) to lower and upper bounds of the transitional probability in the underlying Markov chain. This auxiliary probability distribution can be readily found if the state space is compact (which is true in our model).

As any state-dependent policy induces a Markov process on the state space of the

system that is ergodic, for a given state-dependent policy π , there exists a long-run stationary probability α_{ijk}^π for accepting a class- k customer who wishes to start service in i units of time for j units of time, which is equal to the long-run proportion of accepted customers of this type while running the policy π . In other words, any state-dependent policy π is associated with the stationary probabilities α_{ijk}^π for all possible reservation time i , service time j , and class k . Let $\lambda_{ijk} \triangleq \lambda_k \mathbb{P}(D_k = i, S_k = j)$ be the arrival rate of class- k customers with reservation time i and service time j . Therefore, the mean arrival rate of accepted class- k customers with reservation time i and service time j is $\alpha_{ijk}^\pi \lambda_{ijk}$. By Little's Law and PASTA (see, e.g., Gallager [83]), the expected number of class- k customers with reservation time i and service time j being served in the system under π is $\alpha_{ijk}^\pi \lambda_{ijk} j$. It follows that, under π , the expected long-run average number of resource units being used to serve customers can be expressed as $\sum_{k=1}^M \sum_{i,j} \alpha_{ijk}^\pi \lambda_{ijk} j$. This gives rise to the following *knapsack* LP:

$$\max_{\alpha_{ijk}} \sum_{k=1}^M \sum_{i,j} r_k \alpha_{ijk}^\pi \lambda_{ijk} j, \quad (18)$$

$$\text{s.t. } \sum_{k=1}^M \sum_{i,j} \alpha_{ijk}^\pi \lambda_{ijk} j \leq C, \quad (19)$$

$$0 \leq \alpha_{ijk}^\pi \leq 1, \quad \forall i, j, k. \quad (20)$$

Note that, for each feasible state-dependent policy π , the vector $\alpha^\pi = \{\alpha_{ijk}^\pi\}$ is a feasible solution for the LP with objective value equal to the expected long-run average revenue of policy π . In fact, the LP enforces the capacity constraint (19) of the system only in expectation, whereas in the original problem, this constraint has to hold, for each sample path. It follows that the LP relaxes the original problem and provides an upper bound on the best-obtained expected long-run average revenue. The LP can be solved optimally by applying the following greedy rule: Without loss of generality, assume that classes are renumbered such that $r_1 \geq r_2 \geq \dots \geq r_M$. Then, for each $k = 1, \dots, M$, we sequentially

set $\alpha_{ijk} = 1$ for all i and j as long as constraint (19) is satisfied. If there exists a class $M' \leq M$ such that

$$C = (1 - \gamma) \sum_{k=1}^{M'-1} \sum_{i,j} \lambda_{ijk} j + \gamma \sum_{k=1}^{M'} \sum_{i,j} \lambda_{ijk} j,$$

for some $\gamma \in (0, 1)$, we set $\alpha_{ijM'} = \gamma$ for all i and j . Note that, for each class k , the values of α_{ijk} are all equal regardless of i and j . We abuse the notation and drop the subscripts i and j of α_{ijk} . Then, the optimal solution reduces to: for $k = 1, \dots, M' - 1$, $\alpha_k = 1$; $\alpha_{M'} = \gamma$; and for $k = M' + 1, \dots, M$, we have $\alpha_k = 0$.

Next, we shall use the optimal solution of the knapsack LP to construct a very simple admission policy. Let $\alpha^* = \{\alpha_k^*\}$ be the optimal solution of the knapsack LP. We propose a simple policy that is called *CSP*. Consider an arrival of a class- k customer ($k = 1, \dots, M$). For each $k = 1, \dots, M' - 1$, *accept* the customer on arrival (regardless of the reservation time and the service time) as long as there is sufficient unreserved capacity throughout the requested service interval. If $k = M'$, *accept* the customer with probability γ (regardless of the reservation time and the service time) and as long as there is sufficient unreserved capacity throughout the requested service interval. For each $k = M' + 1, \dots, M$, *reject*.

The CSP has a very simple structure. It always admits customers from the classes for which the corresponding value α_k^* in the optimal LP solution equals to one as long as capacity permits. It never admits customers from classes for which the corresponding value α_k^* equals to zero, and it flips a coin for the possibly one class with fractional value $\alpha_{M'}^* = \gamma$. The CSP is conceptually very intuitive; in that, it splits the classes into profitable and nonprofitable that should be ignored. In fact, we can assume, without loss of generality, that there is no fractional variable in the optimal solution α^* , that is, for each $k = 1, \dots, M'$, $\alpha_k^* = 1$. (If $\alpha_{M'}^* = \gamma$ is fractional, we think of class M' as having an arrival rate $\lambda'_{M'} = \gamma \lambda_{M'}$ and then eliminate the fractional variable from α^* .)

Worst-Case Performance Analysis

In this section, we discuss performance analysis of the CSP under models with advanced reservations. The CSP induces a well-structured stochastic process called *loss networks* with advanced reservations (i.e., a $M/G/C/C$ loss system with advanced reservations). Each class $k = 1, \dots, M$ induces a Poisson arrival stream with respective rate $\alpha_k^* \lambda_k$, $1 \leq k \leq M$. Thus, for each class k with $\alpha_k^* = 1$, the arrival process is identical to the original process, and each class k with $\alpha_k^* = 0$ can be ignored. For each class $k = 1, \dots, M'$, let S_k and D_k be the service and reservation distributions of class- k customers, respectively. We want to characterize the long-run *blocking probability* of class- k customers with reservation time i and service time j under the CSP, that is, the stationary probability that a class- k customer with reservation time i and service time j arrives at a random time to the system and is rejected by the CSP because there is no available capacity at some point within the requested service interval. For each $k = 1, \dots, M'$, let Q_{ijk} be the stationary probability of blocking a class- k customers with reservation time i and service time j under the CSP. As the corresponding stochastic process is ergodic, Q_{ijk} is well defined. Thus, the expected long-run average revenue of the CSP can be expressed as $\sum_{k=1}^{M'} \sum_{i,j} r_k \lambda_{ijk} j (1 - Q_{ijk})$. However, $\sum_{k=1}^{M'} \sum_{i,j} r_k \lambda_{ijk} j$ is the optimal value of the LP, which is an upper bound on the best-achievable expected long-run average revenue, denoted by $\mathcal{R}(OPT)$. Thus, a key aspect of the performance analysis of the CSP is to obtain an upper bound on the probabilities Q_{ijk} 's. More specifically, if $1 - Q_{ijk} \geq \xi$, for each i, j , and k , it follows that

$$\begin{aligned} \mathcal{R}(CSP) &= \sum_{k=1}^{M'} \sum_{i,j} r_k \lambda_{ijk} j (1 - Q_{ijk}) \\ &\geq \sum_{k=1}^{M'} \sum_{i,j} r_k \lambda_{ijk} j \xi \geq \xi \mathcal{R}(OPT). \end{aligned} \quad (21)$$

We want to upper bound probabilities Q_{ijk} 's and analyze their asymptotic behavior under the Halfin–Whitt regime. The *traffic*

intensity $\rho \triangleq \sum_{k=1}^M \sum_{i,j} \lambda_{ijk} j = \sum_{k=1}^M \lambda_k \mu_k$. Under the Halfin–Whitt regime, the capacity C and the arrival rates λ_k as well as the traffic intensity ρ are scaled together to infinity while keeping the service and the reservation distributions fixed (i.e., $C = \rho + \beta \sqrt{\rho} + o(\sqrt{\rho}) \rightarrow \infty$, for some positive spare capacity parameter β).

Theorem 3 (Levi and Shi [66]) *Consider the revenue management model with a single pool of capacitated reusable resources and advanced reservations under the CSP. Let $\Phi(\cdot)$ be the cumulative density function of a standard normal. Then:*

- (a) *For each k and j , the blocking probability Q_{ijk} has the following asymptotic upper bound:*

$$\lim_{\rho \rightarrow \infty} Q_{ijk} \leq \Phi(-\beta); \quad \lim_{\rho \rightarrow \infty} Q_{ijk} = 0, \quad \forall i \geq 1,$$

where $\beta > 0$ is the spare capacity parameter in the Halfin–Whitt regime.

- (b) *The CSP is guaranteed to obtain at least 0.5 of the optimal expected long-run average revenue in the Halfin–Whitt heavy-traffic limit.*

The upper bound on the blocking probability is obtained by considering a counterpart system with infinite capacity, where all customers are admitted into an $M/G/\infty$ system with advanced reservations. The detailed stochastic analysis of this loss queueing system with advanced reservations falls out of scope of this article, and we refer interested readers to Levi and Shi [66].

Some Remarks and Future Directions

There are also several plausible extensions into pricing models. An interesting direction is to study both the static and dynamic pricing model of reusable resources with advanced reservations. The static pricing model allows the arrival rates being affected by prices. Specifically, consider a two-stage decision. At the first stage, we set the

respective prices $r_1, \dots, r_M$ for each class. This determines the respective arrival rates $\lambda_1(r_1), \dots, \lambda_M(r_M)$. (The rate of class- i customers is affected only by price r_i .) Then, given the arrival rates, we wish to find the optimal admission policy that maximizes the expected long-run revenue rate. We may assume that $\lambda_i(r_i)$ is nonnegative, differentiable, and decreasing in r_i for each $1 \leq i \leq M$. One might construct an upper bound on the achievable expected long-run revenue rate through a nonlinear program and then use it to construct a similar policy with the same performance guarantees. In the dynamic pricing model, consider a single-class time-homogenous Poisson arrival process with rate λ . Each customer's reservation and service time are drawn from D and S , respectively. The system offers a price from a fixed price menu $[r_1, \dots, r_n]$ to an arriving customer with d and s , depending on the current state. The state is characterized by the booking profile, d and s . A reservation price distribution R has to be specified, that is, the customer only accepts the offer if the price offered falls below the reservation price. One might construct a new linear program to obtain provably near-optimal randomized policies.

CONCLUSION AND FUTURE DIRECTIONS

Mathematical programming techniques have been used extensively to obtain relaxations and provably near-optimal approximation algorithms for deterministic combinatorial optimization problems (see, Ausiello *et al.* [1], Vazirani [2], and Williamson and Shmoys [3]). As mentioned in our literature review, more recent work has extended these approaches and/or developed novel techniques for two-stage and multistage stochastic optimization problems. These powerful techniques were developed for a wide range of combinatorial optimization problems and could be potentially applied to many core operation management models. We would like to point out some plausible research avenues for future research: (i) *Data-driven models*. Many of the underlying stochastic applications in operations management often involve correlated data (e.g.,

the demands in stochastic inventory control problems are often correlated over periods because of economic and/or seasonal factors). The decision makers may not know the demand distributions exactly and can only rely on collected (unbiased) sampling data. The sampling-based framework developed by Gupta *et al.* [19] and Swamy and Shmoys [23] could be applied to many distribution-free operations management models. These black-box models are desirable in that they allow one to specify distributions with exponentially many scenarios and correlation in a compact way that makes it reasonable to talk about polynomial-time algorithms. (ii) *Models with risk*. The multistage stochastic model measures the expected cost associated with each stage; however, often in applications, one is also interested in the risk associated with each stage decisions, where risk is some measure of the variability in the random cost incurred in later stages. Gupta *et al.* [58] considered the use of budgets that bound the cost of each scenario, as a means of guarding against risk. It would be interesting to explore stochastic operations management models that incorporate risk. One may also consider designing approximation algorithms for the robust variants of these models. (iii) *Models with endogenous uncertainty*. Another interesting and important research avenue, which brings us closer to Markov decision processes, is to investigate problems where the uncertainty is affected by the decisions taken. For instance, in stochastic scheduling problems, the scheduling decisions interact with the evolution of the random job sizes, especially in a preemptive environment. Another salient example is the joint inventory and pricing problem where the pricing decisions influence the demands.

ACKNOWLEDGMENT

We are very grateful to the Topical Editor Simge Küçükyavuz and two anonymous referees for their detailed comments and suggestions, which have helped to significantly improve both the content and the exposition of this article. The research of Cong Shi is

partially supported by NSF grants CMMI-1362619 and CMMI-1451078.

REFERENCES

1. Ausiello G, Protasi M, Marchetti-Spaccamela A, *et al.* Complexity and approximation: combinatorial optimization problems and their approximability properties. 1st ed. New York, Secaucus (NJ): Springer-Verlag; 1999.
2. Vazirani VJ. Approximation algorithms. Berlin: Springer-Verlag; 2001.
3. Williamson DP, Shmoys DB. The design of approximation algorithms. New York: Cambridge University Press; 2010.
4. Ravi R, Sinha A. Hedging uncertainty: approximation algorithms for stochastic optimization problems. *Math Progr* 2006;108(1):97–114.
5. Monien B, Speckenmeyer E. Ramsey numbers and an approximation algorithm for the vertex cover problem. *Acta Inf* 1985;22(1):115–123.
6. Håstad J. Some optimal inapproximability results. *Proceedings of the 29th Annual ACM Symposium on Theory of Computing, STOC '97*. New York: ACM; 1997. p 1–10.
7. Dantzig GB. Linear programming under uncertainty. *Manage Sci* 1955;1(3-4): 197–206.
8. Beale EML. On minimizing a convex function subject to linear inequalities. *J R Stat Soc Ser B* 1955;17:173–184.
9. Birge JR, Louveaux F. Introduction to stochastic programming. Springer series in operations research and financial engineering. New York: Springer-Verlag; 2011.
10. Kall P, Wallace SW. Stochastic programming. Wiley-Interscience series in systems and optimization. Chichester: John Wiley and Sons Ltd; 1994.
11. Stougie L, Van Der Vlerk MH. Stochastic integer programming. In: Dell'Amico M, Maffioli F, Martello S, editors. *Annotated bibliographies in combinatorial optimization*. New York: John Wiley and Sons, Inc.; 1997. p 127–141.
12. Ruszczyński A, Shapiro A. Stochastic programming. Volume 10, *Handbooks in operations research and management science*. Amsterdam: Elsevier; 2003.
13. Möhring RH, Radermacher FJ, Weiss G. Stochastic scheduling problems I: general strategies. *Z Oper Res* 1984;28:193–260.
14. Möhring RH, Radermacher FJ, Weiss G. Stochastic scheduling problems II: set strategies. *Z Oper Res* 1984;29:65–104.
15. Möhring RH, Schulz A, Uetz M. Approximation in stochastic scheduling: the power of LP-based priority policies. *J ACM* 1999;46:924–942.
16. Dye S, Stougie L, Tomasgard A. The stochastic single resource service provision problem. *Nav Res Logist* 2003;50:869–887.
17. Immorlica N, Karger D, Minkoff M, Mirrokni VS. On the costs and benefits of procrastination: approximation algorithms for stochastic combinatorial optimization problems. *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '04*. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2004. p 691–700.
18. Gupta A, Pál M, Ravi R, Sinha A. Boosted sampling: approximation algorithms for stochastic optimization problems. *36TH STOC*; 2004. p 417–426.
19. Gupta A, Pál M, Ravi R, Sinha A. What about wednesday? Approximation algorithms for multistage stochastic optimization. *APPROX*; 2005. p 86–98.
20. Shmoys DB, Swamy C. Stochastic optimization is (almost) as easy as deterministic optimization. *Proceedings of the 45th Annual IEEE Symposium on the Foundations of Computer Science 50*. Washington (DC): IEEE Computer Society; 2004. p 228–237.
21. Shmoys DB, Swamy C. An approximation scheme for stochastic linear programming and its application to stochastic integer programs. *J ACM* 2006;53(6):978–1012.
22. Swamy C, Shmoys DB. Algorithms column: approximation algorithms for 2-stage stochastic optimization problems. *SIGACT News* 2006;37:33–46.
23. Swamy C, Shmoys DB. Sampling-based approximation algorithms for multi-stage stochastic optimization. *46th Annual IEEE Symposium on Foundations of Computer Science, 2005. FOCS 2005*; 2005. p 357–366.
24. Charikar M, Chekuri C, Pál M. Sampling bounds for stochastic optimization. *Proceedings of the 9th RANDOM*. Springer; 2005. p 257–269.
25. Srinivasan A. Approximation algorithms for stochastic and risk-averse optimization. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07*. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2007. p 1305–1313.

26. Dhamdhere K, Goyal V, Ravi R, *et al.* How to pay, come what may: approximation algorithms for demand-robust covering problems. Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, FOCS '05. Washington (DC): IEEE Computer Society; 2005. p 367–378.
27. Gupta A, Nagarajan V, Ravi R. Thresholded covering algorithms for robust and max-min optimization. *Math Progr* 2014; 146(1-2):583–615.
28. Golovin D, Goyal V, Polishchuk V, *et al.* Improved approximations for two-stage min-cut and shortest path problems under uncertainty. *Math Progr* 2014;1–28.
29. So AM, Zhang J, Ye Y. Stochastic combinatorial optimization with controllable risk aversion level. *Math Oper Res* 2009;34(3):522–537.
30. Silver EA, Meal HC. A heuristic selecting lot-size requirements for the case of a deterministic time varying demand rate and discrete opportunities for replenishment. *Prod Invent Manage* 1973;14:64–74.
31. Roundy RO. Efficient, effective lot-sizing for multi-product, multi-stage production systems. *Oper Res* 1993;41:371–386.
32. Levi R, Roundy RO, Shmoys DB. Primal-dual algorithms for deterministic inventory problems. *Math Oper Res* 2006;31(2):267–284.
33. Levi R, Lodi A, Sviridenko M. Approximation algorithms for the multi-item capacitated lot-sizing problem via flow-cover inequalities. *Math Oper Res* 2008;33(2):461–474.
34. Levi R, Roundy RO, Shmoys DB, *et al.* A constant approximation algorithm for the one-warehouse multi-retailer problem. *Manage Sci* 2008;54(4):763–776.
35. Shen ZM, Shu J, Simchi-Levi D, *et al.* Approximation algorithms for general one-warehouse multi-retailer systems. *Nav Res Logist* 2009;56(7):642–658.
36. Cheung M, Elmachetoub AN, Levi R, *et al.* The submodular joint replenishment problem. Working Paper, MIT; 2014.
37. Levi R, Pál M, Roundy RO, *et al.* Approximation algorithms for stochastic inventory control models. *Math Oper Res* 2007;32(4):821–838.
38. Levi R, Roundy RO, Shmoys DB, *et al.* Approximation algorithms for capacitated stochastic inventory models. *Oper Res* 2008;56(5):1184–1199.
39. Levi R, Janakiraman G, Nagarajan M. A 2-approximation algorithm for stochastic inventory control models with lost-sales. *Math Oper Res* 2008;33(2):351–374.
40. Levi R, Shi C. Approximation algorithms for the stochastic lot-sizing problem with order lead times. *Oper Res* 2013;61(3):593–602.
41. Shi C, Zhang H, Chao X, *et al.* Approximation algorithms for capacitated stochastic inventory systems with setup costs. *Nav Res Logist* 2014;61(4):304–319.
42. Truong VA. Approximation algorithm for the stochastic multiperiod inventory problem via a look-ahead optimization approach. *Math Oper Res* 2014.
43. Tao Z, Zhou XS. Approximation balancing policies for inventory systems with remanufacturing. *Math Oper Res* 2014.
44. Chao X, Gong X, Shi C, *et al.* Approximation algorithms for perishable stochastic inventory systems. Working Paper, University of Michigan; 2013.
45. Chao X, Gong X, Shi C, *et al.* Approximation algorithms for capacitated perishable stochastic inventory systems. Working Paper, University of Michigan; 2014.
46. Chao X, Shi C, Zhang H. Approximation algorithms for stochastic lot-sizing systems with perishable products. Working Paper, University of Michigan; 2014.
47. Chu YL, Shen ZM. A power-of-two ordering policy for one-warehouse multiretailer systems with stochastic demand. *Oper Res* 2010;58(2):492–502.
48. Levi R, Roundy RO, Truong VA. Provably near-optimal balancing policies for multi-echelon stochastic inventory control models. Working Paper, MIT; 2012.
49. Levi R, Shi C. Approximation algorithms for the stochastic joint-replenishment problem. Working Paper, MIT; 2013.
50. Levi R, Roundy RO, Shmoys DB. Provably near-optimal sampling-based policies for stochastic inventory control models. *Math Oper Res* 2007;32(4):821–839.
51. Dean BC. Approximation algorithms for stochastic scheduling problems. PhD Thesis, Cambridge: MIT; 2005.
52. Shmoys DB, Sozio M. Approximation algorithms for 2-stage stochastic scheduling problems. In: Fischetti M, Williamson DP, editors. Integer programming and combinatorial optimization. Volume 4513, Lecture notes in computer science. Heidelberg: Springer Berlin Heidelberg; 2007. p 145–157.

53. Bar-Noy A, Bar-Yehuda R, Freund A, *et al.* A unified approach to approximating resource allocation and scheduling. *J ACM* 2001;48(5):1069–1090.
54. Gupta A, Nagarajan V, Ravi R. Approximation algorithms for VRP with stochastic demands. *Oper Res* 2012;60(1):123–127.
55. Bertsimas D. A vehicle routing problem with stochastic demand. *Oper Res* 1992;40(3):574–585.
56. Gørtz I, Nagarajan V, Saket R. Stochastic vehicle routing with recourse. In: Czumaj A, Mehlhorn K, Pitts A, *et al.*, editors. *Automata, Languages, and Programming*. Volume 7391, Lecture notes in computer science. Heidelberg: Springer Berlin Heidelberg; 2012. p 411–423.
57. Shen ZM, Zhan RL, Zhang J. The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *INFORMS Journal on Computing* 2011;23(3):470–482. no.
58. Gupta A, Ravi R, Sinha A. An edge in time saves nine: LP rounding approximation algorithms for stochastic network design. *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*; 2004. p 218–227.
59. Gupta A, Ravi R, Sinha A. LP rounding approximation algorithms for stochastic network design. *Math Oper Res* 2007;32(2):345–364.
60. Krishnaswamy R, Nagarajan V, Pruhs K, *et al.* Cluster before you hallucinate: approximating node-capacitated network design and energy efficient routing. *STOC*, vol. abs/1403.6207; 2014.
61. Dean BC, Goemans MX, Vondrák J. Approximating the stochastic knapsack problem: the benefit of adaptivity. *Math Oper Res* 2008;33(4):945–964.
62. Chan CW, Farias VF. Stochastic depletion problems: effective myopic policies for a class of dynamic optimization problems. *Math Oper Res* 2009;34:333–350.
63. Geunes J, Levi R, Romeijn HE, *et al.* Approximation algorithms for supply chain planning and logistics problems with market choice. *Math Progr* 2011;130(1):85–106.
64. Goyal V, Levi R, Segev D. Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand, Working Paper. MIT; 2012.
65. Levi R, Radovanovic A. Technical note: provably near-optimal LP-based policies for revenue management in systems with reusable resources. *Oper Res* 2010;58(2):503–507.
66. Levi R, Shi C. Revenue management of reusable resources with advanced reservations, Working Paper. MIT; 2013.
67. Levi R, Pál M, Roundy RO, *et al.* Approximation algorithms for stochastic inventory control models. *Math Oper Res* 2007;32(2):284–302.
68. Zipkin PH. *Foundations of inventory management*. New York: The McGraw-Hill Companies; 2000.
69. Halman N, Klabjan D, Mostagir M, *et al.* A fully polynomial time approximation scheme for single-item stochastic lot-sizing problems with discrete demand. *Math Oper Res* 2009;34(3):674–685.
70. Guan Y, Miller AJ. Polynomial time algorithms for stochastic uncapacitated lot-sizing problems. *Oper Res* 2008;56(5):1172–1183.
71. Guan Y, Miller AJ. A polynomial time algorithm for the stochastic uncapacitated lot-sizing problem with backlogging, IPCO 2008. Volume 5035, *Lecture notes in computer science*. Heidelberg: Springer Berlin; 2008. p 450–462.
72. Huang K, Küçükyavuz S. On stochastic lot-sizing problems with random lead times. *Oper Res Lett* 2008;36(3):303–308.
73. Jiang R, Guan Y. An $O(N^2)$ time algorithm for the stochastic uncapacitated lot-sizing problem with random lead times. *Oper Res Lett* 2011;39(1):74–77.
74. Guan Y, Ahmed S, Nemhauser GL, *et al.* A branch-and-cut algorithm for the stochastic uncapacitated lot-sizing problem. *Math Progr* 2006;105(1):55–84 (English).
75. Zhang M, Küçükyavuz S, Goel S. A branch-and-cut method for dynamic decision making under joint chance constraints. *Manage Sci* 2014;60(5):1317–1333.
76. Axsäter S, Lundell P. In-process safety stock. *Proceedings of the 23rd IEEE Conference on Decision and Control*. Las Vegas (NV): IEEE Control Systems Society; 1984. p 839–842.
77. Axsäter S. Simple solution procedures for a class of two-echelon inventory problems. *Oper Res* (1990);38(1):64–69.
78. Key P. Optimal control and trunk reservation in loss networks. *Probab Eng Inf Sci* 1990;4:203–242.

- 79. Iyengar G, Sigman K. Exponential penalty function control of loss networks. *Ann Appl Probab* 2004;14(4):1698–1740.
- 80. Adelman D. Price-directed control of a closed logistics queuing network. *Oper Res* 2007;55(6):1022–1038.
- 81. Sevastyanov BA. An ergodic theorem for markov processes and its application to telephone systems with refusals. *Theory Probab Appl* 1957;2(1):104–112.
- 82. Lu Y, Radovanovic A. Asymptotic blocking probabilities in loss networks with subexponential demands. *J Appl Probab* 2007;44(4): 1088–1102. Preprint. Available at: <http://arxiv.org/abs/0708.4059>. Accessed 2014 Sep 30.
- 83. Gallager RG. *Discrete stochastic processes*. Boston (MA): Kluwer Academic Publishers; 1996.

ASSESSING PROBABILITY DISTRIBUTIONS FROM DATA

VICTOR RICHMOND R. JOSE
McDonough School of Business,
Georgetown University,
Washington, D.C.

INTRODUCTION

The task of assessing probabilities for uncertain events can be cognitively challenging even for the most skilled forecasters. This is most evident when one has limited prior experience with such events. The presence of data, however, can ease the burden on the decision maker by providing a good starting point for understanding the nature of the uncertainty, and such data hopefully would lead to accurate and reliable assessments.

In many cases, it is almost impossible to find information directly related to the uncertainty of concern. Some related data, however, are typically available that could be used to improve our understanding of the uncertainty.

For example, consider an engineer who is studying the failure time of a certain component. She is mainly concerned about whether this component will last until some ordered spares arrive, in order to prevent an disruption in the company's operations. She may study data on the same component or a closely related product for which information is available. This information may be used to generate a probability assessment for this component since it seems to be a reasonable assumption that the life of this component would be similar to that of other identical or similarly designed items.

On the other hand, consider an insurance company which routinely assesses the likelihood that a customer would be involved in a motor vehicle accident. It would be a realistic assumption to say that based on certain demographics such as age, occupation,

and driving experience, some individuals are more likely than others to be involved in an accident. Here, the related information can be used to help predict whether or not the customer would be involved in an accident say, in the next year.

In these examples, data become useful in the probability assessment process. In the engineering example, assessments can be made by using a "historical approach," that is, probabilities are assigned to events based on the assumption that past information is representative of current and future behavior. It works on the premise that the earlier events are drawn from a common distribution or process and that the event being assessed will follow as well. In the insurance example, however, it may no longer be plausible to assume that the chance of a particular individual getting into a motor vehicle accident is a simple random draw from the accident history of everyone in the population. Here, additional explanatory variables are used to make inferences about the event probability.

ASSESSMENTS USING HISTORICAL DATA

In the historical data approach, a probability assessment is made by generating a distribution based on past information and using this for analysis related to the uncertainty of interest. Framed as such, a large literature in statistics is available and has been devoted to address this problem.

Parametric Approach

Traditionally, constructing probability distributions from data often starts by considering a subset of probability measures that are indexed by some parameter $\theta \in \mathbb{R}^n$. By considering functions characterized by this parameter θ , the assessment task is somewhat easier since the focus is then simply on estimating appropriate values for θ .

Choosing a Model. Often, problems can be approached in many different ways and

choosing an appropriate model can be somewhat confusing. Though there is no acid test to clearly determine which distribution or parametric family is appropriate in every occasion, research [1] suggests that considering simple questionnaires about the uncertainty (such as the one listed below) can be helpful in limiting the set of distributions that we need to consider.

1. *Nature of the Sample Space.* Is the random variable discrete, continuous or mixed (i.e., can it take only discrete values, is it continuous over some interval(s) or both)? Are we dealing with a univariate or multivariate distribution?
2. *Bounds.* Is the range of values bounded above? Or below? If so, what are reasonable upper and lower bounds?
3. *Shape.* Is the distribution symmetric? If it is skewed, in which direction is it skewed?
4. *Concentration.* What range of values is of primary interest to us? Is the modeling of the tails critical? Is the distribution unimodal? Or multimodal?
5. *Underlying Process and Dependencies.* Is there an underlying process? Do we expect the variable to be dependent on other variables? Are the variables correlated? Are other forms of dependencies known?

Despite the fact that questionnaires like this one can be used to limit the parametric families that we should consider, it often does not single out a particular parametric family. Often, we still have to choose among several distributions.

Working with a Model. For each of the models which we might consider, the next task for the decision maker is to generate the probability assessments that he or she needs. Consider our engineering example: suppose the records show that the last 15 units of the same component used in the plant lasted 12, 14, 15, 6, 9, 10, 11, 12, 14, 13, 13, 10, 14, 9, and 18 days. If the spares are scheduled to arrive in 10 days, then we can use this to estimate the probability that there will be a disruption.

If we assume that the data we have observed comes from independent draws from a common distribution with the same probability of lasting less than 10 days, then we can say that a binomial distribution (i.e., $f(x) = \theta^x(1 - \theta)^{n-x}$) might be appropriate for this. Having decided on a parametric model would then limit the problem to estimating an appropriate value for θ .

Suppose that over a period of time or a series of experiments, you were able to collect a random sample $X_1, \dots, X_n$ coming from a distribution function $F(x; \theta)$, where θ is an unknown parameter belonging to some set $\Theta \subset \mathbb{R}^m (m \geq 1)$. There are many ways to estimate θ . Perhaps the most commonly used approach used today is the maximum likelihood (ML) approach. Under this method, the data observed is assumed to be generated by some joint distribution L (characterized by the parameter θ) called a *likelihood function*

$$L(\theta) = f(x_1, \dots, x_n | \theta). \quad (1)$$

The estimation is then made by finding the $\hat{\theta}$, which maximizes the value of Equation (1), that is,

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta). \quad (2)$$

The estimate in Equation (2) can be interpreted as the parameter that gives the highest probability (or the one that makes it “most likely”) that the data comes from the said process. A large literature has shown many interesting and useful statistical properties for the ML estimator, making it desirable to many researchers and practitioners. However, one challenge still encountered in this approach is that solving Equation (2) for certain functions can be computationally difficult.

Going back to our example, it can be verified that the ML approach would yield an estimate for θ of $12/15 = 80\%$. Intuitively, this makes sense since 3 instances of the 15 were less than the delivery time of 10 days. Alternative estimation techniques which may yield significantly different results, are also available. Lehmann and Casella [2] provide a detailed discussion on different statistical estimation techniques.

Naturally, an alternative model could also make the estimates significantly different. For example, if we assume that the lifetimes of the component follow say, a normal distribution, then the estimate for the mean and standard deviation using a moment-matching approach would be 12 and 2.95 respectively. This would mean that the probability that the component survives before the spare arrives would be $\mathbb{P}(\text{Component Life} \geq 10) = \mathbb{P}(Z \geq -0.678) = 75.1\%$.

An alternative approach to solving for point estimates is computing for the posterior distribution of the parameter θ and using this information to generate an estimate for the parameter. Given prior information on θ encoded in the distribution $f(\theta)$, Bayes' rule allows us to compute the posterior distribution of θ as follows:

$$f(\theta|\mathbf{x}) = \frac{f(x_1, \dots, x_n|\theta) f(\theta)}{\int f(x_1, \dots, x_n|\theta) f(\theta) d\theta}. \quad (3)$$

Typically, in many instances where no information on θ is available, the statistical literature often relies on the use of noninformative (or diffuse) priors. Once a posterior distribution is attained, we can then use this for analysis. Retrospectively, we can also use this to make probabilistic statements about the uncertainty. The posterior mean of θ could be used as an estimate for θ , similar to what the ML approach tries to achieve. Other statistics (such as the median or other specific quantiles) could be used as well, depending on the problem. Compared to the earlier approach, this Bayesian treatment is a more careful analysis about the event uncertainty since the inherent uncertainties in θ is incorporated in the analysis through the prior distribution $f(\theta)$.

Selecting among Models. If only one model is considered appropriate, then there is no ambiguity in which probability estimate to use. However, when multiple models are considered, we often have to choose which one should be used and reported.

The typical approach in practice is to compare models based on their fit with the empirical data. Models are ranked using goodness-of-fit statistics. Some commonly

used statistics include Pearson's chi-squared (χ^2), Kolmogorov–Smirnov (KS), and Anderson–Darling (AD). D'Agostino and Stephens [3] provide a more detailed treatise on this subject.

Current statistical softwares have the ability to fit data into certain parametric families. In addition, users have the option of ranking the distributions based on fit using some common goodness-of-fit statistics. For example, Crystal Ball, a commonly used Monte Carlo simulation package, estimates parameters for distributions using an ML approach and users have the option to rank distributions based on either the χ^2 , KS, or AD statistic. Figure 1 shows a sample output generated by Crystal Ball for a small financial data set.

For these data-fitting techniques, one important issue that often happens in practice is dealing with small data sets. Research [4] has shown that for several well-known distributions, small differences in the estimates have a large effect on decision and risk analysis studies, especially when the uncertainty has a high relative standard deviation. Extra caution is needed when relying on assessments and analysis that are generated from such small data sets.

Nonparametric Approach

Once in a while, an individual may not have sufficient information to select a class of parametric families to use. In these cases, making certain assumptions on the functional form of the distribution might lead to large errors instead of improvements in the model. In a risk analysis study [5], it was shown that incorrect substitutions in functional forms of probability distributions generated errors that can grow by a “factor on the order of $\sqrt{n}$ for a data set of n measurements”.

An alternative to forcing a choice for a parametric model is to allow the empirical data to determine the distribution on its own. In this nonparametric approach, the empirical distribution is used as the probability distribution in the analysis. This practice is commonly done in Monte Carlo simulation for determining some distribution when none of the standard distributions used seem to

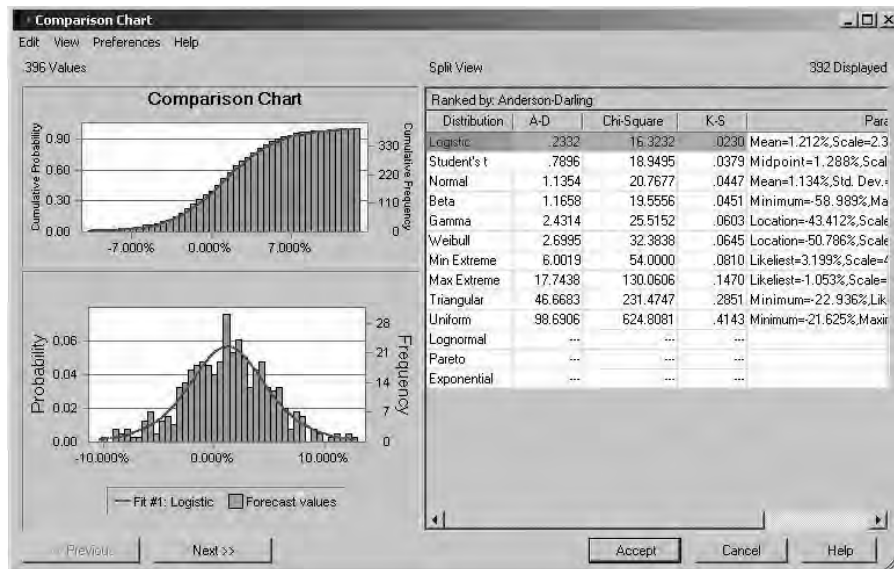


Figure 1. Data-fitting option of Crystal Ball.

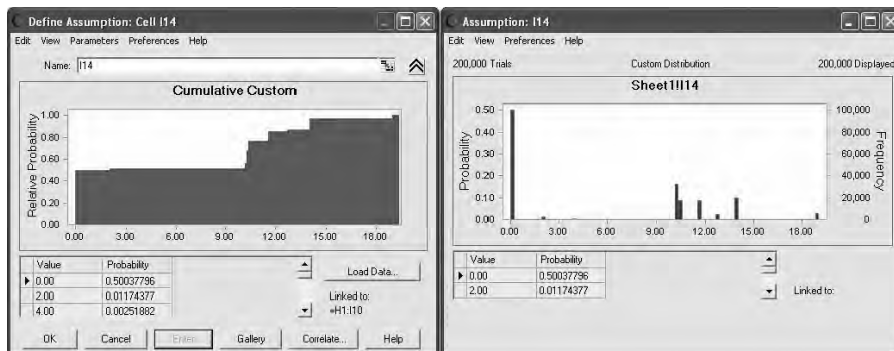


Figure 2. Using empirical distributions in a Monte Carlo setting.

be appropriate. Many Monte Carlo and statistical software packages now allow users to use empirical distributions in their analysis. Figure 2 shows the tool in Crystal Ball that could be used to create a custom distribution.

Though the freedom from choosing a parametric family seems to be a better alternative, the nonparametric approach has its own setback. The main disadvantage of this approach is that the quality and reliability of the analysis exponentially decreases as the number of data points drop. To a lesser extent, tractability may also provide some

problems but with the presence of modern computing this is not a big issue unless the dimension or size of the problem is large.

ASSESSMENTS USING EXPLANATORY VARIABLES

A different and equally valid approach in using data to assess probability distributions is to use historical and current information to predict certain events. For example, consider a firm that has decided to provide a new type of small short-term loans to its customers.

In this case, one useful assessment question that might be asked by this firm is how likely is it that an individual will default a loan if one gets approved.

In this case, historical information on default rates may not be as useful since customers with different backgrounds and profiles may not be likely to default their loans at the same rate. If there is historical data on loans of a similar type, then some analyses could be done to estimate the probability that an individual would default. In particular, information such as age, household income, level of education, amount of credit card debt, and current credit rating score might be useful in predicting the default rates for each individual applicant.

Tools from Categorical Data Analysis

One class of models that could be used to analyze explanatory variables is called *discrete choice models*, which are models that predict what category an uncertain event would materialize as a function of any number of variables that are believed to have an influence on the event.

Binary Models. Consider a Bernoulli random variable Y_i , which depends on a set of explanatory variables $(X_{i1}, \dots, X_{in})$. An initial model that could be considered is a *linear probability model*, which tries to linearly model the probability through the explanatory variables, that is,

$$p_i \equiv E(Y_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in}. \quad (4)$$

Two problems are typical in an approach such as this. First, the estimate that you might get may fall outside the interval $[0,1]$.

The other issue is that the event probability may not be linearly related to the explanatory variable. Typically, you may expect that as you go to the extremities, changes in the probability would start to become smaller. An alternative approach which addresses these issues and has also been popular in the literature is *logistic regression*.

Logistic regressions are models that fit the log-odds ratio (or “logit”) of a binomial random variable as a linear function of the predictor variables $\{X_{ij}\}_{j=1}^n$, that is,

$$\begin{aligned} \text{logit}(p_i) &= \ln\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in}. \end{aligned} \quad (5)$$

Equation (5) is equivalent to

$$p_i = \mathbb{P}(Y_i = 1) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in})}}. \quad (6)$$

The term *logistic* comes from the fact that Equation (6) is similar in form to a logistic curve/function which has the form $f(x) = (1 + \exp(-x))^{-1}$. Figure 3 shows an example comparing a linear probability model with a logistic regression.

To solve for the coefficients in a logistic model, ML estimation is performed. In certain instances the ML estimation procedure can be computationally tedious, especially for

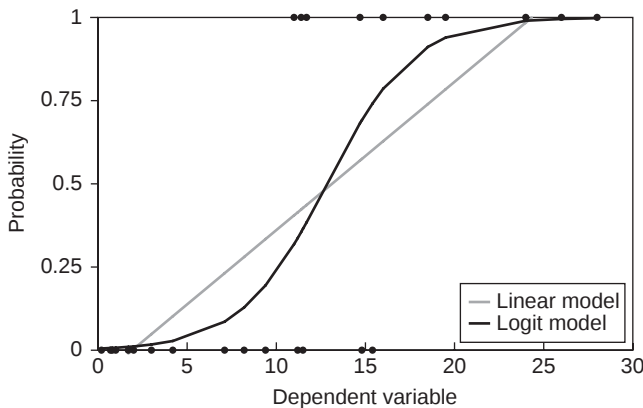


Figure 3. A comparison of the linear probability and logit models.

large problems. However, standard statistical software such as Stata, SAS, and SPSS now have a logistic regression tool built into their standard packages; that does these computations quickly and efficiently.

Multinomial Models. In cases where there are more than two possible states or categories, the binary model can be extended to the multinomial case. If Y_i can take on the values $\{1, 2, \dots, m\}$, then we can write the logit equations as

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j}X_{i1} + \dots + \beta_{nj}X_{in}, \quad (7)$$

for $j \neq i$. Since we also know that $\sum_j p_{ij} = 1$, we only need $m - 1$ equations for Equation (7). Hence, one category (WLOG say state 1) is typically chosen to be the reference state; that is, the log-odds ratio is always compared to the first state making things easier to compare or mathematically,

$$\text{logit}(p_{ij}) = \ln \left[\frac{\mathbb{P}(Y_i = j)}{\mathbb{P}(Y_i = 1)} \right]. \quad (8)$$

Then, Equation (7) can be rewritten as

$$\begin{aligned} p_{i1} &= \mathbb{P}(Y_i = 1) \\ &= \frac{1}{1 + \sum_{k=2}^m \exp(\beta_{0k} + \beta_{1k}X_{i1} + \dots + \beta_{nk}X_{in})}, \end{aligned} \quad (9)$$

$$\begin{aligned} p_{ij} &= \mathbb{P}(Y_i = j) \\ &= \frac{\exp(\beta_{0j} + \beta_{1j}X_{i1} + \dots + \beta_{nj}X_{in})}{1 + \sum_{k=2}^m \exp(\beta_{0k} + \beta_{1k}X_{i1} + \dots + \beta_{nk}X_{in})} \\ &\text{for } j \neq 1. \end{aligned} \quad (10)$$

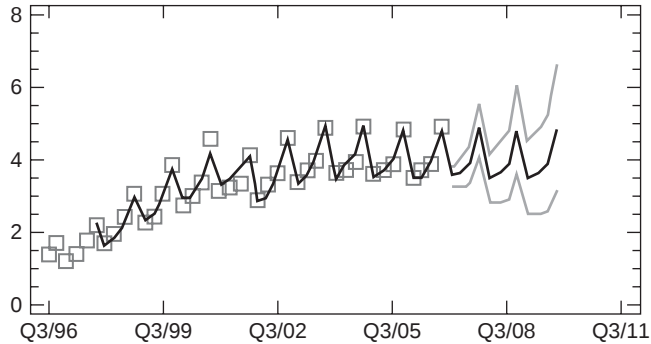
Again, many software packages available today incorporate a multinomial logistic regression tool. In addition, other multinomial logistic models (ordered state space models, conditional logistic models) are available in these packages. An excellent discussion of these more sophisticated models is provided in Refs 6 and 7.

Tools from Time Series Models and Forecasting

Alternatively, we can also use some inferential techniques to generate probability assessments over continuous intervals. Consider a fashion retailer who is trying to provide probabilities for sales for the next few quarters. Given the presence of past data, projections for future revenue streams can be generated say, by using some time series forecasting technique that could be used to provide a predictive distribution for future sales. Figure 4 provides an example of a data series where seasonal patterns are quite evident; draws from historical values might be misleading because of the growth trend that has been experienced by this firm.

Moreover, one may be interested in using other predictors outside the historical data for making predictions. Generalized linear models might be used to predict say, the sales, using other variables such as the advertising

Figure 4. A sample time series forecast output from a statistical package.



budget, lagged sales, market share, economic indicators, and the like. The predictive distribution of the forecast can then be used as the probability distribution for the unknown quantity. Here, the emphasis is not on the point forecast generated by the model but rather the probabilistic information that is contained in the predictive distribution.

SOME CONSIDERATIONS

The use of data can be very useful in determining an appropriate probability distribution for an uncertain event. Historical data about the uncertainty can provide a glimpse of the past behavior of the uncertainties being studied, while related information may provide a good basis for making inferences about these uncertainties.

Despite the many advantages of using data, one must still be cautious of falling into the following traps that are common in this process:

1. *Framing Traps.* Frame blindness refers to the tendency of approaching a problem in an inappropriate manner because one has already accepted a specific framework with little thought leading to the discounting of better alternatives [8]. This narrowness in perspective often blinds decision makers. A common tendency is to lock in on a data set and discard any other theory or model that may reject this data set before doing a reasonable exploration of options. This can be due to a number of reasons such as having an initial cost associated with the data set, the data set being accepted as a common standard in the past, or perhaps its being attributable to a specific individual/source/group.
2. *Illusion of Objectivity.* One particular phenomenon with the use of data is the “illusion of objectivity” [9]. This refers to the distortion in the stakeholder’s understanding of the situation by making him or her believe that the assessments made using the data are purely objective and free from any

input from individuals. In many cases, the data used in the generation of probability assessments are devoid of human inputs but the choice of which data to use and how to model the data (including, for example, distributional assumptions or choice of a model in a regression setting) often involves a decision maker who ascertains why the use of such data over others is appropriate.

3. *Purity of Data.* Related to the notion of objectivity is the “purity of data” phenomenon, where decision makers often eliminate all sources of information that are attributable to individuals. In many instances however, information from individuals can be very useful and insightful. Research has shown that combining forecasts (see **Bayesian Aggregation of Experts’ Forecasts; Combining Forecasts.**) can lead to improvements in forecasting accuracy and this is true regardless of whether the forecasts come from individuals, groups, models or data.

Data can be a powerful tool in providing us more information about uncertainties that we have limited information on. They are useful as long as we keep in mind that data can never be used without a sprinkling of caution and a good dose of critical thought.

REFERENCES

1. Lipton J, Shaw WD, Holmes J, *et al.* Short communication: Selecting input distributions for use in Monte Carlo simulations. *Regul Toxicol Pharm* 1995;21:192–198.
2. Lehmann EL, Casella G. *Theory of point estimation*, Springer texts in statistics. 2nd ed. New York: Springer; 2003.
3. D’Agostino RB, Stephens M. *Statistics: a series of textbooks and monographs*. Volume 68, Goodness-of-fit techniques: New York: Marcel Dekker; 1986.
4. Haas C. Importance of distributional form in characterizing inputs in Monte Carlo risk assessments. *Risk Anal* 1997;17(1):107–113.
5. Seiler FA, Alvarez JL. On the selection of distributions for stochastic variables. *Risk Anal* 1996;16(1):5–18.

6. Agresti A. Categorical data analysis. 2nd ed. Wiley series in probability and statistics. Hoboken (NJ): Wiley-Interscience; 2002.
7. Hosmer DW, Lemeshow S. Applied logistic regression. Wiley series in probability and statistics. Hoboken (NJ): Wiley-Interscience; 2000.
8. Russo JE, Schoemaker PJH. Decision traps: ten barriers to brilliant decision making and how to overcome them. New York: Doubleday; 1989.
9. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Sci* 1988; 76(2):159–165.
- Greene WH. Econometric analysis. 6th ed. New Jersey: Prentice Hall; 2007.
- Hamed M, Bedient P. On the effect of probability distributions of input variables in public health risk assessment. *Risk Anal* 1997;17(1):97–105.
- Hamilton JD. Time series analysis. New Jersey: Princeton University Press; 1994.
- Hattis D, Burmaster DE. Assessment of variability and uncertainty distributions for practical risk analysis. *Risk Anal* 1994;14(5):713–729.
- Kleinbaum DG, Klein M. Logistic regression. 2nd ed. New York: Springer; 2005.
- Lloyd CJ. Statistical analysis of categorical data. New York: Wiley Interscience; 1999.
- Thompson KM. Software review of distribution fitting programs: Crystal Ball and BestFit Add-In to @RISK. *Hum Ecol Risk Assess* 1999; 5(3):501–508.

FURTHER READING

Frey HC, Burmaster DE. Methods for characterizing variability and uncertainty: comparison of bootstrap simulation and likelihood-based approaches. *Risk Anal* 1999;19(1):109–130.

ASSESSING REMAINING USEFUL LIFETIME OF PRODUCTS

SAMI KARA

Life Cycle Engineering and Management
Research Group School of Mechanical &
Manufacturing Engineering, The
University of New South Wales, Sydney,
NSW, Australia

INTRODUCTION

The global society has experienced a growing awareness in the last couple of decade to develop a sustainable society. The two main drivers of this awareness has been the increased problem of resource scarcity and the environmental impact. Manufacturing industry has been the most affected from this due to the fact that they consume the largest amount of natural resources and by far the largest contributor to the environmental impact. As a result, manufacturing industries around the world have been investigating the possibilities of “doing more with less” in order to reduce the resource consumption and the associated environmental impact. A number of strategies have been recently developed to achieve this. One of these strategies is shifting manufacturing focus from selling product to selling service so that the manufacturing companies can keep the ownership of the product from cradle-to-grave [1], and therefore achieving multiple use of out selected components and subassemblies. For instance, companies such as Fuji-Xerox manage to achieve up to seven life-times out of same components. The other strategy is to establish a closed-loop manufacturing system [1]. The underlying idea is that the end-of-life (EOL) products are not waste, but are valuable resources. They are responsible in sustainability development is therefore expanded to include manufacturing firms for taking back their used products for EOL product recovery [2]. Both of these

strategies are mutually supportive of each other, since they both encourage extension of product life hence require EOL decision making at the end of their life-cycle.

Whether it is for selling use or EOL product recovery, EOL decision making consists of three important options such as reuse, recycle, and remanufacture [3]. Among these options, reuse of components is, in many cases, an environmentally effective and economically efficient EOL option [4]. The concept of component reuse stems from the fact that many components have a design life that exceeds the life expectancy of the product itself. For example, Mazhar *et al.* [5] provided empirical evidence that the physical life of an electric motor often exceeds the life expectancy of a washing machine. There are similar findings that have been reported for other products such as one-time-use cameras [6] and fuser roller of photocopiers [7]. Extending the useful life of these components can help minimize the waste and the resource consumption associated with making a whole new product.

While some companies deem reuse of components as a helpful tool for creating a close-loop economy, many of them see the implementation of the strategy as a challenge. One of the major obstacles is the lack of information on the remaining useful life of products. This information is crucial for determining their reuse potential which would justify the subsequent disassembly and reuse decision [8,9]. In addition, the remaining useful life has always been an important issue for maintenance management [2]. From the physical lifetime perspective, the remaining life of a system is defined in terms of physical condition, that is, how long an item will continue to perform its intended functions within a specified usage environment [10]. There have been many studies focused on the remaining physical life assessment such as by using accelerated life testing, predictive maintenance, and nondestructive testing [8]. However, the available methods are either relatively

complex, involving the time-consuming processes such as disassembly and testing of an individual component, or sometimes not applicable to the stationary components.

In addition, further studies reveal that the end of useful life of an item is also subject to the technology improvement [11–13]. As technology improves, there is a chance that underlying components may not be reused. The significance of change can be completely or partially dependent on whether the technology is improving at breakthrough or at incremental progress. However, regardless of the degree of change, technological progress has an impact on market demand of a product and its components. The faster and the greater the penetration of a new innovation, the sooner the old product technology becomes obsolete. Consequently, the remaining useful life of component is not only governed by its physical life, but by its technological life as well [11]. As a result, both these factors should be taken into consideration before an EOL decision is made.

ASSESSING REMAINING USEFUL LIFETIME

The following Fig. 1 shows an integrated model for estimating the remaining useful life of components.

As shown in Fig. 1, the first step in estimating the remaining useful life is the assessment of physical lifetime, T_P , which is the difference between the operating life, T_O , and the usage life, T_U , of the component. As mentioned earlier, the remaining useful life of components is governed not only by the physical, but also the technology life as well. Therefore, the second step in this process is the assessment of the technology life, T_T . Finally, the remaining useful life, T_R , can be estimated as the minimum of these values. The following section will explain these concepts in detail.

ASSESSMENT OF PHYSICAL LIFETIME

Estimation of Operating Life, T_O

The parameter, T_O , is governed by the failure datum, which is the time that a component is

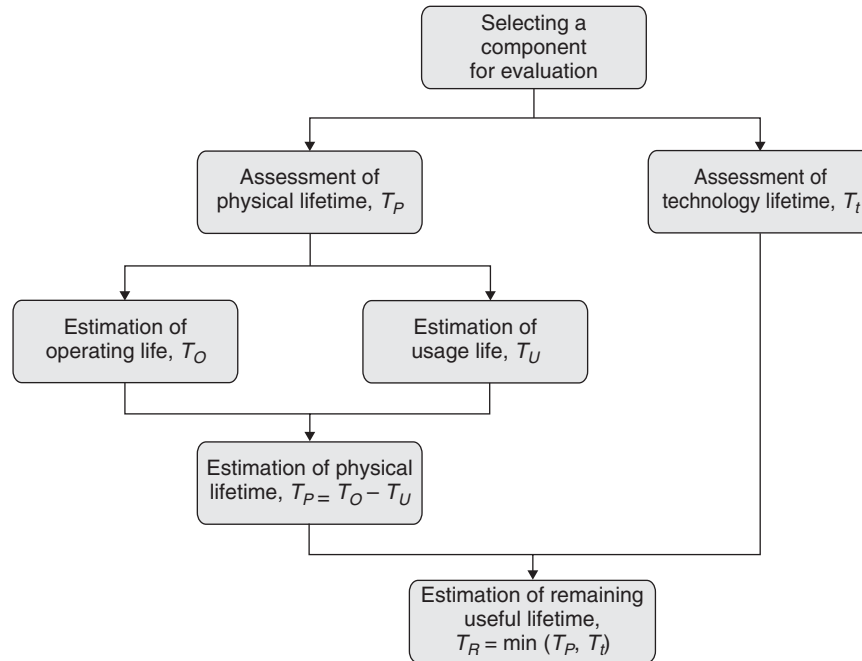


Figure 1. Integrated methodology for the remaining useful lifetime estimation.

expected to experience a failure or disruption so that it cannot resume any of its normal operations. The failure rate of an item is the standard metric for reliability predictions, calculated as the average number of failures per unit of time [14]. For many electronic components, it is possible to consider their failure rate to be constant and characterized by the exponential distribution [15]. However, it is more common to find most of the component failures follow a typical failure rate curve. The “bathtub curve” is often mentioned in this context. However, other curves are more relevant in practice [15]. The failure rate can be characterized by the Weibull distribution [16].

The information on failure rate and the associated probability distribution can be used to derive the time-to-failure (the operating life expectancy) of a particular component [16]. The mean-time-to-failure (MTTF) is a widely used statistical description that specifies the average life or the most likely value to be expected in a group of data [16]. Nevertheless, it has been recommended that the MTTF is to be used in conjunction with the corresponding life parameters (e.g., Weibull parameters) in order to represent meaningful characteristics and uncertainty of a component’s lifetime [16].

As an alternative to the physical testing approach, the field failure data analysis has also been extensively used to assess the life characteristics of components [17,18]. The main advantage of this technique is that it accounts for all the actual field environments and the possible failure mechanisms. The calculated MTTF can be a closer value to the true life expectancy of the component population. Moreover, most companies already have access to the information since it is usually an integral part of their quality control program. There is no new investment required.

The collected data can be analyzed using the Weibull method to determine the life characteristic parameters including the MTTF of the component. Its most common form is the two-parameter Weibull distribution [16]:

$$F(t) = 1 - \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right], \quad (1)$$

where $F(t)$ represents the fraction of units failing and “ t ” is the time-to-failure. The two parameters that characterize the life distribution are the scale parameter, η , which measures the time at which 63.2% population will fail, and the shape parameter, β , which identifies the mode of failure. A value of $\beta < 1$ indicates infant mortality, $\beta = 1$ means a random failure, and $\beta > 1$ describes a wear-out failure. The corresponding MTTF, of the two-parameter Weibull function can be expressed in terms of its scale parameter, η , and the gamma function, Γ , of the shape parameter, β as [16]:

$$\text{MTTF} = \eta \times \Gamma \left[\frac{(\beta + 1)}{\beta} \right]. \quad (2)$$

The most common method used for estimating parameters in Weibull analysis are median rank regression (MRR) and maximum likelihood estimation (MLE). The results provide the estimated values of life distribution parameters η and β , and the corresponding MTTF of the component, hence T_O for a given component [17,18].

Estimation of Usage Life, T_U

The usage life of a product, T_U , is the second element to be considered in estimating the remaining physical lifetime of a component. It is a critical parameter that indicates the actual age measured in operating times (hours, cycles, kilometers, and so on) that a component has been used. One way to determine the age is by measuring in calendar time units (i.e., days, weeks, months, and years) and/or assume a constant usage intensity throughout its lifespan [12,19]. However, this would not produce a realistic estimate, therefore, a number of methodologies have been developed to carry out a more realistic assessment of this critical parameter. Majority of these methodologies utilize life-cycle data collected from various sources. The information on the variation in usage intensity overtime of a product can be collected through a number of ways, for example, using data recording unit in consumer products; such as electronic data log (EDL) [20], life-cycle data acquisition unit—the whitebox [21], life-cycle unit [22], and watchdog [23], or using field surveys.

Statistical Techniques. The basic idea behind this technique is that the usage conditions and level of intensity that may change and can depend on the obsolescence of a product with respect to time. As a product ages, its usage-intensity is expected to be less. For example, the frequency and duration with which a 10-year-old television is used may be less than a 2-year-old television in a household. This may be due to the attractiveness of the new design as well as the superior features and performance of the new television that has over the 10-year-old with malfunctions and inefficient energy consumption.

The main aim is to compile locally designed basic statistics on product usage intensity. The results can then be used to estimate the usage history and its changing pattern in relation to the age of a product. In order to do this, the collected data can be statistically analyzed to derive the following statistical information:

1. the probability function that describes the probability of usage intensity during a particular age of product usage (in year);
2. the average usage intensity during a particular age of product usage, called $\bar{X}_i$ where “ i ” is the order of year 1 to n .

For instance, if it is assumed that a product is likely to spend $\bar{X}_1$ h per day in active

mode during the first year, $\bar{X}_2$ h per day during the second year, $\bar{X}_3$ h per day during the third year, and so forth until $\bar{X}_n$ during the n th year of its service life. The average total hours that a product has been used over n years period, that is, usage-intensity age, T_U , can be calculated as:

$$T_U = \sum_{i=1}^n \bar{X}_i \times 365. \quad (3)$$

Note that the T_U of the components is simply equal to the T_U of the product itself [24,25].

Regression Analysis. In order to use this technique, a data trend analysis needs to be carried out. In the case of a strong trend of data, regression analysis is a useful technique, particularly because of its ability to provide an answer on how each of the explanatory variables contributes toward the estimation of the response variable. The following Fig. 2 shows life-cycle data collected from an electric motor of a washing machine during its lifetime. They show a strong data trend in between the age of the electric motor, speed, and temperature, therefore the regression analysis is particularly suitable in this case.

In its generic form,

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k, \quad (4)$$

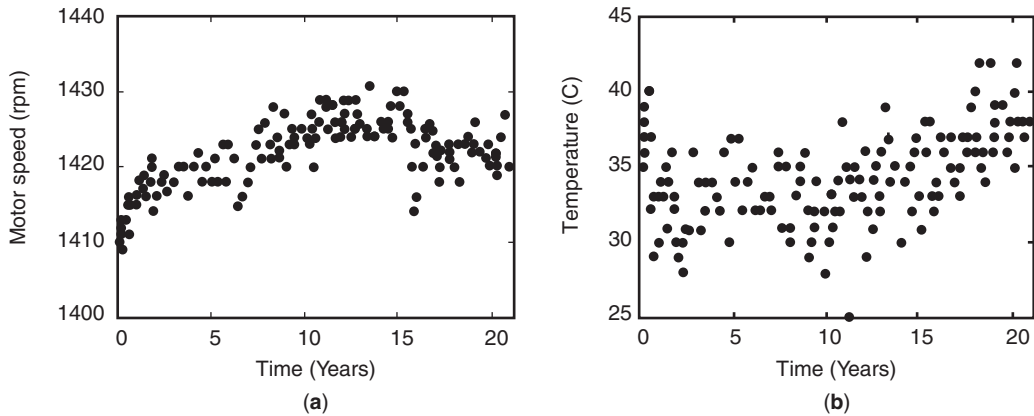


Figure 2. Life-cycle data for an electric motor [26,27].

where Y is the response variable, a and b are regression coefficients, $X_1, X_2, \dots, X_k$ represent the explanatory variables and k is the total number of explanatory variables. There are different types of regression analysis, such as simple and multiple regressions. Simple regression involves two variables, the dependent variable and one independent variable. Multiple regression involves many variables, one dependent variables and many independent variables. The Equation (4) is valid for simple and multiple regressions, whereas the quadratic and higher degree regression equations need to be solved by nonlinear regression analysis.

The purpose of carrying out regression analysis is to know the explanatory variables, also called *predictors* or *independent variables*, such as motor speed and temperature, are related to the response (dependent) variable which is the age of the product, an electric motor, in this case. The primary goal is to estimate or predict machine's life given the current and past values of the explanatory variables. The results of an analysis for the electric motor are shown in Table 1. The effect of each explanatory parameter on the overall result, based on the individual R^2 values, is shown in Table 2.

Table 1. Results of Linear Multiple Regression Analysis [28]

Response Variable	T_U (years)
Explanatory variables	ω, T, P, I, V
Regression equation	$T_U = 10 \times I + 0.81 \times \omega + 0.0606 \times P + 0.0095 \times T - 0.111 \times V - 1152.9$
R^2	0.8232

Table 2. Explanatory Variables Contribution to R^2 [28]

Speed (ω)	Power (P)	Temperature (T)	Voltage (V)	Current (I)
0.8005	0.0221	0.0000	0.0006	0.0000

It can be seen that motor rotation speed (ω) and power (P) are the dominating parameters in the final result. This fact has been confirmed by a very clear rising trend in the monitored values of speed, and power measurements are found reasonably related to the age of the motor during the initial years of operation. No or negligible contributions by voltage (V), current (I), and winding temperature (T) have been found due to a flat response from these parameters over the entire life of the electric motor. These techniques are proved to be inadequate when there is no, reverse, or fluctuating trend in the observed data.

Kriging Techniques. Kriging techniques are modified forms of multiple linear regression. They are often associated with the acronym BLUE Best Linear Unbiased Estimator [27]. Ordinary kriging, also called *punctual kriging*, is the simplest and most common form of kriging. It is suitable for data that shows weak trends. Universal kriging is similar to ordinary kriging but used when a strong trend is present in the data samples. Cokriging is a multivariable extension of kriging. It is capable of estimating one variable from several variables by utilizing the concept of estimating primary and secondary variables at the same time.

Kriging utilizes the information from a variogram to find an optimal set of weight that are used in estimating an unknown value at know condition. A variogram as shown in Fig. 3, summarizes the relationship between differences in pairs of measurement and the distance of the corresponding points from each other.

Sill and Range are defined at the region where the variance levels off (the Sill) after a certain distance (the Range) beyond which observations appear independent, that is, where the variance no longer increases. Nugget is defined when the variogram appears not to go through the origin. There are several different models of variograms including linear, exponential, and spherical and the Gaussian models. The spherical model is widely used and considered as an ideal model since it accounts for all three parameters such as Sill, Range, and Nugget.

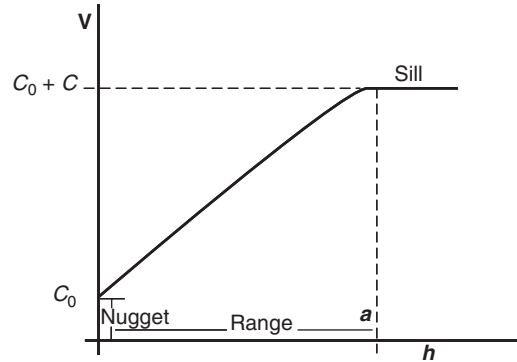


Figure 3. Variogram with Nugget, Sill, and Range [27].

The variogram is computed by measuring the mean-squared difference of a value of interest evaluated at two points, for example, x and $x + h$. This mean-squared difference is the semivariance, v , and is assigned to the value h , which is known as the lag. A plot of the semivariance, v , versus h is the variogram.

Mathematically,

$$v = \begin{cases} C + C \left[1.5 \left(\frac{h}{a} \right) - 0.5 \left(\frac{h}{a} \right)^3 \right], & h < a, \\ C_0 + C, & h > a, \end{cases} \quad (5)$$

where C_0 = Nugget, $C + C_0$ = Sill, a = Range, h = Distance, and v = Semivariance.

When a variogram is used to describe the correlation of different variables it is called *cross-variogram*, which is used in cokriging.

Table 3 below shows the summary of the results of the different kriging techniques for the previous electric motor case.

There are only small differences between the values estimated by universal kriging and Cokriging, in this case. Kriging procedures have a major problem of having singular matrices when there are a lot of data sets with the same data value. This scenario results in a large number of zero entries in the distance

matrix that makes the determinant of the matrix equal to zero. In other words, kriging procedures are more suitable for data sets with distinct data points.

Vibration Analysis. Vibration analysis can extend the knowledge of the machine condition and thus improve the remaining useful life estimate. The vibration analysis is also useful since the suddenly occurring impulse reveals an obvious degradation in a product's performance. Hence the goal of the vibration analysis is to process each vibration signal and extract a set of suitable trending indicators. Later, these trending indicators can be fitted to a suitable model by using one of the techniques mentioned above in order to predict the useful remaining life of products. Although trending can be done by using three indicators, namely, comparing the constant percentage bandwidth (CPB), cepstrum, and envelope analyses, two of those, CPB and cepstrum analysis will be discussed here. Further readings can be found at Refs 29 and 30.

CPB spectra is a robust and versatile fault detection method. Measurement of CPB spectra has been developed even before the fast fourier transformation (FFT) algorithm

Table 3. Results of Kriging Techniques [28]

Response variable	Ordinary Kriging		Universal Kriging	Cokriging
	Years			
Explanatory variables	ω, T	T, P	ω, T	ω, T, P, I, V
R ²	0.80	0.30	0.8592	0.8592

was implemented in vibration analyzers with octave-band filters [29,31]. In order to obtain degradation trend for a given component, a reference spectrum $R[f]$ is generated from the first four CPB spectra by finding a maximum magnitude in each band. The reference spectrum represents the initial state of a product without mechanical faults. Then, a mask $M[f]$ is created by a lateral shift (“widening”) of the reference spectrum $R[f]$ in order to compensate for large speed changes (greater than the percentage bandwidth). The widening is achieved by taking a maximum value of three adjacent frequency components for each center frequency. The mask is then compared with subsequent spectra, and any protrusions above the mask are recorded for trending the fault development. Amplitudes in dB are compared since equal changes in severity are represented by equal changes on a log amplitude scale [29,31]. A significant change corresponds to an increase by 6 dB (ratio of 2:1), while a serious change corresponds to 20 dB (ratio 10:1). This procedure proved to be more reliable than absolute criteria for vibration severity [29,31]. After this, a sequence of difference spectra is computed as

$$Dn[f] = Xn[f] - M[f], \quad n = 5, 6, \dots, N, \quad (6)$$

where $Dn[f]$ is the n th difference spectrum in dB, $Xn[f]$ is the n th CPB spectrum, $M[f]$ is the mask of $R[f]$ and N is the total number of CPB spectra. Only positive differences are of practical interest, thus all negative differences are replaced by 0 dB.

Figure 4a illustrates the slice at 10,293 Hz with an exponential trend exceeding the threshold of 10 dB at day 501, end-of-life of an electric motor T_{EOL} , during a condition monitoring experiment of an electric motor. To attempt prognosis, the trend was computed only from the data up to day 490. The subsequent rapid decline is a consequence of a step change characterized by much lower vibration level at many frequencies. Figure 4b shows the CPB differences at the adjacent center frequency of 9715 Hz. The data exhibit a similar exponential trend reaching the value of 9.3 dB, hence very close to the 10-dB limit. In addition, both trends

exceeded the threshold of 6 dB (indicator of a significant change) almost simultaneously, near the day 425. This suggests that it is the overall trend that quantifies the machine condition, rather than the individual data points, which fluctuate significantly. Therefore, analyzing the CPB differences in a significant band (around 10 kHz in this case) has provided a fault prognosis. Although different curve fitting techniques can be used, two curve fitting methods are shown in Fig. 4 for extracting the trend of each indicator. First, a straight line is fitted by finding a first-order polynomial in the least squares sense. Secondly, an exponential curve $y = Ae^{t+B}$ is fitted by optimizing its parameters using the simplex direct search method. Finally, the method yielding a lower sum of squares error (SSE) can be selected for plotting the trend.

Finally, a CDS is calculated as:

$$CDS(n) = 10 \log_{10} \left[\sum_{f=0}^{f_{\max}} 10^{Dn[f]/10} \right], \quad n = 5, 6, \dots, N, \quad (7)$$

where $CDS[n]$ is the dB increase of the n th spectrum $Xn[f]$ above the mask $M[f]$. The CDS is computed by converting each $Dn[f]$ from dB to the amplitude squared scale, then summing the squared ratio values (integrating over all frequencies) and finally converting the resulting scalar to dB. For this reason, the CDS is a measure of the overall spectrum change, since it quantifies the total increase at all frequencies. This is because the effect of the transfer function is removed, thus allowing to monitor significant changes regardless of the frequency of occurrence. Since the outcome is linear (Fig. 4c), the interpretation is much easier than the others.

The cepstrum analysis can also be used in fault diagnostics for its ability to enhance periodic spectrum structures (harmonics, sidebands) in the same way as spectrum highlights periodicity in time signals [31]. It has been applied to diagnostics of gears [43], bearings, and turbomachinery [31]. The cepstrum is defined as the inverse Fourier transform of a logarithmic spectrum. It is a nonlinear method due to the use

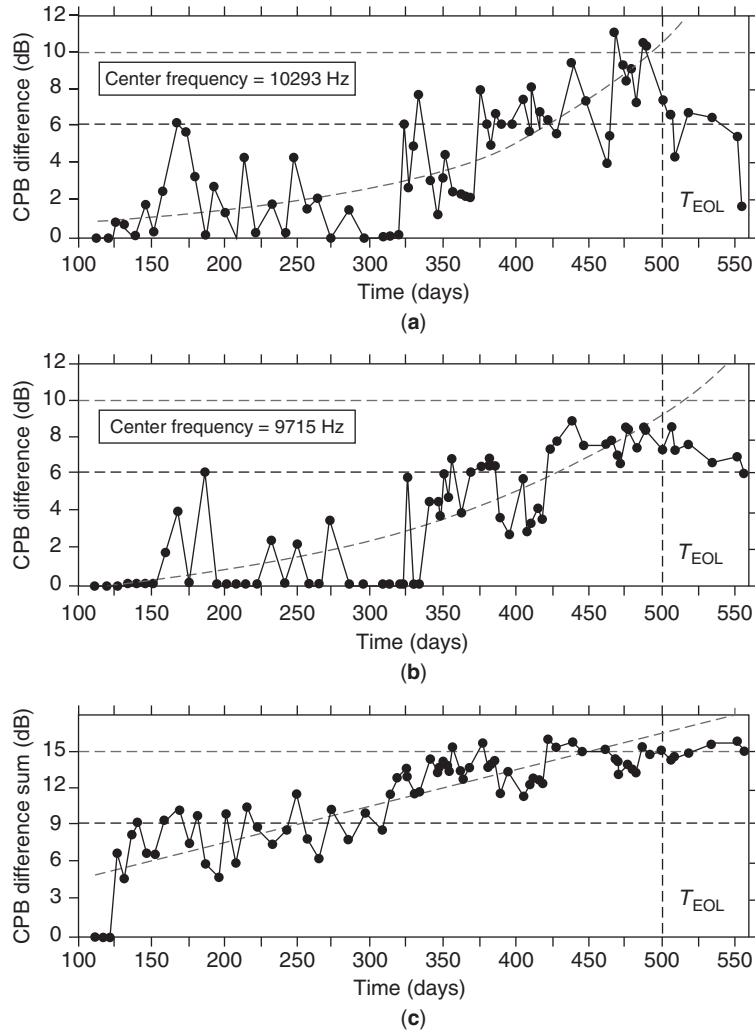


Figure 4. CPB-based trending of electric motors: (a) CPB differences at the center frequency 10,293 Hz, (b) CPB differences at center frequency 9715 Hz, (c) CPB difference sum (CDS) [29,30].

of the logarithm (which determines many cepstrum properties). Two main types of cepstra are commonly used. The real cepstrum is obtained by the inverse DFT of the logarithmic amplitude spectrum, hence it contains no phase information and is used only for analysis. The complex cepstrum is computed by the inverse DFT of the complex logarithmic spectrum. The phase information is retained, thus the complex cepstrum is invertible and allows reconstruction of the time-domain signal.

The following Fig. 5 shows a degradation trend developed for an electric motor by using the cepstrum analysis at 10 ms frequency component. The time axis starts on day 340 in order to show the end-of-life the product. The data exhibits an almost constant trend, while a dramatic increase of 20 dB occurs at T_{EOL} , which indicates the end-of-life the electric motor.

At this point, any previously mentioned linear regression techniques can be used to develop a regression model to predict the

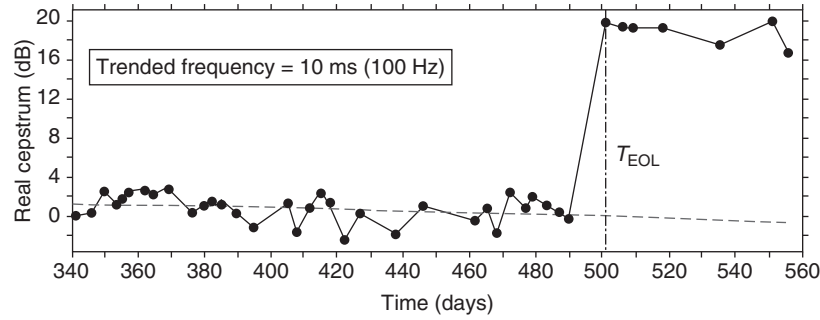


Figure 5. Trending of the real cepstrum at frequency of 10 ms for the electric motor [29,30].

remaining useful life of a component under a given usage condition.

In this section, two possible vibration indicators are used for estimating the remaining useful life of components. However, vibration parameters can fluctuate to some extent, possibly due to load changes caused by the usage condition of the electric motor. Despite this fluctuation, the proposed indicators can be useful for estimating the remaining life components. When the data exhibit a strong trend, the remaining life components can be estimated by fitting a straight line (or exponential curve) and extrapolating the trend to the thresholds of 6 or 10 dB. A more advanced alternative would be to use multiple linear regressions of several vibration parameters. However, as mentioned earlier, these techniques tend to struggle when the trend is weak or fluctuating hence a more accurate model is necessary. The next section will discuss one of these models based on neural networks because of their ability to deal with complex trends and to capture nonlinear input–output relationships.

Artificial Neural Networks. The traditional techniques have been found struggling in certain areas if there are parametric trend fluctuations and a high number of inputs. Research indicates that a neural network is a powerful data-modeling tool that is able to manipulate complex input/output relationships in order to cope with the shortcomings of the previous techniques.

The following Fig. 6 shows an example of neural network model to estimate the remaining useful life of a product by detecting the complex input structures in time series, which is presented to the model network. The features, input, and output of the proposed neural network model are as follows:

- One input layer of 5 neurons—one for each measured variable.
- One hidden layer of 8 neurons.
- One output layer of 1 output neuron.
- Input variables: rpm, temperature, power, current, and voltage.

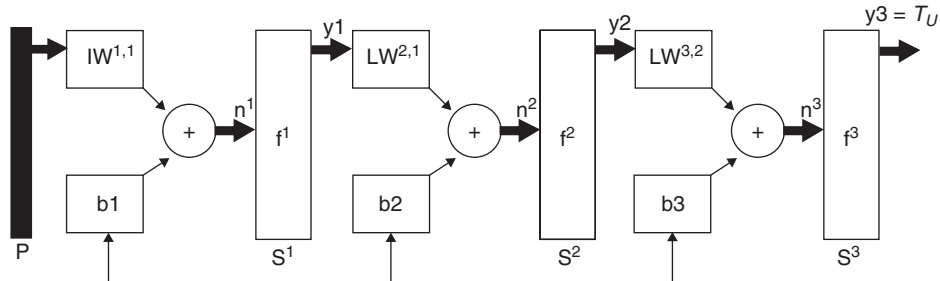


Figure 6. The structure of the neural network model [26].

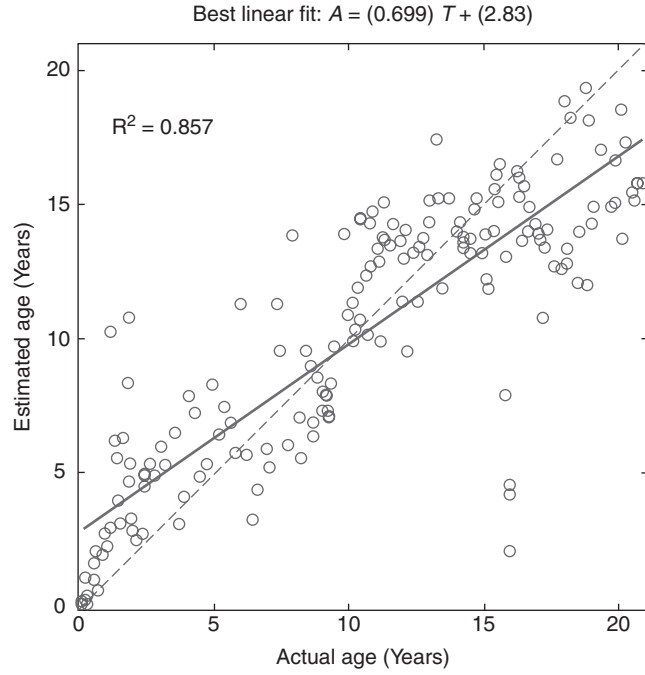


Figure 7. Neural network performance after training [26].

The model's output was obtained and compared to the experimentally measured values by using the postreg function. This function basically performs linear regression between targets (experimentally measured values) and the network response to the presented inputs (condition monitoring data). The model's response to all the three sets of life-cycle data was remarkably accurate especially for the seen and entire data sets as shown in Fig. 7.

Although the R^2 value is still relatively low, once trained, the neural network model yields outputs very closely related to the desired outputs.

Estimation of Remaining Physical Life

As shown in Fig. 1, the estimation of physical lifetime (T_P) requires two variables, namely, operating life T_O , and usage life T_U . The first element, T_O , is an estimated operating life expectancy, which is governed by the failure datum or the expected time-to-failure. The section titled "Estimation of Operating Life, T_O " above explains how this parameter can be obtained. The second element, T_U , is the usage life that represents the true age or used

life of components. The section titled "Estimation of Usage Life, T_U " above explains how these parameters can be estimated by using various methodologies. When the values of the two parameters are obtained, the remaining physical life, T_P , of a component is then calculated as a difference between T_O and T_U . This can be described mathematically as

$$T_P = T_O - T_U. \quad (8)$$

REMAINING TECHNOLOGY LIFETIME ESTIMATE

As mentioned before, the remaining useful life of components is not only governed by its physical life, but by its technological life as well. In order to have a realistic estimate, the technology life of a product must be taken into account, which requires the forecasting of the technology of a given product. Growth curves are commonly used in forecasting the adoption of new technology and hence the obsolescence of its predecessors [32]. This is because the life of a particular product's technology is not random, rather, it forms an

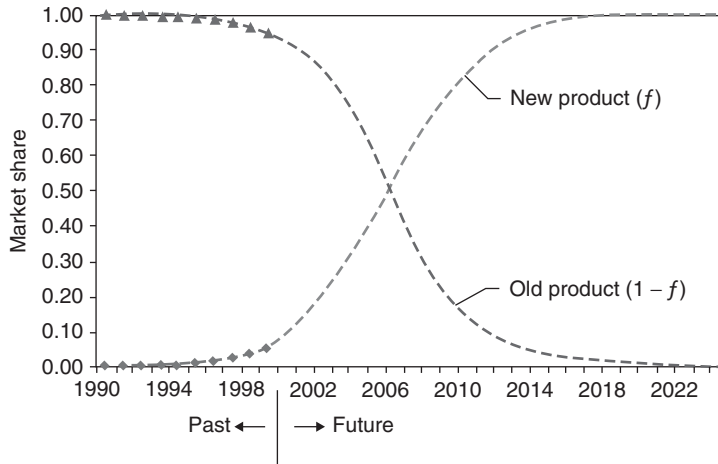


Figure 8. Hypothetical example of product technology substitution [24,25].

S-shaped curve. This enables a mathematical model to be used for forecasting [33].

Market demand is often observed as a related parameter that both captures and drives the technology innovations [34]. The market fraction, f , of a product technology generally increases with time along the S-shaped curve. Similarly, the decay in the fraction represented by the old product generation described a reversed S-curve, as shown in Fig. 8. Therefore, when the historical data is available, a partly established curve can be fitted to the data from its emergence (or the earliest data point) to the present day. It can be assumed that it will continue to grow along an S-curve, and as a result the historical trend will therefore be extrapolated into the future.

There are a number of S-shaped curves which are examined in the forecasting literature and can be used including the Fisher–Pry model [34], Blackman model [35], Floyd model [36], Gompertz curve [37], and the Weibull distribution function [38]. Note that both the most popular models, Fisher–Pry model and Blackman model, are actually different forms of Pearl curves or the standard logistic function (SLF) [39]. However, there is still some debate as to the accuracy and the superiority among all the models. It was observed that, in general, the difference in accuracy and disparity of these models are not greatly significant. It also varies from case to case [40,41]. Therefore,

the simple and the widely used SLF can be used to forecast the technological lifetime of products and their components for the estimation of technology life.

The SLF follows the logistic law of growth that assumes an exponential growth until an upper inherent limit, L , in the system is approached, as shown in Fig. 9.

Mathematically, an SLF can be defined as [39]:

$$p = \frac{L}{1 + ae^{-bt}}, \quad (9)$$

where p is the value of the measuring parameter, which is thus replaced by market fraction, f ; t is the time unit in years; L is the

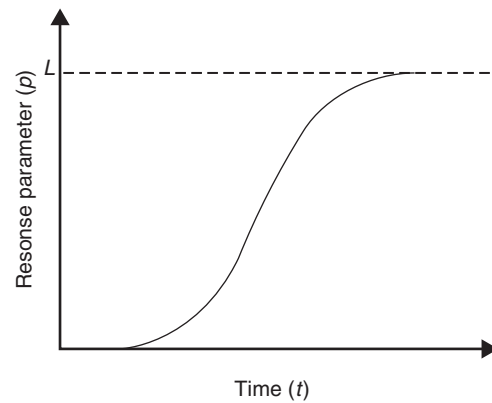


Figure 9. Standard logistic curve (S-shaped curve).

natural limit; a is a constant scale parameter; and b is a shape parameter. Note that the value of a yields p when t is zero and b adjusts how quickly the response changes with changing t as a single unit.

So far, the technology forecast is referred to in terms of market share. How does this predict the remaining technology life? As the expected result indicates the future market share of technologies, a manufacturer needs to identify the market share datum that defines the technological life limit, T_L . This value depends on a manufacturer's individual perspective on the balance between supply and demand of components [42]. For example, a manufacturing company decides that the product under investigation will be obsolete when the demand drops to, say, 20%. A manufacturer makes this individual decision based upon relevant facts and information, including the economical capability.

As a general rule, the remaining technological life, T_t , of a technology and its components is estimated as the outstanding period from present, T_{now} , to the point in time where the demand is predicted to reach the given market share datum, T_L as shown in Fig. 10.

This relationship can be shown mathematically as:

$$T_t = T_L - T_{\text{now}}.$$

Note that the T_t of the components that belong to the old technology is the same as the T_t of the old technology itself. However, the components which appear to remain

applicable in the new technology have their remaining life extended accordingly.

INTEGRATED MODEL FOR THE REMAINING USEFUL LIFETIME ESTIMATION

The concept behind the integrated assessment is that when considering components for reuse, the components should have adequate remaining life in terms of both physical and technological aspects. The integrated model shown in Fig. 1 aims to achieve this. For this reason, the resulting T_R is governed by the shorter value of the two parameters. For instance, if T_P is less than T_t , then T_R cannot be greater and will be equal to the value of T_P itself and vice versa. This relationship can be mathematically written as

$$T_R = \min(T_P, T_t). \quad (10)$$

Furthermore, in order to determine the reuse potential of used components, their remaining life, T_R , should also be at least greater than the average expected service life of a product, T_{AV} , that is, $T_R \geq T_{AV}$.

CONCLUSIONS

An accurate prediction of the remaining useful life has become increasingly important in order to make an appropriate EOL decision for products or components. This is

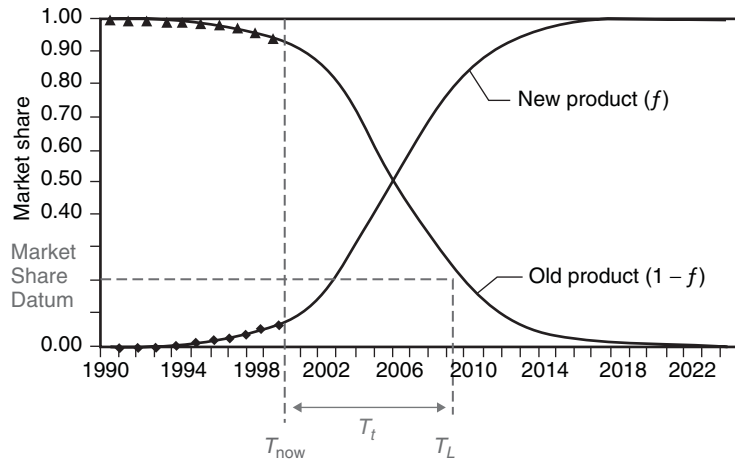


Figure 10. An estimation for the remaining technology life, T_t [24,25].

mainly due to the worldwide drivers such as the increased problem of resource scarcity and the environmental impact. However, the remaining useful life assessment is a complex process due to the uncertainties associated with the condition of the returned product. Therefore an accurate prediction must take into account not only the remaining physical, but also the technology life as well.

In order to achieve this, this article presents an integrated model to estimate the remaining useful life of components. The model integrates the physical life time and the technology life estimation by using various methodologies. This study provides a very comprehensive foundation for further research and decision making on value recovery through reuse. More research into the development of prognostic tools is required in order to develop more reliable EOL decision making tools.

REFERENCES

1. Kaebernick H, Ibbotson S, Kara S. Cradle-to-cradle manufacturing. In: Newton P, editor. Pathways towards sustainable urban development. Collingwood: CSIRO Publishing; 2007. pp. 521–537.
2. Ploog M, Stötting W, Schröter M, *et al.* Efficient closure of material and component loops—substance flow oriented supply chain management. In: Wagner B,ENZLER S, editors. Material flow management—improving cost efficiency and environmental performance. Heidelberg: Physica; 2005. pp. 159–195.
3. Rose CM, Ishii K. Product end-of-life strategy categorisation design tool. *J Electron Manuf* 1999;9(1):41–51.
4. Devoldere T, Dewulf W, Willems B, *et al.* The eco-efficiency of reuse centres critically explored - The washing machine case. Proceedings of the 13th CIRP International Conference on Life Cycle Engineering; Leuven, Belgium; 2006. pp. 219–226.
5. Mazhar MI, Kara S, Kaebernick H. Reuse potential of used parts in consumer products: Assessment with Weibull analysis. Proceedings of the 11th CIRP International Life Cycle Engineering Seminar on Product Life Cycle—Quality Management; Belgrade, Serbia; 2004. pp. 211–216.
6. Fukano A. QuickSnap reusing & recycling system. EcoDesign'99: 1st International Symposium on Environmentally Conscious Design and Inverse Manufacturing; Tokyo, Japan; 1999. pp. 975–978.
7. Kerr W, Ryan C. Eco-efficiency gains from remanufacturing: a case study of photocopier remanufacturing at Fuji Xerox Australia. *J Cleaner Prod* 2001;9(1):75–81.
8. Kara S, Mazhar MI, Kaebernick H. Life-time prediction of components for reuse: an overview. *Int J Environ Tech Manag* 2004;4(4):323–348.
9. Klausner M, Grimm W, Hendrickson MC. Reuse of electric motors in consumer products. *J Ind Ecol* 1998;2(2):89–102.
10. Mathew S, Rodgers P, Evely V, *et al.* A methodology for assessing the remaining life of electronic products. *Int J Perform Eng* 2006;2(4):383–395.
11. Pope SM, Elliott JR. Designing for technological obsolescence and discontinuous change: an evaluation of three successful electronic products. Proceedings of the 1998 IEEE International Symposium on Electronics and the Environment; Oak Brook (IL); 1998. pp. 280–286. DOI: 10.1109/ISEE.1998.675072
12. Daimon T, Kandoh S, Umeda Y. Proposal of decision support method for life cycle strategy by estimating value and physical lifetimes. Proceedings of the 11th CIRP International Life Cycle Engineering Seminar on Product Life Cycle—Quality Management; Belgrade, Serbia; 2004. pp. 49–56.
13. Nes N, Cramer J, *et al.* A practical approach to the ecological lifetime optimization of electronic products. 1st International Symposium on Environmentally Conscious Design and Inverse Manufacturing. EcoDesign'99; Tokyo, Japan; 1999. 108–111.
14. Wood AP. Reliability-metric varieties and their relationships. Proceedings of annual reliability and maintainability Symposium. IEEE; Philadelphia (PA); 2001. pp. 110–115. DOI: 10.1109/RAMS.2001.902451
15. Bertsche B, Lechner G. Zuverlässigkeit im Fahrzeug- und maschinenbau. Ermittlung von Bauteil- und system-zuverlässigkeiten. 3., überarb. und erw. Aufl. Berlin u.a.: Springer; 2004.
16. Leitch RD. Basic reliability engineering analysis. London: Butterworth & Company; 1988.
17. Mazhar MI, Kara S, Kaebernick H. Reuse potential of used parts in consumer products: assessment with Weibull analysis. *Int J Prod Eng Comput* 2004;6(7):113–118.

18. Keller J, Maudie T. Accelerometer lifetime prediction modeling based on field failures. *Reliab Edge* 2001;2(3):20–21.
19. Chalkey AM, Billett E, Harrison D, *et al.* Development of a method for calculation the environmentally optimum lifespan of electrical household products. *Proc Inst Mech Eng Part B: J Eng Manuf* 2003;217(11):1521–1531. DOI: 10.1243/095440503771909890
20. Klausner M, Grimm W, Hendrickson MC, *et al.* Sensor-based recording of use conditions for product take-back. *Proceedings of the IEEE Symposium on Electronics and the Environment*; Oak Brook (IL): 1998. pp. 138–143. DOI: 10.1109/ISEE.1998.675046
21. Simon M, Bee G, Moore P, *et al.* Modelling of the life cycle of products with data acquisition features. *Comput Ind* 2001;45(2):111–122.
22. Seliger G, Buchholz A, Grudzein W. Multiple usage phases by component adaptation. *Proceedings of the 9th CIRP International Seminar on Life Cycle Engineering*; Erlangen, Germany: 2002. pp. 47–54.
23. Ni J, Lee J, Djurdjanovic D. Watchdog—information technology for proactive product maintenance and its implications to ecological product re-use. *Proc Symp Ecol Manuf* 2003;17(3–4):109–125.
24. Rugrungruang F, Kara S, Kaebernick H. An integrated methodology for assessing physical and technological life of products for reuse. *Int J Sus Manuf* 2009;1(4):463–490.
25. Rugrungruang F. An integrated methodology for assessing physical and technological life of products for reuse [PhD Thesis]. The University of New South Wales, Sydney, 2008.
26. Mazhar MI, Kara S, Kaebernick H. Remaining life estimation of used components in consumer products: life cycle data analysis by Weibull and artificial neural networks. *J Oper Manag* 2007;25(6):1184–1193.
27. Mazhar MI. Lifetime monitoring of appliances for reuse [PhD thesis]. The University of New South Wales, Sydney, 2006.
28. Kara S, Mazhar MI, Kaebernick H, *et al.* Determining the reuse potential of used components based on life cycle data. *CIRP Ann* 2005;54(1):1–4.
29. Vass J, Randall B, Kara S. Vibration-based approach to lifetime prediction of electric motors for reuse. *Int J Sus Manuf* 2010;2(1):2–29.
30. Vass J. Vibration-based fault diagnostics for quality control and components reuse [PhD Thesis]. Czech Technical University, Prague, 2008.
31. Randall RB. State of the art in monitoring rotating machinery. *Sound Vib* 2004; 38(3):14–21 (Part 1);38(5):10–17 (Part 2).
32. Barreca SL. Technology life-cycles and technology obsolescence. Birmingham (AL): BCRI Inc.; 1998. pp. 1–16.
33. Twiss BC, Montgomerie GA. Forecasting for technologists and engineers: a practical guide for better decisions. IEE management of technology series 15. London: Peter Peregrinus Limited; 1992.
34. Fisher JC, Pry RH. A simple substitution model of technological change. *Technol Forecast Soc Change* 1971;3:75–88.
35. Blackman AW Jr. A mathematical model of technological change. *Technol Forecast Soc Change* 1972;3:441–452.
36. Floyd A. Trend forecasting: a methodology for figure of merit. In: Bright J, editor. *Technological forecasting for industry and government: methods and applications*. Englewood Cliffs, New Jersey: Prentice-Hall; 1962.
37. Martino JP. *Technological forecasting for decision making*. 3rd ed. New York: McGraw-Hill; 1993.
38. Sharif MN, Islam MN. The Weibull distribution as a general model for forecasting technological change. *Technol Forecast Soc Change* 1980;18(3):247–256.
39. Bengisu M, Nekhili R. Forecasting emerging technologies with the aid of science and technology databases. *Technol Forecast Soc Change* 2006;73(7):835–844.
40. Sultan F, Farley JU, Lehmann DR. A Meta-analysis of applications of diffusion models. *J Market Res* 1990;27(1):70–77.
41. Rai LP, Kumar N. Development and application of mathematical models for technology substitution. *Pranjana* 2003;6:49–60.
42. Kobayashi H, Kumazawa T. A procedure methodology for transition to reuse business. *Proc 13th CIRP Conference on Life Cycle Engineering*; Leuven, Belgium: 2006. pp. 577–582.
43. Randall RB. Cepstrum analysis and gearbox fault diagnosis. *Mainten Manag Int* 1983;3:183–208.

ASSOCIATION OF EUROPEAN OPERATIONAL RESEARCH SOCIETIES

M. GRAZIA SPERANZA
Department of Economics
and Management,
University of Brescia,
C.da S.Chiera 50,
25122 Brescia, Italy

THE FOUNDATION OF EURO

At the International Federation of Operational Research Societies (IFORS) conference in Dublin in 1972, the participating presidents of European Operational Research (OR) societies met and agreed that more could be done to improve communication and cooperation on a European level. Representatives of 11 European countries held a meeting in Düsseldorf on 3 and 4 September, 1973, where it was agreed that European national societies would be contacted for their views on various potential initiatives, including setting up a European coordinating body [1]. With strong support from national societies, a further meeting was held in Amsterdam in May 1974 where it was agreed to “formalize and institutionalize increased European cooperation” [2]. In addition, it was also agreed to organize the First European Conference for OR in January 1975 in Brussels [1]. Despite the short timescale, the conference attracted around 500 participants and feedback suggested that the atmosphere was constructive. “At the very festive and impressive Final Session the draft of the agreement was signed by the Representatives of ten European OR Societies (Belgium, Denmark, Finland, Germany, Great Britain, Greece, Ireland, Netherlands, Sweden, Switzerland)” [1]. This was the foundation of the Association of European Operational Research Societies (EURO) within IFORS. In addition, after the Final Session, the Provisional Council of EURO met and agreed to form a Constitutional Committee [1].

As a regional grouping of IFORS, the aim of EURO is to provide effective opportunities for members of national societies to collaborate and further the theory and practice of OR. In addition to the Association itself, appropriate instruments have also been developed to achieve this aim. This paper gives examples of successful activities that reinforce the validity of such a grouping.



THE HISTORY OF EURO

“On 8 March 1976, the Honorary Secretary [Roger Eddison] announced: ‘I hereby declare that EURO, The Association of European Operational Research Societies within IFORS, is now formally constituted with effect from 5 March 1976 and the draft statutes circulated on 29 June 1975 are effective’” [2].

By 18 June, 1976, “the definitive statutes and by-laws of EURO were ratified in total by the OR societies of: Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, The Netherlands, Spain, Switzerland and the United Kingdom” [3]. The Association comprises a Council of national society representatives, an Executive Committee of EURO officers, and a number of EURO support staff to run the office and provide other

liaison roles. At the first official meeting of the Council of EURO, held on 18 June, 1976 in Brussels, elections took place for the officers of the association. The outcome was

President: Professor Dr. H.-J. Zimmermann

Vice-President: Professor Dr. G. Kreweras

Secretary: Professor Dr. R. Eddison

Treasurer: Professor Dr. J.P. Brans

“It was unanimously agreed that these four would form the executive committee” [3]. Professor Zimmermann remained the EURO President for four years, but it was agreed that subsequent Presidents would retain the post for two years.

The membership of EURO increased as more national societies joined IFORS and consideration was given to countries outside of the European Union. For instance, in November 1987, the Executive Committee received a letter from the President of the Operations Research Society of South Africa requesting membership of EURO. There is no African regional grouping of IFORS and the case was made that collaborative links already existed with some individuals in EURO and there was better access in terms of transport to European countries.

As the Association and its membership grew, it was becoming increasingly important to formalize the administrative arrangements. Although the proposal for a permanent secretariat was discussed on numerous occasions, EURO was not in a financial position to set up an office until 1993 [4]. The EURO office was set up in Brussels and incorporates a number of administrative tasks

that support the function of the Executive Committee.

THE ORGANIZATION OF EURO

EURO (www.euro-online.org) is an association of national OR societies. The members of EURO at present are the national societies of the 30 countries listed in Table 1. The organization of EURO is based on the Executive Committee, the Council, and the staff. The Executive Committee (EC) is composed of seven officers: the President, the President-Elect or the Past-President, the Vice-President responsible for conferences (VP1), the Vice-President responsible for education (VP2), the Vice-President responsible for publications (VP3), a secretary responsible for the organization of meetings and for the minutes, and a treasurer. The term of office of the President and the Vice-Presidents is two years. Their appointment cannot be renewed. Each President is member of the EC one year before the term as President-Elect and one year after as Past-President. The secretary and the treasurer appointment can be renewed. They are confirmed by the Council every two years. The Council is composed of two representatives from each of the member societies and meets once a year, during the EURO or the IFORS conferences. Through the Council, the strategy and the actions of the EC are coordinated with the activities of the member societies. EURO is one of the regional groupings of IFORS, and the IFORS Vice-President (EURO) is invited to the meetings of the EC. The staff includes a manager and a webmaster.

Table 1. List of EURO Members

Austria	Germany	Portugal
Belarus	Hungary	Serbia
Belgium	Iceland	Slovakia
Bulgaria	Ireland	Slovenia
Croatia	Israel	South Africa
Czech Republic	Italy	Spain
Denmark	Lithuania	Sweden
Finland	Netherlands	Switzerland
France	Norway	Turkey
Greece	Poland	United Kingdom

Table 2. List of EURO Presidents

Duration	Name	Country	Duration	Name	Country
1975–1978	Hans-Jürgen Zimmermann	Germany	1995–1996	Paolo Toth	Italy
1979–1980	Birger Rapp	Sweden	1997–1998	Jan Weglarz	Poland
1981–1982	Rolfe Tomlinson	UK	1999–2000	Christoph Schneeweiß	Germany
1983–1984	Jean-Pierre Brans	Belgium	2001–2002	Philippe Vincke	Belgium
1985–1986	Bernard Roy	France	2003–2004	Laureano Escudero	Spain
1987–1988	Dominique de Werra	Switzerland	2005–2006	Alexis Tsoukiàs	France
1989–1990	Jakob Krarup	Denmark	2007–2008	Martine Labbé	Belgium
1991–1992	Jaap Spronk	Netherlands	2009–2010	Valerie Belton	UK
1993–1994	Maurice Shutler	UK	2011–2012	M. Grazia Speranza	Italy

Up to 2012, there have been 18 presidents of 11 nationalities (Table 2), demonstrating a truly international strength in both the discipline and motivation to coordinate across a European community. The President-Elect for 2012 is Gerhard Wäscher from Germany.

CONFERENCES

IFORS conferences have always been triennial events. “During the 6th IFORS Conference in 1972 there was a widespread recognition that, with the venues of the next two IFORS conferences determined, there would not be another international conference in Europe for nine years” [2]. At the meeting held in Amsterdam in May 1974, in addition to agreeing to formalize EURO as a regional grouping of IFORS, the representatives of the European OR societies agreed to “assemble operational researchers from all Western European countries for the First European Conference on Operational Research” [2]. As already mentioned, the first European Conference on Operational Research (EURO I) was held in Brussels in January 1975 and attracted around 500 participants. A number of national societies, as well as IFORS, offered financial support [2]. Following the success of EURO I, a second conference was arranged to be held in Stockholm in April 1976. In 1977, EURO co-sponsored the TIMS (Institute of Management Science) XIII international meeting, which was held in Athens. As EURO did not take a leading part in its organization, it is not recorded as part of the EURO-k conference series, but at the Executive Committee meeting held in May 1978,

it was agreed to participate in a second joint conference with TIMS in 1982 in Lausanne which would also be recorded as the fifth EURO conference. The IFORS triennial conference was held in 1978, and the next EURO conference was held in 1979, and thereafter, EURO conferences have been arranged for the interim two years between IFORS conferences. Table 3 lists the EURO-k conferences held to date.

In the early years, EURO-k conferences attracted 500–600 participants. In 1982, Tilanus [5] looked in detail at the composition of the EURO-k conferences to date and considered factors such as national society membership and equivalent American conferences. He also considered many issues and his opinion was that EURO-k conferences “should, and may, in fact, become larger” [5]. From 1991, EURO-k conferences exceeded 600 participants. Despite some fear that conference participation would reduce with advent of the Internet, participation in fact increased further. In 2007, conference participation exceeded 2000, and this is now the figure typically planned for and achieved. As one of the initial instruments proposed by EURO, the growth in participation demonstrates the benefits the European regional grouping can bring.

WORKING GROUPS

The idea of developing working groups to focus on specific topics within OR was first discussed at the meeting held in Amsterdam in May 1974. Eight working groups were established at the first EURO conference in

Table 3. List of EURO-k Conferences

Year	City	Year	City	Year	City
1975	Brussels	1989	Belgrade	2003	Istanbul
1976	Stockholm	1991	Aachen	2004	Rhodes
1979	Amsterdam	1992	Helsinki (joint with TIMS)	2006	Reykjavik
1980	Cambridge	1994	Glasgow	2007	Prague
1982	Lausanne (joint with TIMS)	1995	Jerusalem	2009	Bonn
1983	Vienna	1997	Barcelona (joint with INFORMS)	2010	Lisbon
1985	Bologna	1998	Brussels	2012	Vilnius
1986	Lisbon	2000	Budapest	2013	Rome (joint with INFORMS)
1988	Paris (joint with TIMS)	2001	Rotterdam		

Brussels in January 1975: OR and Energy Problems; OR Applied to Health Services; Fuzzy Sets; OR in Government and the Public Sector; Multicriteria Decisions; OR in Banking; OR in Regional and Urban Planning; and Education. “In the early years of EURO Germain Kreweras and Jakob Krarup edited a charter providing a structure to the Working Groups within a EURO framework.” [6]. Krarup [7] provided an overview of the EURO working groups in 1984, at which point six working groups were highly active, including four of those established in 1975. There are currently 28 working groups within EURO.

The working group framework encourages collaborative work to be carried out in both established and newer topics in OR. This long-standing instrument is dynamic, is fluid, and can respond to emerging themes. In this case, the benefits of European collaboration on more specialized topics can lead to solutions to new problems.

PUBLICATIONS

The idea of starting a European journal was raised in the first meeting in Amsterdam in 1974 and received “considerable support” [8]. In January 1975, a committee was formed to investigate the feasibility of launching a journal dedicated to OR across Europe. Later that year, it was agreed that the *European Journal of Operational Research (EJOR)* be

established, and the first volume was published in 1977, in time for the second EURO conference. The first volume consisted of six issues published bimonthly with a total of 41 papers, and by 1981, there were three volumes containing 120 papers [9]. EJOR also originally included the *EURO Bulletin* which provided an update of European OR activity. “EJOR became a EURO journal in 1990 and since then has continuously grown in terms of reputation and quality (through measures such as the impact factor) to be one of leading journals of the international OR community” [10]. In 2010, EURO agreed to investigate the option of diversifying its portfolio of journals, and in 2011, three new journals were officially launched: *EURO Journal on Transportation and Logistics*; *EURO Journal on Computational Optimization*; and *EURO Journal on Decision Processes* [10]. These journals are an exciting new instrument, and they will complement EJOR as a means of further enhancing specialized areas of OR.

ADDITIONAL EURO INSTRUMENTS

In 1983, three additional instruments were proposed and “enthusiastically approved by our EURO council” [6]. The EURO Gold Medal was awarded for the first time to Hans-Jürgen Zimmermann at EURO VII in 1985. The award is “not only significant for the laureate but is also very important for

the promotion of OR in the public field by making leading persons and their contributions better known" [11]. The award is made at a EURO-k conference and the laureate is invited to give the opening lecture. Mini conferences were initiated in 1984 to offer a more focused event for specialized themes and with more limited participation. These conferences are "a democratic instrument giving opportunities to all the members of EURO, while a large EURO-k conference can only be organized by a selected country and ... only twice every three years" [6]. The first EURO Summer Institute, on Location Theory, was hugely successful and led onto further collaboration after the event [11]. The Summer Institutes, and subsequently Winter Institutes, are aimed at early stage researchers and places are limited so that the experience is unique and rewarding. The EURO General Support Fund was set up in 1994 as a way of encouraging and funding other appropriate activities. In 2001, ORP³ was introduced as "a forum promoting scientific and social exchanges between the members of the future generation of Operational Research in academic research" [12]. EURO now offers a number of further awards, the latest of which was approved for in July 2011 and will be presented annually to the authors of the best EJOR papers. Further attention to the future generations of researchers in OR is given with recent initiatives on education, from high school to postgraduate. The expansion of the EURO instruments is another indicator of a successful association and a recognition that EURO can also adapt over time.

CONCLUSIONS

The aim of EURO is to offer effective opportunities for members of national societies to collaborate and further the theory and practice of OR. EURO started "with very little money, the tools of EURO-k conferences, European Working Groups, a EURO Bulletin, a party working on a European OR journal and lots of enthusiasm" [1]. As EURO approaches its twenty-fifth conference, it can demonstrate a quantifiable increased its membership, conference participation, and instruments. It has

also been able to embrace new technologies to strengthen its communication capabilities across Europe. This time period has not been without its challenges and there may be many to come, but it is clear that the initial vision was achievable. With the continued commitment and enthusiasm of individual members, national society members, but above all a dedicated Executive Committee and support team over the years, EURO continues to thrive and will do for many years to come.

Acknowledgments

I wish to warmly acknowledge the support of Sarah Fores, EURO manager since October 2011, in exploring the EURO archives, analyzing the documents, deriving the relevant information, and drafting this paper.

REFERENCES

1. Zimmermann H-J. The founding of EURO the Association of European Operational Research Societies within IFORS. *Eur J Oper Res* 1995;87:404–407.
2. Rand GK. Forty years of IFORS. *Int Trans Oper Res* 2001;8:611–633.
3. Brans JP. EURO bulletin 6. *Eur J Oper Res* 1977;1:69–71.
4. Shutler M. Minutes of the EURO executive meeting. s.l.: (confidential to the EURO Executive Committee); 1993.
5. Tilanus CB. The European O.R. Congresses: what are we doing?, where are we going? *Eur J Oper Res* 1982;10:12–21.
6. Brans JP. EURO 1975–1995: a fruitful evolution. *Eur J Oper Res* 1995;87:408–414.
7. Krarup J. Profiles of the European working groups. *Eur J Oper Res* 1984;15:13–37.
8. Eddison RT. Notes of the Meeting of the European Committee Amsterdam 3–4 May. s.l.: (confidential to the EURO Executive Committee) 1974.
9. Tilanus CB, Mercer A, Zimmermann H-J. Editorial. *Eur J Oper Res* 1981;6:1–3.
10. Speranza MG. EURO announces three new journals. *IFORS News* September 2011:28.
11. Brans JP. Editorial. *Eur J Oper Res* 1985;20:294–297.
12. EURO. Available at <http://www.euro-online.org>. 2012.

ASYMPTOTIC BEHAVIOR OF CONTINUOUS-TIME MARKOV CHAINS

EYLEM TEKIN

Department of Industrial and
Systems Engineering, Texas
A&M University, College
Station, Texas

INTRODUCTION

Continuous-Time Markov chains (CTMCs) have a wide variety of applications in the real world, which span telecommunication, computer, queueing, manufacturing and distribution systems, and population growth models among others. In particular, Poisson process and birth and death processes are the classical examples of CTMCs that have been widely studied.

Consider a stochastic process that can be in various states over time. Let $X(t)$ denote the state of the system at time $t \geq 0$. $X(t)$, $t \geq 0$, is a random variable that takes values in a discrete set S , which is called the *state space*. The process $\{X(t), t \geq 0\}$ is a CTMC if

$$P\{X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s\} \\ = P\{X(t+s) = j | X(s) = i\},$$

for all $s, t \geq 0$ and $i, j \in S$. In other words, the stochastic process $\{X(t), t \geq 0\}$ has the Markovian property such that the conditional distribution of the future state $X(t+s)$, given the current state $X(s)$ and the past states $X(u)$, $0 \leq u < s$, depends only on the current state and is independent of the past. A CTMC $\{X(t), t \geq 0\}$ with state space S is called *homogenous* or *stationary* if $P\{X(t+s) = j | X(s) = i\} = P\{X(t) = j | X(0) = i\}$ for all $t \geq 0$.

CTMC models are useful in describing the future statistical properties of systems. In particular, the asymptotic behavior of CTMCs has been important in many

applications. Several performance measures of practical interest can be computed by using the limiting distribution of $X(t)$ as $t \rightarrow \infty$. For example, for queueing systems, these performance measures may include the expected waiting time in queue, the expected number waiting in queue, and server utilization. In inventory systems, the performance measures of interest can be the expected inventory holding cost or expected number of lost sales.

While studying the asymptotic behavior of a CTMC, the imminent questions that arise are as follows: (i) Does the distribution of $X(t)$ approach a limit as $t \rightarrow \infty$? (ii) If the limiting distribution exists, is it unique? (iii) If there is a unique limiting distribution, how do we compute it? This article addresses these questions.

PRELIMINARIES

Consider a homogenous CTMC $\{X(t), t \geq 0\}$ on state space S with transition probability matrix $P(t) = [p_{ij}(t)]$ where

$$p_{ij}(t) = P\{X(t) = j | X(0) = i\} \text{ for } i, j \in S \\ \text{and } t \geq 0.$$

We are interested in studying the behavior of $P(t)$ as $t \rightarrow \infty$. However, the matrix $P(t)$ is hard to specify even for simple CTMCs. Therefore, we will use an alternative approach to investigate the behavior of $P(t)$ as $t \rightarrow \infty$.

We first present an equivalent definition of a CTMC as follows: Suppose that each time a stochastic process enters state i , the amount of time it spends in state i before making a transition to a different state is exponentially distributed with rate v_i , and when it leaves state i , it next enters state j , $i \neq j$, with probability p_{ij} . The time spent (i.e., sojourn time) in state i and the new state depends only on state i and not on the history of the system prior to time t because of

the memoryless property of the exponential distribution. Such a stochastic process is said to be a CTMC. Let $q_{ij} = v_i p_{ij}$ denote the rate of moving (i.e., transition rate) from state i to state j for $i, j \in S$ and $i \neq j$. Let us also define

$$q_{ii} = -\sum_{i \neq j} q_{ij} = -v_i,$$

for $i \in S$. Then, we can write the rate matrix $Q = [q_{ij}]$. This matrix is also called the (*infinitesimal*) *generator* of a CTMC. One important property of the rate matrix is that its rows sum to zero.

As an example, consider a machine that can be either up or down. The machine stays up for a random amount of time that is exponentially distributed with rate μ and it takes an exponentially distributed time with rate λ to fix it. Once it is fixed, it becomes as good as new. Let $X(t)$ be 0 if the machine is down at time t and 1 otherwise. Then, $\{X(t), t \geq 0\}$ is a CTMC with $S = \{0, 1\}$ and $v_0 = \lambda, v_1 = \mu, p_{01} = p_{10} = 1$. The rate matrix is

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

The rate matrix is useful in analyzing the asymptotic behavior of a CTMC.

Furthermore, let S_n denote the time of the n th transition of a CTMC $\{X(t), t \geq 0\}$, and $Y_n = S_n - S_{n-1}$ be the n th sojourn time for $n \geq 1$ and $S_0 = 0$. Define $X_0 = X(0)$ as the initial state and $X_n = X(S_n+)$ as the state of the system immediately after the n th transition. The embedded process $\{X_0, (X_n, Y_n), n \geq 1\}$ satisfies

$$P\{X_{n+1} = j, Y_{n+1} > y | X_n = i, Y_n, X_{n-1}, Y_{n-1},$$

$$\dots, X_1, Y_1, X_0\} = p_{ij} e^{-v_i y}.$$

From the above equation, it is clear that $\{X_n, n \geq 0\}$ is a discrete-time Markov chain (DTMC). It is called the *embedded DTMC* of the CTMC $\{X(t), t \geq 0\}$. The transition matrix $P = [p_{ij}]$ of the embedded DTMC is connected to the rate matrix $Q = [q_{ij}]$ of the CTMC

$\{X(t), t \geq 0\}$ by the following equation:

$$p_{ij} = \begin{cases} q_{ij}/v_i & \text{if } v_i \neq 0, i \neq j, \\ 0 & \text{if } v_i \neq 0, i = j, \\ 0 & \text{if } v_i = 0, i \neq j, \\ 1 & \text{if } v_i = 0, i = j, \end{cases} \quad (1)$$

for $i, j \in S$. This relationship is useful in classifying the states of a CTMC, which in turn lets us to characterize the limiting distribution of a CTMC. The next section discusses the classification of states of a CTMC.

CLASSIFICATION OF STATES OF A CTMC

In this section, we introduce several concepts such as accessibility, communication, irreducibility, recurrence, and transience to classify the states of a CTMC. These concepts form the basis for understanding the asymptotic behavior of CTMCs.

State j is said to be *accessible* from state i if $p_{ij}(t) > 0$ for some $t \geq 0$. If state j is accessible from state i , we write $i \rightarrow j$. Two states i and j that are accessible to each other are said to *communicate*, and we denote this by $i \leftrightarrow j$. The relation of communication has reflexivity ($i \leftrightarrow i$), symmetry ($i \leftrightarrow j \Rightarrow j \leftrightarrow i$), and transitivity ($i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$) properties. A set $C \subset S$ is said to be a *communicating class* if (i) $i, j \in C \Rightarrow i \leftrightarrow j$, and (ii) $i \in C, i \leftrightarrow j \Rightarrow j \in C$. A communicating class C is said to be *closed* if $i \in C$ and $j \notin C$, and then j is not accessible from i . This implies that once a CTMC visits a state in a closed communicating class, it cannot leave that class of states. Hence, we can uniquely partition the state space of a DTMC as $S = C_1 \cup \dots \cup C_k \cup T$ where $C_1, \dots, C_k$ are k disjoint closed communicating classes and T includes the remaining states. A CTMC is said to be *irreducible* if all its states communicate with each other; otherwise, it is called *reducible*.

These definitions are the exact analogs of the corresponding definitions for DTMCs. Furthermore, while classifying the states of a CTMC, there is a one-to-one relationship between the CTMC $\{X(t), t \geq 0\}$ and the corresponding embedded DTMC $\{X_n, n \geq 0\}$. Using Equation (1), we can write the following statements:

1. $i \rightarrow j$ for $\{X(t), t \geq 0\}$ if and only if $i \rightarrow j$ for $\{X_n, n \geq 0\}$.
2. $i \leftrightarrow j$ for $\{X(t), t \geq 0\}$ if and only if $i \leftrightarrow j$ for $\{X_n, n \geq 0\}$.
3. C is a (closed) communicating class for $\{X(t), t \geq 0\}$ if and only if C is a (closed) communicating class for $\{X_n, n \geq 0\}$.
4. $\{X(t), t \geq 0\}$ is irreducible if and only if $\{X_n, n \geq 0\}$ is irreducible.

Therefore, the classification of the states of $\{X(t), t \geq 0\}$ is the same as that of the embedded DTMC $\{X_n, n \geq 0\}$. Next, we introduce the concepts of recurrence and transience. Let T_j denote the first time when the CTMC enters state $j, j \in S$. We can express T_j as follows:

$$T_j = \inf\{t \geq S_1 : X(t) = j\}, j \in S,$$

where $S_1 > 0$ is the first time that the CTMC changes state. We define

$$f_{ij} = P\{T_j < \infty | X(0) = i\} \quad (2)$$

as the probability that the CTMC eventually enters state j and $\mu_{ij} = E[T_j | X(0) = i]$ as the expected time it takes for the CTMC to enter state j starting from state i at time zero, $i, j \in S$. State i is said to be *recurrent* if $f_{ii} = 1$ and *transient* if $f_{ii} < 1$. On the basis of this definition, if state i is recurrent, starting in i , the CTMC will visit state i again. From the definition of a CTMC, the process will be starting over when it revisits state i , and hence, state i will eventually be visited again. Repeating this argument, we can conclude that if state i is recurrent, starting in state i , the CTMC will revisit state i infinitely often. On the other hand, if state i is transient, each time the CTMC visits state i , there is a positive probability of $1 - f_{ii}$ that it will never enter state i again. Hence, it follows that state i is recurrent if and only if, starting in state i , the expected number of times that the CTMC will visit state i is infinite. On the other hand, this expectation is finite for transient states. When $f_{ii} < 1$, it is clear that $\mu_{ii} = \infty$. However, if $f_{ii} = 1$, μ_{ii} can be infinite. A recurrent state i is said to be *positive recurrent* if $\mu_{ii} < \infty$ and *null recurrent* otherwise.

A state i is recurrent (transient) for a CTMC $\{X(t), t \geq 0\}$ if and only if it is recurrent

(transient) for the corresponding embedded DTMC $\{X_n, n \geq 0\}$. Therefore, recurrence and transience are class properties for CTMCs as they are for DTMCs. An irreducible CTMC is recurrent (transient) if and only if the embedded DTMC is irreducible and recurrent (transient).

On the other hand, for positive and null recurrence, there may not be a one-to-one correspondence between the CTMC and the embedded DTMC. More specifically, a state i can be positive recurrent for the CTMC while it is null recurrent for the embedded DTMC or vice versa. However, we can develop a necessary and sufficient condition for positive recurrence of CTMCs as follows: Suppose that $\{X(t), t \geq 0\}$ is an irreducible CTMC with a recurrent embedded DTMC $\{X_n, n \geq 0\}$ with transition probability matrix P and let π be a positive solution to $\pi = \pi P$. Then, the expected time between two consecutive visits to state j can be written as

$$\mu_{jj} = \frac{\sum_{i \in S} \pi_i / q_i}{\pi_j}, j \in S. \quad (3)$$

Hence, the CTMC is positive recurrent if and only if $\sum_{i \in S} \pi_i / q_i < \infty$.

In the remainder of this article, we will use the concepts introduced in this section to classify the CTMCs and analyze the asymptotic behavior based on this classification. We first present some general results as follows: Consider a CTMC $\{X(t), t \geq 0\}$ with a transition matrix $P(t)$. Regardless of the class properties of the CTMC, $P(t)$ always approaches a limit as $t \rightarrow \infty$. This limit can be computed as follows:

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \frac{1}{q_j \mu_{jj}}, \quad (4)$$

for $j \in S$, and

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \frac{f_{ij}}{q_j \mu_{jj}}, \quad (5)$$

for $i, j \in S$ and $i \neq j$. The intuition behind Equation (4) is as follows: The expected time that the CTMC spends in state j during each visit to this state is $1/q_j$. Recall that μ_{jj} is the expected time between two consecutive

visits to state j . Then, $1/q_j\mu_{jj}$ is the long-run fraction of time that the CTMC spends in state j when the starting state is j . Equivalently, it is also the long-run probability of finding the CTMC in state j when the starting state is j . If the starting state is $i \neq j$, then this probability is $f_{ij}/q_j\mu_{jj}$, which is given by Equation (5). If j is a transient or null-recurrent state of a CTMC, then $\lim_{t \rightarrow \infty} p_{ij}(t) = 0$ because $\mu_{jj} = \infty$. The next section discusses the asymptotic behavior of irreducible CTMCs.

ASYMPTOTIC BEHAVIOR OF IRREDUCIBLE CTMCs

An irreducible CTMC consists of a single closed communicating class of states. Hence, the states of such a CTMC can be all transient, or all null recurrent, or all positive recurrent. If an irreducible CTMC is transient or null recurrent, then $\lim_{t \rightarrow \infty} p_{ij}(t) = 0$ for all $i, j \in S$. A transient irreducible CTMC has an infinite state space. For $i \in S$, starting from state i , the expected number of times that state i is visited is finite. As a result, a transient irreducible CTMC will eventually permanently exit any state with probability 1 and the limiting distribution is zero. A null-recurrent irreducible CTMC also has an infinite state space. Unlike the transient case, each state of a null-recurrent CTMC will be visited infinitely often over the infinite horizon. However, the expected time between the two consecutive visits to each state is infinite. Thus, the limiting distribution is also zero for this case.

A positive recurrent irreducible CTMC is called *ergodic*. For an ergodic CTMC $\{X(t), t \geq 0\}$, the limiting distribution of $X(t)$ is independent of the initial state. Let us denote the limiting probability of being in state j by

$$p_j = \lim_{t \rightarrow \infty} P\{X(t) = j | X(0) = i\}, \quad j \in S,$$

and the limiting distribution by $p = [p_j]$. An irreducible CTMC is positive recurrent if and only if there exists a solution to the following

equations:

$$pQ = 0, \quad (6)$$

$$\sum_{j \in S} p_j = 1. \quad (7)$$

Furthermore, if there is a solution, it is unique. Equation (6) is known as the *balance equation*. If we write this equation in scalar form, we obtain the following set of equations:

$$\sum_{j \in S} p_i q_{ij} = \sum_{j \in S} p_j q_{ji} \text{ for } i \in S.$$

The left-hand side is the total rate of transitions out of state i and the right-hand side denotes the total rate of transitions into state i , $i \in S$. In steady state, these rates should be equal. Equation (7) is the *normalizing equation* as the sum of the limiting probabilities should be equal to 1.

An alternative way to compute the limiting distribution of a positive recurrent irreducible CTMC is to use the limiting distribution of the embedded DTMC. Let $\{X_n, n \geq 0\}$ be the embedded DTMC with a vector π satisfying $\pi = \pi P$ where P is the transition probability matrix. Combining Equations (3) and (5) and setting $f_{ij} = 1$ for $i, j \in S$ (because the CTMC is irreducible), we obtain

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \frac{1/q_j}{\mu_{jj}} = \frac{\pi_j/q_j}{\sum_{k \in S} \pi_k/q_k}.$$

Since μ_{jj} denotes the expected time between two visits to state j and the expected time that the CTMC stays in state j is $1/q_j$, the above equation also denotes the fraction of time that the CTMC spends in state j , $j \in S$.

ASYMPTOTIC BEHAVIOR OF REDUCIBLE MARKOV CHAINS

Consider a reducible CTMC that consists of k closed communicating classes as $C_1, \dots, C_k$ and the remaining transient states form the set T . By relabeling the states of the CTMC

(if necessary), we can write the rate matrix as follows:

$$Q = \begin{bmatrix} Q(1) & & & 0 \\ & Q(2) & & \\ 0 & & \ddots & \\ & & & Q(k) \\ & R & & & Q(T) \end{bmatrix},$$

where $Q(r)$ denotes the rate matrix for the r th closed communicating class whose row sums are equal to zero, $r = 1, \dots, k$. $Q(T)$ is a $|T|$ by $|T|$ rate matrix with at least one row sum less than zero and R is a $|T|$ by $|S| - |T|$ non-negative matrix. Accordingly, the transition matrix $P(t)$ is given as follows:

$$P(t) = \begin{bmatrix} P_1(t) & & & 0 \\ & P_2(t) & & \\ 0 & & \ddots & \\ & & & P_k(t) \\ & P_R(t) & & & P_T(t) \end{bmatrix}.$$

Since $P_r(t)$ is the transition probability matrix of an irreducible CTMC, from the previous section, we know the asymptotic behavior of $P_r(t)$, $r = 1, \dots, k$. $\lim_{t \rightarrow \infty} P_T(t) \rightarrow 0$ since T is the set of transient states. Hence, in order to completely characterize the limiting behavior of $P(t)$, we need to investigate the limiting behavior of $P_R(t)$.

Define

$$\alpha_i(r) = P\{X(t) \in C_r \text{ for some } t \geq 0 \mid X(0) = i\},$$

$$i \in T, \text{ and } r = 1, \dots, k. \quad (8)$$

$\alpha_i(r)$ is the probability that starting from a transient state i , the probability that the CTMC eventually gets absorbed in class C_r . By using this definition, we can describe the asymptotic behavior of $P_R(t)$ as follows: If C_r is transient or null-recurrent $P_R(t)(i, j) = 0$ as $t \rightarrow \infty$ for $i \in T$ and $j \in C_r$. If C_r is positive recurrent, $P_R(t)(i, j) = \alpha_i(r)p_j$ as $t \rightarrow \infty$ for $i \in T$ and $j \in C_r$, where p_j is given by the solution to

$$p(r)Q(r) = 0, \sum_{j \in C_r} p_j = 1.$$

The probabilities $\{\alpha_i(r), i \in T\}$ defined by Equation (8) are given by the nonnegative solution to the following system of simultaneous linear equations:

$$\alpha_i(r) = \sum_{j \in C_r} \frac{q_{ij}}{q_i} + \sum_{j \in T, j \neq i} \frac{q_{ij}}{q_i} \alpha_j(r).$$

FURTHER READING

Analysis of the asymptotic behavior of CTMCs can be found in a number of textbooks on stochastic processes. Classical textbooks include those by Karlin and Taylor [1], Kulkarni [2,3], Resnick [4], and Ross [5,6]. In addition, since Markovian queueing systems are modeled as CTMCs, queueing theory books such as those by Gross and Harris [7], Kleinrock [8], and Medhi [9] also discuss the asymptotic behavior of CTMCs among others.

REFERENCES

1. Karlin S, Taylor HM. A first course in stochastic processes. San Diego (CA): Academic Press; 1975.
2. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman & Hall; 1995.
3. Kulkarni VG. Modeling, analysis, design and control of stochastic systems. New York: Springer; 1999.
4. Resnick SI. Adventures in stochastic processes. 4th ed. New York: Birkhäuser Boston; 2005.
5. Ross SM. Stochastic processes. 2nd ed. New York: John Wiley & Sons, Inc.; 1996.
6. Ross SM. Introduction to probability models. 10th ed. London: Elsevier; 2009.
7. Gross D, Harris CM. Fundamentals of queueing theory. 3rd ed. New York: Wiley; 1998.
8. Kleinrock L. Queueing systems. Volume 1: theory. New York: Wiley-Interscience; 1975.
9. Medhi J. Stochastic models in queueing theory. 2nd ed. San Diego (CA): Academic Press; 2003.

ASYMPTOTIC BEHAVIOR OF DISCRETE-TIME MARKOV CHAINS

EYLEM TEKIN

Department of Industrial and
Systems Engineering, Texas
A&M University, College
Station, Texas

INTRODUCTION

Markov chains have long been appropriate models for many real-life applications. Large number of physical, biological, and economic systems can be described by Markov chains. Examples include statistical mechanics, thermodynamics, genetics, sociology, telecommunication systems, computer systems, queueing systems, manufacturing and distribution systems, and pricing models among others.

Consider a system that can be in various states over time and transition from one state to the other occurs at discrete points in time randomly. Let X_n denote the state of the system at time $n = 0, 1, \dots$. $X_n, n \geq 0$, is a random variable that takes values in a set S , which is called the *state space*. The Markov property states that the future probability distribution of the system depends only on the current state and is independent of its past. That is, $P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$ for all $n \geq 0$ and $i, j, i_{n-1}, \dots, i_0 \in S$. The stochastic process $\{X_n, n \geq 0\}$ with Markov property is called a *discrete-time Markov chain* (DTMC). A DTMC $\{X_n, n \geq 0\}$ with state space S is called *time-homogenous* if $P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)$ for all $n \geq 0$. Andrey Markov, a Russian mathematician, introduced the concept of Markov chains and produced the first theoretical results in the beginning of the twentieth century [1,2].

DTMC models are useful in describing the future statistical properties of systems. In particular, the asymptotic behavior

of DTMCs is important in many applications. Several performance measures of practical interest can be computed by using the limiting distribution of X_n as $n \rightarrow \infty$. For example, in designing and operating queueing systems, these performance measures may include the average waiting time in queue, the average number waiting in queue, and server utilization. In telecommunication systems, the performance measures of interest can be the expected number of packets waiting in the buffer or the long-run fraction of time that the buffer is full. The asymptotic behavior of DTMCs is also relevant for the analysis of Markov decision processes [3], especially when the objective is to maximize (minimize) the long-run average reward (cost).

While studying the asymptotic behavior of a DTMC, the imminent questions that arise are: (i) Does the distribution of X_n approach a limit as $n \rightarrow \infty$?; (ii) If the limiting distribution exists, is it unique?; (iii) If there is a unique limiting distribution, how do we compute it? This article addresses these questions.

PRELIMINARIES

Consider a time-homogeneous DTMC $\{X_n, n \geq 0\}$ on state space $S = \{1, 2, \dots\}$ with one-step transition probability matrix $P = [p_{ij}]$ where

$$p_{ij} = P(X_{n+1} = j | X_n = i) \text{ for } i, j \in S \text{ and } n \geq 0.$$

We are interested in studying the behavior of X_n as $n \rightarrow \infty$. Since

$$P(X_n = j) = \sum_{i \in S} P(X_n = j | X_0 = i) P(X_0 = i),$$

$$i, j \in S \text{ and } n \geq 0, \quad (1)$$

the distribution of X_n depends on the initial distribution and the n -step

conditional probabilities of the DTMC. Let us define

$$p_{ij}^{(n)} = P(X_n = j | X_0 = i)$$

and denote the n -step transition probability matrix by $P^{(n)} = [p_{ij}^{(n)}]$ for $i, j \in S$ and $n \geq 0$. It can be shown that $P^{(n)} = P^n$ where P^n is the n th power of P . Letting $a_j^{(n)} = P(X_n = j)$ for $j \in S$ and the corresponding vector as $\mathbf{a}^{(n)} = [a_1^{(n)}, a_2^{(n)}, \dots]$ for $n \geq 0$, Equation (1) can be expressed in matrix-vector notation as

$$\mathbf{a}^{(n)} = \mathbf{a}^{(0)} P^n.$$

Hence, the asymptotic behavior of X_n depends on the asymptotic behavior of P^n .

Let us first define three types of distributions that are useful in analyzing the asymptotic behavior of X_n . The probability mass function of X_n as $n \rightarrow \infty$ is called the *limiting* or *steady-state* distribution, if it exists. We denote the limiting distribution by $\{\pi_j, j \in S\}$ where

$$\pi_j = \lim_{n \rightarrow \infty} P(X_n = j), j \in S.$$

The limiting distribution may not exist for all DTMCs. However, if a limiting distribution exists, it satisfies

$$\pi_j = \sum_{i \in S} \pi_i p_{ij}, j \in S, \quad (2)$$

$$\sum_{j \in S} \pi_j = 1. \quad (3)$$

Equation (2) is known as the *balance equation* since it balances the limiting probability of entering a state with the limiting probability of exiting that state. Equation (3) is the *normalizing equation* as the sum of the limiting probabilities over the state space should be equal to 1.

Next, a distribution $\{\pi_j^*, j \in S\}$ is called a *stationary* distribution, if

$$\begin{aligned} P(X_0 = j) &= \pi_j^* \text{ for } j \in S \Rightarrow \\ P(X_n = j) &= \pi_j^* \text{ for } j \in S \text{ and } n \geq 0. \end{aligned}$$

If the initial distribution of the DTMC is chosen as the stationary distribution, the distribution of it stays the same for all time periods. Stationary distribution may not exist for all DTMCs as well. On the other hand, $\{\pi_j^*, j \in S\}$ is a stationary distribution if and only if it satisfies the balance and normalizing equations given in Equations (2) and (3), respectively. Hence, a limiting distribution, when it exists, is also a stationary distribution.

There is also one more interpretation of the balance and the normalizing equations. Let us denote the expected number of times that state j is visited over a time span of $\{0, 1, \dots, n\}$ starting from state i by $m_{ij}(n)$ for $i, j \in S$. It can be shown that

$$m_{ij}(n) = \sum_{k=0}^n p_{ij}^{(k)}, i, j \in S \text{ and } n \geq 0, \quad (4)$$

and $M(n) = [m_{ij}(n)]$ is called the *occupancy time matrix*. Then, the limit of $m_{ij}(n)/(n+1)$ as $n \rightarrow \infty$ is the long-run fraction of time that the DTMC spends in state j starting from state i . Let us denote the long-run fraction of time that the DTMC spends in state j by $\hat{\pi}_j$. Then, we can define the *occupancy* distribution as $\{\hat{\pi}_j, j \in S\}$. If the occupancy distribution exists, it also satisfies Equations (2) and (3).

Thus, the normalized solution to the balance equations may have three different interpretations as limiting, stationary, and occupancy distributions. In the next sections, we will use these distributions to discuss the asymptotic behavior of DTMCs. The existence and uniqueness of these distributions are based on several characteristics of the DTMC. Next section describes these characteristics, which are essential in classifying DTMCs.

CLASSIFICATION OF STATES OF A DTMC

In this section, we introduce several concepts for classifying the states of a DTMC. These concepts form the basis for understanding the asymptotic behavior of DTMCs.

We begin with the following definitions: State j is said to be *accessible* from state i ,

if $p_{ij}^{(n)} > 0$ for some $n \geq 0$. This implies that state j is accessible from state i if and only if starting in i , it is possible that the system will ever enter state j . Two states i and j that are accessible to each other are said to *communicate*, and we denote this by $i \leftrightarrow j$. Note that each state communicates with itself since $p_{ii}^{(0)} = 1$. The relation of communication has reflexivity ($i \leftrightarrow i$), symmetry ($i \leftrightarrow j \Leftrightarrow j \leftrightarrow i$), and transitivity ($i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$) properties.

A set $C \subset S$ is said to be a *communicating class* if (i) $i, j \in C \Rightarrow i \leftrightarrow j$, and (ii) $i \in C, i \leftrightarrow j \Rightarrow j \in C$. A communicating class C is said to be *closed* if $i \in C$ and $j \notin C$, then j is not accessible from i . This implies that once a DTMC visits a state in a closed communicating class, it cannot leave that class of states. Hence, we can uniquely partition the state space of a DTMC as $S = C_1 \cup \dots \cup C_k \cup T$ where $C_1, \dots, C_k$ are k disjoint closed communicating classes and T includes the remaining states. A DTMC is said to be *irreducible* if S is a closed communicating class; otherwise, it is called *reducible*. In other words, an irreducible DTMC has only one class (i.e., all states communicate with each other).

Consider a three-state DTMC with the following one-step transition probability matrix:

$$P = \begin{bmatrix} \alpha & 1 - \alpha & 0 \\ \beta & \beta' & 1 - \beta - \beta' \\ 0 & \gamma & 1 - \gamma \end{bmatrix}. \quad (5)$$

In the above matrix, when $0 < \alpha, \beta, \beta', \gamma < 1$, and $\beta + \beta' < 1$, the DTMC is irreducible. For example, it is possible to go from state 1 to state 3 since $1 \rightarrow 2 \rightarrow 3$. On the other hand, when $\alpha = 1$, $0 < \beta, \beta', \gamma < 1$, and $\beta + \beta' < 1$, the DTMC is reducible. $C_1 = \{1\}$ is a closed communicating class. $T = \{2, 3\}$ is a communicating class that is not closed. Note that a finite-state DTMC has at least one closed communicating class, whereas a DTMC with an infinite state space may not necessarily have any closed communicating classes.

Next, we introduce the concepts of recurrence and transience. For any state i , let f_i denote the probability that, starting in state i , the DTMC will ever revisit state

i . State i is said to be *recurrent* if $f_i = 1$ and *transient* if $f_i < 1$. Based on this definition, if state i is recurrent, starting in i , the DTMC will eventually visit state i again. From the definition of a DTMC, the process will be starting over when it revisits state i , and hence, state i will eventually be visited again. Repeating this argument, we can conclude that if state i is recurrent, starting in state i , the DTMC will revisit state i infinitely often. On the other hand, if state i is transient, each time the DTMC visits state i , there is a positive probability of $1 - f_i$ that it will never enter state i again. Therefore, starting in state i , the probability that the DTMC is in state i for n time periods is $f_i^{n-1}(1 - f_i)$ for $n \geq 1$. In other words, starting in state i , the number of times that the DTMC will visit state i has a geometric distribution with a finite mean of $1/(1 - f_i)$. Hence, it follows that state i is recurrent if and only if starting in state i , the expected number of times that the DTMC will visit state i is infinite. On the other hand, this expectation is finite for transient states. That is, state i is recurrent (transient) if and only if

$$m_{ii}(\infty) = \sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty (< \infty).$$

where $m_{ij}(n)$ is given by Equation (4) for $i, j \in S$ and $n \geq 0$.

Let us define T_i as the expected time between the two consecutive visits of the DTMC to state i . When $f_i < 1$, it is clear that $T_i = \infty$. However, T_i can be infinite even if $f_i = 1$. A recurrent state i is said to be *positive recurrent* if $T_i < \infty$ and *null recurrent* otherwise. A necessary and sufficient condition for positive and null recurrence can be given as follows: A recurrent state i is positive recurrent (null recurrent) if and only if

$$\lim_{n \rightarrow \infty} \frac{m_{ii}(n)}{n + 1} = \frac{1}{T_i} > 0 (= 0).$$

The above limit can be interpreted as the long-run fraction of time that the DTMC spends in state i .

Recurrence and transience are class properties. More specifically, if $i \leftrightarrow j$ and

i is transient (positive recurrent, null recurrent), then j is transient (positive recurrent, null recurrent). This observation greatly simplifies the task of identifying recurrent and transient states in a DTMC. In a communicating class, if one state is shown to be transient (positive recurrent, null recurrent), then all the remaining states have the same property. Furthermore, a finite communicating class is positive recurrent if it is closed and transient if not. Note that in a finite-state DTMC, not all states can be transient.

The last concept that we will discuss in this section is periodicity. A state i is said to have *period* d if $p_{ii}^{(n)} = 0$ whenever n is not divisible by d , and d is the largest integer with this property. This implies that if state i has period d , then $p_{ii}^{(nd)} > 0$ for $n \geq 0$. A state with period 1 is said to be *aperiodic*. Suppose the set of integers n for which $p_{ii}^{(n)} > 0$ is $\{0, 2, 5, 7, \dots\}$, then state i is aperiodic (i.e., $d = 1$). If this set is $\{0, 2, 4, 6, \dots\}$, then state i is periodic with period 2. If the DTMC never comes back to state i after leaving it (i.e., state i is transient or null recurrent), the period is ∞ . In such cases, periodicity is irrelevant. Periodicity is also a class property. That is, if state i has period d and it communicates with state j , then state j has also period d .

In the remainder of this article, we will use the concepts introduced in this section to classify the DTMCs and analyze the asymptotic behavior based on this classification. Next section presents the asymptotic behavior of irreducible DTMCs.

ASYMPTOTIC BEHAVIOR OF IRREDUCIBLE DTMCs

An irreducible DTMC consists of a single closed communicating class of states. Hence, the states of such a DTMC can be all transient, or all null recurrent, or all positive recurrent. If all states are positive recurrent, the irreducible DTMC can be further classified as either positive recurrent aperiodic or positive recurrent periodic. Recall that periodicity is not relevant when the DTMC is transient or null recurrent. Therefore, in this section, we consider the following four cases:

the DTMC is transient, null recurrent, aperiodic positive recurrent, and periodic positive recurrent.

We first state an important result related to the convergence of the limiting distribution in irreducible DTMCs. For an irreducible DTMC, if state i has period $d \geq 1$,

$$\lim_{n \rightarrow \infty} p_{ii}^{(nd)} = \frac{d}{T_i}, \quad i \in S. \quad (6)$$

Intuitively, in the aperiodic case, the DTMC visits state i in every T_i time periods in the long-run. For the periodic case (i.e., $d > 1$), Equation (6) follows from the fact that state i is visited at times that are integer multiples of period d . These convergence results can be shown by using the discrete renewal theorem [4,5].

The Transient Case

A transient irreducible DTMC has an infinite state space. For $i \in S$, starting from state i , the expected number of times that state i is visited is finite (i.e., $m_{ii}^{(n)} < \infty$). This implies that $\lim_{n \rightarrow \infty} p_{ii}^{(n)} = 0$ for $i \in S$. By using the Chapman–Kolmogorov equations [6], it is also straightforward to show that $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = 0$ for $i, j \in S$. As a result, a transient irreducible DTMC will eventually permanently exit any state with probability 1 and the limiting distribution is zero.

The Null Recurrent Case

A null recurrent irreducible DTMC also has an infinite state space. Unlike the transient case, each state of a null recurrent DTMC will be visited infinitely often over the infinite horizon. However, the expected time between the two consecutive visits to each state is infinite. Thus, the limiting distribution is also zero for this case.

The Positive Recurrent Case

An irreducible DTMC is positive recurrent if and only if there exists a solution to the balance and normalizing equations given by Equations (2) and (3), respectively. Furthermore, if there is a solution, it is unique. We will first discuss the case where such a DTMC is also aperiodic.

a) The Aperiodic Case

An aperiodic positive recurrent irreducible DTMC is called *ergodic*. For an ergodic DTMC, $\lim_{n \rightarrow \infty} p_{ij}^{(n)}$ exists, and it is independent of the starting state i . That is, the limiting probability of state j is given by

$$\pi_j = \lim_{n \rightarrow \infty} p_{ij}^{(n)} = \lim_{n \rightarrow \infty} p_{jj}^{(n)} = \frac{1}{T_j}, \quad i, j \in S,$$

where the second equality follows from Equation (6). Furthermore, $\{\pi_j, j \in S\}$ is the unique solution to the balance and normalizing equations given by Equations (2) and (3), respectively. In this case, limiting distribution is the same as the stationary and occupancy distributions.

Consider the transition probability matrix given in Equation (5) with $\alpha = 0.4$, $\beta = 0.2$, $\beta' = 0$, and $\gamma = 0.6$. It is straightforward to observe that the DTMC that corresponds to P is ergodic. Computing the limit of P^n and $M(n)/(n+1)$ as $n \rightarrow \infty$, we obtain

$$\begin{aligned} \lim_{n \rightarrow \infty} P^n &= \lim_{n \rightarrow \infty} \frac{M(n)}{n+1} \\ &= \begin{bmatrix} 0.125 & 0.5 & 0.375 \\ 0.125 & 0.5 & 0.375 \\ 0.125 & 0.5 & 0.375 \end{bmatrix}. \end{aligned}$$

The limiting and occupancy distributions coincide and independent of the starting state (i.e., the rows of the limiting matrix are the same). If we also solve the normalized balance equations, we get $\pi_1 = 0.125$, $\pi_2 = 0.5$, and $\pi_3 = 0.375$.

b) The Periodic Case

For an irreducible positive recurrent DTMC with period $d > 1$, $\lim_{n \rightarrow \infty} p_{ij}^{(n)}$ does not exist. However, since $T_j < \infty$, from Equation (6), we have

$$\lim_{n \rightarrow \infty} p_{ij}^{(nd)} = \frac{d}{T_j} = d\hat{\pi}_j > 0, \quad j \in S.$$

Hence, the limit of $p_{ij}^{(n)}$ does not exist as $n \rightarrow \infty$ and only d equidistant subsequences of $p_{ij}^{(n)}$ (i.e., $p_{ij}^{(0)}, p_{ij}^{(d)}, p_{ij}^{(2d)}, \dots$) have limits. On

the other hand, the limit for $m_{ij}(n)/(n+1)$ exists as $n \rightarrow \infty$ and is given by

$$\lim_{n \rightarrow \infty} \frac{m_{ij}(n)}{n+1} = \hat{\pi}_j, \quad i, j \in S. \quad (7)$$

The $\{\hat{\pi}_j, j \in S\}$ in Equation (7) is the occupancy distribution and given by the unique solution to the balance and normalizing equations (2) and (3), respectively.

Consider the transition probability matrix given in Equation (5) with $\alpha, \beta' = 0$, $\beta = 0.2$, and $\gamma = 1$. In this case, the underlying DTMC is positive recurrent irreducible with a period of 2. Computing the powers of P , we observe that

$$\begin{aligned} P^{2n} &= \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0 & 1 & 0 \\ 0.2 & 0 & 0.8 \end{bmatrix}, \quad \text{for } n \geq 1 \text{ and} \\ P^{2n+1} &= \begin{bmatrix} 0 & 1 & 0 \\ 0.2 & 0 & 0.8 \\ 0 & 1 & 0 \end{bmatrix}, \quad \text{for } n \geq 0. \end{aligned}$$

The value of $\{P^n, n \geq 0\}$ fluctuates, and hence, the limiting distribution does not exist. On the other hand,

$$\lim_{n \rightarrow \infty} \frac{M(n)}{n+1} = \begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.1 & 0.5 & 0.4 \\ 0.1 & 0.5 & 0.4 \end{bmatrix}.$$

Thus, the occupancy distribution exists and is independent of the initial state. The long-run fraction of time that the DTMC spends in states 1, 2, 3 are $\hat{\pi}_1 = 0.1$, $\hat{\pi}_2 = 0.5$, and $\hat{\pi}_3 = 0.4$, respectively. This distribution can also be computed by solving the normalized balance equations.

To summarize, the balance and the normalizing equations play a key role in analyzing the asymptotic behavior of irreducible DTMCs. If there is a solution to these equations, this always indicates that the DTMC is recurrent. However, the interpretation of the solution is different based on the periodicity of the DTMC. When the DTMC is aperiodic, the solution to the normalized balance equations can be interpreted as (i) the limiting probability of being in state j , or (ii) the stationary probability of being in

state j , or (iii) the long-run fraction of time that the DTMC spends in state j . When the DTMC is periodic, only interpretations (ii) and (iii) remain valid.

ASYMPTOTIC BEHAVIOR OF REDUCIBLE MARKOV CHAINS

When the DTMC is reducible, the balance and the normalizing equations either do not have a solution or have infinitely many solutions. For example, consider the transition probability matrix given in Equation (5) with $\alpha = \beta = \beta' = \gamma = 0$. For this DTMC with two transient and one recurrent communication classes, the normalized balance equations do not have a solution. On the other hand, consider that $\alpha, \beta, \gamma = 0$ and $\beta' = 1$, which results in $p_{ii} = 1$ for $i = 1, 2, 3$. In this case, there are infinitely many solutions to the normalized balance equations.

Consider a reducible DTMC that consists of k closed communicating classes as $C_1, \dots, C_k$ and the remaining transient states form the set T . By relabeling the states of the DTMC (if necessary), we can write the transition probability matrix as follows:

$$P = \begin{bmatrix} P(1) & & & 0 \\ & P(2) & & \\ 0 & & \ddots & \\ & & & P(k) \\ & D & & & Q \end{bmatrix}, \quad (8)$$

where $P(r)$ denotes the transition probability matrix for the r th closed communicating class, $r = 1, \dots, k$, and D and Q are the matrices that give the transition probabilities from transient states to recurrent and transient states, respectively. The n th power of P is

given as follows:

$$P^n = \begin{bmatrix} P^n(1) & & & 0 \\ & P^n(2) & & \\ 0 & & \ddots & \\ & & & P^n(k) \\ & D_n & & & Q^n \end{bmatrix}. \quad (9)$$

Since $P(r)$ is the transition probability matrix of an irreducible DTMC, from the previous section, we know the asymptotic behavior of $P^n(r)$, $r = 1, \dots, k$. $\lim_{n \rightarrow \infty} Q^n \rightarrow 0$ as T is the set of transient states. Hence, in order to completely characterize the limiting behavior of P^n we need to investigate the limiting behavior of D_n .

Define

$$\alpha_i(r) = P(X_n \in C_r \text{ for some } n > 0 | X_0 = i),$$

$$i \in T \text{ and } r = 1, \dots, k. \quad (10)$$

where $\alpha_i(r)$ is the probability that starting from a transient state i , the probability that the DTMC eventually gets absorbed in class C_r . By using this definition, we can describe the asymptotic behavior of D_n as follows: If C_r is transient or null recurrent $D_n(i, j) = 0$ as $n \rightarrow \infty$. If C_r is positive recurrent and aperiodic, $D_n(i, j) = \alpha_i(r)\pi_j$ as $n \rightarrow \infty$ where π_j is given by the solution to

$$\pi_j = \sum_{l \in C_r} \pi_l p_{lj}, \quad j \in C_r, \quad (11)$$

$$\sum_{l \in C_r} \pi_l = 1.$$

If C_r is positive recurrent and periodic, $D_n(i, j)$ does not have a limit as $n \rightarrow \infty$. However,

$$\lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{l=0}^n D_l(i, j) = \alpha_i(r)\pi_j, \quad j \in C_r,$$

where π_j is the solution to Equation (11).

FURTHER READING

Analysis of the asymptotic behavior of DTMCs can be found in a number of textbooks on stochastic processes. Classical textbooks include Karlin and Taylor [4], Kulkarni [7,8], Resnick [9], and Ross [6,10]. In addition to the theoretical foundations, Kulkarni [7] also provides computational algorithms for (i) determining recurrent and transient states in infinite state DTMCs, (ii) determining the period of a state, and (iii) evaluating the limiting distribution for irreducible DTMCs.

REFERENCES

1. Markov AA. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 2-ya seriya, tom 15; 1906. pp. 135–156.
2. Markov AA. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. Reprinted in Appendix B In: Howard R. *Dynamic probabilistic Systems*, volume 1: Markov chains. New York: John Wiley & Sons; 1971.
3. Puterman ML. *Markov decision processes*. New York: John Wiley & Sons, Inc.; 1994.
4. Karlin S, Taylor HM. *A first course in stochastic processes*. San Diego (CA): Academic Press; 1975.
5. Kohlas J. *Stochastic methods of operations research*. London: Cambridge University Press; 1982.
6. Ross SM. *Stochastic processes*. 2nd ed. New York: John Wiley & Sons, Inc.; 1996.
7. Kulkarni VG. *Modeling and analysis of stochastic systems*. London: Chapman & Hall; 1995.
8. Kulkarni VG. *Modeling, analysis, design and control of stochastic systems*. New York: Springer; 1999.
9. Resnick SI. *Adventures in stochastic processes*. 4th ed. New York: Birkhäuser Boston; 2005.
10. Ross SM. *Introduction to probability models*. 10th ed. London: Elsevier; 2009.

**AUSTRIAN SOCIETY OF
OPERATIONS RESEARCH
(OESTERREICHISCHE GESELLSCHAFT
FÜR OPERATIONS RESEARCH,
OEGOR)**

MARION S. RAUNER
School of Business, Economics, and
Statistics, University of Vienna,
Vienna, Austria

JOSEF HAUNSCHMIED
Institute for Mathematical Methods
in Economics, Vienna University
of Technology, Vienna, Austria

INTRODUCTION

The Austrian Society of Operations Research [Oesterreichische Gesellschaft fuer Operations Research (OEGOR)] is a non-profit scientific organization which supports the application and promotes development of operations research (OR) methods in Austria, with particular focus on efficient knowledge transfer among high-profile scientific research and decision makers in all areas. As a nationwide association, OEGOR also represents Austria in the international network of OR experts such as the Association of European Operational Research Societies (EURO) and the International Federation of Operational Research Societies (IFORS).

History and Milestones

OEGOR's predecessor was the Austrian working group on Operations Research (OEAGOR) which was founded by Prof. Christoph Mandl from the Institute for Advanced Studies (IHS). Erich Steinbauer from IBM Austria, who developed the first professional optimization system for power plants in Europe, was as a very active member during those first years. In the mid-1970s, Prof. Mikulas Luptacik moved from the IHS to the Institute of Operations

Research at the Vienna University of Technology. He spread the idea of establishing an Austrian Society of Operations Research. This is why, in fall 1978, OEGOR the OEGOR was founded at the Vienna University of Technology. First executive board members included: Prof. Gustav Feichtinger (Vienna University of Technology), Prof. Otto Gurtner (University of Natural Resources and Applied Life Sciences, Vienna), Peter Harhammer (IBM), Rainer Hasenauer (Institute for Advanced Studies), Alfred Kalliauer (Verbund), Christoph Mandl (Institute for Advanced Studies), Georg Urbanski (Austrian Airlines), and Prof. Guenther Vinek (University of Vienna). Christoph Mandl became OEGOR's founding president.

Shortly after, Christoph Mandl moved to Boston for his research stay at the Operations Research Center at the M.I.T. His successor, Peter Harhammer from IBM, presided over the OEGOR till 1986. He intensified contacts to fellow societies and organized the EURO VI Conference in Vienna in July 1983. Furthermore, he included many practitioners from the energy industry, especially in the working group on optimization and prognosis.

During the presidency of Prof. Rainer Burkard from the Technical University of Graz (1986–1988), a joint meeting with the Swiss Society of Operations Research (SVOR) was organized in July 1987, which proved to be a milestone for future collaboration between Switzerland and Austria. In 1997, Prof. Burkard was awarded the EURO Gold Medal in honor of his outstanding research in combinatorial optimization. He is a frequently invited keynote speaker at international conferences such as the EURO Bonn Meeting in 2009.

His successor, Prof. Gustav Feichtinger from the Vienna University of Technology, Department of Operations Research (1988–1991) successfully organized the huge OR 1990 Conference at the Vienna University of Technology and the Vienna University of Economics and Business, together with his

colleagues (Richard Hartl, Wolfgang Janko, Wolfgang Katzenberger, Adolf Stepan, Alfred Taudes, Alfred Wagenhofer). The conference budget amounted to more than € 300,000, but fortunately many sponsors could be found. With more than 1,200 participants and many distinguished keynote speakers, this event constituted an important milestone for our society. Prof. Feichtinger became an honorary member of both the Austrian and the German Societies of OR. His recent retirement came as a terrible blow to our society and especially to his well-established research group on optimal control theory, especially applied to socio-economic decision problems.

Prof. Wolfgang Katzenberger from the Vienna University of Technology took over the presidency for a few months before Prof. Georg Pflug from the University of Vienna (Department of Statistics and Decision Support Systems) was elected the next president for 1992–1993. In May 1993, under President Pflug, OEGOR finalized and published a comprehensive Software Guide on OR. Furthermore, Pflug intensively recruited new members from industry and academia so that the membership rose to 165. Prof. Mikulas Luptacik took a major part in launching OEGOR's scientific journal, the *Central European Journal for Operations Research and Economics (CEJORE)* during the presidency of Prof. Pflug. *CEJORE* was a successor of the *Czechoslovakian Journal for Operations Research (CSJOR)* published by Kartprint in Bratislava. Prof. Pflug is still very actively involved in the editorial board of this journal. Currently, he is serving as dean of the Faculty of Business, Economics, and Statistics, University of Vienna, Austria and as head of the OEGOR working group on OR in finance. His main research interests lie in computational risk management and he has many successful industrial applications in this field.

Under President of Prof. Ulrike Leopold-Wildburger from the University of Graz, Department for Statistics and Operations Research (1993–1997), the journal *CEJORE* became scientifically and financially well-established. Thanks to her great efforts, our fellow societies of Croatia, Czech Republic,

Hungary, Slovakia, and Slovenia joined this project. In the following years, she continued putting a lot of effort into the journal. She established a laboratory for experimental research and collaborates with internationally respected international scientists such as Prof. Reinhard Selten, a Nobel laureate. In 2008, she was awarded the Great Josef Krainer Prize for Science and Research.

OEGOR's next president was Prof. Mikulas Luptacik from the Vienna University of Technology, Department of Operations Research (1997–1999). Together with Ulrike Leopold-Wildburger, they worked hard on establishing a sound financial basis for the journal *CEJORE* to minimize the resulting financial risks for the OEGOR. In the course of this process, *CEJORE* was renamed to *Central European Journal for Operations Research (CEJOR)*. Today, *CEJOR* is published by Springer and is also ranked in the Science Citation Index (SCI). Currently, Prof. Luptacik is a full professor for Quantitative Economics at the Vienna University of Economics and Business Administration with a main research interest in data envelopment analysis. Furthermore, he is the head of the Industrial Science Institute, Vienna.

Under President Prof. Richard Hartl from the University of Vienna, Department of Production and Logistics (1999–2003), OEGOR launched its Internet website [1] and published its first electronic newsletter. Furthermore, Prof. Franz Rendl from the University of Klagenfurt (Department of Mathematics) organized the OR 2002 Conference, a joint meeting of the Austrian, German, and Swiss Operations Research Societies. For several years, Prof. Hartl served as managing editor of the *OR Spectrum*, one of the journals of the German OR society. He established a strong working group on ant colony optimization and heuristics at the University of Vienna, which are applied in industry and health care, such as ambulance vehicle routing and location for the Austrian Red Cross.

Prof. Immanuel Bomze from the University of Vienna, Department of Statistics and Decision Support (2003–2006) was the next president. He improved OEGOR's webpage by improving content and database management. Furthermore, he intensified

the contacts to practitioners and drew up OEGOR's new statutes. During his presidency, Bank Austria started the sponsorship for the OEGOR prize for young researchers. His research interests are, among others, asymptotic statistics and stochastic modeling, optimization theory, and applications of dynamical systems such as in telecommunication for Telekom Austria. Prof. Bomze serves as head of the OEGOR working group on the theory and practice of optimization.

In 2006, Prof. Marion Rauner from the University of Vienna, Department of Innovation and Technology Management was elected president. The webpage of the society was further improved. In fall 2008, the OEGOR celebrated its 30th anniversary in the old City Hall, an event sponsored by EURO, Bank Austria, the City of Vienna, the University of Vienna, Austrian Telecom, and fin4cast. Fellow societies congratulated OEGOR on its achievements and prolonged their cooperation and financial contributions to our joint journal CEJOR. The neighbor societies have continued their cooperation and financial contribution regarding our joint journal CEJOR. We certainly had reason to celebrate because CEJOR was included in the SCI—a major accomplishment of **editor-in-chief** Prof. Leopold-Wildburger who was supported by **editor-in-chiefs** Profs. Luptacik, Prof. Pflug, and Prof. Vetschera. Prof. Rauner serves as head of the OEGOR working group on health care. She was presented, amongst others, the Pharmig Prize for Health Economics in 2002 for her research on the evaluation of external defibrillators for the Austrian Red Cross together with her masters student, Nikolaus Bajmoczy. Her main research focus is on health technology assessment, prevention policy models, and in-patient reimbursement systems.

Current Organization of the Austrian OR Society

In 2008, the new executive board of OEGOR was elected:

- Prof. Marion Rauner (University of Vienna)—president;

- Prof. Eranda Dragoti-Cela (Technical University of Graz)—OEGOR news;
- Florian Frommlet (University of Vienna)—external communication;
- Josef Haunschmied (Vienna University of Technology)—secretary, promotion of young scientists;
- Georg Kern (Telekom Austria)—vice treasurer, sponsoring;
- Martin Kuehrer (fin4cast)—vice president, sponsoring;
- Gerold Petritsch (e&t Energy Trading Company)—treasurer; and
- Prof. Franz Rendl (University of Klagenfurt)—vice secretary, CEJOR executive.

OEGOR focuses on these main topics:

- supply chain management;
- trans-business logistics-controlling;
- methods and software for production planning;
- efficiency and productivity analysis for industry, commerce, and non-profit organizations;
- financial engineering;
- telecommunication.

These topics are discussed both at the annual meetings and at regular meetings of the different working groups throughout the year. The current working groups of OEGOR include:

- Metaheuristics—Head: Prof. Peter Greistorfer (University of Graz);
- OR in Finance—Heads: Prof. Georg Pflug (University of Vienna) and Martin Kuehrer (fin4cast);
- OR in Health Care—Heads: Prof. Marion Rauner (University of Vienna) and Prof. Margit Sommersguter-Reichmann (University of Graz);
- Production and Logistics—Head: Prof. Manfred Gronalt (University of Natural Resources and Applied Life Sciences, Vienna) and Prof. Werner Jammerneegg (Vienna University of Economics and Business);

- Theory and Practice of Optimization—Head: Prof. Immanuel Bomze (University of Vienna); and
- Mathematical Economics and Optimization in Energy—Head: Gerold Petritsch (e&t, Energy Trading Company).

OEGOR actively collaborates with fellow OR societies and EURO. This is reflected in joint meetings of working groups (e.g., Vienna, Austria, 2005; Graz, Austria, 2007; Neubiberg, Germany, 2007) and joint conferences of the Austrian, German, and Swiss OR Societies (e.g., Klagenfurt, Austria, 2002; Karlsruhe, Germany, 2007; Zuerich, Switzerland, 2011). Furthermore, OEGOR frequently invites distinguished OR specialists to contribute in theory excellence in theory and practice to our working groups. Prof. Ulrike Leopold-Wildburger from the University of Graz, Austria and Prof. Stefan Pickl from the University of Armed Forces, Munich (Neubiberg), Germany chair the EURO working group on Experimental Economics (EWG E-CUBE). In addition, several Austrian researchers are involved in the board of EURO working groups. We are also proud that Prof. Franz Rendl from the University of Klagenfurt will organize a European Summer Institute (ESI) on non-linear methods in combinatorial optimization for EURO in September 2010.

Supporting young scientists is particularly important to our society. Since 1985, many young scientists have been awarded OEGOR prizes which are partly sponsored by, for example, Bank Austria. Many laureates have become associate or full professors at Austrian and foreign universities. For example, Guenter Rote received the OEGOR prize in 1985 for his master thesis on “A systolic array algorithm for the algebraic path problem” written under the supervision of Prof. Rainer Burkard at the TU Graz.

He is currently full professor for computer science at the Free University of Berlin. Herbert Dawid was awarded the OEGOR prize in 1995 for his PhD thesis on genetic learning in economic systems, which he wrote while working as an assistant at the Vienna University of Technology, Department of Operations Research. He is now a distinguished professor at the University of Bielefeld, Department of Mathematical Economics in Germany. OEGOR laureate of 2003 was Marc Reimann, who worked at the Department of Production and Logistics at the University of Vienna, and wrote his PhD thesis on “An ant-based optimization of goods transportation.” He worked at the Department of Production and Logistics at the University of Vienna. In 2009, he was appointed full professor for production and logistics at the University of Graz in Austria. Apart from this, we also sponsor young scientists so as to enable them to participate in winter and summer courses on OR and quantitative modeling.

OEGOR cooperates with fellow OR societies from Czech Republic, Hungary, Slovakia, Slovenia, and Croatia to editing and finance the SCI-indexed journal CEJOR, which is published by Springer. In addition, members of the Austrian OR Society publish scientific books, work as editors of internationally renowned journals, and edit special issues on different topics for OR journals.

Acknowledgment

The authors wish to acknowledge constructive comments on a previous version of this paper made by OEGOR members (especially Prof. Bettina Klinz, Gabriela Sturm-Petritsch, and Gerold Petritsch).

REFERENCES

1. www.oegor.at.

AVAILABILITY ANALYSIS: CONCEPTS AND METHODS

ENRICO ZIO
Ecole Centrale Paris-Supelec,
Paris, France

Politecnico di Milano, Milano,
Italy

Availability is a concept used to characterize the performance of a machine or component (hereafter more generally termed *unit*) with respect to its ability to fulfill the function for which it is operated [1–6]. It applies to units that can be maintained, restored to operation, or renovated upon failure depending on the particular strategy adopted to optimally assure its function:

- off-schedule (corrective) maintenance, that is, replacement or repair of the failed unit;
- preventive maintenance, that is, regular inspections, and possibly repair, based on a structured maintenance plan;
- conditioned maintenance, that is, performance of repair actions upon detection of the degraded conditions of the unit.

It differs from the reliability indicator used to characterize the ability of a unit of achieving the objectives of a specified mission, without failures, within an assigned period (see also the section titled “Reliability and Maintainability” in this encyclopedia).

The main issues to consider when analyzing the availability of a unit are as follows:

1. *Unrevealed failures*, that is, when a unit fails unnoticed; the system goes on without noticing the unit failure until a test on the unit is made or the unit is demanded to function.

2. *Testing/preventive maintenance actions*, that is, when a unit is removed from the system for testing or preventive maintenance.
3. *Repairs*, that is, when a unit is unavailable because under repair.

In this chapter, a general introduction to the quantitative modeling for the availability analysis of components and systems is provided.

AVAILABILITY: QUANTITATIVE DEFINITION

Let $X(t)$ be a binary indicator variable denoting the state of a unit at time t , that is, $X(t) = 1$ if the unit is operating at time t and $X(t) = 0$, if is failed at time t . The *instantaneous availability* $p(t)$ is defined as the probability that the system is operating at time t ; dually, the *unavailability* $q(t)$ is defined as the probability that the unit is failed at time t . Obviously, $p(t) = 1 - q(t)$.

Notice the difference in the meaning of $p(t)$, the probability that the unit is operating at time t , from that of the reliability $R(t)$, the probability that the unit functions free of failures up to time t (see also the section titled “Reliability and Maintainability” in this encyclopedia).

Operatively, the time-dependent availability function of a unit is synthesized by point values as follows:

- For units undergoing corrective maintenance, the *limiting* or *steady-state availability* is computed as the mathematical limit of the availability function $p(t)$ as t grows to infinity. It represents the probability that the unit is functioning at an arbitrary moment of time, after the transient of the failure and repair processes have stabilized. It is obviously undefined for units under periodic maintenance, for which the limit does not exist.

- For units under periodic maintenance, the average availability over a given period of time is introduced as the proper indicator of performance. It represents the expected proportion of time that the system is operating in the considered period of time.

The Availability of an Unattended Unit

An unattended unit (i.e., a unit whose operation is not monitored) will function till its first failure and remain failed after that, because unnoticed. Hence, the probability $p(t)$ that at time t the unit is functioning is equal to the probability that it never failed before t , that is, the reliability at time t .

The Availability of a Continuously Monitored Unit

For a continuously monitored unit it is assumed that restoration starts immediately after its failure and a probabilistic model describing the duration of the repair process is introduced.

Consider N identical units at time $t = 0$ and let $h(t)$ be the conditional probability density function of the unit random failure time given that it is operating at time t (hazard rate, see also **Hazard Rate Function**) and $g(t)$ the probability density function of the random repair duration. At any successive time t , some units will be functioning whereas others will be failed; the following balance equation for the expected number of units functioning at time $t + \Delta t$ can be written as follows:

$$\begin{aligned} N p(t + \Delta t) &= N p(t) - N p(t) h(t) \Delta t \\ &\quad + \int_0^t N p(\tau) h(\tau) \Delta \tau g(t - \tau) \Delta t. \end{aligned} \quad (1)$$

The first term on the right-hand side is the expected number of units functioning at time t ; the second term is the number of units failing in the interval Δt , that is, the loss term in the balance equation; the third term is the expected number of units that had failed prior to t in a generic interval $(\tau, \tau + \Delta \tau)$ and

whose restoration terminates in $(t, t + \Delta t)$, integrated over all values of $\tau \leq t$.

Dividing by $N \Delta t$, subtracting $p(t)$ on both sides, and letting Δt and $\Delta \tau$ tend to zero, the following integral-differential equation is obtained:

$$\begin{aligned} \frac{dp(t)}{dt} &= -h(t) p(t) \\ &\quad + \int_0^t h(\tau) p(\tau) g(t - \tau) d\tau, \end{aligned} \quad (2)$$

where the integral term on the right-hand side of the equation represents the convolution of the instantaneous availability function and the restoration probability density function.

As initial condition, it is usually assumed that the unit is functioning at the initial time $t = 0$, that is, $p(0) = 1$.

The solution to the integral-differential equation (Eq. 2) depends on the functional form of $h(t)$ and $g(t)$. One way to proceed is by applying the Laplace transform to obtain an algebraic equation in the transform $\tilde{p}(s)$ which is then inverse transformed to obtain $p(t)$ to which the final value theorem can be applied for obtaining the limiting availability, p_∞ . The Laplace transform of the probability $p(t)$ is defined as $\tilde{p}(s) = L[p(t)] = \int_0^\infty e^{-st} p(t) dt$.

Consider, for example, a unit with exponential probability distributions of the random failure time and repair duration, with rates λ and μ , respectively, that is, $h(t) = \lambda$ and $g(t) = \mu e^{-\mu t}$; the Laplace transform $\tilde{p}(s)$ of the availability is easily found to be

$$\tilde{p}(s) = \frac{1}{s + \lambda \frac{s}{s + \mu}} = \frac{s + \mu}{s(s + \mu + \lambda)}, \quad (3)$$

which can be inverted to obtain the instantaneous availability in the time domain:

$$p(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda} e^{-(\mu + \lambda)t}. \quad (4)$$

To determine the limiting availability, p_∞ , the final-value theorem can be applied

$$p_\infty = \lim_{t \rightarrow \infty} p(t) = \lim_{s \rightarrow 0} [s \tilde{p}(s)] = \frac{\mu}{\mu + \lambda} \quad (5)$$

or more simply, in this case, one may directly let t tend to zero in Equation (4).

Notice that one can also write Equation (5) as

$$p_{\infty} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \frac{1}{\mu}} = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}}, \quad (6)$$

where MTTF (Mean Time To Failure) and MTTR (Mean Time To Repair) are the expected values of the probability distributions of the unit failure and repair times.

The Availability of a Unit Under Periodic Test

Some types of units are operated in standby until they are called into operation. These units are unattended and their failure is revealed only when tested.

For a unit undergoing a periodic inspection and maintenance plan, the instantaneous availability is a periodic function of time that can be synthesized by the average availability over the period τ between successive inspections:

$$p_{\tau} = \frac{\int_0^{\tau} p(t) dt}{\tau} = \frac{\int_0^{\tau} R(t) dt}{\tau}, \quad (7)$$

where the unit instantaneous availability $p(t)$ is equal to its reliability $R(t)$ within the interval τ in which it is unattended. Notice that the testing and maintenance procedures have been assumed to occur instantaneously every τ and to bring the unit back to its perfect, *as good as new* conditions.

Dually, the average unavailability over the period τ is

$$q_{\tau} = \frac{\int_0^{\tau} q(t) dt}{\tau} = \frac{\int_0^{\tau} F(t) dt}{\tau}, \quad (8)$$

where $F(t) = 1 - R(t)$ is the cumulative distribution function of the unit failure times.

For example, if the unit has exponentially distributed failure times with constant rate λ , $F(t) = 1 - e^{-\lambda t}$, which for the practical cases of interest can be approximated as $F(t) = 1 -$

$e^{-\lambda t} \cong \lambda t$ so that the average unavailability takes the common form:

$$q_{\tau} = \frac{\int_0^{\tau} F(t) dt}{\tau} = \frac{\int_0^{\tau} \lambda t dt}{\tau} = \frac{1}{2} \lambda \tau. \quad (9)$$

Assuming a finite duration τ_R of the test operations, during which the unit is unavailable, the average availability over the entire testing cycle period $\tau + \tau_R$ becomes

$$\bar{p} = \frac{\int_0^{\tau} R(t) dt}{\tau + \tau_R}, \quad (10)$$

and the average unavailability

$$\bar{q} = \frac{\tau_R + \int_0^{\tau} F(t) dt}{\tau + \tau_R}. \quad (11)$$

However, typically in practice the duration τ_R is small compared with the period τ , so that

$$\bar{p} = \frac{\int_0^{\tau} R(t) dt}{\tau} \quad (12)$$

and

$$\bar{q} = \frac{\tau_R + \int_0^{\tau} F(t) dt}{\tau}. \quad (13)$$

AVAILABILITY OF COMPLEX SYSTEMS: MARKOV MODELING

When one is interested by the availability of a system made of many units in complex logic, the previous analysis must be extended to account for the system behavior as described by its multiple states and the transitions among them. The system states are defined by the states of the units comprising the system. The units are not restricted to having only two possible states but rather may have a number of different states such as functioning, in standby, degraded, partially failed, completely failed, and under maintenance; the various failure modes of a unit may also be defined as states. The transitions between

the states occur randomly in time, because caused by various mechanisms and activities such as failures, repairs, replacements, and switching operations, which are random in nature. Common cause failures may also be included as possible transitions occurring randomly in time.

Let us consider a system that may stay in $N + 1$ configurations, $j = 0, 1, 2, \dots, N$. The state variable describing the system configuration at time t is denoted by $X(t)$ and is no longer binary. The system is assumed to start in a specified state at time $t = 0$. The transitions between states are assumed to occur continuously in time as described by a stochastic process $\{X(t); t \geq 0\}$ governed by the transition probabilities.

Under specified conditions, the stochastic process of the system evolution may be described as a Markov process in which the system states and the possible transitions can be depicted with the aid of a state-space diagram, known as a Markov diagram and be mathematically described by a probabilistic Markov system of equations [7–12] (see also the section titled “Continuous-Time Markov Chain” in this encyclopedia). The Markov property states the following: given that a system is in state i at time t [i.e., $X(t) = i$], the probability of reaching state j at time $t + v$ does not depend on the states $X(u)$ visited by the system prior to t ($0 \leq u < t$). In other words, given the present state $X(t)$ of the system, its future behavior is independent of the past:

$$\begin{aligned} P[X(t+v)=j|X(t)=i, X(u)=x(u), 0 \leq u < t] \\ = P[X(t+v)=j|X(t)=i] \end{aligned} \quad (14)$$

The conditional probabilities

$$P[X(t+v)=j|X(t)=i] \quad i, j = 0, 1, 2, 3, \dots, N \quad (15)$$

are called the *transition probabilities* of the Markov process. If the transition probabilities do not depend on time t but only on the time interval v for the transition, then the Markov process is said to be homogeneous or

stationary:

$$\begin{aligned} P[X(t+v)=j|X(t)=i] &= p_{ij}(v), \\ \text{for } t, v > 0 \text{ and } i, j &= 0, 1, 2, \dots, N. \end{aligned} \quad (16)$$

A Markov process with stationary transition probabilities has no memory.

Considering a time step dt sufficiently small that only one event can occur, it is possible to write the one step transition probability from state i to state j as

$$\begin{aligned} p_{ij}(dt) &= P[X(t+dt)=j|X(t)=i] \\ &= \alpha_{ij} dt + o(dt), \end{aligned} \quad (17)$$

where $\lim_{dt \rightarrow 0} \frac{o(dt)}{dt} = 0$.

The parameter α_{ij} is the transition rate from state i to state j . Since α_{ij} is constant, the time T_{ij} that the system stays in state i before making a transition to state j is exponentially distributed with parameter α_{ij} .

A transition probability matrix can be introduced

$$\underline{\underline{A}} = \begin{pmatrix} 1 - dt \sum_{j=1}^N \alpha_{0j} & \alpha_{01} dt & \dots & \alpha_{0N} dt \\ \alpha_{10} dt & 1 - dt \sum_{\substack{j=0 \\ j \neq 1}}^N \alpha_{1j} & \dots & \alpha_{1N} dt \\ \dots & \dots & \dots & \dots \end{pmatrix}, \quad (18)$$

so that the following matrix equation governing the Markov process can be written as

$$\underline{P}(t+dt) = \underline{P}(t) \underline{\underline{A}}, \quad (19)$$

where, for example, the first equation is

$$\begin{aligned} P_0(t+dt) &= \left[1 - dt \sum_{j=1}^N \alpha_{0j} \right] P_0(t) \\ &\quad + \alpha_{10} P_1(t) dt + \dots + \alpha_{N0} P_N(t) dt. \end{aligned} \quad (20)$$

Subtracting $P_0(t)$ on both sides, dividing by dt and in the limit of $dt \rightarrow 0$, one gets:

$$\begin{aligned} \frac{dP_0}{dt} = & - \sum_{j=1}^N \alpha_{0j} P_0(t) \\ & + \alpha_{10} P_1(t) + \dots + \alpha_{N0} P_N(t). \end{aligned} \quad (21)$$

Manipulating in the same way the other equations of the system (Eq. 19), one can write

$$\begin{aligned} \frac{d\underline{P}}{dt} &= \underline{P}(t) \underline{A}^*, \underline{A}^* \\ &= \begin{pmatrix} -\sum_{j=1}^N \alpha_{0j} & \alpha_{01} & \dots & \alpha_{0N} \\ \alpha_{10} & -\sum_{\substack{j=0 \\ j \neq 1}}^N \alpha_{1j} & \dots & \alpha_{1N} \\ \dots & \dots & \dots & \dots \end{pmatrix}. \end{aligned} \quad (22)$$

The above is a system of linear, first-order differential equations in the unknown state probabilities $P_j(t)$, $j = 0, 1, 2, \dots, N$, $t \geq 0$. The matrix $\underline{A}^*$ contains the transition rates of the system; to simplify the notation, from now on the transition rate matrix $\underline{A}^*$ will be simply denoted as $\underline{A}$. Note that $\alpha_{ii} = -\sum_{\substack{j=0 \\ j \neq i}}^N \alpha_{ij}$.

The system of equations (Eq. 22) is to be solved starting from the initial condition $\underline{P}(0) = \underline{C}$ representing probabilities of the initial states of the system units at $t = 0$. The easiest method of solution is again by Laplace transform. The Laplace transform of the state probability $P_j(t)$, $j = 0, 1, 2, \dots, N$, denoted by $\tilde{P}_j(s)$, is defined as $\tilde{P}_j(s) = L[P_j(t)] = \int_0^\infty e^{-st} P_j(t) dt$; correspondingly, the Laplace transform of the time derivative of $P_j(t)$ is

$$L\left(\frac{dP_j(t)}{dt}\right) = s \tilde{P}_j(s) - P_j(0), \quad j = 0, 1, \dots, N.$$

Laplace-transforming the fundamental equation of the Markov process (Eq. 22):

$$s \tilde{\underline{P}}(s) - \underline{C} = \tilde{\underline{P}}(s) \underline{A} \quad (23)$$

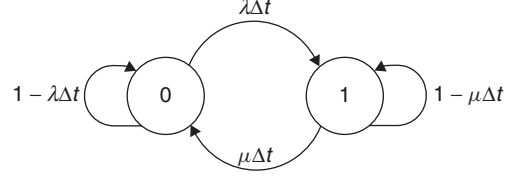


Figure 1. Markov diagram of a binary unit.

from which

$$\tilde{\underline{P}}(s) = \underline{C} [s\underline{I} - \underline{A}]^{-1}, \quad (24)$$

where $\underline{I}$ is the identity matrix. Then, applying the inverse Laplace transformation one may retrieve the state probabilities vector $\underline{P}(t)$.

Furthermore, the steady state probability vector $\underline{\Pi}$ of the system states can be found by simply setting to zero the derivative of $\underline{P}$ in the fundamental equation (Eq. 22):

$$\underline{\Pi} \underline{A} = 0. \quad (25)$$

Taking into account that $\sum_{j=0}^N \Pi_j = 1$, the steady state probabilities are found to be

$$\Pi_j = \frac{D_j}{\sum_{i=0}^N D_i} \quad j = 0, 1, 2, \dots, N, \quad (26)$$

where D_j is the determinant of the square matrix obtained from $\underline{A}$ by deleting the j th row and column.

Consider, for example, the simple case of a unit that can be in only two states, working (0) and failed (1). Let λ and μ be the rates of failure (transitions from state 0 to state 1) and repair (transitions from state 1 to state 0), respectively. The Markov diagram is given by Fig. 1 and the transition matrix takes the form

$$\underline{A} = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

The unit is assumed to be in operation at time $t = 0$, that is, $\underline{C} = [1, 0]$. To compute the transient behavior of the state probability vector from Equation (24), one first computes

the inverse matrix $(s\underline{I} - \underline{A})^{-1}$:

$$\begin{aligned} (s\underline{I} - \underline{A})^{-1} &= \begin{pmatrix} s + \lambda & -\lambda \\ -\mu & s + \mu \end{pmatrix}^{-1} \\ &= \frac{1}{\det[(s\underline{I} - \underline{A})^{-1}]} \begin{pmatrix} s + \lambda & \lambda \\ \mu & s + \mu \end{pmatrix} \\ &= \frac{1}{s^2 + \lambda s + \mu s} \begin{pmatrix} s + \lambda & \lambda \\ \mu & s + \mu \end{pmatrix}. \end{aligned}$$

from which

$$\tilde{P}(s) = \begin{pmatrix} \frac{s + \lambda}{s(s + \lambda + \mu)} & \frac{\lambda}{s(s + \lambda + \mu)} \end{pmatrix}.$$

The roots of the denominator are 0 and $-(\lambda + \mu)$ and applying the inverse Laplace transformation, the state probability vector in the time domain becomes

$$\begin{aligned} \underline{P}(t) &= \begin{pmatrix} \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \\ \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \end{pmatrix}, \end{aligned}$$

where

$$P_0(t) = \frac{\mu}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

is the system instantaneous availability at time t , as found in Equation (4),

$$P_1(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}$$

is the system instantaneous unavailability at time t .

The system steady state probabilities are readily found to be

$$\begin{aligned} \Pi_0 &= \frac{\mu}{\lambda + \mu} \text{ (as in Eq. 3.)} \\ \Pi_1 &= \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

In more complex cases, a system will have many states, some of which correspond to the system functioning according to the required specifications, whereas others represent configurations in which the system fails to perform its function. Letting S denote the subset of states in which the system is functioning

and F the subset of failed states, the system instantaneous availability at time t is computed by simply summing the probabilities of being in a success state at time t , that is,

$$p(t) = \sum_{i \in S} P_i(t) = 1 - q(t) = 1 - \sum_{j \in F} P_j(t). \quad (27)$$

AVAILABILITY OF COMPLEX SYSTEMS: MONTE CARLO SIMULATION

In realistic conditions, the behavior of a system may not respect the Markov property. In these cases, the analysis of the system availability may be effectively carried out by Monte Carlo simulation, which corresponds to performing a virtual experiment in which a large number of identical systems, each one behaving differently due to the stochastic character of the system behavior, are run for test during a given time and their failure occurrences are recorded [13,14] (see also the section titled “Simulation Model Building” in this encyclopedia). This, in principle, is the same procedure adopted in the reliability tests performed on individual units to estimate their failure rates, mean times to failure, or other parameters characteristic of their failure behavior (see also the section titled “Reliability Estimation and Testing” in this encyclopedia); the difference is that for units the tests can be actually done physically in laboratory, at reasonable costs and within reasonable testing times (possibly by resorting to accelerated testing techniques, when necessary; see also **Estimating Intensity and Mean Value Function**), whereas for systems, this is obviously impracticable due to the costs and times involved in systems failures. Thus, instead of making physical tests on a system, the stochastic process of transition among its states is modeled by defining the probabilistic distribution governing the transition process, and a large number of realizations are generated by sampling from it, the times and outcomes of the occurring transitions. Figure 2 shows a number of such realizations on the plane *system configuration* versus *time*: in such

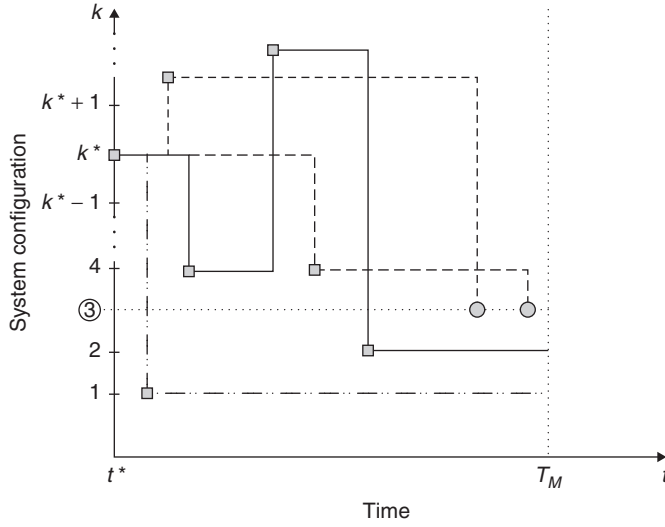


Figure 2. Random walks in the system configuration versus time plane. System configuration 3 is circled as a fault configuration. The squares identify points of transition (t, k) ; the circle bullets identify fault states. The dashed lines identify realizations leading to system failure before the mission time T_M .

a plane, the realizations take the form of random walks made of straight segments parallel to the time axis in-between transitions, when the system is in a given configuration and vertical stochastic jumps to new system configurations at the stochastic times when transitions occur (see also the section titled “Diffusion Processes and Random Walks” in this encyclopedia).

For the purpose of reliability and availability analysis, a subset Γ of the system configurations is identified as the set of fault states. Whenever the system enters one such configuration, its failure is recorded together with its time of occurrence. With reference to a given time t of interest, an estimate of the probability of system failure before such time, that is of the unreliability at time t , can be obtained by the frequency of system failures before t , computed by dividing the number of random walk realizations that record a system failure before t by the total number of random walk realizations simulated.

For the availability analysis, let us consider a generic single realization and suppose that the system enters a failed state $k \in \Gamma$ at time τ_{in} , exiting from it at a next transition at time τ_{out} . The time is suitably discretized in intervals of length Δt and counters are introduced, which accumulate the unavailability contributions in the time channels: a unitary weight is accumulated in the counters for all

the time channels within $[\tau_{in}, \tau_{out}]$. At the end of the simulation of the random walks, the content of each counter divided by the time interval Δt and by the number of random walks simulated, and gives an estimate of the instantaneous unavailability at that counter time.

The Monte Carlo simulation of one single system random walk entails the repeated sampling from the probabilistic transport distribution of the time of occurrence of the next transition and of the new configuration reached by the system as outcome of the transition, starting from the current system configuration. This can be done in two ways, which give rise to the so-called *indirect* and *direct* Monte Carlo approaches [15].

Indirect Simulation Method

The indirect approach consists in sampling first, the time t of a system transition from the conditional probability density $T(t|t', k')$ of the system performing at time t one of its possible transitions out of k' entered at the previous transition at time t' . Then, the transition to the new configuration k actually occurring is sampled from the conditional probability $C(k|t, k')$ that the system enters the new state k given that a transition has occurred at t starting from the system in state k' . The procedure then repeats to the next transition.

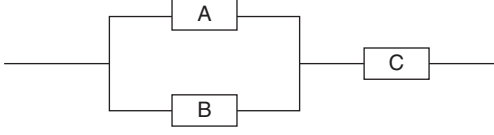


Figure 3. A simple series-parallel logic.

Table 1. Transition rates for the components A, B and C of the system in Figure 3

Initial	Arrival		
	1	2	3
1	0	$\lambda_{1 \rightarrow 2}^{A(B)}$	$\lambda_{1 \rightarrow 3}^{A(B)}$
2	$\lambda_{2 \rightarrow 1}^{A(B)}$	0	$\lambda_{2 \rightarrow 3}^{A(B)}$
3	$\lambda_{3 \rightarrow 1}^{A(B)}$	$\lambda_{3 \rightarrow 2}^{A(B)}$	0

Initial	Arrival			
	1	2	3	4
1	0	$\lambda_{1 \rightarrow 2}^C$	$\lambda_{1 \rightarrow 3}^C$	$\lambda_{1 \rightarrow 4}^C$
2	$\lambda_{2 \rightarrow 1}^C$	0	$\lambda_{2 \rightarrow 3}^C$	$\lambda_{2 \rightarrow 4}^C$
3	$\lambda_{3 \rightarrow 1}^C$	$\lambda_{3 \rightarrow 2}^C$	0	$\lambda_{3 \rightarrow 4}^C$
4	$\lambda_{4 \rightarrow 1}^C$	$\lambda_{4 \rightarrow 2}^C$	$\lambda_{4 \rightarrow 3}^C$	0

Consider for example, the system in Fig. 3, consisting of units A and B in active parallel followed by unit C in series. Units A and B have two distinct modes of operation and a failure state, whereas unit C has three modes of operation and a failure state. For example, if A and B were pumps, the two modes of operation could represent the 50% and 100% flow modes; if C were a valve, the three modes of operation could represent the “fully open,” “half open,” and “closed” modes.

For simplicity of illustration, let us assume that the units’ times of transition between states are exponentially distributed and denoted by $\lambda_{j_i \rightarrow m_i}^i$, the rate of transition of unit i going from its state j_i to the state m_i . Table 1 gives the transition rates matrices in symbolic form for units A, B, and C of the example (with the rate of self-transition $\lambda_{j_i \rightarrow j_i}^i = 0$ by definition).

The units are initially ($t = 0$) in their nominal states, which are labeled with the index

1 (e.g., pumps A and B at 50% flow and valve C fully open), whereas the failure states are labeled with the index 3 for the units A and B and with the index 4 for unit C.

The logic of operation is such that there is one minimal cut set (failure configuration) of order 1, corresponding to unit C in state 4, and one minimal cut set (failure configuration) of order 2, corresponding to both units A and B being in their respective failed states 3.

Let us consider one random walk, starting at $t_0 = 0$ with all units in their nominal states ($j_A = 1, j_B = 1, j_C = 1$). The rate of transition of unit A(B) out of its nominal state 1 is simply

$$\lambda_1^{A(B)} = \lambda_{1 \rightarrow 2}^{A(B)} + \lambda_{1 \rightarrow 3}^{A(B)}, \quad (28)$$

since the transition times are exponentially distributed and states 2 and 3 are the mutually exclusive and exhaustive arrival states of the transition. Similarly, the transition rate of unit C out of its nominal state 1 is

$$\lambda_1^C = \lambda_{1 \rightarrow 2}^C + \lambda_{1 \rightarrow 3}^C + \lambda_{1 \rightarrow 4}^C. \quad (29)$$

It follows, then, that the rate of transition of the system out of its current configuration ($j_A = 1, j_B = 1, j_C = 1$) is

$$\lambda^{(1,1,1)} = \lambda_1^A + \lambda_1^B + \lambda_1^C. \quad (30)$$

The first system transition time t_1 is sampled by applying the inverse transform method for continuous distributions [15]

$$t_1 = t_0 - \frac{1}{\lambda^{(1,1,1)}} \ln(1 - R_t), \quad (31)$$

where

$$R_t \approx U[0, 1).$$

Assuming that $t_1 < T_M$, the system mission time, (otherwise one would proceed to simulate the successive system realization), one needs to determine which transition has occurred, that is, which unit has undergone the transition and to which arrival state. This can be done by resorting to the inverse transform method for discrete distributions [15]. The probabilities of units A, B, C undergoing a transition out of their initial nominal states

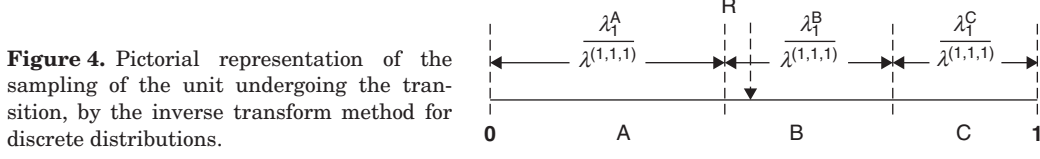
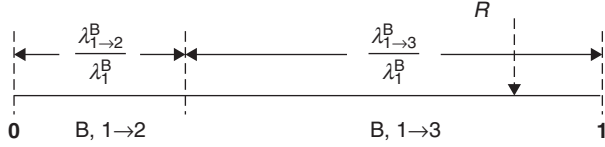


Figure 5. Pictorial representation of the sampling of the arrival state of the transition by the inverse transform method for discrete distributions.



1, given that a transition occurs at time t_1 , are

$$\frac{\lambda_1^A}{\lambda^{(1,1,1)}}, \quad \frac{\lambda_1^B}{\lambda^{(1,1,1)}}, \quad \frac{\lambda_1^C}{\lambda^{(1,1,1)}}, \quad (32)$$

respectively.

Figure 4 shows an example in which the sampled random number $R_C \approx U[0, 1]$ is such that unit B undergoes the transition.

Given that at t_1 unit B undergoes a transition, its arrival state can be sampled by applying again the inverse transform method for discrete distributions, this time, to the set of discrete probabilities $\left\{ \frac{\lambda_{1 \rightarrow 2}^B}{\lambda_1^B}, \frac{\lambda_{1 \rightarrow 3}^B}{\lambda_1^B} \right\}$ of the mutually exclusive and exhaustive arrival states of unit B, as shown in Fig. 5 in which the sampled random number $R_S \approx U[0, 1]$ is assumed to be such that unit B fails (state 3).

As a result of this first transition, at t_1 the system enters configuration (1,3,1). The simulation now continues with the sampling of the next transition time t_2 , based on the updated system transition rate

$$\lambda^{(1,3,1)} = \lambda_1^A + \lambda_3^B + \lambda_1^C. \quad (33)$$

The next transition, then, occurs at

$$t_2 = t_1 - \frac{1}{\lambda^{(1,3,1)}} \ln(1 - R_t), \quad (34)$$

where

$$R_t \approx U[0, 1).$$

Assuming again that $t_2 < T_M$, the unit undergoing the transition and its arrival

state are sampled as before by application of the inverse transform method to the appropriate discrete probabilities.

The trial simulation of the system random walk proceeds through the various transitions from one system configuration to another, until T_M .

As explained earlier, when the system enters a failed configuration $(*, *, 4)$ or $(3, 3, *)$, where the $*$ denotes any state of the unit, its occurrence is recorded. More specifically, from the point of view of the practical implementation into a computer code, the system mission time is subdivided in intervals of length Δt and to each time interval an *unreliability* counter $C^R(t)$ is associated to record the occurrence of a failure: at the time τ when the system enters a fault state, a 1 is collected into all the unreliability counters $C^R(t)$ associated with times t successive to the failure occurrence time, that is $t \in [\tau, T_M]$. After simulating a large number of random walk trials M , an estimate of the system unreliability can be obtained by simply dividing by M and by the time interval Δt , the accumulated contents of the counters $C^R(t)$, $t \in [0, T_M]$. Similarly, the estimation of the system unavailability is obtained by collecting a 1 in the *unavailability* counters $C^A(t)$ associated with all times t successive to the failure, up to the exit of the system from the failed configuration, that is, its repair. As previously explained, the estimate of the system instantaneous unavailability is obtained by dividing by M and by Δt ,

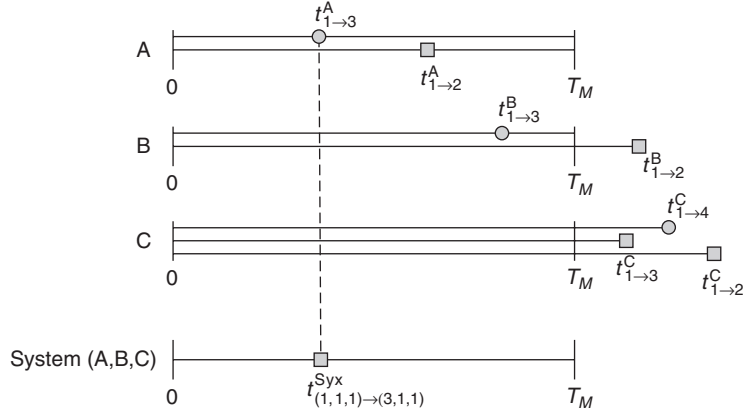


Figure 6. Direct simulation method. The squares identify unit transitions; the circle bullets identify fault states.

the accumulated contents of the counters $C^A(t)$, $t \in [0, T_M]$.

Direct Simulation Method

The direct Monte Carlo simulation method differs from the indirect one in that the system transitions are not sampled by considering the probability distribution governing the transport of the whole system, but rather by sampling directly the times of all possible transitions of all individual units of the system and then arranging the transitions along a timeline in increasing order, in accordance to their times of occurrence. The unit that actually performs the transition is the one corresponding to the first transition in the timeline.

The timeline is updated after each transition occurs, to include the new possible transitions that the transient unit can perform from its new state.

With respect to the previous example of Fig. 5, starting at $t = 0$ with the system in nominal configuration (1,1,1) one would sample the times of all the possible unit transitions (Fig. 6). The simulation then proceeds to the successive times in the list, in correspondence with which, a system transition occurs. After each transition, the timeline is updated by cancelling the times of the transitions relating to the unit, which has undergone the last transition and by inserting the

newly sampled times of the transitions of the same unit from its new state.

Again, when during the trial the system enters a fault configuration 1's are collected in the system unreliability and unavailability counters associated to the time intervals beyond that time, as with the indirect procedure explained above, and in the end, after M trials, the unreliability and unavailability estimates are computed.

Compared to the previous indirect method, the direct approach is more suitable for systems whose units' failure and repair behaviors are represented by different stochastic distribution laws.

On the other hand, it is important to point out that when dependences among units are present (e.g., due to share-load or stand-by configurations), the distribution of the next transition time of a unit may be affected by the transition undergone by another unit, in which case the next transition time of the affected unit (and not only of the transient unit) has also to be resampled after the transition. This can increase the burden, and thus reduce the performance of the direct simulation approach.

REFERENCES

1. Rausand M, Hoyland A. System reliability theory. New York: Wiley; 2004.

2. Ushakov IA. Handbook of reliability engineering. New York: Wiley; 1994.
3. Schneeweiss WG. Reliability modeling. Germany: LiLoLe-Verlag; 2001.
4. Birolini A. Reliability Engineering. Heidelberg: Springer; 2004.
5. Lewis EE. Introduction to reliability engineering. New York: Wiley; 1996.
6. Zio E. Volume 13, An introduction to the basics of reliability and risk analysis. Series in quality, reliability and engineering statistics. Singapore: World Scientific; 2007.
7. Rausand M, Hoyland A. System reliability theory. New York: John Wiley & Sons, Inc.; 2004.
8. Birolini A. Reliability engineering: theory and practice. Heidelberg: Springer; 2004. ISBN 3-540-40287-X.
9. Ushakov IA. Handbook of reliability engineering. New York: John Wiley & Sons, Inc.; 1994.
10. Limnios N, Oprisan G. Semi-Markov processes and reliability. Statistics for industry and technology. Boston: A Birkhäuser book; 2001. ISBN: 978-0-8176-4196-2.
11. Barbu V, Limnios N. Volume 191, Semi-Markov chains and hidden semi-Markov models toward applications. Lecture notes in Statistics. New York: Springer; 2008. ISBN: 978-0-387-73171-1.
12. Howard RA. Dynamic probabilistic systems. New York: John Wiley & Sons, Inc.; 1971.
13. Dubi A. Monte Carlo applications in systems engineering. New York: Wiley; 1999.
14. Marseguer M, Zio E. Basics of the Monte Carlo method with application to system reliability. Germany: LiLoLe- Verlag GmbH (Pbl. Co. Ltd.); 2002.
15. Labeau PE, Zio E. Procedures of Monte Carlo transport simulation for applications in system engineering. Reliab Eng Syst Saf 2002;77:217–228.

AVAILABILITY IN STOCHASTIC MODELS

SOPHIE MERCIER

Laboratoire de Mathématiques et
de leurs Applications–PAU
(UMR CNRS 5142),
Université de Pau et des Pays
de l'Adour, Bâtiment IPRA,
Pau cedex France

This article presents stochastic models and associated tools for the availability assessment of a repairable system. Owing to the limited scope of the article, we concentrate on the classical continuous-time pointwise and asymptotic availabilities: we set $E = \mathcal{U} \cup \mathcal{D}$ with $\mathcal{U} \cap \mathcal{D} = \emptyset$, where $\mathcal{U}$ and $\mathcal{D}$ stand for the up- and down-state sets, respectively; the system state is described by a stochastic process $(X_t)_{t \geq 0}$ with $X_t \in E$, and the point availability and its asymptotic version are defined as

$$A(t) = \mathbb{P}(X_t \in \mathcal{U}) \text{ for } t \geq 0,$$

$$A(\infty) = \lim_{t \rightarrow +\infty} A(t).$$

Other notions of availability may be found in the literature, please see **Point and Interval Availability** and Ref. 1, with lots of references therein.

The article is divided into three sections. The section titled “Availability in Alternating Renewal Models” is devoted to two-state systems, which are considered to be either up or down, with no in-between states. In this case, the system is commonly modeled by a so-called alternating renewal process, which has been extensively studied in the reliability literature [2,3], see also **Alternating Renewal Processes**.

The section titled “Availability in Markov and Semi-Markov Models” deals with systems with finitely many possible states: typically, such systems are formed of components, which can be up or down, leading to more or less degraded up- and down-states. In such a context, the most commonly used

stochastic model is a jump Markov process with a finite state space, which has been thoroughly studied and broadly used in industry [4], see also the section titled “Continuous-time Markov Chains (CTMCs)” in this encyclopedia. Such a model implies that both failure and repair rates should remain constant. This drawback has led to the development and use of semi-Markov processes, which allow for a little more modeling flexibility [5,6]; see also **Semi-Markov Processes**. The section titled “Markov Models” deals with jump Markov processes and the section titled “Semi-Markov Models” with semi-Markov processes.

Finally, the section titled “Availability in Regenerative and Markov Regenerative Models” is devoted to more general systems, which present regenerative and Markov regenerative properties, with no restrictive condition on the state space or other: between (Markov) regeneration points, the system may have a very general behavior, see [7], **Regenerative Processes**, and **Markov Regenerative Processes** for details. As we shall see, the (Markov) regeneration property then allows to concentrate on the behavior of the system between (Markov) regeneration points, to derive both point and asymptotic availabilities.

In order to illustrate the different stochastic models and associated tools, a small and educational example is used all over the article, under various assumptions: a series system is considered, which is formed of two stochastically independent components A and B , with respective failure and repair rates $(\lambda_A(x), \mu_A(x))$ and $(\lambda_B(x), \mu_B(x))$. (The components’ repair durations and times to failure are hence assumed to admit a density with respect to Lebesgue measure). The system always starts from its perfect working state. Owing to its structure, the system is down as soon as one component is down. A repair immediately begins at the failure and puts a component back to its perfect working state (as good as new repairs). According to the cases, the failure of one component

(and hence of the system) may involve the suspension of the other or not, that is by failure of one component, the other one may go on aging and undertake failure or not, with an eventually reduced failure rate. In each case, both components may be entirely renewed (the repair is then said to be complete) or only the down component. Finally, the failure and repair rates may be constant or general, and the components identical or not.

A very good reference for a deeper insight on the models presented here and for other examples is Ref. 7; for the stochastic tools, see also Refs 5 and 8–10.

Throughout the article, if T stands for a generic random variable (r.v.), $\mathbb{P}_T$ stands for its distribution, $\mathbb{E}(T)$ for its expectation, $F_T(t) = \mathbb{P}(T \leq t)$ for its cumulative distribution function, and $\bar{F}_T(t) = \mathbb{P}(T > t) = 1 - F_T(t)$ for its survival function.

AVAILABILITY IN ALTERNATING RENEWAL MODELS

The Example

We assume both components to be identical with a common failure rate $\lambda(x)$ and a repair to be complete, with repair rate $\mu(x)$; we consequently assume that the repair duration always is the same, independent of the degradation state of the system. The length of an up-period is the minimum of two independent r.v.s with common rate $\lambda(x)$ and we have a succession of alternating up- and down independent periods, with respective associate rates $2\lambda(x)$ and $\mu(x)$. The behavior of the system may then be modeled by a so-called alternating renewal process.

The General Case

A system is considered, which evolves according to an alternating renewal process $(X_t)_{t \geq 0}$, where we set $X_t = 1$ when the system is up at time t and $X_t = 0$ when it is down. The system is assumed to start from its perfect working state at time $t = 0$. The successive up-periods are denoted by $U_1, \dots, U_n, \dots$ and the down ones by $V_1, \dots, V_n, \dots$. Setting $T_n = \sum_{i=1}^n (U_i + V_i)$ for $n \geq 1$, the points

$T_0 = 0, T_1, \dots, T_n, \dots$ appear as renewal points for the process $(X_t)_{t \geq 0}$. In order to avoid the trivial case $T_0 = T_1 = T_2 = \dots = 0$ almost surely and following Ref. 5, we assume that $\mathbb{P}(T_1 = 0) < 1$.

With this setting, the system point availability is

$$\begin{aligned} A(t) &= \mathbb{P}(X_t = 1) \\ &= \sum_{n=0}^{+\infty} \mathbb{P}(T_n \leq t < T_{n+1}) \end{aligned}$$

for all $t \geq 0$.

The main tool for its study is the renewal equation (1) fulfilled by $A(t)$, provided just later. To derive it, we classically separate the cases where the first renewal arrives after or before t and we get:

$$\begin{aligned} A(t) &= \mathbb{P}(X_t = 1; t < T_1) + \mathbb{P}(X_t = 1; t \geq T_1) \\ &= \mathbb{P}(t < U_1) \\ &\quad + \int_{[0,t]} \mathbb{P}(X_t = 1 | T_1 = u) \mathbb{P}_{T_1}(du). \end{aligned}$$

Using the regeneration property at time T_1 (see **Renewal Function and Renewal-Type Equations** or Ref. 2 for more details), we easily get

$$\begin{aligned} &\int_{[0,t]} \mathbb{P}(X_t = 1 | T_1 = u) \mathbb{P}_{T_1}(du) \\ &= \int_{[0,t]} A(t - u) \mathbb{P}_{T_1}(du) = (A * \mathbb{P}_{T_1})(t) \end{aligned}$$

and

$$A(t) = \bar{F}_{U_1}(t) + (A * \mathbb{P}_{T_1})(t), \quad (1)$$

where $\mathbb{P}_{T_1}(du) = (\mathbb{P}_{U_1+V_1})(du) = (\mathbb{P}_{U_1} * \mathbb{P}_{V_1})(du)$ and $*$ stands for the standard convolution.

Apart from very special cases (e.g., exponential distributions), the renewal equation (1) cannot be solved explicitly and numerical procedures have to be developed [11,12].

Limit theorems for solutions of renewal equations (see **Limit Theorems for Renewal Processes**) allow to get the asymptotic availability from Equation (1): setting $MUT = \mathbb{E}(U_1)$ to be the Mean Up Time of

the system on a cycle and $MDT = \mathbb{E}(V_1)$ to be its Mean Down Time, we get:

$$\begin{aligned} A(\infty) &= \frac{\int_0^{+\infty} \bar{F}_{U_1}(u) du}{\mathbb{E}(T_1)} \\ &= \frac{\mathbb{E}(U_1)}{\mathbb{E}(U_1) + \mathbb{E}(V_1)} \\ &= \frac{MUT}{MUT + MUT}, \end{aligned} \quad (2)$$

assuming $\mathbb{E}(U_1) + \mathbb{E}(V_1) < +\infty$ and the distribution of $T_1 = U_1 + V_1$ to be nonlattice [13].

Back to the Example

The times to failure of both components are here assumed to be gamma distributed $\Gamma(a_u, b_u)$, with the following p.d.f. (probability density function):

$$f_u(x) = \frac{1}{(b_u)^{a_u} \Gamma(a_u)} x^{a_u-1} e^{-x/b_u} \mathbf{1}_{\mathbb{R}_+}(x)$$

and parameters $(a_u, b_u) = (2, 1/3)$ (mean = $a_u b_u \simeq 0.6667$). The repairs also are gamma distributed with parameters $(a_d, b_d) = (3, 1/36)$ (mean = $a_d b_d \simeq 0.08333$). The point availability has been computed by discretization of Equation (1) and by Monte Carlo simulations. The results of both methods are plotted in Fig. 1, where MC stands for Monte Carlo simulation (5×10^4 histories), MC Sup

and MC Inf stand for the upper and lower bounds of the associated 95% confidence band and DM stands for the discretization method. (Similar notations are used all over the article). The asymptotic availability provided by (2) is also indicated in Fig. 1, with $A(\infty) \simeq 0.833$.

In most cases, a system may have several degraded up-states and/or several down-states and cannot hence be modeled by an alternating renewal process. The next section is devoted to such multistate systems, with Markovian or semi-Markovian degradation.

AVAILABILITY IN MARKOV AND SEMI-MARKOV MODELS

Markov Models

The Example: Case of Constant Failure and Repair Rates. We here come back to our example, in the case of different components with constant failure and repair rates ($\lambda_A(x) \equiv \bar{\lambda}_A$, $\mu_A(x) \equiv \bar{\mu}_A$, $\lambda_B(x) \equiv \bar{\lambda}_B$, $\mu_B(x) \equiv \bar{\mu}_B$). In case of failure of one component, the other one is suspended and cannot fail. The evolution of the system may then be modeled by a continuous-time Markov chain (CTMC) $(X_t)_{t \geq 0}$ with range in $E = \{(1, 1), (1, 0), (0, 1)\}$, where the first place refers to component A and the second one to component B, with an “1” indicating an up component, and an “0” a down one.

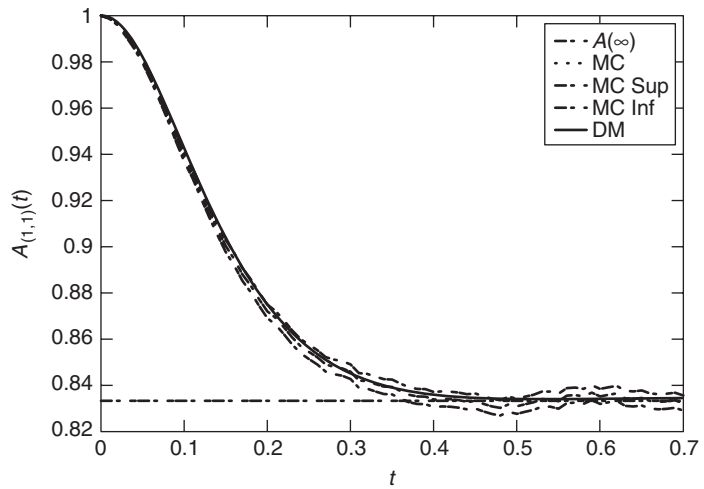


Figure 1. Point and asymptotic availabilities, case of an alternating renewal process.

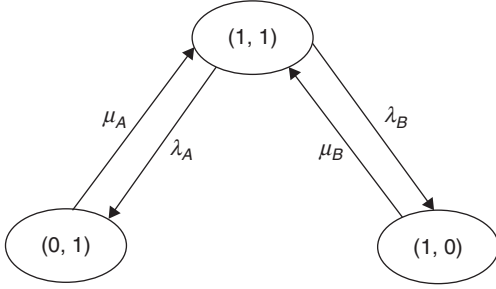


Figure 2. The Markov graph in case of constant failure and repair rates.

The corresponding Markov graph is provided in Fig. 2.

The General Case. In the general case, a system is considered, which evolves in time according to a CTMC $(X_t)_{t \geq 0}$ with range in a finite state space E . We set $\mathcal{A}$ to be the generator matrix of $(X_t)_{t \geq 0}$ and $(P_t)_{t \geq 0}$ to be its transition semigroup:

$$P_t(i, j) = \mathbb{P}_i(X_t = j)$$

for all $i, j \in E$ and all $t \geq 0$, where $\mathbb{P}_i$ is the conditional distribution given that $X_0 = i$, see the section titled “Continuous-time Markov Chains (CTMCs)” in this encyclopedia for more details.

The point availability of the system starting from state i is

$$A_i(t) = \sum_{j \in \mathcal{U}} P_t(i, j) = \sum_{j \in \mathcal{U}} (e^{t\mathcal{A}})(i, j), \quad (3)$$

where $e^{t\mathcal{A}}$ refers to the matrix exponentiation.

In case $(X_t)_{t \geq 0}$ is irreducible (and hence recurrent, because of the finite state space), the CTMC $(X_t)_{t \geq 0}$ admits an unique stationary probability measure π , such that $\pi\mathcal{A} = 0$ and $\sum_{i \in E} \pi(i) = 1$, see **Asymptotic Behavior of Continuous-Time Markov Chains**. The asymptotic availability is then independent of the initial state with

$$A(\infty) = \sum_{i \in \mathcal{U}} \pi(i). \quad (4)$$

Back to the Example. Owing to the Markov graph provided in Fig. 2, the generator matrix of $(X_t)_{t \geq 0}$ is given by:

$$\mathcal{A} = \begin{pmatrix} -\bar{\lambda}_A - \bar{\lambda}_B & \bar{\lambda}_B & \bar{\lambda}_A \\ \bar{\mu}_B & -\bar{\mu}_B & 0 \\ \bar{\mu}_A & 0 & -\bar{\mu}_A \end{pmatrix}.$$

The point availability of the system is

$$\begin{aligned} A_{(1,1)}(t) &= \mathbb{P}_{(1,1)}(X_t = (1, 1)) \\ &= e^{t\mathcal{A}}((1, 1), (1, 1)). \end{aligned}$$

In case $\bar{\lambda}_A = \bar{\lambda}_B = \bar{\lambda}$ and $\bar{\mu}_A = \bar{\mu}_B = \bar{\mu}$, this easily provides

$$\begin{aligned} A_{(1,1)}(t) &= \frac{\bar{\mu}}{2\bar{\lambda} + \bar{\mu}} + \frac{2\bar{\lambda}}{2\bar{\lambda} + \bar{\mu}} e^{-(2\bar{\lambda} + \bar{\mu})t} \quad \text{and} \\ A(\infty) &= \frac{\bar{\mu}}{2\bar{\lambda} + \bar{\mu}}. \end{aligned}$$

In case of different rates for A and B , we get

$$A(\infty) = \frac{\mu_A \mu_B}{\lambda_A \mu_B + \lambda_B \mu_A + \mu_A \mu_B},$$

and an easy but cumbersome expression for $A_{(1,1)}(t)$, which we do not provide.

Both point and asymptotic availabilities are plotted in Fig. 3 for $\bar{\lambda}_A = 1$, $\bar{\lambda}_B = 2$, $\bar{\mu}_A = 10$ and $\bar{\mu}_B = 15$, with $A(\infty) \simeq 0.811$.

In the Markovian case, both point and asymptotic availabilities hence have an easy expression with respect to the CTMC transition probabilities and stationary probability measure, see Equations (3) and (4). Though the transition probabilities have an explicit expression with respect to the generator matrix (Eq. 3), their numerical evaluation

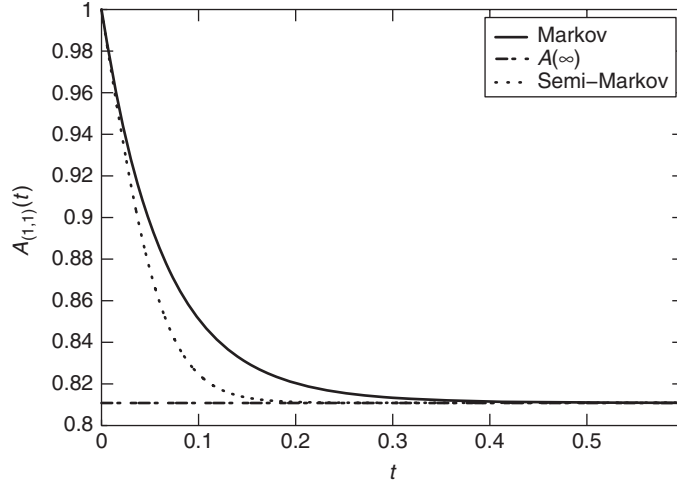


Figure 3. Point and asymptotic availabilities, Markov & semi-Markov cases.

however leads to real difficulties in case of large Markov systems, due to a rapid explosion of the size of the state space with the number of components of the system. This has lead to an extensive literature devoted to their numerical assessment (see *Computational Methods for CTMCs* or Refs 14 and 15)

Another drawback of the Markovian models is the underlying assumption of constant failure and repair rates. Such a restrictive assumption may be partially removed by semi-Markov processes, as shown in the next section.

Semi-Markov Models

The Example: Case of Constant Failure Rates and General Repair Rates. We consider here, the same example as in section titled “The Example: Case of Constant Failure and Repair Rates” except the fact that the repair rates are now general $[\mu_A(x) \text{ and } \mu_B(x)]$, while both failure rates remain constant ($\bar{\lambda}_A$ and $\bar{\lambda}_B$). Just as in section titled “The Example: Case of Constant Failure and Repair Rates”, no further failure is possible when the system is down.

The possible changes in the system state are due to

- the failure of one component (with the other one suspended in its up-state);

- the end of repair of one component (with the other component up).

It is then easy to see that at each transition time, the system forgets its past; indeed, as the failure rates are constant, the repair of one component (with the other one up) puts the system back to its perfect working state. This means that the system fulfills the Markov property each time its state changes and hence behaves according to a semi-Markov process $(X_t)_{t \geq 0}$ (see *Semi-Markov Processes*).

The General Case. In the general case, a system is considered, which evolves in time according to a semi-Markov process $(X_t)_{t \geq 0}$ with range in E (finite) and with semi-Markov transition kernel $(q(i, j, dt))_{i, j \in E}$: we recall that $q(i, j, dt) = \mathbb{P}_i(X_{T_1} = j, T_1 \in dt)$, where T_1 stands for the first jump time of the process $(X_t)_{t \geq 0}$ (see *Semi-Markov Processes* for details). Setting $T_0 = 0 \leq T_1 \leq T_2 \leq \dots$ to be the successive jump times of $(X_t)_{t \geq 0}$, we assume that $\mathbb{P}_i(T_0 = T_1 = T_2 = \dots = 0) = 0$ for all $i \in E$, which here again avoids trivialities.

The point availability of the semi-Markov system starting from state i is

$$A_i(t) = \mathbb{P}_i(X_t \in \mathcal{U}) = \sum_{j \in \mathcal{U}} \mathbb{P}_i(X_t = j)$$

for all $i \in E$. By separating the cases where the first jump arrives after or before t as in the case of an alternating renewal process, we get

$$\begin{aligned} A_i(t) &= \mathbb{P}_i(X_t \in \mathcal{U}, T_1 > t) + \mathbb{P}_i(X_t \in \mathcal{U}, T_1 \leq t) \\ &= \mathbf{1}_{\mathcal{U}}(i) \mathbb{P}_i(T_1 > t) \\ &\quad + \sum_{k \in E} \int_{[0,t]} \mathbb{P}_i(X_t \in \mathcal{U} | T_1 = u, X_{T_1} = k) \\ &\quad \times q(i, k, du) \end{aligned}$$

Applying the Markov property at time T_1 , we have

$$\begin{aligned} \mathbb{P}_i(X_t \in \mathcal{U} | T_1 = u, X_{T_1} = k) \\ = \mathbb{P}_k(X_{t-u} \in \mathcal{U}) = A_k(t-u) \end{aligned}$$

This provides

$$A_i(t) = \mathbf{1}_{\mathcal{U}}(i) \mathbb{P}_i(T_1 > t) + (A * q)(i, t), \quad (5)$$

where we set

$$(A * q)(i, t) = \sum_{k \in E} \int_{[0,t]} A_k(t-u) q(i, k, du). \quad (6)$$

Equation (5) for $i \in E$ and $t \geq 0$ are known as *Markov renewal equations*, which have no explicit solutions in the general case and have to be solved numerically (see **Limit Theorems for Markov Renewal Processes** and Refs 1 and 16).

In the case that the semi-Markov process $(X_t)_{t \geq 0}$ is irreducible and that the sojourn times are nonarithmetic with finite means, the process $(X_t)_{t \geq 0}$ admits a unique stationary probability measure π , which is known to be identical to the unique stationary probability measure of a CTMC $(Y_t)_{t \geq 0}$, which has the same transition matrix $(\mathbb{P}_i(X_{T_1} = j) = \mathbb{P}_i(Y_{T_1} = j))$ and same mean sojourn times $\mathbb{E}_i(T_1)$ as $(X_t)_{t \geq 0}$ (see Refs 5 and 8 for details). This implies that both asymptotic availabilities for the semi-Markov process $(X_t)_{t \geq 0}$ and for the Markov process $(Y_t)_{t \geq 0}$ are identical.

Back to the Example. The semi-Markovian kernel associated with $(X_t)_{t \geq 0}$ is

$$(q(i, j, dt))_{i, j \in E} = \begin{pmatrix} 0 & \bar{\lambda}_B e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} dt & \bar{\lambda}_A e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} dt \\ f_{\mu_B}(t) dt & 0 & 0 \\ f_{\mu_A}(t) dt & 0 & 0 \end{pmatrix},$$

where $f_{\mu_A}(t)$ and $f_{\mu_B}(t)$ stand for the respective p.d.f.s associated with r.v.s with respective hazard rates $\mu_A(t)$ and $\mu_B(t)$, where

$$f_{\mu_A}(t) = \mu_A(t) e^{-\int_0^t \mu_A(u) du},$$

for all $t \geq 0$ and a similar expression for $f_{\mu_B}(t)$, with $\mu_B(t)$ substituted to $\mu_A(t)$.

The Markov renewal equations may here be written as

$$A_{(1,0)}(t) = \int_0^t A_{(1,1)}(t-u) f_{\mu_B}(u) du, \quad (7)$$

$$A_{(0,1)}(t) = \int_0^t A_{(1,1)}(t-u) f_{\mu_A}(u) du, \quad (8)$$

$$\begin{aligned} A_{(1,1)}(t) &= e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} \\ &\quad + \int_0^t A_{(1,0)}(t-u) \bar{\lambda}_B e^{-(\bar{\lambda}_A + \bar{\lambda}_B)u} du \\ &\quad + \int_0^t A_{(0,1)}(t-u) \bar{\lambda}_A e^{-(\bar{\lambda}_A + \bar{\lambda}_B)u} du. \end{aligned} \quad (9)$$

Substituting (7) and (8) in (9) easily provides

$$\begin{aligned} A_{(1,1)}(t) &= e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} \\ &\quad + \int_0^t A_{(1,1)}(v) G(t-v) dv \end{aligned}$$

with

$$\begin{aligned} G(v) &= \int_0^v (\bar{\lambda}_B f_{\mu_B}(v-u) + \bar{\lambda}_A f_{\mu_A}(v-u)) \\ &\quad \times e^{-(\bar{\lambda}_A + \bar{\lambda}_B)u} du, \end{aligned}$$

which we solve by discretization. The results are provided in the same figure as for the Markovian case (Fig. 3), with the same constant failure rates $(\bar{\lambda}_A = 1, \bar{\lambda}_B = 2)$ and gamma repair rates $\Gamma(3, 1/30)$ and

$\Gamma(3.5, 1/52.5)$ with the same means as in the Markovian case from section titled “Back to the Example” in the section titled “Markov Models” (1/10 and 1/15, respectively). As expected, we can observe that the asymptotic availability is the same in both cases, with a slower convergence in the Markovian case.

In a semi-Markov model, the system forgets its past each time it changes state. In the example, assume that the failure rates are not constant any more. In that case, at the end of repair of the down component, the suspended one is not as good as new and it needs to be repaired for the system to be entirely renewed (and for the system to forget its past). If the duration of this repair is independent of the degradation level of the suspended component, the system still is semi-Markovian. However, under the more realistic assumption of a repair depending on the degradation level, this is not true any more. This shows a limitation of the modeling power of the semi-Markov processes. (A classical other limitation is that a parallel two-unit system formed of two independent semi-Markovian components is not semi-Markovian any more). We now come to regenerative and semiregenerative models, which allow for more flexibility.

AVAILABILITY IN REGENERATIVE AND MARKOV REGENERATIVE MODELS

Regenerative Models

The Example. We consider here the case of general failure and repair rates, with suspension of the up component in case of failure. At failure, the system is completely repaired. A single repairman is considered so that the repair durations of both components are added. The repair of one component now depends on its degradation level: for a down component, its repair rate is $\mu_A(x)$ or $\mu_B(x)$, as before. When a component has been functioning for a duration u , the repair rate to bring it back to its perfect working state is some $\tilde{\mu}_A(x, u)$ (or $\tilde{\mu}_B(x, u)$), where $\tilde{\mu}_A(x, u)$ is some decreasing function in u

with $\lim_{u \rightarrow +\infty} \tilde{\mu}_A(x, u) = \mu_A(x)$ (the same for $\tilde{\mu}_B(x, u)$).

Under the previous assumptions, the state $(1, 1)$ is a regenerative state in the sense that each time the system enters this state, it starts again in a similar way as from the beginning and forgets its past. The periods between two successive arrivals in state $(1, 1)$ are called *cycles* and the process $(X_t)_{t \geq 0}$ appears as a regenerative one (see **Regenerative Processes**).

The General Case. In the general case, a system is considered, which evolves in time according to a regenerative process $(X_t)_{t \geq 0}$, with $(T_n)_{n \in \mathbb{N}}$ as regeneration times. Following the definition of Çinlar [5], this means that $(T_n)_{n \in \mathbb{N}}$ are the points of a renewal process such that

1. the T_n 's are stopping times adapted to $(X_t)_{t \geq 0}$ (and $\mathbb{P}(T_1 = 0) < 1$);
2. at each T_n , the future process $(X_{t+T_n})_{t \geq 0}$ given the past up to T_n (namely given the σ -algebra generated by $\{X_u, u \leq T_n\}$), is identically distributed as $(X_t)_{t \geq 0}$.

See **Regenerative Processes** or Ref. 5 for more details. This means that at each T_n , a regenerative system starts again as from the beginning and forgets its past. The evolution of the system between two regeneration points may here be very general.

Using the regeneration property at time T_1 , the point availability satisfies the following renewal equation:

$$\begin{aligned} A(t) &= \mathbb{P}(X_t \in \mathcal{U}, T_1 > t) + \mathbb{P}(X_t \in \mathcal{U}, T_1 \leq t) \\ &= \mathbb{P}(X_t \in \mathcal{U}, T_1 > t) \\ &\quad + \int_{[0, t]} A(t-u) \mathbb{P}_{T_1}(du). \end{aligned} \quad (10)$$

Assuming the distribution of T_1 to be non-lattice and $\mathbb{E}(T_1) < +\infty$, we derive [5]:

$$\begin{aligned} A(\infty) &= \frac{\int_0^{+\infty} \mathbb{P}(X_t \in \mathcal{U}, T_1 > t) dt}{\mathbb{E}(T_1)} \\ &= \frac{\mathbb{E}\left(\int_0^{T_1} \mathbf{1}_{\mathcal{U}}(X_t) dt\right)}{\mathbb{E}(T_1)} = \frac{\text{MUT}}{\text{MUT} + \text{MDT}}, \end{aligned} \quad (11)$$

where MUT and MDT, are respectively, the cumulated Mean Up Time and Mean Down Time of the system on a cycle.

Back to the Example. We assume that both times to failure are gamma distributed with the same means as in section titled “Availability in Markov and Semi-Markov Models” (where the failure rates were constant), and we take $\Gamma(2, 1/2)$ and $\Gamma(2, 1/4)$. The repair durations in case of failure are identically distributed as in the semi-Markovian case: $\Gamma(a_{R_A}, b_{R_A})$ and $\Gamma(a_{R_B}, b_{R_B})$ with $(a_{R_A}, b_{R_A}) = (3, 1/30)$ and $(a_{R_B}, b_{R_B}) = (3.5, 1/52.5)$, and respective means $m_{R_A} = a_{R_A} b_{R_A}$ and $m_{R_B} = a_{R_B} b_{R_B}$. When component A has been functioning for u time units, its repair duration is gamma distributed: $\Gamma(a_{R_A}, b_{R_A}(1 - \frac{1}{1+u^\alpha}))$, with mean $m_{R_A}(1 - \frac{1}{1+u^\alpha})$, where we assume $\alpha = 1/8$. The distribution of the repair duration is similar for component B, with (a_{R_B}, b_{R_B}) substituted to (a_{R_A}, b_{R_A}) and the same α .

A cycle begins with an up-period of length $U_1 = \min(Z_{\lambda_A}, Z_{\lambda_B})$, where Z_v stands for an r.v. with hazard rate $v(t)$, and where Z_{λ_A} and Z_{λ_B} are independent. (Other natural conditions of independence will be assumed further on, which will not be detailed). Then comes a down-period with length

$$\begin{aligned} V_1 &= \left(Z_{\mu_A} + Z_{\mu_B}(\cdot, Z_{\lambda_A})\right) \mathbf{1}_{\{Z_{\lambda_A} < Z_{\lambda_B}\}} \\ &\quad + \left(Z_{\mu_B} + Z_{\mu_A}(\cdot, Z_{\lambda_B})\right) \mathbf{1}_{\{Z_{\lambda_A} \geq Z_{\lambda_B}\}}, \end{aligned}$$

where, given Z_{λ_A} , the distribution of $Z_{\mu_B}(\cdot, Z_{\lambda_A})$ is $\Gamma\left(a_{R_B}, b_{R_B}\left(1 - \frac{1}{1+(Z_{\lambda_A})^\alpha}\right)\right)$ with

mean

$$\mathbb{E}\left(Z_{\mu_B}(\cdot, Z_{\lambda_A}) | Z_{\lambda_A}\right) = m_{R_B} \left(1 - \frac{1}{1 + (Z_{\lambda_A})^\alpha}\right). \quad (12)$$

Similarly,

$$\mathbb{E}\left(Z_{\mu_A}(\cdot, Z_{\lambda_B}) | Z_{\lambda_B}\right) = m_{R_A} \left(1 - \frac{1}{1 + (Z_{\lambda_B})^\alpha}\right).$$

Note that though up- and down periods are alternating, their durations are not independent so that an alternating renewal process cannot model the system evolution.

We have

$$\begin{aligned} \mathbb{P}(X_t \in \mathcal{U}, T_1 > t) &= \mathbb{P}(U_1 > t) \\ &= \bar{F}_{\lambda_A}(t) \bar{F}_{\lambda_B}(t), \end{aligned}$$

and the relation $T_1 = U_1 + V_1$ then allows to get a discretized version of Equation (10), which we solve numerically. The results are plotted in Fig. 4, as well as those by Monte Carlo simulations.

The asymptotic availability is computed via (11), with

$$\begin{aligned} \text{MUT} &= \mathbb{E}(U_1) = \mathbb{E}(\min(Z_{\lambda_A}, Z_{\lambda_B})) \\ &= \int_0^{+\infty} \bar{F}_{\lambda_A}(t) \bar{F}_{\lambda_B}(t) dt \stackrel{\text{def}}{=} U_{AB} \end{aligned}$$

and $\text{MDT} = \mathbb{E}(V_1) = R_{AB} + R_{BA}$, where

$$\begin{aligned} R_{AB} &= \mathbb{E}\left(\left(Z_{\mu_A} + Z_{\mu_B}(\cdot, Z_{\lambda_A})\right) \mathbf{1}_{\{Z_{\lambda_A} < Z_{\lambda_B}\}}\right) \\ &= \mathbb{E}(Z_{\mu_A}) \mathbb{P}(Z_{\lambda_A} < Z_{\lambda_B}) \\ &\quad + \mathbb{E}\left(\mathbb{E}\left(Z_{\mu_B}(\cdot, Z_{\lambda_A}) \mathbf{1}_{\{Z_{\lambda_A} < Z_{\lambda_B}\}} | Z_{\lambda_A}\right)\right) \\ &= m_{R_A} \int_0^{+\infty} f_{\lambda_A}(u) \bar{F}_{\lambda_B}(u) du \\ &\quad + \mathbb{E}\left(\mathbb{E}\left(Z_{\mu_B}(\cdot, Z_{\lambda_A}) | Z_{\lambda_A}\right) \times \mathbb{E}\left(\mathbf{1}_{\{Z_{\lambda_A} < Z_{\lambda_B}\}} | Z_{\lambda_A}\right)\right) \end{aligned}$$

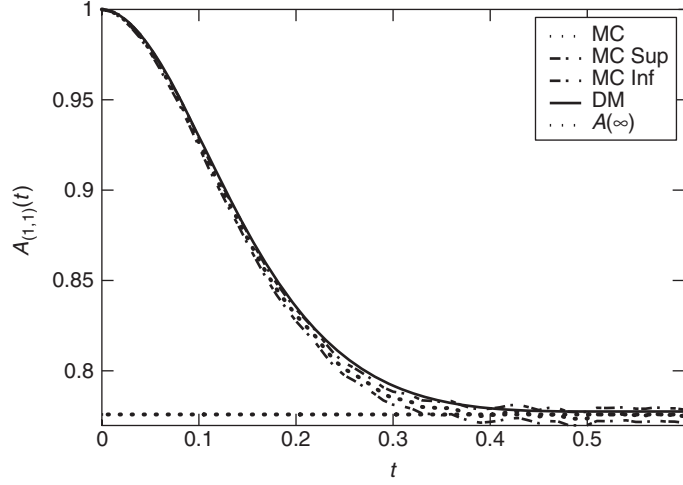


Figure 4. Point and asymptotic availabilities, regenerative case.

$$\begin{aligned}
 &= m_{R_A} \int_0^{+\infty} f_{\lambda_A}(u) \bar{F}_{\lambda_B}(u) du \\
 &\quad + \mathbb{E} \left(m_{R_B} \left(1 - \frac{1}{1 + (Z_{\lambda_A})^\alpha} \right) \bar{F}_{\lambda_B}(Z_{\lambda_A}) \right) \\
 &= \int_0^{+\infty} \left(m_{R_A} + m_{R_B} \left(1 - \frac{1}{1 + u^\alpha} \right) \right) \\
 &\quad \times f_{\lambda_A}(u) \bar{F}_{\lambda_B}(u) du
 \end{aligned} \tag{13}$$

using Equation (12) for (13). A similar expression is valid for R_{BA} and we finally have:

$$A(\infty) = \frac{U_{AB}}{U_{AB} + R_{AB} + R_{BA}}$$

where each quantity is now easily computable. This provides: $A(\infty) \simeq 0.776$.

As already mentioned, regenerative processes allow for a great flexibility as for the system behavior between regeneration points. However, the numerical assessment of the asymptotic availability requires the computation of both mean up time and mean down time on a cycle, which is not always that easy, even in rather simple cases. (As for the point availability, it is generally still more complicated). In some cases, the system may forget its past when entering several states and not only one single state, which may lead to a so-called Markov regenerative process. Though most of the Markov regenerative processes used in reliability also are

regenerative processes, the Markov regeneration property usually leads to different expressions for the point and/or asymptotic availabilities from those obtained with the simple regeneration property, which may be easier to compute. The discussion on Markov Regenerative models follows.

Markov Regenerative Models

The Example. We first come back to our example in a slightly modified version: both failure rates are constant ($\lambda_A(x) \equiv \bar{\lambda}_A$, $\lambda_B(x) \equiv \bar{\lambda}_B$), and the repair rates are general ($\mu_A(x)$ and $\mu_B(x)$) as in section titled “The Example: Case of Constant Failure Rates and General Repair Rates.” However, in case of failure of one component, the other one may now fail while the down component is repaired, at a lower rate however ($\bar{\lambda}'_A$ and $\bar{\lambda}'_B$ with $\bar{\lambda}'_A \leq \bar{\lambda}_A$ and $\bar{\lambda}'_B \leq \bar{\lambda}_B$). There is one single repairman and the repair of one component is completed up to its end, even in case of arrival of a new failure. To model it, we add two new states to our system: state $(0_R, 0)$ which means that component A is down under repair and that component B is down, waiting for a repair; the same for state $(0, 0_R)$.

Entrance in state $(1, 1)$ clearly is a regenerative point and $(X_t)_{t \geq 0}$ is a regenerative process. However, the mean length of a cycle, here, is a little complicated by the fact that lots of different scenarios are

possible, such as $(1, 1) \rightarrow (0, 1) \rightarrow (0_R, 0) \rightarrow (1, 0) \rightarrow (0, 0_R) \rightarrow (0, 1) \rightarrow \dots \rightarrow (1, 1)$. An alternative model is however possible, which leads to simpler computations; indeed, because of the constant failure rates, the system forgets its past each time it enters states $(1, 1)$, $(1, 0)$ and $(0, 1)$. Note that the system, however, does not fulfill the same property when entering states $(0_R, 0)$ and $(0, 0_R)$. In such cases, the repair of one component has indeed already begun, with a general repair rate. The process $(X_t)_{t \geq 0}$ is consequently not semi-Markovian. One can nevertheless consider a new process $(Y_t)_{t \geq 0}$ defined on a new state space $F = \{1, 2, 3\}$, which enters state 1, 2, and 3 each time $(X_t)_{t \geq 0}$ enters the states $(1, 1)$, $(1, 0)$, and $(0, 1)$, respectively. The possible transitions for $(X_t)_{t \geq 0}$ from state $(1, 1)$ are $(1, 1) \rightarrow (1, 0)$ and $(1, 1) \rightarrow (0, 1)$ with respective rates $\bar{\lambda}_B$

$$(q(i, j, dt))_{i, j \in F} = \begin{pmatrix} 0 & \bar{\lambda}_B e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} dt & \bar{\lambda}_A e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} dt \\ p_1 f_{\mu_B}(t) dt & 0 & (1 - p_1) f_{\mu_B}(t) dt \\ p_2 f_{\mu_A}(t) dt & (1 - p_2) f_{\mu_A}(t) dt & 0 \end{pmatrix},$$

where $p_2 = \mathbb{P}(Z_{\mu_A} \leq Z_{\bar{\lambda}_B})$, same notations, and where $f_{\mu_A}(t)$ and $f_{\mu_B}(t)$ have been defined in Subsection 2.2.3. Though the process $(Y_t)_{t \geq 0}$ does not fully describe the system evolution, (the process $(X_t)_{t \geq 0}$ may evolve while $(Y_t)_{t \geq 0}$ remains constant), the process $(X_t)_{t \geq 0}$ forgets its past at each jump of the semi-Markovian process $(Y_t)_{t \geq 0}$. This means that $(X_t)_{t \geq 0}$ is a Markov regenerative process, with $(Y_t)_{t \geq 0}$ as imbedded semi-Markov process (or $(T_n, Y_{T_n})_{n \in \mathbb{N}}$ as imbedded Markov renewal process, equivalently).

The General Case. In the general case, a system is considered, which evolves in time according to a Markov regenerative (or semiregenerative) process $(X_t)_{t \geq 0}$, with state space E . Following Ref. 5, this means that $t \rightarrow X_t(\omega)$ is right continuous and has left-hand limits for almost all ω and that there exists a Markov renewal process $(T_n, Y_n)_{n \in \mathbb{N}}$ on a state space F (here assumed finite) such that (i) the T_n 's are stopping times adapted to $(X_t)_{t \geq 0}$ [with $\mathbb{P}_i(T_0 = T_1 = T_2 = \dots = 0) = 0$ for all $i \in F$]; (ii) for each T_n , the r.v. Y_n is measurable with respect to the

and $\bar{\lambda}_A$. The possible transitions for $(Y_t)_{t \geq 0}$ from state 1 hence are $1 \rightarrow 2$ and $1 \rightarrow 3$ with the same respective rates, $\bar{\lambda}_B$ and $\bar{\lambda}_A$. From state $(1, 0)$, the process $(X_t)_{t \geq 0}$ may go back to state $(1, 1)$ with probability $p_1 = \mathbb{P}(Z_{\mu_B} \leq Z_{\bar{\lambda}_A})$, where $Z_{\bar{\lambda}_A}$ and Z_{μ_B} stand for independent r.v.s with respective hazard rates $\bar{\lambda}_A$ and $\mu_B(t)$. From state $(1, 0)$, the process $(X_t)_{t \geq 0}$ may also go to state $(0, 0_R)$ and next to state $(0, 1)$ with probability $1 - p_1$. This means that from state 2, the possible transitions for $(Y_t)_{t \geq 0}$ are $2 \rightarrow 1$ and $2 \rightarrow 3$, with respective probabilities p_1 and $1 - p_1$. Moreover, the time spent in state 2 by $(Y_t)_{t \geq 0}$ is some r.v. with hazard rate $\mu_B(t)$. A similar reasoning is valid for state $(0, 1)$ and proves that the process $(Y_t)_{t \geq 0}$ is semi-Markovian, with the following semi-Markovian kernel:

σ -algebra $\sigma(X_u, u \leq T_n)$ generated by the past up to T_n ; (iii) at each T_n , the future process $(X_{t+T_n})_{t \geq 0}$ given the past up to T_n (i.e., given $\sigma(X_u, u \leq T_n)$) is identically distributed on $\{Y_n = i\}$ as $(X_t)_{t \geq 0}$ given $\{Y_0 = i\}$ (see **Markov Regenerative Processes** for more details). In other words, this means that the future of a Markov regenerative process after T_n only depends on Y_n and that Y_n only depends on the past of the process up to T_n . As in the case of a regenerative process, the evolution of the system between two jumps of $(Y_n)_{n \in \mathbb{N}}$ may be very general.

Setting $(q(i, j, dt))_{i, j \in F}$ to be the semi-Markov kernel associated to $(T_n, Y_n)_{n \in \mathbb{N}}$ and using the Markov property at time T_1 , the point availability of the system given that $Y_0 = i$ fulfills the following Markov renewal equation:

$$\begin{aligned} A_i(t) &= \mathbb{P}_i(X_t \in \mathcal{U}, T_1 > t) + \mathbb{P}_i(X_t \in \mathcal{U}, T_1 \leq t) \\ &= \mathbb{P}_i(X_t \in \mathcal{U}, T_1 > t) \\ &\quad + \sum_{j \in E} \int_{[0, t]} \mathbb{P}_i(X_t \in \mathcal{U} | T_1 = u, Y_1 = j) \\ &\quad \times q(i, j, du) \end{aligned}$$

$$\begin{aligned}
 &= \mathbb{P}_i (X_t \in \mathcal{U}, T_1 > t) \\
 &\quad + \sum_{j \in E} \int_{[0, t]} A_j (t - u) q (i, j, du) \\
 &= \mathbb{P}_i (X_t \in \mathcal{U}, T_1 > t) + (A * q) (i, t), \quad (14)
 \end{aligned}$$

where $(A * q) (i, t)$ is defined in Equation (6).

Let us now suppose that the Markov renewal process $(T_n, Y_n)_{n \in \mathbb{N}}$ is irreducible with nonarithmetic sojourn times and means $m(i) = \mathbb{E}_i (T_1)$, and let π be the unique stationary probability measure of the Markov chain $(Y_n)_{n \in \mathbb{N}}$. The asymptotic availability $A(\infty)$ then exists and is [5]

$$\begin{aligned}
 A(\infty) &= \frac{\sum_{i \in F} \pi(i) \int_0^{+\infty} \mathbb{P}_i (X_t \in \mathcal{U}, T_1 > t) dt}{\sum_{i \in F} \pi(i) m(i)} \\
 &= \frac{\sum_{i \in F} \pi(i) \mathbb{E}_i \left(\int_0^{T_1} \mathbf{1}_{\mathcal{U}} (X_t) dt \right)}{\mathbb{E}_{\pi} (T_1)} \quad (15)
 \end{aligned}$$

where symbols i and π in $\mathbb{P}_i$, $\mathbb{E}_i$, and $\mathbb{E}_{\pi}$ here refer to the initial distribution of $(Y_n)_{n \in \mathbb{N}}$.

This means that the asymptotic availability is the quotient of the system mean up time between arrivals of $(T_n, Y_n)_{n \in \mathbb{N}}$ divided by the mean interarrival length of $(T_n, Y_n)_{n \in \mathbb{N}}$, when $(T_n, Y_n)_{n \in \mathbb{N}}$ is in its steady state.

Back to the Example. We take $\bar{\lambda}_A = 1$, $\bar{\lambda}_B = 2$ as failure rates, and $\Gamma(3.5, 1/52.5)$,

for the repair distributions, as in the semi-Markovian case. We also take $\bar{\lambda}'_A = 0.5$ and $\bar{\lambda}'_B = 1$.

The Markov renewal Equations (14) may here be written as

$$\begin{aligned}
 A_1(t) &= e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} + \int_0^t A_2(t-u) q(1, 2, du) \\
 &\quad + \int_0^t A_3(t-u) q(1, 3, du) \\
 &= e^{-(\bar{\lambda}_A + \bar{\lambda}_B)t} + \int_0^t (\bar{\lambda}_B A_2(t-u) \\
 &\quad + \bar{\lambda}_A A_3(t-u)) e^{-(\bar{\lambda}_A + \bar{\lambda}_B)u} du, \\
 A_2(t) &= \int_0^t A_1(t-u) q(2, 1, du) \\
 &\quad + \int_0^t A_3(t-u) q(2, 3, du) \\
 &= \int_0^t (p_1 A_1(t-u) \\
 &\quad + (1-p_1) A_2(t-u)) f_{\mu_B}(u) du,
 \end{aligned}$$

with

$$\begin{aligned}
 p_1 &= \mathbb{P}(Z_{\mu_B} \leq Z_{\bar{\lambda}'_A}) = \int_0^{+\infty} f_{\mu_B}(t) e^{-\bar{\lambda}'_A t} dt \\
 &= \frac{1}{(1 + b_{R_B} \bar{\lambda}'_A)^{a_{R_B}}},
 \end{aligned}$$

and a similar equation for $A_2(t)$.

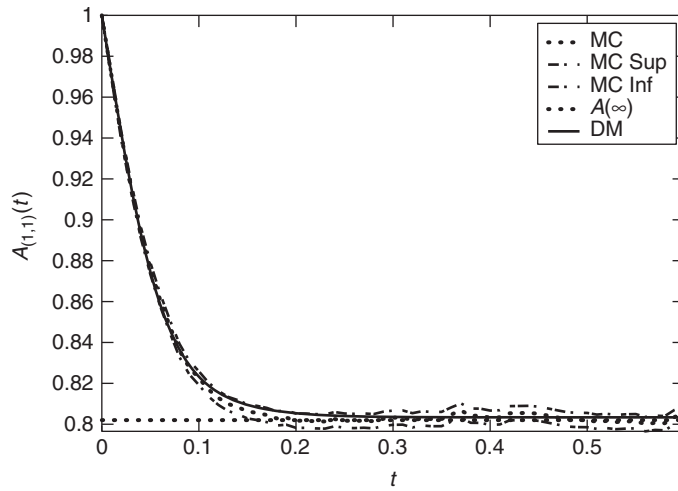


Figure 5. Point and asymptotic availabilities, Markov regenerative case.

This provides a set of three equations, which we discretize for their numerical resolution. Monte Carlo simulations are also performed. The results for $A_{(1,1)}(t)$ ($= A_1(t)$) are provided by both methods in Fig. 5, as well as the asymptotic availability, easily provided by Equation (15) with $A(\infty) \simeq 0.802$.

DISCUSSION AND FURTHER READING

In conclusion, we have presented here, classical stochastic models with (Markov) regenerative properties. For such models, the point availability has been proved to satisfy (Markov) renewal equations. The solving of such equations has, however, been seen to be generally impossible in full form and requires numerical procedures. We have here made the choice to use discretization techniques, which are easy to implement. Also, they can provide upper and lower bounds for the solutions. The results have been checked through Monte Carlo simulations, which are commonly used by practitioners in the reliability field, with generally longer computing times, however. Another numerical method might have been to use Laplace transforms, which are available in full form for the quantities of interest. The great progress made in their inversion in the last decade makes this method very appealing. Numerical difficulties may however arise in case of small or big arguments. Also, the precision of the results is not always available. See Ref. 17 for more details and references on the subject.

As for the asymptotic availability, it is usually simpler to compute than the point availability: it typically requires to evaluate the mean cycle duration and the mean up time on a cycle in the regenerative case, or similar quantities linked to the underlying Markov renewal process in the Markov regenerative case.

The classical models presented here actually sometimes (often?) appear as restrictive in the applications and even very simple systems may not meet with their assumptions: as an example, let us come back to

our two-unit series system, where one component is suspended when the other is down, with general repair and failure rates and only down components repaired. In that case, the system never forgets its past so that the system meets with none of the previous models. Monte Carlo simulations may however be performed, to compute both point and asymptotic availabilities. Another possibility is to use new models coming from dynamic reliability [18], which are presently arriving in “classical” reliability. Such models are called *piecewise deterministic Markov processes* [19] and have been proved to have a great modeling power [20]. Like the models presented here, their numerical assessment may be done by Monte Carlo simulations or by discretization methods [21,22].

Another drawback of the models presented here is that no aging is taken into account in the sense that at (Markov) regeneration points, the future evolution of the system given its present state is a stochastic replica of the past given the same starting state, without any evolution. Other possible models, which take aging into account are geometric processes, where up and down periods alternate with geometrically decreasing and increasing durations [23], nonhomogeneous Markov and semi-Markov processes with finite state spaces [24], or nonhomogeneous piecewise deterministic Markov processes [25].

REFERENCES

1. Csenki A. Stochastic models in reliability and maintenance. In: Osaki S, editor. Transient analysis of semi-markov reliability models - a tutorial review with emphasis on discrete-parameter approaches. Berlin: Springer; 2002. pp. 219–251.
2. Barlow RE, Proschan F. Mathematical theory of reliability. Volume 17, Classics in applied mathematics. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM); 1965. With contributions by Larry C. Hunter, 1996.
3. Høyland A, Rausand M. System reliability theory: models, statistical methods, and applications. Wiley series in probability and statistics. 2nd ed. Hoboken (NJ): Wiley-Interscience

- [John Wiley & Sons, Inc.]; 2004. ISBN 0-471-47133-X.
4. O'Connor PDT, Newton D, Bromley R. Practical reliability engineering. 4th ed. Chichester: John Wiley & Sons, Inc.; 2002.
5. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall Inc.; 1975.
6. Limnios N, Oprisan G. Semi-Markov processes and reliability. Statistics for industry and technology. Boston (MA): Birkhäuser Boston Inc.; 2001. ISBN 0-8176-4196-3.
7. Birolini A. Reliability engineering: theory and practice. Reliability and availability of repairable systems. 5th ed. Berlin Heidelberg: Springer; 2007. pp. 162–276.
8. Coccozza-Thivent C. Processus stochastiques et fiabilité des systèmes. Volume 28, Mathématiques & applications. Berlin: Springer; 1997. In French.
9. Iosifescu M, Limnios N, Oprisan G. Modèles stochastiques. Collection méthodes stochastiques appliquées. Paris: HERMES Science Publishing Ltd; 2007. In French.
10. Asmussen S. Applied probability and queues. Volume 51, Applications of mathematics. 2nd ed. New York: Springer; 2003.
11. Dohi T, Kaio N, Osaki S. Stochastic models in reliability and maintenance. In: Osaki S, editor. Renewal processes and their computational aspects. Berlin: Springer; 2002. pp. 1–30.
12. Mercier S. Discrete random bounds for general random variables and applications to reliability. *Eur J Oper Res* 2007;177(1): 378–405.
13. Ross SM. Stochastic processes. Wiley series in probability and statistics: probability and statistics. 2nd ed. New York: John Wiley & Sons Inc.; 1996. ISBN 0-471-12062-6.
14. Stewart WJ. Introduction to the numerical solution of Markov chains. Princeton (NJ): Princeton University Press; 1994.
15. Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev* 2003;45(1):3–49. (electronic).
16. Mercier S. Numerical bounds for semi-Markovian quantities and application to reliability. *Methodol Comput Appl Probab* 2008;10(2):179–198.
17. Abate J, Whitt W. A unified framework for numerically inverting Laplace transforms. *INFORMS J Comput* 2006;18(4): 408–421.
18. Devooght J. Dynamic reliability. *Adv Nucl Sci Technol* 1997;25,:215–278.
19. Davis MHA. Piecewise deterministic Markov processes: a general class of nondiffusion stochastic models. *J R Stat Soc [Ser B]* 1984;46(3):353–388.
20. Zhang H, Gonzales K, Dufour F, *et al.* Piecewise deterministic Markov processes and dynamic reliability. *J Risk Reliab* 2008;222(4):545–551.
21. Labeau P-E, Zio E. Procedures of Monte Carlo transport simulation for applications in system engineering. *Reliab Eng Syst Saf* 2002;77(12):217–228.
22. Coccozza-Thivent C, Eymard R, Mercier S. A finite-volume scheme for dynamic reliability models. *IMA J Numer Anal* 2006;26(3):446–471.
23. Lam Y. The geometric process and its applications. Hackensack (NJ): World Scientific Publishing Co. Pvt. Ltd.; 2007. ISBN 978-981-270-003-2; 981-270-003-X.
24. Janssen J, Manca R. Semi-Markov risk models for finance, insurance and reliability. New York: Springer; 2007. ISBN 978-0-387-70729-7; 0-387-70729-8.
25. Jacobsen M. Point process theory and applications. Marked point and piecewise deterministic processes: probability and its applications. Boston (MA): Birkhäuser Boston Inc.; 2006.

AVERAGE REWARD OF A GIVEN MDP POLICY

MATTHEW D. BAILEY
School of Management, Bucknell
University, Lewisburg,
Pennsylvania

A Markov decision process (MDP) is a general model for formulating problems that involve a sequence of decisions made under uncertainty to optimize a given performance criterion (e.g., the minimization of costs or the maximization of profits). The sequential nature of the problem allows for the defining of stages or decision epochs. This allows for the separation of the class of MDPs into finite- and infinite-horizon problems. In finite-horizon problems, rewards are received over a finite number of stages whereas infinite horizon problems allow for the indefinite accumulation of rewards. Following the notation of Puterman [1], let S be the defined state space of the MDP. For every state $s \in S$, let the set of feasible decisions or actions be A_s , where for every action $a \in A_s$ the decision maker receives reward $r(s, a)$, where $|r(s, a)| \leq M < \infty$. A transition from state s to state j when action $a \in A_s$ is chosen occurs with probability $p(j | s, a)$. Let a stationary policy $\mu = \{d, d, \dots\}$ be a sequence of identical decision rules, where a decision rule d is a function mapping actions to states such that $d(s) \in A_s$. Such policies are called *stationary Markov deterministic policies*. The application of such a policy induces a discrete-time Markov chain (DTMC) with rewards (a Markov reward process) where X_t is the state of the system at transition t and Y_t is the action chosen at state X_t , so that $Y_t = d(X_t)$ (also refer to **Definition and Examples of DTMCs** and **DTMCs with Costs and Rewards**). The objective of an average reward MDP is to find a policy that maximizes the average expected reward per stage defined as the

gain:

$$g^\mu(s) = \lim_{N \rightarrow \infty} \frac{1}{N} E_s^\mu \left\{ \sum_{t=1}^N r(X_t, Y_t) \right\}, \quad (1)$$

where $E_s^\mu \{\}$ is the expected value under policy μ given the initial state s .

In contrast to discounted reward MDPs, the underlying structure of the DTMC induced by a feasible policy has an impact on the optimal policy for an average reward MDP. As a result, for a given policy, we must classify the underlying DTMC to determine the average reward of the policy. We review the situation where the state space is finite; a discussion of the infinite-state case can be found in Ref. 2. The induced DTMC may be classified as either unichain (a single recurrent communicating class and a potentially empty set of transient states) or multichain (more than one communicating class). In general, a unichain MDP requires that all deterministic stationary policies induce a unichain DTMC. In the case of a finite-state unichain DTMC, it follows from standard DTMC results that all of the recurrent states must be positive recurrent, that is independent of the beginning state; all recurrent states will be visited infinitely often and the expected number of epochs between visits to recurrent states is finite (refer **Definition and Examples of DTMCs**). As a result, the average expected reward of such a system is independent of the beginning state (the gain is constant) and determined by the frequency of visits to each state. Therefore, it can be shown that for any state $i \in S$,

$$g^\mu(i) = \sum_{s \in S} \pi_{(s, d(s))} r(s, d(s)), \quad (2)$$

where $\pi_{(s, d(s))}$ is the long-run fraction of the transitions the DTMC spends in state s under decision rule d . From the theory of DTMCs, these values can be found as the

unique solution to

$$\begin{aligned}\pi_{(j,d(j))} &= \sum_{s \in S} \pi_{(s,d(s))} p(j | s, d(s)), \\ \sum_{s \in S} \pi_{(s,d(s))} &= 1, \\ \pi_{(s,d(s))} &\geq 0.\end{aligned}\quad (3)$$

As seen above, the gain of a policy is defined by the steady-state behavior of the system. In addition, the average rewards received before the system is in steady state can be of interest when comparing policies and utilizing more computationally efficient methods for determining the gain of a policy. To this end, we present the bias, h beginning in state s for a fixed policy to be

$$h^\mu(s) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N E_s^\mu \left\{ \sum_{t=1}^k [r(X_t) - g(X_t)] \right\}, \quad (4)$$

which for an aperiodic DTMC may be simplified to

$$h^\mu(s) = \sum_{k=1}^{\infty} E_s^\mu \{r(X_t) - g(X_t)\}. \quad (5)$$

By the definition of the gain above, when the system reaches steady state the contribution to the bias will be zero. However, while the system is transient the bias will record the average expected deviation from the gain. While two policies may have identical gain, they may differ considerably in bias due to the transient behavior and reward structure of the system (called *bias optimality* [3]). Similar to the gain, the bias for a policy can be determined from the limiting matrix P^* , where

$$P^* = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N P^{t-1}, \quad (6)$$

and P^{t-1} is the $(t-1)$ -step transition probability matrix of the underlying DTMC. The rows of the limiting matrix are identical and the components are given by the $\pi_{(s,d(s))}$ values found from the previous set of

equations. Given P^* , the vector of biases h , can be found by

$$h = (I - P + P^*)^{-1}(I - P^*)r, \quad (7)$$

where r is the vector of rewards for the policy μ and $(I - P + P^*)^{-1}(I - P^*)$ is called the *Drazin inverse* of $(I - P)$.

Example. Let $S = \{s_1, s_2\}$ with stationary policy μ comprise decision rule d , where

$$\begin{aligned}p(s_1 | s_1, d(s_1)) &= 0, \\ p(s_2 | s_1, d(s_1)) &= 1, \\ p(s_1 | s_2, d(s_2)) &= 1/2, \\ p(s_2 | s_2, d(s_2)) &= 1/2,\end{aligned}$$

and

$$r(s_1, d(s_1)) = -2 \quad r(s_2, d(s_2)) = 1.$$

By definition,

$$P = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}$$

and

$$r = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

It can be easily shown that

$$P^* = \begin{bmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{bmatrix}$$

and using Equation (2) we can compute

$$g = [1/3 \quad 2/3] \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Finally, from Equation (7)

$$h = \begin{bmatrix} 4/9 & -4/9 \\ -2/9 & 2/9 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -4/3 \\ 2/3 \end{bmatrix}.$$

However, the above methodology for determining the bias is typically not computationally efficient. When our underlying DTMC is

unichain, the average reward can be determined as the unique solution to

$$r - g + (P - I)h = 0, \quad (8)$$

where g is the unknown vector of gains (each component of the vector is identical by Equation 2). While the gain can be uniquely determined from Equation (8), the vector h is the bias vector if $P^*h = 0$.

In the instance where a fixed policy induces a multichain DTMC, we can use an approach similar to that given above. For each closed irreducible recurrent class, say $R_1, R_2, R_3, \dots, R_m$ the gain will be constant for any state in the same class. If we view each class as an individual unichain DTMC, we can determine the gain for that recurrent class as above. In the case of a state in a transient class T , the gain will be determined by which recurrent class the system eventually enters. Once the gain for the states in the recurrent classes have been determined and by conditioning on the first transition from the transient state, the gains for transient states can be determined by solving the set of equations

$$\begin{aligned} g(s) = & \sum_{k \in R_1 \cup R_2 \dots R_m} p(k | s, d(s))g(k) \\ & + \sum_{j \in T} p(j | s, d(s))g(j) \quad \forall s \in T. \end{aligned} \quad (9)$$

For a more in-depth discussion on determining the average reward for a given MDP policy, we refer the reader to Refs 1, 3, and 4.

REFERENCES

1. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley & Sons, Inc.; 1994.
2. Sennott LI. Average reward optimization theory for denumerable state spaces. In: Feinberg EA, Shwartz A, editors. Handbook of Markov decision processes: methods and applications. Norwell (MA): Kluwer Academic Press; 2003. pp. 153–172.
3. Lewis ME, Puterman ML. Bias optimality. In: Feinberg EA, Shwartz A, editors. Handbook of Markov decision processes: methods and applications. Norwell (MA): Kluwer Academic Press; 2003. pp. 89–111.
4. Heyman DP, Sobel MJ. Volume I, Stochastic models in operations research. Mineola (NY): Dover Publications, Inc.; 1982.

AVIATION SAFETY AND SECURITY

ARNOLD BARNETT
Sloan School of Management,
Operations Research Center
Massachusetts Institute of
Technology, Cambridge,
Massachusetts

Aviation safety and security is a very broad domain, which extends well beyond the contours of operations research and management science (OR/MS). But OR/MS can contribute in this area to two key respects: quantifying the risks that safety and security threats pose to air travelers, and suggesting which measures to reduce aviation risk are most likely to be beneficial. In this brief introductory essay, we offer some illustrations of these points. We restrict our focus on passenger air travel and, in the interests of brevity, limit the discussion to scheduled jet travel.

HOW SAFE?

Surveys suggest that, at one point or another, most air travelers are nervous about their safety in flight. But how dangerous is it to fly? To put the question more pointedly and specifically, what is the chance of being killed during an air journey? We begin with the risks posed by aviation accidents, and turn later to the threats posed by terrorists and other saboteurs.

Finding the appropriate metric for the mortality risk of air travel is not straightforward, as is suggested by considering some of the metrics now in common use. One statistic—used by the US National Transportation Safety Board, among others—is *fatal accidents per million flight hours*. Unfortunately, both the numerator and the denominator of the ratio are problematic. The term “fatal accident” obliterates the distinction between a crash that kills 1 person out of 200 on board and another that kills 200 out of 200. And the measure

gives no weight to safety improvements (e.g., fire-retardant materials) that reduce fatalities but do not prevent them. Moreover, the emphasis on hours flown misses the point that the heavy majority of fatal accidents occurs in the takeoff/climb or descent/landing phases of flight. Boeing reports that, over the period 1998–2007, only 10% of worldwide jet accidents that caused on-board fatalities occurred during the cruise phase of flight [1]. In consequence, the number of flights performed, rather than their durations, is a better guide to exposure to risk.

Another statistic that has long been used is *hull losses per 100,000 flight departures*. A hull loss is an accident that damages the plane beyond repair (the aerial equivalent of “totaling” a car.) In focusing on flight departures rather than on flight hours, this ratio recognizes that the distance or duration of the flight bears scant relationship to its accident risk. If, however, the aim is to gauge passenger death risk, then hull loss is a questionable proxy, because the passenger survival rate in hull losses ranges from 0% to 100%. There have been many instances in which a plane landed with major damage but, because of well-executed emergency procedures, all passengers were evacuated before the plane was engulfed in flames and became a hull loss. If such rescues are becoming more common, then a mere count of hull losses would fail to reflect that salient point.

Yet another statistic of interest is the ratio of *passengers killed to passengers carried*. This indicator has a plausible connection to the chance that a current passenger will be killed, for it literally reports what fraction of passengers were killed in a recent period. But weighting a crash by the number of fatalities can create difficulties. If a jetliner hits a mountain killing all passengers aboard, the implications about air safety are not twice as great if the plane is full rather than one-half full. And 22 deaths out of 22 aboard does not mean the same thing as 22 deaths out of 220 aboard (In the latter case, an excellent crew response may have saved 90% of

the passengers). In other words, risk statistics that use deaths in the numerator are vulnerable to meaningless fluctuations in the proportion of seats occupied, yet insensitive to large differences in the fraction of passengers who survived the crash.

Beyond the three indicators just discussed, there are several others that also seem questionable as guides to airplane mortality risk [2]. Applying operations research reasoning, this author has come up with an alternative measure that, while not perfect, may avoid the shortcomings just mentioned.

THE Q-STATISTIC

The Q-statistic is the answer to the question:

Suppose that a person chose a flight *completely at random* from among the set of interest (e.g., scheduled British domestic jet flights in the 1990s). What is the probability that she would not survive the flight?

The Q-statistic—which is essentially death risk per flight—assumes that there are N flights which can be indexed as $(1, 2, 3 \dots, N)$. We define x_i as the fraction of passengers on flight i who do not survive it (For example, if the flight lands safely, then $x_i = 0$; if it crashes and 20% of the passengers perish, then $x_i = 0.2$).

A formula for the Q-statistic arises if we note that each of the N flights has the same chance of $1/N$ of being selected at random and that, if flight i was selected, the traveler's conditional death risk is x_i . One way of dying because of the "flight lottery" would be to choose flight 1 at random and then to perish; the probability of this event is $(1/N)x_1$. The overall chance of dying would be the sum of the probabilities of all the N mutually exclusive ways that a fatal outcome can arise. In consequence, Q follows the rule:

$$Q = (1/N)x_1 + (1/N)x_2 + \dots (1/N)x_N \\ = \sum x_i \div N \quad (1)$$

Under this formula, the Q-statistic avoids several difficulties of the metrics previously discussed:

- The statistic weights each crash by the *fraction* of passengers killed. Thus,

a crash into a mountain that killed everyone on the plane would be treated the same way regardless of how many people happened to be on board that day. And a high survival rate in a crash would be treated very differently from a low survival rate.

- In accordance with empirical evidence, the calculation gives no weight of the mileage or duration of the flight.
- The calculation of Q is surprisingly easy. N is generally known from publicly available data. The conditional probability x_i is almost always zero and, when it is not, official reports about the crash make clear the value of x_i .
- The intuitive interpretation of Q is straightforward. While Q can only be calculated for a period that has ended, a recent Q-value is a strong approximation of the chance of dying on a flight today.

A disadvantage of the Q-statistic is that it does not reflect differences in risk among the flights within the set of interest. Flights in certain weather conditions or using certain types of aircraft might entail higher risk than others, meaning that death risk for a particular flight can differ from the overall Q-value (Of course, one could calculate Q-statistic by aircraft type if that seems desirable, and could at least approximate death risk per flight in adverse weather).

Some Calculated Q-Values

Table 1 below presents *accidental* death risk per flight for scheduled jet operations for various decades since 1960, which was essentially the start of the jet age (We will turn later to death risk caused by criminal and terrorist acts). The flights are divided into two categories based on the national origins of the airlines that performed them: *First-World* airlines are defined as those from the United States, Southern and Western Europe, the former British Commonwealth (Canada, Australia, and New Zealand), and Japan and Israel from Asia. All other flights are operated by *Developing-World* airlines.

Table 1. Accidental Death Risk per Flight, 1960–2008, for Different Decades and Two Groups of Jet Airlines

Period	Death Risk per Flight (Q-Statistic)
First-World airlines	
1960–1969	1 in 1 million
1970–1979	1 in 2 million
1980–1989	1 in 4 million
1990–1999	1 in 8 million
2000–2008	1 in 20 million
Developing-World airlines	
1960–1969	1 in 100,000
1970–1979	1 in 200,000
1980–1989	1 in 400,000
1990–1999	1 in 500,000
2000–2008	1 in 2 million

Notes: These statistics do not include crashes caused by criminal or terrorist acts. The calculations entail some approximations about the numbers of flights performed; [3,4] for discussions of the methodology and data sources used.

The key patterns in the data are obvious. Throughout the world and without any exceptions, jet travel has consistently become safer decade by decade. Passenger mortality risk fell by 95% between 1960–1969 and 2000–2008, at a surprisingly constant rate of about a factor of 2 per decade. It is also apparent that death risk is far lower on jet flights by First-World airlines than on those by Developing-World carriers. In every decade, the Q-statistic for the Developing-World was about a factor of 10 higher than the First-World statistic.

What does it mean intuitively to say that the accidental death risk per flight on a First-World jet carrier is now about 1 in 20 million? For some perspective, we might note that Barack Obama was elected the 44th president of the United States in 2009, 228 years after the first president was elected. That means the United States elects a new president on average every $(228/44) = 5.3$ years. Over 2000–2007, 4.1 million children were born in the United States per year. At that rate, $5.3 \times 4.1 \approx 22$ million Americans are born over 5.3 years, of whom one on average will be elected president. A randomly chosen American child thus has about a 1 in 22 million chance of being elected President

someday. One way of portraying aviation risk is to note that a child at a US airport has roughly the same chance of eventually becoming President as of perishing in an aviation accident on her next flight.¹

ARE SOME AIRLINES SAFER THAN OTHERS?

The tables might seem to suggest that, if flying between the First-World and the Developing-World, it is safer to travel on a First-World airline than a Developing-World one. But such reasoning could falter because of the statistical problem called the *ecological fallacy* [5]. To understand the fallacy, consider the three statements:

- John is better in mathematics than Bill.
- I have a problem in trigonometry.
- Therefore, I should contact John rather than Bill.

There is a logical problem in this sequence of statements. Perhaps, trigonometry is Bill's strong point, and John's weak point. In that case, I would do well to contact Bill. More generally, a statement that might be true in the aggregate does not have to apply in each specific instance. In the context of aviation, it could be fallacious to suggest that, because First and Developing-World airlines show a large aggregate difference in safety, that difference must apply on every subset of their routes.

To determine whether First-World carriers are safer than Developing-World airlines on routes on which the two groups compete, namely, between First-World and Developing-World cities, one should bypass overall statistics and compare fatal accident records on these specific routes. When Barnett and Wang [4] did so, they found that the Q-statistic was *the same* for both groups of carriers over 1987–1996: approximately 1 in 600,000. When Barnett [6] investigated the issue some years later, the result was the same: the Q-statistic was approximately

¹The risk goes up, however, if we consider criminal/terrorist acts, as we will in Table 2.

1 in 1.5 million on both First-World and Developing-World airlines.

In the United States, there have been suggestions that regional jet flights are less safe than larger jets operated by “mainline” carriers. But the data do not bear those suspicions out: there was one fatal accident with no survivors over 2000–2008 on a US regional jet, and the Q-value for these flights as a group was 1 in 20 million. That statistic is almost identical to the figure for the larger US jets, which suffered two fatal accidents with no survivors on roughly twice as many flights as the regional carriers.

More generally, available data suggest that when two airlines fly nonstop on the same route, very rarely is there an empirical basis for believing that one carrier is safer than the other.

OR/MS AVIATION RISK ANALYSIS

OR/MS can do more than simply document progress in reducing passenger mortality risk. It can help in the decision making that is essential to such progress. OR/MS methods have underlay data analyses and simulations that allow quantification of the extent to which new technologies and policies reduce risk to air travelers. The methods have also allowed estimates of the costs of new measures, and have thus facilitated the cost/benefit analyses that distinguish worthwhile innovations from others.

Aircraft fire safety is one domain in which OR/MS has contributed to major progress. Assessing the effectiveness of a new strategy to prevent, suppress, or slow the spread of fires is a complex endeavor. It involves mathematical modeling, but it especially requires devising, performing, and analyzing appropriate experiments. Such activities have long taken place at the Federal Aviation Administration’s (FAA) Technical Center in Atlantic City, as have other activities that allowed estimation of the costs (as opposed to the benefits) of implementing new strategies [7,8]. One consequence of such efforts is the greater use of fire-retardant materials in aircraft cabins. When Delta Air Lines Flight 1141 crashed on takeoff in Dallas in 1988,

14 people lost their lives. But it has been estimated that, absent changes in cabin materials that slowed the spread of fire, the death toll might have been four times as high.

OR/MS ideas also contributed to an emotional safety debate, on whether small children in planes should be required to sit in child-restraint seats rather than travel in the laps of their parents. These seats would offer greater safety in some emergencies, and would presumably force parents to buy tickets for these children rather than carry them for free. Data analyses indicated that the safety benefit of child-restraint seats was surprisingly small: if mandated, the seats would save the lives of an expected two children every 5 years [9]. And that calculation does not even consider a key point: Raising the cost of air travel with small children would lead some parents to drive rather than fly. That substitution would not only reduce the “lives saved” estimate under the mandate, but would also induce additional child deaths in road travel. Calculations showed that, because travel by car is less safe than travel by air, the net effect of the diversion would be to *increase* child deaths, and thus yield less safety rather than more. In 2005, FAA declined to require child-restraint seats on airplanes on these grounds.

A continuing danger in aviation is the threat of collisions, both on the ground and in the air. OR/MS has long been involved in risk analysis in this area [10]. While hazards like fires on airplanes might largely be things of the past, increases in air traffic raise the fear that collision risk will worsen in the future.

The dangers of collisions start on the ground: the worst aviation accident in history was a 1977 runway crash that killed 583 people. Simple probability models suggest that runway collision risk should vary not with the level of traffic at an airport, but rather with the *square* of the level. Thus, a doubling of traffic would cause a quadrupling of risk. Nor are these models alarmist: analyses of historical data about runway collisions and harrowing near-misses have empirically validated the quadratic model. An OR/MS study using the quadratic model and other

factors yielded a grim mid-range forecast: absent changes in technology or procedures, US airports could be expected to suffer 15 fatal runway collisions over 2003–2022, with a death toll of 700–800 [11].

Midair collisions involving scheduled passenger flights have all but disappeared from First-World skies. The last such event occurred in 1988, more than 200 million flights ago.² But the control arrangements in place now are likely to be changed soon. In Western Europe, there is strong pressure to replace the various national air traffic control systems with a harmonized one (a “single European sky”). In the United States, current airline itineraries—under which planes are confined to a network of prescribed flight paths—are slated to give way gradually to a set of “direct routings” from origin to destination. Such routings would lead to shorter flight times and lesser fuel consumption.

Such changes present safety challenges. Harmonizing dozens of air traffic systems is unlikely to be a straightforward process. And direct routings would certainly complicate any visual display of US flight paths. The moving dots that represent planes on controllers’ screens—which line up today like points on a grid—could in the future resemble gas molecules in random scatter. Moreover, a fundamental notion of industrial sociology is the “learning curve,” under which new procedures beget errors and difficulties that had not been anticipated. Thus, apart from specifics, any major changes in air traffic control in Europe and the United States could pose risk.

However, OR/MS models can sometimes reduce future risk simply by demonstrating that the risk is substantial. Prompted in part by forecasts about runway collisions, the US Federal Aviation Administration has undertaken a major program of technological innovation at the nation’s larger airports, introducing new Airport Surface Detection Equipment-Model X (ASDE-X)

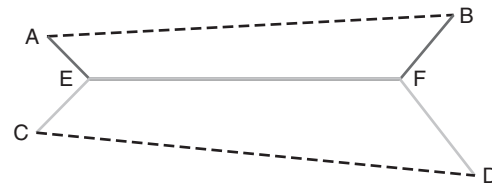


Figure 1. Under direct routings, planes traveling from A to B and from C to D should get nowhere close to each other.

radar systems and runway status lights. Had they been in place, these technologies would have prevented many past runway collisions. The forecast of 700–800 deaths over 2003–2022 might turn out to be a “self-destroying prophecy,” which stimulated an effective response by showing that the price of inaction was intolerable.

As for the hazards posed by direct routings, OR/MS models and data analysis help clarify the situation. The models highlight the important point that while direct routings can increase midair-collision risk, they can also do some things that should reduce risk. Such routings would change the geometry of flight paths in ways that, in themselves, could make collisions less likely. For example, consider Fig. 1, which concerns one plane traveling from A to B and another from C to D. Under the present prescribed routes, the first plane might follow path A-E-F-B and the second, path C-E-F-D. These planes could therefore come in close proximity along their common segment EF. If each could take a direct routing, they would get nowhere near one another.

Moreover, there are reasons to believe that direct routings would reduce the angles at which flight paths would cross at a given altitude [12]. Were such angles to drop, the characteristics of some collision warning systems would mean that pilots might get more time to react to alerts about an impending collision. This extra time should increase the likelihood of avoiding the midair crash.

To be sure, it is not clear that these advantages of direct routings are more consequential than the potential problems. But mathematical models allow more precise thinking about the benefits and drawbacks

²There was, however, a midair collision over Germany in 2002 involving a DHL cargo plane and a Russian charter flight.

of the policy transition, and increase the chance that it will be implemented in a way that minimizes risk.

AVIATION SECURITY

How Great a Threat?

No reader of this essay needs a reminder about 9/11, the worst day ever in the history of commercial aviation. On that day, four hijacked planes crashed, two of them destroying the World Trade Center in New York and a third causing grievous harm at the Pentagon in Washington. In the aftermath of that catastrophe, aviation security changed in ways that have made the entire experience of flying more burdensome. Not everyone is convinced that the added effort has been well spent.

There are those who believe that aviation is no longer at any special risk. Terrorists who wish to “top” the 9/11 achievements could turn to biological, chemical, or nuclear attacks, or to strategies that cause convulsions in cyberspace or the economy. Even if terrorists use conventional weapons, why should they not turn from airplanes to subways, shopping malls, or seaports, which are essentially unprotected? The government’s National Commission on Terrorist Attacks on the United States [[13], the 9/11 Commission, at p. 351] warned that a preoccupation with aviation security because of the 9/11 events could be “fighting the last war.”

Perhaps this viewpoint makes sense, but [14] estimated as follows:

- More US civilians were killed by terrorists during air journeys than during any other form of activity.
- On a per-hour basis, the terrorism death risk was 600 times higher during air journeys than at other times.

And these statistics were based on *the 35-year period prior to 9/11!* During that time, terrorists killed Americans traveling by air on two dozen separate occasions, which included numerous attacks by grenade or bomb (the most famous being the destruction

of Pan Am 103 as it traveled from London to New York).

Nor have attempts to harm air travelers been absent since 9/11. In late 2001, the Shoe Bomber tried to destroy a transatlantic jet. A 2002 shoot-out at Los Angeles International Airport left several dead and injured. A deliberate fire felled a Chinese jet in 2002, while suicide bombers simultaneously destroyed two Russian jets in 2004. In 2006, a plot was uncovered to destroy approximately 10 transatlantic jets with liquid explosives. The year 2007 saw a firebombing at Glasgow airport, while 2008 saw another attempt to crash a Chinese jet plane. President Bush stated in 2007 that a post-9/11 terrorist plot was foiled which would have crashed a passenger jet into the tallest building on the US West Coast.

In short, it seems more plausible to view 9/11 as part of a continuing fascination with aviation than as an isolated, aberrant act. However, recounting lists of terrible events offers less systematic risk analysis than did earlier calculations about passenger death risk tied to accidents. Table 2 strives for greater balance by presenting Q-statistics associated with criminal and terrorist acts against jet aircraft.

We see from the parenthetical percentages in Table 2 that, in all decades and for both groupings of jet flights, deliberate acts against aviation have contributed less to passenger mortality risk than have accidents. We also see in the table some tendency of the absolute passenger risks from terrorism to decrease over time, though less monotonically and dramatically than risks of accidental death.

Yet there is something colossally missing in Table 2: the recognition that the victims of aerial terrorism can go far beyond air travelers. On September 11 2001, there were nine times as many people killed on the ground as on the four planes that were hijacked. There were as many deaths on 9/11 as in all First-World jet accidents *combined* over 1985–2008. The last thing we should do is cite Table 2 as evidence that security is a secondary threat compared to accidents.

Indeed, the terrorist threat to aviation is very different in nature from that tied to

Table 2. Death Risk per Scheduled Jet Flight Caused by Criminal/Terrorist Acts, by Decade 1960–2008

Period	First-World Death Risk per Flight		Developing-World Death Risk per Flight	
1960–1969	1 in 15 million	(7%)	0	(0%)
1970–1979	1 in 25 million (8%)	1 in 1 million	(19%)	
1980–1989	1 in 25 million	(15%)	1 in 3 million	(12%)
1990–1999	1 in 10 billion	(0%)	1 in 4 million	(11%)
2000–2008	1 in 30 million	(70%)	1 in 12 million	(16%)

Notes: Numbers in parenthesis express the preceding Q-value as a *percentage* of the corresponding Q-value for accidents in Table 1. Death risk refers to air travelers only.

accidents. Once an aerial hazard like wind shear has been dealt with, it is essentially gone for good. Terrorists, by contrast, can forever change their methods of attack. The relationship between the historical record and future dangers is much less clear for deliberate attacks on aviation than for accidents. It is certainly conceivable that, in years to come, terrorism will pose a greater threat to aviation than all other dangers combined.

What Should We Do?

It is widely recognized that we cannot pay any price, however enormous, to achieve any increase in security, however minuscule. But this truism does not provide guidance about how finite resources to prevent aerial terrorism might best be allocated. Because OR/MS has a rich tradition in large-scale optimization, it might seem ideally suited for devising an overall antiterrorist strategy. But the great difficulties of specifying costs, constraints, and benefits make the analysis of an aviation security system an inherently speculative venture [see [15] for a discussion about airport security screening].

Instead of trying to develop a sound aviation security policy in one fell swoop, we might do best to develop a policy step-by-step, considering one measure at a time, and assessing whether it might pass a cost-benefit test. Such a sequential analysis is not guaranteed to yield an “optimal” policy in any rigorous sense, but it might represent the best we can reasonably do. In the next section, we apply this sequential approach in two specific situations, reaching conclusions about what measures should (or even could)

be part of effort to protect aviation from terrorists.

No Laptops?

On September 10, 2006, *The New York Times* published an editorial titled “A Ban on Carry-On Luggage.” It noted that “laptops, digital cameras, mobile phones and other electronic devices” could be used to trigger a bomb, and urged that they be prohibited from the passenger cabins of commercial aircrafts. *The Times* acknowledged that “separating people from their laptops during flights would be painful,” but added that “some people could surely use the time to go over reading material, or even revert to pen and paper.” While expressing hope that the ban would eventually give way to technological advances, *The Times* concluded that “for now, the surest way to keep dangerous materials out of the cabin is to keep virtually all materials out of the cabin.”

The Times cannot be faulted for lack of audacity, or for the relentlessness of its logic. The main purpose of air transportation, after all, is to get people to their destinations swiftly and safely; providing a congenial environment en route is a desirable but secondary goal. If carry-on objects that could destroy the plane cannot be reliably detected, then maybe the objects should simply not be carried on.

Obviously, however, a ban on laptops would be highly disruptive for some travelers. The time not spent working on the airplane might result in less time at home with the children, or in leisure or civic activities. And it might be facetious to suggest that modern business activity can be accomplished with “pen and paper.”

How might the discussion extend beyond a shouting match? We can make progress if we are willing to accept two statistics at the outset:

- The RAND Corporation has estimated that another successful terrorist attack against US aviation would cost about \$15 billion to the nation [16].
- The US Congress has estimated that the monetized cost of each minute of air travel delay is 63 cents per passenger [17].

We might note that a typical First-World jet flight is about 1.5 h long [1], during roughly 1.25 h of which the use of laptops is now permitted. It seems clear that the total inconvenience to passengers caused by a laptop ban would be greater than that caused by a 1-min arrival delay. It also seems clear that a 4-h arrival delay would upset passengers more than 1.25 h without laptops. But then there must be some intermediate delay-equivalent X such that an average passenger would be indifferent between a delay of X min and a ban on laptops. Given that the cost of an X -min delay is $\$0.63 X$, we can infer a similar dollar cost for a laptop ban.

Preliminary estimates are that $X \approx 15$ min for US travelers (That figure takes account of the fact that, for those who would not have brought laptops on board, the cost of banning them is zero). That works out to a cost of $15 \times 0.63 = \$9.30$ per traveler which, when multiplied by 650 million US travelers per year, implies an annual cost of banning laptops of about \$6 billion per year.

Comparing \$6 billion with \$15 billion (the cost of successful terrorism) implies that the ban would be “cost-effective” if it prevented one terrorist act *every 2.5 years* (15/6). To put it another way, the ban would be cost-effective if the probability is at least 40% per year that, absent that ban, terrorists would succeed in destroying an aircraft by means that the ban would have prevented. We cannot calculate the relevant probability precisely, but the analysis reduces to the question: is it 40% or greater?

Both historical data and most expert opinion would answer this last question firmly in

the negative. In consequence, a ban on laptops might be so costly that it could be better to bear the risks of not imposing it than to accept the widespread disenchantment it would cause.

Out of Time?

One of the hallmarks of modern terrorism is that an initial act of violence often is followed almost immediately by other acts. 9/11 clearly fits that pattern, as do more recent attacks on transportation systems in Spain, Britain, Russia, and India. Other multiple near-simultaneous attacks have occurred in recent years in Japan, Israel, Morocco, Indonesia, Kenya, Saudi Arabia, Iraq, the Philippines, Jordan, and Egypt. Generally, all the attacks within a given series were of the same form (e.g., time-actuated bombings).

Suppose, therefore, that an airplane is suddenly destroyed by a terrorist bomb. Then there is every reason to fear that other planes aloft are in imminent danger of exploding. But what if anything can be done to save these planes? The situation is desperate: Because time is short, we might assume that any further bombs are highly unlikely to be located, let alone defused.

The first question to consider is: how much time would elapse between the explosion and a warning to the pilots of other endangered planes? The fall to the earth after an explosion takes time: Korean #007 hit the sea 10 min after it was shot down, and Pan Am #103 crashed into Lockerbie 7 min after it exploded. Additional time would pass between the plane’s hitting the ground and the confirmation of a crash, and further time would be required to notify aviation authorities, who could then contact air traffic controllers who in turn would notify pilots. In short, *10 min* is an optimistic estimate of the time between the first airborne explosion and any warning to other aircraft.

But would the endangered planes actually have 10 min? The answer would depend on how simultaneous the bombings in the series would be. In this connection, it is useful to

consider the *major bombings on air/rail systems since 9/11*, all of which claimed many lives:

Year	Place	Number/Timing of Bombings
2003	Russia	1
2004	Madrid	10 in 3 min
2004	Moscow	1
2004	Russia	2 in 2 min
2005	London	3 in 1 min
2006	Mumbai	7 in 11 min

The chart shows that, in four of the six events, there was one or more follow-up bombings. The total number of such subsequent blasts was 18. With a 10-min time lag, only one subsequent bombing—the last one in Mumbai—would not yet have occurred. Moreover, any sudden measures to depressurize an aircraft—probably the only feasible way to reduce the danger posed by a bomb explosion—are highly dangerous in their own right (Loss of cabin pressure has caused nearly 200 deaths in two jet crashes since 2005). Many lives could be lost in the attempts to mitigate a potential explosion, a good fraction of which would probably occur in response to false alarms.

Under these circumstances, we reach an agonizing conclusion: after an initial on-board explosion, the least dangerous response might be to *do nothing*. Most other planes in immediate danger could be beyond help, and rescue efforts could well cost more lives than they save. As optimization researchers know, even the “optimal” strategy against a terrible set of constraints can yield an extremely bad result.

In both of these instances, approximate cost-benefit analysis argued against the security measure under discussion. That situation does not always prevail: this author has argued elsewhere that positive-passenger bag match—under which no luggage would travel on an aircraft baggage compartment if the passenger who checked it did not show up—would be cost effective if it prevented one luggage bombing every 150 years [18]. The broader point is that, when applied on a case-by-case basis,

even simple cost-benefit analysis offers some insight into whether a particular security measure should be adapted into the overall security strategy. If the analysis seems a bit crude, perhaps nothing more elaborate can be defended when the background data are limited and imprecise.

FINAL REMARKS

In terms of preventing accidents, the world’s airlines have made enormous progress throughout the jet age. Indeed, in First-World nations, fatal air accidents are at the brink of extinction, a remarkable circumstance when we consider that, in every jet flight, there are hundreds of things that can go wrong that would kill everyone on board. Such progress reflects well on the airlines, aircraft manufacturers, government regulators, press, and flying publics. OR/MS plays a helpful role both in documenting improvements in safety and in facilitating the many technical analyses that are fundamental to such improvements.

Security threats are inherently more elusive than natural hazards as a menace to air travelers, and OR/MS methods can only reduce the uncertainties to a limited extent. But OR/MS models, often very simple, make it easier to think logically about devising a security strategy, and to distinguish between absurd security measures and others that are highly reasonable. The models work sharply against the paralyzing notion that, unless we can figure out everything about security policy, we can figure out nothing. They better enable us to harness perhaps the strongest weapons that we have against terrorists: our brains.

Acknowledgment

The author thanks the reviewers and editors for their thoughtful and constructive suggestions.

REFERENCES

1. The Boeing Company. Statistical summary of commercial jet airplane accidents 1959–2007. 2008. Available at www.boeing.com/news/techissues; <http://www.docstoc.com/>

- docs/2290522/Statistical-Summary-of-Commercial-Jet-Airplane-Accidents. Accessed 2008.
2. Barnett A. Measure for measure: an analysis of aviation-safety metrics. *AeroSaf World* (Flight Safety Foundation). 2007;2(11):48–52.
 3. Barnett A, Higgins MK. Airline safety: the last decade. *Manage Sci* 1989;35(1):1–21.
 4. Barnett A, Wang A. Passenger mortality-risk estimates provide perspective on aviation safety. *Flight Saf Dig* 2000;27:1–12.
 5. Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 1950;15:351–357.
 6. Barnett A. World airline safety: the century so far. *Flight Saf Dig* 2006;33:14–19.
 7. Sarkos C. Heat exposure and burning behavior of cabin materials during an aircraft post-crash fire. Improved fire and smoke resistant materials for commercial aircraft interiors: a proceedings. Washington (DC): National Academies Press; 1995. pp. 25–36.
 8. Hill RG, Sarkos CP, Marker T. Development of a benefit analysis for an onboard aircraft cabin spray water system. In: Hirschler M, editor. Fire hazard and fire risk assessment. ASTM STP 1150. Philadelphia (PA): American Society for Testing and Materials; 1992. pp.116–127.
 9. Newman TB, Johnston BD, Grossman DC. Effects and costs of requiring child-restraint systems for young children traveling on commercial airplanes. *Arch Pediatr Adolesc Med* 2003;157(10):969–974.
 10. Machol R. An Aircraft Collision Model Management Science 1975;21(10):1089–1101.
 11. Barnett A, Paull G, Iadeluca J. Fatal US runway collisions over the next two decades. *Air Traffic Control Q* 2000;8(4):253–276.
 12. Barnett A. Free-flight and en route air safety: a first-order analysis. *Oper Res* 2000;48(6):833–845.
 13. National Commission on Terrorist Attacks on the United States. The 9/11 commission report. US Government Printing Office; 2004.
 14. Martonosi S, Barnett A. Terror is in the air. *Chance* 2004;17(2):25–27.
 15. Martonosi S, Barnett A. How effective is security screening of airline passengers? *Interfaces* 2006;36(6):545–552.
 16. Chow J, Chisea P, Dreyer M, *et al.* Protecting commercial aviation against the shoulder-fired missile threat, RAND Corporation. Available at www.rand.org/pubs/occasional_papers/2005/RAND_OP106.pdf. 2005.
 17. Joint Economic Committee of US Congress. Your flight has been delayed again. Available at <http://www/jec.senate.gov> 2008 May 22, p.4.
 18. Barnett A. Is it really safe to fly? *Tutorials Oper Res* 2008;5:17–30. a chapter in (INFORMS), Chapter 2.

AXIOMATIC MEASURES OF RISK AND RISK-VALUE MODELS

JIANMIN JIA

Department of Marketing, Faculty
of Business Administration, The
Chinese University of Hong
Kong, Shatin, Hong Kong

JAMES S. DYER

Department of Information, Risk,
and Operations Management,
The McCombs School of
Business, University of Texas at
Austin, Austin, Texas

JOHN C. BUTLER

Department of Finance, The
McCombs School of Business,
University of Texas at Austin,
Austin, Texas

INTRODUCTION

This article provides a review of measures of risk and risk-value models that have been developed over the past 10 years to provide a new class of decision making models based on the idea of risk-value trade-offs. The measurement of risk has been a critical issue in decision sciences, finance, and other fields for many years. We focus on a preference-dependent measure of risk that can be used to derive risk-value models within both an expected utility framework and a non-expected utility framework. Although this measure of risk has some descriptive power for risk judgments, it is more normative in nature. We treat the issue of measures of perceived risk in a separate article in this collection (see *Axiomatic Models of Perceived Risk*).

Intuitively, individuals may consider their choices over risky alternatives by trading off between risk and return, where return is typically measured as the mean (or expected return) and risk is measured by some indicator of dispersion or possible losses. Markowitz [1–3] proposed variance as a

measure of risk, and a mean–variance model for portfolio selection based on minimizing variance subject to a given level of mean return. But arguments have been made that mean–variance models are appropriate only if the investor’s utility function is quadratic or the joint distribution of returns is normal. However, these conditions are rarely satisfied in practice. Markowitz also suggested semivariance as an alternative measure of risk. Some other measures of risk, such as lower partial moment risk measures and absolute standard deviation, have also been proposed in the financial literature [4]. However, without a common framework for risk models it is difficult to justify and evaluate these different measures of risk as components of a decision making process.

Expected utility theory is generally regarded as the foundation of mean-risk models and risk-return models [5–9]. However, expected utility theory has been called into question by empirical studies of risky choice [10–14]. This suggests that an alternative approach regarding the paradigm of risk-return trade-offs would be useful for predicting and describing observed preferences.

In the main stream of decision research, the role of risk in determining preference is usually considered implicitly. For instance, in the expected utility model [15], an individual’s attitude toward the risk involved in choices among risky alternatives is defined by the shape of his or her utility function [16,17]; and in some non-expected utility models, risk (or “additional” risk) is also captured by some nonlinear function over probabilities [12,18–20]. Thus, these decision theories are not, at least explicitly, compatible with the choice behavior based on the intuitive idea of risk-return trade-offs as often observed in practice. Therefore, they offer little guidance for this type of decision making.

In this article, we review our risk-value studies and provide an axiomatic measure of risk that is compatible with choice behavior

based on risk-value trade-offs. In particular, this framework unifies two streams of research: one in modeling risk judgments and the other in developing preference models. This synthesis provides risk-value models that are more descriptively powerful than other preference models and risk models that have been developed separately.

The remainder of this article is organized as follows: the next section describes a preference-dependent measure of risk with several useful examples. The section titled “Frameworks for Risk-Value Trade-Off” reviews the basic framework of our risk-value studies and related preference conditions. The section titled “Generalized Risk-Value Models” presents three specific forms of risk-value models. The concluding section summarizes the applicability of risk-value studies and discusses topics for future research.

STANDARD MEASURE OF RISK

As a first step in developing risk-value models, Jia and Dyer [21] propose a preference-dependent measure of risk, called a *standard measure of risk*. This general measure of risk is based on the converse expected utility of normalized lotteries with zero-expected values, so it is compatible with the measure of expected utility and provides the basis for linking risk with preference.

For lotteries with zero-expected values, we assume that the only choice attribute of relevance is risk. A riskier lottery would be less preferable and vice versa, for any risk-averse decision maker. Therefore, the riskiness ordering of these lotteries should simply be the reverse of the preference ordering. However, if a lottery has a nonzero mean, then we assume that the risk of that lottery should be evaluated relative to a “target” or reference level. The expected value of the lottery is a natural reference point for measuring the risk of a lottery.

Therefore, we consider decomposing a lottery X (i.e., a random variable) into its mean $\bar{X}$ and its standard risk, $X' = X - \bar{X}$, and the standard measure of risk is defined as

follows:

$$R(X') = -E[u(X')] = -E[u(X - \bar{X})], \quad (1)$$

where $u(\cdot)$ is a utility function [15] and E represents expectation over the probability distribution of a lottery.

One of the characteristics of this standard measure of risk is that it depends on an individual’s utility function. Once the form of the utility function is determined, we can derive the associated standard measure of risk over lotteries with zero means. More importantly, this standard measure of risk can offer a preference justification for some commonly used measures of risk so that the suitability of those risk measures can be evaluated.

If a utility function is quadratic, $u(x) = ax - bx^2$, where $a, b > 0$, then the standard measure of risk is characterized by the variance, $R(X') = bE[(X - \bar{X})^2]$. However, the quadratic utility function has a disturbing property; that is, it will be decreasing in x after a certain point and it exhibits increasing risk aversion. Since the quadratic utility function may not be an appropriate description of preference, it follows that variance may not be a good measure of risk (unless the distribution of a lottery is normal).

To obtain a related, but increasing utility function, consider a third-order polynomial (or cubic) utility model, $u(x) = ax - bx^2 + c'x^3$, where $a, b, c' > 0$. When $b^2 < 3ac'$, the cubic utility model is increasing. This utility function is concave, and hence risk averse for low outcome levels (i.e., $x < b/(3c')$), and convex, and thus risk seeking for high outcome values (i.e., $x > b/(3c')$). Such a utility function may be used to model a preference structure consistent with the observation that a large number of individuals purchase both insurance (a moderate outcome—small probability event) and lottery tickets (a small chance of a large outcome) in the traditional expected utility framework [22]. The associated standard measure of risk for this utility function can be obtained as follows:

$$R(X') = E[(X - \bar{X})^2] - cE[(X - \bar{X})^3], \quad (2)$$

where $c = c'/b > 0$. Model 2 provides a simple way to combine skewness with variance into a measure of risk. This measure of risk should be superior to variance alone since the utility function implied by Equation (2) has a more intuitive appeal than the quadratic one implied by variance. Further, since Equation 2 is not consistent with increasing risk aversion, it is more appropriate for prescriptive purposes than variance.

Markowitz [23] noted that an individual with the utility function that is concave for low outcome levels and convex for high outcome values will tend to prefer positively skewed distributions (with large right tails) over negatively skewed ones (with large left tails). The standard measure of risk (Eq. 2) clearly reflects this observation; that is, a positive skewness will reduce risk and a negative skewness will increase risk.

If an individual's preference can be modeled by an exponential or a linear plus exponential utility function, $u(x) = ax - be^{-cx}$, where $a \geq 0$, and $b, c > 0$, then its corresponding standard measure of risk (with the normalization condition $R(0) = 1$) is

$$R(X')E[e^{-c(X-\bar{X})} - 1]. \quad (3)$$

Bell [7] identified $E[e^{-c(X-\bar{X})}]$ as a measure of risk from the linear plus exponential utility model by arguing that the riskiness of a lottery should be independent of its expected value. Weber [24] also modified Sarin's [25] expected exponential risk model by requiring that the risk measure be location free.

If an individual is risk averse for gains but risk seeking for losses as suggested by Prospect Theory [12,26], then we can consider a piece-wise power utility model as follows:

$$u(x) = \begin{cases} ex^{\theta_1}, & \text{when } x \geq 0, \\ -d|x|^{\theta_2}, & \text{when } x < 0, \end{cases} \quad (4)$$

where e, d, θ_1 , and θ_2 are nonnegative constants. Applying Equation (1), the corresponding standard measure of risk is

$$R(X') = dE^{-}[|X - \bar{X}|^{\theta_2}] - eE^{+}[|X - \bar{X}|^{\theta_1}], \quad (5)$$

where $E^{-}[|X - \bar{X}|^{\theta_2}] = \int_{-\infty}^{\bar{X}} |x - \bar{X}|^{\theta_2} f(x) dx$, $E^{+}[|X - \bar{X}|^{\theta_1}] = \int_{\bar{X}}^{\infty} (x - \bar{X})^{\theta_1} f(x) dx$, and $f(x)$ is the probability density of a lottery.

The standard measure of risk (Eq. 5) includes several commonly used measures of risk in the financial literature as special cases. When $d > e > 0, \theta_1 = \theta_2 = \theta > 0$, and the distribution of a lottery is symmetric, then we can have $R(X') = (d - e)E[|X - \bar{X}|^{\theta}]$, which is associated with variance and absolute standard deviation if $\theta = 2$ and $\theta = 1$, respectively. This standard measure of risk is also related to the difference between the parameters d and e , which reflects the relative effect of loss and gain on risk. In general, if the distribution of a lottery is not symmetric, this standard measure of risk will not be consistent with the variance of the lottery even if $\theta_1 = \theta_2 = 2$, but it is still related to the absolute standard deviation if $\theta_1 = \theta_2 = 1$ [27].

Konno and Yamazaki [28] have argued that the absolute standard deviation is more appropriate for use in portfolio decision making than the variance, primarily due to its computational advantages. Dennenberg [29] argues that the average absolute deviation is a better statistic for determining the safety loading (premium minus the expected value) for insurance premiums rather than the standard deviation. These arguments suggest that the absolute standard deviation may be a more suitable measure of risk than variance in some applied contexts.

Another extreme case of Equation (5) arises when $e = 0$ (i.e., the utility function is nonincreasing for gains); then the standard measure of risk is a lower partial moment risk model $R(X') = dE^{-}[|X - \bar{X}|^{\theta_2}]$. When $\theta_2 = 2$, it becomes a semivariance measure of risk [1]; and when $\theta_2 = 0$, it reduces to the probability of loss.

In summary, some other proposed measures of risk are special cases of this standard measure of risk. The standard measure of risk is more normative in nature, as it is independent of the expected value of a lottery. To obtain more descriptive power and to capture perceptions of risk, we have also established measures of perceived risk that are based on a two-attribute structure: the mean of a lottery and its standard risk [30] as described

in a separate article in this encyclopedia (see *Axiomatic Models of Perceived Risk*).

FRAMEWORKS FOR RISK-VALUE TRADE-OFF

When we decompose a lottery into its mean and standard risk, then the evaluation of the lottery can be based on the trade-off between mean and risk. We assume a risk-value preference function $f(\bar{X}, R(X'))$, where f is increasing in $\bar{X}$ and decreasing in $R(X')$ if one is risk averse.

Consider an investor who wants to maximize his or her preference function f for an investment and also requires a certain level μ of expected return. Since f is decreasing in $R(X')$ and $\bar{X} = \mu$ is a constant, then maximizing $f(\bar{X}, R(X'))$ is equivalent to minimizing $R(X')$; that is, $\max \{f(\bar{X}, R(X')) | \bar{X} = \mu\} \Rightarrow \min \{R(X') | \bar{X} = \mu\}$. This conditional optimization model includes many financial optimization models as special cases by choosing different standard measures of risk; that is, Markowitz's mean-variance model, the mean-absolute standard deviation model, and the mean-semivariance model. Some new optimization models can also be formulated based on our standard measures of risk (Eqs 2 and 5).

In the conditional optimization problem, we do not need to assume an explicit form for the preference function f . The problem only depends on the standard measure of risk. However, we may argue that an investor should maximize his or her preference functions unconditionally in order to obtain the overall optimal portfolio. For an unconditional optimization decision, the investor's preference function must be specified. Here, we consider two cases for the preference function f : (i) when it is consistent with the expected utility theory and (ii) when it is based on a two-attribute expected utility foundation.

Let $\mathbf{P}$ be a convex set of all simple probability or lotteries $\{X, Y, \dots\}$ on a nonempty set $\mathbf{X}$ of outcomes, and Re be the set of real numbers (assuming $\mathbf{X} \in \text{Re}$ is finite). We define $\succ$ as a binary preference relation on $\mathbf{P}$.

Definition 1. For two lotteries $X, Y \in \mathbf{P}$ with $E(X) = E(Y)$, if $w_0 + X \succ w_0 + Y$ for some $w_0 \in \text{Re}$, then $w + X \succ w + Y$ for all $w \in \text{Re}$.

This is called the *risk-independence condition*. It means that for a pair of lotteries with a common mean, the preference order between the two lotteries will not change when the common mean changes; that is, preference between the pair of lotteries can be determined solely by the ranking of their standard risks.

Result 1. Assume that the risk-value preference function f is consistent with expected utility theory. Then f can be represented as the following standard risk-value form [21]:

$$f(\bar{X}, R(X')) = u(\bar{X}) - \varphi(\bar{X})[R(X') - R(0)] \quad (6)$$

if and only if the risk-independence condition holds, where $\varphi(\bar{X}) > 0$ and $u(\cdot)$ is a von Neumann and Morgenstern [15] utility function.

This model (Eq. 6) shows that an expected utility model could have an alternative representation if this risk-independence condition holds. If one is risk averse, then $u(\cdot)$ is a concave function and $R(X') - R(0)$ is always positive. $u(\bar{X})$ provides a measure of value for the mean. If we did not consider the riskiness of a lottery X , it would have the value $u(\bar{X})$. Since it is risky, $u(\bar{X})$ is reduced by an amount proportional to the normalized risk measure $R(X') - R(0) \cdot \varphi(\bar{X})$ is a trade-off factor that may depend on the mean. If we further require the utility model to be continuously differentiable, then it must be either a quadratic, exponential, or a linear plus exponential model [21].

There are also some other alternative forms of risk-value models within the expected utility framework under different preference conditions [8,9,31]. In addition, for nonnegative lotteries such as those associated with the price of a stock, Dyer and Jia [32] propose a relative risk-value framework in the form $X = \bar{X} \times (X/\bar{X})$ which decomposes the return into an average return $\bar{X}$ and a percentage-based risk factor $X/\bar{X}$. We find that this form of a risk-value

model is compatible with the logarithmic (or linear plus logarithmic) and the power (or linear plus power) utility functions [32]. Recent empirical studies by Weber *et al.* [33] indicate that this formulation may also be useful as the basis for a descriptive model of the sensitivity of humans and animals to risk.

However, the notion of risk-value trade-offs within the expected utility framework is very limited. Based on model 6, for example, consistency with expected utility imposes very restrictive conditions on the relationship between the risk measure $R(X') = -E[u(X - \bar{X})]$, the value measure $u(\bar{X})$, and the trade-off factor $\phi(\bar{X}) = u''(\bar{X})/u''(0)$ (for continuously differentiable utility models). In particular, the risk measure and the value measure must be based on the same utility function. However, a decision maker may deviate from this “consistency” and have different measures for risk and value if his choice is based on risk-value trade-offs.

In order to be more realistic and flexible in the framework of risk-value trade-offs, we consider a two-attribute structure $(\bar{X}, X')$ for the evaluation of a risky alternative X . In this way we can explicitly base the evaluation of lotteries on two attributes, mean and risk, so that the mean–risk (or risk-value) trade-offs are not necessarily consistent with the traditional expected utility framework.

We assume the existence of the von Neumann and Morgenstern expected utility axioms over the two-attribute structure $(\bar{X}, X')$ and require the risk-value model to be consistent with the two-attribute expected utility model, that is, $f(\bar{X}, R(X')) = E[U(\bar{X}, X')]$, where U is a two-attribute utility function. As a special case, when the relationship between $\bar{X}$ and X' is a simple addition, the risk-value model reduces to a traditional expected utility model, that is $f(\bar{X}, R(X')) = E[U(\bar{X}, X')] = E[U(\bar{X} + X')] = E[U(X)] = aE[u(X)] + b$, where $a > 0$ and b are constants.

To obtain some separable forms of the risk-value model, we need to have a risk-independence condition for the two-attribute structure. Let P^0 be the set of normalized lotteries with zero-expected values,

and $>$ a strict preference relation for the two-attribute structure.

Definition 2. For $X', Y' \in P^0$, if there exists a $w_0 \in \text{Re}$ for which $(w_0, X') > (w_0, Y')$, then $(w, X') > (w, Y')$ for all $w \in \text{Re}$.

This two-attribute risk-independence condition requires that if two lotteries have the same mean and one is preferred to the other, then transforming the lotteries by adding the same constant to all outcomes will not reverse the preference ordering. This condition is generally supported by our recent experimental studies [34].

Result 2. Assume that the risk-value preference function f is consistent with the two-attribute expected utility model. Then f can be represented as the following generalized risk-value form:

$$f(\bar{X}, R(X')) = V(\bar{X}) - \phi(\bar{X})[R(X') - R(0)] \quad (7)$$

if and only if the two-attribute risk-independence condition holds, where $\phi(\bar{X}) > 0$ and $R(X')$ is the standard measure of risk.

In contrast to the standard risk-value model (Eq. 6), the generalized risk-value model (Eq. 7) provides the flexibility of considering $V(\bar{X})$, $R(X')$, and $\phi(\bar{X})$ independently. Thus, we can choose different functions for the value measure $V(\bar{X})$ independent of the utility function. The expected utility measure is used only for the standard measure of risk. Even though expected utility theory has been challenged by some empirical studies for general lotteries, we believe that it should be appropriate for describing risky choice behavior within a special set of normalized probability distributions with the same expected values. For general lotteries with different means, however, our two-attribute risk-value model can deviate from the traditional expected utility preference. In fact, the generalized risk-value model can capture a number of decision paradoxes that violate the traditional expected utility theory [35].

If the utility function u is strictly concave, then $R(X') - R(0) > 0$ and Model 7 will reflect risk-averse behavior. In addition, if $V(\bar{X})$ is increasing and twice continuously differentiable, $\phi(\bar{X})$ is once continuously differentiable and $\phi'(\bar{X})/\phi(\bar{X})$ is nonincreasing; then the generalized risk-value model (Eq. 7) exhibits decreasing risk aversion if and only if $-V''(\bar{X})/V'(\bar{X}) > -\phi'(\bar{X})/\phi(\bar{X})$; and the generalized risk-value model (Eq. 7) exhibits constant risk aversion if and only if $-V''(\bar{X})/V'(\bar{X}) = -\phi'(\bar{X})/\phi(\bar{X})$ is a constant. Thus, if a decision maker is decreasingly risk averse and has a linear value function, then we must choose a decreasing function for the trade-off factor $\phi(\bar{X})$.

The basic form of the risk-value model may be further simplified if some stronger preference conditions are satisfied. When $\phi(\bar{X}) = k > 0$, Model 7 becomes the following additive form:

$$f(\bar{X}, R(X')) = V(\bar{X}) - k[R(X') - R(0)]. \quad (8)$$

When $\phi(\bar{X}) = -V(\bar{X}) > 0$, then Model 7 reduces to the following multiplicative form:

$$f(\bar{X}, R(X')) = V(\bar{X})R(X'), \quad (9)$$

where $R(0) = 1$ and $V(0) = 1$. In this multiplicative model, $R(X')$ serves as a value discount factor due to risk.

We describe measures of perceived risk based on the converse interpretation of the axioms of risk-value models in a companion article in this collection. These perceived risk models are simply a negative linear transformation of the risk-value model (Eq. 7) [30]. Our risk-value framework offers a unified approach to both, risk judgment and preference modeling.

GENERALIZED RISK-VALUE MODELS

According to the generalized risk-value model (Eq. 7), the standard measure of risk, value function, and trade-off factor can be considered independently. Some examples of the standard measure of risk $R(X')$ are provided in the section titled “Standard Measure of Risk.” The value measure $V(\bar{X})$ should be

chosen as an increasing function and may have the same functional form as a utility model. For appropriate risk-averse behavior, the trade-off factor $\phi(\bar{X})$ should be either a decreasing function or a positive constant; for example, $\phi(\bar{X}) = ke^{-b\bar{X}}$, where $k > 0$ and $b \geq 0$. We consider three types of risk-value models, namely, moments risk-value models, exponential risk-value models, and generalized disappointment models. For the corresponding perceived risk of each risk-value model, the reader is referred to the companion article in this collection (see **Axiomatic Models of Perceived Risk**).

Moments Risk-Value Models

People often use mean and variance to make trade-offs for financial decision making because they provide a reasonable approximation for modeling decision problems and are easy to implement [1–3, 36, 37]. In the past, expected utility theory has been used as a foundation for mean–variance models. Risk-value theory provides a better foundation for developing moments models that include the mean–variance model as a special case.

As an example, the mean–variance model, $\bar{X} - kE[(X - \bar{X})^2]$ where $k > 0$, is a simple risk-value model with variance as the standard measure of risk and a constant trade-off factor. Sharpe [36, 37] assumed this mean–variance model in his analysis for portfolio selection and the Capital Asset Pricing Model. However, under the expected utility framework, this mean–variance model is based on the assumptions that the investor has an exponential utility function and that returns are jointly normally distributed.

According to our risk-value theory, this mean–variance model is constantly risk averse. To obtain a decreasing risk-averse mean–variance model, we can simply use a decreasing function for the trade-off factor:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[(X - \bar{X})^2], \quad (10)$$

where $b, k > 0$.

For many decision problems, mean–variance models are an oversimplification.

Based on this risk-value framework, we can specify some richer moment models for risky decision making. First, let us consider the moment standard measure of risk (Eq. 2) for the additive risk-value model (Eq. 8):

$$f(\bar{X}, R(X')) = \bar{X} - k\{E[(X - \bar{X})^2] - cE[(X - \bar{X})^3]\}, \quad (11)$$

where $c, k > 0$. The three-moments model (Eq. 11) can be either risk averse or risk seeking, depending on the distribution of a lottery. For symmetric bets or lotteries not highly skewed (e.g., an insurance policy) such that $E[(X - \bar{X})^2] > cE[(X - \bar{X})^3]$, Model 11 will be risk averse. But for highly positive skewed lotteries (e.g., lottery tickets) such that the skewness overwhelms the variance, that is $E[(X - \bar{X})^2] < cE[(X - \bar{X})^3]$, then Model 11 will exhibit risk-seeking behavior. Therefore, an individual with preferences described by the moments model of Equation (11) would purchase both, insurance and lottery tickets simultaneously.

Markowitz [23] noticed that individuals have the same tendency to purchase insurance and lottery tickets whether they are poor or rich. This observed behavior contradicts a common assumption of expected utility theory that preference ranking is defined over ultimate levels of wealth. But whether our risk-value model is risk averse or risk seeking is determined only by the standard measure of risk, which is independent of an individual's wealth level (refer to the form of risk-value model in Eq. 4). In particular, for the three-moments model (Eq. 10), the change of wealth level only causes a parallel shift for $f(\bar{X}, R(X'))$; this will not affect the risk attitude and the choice behavior of this model. This is consistent with Markowitz's observation.

Exponential Risk-Value Models

If the standard measure of risk is based on exponential or linear plus exponential utility models, then the standard measure of risk is given by Equation (3). To be compatible with the form of the standard measure of risk, we can also use exponential functions, but with

different parameters, for the value measure $V(\bar{X})$ and the trade-off factor $\phi(\bar{X})$ for the generalized risk-value model (Eq. 7), which leads to the following model:

$$f(\bar{X}, R(X')) = -he^{-a\bar{X}} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1], \quad (10)$$

where a, b, c, h , and k are positive constants. When $a = b = c$ and $h = k$, this model reduces to an exponential utility model; otherwise, these two models are different. When $b > a$, Model 10 is decreasingly risk averse even though the traditional exponential utility model exhibits constant risk aversion.

As a special case, when $a = b$ and $h = k$, Model 10 reduces to the following simple multiplicative form:

$$f(\bar{X}, R(X')) = ke^{-a\bar{X}}E[e^{-c(X-\bar{X})}]. \quad (11)$$

This model is constantly risk averse, and therefore has the same risk attitude as an exponential utility model. It has more flexibility since there are two different parameters. This simple risk-value model can be used to explain some well-known decision paradoxes [35]

Choosing a linear function or a linear plus exponential function for $V(\bar{X})$ leads to the following models:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1] \quad (12)$$

and

$$f(\bar{X}, R(X')) = \bar{X} - he^{-a\bar{X}} - ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1]. \quad (13)$$

Model 12 is decreasingly risk averse. Model 13 includes a linear plus exponential utility model as a special case when $a = b = c$ and $h = k$. It is decreasingly risk averse if $b \geq a$.

Generalized Disappointment Models

Bell [38] proposed a disappointment model for decision making under uncertainty. According to Bell, disappointment is a psy-

chological reaction to an outcome that does not meet a decision maker's expectation. Bell used the mean of a lottery as a decision maker's psychological expectation. If an outcome smaller than the expected value occurs, the decision maker would be disappointed. Otherwise, the decision maker would be elated. Although Bell's development of the disappointment model has an intuitive appeal, his model is applicable only to lotteries with two outcomes.

Jia *et al.* [27] use the risk-value framework to develop a generalized version of Bell's disappointment model [38]. Consider the following piece-wise linear utility model:

$$u(x) = \begin{cases} ex & \text{when } x \geq 0, \\ dx & \text{when } x < 0, \end{cases} \quad (14)$$

where $d, e > 0$ are constant. Decision makers who are averse to downside risk or losses should have $d > e$, as illustrated in Fig. 1. The standard measure of risk for this utility model can be obtained as follows:

$$\begin{aligned} R(X') &= dE^-[|X - \bar{X}|] - eE^+[|X - \bar{X}|] \\ &= [(d - e)/2]E[|X - \bar{X}|], \end{aligned} \quad (15)$$

where $E^-[|X - \bar{X}|] = \sum_{x_i < \bar{X}} p_i |x_i - \bar{X}|$ and $E^+[|X - \bar{X}|] = \sum_{x_i > \bar{X}} p_i (x_i - \bar{X})$, and $E[|X - \bar{X}|]$

is the absolute standard deviation. Following Bell's basic idea [38], $dE^-[|X - \bar{X}|]$ represents a general measure of expected disappointment and $eE^+[|X - \bar{X}|]$ represents a general measure of expected elation. The overall psychological satisfaction is measured by $-R(X')$, which is the converse of the standard measure of risk (Eq. 15).

If we assume a linear value measure and a constant trade-off factor, then we can have the following risk-value model based on the measure of disappointment risk (Eq. 15):

$$\begin{aligned} f(\bar{X}, R(X')) &= \bar{X} - \{dE^-[|X - \bar{X}|] - eE^+[|X - \bar{X}|]\} \\ &= \bar{X} - [(d - e)/2]E[|X - \bar{X}|]. \end{aligned} \quad (16)$$

For a two-outcome lottery, Model 16 reduces to Bell's disappointment model. Thus, we call the risk-value model (Eq. 16)

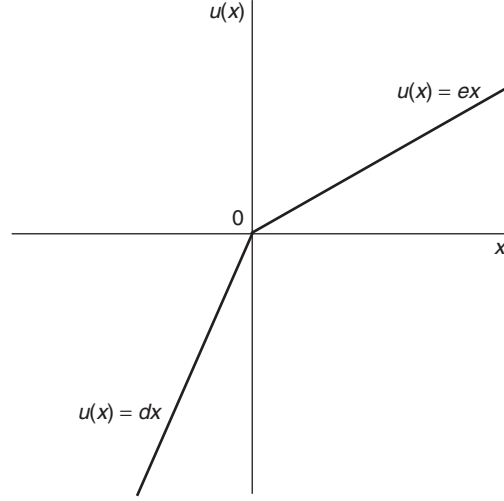


Figure 1. A piece-wise linear utility function.

a “generalized disappointment model.” Note that the risk-value model [16] will be consistent with the piece-wise linear utility model (Eq. 14) when the lotteries considered have the same means. It is a risk-averse model when $d > e$.

Using his two-outcome disappointment model, Bell gave an explanation for the common ratio effect [38]. Our generalized disappointment model (Eq. 16) can explain the Allais Paradox [10,11], which involves an alternative with three outcomes [27]. Bell's model and Equation (16) both imply constant risk aversion and are not appropriate for decreasingly risk-averse behavior. To obtain a disappointment model with decreasing risk aversion, we can use a decreasing function for the trade-off factor:

$$f(\bar{X}, R(X')) = \bar{X} - ke^{-b\bar{X}}E[|X - \bar{X}|]. \quad (17)$$

Bell's disappointment model (Eq. 16) implies that disappointment and elation are proportional to the difference between the expected value and an outcome. We can use risk model (Eq. 5) to incorporate nonlinear functions for disappointment and elation in a more general form of the disappointment

model:

$$f(\bar{X}, R(X')) = \bar{X} - dE^-[|X - \bar{X}|^{\theta_2}] - eE^+[|X - \bar{X}|^{\theta_1}]. \quad (18)$$

When $\theta_1 = \theta_2 = 1$, this model reduces to Model 16. When $e = 0$ and $\theta_2 = 2$, Model 18 becomes a mean-semivariance model.

Finally, our generalized disappointment models are different from Loomes and Sugden's development [39]. In their basic model, disappointment (or elation) is measured by some function of the difference between the utility of outcomes and the expected utility of a lottery. They also assume a linear "utility" measure of wealth and the same sensation intensity for both disappointment and elation, so that their model has the form, $\bar{X} + E[D(X - \bar{X})]$, where $D(x - \bar{X}) = -D(\bar{X} - x)$, and D is continuously differentiable and convex for $x > \bar{X}$ (thus concave for $x < \bar{X}$). Even though this model is different from our generalized disappointment models (Eq. 18), it is a special case of our risk-value model with a linear measure of value, a constant trade-off factor, and a specific form of the standard measure of risk (i.e., $R(X') = -E[D(X - \bar{X})]$, where $D(x - \bar{X}) = -D(\bar{X} - x)$). Loomes and Sugden used this model to provide an explanation for the choice behavior that violates Savage's Sure-Thing Principle [40].

CONCLUSION

We have summarized our efforts to incorporate the intuitively appealing idea of risk-value trade-offs into decision making under risk. The risk-value framework ties together two streams of research: one in modeling risk judgments and the other in developing preference models, and unifies a wide range of decision phenomena including both, normative and descriptive aspects.

This development also refines and generalizes a substantial number of previously proposed decision theories and models, ranging from the mean-variance model in finance to the disappointment models in decision sciences. It is also possible to create many new risk-value models. Specifically, we

have discussed three classes of decision models based on this risk-value theory: moments risk-value models, exponential risk-value models, and generalized disappointment risk-value models. These models are very flexible in modeling preferences. They also provide new resolutions for observed risky choice behavior and the decision paradoxes that violate the independence axiom of the expected utility theory.

The most important assumption in this study is the two-attribute risk-independence condition, which leads to a separable form of risk-value models. Although some other weaker condition could be used to derive a risk-value model that has more descriptive power, this reduces the elegance of the basic risk-value form, and increases operational difficulty. Butler *et al.* [34] conducted an empirical study of this key assumption, and found some support for it. This study also highlighted some additional patterns of choices indicating that the translation of lottery pairs from the positive domain to the negative domain often results in the reversal of preference and risk judgments. To capture this phenomenon, we have extended risk-independence conditions to allow the trade-off factor in the risk-value models to change sign, and therefore to infer risk aversion in the positive domain and risk seeking in the negative domain. These generalized risk-value models provide additional insights into the reflection effects [12] and related empirical results [26,41,42].

Even though some other non-expected utility theories that have been proposed (e.g., Prospect Theory and other weighted utility theories) may produce the same predictions for the decision paradoxes as risk-value theory, it offers a new justification for them based on the appealing and realistic notion of risk-value trade-offs. In particular, since the role of risk is considered *implicitly* in these decision theories and models, they are not compatible with choice behavior that is based on the risk and mean return trade-offs often encountered in financial management, psychology, and other applied fields. Therefore, these theories and models offer little guidance in practice for this type of decision making. We believe that the potential for

contributions of these risk-value models in finance is very exciting. Applications of our risk-value models in other fields such as economics, marketing, insurance, and risk management are also promising.

Acknowledgments

This article summarizes a stream of research on risk and risk-value models. In particular, we have incorporated materials that appeared previously in the following papers: (i) Jia J, Dyer JS. Risk-value theory. Working Paper, Graduate School of Business, University of Texas at Austin, TX 1995; (ii) Jia J, Dyer JS. A standard measure of risk and risk-value models. *Management Science* 1996 42: 1961–1705; (iii) Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Management Science* 1999 45: 519–532; (iv) Jia, J, and Dyer, JS, Decision making based on risk value theory, in *The Mathematics of Preference, Choice and Order: Essays in Honor of Peter C. Fishburn* (Edited by Steven Brams, William V. Gehrlein and Fred S. Roberts), published by Springer Publishing, 2008.

REFERENCES

1. Markowitz HM. Portfolio selection. New York: Wiley; 1959.
2. Markowitz HM. Mean-variance analysis in portfolio choice and capital markets. New York: Basil Blackwell; 1987.
3. Markowitz HM. Foundations of portfolio theory. *J Finance* 1991;XLVI:469–477.
4. Stone B. A general class of 3-parameter risk measures. *J Finance* 1973;28:675–685.
5. Fishburn PC. Mean-risk analysis with risk associated with below-target returns. *Am Econ Rev* 1977;67:116–126.
6. Meyer J. Two-moment decision models and expected utility maximization. *Am Econ Rev* 1987;77:421–430.
7. Bell DE. One-switch utility functions and a measure of risk. *Manage Sci* 1988;34:1416–1424.
8. Bell DE. Risk, return, and utility. *Manage Sci* 1995;41:23–30.
9. Sarin RK, Weber M. Risk-value models. *Eur J Oper Res* 1993;70:135–149.
10. Allais M. Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école américaine. *Econometrica* 1953;21:503–546.
11. Allais M. The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school. In: Allais M, Hagen O, editors. Expected utility hypotheses and the Allais paradox. Dordrecht: D. Reidel; 1979. pp. 27–145.
12. Kahneman DH, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47:263–290.
13. Machina MJ. Choice under uncertainty: problems solved and unsolved. *Econ Perspect* 1987;1:121–154.
14. Weber EU. Decision and choice: risk, empirical studies. In: Smelser N, Baltes P, editors. *International encyclopedia of the social sciences*, Oxford: Elsevier Science Limited; 2001. pp. 13347–13351.
15. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton University Press; 1947.
16. Pratt JW. Risk aversion in the small and in the large. *Econometrica* 1964;32:122–136.
17. Dyer JS, Sarin RK. Relative risk aversion. *Manage Sci* 1982;28:875–886.
18. Quiggin J. A theory of anticipated utility. *J Econ Behav Organ* 1982;3:323–343.
19. Tversky A, Kahneman DH. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5:297–323.
20. Wu G, Gonzalez R. Curvature of the probability weighting function. *Manage Sci* 1996;42:1676–1690.
21. Jia J, Dyer JS. A standard measure of risk and risk-value models. *Manage Sci* 1996;42:1961–1705.
22. Friedman M, Savage LP. The utility analysis of choices involving risk. *J Polit Econ* 1948;56:279–304.
23. Markowitz HM. The utility of wealth. *J Polit Econ* 1952;60:151–158.
24. Weber M. Risikoentscheidungskalküle in der Finanzierungstheorie. Stuttgart: Poeschel; 1990.
25. Sarin RK. Some extensions of Luce's measures of risk. *Theory Decis* 1987;22:25–141.
26. Fishburn PC, Kochenberger GA. Two-piece von Neumann-Morgenstern utility functions. *Decis Sci* 1979;10:503–518.

27. Jia J, Dyer JS, Butler JC. Generalized disappointment models. *J Risk Uncertain* 2001;22:59–78.
28. Konno H, Yamazaki H. Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market. *Manage Sci* 1992;37:519–531.
29. Dennenberg D. Premium calculation: Why standard deviation should be replaced by absolute deviation. *Astin Bull* 1990;20:181–190.
30. Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Manage Sci* 1999;45:519–532.
31. Dyer JS, Jia J. Preference conditions for utility models: a risk-value perspective. *Ann Oper Res* 1998;80:167–182.
32. Dyer JS, Jia J. Relative risk-value models. *Eur J Oper Res* 1997;103:170–185.
33. Weber EU, Shafir S, Blais A. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychol Rev* 2004;111:430–445.
34. Butler J, Dyer J, Jia J. An empirical investigation of the assumptions of risk-value models. *J Risk Uncertain* 2005;30:133–156.
35. Jia J. Measures of risk and risk-value theory [Unpublished Ph.D. Dissertation]: University of Texas at Austin, TX; 1995.
36. Sharpe WF. Portfolio theory and capital markets. New York: McGraw-Hill; 1970.
37. Sharpe WF. Capital asset prices with and without negative holdings. *J Finance* 1991;46:489–509.
38. Bell DE. Disappointment in decision making under uncertainty. *Oper Res* 1985;33:1–27.
39. Loomes G, Sugden R. Disappointment and dynamic consistency in choice under uncertainty. *Rev Econ Stud* 1986;LIII:271–282.
40. Savage LJ. The foundations of statistics. New York: Wiley; 1954.
41. Payne JW, Laughhunn DJ, Crum R. Translation of gambles and aspiration level effects in risky choice behavior. *Manage Sci* 1980;26:1039–1060.
42. Payne JW, Laughhunn DJ, Crum R. Further tests of aspiration level effects in risky choice behavior. *Manage Sci* 1981;27:953–958.

AXIOMATIC MODELS OF PERCEIVED RISK

JIANMIN JIA

Department of Marketing, Faculty of
Business Administration, The Chinese
University of Hong Kong, Shatin,
Hong Kong

JAMES S. DYER

Department of Information, Risk, and
Operations Management, The
McCombs School of Business,
University of Texas at Austin, Austin,
Texas

JOHN C. BUTLER

Department of Finance, The McCombs
School of Business, University of
Texas at Austin, Austin, Texas

INTRODUCTION

In a separate article in this collection (see *Axiomatic Measures of Risk and Risk-Value Models*), we discussed a standard measure of risk based on the converse of the expected utility of lotteries with zero-expected values, and showed how it could be used to derive risk-value models for risky choices within both an expected utility framework and a nonexpected utility framework [refer also [1–3]]. Though this standard measure of risk has some descriptive power for risk judgments, it is more normative in nature. In particular, since the standard measure of risk eliminates the effect of the mean of a lottery, it only measures the “pure” risk of the lottery, and may not be appropriate for modeling perceptions of risk. The purpose of this article is to review previously proposed models of perceived risk, with an emphasis on those perceived risk models that can be related to preference in a natural way. We then describe a two-attribute structure of perceived risk that allows the mean to impact judgments of risk perception and that can also be related to risk-value models for risky choice.

Over the last 30 years, researchers have expended much effort toward developing and testing models of the perceived riskiness of lotteries. Pollatsek and Tversky [4] provide an early summary of risk research that is still meaningful today:

The various approaches to study of risk share three basic assumptions. 1. Risk is regarded as a property of options (e.g., gambles, courses of action) that affects choices among them. 2. Options can be meaningfully ordered with respect to their riskiness. 3. The risk of an option is related in some way to the dispersion, or the variance, of its outcomes.

As stated in the first assumption, risk is a characteristic of a lottery that affects decisions. This is the primary motivation for studying the nature of perceived risk. A measure of perceived risk may be used as a variable in preference models, such as Coombs’ Portfolio theory [5–8] in which a choice among lotteries is a compromise between maximizing expected value and optimizing the level of perceived risk. However, the “risk” measure in Coombs’ Portfolio theory is left essentially undefined, and is considered to be an independent theory. This has stimulated a long stream of research on the measure of perceived risk.

Empirical studies have demonstrated that people are consistently able to order lotteries with respect to their riskiness, and that risk judgments satisfy some basic axioms [9,10]. Thus, as stated in the second assumption, the term “riskiness” should be both a meaningful and measurable characteristic of lotteries.

There have been many refinements to the third assumption as experimental results have exposed how perceived risk judgments change as a function of the characteristics of the lotteries considered. Some stylized facts regarding risk judgments include the following:

- Perceived risk increases when there is an increase in range, variance, or expected loss, for example [11].

- Perceived risk decreases if a constant positive amount is added to all outcomes of a lottery [9,12].
- Perceived risk increases if all outcomes of a lottery with zero-mean are multiplied by a positive constant greater than one [6].
- Perceived risk increases if a lottery with zero-mean is repeated many times [6].

These empirically verified properties provide basic guidelines for developing and evaluating measures of perceived risk.

In the following section, we give a review of previously proposed models of perceived risk and discuss their performance in empirical studies. In the section titled “Two-Attribute Models for Perceived Risk,” we present our measures of perceived risk based on a two-dimensional structure of the standard risk of a lottery and its mean show that it includes many of these previously proposed models as special cases. Finally, in the section titled “Conclusions,” we provide a summary and discuss the implications of the two-attribute perceived risk structure in decision making under risk.

REVIEW OF PERCEIVED RISK MODELS

The literature contains various attempts to define risk based on different assumptions about how perceived risk is formulated and how it evolves as the risky prospects under consideration are manipulated. In this section, we review some previously proposed models for perceived risk and their key assumptions. We focus on those that are closely related to a preference-dependent measure of risk that is compatible with traditional expected utility theory.

A more detailed review of perceived risk studies, including some measures of risk that are not closely related to preference, is provided by Brachinger and Weber [13]. We also exclude from our discussion the “coherent measures of risk” developed in the financial mathematics literature by Arztnar *et al.* [14,15], which produce results in monetary units (e.g., in dollars) that estimate the “expected shortfall” associated

with a portfolio. A recent review of the latter is given by Acerbi [16].

Studies by Coombs and his Associates

In early studies, risk judgments were represented by using the moments of a distribution and their transformations, that is, distributive models. Expected value, variance, skewness, range, and the number of repeated plays have been investigated as possible determinants of risk judgments [6,11,17]. Coombs and Huang [7] considered several composition functions of three indices corresponding to transformations on two-outcome lotteries, and their paper supported a distributive model that is based on a particular structure of the joint effect of these transformations on perceived risk. However, evidence to the contrary of such a distributive model was also found by Barron [18].

Coombs and Lehner [12] used distribution parameters as variables in the distributive model to test if moments of distributions are useful in assessing risk. For a lottery $(b + 2(1 - p)a, p; b - 2pa)$ (this means that the lottery has an outcome $(b + 2(1 - p)a)$ with probability p and an outcome $(b - 2pa)$ otherwise), which has a mean equal to b and range $2a$, the distributive model is represented by

$$R(a, b, p) = [\phi_1(a) + \phi_2(b)] \phi_3(p), \quad (1)$$

where R is a riskiness function and ϕ_1, ϕ_2 and ϕ_3 are real-valued monotonic functions defined on a, b and p , respectively. Coombs and Lehner [12] showed that the distributive model (Eq. 1) is not acceptable as a descriptive model of risk judgment. They concluded that complex interactions between the properties of risky propositions prevent a simple polynomial expression of the variables a, b and p from capturing perceived riskiness.

Coombs and Lehner [19] further considered perceived risk as a direct function of outcomes and probabilities, with no intervening distribution parameters. They assumed a bilinear model. In the case of just three outcomes (positive, negative, and zero), perceived risk is represented by the following

model:

$$R(X) = \phi_1(p)\phi_2(w) + \phi_3(q)\phi_4(l), \quad (2)$$

where w and l represent the positive and negative outcomes, with probabilities p and q ($p + q = 1$), respectively; R and ϕ_i ($i = 1, 2, 3$, and 4) are real-valued functions and X represents the lottery. The model assumes that a zero outcome and its associated probability have no direct effect on perceived risk. The form of model (2) is similar to Prospect theory [20]. Coombs and Lehner's [19] experiment supported the notion that perceived risk can be decomposed into contributions from good and bad components, and the bad components play a larger role than the good ones.

Pollatsek and Tversky's Risk Theory

An important milestone in the study of perceived risk is the axiomatization of risk theory developed by Pollatsek and Tversky [4]. They assumed four axioms for a risk system: (i) weak ordering; (ii) cancellation (or additive independence); (iii) solvability; and (iv) an Archimedean property. Let P denote a set of simple probability distributions or lotteries $\{X, Y, Z, \dots\}$ and $\geq_R$ be a binary risk relation (meaning at least as risky as). For convenience, we will use X, Y , and Z to refer to random variables, probability distributions or lotteries interchangeably. Pollatsek and Tversky showed that the four axioms imply that there exists a real-valued function R on P such that for lotteries X and Y : (i) $X \geq_R Y$ if and only if $R(X) \geq R(Y)$; (ii) $R(X \circ Y) = R(X) + R(Y)$, where " $\circ$ " denotes the binary operation of adding independent random variables; that is, the convolution of their density functions.

Pollatsek and Tversky considered three additional axioms: (v) positivity; (vi) scalar monotonicity; and (vii) continuity. These three additional axioms imply that R is a linear combination of mean and variance:

$$R(X) = -\theta \bar{X} + (1 - \theta)E[(X - \bar{X})^2], \quad (3)$$

where $0 < \theta = 1$.

However, the empirical validity of Equation (3) was criticized by Coombs and

Bowen [21], who showed that factors other than mean and variance, such as skewness, affect perceived risk. In Pollatsek and Tversky's system of axioms, the continuity condition based on the central limit theorem is directly responsible for the form of the mean-variance model (Eq. 3). Coombs and Bowen [21] showed that skewness impacts perceived risk even under multiple plays of a lottery when the effect of the central limit theorem modifies the effect of skewness.

Another empirically questionable axiom is the additive independence condition, which says that, for X, Y and Z in P , $X \geq_R Y$, if and only if $X \circ Z \geq_R Y \circ Z$. Fishburn [22] provides the following example of a setting where additive independence is unlikely to hold. Many people feel that a lottery $X = (\$1000, 0.01; -\$10,000)$ (i.e., X has probability 0.01 of a \$1000 gain, and a \$10,000 loss otherwise) is riskier than $Y = (\$2000, 0.5; -\$12,000)$. Consider another degenerate lottery Z that has a sure \$11,000 gain. Since $X \circ Z$ yields at least a \$1000 gain, whereas $Y \circ Z$ results in a loss of \$1000 with probability 1/2, it seems likely that most people would consider $Y \circ Z$ to be riskier than $X \circ Z$. This risk judgment pattern is inconsistent with additive independence. Empirical studies have also failed to support the independence condition [23,24].

Nevertheless, some of Pollatsek and Tversky's axioms are very appealing, such as positivity and scalar monotonicity. Because they are important to the present article, we briefly introduce them here. According to the positivity axiom, if K is a degenerate lottery with an outcome $k > 0$, then $X \geq_R X \circ K$ for all X in P . In other words, the addition of a positive sure-amount to a lottery would decrease its perceived risk. This quality is considered an essential property of perceived risk and has been confirmed by several empirical studies (e.g., [9,12]).

Another appealing axiom in Pollatsek and Tversky's theory is scalar monotonicity, which says, for all X, Y in P with $E(X) = E(Y) = 0$, (i) $\beta X \geq_R X$ for $\beta > 1$; (ii) $X \geq_R Y$ if and only if $\beta X \geq_R \beta Y$ for $\beta > 0$. This axiom asserts that, for lotteries with

zero expectation, risk increases when the lottery is multiplied by a real number $\beta > 1$ [also refer to [6]], and that the risk ordering is preserved upon a scale change of the lotteries (e.g., dollars to pennies). Pollatsek and Tversky regarded the positivity axiom and part (i) of the monotonicity axiom as necessary assumptions for any theory of risk.

In a more recent study, Rotar and Sholomitsky [25] weakened part (ii) of the scalar monotonicity axiom (coupled with some other additional conditions) to arrive at a more flexible risk model that is a finite linear combination of cumulants of higher orders. This generalized risk model can take into account additional characteristics of distributions such as skewness and other higher order moments. However, because Rotar and Sholomitsky's risk model still retains the additive independence condition as a basic assumption, their model would be subject to the same criticisms regarding the additivity of risk.

Luce's Risk Models and Others

Subsequent to the criticisms of Pollatsek and Tversky's risk measure, Luce [26] approached the problem of risk measurement in a different way. He began with a multiplicative structure of risk. First, Luce considered the effect of a change of scale on risk; multiplying all outcomes of a lottery by a positive constant. He assumed two simple possibilities, an additive effect and a multiplicative effect, presented as follows:

$$R(\alpha * X) = R(X) + S(\alpha) \quad (4)$$

or

$$R(\alpha * X) = S(\alpha)R(X), \quad (5)$$

where α is a positive constant and S is a strictly increasing function with $S(1) = 0$ for Equation (4) and $S(1) = 1$ for Equation (5). Luce's assumptions, Equations (4) and (5), are related to Pollatsek and Tversky's [4] scalar monotonicity axiom. But an important difference is that Pollatsek and Tversky only applied this assumption to the lotteries with zero expected values. As we will see later,

this may explain why Luce's models have not been supported by experiments.

Then, Luce considered two ways in which the outcomes and probabilities of a lottery could be aggregated into a single number. The first aggregation rule is analogous to the expected utility form and leads to an expected risk function:

$$R(X) = \int_{-\infty}^{\infty} T(x)f(x) dx = E[T(X)], \quad (6)$$

where T is some transformation function of the random variable X and $f(x)$ is the density of lottery X . In the second aggregation rule, the density goes through a transformation before it is integrated,

$$R(X) = \int_{-\infty}^{\infty} T^+[f(x)] dx, \quad (7)$$

where T^+ is some nonnegative transformation function. The combinations of the two decomposition rules, Equations (4) and (5), and the two aggregation rules, Equations (6) and (7), yield four possible measures of risk as follows:

1. By Equations (4) and (6),

$$\begin{aligned} R(X) = & a \int_{x \neq 0} \log|x| dx \\ & + b_1 \int_{-\infty}^0 f(x) dx \\ & + b_2 \int_0^{\infty} f(x) dx, \quad a > 0. \end{aligned} \quad (8)$$

2. By Equations (5) and (6),

$$\begin{aligned} R(X) = & a_1 \int_{-\infty}^0 |x|^\theta dx \\ & + a_2 \int_0^{\infty} x^\theta dx, \quad \theta > 0. \end{aligned} \quad (9)$$

3. By Equations (4) and (7),

$$\begin{aligned} R(X) = & -a \int_{-\infty}^{\infty} f(x) \log f(x) dx \\ & + b, \quad a > 0, b = 0. \end{aligned} \quad (10)$$

4. By Equations (5) and (7),

$$R(X) = a \int_{-\infty}^{\infty} f(x)^{1-\theta} dx, \quad a, \theta > 0. \quad (11)$$

Luce's risk models did not receive positive support from an experimental test by Keller, Sarin and Weber [9]. As Luce himself noted, an obvious drawback of the models (Eq. 10 and 11) is that both measures require that risk should not change if we add or subtract a constant amount to all outcomes of a lottery. This is counter to intuition and to empirical evidence [9,12]. Luce's absolute logarithmic measure (Eq. 8) is also neither empirically valid [9] nor prescriptively valid [27]. In Keller *et al.*'s experiment [9], only the absolute power model (Eq. 9) seems to have some promise as a measure of perceived risk.

Following Luce's work, Sarin [27] considered the observation that when a constant is added to all outcomes of a lottery, perceived risk should decrease. He assumed that there is a strictly monotonic function S such that for all lotteries and any real number β ,

$$R(\beta \circ X) = S(\beta)R(X). \quad (12)$$

Together with the expectation principle in Equation (6), this yields an exponential form of risk model,

$$R(X) = kE(e^{cX}), \quad (13)$$

where $kc < 0$.

Luce's and Sarin's models employ the expectation principle, which was first postulated for risk by Huang [28]. The expectation principle—an application of the independence axiom of expected utility theory—means that the risk measure R is linear under convex combinations:

$$R(\lambda X + (1 - \lambda)Y) = \lambda R(X) + (1 - \lambda)R(Y), \quad (14)$$

where $0 < \lambda < 1$. Empirical studies have suggested that this assumption may not be valid for risk judgments [9,19]. Weber and Bottom [10] showed, however, that the independence axiom is not violated for risk judgments,

but that the culprit is the so-called probability accounting principle [29]. These findings cast doubt on any perceived risk models based on the expectation principle, including Luce's logarithmic model (Eq. 8) and power model (Eq. 9) and Sarin's exponential model (Eq. 13).

Sarin also generalized the simple expectation principle using Machina's [30] non-expected utility theory and extended Luce's model (Eqs 8 and 9) into some more complicated risk models. However, since risk judgment is not identical to choice preference under risk, Sarin's proposal needs to be tested empirically.

Luce and Weber [29] proposed a revision of Luce's original power model (Eq. 8) based on empirical findings. This Conjoint Expected Risk (CER) model has the following form:

$$\begin{aligned} R(X) = & A_0 \Pr(X = 0) + A_+ \Pr(X > 0) \\ & + A_- \Pr(X < 0) \\ & + B_+ E[X^{K_+} | X > 0] \Pr(X > 0) \\ & + B_- E[|X|^{K_-} | X < 0] \Pr(X < 0), \end{aligned} \quad (15)$$

where A_0, A_+ , and A_- are probability weights, and B_+ and B_- are weights of the conditional expectations, raised to some positive powers, K_+ and K_- . The major advantage of the CER model is that it allows for asymmetric effects of transformations on positive and negative outcomes. Weber [31] showed that the CER model describes risk judgments reasonably well. One possible drawback of the CER model is that the lack of parsimony provides the degrees of freedom to fit any set of responses.

Weber and Bottom [10] tested the adequacy of the axioms underlying the CER model and found that the conjoint structure assumptions about the effect of change of scale transformations on risk hold for negative outcome lotteries, but not for positive outcome lotteries. This suggests that the multiplicative independence assumption [i.e., for positive (or negative) outcome-only lotteries X and Y , $X \geq_R Y$ if and only if $\beta X \geq_R \beta Y$ for $\beta > 0$] may not be valid. Note that Pollatsek and Tversky's [4] scalar monotonicity axiom is identical to multiplicative independence

but is only assumed to hold for lotteries with zero expected values.

Fishburn's Risk Systems

Fishburn [32,33] explored risk measurement from a rigorous axiomatic perspective. In his two-part study on axiomatizations of perceived risk, he considered lotteries separated into gains and losses relative to a target, say 0. Then the general set P of measures can be written as $P = \{(\alpha, p; \beta, q) : \alpha \geq 0, \beta \geq 0, \alpha + \beta \leq 1, p \in P^-, q \in P^+\}$, where α is the loss probability, p is the loss distribution given a loss, β is the gain probability, q is the gain distribution given a gain, $1 - p - q$ is the probability for the target outcome 0, and P^- and P^+ are the sets of probability measures defined on loss and gain, respectively. The risk measure in this general approach satisfies $(\alpha, p; \beta, q) \geq_R (\gamma, r; \delta, s)$, if and only if $R(\alpha, p; \beta, q) \geq R(\gamma, r; \delta, s)$.

Fishburn assumed that there is no risk if and only if there is no chance of getting a loss, which implies $R = 0$ if and only if $\alpha = 0$. This rules out additive forms of risk measures such as $R(\alpha, p; \beta, q) = R_1(\alpha, p) + R_2(\beta, q)$, but allows forms that are multiplicative in losses and gains; for example, $R(\alpha, p; \beta, q) = \rho(\alpha, p) * \tau(\beta, q)$. If ρ and τ are further decomposable, then the risk measure can be

$$R(\alpha, p; \beta, q) = \left[\rho_1(\alpha) \int_{x < 0} \rho_2(x) dp(x) \right] \times \left[1 - \tau_1(\beta) \int_{y > 0} \tau_2(y) dq(y) \right]. \quad (16)$$

According to Fishburn [33], the first part of this model measures the “pure” risk involving losses and the second measures the effect of gains on the risk. In the multiplicative model in Equation (16), gains proportionally reduce risk independent of the particular (α, p) involved (unless the probability of a loss, α , is zero, in which case there is no risk to be reduced). Fishburn did not suggest functional forms for the free functions

in his models, so it is difficult to test them empirically.

In summary, since the pioneering work of Coombs and his associates' on perceived risk, several formal theories and models have been proposed. But none of these risk models is fully satisfactory. As Pollatsek and Tversky [4] wrote, “. . . our intuitions concerning risk are not very clear and a satisfactory operational definition of the risk ordering is not easily obtainable.” Nevertheless, empirical studies have observed a remarkable consistency in risk judgments [9,10] suggesting the existence of robust measures of perceived risk.

TWO-ATTRIBUTE MODELS FOR PERCEIVED RISK

In this section, we propose two-attribute models for perceived risk based on the mean of a lottery and the standard measure of risk that is discussed in another article in this collection. In particular, our measures of perceived risk can be incorporated into preference models based on the notion of risk-value trade-offs. For the purpose of application, we suggest several explicit functional forms for the measures of perceived risk.

A Two-Attribute Structure for Perceived Risk

A common approach in previous studies of perceived risk is to look for different factors that are responsible for risk perceptions underlying a lottery, such as mean and variance or other risk dimensions, and then consider some separation or aggregation rules to obtain a new measurement model (see Payne [34] for a review). Jia and Dyer [2] decomposed a lottery X into its mean $\bar{X}$ and its standard risk, $X' = X - \bar{X}$, and proposed a standard measure of risk based on expected utility theory:

$$R(X') = -E[u(X')] = -E[u(X - \bar{X})] \quad (17)$$

where $u(\cdot)$ is a von Neumann and Morgenstern [35] utility function. The mean of a lottery serves as a status quo for measuring the standard risk. See the complementary article (see *Axiomatic Measures of Risk*

and Risk-Value Models) in this collection for a summary of this development.

The standard measure of risk has many desirable properties that characterize the “pure” risk of lotteries. It can provide a suitable measure of perceived risk for lotteries with zero expected values as well. However, the standard measure of risk would not be appropriate for modeling people’s perceived risk for general lotteries since the standard measure of risk is independent of expected value or any certain payoffs. That is, if $Y = X + k$, where k is a constant, then $Y' = Y - \bar{Y} = X - \bar{X} = X'$. As we discussed earlier, empirical studies have shown that people’s perceived risk decreases as a positive constant amount is added to all outcomes of a lottery.

To incorporate the effect of the mean of a lottery on perceived risk, we consider a two-attribute structure for evaluating perceived risk; that is, $(\bar{X}, X')$. In fact, a lottery X can be represented by its expected value $\bar{X}$ and the standard risk X' exclusively, for example, $X = \bar{X} + X'$. Thus, $(\bar{X}, X')$ is a natural extension of the representation of the lottery X . This two-attribute structure has an intuitive interpretation in risk judgment. When people make a risk judgment, they may first consider the variation or uncertainty of the lottery, measured by X' , and then take into account the effect of expected value on the uncertainty perceived initially, or vice versa.

To develop our measure of perceived risk formally, let $\mathbf{P}$ be the set of all simple probability distributions, including degenerate distributions, on a nonempty product set, $\mathbf{X}_1 \times \mathbf{X}_2$, of outcomes, where $\mathbf{X}_i \subseteq \text{Re}$, $i = 1, 2$, and Re is the set of real numbers. For our special case, the outcome of a lottery X on $\mathbf{X}_1$ is fixed at its mean $\bar{X}$; thus, the marginal distribution on $\mathbf{X}_1$ is degenerate with a singleton outcome $\bar{X}$. Because $\bar{X}$ is a constant, the two “marginal distributions” $(\bar{X}, X')$ are sufficient to determine a unique distribution in $\mathbf{P}$.

Let $>_{\bar{\mathbf{R}}}$ be a strict risk relation, $\sim_{\bar{\mathbf{R}}}$ an indifference risk relation, and $\geq_{\bar{\mathbf{R}}}$ a weak risk relation on $\mathbf{P}$. We assume a two-attribute case of the expectation principle and other necessary conditions analogous to those of multiattribute utility theory (e.g., Refs 36 and 37) for the risk ordering $>_{\bar{\mathbf{R}}}$ on $\mathbf{P}$ such

that for all $(\bar{X}, X'), (\bar{Y}, Y') \in \mathbf{P}$, $(\bar{X}, X') >_{\bar{\mathbf{R}}} (\bar{Y}, Y')$ if and only if $R_p(\bar{X}, X') > R_p(\bar{Y}, Y')$, where R_p is defined as follows:

$$R_p(\bar{X}, X') = E[U_R(\bar{X}, X')] \quad (18)$$

and U_R is a real-valued function unique up to a positive linear transformation. Note that because the marginal distribution for the first attribute is degenerate, the expectation, in fact, only needs to be taken over the marginal distribution for the second attribute, which in turn is the original distribution of a lottery X normalized to a mean of zero.

Basic Forms of Perceived Risk Models

Model (18) provides a general measure of perceived risk based on two attributes, the mean and standard risk of a lottery. In order to obtain separable forms of perceived risk models, we make the following assumptions about risk judgments:

Assumption 1. For $X', Y' \in \mathbf{P}^0$, if there exists a $w^0 \in \text{Re}$ for which $(w^0, X') >_{\bar{\mathbf{R}}} (w^0, Y')$, then $(w, X') >_{\bar{\mathbf{R}}} (w, Y')$ for all $w \in \text{Re}$.

Assumption 2. For $X', Y' \in \mathbf{P}^0$, $(0, X') >_{\bar{\mathbf{R}}} (0, Y')$, if and only if $X' >_{\mathbf{R}} Y'$.

Assumption 3. For $(\bar{X}, X') \in \mathbf{P}$, then $(\bar{X}, X') >_{\bar{\mathbf{R}}} (\bar{X} + \Delta, X')$ for any constant $\Delta > 0$.

Assumption 1 is an independence condition, which says that the risk ordering for two lotteries with the same mean will not switch when the common mean changes to any other value. Compared with Pollatsek and Tversky’s [4] additive independence condition, Assumption 1 is weaker since it features a pair of lotteries with the same mean and a common constant rather than a common lottery. Coombs [8] considered a similar assumption for a riskiness ordering; that is, $X \geq_{\mathbf{R}} Y$, if and only if $X + k \geq_{\mathbf{R}} Y + k$, where $E(X) = E(Y)$ and k is a constant. However, our formulation is based on a two-attribute structure, which leads to a separable risk function for $\bar{X}$ and X' , as we shall discuss.

Assumption 2 postulates a relationship between the two risky binary relations, $>_{\bar{\mathbf{R}}}$

and $>_R$ (where $>_R$ is a strict risk relation on P^0), so that for any zero expected lotteries, the risk judgments made by $R_p(0, X')$ and by the standard measure of risk $R(X')$ are consistent. The last assumption implies that if two lotteries have the same “pure” risk, X' , then the lottery with a larger mean will be perceived less risky than the one with a lower mean as suggested by previous studies (e.g., Refs 9 and 12).

Result 1. The two-attribute perceived risk model of Equation (18) can be decomposed into the following form [38]:

$$R_p(\bar{X}, X') = g(\bar{X}) + \psi(\bar{X})R(X'), \quad (19)$$

if and only if Assumptions 1–3 are satisfied, where $\psi(\bar{X}) > 0$, $g'(\bar{X}) < -\psi'(\bar{X})R(X')$, and $R(X')$ is the standard measure of risk.

According to this result, perceived risk can be constructed by a combination of the standard measure of risk and the effect of the mean. Result 1 postulates a constraint on the choice of functions $g(\bar{X})$ and $\psi(\bar{X})$ in Equation (19). If $\psi(\bar{X})$ is a constant, then the condition $g'(\bar{X}) < -\psi'(\bar{X})R(X')$ becomes $g'(\bar{X}) < 0$; that is, $g(\bar{X})$ is a decreasing function of $\bar{X}$. Otherwise, a nonincreasing function $g(\bar{X})$ and a decreasing function $\psi(\bar{X})$ should be sufficient to satisfy the condition $g'(\bar{X}) < -\psi'(\bar{X})R(X')$ when $R(X') > 0$.

For risk judgments, we may require that any degenerate lottery should have no risk (e.g., Refs 39–41). The concept of risk would not be evoked under conditions of certainty; no matter how bad a certain loss may be, it is a sure thing and, therefore, riskless. This point of view can be represented by the following assumption.

Assumption 4. For any $w \in \text{Re}$, $(w, 0) \sim_{\tilde{R}} (0, 0)$.

Result 2. The two-attribute perceived risk model of Equation (19) can be represented as follows [38]:

$$R_p(\bar{X}, X') = \psi(\bar{X})[R(X') - R(0)] \quad (20)$$

if and only if Assumptions 1–4 are satisfied, where $\psi(\bar{X}) > 0$ is a decreasing function of the mean $\bar{X}$, $R(X')$ is the standard measure of risk, and $R(0) = -u(0)$ is a constant.

When $g(\bar{X}) = -R(0)\psi(\bar{X})$ as required by Assumption 4, the general risk model of Equation (19) reduces to the multiplicative risk model of Equation (20). This multiplicative risk model captures the effect of the mean on perceived riskiness in an appealing way; increasing the mean reduces perceived riskiness in a proportional manner.

Finally, note that the two-attribute perceived risk models (Eqs 19 and 20) are not simple expected forms; we decompose a lottery into a two-attribute structure and only assume the expectation principle holds for normalized lotteries with zero expected values. For general lotteries with nonzero expected values, the underlying probabilities of lotteries can influence the standard measure of risk in a nonlinear fashion via $g(\bar{X})$, $\psi(\bar{X})$. Thus, models of Equations (19) and (20) avoid the drawbacks of expected risk models because the two-attribute expected utility axioms will not generally result in linearity in probability in the perceived risk models.

Relationship between Perceived Risk and Preference

An important feature of the two-attribute approach to modeling risk is that the derived measures of perceived risk can be treated as a stand alone primitive concept *and* can also be incorporated into preference models in a clear fashion. As summarized in the complimentary article, we proposed a risk-value theory for preference modeling also by decomposing a lottery into a two-attribute structure, the mean of the lottery and its standard risk.

A general form of the risk-value model can be represented as follows:

$$f(\bar{X}, X') = V(\bar{X}) - \phi(\bar{X})[R(X') - R(0)] \quad (21)$$

where $f(\bar{X}, X')$ represents a preference function based on the mean of a lottery and its standard risk, $V(\bar{X})$ is a subjective value measure for the mean of a lottery, $\phi(\bar{X}) > 0$ is

a trade-off factor that may depend on the mean, and the other notations are the same as in Equation (20). In general, a decreasing trade-off factor $\phi(\bar{X})$ is required in risk-value theory, which implies that the intensity of the risk effect on preference decreases as a positive constant amount is added to all outcomes of a lottery. Since the risk-value model (Eq. 21) is based on the two attribute expected utility axioms and the perceived risk model (Eq. 19) is derived by using the reverse interpretation of the same axioms, the two types of models must be a negative linear transformation of each other, that is, $f(\bar{X}, X') = -a R_p(\bar{X}, X') + b$, where $a > 0$ and b are constants. Several previously proposed measures of perceived risk also have the implication that their converse forms may be used for preference modeling (e.g., [4, 27, 30]).

The relationships between the functions in the models (Eqs (19) and (21)) can be clarified by transforming the perceived risk model (Eq. 19) into another representation similar to the risk-value model (Eq. 21). When $X = \bar{X}$, Equation (19) becomes $R_p(\bar{X}, 0) = g(\bar{X}) + \psi(\bar{X})R(0)$. Let $h(\bar{X}) = R_p(\bar{X}, 0)$, then $g(\bar{X}) = h(\bar{X}) - \psi(\bar{X})R(0)$. Substituting this into Equation (19), we obtain an alternative representation of the perceived risk measure, $R_p(\bar{X}, X') = h(\bar{X}) + \psi(\bar{X})[R(X') - R(0)]$. Based on our risk-value theory (Result 1), we can have $h(\bar{X}) = -aV(\bar{X}) + b$, and $\psi(\bar{X}) = a\phi(\bar{X})$, where $a > 0$ and b are constants.

The measure of perceived risk (Eq. 20) has more intuitive appeal in constructing preference based on risk-value trade-offs. Substituting $\phi(\bar{X}) = (1/a)\psi(\bar{X})$ into Equation (21), we have $f(\bar{X}, X') = V(\bar{X}) - (1/a)\psi(\bar{X})[R(X') - R(0)] = V(\bar{X}) - (1/a)R_p(\bar{X}, X')$. This representation is consistent with an explicit trade off between perceived risk and value in risky decision making. This provides a clear link between a riskiness ordering and a preference ordering, and shows an explicit role of risk perceptions in decision making under risk.

Some Examples

When $g(\bar{X})$ is linear, $\psi(\bar{X})$ is constant and $R(X')$ is variance, the perceived risk model (Eq. 19) reduces to Pollatsek and Tversky's

[4] mean-variance model as a special case. But Pollatsek and Tversky's risk model may be considered oversimplified. To obtain Rotar and Sholomitsky's [25] generalized moments model, the standard measure of risk should be based on a polynomial utility model.

We can select some appropriate functional forms for $g(\bar{X})$, $\psi(\bar{X})$ and $R(X')$ to construct specific instances of Equation (19). In our complementary article, we have proposed some explicit models for the standard measure of risk $R(X')$. Those models can be used directly in constructing functional forms of perceived risk models (Eqs 19 and 20). An example for $\psi(\bar{X})$ is $\psi(\bar{X}) = ke^{-b\bar{X}}$, where $k > 0$ and $b \geq 0$ (when $b = 0$, $\psi(\bar{X})$ becomes a constant k), and a simple choice for $g(\bar{X})$ is $g(\bar{X}) = -a\bar{X}$, where $a > 0$ is a constant. Some functional forms of the perceived risk model (Eq. 19) based on these choices of $\psi(\bar{X})$ and $g(\bar{X})$ are the following:

$$R_p(\bar{X}, X') = -a\bar{X} + ke^{-b\bar{X}}E[e^{-c(X-\bar{X})}], \quad (22)$$

$$R_p(\bar{X}, X') = -a\bar{X} + ke^{-b\bar{X}}\{E[(X-\bar{X})^2] - cE[(X-\bar{X})^3]\}, \quad (23)$$

$$R_p(\bar{X}, X') = -a\bar{X} + e^{-b\bar{X}}\{dE^-[|X-\bar{X}|^{\theta_2}] - cE^+[|X-\bar{X}|^{\theta_1}]\}, \quad (24)$$

where $a, b, c, d, e, k, \theta_1$ and θ_2 are constants, $E^-[|X-\bar{X}|^{\theta_2}] = \sum_{x_i < \bar{X}} p_i |x_i - \bar{X}|^{\theta_2}$, $E^+[|X-\bar{X}|^{\theta_1}] = \sum_{x_i > \bar{X}} p_i |x_i - \bar{X}|^{\theta_1}$, and p_i is the probability associated with the outcome x_i . When $b = 0$, these perceived risk models become additive forms. For consistency with their corresponding risk-value models, we refer to Equation (22) as the exponential risk model, Equation (23) as the moments risk model, and Equation (24) as the disappointment risk model. The latter was introduced by Bell [42] and explored in more detail by Jia *et al.* [43].

Similarly, some examples of the multiplicative form of risk model (Eq. 20) are given as follows:

$$R_p(\bar{X}, X') = ke^{-b\bar{X}}E[e^{-c(X-\bar{X})} - 1], \quad (25)$$

$$R_p(\bar{X}, X') = ke^{-b\bar{X}}\{E[(X-\bar{X})^2] - cE[(X-\bar{X})^3]\}, \quad (26)$$

$$R_p(\bar{X}, X') = e^{-b\bar{X}} \{dE^- [|X - \bar{X}|^{\theta_2}] - eE^+ [|X - \bar{X}|^{\theta_1}]\}. \quad (27)$$

Research on financial risk and psychological risk (i.e., perceived risk) has been conducted separately in the past. The risk-value framework is able to provide a unified approach for dealing with the both types of risk. The standard measure of risk is more normative in nature and should be useful in financial modeling. For instance, the standard measure of risk in perceived risk models (Eqs 24 and 27) includes many financial risk measures as special cases [2]. Our perceived risk models show how financial measures of risk and psychological measures of risk can be related. In particular, for a given level of the mean value, minimizing the perceived risk will be equivalent to minimizing the standard risk since the expressions for $g(\bar{X})$ and $\psi(\bar{X})$ in Equation (19) become constants. Our measures of perceived risk provide a clear way to simplify the decision criterion of minimizing perceived risk as suggested, but never operationalized, in Coombs' Portfolio theory.

CONCLUSIONS

In this article, we have reviewed previous studies about perceived risk and focused on a two-attribute structure for perceived risk based on the mean of a lottery and its standard risk. Some of these risk measures also take into account the asymmetric effects of losses and gains on perceived risk. These measures of perceived risk can unify a large body of empirical evidence about risk judgments, and are consistent with the stylized facts regarding risk judgments listed in the introduction. For more details regarding the flexibility provided by the two-attribute structure for perceived risk see Jia *et al.* [38]; for details on the empirical validity of the assumptions behind the models, see Butler, Dyer and Jia [44].

In particular, these measures of perceived risk show a clear relationship between financial measures of risk and psychological measures of risk. They can also be incorporated into preference models in a natural way, based on a trade-off between perceived risk

and expected value. This shows an intuitively appealing connection between perceived risk and preference.

This development uses the expected value of a lottery as the reference point regarding the measures of perceived risk. The expected value is a convenient and probabilistically appealing reference point [2], which makes our risk models mathematically tractable and practically usable. There are other possible reference points that might be considered, such as an aspiration level, a reference lottery, or some other external reference point, such as zero. It would be interesting to consider these alternative reference points in our measures of perceived risk in future research.

Acknowledgment

This article summarizes a stream of research on perceived risk. In particular, we have incorporated materials that appeared previously in Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Management Science* 1999 45: 519–532

REFERENCES

1. Jia J, Dyer JS. Risk-value theory. Working Paper. University of Texas at Austin (TX): Graduate School of Business; 1995.
2. Jia J, Dyer JS. A standard measure of risk and risk-value models. *Manage Sci* 1996;42:1961–1705.
3. Dyer JS, Jia J. Preference conditions for utility models: A risk-value perspective. *Ann Oper Res* 1998;80:167–182.
4. Pollatsek A, Tversky A. A theory of risk. *J Math Psychol* 1970;7:540–553.
5. Coombs CH. Portfolio theory: A theory of risky decision making. Paris: Centre National de la Recherche Scientifique; 1969.
6. Coombs CH, Meyer DE. Risk preference in coin-toss games. *J Math Psychol* 1969;6:514–527.
7. Coombs CH, Huang L. Tests of a portfolio theory of risk preference. *J Exp Psychol* 1970;85:23–29.
8. Coombs CH. Portfolio theory and the measurement of risk. In: Kaplan MF, Schwartz S, editors. *Human judgment and decision processes*. New York: Academic Press, Inc.; 1975. pp. 63–85.

9. Keller LR, Sarin RK, Weber M. Empirical investigation of some properties of the perceived riskiness of gambles. *Organ Behav Hum Decis Process* 1986;38:114–130.
10. Weber EU, Bottom WP. An empirical evaluation of the transitivity, monotonicity, accounting, and conjoint axioms for perceived risk. *Organ Behav Hum Decis Process* 1990;45:253–275.
11. Coombs CH, Huang L. Polynomial psychophysics of risk. *J Math Psychol* 1970;7:317–338.
12. Coombs CH, Lehner PE. An evaluation of two alternative models for a theory of risk: Part 1. *J Exp Psychol Hum Percept Perform* 1981;7:1110–1123.
13. Brachinger HW, Weber M. Risk as a primitive: a survey of measures of perceived risk. *OR Spektrum* 1997;19:235–250.
14. Artzner P, Delbaen F, Eber J-M, *et al.* Thinking coherently. *Risk* 1997;10:68–71.
15. Artzner P, Delbaen F, Eber J-M, *et al.* Coherent measures of risk. *Math Finance* 1999;9:203–228.
16. Acerbi C. Coherent representations of subjective risk-aversion. In: Szego G, editor. *Risk measures for the 21st century*. New York: John Wiley & Sons, Inc.; 2004. pp. 147–206.
17. Coombs CH, Pruitt DG. Components of risk in decision making: Probability and variance preferences. *J Exp Psychol* 1960;60:265–277.
18. Barron FH. Polynomial psychophysics of risk for selected business faculty. *Acta Psychol* 1976;40:127–137.
19. Coombs CH, Lehner PE. Conjoint design and analysis of the bilinear model: An application to judgments of risk. *J Math Psychol* 1984;28:1–42.
20. Kahneman DH, Tversky A. Prospect theory: An analysis of decision under risk. *Econometrica* 1979;47:263–290.
21. Coombs CH, Bowen JN. A test of VE-theories of risk and the effect of the central limit theorem. *Acta Psychol* 1971;35:15–28.
22. Fishburn PC. Foundations of risk measurement. In: *Encyclopedia of statistical sciences*. Volume 8, New York: John Wiley & Sons; 1988. pp. 148–152.
23. Coombs CH, Bowen JN. Additivity of risk in portfolios. *Percept Psychophys* 1971;10:43–46.
24. Nygren T. The relationship between the perceived risk and attractiveness of gambles: A multidimensional analysis. *Appl Psychol Measure* 1977;1:565–579.
25. Rotar IV, Sholomitsky AG. On the Pollatsek-Tversky theorem on risk. *J Math Psychol* 1994;38:322–334.
26. Luce RD. Several possible measures of risk. *Theory Decis* 1980;12:217–228 Correction 1981;13:381.
27. Sarin RK. Some extensions of Luce's measures of risk. *Theory Decis* 1987;22:25–141.
28. Huang LC. The expected risk function. Michigan Mathematical Psychology Program Report 71-6: University of Michigan, Ann Arbor, MI, 1971.
29. Luce RD, Weber EU. An axiomatic theory of conjoint, expected risk. *J Math Psychol* 1986;30:188–205.
30. Machina M. Expected utility analysis without the independence axiom. *Econometrica* 1982;50:277–323.
31. Weber EU. A descriptive measure of risk. *Acta Psychol* 1988;69:185–203.
32. Fishburn PC. Foundations of risk measurement, I, risk as probable loss. *Manage Sci* 1984;30:296–406.
33. Fishburn PC. Foundations of risk measurement, II, effects of gains on risk. *J Math Psychol* 1982;25:226–242.
34. Payne WJ. Alternative approaches to decision making under risk: Moments versus risk dimensions. *Psychol Bull* 1973;80:439–453.
35. von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton (NJ): Princeton University Press; 1947.
36. Fishburn PC. *Utility theory for decision making*. New York: Wiley; 1970.
37. Keeney RL, Raiffa H. *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley; 1976.
38. Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Manage Sci* 1999; 45: 519–532.
39. Bell DE. One-switch utility functions and a measure of risk. *Manage Sci* 1988;34: 1416–1424.
40. Bell DE. Risk, return, and utility. *Manage Sci* 1995;41:23–30.
41. Sarin RK, Weber M. Risk-value models. *Eur J Oper Res* 1993;70:135–149.
42. Bell DE. Disappointment in decision making under uncertainty. *Oper Res* 1985;33:1–27.
43. Jia J, Dyer JS, Butler JC. Generalized disappointment models. *J Risk Uncertain* 2001; 22:59–78.
44. Butler J, Dyer J, Jia J. An empirical investigation of the assumptions of risk-value models. *J Risk Uncertain* 2005;30:133–156.

BACKTRACK SEARCH TECHNIQUES AND HEURISTICS

CHRISTOPHE LECOUTRE
CRIL-CNRS Université
Lille-Nord de France,
Lens, France

Satisfying a set of constraints is known as an instance of the constraint satisfaction problem (CSP), which is the subject of intense research in both artificial intelligence and operations research. Practical solution of CSP instances usually involves *backtrack search*. This is a *complete* approach in which systematic exploration of the search space of an instance finds the full set of solutions or proves that no solution exists. By contrast, *incomplete* approaches, such as those based on local search, are not guaranteed to find a solution or to prove unsatisfiability. Unfortunately, backtrack search is not expected to terminate within polynomial time in the general case, unless $P = NP$. This is why there have been considerable efforts during the last three decades to maximize the practical efficiency of backtrack search. In particular, the development of adaptive heuristics has shown impressive progress in guiding search.

GENERAL DESCRIPTION

A (finite) constraint network (CN), P [1,2] is composed of a finite set of variables, denoted by $vars(P)$, and a finite set of constraints, denoted by $cons(P)$. Each variable x has an associated domain, denoted by $dom(x)$, that contains the finite set of (current) values that can be assigned to x . Each constraint c involves an ordered set of variables, called the *scope* of c . A constraint is defined by a relation, denoted by $rel(c)$, which contains the set of tuples allowed for the variables involved in c . A CN is said to be *satisfiable* iff it admits at least one solution, that is, there is an assignment of a value to every variable such that every constraint is satisfied.

An instance of the CSP, is defined by a CN, which is solved either by finding a solution or by proving unsatisfiability.

For example, the classical queens problem can be stated as follows: can we put n queens on a board of size $n \times n$ such that no two queens attack each other? Two queens attack each other iff they belong to the same row, the same column or the same diagonal. A natural CSP model of this problem involves the introduction of a variable per queen (attached to a column), whose domain contains row numbers. If the i th variable x_i is assigned the value j , it means that the i th queen is put in the square at the intersection of the i th column and the j th row. For the n -queens instance, we have a CN P such that $vars(P) = \{x_1, \dots, x_n\}$ with $dom(x_i) = \{1, \dots, n\}$, $\forall i \in 1 \dots n$, and $cons(P) = \{x_i \neq x_j \wedge |x_i - x_j| \neq |i - j| : i \in 1 \dots n, j \in 1 \dots n, i < j\}$ because we need a binary constraint per pair of variables (queens). Figure 1 shows the two solutions for the 4-queens instance. The first one is given by $\{x_a = 2, x_b = 4, x_c = 1, x_d = 3\}$ where here, variables are denoted by x_a, x_b, x_c , and x_d (instead of x_1, x_2, x_3 , and x_4), to clarify the correspondence with columns.

Within backtrack search [3–5], depth-first exploration instantiates variables and a backtracking mechanism deals with dead-ends. The depth-first search considers a different variable at each level and tries to extend (in turn) different complementary branching decisions concerning this variable. In its simplest form, each branching decision is an assignment of a value to a variable; this is followed by checking that every constraint covered by the current instantiation is satisfied individually. A more sophisticated strategy applies a filtering procedure after each assignment of a value to a variable; this procedure is intended to simplify the subsequent search or to show that a dead-end has been reached, which means that the current set of decisions cannot be extended to a solution. When a dead-end is encountered, one or more decisions must be retracted before continuing the quest for a solution. The process

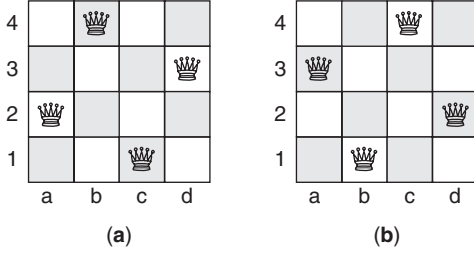


Figure 1. Solutions of the 4-queens instance. (a) First solution. (b) Second solution.

of undoing decisions in order to escape from a dead-end is called *backtracking*.

Backtrack search systems have four main components: *branching* (how and which decisions to take to go forward to a solution?), *propagation* (how and which level of filtering to apply to reduce the search space at each step?), *backtracking* (how to go backward when a dead-end is encountered?) and *learning* (what information to collect during search so as to facilitate subsequent parts of the search?). Each of these components has many possible implementations; there has been much effort to identify the right combinations of implementations. In particular, the interplay between propagation and backtracking techniques has been debated for a long time.

SEARCH TREES AND BRANCHING DECISIONS

Backtrack search algorithms build *search trees*. A search tree is basically a rooted tree that allows us to visualize successive decisions performed by a backtrack search algorithm. Starting at the root node with the initial CN that must be solved, each step in the search derives a new CN. Each node in the search tree is associated with one such CN and each (directed) edge is associated with a search decision. The search tree grows during the search. More specifically, if the search has currently reached node v , then after taking a new (branching) decision δ , we insert into the search tree a new node v' , representing the new step of the search, and a new edge $\{v, v'\}$ labeled with δ . The

new edge $\{v, v'\}$ is directed from v to v' ; v is called the *parent* of v' and v' a *child* of v . $dn(v)$ is the set of decisions that label successive edges in the path from the root to node v . The CN associated with node v is $cn(v) = \phi(P^{\text{init}}|_{dn(v)})$, where P^{init} is the initial CN $P^{\text{init}}|_{dn(v)}$ is P^{init} modified after taking into account all decisions present in $dn(v)$, and ϕ corresponds to the filtering performed during search (i.e., after each taken decision). Typically, ϕ is a property called *local consistency* that allows us to remove some *inconsistent* values; an inconsistent value cannot lead to any solution. If $P' = \phi(P)$ then P' is said to be obtained after enforcing ϕ on P .

Figure 2 provides an example of a search tree. Every node in a search tree is either a *leaf* node or an *internal* node. A leaf node differs from an internal node in that a leaf node has no children. When an inconsistency is detected (typically because a domain becomes empty, which prevents us from reaching a solution) at node v during search, this is denoted by $cn(v) = \perp$ and v is a leaf node called a *dead-end*. The search backtracks when a dead-end is reached. Any node v in the search tree is the root of a subtree obtained by retaining only v and its descendants (with all related edges). A node v is fully explored when the search space of the CN $cn(v)$ has been fully explored. If v is a fully explored internal node such that $cn(v)$ is unsatisfiable, v is called an *internal dead-end*, and the subtree rooted at v is called a *refutation tree* of $cn(v)$. If v' is an internal dead-end, and is a child of node v such that $cn(v)$ is satisfiable, then v' is a *mistake node*. A (sub)tree containing no solution is said to be *fruitless*.

Branching decisions, also called *branching constraints*, split a CN $P = cn(v)$ associated with an internal node v of the search tree into two or more CNs, the union of which is equivalent to P in terms of solutions. Classical branching schemes impose decisions during search under a strategy of *enumeration* or *labeling* [6]. Enumeration and labeling, respectively, correspond to *binary branching* (or *two-way branching*) and *nonbinary branching* (or *d-way branching*). More specifically, with nonbinary branching, at each internal node v , an unfixed variable x (i.e., variable whose domain is not singleton) is

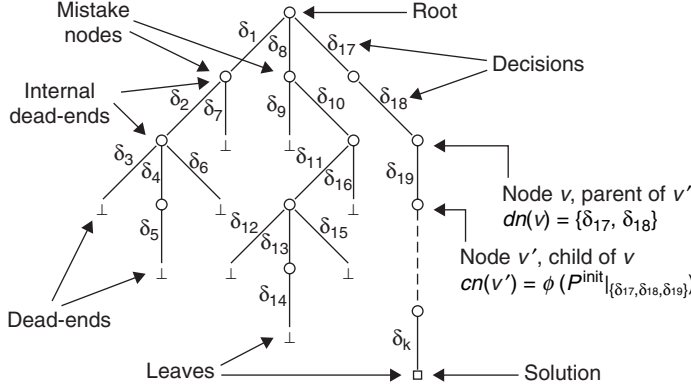


Figure 2. A search tree built by a backtrack search algorithm. Note that decisions are taken depth-first. P^{init} is the instance to be solved initially and ϕ corresponds to the filtering performed during search.

selected, and then for each value a in $\text{dom}(x)$, the assignment $x = a$ is considered, so there are altogether d branches leaving v , where d is the size of $\text{dom}(x)$ at node v . With binary branching, at each internal node v , a pair (x, a) is selected where x is an unfixed variable and a a value in $\text{dom}(x)$, and two cases are considered: the assignment $x = a$ and the refutation $x \neq a$, so there are exactly two branches leaving v . Both of these two schemes guarantee complete exploration of the search space.

When mentioning a backtrack search algorithm, it is important to indicate whether binary branching or nonbinary branching is employed. These two schemes are not equivalent; it has been shown that binary branching is more powerful (to refute unsatisfiable instances) than nonbinary branching [7]. Using the resolution proof system, Hwang and Mitchell show that there exist instances which require exponential search trees for backtracking with d -way branching, but have polynomial search trees for backtracking with two-way branching. Although various other kinds of decision (e.g., membership decisions when splitting domains or nonunary branching constraints) are possible, these are not commonly used in the solution of discrete CSP instances. Recent exceptions are effective use of *partitioning* [8] and *bundling* [9,10], where complementary decisions of the form $x \in D_x$ with $D_x \subset \text{dom}(x)$ are taken.

BACKTRACK SEARCH ALGORITHMS

A (backtrack) ϕ -search algorithm is a backtrack search algorithm that enforces a local consistency ϕ after each decision taken. Algorithm 1 is a general formulation of any backtrack ϕ -search algorithm that employs binary branching. This quite reasonably assumes that ϕ gives a level of consistency which (at least) allows detection of any unsatisfied constraint whose variables are all fixed. For example, if P contains the constraint $x = y$, then after decisions $x = a$ and $y = b$ are taken, ϕ enforced on $P|_{\{x=a, y=b\}}$ will detect an inconsistency (since $a \neq b$). This algorithm starts by enforcing ϕ on the given CN P (which is an input parameter), returning *false* if an inconsistency is detected. If all domains are single valued at line 4 then because of our assumption about ϕ , all constraints are necessarily satisfied, so a solution has been found, and *true* is returned. Each step of the binary branching algorithm selects a pair (x, a) and recursively calls the function *binary- ϕ -search* with decisions $x = a$ and $x \neq a$. Depending on the implementation of the logical operator $\vee$ at line 7, the algorithm finds a single solution (if $\vee$ is managed in short-circuit, that is to say, if $\vee$ does not evaluate the right operand when the left one evaluates to *true*) or finds all solutions (if any).

Algorithm 1. $\text{binary-}\phi\text{-search}(\text{in } P: \text{Constraint Network}): \text{Boolean}$

Output: *true* iff P is satisfiable

```

1  $P \leftarrow \phi(P)$ 
2 if  $P = \perp$  then
3   return false
4 if  $\forall x \in \text{vars}(P), |\text{dom}(x)| = 1$  then
5   // display the solution
6   return true
7 select a pair  $(x, a)$  from  $P$  such that  $|\text{dom}(x)| > 1$ 
8 return  $\text{binary-}\phi\text{-search}(P|_{[x=a]}) \vee \text{binary-}\phi\text{-search}(P|_{[x \neq a]})$ 

```

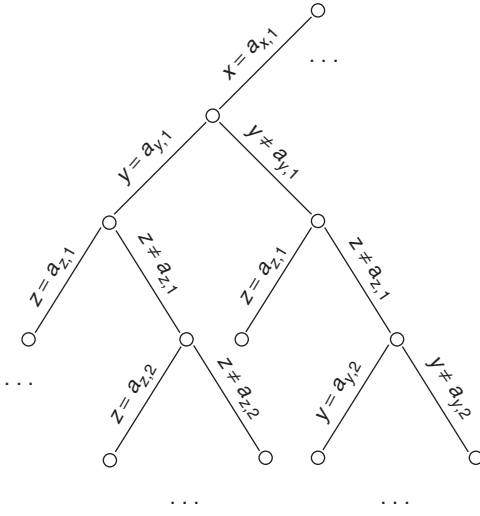


Figure 3. A binary search tree built by a backtracking algorithm. The i th value in the current domain of a variable x is denoted by $a_{x,i}$.

Figure 3 illustrates the binary branching process and shows the systematic exploration of two branches for each selected pair (x, a) . Classically, binary branching algorithms select left branches, which assign values, before right branches, which refute values. The main advantage of binary branching over nonbinary branching is the possibility of selecting, after each refutation, a variable different from the one involved in the last decision. For this reason, heuristics that control the search may be more reactive.

The example in Fig. 3 assumes that an unsatisfiable subtree is explored after assignments $x = a_{x,1}$, $y = a_{y,1}$, and $z = a_{z,1}$. After the refutation $z \neq a_{z,1}$, the next selected variable is again z , making this portion of the search

tree rather similar to one that could be built with nonbinary branching. However, along the path labeled with $x = a_{x,1}$ and $y \neq a_{y,1}$, the next branching decisions involve the variable z which is different from y . Here the search heuristic, perhaps learning from previous explorations of subtrees, has decided to branch on a different variable rather than insisting on y .

Finally, note that CNs are usually processed during a so-called *preprocessing* stage initiated before search. Preprocessing may be limited to enforcing a consistency ϕ , but it may also refer to more sophisticated methods such as those based on structural decomposition [11]. During the preprocessing stage, some data structures may also have to be initialized so as to be used later during search by some algorithms. Sometimes, preprocessing alone is sufficient to solve a CSP instance.

CONSTRAINT PROPAGATION

After each branching decision, a filtering process is run by enforcement of a local consistency that prunes some parts of the search space containing no solution. For example, ϕ as mentioned above, may correspond to (generalized) arc consistency (AC) [12]. This is the strongest form of local reasoning when constraints are considered independently. Intuitively, AC allows us to safely remove a value a for a variable x if there exists a constraint c involving x but accepting no tuple (built from the current domains) with value a for x ; we say that (x, a) has no support on c . For example, if x and y are two variables such that $\text{dom}(x) = \text{dom}(y) = \{1, 2\}$ and we have the constraint $x < y$, then the value 2

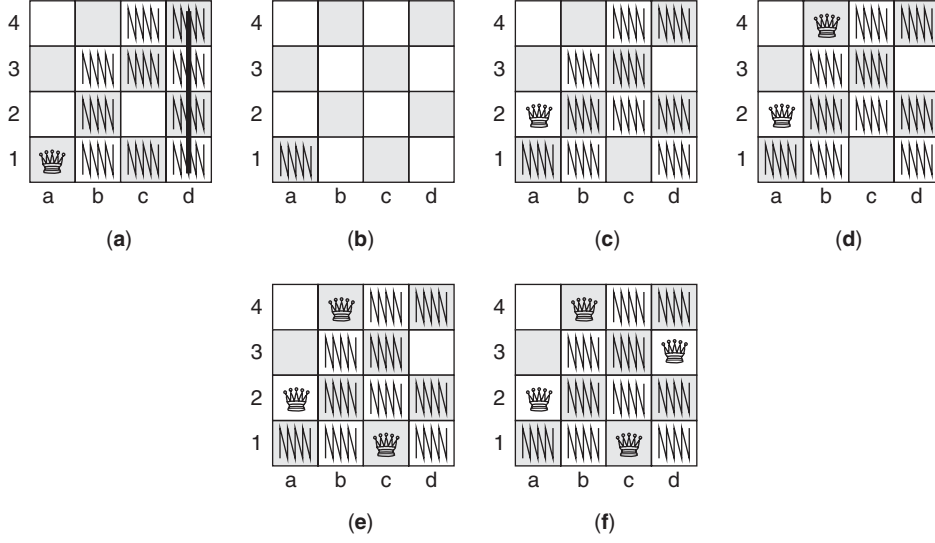


Figure 4. The six steps performed by MAC to solve the 4-queens instance.

for x as well as the value 1 for y can be both removed because they have no support on $x < y$. Interestingly, as soon as a local inference (corresponding to the removal of a value) is performed, the conditions to trigger new inferences may hold since variables are typically shared by several constraints. This mechanism of propagating the results of local inferences from constraints to constraints is called *constraint propagation* [13].

MAC [14–16] is the backtrack search algorithm that maintains (generalized) AC after each decision taken; when the domain of a variable becomes empty (so-called *domain wipe-out*) a dead-end has been reached. Figure 4 shows the search steps performed by MAC to solve the 4-queens instance. Queens (variables) are put (instantiated) in columns from left to right and values are assigned from 1 to 4. Hatched squares represent values removed from variable domains due to constraint propagation. For example, the square at column b and row 1 and the square at column b and row 2 are hatched in Fig. 4a because x_a has just been assigned the value 1 and there is a constraint between the variables x_a and x_b associated with the first two columns defined as: $x_a \neq x_b \wedge |x_a - x_b| \neq 1$. The reason the square at column b and row 3 is also hatched

is due to constraint propagation. Note that all branching decisions are shown. For example, after MAC puts the first queen onto the square of the chessboard at column a and row 1, Fig. 4a, a dead-end is reached and this square is then discarded as shown in Fig. 4b.

BACKTRACKING TECHNIQUES

When a dead-end is reached, conflicting decisions can be reviewed via *eliminating explanations* which are recorded during the search. Instead of backtracking to the most recent previous decision, so-called *chronological backtracking*, it may be helpful to jump back to the most recent decision among those that could possibly have caused the failure. This backward jump (or backjump) is a form of *intelligent backtracking*, also called *backjumping*, and can be managed so as to guarantee that no solution will be missed.

The relationship between *look-back* (efficient escape from dead-ends) and *look-ahead* (simplification of subsequent search), has been the subject of much investigation. MAC was used in the 1970s [14,17] with nonbinary branching, without backjumping and without dynamic variable ordering (presented in the section titled “Search Heuristics”). Nonchronological backtracking

was initiated with dependency-directed backtracking [18–20] and Prolog intelligent backtracking [21]. Early in the 1990s, the forward checking (FC) algorithm (introduced 10 years before [22,23]) associated with the variable ordering heuristic *dom* [23] and the conflict-directed backjumping (CBJ) technique [24] was considered to be the most efficient generic approach to solve CSP instances. In 1994, Sabin and Freuder [16] reintroduced MAC using binary branching and simple chronological backtracking. This algorithm was shown to be more efficient than FC and FC-CBJ; CBJ was considered to be useless to MAC, especially when MAC had a good variable ordering heuristic [25].

The situation subsequently became more confused. First, Bayardo and Shrag [26] showed that many large propositional satisfiability instances derived from real-world problems are easy when CSP look-back techniques are combined with the “Davis–Putnam” procedure [3]. Second, although theoretical results [27] showed that the backward phase is less useful when the forward phase is more advanced, some experiments on hard structured problems showed that combining CBJ with MAC can still produce significant improvements. Third, look-back techniques appeared to be improved by associating an eliminating explanation (or conflict set) with any value rather than with any variable. Indeed, refined eliminating explanations allows a stronger form of backjumping [28] and the possibility of saving much search effort with the principle of dynamic backtracking (DBT) [29]. Experimental results [28,30] showed that MAC can be outperformed by algorithms embedding such advanced look-back techniques. However, the next section will show that some form of learning, perhaps limited to a basic statistical form, is essential to the efficient guidance of search.

The use of eliminating explanations [29], as in CBJ, is now illustrated. An eliminating explanation for a pair (x, a) , denoted by $\text{expl}(x \neq a)$, is a subset of all decisions currently taken by the search algorithm, which explains why the value a has been removed from the domain of x . In Fig. 5, we suppose that we have an eliminating explanation

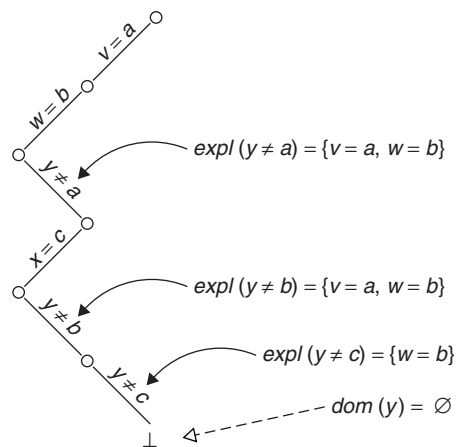


Figure 5. Eliminating explanations computed for values of y . As both branches $y = c$ and $y \neq c$ [c was the last value in $\text{dom}(y)$] have been explored, we have to backtrack.

for each value present in the initial domain of y , and have reached a dead-end. Taking the union of all eliminating explanations for values of y yields a set of decisions, called a *nogood*, that cannot be extended to a solution. Reasoning from this *nogood* allows us to safely backtrack up to $w = b$, the most recent decision in the *nogood*, and refute it globally; see Fig. 6. Further, we can compute an eliminating explanation $\text{expl}(w \neq b)$ from the *nogood*.

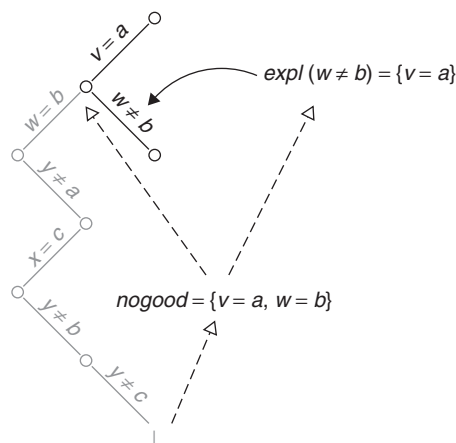


Figure 6. Backtrack guided by the *nogood* extracted from situation of Fig. 5.

SEARCH HEURISTICS

Two choices are made at each step within backtrack search. Before branching, the search algorithm selects first a variable x and then a value a in the current domain of x ; the global selection of a pair (x, a) is also possible but efficient implementations of this approach remains an open challenge. Thus, the search algorithm imposes an ordering on variables and on their values. However, finding an optimal ordering is at least as difficult as solving a CSP instance. This is why, in practice, ordering is determined by *heuristics*. A heuristic is a general guideline rule that is expected to lead to good results, but is not claimed to give optimal outcomes in every situation.

For backtrack search, a first general principle is that it is better to start by assigning variables that belong to the most difficult part(s) of the problem instance. This principle is derived from recognition that there is no point in traversing the easy part(s) of an instance and then backtracking repeatedly when it turns out that the first choices are incompatible with the remaining difficult part(s). Here the underlying *fail-first* principle, according to Haralick and Elliott [23], is: “To succeed, try first where you are most likely to fail.”

Value selection can be based on the *succeed-first* or *promise* principle, which comes from the simple observation that to find a solution quickly, it is better to move at each step to the most promising subtree, primarily by selecting a value that is most likely to participate in a solution. It is preferable to avoid branching on a value that is inconsistent, because this implies exploration of a fruitless subtree, which is clearly a waste of time if there is a solution elsewhere.

Although these two principles are somewhat contradictory (basically because all variables must be given a value whereas only a good value per variable has to be found), variable ordering can to some extent comply with both of them [31–33]. Various different measures of promise of variable ordering heuristics try to assess the ability of the heuristics to avoid making mistakes, that is,

to keep the search on the path to a solution regardless of the value ordering. There appears to be quite a complex relationship between promise and fail-firstness. For value ordering, the extent of adherence to both heuristics can also be assessed; first elements related to this can be found in Szymanek and O’Sullivan [34] and Lecoutre *et al.* [35].

When starting to build the search tree, the initial variable/value choices are particularly important. Bad choices near the root of the search tree may turn out to be disastrous because they lead to exploration of very large fruitless subtrees. To make good initial choices, one strategy is to select the first branching decisions with special care, perhaps calling sophisticated and expensive procedures for this purpose. Another strategy is to restart search several times [36], ideally learning some information each time in order to refine search guidance [37,38].

In what follows, we focus our attention to variable ordering because the order in which variables are assigned by a backtrack search algorithm has been recognized as a key issue for a long time. Using different variable ordering heuristics can drastically affect the efficiency of algorithms solving CSP instances. For example, introducing some form of randomization into a given variable ordering heuristic can cause great variability in performance. *Static* variable ordering heuristics keep the same ordering throughout the search, using only (structural) information about the initial state of search. *Dynamic* variable ordering heuristics take account of the current state of the instance being solved. These heuristics are dynamic because their ordering generally varies during the search. The well-known dynamic heuristic *dom* [5,23] orders variables in sequence of increasing size of domain, so a variable that has the smallest domain size is selected at each step. Variants of *dom* are the heuristics *dom + deg* [39] and *dom/deg* [25] that additionally take degrees of variables into account.

Static and (nonadaptive) dynamic variable ordering heuristics are relatively poor general-purpose heuristics. The remainder of

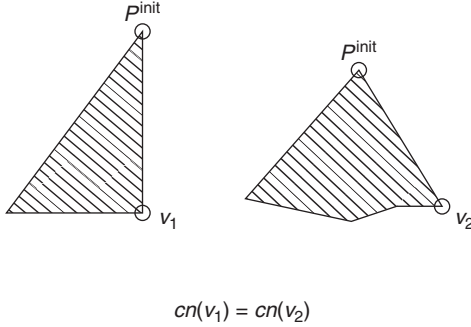


Figure 7. Two partial search trees built from the same instance P^{init} . We have v_1 and v_2 representing similar nodes, that is, nodes such that their associated constraint networks are identical. Unlike a static or dynamic variable ordering heuristic, an adaptive variable ordering heuristic can make a different selection at nodes v_1 and v_2 .

this section presents two adaptive heuristics that can reasonably be considered to be state of the art. A search-guiding heuristic is said to be *adaptive* when it makes choices that depend on the current state of the problem instance as well as previous states. Thus an adaptive heuristic learns, in the sense that it takes account of information concerning the subtree already been explored. Figure 7 illustrates the fact that an adaptive heuristic may behave differently when two similar nodes (i.e., two nodes such that their associated CNs are identical) are reached after exploring different subtrees.

A first family of adaptive heuristics is based on the concept of *impacts*. An impact is a measure of the effect of an assignment. More specifically, an impact is a measure of the relative amount of search space reduction that an assignment is expected to achieve. The impact of a variable is the “sum” of impacts of possible assignments to this variable. Impacts can be used in the heuristic selection of variables and values. The variable with highest impact is typically selected first. For this variable, the value with lowest impact is selected. It is important that impacts can be refined during search, allowing learning from experience. The use of impacts was studied initially by Geelen [40] and has been adaptively revisited by Refalo

[41] inspired by pseudo-costs that are widely used in integer programming.

Another family of adaptive heuristics is based on constraint weighting, which is an efficient mechanism for identifying hard parts of combinatorial problems. It was initially introduced to improve the performance of local search methods and/or Boolean satisfiability (SAT) solving [42–44]. Hybrid search techniques based on dynamic constraint weighting were also proposed in Mazure *et al.* [45] and Eidenberg and Faltings [46]. The principle of constraint weighting is quite simple:

1. associate a counter with each constraint c to denote the *weight* of c ,
2. increment the weight of a constraint c whenever a dead-end occurs due to propagation on c ,
3. at each branching point, select the variable with the highest weighted degree; the weighted degree of a variable x is the sum of the weights of the constraints involving x .

The practical effect of selecting first the variables with greatest weighted degrees is to examine first the locally inconsistent or hard parts of networks, in conformity with the fail-first principle. This variable ordering heuristic is denoted by *wdeg* [47], and a variant is *dom/wdeg*, which selects first the variable having the smallest ratio of current domain size to current weighted degree. These *conflict-directed* heuristics appear to outperform current intelligent backtracking methods [48]. But the robustness of heuristic *dom/wdeg* is best demonstrated by comparing it with classical heuristics on a wide range of problems. Figure 8 shows the results obtained on a large set of instances, including various series of random and structured instances, when using MAC with the adaptive heuristic *dom/wdeg* and with the representative classical heuristic *dom + deg*. In this scatter plot, each dot represents an instance and each axis represents the CPU time required to solve the instances with MAC using the heuristic labeling the axis. Many dots are located on the right side of the scatter plot, which means that *dom/wdeg*

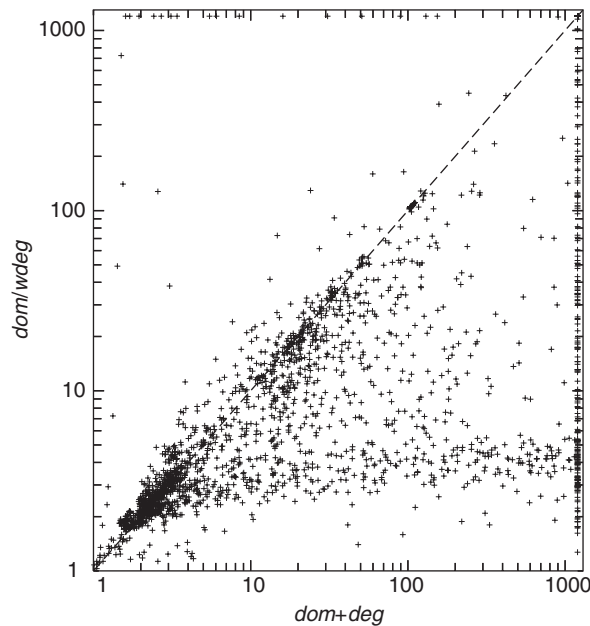


Figure 8. Pairwise comparison (CPU time) of heuristics $dom/wdeg$ and $dom+deg$ when used by MAC to solve the instances chosen as benchmarks of the 2006 constraint solver competition (time out set to 1000 s per instance).

solves far more instances than $dom + deg$ within the allotted time.

REFERENCES

1. Dechter R. Constraint processing. Morgan Kaufmann; 2003.
2. Lecoutre C. Constraint networks: techniques and algorithms. ISTE/Wiley; 2009.
3. Davis M, Logemann G, Loveland DW. A machine program for theorem-proving. Commun ACM 1962;5(7):394–397.
4. Golomb SW, Baumert LD. Backtrack programming. J ACM 1965;12(4):516–524.
5. Bitner JR, Reingold EM. Backtrack programming techniques. Commun ACM 1975;18(11):651–656.
6. Apt KR. Principles of Constraint Programming. Cambridge University Press; 2003.
7. Hwang J, Mitchell DG. 2-way vs d-way branching for CSP. Proceedings of CP'05. Sitges, Spain; 2005. pp. 343–357.
8. van Hoeve WJ, Milano M. Postponing branching decisions. Proceedings of ECAI'04. Valencia, Spain; 2004. pp. 1105–1106.
9. Haselbock A. Exploiting interchangeabilities in constraint satisfaction problems. Proceedings of IJCAI'93. Chambéry, France; 1993. pp. 282–287.
10. Choueiry BY, Davis AM. Dynamic bundling: less effort for more solutions. Proceedings of SARA'02. Kananaskis, Alberta; 2002. pp. 64–82.
11. Gottlob G, Leone N, Scarcello F. A comparison of structural CSP decomposition methods. Artif Intell 2000;124:243–282.
12. Mackworth AK. Consistency in networks of relations. Artif Intell 1977;8(1):99–118.
13. Bessiere C. Constraint propagation. In: Handbook of constraint programming, chapter 3. Elsevier; 2006.
14. Gaschnig J. A constraint satisfaction method for inference making. Proceedings of the 12th Annual Allerton Conference on Circuit and System Theory. Monticello (IL); 1974. pp. 866–874.
15. Nadel BA. Tree search and arc consistency in constraint satisfaction algorithms. In: Kanal L, Kumar V, editors. Search in artificial intelligence. New York: Springer; 1988. pp. 287–342.
16. Sabin D, Freuder EC. Contradicting conventional wisdom in constraint satisfaction. Proceedings of CP'94. Rosario (WA); 1994. pp. 10–20.
17. Ullmann JR. An algorithm for subgraph isomorphism. J ACM 1976;23(1):31–42.
18. Stallman RM, Sussman GJ. Forward reasoning and dependency directed backtracking in

- a system for computer-aided circuit analysis. *Artif Intell* 1977;9:135–196.
19. Doyle J. A truth maintenance system. *Artif Intell* 1979;12(3):231–272.
 20. de Kleer J. An assumption-based TMS. *Artif Intell* 1986;28(2):127–162.
 21. Bruynooghe M. Solving combinatorial search problems by intelligent backtracking. *Inform Process Lett* 1981;12(1):36–39.
 22. McGregor JJ. Relational consistency algorithms and their application in finding subgraph and graph isomorphisms. *Inform Sci* 1979;19:229–250.
 23. Haralick RM, Elliott GL. Increasing tree search efficiency for constraint satisfaction problems. *Artif Intell* 1980;14:263–313.
 24. Prosser P. Hybrid algorithms for the constraint satisfaction problems. *Comput Intell* 1993;9(3):268–299.
 25. Bessiere C, Régin J. MAC and combined heuristics: two reasons to forsake FC (and CBJ?) on hard problems. *Proceedings of CP'96*. Cambridge (MA): 1996. pp. 61–75.
 26. Bayardo RJ, Shrag RC. Using CSP look-back techniques to solve real-world SAT instances. *Proceedings of AAAI'97*. Providence (RI): 1997. pp. 203–208.
 27. Chen X, van Beek P. Conflict-directed backjumping revisited. *J Artif Intell Res* 2001;14:53–81.
 28. Bacchus F. Extending forward checking. *Proceedings of CP'00*. Singapore: 2000. pp. 35–51.
 29. Ginsberg ML. Dynamic backtracking. *J Artif Intell Res* 1993;1:25–46.
 30. Jussien N, Debruyne R, Boizumault P. Maintaining arc-consistency within dynamic backtracking. *Proceedings of CP'00*. Singapore: 2000. pp. 249–261.
 31. Beck JC, Prosser P, Wallace RJ. Variable ordering heuristics show promise. *Proceedings of CP'04*. Toronto, Canada: 2004. pp. 711–715.
 32. Wallace RJ. Heuristic policy analysis and efficiency assessment in constraint satisfaction search. *Proceedings of ICTAI'06*. Washington (DC): 2006. pp. 305–314.
 33. Hulubei T, O'Sullivan B. Failure analysis in backtrack search for constraint satisfaction. *Proceedings of CP'06*. Nantes, France: 2006. pp. 731–735.
 34. Szymanek R, O'Sullivan B. Guiding search using constraint-level advice. *Proceedings of ECAI'06*. Riva del Garda, Italy: 2006. pp. 158–162.
 35. Lecoutre C, Sais L, Vion J. Using SAT encodings to derive CSP value ordering heuristics. *Proceedings of SAT/CP Workshop held with CP'06*. Nantes, France: 2006. pp. 33–47.
 36. Gomes C, Selman B, Crato N, *et al.* Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *J Autom Reason* 2000;24:67–100.
 37. Baptista L, Lynce I, Marques-Silva JP. Complete search restart strategies for satisfiability. *Proceedings of SSA'01 workshop held with IJCAI'01*. Seattle (WA): 2001.
 38. Lecoutre C, Sais L, Tabary S, *et al.* Recording and minimizing nogoods from restarts. *J Satisfiability, Boolean Modeling Comput (JSAT)* 2007;1:147–167.
 39. Brelaz D. New methods to color the vertices of a graph. *Commun ACM* 1979;22:251–256.
 40. Geelen PA. Dual viewpoint heuristics for binary constraint satisfaction problems. *Proceedings of ECAI'92*. Vienna, Austria: 1992. pp. 31–35.
 41. Refalo P. Impact-based search strategies for constraint programming. *Proceedings of CP'04 Toronto*, Canada: 2004. pp. 557–571.
 42. Morris P. The breakout method for escaping from local minima. *Proceedings of AAAI'93*. Washington (DC): 1993. pp. 40–45.
 43. Selman B, Kautz H. Domain-independent extensions to GSAT: solving large structured satisfiability problems. *Proceedings of IJCAI'93*. Chambéry, France: 1993. pp. 290–295.
 44. Thornton JR. Constraint weighting local search for constraint satisfaction. PhD thesis, Griffith University, Australia; 2000.
 45. Mazure B, Sais L, Gregoire E. Boosting complete techniques thanks to local search methods. *Ann Math Artif Intell* 1998;22:319–331.
 46. Eidenberg C, Faltings B. Using the breakout algorithm to identify hard and unsolvable subproblems. *Proceedings of CP'03*. Kinsale, Ireland: 2003. pp. 822–826.
 47. Boussemart F, Hemery F, Lecoutre C, *et al.* Boosting systematic search by weighting constraints. *Proceedings of ECAI'04*. Valencia, Spain: 2004. pp. 146–150.
 48. Lecoutre C, Boussemart F, Hemery F. Backjump-based techniques versus conflict-directed heuristics. *Proceedings of ICTAI'04*. Boca Raton (FL): 2004. pp. 549–557.

BACKWARD AND FORWARD EQUATIONS FOR DIFFUSION PROCESSES

ARKA P. GHOSH

Department of Statistics, Iowa
State University, Ames, Iowa

In this article we discuss two partial differential equations (PDEs) that arise in the theory of continuous-time continuous-state Markov processes, which was introduced by Kolmogorov in 1931. Here, we focus only on Markov diffusion processes (see **Introduction to Diffusion Processes**) and describe the forward and backward equations for such processes. The forward equation is also known as *Fokker–Planck equation* (and was already known in the physics literature before Kolmogorov formulated these). We begin with a brief introduction to continuous-time continuous-state Markov processes, which are continuous analogs of discrete-time Markov chains (DTMCs) and continuous-time Markov chains (CTMCs), discussed earlier in the sections titled “Discrete-Time Markov Chains (DTMCs)” and “Continuous-Time Markov Chains (CTMCs)” in the encyclopedia, followed by some basic properties of Markov processes. Then we state the two equations and provide sketches of the proofs. Finally, we conclude the article with some specific examples and features of these equations.

PRELIMINARIES

Diffusion processes have been discussed in the article titled **Introduction to Diffusion Processes**. For simplicity of the exposition, we consider the following time-homogeneous version of the diffusion process for this section: A (time-homogeneous) Itô diffusion is a stochastic process $\{X(t)\}$ satisfying a stochastic differential equation of the form

$$\begin{aligned} dX(t) &= b(X(t))dt + \sigma(X(t))dW(t), \\ t > 0; X(0) &= x, \end{aligned} \quad (1)$$

where $\{W(t)\}$ is a (standard) Brownian motion and b, σ are functions that satisfy

$$|\sigma(x) - \sigma(y)| < D|x - y|; x, y \in \mathbf{R}.$$

It can be shown that for $\{\mathcal{F}_t^W\}, \{\mathcal{F}_t^X\}$, representing the filtrations generated by W and X ,

$$\mathcal{F}_t^X \subseteq \mathcal{F}_t^W. \quad (2)$$

Markov Property

The diffusion satisfies the *Markov property*: If f is a bounded measurable function, then $E^x[f(X(t+h))|\mathcal{F}_t^W] = E^{X(t)}[f(X(h))]$, where the superscript in the expectation represents that these are conditional expectations given $X(0) = x$. Thus, from Equation (2), we have that: $E^x[f(X(t+h))|\mathcal{F}_t^X] = E^{X(t)}[f(X(h))]$. This intuitively means that the (future) evolution of the diffusion process is completely specified by the current value of the process (and the knowledge of the history of the process is not necessary). Compare this property to the Markov property introduced in the context of Markov chains (discrete-time versions in the section titled “Discrete-Time Markov Chains (DTMCs)” and continuous-time versions in the section titled “Continuous-Time Markov Chains (CTMCs)” in the encyclopedia) above. In fact, the above property holds in a stronger sense, in which the time t can be replaced by a random (stopping) time τ (strong Markov property).

Infinitesimal Generator

For Markov processes, the infinitesimal generator is defined as

$$Af(x) = \lim_{t \downarrow 0} \frac{E^x(f(X(t))) - f(x)}{t}, \quad x \in \mathbf{R}. \quad (3)$$

The set of functions f for which the limit on the right side exists (for all x) is called the *domain of the operator*, and denoted by $\mathcal{D}_A$. For an Itô diffusion that we are concerned with, this operator is a second-order partial differential operator, and can be written

down explicitly as follows:

$$Af(x) = b(x)\frac{df}{dx} + \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2}, \quad (4)$$

and $\mathcal{D}_A$ contains all twice-differentiable functions with compact support. This operator plays a fundamental role in the study of diffusion processes, and is relevant for understanding the forward and backward Kolmogorov equations for diffusions.

Examples

Here are generators of some basic examples of diffusion processes discussed in the article titled **Introduction to Diffusion Processes**. For a (general) Brownian motion (with drift b and diffusion coefficient $\sigma > 0$) the generator is

$$Af(x) = bf'(x) + \frac{1}{2}\sigma^2 f''(x). \quad (5)$$

For geometric Brownian motion process (with parameters b and σ), the generator is

$$Af(x) = bxf'(x) + \frac{1}{2}x^2\sigma^2 f''(x). \quad (6)$$

For Ornstein–Uhlenbeck process (with parameters b and σ), it is

$$Af(x) = -bxf'(x) + \frac{1}{2}\sigma^2 f''(x). \quad (7)$$

FORWARD AND BACKWARD EQUATIONS FOR DIFFUSION PROCESSES

These are two (partial) differential equations that characterize the dynamics of the distribution of the diffusion process. Kolmogorov's forward equation addresses the following: If at time t the state of the system is x , what can we say about the distribution of the state at a future time $s > t$ (hence the term *forward*). The backward equation, on the other hand, is useful when addressing the question that given that the system at a future time s has a particular behavior, what can we say about the distribution at time $t < s$. This imposes a terminal condition on the PDE, which is integrated backward in time, from s to t (hence the term

backward is associated with this). Historically, the forward equation was discovered (as the Fokker–Plank equation) before the backward equation. However, the backward equation is somewhat more general and we will describe that first.

The forward and backward equations are expressed in various equivalent forms. Here, we describe them in the following form, which illustrates the use of the terms forward and backward more clearly: Fix an interval $[0, T]$, and we will deal with $t \in [0, T]$.

Backward Equation

Let $g(x)$ be a bounded smooth (twice continuously differentiable having compact support) function, and let

$$u(t, x) = E^{x,t}(g(X(T))) \equiv E(g(X(T)) | X(t) = x). \quad (8)$$

Then u satisfies

$$\begin{aligned} \frac{\partial u}{\partial t} + Au &= 0, \text{ with the “terminal” condition} \\ u(T, x) &= g(x), \end{aligned} \quad (9)$$

where the right-hand side is to be interpreted as A (as in Equation (4)) applied to the $u(t, x)$ as a function of x .

Forward Equation

In addition, if $X(t)$ has a density $p(t, x)$, then for a probability density function $\mu(\cdot)$, the probability densities satisfy the following:

$$\begin{aligned} \frac{\partial}{\partial t} p(t, x) &= (A^* p)(t, x) \text{ with the initial} \\ \text{condition } p(0, x) &= \mu(x). \end{aligned} \quad (10)$$

Here A^* is the adjoint operator of A , defined as

$$\begin{aligned} A^* v(t, y) &= -\frac{\partial}{\partial y} (b(y)v(t, y)) \\ &\quad + \frac{1}{2} \frac{\partial^2}{\partial y^2} (\sigma^2(y)v(t, y)). \end{aligned} \quad (11)$$

Proof

We now provide a sketch of the proof of the backward equation (9). First, note that for the Markov (diffusion) process $\{X(t)\}$ with some transition probability function

$$P(t, x, A) = P(X(t) \in A | X(0) = x),$$

the following Chapman–Kolmogorov equation holds:

$$P(t + s, x, A) = \int P(t, x, du)P(s, u, A). \quad (12)$$

Recall the Chapman–Kolmogorov equation introduced earlier in the section titled “Discrete-Time Markov Chains (DTMCs)” in the encyclopedia in the context of Markov chains. Also note that for the diffusion in Equation (1), the following properties are satisfied for the transition probabilities

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \int_{\{|x-y| \geq \delta\}} P(t, x, dy) &= 0, \\ \lim_{t \rightarrow \infty} \frac{1}{t} \int_{\{|x-y| < \delta\}} (y-x)P(t, x, dy) &= b(x), \\ \lim_{t \rightarrow \infty} \frac{1}{t} \int_{\{|x-y| \geq \delta\}} (y-x)^2 P(t, x, dy) &= \sigma(x), \end{aligned} \quad (13)$$

for all $\delta > 0$. Informally, it means the following: in a small time interval, it has negligible probability of being away from x ; also, the mean and variance of the “displacements” of the diffusion process is approximately the drift and the diffusion coefficients, respectively. We will use these properties in the proofs of the forward and backward equations here. First, observe that from Equations (12) and (8) it follows that

$$u(t + h, x) = \int P(h, x, dy)u(t, y).$$

Hence,

$$\begin{aligned} \frac{u(t+h) - u(t)}{h} &= \frac{\int P(h, x, dy)[u(t, y) - u(t, x)]}{h} \\ &\approx \frac{1}{h} \int P(h, x, dy) \left[(y-x) \frac{\partial u(t, x)}{\partial x} \right. \\ &\quad \left. + \frac{1}{2} (y-x)^2 \frac{\partial^2 u(t, x)}{\partial x^2} \right], \end{aligned} \quad (14)$$

where the last expression follows from Taylor’s expansion, and the approximate equality is a consequence of Equation (13), for h small. Hence, by taking limit as $h \rightarrow 0$, using 13 and the form of the generator A in Equation (4), the backward equation follows.

Now we sketch the proof of the forward equation (10) using the backward equation as follows: Assume that the random variable $X(t)$ has a density $p(t, x)$ and hence, from Equation (8) and properties of conditional expectations, we have

$$\begin{aligned} E(g(X(T))) &= E(E^{x,t}(g(X(T)))) \\ &= \int u(t, x)p(t, x) dx. \end{aligned}$$

Since the left side is free of t , we get by taking derivatives with respect to t ,

$$0 = \int p(t, x) \frac{\partial u(t, x)}{\partial t} dx + \int \frac{\partial p(t, x)}{\partial t} u(t, x) dx. \quad (15)$$

Since u satisfies the backward equation (9), using the form of A in Equation (4), we get from Equation (15) that

$$\begin{aligned} 0 &= \int p(t, x) \left(-b(x) \frac{\partial u(t, x)}{\partial x} - \frac{1}{2} \sigma^2(x) \frac{\partial^2 u(t, x)}{\partial x^2} \right) dx \\ &\quad + \int \frac{\partial p(t, x)}{\partial t} u(t, x) dx. \end{aligned}$$

Finally, using integration by parts (which replaces x -derivatives of u by those of p), assuming that the integrals decay fast enough (as $|x| \rightarrow \infty$) and using the form of A^* in Equation (11), we get that

$$\int \left(-(A^*p)(x, t) + \frac{\partial p(t, x)}{\partial t} \right) u(t, x) dx = 0.$$

Since this equation should be true for all functions u , we get that $(A^*p)(x, t) = \frac{\partial p(t, x)}{\partial t}$, which proves the forward equation.

EXAMPLES

Here we discuss some special cases of the forward and backward diffusions.

Example 1 [Martingale Property].

When $b \equiv 0$, then the forward equation reduces to

$$\frac{1}{2} \frac{\partial^2}{\partial x^2} (\sigma^2(x)p(t, x)) = \frac{\partial}{\partial t} p(t, x). \quad (16)$$

Hence, using integration by parts, one gets

$$\begin{aligned} \frac{d}{dt} E[X(t)] &= \frac{d}{dt} \int_{-\infty}^{\infty} xp(x, t) dx \\ &= \int_{-\infty}^{\infty} x \frac{\partial}{\partial t} p(t, x) dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{2} \frac{\partial^2}{\partial x^2} (\sigma^2(x, t)p(x, t)) dx \\ &= - \int_{-\infty}^{\infty} \frac{1}{2} \frac{\partial}{\partial x} (\sigma^2(x, t)p(x, t)) dx = 0. \end{aligned}$$

In other words, $E[X(t)]$ does not change with t . In fact, one can prove a stronger statement: the process $\{X(t)\}$ is a martingale, that is, $E[X(t) | \mathcal{F}_s^W] = X(s)$ for all $s < t$.

Example 2. Consider the Ornstein–Uhlenbeck process with parameters b and σ (see *Introduction to Diffusion Processes*), whose generator was discussed earlier (for simplicity, we assume $\sigma = 1$ here). The backward equation for this process follows from the general form of the Equation (9) and the generator in Equation (7)

$$\frac{\partial}{\partial t} u(t, x) = -bx \frac{\partial}{\partial x} u(t, x) + \frac{1}{2} \frac{\partial^2}{\partial x^2} u(t, x), \quad (17)$$

which can be solved explicitly. It turns out, with a change of variable, it reduces to the standard diffusion equation (16) (with $\sigma(x) \equiv 1$). From that, it can be deduced that the transition probability $P(t, x, \cdot)$ is the normal (Gaussian) probability with mean xe^{-bt} and variance parameter $\frac{(1-e^{-2bt})}{2b}$.

Example 3. Similarly, for the geometric Brownian motion process, one can get the forward equation by substituting $b(x) = bx, \sigma(x) = \sigma x$ in Equations (9) and (6). Solving these explicitly yields that the transition probability in this case is a log-normal

probability

$$P(t, x, A) = \int_A \frac{1}{\sigma y \sqrt{2\pi t}} \exp \left(-\frac{\log(y/x) - (b - \frac{1}{2}\sigma^2)t}{2\sigma^2 t} \right) dy.$$

Remark. The backward equation holds whenever b is continuous and $\sigma > 0$. But for the forward equation to hold, one clearly needs the derivatives of these two functions. In that sense, backward equation is more general than the forward equation. However, when such derivatives exist, one can show that there exists a “minimal” solution for the transition function $P(t, x, A)$ such that $u(\cdot, \cdot)$ in Equation (8) solves the backward equation. But this transition function in such cases need not be a “proper” probability. To guarantee that a proper transition probability can be obtained from these equation, one usually assumes additional boundary conditions. Different choices of such boundary conditions (e.g., absorbing boundary condition, reflecting boundary conditions etc.) uniquely characterize the associated process (see Chapter 10 of Ref. 1 for more discussion on this topic).

FURTHER READING

More details about these equations can be found in classical textbooks on probability, such as Ref. 1. Other references for these equations for diffusion processes are Refs 2–5, and so on. For students and researchers that are trying to study this material for the first time, more accessible references could be Refs 6 and 7. A reference for general Markov processes can be found in Ref. 8. For partial differential equations, in general, one is encouraged to consult Ref. 9.

As mentioned earlier, these equations are very important in physics and there is a vast literature [10–12] on this topic from the physics community as well. For applications to financial mathematics, see Refs 13 and 14. A general reference for students in operations research to learn stochastic

systems in general is Ref. 15. For more applications in queueing theory models involving diffusions processes, see Refs 16 and 17.

REFERENCES

1. Feller W. An introduction to probability theory and its applications. Volume 2, 2nd ed. New York: John Wiley & Sons; 1991.
2. Ikeda N. Ito's stochastic calculus and probability theory. New York: Springer; 1996.
3. Itô K, McKean HP. Diffusion processes and their sample paths. New York: Springer; 1996.
4. Protter P. Stochastic integration and differential equations. 2nd ed. New York: Springer; 2005.
5. Stroock DW, Varadhan SRS. Multidimensional diffusion processes. New York: Springer; 2005.
6. Karatzas I, Shreve S. Brownian motion and stochastic calculus. 2nd ed. New York: Springer; 1991.
7. Oksendal BK. Stochastic differential equations: an introduction with applications. 6th ed. Berlin, Heidelberg: Springer, 2003.
8. Ethier SN, Kurtz TG. Markov processes. Characterization and convergence. Wiley series in probability and mathematical statistics. New York: John Wiley & Sons, Inc.; 1986.
9. Gilbarg D, Trudinger NS. Elliptic partial differential equations of second order. New York: Springer; 2001.
10. Berry S. Functional integration and quantum physics. New York: Academic Press; 1979.
11. Kadanoff LP. Statistical physics: statics, dynamics and renormalization. New York: World Scientific; 2000.
12. Risken H, Frank T. The Fokker-Planck equation: methods of solutions and applications. 2nd ed. Berlin, Heidelberg: Springer; 1996.
13. Karatzas I, Shreve S. Methods of mathematical finance. New York: Springer; 2001.
14. Steele JM. Stochastic calculus and financial applications. New York: Springer; 2000.
15. Kulkarni VG. Modeling and analysis of stochastic systems. 1st ed. London: Chapman & Hall; 1996.
16. Whitt W. Stochastic-process limits. New York: Springer; 2002.
17. Kushner HJ. Heavy traffic analysis of controlled queueing and communications networks. 1st ed. New York: Springer; 2001.

BASIC CP THEORY: CONSISTENCY AND PROPAGATION (ADVANCED)

CHRISTIAN BESSIERE
University of Montpellier,
Montpellier, France

Constraint reasoning involves various types of techniques to tackle the inherent intractability of the problem of satisfying a set of constraints. Constraint propagation is one of those types of techniques. Constraint propagation embeds any reasoning which consists in explicitly forbidding values or combinations of values for some variables of a problem because a given subset of the constraints cannot be satisfied otherwise. For instance, in a problem containing two variables x_1 and x_2 taking integer values in $1..10$, and a constraint specifying that $|x_1 - x_2| > 5$, by propagating this constraint we can forbid values 5 and 6 for both x_1 and x_2 . Removing these values is a way to reduce the space of combinations that will be explored by a search mechanism.

The concept of constraint propagation can be found in other fields under different kinds and names. (See, for instance, the propagation of clauses by “unit propagation” in propositional calculus [1].) Nevertheless, it is in constraint reasoning that this concept appears in such a variety of forms, and in which its characteristics have been so deeply analyzed.

Constraint propagation can be presented along two main lines: *local consistencies* and *rules iteration*. Local consistencies define properties that the constraint problem must satisfy *after* constraint propagation. This way, the operational behavior is left completely open, the only requirement being to achieve the given property on the output. The rules iteration approach, on the contrary, describes the process of propagation itself. Rules are conditions on the kind of operations of reduction that can be applied to the problem. In this article, I will present constraint propagation mainly through local

consistency. Rules iteration will be briefly discussed at the end of the article.

BACKGROUND

A *constraint satisfaction problem (CSP)* involves finding solutions to a constraint network, that is, assignments of values to its variables that satisfy all its constraints. Constraints specify combinations of values that given subsets of variables are allowed to take. In this article, we are only concerned with CSPs where variables take their value in a *finite* domain. Without loss of generality, I assume these domains are mapped on the set $\mathbb{Z}$ of integers, and so, I consider only *integer* variables, that is, variables with a domain being a finite subset of $\mathbb{Z}$.

Definition 1 [Constraint]. A *constraint* c is a relation defined on a sequence of variables $X(c) = (x_{i_1}, \dots, x_{i_{|X(c)|}})$, called the *scheme* of c . c is the subset of $\mathbb{Z}^{|X(c)|}$ that contains the combinations of values (or *tuples*) $\tau \in \mathbb{Z}^{|X(c)|}$ that *satisfy* c . $|X(c)|$ is called the *arity* of c . Testing whether a tuple τ satisfies a constraint c is called a *constraint check*.

A constraint can be specified extensionally by the list of its satisfying tuples (or the list of its forbidden tuples), or intensionally by a formula that is the characteristic function of the constraint. Definition 1 allows constraints with an *infinite* number of satisfying tuples. Constraints of arity 2 are called *binary* and constraints of arity greater than 2 are called *nonbinary*.

Example 1. The constraint *alldifferent* $(x_1, x_2, x_3) \equiv (x_1 \neq x_2 \wedge x_1 \neq x_3 \wedge x_2 \neq x_3)$ allows the infinite set of 3-tuples in $\mathbb{Z}^3$ such that all values are different. The constraint $c(x_1, x_2, x_3) = \{(2, 2, 3), (2, 3, 2), (2, 3, 3), (3, 2, 2), (3, 2, 3), (3, 3, 2)\}$ allows the finite set of 3-tuples containing both values 2 and 3 and only them.

Definition 2 [Constraint network]. A *constraint network* (or *network*) is composed of:

- a finite sequence of integer variables $X = (x_1, \dots, x_n)$,
- a domain for X , that is, a set $D = D(x_1) \times \dots \times D(x_n)$, where $D(x_i) \subset Z$ is the finite set of values, given in extension, that variable x_i can take, and
- a set of constraints $C = \{c_1, \dots, c_e\}$, where variables in $X(c_j)$ are in X .

Given a constraint network N , I sometimes use X_N , D_N , and C_N to denote its sequence of variables, its domain, and its set of constraints. Given a set of variables $Y = \{x_{i_1}, \dots, x_{i_q}\} \subseteq X$, D^Y denotes the Cartesian product $D(x_{i_1}) \times \dots \times D(x_{i_q})$. According to Definitions 1 and 2, the variables X_N of a network N and the scheme $X(c)$ of a constraint $c \in C_N$ are sequences of variables, not sets. This is required because the order of the values matters for tuples in c . Nevertheless, it simplifies a lot the notations to consider sequences as sets when no confusion is possible. For instance, given two sequences W and Y , $W \subseteq Y$ means that the sequence W involves only variables that are in the sequence Y , whatever their ordering in the sequence. Given a tuple τ on a sequence Y of variables, and given a sequence $W \subseteq Y$, $\tau[W]$ denotes the restriction of τ to the variables in W , ordered according to W . Given $x_i \in Y$, $\tau[x_i]$ denotes the value of x_i in τ .

Backtracking algorithms are based on the principle of assigning values to variables until all variables are *instantiated*.

Definition 3 [Instantiation]. Given a network $N = (X, D, C)$,

- An *instantiation* I on $Y = (x_1, \dots, x_k) \subseteq X$ is an assignment of values $v_1, \dots, v_k$ to the variables $x_1, \dots, x_k$, that is, I is a tuple on Y . I can be denoted by $((x_1, v_1), \dots, (x_k, v_k))$ where (x_i, v_i) denotes the value v_i for x_i .

- An instantiation I on Y is *valid* if for all $x_i \in Y$, $I[x_i] \in D(x_i)$.
- An instantiation I on Y is *locally consistent* iff it is valid and for all $c \in C$ with $X(c) \subseteq Y$, $I[X(c)]$ satisfies c . If I is not locally consistent, it is *locally inconsistent*.
- A *solution* to a network N is an instantiation I on X which is locally consistent. The set of solutions of N is denoted by $\text{sol}(N)$.

Example 2. Let $N = (X, D, C)$ be a network with $X = (x_1, x_2, x_3, x_4)$, $D(x_i) = \{1, 2, 3, 4, 5\}$ for all $i \in [1..4]$ and $C = \{c_1(x_1, x_2, x_3), c_2(x_1, x_2, x_3), c_3(x_2, x_4)\}$ with $c_1(x_1, x_2, x_3) = \text{alldifferent}(x_1, x_2, x_3)$, $c_2(x_1, x_2, x_3) \equiv (x_1 \leq x_2 \leq x_3)$, and $c_3(x_2, x_4) \equiv (x_4 \geq 2 \cdot x_2)$. $I_1 = ((x_1, 1), (x_2, 2), (x_4, 7))$ is a nonvalid instantiation on $Y = (x_1, x_2, x_4)$ because $7 \notin D(x_4)$. $I_2 = ((x_1, 1), (x_2, 1), (x_4, 3))$ is a locally consistent instantiation on Y because c_3 is the only constraint with scheme included in Y and it is satisfied by $I_2[X(c_3)]$. $I_3 = ((x_1, 1), (x_2, 2), (x_3, 3), (x_4, 5))$ is a solution because it is a locally consistent instantiation on X .

ARC CONSISTENCY

Arc consistency is the oldest and most well-known way of propagating constraints. This is indeed a very simple and natural concept that guarantees every value in a domain to be consistent with every constraint.

Example 3 [From Ref. 2]. Let N be the network that involves the three variables x_1 , x_2 , and x_3 , domains $D(x_1) = D(x_2) = D(x_3) = \{1, 2, 3\}$, and constraints $c_{12} \equiv (x_1 = x_2)$ and $c_{23} \equiv (x_2 < x_3)$. N is not arc consistent because there are some values inconsistent with some constraints. Checking constraint c_{12} does not permit to remove any value. But when checking constraint c_{23} , we see that $(x_2, 3)$ must be removed because there is no value greater than it in $D(x_3)$. We can also remove value 1

from $D(x_3)$ because of constraint c_{23} . Removing 3 from $D(x_2)$ causes in turn the removal of value 3 for x_1 because of constraint c_{12} . Now, all remaining values are compatible with all constraints.

The seminal papers on arc consistency are due to Mackworth, who was the first to clearly define the concept of arc consistency for binary constraints [3], extend definitions and algorithms to nonbinary constraints [4], and analyze the complexity [5].

I give a definition of arc consistency in its most general form, that is, for arbitrary constraint networks. In this case it is often called *generalized arc consistency* (GAC).

Definition 4 [(Generalized) arc consistency ((G)AC)]. Given a network $N = (X, D, C)$, a constraint $c \in C$, and a variable $x_i \in X(c)$,

- A value $v_i \in D(x_i)$ is *consistent with c* in D iff there exists a valid tuple τ satisfying c such that $v_i = \tau[\{x_i\}]$. Such a tuple is called a *support* for (x_i, v_i) on c .
- The constraint c is (generalized) arc consistent on D iff all values of all variables in $X(c)$ have a support on c .
- The network N is (generalized) arc consistent iff all constraints in C are (generalized) arc consistent on D .

By achieving (or enforcing) arc consistency on a network $N = (X, D, C)$, I mean finding the *arc consistent closure* $AC(N)$ of N . $AC(N)$ is the network (X, D_{AC}, C) where $D_{AC} = \cup\{D' \subseteq D \mid (X, D', C) \text{ is arc consistent}\}$. $AC(N)$ is arc consistent and is *unique*. It has the same solutions as N .

Enforcing Arc Consistency

Proposing efficient algorithms for enforcing arc consistency has always been considered as a central question in the constraint reasoning community. A first reason is that arc consistency is the basic propagation mechanism that is used in all solvers. A second reason is that the new ideas that permit to improve

efficiency of arc consistency can usually be applied to algorithms achieving other local consistencies.

The most well-known algorithm for arc consistency is the one proposed by Mackworth in Ref. 3 under the name AC3. It was extended to GAC in arbitrary networks in Ref. 4. I present it in its general version. (See Algorithm 1.)

The main component of GAC3 is the revision of an arc, that is, the update of a domain with respect to a constraint. (The word “arc” comes from the binary case but we also use it on nonbinary constraints.) Updating a domain $D(x_i)$ with respect to a constraint c means removing every value in $D(x_i)$ that is not consistent with c . The function $\text{Revise}(x_i, c)$ takes each value v_i in $D(x_i)$ in turn (line 2), and explores the space $D^{X(c) \setminus \{x_i\}}$, looking for a support on c for v_i (line 3). If such a support is not found, v_i is removed from $D(x_i)$ and the fact that $D(x_i)$ has been changed is flagged (lines 4–5). The function returns true if the domain $D(x_i)$ has been reduced, false otherwise (line 6).

The main algorithm is a simple loop that revises the arcs until no change occurs, to ensure that all domains are consistent with all constraints. To avoid too many useless calls to Revise , the algorithm maintains a list Q of all the pairs (x_i, c) for which we are not sure that $D(x_i)$ is arc consistent on c . In line 7, Q is filled with all possible pairs (x_i, c) such that $x_i \in X(c)$. Then, the main loop (line 8) picks the pairs (x_i, c) in Q one by one (line 9) and calls $\text{Revise}(x_i, c)$ (line 10). If $D(x_i)$ is wiped out, the algorithm returns false (line 11). Otherwise, if $D(x_i)$ is modified, it can be the case that a value for another variable x_j has lost its support on a constraint c' involving both x_i and x_j . Hence, all pairs (x_j, c') such that $x_i, x_j \in X(c')$ must be put again in Q (line 12). When Q is empty, the algorithm returns true (line 13) as we are sure that all arcs have been revised and all remaining values of all variables are consistent with all constraints.

Algorithm 1: AC3 / GAC3

```

function Revise3(in  $x_i$ : variable;  $c$ : constraint): Boolean;
  begin
    1  CHANGE  $\leftarrow$  false;
    2  foreach  $v_i \in D(x_i)$  do
    3    if  $\nexists$  a valid tuple  $\tau \in c$  with  $\tau[x_i] = v_i$  then
    4      remove  $v_i$  from  $D(x_i)$ ;
    5      CHANGE  $\leftarrow$  true;
    6  return CHANGE;
  end

function AC3/GAC3(in  $X$ : set): Boolean;
  begin
    /* initialisation */;
    7   $Q \leftarrow \{(x_i, c) \mid c \in C, x_i \in X(c)\}$ ;
    /* propagation */;
    8  while  $Q \neq \emptyset$  do
    9    select and remove  $(x_i, c)$  from  $Q$ ;
    10   if Revise( $x_i, c$ ) then
    11     if  $D(x_i) = \emptyset$  then return false;
    12     else  $Q \leftarrow Q \cup \{(x_j, c') \mid c' \in C \wedge c' \neq c \wedge x_i, x_j \in X(c') \wedge j \neq i\}$ ;
    13 return true;
  end

```

GAC3 is polynomial in the arity of the constraints. It runs in $O(er^3 d^{r+1})$ time, where r is the greatest arity among constraints. When constraints are all binary, this gives $O(ed^3)$ time. The time complexity of AC3 is not optimal. The fact that function Revise does not remember anything about its computations to find supports for values leads AC3 to do and redo many times the same constraint checks.

A lot of work has been done to improve the worst-case time complexity of AC3. Mohr and Henderson proposed AC4, the first optimal algorithm for arc consistency ($O(ed^2)$ on binary constraints and $O(erd^r)$ in general), at the price of heavy data structures [6,7]. Bessiere and Cordier proposed AC6, a compromise between AC3 and AC4 that maintains a subtle data structure of lists that have appeared to be difficult to integrate in standard constraint solvers [8,9]. Bessiere, Régim, Yap, and Zhang proposed AC2001, which has the advantage of being based on the same framework as AC3 while having optimal time complexity, like AC4 and AC6 [10]. Finally, Lecoutre and Hemery have proposed AC3rm, an algorithm particularly

efficient for maintaining arc consistency during search [11]. AC3rm stores information in a way similar to AC2001, uses the concept of multidirectionality as AC7 does [12], but does not restore information when backtracking to previous states of search. AC3rm thus loses the theoretical time optimality of the algorithms on which it is based but it works well in practice. Several modern constraint solvers are based on ACrm.

CONSISTENCIES STRONGER THAN ARC CONSISTENCY

Arc consistency is not the only way to detect inconsistencies in a network, and as early as in the 1970s, several authors proposed techniques that discover more inconsistencies than arc consistency. Freuder extended the notion of local consistency to a whole class of consistencies stronger than arc consistency, called k -consistencies [13,14]. Enforcing k -consistency makes explicit nogoods of size $k - 1$.

Definition 5 [k -consistency]. Let $N = (X, D, C)$ be a network.

Given a set of variables $Y \subseteq X$ with $|Y| = k - 1$, a locally consistent instantiation I on Y is k -consistent iff for any k th variable $x_{i_k} \in X \setminus Y$ there exists a value $v_{i_k} \in D(x_{i_k})$ such that $I \cup \{(x_{i_k}, v_{i_k})\}$ is locally consistent.

The network N is k -consistent iff for any set Y of $k - 1$ variables, any locally consistent instantiation on Y is k -consistent.

Before Freuder proposed k -consistency, Montanari had proposed a local consistency called *path consistency*. It was defined for binary networks where each pair of variables is involved in at most one constraint [15]. In such networks, path consistency is equivalent to 3-consistency.

Example 4. Consider the network N with variables x_1, x_2, x_3 , domains $D(x_1) = D(x_2) = D(x_3) = \{1, 2\}$, and $C = \{x_1 \neq x_2, x_2 \neq x_3\}$. The pairs $((x_1, 1), (x_3, 2))$ and $((x_1, 2), (x_3, 1))$ are locally consistent because no constraint has its scope included in $\{x_1, x_3\}$, but they cannot be extended to a value of x_2 satisfying both $x_1 \neq x_2$ and $x_2 \neq x_3$. Thus, N is not path/3-consistent. The network $N' = (X, D, C \cup \{x_1 = x_3\})$ is path/3-consistent.

Freuder also defined strong k -consistencies. They are properties that guarantee that the network is j -consistent for $1 \leq j \leq k$. Thus, we can build from scratch a locally consistent instantiation of size k without any backtrack.

Definition 6 [Strong k -consistency]. A network is *strongly k -consistent* iff it is j -consistent for all $j \leq k$.

The maximal amount of simplification we can perform on a network is to reach a network on which *all* locally consistent instantiations can be extended to solutions. Strong n -consistency guarantees that, with $|X| = n$.

A drawback of k -consistency is that enforcing it will produce additional constraints of arity $k - 1$ that were not in C_N

(see Example 4) and that are expensive to store. In Ref. 16, Freuder proposed (i, j) -consistency, a generalization of k -consistency where we do not guarantee that instantiations of size $k - 1$ can be extended to instantiations of size k , but instantiations of size i can be extended to j additional variables. k -consistency is $(k - 1, 1)$ -consistency. Since the main drawback of k -consistencies is the huge space they require to store all forbidden instantiations of size $k - 1$, we can design local consistencies requiring less space by setting i to a small value in (i, j) -consistency.

Consistencies based on Constraints

All strong consistencies we have seen until now are properties of partial instantiations of variables with respect to other variables. They do not take into account the network topology, that is, which sets of variables are linked by a constraint and which are not. This is a limitation for constraint propagation, which creates new constraints everywhere in the network.

Janssen *et al.* proposed a first local consistency based on constraints instead of variables [17]. It was applied from works on relational databases [18].

Definition 7 [Pairwise consistency].

Given a network N , a pair of constraints c_1 and c_2 in C_N is pairwise consistent iff any valid tuple on $X(c_1)$ (respectively, on $X(c_2)$) satisfying c_1 (respectively, c_2) can be extended to a valid instantiation on $X(c_1) \cup X(c_2)$ satisfying c_2 (respectively, c_1). N is *pairwise consistent* iff any pair of constraints in C_N is pairwise consistent.

Example 5. Consider the network with variables x_1, x_2, x_3, x_4 , domains $D(x_1) = D(x_2) = D(x_3) = D(x_4) = \{1, 2\}$ and constraints $c_1(x_1, x_2, x_3) = \{(1, 2, 1), (2, 1, 1), (2, 2, 2)\}$ and $c_2(x_2, x_3, x_4) = \{(1, 1, 1), (2, 2, 2)\}$. This network is generalized arc consistent. However, it is not pairwise consistent because the tuple $(1, 2, 1)$ from c_1 is not compatible with any tuple in c_2 .

Janssen *et al.* showed in Ref. 17 that pairwise consistency is equivalent to 2-consistency on the dual encoding of the

network, where dual variables represent constraints of the original network [19].

Other authors proposed other types of consistencies, that, like k -consistency, can be defined for arbitrary levels. In a database context, Gyssens proposed k -wise consistency, a direct extension of pairwise consistency where we consider k constraints at a time instead of two [20]. Jégou applied this notion to constraint networks [21]. In Ref. 22, Jégou proposed another duality between variables and constraints. He presents hyper k -consistency.

Dechter and van Beek proposed a new form of local consistency which is more bound to schemes of constraints already in the network than hyper k -consistency is. They refer to those new types of consistencies as *relational* consistencies [23].

Definition 8 [Relational (i, m) -consistency]. Let N be a network. A set of constraints $\{c_1, \dots, c_m\} \subseteq C_N$ is relationally (i, m) -consistent iff for every subset of variables $Y \subseteq \bigcup_{i=1}^m X(c_i)$ such that $|Y| = i$, any locally consistent instantiation on Y has an extension to $\bigcup_{i=1}^m X(c_i)$ that satisfies $c_1, \dots, c_m$ simultaneously. N is relationally (i, m) -consistent iff every subset of m constraints in C_N is relationally (i, m) -consistent.

GAC is relational $(1, 1)$ -consistency.

Consistencies that only Prune Values

There exist local consistencies that permit to prune more values than arc consistency while keeping the set of constraints unchanged (as opposed to what is done by k -consistencies and consistencies based on constraints). Several of them are different kinds of reasoning we can apply on triples of variables. Examples are restricted path consistency (RPC) [24], path-inverse consistent (PIC) [25], or max-restricted path consistency (maxRPC) [26].

Another approach to enhance the amount of values pruned consists in trying in turn different assignments of a value to a variable, and performing constraint propagation on the subproblem obtained by this assignment.

If the problem is found to be inconsistent, this means that this value does not belong to any solution and thus can be pruned. This kind of technique was used on the bounds of interval domains in scheduling (“shaving” in Ref. 27) or on continuous CSPs (3B-consistency in Ref. 28). This technique was also used on Boolean variables as a way to derive better variable ordering heuristics in search procedures for formulas in conjunctive normal form (SAT) in Refs 29 and 30. Finally, it was formalized as a class of local consistencies in Refs 31–33 under the name “singleton consistencies.” I give the definition in the case where the amount of propagation applied to each subproblem is arc consistency. In the following, the subnetwork obtained from a network N by reducing the domain of a variable x_i to the singleton $\{v_i\}$ is denoted by $N|_{x_i=v_i}$.

Definition 9 [Singleton arc consistency]. A network $N = (X, D, C)$ is *singleton arc consistent* (SAC) iff for all $x_i \in X$, for all $v_i \in D(x_i)$, the subproblem $N|_{x_i=v_i}$ can be made arc consistent without wiping out a domain.

Several algorithms have been proposed for establishing SAC [31, 34–36].

CONSISTENCIES WEAKER THAN ARC CONSISTENCY

When constraints involve a great number of variables, even arc consistency can become too expensive to be applied at each node of a backtrack search. We can use the fact that domains are composed of integers to perform cheaper propagation. Integer domains inherit the total ordering on z and by consequence they inherit the smallest and greatest values in $D(x_i)$, denoted by $\min_D(x_i)$ and $\max_D(x_i)$ respectively, and called the *bounds* of $D(x_i)$. Bound consistency is a local consistency that uses the ordering on the domains to reduce the cost of propagating a constraint. (It is called BC(Z) in Refs 2 and 37.)

Definition 10 [Bound consistency]. Given a network $N = (X, D, C)$, given a constraint c , a *bound support* τ on c is

a tuple that satisfies c and such that for all $x_i \in X(c)$, $\min_D(x_i) \leq \tau[x_i] \leq \max_D(x_i)$. A constraint c is *bound consistent (BC)* iff for all $x_i \in X(c)$, $(x_i, \min_D(x_i))$ and $(x_i, \max_D(x_i))$ belong to a bound support on c . The network N is bound consistent iff all its constraints are bound consistent.

Example 6 [From Ref. 2]. Consider the network with variables $x_1, \dots, x_6$, domains $D(x_1) = D(x_2) = \{1, 2\}$, $D(x_3) = D(x_4) = \{2, 3, 5, 6\}$, $D(x_5) = \{5\}$, $D(x_6) = [3..7]$, and $C = \{\text{alldifferent}(x_1 \dots, x_6)\}$. BC will only prune value 2 from $D(x_3)$ and $D(x_4)$ because they are the only bounds for which we cannot find a bound support. GAC will additionally prune value 5 from $D(x_3)$ and $D(x_4)$ and values 3, 5, and 6 from $D(x_6)$.

CONSTRAINT PROPAGATION AS ITERATION OF REDUCTION RULES

In previous sections, I presented local consistencies in a generic way without saying what should be done when we have some specific information on the semantics of a constraint. In constraint solvers, many specific constraints are already defined. For these constraints, it is often very inefficient to call a generic propagation algorithm such as the function *Revise* in Algorithm 1. Hence, these solvers attach a specific propagator to the specific types of constraints they contain. The simplest way of designing a propagator is to specify a set of propagation rules. A propagation rule is a condition under which some values can be pruned. The propagation algorithm iterates on the rules until no more changes occur in domains, that is, until a fixpoint has been reached. This rules iteration process may achieve a given level of local consistency or may reach a fixpoint that does not correspond to any properly defined level of consistency. I will just illustrate propagation rules through an example. The full theory on rules iteration can be found in Refs 38–42.

Example 7. Consider the constraint $\text{alldifferent}(x, y, z)$. Instead of calling the

function *Revise* to propagate this constraint, which would cost $O(d^3)$, we can build the following three rules:

- R1: **if** $\exists v, D(x) = \{v\}$ **then** remove v from $D(y)$ and $D(z)$
- R2: **if** $\exists v, D(y) = \{v\}$ **then** remove v from $D(x)$ and $D(z)$
- R3: **if** $\exists v, D(z) = \{v\}$ **then** remove v from $D(x)$ and $D(y)$

We see that this set of rules is very time efficient because it is constant time to apply it each time a domain has been modified. However, it does not achieve arc consistency. Suppose $D(x) = D(y) = \{1, 2\}$ and $D(z) = \{1, 2, 3, 4\}$, these rules do not prune Anything, whereas 1 and 2 would be pruned from $D(z)$ by arc consistency.

Constraint solvers usually provide other facilities to improve the efficiency of propagation. Arithmetic constraints are at the core of most constraint solvers and it appears that when considering arithmetic constraints, a reduction of a domain does not produce the same effect on the other variables of the constraint, depending on if it is the removal of a value in the middle of the domain, if it is the increase of its minimum value, if it is the decrease of its maximum value, or if it is an instantiation to a single value. Then, it is worth differentiating these different types of *events* to be able to propagate exactly as necessary. The events usually recognized by constraint solvers are:

- $\text{RemValue}(x_i)$: when a value v is removed from $D(x_i)$
- $\text{IncMin}(x_i)$: when the minimum value of $D(x_i)$ increases
- $\text{DecMax}(x_i)$: when the maximum value of $D(x_i)$ decreases
- $\text{Instantiate}(x_i)$: when $D(x_i)$ becomes a singleton

With such a differentiation of events, the rule **R3** in the constraint of Example 7 will be put in the list of rules to be woken only when the event $\text{Instantiate}(z)$ is set to true. For instance, if 1 and 3 are removed from $D(z) =$

$\{1, 2, 3, 4\}$, only $\text{RemValue}(z)$ and $\text{IncMin}(z)$ are set to true, thus no rule from constraint $\text{alldifferent}(x, y, z)$ is put in the propagation list. If 1, 2, and 3 are removed from $D(z) = \{1, 2, 3, 4\}$, $\text{Instantiate}(z)$ is set to true. As a result, rule **R3** is put in the propagation list to be woken.

I have briefly presented the basics of what is common to most solvers. However, the art of designing constraint propagators is not a mature science yet, and things can differ from one solver to another. More information can be found in academic publications [43–48] and in manuals of constraint solvers (see also Chapter 14 in Part II of Ref. 49).

GLOBAL CONSTRAINTS

There are “constraint patterns” that are ubiquitous when trying to express real problems as constraint networks. For example, we often need to say that a set of variables must all take different values. The size of the pattern is not fixed, that is, there can be any number of variables in the set. The alldifferent constraint, as introduced in CHIP [50], is not a single constraint but a whole class of constraints. Any constraint specifying that its variables must all take different values is an alldifferent constraint. The conventional wisdom is to name “global constraints” these classes of constraints that are defined by a Boolean function whose domain contains tuples of values of any length. An instance c of a given global constraint is a constraint with a fixed scheme of variables which contains all tuples of length $|X(c)|$ accepted by the function defining the global constraint. Beldiceanu *et al.* proposed an extensive list of global constraints [51].

Example 8. The $\text{alldifferent}(x_1, \dots, x_n)$ global constraint is the class of constraints that are defined on any sequence of n variables, $n \geq 2$, such that $x_i \neq x_j$ for all $i, j, 1 \leq i, j \leq n, i \neq j$. The $\text{NValue}(y, [x_1, \dots, x_n])$ global constraint is the class of constraints that are defined on any sequence of $n + 1$ variables, $n \geq 1$, such that the number of distinct values taken by $[x_1, \dots, x_n]$ must be equal to y , that is, $|\{x_i | 1 \leq i \leq n\}| = y$ [52, 53].

It is interesting to incorporate global constraints in constraint solvers so that users can use them to express the corresponding constraint pattern easily. Because these global constraints can be used with a scheme of any size, it is important to have a way to propagate them without using generic arc consistency algorithms. (Remember that optimal generic arc consistency algorithms are in $O(erd^r)$ for constraints involving r variables—see the section titled “Arc Consistency.”)

The first alternative to the combinatorial explosion of generic algorithms for GAC on a global constraint is to decompose it with “simpler” constraints. A *decomposition* of a global constraint G is a *polynomial time* transformation δ_k (k being an integer) that, given any network $N = (X(c), D, \{c\})$ where c is an instance of G , returns a network $\delta_k(N)$ such that $X(c) \subseteq X_{\delta_k(N)}$, $D(x_i) = D_{\delta_k(N)}(x_i)$ for all $x_i \in X(c)$, $|X(c_j)| \leq k$ for all $c_j \in \bar{C}_{\delta_k(N)}$, and $\text{sol}(N)$ is equal to the projection of $\text{sol}(\delta_k(N))$ on $X(c)$. That is, transforming N in $\delta_k(N)$ means replacing c by some new bounded arity constraints (and possibly new variables), whereas preserving the set of tuples allowed on $X(c)$. Note that by definition, the domains of the additional variables in the decomposition are necessarily of polynomial size.

Example 9. The global constraint $\text{atmost}_{p,v}(x_1, \dots, x_n)$ holds if and only if at most p variables in $x_1, \dots, x_n$ take value v [54]. This constraint can be decomposed with $n + 1$ additional variables $y_0, \dots, y_n$. The transformation involves the constraint $(x_i = v \wedge y_i = y_{i-1} + 1) \vee (x_i \neq v \wedge y_i = y_{i-1})$ for all $i, 1 \leq i \leq n$, and the domains $D(y_0) = \{0\}$ and $D(y_i) = \{0, \dots, p\}$ for all $i, 1 \leq i \leq n$.

It is desirable to design decompositions that *preserve* GAC on the original global constraint. That is, given any instance c of a global constraint G and any initial domain D on $X(c)$, given the corresponding decomposition, we want that for any domain included in D , GAC applied on the decomposition prunes exactly the same values as GAC on c . $\text{atmost}_{p,v}$ is a global constraint

that admits a decomposition preserving GAC (see Example 9).

Unfortunately, it is not always possible to find a decomposition preserving GAC. In Ref. 55, Bessiere *et al.* use tools of computational complexity to decide when a given global constraint has no chance to allow a decomposition preserving GAC. If enforcing GAC on a global constraint G is NP-hard, there does not exist any decomposition that preserves GAC (assuming $P \neq NP$). For instance, enforcing GAC on NValue is NP-hard. This tells us that there is no way to find a decomposition on which GAC always removes all GAC inconsistent values of the original NValue constraint.

Recently, Bessiere *et al.* have shown that NP-hard constraints are not the only class of global constraints for which no decompositions preserving GAC exist. They used circuit complexity theory to show that if a global constraint expresses a pattern that cannot be encoded as a monotone circuit of polynomial size then it cannot be decomposed even if it is not NP-hard to enforce GAC on it [56]. A famous example is the alldifferent constraint. For those constraints, the solution is to build a specialized algorithm that enforces GAC in polynomial time. For instance, there is a strong relationship between GAC on the alldifferent constraint and the problem of finding maximal matchings in a bipartite graph [57,58]. This relationship was used by Régim to propose a polynomial algorithm for GAC on the alldifferent constraint.

REFERENCES

1. Davis M, Putnam H. A computing procedure for quantification theory. *J ACM* 1960;7:201–215.
2. Bessiere C. Constraint propagation. In: Rossi F, van Beek P, Walsh T, editors. *Handbook of constraint programming*. Elsevier; 2006. Chapter 3.
3. Mackworth AK. Consistency in networks of relations. Technical Report 75–3. Vancouver: Department of Computer Science, University of B.C.; 1975. (also in *Artif Intell* 1977;8:99–118).
4. Mackworth AK. On reading sketch maps. *Proceedings IJCAI'77*. Cambridge (MA); 1977. pp. 598–606.
5. Mackworth AK, Freuder EC. The complexity of some polynomial network consistency algorithms for constraint satisfaction problems. *Artif Intell* 1985;25:65–74.
6. Mohr R, Henderson TC. Arc and path consistency revisited. *Artif Intell* 1986;28:225–233.
7. Mohr R, Masini G. Good old discrete relaxation. *Proceedings ECAI'88*. Munchen, FRG; 1988. pp. 651–656.
8. Bessiere C, Cordier MO. Arc-consistency and arc-consistency again. *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI'93)*. Washington (DC); 1993. pp. 108–113.
9. Bessiere C. Arc-consistency and arc-consistency again. *Artif Intell* 1994;65:179–190.
10. Bessiere C, Régim JC, Yap RHC, *et al.* An optimal coarse-grained arc consistency algorithm. *Artif Intell* 2005;165:165–185.
11. Lecoutre C, Hemery F. A study of residual supports in arc consistency. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. Hyderabad; 2007. pp. 125–130.
12. Bessiere C, Freuder EC, Régim JC. Using constraint metaknowledge to reduce arc consistency computation. *Artif Intell* 1999;107:125–148.
13. Freuder EC. Synthesizing constraint expressions. *Commun ACM* 1978;21(11):958–966.
14. Freuder EC. A sufficient condition for backtrack-free search. *J ACM* 1982;29(1):24–32.
15. Montanari U. Networks of constraints: fundamental properties and applications to picture processing. *Inf Sci* 1974;7:95–132.
16. Freuder EC. A sufficient condition for backtrack-bounded search. *J ACM* 1985;32(4):755–761.
17. Janssen P, Jégou P, Nougier B, *et al.* A filtering process for general constraint-satisfaction problems: achieving pairwise-consistency using an associated binary representation. *Proceedings of the IEEE Workshop on Tools for Artificial Intelligence*. Fairfax (VA); 1989. pp. 420–427.
18. Beer C, Fagin R, Maier D, *et al.* On the desirability of acyclic database schemes. *J ACM* 1983;30:479–513.
19. Dechter R, Pearl J. Tree clustering for constraint networks. *Artif Intell* 1989;38:353–366.
20. Gyssens M. On the complexity of join dependencies. *ACM Trans Database Syst* 1986;11(1):81–108.

21. Jégou P. Contribution à l'étude des problèmes de satisfaction de contraintes: algorithmes de propagation et de résolution; propagation de contraintes dans les réseaux dynamiques [PhD thesis]. CRIM: University Montpellier II; 1991. in French.
22. Jégou P. On the consistency of general constraint-satisfaction problems. Proceedings AAAI'93. Washington (DC); 1993. pp. 114–119.
23. Dechter R, van Beek P. Local and global relational consistency. *Theor Comput Sci* 1997;173(1):283–308.
24. Berlandier P. Improving domain filtering using restricted path consistency. Proceedings IEEE Conference on Artificial Intelligence and Applications (CAIA'95). 1995.
25. Freuder EC, Elfe CD. Neighborhood inverse-consistency preprocessing. Proceedings AAAI'96. Portland (OR); 1996. pp. 202–208.
26. Debruyne R, Bessière C. From restricted path consistency to max-restricted path consistency. Proceedings of the 3rd International Conference on Principles and Practice of Constraint Programming (CP'97), LNCS 1330. Linz: Springer; 1997. pp. 312–326.
27. Martin P, Shmoys DB. A new approach to computing optimal schedules for the job-shop scheduling problem. Volume 1084, Proceedings 5th International Conference on Integer Programming and Combinatorial Optimization (IPCO'96), LNCS. Vancouver (BC): Springer; 1996. pp. 389–403.
28. Lhomme O. Consistency techniques for numeric csp. Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93), pp. 232–238, Chambéry, France, 1993.
29. Freeman JW. Improvements to propositional satisfiability search algorithms [PhD thesis]. Philadelphia (PA): University of Pennsylvania; 1995.
30. Li CM, Anbulagan. Heuristics based on unit propagation for satisfiability problems. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI'97). Nagoya, Japan; 1997. pp. 366–371.
31. Debruyne R, Bessière C. Some practicable filtering techniques for the constraint satisfaction problem. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI'97). Nagoya; 1997. pp. 412–417.
32. Prosser P, Stergiou K, Walsh T. Singleton consistencies. Proceedings of the 6th International Conference on Principles and Practice of Constraint Programming (CP'00), LNCS 1894. Singapore: Springer; 2000. pp. 353–368.
33. Debruyne R, Bessière C. Domain filtering consistencies. *J Artif Intell Res* 2001;14:205–230.
34. Barták R, Erben R. A new algorithm for singleton arc consistency. Proceedings FLAIRS'04. Miami Beach (FL): AAAI Press; 2004.
35. Bessière C, Debruyne R. Theoretical analysis of singleton arc consistency and its extensions. *Artif Intell* 2008;172(1):29–41.
36. Lecoutre C, Cardon S. A greedy approach to establish singleton arc consistency. Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05). Edinburgh, Scotland; 2005. pp. 199–204.
37. Choi CW, Harvey W, Lee JHM, *et al.* Finite domain bounds consistency revisited. Dec 2004. Available at <http://arxiv.org/abs/cs.AI/0412021>.
38. Montanari U, Rossi F. Constraint relaxation may be perfect. *Artif Intell* 1991;48:143–170.
39. Benhamou F, McAllester DA, Van Hentenryck P. Clp(intervals) revisited. Proceedings of the International Symposium on Logic Programming (ILPS'94). Ithaca (NY); 1994. pp. 124–138.
40. Benhamou F, Older W. Applying interval arithmetic to real, integer and boolean constraints. *J Logic Program* 1997;32:1–24.
41. Apt KR. The essence of constraint propagation. *Theor Comput Sci* 1999;221(1–2):179–210.
42. Apt KR. Principles of constraint programming. Cambridge University Press; 2003.
43. Laurière JL. A language and a program for stating and solving combinatorial problems. *Artificial Intelligence* 1978;10:29–127.
44. Mohr R, Masini G. Running efficiently arc consistency. In: Ferraté G, *et al.* editors. Syntactic and structural pattern recognition. Berlin: Springer; 1988. pp. 217–231.
45. Van Hentenryck P, Deville Y, Teng CM. A generic arc-consistency algorithm and its specializations. *Artif Intell* 1992;57:291–321.
46. Van Hentenryck P, Saraswat VA, Deville Y. Design, implementation, and evaluation of the constraint language cc(FD). *J Logic Program* 1998;37(1-3):139–164.
47. Laburthe F, Ocre. Choco : implémentation du noyau d'un système de contraintes.

- Proceedings JNPC'00. Marseilles, France; 2000. pp. 151–165.
48. Schulte C, Stuckey PJ. Speeding up constraint propagation. Proceedings of the 10th International Conference on Principles and Practice of Constraint Programming (CP'04), LNCS 3258. Toronto, Canada: Springer; 2004. pp. 619–633.
 49. Rossi F, van Beek P, Walsh T, editors. Handbook of constraint programming. Elsevier; 2006.
 50. Dincbas M, Van Hentenryck P, Simonis H, *et al.* The constraint logic programming language chip. Proceedings of the International Conference on 5th Generation Computer Systems. Tokyo, Japan; 1988. pp. 693–702.
 51. Beldiceanu N, Carlsson M, Rampon JX. Global constraint catalog. Technical Report T2005:08. Kista, Sweden: Swedish Institute of Computer Science; 2005.
 52. Pacht F, Roy P. Automatic generation of music programs. Proceedings of the 5th International Conference on Principles and Practice of Constraint Programming (CP'99), LNCS 1713. Alexandria (VA): Springer; 1999. pp. 331–345.
 53. Beldiceanu N. Pruning for the minimum constraint family and for the number of distinct values constraint family. Proceedings of the 7th International Conference on Principles and Practice of Constraint Programming (CP'01), LNCS 2239. Paphos, Cyprus: Springer; 2001. pp. 211–224.
 54. Van Hentenryck P, Deville Y. The cardinality operator: a new logical connective for constraint logic programming. Proceedings ICLP'91. Paris, France; 1991. pp. 745–759.
 55. Bessiere C, Hebrard E, Hnich B, *et al.* The complexity of reasoning with global constraints. Constraints 2007;12(2):239–259.
 56. Bessiere C, Katsirelos G, Narodytska N, *et al.* Circuit Complexity and Decompositions of Global Constraints. Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09). Pasadena (CA); 2009. pp. 412–418.
 57. Knuth DE, Raghunathan A. The problem of compatible representatives. SIAM J Discr Math 1992;5(3):422–427.
 58. Régis JC. A filtering algorithm for constraints of difference in CSPs. Proceedings AAAI'94. Seattle (WA); 1994. pp. 362–367.

BASIC CP THEORY: SEARCH

LAURENT MICHEL
Computer Science and Engineering,
University of Connecticut, Storrs,
Connecticut

PASCAL VAN HENTENRYCK
Computer Science, Brown University,
Providence, Rhode Island

Constraint programming (CP) has, from its inception, adhered to the following basic tenet

$$\text{CP} = \text{Modeling} + \text{Search},$$

which puts as much emphasis on the ability to specify how to explore the search space as it does on how to model the constraints that any solution must satisfy. Separating the constraint model from the search fosters good modeling practices where one can extend the model with little to no concern for the search and vice versa. Indeed, the addition of new constraints should not force users to reconsider their search to exploit the enhancements. The benefits ought to be transparent. Specifying the search proves instrumental when facing hard problems where one wishes to exploit the semantics of the application to engineer a search procedure that is more effective.

Operations research practitioners well versed in mathematical optimization are accustomed to the idea of developing solutions almost exclusively through modeling. The extent of the control they can exercise on the solution process is through a selection of high level parameters for the solver at hand (e.g., the many parameters of the CPLEX solver). CP offers a different perspective on this issue, and its flexibility and ability to deal with hard problems often stems from sophisticated search procedures that exploit semantics of the problems that are hard or near impossible to model through traditional constraints. The most resounding success stories in scheduling [1], bounded program and model verification [2,3], distributed

system deployment [4], test generation [5], and sport scheduling [6–8], to name just a few, invariably take advantage of the ability to specify the search.

While *black-box* search procedures are undeniably desirable, as they offer a first generic search option that can work without further ado, retaining the key advantage to replace these generic procedures with user-defined ones is critical. The fundamental design challenge for CP languages and toolkits is to enable the specification of a user-definable search procedure while striking the right balance between three conflicting objectives, namely:

1. The clarity of the specification,
2. The flexibility of the specification to accommodate arbitrary levels of control, and
3. The efficiency of the resulting search.

The purpose of this article is to offer insights into what a CP search does, how orthogonal aspects of the search can be clearly and easily specified, and hint at how the underlying engines achieve this result. For clarity's sake, the article illustrates all the ideas with the COMET optimization platform [9,10], as it offers a simple, versatile, and yet powerful search specification language.

The rest of this article is organized as follows. The section titled “Fundamentals” presents the computational model underlying the execution of a search procedure. It presents, at a high level, how the constraint propagation meshes with the search process and provides a simple semantics for what goes on during the search. The section titled “Driving Example: The Steel Mill Slab” presents the driving example that will be used throughout the article. The section titled “Heuristics” introduces the notion of heuristics and how they shape the search tree that embodies the entire search space. The section titled “Strategies” discusses the role of strategies, that is, how to explore the search tree induced by

the heuristics. The section titled “Meta Strategies” touches on metastrategies, that is, the means to alter the overall behavior of the search process through techniques like randomization, restarts [11], or large neighborhood search (LDS) [12]. Finally, the last section concludes the article.

FUNDAMENTALS

CP is typically used to solve *constraint satisfaction problems* (CSP for short) specified with a triplet $P = \langle X, D, C \rangle$ where X is the set of decision variables, D_i is the domain of variable $x_i \in X$ (a finite set of discrete values, typically the integers), and C is a set of constraints over variables from X . $\text{vars}(c)$ denotes the subset of variables from X that appear in constraint c while $|\text{vars}(c)|$ is the arity of c .

Definition 1 [Infeasibility]. A CSP $P = \langle X, D, C \rangle$ is *failed* (*failed*(P)) if and only if $\exists x_i \in X$ s.t. $D_i = \emptyset$.

Definition 2 [Solution]. A CSP $P = \langle X, D, C \rangle$ is *solved* (*solved*(P)) if and only if $\forall x_i \in X : |D_i| = 1$ and C is true with respect to D .

Definition 3 [Solution set]. The solution set of $P = \langle X, D, C \rangle$, written $S(P)$, is $\{T = \langle X, [\{s_0\}, \{s_1\}, \dots, \{s_{n-1}\}], C \rangle$ s.t. $s_0 \in D_0 \wedge s_1 \in D_1 \wedge \dots \wedge s_{n-1} \in D_{n-1} \wedge \text{solved}(T)\}$ (where $|X| = n$).

CP can also accommodate *constraint optimization problems* (COP for short) where an additional objective function $f : X \rightarrow \mathbb{N}$ provides a quality measure for any solution. Optimization problems are, however, dealt with via a reduction back to satisfaction problems. Indeed, a COP $\langle X, D, C, f \rangle$ can first be turned into a CSP $\langle X, D, C \rangle$. A search for all the solutions of this CSP would yield a sequence of solutions $(s_0, f_0), \dots$, where f_0 denotes the value of the function f in s_0 . Given s_0 , one can easily produce and solve a subsequent CSP $\langle X, D, C \cup \{f < f_0\} \rangle$. This new CSP is either infeasible, in which case s_0 is guaranteed to be a global optimum, or it will deliver a solution (s_1, f_1) . The process can be repeated until the k th CSP

becomes infeasible and the process delivers the global optimum s_{k-1} . Therefore, and without loss of generality, the rest of the article focuses exclusively on constraint satisfaction.

Constraint Propagation

As discussed in the article titled *Basic CP Theory: Consistency and Propagation (Advanced)* in this encyclopedia, constraint propagation is a contraction operation which, given a CSP $\langle X, D, C \rangle$, yields new domains D' such that $\forall x_i \in X : D'_i \subseteq D_i$. The contraction is achieved by enforcing a local consistency property on the variables. Generalized arc-consistency (GAC) [13–15] is the most prevalent type of consistency and can be expressed as follows: The CSP $\langle X, D, C \rangle$ is GAC if and only if every constraint $c \in C$ is GAC. A constraint c (of arity k) is GAC if and only if it is GAC with respect to all its variables. Finally, c is GAC with respect to variable $x_j \in \text{vars}(c)$ if and only if

$$\forall v \in D_j, \exists w_0 \in D_0, \exists w_1 \in D_1, \dots, \exists w_{k-1} \in D_{k-1} : c(w_0, w_1, \dots, w_{j-1}, v, w_{j+1}, \dots, w_{k-1}).$$

The prune algorithm uses the GAC definition to eliminate values from D_j that cannot appear in a solution tuple to the constraint c . This contraction process can have three outcomes. First, the contraction can yield new domains D' such that *failed*($\langle X, D', C \rangle$) is true. An empty domain is a manifestation of an inconsistent constraint set. Second, the contraction can yield new domains D' satisfying *solved*($\langle X, D', C \rangle$), in which case the propagation has produced a solution. Finally, the contraction can yield new domains D' where at least one variable has two or more values left in its domain, which indicates that the propagation is inconclusive and the CSP could have zero, one, or more solutions.

Decomposition

When the propagation is inconclusive, it becomes necessary to *branch* to decompose the instance into smaller instances for which the propagation might be able to determine

if the outcome is infeasible, a solution, or in need of further decomposition.

Definition 4 [Branching]. A branching decision is a sequence of k ($k \geq 2$) constraints $[c_0, \dots, c_{k-1}]$ that decomposes a CSP instance $P = \langle X, D, C \rangle$ into k subinstances $P_{j \in 0..k-1} = \langle X, D, C \cup \{c_j\} \rangle$ forming a partition, that is, no solutions are lost: $S(P) = \bigcup_{i=0}^{k-1} S(P_i)$, and the subinstances are nonoverlapping $\forall i, j \in 0..k-1, i \neq j : S(P_i) \cap S(P_j) = \emptyset$.

Example 1 [Binary Branching]. A very simple branching decision is $[x = v, x \neq v]$. Indeed, Given $P = \langle X, D, C \rangle$ and $[x = v, x \neq v]$ for $x \in X$ and $v \in D(x)$, one easily derives $P_0 = \langle X, D, C \cup \{x = v\} \rangle$ and $P_1 = \langle X, D, C \cup \{x \neq v\} \rangle$. The two problems yield a partition as all solutions must have their x either equal to v or different from it. It also yields a nonoverlapping partition since a solution cannot simultaneously satisfy $x = v$ and $x \neq v$.

Once an instance is decomposed, the propagation algorithm (prune) is applied to each subinstance to further contract the domains and eliminate the values that are no longer consistent.

Search Tree

Given a propagation algorithm and a decomposition scheme, it is possible to define a search tree.

Definition 5 [Search Tree]. A search tree for a CSP $P = \langle X, D, C \rangle$ is an ordered k -ary tree rooted at P and defined inductively as follows:

$$tree(P) = failNode(P) \Leftrightarrow failed(P)$$

$$tree(P) = solNode(P) \Leftrightarrow solved(P)$$

$$tree(P) = intNode([P_0, P_1, \dots, P_{k-1}]) \Leftrightarrow$$

$$\begin{cases} \neg failed(P) \wedge \\ \neg solved(P) \wedge \\ branching(P) = [c_0, \dots, c_{k-1}] \wedge \\ \forall i \in 0..k-1 : P_i = \langle X, D, C \cup \{c_i\} \rangle, \end{cases}$$

where $failNode(P)$ and $solNode(P)$ denote leaves of the tree and $intNode([P_0, \dots, P_{k-1}])$ is an internal node. A pictorial representation is shown in Fig. 1. The ordered tree $tree(P)$ is a by-product of the initial CSP and the branching rules (which produces the branching decisions) in effect. Clearly, the branching rule can inspect the CSP P to produce a suitable branching decision (i.e., an ordered sequence of constraints).

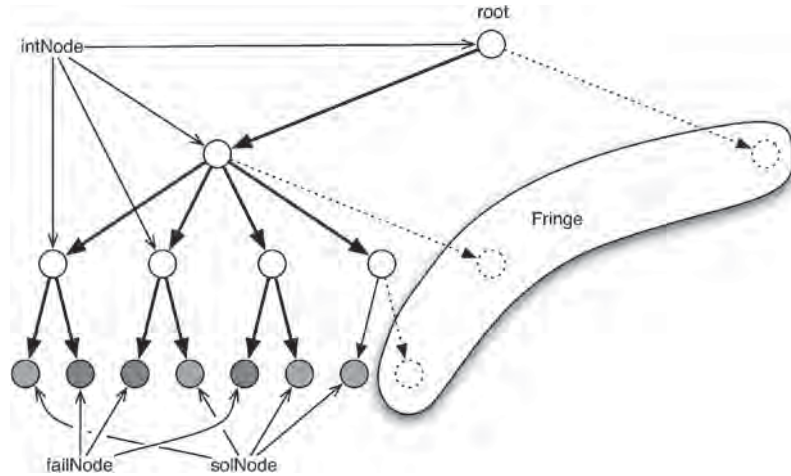


Figure 1. Search tree.

The topology of the tree depends directly on the branching decision which is commonly known as the *variable/value selection* heuristic. The search tree, when rooted at P , structures the entire search space to facilitate the search process per se.

Search Strategy

Given a search tree T for a CSP $P = \langle X, D, C \rangle$, a search *algorithm* carries out the exhaustive exploration of T . The exploration can adopt a traditional technique like depth-first search where the algorithm always first expand the leftmost nonvisited node of the tree recommended by the heuristic. However, the tree exploration is not limited to this *strategy*. Instead, it could elect to first expand any other nonvisited node currently present in the “fringe” (unexpanded region) of the tree.

Richer options include best-first search, limited discrepancy search (LDS), or iterative deepening to name just a few. The motivation to select one strategy or another is intimately linked with the effectiveness of the branching decision. Indeed, if an oracle is available as a branching decision, it will always produce an ordering of the branching constraints that places the solution of the problem on a leftmost path in the search tree, in which case a simple depth-first strategy is perfect. On the other hand, if the branching heuristic is weaker, a more sophisticated node selection strategy can compensate by exploring paths that are distributed throughout the search tree and jump from one node of the fringe to another as the expansions proceed. It could also recognize that the latest failure it encountered is caused by one or more decisions (known as *nogoods* [16]) that were made at shallower depths in the search and jump back over several branching decisions that are not causally related to reconsider what caused the failure. The finding of nogoods can occur prior to the search [17], but more importantly, it can happen dynamically as the search proceeds [18–20]. Nonchronological strategies of this type include, for instance, backjumping [16,21] and conflict-directed backjumping [19]. Readers interested

```

1 search (in  $P = \langle X, D, C \rangle$ ) return  $S$  {
2    $Q = \{P\}$ ;
3    $S = \emptyset$ ;
4   while  $Q \neq \emptyset$  {
5      $I = \text{selectAndExtract}(Q)$ ;
6      $I = \text{prune}(I)$ ;
7     if  $\text{failed}(I)$  then continue;
8     if  $\text{solved}(I)$  then  $S = S \cup \{I\}$ ;
9     else {
10       $[c_0, c_1, \dots, c_{k-1}] = \text{branch}(I)$ ;
11      let  $\langle X, D, C \rangle = I$ 
12         $P_i = \langle X, D, C \cup \{c_i\} \rangle \quad \forall i \in 0..k-1$ 
13      in  $Q = Q \cup \{P_0, P_1, \dots, P_{k-1}\}$ ;
14    }
15  }
16  return  $S$ ;
17 }
```

Figure 2. The search procedure.

in a more in-depth account can consult Rossi *et al.* [22].

The search tree that structures the entire search space is never created explicitly. Instead, the search algorithm maintains a subset of its node, the *fringe* (shown in Fig. 1), representing the boundary between examined and unexplored nodes. Initially, the fringe consist of the root of the search tree ($\text{intNode}(P)$). The search algorithm then repeatedly extracts subinstances from the fringe, propagates the instance and, depending on the outcome, it either records a solution (when $\text{solved}(P)$ is true), discards the node (when $\text{failed}(P)$ is true), or decomposes it and adds the newly generated subinstances back into the fringe. The search strategy governs which node is extracted from the fringe, and that decision affects the performance of the search as well as the memory requirements.

The generic search algorithm, shown in Fig. 2, embodies the core of a search procedure. The selection of different *heuristics* and *strategies* translates into an instantiation of the branch and selectAndExtract subroutines, respectively.

CP languages, such as COMET [9], implement this algorithm and provide the necessary mechanism to specify heuristics and strategies with great ease. The COMET specification of the search heuristics is done directly when describing the search tree through two

important control primitives (**try** and **tryall**). The COMET specification of search strategies is achieved via *search controllers* [23,24] that are orthogonal to the search tree specification. Search controllers are responsible for representing and manipulating the fringe.

DRIVING EXAMPLE: THE STEEL MILL SLAB

The search mechanisms at our disposal will be illustrated throughout the paper on an optimization problem known as the *Steel Mill Slab* production problem. The problem is described in Ref. 25, and early CP models were introduced in Ref. 26.

Steel is produced by casting molten iron into slabs. A steel mill can only produce a finite number of slab capacities (lengths). An order has two properties: a color corresponding to the route required through the steel mill, and a length. Given n input orders, the problem is to assign the orders to slabs such that the total size of steel produced is minimized. This assignment is subject to capacity constraints, that is, the total length of orders assigned to a slab cannot exceed the largest slab capacity. It is also subject to coloring constraints, that is, each slab can contain at most two colors. The color constraints arise because it is expensive to cut up slabs in order to send them to different parts of the mill. The basic model is shown in Fig. 3. The

model starts by defining ranges for the customer orders and the steel slabs. It stores in `colorOrder c` the set of orders that require color c . Line 4 stores in `maxCap` the longest slab that can be produced. Line 5 computes the losses one would incur to produce a slab of length at least c , that is, it is the extra length of steel that would be wasted to reach one of the legal capacities. Lines 7–9 define the solver and the decision variables, namely, $x[i]$ denotes the slab number that order i is assigned to and $l[i]$ is the length of the i th steel slab. The model states the objective function on-line 11: minimize the total losses incurred, and the constraints on lines 13–15, namely: define the length l_i of slab i in term of the length of the orders assigned to it and prohibit more than two colors on the same slab. The search specification would appear inside the **using** block starting at line 16.

HEURISTICS

Heuristics control the topology of the search tree and take the form of a sequence of k constraints meant to decompose the CSP instance. This section provide an overview of the major search heuristics and how one can specify them in COMET. Branching heuristics exert their function through two mechanisms, namely, which variable to branch on and in what order values should be considered for the selected variable.

```

1 int nbOrders = sz; range Orders = 1..nbOrders;
2 int nbSlabs = nbOrders; range Slabs = 1..nbSlabs;
3 set{int} colorOrders[c in Colors] = filter (o in Orders) (color[o] == c);
4 int maxCap = max(i in Caps) capacities[i];
5 int loss[c in 0..maxCap] = min(i in Caps: capacities[i] >= c) capacities[i] - c;
6
7 Solver<CP> m();
8 var<CP>{int} x[Orders](m,Slabs);
9 var<CP>{int} l[Slabs](m,0..maxCap);
10
11 minimize<m> sum(s in Slabs) loss[l[s]]
12 subject to {
13     m.post(multiknapsack(x,weight,l));
14     forall(s in Slabs)
15         m.post(sum(c in Colors) (or(o in colorOrders[c]) (x[o] == s)) <= 2);
16 } using...
```

Figure 3. The steel mill slab problem.

The *variable selection heuristic* is responsible for picking the variable to branch on. Many options are possible ranging from a purely static ordering to a completely dynamic ordering (i.e., dependent on which node of the search tree we are currently on). While the generic branching scheme allows for arbitrary branching constraints, all the heuristics in use select a single variable.

The *value selection heuristic* depends upon the class of branching constraints under consideration. Two common schemes exist: *enumeration* and *splitting*.

Enumeration. The most common approach is to enumerate the values in the domain of variable x_i (the chosen variable) by producing a sequence of constraints $[x_i = v_0, x_i = v_1, \dots, x_i = v_{k-1}]$ where the v 's are values from D_i , and the ordering of the values drives the ordering of the branches in the search.

Splitting. A splitting scheme interprets D_i as a range $L_i..U_i$ where L_i and U_i are the smallest (respectively largest) values in D_i and divides the domain in two halves at $(L_i + U_i)/2$ with the branching $[x_i \leq (L_i + U_i)/2, x_i \geq 1 + (L_i + U_i)/2]$. This gives rise to a binary tree where the same variable is considered multiple times at different depths as its domain is repeatedly split.

The variable and the value selection heuristics have at their disposal the current CSP instance and can examine any property to produce their recommendation. This

```

1  ...
2  } using {
3    forall(i in Orders)
4      tryall<m>(v in x[i].getMin()..x[i].getMax() : x[i].memberOf(v))
5        m.post(x[i] == v);
6    cout << x << endl;
7  }
```

At the first iteration of line 3, $i = 1$. The **tryall** instruction on-line 4 picks the first value in $D_1 = \{1..nbSlabs\}$, namely, 1 and posts the constraints $x[1] == 1$ on-line 5. The addition of the constraint triggers a round of propagation. When the propagation ends, the

section first considers *local heuristics* that base their recommendations on information related exclusively to the variable under consideration. The second section focuses on *global heuristics* that assess the effect of a choice on the entire CSP instance. Global heuristics attempt to be less myopic at the expense of significant efforts.

Local Heuristics

The most common heuristics use local information pertaining to variables alone. The heuristics shown below start from the simplest ideas and then increase in sophistication.

Static Ordering. The simplest local heuristic is static ordering. It simply adopts a fixed ordering of the variables and values. Namely, at depth k , the branching will consider variable x_k , as all the variables between depth 0 and depth $k - 1$ have already been fixed. The value selection heuristic at depth k will consider the domain D_k as an ordered set of discrete values (the natural order for the underlying type, that is, $<$ for the integers). The resulting branching decision at depth k is therefore $[x_k = v_0, \dots, x_k = v_{n-1}]$ where $D_k = \{v_0, v_1, \dots, v_{n-1}\}$ is an ordered set of n values, that is, $v_0 < v_1 < \dots < v_{n-1}$.

While static ordering is trivial to specify and implement, it is clearly wholly ignorant of the current state and fails to steer the search in any particular direction. The following fragment illustrates how to specify this heuristic with COMET for the steel mill slab problem.

execution continues and moves to the second iteration of line 3, namely, $i = 2$. Note that other alternatives for $x[1]$ are still available and, should an inconsistency be detected, the search has the option to return to these internal search tree nodes to explore them. In

essence, the **tryall** is a lazy construction that does not expand all the siblings of a node until they are needed by the strategy. Which node to return to depends on the search strategy which will be discussed in the section titled “Strategies.”

```

1 forall(i in Orders : !x[i].bound())
2   tryall<m>(v in x[i].getMin().x[i].getMax() : x[i].memberOf(v))
3     m.post(x[i] == v);

```

Dynamic Variable Ordering. It is often advantageous to branch first on variables with the smallest domains to discover inconsistencies between choices as early as possible. The so-called *min-dom* heuristic

```

1 forall(i in Orders : !x[i].bound()) by (x[i].getSize())
2   tryall<m>(v in x[i].getMin().x[i].getMax() : x[i].memberOf(v))
3     m.post(x[i] == v);

```

implements this idea. Compared to the static ordering, it adds a **by** clause on the **forall** loop that requires the enumeration of the indices i in increasing order of the expression.

Notice how the search above first commits to a variable selection and subsequently attempts to find a value for the variable. The only way to consider another variable is to either successfully impose a constraint on the chosen variables or exhaust all possibilities

```

1 while(!bound(x)) {
2   selectMin(i in Orders : !x[i].bound())(x[i].getSize()) {
3     int v = x[i].getMin();
4     try<m> m.post(x[i] == v);
5       | m.post(x[i] != v);
6   }
7 }

```

embodies this simple idea. The outer loop (line 1) iterates as long as not all the variables in x are bound. The selector on line 2 selects the index i that identifies a free variable with the smallest domain. Line 3 picks the smallest value in its domain and lines 4–5 create a branching decision $[x_i = v, x_i \neq v]$. In case of failure, when the search algorithm selects the second alternative, it imposes

The static ordering shown above can be slightly improved by skipping the variables that may already be bound as a result of propagation. The fragment below illustrates the addition of a predicate on the **forall** loop to only consider variables that are not bound.

embodies the first-fail principle “To succeed, try first where you are most likely to fail” [27]. It is also easy to implement, as it suffices to order the variables according to their domain size. The fragment

and reconsider an earlier node in the search tree (i.e., backtrack to another node). It might be desirable to have a slightly more flexible setup. If the addition of a constraint $x_k == v$ with $v \in D_k$ leads to a failed CSP, the system has learned the negation of the above, namely, $x_k \neq v$. So it might be valuable to impose this new constraint, as its propagation may very well lead to a contraction of domains and a subsequently *better informed* variable selection. The fragment

the $x_i \neq v$ constraint and propagates the constraint system. Therefore, subsequent iterations of the outer loop can pick a different variable for the next choice.

Dynamic Value Ordering. The heuristics considered thus far enumerate the values in the domain of variable x_k in a fixed, static

order (from the smallest to the largest). Note how the ordering of the values impacts the order of the constraints in the branching sequence created at a node of the search tree. While the variable selection controlled the depth at which a variable is fixed, the value selection controls which path lies on the leftmost edge of the search tree.

```

1 forall(i in 1..n : !q[i].bound()) by (q[i].getSize()) {
2   tryall<m>(v in q[i].getMin()..q[i].getMax() : q[i].memberOf(v)) by (abs(v - n/2)) {
3     m.post(q[i] == v);
4   }
5 }

```

illustrates exactly this strategy. The variable selection is still based on the first-fail principle as it first selects the variable with the smallest domain. The value selection heuristic implemented by the ordering clause on the **tryall** instruction on-line 2 orders the value by increasing distance from the middle of the board with the $\text{abs}(v - n/2)$ expression. The use of this heuristic makes it possible to find the first solutions to much larger instances much faster (under 1 s for sizes up to 1024 queens with as few as five failures when the static value ordering cannot get past 128 queens in the same time.)

Degree Ordering. Variable selection based on the smallest domain (i.e., the *min-dom* heuristic) attempts to cause failures early on in the search. The same outcome could be the result of branching on variables that are subjected to many constraints. A first idea is to exploit the *static* or *dynamic* degree (respectively *deg* and *ddeg*) of a variable in the constraint graph. *deg* is the degree of the variable at the root of the search tree, whereas *ddeg* captures the degree during the search (variables that are fixed by branching decisions and/or propagation can be projected out leaving a smaller constraint graph). Intuitively, a variable that appears in many constraints may have a stronger effect than a variable that only appears in a handful of them. Consequently, branching first on variable with maximal degree seems appropriate.

Definition 6 [max-deg]. The most desirable variable to branch on is the

Example 2. Consider the simple problem of placing n queens on an $n \times n$ chess board so that no two queens attack each other. On this problem, the heuristics considered thus far have a tendency to start from a corner of the board. However, it might be advantageous to start from the middle of the board. The fragment

variable which maximizes its static degree *deg* in the constraint graph.

Definition 7 [max-ddeg]. The most desirable variable to branch on is the variable which maximizes its dynamic degree *ddeg* in the constraint graph.

Unsurprisingly, it is also possible to combine *min-dom* with either *max-deg* or *max-ddeg*. These ideas have been proposed to break the ties that appear with *min-dom*.

Definition 8 [dom/deg and dom/ddeg]. The most desirable variable to branch on is the variable which minimizes the ratio domain size over degree (respectively, dynamic degree) [28,29]. Note that lexicographic orderings based on (*min-dom*, *max-deg*) (respectively (*min-dom*, *max-ddeg*)) are also possible.

A further possibility is to take into account the propensity of variables to lead to conflicts. The resulting heuristics, named *wdeg* and *dom/wdeg* [30], are defined as follows:

Definition 9 [w_c]. The weight of constraint c is a nonnegative integer which, during the search, reflects how often the constraint c detected an inconsistency, that is, emptied the domain of one of its variables.

Definition 10 [wdeg]. The weighted degree of variable $x_i \in X$ is defined as

$$wdeg(x_i) = \sum_{c \in C : x_i \in vars(c) \wedge |utVars(c)| > 1} w_c,$$

where $utVars(c) = \{x \in vars(c) \setminus \{x_i\} \text{ s.t. } |D_x| \geq 2\}$, that is, it is the sum of the weights of the constraints that x_i appears in and that have at least one other uninstantiated variable.

Definition 11 [*dom/wdeg*]. The weighted domain heuristic of variable $x_i \in X$ is defined as the ratio $\frac{|D_i|}{wdeg(x_i)}$, that is, the most desirable variable to branch on has the smallest domain and appears in the most conflicting constraints.

The variable selection heuristic can then focus on first selecting the variable with the largest *wdeg* value, as it corresponds to variables appearing in the most constraining part of the CSP. On the other hand, selecting variables with the smallest *dom/wdeg* dictates the selection of variables according to the first-fail principle and tie-breaking according to *wdeg*.

Global Heuristics

Global heuristics attempt to exploit information that is more global to the entire search space in order to make a more informed decision as to which branching decision to make.

Impacts [31]. These are inspired by *pseudo-costs* [32] and strong branching [33] in integer programming which normally measure the effect of a branching decision on the objective function. Here, impacts offer a measure of a constraint effectiveness in reducing the size of the search space. Impacts consider branching constraints of the form $x_i = v$ with $v \in D_i$ which lead to the biggest reduction in the size of the search space.

The core idea of impacts is to estimate the size of the search tree as $P = |D_0| \times |D_1| \times \dots \times |D_{n-1}|$. The impact of a potential branching decision $x_i = v$ can be characterized as

```

1 while(!bound(x)) {
2   selectMin(i in 1..n : !x[i].bound())(variableImpactOf(x[i])) {
3     selectMax(v in x[i].getMin()..x[i].getMax() : x[i].memberOf(v))(assignmentImpactOf(x[i],v)) {
4       try <m> m.post(x[i] == v);
5         | m.post(x[i] != v);
6     }
7 }

```

$$I(x_i = v) = 1 - \frac{P_{\text{after}}}{P_{\text{before}}}.$$

Namely, if the contraction is very strong ($P_{\text{after}} \ll P_{\text{before}}$), the impacts tends toward 1. Refalo [31] observed that, on average, the impact of an assignment tends to be independent of other assignments and proposed maintaining a running average of the impact of $x_i = v$ over the tree nodes that have been treated from the beginning of the search. If K is the set of nodes seen so far, than the impact $\tilde{I}(x_i = v)$ is defined as

$$\tilde{I}(x_i = v) = \frac{\sum_{k \in K} I^k(x_i = v)}{|K|}$$

with $I^k(x_i = v)$ being the impact of the assignment in the tree node k . One can then easily derive a definition for the impact of a variable (regardless of which value it will be constrained to) and use it as a variable selection heuristic. Refalo's derivation in Ref. 31 yields the following definition:

$$\mathcal{I}(x_i) = \sum_{v \in D_i} 1 - \tilde{I}(x_i = v).$$

Observe that if all the values in D_i have high impact (close to 1), then $\mathcal{I}(x_i)$ tends toward 0. Therefore a variable that effectively contracts the search space is one that minimizes $\mathcal{I}(x_i)$. It should be noted that the impacts can easily be produced at the root of the search tree prior to the actual search via a simple preprocessing step. During the search, the running average of the impact of an assignment $x_i = v$ can be updated each time the constraint is considered for branching. The search strategy itself simply uses the variable and value impacts maintained in a global data structure to drive the variable/value selection processes. The fragment

shows the skeleton of an impact-based search (the initialization is omitted). The data structure behind `variableImpactOf` (that returns $\mathcal{I}(x_i)$) and `assignmentImpactOf` (that returns $\tilde{\mathcal{I}}(x_i = v)$) must also be updated each time a constraint is posted.

Singleton Consistency and Look-Aheads. Singleton consistencies [34] (e.g., singleton arc consistency, restricted singleton arc consistency) offer other opportunities to collect data about variables and assignments to derive a global heuristic. The core idea is to slightly enhance the singleton consistency algorithms to collect measurements of the impacts of the assignments being considered. The core idea is to estimate the size of the search tree for a problem P with n variables x_1 through x_n with domains D_1 through D_n as

$$\sigma(P) = \log \left(\prod_{i=1}^n |D_i| \right),$$

or equivalently

$$\sigma(P) = \sum_{i=1}^n \log |D_i|,$$

in which the estimate is the logarithm of the product of the domain sizes. Note the similarity with the impact-based estimation. Refalo's impacts rely on a measurement of the contraction that is derived from a tentative assignment, whereas Correia's measure is a

direct assessment of the tree size (after an assignment), not how it shrinks.

When the singleton consistency algorithm considers the subproblem $P_{|x_i=a} = P \cup \{x_i = a\}$ to decide whether value $a \in D_i$ can or cannot be discarded, it suffices to compute $\sigma(P_{|x_i=a})$ before proceeding to the next value. Once all the values in D_i are examined, the impact of the variable x_i is easily defined as

$$\mathcal{I}(x_i) = \sum_{a \in D_i} \sigma(P_{|x_i=a}).$$

A global variable selection heuristic can then easily be derived given that the variable associated to the smallest $\mathcal{I}(x_i)$ yields the smallest tree to investigate. Note that other variants can be found in Ref. 34. The value selection heuristic is modeled after the same idea. Once again, the best value is a value delivering the largest subtree and thus most likely to lead to a solution. Given a variable selection x_i , it is preferable to pick a value that produces the largest subtree. Therefore, the heuristic is to select the value $a \in D_i$ that maximizes $\max_{a \in D_i} \sigma(P_{|x_i=a})$.

Note that the $SC_d(\mathcal{X}, \mathcal{C})$ and $RSC_d(\mathcal{X}, \mathcal{C})$ routines (as defined in Ref. 34) are easily instrumented to collect the variable and value selection heuristics discussed above. In these routines, the parameter d controls the subset of variables for which the restricted consistency is checked as the set $\{x \in \mathcal{X} \text{ s.t. } |D_x| \leq d\}$. For instance, the fragment

```

1 class SCH {
2   var<CP>{int} theVar;
3   int theVal;
4   float estimate;
5   SCH() { theVar = null; }
6   void update(var<CP>{int} v,int value,float est) {
7     if (theVar==null || est < estimate) {
8       theVar = v;theVal = value;estimate = est;
9     }
10  }
11 }
12 function SCH RSCd(Solver<CP> m,int d) {
13   var<CP>{int}[] vars = m.getIntVariables();
14   SCH best();
15   forall(i in vars.getRange() : !vars[i].bound() && vars[i].getSize() <= d) {
16     Float vh(0);Float mv(0);
17     Integer bv(vars[i].getMin()-1);

```

```

18   forall(a in vars[i].getMin()..vars[i].getMax() : vars[i].memberOf(a)) {
19       bool b = succeeds(m) {
20           m.label(vars[i],a);
21           float ve = sum(i in vars.getRange()) log(vars[i].getSize());
22           vh := vh + ve;
23           mv := max(mv,ve);
24           if (mv == ve) bv := a;
25       };
26       if (!b) m.diff(vars[i],a);
27   }
28   best.update(vars[i],bv,vh);
29 }
30 return best;
31 }

```

is a complete COMET implementation of restricted singleton consistency which returns a tuple with the selected assignment for the next branching decision. The succeeds statement on-line 19 executes the block spanning lines 20 to 25 which does a tentative assignment $x_i = a$ (line 20) and returns true if this was a success and false otherwise. Line 21 computes $\sigma(P_{|x_i=a})$, while line 22 accumulates the sum of the estimates in vh and lines 23–24 track the maximum estimate mv and its corresponding value bv . Line 26 removes the value a from D_i whenever the tentative assignment $x_i = a$ leads to an inconsistency. Note that, if this removal of a from D_i leads to an inconsistency, the backtracking automatically will return to the last choice. The generic search procedure using $RSC_2(\mathcal{X}, \mathcal{C})$ can then be simply defined as

```

1 SCH sel = RSCd(m,2);
2 while (sel!=null) {
3     try <m> m.label(sel.theVar,sel.theVal);
4         | m.diff(sel.theVar,sel.theVal);
5     sel = RSCd(m,2);
6 }

```

Advanced Heuristics. The arsenal of heuristics available to tame hard problems grows regularly and recent advances include heuristics based on estimating the solution densities for constraints in subtrees to decide how to branch [35] or expectation maximization and belief propagation to find likely assignments [36].

```

1 forall(i in 0..n-1)
2     tryall <m> (v in x[i].getMin()..x[i].getMax() : x[i].memberOf(v))
3         m.post(x[i] == v);

```

STRATEGIES

Heuristics attempt to alter the shape of the tree by placing desirable paths on the leftmost edge of the tree so that the most direct strategy, depth first search, explores these paths first. If heuristics were perfect, this would be sufficient. However, heuristics remain educated guesses as to where to search. Sometimes, heuristics can lead the search astray and limit the ability to find a feasible solution early on. Search strategies can mitigate the weaknesses of heuristics by progressively reducing the algorithm trust in the heuristic and exploring other parts of the search tree. This section discusses two simple strategies, shows how these can be used in COMET, and shows how *search controllers* can be used to implement them.

Chronological

Depth-first search simply expand the deepest leftmost node of the fringe first. If the fringe is represented with a stack data structure, this is trivial to achieve as the last node pushed onto the stack is always the deepest leftmost node. Recall that the node management strategy is completely orthogonal to the tree specification. COMET preserves that separation by parameterizing the instruction responsible for generating nodes by an object responsible for managing them. The fragment

illustrates this separation. The **tryall** instruction is parameterized between the two angle brackets by a *search controller* [37] responsible for storing and managing the nodes. In this specific case, the solver *m* is used, as it simply delegates to its default search controller (a DFS controller). One could explicitly mandate DFS by installing a controller prior to the search as shown below.

```
1 Solver<CP>(m);
2 m.setSearchController(DFSController(m));
```

Controller. The reader has certainly noticed that several instructions in COMET are parameterized with a controller. The binary **try**, as well as **solve** and **solveall**, take a search controller as they either produce search nodes or govern their exploration. A controller is a class that conforms to the simple interface highlighted below.

```
1 interface SearchController
2   void start(Continuation c);
3   void exit();
4   void addChoice(Continuation c);
5   void fail();
6   ...
7 }
```

```
1 try<sc> (left) | (right)      into
```

The implementation creates a continuation *c*, transfers it to the search controller as a new choice, and then executes the left branch of the choice point. On backtracking, that is, when the search calls **fail** to return to a suspended tree node, the continuation *c* is invoked and the right branch is executed since *rb* is true in that context. The compilation of a **tryall** instruction is essentially similar but uses iterators to assign its parameter that is represented as a local variable. The key observation is that the **try** instruction produces an implicit representation of the search node and passes it (via a call to **addChoice**) to the controller. Returning to this node boils down to restoring the state of the solver and calling the continuation to resume.

It stores and manages the search tree nodes produced by branching instructions such as **try** and **tryall**. The flexibility of COMET in implementing search strategies stems from how branching instructions capture the current *state of the control* and entrust its management to the controller. COMET uses *continuations* [38] (common place in functional languages such as scheme [39]) which pair the local state (values of the local variables) with a particular point in the execution flow. The code

```
1 continuation c (body)
```

binds *c* to a continuation $\langle I, S \rangle$, where *I* is the next instruction in the code and *S* is the stack when the continuation is captured. It then executes its body and continues in sequence. The resulting continuation can be invoked at any point with the call **call(c)** which restores the stack *S* and restarts execution from *I*. The immediate benefit appears when considering a branching instruction such as **try**. Its implementation can be seen as the rewriting of

```
1 bool rb = true;
2 continuation c { sc.addChoice(c); rb = false; (left) }
3 if (rb) (right)
```

DFS Node Management Policy. Given the mechanics described above, a DFS controller must simply adhere to a last-in first-out policy when handling the nodes submitted via **addChoice** and explored via **fail**.

The core of the DFS implementation is shown in Fig. 4. The controller constructor simply keeps a reference to the solver and creates the two stacks to hold the continuations and the state descriptions. The **start** and **exit** method are married with the **solve** (and **solveall**) instructions. **start** is called when entering the **solve** and records a continuation representing where to go once the model is solved (i.e., the instruction following the **solve**). The **exit** method, which is called internally by the controller when it has exhausted all its pending nodes, simply

```

1 class DFS implements SearchController {
2   Continuation _exit;
3   Stack{Continuation} _ctrl;
4   Stack{Checkpoint} _state;
5   Solver<CP> _solver;
6   DFS(Solver<CP> m) { _solver = m; ...}
7   void start(Continuation e) { _exit = e;}
8   void exit() { call(_exit);}
9   void addChoice(Continuation c) {
10     _ctrl.push(c);
11     _state.push(new Checkpoint(_solver));
12   }
13   void fail() {
14     if (_ctrl.empty()) exit();
15     _state.pop().restore();
16     call(_ctrl.pop());
17   }
18 }

```

Figure 4. The DFS controller.

calls the exit continuation to resume the execution after the solve block. The **addChoice** method, which is called by the **try** (or the **tryall**), stores the control and the solver state in two stacks. When the search algorithm wishes to return to a pending node (e.g., after the detection of an inconsistency), it simply calls the **fail** method of the controller which pops the topmost node, restores its state, and returns the control to it.

The elegant separation of the search strategy in a controller makes it possible to easily change the strategy without a single modification of the main model.

Nonchronological

While DFS is a natural choice, it blindly follows the heuristics defining the search tree and cannot recover from poor choices made at a shallow depth in the search tree. Nonchronological search strategies offer a departure from the last-in/first-out policy of node management. As alluded to before, data structures like nogoods give rise to strategies that explicitly attempt to return to search nodes that undo at least one branching constraint blamed for the current failure. While interesting, a full treatment of nogoods is beyond the scope of this article; and a partial treatment can be found in Ref. 22. Instead, this section focuses on a strategy

explicitly designed to offset the weaknesses of the variable/value selection heuristic.

Depth-bounded discrepancy search (BDS) [40] is a search strategy that uses a limit on the number of discrepancies (disagreements with the heuristic) with a bias toward the discrepancies that occur at shallow depth in the tree. The strategy, which is related to LDS [41], explicitly attempts to compensate for the potential limitations of the search heuristic. Intuitively, BDS progressively reduces its trust in the heuristic. BDS operates through a number of passes over the search tree. In each pass, it has at its disposal a number of credits that it can spend to disagree with the heuristic and not follow the recommended left branch. In the first pass, BDS has zero credits and is therefore compelled to follow the leftmost path. Once at the bottom of the tree (and without any credits to spend), the first pass is over and the only way to proceed is to repeat the search with one more credit. Consequently, the second pass is authorized to disagree once (at any depth) with the heuristic. The first nodes being considered during the second pass are of course the shallowest in the tree. The process continues until the number of credits allows a full exploration of the search tree without hitting the limit.

Once again, the selection of the search strategy in COMET is Trivial, as it boils down to

```

1 Solver<CP>(m);
2 m.setSearchController(BDSController(m));

```

where the BDSController stores the search tree nodes in a queue data structure. Given that the branching decision produces a sequence (i.e., an ordered collection) of branching constraints, the branches are enqueued left to right by the **try** and **tryall** instructions, which guarantees that nodes with fewer discrepancies are always explored first and nodes enqueued at shallower depths will be the first to be dequeued in the next pass.

The implementation of the DBS is shown in Fig. 5. The essential change is a switch to a queue for the storage of the nodes in the fringe. Indeed, search nodes that are added

```

1 class BDSController implements SearchController {
2   Continuation _exit;
3   Queue<Continuation> _ctrl;
4   Queue<Checkpoint> _state;
5   Solver<CP> _solver;
6   BDSController(Solver<CP> m) { _solver = m; ...}
7   void start(Continuation e) { _exit = e; }
8   void exit() { call(_exit); }
9   void addChoice(Continuation c) {
10    _ctrl.enqueue(c);
11    _state.enqueue(new Checkpoint(_solver));
12  }
13  void fail() {
14    if (_ctrl.empty()) exit();
15    _state.dequeue().restore();
16    call(_ctrl.dequeue());
17  }
18 }

```

Figure 5. The BDS controller.

at shallow depths are enqueued before nodes inserted deeper in the tree, which guarantees the bias toward discrepancies that occur early on. Notice as well how the number of available credits is not needed. Indeed, given a search node $P = \langle X, D, C \rangle$ with a branching $[c_0, \dots, c_{k-1}]$, the method `addChoice` is first called for $\langle X, D, C \cup \{c_1\} \rangle$ right before proceeding with the expansion of $\langle X, D, C \cup \{c_0\} \rangle$. The next sibling will not be explicitly added until the search returns to the CSP $\langle X, D, C \cup \{c_1\} \rangle$ at which point $\langle X, D, C \cup \{c_2\} \rangle$ would be added via `addChoice`. Consequently, a queue with its first-in/first-out policy corresponds exactly to the BDS needs.

Several other nonchronological strategies exist and are briefly described below.

Iterative deepening [42] is where the object is to decompose the entire search in several waves with each wave limited to a maximal depth with the hope that good solutions can be found at shallow depths.

Interleaved depth-first search [43] attempts to compensate for the greediness of depth-first search by interleaving classic depth-first searches on several active subtrees switching from subtrees to subtrees when failures are detected. In essence, Interleaved depth-first search (IDFS) *simulates* a parallel exploration via DFS on a single-core machine.

Limited discrepancy search [41] attempts to compensate for a less than ideal variable selection heuristic. Indeed, if the variable selection heuristic is perfect, following it obviously with depth-first search leads straight to a solution node. However, if the variable selection heuristic is imperfect, a wrong decision taken early on (at a shallow depth) will take a considerable number of branching decision before it gets reconsidered. LDS addresses this difficulty by decomposing the search into several waves where each wave is allowed a bounded number of discrepancies (disagreement with the heuristic). With no discrepancies alone, only the left-most path of the search tree can be followed. With one discrepancy, only one right branch can be followed, *but it can start at any depth*. Consequently, LDS will reconsider early decision much earlier than its DFS cousin.

Best-first search [44] It was designed for optimization problems. Its emphasis is to maintain, for each node in the fringe of the tree, an estimation of the quality of the solution than can be found within the subtree rooted at that node. The node selection heuristic is then bound to give precedence to nodes whose estimate are best and therefore likely to deliver a high quality solution earlier.

Given that all the strategies discussed above define policies to decide which node of the fringe to expand next, they can all be realized through search controllers that implement the given node management policies.

META STRATEGIES

Meta strategies offer the opportunity to alter the basic behavior of a strategy to increase its effectiveness. The changes can take many forms ranging from pruning symmetric subtrees, making the search incomplete, to exploiting randomization or superimposing a local-search component.

Symmetry Breaking

CP usually regards symmetries as a source of difficulties. The existence of symmetries typically translates in the exploration of many subtrees that are isomorphic to each other. Worse, if such a subtree does not contain any solution, the exploration of all its isomorphic cousins is equally doomed and a waste of resources. Considerable efforts have been expended toward the detection and the elimination of symmetries through constraints or the search; see Refs [45–49]. While statically breaking symmetries through constraints is appealing, it often introduces pernicious interactions with search heuristics and prevents them from being effective. For instance, a symmetry breaking constraint might very well eliminate the values that the heuristic would normally recommend. The ability to write the search is then a significant advantage to avoid the problem altogether by breaking the symmetries dynamically during the search after giving the heuristics the opportunity to produce unfettered recommendations.

Value Symmetries. Consider the steel mill slab again. When the search proceeds, it assigns orders to specific slabs. At any point in time, two *classes* of slabs can be clearly recognized:

- slabs that are already in use, with at least one order placed on them;

- slabs that are still completely empty (no orders assigned to them yet).

All the unused slabs in the second class are identical and there is no point in trying all of them. The search should only look for a solution using one member of this class and there is no need to consider the others. This idea gives rise to the following search for the steel mill [50].

```
1 forall(o in Orders) by (x[o].getSize()) {
2   int ms = max(0,maxBound(x));
3   tryall<m>(s in Slabs: s <= ms + 1)
4     m.label(x[o],s);
5 }
```

The fragment still uses the *min-dom* dynamic variable selection heuristic. Its value selection, however, is changed to explicitly limit the enumeration of values (slabs) that are already in use plus one unused value. The instruction on-line 2 sets *ms* (maximum used slab) to the largest slab identifier appearing in variables already bound. The **tryall** instruction on-line 3 simply limits the selection to *ms* plus one “fresh” unused slab identifier.

Limits

To control the runtime of a model, a first alteration is the use of *limits*. In Ref. 51, Perron shows how to build practical solutions for real problem with limits. A limit can be the number of failures seen so far, the total time spent from the beginning, or even the number of solutions constructed. Enforcing limits makes the search incomplete, unless the limit is combined with an iterative scheme that repeatedly increases the limit. For instance, the fragment

```
1 Solver<CP> m();
2 m.limitTime(60);
```

imposes a time limit of 60 s, while

```
1 Solver<CP> m();
2 m.limitFailures(1000);
```

imposes a limit on the number of failed nodes in the search tree. Note that the limit can

be imposed at any point, that is, before the search starts or at any point during the search. For instance, one could impose a limit

```

1 Boolean first(true);
2 solveall<m> {
3   ...// state the model
4 } using {
5   forall(i in 0..n-1 : !x[i].bound()) by (x[i].getSize())
6     tryall<m>(v in q[i].getMin()..q[i].getMax() : q[i].memberOf(v))
7       m.post(q[i] == v);
8   if (first) m.limitFailures(1000);
9   first := false;
10  cout << q << endl;
11 }
```

Restarts

Heuristics that define the topology of the search tree do not always impose a total order on the branching decisions. Sometimes several alternatives (both in the variable and value selection) are indistinguishable by the heuristics and could be produced in any order leading to an inherent randomization of the search. Consider the queens problem. At the root of the search tree, the domains of all the variables are identical ($1..n$) and the variable heuristic could pick any variable to begin with. Ties are therefore typically broken uniformly at random and multiple searches could witness different search trees and yield solutions early or late in the search process. *Restarts* [11] exploit this inherent randomization to improve the efficiency of the search. See Ref. 52 for a survey. Restarts are typically coupled with a limit and repeat the same search. The control of the restarting captured by a *restarting strategy* $S = (t_1, t_2, \dots)$, where each t_i is a positive number representing a limit for the i th round of restart. The limit t_i can take the form of a limit on the number of backtracks, failures, time, or any other concrete measure of the amount of work being performed. The core idea is that if no solution is found during round i when the limit t_i is reached, the search starts round $i + 1$ with limit t_{i+1} . Clearly, the choice of restart strategy S does affect the amount of time needed to produce a solution and the *optimal* strategy depends upon the problem being solved. Luby *et al.* studied in Ref. 53 restarting

on the number of failures after the first solution is found with

strategies for Las Vegas algorithms and showed that if the runtime distribution for an algorithm is known, there exist an optimal fixed limit t^* and associated strategy $S^* = (t^*, t^*, t^*, \dots)$. In practice though, this runtime distribution is not available and practitioners must devise suitable strategies. Luby offers a *universal* strategy $S = (1, 1, 2, 1, 1, 2, 4, 1, 1, 2, 1, 1, 2, 4, 8, 1, 1, 2, \dots)$ which is guaranteed to be within a logarithmic factor of the (unknown) optimal strategy S^* . One caveat is a core assumption made by Luby, namely, all the rounds are independent. This may not hold in practice when learning is used from rounds to rounds to try and improve the heuristic (e.g., when a global heuristic such as impacts is used) so the universal strategy is not a silver bullet. Indeed, Harvey [11] and Gomes *et al.* [54] both use fixed cut-off strategies. Kautz *et al.* [55] revisit this issue and compare several strategies, including a dynamic policy based on empirical observation of running times on limited runs. Walsh proposed in Ref. 56 a restart strategy where the sequence is a geometric progression $S = (1, r, r^2, \dots)$ with $1 \leq r \leq 2$, which works well in practice as it increases the limit faster than the Luby universal strategy.

Restart Theory. A natural question that arises when contemplating the adoption of a restarting strategy is whether restarting can or cannot be helpful. Gomes *et al.* [54,57] establish a connection between the effectiveness of restarting strategies and the nature

of the runtime distribution. Specifically, they show that restarts are beneficial when the runtime distribution is heavy-tailed, that is, when there is a significant probability that a run will be “long.” More formally, a runtime distribution X is heavy-tailed when the tail satisfies $\Pr[X > t] \approx C \cdot x^{-\alpha}$, $t > 0$ where $\alpha > 0$ is a constant, that is, the tail has a power-law decay rather than an exponential decay. Further, Van Moorsel *et al.* [58] offer necessary and sufficient conditions characterizing when restarts are beneficial. Their work still hinges on the independence of the restart rounds though.

Example 3 [The Magic Square]. The magic square problem simply creates a matrix of 10×10 cells with the following requirement: all the rows, columns, and the two main diagonals must each add up to $n \times \frac{n^2+1}{2}$ and all the cells must take different values from $1, \dots, n^2$. A CP model is shown in lines 9–13 in Fig. 6. Without restarts though, a finite domain solver is typically limited to small squares. With restarts, larger instances can be solved without difficulty. The restarting search procedure spans lines 15–22. The while loop (line 16) goes on until all the variables are bound. Each iteration picks (with a semigreed

selectMin selector) a free variable with the smallest or second-smallest domain thereby potentially increasing the number of ties. Clearly, several variables could have the same domain size and the parameter [2] further enlarges the set of candidate variables. The selector is randomized and picks a variable uniformly at random from the set. Line 18 computes the middle of the selected variable domain, and line 19 branches by splitting the domain in half. The restarting strategy is implemented by lines 7 and 22. Line 7 sets up a failure limit of 1000 and states that the search should be restarted when the limit is reached. Line 22 handles a restart by increasing the limit. Upon restarting, the search returns to the root of the tree and repeats the entire process.

Large Neighborhood Search

LNS [12] is another alteration that adds a local search component on top of a restarting strategy. It appeared as “variable-depth search” with Lin and Kernighan [59] for the traveling salesman and as ejection chains in Glover [60]. It has been used extensively in the context of Jobshop scheduling

```

1 Solver<CP> cp();
2 int n = 10;
3 range R = 1..n;
4 range D = 1..n^2;
5 int T = n * (n^2 + 1)/2;
6 var<CP>{int} s[R,R](cp,D);
7 cp.restartOnFailureLimit(1000);
8 solve<cp> {
9   forall(i in R) cp.post(sum(j in R) s[i,j] == T);
10  forall(i in R) cp.post(sum(j in R) s[j,i] == T);
11  cp.post(sum(i in R) s[i,i] == T);
12  cp.post(sum(i in R) s[i,n-i+1] == T);
13  cp.post(alldifferent(all(i in R,j in R) s[i,j]));
14 } using {
15   var<CP>{int} []vars = all(i in R,j in R) s[i,j];
16   while (!bound(vars)) {
17     selectMin[2](i in vars.getRange(): !vars[i].bound())(vars[i].getMin()) {
18       int mid = (vars[i].getMin()+vars[i].getMax()) / 2;
19       try<cp> cp.post(vars[i]<=mid); | cp.post(vars[i]>mid);
20     }
21   }
22 } onRestart cp.setRestartFailureLimit(cp.getRestartFailureLimit() + 100);

```

Figure 6. Restarting magic square.

[61–63], and Perron *et al.* [64] investigated how to define large neighborhoods based on the amount of propagations from constraints.

LNS differs from plain restarts by leaving a subset of the variables fixed so that the subsequent tree search solely focuses on the nonfixed variables. At each restart, a

different subset is fixed. Naturally, LNS is particularly useful in an optimization problem where a stream of solutions of increasing quality is desired. Indeed, one must have a solution handy when restarting to fix a subset of the variables to some value. The fragment

```

1  ...// the search
2 } onRestart {
3   Solution s = m.getSolution();
4   if (s!=null) {
5     forall(i in R)
6       if (lms.get() < 80) cp.post(x[i] == x[i].getSnapshot(s));
7   }
8 }
```

is a schematic LNS procedure. When the search restarts, it first retrieves the incumbent solution in line 3. If there is one, it

loops over all the variables and with an 80% chance, it constraints the *i*th variable to be equal to its value in the incumbent. Clearly,

```

1  int nbOrders = sz; range Orders = 1..nbOrders;
2  int nbSlabs = nbOrders; range Slabs = 1..nbSlabs;
3  set{int} colorOrders[c in Colors] = filter(o in Orders) (color[o] == c);
4  int maxCap = max(i in Caps) capacities[i];
5  int loss[c in 0..maxCap]=min(i in Caps: capacities[i] >= c) capacities[i] - c;
6  UniformDistribution dist(1..100);
7  Solver<CP> m();
8  var<CP>{int} x[Orders](m,Slabs);
9  var<CP>{int} l[Slabs](m,0..maxCap);
10
11 m.lnsOnFailure(4000);
12 minimize<m> sum(s in Slabs)loss[l[s]]
13 subject to {
14   m.post(multiknapsack(x,weight,l));
15   forall(s in Slabs)
16     m.post(sum(c in Colors) (or(o in colorOrders[c]) (x[o] == s)) <= 2);
17 } using {
18   forall(i in Orders)
19     tryall<m>(v in x[i].getMin()..x[i].getMax() : x[i].memberOf(v))
20       m.post(x[i] == v);
21   cout << x << endl;
22 } onRestart {
23   Solution s = CP.getSolution();
24   if (s != null) {
25     tryall<m>(i in 1..1000) {
26       forall(o in Orders)
27         if (dist.get() <= 90-10*i)
28           m.post(x[o] == x[o].getSnapshot(s));
29     }
30     m.setPrimalBound(s.getObjectiveValue());
31   }
32 }
```

Figure 7. The steel mill slab problem with large neighborhood search.

the subsequent search only focuses on the nonfixed variables.

Figure 7 illustrates the steel mill slab model of Refalo [65], which offered the first efficient CP model for this problem. It uses a LNS based on the above schema. The first change appears on-line 11 where the method call `m.lnsOnFailure(4000)`; imposes a limit of 4000 failures between two LNS restarts. The second change spans lines 23–32. It specifies which part of the solution to restore at the onset of an LNS round. It simply attempts (up to a thousand times) to fix some of the variable (randomly selected) to their value in the current solution. Line 30 also requires that the objective function improve. Naturally, the randomly selected subset of fixed variables may be inconsistent with the requirement to improve on the objective, in which case COMET will backtrack and consider another subset.

CONCLUSION

Search in CP is an integral part of the development of a solution to a problem. It provides ample opportunities to exploit the structure of the problem and steer the exploration process toward promising parts of the search space. It has traditionally been supported by languages offering a rich level of control. Search is best captured in terms of the specification of a search tree and its exploration strategy. While heuristics control the topology of the search tree and attempt to place promising nodes on the far left edge of the tree, strategies are mediators that compensate for the weaknesses of heuristics. Even though CP search is most often a complete procedure, meta-strategies deliver incomplete search techniques that are capable of tackling larger instances and taking advantage of randomization.

REFERENCES

1. Baptiste P, Le Pape C, Nuijten W. Constraint-based scheduling. Norwell (MA); Kluwer Academic Publishers; 2001. ISBN: 0-7923-7408-8.
2. Collavizza H, Rueher M, Hentenryck PV. CPBVP: a constraint-programming framework for bounded program verification. In: Stuckey PJ, editor. CP, Volume 5202, Lecture Notes in Computer Science. Berlin, Heidelberg; Springer; 2008. pp. 327–341.
3. Simonis H, Nguyen H, Dincbas M. Verification of digital circuits using CHIP. In: Proceedings of the IFIP WG 10.2 International Working Conference on the Fusion of Hardware Design and Verification. Glasgow, Scotland; 1988 July.
4. Michel L, Shvartsman AA, Sonderegger EL, *et al.* Optimal deployment of eventually-serializable data services. Extended version of the CPAIOR'08 paper. Ann Oper Res 2010. DOI: 10.1007/s10479-010-0684-13; www.springerlink.com/content/a0p64017742753m5/.
5. Bin E, Emek R, Shurek G, *et al.* Using a constraint satisfaction formulation and solution techniques for random test program generation. IBM Syst J 2002;41(3):386–402.
6. McAloon K, Tretkoff C, Wetzel G. Sport league scheduling. In: Proceedings of the 3th Ilog International Users Meeting. Paris, France; 1997.
7. Régim J-C. Sport league scheduling. In: INFORMS. Montreal, Canada; 1998.
8. Lustig I. Scheduling the NFL with constraint programming. Colloquium: Brown University; 2004, 2004.
9. Van Hentenryck P, Michel L. Constraint-based local search. Cambridge (MA): The MIT Press; 2005.
10. Michel L, See A, Van Hentenryck P. Parallelizing constraint programs transparently. In: Proceedings of the 13th International Conference on the Principles and Practice of Constraint Programming (CP-2007). Providence (RI); 2007.
11. Harvey W. Nonsystematic backtracking search [PhD thesis]. Stanford University; 1995.
12. Orlin J, Ahuja R, Ergun O, *et al.* A survey of very large scale neighborhood search techniques. Disc Appl Math 2002;123:75–102.
13. Van Hentenryck P. Consistency techniques in logic programming [PhD thesis]. University of Namur (Belgium); 1987.
14. Deville Y, Van Hentenryck P. An efficient Arc consistency algorithm for a class of CSP problems. In: International Joint Conference on Artificial Intelligence. Sydney, Australia; 1991 Aug.

15. Van Hentenryck P, Deville Y, Teng C. A generic Arc consistency algorithm and its specializations. *Artif Intell* 1992;57(2–3): 291–321.
16. Stallman R, Sussman G. Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. *Artif Intell* 1977;9:135–196.
17. Dechter R. Enhancement schemes for constraint processing: backjumping, learning, and cutset decomposition. *Artif Intell* 1990;41(3):273–312.
18. Ginsberg ML, McAllester DA. Dynamic backtracking. *J Artif Intell Res* 1993;1:25–46.
19. Prosser P. Hybrid algorithms for the constraint satisfaction problem. *Comput Intell* 1993;9:268–299.
20. Schiex T, Verfaillie G. Nogood recording for static and dynamic constraint satisfaction problems. *Int J Artif Intell Tools* 1994;3:48–55.
21. Bruynooghe M. Solving combinatorial search problems by intelligent backtracking. *Inf Process Lett* 1981;12(1):36–39.
22. Rossi F, van Beek P, Walsh T, editors. *Handbook of constraint programming. Backtracking search algorithms*. Amsterdam; Elsevier; 2006. Chapter 4. pp. 85–134.
23. Van Hentenryck P, Michel L. Nondeterministic control for hybrid search. *Proceedings of the 2nd International Conference on the Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimisation Problems (CP-AI-OR'04)*. Prague: Czech Republic; 2005.
24. Michel L, See A, Van Hentenryck P. High-level nondeterministic abstractions in c++. In: *12th International Conference on Principles and Practice of Constraint Programming (CP'06)*, Lecture Notes in Computer Science. Nantes, France; 2006 Sep.
25. Problem 38. Available at www.csplib.org.
26. Frisch AM, Miguel I, Walsh T. Modelling a steel mill slab design problem. In: *Proceedings of the IJCAI-01 Workshop on Modelling and Solving Problems with Constraints*; 2001 Aug 4–10; Seattle, Washington, 2001. pp. 39–45.
27. Haralick R, Elliot G. Increasing tree search efficiency for constraint satisfaction problems. *Artif Intell* 1980;14:263–313.
28. Bessière C, Régin J-C. Mac and combined heuristics: Two reasons to forsake fc (and cbj?) on hard problems. In: *Freduec EC. Proceedings of the Second International Conference on Principles and Practice of Constraint Programming*; 1996 Aug 19–22; Cambridge, MA; Lecture Notes in Computer Science. Berlin: Springer; 1996. pp. 61–75.
29. Smith BM, Grant SA. Trying harder to fail first. *Proceedings 13th European Conference on Artificial Intelligence (ECAI 98)*. John Wiley & Sons, Ltd.; 1997. pp. 249–253.
30. Boussemart F, Hemery F, Lecoutre C, *et al.* Boosting systematic search by weighting constraints. In: *de Mántaras RL, Saitta L, editors. ECAI. IOS Press*; 2004. pp. 146–150.
31. Refalo P. Impact-based search strategies for constraint programming. In: *Wallace [63]*. pp. 557–571.
32. Gauthier J, Ribiere G. Experiments in mixed-integer linear programming using pseudo-costs. *Math Program* 1977;12:26–47.
33. Bixby R, Lee E, Bixby RE, *et al.* Parallel mixed integer programming. Technical report, Rice University Center for Research on Parallel Computation; 1995.
34. Correia M, Barahona P. On the integration of singleton consistencies and look-ahead heuristics. 2008; pp. 62–75.
35. Zanarini A, Pesant G. Solution counting algorithms for constraint-centered search heuristics. *Constraints* 2009;14(3):392–413.
36. Hsu EI, Kitching M, Bacchus F, *et al.* Using expectation maximization to find likely assignments for solving csp's. *AAAI'07: Proceedings of the 22nd National Conference on Artificial intelligence*. Menlo Park (CA): AAAI Press; 2007. pp. 224–230.
37. Michel L, See A, Hentenryck P. Parallelizing constraint programs transparently. In: *Bessiere C. Proceedings of the 13th International Conference on Principles and Practice of Constraint Programming*; 2007 Sep 23–27; Providence, RI; Lecture Notes in Computer Science. Berlin: Springer; 2007. ISBN: 978-3-540-74969-1.
38. Hieb R, Dybvig RK, Bruggeman C. Representing control in the presence of first-class continuations. *Proceedings of the SIGPLAN '90 Conference on Programming Language Design and Implementation*; 2007 Jun 20–22; White Plains, NY; ACM Press; 1990. pp. 66–77.
39. Kelsey R, Clinger W, Rees J, *et al.*, editors. Revised 5 report on the algorithmic language scheme. *ACM SIGPLAN Notices* 1998;33: 26–76.
40. Walsh T. Depth-bounded discrepancy search. In: *Proceedings of the 15th International Joint*

- Conference on Artificial Intelligence. Nagoya, Japan; 1997 Aug.
41. Harvey W, Ginsberg M. Limited discrepancy search. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada; 1995 Aug.
 42. Reinefeld A, Marsland TA. Enhanced iterative-deepening search. *IEEE Trans Pattern Anal Mach Intell* 1994;16(7):701–710.
 43. Meseguer P. Interleaved depth-first search. In: International Joint Conference on Artificial Intelligence. Nagoya, Japan. San Francisco (CA): Morgan Kaufmann Publishers Inc.; 1997. pp. 1382–1387.
 44. Pearl J. Heuristics: intelligent search strategies for computer problem solving. Boston (MA): Addison-Wesley; 1984.
 45. Gent IP, Smith BM. Symmetry breaking in constraint programming. *Proceedings of ECAI-2000*. IOS Press; 2000. pp. 599–603.
 46. Focacci F, Milano M. Global cut framework for removing symmetries. *CP'01: Proceedings of the 7th International Conference on Principles and Practice of Constraint Programming*. London: Springer; 2001. pp. 77–92.
 47. Puget J-F. Symmetry breaking revisited. *Constraints* 2005;10(1):23–46.
 48. Sellman M, Van Hentenryck P. Structural symmetry breaking. In: *IJCAI'05*; 2005 July 30–Aug 5; Edinburgh, Scotland: Professional Book Center; 2005. pp. 298–304. ISBN: 0938075934.
 49. Flener P, Pearson J, Sellmann M, *et al.* Static and dynamic structural symmetry breaking. *Proceedings of CP'06*. Springer; 2006. pp. 695–699.
 50. Hentenryck PV, Michel L. The steel mill slab design problem revisited. In: Perron L, Trick MA, editors. Volume 5015, *CPAIOR, Lecture Notes in Computer Science*. Springer; 2008. pp. 377–381.
 51. Perron L. Search procedures and parallelism in constraint programming. 5th International Conference on the Principles and Practice of Constraint Programming (CP'99). Alexandria (VA): Springer; 1999. pp. 346–360.
 52. Gomes C. Randomized Backtrack Search Constraint and integer programming: towards a unified methodology. Kluwer; 2004. pp. 233–292.
 53. Luby M, Sinclair A, Zuckerman D. Optimal speedup of Las Vegas algorithms. *Inf Process Lett* 1993;47: 173–180.
 54. Gomes CP, Selman B, Crato N, *et al.* Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *J Autom Reason* 2000;24:2000.
 55. Kautz H, Horvitz E, Ruan Y, *et al.* Dynamic restart policies. 2002. pp. 674–681.
 56. Walsh T. Search in a small world. *IJCAI '99: Proceedings of the 16th International Joint Conference on Artificial Intelligence*. San Francisco (CA): Morgan Kaufmann Publishers Inc.; 1999. pp. 1172–1177.
 57. Gomes CP, Selman B, Kautz H. Boosting combinatorial search through randomization. *AAAI '98/IAAI '98: Proceedings of the 15th National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*. Menlo Park (CA): American Association for Artificial Intelligence; 1998. pp. 431–437.
 58. Moorsel APAV, Wolter K. Analysis and algorithms for restart. *Proceedings 1st International Conference on the Quantitative Evaluation of Systems (QEST)*; 2004 Sep 27–30; Enschede, The Netherlands; IEEE Computer Society; 2004. pp. 195–204.
 59. Kernighan B, Lin S. An efficient heuristic procedure for partitioning graphs. *Bell Syst Tech J* 1970;49:291–307.
 60. Gamboa D, Rego C, Glover F. Data structures and ejection chains for solving large-scale traveling salesman problems. *Eur J Oper Res* 2005;160(1):154–171.
 61. Danna E, Perron L. Structured vs. unstructured large neighborhood search: a case study on job-shop scheduling problems with earliness and tardiness costs. In: Rossi F, editor. Volume 2833, *CP Lecture Notes in Computer Science*. Springer; 2003. pp. 817–821.
 62. Applegate D, Cook W. A computational study of the job shop scheduling problem. *ORSA J Comput* 1991;3(2):149–156.
 63. Adams J, Balas E, Zawack D. The shifting bottleneck procedure for job shop scheduling. *Manag Sci* 1988;34(3):391–401.
 64. Perron L, Shaw P, Furnon V. Propagation guided large neighborhood search. In: Wallace [63]. pp. 468–481.
 65. Gargani A, Refalo P. An efficient model and strategy for the steel mill slab design problem. In: Bessiere C. *Proceedings of the 13th International Conference on Principles and Practice of Constraint Programming (CP'07)*; 2007 Sep 23–27; Providence, RI; *Lecture Notes in Computer Science*. Berlin: Springer; 2007. ISBN: 978-3-540-74969-1.

BASIC INTERDICTION MODELS

J. COLE SMITH

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

INTERDICTION BASICS

Network models are prevalent in operations research practice and theory because of their wide applicability to real-world problems. Nodes often represent junctions, connections, or facilities in a network, while arcs that connect these nodes may represent roadways, pipelines, cables, or virtual communication channels. One can model a broad array of scenarios using these concepts. A network may represent a physical entity such as a traffic grid, logistics transportation network, or telecommunication network, or logical entities such as project management schedules or production and inventory decisions.

In order to capture salient features of these networks, several properties are typically assigned to arcs and nodes in a network. The cost of an arc refers to the per-unit charge assessed for flows traversing the arc. In simultaneous flow networks (which are most common in network analysis), there are often flow capacities that restrict the amount of flow traversing the arc. Most flow networks also have supplies and demands associated with each node. Additionally, in many network design models, there are fixed costs for establishing nodes at certain places in the network, and for establishing arcs through which flows may be transmitted.

Given these foundations, it is indeed possible to model a wide array of practical optimization problems. However, these models often assume that data is known with certainty, and that all components of the network operate as intended. Furthermore, optimization models that design networks necessarily do so with limited

redundancy, since costs associated with capital development (e.g., roadways, transmission cables, and facilities) are typically minimized. Therefore, such networks are particularly vulnerable to disruptions such as arc failures.

The goal of network interdiction studies is to assess the vulnerability of networks to various types of failures (see Ref. 1 for an excellent introduction to this field). Most models of this type are Stackelberg games: a *leader* acts first and performs some interdiction actions on the network, and a *follower* responds by performing recourse operations on the network. Since the order and number of actions can vary depending on the model, we refer to these agents as the *interdictor* and *operator*, respectively, in this article, as is often done in the literature. Interdiction models typically have a single objective function on the part of the operator, and a limited disruption budget for the interdictor; this budget may, for example, refer to the number of arcs that can be eliminated in an interdiction act. Information is perfectly known to both agents, and the interdictor's objective is to hinder the operator's objective to the greatest extent possible.

For instance, consider a telecommunication network on which long-distance charges are incurred by routing calls between nodes on the network. Each arc represents a connection between points, and the per-unit flow costs on these arcs represent the cost per call, while their capacities represent the maximum number of calls that can be made across these lines. Point-to-point demands exist on this network, stating the number of calls that must be routed to and from each pair of nodes. The operator seeks to complete all demand requests on the network at a minimum total cost. (This is an example of a *multi-commodity flow network*, since demands are specific to node pairs.) Now, suppose that an interdicting agent has the capability of destroying any k arcs in the network. The interdictor seeks a set of k arcs which, when removed from the system, results in the

largest optimal objective function value for the operator. Note that the interdicator is *maximizing* the minimum cost for the operator, and hence this problem belongs to the class of *maximin* problems (minimax problems can be dealt with similarly).

The most transparent applications of network interdiction invoke military or antiterrorism scenarios. Indeed, as we will discuss later in this article, network interdiction models are very well suited to address problems arising in this context. The interdicator's budget can be tailored to the model under consideration, and can reflect the relative difficulty of interdicting certain network assets. Hence, if a terrorist can reasonably launch a simultaneous attack on four soft targets or two fortified targets, then one can state a constraint that (the number of soft target attacks) + 2 (the number of fortified target attacks) ≤ 4 . It is also a reasonable assumption that terrorists or military enemies will seek to optimally disrupt an operator's objective function value. The assumption of perfect information is valid for two reasons. One, underestimating an enemy's ability to gain information (or to make a lucky guess) could leave the network vulnerable to a devastating attack. Two, in a surprising number of network applications, detailed information is very much available to the public [2].

However, network interdiction is most generally applied to situations in which vulnerability to any kind of disruption is possible. Natural disasters do not intentionally seek to exploit weaknesses in infrastructure, but may very well do so by accident. A network operator may wish to guarantee a certain level of system performance against any of a possible set of scenarios. In the telecommunication network example above, even if terrorism is of no concern, accidents or random equipment failure may drive the operator to consider the strategic addition of redundant infrastructure. To guarantee that the network remain functional even if any three links fail, one may consider a "conceptual" interdicator with a budget of three arcs that can be destroyed. When more general failure probabilities are associated with a network, one can guarantee that the

network remains operational with probability $0 < \tau < 1$ using network interdiction concepts. Let C be the set of network components (including perhaps nodes and arcs), and define x_i to be the interdicator's binary variable that equals 1 if component i of the network is disrupted, and 0 otherwise, for all $i \in C$. Let the probability of component i operating successfully be given by $p_i, \forall i \in C$, and assume that all component failures are independent. The interdicator in this case is constrained to create a failure event whose probability does not exceed $(1 - \tau)$. This means that the interdicator's choice of x must be such that

$$\prod_{i \in C} (1 - p_i)^{x_i} \leq (1 - \tau), \quad (1a)$$

or

$$\sum_{i \in C} \log(1 - p_i) x_i \geq \log(1 - \tau). \quad (1b)$$

The interdiction budget thus takes on the form given in Equation (1b).

One final note on the benefits of interdiction models is relevant at this point. A time-honored tradition in assessing the vulnerability of systems relies on enumerating disaster scenarios and counting all possible things that can go wrong in a network. Considering a network having n components, of which at most k can fail, one would have to enumerate and evaluate $\sum_{\ell=1}^k \binom{n}{\ell}$ possible failure events, which is computationally prohibitive for modest values of n and k . Representing such problems as network interdiction problems results in models that are solvable by much more efficient algorithms than exhaustive enumeration.

In the section titled "Interdiction Problem Classes," we discuss various classes of interdiction models and applications. Then, in the section titled "Fortification against Interdiction," we extend these models to allow the network operator to first design and/or fortify an existing network, keeping in mind the subsequent actions of an interdicator. This article closes with the section titled "Summary and Future Challenges" with a brief discussion of prominent challenges in the field.

INTERDICTION PROBLEM CLASSES

We begin this section by examining a general model for studying interdiction problems. Observe here that interdiction does not need to be specific to network models, but could be applied in any Stackelberg optimization problem. Let the interdictor's decisions be given by vector x , and let the operator's recourse decisions be given by y . Then the interdictor's problem can be given by

$$\sup f(x) \quad (2a)$$

$$\text{s.t. } x \in X, \quad (2b)$$

where X is a set of feasible interdiction solutions, and where $f(x)$ is given by

$$f(x) = \inf g(x, y) \quad (3a)$$

$$\text{s.t. } y \in Y(x), \quad (3b)$$

where $g(x, y)$ is an objective function based on inputs x and decision variables y , and where $Y(x)$ is a feasible region based on inputs x . (Observe that since $f(x)$, $g(x, y)$, X , and $Y(x)$ are defined generally, we thus use “sup” rather than “max” to handle cases in which a maximum does not exist in Equation (2a); similarly, we use “inf” rather than “min” in Equation (2b).)

Just as in any other class of optimization problems, interdiction research examines special structures and assumptions on Models (2) and (3) above. A listing of some such problem specializations is given below.

Maximum Flow Interdiction

The maximum flow interdiction problem takes place on a network with a designated source node s and sink node t . Each arc has a capacity, but no flow cost. The objective is to find a set of flows from s to t that obey flow balance and capacity constraints, and move as much total flow from s to t as possible. Since the operator's goal is to maximize flow, the interdictor has the goal of minimizing the operator's maximum flow (casting this as a minimax, rather than a maximin, problem).

Perhaps not surprisingly, the theory of linear programming duality addresses one

important aspect of interdiction via the *max-flow/min-cut* theorem. Define a cut as a set of arcs that disconnect s from t when removed from the network (i.e., no directed path from s to t would remain in the network), and define the weight of this cut as the sum of all arc capacities in the cut. Then the maximum flow in the network equals the minimum weight cut; if budget is given in terms of total capacity that can be removed from the network, one can easily optimize the maximum flow interdiction problem by identifying this cut (which is polynomially achievable by maximum flow algorithms) and removing capacity from this cut.

Alternatively, if the interdictor is restricted on the *number* of arcs that may be removed from the network, the problem becomes much more difficult. One solution approach considers a dual formulation of the maximum flow problem, which when coupled with interdiction decisions, yields a nonlinear mixed-integer programming problem. The nonlinear terms are quadratic, and since at least one variable in each quadratic term is constrained to be binary valued, the problem can be linearized and solved as a linear mixed-integer program. The work of Wood [3] presents these ideas in technical detail, and lays the foundation for broader network interdiction analysis.

Shortest Path Interdiction

In the shortest path interdiction problem, network arcs are associated with flow costs, but not capacities. The operator seeks a shortest path from a source node s to a sink node t , that is, one whose sum of arc flow costs is minimized over all possible paths. The interdictor, in a maximin framework, wishes to take some action to maximize the operator's shortest path cost. Interdiction in this situation usually refers to eliminating arcs or increasing costs on certain arcs. These problems can be algorithmically addressed in much the same way as the maximum flow problems.

For critical path (CP) networks, one actually seeks the longest path in a network as described above. CP networks model the completion of a complex project by representing individual tasks as arcs, and milestones as

nodes. Before any task represented by an arc (i,j) can commence, all tasks entering node i must be completed. These restrictions model the precedence relationships among tasks. For instance, in publishing a book, one must wait for (i) the arrival of copyright permissions and (ii) the corrected manuscript, before a milestone of having the right to publish is achieved. After this milestone is reached, one can then begin the tasks of (iii) printing the book and (iv) releasing certain specific details of the book to potential consumers. Arc weights refer to corresponding task durations, and the entire project is completed only when all tasks are completed and the last milestone is reached. The last milestone cannot be reached faster than the length of any path from the source to the destination, since each path represents a chain of events that must be done sequentially; hence, the project completion time is given by the longest path in the network. While the longest path problem is typically quite difficult to solve, it is polynomially solvable in this context since CP networks are necessarily acyclic. The interdiction problem here can seek to select a set of possible task duration delays that maximize the longest path in the network.

A closely related problem may analyze a network in which arc weights are bounded between 0 and 1, and refer to the probability of successfully traversing the arc. The operator may thus seek a most-reliable path in the network, assuming independent arc reliabilities. Using a similar operation as shown in Equations (1a) and (1b), we can convert this problem to a shortest path problem (where arc costs are in terms of logarithms of arc probabilities). Interdiction in this context seeks to minimize the maximum probability of traversing the network from the source to the destination. While the interdicator is typically seen as an antagonist seeking to sow destruction in the network, one can easily imagine the operator to be a fugitive, with the interdicator acting as a policing agency reducing the operator's chances of remaining undetected.

Indeed, an excellent example of this application comes in the deployment of nuclear sensors in the former Soviet republics [4].

These nuclear sensors are designed to detect smugglers of nuclear material. With the establishment of the Second Line of Defense (SLD) in 1998, an optimization challenge was presented to deploy sensors on components of the network. The presence of a sensor would decrease the reliability of successfully traversing a link on which the sensor was located (or, if the sensor is placed on a node, it would decrease the probability of successfully traversing any arc coming into the node). We discuss extensions of this research in the context of fortification actions in the section titled "Fortification against Interdiction" (3).

Minimum Cost and Multicommodity Flow Interdiction

The minimum cost flow interdiction problem generalizes the previous models. In the minimum cost flow interdiction problem, arc costs and capacities are given, along with supplies and demands at every node. The operator will exhaust all excess supplies and satisfy all demands at every node, while obeying the capacity constraints on the arcs and minimizing overall flow costs. The interdicator will take an interdiction action to maximize the operator's minimum overall flow cost.

For instance, one could model a military materiel mobilization system as a network in which supply nodes contain reserve materiel, and demand nodes require receipt of this materiel. As long as materiel is considered to be one homogeneous item, issuing supplies across the network can be modeled as a minimum cost flow problem. An enemy may wish to cripple this network by deleting a strategic set of arcs that optimally hinders the deployment of materiel.

On the other hand, if materiel is not homogeneous, then one must consider point-to-point demands of different classes of items (e.g., tanks, artillery, fuel, and personnel). Since different nodes will have varying supplies/demands of these items, one must consider a multicommodity flow version of this problem. (See also the point-to-point telecommunication model given above for another multicommodity flow example.) For this case,

Lim and Smith [5] provide integer programming and duality-based procedures for optimizing multicommodity flow interdiction.

Asymmetric Goals and Information

It is possible that information about the underlying network may be viewed differently by the interdictor and operator, or that the interdictor and operator have different objectives. In this case, the interdictor/operator framework is altered somewhat. Maintaining the follower problem (Eq. 3), we would revise the interdictor's problem as

$$\sup f'(x, y) \quad (4a)$$

$$\text{s.t. } x \in X \quad (4b)$$

$$y \text{ optimizes the operator's} \\ \text{problem (Eq. 3).} \quad (4c)$$

Note that Equation (4c) requires a set of constraints to guarantee that the operator acts optimally. The field of mathematical programming with equilibrium constraints (MPEC) deals with these conditions, often requiring that the y -solution be a part of a Karush–Kuhn–Tucker (KKT) point to Problem (3), if KKT conditions are necessary and sufficient to optimize Problem (3). (See Ref. 6 for a review of MPEC literature.) Also, note that given a value of x , there may be multiple solutions $y \in \hat{Y}(x)$ that optimize Problem (3). The formulation (Eq. 4) optimizes, for the interdictor, over *all* $y \in \hat{Y}(x)$, which implies that the operator is willing to break ties in favor of the interdictor. In this sense, Equation (4) is optimistic for the interdictor [see the work of Bayrak and Bailey [7] or Smith *et al.* [8] for instances in which this assumption has been made].

The problem setup above is often used when the interdictor simply does not have all relevant information, or when the operator has been able to successfully conceal certain aspects of the network components. However, when the interdiction objective differs from the operator's objective, it is possible to represent many different types of nonadversarial behavior. A famous example of this arises in toll-pricing and revenue management [9]. The interdictor may be licensed to

install toll booths and control toll prices at certain points in a network. In response, an assumption is made that the operators (independently operating entities in the system) react to minimize a combination of travel time and tolls. A revenue-seeking interdictor may attempt to maximize revenue from the traffic network. A benevolent interdictor may attempt to improve traffic. It is well known that independent agents across a network will seek their fastest route home and converge to a traffic equilibrium that is sub-optimal, in the sense that the average time required to complete travel is larger than the minimum possible average travel time if the agents were centrally controlled. By strategically interdicting the network and adding prices to roadways (arcs), an interdictor can induce better traffic flows.

FORTIFICATION AGAINST INTERDICTION

Once a proper understanding of interdiction models and techniques is obtained, it is possible to consider fortification decisions on behalf of the operator that anticipate potential interdiction decisions and mitigate their effectiveness. These problems could protect against terrorist actions, improve the reliability of network components, or insure against the loss of certain components. In this section, we provide a brief discussion of fortification problems occurring in the network interdiction literature.

We begin here by discussing a design/fortification problem on multicommodity flow networks, which can be used to analyze the general problem of fortifying networks such as maximum flow, shortest path, critical path, and minimum cost flow networks. These problems now take the form of a three-stage optimization problem. The operator acts first to construct a network, or to fortify existing network infrastructure. In some variations of the problem, the operator may also be required to transmit an initial set of flows through the network. This would be the case when, for instance, an interdiction action is uncertain, and the effectiveness of the network before interdiction should be considered in the decision-making stage.

Following the network design phase, an interdicator acts to eliminate some set of arcs (within an interdiction budget), and the operator sends recourse flows through the remaining network. These models are either “min-max-min” or “max-min-max” models in practice. [Smith *et al.* [8] describe a “max-min-max” model of this form.]

Indeed, as shown in the section titled “Asymmetric Goals and Information,” one need not assume that the interdicator has the same goals as the operator. The interdicator could be acting greedily to destroy components of the network that have the most capacity, or those that transmit the most flows possible. This suboptimal enemy behavior might be due to the desire of the interdicator to make a dramatic attack on the network, or could reflect the interdicator’s lack of ability (due to partial information or time restrictions) to optimize an attack.

These fortification problems are often very difficult due to the three-stage representation. For simplified models, it may be possible to incorporate a compact linear model that represents two simultaneous actions. For instance, consider an emergency relief center location and protection model. There exist a set of client locations in some domain that must be covered by an emergency facility location (e.g., a fire station, hospital, or police station). Once an emergency takes place in any one of the client locations, service is provided from whichever facility is closest to the client. The location model would therefore minimize the expected amount of time to respond to an emergency, weighted perhaps by the importance of serving each of the client nodes and the likelihood of an emergency occurring at the nodes. An interdicator may act first to remove some combination of facilities, thus maximizing the minimum expected time for responding to these emergencies.

In this scenario, the fortification problem would therefore select a set of established facilities that could not be interdicted to minimize expected response time, assuming that optimal interdiction actions follow on unfortified facilities. Instead of modeling this as a three-stage (min-max-min) problem, Church and Scaparra [10] formulate a single mixed-integer linear program that

simultaneously captures the final two stages of the game (interdiction and facility-to-client assignment). This formulation is possible due to the special structure of the problem under investigation; namely, that client nodes are served by a direct assignment to facilities.

One recent development in the field of fortification regards the *prioritization* of fortification assets [11]. An implicit assumption has been made to this point that the operator’s budget for network design/fortification is known. (The same assumption is made for the interdicator, but this is justified owing to the conservative nature of critical network protection problems.) Some real-world implementations of fortification decisions do not implement all fortification decisions at once, but rather do so as budgets permit. Conceptually, considering assets labeled $\{a_1, \dots, a_7\}$, it is possible that the optimal fortification decision with a budget of protecting three assets would protect a_1 , a_2 , and a_3 . However, with a budget of four assets that can be protected, the complement of this subset, $\{a_4, \dots, a_7\}$, may become optimal to protect. It is generally suboptimal to give a priority list to an agency that orders the importance of fortifying assets. However, priority lists are often unavoidable due to practical considerations regarding budget uncertainty. Therefore, an optimal policy given a probability distribution of how much budget will be available for fortifying assets is of considerable practical value.

As a final remark to this section, we note that alternative models exist for certain design/fortification problems having special structures. When link failures are of sufficiently small probability (and nodes are reliable), one may consider only the event in which any single link can fail. While this model fits into the prior discussion of design/fortification problems, one can use problem-specific information to model and solve the problem in a more efficient fashion. One such problem minimizes the cost of adding enough links to a network such that a path exists between every pair of nodes (or, perhaps, a feasible general network flow exists across the entire network) even in the face of any single-arc failure [12]. Another considers the design of synchronous optical

networks (SONETs), which are inherently resistant to single-arc failures [13].

SUMMARY AND FUTURE CHALLENGES

Network interdiction studies are typically multistage games in which an interdicting agent alters certain network components in order to optimally hinder a network operator. These interdiction actions often consist of eliminating network arcs, and are limited in scope by cardinality or budget restrictions. These problems are often solved from the interdictor's perspective by maximizing damage to the network, subject to constraints forcing the interdictor to prepare against optimal operator decisions. These constraints are either logically generated on the basis of the special structure of the problem under consideration, or by considering the dual formulation of the operator's problem when the strong duality theorem applies to the operator's problem. Fortification decisions can then be considered prior to these interdiction models, which help assess where strategic resources should be deployed to protect against an intelligent interdicting agent. It is important to note that the interdicting agent need not be a malicious entity, but may instead be a worst-case accident due to random failures or natural disasters.

There are many avenues for continued development of interdiction models, but perhaps the most pressing regards multistage games, in which fortification and defense actions are alternated over a finite horizon. If these games are not trivial extensions of two- or three-stage problems (i.e., if they do not simply rebuild and destroy the same components over time), they cannot be solved using technology that currently exists. Multistage optimization problems are notoriously difficult, and even three-stage models can be quite difficult to address computationally. Similarly, we note that interdiction does not necessarily occur at discrete points. Especially for a multistage fortification/interdiction application, the timing of interdiction and fortification actions is an important decision aspect. Noting the suboptimality of a priority list, the interdictor

may choose to hold back interdiction actions at the present time, allowing the operator to have a longer time frame in which the network can be used without damage. In return, the interdictor could collect more information about their available interdiction budget, and make a more effective interdiction action later in the scenario. Likewise, if fortification budgets are also uncertain at the beginning of the time horizon, the operator may strategically deploy fortification decisions gradually at certain time points, rather than all at once.

REFERENCES

1. Brown GG, Carlyle WM, Salmeron J, Wood K. Defending critical infrastructure. *Interfaces* 2006;36(6):530–544.
2. Brown GG, Carlyle WM, Salmeron J, Wood K. Analyzing the vulnerability of critical infrastructure to attack and planning defenses. In: Smith JC, editor. *Tutorials in operations research: emerging theory, methods, and applications*. Hanover (MD): INFORMS; 2005. pp. 102–123.
3. Wood RK. Deterministic network interdiction. *Math Comput Model* 1993;17(2):1–18.
4. Morton DP, Pan F, Saeger KJ. Models for nuclear smuggling interdiction. *IIE Trans* 2007;39(1):3–14.
5. Lim C, Smith JC. Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Trans* 2007;39(1):15–26.
6. Luo Z-Q, Pang J-S, Ralph D. *Mathematical programs with equilibrium constraints*. Cambridge (UK): Cambridge University Press; 1996.
7. Bayrak H, Bailey MD. Shortest path network interdiction with asymmetric information. *Networks* 2008;52(3):133–140.
8. Smith JC, Lim C, Sudargho F. Survivable network design under optimal and heuristic interdiction scenarios. *J Glob Optimization* 2007;38(2):181–199.
9. Lawphongpanich S, Hearn DW, Smith MJ, editors. *Mathematical and computational models for congestion charging*. Volume 101, *Applied optimization*. New York: Springer; 2006.
10. Church RL, Scaparra MP. Protecting critical assets: the r -interdiction median problem with fortification. *Geogr Anal* 2006;39(2):129–146.

11. Koc A, Morton DP, Popova E, *et al.* Optimizing project prioritization under budget uncertainty. Proceedings of ICONE 16, ASME 2008 International Conference on Nuclear Engineering; Orlando (FL). 2008.
12. Grotschel M, Monma CL, Stoer M. Polyhedral and computational investigations for designing communication networks with high survivability requirements. *Oper Res* 1995;43(6):1012–1024.
13. Soriano P, Wynants C, Seguin M, *et al.* Design and dimensioning of survivable SDH/SONET networks. In: Sanso B, Soriano P, editors. *Telecommunications network planning*. Chapter 9. Dordrecht: Kluwer Academic Publishers; 1999. pp. 147–167.

BASIC POLYHEDRAL THEORY

VOLKER KAIBEL
Fakultät für Mathematik,
Otto-von-Guericke Universität
Magdeburg, Magdeburg,
Germany

A *polyhedron* is the intersection of finitely many affine halfspaces, where an *affine halfspace* is a set

$$H^{\leq}(a, \beta) = \{x \in \mathbb{R}^n : \langle a, x \rangle \leq \beta\}$$

for some $a \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ (here, $\langle a, x \rangle = \sum_{j=1}^n a_j x_j$ denotes the standard scalar product on $\mathbb{R}^n$). Thus, every polyhedron is the set

$$P^{\leq}(A, b) = \{x \in \mathbb{R}^n : Ax \leq b\}$$

of feasible solutions to a system $Ax \leq b$ of linear inequalities for some matrix $A \in \mathbb{R}^{m \times n}$ and some vector $b \in \mathbb{R}^m$. Clearly, all sets $\{x \in \mathbb{R}^n : Ax \leq b, A'x = b'\}$ are polyhedra as well, as the system $A'x = b'$ of linear equations is equivalent to the system $A'x \leq b'$, $-A'x \leq -b'$ of linear inequalities. A bounded polyhedron is called a *polytope* (where *bounded* means that there is a bound which no coordinate of any point in the polyhedron exceeds in absolute value).

Polyhedra are of great importance for operations research, because they are not only the sets of feasible solutions to linear programs (LP), for which we have beautiful duality results and both practically and theoretically efficient algorithms, but also the solution of (mixed) integer linear programming (MILP) problems can be reduced to linear optimization problems over polyhedra. This relationship has to a large extent formed the backbone of the extremely successful story of MILP and combinatorial optimization over the last few decades.

In the section titled “The Geometry of Polyhedra”, we review those parts of the general theory of polyhedra that are most important with respect to optimization questions,

while in the section titled “Polyhedra and Integrality” we treat concepts that are particularly relevant for integer programming. Most of the “basic polyhedral theory” today is standard textbook knowledge. In the section titled “Pointers to Literature”, for some of the results we provide references to the original papers. There, we also give pointers to proofs of the theorems mentioned in the first two sections, where we mainly refer to the beautiful book by Schrijver [1]. There are, of course, many other excellent treatments of the theory of polyhedra with respect to optimization questions, for example, in the recent survey by Conforti *et al.* [2], in the handbook articles by Schrijver [3] and Burkard [4], as well as in the books by Nemhauser and Wolsey [5], Grötschel *et al.* [6], Bertsimas and Wets [7], Cook *et al.* [8], Wolsey [9], Korte and Vygen [10], or Barvinok [11]. The books by Ziegler [12], and Grünbaum [13] are the most important sources for the general geometric and combinatorial theory of polyhedra, in particular of polytopes. We also refer to the handbook article by Gritzmann and Klee [14] as well as to the one by Bayer and Lee [15].

THE GEOMETRY OF POLYHEDRA

Some Notations

We define $[p] = \{1, \dots, p\}$ and denote by $\mathbb{R}_+ = \{\alpha \in \mathbb{R} : \alpha \geq 0\}$, the set of nonnegative real numbers. A *submatrix* of $M \in \mathbb{R}^{m \times n}$ is a matrix $M_{I,J} \in \mathbb{R}^{I \times J}$ for some $\emptyset \neq I \subseteq [m]$ and $\emptyset \neq J \subseteq [n]$ formed by the rows and columns of M indexed by the elements of I and J , respectively. In particular, $M_{i,j} \in \mathbb{R}$ is the entry in row i and column j . We write $M_{I,*} = M_{I,[n]}$ and $M_{*,J} = M_{[m],J}$, in particular, $M_{i,*} \in \mathbb{R}^n$ and $M_{*,j} \in \mathbb{R}^m$ are the i th row and the j th column of M , respectively. The *kernel* of $M \in \mathbb{R}^{m \times n}$ is $\ker(M) = \{x \in \mathbb{R}^n : Mx = \mathbb{0}\}$. The *identity matrix* $\text{Id}_n \in \mathbb{R}^{n \times n}$ has one-entries on its main diagonal and zeroes elsewhere.

For $x \in \mathbb{R}^n$ and $J \subseteq [n]$, the vector formed by the components of x indexed by elements of J is denoted by $x_J \in \mathbb{R}^J$. We consider, in the

context of matrix multiplication, all vectors as column vectors, and use $(\cdot \cdot \cdot)^t$ to refer to the transposed matrix or vector. We denote the zero vector in $\mathbb{R}^n$ by $\mathbb{0} \in \mathbb{R}^n$ and the standard unit vector by $e_i \in \mathbb{R}^n$, having its i th component equal to one, all other components being zero.

Denoting the set of all determinants of submatrices (formed by arbitrary subsets of rows and columns of equal cardinality, including the empty submatrix, whose determinant is considered to be one) of a matrix $M \in \mathbb{R}^{m \times n}$ by $\delta(M)$, we define

$$\Delta(M) = \left\{ \frac{p}{q} : p, q \in \delta(M) \cup (-\delta(M)), q \neq 0 \right\},$$

and, for every finite set $\emptyset \neq V \subseteq \mathbb{R}^n$, we set $\Delta(V) = \Delta(M)$, where M is any matrix whose set of columns is V . Clearly, for rational matrices $M \in \mathbb{Q}^{m \times n}$ and (finite) sets $V \subseteq \mathbb{Q}^n$ we have $\Delta(M), \Delta(V) \subseteq \mathbb{Q}$.

The *encoding length* of $\alpha = \frac{p}{q} \in \mathbb{Q}$ with $p, q \in \mathbb{Z}$ relatively prime is

$$\langle \alpha \rangle = 1 + \lceil \log_2(|p| + 1) \rceil + \lceil \log_2(|q| + 1) \rceil.$$

For a rational vector $v \in \mathbb{Q}^n$ and a rational matrix $M \in \mathbb{Q}^{m \times n}$, we define

$$\begin{aligned} \langle v \rangle &= n + \sum_{j=1}^n \langle v_j \rangle \quad \text{and} \\ \langle M \rangle &= mn + \sum_{i=1}^m \sum_{j=1}^n \langle M_{i,j} \rangle. \end{aligned}$$

Moreover, we denote the maximum encoding length of any entry in M by $\langle M \rangle_{\max}$, as well as the maximum encoding length of all components of vectors in the finite set $V \subseteq \mathbb{Q}^n$ by $\langle V \rangle_{\max}$.

Basics

Most important, every polyhedron $P \subseteq \mathbb{R}^n$ is *convex*, that is, for all $x, y \in P$ and $\alpha \in [0, 1]$, we have $\alpha x + (1 - \alpha)y \in P$ as well. Moreover, polyhedra are topologically closed subsets of $\mathbb{R}^n$.

As the solution sets to finite systems of linear inequalities, polyhedra generalize affine subspaces, which are the solution sets to

systems of linear equations. The criterion for $Ax = b$ not being solvable via the existence of some multiplier vector $\lambda \in \mathbb{R}^m$ with $\lambda^t A = \mathbb{0}$ and $\langle \lambda, b \rangle \neq 0$ generalizes to systems of linear inequalities in the following way (where both parts of the theorem easily follow from each other).

Theorem 1 [Farkas Lemma]. *For each $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ the following hold:*

- (i) *Either $Ax \leq b$ is solvable or there is some $\lambda \in \mathbb{R}_+^m$ with $\lambda^t A = \mathbb{0}$ and $\langle \lambda, b \rangle < 0$ (but not both).*
- (ii) *Either $Ax = b, x \geq \mathbb{0}$ is solvable or there is some $\lambda \in \mathbb{R}^m$ with $\lambda^t A \geq \mathbb{0}$ and $\langle \lambda, b \rangle < 0$ (but not both).*

It turns out that the “crucial solutions” to systems of linear inequalities are obtained as the unique solutions of certain nonsingular linear equation systems, whose components are well known to be expressible in the following way.

Theorem 2 [Cramer’s rule]. *If $A \in \mathbb{R}^{n \times n}$ is nonsingular, then, for every $b \in \mathbb{R}^n$, the unique solution $x \in \mathbb{R}^n$ with $Ax = b$ is given by*

$$x_j = \frac{\det(A_{*,1}, \dots, A_{*,j-1}, b, A_{*,j+1}, \dots, A_{*,n})}{\det(A)}$$

for all $j \in [n]$.

The following estimates follow from the Leibniz formula for determinants.

Theorem 3. *There is a constant $C > 0$ such that the estimates*

$$\langle \alpha \rangle \leq C \cdot n^2 \cdot \langle M \rangle_{\max} \quad \text{for all } \alpha \in \Delta(M)$$

and

$$\langle \beta \rangle \leq C \cdot n^2 \cdot \langle V \rangle_{\max} \quad \text{for all } \beta \in \Delta(V)$$

hold for all $M \in \mathbb{Q}^{m \times n}$ and for all finite sets $V \subseteq \mathbb{Q}^n$.

Polyhedral and Finitely Generated Cones

A *cone* is a subset $K \subseteq \mathbb{R}^n$ with $\mathbb{0} \in K$ and $\alpha y \in K$ for all $y \in K$ and $\alpha \in \mathbb{R}_+$. A *polyhedral cone* is a polyhedron that is a cone, or, equivalently, a polyhedron $P^\leq(A, \mathbb{0})$ for some $A \in \mathbb{R}^{m \times n}$.

The (convex) *conic hull* of a subset $X \subseteq \mathbb{R}^n$ is the cone

$$\text{ccone}(X) = \left\{ \sum_{x \in X'} \alpha_x x : X' \subseteq X, |X'| < \infty, \right. \\ \left. \alpha_x \geq 0 \text{ for all } x \in X' \right\}$$

(with $\text{ccone}(\emptyset) = \{\mathbb{0}\}$) of all *conic combinations* of the vectors in X . A cone $K \subseteq \mathbb{R}^n$ is *finitely generated*, if there is a finite set $X \subseteq \mathbb{R}^n$ with $K = \text{ccone}(X)$. Every vector in a conic hull can be obtained by a conic combination of few generators.

Theorem 4 [Carathéodory's Theorem, conic version]. *For each $X \subseteq \mathbb{R}^n$ and $y \in \text{ccone}(X)$ there is a linearly independent subset $X' \subset X$ (in particular, $|X'| \leq n$) with $y \in \text{ccone}(X')$.*

The Farkas-Lemma (Part (ii) of Theorem 1) yields a separation theorem for finitely generated cones.

Theorem 5. *If $y \notin \text{ccone}(X)$ for the finite set $X \subseteq \mathbb{R}^n$, then there is some $a \in \mathbb{R}^n$ with*

$$\langle a, x \rangle \leq 0 < \langle a, y \rangle \text{ for all } x \in \text{ccone}(X)$$

(i.e., $\text{ccone}(X) \subseteq H^\leq(a, 0)$, but $y \notin H^\leq(a, 0)$).

The following result implies that every polyhedral cone is finitely generated, which is of utmost importance for the theory of polyhedra.

Theorem 6. *For every matrix $A \in \mathbb{R}^{m \times n}$, there is a finite set $X \subseteq (\Delta(A))^n$ with*

$$P^\leq(A, \mathbb{0}) = \text{ccone}(X).$$

The *polar* of a cone $K \subseteq \mathbb{R}^n$ is the convex cone

$$K^\circ = \{a \in \mathbb{R}^n : \langle a, x \rangle \leq 0 \text{ for all } x \in K\}.$$

The polar of a finitely generated cone $\text{ccone}(X)$ with a finite set $X \subseteq \mathbb{R}^n$ is obviously the polyhedral cone

$$(\text{ccone}(X))^\circ = \{a \in \mathbb{R}^n : \langle x, a \rangle \leq 0 \text{ for all } x \in X\}. \quad (1)$$

From Theorem 5, we also obtain

$$(P^\leq(A, \mathbb{0}))^\circ = \text{ccone}\{A_{1,*}, \dots, A_{m,*}\} \quad (2)$$

for each $A \in \mathbb{R}^{m \times n}$.

Moreover, from Theorem 5, one deduces $(\text{ccone}(X))^{\circ\circ} = \text{ccone}(X)$, from which one finds, by applying Theorem 6 as well as Equation (1) (two times), the following reverse statement to Theorem 6.

Theorem 7. *For every finite set $X \in \mathbb{R}^n$, there is a matrix $A \in (\Delta(X))^{m \times n}$ with*

$$\text{ccone}(X) = P^\leq(A, \mathbb{0}).$$

The Fundamental Structure of Polyhedra

The *homogenization* of a polyhedron $P^\leq(A, b) \subseteq \mathbb{R}^n$ (with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$) is the polyhedral cone

$$\text{homog}(P^\leq(A, b)) = \{(x, \xi) \in \mathbb{R}^n \oplus \mathbb{R} : Ax - \xi b \leq \mathbb{0}, \xi \geq 0\}.$$

Applying Theorem 6 to $\text{homog}(P^\leq(A, b))$ as well as Theorem 7 to the finitely generated cone

$$\text{ccone}(\{(x, 1) \in \mathbb{R}^n \oplus \mathbb{R} : x \in X\} \cup \{(y, 0) \in \mathbb{R}^n \oplus \mathbb{R} : y \in Y\})$$

for finite sets $X, Y \subseteq \mathbb{R}^n$, one obtains the following representation theorem for polyhedra, where

$$\text{conv}(X) = \left\{ \sum_{x \in X'} \alpha_x x : X' \subseteq X, |X'| < \infty, \right. \\ \left. \sum_{x \in X'} \alpha_x = 1, \alpha_x \geq 0 \text{ for all } x \in X' \right\}$$

denotes the *convex hull* of a set $X \subseteq \mathbb{R}^n$, and

$$S + T = \{s + t : s \in S, t \in T\}$$

is the *Minkowski sum* of $S, T \subseteq \mathbb{R}^n$.

Theorem 8 [Weyl–Minkowski Theorem].

- (i) For every $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ there are finite sets $X, Y \subseteq (\Delta(A, b))^n$ with

$$P^{\leq}(A, b) = \text{conv}(X) + \text{ccone}(Y). \quad (3)$$

- (ii) For all finite sets $X, Y \subseteq \mathbb{R}^n$ there are $A \in (\Delta(X \cup Y))^{m \times n}$ and $b \in (\Delta(X \cup Y))^m$ with

$$\text{conv}(X) + \text{ccone}(Y) = P^{\leq}(A, b). \quad (4)$$

Thus, polyhedra can be represented by *outer descriptions* (intersections of finitely many affine halfspaces) and by *inner descriptions* (Minkowski sums of a polytope and a finitely generated cone). Clearly, algebraically an outer description may contain both linear inequalities and linear equations. Depending on the context, one type of description of a polyhedron can be significantly more convenient to deal with than the other. For instance, from inner descriptions one concludes readily that images of polyhedra under linear maps are polyhedra as well (see the section titled “Projections of Polyhedra”). In turn, the fact that also preimages of polyhedra under linear maps are polyhedra is easy to prove via outer descriptions. From outer descriptions of polyhedra, one also finds immediately that intersections of finitely many polyhedra are polyhedra.

A *rational polyhedron* is a polyhedron for which A and b , or, equivalently, X and Y , in Equations (3) and (4) can be chosen to be rational matrices, vectors, and sets, respectively. Denoting, for a rational polyhedron $P \subseteq \mathbb{R}^n$, by $\langle P \rangle_{\max}^{\text{outer}}$ the smallest number α such that there is an outer description $P = P^{\leq}(A, b)$ of P with a rational matrix A and a rational vector b with $\langle(A, b)\rangle_{\max} \leq \alpha$, and by $\langle P \rangle_{\max}^{\text{inner}}$ the smallest number β such that there is an inner description $P = \text{conv}(X) + \text{ccone}(Y)$ of P with rational finite sets

$X, Y \subseteq \mathbb{Q}^n$ with $\langle X \cup Y \rangle_{\max} \leq \beta$, we obtain the following result from Theorem 3.

Theorem 9. *There is a constant $C > 0$ such that*

$$\begin{aligned} \langle P \rangle_{\max}^{\text{inner}} &\leq C \cdot n^2 \cdot \langle P \rangle_{\max}^{\text{outer}} \quad \text{and} \\ \langle P \rangle_{\max}^{\text{outer}} &\leq C \cdot n^2 \cdot \langle P \rangle_{\max}^{\text{inner}} \end{aligned}$$

hold for every rational polyhedron $P \subseteq \mathbb{R}^n$.

In particular, whenever the rational system $Ax \leq b$ (with $A \in \mathbb{Q}^{m \times n}$ and $b \in \mathbb{Q}^m$) has any solution, then it also has a solution whose encoding length is bounded by a polynomial in $\langle(A, b)\rangle$. This shows that the linear programming feasibility problem is contained in the complexity class NP, and, via Theorem 1, also in coNP. (Of course, it is well known that this problem is even solvable in polynomial time).

However, the smallest possible number of inequalities $Ax \leq b$ in an outer description and the smallest possible cardinalities of the sets X and Y in an inner description of a rational polyhedron P are not bounded polynomially by each other, in general.

The *characteristic cone* (or *recession cone*) of a polyhedron $P \subseteq \mathbb{R}^n$ is

$$\begin{aligned} \text{char}(P) = \\ \{y \in \mathbb{R}^n : x + \text{cone}\{y\} \subseteq P \text{ for all } x \in P\}. \end{aligned}$$

Theorem 10. *If $\emptyset \neq P = P^{\leq}(A, b) = \text{conv}(X) + \text{ccone}(Y) \subseteq \mathbb{R}^n$ is a polyhedron (with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $X, Y \subseteq \mathbb{R}^n$ finite), then we have*

$$\text{ccone}(Y) = \text{char}(P) = P^{\leq}(A, \mathbb{O}).$$

The *lineality space* of a polyhedron $P \subseteq \mathbb{R}^n$ is the largest linear subspace

$$\text{lineal}(P) = \text{char}(P) \cap (-\text{char}(P))$$

contained in $\text{char}(P)$.

Theorem 11. *For a polyhedron $\emptyset \neq P = P^{\leq}(A, b)$ (with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$) we have*

$$\text{lineal}(P) = \ker(A).$$

A nonempty polyhedron $\emptyset \neq P \subseteq \mathbb{R}^n$ with lineality space $\text{lineal}(P) = \{\mathbb{O}\}$ is called *pointed*. For most purposes, it is sufficient to consider pointed polyhedra. In fact, if a polyhedron has a nontrivial lineality space $L \neq \{\mathbb{O}\}$, then, in most contexts, it is sufficient to investigate instead of P its orthogonal projection to the orthogonal complement of L , which is a pointed polyhedron. Therefore, subsequently we mainly consider pointed polyhedra. For instance, all polyhedra that are contained in the nonnegative orthant $\mathbb{R}_+^n$ as well as all polytopes are pointed.

Faces of Polyhedra

For $a \in \mathbb{R}^n \setminus \{\mathbb{O}\}$ and $\beta \in \mathbb{R}$, we denote the boundary hyperplane of the affine halfspace $H^\leq(a, \beta)$ by

$$H^=(a, \beta) = \{x \in \mathbb{R}^n : \langle a, x \rangle = \beta\}.$$

A (proper) face of a polyhedron $P \subseteq \mathbb{R}^n$ is the intersection $F = P \cap H^=(a, \beta)$ of P with the boundary hyperplane of some halfspace $H^\leq(a, \beta) \supseteq P$ containing P . The face F is said to be *defined* by the inequality $\langle a, x \rangle \leq \beta$ in this case. In addition, $\emptyset$ and P themselves are considered (*trivial*) faces of P (defined by $\langle \mathbb{O}, x \rangle \leq -1$ and $\langle \mathbb{O}, x \rangle \leq 0$, respectively). Nonempty faces are particularly important for optimization, because they are the (nonempty) sets of optimal solutions to linear optimization problems over polyhedra.

The following result (which is a generalization of Part (i) of Theorem 1) provides a characterization of those inequalities that are valid for (and thus define faces of) a nonempty polyhedron.

Theorem 12. *Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $a \in \mathbb{R}^n$, and $\beta \in \mathbb{R}$ with $P = P^\leq(A, b) \neq \emptyset$. The inequality $\langle a, x \rangle \leq \beta$ is valid for (all x in) P if and only if there is some $\lambda \in \mathbb{R}_+^m$ with $\lambda^t A = a$ and $\langle \lambda, b \rangle \leq \beta$.*

Every face $F \neq \emptyset$ of a polyhedron $P = P^\leq(A, b) = \text{conv}(X) + \text{ccone}(Y)$ (with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $X, Y \subseteq \mathbb{R}^n$ finite) is a polyhedron as well with

$$F = \text{conv}(X \cap F) + \text{ccone}(Y \cap \text{char}(F)).$$

Moreover, with

$$I = \text{Eq}_{Ax \leq b}(F) = \{i \in [m] : F \subseteq H^=(A_{i,*}, b_i)\},$$

we have $F = \{x \in P : A_{I,*}x = b_I\}$, and, conversely, for every $I \subseteq [m]$, the set $F = \{x \in P : F \subseteq H^=(A_{I,*}, b_i)\}$ is a face of P with $I \subseteq \text{Eq}_{Ax \leq b}(F)$.

In particular, every polyhedron has finitely many faces. Partially ordered by inclusion, they form the *face lattice* $\mathcal{L}(P)$ of P (including $\emptyset$ and P). For each face $F \neq \emptyset$ of the polyhedron P we have $\text{lineal}(F) = \text{lineal}(P)$. Hence, the (nonempty) faces of pointed polyhedra are pointed polyhedra as well.

The maximal elements in $\mathcal{L}(P) \setminus \{P\}$ are called the *facets* of P . A face of P is a facet if and only if $\dim(F) = \dim(P) - 1$ holds (where the *dimension* $\dim(P)$ of a polyhedron is the affine dimension of its affine hull $\text{aff}(P)$). The minimal elements in $\mathcal{L}(P) \setminus \{\emptyset\}$ for a pointed polyhedron P are called the *vertices* of P . The vertices of a polyhedron P are the faces of dimension zero, that is, the faces that contain exactly one point (which, of course, is also called *vertex*).

Theorem 13. *For a point $v \in P$ in a polyhedron $P = P^\leq(A, b)$ (with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$) the following statements are pairwise equivalent.*

1. *The point v is a vertex of P .*
2. *There are no two points $x, x' \in P \setminus \{v\}$ with $v \in \text{conv}\{x, x'\}$.*
3. *There is some $I \subseteq [m]$, $|I| = n$ such that $A_{I,*}$ is nonsingular with $v = A_{I,*}^{-1}b_I$.*

A pointed polyhedral cone K has exactly one vertex, namely $\mathbb{O}$, and therefore, one is more interested in the minimal elements of $\mathcal{L}(K) \setminus \{\emptyset, \{\mathbb{O}\}\}$, which are called the *extreme rays* of K . The extreme rays of a (pointed) polyhedral cone are its faces of dimension one. The one-dimensional unbounded faces of a general pointed polyhedron are called its *extreme rays*; the one-dimensional bounded faces are the *edges*. The edges of a pointed polyhedron P are of the form $\text{conv}\{v, w\}$ with two vertices $v \neq w$ of P , and every extreme

ray of P can be written as $v + R$ with a vertex v of P and an extreme ray R of $\text{char}(P)$ (which is pointed if P is pointed).

An outer description $P = \{x \in \mathbb{R}^n : A^\leq x = b^\leq, A^\leq x \leq b^\leq\}$ of a polyhedron P is *irredundant* if removing any equation or any inequality from the system results in a different (larger) polyhedron, and turning any inequality in the system into an equation results in a different (smaller) polyhedron.

Theorem 14. *Let $\emptyset \neq P \subseteq \mathbb{R}^n$ be a polyhedron, $A^{(1)} \in \mathbb{R}^{m(1) \times n}$, $b^{(1)} \in \mathbb{R}^{m(1)}$, $A^{(2)} \in \mathbb{R}^{m(2) \times n}$, and $b^{(2)} \in \mathbb{R}^{m(2)}$ with*

$$P \subseteq \{x \in \mathbb{R}^n : A^{(1)}x = b^{(1)}, A^{(2)}x \leq b^{(2)}\} \quad (5)$$

such that, for all $i \in [m(2)]$, we have $P \not\subseteq H^=(A^{(2)}_{i,}, b_i)$.*

- (i) *Equality in Equation (5) holds if and only if*
 - $\text{aff}(P) = \{x \in \mathbb{R}^n : A^{(1)}x = b^{(1)}\}$, and
 - *for each facet F of P , there is an $i \in [m(2)]$ such that $\langle A^{(2)}_{i,*}, x \rangle \leq b_i$ defines the facet F of P .*
- (ii) *If equality holds in Equation (5), then the right-hand side of Equation (5) is an irredundant outer description of P if and only if*
 - $A^{(1)}$ has full row rank, and
 - *the inequalities in $A^{(2)}x \leq b^{(2)}$ define pairwise distinct facets of P .*

An inner description $P = \text{conv}(X) + \text{ccone}(Y)$ of a polyhedron P is *irredundant* if removing any point from X or any vector from Y results in a different (smaller) polyhedron.

Theorem 15. *Let $\emptyset \neq P \subseteq \mathbb{R}^n$ be a pointed polyhedron, and let $X, Y \subseteq \mathbb{R}^n$ be finite sets with $X \subseteq P$ and $Y \subseteq \text{char}(P)$.*

- (i) *We have $P = \text{conv}(X) + \text{ccone}(Y)$ if and only if*
 - X contains all vertices of P , and
 - Y contains a nonzero vector from each extremal ray of $\text{char}(P)$.

- (ii) *If $P = \text{conv}(X) + \text{ccone}(Y)$ holds, then $\text{conv}(X) + \text{ccone}(Y)$ is an irredundant inner description of P if and only if*

- X is the set of vertices of P , and
- Y contains exactly one nonzero vector from each extremal ray of $\text{char}(P)$.

As for a pointed polyhedron $P = \text{conv}(X) + \text{ccone}(Y)$ (with finite sets $X, Y \subseteq \mathbb{R}^n$, $c \in \mathbb{R}^n$, and $\omega = \max\{\langle c, x \rangle : x \in P\}$, we have $\omega = \infty$ if $\langle c, y \rangle > 0$ for some $y \in Y$, and $\omega = \max\{\langle c, x \rangle : x \in X\}$ otherwise, Theorem 15 implies that a bounded linear optimization problem over a pointed polyhedron P attains its optimum in a vertex of P .

Projections of Polyhedra

We mentioned above that the projection $P = \pi(Q)$ of a polyhedron $Q \subseteq \mathbb{R}^d$ via a linear map $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ with $\pi(y) = Ty$ for a matrix $T \in \mathbb{R}^{n \times d}$ is a polyhedron, as, for all $X, Y \subseteq \mathbb{R}^d$, we have

$$\begin{aligned} \pi(\text{conv}(X) + \text{ccone}(Y)) \\ = \text{conv}(\pi(X)) + \text{ccone}(\pi(Y)). \end{aligned}$$

We say that a face F of Q is π -compatible if it can be defined by an inequality $\langle T^t a, y \rangle \leq \beta$ (valid for Q) for some $a \in \mathbb{R}^n$.

Theorem 16. *For a polyhedron $Q \subseteq \mathbb{R}^d$, a linear projection $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$, and the polyhedron $P = \pi(Q)$, the map defined via $F \mapsto \pi(F)$ is an isomorphism between the sublattice of $\mathcal{L}(Q)$ formed by the π -compatible faces and $\mathcal{L}(P)$.*

An extended formulation for a polyhedron $P \subseteq \mathbb{R}^n$ is an outer description $Q = P^\leq(A, b)$ of some polyhedron $Q \subseteq \mathbb{R}^d$ along with a linear projection $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ with $\pi(Q) = P$. Such an extended formulation of a polyhedron P can be much simpler (e.g., in terms of the number of inequalities) than any outer description of P if the complexity of the facets of P is hidden in lower dimensional parts of the face lattice $\mathcal{L}(Q)$ (see Theorem 16). As owing to

$$\begin{aligned} \max \{\langle c, x \rangle : x \in P\} &= \max \{\langle T^t c, y \rangle : y \in Q\} \\ (\text{for all } c \in \mathbb{R}^n), \end{aligned}$$

linear optimization problems over $P = \pi(Q)$ can be solved by solving linear optimization problems over Q , extended formulations play an important role in modern mathematical optimization [16].

We conclude the first part of the article by considering the question of how to derive an outer description of a polyhedron from an extended formulation. Here, the fundamental result follows via Theorem 12.

Theorem 17. *Let $Q = P^\leq(D, g) \subseteq \mathbb{R}^d$ be a polyhedron with $D \in \mathbb{R}^{q \times d}$ and $g \in \mathbb{R}^q$, suppose the linear projection $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is defined via $\pi(y) = Ty$ for a matrix $T \in \mathbb{R}^{n \times d}$ and all $y \in \mathbb{R}^d$, and let $\bar{T}$ be any matrix whose rows form a basis of $\ker(T)$.*

If $L \in \mathbb{R}_+^{m \times d}$ is a matrix whose rows generate the projection cone

$$\begin{aligned} \left\{ \lambda \in \mathbb{R}_+^q : \lambda^t (D\bar{T}^t) = \mathbb{0} \right\} \\ = \text{ccone} \{ L_{1,*}, \dots, L_{m,*} \}, \end{aligned}$$

then every $A \in \mathbb{R}^{m \times n}$ with $AT = LD$ satisfies

$$\pi(Q) = P^\leq(A, b) \cap \pi(\mathbb{R}^d)$$

with $b = Lg$.

If, in the situation of Theorem 17, the projection $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is the orthogonal projection to the first n coordinates, then the projection cone is simply

$$\left\{ \lambda \in \mathbb{R}_+^q : \langle D_{*,j}, \lambda \rangle = 0 \text{ for all } j \in [d] \setminus [n] \right\}, \quad (6)$$

and A can be chosen to consist of the first n columns of LD . Furthermore, if $n = d - 1$ holds, then a finite generating system for the projection cone (6) can be obtained from the relation:

$$\begin{aligned} \left\{ \lambda \in \mathbb{R}_+^{r+s} : \sum_{i=1}^{r+s} \lambda_i \delta_i = 0 \right\} \\ = \text{ccone} \{ \delta_k e_\ell - \delta_\ell e_k : k \in [r], \ell \in [r+s] \setminus [r] \}, \end{aligned}$$

for all numbers $\delta_1, \dots, \delta_r > 0 > \delta_{r+1}, \dots, \delta_{r+s}$ (which one can easily establish by induction on $r + s$).

Theorem 18 [Fourier–Motzkin elimination]. *For a polyhedron $Q = P^\leq(D, g) \subseteq \mathbb{R}^d$ (with $B \in \mathbb{R}^{q \times d}$ and $g \in \mathbb{R}^q$) and the sets*

$$I^{>0} = \{i \in [q] : D_{i,d} > 0\}$$

$$I^=0 = \{i \in [q] : D_{i,d} = 0\}$$

$$I^{<0} = \{i \in [q] : D_{i,d} < 0\},$$

the polyhedron in $\mathbb{R}^{d-1}$ arising from Q by orthogonal projection to the first $d - 1$ coordinates is the set of all $x \in \mathbb{R}^{d-1}$ that satisfy the following system:

$$\begin{aligned} \langle D_{i,*}, x \rangle &\leq g_i \quad \text{for all } i \in I^{>0} \\ \langle D_{k,d} D_{\ell,*} - D_{\ell,d} D_{k,*}, x \rangle &\leq D_{k,d} g_\ell - D_{\ell,d} g_k \\ &\quad \text{for all } k \in I^{>0}, \ell \in I^{<0}. \end{aligned}$$

By applying the Fourier–Motzkin elimination method (Theorem 18) iteratively, one can compute an outer description of a polyhedron's orthogonal projection to any coordinate subspace from an outer description of the polyhedron. Note that the sizes of the descriptions may blow up exponentially (even if, after each iteration, one removes redundant constraints).

The case of general projections can be solved by Fourier–Motzkin elimination as follows, where we adopt the notation discussed earlier. Let $\tilde{n}$ be the rank of T , and assume that the submatrix $T_{[\tilde{n}], [\tilde{n}]}$ of T is nonsingular. Denoting the nonsingular matrix that equals $T_{\tilde{n},*}$ on its first $\tilde{n}$ rows and that has the vectors e_i , for $i = \tilde{n} + 1, \dots, d$, in its other rows by $\tilde{T} \in \mathbb{R}^{d \times d}$, and the orthogonal projection to the first $\tilde{n}$ coordinates by $\tilde{\pi} : \mathbb{R}^d \rightarrow \mathbb{R}^{\tilde{n}}$, we have $\pi(y)_{[\tilde{n}]} = \tilde{\pi}(\tilde{T}y)$ for all $y \in \mathbb{R}^d$. Thus, with $\tilde{D} = D\tilde{T}^{-1}$ and $\tilde{Q} = P^\leq(\tilde{D}, g)$, we find $P = \{x \in \pi(\mathbb{R}^d) : x_{\tilde{n}} \in \tilde{\pi}(\tilde{Q})\}$.

One reason for the interest in computing projections of polyhedra is that one can convert inner to outer and outer to inner description by means of such computations. Indeed, it follows from the discussion of homogenization and the concept of polarity that all such conversions can be done by any method that computes, for a given finite set $X \subseteq \mathbb{R}^n$, a matrix $A \in \mathbb{R}^{m \times n}$ with $\text{ccone}(X) = P^\leq(A, \mathbb{0})$ (see Theorem 7). Denoting by $T \in \mathbb{R}^{n \times d}$ a

matrix whose set of columns is X , we have, by definition, $\text{ccone}(X) = \pi(Q)$, where $\pi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is the projection defined via $\pi(y) = Ty$ for all $y \in \mathbb{R}^d$ and $Q = \mathbb{R}_+^d$. Thus, computing an outer description of $\pi(Q)$ from the outer description $\mathbb{R}_+^d = P^\leq(-\text{Id}_d, \mathbb{O})$ of Q solves the problem.

POLYHEDRA AND INTEGRALITY

Integral points in polyhedra play a crucial role in integer linear programming and combinatorial optimization. The fundamental concept here is the *integer hull*

$$P_I = \text{conv}(P \cap \mathbb{Z}^n)$$

of a polyhedron $P \subseteq \mathbb{R}^n$. In this second part, we describe the most fundamental results on integer hulls, which are important, since, on the one hand, they satisfy

$$\max\{\langle c, x \rangle : x \in P \cap \mathbb{Z}^n\} = \max\{\langle c, x \rangle : x \in P_I\}$$

for all $c \in \mathbb{R}^n$, while on the other hand, as convex objects, they are much more convenient to deal with than the discrete sets $P \cap \mathbb{Z}^n$. In particular, we treat *integral polyhedra*, that is, polyhedra P with $P_I = P$. Identifying a polyhedron P as integral is particularly pleasant, as in this case integer linear optimization problems over P can be treated as (continuous) linear optimization problems over P .

Theorem 19. *For $A \in \mathbb{Q}^{m \times n}$ and $b \in \mathbb{Q}^m$, defining an integral polyhedron $P^\leq(A, b)$, one can, for every $c \in \mathbb{Q}^n$, find $x^* \in P^\leq(A, b) \cap \mathbb{Z}^n$ with*

$$\langle c, x^* \rangle = \max\{\langle c, x \rangle : x \in P^\leq(A, b) \cap \mathbb{Z}^n\}$$

(or conclude that no such x^* exists) in time bounded by a polynomial in $\langle A \rangle + \langle b \rangle + \langle c \rangle$.

Integer Points in Polyhedra

Similarly to the parameterization (3) of points in a polyhedron by an inner description, for a *rational* polyhedron P , we can also

parameterize the set of all *integral* points in P . For a finite set $Y \subseteq \mathbb{R}^n$, we denote by

$$\text{mono}(Y) = \left\{ \sum_{y \in Y} \alpha_y y : \alpha_y \in \mathbb{N} \text{ for all } y \in Y \right\}$$

(with $\mathbb{N} = \{0, 1, 2, \dots\}$), the submonoid of $\mathbb{R}^n$ which is generated by Y .

Theorem 20. *For every rational polyhedron $P \subseteq \mathbb{R}^n$ there are finite sets $X, Y \subseteq \mathbb{Z}^n$ with*

$$P \cap \mathbb{Z}^n = X + \text{mono}(Y) \quad \text{and}$$

$$\text{char}(P) = \text{ccone}(Y),$$

such that $\langle X \cup Y \rangle_{\max}$ is bounded by a polynomial in n and $\langle P \rangle_{\max}^{\text{inner}}$ (and thus, by a polynomial in n and $\langle P \rangle_{\max}^{\text{outer}}$).

A finite set $H \subseteq \mathbb{Z}^n$ is called an *integral Hilbert basis* (of $\text{ccone}(H)$) if

$$\text{ccone}(H) \cap \mathbb{Z}^n = \text{mono}(H)$$

holds. The existence statement of the following theorem is a special case of Theorem 20.

Theorem 21. *Every rational polyhedral cone K has an integral Hilbert basis; if K is pointed, then the integral Hilbert basis of K is uniquely determined.*

Moreover, Theorem 20 implies that whenever the rational system $Ax \leq b$ (with $A \in \mathbb{Q}^{m \times n}$ and $b \in \mathbb{Q}^m$) has any integral solution, it also has an integral solution whose encoding length is bounded by a polynomial in $\langle (A, b) \rangle$. This shows that the integer linear programming feasibility problem is contained in the complexity class NP (it is, however, not in coNP, unless $\text{NP} = \text{coNP}$).

Finally, from Theorem 20 one can derive that the integral hulls of *rational* polyhedra are rational polyhedra as well.

Theorem 22. *For each rational polyhedron $P \subseteq \mathbb{R}^n$, the integer hull P_I of P is a rational polyhedron, for which $\langle P_I \rangle_{\max}^{\text{outer}}$ is bounded by a polynomial in n and $\langle P \rangle_{\max}^{\text{outer}}$.*

The number of facets of P_I is, however, in general not bounded polynomially in the number of facets of P . Furthermore, the integer hull of a nonrational polyhedron need not even be a polyhedron (see, e.g., $P = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \leq \sqrt{2}x_1, x_1 \geq 1\}$).

A generalization of Theorem 22 to *mixed-integer hulls* of rational polyhedra holds as well: If $P \subseteq \mathbb{R}^n$ is a rational polyhedron, then $\text{conv}\{x \in P : x_J \in \mathbb{Z}^J\}$ is a rational polyhedron for every $J \subseteq [n]$.

We end this section by a crucial criterion for integrality of (pointed rational) polyhedra.

Theorem 23. *A pointed rational polyhedron is integral (i.e., $P_I = P$) if and only if all vertices of P are integral.*

Total Dual Integrality

A rational system $Ax \leq b$ with $A \in \mathbb{Q}^{m \times n}$ and $b \in \mathbb{Q}^m$ defining a nonempty polyhedron $P^\leq(A, b) \neq \emptyset$ is called *totally dual integral (TDI)* if, for every $c \in \mathbb{Z}^n$ with $\max\{c, x\} : Ax \leq b\} < \infty$, there is some $y^* \in \mathbb{N}^m$ with $A^t y^* = c$ and

$$\langle b, y^* \rangle = \min \{ \langle b, y \rangle : A^t y = c, y \in \mathbb{R}_+^m \},$$

that is, the dual problem to $\max\{c, x\} : Ax \leq b\}$ has an integral optimal solution y^* . It is convenient to consider all rational systems $Ax \leq b$ with $P^\leq(A, b) = \emptyset$ TDI as well.

Theorem 24. *A system $Ax \leq b$ with integral matrix $A \in \mathbb{Z}^{m \times n}$ and rational right-hand side vector $b \in \mathbb{Q}^m$ is TDI if and only if, for every face $F \neq \emptyset$ of $P^\leq(A, b)$, the set*

$$\{A_{i,*} : i \in \text{Eq}_{Ax \leq b}(F)\}$$

is an integral Hilbert basis.

Applying Theorem 21 to the rational normal cones

$$\{c \in \mathbb{R}^n : \langle c, x^* \rangle = \max \{ \langle c, x \rangle : x \in P \} \text{ for all } x^* \in F\}$$

of the (minimal) faces F of a rational polyhedron P , one can construct TDI systems as in the following result.

Theorem 25. *Let $P \subseteq \mathbb{R}^n$ be a rational polyhedron.*

1. *There is an integral matrix $A \in \mathbb{Z}^{m \times n}$ and a rational vector $b \in \mathbb{Q}^m$ with $P = P^\leq(A, b)$ such that $Ax \leq b$ is TDI, where b can be chosen to be integral if P is integral.*
2. *If there is a rational matrix $A \in \mathbb{Q}^{m \times n}$ and an integral vector $b \in \mathbb{Z}^m$ with $P = P^\leq(A, b)$ such that $Ax \leq b$ is TDI, then P is integral.*

Thus, every rational polyhedron admits an outer descriptions by a TDI-system with integral (left-hand-side) coefficient matrix, though, in general, this will not be an irredundant description in the sense of the section titled “Faces of Polyhedra”. Nevertheless, such a description can provide much structural insight, for example, within the theory of cutting planes for nonintegral polyhedra. The algorithmic problem, however, to decide whether a given rational system $Ax \leq b$ is TDI is coNP-complete (even for A restricted to the set of node-edge incidence matrices, see the section titled “Total Unimodularity,” of undirected graphs).

Outer descriptions of (integral) polyhedra with integral right-hand side vectors are very important, as, due to Theorem 25(2) (and the definition of TDI), they yield strong duality relations for certain integer linear optimization problems.

Theorem 26. *If the system $Ax \leq b$ is TDI with $A \in \mathbb{Q}^{m \times n}$ and an integral right-hand side $b \in \mathbb{Z}^m$, and $P = P^\leq(A, b) \neq \emptyset$ holds, then, for every $c \in \mathbb{Z}^n$ with $\max\{c, x\} : x \in P\} < \infty$, we have*

$$\begin{aligned} \max\{c, x\} : Ax \leq b, x \in \mathbb{Z}^n\} \\ = \min\{\langle b, y \rangle : A^t y = c, y \geq 0, y \in \mathbb{Z}^m\}. \end{aligned} \quad (7)$$

Total Unimodularity

A matrix $A \in \{-1, 0, 1\}^{m \times n}$ is *totally unimodular (TU)* if all square submatrices of A have determinant -1 , 0 , or 1 . The (version for pointed polyhedra of the) following result is a consequence of Theorem 23 and Theorem 13

(see Part 3) obtained by using Cramer's rule (Theorem 2).

Theorem 27. *If $A \in \{-1, 0, 1\}^{m \times n}$ is totally unimodular and $b \in \mathbb{Z}^m$ is integral, then $P^\leq(A, b)$ is an integral polyhedron.*

If $A \in \{-1, 0, 1\}^{m \times n}$ is TU, then so are all submatrices of A , (A, Id_m) , $(A, -A)$, and A^t , where the latter allows the following conclusion from Theorem 27.

Theorem 28. *If $A \in \{-1, 0, 1\}^{m \times n}$ is totally unimodular, then, for every $b \in \mathbb{Q}^m$, the system $A \leq b$ is TDI.*

In particular, if $A \in \{-1, 0, 1\}^{m \times n}$ is TU, then the strong duality relation (7) holds for all integral $b \in \mathbb{Z}^m$ and $c \in \mathbb{Z}^n$ (for which the respective optimal values are finite).

The following criterion is extremely useful for establishing total unimodularity of matrices.

Theorem 29 [Criterion of Ghouila-Houri]. *A matrix $A \in \{-1, 0, 1\}^{m \times n}$ is totally unimodular if and only if, for every subset $I \subseteq [m]$ of row indices, there is a partitioning $I = I^+ \uplus I^-$ (with $I^+ \cap I^- = \emptyset$) such that*

$$\sum_{i \in I^+} A_{i,*} - \sum_{i \in I^-} A_{i,*} \in \{-1, 0, 1\}^n$$

holds. (Clearly, a similar characterization via all subsets of column indices holds.)

From Theorem 29 one readily deduces that every matrix with entries from $\{-1, 0, 1\}$ that has at most one positive and at most one negative entry per column is TU. In particular, the *node-arc incidence matrix* $\text{inc}(D) \in \{-1, 0, 1\}^{V \times A}$ of a directed graph $D = (V, A)$ (with $A \subseteq V \times V$, defined via

$$\text{inc}(D)_{v,a} = \begin{cases} -1 & \text{if } a = (v, w) \text{ for some } w \in V \\ 1 & \text{if } a = (u, v) \text{ for some } u \in V \\ 0 & \text{otherwise} \end{cases}$$

for all $v \in V$ and $a \in A$, is TU. Thus, the set

$$\{x \in \mathbb{R}^A : \text{inc}(D)x = 0, \ell \leq x \leq u\}$$

of *circulations* in D respecting the *integral* lower and upper bounds $\ell, u \in \mathbb{Z}^A$ is an integral polytope.

The *node-edge incidence matrix* $\text{inc}(G) \in \{0, 1\}^{V \times E}$ of an undirected graph $G = (V, E)$ is defined via

$$\text{inc}(G)_{v,e} = \begin{cases} 1 & \text{if } e = \{v, w\} \text{ for some } w \in V \\ 0 & \text{otherwise} \end{cases}$$

for all $v \in V$ and $e \in E$. Theorem 29 yields that the node-edge incidence matrix of a graph $G = (V, E)$ is TU if and only if the graph is bipartite; that is, there is a partitioning $V = S \uplus T$ (with $S \cap T = \emptyset$) such that $E \subseteq \{\{s, t\} : s \in S, t \in T\}$. In particular, the matching polytope of a bipartite graph $G = (V, E)$ (i.e., the convex hull of the characteristic vectors $\chi(M) \in \{0, 1\}^E$ —with $\chi(M)_e = 1$ if and only if $e \in M$ —of all *matchings* $M \subseteq E$ —with $e \cap e' = \emptyset$ for all $e, e' \in M, e \neq e'$) has

$$\begin{aligned} \sum_{e \in E: v \in e} x_e &\leq 1 && \text{for all } v \in V \\ x_e &\geq 0 && \text{for all } e \in E \end{aligned}$$

as an outer description.

The most important class of TU matrices is formed by the *network matrices*. Such a network matrix $N \in \{-1, 0, 1\}^{A_T \times A}$ arises from a directed graph $D = (V, A)$ and a subset $A_T \subseteq A$ of $|V| - 1$ arcs that, viewed as undirected edges, form a spanning tree on V , by setting, for each $a' \in A_T$ and $a = (v, w) \in A$,

$$N_{a',a} = \begin{cases} 1 & \text{if the } v-w \text{ path in } A_T \text{ uses } a' \text{ in} \\ & \text{forward direction} \\ -1 & \text{if the } v-w \text{ path in } A_T \text{ uses } a' \text{ in} \\ & \text{backward direction} \\ 0 & \text{otherwise} \end{cases}$$

Theorem 30. *Network matrices are totally unimodular.*

In fact, network matrices are the crucial building blocks of the whole class of totally unimodular matrices, as every totally unimodular matrix arises from network

matrices and two special totally unimodular 5×5 matrices by certain operations that preserve total unimodularity. From such structural results, one can derive polynomial time algorithms for testing matrices A for total unimodularity (while, as remarked at the end of the section “Total Dual Integrality,” testing a system $Ax \leq b$ for total dual integrality is coNP-complete).

POINTERS TO LITERATURE

Whenever possible, we provide, for the results mentioned in the article, pointers to proofs in Schrijver [1], which is also a great source for historical notes and many more references than listed here.

The Farkas-Lemma (Theorem 1) dates back to work by Farkas as well as by Minkowski at the end of the 19th century [1, Section 7.3]. There are several possibilities to prove the theorem. In fact, the Farkas-Lemma is a special case of more general separation theorems in convex analysis [17, Chapter 2]. A particularly nice and purely linear algebraic proof was given by Conforti *et al.* [18], who derive the Farkas-Lemma from the corresponding obstruction of the solvability of linear equation systems. Theorem 2 was proposed by Cramer (1750); see Schrijver [1, Section 3.1]. Theorem 3 follows from Schrijver [1, Theorem 3.2]; see also Edmonds [19]. In 1911, Theorem 4 was proved by Carathéodory [20]; see also Schrijver [1, Theorem 7.1]. Theorems 6–8 have their origins in the work of Farkas, Minkowski [21], and Weyl [22]; see also Schrijver [1, Section 7.2]. The statements on the components of the vectors and on the entries of the matrices (as well as Theorem 9) follow, for example, from (the proofs in) Schrijver [1, Section 10.2]. An elementary proof of Theorem 6 can be found in Kaibel [23]; see also Schrijver [1, Corollary 7.1a]. For proofs of Theorems 10 and 11 (on the characteristic cones and lineality spaces of polyhedra), we refer to Schrijver [1, Section 8.2]. Theorem 12 is present as Corollary 7.1h in Schrijver [1]. For the other statements in the section titled “Faces of Polyhedra,” see Schrijver [1, Sections 8.3–8.9]. Theorem 16

is folklore (we are not aware of any other explicit reference, thus we refer to Kaibel [24]). Theorem 17 is usually formulated for orthogonal projections to coordinate subspaces only [2, Section 2.4]. An explicit proof in the general setting can be found in Kaibel [24]. The Fourier–Motzkin method (Theorem 18) is treated in Schrijver [1, Section 12.2]. The method was devised by Motzkin [25], where the idea goes back to the work of Fourier in the early 19th century. For the algorithmic problem of converting representations of polyhedra, we refer to the survey by Seidel [26] and to the software system `polymake` by Gawrilow and Joswig [27] (<http://www.opt.tu-darmstadt.de/polymake/>).

The proof of Theorem 19 relies on the fact that both (continuous) linear programs [28] as well as systems of linear Diophantine equations can be solved in polynomial time [1, Theorem 16.2]. For proofs of Theorems 20–22, we refer to Schrijver [1, Sections 16.2–16.4 and 17.2]. The fact that the integer hull of a rational polyhedron is a rational polyhedron (Theorem 22) was proposed by Meyer [29], and the notion of *Hilbert bases* has been introduced by Giles and Pulleyblank [30], where the ideas of the proof of Theorem 21 date back to Gordan [31]. The concept of *total dual integrality* has been invented by Edmonds and Giles [32,33]. See Schrijver [1, Section 22.3] for proofs of Theorems 24 and 25 (the results being provided by Giles and Pulleyblank [30] and Schrijver [34]). The coNP-hardness of the TDI-property has been established by Ding *et al.* [35]. Proofs of the results on total unimodularity can be found in Schrijver [1, Chapter 19]. The connection between TU matrices and integral polyhedra (Theorem 27 and a similar characterization of total unimodularity) was established by Hoffman and Kruskal [36]. Theorem 29 has been proved by Ghouila-Houri [37]. The total unimodularity of network matrices (Theorem 30) was suggested by Tutte [38]. The decomposition theorem for TU matrices mentioned after Theorem 30 has been proved by Seymour [39]. Cunningham and Edmonds [40] derived a polynomial time test for total unimodularity from that theorem; the asymptotically fastest known algorithm is due to Truemper [41].

REFERENCES

1. Schrijver A. Theory of linear and integer programming, Wiley-Interscience series in discrete mathematics. Chichester: John Wiley & Sons, Ltd.; 1986. A Wiley-Interscience Publication.
2. Conforti M, Cornuéjols G, Zambelli G. Polyhedral approaches to mixed integer linear programming. In: Jünger M., *et al.* editors. 50 Years of integer programming 1958–2008. Berlin: Springer; 2010. pp. 343–385.
3. Schrijver A. Polyhedral combinatorics. Volumes 1 and 2, Handbook of combinatorics. Amsterdam: Elsevier; 1995. pp. 1649–1704.
4. Burkard RE. Convexity and discrete optimization. Volume A and B, Handbook of convex geometry. Amsterdam: North-Holland Publishing Co.; 1993. pp. 675–698.
5. Nemhauser G, Wolsey L. Integer and combinatorial optimization, Wiley-Interscience series in discrete mathematics and optimization. New York: John Wiley & Sons, Inc.; 1999. (Reprint of the 1988 original.)
6. Grötschel M, Lovász L, Schrijver A. Geometric algorithms and combinatorial optimization. Volume 2, Algorithms and combinatorics. 2nd ed. Berlin: Springer; 1993.
7. Bertsimas D, Weismantel R. Optimization over integers. Belmont Massachusetts (MA): Dynamic Ideas; 2005.
8. Cook WJ, Cunningham WH, Pulleyblank WR, *et al.* Combinatorial optimization, Wiley-Interscience series in discrete mathematics and optimization. New York: John Wiley & Sons, Inc.; 1998.
9. Wolsey LA. Integer programming, Wiley-Interscience series in discrete mathematics and optimization. New York: John Wiley & Sons, Inc.; 1998.
10. Korte B, Vygen J. Combinatorial optimization: theory and algorithms. Volume 21, Algorithms and combinatorics. 4th ed. Berlin: Springer; 2008.
11. Barvinok A. A course in convexity. Volume 54, Graduate studies in mathematics. Providence (RI): American Mathematical Society; 2002.
12. Ziegler GM. Lectures on polytopes. Volume 152, Graduate texts in mathematics. New York: Springer; 1995.
13. Grünbaum B. Convex polytopes. In: Kaibel V, Klee V, Ziegler GM, editors. Volume 221, Graduate texts in mathematics. 2nd ed. New York: Springer; 2003.
14. Gritzmann P, Klee V. Mathematical programming and convex geometry. Volume A,B, Handbook of convex geometry. Amsterdam: North-Holland Publishing Co.; 1993. pp. 627–674.
15. Bayer MM, Lee CW. Combinatorial aspects of convex polytopes. In: Volume A,B, Handbook of convex geometry. Amsterdam: North-Holland Publishing Co.; 1993. pp. 485–534.
16. Conforti M, Cornuéjols G, Zambelli G. Extended formulations in combinatorial optimization. Technical report 4OR 8. Padova: Università di Padova; 2010;1–48.
17. Ruszczyński A. Nonlinear optimization. Princeton (NJ): Princeton University Press; 2006.
18. Conforti M, Di Summa M, Zambelli G. Minimally infeasible set-partitioning problems with balanced constraints. Math Oper Res 2007;32(3):497–507.
19. Edmonds J. Systems of distinct representatives and linear algebra. J Res Natl Bur Stand Sect B 1967;71B:241–245.
20. Carathéodory C. Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen. Gesammelte mathematische Schriften Band 3. München: C. H. Beck'sche Verlagsbuchhandlung; 1955. pp. 78–110. Herausgegeben im Auftrag und mit Unterstützung der Bayerischen Akademie der Wissenschaften.
21. Minkowski H. Geometry of numbers. (Geometrie der Zahlen.). Volume 40, Bibliotheca mathematica teubneriana. New York: Johnson Reprint Corp.; 1968. vii. 256 pp.
22. Weyl H. Elementare theorie der konvexen polyeder. Commentarii Math Helvetici 1935;7:290–306.
23. Kaibel V. Another proof of the fact that polyhedral cones are finitely generated. 2009. Available at <http://arxiv.org/abs/0912.2927>.
24. Kaibel V. Two theorems on projections of polyhedra. 2009. Available at <http://www.math.uni-magdeburg.de/~kaibel/Downloads/ProjectPoly.pdf>.
25. Motzkin TS. Beiträge zur Theorie der linearen Ungleichungen [PhD thesis, Dissertation 73 S]. Basel; 1936.
26. Seidel R. Convex hull computations. In: Goodman JE, O'Rourke J, editors. Handbook of discrete and computational geometry. Boca Raton (FL): CRC Press LLC; 2004. Chapter 24. pp. 495–512.
27. Gawrilow E, Joswig M. Polymake: a framework for analyzing convex polytopes. In:

- Kalai G, Ziegler GM, editors. *Polytopes—combinatorics and computation*. Basel, Boston, Berlin: Birkhäuser; pp. 2000. 43–74.
28. Hačijan LG. A polynomial algorithm in linear programming. *Dokl Akad Nauk SSSR* 1979;244(5):1093–1096.
 29. Meyer RR. On the existence of optimal solutions to integer and mixed-integer programming problems. *Math Program* 1974;7: 223–235.
 30. Giles FR, Pulleyblank WR. Total dual integrality and integer polyhedra. *Linear Algebra Appl* 1979;25:191–196.
 31. Gordan P. Ueber die Auflösung linearer Gleichungen mit reellen Coefficienten. *Math Ann* 1873;6(1):23–28.
 32. Edmonds J, Giles R. A min-max relation for submodular functions on graphs. Volume 1, *Studies in integer programming* (Proceedings Workshop, Bonn, 1975), *Annals of Discrete Mathematics*. Amsterdam: North-Holland Publishing Co.; 1977. pp. 185–204.
 33. Edmonds J, Giles R. Total dual integrality of linear inequality systems. *Progress in combinatorial optimization* (Waterloo, Ontario, 1982). Toronto: Academic Press; 1984. pp. 117–129.
 34. Schrijver A. On total dual integrality. *Linear Algebra Appl* 1981;38:27–32.
 35. Ding G, Feng L, Zang W. The complexity of recognizing linear systems with certain integrality properties. *Math Program* 2008; 114(2):321–334.
 36. Hoffman AJ, Kruskal JB. Integral boundary points of convex polyhedra. Volume 38, *Linear inequalities and related systems*, *Annals of Mathematics Studies*. Princeton (NJ): Princeton University Press; 1956. pp. 223–246.
 37. Ghouila-Houri A. Caractérisation des matrices totalement unimodulaires. *C R Acad Sci Paris* 1962;254:1192–1194.
 38. Tutte WT. Lectures on matroids. *J Res Natl Bur Stand Sect B* 1965;69B:1–47.
 39. Seymour PD. Decomposition of regular matroids. *J Combin Theory Ser B* 1980;28(3): 305–359.
 40. Cunningham WH, Edmonds J. A combinatorial decomposition theory. *Can J Math* 1980;32(3):734–765.
 41. Truemper K. A decomposition theory for matroids. V. testing of matrix total unimodularity. *J Combin Theory Ser B* 1990;49(2): 241–281.

BASIS REDUCTION METHODS

GÁBOR PATAKI

Department of Statistics and
Operations Research, UNC
Chapel Hill, Chapel Hill North
Carolina

MUSTAFA TURAL

Institute for Mathematics and Its
Applications, University of
Minnesota, Minneapolis,
Minnesota

INTRODUCTION

Consider the integer programming (IP) feasibility problem in the form

$$\text{Find } x \in \mathbb{Z}^n : x \in P, \quad (\text{IP})$$

where P is a polyhedron described by inequalities. The work of Lenstra in 1979 [1] answered one of the most challenging questions in the theory of integer programming by presenting an algorithm to solve (IP), whose running time is polynomial, when n , the dimension is fixed. This paper pioneered the use of lattice basis reduction in integer programming, and initiated an interest in polynomial results in integer programming under the “fixed dimension” assumption. Somewhat later Kannan developed an improved variant [2,3], which—to date—has the best theoretical complexity for integer programming feasibility.

The goal of this survey is to review lattice-based methods to solve (IP), focusing on Lenstra’s and Kannan’s algorithms, which are by now considered “classical,” and the more recent reformulation methods of Aardal *et al.* [4], and Krishnamoorthy and Pataki [5].

Example 1. As motivation, let us consider a “hard” IP feasibility problem

$$\begin{aligned} 460 &\leq 51x_1 + 49x_2 \leq 489 \\ 0 &\leq x_1, x_2 \leq 10 \\ x_1, x_2 &\in \mathbb{Z}, \end{aligned} \quad (1)$$

shown in Fig. 1. One can see by inspection that it is infeasible.

Let us denote by P the underlying polyhedron. The hyperplanes

$$\{x \mid x_1 = k\} \ (k = 0, \dots, 9)$$

all intersect P , so branch-and-bound trying to prove infeasibility generates 10 subproblems, when branching on x_1 . Branching on x_2 also leads to 10 subproblems.

We will return to this example later, in particular, in the section titled “Reformulation Methods” we show that the rangespace reformulation of Krishnamoorthy and Pataki [5] creates an equivalent feasibility problem, in which the set $\{y \mid y_2 \in \mathbb{Z}\}$ has empty intersection with the underlying polyhedron, hence branching on y_2 solves the problem at the root node.

A lattice L is the set of integral combinations of linearly independent vectors defined as

$$L = \mathbb{L}(B) = \{Bx : x \in \mathbb{Z}^r\} \subseteq \mathbb{R}^n. \quad (2)$$

The matrix B has r independent columns, which are said to form a *basis* of $\mathbb{L}(B)$. We call r the *dimension* of L , and when $r = n$ we say that L is *full dimensional*. Writing $b_1, \dots, b_r \in \mathbb{R}^n$ for the columns of B , we also denote L by $\mathbb{L}(b_1, \dots, b_r)$.

The common theme in lattice-based methods is transforming (IP) into a problem of the form

$$\text{Find } y \in \mathbb{Z}^r : By \in Q, \quad (\text{IP}_{B,Q})$$

where the matrix B and the polyhedron Q are suitably chosen, so the problem of finding a point in $P \cap \mathbb{Z}^n$ is translated into finding one

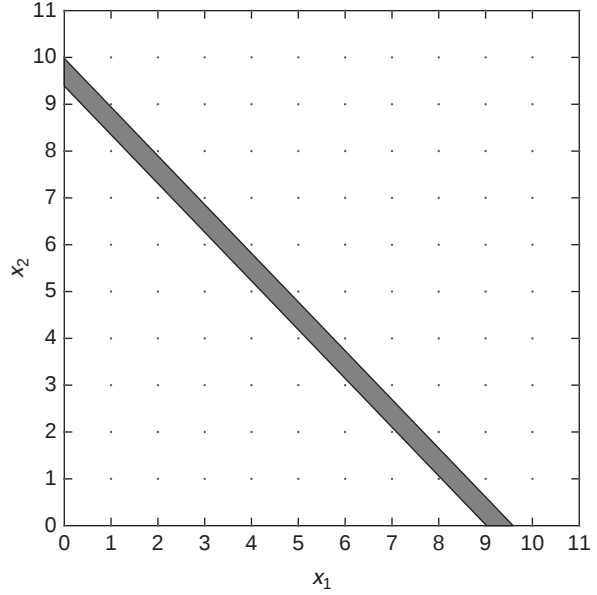


Figure 1. The instance of Example 1.

in $Q \cap \mathbb{L}(B)$. We emphasize that the choice of B and Q is specific to each method, but in all of them, the polyhedron Q has favorable geometry, and B has short and near orthogonal columns, which form a *reduced* basis of $\mathbb{L}(B)$ in a sense to be made precise later. These two facts will allow us to prove good complexity bounds on solving $(IP_{B,Q})$.

The classical (Lenstra’s and Kannan’s) algorithms are recursive. They branch on the last few variables in $(IP_{B,Q})$, that is, enumerate all possible integer values that these can attain, then construct new B matrices and Q polyhedra for each of the resulting subproblems, and so on. However, Mehrotra and Li [6] recently presented a restructuring of the computations in Lenstra’s algorithm, so that branching on the last variable in $(IP_{B,Q})$ is translated back to branching on a suitable hyperplane in the original problem. In contrast, the reformulation methods run a full branch-and-bound algorithm on $(IP_{B,Q})$, and are implemented by simply feeding the formulation to a commercial IP solver.

The *nullspace reformulation method*, which is applicable to equality constrained integer programs was proposed by Aardal *et al.* [4]. Later, Krishnamoorthy and Pataki [5] introduced the *rangespace reformulation method* for general integer programs. One

can analyze these methods on a family of knapsack feasibility problems, with the coefficient vector of the knapsack constraint decomposing as $a = \lambda p + r$, where p and r are integral vectors, and λ is a large integer. There are a variety of instances that fit into this framework with three interesting properties: (i) they are difficult for branch-and-bound that branches on the x_i variables; (ii) their infeasibility is proven by branching on px ; and (iii) when λ is sufficiently large, branching on the last variable in the reformulations has the same effect as branching on px in the original problem.

There is a compelling history of these instances, dating back to Jeroslow’s famous problem from 1974 [7]. An analysis of nullspace reformulation assuming such decomposable structure was given by Aardal and Lenstra [8,9]. Krishnamoorthy and Pataki [5] showed how a variety of knapsack problems starting with Jeroslow’s problem satisfy properties (i) and (ii), presented a “recipe” to generate such instances, and proved Property (iii) for both reformulations.

A recent result of Pataki *et al.* [10] shows that when branch-and-bound is applied to the reformulation of bounded integer programs, the majority of instances get solved without generating any subproblems, that

is, at the root node. This result may seem counterintuitive; however, it fits in nicely with the previous work on “low density” subset sum problems, which have been widely used in cryptography.

The rest of the survey is divided into five sections. In the rest of the introduction we describe key definitions that will be used throughout the paper. In the section titled “Reduced Bases” we describe reduced bases of lattices. In the section titled “The Geometry of Lattices and Convex Sets” we present a lemma due to Lenstra on the geometry of lattices, and convex sets, which will play a role in the analysis of both the classical, and the reformulation methods. In the section titled “Lenstra’s and Kannan’s Algorithms” we review Babai’s improved version [11] of Lenstra’s algorithm, and Kannan’s algorithm. Kannan’s algorithm is less frequently covered in surveys than Lenstra’s method, perhaps because it is more technical: we think that we succeeded in giving a simplified treatment. In the section titled “Reformulation Methods” we review the reformulation methods, with their complexity analyses. In the section titled “Further Reading and Computational Testing” we point to further reading, and review computational results obtained by lattice-based methods.

The emphasis of the survey is providing simple proofs of complexity results, illustrative examples, and a unifying view of the classical and the reformulation methods. For instance, we will continue Example 1, and show how Lenstra’s algorithm, and the rangespace reformulation handle this “difficult” instance. We include exercises, and we think that the survey will be suitable to teach a two- to three-class long segment on lattice-based methods in a course on Integer Programming.

A reader may be interested either only in Lenstra’s and Kannan’s algorithms, or only in the reformulation methods; so for convenience, sections titled “Lenstra’s and Kannan’s Algorithms” and “Reformulation Methods” can be read independently of each other.

Exercise 1. Generalize Example 1 by showing that for positive integers λ and μ with $\lambda \geq 2\mu + 1$, the integer programming problem

$$\begin{aligned} \lambda\mu - \lambda + \mu &\leq (\lambda + 1)x_1 + (\lambda - 1)x_2 \leq \lambda\mu - \mu - 1 \\ 0 &\leq x_1, x_2 \leq \mu \\ x_1, x_2 &\in \mathbb{Z} \end{aligned} \quad (3)$$

is infeasible, and branching either on x_1 or x_2 will generate at least μ subproblems.

Important Definitions

Branch-and-bound, which we abbreviate as B&B, is a classical solution method for (IP); it was first proposed by Land and Doig [12]. It starts with P as the sole subproblem (node). In a general step, one chooses a subproblem P' , a variable x_i , and creates nodes $P' \cap \{x|x_i = \gamma\}$, where γ ranges over all possible integer values of x_i . We repeat this until all subproblems are shown to be empty, or we find an integral point in one of them.

Sometimes we will *branch on hyperplanes* to solve (IP): given a nonzero integral vector p , we let

$$\begin{aligned} k_{\max} &= \max \{px \mid x \in P\}, \\ k_{\min} &= \min \{px \mid x \in P\}, \end{aligned} \quad (4)$$

and create the subproblems

$$P \cap \{x \mid px = \lfloor k_{\max} \rfloor\}, \dots, P \cap \{x \mid px = \lceil k_{\min} \rceil\}.$$

The number of subproblems is the *integer width* of P along p , that is,

$$\text{iwidth}(p, P) = \lfloor k_{\max} \rfloor - \lceil k_{\min} \rceil + 1.$$

In particular, if $\text{iwidth}(p, P) = 0$, then (IP) is infeasible, and we say that *its infeasibility is proven by branching on px* . Analogously, the *width* of P along p is defined as

$$\text{width}(p, P) = k_{\max} - k_{\min}.$$

To emphasize the difference between Land and Doig’s branch-and-bound algorithm, and branching on hyperplanes, we call the former method *ordinary branch-and-bound*.

It is well known [13] that if A is a rational matrix with n columns, then

$$\mathbb{N}(A) := \{x \in \mathbb{Z}^n \mid Ax = 0\} \quad (5)$$

is also a lattice, and we call this set the *null-lattice* of A .

We call a matrix U *unimodular*, if it is square, integral, and has determinant ± 1 .

For a convex set Q we write $r(Q)$ for the radius of the largest ball in the affine hull of Q that can be inscribed in Q , and $R(Q)$ for the radius of the smallest ball that contains Q :

$$r(Q) = \max \{s \mid \exists p \in Q \text{ s.t. } \text{Ball}(p, s) \cap \text{aff}(Q) \subseteq Q\}, \text{ and} \quad (6)$$

$$R(Q) = \min \{s \mid \exists p \in Q \text{ s.t. } \text{Ball}(p, s) \supseteq Q\}, \quad (7)$$

where $\text{Ball}(p, s)$ is the Euclidean ball with center p and radius s .

REDUCED BASES

A lattice $\mathbb{L}(B)$ with dimension $r \geq 2$ has infinitely many bases, and it is well known, that all of them are of the form BU , where U is a unimodular matrix. Multiplying B by a unimodular matrix is equivalent to performing a sequence of three elementary column operations on B : multiplying a column by -1 ; exchanging two columns; and adding an integer multiple of a column to another.

As a lattice may not have an orthonormal basis of unit vectors, we will be interested

in bases comprising “reasonably” short and “near orthogonal” ones. These bases will be called *reduced*.

Example 2. Consider the lattice generated by the vectors

$$b_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix}. \quad (8)$$

Multiplying the matrix $[b_1, b_2]$ with

$$U = - \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

gives another basis

$$b'_1 = -2b_1 - b_2, \quad b'_2 = -b_1 - b_2.$$

Figure 2 shows the lattice points, and the two bases, with the vectors in the first clearly shorter, and closer to being orthogonal.

We describe lattice bases that are reduced in the sense of Lenstra, Lenstra, and Lovász (LLL), and Korkin and Zolotarev (KZ). We will say that $b_1, \dots, b_r$ is an LLL (KZ) reduced basis, if it is such a basis of $\mathbb{L}(b_1, \dots, b_r)$.

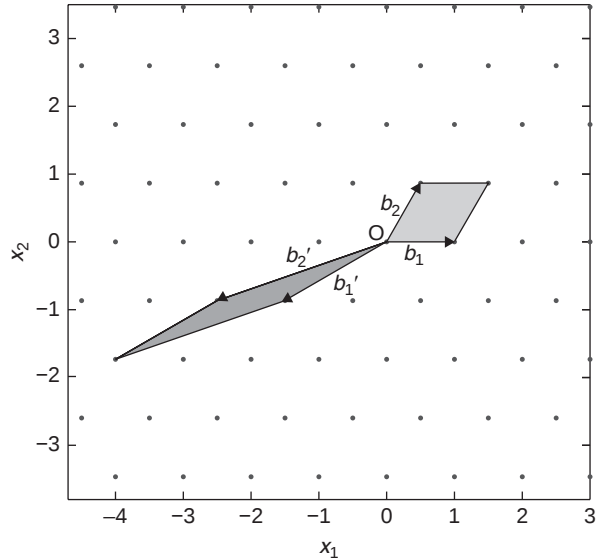


Figure 2. A lattice, with a reduced, and a nonreduced basis.

Given $b_1, \dots, b_r \in \mathbb{R}^n$, the vectors $b_1^*, \dots, b_r^*$ form their *Gram–Schmidt orthogonalization*, if

$$b_1^* = b_1 \quad (9)$$

$$b_i^* = b_i - \sum_{j=1}^{i-1} \mu_{ij} b_j^* \quad (i = 2, \dots, r), \quad (10)$$

with

$$\mu_{ij} = \langle b_i, b_j^* \rangle / \|b_j^*\|^2 \quad (i = 2, \dots, r; j = 1, \dots, i-1). \quad (11)$$

In other words, b_i^* is the projection of b_i onto the orthogonal complement of the linear span of $b_1, \dots, b_{i-1}$. Given $b_1, \dots, b_r$, their Gram–Schmidt vectors are by definition orthogonal, and will be used as reference vectors to express *near* orthogonality of the b_i .

Convention. Given $b_1, \dots, b_r$, we will write $b_1^*, \dots, b_r^*$ for their Gram–Schmidt vectors, and μ_{ij} for the corresponding coefficients without explicitly saying.

Perhaps the best-known reducedness concept in lattices is LLL reducedness. These bases were introduced by Lenstra and colleagues in their seminal 1982 article [14]. They are polynomial time computable when the lattice is generated by rational vectors, and they found uses in numerous contexts other than integer programming, for instance, in number theory and cryptography. For efficient implementations we refer the reader to the LiDia library [15] at the University of Darmstadt, and the NTL library of Victor Shoup [16].

We use Schrijver’s definition [13], which is less restrictive than the classical one [14].

Definition 1. We say that $b_1, \dots, b_r$ is LLL reduced, if

1. $|\mu_{ij}| \leq 1/2$ ($i = 2, \dots, r; j = 1, \dots, i-1$).
2. $\|b_i^*\| \leq \sqrt{2} \|b_{i+1}^*\|$ ($i = 1, \dots, r-1$).

We remark that an LLL reduced basis remains computable in polynomial time, if we replace the factor of $\sqrt{2}$ with $\sqrt{4/3 + \epsilon}$ for an arbitrarily small positive ϵ : the running time is polynomial in the size of the basis and $1/\epsilon$.

Both conditions in Definition 1 naturally correspond to “near orthogonality,” since in an orthogonal basis all μ_{ij} would be zero, and condition (2) could be achieved by a simple rearranging, with the factor $\sqrt{2}$ replaced by 1. On the other hand, it is easy to construct examples of bases having near parallel vectors that violate conditions 1 and 2 by an arbitrary amount.

Exercise 2. Verify that the lattice in Example 2 does not have an orthogonal basis. Show that b_1 and b_2 form an LLL reduced basis, but b'_1 and b'_2 do not.

From the two basic properties of LLL reduced bases, one can derive several other inequalities that show the shortness (which is less straightforward from conditions (1) and (2) of Definition 1), and near orthogonality of the basis vectors. For details we refer to Lenstra *et al.* [14], and we recall an inequality that will be used in the section titled “Reformulation Methods”.

Proposition 1. Let $b_1, \dots, b_r$ be an LLL reduced basis of the lattice L , and $x_1, \dots, x_t$ be arbitrary linearly independent vectors in L . Then

$$\max \{ \|b_1\|, \dots, \|b_t\| \} \leq 2^{(r-1)/2} \max \{ \|x_1\|, \dots, \|x_t\| \}. \quad (12)$$

We remark, that the performance of LLL reduction in practice is much better than predicted by the estimate (12), especially when t is small.

Korkin–Zolotarev (KZ) reduced bases were introduced by Kannan [2,3] as a key ingredient in his integer programming algorithm, and he also showed that they are computable in polynomial time, when the dimension is fixed. It is currently not known, whether they are computable in polynomial time for varying r . Surprisingly, it turned out that KZ reduced bases have been described previously already in the 19th century [17]; however, no algorithms were known to compute them until Kannan’s result. The LiDia [15] and NTL libraries [16]

also contain efficient subroutines to compute KZ reduced bases.

Definition 2. We say that $b_1, \dots, b_r$ is a KZ reduced basis, if

1. b_1 is a shortest vector of $\mathbb{L}(b_1, \dots, b_r)$;
2. $|\mu_{ij}| \leq 1/2$ ($i = 2, \dots, r; j = 1, \dots, i-1$); and
3. letting $b'_2, \dots, b'_r$ be the projection of $b_2, \dots, b_r$ on the orthogonal complement of the line spanned by b_1 , they form a KZ reduced basis.

While KZ reduced bases are harder to compute than LLL reduced ones, they have stronger properties. We recall two inequalities that we will use later.

Proposition 2. Let $b_1, \dots, b_r$ be a KZ reduced basis of the lattice L , and $x_1, \dots, x_t$ arbitrary linearly independent vectors in L . Then

$$\begin{aligned} \max \{ \|b_1\|, \dots, \|b_t\| \} \\ \leq \frac{\sqrt{t+3}}{2} \max \{ \|x_1\|, \dots, \|x_t\| \}. \end{aligned} \quad (13)$$

Furthermore, for all $j \leq r$

$$\|b_j^*\| \leq \sqrt{r-j+1} \left(\|b_j^*\| \dots \|b_r^*\| \right)^{1/(r-j+1)} \quad (14)$$

holds.

The inequality (13) is due to Lagarias *et al.* [18]. To put into context, we can compare the sublinear factor in it with the exponential factor in (12).

Inequality (14) follows from the definition of KZ reducedness, and for the proof we refer to Kannan [3]. Let us note that multiplying the inequalities $\|b_j^*\| \leq \sqrt{2}^i \|b_{j+i}^*\|$ for $i = 0, \dots, r-j$ in an LLL reduced basis would give a similar inequality with a factor of $2^{(r-j)/4}$ only (or with a factor $(4/3 + \epsilon)^{(r-j)/4}$, if we use a strengthened definition of LLL reducedness); so a KZ reduced basis improves inequality (2) in Definition 1 in an

aggregate sense. Exercise 3 shows that the improvement also holds for every i .

Exercise 3. Show that if $b_1, \dots, b_r$ is a KZ reduced basis, then

$$\|b_i^*\| \leq \sqrt{4/3} \|b_{i+1}^*\| \quad (i = 1, \dots, r-1)$$

holds; so a KZ reduced basis is also LLL reduced.

THE GEOMETRY OF LATTICES AND CONVEX SETS

Proposition 3 concerns the solvability of $(\text{IP}_{B,Q})$, and we will use it in the analysis of both the classical, and the reformulation methods. It asserts that if Q is “large” with respect to the norms of the columns of B , then $(\text{IP}_{B,Q})$ is feasible, and if it is “small,” then one can reduce $(\text{IP}_{B,Q})$ to a “small” number of $r-1$ dimensional subproblems.

Proposition 3. Denote the columns of B by $b_1, \dots, b_r$. Then

1. if Q is contained in the linear span of $\{b_1, \dots, b_r\}$, and

$$r(Q) \geq \frac{1}{2} \sqrt{\|b_1^*\|^2 + \dots + \|b_r^*\|^2}, \quad (15)$$

then $(\text{IP}_{B,Q})$ is feasible.

2. If $(\text{IP}_{B,Q})$ is feasible, then y_r must belong to an interval of length at most

$$\left\lfloor \frac{2R(Q)}{\|b_r^*\|} \right\rfloor + 1; \quad (16)$$

hence $(\text{IP}_{B,Q})$ can be reduced to at most this many $r-1$ -dimensional IP feasibility problems.

For a proof of statement 1 of Proposition 3, we refer to Lenstra [1] (Lemma on p. 540). In fact, Lenstra only states the bound with b_i in place of the b_i^* , but his proof does imply the stronger statement. As motivation, consider a lattice in $\mathbb{R}^2$ spanned by orthogonal vectors, that is, $b_i^* = b_i$ for $i = 1, 2$. The point on the plane, which is farthest from the lattice points is at the intersection of diagonals of

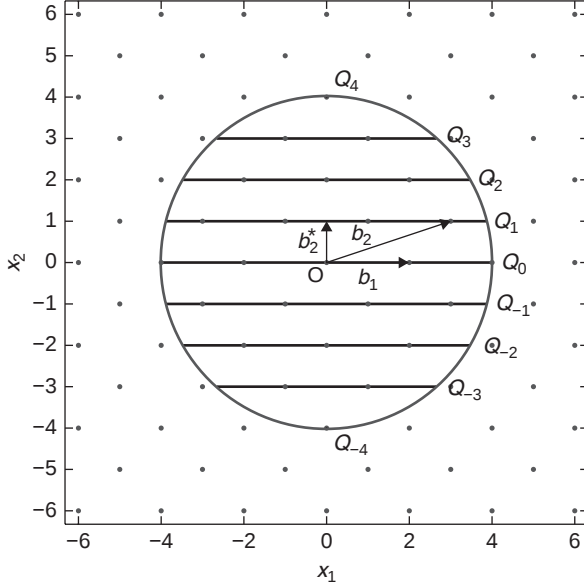


Figure 3. Illustration of statement 2 in Proposition 3.

a grid rectangle, with distance equal to the right-hand side of (15). Thus, a ball around any point in space with this radius will contain a lattice point.

Statement 2 of Proposition 3 is also essentially due to Lenstra, though it is not stated precisely in this form. A geometric proof can be found in Lenstra [1] (p. 541). For motivation, consider Fig. 3. The problem of finding a lattice point in Q is reduced to finding such a point in one of the Q_i , and the distance between the lines containing the Q_i is $\|b_2^*\|$, leading to the desired upper bound. Note that $\|b_2^*\|$ in this example is considerably smaller than $\|b_2\|$.

We give an algebraic proof of statement 2 of Proposition 3. Writing

$$B^{-1}Q = \{y \mid By \in Q\} \quad (17)$$

we show

$$\text{width}(e_r, B^{-1}Q) \leq \frac{2R(Q)}{\|b_r^*\|}. \quad (18)$$

Let $y_{r,1}$ and $y_{r,2}$ denote the maximum and the minimum of y_r over $B^{-1}Q$. Writing $\bar{B}$ for the matrix composed of the first $r-1$ columns of B , and b_r for the last column, it holds that there are $y_1, y_2 \in \mathbb{R}^{r-1}$ such that $\bar{B}y_1 + b_r y_{r,1}$

and $\bar{B}y_2 + b_r y_{r,2}$ are in Q . So,

$$\begin{aligned} 2R(Q) &\geq \|(\bar{B}y_1 + b_r y_{r,1}) - (\bar{B}y_2 + b_r y_{r,2})\| \\ &= \|\bar{B}(y_1 - y_2) + b_r(y_{r,1} - y_{r,2})\| \\ &\geq \|b_r^*\| |y_{r,1} - y_{r,2}| \\ &= \|b_r^*\| \text{width}(e_r, B^{-1}Q) \end{aligned}$$

holds, as required.

Remark 1. If y_r is fixed to an integer k in $(\text{IP}_{B,Q})$, then the corresponding lower-dimensional subproblem is

$$\text{Find } \bar{y} \in \mathbb{Z}^{r-1} : \bar{B}\bar{y} + kb_r \in Q \cap \{y \mid y_r = k\}. \quad (19)$$

Corollary 1. When in $(\text{IP}_{B,Q})$ we branch on $y_r, y_{r-1}, \dots, y_j$ in this order, the number of resulting subproblems on the level of y_j is at most

$$\prod_{i=j}^r \left(\left\lfloor \frac{2R(Q)}{\|b_i^*\|} \right\rfloor + 1 \right). \quad (20)$$

Proof. The formula (19) shows the form of the resulting subproblems, after we

branched on y_r . Using part (2) in Proposition 3 implies that branching on y_{r-1} will create at most

$$\left\lceil \frac{2R(Q)}{\|b_{r-1}^*\|} \right\rceil + 1 \quad (21)$$

subproblems from each of these, and so on, and this completes the proof.

LENSTRA'S AND KANNAN'S ALGORITHMS

Lenstra's Algorithm

Lenstra's paper [1] first disposes with two technicalities. Assuming that P is described as $P = \{x \mid Ax \leq b\}$, with A and b integral, and their components bounded by a in absolute value, Lenstra shows that when (IP) is feasible, it has a solution $\bar{x}$ with $\max_i |\bar{x}_i| \leq (n+1)n^{n/2}a^n$. So we can assume that P is bounded, and Lenstra shows that we can also assume that it is full dimensional.

Next we “round” P by computing in polynomial time an invertible transformation τ , so that

$$R(\tau P)/r(\tau P) \leq c_n := n(n+1) \quad (22)$$

will hold, that is, τP appears relatively “spherical.”

Example 1 (continued). Using Lenstra's algorithm we computed τP , where P is the polyhedron in Example 1. With τ having a transformation matrix

$$\begin{pmatrix} 0.3326 & 0.3101 \\ 0 & 0.0164 \end{pmatrix}, \quad (23)$$

the rounded polyhedron is

$$\begin{aligned} 460 &\leq 153.3333x_1 + 88.5270x_2 \leq 489 \\ 0 &\leq 3.0065x_1 - 56.8034x_2 \leq 10 \\ 0 &\leq 60.9285x_2 \leq 10, \end{aligned} \quad (24)$$

shown in Fig. 4 with the points of the lattice $\tau\mathbb{Z}^2$.

Now (IP) is transformed into trying to find $x \in \tau\mathbb{Z}^n \cap \tau P$. After choosing an LLL reduced basis B for the lattice $\tau\mathbb{Z}^n$, the task becomes

$$\text{Find } y \in \mathbb{Z}^n : By \in \tau P. \quad (\text{IP}_{B,\tau P})$$

Let us write $R := R(\tau P)$, $r := r(\tau P)$, and let $b_1, \dots, b_n$ denote the columns of B . For better intuition, we first describe the algorithm assuming $R = r$ (i.e., that τP is a ball), and that the b_i^* all have the same norm.

By Proposition 3 if

$$r \geq \sqrt{n}\|b_n^*\|/2, \quad (25)$$

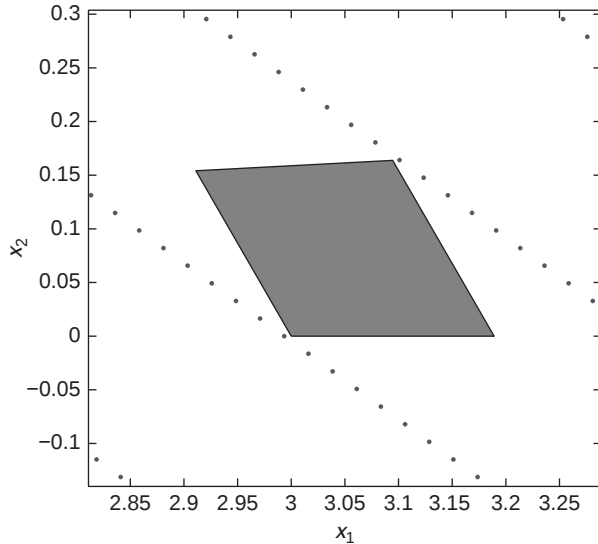


Figure 4. The polyhedron of Example 1 after Lenstra's rounding, and the points of τF^2 .

then $(\text{IP}_{B,\tau P})$ is feasible, and if (25) fails, $(\text{IP}_{B,\tau P})$ can be reduced to at most

$$\left\lfloor \frac{2r}{\|b_n^*\|} \right\rfloor + 1 \leq \sqrt{n} + 1$$

$n - 1$ -dimensional subproblems. On the subproblems we apply the algorithm recursively, by making the underlying polyhedra full dimensional, rounded, and finding a new LLL reduced basis for each one. If we denote by $f(n)$ the number of polyhedra that Lenstra's algorithm must examine, we obtain

$$f(n) \leq (\sqrt{n} + 1)f(n - 1); \quad (26)$$

thus, $f(n) = O(n^{n/2})$. The number of arithmetic operations performed on each polyhedron is polynomial, so the overall complexity is polynomial, when n is fixed.

The roundedness of τP , and the reducedness of B make sure that our unrealistic assumption is not far from the truth. Let j be the index such that $\|b_j^*\|$ is maximal, and consider the two cases:

Case 1.

$$r \geq \frac{1}{2}\sqrt{n}\|b_j^*\|.$$

Then the first statement in Proposition 3 shows that $(\text{IP}_{B,\tau P})$ is feasible.

Case 2.

$$r < \frac{1}{2}\sqrt{n}\|b_j^*\|.$$

Then

$$R < \frac{1}{2}c_n\sqrt{n}\|b_j^*\| < \frac{1}{2}c_n\sqrt{n}\|b_n^*\|2^{(n-j)/2},$$

with the first inequality from (22), and the second from the LLL reducedness of B . Then Proposition 3 implies that branching on y_n in $(\text{IP}_{B,\tau P})$ creates at most

$$\left\lfloor \frac{2R}{\|b_n^*\|} \right\rfloor + 1 = O(2^n) \quad (27)$$

subproblems.

Denoting by $f_L(n)$ the number of polyhedra that must be examined by Lenstra's algorithm, we have

$$f_L(n) = O(2^n)f_L(n - 1) = O(2^{n^2}), \quad (28)$$

and the previous argument shows polynomiality of the method for fixed n .

Exercise 4. Verify that if P is the polyhedron in Example 1, and τ is given in (23), then τP indeed has the description given in (24), and $R(\tau P)/r(\tau P) \leq 3.82$.

In contrast, show that the long and thin nature of P is also shown by the poor ratio of $R(P)$ and $r(P)$, namely $R(P)/r(P) \geq 33$.

Kannan's Algorithm

In Kannan's algorithm the setup of making P bounded, full dimensional, and rounded is the same as in Lenstra's method; then a reduced basis for the lattice $\tau\mathbb{Z}^n$ is found. Hence (IP) is again transformed into $(\text{IP}_{B,\tau P})$, and the handling of Case 1 is identical. However, Kannan uses a Korkin–Zolotarev reduced basis for $\mathbb{L}(B)$ in $(\text{IP}_{B,\tau P})$, and again letting j to be the index for which the norm of the corresponding Gram–Schmidt vector is maximal, we enumerate all possible values for $y_j, \dots, y_n$. So, if $j = n$, Kannan's algorithm behaves like Lenstra's, and if $j = 1$, it does complete enumeration.

The complexity analysis of Kannan's algorithm is more technical than Lenstra's, so we only outline a proof showing that it needs to look at only $O(n^{3n})$ polyhedra, as opposed to Lenstra's $O(2^{n^2})$. We first describe Case 2 in detail.

Case 2.

$$r < \frac{1}{2}\sqrt{n}\|b_j^*\|,$$

that is

$$R < \frac{1}{2}c_n\sqrt{n}\|b_j^*\|.$$

Corollary 1 implies that branching on $y_n, \dots, y_j$ creates at most

$$\prod_{i=j}^n \left(\left\lfloor \frac{2R}{\|b_i^*\|} \right\rfloor + 1 \right).$$

subproblems. Using the fact that b_j^* has the largest norm among the Gram–Schmidt vectors, simple manipulation shows that this upper bound is at most

$$(c_n \sqrt{n} + 1)^{n-j+1} \frac{\|b_j^*\|^{n-j+1}}{\|b_j^*\| \cdots \|b_n^*\|}.$$

We now use the KZ reducedness of the b_i , and plug the estimate (14) into the latter expression. Hence the number of subproblems is at most

$$(c_n \sqrt{n} + 1)^{n-j+1} \left(\sqrt{n-j+1} \right)^{n-j+1},$$

which is $O(n^{3(n-j+1)})$. So, denoting by $f_K(n)$ the number of polyhedra that must be examined by Kannan’s algorithm,

$$f_K(n) = O(n^{3(n-j+1)}) f_K(j-1) = O(n^{3n})$$

holds.

REFORMULATION METHODS

In this section we describe the rangespace and nullspace reformulation methods. It will be convenient to assume that our feasibility problem is given as

$$\begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix} \leq \begin{pmatrix} A \\ I \end{pmatrix} x \leq \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (29)$$

$$x \in \mathbb{Z}^n,$$

where A is an integral m by n matrix. The rangespace reformulation of (29) is

$$\begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix} \leq \begin{pmatrix} A \\ I \end{pmatrix} Uy \leq \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (30)$$

$$y \in \mathbb{Z}^n,$$

where U is a unimodular matrix computed to make the columns of the constraint matrix a reduced basis of the generated lattice. If y is feasible for (30), then Uy is feasible for (29), and if x is feasible for (29), then $U^{-1}x$ is feasible for (30), since U^{-1} is integral. This shows the equivalence of the two problems.

The nullspace reformulation is applicable when $w_1 = \ell_1$. Assuming that the rows of A are linearly independent, it is

$$\ell_2 - x_0 \leq By \leq w_2 - x_0 \quad (31)$$

$$y \in \mathbb{Z}^{n-m},$$

where $x_0 \in \mathbb{Z}^n$ satisfies $Ax_0 = \ell_1$, and B is a reduced basis of $\mathbb{N}(A)$ (recall the definition of the null-lattice from (5)). Since any integral solution of the system $Ax = \ell_1$ can be written as the sum of x_0 , and an element of $\mathbb{N}(A)$, the two formulations are equivalent [13].

For brevity, if the columns of the constraint matrix in (30) form an LLL reduced basis of the generated lattice, we will call (30) the *LLL rangespace reformulation* of (29), and we will similarly talk about the LLL nullspace reformulation, KZ rangespace reformulation, and so on.

Remark 2. In its simplest form, the rangespace reformulation method transforms the feasibility problem

$$Ax \# b, x \in \mathbb{Z}^n$$

into

$$AUy \# b, y \in \mathbb{Z}^n,$$

where U is a unimodular matrix, AU is a reduced basis of the lattice $\mathbb{L}(A)$, and $\#$ denotes an arbitrary mixture of equality and inequality constraints. For the analyses that we present, however, it is convenient to work with a problem in the form of (29).

Also, one can solve an integer programming optimization problem by reducing it to a sequence of feasibility problems, on which a suitable reformulation can be then applied. The simplest direct approach is transforming

$$\max \{cx \mid Ax \# b, x \in \mathbb{Z}^n\}$$

into

$$\max \{cUy \mid AUy \# b, y \in \mathbb{Z}^n\},$$

where again $\#$ is a mixture of equality and inequality constraints, U is a unimodular matrix, and

$$\begin{pmatrix} c \\ A \end{pmatrix} U$$

is a reduced basis of the generated lattice.

Let us note how the reformulated problems fit in the framework of $(IP_{B,Q})$ with the polyhedron Q being simply a box, and B constructed directly from the constraint matrix of (29). This is in contrast with Lenstra's and Kannan's algorithms, where Q is a transformed version of P , and B is a reduced basis of the transformation matrix.

Decomposable Knapsack Problems and Their Reformulations

In this section we are interested in feasibility problems of the type

$$\begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix} \leq \begin{pmatrix} a \\ I \end{pmatrix} x \leq \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (32)$$

$x \in \mathbb{Z}^n,$

with a decomposing as

$$a = \lambda p + r, \quad (33)$$

where a, p , and r are integral row vectors, the components of p relatively prime, and λ is an integer. We will call these *decomposable knapsack problems (DKPs)*, and note that this is not a formal definition, since any knapsack problem with a constraint vector a can be written in this form with $p = a, r = 0$, and $\lambda = 1$. However, in all interesting cases λ will be at least 2, and in most cases relatively large with respect to $\|p\|$ and $\|r\|$.

There are a wide variety of instances, which fit into this framework, with three interesting properties, that we repeat from the introduction:

- (i) they are difficult for ordinary B&B;
- (ii) their infeasibility is proven by branching on px ; and
- (iii) when λ is sufficiently large, branching on the last variable in the reformulations has the same effect as branching on px in the original problem.

Example 1 (continued). This instance fits the mold of (32) with the constraint vector, and its decomposition given by

$$\begin{aligned} a &= (51, 49), & p &= (1, 1), \\ r &= (1, -1), & \lambda &= 50. \end{aligned} \quad (34)$$

The definition of the rest of the data (of ℓ_1, ℓ_2, w_1 , and w_2) is obvious.

With P denoting the underlying polyhedron, we have

$$\begin{aligned} \max \{px \mid x \in P\} &= 489/49 = 9.9796, \\ \min \{px \mid x \in P\} &= 460/51 = 9.0196; \end{aligned} \quad (35)$$

so, branching on px proves the infeasibility of this instance, and it is straightforward to check that the same holds for the generalization of Exercise 1.

The next example is a simplification of Jeroslow's classic problem [7], which established that ordinary B&B may take exponential time even on a trivial problem.

Example 3. Let n be a positive, odd integer. The problem

$$\begin{aligned} 2 \sum_{i=1}^n x_i &= n \\ 0 &\leq x \leq e \\ x &\in \mathbb{Z}^n \end{aligned} \quad (36)$$

is infeasible, and the infeasibility is proven by branching on $\sum_{i=1}^n x_i$. At the same time, ordinary B&B needs to enumerate at least $2^{(n-1)/2}$ nodes to do the same. For a proof of the latter fact we refer to Jeroslow's [7] or Krishnamoorthy and Pataki's paper [5] (p. 247).

This instance also fits into the framework of (32) with

$$a = 2e, \quad p = e, \quad r = 0, \quad \lambda = 2, \quad (37)$$

and $\ell_1 = w_1 = n$.

Another family of instances with similar behavior is related to the famous *Frobenius problem*. For a positive integral vector $a =$

$(a_1, \dots, a_n)$, the *Frobenius number* $\text{Frob}(a)$ is the largest integer for which we cannot make a change using the denominations $a_1, \dots, a_n$; in other words, the largest β for which the integer programming problem

$$\begin{aligned} ax &= \beta \\ x &\geq 0 \\ x &\in \mathbb{Z}^n, \end{aligned} \quad (38)$$

is infeasible. The *Frobenius problem* is computing $\text{Frob}(a)$: for a recent review we refer to a book by Ramirez Alfonsin [19].

The Frobenius problem gives rise to interesting, and difficult integer programming instances. Cornuéjols *et al.* [20] studied knapsack problems with a decomposable structure, and showed how to compute the Frobenius number by solving a sequence of integer programming problems using a test set approach.

Aardal and Lenstra [8,9] considered instances of the form (38), with a decomposing as in (33), and β a positive integer. Following the arguments in Krishnamoorthy and Pataki [5], we give a simplified description of the instances they constructed, and in addition show that they fit Property (ii).

Example 4. Consider a knapsack feasibility problem (38) with a decomposing as in (33), and β a positive integer. We assume that p is componentwise positive, it is not a multiple of r , and

$$q_1 := r_1/p_1 \leq \dots \leq q_n := r_n/p_n. \quad (39)$$

Proposition 4. *Define*

$$\begin{aligned} k &= \lfloor (\lambda + q_1 - 1)/(q_n - q_1) \rfloor - 1, \\ \beta &= \lceil (\lambda + q_n)k \rceil, \end{aligned} \quad (40)$$

and assume that λ is large enough so k and β are both positive. Then the infeasibility of (38) is proven by branching on px .

Proof. Let us consider the interval

$$I = ((\lambda + q_n)k, (\lambda + q_1)(k + 1)), \quad (41)$$

and note that I has length strictly larger than one by the choice of k . Next, let us denote by P the polyhedron of the LP relaxation of (38). Since $a_i = \lambda p_i + r_i$, the ordering in (39) implies

$$a_1/p_1 \geq \dots \geq a_n/p_n, \quad (42)$$

therefore

$$\begin{aligned} \max \{px \mid x \in P\} &= p_1\beta/a_1, \\ \min \{px \mid x \in P\} &= p_n\beta/a_n, \end{aligned} \quad (43)$$

hold. So $\text{iwidth}(p, P) = 0$ holds, when the above maximum and minimum are in the interval $(k, k + 1)$. A simple calculation shows that this happens exactly when $\beta \in I$, which is true because β is the ceiling of the lower end point of I , and $|I| > 1$.

Since (38) is infeasible with a large right-hand side, it is difficult for ordinary B&B, and it is also difficult when the right-hand side is chosen to be $\text{Frob}(a)$.

The following exercises review the recipe from Krishnamoorthy and Pataki [5], and its application to construct Avis's instance from Chvátal [21]:

Exercise 5. Let p and r be arbitrary integral vectors, ℓ_2 and w_2 bound vectors in (32), and k an integer with $0 < k \leq pw_2$. Prove that if we find λ, ℓ_1 , and w_1 to satisfy $\ell_1 \leq w_1$, and

$$\begin{aligned} \max \{rx \mid px \leq k, \ell_2 \leq x \leq w_2\} &+ k\lambda < \ell_1, \\ \text{and} \\ \min \{rx \mid px \geq k + 1, \ell_2 \leq x \leq w_2\} \\ &+ (k + 1)\lambda > w_1, \end{aligned} \quad (44)$$

then the infeasibility of (32) is proven by branching on px .

Show that the instances of Examples 1, 3, and 4 can be obtained this way, with $k = 9$, $k = (n - 1)/2$, and k defined in (40), respectively.

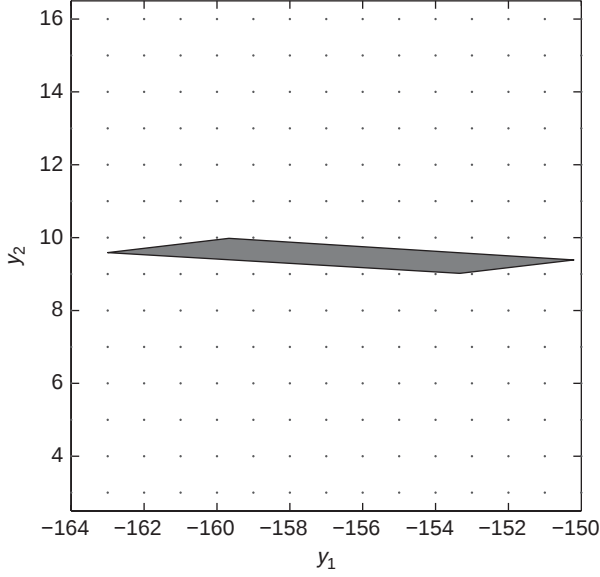


Figure 5. Example 1 after applying the LLL rangespace reformulation.

For a proof of why the above recipe creates hard integer programs, and how the difficulty depends on p, r, λ , and k , we refer to Krishnamoorthy and Pataki [5].

Exercise 6. Using the result of Exercise 5, show that the subset sum problem

$$\begin{aligned} ax &= \beta \\ x &\in \{0, 1\}^n, \end{aligned} \quad (45)$$

where n is odd,

$$\begin{aligned} a &= (n(n+1)+1, \dots, n(n+1)+n), \\ \beta &= \left\lfloor \sum_{i=1}^n a_i/2 \right\rfloor \end{aligned} \quad (46)$$

is infeasible. This instance was proposed by Avis [21], and it is known that ordinary B&B needs at least $2^{(n-1)/2}$ nodes to prove its infeasibility.

Example 1 (continued). To motivate Theorem 1, the main result presented in this section, we computed the LLL rangespace reformulation of Example 1. With

$$U = \begin{pmatrix} -1 & -16 \\ 1 & 17 \end{pmatrix},$$

it is

$$460 \leq 2y_1 + 17y_2 \leq 489$$

$$\begin{aligned} 0 &\leq -y_1 - 16y_2 \leq 10 \\ 0 &\leq y_1 + 17y_2 \leq 10 \\ y_1, y_2 &\in \mathbb{Z}, \end{aligned} \quad (47)$$

shown in Fig. 5. It is interesting to note that the underlying polyhedron is still long and thin, unlike the polyhedron in Fig. 4 produced by the τ transformation in Lenstra's and Kannan's methods, but now infeasibility is proven by branching on y_2 .

Theorem 1. Denote by P the polyhedron of (32), and by Q the polyhedron of its LLL rangespace reformulation. If

$$\lambda > 2^{(n-1)/2}(\|r\| + 1)^2\|p\|, \quad (48)$$

then

$$\begin{aligned} \text{width}(e_n, Q) &= \text{width}(p, P), \text{ and} \\ \text{iwidth}(e_n, Q) &= \text{iwidth}(p, P). \end{aligned} \quad (49)$$

Sketch of proof Let us write

$$A = \begin{pmatrix} a \\ I \end{pmatrix}. \quad (50)$$

First we show that $\mathbb{L}(A)$ has at least $n-1$ vectors with norm bounded by $(\|r\| + 1)\|p\|$.

Indeed, there are $n - 1$ vectors in $\mathbb{N}(p)$ with norm at most $\|p\|$, and if w is such a vector, then

$$\|Aw\| \leq (\|r\| + 1)\|p\|. \quad (51)$$

Next, let U be the unimodular matrix in the LLL rangespace reformulation of (32). Since AU is an LLL reduced basis of $\mathbb{L}(A)$, Proposition 1 shows that its first $n - 1$ columns have norm of at most

$$2^{(n-1)/2}(\|r\| + 1)\|p\|. \quad (52)$$

The integrality of pU , and the choice of λ implies that the first $n - 1$ components of pU are 0, otherwise the corresponding column of AU would have norm larger than as given in (52). For the details we refer to Krishnamoorthy and Pataki [5] (p. 262).

Hence $pU = \delta e_n$ holds for some integer δ , that is, $p = \delta e_n U^{-1}$, and since the components of p are relatively prime, $\delta = \pm 1$.

Finally, the definitions of P and Q imply

$$\max\{px \mid x \in P\} = \max\{pUy \mid y \in Q\}, \quad (53)$$

and the same holds for the minimum. Using $pU = \pm e_n$ with (53), and the definition of width and integer width prove (49).

Remark 3. One can see that an analogous result holds for the KZ rangespace reformulation, even if we replace the lower bound (48) with the weaker

$$\lambda > \sqrt{(n+3)/2}(\|r\| + 1)^2\|p\|. \quad (54)$$

A variant of Theorem 1 about the nullspace reformulation is proven in Krishnamoorthy and Pataki [5].

Remark 4. Another decomposable knapsack instance, similar to Avis's, was described by Todd in Chvátal's 1980 paper [21]. It is interesting that Jeroslow, Avis, and Todd proved that their instances fit property (i), but they did not mention that they fit property (ii) as well.

Aardal and Lenstra [8,9] showed that denoting by b_{n-1} the last column of

the constraint matrix in the nullspace reformulation

$$\|b_{n-1}\| \geq \|b_{n-1}^*\| = \Omega(\lambda)$$

holds, and they argued that $\|b_{n-1}\|$ being long implies that branching on y_{n-1} will generate a small number of subproblems.

Krishnamoorthy and Pataki [5] pointed out a gap in the proof of $\|b_{n-1}^*\| = \Omega(\lambda)$, and constructed an example of a polyhedron $Q = \{y \in \mathbb{R}^r \mid \ell \leq By \leq w\}$, where the columns form an LLL reduced basis of $\mathbb{L}(B)$, but branching on y_r creates $c^r\|b_r\|$ subproblems for some $c > 1$. Furthermore, they proved that the instance of Example 4 fits property (ii).

Analyzing the Reformulation Methods without Assuming Structure

Here we describe an analysis of the reformulation methods based on the paper of Pataki *et al.* [10], without assuming any structure on the matrix A in (29). Interestingly, we will find that ordinary B&B solves the reformulation of the *majority* of the instances *without any branching*. We explain the connection with solving low-density subset sum problems after the proof.

We assume $n \geq 5$, and when a statement is true for all, but at most a fraction of $1/2^n$ of the elements of a set S , we say that it is true for *almost all* elements. For positive integers m, n , and M we denote by $G_{m,n}(M)$ the set of matrices with m rows and n columns, and the entries from $\{1, \dots, M\}$, and by $G'_{m,n}(M)$ the subset of $G_{m,n}(M)$ consisting of matrices with linearly independent rows.

We use a version of ordinary B&B that branches on the variables in reverse order, and call this algorithm *reverse B&B*. If B&B generates at most one node at each level of the tree, we say that it solves an integer feasibility problem *at the root node*. When the system $Ax = \ell_1$ does not have an integral solution, the nullspace reformulation does not exist; for simplicity we still say that in this case the reformulated instance is solved at the root node.

Theorem 2. *If $M \geq (2^{(n+4)/2} \|(w_1; w_2) - (\ell_1; \ell_2)\|)^{n/m+1}$, then for almost all $A \in G_{m,n}(M)$ reverse B&B solves the LLL rangespace reformulation of (29) at the root node.*

Also, if $M \geq (2^{(n-m+4)/2} \|w_2 - \ell_2\|)^{n/m}$, then for almost all $A \in G'_{m,n}(M)$ reverse B&B solves the LLL nullspace reformulation of (29) at the root node.

Proof Sketch We outline a proof of the first statement, and refer the reader to Pataki *et al.* [10] for details, and the proof of the second. For convenience, we shall write $(A; I)$ for the matrix obtained by stacking A on top of I , and the meaning of $(\ell_1; \ell_2)$ and $(w_1; w_2)$ will be analogous.

Let U be the matrix such that the columns of $(A; I)U$ form an LLL reduced basis of the generated lattice. We first use Corollary 3 with $B = (A; I)U$, and $Q = \{y \mid (\ell_1; \ell_2) \leq y \leq (w_1; w_2)\}$ to find that when reverse B&B is applied to (30), the number of B&B nodes on the level of y_j is at most

$$\prod_{i=j}^n \left(\left\lfloor \frac{\|(w_1; w_2) - (\ell_1; \ell_2)\|}{\|b_i^*\|} \right\rfloor + 1 \right),$$

where $b_1^*, \dots, b_n^*$ form the Gram–Schmidt orthogonalization of the columns of $(A; I)U$. Hence if

$$\|b_i^*\| > \|(w_1; w_2) - (\ell_1; \ell_2)\| \quad \forall i = 1, \dots, n, \quad (55)$$

then the problem is solved at the root node.

The definition of LLL reducedness implies

$$\begin{aligned} \|b_i^*\| &\geq \frac{1}{2^{(i-1)/2}} \|b_1\| \\ &\geq \frac{1}{2^{(i-1)/2}} \lambda_1(\mathbb{L}(A; I)), \end{aligned} \quad (56)$$

where $\lambda_1(\mathbb{L}(A; I))$ denotes the length of the shortest nonzero vector in $\mathbb{L}(A; I)$. So (55) holds, when

$$\lambda_1(\mathbb{L}(A; I)) > 2^{(n-1)/2} \|(w_1; w_2) - (\ell_1; \ell_2)\|. \quad (57)$$

Condition (57) does not hold for all A matrices, so let us call a matrix $A \in G_{m,n}(M)$ *bad*, when (57) fails. One can show that for

$r > 0$ the shortest vector in $\mathbb{L}(A; I)$ is strictly longer than r for all, but at most a fraction of

$$\frac{(2\lfloor r \rfloor + 1)^{n+m}}{M^m} \quad (58)$$

matrices in $G_{m,n}(M)$. We refer the reader to Lemma 2.2 in Pataki *et al.* [10] for details.

Using this result, it follows that when M is as given in the first statement of the theorem, the fraction of bad matrices is at most $1/2^n$, and this completes the proof.

Remark 5. A stronger version of Theorem 2 is true, if we use a “more reduced” basis, in particular, a so-called *reciprocal KZ reduced basis*. For details, we refer to Pataki *et al.* [10].

There is an interesting connection with earlier work on subset sum problems, which we outline here. Furst and Kannan [22] based on Lagarias’ and Odlyzko’s [23] and Frieze’s [24] work show that the subset sum problem

$$\begin{aligned} ax &= \beta \\ x &\in \{0, 1\}^n, \end{aligned} \quad (59)$$

is solvable in polynomial time using a simple iterative method for almost all $a \in G_{1,n}(M)$, and all right-hand sides, when M is sufficiently large, and a reduced basis of $\mathbb{N}(a)$ is available. Their bound on M is $2^{n^2/2+2n} n^{3n/2}$, when the basis is LLL reduced, and $2^{(3/2)n \log n + 5n}$, when it is reciprocal KZ reduced.

Subset sum problems with potentially such large coefficients find uses in cryptography. The vector a is a public key, x is a message, and the encoded message that the sender transmits is ax . The wide range of the coefficients of a makes sure that few right-hand sides among the integers in $\{1, \dots, \sum_{i=1}^n a_i\}$ arise as ax for some $x \in \{0, 1\}^n$, and it is rare for two distinct x vectors to map to the same ax . The results of Lagarias and Odlyzko [23], Frieze [24], and a later improvement by Coster *et al.* [25] show that using basis reduction the solution of (59) can be found with high probability, if it is known to exist. In other words, an intercepted message can be decoded for most

of the a public keys. Furst and Kannan [22] go further by finding the solution if it exists, and a proof of infeasibility for instances that do not have a solution.

Our bounds obtained by letting $m = 1$ in Theorem 2 and its variant that uses reciprocal KZ reducedness are comparable to Furst and Kannan's bounds, when the size of M , that is, $\lceil \log(M + 1) \rceil$ is concerned. Precisely, the bound on the size of M is $O(n^2)$, when we use an LLL reduced basis, and $O(n \log n)$, when we use a reciprocal KZ reduced basis.

So these results generalize the solvability results of [22] from subset sum problems to bounded integer programs. It is also interesting that one can prove complexity results via branch-and-bound, an algorithm that has been considered inefficient from the theoretical point of view.

FURTHER READING AND COMPUTATIONAL TESTING

In this section we briefly review results on integer programming in fixed dimension, whose detailed treatment is beyond the scope of our survey, mention other surveys that the reader may find to be of interest, and review computational experience with lattice-based methods. We will write “polynomial time when the dimension is fixed” as *fd-polynomial time* for short.

The generalized basis reduction algorithm of Lovász and Scarf [26] also solves (IP) in *fd-polynomial time*. Instead of rounding the underlying polytope, like Lenstra's and Kannan's algorithms do, at its core is a subroutine that finds an integral vector p such that $\text{width}(p, P)$ is relatively small, by solving a sequence of linear programs. This vector is then used for branching. Kannan [27] presented an algorithm to solve the Frobenius problem in *fd-polynomial time*.

An *fd-polynomial time* algorithm exists even to *count* the number of feasible solutions of (IP). The breakthrough algorithm to achieve this is due to Barvinok [28]. His algorithm was considerably simplified by Dyer and Kannan [29], and successfully implemented by De Loera *et al.* [30]. Koeppel [31]

developed, and implemented a newer primal variant, which in many cases is also faster in practice.

Given $c \in \mathbb{Z}^n$, the integer optimization problem is finding a feasible solution of (IP), which maximizes cx . One can solve this problem by reducing it to a sequence of feasibility problems; however, it is interesting to study direct approaches, which are theoretically efficient. We refer to Eisenbrand [32] for a fast algorithm to solve this problem, under the assumption that the number of variables, and the number of constraints are both fixed. Computing the *integer programming gap* is the problem of finding the maximum difference between the optimal value of an integer programming problem, and its LP relaxation, as the right-hand side varies. An *fd-polynomial time* algorithm for this problem was developed by Hoşten and Sturmfels [33], assuming that the number of constraints is also fixed. Eisenbrand and Shmonin [34] described an *fd-polynomial* algorithm even when the number of constraints is allowed to vary.

Other reviews on the uses of basis reduction and integer programming that the reader may find useful are by Kannan [35], Aardal and Eisenbrand [36], and Eisenbrand [37]: the latter is also a tutorial, with accompanying exercises. A substantial part of the books of Schrijver [13], and Grötschel *et al.* [38] are also devoted to this subject.

There is surprisingly little experience with implementing Lenstra's algorithm. Gao and Zhang [39] described an implementation, and Mehrotra and Li [6] presented and implemented a nonrecursive variant, in which branching is done on a hyperplane in the original formulation. Cook *et al.* [40] reported a successful implementation of the generalized basis reduction method.

There is more computational evidence of the effectiveness of the reformulation methods.

Aardal *et al.* [4] successfully tested their reformulation on knapsack-type feasibility problems that arise from circuit design. Another application of lattice-based methods is on the *marketshare* problems of Cornuéjols and Dawande [41]. Suppose that a company supplies n retailers with m products, with

retailer j receiving a_{ij} units of product i . The company has two divisions. We would like to assign each retailer to one of the divisions, so the retailers in each division receive approximately half of the total supply of each product.

Letting A be a matrix with the (i, j) th entry equal to a_{ij} , and $b \in \mathbb{Z}^m$ with the i th entry equal to

$$\left\lfloor \frac{1}{2} \sum_{j=1}^n a_{ij} \right\rfloor,$$

the problem can be formulated as

$$\begin{aligned} Ax &= b \\ x &\in \{0, 1\}^n, \end{aligned} \quad (60)$$

where x_j is set to 1, if retailer j is assigned to division 1, and 0 otherwise.

Two variants of (60) have also been studied, which are especially interesting, when the original problem is infeasible. The first is an optimization version introduced in Cornuéjols and Dawande [41], which attempts to minimize $\|Ax - b\|_1$. The second is a relaxed version studied in Pataki *et al.* [10], namely,

$$\begin{aligned} b - e &\leq Ax \leq b \\ x &\in \{0, 1\}^n. \end{aligned} \quad (61)$$

This formulation attempts to find a near equal market split (of course if $\sum_{j=1}^n a_{ij}$ is odd, then the i th constraints in (61) are as good as the i th constraint in (60)).

The marketshare problems are exceptionally difficult to solve by commercial integer programming software, and Cornuéjols and Dawande offered them as a challenge to the Integer Programming community. They generated the a_{ij} uniformly at random in the interval $\{1, \dots, 100\}$, with the choice $n = 10(m - 1)$. Aardal *et al.* [42] showed that by using the CPLEX 6.5 commercial Mixed Integer Programming (MIP) solver, the nullspace reformulations of 7×60 instances were solved in a reasonable amount of time, whereas the original formulations of even 5×60 instances could not be handled. Improved results were obtained for the optimization versions as well, and the

authors also derived an approximation for the number of feasible solutions of (60) in terms of m and n . A generalization of the marketshare problem with a matrix variables, and two-sided constraints was studied by Louveaux and Wolsey [43].

A counterintuitive guess based on Theorem 2 is that the reformulations of the marketshare problems should become *easier* in practice, when the a_{ij} are drawn from $\{1, \dots, M\}$, and M grows. This was confirmed by computational experiments by Pataki *et al.* [10]. For instance, the average number of nodes over 12 instances that needed to be enumerated by CPLEX 9 to solve the rangespace reformulation of 5×40 relaxed instances with $M = 100$ was 38865. However, the average number of nodes with $M = 10000$ was just 1976.

A computational study on the Frobenius instances given in Example 4 was carried out by Aardal and Lenstra [8]. Krishnamoorthy and Pataki [5] experimented with more general DKPs, both with feasible and infeasible instances, and using varying bounds. As expected, the reformulations were quite easy to solve, usually requiring less than a hundred nodes, even when λ was not large enough for Theorem 1 (or its version using KZ reducedness) to give theoretical guarantees.

Two other interesting observations were made in Krishnamoorthy and Pataki [5]: first, when the knapsack problem has an equality constraint, and so both reformulations were applicable, there was no difference in their performance. Second, Theorem 1, which asserts that branching on a single variable in the reformulation is equivalent to branching on px in the original problem is verified by another experiment: creating a new variable z , and adding the redundant constraint $z = px$ to the original formulations. Even without specifying a higher priority for branching on the z variable, the *original* instances with this addition solved as fast as the reformulations.

Acknowledgments

Mustafa Tural would like to thank Telcordia Technologies for their hospitality for the duration of this work.

REFERENCES

1. Lenstra HW Jr. Integer programming with a fixed number of variables. *Math Oper Res* 1983;8:538–548. (First announcement (1979).)
2. Kannan R. Improved algorithms for integer programming and related lattice problems. *Proceedings of the 15th Annual ACM Symposium on Theory of Computing*; Boston (MA). New York: The Association for Computing Machinery; 1983. pp. 193–206.
3. Kannan R. Minkowski’s convex body theorem and integer programming. *Math Oper Res* 1987;12:415–440.
4. Aardal K, Hurkens CAJ, Lenstra AK. Solving a system of linear Diophantine equations with lower and upper bounds on the variables. *Math Oper Res* 2000;25:427–442.
5. Krishnamoorthy B, Pataki G. Column basis reduction and decomposable knapsack problems. *Discrete Optim* 2009;6:242–270.
6. Mehrotra S, Li Z. Branching on hyperplane methods for mixed integer linear and convex programming using adjoint lattices. *J Glob Optim* 2010. DOI: 10.1007/s10898-010-9554-4.
7. Jeroslow RG. Trivial integer programs unsolvable by branch-and-bound. *Math Program* 1974;6:105–109.
8. Aardal K, Lenstra AK. Hard equality constrained integer knapsacks. *Math Oper Res* 2004;29:724–738.
9. Aardal K, Lenstra AK. Erratum to: Hard equality constrained integer knapsacks. *Math Oper Res* 2006;31:846.
10. Pataki G, Tural M, Wong EB. Basis reduction and the complexity of branch-and-bound. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Austin (TX): SIAM; 2010. pp. 1254–1261.
11. Babai L. On Lovász lattice reduction, and the nearest lattice point problem. *Combinatorica* 1986;6:1–13.
12. Land AH, Doig AG. An automatic method for solving discrete programming problems. *Econometrica* 1960;28:497–520.
13. Schrijver A. *Theory of linear and integer programming*. Chichester, UK: Wiley; 1986.
14. Lenstra AK, Lenstra HW Jr, Lovász L. Factoring polynomials with rational coefficients. *Math Ann* 1982;261:515–534.
15. LiDIA. A library for computational number theory. <http://www.cdc.informatik.th-darmstadt.de/TI/LiDIA/>.
16. Shoup V. NTL: A number theory library. <http://www.shoup.net>. 1990.
17. Korkine A, Zolotareff G. Sur les formes quadratiques. *Math Ann* 1873;6:366–389.
18. Lagarias JC, Lenstra HW, Schnorr CP. Korkine-Zolotarev bases and successive minima of a lattice and its reciprocal lattice. *Combinatorica* 1990;10:333–348.
19. Ramirez Alfonsin JL. *The Diophantine Frobenius problem*. Oxford lecture series in mathematics and its applications. New York: Oxford University Press; 2005.
20. Cornuéjols G, Urbaniak R, Weismantel R, *et al.* Decomposition of integer programs and of generating sets. Volume 1284, *Algorithms - ESA 1997. Lecture notes in computer science*. London, UK: Springer; 1997. pp. 92–103.
21. Chvátal V. Hard knapsack problems. *Oper Res* 1980;28:1402–1411.
22. Furst M, Kannan R. Succinct certificates for almost all subset sum problems. *SIAM J Comput* 1989;18:550–558.
23. Lagarias JC, Odlyzko AM. Solving low-density subset sum problems. *J ACM* 1985;32:229–246.
24. Frieze A. On the Lagarias-Odlyzko algorithm for the subset sum problem. *SIAM J Comput* 1986;15:536–540.
25. Coster MJ, Joux A, LaMacchia BA, *et al.* Improved low-density subset sum algorithms. *Comput Complex* 1992;2:111–128.
26. Lovász L, Scarf HE. The generalized basis reduction algorithm. *Math Oper Res* 1992;17:751–764.
27. Kannan R. Lattice translates of a polytope and the Frobenius problem. *Combinatorica* 1992;12:161–177.
28. Barvinok AI. A polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math Oper Res* 1994;19:769–779.
29. Dyer M, Kannan R. On Barvinok’s algorithm for counting lattice points in fixed dimension. *Math Oper Res* 1997;22:545–549.
30. De Loera JA, Hemmecke R, Tauzer J, *et al.* Effective lattice point counting in rational convex polytopes. *J Symb Comput* 2004;38:1273–1302.
31. Koeppe M. A primal Barvinok algorithm based on irrational decompositions. *SIAM J Discrete Math* 2007;21:220–236.
32. Eisenbrand F. Fast integer programming in fixed dimension. 11th Annual European

- Symposium on Algorithms - ESA 2003. Volume 2832, Lecture notes in computer science. Berlin, Germany: Springer; 2003. pp. 196–207.
33. Hoşten S, Sturmfels B. Computing the integer programming gap. *Combinatorica* 2007;27:367–382.
 34. Eisenbrand F, Shmonin G. Parametric integer programming in fixed dimension. *Math Oper Res* 2008;33:839–850.
 35. Kannan R. Algorithmic geometry of numbers. *Ann Rev Comput Sci. Palo Alto (CA)*: 1987;2:231–267.
 36. Aardal K, Eisenbrand F. Integer programming, lattices, and results in fixed dimension. *Discrete optimization. Volume 12, Handbooks in operations research, and management science.* Amsterdam, The Netherlands: Elsevier; 2005. pp. 171–243.
 37. Eisenbrand F. Integer programming and algorithmic geometry of numbers. *50 Years of integer programming 1958–2008.* Berlin, Germany: Springer; 2010. pp. 505–559.
 38. Grötschel M, Lovász L, Schrijver A. *Geometric algorithms and combinatorial optimization.* Volume 2, Algorithms and combinatorics. 2nd corrected ed. Berlin, Germany: Springer; 1993.
 39. Gao L, Zhang Y. Computational experience with Lenstra’s algorithm. Technical Report #TR02–12. Department of Computational and Applied Mathematics, Rice University; 2002.
 40. Cook W, Rutherford T, Scarf HE, *et al.* An implementation of the generalized basis reduction algorithm for integer programming. *ORSA J Comput* 1993;5:206–212.
 41. Cornuéjols G, Dawande M. A class of hard small 0–1 programs. *INFORMS J Comput* 1999;11:205–210.
 42. Aardal K, Bixby RE, Hurkens CAJ, *et al.* Market split and basis reduction: towards a solution of the Cornuéjols-Dawande instances. *INFORMS J Comput* 2000;12: 192–202.
 43. Louveaux Q, Wolsey LA. Combining problem structure with basis reduction to solve a class of hard integer programs. *Math Oper Res* 2002;27:470–484.

BATCH ARRIVALS AND SERVICE—SINGLE STATION QUEUES

RAJA JAYARAMAN

Department of Industrial Engineering,
University of Arkansas, Fayetteville,
Arkansas

TIMOTHY I. MATIS

Department of Industrial Engineering,
Texas Tech University, Lubbock, Texas

INTRODUCTION

Queueing models exhibiting arrival and/or service patterns in batches or groups are commonly classified as bulk queueing models. The earliest studies on batch queueing models were applied to the patient appointment scheduling problem, in which the steady-state queue length distribution was characterized and obtained [1,2]. More recent applications of bulk queueing models include computer high-speed random access networks [3], data packet switching [4], Very Large Scale Integrated (VLSI) circuits [5,6], manufacturing systems [7], and transportation systems [8–10], among several others. Unlike the analysis of traditional queueing models, the study of bulk queues is not straightforward.

In traditional queueing analysis, it is often assumed that customers arrive singly; however, this assumption seems invalid in several real-world applications as customers often arrive in batches. A simple example of *batch arrival* queues includes packages arriving by courier truck at a sorting facility, pallets arriving at a retail warehouse, and so on. Note that in these examples, arrivals do not come singly, but rather in bulk or batches where the size of the batch can be a fixed constant or a random variable. In other instances, customers may arrive singly to a server, but are subsequently served in

batches of a fixed or random size. Examples of such *batch service* queues include vehicles waiting at a traffic light, people in an elevator, passengers in an airplane, and so on.

The analysis of a single-server *bulk service* queue, where customers are served in batches of size k , has three variations: (i) a *full batch* model; (ii) a *partial batch* model, in which customers can join the server while it is in process until a full batch is reached; and (iii) a *Partial Batch* model, in which late arriving customers are not permitted to join the server directly. In a full batch model, the service begins only when k customers are present in the queue; that is, when less than k customers are present, the server waits until k customers have joined the queue and then processes all customers simultaneously. A simple example of a full batch model includes forklift operations, in which only complete pallets are transported that are built-up over time. In a partial batch model, service begins on the batch whether or not k are present once the server becomes free. In type (ii) partial batch model, new arrivals are permitted to join the batch of ongoing service until the threshold k is reached, and in type (iii) new arrivals have to wait until next time when the server is available. Note that, in a partial batch model, the service time is congruent to a complete batch, whether or not it is actually full. A simple example of partial batch model of type (ii) is a guided tour, in which late arriving customers are permitted to join the tour up to the threshold k and finish the tour as a group. An example of type (iii) partial batch model where late arriving passengers are not permitted to join the airplane. Although partial batch models of type (iii) are found in several real-life examples, they present inherent challenge and modeling difficulties. In this article, we do not consider the modeling and analysis of type (iii) partial batch model. In the next section, we present single-server batch models with Markovian arrivals and services, and extensions to general distribution.

SIMPLE BULK MODELS

Single-Server Markovian Batch Arrival Model:
 $M^{[X]}/M/1$

In this section, we study a single-server Markovian bulk arrival queue through the treatment of a simple $M/M/1$ queue as a modified birth–death process. Let us assume the arrival stream occurs in batches of size X , which denotes a random variable of nonnegative values with probability distribution $b_k = \Pr\{X = k\}$, $k \geq 1$. The arrival stream constitutes a compound Poisson process with an average arrival rate λ , and services are performed one at a time according to an exponential distribution with rate μ . Let $N(t)$ denote the number of customers in the system at time t , and P_n denote the probability that there are n customers in the system at time t given by $P_n(t) = \Pr\{N(t) = n\}$, $n \geq 0$. The state space of the system increases by nonunitary values based on the arriving group size and decreases by one with a service completion, it follows that $\{N(t)\}$ forms a modified birth–death process with steady-state balance equations given by

$$\begin{aligned} \lambda P_0 - \mu P_1 &= 0 \\ (\lambda + \mu)P_n - \lambda \sum_{k=1}^n b_k P_{n-k} - \mu P_{n+1} &= 0, \\ \text{for } n \geq 1 \end{aligned}$$

where $P_n = \lim_{z \rightarrow 1} P_n(z)$. To solve these equations, we employ probability generating functions denoted by $P(z)$ for the steady-state probabilities and $B(z)$ for the batch size probabilities, defined by

$$P(z) = \sum_{n=0}^{\infty} P_n z^n$$

and

$$B(z) = \sum_{k=1}^{\infty} b_k z^k \quad \text{with } |z| \leq 1. \quad (1)$$

Multiplying the balance equations by suitable powers of z , interchanging the summations, and simplifying the expressions give

$$\begin{aligned} (\lambda + \mu)P(z) - \mu P_0 - \lambda \sum_{k=1}^{\infty} b_k z^k \sum_{n=k}^{\infty} P_{n-k} z^{n-k} \\ - \mu \sum_{n=0}^{\infty} P_{n+1} z^n = 0. \end{aligned} \quad (2)$$

The middle term in the above equation is merely a convolution of discrete random variables representing the product of steady-state probabilities and batch size probabilities, that is

$$\begin{aligned} B(z)P(z) &= \sum_{k=1}^{\infty} b_k z^k \sum_{n=k}^{\infty} P_{n-k} z^{n-k} \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^n P_{n-k} b_k z^n. \end{aligned}$$

Further simplification of Equation (1) gives $P(z)$ in terms of the batch size probabilities $B(z)$ and P_0 , from whence

$$\begin{aligned} (\lambda + \mu)P(z) - \mu P_0 - \lambda B(z)P(z) \\ - \frac{\mu}{z} \sum_{m=1}^{\infty} P_m z^m = 0 \\ (\lambda + \mu)P(z) - \mu P_0 - \lambda B(z)P(z) \\ - \frac{\mu}{z} [P(z) - P_0] = 0 \\ P(z) = \frac{\mu P_0 (1 - z)}{\mu (1 - z) - \lambda z [1 - B(z)]} \\ \text{with } |z| \leq 1. \end{aligned} \quad (3)$$

To obtain P_0 , we use the normalizing condition $\sum_n P_n = 1$, from whence $\lim_{z \rightarrow 1} P(z) = 1$ yielding the expression

$$P_0 = 1 - \frac{\lambda B'(1)}{\mu} = 1 - \rho,$$

where $B'(1) = \lim_{z \rightarrow 1} \sum_{k=1}^{\infty} k b_k z^{k-1} = E(X) = b$ is the average batch size, and the corresponding traffic intensity based on the average batch size “ b ” is given by $\rho = \frac{\lambda b}{\mu}$. Substituting P_0 into $P(z)$ gives the expression

$$\begin{aligned} P(z) &= \frac{\mu (1 - \rho) (1 - z)}{\mu (1 - z) - \lambda z [1 - B(z)]} \\ &\quad \text{with } |z| \leq 1. \end{aligned} \quad (4)$$

Inverting the probability generating function to obtain P_n is analytically cumbersome, yet numerical methods can be employed to evaluate the system probabilities for a specified batch size distribution. Two special cases of either a constant or geometric batch size distribution are discussed in Gross *et al.* [11, pp. 120–121]. Obtaining the mean and variance of the number of customers in the system, however, is rather straightforward through the application of L' Hospital's rule, since the expression for $P(z)$ contains $(1 - z)$ in the numerator and $(1 - B(z))$ in the denominator. Therefore, the mean number of customers in the system at steady state is given by

$$\begin{aligned} E[N] &= \lim_{z \rightarrow 1} P'(z) = \frac{2\rho + \frac{\lambda}{\mu} B''(1)}{2(1 - \rho)} \\ &= \frac{\rho + \frac{\lambda}{\mu} E[X^2]}{2(1 - \rho)}, \end{aligned} \quad (5)$$

as the average batch size is b and $B''(1) = E[X^2] - E[X]^2$. Similarly, the variance of the number of customers in the system at steady state can be obtained after the repeated application of L' Hospital rule by evaluating $\lim_{z \rightarrow 1} P''(z)$.

Single-Server Markovian Batch Service Model: $M/M^{[X]}/1$

The single-server batch service queue assumes a unity Poisson arrival process with rate λ , and exponentially distributed service times with rate μ for batches of size X . As noted previously, there are two variations of a batch service model, that is, (i) *full batch* case, models where the server waits until exactly k customers are present in the system before processing; and (ii) *partial batch* case, models where the server can start servicing when fewer than k customers are present in the system, thereby permitting all new arrivals to go directly into service until the threshold of k customers is reached. In either case, customers depart as a batch, either full or partial, at a service completion epoch, not singly. Interestingly, depending on the assumption of case (i) or (ii) leads to very different results for the model while

preserving common modeling techniques used to analyze the model. For analysis and results on case (ii) *partial batch* service models, the readers are referred to Gross *et al.* [11, pp. 124]. In the analysis presented here, however, we restrict our attention to case (i) *full batch* service models that are more common in practice.

Let the number of customers in the system be denoted by $\{P_n, n \geq 0\}$, for which the steady-state balance equations of the full service batch model is given by

$$\begin{aligned} \lambda P_0 - \mu P_k &= 0 \\ \lambda P_n - \lambda P_{n-1} - \mu P_{n+k} &= 0 \quad \text{for } 1 \leq n \leq k \\ (\lambda + \mu) P_n - \lambda P_{n-1} - \mu P_{n+k} &= 0 \quad \text{for } n \geq k. \end{aligned} \quad (6)$$

Multiplying the above equations with suitable powers of z , and employing the probability generating functions defined in Equation (1) yield the expression

$$\begin{aligned} \lambda P_0 + \lambda \sum_{n=1}^{k-1} P_n z^n + (\lambda + \mu) \sum_{n=k}^{\infty} P_n z^n - \mu P_k \\ - \lambda \sum_{n=1}^{k-1} P_{n-1} z^n - \mu \sum_{n=1}^{k-1} P_{n+k} z^n - \lambda \sum_{n=k}^{\infty} P_{n-1} z^n \\ - \mu \sum_{n=k}^{\infty} P_{n+k} z^n = 0. \end{aligned}$$

Simplifying further and multiplying with z^k yields

$$\begin{aligned} \left[(\lambda + \mu) z^k - \lambda z^{k+1} - \mu \right] P(z) \\ - \mu \left(z^k - 1 \right) \sum_{n=0}^{k-1} P_n z^n = 0. \end{aligned} \quad (7)$$

Rearranging these terms yields the probability generating function

$$\begin{aligned} P(z) &= \frac{(1 - z^k) \sum_{n=0}^{k-1} P_n z^n}{1 + \left(\frac{\lambda}{\mu} \right) z^{k+1} - \left(1 + \frac{\lambda}{\mu} \right) z^k} \\ &\quad \text{with } |z| \leq 1. \end{aligned} \quad (8)$$

To solve Equation (8) and obtain the state probabilities P_n , we need to evaluate $\sum_{n=0}^{k-1} P_n z^n$. The defined probability generation function $P(z)$ should converge in the interval $|z| \leq 1$, and vanish at all the $(z + 1)$ zeros. Note that, we can use Rouché's theorem to both prove the existence of a certain number of zeros in the domain of regularity for the given function and to satisfy concerns of ergodicity in constructing the solution of the generating function in steady state, as explained in Adan *et al.* [12] and Klimenok [13]. In particular, the denominator of Equation (8) is a $(k + 1)$ degree polynomial, and we need to show that there exist $(k - 1)$ zeros of this polynomial inside $|z| \leq 1$, since $z = 1$ is a zero for both numerator and denominator, leaving exactly one root, say z_0 , outside the unit circle. Note that dividing the denominator of Equation (8) by $(z - z_0)(z - 1)$ will lead to exactly $(k - 1)$ roots in $|z| \leq 1$, and the number of zeros in the numerator of Equation (8) within $|z| \leq 1$ given by $\sum_{n=0}^{k-1} P_n z^n$ must be equal to the roots of denominator, leaving utmost a constant say C ; therefore,

$$\sum_{n=0}^{k-1} P_n z^n = C \frac{1 + \left(\frac{\lambda}{\mu}\right) z^{k+1} - \left(1 + \frac{\lambda}{\mu}\right) z^k}{(z - z_0)(z - 1)}. \quad (9)$$

Substituting the approximation given by Equation (9) into Equation (8) yields the expression

$$P(z) = C \frac{(1 - z^k)}{(z - z_0)(z - 1)} = \frac{C}{(z - z_0)} \sum_{n=0}^{k-1} z^n. \quad (10)$$

Since $P(z)$ is a probability generating function, $P(z = 1) = 1$, which yields $C = \frac{(z_0 - 1)}{K}$ from Equation (10). Substituting this back yields the expression

$$P(z) = \frac{(z_0 - 1) \sum_{n=0}^{k-1} z^n}{K(z_0 - z)}. \quad (11)$$

Expanding the above expression as power series in z gives

$$P_n = \frac{z_0^{n+1} - 1}{K z_0^{n+1}} \quad \text{for } n < k \quad (12)$$

$$P_n = \frac{z_0^k - 1}{K z_0^{n+1}} \quad \text{for } n \geq k. \quad (13)$$

Finding the root z_0 outside the circle $|z| > 1$ is the key to obtaining the system probabilities. Readers are referred to methods and algorithms described in Chaudhry and Templeton [14], Harris *et al.* [15], van Leeuwen and Janssen [16], and Zhao and Campbell [17].

Single-Server Model with General Distribution: $M^{[X]}/G/1$ and $G/M^{[X]}/1$

In the previous sections, we considered bulk queues with Poisson arrival rates and exponential service times. These Markovian assumptions, however, may be relaxed by incorporating general distributions for service containing Poisson batch arrivals, $M^{[X]}/G/1$ queues, and general arrivals having exponential services in batches, $G/M^{[X]}/1$ queues. Due to the non-Markovian nature, both models may be studied using imbedded Markov chains by a straightforward extension of $M/G/1$ and $G/M/1$ queues, [14, Section (2.2.3)]. Additionally, the work of Chaudhry [18] contains an extensive analysis of the $M^{[X]}/G/1$ queue.

EXTENSIONS AND FURTHER READING

An important class of bulk queues includes multiserver (or multichannel) models, in which more than one server is available to process batch arrivals and/or services. The number of available servers can be either finite or infinite; yet, in the latter, only server occupancy will be the measure of interest. Examples of multiserver queues include toll booth, banks, and telephone exchanges. Multiserver bulk models require sophisticated arguments, methods, and involved algebra, yet. Interested readers are referred to Chaudhry and Templeton [14] for a detailed analysis of multiserver models with bulk arrival and service involving finite and infinite servers. In addition, generalizations of Erlangian distributions with a fixed number of phases are extended

to the study of bulk queueing models in Chaudhry and Templeton [14]. They present the analysis of $M/E_k/1$ queue to obtain an implicit solution for bulk input $M^{[X]}/M/1$ queue. The method of phases also helps to analyze non-Markovian models by constructing appropriate phase representation model variables, that is, an $E_k^X/E_r/1$ approximation to a $G^X/G/1$ queue. In general, the technique of matrix geometric methods (MGM) has been heavily used to study bulk queueing models to obtain steady-state performance measures. Most notably, Neuts [19,20,35] provides several batch arrival models along with methods and techniques for their analysis in his seminal work.

Another variant of bulk queues includes the batch arrival of customers with different priority classes, upon which service is given to customers first of a higher priority class both with and without service preemption. As an example, consider a single-server bulk arrival queue with two types of arrivals (say X_1 and X_2), where X_1 has higher priority over X_2 customers, whereupon they are served according to their respective service distribution (G_1, G_2) , that is, an $M^{[X_1, X_2]}/G_1, G_2/1$ model. Such models are often studied using supplementary variables and can be found in Bhat [21,22] and Chaudhry and Templeton [14]. The state probabilities of these queues are obtained using Rouches theorem, and Able's theorem is employed to obtain the joint probability distribution, from whence the waiting time distributions for both types of customers may be found.

A sampling of other notable extensions and literature related to bulk queueing models is as follows. Prabhu [23] studies stochastic ordering properties of three important family of queueing models, namely $M/G^{[s]}/1$, $G^{[s]}/M/1$, and $M/D/s$ for $s \geq 1$, to obtain bounds for the expected queue lengths and number of batches during busy period. Gaver [24] studied the $M^{[X]}/G/1$ using imbedded Markov chain techniques, and an extensive analysis of this model is also discussed in Chaudhry [18]. The imbedded Markov chain technique has been employed to study the $M/G^{[X]}/1$ queue by considering the number of customers in the

system at the n th group departure epoch, and the $G^{[X]}/M/1$ queue by considering the number of customers in the system just before an arrival epoch [21, Section 6.4]. Chiamsiri and Leonard [25] use a diffusion approximation method to study single-server queues with batch arrival and services. Chen and Renshaw [26] study Markovian queues with bulk arrivals, modified to allow mass arrivals and departures when the queue is idle, to derive useful properties in closed form. Dikong and Dshalalow [27] develop analytical solutions of system performance for bulk input model under an N -policy and r -quorum, which incorporates state-dependent services to control the server capacity by limiting the number of switchovers between idle and busy modes of the server. Chang *et al.* [5] discuss the performance analysis of finite-buffer bulk arrival and service queue with a variable server capacity, as well as obtain numerically stable relationships for system probabilities and queue lengths at arrival, departure, and random epochs. Transient solutions for bulk queues have been obtained for a limited number of models due to analytical and modeling complexities. In this area, however, Jaiswal [28] obtains the transient solution of the bulk service problem, and Selim [29] obtains closed form solutions for the transient probability distribution of the number of customers in the system to ultimately predict the optimal service rates. Powell and Humblet [30] additionally analyze bulk service queue with a general control strategy to develop new computational procedures. Finally, Curry and Feldman [31] studied the $M/M/1$ queue under bulk service rules, Alexander [32] develops solutions for $GI/M/1$ systems with unbounded arrival groups, Baba [33] studied the $GI/M/1$ queue with batch arrivals, and Alfa and He [34] develop an algorithmic analysis of discrete-time $GI^X/G^Y/1$ queues.

REFERENCES

1. Bailey NTJ. A study of queues and appointment systems in outpatient departments with special reference to waiting times. *J R Stat Soc Ser B* 1952;14:185–199.

2. Bailey NTJ. Queueing for medical care. New York: Springer; 1954.
3. Spiegel EM, Bisdikian C, Tantawi AN. Characterization of the traffic on high-speed token-ring networks. *Perform Eval* 1994;19: 47–72.
4. Zhao Y, Campbell LL. Performance analysis of a multibeam packet satellite system using random access techniques. *Perform Eval* 1996;24(3):231–244.
5. Chang SH, Choi DW, Kim TS. Performance analysis of a finite-buffer bulk-arrival and bulk-service queue with variable server capacity. *Stoch Anal Appl* 2005;22(5): 1151–1173.
6. Fujimoto RM. VLSI communication components for multicomputer networks. EECS-Technical Report(CSD-83-137). Berkeley (CA): UC Berkely; 1983.
7. Gold H, Tran-Gia P. Performance analysis of a batch service queue arising out of manufacturing system modelling. *Queueing Syst* 1993;14(3-4):413–426.
8. Powell WB, Humblet P. Iterative algorithms for bulk arrival, bulk service queues with Poisson and non-Poisson arrivals. *Transport Sci* 1986;20(2):65–79.
9. Siano HP, Powell WB. Waiting time distributions for transient bulk queues with general vehicle dispatching strategies. *Nav Res Logist* 1988;35(2):285–306.
10. Obilade T. An algorithm proposed for busy-period subcomponent analysis of bulk queues. *Acta Math Appl Sin* 1990;6(1): 35–39.
11. Gross D, Shortle JF, Thompson JM, *et al.* Fundamentals of queueing theory. 4th ed. Hoboken (NJ): John Wiley & Sons Inc.; 2008.
12. Adan IJBF, van Leeuwaardena JSH, Winands EMM. On the application of Rouché's theorem in queueing theory. *Oper Res Lett* 2006;34(3): 355–360.
13. Klimenok V. On the modification of Rouché's theorem for the queueing theory problems. *Queueing Syst* 2001;38:431–434.
14. Chaudhry ML, Templeton JGC. A first course in bulk queues. New York: John Wiley & Sons; 1983.
15. Harris CM, Marchal WG, Tibbs RW. An algorithm for finding characteristic roots of quasitriangular Markov chains. In: Bhat UN, editor, Queueing and related models. New York: Oxford University Press; 1992.
16. van Leeuwaarden JSH, Janssen AJEM. Analytic computation schemes for the discrete-time bulk service queue. *Queueing Syst* 2005;50:141–163.
17. Zhao YQ, Campbell LL. Equilibrium probability calculations for a discrete-time bulk queue model. *Queueing Syst* 1996;22:189–198.
18. Chaudhry ML. The queueing system $M^{[X]}/G/1$ and its ramifications. *Nav Res Logist* 1979;26: 667–674.
19. Neuts MF. A general class of bulk queues with Poisson input. *Ann Math Stat* 1967;38(3): 759–770.
20. Neuts MF. Queues solvable without Rouches theorem. *Oper Res* 1979;27:767–781.
21. Bhat UN. An introduction to queueing theory. New York: Springer Science Business Media; 2008.
22. Bhat UN. Some simple and bulk queueing systems [PhD Dissertation]. The University of Western Australia; 1964.
23. Prabhu NU. Stochastic comparisons for bulk queues. *Queueing Syst* 1987;1:265–277.
24. Gaver DP. Imbedded Markov chain analysis of a waiting line process in continuous time. *Ann Math Stat* 1958;30:698–720.
25. Chiamsiri S, Leonard MS. A diffusion approximation for bulk queues. *Manag Sci* 1981; 27(10):1188–1199.
26. Chen A, Renshaw E. Markovian bulk-arriving queues with state-dependent control at idle time. *Adv Appl Probab* 2004;36(2): 499–524.
27. Dikong EE, Dshalalow JH. Bulk input queues with hysteretic control. *Queueing Syst: Theory Appl* 1999;32(4):287–304.
28. Jaiswal NK. A bulk service queueing problem with variable capacity. *J R Stat Soc Ser B* 1961;23:143–148.
29. Selim SZ. Time dependent solution and optimal control of a bulk service queue. *J Appl Probab* 1997;34(1):258–266.
30. Powell WB, Humblet P. The bulk service queue with a general control strategy: Theoretical analysis and a new computational procedure. *Oper Res* 1986;34(2):267–275.
31. Curry G, Feldman RM. An $M/M/1$ queue with a general bulk service rule. *Nav Res Logist* 1985;32:595–603.
32. Alexander D. Matrix-geometric solutions for bulk $GI/M/1$ systems with unbounded arrival groups. *Stoch Model* 1999;15(3): 547–559.

- 33. Baba Y. A bulk service $GI/M/1$ queue with service rates depending on service batch size. *J Oper Res Soc Japan* 1996;39(1):25–35.
- 34. Alfa AS, He Q-M. Algorithmic analysis of the discrete time $GI^X/G^Y/1$ queueing system. *Perform Eval* 2008;65(9):623–640.
- 35. Neuts MF. Matrix geometric solutions in stochastic models: an algorithmic approach. Baltimore (MD): The John Hopkins University Press; 1981.

BATCH MARKOVIAN ARRIVAL PROCESSES (BMAP)

JAMES D. CORDEIRO

Department of Mathematics and Statistics,
Air Force Institute of Technology,
Wright Patterson AFB, Ohio

JEFFREY P. KHAROUFEH

Department of Industrial Engineering,
University of Pittsburgh, Pittsburgh,
Pennsylvania

INTRODUCTION

The batch Markovian arrival process (BMAP) is a stochastic point process that generalizes the standard Poisson process (and other point processes) by allowing for “batches” of arrivals, dependent interarrival times, non-exponential interarrival time distributions, and correlated batch sizes. The Markovian arrival process (MAP) is a special case of the BMAP in which the batch size is restricted to unity. For a detailed description of the MAP, see the article titled *Markovian Arrival Processes* in this encyclopedia.

The origins of the BMAP can be traced to the development of the *versatile Markovian point process* (VMPP) by Neuts [1] whose primary objective was to extend the standard Poisson process to account for more complex customer arrival processes in queueing models. The VMPP is characterized by three distinct classes of batch arrivals, each of which are determined by the transition type of an external Markov process with m transient states and one absorbing state, $m + 1$. One type of arrival is from a Markov-modulated Poisson process (MMPP), and this type occurs during the sojourn of the exogenous process in any of the m transient states. Another type occurs when the Markov process in state i transitions to state j , $j \neq i$, which is an ordinary transition between two transient states. The third type of arrival occurs when the process in transient state i

transitions to the absorbing state, $m + 1$, and then restarts in state j . This type of transition is called an (i, j) -renewal transition, and by virtue of restarting the Markov process, admits the possibility of a “self-transition” from a transient state i to itself. From this description, it is clear that the VMPP is founded upon the notion of a phase-type (PH) distribution, or the distribution of the time to absorption of an absorbing Markov process. Neuts [2] played a major role in advancing the use of the PH-distribution in queueing theory, culminating ultimately in the development of the VMPP.

Lucantoni *et al.* [3] sought to extend the original definition of the VMPP while simultaneously easing its notational burden by defining the MAP. The MAP also uses the concept of arrival dependence upon an external Markov process but does not distinguish between classes of arrivals. The MAP was generalized to the BMAP in Ref. 4 by permitting batch arrivals. Originally it was thought that the VMPP was a special case of the BMAP, and it was only later that Lucantoni and others [5,6] asserted the equivalence of the VMPP and BMAP. The term “BMAP” has persisted due to its widespread acceptance in the stochastic modeling community.

The analysis of queueing systems is assisted by the fact that they may often be modeled, either directly or via embedding, as structured Markov chains. Structured Markov chains are typically classified as two main types: the $GI/M/1$ -type and the $M/G/1$ -type. A well-known third type, the quasi-birth-and-death (QBD) process, can be viewed as the juxtaposition of the other two. Structured Markov chains help facilitate the use of *matrix-analytic methods* in the steady-state analysis of queueing systems with MAP and BMAP input. The use of matrix-analytic methods in the analysis of queueing systems is detailed in Neuts’ two classic texts [7,8], which describe the theory and method underlying the derivation of the stationary distributions of the structured Markov chains. The first queueing model

to be considered is the single-server model with infinite capacity. Ramaswami [9] incorporated the BMAP (or VMPP), which he called the N -process in honor of Neuts, as an arrival process to a single-server queue with generally distributed service times. From this work, a generalization of the Polleczech–Khinchin formula to the $N/G/1$ queue was derived. Basic results for the steady-state analysis of the $MAP/G/1$ queue are provided in Ref. 3. The $BMAP/G/1$ is subsequently considered in Ref. 4, while the first known transient analysis of the $BMAP/G/1$ queue is presented in Ref. 5. Various aspects of the $BMAP/G/1$ continue to be studied, as are queueing variants such as the $D - BMAP/G/1$ and the BMAP retrial queue. See the section titled “Further Reading” for references that pertain to these subjects.

In the sections that follow, we formally define the continuous-time BMAP and provide some basic results including the generation function of its counting process and its fundamental rate. We likewise define the discrete-time batch Markovian process (D-BMAP) and describe a variety of arrival processes that are special cases of the BMAP and D-BMAP. Finally, we will provide suggestions for further reading on the subject for the interested reader to gain a deeper understanding of these versatile arrival processes.

THE CONTINUOUS-TIME BMAP

Let $J \equiv \{J(t) : t \geq 0\}$ be an irreducible, continuous-time Markov chain (CTMC) with state space $E = \{1, 2, \dots, m\}$, where m is a finite, positive integer. The infinitesimal generator matrix of this CTMC is denoted by $\mathbf{Q}$. Suppose J has just entered state $i \in E$. The process spends an exponentially distributed amount of time in state i with rate $\lambda_i = -q_{ii}$, where q_{ii} is the i th diagonal element of $\mathbf{Q}$. The transition that follows this sojourn can be one of two types. For the first type, an “arrival” of batch size k ($k \geq 1$) occurs, and the process transitions to state $j \in E$ with probability $p_{ij}(k)$, where j may be equal to i . For the second type, the batch size is 0 and the process transitions to state $j \neq i$ with probability $p_{ij}(0)$. For each $i \in E$, the probabilities $p_{ij}(k)$ satisfy

$$\sum_{k=1}^{\infty} \sum_{j=1}^m p_{ij}(k) + \sum_{j \in E \setminus \{i\}} p_{ij}(0) = 1. \quad (1)$$

Next, for $k \geq 0$, define the matrices $\mathbf{D}_k = [d_{ij}(k)]_{i,j \in E}$, where

$$d_{ij}(0) = \begin{cases} -\lambda_i, & j = i; \\ \lambda_i p_{ij}(0), & j \neq i; \end{cases} \quad (2)$$

and

$$d_{ij}(k) = \lambda_i p_{ij}(k), \quad i, j \in E, \quad k \geq 1. \quad (3)$$

The matrix $\mathbf{D}_0$ contains the transition rates of J for which no arrivals occur, and the matrices $\{\mathbf{D}_k : k \geq 1\}$ contain the transition rates for which a batch size k occurs. Assuming $\mathbf{D}_0$ is a stable matrix (i.e., it is nonsingular), then the interarrival times will be finite almost surely, which is equivalent to stating that the BMAP will not terminate. From condition (1) and Equations (2) and (3), it is not hard to see that

$$\mathbf{Q} = \sum_{k=0}^{\infty} \mathbf{D}_k.$$

Now, let $N(t)$ denote the total number of arrivals up to time t . The joint process, $(N, J) \equiv \{(N(t), J(t)) : t \geq 0\}$, is called a BMAP. Obviously, it is a Markov process with state space $\{(n, j) : n \geq 0, j \in E\}$ and infinitesimal generator matrix

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \dots \\ 0 & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \dots \\ 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & \dots \\ 0 & 0 & 0 & \mathbf{D}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

In the context of a BMAP, $\{J(t) : t \geq 0\}$ is normally called the *phase process* and $\{N(t) : t \geq 0\}$ is the *counting process*. The matrices, $\{\mathbf{D}_k : k \geq 0\}$, are said to form a *representation* of the BMAP; that is, the BMAP is completely specified by these matrices.

Let us now consider the joint probability distribution of $(N(t), J(t))$ via its (probability) generating function. Adopting the notation of

Lucantoni [5], denote the transition functions of (N, J) by

$$\begin{aligned} P_{ij}(n, t) &= \mathbb{P}(N(t) = n, \\ J(t) = j \mid N(0) = 0, J(0) = i), \end{aligned}$$

and define the matrix $\mathbf{P}(n, t) = [P_{ij}(n, t)]_{i,j \in E}$. Then, for each $n \geq 0$ and $t \geq 0$, $\mathbf{P}(n, t)$ satisfies the Chapman–Kolmogorov equations

$$\begin{aligned} \frac{d}{dt} \mathbf{P}(n, t) &= \sum_{r=0}^n \mathbf{P}(r, t) \mathbf{D}_{n-r}, \\ \mathbf{P}(0, 0) &= \mathbf{I}, \end{aligned} \quad (4)$$

where $\mathbf{I}$ is the identity matrix of order m . Define the matrix-generating function of $\mathbf{P}(n, t)$ by

$$\mathbf{P}^*(z, t) = \sum_{n=0}^{\infty} \mathbf{P}(n, t) z^n, \quad |z| \leq 1, t \geq 0. \quad (5)$$

Differentiating both sides of Equation (5) with respect to t , substituting Equation (4), and summing shows that

$$\begin{aligned} \frac{d}{dt} \mathbf{P}^*(z, t) &= \mathbf{P}^*(z, t) \mathbf{D}(z), \quad t \geq 0, \\ \mathbf{P}^*(z, 0) &= \mathbf{I}, \end{aligned} \quad (6)$$

where

$$\mathbf{D}(z) \equiv \sum_{k=0}^{\infty} \mathbf{D}_k z^k, \quad |z| \leq 1. \quad (7)$$

The (ordinary) matrix differential equation, Equation (6), has the obvious solution

$$\begin{aligned} \mathbf{P}^*(z, t) &= \mathbf{P}^*(z, 0) \exp(\mathbf{D}(z)t) = \exp(\mathbf{D}(z)t), \\ |z| &\leq 1, t \geq 0, \end{aligned}$$

where $\exp(A)$ is the matrix exponential of a square matrix A defined by

$$\exp(A) = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

We pause here to note the similarity between the generating function of the BMAP and

that of a standard Poisson process, which is given by the scalar function

$$P^*(z, t) = \exp((-\lambda + \lambda z)t).$$

For the BMAP, the exponential term $-\lambda + \lambda z$ is replaced by the matrix $\mathbf{D}(z)$ to account for batch sizes larger than unity.

Using the generating function of $\mathbf{P}(n, t)$, one can obtain the (conditional) expectation of the number of arrivals in the interval $(0, t]$. Define this conditional expectation by $\mathbb{E}_i(N(t)\mathbf{1}(J(t) = j))$, where $\mathbf{1}(B)$ denotes the indicator variable of event B and $\mathbb{E}_i$ denotes expectation with respect to $\mathbb{P}_i$, the probability law of (N, J) given $J(0) = i$, and $N(0) = 0$. Then, $\mathbb{E}_i(N(t)\mathbf{1}(J(t) = j))$ is the (i, j) th entry of the $m \times m$ matrix

$$\left. \frac{d}{dz} \mathbf{P}^*(z, t) \right|_{z=1} = \mathbf{D}(1) \exp(\mathbf{D}(1)t) = \mathbf{Q} \exp(\mathbf{Q}t).$$

The conditional k th-factorial moment can be obtained by taking the k th-order derivative $\mathbf{P}^*(z, t)$ and evaluating at $z = 1$ in the usual way.

The limiting behavior of the continuous-time BMAP is discussed next. Let $\boldsymbol{\pi} = [\pi_1, \dots, \pi_m]$ be the invariant probability vector of the CTMC, $\{J(t) : t \geq 0\}$, with generator matrix $\mathbf{Q}$; that is, $\boldsymbol{\pi}$ is the unique positive solution to the system of equations

$$\boldsymbol{\pi} \mathbf{Q} = \mathbf{0} \quad \text{and} \quad \boldsymbol{\pi} \mathbf{e} = 1,$$

where $\mathbf{0}$ is the zero (row) vector and $\mathbf{e}$ is a (column) vector of ones. Then the *fundamental rate*, or the stationary rate of arrivals in a BMAP, is given by

$$\lambda = \boldsymbol{\pi} \left(\sum_{k=1}^{\infty} k \mathbf{D}_k \right) \mathbf{e}. \quad (8)$$

On the other hand, the arrival rate of *batches* is given by

$$\lambda_g = -\boldsymbol{\pi} \mathbf{D}_0 \mathbf{e},$$

which is never zero since $\mathbf{D}_0$ is assumed to be nonsingular. If all the batch sizes are equal to unity, then the process is a Markovian arrival process (MAP), and $\lambda = \lambda_g$.

In the next section, we describe several arrival processes that are special cases of the BMAP. A working knowledge of PH-distributions is assumed for this discussion. For a thorough treatment of continuous- and discrete-time PH-distributions, the reader should consult Refs 1, 6, 7, 10. A cogent summary of PH-distributions is also provided in **Phase-Type (PH) Distributions**.

COMMON CONTINUOUS-TIME BMAPS

1. *Poisson Process*. If the state space E consists of only a single state (i.e., $m = 1$), the time between “transitions” is exponentially distributed with rate λ , and an arrival of batch size 1 occurs at each transition, then the counting process $\{N(t) : t \geq 0\}$, is a Poisson process with rate λ . In this case, the matrices $\{D_k : k \geq 0\}$ are replaced by scalars $\{D_k : k \geq 0\}$. Specifically, $D_0 = -\lambda$, $D_1 = \lambda$, and $D_k = 0$ for all $k \geq 2$. Then, $\{N(t) : t \geq 0\}$ is a BMAP with generator matrix

$$Q^* = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & 0 & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & 0 & 0 & \dots \\ 0 & 0 & 0 & -\lambda & \lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

2. *Batch Poisson Process*. If we allow a batch size greater than unity in the standard Poisson process with rate λ , the resulting *batch Poisson process* is a BMAP. Let p_k denote the probability that an arrival is of batch size k , $k \geq 1$, and note that $\sum_{k \geq 1} p_k = 1$. For this process, $m = 1$, $D_0 = -\lambda$, and $D_k = \lambda p_k$ for each $k \geq 1$. Then the batch Poisson Process is a BMAP with generator matrix

$$Q^* = \begin{bmatrix} -\lambda & p_1\lambda & p_2\lambda & p_3\lambda & p_4\lambda & p_5\lambda & \dots \\ 0 & -\lambda & p_1\lambda & p_2\lambda & p_3\lambda & p_4\lambda & \dots \\ 0 & 0 & -\lambda & p_1\lambda & p_2\lambda & p_3\lambda & \dots \\ 0 & 0 & 0 & -\lambda & p_1\lambda & p_2\lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Moreover, as noted by Lucantoni [5], if $g(z)$ is the generating function of $\{p_k : k \geq 1\}$, then $D(z) = -\lambda + \lambda g(z)$, $|z| \leq 1$.

3. *Batch MMPP*. Consider a Poisson process whose rate is modulated by an exogenous, irreducible Markov process, $\{J(t) : t \geq 0\}$, with state space $\{1, 2, \dots, m\}$ and generator matrix Q . Whenever $J(t) = i$, arrivals are according to a Poisson process with rate λ_i ($\lambda_i > 0$). Define the vector $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ and let $\Delta(\lambda) = \text{diag}(\lambda)$. Arrivals occur in batches of size k with probability p_k , $k \geq 1$. If $N(t)$ denotes the number of arrivals up to time t , then $\{(N(t), J(t)) : t \geq 0\}$ is a BMAP with $D_0 = Q - \Delta(\lambda)$, $D_k = p_k \Delta(\lambda)$ for $k \geq 1$. (An excellent summary of the standard MMPP is provided by Fischer and Meier-Hellstern [11].)
4. *Batch PH-Renewal Process*. Suppose that arrivals are according to a renewal process, $\{\tau_n : n \geq 0\}$, where τ_n denotes the n th arrival epoch. The PH-renewal process is a renewal process for which the interrenewal times $S_n \equiv \tau_{n+1} - \tau_n$, $n \geq 0$, form an i.i.d. sequence of PH-distributed random variables with representation (α, T) , where T is of order m . Again, let p_k denote the probability that the batch size is k , $k \geq 1$. The batch PH-renewal process is then a BMAP with $D_0 = T$ and $D_k = p_k T^0 \alpha$, $k \geq 1$ where $T^0 = -T e$.
5. *Superposition of Independent BMAPs*. The superposition of N independent BMAPs is again a BMAP. Let $\{D_k(i) : k \geq 0\}$, $i = 1, 2, \dots, N$, denote a collection of N independent BMAPs such that the phase process of the i th BMAP is of order $m(i)$. Let

$$M = \prod_{i=1}^N m(i),$$

and define for each $k \geq 0$ the $M \times M$ matrix D_k by

$$D_k = D_k(1) \oplus \dots \oplus D_k(N),$$

where the operator $\oplus$ is the *Kronecker matrix sum* [8,11]. Then $\{D_k : k \geq 0\}$ is the representation of the superposition of the N independent BMAPs.

THE DISCRETE-TIME BMAP (D-BMAP)

Consider an irreducible discrete-time Markov chain (DTMC) $J \equiv \{J_r : r \geq 0\}$ on the state space $E = \{1, 2, \dots, m\}$ which allows self-transitions. Suppose that the n th transition of J triggers the arrival of a batch of customers of size Y_n ($Y_n \geq 0$), with $Y_0 = 0$. Define the conditional probabilities

$$q_{ij}(k) = \mathbb{P}(X_{n+1} = j, Y_n = k | X_n = i),$$

$$i, j \in E, n \geq 0,$$

which are the joint probabilities of a transition of the discrete-time chain J from i to j and an arrival of batch size $k \geq 0$. Next, define the (substochastic) matrices $\mathbf{D}_k = [q_{ij}(k)]_{i,j \in E}$ and assume that $\mathbf{I} - \mathbf{D}_0$ is nonsingular to ensure that the arrival of one or more customers occurs with probability one. The transition probability matrix of J , denoted by $\mathbf{P}$, is given by

$$\mathbf{P} = \sum_{k=0}^{\infty} \mathbf{D}_k,$$

whose entries are necessarily finite. The matrices $\{\mathbf{D}_k : k \geq 0\}$ completely specify a D-BMAP.

As for the continuous-time BMAP, it is possible to construct a bivariate Markov chain representation of the D-BMAP. For $r \geq 0$, let N_r be the total number of arrivals up to, and including, the r th transition of J . The process $\{N_r : r \geq 0\}$ is the *counting process* of the D-BMAP, which together with the phase process J , allows us to define the bivariate process

$$\{(N_r, J_r) : r \geq 0\},$$

which is a two-dimensional DTMC with transition probability matrix $\mathbf{P}^*$ given by

$$\mathbf{P}^* = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \dots \\ 0 & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \dots \\ 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & \dots \\ 0 & 0 & 0 & \mathbf{D}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Next, we elucidate an important property connected to the counting process as documented in Ref. 12. Define the r -step transition matrix $\mathbf{P}(n, r)$ whose (i, j) th entry is defined as

$$[\mathbf{P}(n, r)]_{ij} = \mathbb{P}(N_r = n, J_r = j | J_0 = i),$$

$$r \geq 1, n \geq 0.$$

The matrix generating function of $\mathbf{P}(n, r)$, denoted by $\mathbf{P}^*(z, r)$, is given by

$$\mathbf{P}^*(z, r) = \sum_{n=0}^{\infty} \mathbf{P}(n, r) z^n, \quad |z| \leq 1.$$

It can be shown that, for $r \geq 1$,

$$\mathbf{P}^*(z, r) = [\mathbf{P}^*(z, 1)]^r = [\mathbf{D}(z)]^r, \quad |z| \leq 1, \quad (9)$$

where $\mathbf{D}(z) = \sum_{k=0}^{\infty} \mathbf{D}_k z^k$. If $\mathbf{D}(z)$ is known explicitly, then the k th-factorial moments of the number of arrivals in r ($r \geq 1$) transitions can be obtained by computing the k th-order derivative of $\mathbf{P}^*(z, r)$ and evaluating at $z = 1$.

Next, we consider the fundamental rate of the D-BMAP. Let $\boldsymbol{\pi} = [\pi_1, \dots, \pi_m]$ be the invariant probability vector of the phase process J ; that is, $\boldsymbol{\pi}$ is the unique positive solution to the system of equations

$$\boldsymbol{\pi} \mathbf{P} = \boldsymbol{\pi} \text{ and } \boldsymbol{\pi} \mathbf{e} = 1.$$

Then the *fundamental* or stationary arrival rate of the D-BMAP is given by

$$\lambda = \boldsymbol{\pi} \left(\sum_{k=1}^{\infty} k \mathbf{D}_k \right) \mathbf{e}.$$

Here, the stationary batch arrival rate may be computed as

$$\lambda_g = \boldsymbol{\pi} (\mathbf{I} - \mathbf{D}_0) \mathbf{e},$$

which is always nonzero due to the assumption that $\mathbf{I} - \mathbf{D}_0$ is nonsingular. As before, we have $\lambda = \lambda_g$ if the maximum possible batch size is one, as in a discrete-time MAP.

In the next section, we discuss a few common D-BMAPs and point the reader toward more extensive references on the subject.

COMMON DISCRETE-TIME BMAPS

1. *Batch Geometric Process.* Arrivals here are considered to be a sequence of independent trials for which the “success” probability p_0 ($0 < p_0 < 1$) corresponds to a batch size of zero. This process is a D-BMAP with $m = 1$, $D_0 = p_0$, and $D_k = p_k(1 - p_0)$, where $\{p_k : k \geq 1\}$ are the batch-size probabilities conditioned upon the arrival of a batch of size $k \geq 1$. For the single-arrival process, we note that $p_1 = 1$ and $p_k = 0$ for $k \geq 2$, thus giving $D_0 = p_0$, $D_1 = 1 - p_0$, and $D_k = 0$ for $k \geq 2$.
2. *Batch Markov-Modulated Bernoulli Process.* The Markov-Modulated Bernoulli Process (MMBP) is the discrete-time analog of the MMPP. For both, the single-arrival and batch versions of the MMBP, arrivals are triggered by the transitions of an m -state DTMC with transition probability matrix $\mathbf{P}$. If the process ends up in state $j \in \{1, \dots, m\}$, then the probability of an arrival is given by $\eta_j \in (0, 1]$, while the probability of a null arrival is given by $1 - \eta_j$. For notational convenience, define the vector

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_m).$$

As with the batch geometric process, the conditional probabilities of batch size are given by the sequence $\{p_k : k \geq 1\}$, with the usual adjustments made for the single-arrival version. The batch MMBP may be expressed as a D-BMAP with elements $\mathbf{D}_0 = \Delta(\mathbf{e} - \boldsymbol{\eta})\mathbf{P}$ and $\mathbf{D}_k = p_k \Delta(\boldsymbol{\eta})\mathbf{P}$ for $k \geq 1$.

3. *Batch PH-Renewal Process.* As before, we consider a renewal process whose renewal epochs are the set of points $\{\tau_n : n \geq 0\}$ with interrenewal times $S_{n+1} \equiv \tau_{n+1} - \tau_n$, $n \geq 0$. Here, the i.i.d. sequence of random variables, $\{S_n : n \geq 1\}$, has a discrete PH distribution with representation $(\boldsymbol{\alpha}, \mathbf{T})$ with $\mathbf{T}$ of order m . The process is then a D-BMAP with $\mathbf{D}_0 = \mathbf{T}$ and $\mathbf{D}_k = p_k \mathbf{T}^0 \boldsymbol{\alpha}$, $k \geq 1$, where $\mathbf{T}^0 = -\mathbf{T}\mathbf{e}$ and $\boldsymbol{\alpha}$ is the vector

of initial probabilities for the discrete PH-distribution.

FURTHER READING

The original formulation of the continuous-time BMAP was the VMPP introduced by Neuts [1] in 1979. The current form of the MAP was developed in Ref. 3 as an arrival process to a single-server queueing system, and the extension to the BMAP is detailed in Ref. 5. The D-BMAP first appeared in the works of Blondia [12,13] and has since pervaded the queueing and computer and communications networking literature, just like its continuous-time predecessor. An excellent summary of the BMAP and D-BMAP, along with examples and selected applications, can be found in Ref. 6.

The analysis of queueing systems with BMAP (or related) input processes has received considerable attention in the stochastic modeling community. Specific examples of single-server queueing models with MAP input can be found in Refs 14–18. Machihara [19] examined single-server queues with batch arrivals and state-dependent service times, while Hofmann [20] considered state-dependent batch arrival rates. Krieger *et al.* [21] studied a Markov-modulated $BMAP/G/1$ queue. Queues with BMAP input and server vacations have received much attention, beginning with Ref. 3. Lucantoni [5] provided a nice summary of a number of important results for the $BMAP/G/1$ system. A sampling of the ensuing literature available on the subject can be found in Refs 22–26. Chydzinsky [27] provided a transient analysis of the $MMPP/G/1/k$ loss system, and in Ref. 28, analyzed the first passage time to buffer overflow in the $BMAP/G/1/k$ queue. The processor-sharing queueing discipline in systems with MAP or BMAP input has been studied extensively in Refs 29–31.

Another fruitful area of research has considered queueing systems with BMAP input and retrials. Retrial queueing systems are extensively used to model systems in which customers retry service after encountering a busy (or failed) server. For example,

they are extremely useful for modeling the retransmission of data packets in computer and communications networks, or the call-back behavior of customers in a customer contact center. Some early examples of the BMAP in retrial queues include Refs 32–35. Chakravarthy and Dudin [36,37] introduced single- and multi-server retrial models with group service and exponential retrials. Breuer *et al.* [38] studied a $BMAP/PH/n$ multiserver retrial system, while Li *et al.* [39] introduced the complication of an unreliable server in the BMAP retrial model.

The BMAP has been applied extensively in a number of areas from inventory management [40] to maintenance models [41]. However, the preponderance of applications lies in computer and communications networking, with the bulk of these assuming D-BMAP input processes. The D-BMAP is often used to model specific characteristics of source signals, such as burstiness, in telecommunications systems. Blondia [13] introduced the D-BMAP model, and in Ref. 12, analyzed the steady-state system size distribution of a $D-BMAP/G/1/N$ queue. Van Houdt and Blondia [42] used a D-BMAP to model packet arrivals in a centralized wireless local area network. In Ref. 43, they examined contention resolution in a network among many users generating signals modeled as a D-BMAP. Zhao *et al.* [44] extended the work of Blondia and Casals [45] on the $D-BMAP/PH/1/N$ queue to the case of prioritized service. Queues with PH-distributed service times have proven useful in modeling video streaming over networks, and the addition of a prioritization scheme enhances the usefulness of these models in networks with heterogeneous data streams.

REFERENCES

1. Neuts MF. A versatile Markovian point process. *J App Probab* 1979;16(4):764–779.
2. Neuts MF. Matrix-analytic methods in queueing theory. *Eur J Oper Res* 1984;15:2–12.
3. Lucantoni DM, Meier-Hellstern KS, Neuts MF. A single-server queue with server vacations and a class of non-renewal arrival processes. *Adv Appl Probab* 1990;22(3):676–705.
4. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Commun Stat Stoch Models* 1991;7(1):1–46.
5. Lucantoni DM. The $BMAP/G/1$ queue: a tutorial. In: Donatiello L, Nelson R, editors. *Models and techniques for performance evaluation of computer and communications systems*. London: Springer; 1993. pp. 330–358.
6. Chakravarthy SR. The batch Markovian arrival process: A review and future work. In: Krishnamoorthy A, Raju N, Ramaswami V, editors. *Advances in probability theory and stochastic processes*. New Jersey: Notable Publications; 2000. pp. 21–39.
7. Neuts MF. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. New York: Dover Publications, Inc.; 1981.
8. Neuts MF. *Structured stochastic matrices of $M/G/1$ type and their applications*. New York: Marcel Dekker, Inc.; 1989.
9. Ramaswami V. The $N/G/1$ queue and its detailed analysis. *J App Probab* 1980;12: 222–261.
10. Latouche G, Ramaswami V. *Introduction to matrix-analytic methods in stochastic modeling*. American Statistical Association and the Society for Industrial and Applied Mathematics (SIAM). Alexandria (VA), Philadelphia (PA); 1999.
11. Fischer W, Meier-Hellstern K. The Markov-modulated Poisson process (MMPP) cookbook. *Perform Eval* 1992;18:149–171.
12. Blondia C. A discrete-time batch Markovian arrival process as B-ISDN traffic model. *Belg J Oper Res Stat Comput Sci* 1993;32:3–23.
13. Blondia C. A discrete-time Markovian arrival process. RACE Document PRLB-123-0015-CD-CC. August 1989.
14. Subramanian V, Srikant R. Tail probabilities of low-priority waiting times and queue lengths in $MAP/GI/1$ queues. *Queueing Syst* 2000;34(1–4):215–236.
15. Shioda S. Departure process of the $MAP/SM/1$ queue. *Queueing Syst* 2003;44(1): 31–50.
16. Adan IJ, Kulkarni VG. Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Syst* 2003;45(2): 113–134.
17. Li QL, Zhao YQ. A $MAP/GI/1$ queue with negative customers. *Queueing Syst* 2004;47(1–2): 5–43.
18. Lee HW, Cheon SH, Lee EY, *et al.* Workload and waiting time analyses of $MAP/GI/1$ queue under D-policy. *Queueing Syst* 2004;48:421–443.

19. Machihara F. A *BMAP/SM/1* queue with service times depending on the arrival process. *Queueing Syst* 1999;33(4):277–291.
20. Hofmann J. The *BMAP/G/1* queue with level-dependent arrivals — an overview. *Telecommun Syst* 2001;16(3–4):347–359.
21. Krieger U, Klimenok VI, Kazimirsky AV, *et al.* A *BMAP/PH/1* queue with feedback operating in a random environment. *Math Comput Model* 2005;41(8–9):867–882.
22. Shin YW, Pearce CEM. The *BMAP/G/1* vacation queue with queue-length dependent vacation schedule. *J Aust Math Soc Ser B* 1998;40:207–221.
23. Ferng HW. Departure processes of *BMAP/G/1* queues. *Queueing Syst* 2001;39(2–3):109–135.
24. Chang SH, Takine T, Chae KC, *et al.* A unified queue length formula for *BMAP/G/1* queue with generalized vacations. *Stoch Models* 2002;18(3):369–386.
25. Lee HW, Park NI. Using factorization for waiting times in *BMAP/G/1* queues with N-policy and vacations. *Stoch Anal Appl* 2004;22(3):755–773.
26. Baek JW, Lee HW, Lee SW, *et al.* A factorization property for *BMAP/G/1* vacation queues under variable service speed. *Ann Oper Res* 2008;160:19–29.
27. Chydzinski A. Transient analysis of the *MMPP/G/1/K* queue. *Telecommun Syst* 2006;32:247–262. DOI: 10.1007/s11235-006-9001-5.
28. Chydzinski A. Time to reach buffer capacity in a BMAP queue. *Stoch Models* 2006;23:195–209. DOI: 10.1080/15326340701300746.
29. D’Apice C, Manzo R, Pechinkin AV. A finite *MAP_k/G_k/1* queueing system with generalized foreground-background processor-sharing discipline. *Automat Remote Control* 2004;65(11):1793–1799.
30. Li QL, Lian Z, Liu L. An *RG*-factorization approach for a *BMAP/M/1* generalized processor-sharing [3] queue 2005;21:507–530. DOI: 10.1081/STM-200056223.
31. D’Apice C, Manzo R, Pechinkin AV. A finite capacity *BMAP_k/G_k/1* queueing system with generalized foreground-background processor-sharing discipline. *Automat Remote Control* 2006;67(3):428–434.
32. Choi BD, Chang Y. *MAP₁, MAP₂/M/c* retrial queue with the retrial group of finite capacity and geometric loss. *Math Comput Model* 1999;30(3–4):99–113. DOI: 10.1016/S0895-7177(99)00135-1.
33. Dudin A, Klimenok V. A retrial *BMAP/SM/1* system with linear repeated requests. *Queueing Syst* 2000;34:47–66.
34. Choi BD, Chung YH, Dudin AN. The *BMAP/SM/1* retrial queue with controllable operation modes. *Eur J Oper Res* 2001;131(1):16–30.
35. Klimenok VI. A multiserver retrial queueing system with batch Markov arrival process. *Automat Remote Control* 2001;62(8):1312–1322.
36. Chakravarthy SR, Dudin AN. A single server retrial queueing model with batch arrivals and group services. In: Artalejo JR, Krishnamoorthy A, editors. *Advances in stochastic modeling*. New Jersey: Notable Publications; 2002. pp. 1–21.
37. Chakravarthy SR, Dudin AN. A multi-server retrial queue with *BMAP* arrivals and group services. *Queueing Syst* 2002;42:5–31.
38. Breuer L, Dudin A, Klimenok V. A retrial *BMAP/PH/N* system. *Queueing Syst* 2002;40:433–457.
39. Li Q, Ying Y, Zhao YQ. A *BMAP/G/1* retrial queue with a server subject to breakdowns and repairs. *Ann Oper Res* 2006;141:233–270. DOI: 10.1007/s10479-006-5301-0.
40. Yadavalli VSS, Sivakumar B, Arivarignan G. Stochastic inventory management at a service facility with a set of reorder levels. *ORiON (J Oper Res Soc S Afr ORSSA)* 2007;23(2):137–149.
41. Lee HW, Park NI. A maintenance model for manufacturing lead time in a production system with BMAP input and bilevel setup control. *Asia Pac J Oper Res* 2008;25(6):807–825.
42. Van Houdt B, Blondia C. Robustness properties of FS-ALOHA++: a contention resolution algorithm for dynamic bandwidth allocation. *Mob Netw Appl* 2003;8:237–253.
43. Van Houdt B, Blondia C. Throughput of q-ary splitting algorithms for contention resolution in communication networks. *Commun Inf Syst* 2005;4(2):135–164.
44. Zhao J, Li B, Cao X, *et al.* A matrix-analytic solution for the *DBMAP/PH/1* priority queue. *Queueing Syst* 2006;53:127–145. DOI: 10.1007/s11134-006-8306-0.
45. Blondia C, Casals O. Statistical multiplexing of VBR sources: a matrix-analytic approach. *Perform Eval* 1992;16(1–3):5–20.

BAYESIAN AGGREGATION OF EXPERTS' FORECASTS

KENNETH C. LICHTENDAHL, JR.
Darden School of Business,
University of Virginia,
Charlottesville, Virginia

INTRODUCTION

A decision maker when faced with an uncertain future often consults multiple experts to help better forecast the future. For example, a supply chain manager may ask several retail managers, who have greater knowledge of consumer demand, to forecast the future demand for a new product. With these forecasts, the supply chain manager has a modeling choice to make: how does he or she aggregate the retail managers' forecasts into a single forecast? This modeling choice is important because the supply chain manager will use the aggregate forecast in the analysis of his or her inventory decision.

When faced with the question of how to aggregate disparate opinions, decision makers routinely average the forecasts reported by experts. The idea that the average forecast produces a good aggregate forecast goes back at least to Galton. In 1906, Galton [1] went to the West of England Fat Stock and Poultry Exhibition and observed that the median of 787 guesses of an ox's weight was within nine pounds of the ox's actual weight. (See Surowiecki [2] for more details on Galton's trip to the country fair.) Although Galton [3] advocated for the "middlemost" (or median) estimate as the voice of the group, Hooker [4] asked Galton [5], "... is the median a nearer approximation to the truth than the mean?" Pearson [6] concurred, noting that "the trustworthiness of a democratic judgment is ... more than confirmed, if the material be dealt with by the 'average' method, not the 'middlemost' judgment..." After all, in the case of Galton's weight-judging competition, the

average estimate outperformed the median estimate, coming within one pound of the of the ox's actual weight.

In a decision analysis, however, the usefulness of the average of some best guesses is limited. Such an average produces an aggregate point estimate, when the decision maker is in need of a full probability distribution. Without a complete distribution, the decision maker risks committing the "flaw of averages"—a version of Jensen's Inequality [7]. For instance, the supply chain manager, without a complete distribution for demand, may mistake the profit at the average demand forecast for the average profit.

A common approach that addresses the limitation of the aggregate point estimate uses the experts' point estimates themselves to estimate the mean and standard deviation of a normally distributed quantity of interest. In a supply chain problem, Fisher and Raman [8] use this approach to estimate demand uncertainty from several experts' point estimates. Their aggregate forecast for demand is a normal distribution with a mean equal to the average of experts' point estimates and a standard deviation proportional to the sample standard deviation of the experts' point estimates. With this approach, when experts rely heavily on a common information source to form their individual beliefs, the aggregate forecast will typically have a low variance. This will occur even though the common information source may indicate that the future is highly uncertain. In the extreme, all experts look at the same data and report the same point estimate. In this situation, the aggregate forecast in this situation would contain no uncertainty about the future. Hence, there is a theoretical shortcoming in the approach of Fisher and Raman [8]. The decision maker may inadvertently discard valuable information about the degree to which the experts' information sources overlap.

The Bayesian approach, on the other hand, asks a decision maker to explicitly model the experts' forecasts as potentially dependent

data [9,10]. First, the decision maker assesses her prior beliefs for the quantity of interest, and then, conditional on the quantity of interest, she assesses the likelihood of the possible experts' forecasts. If there is an overlap in the information the experts use to form their individual beliefs, then the decision maker incorporates this dependence in the likelihood [10]. We call this likelihood the expert likelihood. It is the conditional distribution of the experts' forecasts given the quantity of interest. Finally, once the decision maker hears the experts' forecasts, the decision maker applies Bayes' theorem to update her prior beliefs for the quantity of interest. The result of this updating process is the aggregate forecast, or posterior distribution. It is the conditional distribution of the quantity of interest, given the experts' forecasts.

Although the Bayesian approach to aggregating individual beliefs is straightforward and theoretically sound, assessing the expert likelihood function can be difficult. The decision maker must assess a separate distribution of the experts' forecasts for each possible state of the quantity of interest. One advantage of the Bayesian approach is that the decision maker can form an aggregate distribution from just about anything the experts might report—from point estimates to quantiles, all the way up to full probability distributions. But, because of the intensive nature of the Bayesian's assessment task, many researchers and practitioners prefer to work with linear opinion pools. The linear opinion pool is intuitively appealing because it involves a simple weighted average of the experts' probability distributions [11]. The linear opinion pool, however, shifts most of the assessment burden onto the experts; each expert must report a full probability distribution. For more information on the Bayesian and opinion pooling approaches to aggregating expert opinion, see Clemen and Winkler [12], Cooke [13], and Genest and Zidek [14].

In the next three sections, we present a series of progressively more complicated examples—a kind of tutorial on Bayesian aggregation. The examples we consider are prototypical. They are meant to illustrate how the theory of Bayesian aggregation

applies in three important forecasting settings: forecasts for discrete events, parametric forecasts for continuous quantities, and nonparametric forecasts for continuous quantities. In the section titled "Aggregation of Probabilities for a Discrete Event," we present several Bayesian aggregation models of experts' probabilities for a discrete event. In the section titled "Aggregation of Parametric Forecasts for a Continuous Quantity," we examine the classic Bayesian aggregation model for experts who report parametric distributions for a continuous quantity. In the section titled "Aggregation of Nonparametric Forecasts for a Continuous Quantity," we provide two Bayesian aggregation models for experts who report nonparametric forecasts—either probability or quantile forecasts—for a continuous quantity. In each setting, we break down the assessment of an expert likelihood into the decision maker's judgments of expert discrimination and dependency. An expert's discrimination, or forecasting skill, arises from the dependency between the event of interest and the expert's forecast; and expert dependency arises from the amount of overlap the decision maker thinks there is in the experts' information sources. Finally, in the section titled "Summary," we summarize these Bayesian approaches with a view toward how these assessments can be made in practice.

AGGREGATION OF PROBABILITIES FOR A DISCRETE EVENT

The simplest Bayesian aggregation model involves a decision maker who updates her prior probability distribution for a discrete event, for example, rain or no rain. She updates her prior beliefs based on what one or more experts report are their beliefs for the same event. In this section, we present four examples that illustrate several important features in this updating process. In the first two examples, we move from one expert reporting on a binary event to two experts reporting on a binary event. In both of these examples, the experts have very coarse information sources from which to draw their opinions. In the second two examples,

we move from two experts reporting on a binary event to multiple experts reporting on a general discrete event. In these second two examples, the experts have more refined sources of information from which to draw their opinions.

Model for a Binary Event and One Expert

Suppose a decision maker is planning an outdoor festival in a new city three days from now. The event of interest x_3 is the weather on the day of the festival: either rain ($x_3 = 1$) or no rain ($x_3 = 0$). The decision maker has her prior beliefs $p_0 = \Pr_0(x_3 = 1)$ about rain on the day of the festival and considers sending an observer to the new city for the next two days. When the observer returns from this trip at the end of two days, he will report his beliefs $p_1 = \Pr_1(x_3 = 1)$ about rain on the day of the festival. How should the decision maker update her beliefs about rain on the day of the festival once she hears the observer's report?

To answer this question, suppose the decision maker and observer both believe rain on each of the next three days x_i is identically and independently distributed (iid) according to a Bernoulli distribution $\text{Br}(\theta)$ with parameter $0 < \theta < 1$. The Bernoulli distribution has probability mass function $\Pr_{\text{Br}}(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$. In addition, both the decision maker and observer are Bayesian and share the belief that the Bernoulli distribution's parameter θ is distributed according to a beta distribution $\text{Be}(\alpha, \beta)$ with hyperparameters $0 < \alpha, \beta < 1$. The beta distribution has probability density function $f_{\text{Be}}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ where Γ is the gamma function. We represent the inference model for this Bernoulli process with the following hierarchy:

$$\begin{aligned}\theta &\sim \text{Be}(\alpha, \beta) \\ x_i|\theta &\sim_{\text{iid}} \text{Br}(\theta), \quad i = 1, \dots, n,\end{aligned}$$

where “ $\sim$ ” means *is distributed according to*.

To these two Bayesians, the distribution of $(x_{n+1}|x_n, \dots, x_1)$ becomes the beta-binomial distribution $\text{Bb}(\alpha + r, \beta + n - r, 1)$ where $r = \sum_{i=1}^n x_i$ is a sufficient statistic for $(x_n, \dots, x_1)$.

The beta-binomial distribution has the probability mass function

$$\Pr_{\text{Bb}}(x|a, b, c) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+c)} \binom{c}{x} \times \Gamma(a+x)\Gamma(b+c-x).$$

The distribution of any x_i alone is the beta-binomial distribution $\text{Bb}(\alpha, \beta, 1)$. (For more information on this and other Bayesian inference models, see Bernardo and Smith [15].)

If both the decision maker and observer further agree that $(\alpha, \beta) = (1, 1)$, then the decision maker has $p_0 = \Pr_{\text{Bb}}(x_3 = 1|1, 1, 1) = \frac{1}{2}$, and the observer will report $p_1 = \Pr(x_3 = 1|x_2, x_1) = \Pr_{\text{Bb}}(x_3 = 1|1+r, 3-r, 1)$. In Table 1, we see the three possible ways that the observer might report, depending on the outcomes for rain on the first two days.

To update the decision maker's prior beliefs based on the observer's report, the decision maker uses Bayes' theorem to find the conditional distribution of $(x_3|p_1)$ as follows:

$$\Pr(x_3|p_1) = \frac{\Pr(x_3)\Pr(p_1|x_3)}{\Pr(p_1)}.$$

She deduces the distribution of $(p_1|x_3)$ from her beliefs about $(x_1, x_2|x_3)$. This deduction requires a thought experiment—if the decision maker alone knew the outcome on day three what would her beliefs about rain on the first two days be? The distribution of $(x_1, x_2|x_3)$ is given by

$$\begin{aligned}\Pr(x_1, x_2|x_3) &= \frac{\Pr(x_1, x_2, x_3)}{\Pr(x_3)} \\ &= \frac{2\Gamma(1+x_1+x_2+x_3)\Gamma(4-x_1-x_2-x_3)}{\Gamma(5)},\end{aligned}$$

Table 1. Rain Observations and One Observer's Reports

Day 1 Outcome (x_1)	Day 2 Outcome (x_2)	Sufficient Statistic ($r = \sum_{i=1}^2 x_i$)	Observer's Report (p_1)
0	0	0	0.25
0	1	1	0.50
1	0	1	0.50
1	1	2	0.75

Table 2. Rain Observations, Distribution of $(x_1, x_2 | x_3)$, and One Observer's Reports

Day 1 Outcome (x_1)	Day 2 Outcome (x_2)	Day 3 Outcome (x_3)	$\Pr(x_1, x_2 x_3)$	Observer's Report (p_1)
0	0	0	1/6	0.25
0	1	0	1/6	0.50
1	0	0	1/6	0.50
1	1	0	3/6	0.75
0	0	1	3/6	0.25
0	1	1	1/6	0.50
1	0	1	1/6	0.50
1	1	1	1/6	0.75

where the second equality follows from substitution according to $\Pr(x_3) = \frac{1}{2}$ and

$$\Pr(x_1, x_2, x_3) = \int_0^1 f_{\text{Be}}(\theta | 1, 1) \theta^{x_1} (1 - \theta)^{1-x_1} \theta^{x_2} \times (1 - \theta)^{1-x_2} \theta^{x_3} (1 - \theta)^{1-x_3} d\theta.$$

The distribution of $(x_1, x_2 | x_3)$, which is listed in Table 2, leads to the two conditional distributions in Fig. 1 that make up the expert likelihood $\Pr(p_1 | x_3)$.

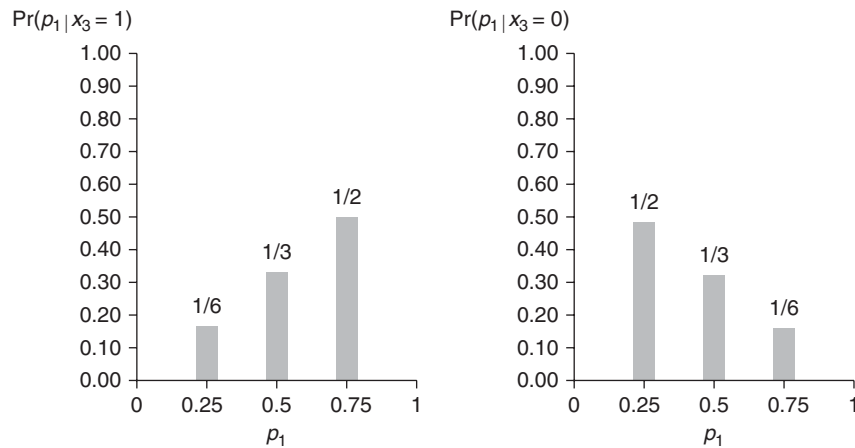
The conditional distributions depicted in Fig. 1 look as we would expect them to look. If it is going to rain on the third day, then the decision maker believes it is more likely the expert will report a high p_1 . If it is not going to rain on the third day, then the decision maker believes it is more likely the expert will report a low p_1 . This feature of the expert likelihood is called *discrimination*

[16]. The less similar the expert likelihood's two conditional distributions are, the more discriminating or skilled, we say the expert is. Another way to say this is: the more correlated (either negatively or positively) the quantity of interest and the expert's report are, the more discriminating the expert is.

Finally, using the fact that the distribution of p_1 is $\Pr(p_1) = \Pr(x_3 = 1) \Pr(p_1 | x_3 = 1) + \Pr(x_3 = 0) \Pr(p_1 | x_3 = 0)$, the decision maker updates her beliefs to find, not surprisingly, that the aggregate distribution is $\Pr(x_3 | p_1) = p_1$. It is as if the decision maker observed the first two days of the weather herself.

Model for a Binary Event and Two Overlapping Experts

To incorporate the idea that the information experts access may overlap, we consider the

**Figure 1.** Expert likelihood for one observer.

situation in the section titled “Model for a Binary Event and One Expert” with a second observer who visits the new city on the second and third days. The decision maker is now planning the outdoor event in the new city on the fourth day. When the two observers return from their trips—the first observer at the end of two days and the second observer at the end of three days—they will report their beliefs p_1 and p_2 about rain on the fourth day. How should the decision maker update her beliefs about rain on the fourth day once she hears the observers’ report?

To answer this question, we suppose the decision maker and the observers agree on the inference model for the Bernoulli process in the section titled “Model for a Binary Event and One Expert” with the same hyperparameters $(\alpha, \beta) = (1, 1)$. Then, the decision maker has $p_0 = \Pr(x_4 = 1) = \Pr_{\text{Bb}}(x_4 = 1 | 1, 1, 1) = \frac{1}{2}$. Observer 1 will report $p_1 = \Pr(x_4 = 1 | x_2, x_1) = \Pr_{\text{Bb}}(x_4 = 1 | 1 + r_1, 3 - r_1, 1)$ where $r_1 = \sum_{i=1}^2 x_i$. Observer 2 will report $p_2 = \Pr(x_4 = 1 | x_3, x_2) = \Pr_{\text{Bb}}(x_4 = 1 | 1 + r_2, 3 - r_2, 1)$ where $r_2 = \sum_{i=2}^3 x_i$. In Table 3, we see the eight possible ways the observers might report, depending on the outcomes for rain on the first three days.

To update the decision maker’s prior beliefs based on the observers’ reports, the decision maker uses Bayes’ theorem to find the conditional distribution of $(x_4 | p_1, p_2)$ as follows:

$$\Pr(x_4 | p_1, p_2) = \frac{\Pr(x_4) \Pr(p_1, p_2 | x_4)}{\Pr(p_1, p_2)}.$$

She deduces the distribution of $(p_1, p_2 | x_4)$ from her distribution of $(x_1, x_2, x_3 | x_4)$:

$$\begin{aligned} \Pr(x_1, x_2, x_3 | x_4) &= \frac{\Pr(x_1, x_2, x_3, x_4)}{\Pr(x_4)} \\ &= \frac{2\Gamma\left(1 + \sum_{i=1}^4 x_i\right) \Gamma\left(5 - \sum_{i=1}^4 x_i\right)}{\Gamma(6)}. \end{aligned}$$

In Fig. 2, we see the conditional distributions that make up the expert likelihood $\Pr(p_1, p_2 | x_4)$. When p_1 (or p_2) is “integrated” out of this joint likelihood $\Pr(p_1, p_2 | x_4)$, the distributions of $p_1 | x_4$ (or $(p_2 | x_4)$) is the same as those depicted in Fig. 1. Thus, the second expert is as discriminating as the first expert. However, the experts’ reports are not independent; in either case ($x_4 = 1$ or $x_4 = 0$), the experts’ reports have a correlation, $\text{Corr}[p_1, p_2 | x_4]$, of 0.7. This dependency between the experts’ reports is due to two facts: (i) the data themselves are dependent, and (ii) both experts observe the weather on the second day. If expert 2 were not to observe the weather on day 2, the experts’ reports would have the lower correlation of 0.267. Note that without informational overlap, the experts’ reports are still dependent because beliefs about x_3 are affected by knowledge of (x_1, x_2) via the inferential model for the Bernoulli process above. Without this dependency, the experts’ report would have no relevance to the quantity of interest, and consulting the experts would be pointless.

Finally, once the decision maker updates her beliefs, she finds the aggregate forecast is one of the probabilities listed in Table 4.

Table 3. Rain Observations and Two Observer’s Reports

Day 1 Outcome (x_1)	Day 2 Outcome (x_2)	Day 3 Outcome (x_3)	Sufficient Statistic ($r_1 = \sum_{i=1}^2 x_i$)	Observer 1’s Report (p_1)	Sufficient Statistic ($r_2 = \sum_{i=2}^3 x_i$)	Observer 2’s Report (p_2)
0	0	0	0	0.25	0	0.25
0	0	1	0	0.25	1	0.50
0	1	0	1	0.50	1	0.50
0	1	1	1	0.50	2	0.75
1	0	0	1	0.50	0	0.25
1	0	1	1	0.50	1	0.50
1	1	0	2	0.75	1	0.50
1	1	1	2	0.75	2	0.75

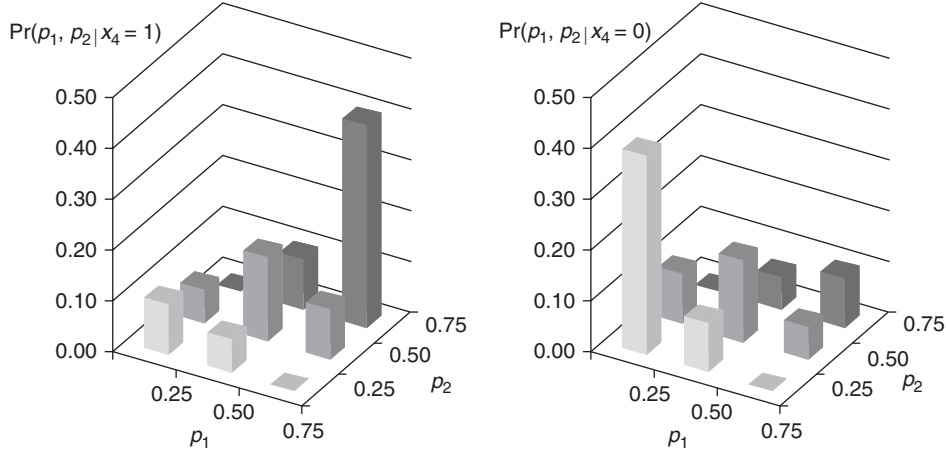


Figure 2. Expert likelihood for two observers.

Table 4. Aggregate Probabilities for Two Observers where $i = 1, 2$ and $i \neq j$

Observer i 's Report (p_i)	Observer j 's Report (p_j)	Aggregate Probability $\Pr(x_4 p_i, p_j)$
0.25	0.25	0.20
0.25	0.50	0.40
0.50	0.50	0.50
0.50	0.75	0.60
0.75	0.75	0.80

At the extremes, when the experts both report either 0.25 or 0.75, the aggregate forecast is even more extreme—either below 0.25 or above 0.75, respectively. These extreme aggregate forecasts make sense because in these cases, the experts observe either no days of rain or all days of rain. In between these extremes, the decision maker finds some compromise, and even unanimity, among the two experts. For more detail on the issues of unanimity and compromise among forecasters, see Clemen and Winkler [17].

Model for a Binary Event and Two Experts with More Refined Information Sources

Suppose the decision maker believes her experts' information is more refined than observing rain on two days. In the previous

section, each of the experts' reports p_1 and p_2 could take one of only three possible values. These experts relied on a coarse information source—rain or not on two previous days. With more refined information on the previous days' temperature, pressure, and humidity, the decision maker might believe each of the experts' reports p_1 and p_2 can take any value between 0 and 1. In this situation, as in most real-world situations, the decision maker will want an expert likelihood that is a continuous distribution.

One of the leading models of a continuous expert likelihood uses the experts' log odds ratios, $l_1 = \log \frac{p_1}{1-p_1}$ and $l_2 = \log \frac{p_2}{1-p_2}$, instead of their probabilities, (p_1, p_2) . The reason a decision maker may want to work with the experts' log odds ratios is so that she can use the multivariate normal distribution as an expert likelihood. Whereas probabilities take values on the interval $[0, 1]$, log odds ratios take values on the same interval that normally distributed quantities take, namely $(-\infty, \infty)$. The main advantage of the multivariate normal distribution is its tractable dependency structure. With the multivariate normal distribution, the decision can assess the dependency between her experts' reported log odds with a single parameter ρ_x as follows:

$$(l_1, l_2)' | x \sim N(\mu_x, \Sigma_x), \quad (1)$$

where $\mu_x = \begin{pmatrix} \mu_{x,1} \\ \mu_{x,2} \end{pmatrix}$ is its mean vector and $\Sigma_x = \begin{pmatrix} \sigma_{x,1}^2 & \rho_x \sigma_{x,1} \sigma_{x,2} \\ \rho_x \sigma_{x,1} \sigma_{x,2} & \sigma_{x,2}^2 \end{pmatrix}$ is its correlation matrix. The higher the decision maker sets $\mu_{1,i}$ relative to $\mu_{0,i}$, the more discriminating she believes her experts are. With the correlation coefficient $\rho_x = \text{Corr}[l_1, l_2|x]$, the decision maker incorporates her view of the dependency between the experts' reports. The more she thinks the experts' information sources overlap, the higher she might set this correlation coefficient.

After hearing the experts' forecasts for rain and computing their reported log odds ratios, the decision maker updates her prior beliefs $p_0 = \Pr(x = 1)$ to find the aggregate forecast:

$$\Pr(x = 1|l_1, l_2) = \frac{\Pr(x = 1)f(l_1, l_2|x = 1)}{f(l_1, l_2)} \\ = \frac{p_0 f_N(l_1, l_2|\mu_1, \Sigma_1)}{p_0 f_N(l_1, l_2|\mu_1, \Sigma_1) + (1-p_0)f_N(l_1, l_2|\mu_0, \Sigma_0)}.$$

This model can be scaled up to handle more than two experts [12]. In addition, a decision maker can use the logistic-normal distribution, which is a distribution for more than one log odds ratio [18], to handle discrete events with more than two outcomes. In the following section, we discuss an alternative model for aggregating experts' forecasts of a discrete event with more than two outcomes.

Model for a General Discrete Event and Multiple Experts

Since Lindley and Smith [19], Bayesian hierarchical modeling has become an attractive way to model the dependency between data [15,20]. In the aggregation problem, because the experts' probability forecasts are data, a decision maker may use a Bayesian hierarchical model to aggregate experts' probabilities for a general discrete event. At the bottom level of the hierarchy, the decision maker might use the popular Dirichlet for each expert's probabilities.

Suppose a decision maker uses the following hierarchical model for aggregating beliefs about a discrete event x , taking values $0, 1, \dots$, or m :

$$x \sim p_0 \\ \alpha(j)|x \sim \text{iid } \text{Ga}(a_x(j), b_x(j)), \quad j = 0, 1, \dots, m \\ p_i|\alpha, x \sim \text{iid } \text{Dir}(\alpha), \quad i = 1, \dots, k$$

where (i) the decision maker's and k experts' prior beliefs are given by $p_i = (\Pr_i(x = 0), \dots, \Pr_i(x = m))$ for $i = 1, \dots, k$, (ii) $\alpha_x = (a_x(0), \dots, a_x(m))$ and $b_x = (b_x(0), \dots, b_x(m))$ are the hyperparameters in the gamma distributions $\text{Ga}(a_x(j), b_x(j))$ with probability density function $f_{\text{Ga}}(\alpha(j)|a_x(j), b_x(j)) = \frac{b_x(j)^{a_x(j)}}{\Gamma(a_x(j))} \alpha(j)^{a_x(j)-1} e^{-b_x(j)\alpha(j)}$ for $j = 0, 1, \dots, m$, and (iii) $\alpha = (\alpha(0), \dots, \alpha(m))$ is the common parameter in the expert likelihood, a Dirichlet distribution with probability density function

$$f_{\text{Dir}}(p_0, \dots, p_{m-1}|\alpha) \\ = \frac{\Gamma\left(\sum_{j=0}^m \alpha(j)\right)}{\Gamma(\alpha(0))\Gamma(\alpha(1))\dots\Gamma(\alpha(m))} p_0^{\alpha(0)-1} p_1^{\alpha(1)-1} \dots \\ p_{m-1}^{\alpha(m-1)-1} \left(1 - \sum_{j=0}^m p_j\right)^{\alpha(m)-1}$$

In this model, the larger the decision maker sets the means $\frac{a_x(j)}{b_x(j)}$ for $j = x$, the more dependent the quantity of interest and the experts' reports are, and thus, the more discriminating she thinks the experts are. In addition, the larger she sets the variances $\frac{a_x(j)}{b_x(j)^2}$ for $j = 0, 1, \dots, m$, the more dependent she thinks the experts' reports are. Although the decision maker assumes the experts' forecasts are exchangeable in the model above, she can extend the model to allow for expert-specific α_i , $a_{x,i}$, and $b_{x,i}$ for $i = 1, \dots, k$. Such a model might include another level in the hierarchy so that the experts' hyperparameters $a_{x,i}$, and $b_{x,i}$ for $i = 1, \dots, k$ are exchangeable.

One of the earliest Bayesian aggregation models is a version of the Dirichlet model above. Morris [9] proposes a beta model where $m = 1$ and $\Pr_0(x = 0) = \frac{1}{2}$. (When $m = 1$, the Dirichlet distribution and the beta distribution are one and the same.) Instead of the gamma distribution above, he uses

a fixed parameter $\alpha = (2(1-x) + 1, 2x + 1)$. Because he assumes no variance in α , Morris' experts are independent and the conditional distributions of the expert likelihood are beta distributions and are similar in shape, although continuous, to those in Fig. 1 from the section titled "Model for a Binary Event and One Expert." If a decision maker thought her experts' reports were conditionally independent and their sources of information quite refined, then she might employ Morris' beta model.

AGGREGATION OF PARAMETRIC FORECASTS FOR A CONTINUOUS QUANTITY

When an uncertain quantity of interest can take any value in a continuous range of possible values, a decision maker may ask her experts to report their means for this continuous quantity. The leading Bayesian model for this sort of aggregation is Winkler's [10] normal consensus model:

$$x \sim \mathbf{v} \\ (\mu_1, \dots, \mu_k)' | x \sim N((x, \dots, x)', \Sigma),$$

where $\mathbf{v}$ is the noninformative (or vague) prior distribution with a constant probability density function $f_{\mathbf{v}}(x) = c$ on the entire real line, and the experts' means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)'$ are jointly normally distributed with mean vector $(x, \dots, x)'$ and covariance matrix Σ , regardless of the value of x . In this model, the decision maker lacks a proper point of view: the integral of her prior density over the entire real line does not exist. Here, the multivariate normal distribution constitutes the expert likelihood. After Bayesian updating, the aggregate forecast for x is again normal:

$$x | \mu_1, \dots, \mu_k \sim N\left(\frac{\mathbf{e}' \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{e}' \Sigma^{-1} \mathbf{e}}, (\mathbf{e}' \Sigma^{-1} \mathbf{e})^{-1}\right)$$

where $\mathbf{e}' = (1, \dots, 1)$ is a $(1 \times k)$ vector.

Lindley [21,22] extends Winkler's model to handle experts' measures of location, spread, and skewness. In many cases, reported location and spread measures, for example, mean and variance, will fully specify an

expert's probability distribution. In these cases, an expert's location and spread measures completely parameterize the expert's forecast (e.g., normal, gamma, beta, etc.), and we say the expert reports a parametric forecast for a continuous quantity.

Importantly, the expert likelihood above may arise from experts who access overlapping information that is exchangeable with the quantity of interest [10,23,24]. In the following example, we consider a related situation where the decision maker has a proper point of view, that is, her prior distribution of the quantity of interest is a proper distribution [25]. Suppose the decision maker and two experts share the belief that the data $(x_1, \dots, x_{n+1})$ follows a normal process with an unknown, normal mean θ and a known variance σ^2 :

$$\theta \sim N(\mu_0, \sigma_0^2) \\ x_i | \theta \sim_{\text{iid}} N(\theta, \sigma^2), \quad i = 1, \dots, n+1$$

These data might be temperatures on four consecutive days in a new city. The decision maker plans to aggregate the experts' forecasts for the temperature on the fourth day, once the experts return from their overlapping trips to the new city. Expert 1 observes the temperature on day 1, expert 2 observes the temperature on day 3, and both experts observe the temperature on day 2.

After observing their respective samples, expert 1 holds the following posterior-predictive beliefs [15, p. 439]:

$$x_4 | x_2, x_1 \sim N\left(\frac{\lambda_0 \mu_0 + \lambda(x_1 + x_2)}{\lambda_0 + 2\lambda}, \frac{\lambda_0 + 3\lambda}{\lambda(\lambda_0 + 2\lambda)}\right),$$

where $\lambda_0 = \frac{1}{\sigma_0^2}$ and $\lambda = \frac{1}{\sigma^2}$ are the precisions of the process, and similarly, expert 2 holds the following posterior-predictive beliefs:

$$x_4 | x_3, x_2 \sim N\left(\frac{\lambda_0 \mu_0 + \lambda(x_2 + x_3)}{\lambda_0 + 2\lambda}, \frac{\lambda_0 + 3\lambda}{\lambda(\lambda_0 + 2\lambda)}\right).$$

Thus, each expert can report the mean μ_i of his posterior-predictive distribution without any loss of information. This situation of informational overlap parallels the one in the

section titled “Model for a Binary Event and Two Overlapping Experts” where two experts observed data from a Bernoulli process.

To update the decision maker's prior beliefs based on the experts' reports, the decision maker uses Bayes' theorem to find the conditional distribution of x_{n+1} given the experts' reports (μ_1, μ_2) :

$$f(x_4|\mu_1, \mu_2) = \frac{f(x_4)f(\mu_1, \mu_2|x_4)}{f(\mu_1, \mu_2)}$$

In this case, the decision maker deduces both the distribution of $(\mu_1, \mu_2|x_4)$ and the distribution of $(x_4|\mu_1, \mu_2)$ directly from her joint predictive distribution:

$$(x_1, x_2, x_3, x_4)' \sim N \left((\mu_0, \mu_0, \mu_0, \mu_0)', \begin{pmatrix} \sigma^2 + \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma^2 + \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 & \sigma^2 + \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \sigma^2 + \sigma_0^2 \end{pmatrix} \right).$$

Because the data are exchangeable in this model, the correlation coefficients $\rho_{ij} = \text{Corr}[x_i, x_j]$ in this joint predictive distribution are all the same $\rho_{ij} = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$. With a change of variables from $(x_1, x_2, x_3, x_4)'$ to $(x_4, \mu_1, \mu_2, x_2)'$, according to the linear transformation $(x_4, \mu_1, \mu_2, x_2)' = \mathbf{a} + \mathbf{B}(x_1, x_2, x_3, x_4)'$

where $\mathbf{a} = \left(0, \frac{\lambda_0 \mu_0}{\lambda_0 + 2\lambda}, \frac{\lambda_0 \mu_0}{\lambda_0 + 2\lambda}, 0\right)'$ and

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ \frac{\lambda}{\lambda_0 + 2\lambda} & \frac{\lambda}{\lambda_0 + 2\lambda} & 0 & 0 \\ 0 & \frac{\lambda}{\lambda_0 + 2\lambda} & \frac{\lambda}{\lambda_0 + 2\lambda} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

we have

$$(x_4, \mu_1, \mu_2, x_2)' \sim N \left((\mu_0, \mu_0, \mu_0, \mu_0)', \begin{pmatrix} \frac{\lambda_0 + \lambda}{\lambda_0 \lambda} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{1}{\lambda_0} \\ \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)^2} & \frac{1}{\lambda_0} \\ \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)^2} & \frac{1}{\lambda_0} \\ \frac{1}{\lambda_0} & \frac{1}{\lambda_0} & \frac{1}{\lambda_0} & \frac{\lambda_0 + \lambda}{\lambda_0 \lambda} \end{pmatrix} \right)$$

This distribution follows from the standard formula for a linear transformation of a jointly normal vector [26, p. 637].

Because $(x_4, \mu_1, \mu_2, x_2)'$ is jointly normally distributed, we can find both the expert likelihood and the aggregate distribution

by applying the standard formulas to find the marginal and conditional distributions of partitions of a jointly normal vector [26, p. 637]. First, the marginal distribution of the partition $(x_4, \mu_1, \mu_2)'$ is given by

$$(x_4, \mu_1, \mu_2)' \sim N \left((\mu_0, \mu_0, \mu_0)', \begin{pmatrix} \frac{\lambda_0 + \lambda}{\lambda_0 \lambda} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} \\ \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)^2} \\ \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)} & \frac{2\lambda}{\lambda_0(\lambda_0 + 2\lambda)^2} \end{pmatrix} \right)$$

The quantity of interest and either expert's mean are positively correlated with correlation coefficient $\frac{\sqrt{2\lambda}}{\sqrt{(\lambda_0 + 2\lambda)(\lambda_0 + \lambda)}}$, which increases as uncertainty σ_0^2 about the unknown mean of the data increases.

The higher this correlation, the more discriminating the expert is.

Next, the expert likelihood is given by the conditional distribution of the partition $(\mu_1, \mu_2)'$ given the partition x_4 of the jointly normal vector $(x_4, \mu_1, \mu_2)'$:

$$(\mu_1, \mu_2)' | x_4 \sim N \left(\begin{pmatrix} \frac{\lambda_0(\lambda_0+3\lambda)\mu_0+2\lambda^2x_4}{(\lambda_0+2\lambda)(\lambda_0+\lambda)} \\ \frac{\lambda_0(\lambda_0+3\lambda)\mu_0+2\lambda^2x_4}{(\lambda_0+2\lambda)(\lambda_0+\lambda)} \end{pmatrix}, \begin{pmatrix} \frac{2\lambda(\lambda_0+3\lambda)}{(\lambda_0+2\lambda)^2(\lambda_0+\lambda)} & \frac{\lambda(\lambda_0+5\lambda)}{(\lambda_0+2\lambda)^2(\lambda_0+\lambda)} \\ \frac{\lambda(\lambda_0+5\lambda)}{(\lambda_0+2\lambda)^2(\lambda_0+\lambda)} & \frac{2\lambda(\lambda_0+3\lambda)}{(\lambda_0+2\lambda)^2(\lambda_0+\lambda)} \end{pmatrix} \right)$$

In part, because the experts' information sources overlap, the correlation coefficient of the experts' means, given the quantity of interest, is $\frac{\lambda_0+5\lambda}{2\lambda_0+6\lambda}$.

Finally, the aggregate distribution is given by the conditional distribution of the partition x_4 given the partition $(\mu_1, \mu_2)'$:

$$x_4 | \mu_1, \mu_2 \sim N \left(\frac{2(\lambda_0+2\lambda)(\mu_1+\mu_2)-\lambda_0\mu_0}{(3\lambda_0+8\lambda)}, \frac{3\lambda_0+11\lambda}{\lambda(3\lambda_0+8\lambda)} \right).$$

Had the decision maker consulted with only expert 1, her aggregate distribution would not surprisingly, be given by $x_4 | \mu_1 \sim N \left(\mu_1, \frac{\lambda_0+3\lambda}{\lambda(\lambda_0+2\lambda)} \right)$, which is identical to expert 1's report. Similar to the observations in the sections titled "Model for a Binary Event and One Expert" and "Model for a Binary Event and Two Overlapping Experts," it is as if she observed the temperature on the first two days herself.

A decision maker who wishes to aggregate experts' reported means (or location parameters from a parametric distribution) may use the above example in a thought experiment. She might assess expert discrimination by thinking about each expert's experience in terms of the quantity and quality of the relevant data that the expert has observed. In addition, she might assess expert dependency by thinking about her experts in terms of how much overlap there is in the experts' experiences.

AGGREGATION OF NONPARAMETRIC FORECASTS FOR A CONTINUOUS QUANTITY

Often times a decision maker will find the assumption that an expert reports according to a parametric distribution too restrictive. In these situations, the decision maker may believe an expert's forecast comes from a highly customized, or nonparametric, distribution. In practice, a decision maker typically elicits such a nonparametric forecast by one

of two methods: either (i) by asking for several quantiles from the expert's quantile function, or (ii) by asking for several probabilities from the expert's cumulative distribution function. In this section, we present two models (one for each method) a decision maker may use to assess an expert likelihood for nonparametric forecasts.

Model for Nonparametric Quantile Forecasts

The earliest model for updating upon hearing an expert's nonparametric quantile forecasts is the following model [27]:

$$x \sim F_0 \\ G_x(Q_1) | x \sim \text{DP}(c\lambda)$$

where (i) F_0 is the decision maker's prior distribution with density $f_0(x)$, (ii) Q_1 is the expert's reported quantile function, $Q_1(u) = \min\{u : F_1(x) \geq u\}$, which is derived from the expert's cumulative distribution function F_1 , (iii) G_x is a distribution function, indexed by x , and (iv) $\text{DP}(c\lambda)$ is a Dirichlet process with a parameter $c\lambda$ that measures the interval $(u_i, u_j]$ as follows:

$$c\lambda((u_i, u_j]) = c(u_j - u_i)$$

where c is a scalar.

The Dirichlet process is the leading model used in Bayesian nonparametric statistics to describe uncertainty about a random distribution function [28,29]. The Dirichlet process is a stochastic process F defined by its finite-dimensional Dirichlet distributions. In particular, the increments of the random distribution function F evaluated at a set of points in any partition of x 's state space $x_0 < x_1 < x_2 < \dots < x_m < x_{m+1}$ are distributed according to a Dirichlet distribution:

$$(F(x_1) - F(x_0), F(x_2) - F(x_1), \dots, F(x_{m+1}) - F(x_m)) \sim \text{Dir}(\alpha((x_0, x_1]), \dots, \alpha((x_m, x_{m+1}]),$$

where α called the process's centering measure because $E[F(x_{j+1}) - F(x_j)] = \alpha(x_j, x_{j+1})$.

In West's [27] model, the Dirichlet process is used in a related way to describe uncertainty about a function G_x of a random quantile function Q_1 .

In West's [27] model above, the distribution function G_x is called a *target distribution* and is the main ingredient in the expert likelihood. For example, suppose $G_{x,\theta}$ is a beta target distribution with density

$$g_{x,\theta}(q) = f_{\text{Be}}(F_0(q)|\theta F_0(x) + 1, \theta(1 - F_0(x)) + 1)f_0(q),$$

where $\theta > -1$ is a discrimination parameter that further indexes the target [30, pp. 76–81]. When $\theta = 0$, $g_{x,\theta}(q) = f_0(q)$: the decision maker believes the expert is noninformative. As $\theta \rightarrow \infty$, $G_{x,\theta}$ goes to a point mass of 1 at x : the decision maker believes the expert is a clairvoyant. For $\theta \in (0, \infty)$, the decision maker expects the expert to report more of his probability around x than she assessed with her prior.

For the expert who reports his m quantiles $Q_1(u_1), \dots, Q_1(u_m)$ where $0 = u_0 < u_1 < \dots < u_m < u_{m+1} = 1$, West's [27] expert likelihood becomes

$$G_{x,\theta}(Q_1(u_1)) - G_{x,\theta}(Q_1(u_0)), \dots, G_{x,\theta}(Q_1(u_{m+1})) - G_{x,\theta}(Q_1(u_m)) \sim \text{Dir}(c\lambda),$$

where $c\lambda = (c\lambda((u_0, u_1]), \dots, c\lambda((u_m, u_{m+1}]))$. After Bayesian updating, the density $f(x|Q_1(u_1), \dots, Q_1(u_m))$ is proportional to

$$f(x) f_{\text{Dir}}(G_{x,\theta}(Q_1(u_1)) - G_{x,\theta}(Q_1(u_0)), \dots, G_{x,\theta}(Q_1(u_{m+1})) - G_{x,\theta}(Q_1(u_m)) | c\lambda) \times \prod_{j=1}^m g_{x,\theta}(Q_1(u_j)).$$

West's [27] model easily extends to incorporate more than just one expert's quantile forecasts. For example, the following hierarchical model, which is similar to the one discussed in the section titled "Model for a General Discrete Event and Multiple Experts," can be used to aggregate multiple experts' nonparametric quantile forecasts:

$$\begin{aligned} x &\sim F_0 \\ (\theta + 1)|x &\sim \text{Ga}(a, b) \\ G_{x,\theta}(Q_i)|c, x &\sim \text{iid DP}(c\lambda), \quad i = 1, \dots, k. \end{aligned}$$

Although West's [27] model involves a straightforward application of the Dirichlet process, we offer a technical note of caution about the Dirichlet process in this context. The Dirichlet process puts all of its probability on a set of discrete distribution functions. That is, it inherits the properties of a stochastic process with an infinite number of jumps occurring at random places in x 's state space with random heights [28]. In the limit, as the decision maker elicits an expert's entire continuous quantile function, the Dirichlet process breaks down. Contrary to the claim in West [27, Theorem 2], the decision maker cannot use the Dirichlet process as a proper likelihood for an expert's entire continuous quantile function [30, pp. 35–37]. In other words, she cannot update on an event that was assigned probability zero.

Nonetheless, the Dirichlet process is a good approximation for the distribution of a random continuous distribution function or, as West [27] uses it, the distribution of a function of a random continuous quantile function. West's [27] model works perfectly fine for any finite number of reported quantiles. This is because it assigns a reasonable continuous distribution (as an expert likelihood) to any finite number of reported quantiles.

Model for Nonparametric Probability Forecasts

We offer the following model for a decision maker who elicits a finite number of expert's probabilities $\mathbf{p}_i = (F_i(x_1), F_i(x_2) - F_i(x_1), \dots, 1 - F_i(x_m))$ over any partition of x 's state space $x_0 < x_1 < x_2 < \dots < x_m < x_{m+1}$, which is analogous to the one in the section titled "Model for a General Discrete Event and Multiple Experts":

$$\begin{aligned} x &\sim F_0 \\ \alpha|x &\sim \text{GP}(a_x, b_x) \\ \mathbf{p}_i|\alpha, x &\sim \text{iid DP}(\alpha), \quad i = 1, \dots, k. \end{aligned}$$

This model assigns a gamma processes to the common measure

$$a = (a((x_0, x_1]), \dots, a((x_m, x_{m+1}))),$$

that is,

$$\begin{aligned} \alpha((x_j, x_{j+1}]) | x &\sim_{\text{iid}} \text{Ga}(a_x((x_j, x_{j+1}]), \\ b_x((x_j, x_{j+1}))), \quad j &= 0, 1, \dots, m \end{aligned}$$

and independent Dirichlet processes to the experts' probability measures where α is each process's centering parameter.

As in the case of West's [27] model for nonparametric quantile forecasts, the model above assigns a reasonable continuous distribution (as an expert likelihood) for any finite number of probabilities. Consequently, one might expect such a model for random probability measures to induce a reasonable continuous distribution for the measure's corresponding quantiles (and vice versa). Interestingly though, this model does not induce a continuous distribution for its corresponding quantiles. A Dirichlet process used to describe the uncertainty in a random continuous distribution function induces a mixed finite-dimensional distribution for its corresponding quantiles—part discrete and part continuous distribution [31]. Thus, when aggregating nonparametric forecasts, a decision maker will want to identify which type of forecast—either quantile or probability—her experts will report. Once she knows the type of forecast, she can assess the appropriate expert likelihood.

SUMMARY

When a decision maker faces a decision, she may wish to consult others for their forecasts of an important uncertainty. Once she hears the experts' forecasts, the decision maker will look to combine them into a single coherent forecast. The principle behind the Bayesian approach to this problem is both simple and logically consistent. From a joint distribution of the quantity of interest and the experts' forecasts, calculate the conditional distribution of the quantity of interest given the experts' forecasts. This is the easy part. The difficult part is the assessment of the joint distribution.

The decision maker can break the assessment of this joint distribution into two parts: (i) assess her own prior distribution for the quantity of interest (which she would need in any case), and (ii) assess an expert likelihood, the conditional distribution of the experts' forecasts given the quantity of interest. The real challenge in Bayesian aggregation is the assessment of the expert likelihood. In the preceding sections, we saw this challenge decomposed into two more manageable assessment tasks: (i) judge each expert's discrimination ability, and (ii) judge the dependency between the experts' forecasts.

First, the assessment of an expert's discrimination ability amounts to the estimation of the dependency between the event of interest and the expert's forecast. Does the expert tend to report higher probabilities for the event when the event occurs and vice versa? Second, the assessment of the dependency between experts' forecasts amounts, in part, to an estimation of the degree to which the experts' information sources overlap. Do the experts share information and experiences in common? In the section titled "Model for a Binary Event and Two Overlapping Experts," we saw these two tasks come together in the context of forecasting a binary event. There, the decision maker's beliefs about these dependencies were affected by a shared inference model and some informational overlap. Similarly, in the section titled "Aggregation of Parametric Forecasts for a Continuous Quantity," we saw how these two assessment tasks played out in the context of forecasting a normal continuous quantity.

In the examples from the sections titled "Model for a Binary Event and Two Overlapping Experts" and "Aggregation of Parametric Forecasts for a Continuous Quantity," the aggregate distribution followed from shared beliefs about two explicit, small-scale, physical processes. The physics of these processes involved two experts who first collected a few pieces of exchangeable data, some of which overlapped, and then reported their updated beliefs. The decision maker deduced the aggregate distribution directly from the group's shared beliefs. In each case, a conjugate Bayesian inference model constituted their shared beliefs. We looked at the

most popular Bayesian inference models for a binary event and a continuous quantity: the beta-binomial model and the normal-normal model, respectively. Elsewhere, we showed some important extensions of these two models. These extensions used more free-form expert likelihoods. They anticipated hearing from experts with more refined and more complicated information sources—situations in which experts report multiple summary or nonparametric statistics.

In practice, a decision maker may reflect on the similarities the small-scale models presented in the section titled “Model for a Binary Event and Two Overlapping Experts” and “Aggregation of Parametric Forecasts for a Continuous Quantity,” share with actual large-scale problems. In the real world, expert information and experiences are more refined and messier than could be described with a small-scale model of exchangeable data. Nonetheless, a real-world expert likelihood will likely have the same statistical properties our small-scale expert likelihoods have. They will exhibit dependencies between forecasts and events. To assess these dependencies in the real world, our small-scale examples become powerful thought experiments. To assess expert discrimination, a decision maker might think about expert experience in terms of the quantity and quality of relevant data an expert has observed. To assess expert dependency, she might consider the degree to which there is overlap in the experts’ experiences.

In the end, to assess the dependencies between forecasts and events well is a skill best developed through repeated interactions with data. In environments where experts forecast events frequently, decision makers have the opportunity to collect valuable data on forecast and event outcomes. With enough data, a decision maker can build histograms (similar to Fig. 1 and 2) of forecasts given event outcomes. These histograms can informally guide the assessment of an expert likelihood. Alternatively, a decision maker may work directly with a hierarchical model, such as the one discussed in the section titled “Model for a General Discrete Event and Multiple Experts,” to update beliefs, given the historical data about

the hierarchical parameters in the model. These days, Bayesian simulation techniques and modern computing power make such inference possible [32]. On one level, this conclusion amounts to saying, let the data speak for themselves through our models.

REFERENCES

1. Galton F. Vox populi. *Nature* 1907;75(1949): 450–451.
2. Surowiecki J. *The wisdom of crowds*. New York: Anchor Books; 2004. pp. XI–XIII.
3. Galton F. One vote, one value. *Nature* 1907;75(1948):414. (Letters to the Editor).
4. Hooker RH. Mean or median. *Nature* 1907;75(1951):487–488. (Letters to the Editor).
5. Galton F. The ballot-box. *Nature* 1907; 75(1952):509–510. (Letters to the Editor).
6. Pearson K. Volume 2, *The life, letters, and labours of Francis Galton. Researches of middle life*. Cambridge: Cambridge University Press; 1924. pp. 403–405.
7. Savage S. *The flaw of averages: why we underestimate risk in the face of uncertainty*. Hoboken (NJ): Wiley; 2009.
8. Fisher M, Raman A. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper Res* 1996;44(1):87–99.
9. Morris PA. Decision analysis expert use. *Manage Sci* 1974;20(0):1233–1241.
10. Winkler RL. Combining probability distributions from dependent information sources. *Manage Sci* 1981;27(4):479–488.
11. Stone M. The opinion pool. *Ann Math Stat* 1961;32:1339–1342.
12. Clemen RT, Winkler RL. Aggregating probability distributions. In: Edwards W, Miles RF, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007. pp. 154–176.
13. Cooke RM. *Experts in uncertainty: opinion and subjective probability in science*. New York: Oxford University Press; 1991.
14. Genest C, Zidek JV. Combining probability distributions: a critique and an annotated bibliography. *Stat Sci* 1986;1(1):11–135.
15. Bernardo J, Smith AFM. *Bayesian Theory*. West Sussex: Wiley; 2000. (Appendix A).

16. Winkler RL, Poses RM. Evaluating and combining physician's probabilities of survival in an intensive care unit. *Manage Sci* 1993;39(12):1526–1543.
17. Clemen RT, Winkler RL. Unanimity and compromise among probability forecasters. *Manage Sci* 1990;36(7):767–779.
18. Aitchison J, Shen SM. Logistic-normal distributions: some properties and uses. *Biometrika* 1980;67(2):261–272.
19. Lindley DV, Smith AFM. Bayes estimates for the linear model (with discussion). *J R Stat Soc B* 1972;34:1–41.
20. Zaslavsky AM. Hierarchical bayesian modeling. In: Press SJ, editor. *Subjective and objective Bayesian statistics*. Hoboken (NJ): Wiley; 2003. pp. 336–356.
21. Lindley DV. Reconciliation for probability distributions. *Oper Res* 1983;31(5):866–880.
22. Lindley DV. Using expert advice on a skew judgmental distribution. *Oper Res* 1987;35(5):716–721.
23. Zeckhauser R. Combining overlapping information. *J Am Stat Assoc* 1971;66(333):91–92.
24. Clemen RT. Combining overlapping information. *Manage Sci* 1987;33(3):373–380.
25. Jose VRR, Lichtendahl KC Jr, Nau RF, *et al.* Objective Bayesian analysis for the multivariate normal model: comment. In: Bernardo JM, *et al.* editors. *Bayesian statistics 8*. Oxford: Oxford University Press; 2007. pp. 557–558.
26. West M, Harrison J. *Bayesian forecasting and dynamic models*. New York: Springer; 1997.
27. West M. Modelling expert opinion. In: Bernardo JM, DeGroot MH, *et al.* editors. *Bayesian Statistics 3*. Oxford: Oxford University Press; 1988. pp. 493–508.
28. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann Stat* 1973;1(2):209–230.
29. Hjort NL, Holmes C, Muller P, *et al.* editors. *Bayesian nonparametrics*. Cambridge: Cambridge University Press; 2010.
30. Lichtendahl KC Jr. *Bayesian models of expert forecasts* [PhD. dissertation]. Durham (NC): Duke University; 2006.
31. Lichtendahl KC Jr. Random quantiles of the Dirichlet process. *Stat Probab Lett* 2009;79(4):501–507.
32. Gamerman D, Lopez HF. *Markov chain Monte Carlo: stochastic simulation for bayesian inference*. Boca Raton (FL): Chapman & Hall/CRC; 2006.

BAYESIAN NETWORK CLASSIFIERS

MOISES GOLDSZMIDT
Microsoft Research, Mountain
View, California

The aim of classification is to assign objects, described in terms of their properties called *features*, to one of a finite number of discrete categories called *classes*. One example is deciding whether a disk will fail (stop working) within the next three days, (*class*), given a vector of measurements such as latency of reads/writes, size of the queues, number of retries, and number of reallocated sectors (*features*). Such a classifier can be used to initiate a disk replacement procedure in time without disruption of the service the disk is supporting. Another example is identifying a digit from the set 0 to 9, given information about the luminosity, color, and location of the pixels in a scan of a check. This capability can be used to automatically scan the amounts of checks at ATM machines or to automatically sort mail by recognizing zip codes. Yet another example is mapping documents to a predefined set of topics, based on the number of times each word in the dictionary appears in the document. This classifier is useful in a number of contexts including novelty detection in web news services.

Although classifiers can be built by hand through careful engineering, it makes sense to take advantage of the large amounts of data that are being collected today and build them automatically. Data about disk behavior are being regularly monitored in data centers. There are also large corpus of scanned checks and letters providing examples of digits in different contents; and to maintain search engines' working and news content flowing, there are numerous crawls of documents in the web which provide data for a classifier that detects novelty. This automation brings the additional benefit of making these classifiers adaptive to new data and behaviors.

The task of inducing classifiers in an automated fashion from data is the domain of pattern classification, statistics, and machine learning. It is one of the most successful tasks in these disciplines in the sense that (i) it is relatively well understood, and (ii) we have examples of high quality classifiers in various domains [1–5]. Consequently, there are a large number of techniques and algorithms, including decision trees [6], support vector machines [7], logistic regression [8], neural networks [5], and Bayesian network classifiers [9,10] to name a few.

Bayesian network classifiers work by representing a joint probability distribution model of the domain of interest. What make Bayesian network classifiers unique is that they arguably offer all of the following advantages:

1. Bayesian networks can combine both expert knowledge and statistical data [11]. Thus, it is possible to incorporate explicit expert knowledge about the domain into the classifier. For example, we can (both) explicitly encode and/or automatically discover relations about the (in)dependence among the features.
2. We can also encode knowledge about the probability distribution in the form of priors [12].
3. We can perform sound inference, both in fitting the probability distribution and when making a classification decision, even in those cases where some of the features are not measured [12,13].
4. As we have a complete distribution of the domain, we can make inferences regarding the most important features determining the class (decision) in each instance, and also make inferences regarding the importance of each of the features on expectation. This can be useful in data collection, in experimental design, and in justifying the rationale for decisions (see Cohen *et al.* [14] for an application where this was required).

5. We can also use the probability of any given decision (or the odds expressed in Equation (1)) as an indication of the confidence of the decision.

The justification for using Bayesian networks as classifiers is rooted in Bayesian decision theory (BDT) [1]. Consider the disk failure prediction task mentioned above. The class represented by C takes two values, c^1 for “the disk will fail within the next three days,” and c^0 for “the disk will not fail within the next three days.” The vector of features is denoted by $\mathbf{F}$. Let $P(C|\mathbf{F})$ denote the probability that the disk will or will not fail within the next three days. The uncertainty that gives rise to this probability comes from at least three factors: (i) we may not have all the necessary information in the feature vector $\mathbf{F}$ for the prediction (and, as it is usually the case, it may be impossible to directly measure such information); (ii) we may not have a precise notion of the functional relationship between the feature vector and the class; and (iii) there may be errors in the measurements of $\mathbf{F}$. The point of using a classifier is to take action or, in BDT terms, to make a decision. In our case, this may be to change the disk and use the period of three days to back up its contents in response to the signal from the classifier. Clearly, there is a cost for executing these actions: if the disk indeed fails in three days and we execute the back up and change the disk, we have the cost of a new disk and the resources expended in the backup; let us use r_{11} to denote this cost. However, if we do not replace the disk and we do not save the data, we will have the cost of the lost data, plus the fact that the service that depends on the disk will be unavailable. We use r_{10} to denote this cost. Let us use r_{00} for the cost of “doing nothing” (normal operation) and r_{01} for the unnecessary cost of backup plus replacement of a healthy disk. Invoking Bayes rule, and under the reasonable assumption that $r_{10} > r_{00}$, it is optimal (in the sense of minimizing risk) to decide to do nothing *if and only if*

$$\frac{p(\mathbf{F}|c^0)}{p(\mathbf{F}|c^1)} > \frac{(r_{01} - r_{11}) P(c^1)}{(r_{10} - r_{00}) P(c^0)}, \quad (1)$$

where p denotes a density function and P denotes probability. This essentially says that the optimal decision depends on the odds of how likely is the state of the monitored features under normal conditions (disk will not fail) versus the same state under the condition that the disk will fail. Note that the problem of inducing a classifier from data is now transformed to the problem of fitting the density (or probability) $p(\mathbf{F}|C)$ (as well as the terms relating to $P(C)$ on the right-hand side).

The rest of this article is organized as follows. We first formally introduce Bayesian networks in the section titled “Bayesian Networks,” and then in the section titled “Learning Bayesian Networks from Data,” we explain how they are induced automatically from data. We then address the learning of Bayesian network classifiers in the section titled “Bayesian Networks as Classifiers.” The section titled “Final Remarks” concludes this article with a set of final remarks.

BAYESIAN NETWORKS

A Bayesian network is an efficient encoding of a joint probability distribution defined over a set of random variables. The efficiency comes from explicitly representing relations of probabilistic independence, and providing algorithms to reason about these relations [15]. A Bayesian network for a set of random variables $\mathcal{X} = \{X_1, \dots, X_n\}$ consists of (i) a directed acyclic graph (DAG) $\mathcal{G}$ and (ii) a set of local probability distributions/densities p associated with each random variable. Together, these components define a unique joint probability distribution over $\mathcal{X}$. The nodes in the DAG are in one to one correspondence with the elements in the set $\mathcal{X}$. Let us use X_i to denote both the random variable and its corresponding node in $\mathcal{G}$. Let us further use Pa_i to denote the parents of node X_i in $\mathcal{G}$ (as well as the random variables corresponding to those parents). The structure of the DAG encodes the following statements about conditional independence: given the state of its parents Pa_i , X_i is conditionally independent of all its other nondescendants in the graph. Consequently, given the structure of

S , the joint probability distribution for $\mathcal{X}$ is given by

$$p(\mathcal{X}) = \prod_{i=1}^n p(X_i | Pa_i). \quad (2)$$

To illustrate the benefits of this factorization, we consider again the example of using a classifier to predict a disk failure. For the sake of this illustration, let us assume that each one of the variables $\{F_1, \dots, F_n\}$ is binary. A realistic scenario may be that these features are monitored signals in the disk and that they can be in one of the two states: *normal* or *abnormal*. In order to apply the classifier and make decisions, we need to compute the ratio in Equation (1). This implies that we need to have the values for the 2^n parameters that comprise $P(F_1, \dots, F_n | C)$.

Now assume that an “expert” builds a Bayesian network similar to the one in Fig. 1 for this domain. This network encodes the independence statement that each measurement F_i is independent of all the other measurements in the set $\{F_1, \dots, F_n\}$, given the value of the class. According to Equation (2), this translates into

$$p(F_1, \dots, F_n | C) = \prod_{i=1}^{i=n} p(F_i | C), \quad (3)$$

which requires only $2 \times n$ parameters. Using the Bayesian network in Fig. 1 we have gained exponential savings in terms of the parameterization of the domain!

Note that there is nothing that prevents us from representing further dependencies between features in $\mathbf{F}$. For example, our expert may determine (or we may gather evidence from data) that the signal that monitors the number of I/O retries (F_r) is correlated with the signal that monitors the length of the I/O queue (F_q). To include this dependence in the model, we add an edge from F_r to F_q in the Bayesian network, and then replace the term $p(F_q | C)$ with $p(F_q | F_r, C)$ in Equation (3). Note that this term now requires four parameters instead of two. More complex relationships can be added accordingly.

Having formally defined Bayesian networks and illustrated its benefits in terms

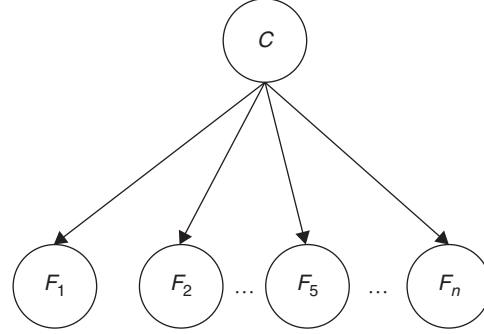


Figure 1. Naive Bayes network structure. The top node C represents the class and the nodes $F_1, \dots, F_n$ represent the features.

of the efficiency of representing a probability distribution, we concentrate on the methods by which we can induce one directly from data.

LEARNING BAYESIAN NETWORKS FROM DATA

Given the structure of the Bayesian network, such as the one in Fig. 1 and a data set of tuples $\langle C, F_1, \dots, F_n \rangle$, fitting the parameters of the conditional densities (probabilities) defined between a node and its parents in the network is conceptually a well-studied statistical problem of parameter estimation [12,16,17]. This problem can be approached in several ways. Two common and reasonable procedures are based on *maximum Likelihood* estimation and *Bayesian* estimation. The procedures are quite different conceptually, and the results obtained by these two procedures can be quite different as well. Maximum likelihood method considers the parameters as quantities whose values are fixed but unknown. The best estimate of the parameters is defined to be the one that maximizes the probability of obtaining the samples actually observed [16,18]. Bayesian methods view the parameters as random variables having some known *a priori* distribution. Observation of the samples converts this to an a posteriori density, thereby revising the true value of the parameters [17,19]. Let us go through the computations needed to fit the parameters of Fig. 1, for

both the maximum likelihood and Bayesian procedures. As an illustration, we take one of the terms, say $P(F_i|c^1)$, from Equation (3). We assume that we have a set of samples $S = \{f_i^1, f_i^2, \dots, f_i^j\}$, with the samples in S having been drawn independently according to the probability law $P(F_i|c^1)$. We assume that $P(F_i|c^1)$ has a known parametric form, and is therefore determined uniquely by the value of a parameter vector θ_i . In our example, given the fact that F_i can take one of the two values, it is reasonable to assume that the distribution is binomial and θ_i is a singleton. If we had assumed that the distribution is normal, then θ_i would be composed of a mean and a variance, which are the parameters of the normal distribution. We start with the approach based on maximum likelihood. To show the dependence of $P(F_i|c^1)$ on θ_i explicitly, we write $P_{\theta_i}(F_i|c^1)$. Our problem is to use the information provided by the samples to obtain good estimates for the unknown parameter θ_i . Let us take the standard assumption that the parameters for the different classes and the different $p(F_k|c^1)$ are functionally independent (namely, the samples do not provide any information to other parameters in the model). Then, since the samples were drawn independently we get that $p_{\theta_i}(S|c^1) = \prod p_{\theta_i}(f_i^k|c^1)$. As a function of θ_i , $p_{\theta_i}(S|c^1)$ is called the *likelihood* of θ_i with respect to the set of samples. The *maximum likelihood* estimate of θ_i is by definition the value $\hat{\theta}_i$ that maximizes $p_{\theta_i}(S|c^1)$. Intuitively, it corresponds to the value of θ_i that best agrees with the observed samples. If $p_{\theta_i}(S|c^1)$ is a well-behaved, differentiable function of θ_i , $\hat{\theta}_i$ is usually found by the standard methods of differential calculus.

The Bayesian methodology starts from the point of view that we should use all the information available in estimating $p(F_i|c^1)$, which in this case is the set of samples S . The goal then is to compute $p(F_i|S, c^1)$ which is as close as we can come in obtaining $p(F_i|c^1)$. Let us assume (as in the case of the maximum likelihood method) a specific functional parameterization of $p(F_i|c^1)$ given by θ_i . Then we compute $p(F_i|S, c^1)$ as an expectation over all possible values of the parameter θ_i

$$p(f_i|S, c^1) = \int p(f_i|\theta_i, c^1)p(\theta_i|S, c^1) d\theta \quad (4)$$

where the first term in the integral comes about because knowing θ_i renders f_i conditionally independent of the actual samples S . Note that θ_i is regarded as another random variable subject to a density/distribution $p(\theta|S)$. Thus, it can be explicitly represented as another node in the Bayesian network, where the assumptions of independence between parameters follow from the structure of the network. The computation of the integral in Equation (4) for a large set of common parameterizations and prior/posterior distributions on the parameters can be obtained in closed form [17]. Otherwise, Markov chain Monte Carlo techniques are used to estimate it [20].

It is beyond the scope of this article to describe in detail these two approaches, and their advantages and disadvantages. The answer to the question of which one is “correct” depends on the definition of the term “correct.” All estimators will have different formal properties and the estimator to be selected will depend on the task at hand. For the classification task, we may compare different estimates on the basis of the error rate of the resulting model, namely, the rate by which the classifier predicts the correct class given an agreed upon data set D to be used as a benchmark (see the section titled “Final Remarks”). It is important to note two things. First, in spite of having *Bayesian* in the name, Bayesian networks are not tied to any particular way of estimating the parameters and the user can select the method that is more convenient for the task at hand. Second, under the assumptions of independence between the parameters outlined above, the learning task decomposes following the structure of the network.

At this point, we know how to learn the “probabilities” given the structure of the network. Let us consider the problem of learning both the structure and probabilities given a data set of tuples $\langle C, F_1, \dots, F_n \rangle$. There are basically three approaches to do this. The first one takes a purely Bayesian approach. Note that, in principle, we can introduce a random variable G encoding possible graph structures and extend Equation (4) with the appropriate terms to compute the expectation over the set of possible structures G .

There are many practical problems, given the combinatorial nature of the space of possible structures G , with computing some of the required averages (see Friedman and Koller [21] for an approach based on Markov chain Monte Carlo techniques). More popular approaches treat this learning task as a combinatorial search problem, guided by a score, which in turn is computed using a maximum likelihood, maximum a posteriori, or Bayesian estimation techniques [11,12]. Thus, given a particular structure G' , the parameters are estimated using statistical techniques. Then, a score on the complete network is computed and used to select the next possible network structure G' . Different criteria and search methods are studied extensively in Chickering [22] and references therein. The third approach takes the definition of independence encoded in the structure as a starting point, and relies on statistical independence testing between the different random variables to discover the structure of the network [23,24]. More information about the theory and principles behind graphical models, inference, and learning in a text-book form can be found in Koller and Friedman [25]. The main point in the previous paragraph is that the induction of a network structure is formally a hard combinatorial problem [26], and as we see below, it has consequences for the classification task.

BAYESIAN NETWORKS AS CLASSIFIERS

One of the most important parts of the classification process is the evaluation of the performance of a given classifier. A natural and very common way to do this is based on estimating the *error rate* of a classifier on a set of instances taken from a data set D . The classifier predicts the class of each instance: if it is correct, it qualifies as a success; if not, it is an error. The error rate is just the proportion of errors made over a set of trial instances. This particular error rate is known as the 0–1 loss error rate, and it essentially entails that $r_{00} = r_{11} = 0$ and $r_{01} = r_{10} = 1$ in Equation (1).

Getting back to the automated learning of a Bayesian network classifier from data.

Based on the discussion above, a reasonable strategy consists of using the methods in the previous section to fit the distribution/density $p(\mathbf{F}|C)$, and then use Equation (1) to make decisions. In Friedman *et al.* [9] the author of this article and colleagues applied such a strategy to 25 data sets that were considered to be benchmarks at that time. We then compared the error rate of the Bayesian network classifiers thus created to ones that have the basic structure as displayed in Fig. 1. A classifier that follows that structure is called a *Naive Bayes classifier* (NBC), because it assumes that each feature F_i is independent of the rest of the features in $\mathbf{F}$, given the class C . Our original hypothesis when we started that experiment was to use the NBC as a baseline. To our surprise, we found the performance of the NBC to be superior (on average) to that of classifiers with general Bayesian network structure. We first examine the main reasons for these results, and then review on a strategy to improve on the NBC.

Most of the challenges of inducing good classifiers have as root the unavoidable problem that the data at hand is finite. All estimators provide strong guarantees of convergence to the actual distribution in the data as the number of samples available tend to infinity. This is hardly the case in practice. Moreover, the demand for a larger number of samples grows exponentially with the dimensionality of the feature space. This limitation is related to what Bellman called *the curse of dimensionality* [27].

With this in mind, we remind the reader that even though we are following the dictates of BDT when making a decision based on a Bayesian network classifier using Equation (1), we cannot guarantee the optimality of the error rate as we do not have the distribution p at the time of applying this rule; we have an *estimate* $\hat{p}$ of the required distribution.

The error in the estimate of $\hat{p}$ can come from multiple sources. There are of course issues with selecting the wrong parameterization (or functional form) when estimating the parameters of p . Most importantly, recall that the induction of the structure of the network requires a search process over a

really large combinatorial space. In addition, there are large sets of networks that are equivalent with respect to the statements of independence that they encode, yet they present different graph structures [28]. Thus, it may be the case that during the automated fitting of $p(F_1, \dots, F_n, C)$, the algorithms perform a really good job in terms of the score selected to evaluate the network because it is fitting the parameters of maximum likelihood of a particular subspace of the joint distribution, while ignoring more important subspaces that are more relevant for the error rate. One way to address this problem is based on the following considerations. By rearranging the terms of Equation (1), and assuming the 0-1 loss error rate, we note that we are still maintaining the optimality in our decision if we use the following rule: decide c^0 if and only if $\frac{p(c^0|f_1, \dots, f_n)}{p(c^1|f_1, \dots, f_n)} > 1$. This begs the following question: why not fit $p(C|F_1, \dots, F_n)$ directly? Indeed this is what logistic regression does [8]. The trade-off is that one misses the advantages of having a full Bayesian network classifier model (stated in the introduction).

There is a large literature explaining the success of the NBC given its strong assumption of independence among features, even in domains where this assumption does not hold [1,10,29]. The analysis is subtle and nontrivial. Here, we provide some intuition and recommend the readers to the papers just cited for the formalization of these arguments. First, we remark that because it is fitting components of the form $p(F_i|C)$ it is guaranteed to maintain the focus of the statistical fitting process on the relationship between the features and the class. Thus, it does not run the risk of falling into an “uninteresting” region of the joint density between features, during the search for structure mentioned in the previous paragraph. Second, we note that it is very robust to the inherent variance in the data, and even though it does not fit a precise probability of the domain, it is able to find a good separating surface for the data. We illustrate this point following an example taken from Duda and Hart [1]. Consider fitting a polynomial to a set of points. Fitting a straight line may not follow the exact location of all the points; yet, it

will find the (linear) trend in the data and it will be robust to noise and general variance in the data samples. One would have to have a large set of points in order to attempt a larger degree polynomial, and this fitting would be more sensitive to variations and noise. Similarly, the NBC will find a highly biased separation between the classes (given by its fixed structure, which is related to a linear separator [1]), but it will be robust to noise, and unimportant variations in the (always) limited sample of data. Because of the high bias, this classifier does not produce calibrated probabilities on the classes given a random sample, and thus the probabilistic output should be handled with care [30]. The terms “bias” and “variance” have been introduced in a very informal fashion, yet they constitute mathematical decompositions of various error measures. Formal analyses can be found in Domingos and Pazzani [10] and Friedman [31].

A well-known approach for improving the performance of the NBC is based on keeping the general structure of the naive Bayes, but enhancing its modeling of dependencies between the features by allowing a limited set of dependencies between them. The idea is to maintain the desirable properties of the NBC, namely, low variance, and no search for structure, while improving on the modeling assumptions of independence. One such approach is called TAN (tree augmented naive Bayes) [9], and it is based on an algorithm proposed in Chow and Liu [32]. This algorithm finds a tree of dependencies among a set of random variables in polynomial time. This tree is guaranteed to be the tree of maximum likelihood with respect to the data among all possible trees. The resulting structure of the classifier is depicted in Fig. 2. It was shown empirically, on the relevant set of benchmarks at that time, that (on average) the TAN classifier performs better than the NBC [9]. Since then, there has been a large number of proposals for Bayesian network classifiers exploring different trade-off and benefits, improving on TAN and NBC, and also applying these in different domains and contexts. TAN and NBC remain extremely popular among both researchers and practitioners.

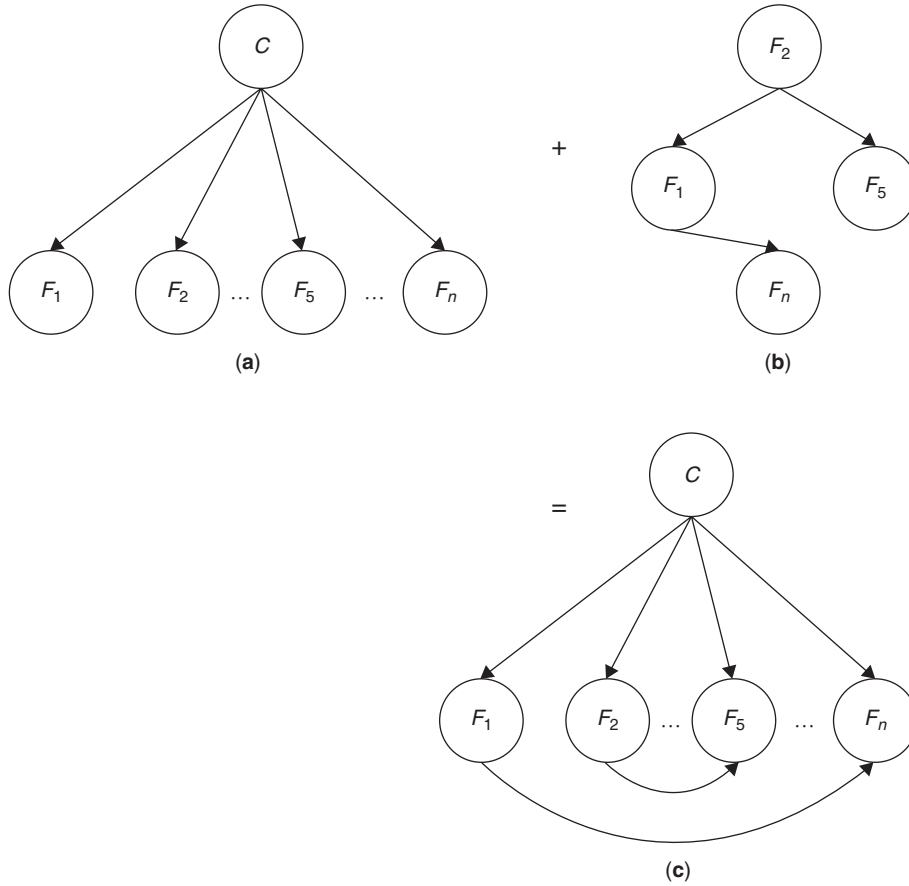


Figure 2. The combination of a naive Bayes structure (a) with a tree dependency structure on the features (b) yields a tree augmented naive Bayes (TAN) network structure (c).

Some references are given in the section titled “Final Remarks,” yet the reader is encouraged to look on-line for the latest approaches and improvements.

FINAL REMARKS

Besides defining an error rate, the evaluation and comparisons of different classifiers require the emphasize of such rate. In the selection of both the proper criteria for evaluation and the right methodology for estimation, the field of pattern classification borrows heavily from statistics. A good general introductory book that focuses on the pattern classification task is Witten and Frank [33]. More in-depth coverage of

topics and issues such as hypothesis testing, cross validation, Receiver operating characteristic (ROC) curves, and model selection can be found in Hastie *et al.* [3], Wasserman, [16], Efron and Tibshirani [34] Demšar [35], Kohavi, [36], and Lachiche, [37].

The induction and proper evaluation of a classifier are just two of the tasks involved in the enterprise of actually applying the classifier in a given domain. There are the data collection task and the feature extraction task. For example, once the raw data about the sizes of the queues of ten thousand disks in the data center are collected, we might decide to compute the median and use the deviation of the median as a feature for our classifier. There is also feature transformation, which may involve normalization

and other mathematical operations on the features that may include the transformation of the whole space using kernels [2,4,7]. Other tasks include dimensionality reduction and feature selection, which are essentially about eliminating features that are irrelevant or may introduce noise in the induction process [1,3].

With the continuous increase in computational power, the availability of data, and the continuous improvement in data collection and storage techniques, there is a strong interest in methods and techniques for automatically extracting actionable information from data. Classifiers play a crucial role in this endeavor, and Bayesian network classifiers in particular are used extensively both in academia and in a wide range of industrial applications and domains. The reasons stem from the ease of implementation, interpretability of the resulting model, the transparency by which users can understand and analyze their performance, and the relative ease to extend the models in several ways and adapt them to different domains. All of these factors are derived from the fundamental representational properties of Bayesian networks, which are listed in the introduction. At this point, it should be clear that a classifier can be induced from data, starting from a given structure, or from a partial structure, and the same applies to the specification of the parameters (and in the case of Bayesian statistics, from priors on all these). This covers the first two advantages given in the introduction. The last three advantages stem from the fact that a Bayesian network classifier is a complete distribution of the domain (classes and features). So, the computation of one of the variables in terms of the other is a matter of applying the well-known rules of probability theory (including those cases where the data contain missing values [12,13]).

A complete list of applications and domains where Bayesian network classifiers play a crucial role is beyond the scope of this article. We will be content to mention just a few for illustration purposes. Early work on automated spam filters, for example, relied on naive Bayes [38]; more recent work in the domain of natural language processing is

described in Peng *et al.* [39], which augments the basic NBC with statistical language models for diverse tasks such as authorship attribution, text genre classification, and topic detection (in several languages). Applications to computer systems (similar to the example used in this article) are described in Cohen *et al.* [14] and Moore and Zuey [40]. In Cohen *et al.* [14], the authors use the modeling capabilities of the Bayesian network classifiers to point to likely indicators of performance problems for diagnosis. In Moore and Zuey [40], the NBC is used to automatically classify traffic in computer networks for security monitoring, accounting, and quality of service. There are plenty of examples in other domains such as biology and bioinformatics [41–43], medicine [44,45], computer vision [46–48], and finance [49,50]. This brief enumeration does not pretend to be either exhaustive or authoritative. It is simply designed to show the breadth of applications and is focused on publications in the last six years to highlight the freshness of the techniques. The reader is advised to consult many of the available on-line sources when looking for a particular domain and/or application.

REFERENCES

1. Duda R, Hart P. Pattern classification and scene analysis. New York: Wiley; 1973.
2. Fukunaga K. Introduction to statistical pattern recognition. New York: Academic Press; 1990.
3. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer; 2001.
4. Bishop C. Pattern Recognition and Machine Learning. New York: Springer; 2006.
5. Bishop C. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
6. Breiman L, Friedman J, Stone C, *et al.* Classification and regression trees. London: Chapman and Hall; 1984.
7. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning methods. Cambridge, UK: Cambridge University Press; 2000.
8. Hosmer D, Lemeshow S. Applied logistic regression. New York: Wiley; 2000.
9. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29:131–163.

10. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach Learn* 1997;29:103–130.
11. Heckerman D, Geiger D, Chickering M. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach Learn* 1995;20:197–243.
12. Heckerman D. A tutorial on learning Bayesian networks. In: Jordan M, editor. *Learning in graphical models*. Dordrecht, The Netherlands: MIT Press; 1997.
13. Friedman N. Learning belief networks in the presence of missing values and hidden variables. *ICML '97: Proceedings of the International Conference on Machine learning*. Nashville (TN): ACM; 1997.
14. Cohen I, Goldszmidt M, Kelly T, *et al.* Correlating instrumentation data to systems states: a building block for automated diagnosis and control. *OSDI '04: Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation*. San Francisco, (CA): USENIX Association; 2004.
15. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference causality: models, reasoning, and inference. San Francisco, (CA): Morgan Kaufmann; 1988.
16. Wasserman L. *All of statistics*. New York: Springer; 2004.
17. Gelman A, Carlin J, Stern H, *et al.* *Bayesian data analysis*. London, (UK): Chapman & Hall; 1995.
18. Casella G, Berger R. *Statistical inference*. Pacific Grove, (CA): Duxbury Press; 2001.
19. Bernardo J, Smith A. *Bayesian theory*. New York: Wiley; 2000.
20. Gilks W, Richardson S, Spiegelhalter D. *Boca Raton, (FL): Markov chain Monte Carlo in practice*. Chapman & Hall; 1995.
21. Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach Learn* 2003;50:95–125.
22. Chickering DM. Optimal structure identification with greedy search. *J Mach Learn Res* 2002;3:507–554.
23. Spirtes P, Glymour RSC. *Causation, prediction, and search*. New York: Springer; 1993.
24. Pearl J, Verma T. *A theory of inferred causation. Knowledge representation and reasoning*. Cambridge, (MA): Morgan Kaufmann; 1991.
25. Koller D, Friedman N. *Probabilistic graphical models: principles and techniques*. Cambridge, (MA): MIT Press; 2009.
26. Chickering DM, Heckerman D, Meek C. Large-sample learning of Bayesian networks is np-hard. *J Mach Learn Res* 2004;5: 1287–1330.
27. Bellman RE. *Adaptive control processes*. Princeton (NJ): Princeton University Press; 1961.
28. Chickering DM. Learning equivalence classes of Bayesian-network structures. *J Mach Learn Res* 2002;2:445–498.
29. Hand DJ, Yu K. Idiot's Bayes: not so stupid after all? *Int Stat Rev/Rev Int Stat* 2001;69:385–398.
30. Cohen I, Goldszmidt M. Properties and benefits of calibrated classifiers. *European Conference on Principles and Practice of Knowledge Discovery in Databases*. Pisa, Italy: Springer; 2004.
31. Friedman J. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Min Knowl Disc* 1997;1:79–119.
32. Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. *IEEE Trans Inform Theor* 1968;14:462–467.
33. Witten I, Frank E. *Data mining, practical machine learning tools and techniques*. San Francisco, (CA): Elsevier; 2005.
34. Efron B, Tibshirani R. *An introduction to the bootstrap*. Boca Raton (FL): Chapman & Hall; 1993.
35. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006;7:1–30.
36. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*. Montreal, Canada: IJCAI; 1995.
37. Lachiche N, Flach P. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. *20th International Conference on Machine Learning (ICML03)*; Washington, (DC): 2003.
38. Sahami M, Dumais S, Heckerman D, *et al.* A Bayesian approach to filtering junk e-mail. *Learning for Text Categorization—Papers from the AAAI Workshop*; Madison, (WI): 1998.
39. Peng F, Schuurmans D, Want S. Augmenting naive Bayes classifiers with statistical language models. *Inform Ret* 2004;7:317–345.

40. Moore A, Zuey D. Internet traffic classification using Bayesian analysis techniques. Proceedings of the International Conference on Measurement and Modeling of Computer Systems. Banff, Canada: ACM SIGMETRICS; 2005.
41. Jansen R, Yu H, Geenbaum D, *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;302:449–453.
42. Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* 2000;301:1059–1075.
43. Myers C, Troyanskaya O. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* 2007;23:2322–2330.
44. Kazmierska J, Malicki J. Application of the naive Bayesian classifier to optimize treatment decisions. *Radiother Oncol* 2008;86:211–216.
45. Gaag LC, Renooij S, Feelders A, *et al.* Aligning Bayesian network classifiers with medical contexts. *MLDM '09: Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition*. Leipzig, Germany: Springer; 2009.
46. Schneiderman H. Learning a restricted Bayesian network for object detection. Computer Society Conference on Computer Vision and Pattern Recognition; Washington, (DC): 2004.
47. Rehg J, Pavlovic V, Huang T, *et al.* Special section on graphical models in computer vision. *IEEE Trans Pattern Anal Mach Intell* 2003;25:785–786.
48. Li LJ, Socher R, Fei-Fei L. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. *Computer Vision and Pattern Recognition (CVPR)*; Miami, (FL): 2009.
49. Baesens B, Verstraeten G, Poel D, *et al.* Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *Eur J Oper Res* 2004;156:508–523.
50. Baesens B, Gestel TV, Viaene S, *et al.* Benchmarking state-of-the-art classification algorithms for credit scoring. *J Oper Res Soc* 2003;54:627–635.

BEHAVIORAL ECONOMICS AND GAME THEORY

MIGUEL A. COSTA-GOMES
University of Aberdeen Business
School, Aberdeen, UK

Game theory is a powerful tool to analyze interactions among a small number of agents. The main goal of game theory is to make predictions about which strategies agents play. Its main solution or prediction concept is Nash equilibrium. A combination of players' strategies (one for each player) is a Nash equilibrium if each player's strategy is optimal in the sense of maximizing her expected payoff, given the others' strategies. An example helps to illustrate this concept of equilibrium.

Consider a static (or simultaneous-move) game described in detail in Costa-Gomes and Weizsäcker [1]. It is a two-person three-by-three payoff matrix game (Fig. 1). The numbers in each cell denote the numbers of points (referred to as *payoffs*) that each of the two players earns in the cell's corresponding outcome.

In this game, the only Nash equilibrium is {B, R} because it is the only combination of actions in which each of the players plays the strategy that earns her the most points given the other player's strategy. Thus, none of the players can increase her number of points by changing her strategy given the other's strategy.

However, the concept of Nash equilibrium is sometimes a poor predictor of human behavior, as documented by Costa-Gomes and Weizsäcker [1], among many others. They asked a group of human subjects who were assigned the roles of Row and Column to play this game once. Costa-Gomes and Weizsäcker observed that out of 66 Row subjects, although 38 played B, their equilibrium action, 7 played T and 21 played M. Only 13 out of 62 Column subjects played their equilibrium action, R, and 24 played L while 25 of them played M.

The purpose of this article is to briefly introduce the modeling approaches that have been proposed to explain why behavior in games deviates from equilibrium.

The search for explanations has relied to a great extent on *behavioral economics*. Behavioral economics acknowledges the fact that rationality is a very strong assumption when modeling people's decisions and relaxes it by paying close attention to a variety of empirically based concepts in psychology. The use of the behavioral economics approach to understand how people play games has resulted in a new variety of game theory, which many refer to as *behavioral game theory*, a term coined by Camerer [2,3]. Although, behavioral game theory's mathematical apparatus is firmly grounded in game theoretical concepts that are used to describe and analyze games such as utility, beliefs, and decision-theoretic rationality, it departs from it by selectively incorporating different ideas from psychology that give rise to different kinds of explanations for the observed deviations from equilibrium.

It is important to note that most of the empirical evidence on the limitations of equilibrium in describing behavior is gathered from studies where people, typically university students, take on the roles of players in a game, and receive monetary payments that are directly related to the point payoffs they earn in the games they play. Such studies are known as *economics experiments*, a term used to describe the use of experimental methods to study situations with economic content. Economics experiments allow for tests of the behavioral predictions of equilibrium as long as one is willing to assume that human subjects decide according to the preferences of the players in the game that is described to them. In other words, tests are possible as long as the use of monetary rewards leads to the control of subjects' preferences. Experiments are useful because by changing the game being played in different ways it is possible to observe and document systematic patterns in behavior.

		Column		
		L	M	R
Row	T	30 47	32 94	38 36
	M	69 38	83 81	20 27
	B	58 80	11 72	67 63

Figure 1. Costa-Gomes and Weizsäcker's game 8.

This article briefly summarizes the broadly defined categories of behavioral explanations and their links to ideas in psychology, whenever appropriate.

GAME THEORETIC REASONING

To better understand why people do not always play equilibrium strategies, one has to explore the logic that underpins game theoretic reasoning, and why in some games one has to use concepts that go beyond equilibrium to narrow down predictions about play.

To do this, we revisit the game described above and present an example of a dynamic (sequential) game along with some play data.

How does one identify the equilibrium (or equilibria, in case there is more than one) in a game? There are different ways of doing this.

One way is to consider each of the possible combinations of strategies and test whether any of the players would want to play something else given the others' strategies so as to increase her payoff. An equilibrium is found when none of the players wants to change her strategy given the others' strategies, that is, when we end up with the same combination of actions that we started with. Another way is known as *best-response dynamics*. This procedure is easy to describe in a game involving two players. One starts from a hypothetical strategy for one of the players and then determines the other player's best-response, that is, the action of the other player that maximizes her payoff. This procedure is iterated until players' best-responses stop changing.

Both of these procedures illustrate the circularity logic of equilibrium. Although equilibrium is a beautiful logically consistent concept, its circularity makes it an unnatural cognitive process.

In some games it is possible to use a different procedure to find a game's equilibrium, or to simply narrow down the set of strategies for each player that can be part of an equilibrium. Such a procedure is known as *iterated dominance*, or *iterated elimination* of strictly dominated strategies. A player's strategy is dominated by another of her strategies if it yields a strictly lower payoff for each of the other player's strategies. A rational player does not play a dominated strategy, and therefore it can be disregarded or eliminated. Iterated dominance is the process of elimination of dominated strategies of all players in an iterative manner. The game in Fig. 1 is used to illustrate iterated dominance and its logic.

In this game, M is dominated by T for the Row player, because she can earn more points by choosing the latter strategy over the former one, regardless of the Column's choice. A rational Row player will not choose M. By assuming that Column knows that Row is rational, we can conclude that a rational Column player will not choose M or L herself, because each earns her fewer points than R, independently of whether Row plays T or B. If we further assume that Row knows both that Column knows that Row is rational and that Column is rational, Row can iterate the deletion of conditionally dominated strategies one step further, and conclude that B will earn

him more points than T, given that Column plays R. In this game the use of iterated dominance leads to a unique strategy for each of the players, B and R, which together constitute the Nash equilibrium strategy profile of the game. This example shows how iterated dominance is related to assumptions about players' rationality, their knowledge of each other's rationality, and so forth. This example illustrates the delicate nature of iterated dominance, and that it breaks down when any of its assumptions does not hold.

Now, an example of a dynamic game to show that sometimes game theory needs to go beyond the idea of equilibrium to produce more precise predictions is presented.

The dynamic game is a three-stage alternating offers-bargaining game presented in Johnson *et al.* [4] and depicted in Fig. 2. In this game, two players have to agree on how to divide an amount of money that shrinks by half from one stage to the next in case of no agreement. The first mover is player 1 who makes an offer, x_1 , to player 2, an amount between \$0 and \$5, who decides whether to accept or reject it. If player 2 accepts the offer, the game ends with her earning x_1 and player 1 ($5 - x_1$). Otherwise, after her rejection, she gets her turn to make an offer, x_2 , to player 1, an amount between \$0 and \$2.5, who decides whether to accept or reject it. Acceptance by player 1 leads to earnings of x_2 for player 1 and ($2.5 - x_2$) for player 2. Rejection by player 1 earns him the right

to make the final offer, x_3 . The game ends with either acceptance, with players 1 and 2 earning $(1.25 - x_3)$ and x_3 , respectively, or rejection that earns both players \$0.

Because this game has many equilibria, a more precise prediction requires one or some equilibria to be selected over the others. Thus, game theory has to go beyond the concept of equilibrium. In sequential games where players observe each other's moves, the concept used is backward induction. The analysis of the game starts at the last stage and proceeds backward to the first stage. At each stage of the procedure, the player who has to choose plays the action that gives her the highest payoff.

The working of the backward induction is illustrated here by applying it to this game. In the last stage, player 2 is indifferent between accepting or rejecting \$0 and accepts \$0.01 rather than rejecting it and earning \$0. Thus, player 1 offers \$0 (when player 1 accepts \$0) or \$0.01 (when player 1 rejects \$0) to player 2 in the third stage which she accepts, and keeps \$1.24 or \$1.25 to himself. Since player 1 is guaranteed \$1.24 or \$1.25 in the last stage, in the second stage player 2 offers between \$1.24 and \$1.26 to player 1, which he accepts. Player 2 keeps between \$1.24 and \$1.26 to herself. Consequently, in the first stage player 1 offers an amount between \$1.24 to \$1.27 to player 2, which she accepts, keeping between \$3.73 and \$3.76 to himself.

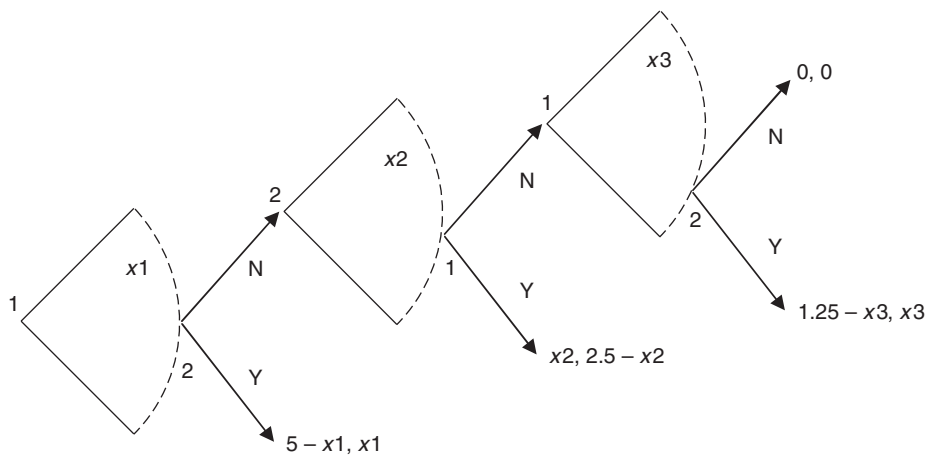


Figure 2. Johnson, Camerer, Sen, and Rymon's three-stage alternating-offers bargaining game.

However, these predictions differ from how people play this game. In one of the experimental treatments reported in Johnson *et al.* [4], the average opening offer is \$2.11, and player 2 rejects offers 10.8% of the time. In the second round, player 2 makes disadvantageous offers (that is offers that earn her less than the amount offered by player 1 in the first stage) 85% of the time. Player 1 rejects 21% of the second stage offers, resulting in him making a final offer that is rejected by player 2 two-thirds of the time.

In short, the data shows that backward induction equilibria offers are observed infrequently and that in many games players' strictly positive offers in all three stages are rejected. But even in a Nash equilibrium not necessarily consistent with backward induction, one of the three offers has to be an equilibrium one, which is then accepted. These and other similar results in sequential games challenge equilibrium and refined versions of it such as backward induction as good predictors of play.

WHAT IF NOT EQUILIBRIUM?

Broadly defined, there are three nonmutually exclusive ways to explain behavioral deviations from equilibrium. One of them links nonequilibrium play to the nonstrategic elements of the decision process, often using decision-theoretic models of bounded rationality. Another interprets the deviations as the manifestation of people's bounded rationality in analyzing the strategic elements of the decision situation. Finally, a third one relies on the idea that players' preferences go beyond the material payoffs they earn in the game. We now provide brief descriptions of the different approaches sometimes using empirical evidence from the games discussed above to substantiate them.

Models of Play When Playing Nature Rather than Others

One class of explanations puts aside the strategic aspect of the players' interaction and approaches the player's decision process as if she faces a dummy player (referred as Nature) whose actions are given by a

known probability distribution, because no attempt is made to explain how the player arrives at that probability distribution that represents the beliefs she holds about the play of her opponents. Mostly, these explanations revolve around departures of the standard model of a decision maker (a subjective expected utility maximizer with a certain attitude toward risk). Under the assumption that a player's utility is a function of her own monetary earnings (which add to her wealth) and that she is risk-neutral, this model predicts that a subject plays the action or strategy that maximizes her expected monetary earnings given her probabilistic belief about the actions of her opponent. Therefore, when only actions (but not beliefs) are observed, failures of compliance with subjective expected utility maximization in a game are hard to detect, except for the play of dominated actions. Such limitation has prompted researchers to resort to the elicitation of beliefs to gather data to complement action data in order to better document failures of subjective expected utility maximization in games. For example, Costa-Gomes and Weizsäcker [1] analyze actions and stated beliefs jointly, finding that in the game above and another 13 similar games more than one-third of the time subjects do not play the action that maximizes their expected monetary earnings given their stated beliefs. In the context of Costa-Gomes and Weizsäcker [1]'s experimental design the observed inconsistencies between actions and stated beliefs are a rejection of the joint hypothesis that a subject is risk-neutral and expected utility maximizer. However, as stated by Costa-Gomes and Weizsäcker [1], stated beliefs cannot be taken at their face value because elicitation mechanisms are truth revealing only under particular assumptions; they are responses to the monetary incentives in experiments and therefore when reporting other than the true belief, a subject loses money in the same manner as he loses by not playing the action that maximizes its subjective expected utility; they might not be accessible at all because some subjects might not hold beliefs. Hence, although the evidence generated by such inconsistencies prompts researchers to search for models that can

rationalize behavior, the net has to be cast wide. We now briefly describe some of those attempts.

A simple explanation questions the assumption that subjects are risk-neutral even though they are expected money maximizers. Although equilibria in pure strategies are insensitive to players' attitudes toward risk, risk aversion can explain the deviations from the unique mixed-strategy equilibrium in generalized matching pennies games, as shown by Goeree *et al.* [5]

A different dimension of the decision process deals with the possibility that subjects in experiments exhibit ambiguity in the sense that they cannot assign probabilities to the possible outcomes of the random variable that describes the choice of their opponent. Ambiguity makes the probabilities that subjects assign to the play of their opponents to not necessarily sum up to one. Instead, they can add up to less than one in the case of ambiguity aversion or to more than one in the case of ambiguity loving (see for example, the formulation of ambiguity proposed by Schmeidler [6]). Camerer and Karjalainen [7] find that a considerable fraction of their subjects exhibit ambiguity aversion in simultaneous-move games.

Prospect theory, a concept coined by cognitive psychologists Kahneman and Tversky [8] has also been used to explain game playing by experimental subjects. Its core ideas are that individuals treat gains and losses differently and weigh probabilities in a nonlinear manner. In this theory, an individual evaluates outcomes in relation to a reference point that reflects her aspiration level which could arise out of adaptation or experience. Thus, the individual assigns values to the monetary payoffs in a way that reflects their position in relation to the reference point. In addition, while in the loss domain the marginal utility increases as the outcome improves, in the gain domain the marginal utility decreases. The individual is risk-averse above the reference point, but risk-loving below it. Furthermore, values are assigned in a way that the positive value of a gain is smaller (and according to many estimates not far from a half) than the absolute value of an equal sized loss, a feature known

as *loss aversion*. Besides evaluating outcomes the individual also weighs probabilities. The effect of probability weighting on individuals is that they may end up perceiving them differently from their objective values. The empirical evidence suggests that people overweight small probabilities but underweight large probabilities. Prelec [9] provides an axiomatic derivation of a parametric weighting function. Goeree *et al.* [5] combine nonlinear probability weighting with risk aversion although not considering the other features of prospect theory in their data analysis of generalized matching pennies games. However, they find that the weighting function inferred from the data does not exhibit the "inverted" S-shape the theory posits it has.

Boundedly Rational Models of Strategic Thinking

In a different class of models special attention is paid to the modeling of players' beliefs about the play of their opponents. The different specifications of players' beliefs reflect the varying degrees to which players' decisions take into account the incentives of their opponents. Although in some models beliefs are tied down to equilibrium reasoning, in others they simply reflect players' boundedly rational strategic thinking, or as ascribed by them to their opponents. In addition, very often these models consider a nonstrategic dimension of bounded rationality as they consider that the act of best-responding is itself subject to errors.

One of the leading models is McKelvey and Palfrey's [10] quantal response equilibrium (henceforth QRE). Its boundedly rational interpretation results in a generalization of the concept of Nash equilibrium by replacing best-response behavior with better-responding behavior. While in a Nash equilibrium each player always plays an action that earns her the highest expected payoff given the behavior of the other players, in a QRE, players play actions with higher expected payoffs more frequently than actions with lower expected payoffs. That is, in the latter concept an expected payoff maximizing action is not always played, but is played more frequently than actions with lower expected payoffs. The

rationale for it is that people are imperfect expected utility maximizers because they make mistakes, which are more frequent the smaller their costs are. In addition, in a QRE, players model their opponents as being exactly as imprecise decision makers as they themselves are, and take this into account when computing the expected utility of their actions; the same is true of their opponents. Although, the theoretical construct is relatively general, the choice specification that embodies mistakes that is mostly used to fit the model to the data is the logit choice rule, which results in the concept of logit quantal response equilibrium.

Weizsäcker [11] adapts QRE by allowing players to be wrong about the level of imprecision of others, and calls it asymmetric QRE. Although players' reasoning is circular as required by the equilibrium, because they are allowed to be wrong about the level of imprecision of others, their beliefs about their opponents' behavior can be incorrect.

Interestingly, the QRE concept lends itself to a different interpretation. Here, the players are not boundedly rational, but are playing a Bayes–Nash equilibrium of a game different from the one that is represented by the material payoffs, because of perturbations to such payoffs that can be seen as accounting for nonpecuniary payoff considerations, and which can arise for many reasons such as mood effects, and so on. The concept of QRE has been extended to dynamic games by McKelvey and Palfrey [12].

A different strand of the literature posits that people use rules of thumb that are not cognitively taxing. In general, these rules dispense with the circularity of equilibrium reasoning because players' models of their opponents make them redundant. In these models, players are self-interested material payoff maximizers and hold beliefs about their opponents which emerge out of the rule of thumb they follow. Its leading example is what is known as *level- k thinking*, which is a model of iterated best-responses usually anchored on a prior belief. Although in many applications the anchoring belief is the uniform prior, as suggested by the intuitive cognitive principle of insufficient reason (with the justification that the player is either

unable to put herself in the shoes of her opponent or does not know that her opponent is rational at all), in general, the anchoring belief captures a player's instinctive reaction on how to play the game. A level-1 player is someone who best responds to a uniform random opponent called level-0, while a level-2 player best responds to an opponent who is best-responding to someone who randomizes uniformly, that is, it best responds to a level-1 player, and so forth. Low level- k thinking can mimic equilibrium when the game is dominance solvable in a few steps. For example, in the game in Fig. 1, a level-3 Row player or higher plays B, the equilibrium action, while a level-2 or higher Column player plays the equilibrium action R. Different versions of the level- k model take different views on how a level- k player models his opponents. While some, such as Costa-Gomes *et al.* [13] and Costa-Gomes and Crawford [14], consider that a level- k models her opponents as being level $k - 1$ (like Nagel's [15] nonstructural level- k model), thus opting for simplicity of the rule of thumb, others such as Stahl and Wilson [16,17] and Camerer *et al.* [18] consider that she believes that her opponents are heterogeneous with some behaving as level-0, others as level-1, and so forth, up to level $k - 1$, according to the conditional relative frequencies of the different levels or types given that she is level- k . Consequently, the calculation of level k 's best-response cannot be reduced to the outcome of k iterated best-responses to a uniform prior. Although the latter view of the level- k model makes it a more cognitively taxing model, it usually moves the behavior of higher levels closer to the aggregate behavior of the lower levels in the game being played.

The typical experiment has each subject playing a series of games with her sequence of decisions used to identify the rule of thumb, if any, that she uses when playing novel games. The usefulness of the basic level- k model is confirmed by data gathered using information search tracking methods that shed light on the cognitive processes subjects use while making decisions. In fact, level- k rules have been compared to other rules of thumb such as iterated deletion of dominated strategies, or D- k rules, which are more closely aligned

with game theoretic reasoning [13,14]. Subjects' decisions and information search when analyzed separately or jointly reveal that D- k rules are not good descriptors of behavior.

The level- k model has been extended to games of incomplete information by Crawford and Iriberri [19], who use it to explain behavior in private-value and common-value actions. Its extension to dynamic games raises lot of issues (e.g., how to model the fact that a fraction of subjects do not even look up the final-stage elements of the game such as payoffs before they initiate play, as shown in Johnson *et al.* [4]) and although no general way of doing so has yet emerged, Crawford [20] and Östling and Ellingsen [21] use it to analyze communication in games.

Models of Nonpecuniary Payoffs

There is a class of models that posits that material payoffs do not fully capture subjects' preferences over outcomes. Its core idea is that elements of the game other than a subject's material payoffs have an effect on her preferences in a systematic rather than a random manner. Therefore, modeling such effects is a worthwhile endeavor as it can help to predict behavior in novel games. In other words, by uncovering how the different features of a game influence subjects' preferences, control of subjects' preferences can still be achieved, even if it is not solely driven by own material payoffs.

Nonpecuniary payoff considerations can influence subjects' preferences over the game's outcomes in different ways: the material payoffs of a player's opponents in an outcome can influence her payoff in that outcome; a player's payoff in an outcome can be a function of procedural concerns; a player's payoff in an outcome can be a function of her own or her opponents' material payoffs in other outcomes; a player's payoff can depend on her belief about the actions chosen by her opponents; a player's payoff can also depend on high order beliefs such as what she believes about what her opponent believes she is going to play. Sobel [22] provides a description and comparison of these different approaches.

The class of outcome-based models of preferences defines a player's utility or payoff

in an outcome as a function of the player's own material payoff and some or all of the other players' material payoffs in that outcome. Models such as those of Fehr and Schmidt [23] and Bolton and Ockenfels [24] explore the idea that people are averse to the inequality of players' material payoffs, both when they are at an advantage and when they are at a disadvantage compared to others. These two models differ on players' perceptions of inequality: according to Fehr and Schmidt [23] a player has a self-centered view of inequality and compares her material payoff with those of all the other players, while according to Bolton and Ockenfels [24] a player compares her material payoff with the average material payoff of the other players. Charness and Rabin [25] add efficiency considerations as well as concerns for the less well-off individual (giving rise to maximin preferences) to inequality aversion.

A variety of experiments have revealed that aggregate behavior is sensitive to all these concerns as well as to others. But, within a particular game one or a few concerns tend to be more salient than the others, as an individual's perception of what is fair is situation-dependent. Most of the clearer existing evidence comes from one-person games, that is, situations where one player chooses one of several allocations of material payoffs for herself and the others.

Another approach takes into consideration that a player's preferences over outcomes are context-dependent. While in outcome-based models a player derives the same utility in two outcomes with the same profiles of players' material payoffs; they can be different when preferences are context-dependent because they can be influenced by the player's or her opponents' material payoffs in other outcomes. Striking evidence for such preferences is provided in Falk *et al.* [26]. This study compares two games that are identical to each other except that one of them has two additional outcomes. They find that subjects' aggregate choices over a pair of outcomes common to the two games depend on whether or not the game has the two additional outcomes. An intuitive and appealing example of context-dependent preferences is provided

by models that extend outcome-based models by allowing players' preferences in outcomes to depend on procedural concerns. In such models a player's preferences over two outcomes depend on the procedure that generates the situation where the player has to choose between them. Bolton *et al.* [27] provide evidence that a player is more willing to accept outcomes where she receives considerably less than her opponent if the procedure that puts her in that choice situation is unbiased, in the sense that it could put her opponent in the same exact situation with the same probability. Bolton *et al.* [27] proposes an extension of the model of inequality aversion [24] that makes players' utilities a function of the procedure.

A different class of models assumes that players' beliefs about actions or beliefs about beliefs about actions (i.e., high order beliefs) influence a player's utility. For example, a player's belief about her opponent's action can be interpreted by the player as the intention of the other toward her, namely her perception of how kind or unkind he is being toward her. On the other hand, an emotion such as guilt is better modeled through a second order belief, as explained by Battigalli and Dufwenberg [28], since guilt is an emotion related to what a player believes about what her opponent believes he will play. Although a variety of emotions can be incorporated into players' utility functions, reciprocity has been the main one modeled by researchers. In addition, conformance to norms and other forms of socially driven behavior can be usefully incorporated into players' payoffs via beliefs. The framework that has been used to do all the above is psychological game theory, which has been reinvigorated by Battigalli and Dufwenberg [28], who have developed it further, but who have also highlighted its current limitations.

MODELING ADAPTIVE DYNAMICS

Games are often played repeatedly. Although in many games play converges to equilibrium, in many others deviations, although not necessarily stable, never go away. To comprehend this diversity in behavior, researchers

have put forward models that try to capture how people adjust their behavior from one period to the next. This literature is usually referred to as the *learning in games literature*, thus effectively borrowing the term from game theoretic formal models of learning how to play equilibria. Because it is impossible to infer from the data that people are learning to play equilibria, I use the term adaptive dynamics to refer to this literature. The different generations of adaptive dynamics models show how insights from psychology have helped shape its progress.

One strand of model is belief-based. The core idea of this strand is that a player forms beliefs about the actions she expects her opponents to play in the next period on the basis of what they played in the past. She then plays the action with the highest expected payoff. Models differ from each other according to the weights attached to different periods of past play. The weights determine a player's belief about her opponent's play in the current period. In *fictitious play*, a player's distant and recent past play are weighed equally. As a result, beliefs are given by the opponent's empirical relative frequency of past play. Cheung and Friedman's [29] model of empirically weighted beliefs is a more flexible approach as the weights attached to an opponent's past play decrease exponentially with the number of periods that have elapsed, thus conforming to the psychological insight of recency, that is, the recent past counts more than the distant past. A different approach is proposed by Crawford [30] who suggests that players adjust their choice toward what would have been optimal for the last observation of the outcome with the adjustment fluctuating up and down as play proceeds. In all these models a player's behavior in a period is uniquely determined by the history of play of her opponents. An alternative class of belief-based model receives the label of *sophisticated learning*, because it has a forward flavor to it, as a player takes into consideration how her future play will influence the play of her opponents in future periods.

A different class of learning model borrows heavily from insights in psychology, and it is known as *reinforcement-learning* [31].

The main insight used is that experienced rewards are central to decision making. Therefore, experienced payoffs (the payoff a player earns by playing an action) rather than counterfactual payoffs (the payoff a player could have earned by playing a different action), are the key variables in determining how players adjust their behavior. Strategies played are reinforced, while strategies not played are not. Reinforcements can cumulate or be averaged, and define propensities that determine the probabilities with which the different strategies are played. Enriched versions of reinforcement-learning incorporate other insights from psychology: people evaluate experienced payoffs in relation to reference points or aspiration levels (which are the product of experience themselves, and hence vary), thus implying that reinforcements are defined in relation to a reference point; experimentation, that is, people now and then like to experiment and thus play actions rarely played in the past; and recency, alluded to in the discussion of belief-based models.

A third class of models, initially proposed by Camerer and Ho [32], is known as *experienced weighted attraction (EWA)*, and is a synthesis of the basic features of the belief-based and reinforcement-learning approaches. It encompasses a variety of formulations of either of them. A more elaborate version of it is called self-tuning EWA [33], because the weights evolve as play unfolds.

Other classes of models consider other aspects of adaptive behavior. For example, in symmetric games, it is quite plausible that people resort to imitating the play of others [34]. A different perspective is that of Stahl [35] who pursues the idea that people behave according to rules, switching between rules according to their relative performance.

Although different models approach adaptive dynamics from different angles, in many games they end up making very similar predictions. This observational equivalence in actions tells us that a deeper understanding of the issues might have to rely on observing and modeling other aspects of people's

decision process. A way of doing so involves tracking people's informational searches of elements of the game such as their own payoffs or their opponents' payoffs as well as the actions that their opponents play as the game unfolds, as in Knoepfle *et al.* [36]. Knoepfle *et al.* [36] study generates intriguing results because while the analysis of the information looked at by the subjects suggests that they engage in sophisticated learning (i.e., they take into account how their actions will influence the actions of their opponents in future periods), rather than adapting to the past as in standard belief-based or reinforcement-learning models or mixtures of both, the action data do not support this hypothesis. The intricacies of behavior highlighted in this study tell us that we will need to consider the different dimensions of the decision process to clarify the issues and produce a synthesis of the different approaches.

CONCLUSIONS

Tests of game theory's behavioral predictions using economics experiments have produced overwhelming and systematic empirical evidence that people play games differently from what the theory predicts. This article provides brief summaries of different modeling approaches that incorporate empirically-based concepts into the theory. Although a lot of progress has been made, the challenge for game theory remains the same: to put forward models of behavior that can explain equilibrium play in some situations and nonequilibrium in others, but that are not tailored to each and every game in a way that renders them ad hoc; instead they should be parsimonious, even if their parameters have to be appropriately fine-tuned across different strategic situations.

REFERENCES

1. Costa-Gomes MA, Weizsäcker G. Stated beliefs and play in normal form games. *Rev Econ Stud* 2008;75(3):729–762.
2. Camerer CF. Progress in behavioral game theory. *J Econ Perspect* 1997;11(4):167–188.

3. Camerer CF. Behavioral game theory. Princeton (NJ): Princeton University Press; 2003.
4. Johnson EJ, Camerer CF, Sen S, *et al.* Detecting failures of backward induction: monitoring information search in sequential bargaining. *J Econ Theory* 2002;104(1): 16–47.
5. Goeree JK, Holt CA, Palfrey TR. Risk averse behavior in generalized matching pennies games. *Games Econ Behav* 2003;45(1): 97–113.
6. Schmeidler D. Subjective probability and expected utility without additivity. *Econometrica* 1989;57(3):571–587.
7. Camerer CF, Karjalainen R. Ambiguity-aversion and Non-additive Beliefs in Non-cooperative games: experimental evidence. In: Munier B, Machina M, editors. Models and experiments on risk and rationality. Dordrecht: Kluwer Academic Publishers; 1994. pp. 325–358.
8. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
9. Prelec D. The probability weighting function. *Econometrica* 1998;66(3):497–527.
10. McKelvey RD, Palfrey TR. Quantal response equilibrium for normal form games. *Games Econ Behav* 1995;10(1):6–38.
11. Weizsäcker G. Ignoring the rationality of others: evidence from experimental normal form games. *Games Econ Behav* 2003;44(1): 145–171.
12. McKelvey RD, Palfrey TR. Quantal response equilibrium for extensive form games. *Exp Econ* 1998;1(1):9–41.
13. Costa-Gomes M, Crawford VP, Broseta B. Cognition and behavior in normal-form games: an experimental study. *Econometrica* 2001;69(5):1193–1235.
14. Costa-Gomes MA, Crawford VP. Cognition and behavior in guessing games: an experimental study. *Am Econ Rev* 2006;96(5): 1737–1768.
15. Nagel R. Unravelling in guessing games: an experimental study. *Am Econ Rev* 1995; 85(5):1313–1326.
16. Stahl DO, Wilson PW. Experimental evidence on players' models of other players. *J Econ Behav Organ* 1994;25(3):309–327.
17. Stahl DO, Wilson PW. On players' models of other players: theory and experimental evidence. *Games Econ Behav* 1995;10(1):218–254.
18. Camerer CF, Ho T-H, Chong J-K. A cognitive hierarchy model of games. *Q J Econ* 2004;119(3):861–898.
19. Crawford VP, Iriberri N. Level-k auctions: can boundedly rational strategic thinking explain the winner's curse and overbidding in private-value auctions. *Econometrica* 2007;75(6):1721–1770.
20. Crawford VP. Lying for strategic advantage: rational and boundedly rational misrepresentation of intentions. *Am Econ Rev* 2003;93(1):133–149.
21. Östling R, Ellingsen T. When does communication improve coordination? *Am Econ Rev* 2010. In press.
22. Sobel J. Interdependent preferences and reciprocity. *J Econ Lit* 2005;43(2):392–436.
23. Fehr E, Schmidt KM. A theory of fairness, competition, and cooperation. *Q J Econ* 1999;114(3):817–868.
24. Bolton GE, Ockenfels A. ERC: a theory of equity, reciprocity, and competition. *Am Econ Rev* 2000;90(1):166–193.
25. Charness G, Rabin M. Understanding social preferences with simple tests. *Q J Econ* 2002;117(3):817–869.
26. Falk A, Fehr E, Fischbacher U. On the nature of fair behaviour. *Econ Inq* 2003;41(1): 20–26.
27. Bolton GE, Brandts J, Ockenfels A. Fair procedures: evidence from games involving lotteries. *Econ J* 2005;115(506): 1054–1076.
28. Battigalli P, Dufwenberg M. Dynamic psychological games. *J Econ Theory* 2009; 144(1):1–35.
29. Cheung Y-W, Friedman D. Individual learning in normal form games: some laboratory results. *Games Econ Behav* 1997;19(1): 46–76.
30. Crawford VP. Adaptive dynamics in coordination games. *Econometrica* 1995; 63(1):103–143.
31. Roth AE, Erev I. Learning in extensive form games: experimental data and simple dynamic models in the intermediate term. *Games Econ Behav* 1995;8(1):164–212.
32. Camerer CF, Ho T-H. Experience-weighted attraction learning in normal form games. *Econometrica* 1999;67(4):837–874.
33. Ho T-H, Camerer CF, Chong J-K. Self-tuning experience-weighted attraction learning in games. *J Econ Theory* 2007;133(1):177–198.

34. Apesteguia J, Hück S, Oechssler J. Imitation: theory and experimental evidence. *J Econ Theory* 2007;136(1):217–235.
35. Stahl DO. Boundedly rational rule learning in a guessing game. *Games Econ Behav* 1996;16(2):303–330.
36. Knoepfle DT, Wang JT, Camerer CF. Studying learning in games using eye-tracking. *J Eur Econ Assoc* 2009;7(2–3):388–398.

BEHAVIORAL OPERATIONS: APPLICATIONS IN SUPPLY CHAIN MANAGEMENT

KAREN DONOHUE
ENNO SIEMSEN
Carlson School of Management,
University of Minnesota,
Minneapolis, Minnesota

How do trust and fairness factor into supply chain relationships? Why do supply chain professionals often place orders that do not correspond to the recommendations made by normative models? When does a stock-out upset customers and what impact should this have on supply chain execution?

Such questions fall under the banner of “Behavioral Operations,” a relatively new term for the study of human behavior as it impacts the performance and management of operations. While our purpose in this article is to focus on one segment of this field, namely, applications relevant to supply chain management, we begin with some background on the field as a whole.

A previous review defines behavioral operations as any study of operations “that explicitly incorporates social and cognitive psychology theory” (see Ref. 1, p. 679). Others extend its scope to include “the effects of human behavior in process performance, influenced by cognitive biases, social preferences, and cultural norms” (see Ref. 2, p. 15). Humans engage with operations in many ways: as workers performing a production or service activity, as managers setting operational policies or incentive schemes, and as customers judging and experiencing operational outcomes. While theory in the past has focused on establishing normative solutions within these contexts, behavioral operations is more descriptive in nature by aiming to deepen our understanding of these human interactions and decision making processes. It also follows the prescriptive goal of providing guidance on how to improve operations in light of this behavior.

One stream of literature within behavioral operations focuses on uncovering settings where humans make operational decisions, and interact with operational processes, in ways that differ from normative theory (see Ref. 2 for a summary). Understanding the causes of this behavior, and how they influence operational performance, is a major research goal. The cause of these deviations may be traced to specific decision biases, cognitive limitations, bounded rationality, social preferences, motivational issues, or other behavioral factors. This research draws from empirical observations made in the field or through controlled laboratory experiments. Analytical models are also used to test the impact of new behavioral assumptions and extend normative theory. Behavioral theories developed at a micro level draw heavily from findings in psychology, cognitive science, and behavioral economics, as well as other business disciplines such as consumer behavior and behavioral finance.

Another stream of research within behavioral operations examines how operations-specific policies or institutions should be structured in light of human dynamics that may have been uncovered in prior micro level research. Questions related to the design of procurement auctions, collaborative planning rules, and matching mechanisms fall into this category. This research tends to be prescriptive in nature. Here recommendations are often developed and tested analytically, based on specific assumptions of how individuals behave under the rules of the policy or institution. The robustness of the policy (and the underlying assumptions) is then tested in a series of studies, using a combination of numerical analysis and controlled laboratory experiments. Less work has been done at this more “macro” level, which we view as a fertile area for future research in behavioral operations. Excellent examples from economics include behavioral studies of alternative market designs [3] and matching mechanisms [4].

In addition to the previously mentioned reviews [1,2], two other references are useful for gaining a basic introduction to the general area of behavioral operations. The first introduces the underlying bodies of knowledge from other behavioral fields (including behavioral economics, judgment and decision making, social psychology, group dynamics, and systems dynamics) that are most germane to behavioral operations [5]. The second reference, reviews research within behavioral operations involving controlled experiments, and introduces a framework to organize this work based on the type of behavioral assumptions being investigated [6].

We emphasize that behavioral research is a substantive and not a methodological choice. Behavioral research is research that explores decisions that deviate from a normative, rational benchmark. It happens that experiments are a methodology that often fits research questions in behavioral research. Such experiments have two essential benefits [7]: they serve as a rigorous empirical pretest of theory prior to collecting field data, and they serve as an empirical feedback mechanism to further our theoretical developments. Unlike field data, experimental data is comparatively cheap. From an epistemological perspective, we can rarely adjust our theory based on refuted theoretical predictions from field data, since it is often impossible or prohibitively expensive to collect a new field data set to retest our theoretical adjustments. Experimental data, however, can be easily collected, and thus better serves as an efficient feedback mechanism in theory development. As Smith writes, “the fact that one can always run a new experiment means that it is never tautological to modify the model in ways suggested by the results of the last experiment” (see Ref. 7, p. 274).

Supply chain management is a particularly active area of application for experimental work. Of 52 publications identified as experimental research falling within the behavioral operations field in the 20 years spanning 1986–2006, more than a third can be categorized as supply chain related (see Ref. 6 for details).

While the line between the disciplines of supply chain management and operations is

somewhat blurred, for the purposes of this article, we limit the scope of supply chain management to the topics outlined in the section titled “Supply Chain Management” in this encyclopedia. This includes topics within supply chain design, product design, inventory management and control, supplier management and contracts, and supply chain collaboration. We divide our coverage of behavioral research within supply chain management into two major areas based on the unit of analysis.

The first area, outlined in the section titled “Individual Decision Making in Supply Chains,” considers the behavior of *individual* decision makers within supply chains. These include tactic decisions related to demand forecasting and inventory planning, as well as more strategic decisions involving product design. This individual decision context is characterized by a lack of direct interaction with other decision makers. Note that this does not imply that interactions do not occur, but rather that they take a back seat when compared to the individual decision biases and errors in choice that are the main focus of study.

The second area, outlined in the section titled “Interactions in Supply Chains,” examines behavior that occurs as a part of supply chain *interactions*. These include interactions within an organization and across organization boundaries (e.g., vertical interaction between a supplier–buyer or horizontal interaction across retailers or manufacturers). The development of this interaction-based behavioral research follows closely from the evolution of behavioral economics, which began with a focus on individual behavior and later developed into a subfield focused on behavioral game theory.

The section titled “Conclusion” concludes the article with a description of resources for learning more about the concepts and methodologies used in this rapidly growing field.

INDIVIDUAL DECISION MAKING IN SUPPLY CHAINS

Many supply chain activities involve making repetitive decisions at an individual level.

For example, an inventory analyst may develop demand forecasts and make order quantity recommendations for hundreds of products each month. Other decisions, such as the choice of product to launch or what design features to include, are encountered less frequently but have significant implications on supply chain performance. This section gives some background on why the behavioral assumptions of rational choice may breakdown in these settings, focusing specifically on applications in forecasting, inventory management, and product design.

Some Background: Rational versus Behavioral Decisions

The field of supply chain management is rooted in the theory of industrial organization, and so it is not surprising that it initially inherited the rational choice paradigm from this subfield of microeconomics. Rational choice proposes rationality of outcomes more than processes [8], and encompasses four behavioral assumptions: (i) that behavior is mostly motivated by self-interested and stable monetary concerns, or by the utility derived from amassing wealth; (ii) that behavior is based on conscious, cognitive, and deliberate decisions; (iii) that such decisions are based on all available information; and (iv) that these decisions optimize a given objective function. Any deviations in human behavior from assumptions (i) to (iv), for example, the study of alternative objective functions to pure monetary concerns, or the study of decisions based on heuristics rather than optimal behavior, can be seen as deviations from rational choice, and are therefore elements of a behavioral research framework.

The pursuit of such deviations of rationality became possible within economics with the seminal work of Kahneman and Tversky [9,10]. Prospect theory, which reintroduces a stronger process perspective to the analysis of decisions, essentially deviates from rational choice by, on the one hand, emphasizing the role of framing into decision making, and on the other by insisting that utility not necessarily derives from accumulating wealth, but from gaining and losing wealth. Ever since, many more such systematic deviations from rational choice have been reported,

like the anchor and insufficient adjustment heuristic, where people incorporate irrelevant information into their decisions [11], or the problem of representativeness, where people believe that the law of large numbers applies to small samples as well [12]. Interested readers are referred to the section titled “Psychological Basis of Decision Making under Uncertainty and Risk” in this encyclopedia, which examines the psychological basis of decision making under uncertainty and risk.

Applications to Forecasting

Forecasting, in specific time series analysis, is essential for the success of any planning process. Since most companies still rely extensively on human judgment when preparing forecasts [13], an understanding of how humans react to time series data is a behavioral kernel of successful supply chain management.

One stream of research that has analyzed human reactions to time series is the literature on judgmental forecasting (see Ref. 14 for a recent review). While research in this area has mostly focused on comparing the performance of humans in analyzing time series with the performance of computer algorithms, some of this research has been devoted to understanding the behavior of forecasters and analyzing their biases [15,16]. An anchor and “excessive” adjustment heuristic seems to fit the observed forecast data well [17]. The precise nature of human behavior when faced with time series, though, seems to depend on the actual time series being analyzed [18].

A related stream of research focuses on the detection of regime-shifts, that is, structural changes in a data generating process [19]. Subjects in such experiments observe a series of data, and need to decide at what time, within the series, the data generating process switches to a different distribution. Results show that sometimes people tend to overreact [20], that is, indicate a change where in fact no change has occurred, and other times underreact [21], that is, not indicate a change even though it occurred. In a recent study, these thoughts were integrated into a

system-neglect hypothesis, which supported the notion that people underreact when the observed signal is precise and the underlying series is unstable, whereas they overreact to a stable time series with a noisy signal [22].

For a more complete understanding of behavioral forecasting, it is necessary to combine these thoughts on the detection of regime change with judgmental time series analysis [23]. Single exponential smoothing adequately describes subject behavior in many time series. In unstable time series, subjects tend to mistake successive random level changes for a trend. Further, conforming with the system-neglect hypothesis, subjects in stable time series tend to overreact to their forecast error, whereas in unstable time series, they tend to underreact to it. For more information on forecasting techniques, see the section titled “Forecasting Techniques” in this encyclopedia. The article titled ***Product/Service Design Collaboration: Managing the Product Life Cycle*** also provides an overview of collaborative forecasting methods within a supply chain environment.

Applications to Inventory Planning

Much of the behavioral research on inventory planning has focused on the problem of ordering for uncertain demand, rather than on the problem of balancing fixed ordering cost with variable holding cost. Since behavioral research in economics has focused on the question of judgment under uncertainty, this focus is not surprising. In a seminal experimental study, researchers subjected graduate students to the task of finding order quantities when faced with random demand with a constant unit cost of ordering too much and ordering too little, a scenario commonly referred to as the *newsvendor problem* [24]. A central finding of this study, which has been replicated many times [25,26], is that participants in these experiments on average order too little when faced with high profit products (i.e., high underage costs), and too much if faced with low profit products (i.e., low underage costs), compared to the optimal solution. This bias appears robust to differences in framing in terms of gains and losses (i.e., the

predictions of prospect theory do not apply) [27]. While this effect decreases with subject experience and training [28], no technique has been found to completely overcome this decision bias.

Many explanations have been put forward to explain this behavioral phenomenon. Schweitzer and Cachon favor the explanations that participants have some utility of reducing *ex post* inventory error, and that they anchor their decisions upon mean demand, and insufficiently adjust toward the optimal order quantity [24]. Other experiments have shown that subjects use a multitude of different anchors if provided in the experiment [29]. Analytical research has shown that such behavior can also be explained by adding noise into the participant's decision, such that subjects tend only to stochastically favor an optimal solution, in the sense that such a solution has a higher likelihood of being chosen, without choice being deterministic in these regards [30]. This notion has also been supported with experimental evidence [31]. Other research measured these sources of error in terms of level and adjustment bias, and introduce the notion of observation bias that may occur when demand is censored (i.e., when the level of lost sales is not known) [32]. While most of this research focuses on understanding aggregate behavior across a population, the data suggests that behavior varies significantly at the individual level. This begs the question of whether there are individual characteristics or cognitive tendencies that one can use *a priori* to predict performance. Research suggests that an individual's level of cognitive reflection (measured by the Cognitive Reflection Test) is a strong predictor of newsvendor performance [33]. Interested readers are referred to the section titled “Inventory Management and Control” in this encyclopedia for more information on inventory management, including the article titled ***Newsvendor Models***, which focuses on the newsvendor problem.

Applications to Product Development

Decisions made in product development have crucial implications for supply chains, such

as determining which components will be produced by whom within the supply chain, what assembly process will be used, and how the supply chain should be configured [34]. One crucial component of modern product development is the use of cross-functional teams to speed up and better integrate the necessary information flow. As such, much research exists that analyzes the behavioral implications of such teams, such as the factors within these teams that create innovativeness [35], how working in such teams leads to job stress and less cohesion [36], and how collaboration among team members can be induced by the perceived procedural fairness of top management decisions [37].

In addition to this literature on product development teams, research has also analyzed the decisions of individuals involved in the product development process. While the field has an extensive normative literature on areas like optimal design and optimal testing, less research exists that explicitly complements this normative perspective with a behavioral point of view. With respect to the behavior of designers, an objective to establish ones reputation in an organization can lead to a preference for more complex and difficult designs than necessary [38]. Decisions of designers to rework parts of their design depend on the knowledge diversity that exists in the direct network around a designer [39]. Further, designers tend to be present-biased, causing them to postpone possibly crucial design tasks to a later time [40]. Finally, providing designers with explicit cost data focuses their decision making on cost at the possible detriment of other design objectives [41]. The section titled “Supply Chain Design” in this encyclopedia offers more information on product design, while the section titled “Product Design and Life Cycle Management” in this encyclopedia focuses on the interaction between product design and life cycle management.

Future Research

The previous examples highlight that current individual level behavioral research in supply chain management is focused on a small number of mostly stochastic decision problems like the newsvendor, whereas the

pantheon of decision contexts that exist in supply chain management is large, diverse, and often unexamined through a behavioral lens. The potential for more behavioral work in the area is therefore large.

Take aggregate production planning and strategic capacity management as an example: How do people deal psychologically with the pressure to adjust capacity to fluctuating market requirements? Will decision makers tend to overreact when faced with demand fluctuations? What is the impact of a crises and disruption on decision making in a supply chain? Will people correctly identify the benefits of pooling for supply chain design [42]? How do people value information in a supply chain? These are important supply chain questions for which little behavioral theory exists.

INTERACTIONS IN SUPPLY CHAINS

A common research question in supply chain management is how should interactions between supply chain partners be designed and managed. The rules of interaction take many forms, including formal supply contracts, data sharing schemes, and collaborative planning, forecasting, and replenishment programs. Many of these interactions contain a game theoretic dynamic where individuals attempt to maximize their own local objectives and, in the process, may undermine the performance of the supply chain as a whole. This section provides background on behavioral issues that may arise in such interactions and highlights possible applications to multiechelon inventory systems, buyer–supplier relations, and procurement markets.

Some Background: Fairness, Trust, and Game Theoretic Behavior

When two or more individuals interact, new psychological factors may arise in addition to the biases and errors in judgment found in individual decisions. Examples include other-regarding preferences (such as fairness), dependences (such as trust), and a lack of understanding of game theoretic dynamics.

The behavioral economics literature identifies many forms of fairness, including distributive fairness, distributional fairness, and peer-induced fairness [43]. The form most applicable to supply chain management is, perhaps, distributional fairness, which implies that the utility of a given supply chain member is impacted by their individual payoff as well as the payoff distribution of others. This concern for fairness is often captured through an equity aversion utility function [44]. Most studies of fairness focus on the fairness of outcomes, although some also consider fairness in terms of the intentions behind the outcomes [45] or the process involved [46]. Closely related to fairness is the concept of reciprocity, which is the act of rewarding or punishing others in response to their behavior.

Trust differs from fairness in that it implies a relationship of reliance. In developing a construct for trust previous scholars [47], define three dimensions: (i) ability: perception of the skills and competencies of the trustee; (ii) benevolence: the extent to which the trustee is believed to want to do good for the other party; and (iii) integrity: perception that the trustee is honest and fulfills its promises. The importance of each dimension varies by problem context. How trust is regulated also differs with the stage of the relationship [48]. For example, early in the relationship parties may exhibit “calculus-based” trust, regulated solely by the relative benefits of keeping a promise versus the penalties of cheating. After repeated interactions predictability is paramount, making trust more “knowledge-based.” Finally, in long-term strategic relationships trust is “identification-based,” implying that the parties have fully internalized each others desires and intentions. Interested readers are referred to the section titled “Collective Choice and Social Utility Theory” in this encyclopedia, which examines collective choice and social utility theory.

When applying game theoretic models to a problem context, one must make specific assumptions about how players will behave in order to identify whether an equilibrium (Nash or otherwise) exists for the game. In reality, humans who interact in game-like

settings usually have limited knowledge of how others will behave. Because of this, strategic decision makers need to make a prediction about others behavior and then devise a “best response” strategy based on their beliefs. This raises several interesting questions for laboratory and field research, including: What type of predictions do people make about others’ behavior? How robust is ones “best response” strategy to errors in these predictions? Is there any evidence that people learn and converge to equilibrium after repeated interactions? Behavioral economists continue to develop new tools to model strategic interactions in light of these behavioral anomalies [49]. The section titled “Behavioral Economics and Game Theory” in this encyclopedia provides more detail on how the assumptions and descriptive power of game theoretic models can break down, and ways to remedy these gaps.

Applications to Multiechelon Inventory Systems

Managing multiechelon inventory systems is a critical internal function for any global product-based company. Many of the biases that occur when making order quantity decisions (as highlighted in the section titled “Applications to Inventory Planning”) continue to apply as the number of decision makers increases. However, errors in processing feedback and lack of trust introduce additional psychological factors that can further degrade performance.

The majority of behavioral research on multiechelon inventory systems focuses on the bullwhip effect, a phenomenon first introduced by Forrester [50]. The effect refers to the tendency for orders to increase in variability as one travels up a supply chain, away from the final customer. A seminal paper in the area [51], reports the results of a human experiment based on the Beer Distribution Game. This game mimics the dynamics of a decentralized, four-stage, serial supply chain with order and fulfillment delays between each stage. In this experiment, each supply chain consists of four people, taking the role of retailer, wholesaler, distributor, and manufacturer, who make inventory ordering decisions. (Note that when utilizing

the Beer Distribution Game for classroom learning purposes, other protocols are often followed, including assigning more than one person to each stage). Demand is unknown but all other operational causes of the bullwhip effect are controlled for, implying that any evidence of the bullwhip effect can be attributed to behavioral causes. The author finds that the resulting bullwhip effect is significant and points to individuals' failure to account for orders that have been placed but not yet received (i.e., underweighting the supply line) as the behavioral cause.

Other researchers have used a similar laboratory setting to test the impact of different demand patterns [52] and different institutions for reducing bullwhip behavior, including increasing information transparency in the form of point of sale data [52,53] and inventory information [54], and increasing ability-based trust [55]. See Ref. 53 for a review of research using the Beer Game prior to 2002. More recently, laboratory experiments have been used to test conditions thought to trigger order smoothing (rather than the order amplification inherent in the bullwhip effect). These laboratory results [56], along with field-based evidence [57], suggest that retailers smooth orders in response to seasonal demand, particularly when there are nontrivial costs to changing orders. Article titled *Information Sharing in Supply Chains* in this encyclopedia provides further insights into the challenges and opportunities offered by sharing information in supply chains.

Applications to Buyer–Supplier Interactions

In the supply chain literature, buyer–supplier interactions are often modeled as a Stackelberg game. In these games, the supplier traditionally acts as leader, making price or other contract parameter decisions subject to the ordering or retail price setting behavior of the buyer. Understanding this basic two-party dynamic is a critical building block for understanding strategic interactions in more complex supply networks.

One behavioral question that arises in this context is how does the form and framing of the contract influence the decision outcome? For example, suppose the supplier

has a choice of using a wholesale price-only, buyback, or revenue sharing contract (see the section titled “Supply Chain Scheduling” in this encyclopedia for more details on these and other contract structures). Analytical research based on rational choice assumptions shows that the buyback and revenue sharing contracts are equivalent in outcomes and dominate the wholesale price-only contract in terms of the average channel profit and supplier profit. Recent analytical research [58] finds that when incorporating fairness considerations into the buyer's utility function, the efficiency of the wholesale price-only contract improves. However, the level of efficiency achieved is impacted, in part, by the supplier's knowledge of the buyer's fairness concerns. When the buyer's fairness concern is private information, Katok and Pavlov [59] prove analytically that the resulting wholesale price-only contract may be rejected by the buyer with some positive probability. This result is also confirmed empirically through a series of experiments. Other authors [60] compare the performance of buyback and revenue sharing contracts in the laboratory and find that differences in the timing of payment across the two contracts matters to suppliers, at least initially. As a result, suppliers induce lower order quantities under a buyback contract—a behavior consistent with loss aversion. In a different channel setting, where demand is fixed and the buyer is charged with setting the market price, Ho and Zhang [61] find that framing a fixed fee as a quantity discount versus a two-part tariff leads to different pricing behavior, again consistent with loss aversion and prospect theory. In a similar setting, Lim and Ho [62] show that the buyer is more sensitive to the number of price blocks provided than normative theory suggests. In particular, theory suggests that two price breaks are optimal while laboratory studies show that offering additional price breaks further increases the supplier's profit.

Another important question is how will the parties share information? For example, if the retailer has better demand forecast information or the manufacturer has better production cost information, will this

information be passed to the other party without distortion? Game theory provides important insights, in the form of signaling and screening games, for developing contracts that induce truth telling. However, recent laboratory results suggest that such contracts may not always be necessary since human decision makers appear more trustworthy than rational choice theory would suggest. A series of laboratory experiments [63] find that even when a buyer has no financial obligation tied to his forecast (i.e., when the forecast is “cheap talk”), the level of forecast distortion is remarkably low. Other researchers [64] examine a screening game in a laboratory setting, where the buyer’s type (in terms of high, medium, or low demand) is not known to the supplier. They find that the recommended menu of contracts, which takes the form of a series of quantity discounts, can actually reduce the supplier’s profit relative to a wholesale price-only contract in some cases. It appears that the additional complexity introduced by a truth-inducing contract makes it difficult for the supplier to price accurately and for the buyer to interpret these prices.

Recently, Loch and Wu [65] found that informal interactions between parties, something as simple as having a supplier and buyer meet and shake hands, can lead to outcomes with higher distributional fairness. In a multiechelon context, Wu and Katok [66] also find that allowing a brief meeting with supply chain partners leads to decisions with higher payoffs for the supply chain as a whole.

Applications to Procurement Markets

There is a rich literature in behavioral economics on negotiations [67] and the performance of alternative mechanism designs for markets such as auctions (see Refs 68 and 69 for reviews as well as the section titled “Game Theory: Extension and Development” in this encyclopedia). However, the requirements of procurement markets, which are most relevant to the supply chain discipline, uncover some new and interesting complications.

For example, many companies now use reverse auctions to purchase some fraction of their direct and indirect materials. In setting up a reverse auction, a buyer must decide

on the type of scoring function to use to evaluate bids and provide feedback to the suppliers. Devising a scoring function that supports the preferences of the buyer can be difficult when the buyer values nonprice attributes that are multidimensional or difficult to communicate, such as service or quality levels. Even when a proposed function is theoretically proven to be efficient, there is no guarantee that it will perform well in practice, since suppliers in the field may have difficulty interpreting its more complex form. Controlled laboratory experiments with human subjects provide an important proof of concept test for such multiattribute reverse auctions. Laboratory studies provide insight into how factors, such as number of suppliers, or correlation between price and quality, impact the efficiency of reverse auctions under different scoring rules [70–72].

Procurement markets also differ from other exchanges in the types of costs incurred by the bidders/suppliers. For example, suppliers often face fixed costs that are generally independent of the quantity won, but avoidable when the quantity won is zero. This is typical when there is a fixed cost involved in setting up a production line or hiring employees to support new demand needs. Such a cost structure is more difficult to handle within an auction because it implies a different demand threshold level (to cover fixed costs) for each possible bid price. A recent study [73] sheds light on how enriching the bidding language (through multipart bids) and enriching market feedback (through quantity-dependent pricing) impacts bidding behavior in a laboratory environment.

Future Research

The previous examples are a small sample of the type of supply chain interactions that may benefit from a behavioral lens. The field is still at its infancy in understanding how other-regarding preferences such as fairness and social status impact supply chain interactions, especially within organizations where collaborative planning processes take place. The role of reciprocity within supply chains is also not well understood. What causes a supply chain member to punish or reward other

members? What form does such reciprocity take?

Our understanding of trust within supply chain is also incomplete, with most studies so far focusing on “cheap talk” settings. More laboratory and field-based research is needed to understand when monitoring mechanisms (such as truth-inducing contracts or scorecards) are really required to ensure trustworthy behavior in a supply chain. It is also unclear how the stage of the supply chain relationship (captured through repeated interactions and reputation building) impact trust. Studies so far have focused mostly on “calculus-based” trust, typical of early stages.

Finally, we hope to see more strides in the future to implement natural experiments in the field, and to improve supply chain practice by designing better institutional mechanisms, and validating these using experiments. In this way, a number of research methodologies can be applied to provide a more complete picture of the behavioral implications of supply chain interactions.

CONCLUSION

Given the prominence that the behavioral approach to decision making has achieved within other business disciplines such as finance, accounting, and marketing, it is not surprising that research in supply chain management has shifted toward a more behavioral approach as well. Within this development, it is important to emphasize that behavioral research in supply chain management is not limited to the narrow psychological perspective of individual judgment and decision biases. If behavioral research is to capture all deviations from the traditional rational choice paradigm it must be broad and include, for example, the analysis of social interactions and cultural norms inherent in supply chain interactions. Our outline of supply chain applications highlights behavioral issues that arise at both the level of an individual decision maker and within the interaction of two or more supply chain members. We hope that

these examples stimulate more researchers to join in the pursuit of viewing supply chain decisions through a behavioral lens.

The following resources are helpful for learning more about the methodologies underlying behavioral operations. In addition to the review articles mentioned earlier [1,2,5,6], Friedman *et al.* [74] offer an overview of experimental economics, Kagel and Roth [75] give background on designing and running human experiments, and Ho *et al.* [76] provide insight into behavioral modeling and its application to business contexts such as marketing. Finally, Boudreau *et al.* [77] outlines opportunities for research on the interface between operations and human resources management.

REFERENCES

1. Gino F, Pisano G. Toward a theory of behavioral operations. *Manuf Serv Oper Manag* 2008;10(4):676–691.
2. Loch C, Wu Y. Behavioral operations management. *Found Trends Technol Inf Oper Manag* 2007;1(3):1–128.
3. Roth A. The economist as engineer: game theory, experimentation, and computation as tools for design economics. *Econometrica* 2002;70(4):1341–1378.
4. Roth A, Oliveira Sotomayor M. Two-sided matching: a study in game-theoretic modelling and analysis. Cambridge: Cambridge University Press; 1990.
5. Bendoly E, Croson R, Goncalves P, *et al.* Bodies of knowledge for behavioral operations. Working Paper, Goizueta Business School, Emory University. 2009.
6. Bendoly E, Donohue K, Schultz K. Behavior in operations management: assessing recent findings and revisiting old assumptions. *J Oper Manag* 2006;24:737–752.
7. Smith V. Experimental economics: induced value theory. *Am Econ Rev* 1974;66(2):274–279.
8. Simon HA. Rationality in psychology and economics. Chicago, IL: The University of Chicago Press; 1986. pp. 25–40.
9. Tversky A, Kahnemann D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185(4157):1124–1131.
10. Kahnemann D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.

11. Chapman GB, Johnson EJ. Incorporating the irrelevant: anchors in judgment of belief and value. Cambridge, MA: Cambridge University Press; 2002. pp. 120–138.
12. Gilovich T, Vallone R, Tversky A. The hot hand in basketball: on the misperception of random sequences. *Cognit Psychol* 1985;17:295–314.
13. Sanders NR, Marodt KB. The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega* 2003;31:511–522.
14. Lawrence M, Goodwin P, O'Connor M, *et al.* Judgmental forecasting: a review of progress over the last 25 years. *Int J Forecast* 2006;22:493–518.
15. Andreassen PB, Kraus SJ. Judgmental extrapolation and the salience of change. *J Forecast* 1990;9:347–372.
16. Lawrence M, O'Connor M. Exploring judgmental forecasting. *Int J Forecast* 1992;8:15–26.
17. Lawrence M, O'Connor M. The anchor and adjustment heuristic in time-series forecasting. *J Forecast* 1995;14:443–451.
18. Harvey N, Bolger F, McClelland A. On the nature of expectations. *Br J Health Psychol* 1994;85:203–229.
19. Rapoport A, Stein WE, Burkheimer GJ. Response models for detection of change. Dordrecht: D. Reidel; 1979.
20. Estes W. Global and local control of choice behavior by cyclically varying outcome probabilities. *J Exp Psychol Learn Mem Cogn* 1984;10:258–270.
21. Barry DM, Pitz GF. Detection of change in nonstationary, random sequences. *Organ Behav Hum Perform* 1979;24:111–125.
22. Massey C, Wu G. Detecting regime shifts: the causes of under- and overreaction. *Manag Sci* 2005;51(6):932–947.
23. Kremer M, Moritz B, Siemsen E. Demand forecasting behavior: system neglect and change detection. Working Paper, University of Minnesota. 2009.
24. Schweitzer ME, Cachon GP. Decision bias in the newsvendor problem with a known demand distribution: experimental evidence. *Manag Sci* 2000;46(3):404–420.
25. Bostian AA, Holt CA, Smith AM. Newsvendor 'Pull-to-Center' effect: adaptive learning in a laboratory experiment. *Manuf Serv Oper Manag* 2008;10(4):590–608.
26. Lurie N, Swaminathan JM. Is timely information always better? The effect of feedback frequency on decision making. *Organ Behav Hum Decis Process* 2008;108:315–329.
27. Thomas LJ, McClain J, Robinson K, *et al.* The use of framing in inventory decisions. Johnson School Research Paper Series, Cornell University. 2007.
28. Bolton GE, Katok E. Learning by doing in the newsvendor problem: a laboratory investigation of the role of experience and feedback. *Manuf Serv Oper Manag* 2008;10(3):519–538.
29. Gavirneni S, Xia Y. Anchor selection and group dynamics in newsvendor decisions—a note. *Decis Anal* 2009;6(2):87–97.
30. Su X. Bounded rationality in newsvendor models. *Manuf Serv Oper Manag* 2008;10(4):566–589.
31. Kremer M, Minner S, van Wassenhove LN. The behavioral newsvendor—heuristic and preference-based accounts in a complex decision task. Working Paper, Penn State University. 2009.
32. Rudi N, Drake D. Level, adjustment and observation bias in the newsvendor model. Working Paper, INSEAD. 2009.
33. Moritz B, Hill A, Donohue K. Cognition and individual differences in the newsvendor problem: behavior under dual process theory. Working Paper, University of Minnesota. 2009.
34. Krishnan V, Ulrich KT. Product development decisions: a review of the literature. *Manag Sci* 2001;47(1):1–21.
35. Sethi R, Smith DC, Park CW. Cross-functional product development teams, creativity, and the innovativeness of new consumer products. *J Market Res* 2001;38(2):73–85.
36. Keller RT. Cross-functional project groups in research and new product development: diversity, communications, job stress, and outcomes. *Acad Manag J* 2001;44(3):547–555.
37. Li H, Bingham JB, Umphress EE. Fairness from the top: perceived procedural justice and collaborative problem solving in new product development. *Org Sci* 2007;18(2):200–216.
38. Siemsen E. The hidden perils of career concerns in R&D organizations. *Manag Sci* 2008;54(5):863–877.
39. Sosa M. Why do we rework? Organizational dyadic interactions and design rework in software development. Working Paper, INSEAD. 2009.
40. Wu Y, Ramachandran K, Krishnan V. Managing projects with present-biased agents. Work-

- ing Paper, National University of Singapore. 2009.
41. Booker DM, Drake AR, Heitger DL. New product development: how cost information precision affects designer focus and behavior in a multiple objective setting. *Behav Res Account* 2007;19:19–41.
 42. Ho T, Lim N, Cui T. Is inventory centralization profitable? An experimental investigation. Working Paper, University of California-Berkeley. 2009.
 43. Ho T, Su X. Peer-induced fairness in games. *Am Econ Rev* 2009;99(5):2022–2049.
 44. Fehr E, Schmidt KM. A theory of fairness, competition, and cooperations. *Q J Econ* 1999;114(3):817–868.
 45. Rabin M. Incorporating fairness into game theory and economics. *Am Econ Rev* 1993;83:1281–1302.
 46. Wu Y, Loch C, van der Heyden L. A model of fair process and its limits. *Manuf Serv Oper Manag* 2008;10(4):637–653.
 47. Mayer R, Davis J, Schoorman FD. An integrative model of organizational trust. *Acad Manag Rev* 1995;20(3):709–734.
 48. Xia L, Monroe K, Cox J. The price is unfair! A conceptual framework of price fairness perceptions. *J Market* 2004;68:1–15.
 49. Camerer C. Behavioral game theory: experiments in strategic interaction. Princeton, NJ: Princeton University Press; 2003.
 50. Forrester J. Industrial dynamics: a major breakthrough for decision makers. *Harv Bus Rev* 1958;36:37–66.
 51. Sterman J. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Manag Sci* 1989;35:321–339.
 52. Steckel J, Gupta S, Banerji A. Supply chain decision making: will shorter cycle times and shared point-of-sale information necessarily help? *Manag Sci* 2004;52(4):458–464.
 53. Croson R, Donohue K. Experimental economics and supply chain management. *Interfaces* 2002;32(5):74–82.
 54. Croson R, Donohue K. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Manag Sci* 2006;52(3):323–336.
 55. Croson R, Donohue K, Katok E, *et al.* Order stability in supply chains: coordination risk and the role of coordination stock. Working Paper, University of Minnesota. 2008.
 56. Cantor D, Katok E. The bullwhip effect and order smoothing in a laboratory beer game. Working Paper, University of North Florida. 2008.
 57. Cachon G, Randall T, Schmidt G. In search of the bullwhip effect. *Manuf Serv Oper Manag* 2007;9(4):457–479.
 58. Cui TH, Raju S, Zhang ZJ. Fairness and channel coordination. *Manag Sci* 2007;53(8):1303–1314.
 59. Katok E, Pavlov V. Fairness and coordination failures in supply chain contracts. Working Paper, Penn State University. 2009.
 60. Katok E, Wu D. Contracting in supply chains: a laboratory investigation. Working Paper, University of Kansas. 2008.
 61. Ho T, Zhang J. Designing pricing contracts for boundedly rational customers: does the framing of the fixed fee matter? *Manag Sci* 2008;54(4):686–700.
 62. Lim N, Ho T. Designing contracts for boundedly rational customers: does the number of blocks matter? *Market Sci* 2007;26(3):312–326.
 63. Ozalp O, Zheng Y, Chen KY. Trust in forecast information sharing. Working Paper, Stanford University. 2008.
 64. Kalkanci B, Chen KY, Erhun F. Contract complexity and performance under asymmetric demand information: an experimental evaluation. Working Paper, Stanford University. 2008.
 65. Loch C, Wu Y. Social preferences and supply chain performance: an experimental study. *Manag Sci* 2008;54(11):1835–1849.
 66. Wu D, Katok E. Learning, communication, and the bullwhip effect. *J Oper Manag* 2006;24:839–850.
 67. Roth A. Bargaining experiments. In: Kagel J, Roth A, editors. *The handbook of experimental economics*. Princeton, NJ: Princeton University Press; 1995. pp. 253–348.
 68. Kagel J. Auctions: a survey of experimental research. In: Kagel J, Roth A, editors. *The handbook of experimental economics*. Princeton, NJ: Princeton University Press; 1995. pp. 501–585.
 69. Kagel J, Levin D. Auctions: a survey of experimental research, 1995–2008. In: Kagel J, Roth A, editors. *Volume 2, The handbook of experimental economics*. Princeton, NJ: Princeton University Press; 2010.
 70. Bichler M. An experimental analysis of multi-attribute auctions. *Decis Support Syst* 2000;29:249–268.

71. Chen-Ritzo C, Harrison T, Kwasnica A, *et al.* Better, faster, cheaper: an experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback. *Manag Sci* 2005;51(12):1752–1762.
72. Engelbrecht-Wiggans R, Haruvy E, Katok E. A comparison of buyer-determined and price-based multi-attribute mechanisms. *Market Sci* 2007;26(5):611–628.
73. Elmaghraby W, Larson N. Procurement auctions with avoidable fixed costs: an experimental approach. Working Paper, University of Maryland. 2008.
74. Friedman D, Sunder S. Experimental methods: a primer for economists. Cambridge: Cambridge University Press; 1994.
75. Kagel J, Roth A. The handbook of experimental economics. Princeton, NJ: Princeton University Press; 1995.
76. Ho T, Lim N, Camerer C. Modelling the psychology of consumer and firm behavior with behavioral economics. *J Market Res* 2006;XLIII:307–331.
77. Boudreau J, Hopp W, McClain J, *et al.* On the Interface between Operations and Human Resources Management. *Manuf Serv Oper Manag* 2003;5(3):179–202.

BENDERS DECOMPOSITION

ZEKI CANER TAŞKIN

Department of Industrial Engineering,
Boğaziçi University, Bebek, Istanbul,
Turkey

An important concern regarding building and solving optimization problems is that the amount of memory and the computational effort needed to solve such problems grow significantly with the number of variables and constraints. The traditional approach, which involves making all decisions simultaneously by solving a monolithic optimization problem, quickly becomes intractable as the number of variables and constraints increases. Multistage optimization algorithms such as Benders decomposition [1], have been developed as an alternative solution methodology to alleviate this difficulty. Unlike the traditional approach, these algorithms divide the decision-making process into several stages. In Benders decomposition, a first-stage master problem is solved for a subset of variables, and the values of the remaining variables are determined by a second-stage subproblem *given* the values of the first-stage variables. If the subproblem determines that the proposed first-stage decisions are infeasible, then one or more constraints are generated and added to the master problem, which is then re-solved. In this manner, a series of small problems are solved instead of a single large problem, which can be justified by the increased computational resource requirements associated with solving larger problems.

The remainder of this article is organized as follows: We will first describe Benders decomposition formally in the section titled “Formal Derivation.” We will then discuss some extensions of Benders decomposition in the section titled “Extensions.” Finally, we will illustrate the decomposition approach on a problem encountered in Intensity Modulated Radiation Therapy (IMRT) treatment planning in the section titled “Illustrative Example,” and give a numerical example.

FORMAL DERIVATION

In this section, we describe Benders decomposition algorithm for linear programs. Consider the following problem:

$$\text{Minimize } c^T x + f^T y \quad (1a)$$

$$\text{subject to: } Ax + By = b \quad (1b)$$

$$x \geq 0, y \in Y \subseteq \mathbb{R}^q, \quad (1c)$$

where x and y are vectors of continuous variables having dimensions p and q , respectively, Y is a polyhedron, A, B are matrices, and b, c, f are vectors having appropriate dimensions. Suppose that y variables are “complicating variables” in the sense that the problem becomes significantly easier to solve if y variables are fixed, perhaps due to a special structure inherent in matrix A . Benders decomposition partitions Problem (1) into two problems: (i) a master problem that contains the y variables, and (ii) a subproblem that contains the x variables. We first note that Problem (1) can be written in terms of the y variables as follows:

$$\text{Minimize } f^T y + q(y) \quad (2a)$$

$$\text{subject to: } y \in Y, \quad (2b)$$

where $q(y)$ is defined to be the optimal objective function value of

$$\text{Minimize } c^T x \quad (3a)$$

$$\text{subject to: } Ax = b - By \quad (3b)$$

$$x \geq 0. \quad (3c)$$

Formulation (3) is a linear program for any given value of $y \in Y$. Note that if Problem (3) is unbounded for some $y \in Y$, then Problem (2) is also unbounded, which in turn implies unboundedness of the original problem (Eq. 1). Assuming boundedness of Problem (3), we can also calculate $q(y)$ by solving its dual. Let us associate dual variables α with constraints (3b). Then, the dual of

Problem 3 is as follows:

$$\text{Maximize } \alpha^T(b - By) \quad (4a)$$

$$\text{subject to: } A^T \alpha \leq c \quad (4b)$$

$$\alpha \text{ unrestricted.} \quad (4c)$$

A key observation is that feasible region of the dual formulation does not depend on the value of y , which only affects the objective function. Therefore, if the dual feasible region (Eqs 4b and 4c) is empty, then either the primal problem (Eq. 3) is unbounded for some $y \in Y$ [in which case the original problem (Eq. 1) is unbounded], or the primal feasible region (Eqs 3b and 3c) is also empty for all $y \in Y$ (in which case Eq. 1 is also infeasible). Assuming that the feasible region defined by Equations (4b) and (4c) is not empty, we can enumerate all extreme points $(\alpha_p^1, \dots, \alpha_p^I)$, and extreme rays $(\alpha_r^1, \dots, \alpha_r^J)$ of the feasible region, where I and J are the numbers of extreme points and extreme rays of Equations (4b) and (4c), respectively. Then, for a given $\hat{y}$ vector, the dual problem can be solved by checking (i) whether $(\alpha_r^j)^T(b - B\hat{y}) > 0$ for an extreme ray α_r^j , in which case the dual formulation is unbounded and the primal formulation is infeasible, and (ii) finding an extreme point α_p^i that maximizes the value of the objective function $(\alpha_p^i)^T(b - B\hat{y})$, in which case both primal and dual formulations have finite optimal solutions. On the basis of this idea, the dual problem (Eq. 4) can be reformulated as follows:

$$\text{Minimize } q \quad (5a)$$

$$\text{subject to: } (\alpha_r^j)^T(b - By) \leq 0 \quad \forall j = 1, \dots, J \quad (5b)$$

$$(\alpha_p^i)^T(b - By) \leq q \quad \forall i = 1, \dots, I \quad (5c)$$

$$q \text{ unrestricted.} \quad (5d)$$

Note that Problem (5) consists of a single variable q , and typically, a large number of constraints. Now we can replace $q(y)$ in Equation (2a) with Problem (5) and obtain a reformulation of the original problem in

terms of q and y variables.

$$\text{Minimize } f^T y + q \quad (6a)$$

$$\text{subject to: } (\alpha_r^j)^T(b - By) \leq 0 \quad \forall j = 1, \dots, J \quad (6b)$$

$$(\alpha_p^i)^T(b - By) \leq q \quad \forall i = 1, \dots, I \quad (6c)$$

$$y \in Y, q \text{ unrestricted.} \quad (6d)$$

Since there is typically an exponential number of extreme points and extreme rays of the dual formulation (Eq. 4), generating all constraints of type (6b and 6c) is not practical. Instead, Benders decomposition starts with a subset of these constraints, and solves a “relaxed master problem,” which yields a candidate optimal solution (y^*, q^*) . It then solves the dual subproblem (Eq. 4) to calculate $q(y^*)$. If the subproblem has an optimal solution having $q(y^*) = q^*$, then the algorithm stops. Otherwise, if the dual subproblem is unbounded, then a constraint of type (6b) is generated and added to the relaxed master problem, which is then re-solved. [Constraints of type (6b) are referred to as *Benders feasibility cuts* because they enforce necessary conditions for feasibility of the primal subproblem (Eq. 3).] Similarly, if the subproblem has an optimal solution having $q(y^*) > q^*$, then a constraint of type (6c) is added to the relaxed master problem, and the relaxed master problem is re-solved. (Constraints of type (6c) are called *Benders optimality cuts* because they are based on optimality conditions of the subproblem). Since I and J are finite, and new feasibility or optimality cuts are generated in each iteration, this method converges to an optimal solution in a finite number of iterations [1].

EXTENSIONS

Benders decomposition is closely related to other decomposition methods for linear programming (see *Relationship Among Benders, Dantzig-Wolfe, and Lagrangian Optimization*). Furthermore, Benders decomposition can be applied to a broader class of problems, some of which are

described in this section. We first observe that only linear constraints are added to the master problem throughout the iterations of Benders decomposition. Therefore, the master problem does not have to be a linear program, but can take the form of an integer (Ref. 1 and the section titled “Illustrative Example”), a nonlinear [2] or a constraint programming problem [3]. Also note that the subproblem is only used to obtain dual information in order to generate Benders cuts. Therefore, the subproblem does not have to be a linear program, but can also be a convex program since dual multipliers satisfying strong duality conditions can be calculated for such problems [4] for detailed information about convex optimization). The extension of Benders decomposition that allows for nonlinear convex programs to be used as subproblems is referred to as *generalized Benders decomposition* [5]. Similarly, “logic-based Benders decomposition” generalizes the use of linear programming duality in the subproblem to “inference duality,” which allows the use of logic-based methods for solving the subproblem and generating Benders cuts (see **Constraint Programming Links with Math Programming** for more information about constraint programming and its relationships with mathematical programming). In some applications subproblems can be solved efficiently by specialized algorithms instead of the explicit solution of linear programs [6,7]. If the subproblem is a linear feasibility problem (i.e., a linear programming problem having no objective function), cuts based on irreducible infeasible subsets of constraints can be generated using a technique known as *combinatorial Benders decomposition* [8].

It is often the case that decisions for several groups of second-stage variables can be made independently, given the first-stage decisions. In such cases, multiple subproblems can be defined and solved separately. For instance, in stochastic programming models, some action needs to be taken in a first stage, which is followed by the occurrence of a random event (typically modeled by a number of scenarios) that affects the outcome of the first-stage decision. A recourse decision can then be made

in a second stage after the uncertainty is resolved (see sections titled “Models” and “Two-Stage Stochastic Programming” of the encyclopedia for more information about two-stage stochastic programming models). In such problems, second-stage recourse problems can be solved independently given the first-stage decisions, and hence are amenable to parallel implementations [9].

ILLUSTRATIVE EXAMPLE

Problem Definition

In this section, we consider a matrix segmentation problem arising in IMRT treatment planning, which is described in detail in Ref. 10 (see also **Optimization Models for Cancer Treatment Planning** for an introduction to optimization models in cancer treatment). The problem input is a matrix of intensity values that are to be delivered to a patient from some given angle, under the condition that the IMRT device can only deliver radiation through rectangular apertures. An aperture is represented as a binary matrix whose ones appear consecutively in each row and column, and hence form a rectangular shape. A feasible segmentation is one in which the original desired intensity matrix is equal to the weighted sum of a number of feasible binary matrices, where the weight of each binary matrix is the amount of intensity to be delivered through the corresponding aperture. We seek a matrix segmentation that uses the smallest number of aperture matrices to segment the given intensity matrix. This goal corresponds to minimizing setup time in the IMRT context [10]. The example below shows an intensity matrix and a feasible segmentation using three rectangular apertures:

$$\begin{bmatrix} 2 & 7 & 0 \\ 2 & 10 & 3 \\ 0 & 8 & 3 \end{bmatrix} = 2 \times \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ + 3 \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} + 5 \times \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

We will denote the intensity matrix to be delivered by an $m \times n$ matrix B , where the element at row i and column j , (i, j) requires $b_{ij} \in \mathbb{Z}^+$ units of intensity. Let R be the set of all $O(m^2 n^2)$ possible rectangular apertures (i.e., binary matrices of size $m \times n$ having contiguous rows and columns) that can be used in a segmentation of B . For each rectangle $r \in R$, we define a continuous variable x_r that represents the intensity assigned to rectangle r , and a binary variable y_r that equals 1 if rectangle r is used in decomposing B (i.e., if $x_r > 0$), and equals 0 otherwise. We say that element (i, j) is “covered” by rectangle r if the (i, j) element of r is 1. Let $C(r)$ be the set of matrix elements that is covered by rectangle r . We define $M_r = \min_{(i,j) \in C(r)} \{b_{ij}\}$ to be the minimum intensity requirement among the elements of B that are covered by rectangle r . Furthermore, we denote the set of rectangles that cover element (i, j) by $R(i, j)$. Given these definitions, we can formulate the problem as follows:

$$\text{Minimize } \sum_{r \in R} y_r \quad (7a)$$

$$\text{subject to: } \sum_{r \in R(i, j)} x_r = b_{ij} \quad \forall i = 1, \dots, m, \\ j = 1, \dots, n \quad (7b)$$

$$x_r \leq M_r y_r \quad \forall r \in R \quad (7c)$$

$$x_r \geq 0, y_r \in \{0, 1\} \quad \forall r \in R. \quad (7d)$$

The objective function (Eq. 7a) minimizes the number of rectangular apertures used in the segmentation. Constraints (7b) guarantee that each matrix element receives exactly the required dose. Constraints (7c) enforce the condition that x_r cannot be positive unless $y_r = 1$. Finally, Equation (7d) states bounds and logical restrictions on the variables. Note that the Objective (7a) guarantees that $y_r = 0$ when $x_r = 0$ in any optimal solution of Problem (7).

Decomposition Approach

Formulation (7) contains two variables and a constraint for each rectangle, resulting in a large-scale mixed-integer program for problem instances of clinically relevant sizes. Furthermore, the M_r terms in constraints (7c)

lead to a weak linear programming relaxation due to the “big-M” structure. These difficulties can be alleviated by employing a Benders decomposition approach. Our decomposition approach will first select a subset of the rectangles in a master problem, and then check whether the input matrix can be segmented using only the selected rectangles in a subproblem. Let us first reformulate the problem in terms of the y variables.

$$\text{Minimize } \sum_{r \in R} y_r \quad (8a)$$

$$\text{subject to: } y \text{ corresponds to a} \\ \text{feasible segmentation} \quad (8b)$$

$$y_r \in \{0, 1\} \quad \forall r \in R, \quad (8c)$$

where we will address the form of Equation (8b) next. Given a vector $\hat{y}$ that represents a selected subset of rectangles, we can check whether constraint (8b) is satisfied by solving the following subproblem:

$$\text{SP}(\hat{y}) : \text{Minimize } 0 \quad (9a)$$

$$\text{subject to: } \sum_{r \in R(i, j)} x_r = b_{ij} \quad \forall i = 1, \dots, m, \\ j = 1, \dots, n \quad (9b)$$

$$x_r \leq M_r \hat{y}_r \quad \forall r \in R \quad (9c)$$

$$x_r \geq 0 \quad \forall r \in R. \quad (9d)$$

If $\hat{y}$ corresponds to a feasible segmentation then $\text{SP}(\hat{y})$ is feasible, otherwise it is infeasible. Note that Formulation (8) is a pure integer programming problem (since it only contains the y variables), and $\text{SP}(\hat{y})$ is a linear programming problem (since it only contains the x -variables). Furthermore, constraints (9c) reduce to simple upper bounds on x -variables for a given $\hat{y}$, which avoids the “big-M” issue associated with constraints (7c). Given a $\hat{y}$ -vector, if $\text{SP}(\hat{y})$ has a feasible solution $\hat{x}$, then $(\hat{x}, \hat{y})$ constitutes a feasible solution of the original problem (Eq. 7). On the other hand, if $\text{SP}(\hat{y})$ does not yield a feasible solution, then we need to ensure that $\hat{y}$ is eliminated from the feasible region of Problem (8). Benders decomposition uses the theory of linear programming duality to achieve this goal.

Let us associate variables α_{ij} with Equation (9b), and β_r with Equation (9c). Then, the dual formulation of $\text{SP}(\hat{y})$ can be given as:

$$\begin{aligned} \text{DSP}(\hat{y}) : \text{Maximize } & \sum_{i=1}^m \sum_{j=1}^n b_{ij} \alpha_{ij} \\ & + \sum_{r \in R} M_r \hat{y}_r \beta_r \end{aligned} \quad (10a)$$

$$\text{subject to: } \sum_{(i,j) \in C(r)} \alpha_{ij} + \beta_r \leq 0 \quad \forall r \in R \quad (10b)$$

$$\begin{aligned} \alpha_{ij} \text{ unrestricted } & \quad \forall i = 1, \dots, m, \\ j = 1, \dots, n & \quad (10c) \end{aligned}$$

$$\beta_r \leq 0 \quad \forall r \in R. \quad (10d)$$

Our Benders decomposition strategy first relaxes constraints (8b) and solves Problem (8) to optimality, which yields $\hat{y}$. If $\text{SP}(\hat{y})$ has a feasible solution $\hat{x}$, then $(\hat{x}, \hat{y})$ corresponds to an optimal matrix segmentation. On the other hand, if $\text{SP}(\hat{y})$ is infeasible, then the dual formulation $\text{DSP}(\hat{y})$ is unbounded [since the all-zero solution is always a feasible solution of $\text{DSP}(\hat{y})$]. Let $(\hat{\alpha}, \hat{\beta})$ be an extreme ray of $\text{DSP}(\hat{y})$ such that $\sum_{i=1}^m \sum_{j=1}^n b_{ij} \hat{\alpha}_{ij} + \sum_{r \in R} M_r \hat{y}_r \hat{\beta}_r > 0$. Then, all y vectors that are feasible with respect to Equation (8b) must satisfy

$$\sum_{i=1}^m \sum_{j=1}^n b_{ij} \hat{\alpha}_{ij} + \sum_{r \in R} (M_r \hat{\beta}_r) y_r \leq 0. \quad (11)$$

We add Equation (11) to Problem (8), and resolve it in the next iteration to obtain a new candidate optimal solution.

Now let us consider a slight variation of the matrix segmentation problem, where the goal is to minimize a weighted combination of the number of matrices used in the segmentation (corresponding to setup time) and the sum of the matrix coefficients (corresponding to “beam-on-time”). In IMRT treatment planning context, this objective corresponds to minimizing the total treatment time [10]. In order to incorporate this change in our model, we simply replace the objective function (Eq. 7a) with

$$\text{Minimize } w \sum_{r \in R} y_r + \sum_{r \in R} x_r, \quad (12)$$

where w is a parameter that represents the average setup time per aperture relative to the time required to deliver a unit of intensity.

The Benders decomposition procedure, discussed above, needs to be adjusted accordingly. We first add a continuous variable t to Formulation (8), which “predicts” the minimum beam-on-time that can be obtained by the set of rectangles chosen. The updated formulation can be written as follows:

$$\text{Minimize } w \sum_{r \in R} y_r + t \quad (13a)$$

subject to: y corresponds to a

$$\text{feasible segmentation} \quad (13b)$$

$$t \geq \text{minimum beam-on-time}$$

$$\text{corresponding to } y \quad (13c)$$

$$t \geq 0, y_r \in \{0, 1\} \quad \forall r \in R. \quad (13d)$$

Given a vector $\hat{y}$, we can find the minimum beam-on-time for the corresponding segmentation, if one exists, by solving:

$$\text{SPTT}(\hat{y}) : \text{Minimize } \sum_{r \in R} x_r \quad (14a)$$

$$\text{subject to: } \sum_{r \in R(i,j)} x_r = b_{ij} \quad \forall i = 1, \dots, m,$$

$$j = 1, \dots, n \quad (14b)$$

$$x_r \leq M_r \hat{y}_r \quad \forall r \in R \quad (14c)$$

$$x_r \geq 0 \quad \forall r \in R. \quad (14d)$$

Let α_{ij} and β_r be dual multipliers associated with constraints (14b) and (14c), respectively. Then, the dual of $\text{SPTT}(\hat{y})$ is:

$$\begin{aligned} \text{DSPTT}(\hat{y}) : \text{Maximize } & \sum_{i=1}^m \sum_{j=1}^n b_{ij} \alpha_{ij} \\ & + \sum_{r \in R} M_r \hat{y}_r \beta_r \end{aligned} \quad (15a)$$

$$\text{subject to: } \sum_{(i,j) \in C(r)} \alpha_{ij} + \beta_r \leq 1 \quad \forall r \in R \quad (15b)$$

$$\begin{aligned} \alpha_{ij} \text{ unrestricted } & \quad \forall i = 1, \dots, m, \\ j = 1, \dots, n & \quad (15c) \end{aligned}$$

$$\beta_r \leq 0 \quad \forall r \in R. \quad (15d)$$

Note that $\text{SPTT}(\hat{y})$ is obtained by simply changing the objective function of $\text{SP}(\hat{y})$, and $\text{DSPTT}(\hat{y})$ is obtained by changing the right hand side of Equation (10b) in $\text{DSP}(\hat{y})$. If $\text{DSPTT}(\hat{y})$ is unbounded, then we add a Benders feasibility cut of type (11) as before, and re-solve Problem (13). Otherwise, let the value of t in Problem (13) be $\hat{t}$, and the optimal objective function value of $\text{DSPTT}(\hat{y})$ be t^* . If $\hat{t} = t^*$, then $(\hat{y}, \hat{t})$ is an optimal solution of Problem (13), which minimizes the total treatment time. However, if $\hat{t} > t^*$, then we need to add a constraint that satisfies the following properties: (i) the optimal value of $t = \hat{t}$ if $\hat{y}$ is generated by Problem (13) in a future iteration, and (ii) the optimal value of $t \leq \hat{t}$ for all y . Benders decomposition, once again, uses linear programming duality theory to generate such a constraint. Let $\hat{\alpha}_{ij}$ and $\hat{\beta}_r$ be optimal dual multipliers. It can be seen that the following constraint satisfies both requirements.

$$t \geq \sum_{i=1}^m \sum_{j=1}^n b_{ij} \hat{\alpha}_{ij} + \sum_{r \in R} (M_r \hat{\beta}_r) y_r. \quad (16)$$

Numerical Example

In this section, we give a simple numerical example illustrating the steps of Benders decomposition approach on our matrix segmentation problem. Consider the input matrix $B = \begin{bmatrix} 8 & 3 \\ 5 & 0 \end{bmatrix}$. The set of rectangular apertures that can be used to segment B is:

$$R = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \right\}$$

Let the average setup time per aperture relative to the time required to deliver a unit of intensity be $w = 7$. Defining an x_r and a y_r -variable for each rectangle $r = 1, \dots, 5$, the problem of minimizing total treatment time can be expressed as the following mixed-integer program:

$$\begin{aligned} \text{Minimize } & 7 \times (y_1 + y_2 + y_3 + y_4 + y_5) \\ & + x_1 + x_2 + x_3 + x_4 + x_5 \end{aligned} \quad (17a)$$

$$\text{subject to: } x_1 + x_4 + x_5 = 8 \quad (17b)$$

$$x_2 + x_5 = 3 \quad (17c)$$

$$x_3 + x_4 = 5 \quad (17d)$$

$$x_1 \leq 8y_1, x_2 \leq 3y_2, x_3 \leq 5y_3,$$

$$x_4 \leq 5y_4, x_5 \leq 3y_5 \quad (17e)$$

$$x_r \geq 0, y_r \in \{0, 1\} \quad \forall r = 1, \dots, 5. \quad (17f)$$

For a given $\hat{y}$ -vector, the primal subproblem $\text{SPTT}(\hat{y})$ can be given as

$\text{SPTT}(\hat{y})$:

$$\text{Minimize } x_1 + x_2 + x_3 + x_4 + x_5 \quad (18a)$$

$$\text{subject to: } x_1 + x_4 + x_5 = 8 \quad (18b)$$

$$x_2 + x_5 = 3 \quad (18c)$$

$$x_3 + x_4 = 5 \quad (18d)$$

$$x_1 \leq 8\hat{y}_1, x_2 \leq 3\hat{y}_2,$$

$$x_3 \leq 5\hat{y}_3, x_4 \leq 5\hat{y}_4,$$

$$x_5 \leq 3\hat{y}_5 \quad (18e)$$

$$x_r \geq 0 \quad \forall r = 1, \dots, 5. \quad (18f)$$

Associating dual variables α_{11} with Equation (18b), α_{12} with Equation (18c), α_{21} with Equation (18d), and $\beta_1, \dots, \beta_5$ with Equation (18e), we get the dual subproblem $\text{DSPTT}(\hat{y})$.

$\text{DSPTT}(\hat{y})$:

$$\text{Maximize } 8\alpha_{11} + 3\alpha_{12} + 5\alpha_{21}$$

$$+ 8\hat{y}_1\beta_1 + 3\hat{y}_2\beta_2 + 5\hat{y}_3\beta_3$$

$$+ 5\hat{y}_4\beta_4 + 3\hat{y}_5\beta_5 \quad (19a)$$

$$\text{subject to: } \alpha_{11} + \beta_1 \leq 1 \quad (19b)$$

$$\alpha_{12} + \beta_2 \leq 1 \quad (19c)$$

$$\alpha_{21} + \beta_3 \leq 1 \quad (19d)$$

$$\alpha_{11} + \alpha_{21} + \beta_4 \leq 1 \quad (19e)$$

$$\alpha_{11} + \alpha_{12} + \beta_5 \leq 1 \quad (19f)$$

$$\alpha_{11}, \alpha_{12}, \alpha_{21} \text{ unrestricted} \quad (19g)$$

$$\beta_r \leq 0 \quad \forall r = 1, \dots, 5. \quad (19h)$$

Iteration 1. We first relax all Benders cuts in the master problem, and solve

$$\begin{aligned} \text{Minimize } & 7 \times (y_1 + y_2 + y_3 + y_4 + y_5) \\ & + t \end{aligned} \quad (20a)$$

$$\begin{aligned} \text{subject to: } & t \geq 0, y_r \in \{0, 1\} \\ & \forall r = 1, \dots, 5. \end{aligned} \quad (20b)$$

The optimal solution of Problem (20) is $\hat{y} = [0, 0, 0, 0, 0], \hat{t} = 0$. In order to solve the subproblem corresponding to $\hat{y}$, we set the objective function (Eq. 19h) to Maximize $8\alpha_{11} + 3\alpha_{12} + 5\alpha_{21}$, and solve DSPTT($\hat{y}$). DSPTT($\hat{y}$) is unbounded having an extreme ray $\alpha_{11} = 2, \alpha_{12} = -1, \alpha_{21} = -1, \beta_1 = -2, \beta_2 = 0, \beta_3 = 0, \beta_4 = -1, \beta_5 = -1$, which yields the Benders feasibility cut $8 - 16y_1 - 5y_4 - 3y_5 \leq 0$.

Iteration 2. We add the generated Benders feasibility cut to our relaxed master problem, and solve

$$\begin{aligned} \text{Minimize } & 7 \times (y_1 + y_2 + y_3 + y_4 + y_5) \\ & + t \end{aligned} \quad (21a)$$

$$\begin{aligned} \text{subject to: } & 8 - 16y_1 - 5y_4 - 3y_5 \leq 0 \\ & (21b) \end{aligned}$$

$$\begin{aligned} & t \geq 0, y_r \in \{0, 1\} \\ & \forall r = 1, \dots, 5. \end{aligned} \quad (21c)$$

An optimal solution of Problem (21) is $\hat{y} = [1, 0, 0, 0, 0], \hat{t} = 0$. We set the objective function (Eq. 19a) to Maximize $8\alpha_{11} + 3\alpha_{12} + 5\alpha_{21} + 8\beta_1$, and solve DSPTT($\hat{y}$). DSPTT($\hat{y}$) is, again, unbounded. An extreme ray is $\alpha_{11} = 0, \alpha_{12} = 0, \alpha_{21} = 1, \beta_1 = 0, \beta_2 = 0, \beta_3 = -1, \beta_4 = -1, \beta_5 = 0$, which yields the following Benders feasibility cut: $5 - 5y_3 - 5y_4 \leq 0$.

Iteration 3. We update our relaxed master problem by adding the generated Benders feasibility cut:

$$\begin{aligned} \text{Minimize } & 7 \times (y_1 + y_2 + y_3 + y_4 + y_5) \\ & + t \end{aligned} \quad (22a)$$

$$\begin{aligned} \text{subject to: } & 8 - 16y_1 - 5y_4 - 3y_5 \leq 0 \\ & (22b) \end{aligned}$$

$$5 - 5y_3 - 5y_4 \leq 0 \quad (22c)$$

$$\begin{aligned} & t \geq 0, y_r \in \{0, 1\} \\ & \forall r = 1, \dots, 5. \end{aligned} \quad (22d)$$

An optimal solution of Problem (22) is $\hat{y} = [0, 0, 0, 1, 1], \hat{t} = 0$. We set the objective function (Eq. 19a) to Maximize $8\alpha_{11} + 3\alpha_{12} + 5\alpha_{21} + 5\beta_4 + 3\beta_5$, and solve DSPTT($\hat{y}$). This time DSPTT($\hat{y}$) has an optimal solution $\alpha_{11} = 1, \alpha_{12} = 1, \alpha_{21} = 1, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = -1, \beta_5 = -1$, and the corresponding objective function value is $t^* = 8$. Since $t^* > \hat{t}$, we generate the following Benders optimality cut: $16 - 5y_4 - 3y_5 \leq t$.

Iteration 4. The updated relaxed Benders master problem is:

$$\begin{aligned} \text{Minimize } & 7 \times (y_1 + y_2 + y_3 + y_4 + y_5) \\ & + t \end{aligned} \quad (23a)$$

$$\begin{aligned} \text{subject to: } & 8 - 16y_1 - 5y_4 - 3y_5 \leq 0 \\ & (23b) \end{aligned}$$

$$5 - 5y_3 - 5y_4 \leq 0 \quad (23c)$$

$$16 - 5y_4 - 3y_5 \leq t \quad (23d)$$

$$\begin{aligned} & t \geq 0, y_r \in \{0, 1\} \\ & \forall r = 1, \dots, 5. \end{aligned} \quad (23e)$$

An optimal solution of Problem (23) is $\hat{y} = [0, 0, 0, 1, 1], \hat{t} = 8$. Note that $\hat{y}$ is equal to the solution generated in the previous iteration, and therefore $t^* = 8$. Since $t^* = \hat{t}$, optimality has been reached and we stop.

REFERENCES

1. Benders JF. Partitioning procedures for solving mixed-variables programming problems. *Numer Math* 1962;4(1):238–252.
2. Cai X, McKinney DC, Lasdon LS, *et al.* Solving large nonconvex water resources management models using generalized Benders decomposition. *Oper Res* 2001;49(2):235–245.
3. Eremin A, Wallace M. Hybrid Benders decomposition algorithms in constraint logic programming. In: Walsh T, editor. Volume 2239, Principles and practice of constraint programming—CP 2001, Lecture Notes in Computer Science. Berlin-Heidelberg, Germany: Springer; 2001. pp. 1–15.
4. Bazaraa MS, Sherali HD, Shetty CC. Nonlinear programming: theory and algorithms. 2nd ed. Hoboken (NJ): John Wiley & Sons, Inc.; 1993.
5. Geoffrion AM. Generalized Benders decomposition. *J Optim Theory Appl* 1972;10(4):237–260.
6. Costa AM. A survey on Benders decomposition applied to fixed-charge network design problems. *Comput Oper Res* 2005;32(6):1429–1450.
7. Andreas AK, Smith JC. Decomposition algorithms for the design of a nonsimultaneous capacitated evacuation tree network. *Networks* 2009;53(2):91–103.
8. Codato G, Fischetti M. Combinatorial Benders' cuts for mixed-integer linear programming. *Oper Res* 2006;54(4):756–766.
9. Nielsen SS, Zenios SA. Scalable parallel Benders decomposition for stochastic linear programming. *Parallel Comput* 1997;23(8):1069–1088.
10. Taşkın ZC, Smith JC, Romeijn HE. Mixed-integer programming techniques for decomposing IMRT fluence maps using rectangular apertures. Technical report. Gainesville (FL): Department of Industrial and Systems Engineering, University of Florida; 2009.

BICLUSTERING: ALGORITHMS AND APPLICATION IN DATA MINING

PETROS XANTHOPOULOS

NIKITA BOYKO

NENG FAN

PANOS M. PARDALOS

Department of Industrial and Systems
Engineering and Biomedical Engineering,
University of Florida, Center for Applied
Optimization, Gainesville, Florida

INTRODUCTION

Clustering is the grouping of objects (samples) that possess similar properties (features). The groups are called *clusters*. In other words, the problem is to “guess” some kind of optimal grouping without having any information about the labels of the objects. On the other hand, in classification a labeled set of samples is given, known as *training set*. The problem is to establish a decision rule in order to classify an unlabeled set of new samples, known as the *test set*.

Biclustering, also known as *coclustering*, differs from clustering and classification, but still shares the basic concepts of samples and features and can be supervised or unsupervised. Input for this problem can be given in an $n \times m$ data matrix A in which every line corresponds to one sample and every column corresponds to one feature. The problem is to rearrange the lines and the rows of this matrix in a way that samples with “similar” expression profile are grouped together, and features that are expressed in the same sample are grouped together as well. Then, the numeric value of the matrix a_{ij} corresponds to the amount of expression of the j^{th} feature in the i^{th} sample.

A very common way to visualize biclustering data is heatmapping. A heatmap is an $n \times m$ checkerboard, where every square corresponds to a different element of data in matrix A . Every square is color coded and corresponds to a numeric value, which

determines the sample’s feature expression. Figure 1 is an example of a heatmap.

An excellent example illustrating the idea of biclustering comes from computational biology and more specifically from microarray analysis. In this example, every line of the $n \times m$ data matrix A corresponds to a small part of DNA and every column corresponds to a sequence. The numeric values of the data matrix indicate if the column sequence exists in every sample. Then the problem is to see which features (short sequences) are associated with which samples and group them into biclusters. This skill is essential for genotyping.

The organization of this paper is as follows: In section titled “Mathematical Formulation of Biclustering,” we give a mathematical description of the biclustering problem, in section titled “Algorithmic Approaches,” we present the main algorithmic approaches for solving biclustering, and in the last section we illustrate some additional applications taken from the field of data mining and computational forecasting.

MATHEMATICAL FORMULATION OF BICLUSTERING

In this section, we present a mathematical formulation of the biclustering problem. Let A be an $n \times m$ data matrix. Each element a_{ij} corresponds to the numeric value of the j^{th} feature for the i^{th} sample.

Let $I \subseteq \{1, 2, \dots, n\}$ represent a subset of row indices and $J \subseteq \{1, 2, \dots, m\}$ be a subset of column indices. Then the submatrix A_{IJ} of A that contains the elements from the selected rows and columns is called a *biclus-ter*. A collection of biclusters $A_{I_s J_s}, s = 1, \dots, S$ is called *biclustering*. It is natural to demand that the data within a bicluster possess certain property or criterion such as uniformity, high or low value, and so on. A typical goal of biclustering is to find a maximum size bicluster or biclusters without violating this criterion. Thus, biclustering is formulated as an optimization problem and presently is a

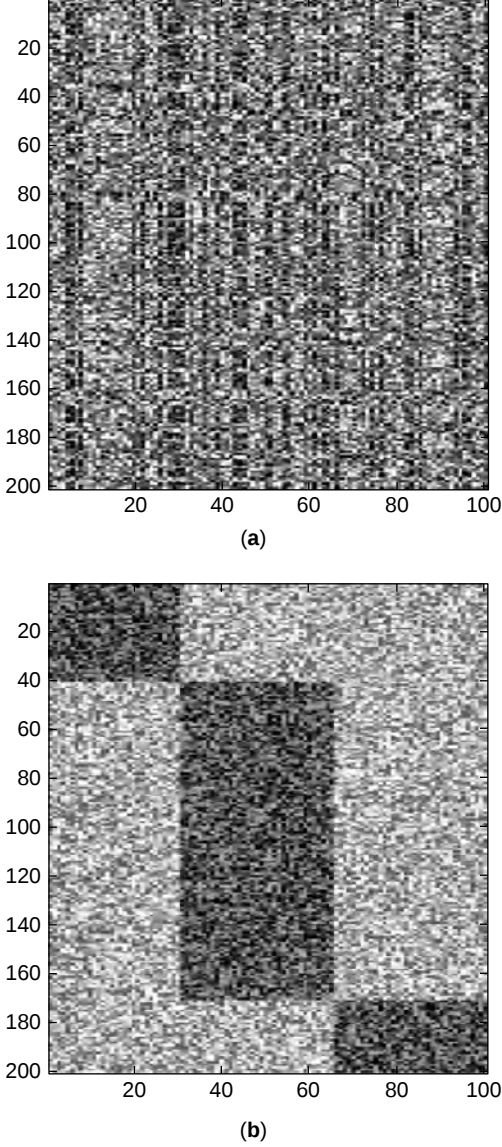


Figure 1. (a) Heatmap for a matrix before applying biclustering. (b) The same matrix permuted in such a way that the bicluster structure is clear.

promising emerging field of study in the operations research community.

The above definition allows two biclusters to share a common feature or sample. In some application, nonoverlapping partitioning is required. Thus, a more strict definition of biclustering is often used [1].

Consider a clustering of m samples into r nonoverlapping clusters $S_1, S_2, S_3, \dots, S_r$ and a clustering of n features in r other nonoverlapping clusters $F_1, F_2, \dots, F_r$. A biclustering B can be defined as the collection of groupings of sample sets and feature sets $B = \{(S_1, F_1), (S_2, F_2), \dots, (S_r, F_r)\}$; each pair of clusters (S_i, F_i) , $i = 1, \dots, r$ is called a *bicluster* or *cocluster*.

ALGORITHMIC APPROACHES

This section briefly discusses some well-known algorithms. The approaches herein address different aspects of biclustering classification, such as definition of the bicluster, the notion of good versus best biclustering, probabilistic versus stochastic structure of the data set, and so on.

Many of the optimization problems for biclustering are NP-hard [2]. Thus, finding the best or “nearly optimal” biclustering classification for large real-world problems is a challenging task requiring expertise in combinatorial optimization. A detailed discussion on complexity in biclustering can be found in surveys [3–5]. Some of the algorithms presented here are heuristics for finding sub-optimal classification, while others aim at finding exact optimal biclustering.

Let us start by considering a bicluster to be a “uniform” submatrix. To quantify the degree of similarity within a bicluster, Hartigan [6] introduces variance as:

$$\text{Var}(S_k, F_k) = \sum_{i \in F_k} \sum_{j \in S_k} (a_{ij} - \mu_k)^2,$$

where

$$\mu_k = \frac{\sum_{i \in F_k} \sum_{j \in S_k} a_{ij}}{|F_k| |S_k|}$$

is the average value of a bicluster. Assuming that biclusters with lower variance are better than those with higher variance leads to

minimizing:

$$\begin{aligned}\text{Var}(S, F) &= \sum_{k=1}^r \text{Var}(S_k, F_k) \\ &= \sum_{k=1}^r \sum_{i \in F_k} \sum_{j \in S_k} (a_{ij} - \mu_k)^2,\end{aligned}$$

where r is a predefined number of biclusters. Hartigan also considers other objective functions representing specific patterns (i.e., variance within the column). Cheng and Church [7] propose to measure mean squared residue score

$$\sum_{i \in F_k} \sum_{j \in S_k} (a_{ij} - \mu_{ik}^r - \mu_{jk}^c + \mu_k)^2,$$

where $\mu_{ik}^r = \frac{1}{|S_k|} \sum_{j \in S_k} a_{ij}$ is the mean of i^{th} row of bicluster k and $\mu_{jk}^c = \frac{1}{|F_k|} \sum_{i \in F_k} a_{ij}$ is the corresponding mean for the j^{th} column. The objective would be to find the maximum size square bicluster having low $H_k < \delta$ or to minimize sum square residue for a fixed number of clusters. Since these problems are NP-hard, greedy, and local search type heuristics, their combinations are used to find a local solution of good quality. One of the widespread residue minimization algorithms is the approach of Yang *et al.* [8] known as *flexible overlapping clustering* (FLOC). It is generalized for the case when some data are missing and based on a simple polynomial heuristic aimed at finding the biclusters with low residue values.

Spectral biclustering algorithms are based on singular value decomposition (SVD) of data set matrix. A is represented as a sum of linear combination of rank-1 matrices:

$$A = \sum_{i=1}^k \sigma_i u_i v_i^T.$$

Kluger [9] suggested the algorithm based on the observation that checkerboard pattern in the data set can be detected by finding piecewise blocks in the pairs of vectors u_i and v_i corresponding to singular values σ_i of high magnitude.

There are methods that are not based directly on SVD, but use this concept for

solving related optimization problems. An example of this is the algorithm suggested by Dhillon in Ref. 10. If the data set is represented as a bipartite weighted graph, then the biclustering is associated with the cut of minimal weight. Finding such a cut is NP-hard. To find an approximated solution, the heuristic utilizes the SVD of a matrix of special structure. Biclustering can be done using conventional clustering algorithms such as k -means or self-organizing maps (SOM). Double conjugated clustering (DCC) [11] consequently clusters feature space and sample space and iteratively updates spaces based on preceding clustering. The sequence of successive clustering terminates when classifications stop changing.

To ensure reliability of resulting biclustering Busygin, Prokopyev, and Pardalos introduced a notion of consistency [1]. Assuming we have a biclustering classification, we can compute the matrix of centroid vectors containing mean values of the features for each sample class

$$c_{ik} = \frac{\sum_{j \in S_k} a_{ij}}{|S_k|}.$$

Feature i is classified into class $\hat{k}$, if its mean value is the highest in the class ($c_{i\hat{k}} > c_{ik}$ for all $k \neq \hat{k}$). Similarly, the centroid matrix can be computed for feature vectors

$$d_{jk} = \sum_{i \in F_k} \frac{a_{ij}}{|F_k|}.$$

Sample j is classified to class k if $d_{j\hat{k}} > d_{jk}$ for all $k \neq \hat{k}$. If the inequalities $c_{i\hat{k}} > c_{ik}$ and $d_{j\hat{k}} > d_{jk}$ are held simultaneously, then the biclustering classification is consistent. The data set is called *conditionally biclustering* if consistent biclustering exists with respect to a predefined partial classification. A supervised biclustering algorithm based on solving 0-1 fractional problems, is introduced in Ref. 1. This algorithm finds a maximum cardinality subset of features that make the data set conditionally biclustering admitting.

Biclustering data set A can be thought of as a joint probability distribution of rows and

columns $p(x = i, y = j) \propto a_{ij}$. Dhillon *et al.* [12] proposed a biclustering algorithm that finds a local minimum for the loss of mutual information. The algorithm is based on the assumption that keeping proper features and samples in the same class is beneficial from the view of one word point of information theory.

Biclustering via Gibbs sampling [13] allows for assigning certain probabilities of the bicluster to rows and columns. The probabilities are iteratively updated using the Bayesian framework. Eventually, the bicluster consists of rows and columns in which estimated probability values exceed a certain threshold.

Another probabilistic approach is the statistical algorithmic method for biclustering analysis (SAMBA) [14]. This algorithm uses statistical reasoning to formulate biclustering as a maximum bounded biclique problem.

Coupled two-way clustering is an analog of the hierarchical algorithm for convenient clusters [15]. Set of samples and features is clustered into two classes each resulting in four biclusters. This process is repeated until certain stopping criteria (i.e., some statistical characteristics) hold.

The plaid model method [16] assumes that an ideal biclustering has diagonal dominating checkboard pattern and the same

background color outside the clusters. On the basis of this assumption, biclustering can be reduced to an optimization problem with objective function reaching the minimum on the above-mentioned pattern.

The surveys [3–5] contain comprehensive introduction to the state of art for biclustering algorithms.

APPLICATIONS

In the introductory part, we described the intuition behind the biclustering problem by showing a computational biology application of the problem. A lot of research works deals with data sets obtained from microarray data [7,11,13,15,17–21]. Other than microarray analysis, biclustering has been used for analyzing drug activity data [22] and nutritional data [16]. Biclustering is also used in text mining. In these instances, every line of the data matrix A corresponds to a dictionary word and every column to a specific document. Then the numerical value a_{ij} is the number of times that word i appears in document j [10,12,23]. Another interesting application of biclustering is in collaborative searching. In this case, biclustering attempts to identify users with similar preferences and make suggestions that fit the user's taste. Example for services that use collaborative

Table 1. Overview of Biclustering Applications in Several Engineering/Science Areas

Application	References	Rows of Matrix A	Columns of Matrix A	a_{ij}
Microarray analysis	7,11,13,17 15,18–21	Gene	Specimen	Amount of gene expression
Drug activity	22	Chemical compounds	Chemical compound descriptor	Value of the chemical descriptor
Text mining (bag of words)	10,12,23	Word	Document	Number of times that i^{th} word appears in j^{th} document
Collaborative filtering	8,24–26	Users	Preferred movies	Rating of the i^{th} user for the j^{th} movie
Dimensionality reduction (databases)	27	Database queries	sources	Association of i^{th} query with i^{th} source
e-Politics	6	A citizen	Topics	Opinion of i^{th} citizen on j^{th} topic

The third, fourth, and fifth columns indicate the physical interpretation of every matrix row, column, and element numerical value for the specific application.

search are movie rental websites and on-line bookstores. An overview of the existing applications of biclustering can be found in Table 1.

CONCLUSION

Biclustering is a problem with uses in many emerging fields of science and engineering. In this article, we gave an overview of the major engineering/science applications, where biclustering is useful, and the most important algorithms for addressing this problem. Since microarray data analysis and text mining are developing rapidly, more theoreticians and practitioners will demonstrate an interest in biclustering. Further theoretical justification of the existing methods and a more thorough analysis of the algorithm (with respect to the quality of the solution) is definitely needed. As pointed out by Busygin *et al.* [3] in the conclusion of his review paper, “future development of biclustering should involve more theoretical studies of biclustering methodology and formalization of its quality criteria.”

REFERENCES

1. Busygin S, Prokopyev OA, Pardalos PM. Feature selection for consistent biclustering via fractional 0–1 programming. *J Comb Optim* 2005;10(1):7–21.
2. Garey MR, Johnson DS. Guide to the theory of NP-completeness. San Francisco (CA): W. H. Freeman; 1979.
3. Busygin S, Prokopyev OA, Pardalos PM. Biclustering in data mining. *Comput Oper Res* 2008;35(9):2964–2987.
4. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 2004;1(1):24–45.
5. Tanay A, Sharan R, Shamir R. Biclustering algorithms: a survey. *Handbook of Computational Molecular Biology*. 1st ed. Boca Raton(FL): Chapman & Hall/CRC (Taylor & Francis Group); 2004.
6. Hartigan JA. Direct clustering of a data matrix. *J Am Stat Assoc* 1972; 67(337):123–129.
7. Cheng Y, Church GM. Biclustering of expression data. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. La Jolla (CA): AAAI Press; 2000. pp. 93–103.
8. Yang J, Wang H, Wang Wei, *et al.* Enhanced biclustering on expression data. In: *BIBE '03: Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*. Washington (DC): IEEE Computer Society; 2003. pp. 321–.
9. Kluger Y, Basri R, Chang JT, *et al.* Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 2003;13(4):703–716.
10. Dhillon IS. Co-clustering documents and words using bipartite spectral graph partitioning. In: *KDD '01: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2001. pp. 269–274.
11. Busygin S, Jacobsen G, Krämer E. Double conjugated clustering applied to leukemia microarray data. In: *SIAM Data Mining Workshop on Clustering High Dimensional Data and Its Applications*; Arlington (VA). 2002.
12. Dhillon IS, Mallela S, Modha DS. Information-theoretic co-clustering. In: *KDD '03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2003. pp. 89–98.
13. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. *Bioinformatics* 2003;19:196–205.
14. Tanay A, Sharan R, Shamir R. Biclustering gene expression data. In: *Proceedings of ISMB 2002*; Edmonton, Canada. 2002. pp. 136–144.
15. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA* 2000;97:12079–12084.
16. Lazzaroni L, Owen A. Plaid models for gene expression data. *Stat Sin* 2002;12:61–86.
17. Cho H, Dhillon IS, Guan Y, *et al.* Minimum sum-squared residue co-clustering of gene expression data. In: *Proceedings of the 4th SIAM International Conference on Data Mining*; Orlando (FL). 2004.
18. Tanay A, Sharan R, Kupiec M, *et al.* Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* 2004;101:2981–2986.
19. Ben-Dor A, Chor B, Karp R, *et al.* Discovering local structure in gene expression data:

- the order-preserving submatrix problem. In: Proceedings of the 6th Annual International Conference on Computational Biology (RECOMB 002). New York: ACM Press; 2002. pp. 49–57.
20. Ben-Dor A, Chor B, Karp R, *et al.* Discovering local structure in gene expression data: the order-preserving submatrix problem. *J Comput Biol* 2003;10:373–384.
 21. Califano A, Stolovitzky G, Tu Y. Analysis of gene expression microarrays for phenotype classification. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. La Jolla (CA): AAAI Press; 2000. pp. 75–85.
 22. Liu J, Wang W. Op-cluster: clustering by tendency in high dimensional space. In: Proceedings of the 3rd IEEE International Conference on Data Mining; Melbourne (FL). 2003. pp. 187–194.
 23. Banerjee A, Dhillon I, Ghosh J, *et al.* A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. In: KDD '04: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2004. pp. 509–514.
 24. Ungar L, Foster DP. A formal statistical approach to collaborative filtering. In: Proceedings of the Conference on Automated Learning and Discovery (CONALD' 98); Pittsburgh (PA). 1998.
 25. Hofmann T, Puzicha J. Latent class models for collaborative filtering. In: IJCAI '99: Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Francisco (CA): Morgan Kaufmann Publishers Inc.; 1999. pp. 688–693.
 26. Wang H, Wang W, Yang J, *et al.* Clustering by pattern similarity in large data sets. In: SIGMOD '02: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. New York: ACM; 2002. pp. 394–405.
 27. Agrawal R, Gehrke J, Gunopulos D, *et al.* Automatic subspace clustering of high dimensional data for data mining applications. In: SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York: ACM; 1998. pp. 94–105.

BILEVEL NETWORK INTERDICTION MODELS: FORMULATIONS AND SOLUTIONS

R. KEVIN WOOD
Department of Operations Research,
Naval Postgraduate School,
Monterey, California

INTRODUCTION

A dictionary definition of *interdict*, in the military sense, is

to destroy, cut or damage by ground or aerial firepower (enemy lines of reinforcement, supply, or communication) in order to stop or hamper enemy movement and to destroy or limit enemy effectiveness [1].

This definition is unnecessarily restrictive—for instance, it should also include ship-based firepower—but the essence is reasonable: interdiction connotes preemptive attacks that limit an enemy's subsequent ability to wage war, or carry out other nefarious activities. The mathematical study of interdiction has focused primarily on *network interdiction*, in which an enemy's activities are modeled using the constructs of network optimization (e.g., maximum flows, multi-commodity flows, and shortest paths), and in which attacks target the network's components to disrupt the network's functionality. Depending on the type of network, targeted components can include bridges, road segments or interchanges, communications links or switches, and so on. This article focuses on the *bilevel network interdiction problem* (BNI), but the reader should note that much of the presentation extends to interdiction of more general systems.

Examples of what we now call network interdiction date from antiquity. Herodotus [[2] 9.49–50] describes how the Persian cavalry (in 479 BC) cut Greek supply lines and routes to water sources in a battle near the

Greek city of Plataea; Livy [3, 22.8] reports that (in 218 BC) the Roman Senate ordered bridges near Rome to be destroyed to slow the advance of Hannibal and his troops; Polybius [4, 9.7] gives a different chronology for the latter incident, but the interpretation as “network interdiction” remains. Two millennia later, the American Civil War is replete with examples of both Confederate and Union forces attacking roads, bridges, rail lines, and telegraph lines to hamper the enemy's resupply, movement, and communications [5, Chapter 4]. In World War II, German submarines interdicted, that is, sank, hundreds of Allied petroleum tankers that were traveling the sea lanes of the Atlantic Ocean and elsewhere [6, Appendix 17].

Allied bombing attacks during World War II on German-controlled oil refineries and synthetic-fuel plants exemplify a more general type of *system interdiction* or “economic warfare.” The German military was crippled by the lack of fuel and lubricants caused by the “Oil Plan,” as the attack strategy was called. Interestingly, an acrimonious debate was waged between the proponents of the Oil Plan (system interdiction) and proponents of the “Rail Plan” (network interdiction). The Rail Plan sought to destroy rail lines and other transportation assets in Europe to restrict the movement of the German troops and equipment that would counter the Allied D-Day offensive. Ultimately, parts of both plans were implemented [7, pp. 75–78, 174–175].

Network interdiction is an important part of modern warfare where attacks on key civilian and military infrastructure can help reduce an enemy's fighting effectiveness, while incurring only limited casualties to “friendly forces” [8,9]. When planning for such attacks, the “interdictor” is typically faced with this question: Given limited attack resources and possibly other restrictions (e.g., political considerations), which network components should be attacked to reduce the enemy's war-fighting capabilities most effectively? BNI addresses this question.

To provide background on the modern, mathematical study of network interdiction, we first make the following definitions and assumptions [10]:

1. The interdictor, called *attacker* hereafter, acts first by using limited interdiction resources to attack components of his enemy's network.
2. The enemy, or *defender*, observes the damage caused by the attack(s) and then operates the damaged network so as to maximize his own, well-defined objective-function value.
3. The attacker understands the defender's capabilities and goals, and chooses attacks that minimize the defender's maximum achievable objective-function value.

These standard assumptions lead to the formulation of BNI and more general system-interdiction models as a type of Stackelberg game: a two-person, zero-sum, sequential-play game with two stages [11,12].

Wollmer [13] studies the problem of finding the single “most vital link” in a capacitated flow network, in his case, the arc whose deletion minimizes the maximum s - t flow in the network. (The reader who is unfamiliar with the terminology of network flows, e.g., “ s - t flows,” may wish to consult a standard text such as Ahuja *et al.* [14].) Here, the attacker has enough interdiction resource to attack and destroy exactly one arc, and the defender's objective is to maximize s - t flow. Wollmer's work may represent the earliest mathematical investigation of an instance of BNI, although as early as 1955 researchers were investigating a simpler “single-level” network interdiction problem that seeks to eliminate all s - t flow efficiently [15].

Danskin [16] presents some of the fundamental theory of “max–min models,” which may be viewed as a generalization of BNI to system interdiction. (His “min” and “max” are reversed compared to our convention). Confusingly, “max–min” and similar terms are also used in the context of the more common two-person, zero-sum, simultaneous-play games [17, pp. 143–165]. For example, von Neumann [18] proves the famous “minimax theorem” for such games.

Mathematical studies of BNI began in earnest during the Vietnam War, with models applied to disrupt the flow of enemy troops and materiel [19,20]. Fulkerson and Harding [21] and Golden [22] investigate the problem of maximizing the length of the shortest path in a network—to slow enemy reinforcements, say—using models in which the length of each network arc can be increased linearly, within limits, based on the amount of interdiction resource applied to it. (This is a “max–min” variant of BNI.) These models are solved as parametric linear programs (LPs). The k -most-vital-arcs problem [23,24] is similar, but at most k arcs may be interdicted and interdiction decisions are discrete. In particular, an arc is attacked and destroyed and its length becomes infinite, or it is left untouched and keeps its nominal length. Ball *et al.* [25] show that problem to be NP-complete. Israeli and Wood [26] extend the k -most-vital-arcs problem to general resource constraints: their study of the *shortest-path network interdiction problem* (or “maximizing the shortest path”), makes important theoretical and computational contributions to the solution of BNI, in general.

Ratliff *et al.* [27] extend Wollmer's [13] model to the problem of finding a set of n arcs in a capacitated network whose deletion minimizes the maximum s - t flow. While investigating problems of drug interdiction, Wood [28] generalizes that model to allow general interdiction resource constraints. (Phillips [29] considers a similar model, but allows some continuous interdiction effort; see also Steinrauf [30].) Wood shows that this *maximum-flow interdiction problem* is NP-complete, even when attacks are constrained only in cardinality. Thus, even the simpler model of Ratliff *et al.* appears to be difficult.

More general system-interdiction issues arise in the work of Grötschel *et al.* [31] and Medhi [32] who seek to evaluate the vulnerability of information networks to interdiction. Chern and Lin [33] study the interdiction of a system represented as a minimum-cost network-flow model.

Similar to the BNI studied by Wood [28], the network interdiction model of Washburn and Wood [34] aims at disrupting drug smuggling, and its efficient solution

involves maximum flows. But this is a two-person, zero-sum, simultaneous-play (Cournot) game, and its purpose is quite different from BNI. Specifically, an interdicator controls one or more “inspectors” who must be placed strategically on the arcs of a transportation network to maximize the probability of detecting a drug smuggler moving surreptitiously through that network. (If the smuggler traverses arc k when an inspector is present, the smuggler is detected with known probability p_k ; otherwise he goes undetected.) In a simultaneous-play model, neither player can observe the other’s actions before acting himself and, consequently, solutions define probabilistic (“mixed”) strategies for both players. In this case, the interdicator’s strategy defines a probability distribution over the inspectors’ locations, and the smuggler’s strategy defines a probability distribution over paths through the network. In contrast, a solution to BNI prescribes deterministic (“pure”) strategies for both players.

Deterministic strategies do not imply that BNI cannot incorporate uncertainty and probability, however. Cormican *et al.* [35] develop stochastic-programming versions of the maximum-flow interdiction model to handle uncertain interdiction successes and uncertain arc capacities. Whiteman [36], studying interdiction problems faced by the US Strategic Command, addresses uncertainty through Monte Carlo simulations of maximum-flow interdiction models. Pan *et al.* [37] maximize the expected probability of detecting a smuggler trying to transport stolen nuclear materials out of a country: nominal probabilities of detection can be improved by installing a limited number of radiation detectors at border crossings. Maximizing probability of detection is related, through a logarithmic transformation in the objective function, to shortest-path interdiction, and in that sense the model is deterministic. The model is stochastic, however, in that a probability distribution describes the smuggler’s origin in the network [38].

Brown *et al.* [39,10] develop a taxonomy for bilevel system-interdiction and system-defense models, as well as for trilevel system-defense models. The bilevel and

trilevel models are two-stage and three-stage Stackelberg games, respectively. BNI is an instance of a two-stage “attacker–defender model”; the trilevel defense models are “defender–attacker–defender models.” In the latter case, a defender wishes to employ his limited defensive resources as efficiently as possible to “interdict the interdicator,” with effectiveness being evaluated by solving a bilevel interdiction model. Bilevel system-defense models can be constructed, also; these are “defender–attacker models”; they apply when the value of attacking a system component is a fixed or easily computed value; they can be solved using the techniques described in this article; but will not be discussed further. We note that Brown *et al.* [10,40] also discuss solution techniques for all these model types and describe a number of applications to infrastructure protection. Indeed, vulnerability analysis for infrastructure is an important new application area for BNI.

Most recently, bilevel interdiction and defense models have been developed for a number of interesting applications: theater ballistic missile defense [41], planning attacks on multicommodity flow networks [42], planning attacks on communications networks [43], delaying the nuclear-weapons project of a “rogue state” [44], and attacking and defending electric-power grids [45]. The theory of “global Benders decomposition” developed in the last paper promises to be a useful computational technique for BNI (and other optimization problems), and is discussed later in this article.

The goal of the rest of this article is to describe basic theoretical models and solution techniques for BNI. A lack of space prohibits further detailed discussion of applications. More information on advanced computational techniques can be found in Magnanti and Wong [46], Israeli and Wood [26], Salmerón *et al.* [45], and Smith *et al.* [47].

A BASIC, BILEVEL, INTERDICTION MODEL

In abstract form, BNI may be stated as the following *attacker–defender model*:

$$\begin{aligned}
[\text{AD0}] \quad & \min_{\mathbf{x} \in X} z_0(\mathbf{x}), \text{ where} \\
& z_0(\mathbf{x}) \equiv \max_{\mathbf{y} \in Y(\mathbf{x})} f(\mathbf{x}, \mathbf{y}), \quad (1)
\end{aligned}$$

and where (i) $\mathbf{x} \in X$ denotes a binary vector of attack decisions that is limited by resources and perhaps logical restrictions (e.g., targets k and k' both cannot be attacked); (ii) $\mathbf{y} \in Y(\mathbf{x})$ denotes the activities that the defender will carry on after the attack, typically restricted by effects of the attack $\mathbf{x}$; and (iii) the objective function $f(\mathbf{x}, \mathbf{y})$ measures the functionality of the defender's network after the attack. Thus, the attacker seeks to minimize the functionality of the network which the defender is assumed to maximize. Of course, by switching the min and the max, we can model an attacker who seeks to maximize the cost of the defender's operations.

[AD0] is a special case of a Stackelberg game in which a leader (attacker) takes some action, and a follower (defender) observes that action and its effects, and then responds optimally given that information (12). [AD0] is a two-stage game and is finished after the follower responds; more general Stackelberg games may have many stages and/or players.

For the sake of concreteness and simplicity, further development of [AD0] assumes that

1. The defender's activities take place on network arcs, which are indexed by k , and there is a one-to-one correspondence with the attacker's potential targets.
2. The defender nominally optimizes network operation by solving the following LP which is feasible for any $\mathbf{u} \geq \mathbf{0}$

$$\max_{\mathbf{y}} \quad \mathbf{c}^T \mathbf{y} \quad (2)$$

$$\text{s.t.} \quad \mathbf{A} \mathbf{y} \leq \mathbf{b} \quad (3)$$

$$0 \leq \mathbf{y} \leq \mathbf{u}, \quad (4)$$

3. Restrictions on the attacker assume $\mathbf{x} = \mathbf{0}$ is feasible, and are represented by

$$X = \{\mathbf{x} \in \{0, 1\}^n \mid H\mathbf{x} \leq \mathbf{h}\}, \text{ and } (5)$$

4. $x_k = 1$ implies that activity k is attacked, its level forced to 0, that is, $y_k = \mathbf{0}$.

Then, defining $U = \text{diag}(\mathbf{u})$, [AD0] takes on the following specific form

$$[\text{ADLP1}] \quad z_1^* = \min_{\mathbf{x} \in X} z_1(\mathbf{x}), \text{ where} \quad (6)$$

$$z_1(\mathbf{x}) \equiv \max_{\mathbf{y}} \quad \mathbf{c}^T \mathbf{y} \quad (7)$$

$$\text{s.t.} \quad \mathbf{A} \mathbf{y} \leq \mathbf{b} \quad (8)$$

$$0 \leq \mathbf{y} \leq U(\mathbf{1} - \mathbf{x}). \quad (9)$$

The inner LP in [ADLP1] might represent a simple maximum-flow problem [27,28], the optimal deployment of the defenders' armed forces [48], or the production and distribution of oil or natural gas in a belligerent country [49]. [ADLP1] extends easily to attacks on nodes, attacks on groups of arcs and/or nodes, attacks that reduce capacity only partially, and so on, but such extensions are straightforward and not considered here.

A Stackelberg game with mixed-integer variables and having just two levels of decision making is called a *bilevel mixed-integer program* (BLMIP) [50]. However, the leader's and follower's objective functions in a BLMIP are not usually diametrically opposed as they are in [AD0]; for example, see Bard and Moore [51], Wen and Yang [52], and Hansen *et al.* [53]. In fact, most algorithms developed for BLMIPs assume a strong positive correlation between the leader's and follower's objective functions. Thus, the theory and algorithms for BLMIPs do not seem well-suited for handling [ADLP1], while the special-purpose methods described in this article have had demonstrated successes.

Standard LP theory tells us that $z_1(\mathbf{x})$ is a concave function in (continuous) $\mathbf{x}$. Thus, [ADLP1] is a difficult, nonconvex minimization problem. The problem can be "convexified," however, by moving $\mathbf{x}$ into the objective of the inner maximization.

Proposition 1 [54]. *Let $\bar{r}_k$ be an upper bound on the optimal dual variable for the*

constraint $y_k \leq u_k(1 - x_k)$ in [ADLP1] taken over all $\mathbf{x} \in X$. Let $\bar{\mathbf{r}} = (\bar{r}_1 \dots \bar{r}_n)^T$ and $\bar{R} = \text{diag}(\bar{\mathbf{r}})$, and define

$$[\text{ADLP2}] \quad z_2^* = \min_{\mathbf{x} \in X} z_2(\mathbf{x}), \text{ where} \quad (10)$$

$$z_2(\mathbf{x}) \equiv \max_{\mathbf{y}} (\mathbf{c}^T - \mathbf{x}^T \bar{R}) \mathbf{y} \quad (11)$$

$$\text{s.t. } A\mathbf{y} \leq \mathbf{b} \quad [\mathbf{q}] \quad (12)$$

$$0 \leq \mathbf{y} \leq \mathbf{u}. \quad [\mathbf{r}]. \quad (13)$$

(The vectors $\mathbf{q}$ and $\mathbf{r}$, used later, denote dual variables for their respective constraint sets when $\mathbf{x}$ is fixed.) Then, [ADLP1] and [ADLP2] are equivalent in the sense that $z_1^* = z_2^*$, and $\mathbf{x}^*$ solves [ADLP2] if and only if it also solves [ADLP1].

Note also that $z_2(\mathbf{x})$ is a convex function in continuous $\mathbf{x}$.

That [ADLP2] is equivalent to [ADLP1] is intuitively clear, at least when the $\bar{r}_k$ are strict upper bounds: $(\mathbf{x}^*, \mathbf{y}^*)$ solves [ADLP1] if and only if it also solves [ADLP2]. Nonstrict bounds can lead to cases where $(\mathbf{x}^*, \mathbf{y}^*)$ is optimal [ADLP2] but infeasible to [ADLP1]. In this case, however, there must exist some $\mathbf{y}^{**}$ such that $(\mathbf{x}^*, \mathbf{y}^{**})$ is optimal to [ADLP1].

To solve [ADLP2], (i) temporarily fix the variables $\mathbf{x}$ in [ADLP2] (i.e., treat $\mathbf{x}$ as data); (ii) take the dual of the resulting LP; and (iii) then release $\mathbf{x}$. The following, equivalent mixed-integer program (MIP) results:

$$[\text{ADMIP2}] \quad \min_{\mathbf{x} \in X, \mathbf{q}, \mathbf{r}} \mathbf{b}^T \mathbf{q} + \mathbf{u}^T \mathbf{r} \quad (14)$$

$$\text{s.t. } A^T \mathbf{q} + I\mathbf{r} + \bar{R}\mathbf{x} \geq \mathbf{c} \quad (15)$$

$$\mathbf{q} \geq \mathbf{0}, \mathbf{r} \geq \mathbf{0} \quad (16)$$

A standard LP-based branch-and-bound algorithm will solve [ADMIP2] if that model is not too large.

Good dual bounds $\bar{\mathbf{r}}$ are important for solving [ADMIP2] directly, but are not easy to come by except in a few instances. For instance, if the inner LP in [ADLP] corresponds to a maximum-flow model with integral capacity vector $\mathbf{u}$, then $\bar{r}_k = 1$ is valid

and tight because the value of an extra unit of arc capacity in a maximum-flow problem is 0 or 1 [35]. Theoretically, we could also use $\bar{r}_k = 100$ in this case, but the resulting LP relaxation of [ADMIP2] would be weak and solution times would suffer.

The decomposition solution method for [ADLP2] described next does not eliminate the need for good bounds on dual variables, but ancillary techniques can alleviate some of the difficulties caused by weak bounds, and decomposition has some key advantages over a branch-and-bound solution of [ADMIP2].

1. In the context of network interdiction, various studies [55,26], have demonstrated that decomposition typically solves [ADMIP2] much faster than does branch-and-bound,
2. As we shall see, decomposition can be extended to solve BNI even when the defender's optimization model is more general than an LP, and
3. Decomposition methods typically solve a sequence of "defender subproblems" in a familiar, user-friendly form, which obviates the complicated, unfamiliar dual constructs of [ADMIP2]. For instance, Salmerón *et al.* [45] evaluate the effects of an attack plan $\mathbf{x}$ on a large, regional electric-power grid using a standard electric-power model. In contrast, a MIP formulation for BNI in this case becomes unwieldy (and can only be solved for small, unrealistic test problems) [56].

BENDERS DECOMPOSITION

The decomposition algorithm described here may be viewed as solving [ADLP2] by applying Benders decomposition to [ADMIP2] [57]. The Benders methodology for solving a minimizing MIP first converts the MIP into a min-max problem by reversing the steps that we used to create [ADMIP2] from [ADLP2]: (i) temporarily fix the integer variables, (ii) take the dual of the resulting LP, and (iii) release the integer variables [58, pp. 135–143]. In the case of [ADMIP2], this

conversion returns us to the more natural starting point of [ADLP2].

An optimal solution in $\mathbf{y}$ to [ADLP1] occurs at an extreme point of the (bounded) feasible region of that problem's inner LP. Because of the essential equivalence of the problems, the same holds true for solutions in $\mathbf{y}$ to [ADLP2]. Let Y denote the full, finite set of extreme points for the latter problem. Then, [ADLP2] may be expressed as this *equivalent master problem*

$$z_2^*(Y) = \min_{\mathbf{x} \in X} z_2(\mathbf{x}, Y), \text{ where} \\ z_2(\mathbf{x}, Y) \equiv \max_{\hat{\mathbf{y}} \in Y} \left(\mathbf{c}^T - \mathbf{x}^T \bar{\mathbf{R}} \right) \hat{\mathbf{y}}. \quad (17)$$

That model has this formulation as a MIP

[ADMP2(Y)]

$$z_2^*(Y) = \min_{\mathbf{x}, z} z \quad (18)$$

$$\text{s.t. } z + \hat{\mathbf{y}}^T \bar{\mathbf{R}} \mathbf{x} \geq \mathbf{c}^T \hat{\mathbf{y}} \quad \forall \hat{\mathbf{y}} \in Y \quad (19)$$

$$\mathbf{x} \in X. \quad (20)$$

Benders decomposition dynamically generates constraints (19) called *Benders cuts*. It solves or approximately solves the original problem by finding $\hat{Y} \subset Y$, with $|\hat{Y}| \ll |Y|$ it is hoped, so that [ADMP2($\hat{Y}$)] is an adequate approximation of the equivalent master problem.

Algorithm A-1. Basic Benders decomposition algorithm to solve [ADLP2]

Input: An instance of [ADLP2] and allowable optimality gap $\epsilon \geq 0$.
Output: An ϵ -optimal interdiction plan for [ADLP2], and associated objective value;
Step 0: $\hat{Y} \leftarrow \emptyset$; $\underline{z} \leftarrow -\infty$; $\bar{z} \leftarrow \infty$; $\hat{\mathbf{x}} \leftarrow \mathbf{0}$; $\hat{\mathbf{x}}^* \leftarrow \mathbf{0}$; $gap \leftarrow \infty$;
Step 1: Fix $\mathbf{x} = \hat{\mathbf{x}}$ in [ADLP2] and solve for $\hat{\mathbf{y}}$ and objective value $\hat{z} \equiv z_2(\mathbf{x})$; $\hat{Y} \leftarrow \hat{Y} \cup \{\hat{\mathbf{y}}\}$;
If $\hat{z} < \bar{z}$ then $\bar{z} \leftarrow \hat{z}$ and $\hat{\mathbf{x}}^* \leftarrow \hat{\mathbf{x}}$;
If $\bar{z} - \underline{z} \leq \epsilon$ then go to Step 3;
Step 2: Solve [ADMP2($\hat{Y}$)] for $z_2^*(\hat{Y})$ and $\hat{\mathbf{x}}$; $\underline{z} \leftarrow z_2^*(\hat{Y})$;
If $\bar{z} - \underline{z} > \epsilon$ then go to Step 1;
Step 3: Print "Approximate solution is," $\hat{\mathbf{x}}^*$, "with objective value," $\bar{z}$;
Print "Provable optimality gap is", $\bar{z} - \underline{z}$;
Stop;

End of Algorithm A-1.

Algorithm A-1, or simply "A-1," is actually a special case of Benders decomposition that does not require "feasibility cuts" [59]. Such cuts are needed if the subproblems can become infeasible for certain $\mathbf{x} \in X$, which ours cannot by assumption. The correctness of the algorithm is easy to see: (i) The upper bound $\bar{z}$ is valid because it corresponds to some feasible solution of [ADLP2] for the minimizing attacker; (ii) the lower bound $\underline{z}$ is valid because it corresponds to a relaxation of the equivalent master problem [ADMP2(Y)]; (iii) if a solution $\hat{\mathbf{x}}$ ever repeats, it follows that $\bar{z} = \underline{z}$ and the algorithm must terminate; and (iv) the termination criterion is satisfied with $\underline{z} \neq \bar{z}$ or a solution repeats in a finite number of steps because X is a discrete, finite set.

The following section describes some enhancements to A-1 that may improve

solution speeds, and shows how global Benders decomposition can actually solve more general problems than [ADLP2].

IMPROVING AND GENERALIZING BENDERS DECOMPOSITION

Faster Solutions with Super-Valid Inequalities

The (*relaxed*) master problem [ADMP1($\hat{Y}$)] can be strengthened in some instances by adding *super-valid inequalities* (SVIs). Intuitively, this strengthening can help alleviate some of the difficulties caused by weak dual bounds $\bar{\mathbf{r}}$. SVIs are similar to valid inequalities of integer-programming theory [60, pp. 205–295] except that they may, and typically do, eliminate feasible solutions from the

master problem. We define SVIs with respect to general MIPs.

Definition. Let $\mathbf{x}$ and $\mathbf{y}$ denote the vectors of integer and continuous variables, respectively, in a MIP. The inequality $\mathbf{w}_1^T \mathbf{x} + \mathbf{w}_2^T \mathbf{y} \geq w_0$ is *super-valid* for this MIP if (i) adding that inequality to the MIP does not eliminate all optimal solutions, or (ii) an incumbent solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is (already) optimal for the MIP.

With proper precautions, SVIs may be used within a branch-and-bound algorithm for [ADMIP2] as well as within Benders decomposition for solving [ADLP2]. Suppose, for instance, that we add a single SVI in the course of solving a MIP by either technique. If case (i) is true for that SVI, then an optimal solution will still be found via enumeration because (a) some optimal solution is still feasible, and (b) any lower bound obtained from a relaxation of the SVI-modified MIP is still a valid lower bound on z_2^* . Thus, standard fathoming tests within a branch-and-bound algorithm and the convergence tests in A-1 are valid. If case (ii) is true when we add the SVI, we simply want our algorithm to halt with a message that the incumbent $\hat{\mathbf{x}}^*$ is optimal, and this is easy to arrange. After adding the SVI

1. If the MIP is found to be infeasible, or $\bar{z} > z(\hat{\mathbf{x}}^*)$, we declare the incumbent optimal, which it is, or
2. If $\bar{z} - \underline{z} \leq \epsilon$ occurs, we declare the incumbent to be ϵ -optimal, which it is. (As often happens, our incumbent is optimal, but we only prove it to be ϵ -optimal.)

By induction, it follows that an enumeration algorithm incorporating a finite number of SVIs will also terminate correctly and finitely.

One type of SVI for our enhanced version of A-1 applied to [ADLP2] is easy to derive.

Proposition 2 [26]. Let $z + \hat{\mathbf{y}}^T \bar{\mathbf{R}} \mathbf{x} \geq \mathbf{c}^T \hat{\mathbf{y}}$ denote a Benders cut from Algorithm A-1 being used to solve [ADLP2].

$$I_k(\hat{\mathbf{y}}) = \begin{cases} 1 & \text{if } \hat{y}_k > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

Then, the following inequality is super-valid:

$$\mathbf{I}(\hat{\mathbf{y}})^T \mathbf{x} \geq 1. \quad (22)$$

Suppose that $\bar{\mathbf{R}}$ derives from strict dual bounds and that $\hat{\mathbf{y}}$ is the response to the feasible interdiction plan $\hat{\mathbf{x}}$. It follows that $\hat{\mathbf{y}}^T \bar{\mathbf{R}} \hat{\mathbf{x}} = \mathbf{I}(\hat{\mathbf{y}})^T \hat{\mathbf{x}} = 0$, and that $\hat{\mathbf{x}}$ is made infeasible by the SVI $\mathbf{I}(\hat{\mathbf{y}})^T \mathbf{x} \geq 1$. That is, the inequality $\mathbf{I}(\hat{\mathbf{y}})^T \mathbf{x} \geq 1$ is not valid in the standard sense. Note also that $\hat{\mathbf{x}}$ could be an optimal solution which is made infeasible by the inequality, but if we already have an optimal solution in hand, we have free reign to restrict the solution in any way we like.

A simple extension of Proposition 2 leads to

Corollary 1. For every Benders cut $z + \hat{\mathbf{y}}^T \bar{\mathbf{R}} \mathbf{x} \geq \mathbf{c}^T \hat{\mathbf{y}}$, the SVI of Proposition 2 can be tightened to $\mathbf{I}(\hat{\mathbf{y}})^T \mathbf{x} \geq 2$ if $\mathbf{c}^T \hat{\mathbf{y}} - \max_i \bar{r}_i \hat{y}_i > \bar{z}$, can be tightened to $\mathbf{I}(\hat{\mathbf{y}})^T \mathbf{x} \geq 3$ if $\mathbf{c}^T \hat{\mathbf{y}} - \max_{k \neq k'} \{\bar{r}_k \hat{y}_k + \bar{r}_{k'} \hat{y}_{k'}\} > \bar{z}$, and so on.

Of course, as $\bar{z}$ changes during the course of A-1, it may be possible to tighten previously generated SVIs.

Modifications of A-1 to incorporate SVIs are straightforward, and SVIs may improve solution times substantially. Israeli and Wood [26] demonstrate this and show how to (i) add heuristically generated SVIs to an instance of [ADMIP2] to improve branch-and-bound solution times, (ii) generalize SVIs to “ ϵ -SVIs” that guarantee not to eliminate all ϵ -optimal solutions, and (iii) solve [ADLP1] and [ADLP2] in a decomposition algorithm whose master problem constraints consist solely of SVIs. The “covering algorithm” alluded to in (iii) uses no dual bounds $\bar{\mathbf{r}}$ at all, and converts Benders decomposition into a purely combinatorial procedure.

Standard Computational Enhancements for Benders Decomposition

In addition to employing SVIs, Algorithm A-1 can benefit from more standard techniques

used to improve solution speeds for Benders decomposition [46]. Some of these techniques are discussed briefly below.

1. For fixed $\mathbf{x}$, [ADLP2] may have multiple extreme-point solutions $\mathbf{y}$ and a different Benders cut can be generated for each. Adding too many such cuts can slow down solutions of [ADMP2], but adding them judiciously can improve solution times greatly. Israeli and Wood [26] use this technique in the shortest-path interdiction problem, where they enumerate multiple shortest paths for a single attack plan $\mathbf{x}$ and generate cuts for each.
2. The master problem need not be solved to optimality if “sufficient progress” is made after each cut is added [61].
3. Cuts derived from interior-point subproblem solutions $\hat{\mathbf{y}}$ may prove better than those derived from extreme-point solutions. For instance, an arc k with a large flow $\hat{y}_k$ on it appears as an attractive candidate for interdiction in the solution of the maximum-flow interdiction problem. That is, it generates a cut with a large-magnitude entry in position k . But $\hat{y}_k$ may be large primarily because the solution is an extreme-point solution, not because it must be large to achieve a maximum s - t flow. So, A-1 may waste time exploring solutions with $x_k = 1$. In contrast, an interior-point solution “spreads flow around” the network, and $\hat{y}_k$ will tend to be large only if it needs to be in order to achieve a maximum s - t flow. Consequently, better guidance and better cuts may be derived from such solutions.
4. Some Benders cuts can be dominated (implied) by others, and the nondominated ones ought to be used for the sake of efficiency. Magnanti and Wong [46] provide guidance on this topic. The related work of Smith *et al.* [47] may also prove useful: that paper shows how a polynomial-sized reformulation of the master problem in an interdiction model can yield cuts that dominate an exponential number of cuts from the original formulation.

Global Benders Decomposition

Algorithm A-1 can be extended to solve instances of [AD0], in which the defender’s operational model is more complicated than an LP. For example, let “[ADIP2]” denote a model identical to [ADLP2] except that $\mathbf{y}$ is required to be integral. A-1 clearly solves this problem because we can replace “the finite set of extreme points Y ” used to define the equivalent master problem [ADMP2] with “the finite set of integer solutions” for [ADIP2]. Geoffrion [62] coins the phrase “generalized Benders decomposition” to describe extensions of Benders decomposition to nonlinear models analogous to [ADLP2] with convex objective functions $z_2(\mathbf{x})$; Salmerón *et al.* [45] therefore use the phrase “global Benders decomposition” to describe the solution of other models like [ADIP2], in which the issues of convexity may even be irrelevant.

In [ADIP2], $\bar{r}_k$ no longer corresponds to a bound on a dual variable. Rather, that datum must comprise an integral part of the original formulation. For instance, [ADIP2] might correspond to a max–min instance of BNI in which an attacker seeks to delay completion of a defender’s project, which is modeled through the constructs of a resource-constrained PERT network. (Davis [63] discusses such PERT networks, and Brown *et al.* [40,44] discuss interdicting them.) If $x_k = 1$, task k in the project is attacked and delayed by a fixed amount: that is what $-\bar{r}_k$ would represent in [ADIP2]; otherwise $x_k = 0$ and task k requires some nominal time to complete, corresponding to c_k in that model.

We can generalize further.

Proposition 3 [45]. *Suppose that BNI has the following form:*

$$\begin{aligned}
 \text{[AD3]} \quad & \min_{\mathbf{x} \in X} z_3(\mathbf{x}), \text{ where} \\
 & z_3(\mathbf{x}) \equiv \max_{\mathbf{y} \in Y} f(\mathbf{x}, \mathbf{y}), \quad (23)
 \end{aligned}$$

and where X is defined as in Equation (5), $\mathbf{y} \in Y$ can be discrete and/or continuous, and $f(\mathbf{x}, \mathbf{y})$ has a general form. Furthermore, suppose that penalty vectors $\mathbf{v}(\mathbf{x})$ can be defined so that

$$z_3(\mathbf{x}) \geq z_3(\hat{\mathbf{x}}) + \mathbf{v}^T(\hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}}) \quad \forall \mathbf{x}, \hat{\mathbf{x}} \in X. \quad (24)$$

Then, the following master problem is equivalent to [AD3]:

$$\begin{aligned} & \min_{\mathbf{x} \in X, z} z \\ & \text{s.t. } z - \mathbf{v}(\hat{\mathbf{x}})^T \mathbf{x} \geq z_3(\hat{\mathbf{x}}) - \mathbf{v}^T(\hat{\mathbf{x}})\hat{\mathbf{x}} \quad \forall \hat{\mathbf{x}} \in X \end{aligned} \quad (25)$$

Given the existence of an equivalent master problem, [AD3] may be solved via a modified version of A-1. We may assume that the attacker has an efficient method for computing $z_3(\hat{\mathbf{x}})$, that is, for evaluating the effects of an attack plan $\hat{\mathbf{x}}$ through the solution of the defender's subproblem. For instance, to evaluate the effects of attack plan $\hat{\mathbf{x}}$ on an electric-power transmission grid, the attacker can solve a nonlinear "AC optimal power-flow model" or a standard, faster, LP approximation, a "DC optimal power-flow model" [64]. Thus, the difficult part here will be defining and computing appropriate penalty vectors $\mathbf{v}(\mathbf{x})$. That task will be problem-dependent, so we expand upon the power-grid example to illustrate.

An attacker wishes to maximize the short-term, unserved demand for power in a defender's transmission grid. Thus, [ADMP3] must be converted to a maximization problem, and the inequality in Equation (26) reversed. When $\hat{x}_k = 0$, $v_k(\hat{\mathbf{x}})$ should bound the amount of unserved demand that will accrue if the status of grid component k is changed from "unattacked and functional" to "attacked and nonfunctional." The power-handling capability of the component provides a simple bound which is usually valid. If $\hat{x}_k = 1$, $v_k(\hat{\mathbf{x}})$ must reflect how much unserved demand will be eliminated if component k 's status is changed in the opposite direction. Because of the existence of series components, $v_k(\hat{\mathbf{x}}) = 0$ is a reasonable, albeit crude, approximation. (Actually, unserved demand can increase after repairing a component, but this does not normally cause difficulties [45]).

The subproblem in this example is merely a LP, and an attack on component k does force its capacity to 0 as in [ADLP1]. A

complication arises, however, because the destruction of a component can also improve power flow by eliminating one or more "susceptance constraints" between power lines having common end points [64]. Thus, $z_0(\mathbf{x})$ is neither concave nor convex in this application. However, this function tends to be well-behaved in practice, and the corresponding penalty vector $\mathbf{v}(\mathbf{x})$ is easily computed. The definition of that vector may seem simplistic, but Salmerón *et al.* [45] use it within global Benders decomposition to solve interdiction models on full-scale, regional transmission grids. An added benefit of the global Benders approach is that it extends to the trilevel network-defense problem for a power grid.

CONCLUSIONS

This article has described mathematical techniques for modeling and solving a BNI. BNI is a two-person, zero-sum, two-stage, sequential-play (Stackelberg) game whose solution prescribes an optimal application of limited resources to attack components of an enemy's network, and thereby limit that network's usefulness to the enemy. When a LP suffices to model optimal network operation, we show that BNI can be converted to and solved as a MIP. But, we describe special decomposition techniques that typically solve these problems more efficiently and, importantly, can solve more general problems.

REFERENCES

1. Webster. Webster's third new international dictionary. Springfield (MA): Merriam-Webster, Inc.; 1993.
2. Herodotus. Translated by Grene D, editor. The history. Chicago: University of Chicago Press; 1987.
3. Livy (Titus Livius). Translated by de Sélincourt A, editor. The war with Hannibal. Middlesex, England: Penguin Books; 1965.
4. Polybius. Translated by Scott-Kilvert I, editor. The rise of the Roman Empire. London: Penguin Books; 1979.
5. Foote S. The civil war. New York: Random House; 1974.

6. Blair C. Hitler's U-Boat war. New York: Random House; 1996.
7. MacIsaac D. Strategic bombing in world war two. New York: Garland Publishing, Inc.; 1976.
8. Joint Chiefs of Staff. Doctrine for joint operations. Joint Pub 3-0; Available at http://www.dtic.mil/doctrine/jel/new_pubs/jp3_0.pdf. Accessed 1995.
9. Joint Chiefs of Staff. Doctrine for joint interdiction operations. Joint Pub 3-03; Available at http://www.dtic.mil/doctrine/jel/new_pubs/jp3_03.pdf. Accessed 1997.
10. Brown G, Carlyle M, Salmerón J, *et al*. Defending critical infrastructure. *Interfaces* 2006;36:530–544.
11. von Stackelberg H. The theory of the market economy. London: German, William Hodge & Co.; 1952.
12. Simaan M, Cruz JB. On the stackelberg strategy in nonzero-sum games. *J Optim Theor Appl* 1973;11:533–555.
13. Wollmer R. Removing arcs from a network. *Oper Res* 1964;12:934–940.
14. Ahuja RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms, and applications. Upper Saddle River (NJ): Prentice-Hall; 1993.
15. Harris TE, Ross FS. Fundamentals of a method for evaluating rail net capacities. Research Memorandum RM-1573, Santa Monica (CA): The RAND Corporation; 1955.
16. Danskin JM. The theory of max-min, with applications. *SIAM J Appl Math* 1966;14:641–664.
17. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton University Press; 1953.
18. (a) von Neumann J. Zur theories der gesellschaftsspiele. *Math Ann* 1928;100:295–320; (b) translated by Bergmann S, Luce RD, Tucker AW, editors. Contributions to the theory of games IV. Princeton (NJ): Princeton University Press; 1959. pp. 13–42.
19. McMasters AW, Mustin TM. Optimal interdiction of a supply network. *Naval Res Logist Q* 1970;17:261–268.
20. Ghare PM, Montgomery DC, Turner TM. Optimal interdiction policy for a flow network. *Naval Res Logist Q* 1971;18:37–45.
21. Fulkerson DR, Harding GC. Maximizing the minimum source-sink path subject to a budget constraint. *Math Program* 1977;13:116–118.
22. Golden B. A problem in network interdiction. *Naval Res Logist Q* 1978;25:711–713.
23. Corley HW, Shaw DY. Most vital links and nodes in weighted networks. *Oper Res Lett* 1982;1:157–160.
24. Malik K, Mittal AK, Gupta SK. The k-most vital arcs in the shortest path problem. *Oper Res Lett* 1989;8:223–227.
25. Ball MO, Golden BL, Vohra RV. Finding the most vital arcs in a network. *Oper Res Lett* 1989;8:73–76. (NP-completeness of BSI).
26. Israeli E, Wood RK. Shortest-path network interdiction. *Networks* 2002;40:97–111.
27. Ratliff HD, Sicilia GT, Lubore SH. Finding the n most vital links in flow networks. *Manage Sci* 1975;21:531–539.
28. Wood RK. Deterministic network interdiction. *Math Comput Model* 1993;17:1–18.
29. Phillips CA. The network inhibition problem. STOC '93: Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing. New York: ACM Press; 1993. pp. 776–785.
30. Steinrauf R. A network interdiction model [Master's Thesis]. Monterey (CA): Operations Research Department, Naval Postgraduate School; 1991.
31. Grötschel M, Monma C, Stoer M. Computational results with a cutting plane algorithm for designing communication networks with low-connectivity constraints. *Oper Res* 1992;40:309–330.
32. Medhi D. A unified approach to network survivability for teletraffic networks: models, algorithms and analysis. *IEEE Trans Commun* 1994;42:534–548.
33. Chern MS, Lin KC. Interdicting the activities of a linear program - a parametric analysis. *Eur J Oper Res* 1995;86:580–591.
34. Washburn AR, Wood RK. Two-person zero-sum games for network interdiction. *Oper Res* 1994;43:243–251.
35. Cormican KJ, Morton DP, Wood RK. Stochastic network interdiction. *Oper Res* 1998;46:184–197.
36. Whiteman PS. Improving single strike effectiveness for network interdiction. *Mil Oper Res* 1999;4(1):15–30.
37. Pan F, Charlton W, Morton D. A stochastic program for interdicting smuggled nuclear material. In: Woodruff DL, editor. Network interdiction and stochastic integer programming. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2003. pp. 1–20.
38. Morton DP, Pan F, Saeger KJ. Models for nuclear smuggling interdiction. *IIE Trans* 2007;39:3–14.

39. Brown G, Carlyle M, Royset J, *et al.* On the complexity of delaying an adversary's project. In: Golden B, Raghavan S, Wasil E, editors. *The next wave in computing, optimization and decision technologies*. New York: Springer; 2005. pp. 3–17.
40. Brown G, Carlyle M, Salmerón J, *et al.* Analyzing the vulnerability of critical infrastructure to attack, and planning defenses. In: Greenberg H, Smith J, editors. *Tutorials in operations research: emerging theory, methods, and applications*. Hanover (MD): Institute for Operations Research and Management Science; 2005.
41. Brown G, Carlyle M, Diehl D, *et al.* A two-sided optimization for theater ballistic missile defense. *Oper Res* 2005;53:263–275.
42. Lim C, Smith JC. Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Trans* 2007;39:15–26.
43. Barkley TR. An attacker-defender model for IP-based networks [Master's Thesis]. Monterey (CA): Operations Research Department, Naval Postgraduate School; 2008.
44. Brown G, Carlyle M, Harney R, *et al.* Interdicting a nuclear weapons project. *Oper Res* 2009;57:866–877.
45. Salmerón J, Wood K, Baldick R. Worst-case interdiction analysis of large-scale electric power grids. *IEEE Trans Power Syst* 2009;24:96–104.
46. Magnanti TL, Wong RT. Accelerating Benders decomposition: algorithmic enhancement and model selection criteria. *Oper Res* 1981;29:464–484.
47. Smith JC, Lim C, Alptekinoglu A. Optimal mixed-integer programming and heuristic methods for a bilevel stackelberg product introduction game. *Naval Res Logist* 2009;56:714–729.
48. Morton D, Rosenthal RE, Lim T. Optimization modeling for airlift mobility. *Mil Oper Res* 1997;1(4):49–68.
49. Mehring JS, Gutterman MM. Supply and distribution planning support for Amoco (U.K.) Limited. *Interfaces* 1990;20(4): 95–104.
50. Ben-Ayed O. Bi-level linear programming. *Comput Oper Res* 1993;20:485–501.
51. Bard J, Moore J. A branch and bound algorithm for the bi-level programming problem. *SIAM J Sci Stat Comput* 1990;11: 281–292.
52. Wen U, Yang Y. Algorithms for solving the mixed integer two-level linear programming problem. *Comput Oper Res* 1990;17:133–142.
53. Hansen P, Jaumard B, Savard G. New branch-and-bound rules for linear bi-level programming. *SIAM J Sci Stat Comput* 1992;13:1194–1217.
54. Morton DP, Wood RK. Restricted recourse bounds for stochastic linear programming. *Oper Res* 1999;47:943–956.
55. Cormican KJ. Computational methods for deterministic and stochastic network interdiction problems [Masters Thesis]. Monterey (CA): Operations Research Department, Naval Postgraduate School; 1995.
56. Motto AL, Arroyo JM, Galiana FD. A mixed-integer LP procedure for the analysis of electric grid security under terrorist threat. *IEEE Trans Power Syst* 2005;20:1357–1365.
57. Benders JF. Partitioning procedures for solving mixed integer variables programming problems. *Numer Math* 1962;4:238–252.
58. Garfinkel RS, Nemhauser GL. *Integer Programming*. New York: John Wiley & Sons; 1972.
59. Birge JR. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper Res* 1985;33:989–1007.
60. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. New York: Wiley-Interscience; 1988.
61. Brown GG, Graves G, Honczarenko M. Design and operation of a multicommodity production/distribution system using primal goal decomposition. *Manage Sci* 1987;33: 1469–1480.
62. Geoffrion A. Generalized Benders decomposition. *J Optim Theor Appl* 1972;10:237–260.
63. Davis EW. Project scheduling under resource constraints: historical review and categorization of procedures. *AIIE Trans* 1973;5: 297–313.
64. Overbye TJ, Cheng X, Sun Y. A comparison of the AC and DC power flow models for LMP calculations. *Proceedings, 37th Hawaii International Conference on System Sciences*. Hawaii: 2004.

BILINEAR OPTIMIZATION

CHURLZU LIM
University of North Carolina at
Charlotte, Charlotte, North
Carolina

PROBLEM DESCRIPTION

Bilinear optimization (or *bilinear programming*) problem (BLP) is a specially structured quadratic programming problem for general quadratic programming, where all quadratic terms are bilinear with respect to two disjoint sets of variables. Let $x \in R^n$ and $y \in R^q$ denote these two sets of variables. Then, a commonly used form of BLP can be written as follows: (More general forms of BLP problems will be discussed later.)

BLP:

$$\text{Minimize } f(x,y) = c^T x + d^T y + x^T Q y \quad (1a)$$

$$\text{subject to } x \in X = \{x \in R^n : Ax \geq b, x \geq 0\} \quad (1b)$$

$$y \in Y = \{y \in R^q : Fy \geq g, y \geq 0\}, \quad (1c)$$

where $c \in R^n$, $d \in R^q$, $Q \in R^{n \times q}$, $A \in R^{m \times n}$, $b \in R^m$, $F \in R^{p \times q}$, and $g \in R^p$. We assume that all vectors are column vectors unless otherwise noted, and T denotes the transpose operator. Note that there are quadratic terms in the objective function. Furthermore, x and y appear exclusively in the respective sets of constraints in Equations (1b) and (1c). Therefore, if the quadratic term in the objective function is removed, the problem is simply two separable linear programming problems. Due to these quadratic terms, however, the objective function becomes nonconvex, and hence, the problem becomes significantly difficult to solve. In particular, BLP is known to be *strongly NP-hard* [1].

Konno [2] lists important application problems that can be formulated as BLP. Such applications include game-theoretic problems (e.g., constrained bimatrix game, sequential game under perfect information), Markovian decision process (e.g., multistage Markovian assignment problem, multistage production, and sales optimization), complementary planning (e.g., complementary flows in a network, orthogonal production scheduling), and reformulation of other mathematical programming problems (e.g., concave quadratic minimization, 0–1 integer program). Other application problems include an inverse optimal value problem [3], estimation of population in finite-source election queues [4], and multicommodity network interdiction [5], to name a few.

CHARACTERIZATION OF BLP

Without loss of generality, assume that X and Y are nonempty and bounded. (Note that the problem can be unbounded if one or both of X and Y are unbounded. We will discuss how to check the unboundedness of BLP for such cases later in this section.) Therefore, there exists an optimal solution having a finite objective function value. One important property of BLP is that there exists an extreme point optimal solution (x^*, y^*) where x^* and y^* are extreme points of X and Y , respectively [6–8]. To be more specific, suppose that $(\bar{x}, \bar{y})$ is a nonextreme point optimal solution of BLP. Then, fixing $\bar{y}$ and solving BLP, we can find an extreme point solution x^* such that $f(x^*, \bar{y}) = f(\bar{x}, \bar{y})$ since it is a bounded linear programming problem. In turn, fixing x^* and solving BLP will yield an extreme point solution y^* such that $f(x^*, y^*) = f(x^*, \bar{y})$ due to the same reason as before. In consequence, there exists an extreme point optimal solution (x^*, y^*) of BLP.

Next, BLP can be interpreted as a concave minimization problem of either variable x or variable y . To illustrate this interpretation, consider the following problem, where

decision on y is delayed.

$$\begin{aligned} \text{Minimize} \quad & f_x(x) = c^T x + \left[\min_{y \in Y} d^T y + x^T Q y \right] \\ & (2a) \end{aligned}$$

$$\text{subject to} \quad x \in X. \quad (2b)$$

Since Y is nonempty and bounded, the above problem can be written as follows:

$$\text{Minimize} \quad \min_{y \in Y_E} \{ (c + Qy)^T x + d^T y \} \quad (3a)$$

$$\text{subject to} \quad x \in X, \quad (3b)$$

where Y_E denotes the set of all extreme points of Y . Note that the objective function is the minimum of a finite number of affine functions, which is a (multidimensional) piecewise affine concave function [9]. (Symmetrically, BLP can be interpreted as a concave minimization problem in variable y .)

BLP can be also viewed as a min-max problem. Let $v \in R^p$ denote the dual variable of the inner linear program in the problem (2). Then, since Y is nonempty and bounded, we can equivalently write the inner minimization problem as follows:

$$\begin{aligned} \text{Minimize} \quad & c^T x + \left[\begin{array}{ll} \max & g^T v \\ \text{s.t.} & F^T v \leq d + Q^T x \\ & v \geq 0 \end{array} \right] \\ & (4a) \end{aligned}$$

$$\text{subject to} \quad x \in X. \quad (4b)$$

Symmetrically, we can write an equivalent min-max problem by delaying decisions on x and taking the dual linear program of the inner problem in x .

Among numerous solution methods proposed in literature during the last four decades, we can observe two main-stream approaches; one is to exploit *Tuy cuts* (also referred to as *concavity cuts*) [10], and the other is to implicitly enumerate extreme points.

SOLUTION APPROACHES VIA TUY CUT AND ITS VARIANTS

Tuy cut was originally introduced for solving concave minimization problems in which a

concave objective function is minimized over a bounded polyhedral set [10], and it has played an important role in various algorithms for solving BLP as well as concave minimization problems. For illustrating Tuy cuts, consider the problem (2). Then, a Tuy cut in the x -variable space is generated by constructing a simplex having $n + 1$ vertices, at which objective function values are greater than or equal to the best available objective value, called the *incumbent value*. Due to the concavity of the objective function, it is guaranteed that the objective function value at a point in the simplex does not exceed the incumbent value. To be more specific, this simplex can be constructed as follows:

Consider a nondegenerate extreme point, x_1 , at which the objective function value is the incumbent value of $\bar{f}_x$. Due to nondegeneracy, there are n adjacent extreme points, $x_2, x_3, \dots, x_{n+1}$, which can be reached after single pivot. Furthermore, $x_1, x_2, \dots, x_{n+1}$ are affinely independent (i.e., $(x_2 - x_1), (x_3 - x_1), \dots, (x_{n+1} - x_1)$ are linearly independent). Assume that $f_x(x_j) \geq \bar{f}_x$ for $j = 2, 3, \dots, n + 1$. That is, x_1 is a local minimum in reference to its adjacent extreme points. Let $\lambda_j = \max\{\lambda > 0 : f(x_1 + \lambda(x_j - x_1)) = \bar{f}_x\}$. Furthermore, let $x'_j = x_1 + \lambda_j(x_j - x_1)$, that is, x'_j is on the ray $x_1 + \lambda(x_j - x_1)$ and is the farthest point from x_1 among points having the objective function value as $\bar{f}_x$. When such points exist on all n rays, the simplex is constructed by $(n + 1)$ vertices, $x_1, x'_2, x'_3, \dots, x'_{n+1}$. Accordingly, the Tuy cut eliminates the (open) halfspace which is formed by the hyperplane passing through $x'_2, x'_3, \dots, x'_{n+1}$ and by the region containing x_1 (see Fig. 1(a) for an illustrative example of $n = 2$). Tuy [10] exploits these cuts in the context of his *cone partitioning algorithm*, in which the cone defined by x_1 and n rays is recursively partitioned into subcones by adding a ray passing through a feasible point located farthest from the extreme point with respect to the Tuy cut until all subcones are covered by their own Tuy cuts. (See Fig. 1(b) for illustration of cone partitioning.)

One of the earliest adoptions of Tuy cuts within the context of BLP is made by Konno [6,8] who employs Tuy cuts to find ε -optimal solutions. Gallo and Ülkücü [11] employed

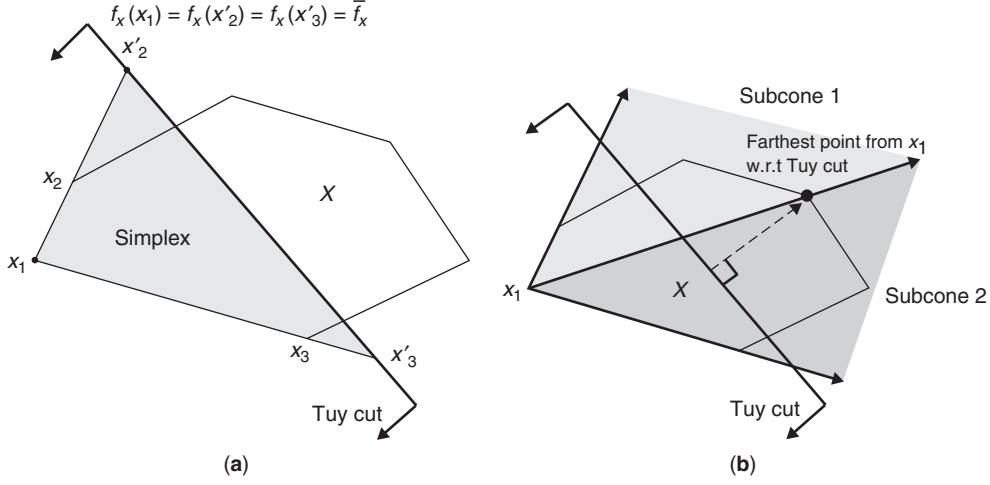


Figure 1. Illustrative examples for Tuy cut generation and cone partitioning. (a) A Tuy cut is generated at an incumbent extreme point x_1 . Due to concavity, $f_x(x) \geq \bar{f}_x$ for each x in the simplex. (b) Cone partitioning is made by adding a ray connecting x_1 and the extreme point of X which is farthest from x_1 with respect to the Tuy cut [23].

Tuy's cone partitioning algorithm in a straightforward manner. In Serali and Shetty [12], this type of cut is further enhanced by generating deeper cuts exploiting negative edge extension. Furthermore, in order to ensure finite convergence, their algorithm additionally generates disjunctive face cuts. Their negative edge extension scheme can be further enhanced by tighter bounding schemes of Ding and Al-Khayyal [13].

IMPLICIT ENUMERATION

Another main stream of solution efforts in BLP is implicit enumeration of extreme points. One notable approach is to solve the min-max formulation (4) via branch-and-bound methods. Note that the feasible region of (x, v) in the min-max problem is $S_{xv} = \{(x, v) : Ax \geq b, F^T v - Q^T x \leq d, x \geq 0, v \geq 0\}$. Furthermore, it has an optimal solution that is an extreme point of S_{xv} [7].

Consider the following relaxation of the min-max problem where the inner "maximization" is removed from the min-max problem.

$$\text{Minimize } c^T x + g^T v \quad (5a)$$

$$\text{subject to } F^T v - Q^T x \leq d \quad (5b)$$

$$Ax \geq b \quad (5c)$$

$$x, v \geq 0. \quad (5d)$$

Note that the optimal objective function value of this linear program is a lower bound of the optimal objective function value of the min-max problem (4). Let $\hat{x}$ denote the solution of x -variable to the linear program. Furthermore, let $\hat{y}$ be the solution to the following inner maximization problem.

$$\text{Maximize } g^T v \quad (6a)$$

$$\text{subject to } F^T v \leq d + Q^T \hat{x} \quad (6b)$$

$$v \geq 0. \quad (6c)$$

Since $(\hat{x}, \hat{y}) \in S_{xv}$, $c^T \hat{x} + g^T \hat{y}$ is an upper bound of the optimal objective value of the min-max problem (4). Hence, when the values of Equations (5) and (6) coincide, $(\hat{x}, \hat{y})$ is a global optimum of the min-max problem. Branching is triggered by designating the basic variable(s) in $\hat{x}$ as nonbasic. Note that there can be up to $q + m$ branches (see Equations (5b) and (5c)).

This branch-and-bound concept is also employed by Hansen *et al.* [1] for solving

linear bilevel programming problems via dichotomous branching, which is based on the complementary slackness of the inner linear program. Audet *et al.* [14] (and Alarie *et al.* [15] subsequently) employed the similar branching and bounding ideas to both symmetric min–max formulations as well as symmetric min–min formulations (i.e., delaying x or y as in Equation (2)).

LOCAL SEARCH

As far as approximation algorithms are concerned, so-called *mountain climbing* and *augmented mountain climbing* methods of Konno [6,8] are perhaps the most commonly practiced in literature as a local search procedure. Note that, given $x \in X$ (or $y \in Y$), BLP is a linear program in y (or x). Let $LP_{y|x}$ and $LP_{x|y}$ denote these linear programs, respectively. Then, starting at an initial point $x^0 \in X$ (typically, a basic feasible solution of X), the mountain climbing method alternately solves $LP_{y|x^i}$ and $LP_{x|y^i}$, where y^i is the solution to $LP_{y|x^i}$ and x^{i+1} is the solution to $LP_{x|y^i}$. The augmented mountain climbing method additionally employs a local search from the final solution of the mountain climbing method to find a better adjacent pair of extreme points, where the mountain climbing is employed again. The entire procedure is repeated until there is no better adjacent pair of extreme points.

BILINEAR PROGRAMMING AND GAME THEORY

While BLP has been used in many applications, its broad applicability comes from two types of game-theoretic problems; bimatrix game with side constraints and sequential game under perfect information. Bimatrix game is a two-person nonzero-sum and noncooperative game [16], where two players (denoted by P1 and P2, respectively) maximize their own payoffs without collaboration between the players. Let $X_G = \{x \in R^n : Ax \leq b, e_n^T x = 1, x \geq 0\}$ and $Y_G = \{y \in R^q : Fy \leq g, e_q^T y = 1, y \geq 0\}$ denote the sets of feasible mixed strategies of two players, where e_k is a vector of k ones. That is,

$x \in X_G$ (or $y \in Y_G$) is a vector of probabilities of n (q) pure strategies, where x (y) has additional side constraints $Ax \leq b$ ($Fy \leq g$). Furthermore, let $Q^1 = [q_{ij}^1] \in R^{n \times q}$ and $Q^2 = [q_{ij}^2] \in R^{n \times q}$ denote respective payoff tables of P1 and P2. Hence, q_{ij}^1 (or q_{ij}^2) is the payoff given to P1 (or P2), when pure strategies of P1 and P2 are i and j , respectively.

Note that, given (x, y) , expected payoffs of P1 and P2 are $x^T Q^1 y$ and $x^T Q^2 y$, respectively. A Nash equilibrium of this bimatrix game is mixed strategies (x^*, y^*) such that

$$(x^*)^T Q^1 y^* = \max \{x^T Q^1 y^* : x \in X_G\} \quad (7a)$$

$$(x^*)^T Q^2 y^* = \max \{(x^*)^T Q^2 y : y \in Y_G\}. \quad (7b)$$

The maximization problems in Equation (7) are linear programs. Loosely speaking, the idea to formulate BLP formulation is to put both maximization problems into a single problem, where the sum of duality gaps is minimized. To be more specific, the Nash equilibrium can be attained from the solution (x^*, y^*) of the following BLP problem [2]:

$$\text{Minimize } b^T u + \mu + g^T v + v - x^T (Q^1 + Q^2) y \quad (8a)$$

$$\text{subject to } A^T u + e_n \mu - Q^1 y \geq 0 \quad (8b)$$

$$F^T v + e_q v - (Q^2)^T x \geq 0 \quad (8c)$$

$$x \in X_G, y \in Y_G \quad (8d)$$

$$u, v \geq 0. \quad (8e)$$

Observe that constraint (8d) check primal feasibilities of Equations (7a) and (7b), and constraints (8b) and (8c) together with Equation (8e) enforce dual feasibilities of those linear programs. Furthermore, the objective function represents the sum of respective duality gaps of Equations (7a) and (7b), which will become zero at an optimal solution (see Konno [2] for more rigorous derivation).

Next, a sequential game under perfect information is a zero-sum game played by two players; (i) *leader* P1 who makes a feasible move $x \in X = \{x \in R^n : Ax \leq b, x \geq 0\}$, and (ii) *follower* P2 who then makes a move after fully observing the leader's move and its

impact. Due to the sequential order of decision making, the move of P2 (denoted by $v \in R^p$) is subject to the move of P1. In particular, let $V(x) = \{v \in R^p : F^T v \leq d + Q^T x, v \geq 0\}$ be the feasible region of moves of P2 (i.e., the move of P1 alters the resource vector of P2 by $Q^T x$). Assume, for simplicity, that $V(x)$ is nonempty and bounded for each $x \in X$. Since this is a zero-sum game, the gain of P2 is equivalent to the loss of P1. This gain (loss) is determined by the pair of moves x and v , and assume that the gain of P2 is $c^T x + g^T v$. Hence, given $x \in X$, P2 solves

$$\text{Maximize } c^T x + g^T v \quad (9a)$$

$$\text{subject to } F^T v \leq d + Q^T x \quad (9b)$$

$$v \geq 0, \quad (9c)$$

where the optimal objective function value is the loss of P1. Since the leader wants to minimize this loss, P1 solves

$$\text{Minimize } \begin{bmatrix} \max & c^T x + g^T v \\ \text{s.t.} & F^T v \leq d + Q^T x \\ & v \geq 0 \end{bmatrix} \quad (10a)$$

$$\text{subject to } x \in X. \quad (10b)$$

This problem is exactly the same as that of min-max problem (4), which is equivalent to Equation (1). In consequence, the sequential game with perfect information can be formulated as BLP.

UNBOUNDED FEASIBLE REGIONS

Earlier, we assumed that both X and Y are nonempty and bounded so that extreme point optimality is assured. However, if one (or both) of X and Y is unbounded, BLP could have an unbounded optimal objective value. Audet *et al.* [14] present how one can check boundedness of BLP by solving bounded BLP(s). While assuming X and Y are still nonempty, we first assume that only one feasible region is unbounded. Without loss of generality, assume that Y is unbounded. Since X is bounded, BLP becomes unbounded if and only if the inner minimization in Equation (2) is unbounded. In turn, from the relationship between primal

and dual linear programs in conjunction with $Y \neq \emptyset$, the inner minimization problem in Equation (2) is unbounded if and only if the feasible region of its dual linear program is empty. Hence, the equivalent condition is $V(x) = \emptyset$ for some $x \in X$, where $V(x) = \{v \in R^p : F^T v \leq d + Q^T x, v \geq 0\}$. Given $x \in X$, consider the following homogeneous form of the inner minimization problem:

$$\text{Minimize } d^T y + x^T Q y \quad (11a)$$

$$\text{subject to } F y \geq 0 \quad (11b)$$

$$y \geq 0. \quad (11c)$$

This problem has a nonempty feasible region (i.e., $y = 0$ is feasible). Hence, this problem is unbounded if and only if its dual problem has no feasible solution, which in turn implies the unboundedness of BLP as discussed above. Adding a constraint $e_q^T y \leq 1$ to Equation (11), let $Y_0 = \{y \in R^q : F y \geq 0, e_q^T y \leq 1, y \geq 0\}$. Then, the above homogeneous problem is unbounded if and only if the following bounded problem has a negative optimal objective function value [17]:

$$\min_{y \in Y_0} d^T y + x^T Q y. \quad (12)$$

Therefore, BLP is unbounded if and only if the following bounded BLP problem has a negative optimal objective value

$$\min_{x \in X, y \in Y_0} d^T y + x^T Q y. \quad (13)$$

Symmetrically, when X is unbounded and Y is bounded, BLP is unbounded if and only if the following problem has a negative optimal objective value

$$\min_{x \in X_0, y \in Y} c^T x + x^T Q y, \quad (14)$$

where $X_0 = \{x \in R^n : A x \geq 0, e_n^T x \leq 1, x \geq 0\}$.

When both X and Y are unbounded, Audet *et al.* [14] show that BLP is unbounded if and only if at least one of the following three BLP problems with bounded feasible regions has

a negative optimal objective value:

$$\begin{aligned}
\text{(I)} \quad & \min_{x \in \text{conv}(X_E), y \in Y_0} d^T y + x^T Q y \\
\text{(II)} \quad & \min_{x \in X_0, y \in \text{conv}(Y_E)} d^T y + x^T Q y \\
\text{(III)} \quad & \min_{x \in X_0, y \in Y_0} x^T Q y,
\end{aligned}$$

where $\text{conv}(\cdot)$ denotes convex hull of the set. (Recall that X_E and Y_E are sets of extreme points.) The first two problems (I) and (II) are direct derivatives of Equations (13) and (14). Solving (I) and (II), we can verify whether the unboundedness of BLP occurs when one of x and y is finite. The third case (III) is for checking unboundedness of BLP under the circumstance where $f(x, y) \rightarrow -\infty$ only if $\|x\| \rightarrow \infty$ and $\|y\| \rightarrow \infty$. Note that (III) can be solved via any solution methods discussed earlier since X_0 and Y_0 are nonempty and bounded. Furthermore, one can solve (I) and (II) via solution methods that find a best extreme point solution (e.g., implicit enumeration of Falk and Audet *et al.* [7, 14]), after replacing $\text{conv}(X_E)$ and $\text{conv}(Y_E)$ by the original polyhedral sets X and Y , respectively.

GENERALIZED FORMULATION

When variables x and y appear in the same constraint, the problem becomes *jointly constrained bilinear programming problem* (JBLP), and can be written as follows [18]:

$$\text{JBLP: Minimize } c^T x + d^T y + x^T Q y \quad (15a)$$

$$\text{subject to } Ax + Fy \geq b \quad (15b)$$

$$x, y \geq 0, \quad (15c)$$

where F is now $F \in R^{m \times q}$. When variables are jointly constrained, the extreme point optimality does not hold any more [18]. Instead, an optimal solution can be found along the boundary of the feasible region. Al-Khayyal and Falk [18] consider a linearly transformed version of JBLP, where $n = q$, $Q = I_n$ (i.e., n -dimensional identity matrix), and x and y explicitly bounded by a hyperrectangle. They employ a branch-and-bound method, where branching is performed by

partitioning a bounding hyperrectangle into subhyperrectangles, and where lower and upper bounds are derived by the minimum of a piecewise-linear convex underestimating function and by the original objective function value at the point, respectively. This branch-and-bound scheme is further enhanced by Sherali and Alameddine [19], who consider the original form of JBLP and apply the *reformulation-linearization technique* (RLT) to generate lower-bounding linear programs. RLT is originally designed to produce tighter linear programming relaxations for mixed-integer programming problems [20]. In their implementation for solving JBLP, the problem is first reformulated to an equivalent form by generating quadratic valid constraints. Then, quadratic terms are linearized by defining new variables to yield a linear program, which is shown to provide a tighter lower bound.

The most generalized form of bilinear program (GBLP) in the literature allows bilinear terms to appear in the constraint set of JBLP as follows [21]:

$$\text{GBLP: Minimize } c^T x + d^T y + x^T Q y \quad (16a)$$

$$\text{subject to } a_i x + f_i y + x^T P_i y \geq b_i \quad (16b)$$

$$x, y \geq 0, \quad (16c)$$

where $a_i \in R^n$ and $f_i \in R^q$ denote rows of A and F , respectively, and $P_i \in R^{n \times q}$. Linderoth [22] extends the branch-and-bound method of Al-Khayyal and Falk [18] by partitioning the feasible region into Cartesian product of triangles and rectangles. The branch-and-bound in conjunction with RLT for solving JBLP [19] is also extended by Sherali and Tuncbilek [23] for polynomially constrained polynomial programming problems, which subsume GBLP.

REFERENCES

1. Hansen P, Jaumard B, Savard G. New branch and bound rules for linear bilevel programming. *SIAM J Sci Comput* 1992;13(5): 1194–1217.

2. Konno H. Bilinear programming: Part II. Application of bilinear programming. Technical report 17-10, Stanford University; August, 1971.
3. Ahmed S, Guan YP. The inverse optimal value problem. *Math Program* 2005;102(1):91–110.
4. Belenky AS. Estimating the size of the calling population in finite-source election queues by bilinear programming techniques. *Math Comput Model* 2007;45(7-8):873–882.
5. Lim C, Smith JC. Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Trans* 2007;39(1):15–26.
6. Konno H. Bilinear programming: Part I. Algorithm for solving bilinear programs. Technical report 17-9, Stanford University; August, 1971.
7. Falk JE. A linear max-min problem. *Math Program* 1973;5(1):169–188.
8. Konno H. A cutting plane algorithm for solving bilinear programs. *Math Program* 1976;11(1):14–27.
9. Thieu TV. A note on the solution of bilinear-programming problems by reduction to concave minimization. *Math Program* 1988;41(2):249–260.
10. Tuy H. Concave programming under linear constraints. *Sov Math* 1964;5:1437–1440.
11. Gallo G, Ülkücü A. Bilinear programming: an exact algorithm. *Math Program* 1977;12(1):173–194.
12. Sherali HD, Shetty CM. A finitely convergent algorithm for bilinear programming problems using polar cuts and disjunctive face cuts. *Math Program* 1980;19(1):14–31.
13. Ding X, Al-Khayyal F. Accelerating convergence of cutting plane algorithms for disjoint bilinear programming. *J Global Optim* 2007;38(3):421–436.
14. Audet C, Hansen P, Jaurnard B, Savard G. A symmetrical linear maxmin approach to disjoint bilinear programming. *Math Program* 1999;85(3):573–592.
15. Alarie S, Audet C, Jaurnard B, Savard G. Concavity cuts for disjoint bilinear programming. *Math Program* 2001;90(2):373–398.
16. Nash J. Non-cooperative games. *Ann Math* 1951;54(2):286–295.
17. Bazaraa MS, Jarvis JJ, Sherali HD. *Linear programming and network flows*. 3rd ed. New York, NY: Wiley-Interscience; 2005.
18. Al-Khayyal FA, Falk JA. Jointly constrained biconvex programming. *Math Oper Res* 1983;8(2):273–286.
19. Sherali HD, Alameddine A. A new reformulation-linearization technique for bilinear programming problems. *J Global Optim* 1992;2(4):379–410.
20. Sherali HD, Adams WP. A hierarchy of relaxations between the continuous and convex-hull representations for zero-one programming-problems. *SIAM J Discrete Math* 1990;3(3):411–430.
21. Al-Khayyal FA. Generalized bilinear programming: Part I. models, applications and linear programming relaxation. *Eur J Oper Res* 1992;60(3):306–314.
22. Linderoth J. A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. *Math Program* 2005;103(2):251–282.
23. Sherali HD, Tuncbilek CH. A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *J Global Optim* 1992;2(1):101–112.

BIOSURVEILLANCE: DETECTING, TRACKING, AND MITIGATING THE EFFECTS OF NATURAL DISEASE AND BIOTERRORISM

RONALD D. FRICKER, JR.
Operations Research Department,
Naval Postgraduate School,
Monterey, California

INTRODUCTION

Bioterrorism is not a new threat in the twenty-first century—at least a thousand years ago, the plague and other contagious diseases were used in warfare [1,2] but today the potential for catastrophic outcomes is greater than it has ever been. To address this threat, the medical and public health communities are putting various measures in place, including biosurveillance systems designed to proactively monitor populations for possible disease outbreaks. The goal is to improve the likelihood that a disease outbreak, whether man-made or natural, is detected as early as possible so that the medical and public health communities can respond as quickly as possible. An ideal biosurveillance system analyzes population health-related data in near-real time to identify subtle trends not visible to individual physicians and clinicians. As they sift through data, many of these systems use one or more statistical algorithms to look for anomalies to trigger detection, investigation, quantification, localization, and outbreak management.

What is Biosurveillance?

Homeland Security Presidential Directive 21 (HSPD-21) defines biosurveillance as “the process of active data-gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health—whether infectious, toxic, metabolic,

or otherwise, and regardless of intentional or natural origin—in order to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity” [3]. As shown in Fig. 1, “biosphere data” can be divided into information about human, animal, and agricultural populations, and biosurveillance thus consists of health surveillance on each of these populations.

One particular type of biosurveillance is *epidemiologic surveillance* which HSPD-21 defines as “the process of actively gathering and analyzing data related to human health and disease in a population in order to obtain early warning of human health events, rapid characterization of human disease events, and overall situational awareness of disease activity in the human population.” Thus, epidemiologic surveillance addresses that subset of biosurveillance as it applies to human populations.

As shown in Fig. 2, epidemiologic surveillance is but one element of public health surveillance. Public health surveillance encompasses the surveillance of adverse reactions to medical interventions (particularly drugs and vaccines) and how health services are used, as well as epidemiologic surveillance. *Syndromic surveillance* is a specific type of epidemiological surveillance that has been defined as “the ongoing, systematic collection, analysis, interpretation, and application of real-time (or near-real-time) indicators of diseases and outbreaks that allow for their detection before public health authorities would otherwise note them” [4]. Thus, syndromic surveillance is epidemiologic surveillance restricted to using leading indicators of disease. In particular, syndromic surveillance is based on the notion of a *syndrome* which is a set of nonspecific prediagnosis medical and other information that may indicate the release of a bioterrorism agent or natural disease outbreak. See, for example, syndrome definitions for diseases associated with critical bioterrorism-associated agents [5].

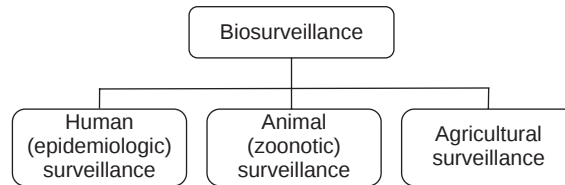


Figure 1. Biosurveillance consists of human, animal, and agricultural surveillance. Animal surveillance may consist of monitoring certain animal populations for unusual behavior or excessive mortality as a leading indicator of an outbreak as well as zoonotic surveillance which refers specifically to animal diseases that can pass to humans. Agricultural surveillance may include monitoring livestock and plant diseases important to the human food chain.

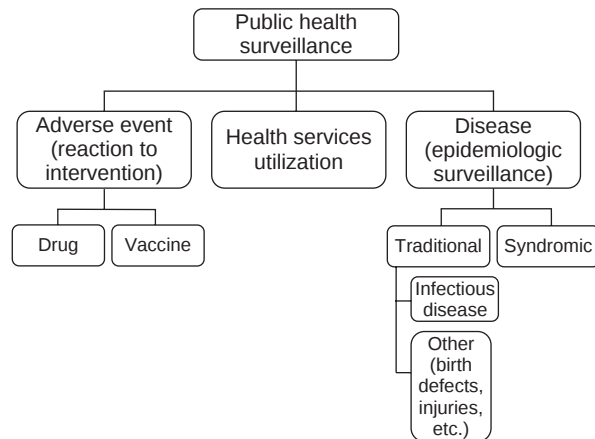


Figure 2. A taxonomy of public health showing that epidemiologic surveillance is but one part of a broader set of surveillance activities. Adapted from Rolka and O'Connor [6].

Syndromic surveillance differs from traditional epidemiologic surveillance in a number of important ways. For example, syndromic surveillance often uses nonspecific health and health-related data (e.g., daily number of individuals seeking health care in an emergency room with sore throats), whereas traditional notifiable disease reporting is based on confirmed cases (e.g., daily number of individuals with laboratory-confirmed diagnoses). In addition, while in conventional public health surveillance it is unusual to initiate active surveillance without a known or suspected outbreak, syndromic surveillance systems actively search for evidence of possible outbreaks well before there is any suspicion of an outbreak.

Biosurveillance Objectives

Syndromic surveillance has also been defined as "...surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response" [7,8]. This definition focuses on a number of ideas important to biosurveillance.

- First, biosurveillance is health surveillance, not military, regulatory or intelligence surveillance. It may use a wide variety of types of data, from case diagnoses to health-related data such as counts derived from chief complaints.
- Second, the data and associated surveillance are generally intended to precede

diagnosis or case confirmation in order to give early warning of a possible outbreak. Clearly, once a definitive diagnosis of a bioagent has been made, the need for detection becomes moot though tracking the location and spread of a potential outbreak is still important whether an outbreak has been confirmed or not.

- Third, the process must provide a signal of “sufficient probability” to trigger “further public health response.” Often the goal is not to provide a definitive determination that an outbreak is occurring but rather to signal that an outbreak *may* be occurring. Such a signal indicates further information is warranted through more detailed investigation by public health officials.

The motivation for biosurveillance, and syndromic surveillance systems in particular, is that some bioagents have symptoms in their prodromal stages similar to naturally occurring diseases. For example, in the first week or two after exposure to smallpox, individuals tend to have symptoms similar to the

flu such as fever, malaise, aches, nausea, and vomiting [9].

Biosurveillance systems have two main objectives: to support public health *situational awareness* (SA) and to enhance outbreak *early event detection* (EED). The CDC [10] defines them as follows:

- **SA** is the ability to utilize detailed, real-time health data to confirm, refute, and provide an effective response to the existence of an outbreak. It also is used to monitor an outbreak’s magnitude, geography, rate of change, and life cycle.
- **EED** is the ability to detect, at the earliest possible time, events that may signal a public health emergency. EED comprises case and suspect case reporting along with statistical analysis of health-related data. Both, real-time streaming of data from clinical care facilities as well as batched data with a short time delay are used to support EED efforts.

As illustrated in Fig. 3, biosurveillance systems are supposed to improve the chances of the medical and public health communities catching a disease outbreak early. The

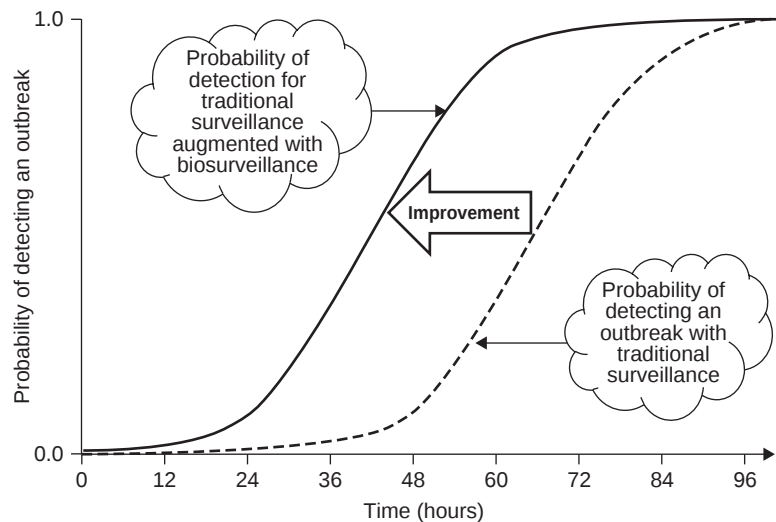


Figure 3. An illustration of how biosurveillance is intended to improve the probability of detecting a disease outbreak, whether man-made or natural. The vertical axis is in terms of probability specifically because early event detection is a stochastic phenomenon. That is, whether and when a biosurveillance system detects an outbreak is a function of both, the aspects of the specific situation and chance.

more biosurveillance improves the probability of detecting an outbreak, the more a biosurveillance system is likely to enhance SA and EED. The goal is "...deployment of surveillance systems that can rapidly detect and monitor the course of an outbreak and thus minimize associated morbidity and mortality" [11].

When assessing biosurveillance systems, speed of (true positive) detection is one of three dimensions critical for completely characterizing performance. The other two dimensions are the rate of false positives and the probability of successfully detecting an outbreak. In the biosurveillance literature, these dimensions are often generically referred to as timeliness, specificity, and sensitivity.

All three dimensions are necessary and they trade off. For example, for a given EED methodology, improving the speed of detecting an outbreak generally comes at the cost of increasing the rate of false positives. Similarly, increasing the probability of detection usually comes at the expense of the speed of detection. These trade-offs are similar to the Type I and Type II error trade-offs inherent in classical hypothesis testing, though the sequential decision-making aspect of biosurveillance adds an additional level of complexity.

See Fricker [12], particularly the rejoinder, for a more detailed discussion about evaluating biosurveillance system performance and about the appropriate metrics for quantifying biosurveillance system performance.

BIOSURVEILLANCE SYSTEMS

As HSPD-21 states, "A central element of biosurveillance must be an epidemiologic surveillance system to monitor human disease activity across populations. That system must be sufficiently enabled to identify specific disease incidence and prevalence in heterogeneous populations and environments and must possess sufficient flexibility to tailor analyses to new syndromes and emerging diseases."

The use of biosurveillance, particularly in the form of syndromic surveillance, is widespread. In 2003, it was estimated that

approximately 100 state and local health jurisdictions were conducting some form of syndromic surveillance [4]. In 2004, Bravata *et al.* [11] conducted a systematic review of the publicly available literature and various web sites from which they identified 115 surveillance systems, of which they found 29 that were designed specifically for detecting bioterrorism. In 2007–2008, Buehler *et al.* [13] sent surveys to public health officials in 59 state, territorial, and large local jurisdictions. Fifty-two officials responded (an 88% response rate) representing jurisdictions containing 94% of the US population. They found that 83% reported conducting syndromic surveillance for a median of three years and two-thirds said they are "highly" or "somewhat" likely to expand the use of syndromic surveillance in the next two years.

Components

As depicted in Fig. 4, a biosurveillance system has four main functions: data collection, data management, analysis, and reporting. As illustrated in the figure, raw data enters the system at the left and as it flows through the system it becomes actionable information at the right.

Expanding on Rolka [14], the ideal system contains the following components:

- the original data, to which access is gained only after appropriately addressing legal and regulatory requirements, as well as personal privacy and proprietary issues;
- computer hardware and information technology for (near) real-time assembly, recording, transfer, and preprocessing of data;
- subject matter experts, data management, and data knowledge experts, as well as software and techniques for processing incoming data into analytic databases, including processes and procedures for managing and maintaining these databases;
- statistical algorithms to analyze the data for possible outbreaks over space and time that are of sufficient sensitivity to provide signals within an

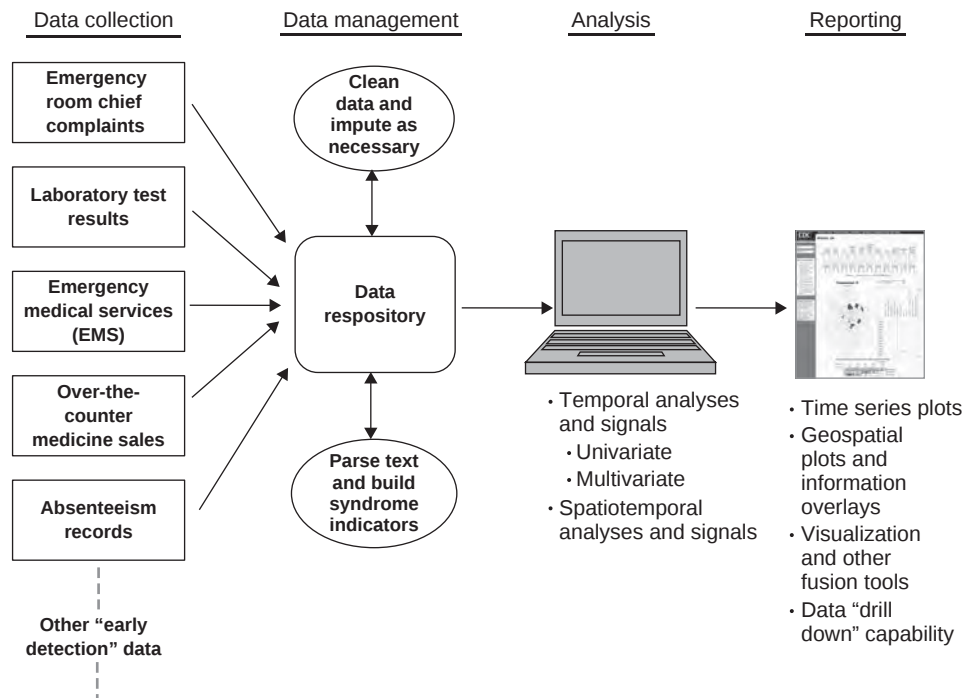


Figure 4. A biosurveillance system has four main functions: data collection, data management, analysis, and reporting. Raw data enters the system at the left and flows through the system to become actionable information at the right.

- actionable time frame while simultaneously limiting false positive signals to a tolerable level;
- public health experts with sufficient statistical expertise that can appropriately choose and apply the algorithms most relevant to their jurisdiction and appropriately interpret the signals when they occur;
- data display and query software, as well as the necessary underlying data, that facilitates rapid and easy investigation and adjudication of signals by public health experts, including the ability to “trace back” from signal to likely source;
- other data displays, combined with decision support and communication tools, to support SA during an outbreak and to facilitate effective and efficient public health response;
- report production processes and access to supporting supplementary information;

Mandl *et al.* [15], in *Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience*, provide a detailed discussion of what is required and guidance about how to implement biosurveillance systems.

Examples

Below are brief descriptions of three biosurveillance systems chosen to illustrate large-scale systems currently in operation. The first two are true systems, in the sense that they are comprised of both, dedicated computer hardware and software. The third is more properly described as a set of software programs that can be freely downloaded and implemented by any public health organization.

- BioSense.** Developed and operated by the Centers for Disease Control and Prevention (CDC), BioSense is intended to be a United States-wide

biosurveillance system. Begun in 2003, BioSense initially used the outpatient data of the Department of Defense and Department of Veterans Affairs, along with medical laboratory test results from a nationwide commercial laboratory. In 2006, BioSense began incorporating data from civilian hospitals as well. The primary objective of BioSense is to “expedite event recognition and response coordination among federal, state, and local public health and health-care organizations” [16,17].

- **ESSENCE.** An acronym for electronic surveillance system for the early notification of community-based epidemics, ESSENCE was developed by the Department of Defense in 1999. ESSENCE IV now monitors for infectious disease outbreaks at more than 300 military treatment facilities worldwide on a daily basis, using data from patient visits to the facilities and pharmacy data. For the Washington DC area, ESSENCE II monitors military and civilian outpatient visit data as well as over-the-counter (OTC) pharmacy sales and school absenteeism [18–20]. Components of ESSENCE have been adapted and used by some public health departments.
- **EARS.** An acronym for early aberration reporting system, EARS was and continues to be developed by the CDC. EARS was originally designed for monitoring for bioterrorism during large-scale events that often have little or no baseline data (i.e., as a short-term “drop-in” surveillance method) [21]. For example, the EARS system was used in the aftermath of Hurricane Katrina to monitor communicable diseases in Louisiana [22], for syndromic surveillance at the 2001 Super Bowl and World Series, as well as at the Democratic National Convention in 2000 [23]. Though developed as a stand-alone, portable surveillance analytic method, EARS data management procedures and algorithms have been adapted for use in many syndromic surveillance systems.

State and local biosurveillance systems may use EARS, BioSense, and in some cases ESSENCE, while some localities have instituted their own systems. Other biosurveillance and health surveillance systems include the real-time outbreak and disease surveillance system [24], the national notifiable diseases surveillance system [25], and the national electronic telecommunications system for surveillance [26]. Descriptions of some of these state and local systems (as well as other information) can be found in volume 53 of the CDC’s *Morbidity and Mortality Weekly Report* and the annotated bibliography for syndromic surveillance [27].

Among the more recent and unique surveillance efforts is Google Flu which is designed to track “health seeking” behavior in the form of search engine queries for flu-related information. As with syndromic surveillance systems using OTC medicine sales, the idea is that sick people first attempt to self-treat before seeking medical attention and, often, the first step is a search engine query for information. Figure 5 is a graph published in the *New York Times* comparing Google Flu’s estimated flu incidence based on search queries for flu-related terms to the CDC sentinel physician data. The figure shows a clear correspondence between the two time series. Ginsberg *et al.* [28, p. 1012] say, “Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day.”

Another unique surveillance effort is HealthMap, a “multistream real-time surveillance platform that continually aggregates reports on new and ongoing infectious disease outbreaks” [29]. HealthMap extracts information from web-accessible information sources such as discussion forums, mailing lists, government web sites, and news outlets. It then filters, categorizes, and integrates the information and, as shown in Fig. 6, plots the information on a map. The goal of HealthMap is to provide real-time information about emerging infectious

PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS *Mid Atlantic region*

Using Google to Monitor the Flu

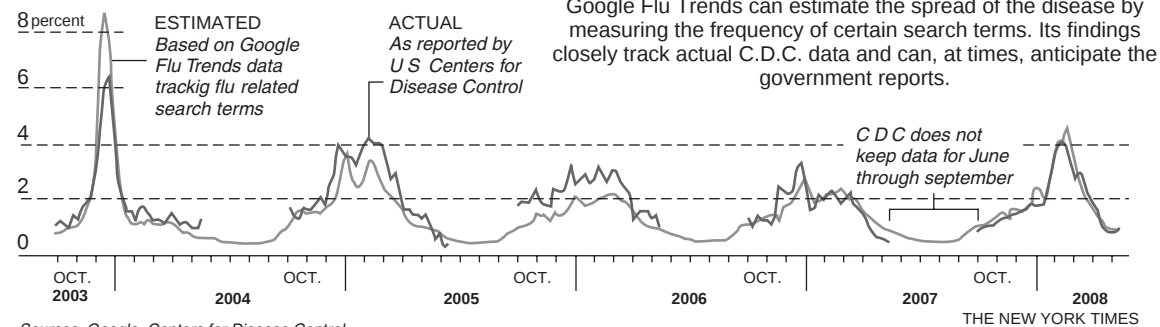


Figure 5. An excerpt from a *New York Times* article from November 12, 2008 about Google Flu.



Figure 6. Screen shot of HealthMap map from June 5, 2010 [30].

diseases and is intended for use by both, public health officials and the traveling public [30]. As Brownstein *et al.* [29, p. 1019] say, “Ultimately, the use of news media and other nontraditional sources of surveillance data can facilitate early outbreak detection, increase public awareness of disease outbreaks prior to their formal recognition, and provide an integrated and contextualized view of global health information.”

BIOSURVEILLANCE UTILITY AND EFFECTIVENESS

In spite of the widespread and potentially expanding use of biosurveillance, questions remain about its utility and effectiveness. There are several examples of papers that discuss various issues, challenges, and important research needs associated with

effective implementation and operation of biosurveillance systems [4,7,11–14,31–35].

In case studies of health departments in eight US states, Uscher-Pines *et al.* [32] found that fewer than half had written response protocols for responding to biosurveillance system alerts and the health departments reported conducting in-depth investigations on fewer than 15% of biosurveillance system alerts. Further, Uscher-Pines *et al.* [32] said, “Although many health departments noted that the original purpose of syndromic surveillance was early warning/detection, no health department reported using systems for this purpose. Examples of typical statements included the following: ‘I was a big supporter of syndromic surveillance for early warning early on, but now I am more realistic about the system’s limitations.’ ”

In the literature, Reingold [35] suggested that a compelling case for the implementation of biosurveillance systems has yet to be made. Cooper [36] said, “To date no bioterrorist attack has been detected in the United Kingdom, or elsewhere in the world using syndromic surveillance systems.” Stoto *et al.* [34] questioned whether biosurveillance systems can achieve an effective early detection capability. And Green [33] said, “Syndromic surveillance systems, based on statistical algorithms, will be of little value in early detection of bioterrorist outbreaks. Early on in the outbreak, there will be cases serious enough to alert physicians and be given definitive diagnoses.”

The research challenges span many disciplines and problems:

- legal and regulatory challenges in order to gain access to data;
- technological challenges related to designing and implementing computer hardware and software for collecting and assembling data;
- ethical and procedural issues inherent in managing and safeguarding data;
- analytical challenges of assessing the likelihood of outbreaks and of displaying data to enhance SA;
- managerial challenges of effectively assembling and operating the entire system.

These research challenges are not necessarily of equal importance nor are they listed in any type of priority order. Furthermore, little is known about how they should be prioritized in terms of their contributions to improving the utility and/or effectiveness of biosurveillance. However, it is clear that improvements are necessary in all of these disciplines to achieve biosurveillance systems that are maximally useful and effective.

Much of the continuing controversy surrounding biosurveillance stems from its initial focus on EED, a use that requires a number of unproven assumptions, including

adequate strength such that they are statistically detectable with satisfactory power;

- leading indicators occur sufficiently far in advance of clinical diagnoses so that, when found, they provide the public health community with enough advance notice to take action;
- statistical detection algorithms exist that produce signals reliable enough to warrant continued dedication of public health resources to investigate the signals.

Of course, a myopic focus only on EED in biosurveillance systems misses important benefits such systems can provide, particularly the potential to significantly advance and modernize the practice of public health surveillance. For example, whether or not biosurveillance systems prove effective at the early detection of bioterrorism, they are likely to have a significant and continuing role in the detection and tracking of seasonal and pandemic flu, as well as other naturally occurring disease outbreaks. This latter function is echoed in an Institute of Medicine report on Microbial Threats to Health by Smolinski *et al.* [37]: “[S]yndromic surveillance is likely to be increasingly helpful in the detection and monitoring of epidemics, as well as the evaluation of health-care utilization for infectious diseases.” In a similar vein, Uscher-Pines *et al.* [32] quote a public health official: “Health departments should not be at the mercy of alerts; they need to develop their own uses for syndromic surveillance.”

In terms of bioterrorism, Stoto [38] states that biosurveillance systems build links between public health and health-care providers—links that could prove to be critical for consequence management should a bioterrorism attack occur. Furthermore, Sosin [4] points out that biosurveillance systems can act as a safety net should existing methods of detection fail to detect an attack. As such, biosurveillance can provide additional lead time to public health authorities so they can take more effective public health actions. For example, a Dutch biologist conducting automated salmonella surveillance related that the surveillance

- leading indicators of outbreaks exist in prediagnosis health-related data of

system detected an outbreak whose occurrence was somehow missed by sentinel physicians [39]. And unusual indicators in a biosurveillance system (not necessarily a signal from an EED algorithm) may give public health organizations time to begin organizing and marshaling resources in advance of a confirmed case and/or provide critical information about how and where to apply resources.

QUANTITATIVE METHODS USED IN BIOSURVEILLANCE SYSTEMS

Biosurveillance systems use a variety of temporal and spatial methods for EED. As is discussed in more detail below, most of these have been adapted from other fields and applications. Buckeridge *et al.* [40] provides a useful classification of common surveillance scenarios and a mapping of some EED methods to those scenarios.

SA is an emerging concept in biosurveillance and, as such, specific methods for assessing and displaying relevant information have not yet been incorporated into the systems. Currently SA is limited to displaying the spatial distribution of data via geographic information system (GIS) software. See, for example, Li *et al.* [41].

Temporal Methods

Most syndromic surveillance systems apply variants of the standard univariate statistical process control (SPC) methods: Shewhart, cumulative sum (CUSUM), and/or exponentially weighted moving average (EWMA) charts. Woodall [42] provides a comprehensive overview of the application of control charts to health surveillance. Montgomery [43] is an introduction to these methods in an SPC setting and Fricker [44] is a primer on how to apply these and other statistical methods to biosurveillance, both for EED and SA. Fricker [45], Shmueli and Fienberg [46], and Shmueli and Burkom [31] also review these and other methods potentially applicable to EED in a biosurveillance setting.

The challenge in applying these methods is that syndromic surveillance generally violates classical SPC assumptions. In SPC it is often reasonable to assume that

- since one controls the manufacturing process, the in-control distribution is (or can reasonably be assumed to be) stationary;
- observations can be drawn from the process so they are independent (or nearly so);
- monitoring the process mean and standard deviation is usually sufficient;
- the asymptotic distributions of the statistics being monitored are known and thus can be used to design appropriate control charts;
- shifts, when they occur, remain until they are detected and corrective action taken;
- temporal (as opposed to spatial) detection is the critical problem.

However, the general biosurveillance problem violates many, if not all, of these assumptions. For example,

- there is little or no control over disease incidence and thus the distribution of disease incidence is usually nonstationary;
- observations (often daily counts) are autocorrelated, and the need for quick detection works against the idea of taking measurements far enough apart to achieve (near) independence;
- in biosurveillance there is little information on what types of statistics are useful for monitoring—one is often looking for anything that seems unusual;
- because individual observations are being monitored, the idea of asymptotic sampling distributions does not apply, and the data often contain significant systematic effects that must be accounted for;
- outbreaks are transient, with disease incidence returning to its original state once an outbreak has run its course;

- identifying both spatial and temporal deviations is often critical.

This gap between existing SPC methods and the biosurveillance problem provides an area that is ripe for new research.

In spite of the gap, the standard SPC methods are sometimes applied to biosurveillance data with little modification [47,48]) and in some cases, the methods are modified to attempt to account for the autocorrelation. For example, EARS applies variants of the Shewhart control chart (Eq. 1) and 2 below) and a cumulative method (Eq. 3). These methods use various moving windows of data to estimate the process mean and standard deviation [23] and they are intended to be used when little historic (“baseline”) information is available.

The EARS methods are called “C1,” “C2,” and “C3” and are defined as follows [49]: Let $Y(t)$ be the observed count for period t representing, for example, the number of individuals arriving at a particular hospital emergency room with a specific syndrome on day t . The C1 calculates the statistic $C_1(t)$ as

$$C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{S_1(t)}, \quad (1)$$

where $\bar{Y}_1(t)$ and $S_1(t)$ are the moving sample mean and standard deviation, respectively:

$$\begin{aligned} \bar{Y}_1(t) &= \frac{1}{7} \sum_{j=t-1}^{t-7} Y(j) \\ \text{and } S_1^2(t) &= \frac{1}{6} \sum_{j=t-1}^{t-7} [Y(j) - \bar{Y}_1(j)]^2. \end{aligned}$$

If $S_1^2(t) = 0$, then EARS sets it to a small positive number. As implemented in the EARS system, the C1 signals on day t when the C_1 statistic exceeds a threshold h , which is fixed at three sample standard deviations above the sample mean: $C_1(t) > 3$.

The C2 is similar to the C1, but incorporates a two-day lag in the mean and standard deviation calculations. Specifically, it calculates

$$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)}, \quad (2)$$

where

$$\begin{aligned} \bar{Y}_3(t) &= \frac{1}{7} \sum_{j=t-3}^{t-9} Y(j) \\ \text{and } S_3^2(t) &= \frac{1}{6} \sum_{j=t-3}^{t-9} [Y(j) - \bar{Y}_3(j)]^2. \end{aligned}$$

If $S_3^2(t) = 0$, then EARS sets it to a small positive number and the C2 method signals on day t when $C_2(t) > 3$.

The C3 combines current and historical data from day t and the previous two days, calculating the statistic $C_3(t)$ as

The C3 method described above, as implemented in EARS V4.5, is different from the original EARS implementation. See Fricker [45] for a description of the original C3 method.

$$C_3(t) = \sum_{i=t}^{t-2} \max[0, C_2(i) - 1]. \quad (3)$$

It signals when $C_3(t) > 2$. See Fricker [45] for additional description.

BioSense originally implemented the C1, C2, and C3 methods, but has since modified the C2. Calling the new method “W2,” it calculates the mean and standard deviation separately for weekdays and weekends (using the relevant last seven days with a two-day lag). Following the suggestion of Buckeridge *et al.* [40], it uses an empiric method to define the threshold and users of the system can select the threshold as an option [50].

To date, multivariate SPC methods have not been incorporated into operational biosurveillance systems. Some research has been conducted to define and evaluate directional multivariate methods, including Joner *et al.* [51], Fricker [47], and Stoto *et al.* [48]. Whether or not multivariate methods provide additional sensitivity to detect outbreaks as compared to the current practice of using multiple simultaneous univariate methods has yet to be conclusively demonstrated. Various methods have also been proposed to combine and/or adjust for the application of multiple univariate methods in multivariate settings. See the discussion in Rolka *et al.* [52] about parallel

and consensus monitoring methods and Stoto *et al.* [34,48] for a discussion and an evaluation of some of these methods.

Regression and time series methods have also been proposed to explicitly model, and hence account for, seasonality. The basic idea is to model the disease incidence process, perhaps including terms in the model for annual seasonal variations, monthly variations, and even day of the week and holiday variations. The model is then used to either (i) predict the expected number of events and the difference between the expected number and observed number is monitored for excessive deviations or, (ii) model and then remove the explainable/known effects (sometimes called *preconditioning*) and then the residuals are monitored using traditional SPC methods. This is consistent with recommendations by Montgomery [43] for autocorrelated data. Examples in the literature include Brillman *et al.* [53], who apply the CUSUM to the prediction errors, the CDC's cyclical regression models discussed in Hutwagner *et al.* [54], log-linear regression models in Farrington *et al.* [55], and time series models in Reis and Mandl [56]. See Shmueli and Burkom [31] for additional discussion of the use of regression and time series methods for syndromic surveillance and Burkom *et al.* [57] for a comparison of two regression-based methods and an exponential smoothing method applied to biosurveillance forecasting. Also see Lotze *et al.* [58] for a detailed discussion of preconditioning applied to syndromic surveillance data.

Other temporal methods that have been proposed or are in use for syndromic surveillance include wavelets [59–61, and the discussion in Shmueli and Burkom 31]; Bayesian networks [52,62]; hidden Markov models [63]; Bayesian dynamic models [64], and rule-based methods [65].

Spatial and Spatiotemporal Methods

Kleinman *et al.* [66] and Lazarus *et al.* [67] proposed a generalized linear mixed model (GLMM) to simultaneously monitor disease counts over time in a region divided into smaller subareas (zip codes). It is statistically attractive because it uses information

across the entire region while appropriately adjusting for the smaller areas.

As described in Kleinman *et al.* [66], there are two forms of the model depending on whether individual data and covariates are available versus aggregated counts and covariates by zip code. In the former case, the model is

$$E(y_{ijt} | b_i) = p_{ijt} \text{ and } \text{logit}(p_{ijt}) = \mathbf{x}_{ijt}\beta + b_i, \quad (4)$$

where y_{ijt} is an indicator for whether or not person j in area i is a case on day t , p_{ijt} is the probability he or she is a case, $\mathbf{x}_{ijt}$ is a vector of observed covariates on person j and/or area i over time up to and including day t , β is a vector of fixed effects, and b_i is a random effect for area i . When no individual level covariate information is available, the most likely situation, the model is

$$E(y_{it}|b_i) = p_{it} \text{ and } \text{logit}(p_{it}) = \mathbf{x}_{it}\beta + b_i, \quad (5)$$

where $y_{it} = \sum_{j=1}^{n_{it}} y_{ijt}$. In this model, p_{it} can be thought of as the probability that an individual in area i will be a case on day t .

Having fitted the model in Equation (5), z cases are observed on day $t + 1$. The rarity of the observed count is assessed by calculating

$$\Pr(Z \geq z \text{ cases}) = 1 - \sum_{k=1}^{z-1} \binom{n_{it}}{k} \hat{p}_{it}^k (1 - \hat{p}_{it})^{n_{it}-k}, \quad (6)$$

where $\hat{p}_{it}$ is calculated from the estimated coefficients in the usual way for logistic regression and $1/(\hat{p}_{it} \times \# \text{ tests conducted})$ is proposed as the recurrence interval: the number of time periods for which the expected number of counts of z or more cases is one [68]. Waller [69] recommends an alternate calculation for the recurrence interval and Woodall *et al.* [68] take issue with both the use of and recommended calculations for the recurrence interval.

The small area regression and testing (“SMART”) method in BioSense (see the BioSense User Guide [70]) is based on the Kleinman *et al.* [66] and Lazarus *et al.* [67] GLMM approach. However, as implemented

in BioSense, it only uses spatial information to bin data into separate time series, the output of which are subsequently combined using a Bonferroni correction. Hence, the BioSense SMART method is properly classified as a temporal method.

The most commonly used spatial method is the scan statistic, particularly as implemented in the SaTScan software (www.satscan.org). Originally developed to retrospectively identify disease clusters [71], the method is now regularly used prospectively in biosurveillance systems [72]. For example, it was used as part of a drop-in syndromic surveillance system in New York City after the 9/11 attack [73]. While it has been studied by the BioSense program, it has not yet been implemented in the BioSense system interface [74] and the BioSense User Guide [70].

The basic idea in SaTScan is to count the number of cases that occur in a cylinder, where the circle is the geographic base and the height of the cylinder corresponds to time. The cylinder is passed over space, varying the radius of the circle (up to a maximum radius that includes 50% of the monitored population) and the height of the cylinder and counts of cases for those geographic regions whose centroids fall within the circle for the period of time specified by the height of the cylinder are summed. When used for prospective biosurveillance, the start date of the height of the cylinder is varied but the end date is fixed at the most current time period. Conditioning on the expected spatial distribution of observations, SaTScan reports the most likely cluster (in both space and time) and its p -value.

Though widely used, some aspects of the prospective application of the SaTScan methodology have been questioned, particularly the use of recurrence intervals and performance comparisons between SaTScan and other methods. See Woodall *et al.* [68] for further details. Also, see Kulldorff [72] for other methods for disease mapping and for testing whether an observed pattern of disease is due to chance.

Fricker and Chang [75] introduced a new spatiotemporal methodology for biosurveillance called the repeated two-sample rank

(RTR) procedure. It is designed to sequentially incorporate information from individual observations and thus can operate on data in real-time as it arrives into an automated biosurveillance system. In addition, upon a signal of a possible outbreak, the methodology suggests a way to graphically indicate the likely outbreak location, and the output can subsequently be used to track the spread of the outbreak. Thus, the methodology can be used for both, EED and SA in automated biosurveillance systems.

Olson *et al.* [76] and Forsberg *et al.* [77] take an alternate approach to assessing possible disease clusters using M -statistics based on the distribution of pairwise distances between cases. That is, let $\mathbf{X} = \{X_1, \dots, X_n\}$ represent the locations of n cases on the plane and let $\mathbf{d} = \{d_1, \dots, d_{\binom{n}{2}}\}$ be the $\binom{n}{2}$ interpoint distances, then the M -statistic is

$$M = (\mathbf{o} - \mathbf{e})^T \hat{S}^{-1} (\mathbf{o} - \mathbf{e}), \quad (7)$$

where $\mathbf{o}$ and $\mathbf{e}$ are vectors of observed and expected counts of binned interpoint distances and $\hat{S}^{-1}$ is the inverse of the estimate of the covariance matrix. However, Forsberg *et al.* [78] state, "...these methods still require much refinement and further research."

Other approaches to spatial and spatiotemporal biosurveillance methods include the automated epidemiologic geotemporal integrated surveillance system (AEGIS) by Olson *et al.* [76] and the application of CUSUM methods to the spatial distribution of cases [79]. See Lawson and Kleinman [80] for additional exposition and methods, and Mandl *et al.* [15] for further discussion on spatial and spatiotemporal modeling issues. For spatial methods with application to more traditional public health data and problems, see Waller and Gottway [81].

CONCLUSION

Biosurveillance systems are being developed and implemented around the world. They are motivated by a need for improved public health surveillance, not only for bioterrorism,

but also to improve detection and responsiveness to natural disease outbreaks such as H1N1, avian influenza, and SARS, and as such they hold great promise as public health tools. For additional development and discussion, see Fricker [44], Lombardo and Buckeridge [82], M'ikanatha *et al.* [83], Wager *et al.* [84], Waller and Gottway [81], and Stroup *et al.* [85].

REFERENCES

- Gottfried RS. The Black Death: natural and human disaster in medieval Europe. New York (NY): Free Press; 1985.
- Deaux G. The Black Death, 1347. London, England: David McKay Company; 1969.
- U.S. Government. Homeland security presidential directive 21: public health and medical preparedness. 2007. Available at www.fas.org/irp/offdocs/nspd/hspd-21.htm. Accessed 2009 Sept 29.
- Sosin DM. Syndromic surveillance: the case for skillful investment view. *Biosecure Bioterror* 2003;1:247–253.
- CDC. Syndrome Definitions for Diseases Associated with Critical Bioterrorism-associated Agents dated 2003 Oct 23. 2003. Available at www.bt.cdc.gov/surveillance/syndromedef/. Accessed 2006 Nov 21.
- Rolka H, O'Connor Jean. Real-time public health biosurveillance: systems and policy considerations. *Infectious disease informatics and biosurveillance: research, systems and case studies*. New York (NY): Springer; 2010.
- Fricker RD Jr, Rolka HR. Protecting against biological terrorism: statistical issues in electronic biosurveillance. *Chance* 2006;19:4–13.
- CDC. 2006a. Available www.cdc.gov/biosense/publichealth.htm. Accessed 2006 Nov 16.
- Zubay G, editor. Agents of bioterrorism: pathogens and their weaponization. New York (NY): Columbia University Press; 2005.
- CDC. 2008. Available at www.cdc.gov/BioSense/publichealth.htm. Accessed 2008 Oct 11.
- Bravata DM, McDonald KM, Smithe WM, *et al.* Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Ann Intern Med* 2004;140(11):910–922.
- Fricker RD Jr. Some methodological issues in biosurveillance (with discussion and rejoinder). *Stat Med*. In press.
- Buehler JW, Sonricker A, Paladini M, *et al.* Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments. *Adv Dis Surveill* 2008;6(3):1–20.
- Rolka HA. Data analysis research issues and emerging public health biosurveillance directions. In: Wilson A, Wilson G, Olwell DH, editors. *Statistical methods in counterterrorism: game theory, modeling, syndromic surveillance, and biometric authentication*. New York (NY): Springer; 2006. pp. 101–107.
- Mandl KD, Overhage JM, Wagner MW, *et al.* Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Informatics Assoc* 2004;11:141–150.
- Tokars J. The BioSense application. Presentation at the 2006 PHIN Conference. Atlanta (GA); 2006. Available at [http://0-www.cdc.gov.mill1.sjlibrary.org/biosense/files/Jerry_Tokars.ppt#387,1,The BioSense Application](http://0-www.cdc.gov.mill1.sjlibrary.org/biosense/files/Jerry_Tokars.ppt#387,1,The%20BioSense%20Application). Accessed 2006 Nov 27.
- CDC. BioSense. 2006c. Available at www.cdc.gov/biosense/. Accessed 2006 Nov 27.
- DoD. 2006. Available at www.geis.fhp.osd.mil/GEIS/SurveillanceActivities/ESSENCE/ESSENCE.asp. Accessed 2006 Nov 27.
- Lombardo JS, Burkom H, Pavlin J. ESSENCE II and the framework for evaluating syndromic surveillance systems. *MMWR Morb Mortal Wkly Rep* 2004;53(suppl):159–165.
- OSD. ESSENCE IV improves Nation's bio-surveillance capability. 2005. Available at http://deploymentlink.osd.mil/news/jan05/news_20050125_001.shtml. Accessed 2006 Nov 27.
- CDC. Early aberration reporting system. 2007. Available at www.bt.cdc.gov/surveillance/ears. Accessed 2007 April 30.
- Toprani A, Ratard R, Straif-Bourgeois S, *et al.* Surveillance in hurricane evacuation centers—Louisiana. *MMWR Morb Mortal Wkly Rep* 2006;55:32–35.
- Hutwagner L, Thompson W, Seeman GM, *et al.* The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *J Urban Health Bull N Y Acad Med* 2003a;80(2 Suppl 1):89i–96i.
- RODS. RODS Laboratory website. 2010. Available at <https://www.rods.pitt.edu/site/>. Accessed 2010 June 5.

25. CDC. NNDSS website. 2010b. Available at www.cdc.gov/ncphi/diss/nndss/nndsshis.htm. Accessed 2010 June 5.
26. CDC. NETSS website. 2010a. Available at www.cdc.gov/ncphi/diss/nndss/netss.htm. Accessed 2010 June 5.
27. CDC. Annotated bibliography for syndromic surveillance. 2006b. Available at www.cdc.gov/EPO/dphi/syndromic/evaluation.htm. Accessed 2006 Nov 28.
28. Ginsberg J, Mohebbi MH, Patel RS, *et al.* Detecting influenza epidemics using search engine query data. *Nature* 2009;457:1012–1014.
29. Brownstein JS, Freifeld CC, Reis BY, *et al.* Surveillance Sans Frontières: internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med* 2008;5(7):1019–1024.
30. Freifeld C, Brownstein J. HealthMap website. 2010. Available at <http://healthmap.org/en/>. Accessed 2010 June 5.
31. Shmueli G, Burkom HS. Statistical challenges facing early outbreak detection in biosurveillance. 2009. Available at www.rhsmith.umd.edu/faculty/gshmueli/web/images/statchallengesbiosurveillancevised-iii.pdf. Accessed 2009 Sep 27.
32. Uscher-Pines L, Farrell CL, Babin SM, *et al.* Framework for the development of response protocols for public health syndromic surveillance systems: case studies of 8 US States. *Disaster Med Public Health Preparedness* 2009;3:S29–S36.
33. Green M. Syndromic surveillance for detecting bioterrorist events - the right answer to the wrong question? Presentation given at the Naval Postgraduate School; 2008 June 9. Monterey (CA); 2008.
34. Stoto MA, Schonlau M, Mariano LT. Syndromic surveillance: is it worth the effort? *Chance* 2004;17:19–24.
35. Reingold A. If syndromic surveillance is the answer, what is the question? *Biosecure Bioterror* 2003;1:1–5.
36. Cooper DL. Can syndromic surveillance data detect local outbreaks of communicable disease? A model using a historical cryptosporidiosis outbreak. *Epidemiol Infect* 2006;134:13–20.
37. Smolinski MS, Hamburg MA, Lederberg J, editors. *Microbial threats to health: emergence, detection, and response*. Washington (DC): National Academies Press; 2003.
38. Stoto MA. Syndromic surveillance in public health practice. Presentation to Institute of Medicine Forum on Microbial Threats; 2006 Dec 12.
39. Burkom H. Personal communication 2006 Dec 22.
40. Buckeridge DL, Burkom H, Campbell M, *et al.* Algorithms for rapid outbreak detection: a research synthesis. *J Biomed Inform* 2005;99–113.
41. Li H, Faruque F, Williams W, *et al.* Real-time syndromic surveillance. *ArcUser: the Magazine for ESRI Software Users*. 2006. January-March issue. pp. 17–19.
42. Woodall WH. The use of control charts in health-care and public-health surveillance. *J Qual Technol* 2006;38:1–16.
43. Montgomery DC. *Introduction to statistical quality control*. 5th ed. New York (NY): John Wiley & Sons, Inc.; 2004.
44. Fricker RD Jr. *Introduction to statistical methods for biosurveillance*. Cambridge University Press; 2010a. In press.
45. Fricker RD Jr. Syndromic surveillance. In: Melnick E, Everitt B, editors. *Encyclopedia of quantitative risk analysis and assessment*. John Wiley & Sons, Ltd; 2008. pp. 1743–1752.
46. Shmueli G, Fienberg SE. Current and potential statistical methods for monitoring multiple data streams for biosurveillance. In: Wilson A, Wilson G, Olwell DH, editors. *Statistical methods in counterterrorism: game theory, modeling, syndromic surveillance, and biometric authentication*. New York (NY): Springer; 2006. pp. 109–140.
47. Fricker RD Jr. Directionally sensitive multivariate statistical process control methods with application to syndromic surveillance. *Adv Dis Surveill* 2007;3(1):1–17. Available at www.isdsjournal.org.
48. Stoto MA, Fricker RD Jr, Jain A, *et al.* Evaluating statistical methods for syndromic surveillance. In: Wilson A, Wilson G, Olwell DH, editors. *Statistical methods in counterterrorism: game theory, modeling, syndromic surveillance, and biometric authentication*. New York (NY): Springer; 2006. pp. 141–172.
49. Hutwagner LC, Browne T, Seaman GM, *et al.* Comparing aberration detection methods with simulated data. *Emerg Infect Dis* 2005;11:314–316.
50. CDC. BioSense bulletin. 2006d Sep issue.
51. Joner MD Jr, Woodall WH, Reynolds MR Jr, *et al.* A one-sided MEWMA chart for

- health surveillance. *Qual Reliab Eng Int* 2008;24:503–519.
52. Rolka H, Burkom H, Cooper GF, *et al.* Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs. *Stat Med* 2007;26:1834–1856.
 53. Brillman JC, Burr T, Forslund D, *et al.* Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC Med Inform Decis Mak* 2005;5:4–18.
 54. Hutwagner L, Thompson W, Seeman GM, *et al.* The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *J Urban Health Bull N Y Acad Med* 2003b;80:89i–96i.
 55. Farrington CP, Andrews NJ, Beale AD, *et al.* A statistical algorithm for the early detection of outbreaks of infectious disease. *J R Stat Soc Ser A Stat Soc* 1996;159:547–563.
 56. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. *BMC Med Informatics Decis Mak* 2003;3.
 57. Burkom HS, Murphy SP, Shmueli G. Automated time series forecasting for biosurveillance. *Stat Med* 2006;4202–4218.
 58. Lotze T, Murphy SP, Shmueli G. Implementation and comparison of preprocessing methods for biosurveillance. *Adv Dis Surveill* 2008;6:1–14.
 59. Goldenberg A, Shmueli G, Caruana RA, *et al.* Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proc Natl Acad Sci U S A* 2002;99:5237–5240.
 60. Zhang J, Tsui F, Wagner M, *et al.* Detection of outbreaks from time series data using wavelet transform. *AMIA Annual Symposium Proceedings*; 2003. pp. 748–752.
 61. Shmueli G. Wavelet-based monitoring for modern biosurveillance. Technical Report RHS-06-002. College Park (MD): University of Maryland, Robert H. Smith School of Business; 2005.
 62. Wong W, Cooper G, Dash D, *et al.* Use of multiple data streams to conduct bayesian biologic surveillance. *MMWR Morb Mortal Wkly Rep* 2005;54(suppl):63–69.
 63. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden markov models. *Stat Med* 1999;18:3463–3478.
 64. Sebastiani P, Mandl KD, Szolovits P, *et al.* A bayesian dynamic model for influenza (with discussion). *Stat Med* 2006;25:1803–1825.
 65. Wong W, Moore A, Cooper G, *et al.* WSARE: What's strange about recent events? *J Urban Health Bull N Y Acad Med* 2003;80(suppl):66i–75i.
 66. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am J Epidemiol* 2004;159:217–224.
 67. Lazarus R, Kleinman K, Dashevsky I, *et al.* Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg Infect Dis* 2002;8:753–760.
 68. Woodall WH, Marshall B, Joner MD, *et al.* On the use and evaluation of scan methods for health-related surveillance. *J R Stat Soc Ser A Stat Soc* 2008;171:223–237.
 69. Waller LA. Invited commentary: syndromic surveillance-some statistical comments. *Am J Epidemiol* 2004;159:225–227.
 70. CDC. BioSense User Guide, Version 2.0. 2006e. Available at http://0-www.cdc.gov.mill1.sjlibrary.org/biosense/files/CDC_BioSense_User_Guide_v2.0.pdf. Accessed 2006 Nov 28.
 71. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods* 1997;26:1481–1496.
 72. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *J R Stat Soc Ser A Stat Soc* 2001;164:61–72.
 73. Ackelsberg J, Balter S, Bornschelgel K, *et al.* Syndromic surveillance for bioterrorism following the attacks on the world trade center - New York City, 2001. *MMWR Morb Mortal Wkly Rep* 2002;51(Special Issue):13–15.
 74. Bradley C. Visualizing and Monitoring Data in BioSense Using SaTScan. Syndromic Surveillance Conference presentation. 2005. Available at www.cdc.gov/biosense/files/SaTScan_Presentation_2005.ppt#501,1, Colleen Bradley. MSPH Syndromic Surveillance Conference. Accessed 2006 Nov 28.
 75. Fricker RD Jr, Chang JT. A spatio-temporal methodology for real-time biosurveillance. *Qual Eng* 2008;20:465–477.
 76. Olson KL, Bonetti M, Pagano M, *et al.* Real time spatial cluster detection using interpoint distances among precise patient locations. *BMC Med Inform Decis Mak* 2005;5.

77. Forsberg L, Jeffery C, Ozonoff A, *et al.* A spatiotemporal analysis of syndromic data for biosurveillance. In: Wilson A, Wilson G, Olwell DH, editors. Statistical methods in counterterrorism: game theory, modeling, syndromic surveillance, and biometric authentication. New York (NY): Springer; 2006. pp. 173–191.
78. Forsberg L, Bonetti M, Jeffery C, *et al.* Distance-based methods for spatial and spatio-temporal surveillance. In: Lawson AB, Kleinman K, editors. Spatial & syndromic surveillance for public health. The Atrium, Southern Gate, Chichester: John Wiley & Sons; 2005. pp. 133–152.
79. Rogerson PA, Yamada I. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Stat Med* 2004;23:2195–2214.
80. Lawson AB, Kleinman K, editors. Spatial & syndromic surveillance for public health. The Atrium, Southern Gate, Chichester: John Wiley & Sons; 2005.
81. Waller LA, Gottway CA, editors. Applied spatial statistics for public health data. Hoboken (NJ): John Wiley & Sons; 2004.
82. Lombardo JS, Buckeridge DL, editors. Disease surveillance: a public health informatics approach. Hoboken (NJ): John Wiley & Sons; 2007.
83. M'ikanatha NM, Lynfield R, Beneden CAVan, *et al.*, editors. Infectious disease surveillance. 1st ed. Malden (MA): Blackwell Publishing; 2007.
84. Wagner MM, Moore AW, Aryel RM, editors. Handbook of biosurveillance. New York (NY): Elsevier Academic Press; 2006.
85. Stroup DF, Williamson GD, Herndon JL, *et al.* Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Stat Med* 1989;8:323–329.

BIRTH-AND-DEATH PROCESSES

GUVENÇ DEĞIRMENCI
Department of Industrial
Engineering, University
of Pittsburgh,
Pittsburgh,
Pennsylvania

INTRODUCTION

A birth-and-death (BD) process is a continuous-time Markov chain (CTMC), $\{X(t) : t \geq 0\}$, defined on the countable state space $S = \{0, 1, 2, \dots\}$ and whose transitions are restricted to only its nearest neighbors. In other words, if at any time the process is in state $i \in S \setminus \{0\}$, then in an infinitesimal time interval, it will next transition to either state $i + 1$ or to state $i - 1$. If the current state is 0, then it can only transition to state 1. BD processes have been used to model population dynamics, queueing systems, inventory systems, computer and communications networks, and biological systems, to name only a few. For example, a BD process can be used to describe the temporal and stochastic evolution of the population of a geographical region when the birth and death rates of the region can be estimated from historical data. In such models, the random variable $X(t)$ represents the population of the region at time t . In a queueing context, the BD process can be used to describe the total number of customers in a stochastic service system (i.e., the number of customers currently in service and those waiting to receive service) when customers arrive according to a Poisson process and bring an exponentially distributed service requirement.

The BD process can be analyzed in a transient sense (for any finite t such that $t \geq 0$), or in an asymptotic sense (as $t \rightarrow \infty$). The asymptotic behavior is often used to highlight the salient features of the system's dynamics. Moreover, asymptotic results are often

mathematically tractable, easier to obtain, and serve as reasonable approximations to the transient behavior when t is large. The ergodic theorem for CTMCs gives the conditions under which the limiting distribution of a BD process coincides with the long-run fraction of time that the process occupies any state in S . For instance, in a high-speed communications network, it is important to know the fraction of time each node in the network is in a congested state in order to characterize important delay and congestion measures. When the joint process of buffer content at all the nodes can be modeled as a multivariate BD process, the distribution can be obtained in a simple, closed form. The limiting behavior can also be used to design new systems that satisfy quality-of-service guarantees.

In the remainder of this article, we formally define the standard BD process, review some basic facts about its transient and asymptotic behavior, and discuss some important special cases of the BD process. In addition, we highlight two applications in queueing theory, discuss some extensions, and provide guidance for further reading.

FORMAL MODEL DESCRIPTION

A BD process, $\{X(t) : t \geq 0\}$, is a CTMC on the countable state space $S = \{0, 1, 2, \dots\}$ whose distinguishing characteristic is that it transitions only to its nearest neighbors. Specifically, from any state $i \in S \setminus \{0\}$, the process may transition to state $i - 1$, or to state $i + 1$; however, it may not transition to any state $i \pm k$ where $k \geq 2$. If $i = 0$, it may only transition to $i + 1$. Intuitively, one may think of a BD process as representing the evolution of a population process, and the state, $X(t)$, is the population size at time t . An increase in the population size (by unit increment) represents a "birth" and a decrease (by unit decrement) represents a "death." If the birth and/or death rates depend on the current population, then the BD process is called *state dependent*. In other words, whenever $X(t) = i$,

the birth rate is a positive number λ_i , and the death rate is μ_i , $i \in S$.

For any $s, t \geq 0$ and $i, j \in S$, the transition probability functions of a BD process are given by

$$p_{ij}(s, t) \equiv \mathbb{P}(X(t + s) = j | X(s) = i),$$

and the BD process is called *time-homogeneous* if $p_{ij}(s, t) = p_{ij}(0, t) =: p_{ij}(t)$ for any $s \geq 0$. For $i > 0$, and a small interval of length t , the transition functions of the time-homogeneous version satisfy the following equations:

$$p_{ij}(t) = \begin{cases} \lambda_i t + o(t), & j = i + 1, \\ \mu_i t + o(t), & j = i - 1, \\ 1 - (\lambda_i + \mu_i)t + o(t), & j = i, \\ o(t), & \text{otherwise,} \end{cases}$$

where $o(t)/t \rightarrow 0$ as $t \rightarrow 0$. These probabilities are interpreted as follows. Given that the current state is i ($i > 0$), in an infinitesimal time interval, the next event is either a birth with probability $\lambda_i t + o(t)$ or a death with probability $\mu_i t + o(t)$, and the population size remains the same with probability $1 - (\lambda_i + \mu_i)t + o(t)$. The probability that a “jump” magnitude exceeds unity is of order $o(t)$. Practically speaking, in a small interval of time, multiple births, multiple deaths, and simultaneous births and deaths do not occur. Whenever $i = 0$, the process next transitions to state 1 after an exponentially distributed amount of time with mean $1/\lambda_0$. It is not hard to see that the *infinitesimal transition rates* of the process are given by the birth rates $\{\lambda_i\}$ and the death rate $\{\mu_i\}$. The transition rate diagram of this CTMC is depicted in Fig. 1.

The infinitesimal generator matrix of the BD process, denoted by $\mathbf{Q}$, is given by

$$\mathbf{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 \dots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 \dots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 \dots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \lambda_3 \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}, \quad (1)$$

which is seen to possess a tridiagonal structure. This structure provides some computational advantages, particularly when evaluating the asymptotic behavior of the process (as $t \rightarrow \infty$). The transition functions, $p_{ij}(t)$, can be shown to satisfy the Kolmogorov backward equations, a set of ordinary differential equations given by

$$\frac{dp_{ij}(t)}{dt} = \begin{cases} \lambda_0 [p_{1j}(t) - p_{ij}(t)], & i = 0, \\ \lambda_i p_{i+1,j}(t) + \mu_i p_{i-1,j}(t) - (\lambda_i + \mu_i) p_{ij}(t), & i > 0. \end{cases}$$

Suppose we let $\mathbf{P}(t) = [p_{ij}(t)]$, $i, j \in S$, denote the matrix of transition functions for any $t \geq 0$. Obtaining $\mathbf{P}(t)$ for a generic CTMC with finite state space S is not too difficult. The Kolmogorov backward equations can be written in a convenient matrix form as follows:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q} \mathbf{P}(t).$$

This (matrix) differential equation has the obvious solution

$$\mathbf{P}(t) = \mathbf{P}(0) \exp(\mathbf{Q}t), \quad t \geq 0,$$

where $\exp(\mathbf{A})$ denotes matrix exponentiation of the matrix $\mathbf{A}$, and $\mathbf{P}(0) = I$, the identity matrix. Hence, $\mathbf{P}(t)$ can be obtained by employing any number of numerical routines for performing matrix exponentiation.

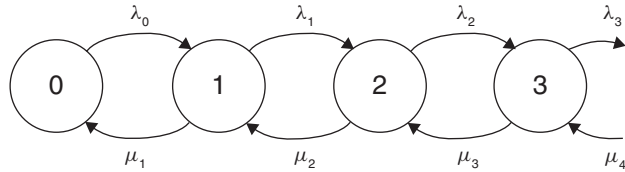


Figure 1. Transition rate diagram of a standard birth-and-death process.

However, because the BD process is defined on $S = \{0, 1, 2, \dots\}$, solving for the transition functions is generally challenging except in a few special cases. For instance, if $\{X(t) : t \geq 0\}$ is a homogeneous Poisson process with $\lambda_i = \lambda$ and $\mu_i = 0$ for all $i \in S$, then the Kolmogorov backward equations can be solved exactly using standard Laplace transform techniques (cf. Kulkarni [1]). For a general BD process (with state-dependent births and deaths), the method of continued fractions has been successfully applied to analyze the transient behavior. Three representative examples of the use of continued fractions in analyzing BD processes are given in Parthasarathy [2–4]. Uniformization methods can also be used, particularly if the state space can be truncated at a finite level [1,5].

Some Special Cases

It is easy to see that some well-known stochastic processes are special cases of the standard BD process. Specifically, if $\mu_i = 0$ for all $i \in S$, then $\{X(t) : t \geq 0\}$ is called a *pure birth process*. In case $\mu_i = 0$ and $\lambda_i = \lambda$ for all $i \in S$, the BD process is a homogeneous Poisson process (see **Poisson Process and its Generalizations**). A prototypical example of the pure birth process is cell reproduction where $X(t)$ is the number of cells living in a colony at time t . Each cell in the colony acts independent of the others and splits into two cells after an exponentially distributed amount of time with parameter λ . Therefore, the birth rates are $\lambda_i = i\lambda$, $i \in S$, since each of the i members gives birth with the same exponential rate λ . In this case, $\{X(t) : t \geq 0\}$ is called a *Yule process*.

A *pure death process* is a BD process with $\lambda_i = 0$ for all $i \in S$ (i.e., only deaths occur until the process is absorbed at state 0). For example, consider a machine that operates initially with i ($i > 0$) components in parallel such that each component has an exponentially distributed lifetime with mean $\mu^{-1} < \infty$, and suppose $X(t)$ denotes the number of functioning components at time t . The component lifetimes are assumed to be mutually independent, and the system can continue to function as long as at least one component is functioning. At the time of a

component failure, the system state is decremented by unity, and the failed component cannot be repaired. Here, the death rates are given by $\mu_i = i\mu$, $i \in S$, since each component fails at the same exponential rate μ .

LIMITING BEHAVIOR

In this section, we discuss the limiting behavior of the transition probabilities $p_{ij}(0, t)$ as $t \rightarrow \infty$. Initially, let us assume the existence of a set of values, $\{p_j : j \in S\}$, such that

$$\begin{aligned} \lim_{t \rightarrow \infty} p_{ij}(0, t) &= \lim_{t \rightarrow \infty} \mathbb{P}(X(t) = j | X(0) = i) \\ &= p_j, \quad j \in S, \end{aligned} \quad (2)$$

and note that p_j is independent of the initial state i . If the limit in Equation (2) exists, then the row vector $\mathbf{p} \equiv (p_0, p_1, p_2, \dots)$ is called the *limiting distribution* of $\{X(t) : t \geq 0\}$. The existence of the limit is ensured when $\{X(t) : t \geq 0\}$ is irreducible (i.e., all states in S communicate) and positive recurrent (i.e., starting from any state, the expected time to return to that state is finite). The following result from the asymptotic analysis of positive recurrent CTMCs shows how to obtain the limiting distribution $\mathbf{p}$.

Theorem 1. *If $\{X(t) : t \geq 0\}$ is an irreducible and positive recurrent CTMC, then its limiting distribution $\mathbf{p}$ exists and is the unique positive solution to the system of equations*

$$\mathbf{p}\mathbf{Q} = \mathbf{0}; \quad \sum_{j \in S} p_j = 1, \quad (3)$$

where $\mathbf{0}$ denotes the zero vector.

Whenever these probabilities exist, the BD process is said to be *ergodic*, and the limit in Equation (2) has two important interpretations. First, by definition, p_j is the limiting probability that the process $\{X(t) : t \geq 0\}$ is in state j . Second, p_j can be interpreted as the long-run proportion of time that the process spends in state j . For the BD process, the matrix $\mathbf{Q}$ takes the form of Equation (1).

Solving the system of equations (3) for the BD process, it is not difficult to see that

$$p_k = p_0 \left[\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right], \quad k \geq 1. \quad (4)$$

Applying the normalization condition,

$$\sum_{k=0}^{\infty} p_k = 1,$$

the probability p_0 is obtained by

$$p_0 = \left[1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \right]^{-1}. \quad (5)$$

Equation (4) shows that each of the probabilities depends explicitly on the limiting probability p_0 that the process is in state 0. Hence, the limiting distribution exists if and only if $p_0 > 0$, that is, if

$$\sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{k-1}}{\mu_1 \mu_2 \cdots \mu_k} < \infty.$$

Hence, the ergodicity of the process depends explicitly on the convergence of the above infinite series. Should the series diverge to $(+\infty)$, then $p_0 = 0$ and a limiting distribution does not exist.

APPLICATIONS IN QUEUEING THEORY

BD processes are commonly used to model Markovian queueing systems. Here, we discuss the application of BD processes in two of the most basic queueing systems.

The $M/M/1$ Queue

In the $M/M/1$ queueing system, customers arrive according to a Poisson process with rate λ , and each customer brings a service requirement that is exponentially distributed with parameter μ . The system has only a single server for processing customers, and it has an infinite waiting room for customers to wait for service (see **The $M/M/1$ Queue** for further details). Let $X(t)$ denote the number of customers in the system at time t (i.e., in

service or in the waiting area) and note that $\{X(t) : t \geq 0\}$ is a BD process with constant birth and death rates $\lambda_j = \lambda$ for $j \geq 0$ and $\mu_j = \mu$ for $j \geq 1$, respectively. The ratio of the arrival rate to the service rate is the traffic intensity given by $\rho = \lambda/\mu$. Then using Equations (4) and (5), we obtain

$$p_0 = \left[1 + \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu} \right)^k \right]^{-1}.$$

The infinite series converges if and only if $\rho < 1$; therefore, the system is stable if and only if $\lambda < \mu$. In such a case, the limiting probabilities are given by

$$p_k = p_0 \left(\frac{\lambda}{\mu} \right)^k = (1 - \rho) \rho^k, \quad k \geq 1.$$

The $M/M/s$ Queue

The $M/M/s$ is similar to the $M/M/1$ queue except that it has s ($s \geq 1$) independent and identical servers (see **The $M/M/s$ Queue**). Since there are s servers, the death rates of this BD process are given by

$$\mu_j = \begin{cases} j\mu, & \text{if } j < s, \\ s\mu, & \text{if } j \geq s. \end{cases}$$

The traffic intensity for this system is $\rho = \lambda/s\mu$. Using Equations (4) and (5), the limiting probabilities are given by

$$p_k = \begin{cases} p_0 \frac{(s\rho)^k}{k!}, & 0 \leq k \leq s, \\ p_0 \frac{\rho^k s^s}{s!}, & k \geq s. \end{cases}$$

and

$$p_0 = \left[\sum_{k=0}^{s-1} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!} \frac{1}{1 - \rho} \right]^{-1}.$$

Therefore, the limiting distribution of the number of customers in the system exists if $\rho < 1$, that is, if $\lambda < s\mu$. This means that the maximum service rate must exceed the arrival rate of customers for the system to remain stable.

EXTENSIONS AND FURTHER READING

An important extension of the BD process is the quasi-birth-and-death (QBD) process [6–8]. The QBD is a bivariate Markov process on the state space $S = \{(i, j) : i \geq 0, 1 \leq j \leq M_i\}$, ($0 < M_i < \infty$), where i is the *level* of the process and j is referred to as the *phase*. Its block tridiagonal infinitesimal generator matrix is given by

$$\mathbf{Q} = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & 0 & 0 & 0 & \dots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & 0 & \dots \\ 0 & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & 0 & \dots \\ 0 & 0 & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The positive integer M_i is the number of phases at the i th level. The QBD process is called *level independent* if, for $i \geq 0$, $M_i = M$, $A_0^{(i)} = A_0$ and for $i \geq 1$, $A_1^{(i)} = A_1$ and $A_2^{(i)} = A_2$ (see **Level-Independent Quasi-Birth-and-Death Processes**). Otherwise, the QBD process is called *level dependent* (see **Level-Dependent Quasi-Birth-and-Death Processes**). The analysis of QBDs is facilitated by the matrix-geometric method, developed primarily by Neuts [6]. By employing matrix-geometric methods, many researchers have addressed the transient and asymptotic analysis of (level-independent) QBD processes, as well as their many and varied applications. A method for computing the stationary distribution in the level-dependent case is given in Bright and Taylor [9].

The multivariate generalization of a BD process is called a *Markov population process* [10]. Let S^m be the set of all m -dimensional vectors $\mathbf{v} = (v_1, v_2, \dots, v_m)$ whose components $v_k \in S$, $1 \leq k \leq m$ are nonnegative integers. Intuitively, one may think of $\mathbf{v}$ as the population vector of a system consisting of m colonies with v_k as the population of the k th colony. Let u and ω be indicator variables such that $u = 1$ ($\omega = 1$), if there is an arrival to (departure from) the k th colony. For any $s, t \geq 0$ and $v_k \in S$, $1 \leq k \leq m$, the transition probability functions of a Markov population

process are given by

$$\mathbb{P}(\mathbf{X}(s+t) = \mathbf{v} + u\mathbf{e}_k - \omega\mathbf{e}_l | \mathbf{X}(s) = \mathbf{v}) = \begin{cases} \lambda_k(\mathbf{v})t + o(t), & u=1, \omega=0, \\ \mu_l(\mathbf{v})t + o(t), & u=0, \omega=1, \\ \gamma_k(\mathbf{v})t + o(t), & u=\omega=1, \\ 1 - [\sum_k \lambda_k(\mathbf{v}) + \sum_l \mu_l(\mathbf{v}) \\ + \sum_k \sum_l \gamma_{kl}(\mathbf{v})]t + o(t), & u=\omega=0, \\ o(t), & \text{otherwise,} \end{cases}$$

where $o(t)/t \rightarrow 0$ as $t \rightarrow 0$ and $\mathbf{e}_k$ is the k th unit vector. The transition from state $\mathbf{v}$ to $\mathbf{v} + \mathbf{e}_k$ may be described as an arrival to the k th colony, the transition from $\mathbf{v}$ to $\mathbf{v} - \mathbf{e}_k$ as a departure from k , and the transition from $\mathbf{v}$ to $\mathbf{v} + \mathbf{e}_k - \mathbf{e}_l$ as a transfer from l to k , $l \neq k$ (assuming $m \geq 2$). For certain classes of Markov population processes, Kingman [10] obtained the equilibrium distributions as $t \rightarrow \infty$. The reader is referred to McNeil and Schach [11], Barbour [12], and Klebaner [13] for further details on this subject.

One of the earliest descriptions of the BD process, its practical implications, and the limiting distribution were provided by Feller [14]. Classical textbooks that provide further details include Ross [15], Kulkarni [1], and Asmussen [16]. Properties of the BD process and its transient analysis are described in Karlin and McGregor [17–19] and more recently by Jouini and Dallery [20]. The BD process has been extended to include more complex dynamics. For instance, Kendall [21] developed the non-Markovian birth process. Age-dependent BD processes are analyzed by Waugh [22] and Weiner [23]. Griffiths [24,25] analyzed bivariate and multivariate BD processes and discussed transient properties. The BD process with immigration and emigration is studied by Zheng *et al.* [26], Swift [27], and Aksland [28]. Karlin and Tavar [29], and van Doorn and Zeifman [30] analyzed the BD process with killing.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman and Hall; 1995.
2. Parthasarathy PR. Some unusual birth-and-death processes. Math Sci 2003;28:79–90.

3. Parthasarathy PR, Lenin RB. Birth-and-death process (BDP) models with applications—queueing, communication systems, chemical models, biological models: the state-of-the-art with a time-dependent perspective. New York: American Scienc Press; 2004.
4. Parthasarathy PR. Exact transient solution of a state-dependent birth-death process. *J Appl Math Stoch Anal* 2006;2006:1–16.
5. Heyman DP, Sobel MJ. Volume I, Stochastic models in operations research. New York: McGraw Hill; 1982.
6. Neuts MF. Matrix-geometric solutions in stochastic models. Baltimore (MD): Johns Hopkins University Press; 1981.
7. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability. Philadelphia (PA): SIAM; 1999.
8. Zhang J, Coyle EJ. Transient analysis of quasi-birth-death processes. *Stoch Models* 1989;5:459–496.
9. Bright LW, Taylor PG. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Commun Stat—Stoch Models* 1995;11:497–525.
10. Kingman JFC. Markov population processes. *J Appl Probab* 1969;6:1–18.
11. McNeil DR, Schach S. Central limit analogues for Markov population processes. *J R Stat Soc B* 1973;35:1–23.
12. Barbour AD. Equilibrium distributions Markov population processes. *Adv Appl Probab* 1980;12:591–614.
13. Klebaner FC. Asymptotic behaviour of Markov population processes with asymptotically linear rate of change. *J Appl Probab* 1994;31:614–625.
14. Feller W. An introduction to probability theory and its applications. New York: John Wiley & Sons, Inc.; 1950.
15. Ross S. Stochastic processes. New York: John Wiley & Sons, Inc.; 1996.
16. Asmussen S. Applied probability and queues. New York: Springer; 2003.
17. Karlin S, McGregor J. The classification of birth and death processes. *Trans Am Math Soc* 1957;86:366–3400.
18. Karlin S, McGregor J. The differential equations of birth-and-death processes and the Stieltjes moment problem. *Trans Am Math Soc* 1957;85:489–546.
19. Karlin S, McGregor J. Linear growth, birth and death processes. *J Math Mech* 1958;7:643–662.
20. Jouini O, Dallery Y. Moments of first passage times in general birth–death processes. *Math Methods Oper Res* 2008;68:49–76.
21. Kendall DG. On the generalized birth-and-death process. *Ann Math Stat* 1948;19:1–15.
22. O’N Waugh WA. An age-dependent birth and death process. *Biometrika* 1955;42:291–306.
23. Weiner HJ. Applications of the age distribution in age dependent branching processes. *J Appl Probab* 1966;3:179–201.
24. Griffiths DA. A bivariate birth-death process which approximates to the spread of a disease involving a vector. *J Appl Probab* 1972;9:65–75.
25. Griffiths DA. Multivariate birth-and-death processes as approximations to epidemic processes. *J Appl Probab* 1973;10:15–26.
26. Zheng Y, Chao X, Ji X. Transient analysis of linear birth-death processes with immigration and emigration. *Probab Eng Inf Sci Arch* 2004;18:141–159.
27. Swift RJ. Transient probabilities for a simple birth-death-immigration process under the influence of total catastrophes. *Int J Math Math Sci* 2001;25:689–692.
28. Aksland M. A birth, death and migration process with immigration. *Adv Appl Probab* 1975;7:44–60.
29. Karlin S, Tavaré S. Linear birth and death processes with killing. *J Appl Probab* 1982;19:477–487.
30. van Doorn EA, Zeifman AI. Extinction probability in a birth-death process with killing. *J Appl Probab* 2005;42:185–198.

BLOCK REPLACEMENT POLICIES

FÉLIX BELZUNCE
Departamento Estadística e
Investigación Operativa,
Universidad de Murcia,
Murcia, Spain

MOSHE SHAKED
Department of Mathematics,
University of Arizona,
Tucson, Arizona

Block replacement policies arise in the context of reliability theory, where the interest is focused on the random time to fail of a unit or a system [1]. In this context, preventive maintenance of units or systems are used to reduce the number of failures for fixed interval times or to increase the time to failure in some stochastic sense. Block replacement policies are among the most commonly used planned replacement policies. Under a block replacement policy, a unit is replaced upon failure and additionally at times $T, 2T, 3T, \dots$. Therefore replacements are scheduled in advance and they make up a simple preventive maintenance program. Several issues have been considered for block replacement policies. In this article, we consider results for the comparison of block replacement policies with alternative preventive maintenance policies. More precisely, we consider conditions under which the number of unplanned failures (or the time at which unplanned failures occur) are reduced (increased) under a block replacement policy compared with some other preventive maintenance policies. Further extensions of the usual block replacement policy and cost analysis are also discussed in this article.

In this article, “increasing” and “decreasing” stand for “nondecreasing” and “nonincreasing,” respectively.

UNIVARIATE COMPARISONS WITH OTHER REPLACEMENT POLICIES

Throughout this article, unless stated otherwise, we will consider a unit or a

system which is subject to failures and we will denote by X , the random time to failure. Under a block replacement policy, the unit is replaced by a new one upon failure and at times $T, 2T, 3T, \dots$. Let us denote by $N_T^B \equiv \{N_T^B(t), t \geq 0\}$, the counting process that counts the number of failures or unplanned replacements, that is, $N_T^B(t)$ denote the number of failures in the interval $[0, t]$. The arrival times at which the n th failure occurs will be denoted by $S_{n,T}^B$. Given that the block replacement policy is introduced to reduce the number of unplanned failures, it is natural to compare this policy with the usual replacement policy in which a unit is replaced by a new one only upon failure and no other replacements are considered. This leads to a renewal process which will be denoted by $N \equiv \{N(t), t \geq 0\}$, where $N(t)$ denotes the number of failures in the interval $[0, t]$ and the arrival times at which the n th failure occurs will be denoted by S_n . The renewal function for this process will be denoted by $M(t) \equiv E[N(t)]$ (see *Definition and Examples of Renewal Processes* and *Renewal Function and Renewal-Type Equations*).

The natural framework where the comparisons can be stated is the context of stochastic orders and aging notions. These provide several tools for the purpose of stochastically comparing random quantities and describe the aging process of the units or systems that are involved in the maintenance.

First, we recall several notions of stochastic orders that will be used in the sequel; for general references on the subject the reader can look at Refs 2 and 3 as well as *Aging, Characterization, and Stochastic Ordering*.

Let X and Y be two absolutely continuous nonnegative random variables. The random variable X is said to be smaller than the random variable Y in the ordinary stochastic order (denoted as $X \leq_{\text{st}} Y$) if $P(X > x) \leq P(Y > x)$ for all x . Thus, the comparison in the ordinary stochastic order indicates that in some stochastic sense, the random variable Y tends to take on larger values than the

random variable X . Additionally, the stochastic order can be characterized as follows: $X \leq_{\text{st}} Y$ if, and only if, $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing functions ϕ for which previous expectations are defined. Therefore, the stochastic order not only compares random lifetimes in terms of survival probabilities but also in terms of expected values of increasing transformations such as expected utilities, benefits, or costs, associated with the random lifetimes.

In case the transformation ϕ is increasing and convex, this leads to a weaker comparison. The random variable X is said to be smaller than the random variable Y in the increasing convex order (denoted as $X \leq_{\text{icx}} Y$) if $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing convex functions ϕ for which the expectations are defined.

Another possibility is to compare the Laplace transforms. The random variable X is said to be smaller than the random variable Y in the Laplace transform order (denoted as $X \leq_{\text{Lt}} Y$), if $E[\exp\{-sX\}] \geq E[\exp\{-sY\}]$ for all $s > 0$.

The following implications hold among the above notions:

$$\begin{array}{ccc} X \leq_{\text{st}} Y & \Rightarrow & X \leq_{\text{icx}} Y \\ \Downarrow & & \Downarrow \\ X \leq_{\text{Lt}} Y & \Rightarrow & E[X] \leq E[Y]. \end{array}$$

Next, we recall some aging notions. Let X be a nonnegative, absolutely continuous random variable. Denote by F and f , its distribution and density functions, respectively.

The random variable X is said to have the aging property of increasing failure rate (IFR) if $1 - F$ is logconcave. It has the property of decreasing failure rate (DFR) if $1 - F$ is logconvex on its support. Denoting the hazard rate function of X by $r \equiv f/(1 - F)$, it holds that X is IFR [or DFR] if, and only if r is increasing [or decreasing].

The random variable X is said to be new better than used (NBU) [or, new worse than used (NWU)] if, and only if, $X \geq_{\text{st}} [X - t | X > t]$ for all $t > 0$. Another characterization of this aging class can be given in terms of $-\log(1 - F)$, that is, X is NBU [NWU], if and only if, $-\log(1 - F)$ is a superadditive [subadditive] function, where we say that a

function ϕ is superadditive [subadditive], if $\phi(x + y) \geq [\leq] \phi(x) + \phi(y)$ for all $x, y \geq 0$.

The random variable X is said to be new better than used in the convex order (NBUC), if $X \geq_{\text{icx}} [X - t | X > t]$ for all $t > 0$.

The random variable X is said to be new better than used in the Laplace transform order (NBUL) if $X \geq_{\text{Lt}} [X - t | X > t]$ for all $t > 0$.

Finally, the random variable X with a finite mean is said to be new better than used in expectation (NBUE) if $E[X] \geq E[X - t | X > t]$ for all $t \geq 0$.

Among these aging notions, we have the following relationships:

$$\begin{array}{ccccc} \text{IFR} & \Rightarrow & \text{NBU} & \Rightarrow & \text{NBUC} \\ & & \Downarrow & & \Downarrow \\ & & \text{NBUL} & \Rightarrow & \text{NBUE}. \end{array}$$

The first result in the literature that stochastically compares maintenance policies was provided by Barlow and Proschan [4]; it states that under the IFR assumption we have $N_T^B(t) \leq_{\text{st}} N(t)$ for all $t \geq 0$. The result was improved by Marshall and Proschan [5], replacing the IFR assumption by the weakest NBU assumption and states the following.

Theorem 1. *The comparison $N_T^B(t) \leq_{\text{st}} [\geq_{\text{st}}] N(t)$ holds for all $t \geq 0$ and $T > 0$, if, and only if, X is NBU [NWU].*

This result tells us that NBU [NWU] distributions make up the largest class for which any block replacement policy reduces [increases], in the sense of the ordinary stochastic order, the number of failures in any interval $[0, t]$ compared with a usual replacement policy. Marshall and Proschan [5] also proved the following result—it shows that NBU [NWU] distributions also make up the largest class for which an increase in the frequency of block replacements stochastically reduces [increases] the number of unplanned replacements in any interval $[0, t]$.

Theorem 2. *The comparison $N_{kT}^B(t) \leq_{\text{st}} [\geq_{\text{st}}] N_{kT}^B(t)$ holds for all $t \geq 0$, $T > 0$, and $k = 1, 2, \dots$, if, and only if, X is NBU [NWU].*

From Theorem 1 and the definition of the ordinary stochastic order, it is clear that if X is NBU [NWU] then $E[N_T^B(t)] \leq [\geq] M(t)$ for all $t \geq 0$ and $T > 0$. Shaked and Zhu [6], in the next theorem, provided a characterization of the comparison of expected number of failures in terms of the superadditivity of the renewal function. That is, superadditivity [subadditivity] of M is a necessary and sufficient condition for the practically useful comparisons $E[N_T^B(t)] \leq [\geq] M(t)$ and $E[N_T^B(t)] \leq [\geq] E[N_{kT}^B(t)]$.

Theorem 3. *The following statements are equivalent:*

- (i) $M(t)$ is superadditive [subadditive].
- (ii) $E[N_T^B(t)] \leq [\geq] M(t)$ for all $t \geq 0$ and $T > 0$.
- (iii) $E[N_T^B(t)] \leq [\geq] E[N_{kT}^B(t)]$ for all $t \geq 0$, $T > 0$, and $k = 1, 2, \dots$

Shaked and Zhu [6] also stated the following result, which relates the monotonicity in T of the expected number of failures with the concavity (convexity) of the renewal function. Note that convexity [concavity] is a stronger condition than superadditivity [subadditivity], and under this condition on M , the practically useful fact that $E[N_T^B(t)]$ is increasing [decreasing] in T is obtained.

Theorem 4. *The following statements are equivalent:*

- (i) $M(t)$ is convex [concave].
- (ii) $E[N_T^B(t)]$ is increasing [decreasing] in $T > 0$ for each fixed $t \geq 0$.

If we consider the monotonicity in T , in the stochastic order of $N_T^B(t)$, then Ref. 6 provided the following interesting result, which indicates that the practically useful comparison $N_{T_1}^B(t) \leq_{\text{st}} [\geq_{\text{st}}] N_{T_2}^B(t)$ holds only for some IFR [DFR] lifetimes.

Theorem 5. *If $N_{T_1}^B(t) \leq_{\text{st}} [\geq_{\text{st}}] N_{T_2}^B(t)$ for all $0 < T_1 \leq T_2$, for each fixed $t \geq 0$, then X is IFR [DFR].*

The reversed implication is not true for IFR distributions (counterexample given in Ref. 6), but for DFR distributions it is still unknown if the reversed implication holds. In this case, Brown [7] proved that if X is DFR then $M(t)$ is concave, and therefore from Theorem 4, we have the following practically useful result for DFR lifetimes.

Theorem 6. *If X is DFR then $E[N_T^B(t)]$ is decreasing in $T > 0$ for each fixed $t \geq 0$.*

From Theorem 1, we can also have comparisons of the times in which unplanned replacements occur. Given any counting process $C = \{C(t), t \geq 0\}$, with arrival times $C_n = \inf\{t : C(t) > n\}$, then $\{C(t) < n\} \Leftrightarrow \{C_n > t\}$. From this observation we have under the NBU [NWU] assumption, that $S_n \leq_{\text{st}} [\geq_{\text{st}}] S_{n,T}^B$ for all $n = 1, 2, \dots$

Along these lines, several authors have provided additional results under weaker aging properties.

For instance, the following result follows from results in Yue and Cao [8], and it indicates a useful weak condition under which S_n and $S_{n,T}^B$ can be compared.

Theorem 7. *If X is NBUL then $S_n \leq_{\text{Lt}} S_{n,T}^B$ for all $T > 0$ and $n = 1, 2, \dots$*

If we replace the NBUL assumption in Theorem 7 by the NBUC, we have the following result that was provided by Belzunce *et al.* [9]. This result indicates another useful weak condition under which S_n and $S_{n,T}^B$ can be compared.

Theorem 8. *If X is NBUC then $S_n \leq_{\text{icx}} S_{n,T}^B$ for all $T > 0$ and $n = 1, 2, \dots$*

Under an even weaker condition, we can compare the expected values of S_n and $S_{n,T}^B$. That is, under the NBUE assumption we have the following result by Belzunce *et al.* [10].

Theorem 9. *If X is NBUE then $E[S_n] \leq E[S_{n,T}^B]$ for all $T > 0$ and $n = 1, 2, \dots$*

The block replacement policy can also be compared with the age replacement policy

(see **Age Replacement Policies**). Under an age replacement policy, a unit is replaced upon failure or at age T , whichever comes first. Let $N_T^A(t)$ denote the number of failures in $[0, t]$ under an age replacement policy. It is of practical interest to be able to tell when $N_T^B(t)$, $N_T^A(t)$, and $N(t)$ can be compared. Barlow and Proschan [4] obtained the following result.

Theorem 10. *If X is IFR then $N_T^B(t) \leq_{\text{st}} N_T^A(t) \leq_{\text{st}} N(t)$ for all $t \geq 0$ and $T > 0$.*

On the other hand, Block *et al.* [11] showed the following result for DFR random variables.

Theorem 11. *If X is DFR then $N_T^B(t) \geq_{\text{st}} N_T^A(t)$ for all $t \geq 0$ and $T > 0$.*

FURTHER EXTENSIONS AND COMPARISONS WITH OTHER MAINTENANCE POLICIES

In this section, we consider some extensions of the usual block replacement policy and provide additional comparisons of entire processes instead of just univariate comparisons, as were considered in the previous section.

The block replacement policy has been mainly extended in three directions in the literature. First note that in the usual block replacement policy, when the unit fails it is repaired by replacing the failed unit with a new one which leads to a renewal process, but one can consider a different type of repair, and therefore, consider a different counting process $C \equiv \{C(t), t \geq 0\}$ which describes the times at which the repairing of unplanned failures occur. A second common extension is to consider that planned replacements occur at times $Z = \{z_1, z_2, \dots, z_n, \dots\}$ with $z_n \rightarrow \infty$. The list Z is often called *block* or *replacement schedule*. In the usual block replacement policy $z_n = nT$. At the end of this section, we also mention a third possible generalization which is to consider that the units in use are not identically distributed.

Following Block *et al.* [12,13], and especially Block and Savits [14], we start by describing how to construct, from an arbitrary counting process, a general

stochastic process that takes into account the underlying block replacement schedule. First, we fix the following notation. For an interval I of the form $[a, b]$, $[a, b)$, or $[a, \infty)$, we will denote by $S(I)$ the set of all right-continuous step-functions from I into the nonnegative integers, which start at zero and only increase by jumps of size one. Let us consider the sequence of numbers $0 = z_0 < z_1 < z_2 < \dots < z_n < \dots$, with $z_n \rightarrow \infty$, which represent the times at which planned replacements are going to be done, and let us denote by $C \equiv \{C(t), t \geq 0\}$, the counting process which describes the times at which the repairing of unplanned failures occur under the above-block replacement schedule. Next define the mapping

$$\Psi : \prod_{n=1}^{\infty} E_n \rightarrow S([0, \infty)),$$

where $E_n = S([0, z_n - z_{n-1}])$, by

$$\Psi(s_n)(t) = \begin{cases} s_1(t), & \text{if } 0 \leq t < z_1; \\ \sum_{j=1}^{i-1} s_j(z_j - z_{j-1}) + s_i(t - z_{i-1}), & \text{if } z_{i-1} \leq t < z_i; \end{cases}$$

where $s_n \in E_n$. Now we can construct a new counting process based on C , denoted by $C(Z)$, by setting $C(Z) = \Psi(C_n)$, where $C_n \equiv \{C_n(t), 0 \leq t \leq z_n - z_{n-1}\}$ are independent and identical copies of the counting process C restricted to the corresponding intervals $[0, z_n - z_{n-1}]$. The idea behind this construction is the following: until the time of the first planned replacement z_1 , the number of unplanned repairs is governed by the counting process C . Once the first planned replacement occurs, the counting process is restarted at time 0 and counts the number of unplanned repairs between the time of the first planned repair z_1 and z_2 , and so on.

At some points in the discussion that follows, we will replace the counting process C by the renewal process N considered in the previous section; this generates the new process $N(Z)$, which counts the number of unplanned repairs under a replacement

policy and planned replacements at times $z_1, z_2, \dots$. At other places in the discussion below, we will replace C by a nonhomogeneous Poisson process (see **Poisson Process and its Generalizations**). We note that nonhomogeneous Poisson processes arise in the context of reliability theory, when a minimal repair policy is utilized. Under a minimal repair policy, when a unit fails, the unit is restored to its working condition just prior to the failure. That is, if the unit fails at time t and is minimally repaired then, the time until its next failure (provided it occurs before the next block schedule instance) is stochastically the same as $[X - t | X > t]$ [15,16]. The counting process $N_m \equiv \{N_m(t), t \geq 0\}$ which counts the number of minimal repairs in any interval, is a nonhomogeneous Poisson process with mean function $E[N_m(t)] = -\log P(X > t)$. In the previous construction, if we replace the counting process C by a minimal repair process, then we have the counting process $N_m(Z)$ which counts the number of unplanned repairs under a minimal repair policy and planned replacements at times $z_1, z_2, \dots, z_n, \dots$.

In order to provide results for the comparison of entire processes, we recall two notions that allow us to compare random vectors and stochastic processes.

Given two n -dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$, the random vector $\mathbf{X}$ is said to be smaller than the random vector $\mathbf{Y}$ in the multivariate stochastic order (denoted as $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$), if $E[\phi(\mathbf{X})] \leq E[\phi(\mathbf{Y})]$ for all increasing functions $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ for which previous expectations exist.

Furthermore, given two stochastic processes $C \equiv \{C(t), t \geq 0\}$ and $D \equiv \{D(t), t \geq 0\}$, the stochastic process C is said to be smaller than the stochastic process D in the ordinary stochastic order (denoted by $C \leq_{\text{st}} D$) if for all choices of an integer n and $0 \leq t_1 < t_2 < \dots < t_n \in \mathbb{R}$, we have

$$(C(t_1), C(t_2), \dots, C(t_n)) \leq_{\text{st}} (D(t_1), D(t_2), \dots, D(t_n)).$$

The following theorem gives several results which stochastically compare, as described above, the replacement policies discussed above and the ones considered in

the previous section. These results can be found in the papers by Block *et al.* [12–14]. They indicate various potentially useful stochastic comparisons, of the counting processes of unplanned replacements, when the item lifetime is NBU [NWU].

Theorem 12. *If X is NBU [NWU] then for any $Z = \{z_1, z_2, \dots, z_n, \dots\}$, where $0 < z_1 < z_2 < \dots < z_n < \dots$, we have the following comparisons:*

- (i) $N(Z) \leq_{\text{st}} [\geq_{\text{st}}] N$,
- (ii) $N(Z) \leq_{\text{st}} [\geq_{\text{st}}] N_m(Z)$,
- (iii) $N(Z) \leq_{\text{st}} [\geq_{\text{st}}] N_m$,
- (iv) $N_m(Z) \leq_{\text{st}} [\geq_{\text{st}}] N_m$.

If any of the (nonparenthesized) inequalities in (i), (iii), and (iv) above hold for all block replacement schedules Z , then X is NBU. It is not known if this is also the case for (ii).

The results in Theorem 12 were extended to more general settings in Ref. 17. They derived a host of comparisons of replacement policies via the general theory of point processes. Another extension, where the repair upon failure is imperfect, was obtained in Refs 18 and 19.

It is also natural to consider in which sense different sets of times for planned replacement (that is, different block replacement schedules) modify the underlying stochastic processes. Some results in this vein were given by Block *et al.* [12]. Before stating them, we need to recall the definition of refinement. Given two block replacement schedules $Z = \{z_1, z_2, \dots, z_n, \dots\}$ and $V = \{v_1, v_2, \dots, v_n, \dots\}$, where $0 < z_1 < z_2 < \dots < z_n < \dots$ and $0 < v_1 < v_2 < \dots < v_n < \dots$, we say that V is a refinement of Z if $V \supset Z$. If V and Z are different sets of times for planned replacements, then a refinement of a block policy results in the addition of new times of planned replacements. With this notion we have the following results. They can be found in Refs 12, 13, and 20.

Theorem 13. *The random variable X is NBU, if and only if, any one of the following equivalent conditions holds:*

- (i) $N(V) \leq_{\text{st}} N(Z)$ for every $V \supset Z$.

(ii) $N_m(V) \leq_{\text{st}} N_m(Z)$ for every $V \supset Z$.

Again, it is worthwhile to mention that Shaked and Szekli [17] derived some variations of Theorem 13. For example, if X in Theorem 13 is IFR, rather than just NBU, then Shaked and Szekli [17] obtained stronger comparisons between $N(V)$ and $N(Z)$, and between $N_m(V)$ and $N_m(Z)$, than the comparisons given in Theorem 13. Another extension, where the repair upon failure is imperfect, was obtained by Li and Shaked [18]. In the latter paper, the authors also showed for the process that counts the failures under block replacement policy with imperfect repair, that it is monotone with respect to the order $\leq_{\text{st}}$ as a function of the quality of the imperfect repair procedure.

It is possible to provide additional results if we replace the unit under a block replacement policy, with random lifetime X , by another one with random lifetime Y . In this case, we want to denote the dependence of the resulting processes on the distribution functions of X and Y . Thus, for the processes $N(Z)$ and $N_m(Z)$, when the underlying random variable X has the distribution function F , we will write $N(Z, F)$ and $N_m(Z, F)$. Whereas, when the underlying random variable Y has the distribution function G , then we will write $N(Z, G)$ and $N_m(Z, G)$, respectively. For this case, we have the following result by Block *et al.* [12]. Again, these results give a condition under which some useful comparisons of counting processes of unplanned replacements can be made.

Theorem 14. *The random variables X and Y satisfy $X \leq_{\text{st}} Y$ if, and only if, any one of the following equivalent conditions holds:*

- (i) $N(Z, G) \leq_{\text{st}} N(Z, F)$ for all block replacement schedules Z .
- (ii) $N_m(Z, G) \leq_{\text{st}} N_m(Z, F)$ for all block replacement schedules Z .

Once again, it is worthwhile to mention that Shaked and Szekli [17] obtained a version of Theorem 14 in which under stronger

conditions on X and Y , one obtains stronger conclusions.

As mentioned in the beginning of this section, another extension of the block replacement policy $N(Z)$ is to consider that any time that a unit is replaced, it is replaced by a unit that is not identically distributed as the failed one. Thus, instead of having just one random variable X , we have a sequence of random lifetimes $X_1, X_2, \dots, X_n, \dots$. The sequence of random variables $X_1, X_2, \dots, X_n, \dots$ is said to be NBU [NWU] in sequence if for any $t \geq 0$ and any nonnegative integer n we have

$$[X_n - t | X_n > t] \leq_{\text{st}} [\geq_{\text{st}}] X_{n+1}.$$

For this case, Langberg [21] provided some results similar to the previous ones; we do not give the details here.

COST ANALYSIS

A topic of interest involving block replacement policies is the determination of the optimal time T . In this case, the problem is to determine the expected long-run cost per unit time and then find the value T that minimizes this quantity. For more information on this topic, the reader is directed to Ref. 22 (see also **Optimal Replacement and Inspection Policies**). Here we just give the expected long run cost for block replacement policy under a usual replacement and under a minimal repair.

In order to present the problem, we fix some notation: we will denote by c_f , the cost of replacement of a failed unit and by c_b , the cost of a planned replacement at times $T, 2T, \dots$. In this case, the expected long run cost per unit is

$$\frac{c_f M(T) + c_b}{T},$$

where $M(t)$ is the renewal function.

If instead of the usual replacement policy, we consider a minimal repair for failed units, and the cost of a minimal repair is denoted by c_m , then expected long run cost per unit is

given by

$$\frac{c_m \int_0^T r(u) du + c_b}{T},$$

where r is the failure rate function of X .

Some further analysis can be found in Refs [23–25], and more recently, in Ref. 26 and the references therein.

REFERENCES

1. Barlow RE, Proschan F. Statistical theory of reliability and life testing: probability models. New York: Holt, Rinehart and Winston; 1981.
2. Müller J, Stoyan D. Comparison methods for stochastic models and risks. Chichester: Wiley; 2002.
3. Shaked M, Shanthikumar JG. Stochastic orders. New York: Springer; 2007.
4. Barlow RE, Proschan F. Comparison of replacement policies and renewal theory implications. *Ann Math Stat* 1964;35:577–589.
5. Marshall AW, Proschan F. Classes of distributions applicable in replacement with renewal theory implications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*; Volume 1; 1972. pp. 395–415.
6. Shaked M, Zhu H. Some results on block replacement policies and renewal theory. *J Appl Probab* 1992;29:932–946.
7. Brown M. Bounds, inequalities, and monotonicity properties for some specialized renewal processes. *Ann Probab* 1980;8:227–240.
8. Yue D, Cao J. The NBUL class of life distribution and replacement policy comparisons. *Nav Res Log* 2001;48:578–591.
9. Belzunce F, Ortega E-M, Ruiz JM. A note on replacement policy comparisons from NBUC lifetime of the unit. *Stat Pap* 2005;46:509–522.
10. Belzunce F, Ortega E-M, Ruiz JM. Comparison of expected failure times for several replacement policies. *IEEE Trans Reliab* 2006;55:490–495.
11. Block HW, Langberg NA, Savits TH. Repair replacement policies. *J Appl Probab* 1993;30:194–206.
12. Block HW, Langberg NA, Savits TH. Maintenance comparisons: block policies. *J Appl Probab* 1990a;27:649–657.
13. Block HW, Langberg NA, Savits TH. Comparisons for maintenance policies involving complete and minimal repair. In: Block HW, Sampson AR, Savits TH, editors. Volume 16, *Topics in statistical dependence, IMS lecture notes - monograph series*. Hayward (CA): Institute of Mathematical Statistics; 1990b. pp. 57–68.
14. Block HW, Savits TH. Comparisons of maintenance policies. In: Shaked M, Shanthikumar JG, editors. *Stochastic orders and their applications*. Boston (MA): Academic Press; 1994. pp. 463–483.
15. Barlow RE, Hunter LC. Optimum preventive maintenance policies. *Oper Res* 1960;8:90–100.
16. Ascher H, Feingold H. Repairable systems reliability. New York: Decker; 1984.
17. Shaked M, Szekli R. Comparison of replacement policies via point processes. *Adv Appl Probab* 1995;27:1079–1103.
18. Li H, Shaked M. Imperfect repair models with preventive maintenance. *J Appl Probab* 2003;40:1043–1059.
19. Li H, Xu SH. On the coordinated random group replacement policy in multivariate repairable systems. *Oper Res* 2004;52:464–477.
20. Shaked M, Shanthikumar JG. Some replacement policies in a random environment. *Probab Eng Inform Sci* 1989;3:117–134.
21. Langberg NA. Comparison of replacement policies. *J Appl Probab* 1988;25:780–788.
22. Dekker R. Block replacement. In: Ruggeri F, Faltin F, Kenett R, editors. *Encyclopedia of statistics in quality and reliability*. London: Wiley; 2007. pp. 229–233.
23. Berg M. A marginal cost analysis for preventive replacement policies. *Eur J Oper Res* 1980;4:135–142.
24. Berg M, Cleroux R. The block replacement model with minimal repair and random repair costs. *J Stat Comput Simul* 1982;15:1–7.
25. Savits TH. A cost relationship between age and block replacement policies. *J Appl Probab* 1988;25:789–796.
26. Sheu SH, Griffith WS. Extended block replacement policy with shock models and used items. *Eur J Oper Res* 2002;140:50–60.

BRANCH AND CUT

JOHN E. MITCHELL
Department of Mathematical Sciences,
Rensselaer Polytechnic Institute,
Troy, New York

Combinatorial optimization problems can often be formulated as mixed integer linear programming problems, as discussed in the article titled **Formulating Good MILP Models** in this encyclopedia. They can then be solved using *branch-and-cut*, which is an exact algorithm combining branch-and-bound (see **Branch-and-Bound Algorithms** and cutting planes (see the section titled “Cutting Planes”, in the encyclopedia and its articles). The basic idea is to take a linear programming relaxation of the problem, solve the relaxation, and then either improve the relaxation by adding additional valid constraints or split the problem into two or more subproblems and repeat the process.

Gomory first proposed strengthening linear programming relaxations of integer programming problems by incorporating extra constraints (or *cutting planes*) in the 1950s [1]. These cutting planes are derived from the optimal simplex tableau, so they are broadly applicable. However, they fell into disfavor for many years because they seemed to get stuck and run out of power. The cuts are now used in more sophisticated ways and are incorporated into the major commercial packages for integer programming. These packages also include several other families of general cutting planes. General cutting planes are discussed in the sections titled “General Cutting Planes” and “Cutting Planes” in this encyclopedia.

Land and Doig [2] proposed a branch-and-bound approach in 1960. In branch-and-bound, the linear programming relaxation of the integer program is solved. If the solution is fractional, then the problem is split into two subproblems and the process is repeated, creating a tree of subproblems. The value

of the linear programming relaxation gives a lower bound on the optimal value of the integer program at the corresponding node of the tree, for a minimization problem. If this lower bound is greater than the value of a known feasible solution, then the node can be pruned, which greatly reduces the total size of the tree. Branch-and-bound became more popular than cutting plane methods for many years, because of the computational difficulties with the latter.

Interest in cutting planes resurfaced in the 1980s. Crowder *et al.* [3] showed the strength of general cuts obtained by regarding a single row of an integer program as a knapsack problem. In addition, the polyhedral theory of many classes of problems was derived, and this led to problem-specific cutting planes that were very successful, leading to great speed-ups in computational time when compared to using branch-and-bound. For many classes of problems (e.g., the traveling salesman problem), an initial integer programming formulation contains a large number of constraints, possibly even an exponential number. In such a situation, it is not computationally attractive to explicitly include all of these constraints in the LP relaxation, and they can be added selectively as cutting planes. Problem-specific cutting planes are the subject of the section titled “Problem Specific Cutting Planes” of this article.

In theory, pure cutting plane methods can be used to solve integer programs, without the need to employ branching. In practice, cutting plane methods appear to tail-off, and so it becomes faster to combine together the two approaches. Initially, cutting planes were employed only at the root node of the tree, in an approach now called cut-and-branch. Examples include Ref. 3 as well as work on the traveling salesman problem [4]. The set of cuts generated at the root node is not exhaustive, so it is possible that the subsequent branch-and-bound approach leads to an integer solution that is not actually feasible in the integer program. In such a situation, it

is then necessary to add additional cuts and restart the process.

Later in the 1980s, cutting planes were employed throughout the tree. The best known of these results is for the traveling salesman problem, starting with the work of Padberg and Rinaldi [5]; an excellent discussion of this problem is contained in the book by Applegate *et al.* [6]. Other notable early work includes the research of Grötschel *et al.* on the linear ordering problem [7] and on the maximum cut problem [8]. In the 1990s, it was discovered that the general cutting planes can actually be very effective in branch-and-cut approaches to integer programming problems, thanks to the work of Balas *et al.* [9,10]. The integration of cutting planes with branch-and-bound is discussed in more detail in the section titled “The Branch-and-Cut Algorithm” of this article.

Refinements and extensions are discussed in the section titled “Refinements and Extensions.” For example, it is possible to generalize the branch-and-cut approach to solve mixed integer nonlinear programming problems and even nonlinear programs without integrality constraints. Also in this section, we consider exploitation of parallel computational hardware. In addition to parallel computers, it is common for problems to be solved on clusters of computers (including cloud computers) or on the multicore processors, now frequent in desktop and even in laptop computers. Branch-and-cut algorithms can be parallelized by solving different nodes of the tree on different processors.

The Lanchester Prize winning book by Nemhauser and Wolsey [11] contains an excellent discussion of polyhedral theory, integer programming, and branch-and-cut. Other very good and relevant books are those by Wolsey [12] and Lee [13]. The three-volume text by Schrijver [14] is also an excellent reference. Various surveys of branch-and-cut have appeared over the years, including Refs 15–18. The aforementioned text on the traveling salesman problem by Applegate *et al.* [6] provides a very accessible development of integer programming, including branch-and-cut.

GENERAL CUTTING PLANES

Our standard form integer programming problem is the following:

$$\begin{array}{ll} \min & c^T x \\ \text{subject to} & Ax \geq b \\ & x \geq 0 \\ & x_i \text{ integer } \forall i \in I, \end{array} \quad (\text{ILP})$$

where x and c are n -vectors, b is an m -vector, A is an $m \times n$ matrix, and I is a subset of the indices $\{1, \dots, n\}$. Any upper bound constraints on the variables are included in the inequality constraints $Ax \geq b$. The optimal value of (ILP) is denoted by z^* . Any feasible solution $\bar{x}$ to (ILP) provides an upper bound $c^T \bar{x}$ on the optimal value z^* of the problem. A lower bound can be obtained by solving a relaxation of (ILP). In this article, we are concerned with linear programming (LP) relaxations, which are obtained by relaxing the integrality restriction.

The lower bound provided by the LP relaxation can be improved by tightening up the relaxation through the addition of valid linear constraints. Typically, these constraints are satisfied by all feasible solutions to (ILP) but violated by the optimal solution to the LP relaxation. The LP relaxation can be solved again after the addition of the constraints, and the process is repeated.

Cutting planes are discussed in detail in the section titled “Cutting Planes” in this encyclopedia, and its articles. In this section, we summarize some of the cutting planes that have been used to solve general integer programming problems, and which are now included in commercial integer programming packages (see Bixby and Rothberg [19] and Ashford [20]).

Gomory cuts are derived from a row of the optimal simplex tableau for the LP relaxation [21]. These were generalized by Chvátal [22], giving Chvátal–Gomory cuts, which can be derived from any nonnegative linear combination of the linear constraints of (ILP). For simplicity, we consider the case where all the variables are required to be integer. In particular, if $u \in \mathbf{R}^m$ is nonnegative then the constraint

$$u^T Ax \geq u^T b$$

is valid for the LP relaxation of (ILP). Since $x \geq 0$, this constraint can be weakened to

$$\lceil u^T A \rceil x \geq u^T b,$$

where $\lceil u^T A \rceil$ is the n -dimensional row vector obtained by rounding up each entry of the row vector $u^T A$. Since each entry of x is non-negative, the left-hand side of this inequality must be integral for any feasible solution to (ILP). Hence, we can round up the right hand side to obtain the valid constraint

$$\lceil u^T A \rceil x \geq \lceil u^T b \rceil,$$

a Chvátal–Gomory cutting plane. Chvátal showed that any valid inequality for (ILP) can be obtained by repeatedly applying this rounding procedure [22].

Gomory cutting planes fell out of favor for many years, but computational results in the 1990s [10,23] showed that they could be very helpful. According to Bixby and Rothberg [19], they are the most useful of the general cutting planes. Fischetti and Lodi [24] showed that just one round of generating every possible inequality from the original constraints $Ax \geq b$ can give a very good approximation to the convex hull of the set of feasible solutions. Letchford [25] described a method for generating deep Chvátal–Gomory cutting planes. One of the problems with Gomory cutting planes is that eventually dual degeneracy is encountered, which can lead to a basis matrix with a large condition number if care is not taken in the generation of the cuts. Zanette *et al.* [26] demonstrated that employing lexicographic cut generation rules can lead to a set of Gomory cutting planes that interact well with one another and allow a pure cutting plane method to work effectively; see also the related paper in Ref. 27. Gomory cutting planes can also be derived for mixed integer programs; see Marchand and Wolsey [28] for some computational results. Chvátal–Gomory cutting planes are considered in far more detail in the article titled **Gomory Cuts** in this encyclopedia.

Cover inequalities are inequalities that are valid for knapsack problems. Crowder *et al.* [3] considered each row of an integer

program as a separate knapsack problem, and then generated cover inequalities for the individual rows. They showed that this powerful technique could be employed in a cut-and-branch algorithm to solve general integer programming problems. Their approach has been considerably refined in recent years, becoming a standard part of branch-and-cut implementations. These inequalities are discussed in detail in the article titled **Cover Inequalities** in this encyclopedia.

The theory of disjunctive inequalities [29] gives a methodology for generating general cutting planes that can be powerful on certain hard integer programs [9,30]. The theory was originally developed for binary variables and has been extended. Given a binary variable x_i , the process is to find the convex hull of two sets: the set of feasible points in the LP relaxation with $x_i = 0$, and the set of feasible points in the LP relaxation with $x_i = 1$. The process is then iterated over all the binary variables. The theory gives a method to systematically construct the convex hull of the set of feasible solutions to an integer program. Generation of a cut may require solution of a linear program, so in practice this method is used selectively. There are methods for generating disjunctive cuts using the optimal simplex tableau for the relaxation; see Balas and Perregaard [31,32]. For more information, see the articles titled **Lift-and-Project Inequalities** and **Disjunctive Programming** in this encyclopedia.

Recently, there has been interest in deriving cuts using two rows of the simplex tableau, which potentially could be stronger than cuts derived from just a single row. These methods use ideas from lattice theory and group theory. For more details, see Andersen *et al.* [33] and also the survey article by Dey and Tramontani [34].

Computational experience with different classes of general cutting planes is detailed by Bixby and Rothberg [19]. One point the authors make is that different families of cutting planes can interact with each other, so the benefit of using all of several different families of cuts is not necessarily equal to the product of the benefits of using each family individually.

PROBLEM-SPECIFIC CUTTING PLANES

Many combinatorial optimization problems can only be expressed as integer linear programming problems with an exponential number of constraints. For example, a standard formulation of the traveling salesman problem uses degree constraints to ensure each city is visited exactly once. It also requires the inclusion of subtour elimination constraints to ensure that any integral solution corresponds to a tour that connects all the cities, and the number of subtour elimination constraints is exponential in the number of cities. Thus, it is impractical to include all of these constraints in the integer programming formulation, and they should be added as cutting planes. The first demonstration of the strength of this cutting plane approach was by Dantzig *et al.* [35], who showed that a problem with 42 cities could be solved to optimality by adding just a limited number of subtour elimination constraints, and some other cutting planes. Cutting plane methods for the traveling salesman problem were revisited in the 1980s [36,37], and subsequently the work of Applegate *et al.* [6] has realized an algorithm that can find provably optimal solutions to problems with as many as 85,900 cities. Their code Concorde is freely available.

The convex hull of the set of feasible integer solutions to a combinatorial optimization problem is a polyhedron. (See **Basic Polyhedral Theory** for more details.) If a linear programming description of this polyhedron is known, then the problem can be solved effectively. However, the number of facets of the polyhedron is large (often exponential) for interesting combinatorial optimization problems. Thus, it is necessary to add the constraints selectively. The strongest cutting planes correspond to facets, and families of facets have been determined for many different problems. For example, the subtour elimination constraints mentioned earlier define facets.

Other problems, for which cutting plane methods have been developed, include the linear ordering problems [7,38] with triangle inequalities and other classes of facet defining inequalities, the maxcut problem

[8,39–44] with cycle-odd subset inequalities, matching problems [45,46], clique and coloring problems [47,48], fixed charge network flow problems [49], vehicle routing problems [50–52], and facility location problems [53].

Knowledge of a strong family of cutting planes is only useful in practice if effective separation routines are also developed, which can find violated constraints in the family efficiently. These separation routines can be simple or involved, even for the same class of constraints. For example, cycle-odd subset inequalities for maxcut problems can be checked by enumeration for all short cycles, but in order to guarantee that any violated inequality can be found it is necessary to use a max-flow algorithm for a graph derived from the original one [41]. Violated subtour elimination constraints for the traveling salesman problem can be found by searching for connected components in the solution to the LP relaxation, but this may not find all violated constraints, so more expensive routines have also been developed [6].

Let X be a feasibility integer program and let x be a point. The *separation problem* for X and x is to find a cutting plane that separates x from the convex hull of X , or determine that x is in this convex hull. If the separation problem for any point x can be solved in time no greater than $g(X)$, then problem X itself can be solved in time polynomial in $g(X)$ using the ellipsoid algorithm. This observation can be generalized as the equivalence of separation and optimization problems [54]. It follows that for an NP-Complete problem, it will not be possible to find a cutting plane for each point not in the convex hull in polynomial time (unless $P = NP$).

THE BRANCH-AND-CUT ALGORITHM

A branch-and-cut algorithm is outlined in Algorithm 1. The set of active nodes in the branch-and-cut tree is denoted by L . The value of the best known feasible point for (ILP) is stored as $\bar{z}$, and provides an upper bound on the optimal value of the integer program. This point is called the *incumbent solution*. We use $\underline{z}_l$ to denote a lower bound on the optimal value of the current subproblem l , under consideration. This lower bound

is initialized to the value of the parent node, and is then updated to the value of the LP relaxation of the subproblem.

Algorithm 1. A general branch-and-cut algorithm.

1. *Initialization:* Denote the initial integer programming problem by ILP^0 and define the set of active nodes to be $L = \{ILP^0\}$. Let $\bar{z} = +\infty$. Set $z_l = -\infty$ for the initial problem $l \in L$.
2. *Termination:* If $L = \emptyset$, then STOP. If $\bar{z} = \infty$ then (ILP) is infeasible; else, the solution x^* which yielded the incumbent objective value $\bar{z}$ in Step 7(b) or Step 5 is optimal.
3. *Problem selection:* Select and delete a problem ILP^l from L .
4. *Relaxation:* Solve the LP relaxation of ILP^l . If the relaxation is infeasible, set $z_l = +\infty$ and go to Step 7. If the relaxation is unbounded set $z_l = -\infty$. If the relaxation has a finite optimal value let x^{LR} be an optimal solution and set $z_l = c^T x^{LR}$.
5. *Heuristic Rounding:* If x^{LR} is not integral, and if desired, use a rounding approach or a heuristic approach to construct a feasible integral solution x^{IH} . Update $\bar{z} = \min\{c^T x^{IH}, \bar{z}\}$.
6. *Add cutting planes:* If desired, search for cutting planes that are violated by x^{LR} ; if any are found, add them to the relaxation and return to Step 4.
7. *Fathoming and Pruning:*
 - a. *Fathom by bounds or infeasibility:* If $z_l \geq \bar{z}$ go to Step 2.
 - b. *Fathom by integrality:* If $z_l < \bar{z}$ and x^{LR} is integral feasible, update $\bar{z} = z_l$, delete from L all problems with $z_l \geq \bar{z}$, and go to Step 2.
8. *Partitioning:* Let $\{S^{lj}\}_{j=1,\dots,k}$ be a partition of the constraint set S^l of problem ILP^l . Add problems $\{ILP^{lj}\}_{j=1,\dots,k}$ to L , where ILP^{lj} is ILP^l with feasible region restricted to S^{lj} , and set $z_{lj} = z_l$ for $j = 1, \dots, k$. Return to Step 2.

Without the inclusion of Step 6, this becomes a branch-and-bound algorithm.

A crucial point with branch-and-bound is that a subproblem l can be discarded once $z_l \geq \bar{z}$, since it is then known that no feasible solution to the subproblem can be better than the incumbent solution. The other method for fathoming in Step 7 is when the optimal solution to the LP relaxation of the subproblem is feasible in the integer program, since this LP solution then solves the subproblem. The reader is referred to the article titled **Branch-and-Bound Algorithms** in this encyclopedia for far more discussion of branch-and-bound, including preprocessing, options for branching, and reduced cost fixing and its exploitation.

The particular procedure employed in Step 5 can be a generic rounding procedure, or it can be a rounding procedure modified to exploit the particular structure of the problem, or it can be a heuristic initiated either at the point x^{LR} or at a rounded version of this point.

The Step 6 decision of when to add cutting planes and when to branch can probably only be resolved through computational experimentation. The conclusion is dependent on the particular class of integer program and on the types of cutting planes considered. Different types of cutting planes interact with each other, so care is needed in experimentation in order to determine reproducible benefits.

Cutting planes generated at one node of the tree may not be valid at another node. One option is to treat the cuts as local, and only use them for descendants of the node where they are generated. The disadvantage of this approach is that it becomes necessary to store several different sets of constraints, for different parts of the tree. Alternatively, the constraints can be modified to make them valid throughout the tree, using a process called *lifting*. In lifting, the value of the slack in the constraint is checked in the remainder of the tree, either by solving integer programs to get strong liftings or LP relaxations to get somewhat weaker lifted inequalities. Lifting is a general process for strengthening and modifying constraints and is discussed in the article titled **Lifting Techniques for Mixed Integer Programming** in this encyclopedia.

One important aspect of a general integer programming code is preprocessing. One aspect of preprocessing is to tighten bounds on the variables and constraints by logical arguments and possibly solving LPs. This may lead to variable fixing or constraint elimination, and can have a dramatic impact on runtime (an average improvement of a factor of 10 for the problems considered in Ref. 19). Far more about preprocessing can be found in the article titled ***Branch-and-Bound Algorithms*** in this encyclopedia.

REFINEMENTS AND EXTENSIONS

Computational implementations of branch-and-cut are now very sophisticated and include many ideas from the research literature of the last 30 years. Commercial codes include CPLEX and GuRoBi [19], and XPRESS-MP [20]. Recent high-quality free software includes the COIN-OR branch-and-cut package Cbc [55], and the packages ABACuS [56] and MINTO [57]. In this section, we consider some possible enhancements to current integer programming solvers.

Before discussing enhancements, we note that the computing environment is becoming ever more parallel. There have been sophisticated parallel computers for many years, and these have become more widespread, with local machines with at least 100 processors available to many users. In addition, there are now clusters of homogeneous or heterogeneous processors linked together using software, there is the availability of cloud computing, and multicore processors are common in desktop and even in laptop computers. This is an environment that must be exploited for a branch-and-cut implementation to remain competitive. Fortunately, the branching aspect of these algorithms leads to a natural way to parallelize: different subproblems in the tree are solved in different processors of the machine. Load balancing requires careful consideration, but in principle branch-and-cut algorithms should flourish in a world of parallel computers. For more concrete discussion of these issues, see Ref. 55.

A standard model for the traveling salesman problem is to use one variable for

each edge, so if the graph has n vertices then there are $O(n^2)$ variables. Based just on the objective function coefficients, it is clear that the great majority of the variables can be (at least temporarily) fixed at zero. Thus, we can work with integer and linear programs where the number of variables is $O(n)$. Before fathoming any node of the tree, the eliminated edges can be checked using reduced costs, to see if they would have been helpful. This pricing step may lead to the introduction of variables, and the resulting algorithm is a form of branch-and-price-and-cut. For more on algorithms of this type, see the article titled ***Branch-Price-and-Cut Algorithms*** in this encyclopedia.

Many integer programming formulations possess a natural symmetry. For example, when scheduling jobs on several identical machines, the important decision is determining which set of jobs go together on a particular machine and then sequencing those jobs. Which machine performs which particular set of jobs does not matter. This poses difficulties for a standard branch-and-cut approach, because many variables have to be fixed in the branching tree before the symmetry is broken. There has been research on methods for breaking symmetry, using ideas from group theory and algebra [58,59]. See the article titled ***Symmetry Handling in Mixed-Integer Programming*** in this encyclopedia for more information.

Classically, cutting planes are satisfied by all feasible solutions. Cuts could potentially be strengthened by requiring only that all *optimal* solutions satisfy them. A similar possibility is noted in dual stabilization of column generation algorithms; see Ref. 60 for example.

Fischetti *et al.* [61] note that it is possible to use sophisticated mixed integer programming techniques within a branch-and-cut solver. For example, the separation problems for some classes of cutting planes are themselves hard integer programs and so it is beneficial to use whatever MIP techniques are available in order to find strong cutting planes; see Ref. 47 for example. Construction of an initial feasible solution can also be performed by using integer programming techniques such as local branching [62].

It is superior to use interior point methods instead of the simplex method for some problems, at least in certain parts of the branch-and-cut process. Interior point methods have two potential advantages: first, cuts are generated from a more central solution, which leads to deeper cuts; secondly, interior point methods can solve large problems more quickly than simplex, and when many cuts are added at once the warm-start benefit enjoyed by simplex is no longer so advantageous. See Ref. 63 for a survey, and Ref. 44–46 for computational results. Branch-and-cut can also be integrated with convex relaxations of the integer program, such as semidefinite relaxations [63,65–67].

Branch-and-cut algorithms have also been developed for mixed integer nonlinear programming problems. The cuts used in these approaches are typically disjunctive cuts. See Bonami *et al.* [68], which describes several different possible branch-and-cut approaches.

Branch-and-cut can even be used for problems without integrality restrictions. For example, Tawarmalani and Sahinidis [69,70] describe an approach for global optimization of general nonlinear programs, and Vandenbussche and Nemhauser [71] show how branch-and-cut can be used to solve a nonconvex quadratic program by exploiting the optimality conditions.

CONCLUSIONS

The performance of branch-and-bound methods for integer programming has been dramatically improved by the inclusion of cutting planes, leading to branch-and-cut. The resulting exact methods have been successfully implemented in powerful general purpose solvers for mixed integer programs, and they are the method of choice for solving hard integer programs to optimality. The software has improved by several orders of magnitude in the last few years. Branch-and-cut solvers have also been developed for specific problems such as the traveling salesman problem, with such a code currently able to solve larger problems to optimality than any other approach. Branch-and-cut methods

are still the subject of active research, with various ideas showing promise. The methods have also been extended to solve mixed integer nonlinear programs, and other classes of optimization problems.

Acknowledgment

The work of this author was supported in part by the National Science Foundation under grant DMS-0715446.

REFERENCES

1. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64:275–278.
2. Land AH, Doig AG. An automatic method of solving discrete programming problems. *Econometrica* 1960;28:497–520.
3. Crowder HP, Johnson EL, Padberg M. Solving large-scale zero-one linear programming problems. *Oper Res* 1983;31:803–834.
4. Crowder HP, Padberg M. Solving large-scale symmetric travelling salesman problems to optimality. *Manag Sci* 1980;26:495–509.
5. Padberg M, Rinaldi G. Optimization of a 532-city traveling salesman problem by branch and cut. *Oper Res Lett* 1987;6:1–8.
6. Applegate D, Bixby R, Chvátal V, *et al.* The traveling salesman problem: a computational study. Princeton, NJ: Princeton University Press; 2006.
7. Grötschel M, Jünger M, Reinelt G. A cutting plane algorithm for the linear ordering problem. *Oper Res* 1984;32:1195–1220.
8. Grötschel M, Jünger M, Reinelt G. Calculating exact ground states of spin glasses: A polyhedral approach. In: van Hemmen JL, Morgenstern I, editors. *Proceedings of the Heidelberg Colloquium on Glassy Dynamics*. Berlin: Springer; 1987. pp. 325–353.
9. Balas E, Ceria S, Cornuéjols G. Mixed 0–1 programming by lift-and-project in a branch-and-cut framework. *Manag Sci* 1996;42(9):1229–1246.
10. Balas E, Ceria S, Cornuéjols G, *et al.* Gomory cuts revisited. *Oper Res Lett* 1996;19:1–9.
11. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. New York: John Wiley; 1988.
12. Wolsey LA. *Integer programming*. New York: John Wiley; 1998.

13. Lee J. A first course in combinatorial optimization. Cambridge: Cambridge University Press; 2004.
14. Schrijver A. Combinatorial optimization – polyhedra and efficiency. Berlin: Springer; 2003.
15. Caprara A, Fischetti M. Branch and cut algorithms. In: Dell’Amico M, Maffioli F, Martello S, editors, Annotated bibliographies in combinatorial optimization, Chapter 4. Chichester: John Wiley; 1997.
16. Jünger M, Reinelt G, Thienel S. Practical problem solving with cutting plane algorithms in combinatorial optimization. In: Combinatorial optimization: DIMACS series in discrete mathematics and theoretical computer science, volume 20. Providence, RI: AMS; 1995. pp. 111–152.
17. Marchand H, Martin A, Weismantel R, *et al.* Cutting planes in integer and mixed integer programming. *Discrete Appl Math* 2002; 123:397–446.
18. Mitchell JE. Branch-and-cut algorithms for combinatorial optimization problems. In: Pardalos PM, Resende MGC, editors. Handbook of applied optimization. Oxford University Press; 2002. pp. 65–77.
19. Bixby RE, Rothberg E. Progress in computational mixed integer programming—a look back from the other side of the tipping point. *Ann Oper Res* 2007;149:37–41.
20. Ashford R. Mixed integer programming: a historical perspective with xpress-mp. *Ann Oper Res* 2007;149:5–17.
21. Gomory RE. An algorithm for integer solutions to linear programs. In: Graves RL, Wolfe P, editors. Recent advances in mathematical programming. New York: McGraw-Hill; 1963. pp. 269–302.
22. Chvátal V. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete Math* 1973;4:305–337.
23. Ceria S, Cornuéjols G, Dawande M. Combining and strengthening Gomory cuts. Volume 920, In: Balas E, Clausen J, editors. Lecture Notes in Computer Science. Heidelberg: Springer; 1995.
24. Fischetti M, Lodi A. Optimizing over the first Chvátal closure. *Math Program* 2007; 110(1):3–20.
25. Letchford AN. Totally tight Chvátal-Gomory cuts. *Oper Res Lett* 2002;30(2):71–73.
26. Zanette A, Fischetti M, Balas E. Lexicography and degeneracy: can a pure cutting plane algorithm work?. *Math Program* 2010. DOI: 10.107/s1007-009-0335-0.
27. Balas E, Fischetti M, Zanette A. (2009) On the enumerative nature of Gomory’s dual cutting plane method. Technical report, DEI, Dipartimento di Ingegneria dell’Informazione, University of Padova, Italy.
28. Marchand H, Wolsey LA. Aggregation and mixed integer rounding to solve MIPs. *Oper Res* 2001;49(3):363–371.
29. Balas E. Disjunctive programming. *Ann Discrete Math* 1979;5:3–51.
30. Ceria S, Pataki G. Solving integer and disjunctive programs by lift-and-project. In: Proceedings of the Sixth IPCO Conference. 1998.
31. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0–1 programming. *Math Program* 2003;94(2-3):221–245.
32. Perregaard M. Generating disjunctive cuts for mixed integer programs. PhD thesis, Carnegie Mellon University, Graduate School of Industrial Administration, Pittsburgh, PA, 2003.
33. Andersen K, Louveaux Q, Wolsey L, *et al.* Inequalities from two rows of a simplex tableau. *LNCS* 4513 2007;1–15.
34. Dey SS, Tramontani A. Recent developments in multi-row cuts. *Optima* 2009;80:2–8.
35. Dantzig GB, Fulkerson DR, Johnson SM. Solutions of a large-scale travelling salesman problem. *Oper Res* 1954;2:393–410.
36. Grötschel M, Holland O. Solution of large-scale travelling salesman problems. *Math Program* 1991;51(2):141–202.
37. Padberg M, Rinaldi G. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM Rev* 1991;33(1):60–100.
38. Grötschel M, Jünger M, Reinelt G. Facets of the linear ordering polytope. *Math Program* 1985;33:43–60.
39. Barahona F, Grötschel M, Jünger M, *et al.* An application of combinatorial optimization to statistical physics and circuit layout design. *Oper Res* 1988;36(3):493–513.
40. Barahona F, Jünger M, Reinelt G. Experiments in quadratic 0-1 programming. *Math Program* 1989;44(2):127–137.
41. Barahona F, Mahjoub AR. On the cut polytope. *Math Program* 1986;36:157–173.
42. De Simone C, Diehl M, Jünger M, *et al.* Exact ground states of two-dimensional $\pm J$ Ising spin glasses. *J Stat Phys* 1996;84:1363–1371.

43. Liers F, Jünger M, Reinelt G, *et al.* Computing exact ground states of hard Ising spin glass problems by branch-and-cut. In: Hartmann A, Rieger H, editors. *New optimization algorithms in physics*. Chichester: John Wiley; 2004. pp. 47–68.
44. Mitchell JE. Computational experience with an interior point cutting plane algorithm. *SIAM J Optim* 2000;10(4):1212–1227.
45. Edmonds J. Maximum matching and a polyhedron with 0, 1 vertices. *J Res Natl Bur Stand* 1965;69B:125–130.
46. Grötschel M, Holland O. Solving matching problems with linear programming. *Math Program* 1985;33:243–259.
47. Ji X, Mitchell JE. Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement. *Discrete Optim* 2007;4(1):87–102.
48. Méndez-Díaz I, Zabala P. A branch-and-cut algorithm for graph coloring. *Discrete Appl Math* 2006;154(5):826–847.
49. Ortega F, Wolsey LA. A branch-and-cut algorithm for the single-commodity, uncapacitated, fixed-charge network flow problem. *Networks* 2003;41(3):143–158.
50. Hadjar A, Marcotte O, Soumis F. A branch-and-cut algorithm for the multiple depot vehicle scheduling problem. *Oper Res* 2006;54(1):130–149.
51. Lysgaard J, Letchford AN, Eglese RW. A new branch-and-cut algorithm for the capacitated vehicle routing problem. *Math Program* 2004;100(2):423–445.
52. Naddef D, Rinaldi G. Branch-and-cut algorithms for the capacitated VRP. In: Toth P, Vigo D, editors. *The vehicle routing problem*. Chapter 3, Number 9 in *Monographs on Discrete Mathematics and Applications*. Philadelphia: SIAM; 2002. pp. 53–84.
53. Labbé M, Yaman H, Gourdín E. A branch and cut algorithm for hub location problems with single assignment. *Math Program* 2005;102(2):371–405.
54. Grötschel M, Lovász L, Schrijver A. *Geometric algorithms and combinatorial optimization*. Berlin, Germany: Springer; 1988.
55. Xu Y, Ralphs TK, Ladányi L, *et al.* Computational experience with a software framework for parallel integer programming. *INFORMS J Comput* 2009;21(3):383–397.
56. Jünger M, Thienel S. Introduction to ABACUS—A Branch-And-CUT System. *Oper Res Lett* 1998;22:83–95.
57. Nemhauser GL, Savelsbergh MWP, Sigismondi GC. MINTO, a Mixed INTEger Optimizer. *Oper Res Lett* 1994;15:47–58.
58. Margot F. Exploiting orbits in symmetric ILP. *Math Program* 2003;98(1-3):3–21.
59. Ostrowski J, Linderoth J, Rossi F, *et al.* Orbital branching. *Math Program*; DOI: 10.107/s1007-009-0273-x.
60. Lübbecke ME, Desrosiers J. Selected topics in column generation. *Oper Res* 2005;53(6):1007–1023.
61. Fischetti M, Lodi A, Salvagnin D. Just MIP it. *Ann Inf Syst* 2009;10:39–70.
62. Fischetti M, Lodi A. Repairing MIP infeasibility through local branching. *Comput Oper Res* 2008;35:1436–1445.
63. Mitchell JE. Cutting plane methods and subgradient methods. Chapter 2, In: Oskoorouchi M, editor. *TutORials in Operations Research*. Hanover, MD: INFORMS; 2009. pp. 34–61.
64. Mitchell JE, Borchers B. Solving linear ordering problems with a combined interior point/simplex cutting plane algorithm, Chapter 14. In: Frenk HL, *et al.*, editors. *High Performance Optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000. pp. 349–366.
65. Fischer I, Gruber G, Rendl F, *et al.* Computational experience with a bundle approach for semidefinite cutting plane relaxations of max-cut and equipartition. *Math Program* 2006;105(2-3):451–469.
66. Helmberg C, Rendl F. Solving quadratic (0,1)-problems by semidefinite programs and cutting planes. *Math Program* 1998;82:291–315.
67. Rendl F, Rinaldi G, Wiegele A. Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. *Math Program* 2010;121(2):307–335.
68. Bonami P, Biegler LT, Conn AR, *et al.* An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optim* 2008;5(2):186–204.
69. Tawarmalani M, Sahinidis N. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*. Dordrecht, The Netherlands: Kluwer; 2002.
70. Tawarmalani M, Sahinidis N. A polyhedral branch-and-cut approach to global optimization. *Math Program* 2005;103(2):225–249.
71. Vandenbussche D, Nemhauser GL. A branch-and-cut algorithm for nonconvex quadratic programs with box constraints. *Math Program* 2005;102(3):559–575.

BRANCH-AND-BOUND ALGORITHMS

KIAVASH KIANFAR
Department of Industrial and Systems
Engineering, Texas A&M University,
College Station, Texas

For most discrete optimization problems, complete (or explicit) enumeration of solution space to find the optimal solution is out of question because even for small problem sizes the number of solution points in the feasible region is extremely large (e.g., even if a single point can be enumerated in 10^{-10} s, enumerating all points in a relatively small problem with only 75 binary variables will take about 120,000 years!). *Branch-and-bound* is a general purpose approach to *implicitly* enumerate the feasible region and was first introduced by Land and Doig [1] for solving integer programming problems. It works based on a few simple principles to avoid enumerating every solution point explicitly.

Although branch-and-bound is often discussed in the context of integer programming, it is actually a general approach that can also be applied to solve many combinatorial optimization problems even when they are not formulated as integer programs. For example, there are efficient branch-and-bound algorithms specifically designed for job shop scheduling or quadratic assignment problems, which are based on the combinatorial properties of these problems and do not use their integer programming formulations. Nevertheless, it is true that almost all integer programming solvers use branch-and-bound to solve IP problems and, therefore, application of branch-and-bound in the context of IP is of special importance. In this article we present the main concepts of branch-and-bound in a general setting and then discuss some problem-specific details in the context of integer programming. Many of these details can be extended to more general settings other than IP depending on the problem that is being solved.

BRANCH-AND-BOUND BASIC IDEAS

In this section we explain the main concepts of branch-and-bound as a general discrete optimization approach. Many of branch-and-bound concepts discussed here are based on Nemhauser and Wolsey [2] and Wolsey [3] the reader can refer to these resources for further reading. Consider a general combinatorial optimization problem like

$$z = \max \{f(x) : x \in S\}.$$

Branch-and-bound is a divide and conquer strategy, which decomposes the problem to subproblems over a *tree* structure, which is referred to as *branch-and-bound tree*. The decomposition works based on a simple idea: If S is decomposed into S^1 and S^2 such that $S = S^1 \cup S^2$, and we define subproblems $z^k = \max\{f(x) : x \in S^k\}$ for $k = 1, 2$, then $z = \max_k z^k$. Each subproblem represents a *node* on the tree. Fig. 1 shows a schematic of a branch-and-bound tree. The main problem with feasible region S (for simplicity we call it problem S) is at the *root node*, and is then divided into two subproblems (with feasible regions) S^1 and S^2 , where we have $S^1 \cup S^2 = S$. The process of dividing a node subproblem into smaller subproblems is called *branching* and subproblems S^1 and S^2 are called *branches* created at node S . In Fig. 1 subproblem S^1 is further branched into smaller subproblems and so on. The branching does not necessarily have to be two-way, and multiway branching is also possible.

Observe that branching indefinitely will only result in explicit enumeration of the feasible region of S . Therefore to avoid explicit enumeration, in branch-and-bound whenever possible a branch is pruned (or fathomed), meaning that its subproblem is not divided anymore. In other words, the feasible region of the node subproblem is implicitly enumerated without branching any deeper. But when can we prune a branch? The main idea is to use bounds on the objective value of subproblems intelligently to prune branches. That is why the method is called *branch-and-bound*.

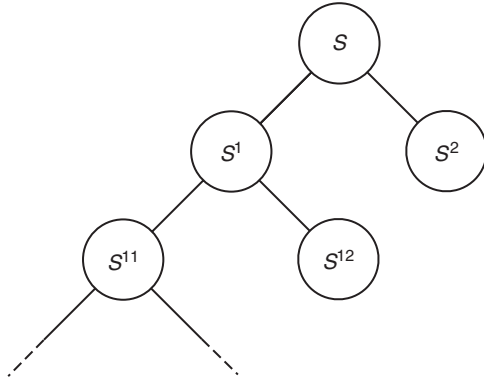


Figure 1. Subproblems in a branch-and-bound tree.

Fig. 2 shows the flow chart of the branch-and-bound algorithm in a general context. For a maximization problem, a lower bound $\underline{z}$ is updated throughout the algorithm and is used in pruning the nodes. The updating of the lower bound is normally triggered by finding an optimal solution with a better objective value at a node subproblem. Clearly, an optimal solution at a node problem gives a lower bound on z . Additionally, the branch-and-bound algorithm must have a mechanism to calculate an upper bound for the objective value of a node subproblem. For example, in IP solving the linear programming (LP) relaxation of the subproblem gives an upper bound on the IP objective value. As observed in Fig. 2, in general, a node is pruned if one of the following cases happens:

Pruning by infeasibility. If the feasible region of a node subproblem is empty, the node is naturally pruned.

Pruning by bound. If the upper bound calculated for a node subproblem is not greater than the lower bound on z , that is, $\bar{z}^i \leq \underline{z}$, then the node is pruned because there is no point in searching the feasible region of that node when we know that the best objective value we can obtain is not better than a solution we already know.

Pruning by optimality. When it is possible to find the optimal solution to a node subproblem S^i , then the node is

pruned and its solution is stored as the incumbent if its objective value is better than the best we know so far. The lower bound $\underline{z}$ is also updated in this case.

If a node cannot be pruned based on any of the above conditions then new branches are created to decompose the node subproblem into smaller problems. The algorithm stops when all nodes are pruned, that is, there is no active node remaining. The optimal solution will be the incumbent solution with objective value $\underline{z}$. The finiteness of the number of steps in a branch-and-bound algorithm has been theoretically studied and proved within a general formulation in works such as Refs 4 and 5.

This provides a complete high level view of branch-and-bound. In a more technical level there are many detailed questions that must be addressed in implementing the algorithm. The answers to many of these questions depend upon the particular problem being solved. Some of these questions are as follows [3]:

- How should the upper bounds be calculated?
- What is the appropriate balance between the time spent to find an upper bound and the strength of the upper bound?
- How should we do the branching, that is, how should a node subproblem be decomposed into smaller problems?
- How many branches should be created at each branching? Two or more?
- Should the branching rule be an a priori rule or should it adapt as the algorithm proceeds?
- How should we choose the next active node to consider?

As mentioned, almost all solvers for IP problems use branch-and-bound. In the next section, we discuss the LP-based branch-and-bound algorithm for solving an IP problem specifically and address the questions above in the context of solving the IP problem.

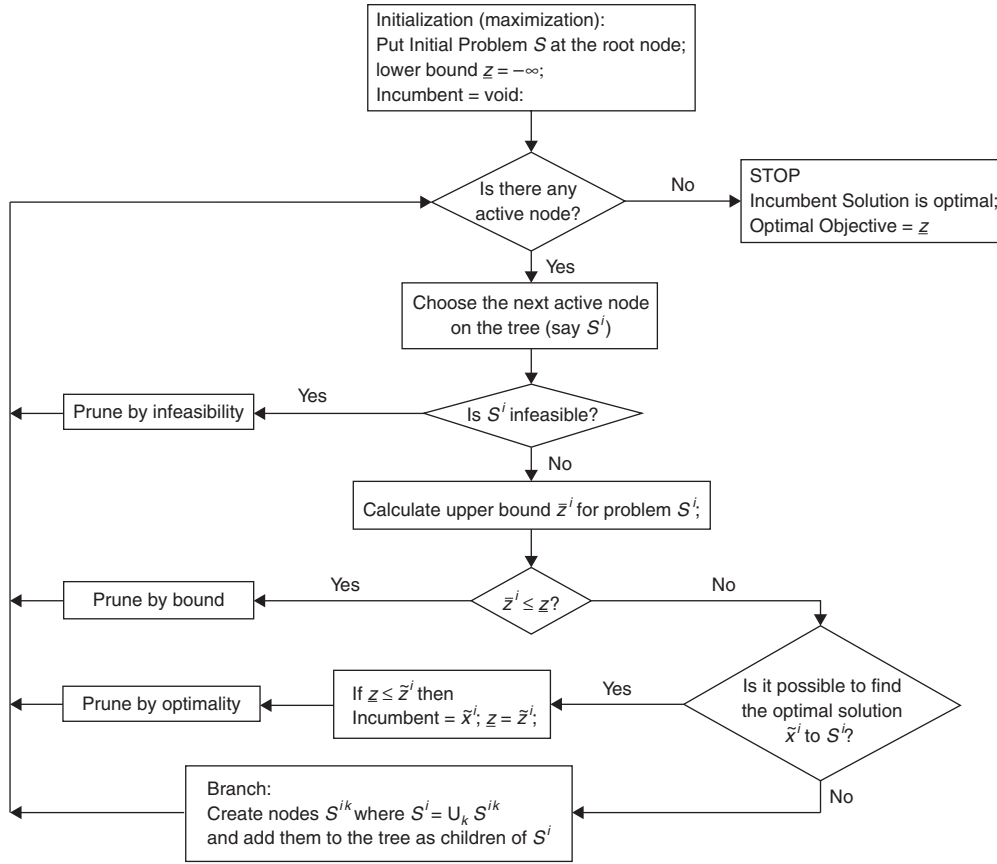


Figure 2. General branch-and-bound algorithm.

LP-BASED BRANCH-AND-BOUND TO SOLVE INTEGER PROGRAMMING PROBLEMS

Consider an integer programming problem like

$$z = \max \{cx : x \in S\},$$

where $S = \{x : x \in Z^n \cap P\}$ and P is a polyhedron. As before we refer to this problem as problem S . For simplicity we talk about pure integer programming problems but what follows can be easily extended to mixed integer programming problems too. Figure 3 shows the flow chart of the branch-and-bound algorithm for solving the IP problem S . This flow chart is a special case of the general flow chart of Fig. 2 customized in order to solve the IP problem.

Illustrative Example. Consider the IP problem S :

$$\begin{aligned} \max z &= 3x_1 + 2x_2 \\ 3x_1 + 4x_2 &\leq 12 \\ 2x_1 + x_2 &\leq 5 \\ x_1, x_2 &\geq 0 \text{ and integer.} \end{aligned}$$

The branch-and-bound tree for solving this problem is shown in Fig. 4.

Observe in Fig. 4 how the calculation of bound in the right branch helps us to prune the left branch and therefore saves us the effort of further enumerating the solutions in that branch.

Now let us address in more detail some of the questions we posed at the end of the section titled “Branch and Bound Basic

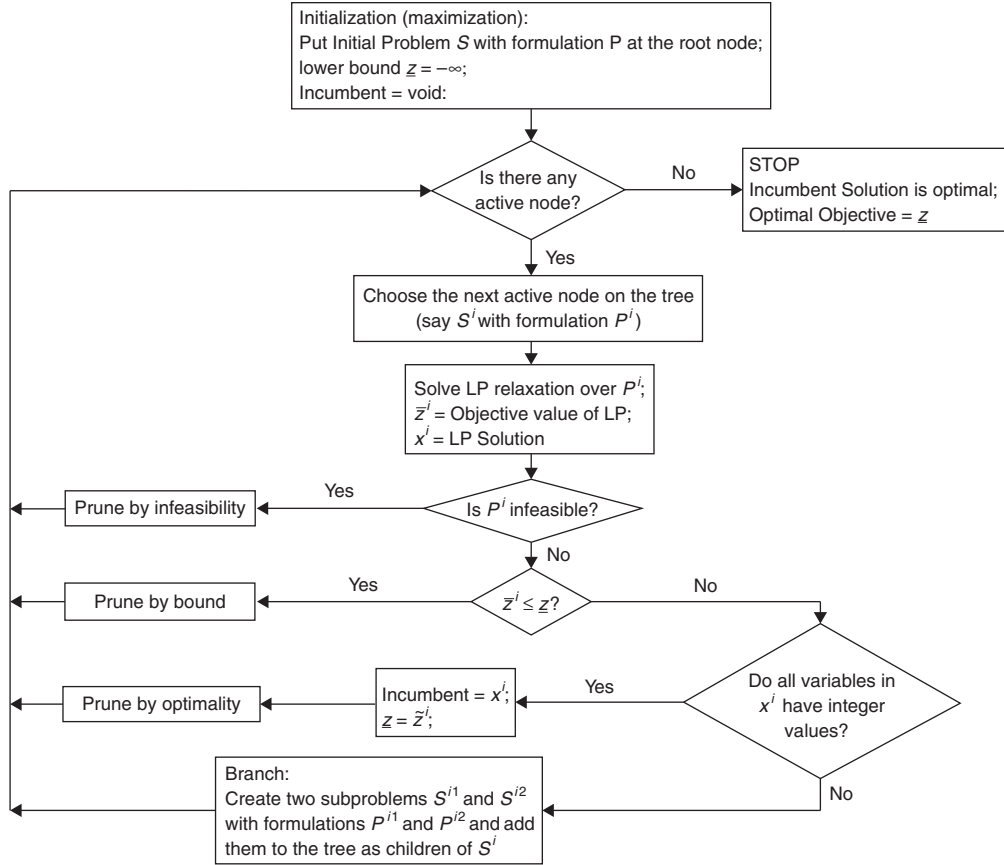


Figure 3. Branch-and-bound flow chart for solving an IP problem.

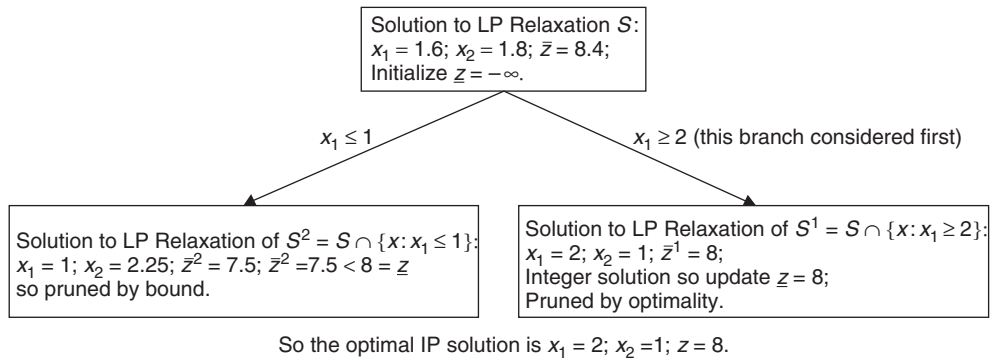


Figure 4. Branch-and-bound tree for the example problem.

Ideas" with respect to branch-and-bound for solving the IP problem.

Bounding

Solving the LP relaxation of any IP subproblem S^i gives an upper bound on its objective value. This is the bound that is most commonly used in practice. The LP is usually solved using simplex-based algorithms. On very large models interior point methods may be best for solution of the first LP [3]. A desirable feature of LP relaxation with simplex is that an optimal or near-optimal basis of the problem can be stored so that the LP relaxation in subsequent nodes can be reoptimized rapidly.

Branching

An important question is how to branch, meaning how to split a subproblem into smaller subproblems. Here we review the most common methods.

Single Variable Branching. The simplest idea to split the feasible region of a subproblem S^i with formulation P^i is to pick an integer variable with fractional value in the LP relaxation optimal solution, say x_j with value $\bar{x}_j$, and create branches by adding simple linear constraints to formulation P^i as follows:

$$\begin{aligned} S^{i1} &= S^i \cap \{x : x_j \leq \lfloor \bar{x}_j \rfloor\} \\ S^{i2} &= S^i \cap \{x : x_j \geq \lceil \bar{x}_j \rceil\}. \end{aligned}$$

S^{i1} is usually referred to as the *left (down) branch* and S^{i2} is referred to as the *right (up) branch*. This is a desirable choice because clearly we have $S^i = S^{i1} \cup S^{i2}$ and $S^{i1} \cap S^{i2} = \emptyset$, and furthermore, the current LP solution is not feasible for any of S^{i1} and S^{i2} so in the absence of multiple optima for LP the upper bound would strictly decrease in each of these branches. This branching scheme was first introduced in Ref. 6.

A generalization of the above branching scheme is using branches such as $S^{i1} = S^i \cap \{x : dx \leq d_0\}$ and $S^{i2} = S^i \cap \{x : dx \geq d_0 + 1\}$ in which d and d_0 are integer. However, in practice, only the simplest form which is the

above single variable branching is used in all solvers.

With the single variable branching strategy, an important question is which variable should be chosen out of all variables with fractional value? A recent comprehensive study on this issue is done in Ref. 7; also see Refs 2, 3, 8, 9. An old and common rule is the *most fractional variable*. If C is the set of all integer variables with fractional LP relaxation value then the chosen variable is

$$\operatorname{argmax}_{j \in C} \min \{f_j, 1 - f_j\},$$

where $f_j = \bar{x}_j - \lfloor \bar{x}_j \rfloor$. When branching is based on this rule, it is recommended that the $x_j \leq \lfloor \bar{x}_j \rfloor$ branch is considered first only if $f_j \leq 1 - f_j$. Otherwise the $x_j \geq \lceil \bar{x}_j \rceil$ should be selected first. Example in Fig. 4 obeys this rule. Accordingly the most fractional variable, that is, x_1 , is chosen as the branching variable and then the right branch is selected after the root node because $f_1 = x_1 - \lfloor x_1 \rfloor = 0.6 > 0.4 = 1 - (x_1 - \lfloor x_1 \rfloor) = 1 - f_1$.

The most fractional rule is simple to implement but in Ref. 7, it is shown computationally that this rule is not better than just any random variable selection rule. More effective strategies have been proposed and studied over the years, which are more sophisticated. The method of *pseudocost branching* goes back to [10] and works based on calculating a pseudocost by keeping a history of success (change in LP relaxation value) of the left and right branching performed on each variable. The branching variable with the highest pseudocost is chosen each time. Different functions have been proposed for the pseudocost such as maximum of sum of left and right *degradations* (decrease in upper bound) [11], maximum of the smaller of the left and right degradations [10]; also see Ref. 8.

Strong branching [8,12] is in a sense an extreme effort to find the best branching variable. In its full version, at any node one tentatively branches on each candidate variable with fractional value and solves the LP relaxation for right and left branches for each variable either to optimality or for a specified number of dual simplex iterations. The degradation in bound for left and right

branches are calculated and the variable is picked on the basis of the function of these degradations [3,8]. In addition to a number of dual simplex method iterations, the computational effort of strong branching can also be limited by defining a much smaller set of candidate branching variables.

An effective variable selection strategy was proposed in Ref. 7 called *reliability branching*. This method integrates strong and pseudocost branching. A threshold is defined on the number of previous branchings involving a variable. Below this threshold the pseudocost is considered unreliable and strong branching is used but above the threshold the pseudocost method is used.

We also note that many solvers provide the user with the option of assigning *user-specified priorities* to the integer variables. When branching, the solver picks the variable with a fractional value that has the highest assigned priority. This is especially useful when the user has some insight regarding the importance of variables based on what they represent in the application underlying the model [3].

GUB Branching. Many IP models contain the so-called *Generalized Upper Bound* (GUB) or *Special Ordered Set* (SOS) constraint of the form

$$\sum_{j=1}^n x_j = 1$$

with $x_j \in \{0, 1\}$ for $j = 1, \dots, n$. If the single variable branching on one of the variables $x_j, j = 1, \dots, n$, is performed then the two branches will be $S^{i1} = S^i \cap \{x : x_j = 0\}$ and $S^{i2} = S^i \cap \{x : x_j = 1\}$. Because of the GUB constraint, $\{x : x_j = 0\}$ will leave $n - 1$ possibilities $\{x : x_i = 1\}_{i \neq j}$ while $\{x : x_j = 1\}$ leaves only one possibility. So S^{i1} is usually much larger than S^{i2} and the tree is unbalanced. A more balanced split is desired and GUB branching proposed in Ref. 13 is a way to get a balanced tree which works as follows: The user provides a special order of variables in the GUB set, say $j_1, \dots, j_n$. Then the two

branches will be as follows:

$$S^{i1} = S^i \cap \{x : x_{j_i} = 0, i = 1, \dots, r\}$$

$$S^{i2} = S^i \cap \{x : x_{j_i} = 0, i = r + 1, \dots, n\},$$

where $r = \min \{t : \sum_{i=1}^t \bar{x}_{j_i} \geq \frac{1}{2}\}$. The number of nodes is usually significantly reduced with this branching scheme compared to the single variable branching.

Node Selection

Having a list of active (unpruned) nodes, the question is which node should be examined next. There are two categories of rules for this purpose: *static rules* that determine in advance the order of node selection and *adaptive rules* which use the information about the status of active nodes to choose a node. For a complete review of different node selection strategies see Ref. 8. Among static rules the *depth-first* and the *best-bound* (or *best-first*) are the two well-known extremes.

Depth-First Node Selection. Depth-first rule also known as *last in, first out* (LIFO) is as follows: if the current node is not pruned the next node is one of its children; and if it is pruned the next node is found by *backtracking* which means the next node is the child of the first node on the path from the current node to the root node with an unconsidered child node. Obviously, this is a completely a priori rule if a rule is specified to select between left and right children of a node. This rule has known advantages:

1. For pruning the tree a good lower bound is needed. The depth-first method descends quickly in the tree to find a first feasible solution which gives a lower bound that hopefully causes the pruning of many future nodes.
2. The depth-first method tends to minimize the memory requirements for storing the tree at any given time during the algorithm.
3. Passing from a node to its immediate child has the advantage that the LP relaxation can be easily resolved by adding just an additional constraint.

However, the depth-first search can result in an extremely large search tree. This is the result of the fact that we may need to consider many nodes that would have been fathomed if we had a better lower bound.

Best-Bound Node Selection. In this rule the next node is the active node with the best (largest) upper bound. With this rule one would never branch a node whose upper bound is less than the optimal value. As a result this rule minimizes the number of the nodes considered. However, the memory requirements for this method may become prohibitive if good lower bounds are not found early leading to relatively little pruning. In terms of reoptimizing the LP relaxations this method is also at a disadvantage because one LP problem has little relation to the next one.

Adaptive Node Selection. Adaptive methods make intelligent use of information about the nodes to select the next node. The *estimate-based methods* attempt to select nodes that may lead to improved integer feasible solutions. The *best projection* criterion [14] and the *best estimate* criterion found in Refs 10 and 15 are among these methods.

Many adaptive methods are *two-phase* methods that essentially use a hybrid of depth-first and best-bound searches. At the beginning the depth-first strategy is used to find a feasible solution and then the best-bound method is used [16]. Variations of two-phase methods using estimate-based approaches are also proposed [15,17]. Some other suggested rules use an estimation of the optimal IP solution to avoid considering *superfluous* nodes, that is, the nodes in which $\bar{z}^i < z$. The tree is searched in a depth-first fashion as long as $\bar{z}^i$ is greater than the estimate. After that a different criterion such as best-bound is used [10,11].

EXTENSIONS OF BRANCH-AND-BOUND

In solving integer programming *problems*, pure branch-and-bound is seldom used. In most cases cutting planes are added to the

root problem or the node subproblems to tighten the feasible region of the LP relaxation and hence obtain better bounds faster. A branch-and-bound algorithm in which cutting planes are used is known under the general name of *branch-and-cut* [2,3,18].

In branch-and-cut at any node, after optimizing the LP relaxation, a *separation* problem is solved to find valid inequalities for feasible integer solutions, which are violated by the LP relaxation solution. These valid inequalities are then added to the problem and the problem is reoptimized to improve the LP relaxation bound. Branching happens when no further valid inequalities can be found. Of course the amount of effort spent to find valid inequalities is one of the parameters of the algorithm that should be decided. We refer the reader to the article titled **Branch and Cut** in this encyclopedia for further information.

Another extension of branch-and-bound is *branch-and-price* [2,3,19]. When the number of variables in an integer program is huge, a solution method is using *column generation* within the branch-and-bound framework. More specifically, implicit pricing of nonbasic variables is used to generate new columns or to prove LP optimality at a node of the branch-and-bound tree. Branching happens when no columns price out to enter the basis and the LP solution does not satisfy integrality constraints. We note that if cutting planes are also used in branch-and-price the algorithm is called *branch-cut-price*. We refer the reader to the article titled **Branch-Price-and-Cut Algorithms** in this encyclopedia for further information.

PARALLEL BRANCH-AND-BOUND

The divide and conquer nature of branch-and-bound makes it a suitable framework for attacking huge problems using today's parallel computing capabilities. For a survey of parallel branch-and-bound algorithms refer to Ref. 20. There are three types of parallelism that can be implemented for a branch-and-bound algorithm: Type 1 is parallel execution of operations on generated subproblems, for example, parallel bounding

operations for each subproblem to accelerate execution. Type 2 consists of building the tree in parallel by performing operations on several subproblems simultaneously. Type 3 is building several trees in parallel. In each tree some operation such as branching, bounding or selection is performed differently but the trees share their information and use the best bounds among themselves. For further reading refer to Ref. 20.

REFERENCES

1. Land AH, Doig AG. An automated method of solving discrete programming problems. *Econometrica* 1960;28(2):497–520.
2. Nemhauser GL, Wolsey LA. Integer and combinatorial optimization. New York: Wiley-Interscience; 1988.
3. Wolsey LA. Integer programming. New York: Wiley; 1998.
4. Bertier P, Roy B. Procédure de résolution pour une classe de problèmes pouvant avoir un caractère combinatoire. *Cah Cent Étud Rech Oper* 1964;6:202–208.
5. Balas E. A note on the Branch-and-bound Principle. *Oper Res* 1968;16(2):442–445.
6. Dakin RJ. A tree search algorithm for mixed integer programming. *Comput J* 1965;8: 250–255.
7. Achterberg T, Koch T, Martin A. Branching rules revisited. *Oper Res Lett* 2005;33:42–54.
8. Linderoth JT, Savelsbergh MWP. A computational study of search strategies for mixed integer programming. *INFORMS J Comput* 1999;11:173–187.
9. Lodi A. Mixed integer programming computation. In: Junger M, Liebling T, Naddef D, *et al.*, editors. 50 years of integer programming 1958–2008. Berlin: Springer; 2010. pp. 619–645.
10. Benichou M, Gauthier JM, Girodet P, *et al.* Experiments in mixed-integer programming. *Math Program* 1971;1:76–94.
11. Gauthier JM, Ribière G. Experiments in mixed-integer linear programming using pseudocosts. *Math Program* 1977;12:26–47.
12. Applegate D, Bixby RE, Chvátal V, *et al.* The traveling salesman problem: a computational study. Princeton (NJ): Princeton University Press; 2007.
13. Beale EML, Tomlin JA. Special facilities in a generalized mathematical programming system for nonconvex problems using ordered sets of variables. In: Lawrence J, editor. *Proceedings of the 5th Annual Conference on Operational Research*. London: Tavistock Publications; 1970. pp. 447–457.
14. Mitra G. Investigation of some branch-and-bound strategies for the solution of mixed integer linear programs. *Math Program* 1973;4:155–170.
15. Forrest JJH, Hirst JPH, Tomlin JA. Practical solution of large scale mixed integer programming problems with UMPIRE. *Manage Sci* 1974;20:736–773.
16. Eckstein J. Parallel branch-and-bound algorithms for general mixed integer programming on the CM-5. *SIAM J Optim* 1994;4:794–814.
17. Beale EML. Branch-and-bound methods for mathematical programming systems. In: Hammer PL, Johnson EL, Korte BH, editors. *Discrete optimization II*. Amsterdam: North Holland Publishing Co.; 1979.
18. Caprara A, Fischetti M. Branch-and-cut algorithms. In: Dell’Amico M, Maffioli F, Martello S, editors. *Annotated bibliographies in combinatorial optimization*. New York: Wiley; 1997. pp. 45–63.
19. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. *Oper Res* 1998;46(3):316–329.
20. Gendron B, Crainic TG. Parallel branch-and-bound algorithms: survey and synthesis. *Oper Res* 1994;42(6):1042–1066.

BRANCH-PRICE-AND-CUT ALGORITHMS

JACQUES DESROSIERS
HEC Montréal and GERAD,
Montréal, Canada

MARCO E. LÜBBECKE
Chair of Operations Research,
RWTH Aachen University,
Aachen, Germany

Decompositions and reformulations of mixed integer programs are classical approaches to obtaining stronger relaxations and reduce symmetry. These often entail the dynamic addition of variables (columns) and/or constraints (cutting planes) to the model. When the linear relaxation in each node of a branch-and-bound tree is solved by column generation, one speaks of branch-and-price. Optionally, as in standard branch-and-bound, cutting planes can be added in order to strengthen the relaxation, and this is called *branch-price-and-cut*. Now, having understood and familiarized oneself with the three concepts, branch-and-bound (see **Branch-and-Bound Algorithms**), column generation (see **Column Generation**), and cutting planes (see the section titled “Automatic Convexification” in this encyclopedia); one may think that the above explanation is the end of the story. Actually, this understanding is precisely where the story begins. Strengthening the relaxation by means of cutting planes and performing branching on fractional variables both interfere with column generation and may entirely ruin the mentioned advantages of a decomposition when done naively.

In early years, one attempted to circumvent these complications in an *ad hoc* fashion [1] but over time a generic theoretical understanding developed. The state-of-the-art relies on the relation between the original problem and its extended reformulation, as first used in Desrosiers *et al.* [2].

There are very successful applications of branch-and-price in industry (see Desrosiers

and Lübbecke [3], and also the section titled “Vehicle Routing and Scheduling” in this encyclopedia) and also to generic combinatorial optimization problems like bin packing and the cutting stock problem [4], graph coloring [5], machine scheduling [6], the p -median problem [7], the generalized assignment problem [8], and many others. The method today is an indispensable part of the integer programming toolbox.

COLUMN GENERATION

Consider the following *integer master problem*

$$\begin{aligned} \min \quad & \sum_{j \in J} c_j \lambda_j \\ \text{subject to} \quad & \sum_{j \in J} \mathbf{a}_j \lambda_j \leq \mathbf{b} \\ & \lambda \in \mathbb{Z}_+^{|J|}. \end{aligned} \quad (1)$$

In many applications, $|J|$ is huge (but always finite) and we solve the linear relaxation of problem (1), called the *master problem*, by column generation as follows (see **Column Generation**). The *restricted master problem* (RMP) contains only a subset $J' \subseteq J$ of variables, initially possibly none. In each iteration, (i) we obtain λ^* and π^* , the primal and dual optimal solutions to the RMP, respectively. Then, (ii) the following pricing problem (subproblem) is to be solved

$$v := \min_{\mathbf{x} \in X} \{c(\mathbf{x}) - \pi^* a(\mathbf{x})\}, \quad (2)$$

where $c_j = c(\mathbf{x}_j)$ and $\mathbf{a}_j = a(\mathbf{x}_j)$ reflect that each column $j \in J$ is associated with an element $\mathbf{x}_j \in X$ from a domain X over which we can optimize, often a set of combinatorial objects like paths or other subgraphs (so, usually, the $\mathbf{x}_j$ bears much more information than just the column $a(\mathbf{x}_j)$). When $v < 0$ the variable λ_j and its coefficient column $(c_j, \mathbf{a}_j)$ corresponding to a minimizer $\mathbf{x}_j$ are added to the RMP, and the process is

iterated. Otherwise, $v \geq 0$ proves that there is no such improving variable, and the current λ^* is an optimal solution to the master problem.

DECOMPOSITIONS AND REFORMULATIONS

In general, when solving mathematical programs involving integer variables, a *good* model is of utmost importance, refer to Nemhauser and Wolsey [9] for some general considerations (see also the section titled “Models and Algorithms” in this encyclopedia). After all, the whole integer programming machinery is about better describing the convex hull of feasible integer solutions: tighter relaxations from a different modeling choice and cutting planes fulfill this purpose. Also, branch-and-price is motivated by the perspective for better dual bounds and reduced problem symmetry. In the following we describe the groundwork.

Dantzig–Wolfe Decompositions for Integer Programs

The Dantzig–Wolfe decomposition principle in linear programming [10] was devised to exploit sparsity and special structure in linear programs (see also **Dantzig–Wolfe Decomposition**). However, it reveals its true strength only when adapted to integer programs. We assume that optimizing over the mixed integer set $X = \{\mathbf{x} \in \mathbb{Z}_+^n \times \mathbb{Q}_+^q \mid D\mathbf{x} \leq \mathbf{d}\}$ is “relatively easy,” which however, is overshadowed by additional *complicated* constraints, that is,

$$\begin{aligned} \min \quad & \mathbf{c}\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \in X \end{aligned} \quad (3)$$

is rather hard to solve when the “hidden” structure X is not exploited. The mixed integer program (3) is called the *original* problem in this context and we call $\mathbf{x}$ the *original* variables. As we will see, a decomposition will lead to a master and a pricing problem as above. It was noted already by Geoffrion [11] in the context of Lagrangean relaxation (see also **Lagrangian Optimization for LP**)

that decomposing problem (3) by solving over X as a mixed integer program (and penalizing the violation of $A\mathbf{x} \leq \mathbf{b}$ in the objective function) may yield a stronger dual bound than the standard linear relaxation when the convex hull $\text{conv}(X)$ is not an integral polyhedron. In this article, X is assumed to be a pure integer set, that is, $q = 0$; we only point to the mixed integer generalization when needed.

Convexification. The classical decomposition approach builds on the representation theorems by Minkowski and Weyl [12] and *convexifies* X , hence the name. Each $\mathbf{x} \in X$ can be expressed as a convex combination of finitely many extreme points $\{\mathbf{x}_p\}_{p \in P}$ plus a nonnegative combination of finitely many extreme rays $\{\mathbf{x}_r\}_{r \in R}$ of $\text{conv}(X)$, that is,

$$\begin{aligned} \mathbf{x} &= \sum_{p \in P} \mathbf{x}_p \lambda_p + \sum_{r \in R} \mathbf{x}_r \lambda_r, \\ \sum_{p \in P} \lambda_p &= 1, \quad \lambda \in \mathbb{Q}_+^{|P|+|R|}. \end{aligned} \quad (4)$$

Substituting for $\mathbf{x}$ in the original problem (3) and applying the linear transformations $c_j = \mathbf{c}\mathbf{x}_j$ and $\mathbf{a}_j = A\mathbf{x}_j, j \in P \cup R$ one obtains an *extended formulation* equivalent to problem (3), which is an *integer master problem* as Equation (1)

$$\begin{aligned} \min \quad & \sum_{p \in P} c_p \lambda_p + \sum_{r \in R} c_r \lambda_r \\ \text{subject to} \quad & \sum_{p \in P} \mathbf{a}_p \lambda_p + \sum_{r \in R} \mathbf{a}_r \lambda_r \leq \mathbf{b} \\ & \sum_{p \in P} \lambda_p = 1 \\ & \lambda \geq \mathbf{0} \\ & \mathbf{x} = \sum_{p \in P} \mathbf{x}_p \lambda_p + \sum_{r \in R} \mathbf{x}_r \lambda_r \\ & \mathbf{x} \in \mathbb{Z}_+^n. \end{aligned} \quad (5)$$

The constraints involving only λ variables are called *coupling constraints* and *convexity constraints*, respectively. It is important to note that integrality is still imposed on the original $\mathbf{x}$ variables. Owing to the large

cardinality of $P \cup R$ the LP relaxation of Equation (5) is solved by column generation (see **Column Generation** and the section titled “Column Generation”), where the constraints linking $\mathbf{x}$ and λ variables can be dropped. Thus, only the dual variables π and π_0 remain relevant where π_0 corresponds to the convexity constraint. The pricing problem is the integer program

$$\min \{ \mathbf{c}\mathbf{x} - \pi\mathbf{A}\mathbf{x} - \pi_0 \mid \mathbf{x} \in X \}, \quad (6)$$

for which, ideally, a tailored combinatorial algorithm is available; and if not, we need to solve it with some standard IP solver which may be very time consuming.

Discretization. In contrast to convexification where $\text{conv}(X)$ is reformulated, *discretization* is a reformulation of X itself. It enables us to require integrality on the master variables, which is not valid in Equation (5). Vanderbeck [13] introduced the concept since “it allows for the development of a unifying and complete theoretical framework to deal with all relevant issues that arise in the implementation of a branch-and-price algorithm.” As we will see, he was in particular thinking of cutting planes and branching. We need the fact [9] that every integer $\mathbf{x} \in X$ can be written as an *integral* combination

$$\begin{aligned} \mathbf{x} &= \sum_{p \in P} \mathbf{x}_p \lambda_p + \sum_{r \in R} \mathbf{x}_r \lambda_r, \quad \sum_{p \in P} \lambda_p = 1, \\ \lambda &\in \mathbb{Z}_+^{|P|+|R|} \end{aligned} \quad (7)$$

with finite sets of *integer* points $\{\mathbf{x}_p\}_{p \in P} \subseteq X$ and *integer* rays $\{\mathbf{x}_r\}_{r \in R}$ of X . Note that we slightly abused the notation here since the set P of *generators* is usually not identical to the corresponding extreme points of the convexification approach, and rays in R are scaled to be integer. Substitution for $\mathbf{x}$ in Equation (3) yields an *integer* master problem

$$\min \sum_{p \in P} c_p \lambda_p + \sum_{r \in R} c_r \lambda_r$$

$$\begin{aligned} \text{subject to } & \sum_{p \in P} \mathbf{a}_p \lambda_p + \sum_{r \in R} \mathbf{a}_r \lambda_r \leq \mathbf{b} \\ & \sum_{p \in P} \lambda_p = 1 \\ & \lambda \in \mathbb{Z}_+^{|P|+|R|}, \end{aligned} \quad (8)$$

where integrality is now imposed on the master variables λ , and again $c_j = \mathbf{c}\mathbf{x}_j$ and $a_j = \mathbf{A}\mathbf{x}_j, j \in P \cup R$. Note that in fact $\lambda_p \in \{0, 1\}$ for $p \in P$. It is known [13] that the LP relaxation gives the same dual bound as problem (5). When solving problem (8) by column generation the pricing problem is the same as above, but it needs to be able to generate integer solutions in the interior of X . When X is bounded, Equation (8) is a linear integer program even when the original problem (3) has nonlinear cost function $c(\mathbf{x})$: Because of the convexity constraint, variables λ are binary, thus, $c(\mathbf{x}) = c(\sum_{p \in P} \mathbf{x}_p \lambda_p) = \sum_{p \in P} c_p \lambda_p$ turns into a linear objective function.

The discretization approach generalizes to a mixed integer set X in that integer variables are discretized and continuous variables are reformulated using the convexification approach [14]. For general integer variables this is not straight forward using the convexification approach, but in the important special case $X \subseteq [0, 1]^n$ of combinatorial optimization, convexification, and discretization coincide. Both approaches yield the same dual bound which equals that of Lagrangean relaxation (see **Relationship Among Benders, Dantzig-Wolfe, and Lagrangian Optimization** and Vanderbeck [15]).

Bordered Block-Diagonal Matrices

For many problems, $X = X_1 \times \dots \times X_K$ (possibly permuting variables), that is, X decomposes into $X_k = \{\mathbf{x}_k \in \mathbb{Z}_+^{n_k} \times \mathbb{Q}_+^{q_k} \mid D_k \mathbf{x}_k \leq \mathbf{d}_k\}$, $k = 1, \dots, K$, with all matrices and vectors of compatible dimensions and $\sum_k n_k = n$, $\sum_k q_k = q$. This is another way of saying that D can be brought into a block-diagonal form, so that the original problem (3) reads

$$\min \sum_k \mathbf{c}_k \mathbf{x}_k$$

$$\begin{aligned} \text{subject to } \sum_k A_k \mathbf{x}_k &\leq \mathbf{b} \\ \mathbf{x}_k &\in X_k, \quad k = 1, \dots, K. \end{aligned} \quad (9)$$

This is the classical situation for applying a Dantzig–Wolfe decomposition. Each $\mathbf{x}_k$ is expressed using Equation (4) or (7), with the introduction of λ_{kj} variables, $j \in P_k \cup R_k$, where $P = \bigcup_{k=1}^K P_k$ and $R = \bigcup_{k=1}^K R_k$. It is an important special case that some or all $(D_k, \mathbf{d}_k)$ are identical, for example, for bin packing, vertex coloring, or vehicle-routing problems. This implies a symmetry since “the same” solution can be expressed in many different ways by permuting the k indices. Symmetry is beautiful in many areas of mathematics, however, for an integer program it may be a major source of inefficiency in a branch-and-bound algorithm and should be avoided by all means. Typically, one aggregates (sums up) the λ_{kp} variables, substituting $v_p := \sum_k \lambda_{kp}$, and adding up the K convexity constraints. Extreme rays need no aggregation. Choosing a representative P_1 , we obtain the *aggregated* extended formulation

$$\begin{aligned} \min \quad & \sum_{p \in P_1} c_p v_p + \sum_{r \in R} c_r \lambda_r \\ \text{subject to } \quad & \sum_{p \in P_1} \mathbf{a}_p v_p + \sum_{r \in R} \mathbf{a}_r \lambda_r \leq \mathbf{b} \\ & \sum_{p \in P_1} v_p = K \quad (10) \\ & \mathbf{v} \in \mathbb{Z}_+^{|P_1|} \\ & \lambda \in \mathbb{Z}_+^{|R|}, \end{aligned}$$

here in the discretization version. This also condenses the original $\mathbf{x}_k$ variables into aggregated original variables $\mathbf{z} = \sum_k \mathbf{x}_k$.

Extended Reformulations

Lifting a mixed integer problem to a higher-dimensional space, obtaining a stronger formulation there, and projecting this back to the original variables’ space is a well-known concept in integer programming (see *Symmetry Handling in Mixed-Integer Programming, Disjunctive Programming*). Not only in the context

of branch-and-price it is helpful to know some basic ideas. The very recommendable survey [16] presents decomposition approaches in this context.

A polyhedron $Q = \{(\mathbf{x}, \lambda) \in \mathbb{R}^n \times \mathbb{R}^\ell \mid A\mathbf{x} + L\lambda \leq \mathbf{b}\}$ is called an *extended formulation* of a polyhedron $O \subseteq \mathbb{R}^n$ if Q can be projected to O , that is, if $O = \text{proj}_{\mathbf{x}}(Q)$, where $\text{proj}_{\mathbf{x}}(Q)$ denotes the projection of Q on the $\mathbf{x}$ variables, that is, $\text{proj}_{\mathbf{x}}(Q) = \{\mathbf{x} \in \mathbb{R}^n \mid \exists \lambda \in \mathbb{R}^\ell : (\mathbf{x}, \lambda) \in Q\}$. Polyhedron Q is an extended formulation of the integer set X if $X = \text{proj}_{\mathbf{x}}(Q) \cap \mathbb{Z}_+^n$. More generally, Q may itself be a mixed integer set, and we call Q an extended formulation of the mixed integer set X if $X = \text{proj}_{\mathbf{x}}(Q)$. The discretization approach provides an example. We also speak of a *problem* being an extended formulation, for example, we call the linear relaxation of the master problem (5) an extended formulation of the original problem (3).

As stated above, extended formulations are typically stronger than their original counterparts, and the special interest in Dantzig–Wolfe type extended reformulations Q of mixed integer sets X lies in the fact that they are *tight*, that is, $\text{proj}_{\mathbf{x}}(Q) = \text{conv}(X)$. That is, they are best in a well-defined sense at the expense that they may contain an exponential number of variables.

It is of main interest in our context that some experience and creativity with projections and extended formulations may help us reformulating a problem *before* a Dantzig–Wolfe decomposition is applied, be it implicit or explicit. This becomes important when formulating cutting planes (see the section titled “Cutting Planes”) and branching rules (the section titled “Branching”). In order to provide some intuition, consider a flow-based formulation for the minimum spanning tree problem in a graph $G = (V, E)$ where one sends one unit of flow from a designated root node to all other nodes, that is, $|V| - 1$ units in total. An integer variable x_{ij} represents the amount of flow on edge $(i, j) \in E$, and binary variables y_{ij} indicate whether $(i, j) \in E$ is in the spanning tree. Edge $(i, j) \in E$ may carry flow only when it is part of the tree, that is, $x_{ij} \leq (|V| - 1) \cdot y_{ij}$. This “big- M ” formulation can be much improved by extending the

formulation by introducing a separate flow commodity $k = 1, \dots, |V|$ for each node, that is, reformulate $x_{ij} = \sum_k x_{ij}^k$. The above mentioned constraint becomes $x_{ij}^k \leq y_{ij}$ which obviously better reflects the integrality of the y_{ij} variables, and thus gives stronger branching and cutting opportunities.

Reducing problem symmetry can be another reason to consider an extended formulation [17]. For problems like bin packing, graph coloring, and many others, binary variables x_{ij} assign items i to identical entities j . The symmetry in index j can be broken, for example, by binary variables z_{ij} which reflect the assignment of items i and j to the same entity but not items $i < k < j$. Cutting planes are reported to be effective at the cost of a more complicated pricing problem.

Reversing Dantzig–Wolfe Decomposition

Column generation can be applied without a prior decomposition. Examples are set covering or set partitioning problems like in the classical cutting stock problem or in vehicle and crew scheduling (see the section titled “Vehicle Routing and Scheduling” in this encyclopedia). One directly formulates an integer master problem (1) together with a pricing problem (2). Without a decomposition, there is no *original* problem. However, as we describe later, such an original problem can be very helpful in designing branching rules and cutting planes. Consequently, we would like to construct a corresponding original problem from Equations (1) and (2). That is, in a sense, one aims at *reversing* the Dantzig–Wolfe decomposition. Under a mild assumption, which is typically true, this can be done [18]. The idea exploits the fact that the pricing problem is formulated in original variables, for example, when the pricing problem constructs a path in a network, the decisions taken are typically whether an edge is included in the path or not. Conforming with our previous notation, we assume that we know an integer upper bound K on $\sum_{j \in J} \lambda_j$ in Equation (1) (the maximum number of vehicles, paper rolls, bins, etc.). When the zero column $\mathbf{a}_0 = \mathbf{0}$ has cost $c_0 = 0$, we can assume equality constraints in an original formulation. We duplicate the pricing

problem’s domain K times, that is, we set $X_k = X$, $k = 1, \dots, K$, and obtain a bordered block-diagonal form

$$\begin{aligned} \min \quad & \sum_{k=1}^K c(\mathbf{x}_k) \\ \text{subject to} \quad & \sum_{k=1}^K a(\mathbf{x}_k) = \mathbf{b} \\ & \mathbf{x}_k \in X_k \quad k = 1, \dots, K. \end{aligned} \quad (11)$$

Performing a Dantzig–Wolfe decomposition on Equation (11) one obtains a formulation equivalent to Equation (1) where each of the K pricing problems contributes, due to the convexity constraints, at most one unit to the master solution. Note that this introduces the symmetry of identical subproblems by design of the construction and one needs to aggregate the $\mathbf{x}_k$ variables. We have produced a projection to the variables implicitly given in the pricing problem. This is just one option; symmetry avoiding projections are preferable, for example, by explicitly producing different pricing problems.

CUTTING PLANES

Just as in standard branch-and-cut adding valid inequalities can strengthen the LP relaxation, yielding what is known as *branch-price-and-cut*. In the earlier days, this combination has been considered problematic, since the pricing problem must be aware of the coefficients in cutting planes as these need to be lifted when new variables are generated. With the introduction of the two viewpoints presented in this section the notion of *compatibility* became obsolete.

In the convexification approach (Eq. 5) integrality is required on the $\mathbf{x}$ variables just as in the original problem (3). Thus, it is only natural to formulate valid inequalities on these variables. On the other hand, remembering the motivation for extended formulations (see the section titled “Extended Reformulations”), this ignores the potential of the higher dimensional space, or in other words: We would like to formulate valid inequalities on the integer master variables

of the discretization reformulation (8) as well. Our presentation follows Desaulniers *et al.* [19], and several examples can be found therein.

Cutting Planes on the Original Variables

Assume that we know a set $F\mathbf{x} \leq \mathbf{f}$ of inequalities valid for the original problem (3), that is,

$$\begin{aligned} \min \quad & \mathbf{c}\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq \mathbf{b} \\ & F\mathbf{x} \leq \mathbf{f} \\ & \mathbf{x} \in X \end{aligned} \quad (12)$$

has the same integer feasible solutions. Via a Dantzig–Wolfe reformulation these inequalities directly transfer to the master problem, both in convexification and discretization

$$\sum_{p \in P} \mathbf{f}_p \lambda_p + \sum_{r \in R} \mathbf{f}_r \lambda_r \leq \mathbf{f} \quad (13)$$

with the linear transformations $\mathbf{f}_j = F\mathbf{x}_j$ for $j \in P \cup R$. The dual variables α of the additional inequalities (13) can be easily taken care of in the pricing problem, (6)

$$\min\{\mathbf{c}\mathbf{x} - \pi A\mathbf{x} - \alpha F\mathbf{x} - \pi_0 \mid \mathbf{x} \in X\} \quad (14)$$

or in other words, with the dynamic addition of valid inequalities formulated in the original $\mathbf{x}$ variables, only the pricing problem’s objective function needs to be updated. We may alternatively enforce the cutting planes in the pricing problem by reducing its domain to $X_F = \{\mathbf{x} \in X \mid F\mathbf{x} \leq \mathbf{f}\}$. The reason for doing so is that inequalities added to X are convexified when we solve the pricing problem as an integer program, and thus we may hope for a stronger dual bound. The pricing problem (6) becomes

$$\min\{\mathbf{c}\mathbf{x} - \pi A\mathbf{x} - \pi_0 \mid \mathbf{x} \in X_F\}, \quad (15)$$

which, interestingly, may sometimes be of the same structure as before the modification. Sets P and R need to be updated in the master problems (5) or (8); moreover, variables λ_j with $F\mathbf{x}_j > \mathbf{f}$, $j \in P \cup R$, have to be eliminated.

Generic cutting planes (see also the section titled “Automatic Convexification” in this encyclopedia) formulated on the original $\mathbf{x}$ variables sometimes seem to have little or no effect on improving the usually already strong dual bound (see *Decomposition Methods for Integer Programming*). This may also be due to the fact that usually no basic solution to the original problem is available on which several generic cutting planes rely (a crossover might help here). Currently, problem-specific valid inequalities are the alternative to go.

It is known that separation of particularly structured (e.g., integer) points may be considerably easier than separating arbitrary fractional solutions. This motivates to use the decomposition to aid separation: A fractional master solution is a convex combination of integer solutions to the pricing problems which can be recovered and separated separately. In Ralphs and Galati [20] this is called *structured separation* or *decompose and cut*.

Cutting Planes on the Master Variables

In many applications, and in particular, in the discretization approach, the master variables are integer variables. It is clear that not every inequality in the master λ variables can be derived from the original $\mathbf{x}$ variables via a Dantzig–Wolfe decomposition, and this complicates matters as we will see. Assume that we already separated a set $G\lambda \leq \mathbf{g}$ of valid inequalities in the master problem, that is,

$$\begin{aligned} \min \quad & \sum_{p \in P} c_p \lambda_p + \sum_{r \in R} c_r \lambda_r \\ \text{subject to} \quad & \sum_{p \in P} \mathbf{a}_p \lambda_p + \sum_{r \in R} \mathbf{a}_r \lambda_r \leq \mathbf{b} \\ & \sum_{p \in P} \mathbf{g}_p \lambda_p + \sum_{r \in R} \mathbf{g}_r \lambda_r \leq \mathbf{g} \\ & \sum_{p \in P} \lambda_p = 1 \\ & \lambda \in \mathbb{Z}_+^{|P|+|R|}. \end{aligned} \quad (16)$$

The dual variables β of these cuts need to be respected when calculating reduced costs in the pricing problem—if we do not—we

may regenerate “cut off” variables and end up in an infinite loop of separation and pricing. Certainly, we would lose the strength of a cut if we did not lift it. If we think of the cuts’ coefficients of a variable $\lambda_j, j \in P \cup R$, as the result of a function $\mathbf{g}_j = g(\mathbf{a}_j)$, the pricing problem reads

$$\min\{\mathbf{c}\mathbf{x} - \pi\mathbf{A}\mathbf{x} - \beta g(\mathbf{A}\mathbf{x}) - \pi_0 \mid \mathbf{x} \in X\}. \quad (17)$$

Function g can be quite complicated; it may be nonlinear as in the case of a Chvátal–Gomory rank-1 cut [9,21] with rational multipliers $\mathbf{u} \in [0, 1]^n$

$$\sum_{p \in P} \lfloor \mathbf{u}\mathbf{a}_p \rfloor \lambda_p + \sum_{r \in R} \lfloor \mathbf{u}\mathbf{a}_r \rfloor \lambda_r \leq \lfloor \mathbf{u}\mathbf{1} \rfloor.$$

It can be helpful from a conceptual viewpoint to introduce new variables $\mathbf{y} = g(\mathbf{A}\mathbf{x})$ to compute the coefficients from a solution to the pricing problem. An example is to introduce new resources, one for each cutting plane, in the resource-constrained shortest path pricing problem typically used in vehicle-routing problems. As just mentioned we cannot hope that g is linear, but if it is, that is, $\mathbf{y} = g(\mathbf{A}\mathbf{x}) = F\mathbf{x}$ we immediately see that cutting planes on the master variables contain those which can be derived from the original variables.

Desaulniers *et al.* [19] note that introducing additional variables in the pricing problem hints at an extended original (possibly nonlinear) formulation from which cutting planes in the master variables follow Dantzig–Wolfe decomposition (the section titled “Reversing Dantzig–Wolfe Decomposition”):

$$\begin{aligned} \min \quad & \mathbf{c}\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{y} \leq \mathbf{g} \\ & \mathbf{x} \in X \\ & \mathbf{y} = g(\mathbf{A}\mathbf{x}), \end{aligned} \quad (18)$$

where $\mathbf{y} \leq \mathbf{g}$ remains in the master problem while $\mathbf{y} = g(\mathbf{A}\mathbf{x})$ goes in the pricing problem.

Cutting planes on the master variables is a recent topic. So far a successful separation of clique inequalities [22], Chvátal–Gomory rank-1 cuts [21] (and subsets [23]) has been reported only for a very few problems. The modified subproblems become harder, in particular, when the computation of cut coefficients requires severe modifications in the pricing problem.

Using an Extended Formulation

As noted in the section titled “Extended Reformulations”, extended formulations may give rise to tighter relaxations as they may “better reflect integrality requirements of the problem.” Column generation based algorithms often offer natural candidates for such extended formulations when the pricing problem is solved via dynamic programming. Simple examples are the bin packing [24] or cutting stock [25] master problems where the subproblem is a knapsack problem. The dynamic program for the knapsack problem with capacity B can be formulated as a longest path problem in an acyclic network of pseudopolynomial size, namely with B nodes representing the used capacity of the knapsack. Arcs between vertices i and j represent picking an item of size $j - i$ when already a capacity of i is used. Zero-cost arcs between consecutive vertices represent unused capacity. Using these arcs as variables for a reformulation of the pricing problem one obtains a network flow problem.

In general, variables represent state transitions of the dynamic program and this may allow to formulate complex cutting planes, expressed in a simple way and without significant changes to the pricing problem. In the capacitated vehicle-routing problem [26] this approach leads to variables x_{ij}^d which state that some vehicle arrives in j , coming from i , with a remaining capacity of d . Besides such “capacity-indexed formulations,” for example, time-indexed formulations are used in scheduling problems. Flow conservation reformulated in these variables gives what they call a *base equality* from which many families of valid inequalities can be derived.

There is good experience with such kinds of cutting planes also for the capacitated

minimum spanning tree problem [27], machine scheduling problems [28], and several more [16].

BRANCHING

Even though branching decisions can be imposed by additional constraints (which we know how to do) we still have to *find* good branching rules. Disjunctive branching on the master variables is either not feasible (in convexification nobody asks for integer master variables) or not advisable: In discretization branching a master variable to zero has essentially no effect on the dual bound, while the up-branch significantly changes the solution, and thus potentially the dual bound. This produces an unbalanced search tree. Moreover, down-branching forbids certain solutions to the pricing problem to be regenerated. This problem brought up the notion of *compatibility* between pricing problem and branching rule which means that the pricing problem should not complicate after branching. Working simultaneously with original and master formulation, that is, branching on original variables helped a lot, but introduces new problems, in particular, when pricing problems are identical; we follow the classification of Vanderbeck and Wolsey [16], which contains a thorough exposition. Let λ^* denote an optimal solution to the restricted master problem.

Convexification: Branching on Original Variables

When all pricing problems are distinct, in particular, when there is only one pricing problem, the convexification approach with its integrality requirement on original variables is the natural way to go. As is immediate by Equation (5), branching candidates are all original x_i variables with $x_i^* = \sum_{j \in P \cup R} x_{ji} \lambda_j^* \notin \mathbb{Z}_+$, where x_{ji} denotes the i th component of $\mathbf{x}_j$, $j \in P \cup R$. Dichotomic branching on x_i creates two new problems, on the down-branch by imposing $x_i \leq \lfloor x_i^* \rfloor$, and on the up-branch by requiring $x_i \geq \lceil x_i^* \rceil$. There are two general options on how to enforce the branching decision, either in the

master problem or in the pricing problem. We discuss only the down-branch here, the up-branch is handled analogously.

Master Problem. The branching constraint $x_i \leq \lfloor x_i^* \rfloor$ is reformulated via convexification, that is, we add to the master problem (5) the constraint

$$\sum_{p \in P} x_{pi} \lambda_p + \sum_{r \in R} x_{ri} \lambda_r \leq \lfloor x_i^* \rfloor. \quad (19)$$

The additional dual variable α_i is respected in the pricing problem as in Equation (14), that is, only its objective function needs modification. However, no integer points in the interior of $\text{conv}(X)$ can be obtained from the pricing problem, and we may miss an optimal solution in the case of general integer variables $\mathbf{x}$.

Pricing Problem. The second option is to change the bound $x_i \leq \lfloor x_i^* \rfloor$ directly in the pricing problem, which forbids the generation of extreme points and rays which violate the branching decision. Master variables already present but incompatible with the branching decision need to be eliminated. This can be done by adding

$$\begin{aligned} \sum_{\substack{j \in P \cup R: \\ x_{ji}=1}} \lambda_j &= 0 \quad \text{or equivalently,} \\ \sum_{\substack{j \in P \cup R: \\ x_{ji}=0}} \lambda_j &= 1, \end{aligned} \quad (20)$$

to the master problem (5) which can be seen as modifying the convexity constraint (its dual variable is still denoted by π_0). The pricing problem then becomes

$$\min\{\mathbf{c}\mathbf{x} - \pi\mathbf{A}\mathbf{x} - \pi_0 \mid \mathbf{x} \in X \cap \{\mathbf{x} \mid x_i \leq \lfloor x_i^* \rfloor\}\}. \quad (21)$$

This may complicate the pricing problem, however, if it stays tractable, this option is to be preferred. The main reason is the potentially stronger dual bound from the master

problem relaxation since the bound change is convexified

$$\begin{aligned} & \min\{\mathbf{c}\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \text{conv}(X), x_i \leq \lfloor x_i^* \rfloor\} \\ & \leq \min\{\mathbf{c}\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \text{conv}(X) \\ & \quad \cap \{\mathbf{x} \mid x_i \leq \lfloor x_i^* \rfloor\}\}. \end{aligned} \quad (22)$$

Furthermore, adding disjunctive bounds to the pricing problem, that is, partitioning its domain allows to generate points in the interior of $\text{conv}(X)$ after branching.

Our presentation discussed the root node but both options directly extend to any node in the tree. One only needs to keep track of the modifications to the master and pricing problems which are local to subtrees. Both options generalize to the case of mixed integer programs.

Discretization: Avoiding the Symmetry

Branching on original variables works well when all pricing problems are distinct since

$$\mathbf{x}_k = \sum_{p \in P_k} \mathbf{x}_p \lambda_{kp}, \quad (23)$$

defines a unique projection from the λ variables into the original $\mathbf{x}$ variable space.

As pointed out in the section titled “Bordered Block-Diagonal Matrices”, the case of identical pricing problems bears a symmetry which should be avoided. One may aggregate original variables $\sum_k \mathbf{x}_k = \mathbf{z}$ and obtain a single pricing problem as above. Branching decisions disaggregate variables again and create distinct pricing problems [18]. However, as any permutation of the index set $\{1, \dots, K\}$ gives an equivalent solution, this does not eliminate the symmetry in the $\mathbf{x}_k$ variables [29].

Consider the aggregated master problem (10) of the discretization approach. Disaggregation of the variables $v_p = \sum_k \lambda_{kp}$ and using Equation (23) to obtain an original $\mathbf{x}$ solution is neither unique nor does integrality of $\mathbf{v}$ necessarily imply integrality of $\mathbf{x}$. The trick to avoid these shortcomings and symmetry at the same time is to present a projection from the master into the original variable space which does not use the one-to-one correspondence (Eq. 23) between λ_{kp} and

$\mathbf{x}_k$ variables. In other words, the grouping of λ variables is only implicit. Vanderbeck [29] (see also Vanderbeck and Wolsey [16]) proposed to obtain values $\mathbf{x}_1^*, \dots, \mathbf{x}_K^*$ (in that order) by summing variables λ_{kp} in lexicographic order of the corresponding $\mathbf{x}_p$, where $\mathbf{x}_q < \mathbf{x}_p$ means that $\mathbf{x}_q$ precedes $\mathbf{x}_p$ in that ordering. For all $k = 1, \dots, K$ and $p \in P$ let

$$\begin{aligned} \lambda_{kp}^* = \min & \left\{ 1, v_p - \sum_{\kappa=1}^{k-1} \lambda_{\kappa p}^*, \right. \\ & \left. \max \left\{ 0, k - \sum_{q: \mathbf{x}_q < \mathbf{x}_p} v_q^* \right\} \right\}. \end{aligned} \quad (24)$$

We obtain $\mathbf{x}_k^* = \sum_{p \in P} \mathbf{x}_p \lambda_{kp}^*$. The lexicographic sorting guarantees that we always work with a unique representative solution $\mathbf{x}$ out of the many symmetric possibilities. This is a standard trick which has been used in symmetry breaking of integer programs recently [30].

Aggregate Original Variables. When there happens to be a fractional aggregate original variable value $y_i^* = \sum_{p \in P} x_{pi} v_p^* \notin \mathbb{Z}_+$, which need not be the case in general, branching can be performed on such a variable by imposing

$$\begin{aligned} y_i &= \sum_{p \in P} x_{pi} v_p^* \leq \lfloor y_i^* \rfloor \quad \text{or} \\ y_i &= \sum_{p \in P} x_{pi} v_p^* \geq \lceil y_i^* \rceil \end{aligned} \quad (25)$$

in the master. This only affects the pricing problem’s objective function but this may considerably change its character. This simple rule may give only little improvement on the dual bound [16].

Auxiliary Original Variables. When the previous rule fails, that is, when an integer $\mathbf{y}$ does not yield an integer $\mathbf{x}$ (“the set of branching objects is not rich enough” [16]), one may try to work with an extended original formulation by introducing auxiliary variables to branch on, see the section titled “Extended Reformulations.” As an example consider the

set partitioning problem

$$\min \left\{ \sum_{p \in P} \mathbf{c} \mathbf{x}_p v_p \mid \sum_{p \in P} \mathbf{x}_p v_p = \mathbf{1}, \right. \\ \left. \sum_{p \in P} v_p = K, v_p \in \{0, 1\}^{|P|} \right\}. \quad (26)$$

Many problems lead to such a Dantzig–Wolfe reformulation like bin packing or vertex coloring. For the latter problem, original variables $x_{ki} \in \{0, 1\}$ state whether vertex i receives color k and columns $\mathbf{x}_p$ correspond to independent sets in the underlying graph. Aggregate variables $y_i^* = \sum_{p \in P} x_{pi} v_p^* = 1$ for any master solution, so the previous rule does not apply. However, it is well-known that in a fractional master solution there must exist rows i and j with

$$\sum_{\substack{p \in P \\ x_{pi} = x_{pj} = 1}} v_p^* =: w_{ij}^* \notin \{0, 1\}. \quad (27)$$

We (conceptually) introduce w_{ij} as auxiliary variables in the original problem for every pair (i, j) of vertices (and in the pricing problem as well) and branch on these variables. It may not be straight forward to impose the branching decision in the pricing problem directly; Ryan and Foster branching [31] is an example in which $w_{ij} \in \{0, 1\}$ is enforced by letting $x_i = x_j$ in one branch and $x_i \neq x_j$ in the other. When the pricing problem is solved via a combinatorial algorithm, often a dynamic program, this naturally suggests an extended formulation of the pricing problem which translates to an extended original formulation [16], see also the section titled “Using an Extended Formulation.”

Nested Partition of the Convexity Constraint.

The most general rule is to split the master variables by modification of the convexity constraint [29]. If

$$\sum_{p \in P: x_{pi} \geq \ell_i} v_p = \delta \notin \mathbb{Z}_+ \quad (28)$$

for an index i (corresponding to original variable x_i) and an integer bound ℓ_i , one creates

two branches with

$$\sum_{p \in P: x_{pi} \geq \ell_i} v_p \geq \lceil \delta \rceil \quad \text{or} \\ \sum_{p \in P: x_{pi} \leq \ell_i - 1} v_p \geq K - \lfloor \delta \rfloor. \quad (29)$$

The pricing problems must respect the variable bounds $x_i \geq \ell_i$ and $x_i \leq \ell_i - 1$, respectively. In order to guarantee a fractional δ in Equation (28) one may need to impose bounds on a set S of original variables. In fact, such sets are found recursively, and this generalizes the partition in Equation (29) in a nested way, producing more than two branches in general. The pricing problems must respect the bounds on variables in S and the respective complementary sets. There are several technical details to consider for which we refer to Vanderbeck [29].

This last rule provides the strongest dual bound among the above proposals and implies the smallest impact in the pricing problem (only bound changes). It should be noted that points in the interior of $\text{conv}(X)$ can be generated with this generic rule and that the depth of the search tree is polynomially bounded.

IMPLEMENTATION ISSUES

Even when the whole lot of work of implementing a branch-price-and-cut algorithm will pay off, it is a whole lot of work, even in 2010. We have seen that the freedom of choice can be enormous and some experience will certainly help. Nonetheless, it has never been easier than today with the great body of literature available.

At least when working with a convexification approach, one needs access to the values of the original variables. A trivial (but probably not efficient) way of doing this is to keep the constraints linking them to the master variables in the formulation (keeping the master constraints on the original variables is called the *explicit master* [17]). This facilitates, for example, branching on binary original variables since a simple bound change on the $\mathbf{x}$ variables implies eliminating incompatible master variables since their upper bounds are automatically changed to zero.

Even though an original (fractional) solution $\mathbf{x}^*$ is available, there is a drawback that has not been satisfactorily addressed so far: Typically, $\mathbf{x}^*$ is not a basic solution. This is important since several generic cutting planes and primal heuristics rely on that. Performing a crossover has been suggested (Matthew Galati in personal communication referred to John Forrest) as a possible remedy but there are no experiences reported on this yet.

Frameworks

There are several frameworks which support the implementation of branch-and-price algorithms like ABACUS [32], BCP [33], MINTO [34], SCIP [35], and SYMPHONY [36], to name only a few. In addition there are codes which perform a Dantzig–Wolfe decomposition of a general (mixed) integer program, and handle the resulting column generation subproblems in a generic way. BaPCod [37] is a “prototype code that solves mixed integer programs (MIPs) by application of a Dantzig–Wolfe reformulation technique.” The COIN-OR initiative (www.coin-or.org) hosts a generic decomposition code, called DIP [38] (formerly known as DECOMP), which is a “framework for implementing a variety of decomposition-based branch-and-bound algorithms for solving mixed integer linear programs” as described in Ralphs and Galati [20]. The constraint programming G12 project develops “user-controlled mappings from a high-level model to different solving methods,” one of which is branch-and-price [39]. The attempt to turn the branch-price-and-cut framework SCIP into a branch-price-and-cut solver is called GCG [40].

When Everything Fails . . .

Many problems which are approached by branch-price-and-cut are so large and complex that optimal solutions are out of reach in a reasonable computation time. What do we do then? The most honorable answer to that is: Research your problem! Is there any particular structure you can exploit, for example, by using a combinatorial algorithm to solve your pricing problems (instead of

solving them as MIPs); by formulating cutting planes; or by rethinking the entire Formulation. After all, this is what drives the innovations! The most practicable answer probably is to run a profiler to check where your code spends the CPU time, and search the bag of tricks for accelerating the weak spots. In particular, you may consider acceleration techniques [41] for solving the relaxations by column generation. The quickest (and sometimes the most promising, but certainly the dirtiest) answer is to go with a heuristic. In particular, in practical applications you may not need to close the last percents of the optimality gap. Many practitioners will use *price-and-branch*, that is, column generation is used only in the root node. In particular, if the problem is of set covering type (which it often is), one may branch on master variables and do not care about the theory. This is not elegant, but it often works. However, if the solver allows this, one should try to fix some variables (e.g., by branching) and generate further columns; this usually perceptively improves the solution quality.

A REMARK AND RECOMMENDATIONS FOR FURTHER READING

A final remark on the notion *branch-price-and-cut*. There is consent on using *branch-and-cut* and *branch-and-price*; so consequently the integration was named *branch-and-cut-and-price* in the first references. Adam Letchford (personal communication, 2005) remarked that a better style English is to omit the first *and*. In addition, exchanging *cut* and *price* reflects their order when solving the relaxation in each node, so we suggest to use *branch-price-and-cut*.

The classical—by now a bit outdated—survey on branch-and-price is by Barnhart *et al* [42]. The book [43] on column generation is, in fact, a book on branch-and-price and contains a lot of applications, in particular, vehicle routing, the cutting stock problem, and machine scheduling. The book also contains an introductory text to the topic [44] with a focus on convexification. Implementation issues can be found in Vanderbeck [45].

REFERENCES

1. Nemhauser GL, Park S. A polyhedral approach to edge coloring. *Oper Res Lett* 1991;10(6):315–322.
2. Desrosiers J, Soumis F, Desrochers M. Routing with time windows by column generation. *Networks* 1984;14:545–565.
3. Desrosiers J, Lübbecke ME. Selected topics in column generation. *Oper Res* 2005;53(6):1007–1023.
4. Vanderbeck F. Computational study of a column generation algorithm for bin packing and cutting stock problems. *Math Program* 1999;86(3):565–594.
5. Mehrotra A, Trick MA. A column generation approach for graph coloring. *INFORMS J Comput* 1996;8(4):344–354.
6. van den Akker JM, Hoogeveen JA, van de Velde SL. Parallel machine scheduling by column generation. *Oper Res* 1999;47(6):862–872.
7. Ceselli A, Righini G. A branch-and-price algorithm for the capacitated p -median problem. *Networks* 2005;45(3):125–142.
8. Savelsbergh MWP. A branch-and-price algorithm for the generalized assignment problem. *Oper Res* 1997;45(6):831–841.
9. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. Chichester: John Wiley & Sons, Inc.; 1988.
10. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
11. Geoffrion AM. Lagrangean relaxation for integer programming. *Math Program Stud* 1974;2:82–114.
12. Schrijver A. *Theory of linear and integer programming*. Chichester: John Wiley & Sons, Inc.; 1986.
13. Vanderbeck F. On Dantzig-Wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm. *Oper Res* 2000;48(1):111–128.
14. Vanderbeck F, Savelsbergh MWP. A generic view of Dantzig-Wolfe decomposition in mixed integer programming. *Oper Res Lett* 2006;34(3):296–306.
15. Vanderbeck F. A generic view of Dantzig-Wolfe decomposition in mixed integer programming. *Oper Res Lett* 2006;34(3):296–306.
16. Vanderbeck F, Wolsey L. Reformulation and decomposition of integer programs. In: Jünger M, Liebling ThM, Naddef D, *et al.*, editors. 50 years of integer programming 1958–2008. Berlin: Springer; 2010.
17. Aragão MP, Uchoa Eduardo. Integer program reformulation for robust branch-and-cut-and-price. *Annals of mathematical programming in Rio. Brazil: Búzios*; 2003. pp. 56–61.
18. Villeneuve D, Desrosiers J, Lübbecke ME, *et al.* On compact formulations for integer programs solved by column generation. *Ann Oper Res* 2005;139(1):375–388.
19. Desaulniers G, Desrosiers J, Spoorendonk S. Cutting planes for branch-and-price algorithms. 2009. Les Cahiers du GERAD G–2009–52, HEC Montréal, Forthcoming in *Networks*.
20. Ralphs TK, Galati MV. Decomposition and dynamic cut generation in integer linear programming. *Math Program* 2006;106(2):261–285.
21. Petersen B, Pisinger D, Spoorendonk S. Chvátal-gomory rank-1 cuts used in a Dantzig-Wolfe decomposition of the vehicle routing problem with time windows. In: Golden B, Raghavan S, Wasil E, editors. *The vehicle routing problem: latest advances and new challenges*. Berlin: Springer; 2008. pp. 397–419.
22. Spoorendonk S, Desaulniers G. Clique inequalities for the vehicle routing problem with time windows. *INFOR*. In press.
23. Jepsen M, Petersen B, Spoorendonk S, *et al.* Subset-row inequalities applied to the vehicle-routing problem with time windows. *Oper Res* 2008;56(2):497–511.
24. Valério de Carvalho JM. Exact solution of bin-packing problems using column generation and branch-and-bound. *Ann Oper Res* 1999;86:629–659.
25. Valério de Carvalho JM. Exact solution of cutting stock problems using column generation and branch-and-bound. *Int Trans Oper Res* 1998;5(1):35–44.
26. Fukasawa R, Longo H, Lysgaard J, *et al.* Robust branch-and-cut-and-price for the capacitated vehicle routing problem. *Math Program* 2006;106(3):491–511.
27. Uchoa E, Fukasawa R, Lysgaard J, *et al.* Robust branch-cut-and-price for the capacitated minimum spanning tree problem over a large extended formulation. *Math Program* 2008;112(2):443–472.
28. Pessoa A, Uchoa E, Poggi de Aragão M, *et al.* Algorithms over arc-time indexed formulations for single and parallel machine scheduling problems. Report RPEP Vol. 8

- no. 8, Universidade Federal Fluminense, 2008.
29. Vanderbeck F. Branching in branch-and-price: a generic scheme. Math Program 2010. In press.
30. Margot F. Symmetry in integer linear programming. In: Jünger M, Liebling ThM, Naddef D, *et al.*, editors. 50 years of integer programming 1958–2008. Berlin: Springer; 2010.
31. Ryan DM, Foster BA. An integer programming approach to scheduling. In: Wren A, editor. Computer scheduling of public transport urban passenger vehicle and crew scheduling. Amsterdam: North Holland Publishing Co.; 1981. pp. 269–280.
32. Jünger Michael, Thienel Stefan. The ABA-CUS system for branch-and-cut-and-price algorithms in integer programming and combinatorial optimization. Softw Pract Exp 2000;30(11):1325–1352.
33. Ralphs TK, Ladányi L. COIN/BCP User's Manual. 2001. Available at <http://www.coin-or.org/Presentations/bcp-man.pdf>. Accessed 2010.
34. Nemhauser GL, Savelsbergh MWP, Sigismondi GS. MINTO, a mixed integer optimizer. Oper Res Lett 1994;15:47–58.
35. Achterberg T. SCIP: solving constraint integer programs. Math Program Comput 2009;1(1):1–41.
36. Ralphs TK, Mahajan A, Güzelsoy M. SYMPHONY Version 5.2 User's Manual. COR@L Laboratory, Lehigh University, 2010. <http://www.coin-or.org/SYMPHONY/doc/SYMPHONY-5.2.3-Manual.pdf>. Accessed 2010.
37. Vanderbeck F. BaPCod—a generic branch-and-price code. 2005. Available at <https://wiki.bordeaux.inria.fr/realopt/pmwiki.php/Project/BaPCod>. Accessed 2010.
38. Ralphs TK, Galati MV. DIP—decomposition for integer programming. 2009. Available at <https://projects.coin-or.org/Dip>. Accessed 2010.
39. Puchinger J, Stuckey PJ, Wallace MG, *et al.* Dantzig-Wolfe decomposition and branch-and-price solving in G12. Constraints 2010. In press.
40. Gamrath G, Lübbecke ME. Experiments with a generic Dantzig-Wolfe decomposition for integer programs. In: Festa P, editor. Volume 6049, Proceedings of the 9th International Symposium on Experimental Algorithms (SEA), Lecture notes in computer science. Berlin: Springer; 2010. pp. 239–252.
41. Desaulniers G, Desrosiers J, Solomon MM. Accelerating strategies in column generation methods for vehicle routing and crew scheduling problems. In: Ribeiro CC, Hansen P, editors. Essays and surveys in metaheuristics. Boston (MA): Kluwer; 2001. pp. 309–324.
42. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. Oper Res 1998;46(3):316–329.
43. Desaulniers G, Desrosiers J, Solomon MM, editors. Column generation. Berlin: Springer; 2005.
44. Desrosiers J, Lübbecke ME. A primer in column generation. pp. 1–32. In Desaulniers *et al.* 2005.
45. Vanderbeck F. Implementing mixed integer column generation. pp. 331–358. In Desaulniers *et al.* 2005.

BRANCH-WIDTH AND TANGLES

ILLYA V. HICKS

Department of Computational
and Applied Mathematics,
Rice University, Houston,
Texas

SANG-IL OUM

Department of Mathematical
Sciences, KAIST, Daejeon,
Republic of Korea

Branch-width, introduced by Robertson and Seymour [1], is a general concept to describe the difficulty of decomposing finitely many objects into a tree-like structure by partitioning them into two parts recursively, while maintaining each cut to have small connectivity measure. Branch-width normally is defined for graphs or hypergraphs, as discussed by Robertson and Seymour [1], but it is easy to be extended for other combinatorial objects such as matroids and any integer-valued symmetric submodular functions.

Roughly speaking, branch-decomposition is a description on a maximal collection of nonoverlapping partitions of a finite set E . The width of a branch-decomposition is the maximum “complexity” of each part appearing in the branch-decomposition, where the “complexity” is given by some function on subsets of E . The branch-width is the minimum possible width over all possible branch-decompositions of E . Precise definition will be discussed in the following section.

To show that branch-width is small, we need to illustrate how to decompose nicely; in other words, we need to present a branch-decomposition of small width in order to certify that branch-width is small. On the other hand, if we want to certify that branch-width is large, a naive approach would be trying all possible branch-decompositions, which is too time-consuming. For that purpose we use tangles. A tangle is a dual notion of branch-width which certifies why the branch-width

is large. It was also defined by Robertson and Seymour in the same article.

In this article, we explain those definitions and list their algorithmic properties.

BRANCH-WIDTH

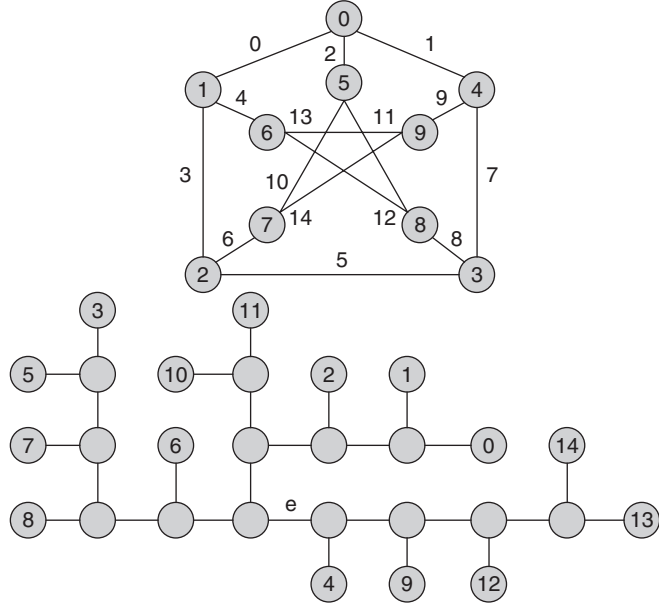
Usually branch-width is defined for graphs and hypergraphs. But for the sake of generality, we define it for integer-valued symmetric submodular functions first. An integer-valued function f on subsets of a finite set E is *symmetric* if $f(X) = f(E - X)$ for all subsets X of E and f is called *submodular* if $f(X) + f(Y) \geq f(X \cap Y) + f(X \cup Y)$ for all subsets X, Y of E .

Let us now assume that an integer-valued symmetric submodular function f on subsets of a finite set E is given. We call a tree *subcubic* if every vertex has degree 3 or 1. A *branch-decomposition* (T, τ) of f consists of a subcubic tree T and a bijection τ from the set of leaves of T to E . Then the *width* of an edge e of T is defined to be $f(\tau(A_e))$ when (A_e, B_e) is a partition of the set of leaves of T given by $T \setminus e$. Notice that this is well-defined because $f(\tau(A_e)) = f(\tau(B_e))$. The *width* of a branch-decomposition (T, τ) is the maximum width of all edges of T . The *branch-width* of f , denoted by $\text{bw}(f)$, is the minimum width of all possible branch-decompositions of f . If $|E| \leq 2$, then there are no branch-decompositions and so we just define branch-width to be $f(\emptyset)$.

By choosing an appropriate set E and an integer-valued symmetric submodular function, we can generate various notions of width parameters. Let us present some of them here.

Branch-Width of Graphs and Hypergraphs

Branch-width was first introduced by Robertson and Seymour [1] for graphs and hypergraphs. For a graph (or a hypergraph) G and a subset X of edges, let $\eta_G(X)$ be the number of vertices which are incident with an edge in X as well as an edge in $E(G) - X$. It is



studied by Dharmatilake [5] and has played an important role in the development of the matroid structure theory by Geelen *et al.* [6,7].

If a graph G has at least one cycle of length at least 2, then G and its cycle matroid $M(G)$ has the same branch-width, shown by Hicks and McMurray, Jr. [8] and independently by Mazoit and Thomassé [9] later.

Carving-Width of Graphs

Carving-width of graphs was introduced by Seymour and Thomas [10]. For a graph G and a subset A of vertices, we write $\delta_G(A)$ to denote the set of all edges joining a vertex in A with a vertex in $V(G) - A$. Let $p_G(X) = |\delta_G(A)|$. Again p_G is symmetric submodular. The *carving-width* of a graph is the branch-width of p_G . Carving-width is a useful tool for the branch-width of a planar graph because the branch-width of a planar graph is exactly half of the carving-width of its medial graph [10].

TANGLES

Tangles are introduced as a means to certify that the branch-width is large. If we wish to convince that branch-width is small, we can simply present a branch-decomposition of small width. However, we do not want to try all possible branch-decompositions in order to convince that branch-width is big. Tangles play such a role; if a tangle is presented, then no branch-decomposition of small width can exist.

For an integer-valued symmetric submodular function f on subsets of a finite set E , an f -tangle of order $k + 1$ is a collection $\mathcal{T}$ of subsets of E satisfying the following three axioms.

- (T1) For all $A \subseteq E$, if $f(A) \leq k$, then either $A \in \mathcal{T}$ or $E - A \in \mathcal{T}$.
- (T2) If $A, B, C \in \mathcal{T}$, then $A \cup B \cup C \neq E$.
- (T3) For all $e \in E$, we have $E - \{e\} \notin \mathcal{T}$.

Let us call a set X in a tangle $\mathcal{T}$ *small* and the complement $E - X$ *large*. Informally speaking, a large set is a “highly connected” set so that it is impossible to decompose

a large set properly to construct a branch-decomposition of small width. In Fig. 2, we illustrate a large set in a tangle of order 3 for the Petersen graph. Edges shown in Fig. 2 form a large set.

Robertson and Seymour introduced tangles and proved lots of useful properties. The following duality theorem is very useful. It was implicitly proved by Robertson and Seymour [1, (3.5)]. Geelen *et al.* [11, Theorem 3.2] rewrote the proof.

Theorem 1. *Let f be an integer-valued symmetric submodular function on subsets of E . Then no f -tangle of order $k + 1$ exists if and only if the branch-width of f is at most k .*

This allows us to define the branch-width from tangles; the branch-width is equal to the maximum k such that a tangle of order k exists. And to show that $\text{bw}(f) = k$ for an integer k , we frequently construct both a branch-decomposition of width k for an upper bound on the branch-width and an f -tangle of order k for a lower bound.

Providing a lower bound for the branch-width is generally harder than finding an upper bound. Therefore, much of the work to find the exact branch-width is usually devoted to finding a tangle. For the branch-width of the $n \times n$ grid, Kleitman and Saks (in Ref. 1) presented a tangle of order n , thus proving that the branch-width of the $n \times n$ grid is n . Geelen *et al.* [12] used tangles to prove that the branch-width of the cycle matroid of the $n \times n$ grid is n . For the rank-width of the $n \times n$ grid G , Jelínek [13]

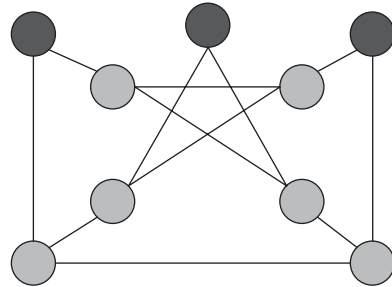


Figure 2. A “large” part in an order-4 tangle of the Petersen graph.

presented a ρ_G -tangle of order $n - 1$, thus certifying that the rank-width of the $n \times n$ grid is $n - 1$.

Roughly speaking a set of maximal tangles is used to identify highly connected pieces in a combinatorial structure. Robertson and Seymour [1] (see also Geelen *et al.* [12]) showed that any symmetric submodular function on E has at most $(|E| - 2)/2$ maximal tangles, which are displayed by a tree structure. That tree structure has been used to describe and prove the structure of graphs or binary matroids without some fixed minor.

COMPUTING BRANCH-WIDTH

One of the most natural questions after defining branch-width is the complexity of computing the branch-width of integer-valued symmetric submodular functions on subsets of a finite set E . Since we may need 2^n values of f for all subsets of E in order to input f , we will assume that f is given by an oracle so that we can query the oracle to compute $f(X)$ for the input set X at a unit time.

Hardness Results

In general, it is hard to decide whether branch-width is at most k for an integer-valued symmetric submodular function f given by an oracle and an input k in time polynomial in n . Seymour and Thomas [10] showed that it is NP-hard to compute branch-width or carving-width of a graph. Kloks *et al.* [14] proved that computing branch-width is NP-hard even for bipartite graphs or split graphs. Computing branch-width of a matroid given as a matrix representation is also NP-hard and computing rank-width of a graph is also NP-hard, because of the relationship between branch-width of graphs and branch-width of cycle matroids [8,9].

Exact Exponential-Time Algorithms

For the efficient exact algorithm, Oum [15] found an $O^*(2^{|E|})$ -time algorithm to compute the branch-width of any integer-valued symmetric submodular function f given by an oracle as above. (Here, $O^*(2^{|E|})$ means

$O(2^{|E|}|E|^{O(1)})$.) It is not known whether $O^*(2^{|E|})$ can be improved to $O^*(c^{|E|})$ for some $1 < c < 2$. For graphs $G = (V, E)$, branch-width can be computed in time $O^*((2\sqrt{3})^{|V|})$, shown by Fomin *et al.* [16].

Exact Polynomial-Time Algorithms for Special Classes

When we restrict inputs, the branch-width can sometimes be computed efficiently. Branch-width can be computed in polynomial time for circular arc graphs [17] and interval graphs [14,18]. For planar graphs, branch-width and carving-width can be computed in polynomial time, shown by Seymour and Thomas [10]. More precisely, their algorithm can decide in time $O(n^2)$ whether a given planar graph has branch-width at most k for a given k and output an optimal decomposition in time $O(n^4)$. Gu and Tamaki [19] improved that result to construct an $O(n^3)$ -time algorithm to output an optimal carving-decomposition or an optimal branch-decomposition of n -vertex planar graphs.

Testing Branch-Width at most k for Fixed k

As we discussed above, we cannot hope to have a polynomial-time algorithm to test whether branch-width is at most k for an input k . However, if we fix k as a constant, then the situation is different. Oum and Seymour [20] proved that for any fixed constant k , one can answer whether the branch-width is at most k in time $O(|E|^{8k+c})$ where c only depends on $f(\emptyset)$. Moreover, one can construct a branch-decomposition of width at most k in time $O(|E|^{8k+c+3})$.

For many applications on fixed-parameter tractable algorithms, it is desirable to have an algorithm which runs in time $O(g(k)n^c)$ for some function g and a constant c independent of k . Such an algorithm is called a *fixed-parameter tractable* algorithm with parameter k . It is still unknown whether there is a fixed-parameter tractable algorithm to decide whether branch-width of f is at most k when f is an integer-valued symmetric submodular function given as an oracle.

Fortunately, fixed-parameter tractable algorithms are known for most interesting classes of integer-valued symmetric submodular functions. Bodlaender and Thilikos [21,22] constructed a linear-time algorithm to test whether branch-width of an input graph is at most k for fixed k . Thilikos *et al.* [23] constructed a linear-time algorithm to decide whether carving-width is at most k for fixed k . Hliněný and Oum [24] showed that there exists a cubic-time algorithm to decide whether rank-width of a graph is at most k for fixed k . Their algorithm also works for branch-width of matroids represented over a fixed finite field. All of these algorithms mentioned above can output the corresponding branch-decomposition as well.

Fixed-Parameter Tractable Approximation Algorithms

For applications on fixed-parameter tractable algorithms with the branch-width as a parameter, we often need a fixed-parameter tractable algorithm to construct a branch-decomposition of small width in order to use the dynamic programming approach. So far, we do not know the existence of a fixed-parameter tractable algorithm that can output a branch-decomposition of width at most k if such a branch-decomposition exists, for an integer-valued symmetric submodular function given by an oracle. As we discussed above, the best algorithm known runs in time $O(|E|^{8k+c+3})$.

As a workaround, Oum and Seymour [2] constructed the following algorithm: for each fixed k , it runs in time $O(|E|^7 \log |E|)$ to either output a branch-decomposition of width at most $3k + c'$ or confirm that the branch-width is larger than k , where c' only depends on $f(\emptyset)$ and $\max\{f(\{e\}) : e \in E\}$. (In fact, the article [2] only discusses the case when $f(\emptyset) = 0$ and $f(\{e\}) \leq 1$ for all $e \in E$. But its argument can be modified to accommodate the case when there is an element $e \in E$ such that $f(\{e\}) - f(\emptyset) > 1$.) This allows us to construct a branch-decomposition of small width from the given adjacency list of a graph, and this branch-decomposition can be used to solve other algorithmic problems by the dynamic programming technique.

There are similar algorithms for branch-width of matroids represented over a finite field [25].

Heuristics

Cook and Seymour [26,27] gave a heuristic algorithm to produce branch-decompositions of graphs and used it in their work on the ring-routing problem and the traveling salesman problem. Hicks [28] also found another branch-width heuristic that was comparable to the heuristic of Cook and Seymour. Recently, Ma and Hicks [29] found two heuristics to derive near-optimal branch-decompositions of linear matroids; one of the heuristics uses classification techniques and the other one is similar to the heuristics for graphs which use flow algorithms.

ALGORITHMIC APPLICATIONS

Branch-Width of Graphs

There are many graph-theoretic algorithmic problems that are shown to be polynomial-time solvable on the class of graphs of bounded branch-width. Many of them actually run their algorithms based on tree-width. We refer to the section on tree-width for such applications.

Branch-width is used to design exact subexponential-time algorithms or efficient parameterized algorithms on the class of planar graphs or the class of graphs with no fixed minor [30–35].

Branch-Width of Matroids

Hliněný [36] extended Courcelle's theorem on graphs of bounded tree-width or branch-width to matroids represented over a fixed finite field. Namely, for a fixed finite field F and a given monadic second-order formula φ on matroids, one can test whether an input F -represented matroid of bounded branch-width satisfies φ in time polynomial in the size of the matroid. The requirement that the matroid has to be represented over a finite field cannot be relaxed unless $\text{NP}=\text{P}$, shown by Hliněný [37].

Hliněný [38] also found a fixed-parameter tractable algorithm to evaluate the Tutte polynomial of an input matroid represented over a fixed finite field of bounded branch-width.

Rank-Width of Graphs

Rank-width is a sibling of better known clique-width, that is a kind of a generalization of tree-width. Oum and Seymour [2] proved not only that for every class of graphs, rank-width is bounded if and only if clique-width is bounded, but also that one can translate a rank-decomposition into a decomposition for clique-width and vice versa in polynomial time. It had been known that many algorithmic properties of tree-width could be generalized to graphs of bounded clique-width, even before rank-width was introduced and it is easy to see that all of such algorithmic results on graphs of bounded clique-width apply to rank-width.

Here is one of the most important theorems for graphs of bounded rank-width. Courcelle *et al.* [39] proved that there is a cubic-time algorithm to decide whether a fixed monadic second-order formula without edge-set quantification is satisfied by an input graph of bounded rank-width. As a corollary, many hard problems such as 3-colorability are solvable in a cubic time for graphs of bounded rank-width.

Practical Algorithms

Although theory indicates the fruitful potential of these algorithms, the number of practical algorithms in the literature is scant. Most notable is the work of Cook and Seymour [27], who produced the best known solutions for the 12 unsolved problems in TSPLIB95, a library of standard test instances for the traveling salesman problem [40]. Hicks presented a practical algorithm for general graph minor containment [41] and for constructing optimal branch decompositions [42]. One is also referred to the work of Christian [43].

Based on branch-width of matroids, Cunningham and Geelen [44] proposed a

pseudopolynomial-time algorithm to solve an integer programming problem $\max(c^T x : Ax = b, x \geq 0, x \in \mathbb{Z}^n)$, when A is nonnegative and the matroid represented by A has bounded branch-width. Their algorithm shows some hope to make branch-width much more useful for practical applications, as many problems are modeled as an integer programming algorithm.

Acknowledgments

The first author was partially supported by National Science Foundation CMMI-0926618. The second author was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0001655) and by TJ Park Junior Faculty Fellowship.

REFERENCES

1. Robertson N, Seymour P. Graph minors X. Obstructions to tree-decomposition. *J Combin Theor Ser B* 1991;52(2):153–190.
2. Oum S, Seymour P. Approximating clique-width and branch-width. *J Combin Theor Ser B* 2006;96(4):514–528.
3. Courcelle B, Olariu S. Upper bounds to the clique width of graphs. *Disc Appl Math* 2000;101(1–3):77–114.
4. Oum S. Rank-width is less than or equal to branch-width. *J Graph Theor* 2008;57(3):239–244.
5. Dharmatilake JS. Binary matroids of branch-width 3 [PhD thesis]. Ohio State University; 1994.
6. Geelen J, Gerards B, Whittle G. Towards a structure theory for matrices and matroids. Volume III, International Congress of Mathematicians. Zürich: The European Mathematical Society; 2006. pp. 827–842.
7. Geelen J, Gerards B, Whittle G. Towards a matroid-minor structure theory. Volume 34, Combinatorics, complexity, and chance, Oxford Lecture Series in Mathematics and its Applications. Oxford: Oxford University Press; 2007. pp. 72–82.
8. Hicks IV, McMurray NB Jr. The branchwidth of graphs and their cycle matroids. *J Combin Theor Ser B* 2007;97(5):681–692.

9. Mazoit F, Thomassé S. Branchwidth of graphic matroids. In: Hilton A, Talbot J, editors. Volume 346, Surveys in combinatorics 2007, London Mathematical Society Lecture Note Series. Cambridge: Cambridge University Press; 2007. pp. 275–286.
10. Seymour P, Thomas R. Call routing and the ratcatcher. *Combinatorica* 1994;14(2): 217–241.
11. Geelen JF, Gerards B, Robertson N, *et al.* Obstructions to branch-decomposition of matroids. *J Combin Theor Ser B* 2006;96(4): 560–570.
12. Geelen J, Gerards B, Whittle G. Tangles, tree-decompositions and grids in matroids. *J Combin Theor Ser B* 2009;99(4):657–667.
13. Jelínek V. The rank-width of the square grid. *Disc Appl Math* 2010;158(7):841–850.
14. Kloks T, Kratochvíl J, Müller H. Computing the branchwidth of interval graphs. *Disc Appl Math* 2005;145(2):266–275.
15. Oum S. Computing rank-width exactly. *Inf Process Lett* 2009;109(13):745–748.
16. Fomin F, Mazoit F, Todinca I. Computing branchwidth via efficient triangulations and blocks. *Disc Appl Math* 2009;157(12): 2726–2736.
17. Mazoit F. The branch-width of circular-arc graphs. Volume 3887, LATIN 2006: theoretical informatics, Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 727–736.
18. Paul C, Telle JA. Branchwidth of chordal graphs. *Disc Appl Math* 2009;157(12): 2718–2725.
19. Gu Q-P, Tamaki H. Optimal branch-decomposition of planar graphs in $O(n^3)$ time. *ACM Trans Algorithms* 2008;4(3):13. Article 30.
20. Oum S, Seymour P. Testing branch-width. *J Combin Theor Ser B* 2007;97(3):385–393.
21. Thilikos DM, Bodlaender HL. Constructive linear time algorithms for branchwidth. Technical Report UU-CS 2000-38. Universiteit Utrecht; 2000.
22. Bodlaender HL, Thilikos DM. Constructive linear time algorithms for branchwidth. Volume 1256, Automata, languages and programming (Bologna, 1997), Lecture Notes in Computer Science. Berlin: Springer; 1997. pp. 627–637.
23. Thilikos DM, Serna MJ, Bodlaender HL. Constructive linear time algorithms for small cutwidth and carving-width. Volume 1969, Algorithms and computation (Taipei, 2000), Lecture Notes in Computer Science. Berlin: Springer; 2000. pp. 192–203.
24. Hliněný P, Oum S. Finding branch-decompositions and rank-decompositions. *SIAM J Comput* 2008;38(3):1012–1032.
25. Hliněný P. A parametrized algorithm for matroid branch-width. *SIAM J Comput* 2005; 35(2):259–277. loose erratum (electronic).
26. Cook W, Seymour P. An algorithm for the ring-router problem. Technical Report. Bellcore; 1994.
27. Cook W, Seymour P. Tour merging via branch-decomposition. *INFORMS J Comput* 2003;15(3):233–248.
28. Hicks IV. Branchwidth heuristics. *Congr. Numer.* 2002;159:31–50.
29. Ma J, Hicks IV. Branchwidth heuristics for linear matroids. 2009. Preprint.
30. Dorn F, Penninkx E, Bodlaender HL, *et al.* Efficient exact algorithms on planar graphs: Exploiting sphere cut branch decompositions. Volume 3669, Proceedings of the 13th Annual European Symposium on Algorithms (ESA 2005), Lecture Notes in Computer Science. Berlin: Springer; 2005. pp. 95–106.
31. Dorn F, Fomin FV, Thilikos DM. Fast subexponential algorithm for non-local problems on graphs of bounded genus. Volume 4059, Algorithm theory—SWAT 2006, Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 172–183.
32. Dorn F. Dynamic programming and fast matrix multiplication. Volume 4168, Algorithms—ESA 2006, Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 280–291.
33. Fomin FV, Thilikos DM. Dominating sets in planar graphs: branch-width and exponential speed-up. *SIAM J Comput* 2006;36(2): 281–309. (electronic).
34. Dorn F, Fomin FV, Thilikos DM. Subexponential parameterized algorithms. Volume 4596, Automata, languages and programming, Lecture Notes in Computer Science. Berlin: Springer; 2007. pp. 15–27.
35. Dorn F, Fomin FV, Thilikos DM. Catalan structures and dynamic programming in H -minor-free graphs. Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms. New York: ACM; 2008. pp. 631–640.
36. Hliněný P. Branch-width, parse trees, and monadic second-order logic for matroids. *J Combin Theor Ser B* 2006;96(3):325–351.

- 37. Hliněný P. On some hard problems on matroid spikes. *Theor Comput Syst* 2007;41(3): 551–562.
- 38. Hliněný P. The Tutte polynomial for matroids of bounded branch-width. *Combin Probab Comput* 2006;15(3):397–409.
- 39. Courcelle B, Makowsky JA, Rotics U. Linear time solvable optimization problems on graphs of bounded clique-width. *Theor Comput Syst* 2000;33(2):125–150.
- 40. Reinelt G. TSPLIB - a traveling salesman library. *ORSA J Comput* 1991;3:376–384.
- 41. Hicks IV. Branch decompositions and minor containment. *Networks* 2004;43(1):1–9.
- 42. Hicks IV. Graphs, branchwidth, and tangles! Oh my! *Networks* 2005;45(2):55–60.
- 43. Christian W. Linear-time algorithms for graphs with bounded branchwidth [PhD thesis]. Rice University; 2003.
- 44. Cunningham WH, Geelen J. On integer programming and the branch-width of the constraint matrix. In: Fishetti M, Williamson D, editors. Volume 4513, *Proceedings of the 13th International IPCO Conference*, Ithaca (NY) June 25–27, 2007, *Lecture Notes in Computer Science*. Berlin: Springer; 2007. pp. 158–166.

BRAZILIAN SOCIETY OF OPERATIONAL RESEARCH

ANNIBAL PARRACHO SANT'ANNA
Escola de Engenharia,
Universidade Federal
Fluminense, Niterói, Rio de
Janeiro, Brazil

The Brazilian Society of Operational Research (SOBRAPO) was founded in 1969, after the completion of the First Symposium of Operational Research in 1968, held in the Aeronautics Technological Institute (ITA) in São José dos Campos-SP, by Oswaldo Fadigas Sources Torres, Ricardo Alberto Von Ellenrieder, Roberto Gomes da Costa, Ruy Braga Vianna, Alfredo Otto Brockmeyer, Mario Rosenthal, Ricardo Augusto France Leme, Ellery Gyro Sergio Barroso, Ramiro A. de Almeida Sobrinho, Jointly Rodolpho Teixeira, Carlos Sigfrido Mazza, Nelson Ortega da Cunha, Antonio Campos Salles Filho, Celso Pascoli Bottura, José Luiz Fabiani, Itiro Iida, Claus Warzhariet, Sergio Grinberg, Pedro Rodrigues Bueno Neto, Sergio Viana Domingues, and Grystz Israel.

Since then, SOBRAPO has been aggregating a large community of professionals in the area of operational research in Brazil, in the universities, in business, and in the public sector. Now the society has among their associates not only people but also institutions of these three sectors of activity.

In these four decades, SOBRAPO has had 16 presidents: Oswaldo Fadigas Fontes Torres (1969–1970), Roberto Gomes da Costa (1971–1972), Ricardo Alberto Von Ellenrieder (1973–1974), Nelson Maculan Filho (1975–1976), Ramiro de Araújo Sobrinho (1977–1978), Izaltino Camozzato (1979–1980), Roberto Dieguez Galvão (1981–1982), Newton Paciornik (1983–1984), Alberto Gabbay Canen (1985–1986), Rogério de Miranda Freire (1987–1988), Celso Cruz Carneiro Ribeiro (1989–1990), Paulo Roberto Oliveira (1991–1992), Reinaldo Castro Souza

(1993–1998), Luiz Flavio Autran Monteiro Gomes (1999–2002), Nélío Domingues Pizzolato (2003–2006), and Annibal Parracho Sant'Anna (2007–2010).

Affiliated to IFORS (International Federation of Operational Research Societies) and ALIO (Asociación Latino-Ibero-Americana de Investigación Operativa), SOBRAPO by its collaboration with international institutions and by directly publishing and enhancing expertise, maintains contact with the world in general, helping to disseminate, in conferences and journals, the scientific work of Brazilian researchers. Internally, SOBRAPO is a member society of the Brazilian Society for Scientific Enhancement (SBPC).

SOBRAPO publishes its own scientific periodicals, *Pesquisa Operacional* and *Pesquisa Operacional para o Desenvolvimento*. *Pesquisa Operacional* completes its third decade, now with English as its language. It offers three issues a year, with around ten refereed articles each. It is indexed in the International Abstracts in Operations Research, and is published on-line in SciELO since 2002. *Pesquisa Operacional para o Desenvolvimento* was created in 2009 to publish on-line texts in Portuguese and in Spanish.

The editor of *Pesquisa Operacional* for the past 10 years was Horacio Hideki Yanasse (Space Research National Institute – INPE). He was preceded by Roberto Dieguez Galvao. The Editorial Board included as associate editors, Andres Weintraub (Universidade do Chile), Antonio Galvão Novaes (Universidade Federal de Santa Catarina), Basílio de Bragança Pereira (Universidade Federal do Rio de Janeiro), Brian T. Boffey (University of Liverpool), Derek Bunn (London Business School), Geraldo Robson Mateus (Universidade Federal de Minas Gerais), Gerson Lachtermacher (Instituto Brasileiro de Mercado de Capitais), Gilbert Laporte (HEC Montreal), John Beasley (Imperial College of Science, Technology and Medicine), José Mario Martinez (Universidade Estadual de

Campinas), Marcos Nereu Arenales (Universidade de São Paulo em São Carlos), Maurício G. C. Resende (AT&T Labs—Research), Nair Maria M. de Abreu (Universidade Federal do Rio de Janeiro), Nelson Maculan Filho (Universidade Federal do Rio de Janeiro), Paulo Renato de Moraes (Universidade do Vale do Paraíba), Reinaldo Castro Souza (Pontifícia Universidade Católica do Rio de Janeiro), Reinaldo Morabito (Universidade Federal de São Carlos), and Vinícius Amaral Armentano (Universidade Estadual de Campinas).

Among the topics more frequently covered in the more recent volumes of *Pesquisa Operacional* are container loading problems, cutting and packing, efficiency evaluation, facilities location, economic forecasting, genetic algorithms, graphs characterization, GRASP procedures, interior point methods, multicriteria methods, quality control, and reliability assessment.

In addition, SOBRAPO hosts annual congresses, the Brazilian Symposia on Operational Research (SBPO), with an average of 500 participants and more than 200 communications each year. The articles submitted are accepted only after careful refereeing and, if presented at the symposium, are published in the *Annals of SBPO*. SBPO calls for papers in theoretical research fields such as mathematical programming, combinatorial optimization, metaheuristics, simulation, statistics, graph theory, multicriteria decision analysis, and data envelopment analysis. From the point of view of applications, the communications include applications to education, electric energy, finance, health, information, logistics, military operations, networks, oil and gas, quality control, production administration, and transportation, among other sectors.

In 2008, to celebrate 40 years of SBPO, SOBRAPO offered prizes for the best papers presented at the symposium, in three distinct areas: optimization, decision, and management. The prize for decision was awarded to the paper “New product development projects evaluation under time uncertainty,” by Leonardo Santiago (Universidade Federal de Minas Gerais) and Thiago Silva (Universidade Federal de Minas Gerais). The prize for management was shared between the papers

“A study on the universal access to vaccines in Brazil,” by Fabio Dias Fagundes (Universidade Federal do Rio de Janeiro), João Lauro Dorneles Facó (Universidade Federal do Rio de Janeiro), Roberto Medronho (Universidade Federal do Rio de Janeiro), Adilson Xavier (Universidade Federal do Rio de Janeiro), and Leandro Xavier (Instituto Oswaldo Cruz) and “Monitoring bivariate processes,” by Antonio Fernando Branco Costa (Universidade Estadual Paulista), Fernando Antonio Elias Claro (Universidade Estadual Paulista), and Marcela Aparecida Guerreiro Machado (Universidade Estadual Paulista). The prize for optimization was also awarded to two papers: “An improved algorithm for the generalized quadratic assignment problem,” by Artur Alves Pessoa (Universidade Federal Fluminense), Monique Guignard (University of Pennsylvania), Peter Hahn (University of Pennsylvania), and Yi-Rong Zhu (University of Pennsylvania) and “A branch-and-cut SDP-based algorithm for minimum sum-of-squares clustering,” by Daniel Aloise (École Polytechnique de Montreal) and Pierre Hansen (GERAD).

Other articles selected for the final round were “A column generation approach for shared protection schemes in WDM mesh networks,” by Brigitte Jaumard (Concordia University) and Caroline Rocha (Université de Montréal); “A compact code for k -trees,” by Lilian Markenzon (Universidade Federal do Rio de Janeiro), Paulo Renato da Costa Pereira (Instituto Militar de Engenharia), and Oswaldo Vernet (Universidade Federal do Rio de Janeiro); “A fuzzy cultural immune system for economic load dispatch with non-smooth cost function,” by Myriam Delgado (Universidade Federal Tecnológica do Paraná), Carolina de Almeida (Universidade Estadual do Centro-Oeste), Ricardo Gonçalves (Universidade Estadual do Centro-Oeste/Universidade Federal Tecnológica do Paraná), Josiel Kuk (Universidade Federal Tecnológica do Paraná), and Nátalli Rodrigues (Universidade Estadual do Centro-Oeste); “An efficient iterated local search algorithm for the vehicle routing problem with simultaneous pickup and delivery,” by Lucídio Cabral (Universidade Federal da Paraíba), Luiz Satoru

Ochi (Universidade Federal Fluminense), Anand Subramanian (Universidade Federal Fluminense); “Applying a new approach methodology with ZAPROS,” by Plácido Rogério Pinheiro (Universidade de Fortaleza) and Isabelle Tamanini (Universidade de Fortaleza); “Decision Analysis for the exploration of gas reserves: merging TODIM and THOR,” by Luiz Flavio Autran Monteiro Gomes (Ibmec-RJ), Carlos Francisco Simões Gomes (Ibmec-RJ) and Francisco Maranhão (OGX Oil & Gas); “Decision theory with multiple criteria: an application of ELECTRE IV and TODIM to SEBRAE/RJ,” by Luis Alberto Duncan Rangel (Universidade Federal Fluminense, Rogério Amadel Moreira (Universidade Federal Fluminense) and Luiz Flavio Autran Monteiro Gomes (Ibmec-RJ); “Design of robust financial products using system dynamics and multi-objective genetic algorithms,” by Eder Abensur (Universidade Federal do ABC); “Safe sex and the spread of HIV,” by Valter de Senna (SENAI-CIMATEC), Hernane Pereira (SENAI-CIMATEC/Universidade Estadual de Feira de Santana), and Israel Vieira (University of Southampton); and “Tests and preventive maintenance scheduling optimization for aging systems modeled by GRP,” by Vinicius Damaso (Centro Tecnológico do Exército) and Pauli Garcia (Universidade Federal Fluminense).

Every year, SOBRAPO also awards a prize for the student or group of students submitting the best Scientific Initiation Work Report. In 2009, five reports were selected for the prize: “Intensive local search: a new meta-heuristics for restricted continuous global optimization,” submitted by Wendel Melo (Universidade Federal do Rio de Janeiro); “MDM-GRASP: a hybrid and adaptive meta-heuristic,” submitted by Richard Fuchshuber (Universidade Federal Fluminense); “Meta-heuristics with local search do the problem of just-in-time job-shop scheduling,” submitted

by Rodolfo Araujo (Universidade Federal de Viçosa); “Optimizing trajectories of nonholonomic vehicles,” submitted by André Cesar de Souza Medeiros (Universidade Federal de Minas Gerais); and “PLIM Model with linear power flow for heat exchange in electric energy distribution networks,” submitted by Lucas El G. de Lara (Universidade Federal Tecnológica do Paraná).

In the same way, every year, small-term courses are offered for the participants of SBPO. In 2009, these courses focused on economics of the health market, by Sebastian Flourier (Universidade Federal da Bahia); fractionary graph theory, by Samuel Jurkiewicz (Universidade Federal do Rio de Janeiro); multicriteria decision aid for the public sector, by Carlos Bana e Costa (Instituto Superior Técnico of Lisbon, Portugal); and optimization in management and scheduling of sports events, by Celso Carneiro Ribeiro (Universidade Federal Fluminense).

SOBRAPO also contributed toward the International Conferences on Operations Research for Development, which were held twice in Brazil, in 1996 and 2007. SOBRAPO has also participated in the organization of the Symposia of Operation Research of the Marine, held in CASNAV and CASOP, research units of the Brazilian Marine.

The society also has among its goals the development of regional initiatives around the country. Therefore, regional meetings have been held on operational research in the Northeast, in the cities of Recife-PE and Natal-PE, and in the South, in the city of Foz do Iguaçu-PR and Porto Alegre-RS. From the efforts of the SOBRAPO community, important enterprises in Brazil, like PETROBRAS and VALE, and companies of the electric energy sector have a tradition of maintaining strong research groups in operational research.

BROWNIAN MOTION AND QUEUEING APPLICATIONS

DAVID D. YAO
IEOR Department, Columbia
University, New York, New York

BASIC PROPERTIES

Let $\{B(t), t \geq 0\}$ denote the standard Brownian motion (BM). Let

$$dB(t) := B(t + dt) - B(t)$$

denote an increment of the BM, with $dt > 0$ being a time increment. Let $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2 ; let Z denote the random variable following the standard normal distribution $N(0, 1)$.

Here are some of the key properties of BM. It is a Markov process, with a continuous trajectory (sample path) over time, starting at $B(0) = 0$. It has independent and stationary increments; specifically, $dB(s)$ and $dB(t)$ are independent, for any $s + ds \leq t$, and $dB(t)$ follows a normal distribution $N(0, dt)$, which depends only on the length of the increment, but not on where it starts (hence, “shift invariant,” or stationary). Also note that it is the variance, not the standard deviation, that is proportional to the length of the increment.

Fix a time $t > 0$, and divide the interval $[0, t]$ into 2^n equal segments. Denote

$$\Delta_k(n) := B\left(\frac{k}{2^n}t\right) - B\left(\frac{k-1}{2^n}t\right).$$

Then, $\Delta_k(n)$, for $k = 1, \dots, 2^n$, are i.i.d. $N(0, t/2^n)$ random variables. The total variation (TV) and quadratic variation (QV) of the BM are defined as follows:

$$\begin{aligned} (\text{TV}) &:= \lim_{n \rightarrow \infty} \sum_{k=1}^{2^n} |\Delta_k(n)|, \\ (\text{QV}) &:= \lim_{n \rightarrow \infty} \sum_{k=1}^{2^n} [\Delta_k(n)]^2. \end{aligned} \quad (1)$$

It is known that

$$(\text{TV}) = \infty \quad \text{and} \quad (\text{QV}) = t.$$

A continuous function of time (such as what BM is) that has the above path properties—an infinite TV and a finite QV—implies that it must have an infinite number of ups and downs over any finite time interval, with the size of upward and downward movements being infinitesimally small. This essentially depicts what a BM path looks like over time.

That $(\text{QV}) = t$ also leads to the following property:

$$[dB(t)]^2 - dt = 0, \quad \text{as } dt \rightarrow 0. \quad (2)$$

Note that whereas $\mathbf{E}([dB(t)]^2) = dt$, we have,

$$\begin{aligned} \mathbf{Var}([dB(t)]^2) &= \mathbf{Var}(dt \cdot Z^2) \\ &= (dt)^2 \mathbf{Var}(Z^2) = 2(dt)^2. \end{aligned}$$

The relation in Equation (2) is, hence, intuitively appealing: the random variable $[dB(t)]^2$ has a variance that is a higher-order infinitesimal than its mean; thus, as $dt \rightarrow 0$, the random variable degenerates to a deterministic quantity (naturally, its mean dt).

Let $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ be an increasing family of sigma-fields (or, “filtration”); that is, $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s > t$. Let $X = \{X(t), t \geq 0\}$ be a process that is adapted to $\mathcal{F}$, that is, $X(t) \in \mathcal{F}_t$ for every t . (Here, $X(t) \in \mathcal{F}_t$ is shorthand for $\{X(t) \leq x\} \in \mathcal{F}_t$ for any x .) X is called a *martingale* (w.r.t. $\mathcal{F}$) if $\mathbf{E}|X(t)| < \infty$ and $\mathbf{E}[X(t)|\mathcal{F}_s] = X(s)$ for all $s < t$.

Direct verification shows that both $\{B(t)\}$ and $\{\exp[\theta B(t) - \frac{1}{2}\theta^2 t]\}$ (where θ is a deterministic parameter) are martingales. (Here, $\mathcal{F}_t$ is taken to be the process history, the sigma-field associated with the BM up to t .) For instance, to verify the martingale property of the second process, we have, for

$s < t$,

$$\begin{aligned}\mathbf{E} \exp[\theta(B(t) - B(s)) | \mathcal{F}_s] &= \mathbf{E} \exp[\theta(B(t) - B(s))] \\ &= \exp\left[\theta\sqrt{t-s}Z\right] \\ &= \exp\left[\frac{1}{2}\theta^2(t-s)\right],\end{aligned}$$

where the first equality follows from the independent increment property, and the last equality follows from the known result (generating function of Z), $\mathbf{E}(e^{\theta Z}) = e^{\theta^2/2}$.

Taking expectation on both sides of $[X(t) | \mathcal{F}_s] = X(s)$ yields $\mathbf{E}X(t) = \mathbf{E}X(s)$ for any $t > s$; in particular, $\mathbf{E}X(t) = \mathbf{E}X(0)$ for any $t > 0$. That is, the martingale property implies the process has a constant mean over time. The so-called martingale optional stopping theorem extends (under certain technical conditions) this constant mean property to a stopping time, T : $\mathbf{E}X(T) = \mathbf{E}X(0)$. (Recall T is a stopping time if $\{T \leq t\} \in \mathcal{F}_t$ for any t .)

Let $X(t) = \mu t + \sigma B(t)$ be a BM with drift, where μ and $\sigma > 0$ are constant parameters. Consider the stopping time, $T = \inf\{t : X(t) \notin (-a, b)\}$, where a and b are positive constants. We want to derive p_b , the probability of $X(t)$ hitting b before $-a$. Applying the optional stopping theorem to the exponential martingale, we have

$$\begin{aligned}\mathbf{E}\left[\exp(\theta B(T) - \frac{1}{2}\theta^2 T)\right] &= 1 \\ \text{or} \\ \mathbf{E}\left[\exp\left(\frac{\theta}{\sigma}(X(T) - \mu T) - \frac{1}{2}\theta^2 T\right)\right] &= 1.\end{aligned}$$

Letting $\theta = -\frac{2\mu}{\sigma}$ yields $-\frac{\theta}{\sigma}\mu T - \frac{1}{2}\theta^2 T = 0$; hence, the second equation above reduces to

$$\begin{aligned}1 &= \mathbf{E}\left[\exp\left(-\frac{2\mu}{\sigma^2}X(T)\right)\right] \\ &= e^{2a\mu/\sigma^2}(1 - p_b) + e^{-2b\mu/\sigma^2}p_b.\end{aligned}$$

Thus,

$$p_b = \frac{e^{2a\mu/\sigma^2} - 1}{e^{2a\mu/\sigma^2} - e^{-2b\mu/\sigma^2}}. \quad (3)$$

Now, suppose $\mu < 0$, that is, the BM has a negative drift. Letting $a \rightarrow +\infty$ in Equation (3), we have $p_b = e^{-2b|\mu|/\sigma^2} < 1$; that is, the

probability of ever reaching b is less than 1. We can also write this as

$$\mathbf{P}\left[\sup_{0 \leq t < \infty} X(t) \geq b\right] = e^{-2b|\mu|/\sigma^2}, \quad \mu < 0, b \geq 0. \quad (4)$$

Note that $\sup_{0 \leq t < \infty} X(t)$ can be viewed as the limit, as $t \rightarrow \infty$, of $M(t) := \sup_{0 \leq s \leq t} X(s)$, the latter being nonnegative and nondecreasing. Hence, the maximum of a BM with a negative drift has a limiting distribution that is exponential with mean $\sigma^2/(2|\mu|)$.

DONSKER'S THEOREM

Consider the partial sums:

$$S_0 := 0, \quad S_k := \xi_1 + \cdots + \xi_k, \quad k = 1, 2, \dots, \quad (5)$$

where $\{\xi_i, i \geq 1\}$ is a sequence of i.i.d. non-negative random variables. Assume that $\mathbf{E}(\xi_1) = 1/\mu$ and $\mathbf{Var}(\xi_1) = \sigma^2$. Define for $t \geq 0$,

$$\begin{aligned}\alpha(t) &= S_{[t]} = \sum_{i=1}^{[t]} \xi_i, \\ \beta(t) &:= \sup\{s \geq 0 : \alpha(s) \leq t\}.\end{aligned} \quad (6)$$

Note in the context of renewal theory, with $\{\xi_i\}$ being the i.i.d. interevent times, $\alpha(t)$ is the $[t]$ -th arrival epoch and $\beta(t)$ is the associated counting process. From standard renewal theory, we know, as $n \rightarrow \infty$,

$$\begin{aligned}\bar{\alpha}^n(t) &:= \frac{1}{n}\alpha(nt) \rightarrow \frac{t}{\mu} := \bar{\alpha}(t), \\ \bar{\beta}^n(t) &:= \frac{1}{n}\beta(nt) \rightarrow \mu t := \bar{\beta}(t).\end{aligned} \quad (7)$$

Here, the mode of convergence is a.s. (almost surely, or with probability 1). More precisely, it is u.o.c., uniformly on all compact subsets of any time interval $[0, T]$, with a given T . The limits in Equation (7) can be viewed in the spirit of strong law of large numbers. For instance, for each given t , the average $\bar{X}^n(t)$ approaches its mean $\frac{t}{\mu}$. As the time

index t changes, the limit takes the form of a deterministic function of time. Hence, these limits are often referred to as a *functional strong law of large numbers* (FSLLN).

Note in Equation (7), a common scaling factor n is applied to both time and space. This resembles the “zoom-out” operation when we view a Google map, for instance, as the same scaling is applied to both the horizontal and vertical axes (of the map). The net effect of this scaling is that the random fluctuation around the mean disappears, resulting in a straight line. In order to preserve the random fluctuation, what we can do is to scale the space by a factor of a lesser order, $\sqrt{n}$, and still scale time by n . This way, the fluctuation around the mean is made more prominent. This is in the same spirit as how the central limit theorem refines the strong law of large numbers. The result under this scaling is the well-known *Donsker’s Theorem* [1], which is a functional central limit theorem (FCLT), with the (functional) limit being BM:

$$\begin{aligned}\hat{\alpha}^n(t) &:= \frac{1}{\sqrt{n}} \left[\alpha(nt) - \frac{nt}{\mu} \right] \\ &= \sqrt{n} [\bar{\alpha}^n(t) - \bar{\alpha}(t)] \\ &\Rightarrow \sigma B(t) := \hat{\alpha}(t),\end{aligned}\quad (8)$$

$$\begin{aligned}\hat{\beta}^n(t) &:= \frac{1}{\sqrt{n}} [\beta(nt) - \mu(nt)] \\ &= \sqrt{n} [\bar{\beta}^n(t) - \bar{\beta}(t)] \\ &\Rightarrow \mu^{3/2} \sigma B(t) := \hat{\beta}(t).\end{aligned}\quad (9)$$

The convergence mode above ($\Rightarrow$) is the so-called “weak convergence,” that is, the probability laws (distributions) of a sequence of processes converge to the probability law (distribution) of another process, the limit. In fact, the convergence is *joint*, that is, $(\hat{\alpha}^n, \hat{\beta}^n) \Rightarrow (\hat{\alpha}, \hat{\beta})$; and the limit of $\hat{\beta}^n$, $\hat{\beta}(t) = -\mu \hat{\alpha}(\mu t)$, follows from the limit of $\hat{\alpha}^n(t)$, based on the inverse functional relationship between α and β .

THE SKOROHOD PROBLEM

Let $x := \{x(t), t \geq 0\}$ be a process (deterministic or stochastic) that is right continuous with left limit (RCLL). Let y and z denote two other RCLL processes. Let $\mathcal{D}$ denote the functional space of all RCLL processes. (Note $\mathcal{D}$ includes all continuous processes as a subclass.) The Skorohod problem refers to the following: given the process x , find a pair (y, z) that satisfies

$$z = x + y \geq 0; \quad dy \geq 0, \quad y(0) = 0; \quad z dy = 0. \quad (10)$$

It can be directly verified that the solution to the above problem does exist and is unique; furthermore, with the solution denoted as $y = \Psi(x)$ and $z = \Phi(x)$, the mappings Ψ and Φ are both Lipschitz continuous. In fact, in the single-dimensional case (i.e., when x is a scalar process), the unique y and z can be explicitly expressed as follows:

$$\begin{aligned}y(t) &= \sup_{0 \leq s \leq t} [-x(s)]^+, \\ z(t) &= x(t) + \sup_{0 \leq s \leq t} [-x(s)]^+.\end{aligned}\quad (11)$$

When $x(t) := X(t) = \theta t + \sigma B(t)$, the BM with drift discussed earlier, with $\theta < 0$. Then, the solution to the above Skorohod problem is (note $X(0) = 0$)

$$\begin{aligned}y(t) &= \sup_{0 \leq s \leq t} \{-X(s)\} := Y(t), \\ z(t) &= \sup_{0 \leq s \leq t} \{X(t) - X(s)\} := Q(t).\end{aligned}\quad (12)$$

Note that

$$Q(t) \stackrel{d}{=} \sup_{0 \leq s \leq t} \{X(s)\} = M(t),$$

where $M(t)$ is the maximum BM discussed earlier (Eq. 4). $Q(t)$ is called a *reflected Brownian motion*, denoted by $\text{RBM}(\theta, \sigma^2)$. It has a stationary (or limiting) distribution, which is exponential, with mean $\sigma^2/(2|\theta|)$.

THE GI/GI/1 QUEUE

Now, consider a single-server queue, with a renewal arrival process and i.i.d. service times following a general distribution, the GI/GI/1 model. It is well known in queueing theory that this model does not have closed-form solutions to even the most basic performance measures such as steady-state mean queue length and average waiting time. Below, we first present the equations that govern the dynamics of this model, and then connect the equations via diffusion scaling to the Skorohod problem, which has an explicit solution, the RBM. This way, the GI/GI/1 model, under diffusion scaling, becomes very accessible analytically.

Let $A(t)$ and $S(t)$ be the renewal counting processes associated with i.i.d. interarrival times and i.i.d. service times. Let λ and μ be the arrival and service rates. Let $Q(t)$ be the queue-length process, that is, the total number of jobs in the systems at time t . Then, we have

$$\begin{aligned} Q(t) &= Q(0) + A(t) - S(D(t)), \quad \text{with} \\ D(t) &= \int_0^t \mathbf{1}\{Q(s) > 0\} ds. \end{aligned} \quad (13)$$

Here, $D(t)$ denotes the cumulative amount of time up to t when work is being depleted from the system (or equivalently, the server is busy). We can rewrite the above as

$$Q(t) = X(t) + Y(t), \quad (14)$$

with

$$\begin{aligned} X(t) &:= Q(0) + (\lambda - \mu)t + [A(t) - \lambda t] \\ &\quad - [S(D(t)) - \mu D(t)], \end{aligned} \quad (15)$$

$$Y(t) := \mu[t - D(t)]. \quad (16)$$

Furthermore, the following relations must hold: for all $t \geq 0$,

$$\begin{aligned} Q(t) &\geq 0; \quad dY(t) \geq 0, \quad Y(0) = 0; \\ Q(t) dY(t) &= 0. \end{aligned} \quad (17)$$

Note $Y(t)$ is the cumulative unused service capacity (service rate times idle time) up to t .

Hence, it must be nondecreasing ($dY(t) \geq 0$). Furthermore, $Q(t) dY(t) = 0$ means that the server must not be idling while the queue length is positive. Comparing the queueing dynamics in Equations (14) and (17) against Equation (10), we conclude that the former too constitute a Skorohod problem. In this case, however, the solution form in Equation (12) is not particularly helpful, as X is not quite a primitive process (input data), as it involves Q via D . Nevertheless, it turns out that a sequence of scaled queueing processes will converge to a limit that is characterized by the Skorohod problem, and with an explicit solution.

To illustrate, we first need the following result. Let $\{\theta_n, n \geq 1\}$ be a sequence of real values and $\{x_n, n \geq 1\}$ be a sequence in $\mathcal{D}$ with $x_n(0) \geq 0$. Let

$$\begin{aligned} y_n(t) &= \sup_{0 \leq u \leq t} [-x_n(u) - \theta_n u]^+, \\ z_n(t) &= x_n(t) + \theta_n t + y_n(t). \end{aligned} \quad (18)$$

Suppose that as $n \rightarrow \infty$, we have $\theta_n \rightarrow \theta < \infty$ and $x_n \rightarrow x$ u.o.c., with the limit $x := \{x(t)\}$ being a *continuous* function. Then, $(y_n, z_n) \rightarrow (y, z)$ u.o.c. as $n \rightarrow \infty$, where

$$\begin{aligned} y(t) &= \sup_{0 \leq u \leq t} [-x(u) - \theta u], \\ z(t) &= x(t) + \theta t + y(t). \end{aligned} \quad (19)$$

THE DIFFUSION LIMIT

Continue with the GI/GI/1 model introduced above. Applying fluid scaling to the primitives, the arrival and service counting processes, we assume the following limits exist as $n \rightarrow \infty$:

$$\begin{aligned} \bar{A}^n(t) &= \frac{1}{n} A(nt) \rightarrow \lambda t, \\ \bar{S}^n(t) &= \frac{1}{n} S(nt) \rightarrow \mu t. \end{aligned}$$

Assume in addition that $\bar{Q}^n(0) := \frac{1}{n} Q^n(0) \rightarrow \bar{Q}(0)$ as $n \rightarrow \infty$. Then, applying the same fluid scaling to Equation (15), and using the same

“bar” notation as in $\bar{A}^n(t)$ and $\bar{B}^n(t)$ above to all the processes concerned, we have

$$\begin{aligned}\bar{X}^n(t) &= \bar{Q}^n(0) + (\lambda - \mu)t + [\bar{A}^n(t) - \lambda t] \\ &\quad - [\bar{S}^n(\bar{D}^n(t)) - \mu \bar{D}^n(t)] \\ &\rightarrow \bar{Q}(0) + (\lambda - \mu)t := \bar{X}(t).\end{aligned}$$

Note here $\bar{S}^n(\bar{D}^n(t)) - \mu \bar{D}^n(t) = \frac{1}{n}[S(D(nt)) - \mu D(nt)] \rightarrow 0$, since

$$\begin{aligned}\frac{1}{nt} |S(D(nt)) - \mu D(nt)| &= \left| \frac{S(D(nt))}{D(nt)} - \mu \right| \\ \cdot \frac{D(nt)}{nt} &\leq \left| \frac{S(D(nt))}{D(nt)} - \mu \right| \rightarrow 0,\end{aligned}$$

where the inequality follows from $0 \leq D(nt) \leq nt$, and the limit follows from SLLN, taking into account that $D(nt) \rightarrow \infty$ as $n \rightarrow \infty$.

Consequently, we have

$$\begin{aligned}\bar{Y}(t) &= [-\bar{Q}(0) - (\lambda - \mu)t]^+, \\ \bar{Q}(t) &= [\bar{Q}(0) + (\lambda - \mu)t]^+, \end{aligned}$$

taking into account $\bar{Y}(t) = \sup_{0 \leq s \leq t} [-\bar{X}(s)]^+ = [-\bar{X}(t)]^+$. We also have $\bar{D}^n(t) \rightarrow \bar{D}(t) = t - \frac{1}{\mu} \bar{Y}(t)$.

Under diffusion scaling, the arrival and service counting processes converge to BM (via Donsker’s Theorem) as follows:

$$\begin{aligned}\hat{A}^n(t) &= \frac{A(nt) - \lambda nt}{\sqrt{n}} = \sqrt{n} [\bar{A}^n(t) - \lambda t] \\ &\Rightarrow \sqrt{\lambda} c_a B_1(t) := \hat{A}(t), \\ \hat{S}^n(t) &= \frac{S(nt) - \mu nt}{\sqrt{n}} = \sqrt{n} [\bar{S}^n(t) - \mu t] \\ &\Rightarrow \sqrt{\mu} c_s B_2(t) := \hat{S}(t),\end{aligned}$$

where B_1 and B_2 are two independent standard BMs, and $c_a := \lambda \sigma_a$ and $c_s := \mu \sigma_a$ are the coefficients of variation of the interarrival times and the service times, respectively.

Applying diffusion scaling to the queue-length process, we have

$$\hat{Q}^n(t) = \hat{X}^n(t) + \hat{Y}^n(t), \quad (20)$$

where (assuming $\hat{Q}^n(0) = 0$ for simplicity)

$$\hat{X}^n(t) = (\lambda - \mu)\sqrt{nt} + \hat{A}^n(t) - \hat{S}^n(\bar{D}^n(t)), \quad (21)$$

and $\hat{Y}^n(t) = \sqrt{n}\mu[t - \bar{D}^n(t)]$. Note the drift term in $\hat{X}^n(t)$ has a factor of $\sqrt{n}$. Consequently, $\hat{X}^n \rightarrow +\infty$ (and $\hat{Q}^n \rightarrow +\infty$) if $\lambda > \mu$, whereas $\hat{X}^n \rightarrow -\infty$ (and $\hat{Q}^n \rightarrow 0$) if $\lambda < \mu$. Hence, a nontrivial limit requires $\lambda = \mu$, the so-called *heavy traffic condition*. In this case, following the fluid limit derived above, we have $\bar{D}^n(t) \rightarrow \bar{D}(t) = t$. Hence, $\hat{X}^n(t) \Rightarrow \hat{A}(t) - \hat{S}(t)$.

A refinement is achieved by considering a sequence of queues indexed by n . Let λ_n and μ_n be the arrival rate and the service rate, respectively, for the n th queue. Assume the following:

$$\begin{aligned}\sqrt{n}[\lambda_n - \mu_n] &\rightarrow \theta < 0, \quad \text{and} \\ \lambda_n &\rightarrow \lambda, \quad \mu_n \rightarrow \mu.\end{aligned} \quad (22)$$

The diffusion-scaled arrival and service processes above require only minor modifications as follows:

$$\begin{aligned}\hat{A}^n(t) &= \frac{A^n(nt) - \lambda_n nt}{\sqrt{n}} = \sqrt{n} [\bar{A}^n(t) - \lambda_n t] \\ &\Rightarrow \sqrt{\lambda} c_a B_1(t) := \hat{A}(t), \\ \hat{S}^n(t) &= \frac{S^n(nt) - \mu_n nt}{\sqrt{n}} = \sqrt{n} [\bar{S}^n(t) - \mu_n t] \\ &\Rightarrow \sqrt{\mu} c_s B_2(t) := \hat{S}(t).\end{aligned}$$

That is, the diffusion scaling is now applied to (the primitive processes of) the sequence of queues, and so is the fluid scaling $\bar{A}^n(t) = \frac{1}{n}A^n(nt) \rightarrow \lambda t$ and $\bar{S}^n(t) = \frac{1}{n}S^n(nt) \rightarrow \mu t$.

Consequently, Equation (21) becomes

$$\begin{aligned}\hat{X}^n(t) &= (\lambda_n - \mu_n)\sqrt{nt} + \hat{A}^n(t) - \hat{S}^n(\bar{D}^n(t)) \\ &\Rightarrow \theta t + \hat{A}(t) - \hat{S}(t).\end{aligned} \quad (23)$$

Therefore, in this case, $\hat{Q}^n(t) \Rightarrow \text{RBM}(\theta, \lambda c_a^2 + \mu c_s^2)$.

Note that the assumed conditions in Equation (22) imply $\lambda = \mu$; hence, the heavy traffic condition remains in force. However, the above enables us to view the GI/GI/1 queue in question as a member of the

sequence of queues under scaling, so that we can *approximate* it, when $\lambda \neq \mu$, by the $\text{RBM}(\theta, \lambda c_a^2 + \mu c_s^2)$, with $\theta = \lambda - \mu$. When $\lambda < \mu$ (or, $\theta < 0$), we know the queue has a stationary distribution, which can also be approximated by the stationary distribution of the RBM, which we know is exponential with a mean $(\lambda c_a^2 + \mu c_s^2)/(2|\theta|)$.

BIBLIOGRAPHICAL NOTES

Introductory materials for BM can be found in Karlin and Taylor [2] (Chapter 8) and in Ross [3] (Chapter 8). Harrison [4] covers both the basic theory and essential applications of BM in inventory and queueing systems. Ref. 1 is a standard reference on weak convergence; also refer to Ref. 5 for a survey of weak convergence in the $\mathcal{D}$ space and related diffusion approximations.

Queueing systems that have closed-form analytical solutions are limited to models that have Poisson arrivals and/or i.i.d. exponential service times, and in most cases, it is also limited to a single server. Once outside the Markov chain (or embedded Markov chain) setting, the most effective analytical approach is asymptotic analysis, via fluid and diffusion scalings.

Kingman [6–8] pioneered the diffusion approximation of single-server queues; also refer to Borovkov [9]. Iglehart and Whitt [10,11] developed the heavy-traffic theory for many-server queues. Halfin and Whitt [12] applied a different kind of diffusion scaling to the many-server queue model with exponential service times, leading to a limiting regime that is a combination of BM and Orstein–Uhlenbeck process.

This article is adapted from Chapters 5 and 6 of Ref. 13. Later chapters of the book discuss applications of Brownian models to networks of queues, as opposed to the stand-alone (i.e., single-node) queue model presented here. In this regard, Refs 14 and 15 are the first works that extend the heavy-traffic theory to single-class queueing networks (generalized Jackson networks). Harrison [16] presents an overview of the Brownian models of general stochastic processing networks with multiclass jobs

and operating under various routing and service mechanisms. Dai [17] demonstrates that the fluid model is a critical tool in characterizing the stability of multiclass networks. Williams [18] establishes the general theory of reflection mapping, or Skorohod problem in higher dimension, which is the key to establish limiting regimes for networks. Reference 19 is one example, among many others in the literature, that illustrates the advantage of asymptotic analysis not only in performance evaluation but also in identifying asymptotically optimal scheduling policies (i.e., optimal under diffusion scaling). More recent studies by Kang *et al.* [20] and Ye and Yao [21] provide further examples of the power of Brownian models in handling certain new features in contemporary stochastic networks, such as resource controls involved in Internet protocols.

REFERENCES

1. Billingsley P. Convergence of probability measures. 2nd ed. New York: Wiley; 1999.
2. Karlin S, Taylor H. A first course in stochastic processes. 2nd ed. New York: Academic Press; 1975.
3. Ross S. Stochastic processes. 2nd ed. New York: John Wiley & Sons, Inc.; 1996.
4. Harrison JM. Brownian motion and stochastic flow systems. New York: Wiley; 1985.
5. Glynn PW. Diffusion approximations. In: Heyman DP, Sobel MJ, editors. Volume 2, Handbooks in operations research and management science: stochastic models. New York: Elsevier; 1990. pp. 145–198.
6. Kingman JFC. On queues in heavy traffic. *Proc Camb Philos Soc* 1961;57:902–904.
7. Kingman JFC. The single server queue in heavy traffic. *J R Stat Soc [Ser B]* 1962;24:383–392.
8. Kingman JFC. The heavy traffic approximation in the theory of queues. In: Smith W, Wilkinson W, editors. *Proceedings of Symposium on Congestion Theory*. Chapel Hill: University of North Carolina Press; 1965. pp. 137–159.
9. Borovkov AA. Asymptotic methods in queueing theory. New York: Wiley; 1984.

10. Iglehart DL, Whitt W. Multiple channel queues in heavy traffic, I. *Adv Appl Probab* 1970;2:150–177.
11. Iglehart DL, Whitt W. Multiple channel queues in heavy traffic, II. *Adv Appl Probab* 1970;2:355–364.
12. Halfin S, Whitt W. Heavy traffic limits for queues with many exponential servers. *Oper Res* 1981;29:567–588.
13. Chen H, Yao DD. *Fundamentals of queueing networks: performance, asymptotics, and optimization*. New York: Springer; 2001.
14. Reiman MI. Open queueing networks in heavy traffic. *Math Oper Res* 1984;9:441–458.
15. Harrison JM, Reiman MI. Reflected Brownian motion on an orthant. *Ann Probab* 1984;9:302–308.
16. Harrison JM. A broader view of Brownian networks. *Ann Appl Probab* 2003;13:1119–1150.
17. Dai JG. On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann Appl Probab* 1995;5:49–77.
18. Williams RJ. An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Syst Theory Appl* 1998;30:5–25.
19. Mandelbaum A, Stolyar AL. Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$ -rule. *Oper Res* 2004;52:836–855.
20. Kang WN, Kelly FP, Lee NH, *et al.* State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy; 2007. In press.
21. Ye H, Yao DD. Heavy traffic optimality of a stochastic network under utility-maximizing resource control. *Oper Res* 2008;56:453–470.

BUSINESS PROCESS OUTSOURCING

ANDY A. TSAY

OMIS Department, Leavey School of
Business, Santa Clara University,
Santa Clara, California

This article describes a practice that is implicitly considered by every individual or organization every day, is central to the strategic business models of many modern firms, and has even become a mainstay of lay conversation (albeit with sometimes incorrect usage). We first introduce terminology necessary to explain “Business Process Outsourcing” (BPO) in general (henceforth, simply “Outsourcing”), discuss the decision process for choosing whether or not to outsource an activity, and then summarize best practices for managing the service providers. We conclude with a comment about ORMS research on outsourcing. This article is intended to serve as a tutorial, and will not provide a comprehensive review of the research literature. A follow-up article in this encyclopedia (see *Supply Chain Outsourcing*) extends this BPO dialogue to the outsourcing of manufacturing/production/assembly, procurement/sourcing, logistics, and product design/development.

TERMINOLOGY

The Oxford English Dictionary (OED) offers this definition:

Outsource: to obtain (goods, a service, etc.) by contract from an outside source; to contract (work) out.

The OED cites as the term’s earliest appearance in print a 1979 item in the *Journal of the Royal Society of Arts*, in the sentence “We are so short of professional engineers in the motor industry that we are having to outsource design work to Germany.”

Some interpret the term to specifically connote the act of shifting an internal activity to an outside party. However, the above definition, which we follow in this article, does not stipulate where the activity might have been formerly performed. For instance, firms are correctly said to be outsourcing their manufacturing even if they opted from day one to focus solely on designing and marketing their products, so that at no point ever possessed any manufacturing capabilities [1].

The antonym of “outsource” is “insource,” which thus means to perform an activity internally. Likewise this does not require that the activity was ever previously outsourced. The OED shows both words to have begun appearing in print around the same time.

The act of outsourcing involves two main participants, neither of which has a prevailing name. Some possibilities for the one receiving the good or service are “buyer,” “client,” “service recipient,” or “outsourcer.” The providing party can be “supplier,” “vendor,” “service provider,” or “outsourcee.” Of these, the mainstream usages of “buyer,” “supplier,” and “vendor” slightly hint at the selling of packaged product rather than services, although nothing in the formal definitions specifies this. “Outsourcer” and “outsourcee” draw specific attention to the nature of the relationship. The latter is not commonly used, perhaps since it could be misunderstood to be the internal employee laid off when his/her function was outsourced. To add to the confusion, the firm on the selling side is occasionally labeled as an “outsourcer.” In this article, we will generally identify these two parties as the “outsourcing party” and “service provider,” since these are sufficiently neutral and clear. The latter also has support in the labels applied to such emerging specialist categories as “Procurement Service Providers” (PSP) or “Manufacturing Service Providers” (MSP). This article will use language descriptive of outsourcing performed by organizations for business purposes, although most of

the concepts will be just as relevant when individuals outsource or when the objectives are noncommercial.

Besides naming the actors and their actions, we also need vocabulary to identify the constellation of linked partners that results from extensive outsourcing. A nonexhaustive list includes “virtual supply chain,” “virtual value chain,” “virtual integration,” and “extended enterprise.” The first two differ in the subtle distinction between a supply chain, which describes the parties along a physical path of flow, and a value chain, which highlights the activities performed but does not necessarily map to a physical or chronological sequence or have a crisp division of labor. “Virtual integration” forms a dyad with “vertical integration.” “Extended enterprise” may be the least explicitly suggestive of outsourcing. This simply encompasses the full ecosystem of parties needed to provide a product or service, but does not allude to a consolidated alternative. Terms of this ilk are disparaged by some as business jargon, and those mentioned here may very well be *passé* by the time this article appears.

Two additional keywords merit elaboration, since they arise in nearly every discussion of outsourcing. They are “offshoring,” which is a distinct but sometimes related business action, and “core competence” (CC), which is central to one of the popular rationales for outsourcing.

Offshoring

“Outsourcing” is sometimes misused in place of “offshoring,” especially in political commentaries that unfairly disparage the former for endangering the jobs of hard-working local citizens. In fact, while offshoring moves work to another country, outsourcing only shifts tasks to another organization and need not entail a location change at all. The employees of service providers sometimes work alongside the client’s internal staff, wearing the same uniforms, checking email on the same servers, and living and paying taxes in the same communities. Offshoring typically seeks to leverage an internationally based workforce that is cheaper and/or better suited for a task, but may also reduce

taxes and duties, and offer proximity to end-customers and input suppliers.

In this age of global free trade and increasingly complete marketplaces for virtually every imaginable product or service, a firm can outsource without going offshore, and vice versa. Nearly every multinational corporation outsources some activities to domestic vendors and insources other activities via wholly owned facilities that may be spread across many countries. For instance, GM outsources aspects of production to vendors in the United States, as well as Canada and Mexico. Meanwhile, Toyota and BMW own production facilities in the United States. Even in the highly outsourced mobile phone sector, the majority of Nokia’s production occurs at the Finnish firm’s own factories around the world (including in Finland and the United States) [2]. In 2007, Wipro, an India-based leading provider of outsourced IT services, announced plans to open four software development centers in the United States. Through this, Wipro will offshore without outsourcing, while Wipro’s American clients will be outsourcing without offshoring.

Outsourcing and offshoring do sometimes occur simultaneously, for which the unambiguous label is “offshore outsourcing.” This strategy is motivated by a belief that the shortest path to the benefits offered by an offshore solution is to outsource to a service provider with expertise and resources in the appropriate geographies. Everything from low end manual labor to high end knowledge work is a candidate for offshore outsourcing these days.

Should outsourcing take activities offshore, the risk factors and challenges detailed in the section titled, “Advantages and Disadvantages of Outsourcing” will only be intensified by any cultural or language barriers, differences in legal codes and enforcement practices (especially vis-à-vis the protection of intellectual property), or misalignment in attitudes toward environmental and human rights issues. And geographic distance only complicates the monitoring needed to assure that a service provider’s actions are true to its customer’s intentions. These issues are particularly

salient when offshore outsourcing involves emerging economies.

Core Competence

Prahalad and Hamel popularized the notion of CC in a 1990 *Harvard Business Review* article [3]. Their CCs, of which most firms will have not more than five or six, are defined by three key attributes:

- They provide potential access to a wide variety of markets.
- They make a significant contribution to perceived customer benefits of the end product.
- They are difficult for competitors to imitate.

The basic message in that article is that an organization can maximize its competitive advantage by identifying its CCs and organizing activities around them. These authors deem the outsourcing of CCs to be a strategic error of the highest order, but make no pronouncement about how to handle the noncore activities.

Quinn and Hilmer [4] articulate the connection between CCs and outsourcing that has become central to the modern business zeitgeist, paraphraseable as “Focus on your CCs, and outsource everything else.” By their definition, CCs are

- skill or knowledge sets, not products (which can be reverse-engineered) or functions (since CCs tend to cut across traditional functions, e.g., production, engineering, sales, finance);
- flexible, long-term platforms that are capable of adaptation or evolution;
- limited in number to perhaps two or three (more than one, but fewer than five);
- unique sources of leverage in the value chain;
- areas where the company can dominate;
- elements important to customers in the long run;
- embedded in the organization’s systems (rather than dependent upon key individuals).

In the eyes of both sets of authors, CCs are *not* “things we do very well or very often,” but instead *are* “things that are strategically important.” These are rarely confined to individual product departments or functional areas. Given this, current usage has become somewhat of a perversion of what these articles expound, as evidenced by commonly heard statements such as “We outsource manufacturing because design and marketing are our CCs.” Perhaps this can be reconciled through the way the term’s meaning has evolved since the early 1990s, which is captured in far too many articles to document here.

A semantic matter is whether the second C in the term should stand for “competence” or “competency.” The OED views these as interchangeable. Neither version of CC appears in the OED as of 2008. Google searches on November 2, 2008 provided the following numbers of results:

“core competence” and “core competences”:
 ~ 463,000 and ~ 122,000, respectively
 “core competency” and “core competencies”:
 ~ 754,000 and ~ 1,980,000, respectively

Hence, both terms are commonplace, but “competency” seems more prevalent.

ADVANTAGES AND DISADVANTAGES OF OUTSOURCING

Even if the term “outsourcing” might be fairly new, the actual practice is not. Because no organization can do everything itself, each one must choose a division of labor in every endeavor, defining its own roles and ceding any remaining duties to other parties. The key questions are which activities and to what extent.

Proponents commonly emphasize the outsourcing party’s resulting ability to focus on those activities deemed CCs for their strategic significance, as noted earlier. Converting some fixed costs to variable costs can increase financial and operational flexibility, and improve return on assets. Tax benefits may also accrue on moving certain activities to outside parties. Outsource service providers ostensibly enjoy superior cost

structures due to specialization and scale economies, and lower risk because they can balance the peaks in some customers' needs with valleys in others'. Some argue that outsiders provide better service with fewer headaches than would a company's own employees, as outsiders are easier to terminate and therefore ought to be more willing to please. But outsourcing need not be about replicating an existing function at lower cost or with improved quality. An outside party may offer transformative capabilities that are unavailable any other way [4].

Through outsourcing, firms risk eroding critical capabilities, institutional and tacit knowledge, and long-term relationships. Communication and coordination among internal and outsourced functions can be difficult and costly. Dependence leaves firms susceptible to service providers' underperformance, holding hostage of critical assets (like scarce parts or custom tooling), using their clients' product or process knowledge to benefit the firms' competitors, or even themselves becoming competitors. Outsourcing complicates decision making as power is distributed across a constellation of independently controlled firms whose relationships are shorter-term and more transactional. In many cases, the outcome has been disappointing [5–9].

Some of these difficulties of outsourcing result from the complexity, fragmented decision making, and broken information flows that come from decentralizing, which can be countered by process redesign and enhancement of information technologies. Others, however, reflect deliberate actions by service providers that are not in their clients' best interests. This possibility exists because of limitations in the client's ability to dictate and monitor the provider's actions (which are only exacerbated by any geographic or cultural separation). All this is particularly baffling for organizations whose institutional knowledge of the intricacies of the outsourced activity have been lost over time, or never existed in the first place [10].

Many processes conducted in-house also suffer from some variant of these challenges, but at least these play out under the auspices of the company's own internal checks and

balances. However, many companies equate outsourcing with reductions in resource and staff requirements, and fail to recognize that investments in business controls must actually increase to address the new risks. For some activities, properly overseeing the service provider may require such intimate involvement that the firm may be better off not outsourcing.

Many of the aforementioned costs and risks are manifestations of what economists classify as "transactions costs" (e.g., costs of search, contracting, negotiating, monitoring, and dealing with changes/disagreements), which are often invoked as a determinant of an industry's extent of vertical integration in the literature termed *transactions costs economics* (TCE) [11–13]. Constructs from Principal Agent (aka Agency) Theory, which focuses on relationships in which one party (the principal) delegates work to another (the agent), have been used to analyze these types of transactions costs. This framework highlights the "moral hazard" inherent in any relationship in which the principal's goals conflict with the agent's goals, and the principal has difficulty verifying the agent's actions (i.e., incomplete information).

INSOURCE-VERSUS-OUTSOURCE DECISION FRAMEWORKS

Defining which activities a firm should perform is among the most fundamental and profound of management duties, with consequences felt in every day of operation. Skill at making this decision is itself strategically critical enough to merit consideration as a CC [14,15].

"Make-versus-buy" is a traditional term for this challenge, and appears in the index of many business textbooks, especially in accounting, economics, and operations. To avoid the slight materials centrism in that term, this article will use "insource-versus-outsource" since many such evaluations concern the procurement of services rather than goods.

The academic and practitioner literatures overflow with commentaries on this topic. Most provide qualitative lists of issues to

consider or questions to ask, but leave to the decision maker any specific quantification of the multidimensional trade-offs (or authority to make a judgment call). This is not a criticism of the extant work, but an acknowledgment of the complexity and context-specificity of the problem. Here, we will simply sketch as an illustrative example one such insource-versus-outsource decision framework.

We earlier mentioned the high level strategic sound bite advocating for focusing on CCs and outsourcing everything else, a thread that can be traced through Refs 3 and 4. An oft-invoked variant of this is the notion of “core-versus-context” articulated by Geoffrey Moore of The Chasm Group [16], which has influenced strategy at firms like Cisco Systems. This defines “core” as those activities that differentiate a company in the marketplace and thereby drive the company stock’s valuation, whereas “context” is everything else the company does, and advises assigning the best people to the core while outsourcing as much of the context as possible.

An example of how these ideas might be operationalized is the seven-part framework of Quinn and Hilmer [4]:

1. Do we really want to produce the good or service internally in the long run? If we do, are we willing to make the back-up investments necessary to sustain a best-in-world position? Is it critical to defending our CC? If not,
2. Can we license technology or buy know-how that will let us be best on a continuing basis? If not,
3. Can we buy the item as an off-the-shelf product or service from a best-in-world supplier? Is this a viable long-term option as volume and complexity grow? If not,
4. Can we establish a joint development project with a knowledgeable supplier that ultimately will give us the capability to be best at this activity? If not,
5. Can we enter into a long-term development or purchase agreement that gives us a secure source of supply and a proprietary interest in knowledge or other property of vital interest to us and the supplier? If not,

6. Can we acquire and manage a best-in-world supplier to advantage? If not, can we set up a joint venture or partnership that avoids the shortcomings we see in each of the above? If so,
7. Can we establish controls and incentives that reduce total transaction costs below those of producing internally?

This set of questions implies a flowchart terminating in a spectrum of possible structures (“full ownership,” “partial ownership,” “joint development,” “retainer,” “long-term contract,” “call option,” and “short-term contract”) that exchange control (greatest with full ownership) for flexibility (greatest with short-term contract). A key message in this is that insource-versus-outsource is not a binary decision. For a single activity, a firm may even choose to outsource a portion while performing the rest in-house. This risk-mitigation strategy is sometimes termed *partial integration*, *taper(ed) integration*, or simply *make-and-buy* [17,18]. Furthermore, significant activities invariably contain many subtasks. A firm should consider various permutations of these that differ in divisions of labor, relationship lengths, and ownership of assets and liabilities.

Further complicating the matter is that the factors that drive the insource-versus-outsource decision are constantly in flux, so that the correct decision will be a moving target. Linder [1] and many others [19,20] emphasize the dependence on the stage in the life cycle of the company and industry. Even within a given industry, a particular set of environmental stimuli might elicit disparate responses from direct competitors. For instance, the recent global economic slowdown has directly led some consumer electronics firms to insource more production activities (to maintain utilization of existing in-house capacity) while others increased their outsourcing (to lower costs and achieve flexibility for responding to demand volatility).

Robust quantitative frameworks are elusive here for many of the reasons that apply to all complex managerial decisions with strategic impact. The individual consequences, such as a sharpening of

organizational focus or the atrophy of the knowledge and capabilities that are preserved only by regularly doing a task oneself, are very hard to translate into dollars and cents. Measuring the true cost of coordination across organization boundaries is also thorny. Certainly the contract delineates explicitly the transfer of funds, and salary impacts can be tallied. But how does one quantify an increase in the difficulty in communication? How does one measure the increased risk of opportunistic behavior by service providers, the possibilities of which are only limited by one's imagination? Existing accounting frameworks, which already struggle to assess the true cost of performing activities in-house, are stressed even further by outsourcing. McIvor [21] articulates an "overhead allocation fallacy" in standard cost accounting: when an activity is partially outsourced, certain overhead costs (which were not liquidated in the course of outsourcing) tend to be allocated to the activities that remain in-house, making those activities look even worse relative to outside alternatives. This can encourage further outsourcing and thereby perpetuate the fallacy.

ADVICE ON MANAGING THE OUTSOURCING RELATIONSHIP

The notion of best practices is largely idiosyncratic to the type of activity being considered for outsourcing, and attempting to unify all these context-specific details would go beyond the charter of this article. Here, we simply summarize the themes that overarch the extant body of academic and practitioner knowledge.

The obvious, yet profound, first concern is to carefully evaluate whether to outsource the particular activity at all. Outsourcing is a strategic action that must not be undertaken with the single-minded objective of reducing costs, or under the influence of herd mentality. The decision maker must be open to the possibility that outsourcing may actually increase overall costs, but might willingly proceed anyway if the structural change adds new capabilities or enhances existing ones. Contemplation of the full range

of pros and cons, such as those articulated in the section titled "Advantages and Disadvantages of Outsourcing", will affirm that outsourcing is no panacea. It is most prudently viewed as exchanging one set of headaches for another. Doig *et al.* [6] caution, "Don't assume that it is easier to manage suppliers than to improve your company's own performance."

After deciding to proceed, the outsourcing party must exercise caution and vigilance, which means due diligence on service providers up front, determining whether to single-source or multisource and the closeness of the resulting relationship(s), carefully writing specifications, contemplating and structurally addressing potential incentive conflicts, and installing appropriate monitoring mechanisms. In this spirit, Allen and Chandrashekar [22] and Aron and Singh [23] discourage outsourcing a process until it is well-understood and has coherent metrics. This can be harder to achieve for procured services than for procured materials, in part because the intangibility of what is being purchased complicates quality assessment and retrospective attribution of liability for problems [22]. Consequently, the outsourcing party must accept the need to invest resources (and maybe even add new headcount) in new control processes, which must be cost-justified based on the value delivered over the lifetime of the sourcing relationship. Priority shifts to skills such as relationship-building, negotiation, program and project management, and contract management. Peisch [24] points out that "Managing external resources requires an entirely different set of skills than managing the same services internally."

These challenges fall under the purview of the well-established discipline of purchasing and supply management. This community has established active professional organizations (e.g., the Institute for Supply Management (ISM), founded in 1915), certifications [e.g., the ISM's Certified Purchasing Manager (CPM) and Certified Professional in Supply Management (CPSM) credentials], university undergraduate and graduate degree programs, and a rich body of practitioner and academic literature (e.g.,

textbooks such as Monczka *et al.* [25] and numerous journals).

ORMS RESEARCH ON OUTSOURCING

The body of existing ORMS research on outsourcing is either too vast to survey in one article, or nascent, depending on one's definition of ORMS and the criteria used to determine what counts as work on outsourcing.

Mathematical modeling approaches popular among those who identify with the ORMS community are also used in other academic disciplines, including economics, accounting, and finance. All of these have studied issues relevant to the outsourcing decision. For reasons of tractability, the analytical work tends to focus on the trade-offs among a very small number of factors, primarily the ones easier to quantify. The limitations of this should be apparent from the preceding discussion. Questionnaire or interview-based descriptive surveying is a more popular format (many of this article's citations are of this sort), but this is not traditionally viewed as ORMS.

What counts as research on outsourcing? In the broadest sense, any model that includes a transaction between a supplier and a buyer firm could qualify. Such research has gone on for decades, and has generated thousands of publications. However, this author's position is that the outsourcing literature should be defined more narrowly as those works that consider the design, management, and control of an outsourcing relationship and give guidance about addressing some problem explicitly ascribable to the outsourcing.

A wish-list of specifications for an analytical, prescriptive research piece about outsourcing might include the following set of features, which does seem mathematically intractable:

- multiple parties: buyer, service provider, possibly a materials supplier, and competition for each;
- conflicting agendas, possibly also with internal conflict among agents within each firm;
- multi-attribute objective functions;

- private information that renders complete monitoring of the service provider impossible, so as to allow the possibility of deliberate deception;
- a cost model for buyer activities that reflects changes in organizational complexity, since outsourcing reduces complexity in some respects (in enabling focus on CCs) but increases it in others (for managing the service provider);
- institutional knowledge, since outsourcing jeopardizes the retention of this;
- power, since outsourcing creates dependence on outside parties.

Even this challenging list is not complete, since it does not address numerous other issues presented throughout this article, phenomena that are often difficult to quantify. Also, the best wisdom available is that firms must think of these factors strategically, and not be overly focused on short-term financial impact. This necessitates a longer-term (and more difficult to define) objective function.

This discussion should make clear why ORMS work that could truly be said to capture the essence of the outsourcing phenomenon is still sparse. Since this article was not meant to be a literature review, we will simply end here urging the reader to consider this as a roadmap to many important research opportunities.

ADDITIONAL READING

Due to publication restrictions, the bibliography for this article was limited to a small number of references. A fully annotated version containing more than 65 references is available for download at the author's university webpage.

REFERENCES

1. Linder JC. Transformational outsourcing. *Sloan Manage Rev* 2004;45(2):52–58.
2. Reinhardt A. Nokia's magnificent mobile-phone manufacturing machine. *Bus Week* 2006. Available at http://www.businessweek.com/globalbiz/content/aug2006/gb20060803_618811.htm.

3. Prahalad CK, Hamel G. The core competence of the corporation. *Harv Bus Rev* 1990;68(3):79–90.
4. Quinn JB, Hilmer FG. Strategic outsourcing. *Sloan Manage Rev* 1994;35(4):43–55.
5. Earl MJ. The risks of outsourcing IT. *Sloan Manage Rev* 1996;37(3):26–32.
6. Doig SJ, Ritter RC, Speckhals K, *et al.* Has outsourcing gone too far? *McKinsey Q* 2001;26(4):25–37.
7. Lakenan B, Boyd D, Frey E. Why Cisco fell: outsourcing and its perils. *Strategy Bus* 2001;Q3(24):54–65.
8. Barthelemy J. The seven deadly sins of outsourcing. *Acad Manage Exec* 2003;17(2):87–100.
9. Thurm S. Behind outsourcing: promise and pitfalls. *Wall St J* 2007;B3.
10. Anderson EG Jr, Parker GG. The effect of learning on the make/buy decision. *Prod Oper Manage* 2002;11(3):313–339.
11. Coase RH. The nature of the firm. *Economica* 1937;4(16):386–405.
12. Williamson OE. Markets and hierarchies: analysis and antitrust implications. New York: The Free Press; 1975.
13. McIvor R. How the transaction cost and resource-based theories of the firm inform outsourcing evaluation. *J Oper Manage* 2009;27(1):45–63.
14. Fine CH, Whitney DE. Is the make-buy decision process a core competence? In: Muffatto M, Pawar K, editors. *Logistics in the information age*. Padova: Servizi Grafici Editoriali; 1999. pp.31–63.
15. Gottfredson M, Puryear R, Phillips S. Strategic sourcing: from periphery to the core. *Harv Bus Rev* 2005;83(2):132–139.
16. Moore GA. Living on the fault line: managing for shareholder value in the age of the Internet. HarperBusiness, New York: Collins Business; 2000.
17. Porter ME. Competitive strategy: techniques for analyzing industries and competitors. New York: Free Press; 1980.
18. Harrigan KR. Formulating vertical integration strategies. *Acad Manage Rev* 1984;9(4):638–652.
19. Fine CH. Clockspeed: winning industry control in the age of temporary advantage. Reading (MA): Perseus Books; 1998.
20. Christensen CM, Raynor ME, Verlinden MC. Skate to where the money will be. *Harv Bus Rev* 2001;79(10):72–81.
21. McIvor R. A practical framework for understanding the outsourcing process. *Supply Chain Manage Int J* 2000;5(1):22–36.
22. Allen S, Chandrashekar A. Outsourcing services: the contract is just the beginning. *Bus Horiz* 2000;43(2):25–34.
23. Aron R, Singh JV. Getting offshoring right. *Harv Bus Rev* 2005;83(12):135–143.
24. Peisch R. When outsourcing goes awry. *Harv Bus Rev* 1995;73(3):24–37.
25. Monczka RM, Handfield RB, Giunipero LC, *et al.* Purchasing and supply chain management. Mason (OH): South-Western College/West; 2008.

BYELORUSSIAN OPERATIONAL RESEARCH SOCIETY (ByORS)

VALERY S. GORDON*
NIKOLAI N. GUSCHINSKY
United Institute of Informatics
Problems, National Academy of
Sciences of Belarus, Minsk,
Belarus

The Byelorussian Operational Research Society (ByORS) is a nonprofit scientific organization which supports application and promotes development of operations research (OR) methods in the Republic of Belarus. It unites experts in the fields of OR, applied mathematics, mathematical economy, and their applications. As a nationwide association, the ByORS also represents Belarus in the international network of experts in the field of operations research such as the Association of European Operational Research Societies (EURO) and the International Federation of Operational Research Societies (IFORS). The objectives of the BYORS are

- encouraging theoretical and applied research in the field of operational research in Belarus;
- promoting international connections and gaining the prestige of the Byelorussian scientists and experts who work in the field of OR.

To realize these objects, ByORS concentrates on the following:

- conducting scientific research and applied work in the field of OR;
- forecasting and promoting the new perspective branches of OR;

- conducting the independent scientific expertise of projects in the field of OR, formulating the reasonable recommendations for their practical application;
- dissemination of achievements in the field of OR by means of giving lectures, publishing books and articles, creating popular scientific programs, and using mass media;
- editorial activities in publishing the scientific, popular scientific literature, handbooks in the field of OR;
- organizing conferences, seminars, exhibitions, and meetings in the field of OR;
- encouraging the teaching of OR.

In order to achieve these objectives, the activities of the ByORS include editing and advertising; publishing the results of scientific research; organizing interconnections with scientific and industrial associations; promoting cooperation with firms, universities, and institutions; participating in international scientific congresses, conferences, symposia, meetings, exhibitions, and other events.

HISTORY AND MILESTONES

The ByORS was founded in November 1996 at the Institute of Engineering Cybernetics of the National Academy of Sciences of Belarus (NASB). Foundation of the ByORS was initiated by Vyacheslav S. Tanaev who was a full member of the NASB and the director of the United Institute of Informatics Problems (UIIP) of NASB. Tanaev became the first President of the ByORS and remained in this position until his death in the year 2002. What follows is a citation from the paper "Vyacheslav Tanaev: Contributions to Scheduling and Related Areas" by Valery S. Gordon, Mikhail Y. Kovalyov, Genrikh M. Levin, Yakov M. Shafransky, Yuri N. Sotskov, Vitaly A. Strusevich, and Alexander V. Tuzikov submitted to a special issue

*Deceased June 4, 2010.

of “Journal of Scheduling” (2010) which is dedicated to the memory of V.S. Tanaev on the occasion of his 70th birthday. Vyacheslav Tanaev (1940–2002) was born in village Akulovo, Tver region, Russian Federation. He received his Candidate of Sciences (PhD equivalent) degree for his work on scheduling from the Institute of Mathematics of the NASB in 1965. The degree of Doctor of Sciences (Habilitation Doctor) was awarded to V.S. Tanaev after a successful defense of the thesis on parametric decomposition of optimization problems at the Computer Center of the Academy of Sciences of the USSR, Moscow, in 1977. In 1987 he became the director of the UIIP of NASB, and in 2000 was elected a full member of NASB, which is the highest scientific rank in the states of the former Soviet Union. Scientific heritage of Vyacheslav Tanaev includes more than 130 research publications among which there are 10 monographs. His scientific interests included scheduling theory, discrete and continuous optimization, computer-aided design (CAD); he coordinated research in geo-information systems, development of supercomputers, and applications of informatics to medicine. He supervised 18 candidates of sciences among which 6 became Doctors of sciences. Vyacheslav Tanaev is the author of the first papers on scheduling in Russian, and his early research stimulated the study in this area in the Soviet Union and the countries of Eastern Europe. Several generations of Russian-speaking researchers benefited from becoming familiar with the major results on scheduling by studying his monographs.

From October 2002, Valery S. Gordon (Prof., Dr., Principal researcher of UIIP) served as president of ByORS. There are three vice presidents of ByORS: Prof. Dr. Vladimir A. Golovko from Brest State Technical University, Prof. Dr. Mikhail Y. Kovalyov (Deputy General Director of UIIP), and Dr. Genrikh M. Levin (Head of OR laboratory of UIIP). Dr. Nikolai N. Guschinsky is the secretary of ByORS.

MAIN ACTIVITY

The ByORS focuses on the following OR topics:

- theory and practice of optimization;
- scheduling theory;
- supply chain management;
- multicriteria optimization;
- graph optimization problems;
- decision support system problems;
- methods and software for production planning;
- methods for project management and scheduling;
- decomposition methods of optimization problems;
- synthesis of logical circuits;
- OR in manufacturing;
- OR in CAD\CAE\CAM;
- OR in finance;
- OR in health care.

Members of ByORS give courses and lectures on topics in OR at the Belarus State University (BSU) and the Belarus National Technical University as well as supervise postgraduate students in the field of OR at the BSU and the NASB. Members of ByORS conduct their scientific projects and publish papers in the field of OR, and they take part in scientific councils and juries on thesis defenses.

The ByORS organized the international workshop “Discrete optimization methods in scheduling and computer-aided design” (Minsk, Belarus, 2000), the first international workshop on “Discrete optimization methods in production and logistics” (Minsk, 2002), and took part in the organization of the Conference of the European Chapter on Combinatorial Optimization (ECCO) “Combinatorics for modern manufacturing, logistics, and supply chains” (ECCO XVIII, Minsk, 2005). Members of the ByORS also took part in organizing the second international workshop “Discrete optimization methods in production and logistics” (Omsk—Irkutsk, Russia, 2004), the international scientific

conference “Discrete mathematics, algebra and applications” (Minsk, 2009).

Beginning from 2003, ByORS has organized the biannual international conference “The Tanaev readings” in Minsk. The conference is dedicated to the memory of Vyacheslav Tanaev (the first president of ByORS) and brings together researchers, academicians, practitioners, and students interested in all branches of operational research. The topic of the first of these conferences was “Scheduling theory and decomposition methods,” and the scope of the conference was later broadened to include other branches of OR and information technology. The 4th international conference “The Tanaev readings” took place in Minsk in March 2010.

President of ByORS V.S. Gordon, Vice President M.Y. Kovalyov, and member Prof. V.A. Strusevich are the editors of a special issue “Scheduling: new branches, old roots” of the *Journal of Scheduling* which is dedicated to the memory of Vyacheslav Tanaev on the occasion of his 70th birthday.

Members of the ByORS took part in most global (IFORS) and European conferences in the field of OR, being Chairpersons of sessions and streams or members of program committees (in particular, president of ByORS Prof. V.S. Gordon was a member of the program committee of the 20th ECOR, Rhodes, Greece, July 4–7, 2004; Vice President Dr. G.M. Levin was a member of the Program Committee of the 13th IFAC symposium on information control problems in manufacturing INCOM 2009, Moscow, June 3–5, 2009; Vice President

Prof. M.Y. Kovalyov was a chairperson and V.S. Gordon was a cochairperson of the program and organizing committees of the international conference ECCO XVIII, Minsk, May 26–28, 2005; member of the ByORS Prof. A.B. Dolgui was a member of the advisory, scientific or program committees of several international conferences. He was also an international program committee chair of the symposiums INCOM 2006 (Saint Etienne, France, May 17–19, 2006) and INCOM 2009.

Members of ByORS serve in the editorial boards of the international journals: M.Y. Kovalyov and A.B. Dolgui are associate editors for “OMEGA—The International Journal of Management Science” and “Asia-Pacific Journal of Operational Research”; M.Y. Kovalyov is a member of the editorial advisory board for “European journal of operational research”; V.A. Strusevich and M.Y. Kovalyov are members of editorial advisory board for “Computers and operations research”, A.B. Dolgui is an area editor for the journal “Computers and industrial engineering”, member of the editorial board for the “International journal of production economics,” “International journal of systems sciences,” “Journal of mathematical modeling and algorithms,” and “Journal of decision systems.”

Members of the ByORS work at the universities, institutes, scientific and project organizations and firms in Minsk, Grodno, Brest, Gomel, and Mogilev (Belarus) as well as in London, Leeds (Great Britain), Saint Etienne (France), and Olsztyn (Poland).

CALCULATING REAL OPTION VALUES

MARCO ANTONIO GUIMARÃES DIAS
Department of Industrial Engineering,
PUC-Rio, Rio de Janeiro, Brazil
Financial Planning, Petrobras, Brazil

INTRODUCTION

A real option (RO) is the right, but not the obligation, an agent has for making decisions over a real asset. The agent can be a manager, a consumer, a social planner or any other decision maker. The real asset can be an investment project opportunity or an already existing asset like a factory. RO highlights the *flexibility* value, which is more valuable under conditions of uncertainty. RO can be viewed as a problem of *optimization under uncertainty*: maximize the real asset value by exercising optimally the relevant options subject to the uncertainties and physical and other constraints. Typically, RO considers the optimal *timing* to exercise the option(s). RO uses *dynamic programming* and *optimal control* concepts from *operations research* school, but RO incorporates concepts from modern finance, for example, asset pricing models must be free of *arbitrage* opportunity, which has an impact on discount rates and probabilities used in the quantitative model.

Myers [1] coined the term *real options* in the 1970s for the growth opportunities faced by firms, adapting the concept of financial options pioneered by Black and Scholes [2], and Merton [3]. The value of an RO is linked with the optimal decision (exercise rule). While *financial* options/derivatives theory is the *valuation* root of RO, the *decision* roots of RO are *decision analysis* [4] and *environmental economics* [5].

The first mathematical model appeared in 1979 [6], but textbooks on RO appeared only in the 1990s [7,8]. More than 50 books have

been published now (textbooks and edited books) on RO; books on corporate finance and valuation [9,10], as well as several books on derivatives [11,12], usually have one or more chapters on real options.

The traditional economic valuation approach for projects is the *discounted cash flow* (DCF) that uses a *risk-adjusted discount rate* (μ) to calculate the present value of *expected* cash flows, thus obtaining the (expected) *net present value* (NPV). If the NPV is positive, the project can be accepted. Let V be the value of the operating project and I the investment value so that

$$\text{NPV} = V - I. \quad (1)$$

In cash flow terms, V can be interpreted as the present value of expected operational cash flows (revenues net of operational costs and taxes), whereas I is the present value of the expected investment (net of tax benefits like depreciation).

Consider two *mutually exclusive alternatives* to invest in a project, investment at $t = 0$ with $\text{NPV}_0 = V(t = 0) - I$, and investment at one period later ($t = 1$) with $\text{NPV}_1 = e^{-\mu} E[V(t = 1) - I]$, where the first term is the discount factor in continuous time, V is stochastic, and I is deterministic. By the traditional DCF approach the value of this investment opportunity is

$$\begin{aligned} W &= \text{Max}[\text{NPV}_0, \text{NPV}_1, 0] \\ &= \text{Max}[V(t = 0) - I, e^{-\mu} E[V(t = 1) - I], 0]. \end{aligned} \quad (2)$$

Consider the following numbers: $V(t = 0) = 100$, $I = 100$, and two equiprobable scenarios for $V(t = 1)$: $V^+ = 120$ and $V^- = 80$. With these numbers in Equation (2), the DCF value is $W = 0$, because $\text{NPV}_0 = \text{NPV}_1 = 0$.

The same example under the RO approach shows a very different result. Consider $\mu = 10\%$. If we wait until $t = 1$ and consider the options (invest or not) in *each scenario* at $t = 1$ (instead $E[V(t = 1)]$), the RO value at

$t = 1$ is $F(t = 1) = E[\text{Max}\{V - I, 0\}] = 50\% \times \text{Max}\{V^+ - I, 0\} + 50\% \times \text{Max}\{V^- - I, 0\} = 0.5 \times 20 + 0.5 \times 0 = 10 > 0$. The RO value of this investment opportunity at $t = 0$ could be calculated as

$$\begin{aligned} F &= \text{Max}\{\text{NPV}_0, e^{-\mu} F(t = 1)\} \\ &= \text{Max}\{V(t = 0) - I, \\ &\quad e^{-\mu} E[\text{Max}\{V(t = 1) - I, 0\}]\}. \end{aligned} \quad (3)$$

Substituting the numbers, $F = 9.05 > W = 0$. The mathematical difference between Equations (2) and (3) is that the latter uses $E[\text{Max}\{.\}]$ whereas the former uses $\text{Max}\{E[.]\}$. This result is very general: by the *Jensen's inequality* if $f(x)$ is a strictly *convex function* and x is a random variable (rv), then

$$E[f(x)] > f(E[x]). \quad (4)$$

Note that $\text{Max}\{V(t = 1) - I, 0\}$ is a convex function in V , as in almost all RO cases. The difference between the two sides of this inequality generally increases with the variance of x . This explains the stylized fact that the option value increases with volatility. This intuitive discussion shows that optionality generates value and it does not depend on *how* an RO is calculated. This example is similar to examples from Dixit and Pindyck

[7, Chapter 2], but it is used only for incomplete markets. The section titled “The Two Main Derivatives Valuation Ideas” addresses the method of calculation using finance theory, for example, discussing the option discount rates problem. Incomplete market is discussed in the section titled “Other Real Options Issues and Concluding Remarks”.

The previous example is the *investment timing* option, but there are many types of ROs and many ways to classify the types of ROs. Trigeorgis [8, p. 145] divides the cases of *proprietary* (only one agent owns) versus *shared* RO, as well as the cases *simple* versus *compound* (option on another option) and *expiring* versus *deferrable* RO. We could also consider cases of *European* RO (only exercisable at the expiration) versus *American* RO (exercisable at any date including the expiration), *perpetual* (never expires) versus *finite-lived* RO. But the article follows a more practical RO classification to ease the identification of the project's relevant ROs and the selection of the valuation method. Figure 1 proposes this classification.

Figure 1 shows the three main types of ROs. The first group is related to the investment itself; the second is *after* the investment, and the third is the learning RO. Timing RO (or *option to wait* or *option to delay*) can be simple, as in the previous example (the general case is the classical

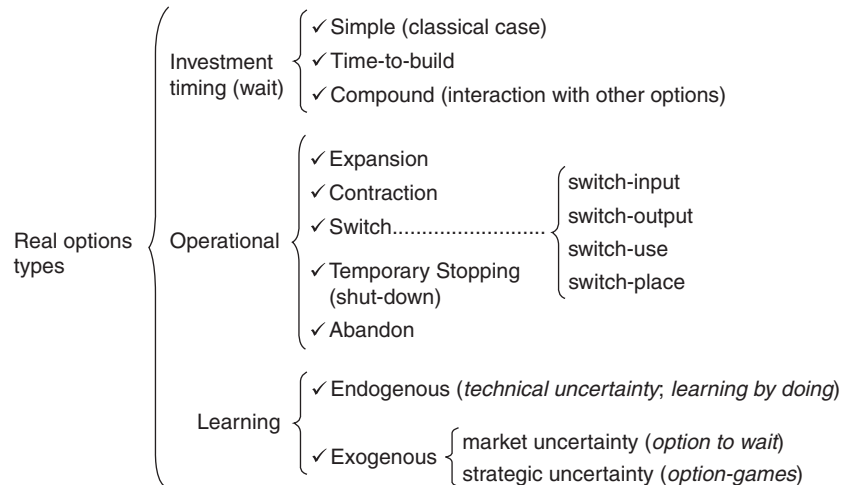


Figure 1. Classification of real options.

RO; see McDonald and Siegel [13] and the section titled “Continuous-Time Approach: Risk-Free Portfolio”), considering the time to build the project (used for applications with a long investment period before the revenues as in nuclear plants, see Majd and Pindyck [14] and Pindyck [15]) and compound options (e.g., timing to invest in a project with operational RO like abandonment). Operational RO includes expansion (with additional investment) and contraction (decreasing the operational cost), temporary and permanent (abandonment) stopping, and many switch ROs (input, output, use, and location). Learning RO generally is performed before considering industrial-scale investments and typically uses Bayesian methods. Some authors use the term *growth options* for operational-expansion options or for learning options (investment in R&D can be labeled as growth option or as learning option). Exogenous learning with market uncertainty is just another label for the option to wait. The case of strategic uncertainty about the type of an agent, for example, uncertainty about the behavior of a competitor in a market, can be modeled with Bayesian game theory combined with RO theory: a real option game model [16]. All shared ROs, for example, any two firms considering an entry into a new market (shared-timing options), can use option game models [17].

In many cases it is possible to design a business (with additional investment) to ease the exercise of operational options, for example, by acquiring a neighboring vacant land when investing in a new factory in order to make the exercise of the RO to expand production in case of high demand cheaper. RO quantification can be necessary to decide whether or not to make the additional investment to provide flexibility in a plant, for example, dual-fuel electricity generation (*switch-input* option) versus single-fuel generation. Although limited, petroleum refinery and GTL (gas-to-liquid) plants are examples of *switch-output* RO. One example of *switch-use* RO is an urban land with a house; the land can be redeveloped to build a hotel. Examples of *switch-place* RO are the mobile thermogenerators (typically ~ 100 MW) in containers, trucks or boats that are

used in remote regions or in regions with temporary shortage of electricity supply.

One interesting real-life switch RO case occurred in the Brazilian automobile industry. Owing to the petroleum price shocks in the 1970s, Brazil initiated the *ethanol-fuel automobile* production in the 1980s. But with the low petroleum prices in 1990s (dropping the fuel-ethanol prices), the owners of sugar mills preferred to make sugar instead of ethanol (*switch-output* RO), leaving the service station without ethanol for the customers. The ethanol car fabrication practically disappeared with the fall of the consumer confidence in ethanol automobiles. But in the 2000s, a new technology appeared in Brazil: the *flex-fuel car* using gasoline or ethanol (in some cases the natural gas as third fuel option). The flex-fuel car provides *switch-input* RO for the *consumer* so that they do not have to fear anymore about the producer switch-output (sugar–ethanol) option. By 2010, almost all automobiles sold in Brazilian market (>3 million per year) were flex-fuel. The best antidote to the producers’ switch option was the consumers’ switch option! The flex-fuel technology increased the consumer confidence and boosted the automobile market demand, leaving everybody better: ethanol producers, automobile producers, and the consumers. An RO that had a happy ending!

THE TWO MAIN DERIVATIVES VALUATION IDEAS

A very important concept for calculating the fair value of an asset is *arbitrage*, which is the possibility to get a riskless profit without investing money. Markets in equilibrium must be free of arbitrage opportunities as well as have good asset pricing models. Before the seminal derivatives articles [2,3] appeared, the key question was the option/derivatives risk-adjusted discount rate. While the risk of a derivative $F(V)$ is linked with the risk of the underlying asset V , clearly the risks are different (derivative can amplify or reduce the underlying risk) and so is the option’s risk-adjusted discount rate. The derivatives discount rate is a

complex problem, but the following idea bypasses this problem. If we can build a riskless portfolio with a certain combination of derivatives and basic assets, then the expected return (and so the discount rate) must be the risk-free discount rate; otherwise we can get an arbitrage opportunity [11,12].

The article now discusses the concepts of *riskless portfolio* and the *change of probability measure*, which constitute the two main derivatives calculation concepts underlying many RO valuation approaches.

Consider the following simple discrete-time options framework (Fig. 2), where the basic asset V is known at $t = 0$, but uncertain at $t = 1$. The aim is to calculate the fair value of the derivative $F(V)$ at $t = 0$. Figure 2a shows that the asset V (for example, the value of an operating factory) at $t = 1$ can rise to V^+ with probability p or can decrease to V^- with probability $(1 - p)$. In addition, V generates cash flow (analogous to “dividends” of financial assets) with values c^+ and c^- in each state at $t = 1$. The derivative $F(V)$ is a function of V . Hence, given the future values V^+ and V^- , F^+ and F^- are known. The unknown is the *present value* of F because we do not know its discount rate. With V, V^+, V^-, c^+, c^- we can calculate the total return μ of the underlying asset V , which (capital asset pricing model, CAPM) in equilibrium is also the risk-adjusted discount rate for V , but this discount rate of $F(V)$ is not the same and it is a complex problem. The initial idea to solve

this problem is by building a riskless portfolio Φ so that (by nonarbitrage argument) the return must be the risk-free discount rate r . Figure 2b proposes this portfolio:

$$\Phi = F - nV, \quad (5)$$

where n is chosen so that the portfolio is riskless, that is, the value at $t = 1$ is the same: $\Phi^+ = \Phi^-$. The value of n is known as *delta-hedge* and is written as

$$n = \frac{\Delta F}{\Delta V} = \frac{F^+ - F^-}{(V^+ + c^+) - (V^- + c^-)}. \quad (6)$$

By substituting n in Φ^+ and Φ^- (see Fig. 2b), they are equal (riskless portfolio) as desired.

$$\Phi^+ = \Phi^- = \frac{F^-(V^+ + c^+) - F^+(V^- + c^-)}{(V^+ + c^+) - (V^- + c^-)}. \quad (7)$$

Letting $u = \text{upside} = (V^+ + c^+)/V$ and $d = \text{downside} = (V^- + c^-)/V$, we can rewrite Equation (7) as

$$\Phi^+ = \Phi^- = \frac{uF^- - dF^+}{u - d}. \quad (8)$$

So, the present value of this *riskless* portfolio is easily calculated with the risk-free discount rate r :

$$\Phi(t=0) = \frac{uF^- - dF^+}{(u - d)(1 + r)}. \quad (9)$$

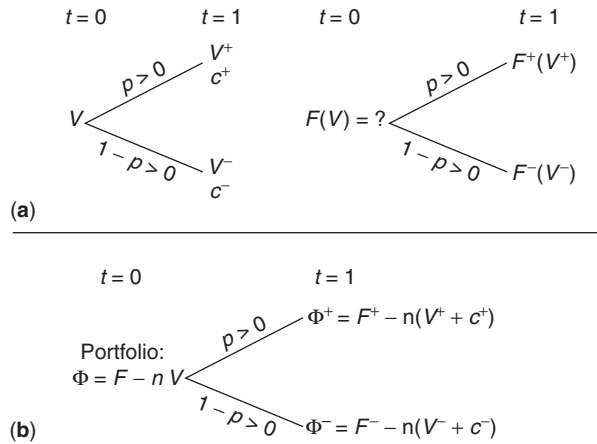


Figure 2. Discrete-time options and the risk-free portfolio.

Because we know all the inputs of Equation (9), we know $\Phi(t=0)$. Hence, we can calculate the derivative $F(t=0)$ by using Equation (5):

$$F(t=0) = \frac{uF^- - dF^+}{(u-d)(1+r)} + nV.$$

Using Equation (6) and the definition of u and d , the derivative value can be calculated with

$$F(t=0) = \frac{uF^- - dF^+}{(u-d)(1+r)} + \frac{F^+ - F^-}{u-d}. \quad (10)$$

It is the first way to calculate the derivative value without requiring the derivative discount rate. It uses the concept of nonarbitrage. The riskless portfolio is like a riskless bond, $\Phi(t=0) = B$, and this suggests that we can replicate a derivative with a bond B and n underlying assets V by rewriting Equation (5) as

$$F = B + nV. \quad (11)$$

The analysis above can be generalized for several underlying assets. For example, in the case of two assets V and W , the risk-free portfolio will be $\Phi = F(V, W) - nV - mW$, where $n = \Delta F / \Delta V$ and $m = \Delta F / \Delta W$.

The second calculation concept is the change of probability: the total return on V under *real* probability measure P (here p and $1-p$) is μ , but we can change the probability measure in a way that instead of μ , the *total* return (including the cash flows or dividends) is the risk-free return r . This probability measure Q (here q and $1-q$) is named *risk-neutral probability* or *equivalent martingale measure*. The risk-free return under Q is

$$\begin{aligned} r &= \frac{E^Q[V(t=1) + c(t=1)]}{V(t=0)} - 1 \\ &= \frac{q(V^+ + c^+) + (1-q)(V^- + c^-)}{V} - 1. \end{aligned}$$

Using the definitions of u and d , we get the risk-neutral probability q for the asset V :

$$q = \frac{1+r-d}{u-d}. \quad (12)$$

The key issue is that the risk-neutral probability for V also makes the return of the derivative or function $F(V)$ be the risk-free return r . In order to see this, let q' and $(1-q')$ be the probabilities that make the derivatives' rate of return equal to r , that is,

$$r = \frac{q'F^+ + (1-q')F^-}{F(t=0)} - 1. \quad (13)$$

Using the replication portfolio, that is, Equation (11) for F and equations from Fig. 2b for F^+ and F^- in Equation (13), we get the risk-neutral probability q' for the derivative $F(V)$:

$$q' = \frac{1+r-d}{u-d}, \quad (14)$$

which is exactly the same risk-neutral probability of V , that is, $q' = q$. Hence, the probability measure that makes the underlying asset V return the risk-free rate r is the same probability that makes the function $F(V)$ return the risk-free rate r . Hence, under risk-neutral probabilities q and $1-q$, the correct discount rate for $F(V)$ is also r . Hence, there is another way to calculate the present value of the derivative $F(V)$:

$$F(t=0) = \frac{qF^+ + (1-q)F^-}{(1+r)}. \quad (15)$$

Note that the risk-neutral probability q depends only on the dynamics of the underlying asset V (information about the dynamics of F is redundant and not required). In many cases, this method is easier to apply than the riskless portfolio presented before. On the basis of this risk-neutral approach, the method named *binomial* was developed [18], with a recombining tree in order to solve multiperiod derivatives valuation (Fig. 3). The tree is solved *backward*, as in *dynamic programming*, where each node compares the present value of waiting (using Eq. 15) with the immediate exercise payoff (e.g., Eq. 1), rolling back until the initial date. For a bivariate binomial application with another model for the uncertainties (mean-reversion), see Bastian-Pinto *et al.* [19]. Other *lattices* with similar approaches are also used in ROs, for example, *pentanomial* [20]. The binomial/lattice approach

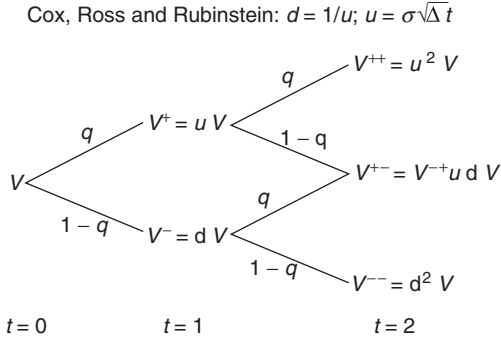


Figure 3. Recombining tree: the binomial diffusion.

is popular because it is more intuitive than continuous-time approach, but it is not good for perpetual ROs (differential equation or integral methods that get closed-form solutions in most cases are preferable, see the sections titled “Continuous-Time Approach: Risk-Free Portfolio” and “Continuous-Time Approach: Change Of Probability Approaches”), and to model three or more sources of uncertainties (Monte Carlo simulation is preferable, see the section titled “Continuous-Time Approach: Change Of Probability Approaches”).

These ideas allow us to calculate present values of derivatives, including RO. For example, a manager has the option to invest I in $t = 0$ or in $t = 1$ or not to invest. The NPV of immediate investment is $V(t = 0) - I$. Consider the investment deterministic, but the underlying asset value V is stochastic (as in Fig. 2). The values of F^+ and F^- are

$$F^+ = \text{Max}[V^+ - I, 0], \quad (16)$$

$$F^- = \text{Max}[V^- - I, 0]. \quad (17)$$

The present value of waiting can be calculated with either Equation (10) or Equation (15). The RO value is simply the maximum value between the immediate exercise and the present value of waiting:

$$\text{RO value} = \text{Max}[V - I, F(t = 0)]. \quad (18)$$

Note that $F(t = 0)$ will always be nonnegative, so that in many cases a (small) strictly

positive NPV is not enough to decide on investing immediately because the waiting value is higher. It is not difficult to see, because of the discount factor < 1 , that there exists a sufficiently high current value of V , say V^* , for which we are *indifferent* between waiting and immediate option exercise, that is, $F(t = 0) = V^* - I$. This suggests an optimal investment rule: invest immediately if $V(t = 0) \geq V^*$. This critical value V^* is known as *threshold* and is a very important RO concept. For the more general case of multiple periods we can calculate a threshold *curve* $V^*(t)$. The section titled “Continuous-Time Approach: Risk-Free Portfolio” presents a continuous-time framework with a threshold curve for finite-lived RO.

Most RO valuation methods are based on the two ideas presented above that can be generalized for the continuous-time framework.

Figure 4 presents the main RO solution methods. The risk-free portfolio in discrete time was presented above. The next section presents the continuous-time case, which results in a differential equation that describes the arbitrage-free relation between the derivative and its underlying asset. From the change of probability concept we have, in addition to the binomial approach presented above, many other approaches. One powerful method using this concept is the risk-neutral Monte Carlo simulation of the underlying assets’ stochastic processes together with an optimal decision rule, which is useful for the case of many stochastic variables (see the section titled “Continuous-Time Approach: Change Of Probability Approaches”). Dynamic programming under uncertainty but using the risk-neutral expectation operator and discounting with risk-free rate is another approach, which in continuous-time results in the same differential equation (see the section titled “Continuous-Time Approach: Risk-Free Portfolio”). The integral method is based on the first moment that a risk-neutral stochastic process hits a threshold for optimal investment and is particularly useful for perpetual options (at the end of the section titled “Continuous-Time Approach: Risk-Free Portfolio”).

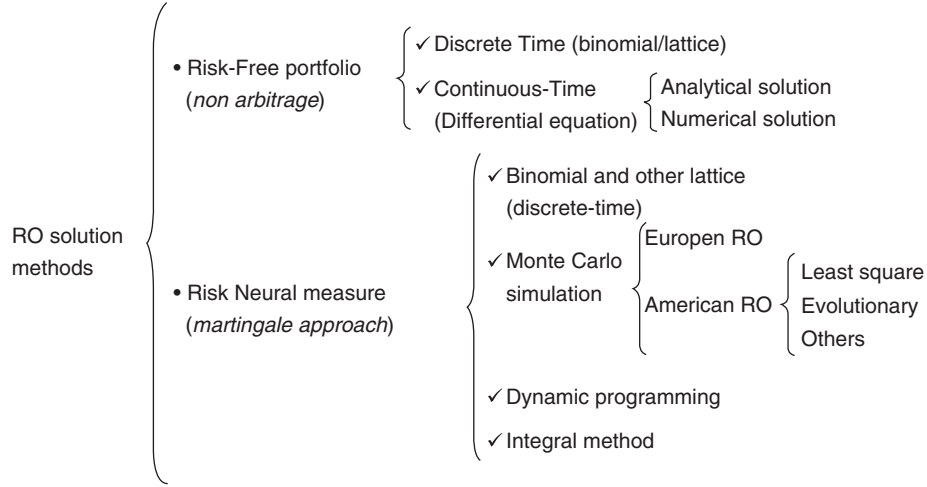


Figure 4. Real option solution methods.

CONTINUOUS-TIME APPROACH: RISK-FREE PORTFOLIO

Now, consider a more general mapping of uncertainties described by a *continuous-time stochastic process*. Let V be the value of a project that follows an *Itô process* described by the differential equation

$$dV = a(V, t) dt + b(V, t) dz, \quad (19)$$

where $a(V, t)$ is the drift of this process (function of V and time), $b(V, t) dz$ is the random term, and dz is the *Wiener increment* denoted by

$$dz = N(0, 1)\sqrt{dt}, \quad (20)$$

where $N(0, 1)$ is the standard normal distribution. The most popular stochastic process used in ROs and in finance is the geometric Brownian motion (GBM), denoted by

$$dV = \alpha V dt + \sigma V dz. \quad (21)$$

Hence, it is an Itô process with $a(V, t) = \alpha V$ and $b(V, t) = \sigma V$, where α is the exponential growth rate of V (the capital gain) and σ is the volatility, which is the standard deviation of dV/V . In Equation (19), V is the (known) current asset value, but dV is stochastic.

Another important particular case used in many cases to model commodity prices is the *mean reverting process* that has some variations, and one of them is

$$dV = \eta (\bar{V} - V) dt + \sigma V dz, \quad (22)$$

where η is the reversion speed and $\bar{V}$ is the long-run equilibrium level of V . In some applications, the Itô process is combined with a Poisson (jump) process, as discussed in Dias and Rocha [21].

The main *stochastic calculus* tool used in RO is the *Itô Lemma* [7, Chapter 3]. For the case of two state variables, V and t , where V follows Equation (21), the Itô Lemma for $F(V, t)$ is as follows:

$$dF = \frac{\partial F}{\partial V} dV + \frac{1}{2} \frac{\partial^2 F}{\partial V^2} (dV)^2 + \frac{\partial F}{\partial t} dt. \quad (23)$$

Because $(dz)^2 = dt$ [7, Chapter 3], $(dV)^2$ is not negligible if V is stochastic. For $n > 1$ rv, see Dixit and Pindyck [7, Chapter 3]. The second term in the right side quantifies the Jensen's inequality effect (if F is convex, this term is positive). If V follows Equation (19)

$$(dV)^2 = [b(V, t)]^2 dt. \quad (24)$$

Substituting Equations (19) and (24) into Equation (23)

$$dF = \left[a(V, t) \frac{\partial F}{\partial V} + \frac{1}{2} b^2(V, t) \frac{\partial^2 F}{\partial V^2} + \frac{\partial F}{\partial t} \right] dt + b(V, t) \frac{\partial F}{\partial V} dz. \quad (25)$$

Note that the partial derivatives are deterministic (V is known at t), and only dF and dz are stochastic. Equation (25) shows that $F(V, t)$ follows another Itô process:

$$dF = g(V, t) dt + h(V, t) dz. \quad (26)$$

Now, it is easy to prove that the value of n that makes the portfolio $\Phi = F - nV$ riskless is the partial derivative $\partial F / \partial V$. Consider the more general case where both the derivative $F(V, t)$ and the underlying asset V generate cash flows (dividends). In an infinitesimal time interval dt , the variations (capital gain plus dividends) for the portfolio components F and V are $dF + \delta_F F dt$ and $dV + \delta_V V dt$, respectively (δ_X is named *dividend yield* of X) so that the portfolio variation in dt is

$$d\Phi = dF + \delta_F F dt - n(dV + \delta_V V dt). \quad (27)$$

Substituting Equations (19) and (26) into Equation (27)

$$d\Phi = [k(V, t) - nc(V, t)] dt + [h(V, t) - nb(V, t)] dz, \quad (28)$$

where $k(V, t) = g(V, t) + \delta_F F$ and $c(V, t) = a(V, t) + \delta_V V$. Riskless portfolio means nonrandom $d\Phi$. Hence, the second term of Equation (28) must be zero. The value of n that makes the random term zero is $n = h(V, t) / b(V, t)$. But $h(V, t) = b(V, t) \partial F / \partial V$ (compare Eqs 25 and 26). So, the n that makes the portfolio Φ riskless is

$$n = \frac{\partial F}{\partial V}. \quad (29)$$

This result is analogous to the discrete-time result (in Eq. 6 we allow stochastic dividends for V).

If the portfolio is riskless with this choice of n , the portfolio return must be the risk-free

interest rate r because of nonarbitrage requirement. Hence, we can generate the nonarbitrage relation between F and V illustrated with the following example that results in the classical Black–Scholes–Merton *partial differential equation* (PDE).

Let V be the underlying asset that follows a GBM (Eq. 21) with dividend yield $\delta > 0$. Let $F(V, t)$ be a derivative that does not generate cash flow. With the convenient choice of n , the portfolio return is risk-free and equal to $r \Phi dt$ in the interval dt . But the portfolio return is also the variations of its components, so that we can write

$$r(F - nV) dt = dF - n(dV + \delta V dt). \quad (30)$$

The value of dF is given by the Itô Lemma (Eq. 23), where $(dV)^2 = \sigma^2 V^2 dt$ for GBM. Hence,

$$dF = \frac{\partial F}{\partial V} dV + \frac{1}{2} \sigma^2 V^2 \frac{\partial^2 F}{\partial V^2} dt + \frac{\partial F}{\partial t} dt. \quad (31)$$

Substituting Equations (31) and (29) into Equation (30) and rearranging, we get the classical PDE:

$$\frac{1}{2} \sigma^2 V^2 \frac{\partial^2 F}{\partial V^2} + (r - \delta) V \frac{\partial F}{\partial V} - rF + \frac{\partial F}{\partial t} = 0. \quad (32)$$

This PDE just gives the *nonarbitrage relation between F and V* . We do not yet say if F is an option, if it is a call or a put, if it is an American or European option, and so on. The *boundary conditions* make the case more specific.

Now, consider that V is the value of an operating project that can be obtained if the manager invests an amount I (deterministic). This RO is like a *financial call option*. Let T be the expiration of this RO (e.g., in petroleum sector, there is generally a last legal date that the oil company can develop a discovered oilfield), but the RO can be exercised at any date until T (American RO). This specific case

demands four boundary conditions (bc):

$$\text{If } V = 0, F(0, t) = 0, \quad (33)$$

$$\text{If } t = T, F(V, T) = \text{Max}[V - I, 0], \quad (34)$$

$$\text{If } V = V^*, F(V^*, t) = V^* - I, \quad (35)$$

$$\text{If } V = V^*, \left. \frac{\partial F(V, t)}{\partial V} \right|_{V=V^*} = \left. \frac{\partial(V - I)}{\partial V} \right|_{V=V^*} = 1. \quad (36)$$

Equation (33) is the *trivial* bc: if $V = 0$, Equation (21) shows that $dV = 0$, that is, V remains at zero forever (*absorbing barrier*) so that the right to invest $I > 0$ is worthless. Equation (34) shows that at expiration it is optimal to invest only with positive NPV (see Eqs 16 and 17 for the discrete-time case). Equation (34) is known as *value-matching* bc and shows that at any time it is optimal to invest at the threshold level V^* , which is defined as the indifference point where the waiting value F is equal (matches) to the exercise payoff $V - I$ (see discussion after Eq. 18). Equation (36) is known as *smooth-pasting* (or *high-contact*) condition and shows that not only the waiting value (F) and the exercise payoff (here $V - I$) values but also the slopes of the two functions match at V^* . The smooth-pasting is a *sufficient* condition (but not *necessary*) for the optimal option exercise (see Brekke and Oksendal [22] for a general proof).

The PDE (Eq. 32) and its bc (Eqs 33–36) are solved with numerical methods like *finite-differences* or with *analytical approximations*

[23] to calculate both the RO value $F(V, t)$ and the optimal exercise rule $V^*(t)$.

Figure 5 shows the $F(V, t)$ curve for some numerical parameters ($I = \$100$ million; $r = \delta = 4\%$ pa; $\sigma = 25\%$ pa) 2 years ($\tau = T - t = 2$) before expiration (dotted line) and at expiration ($t = T$, continuous line). Note the smooth-pasting contact of the dotted line with the exercise payoff line at V^* .

Figure 6 shows the threshold curve $V^*(t)$ for the same case. Note that at expiration $V^*(T) = I$, whereas 2 years before the expiration we have $V^*(t = 0) = \$155$ million $> I$ (shown also in Fig. 5).

This RO is a finite-lived American call option and was used in Paddock *et al.* [24] for oilfield development decision, where there is a legal period to develop the oilfield. But, in many cases the RO is *perpetual*, for example, the option to develop an urban land, where $F(V)$ is the land value and V is the house value. In this perpetual case, the partial derivative $\partial F / \partial t = 0$ so that the PDE (Eq. 32) simplifies to an *ordinary differential equation* (ODE):

$$\frac{1}{2} \sigma^2 V^2 \frac{\partial^2 F}{\partial V^2} + (r - \delta) V \frac{\partial F}{\partial V} - rF = 0. \quad (37)$$

In this perpetual option case, the threshold V^* is not a function of time anymore. Because there is no expiration, it remains valid only in Equations (33), (35), and (36) as boundary conditions. This ODE is *homogeneous* (all terms with F or their partial derivatives) and

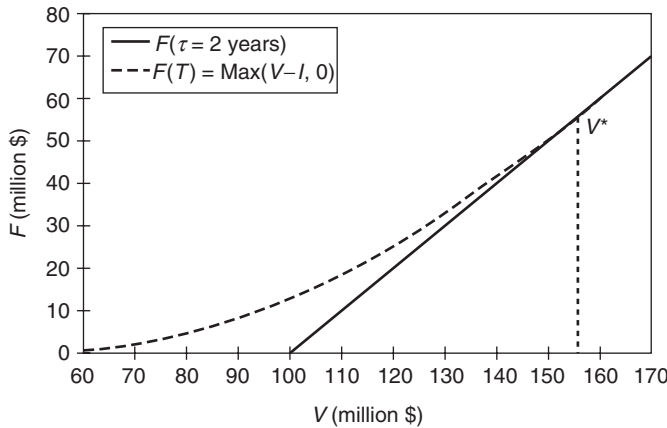


Figure 5. Real option values $F(V)$.

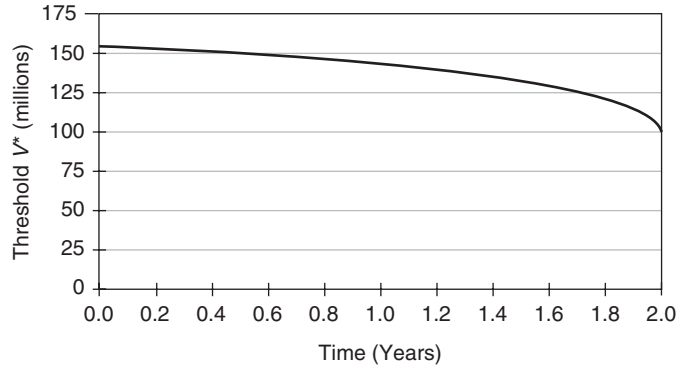


Figure 6. Optimal exercise threshold curve $V^*(t)$.

has *analytical solution* of the type

$$F(V) = AV^\beta, \quad (38)$$

where A is a constant to be determined. Substituting Equation (38) and its derivatives into Equation (37), we get a quadratic equation for β , which has two roots:

$$\beta_1 = \frac{1}{2} - \frac{r - \delta}{\sigma^2} + \sqrt{\left(\frac{r - \delta}{\sigma^2} - \frac{1}{2}\right)^2 + \frac{2r}{\sigma^2}}, \quad (39)$$

$$\beta_2 = \frac{1}{2} - \frac{r - \delta}{\sigma^2} - \sqrt{\left(\frac{r - \delta}{\sigma^2} - \frac{1}{2}\right)^2 + \frac{2r}{\sigma^2}}, \quad (40)$$

where $\beta_1 > 1$ and $\beta_2 < 0$ [7, pp. 142–144]. With two roots, the solution is the linear combination

$$F(V) = A_1 V^{\beta_1} + A_2 V^{\beta_2}. \quad (41)$$

But due to Equation (33) the constant $A_2 = 0$ because with $\beta_2 < 0$ when $V \rightarrow 0$ the term in A_2 goes to infinity if $A_2 \neq 0$. Hence,

$$F(V) = A_1 V^{\beta_1} \quad (42)$$

The two unknowns that remain are the constant A_1 and the threshold V^* . We get these values by applying the solution of Equation (42) into Equation (35) and (36). This set of

two equations with two unknowns results in

$$V^* = \frac{\beta_1}{\beta_1 - 1} I, \quad (43)$$

$$A_1 = \frac{V^* - I}{(V^*)^{\beta_1}}. \quad (44)$$

Equations (42)–(44), and (39) solve this perpetual RO problem completely.

CONTINUOUS-TIME APPROACH: CHANGE OF PROBABILITY APPROACHES

This section discusses the application of change of probability measure for the continuous-time case. It uses *risk-neutral stochastic processes* for the underlying assets. The risk-adjusted discount rate μ is the risk-free interest rate plus a risk premium π (given generally by the CAPM):

$$\mu = r + \pi. \quad (45)$$

Finance theory says that in equilibrium μ is also the total expected rate of return, which is the sum of the expected capital gain rate (α) and the expected dividend yield (or cash flow yield) δ

$$\mu = \alpha + \delta. \quad (46)$$

Equating Equations (45) and (46) we get

$$\alpha - \pi = r - \delta. \quad (47)$$

While in GBM (Eq. 21) α is the drift, the left side of Equation (47) presents a drift

penalized by a risk premium. This penalized drift (as well as the right side $r - \delta$) is known as *risk-neutral drift* or *drift under martingale measure* Q . In order to give more intuition, we could substitute α in Equation (21) by Equation (46):

$$dV = (\mu - \delta)V dt + \sigma V dz. \quad (48)$$

Equation (48) is under *real* probability measure P . It may be recalled that the measure Q is so that the total return of V changes to the risk-free return r (replacing the total return μ). Hence, the GBM under risk-neutral measure Q is simply

$$dV = (r - \delta)V dt + \sigma V dz \quad (49)$$

or

$$dV = (\alpha - \pi)V dt + \sigma V dz. \quad (50)$$

Hence, the risk-neutral GBM is the same but with the drift penalized with a risk premium. Risk premium is subtracted from the drift also when building any other risk-neutral Itô process (Eq. 19). A more rigorous proof uses the Girsanov's theorem [25].

The *dynamic programming* (DP) approach breaks the sequence of decisions into two components, the *immediate decision* (e.g., invest or wait) and a *value function* that considers the consequences of all posterior decisions. Typical RO applications use the binary version of DP named *optimal-stopping*, where “stopping” is the exercise of an RO (stop the “wait and see” policy).

DP in continuous time is well discussed in [7, Chapter 4], but these authors use DP for incomplete markets using an exogenous discount rate. Here, we discount with risk-free rate r but use the probability measure Q .

Consider the GBM under Q for the underlying asset V (Eq. 49). We can extend the discrete-time result (see the titled section “The Two Main Derivatives Valuation Ideas”) to continuous time by showing that under Q the correct discount rate for the RO $F(V)$ is r . But initially let μ_F be the discount rate for F under Q . For a small time interval, we

can write $e^{-\mu_F dt} \cong (1 + \mu_F dt)^{-1}$. The waiting (or *continuation*) value is

$$\begin{aligned} F(V, t) &= \frac{1}{1 + \mu_F dt} E^Q [F(V + dV, t + dt) | V(t)] \\ &= \frac{1}{1 + \mu_F dt} \left\{ F(V, t) + E^Q [dF(V, t)] \right\}, \quad (51) \\ &\Rightarrow F(V, t) \mu_F dt = E^Q [dF(V, t)], \quad (52) \end{aligned}$$

where dF can be obtained with the Itô Lemma, but for dV following a GBM under Q measure. Substituting Equation (49) for dV into Equation (23), taking expectations under Q , and noting that $E[dz] = 0$ we get

$$E^Q [dF] = \left\{ (r - \delta)V \frac{\partial F}{\partial V} + \frac{1}{2} \sigma^2 V^2 \frac{\partial^2 F}{\partial V^2} + \frac{\partial F}{\partial t} \right\} dt \quad (53)$$

Substituting Equation (53) into Equation (52) and rearranging, we get

$$\frac{1}{2} \sigma^2 V^2 \frac{\partial^2 F}{\partial V^2} + (r - \delta)V \frac{\partial F}{\partial V} - \mu_F F + \frac{\partial F}{\partial t} = 0. \quad (54)$$

But we know that Equation (32) gives the nonarbitrage relation between F and V . Comparing Equation (54) with Equation (32), in order to get an arbitrage-free price for F , the correct discount rate for F under Q is the risk-free discount rate, that is, $\mu_F = r$. This analysis extends the result obtained in discrete time (recall discussion on Eq. 14).

Another way to solve the perpetual RO problem is by the integral method. Suppose $P(t)$ is the price of a commodity that follows a GBM. Assume that the unitary operational cost C and the investment I to get a plant producing $Q(t)$ are deterministic. Let t^* be the first time that $P(t)$ hits the threshold P^* where the option exercise is optimal. The RO value can be written as

$$\begin{aligned} F(t = 0) &= E^Q \left[\int_{t^*}^{\infty} e^{-rt} (P(t) - C) Q(t) dt \right] \\ &\quad - E^Q [e^{-rt^*}] I. \quad (55) \end{aligned}$$

The integral method has been used in *real option games* [7, pp. 309–314] and in general perpetual options [26]. It uses conventional optimization (below) and some results like (for proof, see Dixit and Pindyck [7, pp. 315–316])

$$\mathbb{E}^Q \left[e^{-rt^*} \right] = \left(\frac{P}{P^*} \right)^{\beta_1}, \quad (56)$$

$$\mathbb{E}^Q \left[\int_0^{t^*} e^{-rt} P(t) dt \right] = \frac{P}{\delta} \left[1 - \left(\frac{P}{P^*} \right)^{\beta_1 - 1} \right], \quad (57)$$

where P follows a GBM. Equation (57) is useful for solving Equation (55) because the integral with integration limits from t^* to ∞ can be written as the difference of two integrals, one with integration limits from 0 to ∞ (it is a perpetuity with growth rate $= \alpha$ and discount rate $= r$) and other with limits from 0 to t^* . The dividend yield for commodities is interpreted as convenience yield and estimated from future markets [12, pp. 182–184].

The classical perpetual RO case of investing I to get the project with value V uses Equation (56) (with V instead P) and the first-order condition (FOC: $\partial F / \partial V = 0$ at V^*) to solve the maximization problem

$$F(V) = \max_{V=V^*} \mathbb{E}^Q \left[e^{-rt} (V - I) \right]. \quad (58)$$

The optimal V^* is nonrandom, but given V^* , t^* is random, so

$$F(V) = \mathbb{E}^Q \left[e^{-rt^*} \right] (V^* - I). \quad (59)$$

There is a trade-off in choosing V^* : if we choose a very high V^* , $(V^* - I)$ will be large, but it will take much time to hit V^* so that $\mathbb{E}^Q[\cdot]$ will be small; whereas if we choose a small V^* , $\mathbb{E}^Q[\cdot]$ will be large, but $(V^* - I)$ will be small. With the FOC in Equation (59) we get V^* [26]. With Equation (59) we get $F(V)$, which is the same expression obtained before (Eqs 38 and 44).

A practical approach for RO using the change of measure concept is the *Monte Carlo simulation* (MCS) of risk-neutral stochastic processes combined with an optimal decision rule. For European options it is simple,

because the rule is easy to set, for example, $F(T) = \text{Max}[V(T) - I, 0]$. The European RO value is

$$F(t=0) = e^{-rT} \mathbb{E}^Q [\text{Max}[V(T) - I, 0] | V(0)]. \quad (60)$$

By simulating the risk-neutral process for V at $t = T$, starting from $V(0)$ at $t = 0$, and using Equation (60), we get the RO value $F(t=0)$. MCS is useful mainly for the case of *many stochastic variables* following correlated stochastic processes, which are easy to simulate but hard to solve using differential equations or lattice methods (the *curse of dimensionality*).

Many ROs can be viewed as a *sequence of European options*. For example, a *multivegetable biodiesel plant* has the option to use soybean, cotton, castorbean, pinion, or palm at each quarter. We simulate the correlated risk-neutral stochastic processes for these vegetable prices and at each quarter we choose the vegetable that provides the maximum profit in that quarter (including the option to stop production). At each quarter, one European option of this kind is expiring. Because the switch-cost is negligible, we can value this plant as the sum of a sequence of European options.

MCS can also be used for *American options*. But the optimal decision rule is not simple as in the European case because of the earlier exercise possibility. In cases with more than three stochastic variables, it can be the best practical way to get the RO value. Since the 1990s some methods for using MCS with American options have appeared. One very popular method is the use of *least square approach* [27] to estimate the waiting (continuation) value, so that this value is compared with the immediate exercise at each t .

Another way is the *evolutionary real option approach* [28], for example, by using genetic algorithm to evolve a population of optimal decision rules (e.g., a threshold curve) and using risk-neutral simulations (and discounting with r) to evaluate the decision rules. The more valuable decision rules are selected as parents of a new generation of decision rules, which are again evaluated with MCS. When the gain in the last

generation for the best rule is small, the program stops and we get a *near optimal* solution with the associated RO value. This approach is computationally very time consuming, but its simplicity makes it a good practical alternative for future complex problems.

OTHER REAL OPTIONS, ISSUES, AND CONCLUDING REMARKS

There are many open questions that demand additional research in ROs. Some of them are listed here.

The effect of strategic iterations of competition and cooperation, when the exercise of an RO by one agent has an impact over the RO of another agent, is an important topic requiring models that integrate RO theory with game theory [16,29,30].

Most RO models assume *complete markets* at least as an approximation. When the market is complete, the martingale measure Q is unique (*fundamental theorem of asset pricing*: [31, Section 19-2]), but when the market is *incomplete*, there are *multiple martingale measures (range)* that can be used for pricing without producing arbitrage opportunities. In this case, it is necessary to select a measure Q from this range, and some *preference theory*, like the investor's *expected utility*, is used to select a value from this range. Exponential utility function has been used in articles on ROs [32,33] when incomplete market arises because the investor is not well diversified so that the private risk is not eliminated by diversification. A simple approach for incomplete markets is to choose the dynamic programming framework with an exogenous discount rate as in Dixit and Pindyck [7], but perhaps some preference theory could be used.

Technical uncertainty even with diversified investors is another topic that demands more research in ROs. For example, the uncertainty with respect to the existence, volume, and economic quality of petroleum reserves is very important for oil companies that have learning options to evaluate. Another relevant application is the R&D project. Methods in which learning is modeled as a *process of variance reduction*

indexed by events (not time) [15], where *events* are the exercise of learning options, and models incorporating *learning measures* [34] are promising paths. RO models incorporating both strategic interaction and technical uncertainty [35] are another topic for research.

The main drawback of an RO is its higher complexity compared to DCF, which demands the identification of the relevant options, modeling of the relevant uncertainties including the best approximation of market values for the input parameters, and optimization under uncertainty algorithm or software. This generates many different ways to apply the key RO ideas and the problem of many different RO values for the same real-life application. For a discussion on the known (but limited) ways to implement the ideas in some popular RO models, see Borison [36] and Copeland and Antikarov [37]. The strength of an RO is its strong conceptual background, but a wider use of ROs is expected.

In this article, the two main calculation concepts used in almost all RO solution methods, the construction of a risk-free portfolio and the change of probability measure, were presented. The article showed the application of these concepts in both discrete and continuous time. As a matter of fact, it is possible to use other methods such as the intertemporal CAPM to solve RO problems [38], but these are rarely used. There are many challenges in developing more realistic RO models, but at the same time they should be simple enough for practitioners. However nowadays RO is considered as a consolidated approach for capital budgeting decisions.

REFERENCES

1. Myers SC. Determinants of corporate borrowing. *J Financ Econ* 1977;5(2):147–175.
2. Black F, Scholes M. The pricing of options and corporate liabilities. *J Polit Econ* 1973; 81(3):637–659.
3. Merton RC. Theory of rational option pricing. *Bell J Econ Manage Sci* 1973;4(1):141–183.
4. Howard R. Decision analysis: applied decision theory. *Proceedings of the 4th International Conference on Operational Research*; 1966 Aug 21–Sep 2; Boston. 1966. pp. 55–71.

5. Henry C. Investments decisions under uncertainty: the irreversibility effect. *Am Econ Rev* 1974;64(6):1006–1012.
6. Tourinho OAF. The valuation of reserves of natural resources: an option pricing approach [PhD dissertation]. Berkeley (CA): University of California; 1979.
7. Dixit AK, Pindyck RS. Investment under uncertainty. Princeton (NJ): Princeton University Press; 1994.
8. Trigeorgis L. Real options—managerial flexibility and strategy in resource allocation. Cambridge (MA): MIT Press; 1996.
9. Brealey RA, Myers SC. Principles of corporate finance. 6th ed. New York: McGraw-Hill; 2000.
10. Titman S, Martin JD. Valuation—the art & science of corporate investment decisions. Boston (MA): Pearson Education/Addison Wesley; 2007.
11. Wilmott P. Paul Wilmott on quantitative finance. 2nd ed. Chichester: John Wiley & Sons, Ltd.; 2006.
12. McDonald RL. Derivatives markets. 2nd ed. Boston (MA): Pearson Education/Addison Wesley; 2006.
13. McDonald RL, Siegel D. The value of waiting to invest. *Q J Econ* 1986;101(4):707–727.
14. Majd S, Pindyck RS. Time to build, option value, and investment decisions. *J Financ Econ* 1987;18(1):7–27.
15. Pindyck RS. Investments of uncertain cost. *J Financ Econ* 1993;34(1):53–76.
16. Lambrecht BM, Perraudin WRM. Real options and preemption under incomplete information. *J Econ Dyn Control* 2003;27(4):619–643.
17. Smit HTJ, Trigeorgis L. Strategic investment—real options and games. Princeton (NJ): Princeton University Press; 2004.
18. Cox JC, Ross SA, Rubinstein M. Option pricing: a simplified approach. *J Financ Econ* 1979;7(3):229–263.
19. Bastian-Pinto C, Brandão L, Hahn WJ. Flexibility as a source of value in the production of alternative fuels: the ethanol case. *Energy Econ* 2009;31(3):411–422.
20. Bollen NPB. Real options and product life cycles. *Manage Sci* 1999;45(5):670–684.
21. Dias MAG, Rocha KMC. Petroleum concessions with extendible options: investment timing and value using mean reversion with jumps to model oil prices. Paper presented at the Workshop on Real Options; 1998 May; Stavanger. 1998. Available at <http://www.puc-rio.br/marco.ind/extend.html>.
22. Brekke KA, Oksendal B. The high contact principle as a sufficiency condition for optimal stopping. In: Lund D, Oksendal B, editors. Stochastic models and options values. Amsterdam, The Netherlands: Elsevier Science; 1991. pp. 187–208.
23. Bjerk Sund P, Stensland G. Closed-form approximation of American options. *Scand J Manage* 1993;9(Suppl 1):87–99.
24. Paddock JL, Siegel DR, Smith JL. Option valuation of claims on real assets: the case of offshore petroleum leases. *Q J Econ* 1988;103(3):479–508.
25. Tavella D. Quantitative methods in derivatives pricing—an introduction to computational finance. Hoboken (NJ): John Wiley & Sons, Inc.; 2002.
26. Dixit AK, Pindyck RS, Sodal S. A markup interpretation of optimal investment rules. *Econ J* 1999;109(455):179–189.
27. Longstaff FA, Schwartz ES. Valuing American options by simulation: a simple least-square approach. *Rev Financ Stud* 2001;14(1):113–147.
28. Dias MAG. Selection of alternatives of investment in information for oilfield development using evolutionary real options approach. Paper presented at the 5th Annual International Conference on Real Options; 2001 Jul 13–14; Los Angeles. 2001. Available at <http://www.puc-rio.br/marco.ind/multimid.html#LA> 2001.
29. Huisman KJM. Technology investment: a game theoretic real options approach. Boston (MA): Kluwer Academic Publishers; 2001.
30. Dias MAG, Teixeira JP. Continuous-time option games: review of models and extensions. *Multin Fin J* 2010;14(1–2). In press.
31. Epps TW. Quantitative finance—its development, mathematical foundations, and current scope. Hoboken (NJ): John Wiley & Sons, Inc.; 2009.
32. Smith JE, Nau RF. Valuing risky projects: option pricing theory and decision analysis. *Manage Sci* 1995;14(5):795–816.
33. Henderson V. Valuing the option to invest in an incomplete market. *Math Financ Econ* 2007;1(2):103–128.
34. Dias MAG. Real options, learning measures, and Bernoulli revelation processes. Paper presented at the 9th Annual International Conference on Real Options; 2005 Jun 23–25;

- Paris. 2005. Available at <http://www.puc-rio.br/marco.ind/multimid.html#Paris2005>.
35. Dias MAG, Teixeira JP. Continuous-time option games: war of attrition and bargaining under uncertainty in oil exploration. In: Pitt ER, Leung CN, editors. OPEC, oil prices and LGN. New York: Nova Science Pub., Inc.; 2009. pp. 73–105.
36. Borison A. Real options analysis: where are the Emperor's clothes? *J Appl Corp Financ* 2005;17(2):17–31.
37. Copeland T, Antikarov V. Real options: meeting the Georgetown challenge. *J Appl Corp Financ* 2005;17(2):32–51.
38. Sick G. Real options. In: Jarrow RA, Maksimovic V, Ziemba WT, editors. *Finance*. Amsterdam, The Netherlands: North-Holland Publishing Co.; 1995. pp. 631–691.

CALL CENTER MANAGEMENT

VIJAY MEHROTRA
THOMAS A. GROSSMAN
School of Business and
Professional Studies,
University of San Francisco,
San Francisco, California

DOUGLAS A. SAMUELSON
InfoLogix, Annandale, Virginia

A call center is an organization where agents of a company talk on the telephone to customers or potential customers. Call centers are central to operations for a broad range of businesses, including travel reservations, product support, help desk services, order taking, emergency services dispatch, and financial transaction processing. Call centers are a strategic asset which provides firms with a direct line to customers, drive customer perception of quality, and generate significant numbers of transactions.

Call centers are a large global industry. It is estimated that for 2008 the United States had approximately 47,000 call centers and 2.7 million agents; Europe, the Middle East, and Africa together had 45,000 centers and 2.1 million agents; and Canada and Latin America had 35,000 centers and 730,000 agents. The demand for agents in India is predicted to exceed one million and there is a shortage of qualified labor.

Call center terminology originated in the traffic engineering work of Erlang [1], and can be different from queueing theory terminology (see the section titled “Queueing Theory and Queueing Networks” in this encyclopedia). Call center managers use “Erlang B” and “Erlang C” to refer to the M/M/s/s (see *The M/M/s Queue*) queues, respectively, and use the dimensionless unit “Erlangs” to refer to “offered load” (or equivalently “traffic intensity”) = arrival rate/service rate.

Call centers can be inbound call centers (which answer calls from customers), outbound call centers (which originate calls to customers, generally for telemarketing

purposes), or “blended” centers that do both. For inbound call centers, call volume and duration are random variables whose parameters and distributions are often estimated from historical data. These estimates are used to determine staffing levels and agent schedules. In contrast, outbound call center managers decide the pace and volume of calls that are placed to customers based on the number of available agents as well as the estimated likelihood of reaching a customer once a call has been placed.

Advances and cost reductions in information technology and telecommunications are creating new call center management approaches. For example, large call center operations are often housed in multiple locations and managed as a single virtual call center. In addition, a rich workflow is possible, including call-specific routing across agents and physical sites, automated interactions with customers on hold, and call messaging that results in an automatic callback to a customer when an agent becomes available. Call center operations are routinely outsourced to third parties. Call centers can be located around the globe to access specialized or less expensive labor resources.

There are several recent survey papers. Aksin *et al.* [2] discuss the business and operational issues facing call center managers, and highlight gaps between industry practice and the operations research literature. Gans *et al.* [3] cite 164 papers associated with call centers and an expanded on-line bibliography [4] includes over 450 papers along with dozens of case studies and books. There are also more specialized surveys. Koole and Mandelbaum [5] focus on queueing models for call centers. L’Ecuyer [6] focuses on optimization problems for call centers. Koole and Pot [7] and Aksin *et al.* [2] focus on multiskill call centers. Human resource issues are important to call center managers but tend to be missing from call center models; Holman [8] and Aksin *et al.* [2] discuss human resources in call centers.

Significant nonpublished call center research and application is performed by consulting firms, services research firms, and software vendors. There is a need for more public-domain work.

INBOUND SYSTEMS

Inbound call centers are driven by random customer call arrivals, usually to a telephone number that is free to the caller. In the simplest situation, a customer is routed to an available agent or, if no agent is available, the customer is routed to a hold queue. After some period, the customer is connected with an agent and speaks to the agent for some random time until the call is completed. It may not be so simple. The customer may encounter a busy signal and be “blocked.” The customer may abandon the hold queue by hanging up upon entry or after a period of waiting. After speaking to an agent, the customer might be transferred to another agent or queue for further assistance. The customer’s problem might not be resolved, leading to a follow-on call in the future. A good general business overview of inbound call center management is Cleveland [9].

Inbound call center managers need to balance cost with service quality. Cost is easy to measure. The largest expense is people, and cost performance is typically measured by agent utilization. Agent costs usually comprise 60–80% of a call center operating budget. Telecommunications costs can be an issue when telephone charges are high and queues are long.

Service quality is hard to measure. Customer waiting time in the hold queue has traditionally been used as a proxy for service quality. Waiting times vary across customers; hence, metrics are a function of the waiting time distribution. The two most popular metrics are average speed of answer (ASA), the mean customer waiting time; and service level (SL), the percentage of calls answered within a target time.

Researchers and call center managers are starting to focus on customer renegeing, also called *abandonment*. The customer abandonment rate (CAR) is becoming an

important metric. Abandonment leads to perverse incentives. Customers who abandon the queue are likely dissatisfied with the service encounter, which is bad for business. However, when a customer abandons the hold, queue shrinks, which is good for the call center’s waiting time metrics [10].

Metrics based on waiting time and abandonment measure a customer’s experience *prior* to service. There is growing interest in metrics for the quality of the customer’s experience *during* service, and the quality of service delivered by individual agents and agent groups. Hence, call centers are starting to use the “first call resolution” metric as well as traditional customer satisfaction surveys.

Creating Agent Schedules

In day-to-day operations, inbound call centers are managed using short-interval (15–60 min) “time blocks” for forecasting call arrivals and scheduling agents. Managers must staff each time block to assure adequate service quality without incurring excessive cost. This difficult managerial and technical problem seems first to appear in the literature in Edie [11], who examined toll booths.

With details varying across call centers, a five-activity process is used to schedule individual agents: (i) “forecasting” to predict call arrival rates by time block; (ii) “performance estimation” of key metrics given alternative agent staff levels by time block; (iii) “staffing” to determine the desired agent staff level by time block; (iv) “shift scheduling” to devise a set of multi-time block agent shifts that in aggregate approximate the desired agent staff level in each time block, considering legal, contractual, and customary work rules; and (v) “rostering” to assign individual agents to shifts, considering work rules and individual agent preferences. Each of these activities has an associated research area and there are interactions across them.

Activity 1, Forecasting. Standard forecasting techniques (see the section titled “Forecasting Techniques” in this encyclopedia) such as Winters’ Method, ARIMA, and regression are widely used for forecasting call arrival rates. However, Gans *et al.* [3] assert

that call forecasting was “still in its infancy.” Challenges specific to call centers include the small size of the time blocks (making the data noisy); the need for a forecast of each queue in a multiqueue center; complex call patterns, demand spikes, and lag effects that may occur at random or without the same effect every time they occur; and the effect of lost demand (abandoned and blocked calls) on future arrival rates. In addition, there will be opportunities for intraday forecast updating.

Brown *et al.* [12] feature arrival rate forecasting models as part of a much larger empirical study of call center operations. Channouf *et al.* [13] test a variety of forecasting models to predict incoming phone calls for an emergency medical system. Taylor [14] empirically study a range of forecasting methods and concludes that there is no clear winner because different methods proved to be more effective under different lead times and workloads. Weinberg *et al.* [15] propose a model for forecasting Poisson arrival rates for short time blocks, which is valuable from an operational perspective.

Avramidis *et al.* [16] discuss the correlation of call volumes across periods of the same day and suggest that information about call volumes in early periods can be used to improve the quality of forecasts for subsequent periods within the same day. Shen and Huang [17] develop a model based on singular value decomposition that is more accurate than standard industry practice and has certain advantages over Weinberg *et al.* [15], while also providing the capability for accurate intraday forecast updating.

Activity 2, Performance Estimation. Performance estimation computes the value of key performance metrics for alternative agent staff levels in each time block. Metrics can be customer-oriented, for example, ASA, SL, and CAR. Metrics can be cost-oriented, such as average agent utilization. Metrics are typically estimated using analytic models or discrete-event simulation.

The simplest situation is a single queue of homogeneous calls handled by homogeneous agents. The standard approach is to model call arrivals as a time-dependent Poisson process (*Poisson Process and*

its Generalizations) with arrival rates determined by the forecasting activity; to model service times as exponential with the service rate forecasted based on historical time-dependent averages; and to use the M/M/s steady-state waiting time distribution (see *The M/M/s Queue*) to determine ASA or SL. This “Erlang C” model is widely deployed in software used by managers with no knowledge of operations research via spreadsheet tools and small stand-alone products. The Erlang C is embedded in virtually every commercial call center resource planning system.

However, the application of the Erlang C model to call centers is problematic for many reasons, including abandonment; complex routings; random arrival rates; and violations of the steady-state assumptions of queueing theory due to short time blocks with block-dependent arrivals and staffing levels [18].

Palm [19] and more recently Mandelbaum and Zeltyn [10] have developed the Erlang-A model to help determine the waiting time distribution as a function of arrival rates, service rates, staffing levels, and the distribution of the time until customers choose to abandon. Random arrival rates are observed empirically by Avramidis *et al.* [16], Brown *et al.* [12], and Steckley *et al.* [20]. They conclude that in the presence of random arrival rates, the Erlang C systematically overestimates waiting time because the assumption of Poisson arrivals with a known arrival rate understates the variability of the call arrival patterns.

Modern call centers can route calls among multiple queues. A common approach is “skill-based routing,” where multiple types of incoming calls are handled by multiple agent pools, each capable of handling a unique subset of call types. Skill-based routing makes performance estimation much more complex, in part, because call arrivals at each agent pool depends on the staffing of the other agent pools. Discrete-event simulation has often been used to address this situation, as discussed in Mehrotra and Fama [21].

Activity 3, Staffing. Staffing sets the desired number of agents in each agent pool for each time block. Staffing is typically determined by minimizing agents subject to

a performance measure goal, such as ASA or SL. Skill-based routing make staffing more challenging.

When management seeks to obtain relatively short ASA and relatively high agent utilization rates, a simple model of practical value is the “square root safety-staffing rule” first observed by Erlang [1] and later formalized by Halfin and Whitt [22]. This rule stipulates that for sufficiently large R (where R is “offered load” or “traffic intensity” = arrival rate/service rate), staffing the system with $R + \beta\sqrt{R}$ servers will perform well for some parameter β (see Aksin *et al.* [2, Section 2.3] for additional discussion on this rule). Steckley *et al.* [20] propose an alternative method for determining staffing levels in the presence of random arrival rates, where the arrival rate variability is either known or is estimated from empirical data.

Activity 4, Shift Scheduling, and Activity 5, Rostering. Shift scheduling and rostering are challenging for call centers. It might not be cost-effective or even feasible to achieve the staffing activity’s desired number of agents in each time block. For example, it can be difficult to devise a schedule with very different adjacent time blocks. (Note that the number of time blocks can be large: with 15-min time blocks in a 24-h center, there are 96 time blocks to be managed.) Creating shift schedules that meet target requirements; satisfy legal, contractual, and customary work rules; and do not lead to the under-utilization of agents is an inherently complex problem. Considering individual agent preferences is also difficult. These topics are discussed in *Tour Scheduling and Rostering* and *Nurse Scheduling Models*.

Integration of the Five Activities. It is a common practice to perform each of the five activities in the staffing process sequentially and independently from the others. However, separating these activities can cause difficulties since, for example, highly variable staff requirements can make shift scheduling more difficult. Integration of these activities is an emerging research theme [2].

In the context of skill-based routing, Avramidis *et al.* [23] show that separating the staffing and scheduling steps can lead to very poor solutions and propose an integrated solution that performs well. Cezik and L’Ecuyer [24] propose a scheduling methodology that combines staffing optimization with skill-based routing using linear programming and simulation, while Fukunaga *et al.* [25] describe a commercially implemented technique that combines artificial intelligence with simulation.

OUTBOUND SYSTEMS

In outbound systems, a computer automatically dials calls and routes them to agents. Typically, the computer predicts when agents will become free and dials in anticipation of agent availability, thereby reducing the time agents wait for the dialed party to pick up the phone or to not answer. The system automatically processes busy signals, no-answers, and telephone company messages.

A key analytical challenge is to determine the “pacing,” or when to dial the next outbound call. If the pacing is too slow, agent time is wasted. If the pacing is too fast a called party answers when no agent is available, creating a nuisance for the called party and a wasted expense for the system.

Research in this area is mostly proprietary and there is scant research literature. The first US patent based on queueing theory [26] was granted for a method [27] that estimated service durations, times from dialing to answer, and proportions of dial attempts that result in answers and synchronized dialing attempts to finish shortly after predicted agent service completions. Other patents, such as David [28], expanded and extended this approach, and there at least 48 additional patents in this area.

Two interesting problems merit research. First is pacing for multiple campaigns, in which some agents may be shifted among different calling campaigns in real time. Second is pacing for systems, in which the called party is switched to a recorded message then back to a live operator; this approach is popular among some charitable organizations that

use celebrity fund-raising appeals, and debt collectors who want proof of what they said.

Blended Systems

Blended call centers allow agents to be switched in real time between inbound and outbound calls. Bhulai and Koole [29] present a queueing theory model which yields a threshold policy for assigning agents to outbound calls. Deslauriers *et al.* [30] provide a set of Markov chain models for a call center where outbound agents can be diverted to serve inbound calls. A few patents, notably those by Szlam *et al.* [31] and Villena *et al.* [32], address aspects of these systems. Call center managers believe that frequent switching between inbound and outbound calls degrades agent performance for both types of calls, and common practice is to make reassignments for blocks of time rather than call by call.

OPERATIONAL TRENDS AND RESEARCH OPPORTUNITIES

Some traditional call center assumptions are being questioned in the operations research literature [2]. One approach is to replace the standard point-forecast of arrival rates for a short time block with a stochastic forecast. It is possible to relax the assumption of independent time block call arrivals and model correlation of arrivals across time blocks. More general assumptions on arrival rates can affect the scheduling and rostering problems, with Steckley *et al.* [20], Robbins [33], and Gans *et al.* [34] taking some early steps in this area, but there are still significant research opportunities. Bassamboo *et al.* [35] propose a methodology for capacity planning and dynamic system control in the presence of random arrival rates and multiple inbound call types.

Scheduling

Call center workforce scheduling decisions can be dynamic. As updated data on call arrivals and agent availability become available over time, short-term forecasts and agent schedules can be adjusted. Mehrotra

et al. [36] have developed a methodology for intraday forecast and schedule updating, while Gans *et al.* [34] have suggested a stochastic programming model with recourse to account for both random arrival rates and intraday schedule updates. However, there are still significant research opportunities.

Use of Real-Time Data

Call center models generally assume that agents' service times are identically distributed for a given class of customer. Outbound models generally assume that the proportion of called parties who answer changes relatively smoothly. However, actual call center data indicate persistent difference among agent service times, even for probabilistically identical customers, and runs of high or low proportions of good contacts and of live answers. Therefore, using real-time data to adjust call center operations could produce improvement in performance, although call center managers are quick to point out that efficiency must be balanced against robustness.

Research Data

Researchers who wish to test their models need data. Data tend to be aggregated into time-based averages, which is problematic from a queueing science perspective. Fortunately, the DataMOCCA Project [37] provides a clean source of high-granularity, call-based customer call data from several sources.

Call Routing

Skill-based routing, in which different agents are capable of handling different subsets of calls in an environment with multiple call types, is a major trend in the call center industry, as discussed in L'Ecuyer [6]. These systems route customers to different agents depending on their needs and support the creation of a hierarchy of agents with highly skilled (and highly paid) personnel handling only the most challenging calls. In the past decade, there have been a number of patents in this area, notably Crockett and Leamon [38]. There is an opportunity for

research regarding design and appropriate performance measures in such systems, and in the dependency and interaction among staffing, scheduling, and routing.

When there are multiple types of calls and multiple types of agents, performance modeling, staffing, scheduling, and rostering problems all become significantly more complex, which leads to many interesting and important research problems such as those addressed by Fukunaga *et al.* [25] and Avramidis *et al.* [39].

Resource Acquisition

Call center resource acquisition is an emerging area of interest. There is a need for additional research on long-term forecasting, personnel planning for general multiskill call centers in the presence of both learning and attrition [40], and for complex networks of service providers [[2], Section 2.2].

Companies routinely outsource call center operations to third party service providers. Hasija *et al.* [41], Ren and Zhou [42], and Milner and Olson [43] explore issues associated with establishing and managing these relationships.

From Call Center to Customer Contact Center

Some call centers are transforming into customer contact centers, which integrate telephone calls with callbacks and voice mail, as well as with email, internet chat, and web-based voice. Customer contact centers face challenges similar to call centers in balancing cost with quality of service, and must grapple with the challenges of predicting customer contact volumes, and scheduling people to communicate with customers. However, from an operations management perspective, such multichannel centers provide additional operational challenges and research opportunities as touched on in Deslauriers *et al.* [30].

ISSUES FOR OR PROFESSIONALS

The highest call center expense is direct labor, so managerial attention is most readily available for techniques that reduce

call volumes, shorten talk time, or improve workforce scheduling. It can be challenging to sell stand-alone operations research (OR) software solutions because call center managers need systems to work together. It is widely accepted in the industry that software sells on the basis of workflow, not algorithms or mathematical models. Thus, “hard” OR innovations need to be embedded in software integrated with the existing call center systems. These systems capture and process data from the ACD and other sources to provide analysis and decision support and to generate workforce schedules. Leading call center management software packages that include or interact with OR techniques include Nice/IEX, Verint/Blue Pumpkin, Genesys/Alcatel–Lucent, and Aspect Telecommunications. There are many minor software vendors.

Certain “OR process” skills and the technique of call content analysis, which prevents future calls by using customer complaints as a driver to improve products, manuals, help facilities, websites, and agent training, can be powerful tools for OR consultants working in the call center space [44]. Researchers in the field of services marketing bring a behavioral science perspective to studying service operations, and pose a set of research questions about call centers that are different than the operations research community; see Bitran *et al.* [45] for more on this emerging area.

REFERENCES

1. Erlang AK. On the rational determination of the number of circuits. In: Brockmeyer E, Halstrom HL, Jensen A, editors. The life and works of A. K. Erlang. Copenhagen: The Copenhagen Telephone Company; 1948.
2. (a) Aksin Z, Armony M, Mehrotra V. The modern call center: a multi-disciplinary perspective on operations management research. *Prod Oper Manage* 2007;16(6):665–688; (b) Aksin OZ, Karaesmen F, Ormeci EL. A review of workforce cross-training in call centers from an operations management perspective. In: Nembhard D, editor. *Workforce cross training handbook*. Boca Raton (FL): CRC Press; 2007.
3. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review and research

- prospects. *Manuf Serv Oper Manage* 2003; 5(2):79–141.
4. Mandelbaum A. 2004. Call centers (centres) research bibliography with abstracts. Available at <http://iew3.technion.ac.il/serveng/References/ccbib.pdf>. Accessed 2009 Jul 29.
 5. Koole G, Mandelbaum A. Queueing models of call centers: an introduction. *Ann Oper Res* 2002;113(1–4):41–59.
 6. L'Ecuyer P. Modeling and optimization problems in contact centers. Proceedings of the 3rd International Conference on the Quantitative Evaluation of Systems (QEST 2006); 2006 Sep 11–14; University of California, Riverside. Washington, DC: IEEE Computing Society; 2006. pp. 145–154.
 7. Koole G, Pot A. An overview of routing and staffing algorithms in multi-skill customer contact centers. Working paper. Amsterdam: Department of Mathematics, Vrije Universiteit Amsterdam; 2006.
 8. Holman D. Call centers. In: Holman D, Wall TD, Clegg CW, *et al.*, editors. *The essential of the new workplace: A guide to the human impact of modern work practices*. New York: Wiley; 2005.
 9. Cleveland B. Call center management on fast forward: succeeding in today's dynamic inbound environment. Colorado Springs (CO): ICM Press; 2006.
 10. Mandelbaum A, Zeltyn S. Service engineering in action: the Palm/Erlang-a queue, with applications to call centers. In: Spath D, Fähnrich K-P, editors. *Advances in services innovations*. Berlin-Heidelberg: Springer; 2007. pp. 17–48.
 11. Edie LC. Traffic delays at toll booths. *J Oper Res Soc Am* 1954;2(2):107–138.
 12. Brown L, Gans N, Mandelbaum A, *et al.* Statistical analysis of a telephone call center. *J Am Stat Assoc* 2005;100(469):36–50.
 13. Channouf N, L'Ecuyer P, Ingolfsson A, *et al.* The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Manage Sci* 2007;10(1):25–45.
 14. Taylor JW. A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Manage Sci* 2008; 54(2):253–265.
 15. Weinberg J, Brown LD, Stroud JR. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J Am Stat Assoc* 2007;102:1185–1199.
 16. Avramidis AN, Deslauriers A, L'Ecuyer P. Modeling daily arrivals to a telephone call center. *Manage Sci* 2004;50(7):896–908.
 17. Shen H, Huang JZ. Interday forecasting and intraday updating of call center arrivals. *Manuf Serv Oper Manage* 2008;10(3): 391–410.
 18. Ingolfsson A, Akhmetshina E, Budge S, *et al.* A survey and experimental comparison of service level approximation methods for non-stationary M/M/s queueing systems. *INFORMS J Comput* 2007;19:201–214.
 19. Palm C. Research on telephone traffic carried by full availability groups. *Tele* 1957;1(1):107.
 20. Steckley S, Henderson S, Mehrotra V. Forecast errors in service systems. *Probab Eng Inform Sci* 2009;23(2):305–332.
 21. Mehrotra V, Fama J. 2003 Call center simulation modeling: methods, challenges, and opportunities. In: Chick S, Sánchez PJ, Ferrin D, Morrice DJ editors, *Proceedings of the 2003 Winter Simulation Conference*; 2003. pp. 135–143.
 22. Halfin S, Whitt W. Heavy-traffic limits for queues with many exponential servers. *Oper Res* 1981;29(3):567–588.
 23. Avramidis AN, Gendreau M, L'Ecuyer P, *et al.* Optimizing daily agent scheduling in a multiskill call center. *Eur J Oper Res* 2010;200(3):822–832. DOI: 10.1016/j.ejor.2009.01.042.
 24. Cezik MT, L'Ecuyer P. Staffing multiskill call centers via linear programming and simulation. *Manage Sci* 2008;54(2):310–323.
 25. Fukunaga A, Hamilton E, Fama J, *et al.* Staff scheduling for inbound call centers and customer contact centers. In: Dechter R, Kearns M, Sutton R, editors. *18th National Conference on Artificial Intelligence*; 2002 Jul 28–Aug 1; Edmonton, Alberta. Menlo Park (CA): American Association for Artificial Intelligence; 2002. pp. 822–829.
 26. Samuelson DA. System for regulating arrivals of customers to servers. US Patent 4,858,120. 1989.
 27. Samuelson DA. Call attempt pacing for outbound telephone dialing systems. *Interfaces* 1999;29(5):66–81.
 28. David JE. Outbound call pacing method which statistically matches the number of calls dialed to the number of available operators. US Patent 5,640,445. 1997.
 29. Bhulai S, Koole G. A queueing model for call blending in call centers. *IEEE Trans Autom Contr* 2003;48(8):1434–1438.

30. Deslauriers A, L'Ecuyer P, Pichitlamken J. *et al.* Markov chain models of a telephone call center with call blending. *Comput Oper Res* 2007;34(6):1616–1645.
31. Szlam A, Crooks JW Jr, Harris D. Method and apparatus for dynamic and interdependent processing of inbound and outbound calls. US Patent RE36416. 1999.
32. Villena J, Tellez A, Mathur M, *et al.* Blended agent contact center. US Patent 6,775,378. 1999.
33. Robbins TR. Managing service capacity under uncertainty [PhD dissertation]. Pennsylvania State University; 2007.
34. Gans N, Shen H, Zhou Y-P, *et al.* Parametric stochastic programming for call-center workforce scheduling. Wharton School Working Paper. 2009.
35. Bassamboo A, Harrison JM, Zeevi A. Pointwise stationary fluid models for stochastic processing networks. *Manuf Serv Oper Manage* 2009;11(1):70–89.
36. Mehrotra V, Ozluk O, Saltzman RM. Intelligent procedures for intra-day updating of call center agent schedules. *Prod Oper Manage* 2010;19(3):353–367.
37. DataMOCCA Project. 2006. Available at <http://iew3.technion.ac.il/serveng/References/DataMOCCA>. Accessed 2009 Jul 28.
38. Crockett G, Leamon P. Skills-based scheduling for telephone call centers. US Patent 6,044,355. 2000.
39. Avramidis AN, Chan W, L'Ecuyer P. Staffing multi-skill call centers via search methods and a performance approximation. *IEE Trans* 2009;41:483–497.
40. Ryder G, Ross K, Musacchio J. Optimal service policies under learning effects. *Int J Serv Oper Manage* 2008;4(6):631–651.
41. Hasija S, Pinker E, Shumsky R. Call center outsourcing contracts under information asymmetry. *Manage Sci* 2008;54(4):793–807.
42. Ren Z, Zhou YP. Call center outsourcing: coordinating staffing level and service quality. *Manage Sci* 2008;54(2):369–383.
43. Milner J, Olson T. Service-level agreements in call centers: perils and prescriptions. *Manage Sci* 2008;54(2):238–252.
44. Mehrotra V, Grossman TA. OR process skills transform an out of control call center into a strategic asset. *Interfaces* 2009;39(4):346–352.
45. Bitran GR, Ferrer JC, Rocha e Oliveira P. Managing customer experiences: perspectives on the temporal aspects of service encounters. *Manuf Serv Oper Manage* 2008;10(1):61–83.

CAMPAIGN ANALYSIS: AN INTRODUCTORY REVIEW

JEFFREY KLINE

WAYNE HUGHES

DOUGLAS OTTE

Operations Research Department,
Naval Postgraduate School,
Monterey, California

“As for military methods: the first is termed measurement, the second, estimation [of forces]; the third, calculation [of numbers of men]; the fourth, weighing [relative strength]; and the fifth, victory. Terrain gives birth to measurement; measurement produces the estimation [of forces]. Estimation [of forces] gives rise to calculating [the numbers of men]. Calculating [the number of men] gives rise to weighting [strength]. Weighting [strength] gives birth to victory.” Sun Tzu, *The Art of War*

Deriving insights in military operations through analytical methods is as old as the writings of Sun Tzu, such as the 2500-year-old book *The Art of War* [1]. Discussing considerations for military force disposition, Sun Tzu demonstrates the use of quantitative thinking in military decision making. His writings show that campaign analysis is far older than the origins of operations research in World War II, or Lanchester’s derivation of force-on-force exchange equations from World War I. Broad in nature, campaign analysis is a field of application, rather than a focused academic discipline. Hughes defines campaign analysis as a study of conflict between heterogeneous forces in a series of encounters, over time and a wide geographic area [2]. Its purpose is to develop insight into relationships between risk(s) engaged, resources committed, and campaign analysis. While borrowing from tools of operations research, campaign analysis is simultaneously informed by analogous conflicts in history, geographical literacy, awareness of dynamic social issues and economic interdependencies, and respect for operational military experience. Value is provided to

military decision makers when campaign analysis inspires focused discussion of synthesized information while offering informed, quantitative, specific, yet incomplete advice. *An appreciation for levels of risk versus resources is one product of campaign analysis, but detailed predictions of outcomes are not.*

While the focus of campaign analysis is on quantitative insights in support of operations before and during conflict, the analyses of past campaigns are a vital element of the field that provide guidance, data, and models for present applications. Models and lessons from McCue’s *U-boats in the Bay of Biscay* have application in antisubmarine operational plan (OPLAN) development today [3]. Likewise, Morse and Kimball’s *Methods of Operations Research* chapter on strategic kinematics provides useful models and data developed from their World War II analyses [4]. Additionally, analyses of past campaigns provide tools like Lanchester force exchange equations and Hughes’ naval salvo equations [5]. *Numbers, Predictions, and War* by Colonel Trevor Dupuy is another example of historical data analysis that gives insight into exchange rates between the competing forces [6]. Derived from analyzing historical campaigns, these works are used to help design and adjudicate war game interactions, improve combat simulations, and compute exchange and delivery rates between the conflicting forces. *In campaign analysis, we use the past to inform the future.*

The rest of this article discusses a campaign analysis process and its elements, where analysis has been informative in past campaigns, how campaign analysis is used today, and how it is evolving for future work.

CAMPAIGN ANALYSIS PROCESS

Campaign analysis cannot predict outcomes with assurance. We start with messy, complex problems; rely on performance estimates for most of our data; and deal with complex, dynamic activities, where

opponents have multiple courses of action from which to choose. The analyst's hope is to identify patterns of activity and isolate important conditions for a desirable campaign outcome.

Like most systems analysis studies or detailed operations research models, campaign analysis may be viewed as an IF-THEN statement with feedback loops. *If* we accept these assumptions; *if* we use these data about friendly and enemy forces, weapons, and their dispositions; *if* we select these variables to study; and *if* we use these models; *then*, we obtain results in our campaign, or revisit our original assumptions, data, and models and revise our battle plans.

William "Forrest" Crain describes several general approaches to theater campaign analysis that parallel military staff planning [7]. The first of these is to start at the desired end state of the campaign's objectives, and then plan and analyze back in time to develop conditions and actions that will allow that end state to be achieved. The next approach is to analyze alternatives within bounded sets defined by feasibility, suitability, and acceptability. Similarly, the final approach is to develop and analyze a single best course of action to achieve the desired end state. All variations of campaign analysis are essentially a systems analysis process. A campaign's objectives, concept of operations, and measures of effectiveness are identified; assumptions about the battle environment, force capabilities, and dispositions made; important variables identified; models selected and implemented; results analyzed and evaluated; sensitivity analysis conducted; conclusions warranted by results, and reports delivered to the commander and planning staff.

The campaign's objectives come from the direction provided by political and military leadership at the national level. Derivation of a concept of operations to achieve those objectives, and metrics to measure their achievement, is done in collaboration with the commander and his or her staff. Assumptions are agreed to and provide a bound on the analytical study. We must assume some enemy objective and capability in a region of conflict—and the weather conditions or

season—often making statistical simplifications that we know are not strictly true, such as the independence among probabilities. A common and valuable study output is the identification of critical assumptions or those assumptions whose change would modify our advice to a commander. These critical assumptions may potentially define initial conditions for the commander to establish before campaign execution, such as the type and disposition of his forces, and the security of his sustainment pipeline.

Data are difficult and suspect. Peacetime test data are only estimates of combat performance in a particular future situation and condition. Sensor performance, weapons' effectiveness, platforms' combat range, weapons' load out, and turnaround times are just a few examples of purported "data" that can quickly degenerate from "facts" to assumptions. An estimate or probabilistic distribution, however, is required for each friendly and enemy capability, if only as a first-order approximation. In addition, although we can control (somewhat) the levels, capacities, and disposition of our own forces, the enemy's performance data must be an educated guess, pieced together from intelligence estimates and our assumptions about his objectives. The analyst will mark these estimates for later sensitivity analysis to evaluate which assumptions are critical to the results.

While assembling imperfect, but sufficient inputs, we concurrently select a model or series of models to represent the campaign environment. Broadly speaking, models bound the campaign in either a series of engagements (pulses of power) or a continuous operation where many small engagements create a larger effect (cumulative warfare). The battles across Europe in World War II represent a series of engagements, while insurgent warfare, submarine warfare, and blockades are examples of the continuous cumulative model. Model selection is informed by the scenario and rough concept of operation outlined by the commander and staff.

Model categories range from closed-form probabilistic equations, computer simulations, optimization, and war games

to field experiments and operational rehearsals. A field experiment, the rehearsal of an amphibious landing or coordinated air strike, is a model in the sense that it is not reality. In his book, *The Stress of Battle*, David Rowland observes that ground “combat” on an instrumented range can differ from (exceed) casualty generation rates in similar real battles by a factor of 2 or more [8]. Real bullets are more serious to soldiers than “death” by laser in training. This is an example how actual battle data can inform our modeling efforts.

Various models may be used at each stage of the combat analysis. For example, a war game may help to develop concepts of operation and employment for the opposing sides. The war game’s interactions may be adjudicated by tactical simulations, equations, historic engagements or professional judgment. Once an employment concept or course of action is generated, it may be programmed in a larger campaign simulation to conduct analysis on many model variations. Optimization may be used to analyze the best resource allocation for particular courses of action. This implies some set of campaign measures of effectiveness, agreed to by the commander, to guide planning and evaluate the model outputs, and requires constant dialog between the commander, the planning staff, and the analyst. The dialog also facilitates matching the model’s metrics to those collected during the campaign’s execution, thereby better informing determinations regarding progress toward campaign objectives. In the Battle of the Atlantic in World War II, the exchange rate—merchant vessels sunk per submarine sunk—was popular, convenient, clean, and easy to measure. However, as was pointed out by historians, the “true” measure of accomplishment was successful merchant transits; pushing shipping safely across the ocean was aided by slowing or pinning down the U-boats, without actually sinking them.

Models may be used in a hierarchy representing an extended combat process. The results of engineering-level models on weapons’ performance (e.g., radar performance against an enemy missile) may provide data to an engagement model that

focuses on a single contest between an air defense system and enemy missiles. The engagement model’s output may then be used in a mission-level model that represents the effectiveness of the air defense mission in a particular area. The mission-level output, in turn, can feed a campaign-level model to provide an estimate of air defense’s contributions to the total campaign.

Once the models’ outputs are received, the campaign analyst’s art is tested to its utmost. Interacting with the commander and staff, the campaign analyst must identify important messages from the models’ output that affect the campaign, possibly repeating some of the modeling to identify critical assumptions affecting the results. He/she must evaluate the quantifiable output based on his/her own operational experience and tailor the advice using other considerations, such as the social and economic conditions, or historical regional tensions. This last step is a final dialog between the combat analyst, staff, and commander. The analyst must present his/her contributions in relevant and understandable, jargon-free ways. But successful communication requires a commander’s astute appreciation of campaign analysis in both its power and limitations. *In the end, the value of the campaign analysis is judged by its contribution in aiding the commander to develop his/her intent, strategy, and operational course of action.*

CAMPAIGN ANALYSES’ CONTRIBUTIONS TO MILITARY JUDGMENT

A superb reference on analytical techniques for insurgent warfare, Lieutenant General Julian Ewing and Ira Hunt’s book, *Sharpening the Combat Edge, the Use of Analysis to Reinforce Military Judgment* [9], relates how analysis was used to guide operations during the Vietnam War. The book’s title conveys a balanced message on the contributions campaign analysis can make to planned and ongoing operations. Campaign analysis can reinforce or challenge military judgment, but never replace it. We illustrate with three examples from World War II where analysis was used to shape and reinforce strategic planning or ongoing operations.

In 1938, German submariner Admiral Doenitz was asked by the German high command to estimate the number of undersea boats (U-boats) he would need to cut off transatlantic supplies from North America to Britain when war began. To obtain such a figure, the Admiral and his staff employed war gaming to estimate the tonnage of allied shipping to sink to starve Britain's need for fuel, ammunition, war supplies, and food. Their estimate of 600,000 tons per month was undoubtedly based on assumptions of the intensity of future land and air combat with corresponding supply requirements; allied shipping capacity, speeds, and resupply rates; merchant crew availability; and British shipping replacement rates [10]. To obtain this level of effectiveness, they believed 300 U-boats were needed. Again, assumptions on submarine effectiveness, sea coverage capabilities, and on-station times were considered necessary to derive this number. The German Navy was forced to start the war with only 57 U-boats in 1939. Admiral Doenitz's campaign analysis prior to the war was, in hindsight, probably about right. In any event, it undoubtedly shaped the German Navy's operational thinking during the Battle of the Atlantic [11].

On the other side of the Atlantic, to gain strategic lessons in a potential conflict with Japan, a series of strategic war games was played between the world wars at the United States Naval War College. This campaign analysis used serial war gaming to educate and inform the strategic planners of War Plan Orange—an evolving strategic plan, decades in development. In case of war with Japan, an initial strategy proposed moving the US fleet quickly across the Pacific to engage the Japanese Navy, if Japan moved against the US Philippine territory. After several war games and tactical assessments on the War College's game floors, planners realized that despite the nominal 5:3 advantage in capital ships tonnage enjoyed by the US Navy, too few ships would be in position to conduct an early battle for mastery of the seas [12,13]. The island-hopping campaign to roll back Japanese forces then emerged as a more feasible strategy, and was later executed during the war. In his book, *War Plan*

Orange, Miller relates how planners in Newport and Washington studied logistic requirements and capabilities, force availability, and force-on-force exchange evaluations through capability comparisons, and force buildups to quantitatively update the debate on how to defeat Japan.

Best known to operations research professionals is the birth of their discipline during World War II: England and the United States brought together scientists from diverse fields to develop the best coordination methods of radar and fighter aircraft during the Battle of Britain, and to make valuable contributions to the Battle of the Atlantic antisubmarine campaign. In both cases, scientists evaluated how best to employ new technologies to shape ongoing campaigns. In their seminal work, *Methods of Operations Research*, Morse and Kimball quote Admiral E.J. King, the American Chief of Naval Operations (CNO), regarding operations research contributions to the allies' response to new German submarine technology and tactics: "Operations Research, bringing scientists in to analyze the technical import of the fluctuations between measure and counter-measure, made it possible to speed up our reaction rate in several critical cases" [14]. In the same 1945 report, King makes clear the necessity for scientists to work under the direction of, or have close personal contact with, the military officers planning and carrying out the war. This relationship ensured that mathematical tools developed by the scientists would be influenced by campaign planners in a close relationship of mutual trust. Search theory, tactical analysis, and combat exchange rate equations, refined and evaluated by scientists during the war, are all tools in the campaign analyst's toolbox.

CURRENT USES OF CAMPAIGN ANALYSIS

Force Structure Analysis

Campaign analysis has long been the cornerstone of the US Department of Defense (DoD) staff requirements generation and force structure development. Although many countries use similar processes to establish force requirements, we focus on the United

States defense department as the example most familiar to the authors. Using strategic guidance focused on potential major conflicts and current operations, campaign analysis provides critical insights into capability gap assessments and helps decision makers understand the risk associated with procurement decisions. Currently the analysis generates the foundation of resource decisions within the biannual DoD Program Objective Memorandum. It is conducted within the context of comprehensive joint and combined warfare analysis.

Campaign analysis to support force structure decisions starts with a review of national-level guidance. The DoD's strategic goal is to respond to the security requirements articulated by the President through the National Security Council. The Defense Department translates the strategic priorities of the National Security Strategy into defense documents, such as the Guidance for Development of the Force. Those articulate specific war-fighting capabilities and requirements within a fiscally constrained budget. Defense Planning Scenarios (DPSs) outline analytic baselines by identifying possible future conflicts the United States' armed forces may be required to face. The DPSs establish the future war-fighting baseline, assess force flows associated with each campaign scenario, and attempt to anticipate the combat characteristics of potential adversaries.

Every US armed service conducts its campaign analysis a bit differently, but converting national guidance to force requirements via campaign analysis of the DPSs is the underlying goal for each service. To illustrate, we focus on the Navy's process. The CNO augments national strategic directives with pertinent maritime guidance that focuses campaign analysis on naval contributions to the joint and coalition war fight. In 2007, *A Cooperative Strategy for the 21st Century Seapower*, served this purpose [15]. Signed by the service chiefs of the US Navy, US Marine Corps, and US Coast Guard, it articulates expectations for future operations of maritime forces. Supplemented by the Naval Operational Concept, Navy Strategic Plan, specific CNO guidance, and a

maritime-specific intelligence update, these documents collectively describe activities and associated assumptions for Navy campaign analysis. Merged with the DoD DPSs, they provide the foundation for the Navy's analytical work in force structure analysis.

A maritime, DPS-based analysis is composed of the same steps from the general campaign analysis method previously discussed. Figure 1 shows this procedure for the Navy staff. Within each scenario, the maritime campaign objectives are identified (e.g., protect friendly shipping). Tasks are then assigned to joint forces to obtain that campaign objective (e.g., clear minefields). Campaign metrics are selected and criteria established within the campaign plan (e.g., time needed to restore military and commercial shipping's access into a theater). Finally, specific force contributions to meet those mission criteria are analyzed (e.g., capability to find, fix, and clear the mines). In this way, current force capabilities are assessed to meet future mission requirements.

The practical goal of naval campaign analysis, as conducted by the headquarters staff, is to identify existing war-fighting capability and capacity gaps, or shortfalls, in specific weapon systems and force structure components, which will form future requirements and ultimately affect program budget decisions. Since the analysis is grounded in DoD-approved war-fighting scenarios, which stress all aspects of maritime warfare, the campaign analysis provides a comprehensive assessment of all maritime warfare areas and their interdependencies within the campaign. The individual DPSs test specific war-fighting capabilities such as antisubmarine warfare, strike warfare, and antiship missile defense, as well as theater security cooperation and lesser contingency mission sets. Finally, the war-fighting requirements highlighted in the current analysis are then assessed against capabilities addressed in acquisition programs in previous budget Program Objective Memorandums, to determine potential capability and capacity gaps.

Gap analysis assisted by campaign analysis generally takes two forms. First, it provides insights as to how well our procurement objectives keep pace with the

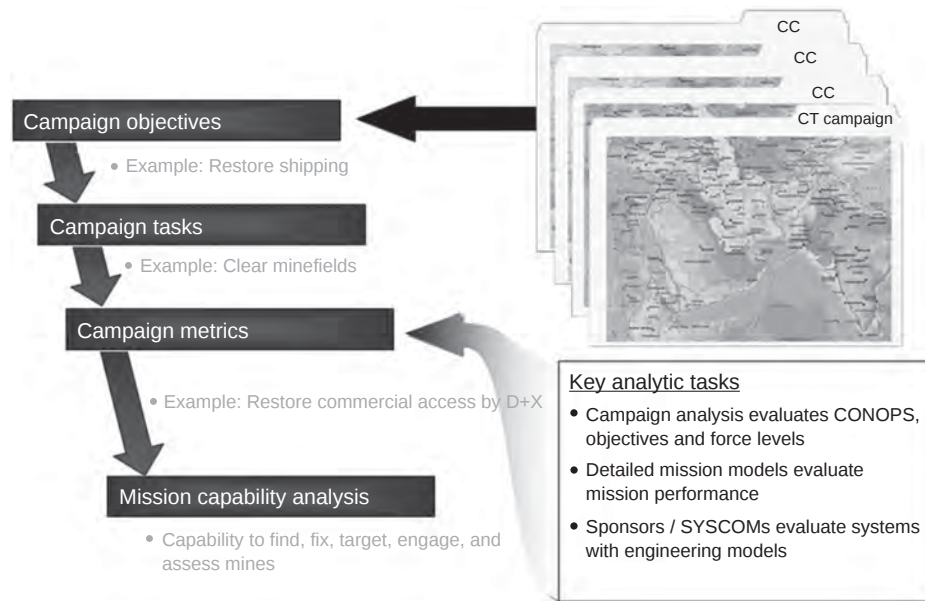


Figure 1. The scenario-based war-fighting analysis.

enemy's weapons advancements. The gaps can show divergence in the strategic outlook (and, hence, the need for improvement of the forces), focus advancement on a particular enemy war-fighting capability that out-paces our potential technological advancements, or atrophy of a particular war-fighting capability that can be exploited by a potential enemy. Campaign analysis helps quantify the magnitude of the capability gap and shows the propensity of specific technologies to overcome it.

Second, campaign analysis helps to describe the enemy's war-fighting capabilities over time. Our procurement programs may reflect the enemy's buildup or technological advancements, but not at a rate that meets critical parameters of the campaign. Capacity gap analysis is critical in the development of new strategies, such as the need to build partnerships for theater security, as defined in *A Cooperative Strategy for 21st Century Seapower*.

The insights that campaign analysis provides are not complete without concurrently addressing risk assessment. Assessing risk aids the calculus of limited resource allocation decisions. It can help shape

command decisions regarding the primacy of addressing short-term gaps versus providing for future capacity and flexibility.

In summary, the services build their campaign analyses on the foundation of national security guidance and within the context of DoD-approved planning scenarios. Although each service focuses on its domain, these campaigns include the joint and coalition capabilities that the other services contribute. Campaign analysis provides insights into current and future capabilities and force structure requirements, while providing decision makers with the appropriate risk assessments. Joint campaign analysis is the bedrock of resource allocation decisions within the DoD.

Operational Planning and Execution

Campaign analysis also contributes to the development of current OPLANs. The primary difference between analyses to support future force structure planning and current operational planning is that in current OPLAN development, the resources available to the planners and enemy are immediate, not years in the future. Bertha

and Shelton provide general approaches to provide combat operations analysis [16]. For the most part, these parallel the analytical methods already described, but they focus on providing the commander a quantitative assessment of important issues related to resource allocation in time, distance, location, cost, and effectiveness. These broad questions may be developed in advance and identified as the commander's essential elements of analysis, with related measures of effectiveness. If the plan is executed, collecting these measures during the campaign's phases becomes an important way to assess goal achievement.

Campaign analysis is also used during the execution of a military campaign to redirect resources, as the enemy's actions and our own performance impact the original plan. Collecting metrics related to achievement of a campaign's objective, assessing engagement results, and modeling new courses of action are activities conducted by campaign analysts during ongoing operations.

NEW OPPORTUNITIES AND CHALLENGES

When threats and strategies change, new military missions emerge and campaign analysis must develop new tools to provide quantitative contributions to aid decision making. Following the fall of the Soviet Union and accompanying the rise of the terrorist threat, the past 30 years have been a time of ferment. Examples of new missions and resulting technological advancements include expanding interest in social studies, the impact of robotic swarms on the battlefield, the influence of cyber war, and the ability to exploit computing power through data mining.

The US military is invigorating an interest in how civil and constabulary actions by armed forces can shape social outlooks to counter enemy insurgents. These actions may include digging wells, providing security for a marketplace, giving assistance after a tsunami, or intercepting pirates on the high seas. Understanding and measuring the effects of their actions on local population perceptions and attitudes requires an appreciation for local societal norms, invoking the

assistance of social sciences. In some cases, direct measures are too difficult to obtain, so indicators are selected as surrogates. For example, school attendance in a region as an indicator of achieving stability may be selected. By teaming with regional experts, planners, and social scientists, campaign analysts can strive to develop measures or indicators to help quantify the effects of civil interactions. Modeling these effects to obtain quantitative insights is challenging. Human values and attitudes are complex, changing, and not always "rational." Lessons from social sciences, strategic communications, and even advertising may prove valuable in attempts to model issues related to shaping perceptions through civil actions.

With increased capabilities in cyber networking and the potential for numerous unmanned vehicles, regular warfare is becoming more complex. Force-on-force sequential warfare and cumulative warfare have the potential to merge across a wider geographic region, where simultaneous asymmetric engagements may create emerging conditions for victory [17]. This increase in complexity poses a new requirement for the campaign analyst. Casualty exchange models representing either large, sequential engagements or smaller, continuous actions may not be sufficient. Use of agent-based simulation to model each robot or attack vehicle in a basic engagement may provide a better understanding of networked and robotic warfare. Agent-based modeling is not new, but its use to model the effects of emerging technologies in a long campaign is a method that is promising for the future.

"Data farming" aids new simulations by distilling the large databases that are generated. Modern computing power allows many thousands of simulations of a campaign, while varying numerous parameters. Through design of experiments, select parameters can be varied across their possible range. Data analysis estimates the correlation between parameters and potential outcomes. The goal is a better understanding of initial conditions (weapon performance, logistic capability, and geography) required to obtain the desired

campaign objectives. Computing power also allows solutions to new “Attacker–Defender” optimization models that provide planners with the best friendly courses of action, given the enemy may observe our preparations before mounting his most effective defense against us (or Defender–Attacker models, where our defensive preparations may be visible to the attacking enemy). Performance data assumptions can be quickly modified to see how the initial conditions affect decisions by both enemy and friendly forces. For example, best placement of theater air defense platforms to defend a number of cities and military installations may be derived to counter the enemy’s worst-case launch plan, assuming the enemy will observe the positioning of our defensive interceptor platforms. If additional friendly platforms are provided, or if new enemy capabilities are discovered, the optimization can be quickly rerun to change force distribution and see how this may affect overall defender performance [18].

Campaign analysis may also benefit from a more aggressive adoption of system dynamics modeling, which provides the potential to capture a better understanding of the complexity associated with changing conditions and feedback loops in the execution of a campaign. Uncovering possible unintended consequences from our actions and discovering interrelationships within the target system we are trying to affect are two benefits that may be derived. For example, Grynkewich and Reifel [19] used a systems dynamics model to better understand the relationship between time, funding, and operational aspects of the radical Islamic Salafist Group for Preaching and Combat (SGPC) in Algeria. Although, by their own admission, their resulting model had limitations and will benefit from additional research, the act of creating the model provided a deeper understanding of the SGPC’s funding sources and their relationship with radical activities.

CONCLUSIONS

Campaign analysis is an applied field of endeavor designed to provide quantitative

insights to a decision maker on how to best use military forces to achieve strategic and operational goals. Analysis at the campaign level may aid in concept generation and course of action selection. But because of the sheer number of variables and because a campaign is conducted against a thinking, adaptive enemy, it can only supplement, never replace, experienced military and naval judgment. Campaign analysis adapts tools from operations research, social science, and systems analysis. It applies simulation, optimization, data analysis, economics, game theory, statistics, and psychology as well as an appreciation of historical, social, and military experience. In the end, however, the value of campaign analysis is judged solely by its ability to provide the commander with timely, useful, and measurable information to guide decisions.

REFERENCES

1. Tzu S. The art of war [translation and introduction by Ralph D. Sawyer]. Boulder: Westview Press; 1994. p. 184.
2. Hughes WP. Joint campaign analysis student text book 1. Monterey: Naval Postgraduate School; 1999.
3. McCue B. U-boats in the Bay of Biscay: an essay in operations analysis. Washington, DC: National Defense University Press; 1990.
4. Morse PM, Kimball GE. Methods of operations research. New York: Technology Press of Massachusetts Institute of Technology and John Wiley & Sons; 1951.
5. Hughes WP. Fleet tactics and coastal combat. 2nd ed. Annapolis: Naval Institute Press; 1999.
6. Dupuy TN. Numbers, predictions and war: using history to evaluate combat factors and predict the outcome of battles. Indianapolis/New York: The Bobb-Merrill Company, Inc.; 1979.
7. Crain WF. Theater campaign analysis. In: Loerch AG, Rainey LB, editors. Methods for conducting military operational analysis. Washington, DC: Military Operations Research Society; 2007. pp. 13–49.
8. Rowland D. The stress of battle: quantifying human performance in combat. MinDef, Great

- Britain: Defense Science and Technology Lab; 2006.
9. Ewing JJ, Hunt IA. Sharpening the combat edge: the use of analysis to reinforce military judgment. Washington, DC: Department of the Army; 1995. Available at <http://www.history.army.mil/books/Vietnam/Sharpen/index.htm>.
 10. Doenitz K. *Memoirs, ten years and twenty days*. New York: Da Capo Press; 1997. pp. 34–40.
 11. Perla PP. *The art of war gaming: a guide for professionals and hobbyists*. Annapolis, MD: Naval Institute Press; 1990.
 12. Miller ES. *War Plan Orange: the U.S. strategy to defeat Japan, 1897–1945*. Annapolis, MD: Naval Institute Press; 1991.
 13. Vlahos M. *The Blue Sword, The National War College and the American Mission, 1919–1941*. Newport, RI: Naval War College Press; 1980.
 14. Morse PM, Kimball GE. *Methods of operations research*. New York: Technology Press of Massachusetts Institute of Technology and John Wiley & Sons; 1951. p. 3.
 15. Allen TW, Conway JT, Roughead Gary. A cooperative strategy for 21st century seapower. 2007. A joint U.S. Navy, U.S. Marine Corps, and U.S. Coast Guard document available at www.navy.mil/maritime/MaritimeStrategy.pdf.
 16. Bertha RL, Shelton RL. Combat operations analysis. In: Loerch AG, Rainey LB, editors. *Methods for conducting military operational analysis*. Washington, DC: Military Operations Research Society; 2007. pp. 51–73.
 17. Kline JE. Joint vision 2010 and accelerated cumulative warfare. In: Somerville MA, editor. *Essays on strategy XV*. Washington, DC: National Defense University Press; 1999.
 18. Brown G, Carlyle M, Diehl D *et al.* A two-sided optimization for theater ballistic missile defense. *Oper Res* 2005;53(5):745–763.
 19. Gynkewich A, Reifel C. Modeling jihad: a system dynamics model of the Salafist group for preaching and combat financial subsystem. *Strat Insights* 2006;V(8). Available at <http://www.ccc.nps.navy.mil/si/2006/Nov/grynkewichNov06.asp>.

CAPACITY ALLOCATION IN SUPPLY CHAIN SCHEDULING

NICHOLAS HALL

Department of Management Sciences,
Fisher College of Business, The Ohio
State University, Columbus, Ohio

ZHIXIN LIU

Department of Management Studies,
College of Business, University of
Michigan-Dearborn, Dearborn,
Michigan

Capacity allocation problems arise frequently in manufacturing systems, where one or more manufacturers supply one or more distributors (see **Capacity Allocation**). We adopt this notation throughout, in place of, for example, “suppliers” and “retailers,” since the scheduling costs we consider are most typically incurred by manufacturers and distributors. A capacity allocation problem can be viewed as either (i) a deficiency or excess of capacity allocated to a single-product line or distributor, or (ii) a misallocation of capacity among multiple product lines or distributors. Moreover, a misallocation of capacity can lead to excess capacity in a part of the system and a deficiency elsewhere, resulting in high scheduling cost and poor supply chain profitability. As we discuss, capacity allocation and scheduling issues are closely related, and decisions involving them need to be coordinated.

Capacity allocation problems can originate on either the demand side or the supply side. On the demand side, certain industries naturally experience volatile demand, which makes capacity planning more difficult. Examples include industries with high fashion content or occasional demand surges, and industries with rapid technological development including telecommunications and consumer electronics. The supply side also generates capacity allocation problems, for example in industries that are capital-intensive, where additional capacity

cannot be added quickly in order to meet demand. Capacity allocation and scheduling problems are more likely to arise in make-to-order environments, where production is stimulated by an order. This is because, in make-to-stock environments, inventory can be used as a buffer to resolve differences between demand and supply. Moreover, it is in make-to-order environments that detailed scheduling is most valuable.

Capacity allocation problems can most easily be discussed in simplified, single-product environments. Although capacity allocation among products is not an issue here, there may be inefficiencies that result from over-ordering or under-ordering, where distributors distort their demand information and orders for their own interests. Capacity allocation issues within multiple product environments are often considered as manufacturing flexibility, which is the ability to use the same capacity to produce different products. Capacity allocation among distributors can be performed either product by product or in total, across all products. There are advantages to each approach. When manufacturers allocate capacity product by product, they more closely control their schedule and therefore also their costs and profitability. The allocation of total capacity across products is preferred by distributors, since it gives them more flexibility. Also, since distributors are closer to the market than manufacturers are, they may have better market information. This can enable them to make more efficient choices about how to use the capacity that is allocated to them.

Capacity allocation problems are often studied in coordination with other operational decisions in supply chains. We focus on capacity allocation decisions within the scope of *supply chain scheduling*, which studies the coordination of scheduling and other operational decisions within supply chains. Within supply chain scheduling, the benefits of coordinating the entire supply chain are evaluated exactly, by using

optimal algorithms. Chen (see *Coordination of Production and Delivery in Supply Chain Scheduling*) and Hall (see *Supply Chain Scheduling: Origins and Application to Sequencing, Batching and Lot Sizing*) provide an overview of the coordination of scheduling and other operational decisions, such as sequencing, batching, lot sizing, and delivery, within supply chain scheduling.

Scheduling is important in capacity allocation decisions because orders from distributors typically specify not only quantities, but also delivery times. Further, capacity is often time-sensitive, so manufacturers incur different scheduling costs under different production schedules. Traditionally, scheduling decisions are made only after production capacity has been allocated. However, with increasing competition and declining profit margins, it is necessary for companies to consider capacity allocation and scheduling decisions simultaneously, especially in make-to-order environments [1]. For example, in the personal computer industry, Dell, Apple, and Gateway have responded effectively to volatile customer demand by

managing capacity and production schedules simultaneously [1].

Capacity allocation problems are often modeled as cooperative or noncooperative games. In Table 1, we classify 20 related works. Our classification scheme focuses on modeling characteristics. In Table 1, the column “Game” shows whether cooperative (C) or noncooperative (N) game theory is used; the column “Capacity” indicates whether capacity is sufficient (S), deficient (D), or both (DS); the column “Product” indicates whether a single (S) or multiple (M) product environment is considered; the column “Scheduling” indicates whether scheduling is considered or not (Y: yes; N: no); the column “Mechanism” shows the mechanism by which capacity is allocated (A: auction; C: contract; P: pricing; R: rule); finally, the column “Truth” indicates, for noncooperative distributors, whether the capacity allocation mechanism finds the same allocation as when all the distributors reveal their true information (Y: yes; N: No; YN: multiple mechanisms with both yes and no). Under “Mechanism,” a rule can be

Table 1. Classification of Capacity Allocation Papers

Paper	Game	Capacity	Product	Scheduling	Mechanism	Truth
Curiel <i>et al.</i> [2]	C	S	M	Y	R	—
Curiel <i>et al.</i> [3]	C	S	M	Y	R	—
Slikker [4]	C	S	M	Y	R	—
Maniquet [5]	C	S	M	Y	R	—
Katta and Sethuraman [6]	C	S	M	Y	R	—
Aydinliyim and Vairaktarakis [7]	C	S	M	Y	R	—
Cai and Vairaktarakis [8]	C	S	M	Y	R	—
Vairaktarakis and Aydinliyim [9]	C	S	M	Y	R	—
Hall and Liu [10]	C	D	M	Y	R	—
Cachon and Lariviere [11]	N	D	S	N	R	YN
Cachon and Lariviere [12]	N	D	S	N	R	N
Cachon and Lariviere [13]	N	D	S	N	R	YN
Fang and Whinston [14]	N	D	S	N	C	Y
Ganesh <i>et al.</i> [15]	N	D	S	N	P	Y
Wellman <i>et al.</i> [16]	N	DS	S	Y	A	N
Reeves <i>et al.</i> [17]	N	DS	S	Y	A	N
Hall and Liu [18]	N	DS	S	Y	A	N
Hain and Mitra [19]	N	S	S	Y	R	Y
Bukchin and Hanany [20]	N	DS	M	Y	R	Y
Vairaktarakis [21]	N	S	S	Y	R	Y

used for capacity allocation, for scheduling, or for both.

This article is organized as follows: the section titled “Capacity Allocation Mechanisms” reviews the papers listed in Table 1. The middle section presents a list of potential future research directions, and the final section provides a summary.

CAPACITY ALLOCATION MECHANISMS

In the section titled “Cooperative Capacity Allocation,” we discuss cooperative games that involve capacity allocation and scheduling issues. We focus on the existence of fair allocations of savings that ensure the cooperation of all the players. The subsequent section discusses noncooperative capacity allocation games. Some representative works on capacity allocation without scheduling issues are introduced, followed by a discussion about the allocation of scheduling capacity among competitive distributors.

Cooperative Capacity Allocation

Cooperative capacity allocation considers a central planner which finds an allocation of production capacity that achieves an optimized or nearly optimized system value, and then distributes the resulting total profit or cost saving among the manufacturers and distributors in a mutually beneficial way. Thus, we consider a cooperative game where all the players work together in a grand coalition. However, in order for this coalition to be stable, each subset of players must receive at least as much profit as they could obtain independently. Profit divisions that ensure stability of the grand coalition by achieving this are said to be in the core. Core solutions compensate some players for foregoing other profitable opportunities, in order to sustain the grand coalition. In the following discussion, we focus on cooperative capacity allocation games with scheduling costs. An important distinction within the related literature is whether total capacity, if properly allocated, is sufficient.

We first discuss problems where total capacity is sufficient. However, the profitability of the orders, referred to in the scheduling literature as jobs, depends on which part of the capacity is allocated to them, since this determines their completion times. Within this literature, two types of jobs are studied. For preemptive jobs, the processing of a job can be stopped and resumed at a later time; for nonpreemptive jobs, processing cannot be interrupted until completion. In our discussion, jobs are nonpreemptive unless specified otherwise.

The simplest cooperative scheduling games are the sequencing games defined by Curiel *et al.* [2]. Suppose that a sequence of nonpreemptive jobs, without idle time among them, is given. In practice, this is typically the order in which the jobs arrived. Each job incurs a cost that increases linearly with its completion time. Thus, a pairwise interchange of two adjacent jobs may result in cost savings. A sequencing game considers the various possible divisions of such savings. It is shown that divisions which result from dividing the savings equally between the two adjacent jobs are in the core. This valuable insight has led to several generalizations.

The first generalization, by Curiel *et al.* [3], considers cost as a general nondecreasing function of the job completion times. They identify a particular division of savings, and show that it is always in the core. Under this division, the payoff to player i is the average of (i) the incremental value from adding i to the coalition of players that precede i in the initial sequence, and (ii) the incremental value from adding i to the coalition of players that follow i in the initial sequence. The second generalization, by Slikker [4], allows nonadjacent job interchanges. Maniquet [5] discusses a related problem of sequencing agents who require a service. The initial sequence is given by their arrival order, but efficiency requires that they be served in nondecreasing order of waiting cost. An axiomatic justification is given for using the Shapley value to determine the savings division. The Shapley value is the incremental value added when a player joins

a coalition, averaged over all possible joining sequences.

Another natural generalization of the basic sequencing game model allows for no initial sequence, as discussed by Katta and Sethuraman [6]. A single machine is used to process jobs from multiple agents. Each agent has a single job with scheduling cost in proportion to its completion time. A reasonable consideration of fairness is to treat every job equally, that is, to sequence the jobs randomly. On the basis of the definition of the expected cost of a coalition, the authors propose two solution approaches to derive upper bounds on the allocation of the cost of a centralized optimal schedule to any coalition of agents. One of these approaches guarantees a nonempty core, while the other does not.

We now consider works where the total capacity is insufficient. One solution to this is outsourcing, as commonly used in large-scale electronics manufacturing. Aydinliyim and Vairaktarakis [7] consider a group of manufacturers who outsource their operations to a single third party. They describe a closed-form expression for a particular division of savings, and show that it is a core solution of the game. Using this division, the total cost can be reduced by an average of 32%. Cai and Vairaktarakis [8] provide a similar analysis, allowing for preemptive processing. They show that cooperation among manufacturers reduces costs by about 20%. Vairaktarakis and Aydinliyim [9] consider manufacturers which individually subcontract part of their orders, and minimize their total job completion time. The third party's rescheduling solution provides substantial savings, relative to the Nash equilibrium or first-come-first-served solutions. A Nash equilibrium is a solution in which no player has an incentive to change its decisions, assuming that the other players do not. The authors describe a core solution of the game. An interesting insight is that decentralized solutions tend to underutilize third-party capacity, relative to those chosen by a central planner.

Another solution to a total capacity shortfall is the rationing of capacity through a capacity allocation rule. This solution is applied to a two-stage supply chain involving

a manufacturer and several distributors by Hall and Liu [10]. Three steps occur: (i) the manufacturer allocates capacity and a set of orders that can be revised and resubmitted, to the distributors; (ii) the distributors revise their orders to maximize profit, either by coordinating among themselves or not; and (iii) the manufacturer schedules its revised orders, subject to its allocated capacity and resubmittable orders. The distributors' capacity-sharing problem is modeled as a cooperative game. Three coordination issues are studied: first, the manufacturer's coordination of production and capacity allocation, which generates 10.66% more profit on average for the manufacturer than a typical result from the proportional and linear capacity allocation mechanisms [11,12]; second, the distributors' coordination of order revisions, which generates 2.58% more profit for them; and third, coordination between the manufacturer and distributors, which generates 3.44% more profit for the supply chain.

Noncooperative Capacity Allocation

Noncooperative decision environments are characterized by a lack of centralized coordination, imperfect information, and conflicts between manufacturers and distributors. In such environments, it is difficult to find an efficient allocation of capacity, even when scheduling issues are not considered. Capacity allocation mechanisms without consideration of scheduling issues typically use simple allocation rules. Examples include the lexicographic, proportional, linear, relaxed linear, and uniform allocation rules [11,12]. More complex allocation mechanisms, such as a turn-and-earn policy [13], option contracts [14], and pricing mechanisms [15] are also used in practice. When capacity allocations are time-sensitive, scheduling issues arise and make capacity allocation decisions more difficult. In this case, auctions and complex allocation mechanisms help to find well-coordinated solutions for capacity allocation problems. This may require the elicitation of true information from competitive distributors.

We first introduce noncooperative capacity allocation models without consideration

of scheduling costs. An important concern here is strategic ordering. For example, distributors may order more than their ideal requirements, in order to receive larger order allocations. Cachon and Lariviere [11] show that capacity allocation mechanisms using proportional and linear allocation rules are subject to order inflation, whereas those using lexicographic and uniform allocation rules are truth inducing. Strategic ordering increases the manufacturer's profits, but this may be more than offset by reduced profit at the distributors. The net effect on overall supply chain profit tends to be positive when the wholesale price is high, and strongly negative when it is low. Cachon and Lariviere [12] consider two types of capacity allocation mechanisms: those that induce truth telling by the distributors, and those that are vulnerable to strategic ordering. It is shown that truth telling helps the allocation of capacity among the distributors, but may distort the choice of total capacity level.

Capacity allocation schemes based on past sales may benefit manufacturers at the cost of distributors. Cachon and Lariviere [13] consider a manufacturer which sells to two distributors over a two-period planning horizon. Demand in each period has two possible states, high and low. The manufacturer chooses a wholesale price and a capacity level. The distributors choose their order quantities. The authors demonstrate that a turn-and-earn policy based on past sales may benefit the manufacturer more than a fixed capacity allocation scheme based on order size. However, the distributors and the overall supply chain may be less profitable. This occurs because the turn-and-earn policy may induce the distributors to sell more than is optimally profitable when the demand is low, in order to increase their future capacity allocations. The manufacturer may change its wholesale price and capacity level to account for the distributors' increased sales, which can compensate the distributors' and overall supply chain profits. This compensation is sufficient to improve overall supply chain performance, except where capacity is extremely tight.

More complex allocation schemes may also be useful. Fang and Whinston [14] design an

option contract for a supply chain where, from the supplier's perspective, the value of capacity for each retailer is either high or low. The supply chain with the option contract achieves the same expected total return as if the supplier knows the number of retailers of each type before investing in capacity. Ganesh *et al.* [15] develop a congestion pricing mechanism for allocating bandwidth in communication networks. It is shown that there exists a unit price such that if all users predict that price and select their transmission rates accordingly, then the resulting price coincides with the prediction.

The works [11–15] do not consider scheduling cost. However, if capacity is time-sensitive as in many make-to-order environments, additional cost savings are obtainable by including scheduling issues in capacity allocation decisions. As discussed in the previous section, a centralized optimal solution can save up to about 30% of scheduling cost, relative to a typical capacity allocation that does not consider scheduling. Since in many practical situations centralized decisions are not possible, we next consider noncooperative capacity allocation problems that include scheduling issues. A typical assumption that promotes problem solvability is that each distributor requests production capacity that can be modeled as a single order or job, as shown in Table 1, in the column "Product."

We introduce the use of auctions to allocate production capacity, and then discuss examples of mechanisms that seek true information from distributors in order to find a centralized allocation of capacity. Wellman *et al.* [16] consider a capacity allocation problem involving several agents, each with a single job that allows preemption. The agents maximize their profit, which is the value of their scheduled jobs, less their scheduling cost and their cost of purchasing capacity. The facility owner maximizes its total revenue. The authors describe ascending auction mechanisms with two alternative market goods: time slots and time slot bundles. A time slot is a single unit of capacity at a fixed time, and a time slot bundle consists of a set of slots with the last one at a

fixed time. If time slots are chosen as market goods, then an equilibrium solution is globally optimal. It is possible that no equilibrium solution exists, due to the presence of complementarities. Secondly, with time slot bundles as market goods, the system value of an equilibrium solution is sometimes not good enough to be useful in practice. For a similar problem, Reeves *et al.* [17] illustrate the difficulty of making strategic choices in a simple ascending auction game. They show that straightforward bidding policies and their variants can provide capacity allocations that are not close to optimal.

Hall and Liu [18] study a capacity allocation and scheduling problem involving several buying agents and a single facility, where job preemption is not allowed. The objective of each agent is to maximize its profit, that is, its revenue less cost of capacity and scheduling. The objective of the facility owner is to maximize its revenue. Capacity is allocated using one of three market goods: time slots, fixed time blocks, and flexible time blocks. A fixed time block is any block of consecutive time slots that ends at a specific time, whereas a flexible time block is any block of consecutive time slots that ends no later than a specific time. The flexible time block auction typically provides more efficient schedules and better supply chain value than auctions using the other market goods, and on average more than 94% of the profit found by a centralized planning approach. Also, on average, flexible time blocks provide more profit for the distributors, and thus attract more buying agents to bid higher prices. With only a 6% increment in the number of agents bidding, or a 25% increment in bid prices, the facility owner achieves more revenue in the flexible time block auction than in the time slot auction. However, for both the auctions with fixed and flexible time blocks, there may not exist an equilibrium solution that is globally optimal.

Hain and Mitra [19] consider a problem where a manufacturer allocates capacity at a facility to multiple agents, each of which has a single job to process. The jobs have the same scheduling cost function that is publicly known, but a processing time known only to their agents. If the cost function is strictly

increasing and concave, then there exists a mechanism under which truth telling is a best response of an agent whenever other agents are truthful.

Bukchin and Hanany [20] analyze a dispatching and sequencing model where each department has a set of jobs and decides whether to process each job using an in-house resource with limited capacity or using a less efficient subcontractor. The objective is to minimize the total completion time of the jobs. A centralized schedule may not be a decentralized Nash equilibrium schedule. The decentralized Nash equilibrium schedule may not be unique, and finding it is an intractable problem. The ratio between the Nash equilibrium cost and the cost of a centralized optimal solution ranges from 1.00 to 1.35. A coordination mechanism is designed to guarantee that at equilibrium, a centralized optimal solution is found.

Vairaktarakis [21] considers multiple competing manufacturers which outsource their operations to a third-party capacity provider. Each manufacturer decides the amount of its workload to be outsourced, so as to minimize the completion time of its in-house and outsourced workloads. The paper develops pure Nash equilibrium schedules under several production specifications from the third party and information-sharing schemes of the manufacturers. Near-optimal supply chain performance is achieved if the third party chooses an appropriate incentive rule and the manufacturers share sufficient information.

FUTURE RESEARCH

For capacity allocation in supply chain scheduling that involves cooperative players, we propose the following interesting directions for further research:

1. Cooperation and cost sharing between manufacturers and distributors at the capacity planning stage may provide an improvement over uncoordinated decision making where manufacturers make capacity allocation decisions based on the cost of capacity, and

then distributors submit orders after the capacity level is determined (see **Capacity Allocation**).

2. It is valuable to study incentives, stability of cooperation, strategic behavior, and capacity planning within supply chain scheduling (see **Coordination of Production and Delivery in Supply Chain Scheduling: Origins and Application to Sequencing, Batching and Lot Sizing**).
3. It is interesting to extend sequencing games to consider more general scheduling environments and to find implementable cooperative mechanisms. See Refs 4–6.
4. Capacity-sharing games, involving both multiple manufacturers and multiple distributors, deserve careful exploration. See Refs 7–10.
5. Testing whether a given instance has a fair payoff division, and testing whether a given payoff division is fair, have been studied to only a limited extent. See Ref. 10.
6. Cooperative schedule planning games arise when various scheduling resources are owned by different agents. More investigation of such games is needed.
7. It is interesting to study cooperation with limited information sharing. Further, it should be possible to design mechanisms and incentives to encourage the players to release full and true information.
8. General theoretical structures may play an important role in successfully solving a specific game. For example, linear programming and duality have been applied to solve several important cooperative games within supply chains.
9. Several payoff divisions such as the Shapley value, nucleolus, and others, have not been studied extensively within capacity allocation games. Also, desirable properties of fair solutions, such as population monotonicity, should be studied.

10. One important problem is finding weakly fair payoff divisions, using concepts such as the ϵ -core and least core value, when a fair payoff division does not exist.

For capacity allocation in supply chain scheduling with noncooperative players, we propose the following interesting directions for further research.

1. Noncooperative capacity planning games involving multiple manufacturers should be studied. See Refs 11 and 12.
2. Competition between the distributors for demand from all the customers should be studied in the context of capacity allocation games. See Refs 11 and 12.
3. Multiechelon capacity allocation games with inventory-holding distributors model additional practical issues and therefore deserve investigation. See Ref. 13.
4. Capacity allocation with discrimination, either in prices or in capacity allocation mechanisms, or in both, is an interesting topic for future study. See Refs 13 and 14.
5. It is interesting to consider generalized games where distributors place orders that compete for both production capacity and order delivery time. See Ref. 15.
6. Combinatorial auctions are naturally useful for the allocation of discrete capacity and other scheduling resources. Research is needed to develop effective ways of solving the bid determination and winner determination problems. See Refs 16–18.
7. It is interesting to investigate heuristic-based equilibrium for scheduling jobs, due to the intractability of noncooperative scheduling games. See Refs 16–18.
8. It is important to conduct further studies of auctions of scheduling resources, since few practically effective auction mechanisms have been developed. See Refs 16–18.

9. Since it is difficult to predict the strategies of bidders, it is important to design robust auction mechanisms that are insensitive to players' bidding strategies and policies. See Refs 16–18.
10. Opportunities exist for modeling manufacturers' scheduling costs in various relevant ways while analyzing capacity allocation decisions. See Refs 19–21.

SUMMARY

Capacity allocation problems in manufacturing can be studied using both cooperative and noncooperative games. Such problems include under-allocation and over-allocation of capacity, and originate from either the demand or the supply side. Cooperative game models for capacity allocation identify profit divisions that ensure a stable coalition of the players. Recent research suggests various ways in which the extra value or cost saving from a centralized solution approach can be distributed. Noncooperative game models address the problem of strategic ordering, where players introduce false information in order to gain an advantage. Incentives need to be built into the manufacturing system to prevent this. Naturally occurring allocations of capacity often form a Nash equilibrium. However, Nash equilibrium solutions may not be close to an optimal solution for the overall supply chain. Within noncooperative games, auctions offer several useful features, especially the need to share only very limited information, such as a price, with competitors. Auctions need to be carefully designed in capacity allocation problems, since the resulting supply chain performance is highly sensitive to the choice of market goods. Moreover, different measures of supply chain performance recommend different choices for market goods.

REFERENCES

1. Gunasekaran A, Ngai EWT. Build-to-order supply chain management: A literature review and framework for development. *J Oper Manage* 2005;23:423–451.
2. Curiel I, Pederzoli G, Tijs S. Sequencing games. *Eur J Oper Res* 1989;40:344–351.
3. Curiel I, Potters J, Prasad R, *et al.* Sequencing and cooperation. *Oper Res* 1994;42:566–568.
4. Slikker M. Relaxed sequencing games have a nonempty core. *Nav Res Log* 2006;53:235–242.
5. Maniquet F. A characterization of the Shapley value in queueing problems. *J Econ Theory* 2003;109:90–103.
6. Katta AK, Sethuraman J. Cooperation in queues. Working Paper. New York: Department of Industrial Engineering and Operations Research, Columbia University; 2006.
7. Aydinliyim T, Vairaktarakis GL. Coordination of outsourced operations to minimize weighted flow time and capacity booking costs. *Manuf Serv Oper Manage* 2010;12:236–255.
8. Cai X, Vairaktarakis GL. Cooperative strategies for manufacturing planning with negotiable third-party capacity. Working paper. Hong Kong, China: Department of Systems Engineering & Engineering Management, Chinese University of Hong Kong; 2007.
9. Vairaktarakis GL, Aydinliyim T. Centralization vs. competition in subcontracting operations. Working paper. Cleveland (OH): Department of Operations, Case Western Reserve University; 2008.
10. Hall NG, Liu Z. Capacity allocation and scheduling in supply chains. *Operations Research* 2010. In press.
11. Cachon GP, Lariviere MA. An equilibrium analysis of linear, proportional and uniform allocation of scarce capacity. *IIE Trans* 1999;31:835–849.
12. Cachon GP, Lariviere MA. Capacity choice and allocation: Strategic behavior and supply chain performance. *Manage Sci* 1999;45:1091–1108.
13. Cachon GP, Lariviere MA. Capacity allocation using past sales: when to turn-and-earn. *Manage Sci* 1999;45:685–703.
14. Fang F, Whinston A. Option contracts and capacity management: enabling price discrimination under demand uncertainty. *Prod Oper Manage* 2007;16:125–137.
15. Ganesh A, Laevens K, Steinberg R. Congestion pricing and noncooperative games in communication networks. *Oper Res* 2007;55:430–438.
16. Wellman MP, Walsh WE, Wurman PR, MacKie-Mason JK. Auction protocols for decentralized scheduling. *Games Econ Behav* 2001;35:271–303.

17. Reeves DM, Wellman MP, MacKie-Mason JK, *et al.* Exploring bidding strategies for market-based scheduling. *Decis Support Syst* 2005;39:67–85.
18. Hall NG, Liu Z. Auctions for competitive capacity allocation and scheduling. Working paper. Columbus (OH): Department of Management Sciences, Ohio State University; 2008.
19. Hain R, Mitra M. Simple sequencing problems with interdependent costs. *Games Econ Behav* 2004;48:271–291.
20. Bukchin Y, Hanany E. Decentralization cost in scheduling: a game-theoretic approach. *Manuf Serv Oper Manage* 2007;9:263–275.
21. Vairaktarakis GL. Non-cooperative outsourcing games. Working paper. Cleveland (OH): Department of Operations, Case Western Reserve University; 2007.

CAPACITY ALLOCATION

MARTIN A. LARIVIERE
Kellogg School of Management,
Northwestern University,
Evanston, Illinois

Seldom in life is it always possible to make everyone happy. Supply chains are no different. When an upstream supplier in a distribution system has limited capacity or inventory, downstream demand may exceed available supply. The supplier must then employ a *capacity allocation mechanism* to convert an infeasible set of demands into a feasible set of output assignments. Such a scheme inherently divides the downstream markets into winners and losers as it is possible to increase the allocation to one market only by shorting another. In an integrated system, the decision maker can dole out available output in order to maximize the system-wide return. Here we will focus on decentralized supply chains in which the supplier and the buyers are all independent firms seeking to maximize their own profits. In such a setting, the allocation mechanism does not exist in a vacuum and buyers may distort their orders or other actions in order to gain a more favorable allocation. The allocation scheme may consequently be seen as part of the supplier's marketing mix for influencing the behavior of its channel partners.

ALLOCATION MECHANISMS IN PRACTICE

Allocation schemes have wide application and considerable consequences. At one time or another, pharmaceuticals [1], computers [2], paper towels, and liquid detergent [3] have all been on allocation. Allocation of scarce capacity also plays a major role in the automotive industry, and a significant body of research has focused on this setting. There are several reasons for this. First, automakers are much bigger than their dealers. Where a packaged goods manufacturer might have to negotiate with powerful

retailers, automakers are able to dictate their terms. Second, several government bodies in the United States regulate dealings between automakers and their dealers, so automakers must explicitly formulate and announce their allocation schemes. Finally, because vehicles are expensive and have large profit margins, receiving more of a hot product makes a nontrivial difference in a dealership's profitability.

The case of Dave Smith Motors illustrates the impact allocation schemes can have. Located in Kellogg, Idaho, a town of less than 3000 people, the firm sold over 4000 Dodge pickups in 1996, more than any other dealer in the nation. Key to Dave Smith's ability to sell so many vehicles was Dodge's allocation system. Like most automakers in the US market, Dodge used a variant of *turn-and-earn*. Under this scheme, an initial allocation of vehicles is made and a dealership earns more product by turning (i.e., selling) units. In the case of Dave Smith Motors, its willingness to price aggressively—and Dodge's willingness to keep sending it more vehicles—allowed it to dominate its region [4]. Indeed, other dealers in the Pacific Northwest threatened to boycott Dodge unless the firm altered its allocation scheme to reign in Dave Smith Motors. The Federal Trade Commission ultimately intervened, siding with Dave Smith Motors [5].

As the example demonstrates, allocation schemes impact supply-chain performance. Several points are worth noting. First, turn-and-earn creates an incentive for dealers to "move the metal" and ramp up their sales rate [6]. This ability to direct product to where it is selling well is seen as one of the main virtues of turn-and-earn allocation. Toyota moved its Chinese dealership network to turn-and-earn allocation for this very reason. Toyota previously based its allocations on forecasts. When realized sales deviated from forecasts, there was little ability to redirect stock. Some dealers had excess inventory while others were turning away customers [7]. However, this

incentive to drive up sales is arguably a drawback of turn-and-earn since it may lead to high pressure tactics that increase sales today but risk long-term loyalty and brand value [8]. Further, turn-and-earn has been linked to commercial fraud. Dealers have falsified sales in order to win better allocations [9].

Turn-and-earn favors dealerships that somehow achieve scale and may create long-lasting asymmetries between dealers. A dealership cannot sell what it does not have, and it cannot receive more of a popular product unless it sells. Much like Dave Smith Motors, Woodhouse Ford of Blair, Nebraska, became a regional powerhouse by aggressively discounting full-sized pickup trucks. It is the third-largest Ford dealership in the nation, sells more F-series pickups than anyone else and outsells the next biggest dealership in the area by over 4000 vehicles a year. Again the automaker's reliance on turn-and-earn fueled the growth of a "super" dealer [10]. Beyond allowing some dealers to develop an entrenched lead, turn-and-earn can create regional biases in how firms allocate inventory. Toyota has traditionally been stronger on the West Coast of the United States than in the Midwest, making it hard for Midwestern dealers to get enough supply of popular models. Some Midwestern dealers have viewed national economic slowdowns as opportunities to expand their allocations [11].

Finally, it is worth noting that while allocation systems impact the actions of retailers and the evolution of markets, it is the supplier that gets to choose and control the allocation system. Dave Smith Motors is not merely a Dodge dealer. The firm owns franchises for a number of brands including GMC which like Dodge, counts on pickup trucks for much of its volume. Dave Smith Motors follows the same advertising and pricing practices for its GMC and Dodge dealerships. The GMC franchise has been successful but has not been as dominant (or troubling to other dealers) as the Dodge dealership. This has been in part because GMC has maintained tighter control over its allocation policies and not sent as much product to the dealer.

CAPACITY ALLOCATION AND THE BULLWHIP EFFECT

Academic interest in allocation schemes can be traced to the seminal work by Lee *et al.* [12] on the bullwhip effect. The bullwhip effect is the empirical supply-chain phenomenon in which the variance of order placed on upstream firms exceeds the variance of demand seen by downstream firms (see *Information Sharing in Supply Chains*). A potential cause of the bullwhip effect is shortage gaming. Consider a supplier selling to symmetric retailers. Each retailer is a monopolist in its market and faces a newsvendor problem. The supplier's total capacity is uncertain, and the retailers must order before knowing their individual demand realization or the total available capacity. The supplier sells at a fixed wholesale price and commits to allocating capacity in proportion to orders. If realized capacity is insufficient to meet all orders, retailers receive a fraction of the available capacity equal to the fraction of total orders they submitted. Thus, if a retailer's order represented one-fourth of all orders, he will receive one-fourth of realized capacity when capacity is tight.

The allocation mechanism forces the retailers into a game. They are monopolists in their markets but must compete for capacity. How a retailer should order depends on how every other retailer orders. Lee *et al.* [12] characterize the equilibrium of this game and show that the possibility of being placed on allocation induces the retailers to over-order, that is, to order more than they would if supply were certain to be ample and the sole concern were maximizing profits. The bullwhip effect thus appears. They also suggest a countermeasure to mitigate the bullwhip effect: allocate capacity based on past sales as with the auto industry's turn-and-earn system. Turn-and-earn allocation enforces a maximum allocation the retailer can receive and hence limits the ability to over-order.

This simple model and possible remedy has shaped the research agenda on capacity allocation mechanisms. Two streams of work have followed from it: one has focused on

one-period models and examined the properties of simple allocation mechanisms, while the other has looked at multiperiod models and the impact of turn-and-earn.

CAPACITY ALLOCATION IN A ONE-PERIOD MODEL

In examining single-period allocation models, we follow the formulation of Ref. 13. A single supplier sells to N buyers, usually thought of as retailers. The supplier has a fixed capacity K . We will consider having the supplier-set capacity at the end of this section. Like Lee *et al.* [12], we restrict the supplier to using a price-only contract with wholesale price w per unit. We will say something about alternative terms of trade in the final summary. Also, we ignore any scheduling costs or timing issues. For a discussion of capacity allocation in the context of scheduling problems, (see **Information Sharing in Supply Chains**). Finally, we assume that today's allocation depends only on today's orders and does not consider other factors such as customer satisfaction scores or the length of the retailer-supplier relationship although these may be relevant in practice.

Each retailer is a local monopolist. (We say something about competition later.) Retailer i 's revenue $R_i(y_i, \theta_i)$ depends on the amount of inventory (he stocks) and some market-specific parameter θ_i . The retailer may, for example, face a newsvendor problem and θ_i would be a parameter of the demand distribution. Only retailer i observes θ_i . $R_i(y_i, \theta_i)$ is concave in y_i and retailer i 's profit is maximized at $y_i^*(\theta_i)$. Let m_i denote the quantity retailer i orders from the supplier. $\Theta = \{\theta_1, \dots, \theta_N\}$ and $\mathbf{m} = \{m_1, \dots, m_N\}$. $F(\Theta)$ is the joint distribution of Θ , which is common knowledge. In contrast to Ref. 12, uncertainty about whether capacity will be adequate is driven by the retailers' needs and not uncertainty in the supplier's capacity.

The assumed sequence of events begins with the supplier announcing her allocations scheme $A(\mathbf{m}) = \{a_1(\mathbf{m}), \dots, a_N(\mathbf{m})\}$. Retailers then observe their respective pieces of market information and order. Next, the supplier allocates capacity according to $A(\mathbf{m})$,

receiving $wa_i(\mathbf{m})$ from retailer i . Retailer i then receives revenue $R_i(a_i(\mathbf{m}), \theta_i)$. Observe that the model implicitly prohibits transshipments between retailers.

$A(\mathbf{m})$ is a mapping from the set of orders to feasible allocations and must conform to three basic rules. First, if $\sum_{j=1}^N m_j \leq K$, $a_i(\mathbf{m}) = m_i$. Second, if $\sum_{j=1}^N m_j \geq K$, $\sum_{j=1}^N a_j(\mathbf{m}) = K$. Finally, $a_i(\mathbf{m}) \leq m_i$. The first assumption simply supposes that the allocation scheme is relevant only when capacity binds while the second assures that no capacity is wasted. The last assumption keeps the supplier from sending the retailer more than he requested.

A difference between the usual analysis of capacity allocation schemes and mechanism design models in the economics literature [14] deserves mention. The economics literature typically appeals to the revelation principle and concentrates on direct mechanisms in which agents truthfully report their private information. The question is then, what mechanism to use. The literature on allocation mechanisms generally fixes the mechanism and examines the behavior the scheme induces. In particular, the analysis studies whether it is ever the case that the retailer submits an order that exceeds his profit-maximizing quantity $y_i^*(\theta_i)$.

A frequently studied allocation function is *proportional allocation* [12,15]. Here $a_i(m_i) = \frac{m_i}{\sum_{j=1}^N m_j} K$. Under proportional allocation, each retailer loses the same fraction of his order. Alternatively, all retailers could have their order reduced by the same absolute amount. This leads to *linear allocation*. Ideally, all retailers would have their order reduced by $\sum_{j=1}^N m_j - K$. Obviously, a retailer with a very small order may be given a negative allocation. Linear allocation deals with this possibility by sequencing orders in decreasing size (i.e., $m_1 \geq \dots \geq m_N$) and fixing a value n' . We then have

$$a_i(\mathbf{m}, n') = \begin{cases} m_i - \frac{1}{n'} \left(\sum_{j=1}^{n'} m_j - K \right), & \text{if } i \leq n', \\ 0, & \text{if } i > n'. \end{cases}$$

n' is the largest integer less than or equal to N such that all retailers receive nonnegative allocations.

Both linear and proportional allocation are *individually responsive* [13]. If retailer i receives a nonnegative allocation when he submits m_i , he will receive a larger allocation if he submits $m'_i > m_i$ (assuming the other retailers' orders are unchanged). Many allocation rules share this property. In particular, Cachon and Lariviere [13] show that if the supplier knew each retailer's private information and wanted to allocate stock to maximize the sum of total profits, the resulting allocation rule would be individually responsive.

However, not all allocation rules are individually responsive. For example, consider *lexicographic allocation* or *uniform allocation*. Under lexicographic allocation, retailers are sequenced in a manner independent of their order size (say via a lottery) and retailer 1 receives $a_1(\mathbf{m}) = \min\{m_1, K\}$ and each subsequent retailer receives $a_i(\mathbf{m}) = \min\{m_i, K - \sum_{j=1}^{i-1} m_j, 0\}$. Uniform allocation begins like linear allocation by sequencing retailers such that $m_1 \geq \dots \geq m_N$ and fixing an integer $\tilde{n}$. Retailer i is then allocated

$$a_i(\mathbf{m}, \tilde{n}) = \begin{cases} \frac{1}{\tilde{n}} \left(K - \sum_{j=\tilde{n}+1}^N m_j \right), & \text{if } i \leq \tilde{n}, \\ m_i, & \text{if } i > \tilde{n}. \end{cases}$$

$\tilde{n}$ is the largest integer less than or equal to N such that $a_{\tilde{n}}(\mathbf{m}, \tilde{n}) \leq m_{\tilde{n}}$. Intuitively, uniform allocation first offers each retailer K/N units. If this exceeds some retailer's request, that retailer receives her full order and the excess capacity is split evenly among the other retailers.

Neither lexicographic nor uniform allocation spreads the pain equally. Under lexicographic allocation, at least one retailer will get as close as possible to his order while others may get nothing. Uniform allocation favors smaller markets. Small orders are filled completely while those placing larger orders all receive the same quantity.

We now turn to how retailers order, given an allocation mechanism. It is natural to

consider a Bayesian equilibrium in which (roughly speaking) each player seeks to maximize his payoff given his private information and his expectation of how others will play. Such an equilibrium may not necessarily exist. For an analysis of when an equilibrium exists, see Ref. 15. Suppose that an equilibrium exists, Cachon and Lariviere [13] show that it cannot be the case that all retailers truthfully report their optimal quantities $y_i^*(\theta_i)$ if the allocation mechanism is individually responsive. Further, any equilibrium under such a mechanism must involve some retailers inflating their orders.

Over-ordering and the bullwhip effect thus plague a large class of allocation mechanisms. In particular, if the supplier wanted to allocate inventory in order to maximize supply-chain profits, the retailers will inflate their orders and distort the allocation. That, of course, begs the question of whether there are any allocation mechanisms that result in the retailers truthfully submitting their optimal quantities. Clearly, the scheme cannot be individually responsive. A stronger condition delivers the desired sufficient condition. Let $(m'_i, m_{-i}) = \{m_1, \dots, m_{i-1}, m'_i, m_{i+1}, \dots, m_N\}$. Then suppose that for a given allocation mechanism $A(\mathbf{m})$ there does not exist an m'_i such that $a_i(m'_i, m_{-i}) > a_i(m_i, m_{-i})$ for all i and (m_i, m_{-i}) such that $a_i(m_i, m_{-i}) < m_i$. Then all retailers truthfully reporting their optimal stocking quantities (i.e., $m_i = y_i^*(\theta_i)$) is an equilibrium.

Both lexicographic and uniform allocation meet the stated requirements. Intuitively, if a retailer receives less than his full order under these schemes, no action can increase his order. (The assumption that the sequencing under lexicographic order is independent of the order size is essential. If orders were filled from, say, smallest to largest, this would not hold.) Under either lexicographic or uniform allocation, asking for more increases a retailer's allocation only if he is already receiving all of his order.

This result is in some sense robust. Cachon and Lariviere [13] show that under such a mechanism, truthful reporting is a dominant strategy. Retailer i will order $y_i^*(\theta_i)$ even if others deviate from the equilibrium. Further, it remains an equilibrium even

when all retailers know that capacity is certain to bind, a setting in which proportional or linear allocation would fail to deliver an equilibrium.

This is not to say that a blanket endorsement of either rule is in order. First, these schemes may not always induce truthful revelation when the setup of the model is tweaked slightly. For example, consider the basic issue of the bullwhip effect. Having distorted information is problematic when the upstream supplier plans on using that information. Here, information is simply being used to allocate current output. It is not being used for other decisions such as future capacity expansions. Obviously, if the retailers realize that today's orders will influence tomorrow's availability, they may well alter their orders.

Truthful revelation will also not necessarily withstand the introduction of competition for customers. Consider a market with two retailers engaged in quantity competition. The characteristics of the market are fixed and commonly known. Given that they bring quantities y_1 and y_2 to the market, the retail price will be $P(y_1, y_2) = \theta - (y_1 + y_2)$. If capacity were certain to be available and the retailers order simultaneously, they would each order the standard Cournot quantity $y_C = (\theta - w)/3$, where w is again the supplier's wholesale price. Now, suppose that capacity is less than $\theta - w$ and that the supplier commits to using lexicographic allocation where the ordering will be determined randomly after orders are submitted. For simplicity, assume that the retailers cannot withhold stock from the market (i.e., y_i must equal to $a_i(\mathbf{m})$). Having $m_i = y_C$ is not necessarily an equilibrium. Specifically, suppose that $m_1 = K$ but that $m_2 < K$. If retailer 2 is sequenced second by the supplier, he receives nothing and earns zero profit. If he is sequenced first, he will receive m_2 but the prevailing price will be $\theta - K$ since retailer 1 is certain to take all the remaining capacity. Given that the retail price exceeds w and is fixed given $m_1 = K$, retailer 2's best response is to also order K . Given competition, lexicographic allocation results in extreme order inflation and a *de facto* monopoly in the retail

market. For more on allocation with competition, see Refs 16–18.

A second consideration is that a one-period model is in some ways inadequate for the questions being asked. For example, in a one-period setting, how can order inflation hurt the supplier? It is well known that a supply chain operating under a simple wholesale price contract suffers from double marginalization (see **Supply Chain Coordination**). Because the supplier (presumably) sells above his or her marginal cost of production, each retailer's optimal sales quantity is too low from a supply-chain perspective. Competition in the retail market is one way of addressing the resulting inefficiency. Alternatively, an individually responsive allocation mechanism introduces competition for capacity, inducing each retailer to order more and (in some states of the world) sell more than they would if a nonindividually responsive scheme were used.

Further, higher orders imply higher sales for the supplier, increasing her profit. The supplier's gain must come at the expense of the retailers. The overall outcome for the supply chain is unclear. In a numerical study, Cachon and Lariviere [13] show that supply-chain profits generally improve when the allocation method is switched from uniform to linear if capacity is fixed. It is even possible that average retailer profits improve with the move. For intuition on why supply-chain profit increases, note that the derivative of a retailer's profit function at $y_i^*(\theta_i)$ is zero. Hence, increasing the order size is initially costless to the retailer but strictly beneficial to the supplier. To understand how the retailers may benefit from being forced to compete more aggressively for stock, observe that uniform allocation leaves open the possibility of profitable trades between the retailers. A retailer in a small market receives his full order and values the last unit at zero while a retailer in a large market is shorted and would pay a premium over the wholesale price for another unit. Linear allocation reduces the extremes of these marginal valuations. Retailers in smaller markets will see their profits fall but the gains captured by

those in large markets may swamp the small market losses.

Allowing the supplier to choose his or her capacity adds a wrinkle. If the supplier uses a truth-inducing mechanism such as uniform allocation and has a constant marginal cost of capacity c , she faces a newsvendor problem (see **News vendor Models**). Let $\Psi_U(x)$ denote the distribution of total orders the supplier receives when she uses uniform allocation. $\Psi_U(x)$ does not depend on K . Consequently, the optimal capacity for uniform allocation K_U solves $\Psi_U(K_U) = \frac{w-c}{w}$. The distribution of total orders under linear allocation $\Psi_L(x|K)$ does depend on available capacity. Cachon and Lariviere [13] shows that each retailer orders less as capacity increases, implying that $\Psi_L(x|K_0) \leq \Psi_L(x|K_1)$ for $K_0 \leq K_1$. They further show that for the optimal capacity under linear allocation K_L , it must be that $\Psi_L(K_L|K_L) \leq \frac{w-c}{w}$. Thus, when retailers are ordering strategically, it will seem as if the supplier undershoots the relevant newsvendor fractile. However, if she were to expand capacity, demand would collapse.

These results do not guarantee an ordering of K_U and K_L . Numerically, one can show that moving from uniform to linear allocation can result in either an increase or a decrease in capacity. An increase in capacity is likely when the cost of capacity is high relative to the wholesale price. Under uniform allocation, such settings result in very low capacity. Moving to linear allocation allows the supplier to expand capacity while still generating enough competition so that system capacity is nearly fully utilized. This is yet another way that an allocation mechanism that induces over-ordering can improve supply-chain profit: more aggressive ordering may result in more capacity being built.

TURN-AND-EARN: BASING ALLOCATION ON SALES HISTORIES

We now examine the implications of tying current capacity allocation to prior sales as exemplified by the auto industry's practice of turn-and-earn. Any model of such a system must have at least two sales periods. Also, there must be some variation in demand so

that capacity is not binding (assuming that retailers order myopically) in some states but is binding in others. Low demand states allow one to study how the possibility of being put on allocation in the future induces retailers to alter their current sales policies in order to capture a larger share of the capacity in the future.

This is the basic structure assumed in Ref. 19. The model has two retailers who are both local monopolists. Each market is characterized by a linear demand curve. Given a sales quantity y , the retail price will be $P(y) = \theta - y$, and a retailer's current profit is maximized by selling $y(\theta) = \frac{\theta-w}{2}$. The intercept θ can vary by period and takes one of two values. In the high demand state $\theta = 1$ and in the low demand state $\theta = \alpha < 1$. The markets are perfectly, positively correlated: when demand is strong in one market, it is strong in the other.

The supplier has a fixed capacity K and $\alpha - w < K < 1 - w$. When demand is low, there is enough capacity for both retailers to receive their profit-maximizing quantities. When demand is high, the supply chain will be capacity constrained. There are two periods, and the wholesale price is fixed over the horizon. For simplicity, ignore discounting and suppose first-period demand is low ($\theta = \alpha$) while second-period demand is certain to be high ($\theta = 1$). (It is straightforward to allow second-period demand uncertainty.) Further assume that the retailers cannot carry inventory from one period to the other. We will explain how carrying inventory changes the model momentarily.

Two allocation methods are considered, fixed allocation and turn-and-earn. Under fixed allocation, each retailer has a "guaranteed allocation" of half of available capacity (i.e., he or she is certain to receive $K/2$ if he orders at least $K/2$). He may also claim any capacity that the other retailer does not want. Hence, given orders m_i and m_j , retailer i is allocated $a_i(m_i, m_j) = \min\{m_i, K/2 + [K/2 - m_j]^+\}$, where $x^+ = \max\{x, 0\}$. Fixed allocation is arguably a mere straw man, but it is simple, stationary, and guarantees symmetric treatment of the symmetric retailers. It thus forms a useful base case.

Analysis of the retailers' actions is straightforward. Fixed allocation functions much like uniform allocation in the single-period setting and thus, there is no reason for a retailer to distort his order. Each retailer orders $y(\alpha)$ in the first period and $y(1)$ in the second. Given our assumptions on capacity, each receives $y(\alpha)$ in the first period but gets only $K/2$ in the second.

Turn-and-earn is more involved. First-period allocations are the same as under fixed allocation, but second-period allocations depend on first-period sales. Let q_i denote retailer i 's first-period sales quantity and suppose $q_i \geq q_j$. Retailer i is then designated as the leader and will receive a favorable allocation in the second period. Specifically, his guaranteed allocation rises from $K/2$ to $K/2 + q_i - q_j$ and given orders, he receives

$$a_i(m_i, m_j) = \min\{m_i, K/2 + q_i - q_j + [K/2 + q_j - q_i - m_j]^+\}.$$

Retailer i 's gain must come at the expense of retailer j . As the follower, his guaranteed allocation is $K/2 + q_j - q_i$ and

$$a_j(m_j, m_i) = \min\{m_j, K/2 + q_j - q_i + [K/2 + q_i - q_j - m_i]^+\}.$$

As the players move into the high demand second period, there are distinct advantages to being the leader. Capacity is certain to bind and at least the follower (if not both players) will have allocations below $y(1)$. Additional supply would then have a positive marginal valuation. This leads to a distortion in first-period sales. If retailer j anticipates being stuck as the follower in the second period, increasing his first-period sales above $y(\alpha)$ has zero initial marginal cost but has a positive marginal benefit (holding retailer i 's sales quantity fixed).

Cachon and Lariviere [19] show that the retailers do, indeed, increase their first-period sales. The supplier then certainly benefits from turn-and-earn. His or her first-period sales increase relative to fixed allocation and the capacity remains fully utilized in the second period. The news for the retailers is less optimistic. They effectively

play to a draw. The unique equilibrium has them selling the same quantity in the first period. Increasing sales is more a defensive necessity than a means to a second-period advantage. The impact on supply-chain profit is consequently unclear.

Moving from fixed allocation to turn-and-earn also has an ambiguous impact on the supplier's choices when the supplier is free to set capacity and wholesale price. Suppose that demand in the second period is not certain to be high but will be high with probability ϕ . The supplier under fixed allocation will choose either $K = 2y(\alpha)$ or $K = 2y(1)$. Turn-and-earn admits a third possibility $k(w)$ such that $2y(\alpha) < k(w) < 2y(1)$. $k(w)$ is the largest capacity that will be fully utilized in the first period under turn-and-earn. This raises the possibility that the supplier brings more capacity to the market under turn-and-earn. He or she may also lower the wholesale price. Either of these changes may produce a sufficiently large benefit for the retailers so that they prefer competing under turn-and-earn to operating under fixed allocation.

Given that supply may be tight in the second period, having the retailers carry inventory between periods would be desirable. When this is added to the model, the retailers continue to increase their first-period sales rate under turn-and-earn. They may also distort the inventory decision. When capacity is very tight, the retailers reduce their inventory in order to bolster their first-period sales. This is clearly bad for the supply chain; the supplier's sales remain fixed (because capacity binds in both periods) and the retailers are selling more at a low price in the first period instead of saving inventory for the larger, more profitable second-period market.

In extending this model, it is straightforward to see that if there is another action that will increase sales, the retailers will increase this effort as well. For example, the retailers might face newsvendor problems instead of deterministic demand curves. If they could stochastically increase the demand distributions they face via advertising, they would advertise more under turn-and-earn than under fixed allocation. Similarly, if the retailers sold substitute products such as new and used cars, they would sacrifice sales of an

unconstrained product (used cars) to boost sales of a capacity-constrained product (new cars). This clearly benefits the supplier. However, suppose that the supplier provides both products (say, two car models in the same product line). Now if allocation of the hot-selling, capacity-constrained model depends only on its sales, the retailers have an incentive to reduce sales of the other product, which may harm the supplier. Here, the supplier would benefit from tying the allocation of a hot product to the total sales of the product line. See Ref. 20 for more on this logic.

The basic model of Ref. 19 captures the incentive to move the metal inherent in turn-and-earn but leaves a number of important issues unexamined. The symmetric, deterministic nature of the model keeps one retailer from gaining an advantage in equilibrium. Even if it were possible to gain an advantage, its limited two-period horizon does not allow one to examine the extent to which a leading retailer would fight to maintain his allocation. Finally, interest in turn-and-earn followed from research on the bullwhip effect but nothing in the model allows one to examine whether this allocation scheme helps reduce the bullwhip effect.

Lu and Lariviere [21] tackle a number of these issues. They consider an infinite horizon model in which the retailers face linear demand curves. The demand intercepts now have two components. The first is similar to the original [19] and is correlated across markets. This again reflects common market conditions but can now take three values. Under fixed allocation, capacity would bind only in the highest demand state. The common market condition evolves according to a Markov process. The second intercept term is a local shock that is drawn independently for each market. The shock for market i is only observed by retailer i . These modifications to the model allow for a richer story. Three common demand states imply that turn-and-earn induce higher sales in some states with excess capacity but not necessarily in all such states. The local shocks create the possibility that one retailer might gain an advantage. Further, the privately observed noise allows

one to evaluate the impact of turn-and-earn on the variance of demand placed on the supplier.

The analysis begins by finding a Markov perfect equilibrium between the retailers. As one would expect, the retailers still increase their sales relative to fixed allocation. More interesting is what happens when one retailer gains an advantage. The leader is willing to incur some cost (through selling more in less favorable markets) in order to maintain his lead. How long leadership persists depends on capacity tightness and demand volatility. Limited capacity makes leadership more valuable. Limited demand volatility keeps the leader from having to face the lowest demand state. Once the market enters the lowest demand state, the leader gives up trying to maintain his position. Stated another way, a recession re-levels the playing field. Finally, Lu and Lariviere [21] show that turn-and-earn can mitigate the bullwhip effect. By creating an incentive to boost sales, turn-and-earn induces the retailers to absorb unfavorable demand shocks as opposed to passing them on to the supplier. Thus turn-and-earn not only divorces allocation from possibly inflated orders, it actively counters the bullwhip effect.

SUMMARY

The choice of a capacity allocation mechanism is an effective—and to some extent underappreciated—means for a supplier to influence the behavior of downstream supply-chain partners. It influences both, how downstream buyers order and how they act. A well-designed allocation scheme has the potential to increase the supplier's profit as well as total supply-chain performance. The buyers may also benefit, especially when the allocation scheme induces the supplier to provide more capacity.

There are a number of open research issues in capacity allocation. One is the impact of alternative terms of trade. Essentially, all work on allocation schemes assumes that transactions are governed by wholesale price-only contracts. Consequently, given a total amount of orders,

the supplier's profit is fixed. This would not be the case under, say, a returns policy if retailers faced stochastic demand that varied across markets. If the supplier were responsible for unsold stock, she would have a greater interest in efficiently allocating capacity.

Another open issue is how to deal with asymmetric retailers. Some existing models allow retailers to be privately differentiated ex-post. The supplier cannot take advantage of this information upfront in designing the allocation scheme. In reality, a supplier knows which retailer has a favorable location in a large city and which is barely getting by in a rural town or which is a large chain representing a significant part of the supplier's sales and which is a mom-and-pop operation. A well-conceived allocation scheme should be able to take advantage of this knowledge. How this can be done for the benefit of the supplier or the supply chain while conforming to relevant antitrust laws has not been studied.

Finally, existing work has assumed a monopolist supplier. This is sensible to the extent that, say, a car dealer has only one source for a hot model. However, products exist in competitive markets and retailers have a choice over which products to stock or for which to exert promotional effort. How allocation schemes impact these decisions is an open question.

REFERENCES

1. Hwang S, Valeriano L. Marketers and consumers get the jitters over severe shortages of nicotine patches. *Wall St J* 1992; May 22: B-1.
2. Zarley C, Damore K. Backlogs plague HP: resellers place phantom orders to get more products. *Comput Reseller News* 1996; May.
3. Associated Press. Shortage has P&G allocating detergent. *Associated Press Newswire* 1997; Aug 6.
4. Jackson K. Internet ad has dealers fuming. *Automot News* 1997; Jan 6.
5. Associated Press. Federal Trade Commission settles charges alleging boycott threat. *Associated Press Newswire* 1998; Aug 6.
6. LaReau J. GM to increase enclave output. *Automot News* 2008; Jan 21.
7. Shirouzu N. Foreign models: in Chinese market, Toyota's strategy is made in U.S.A. *Wall St J* 2006; May 26.
8. Lavin D. Youwannadeal? Bucking Detroit trend, landmark Chevrolet still uses the hard sell. *Wall St J* 1994; July 8.
9. Lynch S. *Arrogance and accords*. Dallas (TX): Pecos Press; 1984.
10. Mortimer J, Wilson A, Shereffkin R. How small-town Ford store dominates region, rankles rivals. *Automot News* 2006; May 15.
11. Rehtin M. Toyota division inventories soar to record high. *Automot News* 2008; Apr 28.
12. Lee HL, Padmanabhan V, Whang S. Information distortion in a supply chain: the bullwhip effect. *Manage Sci* 1997;43(4):546-558.
13. Cachon GP, Lariviere MA. Capacity choice and allocation: strategic behavior and supply chain performance. *Manage Sci* 1999;45(8): 1091-1108.
14. Salanié B. *The economics of contracts*. Cambridge (MA): The MIT Press; 1997.
15. Cachon GP, Lariviere MA. An equilibrium analysis of linear and proportional allocation of scarce capacity. *IIE Trans* 1999; 31(9):835-850.
16. Chen F, Li J, Zhang H. Retail competition, capacity allocation, and supply performance. Working paper, Columbia University; 2007.
17. Furuhata M, Perrussel L, Zhang D. Mechanism design for capacity allocation under price competition. Working paper, University of Western Sydney; 2007.
18. Furuhata M, Zhang D. Capacity allocation with competitive retailers. *Proceedings of the 8th International Conference on Electronic Commerce*; 2006 Aug 14-16; Fredericton, Canada. 2006. pp. 31-37.
19. Cachon GP, Lariviere MA. Capacity allocation using past sales: when to turn-and-earn. *Manage Sci* 1999;45(5):685-703.
20. Purohit D, Vernik D. Turn-and-earn in a product line: The impact of product substitutability. Working paper, Duke University; 2008.
21. Lu LX, Lariviere MA. Capacity allocation over a long horizon: the return on turn-and-earn. Working paper, University of North Carolina; 2008.

CAPACITY PLANNING IN HEALTH CARE

JONATHAN PATRICK
PEDRAM NOGHANI
University of Ottawa, Ottawa,
ON, Canada

At one level, one could argue that capacity planning has been made relatively simple by long-standing equations developed in the queuing theory literature. Given a performance metric of the form, $x\%$ of patients need to be seen within y units of time, queuing theory can easily determine the necessary capacity to make this a reality. In truth, if even these simple formulations were used more frequently in health organizations, we would not see the kind of demand/supply mismatches that are so common and lead to such incredibly long wait times. An excellent example of the benefit of these simple queuing methodologies can be found in the research of Green [1, 2].

The importance of the form of the performance metric, though well known in the literature [3], often goes unrecognized by healthcare managers. Clearly without a performance metric there is no required capacity while as soon as one is imposed (most helpfully in the form given above) then implicitly a capacity requirement has been imposed as well. Back of the envelope rough estimations using queuing theory are helpful but obviously a lot more can be said about the additional challenges that complicate most capacity planning problems in health care. The ones we will touch on briefly in this article are patient behavior, the presence of multiple patient classes, and the strong possibility of downstream blocking. These are by no means an exhaustive list but they are perhaps three of the more common and significant challenges. We will first discuss the impact of these additional challenges and then circle back to discuss the attempts to address them and the challenges that yet

remain. This article can hardly hope to provide an exhaustive look at the literature but hopefully will provide some lay of the land as well as some hints as to potential solutions and future directions.

CHALLENGES TO CAPACITY PLANNING IN HEALTH CARE

Patients are not widgets. This quite obvious statement has a number of implications for capacity planning in health care. It means that one can assume that each patient coming through the door is going to have some unique features in terms of the type of resources required to treat them as well as the urgency with which they need to access those resources. Moreover, the non-widget like nature of patients means that they will not necessarily behave as one would like or imagine.

Patient Behavior

Outpatient clinics, for instance, are constantly dealing with idle capacity created by patients failing to show for an appointment. This has led to a prolific stream of literature on optimal overbooking policies as well as a strong push for what is called *open access* or *advanced access* scheduling. However, while this is the most commonly researched impact of patient behavior, other impacts are also present. Most research, for instance, assumes that patients will accept the appointment offered to them. In reality, wait times are often artificially increased by patients who refuse to take the first appointment offered to them either because it does not fit their own schedule or because they choose to wait for their preferred provider. These user-side delays are rarely recorded in the data sets leading to an inflated sense of the actual wait times for a given health service (though how inflated is a matter of contention) and posing a significant challenge in terms of planning capacity.

Presence of Multiple Patient Classes

Many health needs are not planned months in advance where one has the leisure of smoothing out demand through intelligent advanced scheduling. Health needs are often sudden and urgent leading to many health services having a significant proportion of demand that must be served close to, or even on the day of, the original request. Such a reality greatly hampers the ability of a health service to manage the available capacity intelligently.

The individuality of patients also becomes a significant challenging factor as patients arrive to a health service not only with varying levels of urgency but also with varying conditions that require differing amounts of resource consumption as well as potentially different types of resources. For example, many hospitals have close to 1000 different surgical types each with their own separate distribution for length of surgery as well as length of recovery time and some of which require specialized equipment. How much capacity is needed in such a situation is clearly dependent on how these patients are scheduled. Poor scheduling, for instance, could lead to peaks and valleys in the distribution of occupancy rates in the wards leading to a higher capacity requirement. A more intelligent scheduling policy could meet the same performance target with significantly less capacity. In addition, the larger the urgent class and the shorter the window within which to book them, the less flexibility one has in scheduling and thus, again, the greater the required capacity in order to meet the same performance target.

Downstream Blocking

A third complicating factor is that patient care rarely involves a single service. More frequently, patients move through a series of services and not necessarily in a serial manner as they may loop back to revisit a previous service due to a complication. Thus, for instance, a patient may enter the emergency department of a hospital, be transferred to a ward bed for a period of time and then require time in a rehabilitation facility before finally ending up in a long-term care home. From the

long-term care home, they may relapse and require a second visit to the hospital. What happens all too frequently is that a patient gets stalled in this journey due to a lack of capacity at the next stage in their care. In much of the developed world, there is a significant back log of patients who remain in the hospital after the acute phase of treatment is finished simply because there is no capacity at the community care facility (i.e., long-term care home) at the next stage in their care pathway. These patients, therefore, block acute care (AC) beds delaying patients in the emergency department who are awaiting admittance to the wards. Treating a single facility in isolation (as the simple queuing equations mentioned at the outset would do) is therefore often insufficient as the impact of downstream blocking may mean a significant increase in the required capacity to meet the specified performance target.

Surge Capacity

Finally, it is crucial to remember that it is people being serviced and that failure to receive service in a timely manner can have dire consequences to their well-being. Thus, one cannot simply write-off the tail of the wait-time distribution. The common practice used in other settings of refusing service in order to deal with surges in demand is often not an option for a health service. This leads to the absolute necessity for health services to invest in surge capacity—capacity that an organization can turn to in periods of high demand. Surge capacity can take on a number of forms but the two most common are overtime and transfers to another facility. These typically come with a higher service cost but can often be more cost-effective than continually carrying the additional capacity required to meet the peaks in demand.

What the above discussion hopefully demonstrates is twofold. First, health care capacity planning is often complicated by additional factors that make the simple capacity requirement calculations less effective than in other settings. Second, the required capacity is determined not only by the performance target that has been put in place but also by how the complicating

factors are dealt with both through investment in surge capacity and the choice of scheduling policy. Thus, any discussion on capacity planning needs to make mention of how that capacity is managed. The primary manner by which capacity is managed is through the schedule.

The health care scheduling literature can roughly be divided into *advanced scheduling* and *appointment scheduling*. Advanced scheduling refers to the policy that determines how far in advance to book each request for service. It is through advanced scheduling that the daily fluctuations in demand are smoothed out—an advantage that can play a significant role in reducing the capacity required to meet a given performance target. Appointment scheduling refers to the policy that determines the order and the start times of each patient booked into a given day. It is through the appointment schedule that the wait times on the day of service, the idle time, and the use of overtime are controlled.

ADDRESSING THE CHALLENGES

Patient Behavior

Patient choice is a challenging area of research because (i) it is unpredictable by nature and (ii) very few health services collect the data that would allow you to model it. For instance, few health services collect information on what booking appointments were refused by a patient and for what reason. Gupta and Wang [4,5] developed multiple Markov decision process (MDP) models that mimic patient choice both in terms of provider preference (loyalty to their own physician in a multi-physician clinic) and in terms of time of day preference. They examine a single day's booking from a revenue management perspective with the decision variables being whether to accept or reject an incoming request for service. They were able to partially describe the form of the optimal policy but implementation remains a challenge as the policy depends on data that is rarely collected. Feldman *et al.* [6] provide some evidence for the time of day preference of patients based on a

discrete choice experiment conducted at a large health center. They build a model for appointment scheduling that incorporates these choice decisions where the decision of the provider is which appointment slots to offer to each arriving demand. They incorporate both cancellations and no-shows and demonstrate the superiority of a dynamic system over a static one. In other words, they demonstrate improved performance associated with taking into account the state of the system (i.e., current bookings) and patient behavior in determining the booking policy.

Data on patient no-shows are more commonly collected with studies that have linked no-shows to the length of the lead time (time between request and service), the postal code of the patient, and the weather to name just three [7]. An impressive amount of research has led to a number of policies based on different formulations of the problem all telling a very similar and somewhat predictable story—no-shows can be mitigated by overbooking but how that overbooking is best implemented is an open question [8–11]. Dome-shaped policies that book clients closer together early and late in the day are generally accepted to work well provided that service times come from the same distribution for all patients [12,13]. Double booking early in the day has also demonstrated fairly good performance as has the Bailey–Welch rule that schedules a number of patients right at the beginning of the day and then separates further appointments by the average service time [14]. Of course, which policy is deemed optimal depends in large part on the relative importance one places on the relevant performance metrics. Most papers in this vein use some combination of patient wait time, server idle time, and overtime. The actual weightings play a significant part in terms of the shape of the optimal policy. For instance, increased weight on the wait time of patients will clearly be detrimental to the Bailey–Welch rule, whereas if overtime and idle time are the primary concerns then the Bailey–Welch rule may do fairly well. It is worth noting that, in most cases, the capacity is deemed to be fixed when determining the optimal appointment schedule, whereas

it is quite likely that the optimal appointment schedule may in fact be a function of both the available regular-hour capacity and the surge capacity options. This is certainly an avenue for future research. Also lacking at present is a model that determines both the order and the start times of a set of patients booked on a given day. Most appointment scheduling models either assume the order is known [15] or else assume patients are of one type (thus negating the importance of the order). One heuristic that appears to perform well is to book patients starting with the one with the least variable service time and working down to the one with the most variable service time. However, this has the detrimental side-effect of leaving the more complex patients to later in the day.

The detrimental impact of no-shows combined with patient dissatisfaction with over-booking has led some researchers to advocate for a form of scheduling called *open access* that essentially books patients on the day the request is placed [16]. The rationale is that no-shows are much less likely if the lead time is zero and moreover same day access is generally seen as an important benefit for the patient. However, it is clear that advanced scheduling provides significant advantages in terms of both the amount of regular-hour capacity required and the amount of surge capacity utilized. Even a small booking window of one day [17] or a few days [18] has been shown to significantly improve the efficiency of a clinic compared to “open access.” Models that determine the optimal (potentially dynamic) booking window that would allow a health service to take advantage of advanced scheduling without unnecessarily inflating wait times would certainly add value to the literature.

Presence of Multiple Patient Classes

Less widely studied has been the impact of multiple patient classes. This is unfortunate as the impact of multiple patient classes is perhaps of much greater import than the impact of “no-shows” while at the same time not suffering from the lack of data that compromises much of the research on patient behavior. Patient classes are generally the

result of differences in urgency and/or differences in resource consumption. The general challenge is how to manage capacity in such a way that each class receives service in a timely manner and such that the available capacity is used efficiently. Gerchak and Gupta [19] looked at this problem in the context of elective and emergency surgeries with the decision simply being the number of elective surgeries to book on a given day. They demonstrated that the optimal policy is not simply a cut-off policy that reserves a fixed number of slots for emergency surgeries but rather that the number of elective surgeries to book depends on the length of the elective wait list. However, they do not provide an actual schedule for the elective surgeries. Models that seek to book patients from multiple classes in advance quickly run into tractability issues. Thus, work has focused on approximation techniques [20, 21]. One of the take home messages from this line of research is that the driving force behind the amount of capacity required is the size of the highest priority class and even more importantly the length of the window within which they have to be booked. In other words, flexibility in the booking of the highest priority class is the best means of reducing capacity requirements [22]. Second, intelligent use of short bursts of surge capacity can greatly reduce the necessary capacity and significantly reduce the cost of meeting a given performance target [21]. As mentioned earlier, one cannot simply ignore the tail of the wait-time distribution and thus fixed wait-time targets are often assigned instead of the probabilistic rule mentioned at the outset. However, rather than seeking to build sufficient regular hour capacity to meet the target, it is often much more beneficial to address the tail through surge capacity. More recent work, based on more sophisticated approximation techniques, has demonstrated significant improvements in capacity utilization (in terms of reduced idle time and surge capacity utilization) over previous attempts yielding policies that will on occasion book patients late (if the downstream congestion is minimal) or will act more pro-actively to initiate surge capacity in states where anticipated congestion is significant [23].

One significant area of potential research involves the intelligent merger of the appointment scheduling and advanced scheduling problems. Currently, these problems are treated as independent problems by solving advanced scheduling problems as though service times were deterministic (thereby avoiding the impact of the appointment schedule) or else by solving appointment scheduling problems as though the number of patients to be scheduled was an exogenous variable outside of the control of the manager. Ideally, these two scheduling problems would be solved together as the performance of the one is clearly dependent on the form of the other.

Downstream Blocking

More challenging still are those settings where appropriate capacity planning requires the simultaneous consideration of multiple resources and where each patient's flow through the various services is far from straightforward. Two excellent examples of this challenge are surgical capacity planning (where patients require both time in the operating room (OR) and recovery time in the wards) and patient flow through acute care (AC) and into community care (where patients need access to certain community services before they can be discharged from AC). Tackling the first has led to a number of models that attempt to minimize the peak load in the ward based on restrictions on the performance of the ORs—usually in the form of ensuring that a certain throughput is maintained or a given threshold utilization is achieved [24–26]. Seeking to optimize capacity in the ORs and the wards simultaneously becomes a multi-objective problem where the desire is to meet at least three competing performance targets—wait-time thresholds for patients, efficient use of OR time, and efficient use of ward capacity. In this setting, being significantly under capacity is often viewed as just as much an issue as being over capacity. One attempt to solve the combined problem can be found in Astaraky and Patrick [27] who develop an MDP model that seeks to minimize precisely this three-pronged objective. Tractability issues force them to apply

approximation methods that are nonetheless able to demonstrate significant improvement in wait-time management without significant increases in overtime in the OR or congestion in the wards. Key to their results is the grouping of patients within a surgical specialty through clustering techniques so that within each group service time distributions (both OR time and recovery time) are reasonably tight and well-defined. Implementation would involve wresting away some of the control of the scheduling process that currently resides with the surgeon. With only the block schedule (which surgeon receives OR time on what day) under the control of most hospitals, capacity planning that adequately addresses both OR capacity and ward capacity is severely hampered.

The progression of patients out of AC and into community care has not received nearly enough attention in the operations research literature considering the considerable shift in emphasis within health services away from the hospital and toward the community. Day surgery has vastly increased followed by extensive rehabilitation services provided in the community. The increasing number of elderly patients who require 24-h (non-acute) care has led to significant increases in the need for long term care facilities and supportive housing. The lack of appropriate capacity plans for these facilities has led to significant backlogs in hospitals as patients whose AC treatment is finished block acute beds for want of the appropriate services in the community [28].

The importance of this issue in health care was highlighted at least 25 years ago with the work of Weiss and McClain [29]. They developed a queuing system with state-dependent service rates in order to model discharges from the hospital that require community service. Weiss and McClain predict the probability distribution of the number of clients waiting in the hospital for transfer to a community service. Their model, however, does not provide a means of predicting the impact of downstream blocking.

There is, however, a stream of literature that has progressively improved the ability of queuing network models to predict the

impact of downstream blocking. This literature can be described as a series of progressively improved heuristics for estimating the probability of blocking given a capacity plan for a network of services. All of the heuristics provide a means of estimating the blocking probability and the estimated size of the wait list at each node. The improvements in the heuristics are twofold: better accuracy in small networks where the optimal solution can be computed and a reduction in the limiting assumptions that allow for the application to a wider variety of scenarios. The first of these heuristics was developed by Takahashi *et al.* [30] using an open queuing network system with feed-forward flows and finite buffers between stations. They introduce an approximation method for open restricted queuing networks in which the service time and arrivals have exponential and Poisson distributions, respectively. The approximation method is capable of providing various performance measures such as blocking probabilities and output rates of general open restricted queuing networks.

In a more recent health applications study, Koizumi *et al.* [31] use a queuing network system with blocking to analyze congestion in a mental health system in Philadelphia. They use a multi-server model to present both mathematical and simulation results. Their system consists of three types of psychiatric institutions. Patients enter the network from either an acute hospital or the community and proceed through the network in a sequential manner (no feedback loops). They use the same approximation methodology that Takahashi *et al.* developed but extended to a multi-server model. Two performance indicators are used in their model: number of patients waiting to enter each facility of the system and the associated waiting time in steady state.

Finally, a recent study by Brettahauer *et al.* [32] presents a new heuristic method for queuing networks with blocking. They use a queuing network system to model patient flow between intensive care (ICU), step-down (SD), AC, and post-acute care units (PAC). Of note is that the heuristic can be applied to scenarios where patient flow involves feedback loops or where patients may enter

the system at any node in the network. The aim of their study is to find the best mix of beds across the hospital's inpatient units in order to provide the highest quality of care possible given a limited budget. Since the approximation of the blocking probabilities is determined algorithmically, there is a significant challenge presented in utilizing these heuristics to determine the optimal capacity at each node (as there is no closed form solution for the blocking probability). Brettahauer *et al.* resort to enumeration to find the optimal bed allocation (in order to minimize a weighted sum of the probability of blocking at each node) between the services given a fixed budget. A method for optimally determining capacity that nonetheless makes use of the performance metrics outlined in algorithms such as Brettahauer *et al.*'s would certainly add to the literature.

What all three heuristics demonstrate is the significant impact of downstream blocking. Healthcare managers have time and again neglected downstream blocking effects, choosing to focus on the high impact or critical resources (such as an emergency department) often at the expense of health resources that are of less immediate consequence (such as long term care). Such short-sighted planning has led to significant issues in precisely those high impact regions of the health system—not because they are under-capacitated but simply because of the inability to off-load patients in a timely manner to underfunded, under-resourced downstream facilities.

INSIGHTS FOR HEALTHCARE MANAGERS

It is understandable, given the complexity of managing capacity in a health care setting, that managers have often felt overwhelmed and have been limited to reacting to crisis rather than anticipating it. What managers need to be aware of is that there are viable models that can help determine and minimize capacity requirements through intelligent scheduling—even in complex situations where multiple patient classes are involved with potential no-shows and with multiple resources consumed in sequence.

Some of these models do require sophisticated software for implementation but others have developed simple heuristic policies that could be easily implemented. Managers also need to realize that limited use of overtime is not a negative and may in fact be the most cost-effective means of meeting a given performance target. Certainly some form of surge capacity is an essential component of an efficiently run health service.

For the operations researcher, there is a wealth of potential in this field for future avenues of research many of which have been mentioned above. One key message is that we need to stop looking at part of the health system in isolation but rather to recognize the inter-connectedness between the various parts and the impact that inter-connectedness has on how we plan and manage capacity. We also need to provide a more reflective assessment of health care capacity planning that looks at long term trends and does not necessarily focus on the current point of crisis that may in fact be merely a symptom of a much larger problem.

CONCLUSION

The individuality of patients with each having unique health service needs has led many health managers to believe that quantitative models simply cannot be of much use. This is, in my mind, an over-reaction to what is a legitimate concern. Operations researchers working in the field of healthcare management need to recognize this individuality and the significant complexity it adds to scheduling and capacity planning. It ought to instill in us a certain amount of humility and cautiousness in presenting the results of our models, recognizing the difficulty of accurately representing how people behave and the necessarily simplified representation of reality upon which our model depends. Einstein once wrote that “as far as the propositions of mathematics refer to reality, they are not certain and as far as they are certain they do not refer to reality.” As operations researchers, we are always seeking to find that balance between making the model complex enough to be useful and yet simple enough to be tractable. This unavoidable

trade-off is certainly evident in the numerous attempts to plan capacity in the health care setting. Approximation techniques are one means of increasing complexity without losing tractability though of course at the loss of optimality! Unfortunately, that loss of optimality may be necessary if our models are to contain the amount of complexity required in order to be of any use to healthcare managers.

REFERENCES

1. Green L. How many hospital beds? *Inquiry* 2002;39:400–412.
2. Green L, Nguyen V. Strategies for cutting hospital beds: the impact on patient service. *Health Serv Res* 2001;36(2):421–442.
3. Cohen M, Hershey J, Weiss E. Analysis of capacity decisions for progressive patient care hospital facilities. *Health Serv Res* 1980;15(2):145–160.
4. Gupta D, Wang L. Revenue management for a primary-care clinic in the presence of patient choice. *Oper Res* 2005;56(3):576–592.
5. Wang W, Gupta D. Adaptive appointment systems with patient preferences. *Manuf Serv Oper Manag* 2011;13(3):373–389.
6. Feldman J, Liu N, Topaloglu H, *et al.* Appointment scheduling under patient preference and no-show behaviour. *Oper Res* 2014;62(4):794–811.
7. Kopach R, DeLaurentis P, Lawley M, *et al.* Effects of clinical characteristics on successful open access scheduling. *Health Care Manag Sci* 2007;10:111–124.
8. Kim S, Giachetti R. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Trans Syst Man Cybern A Syst* 2006;36:1211–1219.
9. LaGanga L, Lawrence S. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sci* 2007;38:251–276.
10. Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans* 2008;40: 820–837.
11. Zeng B, Turkcan A, Lin J, *et al.* Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Ann Oper Res* 2009;178:121–144.
12. Wang P. Static and dynamic scheduling of customer arrivals to a single-server system. *Nav Res Log* 1993;40:345–360.

13. Denton B, Gupta D. A sequential bounding approach for optimal appointment scheduling. *IIE Trans* 2003;35:1003–1016.
14. Ho C, Lau H. Minimizing total cost in scheduling outpatient appointments. *Manag Sci* 1992;38:1750–1764.
15. Begen M, Queyranne M. Appointment scheduling with discrete random durations. *Math Oper Res* 2011;36(2):240–257.
16. Murray M, Tantau C. Redefining open access to primary care. *Manag Care Q* 1999;7:45–51.
17. Robinson L, Chen R. Traditional and open-access appointment scheduling policies: the effects of patient no-shows. *Manuf Serv Oper Manag* 2009;12:330–346.
18. Patrick J. A Markov decision model for determining optimal outpatient scheduling. *Health Care Manag Sci* 2012;15(2):91–102.
19. Gerchak Y, Gupta D, Henig M. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manag Sci* 1996;42:321–334.
20. Erdelyi A, Topaloglu H. Computing protection level policies for dynamic capacity allocation problems by using stochastic approximation methods. *IIE Trans* 2009;41(6):498–510.
21. Patrick J, Puterman M, Queyranne M. Dynamic multi-priority patient scheduling. *Oper Res* 2008;56(6):1507–1525.
22. Patrick J, Puterman M. Improving resource utilization for diagnostic services through exible inpatient scheduling. *J Oper Res Soc* 2007;58:235–245.
23. Saure A, Patrick J, Puterman M. (2014) Simulation-based approximate policy iteration with generalized logistic functions (in press with *INFORMS Journal on Computing*).
24. Belien J, Demeulemeester E. Building cyclic master surgery schedules with leveled resulting bed occupancy. *Eur J Oper Res* 2005;176:1185–1204.
25. Santibanez P, Begen M, Atkins A. Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. *Health Care Manag Sci* 2007;10(3):269–282.
26. Chow V, Puterman M, Salehirad N, *et al.* Reducing surgical ward congestion through improved surgical scheduling and uncappeditated simulation. *Prod Oper Manag* 2011;20:418–430.
27. Astaraky D, Patrick J. A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling. *Eur J Oper Res* 2015;245:309–319.
28. Patrick J. Access to long-term care: the true cause of hospital congestion? *Prod Oper Manag* 2011;20(3):347–358.
29. Weiss E, McClain J. Administrative days in acute care facilities: a queueing-analytic approach. *Oper Res* 1987;35(1):35–44.
30. Takahashi Y, Miyahara H, Hasegawa T. An approximation method for open restricted queueing networks. *Oper Res* 1980;28(3):594–602.
31. Koizumi N, Kuno E, Smith T. Modeling patient flows using a queueing network with blocking. *Health Care Manag Sci* 2005;8(1):49–60.
32. Bretthauer K, Heese H, Pun H, *et al.* Blocking in healthcare operations: a new heuristic and an application. *Prod Oper Manag* 2011;20(3):375–391.

CAPACITY PLANNING

SAMPATH RAJAGOPALAN
Marshall School of Business,
University of Southern
California, Los Angeles,
California

Capacity planning and expansion is an important activity for a firm for several reasons. First, capacity levels determine how much output a firm can produce which in turn determines the level of demand that can be satisfied. Also, capacity levels impact how quickly a firm can respond to customer requests or supply chain disruptions. Thus, capacity levels impact sales volume, revenues, and profits both in the short and long run. Second, capacity decisions involve significant capital commitments that are either partially or fully irreversible. For instance, once an oil company has built a refinery, it cannot easily change its decision and recover its costs fully. Third, capacity expansion decisions in many industries involve significant scale economies and therefore are made infrequently. So, they are based on long term demand forecasts which tend to be uncertain. For instance, capacity decisions in the power sector are based upon demand forecasts 20 years into the future [1]. Together with the irreversibility of these decisions, this implies that making a sound capacity decision is critical for a firm's operations. In this article, we discuss the main elements and trade-offs in capacity decisions and some typical models used in capacity planning.

MAIN ELEMENTS OF CAPACITY DECISIONS

The main elements of a capacity decision involve issues related to size, timing, type, and location of the capacity addition, and we discuss these in some detail next. Also, a facility often contains multiple resources and in some industries, capacity levels may

depend on the level of all these resources. For instance, at an airline, capacity is a function of the type and number of planes, number of gates at the airports it operates, pilots, flight crew, ground crew, and so on. So, careful consideration needs to be paid to all these factors. It should be kept in mind that several elements of capacity decisions are often intertwined but we discuss some of them separately for clarity in exposition.

Capacity can be defined as the *maximum output of a system or resource per unit time*. While this may appear to be precise, there are several measurement issues in practice. A steel plant's capacity may be defined as 500,000 tons per year but this may be a *theoretical* or *design* capacity (also referred to as *rated* or *engineering* capacity) that can be achieved only under ideal conditions. The *effective* capacity may be lower, say because the plant requires downtime for scheduled maintenance. The *actual* capacity may be even lower due to yields, defects, equipment breakdowns, and so on.

Size and Timing

Firms add to existing capacity in response to (unexpected) past demand growth and in anticipation of future demand growth. When a firm builds a plant of a certain capacity to meet this demand growth, the added capacity may be fully utilized after some years and it may have to add capacity again. Thus, the size of the current capacity addition impacts the time at which the next capacity addition is made and so the size and timing decisions are inextricably linked. For example, Genentech [2] found that demand for its drugs was growing fast and it had to make several capacity expansion decisions during the past decade and a half. The size of each expansion has impacted the timing of future expansions. One of the main benefits of a large expansion is economies of scale, that is, average cost per unit of capacity decreases as the expansion size increases. For instance, the capacity cost in heavy process industries such as chemicals and petroleum exhibit

substantial scale economies [3,4]. Scale economies arise due to the presence of fixed costs that do not vary with the volume of expansion or marginal costs that decline with volume. So, the capacity cost function is often represented as a concave increasing function of volume to depict the decreasing average unit costs of capacity. A drawback of a large expansion is that some of the capacity acquired is unlikely to be utilized for a long time. Given the high capital expenditures for capacity expansions, this implies a lower return on invested capital considering the time value of money. Equally important, such a large expansion is based on demand growth in the distant future which is more uncertain and may not materialize. For instance, Eskom, a power generator based in South Africa, committed to units of six 600 MW generators in the mid-1970s based on certain demand projections [5]. This resulted in an excess capacity of over 40% even after 15 years (i.e., in 1990) because the demand growth they had projected did not materialize and they had committed to irreversible contracts. On the other hand, it may be worthwhile to have significant excess capacity if the cost of lost sales is high and demand is very uncertain. This is the situation at Genentech [2] where the gross margin is high and there is considerable uncertainty about future demand for drugs, since they may or may not be approved for new uses. Similarly, when Nintendo launched the sales of Wii in late 2006, it was an instant hit and demand came from unexpected sources such as senior citizens. Nintendo's supply could not meet the demand, and there was a substantial supply shortage. The company had to expand its production capacity for the gaming console and had to make a critical capacity sizing decision [6]. If Nintendo substantially increased production capacity but the demand growth did not materialize, the company would absorb excess capacity cost and its profits would be impaired. However, if Nintendo underestimated the demand and did not expand its capacity enough, then the supply shortages and unmet demand would continue, possibly resulting in lost sales and customers.

Type

Capacity expansion planning also requires determining the types of resources to be added. We focus on two important aspects of the type of capacity: technology and flexibility. The technology of an equipment or facility determines several critical aspects of the products produced using that equipment: quality, speed, and unit cost. For instance, successive generations of computed tomography (CT) equipment have undergone revolutionary changes, with successive generations of equipment resulting in higher patient throughput and better image quality. Firms have to decide whether to buy the latest technology which may be more expensive and also whether to wait for some better technology. This is especially true in industries such as electronics where there is considerable technological change. Also, note that the size of the capacity additions is linked to the technology decision—large capacity expansions may not allow a firm to take advantage of frequent improvements in equipment technology.

Another important aspect of a resource type is the degree of flexibility of the resources in terms of their ability to produce different products, perform various tasks or operations. For example, an auto assembly plant may be able to assemble only one or multiple models of cars. Similarly, some flexible facilities are capable of producing a variety of different products. For instance, in the early 1990s, Eli Lilly had to decide whether to continue its past strategy of building dedicated capacity that could produce one product or adopt flexible facilities that could produce multiple products [7]. Flexible resources have the advantage of enabling a firm to adapt to fluctuations in demand for different products, thus increasing revenues and decreasing costs. However, they also require higher capital investments as flexible assets may be more expensive and may involve superior technologies. In some contexts, flexibility may require a more labor-intensive production process and this has implications for the type and size of capacity acquired as scale economies are very different in labor-intensive versus capital-intensive processes.

Location

In some contexts, firms have to make decisions about where to add capacity and in how many locations. Factors involved in making such location decisions are: customer locations, raw material supply locations, labor costs, transportation costs, exchange rates, tariffs, and so on. The focus of this article is not on location decisions; so we do not discuss this topic here (please refer to the section titled “Location Analysis” in this encyclopedia). But it is important to keep in mind that location issues are an important part of the capacity expansion decision. For instance, if shipping costs are substantially high, it is better for the firms to have capacities in multiple locations. For example, a soft drink producer will have numerous bottling plants as transportation costs are likely to be high. In this case, transportation costs between different locations need to be considered and minimized by determining optimal network configurations. If shipping costs are not as significant, firms can better utilize economies of scale by having capacity in a single location.

Alignment with Resources/Strategies

Organizations utilize many different types of resources and assets to perform various business activities at different stages of a production process. Capacity planning differs significantly depending on the type of production system and processes at a firm. In continuous flow processes such as steel mills, the production system is tightly integrated and capacity planning involves a simultaneous and proportional change in all the resources. On the other hand, in batch processes and job shops, capacity planning involves identifying appropriate capacity levels for different types of resources. For example, in apparel manufacturing, a firm will have to determine capacity levels of resources such as workers, cutting machines, dyeing capacity, and so on and this will vary with the type and mix of products produced. In such cases, capacity decision models usually assume that there is a single bottleneck (i.e., most utilized) resource, and capacity planning is based on this bottleneck resource. The facilities may be designed

around the bottleneck resource or stage, often the most expensive resource. In a job shop, such as a custom furniture manufacturer, capacity may be measured directly in terms of resource needs (number of lathes, saws, sanders, etc.) rather than products shipped.

Supply Chain Capacity Coordination

Capacity decisions often impact all the members of a supply chain, but each of these members may be independent entities. So, coordination and cooperation within the supply chain is important. In the Wii example cited earlier, some reports suggested that Nintendo could not produce enough consoles due to the tight supply of integral components such as IC chips and PCBs. Thus, even though Nintendo may have had sufficient assembly capacity, its suppliers did not have enough capacity to support their customer's production decision. This example illustrates the importance of coordination of the entire supply chain for capacity expansion. Ideally, capacity decisions need to be coordinated across a supply chain to prevent or reduce negative outcomes such as excess inventory, supply shortages, and congestion costs. Such issues are especially difficult to coordinate and manage in complex supply chains used to produce automobiles, apparel, and so on (see *Coordination of Production and Delivery in Supply Chain Scheduling* for additional details).

MODELS FOR CAPACITY PLANNING

Models are convenient ways to abstract reality and provide guidance for managers to make decisions. Several models have been developed to capture the main trade-offs in capacity expansion decisions and we present some of these models in the next few sections.

Deterministic Demand Models

In this section, we consider models where future demand is assumed to be known and in the next section, we consider uncertain future demand. While it may seem unrealistic to assume known future demand, we are really assuming a known demand forecast; so, it

is imperative that the use of such deterministic models be accompanied by substantial sensitivity or scenario analysis, as discussed in detail later. We assume that capacity shortages are not allowed unless specified otherwise. We now consider an infinite horizon model that captures the trade-off between making large capacity additions which result in lower purchase costs per unit of capacity versus small capacity additions, which result in a lower discounted cost of unused capacity (since early investments are more costly). Suppose demand is growing linearly over time at a rate of δ per year, that is, demand at time t is $D_t = \mu + \delta t$. Also, for convenience assume that capacity at time 0 is equal to μ . Then, at time 0, the firm will have to add capacity assuming they do not want any capacity shortages. Now, assume further that equipment or capacity purchased is infinitely durable. Suppose the same amount of capacity is added each time, and this amount is represented as x . Since demand is growing linearly at rate δ , note that $T = x/\delta$ is the time at which the next capacity addition will be made. At this time, again there will be no excess capacity and such a point is referred to as a *regeneration point*.

Let the cost of purchasing and installing capacity increment of size x be given by $g(x)$. Let the discount rate be " r " and so e^{-r} is the present value of a dollar at time t in the future. Let $C(x)$ denote the sum of all discounted future costs associated with purchase of current and future capacity increments. Since we have an infinite horizon problem, at every regeneration point, we will add a capacity unit x and the future costs are $C(x)$. We then have the following recursive equation

$$C(x) = g(x) + e^{-\frac{rx}{\delta}} C(x). \quad (1)$$

The first term on the right side of the equation is the purchase cost for the capacity increment installed now and the second term is the sum of all future capacity purchase costs discounted by the term $e^{-rx/\delta}$. Note that capacity additions of size x are made every time demand is equal to capacity, that is, at the regeneration point, and these occur at time intervals x/δ . From Equation (1), we

have

$$C(x) = \frac{g(x)}{\left(1 - e^{-\frac{rx}{\delta}}\right)}.$$

Example. If $g(x)$ is given by the power function kx^a where $k > 0, 0 < a < 1$, then

$$\frac{C(x)}{k} = \frac{x^a}{\left(1 - e^{-\frac{rx}{\delta}}\right)}.$$

Then the optimal $\hat{x}$ is obtained by differentiating $\log C(x)$ with respect to x and setting it equal to zero. We then have

$$\begin{aligned} \frac{d \log C(x)}{dx} &= \frac{a}{x} - \frac{\left(\frac{r}{\delta}\right) e^{-\frac{rx}{\delta}}}{\left(1 - e^{-\frac{rx}{\delta}}\right)} = 0 \\ \frac{\delta a}{r} &= \frac{\hat{x}}{\left(e^{\frac{r\hat{x}}{\delta}} - 1\right)}. \end{aligned}$$

Alternatively the optimal time $\hat{t} = \hat{x}/\delta$ between capacity additions is obtained by solving

$$\left(e^{r\hat{t}} - 1\right) = \frac{r\hat{t}}{a}.$$

For more general capacity cost functions, the optimal capacity additions can be obtained by a similar approach, that is, minimizing the discounted present value of capacity costs. While the linear demand function is easy to analyze and may represent reality in some contexts, note that the increase in demand is constant over time and does not increase as the total demand increases. In some contexts, total demand may grow at a constant percentage rate, that is, incremental demand is proportional to the volume. Such a demand pattern can be represented by $D_t = \mu e^{\delta t}$; note that demand growth is proportional to the demand volume at any point in time. Another demand function used in capacity models is

$$D_t = \beta(1 - e^{\delta t}).$$

This function represents demand with a decreasing growth rate over time, with a saturation level of β .

It has been shown [4] that the optimal policy is to add capacity at equal time intervals between expansions (although the size of each capacity addition may vary over time) when the capacity cost function is given by $g(x) = kx^a$ and demand is given by one of the three demand functions identified earlier or when cost function has a fixed plus variable cost structure, that is, $g(x) = F + vx$ and demand is increasing linearly.

Discrete-Time Model

The model presented in the previous section is a continuous-time model (i.e., demand, costs, etc. are computed continuously over time rather than at discrete-time points) which assumed a certain rate at which the demand varies over time. More flexible models allow demand to change at arbitrary rates over time and discrete-time models are used to represent such scenarios. We present a simple example of a discrete-time model next, assuming that the market is growing and only capacity additions are considered and shortages are not allowed. Let d_t, x_t , and I_t respectively denote the demand increment, capacity addition (or expansion), and excess capacity at end of period t ($t = 1, 2, \dots, T$). Note that d_t is the demand increment in a period and $d_t \geq 0$, so total demand is assumed to be nondecreasing over time. Let $f_t(x_t)$ denote the costs of purchasing and installing capacity and $h_t(I_t)$ the costs associated with excess capacity. The cost of capital or the appropriate discount factors are assumed to be incorporated in the functions $f_t(x_t)$ and $h_t(I_t)$ and are not identified explicitly. For simplicity, assume initial and final excess capacities are equal to 0. The formulation is

$$\begin{aligned} & \text{Minimize} \quad \sum_{t=1}^T (f_t(x_t) + h_t(I_t)) \\ & \text{subject to:} \quad I_t = I_{t-1} + x_t - D_t \quad \forall t \\ & \quad \quad \quad I_t, x_t \geq 0 \quad \forall t \\ & \quad \quad \quad I_0 = I_T = 0. \end{aligned}$$

While one could solve this problem directly as a nonlinear program, a simpler approach is to reformulate it as a dynamic program

(see the section titled “Fundamentals” in this encyclopedia) where the periods of the model are equivalent to the stages of the dynamic program. Let the state variable in the dynamic program be the excess capacity I_t . Let C_t denote the minimum total discounted cost during periods t through T . Then, the recursive equation is given by

$$\begin{aligned} C_t(I_t) = \min_{x_t} & (f_t(x_t) + h_t(I_t) \\ & + C_{t+1}(I_t + x_t - d_t)). \end{aligned}$$

While the dynamic program may be difficult to solve in general due to the curse of dimensionality, special cases of this formulation can be solved very efficiently. For instance, if the functions $f_t(x_t)$ and $h_t(I_t)$ are concave, the optimal solution is such that the capacity additions are equal to the demand increment in an integral number of periods, that is, $x_t = \sum_{\tau=1}^t d_\tau$. This dramatically reduces the state space and simplifies the solution approach. Furthermore, this model is equivalent to the economic lot sizing problem [8] (see also **Lot-Sizing**) and so many of the results obtained for that problem also apply here. This formulation and solution approach can be extended to incorporate negative demand increments, capacity disposals, capacity shortages, and so on [9]. This type of formulation has also been extended to consider discrete capacity sizes, capacity deterioration, multiple capacity types, and so on [9,10]. Also, such a formulation could be extended to consider the possibility of using the excess capacity to build inventory and carry it to future periods. Finally, in many real-world environments, equipment is not infinitely durable as we assumed earlier. So, equipment replacement decisions are often made together with capacity expansion decisions. The discrete-time model presented can be generalized to consider replacement together with capacity expansion decisions as discussed in Ref. 9.

The primary advantage of discrete-time models compared to the continuous-time models presented earlier is that they are more flexible in modeling the variation of demand over time. Also, various features of real capacity problems such as disposals and

shortages can be modeled easily. However, a drawback of deterministic models presented so far is that they do not explicitly consider the uncertainty in demand that is a key element of capacity expansion decisions.

Stochastic Demand Models

In this section, we consider models where future demand (or other parameters) is not known with certainty. In practice, many approaches have been proposed to deal with uncertainty such as decision trees, scenario planning [11], stochastic programming, and providing capacity cushions or buffers [10,12]. In the capacity cushions approach, we solve deterministic models with forecasted demands using the approach described in the section titled “Deterministic Demand Models” and use capacity buffers to deal with uncertainty. Such buffers can be computed using “newsvendor” type models [12] (see also **Supply Chain Coordination**) that trade-off the cost of lost sales with the cost of excess capacity. We do not discuss decision trees as they are similar in spirit to a special case of the stochastic programming approach discussed later in this section. Several models have been proposed that explicitly incorporate stochastic evolution of future demand and these require knowledge about the demand distributions. In this section, we discuss a basic model of this type.

An early seminal model to consider capacity expansion decisions when demand growth is probabilistic is due to Manne [1]. Manne was able to show that if demand follows a Brownian motion (see the section titled “Diffusion Processes and Random Walks” in this encyclopedia) with positive drift and capacity shortages are not allowed, then the stochastic model is equivalent to a deterministic model with a lower discount rate. Let us consider this model which assumes an infinite horizon. As in the deterministic case, let the cost of purchasing and installing capacity increment of size x be given by $f(x)$, discount rate be “ r ” and $C(x)$ denote the sum of all discounted future costs associated with the purchase of current and future capacity increments. Let the demand process be represented by a Brownian motion with the drift parameter

μ and variance σ^2 ; that is, demand is random and the demand increment at any time t is a random variable which is normally distributed with mean μt and a variance $\sigma^2 t$ [1]. We assume that capacity shortages are not allowed and capacity can be added instantaneously, that is the lead time for adding capacity is zero.

Manne shows that this stochastic capacity expansion problem can be written as the following problem

$$C(x) = f(x) + \int_{t=0}^{t=\infty} u(t, x) e^{-rx} C(x) dt, \quad (2)$$

where $u(t, x) dt$ is the probability that t is the time difference between two successive capacity additions (or regeneration points) between which total demand grows by x units. Manne further shows that

$$\int_{t=0}^{t=\infty} u(t, x) e^{-rx} C(x) dt = e^{\lambda x},$$

where $\lambda = \frac{\mu}{\sigma^2} \left(1 - \sqrt{\frac{1+2r\sigma^2}{\mu^2}}\right)$. Hence, the recursive Equation (2) simplifies to

$$C(x) = \frac{f(x)}{(1 - e^{\lambda x})}. \quad (3)$$

Comparing Equations (1) and (3), the stochastic problem is equivalent to the deterministic problem (1) with λ replacing $-r$, that is the negative of the discount rate r . It is straightforward to see from the expression for λ that an increase in the variance σ^2 results in a higher value of λ (note that λ is negative), that is, a smaller adjusted discount rate. Two important observations then follow: an increase in the variance σ^2 will result in higher costs $C(x)$ and the optimal capacity increment $\hat{x}$ will be larger. Thus, greater variance in the demand process results in larger capacity acquisitions.

Several subsequent articles have shown that this important transformation holds under more general conditions. Bean *et al.* [13] showed that if demand is a transformed (nonlinear) Brownian motion or a regenerative birth-and-death process (see the section

titled “Stochastic Processes” in this encyclopedia), then the problem can be transformed to an equivalent deterministic problem with a lower adjusted discount rate. It may be difficult in reality to estimate exactly the adjusted discount rate for the regenerative birth-and-death process as it requires knowledge of the complete distribution of the first passage time in the underlying process. But Bean *et al.* [13] point out that a very good second-order approximation of the discount rate can be obtained using only the coefficient of variation of the demand process. Furthermore, they point out that when the coefficient of variation is small, the adjusted discount rate is only slightly lower than the unadjusted discount rate. In such situations, it may not be costly to ignore the effects of demand uncertainty in making capacity decisions. However, if the demand variation is high, then it is critical to consider the impact of demand uncertainty.

Appropriate discount rate “r”: even in a deterministic problem, as Luss [4] points out, “since the planning horizon for capacity expansion problems is usually long, it is virtually impossible to forecast the *appropriate* discount rate over that period.” This issue becomes even trickier in a stochastic environment as we have to adjust the discount rate downwards to account for demand variance as observed earlier. Furthermore, technological changes impact future capacity purchase and operating costs, often resulting in lower costs. This can be incorporated by adjusting the discount rate appropriately but since effects of technological changes are difficult to forecast over long periods, it is difficult to estimate the appropriate adjustments to the discount rate. Therefore, extensive sensitivity analysis studies are common in capacity planning decisions. Next, we provide a systematic approach to capacity planning that is based on considering various future scenarios.

Stochastic Programming Model

An alternative approach proposed to deal with uncertainty in demand and other parameters is a stochastic programming based approach, which uses scenarios to model the uncertainty within large-scale mathematical programs. The advantage of

these approaches is that one can consider fairly general models wherein multiple resources used to make multiple products can be considered. Also, one could potentially consider scenarios incorporating different possible future realizations of demand, costs, and so on. The drawback of these approaches is that one does not obtain analytical solutions for the optimal capacity size and the problem size grows substantially with the number of scenarios considered which makes it difficult to solve the problems. We describe a generic version of the approaches proposed in Refs 11,14–16—we focus on one resource producing one product to keep the exposition simple. It is straightforward to extend the approach to multiple resources and products, although it increases the problem size and the corresponding difficulty in solving the model. The model is similar to the discrete-time model presented earlier except that we use different scenarios for demand and we use a fixed plus variable cost function for capacity costs. Also, such models incorporate production decisions where production takes place after demand is known and production is constrained by capacity acquisition decisions which are made before the demand is known.

We define variables x_t^s for capacity purchased, y_t^s for production level, and I_t^s for excess capacity in period t under scenario s . Let z_t^s denote a binary variable equal to 1 or 0 depending on whether or not capacity is purchased in period t in scenario s . The capacity acquisition costs comprise of fixed costs $f_t^s z_t^s$ and variable costs $v_t^s x_t^s$, excess capacity costs are $h_t^s(I_t^s)$ and the production costs are $c_t^s(y_t^s)$ corresponding to period t and scenario s . A scenario s corresponds to a particular realization of the parameters and p_s is the probability of scenario s ($\sum_{s=1}^S p_s = 1$). However, note that parameters get realized incrementally over time. So we have a scenario tree where the nodes at stage (or level) t of the tree, denoted as B_t , constitute the states of the world or scenarios, say $s_i \in B_t$, that can be distinguished by information available up to time stage t (we do not add an index t to s_i to keep the notation simple). Therefore, in the formulation below, s, s_i , and s_j all represent

different scenarios. The formulation is

Minimize

$$\sum_{s=1}^S p_s \left(\sum_{t=1}^T (f_t^s z_t^s + v_t^s x_t^s + c_t^s (y_t^s) + h_t^s (I_t^s)) \right) \quad (4)$$

subject to:

$$I_t^s = I_{t-1}^s + x_t^s - D_t^s \quad \forall t, s \quad (5)$$

$$y_t^s \leq \sum_{\tau=1}^t x_\tau^s \quad \forall t, s \quad (6)$$

$$x_t^s \leq M z_t^s \quad \forall t, s \quad (7)$$

$$z_t^{s_i} = z_t^{s_j}, x_t^{s_i} = x_t^{s_j}, y_t^{s_i} = y_t^{s_j}, I_t^{s_i} = I_t^{s_j} \quad \forall (s_i, s_j) \in B_t, i \neq j, \forall t \quad (8)$$

$$z_t^s \in \{0, 1\}, x_t^s, I_t^s, y_t^s \geq 0 \quad \forall t, s. \quad (9)$$

Constraint (5) is the excess capacity balance constraint. Constraint (6) enforces the condition that the production level in a period does not exceed the installed capacity. Constraint (7) ensures that capacity is added in period t in scenario s only if $z_t^s = 1$ (where M is a large number). At time t , the decision maker cannot distinguish between two scenarios s_i and s_j that belong to the same node B_t of the scenario tree. Consequently, the decisions corresponding to scenarios s_i and s_j have to be identical and this is imposed through the constraints (8), which are referred to as *nonanticipativity* constraints in Ref. 17. Nonnegativity and binary restrictions of the variables are enforced through (9).

Note that this problem is a stochastic mixed integer linear programming problem and is therefore difficult to solve. Hence, heuristic solution approaches are proposed and we briefly summarize the heuristic approach proposed in Ref. 14 as an example of a typical approach, although there are variations in the models and corresponding solution approaches proposed. In the first phase, the integrality constraints are relaxed and so we have a large stochastic linear program which can be solved using a variety of approaches developed specifically for large-scale linear programs. If this solution is also integral, then we have the optimal solution. Otherwise, in the second phase,

the *nonanticipativity* constraints are relaxed and this results in S instances of the deterministic capacity expansion problems, which can be solved optimally using conventional approaches such as the Wagner–Whitin procedure [8], if the subproblems are simple (e.g., single resource, single product with continuous capacity purchases possible) or more complex approaches depending on the nature of the subproblems. If these solutions to the subproblems satisfy the *nonanticipativity* constraints, then we have the optimal solution. Otherwise, we go to the third phase, wherein the solution obtained in the second phase is heuristically modified to ensure that it satisfies the *nonanticipativity* constraints. Ahmed and Shinidis [14] show that this approach works quite well. Huang and Ahmed [16] generalize and improve upon this approach by exploiting properties of the subproblems. They provide an efficient approximation scheme and an asymptotic optimality guarantee for the approximation scheme. So, it appears that the stochastic programming approach is potentially attractive in solving capacity expansion problems and one can expect further refinements and improvements in these approaches.

Most of the literature in this area has focused on capacity expansion in scenarios with demand growth. In reality, demand may decline too and capacity may have to be decreased. In many industries, the salvage value of capacity is often negligible, that is, only a fraction of the investments in capacity can be recovered if the facilities are sold. This is primarily for two reasons. First, investments may be firm-specific in which case there is no clear market value for the assets. Second, if the assets are more general-purpose, firms will often try to dispose of assets when industry demand is low and likely to remain low in the near future, but in those circumstances, other firms are facing the same industry conditions and so demand for such capacity is low and this leads to low salvage values for the assets. In scenarios where demand may rise or decline due to economic cycles, firms may invest, stay put or disinvest—Eberly and Van Mieghem [18] refer to this as an Invest/Stay put/Disinvest (ISD) policy and show that such a policy is

optimal when the operating profit functions are concave. Such a policy involves computing two triggers, say L_t and H_t , which are functions of the problem parameters and the following three action zones in each period t : (i) Invest, that is, increase in current capacity to L_t if it is less than L_t ; (ii) Stay put, that is, no action is taken if current capacity is between L_t and H_t ; (iii) Disinvest, that is, decrease in current capacity to H_t if it is greater than H_t . The width of the region $(H_t - L_t)$ is called a *hysteresis zone* and increases with the amount of irreversibility in investment costs, level of fixed costs of capacity adjustment, and indivisibility or lumpiness in capacity purchases and disposal.

Acknowledgment

I am grateful to David Cho, Marshall MBA 2008 for his assistance in the preparation of this manuscript.

REFERENCES

1. Manne AS. Capacity expansion and probabilistic growth. *Econometrica* 1961;29: 632–649.
2. Snow DC, Wheelwright SC, Wagonfeld AB. Genentech—Capacity planning, Harvard Business School Publishing, Product # 606052, 2006.
3. Erlenkotter D, Manne AS. Capacity expansion for India's nitrogenous fertilizer industry. *Manag Sci* 1968;14(10):B553–B572.
4. Luss H. Operations research and capacity expansion problems: a survey. *Oper Res* 1982;30:907–947.
5. Aberdeen D. Incorporating risk into power station investment decisions in South Africa S.M. Thesis. MIT; 1994.
6. Ayers C. Why the Wii shortage is legit, Wii.combo.com, Editorial, November 29, 2007.
7. Pisano GP, Rossi S. Eli Lilly and Co.: The flexible facility decision—1993. Harvard Business School Publishing, Product # 9-694-074; 1994.
8. Wagner HM, Whitin T. Dynamic version of the economic lot size model. *Manag Sci* 1958; 5:1770–1774.
9. Rajagopalan S. Capacity expansion and equipment replacement: a unified approach. *Oper Res* 1998;46(6):846–857.
10. Van Mieghem JA. Capacity management, investment, and hedging: review and recent developments. *Manuf Serv Oper Manag* 2003;5(4):269–302.
11. Eppen GD, Martin RK, Schrage L. A scenario approach to capacity planning. *Oper Res* 1989;37:517–527.
12. Hayes RH, Pisano GP, Upton DM, Wheelwright SC. Pursuing the competitive edge. Chapter 3. Hoboken (NJ): John Wiley; 2004.
13. Bean JC, Hagle JL, Smith RL. Capacity expansion under stochastic demands. *Oper Res* 1992;40:S210–S216.
14. Ahmed S, Sahinidis NV. An approximation scheme for stochastic integer programs arising in capacity expansion. *Oper Res* 2003;51:461–471.
15. Chen Z, Li S, Tirupati D. A scenario-based stochastic programming approach for technology and capacity planning. *Comput Oper Res* 2002;29(7):781–806.
16. Huang K, Ahmed S. The value of multi-stage stochastic programming in capacity planning under uncertainty. Working paper, Atlanta (GA): Georgia Institute of Technology; 2005.
17. Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
18. Eberly JC, Van Mieghem JA. Multi-factor dynamic investment under uncertainty. *J Econ Theor* 1997;75(2):345–387.

CATEGORY AND INVENTORY MANAGEMENT

YASIN ALAN
VISHAL GAUR
Johnson Graduate School of
Management, Cornell
University, Ithaca,
New York

SHELF SPACE ALLOCATION

Problem Description

Retail stores have a limited amount of space and many products to display. The amount of shelf space allocated to an item affects its frequency of replenishment, incidence of stockouts, and demand rate. Therefore, finding the optimal amount of shelf space to allocate to each item becomes a key factor for success. Effective shelf space allocation leads to higher profits by increasing sales and customer satisfaction, creating better product visibility and brand exposure, and reducing inventory-related costs and stockouts [1–3]. Indeed, effective shelf space allocation has become harder and more critical in recent years because of increases in product variety and competition. For instance, on average, the number of consumer-packaged stock keeping units (SKUs) in the marketplace increased by 16% per year between 1985 and 1992, whereas shelf space expanded by only 1.5% per year during the same period [4]. A modern conventional supermarket offering major food departments, nonfood grocery, and limited general merchandise products has 20,000 to 30,000 sq. ft of floor space and it carries 20,000 to 40,000 SKUs [[1], p. 40].

Shelf space allocation is the process of apportioning the amount of space to each product in order to maximize the total store profit or another well-defined objective function subject to limited store space and other financial and operational constraints. Shelf space is usually measured in linear terms

such as the number of *facings* and the length of each facing. The depth of a shelf and the dimensions of an item are also used in order to compute the shelf capacity allocated to each item.

Researchers from marketing, operations management, and other related fields have been developing tools and techniques for more effective shelf space allocation [5]. From a marketing perspective, the visual appeal of the display, the amount of product variety, and the location of each item in the store are important in order to maximize total sales. In particular, researchers define a metric called the *shelf space elasticity* of demand as the ratio of relative change in unit sales to relative change in shelf space. They measure shelf space elasticity for various types of products to evaluate the effect of display on sales. The key factors from an operations perspective are demand estimation, shelf replenishment frequency, inventory holding cost, the cost of stockouts, replenishment-related labor time and cost, and the impact of case pack sizes on inventory management policies. The operations perspective is more relevant for supermarkets, convenience stores, drug stores, and discount retailers because operational efficiency can lead to lower costs, which, in turn, increases competitiveness. The marketing issues are dominant for other types of retailers such as apparel stores in which the visual appeal of the display has a substantial demand-stimulating effect. Conflicting interests of the marketing and operations strategies also pose challenging questions.

In this article, we describe the measurement of shelf space elasticity of demand, the mathematical models and heuristic approaches to allocate shelf space, and commercial packages used in the industry.

Measurement of Shelf Space Elasticity

The amount of shelf space allocated to a product influences its unit sales. In one of the earliest studies on shelf space allocation, Cairns [[6], p. 43] states that “the more space

allocated to an item, the more likely it is to be seen by a shopper and, hence, the more likely it is to be purchased.” However, the shelf space elasticity of demand can vary across products. Brown and Tucker [7] identify the following three classes of products with varying shelf space elasticity of demand:

1. Unresponsive products, for which changes in shelf space allocation have no impact on sales rate (e.g., salt and spices). These products are also generally price inelastic.
2. General use products, for which increasing shelf space leads to increase in sales but at a diminishing rate (e.g., breakfast foods and canned fruit and vegetables).
3. Occasional purchase products, for which shelf space has a step-function or threshold effect on sales. Sales increase slowly with shelf space at first, until a large display causes a steep increase in sales to a point of diminishing returns. These include impulse buys (e.g., candy).

Following Brown and Tucker [7], researchers have sought to measure shelf space elasticity for various types of products. Both retailers and manufacturers are interested in this measurement, but for different reasons. Retailers care about total profits across all products. They benefit because they can allocate different amounts of shelf space to different products depending on their space elasticities and gross margins. Manufacturers care about the profit from their own products. They benefit because they can build shelf space allocation into their merchandizing discussions with retailers. For example, it has been found that private-label or store-brand products have higher shelf space elasticity than competing national brand products [8]. Thus, an easy way for a retailer to shift sales from national to store brands is to give more space to store brands on its shelf display.

Among early works in this area, Kotzan and Evanson [9] have conducted an experiment at a drug store chain to evaluate the impact of changes in shelf space on

unit sales. Their experiment was conducted on four products that met certain criteria related to demand uncertainty and availability of inventory. These products were tested in three stores for three weeks, where the number of facings allocated to each product was changed each week from 1, 2, or 3 facings. The authors discovered that three of the four products, a family-size Crest toothpaste, Hook’s Red Mouth Wash, and Johnson and Johnson Assorted Band Aids, had statistically significant positive shelf space elasticities. For example, the sales of Crest toothpaste with 1, 2, and 3 facings were 219, 291, and 294 tubes, respectively. The results for the fourth product (Preparation H Suppositories) were inconclusive.

Kotzan and Evanson did not investigate why shelf space elasticity varied across these products. However, the academic literature suggests many reasons for such variation. For instance, Curhan [8] relates shelf space elasticity to product characteristics. He hypothesizes that items with smaller package sizes, lower prices, smaller market shares, private-label brands, and higher sales rates (i.e., fast moving items) would have higher shelf space elasticities. He also hypothesizes that greater product variety, more availability of substitutes, and lower repurchase frequency would lead to higher shelf space elasticities. He tests these hypotheses by conducting an experiment for grocery products in supermarket stores. In this experiment, shelf space is changed for 500 items and their unit sales are observed for 5 weeks before the change and 12 weeks after the change. Shelf space elasticity is measured as the ratio of percent change in unit sales to the percent change in shelf space. Curhan [8] obtains very low R^2 values. He concludes that the product characteristics do not satisfactorily explain the observed variation in shelf space elasticities because the impact of shelf space on unit sales is very small relative to the effects of other environmental variables, leading to a failure of the model. Nonetheless, the average space elasticity across all items in his dataset is 0.212, showing a positive effect of shelf space on sales. In another study, Frank and Massy [10] show that environmental

variables such as store size, number of shelf rows, and shelf levels have a substantial impact on sales. Cox [11] also conducts an in-store experiment and provides evidence that staple product brands (e.g., salt brands) and impulse product brands that have low consumer acceptance are unresponsive to changes in shelf space. On the contrary, sales of an impulse product brand that has high consumer acceptance (e.g., Coffeemate) increase in shelf space.

Various conclusions are drawn in the literature from this early research. One is that shelf space elasticity is difficult to measure because a retail store is a dynamic environment and it is almost impossible to control factors such as retail prices, advertising, and addition and deletion of products, which have direct effects on sales [12,13]. Another is that shelf space elasticity is not large enough in magnitude to be managerially relevant. Instead, shelf space allocation should emphasize operational considerations such as the labor cost of restocking shelves and avoidance of stockouts. On the positive side, it is well recognized that shelf space elasticity varies across products, and is more important for private-label products and impulse-purchase items.

Research has also addressed how shelf space elasticity can be used by a retailer to increase sales and profits. For instance, Anderson [14] models the relationship between a product's market share and its share of shelf space using a logistic regression in order to find the profit-maximizing shelf space allocation. Dreze *et al.* [15] conduct experiments comparing two types of shelf management at a supermarket chain. In the first of these experiments, they change the shelf space allocation for each product to be proportional to the historical sales rate of the product in similar stores. Thus, in this method, they customize shelf space allocation in each store according to its historical sales. They note that this contrasts with the existing practice, which allocates shelf space in all stores of the chain identically regardless of differences in their sales mix. In these authors' second experiment, they reorganize the planogram for the store to facilitate cross-category merchandizing by

placing complementary product categories closer to each other. The experiment shows a 4% increase in sales and profits due to customized shelf space and 5–6% increase due to planogram reorganization. They use the results of their experiment in an optimization model and estimate that there is a potential for 15% increase in sales by optimizing the shelf space allocation to each item using the estimated parameters. They conjecture that the increase in sales is driven by customers increasing their share of purchases at the subject supermarket store when they are presented with a better shelf space allocation.

Thus, we see that experiments have been widely used to study shelf space allocation. Next, we turn our attention to using the results of such experiments in optimization models.

Optimization Models

We explain the optimization of shelf space using a model given by Corstjens and Doyle [13]. This model has been widely used and improved upon since 1981. We summarize some of the later developments after presenting the model.

Consider a retailer with total available shelf space S^* . The retailer seeks to allocate this space among K products in order to maximize its total profit. Let s_i be the shelf space allocated to product i , β_i be the direct elasticity of sales of product i with respect to its shelf space s_i , and δ_{ij} be the cross-space elasticity of the sales of product i with respect to the shelf space s_j allocated to product j . δ_{ij} can be positive or negative, and need not be equal to δ_{ji} . Then, the total sales for product i are written as

$$q_i = \alpha_i s_i^{\beta_i} \prod_{j=1, \dots, K, j \neq i} s_j^{\delta_{ij}}. \quad (1)$$

The gross margin from product i is written as $w_i q_i$, where w_i is the percent gross margin, and the variable store expense for product i given the sales quantity is written as $C_i = \gamma_i q_i^{\tau_i}$, where τ_i is the operating cost elasticity associated with the sales of product i . The retailer seeks to maximize its total profit, which is equal to the difference between the total gross margin and the total variable store

expense. This problem is formulated as the following constrained nonlinear program:

$$\begin{aligned} \max \sum_{i=1}^K w_i & \left[\alpha_i s_i^{\beta_i} \prod_{j=1, \dots, K, j \neq i} s_j^{\delta_{ij}} \right] \\ - \sum_{i=1}^K \gamma_i & \left[\alpha_i s_i^{\beta_i \tau_i} \prod_{j=1, \dots, K, j \neq i} s_j^{\delta_{ij} \tau_i} \right] \quad (2) \end{aligned}$$

subject to

$$\sum_{i=1}^K s_i \leq S^*, \quad (3)$$

$$\alpha_i s_i^{\beta_i} \prod_{j=1, \dots, K, j \neq i} s_j^{\delta_{ij}} \leq Q^* \quad i = 1, \dots, K, \quad (4)$$

$$s_i^L \leq s_i \leq s_i^U \quad i = 1, \dots, K, \quad (5)$$

$$s_i \geq 0 \quad i = 1, \dots, K. \quad (6)$$

Here, the first constraint represents the upper limit on available shelf space in the retail store. The second constraint represents an upper limit on the amount of sales that can be achieved for each product in the store. The third constraint restricts the amount of shelf space that can be allocated to each item to lie between two control limits.

This problem is extremely difficult to solve because the objective function and one of the constraints are nonlinear. Corstjens and Doyle [13] use a geometric programming method to solve this problem. They illustrate their model using data from a retail chain selling various types of candy, ice cream, and greeting cards. First they obtain data on sales and facings for each product across 140 stores. They fit equation (1) to each product category to estimate the direct and cross-shelf space elasticities. They also obtain cost data from the management to estimate gross margin and variable store expenses. They then optimize shelf space using these estimated parameters. The results lead to substantial changes in shelf space allocation from the existing allocations. For example, the results show that the optimal allocations for large and small stores differ from each other due to variation in sales mix. Thus, the model incorporates direct and cross-space elasticities, profit margins, and operating

costs to improve shelf space allocation. Corstjens and Doyle [16] later introduced a dynamic version of this model. It is a multiperiod model that takes into account anticipated changes in customer preferences and sales growth and decline of products. This model presents a long-term strategic view that encourages retailers to sacrifice short-term profits in order to maximize profits in the long run by taking shelf space away from declining products and allocating it to products with high growth potential. The optimization model of Corstjens and Doyle [13] has been further improved upon in subsequent research using many techniques, such as marginal analysis [17], dynamic programming [18], generalized Lagrange multiplier methods [19], and multistage optimization [20].

Most of the mathematical shelf space allocation models are nonlinear, mixed-integer problems that are computationally expensive for moderate-sized cases; even linear versions of this problem are NP-hard [2]. These difficulties have generated interest in heuristic solution approaches for these problems. For example, Urban [21] proposes a greedy heuristic to solve a nonlinear mixed-integer shelf space allocation problem. His algorithm starts with an initial solution including all items in the assortment plan. It then iteratively removes one item at a time from the assortment based on the greatest improvement in net profit estimated using a generalized reduced gradient. The algorithm stops when net profits cannot be improved any further by removing another item. Other techniques such as genetic algorithms [21–23], knapsack heuristics [24], local search methods [2], goal programming [25], simulated annealing [26], and greedy heuristics [27] are also used to solve shelf space allocation problems. In short, methods that are used to solve nonlinear programs can be tailored to solve shelf space allocation problems.

Another important consideration in shelf space allocation is its effect on inventory costs. Two items differing in shelf facings and/or service level requirements will have different inventory costs. For instance, an item with less shelf space will have more frequent stockouts and replenishments, which

will lead to higher stockout and labor costs, *ceteris paribus*. The shelf space allocation models that we have discussed thus far assume that the shelves are always fully stocked. Thus, they do not address the role of inventory costs on shelf space allocation. Freund and Matsuo [28] study this aspect by modeling inventory replenishment as a periodic order-up-to policy and explicitly defining holding costs, review costs, and stockout risk. They show that the importance of considering inventory costs increases as the desired service level increases. Furthermore, they conclude that higher service level requirements force the retailer to have a smaller assortment because the net profit suffers due to high operating costs when there is a wide product variety.

In many retail stores, inventory is split into “display inventory” and “backroom inventory.” Urban [21] distinguishes between the backroom and displayed inventories by tracking them separately. In his model, demand is a function of displayed inventory, whereas the backroom inventory allows the retailer to achieve economies of scale by ordering more than the shelf capacity and storing the excess units in the backroom. Thus, a backroom allows the retailer to stock more SKUs. Maiti and Maiti [23] adopt a similar framework. They provide a contractive mapping genetic algorithm to simultaneously solve the inventory management and shelf space allocation problems. See Urban [29] for an overview of the interdependencies between the inventory and shelf space allocation decisions.

Hwang *et al.* [22] analyze a model in which the demand rate is a function of the display location and quantity displayed. Their model determines shelf space allocation, order quantities, and the location where each brand should be displayed. Recent research by K  k and Fisher [30] has focused on integrating not only shelf space allocation and inventory decisions but also product assortment decisions in a single model. They determine what products should be selected to be put in the assortment and how much shelf space should be allocated to each item given inventory costs, product substitution, and shelf space constraints. They implement

their results at a supermarket chain. It is noteworthy that earlier researchers have also sought to integrate shelf space allocation and assortment decisions into one model. For example, Anderson and Amato [31] determine product assortment and shelf space allocation decisions jointly, but without considering inventory costs.

Commercial Software

Early commercial software such as PRO-GALI, OBM, CIFRINO, SLIM, COSMOS, and HOPE are based on simple rules of thumb such as allocating more space to products with the highest average sales or profit margins. None of these tools explicitly considers elasticities [16–18]. The next-generation commercial solutions are hybrid knowledge-based systems that can integrate human expertise and algorithmic techniques. One of the earliest hybrid knowledge-based decision support systems (DSS) used for shelf space allocation is Resource-opt [32]. It takes past sales, market research information, and managerial intuition as user input and extends Corstjens and Doyle [13] to provide store managers with a user-friendly DSS. It has been utilized to redesign a hypermarket in France, and perform shelf space allocation for three departments of a Scandinavian store as well as an oil company’s 2000 store franchise in Europe.

Presently it is common practice to use software assistance to generate visual diagrams (planograms) that show where every product in a retail store should be placed. For instance, PC-based systems such as Appollo (IRI) and Spaceman (Nielsen) are widely used at a strategic level. However, their operational level use is still very restricted due to their limited decision support functionality [20]. Generic optimization software packages such as LINGO (LINDO Systems) can be used at a tactical and/or operational level once the shelf space allocation problem is modeled as a nonlinear program [33].

Discrepancies occurring due to the use of planograms pose another problem as store managers have a tendency to deviate from recommended allocations. Most large retailer chains design planograms at the corporate level and give local store managers

some degree of freedom to modify the recommended design. Incentives for store managers are frequently tied to stockouts on the shelf and discipline in adhering to the shelf space allocation decided by the corporate office. However, managers also need to consider other issues such as the sales effect of fully stocked shelves and holding labor and replenishment costs (most of which are not captured by high level planograms). Furthermore, they need to react to local campaigns run by competitors. van Woensel *et al.* [34] analyze the possible causes for discrepancies between the recommended planograms and actual allocations and discuss the negative impact of these deviations on marketing efforts and operational efficiency.

Moreover, since shelf space is a scarce resource, it is natural for competing suppliers to try to influence the retailers' allocation decisions via negotiations and contract terms in order to obtain more shelf space. Martin-Herran *et al.* [35] model the interaction between two suppliers and a retailer as a Stackelberg game where two suppliers are leading and competing for the follower's (i.e., the retailer's) shelf space. They focus on the effect of the suppliers' advertisement strategies and show that the Stackelberg open-loop equilibrium is time-consistent.

In conclusion, researchers have been addressing various aspects of the shelf space allocation problem for more than 40 years. However, other issues remain to be addressed, such as the impact on shelf space management of the introduction of new tools and techniques for inventory management (e.g., RFID tags, contracting, and vendor managed inventory systems), poor processes for replenishment from backroom to shelf, inventory data inaccuracy, and misplaced SKUs [36,37].

MODELS WITH INVENTORY DEPENDENT DEMAND

Problem Description

In classical inventory models, we minimize inventory-related costs under the assumption that the demand is exogenous. However, in retailing, where the inventory

is visible to the end consumer, the rate of demand is often increasing in the amount of inventory stocked. As mentioned in the previous section, displayed stocks can stimulate demand [13]. Similarly, Wolfe and Little [38] provide evidence that the sales of style merchandise goods such as women's dresses are proportional to the amount of inventory displayed. In fact, the concept of psychic stock, defined as retail display inventory for stimulating demand, is motivated by the prevalence of this effect [39]. The demand-stimulating effect of inventories has motivated research to determine optimal inventory levels by solving the trade-off arising from the benefits and costs of holding more inventories. Early studies in the literature focus on developing inventory models for items with inventory-level-dependent rates. Recent literature has started to address strategic and tactical issues.

Economic Order Quantity Type Models and their Extensions

Economic order quantity (EOQ) type models are used to determine the optimal inventory policy when the demand is deterministic and influenced by inventory. We follow the research paper by Balakrishnan *et al.* [40] to introduce such models. Let the demand rate vary continuously as a function of the current inventory. If I denotes current inventory, then the demand rate is written as a nondecreasing concave function of I , denoted as $\lambda(I)$. For example, $\lambda(I) = \alpha(I + \phi)^\beta$, where $\alpha > 0$ is a scaling constant, $0 < \beta \leq 1$ is called the inventory elasticity of demand, and ϕ is a nonnegative base demand factor.

The mechanics of the model are as follows. Let $\hat{I}$ denote the inventory at the beginning of a replenishment cycle and T denote the length of the replenishment cycle. Inventory depletes at the rate of demand during the cycle. The inventory level at any time t during the cycle is obtained by integrating the inventory balance equation, $dI(t)/dt = -\lambda(I(t))$. For example, if $\lambda(I) = \alpha I^\beta$, then the inventory at time t is equal to $I(t) = (\hat{I}^{1-\beta} - \alpha t)^{1/(1-\beta)}$. Let $D(\hat{I}, T)$ denote the total demand during the replenishment cycle. It is equal to the difference between $\hat{I}$ and $I(T)$. The length of the

replenishment cycle must be smaller than the time taken to deplete all the inventory. Let $T_{\text{runout}}(\hat{I})$ denote the time at which inventory will run out if not replenished.

The firm seeks to determine the order-up-to level $\hat{I}$ and the replenishment interval T in order to maximize its average profit per unit time. Let h denote the inventory holding cost rate, S the ordering cost, and r the per unit profit contribution. The profit per unit time is given by

$$\pi(\hat{I}, T) = \frac{1}{T} \{rD(\hat{I}, T) - S\} - h\bar{I}(\hat{I}, T), \quad (7)$$

where $\bar{I}(\hat{I}, T)$ is the average inventory per unit time and can be written as

$$\bar{I}(\hat{I}, T) = \frac{1}{T} \int_0^T I(t)dt = \hat{I} - \frac{1}{T} \int_0^T D(\hat{I}, t)dt.$$

The firm's demand-stimulating inventory problem, thus, involves maximizing (7) subject to the constraint that $T \leq T_{\text{runout}}(\hat{I})$. This problem differs from the EOQ problem because it entails profit maximization rather than cost minimization and it allows replenishment to take place when the inventory is not fully depleted. Indeed, Balakrishnan *et al.* [40] distinguish two types of products depending on the optimal policy: early-replenishment products are those for which $T < T_{\text{runout}}(\hat{I})$ and runout-replenishment products are those for which the constraint is binding.

We illustrate the optimal solution obtained by Balakrishnan *et al.* [40] for a specific demand function which they call the reference demand model. In this model, the inventory elasticity of demand is equal to 0.5 and the demand rate is specified as $\lambda(I) = \alpha\sqrt{I} + \phi$. This yields the cumulative demand function $D(\hat{I}, t) = \alpha t\sqrt{\hat{I} + \phi} - (\alpha t/2)^2$ and $T_{\text{runout}}(\hat{I}) = (2/\alpha)(\sqrt{\hat{I} + \phi} - \sqrt{\phi})$. Substituting these expressions into the profit function and solving for the optimal $\hat{I}$ and T , we find that the early-replenishment strategy gives the optimal solution if the ordering cost S is below a certain threshold and the runout-replenishment strategy is optimal otherwise. The optimal replenishment cycle

and order-up-to level for early-replenishment products are obtained as

$$T^* = 2 \left(\frac{3S}{\alpha^2 h} \right)^{1/3}$$

and

$$\hat{I}^* = \frac{\alpha^2}{4} \left(\frac{r}{h} + \left(\frac{3S}{\alpha^2 h} \right)^{1/3} \right)^2 - \phi.$$

The optimal order-up-to level for runout-replenishment products is obtained similarly.

Many insights can be obtained from this solution. As in the classical EOQ model, in this model T^* depends on the ratio S/h . Moreover, T^* does not depend on r . Thus, as in the EOQ model, the length of the replenishment cycle is determined by the trade-off between the ordering cost and the holding cost. The optimal order-up-to level, however, depends on r/h . Hence, it is determined by the trade-off between the contribution and the holding cost. When the contribution increases, the replenishment cycle remains unchanged, but the firm carries more inventory and places its orders at higher reorder points. Balakrishnan *et al.* [40] also construct a heuristic and refer to it as the adaptive EOQ policy. In this heuristic, the firm uses the EOQ formula to determine the order quantity in each cycle, but recalibrates the demand rate parameter λ using the observed average demand rate in the previous cycle. This heuristic differs from the optimal policy because it does not allow orders to take place before the inventory runs out. However, it enables the firm to learn about the dependence of demand on inventory from historical data. Balakrishnan *et al.* [40] show that this heuristic converges to an equilibrium order quantity. However, it performs poorly because (i) it waits for the inventory to run out before placing a reorder and (ii) in each cycle, it orders too little and too frequently. For example, their analysis shows that, if $\lambda(I) = \alpha I^{0.5}$ and the product is an early-replenishment product, then the profit from the adaptive EOQ policy is always less than 40% of the optimal profit.

The demand function $\lambda(I) = \alpha I^\beta$ is appealing because it is similar to functions used to model the dependence of demand on shelf

space [13]. This functional form makes the parameter estimation relatively easy because we can do a logarithmic transformation and fit a linear regression model to estimate α and β . It is possible to use many similar functional forms in a model to capture the demand-stimulating effect of inventory. Gupta and Vrat [41] assume that the demand rate depends only on the initial inventory, and remains constant at this rate. Baker and Urban [42] and Urban [43] allow demand to vary continuously with inventory and show that it may not be optimal to wait to place an order until the inventory level reaches zero. Urban [44] gives a recent survey of the literature on this topic.

Modeling the inventory dependent demand for perishable items require an additional structure since demand for a perishable item decreases not only with inventory but also due to loss of product freshness and/or an approaching expiration date. Such models have been studied in the literature [45–48]. For example, Balkhi and Benkherouf [48] model demand as $\lambda(I, t) = I^\beta G(t)$, where I is the current inventory, t denotes time, and $G(t)$ is an increasing function of time, and the inventory is allowed to deteriorate continuously over time at a fixed rate θ . Instead of deterioration of inventory, another feature to consider in certain situations is that the value of the unsold inventory may decrease over time [49]. This happens for seasonal or fashion products.

Strategic and Tactical Implications

Strategic consumers indirectly observe retailers' inventory policies over time. If a consumer's observations lead her to believe that she has a low probability of finding a desired product at a given retailer, she might choose to shop at a competitor instead. For instance, Anderson *et al.* [50] present empirical evidence showing that consumers who experience a stockout are less likely to place an order; these consumers also order fewer items, and spend less. These authors also show that the retailer might be able mitigate the cost of a stockout if an item is out of stock due to high popularity as consumers are more willing to backorder

scarce items. On the contrary, consumers may be willing to pay a premium price for consistently high service rates. For instance, Dana [51] describes a small experiment on video stores and argues that Blockbuster's advertised claims of high availability allow the retailer to charge higher prices. These empirical results show that the retailer's stocking decisions and ability to manage stockouts shape its reputation which affects its future demand and pricing power.

Dana and Petruzzi [52] present a model in which consumers form beliefs about the service level offered by a retailer. They then decide whether to visit the retailer based on these beliefs. This model describes a single period, which can be interpreted as a steady-state representation of a repeated game between the consumers and the retailer. Their analysis shows that, when the retailer recognizes the effect of its service level on demand, it stocks higher inventory levels. As a consequence, the retailer attracts more customers and earns a higher expected profit. Similarly, many models have been developed in the recent research describing consumers who behave strategically by deciding when to visit a given retailer. Their strategic behavior can be affected by product availability, expectation of markdowns, and so on. We do not describe these models in this article, since they form a large body of literature by themselves.

Overlooking the demand-stimulating effect of inventory also creates incentive misalignment issues in retail operations, as most of the automated inventory replenishment systems try to minimize inventory-related costs whereas store managers are assessed on revenues. van Donselaar *et al.* [53] empirically show that retail stores managers have a tendency to deviate from order advices generated by an automated inventory replenishment system. Their analysis illustrates that the store managers improve the automated replenishment system's performance by systematically modifying the order advices because they are able to capture the demand-stimulating effect of inventories which is ignored by the system. Similarly, Kesavan *et al.* [54] estimate the effect of gross margin and inventory on

each other and empirically show that more accurate firm level sales forecasts can be achieved by incorporating its relationship with inventory and gross margin.

In conclusion, capturing the demand-stimulating effect of inventory is necessary for effective inventory management. However, it is not sufficient to achieve firm level success. Identifying the relationship between demand and inventory should be seen as a part of a broader action plan called demand-based (supply chain) management [55]. This approach seeks to maximize the total value (i.e., profits and/or other well-defined objectives of the entire supply chain) by addressing the interdependencies between inventory, demand, pricing, and marketing. Overlooking this approach and tackling the demand inventory relationship independently of pricing and marketing leads to suboptimal solutions. Put differently, pricing, promotional marketing, and/or other demand manipulation tools should be used to mitigate the negative consequences of the relationship between demand and inventory.

MODELS OF RETAIL COMPETITION

The actions of one retailer directly affect not only its own demand but also the demand faced by its competitors. Promotions, price discounts, and increases in service level can increase its demand due to customers switching their purchases from competitors to its stores. On the contrary, core service failures (e.g., stockouts, billing mistakes, and service catastrophes), service encounter failures (e.g., impolite and unknowledgeable employees), and pricing failures (e.g., high prices and deceptive pricing strategies) can lead to customer losses [56]. These changes in demand can be temporary or permanent. Therefore, a retailer should determine its operational policies taking into account their competitive implications.

Retail competition can be modeled in many ways of varying complexity depending on the extent of competitive interaction among firms in the marketplace. For example, we may consider

1. *single-period models*, in which consumers switch from one retailer to another in the immediate period upon experiencing a stockout;
2. *multiperiod models*, in which consumers switch from one retailer to another in a future period upon experiencing a stockout;
3. *models of full or partial information*, in which consumers have some prior knowledge of inventories at competing retailers and choose whether to visit a retailer upon evaluating their chances of finding a product in stock;
4. *learning models*, in which consumers do not have knowledge of product availability but learn about their probability of finding a product in stock based on their own historical experience with a retailer, and then modify their future shopping behavior accordingly;
5. *multidimensional models*, in which consumers select the retailer to visit based on many service dimensions, such as price, promotions, and service level.

At the simplest level (in (1) above), consider a model studied by Lippman and McCardle [57]. In this model, there are two retailers competing with each other in a single period. The price and cost parameters are given and are equal across retailers. The retailers compete only on the basis of their inventory levels y_i . The total demand is a random variable denoted as D . It is allocated between the retailers via some splitting rule; let D_i denote the initial allocation of demand to firm i . If the demand allocated to firm j is greater than its inventory, that is, if $D_j > y_j$, then a fraction a_i of the excess demand is reallocated to firm i . Thus, the effective demand at firm i , including its initial demand and reallocation, is given by $R_i = D_i + a_i \max\{0, D_j - y_j\}$, where $0 \leq a_i \leq 1$. Each retailer seeks to determine the inventory level that maximizes its expected profit. Note that each retailer's choice of inventory affects the demand faced by its competitor because of substitution or reallocation of excess demand.

Lippman and McCardle [57] describe many ways by which the initial demand for

each retailer may be obtained. For example, a deterministic splitting is one in which total demand is allocated between the retailers in a fixed proportion. A random splitting can be obtained in many ways: if each customer flips a coin to choose a retailer, then it is called an incremental random splitting of demand; if the first customer flips a coin to choose a retailer and each subsequent customer follows the previous customer, then it leads to a simple random splitting or herd behavior, that is, all of the demand is allocated to one or the other retailer with probability 0.5; finally, the initial demand at the two retailers may be given by independent random variables. Lippman and McCardle [57] present a beautiful example of a tourist bus visiting a chateau to illustrate the demand models. They show that there exists a pure strategy Nash equilibrium in inventory levels in this competitive game. The equilibrium need not be unique. Moreover, competition can be detrimental to the firms because it can drive down the industry profit to zero and drive up the total inventory level in the industry to be higher than the monopolist's optimal inventory, that is, the optimal inventory quantity if there were a single firm serving the entire demand.

Whereas single-period models allow us to study how competition affects the demand faced by a retailer in the current time period, multiperiod models (in (2) above) are useful to study the effect on demand in subsequent time periods. A multiperiod model of competition can allow us to measure the future goodwill cost of losing a customer if the service level in the current period is reduced. Following Hall and Porteus [58], suppose that there are two firms competing in a marketplace over a finite time horizon of T periods. Let D_{it} be the demand and y_{it} be the inventory level for firm i in period t . The unsatisfied demand is then given by $\max\{0, D_{it} - y_{it}\}$, and represents the number of customers who experience a stockout. Suppose that a fraction γ_i of these customers switch to the competitor in the next time period. Thus, if a firm were to increase its inventory, it would lose fewer customers to its competitor. The objective of the firm in the model is to determine its inventory

levels in order to maximize the total expected profits over the time horizon. With some additional simplifying assumptions, Hall and Porteus [58] show that this problem yields a unique subgame perfect Nash equilibrium. Their solution gives an imputed lost goodwill cost for each firm, which is a function of the present value of an additional customer in the next time period and the probability of losing that customer due to a stockout. The retailer with less sensitive customers, that is, a smaller value of γ_i , faces a lower imputed goodwill cost. Such a retailer will provide a lower service level but still enjoy less defections and a larger market share.

A different way to model competition due to inventory is to define fill rate or service level as a dimension of quality. Fill rate is defined as the fraction of demand satisfied by a retailer from stock. It is equal to the ratio of expected sales to mean demand; if y is the amount of inventory stocked by a retailer and D is the random demand faced by it with mean $E[D] = \mu$, then fill rate is equal to $E[\min\{D, y\}]/\mu$. The numerator in this expression is the expected sales of the firm, which is given by the minimum of demand and inventory. When the fill rate is high, fewer customers experience stockout, and therefore fewer customers are likely to be dissatisfied with the retailer. Therefore, the demand faced by a retailer becomes a function of its fill rate as well as the fill rates of its competitors. This concept is usually expressed in models of type (3)–(5) listed above by defining the market share of a retailer as

$$\frac{f_i}{\sum_j f_j},$$

where f_i is the service level offered by the subject retailer. In models of full information, customers know the inventory level stocked by each retailer and choose which firm to visit by weighing their probability of finding the product in stock [51]. In models of partial information, customers do not know the inventory levels of the retailers, and use other cues such as price to form expectations about the fill rates provided by retailers [51, 59]. Learning models are multiperiod models in which customers learn about the fill rates

from their past shopping experience at each firm [60,61].

Consumers' reaction to price and service quality variations might be asymmetric. That is, consumers might weigh negative experiences (losses) more than equivalent positive experiences (gains). For instance, Hardie *et al.* [62] argue that loss aversion and the position of brands relative to multiattribute reference points (e.g., price and quality) influence the brand choice. They empirically support this claim by estimating and comparing gain and loss coefficients for price, quality, loyalty, and the presence of advertising for orange juice sales using scanner data. These analyses can be utilized to model the effect of asymmetric consumer behavior on the new product introduction and price promotions. Similar asymmetry also arises in services based on satisfying and unsatisfying service experiences. For instance, consumers react more strongly to a stockout of a necessity items such as bread because they always expect to see such items in stock (negative bias). On the other hand, they might have positive bias for prestige products such as designer suits because they expect to search for these items. Gaur and Park [61] show that the effect of competition on total inventory levels and total industry profits depends on the type of bias exhibited by consumers. Moreover, when retailers have different costs, the difference in market shares of the retailers also depends on the type of bias. The lower cost retailer enjoys greater market share and profit differential when consumers have a negative bias, whereas a positive bias tends to attenuate the effect of competition for the higher cost retailer.

Competition does not necessarily occur in a single dimension. For instance, two retailers selling an identical product might compete on selling price and service. Service can be measured with a single proxy such as the fill rate [63] or it can be a performance measure which aggregates many aspects of the shopping experience such as promotions, advertising, and customer relations into a single decision variable [64]. In both cases, the product can be treated as a bundle of two attributes, price and service. Tsay and Agrawal [64] consider a single period setting

in which two competing retailers obtain a product from a common manufacturer, and discover that there are cases under which both retailers prefer an increase in competitive intensity because adding a small amount competition in one dimension mitigates the competitive intensity in the other. Bernstein and Federgruen [63] characterize an infinite-horizon, stochastic general equilibrium model for competing firms under different competition scenarios and demand processes.

Competition can affect firms in many other ways. A retailer might be willing to offer a monetary incentive to a customer in order to convince her to backorder instead of switching to another retailer [65]. The option to backorder reduces competition because unsatisfied demand is not necessarily lost. Another aspect of retail competition is the speed of delivery. If firms compete on the delivery time, then holding inventories is utilized as a tactic to reduce the customer waiting time and increase sales at the expense of high inventory holding costs. In fact, firms are more likely to switch from a make-to-order policy to a make-to-stock policy when the number of competitors increases. This type of competition increases consumer welfare and reduces retailers' profits [66]. Factors such as the firm and consumer characteristics, service quality, and searching costs also affect retail competition. For instance, McGahan and Ghemawat [67] study the relationship between firm sizes and competition to retain customers in a two stage game. In the first stage, firms try to build up loyalty among existing customers. In the second stage, they compete on price. Their analysis shows that large firms are likely to exhibit greater customer retention rates than their small rivals in equilibrium. Lastly, if customer search costs are low in an oligopolistic price competition setting, profit-maximizing firms may choose to have occasional stockouts to reduce competition. Reduced competition allows the firms to charge higher prices which might offset the effect of lost sales due to stockouts [68].

In conclusion, retailer competition has various aspects including, but not limited to, quantity, price, and service quality. Taking competition into account helps retailers

create more accurate models that can capture market dynamics and consumer choice, which will lead to more effective pricing and stocking decisions.

SUBSTITUTION AND TRANSSHIPMENT MODELS

Adoption of modern information technology tools such as ERP (enterprise resource planning) systems and web-based inventory tracking applications has led to remarkable improvements in retail supply chain transparency. Nowadays, it is possible to obtain real-time information regarding on-hand and in-transit inventory quantities for each location in a retailer's supply chain. In some systems, it is even possible to track lost sales due to shortages so that more accurate demand forecasts can be made. Reduced information and transportation costs and shorter transportation lead times have enabled companies to move items not only from an upper installation (e.g., a warehouse) to a lower installation (e.g., a store) but also between any two lateral points in the system (e.g., from one store to another) [69].

These technologic advancements have allowed companies to utilize two risk pooling techniques more effectively: product substitution and transshipment. Substitution redistributes demand from a stocked-out product to another product with excess inventory. Lateral transshipment redistributes inventory from stores with excess on-hand inventory to stores facing shortages or low inventory levels [70]. These techniques are complementary because substitution can be utilized when the consumer is willing to purchase a similar item instead of waiting for her favorite item to be restocked, whereas lateral transshipment is more appropriate when the consumer is willing to delay her purchase. For instance, a consumer might be willing to purchase a 13.5-oz shampoo of the same brand when she could not find the 25.4-oz size. However, she might be willing to wait for designer shoes if her size is temporarily out of stock. The mathematical models to compute solutions that provide substitution

and transshipment capabilities are similar. Below, we first summarize the effect of substitution on inventory management. Then, we describe transshipment policies. The topic of assortment planning is similar to substitution models and is discussed in a separate section. Typically, substitution models have a fixed number of available products, whereas the number of products to stock is a decision variable in assortment planning models.

Substitution

The implications of substitution on retail profits and inventory levels are important to study because consumers are often willing to purchase substitute items when they face stockouts. According to a survey conducted by Food Marketing Institute, more than 80% of the survey participants would be willing to buy a substitute item if their favorite item were not available [71]. Although demand substitution mitigates the effect of lost sales by switching demand from one item to another, it complicates inventory management because the sales of each item now depend not only on its own inventory and demand but also on the inventory and demand of all other items. Therefore, inventory policies developed without taking the substitution effect into account can lead to large profit losses.

There are two types of substitution phenomena, retailer driven and consumer driven. Under the first scenario, the retailer satisfies the demand for one product using another, possibly higher quality, product to mitigate stockouts. For instance, a downward substitution takes place when a high quality item can be downgraded and used as a substitute for a low quality item but not vice versa. Downward substitution can be understood using the model in Bassok *et al.* [71] as follows. Suppose that there are N products and N demand classes. The demand from class i can be satisfied using inventory of any product j such that $j \leq i$, thus representing downward substitution. The retailer earns revenue p_i for meeting demand of type i , incurs backorder cost π_i for shortages in class i , and incurs a cost of substitution b if demand for class i

is satisfied using inventory of some other product $j < i$. Product i can be purchased at cost c_i and its excess inventory can be salvaged at s_i . The retailer's profit maximization problem can be decomposed into two parts. In the first stage, the retailer determines the amount of inventory y_i of each product to stock. Then, a random amount of demand of each class is received. Given this demand, the retailer determines how to allocate the available inventory of various products among the demand classes in order to maximize its profit. The second stage problem can be formulated as the following linear program:

$$\begin{aligned} & G(y_1, \dots, y_N, D_1, \dots, D_N) \\ &= \max_{u_i, v_i, w_{ji}} \sum_{i=1}^N \left[p_i w_{ii} + \sum_{j=1}^{i-1} (p_j - b) w_{ji} \right. \\ & \quad \left. + s_i v_i - \pi_i u_i \right], \end{aligned}$$

subject to

$$\begin{aligned} u_i + \sum_{j=1}^i w_{ji} &= d_i \quad i = 1, \dots, N \\ v_i + \sum_{j=1}^N w_{ij} &= y_i \quad i = 1, \dots, N \\ w_{ij} &\geq 0 \quad i, j = 1, \dots, N \\ u_i, v_i &\geq 0 \quad i = 1, \dots, N. \end{aligned}$$

The decision variables in this formulation are w_{ij} , which is the amount of product i used to satisfy demand class j ; u_i , the amount of shortage in demand class i ; and v_i , the excess inventory of product i . The objective function consists of the revenue from meeting demand, the salvage value of excess inventory, and the cost of shortages. The constraints implement the balance of flow between supply and demand within the downward substitution structure. Given the solution of the second stage problem for any inventory level and demand realization, the first stage consists of determining the inventories to maximize

total expected profit, that is,

$$\max_{y_1, \dots, y_N} E[G(y_1, \dots, y_N, D_1, \dots, D_N)] - \sum_{i=1}^N c_i y_i.$$

Bassok *et al.* [72] derive a solution for this problem by showing that the profit function in the first stage problem is concave and submodular in inventory levels under mild assumptions on the price and cost parameters. Besides being suitable for many kinds of applications, this problem formulation has also been used to give an upper bound on the profit function of the substitution problem under consumer-driven substitution [73].

In consumer-driven substitution, the excess demand from one product is reallocated to other products according to some fixed substitution rule. We describe this model using Netessine and Rudi [74]. Consider the same notation as above. However, now, demand for class i must be satisfied first by available inventory of product i . Then, unsatisfied demand of class i is allocated to all the other products in a fixed proportion. Let a_{ij} denote the fraction of unsatisfied demand of class i allocated to product j , where $a_{ij} \in [0, 1]$ and $\sum_{j=1}^N a_{ij} \leq 1$. Therefore, the effective demand for product i is equal to

$$D_i + \sum_{j \neq i} a_{ji} (D_j - y_j)^+,$$

and the leftover inventory of product i is

$$\left(y_i - D_i - \sum_{j \neq i} a_{ji} (D_j - y_j)^+ \right)^+.$$

Here, $(D_j - y_j)^+$ denotes $\max\{0, D_j - y_j\}$. Thus, the retailer seeks to maximize its expected profit, which is a function of the total sales revenue, the salvage value of leftover inventory, and the purchasing cost of the inventory across all products. Note that the substitution is up to "first-level," that is, when excess demand of product i is reallocated to product j , then a stockout of product j does not lead to further reallocation of the leftover demand to other products. Even so, this problem is extremely difficult to solve. Netessine and Rudi [74] derive

first-order necessary optimality conditions for this problem, but show that there may be multiple local maxima. For simpler two- and three-product problems, the profit function is, however, concave [75,76].

Substantial savings in inventory-related costs can be achieved when items in the product portfolio are highly substitutable, that is, the probability of accepting another item is relatively high for a customer that cannot find her favorite item [77]. However, demand substitution does not necessarily lead to decrease in the total inventory level. For instance, for a two-product case, it has been shown that the total inventory level may increase when only one item is substitutable [78].

Transshipment

The purpose of transshipment is to redistribute inventory so that the right quantities are available in the right location [79]. Lateral transshipment can be divided into emergency lateral transshipment (ELT) and preventive lateral transshipment (PLT) [80]. ELT mandates emergency transfers from a retailer with excess stock to a retailer that has a stockout [81,82]. This policy responds to stockout incidents after the realization of demand. On the other hand, under a PLT policy, items are transferred among locations in order to balance inventories in anticipation of stockout [83]. PLT has a nonmyopic view which tries to reduce the risk of future stockouts by redistributing the inventory [70]. In both ELT and PLT policies, the initial inventory at each location is planned with the view to allowing transshipment in the future. In the most primitive case, if the retailer does not conduct any transshipment, the inventory decision for each location can be made independently. Thus, transshipment requires the inventories at all locations to be determined jointly, adding considerable complexity to the problem. The benefits are that it provides risk pooling across locations, and generally reduces the total inventory requirement.

Most of the early literature is concentrated on ELT policies for repairable items because they have low demand rates and high backorder costs and thus could benefit the most

from transshipment. The models used in this context are one-for-one continuous-review models similar to the seminal METRIC model of Sherbrooke [84]. For instance, Lee [81] studies a one-for-one multiechelon model for repairable items which allows lateral transshipments between identical retailers. He derives approximations for the commonly used performance measures such as the backorder levels and the number of lateral transshipments. These approximations are used to determine the optimal stocking levels. He shows that the use of lateral transshipment leads to large cost savings. Axsäter [82] extends these analyses under a scenario which the retailers are not identical. Archibald *et al.* [85] utilize Markov decision processes to study a two-location, multi-period, multi-item transshipment problem subject to capacity constraints. They show that the order-up-to policy is optimal when the demand for each item arises according to independent Poisson processes at each location. Although it is, in theory, possible to represent an inventory system as a Markov model and derive the steady-state probabilities, this approach may not always be computationally feasible since the state space grows exponentially with system size. Dada [86] focuses on developing a fast procedure to approximate the steady-state expected performance of a two-echelon system with transshipment. His model provides tight bounds on the system performance.

Due to the complexity of the transshipment problem, it is useful to devise simple heuristics. For instance, when shortage costs differ among locations, a simple yet effective technique is to allow unidirectional transshipments, that is, transshipments from locations with lower shortage cost to locations with higher shortage cost, but not in the opposite direction [87]. Another heuristic takes reorder points and batch quantities as given and tries to fulfill the excess demand at a retailer by a lateral transshipment from the retailer with most stock on hand [88]. This heuristic is useful under complex decision situations because it does not jointly optimize inventories across locations.

ELT policies for a supply chain with multiple retailers with different cost structures and demand parameters have also been studied under a continuous-review one-for-one inventory policy [89] as well as for a periodic-review system with a base stock inventory policy [90].

A PLT policy is modeled by Das [83] for a two-location stochastic inventory problem under centralized decision making. He implements this policy by setting a predetermined time point within each period at which the decision maker can move items from the overstocked location to the understocked location. He shows that a base stock policy with a transfer to the understocked location to bring its inventory level closest to its base stock level without decreasing the inventory level of the overstocked location below its base stock level is optimal.

Lee *et al.* [80] study a periodic-review model and develop a transshipment policy which can be classified as a combination of ELT and PLT. Namely, they define predetermined and fixed upper, lower, and target service levels which are used to compute the lateral transshipment quantities. At the end of each review period, retailers with inventory levels exceeding the corresponding upper service level send their excess inventories to retailers with low inventory levels or stockouts. A retailer with a low inventory level determines the amount of inventory that it can receive using the difference between the low and target service levels. Thus, this policy performs inventory balancing as well as emergency transshipments.

Two extreme transshipment policies are to never transship and always transship when there is a shortage at one location and stock is available at another location. It has been observed that choosing the better of these two extreme policies leads to a performance which is almost as good as a complex policy that takes the future impact of a stock transfer into account [91].

When a supply chain is centralized, the objective of transshipment is to optimize the overall system performance. However, when locations within a supply chain belong to

different organizations, transfer prices affect each location's profitability and willingness to participate in a transshipment activity. In general, when each location tries to maximize its own profits, the resulting Nash equilibrium will not maximize joint profits. However, it is possible to set transshipment prices to create supply chain coordination such that the decisions made by each location are consistent with joint-profit maximization [92]. Furthermore, in decentralized supply chains, it might be necessary to offer some incentives to supply chain partners in order to prevent free-riding and implement effective inventory distribution and transshipment policies [69]. Dong and Rudi [93] study a transshipment model in which an external supplier sells to multiple retail stores owned and operated by the same firm. They show that stores' order quantities are less sensitive to the wholesale price under a transshipment policy due to risk pooling. Hence, the manufacturer benefits from transshipments at the expense of the retailers because it can charge a higher wholesale price. Zhao *et al.* [94] consider a decentralized system with a large number of independent retailers and prove that a threshold requesting and rationing policy is optimal. Under this policy, there exist thresholds $Z_i \leq K_i \leq S_i$ for each retailer i , denoting the optimal requesting, rationing, and base stock levels, respectively, for retailer i . It is optimal to send a transshipment request to another retailer only if the inventory level is below the requesting level Z_i , and to fill a received transshipment request only if the inventory level is above the rationing level K_i .

To sum up, lateral transshipment is a way to perform risk pooling and satisfy demand especially for low demand, high stockout cost items. However, one should take the transshipment and replenishment lead times; related ordering, holding, transportation, and stockout costs; structure of the supply chain (e.g., centralized vs decentralized); alternative risk pooling techniques (e.g., substitution); and the underlying demand distributions into account in order to develop an effective transshipment policy. Chiu [95]

presents an extensive survey of the academic literature in this area.

REFERENCES

1. Levy M, Weitz BA. Retailing management. 6th ed. New York: McGraw-Hill/Irwin; 2007.
2. Lim A, Rodrigues B, Zhang X. Meta-heuristics with local search techniques for retail shelf-space optimization. *Manage Sci* 2004;50(1):117–131.
3. Buttle F. Retail space allocation. *Int J Phys Distrib Mater Manage* 1984;14(4):3–23.
4. Quelch JA, Kenny D. Extend profits, not product lines. *Harvard Bus Rev* 1994; 72(5):153–160.
5. Corstjens J, Corstjens M. Store wars: the battle for mindshare and shelf space. Chichester, UK: Wiley; 1995.
6. Cairns JP. Allocate space for maximum profits. *J Retailing* 1963;39(2):41–45.
7. Brown W, Tucker WT. The marketing center: vanishing shelf space. *Atlanta Econ Rev* 1961;11(10):9–13.
8. Curhan RC. The relationship between shelf space and unit sales in supermarkets. *J Mark Res* 1972;9(4):406–412.
9. Kotzan JA, Evanson RV. Responsiveness of drug store sales to shelf space allocations. *J Mark Res* 1969;6(4):465–469.
10. Frank RE, Massy WF. Shelf position and space effects on sales. *J Mark Res* 1970;7(1):59–66.
11. Cox KK. The effect of shelf space upon sales of branded products. *J Mark Res* 1970; 7(1):55–58.
12. Curhan RC. Shelf space allocation and profit maximization in mass retailing. *J Mark* 1973; 37(2):54–60.
13. Corstjens M, Doyle P. A model for optimizing retail space allocations. *Manage Sci* 1981;27(7):822–833.
14. Anderson EE. An analysis of retail display space: theory and methods. *J Bus* 1979; 52(1):103–118.
15. Dreze X, Hoch SJ, Purk ME. Shelf management and space elasticity. *J Retailing* 1994;70(4):301–326.
16. Corstjens M, Doyle P. A dynamic model for strategically allocating retail space. *J Oper Res Soc* 1983;34(10):943–951.
17. Bultez A, Naert P. SHARP: shelf allocation for retailers' profit. *Mark Sci* 1988;7(3):211–231.
18. Zufryden FS. A dynamic programming approach for product selection and supermarket shelf-space allocation. *J Oper Res Soc* 1986;37(4):413–422.
19. Hansen P, Heinsbroek H. Product selection and space allocation in supermarkets. *Eur J Oper Res* 1979;3(6):474–484.
20. Yang MH, Chen WC. A study on shelf space allocation and management. *Int J Prod Econ* 1999;60:309–317.
21. Urban TL. An inventory-theoretic approach to product assortment and shelf-space allocation. *J Retailing* 1998;74(1):15–35.
22. Hwang H, Choi B, Lee MJ. A model for shelf space allocation and inventory control considering location and inventory level effects on demand. *Int J Prod Econ* 2005;97(2):185–195.
23. Maiti MK, Maiti M. Multi-item shelf-space allocation of breakable items via genetic algorithm. *J Appl Math Comput* 2006;20(1–2):327–343.
24. Yang MH. An efficient algorithm to allocate shelf space. *Eur J Oper Res* 2001; 131(1):107–118.
25. Reyes PM, Frazier GV. Goal programming model for grocery shelf space allocation. *Eur J Oper Res* 2007;181(2):634–644.
26. Borin N, Farris P, Freeland J. A model for determining retail product category assortment and shelf space allocation. *Decis Sci* 1994;25(3):359–384.
27. Bai R, Burke EK, Kendall G. Heuristic, meta-heuristic and hyper-heuristic approaches for fresh produce inventory control and shelf space allocation. *J Oper Res Soc* 2008;59(10):1387–1397.
28. Freund R, Matsuo H. Retail inventory and shelf space allocation. Working paper. Department of Management, University of Texas at Austin; 1997.
29. Urban TL. The interdependence of inventory management and retail shelf management. *Int J Phys Distrib Logist Manage* 2002;32(1):41–58.
30. K  k AG, Fisher ML. Demand estimation and assortment optimization under substitution: methodology and application. *Oper Res* 2007;55(6):1001–1021.
31. Anderson EE, Amato HN. A mathematical model for simultaneously determining the optimal brand collection and display-area allocation. *Oper Res* 1974;22(1):13–21.
32. Singh MG, Cook R, Corstjens M. A hybrid knowledge-based system for allocating retail

- space and for other allocation problems. *Interfaces* 1988;18(5):13–22.
33. Hariga MA, Al-Ahmari A, Mohamed AA. A joint optimisation model for inventory replenishment, product assortment, shelf space and display area allocation decisions. *Eur J Oper Res* 2007;181(1):239–251.
 34. van Woensel T, Broekmeulen R, van Donseelaar K, *et al.* Planogram integrity: a serious issue. Available at <http://home.tn.tue.nl/tvwoense/files/PlanogramIntegrity.pdf>. Accessed July 3, 2010.
 35. Martin-Herran G, Taboubi S, Zaccour G. A time-consistent open-loop Stackelberg equilibrium of shelf-space allocation. *Automatica* 2005;41(6):971–982.
 36. DeHoratius N, Raman A. Inventory record inaccuracy: an empirical analysis. *Manage Sci* 2008;54(4):627–641.
 37. Raman A, DeHoratius N, Ton Z. Execution: the missing link in retail operations. *Calif Manage Rev* 2001;43(3):136–152.
 38. Wolfe HB, Little AD. A model for control of style merchandise. *Ind Manage Rev* (now *Sloan Manage Rev*) 1968;9(2):69–82.
 39. Larson PD, DeMarais RA. Psychic stock: an independent variable category of inventory. *Int J Phys Distrib Logist Manage* 1999;29:495–507.
 40. Balakrishnan A, Pangburn MS, Stavroulakis E. Stack them high, let'em fly: lot-sizing policies when inventories stimulate demand. *Manage Sci* 2004;50(5):630–644.
 41. Gupta R, Vrat P. Inventory model for stock-dependent consumption rate. *Oper Res* 1986;23(1):19–24.
 42. Baker RC, Urban TL. A deterministic inventory system with an inventory-level-dependent demand rate. *J Oper Res Soc* 1988;39(9):823–831.
 43. Urban TL. An inventory model with an inventory-level-dependent demand rate and relaxed terminal conditions. *J Oper Res Soc* 1992;43(7):721–724.
 44. Urban TL. Inventory models with inventory-level-dependent demand: a comprehensive review and unifying theory. *Eur J Oper Res* 2005;162(5):792–804.
 45. Padmanabhan G, Vrat P. EOQ models for perishable items under stock dependent selling rate. *Eur J Oper Res* 1995;86(2):281–292.
 46. Giri BC, Pal S, Goswami A, *et al.* An inventory model for deteriorating items with stock-dependent demand rate. *Eur J Oper Res* 1996;95(3):604–610.
 47. Sarker BR, Mukherjee S, Balan CV. An order-level lot size inventory model with inventory-level dependent demand and deterioration. *Int J Prod Econ* 1997;48(3):227–236.
 48. Balkhi ZT, Benkherouf L. On an inventory model for deteriorating items with stock dependent and time-varying demand rates. *Comput Oper Res* 2004;31(2):223–240.
 49. Giri BC, Chaudhuri KS. Deterministic models of perishable inventory with stock-dependent demand rate and nonlinear holding cost. *Eur J Oper Res* 1998;105(3):467–474.
 50. Anderson ET, Fitzsimons GJ, Simester D. Measuring and mitigating the costs of stock-outs. *Manage Sci* 2006;52(11):1751–1763.
 51. Dana JD. Competition in price and availability when availability is unobservable. *RAND J Econ* 2001;32:497–513.
 52. Dana JD Jr, Petrucci NC. Note: the news-vendor model with endogenous demand. *Manage Sci* 2001;47(11):1488–1497.
 53. van Donselaar KH, Gaur V, van Woensel T, *et al.* Ordering behavior in retail stores and implications for automated replenishment. *Manage Sci* 2010;56(5):766–784.
 54. Kesavan S, Gaur V, Raman A. Do inventory and gross margin data improve sales forecasts for U.S. public retailers?. *Manage Sci* 2010;56(9):1519–1533.
 55. Lee HL. Ultimate enterprise value creation using demand-based management. Working paper. Stanford Global Supply Chain Management Forum. Stanford University; 2001.
 56. Keaveney SM. Customer switching behavior in service industries: an exploratory study. *J Mark* 1995;59(2):71–82.
 57. Lippman SA, McCardle KF. The competitive newsboy. *Oper Res* 1995;45(1):54–65.
 58. Hall J, Porteus E. Customer service competition in capacitated systems. *Manuf Serv Oper Manage* 2001;2(2):144–165.
 59. Deneckere R, Peck J. Competition over price and service rate when demand is stochastic: a strategic analysis. *RAND J Econ* 1995;26(1):148–162.
 60. Gans N. Customer loyalty and supplier quality competition. *Manage Sci* 2002;48(2):207–221.
 61. Gaur V, Park Y. Asymmetric consumer learning and inventory competition. *Manage Sci* 2007;53(2):227–240.
 62. Hardie BGS, Johnson EJ, Fader PS. Modeling loss aversion and reference dependence effects on brand choice. *Mark Sci* 1993;12(4):378–394.

63. Bernstein F, Federgruen A. A general equilibrium model for industries with price and service competition. *Oper Res* 2004; 52(6):868–886.
64. Tsay A, Agrawal N. Channel dynamics under price and service competition. *Manuf Serv Oper Manage* 2000;2(4):372–391.
65. Netessine S, Rudi N, Wang Y. Inventory competition and incentives to back-order. *IIE Trans* 2006;38:883–902.
66. Li L. The role of inventory in delivery-time competition. *Manage Sci* 1992;38(2):182–197.
67. McGahan AM, Ghemawat P. Competition to retain customers. *Mark Sci* 1994;13(2):165–176.
68. Balachander S, Farquhar PH. Gaining more by stocking less: a competitive analysis of product availability. *Mark Sci* 1994;13(1):3–22.
69. Grahovac J, Chakravarty A. Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Manage Sci* 2001;47(4):579–594.
70. Tagaras G. Pooling in multi-location periodic inventory distribution systems. *Omega* 1999;27:39–59.
71. Variety or duplication: a process to know where you stand. Washington, DC: The Research Department, Food Marketing Institute; 1993.
72. Bassok Y, Anupindi R, Akella R. Single-period multi-product inventory models with substitution. *Oper Res* 1999;47(4):632–642.
73. Gaur V, Honhon D. Assortment planning and inventory decisions under a locational choice model. *Manage Sci* 2007;52(10):1528–1543.
74. Netessine S, Rudi N. Centralized and competitive inventory models with demand substitution. *Oper Res* 2003;51(2):329–335.
75. Parlar M, Goyal SK. Optimal ordering decisions for two substitutable products with stochastic demands. *Opsearch* 1984;21:1–15.
76. Ernst R, Kouvelis P. The effects of selling packaged goods on inventory decisions. *Manage Sci* 1999;45(8):1142–1155.
77. McGillivray A, Silver EA. Some concepts for inventory control under substitutable demands. *INFOR* 1978;16:47–63.
78. Pasternack B, Drezner Z. Optimal inventory policies for substitutable commodities with stochastic demand. *Naval Res Logist* 1991;38:221–240.
79. Needham PM, Evers PT. The influence of individual cost factors on the use of emergency transshipments. *Transp Res E* 1998; 34:149–160.
80. Lee YH, Jung JW, Jeon YS. An effective lateral transshipment policy to improve service level in the supply chain. *Int J Prod Econ* 2007;106:115–126.
81. Lee HL. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Manage Sci* 1987;33(10):1302–1316.
82. Axsäter S. Modeling emergency lateral transshipments in inventory systems. *Manage Sci* 1990;36(11):1329–1338.
83. Das C. 1975. Supply and redistribution rules for two-location inventory systems: one period analysis. *Manage Sci* 1975;21(7):765–776.
84. Sherbrooke CC. METRIC: a multi-echelon technique for recoverable item control. *Oper Res* 1968;16(1):122–141.
85. Archibald TW, Sassen SA, Thomas LC. An optimal policy for a two depot inventory problem with stock transfer. *Manage Sci* 1997; 43(2):173–183.
86. Dada M. A two-echelon inventory system with priority shipments. *Manage Sci* 1992;38(8):1140–1153.
87. Axsäter S. Evaluation of unidirectional lateral transshipments and substitutions in inventory system. *Eur J Oper Res* 2003;149: 438–447.
88. Axsäter S. New decision rule for lateral transshipments in inventory systems. *Manage Sci* 2003;49(9):1168–1179.
89. Kukreja A, Schmidt CP, Miller DM. Stocking decisions for low-usage items in multi-location inventory system. *Manage Sci* 2001;47(10):1371–1383.
90. Robinson LW. Optimal and approximate policies in multiperiod, multilocation inventory models with transshipments. *Oper Res* 1990;38(2):278–295.
91. Minner S, Silver EA. Evaluation of two simple extreme transshipment strategies. *Int J Prod Econ* 2005;93–94:1–11.
92. Rudi N, Kapur S, Pyke D. A two-location inventory model with transshipment and local decision making. *Manage Sci* 2001;47(12):1668–1680.
93. Dong L, Rudi N. Who benefits from transshipment? Exogenous vs. endogenous wholesale prices. *Manage Sci* 2004;50(5): 645–657.

94. Zhao H, Deshpande V, Ryan JK. Emergency transshipment in decentralized dealer networks: when to send and accept transshipment requests. *Naval Res Logist* 2006;53:547–567.
95. Chiou CC. Transshipment problems in supply chain systems: review and extensions. In: Kordic V, editor. *Supply chain, theory and applications*. Vienna: I-Tech Education and Publishing; 2008.

CENTRAL PATH AND BARRIER ALGORITHMS FOR LINEAR OPTIMIZATION

KES ROOS

Faculty of Electrical Engineering,
Mathematics and Computer
Science, Delft University of
Technology, Delft,
The Netherlands

We consider the linear optimization (LO) problem in the standard form

$$\min \{c^T x : Ax = b, x \geq 0\},$$

with its dual problem

$$\max \{b^T y : A^T y + s = c, s \geq 0\}.$$

Here, $A \in \mathbf{R}^{m \times n}$, $b, y \in \mathbf{R}^m$, and $c, x, s \in \mathbf{R}^n$. Without loss of generality we assume that $\text{rank}(A) = m$. The vectors x, y , and s are the vectors of variables.

Our aim is to show how this primal–dual pair of LO problems can be solved in polynomial time by interior-point methods (IPMs). In most IPMs, the central path acts as a guideline to the set of optimal solutions. Assuming that a feasible solution is known that is close enough to the central path of a given LO problem, the problem can be solved by a simple numerical procedure that generates iterates on (or close to) the central path until we are close enough to the optimal set. In this way, one may find in polynomial time a solution with prescribed accuracy. One distinguishes between primal IPMs, dual IPMs, and primal–dual IPMs. Primal (dual) IPMs generate only primal (dual) feasible iterates, whereas primal–dual methods generate feasible iterates for both (P) and (D). In all cases, a crucial question concerns the availability of a starting point for the algorithm, that is, a feasible solution (close enough to the central path) of the given LO problem. If such a solution is at hand, the problem can be solved

with the just described method, which is then called a *feasible IPM*. In most cases, however, such a starting solution is not available. This certainly occurs if the given LO problem is infeasible or unbounded. In such cases, one may construct an artificial problem (see *Self-Dual Embedding Technique for Linear Optimization*) that can be solved by a feasible IPM and whose solution either informs us that the given LO problem is infeasible or unbounded, or it yields an optimal solution for the given problem. In this article, we leave out the issue of finding a starting point on the central path or close to it. Rather, we concentrate on the definition and main properties of the central path.

The theory underlying IPMs for LO has been developed in the last 25 years and is the result of the common effort of many researchers. In this article, we focus on elementary parts of this theory; nevertheless, trying to get a rather complete treatment of its main parts. Despite this, some topics could not be covered. For example, we do not touch the so-called *infeasible* IPMs; neither do we discuss sensitivity analysis based on the use of optimal partitions nor the so-called *potential reduction* IPMs and *target-following* methods. For these and related topics, we refer to the survey papers [1–3] and to some books devoted to IPMs [4–8].

We conclude this introduction with a simple lemma that will be used frequently in the sequel.

Lemma 1. *If x is feasible for (P) and (y, s) for (D) then $c^T x - b^T y = x^T s \geq 0$.*

Proof. Using the feasibility of x and (y, s) , we may write

$$\begin{aligned} c^T x - b^T y &= c^T x - (Ax)^T y = c^T x - x^T A^T y \\ &= c^T x - x^T (c - s) = x^T s. \end{aligned}$$

Since x and s are nonnegative, $x^T s \geq 0$. Hence the lemma follows.

This lemma implies the so-called *weak duality property*: If x is feasible for (P) and (y, s) for (D) then $c^T x \geq b^T y$. The difference $c^T x - b^T y$ is called the *duality gap* for the triple (x, y, s) .

NOTATIONS

The set of real numbers is denoted as $\mathbb{R}$. $\mathbb{R}^n$ and $\mathbb{R}_+^n$ denote the set of real vectors of length n and its subset of nonnegative vectors, respectively. The zero vector and all-one vector in $\mathbb{R}^n$ are denoted as 0_n and e_n , respectively; if this gives no rise to confusion we drop the subscript and simply write 0 or e . If f is a univariate function and $x \in \mathbb{R}^n$ then $f(x)$ is the vector with coordinates $f(x_i)$. This defines, for example, for any positive vector x the vectors x^2 , x^{-1} and $\sqrt{x}$. To any vector $x \in \mathbb{R}^n$, we associate the diagonal matrix X whose diagonal entries are the elements of x , in the same order. If also $s \in \mathbb{R}^n$, then Xs will be denoted in short as xs . So xs is the so-called Hadamard product: its entries are obtained by componentwise multiplication of x and s . Finally, if s is positive, then we denote xs^{-1} also as x/s .

PRIMAL METHOD

The main source of difficulties in solving the problem (P) is the presence of the inequality constraints $x \geq 0$. A basic idea in IPMs is to replace these inequalities by a barrier function and to reduce the solution of (P) to solving a sequence of minimization problems subject to the equality constraints $Ax = b$ only. For a detailed discussion of the history of this technique, we refer to the classic book of Fiacco and McCormick [9]. Nowadays it is commonly accepted that the logarithmic barrier function $-\sum_{i=1}^n \log x_i$ is very useful for our purpose. It was introduced in 1955 by Frisch [10]. Its domain is the interior of the nonnegative orthant, $\mathbb{R}_+^n$, and it reaches infinity if one approaches the boundary of $\mathbb{R}_+^n$. So, instead of (P), we consider the problem

$$\min \left\{ f_P(x; \mu) := \frac{c^T x}{\mu} - \sum_{i=1}^n \log x_i : Ax = b \right\}, \quad (\text{P}_\mu)$$

where μ may be any positive number. If μ approaches zero then the term containing $c^T x$ is given so much emphasis in the objective $f_P(x; \mu)$ that we may hope that the optimal solution of (P_μ) will provide a good approximation of an optimal solution of (P). We show below that the method works under the assumption that a strictly feasible solution x^0 , close to the minimizer of $f_P(x; \mu^0)$ for some $\mu^0 > 0$, is readily available. See the article titled **Self-Dual Embedding Technique for Linear Optimization** describes how to get into this situation.

Primal Newton Step

The working horse in all primal (and dual) IPMs is Newton's method for minimizing a convex function. Indeed, $f_P(x; \mu)$ is convex in x . One easily verifies this by computing the gradient and Hessian, which are as follows:

$$\nabla f_P(x; \mu) = \frac{c}{\mu} - x^{-1} \quad \text{and} \quad \nabla^2 f_P(x; \mu) = X^{-2}.$$

Since X is a positive diagonal matrix, it is positive definite, and so is X^{-2} , whence the convexity of $f_P(x; \mu)$ follows. The Newton step Δx at x is obtained by minimizing the second-order Taylor expansion of $f_P(x; \mu)$ subject to the constraint $A\Delta x = 0$:

$$\min \left\{ f_P(x; \mu) + \left(\frac{c}{\mu} - x^{-1} \right)^T \Delta x + \frac{1}{2} \Delta x^T X^{-2} \Delta x : A\Delta x = 0 \right\}.$$

Using a vector ξ as Lagrange multiplier, the first-order optimality conditions for this problem are

$$\begin{aligned} \frac{c}{\mu} - x^{-1} + X^{-2} \Delta x &= A^T \xi, \\ A\Delta x &= 0. \end{aligned}$$

Multiplying the first equation from the left with X and rearranging the terms give the equivalent system

$$\begin{aligned} e - \frac{Xc}{\mu} &= X^{-1} \Delta x - (AX)^T \xi, \\ AX \cdot X^{-1} \Delta x &= 0. \end{aligned}$$

The second equation shows that $X^{-1}\Delta x$ belongs to the null space of the matrix AX . Since $(AX)^T\xi$ belongs to the row space of AX , which is orthogonal to its null space, it follows that the first equation represents an orthogonal decomposition of the vector $e - \frac{Xc}{\mu}$ along the row space and the null space of AX . Denoting the orthogonal projection onto the null space of AX as P_{AX} implies that

$$X^{-1}\Delta x = P_{AX} \left(e - \frac{Xc}{\mu} \right). \quad (1)$$

An immediate consequence is that $X^{-1}\Delta x$ is the shortest vector in the 2-norm in the affine space containing $e - \frac{Xc}{\mu}$ and parallel to the row space of AX . The vectors in this affine space have the form

$$\begin{aligned} e - \frac{Xc}{\mu} + (AX)^T\xi &= e - \frac{X(c - A^T(\mu\xi))}{\mu}, \quad \xi \in \mathbb{R}^m \\ &= e - \frac{X(c - A^Ty)}{\mu}, \quad y \in \mathbb{R}^m. \end{aligned}$$

Now let $y(x, \mu)$ denote the vector y that minimizes the 2-norm of the last expression, and $s(x, \mu) = c - A^Ty(x, \mu)$. In other words, with $\|\cdot\|$ denoting the 2-norm,

$$s(x, \mu) := \operatorname{argmin}_s \left\{ \left\| e - \frac{Xs}{\mu} \right\| : A^Ty + s = c \right\}. \quad (2)$$

Then it follows that the Newton direction satisfies

$$X^{-1}\Delta x = e - \frac{Xs(x, \mu)}{\mu}. \quad (3)$$

This characterization of the Newton direction turns out to be extremely useful, as it will become clear in the analysis below.

Primal Proximity Measure and Quadratic Convergence

For the moment we assume that $f_P(x; \mu)$ has a minimizer, which we denote as $x(\mu)$. To measure progress of the Newton process, we need a measure for the distance of x to $x(\mu)$. Owing to the convexity of $f_P(x; \mu)$, x is a minimizer if and only if the Newton step at x equals the zero vector. A natural candidate

for measuring proximity to $x(\mu)$ is therefore the norm of the Newton step Δx . Instead, before taking the norm, we scale Δx to $X^{-1}\Delta x$, and define

$$\delta(x, \mu) := \|X^{-1}\Delta x\| = \left\| e - \frac{Xs(x, \mu)}{\mu} \right\|, \quad (4)$$

where the last equality is due to Equation (3). Note that $\delta(x, \mu) = 0$ if $x = x(\mu)$, because then $\Delta x = 0$, and otherwise $\delta(x, \mu) > 0$. The point that results by taking a (full) Newton step at x is denoted as x^+ . According to Equation (3), we may write

$$\begin{aligned} x^+ &= x + \Delta x = x + X \left(e - \frac{Xs(x, \mu)}{\mu} \right) \\ &= X \left(2e - \frac{Xs(x, \mu)}{\mu} \right). \end{aligned} \quad (5)$$

Using this we can show that the Newton process is quadratically convergent in terms of the proximity measure $\delta(x, \mu)$, provided that x is close enough to $x(\mu)$.

Lemma 2. *If $\delta(x, \mu) < 1$ then x^+ is a strictly feasible solution of (P). Moreover, $\delta(x^+, \mu) \leq \delta(x, \mu)^2$.*

Proof. It will be convenient to introduce the vector w as follows:

$$w := \frac{Xs(x, \mu)}{\mu}. \quad (6)$$

Then we have $\delta(x, \mu) = \|e - w\|$. Hence, $\delta(x, \mu) < 1$ implies that $\|e - w\| < 1$. Therefore, $0 < w_i < 2$, for each i , which implies that $2e - w > 0$. Using Equation (5), we get $x^+ = X(2e - w) > 0$, proving that x^+ is positive. The relation $Ax^+ = b$ also holds, because $Ax = b$ and $A\Delta x = 0$. This proves the first statement in the theorem.

The definition of $s(x^+, \mu)$ implies the following:

$$\begin{aligned} \delta(x^+, \mu) &= \left\| e - \frac{Xs(x^+, \mu)}{\mu} \right\| \leq \left\| e - \frac{Xs(x, \mu)}{\mu} \right\| \\ &= \|e - X^{-1}X^{-1}w\|. \end{aligned}$$

Using once more that $x^+ = X(2e - w)$ we find

$$\begin{aligned} e - X + X^{-1}w &= e - X(2I - W)X^{-1}w \\ &= e - (2I - W)w \\ &= e - 2w + Ww = (e - w)^2. \end{aligned}$$

In this chain of equalities, we used the fact that diagonal matrices commute. We also recall that, according to our convention, $(e - w)^2$ denotes the Hadamard product of $(e - w)$ with itself. Substituting this we obtain

$$\delta(x^+, \mu)^2 = \|(e - w)^2\|^2 \leq \|w - e\|^4 = \delta(x, \mu)^4,$$

where we used that $\|z^2\| \leq \|z\|^2$, for any $z \in \mathbb{R}^n$. Hence, it follows that $\delta(x^+, \mu) \leq \delta(x, \mu)^2$, which was to be shown.

The importance of Lemma 2 is clear: if we repeat doing Newton steps, the proximity measure converges quadratically fast to zero, which implies that the iterates converge to $x(\mu)$. As a corollary we may state the following.

Corollary 1. *If $\delta(x, \mu) < 1$ for some (strictly feasible) x then $f_P(x; \mu)$ has a minimizer.*

Proof. If we repeat doing Newton steps, then the values of $\delta(x, \mu)$ converge (quadratically fast) to zero. The limit point $\bar{x}$ satisfies $\delta(\bar{x}, \mu) = 0$, which implies that the Newton step $\Delta\bar{x}$ at $\bar{x}$ vanishes; this means that $\bar{x}$ is a minimizer of $f_P(x, \mu)$.

Updating the Barrier Parameter

Using Corollary 1, we show in this section that if $f_P(x; \mu)$ has a minimizer for some positive μ then $f_P(x; \mu)$ has a minimizer for every positive μ . For this we need the following lemma. Note that in this lemma, we allow θ to be negative.

Lemma 3. *Let $\theta < 1$ and $\mu^+ = (1 - \theta)\mu$. Then*

$$\delta(x, \mu^+) \leq \frac{\delta(x, \mu) + |\theta|\sqrt{n}}{1 - \theta}.$$

Proof. With w , as defined in Equation (6), the definition of $s(x, \mu)$ implies the following:

$$\begin{aligned} \delta(x, \mu^+) &= \left\| e - \frac{Xs(x, \mu^+)}{\mu^+} \right\| \leq \left\| e - \frac{Xs(x, \mu)}{\mu^+} \right\| \\ &= \left\| e - \frac{w}{1 - \theta} \right\| = \frac{\|(e - w) - \theta e\|}{1 - \theta}. \end{aligned}$$

Since $\|e - w\| = \delta(x, \mu)$ and $\|\theta e\| = |\theta|\sqrt{n}$, the lemma follows by applying the triangle inequality.

Corollary 2. *If $\delta(x, \mu) < 1$ for some (strictly feasible) x and some $\mu > 0$, then $f_P(x; \mu)$ has a minimizer for every $\mu > 0$.*

Proof. Suppose $\delta(x, \mu^0) < 1$ for some $\mu^0 > 0$. Then, by Corollary 1, $f_P(x; \mu^0)$ has a minimizer. We express this by saying that $x(\mu^0)$ exists. Taking $x = x(\mu^0)$, we then have $\delta(x, \mu^0) = 0$. Using Lemma 3 one easily deduces from this that $\delta(x, \mu) < 1$ whenever

$$\frac{\sqrt{n}}{\sqrt{n} + 1} \mu^0 < \mu < \frac{\sqrt{n}}{\sqrt{n} - 1} \mu^0. \quad (7)$$

Using Corollary 1 again, we conclude that $x(\mu)$ exists for any such μ . Thus, we see that whenever $x(\mu)$ exists for some $\mu = \mu^0 > 0$ then $x(\mu)$ also exists for all μ in the open interval around μ^0 as given by Equation (7). By applying the same argument to the points in this interval one easily understands that $x(\mu)$ exists for all μ satisfying

$$\left(\frac{\sqrt{n}}{\sqrt{n} + 1} \right)^2 \mu^0 < \mu < \left(\frac{\sqrt{n}}{\sqrt{n} - 1} \right)^2 \mu^0.$$

By repeating this argument k times, we get that $x(\mu)$ exists for all μ satisfying

$$\left(\frac{\sqrt{n}}{\sqrt{n} + 1} \right)^k \mu^0 < \mu < \left(\frac{\sqrt{n}}{\sqrt{n} - 1} \right)^k \mu^0.$$

By taking k large enough we can reach every positive value of μ , and hence the corollary follows.

Central Path

We are now in a position to define the central path of (P). Assuming again that $\delta(x, \mu) < 1$ for some (strictly feasible) x and some $\mu > 0$, we have shown that $f_P(x; \mu)$ has a minimizer, denoted as $x(\mu)$, for every positive μ . If μ runs through all positive real numbers, then $x(\mu)$ runs through a curve in the interior of the feasible region of (P). This curve is called the *central path* of (P). Below we show that if μ approaches zero, then the central path converges to an optimal solution of (P). As a consequence, if (P) has a unique optimal solution, which then is a vertex of the feasible region, then the central path converges to this vertex. If, however, there are more optimal solutions, then the limit of the central path is a specific optimal solution, called the *analytic center* of the optimal set, which is a strictly complementary solution, certainly not a vertex. This reveals a remarkable difference with the simplex method, which always generates a vertex solution.

In this section, we discuss some important properties of the central path. We start with an enlightening characterization of the points on the central path. As we discussed earlier, due to the convexity of $f_P(x; \mu)$, x is a minimizer if and only if the Newton step equals the zero vector. Because of Equations (2) and (3) this happens if and only if $e - \frac{xs(x, \mu)}{\mu} = 0$, or, equivalently, $xs(x, \mu) = \mu e$. We conclude that x minimizes $f_P(x; \mu)$ if and only if there exist y and s such that

$$Ax = b, x \geq 0, \quad (8)$$

$$A^T y + s = c, s \geq 0, \quad (9)$$

$$xs = \mu e. \quad (10)$$

Note that Equation (8) is the condition for primal feasibility and Equation (9) is the condition for dual feasibility. Condition (10) is called the *centering condition*. Since the system (Eqs 8–10) represents the first-order optimality conditions for a minimizer of $f_P(x; \mu)$; it therefore is also known as the *Karush–Kuhn–Tucker* (KKT) system for (P).

Now let the triple (x, y, s) solve this system (for some fixed $\mu > 0$). Since $f_P(x; \mu)$ has a unique minimizer, x is the unique solution of the system (Eqs 8–10). But, in view of

centering condition (10) s is also unique, so is y because of Equation (9) and because A has full row rank. This shows that the KKT system has a unique solution, for every $\mu > 0$. This solution is denoted as $(x(\mu), y(\mu), s(\mu))$, while $x(\mu)$ is called the μ center of (P), and $(y(\mu), s(\mu))$ the μ center of (D). Indeed, as will become clear in the section titled “Dual Method”, the dual problem also has a central path, and it consists of the pairs $(y(\mu), s(\mu))$, with $\mu > 0$.

We conclude this section with four important remarks. First, in see the article titled ***Self-Dual Embedding Technique for Linear Optimization*** it is shown that every primal–dual pair of problems (P) and (D) can be embedded in a (self-dual) artificial problem for which a solution of Equations (8)–(10) is at hand with $\mu = 1$. This means that the artificial problem has a central path. The importance of this artificial problem is that the information at the limit of its central path can be used to solve (P) and (D).

The second remark concerns the condition that we used so far to guarantee the existence of the central path of (P): the assumption that $\delta(x, \mu) < 1$ for some (strictly feasible) x and some $\mu > 0$. This condition can be weakened considerably, however. Note that the centering condition implies that $x(\mu) > 0$ and $s(\mu) > 0$ for every $\mu > 0$. This means that the central path can exist only if (P) has a strictly feasible solution x , that is, satisfying $Ax = b, x > 0$ and (D) has a strictly feasible solution (y, s) , that is, satisfying $A^T y + s = c, s > 0$. We want to point out that the converse is also true: if both (P) and (D) have a strictly feasible (or interior) solution, then the central path exists. In this presentation we do not need this fact, for its proof we refer to Refs 5–8. As a consequence, we may state that the central path exists if and only if both (P) and (D) have a strictly feasible solution. We refer to this condition as the *interior-point condition* (IPC).

Third, we want to emphasize the surprising fact that one may associate a dual feasible solution (y, s) to any primal feasible solution x on (or close to) the central path. This is now obvious if x is on the central path, say $x = x(\mu)$, because then we may take $y = y(\mu)$ and $s = s(\mu)$. The duality gap for

these solutions can be obtained from Lemma 1 and the centering equation

$$\begin{aligned} c^T x(\mu) - b^T y(\mu) &= x(\mu)^T s(\mu) = e^T(x(\mu)s(\mu)) \\ &= e^T(\mu e) = n\mu. \end{aligned} \quad (11)$$

The following lemma generalizes this result to the case where x is close to the μ center $x(\mu)$.

Lemma 4. *If $\delta(x, \mu) \leq 1$, then the pair $(y(x, \mu), s(x, \mu))$ is dual feasible. Moreover,*

$$\begin{aligned} \mu(n - \delta(x, \mu)\sqrt{n}) &\leq c^T x - b^T y(x, \mu) \\ &\leq \mu(n + \delta(x, \mu)\sqrt{n}). \end{aligned}$$

Proof. As noted earlier, if w is defined by Equation (6), then $\delta(x, \mu) = \|e - w\|$. As it was pointed out earlier, $\delta(x, \mu) \leq 1$ implies that w is nonnegative. Hence, $Xs(x, \mu)$ is nonnegative. Since X is a positive diagonal matrix, it follows that $s(x, \mu) \geq 0$, proving the first statement in the lemma. Owing to Lemma 1, we therefore may write

$$c^T x - b^T y(x, \mu) = x^T s(x, \mu).$$

Furthermore, application of the Cauchy–Schwarz inequality gives

$$\begin{aligned} \delta(x, \mu)\sqrt{n} &= \left\| e - \frac{Xs(x, \mu)}{\mu} \right\| \|e\| \\ &\geq \left| e^T \left(e - \frac{Xs(x, \mu)}{\mu} \right) \right| \\ &= \left| n - \frac{x^T s(x, \mu)}{\mu} \right|. \end{aligned}$$

From this we derive

$$-\delta(x, \mu)\sqrt{n} \leq n - \frac{x^T s(x, \mu)}{\mu} \leq \delta(x, \mu)\sqrt{n},$$

which implies the inequalities in the lemma.

Fourth, and last, along the primal central path the primal objective value is monotonically increasing with respect to μ . This can be understood easily. Letting

$0 < \mu_1 \leq \mu_2$ and $x^1 = x(\mu_1)$ and $x^2 = x(\mu_2)$, we claim that $c^T x^1 \leq c^T x^2$. Since x^1 minimizes $f_P(x; \mu_1)$, we have $f_P(x^1; \mu_1) \leq f_P(x^2; \mu_1)$. For a similar reason, $f_P(x^2; \mu_2) \leq f_P(x^1; \mu_2)$. In other words,

$$\begin{aligned} \frac{c^T x^1}{\mu_1} - \sum_{i=1}^n \log x_i^1 &\leq \frac{c^T x^2}{\mu_1} - \sum_{i=1}^n \log x_i^2, \\ \frac{c^T x^2}{\mu_2} - \sum_{i=1}^n \log x_i^2 &\leq \frac{c^T x^1}{\mu_2} - \sum_{i=1}^n \log x_i^1. \end{aligned}$$

The sums in these inequalities can be eliminated by adding the above inequalities, which gives

$$\frac{c^T x^1}{\mu_1} + \frac{c^T x^2}{\mu_2} \leq \frac{c^T x^2}{\mu_1} + \frac{c^T x^1}{\mu_2},$$

which is equivalent to

$$\left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) (c^T x^1 - c^T x^2) \leq 0.$$

Since $\frac{1}{\mu_1} - \frac{1}{\mu_2} > 0$, we obtain $c^T x^1 \leq c^T x^2$, proving the claim. A similar result holds for the dual central path: the dual objective value is monotonically decreasing with respect to μ . The proof is easy and similar to the previous case. We shall dispense with it because it involves the dual logarithmic barrier function that is introduced in the section titled “Dual Methods”. Since $c^T x(\mu) - b^T y(\mu) = n\mu$, from Equation (11), it follows that if μ approaches zero then $c^T x(\mu)$ and $b^T y(\mu)$ converge to the same value, which is the common optimal objective value of (P) and (D).

Primal Interior-Point Method

The following lemma provides a numerical method to follow the primal central path (approximately).

Lemma 5. *Let $\delta(x, \mu) \leq \frac{1}{2}$. If $\theta = \frac{1}{6\sqrt{n}}$ and x^+ and μ^+ are as defined before, then $\delta(x^+, \mu^+) \leq \frac{1}{2}$.*

Proof. Using Lemmas 2 and 3, successively we may write

$$\delta(x^+, \mu^+) \leq \frac{\delta(x^+, \mu) + \theta\sqrt{n}}{1 - \theta} \leq \frac{\delta(x, \mu)^2 + \theta\sqrt{n}}{1 - \theta}.$$

Since $\delta(x, \mu) \leq \frac{1}{2}$ and $\theta\sqrt{n} = \frac{1}{6}$, it follows that

$$\delta(x^+, \mu^+) \leq \frac{\frac{1}{4} + \frac{1}{6}}{\frac{5}{6}} = \frac{1}{2},$$

proving the lemma.

The above lemma makes clear that by repeatedly performing the Newton step and a barrier update with $\theta = \frac{1}{6\sqrt{n}}$, we can get as close to the optimal set as we wish. A more precise formulation of the algorithm is as follows:

Algorithm 1. A full-Newton step primal IPM for LO

Input:

(x^0, μ^0) , with x strictly feasible, $\mu^0 > 0$, and $\delta(x^0, \mu^0) \leq \frac{1}{2}$;
accuracy parameter $\varepsilon > 0$;
barrier update parameter θ , $0 < \theta < 1$;

begin

$x := x^0; \mu := \mu^0$.

while $n\mu > \varepsilon$

$x := x + \Delta x$;

$\mu := (1 - \theta)\mu$.

endwhile

end

Theorem 1. Taking $\theta = \frac{1}{6\sqrt{n}}$, the algorithm requires no more than

$$6\sqrt{n} \log \frac{n\mu^0}{\varepsilon}$$

iterations. The last generated point x is strictly feasible, whereas $y(x, \mu)$ is dual feasible. Moreover, $c^T x - b^T y(x, \mu) \leq \frac{3}{2}\varepsilon$.

Proof. By Lemma 5, after each iteration of the algorithm, the vector x will be strictly feasible and $\delta(x, \mu) \leq \frac{1}{2}$. After the k th iteration, we have $\mu = (1 - \theta)^k \mu^0$. Hence, the algorithm must have stopped if k is such that $(1 - \theta)^k n\mu^0 \leq \varepsilon$. By taking logarithms on both sides, the inequality reduces to

$$-k \log(1 - \theta) \geq \log \frac{n\mu^0}{\varepsilon}.$$

Since $-\log(1 - \theta) > \theta$, this will certainly hold if $k\theta \geq \log \frac{n\mu^0}{\varepsilon}$. Substituting the value of θ , we find the iteration bound in the theorem.

Now let x be the last generated point and μ the final value of the barrier parameter.

Then $n\mu \leq \varepsilon$. From Lemma 4, we obtain that $y(x, \mu)$ is dual feasible. Finally, using $n \geq 1$, we may write

$$\begin{aligned} c^T x - b^T y(x, \mu) &\leq \mu(n + \delta(x, \mu)\sqrt{n}) \\ &\leq \frac{\varepsilon}{n}(n + \delta(x, \mu)\sqrt{n}) \\ &= \varepsilon \left(1 + \frac{\delta(x, \mu)}{\sqrt{n}} \right) \leq \frac{3}{2}\varepsilon. \end{aligned}$$

This completes the proof.

DUAL METHOD

It will be no surprise that we can apply a similar approach to solve the dual problem as described for the primal problem in the previous section. In this section, we briefly describe how this can be worked out, without going into too much detail.

First we eliminate the nonnegativity constraint $s \geq 0$ by using the logarithmic barrier function $-\sum_{i=1}^n \log s_i$. So, to solve (D), we

consider the auxiliary problem

$$\max \left\{ f_D(y; \mu) := \frac{b^T y}{\mu} + \sum_{i=1}^n \log s_i : A^T y + s = c \right\}, \quad (D_\mu)$$

where μ may be any positive number. Since A has full row rank, we feel free to represent a dual feasible pair (y, s) by either y or s , because either of these uniquely determines the other.

The gradient and Hessian, with respect to y , are as follows:

$$\nabla_y f_D(y; \mu) = \frac{b}{\mu} - A s^{-1}$$

and

$$\nabla_y^2 f_D(y; \mu) = A S^{-2} A^T,$$

where s^{-1} denotes the vector whose entries are s_i^{-1} and $S = \text{diag}(s)$. Hence, the Newton step is given as

$$\begin{aligned} \Delta y &= (A S^{-2} A^T)^{-1} \left(\frac{b}{\mu} - A s^{-1} \right), \\ \Delta s &= -A^T \Delta y. \end{aligned}$$

Let $\bar{x}$ be such that $A\bar{x} = b$. Then, we may write

$$\begin{aligned} S^{-1} \Delta s &= -S^{-1} A^T (A S^{-2} A^T)^{-1} \left(\frac{A\bar{x}}{\mu} - A s^{-1} \right) \\ &= -S^{-1} A^T (A S^{-2} A^T)^{-1} A S^{-1} \left(\frac{S\bar{x}}{\mu} - e \right). \end{aligned}$$

At this stage we recognize $S^{-1} A^T (A S^{-2} A^T)^{-1} A S^{-1}$ as the matrix of the orthogonal projection onto the row space of $A S^{-1}$. Denoting this matrix as $P_{A S^{-1}}^\perp$, we thus have

$$S^{-1} \Delta s = P_{A S^{-1}}^\perp \left(e - \frac{S\bar{x}}{\mu} \right). \quad (12)$$

Defining

$$x(s, \mu) := \operatorname{argmin}_x \left\{ \left\| e - \frac{Sx}{\mu} \right\| : Ax = b \right\}, \quad (13)$$

one easily understands, just using the same arguments as in the primal case, that

$$S^{-1} \Delta s = e - \frac{Sx(s, \mu)}{\mu}. \quad (14)$$

The point that results by taking a (full) Newton step at s is denoted as s^+ . According to Equation (14), we may write

$$\begin{aligned} s^+ &= s + \Delta s = s + S \left(e - \frac{Sx(s, \mu)}{\mu} \right) \\ &= S \left(2e - \frac{Sx(s, \mu)}{\mu} \right). \end{aligned} \quad (15)$$

As proximity measure, we use

$$\delta(y, \mu) := \|S^{-1} \Delta s\| = \left\| e - \frac{Sx(s, \mu)}{\mu} \right\|. \quad (16)$$

We can now state the following results, whose proofs are omitted since they are completely similar to the proofs of the corresponding results for the primal case.

Lemma 6. *If $\delta(y, \mu) < 1$, then $y^+ := y + \Delta y$ is a strictly feasible solution of (D). Moreover, $\delta(y^+, \mu) \leq \delta(y, \mu)^2$.*

Corollary 3. *If $\delta(y, \mu) < 1$ for some (strictly feasible) y , then $f_D(y; \mu)$ has a maximizer.*

Lemma 7. *Let $\theta < 1$ and $\mu^+ = (1 - \theta)\mu$. Then*

$$\delta(y, \mu^+) \leq \frac{\delta(y, \mu) + |\theta| \sqrt{n}}{1 - \theta}.$$

Corollary 4. *If $\delta(y, \mu) < 1$ for some (strictly feasible) y and some $\mu > 0$, then $f_D(x; \mu)$ has a maximizer for every $\mu > 0$.*

On the basis of these results we may conclude that just as in the primal case, the (unique) maximizer of $f_D(y; \mu)$ is the pair $(y(\mu), s(\mu))$ which is determined by the system (8)–(10).

Lemma 8. *If $\delta(y, \mu) \leq 1$, then $x(s, \mu)$ is primal feasible. Moreover,*

$$\begin{aligned}\mu(n - \delta(x, \mu)\sqrt{n}) &\leq c^T x(s, \mu) - b^T y \\ &\leq \mu(n + \delta(x, \mu)\sqrt{n}).\end{aligned}$$

Lemma 9. *Let $\delta(y, \mu) \leq \frac{1}{2}$. If $\theta = \frac{1}{6\sqrt{n}}$ and (y^+, s^+) and μ^+ are as defined before, then $\delta(y^+, \mu^+) \leq \frac{1}{2}$.*

Owing to these results, the dual algorithm is quite similar to the primal algorithm (Algorithm 1.5), as shown below.

Algorithm 2. A full-Newton step dual IPM for LO

Input:

(y^0, s^0, μ^0) , with (y^0, s^0) strictly feasible, $\mu^0 > 0$, and $\delta(y^0, \mu^0) \leq \frac{1}{2}$;
accuracy parameter $\varepsilon > 0$;
barrier update parameter θ , $0 < \theta < 1$;

begin

$y := y^0; s := s^0; \mu := \mu^0$.

while $n\mu > \varepsilon$

$(y, s) := (y, s) + (\Delta y, \Delta s)$;

$\mu := (1 - \theta)\mu$.

endwhile

end

PRIMAL–DUAL METHOD

The primal and dual IPMs discussed so far are very much in the spirit of the barrier methods that received much attention in the 1970s. At that time much emphasis was given to proving convergence of such methods. Only after the work of Karmarkar it became clear (in the late 1980s) that by a proper setting of the parameters in the algorithm we can not only prove convergence but also obtain polynomial iteration bounds [11]. More importantly, a new class of methods was developed that take much more profit of the duality properties in LO. As we know, every LO problem has a dual problem. We also have seen, more or less to our surprise, that when solving an LO problem with a primal (or dual) IPM, the output of the algorithm is not only a primal optimal solution (with prescribed accuracy) but also a dual optimal solution (with the same accuracy). Note that this phenomenon also occurs when solving an LO problem with the simplex method. It has turned out, however, that IPMs can

Theorem 2. *Taking $\theta = \frac{1}{6\sqrt{n}}$, the algorithm requires no more than*

$$6\sqrt{n} \log \frac{n\mu^0}{\varepsilon}$$

iterations. The last generated pair (y, s) is strictly feasible, whereas $x(s, \mu)$ is primal feasible. Moreover, $c^T x(s, \mu) - b^T y \leq \frac{3}{2}\varepsilon$.

profit much more from the inherent duality in LO; namely, by designing methods that generate iterates both in the primal space and the dual space. These are the so-called primal–dual methods that we discuss in this section.

Primal–Dual Search Direction

Primal–dual methods heavily rely upon the KKT system (Eqs 8–10). Suppose we have a strictly feasible triple (x, y, s) —that is, x is strictly feasible for (P) and (y, s) strictly feasible for (D)—and that we want to find the μ centers of (P) and (D), where μ is an arbitrary positive number. Denoting $\Delta x = x(\mu) - x$, $\Delta y = y(\mu) - y$, and $\Delta s = s(\mu) - s$, this means that we are looking for a triple $(\Delta x, \Delta y, \Delta s)$ such that

$$A\Delta x = 0,$$

$$A^T \Delta y + \Delta s = 0,$$

$$(x + \Delta x)(s + \Delta s) = \mu e.$$

Unfortunately the third equation is nonlinear in the unknown vectors Δx and Δs , because of the quadratic term $\Delta x \Delta s$ in

$$(x + \Delta x)(s + \Delta s) = xs + x\Delta s + s\Delta x + \Delta x \Delta s.$$

At this stage we use a fundamental idea (which is again due to Newton) for solving nonlinear equation systems, namely, to ignore the nonlinear term. This leads to the following linear system of equations in the search directions Δx , Δy , and Δs :

$$A\Delta x = 0, \quad (17)$$

$$A^T \Delta y + \Delta s = 0, \quad (18)$$

$$x\Delta s + s\Delta x = \mu e - xs. \quad (19)$$

Lemma 10. *The system (Eqs 17–19) has a unique solution.*

Proof. The proof can be given in several ways. Our proof is based on a rescaling scheme that is also useful in the analysis of the algorithm that we present later. It uses the vector v defined by

$$v := \sqrt{\frac{xs}{\mu}}. \quad (20)$$

Obviously v equals the all-one vector if and only if $xs = \mu e$. In other words, $v = e$ holds if and only if $x = x(\mu)$ and $s = s(\mu)$. Using v , we define scaled versions $\dot{x}$ and $\dot{s}$ of the search directions Δx and Δs , respectively, according to

$$\dot{x} = \frac{v\Delta x}{x} \quad \text{and} \quad \dot{s} = \frac{v\Delta s}{s}. \quad (21)$$

In other words, with $V = \text{diag}(v)$, one has $\Delta x = V^{-1}X\dot{x}$ and $\Delta s = V^{-1}S\dot{s}$. Therefore, we may rewrite the system (Eqs 17–19) as follows:

$$\begin{aligned} AV^{-1}X\dot{x} &= 0, \\ A^T \Delta y + V^{-1}S\dot{s} &= 0, \\ xv^{-1}s\dot{s} + sv^{-1}x\dot{x} &= \mu e - xs. \end{aligned}$$

Since $xs = \mu v^2$, the third equation simplifies to $\mu v(\dot{x} + \dot{s}) = \mu(e - v^2)$, which is equivalent to $\dot{x} + \dot{s} = v^{-1} - v$. On the other hand, defining

$$d = \sqrt{\frac{x}{s}},$$

we have $V^{-1}X = \sqrt{\mu}D$ and $V^{-1}S = \sqrt{\mu}D^{-1}$, where $D = \text{diag}(d)$. So, the first equation is equivalent to $AD\dot{x} = 0$ and the second equation to $A^T \Delta y + \sqrt{\mu}D^{-1}\dot{s} = 0$, which can be rewritten as $(AD)^T \frac{\Delta y}{\sqrt{\mu}} + \dot{s} = 0$. Thus, with $\bar{A} = AD$, we obtain that

$$\bar{A}\dot{x} = 0, \quad (22)$$

$$\bar{A}^T \frac{\Delta y}{\sqrt{\mu}} + \dot{s} = 0, \quad (23)$$

$$\dot{x} + \dot{s} = v^{-1} - v. \quad (24)$$

From the first two equations we conclude that $\dot{x}$ belongs to the null space of $\bar{A}$ and $\dot{s}$ to the row space of $\bar{A}$. Hence, we derive from the third equation that $\dot{x}$ and $\dot{s}$ are the components of $v^{-1} - v$ in these two (orthogonal) spaces. Thus, we obtain that

$$\dot{x} = P_{\bar{A}}(v^{-1} - v) \quad \text{and} \quad \dot{s} = P_{\bar{A}}^\perp(v^{-1} - v), \quad (25)$$

where $P_{\bar{A}}$ denotes the orthogonal projection onto the null space of $\bar{A}$, and $P_{\bar{A}}^\perp$ the orthogonal projection onto orthogonal complement of the null space of $\bar{A}$, which is the row space of $\bar{A}$. This proves that $\dot{x}$ and $\dot{s}$ are unique. Since A has full row rank, so does $\bar{A}$. Therefore, it follows from Equation (23) that Δy is also unique. Hence, the lemma has been proved.

The search directions defined by Equations (17)–(19) are used in all primal–dual IPMs. Because of Equation (21) we may rewrite Equation (25) as follows:

$$\begin{aligned} \Delta x &= V^{-1}XP_{AD}(v^{-1} - v), \\ \Delta s &= V^{-1}SP_{AD}^\perp(v^{-1} - v). \end{aligned}$$

A natural question is how these directions compare with the primal and dual directions that we have seen before, as given

by Equations (1) and (12). Obviously, if $v = e$, that is, if the triple (x, y, s) consists of the μ centers of (P) and (D), then all directions vanish, and therefore are equal, but in general the directions will be different.

We denote the result of a (full) Newton step at the triple (x, y, s) as (x^+, y^+, s^+) . So, we have

$$\begin{aligned} x^+ &= x + \Delta x, \\ y^+ &= y + \Delta y, \\ s^+ &= s + \Delta s. \end{aligned}$$

Since we omitted the quadratic term in the definition of the Newton directions, (x^+, y^+, s^+) will in general not coincide with the triple of μ centers. But, as it is shown in the next section, if (x, y, s) is close enough to $(x(\mu), y(\mu), s(\mu))$ then (x^+, y^+, s^+) will be significantly closer. For the moment we establish a related, but weaker, result, namely, that after a Newton step the duality gap is equal to the duality gap at the μ center.

Lemma 11. *One has $(x^+)^T s^+ = n\mu$.*

Proof. Because of Equations (19) and (21), we have

$$\begin{aligned} x^+ s^+ &= (x + \Delta x)(s + \Delta s) \\ &= xs + (x\Delta s + s\Delta x) + \Delta x\Delta s \\ &= \mu e + \Delta x\Delta s = \mu(e + \dot{x}\dot{s}). \end{aligned}$$

Since the vectors $\dot{x}$ and $\dot{s}$ are orthogonal, we get

$$\begin{aligned} x^{+T} s^+ &= e^T (x^+ s^+) = \mu e^T (e + \dot{x}\dot{s}) \\ &= \mu(e^T e + \dot{x}^T \dot{s}) = \mu n, \end{aligned}$$

proving the lemma.

Primal–Dual Proximity Measure and Quadratic Convergence

In order to analyze the process of taking Newton steps, our next task is to define a suitable proximity measure for the current case. Since a strictly feasible triple (x, y, s) consists of the μ centers of (P) and (D) if

and only if $v = e$, the measure must vanish if and only if $v = e$, which happens if and only if $v^{-1} - v = 0$. Indeed, the norm of $v^{-1} - v$ seems to be a natural candidate for measuring proximity to the μ centers. Therefore, we define

$$\delta(x, s, \mu) := \frac{1}{2} \|v^{-1} - v\|. \quad (26)$$

Recall that $\dot{x}$ and $\dot{s}$ are orthogonal. Hence, from Equation (24), we obtain

$$\delta(x, s, \mu) = \frac{1}{2} \|\dot{x} + \dot{s}\| = \frac{1}{2} \sqrt{\|\dot{x}\|^2 + \|\dot{s}\|^2}.$$

Since $\delta(x, s, \mu)$ depends only on the Hadamard product xs and μ , we feel free to use the notation $\delta(xs, \mu)$ instead.

We proceed by showing that if $\delta(xs, \mu) \leq 1$ (< 1) then after a Newton step the new iterates are (strictly) feasible. This can also be expressed by saying “the Newton step is (strictly) feasible.” For the proofs in this section we need the following technical lemma.

Lemma 12. *If $\delta := \delta(xs, \mu)$, then one has*

- (i) $\|\dot{x}\dot{s}\|_\infty \leq \delta^2$;
- (ii) $\|\dot{x}\dot{s}\| \leq \sqrt{2}\delta^2$.

Proof. We may write

$$\dot{x}\dot{s} = \frac{1}{4}((\dot{x} + \dot{s})^2 - (\dot{x} - \dot{s})^2). \quad (27)$$

From this we derive the componentwise inequality

$$-\frac{1}{4}(\dot{x} - \dot{s})^2 \leq \dot{x}\dot{s} \leq \frac{1}{4}(\dot{x} + \dot{s})^2.$$

This implies

$$-\frac{1}{4}\|\dot{x} - \dot{s}\|^2 e \leq \dot{x}\dot{s} \leq \frac{1}{4}\|\dot{x} + \dot{s}\|^2 e.$$

Since $\dot{x}$ and $\dot{s}$ are orthogonal, the vectors $\dot{x} - \dot{s}$ and $\dot{x} + \dot{s}$ have the same norm. Since this norm equals 2δ the first inequality in the

lemma follows. For the second inequality we derive from Equation (27) that

$$\begin{aligned}\|\dot{x}\dot{s}\|^2 &= e^T(\dot{x}\dot{s})^2 = \frac{1}{16}e^T((\dot{x} + \dot{s})^2 - (\dot{x} - \dot{s})^2)^2 \\ &\leq \frac{1}{16}e^T((\dot{x} + \dot{s})^4 + (\dot{x} - \dot{s})^4).\end{aligned}$$

Since $e^T z^4 \leq \|z\|^4$ for any $z \in \mathbb{R}^n$, we obtain

$$\|\dot{x}\dot{s}\|^2 \leq \frac{1}{16}(\|\dot{x} + \dot{s}\|^4 + \|\dot{x} - \dot{s}\|^4). \quad (28)$$

Using again that $\|\dot{x} - \dot{s}\| = \|\dot{x} + \dot{s}\| = 2\delta$, the second inequality follows.

Lemma 13. *If $\delta(xs, \mu) \leq 1$, then $e + \dot{x}\dot{s} \geq 0$. Moreover, if $\delta(xs, \mu) < 1$, then $e + \dot{x}\dot{s} > 0$.*

Proof. This is an immediate consequence of part (i) of Lemma 12.

Lemma 14. *The Newton step is feasible if $\delta(xs, \mu) \leq 1$ and strictly feasible if $\delta(xs, \mu) < 1$.*

Proof. To prove (strict) feasibility of x^+ it suffices to show that x^+ is nonnegative (positive), and a similar statement for s^+ holds. Using Equation (21) we obtain

$$x^+ = x + \Delta x = x + xv^{-1}\dot{x} = xv^{-1}(v + \dot{x}) \quad (29)$$

$$s^+ = s + \Delta s = s + sv^{-1}\dot{s} = sv^{-1}(v + \dot{s}). \quad (30)$$

Since x , s , and v are positive vectors, it therefore suffices to show that if $\delta(xs, \mu) \leq 1 (< 1)$ then $v + \dot{x}$ and $v + \dot{s}$ are nonnegative (positive). To this end, we introduce a step size α with $0 \leq \alpha \leq 1$, and define

$$v_x^\alpha := v + \alpha\dot{x}$$

$$v_s^\alpha := v + \alpha\dot{s}.$$

Then, we have $v_x^0 = v_s^0 = v$, and $v_x^1 = v + \dot{x}$ and $v_s^1 = v + \dot{s}$. It follows from Equation (24) that

$$\begin{aligned}v_x^\alpha v_s^\alpha &= (v + \alpha\dot{x})(v + \alpha\dot{s}) \\ &= v^2 + \alpha v(\dot{x} + \dot{s}) + \alpha^2 \dot{x}\dot{s} \\ &= (1 - \alpha)v^2 + \alpha e + \alpha^2 \dot{x}\dot{s}.\end{aligned}$$

Now let $\delta(xs, \mu) \leq 1$. By Lemma 13, we then have $\dot{x}\dot{s} \geq -e$. Substitution gives

$$\begin{aligned}v_x^\alpha v_s^\alpha &\geq (1 - \alpha)v^2 + \alpha e - \alpha^2 e \\ &= (1 - \alpha)(v^2 + \alpha e).\end{aligned}$$

If $0 \leq \alpha < 1$, the last vector is positive. So, we have

$$v_x^\alpha v_s^\alpha > 0, \alpha \in [0, 1).$$

Since the entries of v_x^α and v_s^α depend continuously on α , and they are positive for $\alpha = 0$, it follows that they will be positive for $\alpha \in [0, 1)$, and therefore nonnegative for $\alpha \in [0, 1]$. Hence, $v_x^1 = v + \dot{x} \geq 0$ and $v_s^1 = v + \dot{s} \geq 0$, proving the first statement in the lemma.

Finally, if $\delta(xs, \mu) < 1$, then Lemma 15 implies that $\dot{x}\dot{s} > -e$, whence we obtain $v_x^\alpha v_s^\alpha > 0$ for $\alpha \in [0, 1]$. The same continuity argument yields that $v_x^1 = v + \dot{x} > 0$ and $v_s^1 = v + \dot{s} > 0$. Hence, the proof is complete.

We are now ready to prove that the Newton process is quadratically convergent.

Lemma 15. *If $\delta := \delta(xs, \mu) < 1$, then*

$$\delta(x^+ s^+, \mu) \leq \frac{\delta^2}{\sqrt{2(1 - \delta^2)}}.$$

Proof. According to the definition of $\delta(x^+ s^+, \mu)$, we have

$$\begin{aligned}2\delta(x^+ s^+, \mu) &= \|(v^+)^{-1} - v^+\| \\ &= \|(v^+)^{-1}(e - (v^+)^2)\| \\ &\leq \frac{\|e - (v^+)^2\|}{v_{\min}^+},\end{aligned}$$

where $v_{\min}^+$ denotes the smallest entry in v^+ , the (positive!) vector given by

$$v^+ = \sqrt{\frac{x^+ s^+}{\mu}}.$$

Using Equations (29) and (30), and $xs = \mu v^2$, we get

$$\begin{aligned} x^+ s^+ &= xv^{-1}(v + \dot{x}) \cdot sv^{-1}(v + \dot{s}) \\ &= \mu(v + \dot{x})(v + \dot{s}), \end{aligned}$$

whence it follows that $(v^+)^2 = (v + \dot{x})(v + \dot{s})$. Also using Equation (24), we obtain

$$\begin{aligned} (v^+)^2 &= v^2 + v(\dot{x} + \dot{s}) + \dot{x}\dot{s} \\ &= v^2 + v(v^{-1} - v) + \dot{x}\dot{s} \\ &= e + \dot{x}\dot{s}. \end{aligned}$$

This implies that $e - (v^+)^2 = -\dot{x}\dot{s}$ and $v_{\min}^+ \geq \sqrt{1 - \|\dot{x}\dot{s}\|_\infty}$. Substitution gives

$$2\delta(x^+ s^+, \mu) \leq \frac{\|\dot{x}\dot{s}\|}{\sqrt{1 - \|\dot{x}\dot{s}\|_\infty}}.$$

Now applying the bounds in Lemma 12, we obtain

$$2\delta(x^+ s^+, \mu) \leq \frac{\sqrt{2}\delta^2}{\sqrt{1 - \delta^2}},$$

which implies the inequality in the lemma.

As a result we have the following corollary.

Corollary 5. *If $\delta(xs, \mu) < \frac{1}{\sqrt{2}}$, then $\delta(x^+ s^+, \mu) \leq \delta(xs, \mu)^2$.*

Updating the Barrier Parameter

We proceed by considering the effect on the proximity measure of an update of the barrier parameter from μ to $\mu^+ = (1 - \theta)\mu$. Note that contrary to the corresponding results for the primal and dual method (cf. Lemmas 3 and 7), this time we have an equality in the “updating” lemma.

Lemma 16. *Let (x, y, s) be a strictly feasible triple and $\mu > 0$ such that $x^T s = n\mu$. Moreover, let $\delta := \delta(xs; \mu)$ and $\mu^+ = (1 - \theta)\mu$, with $\theta < 1$. Then*

$$\delta(xs; \mu^+)^2 = (1 - \theta)\delta^2 + \frac{\theta^2 n}{4(1 - \theta)}.$$

Proof. Let $\delta^+ := \delta(xs; \mu^+)$ and, as usual, $v = \sqrt{xs/\mu}$. Then, by definition,

$$\begin{aligned} 2(\delta^+)^2 &= \left\| \sqrt{1 - \theta}v^{-1} - \frac{v}{\sqrt{1 - \theta}} \right\|^2 \\ &= \left\| \sqrt{1 - \theta}(v^{-1} - v) + \frac{\theta v}{\sqrt{1 - \theta}} \right\|^2. \end{aligned}$$

From $x^T s = n\mu$, it follows that $\|v\|^2 = n$. Hence, v is orthogonal to $v^{-1} - v$, because

$$v^T(v^{-1} - v) = n - \|v\|^2 = 0.$$

Therefore,

$$4(\delta^+)^2 = (1 - \theta)\|v^{-1} - v\|^2 + \frac{\theta^2 \|v\|^2}{1 - \theta}.$$

Finally, since $\|v^{-1} - v\| = 2\delta$ and $\|v\|^2 = n$, the result follows.

Primal–Dual Interior-Point Method

Assuming that we have strictly feasible triple (x, y, s) and $\mu > 0$ such that $x^T s = n\mu$ and $\delta(xs, \mu) < 1/2$, the following lemma provides a numerical method to follow the central path approximately.

Lemma 17. *Let $x^T s = n\mu$ and $\delta(xs, \mu) \leq 1/2$. If we first update μ to $\mu^+ = (1 - \theta)\mu$, with $\theta = 1/\sqrt{2n}$ and then perform a Newton step, the resulting triple (x^+, y^+, s^+) is such that $(x^+)^T s^+ = n\mu^+$ and $\delta(x^+ s^+, \mu^+) \leq 1/2$.*

Proof. After the barrier parameter update we have, by Lemma 16,

$$\begin{aligned} \delta(xs; \mu^+)^2 &= (1 - \theta)\delta(xs, \mu)^2 + \frac{\theta^2 n}{4(1 - \theta)} \\ &\leq \frac{1 - \theta}{4} + \frac{1}{8(1 - \theta)}. \end{aligned}$$

The right-hand side is convex in θ for $\theta < 1$. It achieves its maximum value $3/8$ on $[0, 1/2]$ at one of the end points of the interval. We conclude from this that $\delta(xs; \mu^+)^2 \leq 3/8$. Since $\delta(xs; \mu^+) \leq \sqrt{3/8} < 1/\sqrt{2}$ we may apply Corollary 5, which gives $\delta(x^+ s^+, \mu^+) \leq \delta(xs; \mu^+)^2$. Hence, we have $\delta(x^+ s^+, \mu^+) \leq \frac{1}{2}$.

By Lemma 11, we also have $(x^+)^T s^+ = n\mu^+$. Hence, the proof is complete.

The primal–dual algorithm can now be stated as shown below.

Algorithm 3. A full-Newton step primal–dual IPM or LO

Input:

(x^0, y^0, s^0) strictly feasible and $\mu^0 > 0$ such that $\delta(x^0 s^0, \mu^0) \leq \frac{1}{2}$ and $(x^0)^T s^0 = n\mu^0$;
accuracy parameter $\varepsilon > 0$;
barrier update parameter $\theta, 0 < \theta < 1$;

begin

$x := x^0; y := y^0; s := s^0; \mu := \mu^0$;

while $x^T s > \varepsilon$

$\mu := (1 - \theta)\mu$;

$(x, y, s) := (x, y, s) + (\Delta x, \Delta y, \Delta s)$;

endwhile

end

We conclude this section by deriving an iteration bound for the algorithm.

Theorem 3. Taking $\theta = \frac{1}{\sqrt{2n}}$, the algorithm requires no more than

$$\sqrt{2n} \log \frac{n\mu^0}{\varepsilon}$$

iterations. The last generated triple (x, y, s) is strictly feasible, whereas $x^T s \leq \varepsilon$.

Proof. By Lemma 17, the triple (x, y, s) will be strictly feasible and $\delta(xs, \mu) \leq \frac{1}{2}$, after each iteration of the algorithm and, moreover, $x^T s = n\mu$. Therefore, the algorithm will stop if $n\mu \leq \varepsilon$. After the k th iteration, we have $\mu = (1 - \theta)^k \mu^0$. Using exactly the same arguments as in the proof of Theorem 1, the iteration bound in the current theorem follows. Owing to the stopping criterion in the algorithm at termination, we have $x^T s \leq \varepsilon$. Hence, the proof is complete.

SPEEDING UP THE ALGORITHMS

The iteration bounds of the primal, dual, and primal–dual algorithms that we presented stem from the reduction of the parameter μ at each iteration by the factor $1 - \theta$. In practical situations, n may be (very) large and then this factor is so close to 1 that the number of iterations that is necessary to

achieve a sufficiently accurate solution will be very large. Owing to the high computational cost per iteration, that is required to compute the search direction, the algorithms may turn out to be far from efficient. In this section we discuss some remedies for the primal–dual algorithm, but similar arguments apply to the primal algorithm and the dual algorithm. We choose the primal–dual algorithm in our discussion; it is the most popular algorithm because the most efficient in many cases.

Adaptive Updating

When implementing the primal–dual IPM presented in the section titled “Primal–Dual Method”, it turns out that Newton’s method is much more accurate than predicted by the theory. As a result, the value of the proximity measure after each iteration is usually much smaller than $\frac{1}{2}$. Note that by Lemma 17 the algorithm remains well defined as long as after each iteration $\delta(xs, \mu) \leq 1/2$ holds. This indicates that larger reductions of the barrier parameter are possible, provided that one takes care that the above condition on the proximity measure remains satisfied. For a detailed discussion of this idea, we refer to Roos *et al.* [5, Section 7.6].

Large-Updates

We may enforce much larger reductions in μ by taking, for example, $\theta = 0.5$ or even $\theta = 0.99$, at each iteration. This method, which is most used in practice, significantly

reduces the number of iterations. Unfortunately, it is then no longer possible to guarantee that the new iterates remain in the vicinity of the center $x(\mu)$, as measured by the distance $\delta(x, \mu)$. Even worse, if $\delta(x, \mu) > 1$, it may happen that the new iterates are no longer feasible (because having negative entries). This may be prevented by not taking the full-Newton step, but use a damping factor that forces the iterates to remain strictly feasible. In this approach, $\delta(x, \mu)$ becomes useless for measuring the distance to the μ center, because we have convergence results only for $\delta(x, \mu) \leq 1$. However, we may take profit of the fact that the Newton direction is a descent direction for the so-called primal-dual logarithmic barrier function. This function is obtained by adding the primal logarithmic barrier function and minus the dual logarithmic barrier function:

$$\begin{aligned} f_{PD}(x, y, s, \mu) \\ &:= f_P(x; \mu) - f_D(y; \mu) \\ &= \frac{c^T x - b^T y}{\mu} - \sum_{i=1}^n \log x_i - \sum_{i=1}^n \log s_i. \end{aligned}$$

Since $f_P(x; \mu)$ is strictly convex and $f_D(y; \mu)$ strictly concave, this function is strictly convex and its unique minimizer on the feasible domain is the triple $(x(\mu), y(\mu), s(\mu))$. So, its minimal value is given by

$$\begin{aligned} f_{PD}(x(\mu), y(\mu), s(\mu), \mu) &= \frac{c^T x(\mu) - b^T y(\mu)}{\mu} \\ &\quad - \sum_{i=1}^n \log x_i(\mu) s_i(\mu) \\ &= n - n \log \mu. \end{aligned}$$

where we used Equation (11) and $x_i(\mu)s_i(\mu) = \mu$ for each i . Hence, defining

$$\begin{aligned} \Psi(x, s, \mu) &:= f_{PD}(x, y, s, \mu) \\ &\quad - f_{PD}(x(\mu), y(\mu), s(\mu), \mu), \end{aligned}$$

and using Lemma 1 we obtain

$$\Psi(x, s, \mu) = \frac{x^T s}{\mu} - \sum_{i=1}^n \log x_i s_i - n + n \log \mu. \quad (31)$$

This barrier function has the triple $(x(\mu), y(\mu), s(\mu))$ as its minimizer on the feasible domain, where its value equals zero.

Assuming $\delta(xs, \mu) \leq 1$, after updating μ to $(1 - \theta)\mu$, we have $\Psi(x, s, \mu) = O(n)$. It has been shown by many authors that by doing a line search along the primal-dual Newton direction, the value of $\Psi(x, s, \mu)$ can be decreased by at least a fixed constant, independent of (x, y, s) and n . The line search will keep the iterates certainly strictly feasible, because at the boundary of the feasible domain $\Psi(x, s, \mu)$ becomes infinite. It follows that only $O(n)$ damped recentering steps will suffice to reach an iterate x in the vicinity of the μ center, for example, for which $\delta(xs, \mu) < 1$. The resulting iteration bound of a suitable variant of this method is $O(n \log(n\mu^0/\varepsilon))$. For a detailed discussion of this method, we refer to Roos *et al.* [5], Vanderbei [6], Wright [7], and Ye [8] and their references.

Despite the worse theoretical iteration bound (by a factor $\sqrt{n}$), the so-called *large-update methods* are in practice much more efficient than the full step methods presented in the previous sections. This phenomenon has been called the *irony of IPMs* [12].

The large-update variant of the primal method is very much in the spirit of the classical barrier method, as presented by Gill *et al.* [13], though these authors did not find a polynomial-time iteration bound for their method.

REFERENCES

1. Todd MJ. Potential-reduction methods in mathematical programming. Math Program 1997;76:3–45.
2. Anstreicher KM. Potential reduction algorithms. In: Interior point methods of mathematical programming. Volume 5. Applied Optimization. Dordrecht: Kluwer Academic Publishers; 1996. pp. 125–158.

3. Gonzaga CC. Path following methods for linear programming. *SIAM Rev* 1992;34(2): 167–227.
4. Kojima M, Megiddo N, Noma T. *et al.* A unified approach to interior point algorithms for linear complementarity problems. Volume 538. Lecture Notes in Computer Science. Berlin, Germany: Springer; 1991.
5. Roos C, Terlaky T, Vial J-P. Theory and algorithms for linear optimization. Chichester, UK: Springer; 2005. [1st ed. Theory and algorithms for linear optimization. An interior-point approach. John Wiley & Sons; 1997.]
6. Vanderbei RJ. Linear programming: foundations and extensions. Boston (MA): Kluwer Academic Publishers; 1996.
7. Wright SJ. Primal-dual interior-point methods. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM); 1997.
8. Ye Y. Interior point algorithms: theory and analysis. Wiley-Interscience series in discrete mathematics and optimization. New York: John Wiley & Sons Inc.; 1997.
9. Fiacco AV, McCormick GP. Nonlinear programming. Sequential unconstrained minimization techniques. New York: John Wiley & Sons Inc.; 1968. [Reprint: Volume 4. SIAM classics in applied mathematics. Philadelphia (PA): SIAM Publications; 1990.]
10. Frisch R. The logarithmic potential method for solving linear programming problems. [Memorandum]. Oslo: University Institute of Economics; 1955.
11. Anstreicher KM. On long step path following and SUMT for linear and quadratic programming. *SIAM J Optim* 1996;6(1):33–46.
12. Renegar J. A mathematical view of interior-point methods in convex optimization. MPS/SIAM series on optimization. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM); 2001.
13. Gill PE, Murray W, Saunders MA, *et al.* On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method. *Math Program* 1986;36:183–209.

CHILEAN INSTITUTE OF OPERATIONS RESEARCH

JUAN GUTIERREZ
Department of Industrial
Engineering, Universidad de
Santiago de Chile, Santiago,
Chile

VICTOR PARADA
Department of Computer Science,
Universidad de Santiago de
Chile, Santiago, Chile

RICHARD WEBER
Department of Industrial
Engineering, Universidad de
Chile, Santiago, Chile
Statistical Laboratory, University
of Cambridge, Cambridge, UK

HISTORY OF THE CHILEAN INSTITUTE OF OPERATIONS RESEARCH

In 1966 ENDESA's¹ CEO Renato Salazar together with managers from other important companies or institutions, such as IANSA, CAP, ENAP, and CIENES invited professionals working with OR models to a meeting where they wanted to share experiences related to mathematical models for decision making in Chilean industry. Participants at this meeting agreed on the creation of a society promoting applied OR models; ICHIO, the Chilean Institute of Operations Research was born.

In 1967, after having organized its first congress, ICHIO's first president, secretary, and board of directors were elected as shown in the following table (Table 1).

This team worked on ICHIO's legal structure which was presented to the Ministry of Justice on April 30, 1968. In 1975, ICHIO

was accepted as member society of the International Federation of Operations Research (IFORS).

Today, ICHIO is governed by its president, past president, secretary, treasurer, and two directors. Its approximately 40 individual members come mainly from 15 Chilean universities.

ACCOMPLISHMENTS

ICHIO serves as a facilitator for the development of OR in Chile. Its meetings gather Chilean researchers working in related areas. ICHIO's journal has published mostly scientific articles, describing latest developments from operations research (OR) and neighboring fields. It has undergone a major redesign and will be published from 2010 on as an electronic journal.

In 2008, the first student chapter within ICHIO was created by students from the Universidad de Concepción. We plan to have such chapters in the main Chilean cities where OR is taught. The main goal of these chapters is to attract students interested in OR-related topics to ICHIO's activities at an early stage of their university education.

In 2009, ICHIO launched a new version of its web site (www.ichio.cl) with interactive material regarding different topics of OR, free for public use. Each such area is maintained by its editor who is responsible for the respective content.

MEETINGS OF THE CHILEAN INSTITUTE OF OPERATIONS RESEARCH

The main meeting hosted by ICHIO is its biannual congress Optima with approximately 150 to 200 participants. At these events usually there are plenary and semiplenary presentations, regular sessions, and poster presentations. Participants come mainly from Chile but also from several

¹All abbreviations in this text are expanded in the appendix.

Table 1. ICHIO's First President, Secretary, and Board of Directors

President	Oscar Barros, Department of Industrial Engineering, Universidad de Chile
Secretary	Andrés Weintraub, Department of Industrial Engineering, Universidad de Chile
Directors	Sergio Álvarez, CAP S.A Enrique Cansado, CIENES Juan Gutiérrez de, IANSA S.A Hernán Santa María, Pontificia Universidad Católica de Chile Roberto Fuenzalida, Empresas El Mercurio

Latin-American countries, and in some cases from northern America, Europe, or Australasia. The beauty of the Chilean landscape makes it particularly attractive for foreign scientists to combine the congress with some days of vacations.

As ICHIO's strategy, since 1995 its congresses generally take place outside the capital Santiago where most of the scientific activities are concentrated at the major universities. This way a better dissemination of OR to remote places of this highly centralized and densely populated country is achieved.

Consequently, the Optima congresses are particularly interesting for local faculty members, professionals, and students from engineering, mathematics, computer science, and related areas. The following table (Table 2) displays information about the series of past Optima congresses.

The congress Optima 2011 will take place in October 2011 in *Pucon* organized by the *Universidad de la Frontera*.

The Latin-American operations research society (ALIO) organizes summer schools mainly for graduate students from OR and related fields. Chile hosted these summer schools twice: in 1994 in *Punta de Talca* and in 2001 in *Viña del Mar*.

Approximately twice a year, ICHIO organizes its local colloquia with usually two academic talks open to the general public and a subsequent meeting of its members. These talks are particularly interesting for students and faculty members at the host university.

Table 2. Optima Congresses Hosted by ICHIO (between 1968 and 1994 the Optima Congresses were organized together with the "Taller de Ingeniere Sistemas" at the Universidad de Chile in Santiago)

Place	Organized By	Year
Santiago	Universidad de Chile	1967
Santiago	Universidad de Chile	1995
Concepción	Universidad de Concepción	1997
Arica	Universidad de Tarapacá	1999
Curicó	Universidad de Talca	2001
Valparaíso	Universidad Técnica Federico Santa María	2003
Valdivia	Universidad Austral	2005
Puerto Montt	Universidad de Los Lagos	2007
Termas de Chillan	Universidad del Bio Bio	2009

SUCCESSFUL APPLICATIONS

During the first Optima Congress organized by ICHIO in 1967 the following applications were presented:

- "Planning model for expanding the beet sugar industry in Chile" by Professor Juan Gutierrez
- "Logistic models for material handling in the Chilean steel industry" by Sergio Alvarez
- "Optimizing process in the Chilean oil refinery industry" by George Nascimento
- "OR: A necessary discipline for engineering education" by Dr. Enrique Cansado.

From then on, ICHIO members developed and implemented many successful

applications in industry and government. The following list of papers from the journal "Interfaces" gives an idea of the impact OR had in Chile in a broad sense. In each case we list title, reference, and a shortened version of the abstract as it appears in the paper.

- *Competitive Cost Dynamics* [1]. This is the first of three articles about some popular tools that have been widely used since the early 1970s to support strategic decision making.
- *Growth–Share Matrix in Strategic Planning* [2]. In the second of a series of three tutorials, the methodology and strategic implications of the portfolio business matrix are analyzed and illustrations given of the use of the growth–share matrix.
- *Industry Attractiveness–Business Strength Matrix in Strategic Planning* [3]. In the third and final paper in a series of tutorials on strategic planning, the industry attractiveness–business strength matrix is presented with a detailed procedure for its application, variety of examples, and a discussion of its strengths and weaknesses.
- *Corporate Strategic Planning Process* [4]. Described is an approach to develop a formal corporate strategic planning process in a business firm.
- *Concept of Strategy and the Strategy Formation Process* [5]. The concept of strategy is presented as a normative model that has validity for all firms.
- *Truck Scheduling* [6]. An operative and computerized system, ASICAM, based on a simulation process with heuristic rules, to support daily truck scheduling decisions has been developed.
- *OR Models and the Management of Agricultural and Forestry Resources* [7]. Over the last two decades, forest land management practices have changed in response to ecological issues and the need to improve efficiency to remain competitive in emerging global markets. Decision processes have become more open and complex as information and communication technologies change. The OR/MS community is meeting these challenges by developing new modeling strategies, algorithms, and solution procedures that address spatial requirements, multiresource planning, hierarchical systems, multiple objectives, and uncertainty as they pertain to both industrial timberlands and public forests.
- *Optimal Investment Policies for Pollution Control in the Copper Industry* [8]. A decision support system for investment projects in pollution abatement plants at state-owned copper smelters in Chile has been developed. A mixed-integer linear model was formulated to determine the optimal investment policy for this interrelated production–environmental problem.
- *Strategic Valuation of Investment under Competition* [9]. The most serious problem with the widely used discounted cash flow (DCF) methods of investment valuation is that they are commonly applied without explicit regard to competition. As a result, the prescriptions they afford are often inconsistent with those given by competitive strategy frameworks. To remedy this, such frameworks have been integrated with DCF methods to value investments (and disinvestments) in competitive and uncertain contexts.
- *OR Systems in the Chilean Forest Industries* [10]. The Chilean forestry sector is composed of private firms that combine large timber-land holdings of mostly pine plantations and some eucalyptus with sawmills and pulp plants. Since 1988, to compete in the world market, the main Chilean forest firms, which have sales of about \$1 billion, have started implementing OR models developed jointly with academics from the University of Chile. These systems support decisions on daily truck scheduling, short-term harvesting, location of harvesting machinery and access roads, and medium- and long-term forest planning. Approaches used in solving these complex problems include simulation,

linear programming with column generation, mixed-integer LP formulations, and heuristic methods. The systems have led to a change in organizational decision making and to estimated gains of at least US \$20 million per year.

- *Production Planning Decision Support System (DSS) [11]*. An optimization-based decision support systems (OBDSSs) to support production planning at CTI (an appliance-manufacturing firm) using new methodologies and tools based on structured modeling has been developed and implemented.
- *Assignment of Catering Contracts [12]*. Chile's school system is using mathematical modeling to assign catering contracts in a single round sealed-bid combinational auction.
- *Mathematical Programming Combined with Expert Systems Improves Customer Service [13]*. Determining what fertilizer mix to apply to certain soils is complicated and time consuming. A hybrid tool using expert system technology and mixed-integer linear programming models that increased fertilizer sales has been proposed.
- *Programming of the Copper Smelting Process [14]*. A model to maximize production while respecting metallurgic, operational, environmental, and other restrictions has been developed. Two methods of planning daily activities for the plant were proposed.
- *OR Models and the Management of Agricultural and Forestry Resources [15]*. OR has helped people to understand and manage agricultural and forestry resources during the last 40 years. Its use to assess the past performance of OR models in this field and to highlight current problems and future directions of research and applications has been analyzed.
- *Mixed-Integer Programming Models for Sports Scheduling [16]*. Since 2005, Chile's professional soccer league has used a game-scheduling system that is based on an integer linear programming model. The Chilean league managers considered several operational, economic, and sporting criteria for the final tournaments' scheduling. Thus, they created a highly constrained problem that had been, in practice, unsolvable using their previous methodology. This led to the adoption of a model that used some techniques that were new in soccer league sports scheduling. The schedules they generated provided the teams with benefits such as lower costs, higher incomes, and fairer seasons. In addition, the tournaments were more attractive to sports fans. The success of the new scheduling system has completely fulfilled the expectations of the "Asociación Nacional de Fútbol Profesional" (ANFP), the organization for Chilean professional soccer.
- *Operations Research in Mine Planning [17]*. Applications of operations research to mine planning date back to the 1960s. Since that time, optimization and simulation, in particular, have been applied to both surface and underground mine planning problems, including mine design, long- and short-term production scheduling, equipment selection, and dispatching, inter alia. This article reviews several decades of such literature with a particular emphasis on more recent work, suggestions for emerging areas, and highlights of successful industrial applications.
- *Course Scheduling System for Executive Education [18]*. Each October, the Executive Education Unit at the Universidad de Chile develops its course schedules for the following year. By 2008, the complexities of increasing enrollments and course offerings had rendered its manual timetabling process unmanageable. This article presents an automated computational system that generates optimal timetables and classroom assignments for all the unit's courses, minimizing both operating costs and schedule conflicts.

FUTURE CHALLENGES

According to the most recent study performed following the Knowledge Assessment Methodology developed by the World Bank (www.worldbank.org/kam), Chile ranked 43 in Innovation and 47 in Education. There is a clear need to improve in those sectors if Chile wants to increase its productivity and take the step to become a developed country. Advanced OR and its applications play a crucial role on this road and ICHIO will contribute to reach this goal by dissemination of OR techniques and applications.

For the period 2010–2015 ICHIO has the following goals.

- ICHIO will try to increase membership. By mid-2009 ICHIO had about 50 individual members. At the biannual congress Optima 2009 membership of students as well as organizations was introduced. By the end of 2015 we want to have at least 50 student members, 100 individual professional members, and 5 corporate members.
- ICHIO will improve its links with industry since more fluent communication between academia and industry is necessary in order to inject OR into practice on the one hand. On the other hand, this integration helps to better identify real problems that need advanced solutions. The aforementioned corporate membership is one step in this direction. We will also have colloquia at companies and increase practitioners' participation at our Optima congresses.
- ICHIO will establish exchange agreements with its sister societies in Latin-American countries. A first contract with EPIO from Argentina has been signed recently in 2009; the agreement with the Brazilian OR society SOBRAPO is currently under development and will be signed in 2010. This kind of agreement will improve dissemination of research results via increased participation in the respective conferences and journals.
- Along the lines of the previous item, it is intended to have joint conferences

hosted by ICHIO and some affiliated societies.

- The journal published by ICHIO has been redesigned and will appear from 2010 on as an electronic journal. It is our goal to get at least an indexation in SCIELO by 2015.

Acknowledgments

Many ICHIO members have contributed to the developments described in this short article. They deserve special thanks. Support from the Chilean "*Instituto Sistemas Complejos de Ingeniería*" (ICM: P-05-004-F, CONICYT: FBO16) is gratefully acknowledged; (www.sistemasdeingenieria.cl).

APPENDIX

List of Abbreviations

ALIO	Asociación Latino-Iberoamericana de Investigación Operativa (http://www-2.dc.uba.ar/alio/)
CAP	Compañía de Acero del Pacco S.A. (www.cap.cl)
CIENES	Centro Interamericano de Enseñanza de Estadística
ENAP	Empresa Nacional del Petróleo (www.enap.cl)
ENDESA	Empresa Nacional de Electricidad S.A. (www.endesa.cl)
EPIO	Escuela de Perfeccionamiento en Investigación Operativa (www.epio.org.ar)
IANSA	Industria Azucarera Nacional Sociedad Anónima (www.iansa.cl)
ICHIO	Instituto Chileno de Investigación Operativa (www.ichio.cl)
IFORS	International Federation of Operational Research Societies (www.ifors.org)
SOBRAPO	Sociedade Brasileira de Pesquisa Operacional (www.sobrapo.org.br)

REFERENCES

1. Hax AC, Majluf NS. Competitive cost dynamics: the experience curve. *Interfaces* 1982; 12:50–61.

2. Hax AC, Majluf NS. The use of the growth-share matrix in strategic planning. *Interfaces* 1983;13:46–60.
3. Hax AC, Majluf NS. The use of the industry attractiveness-business strength matrix in strategic planning. *Interfaces* 1983; 1983:54–71.
4. Hax AC, Majluf NS. The corporate strategic planning process. *Interfaces* 1984;14:47–60.
5. Hax AC, Majluf NS. The concept of strategy and the strategy formation process. *Interfaces* 1988;18:99–109.
6. Weintraub A, Epstein R, Morales R, *et al.* A truck scheduling system improves efficiency in the forest industries. *Interfaces* 1996;26:1–12.
7. Weintraub A, Bare BB. New issues in forest land management from an operations research perspective. *Interfaces* 1996; 26:9–25.
8. Mondschein SV, Schilkrot A. Optimal investment policies for pollution control in the copper industry. *Interfaces* 1997;27:69–87.
9. Del Sol P, Ghemawat P. Strategic valuation of investment under competition. *Interfaces* 1999;29:42–56.
10. Epstein R, Morales R, Serón J, Weintraub A. Use of or systems in the Chilean forest industries. *Interfaces* 1999;29:7–29.
11. Gazmuri P, Maturana S. Developing and implementing a production planning DSS for CTI using structured modeling. *Interfaces* 2001;31:22–36.
12. Epstein R, Henríquez L, Catalán J, *et al.* A combinatorial auction improves school meals in Chile. *Interfaces* 2002;32:1–14.
13. Angel AM, Taladriz LA, Weber R. Soquimich uses a system based on mixed-integer linear programming and expert systems to improve customer service. *Interfaces* 2003;33:41–52.
14. Pradenas L, Zúñiga J, Parada V. CODELCO, Chile programs its copper smelting operations. *Interfaces* 2006;36:296–301.
15. Weintraub A, Romero C. Operations research models and the management of agricultural and forestry resources: a review and comparison. *Interfaces* 2006;36:446–457.
16. Durán G, Guajardo M, Miranda J, *et al.* Scheduling the Chilean soccer league by integer programming. *Interfaces* 2007; 37:539–552.
17. Newman AM, Rubio E, Caro R, *et al.* A review of operations research in mine planning. *Interfaces* 2010. In press.
18. eClasSkeduler MJ. A course scheduling system for the executive education unit at the Universidad de Chile. *Interfaces* 2010. In press.

CHINESE POSTMAN PROBLEM

WANYAN YU

RAJAN BATTA

Department of Industrial and
Systems Engineering,
University at Buffalo (State
University of New York),
Buffalo, New York

CHINESE POSTMAN PROBLEM BASICS

Chinese postman problem (known as *CPP*) is a branch of arc routing problems. It was first studied by the Chinese mathematician Meigu Guan (or Kwan Mei-ko) when he worked as a post office worker during the Chinese cultural revolution. Guan stated: "A mailman has to cover his assigned segment before returning to the post office. The problem is to find the shortest walking distance for the mailman." [1]. In general, the CPP is to determine a closed walk of minimum cost or length covering each arc at least once. The earliest origin of CPP is the Königsberg bridge problem (see Fig. 1), which is to determine whether there exists a closed walk traversing each of the seven bridges exactly once. Leonhard Euler [2] solved the problem in 1736 by finding a necessary and sufficient condition for the existence of such a walk on any connected undirected graph, and proved that there is no solution for this particular problem. A graph with the same property is therefore called *Eulerian graph* and the closed path found on the graph is called *Eulerian tour* (Eulerian circuit or Eulerian cycle). A formal definition of an Eulerian graph is as follows: an Eulerian graph indicates that there exists a closed path on the graph that contains each arc exactly once and each vertex at least once. We note that the definition of an Eulerian graph is not restricted to a connected undirected graph.

The relationship between a CPP solution and an Eulerian tour is explained below. If

a graph is an Eulerian graph, the solution of a CPP (named *CP tour*) for the graph is simply the Eulerian tour. Otherwise, some edges must be traversed more than once in order to cover the graph, and the CP tour is the shortest among these edge-covering tours. Therefore, if a graph can be revised to an Eulerian graph by adding a set of arcs with minimum cost (called a *subproblem of augmentation*), then the solution of a CPP for the original graph can be found by finding an Eulerian tour on the revised graph. The scope of this article includes a discussion of CPP on different types of graphs (e.g., directed graphs and mixed graphs), and we will examine how CPP can be solved by converting a graph to a Eulerian graph for each graph type.

As mentioned above, algorithms for the CPP include two stages. The first stage is to find the least-cost set of repeated arcs or edges which makes the graph Eulerian. The second stage is to find the actual sequence of the traversed arcs or edges. It is proved that given the set found in stage one, the traversing sequence can be determined in polynomial time [3]. Therefore, the computational complexity depends on the subproblem of augmentation in the first stage.

Most applications of CPP relate to arc routing problems, such as mail delivery, garbage collection, snow plows, highway lawnmowers, and school bus routing [4]. Other than conventional routing applications, there are several other problems that have a similar mathematical model to that of the CPP. Examples of these problems include topological testing of computer networks, design of VLSI (very-large-scale-integration) tours of integrated circuit to minimize the number of ways of connecting different layers [5], analysis of DNA [6], and transmission line inspections.

The next section provides a classification for CPP and presents the available mathematical models and exact algorithms or heuristics for each class of problems. We

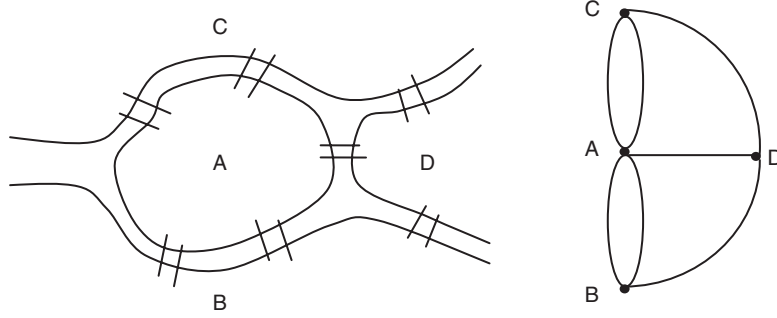


Figure 1. The seven bridges of Königsberg.

note that this classification is similar to that found in the paper by Eiselt *et al.* [3].

CHINESE POSTMAN PROBLEM CLASSES

Based on the type of graphs it solves, CPP can be classified into the following five types: the undirected Chinese postman problem (UCPP), the directed Chinese postman problem (DCPP), the mixed Chinese postman problem (MCP), the windy postman problem (WPP), and the hierarchical CPP (HPP). For instance, MCP solves CPP on mixed graphs that include both undirected and directed arcs. WPP refers to the problem of undirected graphs, while the cost of traversing an edge is different for the two directions of travel. The last type is hierarchical CPP, where a precedence relation constraint is enforced to the set of arcs. This precedence relation defines either a partial or a complete order of the arcs needed to be traversed. In the literature, the UCPP and DCPP have been well studied and solved in polynomial time. However, the other types of CPP are proved to be NP-hard problems and are still relatively unexplored.

As a general representation of graphs, let $G = (V, A)$ be a connected graph without loops, where $V = \{v_1, \dots, v_n\}$ is the vertex set and $A = \{(v_i, v_j) : v_i, v_j \in V \text{ and } i \neq j\}$ is the arc set. If a graph is undirected, the arc set is equivalent to an edge set. For every arc, there is a nonnegative cost (distance, or length) c_{ij} associated with it, where $c_{ij} = \infty$ if the arc is not defined. These five problems

are discussed in the following subsections, respectively.

Undirected Chinese Postman Problem (UCPP)

The CPP defined on an undirected graph is classified as an undirected CPP (UCPP). The Eulerian property is central to the solution of the UCPP, where an Eulerian graph indicates that there exists a closed path in G that contains each arc exactly once.

Guan [1] observed that G always has an even number of odd-degree vertices and that an Eulerian graph G' can be derived from G by adding edges to link odd-degree vertices. Guan proved that the necessary and sufficient condition for the optimality of an Eulerian tour on G' is that there is no redundancy, that is, a CP tour will never utilize an edge more than twice. Therefore, the length of the CP tour will not exceed twice of the length of graph G .

The UCPP is usually formulated as an integer programming model in the first stage and the objective is to determine a minimum increase of cost of G into G' such that all vertices of G' have an even degree. Let x_{ij} be the number of additional copies of (v_i, v_j) in the tour. Let $\delta(i)$ be the set of edges incident to v_i , and let $T \subseteq V$ be the set of odd-degree vertices of V . Then, the formulation for UCPP is given below.

UCPP Model 1

$$\text{Minimize } \sum_{(v_i, v_j) \in A} c_{ij} x_{ij} \quad (1)$$

$$\begin{aligned}
&\text{subject to } \sum_{(v_i, v_j) \in \delta(i)} x_{ij} \\
&= \begin{cases} 1(\text{mod } 2) & \text{if } v_i \in T \\ 0(\text{mod } 2) & \text{if } v_i \in V \setminus T \end{cases} \\
&\quad (2)
\end{aligned}$$

$$x_{ij} \in \{0, 1\} \quad ((v_i, v_j) \in A). \quad (3)$$

This model can be solved as a minimum weight perfect matching problem over the odd-degree vertex set T . Graph G' with the shortest length of tour is then obtained by adding edges in the matching to G . The matching problem is solved by Lawler [7] using an $O(|V|^3)$ algorithm. More results with less computational complexity can be found in Refs 8 and 9.

An equivalent model of UCPP is proposed by Edmonds and Johnson [10]. A set of edges $E(S) = \{(v_i, v_j) : v_i \in S, v_j \in V \setminus S \text{ or } v_i \in V \setminus S, v_j \in S\}$ is defined such that any edge in the set meets one vertex in S and one vertex not in S , where S is a nonempty proper subset of V containing an odd number of odd-degree vertices. Then, the model is given below.

UCPP Model 2

$$\text{Minimize } \sum_{(v_i, v_j) \in A} c_{ij} x_{ij} \quad (4)$$

$$\begin{aligned}
&\text{subject to } \sum_{(v_i, v_j) \in E(S)} x_{ij} \geq 1 \\
&\quad (S \subset V, S \text{ odd}) \quad (5)
\end{aligned}$$

$$x_{ij} \geq 0 \quad ((v_i, v_j) \in A) \quad (6)$$

$$x_{ij} \text{ integer } ((v_i, v_j) \in A). \quad (7)$$

Constraint (5), known as *blossom inequalities*, are defined for every odd set S . It is equivalent to the statement that each odd-degree vertex must be made even by adding edges incident to it. Edmonds and Johnson solved this model by adapting Edmonds' blossom algorithm for matching problems.

Two algorithms for finding the Eulerian tour after obtaining the Eulerian

graph G' are presented below. The first algorithm is referred to as *Fleury's algorithm*, which dates to 1883. The second algorithm is developed by Edmonds and Johnson in Ref. 10.

Fleury's Algorithm.

Step 0. Select an arbitrary vertex v_i .

Step 1. Select an edge (v_i, v_j) such that the deletion of this edge does not disconnect the graph. Then delete edge (v_i, v_j) from the graph. Set $v_i = v_j$.

Step 2. Repeat step 1 until all edges have been deleted. These deleted edges in sequence form an Eulerian tour.

End-Pairing Algorithm.

Step 0. Trace out a simple tour which may not include all edges. If the tour includes all edges, stop and declare it to be Eulerian.

Step 1. Begin at any vertex v on the tour incident to edges not in the tour and complete the second simple tour not including any edge on the first tour.

Step 2. Inject the second tour into the first tour by swapping the edge pairing in the two tours. That is, if e_0 is the first edge leaving vertex v and e_L is the last edge entering vertex v , then for any edge pair (e_1, e_2) of the first tour meeting v , swap the edge pairings by replacing (e_1, e_2) by (e_1, e_0) and (e_L, e_2) . If all edges are included in the newly formed tour, stop. Otherwise, go to step 1.

Fleury's algorithm is straightforward but time consuming. However, the algorithm proposed by Edmonds and Johnson improves the computational complexity to $O(|V|)$. Readers are referred to [11] for additional algorithms for finding Eulerian tours on an Eulerian graph.

Directed Chinese Postman Problem (DCPP)

The CPP defined on a directed graph is classified into this type. Note that different from the UCPP that always has a solution, the

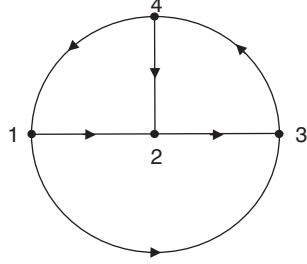


Figure 2. A strongly connected graph.

directed graph must be strongly connected for the existence of a solution to the DCP. An example of a strongly connected graph is shown in Figure 2, where there is a path from each vertex on the graph to every other vertex. In a directed graph, a vertex is called *unbalanced* if the number of arcs coming into it does not equal that going out from it. For example, all the vertices in Fig. 2 are unbalanced. Define I as the set of unbalanced vertices v_i with an excess of s_i incoming arcs and J the set of unbalanced vertices v_j with an excess of d_j outgoing arcs. Let c_{ij} be the length of the shortest path between v_i and v_j . Edmonds and Johnson [10], Orloff [12], and Beltrami and Bodin [13] showed that this problem can be modeled as a least-cost transportation problem where the flow on each arc has to be at least 1. The formulation is given below.

DCPP Model

$$\text{Minimize } \sum_{v_i \in I} \sum_{v_j \in J} c_{ij} x_{ij} \quad (8)$$

$$\text{subject to } \sum_{v_j \in J} x_{ij} = s_i \quad (v_i \in I) \quad (9)$$

$$\sum_{v_i \in I} x_{ij} = d_j \quad (v_j \in J) \quad (10)$$

$$x_{ij} \geq 0 \quad (v_i \in I, v_j \in J). \quad (11)$$

By adding arcs represented by solution x_{ij} to the above model, the original graph G is transferred to an Eulerian graph G' . Fleury's algorithm can then be adapted to find an Eulerian tour on G' . Another

popular algorithm to solve the Eulerian tour problem for directed graph is provided by Aardenne-Ehrenfest and de Bruijn [14]. The algorithm is given below.

van Aardenne-Ehrenfest and de Bruijn's Algorithm [14].

Step 0. Find a spanning arborescence of G rooted at any vertex v_r , where an arborescence is a directed, rooted tree such that all the edges are directed away from the root.

Step 1. At any vertex v_i except for the root v_r of the arborescence, specify any order and label the arcs directed away from v_i so long as the arc used in the arborescence is last in the ordering. For the root vertex v_r , specify any order and label the arcs directed away from it.

Step 2. Obtain an Eulerian tour by following the ordered arc from an arbitrary vertex, that is, whenever a vertex is entered, it is left through the arc not yet traversed according to an ascending order of labels. Stop until all the arcs have been included in the tour.

The proof of the algorithm providing an Eulerian tour is presented by Edmonds and Johnson in Ref. 10.

The Mixed Chinese Postman Problem (MCP)

If the CPP is defined on a graph containing both directed arcs and undirected edges, this type of CPP is called the *MCP*. Similar to previous CPP, solving MCP involves finding a minimum-cost augmentation of graph G to satisfy the necessary and sufficient conditions of an Eulerian graph, and then to determine an Eulerian path. Here, G needs to be a strongly connected graph represented as $G = (V, A \cup E)$, where A represents an arc set and E represents an edge set.

Some terminologies need to be defined over mixed graphs before presentation of MCP. A mixed graph is *even* if the total number of arcs and edges to any vertices on it is even. It is *symmetric* when any vertex has

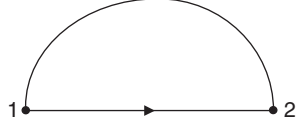


Figure 3. Example of a graph that is not symmetric but Eulerian [3].

an equal number of incoming arcs and outgoing arcs. It is *balanced* if the balanced set conditions are satisfied [15], that is, for every $S \subseteq V$, the difference between the number of directed arcs from S to $V \setminus S$ and the number of directed arcs from $V \setminus S$ to S must be less than or equal to the number of undirected arcs joining S and $V \setminus S$. A mixed graph is balanced if it is even and symmetric. A mixed graph is Eulerian if and only if it is even and balanced, or if it is even and symmetric. However, it is noted that the symmetry is not a necessary condition for a directed mixed graph being Eulerian as described above; see Fig. 3 for an example. If a given graph G is Eulerian, the problem now becomes determination of an Eulerian tour on G . Three steps are available to achieve this goal. The first step is to assign directions to some edges in order to make G symmetric. The procedure was proposed by Ford and Fulkerson [15] to transform a mixed graph to a symmetric graph. Assigning directions to the remaining edges is completed in the second step followed by the procedure described in Eiselt *et al.* [3]. Finally, the third step is to determine an actual Eulerian tour on G by applying the van Aardenne-Ehrenfest and de Bruijn algorithm [14]. It should be noticed that it is not always feasible to determine a minimum-cost augmentation in order to solve the MCPP since some graphs cannot become Eulerian [3].

One of the formulations for MCPP is stated by Christofides *et al.* [16]. The notation needed is as follows: let (i) $A_k^+ = \{(v_i, v_j) \in A : v_i = v_k\}$, $A_k^- = \{(v_i, v_j) \in A : v_j = v_k\}$, and V_k be the set of all vertices linked to v_k by an edge; (ii) x_{ij} be the number of extra times arc (v_i, v_j) is traversed in the optimal solution; (iii) y_{ij} be the total number of times edge (v_i, v_j) is traversed from v_i to v_j ; (iv) p_k be a binary constant that equals 1 if and only if the degree of vertex v_k is odd, and 0,

otherwise; and (v) z_k be an integer variable. Then the formulation is given below.

MCPP Model 1

Minimize

$$\sum_{(v_i, v_j) \in A} c_{ij}(1 + x_{ij}) + \sum_{(v_i, v_j) \in E} c_{ij}(y_{ij} + y_{ji}) \quad (12)$$

subject to

$$\begin{aligned} & \sum_{(v_i, v_j) \in A_k^+} (1 + x_{ij}) - \sum_{(v_i, v_j) \in A_k^-} (1 + x_{ji}) \\ & + \sum_{v_j \in V_k} y_{kj} - \sum_{v_j \in V_k} y_{jk} = 0 \quad (v_k \in V) \end{aligned} \quad (13)$$

$$\begin{aligned} & \sum_{(v_i, v_j) \in A_k^+} x_{ij} + \sum_{(v_i, v_j) \in A_k^-} x_{ji} \\ & + \sum_{v_j \in V_k} (y_{kj} + y_{jk} - 1) \\ & = 2z_k + p_k \quad (v_k \in V) \end{aligned} \quad (14)$$

$$y_{ij} + y_{ji} \geq 1 \quad ((v_i, v_j) \in E) \quad (15)$$

$$z_k, x_{ij}, y_{ij}, y_{ji} \geq 0 \text{ and integer.} \quad (16)$$

An enumerative algorithm was used to solve this problem, where two different lower bounds are calculated at each node in the search tree. One is obtained by solving the minimum-cost perfect matching problem via the Lagrangian relaxation of the first constraint (13) in the above formulation, and the other is obtained by solving the minimum-cost flow problem via the Lagrangian relaxation of the third constraint (15).

Another formulation and algorithm of the MCPP were found in Nobert and Picard [17] in which they used only one variable y_{ij} for each edge of E to represent the number of copies of edge (v_i, v_j) added to the graph to make it Eulerian. For any subset S of V , following sets are defined:

$$\begin{aligned} A^+(S) &= \{(v_i, v_j) \in A : v_i \in S, v_j \in V \setminus S\}, \\ A^-(S) &= \{(v_i, v_j) \in A : v_i \in V \setminus S, v_j \in S\}, \end{aligned}$$

$$E(S) = \{(v_i, v_j) \in E : v_i \in S, v_j \in V \setminus S \text{ or } v_i \in V \setminus S, v_j \in S\}. \quad (17)$$

Furthermore, let $u(S) = |A^+(S)| - |A^-(S)| - |E(S)|$. Then the formulation is given below.

MCPM Model 2

$$\text{Minimize} \quad \sum_{(v_i, v_j) \in A} c_{ij} x_{ij} + \sum_{(v_i, v_j) \in E} c_{ij} y_{ij} \quad (18)$$

$$\text{subject to} \quad \sum_{(v_i, v_j) \in A} x_{ij} + \sum_{(v_i, v_j) \in E} y_{ij} = 2z_k + p_k (v_k \in V) \quad (19)$$

$$\begin{aligned} & - \sum_{(v_i, v_j) \in A^+(S)} x_{ij} + \sum_{(v_i, v_j) \in A^-(S)} x_{ij} \\ & + \sum_{(v_i, v_j) \in E(S)} y_{ij} \geq u(S) \\ & (S \subset V, S \neq \emptyset) \end{aligned} \quad (20)$$

$$\begin{aligned} & \sum_{(v_i, v_j) \in A^+(S)} x_{ij} + \sum_{(v_i, v_j) \in A^-(S)} x_{ij} \\ & + \sum_{(v_i, v_j) \in E(S)} y_{ij} \geq 1 (S \subset V, S \text{ odd}) \end{aligned} \quad (21)$$

$$z_k, x_{ij}, y_{ij} \geq 0 \text{ and integer.} \quad (22)$$

In this formulation, the second constraints (20) make sure that all nonempty proper subsets S of V become balanced and the third constraints (21) are a generalized form of blossom inequalities. An algorithm to solve this formulation was provided by Nobert and Picard [17]. Initially, the problem includes all nonnegativity constraints, balanced set constraints (Eqs 19 and 20), and generalized blossom inequalities (Eqs 20 and 21). The algorithm is preceded by generating additional balanced set and generalized blossom inequalities that are found to be violated. A number of Gomory cuts are also added to encourage integrality. If the solution is an integer and satisfies all

constraints, a minimum-cost Eulerian graph is identified and the procedure terminates. Otherwise, a branching process is to be initiated.

The Windy Chinese Postman Problem (WPP)

The WPP was first introduced by Minieka [18]. The feature of the WPP is that the cost of traversing an edge depends on the direction of travel: in one direction, the postman can walk with the wind resulting in a lower cost, whereas in the other direction, he must walk against the wind with more cost. It can be seen that UCPP, DCPM, and MCPM are special cases of WPP by using an appropriate definition of edge cost. For example, DCPM can be modeled by an infinite cost of travel on opposite directions of the arcs. An example of cost depending on the direction of travel is the different standby air fares between New York and London depending upon direction [18]. Previous studies [19] have shown that the WPP is an NP-hard problem. However, it has been proved that it is solvable in polynomial time if G is Eulerian [20], or if the two orientations (clockwise and counterclockwise) of every cycle have the same cost [21].

Let $\delta(i)$ be the set of edges incident to vertex v_i and $E(S)$ be the set of edges where $E(S) = (v_i, v_j) : v_i \in S, v_j \in V \setminus S$ and $S \subset V$. Furthermore, let x_{ij} be the number of times edge (v_i, v_j) is traversed from v_i to v_j . The WPP is represented as follows:

WPP Model 1

$$\text{Minimize} \quad \sum_{(v_i, v_j) \in A} (c_{ij} x_{ij} + c_{ji} x_{ji}) \quad (23)$$

$$\text{subject to} \quad x_{ij} + x_{ji} \geq 1 \quad ((v_i, v_j) \in A) \quad (24)$$

$$\sum_{(v_i, v_j) \in \delta(i)} (x_{ij} - x_{ji}) = 0 \quad (v_i \in V) \quad (25)$$

$$x_{ij}, x_{ji} \geq 0 \quad ((v_i, v_j) \in A) \quad (26)$$

$$x_{ij}, x_{ji} \text{ integer } ((v_i, v_j) \in A). \quad (27)$$

Let $P(G)$ be the convex hull of the feasible solutions to the integer program above, and let $Q(G)$ be the set of feasible solutions to its linear programming relaxation. Win [22] proved that every extreme point x of the polyhedron $Q(G)$ has components whose values are either $\frac{1}{2}$ or a nonnegative integer. Furthermore, $Q(G)$ is integral if and only if G is even. Let $S \subset V$ be such that $|E(S)|$ is odd. Grötschel and Win [23] showed that the following odd cut inequalities are valid for $P(G)$:

$$\sum_{(v_i, v_j) \in E(S)} (x_{ij} + x_{ji}) \geq |E(S)| + 1(S \subset V) \quad (28)$$

$$\sum_{v_i \in S, v_j \notin S} x_{ij} \geq \frac{1}{2}(|E(S)| + 1)(S \subset V) \quad (29)$$

$$\sum_{v_i \in S, v_j \notin S} x_{ji} \geq \frac{1}{2}(|E(S)| + 1)(S \subset V). \quad (30)$$

It is noticed that although there exists an exponential number of constraints in the above model and it has odd cut inequalities, it can still be solved in polynomial time by the Padberg and Rao procedure [24]. In general, the WPP has been solved by applying a cutting plane algorithm. For details, see Grötschel and Win [25], Win [20,22], and Gendreau *et al.* [26].

The Hierarchical Postman Problem (HPP)

If there is a predetermined order of how arcs need to be traversed, the WPP problem becomes the hierarchical postman problem (HPP). Take snow plowing operations, for example; emergency evacuation routes have higher priority and need to be serviced first. HPP can be defined on either a directed or an undirected graph, in which arcs or edges are partitioned into clusters $\{A_1, \dots, A_p\}$ with $p > 1$ such that precedence relations specify the order in which the clusters are to be traversed starting from and ending at the given depot. An example of the precedence relationship of arcs can be $A_1 \ll A_2 \ll \dots \ll A_p$, where $\ll$ represent a linear order.

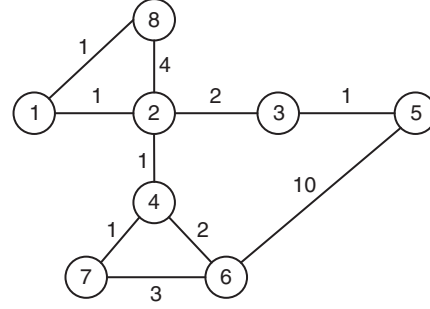


Figure 4. Example graph to illustrate an HPP.

If the clusters obey a linear order, the precedence relationship is said to be completely defined. The objective of a HPP is to determine a least-cost tour of G starting from and ending at a given depot, and serving the arcs of each cluster in a manner where no edges in A_i can be served before the service of all edges in A_{i-1} is completed. For the graph depicted in Fig. 4, if vertex 1 is the depot, $A_1 = \{(1,2), (2,3), (2,4)\}$; $A_2 = \{(3,5), (5,6), (4,6), (4,7), (6,7)\}$; $A_3 = \{(2,8), (1,8)\}$ and $A_1 \ll A_2 \ll A_3$, then a HPP tour is $1 \rightarrow 2 \rightarrow 4 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 4 \rightarrow 6 \rightarrow 4 \rightarrow 2 \rightarrow 8 \rightarrow 1$. This variant of CPP has been applied in applications such as flame cutting [27], waste collection [28], and snow plowing operations in an urban or rural setting [29].

Consider subgraphs $G_i = (V_i, A_i)$ derived from graph G induced by the sets A_i ; the problem can be solved in polynomial time if each subgraph is connected and the precedence relation is completely defined [30]. On the other hand, when subgraphs are not connected or if the precedence relation is partial, the problem becomes NP-hard. Dror *et al.* [30] and G  linas [31] proved this by transformation of the HPP from the Hamiltonian path problem, which is known to be strongly NP-complete in both the undirected and directed cases.

Algorithms to solve HPP have been developed in the past years. G  linas [31] proposed an exact enumerative algorithm for an undirected HPP with given starting and ending vertices, and general precedence relation, in which all subgraphs are strongly connected. As opposed to linear precedence relation, which requires a unique partial

ordering relation between classes, general precedence relation imposes a weak partial ordering relation between classes. This is the situation where, for example, all arcs of A_{i-2} must be serviced before all arcs of A_i , but the arcs of A_{i-1} can be serviced either before or after some arcs of A_{i-2} and A_i . The algorithm to this problem was based on a dynamic program in which states correspond to the subgraphs that have been traversed. Dror [30] proposed an algorithm of $O(p|V|^5)$ if a graph is undirected and all subgraphs induced by A_i are connected. Korteweg and Volgenant [32] improved that result to $O(p|V|^4)$ by a layer algorithm for the same problem. Ghiani and Improta [33] proposed a lower computational complexity algorithm that solves the problem as matching on an auxiliary graph with $O(p|V|)$ computational time. Their algorithm is described below.

Ghiani and Improta's Algorithm.

- Step 0.* For each subgraph $G_i, i = 1, \dots, p$, determine V' and V'' , where V' is the subset of vertices of V_i incident to at least one edge in $\bigcup_{k=1}^{i-1} A_k$, and V'' is the subset of vertices of V_i incident to at least one edge in $\bigcup_{k=1}^{i-1} A_k \cup A_{i+1}$.
- Step 1.* Construct the auxiliary $(p+2)$ partite graph G^* as described in their paper.
- Step 2.* Find a minimum weight perfect matching on G^* .
- Step 3.* Introduce in G the shortest paths P corresponding to the solution found in step 2; let G' be the resulting graph.
- Step 4.* Determine an optimal tour on G' using the end-pairing algorithm.

OTHER VARIANTS OF THE CHINESE POSTMAN PROBLEM

There exist several variations of the CPP. The first related problem is called *open CPP*, where the starting vertex and ending vertex of the path do not have to be the same. Note that the sufficient and necessary condition for the existence of an Eulerian path on the undirected graph is that there are at most two odd-degree vertices. The open CPP is

also important in real-world applications. For example, finding the optimal test sequence for a website does not require returning to the homepage. In this case, the objective is to find the shortest length open Eulerian path, which is an open path that covers all the arcs on the graph.

If not all arcs of the graph are required to be covered, the problem is called *rural postman problem*. This problem comes from a rural area setting: there are a number of villages whose streets have to be serviced but the links between villages do not need to be serviced but may be used for traveling. For a recent survey of the applications and algorithms of rural postman problem (RPP), refer to Eiselt *et al.* [34].

Another problem related to the undirected CPP is attributed to Alpern and it is referred to as the utilitarian postman (UP) problem. The UP problem's objective is to minimize the mean time of delivery of mail to all customers. More generally, a UP path is the one that minimizes the expected time required to find a random (uniformly distributed) point on a network. The optimal UP path is one of the open Eulerian paths but may be longer than the CP path. Furthermore, unlike a CP tour on an undirected graph whose length is bounded by twice the length of the graph, a UP path may traverse an edge more than twice and therefore no similar bound exists for a general graph. Recent papers devoted to this problem include Refs 35 and 36.

Numerous covering problems associated with the undirected CPP and directed CPP have also attracted the attention of researchers. An m -vehicle version of the undirected CPP with the objective to minimize the length of the longest route was proposed by Frederickson *et al.* [37]. Cross [38] considered the problem of covering an undirected graph by undirected cycles where the total length of the cover is minimized. For the directed case, several covering criteria such as covering the graph with star trees, and simple paths or circuits have been studied [39].

FURTHER READING

Classical books for an introduction of the CPP from a graph theoretical perspective

include: Assad and Golden [40], Christofides [41], Evans and Minieka [42], and Fleischner [11]. More recently, a book on the subject was edited by Dror [43]. Eiselt *et al.* [3] proposed a detailed survey paper for the CPP. Furthermore, Laporte *et al.* [44] provides a list of 500 references on four classical routing problems, including the CPP.

For an overall understanding of exact algorithms for the undirected CPP, the directed CPP and the mixed CPP, refer to Edmonds and Johnson [10]. Heuristic algorithms to solve the mixed CPP can be found in Refs 16 and 37. More recent literature on WPP and HPP include Refs 29, 32, 45–48 and Frederickson [49].

REFERENCES

1. Guan M. Graphic programming using odd and even points. *Chin Math* 1962;1:273–277.
2. Euler L. Solution problematis ad ceometrian situs pertinentis. *CASP* 1736;8:128–140.
3. Eiselt HA, Gendreau M, Laporte G. Arc routing-problems. 1. The Chinese postman problem. *Oper Res* 1995;43(2):231–242.
4. Larson RC, Odoni AR. *Urban Operations Research*. Belmont (MA): Dynamic Ideas; 2007.
5. Barahona F. On some applications of the chinese postman problem. In: Korte B, Lovasz L, Promel HJ, *et al.*, editors. Volume 9, Algorithms and combinatorics, paths, flows and VLSI-layout. Berlin: Springer-Verlag; 1990.
6. Thimbleby H. The directed Chinese postman problem. *Softw Pract Exp* 2003;33(11):1081–1096.
7. Lawler EL. *Combinatorial optimization: networks and matroids*. New York: Holt, Rinehart & Winston; 1976.
8. Galil Z, Micali S, Gabow H. An $O(E \log V)$ algorithm for finding a maximal weighted matching in general graphs. *Siam J Comput* 1986;15(1):120–130.
9. Derigs U, Metz A. Solving (large-scale) matching problems combinatorially. *Math Program* 1991;50(1):113–121.
10. Edmonds J, Johnson EL. Matching, Euler tours and the Chinese postman problem. *Math Program* 1973;5:88–124.
11. Fleischner H. Eulerian graphs and related topics (Part 1, Volume 1). Volume 50, *Annals of discrete mathematics*. Amsterdam: North-Holland; 1991.
12. Orloff CS. A fundamental problem in vehicle routing. *Networks* 1974;4(1):35–64.
13. Beltrami EL, Bodin LD. Networks and vehicle routing for municipal waste collection. *Networks* 1974;4(1):65–94.
14. Aardenne-Ehrenfest Tv, Bruijn NGD. Circuits and trees in oriented linear graphs. *Simon Stevin* 1951;28:203–217.
15. Ford LR, Fulkerson DR. *Flows in networks*. Princeton (NJ): Princeton University Press; 1962.
16. Christofides N, Benavent E, Campos V, *et al.* An optimal method for the mixed postman problem. In: Thoft-Christensen P, editor. Volume 59, *System modeling and optimization, Lecture Notes in control and Information Sciences*. Berlin: Springer; 1984.
17. Nobert Y, Picard JC. An optimal algorithm for the mixed Chinese postman problem. Publication No. 799. Montreal, Canada: Centre de recherche sur les transports; 1991.
18. Minieka E. Chinese postman problem for mixed networks. *Manage Sci* 1979;25(7):643–648.
19. Brucker P. The Chinese postman problem for mixed graphs. In: Noltemeier H, editor. *Graph theoretic concepts in computer Science*. Berlin: Springer; 1981. pp. 354–366.
20. Win Z. On the windy postman problem on Eulerian graphs. *Math Program* 1989;44(1):97–112.
21. Guan M. On the Windy postman problem. *Disc Appl Math* 1984;9(1):41–46.
22. Win Z. Contributions to routing problems [Doctoral Dissertation]. Universität Augsburg; 1987.
23. Grötschel M, Win Z. On the Windy postman polyhedron. Report No. 75. Germany: Schwerpunkt-program der Deutschen Forschungsgemeinschaft, Universität Augsburg; 1988.
24. Padberg MW, Rao MR. Odd minimum cut-sets and B-matchings. *Math Oper Res* 1982;7(1):67–80.
25. Grötschel M, Win Z. A cutting plane algorithm for the windy postman problem. *Math Program* 1992;55(3):339–358.
26. Gendreau M, Laporte G, Zhao Y. The windy postman problem on general graphs Publication No. 698, centre de recherche sur les transports, Montreal, Canada; 1990.

27. Manber U, Israni S. Pierce point minimization and optimal torch path determination in flame-cutting. *J Manuf Syst* 1984;3(1):81–89.
28. Bodin LD, Kursh SJ. Computer-assisted system for routing and scheduling of street sweepers. *Oper Res* 1978;26(4):525–537.
29. Perrier N, Langevin A, Amaya C-A. Vehicle routing for urban snow plowing operations. *Transp Sci* 2008;42(1):44–56.
30. Dror M, Stern H, Trudeau P. Postman tour on a graph with precedence relation on arcs. *Networks* 1987;17(3):283–294.
31. G  linas E. Le probleme du postier chunois avec contraintes generales de preseance [MSc A. Dissertation]. Ecole Polytechnique de Montreal; 1992.
32. Korteweg P, Volgenant T. On the hierarchical Chinese Postman Problem with linear ordered classes. *Eur J Oper Res* 2006;169(1):41–52.
33. Ghiani G, Improta G. An algorithm for the hierarchical Chinese postman problem. *Oper Res Lett* 2000;26(1):27–32.
34. Eiselt HA, Gendreau M, Laporte G. Arc routing-problems. 2. The rural postman problem. *Oper Res* 1995;43(3):399–414.
35. Jotshi A, Batta R. Search for an immobile entity on a network. *Eur J Oper Res* 2008;191(2):347–359.
36. Alpern S, Baston V, Gal S. Searching symmetric networks with utilitarian-postman paths. *Networks* 2009;53(4):392–402.
37. Frederickson GN, Hecht MS, Kim CE. Approximation algorithms for some routing problems. *Siam J Comput* 1978;7(2):178–193.
38. Cross H. Analysis of flow in networks of conduits of conductors. Bulletin No. 286. Urbana (IL): University of Illinois Engineering Experimental Station; 1936.
39. Busacker RG, Saaty TL. Finite graphs and networks. New York: McGraw-Hill; 1965.
40. Assad AA, Golden BL. Arc routing methods and applications. In: Ball M, Magnanti T, Monma C, *et al.*, editors. *Handbook of Operations Research and Management Science: Networks and Distribution*. Amsterdam: North-Holland; 1995.
41. Christofides N. Graph theory. an algorithm approach. London: Academic Press; 1975.
42. Evans JR, Minieka J. Optimization algorithms for networks and graphs. New York: Marcel Dekker; 1992.
43. Dror M, editor. Arc routing: theory, solutions and applications. Boston (MA): Kluwer; 2000.
44. Laporte G, Osman IH. Routing problems: a bibliography. *Ann Oper Res* 1995;61:227–262.
45. Martinez FJZ. Series-parallel graphs are Windy postman perfect. *Disc Math* 2008;308(8):1366–1374.
46. Benavent E, Corberan A, Pinana E, *et al.* New heuristic algorithms for the windy rural postman problem. *Comput Oper Res* 2005;32(12):3111–3128.
47. Benavent E, Carrota A, Corberan A, *et al.* Lower bounds and heuristics for the Windy Rural Postman Problem. *Eur J Oper Res* 2007;176(2):855–869.
48. Cabral EA, Gendreau M, Ghiani G, *et al.* Solving the hierarchical Chinese postman problem as a rural postman problem. *Eur J Oper Res* 2004;155(1):44–50.
49. Frederickson GN. Approximation algorithm for some postman problems. *J Assoc Comput Mach* 1979;26(3):638–554.

CLASSIC FINANCIAL RISK MEASURES

ARCADY NOVOSYOLOV
Institute of Mathematics,
Siberian Federal
University, Krasnoyarsk,
Russia

The goal of risk management or decision making under risk is to provide appropriate decisions in situations when consequences of actions are vague and uncertain. An early mention of a risk management technique is from Daniel de Foe: it may be found in his famous “Robinson Crusoe” [1]. During the first rain storm in the desert island, a sudden flash of lightning had struck Robinson Crusoe with a thought that all his powder might be destroyed in one blast. So after the storm, he separated the powder in not less than a hundred parcels. This is an early example of diversification, which is one of the risk management techniques.

Measurement is a process of ordering, if not quantifying. In the early twentieth century, there was some debate as to whether all risks are amenable to ordering, quantification or measurement. See, for example, Refs 2, 3 or for more modern treatment, Ref. 4. Any discussion of risk measurement should be understood in the context of that caveat.

For the sake of simplicity, we will confine ourselves to considering only a single period setting here, so that decision is made at present time t_0 , and consequences of the decision are to be exposed at some future time t_1 . Dynamic risk measures have attracted significant attention during the last decade, so we classify them as contemporary rather than classic. Thus, the consequences of a decision d are described by a random variable $X = X_d$, its values are interpreted as possible profit/loss values. In some cases it is more convenient to use cumulative distribution functions $F = F_d$ instead. Recall that cumulative distribution function F of a random

variable X is defined over real numbers $\mathbf{R}$ as follows:

$$F(t) = \mathbf{P}(X \leq t), \quad t \in \mathbf{R}.$$

RISK MEASURES VERSUS RISKINESS MEASURES

Unfortunately, there is an ambiguity of terminology in the field of risk measurement. The term “risk measure” is sometimes used to name functionals that measure risk *per se*. These functionals are expected to take positive values for risky actions (consequences), and zero value for risk-free actions (consequences); they never take negative values. Such risk measures take only risk into account and ignore benefits obtained in exchange for the risk. They should take larger values for riskier decisions (the more—the riskier). Variance is a typical example of risk measure of this first kind.

Another usage of the term “risk measure” assumes these functionals take into account not only risk, but also benefits and all other relevant information about consequences of decisions. These functionals take larger values for better decisions (the more—the better), so they represent preferences over actions (consequences). The functionals can usually take arbitrary real values, both positive and negative; zero value of such functionals does not have any specific meaning. These functionals may be also called *decision criteria*, since making decisions may be formalized as maximization of such functionals. Certainty equivalent is a typical measure of this second kind. Note that riskiness measures and decision criteria are often, though not always, related through expectation $\mathbf{E}X$ of risk X by

$$\begin{aligned} \text{riskiness measure}(X) &= \text{decision tool}(X) \\ &\quad - \mathbf{E}X. \end{aligned}$$

Though using the term “risk measure” for the functionals of the first kind is more intuitive

and attractive, it is more often used in the literature for the functionals of the second kind. Well-known examples are expected utility functionals briefly considered in the present article and also presented in a separate article, and coherent risk measures considered in a separate article.

RISK

Financial risk in a one-period setting is often thought of as uncertainty in a financial result (gain or loss). For purposes of measurement, this is modeled with a random variable. In a simple case, the size of loss may be known in advance and fixed, only probability of loss may vary; the loss probability serves as a risk measure of the first kind:

$$\text{amount of risk} = \text{probability of loss.}$$

Note that negative of this functional may be treated as a risk measure of the second kind. If the loss size may vary too, the amount of risk is often measured by

$$\begin{aligned} \text{amount of risk} &= (\text{value of loss}) \\ &\quad \times (\text{probability of loss}) \end{aligned}$$

Denoting v the value of loss and p its probability, we can describe the loss as a random variable X with the following distribution

$$X: \begin{array}{ll} \text{Value} & 0 \quad v \\ \text{Probability} & 1-p \quad p \end{array}$$

In this case, risk might be measured as $f(X) = vp$, the expected value $\mathbf{E}X$ of the loss random variable X . This value may be regarded as a risk measure of the first kind and again, its negative may be considered as a risk measure of the second kind. Such a duality is possible due to a narrow set of allowed risks.

More generally, profit/loss distribution may be represented by a discrete random variable in the form

$$X: \begin{array}{lllll} \text{Value} & x_1 & x_2 & \dots & x_n \\ \text{Probability} & p_1 & p_2 & \dots & p_n \end{array},$$

and the risk of X may be measured by

$$f(X) = \mathbf{E}X = \sum_{i=1}^n X_i p_i.$$

This value $f(X)$ may be regarded as a proxy for the certainty equivalent of X , or the price at which the uncertain loss X may be sold and bought in a financial market; for example, as an insurance policy. When all expenses are included in the values of X , this functional $f(\cdot)$ may be used as a (very crude) decision criterion, that is, the risk measure of the second kind.

ST. PETERSBURG PARADOX

In 1738 Daniel Bernoulli published the paper [5] where expectation was shown to be not appropriate as a risk measure. This view was supported by presenting a game with infinite expected gain, so a mean-driven person would be willing to pay any amount of money to enter the game, while real persons would hardly behave that way, thus giving rise to a paradox.

The game is played as follows: A fair coin is tossed till the first appearance of heads. If the first appearance occurs at the n -th toss, then the gain is equal to 2^n dollars. Since the probability of the first appearance of heads at the n -th toss is 2^{-n} , we have the following gain description of the game:

	Toss of heads appearance	1	2	...	n	...
X :	Value of gain	2	4	...	2^n	...
	Probability of the gain	1/2	1/4	...	2^{-n}	...

so the mean gain is indeed infinite:

$$\mathbf{E}X = \sum_{n=1}^{\infty} 2^n 2^{-n} = 1 + 1 + \dots = \infty.$$

Infiniteness of expectation means that though gain in each game is finite, the average gain in a long series of games tend to increase without bound, although it increases very slowly. Figure 1 illustrates the effect for a series of 1000 games. The data for the graph has been prepared as follows: Consider a sequence of 1000 games, denote X_k the gain

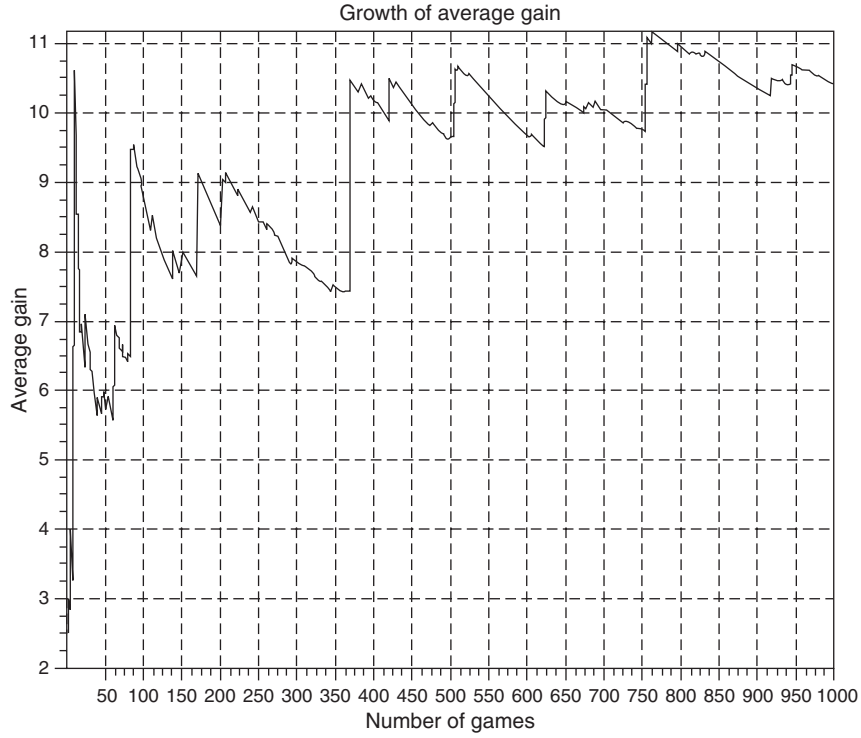


Figure 1. Average gain in m consecutive games, $m = 1, 2, \dots, 1000$.

in the game $k, k = 1, \dots, 1000$, and denote S_m the average gain after m games:

$$S_m = \frac{1}{m} \sum_{k=1}^m X_k, \quad m = 1, 2, \dots, 1000.$$

From a strict probabilistic point of view, S_m is a random process, and Fig. 1 depicts a sample path of that process.

To overcome the paradox, Bernoulli suggested using the mean logarithm of a random variable instead of its expectation as a decision criterion (risk measure), that is, calculating

$$f(X) = \mathbf{E} \log X.$$

Using logarithm base 2 provides for our game

$$\begin{aligned} f(X) &= \mathbf{E} \log_2 X = \sum_{n=1}^{\infty} 2^{-n} \log_2 2^n \\ &= \sum_{n=1}^{\infty} n 2^{-n} = 2. \end{aligned}$$

According to Bernoulli's suggestion, fair price Q for the game is equal to the monetary value, which has the logarithm equal to $f(x) = 2$, that is, $\log_2 Q = 2$, which gives $Q = 4$.

EXPECTED UTILITY

Logarithms are used to calculate utility of wealth in the above consideration, which is why it is called *utility function* in this framework. Using other utility functions U , instead of the logarithms, provides the more general expected utility principle. According to this principle, given a utility function U , one should calculate expected utility $f(X) = \mathbf{E}U(X)$ of the game X and then find the fair price Q for the game X so that $U(Q) = \mathbf{E}U(X)$. In other words, utility of Q should be equal to the expected utility of X .

Typically the utility function is chosen to be strictly increasing, in which case the inverse function U^{-1} exists, so the fair price (certainty equivalent) may be also calculated directly by

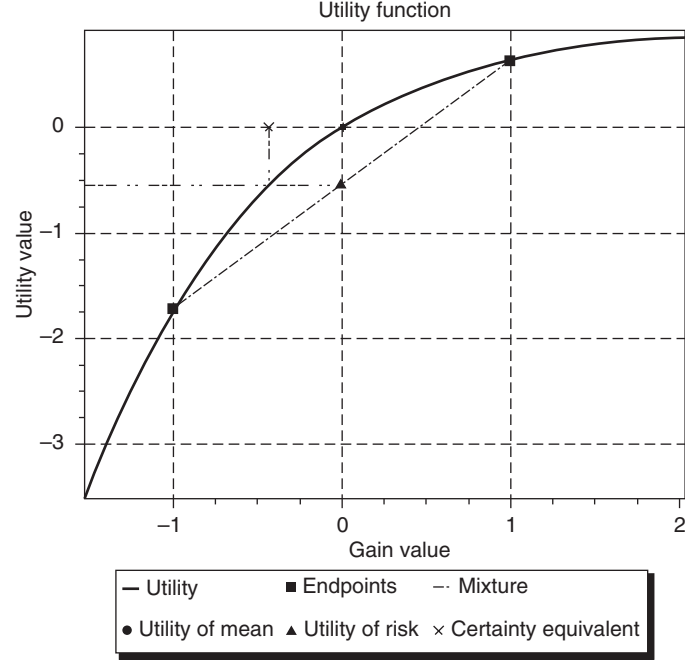


Figure 2. Increasing concave utility function.

$$Q = U^{-1}(\mathbf{E}U(X)).$$

If additionally U is concave (sometimes such functions are also called *convex upwards*), it is said to exhibit risk aversion. Indeed, in this case a certain amount has greater or equal utility than any risky amount with the same mean value. Figure 2 presents an example of typical increasing concave utility function and illustrates the risk aversion concept. Here the utility function takes the form

$$U(x) = 1 - \exp(-x).$$

Consider the risk (random gain) described by the table

X :	Value	-1	1
	Probability	$\frac{1}{2}$	$\frac{1}{2}$

Since $U(-1) = -1.718$ and $U(1) = 0.632$, the expected utility of the risk X equals

$$\mathbf{E}U(X) = \frac{1}{2}U(-1) + \frac{1}{2}U(1) = -0.543.$$

Next, $\mathbf{E}U(X)$ is less than $U(\mathbf{E}X) = U(0) = 0$, so the expected utility principle declares the

risk X less preferable than its expected value $\mathbf{E}X$. Moreover, the inverse utility function U^{-1} is uniquely defined:

$$U^{-1}(z) = -\ln(1 - z), \quad z \in (-\infty, 1);$$

thus, we can calculate the certainty equivalent of the risk X as follows:

$$Q = U^{-1}(\mathbf{E}U(X)) = -\ln(1 + 0.543) = -0.434.$$

The points $(-1, U(-1))$ and $(1, U(1))$ are marked in Fig. 2 by squares; the triangle represents the utility of risk point $(\mathbf{E}X, \mathbf{E}U(X))$, and the diagonal cross marks the certainty equivalent of the risk X .

The expected utility principle was supplied with a solid ground in the seminal book by von Neumann and Morgenstern [6], where they built an axiomatic theory of expected utility. In particular, they showed that if preference over probability distributions is linear in some specific sense, then it may be represented by an expected utility functional. To be more precise, for two risks X and Y denote $X \preceq Y$ the fact that Y is at least as preferred as X . If the preference relation $\preceq$ possesses

the linearity property mentioned above, then there exists a utility function U (unique in some sense) such that $\mathbf{E}U(X) \leq \mathbf{E}U(Y)$ if and only if $X \preceq Y$.

Linearity of the preference relation generated by an expected utility functional may be illustrated by a simple example. Consider the set of risks such that each risk can take only three values 1, 2, 3 with corresponding probabilities p_1, p_2, p_3 . Then distribution of any such risk X may be described by only two parameters p_1, p_2 such that $p_1 \geq 0, p_2 \geq 0$ and $p_1 + p_2 \leq 1$. The third probability is clearly equal to $p_3 = 1 - p_1 - p_2$. The set of all admissible parameters p_1, p_2 is represented by a shaded triangle on a plane (Fig. 3). We say that two risks X, Y are equivalent and denote the fact by $X \sim Y$, if $X \preceq Y$ and $X \succeq Y$. An equivalence class K contains only equivalent risks, so that $X, Y \in K$ if and only if $X \sim Y$. Alternatively, equivalence classes may be described by the expected utility condition: $X, Y \in K$ if and only if

$$\mathbf{E}U(X) = \mathbf{E}U(Y) = \alpha_K.$$

Different values of α_K correspond to different equivalence classes. Each dashed line in Fig. 3 represents an equivalent class. Equivalence classes are segments of parallel straight lines; this property is characteristic

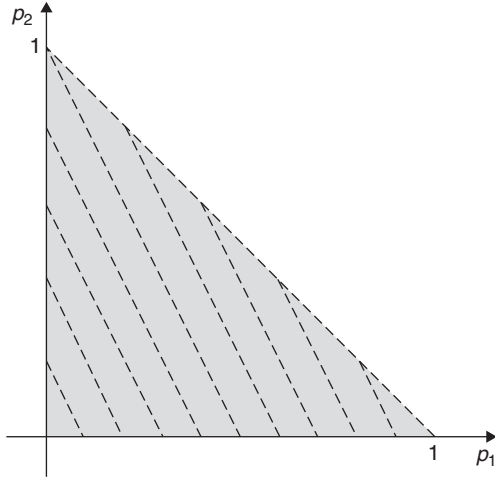


Figure 3. Equivalence classes of a linear preference relation.

for the expected utility principle. This is why we call the corresponding preference relations linear.

MARKOWITZ MEAN-VARIANCE APPROACH

In 1952 Harry Markowitz in his seminal paper [7] introduced the mean-variance approach to portfolio selection. For simplicity we will start with the case of two assets. Let $X = (X_1, X_2)'$ be a random vector, whose component X_i represents return of asset i over a fixed horizon $i = 1, 2$; prime denotes transpose vector. Given weight w , a portfolio containing a portion w of the first asset and a portion $1 - w$ of the second asset, provides a return $P_w = wX_1 + (1 - w)X_2$ over the same horizon. The random variable P_w has expectation and variance

$$\begin{aligned} \mathbf{E}P_w &= w\mathbf{E}X_1 + (1 - w)\mathbf{E}X_2, \\ \sigma^2(P_w) &= w^2\sigma^2(X_1) + 2w(1 - w)r\sigma(X_1)\sigma(X_2) \\ &\quad + (1 - w)^2\sigma^2(X_2), \end{aligned}$$

where r denotes correlation of X_1 and X_2 . The expected return $\mathbf{E}P_w$ may informally be called *return*, and the variance $\sigma^2(P_w)$ is often treated as a proxy for the risk of the portfolio P_w . This is why the approach is often called *risk-return approach*. The essence of this approach is minimizing risk given a fixed return value M or, equivalently, solving the following optimization problem:

$$\sigma^2(P_w) \rightarrow \min_w, \quad (1)$$

$$\mathbf{E}P_w = M. \quad (2)$$

If short-selling is prohibited or impossible, then additional constraint

$$0 \leq w \leq 1 \quad (3)$$

is also included into the optimization problem.

The parameter M describes risk appetite of a decision maker. The more risky decision maker would set larger value of M and get a more risky solution of the Markowitz optimization problem.

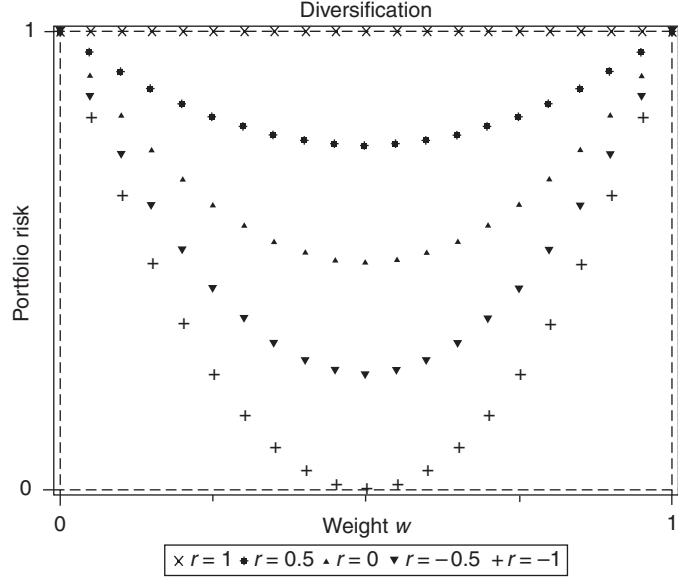


Figure 4. Graphs of risk of two-asset portfolios with different asset correlation.

Note also that along with the problems (1)–(3), Markowitz in Ref. 7 had considered the equivalent problem of maximizing expected return given a level of risk, namely,

$$\begin{aligned} \mathbf{E}P_w &\rightarrow \max_w, \\ \sigma^2(P_w) &= \Sigma^2 \end{aligned}$$

with the additional constraint (3) if short-selling is prohibited or impossible.

Let us illustrate the diversification effect which might be obtained using portfolio selection. To keep things simple, assume that asset returns are equal, $\mathbf{E}X_1 = \mathbf{E}X_2 = R$, and asset variances are also equal, $\sigma^2(X_1) = \sigma^2(X_2) = D^2$. In this case the constraint $\mathbf{E}P_w = M$ does not make sense, since always $\mathbf{E}P_w = R$, so the Markowitz problem reduces to just minimizing risk, which in this case equals

$$\begin{aligned} \sigma^2(P_w) &= D^2 (w^2 + 2w(1-w)r + (1-w)^2) \\ &= D^2 (2w^2(1-r) - 2w(1-r) + 1). \end{aligned}$$

If $r = 1$, then the portfolio risk is always equal to D^2 and there is no diversification in case of perfectly correlated assets. If $-1 \leq r < 1$,

then the minimum risk is clearly attained at $w = 1/2$, and is equal to

$$D^2 \frac{1+r}{2}.$$

We see that the closer r value to -1 , the stronger the diversification. In the ideal case $r = -1$, we obtain zero risk. Figure 4 illustrates this example by presenting graphs of risk vs w for values of $r = -1; -0.5; 0; 0.5; 1$.

The general case of n assets is quite similar. Denote $X = (X_1, \dots, X_n)'$ the random vector with components describing returns of assets on a given horizon and $w = (w_1, \dots, w_n)'$ the vector of weights. Then return of a portfolio has the form

$$P_w = w'X = \sum_{i=1}^n w_i X_i$$

and its expected return and risk are

$$\begin{aligned} \mathbf{E}P_w &= w'm = \sum_{i=1}^n w_i m_i, \\ \sigma^2(P_w) &= w'Vw = \sum_{i,j=1}^n v_{ij} w_i w_j, \end{aligned}$$

where $m = (m_1, \dots, m_n)'$ with $m_i = \mathbf{E}X_i$, $i = 1, \dots, n$ is the vector of expected returns, and

$V = (v_{ij})$ with $v_{ij} = \text{Cov}(X_i, X_j), i, j = 1, \dots, n$ is a variance–covariance matrix. The Markowitz problem still has the form of Equations (1), (2) with the additional constraint

$$w_1 + \dots + w_n = 1, \quad (4)$$

and more additional constraints in case of impossibility of short trades:

$$w_1 \geq 0, \dots, w_n \geq 0.$$

In case of asymmetric distributions, using variance $\sigma^2(P_w) = \mathbf{E}(P_w - \mathbf{E}P_w)^2$ as a riskiness measure may not be natural, because it imposes equal penalty to both downside and upside deviations from the mean. In such circumstances many authors, including Markowitz, suggest using downside measures like semivariance, which is defined as conditional expectation of the form

$$\sigma_-^2(P_w) = \mathbf{E} \left((P_w - \mathbf{E}P_w)^2 \mid P_w \leq \mathbf{E}P_w \right).$$

MARKOWITZ RISK MEASURES

It turns out that solving a Markowitz problem (Eqs 1, 2, and 4) is equivalent to solving the problem

$$\mathbf{E}P_w - \gamma \sigma^2(P_w) \rightarrow \max_w \quad (5)$$

with constraint (4), where $\gamma > 0$ is uniquely defined by the parameter M in Equation (2). The parameter γ describes the risk aversion of a decision maker; smaller values correspond to more risky decision makers. This equivalence invokes using the functional

$$f(X) = \mathbf{E}X - \gamma \sigma^2(X) \quad (6)$$

as a decision criterion, or risk measure of the second kind. Note that the Markowitz problem (Eqs 1 and 2) allows equivalent statement with portfolio standard deviation $\sigma(P_w)$ as a goal function instead of the variance as in Equation (1). This observation leads to another risk measure of the second kind, related to the Markowitz problem. Fix $\alpha > 0$ and consider a functional

$$f(X) = \mathbf{E}X - \alpha \sigma(X).$$

One can easily see that this functional is

- positive homogeneous, that is, $f(tX) = tf(X), t \geq 0$;
- superadditive, that is, $f(X + Y) \geq f(X) + f(Y)$;
- translation invariant, that is, $f(X + a) = f(X) + a$ for real a .

However it is not a monotone, that is for risks X, Y such that $X \leq Y$; both inequalities $f(X) > f(Y)$ and $f(X) < f(Y)$ are possible. To show this, consider a degenerate risk X with $\mathbf{P}(X = 0) = 1$ and a family of Bernoulli risks Y_p with

$$\mathbf{P}(Y_p = 1) = p, \mathbf{P}(Y_p = 0) = 1 - p,$$

where p is a parameter. We have $f(X) = 0$ and $Y_p \geq X$, however,

$$f(Y_p) = p - \alpha \sqrt{p(1-p)}$$

may take both positive and negative values. Indeed,

$$f(Y_p) < 0 \text{ for } 0 < p < \frac{\alpha^2}{1 + \alpha^2}, \quad f(Y_p) > 0 \text{ for } \frac{\alpha^2}{1 + \alpha^2} < p \leq 1,$$

see illustration in Fig. 5 for $\alpha = 0.5$. The properties of positive homogeneity, superadditivity, translation invariance, and monotonicity are closely related to the so-called coherent risk measures which are covered in a separate article.

VALUE-AT-RISK

Value-at-risk (VaR) was introduced into financial applications in the early 1990s by a RiskMetrics technical document, the latest version of which is available in Ref. 8. In a nutshell, VaR means the following: suppose someone tells you that a given portfolio will lose \$10,000 one day in ten. In the language of VaR, he/she is telling you that the portfolio has a one-day 90% VaR of \$10,000—that nine days out of ten, it will lose less than \$10,000 or even gain something.

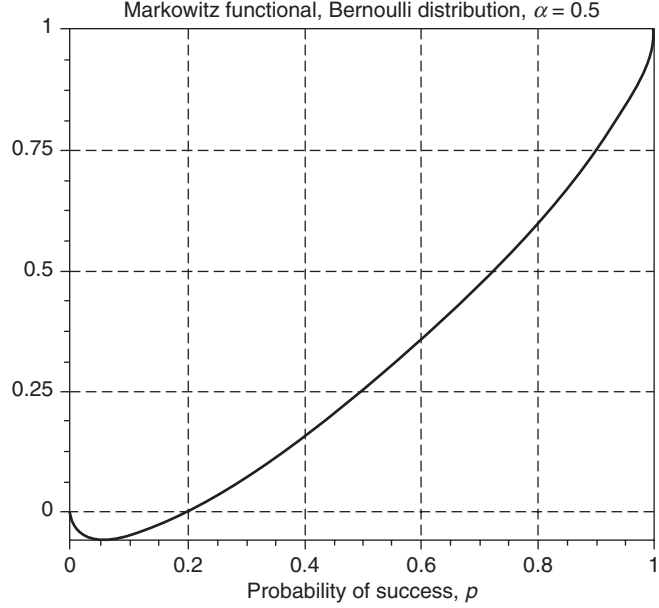


Figure 5. Markowitz functional for a family of Bernoulli risks.

More formally, VaR is a number such that with a given probability, loss over a given horizon will not exceed that number. Fix a horizon, say one day, a confidence level or probability, say 90%, and consider a profit/loss distribution with the density shown in Fig. 6. The one-day 90% VaR is shown in the figure by the square mark. In other words, 90% of the distribution probability belongs to the best side of the distribution, while the remaining 10% belongs to the worst side.

Calculating VaR is largely a process of somehow constructing a profit/loss distribution such as that of Fig. 6 and then taking the desired quantile. Inputs must provide information about both, the volatility (uncertainty) in market prices, interest rates or other market variables, as well as information about the portfolio's exposures to those market variables. For the former, historical values for the market variables are used. The latter is obtained from the portfolio's holdings. See Ref. 9 for a more detailed discussion.

Let us describe a parametric method for one-day VaR calculation in more detail. It usually starts with a factor model

$$A = LB + Z, \quad (7)$$

where B stands for an $n \times 1$ vector of one-day factor returns, which is supposed to have a joint normal distribution with zero mean and variance-covariance matrix C_B , $m \times n$ matrix L stands for loadings, Z denotes normal random vector with zero mean and diagonal variance-covariance matrix C_Z , and is interpreted as asset-specific variance source (noise). Here, A denotes $m \times 1$ vector of asset returns. Factor and noise are assumed to be mutually independent. From factor model assumptions it follows that A has a joint normal distribution with zero mean and variance-covariance matrix $C_A = LC_B L' + C_Z$.

Now consider an $m \times 1$ vector w of weights, and compose a portfolio of assets with these weights; its return takes the form $P_w = w'A$. This is clearly a normal random variable with zero mean and variance $\sigma^2(P_w) = w'C_A w$. Thus, its α -quantile equals $q_\alpha \sqrt{w'C_A w}$, where q_α denotes the α -quantile of the standard normal distribution, for example $q_{0.9} = 1.2816$. Finally, if the current portfolio value equals S_0 , then its one-day α -VaR equals $S_0 q_\alpha \sqrt{w'C_A w}$.

VaR is a rather simple method of risk measurement and communication. It has a number of disadvantages though, especially when

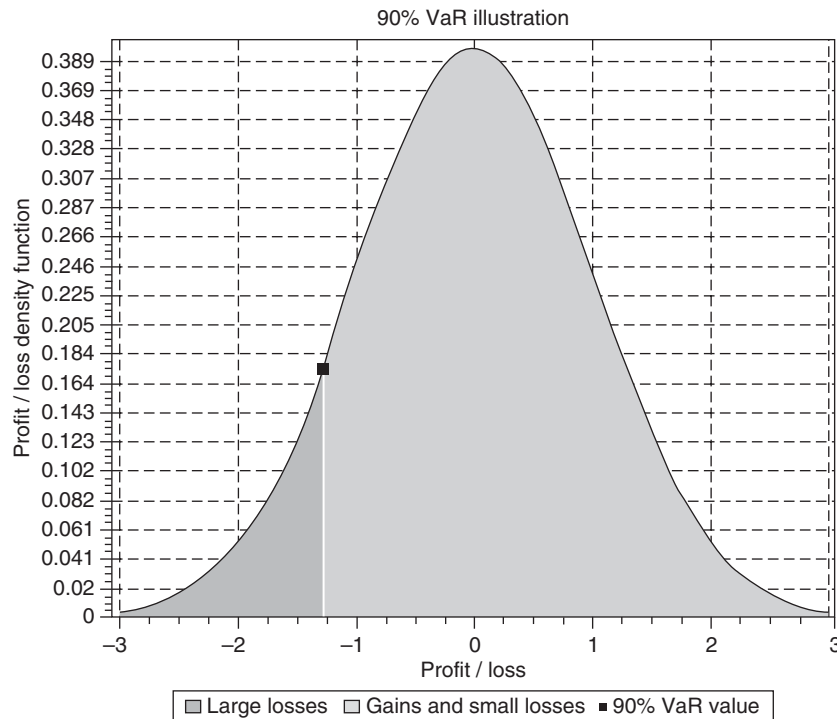


Figure 6. One-day 90% VaR illustration.

normality assumption cannot be justified. First, VaR tells us nothing about tail behavior of the distribution of interest, in particular, it cannot catch the so-called “fat tails” or “heavy tails” [10]. Second, in general, VaR is not a subadditive functional, so it may fail to support decision leading to diversification [2,11]. These features force both researchers and practitioners to look for advanced risk measures. Representatives of those, coherent risk measures, are considered in a separate article.

FURTHER READING

For a general discussion of risk, see Refs 2–4. For deeper understanding of expected utility theory, preferences over distributions, and basics of decision making under risk, the classic monograph [6] is a must-read. Development of the Markowitz mean–variance approach may be found in the book [12], and also in the publications of his colleagues

[13,14]. For calculation and interpretation VaR, see Ref. 9. An interesting class of risk measures, the so-called risk-value functionals, are presented in Ref. 15 and other papers of these authors. A coherent modification of VaR, the conditional VaR, was introduced in Ref. 16. There is also a vast literature on dynamic risk measurement [17]. Some paradoxes of choice among random alternatives are discussed in Ref. 18.

REFERENCES

1. Defoe D. The life and adventures of Robinson Crusoe. 1719. Available at http://www.planetpdf.com/planetpdf/pdfs/free_ebooks/Robinson_Crusoe_BT.pdf.
2. Knight FH. Risk, uncertainty, and profit. New York: Hart, Schaffner, and Marx; 1921.
3. Keynes JM. A treatise on probability. London: Macmillan; 1921.
4. Holton GA. Defining risk. *Financ Anal J* 2004; 60(6):19–25.

5. Bernoulli D. Exposition of a new theory of the measurement of risk. *Econometrica* 1954; 22(1):22–36. (first published in 1738).
6. von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton (NJ): Princeton University Press; 1944.
7. Markowitz H. Portfolio selection. *J Finance* 1952;7(1):77–91.
8. RiskMetrics. Technical Document; 1997. Available at <http://www.riskmetrics.com/publications/techdocs/rmcovv.html>.
9. Holton G. *Value at risk: theory and practice*. San Diego (CA): Academic Press; 2003.
10. Liu M-Y, Wu C-Y, Lee H-F. Fat tails and VaR estimation using power EWMA models. *J Acad Bus Econ* 2004.
11. Ruppert D. *Statistics and data analysis for financial engineering*. New York: Springer; 2004.
12. Markowitz H. *Portfolio selection: efficient diversification of investments*. New Jersey: John Wiley & Sons, Inc.; 1959.
13. Tobin J. Liquidity preference as behavior towards risk. *Rev Econ Stud* 1958;25(67): 124–131.
14. Sharpe W. Capital asset prices: a theory of market equilibrium under conditions of risk. *J Finance* 1964;19(3):425–442.
15. Jia J, Dyer J. A standard measure of risk and risk-value models. *Manage Sci* 1996;42(12): 1691–1705.
16. Tyrrell R, Uryasev S. Optimization of conditional value-at-risk. *J Risk* 2000;2(3): 21–41.
17. Hey J. Experimental economics and the theory of decision making under risk and uncertainty. *GPRIT* 2002;27:5–21.
18. Blyth C. Some probability paradoxes in choice from among random alternatives. *J Am Stat Assoc* 1972;67(338):366–373.

CLIQUE RELAXATIONS

BENJAMIN MCCLOSKEY
Nature Source Genetics,
Ithaca, New York

INTRODUCTION

Researchers have historically studied the clique concept from a variety of perspectives. For example, cliques arise in extremal graph theory [1], complexity theory [2,3], random graph theory [4,5], and perfect graph theory [6–8]. In addition, cliques provide the applied mathematician with a natural framework for modeling and detecting relationships among elements of a system [9–11].

In practice, the precise definition of a clique can be overly restrictive [12], thus motivating the study of clique relaxations. This article discusses various ways to relax the clique concept. The remainder of this section contains a brief survey of applications from the field of social network analysis to motivate the more precise definitions and concepts found in later sections. While applications serve to motivate the present discussion, it should be noted that the theoretical aspects of clique relaxations are an active area of research as well [13–15].

Consider the example of a social network $G = (V, E)$, where G is finite, simple, and undirected. The vertex set V represents people (or *actors*). The edge set E represents mutual relationships between pairs of actors. As the following examples show, social networks have become a popular topic.

For instance, suppose the edges in a social network denote physical contact between people. Health officials use these *contact networks* to study the spread of infectious diseases [16,17]. The rate of disease transmission naturally accelerates in regions with high adjacency levels, and the identification of dense subnetworks can help officials understand how certain actions, such as

closing schools, might affect the rate of transmission [18].

Now suppose the social network edges denote mutual friendship. Advertisers have long had an interest in exploiting friendship networks for marketing purposes [19–21], and research has found a significant and positive impact of friends' purchases on the purchase probability of an individual [22]. In particular, analysts study how the influence of certain actors can propagate throughout the network [23,24]. This spread of influence depends on the structure of the network and is greatly facilitated by the presence of dense subnetworks.

As a final example, suppose the network edges represent collaboration between researchers. These *collaboration networks* have been constructed for mathematicians and computational geometers [25,26]. Collaboration networks are used to determine collaborative distance between researchers, a notion popularized by the concept of Erdős numbers [27].

Each of these applications involves the search for subsets potentially modeled as cliques, but there are other options. The following section introduces alternatives to the clique. The section titled “Combinatorial Algorithms” discusses algorithms for finding certain clique relaxations, and the section titled “Theoretical Results” contains some results related to these objects.

RELAXING CLIQUES

Cliques consist of pairwise adjacent vertices. This section discusses examples of clique relaxations. But first, recall the following standard graph-theoretic terminology [28]. Let $N(v) = \{u \in V : vu \in E\}$ and $\deg(v) = |N(v)|$. Let $d_G(v, u)$ be the length of a shortest vu -path in G , and let $G[S]$ be the subgraph induced by $S \subseteq V$. The diameter of a graph is equal to $\max_{u,v \in V} d_G(v, u)$.

Some possibilities for relaxing cliques include imposing a bound on adjacency level,

graph-based distance, or edge-to-vertex ratio. Notice that cliques have a high adjacency level, small diameter, and high edge-to-vertex ratio. Let $\lambda \in [0, 1]$ and $k \geq 1$ be a fixed integer.

Definition 1 [29]. A k -plex S satisfies $|S \cap N(v)| \geq |S| - k$ for all $v \in S$.

k -plexes are degree-based clique relaxations. The degree requirement must hold at every vertex, so k -plexes tend to be highly structured. Notice also that k -plexes form a superset of cliques (i.e., cliques are 1-plexes). The section titled “Combinatorial Algorithms” contains a maximum k -plex algorithm, and the section titled “Theoretical Results” discusses some aspects of the k -plex polytope.

Definition 2 [30]. A k -core satisfies $|S \cap N(v)| \geq k$ for all $v \in S$.

k -cores relax the degree requirement of cliques in a considerably less stringent way than k -plexes. In particular, the k -core degree bound is independent of S . This independence allows the structure of k -cores to differ greatly from k -plexes and leads to a simple algorithm for finding maximum k -cores discussed in the section titled “Combinatorial Algorithms.”

The k -core and k -plex concepts have a superficial relationship in that any k -plex S is also a $(|S| - k)$ -core. Conversely, however, a maximum $(|S| - k)$ -core could be much larger than $|S|$ and need not be a k -plex. Therefore, this fact has no algorithmic consequences.

Definition 3 [31]. A k -club satisfies $d_{G[S]}(v, u) \leq k$ for all $u, v \in S$.

k -clubs are an example of distance-based clique relaxations. k -clubs have the interesting property that testing their maximality is NP-complete [13]. The section titled “Combinatorial Algorithms” describes a maximum k -club algorithm. The section titled “Combinatorial Algorithms” discusses some further properties of k -clubs.

Definition 4 [32]. A k -clique satisfies $d_G(v, u) \leq k$ for all $u, v \in S$.

k -cliques offer a second example of distance-based clique relaxations. k -cliques and k -clubs have much in common; the main difference being that the distance measurements used to define k -cliques do not depend on the subgraph $G[S]$. Notice that k -clubs form a proper subset of k -cliques.

Definition 5 [33]. A λ -quasi-clique satisfies $|E(G[S])| \geq \lambda \frac{|S|(|S|-1)}{2}$.

Quasi-cliques have received other definitions in the literature [34]. Here they are defined to obtain high edge density, but note that the definition fails to ensure highly structured subsets. Indeed, a quasi-clique can have vertices with very low degree. On the other hand, density does imply connectivity. More precisely, a classic theorem of Mader states that every graph with edge density at least $3k$ and sufficiently many vertices contains a k -connected subgraph with at least r vertices or r pairwise disjoint k -connected subgraphs [35].

The examples in this section offer various approaches to relaxing the clique concept. A more thorough discussion of these objects can be found in Ref. 12. The following section deals with algorithms for finding clique relaxations.

COMBINATORIAL ALGORITHMS

This section discusses combinatorial algorithms for finding the clique relaxations discussed above. Finding a maximum k -core is especially easy. Sequentially deleting all vertices of degree at most $k - 1$ results in either a maximum k -core if one exists or the empty graph [12]. The difficulty of optimizing for quasi-clique varies. On the one hand, finding a subgraph with maximum edge density can be solved in polynomial time through a series of min cut calculations [36,37]. On the other hand, given $k \geq 1$, the problem of finding a subgraph on k vertices with the maximum number of edges is NP-complete [38].

Combinatorial branch and bound algorithms offer a standard approach to the NP-complete problems of finding maximum k -plexes, k -cliques, or k -clubs. Maximum clique has been extensively studied [39–42], so it is natural to start by adapting a well-known clique algorithm [10]. Consider the recursive clique algorithm in Algorithm 1.

Algorithm 1. Clique algorithm

```

function Clique( $U, K$ )
1.  if  $U = \emptyset$ 
2.      if  $|K| > \text{max}$  and  $K$  is feasible
3.           $\text{max} = |K|$ 
4.      end
5.      return
6.  end
7.  while  $U \neq \emptyset$ 
8.      if  $|K| + b(U) \leq \text{max}$ 
9.          return
10.  end
11.  Choose  $v \in U$ 
12.   $U = U \setminus \{v\}$ 
13.  Clique( $U \cap N(v), K \cup \{v\}$ )
14. end
15. return
    
```

For the initial call, set $U = V(G)$ and $K = \emptyset$. Refer to K as the current set and to U as the candidate set. The candidate set contains elements that can extend the current set. The function b in line 8 produces an upper bound on $\max_{S \subseteq U} |S|$ such that $K \cup S$ is a clique. A typical algorithm would use graph coloring for this bound. Line 13 redefines the candidate set as a consequence of adding v to the current set. Note that lines 8 and 13 are the only instructions specific to cliques. Consequently, adapting this algorithm amounts to redefining b and determining how the candidate set changes with respect to changes in the current set.

For k -plex and k -clique, adapting the candidate set is straightforward. Specifically, $K \cup \{u\}$ must be a feasible solution for all $u \in U$. This follows from the fact that k -plexes and k -cliques are closed under set inclusion. More precisely, if there exists a subset $S \subseteq U$ such that S extends K , then all subsets of $K \cup S$ are feasible. In particular, $K \cup \{u\}$ is feasible for all $u \in S$. Notice also that, when applied to k -plexes or k -cliques, the algorithm maintains the feasibility of the current set K .

To adapt the bound function b for k -plex and k -clique, simply partition the candidate set in a way that limits the intersection of a feasible solution with each partition class. Then the number of partition classes produces a bound on the amount that U can extend K . Co- k -plex colorings [14] and distance k -colorings [13] are known partitions for bounding k -plex and k -clique candidate sets, respectively.

In contrast, k -clubs are not closed under set inclusion, and testing k -club maximality is NP-complete [13]. Therefore, constructing a candidate set in terms of elements that extend a k -club must be NP-complete as well. One way to circumvent this computationally intractable subproblem is to relax the algorithm’s feasibility invariant. In other words, allow the current set to become infeasible at certain nodes in the branch and bound tree. In practice, one can define the k -club current set, candidate set, and bound function as in the k -clique algorithm described above [13]. However, it is important that the variable max be set only by feasible k -clubs. The validity of this approach follows from the fact that all k -clubs are also k -cliques, so the algorithm never prunes feasible k -clubs with the potential to improve the incumbent.

A second approach for designing combinatorial algorithms also has its origin in the clique literature [40]. The idea is to reverse the algorithm in Algorithm 1. In the context of cliques, the algorithm iteratively finds a maximum clique in $G[\{v_1\}]$, $G[\{v_1, v_2\}]$, $\dots$, and G . The algorithm maintains the invariant that the candidate set is contained in a graph for which the size of a maximum clique is known. This invariant leads to an effective bounding function.

It is easy to see that such an algorithm generalizes to find a maximum independent set in any independence system, so k -plex and k -clique algorithms follow immediately [14,43]. Using the same algorithmic tricks described above, one could devise a similar approach to find maximum k -clubs as well. The algorithm would essentially run the k -clique version to establish bounds, but it would not advance from $G[\{v_1, v_2, \dots, v_i\}]$ to $G[\{v_1, v_2, \dots, v_{i+1}\}]$ before eliminating the possibility of finding a new incumbent k -club.

through further branch and bound. This is an interesting area for future research.

This section focuses on combinatorial algorithms for finding clique relaxations. However, researchers have studied other approaches for finding k -plexes, k -cliques, and k -clubs including integer programming [12,44], fixed-parameter algorithms [15], and heuristics [45–48].

THEORETICAL RESULTS

This section briefly discusses some theoretical results related to k -plexes, k -clubs, and k -cliques. The content is limited to complexity and polyhedral results.

Concerning complexity, maximum k -club, k -clique, and k -plex are all NP-complete. This follows directly from the fact that for $k = 1$, these problems reduce to maximum clique. Moreover, k -club remains NP-complete even on graphs with diameter at most $k + 1$ [49] and has the additional property that testing maximality is NP-complete [13]. It is also difficult to test for a gap between k -clubs and $k + 1$ -clubs. The same result holds for k -cliques. Let $\bar{\omega}_k(G)$ and $\tilde{\omega}_k(G)$ denote the largest k -club and k -clique in G .

Theorem 1 [50]. *Given $k < l$, it is NP-hard to test if $\bar{\omega}_k(G) = \bar{\omega}_l(G)$ or $\tilde{\omega}_k(G) = \tilde{\omega}_l(G)$.*

One can represent the maximum k -clique problem as the maximum clique problem on the related power graph [44]. Given $G = (V, E)$, let $F = \{u, v \in V : d_G(u, v) \leq k\}$ and $G^k = (V, F)$. Clearly, a clique in G^k is a k -clique in G . Therefore, bounding the size of cliques in G^k is equivalent to bounding the size of k -cliques in G . Note also that this relationship leads to bounds on $\bar{\omega}_k(G)$ since $\bar{\omega}_k(G) \leq \tilde{\omega}_k(G)$. These ideas are further developed in Ref. 13.

Another line of research seeks to bound the size of the search tree. This approach typically relies on data reduction techniques to reduce the complexity of a problem instance. Applying these techniques can lead to results such as the following.

Theorem 2 [15]. *The search tree for testing the existence of a cardinality m*

2-plex can be bounded to have $\mathcal{O}(2.31^m)$ nodes, and maximum 2-plex can be solved in $\mathcal{O}(k^{5/2}2.31^k + kn)$ time.

Integer programming techniques offer another approach to finding maximum k -plexes, k -cliques, and k -clubs. These formulations also induce an associated class of polyhedra. Given $u, v \in V$, let C_{uv}^k be the set of all $u-v$ paths P in G such that $|E(P)| \leq k$. Let C be the set of all paths in G and V_t be the vertex set of path t . Maximum k -club can be formulated as follows [44]:

$$\begin{aligned} \text{Max.} \quad & \sum_{v \in V} x_v \\ \text{s.t.} \quad & x_v + x_u - \sum_{t \in C_{uv}^k} y_t \leq 1 \quad \forall u, v \in V \\ & y_t - x_v \leq 0 \quad \forall t \in C; v \in V_t \\ & x_v, y_t \in \{0, 1\} \quad \forall v \in V; t \in C. \end{aligned}$$

Notice that the first constraint reduces to $x_v + x_u \leq 1$ whenever $C_{uv}^k = \emptyset$. The y variables prevent the consideration of any path not entirely included in the induced subgraph. For the k -clique problem, one can eliminate the y variables to obtain the following formulation [49]:

$$\begin{aligned} \text{Max.} \quad & \sum_{v \in V} x_v \\ \text{s.t.} \quad & x_v + x_u \leq 1 \quad \forall u, v \in V \text{ s.t. } C_{uv}^k = \emptyset \\ & x_v \in \{0, 1\} \quad \forall v \in V. \end{aligned}$$

The k -club formulation was introduced in Ref. 44. The k -clique formulation and a polyhedral analysis of the 2-club polytope were given in Ref. 49. The facial structure of both the k -clique and k -club polytopes offer an interesting topic for future research. Let $N[v] = \{u \in V : uv \in E\} \cup \{v\}$ and $\bar{N}(v) = V \setminus N[v]$. Maximum k -plex can be formulated as follows [12]:

$$\text{Max.} \quad \sum_{v \in V} x_v$$

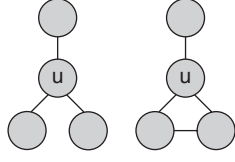


Figure 1. Minimal 2-claws.

$$\begin{aligned} \text{s.t.} \quad & \sum_{u \in \bar{N}(v)} x_u \leq (k-1)x_v + |\bar{N}(v)|(1-x_v) \\ & \forall v \in V \\ & x_v \in \{0, 1\} \quad \forall v \in V. \end{aligned}$$

The facial structure of this polytope was first examined in Ref. 12, and a further investigation was reported in Ref. 14. A successful approach to analyzing this polytope involves generalizing well-known inequalities, such as cycles and webs [51], from the related stable set polytope. Finding further inequalities is an interesting area for future work.

One can also use 2-plexes to characterize a class of integral polytopes. Given G , define the 2-plex clutter matrix A so that the rows of A are the incidence vectors of all maximal 2-plexes in G . Define the minimal 2-claws to be the graphs shown in Fig. 1.

Theorem 3 [14]. *If A is a 2-plex clutter matrix for G , then $\{x \in \mathbf{R}_+^n \mid Ax \leq 2, x \leq 1\}$ is integral if and only if G contains no minimal induced 2-claw or induced cycle C^n where $n \not\equiv 0 \pmod{3}$.*

Since it describes integral polyhedra in terms of excluding induced subgraphs, this result can be viewed as a polyhedral 2-plex analogue to graph perfection. However, combinatorial and polyhedral notions of 2-plex perfection do not seem to coincide. This theorem also implies the existence of a polynomial time algorithm for testing the integrality of polytopes defined by 2-plex clutter matrices. It is unknown if one can test whether A is a k -plex clutter matrix in polynomial time. It is also an interesting open problem to determine if k -plexes can characterize integral systems when $k > 2$.

This section offered a brief survey of results related to k -clique, k -clubs, and k -plexes. These results show that the theoretical aspects of clique relaxations have become an active area of research.

CONCLUSION

The clique concept is useful in many applications, but the definition of clique can be overly restrictive. The clique approach can fail to detect much of the structure present in a graph. This article discussed various ways of relaxing the clique concept, and it focused on k -plexes, k -clubs, and k -cliques.

The section titled “Combinatorial Algorithms” described a class of combinatorial algorithms by adapting a general clique algorithm to find k -plexes, k -clubs, and k -cliques. Such algorithms can benefit from the discovery of new bounding and pruning techniques. This line of work generally involves the development of heuristics, data reduction techniques, and preprocessing methods. These topics are an interesting area for future work.

The section titled “Theoretical Results” reviewed some recent complexity results, presented three integer programming formulations, and discussed polyhedral results. From a complexity standpoint, future work includes the identification of instances that admit polynomial time solutions. There also remains a great deal of work to be done on the study of k -plex, k -club, and k -clique polytopes, including the identification of strong inequalities and the recognition of polyhedra with simple structure.

REFERENCES

1. Turan P. On an extremal problem in graph theory. *Mat Fiz Lapok* 1941;48:436–452.
2. Garey MR, Johnson DS. *Computers and intractability: a guide to the theory of NP-completeness*. New York: W.H. Freeman and Company; 1979.
3. Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors. *Complexity of computer computations*. New York: Plenum; 1972. pp. 85–103.

4. Alon N, Krivelevich M, Sudakov B. Finding a large hidden clique in a random graph. *Random Struct Algorithms* 1998;13(3-4): 457–466.
5. Bollobas B, Erdős P. Cliques in random graphs. *Math Proc Camb Phil Soc* 1976;80(3):419–427.
6. Berge C. Perfect graphs. Six papers on graph theory. Calcutta: Indian Statistical Institute; 1976. pp. 419–427.
7. Chudnovsky M, Robertson N, Seymour P, *et al.* The strong perfect graph theorem. *Ann Math* 2006;164(1):51–229.
8. Lovász L. Normal hypergraphs and the perfect graph conjecture. *Disc Math* 1972;2:253–267.
9. Balasundaram B, Butenko S. Graph domination, coloring and cliques in telecommunications. In: Resende MGC, Pardalos PM, editors. *Handbook of optimization in telecommunications*. New York: Springer; 2006. pp. 865–890.
10. Bomze IM, Budinich M, Pardalos PM, *et al.* The maximum clique problem. In: Du D-Z, Pardalos PM, editors. *Handbook of combinatorial optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999. pp. 1–74.
11. Butenko S, Wilhelm W. Clique-detection models in computational biochemistry and genomics. *Eur J Oper Res* 2006;173:1–17.
12. Balasundaram B, Butenko S, Hicks IV. Clique relaxations in social network analysis: the maximum k-plex problem. *Oper Res*. 2010. To appear.
13. Mahdavi F, Balasundaram B. On inclusion-wise maximal and maximum cardinality k-clubs in graphs. Submitted for publication.
14. McClosky B, Hicks IV. The co-2-plex polytope and integral systems. *SIAM J Disc Math* 2009;23(3):1135–1148.
15. Moser H, Niedermeier R, Sorge M. Algorithms and experiments for clique relaxations - finding maximum s-plexes. Volume 5526, *Proceedings of the 8th International Symposium on Experimental Algorithms (SEA'09), Lecture Notes in Computer Science*. Dortmund: Springer; 2009. pp. 233–244.
16. Lloyd AL, Valeika S, Cintron-Arias A. Infection dynamics on small-world networks. In: Gumel AB, Castillo-Chavez C, Clemence DP, editors. *Mathematical studies on human disease dynamics: emerging paradigms and challenges*. Contemporary Mathematics 410. Providence (RI): American Mathematical Society; 2006. pp. 209–234.
17. Smieszek T, Fiebig L, Scholz RW. Models of epidemics: when contact repetition and clustering should be included. *Theor Biol Med Model* 2009;6:11.
18. Glass LM, Glass RJ. Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health* 2008;8:61.
19. Hill S, Provost F, Volinsky C. Network-based marketing: Identifying likely adopters via consumer networks. *Stat Sci* 2006;22(2): 256–275.
20. Iacobucci D, Hopkins N. Modelling dyadic interactions and networks in marketing. *J Mark Res* 1992;24(1):5–17.
21. Woodside AG, DeLozier MW. Effects of word of mouth advertising on consumer risk taking. *J Adv* 1976;5(4):12–19.
22. Raghuram I, Han S, Gupta S. Do Friends Influence Purchases in a Social Network?. Harvard Business School Working Paper, No. 09-123. 2009.
23. Kempe D, Kleinberg J, Tardos E. Influential nodes in a diffusion model for social networks. *Proceedings of 32nd International Colloquium on Automata, Languages and Programming (ICALP)*. Lisboa, Portugal; 2005.
24. Leskovec J, Singh A, Kleinberg J. Patterns of influence in a recommendation network. *Adv Knowl Discov Data Min* 2006;380–389.
25. Batagelj S, Mrvar A. Pajek datasets. 2006. Available at <http://vlado.fmf.uni-lj.si/pub/networks/data/>.
26. Grossman J, Ion P, De Castro R. The Erdős Number Project. 1995. Available at <http://www.oakland.edu/enp>.
27. Goffman C. And what is your Erdős number? *Am Math Mon* 1969;76:791.
28. Diestel R. Graph theory. Volume 173, *Graduate texts in mathematics*. Heidelberg: Springer; 2005.
29. Seidman SB, Foster BL. A graph theoretic generalization of the clique concept. *J Math Soc* 1978;6:139–154.
30. Seidman SB. Network structure and minimum degree. *Soc Netw* 1983;5:269–287.
31. Alba RD. A graph-theoretic definition of a sociometric clique. *J Math Soc* 1973;3: 113–126.
32. Luce RD. Connectivity and generalized cliques in sociometric group structure. *Psychometrika* 1950;15:169–190.
33. Abello J, Resende MGC, Sudarsky S. Massive quasi-clique detection. In: Rajsbaum S, editor.

- LATIN 2002: theoretical informatics. London: Springer; 2002. pp. 598–612.
34. Jiang D, Pei J. Mining frequent cross-graph quasi-cliques. Volume 2(4), ACM Transactions on Knowledge Discovery in Data. New York: ACM Press; 2009. pp. 16:1–42.
35. Mader W. Existenz n -fach zusammenhängender Teilgraphen in Graphen genügend grosser Kantendichte. Abh Math Sem Univ Hamburg 1972;37:86–97.
36. Goldberg AV. Finding a maximum density subgraph. UC Berkeley report UCB/CSD/84/171; 1984.
37. Lawler EL. Combinatorial optimization: networks and matroids. New York: Holt, Rinehart and Winston; 1976.
38. Feige U, Kortsarz G, Peleg D. The dense k -subgraph problem. Algorithmica 2001;29: 410–421.
39. Balas E, Xue J. Weighted and unweighted maximum clique algorithms with upper bounds from fractional coloring. Algorithmica 1996;15:397–412.
40. Östergard PRJ. A fast algorithm for the maximum clique problem. Disc Appl Math 2002;120:192–207.
41. Tomita E, Seki T. An efficient branch-and-bound algorithm for finding a maximum clique. Lect Notes Comput Sci Ser 2003; 2731:278–289.
42. Wood DR. An algorithm for finding a maximum clique in a graph. Oper Res Lett 1997;21:211–217.
43. Trukhanov S, Balasundaram B, Butenko S. Generalization of Ostergard’s algorithm and an application to the maximum weight k -plex problem. In press.
44. Bourjolly JM, Laporte G, Pesant G. An exact algorithm for the maximum k -club problem in an undirected graph. Eur J Oper Res 2002;138:21–28.
45. Brandenburg F, Edachery J, Sen A. Graph clustering using distance- k cliques. Lect Notes Comput Sci 1999;1731:98–106.
46. Bourjolly JM, Laporte G, Pesant G. Heuristics for finding k -clubs in an undirected graph. Comput Oper Res 2000;27:559–569.
47. Mahdavi F, Balasundaram B. A variable neighborhood search heuristic for k -clubs in graphs; 2009. Manuscript.
48. McClosky B, Hicks IV. Combinatorial algorithms for max k -plex; 2008. In press.
49. Balasundaram B, Butenko S, Trukhanov S. Novel approaches for analyzing biological networks. J Comb Optim 2005;10:23–39.
50. Butenko S, Prokopyev O. On k -club and k -clique numbers in graphs. J Comb Optim 2005;10:23–39.
51. Trotter LE. A class of facet producing graphs for vertex packing polyhedra. Disc Math 1975;12:373–388.

CLOSED-LOOP SUPPLY CHAINS: ENVIRONMENTAL IMPACT

JACQUELINE M. BLOEMHOF
Wageningen University, BH
Wageningen, The Netherlands

CHARLES J. CORBETT
UCLA Anderson School of
Management, Los Angeles,
California

The context referred to under the heading “closed-loop supply chains” has, broadly speaking, two main origins. One deals with commercial returns, that is, products returned by customers for commercial reasons relatively soon after initial purchase; for instance, because the customer did not like the product, or because the product was defective. The other deals with end-of-life or end-of-use returns, where customers return products because they have reached the end of their technical life (for instance, when they become defective), or when they have reached the end of their economic life (for instance, when the customer decides to upgrade to a new model, or when a lease contract expires).

In this article, we focus on environmental impacts in three types of supply chains: (i) the traditional forward supply chains (in which disposal and return of materials are ignored), (ii) the extended forward supply chains (which consider disposal but no form of reprocessing), and (iii) environmental impacts in full closed-loop supply chains (which include forward chains and reprocessing of materials). Environmental impacts are also relevant in a forward supply chain, as the materials extracted from nature return to nature in the form of emissions or landfill and hence, form a closed loop in the natural cycle. Although other fields extensively discuss environmental impacts of systems, they rarely focus explicitly on how modifications in supply chain structures

or processes can improve environmental performance of the system.

The term *closed-loop supply chain* is used loosely in practice, referring either to “closed-loop” in the strict sense of the word, where previously used materials return to the original supply chain to be reused (in some form) in the same supply chain as before, or to “open-loop” systems, in which materials are taken back from customers but not necessarily reprocessed in the same supply chain as before. Reusable cameras that are taken back, reprocessed, and reintroduced as cameras into the same supply chain as before, are an example of a closed-loop supply chain; rubber tires, which are taken back after use and then reprocessed into road-surfacing material, are an example of an open-loop supply chain. We use the broad interpretation of “closed-loop” here.

The primary objective of closed-loop supply chain management is usually firm profitability [1]. That is unambiguous in the case of commercial returns, which were a business issue well before environmentally driven returns [2]. The environmental imperative is a major driver of the rapid growth in interest in closed-loop supply chains, hinging on the assumption that taking back and reprocessing materials is environmentally beneficial. Despite this environmental context, though, the vast majority of the closed-loop supply chain literature assumes that the objective is to increase the amount of materials taken back, or, alternatively, to minimize the cost of taking back materials (which presumably will in turn, increase the amount of materials taken back). Several studies indicate that that assumption does not always hold. For instance, Bloemhof-Ruwaard *et al.* [3] discuss how Scandinavian legislation requiring a certain proportion of paper to be recycled led to the environmentally undesirable outcome of wastepaper being shipped from the Netherlands and Germany to Scandinavia for recycling (refer also to **Recycling**). Similarly, Mayers *et al.* [4] shows that the exclusive focus on increasing take-back

embodied in the European Union's waste electrical and electronic equipment (WEEE) directive can be undesirable from a cumulative energy demand (CED) perspective.

The aim of this article is to introduce and exemplify the importance of environmental impact in closed-loop supply chain management. Many surveys exist of closed-loop supply chains (or green supply chains, or reverse logistics) [5–9]; for example, a special issue planned to appear in the journal *Production and Operations Management* in 2010, and others. However, the papers mentioned in these surveys do not explicitly deal with environmental impacts in a supply chain context, but instead use for example units returned as a measure for environmental impact. Here, we only focus on the explicitly measured environmental impacts in a supply chain context, in basic units such as megawatt hours, tons of CO₂, or others. We build our discussion on a limited set of articles that explicitly measure environmental impacts. It has been argued (see for instance Ref. 10 and the references therein) that including a broader set of metrics can help firms to find opportunities to improve financial performance of their supply chains that they may otherwise have missed. The close link between greenhouse gas emissions and fuel consumption provides many of the most immediate examples of firms achieving significant cost savings as a result of environmental objectives, but many less direct examples of this same mechanism also exist.

We start with a brief discussion of environmental impact measurement in general and provide resources for measuring environmental impacts. The structure of the rest of this article loosely follows the chronological development of environmental impact measurement in the supply chain context: first, we look at studies of environmental impacts in traditional forward supply chains (in which disposal and return of materials are ignored), then we turn to extended forward supply chains (which consider disposal but no form of reprocessing), after which we discuss environmental impacts in full closed-loop supply chains (which include forward chains and reprocessing of materials). The papers we discuss are intended primarily as

illustrations of how environmental impacts can be measured in closed-loop supply chains, rather than aiming to be a tutorial on how to perform such measurements.

MEASURING ENVIRONMENTAL IMPACTS

Many measures of environmental impacts exist, and no single measure unambiguously captures all impacts. Many studies based on life cycle assessment (LCA) methods consider the following seven main categories of environmental impacts: global warming potential (GWP), resource depletion, acidification, eutrophication, tropospheric ozone formation, ozone depletion, and human toxicity, though many of these in turn can be divided into subcategories. The Millennium Ecosystems Assessment (www.millenniumassessment.org, last accessed December 31, 2008) examines the state of 24 “ecosystem services” provided by the environment, including “provisioning services” (several types of food, fibers, genetic resources, biochemicals, natural medicines and pharmaceuticals, and fresh water), “regulating services” (including air quality regulation, global climate, local and regional climate, water regulation, erosion regulation, water purification and waste treatment, disease regulation, pest regulation, pollination, and natural hazard regulation), and “cultural services” (including spiritual and religious values, aesthetic values, and recreation and ecotourism).

In determining the environmental impacts of a closed-loop supply chain, we often want to know not only the direct impacts associated with a product, but also indirect ones, throughout its entire life cycle. This calls for an LCA, the process of defining a functional unit (unit of analysis) and scoping the system (setting the boundaries of impacts to be included), performing a life cycle inventory (LCI) of impacts associated with one functional unit within those boundaries, and then characterizing and interpreting those environmental impacts. For instance, to determine the environmental impacts of the printer supply chain, we would need to know the impacts of the

various mining and processing operations that yield the materials that go into the printer, but also (among many others) those that go into the packaging materials used at various stages in the forward and reverse part of the supply chain. A key decision in any LCA study is determining the scope of the study. For instance, a truly complete LCA would even require knowing the life cycle impacts associated with the production of the trucks used to transport the printers. In most practical cases, many of those indirect impacts can safely be ignored. Setting the boundaries of an LCA study appropriately therefore requires some knowledge of relative magnitudes of the impacts concerned. For that reason, some basic environmental numeracy is essential for scholars who study environmentally motivated problems.

Increasingly, researchers seek to simplify their analyses by aggregating impacts in the disparate categories mentioned above into a single metric, to facilitate further analysis and decision making. One such metric is cumulative fossil energy demand (or fossil CED), which aggregates primary energy used in all stages of a closed-loop supply chain. Given that use of fossil fuel is a major cause of global warming, and of depletion of fossil fuel supplies, CED is one metric that captures several of the key environmental impact categories. Huijbregts *et al.* [11] found that fossil CED also correlates well with most other impact categories, and hence can be a useful primary screening indicator, although they do caution that the high uncertainty in product-specific impact scores makes it difficult to use CED as a standalone impact measure.

A more complete summary representation of LCA results can be given in the form of a materials, energy, toxicity (MET) matrix, which shows the most important impacts in each of those three categories (the columns) associated with production, use, and disposal (the rows). Alternatively, it is common to select a few key environmental performance indicators (EPIs) that are particularly relevant for the given context and focus on those. A range of other tools, such as ecological

footprint (EF), environmental impact assessment (EIA), material flow accounting (MFA), material intensity per unit service (MIPS), and others, are discussed in Ref. 12.

RESOURCES FOR MEASURING ENVIRONMENTAL IMPACTS

Here we introduce some of the LCA approaches used in this field and provide pointers to resources for researchers who aim to include environmental impact measurements in their work.

Two common methods for conducting LCA studies are process-based LCA and economic input–output LCA. Process-based LCA involves specifying the stages of the system under consideration in detail, including all in- and out-flows of material and energy, and then determining (often using existing databases) the corresponding environmental impacts of each element of the process. Several software packages for process-based LCA are widely used in practice, including SimaPro (<http://www.pre.nl/simapro>, last accessed December 31, 2008), GaBi (<http://www.gabi-software.com>, last accessed December 31, 2008), and Umberto (<http://www.umberto.de/en/>, last accessed December 31, 2008). Although one can perform highly detailed LCA studies with these systems, they are relatively straightforward to use and are quite suitable for new users too.

Economic input–output life cycle assessment (EIO-LCA) involves specifying the economic system one is interested in, and modeling all economic flows into and out of that system, and then converting those flows into environmental impacts. A commonly used resource for EIO-LCA is the www.eiolca.net website maintained by the Green Design Initiative at Carnegie-Mellon University (www.eiolca.net). This relies on input–output analysis, and hence includes all direct and indirect environmental impacts associated with a certain volume of final product. One can choose, for instance, “tire manufacturing” from a drop-down menu of product types, and instantly learn that the total GWP associated with final sale of \$1 million worth

of rubber tires is 1090 metric tons of CO₂ (as of December 31, 2008), which can be further broken down into CO₂, N₂O, and other greenhouse gases; similarly for conventional air pollutants, energy, and toxic releases. Note that this includes emissions not just caused by tire manufacturers themselves, but also by their first, second, and higher tier suppliers.

Process-based LCA and EIOLCA both have advantages and disadvantages that are discussed in detail in the references below. Many good resources exist to learn more about environmental impact measurement and LCA:

- SETAC (The Society of Environmental Toxicology and Chemistry, see <http://www.setac.org>, last accessed January 6, 2009) provides a forum where scientists, managers, and other professionals exchange information and ideas for the development and use of multidisciplinary scientific principles and practices leading to sustainable environmental quality.
- The Institute of Environmental Sciences (CML) at the University of Leiden in the Netherlands is often considered one of the origins of LCA; their website (<http://www.leidenuniv.nl/cml/ssp/index.html>, last accessed December 30, 2008) provides some useful perspectives.
- For an extensive list of resources on LCA including a discussion of existing software, see <http://www.epa.gov/nrmrl/lcaccess/index.html> (last accessed December 30, 2008). Guidelines for conducting LCA studies include the ISO 14040:2006 and 14044:2006 standards.
- A site that focuses on LCA in a supply chain context is www.supplychainlca.com. Online courses on LCA are available from the following web sites:
- Royal Melbourne Institute of Technology (using SimaPro; see <http://simapro.rmit.edu.au>, last accessed December 30, 2008);
- Harvard School of Public Health (see www.sciencenetwork.com/lca, last accessed December 30, 2008);
- Professor Dave Allen at the University of Texas, Austin (see <http://www.utexas.edu/research/ceer/greenproduct> and <http://www.utexas.edu/research/ceer/greenmaterial>, last accessed December 30, 2008).

In what follows, we discuss several studies that measure environmental impacts in supply chains, primarily to illustrate the range of ways in which environmental impacts can be measured and how such information is used in the closed-loop supply chain context. There is a vast literature on applications of LCA which, by definition, have a supply chain element; here, we select a few studies that illustrate the range of uses of LCA in supply chains.

ENVIRONMENTAL IMPACTS IN FORWARD SUPPLY CHAINS

We start with some examples of earlier LCA studies in supply chains that do not explicitly consider disposal or recycling.

Table 1 summarizes the main research in this category. All papers make use of LCA, using a multifaceted index to measure the environmental impact in an industry, focusing on fuel choice and product and process design in process industries and transportation.

LCA studies invariably lead to environmental impacts, in addition to economic considerations. Hence, combinations of LCA with OR (operations research) techniques that facilitate dealing with multiple objectives are natural. In one of the earliest studies combining OR and LCA, Bloemhof-Ruwaard *et al.* [13] use LCA, linear programming (see section titled “Linear Programming (LP)” in this encyclopedia) and multicriteria decision making (MCDM) to compare environmental impacts of various fat blends. MCDM facilitates combining the environmental impact measures into a single index, while LP then allows the authors to determine the relative cost increase and environmental impact decrease that can be obtained with several environmentally optimized fat blends. Azapagic and Clift [14]

Table 1. Environmental Impacts in Forward Supply Chains

References	Research Questions	Solution Method	Application
13	How to compare environmental impacts of blends	LP + LCA + MCDM	Process industry (food)
14	How to include environmental impacts in process optimization	LP + LCA + MODM	Process industry (mining and minerals)
15	How to apply OR/MS tools to efficiently incorporate environmental legislation into production processes	NFM + nonlinear integer model	Process industry (copper)
16	How to compare environmental impact across various efficiency scenarios	LCA	Process industry (aluminum)
17	What are environmentally preferred vehicle options	LCA	Transportation (vehicles)
18	How to improve eco-efficiency at the fleet level	Eco-efficiency measure	Transportation (fleet)

apply a combination of LCA and multiobjective optimization to the boron system. This includes mining (for borax and kernite) and various processing steps leading to several boron products, and considers the seven impact categories mentioned before. As there are many ways to operate this system, a linear programming model is formulated to take into account market demand constraints, availability of primary and raw materials, production capacity, and heat requirements.

A nonlinear integer model combined with a network flow model is used by Caldentey and Mondschein [15] to design an optimal supply chain for the smelting and sulfuric acid production stages in the copper industry. Optimizing the entire system and allowing the market price for sulfuric acid to emerge endogenously rather than be imposed exogenously, enables the copper industry to earn substantially higher profits. It also allows decision makers to estimate the cost of government policies, such as new standards for arsenic emissions.

Tan and Khoo [16] report on an LCA study of the primary aluminum supply chain, consisting of mining, refining, smelting, casting, and power generation. The authors calculate GWP, acidification, human toxicity for air, resource consumption, and bulk wastes. They examine a base case, and scenarios with

progressively higher efficiency, starting with the case in which the casting plant reduces scrap metal by 20%, then one in which in addition the smelter implements more sustainable practices, and a final case in which the refinery also reduces bulk waste and the power plant uses clean coal technology.

As transportation is a major source of environmental impacts, many studies examine a range of transportation systems. Van Mierlo *et al.* [17] compare life cycle impacts of vehicles using two different methodologies: EcoScore (which attaches weights to impacts with respect to global warming, human health: respiration and cancer, ecosystems: acidification, buildings, and noise), and Cleaner Drive (which focuses on global warming and air quality). Although the two methodologies exhibit some differences, the insights they yield are largely consistent and both point toward battery-electric, hybrid and compressed natural gas (CNG) vehicles as the environmentally preferred options, with liquid petroleum gas (LPG) scoring fairly well, and petrol and diesel vehicles displaying wide variety in ratings. D'Agosto and Ribeiro [18] analyze improvement opportunities at the fleet level rather than for individual vehicles. They use the eco-efficiency measure (product or service value divided by environmental influence) as their basis for analysis, and focus on

CO₂ emissions from diesel oil, gasoline, ethanol, and CNG. Applying their approach to the fleet operations of Rio de Janeiro International Airport, they define landing and take-off cycles as the driver of service value, and then compare various scenarios, including converting all vehicles to CNG, implementing an eco-efficiency management plan, and replacing gasoline and CNG vehicles with hydrated ethanol.

It is not surprising to see much of the LCA-based work on forward supply chains focusing on the process industries and transportation sector. Both are a major source of emissions, so it is already worthwhile studying environmental impacts within this more limited scope, excluding extended and reverse supply chains. In the remainder, we see a continued focus on process industries and transportation, but also more consideration of industries where excluding extended and even reverse supply chains is less justifiable.

ENVIRONMENTAL IMPACTS IN EXTENDED FORWARD SUPPLY CHAINS

Progressively, authors have extended the boundaries of the supply chain to include end-of-life disposal. The first studies examine energy systems, where a good understanding

of the full energy implications of various materials and processing choices is of course essential. The remaining studies concern food supply chains, the electronics supply chain, and the vehicle life cycle.

Table 2 shows the three main application areas in which environmental impacts have been studied in extended supply chains: biomass, process industry, and transportation. The main issue in biomass research is to find economically and environmentally sound supply chain designs for renewable energy carriers. The main issue in the other areas of application is to assess the relative environmental impacts of all stages in the supply chain.

In view of the strong interest in biofuels as a source of renewable energy, especially when the biomass is derived from agricultural waste, a common question concerns the life cycle impacts of using biofuels when long-distance transportation of the biomass is required. Focusing on the functional unit of 1 MW of electricity delivered to the grid in an importing country, Forsberg [19] compares life cycle impacts in terms of energy expenditure (from electricity and fuel) and air emissions of various forms of biomass energy systems. Specifically, Forsberg [19] compares exporting forest residues after

Table 2. Environmental Impacts in Extended Forward Supply Chains

References	Research Questions	Solution Method	Application
19	What is the life-cycle impact of biofuels?	LCA	Biomass energy system
20	What are economically and environmentally sound ways to transport biomass?	LCA + LCC	Biomass supply chain
21	How to minimize costs, energy, and global warming of a supply chain design?	Simulation	Biomass supply chain
22	Which supply chain process contributes to which environmental impact?	Scenario analysis	Food supply chain
23	What is the environmental impact of the processes in the supply chain?	LCA	Food supply chain
24	What is the environmental impact of the processes in the supply chain?	LCA + MODM	Electronics
25	What is the trade-off between extending lifetime or frequent replacement?	LCA	Vehicle and fuel supply chain

making into bales or pellets in the form of tree sections, or generating and exporting the electricity derived from the biomass, and domestic use of the energy produced, and finds that, provided appropriate technology is used, biomass can be used to transport energy over long distances.

A more comprehensive analysis, including economic considerations, is reported in Ref. 20. They compare various supply chains for biomass, including transportation options for raw biomass (logs, bales, chips), for refined biomass (pellets, pyrolysis oil), and liquid biofuels (methanol) with the original biomass coming from Scandinavia, eastern Europe, or Latin America and being transported to western Europe. They also find that biomass can be transported over long distances in ways that are economically and environmentally sound, and provide recommendations for how to design the supply chain, specifically in which form the biomass should be transported and hence, which processing stages should take place close to where the biomass originates and which should be located close to where the energy is consumed. Sokhansanja *et al.* [21] develop an integrated simulation model to determine the cost, energy, and global warming implications of various supply chain designs defining how biomass is collected, stored, and transported to a processing facility.

The energy used in the food supply chain is often underestimated, as the following studies on milk and poultry illustrate. Soneson and Berlin [22] examine various scenarios for the extended supply chain for milk in Sweden, in the light of several metrics. Transport from retailers to households is easily the largest contributor to global warming, transport to the dairies is the largest contributor to NO_x , waste management has the largest impact in terms of eutrophication, and packaging production usually has the largest impact in terms of acidification. The results for formation of petrochemical oxidants vary more heavily across the scenarios. Pelletier [23] examines the multiple supply chains involved in poultry production, including the degree to which poultry litter displaces fertilizer production, and finds

that 80–98% of environmental impacts are associated with feed production rather than on-farm emissions, which is consistent with other studies cited therein on pork, beef, salmon, and milk.

One of the biggest drivers of interest in closed-loop supply chains is the WEEE Directive adopted by the European Union in 2003, which holds producers responsible for organizing and financing take-back, treatment, and recycling of WEEE. Mayers *et al.* [4] examine, for the case of recycling of HP printers in the United Kingdom, the impacts along seven dimensions of four different scenarios: 100% landfilling, partial or full recycling with no energy recovery from burning the plastics, and 99% recycling with energy recovery. The results are quite mixed: no scenario dominates, and higher recycling or energy recovery rates lead to worse environmental outcomes on some dimensions. Quariguasi Frota Neto *et al.* [24] also use the WEEE directive to illustrate a methodology for assessing eco-efficiency. They choose CED as their multifaceted environmental index in contrast to the single-faceted measure of waste diverted from landfill due to the WEEE directive, and use an optimization model to obtain iso-cost curves depicting the trade-off between CED (in MJ per year) and landfilled waste (in tons per year).

Using the eco-indicator 99 method, balancing the burdens associated with the vehicle and fuel supply chains, Spielmann and Althaus [25] finds that Swiss drivers should extend the life of their cars rather than replace them more frequently in order to benefit from increased fuel efficiency, though some of the results are sensitive to their specific assumptions.

Biomass is heavily represented in this category of extended forward supply chains, relative to the earlier category of forward supply chains, as the whole point of biomass supply chains is precisely in the final stage, where material is converted to energy. While the earlier section focused primarily on process choice and fuel choice, we see an increasing focus on product design and lifetime

extension in this category. The final category will consider more comprehensive supply chain issues, including both process and product design.

ENVIRONMENTAL IMPACTS IN CLOSED-LOOP SUPPLY CHAINS

The final set of studies we discuss consider environmental impacts in entire closed-loop supply chains, including various forms of reprocessing, whether in the form of reuse, recycling, or other. The first studies examine bulk materials (lumber, paper and pulp, packaging, and steel); the subsequent studies examine more complex products (photocopiers, refrigerators, and various electrical and electronic products).

Table 3 shows that the research questions in this category all deal with a combination of improvements that can be made in the forward chain (lifetime extension, design, incineration) and in the reverse chain (recycling, remanufacturing, reuse). As the supply chains under investigation are complex and

both environmental and economic objectives have to be considered, the solution methods are a combination of LCA and Operations Research/Management Science optimization models.

Sathre and Gustavsson [26] examine the global warming effects of various uses for recovered wood lumber (reuse as lumber, reprocessing as particleboard, pulping, and energy recovery). They find that the degree to which forest land is constrained, and hence the degree to which it can be put to alternative uses, is a major factor in determining which cascade is preferred.

Bloemhof-Ruwaard *et al.* [3] model the European paper and pulp supply chain using a linear network flow model, distinguishing various pulping and bleaching techniques and allowing relocation of paper production. They compare a number of regulatory scenarios in terms of the seven main environmental categories and a single environmental index. Quariguasi Frota Neto *et al.* [27] build on this work and also finds that mandatory recycling targets can have limited or even counterproductive

Table 3. Environmental Impacts in Closed-Loop Supply Chains

References	Research Questions	Solution Method	Application
26	What are the global warming effects of uses for recovered product?	Carbon balance	Natural resources (wood lumber)
27	What is the environmental impact of an integrated closed-loop supply chain?	Linear programming	Process industry (paper and pulp)
28	What is the trade-off between single-use and reusable packaging?	EIOLCA	Packaging
29	What are the economic and environmental impacts of reuse constraints?	LCA + LCC	Metals (steel)
30	What is the best option: recycling, incineration or landfilling?	LCA	Packaging
31	What is the best option: recycling or disposal?	Various LCA methods	Lead-acid batteries
32	What is the trade-off between integrated closed-loop networks and sequential forward and reverse networks?	MILP	Paper, electronics
33	What are the trade-offs between cost, waste, and energy use?	MILP	Electronics
34	What is the best option: product design, lifetime extension or remanufacturing?	LCA	Electronics

environmental effects. Matthews [28] illustrates use of the EIOLCA method in a case study at Quantum Corporation comparing a conventional single-use packaging with a packaging designed for reuse, and finds that, even taking the increased transportation into account, reusable packaging can lead to lower costs and substantial improvements in energy usage, GWP, and various air pollution measures.

Quite some work has been done on comparing options for reuse. Geyer and Jackson [29] compare various options for reuse of steel (recovering steel sections through deconstruction and reusing them, or recovering steel via demolition and recycling, or landfill) in terms of life cycle cost and energy use. They find that constraints on reuse have limited impact on environmental performance but significant impact on economic performance. Björklund and Finnveden [30] survey a collection of LCA case studies that compare recycling, incineration, and landfill, finding that for most of the products they consider (glass, metals, plastics, paper, cardboard) the ranking between the three options is often quite clear. Daniel *et al.* [31] carried out an LCA study on alternative end-of-life scenarios for used starter batteries. They focus on the question of which methodology to choose in the impact assessment (IA) step of the LCA. They concluded that the effectiveness of the LCA is improved if as many IA methods as possible are used in a study, so decision makers can choose the method which best fits the specific features of the system examined.

Fleischmann *et al.* [32] compare networks in which the forward and reverse flows are optimized sequentially with those in which both flows are optimized simultaneously. In the case of copiers, where production facilities tend to be relatively close to the markets, a reverse flow can be added to an existing forward network with few complications. In contrast, in the paper industry, production locations are typically located close to natural resources (raw materials) and far from customer markets, so adding a reverse flow prompts a drastically different network design.

Krikke *et al.* [33] modify a case study of a Japanese company's supply chain for refrigerators in Europe to examine trade-offs between cost, waste, and energy use, finding that (for their scenarios) the supply chain network has the most impact on cost, while product design has the most impact on energy use and waste. Moreover, they find that recovery targets based on proposed legislation have an ambivalent impact; they reduce waste but increase energy use and costs. Quariguasi Frota Neto *et al.* [34] explore which phase in a supply chain produces the largest life cycle environmental impacts (measured by CED), in order to determine whether one should focus on improving product design (to increase energy efficiency), longer product life cycles, or more remanufacturing (See **Recycling**). For the two relatively high-tech electronic products they consider (PCs and mobile phones), the CED is driven primarily by the manufacturing stage, so that product life extension and reuse and remanufacturing of components provide the largest environmental improvements. Conversely, for the three relatively low-tech electrical products they consider (washing machines, refrigerators, and televisions), the main impacts are in the usage phase, so that focusing on energy-efficient designs has more promise.

Summarizing, we see that despite the complexity of the supply chains considered, a combination of existing LCA and OR/MS methods can yield valuable insights into optimal design of products and of forward and reverse supply chains.

CONCLUSIONS

From the studies reviewed here, several observations emerge. A wide range of approaches and methodologies exist for combining actual environmental impact measurement in models of closed-loop supply chains, and they yield valuable insights into the key trade-offs between various environmental and economic metrics. On the other hand, many of the findings about which stage in a supply chain causes higher impacts on any given metric are often specific to that metric and to that context, making

it challenging to draw general conclusions about which metrics or supply chain stages to focus on. We believe the rich and rapidly growing set of tools available to scholars today to incorporate explicit environmental impact measures into their models of closed-loop supply chains makes this a promising and crucial area for future growth of the field.

This article aims to show that including explicit environmental impact measurements in models of closed-loop supply chains is not only feasible, it is in fact considerably easier to do than is often assumed. With the examples provided by the papers discussed here, and the resources referred to above, we hope that scholars of closed-loop supply chains will attempt to include more complete environmental impact measures in their models rather than focusing exclusively on the number of units reprocessed. Even if a full LCA study is usually far beyond the scope of most closed-loop supply chain models, the materials summarized here provide examples of how, for any given context, one can obtain preliminary assessments of which stages in a supply chain dominate in terms of specific impact categories.

For products with high environmental impacts during production, extension of product lifetime will reduce environmental impact per unit of time. Therefore, strategies such as repair, refurbishing, and remanufacturing are environmentally beneficial. However, this is very product-specific, as sometimes the higher energy efficiency of new equipment can make it undesirable to extend the life of old equipment. In any event, at the end of its useful life, every product can be recycled in some form. For some products, however, the proportion of virgin materials and energy that can be reclaimed by recycling is very small compared to the effort that is needed (in transportation, shredding, reclaiming, etc.), so that total environmental impact improvement can even be negative. Therefore, the key prerequisite for environmental improvements in closed-loop supply chains is to understand the life cycle economic and environmental impacts of the product in question, at a richer level than is often done so far.

REFERENCES

1. Guide VDR Jr, van Wassenhove LN. Business aspects of closed-loop supply chains. Pittsburgh (PA): Carnegie-Bosch Institute; 2003.
2. Rogers D, Tibben-Lembke RS. Going backwards: reverse logistics trends and practices. Pittsburgh (PA): RLEC Press; 1999.
3. Bloemhof-Ruwaard J, Van Wassenhove L, Gabel HL, *et al.* An environmental life cycle optimization model for the European pulp and paper industry. *Omega* 1996;24(6):615–629.
4. Mayers CK, France CM, Cowell SJ. Extended producer responsibility for waste electronics: an example of printer recycling in the United Kingdom. *J Ind Ecol* 2005;9:169–189.
5. Bloemhof-Ruwaard JM, van Beek P, Hordijk L, *et al.* Interactions between operational research and environmental management. *Eur J Oper Res* 1995;85:229–243.
6. Fleischmann M, Bloemhof-Ruwaard JM, Dekker R, *et al.* Quantitative models for reverse logistics: a review. *Eur J Oper Res* 1997;103(1):1–17.
7. Linton J, Klassen R, Jayaraman V. Sustainable supply chains: an introduction. *J Oper Manage* 2007;25(6):1075–1082.
8. Srivastava SK. Green supply-chain management: a state-of-the-art literature review. *Int J Manage Rev* 2007;9(1):53–80.
9. Seuring S, Müller M. From a literature review to a conceptual framework for sustainable supply chain management. *J Cleaner Prod* 2008;16:1699–1710.
10. Corbett CJ, Klassen R. Extending the horizons: environmental excellence as key to improving operations. *Manuf Serv Oper Manage* 2006;8(1):5–22.
11. Huijbregts MAJ, Rombouts LJA, Hellweg S, *et al.* Is cumulative fossil energy demand a useful indicator for the environmental performance of products? *Environ Sci Technol* 2006;40(3):641–648.
12. Finnveden G, Moberg Å. Environmental systems analysis tools - an overview. *J Clean Prod* 2005;13:1165–1173.
13. Bloemhof-Ruwaard JM, Koudijs HG, Vis JC. Environmental impacts of fat blends: a methodological study combining life cycle assessment, multicriteria decision analysis and linear programming. *Environ Res Econ* 1995;6:371–387.
14. Azapagic A, Clift R. The application of life cycle assessment to process optimisation. *Comput Chem Eng* 1999;23:1509–1526.

15. Caldentey R, Mondschein S. Policy model for pollution control in the copper industry, including a model for the sulfuric acid market. *Oper Res* 2003;51(1):1–16.
16. Tan RBH, Khoo HH. An LCA study of a primary aluminum supply chain. *J Clean Prod* 2005;13(6):607–618.
17. Van Mierlo J, Timmermans J-M, Maggetto G, *et al.* Environmental rating of vehicles with different alternative fuels and drive trains: a comparison of two approaches. *Transp Res D* 2004;9:387–399.
18. D'Agosto M, Ribeiro SK. Eco-efficiency management program (EEMP)—a model for road fleet operation. *Transp Res D* 2004;9:497–511.
19. Forsberg G. Biomass energy transport analysis of bioenergy transport chains using life cycle inventory method. *Biomass Bioenergy* 2000;19:17–30.
20. Hamelinck CN, Suurs RAA, Faaij APC. International bioenergy transport costs and energy balance. *Biomass Bioenergy* 2005;29:114–134.
21. Sokhansanja S, Kumarc A, Turhollowa AF. Development and implementation of integrated biomass supply analysis and logistics model (IBSAL). *Biomass Bioenergy* 2006;30:838–847.
22. Sonesson U, Berlin J. Environmental impact of future milk supply chains in Sweden: a scenario study. *J Clean Prod* 2003;11(3):253–266.
23. Pelletier N. Environmental performance in the US broiler poultry sector: Life cycle energy use and greenhouse gas, ozone depleting, acidifying and eutrophying emissions. *Agric Syst* 2008;98:67–73.
24. Quariguasi Frota Neto J, Walther G, Bloemhof J, *et al.* A methodology for assessing eco-efficiency in logistics networks. *Eur J Oper Res* 2009;93:670–682.
25. Spielmann M, Althaus H-J. Can a prolonged use of a passenger car reduce environmental burdens? Life cycle analysis of Swiss passenger cars. *J Clean Prod* 2007;15:1122–1134.
26. Sathre R, Gustavsson L. Energy and carbon balances of wood cascade chains. *Res Cons Recycl* 2006;47:332–355.
27. Quariguasi Frota Neto J, Bloemhof-Ruwaard JM, van Nunen JAEE, *et al.* Designing and evaluating sustainable logistics networks. *Int J Prod Econ* 2008;111:195–208.
28. Matthews HS. Thinking outside “the box”: designing a packaging take-back system. *Calif Manage Rev* 2004;46(2):105–119.
29. Geyer R, Jackson T. Supply loops and their constraints: the industrial ecology of recycling and reuse. *Calif Manage Rev* 2004;46(2):55–73.
30. Björklund A, Finnveden G. Recycling revisited—life cycle comparisons of global warming impact and total energy use of waste management strategies. *Res Cons Recycl* 2005;44:309–317.
31. Daniel S, Tsoulfas GT, Pappis CP, *et al.* Aggregating and evaluating the results of different environmental impact assessment methods. *Ecol Indic* 2004;4:125–138.
32. Fleischmann M, Beullens P, Bloemhof-Ruwaard JM, *et al.* The impact of product recovery on logistics network design. *Prod Oper Manage* 2001;10(2):156–173.
33. Krikke H, Bloemhof-Ruwaard J, Van Wassenhove LN. Concurrent product and closed-loop supply chain design with an application to refrigerators. *Int J Prod Res* 2003;41(16):3689–3719.
34. Quariguasi Frota Neto J, Walther G, Bloemhof J, *et al.* From closed-loop to sustainable supply chains: The WEEE case. ERIM report series Research in Management ERS 2007-036-LIS, Erasmus University Rotterdam. 2007.

FURTHER READING

- Dekker R, Fleischmann M, Inderfurth K, Van Wassenhove LN, editors. *Reverse logistics: quantitative models for closed-loop supply chains*. Berlin: Springer; 2004. (Chapters 14 and 15 include explicit discussion of environmental impacts.)
- Flapper SD, van Nunen JAEE, Van Wassenhove LN, editors. *Managing closed-loop supply chains*. Berlin: Springer; 2005.

CLUSTERING

CEM IYIGUN
Department of Industrial
Engineering, Middle East
Technical University,
Ankara, Turkey

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait—often proximity according to some defined distance measure. *Data clustering* is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics [1].

The ideas and methods of clustering are used in many areas, including *statistics* [2], *machine learning* [3], *data mining* [4], *operations research* [5,6], *medical diagnostics*, *facility location*, and across multiple application areas including genetics, taxonomy, medicine, marketing, finance and e-commerce (see Refs 7–10 for applications of clustering). It is therefore useful to begin by stating our notation and terminology. We then survey some of the clustering algorithms and methods.

NOTATION AND TERMINOLOGY

Data

The objects of clustering are *data points* (also *observations*, and in facility location, *customers*.) Each data point is an ordered list of *attributes* (or *features*), such as height, weight, blood pressure, and so on. Assuming p attributes, a data point $\mathbf{x}$ is thus a p -dimensional vector, $\mathbf{x} = (x_1, x_2, \dots, x_p)$, with the attributes x_i for components.

The vector analogy cannot be carried too far, since, in general, vector operations (such as vector addition, scalar multiplication) do

not apply to data points. Also, the attributes are not necessarily of the same algebraic type; some may be categorical, and others are reals. However, we can always imbed the data points in a p -dimensional real vector space $\mathbb{R}^p$, and for convenience we denote by $\mathbf{x} \in \mathbb{R}^p$ the fact that the data point $\mathbf{x}$ has p attributes.

We assume N data points $\mathbf{x}_i$, collected in a set

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^p, \quad (1)$$

called the *data set*. We sometimes represent $\mathcal{D}$ by an $N \times p$ matrix

$$D = (x_{ij}), \text{ where } x_{ij} \text{ is the } j\text{th component} \\ \text{of the data point } \mathbf{x}_i. \quad (2)$$

The Problem

Given the data set $\mathcal{D}$, and integer K , $1 \leq K \leq N$, the *clustering problem* is to partition the data set $\mathcal{D}$ into K disjoint clusters

$$\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \dots \cup \mathcal{C}_K, \quad \text{with} \\ \mathcal{C}_j \cap \mathcal{C}_k = \emptyset \text{ if } j \neq k, \quad (3)$$

each cluster consisting of points that are *similar* (in some sense) and points of different clusters being dissimilar. We take here, *similar*, to mean *close* in the sense of *distances* $d(\mathbf{x}, \mathbf{y})$ between points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. The *number of clusters* denoted by K is given; however, there are problems where the “right” number of clusters (to fit best the data) is to be determined. The cases $K = 1$ (the whole $\mathcal{D}$ is one cluster) and $K = N$ (every point is a separate cluster) are included for completeness.

Cluster Membership

A clustering is *hard* (or *crisp*, *rigid*, *deterministic*) if each data point $\mathbf{x}$ is assigned to one and only one cluster $\mathcal{C}$, so that the statement $\mathbf{x} \in \mathcal{C}$ is unambiguous.

A point $\mathbf{x}$ is *labeled* if its cluster $\mathcal{C}$ is known, in which case $\mathcal{C}$ is the *label* of $\mathbf{x}$.

In *soft* (or *fuzzy*, *probabilistic*) clustering the rigid assignment $\mathbf{x} \in C$ is replaced by a *cluster membership function* $u(\mathbf{x}, C)$ representing the *belief* that x belongs to C . The numbers $u(\mathbf{x}, C_k)$ are often taken as probabilities that $\mathbf{x}$ belongs to C_k , so that

$$\sum_{k=1}^K u(\mathbf{x}, C_k) = 1, \text{ and } u(\mathbf{x}, C_k) \geq 0 \\ \text{for all } k = 1, \dots, K. \quad (4)$$

Classification

In *classification* or *supervised learning*, the number K of clusters is given, and a certain subset $\mathcal{T}$ of the data set $\mathcal{D}$ is given as labeled, that is, for each point $\mathbf{x} \in \mathcal{T}$ it is known to which cluster it belongs. The subset $\mathcal{T}$ is called the *training set*.

The information obtained from the training set, is then used to find a rule for classifying the remaining data $\mathcal{D} \setminus \mathcal{T}$ (called the *testing set*), and any future data of the same type, to the K clusters. The *classification rule* r is a function from $\mathcal{D}$ (and by extension $\mathbb{R}^p$) to the integers $\{1, 2, \dots, K\}$, so that

$$r(\mathbf{x}) = k \iff \mathbf{x} \in C_k.$$

In analogy, clustering is called *unsupervised learning* to emphasize the absence of prior information.

Similarity and Dissimilarity

Some cluster analysis techniques begin with transforming the matrix $\mathcal{D}$ into an $N \times N$ matrix of data points of *similarities* or *dissimilarities* (a general term is *proximity*). The *similarity* between two data objects is a numerical measure of the degree to which the objects are alike. Naturally, the more alike the objects are, the higher the similarities are. Similarities are usually nonnegative and often take values between 0 and 1. On the other hand, the *dissimilarity* between two objects is a numerical value measuring the degree to which the objects are different. Dissimilarities are lower for more similar data objects. Although the *distance* is used as synonym for dissimilarities, distance is

a special class of dissimilarities. Dissimilarity measures mostly range from 0 to ∞ , but they also take values in the range of $[0, 1]$. In the following sections, we provide some general examples of similarity and dissimilarity measures.

Similarity Measures

Let similarity between two data points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ be defined by $s(\mathbf{x}, \mathbf{y})$. If the data points contain only binary values the similarity measures $s(\mathbf{x}, \mathbf{y})$, are called *similarity coefficients*. The comparison of two binary data points leads to different similarity measures. All the measures are defined in terms of the counts of matches and mismatches in p variables of two data binary points. One commonly used similarity coefficient, *matching coefficient* is defined as

$$s(\mathbf{x}, \mathbf{y}) = \frac{f_{00} + f_{11}}{\text{number of attributes}}, \quad (5)$$

where f_{00} and f_{11} are the number of attributes that both data points take the value 0 and 1, respectively.

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are two data objects that have different number of attributes. In such cases, the *Jaccard coefficient* is frequently used to handle asymmetric binary attributes. The *Jaccard coefficient* $\mathcal{J}$, is given by the following equation:

$$\mathcal{J}(\mathbf{x}, \mathbf{y}) = \frac{f_{11}}{\text{number of attributes excluding } f_{00}}. \quad (6)$$

The *cosine similarity*, which depends on the *standard inner product* of two vectors is another similarity measure that is commonly used.

Assuming a norm $\|\cdot\|$ on the space $\mathbb{R}^p$, the *standard inner product* of $\mathbf{x} = (x_1, \dots, x_p)$ and $\mathbf{y} = (y_1, \dots, y_p)$ is defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^p x_i y_i. \quad (7)$$

Then the *cosine similarity* is defined as

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}. \quad (8)$$

Actually, cosine similarity is a measure of the angle between $\mathbf{x}$ and $\mathbf{y}$. Thus, if the cosine similarity is 1, the angle between $\mathbf{x}$ and $\mathbf{y}$ is 0° , and $\mathbf{x}$ and $\mathbf{y}$ are the same except for the length. When the cosine similarity is 0, the angle between $\mathbf{x}$ and $\mathbf{y}$ is 90° , and these two data points do not have any similarities.

Dissimilarity and Distance Measures

When all the attributes are continuous, proximities between data points are typically measured by dissimilarity measures or distance measures. *Distance measures* are the important class of dissimilarity measures with certain properties. Generally, a dissimilarity measure $\delta(\mathbf{x}, \mathbf{y})$, fulfilling the metric axioms is termed as a *distance measure*.

Assuming a norm $\|\cdot\|$ on the space $\mathbb{R}^p$, a distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ is defined by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \quad (9)$$

for example, the Euclidean norm gives the distance between $\mathbf{x} = (x_1, \dots, x_p)$ and $\mathbf{y} = (y_1, \dots, y_p)$ as

$$d(\mathbf{x}, \mathbf{y}) := \left(\sum_{j=1}^p (x_j - y_j)^2 \right)^{1/2}, \quad \text{the Euclidean distance}, \quad (10)$$

and the ℓ_1 -norm gives

$$d(\mathbf{x}, \mathbf{y}) := \sum_{j=1}^p |x_j - y_j|, \quad \text{the } \ell_1\text{-distance}, \quad (11)$$

also called the *Manhattan* or *taxicab distance*.

If $\mathbf{Q}$ is a positive-definite $p \times p$ matrix, then $\sqrt{\langle \mathbf{x}, \mathbf{Q} \mathbf{x} \rangle}$ is a norm on $\mathbb{R}^p$, and the corresponding distance is

$$d(\mathbf{x}, \mathbf{y}) := \sqrt{\langle \mathbf{x} - \mathbf{y}, \mathbf{Q}(\mathbf{x} - \mathbf{y}) \rangle}, \quad \text{an elliptic distance}, \quad (12)$$

depending on the choice of $\mathbf{Q}$. For $\mathbf{Q} = \mathbf{I}$, the identity matrix, Equation (12) gives the

Euclidean distance (10). Another common choice is $\mathbf{Q} = \Sigma^{-1}$, where Σ is the covariance matrix of the data in question, in which case Equation (12) gives

$$d(\mathbf{x}, \mathbf{y}) := \langle \mathbf{x} - \mathbf{y}, \Sigma^{-1}(\mathbf{x} - \mathbf{y}) \rangle, \quad \text{the Mahalanobis distance}, \quad (13)$$

that is used commonly in multivariate statistics.

Distances associated with norms satisfy the *triangle inequality*,

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}), \quad \text{for all } \mathbf{x}, \mathbf{y}, \mathbf{z}. \quad (14)$$

However, distance functions violating Equation (14) are also used.

Similarity Data

Given a distance function $d(\cdot, \cdot)$ in $\mathbb{R}^p$, and the data set $\mathcal{D}$, the *dissimilarity* (or *proximity*) *matrix* of the data is the $N \times N$ matrix

$$\delta = (d_{ij}), \quad \text{where } d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \dots, N. \quad (15)$$

It is sometimes convenient to work with the *similarity matrix*,

$$S = (g(d_{ij})), \quad \text{where } g(\cdot) \text{ is an increasing function}. \quad (16)$$

Representatives of Clusters

In many clustering methods a cluster is represented by a typical point, called its *center* (also *representative*, *prototype*, and in facility location, *facility*.) A common choice for the center is the *centroid* of the points in the cluster. The cluster centers can happen to be the data points in the clusters.

The center of the k th cluster C_k is denoted by $\mathbf{c}_k$, and the distance $d(\mathbf{x}, C_k)$ of a point $\mathbf{x}$ from that cluster is defined as its distance

4 CLUSTERING

from the center $\mathbf{c}_k$,

$$d(\mathbf{x}, \mathcal{C}_k) := d(\mathbf{x}, \mathbf{c}_k), \quad (17)$$

and denoted $d_k(\mathbf{x})$, if the center is understood.

Objective Function of Clustering

Sometimes the “goodness” of clustering can be expressed by an *objective function* of the given data $\mathcal{D}$ and the clusters $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ (see the section titled “Dispersion Statistics” for more details). For example,

$$f(\mathcal{D}, \{\mathcal{C}_1, \dots, \mathcal{C}_K\}) = \sum_{k=1}^K \sum_{\mathbf{x}_i \in \mathcal{C}_k} d(\mathbf{x}_i, \mathbf{c}_k) \quad (18)$$

is the sum of distances of data points to the centers of their respective clusters, while

$$f(\mathcal{D}, \{\mathcal{C}_1, \dots, \mathcal{C}_K\}) = \sum_{k=1}^K \sum_{\mathbf{x}_i \in \mathcal{C}_k} d(\mathbf{x}_i, \mathbf{c}_k)^2 \quad (19)$$

is the sum of squares of these distances.

In such cases, clustering reduces to an *optimization problem*, that without loss of

Algorithm 1. *k*-means Clustering Algorithm.

Step 0 Initialization: Given data set $\mathcal{D}$, integer K , $2 \leq K < N$,

select K initial centers $\{\mathbf{c}_k\}$

Step 1 Compute the distances $d(\mathbf{x}_i, \mathbf{c}_k)$, $i = 1, \dots, N$, $k = 1, \dots, K$.

Step 2 Partition the data set $\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \dots \cup \mathcal{C}_K$ by assigning each data point to the cluster whose center is the nearest

Step 3 Recompute the cluster centers.

Step 4 If the centers have not changed or the change is smaller than a threshold, **stop**.
else go to Step 1.

Notes

1. The initial “centers” in Step 0 are just points, and not yet associated with clusters. There are several ways to initialize the cluster centers. One way is to randomly select K points from the data set. The other way is to randomly generate K points evenly distributed in the space spanned by the data set $\mathcal{D}$.
2. In Step 3 the center of each cluster is computed using the points assigned to that cluster.

generality, is considered a *minimization problem*,

$$\min_{\mathcal{C}_1, \dots, \mathcal{C}_K} f(\mathcal{D}, \{\mathcal{C}_1, \dots, \mathcal{C}_K\}), \quad (20)$$

which is often hard (combinatorial, non-smooth), but approximate solutions of Equation (20) may be acceptable.

CENTER-BASED CLUSTERING METHODS

Center-based clustering algorithms construct the clusters using the distances of data points from the cluster centers.

The best-known and most commonly used center-based algorithm is the *k-means algorithm* [11,12], which (implicitly) minimizes the objective

$$\sum_{k=1}^K \sum_{\mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad (21)$$

where $\mathbf{c}_k$ is the centroid of the k^{th} cluster. Other names like hard *k*-means and ISO-DATA [13,14] have also been used in the literature.

3. The stopping rule in Step 4 implies that there are no further reassignments.
4. The center updates in the iterations are computed by

$$\mathbf{c}_k = \frac{\sum_{i=1}^N u_{ik} \mathbf{x}_i}{\sum_{i=1}^N u_{ik}}, \quad k = 1, \dots, K, \quad (22)$$

where $u_{ik} = 1$, if $\mathbf{x}_i \in \mathcal{C}_k$, and $u_{ik} = 0$ otherwise. Equation (22) gives the

centers as the geometrical centroids of the data points of the cluster.

5. Using Euclidean distances, iterating Steps 2 and 3 leads to the minimization of the objective (21).

Variants of the k -Means Algorithm

Several variants of k -means algorithm have been reported in the literature [15,16]. Some of them attempt to select a good initial partition so that the algorithm is more likely to find the global minimum value [17]. An important variant of the algorithm is to permit splitting and merging of the resulting clusters [13] in Step 2 of the algorithm.

Some variants of the algorithm use different criteria. Diday [18] used different representatives of the clusters (other than the cluster centers), and the Mahalanobis distance is used instead of the Euclidean distance in Mao and Jain [19], Cheung [20], and elsewhere.

The *k-modes algorithm* [21] is a recent center-based algorithm for categorical data. It utilizes a simple matching dissimilarity measure, which is the total number of mismatches of the categorical attributes of the objects between two categories [2]. Another variant, the *k-prototypes algorithm* [21], incorporates real and categorical data. It combines the k -means and k -modes processes in order to cluster data with numerical and categorical values. The dissimilarity measure in the k -prototypes algorithm is defined as a weighted sum of the dissimilarity measure on numerical attributes (just the squared Euclidean distance) and the simple matching dissimilarity measure on categorical attributes.

Fuzzy k -Means

The k -means algorithm can be adapted to soft clustering (see section titled “Cluster Membership”). A well-known center-based algorithm for soft clustering is the *fuzzy k -means algorithm* [22,23].

The objective function minimized in this algorithm is

$$f = \sum_{i=1}^N \sum_{k=1}^K u_{ik}^m d_{ik}^2 = \sum_{i=1}^N \sum_{k=1}^K u_{ik}^m \|\mathbf{x}_i - \mathbf{v}_k\|^2,$$

where u_{ik} are the membership functions of $\mathbf{x}_i \in C_k$, and typically satisfy Equation (4), and m is a real number, $m > 1$, known as *fuzzifier*.

The equation for finding the centers is similar to Equation (22) of the k -means algorithm, but u_{ik} takes values between 0 and 1.

$$\mathbf{c}_k = \frac{\sum_{i=1}^N u_{ik}^m \mathbf{x}_i}{\sum_{i=1}^N u_{ik}^m}, \quad k = 1, \dots, K. \quad (23)$$

When m tends to 1, the algorithm converges to the k -means method.

HIERARCHICAL CLUSTERING ALGORITHMS

Hierarchical clustering algorithms are an important class of clustering methods that do not group data into particular clusters represented by centers, but instead a series of partitions to the data set or clusters are merged and a hierarchy of clusters is created.

These algorithms transform a similarity data set into a tree-like structure which is called a *dendrogram* [24]. The root of the dendrogram consists of a single cluster containing all observations (data points), and the leaves correspond to individual observations. In the middle, child clusters partition the points assigned to their common parent according to a similarity measure. This is illustrated in Fig. 1. (Note that the dendrogram is not a binary tree.) Hierarchical clustering methods are generally categorized into two major methods as *agglomerative*, in which one starts at the leaves and successively merges clusters together; and *divisive* [2,10], in which one starts at the root and recursively splits the clusters.

Agglomerative clustering is a bottom-up way of constructing the dendrogram. It starts with N clusters, each has one point (singleton), and recursively merges two or more most similar clusters until all N data points are in a single cluster (which is the root of dendrogram). On the other hand, *divisive* clustering starts from the root, a top-down way of constructing the dendrogram. It starts at the root with one cluster containing all N

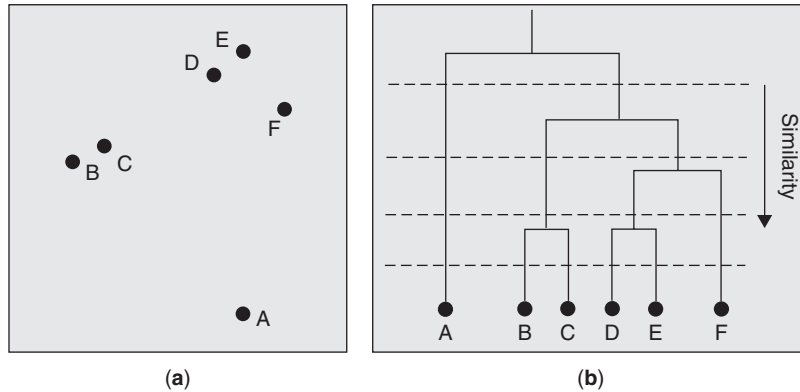


Figure 1. An example of the dendrogram that might be produced by a hierarchical algorithm from the data shown in (a). The dashed lines indicate different partitions at different levels of similarity.

data points and recursively splits the clusters until N clusters are formed or until a stopping criterion (frequently, the requested number k of clusters) is achieved. Agglomerative techniques are more commonly used than divisive techniques, because they are computationally less complex.

For agglomerative hierarchical techniques, the criterion is typically to merge the *closest* pair of clusters, where *close* is defined by a specified measure of cluster proximity. There are three definitions of the closeness (linkage metrics) between two clusters: *single-link*, *complete-link*, and *average-link*. The *single-link* similarity between two clusters is the similarity between the two most similar instances, one of which appears in each cluster. Single link is good at handling nonelliptical shapes, but is sensitive to noise and outliers. The *complete-link* similarity is the similarity between the two most dissimilar points, one from each cluster. Complete link is less sensitive to noise and outliers, but can break large clusters, and it favors globular shapes. The *average-link* similarity is a compromise between the two. In this version, proximity of two clusters is the average of pairwise proximity between points in the two clusters. *Average-link* is also less susceptible to noise and outliers as in *complete-link*. The advantages of agglomerative and divisive algorithms are the following: (i) they do not require to prespecify the number of clusters in advance, (ii) they compute a complete hierarchy of clusters, (iii)

good result visualizations are integrated into the methods, and (iv) a *flat* partition can be derived afterwards (using a cut through the dendrogram). However, both methods suffer from their inability to perform adjustments once the splitting or merging decision is made. If a misclassification is made at one point during the growth of the dendrogram, it is carried on until the end of the process. It is not possible to correct the misclassification during the clustering process. After the tree structure has been constructed, different clustering interpretations can be performed.

Although the number of clusters is not predefined in hierarchical clustering, when to stop the partition in a divisive algorithm or when to stop the merging in an agglomerative algorithm are important questions. At which level should the tree be cut or at which value of similarity (or dissimilarity) should the dendrogram be cut are practical questions. All such questions lead to deciding how many clusters are present. The usual practice being that a large disparity in the levels of dendrogram at which clusters merge indicates the presence of natural groupings. Different heuristics may be used to answer the question. Although it does not always work, choosing a value of dissimilarity such that there is a large *gap* in the dendrogram is commonly used as a heuristic [25]. Moreover, various validity indices for hierarchical clustering are also used for determining where to stop the dendrogram [26]. The index measure is evaluated in different stages of the

hierarchical clustering and the graphing the index values for different levels of dendrogram helps determining the number of clusters.

In hierarchical clustering, the regular data matrix representation (which is point-by-attribute data) is sometimes of secondary importance. Instead, hierarchical clustering frequently deals with the similarity (or dissimilarity) matrix between the data points. It is sometimes called *connectivity matrix*. Linkage metrics are constructed from elements of this matrix.

There are many different hierarchical clustering algorithms presented in the literature. Some of them are Balanced Iterative Reducing and Clustering using Hierarchies *BIRCH* [27], Clustering Using Representatives *CURE* [28], and *CHAMELEON* [29]. Robust Clustering Algorithm for Categorical Attributes (*ROCK*) [30] is another clustering algorithm using the Jaccard coefficient to measure similarity. More recently, a novel incremental hierarchical clustering algorithm, *GRIN*, for numerical data sets is presented in Chen *et al.* [31]. A survey and comparison of these algorithms are in Kotsiantis and Pintelas [32] and Berkhin [33].

OTHER CLUSTERING METHODS

In this section, we briefly describe other clustering methods developed in the data clustering area. For comprehensive explanations and further details, see the cited references.

Probabilistic Clustering

Probabilistic clustering refers to data sets whose points come from a known statistical distribution, whose parameters have to be estimated. Specifically, the data may come from a mixture of several distributions and the weights of the distributions in the mixture, and their parameters, have to be determined.

The best-known probabilistic method is the *expectation-maximization (EM) algorithm* [34] where log-likelihood of the data points drawn from a given mixture model. The underlying probability model and its parameters determine the membership function of the data points. The algorithm starts

with initial guesses for the mixture model parameters. These values are then used to calculate the cluster membership functions for the data points. In turn, these membership functions are used to re-estimate the parameters, and the process is repeated.

Another probabilistic clustering method is *AUTOCLASS* [35], which also uses the mixture model and extends the search to different models and different k distributions. The method relies on Bayesian methodology and is applicable both continuous and categorical data. The algorithm starts with a random initialization of the parameters, then iteratively updates them to find their maximum likelihood estimates. Furthermore, Pena *et al.* [36] assume that there is a hidden variable reflecting the cluster membership for every data point, besides the observed variables. Using the hidden variable, the clustering problem as a supervised learning from incomplete data [37] and a learning algorithm, called *RBMNs* (recursive Bayesian multinets), relying on Bayesian methodology can be used for clustering.

Probabilistic methods depend critically on their assumed priors. If the assumptions are correct, one gets good results. A drawback of these algorithms is that they are computationally expensive. Another problem found in this approach is called the *overfitting* [38].

Density-Based Clustering

Density-based methods consider that clusters are dense sets of data points separated by less dense regions; clusters may have arbitrary shape and data points can be arbitrarily distributed. The concepts of *density*, *connectivity*, and *boundary*, which are related to a point's neighborhood are required in clustering the data points. In density-based clustering, the neighborhood of a data point is defined with a given radius (*Eps*), which contains at least a minimum number of data points (*MinPts*). The algorithm is guided by the density of the points and it can find clusters having arbitrary shapes.

DBSCAN [39] is one of the most common density-based clustering algorithm. It is further improved in Kotsiantis and Pintelas [32], called the *GDBSCAN* algorithm, which generalizes the DBSCAN algorithm

for clustering data points with both numerical and categorical attributes. A parallel version of DBSCAN algorithm is developed in Xu *et al.* [40]. Moreover, the *DBCLASD* algorithm (Distribution-based Clustering of Large Spatial Data Sets), introduced by Xu *et al.* [41], eliminates the need for *MinPts* and *Eps* parameters.

One can consider within the category of density-based methods, the *grid-based solutions*, such as *DENCLUE* [42] or *CLIQUE* [43], mostly developed for spatial data mining. These methods quantize the space of the data points into a finite number of cells (attention is shifted from data points to space partitioning) and only retain for further processing the cells having a high density of points; isolated data points are thus ignored. Quantization steps and density thresholds are common parameters for these methods.

Graph-Theoretic Clustering

Another clustering method is the *graph-theoretic clustering* method, where the data points are represented as nodes in a graph and the dissimilarity between two points is the “length” of the edge between the corresponding nodes. In several methods, a cluster is a subgraph that remains connected after the removal of the longest edges of the graph [10]; for example, in Zahn [44] (the best-known graph-theoretic clustering algorithm) the minimal spanning tree of the original graph is built and then the longest edges are deleted. Some other graph-theoretic methods rely on the extraction of *cliques* and are then more related to center-based methods [45].

Model-Based Clustering

SOM(self-organizing map)net, introduced by Teuvo Kohonen, is a model-based method [46]. It is a type of artificial neural network, which is trained using unsupervised learning. The method can be thought as two layers of neural network using a neighborhood function to preserve the topological properties of the input space [47]. Each neuron of the network is the cluster center represented by the n -dimensional weight vector. SOM algorithm works in two modes: training and mapping. Training creates the map using the input

vectors that are also called *vector quantization*. SOM is trained iteratively. In each training step, one input vector is randomly chosen from the input data set, and the distance between the input vector and all the weight vectors of the SOM is calculated using a distance measure. The neuron whose weight vector is closest to the input is called best-matching unit. After finding the best-matching unit, the weight vectors of SOM are updated such that best-matching unit is moved closer to the input vector. The SOM method is robust and insensitive to outliers. It can easily detect the outliers and can handle the missing data.

Volume-Based Clustering

To overcome the difficulties like clustering with equal size or spherical shapes, we can use Mahalanobis distances (see the section titled “Dissimilarity and Distance Measures”) instead of Euclidean distance [48]. For example, if the covariance Σ is known, then the similarity within that cluster, with center $\mathbf{c}$ would be measured by $\|\mathbf{x} - \mathbf{c}\|_{\Sigma^{-1}}$. This measure is scale invariant and can deal with asymmetric, nonspherical clusters. A difficulty in using Mahalanobis distances is getting a good estimate of the covariance matrices in question.

A promising alternative scale-invariant metric of cluster quality are *minimum volume ellipsoids*, where data points are allocated into clusters so that the volumes of the covering ellipsoids for each cluster is minimized. The problem of finding the minimum volume ellipsoid can be formulated as a *semidefinite programming* problem and an efficient algorithm for solving the problem has been proposed by Sun and Freund [49].

Biclustering

Biclustering is a clustering method in which the rows and columns of a data matrix are grouped simultaneously. It is also named as *co-clustering* or *two-mode clustering* [50]. Given a data matrix with m rows and n columns, the biclustering algorithm generates subset of rows (columns), which show similar behavior across a subset of columns (and rows) in a given data set. Biclustering

has been used in *text mining* and *bioinformatics* (i.e., gene expression analysis [51]).

There are many biclustering algorithms developed for bioinformatics, in the literature. Some of them are block clustering, coupled two way clustering (CTWC) [52], δ -bicluster [51], δ -Cluster, Gibbs, SAMBA [53], and cMonkey [54].

Semisupervised Clustering

In unsupervised clustering no prior information is given; the data points are clustered into disjoint sets using the similarity information. In many cases a small amount of information is available about the data points. The information can be in the form of pairwise (*must-link* or *cannot-link*) constraints, which define the relationship between the data points [55]. A must-link constraint specifies that the two data points should be assigned to the same cluster. A cannot-link constraint specifies that the two data points should not be assigned to the same cluster. The information can also be about the class labels for some points or all points with a reliability level [56]. The process of clustering using a side information is called semisupervised clustering. The set of extra information provides a limited supervision but guides the clustering process. Examples of semisupervised clustering algorithm include *COP-Kmeans* [55], and *PCKmeans* [57].

DISPERSION STATISTICS FOR CLUSTERING

The partitioning (3) of the data points $\mathbf{x}_i$ (which are the rows of the $N \times p$ data matrix D of Equation (2)), gives rise to the $p \times p$ total dispersion matrix,

$$T = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad (24)$$

where the p -dimensional vector $\bar{\mathbf{x}}$ is the mean of all the data points. The total dispersion matrix T can be partitioned as

$$T = W + B, \quad (25)$$

where W is the *within-cluster dispersion matrix*,

$$W = \sum_{k=1}^K \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)' \quad (26)$$

here $\bar{\mathbf{x}}_k$ is the mean of the data points in the cluster C_k , and B is the *between-clusters dispersion matrix*,

$$B = \sum_{k=1}^K N_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})', \quad (27)$$

where N_k is the number of data points in C_k .

For univariate data ($p = 1$), Equation (25) represents the division of the total sum of squares of a variable into the within- and between-clusters sum of squares. In the univariate case, a natural criterion for grouping would be to choose the partition corresponding to the minimum value of the within-group sum of squares or, equivalently, the maximum value of the between-cluster sum of squares.

In the multivariate case ($p > 1$) the derivation of a clustering criterion from the Equation (25) is not as clear-cut as the univariate case, and several alternatives have been suggested.

DISPERSION OBJECTIVES

The dispersion statistics of the section titled "Dispersion Statistics for Clustering" suggest several different objectives for clustering.

Minimization of trace (W)

The trace of the matrix W in Equation (26) is the sum of the within-cluster variances. Minimizing this trace works to make the clusters more homogeneous; thus the problem,

$$\min\{\text{trace } W\}, \quad (28)$$

which, by Equation (25) is equivalent to

$$\max\{\text{trace } B\}. \quad (29)$$

This can be shown to be equivalent to minimizing the sum of the squared Euclidean distances between data points and their cluster mean, which is used in k -means algorithms. The criterion can also be derived on the basis of the distance matrix:

$$E = \sum_{k=1}^K \frac{1}{2N_k} \sum_{i=1}^{N_k} \sum_{j=1, j \neq i}^{N_k} d_{ij}^2, \quad (30)$$

where d_{ij} is the Euclidean distance between i th and j th data points in cluster C_k . Thus the minimization of $\text{trace}(W)$ is equivalent to the minimization of the homogeneity criterion $h_1(C_k)/N_k$ for Euclidean distances and $n = 2$ [58].

Minimization of $\det(W)$

The differences in cluster mean vectors are based on the ratio of the determinants of the total and within-cluster dispersion matrices. Large values of $\det(T)/\det(W)$ indicate that the cluster mean vectors differ. Thus, a clustering criterion can be constructed as the maximization of this ratio

$$\min \left\{ \frac{\det(T)}{\det(W)} \right\}. \quad (31)$$

Since T is the same for all partitions of N data points into K clusters, this problem is equivalent to

$$\min\{\det(W)\}. \quad (32)$$

Maximization of $\text{trace}(BW^{-1})$

A further criterion considered is a combination of dispersion matrices:

$$\max \left\{ \text{trace} \left(\frac{B}{W} \right) \right\}. \quad (33)$$

This criterion is obtained from the product of the between-clusters dispersion matrix and the inverse of the within-clusters dispersion matrix. This function is also a further test criterion used in the context of multivariate analysis of variance, with large values of $\text{trace}(BW^{-1})$ indicating that the cluster mean vectors differ.

Comparison of the Clustering Criteria

The criterion (31) is, perhaps, the most commonly used one among the three clustering criteria presented above. However, it has some serious problems [58]. First, the method is not scale invariant. This means that, different solutions may be obtained from the

raw or standardized data. Clearly, this is of considerable practical importance because of the need for standardization in many applications. Another problem with the use of this criterion is that it may impose a *spherical* structure on the observed clusters even when the natural clusters in the data are of other shapes. The criteria in Equations (28) and (31) are not affected by scaling. Moreover, the criterion in Equation (32), which has also been widely used does not restrict clusters to being spherical. It can also identify *elliptical* clusters. However, this criteria assumes that all clusters in the data have the same shape. Finally, both the criteria in Equations (28) and (32) produce clusters that are of nearly equal size, meaning that clusters contain roughly equal numbers of data points.

REFERENCES

1. Cluster analysis. 2007. Available at http://en.wikipedia.org/wiki/Cluster_analysis.
2. Kaufman L, Rousseeuw P. Finding groups in data: an introduction to cluster analysis. New York: John Wiley & Sons, Inc.; 1990.
3. Fisher D. Knowledge acquisition via incremental conceptual clustering. Mach Learn 1987;2:139–172.
4. Fayyad UM, Piatetsky-Shapiro G, Smyth P, *et al.* Advances in knowledge discovery and data mining. Cambridge (MA): MIT Press; 1996.
5. Bradley PS, Mangasarian OL, Street WN. Clustering via concave minimization. In: Mozer MC, Jordan MI, Petsche T, editors. Volume 9, Advances in neural information processing systems. Cambridge: MIT Press; 1997. pp. 368–374.
6. Hansen P, Jaumard B. Cluster analysis and mathematical programming. Math Program 1997;79:191–215.
7. Bertsimas D, Mersearau AJ, Patel NR. Dynamic classification of online customers. Proceedings of SIAM International Conference on Data Mining. San Francisco (CA); 2003.
8. Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics 1993;49:803–821.
9. Gavrilov M, Anguelov D, Indyk P, *et al.* Mining the stock market: which measure is best?. 6th ACM SIGKDD International Conference

- on Knowledge Discovery and Data Mining. Boston (MA); 2000. pp. 487–496.
10. Jain AK, Dubes RC. Algorithms for clustering data. Upper Saddle River (NJ): Prentice Hall; 1988.
 11. McQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley (CA); 1967. pp. 281–297.
 12. Hartigan J. Clustering algorithms. New York: John Wiley & Sons, Inc.; 1975.
 13. Ball GH, Hall DJ. Isodata, a novel method of data analysis and pattern classification. Technical report no. AD 699616. Stanford (CA): Stanford University; 1965.
 14. Ball GH, Hall DJ. A clustering technique for summarizing multivariate data. Behav Sci 1967;12:153–155.
 15. Forgy EW. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics 1965;21:768–769.
 16. Anderberg MR. Cluster analysis for cluster applications. New York: Academic Press, Inc.; 1973.
 17. Duda RO, Hart PE. Pattern classification and scene analysis. New York: John Wiley & Sons, Inc.; 1973.
 18. Diday E. The dynamic clustering method in non-hierarchical clustering. J Comput Inf Sci 1973;2:61–88.
 19. Mao J, Jain AK. A self-organizing network for hyperellipsoidal clustering. IEEE Trans Neural Netw 1996;7:16–29.
 20. Cheung Y-M. k*-means: a new generalized k-means clustering algorithm. Pattern Recognit Lett 2003;24:2883–2893.
 21. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowl Discov 1998;2:283–304.
 22. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum; 1981.
 23. Sato M, Sato Y, Jain L. Fuzzy clustering models and applications. Volume 9, Studies in fuzziness and soft computing. Physica-Verlag; 1997.
 24. Jardine N, Sibson R. Mathematical taxonomy. New York: John Wiley & Sons, Inc.; 1971.
 25. Jain AK. Cluster analysis. In: Young YT, Fu K-S, editors. Handbook of pattern recognition and image processing. Orlando (FL): Academic Press; 1986.
 26. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst 2001;17(2):107–145.
 27. Zhang T, Ramakrishnan R, Linvy M. Birch: an efficient data clustering method for very large data sets. Data Mining Knowl Discov 1997;1(2):141–182.
 28. Guha S, Rastogi R, Shim K. Cure: an efficient clustering algorithm for large data sets. Proceedings of the ACM SIGMOD Conference. Seattle, Washington; 1998.
 29. Karypis G, Han EH, Kumar V. Chameleon: a hierarchical clustering algorithm using dynamic modeling. Computer 1999;32(7):68–75.
 30. Guha S, Rastogi R, Shim K. Rock: a robust clustering algorithm for categorical attributes. Proceedings of the IEEE Conference on Data Engineering. Sydney, Australia; 1999.
 31. Chen C-Y, Hwang S-C, Oyang Y-J. An incremental hierarchical data clustering algorithm based on gravity theory. Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002, LNCS 2336. Taipei, Taiwan: Springer; 2002.
 32. Kotsiantis SB, Pintelas PE. Recent advances in clustering: a brief survey. Volume 1. WSEAS Transactions on Information Science and Applications; 2004. pp. 73–81.
 33. Berkhin P. Survey of clustering data mining techniques. Technical report. Accrue Software, Inc.; 2003.
 34. McLachlan GJ, Krishnan T. The EM algorithm and extensions. New York: John Wiley & Sons, Inc.; 1997.
 35. Cheeseman P, Stutz J. Bayesian classification (autoclass): theory and results. In: Fayyad UM, Piatetsky-Shapiro G, Smith P and Uthurusamy R, editors. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence. Menlo Park (CA); 1996. pp. 153–180.
 36. Pena J, Lozano J, Larranaga P. Learning recursive Bayesian multinets for data clustering by means of constructive induction. Mach Learn 2002;47:63–89.
 37. Jensen F. An introduction to bayesian networks. Springer; 1996.
 38. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data Mining, Inference and Prediction. New York (NY): Springer; 2003.
 39. Ester M, Krieger H-P, Sander J, *et al.* A density-based algorithm for discovering clusters in large spatial data sets with noise.

- Proceedings 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park (CA); 1996. pp. 226–231.
40. Xu X, Jager J, Kriegel H-P. A fast parallel clustering algorithm for large spatial data sets. *Data Mining Knowl Discov* 1999;3:263–290.
 41. Xu X, Ester M, Kriegel H-P, *et al.* A nonparametric clustering algorithm for knowledge discovery in large spatial data sets. *Proceedings of the 14th IEEE International Conference on Data Engineering*. Orlando (FL): IEEE Computer Society Press; 1998.
 42. Hinneburg A, Keim D. An efficient approach to clustering in large multimedia data sets with noise. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. New York (NY); 1998. pp. 58–65.
 43. Agrawal R, Gehrke J, Gunopulos D, *et al.* Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of the 1998 ACM-SIGMOD Conference On the Management of Data*. Seattle, Washington; 1998. pp. 94–105.
 44. Zahn CT. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans Comput* 1971;C-20(1):68–86.
 45. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. *Proceedings of 14th Advances in Neural Information Processing Systems*. 2002.
 46. Kohonen T. *Self-organizing maps*, Springer Series in Information Sciences. 2nd Extended ed. New York: Springer; 1997.
 47. Self-organizing map. 2010. Available at http://http://en.wikipedia.org/wiki/Self-organizing_map.
 48. Shioda R, Tuncel L. Clustering via minimum volume ellipsoids. *Comput Optim Appl* 2007;37(3):247–295.
 49. Sun P, Freund R. Computation of minimum volume covering ellipsoids. *Oper Res* 2004; 52(5):690–706.
 50. Mechelen IV, Bock H, Boeck PD. Two-mode clustering methods: a structured overview. *Stat Methods Med Res* 2004;13(5):363–94.
 51. Cheng Y, Church GM. Bicustering of expression data. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. La Jolla (CA); 2000. pp. 93–103.
 52. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* 2000;97: 12079–12084.
 53. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 2002;18, Suppl 1): 136–144.
 54. Reiss D, Baliga N, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *Bioinformatics* 2006;2(7):280–302.
 55. Wagstaff K, Cardie C, Rogers S, *et al.* Constrained k-means clustering with background knowledge. *Proceedings of the 18th International Conference on Machine Learning*. Williamstown (MA); 2001. pp. 577–584.
 56. Iyigun C, Ben-Israel A. Semi-supervised probabilistic distance clustering and the uncertainty of classification. In: Fink WSA, Lausen B, Ultsch A, editors. *Advances in data analysis, data handling and business intelligence*. New York: Springer; 2010. pp. 3–20.
 57. Basu S, Banerjee A, Mooney R. Active semi-supervision for pairwise constrained clustering. *Proceedings of SIAM International Conference on Data Mining*. Lake Buena Vista (FL); 2004. pp. 333–344.
 58. Everitt B, Landau S, Leese M. *Cluster analysis*. New York: Oxford University Press Inc.; 2001.

COGNITIVE MAPPING AND STRATEGIC OPTIONS DEVELOPMENT AND ANALYSIS (SODA)

ION GEORGIU
Fundação Getulio Vargas,
Departamento de Informatica e
Metodos Quantitativos, Escola de
Administração de Empresas de
São Paulo, São Paulo,
São Paulo, Brazil

The usefulness of cognitive mapping has gained currency since the mid-1980s [1–7]. A number of variants exist. Social science methodology, in conjunction with information visualization, has put forward concept mapping [8]. The Florida University System incorporates the Institute for Human and Machine Cognition which has devised, and made freely available, the *Cmap tools* program that “empowers users to construct, navigate, share, and criticize knowledge models represented as concept maps” (see <http://www.ihmc.us/>). A string of mind-mapping gurus has emerged, the most famous being Tony Buzan, the self-proclaimed “inventor of mind mapping” ([9]; <http://www.imindmap.com>). Furthermore, irrespective of whether one calls them concept maps, cognitive maps, or mind maps, particular structures have spawned spidergrams, bubble diagrams, logic diagrams, and tree diagrams (to name but a few). Clearly, cognitive mapping is a flourishing industry, one that is driven by a market demand for such tools.

Strategic options development and analysis, or SODA, is one of the most prominent problem structuring methods (PSMs) developed in British operational research [10–13]. As with all PSMs, SODA is used for group decision making in situations characterized by nontrivial uncertainty and complexity that is not amenable to formal algorithmic modeling [14–17]. SODA incorporates a particular version of cognitive mapping as its main tool. The process of designing SODA

maps, as well as their content, offer the users a transparent interface through which they can explore, learn about, and consequently take more confident decisions to improve, or otherwise change, a problematic situation. Although mainly applied in group decision making situations, SODA has also been applied to the analysis of documents [18,19].

What differentiates SODA’s cognitive mapping approach is its basis in George Kelly’s psychological theory of personal constructs [20–22]. Although SODA does not pretend to appropriate Kelly’s theory *en masse*, it does borrow two key ideas: one theoretical and one procedural.

THEORETICAL APPROACH OF SODA

George Kelly’s theory is highly developed, so much so that there are international journals, such as the *International Journal of Personal Construct Psychology*, dedicated singularly to his psychological approach. As the title of his theory indicates, Kelly’s central theme is the manner in which human beings understand the world through mental *constructs*. Unlike a concept, a construct is dichotomously comprised of two poles, the relationship between them being one of contrast or alternative-ness. The use of binary oppositions to facilitate the elicitation of meaning was especially prominent in semiotics and linguistics during the first half of the twentieth century when Kelly was developing his theory. By the time he completed it, the semiotician Greimas [23] had begun his studies on the “semiotic square” that allowed for two simultaneous oppositions, each of which occupied the corner of a square. Although developments such as this one have been found to elicit richer understandings than a bipolar concentration [24], Kelly’s theory continues to attract attention, not least due to its perceived philosophical richness [25] and psychotherapeutical relevance [26].

George Kelly was interested in uncovering the meaning behind what people say so as

to minimize ambiguity. He noted that problems tended to be analyzed or interpreted according to the type of analyst one consulted. So, for example, if one took a problem to a Freudian analyst, it would be structured and analyzed according to Freudian principles. A behaviorist, in turn, would most probably analyze the situation and draw conclusions in terms of conditioning. All of this implies that the frame of reference of the analyst delimits what is perceived, how it is described, and what the ultimate prescription might be. Kelly's objective was to devise a theory, coupled with an analytical technique, which would remove—as far as possible—the analyst's frame of reference and so undertake problem description and resolution from the client's/patient's point of view.

This implies a significant change in the role of the analyst. The analyst, once perceived as some type of specialist in the contents of the mind, was now a process facilitator specializing in structuring the client's thoughts as the client sees them. This view of analysis and of the role of the analyst underpins SODA: what clients need is help in structuring complex perceptions so that clients themselves, with facilitative assistance, can resolve the problem using this structure.

Kelly developed a number of intricate analytical tools for his theory, the most famous of which is the repertory grid [27,28]. SODA does not use such tools [29]. The maps designed through SODA, however, are models whose essential structure is that of a graph (nodes and links) or, more exactly, a directed graph (also known as a digraph) [30,31]. As such, SODA maps are amenable to the powerful analytical tools of digraph theory [32–36], as well as givens–means–ends analysis [7]. SODA maps have also served as a basis for the design of system dynamics models [37–42].

PROCEDURE OF BIPOLAR CONSTRUCT DESIGN

SODA adheres to bipolar construct design, making for a *construct* mapping methodology, as opposed to one that involves concept mapping. The reason for this is that, for SODA,

language is the basic currency that gives meaning to people's concerns but it is also a problematic one. Language consists of words, many of which can have multiple connotations. One manner of minimizing the breadth of interpretation is to offer an alternative word or phrase that can serve to highlight what is meant by the description originally given [28, p. 11]. Say, for example, that a person is described as *pleasant*. In itself, this description is vague, not only because the term *pleasant* has numerous synonyms that open up a field of subtle variations in understanding, but because no alternatives have been put forth against which the meaning of *pleasant* can be deduced. To offer a strictly negative alternative, moreover, such as “not pleasant,” is rather useless in trying to understand what is being meant. Is the person pleasant in the sense that they are polite, or charming, or alluring, or perhaps gentle? Or is the person pleasant as opposed to being rude or perhaps exciting? A more precise alternative is required in order to obtain at least the flavor of what is meant, and this is the function of constructs. Constructs are designed with two poles, whereby the second pole serves to clarify what is meant by the first pole. For instance, to say that the person is pleasant as opposed to alluring, or pleasant as opposed to rude, already offers more precise meanings in each case. SODA would write such constructs as follows:

person is pleasant . . . person is alluring
person is pleasant . . . person is rude

The three dots serve to distinguish the two poles of a construct. The alternative pole in each case serves to contextualize and refine the understanding of the primary pole. In essence, then, the deduction or elicitation of secondary poles is crucial to the use of SODA. In practice, of course, people do not talk or write in constructs, and pushing them for alternative poles can quickly become tedious. Often, secondary poles are deduced by a SODA mapper and clarified later by the client. At times, simple negatives are used due to insuperable ambiguities. In other cases, they are left out altogether but, as will be discussed, this does not do justice to SODA as a distinct methodology.

In summary, SODA dissects an actor's oral or written description and translates it into a set of bipolar constructs. The constructs are then causally connected in a manner that reflects the actor's descriptive logic. When complete, the map can be read independently of its sources. SODA offers a qualitative, bipolar, cartographic approach to complex situations that is amenable to quantitative analysis and simulation modeling.

UNDERSTANDING SODA MAPS

Figure 1 is a SODA map *about* SODA mapping. In considering it, one can therefore appreciate both, the nature of SODA maps, as well as how to read them. To begin with, note the following:

- The numbering of constructs is purely random and serves only to reference them (they will be referenced in italics throughout this article).
- The arrows or links come with a negative sign, or are otherwise unsigned.
 - An unsigned link between two constructs indicates that their respective

primary or secondary poles are to be read in order, from the arrow's tail to the arrow's head.

- An arrow signed with a negative symbol ("–") indicates that, at that point, one must switch poles when following the argument along the link.

As an illustration, begin at construct 3 of Fig. 1 and follow the links through constructs 4, 2, 12, 6, ending at construct 8. An actor attempts to explain a problematic situation, but the description offered is not easy to follow (3). The complexity and uncertainty of the situation inhibit the actor from articulating a logical train of thought, resulting in a storm of information, a tornado of thought, so to speak (4). There is a need to impose some sort of methodological structure on the information offered by the actor (2), calling for a methodologically guided manipulation of the actor's description (12)—in this case, the use of SODA mapping. Through such methodological manipulation, knowledge of the situation is seen in a new light (6).

Note that this reading began by considering the primary pole of construct 3 and

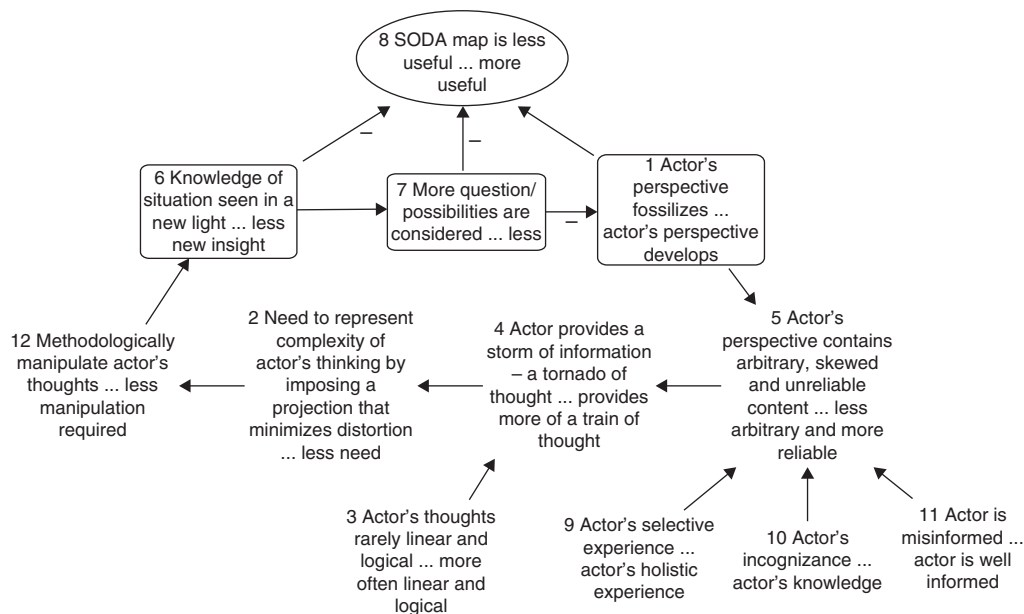


Figure 1. Understanding SODA through a SODA map.

continued by referring to the primary poles of subsequent constructs, the reason being that the arrows in the sequence are unsigned. Due to the negative arrow between constructs 6 and 8, however, there is a switch from the primary pole of construct 6 to the secondary pole of construct 8: as long as knowledge of the situation is seen in a new light (6), the SODA map is more useful (8). If, on the other hand, there is less new insight emerging (secondary pole of construct 6), the SODA map's usefulness decreases proportionally (primary pole of construct 8), or equally, the SODA map has served its purpose.

The constructs in a SODA map are typically decisions available to be taken, or consequences that may result from decisions taken. Apart from making explicit the logical dependencies between constructs, a SODA map also renders explicit the structural significance of constructs. Constructs may be structurally categorized according to certain basic types: *tails*, *heads*, *strategic options*, *implosions*, *explosions*, and *dominants*. They may also participate in *loops*.

Tails

Tails have no constructs leading into them. In the language of graph theory, they are *transmitters* whereby their in-degree is zero and their out-degree is positive [30, p. 17]. In SODA, they are otherwise known as prime causes. In Fig. 1, constructs 3, 9, 10, and 11 are tails. They indicate that SODA mapping is primarily (but not exclusively) useful when actors find difficulty in articulating their thoughts in a linear or logical manner (3), when they have a selective, as opposed to a holistic, appreciation of the situation (9), when their level of knowledge is relatively low (10), and when they have been victims of misinformation (11).

Heads

Heads have no constructs leading out of them. In the language of graph theory, they are *receivers* whereby their out-degree is zero and their in-degree is positive [30, p. 17]. They reflect objectives, outcomes, results, or consequences stemming from the dependency paths of arrows that lead into them. When

first looking at a SODA map, the heads will usually offer a good idea of what it is about. Figure 1 has only one head, construct 8, from which the user quickly infers that the figure is a map about the usefulness of SODA maps. Heads may be highlighted for emphasis (Fig. 1 encircles construct 8 in an oval). Furthermore, large maps of complex situations usually have numerous heads, indicating the requirement to address multiple, equally necessary, and at times conflicting, objectives usually measurable on different dimensions that preclude trade-offs between them.

Strategic Options

In SODA, those constructs with immediate links to a head are called *strategic options*, from which the methodology takes its name (there is no equivalent term in graph theory). They reflect the options available through which a particular result (head) may materialize or, in other words, the immediate influences that will govern which pole of the outcome will eventually happen. We find, for instance, that the usefulness of a SODA map (8) depends upon its

- providing new insight (6);
- stimulating more questions and possibilities about the problematic situation (7);
- mapping developing perspectives about a situation (1).

Due to their perceived immediate influence upon a head, strategic options may also be emphasized diagrammatically (Fig. 1 has employed rounded rectangles).

Implosions

Implosions are constructs with a relatively high number of constructs leading directly into them. In the language of graph theory, they have a relatively high in-degree or *in-bundle* [30, p. 17]. Social network analysis [43, p. 202], with regard to directed networks (i.e., digraphs), would say that such constructs have *degree prestige* that is relatively higher to other constructs. An implosion indicates a major effect. It

is a construct affected by multiple other constructs and, by extension, multiple areas of the map. It is where a number of issues culminate or converge. In Fig. 1, construct 5 has an in-degree of four, while the only other construct that comes close is the head (8) with in-degree of three. The implosion of construct 5 serves to highlight the factors that lead to an actor's arbitrary, skewed, and unreliable understanding of a situation: the actor's limited experience (9), incognizance (10), and misinformation (11), as well as their set (and perhaps inflexible) perspective (1).

Explosions

Explosions are constructs with a relatively high number of constructs directly leading out of them. In the language of graph theory, they have a relatively high out-degree or *out-bundle* [30, p. 17]. Social network analysis [43, p. 199], with regard to directed networks, would say that such constructs have *degree centrality* that is relatively higher to other constructs. An explosion indicates a major cause. It is a construct that affects multiple other constructs and by extension, multiple areas of the map. It is from where a number of multiple issues stem or diverge. In Fig. 1, the three strategic options (6, 7, 1) all share the same, relatively higher out-degree. To take but one example, construct 7 indicates that with the consideration of more questions and possibilities about the problematic situation, the usefulness of SODA maps is increased (8), and an actor will be open to new perspectives (1).

Dominants

Dominants are constructs with a relatively high total number of constructs leading into them and leading out of them. In the language of graph theory, they have a relatively high degree (sum of in-degree and out-degree) [30, p. 17]. Social network analysis [43, p. 173], with regard to the *underlying graph* of the digraph [44, p. 92], would call such constructs *central*. A construct with a relatively high degree indicates cognitive centrality of an issue in an actor's perceptions, and/or central relevance of an issue to the situation in question. Depending on the

balance between in-degree and out-degree, a dominant will affect, and be affected by, multiple constructs and by extension, multiple areas of the map. Whereas heads offer a good initial idea of what a map is about, dominants offer a good indication of the major issues that must be tackled in order to reach the heads. In Fig. 1, construct 5 has the highest degree of the map. It indicates that a major issue in judging the usefulness of SODA maps is their ability to render actors' perspectives less arbitrary and more reliable.

Loops

In SODA, loops are not to be confused with the manner in which this term is understood in digraphs. The equivalent term in digraph theory would be *cycles* [44, p. 95]. In SODA, moreover, loops are of primary interest for identifying feedback dynamics.

It is difficult to perceive feedback dynamics mentally, without the aid of some interface that renders them explicit. And yet, it is exactly such dynamics that can sabotage an otherwise seemingly logical set of decisions. The identification and analysis of feedback loops in any map is therefore important. More ominously, feedback loops can serve to identify areas of uncontrolled degenerative or regenerative dynamics, pointing to the ultimate collapse of the situation under consideration.

To begin with, consider the relatively more benign case of a loop that is self-controlled. This occurs when a loop contains an odd number of negative links. Any perturbation in the state of the variables within the loop will result in stabilizing dynamics to bring activity under control, not unlike a thermostat. This can be illustrated in a simple decision making situation faced by a bus company at various times of the day, such as that depicted in Fig. 2. Here, as the number of passengers per bus increases (17), the pressure to service the route increases (18), thus the decision is taken to make more buses available on the route (19). Eventually, and due to the availability of more buses, the number of passengers per bus decreases (17), easing the service pressure (18), and thus leading the company to reduce the number of buses on the route (19). Of course, as the

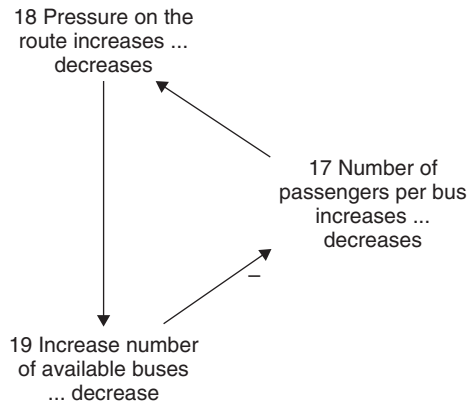


Figure 2. Self-controlled feedback loop.

number of passengers picks up again (17), the company will eventually find itself needing to make available more buses (19). There is a seemingly regular swing between making available more or fewer buses, a swing controlled by the effects on service pressure (18).

This pendulum effect is also apparent in Fig. 1. Beginning from the primary pole of construct 6 and reading through to constructs 7, 1, 5, 4, 2, and 12, leads to the secondary pole of construct 6, and ultimately to SODA maps being less useful (8). Remaining in the same circuit but now beginning with the secondary pole of construct 6, will result in SODA maps being more useful. Indeed, from this loop, one can also infer that the usefulness of a SODA map (8) is a function of the dynamic interrelations between three issues: the provision of insight (6, 7), the state of an actor's understanding (1, 5, 4, 2), and the methodological manipulation afforded by the approach (12).

Self-controlled loops do not present problems as long as the exhibited control is desired. More problematic are those loops constituted completely by unsigned links, or by an even number of negative links. A loop with either of these characteristics suggests either generative growth or degenerative decline, depending on the poles that are followed within each construct. Figure 3 extends the bus company situation by taking into account throughput time (20), passenger satisfaction (21), and the decisions of the public to use or not to use the bus as a viable means of transportation (22). Reading

through the figure of eight, one finds that the number of buses on the route is either increasing or decreasing uncontrollably, depending on which poles are followed. Furthermore, although the bus company might like the idea of having more passengers, hence more revenues, and more growth, the feedback loop highlights that any process without controls will inevitably collapse.

An interesting aspect of Fig. 3 is the dominant construct (17) regarding the number of passengers per bus. This construct participates in two self-controlled loops, one on either side of it. When the loops are conjoined into the figure of eight, however, any controlling dynamics are cancelled out, resulting in the inevitable collapse of this situation either through uncontrolled growth or uncontrolled decline. In other words, when viewing a seemingly self-controlled loop, it is worth taking into account its wider context to ensure that it might not actually be participating in some wider and more dangerous cycle.

GROUP DECISION MAKING

If complexity and uncertainty already pose a considerable challenge for an individual decision maker, then the challenge is compounded in group decision making where different actors, with different views, and perhaps from different organizations or departments, have to work together toward a resolution that accommodates all those concerned. The classic approach of SODA under these circumstances is to begin with the design of individual actors' maps. This allows each group member to at least have a model of their own understanding of the situation. An individual map is usually designed in conjunction with a semistructured interview or discussion with the person concerned. As individual maps are designed and collected together, points of commonality between them will necessarily emerge since all the individuals concerned share the same problematic situation. These commonalities may be explicit constructs or clusters of constructs that identify similar issues. They will form the points that link one individual map to another and hence allow the emergence of a group map.

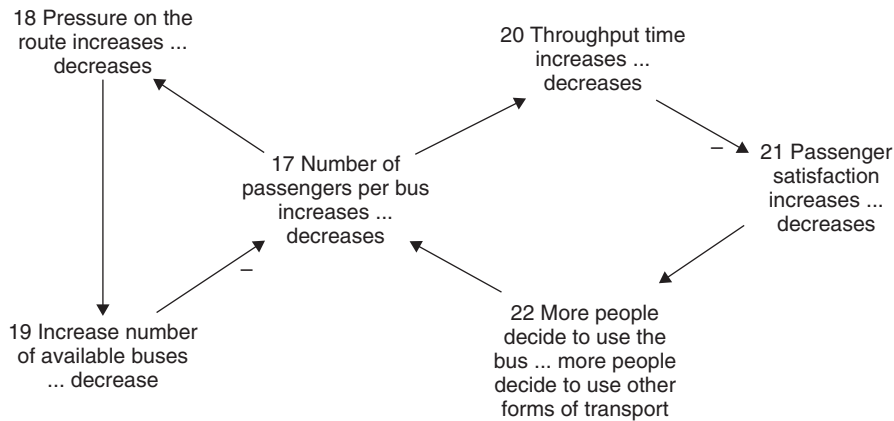


Figure 3. Uncontrolled regenerative/degenerative feedback loop.

This group map, or merged map, will not necessarily reflect how any one individual understands the situation, but will reflect the group understanding. The understanding of a group is always somewhat more abstract and difficult to grasp, and a SODA map offers a means for concretely visualizing such an understanding. It also allows each individual to perceive relationships between decisions that might otherwise have remained obscure, as well as allowing for an appreciation of the understanding of others. As such, the merged SODA map is a tool for organizational, or group, learning that enables a group to take decisions more confidently.

CRITICAL APPRECIATION

SODA is one of the most frequently used PSMs [45]. Computer software, called Decision Explorer[®], has also been designed to facilitate construction and analysis of SODA maps (see www.banxia.com). This is especially useful for the case of large maps that contain hundreds of constructs. Still, there remain significant areas for further research in SODA, especially related to graph theory, diagrammatic representation, and the use of constructs.

As mentioned earlier, recent years have seen some developments in methods for analyzing SODA maps, especially from a graph theoretical point of view. In comparison, however, to other approaches that utilize

graph theory—such as, for example, social network analysis [43]—the SODA literature indicates relatively little use of this branch of mathematics. It can be shown that SODA maps share the four primitives and four axioms of digraphs [30, p. 9]. This being the case, research needs to be carried out on the numerous graph theoretic analyses that are available, in order to uncover which ones are of special relevance to SODA analyses. More research is also required on how matrices, graph mining [46], and blockmodeling [47] can inform SODA map analyses.

The diagramming process itself has also been left largely unaddressed by SODA. Graph drawing [48] is a novel area within graph theory and an interdisciplinary excursion here will undoubtedly benefit the representational design of SODA maps. The broader field of information visualization [49–53] also offers a wealth of research, which might not only be relevant to SODA map representations in general, but also undoubtedly useful for the further development of the Decision Explorer[®] software.

A scan of the literature shows that the use of bipolar constructs in SODA mapping appears less frequently than the use of single-issue concepts. This is surprising for an approach that explicitly draws from Kelly's theory. It also does not serve to differentiate SODA very effectively from other cognitive mapping approaches. One attempt at differentiation was made in the

late 1990s whereby SODA was rechristened as “journey making” and aligned closely with strategic management [54]. In the span of 500 pages, however, the idea of a construct got short shrift and only four diagrams contained any constructs at all (see pp. 96, 287, 291, 295). Since SODA is explicitly concerned with perceptions and meanings, rather than seemingly relegating the idea of construct to the background, research would arguably be better served by exploring this idea to its limits, alongside the contributions of semiotics and perhaps even linguistics.

CONCLUSION

SODA is a cognitive mapping approach whose qualitative content is structured in such a manner as to render it especially amenable to quantitative analysis. Its diagrammatic basis makes it transparent to users, requiring relatively little prior training to understand the approach. The maps themselves offer a means for describing situations and thus facilitate learning and understanding.

The fact that maps are amenable to quantitative analysis must not be taken as in any way minimizing their qualitative content. Measurements, of any sort, do not provide answers in themselves, and much less should they be used as a substitute for thinking through the situation in question. Measurements are to be used in conjunction with a more holistic understanding of the map and the situation it is describing, so that informed conclusions can be drawn.

Furthermore, the maps should be treated less as models of cognition—that is, as psychological tools—and more as means for investigating problematic situations. For, although the perspectives of decision makers are what maps record, ultimately maps describe situations.

Finally, relatively large maps are not necessarily more complex than smaller maps. They may, for example, have no multiple and interrelated feedback loops. Complexity is not dependent on the size of any particular variable, but on the interrelationship of variables.

More generally, faced with complexity and uncertainty, and prior to any prescriptive action, a descriptive and exploratory approach is required for initiating focused debate and targeted exploration, two essentials of effective decision making. This is exactly what SODA offers.

REFERENCES

1. Bryant J. Modelling alternative realities in conflict and negotiation. *J Oper Res Soc* 1984; 35(11):985–993.
2. Daniels K, Johnson G. On trees and triviality traps: locating the debate on the contribution of cognitive mapping to organizational research. *Organ Sci* 2002;23(1):73–81.
3. Fiol CM. Maps for managers: where are we? Where do we go from here? *J Manage Stud* 1992;29(3):267–285.
4. Kitchin RM. Cognitive maps: what are they and why study them? *J Environ Psychol* 1994; 14(1):1–19.
5. Langfield-Smith K. Exploring the need for a shared cognitive map. *J Manage Stud* 1992; 29(3):349–368.
6. Nicolini D. Comparing methods for mapping organizational cognition. *Organ Stud* 1999;20(5):833–860.
7. Tegarden DP, Sheetz SD. Group cognitive mapping: a methodology and system for capturing and evaluating managerial and organizational cognition. *OMEGA Int J Manage Sci* 2003;31(2):113–125.
8. Kane M, Trochim WMK. *Concept mapping for planning and evaluation*. Thousand Oaks (CA): Sage; 2007.
9. Buzan T. *How to mind map*. London: Thorsons; 2002.
10. Eden C. Using cognitive mapping for strategic options development and analysis (SODA). In: Rosenhead J, editor. *Rational analysis for a problematic world: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley; 1989. pp. 21–42.
11. Eden C, Simpson P. SODA and cognitive mapping practice. In: Rosenhead J, editor. *Rational analysis for a problematic world: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley; 1989. pp. 43–70.
12. Eden C, Ackermann F. SODA—the principles. In: Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited*:

- problem structuring methods for complexity, uncertainty and conflict. Chichester: Wiley; 2001. pp. 21–41.
13. Ackermann F, Eden C. SODA—journey making and mapping in practice. In: Rosenhead J, Mingers J, editors. Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict. Chichester: Wiley; 2001. pp. 44–60.
 14. Eden C. Cognitive mapping. *Eur J Oper Res* 1988;36(1):1–13.
 15. Eden C. Analyzing cognitive maps to help structure issues or problems. *Eur J Oper Res* 2004;159(3):673–686.
 16. Eden C, Sims D. Subjectivity in problem identification. *Interfaces* 1981;11(1):68–74.
 17. Eden C, Huxham C. Action-oriented strategic management. *J Oper Res Soc* 1988;39(10):889–899.
 18. Cossette P. Analysing the thinking of FW Taylor using cognitive mapping. *Manage Decis* 2002;40(2):168–182.
 19. Klein JH, Cooper DF. Cognitive maps of decision-makers in a complex game. *J Oper Res Soc* 1982;33(1):63–71.
 20. Kelly GA. The psychology of personal constructs. 2nd ed. London: Routledge; 1955/1991.
 21. Kelly GA. A theory of personality: the psychology of personal constructs. London: Norton; 1963.
 22. Kelly GA. A brief introduction to personal construct theory. In: Bannister D, editor. Perspectives in personal construct theory. London: Academic Press; 1970. pp. 1–30.
 23. Greimas AJ. Structural semantics: an attempt at a method. Lincoln: University of Nebraska Press; 1984.
 24. Lévi-Strauss C. Structural anthropology. New York: Basic Books; 1958.
 25. Warren B. Philosophical dimensions of personal construct psychology. London: Routledge; 1998.
 26. Fransella F, Dalton P. Personal construct counselling in action. London: Sage; 2000.
 27. Fransella F, Bell R, Bannister D. A manual for repertory grid technique. Chichester: Wiley; 2004.
 28. Jankowicz D. The easy guide to repertory grids. Chichester: Wiley; 2004.
 29. Eden C, Jones S. Using repertory grids for problem construction. *J Oper Res Soc* 1984;35(9):779–790.
 30. Harary F, Norman RZ, Cartwright D. Structural models: an introduction to the theory of directed graphs. Chichester: Wiley; 1965.
 31. Bang-Jensen J, Gutin G. Digraphs: theory, algorithms and applications. London: Springer; 2002.
 32. Langfield-Smith K, Wirth A. Measuring differences between cognitive maps. *J Oper Res Soc* 1992;43(12):1135–1150.
 33. Montibeller G, Belton V. Causal maps and the evaluation of decision options—a review. *J Oper Res Soc* 2006;57(7):779–791.
 34. Wang S. A dynamic perspective of differences between cognitive maps. *J Oper Res Soc* 1996;47(4):538–549.
 35. Eden C, Ackermann F, Cropper S. The analysis of cause maps. *J Manage Stud* 1992;29(3):309–324.
 36. Montibeller G, Belton V, Ackermann F, *et al.* Reasoning maps for decision aid: an integrated approach for problem-structuring and multi-criteria evaluation. *J Oper Res Soc* 2008;59(5):575–589.
 37. Howick S. Using system dynamics to analyse disruption and delay in complex projects for litigation: can the modelling purposes be met? *J Oper Res Soc* 2003;54(3):222–229.
 38. Williams T, Ackermann F, Eden C. Structuring a delay and disruption claim: an application of cause-mapping and system dynamics. *Eur J Oper Res* 2003;148(1):192–204.
 39. Howick S, Eden C. The impact of disruption and delay when compressing large projects: going for incentives? *J Oper Res Soc* 2001;52(1):26–34.
 40. Eden C, Williams T, Ackermann F, *et al.* The role of feedback dynamics in disruption and delay on the nature of disruption and delay (D&D) in major projects. *J Oper Res Soc* 2000;51(3):291–300.
 41. Bennett PG, Ackermann F, Eden C, *et al.* Analysing litigation and negotiation: using a combined methodology. In: Mingers J, Anthony G, editors. Multimethodology: the theory and practice of combining management science methodologies. Chichester: Wiley; 1997. pp. 59–88.
 42. Eden C. Cognitive mapping and problem structuring for system dynamics model building. *Syst Dyn Rev* 1994;10(2–3):257–276.
 43. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge: Cambridge University Press; 1994.
 44. Aldous JM, Wilson RJ. Graphs and applications: an introductory approach. London: Springer; 2000.

45. Mingers J, Rosenhead J. Problem structuring methods in action. *Eur J Oper Res* 2004; 152(3):530–554.
46. Cook DJ, Holder LB. Mining graph data. Chichester: Wiley; 2007.
47. Doreian P, Batagelj V, Ferligoj A. Generalized blockmodeling. Cambridge: Cambridge University Press; 2005.
48. di Battista G, Eades P, Tamassia R, *et al.* Graph drawing: algorithms for the visualization of graphs. New Jersey: Prentice-Hall; 1999.
49. Kosslyn SM. Graph design for the eye and mind. Oxford: Oxford University Press; 2006.
50. Kosslyn SM. Elements of graph design. New York: WH Freeman and Company; 1994.
51. Glasgow J, Narayanan NH, Chandrasekaran B, editors. Diagrammatic reasoning: cognitive and computational perspectives. Menlo Park (CA): AAAI Press/MIT Press; 1995.
52. Bertin J. Semiology of graphics: diagrams, networks, maps. Redlands (CA): ESRI Press; 2010.
53. Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire (CT): Graphics Press; 2001.
54. Eden C, Ackermann F. Making strategy: the journey of strategic management. London: Sage; 1998.

COHERENT SYSTEMS

RONALD C. NEATH
Department of Statistics and
Computer Information
Systems, Baruch College,
City University of New York,
New York

ANDREW A. NEATH
Department of Mathematics and
Statistics, Southern Illinois
University, Edwardsville,
Illinois

PROPERTIES OF COHERENT SYSTEMS AT A FIXED POINT IN TIME

The first formal study of coherent systems was conducted by Birnbaum *et al.* [1], whose work led to a unified framework for analyzing an engineering system in terms of its components.

Consider a system with n components, which we will examine at a particular instant in time. For each $i = 1, \dots, n$ define x_i , the state of the i th component at that instant, by

$$x_i = \begin{cases} 1, & \text{component } i \text{ is working;} \\ 0, & \text{component } i \text{ is in failure.} \end{cases} \quad (1)$$

The state of the system depends solely on the states of its components; specifically, the system works if certain combinations of components work.

Example 1. Natvig [2] gives the following example, a power supply system represented by Fig. 1. The system works (is able to provide power) as long as it can (i) generate power, and (ii) transmit it. In Fig. 1, component 1 represents an offsite power source and components 2 and 3 are transformers; component 6, an onsite emergency generator, backs up that part of the process. Components 4, 5, and 7 are cables.

The Structure Function

We define the structure function $\phi : \{0, 1\}^n \rightarrow \{0, 1\}$ by

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{the system is working;} \\ 0, & \text{the system is in failure;} \end{cases} \quad (2)$$

where $\mathbf{x} = (x_1, \dots, x_n)$.

The two simplest systems are the *series system* and the *parallel system*. We state here the structure functions for those two cases, and also consider the *k-out-of-n system*.

A series system works if and only if all of its components work. Thus

$$\phi_{\text{series}}(\mathbf{x}) = \min \{x_1, \dots, x_n\} = \prod_{i=1}^n x_i.$$

A parallel system works if at least one of its components works:

$$\phi_{\text{parallel}}(\mathbf{x}) = \max \{x_1, \dots, x_n\} = 1 - \prod_{i=1}^n (1 - x_i).$$

Define the *cup operator* $\sqcup$ through its Boolean addition properties

$$\begin{aligned} 0 \sqcup 0 &= 0; & 0 \sqcup 1 &= 1; & 1 \sqcup 0 &= 1; \\ 1 \sqcup 1 &= 1. \end{aligned}$$

Then, for any $\mathbf{x} \in \{0, 1\}^n$,

$$\prod_{i=1}^n x_i = x_1 \sqcup x_2 \sqcup \dots \sqcup x_n = 1 - \prod_{i=1}^n (1 - x_i)$$

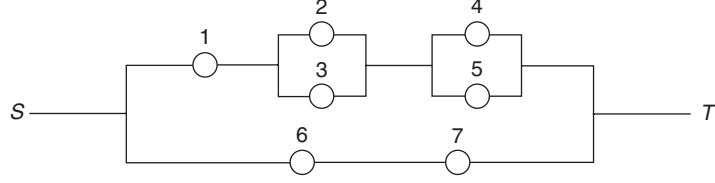
and thus, we can write $\phi_{\text{parallel}}(\mathbf{x}) = \prod_{i=1}^n x_i$.

A *k-out-of-n system* works if and only if at least k of its components work. For example, the structure function for a two-out-of-three system can be written as

$$\phi(\mathbf{x}) = (x_1 x_2) \sqcup (x_1 x_3) \sqcup (x_2 x_3).$$

Note that the *k-out-of-n system* includes the series (*n-out-of-n*) and parallel (*one-out-of-n*) systems as special cases.

Figure 1. Power supply system for Example 1, taken from Ref. 2. System works if and only if there is a working connection from S to T .



Example 2. The structure function for Natvig's [2] power supply system of Example 1 is given by

$$\phi(\mathbf{x}) = (x_1(x_2 \sqcup x_3)(x_4 \sqcup x_5)) \sqcup (x_6x_7).$$

Path and Cut Sets

In this section, we develop a representation of the structure function based on the identification of subsets of the components whose (non-) failure guarantees the system's (non-) failure; see also Ref. 3, Section 1.3, or 4.

Given a system of n components with structure function ϕ , the vector $\mathbf{x} \in \{0, 1\}^n$ is called a *path vector* if $\phi(\mathbf{x}) = 1$, and a *cut vector* if $\phi(\mathbf{x}) = 0$. Define

$$C_1(\mathbf{x}) = \{i : x_i = 1\} \text{ and } C_0(\mathbf{x}) = \{i : x_i = 0\}.$$

If $\mathbf{x}$ is a path vector, $C_1(\mathbf{x})$ is the corresponding *path set*; if $\mathbf{x}$ is a cut vector, $C_0(\mathbf{x})$ is the corresponding *cut set*.

We say that $\mathbf{x}$ is a *minimal path vector* if $\phi(\mathbf{x}) = 1$ and $\phi(\mathbf{y}) = 0$ for any $\mathbf{y}$ such that each $y_i \leq x_i$ with strict inequality for some i . Similarly, $\mathbf{x}$ is said to be a *minimal cut vector* if $\phi(\mathbf{x}) = 0$ and $\phi(\mathbf{y}) = 1$ for any $\mathbf{y}$ satisfying $y_i \geq x_i$ for each i , with strict inequality for some i . Thus, a minimal path vector represents a state in which the system works, but for which the failure of even a single working component would cause system failure. A minimal cut vector represents a failure state in which the repair of even a single non-working component would lead to the system working. If $\mathbf{x}$ is a minimal path vector, $C_1(\mathbf{x})$ is called a *minimal path set*; if $\mathbf{x}$ is a minimal cut vector $C_0(\mathbf{x})$ is called a *minimal cut set*.

Denote the minimal path sets of a system by $\mathcal{P}_1, \dots, \mathcal{P}_r$, and the minimal cut sets by $\mathcal{C}_1, \dots, \mathcal{C}_s$. Then the system is working if and only if the components of at least one minimal path set are all working. We can then write

$$\phi(\mathbf{x}) = \bigsqcup_{j=1}^r \prod_{i \in \mathcal{P}_j} x_i. \quad (3)$$

Similarly, the system is working if and only if at least one component from each minimal cut set is working, thus

$$\phi(\mathbf{x}) = \prod_{j=1}^s \bigsqcup_{i \in \mathcal{C}_j} x_i. \quad (4)$$

Example 3. Consider the power supply system from Example 1. From Fig. 1, the minimal path sets of this system are easily seen to be

$$\mathcal{P}_1 = \{1, 2, 4\}; \quad \mathcal{P}_2 = \{1, 2, 5\}; \quad \mathcal{P}_3 = \{1, 3, 4\}; \\ \mathcal{P}_4 = \{1, 3, 5\}; \text{ and } \mathcal{P}_5 = \{6, 7\}.$$

The minimal cut sets for this system are given by

$$\mathcal{C}_1 = \{1, 6\}; \quad \mathcal{C}_2 = \{1, 7\}; \quad \mathcal{C}_3 = \{2, 3, 6\}; \\ \mathcal{C}_4 = \{2, 3, 7\}; \quad \mathcal{C}_5 = \{4, 5, 6\}; \text{ and } \mathcal{C}_6 = \{4, 5, 7\}.$$

The reader can verify that both Equations (3) and (4) return the expression given in Example 2.

Coherence

Define

$$(1_i, \mathbf{x}) = (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n)$$

and

$$(0_i, \mathbf{x}) = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n).$$

Then for any fixed $i = 1, \dots, n$, a structure function ϕ can be expressed as

$$\phi(\mathbf{x}) = x_i \phi(1_i, \mathbf{x}) + (1 - x_i) \phi(0_i, \mathbf{x}).$$

This representation is known as the *pivotal decomposition* of a structure function.

Definition 1. The i th component of a system with structure function ϕ is *irrelevant* if $\phi(1_i, \mathbf{x}) = \phi(0_i, \mathbf{x})$ for all $\mathbf{x}$. Otherwise the i th component is said to be *relevant*.

Definition 2. A structure function is said to be *monotone* if for any $\mathbf{x}$ and $\mathbf{y}$ such that $x_i \leq y_i$ for each $i = 1, \dots, n$, $\phi(\mathbf{x}) \leq \phi(\mathbf{y})$.

Thus, a component is relevant when there exists at least one case where the state of the system depends on whether the component is working or in failure, and a structure function is monotone when the state of the system can never be improved by the failure of a component. We can now give a formal definition of a coherent system.

Definition 3. A system is *coherent* if each of its components is relevant and its structure function is monotone.

A coherent system consists of components that when working, never harm the system and improve the system in at least one instance. Thus, each working component is beneficial to the system. It is quite rare to find a structure of components designed in practice that is not coherent. For this reason, the study of engineering systems can be narrowed to the study of coherent systems.

For a system of $n = 3$ components, there are five coherent structures. We have encountered three already, in the series, parallel, and two-out-of-three systems. Two more coherent systems are given by $\phi(\mathbf{x}) = x_1 \bigcup (x_2 x_3)$ and $\phi(\mathbf{x}) = x_1 (x_2 \bigcup x_3)$. For a system of $n = 4$ components, there are 20 possible coherent structures. A general rule for the number of coherent systems of n components remains an open question.

We now take up the question of comparison between two coherent systems of n components. It seems reasonable to declare system B with structure function ϕ_B better than system A with structure function ϕ_A if $\phi_A(\mathbf{x}) \leq \phi_B(\mathbf{x})$ for all $\mathbf{x} \in \{0, 1\}^n$, with a strict inequality for at least one $\mathbf{x}$. In other words, system B is better than system A if it works

in all the instances where system A works, and at least one case where system A does not. The following theorem asserts that all coherent systems fall somewhere between the series system and the parallel system by this standard.

Theorem 1. For any coherent system in n components with structure function ϕ ,

$$\prod_{i=1}^n x_i \leq \phi(\mathbf{x}) \leq \bigcup_{i=1}^n x_i \quad (5)$$

for any $\mathbf{x} \in \{0, 1\}^n$.

Proof. Define $\mathbf{0} = (0, \dots, 0)$ and $\mathbf{1} = (1, \dots, 1)$. All coherent systems have the property that $\phi(\mathbf{0}) = 0$ and $\phi(\mathbf{1}) = 1$. To prove Relation (5) we must consider three cases.

1. If $\prod x_i = 1$, then $\bigcup x_i = 1$ and $\mathbf{x} = \mathbf{1}$. Since $\phi(\mathbf{1}) = 1$, Relation (5) holds.
2. If $\bigcup x_i = 0$, then $\prod x_i = 0$ and $\mathbf{x} = \mathbf{0}$. Since $\phi(\mathbf{0}) = 0$, Relation (5) holds.
3. If $\prod x_i = 0$ and $\bigcup x_i = 1$, then Relation (5) reduces to $0 \leq \phi(\mathbf{x}) \leq 1$ and holds trivially.

The Reliability of a System

Up to this point, we have conducted only a deterministic analysis of coherent systems, in which the system in question is either working or not, depending on the states of its components. Beginning here and continuing in the next section, we take up a probabilistic study of coherent systems, wherein the component states, and the state of the system, are treated as random variables. Still holding time fixed, we consider a coherent system of n components, and let p_i denote the probability that the i th component is working for $i = 1, \dots, n$; let $\mathbf{p} = (p_1, \dots, p_n)$. We denote the state of component i by X_i , defined identically to Equation (1), but using capital-letter notation to indicate the stochastic nature of X_i . The state of the system is indicated by the random variable $\phi(\mathbf{X})$, where ϕ is again the structure function defined by Equation (2), now taking as its argument the random vector $\mathbf{X} = (X_1, \dots, X_n) \in \{0, 1\}^n$.

Our first goal is to study the probability that the system works as a function of the probabilities that the individual components work.

Definition 4. The *reliability function* of a coherent system ϕ is a mapping $h_\phi : [0, 1]^n \rightarrow [0, 1]$ defined by $h_\phi(\mathbf{p}) = \Pr(\phi(\mathbf{X}) = 1)$.

Assume the component states are stochastically independent, so the X_i are independent Bernoulli random variables with success probabilities p_i for $i = 1, \dots, n$. The state of the system $\phi(\mathbf{X})$ is likewise a Bernoulli random variable, with success probability $h_\phi(\mathbf{p})$. The reliability function of a series system is given by

$$\begin{aligned} h_{\text{series}}(\mathbf{p}) &= \Pr(\phi_{\text{series}}(\mathbf{X}) = 1) = \Pr\left(\prod_{i=1}^n X_i = 1\right) \\ &= \Pr(X_1 = 1, \dots, X_n = 1) \\ &= \prod_{i=1}^n \Pr(X_i = 1) = \prod_{i=1}^n p_i, \end{aligned}$$

and that of a parallel system is found to be

$$\begin{aligned} h_{\text{parallel}}(\mathbf{p}) &= \Pr(\phi_{\text{parallel}}(\mathbf{X}) = 1) = \Pr\left(\bigsqcup_{i=1}^n X_i = 1\right) \\ &= \Pr\left(1 - \prod_{i=1}^n (1 - X_i) = 1\right) \\ &= \Pr\left(\prod_{i=1}^n (1 - X_i) = 0\right) \\ &= 1 - \Pr\left(\prod_{i=1}^n (1 - X_i) = 1\right) \\ &= 1 - \Pr(X_1 = 0, \dots, X_n = 0) \\ &= 1 - \prod_{i=1}^n \Pr(X_i = 0) = 1 - \prod_{i=1}^n (1 - p_i). \end{aligned}$$

Define the cup operator for a vector of probabilities analogously to our earlier definition for zeros and ones, so that $\bigsqcup_{i=1}^n p_i = 1 - \prod_{i=1}^n (1 - p_i)$. Then $h_{\text{parallel}}(\mathbf{p}) = \bigsqcup_{i=1}^n p_i$.

From these two special cases, it might be tempting to believe that $h_\phi(\mathbf{p})$ can be obtained

from $\phi(\mathbf{x})$ simply by replacing the x_i with p_i . However, this is not the case in general.

A pivotal decomposition can be defined for the reliability function. Let $h_\phi(1_i, \mathbf{p}) = \Pr(\phi(1_i, \mathbf{X}) = 1)$ and $h_\phi(0_i, \mathbf{p}) = \Pr(\phi(0_i, \mathbf{X}) = 1)$. Then,

$$\begin{aligned} \Pr(\phi(\mathbf{X}) = 1) &= \Pr(X_i = 1) \Pr(\phi(\mathbf{X}) = 1 | X_i = 1) \\ &\quad + \Pr(X_i = 0) \Pr(\phi(\mathbf{X}) = 1 | X_i = 0) \\ &= p_i \Pr(\phi(1_i, \mathbf{X}) = 1) + (1 - p_i) \Pr(\phi(0_i, \mathbf{X}) = 1) \end{aligned}$$

or

$$h_\phi(\mathbf{p}) = p_i h_\phi(1_i, \mathbf{p}) + (1 - p_i) h_\phi(0_i, \mathbf{p}). \quad (6)$$

An interesting early application of this representation is attributable to Esary and Proschan [5], who called h_ϕ the *structure reliability function* and employed Equation (6) to study properties of upper and lower bounds on the system reliability for a k -out-of- n system. We will make use of Equation (6) in proving the following result.

Theorem 2. The reliability function h_ϕ of a coherent system ϕ in n independent components is strictly increasing in each p_i for all $\mathbf{p} \in (0, 1)^n$.

Proof. From the pivotal decomposition Equation (6),

$$\begin{aligned} \frac{\partial}{\partial p_i} h_\phi(\mathbf{p}) &= h_\phi(1_i, \mathbf{p}) - h_\phi(0_i, \mathbf{p}) \\ &= \Pr(\phi(1_i, \mathbf{X}) = 1) - \Pr(\phi(0_i, \mathbf{X}) = 1). \end{aligned}$$

Since ϕ is monotone in each component, $\phi(1_i, \mathbf{x}) \geq \phi(0_i, \mathbf{x})$ for all $\mathbf{x}$, and thus $\Pr(\phi(1_i, \mathbf{X}) = 1) \geq \Pr(\phi(0_i, \mathbf{X}) = 1)$. Thus, we have

$$\begin{aligned} \frac{\partial}{\partial p_i} h_\phi(\mathbf{p}) &\geq 0 \quad \text{for all } \mathbf{p} \in [0, 1]^n, \\ &\text{for each } i = 1, \dots, n. \end{aligned}$$

Further, since every component of ϕ is relevant, there exists for each i a case $\mathbf{x}^*$ such

that $\phi(1_i, \mathbf{x}^*) > \phi(0_i, \mathbf{x}^*)$; for any $\mathbf{p} \in (0, 1)^n$, we have

$$\Pr(\mathbf{X} = \mathbf{x}^*) = \prod_{i=1}^n p_i^{x_i^*} (1 - p_i)^{1-x_i^*} > 0.$$

Thus,

$$\frac{\partial}{\partial p_i} h_\phi(\mathbf{p}) > 0 \text{ for all } \mathbf{p} \in (0, 1)^n, \\ \text{for each } i = 1, \dots, n.$$

Each working component is beneficial to the system, so an increase in the probability that a component works necessarily increases the probability that the system works.

BEHAVIOR OF COHERENT SYSTEMS OVER TIME

We now consider system reliability as a function of time. In particular, we are interested in what can be deduced about the lifetime of a system from the probability distributions of its components' lifetimes. To that end, consider a system of n components, and let T_i denote the lifetime of component i for $i = 1, \dots, n$. We assume that the T_i are independent random variables with distribution functions $F_i(t) = \Pr(T_i \leq t)$. We denote the *survival functions* by $\bar{F}_i$, so

$$\bar{F}_i(t) = \Pr(T_i > t) = 1 - F_i(t),$$

for each $i = 1, \dots, n$.

Suppose we start the system running at time $t = 0$ with all components working, that is, we assume that each $F_i(0) = 0$. Then the i th component is still working as of time $t > 0$ if and only if its lifetime surpasses t (we assume that once a component fails, it remains in failure until the system is restarted). If we let

$$X_i(t) = \begin{cases} 1, & \text{component } i \text{ working at time } t; \\ 0, & \text{otherwise;} \end{cases}$$

then the probability that component i still works at time t is given by $p_i(t) = \Pr(X_i(t) = 1) = \bar{F}_i(t)$ for each $i = 1, \dots, n$.

Let T denote the lifetime of a coherent system with independent components, structure

function ϕ , and system reliability function h . Define $\mathbf{X}(t) = (X_1(t), \dots, X_n(t))$, and the probability that the system still works at time t is given by $\Pr(\phi(\mathbf{X}(t)) = 1) = h(p_1(t), \dots, p_n(t))$. Since the system still works at time t if and only if $T > t$, we have

$$\Pr(T > t) = h(\bar{F}_1(t), \dots, \bar{F}_n(t)),$$

an expression obtained by Esary and Proschan [5]. Thus, the system survival function $\bar{F}_T(t) = \Pr(T > t)$ depends only on the survival functions of the individual components and the reliability function of the system.

The Signature Vector

Continue to imagine we start the system running at time $t = 0$ with all components working, and suppose we observe the system until every component's failure. Here, we further assume that the individual component lifetimes are independent and identically distributed (i.i.d.) random variables, thus $F_i(t) \equiv F(t)$ for each $i = 1, \dots, n$. Let $\bar{F}(t) = 1 - F(t)$ denote the common survival function.

Let $T_{(i)}$ denote the time of the i th component failure for $i = 1, \dots, n$. Thus, $0 < T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$ are the *order statistics* for the i.i.d. sample of component lifetimes. From the distribution theory for order statistics,

$$\Pr(T_{(k)} > t) = \sum_{j=0}^{k-1} \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j}.$$

We now give an alternative characterization of the survival function of a coherent system.

Definition 5. Let $\pi_k = \Pr(T = T_{(k)})$ denote the probability that the system fails at the k th failure of a component. The *signature* of a coherent system is the probability vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$.

If F is continuous, then $\boldsymbol{\pi}$ does not depend on F and the system survival function can be written as

$$\begin{aligned}
\bar{F}_T(t) &= \sum_{k=1}^n \Pr(T = T_{(k)}) \Pr(T > t | T = T_{(k)}) \\
&= \sum_{k=1}^n \pi_k \Pr(T_{(k)} > t) \\
&= \sum_{k=1}^n \sum_{j=0}^{k-1} \pi_k \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j}. \quad (7)
\end{aligned}$$

This representation, first given by Samaneigo [6], was recently extended to systems with exchangeable components by Navarro and Rychlik [7] and Navarro *et al.* [8].

Consider a coherent system of $n = 3$ components. We noted above that there are five such systems. Here, we will derive the signature of each.

1. If the system components are in series, so $\phi(\mathbf{x}) = x_1 x_2 x_3$, then the system fails at exactly the time of the first component failure. Thus $T = T_{(1)}$ with probability 1 and thus, $\pi = (1, 0, 0)$.
2. If the components are in parallel, so $\phi(\mathbf{x}) = x_1 \sqcup x_2 \sqcup x_3$, the system works until the last component failure. Here $T = T_{(3)}$ with probability 1 and thus, $\pi = (0, 0, 1)$.
3. The two-out-of-three system $\phi(\mathbf{x}) = (x_1 x_2) \sqcup (x_1 x_3) \sqcup (x_2 x_3)$ fails at precisely the time of the second component failure; that is, $T = T_{(2)}$ with probability 1 and thus, $\pi = (0, 1, 0)$.
4. Consider $\phi(\mathbf{x}) = x_1 \sqcup (x_2 x_3)$. This system fails at the earlier of the failure of components 1 and 2 and the failure of components 1 and 3; that is, the system lifetime T satisfies

$$T = \min \{ \max(T_1, T_2), \max(T_1, T_3) \}.$$

Thus, the system is certain to still be running at the first component failure, and will survive until the third component failure if the two to fail are components 2 and 3. The signature vector for this system is $\pi = (0, \frac{2}{3}, \frac{1}{3})$.

5. Consider $\phi(\mathbf{x}) = x_1(x_2 \sqcup x_3)$, a system that fails at the earlier of the failure of component 1 and the failure of both

components 2 and 3. Here we have

$$T = \min \{ T_1, \max(T_2, T_3) \}.$$

This system will survive the first component failure if that failure is component 1, but it cannot possibly survive the second. The signature vector is $\pi = (\frac{1}{3}, \frac{2}{3}, 0)$.

For the signature vectors of the 20 coherent systems of $n = 4$ components, see Ref. 9. A nice review of the applications of signatures is given by Kochar *et al.* [10]; for a book length treatment of the subject, the reader is referred to Ref. 11.

Stochastic Ordering of Coherent Systems

The signature vector leads to a method for comparison of two coherent systems. Let T_A and T_B denote the system lifetimes for systems A and B, respectively, and define their survival functions $\bar{F}_A(t) = \Pr(T_A > t)$ and $\bar{F}_B(t) = \Pr(T_B > t)$. We say that T_A is stochastically less than or equal to T_B , and write $T_A \leq^{st} T_B$, if $\bar{F}_A(t) \leq \bar{F}_B(t)$ for all $t > 0$. Thus if systems A and B are both started at time $t = 0$, then at any point in time, the probability that system B is still running is never less than the probability that system A is still running.

A deterministic ordering between two coherent systems guarantees a stochastic ordering of the system survival times, that is, if $\phi_A(\mathbf{x}) \leq \phi_B(\mathbf{x})$ for all $\mathbf{x} \in \{0, 1\}^n$, then $T_A \leq^{st} T_B$. The following theorem gives a sufficient condition for the stochastic ordering of system lifetimes in terms of their respective signatures.

Theorem 3. *Let A and B be two coherent systems of n components each, with common component-wise distribution function F and survival function $\bar{F}$. Denote the signature vectors of systems A and B by $\pi^A = (\pi_1^A, \dots, \pi_n^A)$ and $\pi^B = (\pi_1^B, \dots, \pi_n^B)$, respectively. If*

$$\sum_{k=j}^n \pi_k^A \leq \sum_{k=j}^n \pi_k^B \text{ for each } j = 1, \dots, n,$$

then $T_A \leq^{st} T_B$.

Proof. From Equation (7),

$$\begin{aligned}
\bar{F}_A(t) &= \sum_{k=1}^n \sum_{j=1}^{k-1} \pi_k^A \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \\
&= \sum_{j=0}^{n-1} \left(\sum_{k=j+1}^n \pi_k^A \right) \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \\
&\leq \sum_{j=0}^{n-1} \left(\sum_{k=j+1}^n \pi_k^B \right) \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \\
&= \sum_{k=1}^n \sum_{j=0}^{k-1} \pi_k^B \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \\
&= \bar{F}_B(t),
\end{aligned}$$

and thus $T_A \leq^{st} T_B$.

Denote the five coherent systems for $n = 3$ components by A, B, C, D, and E in the order considered in the section titled “The Signature Vector,” so that A corresponds to three components in series, B corresponds to three components in parallel, C is the two-out-of-three case, and so on. Using Theorem 3, it is easy to verify the stochastic ordering

$$T_B \leq^{st} T_E \leq^{st} T_C \leq^{st} T_D \leq^{st} T_A.$$

Of course, it is not always possible to establish such a relationship between two coherent systems, as the following example from Kochar *et al.* [10] illustrates. Suppose $n = 4$ and let

$$\phi_A(\mathbf{x}) = x_1 \left(x_2 \coprod x_3 \coprod x_4 \right)$$

and

$$\phi_B(\mathbf{x}) = (x_1 x_2) \coprod (x_3 x_4).$$

The signatures of these systems are given by $\pi^A = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0)$ and $\pi^B = (0, \frac{2}{3}, \frac{1}{3}, 0)$. Then

$$\pi_3^A + \pi_4^A > \pi_3^B + \pi_4^B$$

but

$$\pi_2^A + \pi_3^A + \pi_4^A < \pi_2^B + \pi_3^B + \pi_4^B$$

and thus the lifetime of system B is neither stochastically greater nor stochastically

less than that of system A. Indeed, of the $\binom{20}{2} = 190$ possible pairwise comparisons among the 20 coherent systems of $n = 4$ components, in 10 cases there is not a clear-cut winner.

For more on the stochastic ordering of coherent systems, the reader is again referred to Kochar *et al.* [10], who obtain conditions to compare systems in the hazard rate and likelihood ratio orderings as well. For comparisons between systems of exchangeable component lifetimes, or systems of different numbers of components, see Navarro *et al.* [8].

Acknowledgments

The authors would like to express their appreciation to the referees for providing thoughtful and insightful comments which helped to improve the original version of this article.

REFERENCES

1. Birnbaum ZW, Esary JD, Saunders SC. Multi-component systems and structures and their reliability. *Technometrics* 1961;3:55–77.
2. Natvig B. Reliability analysis. *Encyclopedia of Actuarial Science*. Teugels JL, Sundt J, editors. John Wiley and Sons; 2004.
3. Barlow RE, Proschan F. *Statistical Theory of Reliability and Life Testing*. New York: Holt, Rinehart, and Winston; 1975.
4. Bergman B. On reliability theory and its applications (with discussion). *Scandinavian J Stat* 1985;12:1–41.
5. Esary JD, Proschan F. Relationship between system failure rate and component failure rates. *Technometrics* 1963;5:183–189.
6. Samaniego FJ. On closure of the IFR class under formation of coherent systems. *IEEE Trans Reliab* 1985;34:69–72.
7. Navarro J, Rychlik T. Reliability and expectation bounds for coherent systems with exchangeable components. *J Multivariate Anal* 2007;98:102–113.
8. Navarro J, Samaniego FJ, Balakrishnan N, Bhattacharya D. On the application and extension of system signatures in engineering reliability. *Nav Res Logist* 2008;55:313–327.
9. Shaked M, Suarez-Llorens A. On the comparison of reliability experiments based on

- the convolution order. *J Am Stat Assoc* 2003;98:693–701.
10. Kochar S, Mukerjee H, Samaniego FJ. The “signature” of a coherent system and its application to comparisons among systems. *Nav Res Logist* 1999;46:507–523.
 11. Samaniego FJ. *System Signatures and their Applications in Engineering Reliability*, Number 110 in *International Series in Operations Research & Management Science*. Springer; New York: 2007.

COLLABORATIVE PROCUREMENT

FERYAL ERHUN

Department of Management Science and
Engineering, Stanford University,
Stanford, California

INTRODUCTION

In today's vertically disintegrated business environment, manufacturers often rely on external suppliers to provide (complex) components to produce their products. Original equipment manufacturers in the computer and communications industry outsource about 70% of their manufacturing to the electronics manufacturing services industry [1]. In the aircraft industry, Boeing has outsourced about 70% of the parts for its 787 Dreamliner while Airbus relies on subcontractors for about 50% of its A350 planes [2]. Overall, for every dollar of sales revenue, 50 cents are spent on component procurement in the US manufacturing industry [3]. It is therefore crucial that firms manage their procurement effectively to maintain a competitive edge in the market. One way of achieving this goal is to move toward collaboration in relationships that were once considered to be adversarial [4,5]:

Managing this extended network of relationships [due to outsourcing] requires more transparency, better communication, greater trust, and genuine reciprocity. In a nutshell, success in this environment will hinge heavily on shifting the customer–service provider relationship from adversarial to collaborative; from one based on procurement to one grounded in partnership. (Pat McArdle, Global Outsourcing Partner at PwC [6])

Given the impact of procurement on the bottom line, collaboration in procurement promises several potential advantages. Through collaborative procurement, planners can achieve cost reductions and increase

the efficiency of their supply chains considerably by leveraging their buying power and streamlining their procurement processes.

Cohen and Roussel [7] define collaboration as “the means by which companies within the supply chain work together toward mutual objectives through the sharing of ideas, information, knowledge, risks, and rewards.” There are two key aspects of this definition. First, collaboration requires common goals. Unless everyone involved benefits from the collaboration, one cannot talk about a true collaboration. Secondly, companies should work together to achieve these common goals, which involves sharing not only the benefits but also the risks.

Collaboration can occur among different functionalities of an enterprise (i.e., *intra-enterprise collaboration*) or among independent companies, shifting the nature of their relationship from adversarial to collaborative (i.e., *inter-enterprise collaboration*); it can also occur among multiple entities (different departments of an enterprise or different enterprises), which are responsible for the same function and may potentially have conflicting objectives (i.e., *horizontal collaboration*) or among different functional areas within an enterprise or a supply chain (i.e., *vertical collaboration*) [4]. Next, we discuss collaborative procurement in the context of this classification (Fig. 1).

Intra-Enterprise Horizontal Collaborative Procurement. Within an enterprise, centralizing procurement has many advantages; for example, it reduces duplication of effort, potentially decreases the number of suppliers in the supply base and enhances the relationships with suppliers, leverages the buying power, decreases costs through scale economies (such as quantity/volume discounts or logistics consolidation), and improves inventory control through risk pooling by aggregating forecasting and procurement of different business units within an enterprise.

Vertical	• Coordinates procurement decisions with other functional areas • Eliminates inefficiencies due to silo planning	• Enables information sharing between supply chain partners • Eliminates inefficiencies due to decentralization • Enables firms to focus on their core competencies
Horizontal	• Reduces duplication of effort • Decreases the size of supply base • Enhances relationships with suppliers • Leverages buying power • Enables scale economies • Enables risk pooling	• Leverages buying power • Enables scale economies • Standardizes purchasing process • Enables firms to focus on their core competencies
	Intra-enterprise	Inter-enterprise

Figure 1. Advantages of different facets of collaboration.

As a result, significant cost savings can be realized through centralization. For example, IBM “reinvented” its purchasing by centralizing its procurement and setting up commodity councils:

Commodity councils that leverage IBM purchases worldwide have resulted in IBM sourcing parts at pricing that’s 5%–10% below industry averages. By developing close relationships with suppliers, IBM has been successfully guaranteed supply of leading-edge parts to its line of PCs, servers, workstations, and mainframes. IBM’s Internet initiative with suppliers saved the company \$70 million [in 1998] and is expected to save \$240 million [in 1999] by making the procurement process more efficient. The strategies employed have had a big impact on IBM, which [in 1998] spent about \$41.5 billion with suppliers. IBM executives say procurement has saved Big Blue hundreds of million of dollars over the past several years and helped IBM return to profitability after some near-ruinous years earlier in the decade when it was bleeding red ink [8].

Intra-Enterprise Vertical Collaborative Procurement. The purchasing department’s goal in many companies is to procure goods at minimum cost. However, this focus on cost may do more harm than good:

Grackin [Managing Director of the Supply Chain Intelligence Service of Marsh Inc.] often sees a disconnect between procurement and logistics. The first is out to secure the cheapest parts regardless of location, while the second must deal with the realities of transportation, such as port congestion. Often the savings derived from buying the lowest-price part will be more than offset by higher transportation costs [9].

Silo planning is a common practice and procurement is especially prone to its effects. Intra-enterprise vertical collaboration breaks silo planning by enabling a “net landed cost” approach. It eliminates inefficiencies by coordinating procurement decisions with other functional areas, such as manufacturing, planning, inventory control, and logistics, and by introducing common goals for these different functionalities [10]:

In late 2003, Aetna’s senior management sent down the mandate—reduce selling, general and administrative (SG&A) expenses through a more integrated spend-management program. . . . One area where the spend analysis was most effective was in legal services, “a silo that traditionally procurement was not invited

into.” . . . While the measures [a more in depth and formalized RFP (Request for Proposal) process, caps on reimbursable expenses, etc.] were initiated in the third quarter of [2004], Aetna began to see the positive results by the end of 2004. Legal services realized an overall savings of 10% by the end of 2004, with some silos within legal services achieving 30–40% reductions [11].

Inter-Enterprise Horizontal Collaborative Procurement. In inter-enterprise horizontal collaborative procurement (also known as *cooperative*, *group purchasing*, or *consortium purchasing*) two or more independent organizations (potential competitors) with similar products or services join together (either formally or informally, or through an independent third party) for the purpose of combining their purchase of materials, services, and capital goods. This is not a new concept; it is commonly used in many industries, such as health-care, chemicals, telecommunications, transportation, and services. For example, in the health-care industry, group purchasing organizations (GPOs) manage about 70–80% of the purchases that hospitals make, which is around \$195 billion annually. By taking advantage of large volumes, GPOs save hospitals “between 10 to 15% off their purchasing costs” [12]. When combined with other benefits, such as cost reductions due to standardization of the purchasing process and reductions in personnel, group purchasing increases the procurement efficiency considerably:

Last year, 25 Council of Industry members signed fixed-priced supply contracts with Pepco Energy Services to collectively purchase an estimated 80,000 megawatt hours of electricity annually. Over the first eight months of 2008, the Council of Industry energy purchasing consortium has saved over \$1 million, compared to the applicable utility rates. . . . Based on current market conditions, Pepco Energy Services estimates that the

Council of Industry energy purchasing consortium will save an additional \$504,000 for a total savings of over \$1.5 million for the 2008 calendar year [13].

Inter-Enterprise Vertical Collaborative Procurement. Collaboration between buyers and suppliers, which is based on long-term purchasing relationships, has been successfully adopted by many companies in various industries. Such a collaborative relationship requires full commitment from all parties involved as well as joint efforts to develop processes to support the new relationship:

There are no quick fixes in this game, and the shared trust, honesty, integrity, and objective focus on results must be maintained over a long time frame, unimpeded by sub-optimal squeezing of carefully chosen key suppliers. The long time frame is, in essence, used for a continuing set of business process re-engineering activities in both the customer and supplier companies, where cost and value are continually enhanced. For example, the vice-president of procurement for Honda of America at the time decided to develop one of its super-supplier relationships in sheet metal parts. After analysing existing suppliers in this area, he approached a small sheet metal company run by two brothers who had the “correct attitude.” His criteria were to find a supplier that would be excellent at collaboration, trustworthy and share a mutual motivation for results. He approached the owners and took them into his firm’s internal operations, shared best practices and taught them how to develop process engineering techniques, benchmarking and rapid knowledge implementation. They continued to collaborate, share information and provide feedback to one another, which resulted in this supplier becoming one of the most competitive in the industry. After three years, the supplier had cut the costs of its materials by half, which was significant because it had previously been competitive anyway [14].

Though difficult to attain and maintain, inter-enterprise vertical collaborative procurement provides tangible benefits. For example, in an environment where 70–80% of the manufacturing costs are due to suppliers, collaborative relationships reduced Honda’s costs by 25% in the 1990s; since “many of the cost-cutting ideas that made [Accord] so successful came from suppliers” [15].

Vertical and Horizontal Collaborative Procurement. More and more enterprises combine vertical *and* horizontal approaches to procurement to improve the overall process. Electronic marketplaces, which bring buyers and suppliers together in order to create an on-line transaction platform, are commonly used for this goal and have emerged in many industries. In the aerospace industry, Exostar, which was founded in 2000 by some of the world’s largest aerospace and defense companies—BAE Systems, Boeing, Lockheed Martin, Raytheon, and Rolls-Royce—creates an on-line collaboration platform for aerospace and defense industry partners. Exostar has over 40,000 registered companies worldwide, with an annual throughput of \$38+ billion, and electronic Request for Quotations (eRFQs) of around 72,000 [16]. Covisint is another on-line platform, which was founded in 1999 by a consortium of the world’s largest automakers—General Motors, Ford Motor Company, DaimlerChrysler, Nissan, and Renault—to manage increasing costs and inefficiencies in the automotive industry through on-line collaboration. Today, the platform supports over 45,000 organizations worldwide in the global automotive, health-care, public sector, and financial services industries [17].

In the rest of this article, we will focus on horizontal collaborative procurement. We refer interested readers to different articles in this encyclopedia for the analysis of vertical collaboration. Vertical collaboration relies on trust among parties involved;

reviews of mechanisms to build trust and incentives among supply chain partners, as well as mechanisms that enable supply chain coordination, can be found in the article titled **Supply Chain Coordination** and in the section titled “Supply Chain Collaboration” in this encyclopedia. Many supply chain partners now practice collaborative methods, which are closely related to collaborative procurement, such as Collaborative Planning, Forecasting, and Replenishment (CPFR) and Vendor-Managed Inventory (VMI). These practices eliminate supply chain inefficiencies, share risks, and enable collaboration. We refer interested readers to the articles titled **Vendor-Managed Inventory** in this encyclopedia for reviews of VMI and CPFR, respectively.

QUANTITY DISCOUNTS AND HORIZONTAL COLLABORATION

One of the major benefits of horizontal collaborative procurement is reduced purchasing costs since the firm(s) can leverage its (their joint) buying power through intra-(inter-)enterprise collaboration. This is possible as suppliers often offer quantity discounts to their buyers. We provide a simple example below that highlights the role of quantity discounts in collaborative procurement.

Two manufacturers, **M1** and **M2**, are selling the same product in two different markets. The market price for the product is set as a function of the quantity each manufacturer produces as follows: $P_i = 10.5 - q_i$, where $i = 1, 2$; q_1 and q_2 are the respective production quantities of **M1** and **M2**; and P_i is the price in market i . The manufacturers procure a subcomponent from the same supplier (one subcomponent is required for every unit produced and any unused subcomponents cannot be disposed of) who offers the following all-unit quantity discount scheme with a minimum order commitment:

$$\begin{aligned} 1 < \text{order size} \leq 3 \text{ units} &\Rightarrow \text{wholesale price} \\ &= \$3/\text{unit} \end{aligned}$$

$3 < \text{order size} \leq 5 \text{ units} \Rightarrow \text{wholesale price}$
 $= \$2.25/\text{unit}$
 $5 < \text{order size} \Rightarrow \text{wholesale price}$
 $= \$2/\text{unit}$

The supplier incurs a fixed cost of $F = \$3.50$ per order independent of the order size. The manufacturers incur a production cost of c_i per unit, where $c_1 = \$1.25$ and $c_2 = \$0.50$. Owing to restrictions on batch sizes and capacities, the manufacturers can produce 2, 4, or 6 units. Assuming no collaboration between these manufacturers, how many units should each manufacturer produce to maximize her own profit?

We can formulate the problem that each manufacturer faces as follows:

$$\max_{q_i} \Pi_i(q_i) = \max_{q_i} \{(P_i - C(q_i) - c_i) \times q_i\}, \quad (1)$$

where $C(q_i)$ is the wholesale price for an order of size q_i and $\Pi_i(q_i)$ is the profit of manufacturer i given her production quantity q_i . Specifically, we can write **M1**'s and **M2**'s respective profits for different production quantities as follows:

$$\begin{aligned}
 q_i = 2: \Pi_1 &= (10.5 - 2 - 3 - 1.25) \times 2 = 8.5, \\
 \Pi_2 &= (10.5 - 2 - 3 - 0.5) \times 2 = 10, \\
 q_i = 4: \Pi_1 &= (10.5 - 4 - 2.25 - 1.25) \times 4 = 12, \\
 \Pi_2 &= (10.5 - 4 - 2.25 - 0.5) \times 4 = 15, \\
 q_i = 6: \Pi_1 &= (10.5 - 6 - 2 - 1.25) \times 6 = 7.5, \\
 \Pi_2 &= (10.5 - 6 - 2 - 0.5) \times 6 = 12.
 \end{aligned}$$

Each manufacturer will choose the quantity that maximizes her profit. Thus, both **M1** and **M2** produce 4 units, and earn profits of \$12 and \$15, respectively. As **M2** has a significant cost advantage over **M1**, she enjoys a higher profit than **M1** although both manufacturers produce the same amount. The supplier's profit is $C(q_1) \times q_1 + C(q_2) \times q_2 - 2 \times F = (2.25 \times 4 - 3.50) \times 2 = \11 .

Now, let us consider the case where the manufacturers engage in collaborative procurement. In this case, the two manufacturers order jointly, hence the unit cost is

determined by their total quantity (i.e., if both manufacturers produce 2 units, the unit cost is \$2.25 per unit; otherwise, it is \$2 per unit). We observe that both manufacturers continue to produce 4 units, but they now pay only \$2 per unit instead of \$2.25 per unit, thus they increase their profits to \$13 and \$16, respectively. Through collaborative procurement the manufacturers present a unified front to the supplier and subsequently increase their buying power. When the manufacturers collaborate, the supplier's profit is $C(q_1 + q_2) \times (q_1 + q_2) - F = \12.50 . Since the supplier has to process one large order rather than two smaller orders, he benefits from economies of scale. He may pass some of these benefits to buyers by lowering his prices [18].

The fact that the manufacturers benefit from collaboration when they serve two different markets is quite intuitive. However, even when the manufacturers *compete in the same market*, they may choose to collaborate and help decrease each other's costs. This is true even for a manufacturer who has a significant cost advantage over her competitor. To see why, suppose that the manufacturers choose their production quantities simultaneously and the market price is then set as a function of the total quantity as follows:

$$P = (10.5 - (q_1 + q_2))^+ \quad (2)$$

$$= \max\{0, 10.5 - (q_1 + q_2)\}, \quad (3)$$

where q_1 and q_2 are the respective production quantities of **M1** and **M2** and P is the market price. This game is Cournot competition (Cournot competition models a setting where competing companies choose their output quantities independently and simultaneously; with a slight twist: instead of charging a constant per unit price for the subcomponent, the supplier offers a discount to each manufacturer).

Using the techniques covered in the article titled *Nash Equilibrium (Pure and Mixed)* in this encyclopedia, we can find the unique Nash equilibrium of this game under the cases of competition and collaboration. The equilibrium quantities for **M1** and **M2** are 2 units and 4 units, respectively, under both cases. When the manufacturers

do not collaborate, their profits are \$0.50 and \$7, respectively. The supplier's profit is \$8. When the manufacturers collaborate, their profits are \$2.50 and \$8, respectively. That is, when they collaborate, both manufacturers are better off; namely, collaboration Pareto dominates competition. Note that the manufacturers' profits increase disproportionately: **M1**, who has a cost disadvantage, benefits from collaboration more than her competitor **M2** does. The supplier benefits from economies of scale as before, and increases his profits to \$8.50 when the manufacturers collaborate.

This simple example leads to several questions; for example, under what demand conditions would manufacturers prefer collaboration? When would a supplier be better off providing a discount scheme that enables collaboration among manufacturers? What about the impact of collaboration on social welfare? Such questions are the premise of several academic papers that we review next.

LITERATURE REVIEW ON HORIZONTAL COLLABORATION

Collaboration and Social Welfare

Mathewson and Winter [19] study a setting where a group of buyers collaborates to lower their costs and contracts with a group of suppliers. Such coalitions may be considered legally troublesome; they impose negative externality to buyers outside the coalition as these buyers face higher prices and lower availability. However, the authors conclude that coalitions leading to group purchasing may actually increase social welfare.

Value of Collaboration for Buyers

Keskinocak and Savaşaneril [20] study group purchasing among competing buyers with a single-supplier, multiple-buyers model. The supplier offers a quantity discount scheme to the buyers; that is, multiple-buyers can procure together to obtain lower prices by increasing their buying power. The authors show that when the buyers do not have capacity constraints (and are identical in terms of

costs), they always earn more profits by collaborating. Interestingly, the supplier also prefers to sell to collaborating buyers. However, when the buyers have limited capacity, collaboration may not always be preferable: a buyer is willing to collaborate if collaboration increases the quantity that he purchases.

Griffin *et al.* [21] study alternative buyer strategies in markets where procurement costs are affected by economies of scale in the suppliers' production costs and by economies of scope in transportation. They consider buyer strategies with different types of collaboration, namely, (i) no collaboration among buyers or buyer divisions, (ii) intra-enterprise collaboration among the purchasing organizations of the same buyer enabled by an internal intermediary, and (iii) inter-enterprise (full) collaboration among multiple-buyers enabled by a third party intermediary. They find that when the potential benefits from economies of scope are high, intra-enterprise collaboration performs very well. When the potential benefits from economies of scale are high, the authors observe that buyer strategies need to consider potential future trades in the market by other buyers while contracting with a supplier. Their computational analysis indicates that the potential benefits of collaboration are highest in capacitated markets with high fixed production and/or transportation costs.

Value of Collaboration for Suppliers

Like Keskinocak and Savaşaneril [20], Anand and Aron [22] study collaboration between multiple-buyers who purchase from a single-supplier. Different from Keskinocak and Savaşaneril, their model analyzes group purchasing under demand uncertainty. In Anand and Aron's model, the supplier posts a quantity discount scheme. Then, buyers arrive and demand single units. As the total number of units demanded increases, the price buyers pay drops. Eventually all buyers pay the same price for the product. The supplier does not know the demand function prior to setting the discount scheme. Anand and Aron's goal is to compare the performance of group buying to posted prices. They show that the supplier prefers

group buying to posted prices (i) as demand heterogeneity increases and (ii) if he can postpone the production decision until after demand is realized *and* if he enjoys scale economies in production.

Product Substitutability and Value of Collaboration

Using concepts from cooperative game theory (see the section titled “Cooperative Games” in this encyclopedia), Granot and Sošić[23] study collaboration in electronic marketplaces. The authors characterize the conditions under which a firm would prefer to collaborate with a model of three retailers whose products may have a certain degree of substitutability. They show that all three retailers collaborate either when all products are nonsubstitutable or when all three products are highly substitutable and the retailers benefit from collaboration equally.

CONCLUSIONS AND KEY MANAGERIAL INSIGHTS

The advantages of successful collaborative procurement are numerous, including lower costs, reduced inventory, increased sales, increased revenue, improved customer service, and more efficient use of resources. Despite its many advantages, collaboration does not come easily to companies. Establishing a successful collaboration requires *commitment* from all involved parties. Collaboration should start internally before extending to external partners. Companies should know *why* as well as *with whom* they want to collaborate. *Trust* is key for a true collaboration. Partners should be ready to *share* practices, processes, and information, even proprietary information, when necessary. Roles, responsibilities, expectations, and goals should be clearly documented. As is the case with any process improvement, setting short-term and long-term *performance measures* to evaluate the success of collaboration is essential. Furthermore, companies should be ready to redefine the scope of their relationships based on these performance measures and as their partnership evolves.

It is important that the academic literature continues to guide companies in their quests for collaboration. Understanding the role of collaboration in environments with uncertainty is important. Decision support tools and models for net landed cost would be especially beneficial. Collaboration requires truthfulness and trust. Conditions under which such truthfulness can be achieved, especially when companies have private information, should be characterized. Long-term and relational models to study collaborative procurement would be valuable.

REFERENCES

1. Carbone J. EMS industry goes car shopping. *Purchasing.com* October 18, 2007.
2. Hise P. The remarkable story of Boeing's 787. *CNNMoney.com*, July 9, 2007.
3. U.S. Department of Commerce. Statistics for industry groups and industries: annual survey of manufacturers. Available at <http://www.census.gov/prod/2006pubs/am0531gs1.pdf>. May 2009. 2006.
4. Erhun F, Keskinocak P. Collaborative supply chain management. In: Kempf K, Keskinocak P, Uzsoy R, editors, *Handbook of production planning*. Kluwer International Series in Operations Research and Management Science. Kluwer Academic Publishers; In press.
5. Wilkinson S, Shestakova Y. Collaborative procurement on the rise. *Build*. December 2006. pp. 70–71.
6. PricewaterhouseCoopers Global Sourcing. Partnership rather than procurement is new path to effective outsourcing, according to PricewaterhouseCoopers' new 2007 global outsourcing survey. 2007. Available at <http://www.pwc.com/>. May 21.
7. Cohen S, Roussel J. Strategic supply chain management: the five disciplines for top performance. Chapter core discipline 4: Build the right collaborative model. New York, NY: McGraw-Hill; 2005. pp. 139–167.
8. Carbone J. Reinventing purchasing wins the Medal for BIG BLUE. *Purchasing.com*, September 16, 1999.
9. Bowman RJ. Free the enterprise! Bust the silos in the supply chain! *SupplyChainBrain.com*, September 10, 2008.

10. Erhun F, Tayur S. Enterprise-wide optimization of total landed cost at a grocery retailer. *Oper Res* 2003;51(3):343–353.
11. Forrest W. Aetna's silo-busting strategy corals SG&A spend. *Purchasing.com*, November 3, 2005.
12. RedOrbit News. Group purchasing organizations enable hospitals to save USD33 billion each year through lower product prices. 2006. Available at <http://www.redorbit.com/>. Jul 31.
13. RedOrbit News. Pepco energy services saves \$1 million for council of industry energy purchasing consortium in New York. 2008. Available at <http://www.redorbit.com/>, Sept. 23.
14. Billington C, Cordon C, Vollmann T. Developing the super supplier. *CPO Agenda*, Spring 2006.
15. Knowledge@W.P. Carey. Deep supplier relationships drive automakers' success. 2005. Available at <http://knowledge.wpcarey.asu.edu/article.cfm?articleid=1061#>. Jul 06.
16. Exostar. www.exostar.com. May 2009.
17. Covisint. www.covisint.com. May 2009.
18. Melymuka K. Efficient? Become superefficient. *Computerworld*, September 10 2001.
19. Mathewson F, Winter RA. Buyer groups. *Int J Ind Organ* 1996;15(1):137–164.
20. Keskinocak P, Savaşaneril S. Collaborative procurement among competing buyers. *Nav Res Logist* 2003;55(6):516–540.
21. Griffin P, Keskinocak P, Savaşaneril S. The role of market intermediaries for buyer collaboration in supply chains. In: Akcali E, Geunes J, Pardalos P, *et al.*, editors. *Applications of supply chain management and e-commerce research in industry*. Chap. 3, New York, NY: Springer; 2005. pp. 87–118.
22. Anand KS, Aron R. Group buying on the web: a comparison of price discovery mechanisms. *Manag Sci* 2003;49(11):1547–1564.
23. Granot D, Sošić G. Formation of alliances in Internet-based supply exchanges. *Manag Sci* 2005;51(1):92–105.

COLUMN GENERATION

MARCO E. LÜBBECKE
Chair of Operations Research,
RWTH Aachen University,
Aachen, Germany

Column generation is a classical technique to solve a mathematical program by iteratively adding the variables of the model [1]. Typically, only a tiny fraction of the variables is needed to prove optimality, which makes the technique interesting for problems with a huge number of variables. The method is often said in one sentence with Dantzig–Wolfe decomposition [2] (see also **Dantzig–Wolfe Decomposition**), as it is particularly effective when the matrix has a special structure like bordered block-diagonal or staircase forms.

We wish to clarify one point right at the start. Even though the method was termed *generalized linear programming* in the early days, it never became competitive for solving linear programs, except for special cases [3]. In addition, tailored implementations were designed to exploit matrix structures, but they did not perform better than the simplex method [4]. In contrast, column generation is a real winner in the context of integer programming (see **Branch-Price-and-Cut Algorithms**). This made the powerful method a *must-have* in the computational mixed integer programming “bag of tricks.”

We assume that the reader is familiar with basic linear programming duality and the simplex method (see also the section titled “Fundamental Techniques” in this encyclopedia).

COLUMN GENERATION

We would like to solve a linear program, called the *master problem* (MP),

$$v(\text{MP}) := \min \sum_{j \in J} c_j \lambda_j$$

$$\text{subject to } \sum_{j \in J} \mathbf{a}_j \lambda_j \geq \mathbf{b} \quad (1)$$

$$\lambda_j \geq 0, \quad j \in J,$$

with $|J| = n$ variables and m constraints. In many applications, n is exponential in m and working with (1) explicitly is not an option because of its sheer size. Instead, consider the *restricted master problem* (RMP), which contains only a subset $J' \subseteq J$ of variables. An optimal solution λ^* to the RMP need not be optimal for the MP, of course. Denote an optimal dual solution to the RMP by π^* . In the *pricing step* of the simplex method (see also **The Simplex Method and Its Complexity**), we look for a nonbasic variable of negative reduced cost to enter the basis. To accomplish this in column generation, one solves the *pricing problem* (or *subproblem*) PP:

$$v(\text{PP}) := \min \{c_j - \pi^* \mathbf{a}_j \mid j \in J\}. \quad (2)$$

When $v(\text{PP}) < 0$, the variable λ_j and its coefficient column $(c_j, \mathbf{a}_j)$ corresponding to a minimizer j are added to the RMP; this is solved to optimality to obtain optimal dual variable values, and the process iterates until no further improving variable is found. In this case, λ^* optimally solves the MP (1) as well. In particular, column generation inherits finiteness and correctness from the simplex method, when cycling is taken care of.

It does not seem clear why Equation (2) should be of any help when $|J|$ is large. However, in almost every application, indices in J enumerate entities, which can well be described as the feasible domain X of an optimization problem

$$\min_{\mathbf{x} \in X} \{c(\mathbf{x}) - \pi^* a(\mathbf{x})\}, \quad (3)$$

where $c_j = c(\mathbf{x}_j)$ and $\mathbf{a}_j = a(\mathbf{x}_j)$, and $\mathbf{x}_j \in X$ corresponds one-to-one to $j \in J$. That is, instead of *explicitly* pricing all candidate variables, we solve a typically well-structured optimization problem, making the search for a variable of negative reduced cost *implicit*.

Technically, our notation suggests that X be finite, but this need not be the case.

Consider the one-dimensional cutting-stock problem, the classical example in column generation introduced in Gilmore and Gomory [5]. We are given paper rolls of width W , and m demands $b_i, i = 1, \dots, m$, for orders of width w_i . The goal is to minimize the number of rolls to be cut into orders, such that the demand is satisfied. A standard formulation is

$$\min \left\{ \mathbf{1}\lambda \mid A\lambda \geq \mathbf{b}, \lambda \in \mathbb{Z}_+^{|J|} \right\}, \quad (4)$$

where A encodes the set of $|J|$ feasible cutting patterns, that is, $a_{ij} \in \mathbb{Z}_+$ denotes how often order i is obtained when cutting a roll according to $j \in J$. From the definition of feasible patterns, the condition $\sum_{i=1}^m a_{ij}w_i \leq W$ must hold for every $j \in J$, and λ_j determines how often the cutting pattern $j \in J$ is used. The linear relaxation of Equation (4) is then solved via column generation, where the pricing problem is a *knapsack problem*.

Dual Bounds. During column generation we have access to a dual bound on $v(\text{MP})$ so that we can terminate the algorithm when a desired solution quality is reached. Let $v(\text{RMP})$ denote the optimum of the current RMP. When we know that $\sum_{j \in J} \lambda_j \leq \kappa$ for an optimal solution of the MP, we cannot improve $v(\text{RMP})$ by more than κ times the smallest reduced cost $v(\text{PP})$. Hence

$$v(\text{RMP}) + \kappa \cdot v(\text{PP}) \leq v(\text{MP}). \quad (5)$$

An important special case is $\kappa = 1$ when a *convexity constraint* is present; see the section titled “Dantzig–Wolfe Decomposition.” This bound is tight as $v(\text{PP}) = 0$ when column generation terminates. Note that $v(\text{PP})$ is not available when the pricing problem is solved heuristically. When the objective function is a sum of all variables, that is, $\mathbf{c} \equiv \mathbf{1}$, we use $\kappa = v(\text{MP})$ and obtain $v(\text{RMP})/(1 - v(\text{PP})) \leq v(\text{MP})$. There are other proposals, for example, such as that of Farley [6], and also tailored bounds for special problems, for example, such as that of Valério de Carvalho [7]. In general, the dual bound is not monotone over the iterations (*yo-yo effect*).

Dantzig–Wolfe Decomposition

The classical scenario of column generation is set in the context of Dantzig–Wolfe decomposition [2] in which a special structure of the typically very sparse coefficient matrix is exploited (see also **Dantzig–Wolfe Decomposition**). Consider a linear program, called the *original formulation* in this context:

$$\begin{aligned} \min \quad & \mathbf{c}\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \geq \mathbf{b} \\ & D\mathbf{x} \geq \mathbf{d} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned} \quad (6)$$

Let $X = \{\mathbf{x} \in \mathbb{Q}_+^n \mid D\mathbf{x} \geq \mathbf{d}\}$. Using the representation theorems for convex polyhedra by Minkowski and Weyl [8], we can write each $\mathbf{x} \in X$ as a finite convex combination of extreme points $\{\mathbf{x}_p\}_{p \in P}$ plus finite nonnegative combination of extreme rays $\{\mathbf{x}_r\}_{r \in R}$ of X , that is,

$$\begin{aligned} \mathbf{x} &= \sum_{p \in P} \mathbf{x}_p \lambda_p + \sum_{r \in R} \mathbf{x}_r \lambda_r, \quad \sum_{p \in P} \lambda_p = 1, \\ \lambda &\in \mathbb{Q}_+^{|P|+|R|}. \end{aligned} \quad (7)$$

Substituting for $\mathbf{x}$ in Equation (6), thereby eliminating constraints $D\mathbf{x} \geq \mathbf{d}$, and letting $c_j = \mathbf{c}\mathbf{x}_j$ and $\mathbf{a}_j = A\mathbf{x}_j, j \in P \cup R$, we obtain an equivalent *extended formulation*:

$$\begin{aligned} \min \quad & \sum_{p \in P} c_p \lambda_p + \sum_{r \in R} c_r \lambda_r \\ \text{subject to} \quad & \sum_{p \in P} \mathbf{a}_p \lambda_p + \sum_{r \in R} \mathbf{a}_r \lambda_r \geq \mathbf{b} \\ & \sum_{p \in P} \lambda_p = 1 \\ & \lambda \geq \mathbf{0}, \end{aligned} \quad (8)$$

which is solved by column generation. Let π^*, π_0^* denote a dual-optimal solution to the RMP obtained from Equation (8), where variable π_0 corresponds to the *convexity constraint* $\sum_{p \in P} \lambda_p = 1$. The subproblem (3) is to check whether $\min_{j \in P \cup R} \{c_j - \pi^* \mathbf{a}_j - \pi_0^*\} < 0$. By our previous linear transformation, this results in solving the linear program

$$\min \{(\mathbf{c} - \pi^* A)\mathbf{x} - \pi_0^* \mid D\mathbf{x} \geq \mathbf{d}, \mathbf{x} \geq \mathbf{0}\}. \quad (9)$$

When the minimum is negative and finite, an optimal solution to Equation (9) is an extreme point $\mathbf{x}_p$ of X , and we add a variable with coefficient column $[\mathbf{c}\mathbf{x}_p, (A\mathbf{x}_p), 1]$ to the RMP. When the minimum is minus infinity, we obtain an extreme ray $\mathbf{x}_r$ of X as a homogeneous solution to Equation (9), and we add the column $[\mathbf{c}\mathbf{x}_r, (A\mathbf{x}_r), 0]$ to the RMP. It is particularly interesting that the MP stays a linear program even when the subproblem is nonlinear.

The usefulness of Dantzig–Wolfe decomposition becomes more apparent in the practically relevant case where D has a block-diagonal structure, that is,

$$D = \begin{pmatrix} D^1 & & \\ & D^2 & \\ & & \ddots \\ & & & D^K \end{pmatrix} \quad \mathbf{d} = \begin{pmatrix} \mathbf{d}^1 \\ \mathbf{d}^2 \\ \vdots \\ \mathbf{d}^K \end{pmatrix}. \quad (10)$$

Each $X^k = \{D^k \mathbf{x}^k \geq \mathbf{d}^k, \mathbf{x}^k \geq \mathbf{0}\}$, $k = 1, \dots, K$, gives rise to a representation as in Equation (7). The decomposition yields K subproblems, each with its own convexity constraint and associated dual variable π_0^k :

$$\min \left\{ (\mathbf{c}^k - \pi A^k) \mathbf{x}^k - \pi_0^k \mid \mathbf{x}^k \in X^k \right\}, \quad k = 1, \dots, K, \quad (11)$$

where $\mathbf{c}^k$ and A^k correspond to variables $\mathbf{x}^k$. An optimal solution to the RMP is found when no minimum in Equation (11) is negative. The dual bound (5) can be adapted.

There are other special matrix structures that can be exploited, for example, the so-called staircase form of matrices, which arises in multiperiod or multistage planning problems, in particular in stochastic programming. In the easiest case, the matrix of the pricing problem has bordered block-diagonal structure again, and the Dantzig–Wolfe decomposition can be iteratively applied (also known as *nested column generation*).

Lagrangian Relaxation

For a bordered block-diagonal matrix, in particular, Dantzig–Wolfe decomposition

can be interpreted as keeping complicating constraints in the MP while exploiting a particular structure in the subproblems. *Lagrangian relaxation* [9] proceeds the other way round: the complicating constraints $A\mathbf{x} \geq \mathbf{b}$ are relaxed and their violation is penalized in the objective function via multipliers $\pi \geq \mathbf{0}$ (see also **Lagrangian Optimization for LP**). This results in the *Lagrangian subproblem*

$$L(\pi) := \min_{\mathbf{x} \in X} \mathbf{c}\mathbf{x} - \pi(A\mathbf{x} - \mathbf{b}), \quad (12)$$

which gives a lower bound on the optimum in Equation (6) for any $\pi \geq \mathbf{0}$. We obtain the best such bound by solving the *Lagrangian dual problem*

$$\max_{\pi \geq \mathbf{0}} L(\pi). \quad (13)$$

The Lagrangian function $L(\pi)$ is piecewise linear, concave, and subdifferentiable (but not differentiable). Since they are very easy to implement, the most popular choice to obtain optimal or near-optimal multipliers are the subgradient algorithms (see also **Subgradient Optimization**). By duality, in the optimum $v(\text{RMP}) = \pi^* \mathbf{b}$, and Equation (12) can be written as

$$\begin{aligned} L(\pi) &= \pi^* \mathbf{b} + \min_{\mathbf{x} \in X} (\mathbf{c} - \pi A)\mathbf{x} \\ &= v(\text{RMP}) + v(\text{PP}), \end{aligned}$$

that is, the dual bound in Dantzig–Wolfe decomposition and the Lagrangian bound coincide (see also **Relationship Among Benders, Dantzig–Wolfe, and Lagrangian Optimization**).

Row and Column Generation

Linear programs may have not only a large number of variables but also (too) many rows, for example, when constraints are formulated on all subsets of a given ground set (like sub-tour elimination constraints for the TSP). In such cases, one iteratively adds only those constraints that are violated by the current solution. The identification of a violated constraint (or the detection that none exists) is called *separation*. Embedded in a branch-and-bound algorithm, *cutting-plane methods*

became instrumental (and thus the standard) in solving mixed *integer* programs. Now, row and column generation obviously cannot be viewed independently. Even though some general ideas exist on how the pricing problem needs to be modified in order to cope with the dual variables from the additional rows, such approaches are still mainly problem specific (see also **Branch-Price-and-Cut Algorithms**).

Mixed Integer Programs

When solving a mixed integer program by branch-and-bound, the (linear) relaxation serves the purpose of providing a dual bound on the optimal objective function value. When the relaxation is solved by column generation in each node, one is referring to branch-and-price (see also **Branch-Price-and-Cut Algorithms**). We cannot overstate the fact that the primary use of column generation is in this context, and it is becoming increasingly popular as column generation reformulations often give much stronger bounds than the original LP relaxation. Many people actually refer to solving an integer program when they speak of column generation.

The Dantzig–Wolfe decomposition principle can be generalized to mixed integer programs in several ways. However, the basic column generation procedure to solve the linear relaxation remains the same. One drawback, the slow convergence (see the section titled “Stabilization of Dual Variables”), may even become less severe. When a dual bound LB is available, and the objective function coefficients are all integers, that is, $c_j \in \mathbb{Z}$, $j \in J$, column generation can be stopped as soon as $\lceil LB \rceil = \lceil z \rceil$. When one is aiming for quick integer solutions one may even terminate the process prematurely, and take a branching decision as soon as column generation starts tailing off. In this case the node’s dual bound is not valid, so it is set to that of the father node, and this *early termination* even is exact in principle.

ALGORITHMIC ISSUES

The dual of the RMP is the dual of the MP with rows omitted, and hence a relaxation.

Therefore, the pricing problem is a *separation problem* for the dual; column generation is a cutting plane method to solve the Lagrangian dual (13). This explains why many researchers relate it to the Kelley [10] and Cheney-Goldstein [11] cutting plane methods that are known for maximizing a concave continuous function. The dual point of view (see Briant *et al.* [12] for a more detailed discussion) revealed central algorithmic issues in column generation. In particular, one should re-read this section after having read the section titled “Stabilization of Dual Variables” on dual variable stabilization.

Note that there is a theoretical consequence from the equivalence of separation and optimization [13]. Even exponential size RMPs (linear programs) are solvable in polynomial time (in theory by the Ellipsoid method) when the pricing problem is solvable.

Master Problem: Computing Primal and Dual Solutions

The purpose of the RMP is to provide dual variable values: To communicate to the pricing problem which primal variables are needed to come closer to dual feasibility, and thus primal optimality. Note that we never need a primal feasible solution before optimality is reached, not even to calculate dual bounds. That is, the RMP serves the same purpose as, for example, subgradient methods in Lagrangian relaxation, and this connection can be exploited.

Initialization, Infeasibility, and Farkas Pricing. Even when the MP has a feasible solution, there are two important situations when the RMP is not feasible: in the beginning when no variables have been generated yet, and after branching when solving an integer program. In the traditional “phase I” approach [14] artificial variables with a “big M ” penalty cost are introduced. A smaller M gives a tighter upper bound on the respective dual variables, and may reduce the *heading-in effect* [15] of initially producing irrelevant columns. Heuristic estimates of the optimal dual variable values can be used for this purpose [16]. Furthermore, one may warm-start

from a previous similar run [17] or use a primal heuristic to produce an initial solution.

Column generation provides another way for turning an infeasible RMP feasible via the well-known fact that the dual of an infeasible linear program is unbounded (if not infeasible). This is formalized in Farkas' Lemma, which states that either $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ is feasible or there is a vector $\boldsymbol{\pi}$ with $\boldsymbol{\pi}A \leq \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{b} > 0$. Such a vector $\boldsymbol{\pi}$, which is interpreted as a ray in the dual, *proves* the infeasibility of the first system as the condition $\boldsymbol{\pi}A\mathbf{x} = \boldsymbol{\pi}\mathbf{b}$ cannot be fulfilled. The idea is now to add a variable to A with coefficient column $\mathbf{a}$ with $\boldsymbol{\pi}\mathbf{a} > 0$, which thus *destroys* this proof of infeasibility. Such a variable can be found (or concluded that none exists) by solving

$$\max_{\mathbf{x} \in X} \{\boldsymbol{\pi}^* \mathbf{a}(\mathbf{x})\} = \min_{\mathbf{x} \in X} \{-\boldsymbol{\pi}^* \mathbf{a}(\mathbf{x})\}, \quad (14)$$

which is nothing else but the standard pricing problem (3) with cost coefficients $c(\mathbf{x}) = 0$. The dual ray $\boldsymbol{\pi}^*$ is typically provided by the LP solver in the case of an infeasible linear program. While this method appears to belong to the *folklore*, the name *Farkas pricing* has been introduced only recently in Achterberg [18] within the SCIP framework (see section titled "Acceleration Techniques and Implementation Issues").

Algorithms: Pivots, Subgradients, Bundles, and Volumes. As for any linear program, it is not *a priori* clear which method of solving the RMP will perform "best." This may depend on the problem and the available solvers. Traditionally, primal or dual simplex methods are used (see Lasdon [19] for general comments on their suitability), but there are many alternatives. The *sifting method* [20], which is some sort of static column generation, can be a reasonable complement for large-scale RMPs [17,21]. Interior point methods like the *barrier method* can prove effective, although there is no warm-start (yet). Also, the *analytic center cutting plane method* [22] is advantageous as it produces interior point dual solutions. In addition to these general-purpose methods, one may stronger exploit duality.

As we have stressed before, the RMP should furnish dual multipliers. After some

initial iterations, a simplex method may produce relevant dual solutions which lead to progress, but then switching to a subgradient or more elaborate method to improve the dual solution may produce better dual bounds, and thus faster termination [23–25]. This can be cheaper and more stable (see the section titled "Stabilization of Dual Variables"), and may considerably reduce computation times. As the literature on Lagrangian relaxation is rich, there are many proposals for multiplier adjustment in subgradient methods that can be adapted to the column generation context. The RMP may itself be solved by subgradient algorithms by relaxing all its constraints in the objective function. This can be used as a primal heuristic as well, as proposed for set covering applications [26–28].

Subgradient algorithms suffer from a very restricted information; only the current subgradient is available. *Bundle methods* [29,30] therefore work with a set of subgradients, the *bundle*, from which the name is derived. It is true that a simplex method maintains a kind of bundle as well (the variables in the basis) but bundle methods may be more flexible. Bundle methods apply the proximal point idea of (quadratically) penalizing a deviation of the next iterate from the current best one in terms of the dual bound. This makes them attractive in the context of the section titled "Stabilization of Dual Variables" and explains their use in column generation [31]. It usually takes only a few iterations to produce an approximately optimal primal–dual pair.

The *volume algorithm* [32] is another extension of subgradient algorithms that also rapidly produces good approximations. It is so named because of a new way of looking at linear programming duality: using volumes below the active faces to compute the dual variable values and the direction of movement. The pricing subproblem is called with a dual solution "in a neighborhood" of an optimal dual solution. One can compute the probability that a particular column (which induces a face of the dual polyhedron) is generated. A modified subgradient method furnishes estimates of these probabilities, that is, approximate primal solutions. Primal feasibility may be mildly violated.

The volume algorithm, when used in alternation with the simplex method, produces dual solutions with a large number of nonzero variables [17], which may accelerate column generation. The computational experience has been promising [23,33] for various combinatorial optimization problems. Advantages of the volume algorithm are straight forward implementation with small memory requirements, numerical stability, and fast convergence.

Row Aggregation for Set Partitioning Problems. Primal degeneracy is an efficiency issue also in column generation, for example, for large-scale set partitioning problems. Because of the degenerate pivots, dual variables yield less reliable information for the pricing problem. A possible remedy is to group *similar* constraints and aggregate them into one [34], thus working with an RMP with much less rows. The intuition is that in applications like vehicle routing and crew scheduling, some activity sequences are more likely to occur than others: In airline crew scheduling a pilot usually stays on the same aircraft for several flight legs. Since aircraft itineraries are known prior to solving the crew pairing problem, it is natural to “guess” some aggregation of the flights to cover.

The method is not particular to column generation but can be used in this context. Most importantly, an aggregated RMP gives aggregated dual variables that need to be disaggregated. This should (and can) be done carefully so that the disaggregated dual solution fulfills many of the dual constraints. To ensure proper convergence and optimality, the aggregation is dynamically updated throughout the solution process. Tests conducted on the linear relaxation of the simultaneous vehicle and bus driver scheduling problem in urban mass transit show that this solution approach significantly reduces the size of the MP, the degeneracy, and the solution times, especially for larger problems: for an instance with 1600 set partitioning constraints, the RMP solution time is reduced by a factor of 8. A partial pricing strategy, called *multiphase dynamic constraint aggregation* [35], gives further significant speedup.

The Pricing Problem

The pricing problem provides a column that prices out profitably or proves that none exists. Any variable with negative reduced cost will do, be it obtained by an exact, approximate, or heuristic algorithm (the latter are first choice in terms of speed). One may even add positive reduced cost variables (possibly to a pool first). Sometimes relaxations of the pricing problem are solved, at the expense of a weaker dual bound, such as for vehicle routing problems [36]. Highly complex pricing problems (like in staff and duty scheduling) may be better solved by constraint programming as this offers a strong expressiveness of the model [37].

Pricing Schemes and Pricing Rules. For the simplex method many proposals have been made as to which columns to consider and according to which rule to choose when selecting a variable to enter the basis. Schemes like *full*, *partial*, or *cyclic* pricing find their analogs in column generation pricing. When there are *many* subproblems, it may be sensible to use partial/cyclic pricing in order to avoid the generation of many similar columns [38], but the number of iterations may increase. Dantzig’s classical most-negative reduced cost pricing rule is not the only choice. The Devex rule [39] (a practical variant of *steepest-edge* [40,41]) is reported to perform particularly well for set partitioning RMPs [42]. The dual analog, the *deepest-cut* rule [43], tries to cut away as much of the dual space as possible. It can be implemented heuristically and is reported to offer some speedup [44].

While steepest-edge is inherently based on the simplex method, deepest-cut is more independent from a particular solution method. This leads to the *lambda pricing* rule [20]. Assume that $c_j \geq 0, j \in J$. Clearly, the reduced cost $c_j - \pi^* \mathbf{a}_j$ is nonnegative for all $j \in J$ iff

$$\min_{j \in J} \left\{ \frac{c_j}{\pi^* \mathbf{a}_j} \mid \pi^* \mathbf{a}_j > 0 \right\} \geq 1. \quad (15)$$

At first glance, this is just a reformulation. However, Equation (15) takes advantage of

structural properties of (particular) set partitioning problems: picking columns with a small ratio accounts for smaller cost coefficients as well as for more nonzero entries in $\mathbf{a}_j$.

It is common in cutting plane algorithms to fill a *cut pool* first and select a *good* subset of cuts from it according to criteria like efficiency, orthogonality, sparsity, and others [18]. It remains to be seen how such criteria can be defined and applied for selecting good columns. Attempts to characterize dual facets [42] do not appear to have had any practical impact so far. It would be interesting to see other pricing rules *particular* to column generation, for example, with the aim of stabilization.

Pricing Problems when Solving Integer Programs. When the subproblem's domain X in Equation (3) is a mixed integer set, for example, when a Dantzig–Wolfe type decomposition is applied to a mixed integer original problem (6), pricing problems become mixed integer programs themselves (see also **Branch-Price-and-Cut Algorithms**). It is well known [9] that the dual bound from the RMP can be stronger than the LP relaxation only when the subproblem does not possess the *integrality property*. That is, the linear relaxation of the pricing problem should not give an integer solution. The trade-off in choosing a decomposition is between a strong dual bound (by adding also complicating constraints to the subproblem) and the manageability of the subproblem (by avoiding this). Sometimes a combinatorial algorithm is available for the pricing problem and a faster alternative to an integer program; often this is a dynamic program (like for resource constrained shortest path problems in routing applications), which has the advantage of providing more than one solution to the pricing problem. The latter can be achieved with integer programs as well as by using the *solution pool* that state-of-the-art solvers offer.

In particular, with the help of pricing heuristics, one often generates columns that resemble a good *integer* solution rather than an optimal *fractional* one (which may be much harder to characterize). One should keep in mind that what helps the integer

program need not help the linear program. Still, for example, for set partitioning RMPs a reasonable strategy is to generate columns of a rich diversity [45] (*complementary columns*).

STABILIZATION OF DUAL VARIABLES

Column generation is known to suffer from *tailing off* [46], that is, there is only incremental progress per iteration as we get closer to the optimum, in particular, for large and degenerate problems. There are several partial explanations (see Desrosiers and Lübbecke [47] for a summary), but the main reason lies in the *unstable* behavior of the dual variables. A dual solution may be far apart from the previous one (*bang-bang effect*, in Briant *et al.* [12] an example provided by Nemirovskii is cited, which drastically shows this behavior). *Stabilization* of the dual variables tries to reduce this effect. The principles are well established in the nonlinear programming world; choosing *good separation points* in cutting plane algorithms is the analogous concept [48].

It should be noted that in the case that stabilization is successful, regardless of the method employed, one typically observes a reduction in the number of column generation iterations. The downside of it is that the pricing problems become harder to solve on an average. However, among more sophisticated implementation techniques, stabilization may provide the largest performance gains [15].

Interior Point Stabilization

Solving the RMP by a simplex method gives an extreme point of the optimal face of the dual polyhedron. When this face has a large dimension, for example, when the primal is highly degenerate, there may be many extreme points, and the one obtained is essentially a “random choice” [20]. This extreme point is cut off in the next iteration; however, one would rather like to cut off the whole optimal face. In that sense, a simplex method may yield a “bad representative” of the optimal face. An immediate remedy to this may be to use an interior point

method instead, as one would cut off an interior point of the optimal face. Particular proposals have been using *analytic centers* [49], *volumetric centers*, and *central paths* [50], among others. Such concepts have been discussed for cutting plane algorithms as well [48].

A simplex-method-based approach to obtain a solution in the interior of the dual-optimal face is taken in Rousseau *et al.* [51]. It works in two steps and exploits the extremity of basic solutions. First, the RMP is solved and the objective function value is fixed to the optimum by adding an additional constraint. Then, several random objective functions $\mathbf{c}$ are chosen (and also the opposite direction $-\mathbf{c}$), each of which produces an extreme point of the optimal face. The final dual solution is a convex combination of all extreme points obtained. This approach is computationally expensive but easy to implement.

Boxstep Method

Instead of producing rather arbitrary interior points, one may introduce a control of the dual solution's trajectory. By imposing lower and upper bounds, dual variables are constrained to lie "in a box around" the previous dual solution π^* . The RMP that is thus restricted is reoptimized. If the new dual optimum is attained on the boundary of the box, we have a direction toward which the box should be relocated. Otherwise, the optimum is attained in the box's interior, producing the sought global optimum. This is the principle of the *Boxstep* method [52,53] and the basic idea of using a *stability center*, that is, our current best guess of an optimal dual solution that is in some sense "more reliable" than the other dual solutions. This is well known, for example, in trust-region methods, and it is the underlying mechanism of in what follows.

Polyhedral Penalty Terms

A hard-coded box is not very flexible. Instead, *stabilized column generation* [54] automates the recentering of the box to the current dual solution. Consider the following linear

program:

$$\begin{aligned} \min \quad & \mathbf{c}\lambda - \delta_- \mathbf{y}_- + \delta_+ \mathbf{y}_+ \\ \text{subject to} \quad & A\lambda - \mathbf{y}_- + \mathbf{y}_+ = \mathbf{b} \\ & \mathbf{y}_- \leq \boldsymbol{\varepsilon}_- \\ & \mathbf{y}_+ \leq \boldsymbol{\varepsilon}_+ \\ & \lambda, \mathbf{y}_-, \mathbf{y}_+ \geq \mathbf{0} \end{aligned} \quad (16)$$

and its dual:

$$\begin{aligned} \max \quad & \pi \mathbf{b} - \boldsymbol{\varepsilon}_- \mathbf{w}_- - \boldsymbol{\varepsilon}_+ \mathbf{w}_+ \\ \text{subject to} \quad & \pi A \leq \mathbf{c} \\ & -\pi - \mathbf{w}_- \leq -\delta_- \\ & \pi - \mathbf{w}_+ \leq \delta_+ \\ & \mathbf{w}_-, \mathbf{w}_+ \geq \mathbf{0}. \end{aligned} \quad (17)$$

Surplus and slack variables $\mathbf{y}_-$ and $\mathbf{y}_+$, respectively, perturb $\mathbf{b}$ by $\boldsymbol{\varepsilon} \in [-\boldsymbol{\varepsilon}_-, \boldsymbol{\varepsilon}_+]$, which helps to reduce degeneracy. The interpretation of Equation (17) is more interesting. The dual variables π are restricted to the interval $[\delta_- - \mathbf{w}_-, \delta_+ + \mathbf{w}_+]$ that is, deviation of π from the soft interval $[\delta_-, \delta_+]$ is allowed but penalized by an amount of $\boldsymbol{\varepsilon}_-, \boldsymbol{\varepsilon}_+$ per unit, respectively. From Equation (16) we obtain an optimal solution to the unperturbed problem $\min\{\mathbf{c}\lambda \mid A\lambda = \mathbf{b}, \lambda \geq \mathbf{0}\}$ when $\boldsymbol{\varepsilon}_- = \boldsymbol{\varepsilon}_+ = \mathbf{0}$ or $\delta_- < \hat{\pi} < \delta_+$, where $\hat{\pi}$ is an optimal solution to Equation (17). Therefore, the stopping criteria of a column generation algorithm become $v(PP) = 0$ and $\mathbf{y}_- = \mathbf{y}_+ = \mathbf{0}$.

This approach may need some parameter tuning, but it offers considerably speedup for some problems [54]. The change in the RMP requires addition of upper bounded artificial variables only, which does not increase the size of the basis. It can be easily generalized to piecewise linear penalty functions with more pieces, where five pieces appear to give a good compromise [55], with a stronger penalty further away from the stability center. Note that Equation (16) is a relaxation of the unperturbed RMP, and it may be computed faster.

Bundle Methods: Quadratic Penalty Term

The aim of the penalty terms is to encourage a dual solution to stay close to the stability center; so the penalty is larger the further

away we go. Pictorially, a quadratic penalty function can achieve this goal better than a piecewise linear penalty, and *bundle methods* do precisely this: penalizing the Euclidean distance to the stability center. There is an extensive comparison between bundle methods and “classical” stabilization techniques in Briant *et al.* [12], and the current conclusion is that there is no clear winner. The situation may change in favor of bundle methods when future developments bring improvements, for example, in quadratic programming.

Convex Combinations with Previous Dual Solutions

A different approach to avoid (too) large steps in the dual space does not need any modification to the RMP at all, but convex combines the current dual solution π^* with a previous one $\hat{\pi}$, that is, the pricing problem is called with $\alpha\hat{\pi} + (1 - \alpha)\pi^*$ for $0 \leq \alpha \leq 1$. When a column is found it is added to the RMP only when it has negative reduced cost with respect to π^* . The dual bound is updated whenever $L(\alpha\hat{\pi} + (1 - \alpha)\pi^*) > L(\hat{\pi})$. An interesting property is that even in the case when no column was added to the RMP (a *mis-price*) it holds that the dual bound improves to at least $L(\hat{\pi}) + \alpha(v(\text{RMP}) - L(\hat{\pi}))$ [56]. As a consequence the duality gap $v(\text{RMP}) - L(\hat{\pi})$ is reduced at least by a factor $(1 - \alpha)^{-1}$, that is, the method not only converges but does so at a proven rate. Only a single parameter has to be calibrated; however, because of this static choice of α , the stability center moves with less flexibility than in the previous proposals.

The convex combination with a dual solution that produced the current best dual bound is a rediscovery of the *weighted Dantzig–Wolfe decomposition* method [57], in which α is updated in each iteration. The stability center $\hat{\pi}$ becomes more reliable (larger α) the more often it leads to an improvement of the dual bound.

Valid Inequalities in the Dual Space

A complementary stabilization technique is to add valid inequalities to the dual. A simple

proposal is the relaxation of RMP equalities to inequalities (when possible), which imposes sign constraints to the dual variables [5]. The concept of *dual-optimal inequalities* [58,59] is more refined. One adds constraints $\pi E \leq \mathbf{e}$, which are valid for the optimal face of the dual polyhedron. The consequence in the primal is that additional variables are introduced, and the RMP becomes $\min\{\mathbf{c}\lambda + \mathbf{e}\mathbf{y} \mid A\lambda + E\mathbf{y} \geq \mathbf{b}, \lambda, \mathbf{y} \geq \mathbf{0}\}$. Deep dual-optimal inequalities [59] may even cut away dual-optimal solutions except at least one.

As an example, consider the one-dimensional cutting stock problem (4). It can be easily shown that if the orders are ranked such that $w_1 < w_2 < \dots < w_m$, then the dual variables satisfy the *ranking constraints* $\pi_1 \leq \pi_2 \leq \dots \leq \pi_m$. These $m - 1$ dual constraints can be generalized to larger sets [58,59]. Let $S_i = \{s \mid w_s < w_i\}$. Then

$$\sum_{s \in S} w_s \leq w_i \Rightarrow \sum_{s \in S} \pi_s \leq \pi_i, \quad S \subset S_i, \quad (18)$$

which significantly reduces the number of iterations in difficult instances [59].

As dual inequalities relax the primal RMP, one has to ensure primal feasibility of the final λ^* , which can be done by slightly perturbing the RMP [59]. The usefulness of adding valid dual inequalities has been demonstrated by constraining the dual variables to a small interval around their optimal values [55,59] (or a heuristic good guess). Such *perfect dual information* is available, for example, for the cutting stock triplet-problems, where each roll is cut into exactly 3 orders without any wastage, where $\pi_i = w_i/W$, $i = 1, \dots, m$ is dual optimal. It is further known that restricting the dual space can reduce primal degeneracy [58].

ACCELERATION TECHNIQUES AND IMPLEMENTATION ISSUES

Column generation is easily understood in theory but an implementation may suddenly reveal that there are many small pieces that need to fit together. We mention some of these in the sequel.

Libraries, Frameworks, Modeling Languages

Most people who implement a column generation code will at least rely on some package that provides an efficient simplex algorithm. There are plenty of packages available, both commercial and open-source, such as CLP [60], GLPK [61], and Soplex [62]. As noted above, there are alternatives (at least complements) to the simplex method, like the bundle method [63]. When we implement only column generation, the main loop is quickly written. The major implementation effort then probably remains for the pricing problem. The situation is a little different when doing branch-and-price, but there are several frameworks that support its implementation (and thus in particular column generation) such as ABACUS [64], BCP [65], SCIP [18], and SYMPHONY [66].

Frameworks have the advantage that they may automatically take care of features like using a *column pool*, which contains variables from previous pricing rounds, or *lazy constraints*, which are separated only when needed. This can be useful for constraints that are unlikely to be tight at optimality [67]. The main benefit from a framework is that it manages the branch-and-price tree, and that standard branching schemes, and others are available.

It is a little less known fact that column generation can also be implemented within several modeling languages like GAMS [68] or OPL [69], but a true branch-and-price is usually not supported. Since the user does not have access to all the internals, this option is probably not quite suitable for exactly solving very difficult problems, but it can be useful for practitioners working with the modeling language.

Suboptimal Solutions

Column generation and branch-and-price are exact methods, that is, in theory we obtain an optimal solution. The crux is that in practice, this may happen after a too long computation time, and one may wish to resort to a suboptimal solution. Fortunately, the dual bound gives a guarantee on its quality at any time. Heuristics should be used to construct or improve primal and dual solutions

as often as it seems useful. This point cannot be overestimated.

Numerical computations on a computer are in limited precision and there are several tolerances to be thought of: What is negative reduced cost? When comparing against 0.0, one easily ends up in an infinite loop because of numerical inaccuracies. When does the primal bound match the dual bound closely enough? When an explicit perturbation of the right-hand sides is used, of what magnitude will it be? Typically, for each of these tolerances one chooses some small value in the order of 10^{-3} to 10^{-6} . One can access the topic a bit more rigorously using the notion of ε -optimality [12]. An alternative is to resort to exact (rational) arithmetic; but due to performance reasons this is only advisable for mission-critical linear and integer programs.

Practitioners interested in primal solutions (found quickly) may choose some sort of price-and-branch, that is, pregenerate a reasonable set of variables in several rounds, and then solve the resulting program with standard branch-and-bound.

Some Simple Ideas that Often Work

Again, think of heuristics everywhere. Preprocess your problem well, in particular, when solving integer programs. For many problems on networks, the graph can be significantly reduced. Use a profiler to identify which part of the algorithm is the bottleneck. Typically, this will be the pricing problem but sometimes reoptimizing the RMP can be extremely time-consuming as well. Try solving the RMP only approximately and improve the dual solution with some iterations with a method from the Lagrangian world. Try dual variable stabilization; see the section titled “Stabilization of Dual Variables.” Try to avoid solving the pricing problem to optimality too often. Again, use heuristics first, maybe even a cascade of heuristics of increasing complexity. Relaxations serve the same purpose. Experiment with different parameters, in particular, how many columns are added to the RMP in each iteration; too few do not yield enough progress, and too many slow down computations (a combinatorial algorithm, or the

solution pool of your solver can return more than one column). For large-scale problems, it pays to remove columns that are nonbasic for too many iterations.

Many acceleration techniques are problem-dependent, but can often be adapted. The survey [70] in the context of vehicle routing and crew scheduling is very helpful in this respect. Re-read about the algorithmic alternatives in the section titled “Algorithmic Issues,” all of which can be (and have been) modified and combined (see also the section titled “Nonsimplex Algorithms for LP” in this encyclopedia). When everything fails, you need to research your problem (more thoroughly)! A proof that an optimal primal (or dual) solution you are looking for has a particular structure may restrict the search a lot.

CONCLUSIONS

Despite the obvious similarity to cutting plane techniques—both methods dynamically extend the model—column generation has significant differences. While cutting planes *can optionally be* added to the (already optimally solved) linear relaxation in order to strengthen it, one *has to* add negative reduced cost variables for, otherwise, one does not obtain a valid dual bound. This makes the competition a bit unfair but we believe that the future lies in integrating the two methods into one anyway.

Even though column generation was inceptioned more than half a century ago, the last decade was the most active in research and implementation. The availability of powerful computers and electronic large-scale data of hard practical problems challenged the community. The influence of nondifferential convex analysis, in particular, the idea of dual variable stabilization, was beneficial for the field. Still, column generation and branch-and-price are not available as generic implementations, and we are eager to see this change.

Until this happens, there are very elaborate suggestions for tailoring the method to particular problems, sometimes even particular problem instances. While this is

questionable in terms of general-purpose applicability, it is the driving force for pushing the border. Many interesting developments will certainly follow.

Column generation is clearly a success story in large-scale *integer* programming. The dual bound obtained from an extended reformulation is often stronger, the tailing off effect can be lessened, and the knowledge of the original formulation provides us with a guide for branching and cutting decisions in the search tree. Today we are in a position where branch-and-price codes solve many large-scale problems of “industrial difficulty,” that no standard commercial solver could cope with.

FURTHER READING

Previous general reviews on column generation include Barnhart *et al.* [71], Desrosiers *et al.* [72], Soumis [73], and Wilhelm [74]. This article is based on Desrosiers and Lübbecke [47]. The literature on applications of branch-and-price and column generation grew so quickly in recent years (see the book by Desaulniers *et al.* [75]) that it is likely that someone already proposed at least a partial solution to the application you have in mind. For further reading on branch-and-price refer to Desrosiers and Lübbecke [76] and the article titled ***Branch-Price-and-Cut Algorithms***.

REFERENCES

1. Ford LR, Fulkerson DR. A suggested computation for maximal multicommodity network flows. *Manage Sci* 1958;5:97–101.
2. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
3. Ogtilduz O. Generalized column generation for linear programming. *Manage Sci* 2002;48(3):444–452.
4. Ho JK, Louie E. Computational experience with advanced implementation of decomposition algorithms for linear programming. *Math Program* 1983;27:283–290.
5. Gilmore PC, Gomory RE. A linear programming approach to the cutting-stock problem. *Oper Res* 1961;9:849–859.

6. Farley AA. A note on bounding a class of linear programming problems, including cutting stock problems. *Oper Res* 1990;38(5): 922–923.
7. Valério de Carvalho JM. A note on branch-and-price algorithms for the one-dimensional cutting stock problem. *Comput Optim Appl* 2002;21(3):339–340.
8. Schrijver A. *Theory of linear and integer programming*. Chichester: John Wiley & Sons; 1986.
9. Geoffrion AM. Lagrangean relaxation for integer programming. *Math Program Stud* 1974;2: 82–114.
10. Kelley JE Jr. The cutting-plane method for solving convex programs. *J Soc Ind Appl Math* 1961;8(4):703–712.
11. Cheney EW, Goldstein AA. Newton's method for convex programming and Tchebycheff approximation. *Numer Math* 1959;1(1): 253–268.
12. Briant O, Lemaréchal C, Meurdesoif Ph, *et al.* Comparison of bundle and classical column generation. *Math Program Ser A* 2008;113(2): 299–344.
13. Grötschel M, Lovász L, Schrijver A. *Geometric algorithms and combinatorial optimization*. Berlin: Springer; 1988.
14. Chvátal V. *Linear programming*. New York: W.H. Freeman and Company; 1983.
15. Vanderbeck F. Implementing mixed integer column generation. In: Desaulniers G, Desrosiers G, Solomon MM, editors. *Column generation*. Berlin: Springer; 2005. pp. 331–358.
16. Agarwal Y, Mathur K, Salkin HM. A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks* 1989;19: 731–749.
17. Anbil R, Forrest JJ, Pulleyblank WR. Column generation and the airline crew pairing problem. *Proceedings of the International Congress of Mathematicians*; August 1998; Berlin. *Documenta Mathematica Extra Volume ICM III* 1998. pp. 677–686.
18. Achterberg T. SCIP: Solving constraint integer programs. *Math Program Comput* 2009;1 (1):1–41.
19. Lasdon LS. *Optimization theory for large systems*. London: Macmillan; 1970.
20. Bixby RE, Gregory JW, Lustig IJ, *et al.* Very large-scale linear programming: a case study in combining interior point and simplex methods. *Oper Res* 1992;40(5):885–897.
21. Chu HD, Gelman E, Johnson EL. Solving large scale crew scheduling problems. *Eur J Oper Res* 1997;97:260–268.
22. Goffin J-L, Haurie A, Vial J-Ph. Decomposition and nondifferentiable optimization with the projective algorithm. *Manage Sci* 1992;38(2):284–302.
23. Barahona F, Jensen D. Plant location with minimum inventory. *Math Program* 1998;83: 101–111.
24. Ceselli A, Righini G. A branch-and-price algorithm for the capacitated p -median problem. *Networks* 2005;45(3):125–142.
25. Mahey P. A subgradient algorithm for accelerating the Dantzig-Wolfe decomposition method. *Proceedings of the X. Symposium on Operations Research, Part I: Sections 1–5, Volume 53 of Methods Operations Research*, Königstein/Ts: Athenäum/Hain/Scriptor/Hanstein; 1986. pp. 697–707.
26. Caprara A, Fischetti M, Toth P. A heuristic method for the set covering problem. *Oper Res* 1999;47:730–743.
27. Caprara A, Fischetti M, Toth P. Algorithms for the set covering problem. *Ann Oper Res* 2000;98:353–371.
28. Wedelin D. An algorithm for large scale 0-1 integer programming with application to airline crew scheduling. *Ann Oper Res* 1995;57:283–301.
29. Lemaréchal C. An algorithm for minimizing convex functions. In: Rosenfeld JL, editor. *Information Processing '74*. Amsterdam: North Holland Publishing Co.; 1974. pp. 552–556.
30. Kiwiel KC. An aggregate subgradient method for nonsmooth convex minimization. *Math Program* 1983;27:320–341.
31. Kiwiel KC, Lemaréchal C. An inexact conic bundle variant suited to column generation. *Math Program* 2009;118(1):177–206.
32. Barahona F, Anbil R. The volume algorithm: producing primal solutions with a subgradient method. *Math Program* 2000;87(3):385–399.
33. Barahona F, Anbil R. On some difficult linear programs coming from set partitioning. *Discrete Appl Math* 2002;118(1–2):3–11.
34. Elhallaoui I, Villeneuve D, Soumis F, *et al.* Dynamic aggregation of set partitioning constraints in column generation. *Oper Res* 2005;53(4):632–645.
35. Elhallaoui I, Metrane A, Soumis F, *et al.* Multi-phase dynamic constraint aggregation for set partitioning type problems. *Math Program* 2010;123(2):345–370.

36. Desrochers M, Desrosiers J, Solomon MM. A new optimization algorithm for the vehicle routing problem with time windows. *Oper Res* 1992;40(2):342–354.
37. Junker U, Karisch SE, Kohl N, *et al.* A framework for constraint programming based column generation. In: Joxan Jaffer, editor. *Principles and practice of constraint programming*. Volume 1713, *Lecture Notes Computer Science*. Berlin: Springer-Verlag; 1999. pp. 261–275.
38. Gamache M, Soumis F, Marquis G, *et al.* A column generation approach for large-scale aircrew rostering problems. *Oper Res* 1999;47(2):247–263.
39. Harris PMJ. Pivot selection methods of the Devex LP code. *Math Program* 1973;5:1–28.
40. Forrest JJ, Goldfarb D. Steepest-edge simplex algorithms for linear programming. *Math Program* 1992;57:341–374.
41. Goldfarb D, Reid JK. A practicable steepest-edge simplex algorithm. *Math Program* 1977;12:361–371.
42. Sol M. Column generation techniques for pickup and delivery problems [PhD thesis]. Eindhoven University of Technology; 1994.
43. Vanderbeck F. Decomposition and column generation for integer programs [PhD thesis]. Université catholique de Louvain; 1994.
44. Papadakos N. Integrated airline scheduling. *Comput Oper Res* 2009;36:176–195.
45. Ghoniem A, Sherali HD. Complementary column generation and bounding approaches for set partitioning formulations. *Optim Lett* 2009;3(1):123–136.
46. Gilmore PC, Gomory RE. A linear programming approach to the cutting stock problem—Part II. *Oper Res* 1963;11:863–888.
47. Desrosiers J, Lübbecke ME. Selected topics in column generation. *Oper Res* 2005;53(6):1007–1023.
48. Ben-Ameur W, Neto J. Acceleration of cutting-plane and column generation algorithms: applications to network design. *Networks* 2007;49(1):3–17.
49. Elhedhli S, Goffin J-L. The integration of an interior-point cutting-plane method within a branch-and-price algorithm. *Math Program* 2004;100(2):267–294.
50. Kirkeby Martinson R, Tind J. An interior point method in Dantzig-Wolfe decomposition. *Comput Oper Res* 1999;26(12):1195–1216.
51. Rousseau L-M, Gendreau M, Feillet D. Interior point stabilization for column generation. *Oper Res Lett* 2007;35(5):660–668.
52. Marsten RE. The use of the boxstep method in discrete optimization. *Math Program Stud* 1975;3:127–144.
53. Marsten RE, Hogan WW, Blankenship JW. The BOXSTEP method for large-scale optimization. *Oper Res* 1975;23:389–405.
54. du Merle O, Villeneuve D, Desrosiers J, *et al.* Stabilized column generation. *Discrete Math* 1999;194:229–237.
55. Ben Amor HMT, Desrosiers J, Frangioni A. On the choice of explicit stabilizing terms in column generation. *Discrete Appl Math* 2009;157(6):1167–1184.
56. Pessoa A, Uchoa E, Poggi de Aragão M, *et al.* Algorithms over arc-time indexed formulations for single and parallel machine scheduling problems. Report RPEP Volume 8 no. 8. Universidade Federal Fluminense; 2008.
57. Wentges P. Weighted Dantzig-Wolfe decomposition of linear mixed-integer programming. *Int Trans Oper Res* 1997;4(2):151–162.
58. Valério de Carvalho JM. Using extra dual cuts to accelerate column generation. *INFORMS J Comput* 2005;17(2):175–182.
59. Ben Amor H, Desrosiers J, Valério de Carvalho JM. Dual-optimal inequalities for stabilized column generation. *Oper Res* 2006;54(3):454–463.
60. COIN-OR linear programming. <https://projects.coin-or.org/Clp>. 2010.
61. GNU linear programming kit. 2008. Available at <http://www.gnu.org/software/glpk>.
62. Sequential object-oriented simplex. 2010. Available at <http://soplex.zib.de>.
63. Helmberg C. ConicBundle library for convex optimization. 2009. Available at <http://www-user.tu-chemnitz.de/helmberg/ConicBundle>.
64. Jünger M, Thienel S. The ABACUS system for branch-and-cut-and-price algorithms in integer programming and combinatorial optimization. *Softw Pract Exper* 2000;30(11):1325–1352.
65. Ralphs TK, Ladányi L. COIN/BCP User's Manual. 2001. Available at <http://www.coin-or.org/Presentations/bcp-man.pdf>.
66. Ralphs TK. Symphony version 5.1 user's manual. Corl Laboratory Technical Report. 2006.
67. Cordeau J-F, Desaulniers G, Lingaya N, *et al.* Simultaneous locomotive and car assignment at VIA Rail Canada. *Transp Res B* 2001;35:767–787.
68. General algebraic modeling system. 2010. Available at <http://www.gams.com>.

69. IBM ILOG CPLEX optimization studio. 2010. Available at <http://www-01.ibm.com/software/integration/optimization/cplex-optimization-studio>.
70. Desaulniers G, Desrosiers J, Solomon MM. Accelerating strategies in column generation methods for vehicle routing and crew scheduling problems. In: Ribeiro CC, Hansen P, editors. *Essays and surveys in metaheuristics*. Boston (MA): Kluwer Academic publishers; 2001. pp. 309–324.
71. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. *Oper Res* 1998;46(3):316–329.
72. Desrosiers J, Dumas Y, Solomon MM, *et al.* Time constrained routing and scheduling. In: Ball MO, Magnanti TL, Monma CL, *et al.*, editors. *Network routing*. Volume 8, Handbooks in operations research and management science. Amsterdam: North-Holland Publishing Co.; 1995. pp. 35–139.
73. Soumis F. Decomposition and column generation. In: Dell'Amico M, Maffioli F, Martello S, editors. *Annotated bibliographies in combinatorial optimization*. Chichester: John Wiley & Sons; 1997. pp. 115–126.
74. Wilhelm WE. A technical review of column generation in integer programming. *Optim Eng* 2001;2:159–200.
75. Desaulniers G, Desrosiers J, Solomon MM, editors. *Column generation*. Berlin: Springer; 2005.
76. Desrosiers J, Lübbecke ME. A primer in column generation. In: Desaulniers G, Desrosiers J, Solomon MM, editors. *Column generation*. Berlin: Springer; 2005. pp. 1–32.

COMBINATORIAL AUCTIONS: COMPLEXITY AND ALGORITHMS

MARTIN BICHLER
Department of Informatics,
Technische Universität
München, Munich, Germany

An auction can be defined as “a market institution with an explicit set of rules determining resource allocation and prices on the basis of bids from the market participants” [1]. The competitive process serves to aggregate the scattered information about bidders’ valuations and to dynamically set a price. The auction format determines the rules governing when and how a deal is closed [2]. Auctions are typically evaluated using two main criteria, (*allocative*) *efficiency* and *revenue*. The former measures whether the objects end up with those bidders who value them most, while the latter focuses on the expected selling price.

Multiple object auctions can be divided into those, where *multiple units* of a single item are sold or where *multiple items* are sold. Of course, other combinations are also possible, where large quantities (multiple units) of different items get sold or bought, such as large quantities of different types of hard disk drives. Combinatorial auctions are a means to buy or sell multiple items. They have found application in a variety of domains such as the auctioning of spectrum licenses [3], truck load transportation [4], bus routes [5], and industrial procurement [6]. Original designs have been proposed by Rassenti *et al.* [7] for the allocation of airport time slots.

Combinatorial auctions address fundamental questions regarding efficiency and prices in markets [8,9]. These questions have been at the core of algorithmic mechanism design, a discipline at the intersection of Computer Science, Economics, and Operations Research. In this article, we focus on the design of combinatorial auctions, and also address some related types of auctions such as volume discount and multiattribute

auctions. We mainly look at the efficiency of auction formats as objective, as most of the literature in this area does. The article provides a concise introduction and is in parts based on publications such as Cramton *et al.* [10], which we refer to for a more detailed discussion. We assume that the reader has a basic knowledge about single-object auctions and the respective theory.

COMPLEXITY IN COMBINATORIAL AUCTIONS

Combinatorial auctions have been discussed in the literature, as they allow selling or buying a set of heterogeneous items to or from multiple bidders. Bidders can specify bundle bids, that is, a price is defined for a subset of the items for auction [10]. The price is only valid for the entire set and the set is indivisible. For example, in a combinatorial auction a bidder might want to buy 10 units of item x and 20 units of item y for a bundle price of \$100, which might be more than the total of the prices for the items x and y sold individually. We will refer to a bidding language as a set of allowable bid types (e.g., bundle bids or bids on price and quantity) in an auction. A bidding language allowing for bundle bids is also useful in procurement markets with economies of scope, where suppliers have cost complementarities due to reduced production or transportation costs for a set of items. In this case, we will discuss either about a combinatorial procurement auction or a combinatorial reverse auction.

Combinatorial auctions have been intensively discussed for the sale of spectrum licenses by the US Federal Communications Commission (FCC) [11]. The FCC divides licenses into different regions. Bidders - usually large telecom companies - often have superadditive preferences for licenses that are adjacent to each other. This can have advantages in advertising a service to the end customer, and also in the infrastructure that needs to be set up. In simultaneous

auctions where no bundle bids are allowed, bidders incur the risk that they only win a subset of items from a set of items that they are interested in, and that they end up paying too much for the subset. This is also called the *exposure problem*. These types of preferences can easily be considered in combinatorial auctions. However, the design of combinatorial auctions is such that several types of complexity can arise:

- The auctioneer faces *computational complexity* when determining an optimal allocation. The winner determination problem (WDP) in combinatorial auctions is an NP-hard problem [12]. In addition, the auctioneer needs to derive ask prices in iterative auctions, which is typically a hard computational problem as well.
- A bidder needs to determine his valuations for $2^m - 1$ bundles, where m is the number of items. We will refer to this as *valuation complexity*. Without restrictions, this would require to elicit 1023 valuations for an auction with only 10 items of interest.
- Even if the bidders knew their valuations perfectly, they would still need to decide how to respond during the auction. The issues relate to when and how they reveal their preferences. We will describe this as *strategic complexity*. Researchers have proposed different auction formats that exhibit various degrees of strategic complexity for bidders [13].
- Finally, *communication complexity* describes the number of messages that need to be exchanged between the auctioneer and the bidders in order to determine the optimal allocation. It has been shown that the communication complexity in combinatorial auctions is exponential in the number of items [14].

We focus on “Computational Complexity” in the next section and “Strategic Complexity” in the following section of this article. In the section titled “Combinatorial Auction Formats”, we discuss different auction

formats that have been suggested in the literature, and how they address these complexities.

COMPUTATIONAL COMPLEXITY

First, we will concentrate on the winner determination problem in combinatorial auctions [15–18]. It is a good example of the types of optimization problems that one encounters in various multiobject auctions. The following example with four bids and three items illustrates a simple procurement application (see Table 1). The buying organization needs different quantities of grain in different production sites. In this case, the buyer aggregates demand for multiple production sites, as suppliers might be able to provide better prices due to reduced production and transportation costs. Suppliers bid on subsets of the locations and each subset has a bundle price. In this article, we assume suppliers to provide the entire quantity for an item or location. In case they can provide subsets of the quantity, for example, only 500t of grain for Berlin, this is referred to as a multiunit combinatorial auction.

Given the bidder valuations for all possible bundles, the efficient allocation can be found by solving WDP. Let $\mathcal{K} = \{1, \dots, m\}$ denote the set of items indexed by k and $\mathcal{I} = \{1, \dots, n\}$ denote the set of bidders indexed by i with private valuations $v_i(S) \geq 0$ for bundles $S \subseteq \mathcal{K}$, and p as the price. This means, each bidder i has a valuation function $v_i : 2^{\mathcal{K}} \rightarrow \mathbb{R}_0^+$ that attaches a value $v_i(S)$ to any bundle

Table 1. Example with Bundle Bids

Line	Bids	Bids			
		B1	B2	B3	B4
1	1000t grain in Berlin	1	0	1	1
2	800t grain in Munich	0	1	1	1
3	800t grain in Vienna	1	1	1	0
4	Bid price (in thousands)	\$150	\$125	\$300	\$125

$S \subseteq \mathcal{K}$. In addition, we assume bidder values $v_i(S)$ to be independent and private (i.e., only known to the bidder), the bidders' utility function to be quasi-linear (i.e., the payoff of a bidder $\pi_i(S) = v_i(S) - p$) with free disposal (i.e., if $S \subset T$ then $v_i(S) \leq v_i(T)$). There are other situations where valuations are interdependent, such as the sale of a tract of land with an unknown amount of oil underground, where the bidders may have different estimates of the amount of oil based on privately conducted tests, but the final value is the same for all bidders. We focus on independent and private valuations in this article.

The WDP in a forward auction can be formulated as a binary program using the decision variables $x_i(S)$ which indicate whether the bid of the bidder i for the bundle S belongs to the allocation:

$$\begin{aligned} \max_{x_i(S)} \quad & \sum_{S \subseteq \mathcal{K}} \sum_{i \in \mathcal{I}} x_i(S) v_i(S) \\ \text{s.t.} \quad & \sum_{S \subseteq \mathcal{K}} x_i(S) \leq 1 \quad \forall i \in \mathcal{I} \quad (\text{WDP}) \\ & \sum_{S: k \in S} \sum_{i \in \mathcal{I}} x_i(S) \leq 1 \quad \forall k \in \mathcal{K} \\ & x_i(S) \in \{0, 1\} \quad \forall i, S \end{aligned}$$

The first set of constraints guarantees that any bidder can win at most one bundle, which is only relevant for the XOR bidding language. Without this constraint, bidders can win multiple bundles, which is referred to as an *OR bidding language*.

The XOR language is used because it is fully expressive compared to the OR language, that is, a bidder can express values for all possible subsets of items. Subadditive valuations, where a bundle is worth less than the sum of individual items, cannot be described appropriately without an exposure risk using the OR bidding language. The second set of constraints ensures that each item is only allocated once. Very early, it has been shown that the WDP (with an OR bidding language) is NP-hard by reducing it to the weighted set packing problem [15].

Theorem 1 [15]. *The decision version of the WDP with an OR bidding language is NP-complete, even if restricted to instances*

where every bid has a value equal to 1, and every bidder bids only on subsets of size of at most 3.

The same holds for an XOR bidding language, where bidders only bid on subsets of size of at most 2 [19]. The decision version of the WDP refers to the WDP, in which the auctioneer only wants to know, if there is an allocation with a revenue larger than a particular amount. Given those hardness results, one could try to approximate the WDP. Approximation algorithms are polynomial time algorithms with a provable performance guarantee on the deviation from the optimal solution. Unfortunately, it has been shown that for OR and XOR bidding languages, there are no polynomial algorithms that approximate the WDP within certain bounds. A comprehensive overview of complexity results in this area is given in Lehmann *et al.* [18].

There are, however, tractable cases if we restrict bids or valuations in a way that gives the bids a structure to allow for efficient solution methods. For example, the *goods are substitutes property* (aka. *substitutes condition*) leads to integral solutions of the LP-relaxation of the WDP.

Definition 1 [Substitutes condition [13]]. Bidder i considers the objects in $\mathcal{K}$ to be substitutes if for all $A \in \mathcal{K}$ and packages S and T not containing A , such that $S \subset T$, $v_i(S \cup \{A\}) - v_i(S) \geq v_i(T \cup \{A\}) - v_i(T)$

In other words, if items being sold are substitutes, the marginal value of obtaining a particular object A is smaller if the set of objects T already owned is larger than another set S . The substitutes condition would allow for additive valuations but not for complements or superadditive valuations, and will also play a role when determining ask prices in combinatorial auctions as discussed in the section titled "Combinatorial Auction Formats".

Theorem 2 [20]. *Let for all $i \in \mathcal{I}$ the bid values $v_i(S)$, $S \subseteq \mathcal{K}$ of the WDP with an XOR*

bidding language satisfy the substitutes condition, then the LP-relaxation of the WDP has an integral optimal solution.

A good overview of tractable cases of the WDP is provided in Mueller [21]. Unfortunately, the restrictions on tractable cases are so severe that auctioneers cannot rely on them in most applications of combinatorial auctions. Independent of this, extensive analyses of the empirical hardness of the WDP [22] illustrate that satisfactory performance can be obtained for problem sizes and structures occurring in practice. The problem sizes in many real-world applications have shown to be tractable within acceptable time limits [6].

Apart from bundle bids, other types of advanced bidding languages and respective auction formats have shown to be useful. In addition to traditional multiunit auctions, which allow for the specification of a price for a particular quantity, *volume discount bids* allow to specify supply curves, that is, unit prices for different quantities of an item sold. Suppliers can express economies of scale when bidding on very large quantities (e.g., \$500 per unit until 1000 units and \$450 per unit for more than 1000 units). Also here, buyers need to consider various business constraints when selecting such bids. For example, there might be limits on the spend per bidder or group of bidders, and upper and lower bounds on the number of winners. These side constraints, as well as limited capacity of suppliers, turn the WDP into a hard computational problem [23,24].

Multiattribute auctions allow bids on price and qualitative attributes such as delivery time or warranty. In contrast to request for quotes or tenders as they are regularly used in procurement, the purchasing manager specifies a scoring function that is used to evaluate bids. This enables competitive bidding with heterogeneous, but substitutable offers. Multi-attribute auctions differ in the types of scoring rules or functions used, and in the type of feedback that is provided to bidders. Depending on the type of bids submitted, and on the type of scoring function, the auctioneer faces different optimization problems [25].

STRATEGIC COMPLEXITY

In this section, we discuss about auction formats, which elicit bidders' preferences to an extent that the optimal solution to the WDP, that is, the economic efficient outcome can be selected by the auctioneer. As outlined in the first paragraph of this article, a central auction design objective is to obtain an *efficient allocation* $X^* = (S_1^*, \dots, S_n^*)$, where S_i^* is bidder i 's bundle in this allocation.

Definition 2 [Allocative efficiency]. Allocative efficiency is measured as the ratio of the total valuation of the auction outcome X to the maximum possible valuation of an allocation (i.e., the efficient allocation) X^* :

$$E(X) = \frac{\sum_{i \in \mathcal{I}} v_i \left(\bigcup_{S \subseteq \mathcal{K}: x_i(S)=1} S \right)}{\sum_{i \in \mathcal{I}} v_i \left(\bigcup_{S \subseteq \mathcal{K}: x_i^*(S)=1} S \right)}$$

Since typically in auctions the bidder valuations are not given and bidders have incentives to lie about their true preferences, strategic complexity is a concern in the design of combinatorial auctions. Strategic complexity is concerned with the effort it takes for a bidder to determine his optimal bidding strategies. In some auction formats bidders might not be willing to reveal their true preferences and rather speculate, which is one of the main sources of inefficiency in auctions.

Incentive compatibility and *strategy proofness* are properties that should lead bidders to reveal their true private valuations to an extent that the auctioneer can determine the efficient allocation, without the need for further speculation about other bidders' preferences. An auction is incentive compatible, if truthful revelation is a Bayes–Nash equilibrium. In other words, truth revelation is optimal for a bidder, if and only if, all other bidders in a game with uncertainty about the types of other bidders reveal their valuations truthfully. An auction is strategy proof, if truth revelation is a dominant strategy for bidders, that is, it is the bidder's best strategy independent of other bidders' types and

strategies. In these cases, the strategic complexity of an auction is reduced to a minimum and speculation is not necessary.

Traditional single-object auction theory distinguishes at least four different types of auction formats: first-price sealed bid, Dutch, English, and second-price sealed bid auctions [13]. The first-price sealed bid and the Dutch auction are strategically equivalent, as are the English and the second-price sealed bid. The second-price sealed-bid auction or Vickrey auction has a dominant strategy, and the same holds for a simple implementation of the English auction, in which the auctioneer is replaced by an upward ticking clock, and bidders cannot place jump bids, but only drop out at a certain price level. This is often referred to as a Japanese or clock auction. The clock auction can be described as iterative or ascending auction format, where a bidder learns about the willingness-to-pay of other bidders during the course of the auction. Efficiency in dominant strategies is a desirable property of auction mechanisms. There is a generalization of the second-price sealed bid auctions (aka. Vickrey-Clarke-Groves auction) to multiple-item auctions, which maintains its dominant strategy property (see section “Combinatorial Auction Formats”). It is not obvious that a generalization of the clock auction or any other iterative auction format has similar properties.

General equilibrium models have been developed in Economics to show that in markets with multiple items, the Walrasian price mechanism also known as *tâtonnement*, which uses item-level or linear prices, actually yields the efficient allocation [26] while communicating as few real variables as possible [27,28]. As a consequence, the First Welfare Theorem shows Pareto-efficiency of allocations obtained at those equilibrium prices. The *tâtonnement* works as follows: Prices are cried, and agents register how much of each good they would like to offer or purchase. No transactions and no production take place at disequilibrium prices. Then, prices are lowered for goods with positive prices and excess supply, and prices are raised for goods with excess demand until no agent wants to deviate from his allocation.

However, these results assume that all production sets and preferences are convex. The results do not carry over to nonconvex economies with indivisible items, such as they often occur in combinatorial auctions. The question is, whether a combinatorial auction mechanism can be fully efficient, and, if so, what types of equilibrium prices are necessary. We will introduce the notion of “Competitive Equilibrium” for the following discussion.

Definition 3 [Competitive equilibrium, CE [29]]. Prices $\mathcal{P}$, and allocation X^* are in competitive equilibrium if allocation X^* maximizes the payoff of every bidder and the auctioneer revenue given prices $\mathcal{P}$. The allocation X^* is said to be *supported* by prices $\mathcal{P}$ in CE.

The first approach would be to use the same Walrasian price mechanism and see if it produced efficient outcomes in combinatorial auctions, where indivisibilities are present. Unfortunately, without convexity assumptions full efficiency cannot be achieved with simple linear competitive equilibrium prices in a combinatorial auction with unrestricted bidder valuations. It has been shown that a CE always exists in combinatorial auctions, but it possibly requires nonlinear and nonanonymous prices [30,31]. Prices are *nonlinear* if the price of a bundle is not equal to the sum of prices of its items, and prices are *nonanonymous* or *personalized* if prices for the same item or bundle differ across bidders. This leads to the following classification of CE prices:

1. linear anonymous prices $\mathcal{P} = \{p(k)\}$
2. linear personalized prices $\mathcal{P} = \{p_i(k)\}$
3. nonlinear anonymous prices $\mathcal{P} = \{p(S)\}$
4. nonlinear personalized prices $\mathcal{P} = \{p_i(S)\}$

Indeed, there have been proposals for *ascending* combinatorial auctions with nonlinear and personalized prices, which have been shown to be fully efficient if bidders follow a straightforward bidding strategy [32]. Such a strategy assumes

that bidders bid only on those bundles, which maximize their payoff in each round. Unfortunately, straightforward bidding is only a best response for bidders in pure ascending combinatorial auction formats, if bidders' valuations are restricted (see section "Combinatorial Auction Formats").

More generally, it is known that the only efficient mechanisms in which honest revelation is a dominant strategy for each agent is the Vickrey-Clarke-Groves (VCG) mechanism [33]. VCG mechanisms, however, exhibit significant problems in practical applications [34,35]. Among others, the VCG mechanism can lead to low seller revenues, nonmonotonicity of the seller's revenues in the set of bidders, and is susceptible to collusion. Apart from this, all bidders would need to submit all their valuations for an exponential number of bundles, which is not practical for all but very small auctions with only a few items.

In summary, designing strategy-proof and practical combinatorial auction formats turns out to be a formidable task. The VCG auction does not seem practical in most applications and iterative forms of combinatorial auctions are bound to nonlinear and personalized competitive equilibrium prices for full efficiency.

COMBINATORIAL AUCTION FORMATS

In the following, we provide an overview of well-known combinatorial auction formats and discuss some of the concepts from the overview in the previous section in more detail.

The Vickrey-Clarke-Groves Auction

VCG mechanisms describe a class of strategy-proof economic mechanisms [36,37], where sealed bids are submitted to the auctioneer. The winners are also determined by the WDP. However, rather than paying the bid prices, the winners pay a discounted price. This price is calculated in the following manner.

$$p_i^{VCG} = v_i(X^*) - [w(\mathcal{I}) - w(\mathcal{I}_{-i})]$$

Here p_i^{VCG} describes the Vickrey price, while $w(\mathcal{I})$ is the objective value the WDP

with the valuations of all bidders, and $w(\mathcal{I}_{-i})$ is the objective value to the WDP with all bidders except the winning bidder i . If the auction is modeled as a coalitional game, $w(\bullet)$ can also be referred to as the *coalitional value function*, that is, the outcome of the auction game with a certain set of bidders. In a combinatorial auction, this means that a bidder needs to submit bids on all possible bundles, a number which is exponential in the number of items. Each winning bidder receives a Vickrey payment, which is the amount that he has contributed to increasing the total value of the auctioneer.

Let us take an example with two items x and y which are to be sold in a combinatorial auction. The bids of bidder 1 and 2 are described in Table 2. The total value will be maximized at \$34, while selling x to bidder 1 and y to bidder 2. Bidder 1 bids \$20 for x , but he receives a Vickrey payment of \$34 - \$29 = \$5, since without his participation the total value would be \$29. In other words, the net payment or Vickrey price p_1^{VCG} bidder 1 has to pay to the auctioneer is (\$20 (bid price) - \$5 (Vickrey payment) =) \$15. Bidder 2 bids \$14 on y , but receives a Vickrey payment of \$34 - \$33 = \$1, because without his participation the total valuation of this auction would be \$33. Auctioneer revenue would then be \$15 + \$13 = \$28 in this auction.

In this auction bidders have a dominant strategy of reporting their true valuations $b_i(S) = v_i(S)$ on all bundles S to the auctioneer, who then determines the allocation and respective Vickrey prices. As already introduced in the previous section, the VCG design suffers from a number of practical problems.

The decisive fault of the VCG is best understood if the auction is modeled as a coalitional game [34]. (N, w) is the coalitional game derived from trade between the seller and bidders. Let N denote the set of all bidders $\mathcal{I}$ plus the auctioneer with $i \in N$, and

Table 2. Bids Submitted in a VCG Auction

Items	Bids		
	{x}	{y}	{x,y}
Bidder 1	20*	11	33
Bidder 2	14	14*	29

Table 3. Bids Submitted in a VCG Auction

Items	Bids		
	$\{x\}$	$\{y\}$	$\{x, y\}$
Bidder 1	0	2	2
Bidder 2	2	0	2
Bidder 3	0	0	2

$M \subseteq N$ be a coalition of bidders with the auctioneer. Let $w(M)$ denote the coalitional value for a subset M , equal to the objective value of the WDP with all bidders $i \in M$ involved. A core payoff vector Π , that is, payoffs Π_i of the bidders in this auction, is then defined as follows

$$\text{Core}(N, w) = \left\{ \Pi \geq 0 \mid \sum_{i \in N} \pi_i = w(N), \right. \\ \left. \sum_{i \in M} \pi_i \geq w(M) \quad \forall M \subset N \right\}$$

This means, there should be no coalition $M \subset N$, which can make a counteroffer that leaves themselves and the seller at least as well off as the currently winning coalition. Unfortunately, in the VCG auction there can be outcomes which are not in the core. To see this, assume again a combinatorial sales auction with three bidders and two items (see Table 3).

Bidder 1 bids $b_1(x) = \$0$, $b_1(y) = \$2$ and $b_1(x, y) = \$2$. Bidder 2 bids $b_2(x) = \$2$, $b_2(y) = \$0$ and $b_2(x, y) = \$2$. Finally, bidder 3 only has a bid of $b_3(x, y) = \$2$, but no valuation for the individual items. In this situation the net payments of the winners (bidder 2 and 3) are zero, and bidder 3 could find a solution with the auctioneer that makes both better off. It has been shown that there is an equivalence between the core of the coalitional game and the competitive equilibrium for single-sided auctions [31]. Outcomes, which are not in the core lead to a number of problems, such as low seller revenues or nonmonotonicity of the seller's revenues in the set of bidders and the

amounts bid. To see this, just omit bidder 1 from the auction. Also, such auction results are vulnerable to collusion by a coalition of losing bidders. Therefore, it has been argued that the outcomes of combinatorial auctions should be in the core [38].

The *bidders are substitutes condition* (BSC) is necessary and sufficient to support VCG payments in competitive equilibrium [31]. A bidder's payment in the VCG mechanism is always less than or equal to the payment by a bidder at any other CE.

Definition 4 [Bidders are Substitutes Condition, BSC]. The BSC condition requires

$$w(N) - w(N \setminus M) \geq \sum_{i \in M} [w(N) - w(N \setminus i)], \\ \forall M \subseteq N$$

In other words, BSC holds where the incremental value of a subset of bidders to the grand coalition is at least as great as the sum of the incremental contributions of each of its members. When at least one bidder has a nonsubstitutes valuation an ascending CA cannot implement the VCG outcome [39].

Nonlinear Personalized Price Auctions

In this section, we discuss relevant theory with respect to ascending combinatorial auctions using nonlinear and personalized prices (NLPPAs). We have seen that the WDP is a nonconvex optimization problem. By adding constraints for each set partition of items and each bidder to the WDP the formulation can be strengthened, so that the integrality constraints on all variables can be omitted but the solution is still always integral [31, 39]. Such a formulation describes every feasible solution to an integer problem, and is solvable with linear programming. We will refer to this formulation as NLPPA WDP.

$$\begin{aligned} \max \quad & \sum_{i \in \mathcal{I}} \sum_{S \subseteq \mathcal{K}} v_i(S) x_i(S) \\ \text{s.t.} \quad & x_i(S) = \sum_{X: x_i = S} \delta_X \quad \forall i \in \mathcal{I}, \forall S \subseteq \mathcal{K} \quad (p_i(S)) \\ & \sum_{S \subseteq \mathcal{K}} x_i(S) \leq 1 \quad \forall i \in \mathcal{I} \quad (\pi_i) \\ & \sum_{X \in \Gamma} \delta_X = 1 \quad (\pi^s) \\ & 0 \leq x_i(S) \quad \forall S \subseteq \mathcal{K}, \forall i \in \mathcal{I} \\ & 0 \leq \delta_X \quad \forall X \in \Gamma \end{aligned} \quad \text{(NLPPA WDP)}$$

Personalized nonlinear CE prices can now be derived from the dual of the NLPPA WDP. In the first side constraint, $x_i(S)$ is equal to the sum of weights δ_X over all allocations X where bidder i gets bundle S . The dual variables of this constraint are the personalized prices $p_i(S)$. The second side constraint makes sure that each bidder i receives at most one bundle, and the dual variable π_i describes bidder i 's payoff. Finally, the total weight of all selected allocations $X \in \Gamma$ equals 1, such that only one allocation can be selected. Here, Γ describes the set of all possible allocations. The dual variable (π^s) for this side constraint describes the seller's payoff.

From duality theory follows that the *complementary slackness conditions* must hold in the case of optimality. This is equivalent to the CE, where every buyer receives a bundle out of his demand set or demand correspondence $D_i(\mathcal{P})$, that is, the bundles maximizing his payoff at the prices, and the auctioneer selects the revenue maximizing allocation at these prices.

Definition 5 [Demand set]. The demand set $D_i(\mathcal{P})$ of a bidder i includes all bundles which maximize a bidder's payoff π_i at the given prices $\mathcal{P}$:

$$D_i(\mathcal{P}) = \left\{ S : \pi_i(S, \mathcal{P}) \geq \max_{T \subseteq \mathcal{K}} \pi_i(T, \mathcal{P}), \right. \\ \left. \pi_i(S, \mathcal{P}) \geq 0, S \subseteq \mathcal{K} \right\}$$

Complementary slackness provides us with an optimality condition, which also serves as a termination rule for NLPPAs. If bidders follow the straightforward strategy then terminating the auction when each active bidder receives a bundle in his demand set will result in the efficient outcome. Note that a demand set can include the empty bundle. In addition, the starting prices must represent a feasible dual solution. A trivial solution is to use zero prices for all bundles.

Although such nonlinear personalized prices always exist, the NLPPA WDP is huge since one must enumerate all possible feasible coalitions. Nevertheless, it has

provided a guideline to a number of practical auction designs using nonlinear personalized prices. Individual NLPPA formats discussed in the following such as the Ascending Proxy Auction, iBundle, and the dVSV auction have different rules for determining the prices provided to the bidders and for determining how bidders submit new bids based on these announced prices.

iBundle [40] calculates a provisional revenue maximizing allocation at the end of every round and increases the prices based on the bids of nonwinning bidders. Three different versions of iBundle have been suggested [40]: iBundle(2) with anonymous prices, iBundle(3) with personalized prices, and iBundle(d) that starts with anonymous prices and switches to personalized prices for agents who submit bids for disjoint bundles. The *ascending proxy auction* [32] is similar to iBundle(3), but the use of proxy agents is mandatory, which essentially leads to a sealed-bid auction format.

The dVSV auction [39] design differs from iBundle in that it does not compute a provisional allocation in every round but increases prices for one *minimally undersupplied set of bidders*. A set of bidders is minimally undersupplied if each bidder in this set receives a bundle from his demand set, and removing only one of the bidders from the set forfeits this property. Similar to iBundle(3), it maintains nonlinear personalized prices and increases the prices for all agents in a minimally undersupplied set based on their bids of the last round. While the Ascending Proxy Auction can be interpreted as a sub-gradient algorithm, the dVSV auction can be interpreted as a primal-dual algorithm for the NLPPA WDP [39].

Even though the BSC condition is sufficient for VCG prices to be supported in CE, the slightly stronger *bidder submodularity condition (BSM)* is required for a pure ascending combinatorial auction to implement VCG payments [39].

Definition 6 [Bidder submodularity condition, BSM]. BSM requires that for all $M \subseteq M' \subseteq N$ and all $i \in N$ there is

$$w(M \cup \{i\}) - w(M) \geq w(M' \cup \{i\}) - w(M')$$

Here bidders are more valuable, when added to a smaller coalition. Under BSM the NLPPAs yield VCG payments and straightforward bidding is an ex post equilibrium. An ex post equilibrium is stronger than a Bayes-Nash equilibrium, but weaker than a dominant strategy equilibrium. It does not require bidders to speculate about other bidders' types, but requires assumptions about their strategies. When the BSM condition does not hold, the property breaks down and a straightforward strategy is likely to lead a bidder to pay more than the VCG price for the winning bundle, and bidders have an incentive to shade their bids and deviate from straightforward bidding. In case of nonstraightforward bidding the outcome of NLPPAs can deviate significantly from the efficient solution [41].

The restriction to BSM valuations is mainly due to the definition of ascending auctions, such that prices can only increase and no payments from the auctioneer are allowed. The *Credit-Debit auction* is an extension to the dVSV design which achieves the VCG outcome for general valuations by determining payments or discounts from the auctioneer to the bidders at the end. Similarly, *iBEA* is described as an extension of iBundle. Both approaches are based on universal competitive equilibrium (UCE) prices, which are CE prices for the main economy as well as for every marginal economy, where a single buyer is excluded [42]. These auctions terminate as soon as UCE prices are reached and VCG payments are determined as one-time discounts dynamically during the auction. Truthful bidding is an ex post equilibrium in the Credit-Debit auction and iBEA. The auctions are an important contribution to the literature, because they describe fully efficient iterative combinatorial auctions where straightforward bidding is an ex post equilibrium for general valuations. However, they share a central problem of the VCG auction: if buyer submodularity does not hold, the outcomes might not be in the core.

Linear Price Auctions

In many applications of ICAs, linear and anonymous ask prices are essential. For

example, day-ahead markets for electricity sacrifice efficiency for the sake of having linear prices [43]. Also, the main auction formats, which have been tested for selling spectrum in the United States used linear ask prices [44]. Simple examples illustrate that linear anonymous CE prices do not exist for general valuations. It has been shown that the *goods are substitutes* property is a sufficient condition for the existence of the exact linear CE prices [20], as the LP-relaxation of the WDP is integral (see "Computational Complexity") and dual variables can be interpreted as prices. The substitutes condition is, however, very restrictive and not satisfied in most combinatorial auctions. In spite of these negative results, some combinatorial auction designs with linear prices achieved high levels of efficiency in the lab.

The *CCA* (combinatorial clock auction) [45] utilizes anonymous linear ask prices called *item clock prices*. In each round bidders express the quantities desired on the bundles at the current prices. As long as demand exceeds supply for at least one item (each item is counted only once for each bidder) the price clock "ticks" upwards for those items (the item prices are increased by a fixed price increment), and the auction moves on to the next round. If there is no excess demand and no excess supply, the items are allocated corresponding to the last round bids and the auction terminates. If there is no excess demand but there is excess supply (all active bidders on some item did not resubmit their bids in the last round), the auctioneer solves the WDP considering all bids submitted during the auction runtime. If the computed allocation does not displace any bids from the last round, the auction terminates with this allocation, otherwise the prices of the respective items are increased and the auction continues. Note that due to the winner determination the final payments can deviate from the ask prices.

The *RAD* (resource allocation design) proposed in Kwasnica *et al.* [46] uses anonymous linear ask prices. However, instead of increasing the prices in case of over-demand, the auction lets the bidders submit

priced bids and calculates so called pseudo-dual prices based on a restricted dual of the LP relaxation of the WDP [7]. The dual price of each item measures the cost of not awarding the item to whom it has been allocated in the last round. In each round the losing bidders have to bid more than the sum of ask prices for a desired bundle plus a fixed minimum increment. RAD suggests an OR bidding language and only winning bids remain in the auction in its original design. The *ALPS* (approximate linear prices) design [47] is also based on the ideas in Rassenti *et al.* [7], but improves termination rules and the ask price calculation to better balance prices across items and have the auction avoid cycles. Note that in RAD and ALPS, prices can also decrease if the competition shifts to different items.

HPB (hierarchical package bidding) imposes a hierarchical structure of allowed package bids. This hierarchy and an OR bidding language reduces the WDP to a computationally simple problem that can be solved in linear time [15]. If the hierarchy meets the bidder's preferences, the auction is likely to achieve efficient outcomes, and reduces the strategic complexity for bidders. HPB provides a simple and transparent pricing mechanism [48]. It uses a recursive algorithm to determine new ask prices, which starts with the highest bids on every single item as a lower bound, adding a tax if the next level package received a bid higher than the sum of the single item bids contained in the package. The difference is distributed uniformly upon the respective item prices. The algorithm ends evaluating the package(s) of the top level, resulting in new ask prices for each item.

A few other combinatorial auction designs have been suggested, which use linear and nonlinear prices. For example, in the Clock-Proxy auction a clock auction is followed by a best-and-final ascending proxy auction [49]. The approach combines the simple and transparent price discovery of the clock auction with the efficiency of the ascending proxy auction. Progressive adaptive user selection environment (PAUSE) combines the simultaneous multiround auction with bidding on bundles in later

stages. Here, the burden of evaluating a combinatorial bid is transferred to the bidder [50]. Also, alternative ways of pricing and bidder support have shown promising results [51].

Interestingly, experimental research has shown that iterative auction designs with linear prices achieved very high levels of efficiency, even for auctions with up to 18 items [44,48,52]. While linear competitive equilibrium prices do not always exist, linear ask prices used in the combinatorial clock auction [45], HPB [48], or ALPS [47] have shown to be a good guideline to bidders in finding the efficient solution, even though no formal equilibrium analysis is available for any of these auction formats.

CONCLUSIONS

Many theoretical results on combinatorial auctions are negative in the sense that it seems quite unlikely that practical applications would satisfy the assumptions, which would lead to efficiency with a strong game-theoretical solution concept. Nevertheless, experimental results have yielded very high levels of efficiency in the lab. These results suggest that even if full efficiency is not always possible, combinatorial auction designs can achieve very high levels of efficiency, higher than what would be possible in simultaneous or sequential auctions in the presence of complementarities. The results of this research can have significant impact on the design and the efficiency of real-world markets. Further development of practical combinatorial auction designs will probably remain an active and rewarding area of theoretical, experimental, and applied research for the foreseeable future.

Acknowledgments

Special thanks go to Jannis Petrakis, Stefan Schneider, and Pasha Shabalin.

REFERENCES

1. McAfee R, McMillan PJ. Auctions and bidding. *J Econ Lit* 1987;25:699–738.

2. Klemperer P. Auction theory: a guide to the literature. *J Econ Surv* 1999;13(3):227–260.
3. Cramton P, Spectrum auction design. Working paper. University of Maryland, Department of Economics; 2009.
4. Caplice C, Sheffi Y. Combinatorial auctions for truckload transportation. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
5. Cantillon E, Pesendorfer M. Auctioning bus routes: the london experience. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
6. Bichler M, Davenport A, Hohner G, *et al.* Industrial procurement auctions. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. MIT Press; 2006.
7. Rassenti S, Smith VL, Bulfin RL. A combinatorial auction mechanism for airport time slot allocations. *Bell J Econ* 1982;13:402–417.
8. Nisan N, Ronen A. Algorithmic mechanism design. *Games Econ Behav* 2001;35:166–196.
9. Vazirani VV, Nisan N, Roughgarden T, *et al.* *Algorithmic game theory*. Cambridge (UK): Cambridge University Press; 2007.
10. Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
11. Milgrom P. Putting auction theory to work. Cambridge (UK): Cambridge University Press; 2004.
12. Garey MR, Johnson DS, editors. *Computers and intractability - a guide to the theory of NP-completeness*. New York: W. H. Freeman and Company; 1972.
13. Krishna V, editor. *Auction theory*. San Diego (CA): Elsevier Science; 2002.
14. Nisan N, Segal I. The communication requirements of efficient allocations and supporting prices. *J Econ Theory* 2006;129:192–224.
15. Rothkopf MH, Pekec A, Harstad RM. Computationally manageable combinatorial auctions. *Manage Sci* 1998;44:1131–1147.
16. Sandholm T. Approaches to winner determination in combinatorial auctions. *Decis Support Syst* 1999;28(1):165–176.
17. de Vries S, Vohra R. Combinatorial auctions: A survey. *INFORMS J Comput* 2003;15(3):284–309.
18. Lehmann D, Mueller R, Sandholm T. The winner determination problem. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
19. Hoesel S, Mueller R. Optimization in electronic markets: examples in combinatorial auctions. *Netnomics* 2001;3:23–33.
20. Kelso AS, Crawford VP. Job matching, coalition formation, and gross substitute. *Econometrica* 1982;50:1483–1504.
21. Mueller R. Tractable cases of the winner determination problem. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
22. Leyton-Brown K, Nudelman E, Shoham Y. Empirical hardness models: Methodology and a case study on combinatorial auctions. *J ACM* 2009;56:1–52.
23. Davenport A, Kalagnanam J. Price negotiations for procurement of direct inputs. IMA “Hot Topics” Workshop: Mathematics of the Internet: E-Auction and Markets, Volume 127; Minneapolis; 2000. pp. 27–44.
24. Goossens DR, Maas AJT, Spieksma F, *et al.* Exact algorithms for procurement problems under a total quantity discount structure. *Eur J Oper Res* 2007;178:603–626.
25. Bichler M, Kalagnanam J. Configurable offers and winner determination in multi-attribute auctions. *Eur J Oper Res* 2005;160(2):380–394.
26. Arrow KJ, Debreu G. Existence of an equilibrium for competitive economy. *Econometrica* 1954;22:265–290.
27. Mount K, Reiter S. The information size of message spaces. *J Econ Theory* 1974;28:1–18.
28. Hurwicz L. On the dimensional requirements of informationally decentralized pareto-satisfactory processes. In: Arrow K, Hurwicz L, editors. *Studies in resource allocation processes*. New York: Cambridge University Press; 1977.
29. Parkes D. Iterative combinatorial auctions. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.
30. Bikhchandani S, Mamer JW. Competitive equilibrium in an exchange economy with indivisibilities. *J Econ Theory* 1997;74:385–413.
31. Bikhchandani S, Ostroy JM. The package assignment model. *J Econ Theory* 2002;107(2):377–406.
32. Ausubel L, Milgrom P. Ascending proxy auctions. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006.

- editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006.
33. Green J, Laffont J-J. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica* 1977;45:427–438.
 34. Ausubel L, Milgrom P. The lovely but lonely vickrey auction. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006.
 35. Rothkopf MH. Thirteen reasons why the vickrey-clarke-groves process is not practical. *Oper Res* 2007;55:191–197.
 36. Vickrey W. Counterspeculation, auctions, and competitive sealed tenders. *J Finance* 1961; 16(3):8–37.
 37. Groves T. Incentives in teams. *Econometrica* 1973;41:617–631.
 38. Day R, Milgrom P. Core-selecting package auctions. *Int J Game Theory* 2008;38: 393–407.
 39. de Vries S, Schummer J, Vohra R. On ascending vickrey auctions for heterogeneous objects. *J Econ Theory* 2007;132:95–118.
 40. Parkes D, Ungar LH. Iterative combinatorial auctions: theory and practice. 17th National Conference on Artificial Intelligence (AAAI-00); Austin, Texas, USA: 2000.
 41. Schneider S, Shabalin P, Bichler M. On the robustness of non-linear personalized price combinatorial auctions. *European Journal of Operational Research*, 2010;206(1):248–259.
 42. Mishra D, Parkes D. Ascending price vickrey auctions for general valuations. *J Econ Theory* 2007;132:335–366.
 43. Meeus L, Verhaegen K, Belmans R. Block order restrictions in combinatorial electric energy auctions. *Eur J Oper Res* 2009; 196:1202–1206.
 44. Brunner C, Goeree JK, Hold C, *et al.* An experimental test of flexible combinatorial spectrum auction formats *Am Econ J Micro Econ* 2010;2(1):39–57.
 45. Porter D, Rassenti S, Roopnarine A, *et al.* Combinatorial auction design. *Proc Natl Acad Sci U S A* 2003;100:11153–11157.
 46. Kwasnica T, Ledyard JO, Porter D, *et al.* A new and improved design for multi-objective iterative auctions. *Manage Sci* 2005; 51(3):419–434.
 47. Bichler M, Shabalin P, Pikovsky A. A computational analysis of linear-price iterative combinatorial auctions. *Inf Syst Res* 2009; 20(1):33–59.
 48. Jacob KG, Charles A, Holt. Hierarchical package bidding: A paper & pencil combinatorial auction. *Games and Economic Behavior* 2008. DOI: 10.1016/j.geb.2008.02.013
 49. Ausubel L, Crampton P, Milgrom P. The clock-proxy auction: a practical combinatorial auction design. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006.
 50. Kelly F, Steinberg R. A combinatorial auction with multiple winners for universal service. *Manage Sci* 2000;46(4):586–596.
 51. Adomavicius D, Gupta A. Towards comprehensive real-time bidder support in iterative combinatorial auctions. *Inf Syst Res (ISR)* 2005;16:169–185.
 52. Scheffel T, Pikovsky A, Bichler M, *et al.* An experimental comparison of linear and non-linear price combinatorial auctions. *Inf Syst Res* 2010. DOI: 10.1287/isre.1090.0267.

COMBINATORIAL AUCTIONS

DAVID PORTER
STEPHEN RASSENTI
Economic Science Institute,
Chapman University,
Orange, California

INTRODUCTION

A combinatorial auction is a resource allocation process that can be implemented when multiple resources must be simultaneously allocated amongst competing users, and information concerning the values of the various possible uses and constraints affecting those uses is decentralized. For example, the demands for landing and takeoff “slots”¹ at LaGuardia airport in New York outstrip the maximum 45 arrivals per hour servicing constraint. To ensure safety at LaGuardia and to control demand, the Port Authority of New York requires that airlines have in their possession a designated takeoff or landing slot that they have issued prior to time of use. Now imagine that congested airspace becomes a simultaneous reality at many major airports (various reports indicate this day will soon be upon us), and the slot-based regime becomes enforced by all airport authorities. In order to determine the *efficient* (value maximizing) allocation of landing and takeoff slots to competing airlines across the country, the “allocation center” would need to solve the following set packing problem with added logistical constraints²

Maximize : $Z = \sum_j \pi_j x_j \rightarrow$ Maximize the
value of the allocated slots (1)

¹A takeoff/landing slot is usually thought of as a 15 min interval during which a maximum number of airplanes can safely takeoff and/or land.

²One can think of the vector $\langle \pi_j, a_{ij}, d_{kj}, e_k \rangle$ as a bid submitted by an airline.

Subject to : $\sum_j a_{ij} x_j \leq b_i \quad \forall i \rightarrow$ Maximum
number of slots available
imposed by airports (1a)

$\sum_j d_{kj} x_j \leq e_k \quad \forall k \rightarrow$ Logical constraint
on packages imposed airlines
(1b)

$x_j \in \{0, 1\} \rightarrow$ Either the whole package
is accepted or not.

Where:

- $i = 1, \dots, N$, subscripts some particular
takeoff or landing slot at some
particular airport;
- $j = 1, \dots, J$, subscripts some package of
slots that has value for some
particular airline³;
- $k = 1, \dots, K$, subscripts some particular
logistical constraint imposed on a set
of packages by some airline or by the
optimizing center⁴;

³An airline is always interested in combinations (packages) of slots to optimize their own flight resources. A landing slot without a corresponding takeoff slot at a particular airport would be of very little value. Furthermore, a particular aircraft is most efficiently used when it cycles through a series of takeoffs and landings in a given day or set of days. If slots were not allowed to be sold in useful combinations but had to be procured piecemeal (slot by slot and airport by airport), then an airline would face the considerable financial exposure of paying too much for a noncomplementary sets of slots (see Bykowsky *et al.* [1] for more on financial exposure in noncombinatorial auctions).

⁴For example, suppose an airline has one airplane and a value π_z for using it with the cycle of slots in package z , and π_y for using it with the cycle of slots in package y which occurs 30 min later. Then the airline would specify its interest in obtaining package z or y but not both, by submitting the logical constraint $x_z + x_y \leq 1$ to the center. Similarly, the center might wish to impose constraints on allocations to potential users: for example, user u should not be allocated the rights to any more than s slots at airport a .

$$a_{ij} = \begin{cases} 1 & \text{if package } j \text{ includes slot } i \\ 0 & \text{otherwise;} \end{cases}$$

b_i = the number of planes that can takeoff (land) in slot i at some airport;

$d_{kj} \in \{-1, 0, 1\}$, if package j appears in logistical constraint k ⁵;

e_k = an integer;

π_j = the value of obtaining slot package j to the airline that seeks it.

While this is a difficult (but manageable) discrete programming problem, the real problem is that π_j (the value of package j to the airline that seeks it), a_{ij} (whether a particular slot is included in package j), and d_{kj} (whether package j is part of some logistical constraint k) are all parameters that are not known to the “allocation center” but privately known to the competing airlines. That is, the center needs to solve a resource allocation problem where information on the coefficients in the objective function and constraints is decentralized and unknown, except for what may be revealed voluntarily by the potential users during the auction process.

AUCTION DESIGN

Each auction process requires users to supply information to the “optimizing center,” in order to be included in the final allocation. The incentive problem in mechanism design⁶ refers to the process in determining who is allocated which resources at specific prices so that the quality of information provided by users can be used by the center to obtain a desired outcome. The most famous of such mechanisms is the Vickrey auction [3]. This sealed-bid auction uses a

particular pricing scheme to provide users the incentive to reveal truthful information. While Vickrey did not develop his original auction for package bids, his basic principles are easily extended to cover many cases where users would compete for packages of resources subject to user and/or system specific constraints.⁷

For example, in Maximization (1), suppose airline $\mathcal{L}$ submits a vector of bids for L different packages of slots. (Without loss of generality, assume the first L of all J bids recorded by the system are those belonging to airline $\mathcal{L}$.) The Vickrey auction has two parts. First, using all J submitted sealed bids and any constraints imposed on the bids by the airlines or the center, we solve Maximization (1) above and obtain the maximum revenue Z^* and the corresponding allocation $\mathbf{x}^*$ of slot packages to airlines. The next step, crucial in soliciting truthful bid information, is to determine the total price for each auction winner in an incentive compatible manner. The total Vickrey price for all packages j won by airline $\mathcal{L}$ ($j \in \{1, \dots, L\}$) is found by performing the following calculations:

- Remove ALL of airline $\mathcal{L}$ ’s bids and constraints from the system and recalculate the solution to Maximization (1) using all remaining bids and constraints but not $\mathcal{L}$ ’s. Let $Z \sim$ denote the optimal value of the bids without $\mathcal{L}$ ($\sim \mathcal{L}$), and let $\mathbf{x} \sim$ denote the allocation using $\sim \mathcal{L}$.
- Airline $\mathcal{L}$ then pays the following total price $p_{\mathcal{L}}$ for all its winning bids:

$$\begin{aligned} p_{\mathcal{L}} &= \sum_{j=1, L} \pi_j x_j^* - [Z^* - Z \sim] \\ &= Z \sim - \sum_{j=L+1, J} \pi_j x_j^*. \end{aligned} \quad (2)$$

There are two interesting things to note: first, that the total paid for the airline $\mathcal{L}$ ’s winning bids is always less than or equal to the total amount that it bid ($\sum_{j=1, L} \pi_j x_j^*$); and second, that the airline’s bids do not

⁵As an example of a logical constraint suppose we have $e_k = 2$ and $d_{k1} = -1$; $d_{21} = 1$; $d_{31} = 1$ for some k so that we have a constraint $-x_1 + x_2 + x_3 \leq 1$. This means that a necessary condition for both packages 2 and 3 to be part of the solution is that package 1 must also be part of the solution; otherwise, only one of packages 2 or 3 can be included.

⁶See Myerson [2] for more on mechanism design.

⁷See Forsythe and Isaac [4] for an example. In addition, Jackson [5] developed a Vickrey auction for the vertex picking problem, a combinatorial problem.

determine its payment (the RHS of the price calculation involves only the bids $j = L + 1, \dots, J$). Because of these properties, the incentive structure of this auction provides a dominant strategy for the airline to truthfully reveal all its package value information.

At first glance it seems then that we need not proceed any further: the Vickrey auction appears to be the panacea for complex resource allocation problems with decentralized value information. But, though the Vickrey auction has this nice revelation property, it is plagued by several other issues that can cause it to be shunned for field implementation. For this introduction, we discuss a couple of specific shortcomings.⁸ First, it should be noted that Maximization (1) is an NP-complete problem. This is known in the auction literature as the winner-determination problem or, as Michael Rothkopf calls it, the 2^N bogeyman, where N is the number of different resources available for allocation in Maximization (1).⁹ In addition, each price determination is also a 2^N problem. Furthermore, each potential user must determine how many of the 2^N possible bids it should submit.¹⁰ Second, Vickrey prices are not linear in the sense that most users would be familiar with, where each slot (resource) would have its own price. Prices normally perform the very important task in a resource constrained economy of signaling scarcity and value. No price discovery is provided in the Vickrey auction. In the experimental economics literature, it is well documented that auctions that provide feedback and price adjustment, such as the English

auction, perform better than sealed-bid auctions that have no feedback mechanism.¹¹ Third (and importantly!), the Vickrey prices can often bring the center very low amounts of revenue.¹² Fourth, the winning allocations are nonmonotonic in the sense that adding more competition to the Vickrey auction can actually reduce prices and revenue.¹³ And finally, the Vickrey auction does not handle budget constraints well. They require the submission of some large set of side constraints, and not being able to express the true value for a package if its value exceeds a bidder's auction budget can also eradicate the dominant strategy to reveal the truth concerning the value of all packages.

With these shortcomings of the Vickrey auction in mind, we investigate the evolving design of combinatorial auctions that developed to be better suited for field implementation when human agents are bid competitively for the rights to use scarce resources.

COMBINATORIAL AUCTION DESIGN

Finding Slot Prices

The term combinatorial auction was coined by Rassenti, Smith, and Bulfin (RSB), when they first designed and tested an auction for takeoff and landing slots by allowing participants to bid for packages of slots [11].¹⁴ If Maximization (1) were a linear programming

⁸For more details on the other issues with the Vickrey auction see Ausubel and Milgrom [6] and Rothkopf *et al.* [7].

⁹The 2^N is a worst-case scenario. There are useful approximation algorithms that could be used for winner-determination [8]. How such approximating algorithms would perform in a strategic setting is yet to be determined.

¹⁰Practitioners well know that explaining Vickrey pricing rules and allocations to bidders is difficult (Parkes [9] calls this feature *transparency* of the auction design).

¹¹For example, see Smith *et al.* [10].

¹²For example, suppose there are two items to be allocated A and B and User1 is willing to pay \$10 for item A, User2 is willing to pay \$10 for item B, and User3 is willing to pay \$11 for both items A and B. Then the winners would be User1 and User2, and according to Equation (2), they would each pay a Vickrey price of \$1.

¹³Even more troublesome in the example of the previous footnote is that if User1 reduced his willingness to pay for A, say from \$10 to \$4, then User1 would still pay \$1 but User2 would pay \$7, increasing the center's revenue.

¹⁴For more detailed information on combinatorial auctions, see Vries and Vohra [12].

problem, then the dual of Maximization (1) would provide shadow prices for slots. However, integer programs do not normally allow the dual program to find the linear shadow prices for each resource, which completely separate winning from losing bids.¹⁵ However, in keeping with the real-world norm for implementing processes that provide bidders and the center price discovery during resource allocation processes, RSB were interested in devising a way to price individual slots that provided users feedback on their scarcity, and simultaneously provided airports a method to divide the income from the auction. The RSB process allowed airlines to submit package bids of the form as described in Maximization (1), and then used these bids to find the revenue maximizing allocation $(Z^*, \mathbf{x}^*)$. Let $\mathcal{A} = \{j | x_j^* = 1\}$ be the set of accepted bids, and $\mathcal{R} = \{j | x_j^* = 0\}$ be the set of rejected bids given by this solution.

The pricing algorithm implemented required solving a complementary set of pseudo-dual linear programs. Program D_R defined the set of linear prices w_i one for each slot i , such that if $\pi_j > \sum_i w_i a_{ij} x_j$ then the package j was definitely accepted ($j \in \mathcal{A}$):

$$\begin{aligned} & \text{Minimize : } \sum_R y_r \\ & \text{Subject to : } \sum_i w_i a_{ij} \leq \pi_j \quad \forall j \in \mathcal{A} \\ & \quad y_r \geq \pi_r - \sum_i w_i a_{ir} \quad \forall r \in \mathcal{R} \quad (D_R) \\ & \quad y_r \geq 0, w_i \geq 0. \end{aligned}$$

Program D_A defines the set of prices v_i one for each slot i , such that if $\pi_r < \sum_i v_i a_{ir} x_r$ then the package r was definitely rejected ($r \in \mathcal{R}$):

$$\begin{aligned} & \text{Minimize : } \sum_A y_j \\ & \text{Subject to : } \sum_i v_i a_{ir} \leq \pi_r \quad \forall r \in \mathcal{R} \\ & \quad y_j \geq \sum_i v_i a_{ij} - \pi_j \quad \forall j \in \mathcal{A} \quad (D_A) \\ & \quad y_j \geq 0, v_i \geq 0. \end{aligned}$$

¹⁵There are infrequent exceptions which occur when the linear solution to Maximization (1) happens to correspond to an integer solution and a competitive equilibrium exists with separating prices.

RSB used these prices as if they were competitive equilibrium prices, so that for $\pi_j \geq \sum_i w_i a_{ij} x_j$ the winning bidder would pay the w_i prices for the individual slots in the package j . If package k was in the solution to Maximization (1) but $\sum_i w_i a_{ik} x_k \geq \pi_k \geq \sum_i v_i a_{ij} x_j$, then the winning bidder would pay exactly his bid of π_k . RSB tested this simple combinatorial auction process in a laboratory setting across repeated rounds of bidding, and found that it had reasonable incentive properties and very high levels of allocative efficiency when compared to independent auctions for each slot which later required the participants to assemble their own packages. Unfortunately, this is a sealed-bid auction, so there can only be feedback across auctions, and there is no price discovery within an auction. In addition, the 2^N bogeyman was not directly confronted in this design as participants had value for a very limited set of packages and budget constraints were not an issue.

Auctions with Feedback

Banks *et al.* [13] try to solve Maximization (1) using a combinatorial English auction.¹⁶ This auction format allows participants to individually place public bids for packages of items. When a bid is submitted, the center checks to see if the bid can be added to the standing allocation; if not, it goes into a queue where other bidders can view and combine with it to displace standing bids. Table 1 provides an example of this process.

Notice that there is no computational problem for the center here. The entire computational burden is placed on the bidders to find combinations that can displace standing bids. The main feature of this design is the feedback and open package price discovery during the bidding process.

¹⁶The typical English auction is an open outcry process in which the auctioneer accepts increasingly higher bids “from the floor” delivered by participants. The highest bidder at any given moment defines the standing bid, which can only be displaced by a higher bid from another participant. If no participant betters the standing bid within a fixed time, the standing bid becomes the winning bid.

Table 1. Combinatorial English Auction

<i>Standing Bids (Contract 1)</i>			
Bidder	Item X	Item Y	Bid(\$)
1	15	2	50
4	1	10	75
5	4	8	100
	20	20	
↓			
<i>Standby Queue</i>			
Bidder	Item X	Item Y	Bid(\$)
8	5	4	40
↓			
<i>New Potential Bid</i>			
Bidder	Item X	Item Y	Bid(\$)
6	11	8	120
↓			
<i>Combined 8+6 Value</i>			
Bidder	Item X	Item Y	Bid(\$)
8	5	4	40
6	11	8	120
	16	12	160
	20	20	
⇒			
<i>New Standing Bids (Contract 2)</i>			
Bidder	Item X	Item Y	Bid(\$)
8	5	4	40
6	11	8	120
5	4	8	100

The system has 20 units of X and 20 units of Y available for sale. Currently bidder 1 has a bid for 15 units of X and 2 units of Y , and is willing to pay \$50 for that package. Bidders 4 and 5 also have standing bids for packages. The current bids exhaust the available resources. Bidder 8 submits a bid for 5 units of X and 4 units of Y for \$40. This cannot displace any of the standing bids alone unless Bidder 8 raises his bid to more than \$125. Thus, his bid is sent to the standby queue. Bidder 6 sees the Bidder 8 bid and tenders a bid for 11 units of X and 8 units of Y for which he is willing to pay \$120. Together, Bidders 6 and 8 displace Bidders 1 and 4, since their combined bid is higher and fits within the released resources. The new set of bids now become the standing bids.

However, this auction has high cognitive participation costs. More importantly, there is no individual price information to guide decision making. There is no transparency provided to the bidder concerning the minimum amount they need to bid to be included in the optimal allocation. That is, each tentative bidder needs to solve their own 2^N problem in order to understand how to combine properly with existing bids in both the queue and the standing contract.

Subsequent to this design, many other iterative auctions sprang up responding to the need to examine such auction processes in preparation for the allocation of the radio spectrum by the Federal Communications Commission (FCC). This seemed to be an excellent potential use for a combinatorial auction. Indeed, the interest in combinatorial

auctions has mushroomed since the late 1990s as auctions have become the preferred method for the allocation of spectrum around the world. Several extensions of the English auction for package bids have been developed.¹⁷

One auction that reduces the computational burden on bidders but provides feedback is the iBundle mechanism developed by Parkes [15]. This auction proceeds in rounds $t = 1, 2, \dots$, where in each round participants are permitted to submit bids that have an exclusive OR logic: that is, only one of the submitted bids can be accepted in the final allocation. iBundle keeps track of all

¹⁷See Cramton *et al.* [12] and de Vries and Vohra [14] for examples.

submitted bids and the highest bid submitted on a package becomes the ask price for that package. For example, Maximization (1) is solved using the bids, where the bids of each participant are OR bids. The solution to Maximization (1) then becomes the provisional allocation to be bettered. The “ask prices” for round $t + 1$ are the best bids on each package from round t plus an increment. Bids for packages in round $t + 1$ are considered competitive if they meet the ask price. The auction ends when each competitive bidder receives one package (this could be the null package). The basic feature of iBundle, and a typical feature of most iterative mechanisms, is that there is feedback on what makes a particular package bid competitive during the auction process. However, there remains a 2^N optimization problem to solve each round, and a large cognitive load to determine whether my bid might be a winner in the next round when the next optimization is solved.

Auctions with Item Price Feedback

In an attempt to select prices for each item to help guide bidders in an iterative combinatorial auction, Kwasnica *et al.* [16] devised a set prices, one for each item, that determine whether a bid submitted in round t is acceptable. This auction is called the *resource allocation design* or *RAD*. Using prices for the elemental goods, it becomes easy for a bidder to determine the minimum bid required for a package to be acceptable by simply adding the prices of each item in the package. Returning to Maximization (1), the task is to find a set of prices v_i one for each slot $i = 1, \dots, m$ that can guide bidders and move the auction.¹⁸ RAD was motivated as an attempt to improve the noncombinatorial auction process used by the FCC to auction spectrum.¹⁹

RAD proposed three properties that prices must satisfy if bidders were to pay

them: (i) all accepted bids would receive price signals with aggregate costs totaling less than or equal to what they bid; (ii) all rejected bids would receive price signals that resulted in aggregate costs totaling higher than what they bid; and (iii) new bids willing to pay more than the aggregate price of the package should have a good opportunity to become provisionally winning (that is, prices ought to “guide” new bids to packages that will increase revenues). As discussed in RSB, (i) and (ii) are generally impossible to solve simultaneously, so these guidelines can only be executed in an approximate manner.

RAD proceeds in rounds $t = 1, 2, \dots$. In round t , Maximization (1) is solved using the round t submitted bids to generate the current optimal allocation $(^tZ^*, ^tx^*)$ with the set of winning bids $^t\mathcal{W} = \{j | ^tx_j^* = 1\}$ and the set of losing bids $^t\mathcal{L} = \{j | ^tx_j^* = 0\}$. RAD prices tv are then found by solving the following linear program:

$$\begin{aligned} \text{Minimize: } & D(D, ^tv, \mathbf{g}) \\ \text{Subject to: } & \sum_i ^tv_i a_{ij} = ^t\pi_j \quad \forall j \in ^t\mathcal{W} \\ & \sum_i ^tv_i a_{ij} \geq ^t\pi_j - g_j \quad \forall j \in ^t\mathcal{L} \quad (5) \\ & 0 \leq g_j \leq D \quad \forall j \in ^t\mathcal{L} \\ & ^tv_i \geq 0 \quad \forall i. \end{aligned}$$

The solution to program (5) ensures that at prices tv_i the winner of any package j pays his bid, and the variables g_j are selected so that all losing bids are less than their corresponding package prices with the maximum price distortion (D) minimized.²⁰ RAD does not ensure unambiguous price signals (a losing bid can exceed the sum of the prices of items of that package), and the winner-determination problem still looms in each round.

Clock Auctions and Item Price Feedback

One deficiency with iterative auctions in which participants select a bid π_j for their package j is that they do not know how to

¹⁸Recall that Rassenti *et al.* found a pseudo-dual envelope of prices for a single round sealed-bid. These prices could also be used in an iterative auction.

¹⁹See Milgrom [17] for a description of the basic FCC auction design.

²⁰The auction ends when there is no bid greater than the costs defined by the implicit prices.

refine their bids in order to not overshoot competitive prices. McCabe *et al.* [18] found “jump bidding behavior” in their attempts to test Vickrey’s proposal to use English auctions for multiple units. They found that allowing bidders to announce bid prices from the floor is not a good design feature in multiple unit auction environments as it can lead impatient bidders to overshoot competitive prices or aggressive bidders to deliver implicit signals to competitors. One way to eliminate this “strategic” behavior in the bid message space is to use clocks that move price upward automatically based on excess demand for each resource. With a clock-controlled resource price, the bidders need to announce only the quantity of the resource they wish to obtain at the current price. The bidder’s decision is relegated to a much simpler task, and the scope for strategic behavior is dampened significantly. Porter, Rassenti, Roopnarine and Smith (PRRS) extend the notion of a clock auction to a combinatorial environment and report experimental results with high levels of performance [19]. Their idea is simple. Returning to Maximization (1), for each slot $i = 1, \dots, m$ a price ${}^t v_i$ is posted in round $t = 1, 2, \dots$. At the stated clock prices, the participants submit the set of package bids they would be willing to pay for and attach any corresponding constraints (e.g., I’ll pay for package f or g but not both). The center keeps this information on record for each round t , and proceeds as follows:

1. Calculate the current excess demand ${}^t q_i$ for each item i .²¹
2. If ${}^t q_i > b_i$, set ${}^{t+1} v_i \leftarrow {}^t v_i + {}^t \varepsilon_i$; otherwise, set ${}^{t+1} v_i \leftarrow {}^t v_i$.

²¹PRRS counts a bidder’s overlapping packages as contributing multiple units of demand on the overlapped resources. This tends to lead bidders to avoid competition with themselves through round by round submission of a single large package aggregating their most valuable cover, but bidders must remain cognizant of the fact that providing the center discrete alternatives (often overlapping) increases the probability of matching with other bidders when a bidder cannot afford to outbid all others for a large package of resources.

The clock ticks up the price of resource i by ${}^t \varepsilon_i$ in round $t + 1$ if all bidders in aggregate demand more units of resource i than the center has available (b_i) to allocate. The price increment ${}^t \varepsilon_i$ need not be the same for all resources and it need not remain the same during the auction process for any given resource. In the PRRS auction, no information concerning excess demand or how the increments ${}^t \varepsilon_i$ are chosen is provided to the bidders: this makes tacit collusion a very difficult sport. Notice also that there is no winner-determination algorithm required to be run during each round of the auction process. The cognitive costs for the bidder are low and the process is intuitive: participants need only to understand their own private values for packages of the resources being offered. They do not need to estimate what it might take for a package bid to be acceptable: the linear price information is unambiguous. As the rounds proceed and the clock prices tick upward, a bidder can peruse the new clock prices and submit any new vector of package bids and constraints (unconstrained by previous round’s bids) along with stating which of his old rounds’ bids he wishes to keep active in the system at their former clock prices. The auction process continues until round r , when there is no explicit excess demand for any resource being offered (${}^r q_i \leq b_i, \forall i$). The clock auction now enters its close-out phase.

If in round r the explicit demand at the current clock prices exactly equals the supply of all resources (${}^r q_i \leq b_i, \forall i$), then the auction has drawn to a successful conclusion, and all resources are awarded at the current clock prices without ever having to solve the winner-determination problem. If at least one of the resources is undersubscribed at the final prices ($\exists i \text{ } {}^r q_i < b_i$), then a winner-determination algorithm must be conducted using all of the nondominated bids intentionally left by bidders in the system up to round r to generate the revenue maximizing allocation ${}^r(Z^*, x^*)$. If this solution to the winner-determination algorithm discards any package j (i.e., ${}^r x_j^* = 0$), which was a standing package at the round r prices, then there is implicit competition for the resources in the discarded package and the price of at least one of those resources in that package

must be increased. For example, suppose that we have three single items for sale (A, B, and C) and three bidders: the first wants to win

only package AB, the second only package BC, and the third only wants item A. The bidding proceeds as indicated:

Round	Item Prices			Package Bids Submitted			Aggregate Item Demands		
	A	B	C	{AB}	{BC}	{A}	A	B	C
1	100	100	100	✓	✓	✓	2	2	1
2	120	120	100	✓		✓	2	1	0
3	140	120	100	✓			1	1	0
4	160	140	100				0	0	0

Notice in round three there is only one standing bid remaining, {AB} at a total price of 260, but because there are previous bids, {BC} at 200 from round 1 and {A} at 120 from round 2, that can be combined for more revenue to displace the standing {AB} bid, the auctioneer raises the prices on items A and B because of the implicit competition. Notice in round four {AB} quits and {BC} and {A} would be declared the auction winners.

In fact, there may be more than one standing package discarded and more than one resource i , which requires price increases during the close-out phase. The center sets $v_i^{r+1} \leftarrow v_i + \epsilon_i$ for all resources i , for which there is implicit competition, and the auction process continues to the next round $r + 1$. Bidders are again asked to respond to the new prices just as in previous rounds. From the bidder's perspective, it is impossible to tell whether the center is in the close-out phase or not! This process continues until no standing package bids are discarded in the latest winner determination. The final allocation always includes all standing bids at the final clock prices, but may also include old bids from previous rounds that fill demand gaps. Bids from previous rounds that make into the final allocation are charged at the prices from the round from which they were retrieved. This means that prices are non-linear in nature. The final clock prices are akin to the D_R prices in RSB, at which price level no package is rejected. The packages retrieved from previous rounds are akin to those sold exactly at their bid prices below the D_R but above the D_A prices in RSB.

Notice that while the computational burden of the auction has been reduced significantly during the process, there remains the possibility of being required to solve sequentially for winner determination during the close-out phase. However, by the time the close-out phase arrives, much preprocessing can be accomplished by the center to help prune the decision tree and reduce the computational burden of winner determination. Constraints placed on each rounds' package bids, including budget constraints across bids, are considered only when solving for winner determination during the close-out phase.

The simple and natural price discovery format, maintained privacy with regard to each bidder's maximal willingness to pay, and the fact that no demand information is passed on to auction participants (including when the close-out phase is executing) are the apparent keys to the success of PRRS auction.

Another variant of the combinatorial clock auction, the clock-proxy auction, has been proposed by Ausubel, Cramton, and Milgrom (ACM) [20]. This auction has a second stage (in addition to an initial clock phase), in which mutually exclusive package bids from each bidder is submitted through a proxy agent preinformed by the bidder concerning the maximum that it should bid for each package. The clock phase of ACM has several differences from PRRS: agents are privy to excess demand information; price increments are calculated in a predetermined linear manner; there is no close-out phase to resolve implicit demand amongst old bids

for unallocated resources; and there is an activity rule which regulates what you can bid on in any later round given what you bid on in previous rounds. ACM claim that without an activity rule strategic insincere bidding will plague a clock auction; however, PRRS report its clock auction performs well without an activity rule given that bidders are not privy to the demand in the system at any point during the auction and they are uncertain concerning how the close-out phase will play out (whose bids will be retrieved).

The proxy stage of the Clock-Proxy auction is similar to holding a sealed-bid auction after the initial clock auction stage, with all clock stage bids remaining binding. In our previous discussion of the Vickrey auction and its failures, we spent no time discussing issues of implicit collusion by bidders. ACM concludes that a clock auction by itself would do well in certain competitive environments, but that the proxy phase helps prevent tacit collusion in the form of premature demand reduction during their clock stage. However three features of PRRS, a pure clock auction, should make it difficult for bidders to tacitly collude

- no public provision of demand information (who's in for which items at each round);
- no explicit price increment rules need to be provided (increments are the auctioneer's prerogative);
- no indication of having reached the close-out phase (at any round competition for items might be explicit due to excess demand from current bidders or implicit due to former bidders displacing current bidders).

Whether the ACM and PRRS auctions are broadly robust against tacitly collusive activity remains to be thoroughly tested.

Some Design Detail Issues

1. *Computation.* We have discussed the 2^N bogeyman associated with the winner-determination problem. This issue has sparked many papers intended on solving this problem. Some

papers have discussed limiting the type/set of bids that can be submitted so that the computation becomes relatively straight-forward [20,21]. Others have devised clever algorithms (including "simple" greedy methods) to lessen the computational issues for large problems (see Sandholm [22] for a review of these). Lastly, some researchers are examining noncomputational methods to reduce the computational burden [23].

2. *Speed.* The clock auctions of PRRS and ACM are both dependent upon choosing parameters that determine the speed of price changes as the auction progresses. Clearly, if the price increments are larger the auction moves faster, but there is a danger of overshooting the limit prices of bidders and reducing the efficiency of the resulting allocation. The relationship between overaggressive price augmentation and allocative inefficiency is not well studied. PRRS offers little guidance to choosing optimal price increments, but indicates that high levels of efficiency can be achieved without much concern in the 10 bidder/10 resource environments tested. ACM discusses one method for addressing sudden package withdrawals due to gross changes in price by allowing bidders to submit intra-round bids indicating intermediate limit price withdrawal points, but this process can become cumbersome in environments where there are a large number of resources to be allocated. For the purpose of improving actual field implementations, research in this area would probably be most valuable. It is likely that in addition to the version of the clock auction being implemented, the structural features of the delivered bids should be taken into account in optimally determining price increments.

REFERENCES

1. Bykowsky M, Cull R, Ledyard J. Mutually destructive bidding: the Federal Communications

- Commission auction design problem. *J Regul Econ* 2000;17(3):205–228.
2. Myerson R. Mechanism design. The New Palgrave Dictionary of Economics Online. New York: Palgrave Macmillan; 2008.
 3. Vickrey W. Counterspeculation, auctions, and competitive sealed tenders. *J Finan* 1961;16(1):8–37.
 4. Forsythe R, Isaac R. Demand-revealing mechanisms for private good auctions. In: Smith VL, editor. Volume 2, Research in experimental economics. Greenwich: JAI Press, Inc.; 1982.
 5. Jackson C. Technology for spectrum markets [PhD dissertation]. MIT; 1976.
 6. Ausubel LM, Milgrom P. The lovely but lonely Vickrey auction. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006. Chapter 1.
 7. Rothkopf MH, Teisberg TJ, Edward PK. Why are Vickrey auctions rare. *J Polit Econ* 1990;98(1):94–109.
 8. Dobzinski S, Schapira M. An improved approximation algorithm for combinatorial auctions with submodular bidders. Proceedings of the 17th annual ACM-SIAM Symposium on Discrete Algorithms; Miami (FL). 2006. pp. 1064–1073.
 9. Parkes D. Iterative combinatorial auctions. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006. Chapter 2.
 10. Smith VL, Williams A, Bratton WK, *et al.* Competitive market institutions: double auctions vs. sealed bid-offer auctions. *Am Econ Rev* 1982;72:58–77.
 11. Rassenti SJ, Smith VL, Bulfin RL. A combinatorial auction mechanism for airport time slot allocation. *Bell J Econ* 1982;13:402–417.
 12. de Vries S, Vohra R. Combinatorial auctions: a survey. *INFORMS J Comput* 2003; 15:284–309.
 13. Banks J, Ledyard J, Porter D. Allocating uncertain and unresponsive resources: an experimental approach. *Rand J Econ* 1989; 20(1):1–25.
 14. Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006.
 15. Parkes. 2001.
 16. Kwasnica A, Ledyard J, Porter D, *et al.* A new and improved design for multiobject iterative auctions. *Manage Sci* 2005;51(3): 419–434.
 17. Milgrom P. Putting auction theory to work: simultaneous ascending auction. *J Polit Econ* 2000.
 18. McCabe K, Rassenti S, Smith V. Testing Vickrey's and other simultaneous multiple unit versions of the English auction. In: Isaac RM, editor. Volume 4, Research in experimental economics. Greenwich (CT): JAI; 1988.
 19. Porter D, Rassenti S, Roopnarine A, *et al.* Combinatorial auction design. *Proc Nat Acad Sci* 2003;100:11153–11157.
 20. Rothkopf M, Pekec A, Harstad R. Computationally manageable combinational auctions. *Manage Sci* 1998;44:1131–1147.
 21. Goeree J, Holt C. Hierarchical package bidding: a 'paper & pencil' combinatorial auction. *Games Econ Behav* 2007;70(1): 146–169.
 22. Sandholm T. Optimal winner determination algorithms. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006. Chapter 14.
 23. Pekec A, Rothkopf M. Noncomputational approaches to mitigating computational problems in combinatorial auctions. In: Cramton P, Shoham Y, Steinberg R, editors. Combinatorial auctions. Cambridge (MA): MIT Press; 2006. Chapter 16.

COMBINATORIAL TRAVELING SALESMAN PROBLEM ALGORITHMS

CLAUDIA D'AMBROSIO
ANDREA LODI
SILVANO MARTELLO
DEIS, Università di Bologna,
Bologna, Italy

INTRODUCTION AND NOTATION

Given a set of cities, and the cost of traveling between each pair of them, the *traveling salesman problem* (TSP), calls for finding a unique tour visiting all cities at minimum cost. Besides being for sure the best known combinatorial optimization problem, the TSP has many applications, not only in operations research/management science (especially in transportation and logistics), but also in many other fields, such as genome sequencing or drilling of printed circuits boards. In fact, there are no fewer than four books devoted to it, see Lawler *et al.* [1], Reinelt [2], Gutin and Punnen [3], and Applegate *et al.* [4].

Throughout this article, we distinguish between the *symmetric* TSP (STSP), in which the cost of traveling from city A to city B is the same as the cost of traveling in the reverse direction, and the *asymmetric* TSP (ATSP), in which these costs are permitted to be different. The former version can be modeled by a *graph* $G = (V, E)$, where $V = \{1, \dots, n\}$ is a set of *vertices* and E is a set of unordered pairs $[i, j]$ of vertices, called *edges* (with $|E| = m$). The latter version can be modeled by a *digraph* (directed graph) $G = (V, A)$, where A is a set of ordered pairs (i, j) of vertices, called *arcs* (with $|A| = m$).

All graphs and digraphs considered in this article are *simple*, that is, they have no loops and no two of their edges or arcs connect the same pair of vertices (in the same direction, in the case of digraphs).

The *cost* of traveling along edge $e = [i, j]$ (in either direction), or along arc $a = (i, j)$ (from i to j), will be denoted by c_{ij} , or, alternatively, by c_e and c_a , respectively. For a graph $G = (V, E)$, we will denote by $\delta(v)$ ($v \in V$) the set of edges having v as one end vertex. For a digraph $G = (V, A)$, we will denote by $\delta^+(v)$ (and $\delta^-(v)$, respectively) the set of arcs leaving (and entering, respectively) vertex v .

The TSP is $\mathcal{NP}$ -hard in the strong sense, and very difficult to solve in practice. Nevertheless, and surprisingly, large-scale instances arising in practice can often be solved to proven optimality (or near-optimality) by modern solvers. Currently, the most successful technique is the so-called *branch-and-cut* (see **Branch and Cut**), a general technique based on the solution of the linear programming (LP) relaxation, which was indeed developed in the context of the TSP [5,6] (see also **Mathematical Programming Approaches to the Traveling Salesman Problem**).

There are, however, a number of solution approaches alternative to branch-and-cut, most of which are based on graph-theoretic relaxations rather than linear programming. We refer to such approaches as *combinatorial* TSP algorithms. They are relevant both from theoretical and historical points of view, and because they are competitive for the solution of the ATSP.

In the next section, we review some of the TSP ancestors, including the famous Hamiltonian cycle problem (HC), of which the TSP is the weighted version. In the section titled "Formulations," we review the most famous formulations for both symmetric and asymmetric TSP. Then, we consider in the section titled "Algorithms for the Symmetric TSP" combinatorial approaches for the symmetric TSP, whereas the algorithms for the asymmetric TSP are presented in the section titled "Algorithms for the Asymmetric TSP." Finally, in the section titled "Software," we briefly discuss the availability of TSP software and we give some useful pointers.

THE ORIGINS AND THE HAMILTONIAN CYCLE PROBLEM

The origins of the TSP date back to the eighteenth century, when Leonard Euler studied the *knight's tour problem*: a knight is placed on an empty $n \times n$ chessboard and must visit each square exactly once by only using valid chess moves of a knight. The tour is called *closed* if the knight ends on a square from which there is a legal move to the starting square, or *open* otherwise. In a study presented in 1759 (but published only 7 years later), Euler [7] proposed a number of possible solutions to the problem for the classical case $n = 8$.

By defining a graph in which the vertices correspond to the chessboard squares and the edges to the legal knight moves, an open (and closed, respectively) knight's tour corresponds to a path (and a cycle, respectively) that visits every vertex of the graph exactly once. Such paths and cycles are called *Hamiltonian* in honor of Sir William Rowan Hamilton, a great Irish mathematician of the nineteenth century, famous for the invention of hypercomplex numbers known as *quaternions*.

In 1859, Hamilton invented a game that was played on a wooden planar representation of the edges of a dodecahedron, with holes at each of the 20 vertices: The first player stuck 5 pegs in any consecutive vertices, and the second player was requested to stick the remaining 15 pegs so as to complete the resulting path to a cycle visiting each vertex exactly once. The idea of the game, called the *Icosian game*, was sold for £25 to a Dublin toy manufacturer. (It seems that the sales were not satisfactory though.) Many details on the life and achievements of Hamilton can be found in the page maintained by the Hamilton Mathematics Institute, Trinity College Dublin at <http://www.hamilton.tcd.ie/>.

The problem of deciding if a graph or a digraph possesses a cycle that passes through all vertices exactly once is thus known as the Hamiltonian Cycle problem. It is a generalization of the knight's tour problem discussed above, and the special case of the TSP that arises when all edges of E , or all arcs of A ,

have unit cost (while infinite cost is assumed for traveling between vertices not connected by an edge or arc). The HC is one of the first problems proved to be $\mathcal{NP}$ -complete: its two versions (for graphs and digraphs) appear indeed in the list of 21 problems for which Karp [8] proved $\mathcal{NP}$ -completeness in his famous 1972 paper. In the second half of the twentieth century, an impressive amount of results has been produced on this problem.

An important stream of research consists in finding conditions ensuring that if a graph has a sufficient number of edges then it is *Hamiltonian*, that is, it possesses at least one HC. All results below refer to graphs in which the number n of vertices is at least 3. The first seminal result was obtained by G.A. Dirac [9] in 1952, and reads

A graph $G = (V, E)$ such that $|\delta(i)| \geq n/2$ for every $i \in V$ is Hamiltonian.

The result was strengthened by Ore [10] in 1960:

A graph $G = (V, E)$ such that $|\delta(i)| + |\delta(j)| \geq n$ for every pair of nonadjacent vertices $i, j \in V$ is Hamiltonian.

(Analogous results hold for the case of digraphs, see Ghouila-Houri [11] and Meyniel [12].) This flood of research culminated in the necessary and sufficient conditions established in 1976 by Bondy and Chvátal [13]. They first observed that the proofs of the results by Dirac and Ore imply the following one:

Given a graph $G = (V, E)$, let i and j be two nonadjacent vertices of V such that $|\delta(i)| + |\delta(j)| \geq n$. Then G is Hamiltonian if and only if $G' = (V, E \cup \{i, j\})$ is Hamiltonian.

Now, define the *closure* of $G = (V, E)$ as the graph obtained from G by recursively adding to E edges $[i, j]$ connecting pairs of nonadjacent vertices i, j such that $|\delta(i)| + |\delta(j)| \geq n$ until no such pair remains. Then, we have the Bondy–Chvátal theorem:

A graph $G = (V, E)$ is Hamiltonian if and only if its closure is Hamiltonian.

Following the above results an impressive number of extensions and refinements has been developed. The reader is referred to the surveys by Bermond [14] and Gould [15,16] for thorough reviews.

Another relevant stream of research consists in developing exact and heuristic algorithms to find HCs in a graph or a digraph. Most of these algorithms (of both kinds) are based on the enumeration method developed in 1966 by Roberts and Flores [17]. We describe it for a graph, the adaptation to digraphs being straightforward. The general strategy is to progressively extend a simple path, going, say, from vertex s to vertex t , by adding at each iteration a feasible edge $[t, k]$, where feasible means that vertex k is not already in the path. The extension continues until either

1. the path includes n vertices, that is, it is Hamiltonian; or
2. no feasible edge exists.

In the former case, if $[t, s] \in E$ we have obtained a HC. Otherwise, and in case (2), exact algorithms backtrack in a systematic way so as to ensure that at the end all possibilities have been explored, while heuristic algorithms modify the current path so that new edges can hopefully be added.

A number of exact algorithms have been developed by improving the above basic enumeration method [18–21]. The improvements are essentially based on (i) identifying edges which cannot be part of any HC, and deleting them from the current graph; (ii) identifying edges which must be in any HC (forced edges), and including them in the current partial solution. Forced edges can in turn produce forced paths, which can lead to further additions and deletions (*multipath* methods). Concerning heuristic algorithms [22–25], most techniques for modifying a path that cannot be extended any longer are based on the following idea by Pósa [26], known as *rotation* (not adaptable to digraphs). Given the current path $(i_1, i_2, \dots, i_\ell, i_{\ell+1}, \dots, i_k)$ and an (infeasible) edge $[i_k, i_\ell]$, create a new path of the same length by deleting edge $i_\ell, i_{\ell+1}$,

and inserting edge $[i_k, i_\ell]$: the new path is $(i_1, i_2, \dots, i_\ell, i_k, i_{k-1}, \dots, i_{\ell+1})$.

A review of exact and heuristic algorithms has been given by Vandegriend [27]. The HC area includes many other streams of research. The surveys cited above discuss hundreds of results in this exciting field.

FORMULATIONS

The most well-known and used formulation of the TSP was introduced in the seminal paper by Dantzig *et al.* [28]. Given a graph $G = (V, E)$ and a subset $S \subset V$, let $\delta(S)$ denote the set of edges with exactly one end vertex in S . For each edge $e \in E$, define a binary variable x_e , taking the value 1 if and only if edge e is in the tour. The STSP is then equivalent to the following 0–1 LP:

$$\min \sum_{e \in E} c_e x_e \quad (1)$$

$$\text{s.t. } \sum_{e \in \delta(i)} x_e = 2 \quad (\forall i \in V), \quad (2)$$

$$\sum_{e \in \delta(S)} x_e \geq 2 \quad (\forall S \subset V : 2 \leq |S| \leq |V| - 2), \quad (3)$$

$$x_e \in \{0, 1\} \quad (\forall e \in E). \quad (4)$$

Equations (2), called *degree equations*, express the fact that the salesman must arrive at and depart from each city. Inequalities (3), called *subtour elimination constraints* (SECs), ensure that the tour is connected.

The ATSP can be formulated in a very similar way using a digraph $G = (V, A)$. For any subset $S \subset V$, let $\delta^+(S)$ denote the set of arcs leaving S . For each arc $(i, j) \in A$, define a binary variable x_{ij} taking the value 1 if and only if arc (i, j) is in the tour. The ATSP is then equivalent to the following 0–1 LP:

$$\min \sum_{(i,j) \in A} c_{ij} x_{ij} \quad (5)$$

$$\text{s.t. } \sum_{j \in \delta^+(i)} x_{ij} = 1 \quad (\forall i \in V), \quad (6)$$

$$\sum_{i \in \delta^-(j)} x_{ij} = 1 \quad (\forall j \in V), \quad (7)$$

$$\sum_{(i,j) \in \delta^+(S)} x_{ij} \geq 1$$

$$(\forall S \subset V : 2 \leq |S| \leq |V| - 2), \quad (8)$$

$$x_{ij} \in \{0, 1\} \quad (\forall (i, j) \in A). \quad (9)$$

Equations (6) and (7) are called *out-degree* and *in-degree* equations, respectively, while inequalities (8) are again called SECs.

Even if, in both formulations, the SECs are exponential in number, this is not a dramatic issue for enumeration algorithms that can generate them *on the fly*, by using polynomial-time separation procedures (see **Branch and Cut**). However, alternative formulations involving a polynomial number of constraints exist and are discussed in the survey by Öncan *et al.* [29].

ALGORITHMS FOR THE SYMMETRIC TSP

For the STSP, the exact algorithms based on the iterative solution of the linear programming relaxation of model (1)–(4) have been proven to be, by far, the most successful approaches. Such algorithms are described in details in the article titled **Mathematical Programming Approaches to the Traveling Salesman Problem** in this encyclopedia. The current section covers the *combinatorial* algorithms that represent milestones for the resolution of the symmetric TSP including some heuristic approaches.

In 1962, Bellman [30], Gonzales [31], and Held and Karp [32] proposed *dynamic programming* based algorithms specialized for symmetric TSPs. The basic idea is that it is possible to build the optimal solution step by step. At iteration k , all subsets S of cities with cardinality k (not containing vertex 1) are considered, and all paths starting from vertex 1 and ending at each vertex of S are computed. The paths of minimum cost are stored. At iteration $k + 1$, the new minimum-cost paths are obtained from those of the previous iteration. The optimal tour is finally constructed from the minimum-cost paths of subset $V \setminus \{1\}$, by connecting the last visited vertex to vertex 1, and taking the minimum-cost tour. The theoretical time bound of this

kind of algorithms was determined by Held and Karp [32]: an instance of n cities can be solved in time at most proportional to $n^2 2^n$. Although such complexity is nonpolynomial, it compares favorably with that of the complete enumeration of all Hamiltonian circuits, which takes time proportional to $(n - 1)!$. Unfortunately, in practice dynamic programming can only be used to solve small size instances, due to the amount of data which must be stored and examined, proportional to $n 2^n$.

An important role in the resolution of the symmetric TSP has been played by branch-and-bound algorithms (see **Branch-and-Bound Algorithms**). In 1963, Little *et al.* [33] developed the first complete branch-and-bound algorithm for the TSP (and actually were the first to propose such a name) by specializing the framework of Land and Doig [34]. Several earlier enumeration algorithms for the TSP were presented before 1963, the most notable being the one by Eastman [35], whose algorithm can probably fit into the branch-and-bound framework. Little *et al.* [33] based their branch-and-bound approach on the LP relaxation. In particular, the lower bound on the TSP cost is computed by finding an approximate dual solution, that is, a lower bound on the LP relaxation of the problem.

The most famous lower bound, which represented a breakthrough in the study of the TSP, was proposed by Held and Karp [36,37]. Remind that a shortest spanning tree of a graph $G = (V, E)$, with $n = |V|$, is a minimum-cost subset of $n - 1$ edges of E that spans all vertices of V (see **Minimum Spanning Trees**). The peculiarity of the branch-and-bound algorithm by Held and Karp is the use of a lower bound given by the shortest spanning tree of the subgraph of G which does not include vertex 1, plus the two minimum-cost edges connecting vertex 1 to the other vertices. (This particular structure is called *1-tree*.) The advantages of such a lower bound are basically two: the shortest spanning tree can be found in $O(n^2)$ time, and the resulting lower bound turns out to be generally good. The bound is tightened through a technique based on a result by Flood [38]. Suppose that we associate a value λ_i to each vertex $i \in V$,

and modify the cost of each edge $e = [i, j]$ as $c_{ij} := c_{ij} - \lambda_i - \lambda_j$. In every tour, each vertex has exactly two incident edges, hence its cost changes by $\sum_{k \in V} 2\lambda_k$. It follows that the optimal tour remains optimal, while the optimal 1-tree generally changes since in a 1-tree the vertices have different degrees. In order to choose the λ values that provide the tightest 1-tree lower bound, Held and Karp developed an iterative technique which, at each iteration, increases (respectively decreases) λ_k ($k \in V$) if, in the current 1-tree, vertex k has one incident edge (respectively more than two incident edges). This technique is related to the Lagrangean relaxation of the degree constraints (2) and the determination of the corresponding Lagrangean multipliers through a subgradient optimization procedure. The reader is referred to Refs [36,37] and to Chapter 4 of Applegate *et al.* [4] for details.

While the 1-tree lower bound is obtained by relaxing constraints (2), another important lower bound comes from the relaxation of constraints (3). The resulting problem is to find, in general, a set of minimum-cost disjoint tours covering all vertices, and is known as *perfect 2-matching*. Such a problem is solvable in polynomial time [39]. The resulting bound is generally not very tight, although techniques for improving it exist. Other lower bounds for the STSP have been obtained by Houck *et al.* [40] (*n-path relaxation*) and by Cowling and Maffioli [41] (*matroid relaxation*).

Among the heuristic algorithms for the TSP, the most successful is probably the local search procedure proposed in 1973 by Lin and Kernighan [42]. They extended an idea by Flood [38], who had proposed the so-called *2-opt move*: starting from a given tour including edges $[i, j]$ and $[k, l]$ (assume that the vertices are in the same order of visit), it is convenient to replace them by edges $[i, k]$ and $[j, l]$ if $c_{ik} + c_{jl} < c_{ij} + c_{kl}$. Lin and Kernighan generalized the *2-opt move* to the more complex *k-opt move* where k edges are replaced with respect to the initial solution. Considering $k > 3$ is impractical in general, but the key-point of Lin–Kernighan’s heuristic is to avoid fixing k to a prefixed value while instead looking for special *k-opt moves*,

with large k , obtained by carefully selected sequences of *2-opt moves*. The quality of the tours obtained in this way is high, and the Lin–Kernighan algorithm is nowadays employed in most exact algorithms for the TSP, usually in the improved version proposed by Helsgaun [43].

Finally, it is worth presenting a fundamental result by Christofides [44] who, in 1976, proposed a polynomial-time algorithm that provides a tour of cost not greater than $3/2$ times the cost of the optimal one for the metric TSP. A TSP instance is *metric* if, for any $i, j, k \in V$, $c_{ij} \leq c_{ik} + c_{kj}$ (*triangle inequality*) holds. The Christofides algorithm starts by finding the shortest spanning tree T of graph G , and the minimum-cost perfect matching M in the subgraph induced by the vertices of odd degree in T . (A *matching* is a subset of edges with no vertex in common, and it is *perfect* if every vertex is an end point of one edge of the matching.) It then forms an Eulerian path (i.e., a path that uses each edge of the graph exactly once) in the multi-graph produced by all edges of T and M , and shortcuts it by skipping repeated vertices: if edges $[i, k]$ and $[k, j]$ are in the path, and k has already been visited, they are replaced by edge $[i, j]$.

This simple and elegant algorithm still provides the tightest approximation guarantee for the metric TSP.

ALGORITHMS FOR THE ASYMMETRIC TSP

In the asymmetric case, there is no dominance between mathematical programming based algorithms and combinatorial branch-and-bound approaches. In particular, the branch-and-cut approach by Fischetti and Toth [45,46] (see **Branch and Cut** for details on branch-and-cut algorithms, and **Mathematical Programming Approaches to the Traveling Salesman Problem** for the specific branch-and-cut for the ATSP) is the fastest algorithm on the asymmetric instances of the TSPLIB [47,48]. However, if the instances are randomly generated in such a way that the costs c_{ij} and c_{ji} for each pair $i, j \in V$ are uncorrelated, then combinatorial algorithms become competitive

and sometimes outperform branch-and-cut approaches (mainly because they are simpler). Note that uncorrelated instances are not necessarily artificial: they do appear in real-world applications like the stacker crane problems described by Ascheuer [49].

Combinatorial algorithms for the exact solution of the ATSP are based on two well-known relaxations. The first one is the *linear assignment problem* (AP) relaxation, obtained by dropping SECs Eq.(8) from models (5)–(9). The surviving constraints impose that each vertex has exactly one leaving arc and one entering arc. Thus the AP is the graph-theoretic problem of finding the minimum-cost collection of vertex-disjoint subtours visiting all vertices of a digraph $G = (V, A)$. The AP can be solved in $O(n^3)$ time (see the recent book by Burkard *et al.* [50] for a thorough analysis of AP algorithms).

The second important relaxation is the *Spanning r -Arborescence Problem* (r -SAP), obtained by dropping constraints (6) from model (5)–(9). Such a problem corresponds to finding a minimum-cost spanning subdigraph $\bar{G} = (V, \bar{A})$ of G such that: (i) the in-degree of each vertex is exactly one; and (ii) each vertex can be reached from a given root vertex r . Such a relaxation can be solved in $O(n^2)$ time by adding a minimum-cost arc entering vertex r to the *shortest spanning arborescence rooted at r* , that is, to the same sub-digraph but with zero in-degree at the root r (see *Minimum Spanning Trees*).

Note that for both the AP and the r -SAP relaxation one can drop the integrality requirements (9) as their linear relaxation always possesses an integer optimal solution which is easily found by the combinatorial algorithms.

It is not difficult to see how one can build a naïve branch-and-bound algorithm by repeatedly solving either AP or r -SAP relaxations. It is less obvious how to make such algorithm effective in practice, and how to combine the two relaxations into a unique algorithm. These two issues are discussed below.

All the effective AP-based branch-and-bound algorithms are derived from the lowest-first branch-and-bound procedure TSP1 presented by Carpaneto and Toth [51].

At each node h of the decision tree, procedure TSP1 solves a *modified* AP (MAP_h) which is the AP plus the additional variable-fixing constraints associated with the arc subsets of *excluded* and *forced* arcs. Thus, MAP_h can easily be transformed into a standard AP by properly modifying the cost matrix so as to take care of the additional constraints.

If the optimal solution to MAP_h does not define a Hamiltonian directed cycle, then ℓ children nodes are generated from node h according to a modified version of the classical subtour elimination rule by Bellmore and Malone [52], where ℓ is the length of the smallest subtour associated with the MAP_h solution. Precisely, if $a_1, \dots, a_\ell$ are the not fixed arcs of such a subtour, the ℓ nodes are generated as

$$x_{a_1} = 1 \vee (x_{a_2} = 1 \wedge x_{a_1} = 0) \vee \dots \vee (x_{a_\ell} = 1 \wedge x_{a_1} = \dots = x_{a_{\ell-1}} = 0). \quad (10)$$

Based on the above branch-and-bound scheme, two very effective AP-based algorithms have been proposed by Miller and Pekny [53], and Carpaneto *et al.* [54]. Both algorithms include a number of sophisticated techniques aimed, among other issues, at reducing the size of the graph and at finding good heuristic solutions. However, the main improvement is the parametric solution of each MAP_h in $O(n^2)$ time, which speeds up the overall computation significantly. On the whole, the two methods exhibit a comparable performance.

Coming to the combination of the AP and r -SAP relaxations, preliminarily observe that such relaxations are complementary to each other. Indeed, the AP imposes the degree constraints for all vertices, disregarding connectivity constraints, while r -SAP imposes reachability from vertex r to all other vertices, disregarding out-degree constraints for all vertices.

A possible way of combining the two relaxations is to apply the so-called *additive approach* that Fischetti and Toth [55] related to the restricted Lagrangean relaxation approach by Balas and Christofides [56], and later used for the ATSP [57]. Roughly speaking, given a set of bounding procedures for a problem P , one can apply them in

sequence so that the reduced costs output by a procedure are used as an input for the next one. Then, the sum of the solution values of all the procedures is a valid bound for P .

In the specific ATSP case, the AP relaxation is solved first, and the resulting reduced cost matrix defines a new ATSP instance on which the r -SAP is solved. The sum of the values of the optimal solutions of the two relaxations is a valid lower bound for the ATSP.

Finally observe that a third relaxation can be easily defined and used in such a framework, namely the *Spanning r -Antiarborescence Problem* (r -SAAP), which is the same as the r -SAP but with the in-degree constraints (7) dropped instead of the out-degree constraints (6).

We conclude this section with some considerations on the approximation of the ATSP. The existence of a polynomial-time algorithm with worst-case error bounded by a constant is still an open question. The most famous result in this field was obtained by Frieze *et al.* [58], whose polynomial-time algorithm, based on the iterated solution of AP relaxations, provides, for the metric ATSP, a solution with performance guarantee $\log n$. This bound was not improved until 2002. Recently Asadpour *et al.* [59] presented a randomized algorithm that produces a solution within a factor of $O(\log n / \log \log n)$ of the optimum with high probability.

SOFTWARE

As already discussed, the TSP has attracted over the years the attention of a number of researchers and practitioners, often belonging to different communities. As an effect almost all algorithmic techniques, both exact and heuristic, have been tested on the TSP. Often, the codes associated with such algorithms have been made available for research and teaching. The purpose of such software might be very different. There are applications (especially Java) created with didactic purposes, or implementations intended as examples for the use of a special algorithmic environment, or pieces of code available as building blocks for more complex software projects. Of course, exhibiting the fastest

computer code to solve the TSP is always a great motivation and this is the case of Concorde [60] by Applegate *et al.* [4].

Many TSP software codes are classified in Lodi and Punnen [61] and the associated web page—http://www.or.deis.unibo.it/research_pages/tspsoft.html—is continuously maintained.

REFERENCES

1. Lawler EL, Lenstra JK, Rinnooy Kan AHG, editors. The traveling salesman problem. Chichester; Wiley; 1985.
2. Reinelt G. The traveling salesman: computational solutions for TSP applications. Lecture Notes in Computer Science. Berlin: Springer; 1994. pp. 840.
3. Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002.
4. Applegate D, Bixby RE, Chvátal V, *et al.* The traveling salesman problem: a computational study. Princeton (NJ): Princeton University Press; 2007.
5. Padberg MW, Rinaldi G. Optimization of a 532 city symmetric traveling salesman problem by branch-and-cut. *Oper Res Lett* 1987;6:1–7.
6. Padberg MW, Rinaldi G. A branch-and-cut algorithm for the resolution of large-scale symmetric travelling salesman problems. *SIAM Rev* 1991;33:60–100.
7. Euler L. Solution d'une question curieuse qui ne paroît soumise à aucune analyse. *Academy of Sciences of Berlin, Mémoires de l'Académie Royale des Sciences et Belles Lettres*. 1776, Année 1759;15, Berlin. pp. 310–337.
8. Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors. *Complexity of computer computations*. New York: Plenum; 1972. pp. 85–103.
9. Dirac GA. Some theorems on abstract graphs. *Proc Lond Math Soc* 1952;2:69–81.
10. Ore Ø. A note on hamilton circuits. *Am Math Mon* 1960;67:55.
11. Ghouila-Houri A. Une condition suffisante d'existence d'un circuit hamiltonien. *Compt R Acad Sci* 1960;156:495–497.
12. Meyniel M. Une condition suffisante d'existence d'un circuit hamiltonien dans un graphe orienté. *J Comb in Theor B* 1973;14:137–147.
13. Bondy JA, Chvátal V. A method in graph theory. *Discr Math* 1976;15:111–136.

14. Bermond JC. Hamiltonian graphs. In: Beineke L, Wilson RJ, editors. *Selected topics in graph theory*. London: Academic Press; 1972. pp. 127–167.
15. Gould RJ. Updating the Hamiltonian problem - a survey. *J Graph Theor* 1991;15: 121–157.
16. Gould RJ. Advances on the Hamiltonian problem: a survey. *Graphs Comb* 2003;19:7–52.
17. Roberts SM, Flores B. Systematic generation of Hamiltonian circuits. *Commun ACM* 1966;9:690–694.
18. Selby GR. The use of topological methods in computer-aided circuit layout [PhD thesis]. London University; 1970.
19. Christofides N. *Graph theory - an algorithmic approach*. London: Academic Press; 1975.
20. Martello S. An enumerative algorithm for finding Hamiltonian circuits in a directed graph. *ACM Trans Math Softw* 1983; 9:131–138.
21. Kocay W. An extension of the multi-path algorithm for Hamilton cycles. *Disc Math* 1992;101:171–188.
22. Angluin D, Valiant LG. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *J Comput Syst Sci* 1979; 18:155–193.
23. Bollobás B, Fenner TI, Frieze AM. An algorithm for finding Hamilton paths and cycles in random graphs. *Combinatorica* 1987; 7:327–341.
24. Frieze AM. Finding Hamilton cycles in sparse random graphs. *J Comb in Theor A* 1987;44:230–250.
25. Thomason A. A simple linear expected time algorithm for finding a Hamilton path. *Disc Math* 1989;75:373–379.
26. Pósa L. Hamiltonian circuits in random graphs. *Disc Math* 1976;14:359–364.
27. Vandegriend B. Finding Hamiltonian cycles: algorithms, graphs and performance [MSc thesis]. University of Alberta; 1998.
28. Dantzig GB, Fulkerson DR, Johnson SM. Solution of a large-scale traveling salesman problem. *Oper Res* 1954;2:393–410.
29. Öncan T, Altinel IK, Laporte G. A comparative analysis of several asymmetric traveling salesman problem formulations. *Comput Oper Res* 2009;36:637–654.
30. Bellman R. Dynamic programming treatment of the travelling salesman problem. *J Assoc Comput Mach* 1962;9:61–63.
31. Gonzales RH. Solution to the traveling salesman problem by dynamic programming on the hypercube. Technical Report Number 18. Cambridge (MA): Operations Research Center, Massachusetts Institute of Technology; 1962.
32. Held M, Karp RM. A dynamic programming approach to sequencing problems. *J Soc Ind Appl Math* 1962;10:196–210.
33. Little JDC, Murty KG, Sweeney DW, *et al.* An algorithm for the traveling salesman problem. *Oper Res* 1963;11:972–989.
34. Land AH, Doig AG. An automatic method of solving discrete programming problems. *Econometrica* 1960;28:497–520.
35. Eastman WL. *Linear Programming with Pattern Constraints* [PhD thesis]. Cambridge (MA): Department of Economics, Harvard University; 1958.
36. Held M, Karp RM. The traveling-salesman problem and minimum spanning trees. *Oper Res* 1970;18:1138–1162.
37. Held M, Karp RM. The traveling-salesman problem and minimum spanning trees: Part II. *Math Program* 1971;1:6–25.
38. Flood MM. The traveling-salesman problem. *Oper Res* 1956;4:61–75.
39. Edmonds J. Maximum matching and a polyhedron with 0,1-vertices. *J Res Nat Bur Stand* 1965;69B:125–130.
40. Houck DJ, Picard J-C, Queyranne M, *et al.* The traveling salesman problem as a constrained shortest path problem: theory and computational experience. *Opsearch* 1980;17:93–109.
41. Cowling P, Maffioli F. A bound for the symmetric travelling salesman problem through matroid formulation. *Eur J Oper Res* 1995;83:301–309.
42. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling-salesman problem. *Oper Res* 1973;21:498–516.
43. Helsgaun K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. *Eur J Oper Res* 2000;126:106–130.
44. Christofides N. Worst-case analysis of a new heuristic for the traveling salesman problem. Report No 388. Pittsburg (PA): Graduate School of Industrial Administration, Carnegie Mellon University; 1976.
45. Fischetti M, Toth P. A polyhedral approach to the asymmetric traveling salesman problem. *Manag Sci* 1997;43:1520–1536.

46. Fischetti M, Lodi A, Toth P. Exact methods for the asymmetric traveling salesman problem. In: Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002. pp. 169–205.
47. Reinelt G. TSPLIB - a traveling salesman problem library. ORSA J Comput 1991; 3:376–384.
48. TSPLIB. Available at <http://comopt.ifl.uni-heidelberg.de/software/TSPLIB95/>.
49. Ascheuer N. Hamiltonian path problems in the on-line optimization of flexible manufacturing systems [PhD thesis]. Berlin, Germany; Technische Universität Berlin; 1995.
50. Burkard R, Dell'Amico M, Martello S. Assignment problems. Philadelphia (PA): SIAM; 2009.
51. Carpaneto G, Toth P. Some new branching and bounding criteria for the asymmetric traveling salesman problem. Manag Sci 1980;26:736–743.
52. Bellmore JC, Malone M. Pathology of traveling salesman subtour elimination algorithms. Oper Res 1971;19:278–307.
53. Miller DL, Pekny JF. Exact solution of large asymmetric traveling salesman problems. Science 1991;251:754–761.
54. Carpaneto G, Dell'Amico M, Toth P. Exact solution of large-scale asymmetric traveling salesman problems. ACM Trans Math Softw 1995;21:394–409.
55. Fischetti M, Toth P. An additive bounding procedure for combinatorial optimization problems. Oper Res 1989;37:319–328.
56. Balas E, Christofides N. A restricted Lagrangean approach to the traveling salesman problem. Math Program 1981;21:19–46.
57. Fischetti M, Toth P. An additive bounding procedure for the asymmetric travelling salesman problem. Math Program 1992;53:173–197.
58. Frieze AM, Galbiati G, Maffioli F. On the worst-case performance of some algorithms for the asymmetric traveling salesman problem. Networks 1982;12:23–39.
59. Asadpour A, Goemans MX, Madry A, *et al.* An $O(\log n / \log \log n)$ -approximation algorithm for the asymmetric traveling salesman problem. In: Charikar M, editor. Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia (PA): SIAM; 2002. pp. 379–389.
60. Concorde TSP Solver. Available at <http://www.tsp.gatech.edu/concorde.html>
61. Lodi A, Punnen A. TSP Software. In: Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002. pp. 737–749.

COMBINING EXACT METHODS AND HEURISTICS

MARCO A. BOSCHETTI
Department of Mathematics,
University of Bologna, Bologna,
Italy

VITTORIO MANIEZZO
Department of Computer Science,
University of Bologna, Bologna,
Italy

INTRODUCTION

The combination of exact and heuristic methods is as old as mathematical programming (MP) itself, because usually exact methods cannot work properly without a good bound on the optimal cost, which helps prune the search space. However, it is only in very recent years that the combination has also gone the other way round, permitting the use of methods originally conceived for exact problem solution also in heuristic frameworks. Noteworthy exceptions have existed, but mainstream approaches in heuristic and metaheuristic researches have traditionally been quite oblivious to elements like bounds, duality, pruning, cutting planes, column generation, and in general of all the armamentaria that makes effective exact codes. This reflected on the scarce attention given to such topics in most conferences and publication outlets dedicated to heuristics or metaheuristics.

This state of things has changed, mainly because of the steady performance improvement of mixed-integer programming (MIP) solvers. The current effectiveness of MIP solvers—mainly of the commercial ones, but some open source ones are becoming an option, too—makes it viable to use them as subroutines for solving (NP-hard) subproblems arising during heuristic search.

MIP solvers used as subroutines are not the only option for designing exact/heuristic hybrids. In this article, we are particularly

interested in revising the newly identified field of *matheuristics*, which refers in particular to algorithms combining metaheuristics with model grounded, typically MIP, techniques. This choice was made because the coverage of the different ways in which generic heuristics have been used as subroutines inside exact codes would imply, as mentioned, a review of 50 years of research in MP. The same is true when considering how exact codes can be turned into heuristics, since every exact code is a heuristic when run under time or memory constraints. Considering metaheuristics alone, it permits to focus on a wide and well-defined research community, which self-statingly produces the best performing generic algorithms for the real-world, NP-hard combinatorial optimization (CO) problems. Besides the fact that for most CO problems they are the method that produced the best known results in the literature, acknowledged merits of metaheuristics also include the fact that they are usually easy to adapt to the many variants of any given problem and that they are usually also easy to design and implement, given a problem.

Moreover, the set of algorithms, which are classified as metaheuristics has significantly enlarged in recent years. Therefore, despite its novelty, to this date more than 50 research papers have already been published under the *matheuristics* heading, journal special issues and book edited collections are in print, and a dedicated international workshop series exists, besides several special sessions on broad coverage optimization conferences. A first analysis of the state of the art can thus be of interest.

Contributions to this area can be classified into two main groups, “how can metaheuristics help exact methods” and “how can exact methods help metaheuristics.” This last contains contributions consisting either of a driving exact approach used to define new metaheuristic techniques, or of a driving metaheuristic algorithm using an exact code as a subroutine. Following either of

these approaches, it has been shown how metaheuristics and MP can leverage on one another, permitting to improve the state of the art on several problems of theoretical interest or arising from real-world practice.

This article reflects this classification, consisting of three sections, the section titled “introduction,” the section titled “MP for Metaheuristics” on metaheuristics using MP as a subroutine, and the section titled “Metaheuristics for MP” on metaheuristics used by MP as a subroutine. The constraint on the length of this article rules out the possibility of an in-depth analysis of the different contributions. Relevant collections can be found in Refs 1–5, while Caserta and Voß [6] provide a detailed overview, which helps to frame the metaheuristics contributions in the more general context of metaheuristics advances. Throughout the text, unless otherwise specified, we make reference to minimization problems, therefore upper bounds are cost of feasible solutions and lower bounds are possibly better-than-optimal costs.

One area of contribution, which escapes this taxonomy refers to the use of mathematical tools to model and predict the performance of metaheuristic. It does not directly refer to hybrid solution techniques, but we like to mention it here because of the complexity and of the quality of the contributions. For example, Gutjahr [7] gives an overview on techniques for proving convergence of metaheuristics to optimal or sufficiently good solutions and estimation of expected runtime, that is, of the time needed by a metaheuristic to hit for the first time, a solution of a required quality. This can be done for metaheuristics applied both to deterministic and to stochastic CO problems. For related research, see also Refs 8–13.

MP FOR METAHEURISTICS

This is the area most intensively studied, where the potential benefits with respect to the state of the art seem to be more immediate. We can identify two directions along

which research is developing. The first makes use of MP results—mainly MIP solvers—as components to be included in an existing metaheuristic framework, such as tabu search, genetic algorithms (GAs), or ant colony optimization (ACO). The second direction aims at producing new metaheuristic frameworks, directly derived from the internal workings of MP methods.

MP as a Subroutine for Known Metaheuristics

This approach is the one that directly makes most use of the complementary merits of MP and metaheuristic algorithms. Most best-known metaheuristic frameworks have been by now complemented in some of their internal steps by MP codes. This is true from hybrids of local search variants and MP techniques, for which an interesting review can be found in Refs 14 and 15, up to hybrids of the most involved frameworks, such as ACO or particle swarm optimization (PSO).

One of the first investigated areas has been the exact recombination of parents, in the frameworks of GAs or alike (evolution strategies, scatter search, memetic algorithms, etc.). The topic of finding the best possible offspring, as a result of a recombination operator in an evolutionary algorithm (EA), given two parent solutions using binary encoding, is theoretically investigated in Ref. 16, where some cases of polynomial or NP-hard cases for optimal recombination are outlined. Two general approaches are feasible: fixing the solution parts common to both parents and optimizing the rest (see Ref. 17 for an early example) or making the union of the parent solution components and optimizing within that set [18,19].

Liberti *et al.* [20] combine several well-known metaheuristics, namely, variable neighborhood search (VNS), local branching, with sequential quadratic programming and branch-and-bound to obtain a method, called *Relaxed-Exact Continuous-Integer Problem Exploration* (RECIPE), which is applied to general mixed integer nonlinear programming, without hand-tuned parameter configuration.

VNS has been hybridized with MP methods in different ways. In Ref. 21, a

VNS approach is presented, which uses three different neighborhood types. Two of them work in complementary ways in order to maximize search effectiveness. Both are large in the sense that they contain exponentially many candidate solutions, but efficient polynomial-time algorithms are used to identify the best neighbors. For the third neighborhood type, mixed integer programming is applied to optimize local parts within candidate solution trees, considering the generalized version of the classical minimum spanning tree problem where the nodes of a graph are partitioned into clusters; and exactly one node from each cluster must be connected. A different hybridization, this time with local branching, is described in Ref. 22.

ACO took advantage of MP results in different ways. In an approach, named *approximate nondeterministic tree search* (ANTS), [23], lower bounds are used to assess the likelihood of success of the possible ants moves, thus making the overall process quite similar to branch and bound constructions. Another approach, which seems particularly suited for strongly constrained problems, hybridizes ACO functioning with bounded enumeration, or beam-search, expansions [24,25,62]. Here the idea is to expand, at each step, not a single solution but a number of them, retaining for further expansion only the best ones, as determined by lower bound and ants-specific considerations.

GRASP and GA were both integrated with MIP solving by Dolgui *et al.* [26], in the context of a successful application to the problem of balancing transfer lines with multispindle machines. Specifically, they design a GRASP and a GA, which both use a MIP solver as a subroutine for solving subproblems arising in the search process of the metaheuristics.

Tabu search has been hybridized with MIP in different ways. For example, Ulrich-Ngueveu *et al.* [27] studied the *m*-peripatetic vehicle routing problem (VRP), which is a special VRP asking that each arc is used in the solution at most once for each set of *m* periods considered in the plan. Their approach uses a perfect *b*-matching to define

the candidate sets to be passed to a granular tabu search algorithm.

We finally mention here a possibility that, under different forms, will be present in all three sections in which we have partitioned the review part of this article (i.e., the sections titled “MP as a Subroutine for known Metaheuristics,” “MP as a Paradigm for New Metaheuristics,” and “Metaheuristics for MP”), and that we denote as the *core problem* approach. In its essence, it prescribes to collect possible components of the problem solution and to include them in an MIP formulation, usually as columns of a set-partitioning (SP) formulation, possibly with additional constraints. The resulting restricted MIP formulation is then solved in an exact or heuristic way. The roots of this approach can be found in Ref. 28, but see also Refs 29–32 for extensions. In the context of MP as a subroutine for other metaheuristics, this has been used by running the metaheuristic of interest (a GA in Ref. 33, a tabu search in Ref. 34) and collecting elements of the discovered solutions (VRP routes in Ref. 34, node subsets in Ref. 33). These components are then possibly recombined in the MIP solution as a postprocessing of metaheuristic search.

MP as a Paradigm for New Metaheuristics

The difference between a “new” and a significant improvement of an existing framework can be blurred; here we consider to be “new,” a method which could not exist without the MP contribution. This includes both methods, which are essentially local searches, where the neighborhood is defined according to a suitable MIP problem or method, and methods derived from proper MP techniques.

A clear example of how to exploit MIP in neighborhood definition is given by the very large neighborhood search (VLNS) by Ahuja *et al.* [35,36]. This technique was first presented for partitioning problems, a general class of CO problems, which includes vehicle routing, capacitated minimum spanning tree, generalized assignment, graph partitioning, parallel machine scheduling, and location problems [35]. The general consideration is that local search algorithms produce

better results when they are based on large neighborhoods, but the larger the neighborhood the longer the time to explore it, thus the time for the algorithm to get a local optimum. The idea behind VLNS is to consider very large neighborhoods, even growing exponentially with the problem size, but to avoid their explicit exploration. The method can be applied when it is possible to define the neighborhood exploration, in order to identify the best neighbor of the incumbent solution, as a CO problem itself. In this case, it could be possible to solve it efficiently, thus making it viable for the full exploration of exponential neighborhoods. Very good computational results were presented for the capacitated minimum spanning tree [37].

A similar objective, but pursued with totally different means, is at the heart of Dynasearch [38]. Here, again, the authors want to improve the performance of a local search algorithm by searching an exponential size neighborhood in polynomial time. This is attained by allowing a series of moves to be performed at each of the iterations, using dynamic programming to find the best sequence of simple moves to be performed. A condition to check is that the simple moves to combine are mutually independent, that is, that the overall improvement to the objective function that results from the independent moves is equal to the sum of the improvements from the individual moves. Good results were presented for the total weighted tardiness scheduling problem, the traveling salesman problem, and the linear ordering problem.

Yet another way to explore exponential neighborhoods is local branching [39]. This is a general heuristic method for MIP problems, not restricted to CO ones. The method is based on the use of an MIP formulation P of the problem to solve and to start, as any local search, with a first feasible solution x_h of the problem. Then, at each of the iterations, an MIP problem is solved, asking to identify the best feasible solution for P which differs from x_h in at most k variables, where k is a parameter. This corresponds to solve to optimality a suitable k -opt neighborhood of x_h , and it is achieved by introducing

so-called *local branching* constraints in the original formulation P , then using an MIP solver for determining the new incumbent solution. This technique has recently gained a considerable popularity and it has been integrated in several ways. A recent survey on the use of MIP solvers as subroutines for solving NP-hard subproblems arise while solving more complex problems, can be found in Ref. 40. Several real-world applications of local branching are being published [41,42].

A further variation of the idea of solving to optimality, a neighborhood of the incumbent solution is at the heart of the corridor method [43]. Here, the neighborhood is defined according to the method M , which will be used for exploring it, be it an MIP solver, dynamic programming, or any other technique. The neighborhood is a part of the solution space, which can be effectively explored by employing M and it is implemented by imposing exogenous constraints on the original problem. This defines a “corridor” around an incumbent solution along which the solver is forced to move. An application is presented in Ref. 44 on the blocks relocation problem in block stacking systems, with application in the stacking of container terminals in a yard. The proposed algorithm is based on a dynamic programming formulation, defining a two-dimensional “corridor” around the incumbent blocks configuration.

There are also methods rooted in MP, which depart from the optimized neighborhood search approach. Boschetti *et al.* [45, 46] show how it is possible to use decomposition techniques, which were originally conceived as tools for exact optimization, as metaheuristic frameworks. The general structure of each of the best known decomposition approaches (Lagrangian, Benders and Dantzig-Wolfe) can, in fact, be considered as a general structure for a corresponding metaheuristic. Special attention is given to Lagrangian decomposition, which has a long track record as a basis for heuristics, for which a novel, fully distributed subgradient optimizer is envisioned. Results are presented for a problem arising in P2P network design [the Membership Overlay Problem, (MOP)]

and for a more standard network design problems [61].

Bartolini and Mingozzi [47] propose an algorithm which builds on partial enumeration search. The new algorithm, named *F&B* (shorthand for “Forward and Backward”), escapes from local minima by adopting a memory-based look ahead strategy that exploits the knowledge gained in its past search history. It iterates a partial exploration of the solution space by generating a sequence of enumerative trees of two types, called *forward* and *backward trees*. Each node at level h of the trees represents a partial solution X^L containing h items. At each iteration t , the algorithm generates a forward tree if t is odd, or a backward tree if t is even. In generating a tree, each partial solution X^L is extended to a feasible one using as completion candidates the partial solutions generated at the previous iteration, and stored in the reverse-ordered tree. The cost of the resulting solution is used to guess the quality of the best complete solution that can be obtained from X^L .

Finally, we mention the MP-as-metaheuristic variant of the *core problem* approach introduced in the section “MP as a Subroutine for Known Metaheuristics.” In this case, it consists in working directly on the MIP formulation of the problem of interest trying to iteratively generate as many columns as possible and solving, again in an exact or heuristic way, the resulting MIP instance. This idea was advocated first for solving large set covering instances [48,49], then named *Kernel Search* by Speranza *et al.* [50]. Here, an initial solution is obtained over a small set of promising elements, denoted as the problem *kernel*. The kernel is initially built using information provided by the solution of the linear relaxation of the original problem (for a similar approach, see also Ref. 51). Then new promising elements are identified by solving a sequence of small/moderate size MIP problems, each restricted to the previous kernel plus a set of other elements that were identified by the previous MIP problem in the sequence. An application to portfolio optimization is proposed [50].

METAHEURISTICS FOR MP

This is a relatively unexplored area, where contributions are just starting to be published. Puchinger *et al.* [52] propose a first review of this subject, showing how the possibilities for allowing exact codes to take advantage of metaheuristic computations are varied, and range from providing good quality starting solutions to using metaheuristics for cut separation or column generation.

The most obvious way to use metaheuristics in an exact context is to get tight upper bound to prune the search tree. This can be simply done running a metaheuristic before the exact method, but more elaborate proposals exist. For example, Rothberg [53] integrates an EA in a branch and cut, applying it at regular intervals as tree node heuristic. The population of the EA consists of the best nonidentical solutions found so far, which have either been discovered by the MIP tree search or by previous iterations of the EA itself.

The use of heuristics to separate violated cuts has always been common practice when implementing branch-and-cut methods. Occasionally, metaheuristics were used to this end. A first proposal in this direction was made by Augerat *et al.* [54] in the context of the capacitated VRP. They propose a branch-and-cut algorithm in which a set of methods ranging from simple construction to a tabu search are used for separating capacity constraints. The algorithm starts with the fast simple heuristic and switches to a more complex strategy when no more valid cutting planes could be found. A similar procedure was used by Gruber and Raidl [55], in the context of a branch and cut for the bounded diameter minimum spanning tree problem, based on so-called *jump inequalities*. The separation subproblem of identifying violated jump inequalities is difficult, and was approached by two alternative greedy construction heuristics, followed by local and tabu search to identify and strengthen violated cuts. Moreover, they include in their approach a variable neighborhood descent for finding good primal solutions.

This led to very interesting computational results.

A different possibility was explored by Rei *et al.* [56]. They worked on a Benders decomposition of an MIP problem and suggested to start obtaining different heuristic solutions of the problem to solve (using local branching in their case). These solutions, besides providing good upper bounds, allow to derive multiple additional cuts before solving the Benders master problem.

Another area which calls for profitable use of metaheuristics is the column generation phase (or the pricing problem) in a column generation algorithm. Puchinger and Raidl [57] describe a branch-and-price algorithm for a special case of bin packing problem. The authors propose to approach column generation by applying a sequence of four methods such as a greedy heuristic, an EA, the solving of a restricted, a simpler IP-model of the pricing problem using CPLEX within a given time-limit, and finally the solving of a complete IP-model by CPLEX. This proved computationally good, especially in time-constrained solution of large-scale instances.

Finally, we have here again the possibility of using the “core problem” approach of the section titled “MP for metaheuristics.” In this case, we have a metaheuristic that can be seen as a preprocessing phase for the exact approach, usually a column generation, feeding elements of good solutions and thus triggering a warm start of the column generation phase [58]. Notice that the contribution of the metaheuristic is not restricted to ensuring a good starting upper bound, which is anyway helpful for any exact search, but it is also important for stabilizing dual variables, thus speeding up the overall search. Furthermore, metaheuristics can be used to generate negative reduced cost columns. This is shown, for example, in Refs 59 and 60, where Ref. 59 makes use of a tabu Search, while Ref. 60 propose a GRASP and a GA which are able to generate at each call not just only one column, but a set of columns with negative reduced costs.

CONCLUSIONS

We are witnessing a steady increase of interest on exploiting MP and metaheuristics synergies. The area is still new, and researchers are exploring it along many different directions. Though a few successful approaches have already been established, for example, local branching or VLNS, we can expect to see new ones appearing in the near future, together with the improvements of methods for which we now have the idea and little more than a proof-of-concept.

REFERENCES

1. Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
2. Hansen P, Maniezzo V, Voß S. Special issue on mathematical contributions to metaheuristics editorial. *J Heuristics* 2009;15(3):197–199.
3. Jourdan L, Basseur M, Talbi E-G. Hybridizing exact methods and metaheuristics: a taxonomy. *Eur J Oper Res* 2009;199(3):620–629.
4. Raidl GR, Puchinger J. Combining (Integer) linear programming techniques and metaheuristics for combinatorial optimization. In: Blum C, *et al.*, editors. Volume 114, *Hybrid metaheuristics - an emerging approach to optimization: Studies in computational intelligence*. Berlin: Springer; 2008. pp. 31–62.
5. Talbi E-G. A taxonomy of hybrid metaheuristics. *J Heuristics* 2002;8(5):541–565.
6. Caserta M, Voß S. Metaheuristics: intelligent problem solving. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009. pp. 1–38.
7. Gutjahr WJ. Convergence analysis of metaheuristics. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
8. Gutjahr WJ. A provably convergent heuristic for stochastic bicriteria integer programming. *J Heuristics* 2009;15(3):227–258.
9. Oliveto PS, He J, Yao X. Time complexity of evolutionary algorithms for combinatorial optimization: a decade of results. *Int J Autom Comput* 2007;4:281–293.

10. Hoos HH. On the runtime behavior of stochastic local search algorithms for SAT. Proceedings of the 16th National Conference on Artificial Intelligence. Menlo Park (CA): AAAI Press/The MIT Press; 1999. pp. 661–666.
11. Margolin L. On the convergence of the cross-entropy method. *Ann Oper Res* 2005;134: 201–214.
12. Stützle T, Dorigo M. A short convergence proof for a class of ACO algorithms. *IEEE Trans Evol Comput* 2002;6:358–365.
13. Trelea IC. The particle swarm optimization algorithm: convergence analysis and parameter selection. *Inform Process Lett* 2003;85:317–325.
14. Dumitrescu I, Stützle T. Usages of exact algorithms to enhance stochastic local search algorithms. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
15. Dumitrescu I, Stützle T. Combinations of local search and exact algorithms. In: Raidl GR, *et al.*, editors. Volume 2611, Applications of evolutionary computation. LNCS, Berlin: Springer; 2003. pp. 211–223.
16. Ereemeev A. On complexity of optimal recombination for binary representations of solutions. *Evol Comput* 2008;16(1):127–147.
17. Yagiura M, Ibaraki T. The use of dynamic programming in genetic algorithms for permutation problems. *Eur J Oper Res* 1996;92:387–401.
18. Balas E, Niehaus W. Optimized crossover-based genetic algorithms for the maximum cardinality and maximum weight clique problems. *J Heuristics* 1998;4(2):107–122.
19. Aggarwal CC, Orlin JB, Tai RP. An optimized crossover for the maximum independent set. *Oper Res* 1997;45:226–234.
20. Liberti L, Nannicini G, Mladenović N. A good recipe for solving MINLPs. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
21. Hu B, Leitner M, Raidl GR. Combining variable neighborhood search with integer linear programming for the generalized minimum spanning tree problem. *J Heuristics* 2008; 14(5):473–499.
22. Hansen P, Mladenovic N, Urošević D. Variable neighborhood search and local branching. *Comput Oper Res* 2006;33(10):3034–3045.
23. Maniezzo V. Exact and approximate non-deterministic tree-search procedures for the quadratic assignment problem. *INFORMS J Comput* 1999;11(4):358–369.
24. Blum C. Beam-ACO—Hybridizing ant colony optimization with beam search: an application to open shop scheduling. *Comput Oper Res* 2005;32(6):1565–1591.
25. Maniezzo V, Milandri M. In: Dorigo M, *et al.*, editors. Volume 2463, An ant-based framework for very strongly constrained problems. LNCS, ANTS, Berlin: Springer; 2002. pp. 222–227.
26. Dolgui A, Ereemeev A, Guschinskaya O. MIP-based GRASP and genetic algorithm for balancing transfer lines. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
27. Ulrich NS, Prins C, Wolfer Calvo R. A hybrid tabu search for the m-peripatetic vehicle routing problem. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
28. Balas E, Zemel E. An algorithm for large zero-one knapsack problems. *Oper Res* 1980;28(5):1130–1154.
29. Pisinger D. An expanding-core algorithm for the exact 0–1 knapsack problem. *Eur J Oper Res* 1995;87(1):175–187.
30. Pisinger D. Core problems in knapsack algorithms. *Oper Res* 1999;47(4):570–575.
31. Puchinger J, Raidl GR, Pferschy U. The multidimensional knapsack problem: structure and algorithms. *INFORMS J Comput* 2010;22(2):250–265.
32. Huston S, Puchinger J, Stuckey PJ. The core concept for 0/1 integer programming. Volume 77, Proceedings of the 14th Computing: the Australasian Theory Symposium (CATS 2008). Wollongong, Australia: CRPIT, ACS; 2008. pp. 39–47.
33. Ribeiro FG, Nogueira Lorena LA. Constructive genetic algorithm and column generation: an application to graph coloring. In: Chuen LP, editor. Proceedings of the 5th Conference of the Association of Asian-Pacific Operations Research Societies within IFORS. Singapore, IFORS; 2000.
34. De FR, Fischetti M, Toth P. A new ILP-based refinement heuristic for vehicle routing problems. *Math Program* 2006;105:471–499.

35. Ahuja RK, Orlin JB, Sharma D. Very large-scale neighbourhood search. *Int Trans Oper Res* 2000;7(4–5):301–317.
36. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123(1–3):75–102.
37. Ahuja RK, Orlin JB, Sharma D. A composite very large-scale neighborhood structure for the capacitated minimum spanning tree problem. *Oper Res Lett* 2003;31(3):185–194.
38. Congram RK, Potts CN, Van de Velde SL. An iterated dynasearch algorithm for the single-machine total weighted tardiness scheduling problem. *INFORMS J Comput* 2002;14(1):52–67.
39. Fischetti M, Lodi A. Local branching. *Math Program B* 2003;98:23–47.
40. Fischetti M, Lodi A, Salvagnin D. Just MIP it! In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
41. Rodríguez-Martín I, Salazar-González JJ. A local branching heuristic for the capacitated fixed-charge network design problem. *Comput Oper Res* 2010;37(3):575–581.
42. Schmid V, Doerner KF, Hartl RF, *et al.* Hybridization of Local Branching guided by Variable Neighborhood Search for Ready-Mixed Concrete Delivery Problems, *MATHEURISTICS 2* submitted to *Computers & Operational Research*.
43. Sniedovich M, Voss S. The corridor method: a dynamic programming inspired metaheuristic. *Control Cybern* 2006;35(3):551–578.
44. Caserta M, Voß SA. Corridor method-based algorithm for the pre-marshalling problem. Volume 5484/2009, *Applications of evolutionary computing*. LNCS. Berlin: Springer; 2009. pp. 788–797.
45. Boschetti M, Maniezzo V. Benders decomposition, Lagrangean relaxation and metaheuristic design. *J Heuristics* 2009;15(3):283–312.
46. Boschetti M, Maniezzo V, Roffilli M. Decomposition techniques as metaheuristic frameworks. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
47. Bartolini E, Mingozzi A. Algorithms for the non-bifurcated network design problem. *J Heuristics* 2009;15(3):259–281. DOI: 10.1007/s10732-008-9091-1.
48. Ceria S, Nobili P, Sassano A. A Lagrangian-based heuristic for large-scale set covering problems. *Math Program B* 1998;81:215–228.
49. Caprara A, Fischetti M, Toth P. A heuristic method for the set covering problem. *Oper Res* 1999;47(5):730–743.
50. Angelelli E, Mansini R, Speranza MG. Kernel search: a heuristic framework for MILP problems with binary variables. *Matheuristics* 2008, Technical Report of the Department of Electronics for Automation, University of Brescia; R.T. 2007-04-56, 2007.
51. Danna E, Rothberg E, Le Pape C. Exploring relaxation induced neighborhoods to improve MIP solutions. *Math Program A* 2005;102:71–90.
52. Puchinger J, Raidl GR, Pirkwieser S. MetaBoosting: enhancing integer programming techniques by metaheuristics. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
53. Rothberg E. An evolutionary algorithm for polishing mixed integer programming solutions. *INFORMS J Comput* 2007;19(4):534–541.
54. Augerat P, Belenguer JM, Corberan A, *et al.* Separating capacity constraints in the CVRP using tabu search. *Eur J Oper Res* 1999;106(2):546–557.
55. Gruber M, Raidl GR. (Meta-)heuristic separation of jump cuts in a branch&cut approach for the bounded diameter minimum spanning tree problem. In: Maniezzo V, Stützle T, Voß S, editors. *Matheuristics: hybridizing metaheuristics and mathematical programming*, OR/CS Interfaces Series. New York: Springer; 2009.
56. Rei W, Cordeau J-F, Gendreau M, *et al.* Accelerating Benders decomposition by local branching. *INFORMS J Comput* 2009;21(2):333–345.
57. Puchinger J, Raidl GR. Models and algorithms for three-stage two-dimensional bin packing. *Eur J Oper Res* 2007;183:1304–1327.
58. Fernandes MAE, Iori M, Malaguti E, *et al.* Hybridizing population heuristics and column generation for the bin packing problem with conflicts. *Matheuristics* 2008. <http://astarte.csr.unibo.it/Matheuristics2008/>.
59. Taillard ED. A heuristic column generation method for the heterogeneous fleet VRP. *RAIRO* 1999;33(1):1–14.

60. dos Santos AG, Mateus GR. Hybrid approach to solve a crew scheduling problem: an exact column generation algorithm improved by metaheuristics. 7th International Conference on Hybrid Intelligent Systems. Kaiserslautern, Germany: IEEE; 2007. DOI: 10.1109/HIS.2007.58.
61. Boschetti MA, Maniezzo V, Roffilli M. A fully distributed lagrangean solution for a peer-to-peer overlay network design problem. *INFORMS J Comput* 2010; DOI: 10.1287/ijoc.1100.0381.
62. Blum C. Beam-ACO for simple assembly line balancing. *INFORMS J Comput* 2008; 20(4):618–627.

COMBINING FORECASTS

INGRIDA RADZIUKYNIENE

PETROS XANTHOPOULOS

PANOS M. PARDALOS

Department of Industrial and Systems
Engineering and Biomedical
Engineering, University of Florida,
Center for Applied Optimization,
Gainesville, Florida

INTRODUCTION

Several questions can be posed on the usefulness of combining different forecasts. Is it worth combining forecasts? Is it better to combine forecasts or to combine different sources of data sets? Which individual forecasts we should combine in order to achieve better results? How much can one increase the forecasting efficiency by using combining forecasts? For sure, there is no unique answer for all of these questions. The decision is always application dependent. Sometimes the answer might come from the natural constraints of the problem (e.g., when only a fixed data set is available, the combining data sets option is infeasible). Nevertheless, over the years several methods have been proposed for aggregating forecasts. In this article, we present an overview of this general framework, and we try to address questions relevant to the usefulness of combining forecast methodology.

By combining forecasts, we mean all the aggregation methods that aim to produce a better forecast for estimating the value of some random variable or some series of random variables (a stochastic process). Research toward this direction was triggered by the intuitive belief that a combination of many individual forecasts is better than just one. Individual forecasts can be provided by any computational algorithm or might be the result of an expert decision maker. In general, the combined forecast can be expressed as a weighted average of the individual

forecasts. Combining forecasts methodology is about defining the weights so that the prediction accuracy is maximized. Historically, the first modern attempt was made in 1924 when Gordon asked several people to estimate some lifted weights [1]. By averaging the judgment of many judges, she observed that the estimate had higher correlation with the real value. The formal framework of combining forecasts was first presented by Granger and Bates [2]. Since then, there were several research papers that suggested alternative methods for forecast aggregation.

Through the years, with the development of algorithms, there have been many software packages that aim to offer a framework for forecasting to the end user. Many of these software packages incorporate methods for combining forecasting. In the literature, one can find plenty of documentation and comparative studies for such software [3–6].

The rest of the article is organized as follows: first, we present an overview of the most important methods used in the order of ascending complexity (starting from simple linear aggregation methods of averaging and weighted mean, and then describing more involved nonlinear models). In the last section, we cite some computational results and discuss the potential usefulness of combining forecasts.

METHODS

Combining forecasts based on single models can integrate the merits of different single models to enhance the performance. The combining methods can be roughly classified into two groups: “variance-covariance” methods and “regression-based” methods. According to the combining style of single models, there are two kinds of combining forecasts [7]: linear and nonlinear. In this article, we consider the following linear methods: simple combination, discounted mean squared forecast error (MSFE), shrinkage, factor model, and time-varying parameter (TVP) combination forecasts, as well as nonlinear methods based

on expert knowledge, clustering, and fuzzy logic.

The difference among these methods is based on the manner of how the historical information is employed to combine forecasts and the variability of the weight of a particular forecast during that time period.

Linear Methods

The most widely used method for combining forecasts is weighted average of the individual forecasts. It has the following form:

$$f_c = \sum_i \omega_i f_i, \quad \sum_i \omega_i = 1, \quad (1)$$

where f_c is the combined forecast value and f_i are the individual forecasts and ω_i are the weights that correspond to the importance that every individual forecast has.

Simple Combination Forecasts. In the simple combination method, the combined forecast is computed irrespective of the historical performance of the individual forecasts. As mentioned in the introductory part, the first and the most simple way to combine forecasts is to take the average of all the individual f_i s. In this case, ω_i are all equal to $1/N$ (where N is the number of individual forecasts). This is a simple and more secure way to combine forecasts, especially when we do not have any information about which forecast is more accurate. In cases when we are aware that some forecasts have been more accurate (by utilizing previous knowledge), we modify the weights so that greater importance is given to those with higher accuracy.

The second simple combining forecasts method is a very common practical tactic in order to minimize the influence of such outlier forecasts, that is, removing the highest overestimator and the lowest underestimator (this is also known as *trimming*). In this case, if we have N forecasts, the trimmed combined forecast is given by

$$f_c = \sum_{i=2}^{N-1} \omega_i f_i, \quad (2)$$

provided that f_i s are sorted in ascending order. Here, it is worth noting that we do not consider any constraint on signs of the weights on purpose. As mentioned in Ref. 8, combined forecast might have increased performance if some of ω_i are negative. This might be the case when an individual forecast is historically known to give bad predictions.

Discounted MSFE. The discounted MSFE approach computes the weights, which depend inversely on the historical performance of each individual forecast [9]. In this method, a combined forecast is estimated over a time window $[t, t+h]$ by utilizing the information up to time t . The h -step-ahead combined forecast is calculated by the following equation:

$$f_{c,t+h|t} = \sum_{i=1}^n \omega_{it} f_{i,t+h|t}, \quad (3)$$

where $f_{i,t+h|t}$ is the individual forecast over the time window $[t, t+h]$, given the information up to time t , and ω_{it} is the weight on the i th forecast in period t . It is calculated according to the formula

$$\omega_{it} = m_{it}^{-1} / \sum_{j=1}^n m_{jt}^{-1}, \quad \text{where} \\ m_{it} = \sum_{s=T_0}^{t-h} \delta^{t-h-s} E_{it}^2, \quad (4)$$

where δ is the discount factor (when $\delta = 1$, there is no discounting) and E_{it} is the error of the i th forecast at the t th time sample.

Outperformance. Outperformance approach was proposed in Ref. 7. It calculates the forecast combination as $f_c = \sum \omega_i f_i$, where $\omega = [\omega_1, \omega_2, \dots, \omega_N]$ is a simplex of probabilities which can be estimated using the Bayesian approach. In this method, each weight is interpreted as the probability of the corresponding forecast giving the best fit (in a way of the smallest absolute error) on the successive occurrence. Each probability is estimated as the fraction of occurrences, where the corresponding forecasting model has performed the best in the past. Menezes

et al. [10] state that this approach is a robust nonparametric method for achieving differential weights with intuitive meaning, and it performs well when there is relatively little past data, and/or when the decision maker wishes to incorporate expert judgment into the combining weights.

Optimal Method. In the optimal method [2], the calculated linear weights are used to minimize the error variance of the combination. The vector used for combining forecasts is calculated using the following formula:

$$\omega = \frac{S^{-1}e}{e'S^{-1}e}, \quad (5)$$

where e is the $(n \times 1)$ unit vector and S is the $(n \times n)$ covariance matrix of forecast errors. Granger and Ramanathan [11] showed that this method is equivalent to least squares regression in which the constant is suppressed, and the sum of weights has to be equal to one. The shortcoming of this method is that it is necessary to have appropriately estimated S . In practice, it is difficult to evaluate S , because it is often nonstationary. In such a case, it is common to estimate it on the basis of a short history of forecasts, and, as a result, the method becomes an adaptive approach to combining forecasts. In optimal (adaptive) with independence assumption method, the estimate of S in Equation (5) is restricted to be diagonal, comprising only the individual forecast error variance. A case where the optimal formula has the additional restriction that no individual weight can be outside the interval $[0,1]$ is called *optimal* (adaptive) with restricted weights.

Regression. Regression has several variations and extensions; however, overall, it is a simple and flexible method. It is common when combining forecasts to use the constituent forecasts as regressors in the ordinary least squares (OLS) regression with the inclusion of a constant. Granger and Ramanathan [11] stated that it has an advantage over the popular optimal method. The advantage is that an unbiased combined forecast is produced regardless of whether

the constituent forecasts are biased. This can be achieved when one combines the individual forecasts in such a way that their individual errors cancel out. A special case where the least squares regression is done with the inclusion of a constant but the weights are constrained to sum up to one is called *regression with restricted weights*.

Weights are estimated by applying regression on the target variable f_τ on the N -vector of forecast $\hat{f}_{\tau|\tau-1}$ using data over period $\tau = 1, \dots, T$:

$$\hat{\omega}_T = \left(\sum_{\tau=1}^{T-1} \hat{f}_{\tau+1|\tau} \hat{f}_{\tau+1|\tau}' \right)^{-1} \times \sum_{\tau=1}^{T-1} \hat{f}_{\tau+1|\tau} f_{\tau+1}. \quad (6)$$

Many authors have proposed different versions of regression. For example, Granger and Ramanathan [11] consider the following three:

$$\begin{aligned} f_{t+1} &= \omega_t^0 + \omega_t' \hat{f}_{t+1|t} + \varepsilon_{t+1}, \\ f_{t+1} &= \omega_t' \hat{f}_{t+1|t} + \varepsilon_{t+1}, \\ f_{t+1} &= \omega_t' \hat{f}_{t+1|t} + \varepsilon_{t+1}, \text{ s.t. } \omega_t' e = 1. \end{aligned} \quad (7)$$

The first and second of these regressions can be estimated by standard OLS, the only difference being that the second equation omits an intercept term, ω_t^0 . The third equation omits an intercept and it can be estimated through constrained least squares.

The disadvantage of this method is that when applied in unbalanced data sets, the performance is very poor. The estimate of the complete covariance matrix for this type of data is infeasible. The way to deal with such cases is to introduce minimum data requirements and trim the set of forecasts. For instance, it can be done by enforcing the requirements to have forecasts from a certain minimum number of common periods.

In order to partially address this issue, Capistran and Timmermann [12] apply a weighting scheme which, for forecasters with a sufficiently long track record, uses weights that are inversely proportional to their historical MSE-values while using equal weights

for the remaining forecasters (normalized so that the weights add up to one).

Shrinkage. Shrinkage methods have been widely used in forecasting, where the weights are computed as an average of the recursive OLS estimator of the weights [11] and equal-weighting. However, the shrinkage forecasts have the following form:

$$\begin{aligned}\omega_t^i &= \lambda_t \hat{\omega}_t^i + (1 - \lambda_t)(1/N_t), \\ \lambda_t &= \max(0, 1 - \kappa N_t / (T - 1 - N_t - 1)),\end{aligned}\quad (8)$$

where $\hat{\omega}_t^i$ is the least squares estimator of the weight of the i th model in the calculated combined forecast, for example from one of the regression in Equation (7); κ defines the extent of the shrinkage, where the larger values of κ result in a lower λ_t and consequently in a greater degree of shrinkage toward equal weights. When the sample size, T , approaches to the number of forecasts, N , a larger weight is assigned to the least squares estimate. Because this approach is based on the least squares estimator, it faces the same problem as the underlying method. Contrarily, when values of T and N are fixed, greater values of κ coincide with a larger extent of shrinkage leading to equal-weighting (smaller λ_t), and thus this method exhibits some of the problems associated with using equal weights.

Time-Varying Parameter Forecasts. The TVP models usually have a state space form. The state space representation allows unobserved variables to be involved and be evaluated with the observable model. To estimate the time-varying coefficient in forecast regression, the Kalman filter is used.

Stock and Watson [13] modified the approach of combining forecast proposed in Ref. 11, that is, they imposed a zero intercept and extended to have TVPs:

$$f_{s+h}^h = \omega_{1t} \hat{f}_{1,s+h|s}^h + \cdots + \omega_{nt} \hat{f}_{n,s+h|s}^h + \varepsilon_{s+h}^h, \quad (9)$$

where $\omega_{it} = \omega_{it-1} + \eta_{it}$ and η_{it} are serially uncorrelated, uncorrelated with ε_{s+h}^h , and

uncorrelated across i . Essentially, the relative variance $\text{var}(\eta_{it})/\text{var}(\varepsilon_{s+h}^h)$ can be estimated but with many forecasts; its estimation was not reliable enough, so, instead, the authors set the relative variance to $\text{var}(\eta_{it})/\text{var}(\varepsilon_{s+h}^h) = \phi^2/n^2$, where ϕ is a chosen parameter. Larger values of ϕ lead to larger time variation. The initial distribution of ω_{it} settles each weight to $1/n$ with zero variance; taking the limit when $\phi = 0$, the TVP combination forecast reduces to the combined forecast given by simple mean combination.

Nonlinear Methods

The nonlinear combining forecast has been applied widely in recent years [14–16]. In this section, we discuss some of the most prominent methods used for nonlinear combination of forecasts.

Artificial Neural Networks. An artificial neural network (ANN) is a mathematical model inspired by biological neural networks. It is composed of an interconnected group of artificial neurons working together to solve specific problems in fields such as pattern recognition or data classification. In most cases, an ANN is an adaptive system that changes through a learning process.

ANN can be seen as a nonlinear statistical data modeling tool that can be used to model complex nonlinear relationships between a set of individual forecasts and the variable being forecasted. Let $f_{j,t}$ denote the forecast from model j for time t ; $\bar{d}$ and S_d denote the in-sample mean and in-sample standard deviation of the variable being forecasted out-of-sample. In Ref. 17, the following ANN model was applied for forecasts combination:

$$Z_{j,t} = (f_{j,t} - \bar{d})/S_d, j \in \{1, 2\}, \quad (10)$$

$$\begin{aligned}\Psi(z_t, \gamma_i) &= \left(1 + \exp \left[- \left(\gamma_{0,i} + \sum_{j=1}^2 \right) \right. \right. \\ &\quad \left. \left. \times \gamma_{1,i,j} z_{j,t} \right] \right)^{-1},\end{aligned}\quad (11)$$

$$F_t = \beta_0 + \sum_{j=1}^k \beta_j f_{j,t} + \sum_{i=1}^p \delta_i \Psi(z_t, \gamma_i),$$

$$k \in \{0, 2\}, \quad p \in \{0, 1, 2, 3\}. \quad (12)$$

In Equation (10), the individual forecasts are standardized. This procedure and the appropriately selected γ 's undertake that Ψ in Equation (11) usually coincides with the value close to 0.5. The final combined forecast is given in Equation (12). Equation (12) can be easily implemented by neural networks with value of some γ . One way to select γ is to assign a random number with uniform generator from $[-1, 1]$. The parameter δ can be evaluated by applying the least squares technique.

Clustering. Clustering-based forecast methods were applied by some authors [18,19] to sales data analyzed as time series. The common feature of these methods was the usage of past sales data for clustering time series. Kumar [20] proposed the simple clustering method based on the trade-off between decreased variance and increased bias due to combining, where the similarity in the next period forecast and its variance to determine clusters of time series are used. The goal is to minimize the MSE of forecasts. The total MSE for all k clusters is given by

$$\text{MSE}(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{i \in C_j} \left[(\mu_i - \bar{\mu}_{C_j})^2 + \frac{\sigma_i^2}{n_j} \right], \quad (13)$$

where $n_j = |C_j|$ and $\bar{\mu}_{C_j} = \frac{1}{n_j} \sum_{i \in C_j} \mu_i$. Unfortunately, μ_i and σ_i are unknown and they are estimated from data. Replacing μ_i and σ_i in Equation (13) by estimates y_i and s_i^2 , the objective function of clustering is as follows:

$$\min_{C_1, \dots, C_k} = \sum_{j=1}^k \sum_{i \in C_j} \left[(y_i - c_j)^2 + \frac{s_i^2}{n_j} \right], \quad (14)$$

where $c_j = \frac{1}{n_j} \sum_{i \in C_j} y_i$. The first term (i.e., the bias) $\sum_{j=1}^k \sum_{i \in C_j} (y_i - c_j)^2$ decreases with

decreasing number of data points in each cluster, and this leads to a larger number of clusters. The second term (variance) $\sum_{j=1}^k \sum_{i \in C_j} \frac{s_i^2}{n_j}$ decreases with increasing number of data points in each cluster and this leads to a smaller number of clusters. This approach is based on the trade-off between the bias and the variance of combining forecasts from multiple items.

In order to overcome the disadvantages of this simple clustering method, Kumar proposed the weighted average clustering method, where the weight of forecast is given as the inverse of its variance. So the weighted average forecast for items in cluster C_j is calculated as follows:

$$c'_j = \frac{\sum_{i \in C_j} \frac{y_i}{s_i^2}}{\sum_{i \in C_j} \frac{1}{s_i^2}}. \quad (15)$$

In such a case, the new clustering criterion has the following form:

$$\begin{aligned} & \min_{C_1, \dots, C_k} \text{WMSE}(C_1, \dots, C_k) \\ &= \min_{C_1, \dots, C_k} \sum_{j=1}^k \sum_{i \in C_j} \frac{1}{s_i^2} \\ & \quad \times \left[(y_i - c'_j)^2 + \frac{1}{\sum_{i \in C_j} \frac{1}{s_i^2}} \right] \end{aligned} \quad (16)$$

The author showed that simple clustering reduced the overall MSE of forecast by 7.6%, from 26.1% to 18.5%, and weighted clustering reduced it further by 3.7%, from 10.5% to 6.8%.

Fuzzy Systems. Fuzzy control systems are systems based on fuzzy logic, a language that allows incorporation of qualitative knowledge into solving some specific problem. Fuzzy logic-based control systems offer a general framework for implementation of linguistic IF-THEN rules.

In order to combine forecasts, Fiordaliso [21] used the Takagi–Sugeno fuzzy system, a special class of fuzzy systems whose output of each rule is a scalar. A typical single fuzzy rule in first order has the form

$$\text{IF } X \text{ is } A \text{ THEN } Z = h(X), \quad (17)$$

where A is a fuzzy set with membership function m_A and $h(X)$ is a polynomial of order one in the components X_i of the input variable X . The output f_c of such system with r rules [IF X is A_k THEN $Z = B_k(X)$] can be written as follows:

$$f_c = \frac{\sum_{k=1}^r m_{A_k}(X) B_k(X)}{\sum_{k=1}^r m_{A_k}(X)}, \quad (18)$$

where

$$B_k(X) = b_k(0) + b_k(1)X_1 + \dots + b_k(p)X_p. \quad (19)$$

Letting X_j denote the forecasted value $f_j(t)$ and using Equation (19), the combined forecast can be expressed as $f_c = \omega_0(t) + \omega_1(t)f_1(t) + \dots + \omega_p(t)f_p(t)$, where

$$\omega_j(t) = \frac{\sum_{k=1}^r m_{A_k}(f(t)) b_k(j)}{\sum_{k=1}^r m_{A_k}(f(t))}. \quad (20)$$

This property can be achieved only with the special structure of Takagi–Sugeno first order system because of its linear local model.

Every rule in the given system implements a particular combined forecast according to Equation (18). The input of each rule to combining forecasts is determined by the membership function. Fiordaliso [21] used the p -dimensional generalized Gaussian-type membership functions for the coding of the linguistic terms:

$$m_{A_k}(X) = G\left(\|X - \mu_k\|_{S_k}^2\right), \quad (21)$$

where μ_k is the center of the Gaussian, G is defined by $G(x) = \exp(-x)$ and $\|X\|_{S_k}$ is

the weighted norm of X defined by $\|X\|_{S_k}^2 = X^T S_k^T X$, where S_k is a square matrix corresponding to a specific metric in the multidimensional input space, which is responsible for the linear transformation of individual forecasts. In such a case, the system can be considered as a smooth switching regression model, where two regimes can be active at the same time [21]. In order to allow monitoring, the importance of each rule k in the inference process, a nondecreasing function $g(\rho_k)$ with its values in the $[0, 1]$ can be introduced in the final output f_c :

$$f_c = \frac{\sum_{k=1}^r g(\rho_k) m_{A_k}(X) B_k(X)}{\sum_{k=1}^r g(\rho_k) m_{A_k}(X)}. \quad (22)$$

A small $g(\rho_k)$ value can be used to remove a particular rule from a system; that is, this rule will make a very small contribution to the combined forecast.

Palit and Popovic [22] proposed performing a nonlinear forecast combination based not only on fuzzy logic approach but also on neuro-fuzzy approach. In the neuro-fuzzy approach, they used the fuzzy logic system:

$$f(x) = \frac{\sum_{l=1}^M [y^l z^l]}{\sum_{l=1}^M [z^l]}, \quad (23)$$

$$z^l = \prod_{i=1}^n \exp\{-(x_i - c_i^l)^2 / \sigma_i^l\},$$

which is based on the Gaussian membership function, singleton fuzzifier, product inference rule, and center of area defuzzifier. It is equivalent to the multilayer feedforward network, described in Ref. 23 as neuro-fuzzy network, with the means c_i^l , the variances σ_i^l , and the fuzzy region centers y^l as the adjustable parameters of the network.

The superiority of combining forecasts based on neuro-fuzzy approach to any individual forecast involved in the combination was demonstrated by calculating performance indices.

Rule-Based Approach. Some experts [Makridakis and Hibon [24], Bunn and Wright [25]] advise to combine the judgment and quantitative or statistical methods to get better forecast accuracy. Collopy and Armstrong [26] proposed the rule-based forecasting, that is, the procedure that applies forecasting competence and knowledge to make forecasts according to the features of the data. Their proposed rule is based on integrated strategies such as

- using features of the series to establish weights for combining forecasts;
- using heuristics to establish parameters for an exponential smoothing model;
- using separate models for long-range and short-range forecasts;
- damping the trend under certain conditions;
- incorporating domain knowledge in extrapolation.

They created a rule base of 99 rules with reference to report analysis of five experts on forecasting methods. Four extrapolation methods are used to produce combinations of forecasts.

The authors report 42% less errors for rule-based forecasting than those from equal weights combining and this improvement is statistically significant. It should be noted that improvement depends on the condition of the data.

CONCLUSIONS

A very important question is whether combining forecasts is really worth the effort. In other words, one would like to know if the intuitive statement “many predictors are better than one” is always confirmed in practice. Some researchers prefer combining different data sets versus combining different forecasts. In any case, one should be aware that the comparison between combining data sets and combining forecasts might not always be fair, since different sources of data sets might be expensive to obtain or even infeasible.

In all respects, one should be always extremely cautious when combining forecasts. In Ref. 27, Armstrong suggests some practical rules for aggregating forecasting methods: combining at least five methods, using some formal procedure for combining forecasts, using trimmed means, utilizing prior experience to adjust weights when taking means, and using manually adjusted weighted means when there is some strong evidence from the application.

Also, on a total of 30 studies from 1950 to 2000, he reports that there was on average 12.5% error reduction (ranging from 3.4–24.2%) when combining forecasts was employed.

However, there have been some researchers who have reported negative results (meaning that individual forecasts outperformed the combined). Although there is no mathematical proof for the one or the other side, Granger, who was the one who first formally introduced the concept in his review paper [8], says “if that is true (individual forecasts are better than combined), this would be very disturbing for forecasters as it would mean that very simple methods of forecasting are difficult to improve upon.”

Acknowledgments

Authors would like to acknowledge Sibel B. Sonuç for proofreading the manuscript and for providing useful suggestions and corrections.

REFERENCES

1. Gordon K. Group judgments in the field of lifted weights. *J Exp Psychol* 1924;7:398–400.
2. Bates JM, Granger CWJ. The combination of forecasts. *Oper Res Q* 1969;20:451–468.
3. Tashman LJ, Leach ML. Automatic forecasting software: a survey and evaluation. *Int J Forecast* 1991;7(2):209–230.
4. Sanders NR, Manrodt KB. Forecasting software in practice: use, satisfaction, and performance. *Interfaces* 2003;33(5):90–93.
5. Mahmoud E. Preview of selected software for sales forecasting and decision support systems. *J Acad Market Sci* 1988;16(3):104–109.

6. Kusters U, Bell M. The forecast report: a comparative survey of commercial forecasting systems. Brookline (MA): IT Research Corporation; 1999.
7. Bunn DW. A Bayesian approach to the linear combination of forecasts. *Oper Res* 1975;26:325–329.
8. Granger CWJ. Invited review combining forecasts—twenty years later. *J Forecast* 1989;8:167–173.
9. Diebold FX, Pauly P. Structural change and the combination of forecasts. *J Forecast* 1987;6:21–40.
10. de Menezes LM, Bunn DW, Taylor JW. Review of guidelines for use of combined forecasts. *Eur J Oper Res* 2000;120:190–204.
11. Granger CWJ, Ramanathan R. Improved methods of forecasting. *J Forecast* 1984; 3:197–204.
12. Capistran C, Timmermann A. Forecast combination with entry and exit of experts. CREATES Research Paper No. 2008-55. 2007. pp. 1–37.
13. Stock JH, Watson MW. Combination forecasts of output growth in a seven-country data set. *J Forecast* 2004;23(6):405–430.
14. Dong J-R, Yang X-T. Nonlinear combination forecasting method based on fuzzy logic systems. *J Manage Sci China* 1999;3:28–33.
15. Dong J-R, Yang X-T. Research on nonlinear combining exchange rate forecasts. *China J Manage Sci* 2001;9(5):1–7.
16. Han P, Xi Y-M. Combining forecast based on fuzzy neural network in credit risk. *Quant Tech Econ* 2001;5:107–110.
17. Donaldson RG, Kamstra M. Neural network forecast combining with interaction effects. *J Franklin Inst* 1998;336:227–236.
18. Maharaj EA, Inder BA. Forecasting time series from clusters. International Symposium on Forecasting; Washington DC, USA. 1999.
19. Mitchell RJ. Forecasting electricity demand using clustering. Proceedings of the 21st IASTED International Conference on Applied Informatics; Innsbruck, Austria. 2003. pp. 225–230.
20. Kumar M. Combining forecasts using clustering. Rutcor Research Report. 2005. pp. 1–17.
21. Fiordaliso A. A nonlinear forecasts combination method based on Takagi–Sugeno fuzzy systems. *Int J Forecast* 1998;14:367–379.
22. Palit AK, Popovic D. Nonlinear combination of forecasts using artificial neural network, fuzzy logic, and neuro-fuzzy approaches. The Ninth IEEE International Conference on Fuzzy Systems, Volume 2; San Antonio (TX). 2000. pp. 566–571.
23. Palit AK, Popovic D. Forecasting chaotic time series using neuro-fuzzy approach. Proceedings of IJCNN'99; 1999 July 10–16; Washington DC, USA.
24. Makridakis S, Hibon M. Accuracy of forecasting: an empirical investigation. *J R Stat Soc [Ser A]* 1979;142:97–145.
25. Bunn D, Wright G. Interaction of judgmental and statistical forecasting methods: issues and analysis. *Manage Sci* 1991;37:501–518.
26. Collopy F, Armstrong JS. Rule-based forecasting: development and validation of an expert system approach to combining time series extrapolations. *Manage Sci* 1992; 38(10):1394–1414.
27. Armstrong JS. Combining forecasts. Principles of forecasting: a handbook for researchers and practitioners. Norwell (MA): Kluwer Academic Publishing; 2001. pp. 417–439.

COMBINING SCENARIO PLANNING WITH MULTIATTRIBUTE DECISION MAKING

PAUL GOODWIN
School of Management,
University of Bath, Bath,
Somerset, United Kingdom

Companies and other organizations have to make long-term strategic decisions in a highly uncertain environment. Economic booms and slumps, unforeseen technological developments, changes in government policies, the actions of competitors, and many other factors that are beyond the control of the decision maker can all lead to outcomes that were totally unexpected when the original plans were laid. Businesses that have been highly successful for decades can suddenly find themselves in difficulties, while new companies, which are able to exploit changes in the environment, come to dominate their markets.

However, uncertainty is not the only factor that challenges strategic decision makers. Many decisions will be made with the intention of achieving several objectives. Sometimes these objectives will reflect the different perspectives of the various stakeholders in the organization, such as the shareholders, the senior managers, the non-managerial employees, and the customers. Typical objectives may include maximizing profits, maximizing market share, minimizing pollution, maximizing employee welfare, minimizing risk, and maximizing customer satisfaction. Many of these objectives are likely to conflict. For example, minimizing pollution may come at the expense of reduced profits. This means that decision makers will need to make trade-offs.

Research by psychologists suggests that unaided decision makers often have difficulties in handling the complexity presented by both uncertainty and multiple objectives [1, 2]. When confronted with this complexity, they resort to the use of simplifying

mental strategies, or heuristics. While these heuristics allow decisions to be made quickly and with relatively little cognitive effort, they are prone to biases. As a result, probabilities are misjudged and some objectives may receive insufficient attention, while others are overweighted. In addition, decision makers may also be subject to strategic (or cognitive) inertia, which is reflected in a tendency to continue with existing strategies even when changes in the organization's environment mean that these strategies are no longer appropriate [3].

All of this suggests that managers planning their future strategies can benefit from some form of decision support. Multiattribute decision analysis provides this support by breaking complex decisions down into smaller and, hopefully, easier problems, thereby helping managers to address the trade-offs they need to make between their objectives. Scenario planning also provides a structure to help managers to explore the uncertainties that they face, without relying on probability estimation. In the following sections, we consider how the benefits of the two approaches can be combined to allow managers to plan and adapt their strategies with enhanced insight and with a greater awareness of the challenges posed by an ever-changing environment.

SCENARIO PLANNING

Scenario planning avoids the problems that decision makers have in estimating probabilities for future events. Instead, a series of stories are constructed, usually in structured sessions involving groups of managers with a facilitator. Independent-minded experts (referred to as "remarkable people") who might challenge an organization's current thinking may also be involved in the process [4]. Each story describes how the future might unfold between the date when the strategic decision is made and the planning horizon. While the particular combination

of events described by each scenario is extremely unlikely to occur, it is hoped that, collectively, these stories will encompass the range of possible futures.

There are several approaches to the writing of scenarios but all involve a formal structured procedure [5]. The “intuitive logics” method is one approach that is widely used. Wright and Cairns [4] give a detailed and comprehensive explanation of how to apply this method, but the key steps are demonstrated below using an example relating to a manufacturer of electric cars.

1. *Identify the focal issue and time frame which the scenarios are intended to cover.*

The focal issue in this case is: “the viability of the business for the next ten years.”

2. *Identify the driving forces that will impact on the focal issue.*

Driving forces that are relevant here include:

- a. The extent to which there are improvements in battery technology.
- b. Future movements in oil prices.

- c. The extent to which governments provide tax incentives for purchasers of electric cars.

- d. The extent to which battery-recharging points become widely available in public places.

- e. The extent to which there are stricter international targets on greenhouse gas emissions.

3. *Cluster these driving forces into factors by exploring how they interrelate. Give each factor a name.*

This step is intended to consider how the driving forces interact. In some cases, the outcome of one driving force will have an effect on the outcome of another.

Figure 1 shows how driving forces can be arranged into a factor called “level of market penetration of electric cars.” For example, the extent to which there are stricter international targets on emissions will affect the extent to which governments encourage the development of an infrastructure of charging points in public places for electric vehicles. This, in turn, will affect consumers’ relative preference for electric cars over other types of vehicle.

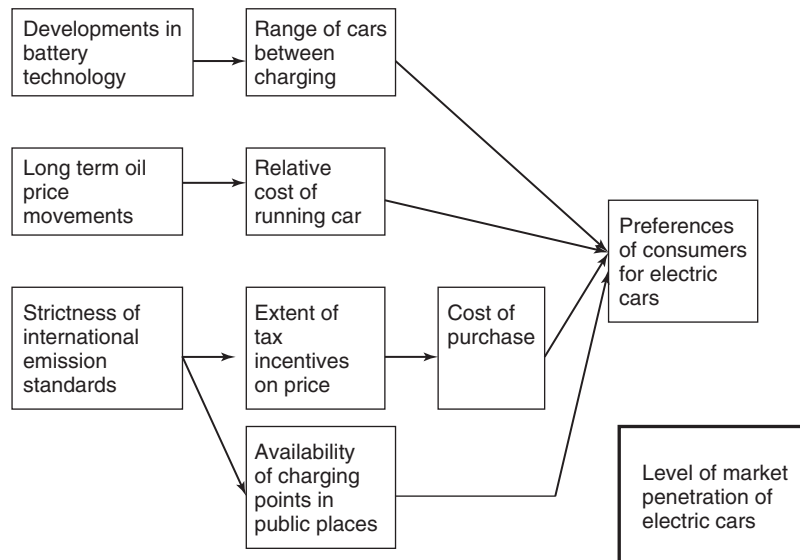


Figure 1. A factor displaying a typical cluster of driving forces.

Usually, several factors will be identified. For example, in this case, we may have other clusters relating to the cost of manufacture, the level of entry into the industry by competing manufacturers, world economic recession or growth, and so on.

4. *Identify extreme outcomes for each factor.*

Examples of extreme outcomes are as follows.

Outcomes leading to high profitability for the manufacturer

Very high market penetration for electric cars.

Unit manufacturing costs fall.

Few manufacturers enter market.

World economic growth.

Outcomes leading to poor profits for the manufacturer

Little or no market penetration for electric cars.

Unit manufacturing costs rise as a result of increased commodity prices.

High levels of international competition with other manufacturers.

World economic recession.

5. *Plot the factors on a graph to identify those that have the maximum potential impact on the focal issue and the greatest uncertainty about this impact.*

Figure 2 shows a typical graph that assumes that we have identified eight factors, including the level of market penetration that we identified earlier.

It can be seen that the two factors, “market penetration” and “world economic recession or growth,” are judged to have both the highest impact and the greatest level of uncertainty. We now list the extreme outcomes associated with these factors and give them a label.

Market penetration: A1: High penetration

A2: Very low penetration

World economy B1: Growth

B2: Recession

We then check that the four combinations of outcomes: A1 & B1, A1 & B2, A2 & B1, and A2 & B2 are plausible, in that they could coexist in the future. If they could not, then we need to review the earlier steps.

6. *Write a scenario based on each of the four combinations and give each scenario a name.*

Each scenario should also include other high impact factors that are regarded as relatively certain. For example, the effect of manufacturing costs should be included in all scenarios, given its importance. The scenario for A1 & B1, which has been named “Green Boom,” is as follows.

“Greater world demand for oil leads to significantly higher prices. In many countries, electric cars also carry low rates of taxation when purchased and licensed. Developments in battery technology lead to longer intervals between charging and higher sales lead to significant falls in the manufacturing costs of vehicles. Environmental concerns over pollution also lead to increased taxes on oil-based fuels in many countries. In addition, governments subsidise the development of infrastructures to support the recharging of electric cars, including the widespread availability of power points in car parks and at supermarkets. World economic growth stimulates international demand, but competing manufacturers are slow to catch up with the technologies needed to develop mass production of electric vehicles.”

Three other scenarios would be developed around the other combinations of extreme outcomes. We will refer to the scenario based on outcomes A1 & B2 as “Electrics dominate in tough times” and those based on A2 & B1 and A2 & B2 as “Boom bypasses electrics” and “Rocky road ahead.” Collectively, these scenarios would attempt to bound the range of possibilities of what might be expected to happen in the specified time frame. They can then be used to inform strategic planning.

Proponents of scenario planning argue that it brings a number of important benefits to strategic decision making. Stories may be a natural way for humans to make sense of

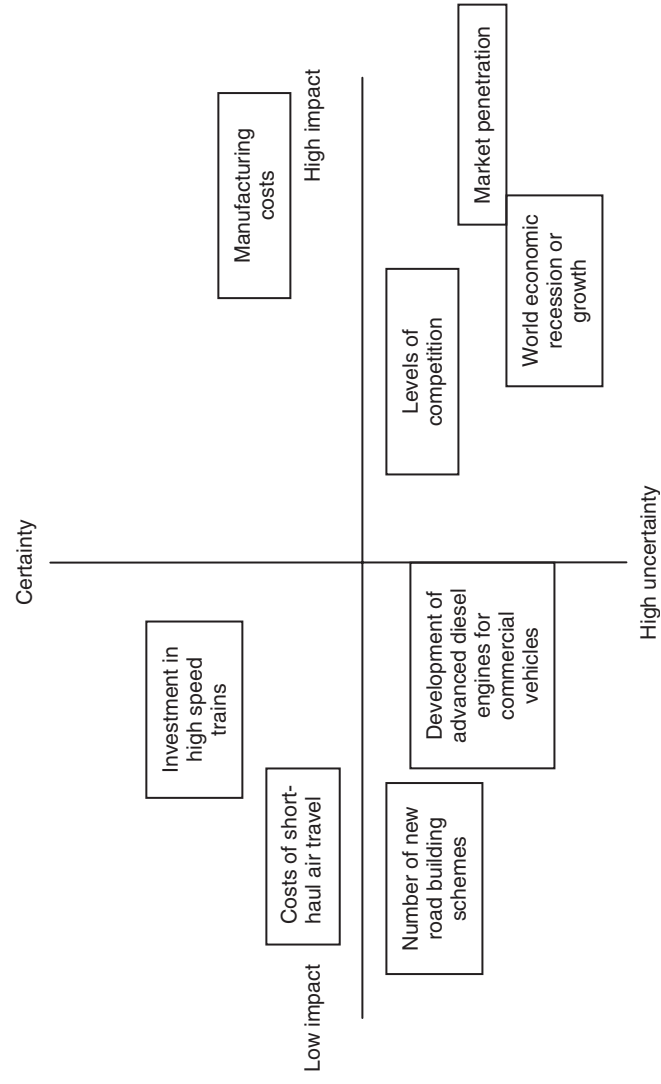


Figure 2. Impact-uncertainty plot.

the world and they provide an explanation of why particular futures might unfold. Once a scenario has been read, it may alert managers that events are moving in a particular direction allowing early contingency plans to be laid to deal with potential problems or new business opportunities to be embraced. In addition, the process of scenario development allows a diversity of views about the future to be reflected so that minority opinions are considered, while the presence of remarkable people may counter strategic inertia.

However, scenario planning lacks a theoretical underpinning and different practitioners tend to construct scenarios in different ways. There is also the danger that, as more detail is added to scenarios, they appear to become more plausible and hence more probable when, in fact, the opposite is the case [1]. Most importantly, the basic scenario planning process does not provide guidance as to how alternative strategies should be evaluated. Usually, the evaluation is informal. The range of scenarios are regarded as being analogous to a wind tunnel within which alternative strategies can be assessed for their robustness and success under the various conditions that might prevail in the future. But where the strategic decision depends on a plurality of objectives, it is difficult for unaided decision makers to carry out such an assessment, given the multiplicity of scenarios and objectives that have to be considered.

COMBINING SCENARIO PLANNING WITH THE SIMPLE MULTIATTRIBUTE RATING METHOD

In order to provide support for decision makers faced with the need to evaluate strategies over multiple objectives and scenarios, Goodwin and Wright [6] suggested an approach that is intended to be simple and transparent. This combines the simple multiattribute rating technique (SMART) method [3, 7] with scenario planning. We illustrate the main stages of the process for the electric car manufacturer.

Stage 1: Identify objectives.

The electric car manufacturer identifies the following three objectives.

- a. Maximize the sustainability of profits in the long term.
- b. Minimize the impact of its operations on the environment.
- c. Maximize its share of the car market.

Stage 2: Design alternative strategies.

The manufacturer is considering one of the following three strategies to pursue.

- a. Produce a limited range of small family electric cars for the domestic market (LIMITED).
- b. Produce a wider range of cars and aim for international sales (WIDE RANGE).
- c. Diversify into hybrid and petrol vehicle production for the international market (DIVERSIFY).

Stage 3: For each objective:

- a. Rank all the strategy/scenario combinations from best to worst.
Table 1 shows these ranks for each objective.
- b. Allocate a score of 100 to the best strategy/scenario combination and 0 to the worst.
- c. Allocate scores between 0 and 100 to measure the performance of intermediate strategy scenario combinations.

These scores are shown in Table 2

Stage 4:

- a. List the objectives and for each one consider the improvement between the worst-performing strategy (scoring 0) and the best (scoring 100). Rank these improvements in order of importance or desirability.

Table 3 shows these ranks.

- b. Attach a weight of 100 to the objective that offers the most important or desirable improvement and obtain weights for the other objectives by comparing the importance of the improvements they offer against the maximum weight of 100. Normalize the weights so that they sum to one by dividing each weight by the sum of the weights.

Table 4 shows the weights obtained for the electric car manufacturer.

Table 1. Ranking the Performance of Strategies for each Objective and Scenario

Strategy	Green Boom	Electrics Dominate in Tough Times	Boom Bypasses Electrics	Rocky Road Ahead
Objective: maximize sustainability of profits in long-term scenarios				
LIMITED	2	5	8	11
WIDE RANGE	1	4	9	12
DIVERSIFY	3	6	7	10
Objective: minimize environmental impact scenarios				
LIMITED	6	5	1 =	1 =
WIDE RANGE	8	7	3 =	3 =
DIVERSIFY	12	11	9 =	9 =
Objective: maximize share of car market scenarios				
LIMITED	5 =	5 =	11 =	11 =
WIDE RANGE	3 =	3 =	9 =	9 =
DIVERSIFY	1 =	1 =	7 =	7 =

1 represents best performance. Equals signs indicate where the ranks are tied.

Table 2. Scoring the Performance of the Strategies on each Objective

Strategy	Green Boom	Electrics Dominate in Tough Times	Boom Bypasses Electrics	Rocky Road Ahead
Objective: maximize sustainability of profits in long-term scenarios				
LIMITED	85	50	25	10
WIDE RANGE	100	60	20	0
DIVERSIFY	70	40	30	15
Objective: minimize environmental impact scenarios				
LIMITED	65	70	100	100
WIDE RANGE	55	60	80	80
DIVERSIFY	0	20	30	30
Objective: maximize share of car market scenarios				
LIMITED	50	50	0	0
WIDE RANGE	60	60	5	5
DIVERSIFY	100	100	20	20

Table 3. Ranking the Importance of the Improvements

Improvement	Rank of Improvement
Worst profit sustainability to best	1
Minimize environmental impact	2
Worst share of electric market to best	3

Table 4. Assigning Weights to the Objectives

Improvement	Rank of Improvement	Weight	Normalized Weight
Worst profit sustainability to best	1	100	0.59
Minimize environmental impact	2	60	0.35
Worst share of electric market to best	3	10	0.06
	Total	170	1

Stage 5: For each strategy / scenario combination, use the performance scores and weights to determine a weighted aggregate score.

For example, Table 5 shows how the score has been obtained for the LIMITED strategy under the Green Boom scenario.

Stage 6: Produce a table of strategy / scenario aggregate scores and use this to assess and compare the strategies' performances, paying particular attention to the robustness of performance over the range of scenarios.

Table 6 shows the resulting scores (which have been rounded to whole numbers).

It can be seen that the DIVERSIFY strategy is dominated by the others. Whichever scenario prevails, its performance is worse than the other two. The LIMITED strategy is likely to be more appealing to a risk averse decision maker. Although it does not perform quite as well as WIDE RANGE in the first two scenarios, it performs much better in the others so it appears to carry less risk.

Stage 7: Perform sensitivity analysis. This is because the scores and weights are likely to be based on rough estimates, or there may be disagreements within

the decision-making team, so it is important to determine the effect of variations in these estimated values.

For brevity, this will not be demonstrated here.

Goodwin and Wright [3], who first proposed the method, suggest that in practice it is likely that the process will need to switch backward and forward between the stages as decision makers develop an increased understanding of the problem and wish to revise their earlier judgments.

EXTENSIONS OF THE METHOD AND ALTERNATIVES

The above-outlined basic method has a number of limitations. In some circumstances, it can require a large number of judgments to be elicited from decision makers. Wright and Cairns [4] have suggested a simpler form that only involves the assessment of ranks for the performances and objectives (this avoids stages 3b, 3c, and 4b). This method is likely to be most appropriate where the gaps between the performances of the strategies in the different scenarios are fairly even and the same applies to the gaps between the weights assigned to the objectives. It will be less appropriate where, for example, strategy A would have been allocated a

Table 5. Calculating a Weighted Aggregate Score for a Strategy

Objective	Score	Weight	Score × Weight
Profit sustainability	85	0.59	50
Environmental impact	65	0.35	23
Market share	50	0.06	3
—	—	—	Weighted aggregate score = 76

Table 6. Weighted Aggregate Scores for each Strategy under each Scenario

Strategy	Scenarios			
	Green Boom	Electricity Dominate in Tough Times	Boom Bypasses Electricity	Rocky Road Ahead
LIMITED	76	57	50	41
WIDE RANGE	82	60	40	28
DIVERSIFY	47	37	29	21

weight of 100, with strategy B coming a close second with a weight of 90, while strategy C would have lagged far behind with a weight of only 20. Karvetski *et al.* [8] also demonstrated a method designed to reduce the number of judgments that needed to be elicited from decision makers. This was also based on ranks, though the way that decision makers' preferences were inferred from these ranks differed from the Wright and Cairns approach.

Goodwin and Wright's method makes the strong assumption that the relative importance of the objectives will be the same under different scenarios. In some situations, this may not be the case. For example, the objective of minimizing dependency on oil may be seen to be relatively less important in a scenario of increased availability of oil, resulting from the discovery of extensive new oil fields, than in a scenario of declining world oil resources. Where this is the case, Montibeller *et al.* [9] suggest extending the approach by developing separate multiattribute models for each scenario. They applied their extended methods to two Italian companies. The application by Karvetski *et al.* [8] was also able to consider changes in preferences across scenarios, while still minimizing the number of judgments required from decision makers.

Some authors recommend that decision makers should select strategies on the basis of how robust they are across the range of scenarios. A strategy that performs *reasonably* well, whatever happens in the future, is deemed to be superior to the one that performs superbly in some scenarios, but disastrously in others. However, researchers have yet to resolve how robustness should be measured and the extent to which it is desirable.

Scenario analysis and multiattribute decision analysis have been combined in several other ways. For example, Durbach and Stewart [10] proposed an alternative approach that involved the integration of goal programming, rather than SMART, with scenarios. Other researchers have combined the two approaches in order to appraise the relative desirability of alternative scenarios (e.g., [11]). In their analysis, it is assumed that the

decision maker has a choice between scenarios, rather than regarding them as outcomes that are largely or wholly beyond the decision maker's control.

CONCLUSIONS

In a volatile and uncertain world, scenario planning can be an attractive tool enabling managers to explore the range of plausible futures that may impact on their organization and sensitizing them to potential opportunities or threats. Alternative strategies can be evaluated under different conditions that may apply in the future, but where there are multiple objectives, this evaluation may be cognitively demanding and lead to biases in the decision-making process. Combining multiattribute value analysis with scenario planning can help to overcome these biases, allowing managers to address the full range of issues that may be relevant to their planning and providing a documented rationale for any decisions that they make.

REFERENCES

1. Kahneman D. Thinking fast and slow. New York: Farrar, Straus and Giroux; 2011.
2. Payne JW, Bettman JR, Johnson EJ. The adaptive decision maker. Cambridge: Cambridge University Press; 1993.
3. Goodwin P, Wright G. Decision analysis for management judgment. Chichester: John Wiley & Sons; 2009.
4. Wright G, Cairns G. Scenario thinking: practical approaches to the future. New York: Palgrave Macmillan; 2011.
5. van der Heijden K. Scenarios: the art of strategic conversation. Chichester: John Wiley & Sons; 1996.
6. Goodwin P, Wright G. Enhancing strategy evaluation in scenario planning: a role for decision analysis. *J Manage Stud* 2001;38:1–16.
7. von Winterfeldt D, Edwards W. Decision analysis and behavioral research. Cambridge: Cambridge University Press; 1986.
8. Karvetski CW, Lambert JH, Keisler JM, Linkov I. Integration of decision analysis and scenario planning for coastal engineering and climate change. *IEEE Trans Syst Man Cybern* 2011;41:63–73.

9. Montibeller G, Gummer H, Tumidei D. Combining scenario planning and multi-criteria decision analysis in practice. *J Multi-Criteria Decis Anal* 2006;14:5–20.
10. Durbach I, Stewart TJ. Integrating scenario planning and goal programming. *J Multi-Criteria Decis Anal* 2003;12:261–271.
11. Kowalski K, Stagl S, Madlener R, Omann I. Sustainable energy futures: methodological challenges in combining scenarios and participatory multi-criteria analysis. *Eur J Oper Res* 2009;197:1063–1074.
- Stewart TJ. Dealing with uncertainties in MCDA. In: Figueira J, Greco S, Ehrgott M, editors. *Multiple criteria decision analysis—state of the art surveys*. New York: Springer; 2005. p 445–470.
- Taleb NN. *The black swan*. New York: Random House; 2007.
- Wright G, Goodwin P. Decision making and planning under low levels of predictability: enhancing the scenario method. *Int J Forecast* 2009;25:813–825.

FURTHER READING

Stewart TJ. Scenario analysis and multicriteria decision making. In: Climaco J, editor. *Multicriteria analysis*. Berlin: Springer; 1997. p 519–528.

COMMON FAILURE DISTRIBUTIONS

SATYANSHU KUMAR UPADHYAY
Department of Statistics, DST Centre
for Interdisciplinary Mathematical
Sciences, Banaras Hindu
University, Varanasi, India

INTRODUCTION

A failure (or life) time distribution is an attempt to describe mathematically the length of life of an item or an organism under study. That is, a probability distribution of a certain form is usually assumed to provide among many other characteristics the reliability or survival probability of the item under consideration at each time t . However, since there are many physical or biological causes that individually or collectively may be responsible for the failure of an item or an organism at any given instant, and since it is not generally possible to isolate these causes, the choice of a failure distribution is definitely an art. The selection of failure distribution in a particular situation is thus subject to uncertainty and is often decided based on empirical considerations, although it does not necessarily imply the absolute correctness of the assumed model. Sometimes, we may have information about the ageing or failure processes in the population that may ultimately suggest several forms of failure distributions but often not able to narrow our consideration to specific family of models [1].

To fix the idea, let T be a nonnegative random variable representing the failure time of an item or an organism. Five different functional forms may normally be used for representing the failure distribution of T . These may be listed as the survival function, the probability density function, the hazard function, the cumulative hazard function, and the mean residual life function, the last one representing the remaining life given survival until time t . If $f(t)$ denotes the probability density function of the random variable T , the explicit expressions for the

survival function, the hazard function, the cumulative hazard function, and the mean residual life function can be given as

$$S(t) = P(T \geq t), \quad (1)$$

$$h(t) = \frac{f(t)}{S(t)}, \quad (2)$$

$$H(t) = \int_0^t h(u) du, \quad (3)$$

$$r(t) = E(T - t | T \geq t), \quad (4)$$

respectively, where the random variable T is assumed to be continuous. It is to be noted that these are not the only representations, the other representations, although not often in practice, may include the moment generating function, the characteristic function, and the Mellin transform. A detailed discussion of these various concepts can be found in Ref. 2.

It is not always clear whether a continuous or a discrete failure model should be used for the random variable T under consideration though the discrete models are advocated only in very few situations. An example can be the assessment of software reliability where the time can be modeled discretely. The five distributional representation concepts mentioned in the preceding lines apply to discrete distributions as well. Except for the probability density function which is now replaced by the probability mass function, the names for the other four representations remain the same.

There are numerous parametric models used in the analysis of failure time data and in the problems related to the modeling of ageing or failure processes. Among univariate class of models, a few particular distributions play a central role because of their demonstrated usefulness in a wide variety of situations. Exponential, Weibull, gamma, lognormal, generalized gamma, and so on, are some commonly available continuous distributions. The important discrete distributions may include geometric, binomial, Poisson, negative binomial, hypergeometric, multinomial, and so on,

among others. Besides, there are several other models that have been used successfully in lifetime contexts based on specific applications or specific hazard rate shapes. Multivariate distributions are also often used especially in situations where two or more related failure time variables are of interest simultaneously. Situations may also arise where the failure times are related to or affected by several concomitant or regressor variables and the experimenter seeks for an appropriate regression model for best representing the data. Such situations arise, for example, in accelerated testing or medical data analyses where the covariates such as age, sex, physical conditions, and reports of pathological or other examinations are expected to affect the patients' conditions. Owing to space restriction, we shall be discussing a very few models named above though the interested readers may refer to Refs 2–5 for other related details.

As mentioned, hazard rate, which represents the instantaneous probability of failure, is an important characteristic to specify a failure model. Models with increasing hazard rate are used the most. One reason for this may be attributed to the fact that the interest often centers on a period in the life of an item or an individual over which some kind of gradual increase in the age takes place, yielding an increasing failure rate. In addition, populations that reflect a bathtub shape failure rate are sometimes purged of weak individuals, leaving a reduced population with an increasing failure rate. For example, manufacturers often use a “burn-in” process in which items are subjected to a brief period of operation before being sent to customers. In this way, flawed items that would fail very early are removed from the population, this frequently leaves a residual population in which individuals, or items or devices exhibit gradual ageing, with an increasing failure rate.

Models with constant failure rate are important and have a particularly simple structure. Models with decreasing failure rate are less common, but are sometimes used especially in situations where early failures dominate the processes. If items (or individuals) are kept over some fairly

long initial period of use, these result in the decreasing failure rate. One can also have situations where the process requires all the three shapes simultaneously giving rise to bathtub-shaped hazard rates. Models having capability of representing all three shapes simultaneously are important but usually difficult in many ways. Nonmonotone failure rate other than bathtub-shaped curves are less common but advocated sometimes in the literature [1,6].

A FEW COMMON CONTINUOUS FAILURE DISTRIBUTIONS

The Exponential Distribution

The exponential distribution has been widely used in the areas ranging from studies on the lifetimes of manufactured items [7] to research involving survival or remission times in some diseases and perhaps it has been the most widely used model in reliability studies and survival analysis in general [1,4]. The desirability of the exponential distribution is due to its simplicity and its inherent association with the theory of Poisson process [4]. The applicability of the distribution is, however, limited to some extent due to its lack of memory property which requires that previous use of an item or an organism does not affect its future use [8]. That is, if a device is functioning at time t , it is as good as a new one and the remaining life has the same exponential distribution.

The exponential distribution characterized by a constant hazard rate λ (> 0) has the probability density function given by

$$f(t) = \lambda \exp(-\lambda t), \quad t > 0. \quad (5)$$

The distribution has mean λ^{-1} and variance λ^{-2} and its survival function is available in closed form, that is, $S(t) = \exp(-\lambda t)$. An alternative formulation of the model is also sometimes used with scale parameter θ replacing the parameter λ^{-1} . The distribution when θ equals unity is called the *unit exponential distribution* that has been shown in Fig. 1. Clearly, if T has the density (Eq. 5), λT has a unit exponential distribution.

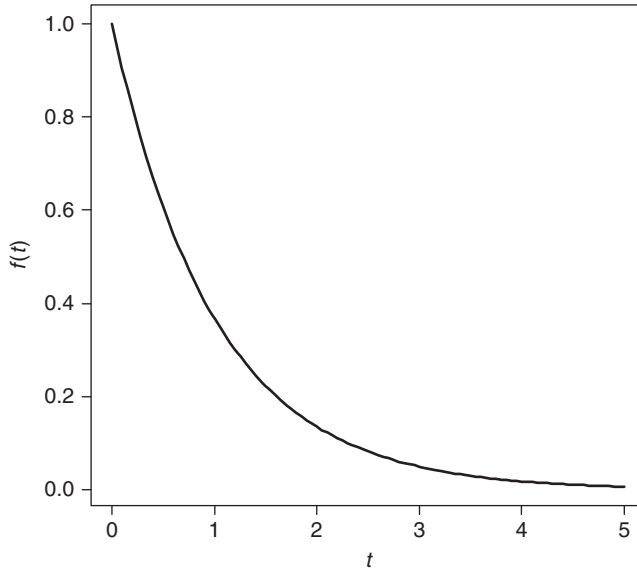


Figure 1. Probability density function of the unit exponential distribution.

Historically, the exponential distribution can be considered as the first widely used model especially in the context of failure time data analyses exactly as normal distribution does in other areas. This was partly because of the availability of simple statistical methods for it [9] and partly because the distribution appeared to provide suitable representation for the lifetimes of many things, such as various types of manufactured items [1]. The assumption of constant hazard rate is, however, a restrictive criterion and it was lately realized that many inferences are sensitive to departures from the exponentiality. This has led to greater caution in the use of the distribution.

Inferences for the exponential distribution are available in bulk both in the context of classical [1,4] and the Bayesian paradigm [10]. The obvious reason in case of former is the availability of closed form expression for the sufficient statistic of the parameter θ . The inferences for the latter paradigm are also often available in closed form or at most it may involve a single dimensional integral for some typical selection of the prior distribution for the parameter θ [10].

The Weibull Distribution

The applicability of the exponential distribution is limited because of the assumption of

a constant hazard rate, whereas the Weibull family can include increasing and decreasing hazard rates as well. The distribution is perhaps the most widely used model for lifetime data analyses. Its suitability in connection with the lifetimes of a wide variety of manufactured items, including vacuum tubes, ball bearings, and electrical insulating materials has been discussed by several authors [1,4]. The distribution has wide applicability in several biomedical studies as well. A few such situations include the time to the occurrence of tumors in human population, survival times of patients suffering from carcinoma after receiving a treatment, say chemotherapy, and so on [1,11]. The probability density function of the Weibull distribution has the form

$$f(t | \alpha, \beta) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha} \right)^{\beta-1} \times \exp \left[- \left(\frac{t}{\alpha} \right)^{\beta} \right], \quad \alpha, \beta > 0, t > 0, \quad (6)$$

where α is the scale parameter and β determines the shape of the distribution. If $\beta = 1$, the distribution reduces to the one-parameter exponential distribution and, therefore, it results in constant hazard rate. It can be easily observed that the hazard

rate of the model is decreasing (increasing) in t if $\beta < 1$ ($\beta > 1$). When $\beta = 2$, the hazard rate is a linearly increasing function of t , and the resulting distribution is known as the *Rayleigh distribution*. The survival and the hazard function of the model can be written as

$$S(t | \alpha, \beta) = \exp \left[- \left(\frac{t}{\alpha} \right)^\beta \right], \quad (7)$$

$$h(t | \alpha, \beta) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha} \right)^{\beta-1}, \quad (8)$$

respectively. The mean and variance of the distribution are

$$E(T) = \alpha \Gamma \left(\frac{\beta + 1}{\beta} \right),$$

$$V(T) = \alpha^2 \left[\Gamma \left(\frac{\beta + 2}{\beta} \right) - \Gamma^2 \left(\frac{\beta + 1}{\beta} \right) \right].$$

The shape of the density and hazard rate depends on the model parameters. The Weibull model exhibits a wide variety of shapes for the density, survival function, and the hazard function. Figure 2 shows a typical Weibull density for the case where $\alpha = 1$. The corresponding hazard function is shown in Fig. 3.

Weibull distribution is an extensively rich family in the sense that a large body of literature on both classical and Bayesian methods has successfully evolved out of it. An important reason for this being the nonexistence of any two-dimensional sufficient statistics, in general, for both α and β and, therefore, an enormous possibility exists for producing inferential procedures for the distribution. Although mathematical tractability is often a serious problem involved with the analysis of the model, it appears that the statistical procedures that are relatively easy to use are now available [1,4,12,13]—thanks to the modern computing tools that enabled the routine development of these procedures.

The Gamma Distribution

The gamma distribution is a natural extension of the exponential distribution which is often used in engineering, science, and business applications to model continuous variables that are always positive and skewed. The distribution can be derived by considering the time to occurrence of n th event in a Poisson process or, equivalently, by considering the n -fold convolution of an exponential distribution [4,14]. Actually the distribution is a continuous analog of negative binomial distribution that can be

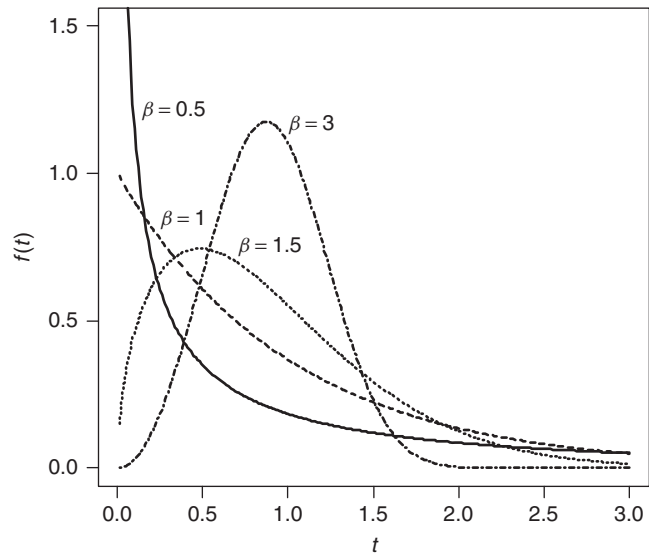


Figure 2. Probability density function of the Weibull distribution when $\alpha = 1$.

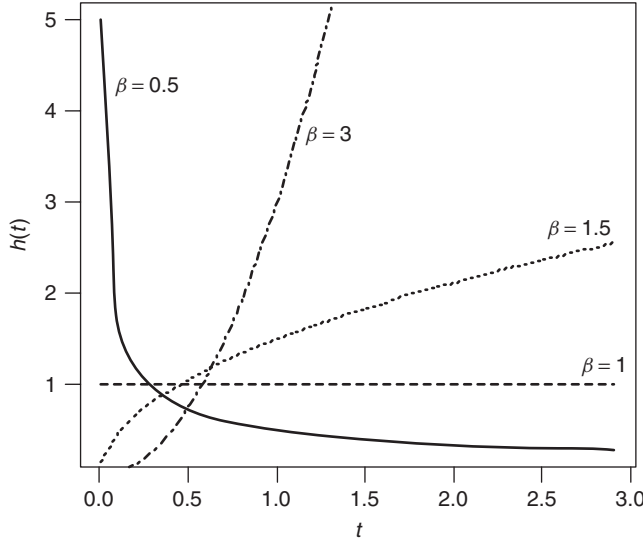


Figure 3. Hazard function for the Weibull distribution when $\alpha = 1$.

obtained by considering the sum of n variables with common geometric distribution. The probability density function of a gamma distribution with parameters α and β is given by

$$f(t | \alpha, \beta) = \frac{1}{\alpha^\kappa \Gamma(\kappa)} t^{\kappa-1} \exp\left(-\frac{t}{\alpha}\right), \quad \alpha, \kappa, t > 0, \quad (9)$$

where $\Gamma(\kappa)$ denotes the complete gamma function. The parameters α and κ are referred to as the *scale* and *shape parameters* of the gamma distribution, respectively.

The survival and the hazard functions of the gamma model are not available in closed forms unless α happens to be integer; however, they may be expressed in terms of a standard incomplete gamma function for which computer subroutines are widely available. The family includes increasing, decreasing, or constant hazard rates depending upon the values of κ . It can be seen that the hazard rate is monotone increasing for $\kappa > 1$, decreasing for $\kappa < 1$, and constant for $\kappa = 1$ and, therefore, the distribution reduces to the one-parameter exponential distribution for the last case. The expressions for the survival and the hazard functions for the

gamma distribution can be written as

$$S(t | \alpha, \kappa) = \frac{\Gamma(\kappa) - \Gamma(\kappa, t/\alpha)}{\Gamma(\kappa)}, \quad (10)$$

$$h(t | \alpha, \beta) = \frac{t^{\kappa-1} \exp(-t/\alpha)}{\alpha^\kappa [\Gamma(\kappa) - \Gamma(\kappa, t/\alpha)]}, \quad (11)$$

where $\Gamma(a, z)$ is the incomplete gamma function given by

$$\Gamma(a, z) = \int_0^z y^{a-1} \exp(-y) dy, \quad a > 0.$$

The probability density function in Equation (9) is bell shaped for $\kappa > 1$, and reverse J shaped for $\kappa \leq 1$. The mean of the distribution is $\kappa\alpha$ and its variance is $\kappa\alpha^2$ and, therefore, the parameter κ may also be interpreted as a square of the reciprocal of the distribution coefficient of variation. The probability density function of the gamma distribution is shown in Fig. 4 for various values of shape parameters, whereas the corresponding hazard rate curves are shown in Fig. 5. These curves are drawn by taking scale parameter α as unity.

Gamma distribution is closely related to chi-square distribution. We can easily see from Equation (9) that if T is distributed according to gamma distribution with shape parameter α and scale parameter unity, $2T$ has chi-square distribution with

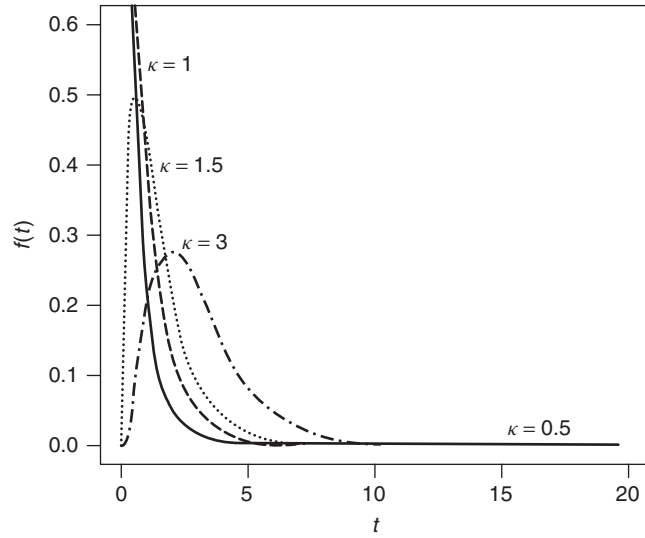


Figure 4. Probability density function of the gamma distribution when $\alpha = 1$.

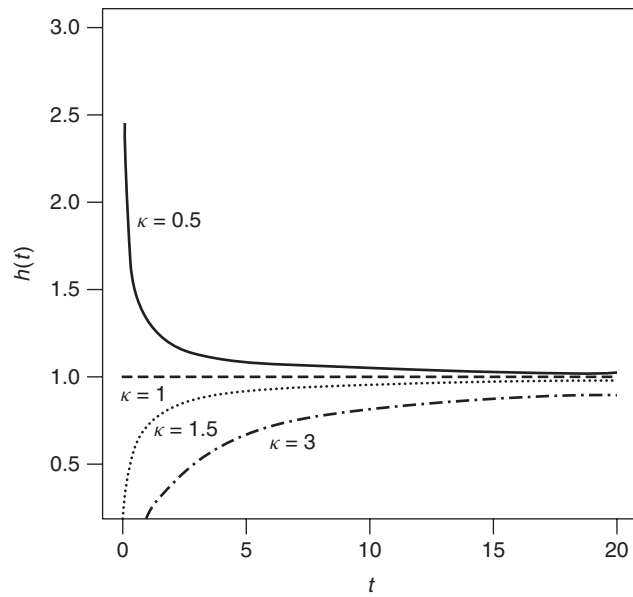


Figure 5. Hazard function for the gamma distribution when $\alpha = 1$.

2α degrees of freedom. For integer values of α , the gamma distribution is also known as *Erlang distribution*. The moments and other properties of gamma distribution can be had from Ref. 15, and the reference also provides the relationship of gamma distribution with several other important distributions.

The use of gamma distribution as a lifetime model is not as common as the corresponding Weibull model. The obvious

reason being the nonavailability of closed form expressions for the survival and the hazard functions. This, in fact, has resulted in more complicated inferential procedures as compared to the same for Weibull distribution. Undoubtedly, the credit goes to the sophisticated computational techniques, most of which were developed in the previous decades especially after the advent of modern computing devices. The classical developments related to the model can

be found in Refs 1 and 4, whereas the corresponding Bayesian results can be seen in Refs 10 and 13 among others.

The Lognormal Distribution

The lognormal distribution initially received little attention in the statistical literature due to its limited applications in some rare situations. However, its importance was lately realized and it became an important model particularly in the area of life testing and survival analysis in spite of its one unattractive feature concerning the hazard rate. References 1, 4, and 12 are worth mentioning texts that detail a systematic accountability of various developments and provide some good references related to the model.

The logarithmic normal distribution implies that the logarithms of the lifetimes are normally distributed. That is, when the random variable T follows two-parameter lognormal distribution with parameters μ , and σ , $\ln(T)$ follows normal distribution with mean μ and variance σ^2 . Thus, the probability density function of T can be written as

$$f(t | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma t} \exp \left[-\frac{1}{2\sigma^2} (\ln t - \mu)^2 \right];$$

$$0 < t < \infty, -\infty < \mu < \infty, \sigma > 0, \quad (12)$$

where $\exp(\mu)$ and σ are the scale and shape parameters, respectively [1].

The hazard rate of the lognormal distribution, as a function of time, is an increasing function followed by a decreasing function, and can be shown to approach to zero for large lifetimes and also at the initial time [4]. Although this feature appears unattractive, the model is found suitable especially when large values of lifetimes are not of interest and the early failures or occurrences dominate. Some of the general properties of this distribution can also be found in Refs 15–17. The expressions for the survival and the hazard functions of the model are not available in closed forms and involve the standard normal distribution function. These expressions can be written as [10]

$$S(t | \mu, \sigma^2) = 1 - \Phi \left(\frac{\log t - \mu}{\sigma} \right), \quad (13)$$

$$h(t | \mu, \sigma^2) = \frac{\phi \left(\frac{\log t - \mu}{\sigma} \right)}{\sigma t - \Phi \left(\frac{\log t - \mu}{\sigma} \right)}, \quad (14)$$

where $\phi(\cdot)$ denotes the standard normal probability density function and $\Phi(z)$ denotes the cumulative distribution function given by

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{x^2}{2} \right) dx.$$

The mean and variance of the distribution are

$$E(T; \mu, \sigma^2) = \exp \left(\mu + \frac{\sigma^2}{2} \right),$$

$$V(T; \mu, \sigma^2) = (\exp(2\mu + \sigma^2))(\exp(\sigma^2) - 1),$$

respectively. Figure 6 shows a typical plot of lognormal probability density function for different values of its shape parameter σ . The corresponding plot for the hazard rate curves is shown in Fig. 7. The parameter μ is taken to be zero throughout since its basic role is to change the scale of the distribution on the time axis and not the shape of the distribution [1]. The moments and other properties of the distribution can be had from Ref. 15.

The inferential procedures for the model written in the (Eq. 12) are well understood from the standpoint of each of the major systems of statistical inferences, since by transforming to logarithms the problem can be reduced to the inferences from simple normal distribution and most of the procedures are routine. References 1 and 4 are the good references where classical procedures have been detailed for the lognormal distribution. Bayes procedures for the two-parameter model are given in Martz and Waller [10] using different prior combinations but most of the results directly follow from the normal distribution using the appropriate transformation. Needless to mention the fact that log lifetimes are normally distributed is often a convenience and is not much justified in the modern computing days.

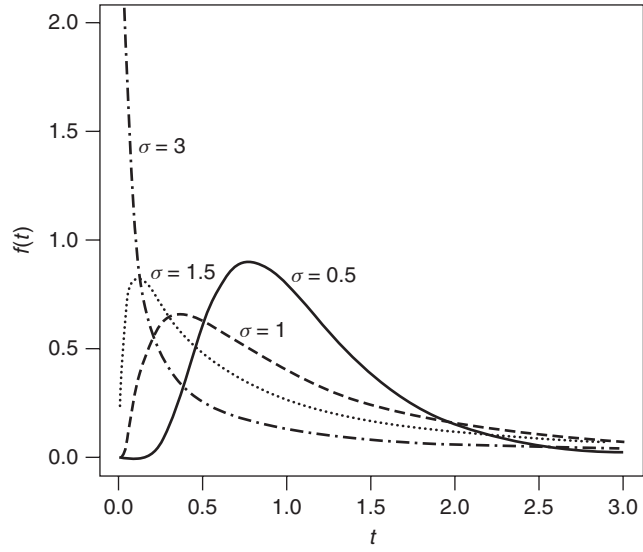


Figure 6. Probability density function of the lognormal distribution when $\mu = 0$.

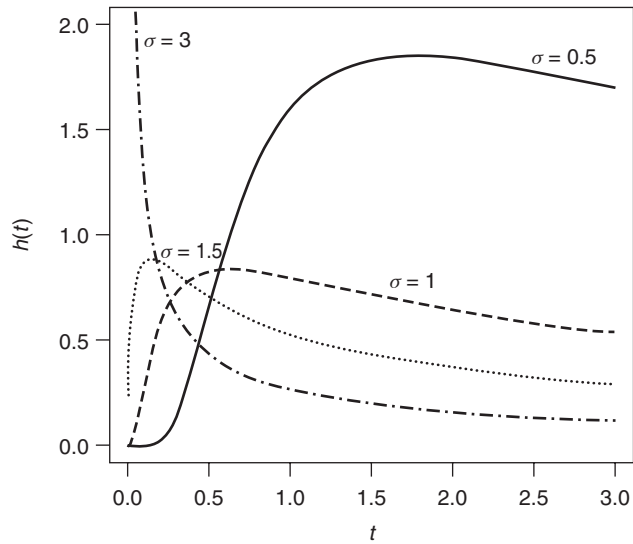


Figure 7. Hazard function for the lognormal distribution when $\mu = 0$.

The Generalized Gamma Distribution

The generalized gamma distribution is a fairly flexible three-parameter family that was first proposed in Ref. 18 and later on independently in Ref. 19. Its probability density function is given by

$$f(t | \alpha, \beta, \kappa) = \frac{\beta}{\Gamma(k)} \frac{t^{\beta\kappa-1}}{\alpha^{\beta\kappa}} \exp \left[- \left(\frac{t}{\alpha} \right)^\beta \right];$$

$$t > 0, \alpha > 0, \beta, \kappa > 0, \quad (15)$$

where α is the scale parameter and both β and κ determine the shape of the distribution. Sometimes $\beta\kappa$ is jointly referred to as the *power parameter*. As the name itself makes it obvious, this is a generalized family and includes as special cases all the important failure time distributions. Thus, it reduces to exponential distribution when both β and κ are set to unity. The Weibull distribution becomes a special case for $\kappa = 1$, and gamma as a special case for $\beta = 1$. In addition, the

lognormal distribution results as a limiting case when $\kappa \rightarrow \infty$. Like gamma distribution here also the survival and the hazard functions are not available in closed forms and involve the incomplete gamma integral. The expression for the survival function can be written as

$$S(t | \alpha, \beta, \kappa) = \frac{\Gamma(\kappa) - \Gamma(\kappa, (t/\alpha)^\beta)}{\Gamma(\kappa)}, \quad (16)$$

and the expression for the corresponding hazard function can be explicitly obtained using Equation (2).

Inferences for the generalized gamma distribution are even more complex due to involvement of additional shape parameter [1]. Much of the difficulties occurred because the usual classical techniques failed to produce any significant solution with the model written in form (15) mainly because of the complex nature of the corresponding likelihood function for any relevant sampling plan. Large sample methods are also found to be unsatisfactory because the limiting normal approximation approaches very slowly. An alternative reparameterized form of the model is given in Ref. 1 that simplifies matter to some extent but definitely does not offer the routine inferences. Bayes inferences are even more unmanageable for any prior-likelihood combination unless the work of Upadhyay and others [20] that appeared only recently (see also Ref. 13).

The distributions given above can be extended with an additional parameter λ known as the *threshold parameter* or the *guarantee time parameter*. The probability density functions can then be written by replacing t with $t - \lambda$ and the range of t then becomes from λ to ∞ , that is, no failure can occur prior to λ . Such models are typically nonregular and the inferences can be even daunting if the parameter λ is assumed to be unknown. Two-parameter exponential distribution may be an exception with regard to its inferential developments.

Other Continuous Distributions

Besides the above important models, there are several other continuous models that have received attention in a variety of

applications of failure time data analyses. Owing to space restriction, we shall not go into the details although a brief discussion supported with a very few important references will be provided for completeness. A few such models can be named as inverse Gaussian distribution [21], log-logistic distribution, Gompertz distribution, Gumbel distribution, Birnbaum–Saunders distribution, and Pareto distribution [1,4]. One can also consider a normal distribution either truncated on left at zero or defined in a way such that its negative region can have negligibly small probability.

Mann *et al.* [4] obtained a general time to failure distribution using a pure death process especially when a unit can suffer either a random or a wear-out failure. Models under the name of mixed or composite distributions have been discussed by several authors in order to allow more flexibility in fitting and explaining the failure data [1,4]. Models capable of representing bathtub and nonmonotone hazard shapes [22] have also been discussed exclusively although a practical difficulty is that these models are often quite difficult to handle statistically. A few models discussed above may be used for characterizing such hazard rate shapes [1] although there has been a practice in the recent years to generalize the existing models by introducing extra parameters so that they are capable of representing nonmonotone and bathtub hazard rate shapes [23].

Most of the models discussed above can be extended to include regressor variables resulting in what we call the regression models in failure time data analyses. Some specific models such as the proportional hazards [24] have been detailed in the literature and these received wide applicability in both the reliability and the survival analyses contexts [1]. Multivariate models in failure time data analysis have also been reported although not as elaborately discussed as the univariate models. The last category of models may arise, for example, in a situation where one can desire to model the joint failure time distribution of parts of an integral system consisting of several parts. Several other important situations for the use of

multivariate models are discussed in Ref. 25 among others.

A FEW COMMON DISCRETE FAILURE DISTRIBUTIONS

Discrete distributions are often used in the failure time data analyses when the random variables are available either in the form of categorical data or in the form of binary responses. Most of these distributions are fundamental statistical distributions in the sense that they have not been derived either using any physical processes or using any transformations on some directly or indirectly obtained random variables. A few such distributions include the geometric distribution, the binomial distribution, the negative binomial distribution, the hyper geometric distribution, the Poisson distribution, and the multinomial distribution, the last one can be considered to be a multivariate analog of the binomial distribution. Let us consider, for instance, a group of patients suffering from gall bladder diseases and one is willing to know the probability that how many of these have developed carcinoma. Obviously, binomial model can be a desired possibility. Similarly, consider a situation where a device is operating and every small period of the operation is identified as a trial. Thus, we may have trials, similar to Bernoulli, with failure-free operation or operation with failure. Suppose $X - 1$ is the number of successive periods of failure-free operation and X is the period at which failure occurs for the first time then X can be considered to have geometric distribution. The multinomial distribution arises, for example, when continuous data have been grouped in a number of classes and each class leads to certain frequencies. An example to this effect can be found in Ref. 1. Situations can be similarly noticed where binomial, Poisson, and other discrete distributions named above can be considered as appropriate distributions. References such as 1 and 4 can provide a few of such situations where these distributions may be important candidates. Reference 26 provides a detailed discussion of these distributions with several important properties.

REFERENCES

1. Lawless JF. Statistical models and methods for lifetime data. New York: Wiley; 1982.
2. Crowder MJ, Kimber AC, Smith RL, *et al.* Statistical analysis of reliability data. London: Chapman & Hall; 1991.
3. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley; 1980.
4. Mann NR, Schafer RE, Singpurwalla ND. Methods for statistical analysis of reliability and life data. New York: Wiley; 1974.
5. Shooman ML. Probabilistic reliability: an engineering approach. New York: Mc Graw-Hill; 1968.
6. Aalen OO, Gjessing HK. Understanding the shape of the hazard rate: a process point of view. *Stat Sci* 2001;16:1–22.
7. Epstein B. The exponential distribution and its role in life-testing. *Ind Qual Contr* 1958;15:2–7.
8. Sinha SK. Reliability and life testing. New Delhi: Wiley; 1986.
9. Epstein B, Sobel M. Life testing. *J Am Stat Assoc* 1953;48:486–502.
10. Martz HF, Waller RA. Bayesian reliability analysis. New York: Wiley; 1982.
11. Qian J. A Bayesian Weibull survival model [Unpublished PhD thesis]: Department of Statistics, Duke University, U.S.A.; 1994.
12. Bain LJ. Statistical analysis of reliability and life-testing models. New York: Marcel Dekker; 1991.
13. Singpurwalla ND. Reliability and risk: a Bayesian perspective. Chichester: Wiley; 2006.
14. Feller W. Volume 1, An introduction to probability theory and its applications. New York: Wiley; 1968.
15. Johnson NJ, Kotz S. Volumes 1 & 2, Continuous univariate distributions. Boston (MA): Houghton Mifflin; 1970.
16. Aitchison J, Brown JAC. The lognormal distribution. Cambridge: Cambridge University Press; 1957.
17. Hill BM. The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. *J Am Stat Assoc* 1963;58:72–84.
18. Stacy EW. A generalization of the gamma distribution. *Ann Math Stat* 1962;33:1187–1192.
19. Cohen AC. A generalization of Weibull distribution. Marshall Space Flight Centre, NASA

- Contractor Report No. 61293, NAS 8-11175; 1969.
20. Upadhyay SK, Vasistha N, Smith AFM. Bayes inference in life testing and reliability via Markov chain Monte Carlo simulation. *Sankhya Ser A* 2001;63:15-40.
 21. Chhikara RS, Folks JL. The inverse gaussian distribution. New York: Marcel Dekker; 1989.
 22. Murthy VK, Swartz G, Yuen K. Realistic models for mortality rates and their estimation, I and II. University of California at Los Angeles, Department of Biomathematics, Technical Report; 1973.
 23. Murthy DNP, Xie M, Jiyang R. Weibull models. New Jersey: Wiley; 2004.
 24. Cox DR. Regression models and life tables. *J R Stat Soc [Ser B]* 1972;34:187-220.
 25. Hougaard P. Analysis of multivariate survival data. New York: Springer; 2000.
 26. Johnson NJ, Kotz S. Discrete distributions. Boston (MA): Houghton Mifflin; 1970.

COMMON RANDOM NUMBERS

LEE W. SCHRUBEN

Department of Industrial Engineering
and Operations Research, University
of California, Berkeley, California

COMMON RANDOM NUMBERS AND SIMULATION RUNS

Random phenomena are conventionally modeled in stochastic computer simulations and Monte Carlo experiments using transforms of one or more sequences (“streams”) of (pseudo)random numbers. Random numbers are regarded as observations of mutually independent and identically distributed uniform random variables strictly between 0 and 1.

A run of a stochastic model with feasible input factor values x transforms sets of random numbers U into samples of random variables or sample paths of stochastic processes $V(x)$, which are in turn transformed by the simulation code into random output data $Y(x)$. This output data is then used to compute estimators $\hat{\theta}(x)$ of interesting system characteristics $\theta(x)$, called the *simulation response* at x . A generic simulation run is

$$U \rightarrow V(x) \rightarrow Y(x) \rightarrow \hat{\theta}(x). \quad (1)$$

The fundamental idea of the variance reduction technique (VRT) of common random numbers (CRNs) is simply to reuse some or all of the random number streams to generate the same or different random processes for two or more runs that do not have identical inputs. U depends on x and the simulation run (1) becomes

$$U(x) \rightarrow V(x) \rightarrow Y(x) \rightarrow \hat{\theta}(x). \quad (2)$$

Some other VRTs including antithetic and control variates also fall within this framework and are collectively called *correlation*

induction methods. Of course, using the same streams to model *all* of the same random processes in two runs with the identical input factor settings, initializations, and durations will produce the same outputs. Something has to be different.

Representations (1) and (2) for a simulation run are abstract and general. Here we consider only simulation experiments in which *runs* are the observational units. (An example of a simulation experiment that is not run-oriented is a frequency domain screening experiment [1].) A simulation run involves more than merely choosing the input factor settings. For example, there are many important decisions that need to be made: how many replications should be run at a particular input setting; how a run should be initialized; how long it should be warmed up before collecting data; and when a run should be terminated [2]. No attempt is made here to address the many details that must be attended to when running a simulation.

The notion here of a simulation input factor is also abstract and general. Using a simulation, the experimenter conceptually can control and observe everything, depending on the flexibility of the simulation modeling software used. Unlike conventional experiments, a simulation input factor can be anything, which may or may not be controllable in the real-world system being modeled. For example, the distributions of random arrivals of customers and service times for a simulated queue are input factors that can be easily changed by the experimenter. In the real-world, these may be unknown and at best only partially controllable or observable. The numbers and types of servers are input factors for both simulations and real-world experiments. An experimenter using a simulation can fully control all inputs and even change them *during* a simulation run! Again, this power may be restricted somewhat by the simulation software. The simulation itself may actually be a collection of different computer programs that model identical, similar, or alternative systems, designs, and/or

operating policies. Thus different input factor settings may model very different systems with the same computer program, or they may model the same system using different simulation computer programs.

COMMON RANDOM NUMBERS AND SIMULATION EXPERIMENTS

At its most basic level, a good simulation experiment is designed and run to try to obtain estimators that have both low bias and low variance, and allows statistical assessment of the quality of the estimators. Designing a simulation experiment consists of carefully choosing sets of input factor settings, called *design points*. The experiment should be based on principles of statistical design of experiments (DOE) [3]. From a classical DOE viewpoint, the random number streams chosen for the runs in the experiment can be regarded as a nuisance factor. An experimental design can be blocked to remove the extraneous effects of random number streams. This can be done in an optimal manner to reduce the variances of interesting effects [4–7]. However, it can also complicate the analysis of the output [8–10].

Using standard DOE notation, the elements of the experimental design matrix X will have as its rows the values of scaled functions of the input factor settings for each design point. The factor settings in each run are the columns of X . The first column of X may all be 1's to include the mean effect in a regression meta-model that is to be fit to the output from the experiment. The setting for input factor i for simulation run j is denoted as $x_{i,j}$. (When there is a single subscript on x_i it will denote an experimental design point, input settings, at which there may be multiple simulation runs.)

When a simulation is run at input setting x_{ij} , $U_k(x_{ij})$ is any set of streams of random numbers that are sufficient to generate the random processes for the run. A more general definition of CRNs than typically found in the literature is running any simulation experiment where at least two streams are reused. We can use random number streams from the same sets of streams to generate some of the

same or different processes in the same or different runs at the same or different design points, as long as these are not *all* the same (which would produce identical output).

Definition : $CRN \Leftrightarrow \exists k, k', i, j, i', j' :$

$$\begin{aligned} U_k(x_{ij}) \cap U_{k'}(x_{i'j'}) &\neq \emptyset \\ \forall k \neq k' \vee i \neq i' \vee j \neq j' \end{aligned} \quad (3)$$

We can use any subset of random number streams in different, or the same, runs even if they are at the same design point as long as they are not used to generate all the same random processes. This broader definition also includes antithetic and control variate VRTs and includes interesting variations like that in Refs 8, 11 and 12.

The intuition behind the CRN simulation strategy is the same as for any real laboratory experiment that compares two or more systems. Simply, subject the tests of the different systems to common sources of extraneous randomness that otherwise merely contaminates the measurements. That is, try to observe the different systems in the same controlled environments. There is some sound theory behind this practice, which at times has been misinterpreted. The most straightforward result is the fact that when the variance of the difference between the response estimators for two different systems $D(x_i, x_j) = \hat{\theta}(x_i) - \hat{\theta}(x_j)$ is

$$\begin{aligned} \text{Var}[D(x_i, x_j)] &= \text{Var}[D(x_j, x_i)] = \text{Var}[\hat{\theta}_i - \hat{\theta}_j] \\ &= \text{Var}[\hat{\theta}_i] + \text{Var}[\hat{\theta}_j] - 2\text{Cov}[\hat{\theta}_i, \hat{\theta}_j] \\ &= \frac{\text{Std}[\hat{\theta}_j]}{\text{Std}[\hat{\theta}_j]} + \frac{\text{Std}[\hat{\theta}_i]}{\text{Std}[\hat{\theta}_j]} - 2\text{Corr}[\hat{\theta}_i, \hat{\theta}_j]. \end{aligned} \quad (4)$$

When the two system responses are run using independent (disjoint sets) of random number streams, the correlation is zero. But, if CRN is used, *and* this results in $\text{Corr}[\hat{\theta}(x_i), \hat{\theta}(x_j)] > 0$, then the variance of their difference is reduced from its value for independent sampling by twice the estimator covariance. CRNs are perfectly correlated ($\text{Corr} = 1$). Since using CRNs do not induce a bias in the response, the hope is that when CRN streams are used, then the estimated responses will still have significant positive

correlation after the sequence of transformations in Equation (1). Unfortunately, this is sometimes not the case, particularly when the systems are far apart according to some metric on the x 's.

CRN Synchronization

A loss in induced correlation magnitude occurs when simulation runs require different numbers of random numbers, and the input random number streams get out of synchronization. Then the same random numbers may be used to model different random processes in the different simulated systems. For example, in comparing two simple simulated queueing systems—like Example 11.1 in the textbook given in Ref. 2—it may be that the same random number that was used to generate a time between customer arrivals in one system is used to generate a service time in the other simulation. These two random variables have opposite effects on the level of congestion (shorter service times result in smaller lines, and shorter times between customer arrivals result in longer lines). Blindly using CRNs hoping to highlight the differences in these systems might be detrimental, resulting in increased variance of the estimator for the differences in performance between the systems—ultimately resulting in a lower probability of being able to identify the better system.

There are several suggestions for improving the chances of inducing positive correlations in simulation estimators. The simplest is to use the same streams to generate the same stochastic processes in runs of different systems [2]. For example, the same stream used to generate the arrivals in one queueing system is used to generate the arrivals in the other system. A different set of streams is reused to generate the service times for both systems. In general, distinct random number streams can be used to simulate components or processes in the simulation models that are identical or closely related, and independent streams (re)used to generate processes that are very different. Table 11.3 on page 587 in Ref. 2 demonstrates that using one stream of random numbers

to generate customer arrivals and a separate stream to generate service times when comparing two queueing systems effectively reduces the variance of the differences in their estimated performance. The result is a significant improvement in the ability to identify the better of two competitive systems. (This experiment was replicated as a homework problem producing similar, but notably less dramatic, results.)

However, in all but the simplest models, it is not easy to keep random number streams synchronized across simulation runs. We will see there is a simpler approach later.

Simulation Experiments

A common goal for a simulation experiment is to fit an additive noise regression meta-model to the response

$$Y = X\theta + \varepsilon(X), \quad (5)$$

where ε is some random noise in the output. There are many alternative, more sophisticated, meta-models and simulation study objectives.

When estimating the average (or equivalently the sum) of two responses for the same or different systems, the positive correlation induced by using CRNs will *increase* the variance since the sign of last term in the variance of the sum, corresponding to Equation (4), would be positive.

Therefore it is important to know what effects in x are of interest. To illustrate, consider the simple case where we are interested in estimating the coefficients of a homoscedastic (constant variance) linear additive-noise response meta-model to a one-dimensional input factor. A simple model is given by

$$Y(x) = \theta_0 + \theta_1 x + \varepsilon(\sigma). \quad (6)$$

Here the system response estimator is three dimensional: the mean effect or intercept $\hat{\theta}_0$, the main effect or slope $\hat{\theta}_1$, and the noise standard deviation $\hat{\sigma}$. We will assume that the noise can be estimated from the pooled outputs from the runs using, for example, standardized time series or batched means [2]. Therefore, we make only

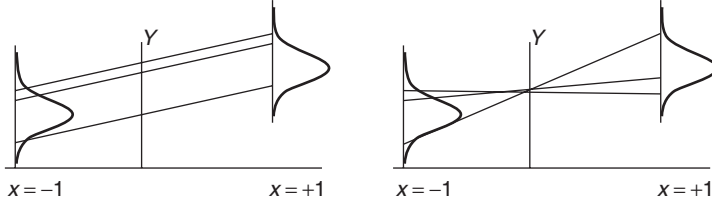


Figure 1. Three possible experiments to fit response model (6) using CRNs and antithetic CRNs (b).

two runs, one at a low value of x and one at a high value of x , which are scaled to be $x_1 = -1$ and $x_2 = +1$. We run the simulation and get the outputs $Y(x_1)$ and $Y(x_2)$. We will denote the common variances of these outputs as σ^2 and the *magnitude* of their correlation (if any) by ρ .

The ordinary least squares (we are assuming that the noise is not dependent on x) estimators are

$$\hat{\theta}_0 = \frac{1}{2}[Y(x_1) + Y(x_2)] = (.5, .5) \begin{pmatrix} Y(x_1) \\ Y(x_2) \end{pmatrix} \quad (7)$$

and

$$\hat{\theta}_1 = \frac{1}{2}[Y(x_1) - Y(x_2)] = (.5, -.5) \begin{pmatrix} Y(x_1) \\ Y(x_2) \end{pmatrix}.$$

The variances of these unbiased estimators are given by

$$\text{Var}[\hat{\theta}_0] = \frac{\sigma^2}{2}(1 + \rho) \quad (8)$$

and

$$\text{Var}[\hat{\theta}_1] = \frac{\sigma^2}{2}(1 - \rho).$$

We see that successful CRNs ($\rho > 0$) reduce the variance of the estimator of the slope but *increase* the estimator of the intercept by *exactly the same amount*. There is no overall net gain in variance reduction by inducing correlations. This is one reason that correlation induction has sometimes been called a *Monte Carlo swindle*. Using antithetic CRNs (where CRNs are subtracted from 1 before use) to try to induce negative correlations in outputs from the two runs has the opposite effect. The variance of the intercept is reduced, but the variance of the slope is increased by the exactly same amount; so, again, there is no net gain in variance reduction.

Intuitively, we can visualize the Monte Carlo swindle going on here in several ways. Figure 1 shows three possible independent outcomes of experiments to estimate the response model given by Equation (6). In the left frame of Fig. 1, a successful application of CRNs was used for each pair of runs in the three different experiments. We see that the pairs of values of the outputs Y for each experiment tend to occur in the same tails of their respective distributions. This results in low observed variability in the three estimated slopes, but high variability in the three intercepts.

In Fig. 1b, *antithetic* CRNs (where each random number is subtracted from 1 before use) were successfully applied so that values of Y tend to occur in the opposite tails of their respective distributions. This results in low variability in the three estimated intercepts, but high variability in the three estimated slopes.

Using the vector inner product representations of the estimators in (7), we can view the estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ for this model geometrically as the lengths of the projections of the output vector $(Y(x_1), Y(x_2))$ onto the lines in the $(0.5, 0.5)$ and $(0.5, -0.5)$ directions. Figure 2 shows an ellipsoid of concentration (proportional to contours of the distribution) of the output vector if these are assumed normal with positive correlations.

Increasing positive correlation causes these contours to become more and more elliptical with a 45° slope for the major axis (this slope is 45° because of the equal estimator variances). The longer possible lengths of the projections of the output vector on the $(0.5, 0.5)$ line cause variation in the average or intercept estimator $\hat{\theta}_0$. The positive correlations in the outputs simultaneously make the lengths of the possible projections of the output vector on the $(0.5, -0.5)$ line,

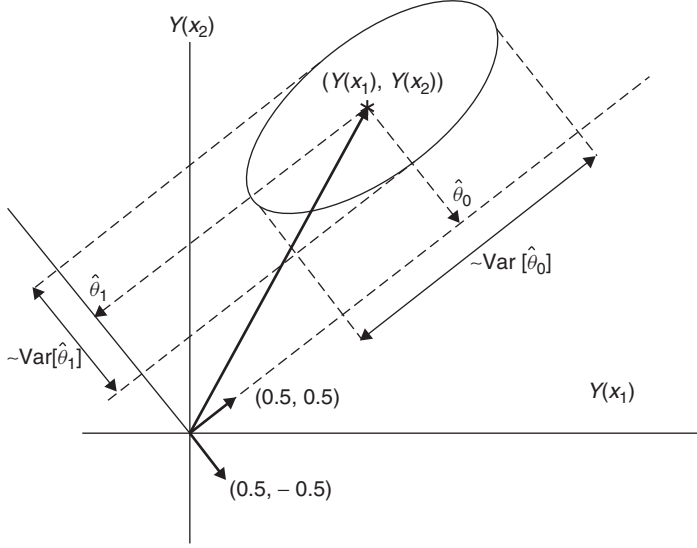


Figure 2. Elliptical contours of the output distributions with positive correlation.

that is, the slope $\hat{\theta}_1$, shorter, resulting in smaller sample variances. If the outputs were negatively correlated, the major axis of the contours would have a -45° slope with the opposite effects on the variability of $\hat{\theta}_0$ and $\hat{\theta}_1$. Figure 2 is complicated, but it is worth studying carefully.

Next consider a system with two input factors where we wish to estimate the meta-model

$$Y(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \varepsilon(\sigma) \quad (9)$$

using a 2^2 full factorial design [3] running two low and two high values of each factor. The design is pictured in Fig. 3, where the same set of CRNs are used in all four runs but their antithetic sets are used in the two orthogonal fractional designs (at opposite corners).

Using the same notation as before, we find that the least square estimator variances are now given by

$$\begin{aligned} \text{Var}[\hat{\theta}_0] &= \frac{1}{4}\sigma^2(1 - \rho) \\ \text{Var}[\hat{\theta}_1] &= \frac{1}{4}\sigma^2(1 - \rho) \\ \text{Var}[\hat{\theta}_2] &= \frac{1}{4}\sigma^2(1 - \rho). \end{aligned} \quad (10)$$

This time, the variances of *all three* estimators are reduced. There appears to be no Monte Carlo swindle. The missing variance

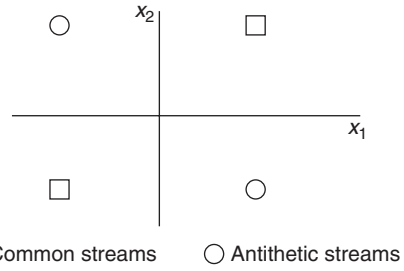


Figure 3. A 2^2 factorial design with common and antithetic $1/2$ fractional blocks.

can be found by considering another possible meta-model that saturates (having the same number of unknown θ s as runs) the experimental design. This model adds an interaction effect between the two factors:

$$Y(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2 + \varepsilon(\sigma). \quad (11)$$

The variances of the least squares estimators for this model are

$$\begin{aligned} \text{Var}[\hat{\theta}_0] &= \frac{1}{4}\sigma^2(1 - \rho) \\ \text{Var}[\hat{\theta}_1] &= \frac{1}{4}\sigma^2(1 - \rho) \\ \text{Var}[\hat{\theta}_2] &= \frac{1}{4}\sigma^2(1 - \rho) \\ \text{Var}[\hat{\theta}_{12}] &= \frac{1}{4}\sigma^2(1 + 3\rho) \end{aligned} \quad (12)$$

The Monte Carlo swindle is back: all of the variance that was taken from the first three estimators in the additive meta-model is in the variance of the interaction estimator. So, considering the saturating meta-model, the there is no net variance reduction.

These observations are more general: For saturated orthogonal experimental designs, the sum of the variances of the independently estimable parameters is a constant equal to the sum of the variances of the observations, regardless of the correlations. This is sometimes called the *total variance* in $(\hat{\theta}_0, \hat{\theta}_1)$ and is the sum of the eigenvalues of its dispersion matrix. The Monte Carlo swindle is a result of the fact that saturating an orthogonal experimental design makes the design matrix X square (the number of estimators and the number of runs n are equal). Therefore, because the rows are also orthogonal

$$XX' = X'X = nI_n. \quad (13)$$

Here I_n is the identity matrix. Let D denote the variance-covariance matrix of the simulation outputs (the $Y(x_i)$'s). Let C denote the variance-covariance matrix of all parameter estimators (the $\hat{\theta}_i$'s). Then the trace, $\text{Tr}[C]$, is the sum of the variances of the estimators. For generalized least squares (also for ordinary least squares when the result in Ref. 13 applies)

$$\begin{aligned} \text{Tr}[C] &= \text{Tr}[(X'D^{-1}X)^{-1}] = 1/n \text{Tr}[D] \\ &= 1/n \sum_{j=1}^n \sigma_j^2. \end{aligned} \quad (14)$$

where σ_j^2 is the variance of the output from the j th run, $Y(x_{ij})$.

Thus the sum of the variances of the estimators computed from the simulation experiment will be equal to the sum of the variances of the outputs from the simulation runs, regardless of their induced correlations. One can induce correlations between the simulation outputs to reallocate the estimator variances to uninteresting estimators (like the interaction term in model (8), but this will not reduce the total variance.

The only possible benefit to using CRN in such situations is if there is a factor, like the

interaction term in model (8), that we are willing to assume is zero or at least of much less important than the other effects. We can then deploy CRNs and their antithetic sets to simply maximize the variance of this unimportant estimator [14].

An obvious strategy then is to add enough uninteresting effects to a meta-model in order to saturate an orthogonal experimental design. Then assign the correlations to maximize the sum of their variances, thus shifting variance away from interesting estimators. This is a more general and simpler strategy for applying CRNs than those in the references and includes many of the optimal rules as special cases.

COMMON RANDOM VARIABLES

When full CRN simulations are run, $U_{ijk} = U_{nmk} \forall x_j \neq x_m$, the random numbers are identical and hence have a perfect correlation of +1 (antithetic CRN streams have perfect negative correlation of -1). When these random numbers are filtered through a simulation run abstracted by Equation (2), the outputs inevitably lose some of this correlation, (unless, say, all the transformations are linear).

A major reason for loss of correlation in simulation outputs comes during the transformation from the uniform random number streams into random variables ($U \rightarrow V$). The maximal correlation between two random variates is obtained by using the inverse CDF transforms with the same random number, $(x, y) = (F_X^{-1}(u), F_Y^{-1}(u))$, which is typically not linear unless uniform random variates are being generated by shifting and/or scaling the random numbers [15]. Furthermore, for the most interesting random variates, these inverse distribution functions need to be approximated and there are faster and exact generation algorithms. Except for a few special random variable generation algorithms, more than one random number is needed for each random variable. For example, in an acceptance/rejection algorithm for generating random variables, a random number of random numbers is used to generate each random variable [2].

Using the general definition of CRNs in Equation (3), it is possible to improve the correlation between generated random variables using acceptance/rejection. For example, one stream of random numbers can be used to nominate values of a random variable and another can be used to test against the acceptance level. The two streams can be easily synchronized at every initiation of algorithm for generating correlated pair variates by discarding unused variates in the shorter stream (corresponding to the streams used to generate the beta variate that needed the fewer streams). An alternative is reported in Ref. 16.

To illustrate this idea, consider generating beta random variables using an acceptance/rejection algorithm. The beta is a popular random variable choice for stochastic simulation since it can take on a wide variety of shapes useful for sensitivity analysis. Some examples of beta variate shapes, controlled by two shape parameters, are given in Fig. 4.

Since the number of random numbers used to generate each beta variate is random, it is very hard to generate correlated

beta variates unless they have the identical shapes and differ at most by a linear transformation (then they are perfectly correlated). When the distributions have even slightly different shapes, the observed correlation between pairs of beta variates using CRNs is essentially not increased if at all. In an experiment, the estimated correlations between 10,000 CRN pairs of betas with the different shapes in Fig. 4 were rarely outside the range from -0.01 to +0.01, with the largest positive correlation only 0.07, essentially zero [16].

To improve the correlation between pairs of random variables, one stream is used to nominate values and the other for testing acceptance. When the betas have the same shape, the generated values are, of course, the same. However, when their shapes are different, the induced correlation degrades only marginally. Table 1 (also from Ref. 16) shows the estimated correlations between 10,000 pairs of beta random variables with the different shapes in Fig. 4. There is a dramatic improvement over the lack of correlations found using straightforward CRNs.

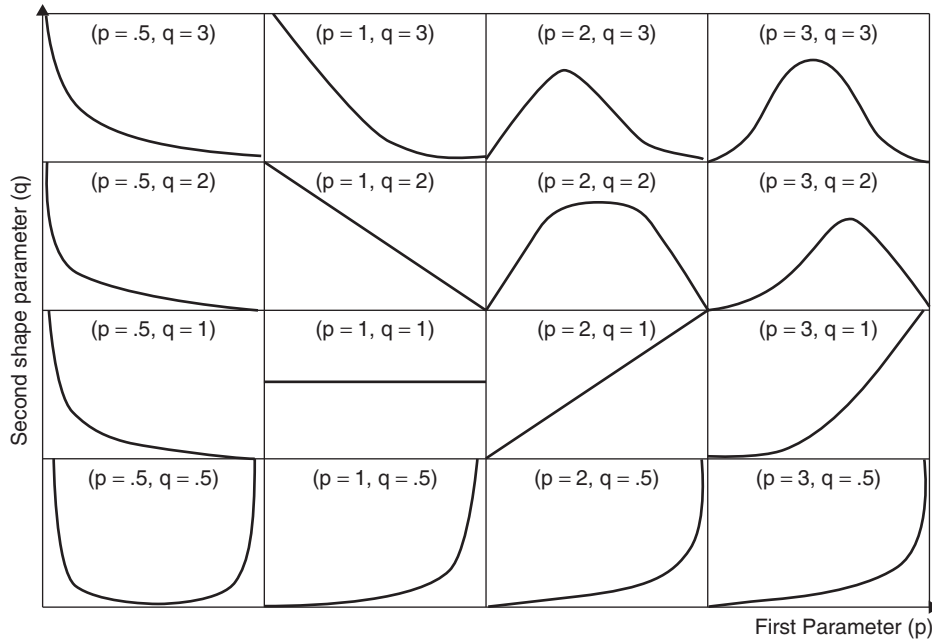


Figure 4. Beta random variates with different shape parameters [17].

Table 1. Correlations for CRN Beta Pairs with Separate Streams for Nomination and Acceptance [24]																		
	Beta (p,q)	0.5, 0.5	0.5, 1.0	0.5, 2.0	0.5, 3.0	1.0,0.5	1.0,1.0	1.0,2.0	1.0,3.0	2.0,0.5	2.0,2.0	2.0,3.0	3.0,0.5	3.0,0.5	3.0,1.0	3.0,1.0	3.0,0.2	3.0,3.0
α	0.5, 0.5	1	0.73	0.59	0.54	0.73	0.69	0.79	0.75	0.59	0.81	0.75	0.76	0.53	0.53	0.76	0.76	0.76
	0.5, 1.0	0.73	1	0.79	0.71	0.62	0.47	0.63	0.72	0.51	0.59	0.51	0.52	0.47	0.47	0.57	0.52	0.52
	0.5, 2.0	0.59	0.79	1	0.88	0.52	0.38	0.48	0.54	0.44	0.48	0.4	0.41	0.4	0.4	0.47	0.42	0.41
	0.5, 3.0	0.54	0.71	0.88	1	0.48	0.33	0.42	0.47	0.4	0.43	0.36	0.36	0.37	0.37	0.43	0.38	0.37
	1.0, 0.5	0.73	0.62	0.52	0.48	1	0.48	0.59	0.58	0.78	0.63	0.51	0.53	0.69	0.69	0.72	0.53	0.53
	1.0, 1.0	0.69	0.47	0.38	0.33	0.48	1	0.75	0.66	0.37	0.74	0.89	0.85	0.32	0.32	0.64	0.86	0.86
	1.0, 2.0	0.79	0.63	0.48	0.42	0.59	0.75	1	0.87	0.47	0.68	0.78	0.84	0.42	0.42	0.61	0.74	0.78
	1.0, 3.0	0.75	0.72	0.54	0.47	0.58	0.66	0.87	1	0.46	0.62	0.69	0.73	0.43	0.43	0.56	0.66	0.69
	2.0, 0.5	0.59	0.51	0.44	0.4	0.78	0.37	0.47	0.46	1	0.47	0.39	0.41	0.87	0.87	0.53	0.39	0.4
	2.0, 0.5	0.81	0.59	0.48	0.43	0.63	0.74	0.68	0.62	0.47	1	0.78	0.73	0.41	0.41	0.85	0.82	0.77
	2.0, 2.0	0.75	0.51	0.4	0.36	0.51	0.89	0.78	0.69	0.39	0.78	1	0.9	0.35	0.35	0.67	0.91	0.96
	2.0, 3.0	0.76	0.52	0.41	0.36	0.53	0.85	0.84	0.73	0.41	0.73	0.9	1	0.36	0.36	0.64	0.82	0.91
	3.0, 0.5	0.53	0.47	0.4	0.37	0.69	0.32	0.42	0.43	0.87	0.41	0.35	0.36	1	1	0.45	0.34	0.35
	3.0, 1.0	0.76	0.57	0.47	0.43	0.72	0.64	0.61	0.56	0.53	0.85	0.67	0.64	0.45	0.45	1	0.7	0.67
	3.0, 2.0	0.76	0.52	0.42	0.38	0.53	0.86	0.74	0.66	0.39	0.82	0.91	0.82	0.34	0.34	0.7	1	0.9
	3.0, 3.0	0.76	0.52	0.41	0.37	0.53	0.86	0.78	0.69	0.4	0.77	0.96	0.91	0.35	0.35	0.67	0.9	1

This method of generating correlated betas was used in the example mentioned earlier in Ref. 2. For this problem, the service times of the faster servers (called *zippy* in Ref. 2) are distributed as $(9/4) \times \text{beta}(2,3)$. This distribution has a mean of 0.9 and a range of $(0,9/4)$. The service times of each of the slower system servers (called *klunky* in Ref. 2) are distributed $3 \times \text{beta}(3,2)$. This distribution has a mean of 1.8 and a range of $(0,3)$. The induced correlations between the system performance estimates increased from essentially zero (-0.04) for independent streams to 0.93, resulting in the estimated variance of the difference in performance being reduced from 7.41 to only 0.41 [16].

The use of one set of streams of random numbers to nominate values and another set to test acceptance can be generalized to Metropolis–Hastings Markov Chain Monte Carlo sampling [18].

COMMON RANDOM MODELING

Some of the difficulties in generating correlated random variables and synchronizing input processes can be avoided by using common modeling components. To illustrate this, we again use the example cited earlier—Example 11.1 in Ref. 2—where the performances of two simple queueing systems are compared. The simple event relationship graph (ERG) model on page 47 in Ref. 2 can be used to model both systems, and is shown explicitly in Fig. 5, where R is the number of servers and the simulation output is the Q process—the number of jobs in the system. Here, t_a represents the random times between job arrivals, and t_s represents the random service times. Conditions (/) required for event scheduling and delay times between scheduled events are shown in Fig. 5, making this a complete model.

The two events in this ERG model merely increment and decrement the Q when jobs ENTER the system and LEAVE the system. (This same ERG model can model waiting times, batch arrivals and processing, service breakdowns and repairs, etc. [17,19] with different event state changes.)

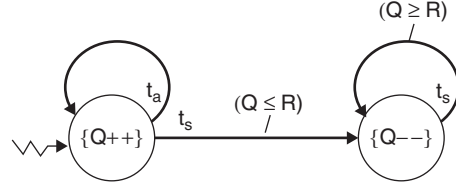


Figure 5. An event relationship graph (ERG) model for a G/G/R queue.

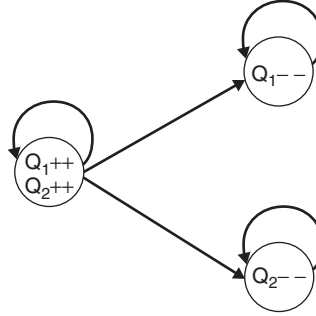


Figure 6. A Single ERG model for two queueing systems with identical arrivals.

Reading the dynamics in an ERG is easily done by interpreting each arc in a single sentence. Hence the complete dynamics for this model requires at most three sentences to define. The dynamics in the ERG is given by a single sentence for each of the three arcs:

1. Unconditionally a new job will ENTER ($Q++$) the system every t_a time units.
2. When a job ENTERs the system and finds an idle server ($Q \leq R$), it does not have to wait and can LEAVE ($Q--$) after its service time of t_s .
3. When a job LEAVES, if there still jobs waiting ($Q \geq R$), the next job will be served and can LEAVE after its service time of t_s .

To create perfectly correlated job arrival times for the two different service systems, like for the example mentioned earlier from Ref. 2, the two simulated systems can be combined into the single ERG in Fig. 6, with arc labels not displayed.

In the ERG model in Fig. 6, the two share the same ENTER event, which increments the Q for both systems. The induced correlation between job arrivals is perfect since they are identical. Service times can be generated and synchronized using some of the methods mentioned earlier in this article. The concept of a single model of several systems can be extended and exploited in many more ways as discussed in Ref. 18. Sharing common modeling components in an ERG model allows CRN to be used even when the seeds for the random number streams are not easily repeatable (like RAND() in Excel). This is a partial answer to the problem at the top of page 580 in Ref. 2: "...irreproducible gimmicks ... would generally preclude the use of CRN as well as many other valuable VRTs."

FURTHER READING

CRNs are by far the most commonly used and researched VRT (the full title of this article had over 50 000 Google references when it was initially published). Significant research papers that show the remarkable breadth and depth of CRN research spanning over a quarter century include Refs 20–23.

REFERENCES

1. Sanchez SM, Moeeni F, Sanchez PJ. So many factors, so little time: simulation experiments in the frequency domain. *Int J Prod Econ* 2006;103(1):149–165.
2. Law AM. *Simulation modeling and analysis*. New York (NY): McGraw-Hill; 2007.
3. Kleijnen JPC. *Design and analysis of simulation experiments*. New York (NY): Springer; 2008.
4. Song WT, Chien-Chou S. A three-class variance swapping technique for simulation experiments. *Oper Res Lett* 1998;23(1–2):63–70.
5. Hussey JR, Myers RH, Houck EC. Correlated simulation experiments in first-order response surface design. *Oper Res* 1987;35(5):744–758.
6. Hussey JR, Myers RH, Houck EC. Pseudorandom number assignments in quadratic response surface designs. *IEE Trans* 1987;19:395–403.
7. Schruben LW, Margolin B. Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *J Am Stat Assoc* 1978;73(363):504–525.
8. Tew JD, Wilson JR. Validation of simulation analysis methods for the Schruben-Margolin correlation-induction strategy. *Oper Res* 1992;40(1):87–103.
9. Nozari A, Arnold SF, Pegden CD. Statistical analysis for use with the Schruben and Margolin correlation induction strategy. *Oper Res* 1987;35(1):127–139.
10. Kleijnen JPC. Analyzing simulation experiments with common random numbers. *Manage Sci* 1988;34(1):65–74.
11. Page ES. On Monte Carlo methods in congestion problems: II simulation of queueing systems. *Oper Res* 1965;13:300–305.
12. Cheng RCH, Traylor L, Sztrik J. Simulation or rare queueing events by switching arrival and service rates. *Proceedings of the 25th Conference on Winter Simulation*; 1993 Dec. Los Angeles (CA). 1993. pp. 317–322.
13. Rao CR. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley (CA). 1965.
14. Schruben LW. Designing correlation induction strategies for simulation experiments. In: Adam N, Dogramaci A, editors. *Current issues in simulation*. New York: Academic Press; 1979. pp. 235–256.
15. Whitt W. Bivariate distributions with given marginals. *Ann Stat* 1976;4:1280–1289.
16. Schmeiser BW, Kachitvichyanukul V. Non-inverse correlation induction: guidelines for algorithm development. *J Comput Appl Math* 1990;31(1):173–180.
17. Schruben D, Schruben L. *Simulation modeling with event graphs. Text and ERG modeling software*. Accessed 2010. Available at www.sigmapwiki.com.
18. Kroese DP, Taimre T, Botev T. *Handbook of Monte Carlo methods* (Scheduled for possible publication in 2011–2012).
19. Schruben LW. *Simulation modeling for analysis*. *ACM Trans Model Comput Simul* 2010;20(1):1–22.
20. Kleijnen JPC. Antithetic variates, common random numbers and optimal computer time allocation in simulation. *Manage Sci* 1975;21(10):1176–1185.

21. Gal S, Rubinstein RY, Ziv A. On the optimality and efficiency of common random numbers. *Math Comput Simul* 1984;26(6):502–512.
22. Glasserman P, Yao DD. Some guidelines and guarantees for common random numbers. *Manage Sci* 1992;38(6):884–908.
23. Ehrlichman SMT, Henderson SG. Comparing two systems: beyond common random numbers. In: Mason SJ, Hill RR, Moench L, *et al.*, editors. *Proceedings of the 2008 Winter Simulation Conference*. Miami (FL): IEEE; 2008. pp. 245–251.
24. Centeno EM. Private communication (homework set for IEOR 261).

COMMUNICATING DECISION INFORMATION TO THE LAY RISK MANAGER: A CONSULTANT'S PERSPECTIVE

REX BROWN
School of Public Policy, George
Mason University, Fairfax,
Virginia

OBJECTIVES

A risk manager often finds he¹ has difficulty using information to form his own judgment or to explain it to others. He may not use it effectively—if at all. In this article, I suggest how an information provider can adapt the form and content of a communication to serve a risk manager.

The purpose may be to help him to choose among identified options (now or in the future), or to decide what new information to seek first. The purpose may also be to justify his action, before or after the fact.

Writer's Perspective

My treatment is selective and personal (and sometimes controversial). This is not intended as a balanced report of the state-of-the-art of information communication

My perspective is based on 50 years of consulting, mainly to government risk managers [1], often at the highest levels.² This experience covers a broad spectrum

¹Throughout this article, “he” means “he” or “she.”

²Including office heads at the Nuclear Regulatory Commission and the Department of Energy, an Assistant Secretary of Defense, a Deputy Assistant Administrator of EPA, a US Senate Committee, a House representative, and midlevel officials at IAEA, federal aviation administration (FAA), Corps of Engineers, and the Department of Commerce.

of risk management situations.³ My work with business executives,⁴ though not generally on risk management, has been quite relevant. The perspective is also based on two research projects, one for NSF⁵ to develop communication methodology, the other for NRC⁶ to test it in practice [3,4]. My technical approach has been lately applied decision theory (ADT) [5].

Elements of the Communication Process

The communication process has several roles. A Provider of information, P (e.g., a risk analysis contractor) produces Data, D (e.g., prescriptive decision analyses, factual risk assessments, value judgments, raw data, personal judgment or observation), and turns it into a Communication, C (e.g., a report or an interactive process). A Translator, T (e.g., decision analyst, who may also be the Provider, P) transmits it to a Risk manager, R (e.g., an executive, legislator or regulator), who is taking some Action, A (e.g., a safety intervention or research activity).

An Illustrative Case: Reactor Safety

A past consulting case will illustrate much of the argument. A senior nuclear regulator R was deciding whether to close down a reactor he suspected was unsafe, or to require

³Including reactor safety, waste disposal, air quality, airline security, business fraud, civil engineering construction, emergency planning, health risk, nuclear proliferation, laboratory accidents, anticrime legislation, gasoline additives, nuclear safeguards, military crisis management, submarine warfare, telecommunications emergencies, and toxic substance control [2].

⁴Including presidents and vice presidents at Ford, Firestone, Perkins Engines, and smaller companies.

⁵Decision and Management Science Program, grant ses84-20732.

⁶Division of Risk Analysis and Operations.

major safety backfit measures, or to do more research first A.

He had received a classic probabilistic risk assessment D, for which the reactor operator had hired a contractor at a cost of \$4 million. The PRA indicated that the reactor was among the safest in the United States. However, R knew of enough accident near misses at the reactor to put it on a warning “watch list”. Although technically sophisticated, R had difficulty extracting from the PRA what he needed to form his own judgment or to justify it, if it were challenged. The PRA also indicated that a “venting backfit” would be the most cost-effective backfit and R found *that* convincing.

My colleagues and I, T, worked with R a few weeks on evaluating the PRA and combining it with other available information, including R’s own observations and the informed judgment of colleagues [6,7]. In the light of this more comprehensive analysis, R decided to permit the reactor to operate, subject to a venting backfit. We developed a report and interactive software C designed to defend the decision if it were challenged, but it was not.

ASSESSING RISK

The Substance: Subjective versus “Objective”

A judgment-avoidance culture is common among information providers P who purport to limit themselves to “objective” information, disregarding critical “soft” information D. This objectivity orientation is legitimate in scientific enquiry, but some of us [8] hold that it does not belong in analysis intended to aid decisions and may seriously distort action. It often leads to analyses embodying simplistic “cop-out” assumptions that imply unrealistic personal judgments.

For example, a proposed regulation is often evaluated *as if* the regulated party were certain to faithfully implement requirements—which is rarely realistic. This has the effect of overvaluing the regulation, since it may be less effective than was assumed. By how much, is the question for informed judgment (but it will not matter too much if

it simply strengthens a case for rejecting the regulation).

Simplified assumptions need not invalidate a communication C provided they are recognized and translator T helps R to make appropriate adjustments to his judgment.

A very common and reasonable simplification is to treat an incremental decision strategy (like putting your toe in the water before taking the plunge) as if it were a once-and-for-all commitment (like plunging without hesitation). The simplification *undervalues* the incremental strategy [since R can retreat if the water is too cold]. A risk management example would be the case where there are signs of an impending reactor accident. The operator R has to decide whether to begin shutting the reactor down, given that he can abort the shutdown if it proves to be a false alarm. He can evaluate the incremental shutdown step *as if* it committed R to shut down irrevocably, and use his informed judgment to adjust the evaluation upward.

A prime example of a judgment-avoiding assumption is classic PRA, referred to in the reactor backfit case. Such PRAs, as prescribed in regulatory guidelines [9], normally address only sources of risk with well-documented evidence (say from experiments or testing) and disregard other evidence (such as unplanned observation).⁷

In the reactor case, the PRA addressed only internally initiated core melt accidents (such as pipe breaks) and ignored external accidents (such as station blackout), as sources of risk (as well as any *informal* evidence about internal accidents). From other sources we sought probabilities of externally initiated events. On the basis of this adjustment, R rejected the PRA finding that the reactor was adequately safe overall and required that a costly backfit be installed A.

The PRA may contain a great deal of relevant information (as in this example), but is not sufficient on its own, to produce a realistic risk assessment. Experienced regulators let PRA guide them in certain cases, such as engineering design, where there is, in fact,

⁷“PRA” might be interpreted as “Partial Reliability Assessment.”

authoritative evidence on all relevant sources of risk and virtually no other useful evidence. In the reactor case, R accepted the PRA finding that a venting backfit was preferred to an alternative, since the PRA omissions did not affect that comparison.

Special Cases

“Unknown Unknowns”⁸. The Code of Federal Regulations [10] requires that “the likelihood” (see section below titled “The Form”) that a certain total radioactive emission level from a waste repository over 10,000 years should not exceed 10%. A DOE study of the proposed Yucca Mountain site found the human intrusion component of that likelihood to be of the order of one in a million. This assumed that the only source of human intrusion was mining for minerals—the only source that the authors could identify. (This would be like the Indians on Manhattan Island 10,000 years ago assuming that the only risk of human intrusion into a sacred burial site by 2000 AD would be the robbery of ancestral bones. . .)

In view of the hopelessness of coming up with a defensible and realistic probability of unvisualized human intrusion, I suggested reformulating the risk of a site along the lines of “the probability of any presently identifiable human intrusion that *differentiates among* contending nuclear waste sites.” In other words, report only risk information *relevant to the decision purpose at hand*. In this case, report only whether the presence of mineral deposits at Yucca Mountain makes it more or less susceptible to human intrusion than contending waste sites. If other sources of intrusion cannot be specified (i.e., unknown unknowns), they cannot affect the current *comparison* of sites, and can be safely ignored.

Sample Data. Sample data, say, of public risk attitudes and system failures, are particularly vulnerable to miscommunication due

to judgment-avoidance culture. Errors, that is, differences between the sample as measured and the population of interest, can arise from a number of causes. Usually only one of them, random sampling fluctuations, can be “objectively” reported. Established statistical procedure can produce a random sampling error, but that is commonly cited as the whole “margin of error.” This is appropriate in those cases where the sample is indeed drawn randomly from a stable repetitive process of interest and can be accurately measured. In the manufacture of standardized mechanical components, the sample may come close enough for the “random sampling error” to be almost all total error.

However, in many important cases, sources of error other than random are much greater. Suppose R wishes to assess what fraction of the US population engages in some risky practice (such as phoning while driving). In a large random phone sample from the DC directory, 5%, say, admit that they have done so in the past week. This 5% is reported to R with a statistically calculated “90% margin of error of, say, $\pm 1\%$ ” that is, 4–6%. R’s realistic uncertainty range could well be, say, 10–40% if R takes account of uncertain biases, because people can be expected to understate their dangerous behavior and/or DC phone subscribers differ from the US population. It is legitimate for P to communicate $\pm 1\%$ as the purely random component of total error and alert R to other errors he must take into account. Alternatively, he can help R to quantify those errors. A *calibration sample*, where the actual behavior of a small subsample is meticulously observed, can introduce some degree of objectivity [11].

Rare Events. Comparing very low probabilities is difficult, but may be critical (e.g., when deciding how much it is worth spending to reduce the risk of a nuclear accident). Equating the probability to an analogy that R is likely to be familiar with (e.g., being killed in a car accident; or 10 heads in a row; or being struck by lightning; or a royal flush, for a card-player) may better convey appropriate uncertainty [12]. Familiar analogs can also be used to convey quantities that are

⁸“There are more things in heaven and earth, Horatio, than are dream’d of in your philosophy”—Hamlet.

mind-bogglingly *large* (e.g., noting that a \$300 billion borrowed from China to make infrastructure safer is equivalent to \$1000 more debt for every US citizen).

The Form

Language for the Layman. Language is an essential component of communication strategy. Misleading or confusing language on its own can—and often does—doom a decision-aiding effort.

I was much chastened by an early career experience. I presented a company president with an analysis of a major decision, couched in conventional technical language. Before I finished, he grunted “gobbledygook!,” stormed out of the room and never sought my help again! Many a decider has been lost to analysis because he had not the time, the patience, or the training to absorb what he was hearing.

I will not enter, here, into the current debate in the decision-aiding literature⁹ on reforming decision vocabulary. Instead, I will adopt existing conventions, which are acceptable when engaging a technical audience. However, I will mention, in passing, a few lay translations, which have helped me avoid confusing and misleading laymen. In particular, I will replace “decision analysis” with the more specific “applied decision theory” [15]. Using “frequency” to characterize what decision analysts call “probability” is meaningless if R’s personal judgment is what is being aided (unless some real iterative process is involved, such as component failure).

Where possible I follow some respected vocabulary precedent.¹⁰ Hopefully, an authoritative professional body will, in time, authorize a standard glossary¹¹ that we can

confidently use when communicating with lay deciders.¹²

Graphical devices for getting across simple uncertainty, expressed in terms of R’s personal probability, have been well addressed at some length by other authors [17]. Devices include various variants of bar charts that show means and credible intervals and probability density curves. (Laymen appear to find cumulative curves more difficult to interpret).

EVOLVING COMMUNICATION FOR PRIVATE DECISION MAKING

Communicating risk management information for R’s private decision making can have two distinct phases: basic and interactive. In the “basic phase,” the communicator identifies a minimum set of information that R will certainly want to know, and which P can prepare ahead of time. This usually takes the form of a basic report, which is typically a written document or a formal briefing. The “interactive phase,” conveys optional information in response to R’s interest that emerges later. This is what I will now discuss.

The Macromodel as an Organizing Device

I have found the “macro model” to be a convenient vehicle for conveying an evolving decision argument in the interactive phase. This is a structurally simple model of a “target” judgment, an extreme example of which would be “net value of option = benefit—cost.”¹³

Any macromodel input can be derived by plural evaluation, that is, making an assessment in several ways and merging possibly inconsistent results. Each component assessment may include R’s intuitive judgment

⁹See Ref. 13 and discussion in Decision Analysis, June 1994, and my tentative proposal in Ref. 14.

¹⁰For example, I adopt Schlaifer’s replacing of “utility” by “preference” to connote its personal nature [16], and accept Howard’s rejection of “expected” implying unnecessary confidence [12]. I tentatively suggest “projected preference” in place of “expected utility”.

¹¹As the Academie Francaise does for the French language.

¹²As well as students—and perhaps, eventually technical specialists.

¹³Where the target is a probability (e.g., of accidental damage), the macromodel could be an inference equation. Its inputs would be unconditional probabilities of conditioning events (such as types of accidents) and the conditional probabilities of damage in each case. They would be combined in a familiar conditioned assessment formula.

Table 1. Macromodel for venting backfit of option

	Baseline	Backfit Impact
<i>Inputs</i>		
Factual judgments		
Core melt/year probability ($\times 1\text{E}-4$)	0.76	0.63
Conditional off-site radiation (man-rem)	23.7	0.6
Off-site damage (\$B)	1.7	0.6
On-site damage (\$B)	4.6	1.0
Industry backfit cost (\$M)	4.8	
Regulator cost (\$M)	0.3	
Plant life (years)	30	
Value judgments		
Radiation—cost trade-off (man-rem)	1000	
<i>Outputs</i>		
Total benefit	\$40.6 M	
Total cost	\$5.1 M	
<i>Net benefit of venting</i>	<i>\$35.5 M</i>	

and the output of “feeder” models or studies, which are then combined.

Table 1 refers to a macromodel we used for the reactor backfit case. It shows estimates for the venting backfit option, resulting in a positive net benefit of \$35.5 M, compared with “do nothing.” An alternative backfit, showed a negative net benefit of \$24.8 M. The uncertainty ranges assessed (not shown here) implied a 90% range for net benefit of \$13.4 M–\$159.6 M¹⁴ R took this into account when deciding to require the venting backfit.

It included an assessment of a core melt accident within the next year with a probability of 0.000076, and a mean estimate of 23.7 man-rem of emitted radiation, which is a determinant of resulting deaths. These numbers drop to 0.000063 and 0.06 man-rem, respectively, with a venting backfit. They also included judgmental adjustments for omissions and biases (see section titled “The Substance: Subjective versus Objective”).

Another project [18] evaluated the cost effectiveness of the Clean Air Act. Inputs to a macromodel included industry costs and impact on various classes of pollution, supplied by specialized contractors (e.g., the US Geologic Survey). However, the contractors

did not report their findings in the form required as input to our macromodel, and “bridging” analyses were needed. The need for bridging analyses slowed down presentation to the Democratic Congress, which had mandated the study, until it was replaced by a Republican Congress, which cancelled the study!

The Interactive Phase

Figure 1 illustrates the role of a macro model in mediating between a target judgment and various sources and levels of input judgment, analysis, and data. It shows schematically the kinds of considerations R took into account, including plural evaluations for each element.

In the reactor case, we presented R with an interactive computer graphic display of a macromodel corresponding to Fig. 1, augmented by input uncertainty ranges. The input values included those shown in Table 1, derived from our digestion of the PRA and whatever other information was available. Each assessment was extracted and second-guessed from various kinds of data (e.g., elements of PRA and ADT analyses, experimental or anecdotal observations, and engineering and health statistics).

R worked with the display, overriding inputs as he saw fit, either from his direct judgment or by referring to other sources

¹⁴The combining formulas were familiar statistics decompositions.

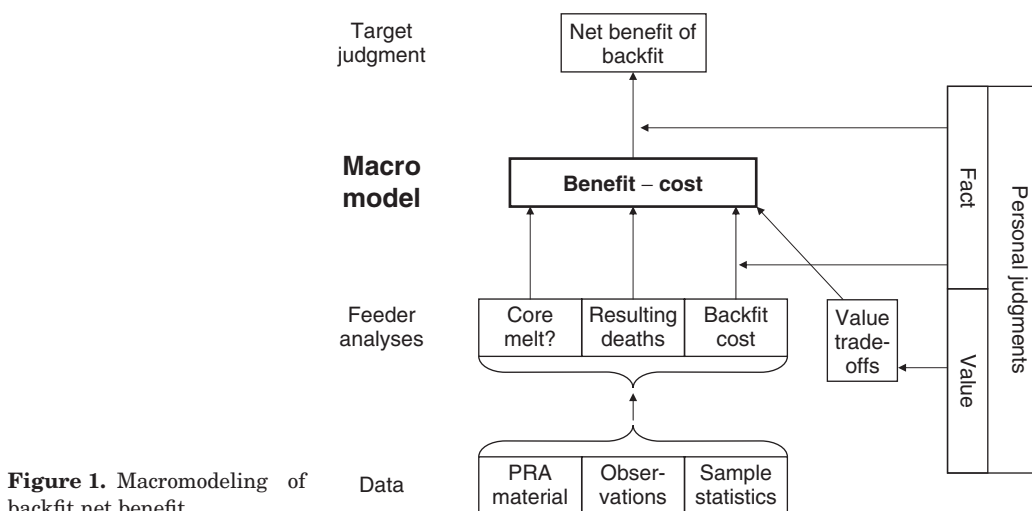


Figure 1. Macromodeling of backfit net benefit.

of data, and noting the resulting impact on the backfit net benefit display. In this way, R dynamically probed deeper and deeper into the available information mass, as need dictated. On the basis of such progressive iterations, he made his final judgment: to require the venting backfit.

Contingent Decision Aid

Sometimes a decision tool is designed for future contingencies. The as-yet-unknown circumstances are inputted to the model structure when the situation arises.

Lab Accident Example. The EPA had our team [19] develop decision aid for chemical laboratory directors R to determine what level of risk control to apply to hazardous chemical operations (e.g., simple lab bench precautions vs. full containment equipment).

A team of chemical engineers had constructed a large generic risk, intended to apply to any lab experiment. It calculated overall probabilities of accidental injury as output of a single complex micromodel, with over 1000 variables (such as features of the experiment and properties of the materials involved). We structured the model as a hierarchy of macromodels whose inputs R could override at any level of aggregation.

For example, in a midlevel macromodel, the micromodel calculated the probability of

a spill in a certain experimental situation as 1%. However, one R noted that in the past 30 experiments there had been three spills, that is, 10%. After conferring with colleagues experienced in comparable situations, R substituted 5% for the 1% as input to the corresponding macromodel. He similarly overrode other variables in the macromodel hierarchy, and accepted as his own risk assessment the output of the top-level macromodel.

COMMUNICATING WITH OUTSIDE PARTIES

The Need to Justify Action

Guidelines for communicating with risk managers generally also apply to other areas of business and government where uncertainty is significant.

However, a distinguishing feature of risk management is that decisions are often controversial and need to be justified, for example, to the public or the courts. In our reactor case, a decision to close down a plant operation or require expensive safety measure is liable to be challenged. R may have good grounds for his choice, but not be able to demonstrate uncontroversially that it is sound, which requires documented evidence (such as engineering tests).

Although businesses have lesser need for a defensible rationale than government, a significant exception is the case of

legal “prudence hearings.” To avoid often crushingly expensive liability, a company may need to defend itself against a charge that it made “imprudent” decisions resulting in major environmental damage, perhaps requiring government superfunding to cleanup.

Criteria for Communication

Like decision rationale for R’s *private* use, a rationale for *public* use must be readily and quickly absorbed. However, it must also be “acceptable” to the audience. In particular, the *apparent* role of subjective judgment must be minimized. Thus, even if a regulator’s decision to, say, close down a reactor were based, largely, on his personal observation of lax safety practices, he would be well-advised to cite just one formal decision analysis that supported the same decision. (If he has reconciled all plural evaluations, they all will point to the same choice.) Some legal experts [20], but not others [21], hold that a decision produced by some systematic process (like ADT) is valid defense against being considered “arbitrary and capricious.”

Value Judgments

A risk management decision depends on typically very subjective value judgments, which can be difficult to validate and are politically sensitive.

Legislators shy away from putting a dollar value on human life, although they do so implicitly whenever they pass legislation intended to save lives. A congressman told me that if he proposed any value of a human life (regardless of what it was), it would incite hostile and emotional public reactions and lose him votes.

In the reactor backfit case, risk could be evaluated in two ways, according to different versions of published regulation [22]. Radiation harm can be evaluated at \$1000 per man-rem, or at \$20 million per life (which are cited in regulations). As shown in Table 1, we opted for the former as less provocative (if less relevant).

Where R has private interests beyond public welfare, he may prefer not to disclose embarrassing value judgments. A nuclear

regulator may, for example, be inclined to make safety requirements stricter than meeting health rules require, in order to avoid the bureaucratic hassle and political outcry of, say, the 1979 Three Mile Island accident. In the backfit case including “hullabaloo” as a criterion would not have indicated a different decision, so nothing was lost by ignoring it.

One approach to politically sensitive value judgments is to provide R with a “parametric” analysis, where he may privately supply value judgments that he would hesitate to declare publicly.

The Senate Judiciary Committee was deciding whether to authorize a Community Anti-Crime Program, which would fund local initiatives (like Neighborhood Watch) and cost \$100 million a year. We presented committee staff with a macromodel in the form of a seven-criterion Multi-attribute Utility Analysis model. Taxpayer cost and impact on crime were the main criteria. The macro model strongly favored the bill, based on a wide range of plausible importance weights. However, we were told that the committee Chairman (Joseph Biden) was sensitive to the goodwill of the politically influential Association of Chiefs of Police, who opposed the program (since it would by-pass the regular criminal justice system). Therefore, we prudently added “electoral security” to the criteria, and passed the model to the chairman to privately enter his own importance weights. The bill failed to pass. . .

GETTING AND USING EVIDENCE

An important risk management judgment is deciding what, if any, new evidence to seek before making the main decision and how to update uncertainties in the light of what is learned. I know of no very satisfactory way to convey the underlying reasoning in information gathering decision cases. Laymen, in particular, seem to have great difficulty understanding the rationale—however expressed—of applying Bayes’ theorem.¹⁵ However, it is feasible and

¹⁵Common use of “Bayesian statistics” to refer to a broad class of methods for making inferences

useful to communicate certain elements of that reasoning and to assure that R concurs with it.

Diagnostic Updating

Prior Probability. Before evidence has been received, assessing a “prior” is the same as making any probabilistic assessment (see section titled “Assessing Risk” above). However, when R comes to update his uncertainty, he is usually already aware of the evidence, so his judgment of what his assessment *would have been* without that knowledge may be fatally contaminated. I do not even try to elicit it.

“Likelihood”. On the other hand, it can be helpful to elicit, with some care, the “likelihood” (“diagnosticity?”)¹⁶ component of Bayesian updating, from R, who will normally be the main source of the judgment.

Suppose the possibility to be predicted is an impending earthquake and the evidence is agitation among farm animals. Something like the following phrasing seems to work quite well: “How surprised would you be to see the animals agitated like this when no earthquake is imminent?” “Very.” “How surprised if an earthquake *is* imminent?” “Somewhat, but about a factor of ten less.”¹⁷

based on quantified judgment is misleading. Applying Bayes’ theorem does not have to involve human judgment and Bayes’ theorem is not the only formulation for making judgmental inferences. I prefer “personalist” inference (and decision theory).

¹⁶The standard statistical meaning of “likelihood” is probability of evidence conditional on hypothesis. However, it is often used in the literature to mean probability (e.g., “... the likelihoods of these outcomes, expressed as probabilities” on p. 20 of Ref. 23 and “... the disposal system ... shall have a likelihood of less than 1 chance in 10 of exceeding ...” on p. 11 of Ref. 9).

¹⁷The following wording, though technically accurate, would be hopelessly confusing to a layman: “What would the ratio of your probability of earthquake conditional on “agitation” be to your probability of “earthquake” conditional on “no agitation.”

Risk managers often insist on replacing of numbers by words, in this way. We did a high-level study¹⁸ recommending that the United States should somewhat lower an embargo on exporting high-powered computers to the then-Soviet bloc, on the grounds that a small increase in military threat would be more than offset by economic and other advantages. A quantitative linear multiattribute utility analysis had to be reduced to a single-page report without numbers before being submitted to the national security advisor (Henry Kissinger), who was known to refuse to read quantitative reports.

Value of Information

The rationale underlying quantitative “Pre-posterior analysis” [25], i.e., Bayesian derivation of the value of imperfect information, is still more complex to communicate to a decision analyst, let alone to a layman.

However, the following line of questioning, though still awkward, might make a recommended information strategy plausible. “How firm is your present risk assessment—that is, how much would perfect information shift it? If you were sure this research would give you that perfect information, what are the chances it would save you from a decision you would regret? How much cost would avoiding that mistake save you? How close to perfect information would this imperfect research bring you?”

A More Ambitious Rationale for an Information Strategy

The reasoning underlying a decision on what new information to seek before making a primary decision can be communicated at greatly differing levels of effort and technical sophistication, depending on circumstances.

A nontechnical congressman will rarely have more than an hour to devote to any one decision, however important (for example, the Clean Air Act evaluation discussed above). In such cases, having R answer a

¹⁸For the President’s Council on International Economic Policy [24].

few questions, as above, may be all that is feasible.

A rigorous mathematical argument would be prohibitively costly and technically challenging. However, if the problem is sufficiently important and R adequately trained, enough of the technical argument may be communicated to help R decide whether to adopt an information strategy. For example, a senior DOE official (Ben Rusche, Head of the Office of Civilian Radioactive Waste Management), a trained scientist, had to decide whether to commission a comparative study of the suitability of alternative nuclear waste sites. He in fact did have one conducted for several hundred thousand dollars [26].

In such a case, a quantitative elaboration of just part of the argument having to do with the firmness of R's risk assessment (see first question above) could be feasible and appropriate. Appendix A proposes a moderately ambitious communication process.

APPENDIX A: QUANTIFYING THE FIRMNESS OF RISK ASSESSMENTS

Several important risk communication tasks depend on the firmness or "shiftability" of R's current judgments.

1. *Plural Evaluation.* The more a risk management evaluation is liable to shift with more information or reflection, the less weight it should be given when combining it with other attempts at the same judgment.
2. *Getting More Information.* The more shiftable an evaluation, the stronger the case for more research or delay.
3. *Vulnerability to Challenge.* The more shiftable an analysis, the more vulnerable it is to legal or other challenges.

In many ways, communicating this firmness is more important than a simple assessment. It has serious action implications and is more difficult to do right, since it involves the elusive concept of second-order probability.

Two interpretations of assessment firmness need to be distinguished: assessment shift with unlimited information; and with

limited, but achievable, information. They require distinctive language and graphical devices to keep them apart. Figure A.1 illustrates them for the case where the uncertainty is about deaths in the event of a reactor accident.

MAXIMUM ASSESSMENT SHIFT WITH UNLIMITED RESEARCH

Outcome assessments, including event probabilities and estimates, may shift with new information analysis or reflection. How far could they conceivably shift with *unlimited* research and information? In particular, could they shift enough to favor a different decision?

Laymen and technical people alike commonly talk as if there were some objective "true risk" of a hazard, which risk assessment tries to approximate. Most personalists reject this view (it has been likened to the medieval theory that fire results from a substance, phlogiston, contained in an object, which escapes when it burns), but I think it has a legitimate interpretation worth communicating. It can be viewed as the point a *personal* probability would shift to if the assessor had enough information to essentially remove subjectivity. In that sense, it becomes an impersonal probability, about which we can be uncertain, and subjectively assess maximum shift distributions (contrasted with *personal ideal probability*), which is the result of perfect analysis of a given person's idiosyncratic, and therefore subjective, knowledge [27,28].

Figure A.1a shows such a distribution. The mean and range of the original risk hazard assessment is shown by a broken line on the left. Where the mean would shift to with maximum information is shown by a fan leading to a solid line. The remaining irreducible hazard uncertainty is illustrated by a new shorter broken line.

PARTIAL ASSESSMENT SHIFT WITH LIMITED RESEARCH

Unlike the familiar decision theory concept of *perfect* information, information does not

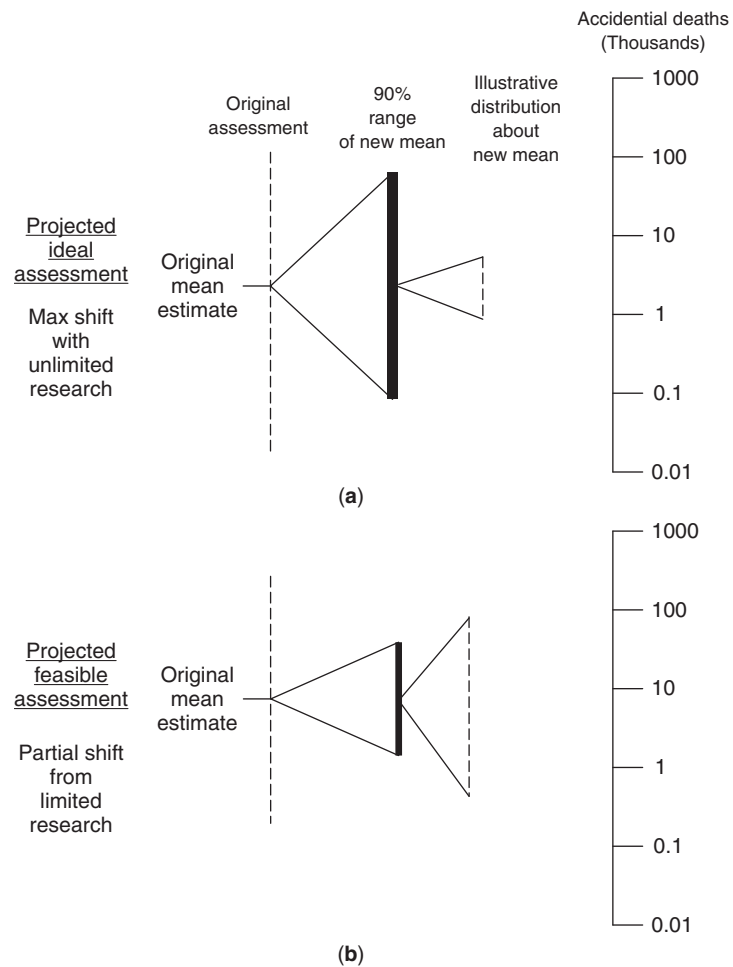


Figure A.1. Firmness of current assessment (projected shift with research).

remove all uncertainty, only as much as is feasible. It sets a tighter, and therefore more useful, bound than perfect information on the value of new information research. (But it is more elusive to define.)

Figure A.1b uses the same graphic conventions to communicate how a hazard assessment might shift with information or analysis that might *actually become available*, say, by waiting for a few years or conducting a particular research project. This is directly relevant to the decision maker in deciding whether in fact to wait for such information. If the potential shift in assessment is unlikely to change the decision, this would argue against waiting.

The limited information can be expected to shift the original hazard assessment less

than maximum information. This is reflected in the shortened vertical line in the left fan, but a longer broken line for remaining uncertainty in the right fan. (The amount of prospective information can be indicated by the thickness of the line).

REFERENCES

1. Brown RV. Working with policy makers on their choices: a decision analyst reminisces. *Decis Anal* 2009;6(1):14–23.
2. Brown RV. Environmental regulation: developments in setting requirements and verifying compliance. In: Sage AP, editor. *Systems engineering and management for sustainable development*, Encyclopedia of life support systems (EOLSS), UNESCO.

- Oxford: EOLSS Publishers; 2002, <http://www.eolss.net; mason.gmu.edu/~rbrown>.
3. Brown RV. Communicating information to risk managers: the interactive phase. In: Lave LB, editor. Risk assessment and management. New York: Plenum Publishing; 1987.
4. Brown RV, Ulvila JW. Communicating uncertainty for regulatory decisions. In: Covello VT, Lave LB, Moghissi A, *et al.*, editors. Uncertainty in risk assessment, risk management, and decision making. New York: Plenum Press; 1987.
5. Brown RV. Rational choice and judgment: decision analysis for the decider. New York: Wiley; 2005.
6. Brown RV, Ulvila JW. Papers on communicating information to risk managers. Falls Church (VA): Decision Science Consortium, Inc.; 1985.
7. Brown RV, Ulvila JW. Does a reactor need a safety backfit? Case study on communicating decision and risk analysis information to managers. Risk Anal 1988;8(2):271–282.
8. Morgan MG. Bad science and good policy analysis. Science 1978;201:971.
9. US Nuclear Regulatory Commission. PRA Procedures Guide, NUREG/CR-23 000. Washington (DC); 1982.
10. Code of Federal Regulations Title 40, Energy Part 191. Office of the Federal Register, National Archives and Records Administration; 1985.
11. Brown RV, Campbell VN, Repici DJ. Analysis of residential fuel conservation behavior: for federal energy administration. Decisions and Designs, Inc.; october 1977.
12. Merkhofer LW. The use of risk comparison to aid the communication and interpretation of risk analyses for regulatory decision making. In: Lave L, editors. Risk assessment and management. Plenum; 1987.
13. Howard RA. Speaking of decisions: toward precision in decision language. Decis Anal 2004;1(2):77–78.
14. Brown RV. A decision glossary for laymen; 2006, [Mason.gmu.edu/~rbrown](http://mason.gmu.edu/~rbrown).
15. Brown RV. Decision analysis: applied decision theory? Decis Anal Soc Newsl. May 6–8 2008.
16. Schlaifer R. Analysis of decisions under uncertainty. New York; McGraw-Hill; 1969.
17. Tufte ER. Visual display of quantitative information. Cheshire, Conn; Graphic Press; 1983.
18. Brown RV. Evaluating the cost-benefit of the clean air act realistically using hierarchically partitioned assessment. Presented to the Society for Risk Analysis national conference. Baltimore (MD); 1991.
19. Mendez W, Brown RV, Bresnick TA. Laboratory level-of-control decision aid. Washington (DC): ICF, Inc.; 1984.
20. Stewart R. Personal communication. 1976.
21. Breyer S. Personal communication. 2009.
22. US Nuclear Regulatory Commission. Safety goals for nuclear power plant operation. Washington (DC); 1983.
23. Raiffa H, Richardson J, Metcalfe D. Negotiation analysis: the science and art of collaborative decision makings. Cambridge (MA); Belknap Press of Harvard University Press; 2003.
24. Watson SR, Brown RV. Issues in the value of decision analysis: an evaluation of export controls on computer sales to the Soviet Bloc. McLean (VA): Decisions and Designs, Inc.; 1975. (NTIS No. AD A019 339).
25. Raiffa H, Schlaifer R. Applied statistical decision theory. Boston (MA): Division of Research Harvard Business School; 1962.
26. Keeney RL. An analysis of the portfolio of sites to characterize a nuclear repository. Risk Anal 1987;7:195–218.
27. Brown RV. Assessment uncertainty technology for making and defending risky decisions. J Behav Decis Making 1990;3:213–228.
28. Brown RV. Impersonal probability as an ideal assessment based on accessible evidence: a viable construct? J Risk Uncertain 1993;7:215–235.

COMPARISONS OF RISK ATTITUDES ACROSS INDIVIDUALS

DANIEL EGAN
GREG B. DAVIES
PETER BROOKS
Barclays Wealth, Behavioural
Finance, London, UK

GOALS IN COMPARING RISK ATTITUDES

In comparing risk attitudes across individuals, a researcher is generally looking to answer one of the three questions. The first is “can I reliably measure the difference between the risk attitudes of individual A and individual B?” This question asks if the measurements or methods that the researcher is using can reliably differentiate between individuals with different risk attitudes. The second is “what is the magnitude of the difference?” This question seeks to determine what range and variance the risk attitudes cover, and the patterns and covariates of dispersion over that range. The final question is “what factors do we need to control for to compare risk attitudes meaningfully?” This question often arises when we are comparing individuals in different circumstances and backgrounds (such as a poor laborer vs a rich financier) and groups of individuals, such as men and women, young and old, or people raised in different cultures.

Comparing these attitudes across individuals requires a clear understanding of how “people differ from one another, but remain true to themselves” [1]. Empirical evidence supporting the possibility of stable risk attitudes comes from “strong evidence that virtually all individual psychological differences, when reliably measured, are moderately to substantially heritable” [2]. Heritability of risk attitudes has specifically been found with a variety of measurement methods. Using actual investment portfolio choice as an outcome measure, roughly 30% of the

variance in choices is attributable to genetic differences, rather than environmental [3]. Tests of covariation in economic preferences roughly attribute 50% to heritable factors [4].

Further evidence of individual stability reflecting both genetic and environmental factors comes from longitudinal studies of the same individuals. Correlations in risky decision making over a three-year period range between 0.20 and 0.38 [5], and persistent individual factors accounted for 75% of the variation in responses to a risky choice regarding income levels [6]. If coarse genetics and longitudinal studies reveal stable individual differences, we may be able to design finer grained methods that reliably measure these individual risk attitudes.

In the pursuit of this goal, we first broadly review the challenges in defining comparable individual risk attitudes in the section titled “Challenges in Comparing Risk Attitudes.” The section titled “Measurements of Risk Attitude and Their Implications” introduces commonly used measurement scales and their implications for comparisons. The section titled “Elicitation Methods” details the evidence on variation of risk attitudes from commonly used elicitation methods. The section titled “Inferences in Comparing Individuals and Groups” concludes with guidance for designing studies intended to make inferences in risk attitude comparisons.

CHALLENGES IN COMPARING RISK ATTITUDES

It is essential that we first define the risk attitude we are attempting to compare. A risk attitude is not a unidimensional construct that can be easily summarized and compared using a single measure. Rather, risk attitudes are complex constructs that may be multidimensional, both across domains and within a specific domain. Moreover, there is significant evidence that empirically inferred risk attitudes are more a function of the choice context, rather than a global inherent

attitude [7–9]. To compare elicited risk attitudes, researchers should have a clearly defined risk attitude they are attempting to elicit, and understand the ramifications of elicitation method for analysis.

A key rule when making comparisons is that measurement scales and elicitation methods must be like-for-like between individuals. The *measurement scale* defines the values that an assessment can take on, and the viable inferences from it. The *elicitation method* specifies the exact mechanism by which an individual's risk attitude is revealed to the researcher. Many measures of risk attitude are notoriously sensitive to small changes in elicitation method, which has led researchers to describe preferences as “constructed” rather than “revealed” [10,11]. Constructed attitudes can be strongly influenced by the *framing* of the decision problem, such as in Kahneman and Tversky's famous Asian disease experiment [12]. Two objectively equivalent choice sets with only different wordings (number of lives potentially *saved*, rather than the number of lives potentially *lost*) caused the majority choice to flip from being risk averse to risk seeking. Thus, when comparing individual risk attitudes, we must ensure we have used the same elicitation method.

We now review factors that can bias comparisons if the elicitation method is not carefully designed.

Domain

The first challenge is defining the appropriate domain. A domain is a specific category of risk—for example, investing, health and safety, gambling, negotiations, purchasing insurance, and recreational activities. The domain dependence of elicited risk attitudes indicates that they must be carefully measured to correctly reflect risk attitudes within a specific domain. Attempting to infer investment risk attitudes from a lottery or gambling is likely to fail, as in these seemingly close domains individuals actually exhibit different risk attitudes [13–15]. Individual risk attitudes are so inconsistent across domains that they do not appear to reflect a stable person trait, implying that there is little basis for identifying a global

risk-attitude assessment instrument [16,17]. However, risk attitudes and behaviors are consistent and predictive within domain [18], indicating that focused risk-attitude assessments are meaningful and useful.

Perception of Risk

A single measure of risk attitude may mask the fact that a number of different features of the choice problem combine to form the overall *perception* of risk. Weber and Millman [19] show that people display consistent risk aversion according to their subjective perceptions of risk, but not according to theoretical measures of risk. In other words, their perceptions of risk may be derived from aspects of the choice other than the researchers' theoretical measures of risk. This divergence may then drive the degree of observed risk aversion. For example, investors show strong preferences for investing in local or well-known stocks, because the sense of familiarity leads to a perception of them being lower risk, even at the expense of diversification benefits from less well-known stocks [20,21]. Determining risky attitudes from choices may thus also require the assessment of the perceived risk of the choice [22]—an area of active research.

Framing and Risk Myopia

As detailed in the Asian disease example above, the presentation of a choice often influences the risk attitudes to a surprising degree. The implications of this in some real-world settings are alarming. Individuals show a poor ability to look at the “big picture,” resulting in suboptimal choices over time [23], over many choices, and over the whole outcome set [24]. Samuelson once noted that if you offer an individual a single 50/50 chance of gaining \$100, or losing \$50, you have few takers. However, if you offer an individual 100 such bets as a set, they are likely to take them (even though backward induction shows that this sequential choice is unsustainable) [25]. Thus, many compartmentalized *individual* decisions can result in extremely high risk aversion compared to cases where one takes a more holistic view. Similarly, individuals may react to compound

long-term prospects such as investing as if they were one-off gambles, and thus express more risk aversion than they would when using the appropriate time horizon [26].

Noise in Responses

When assessing risk attitude in any manner, we also need to consider the noise and inconsistencies that are apparent in many elicitation tasks. For example, experimental subjects frequently respond differently to the same question repeated on two occasions within the same experiment [27]. Retest reliability rates within subjects rarely exceed 70%. We can only say that one individual truly differs from another once the noise in an individual's responses is accounted for, and a number of decision-making models attempt to model such variability explicitly [28–35]. To reach the point where we can confidently assess the consistency of differences in estimates often requires many questions, and assessment of the fit of competing utility functions [36].

Incentive Effects

Inconsistency can be found between experiments using hypothetical rewards and real rewards in choice tasks. Typically, there can be a heightened sense of purpose and impact from a decision task if it is being played for real rewards rather than for hypothetical rewards. This often leads to a greater level of risk aversion being observed in tasks with real rewards [37–39], though the effect is not independent of frame [40].

MEASUREMENTS OF RISK ATTITUDE AND THEIR IMPLICATIONS

Nonutility-Based Measures for Comparing Risk Attitude

Perhaps the most general pure definition of risk aversion is that of Yaari [41], who characterizes it in terms of *acceptance sets*: individual A is said to be more risk averse than B if the set of gambles or investments that A would accept is a subset of the acceptance set of B. Individual B will accept all gambles taken by A, plus additional gambles that

are too risky for A to accept. Note that this provides us with only a rank ordering of individuals, which states that person A is more or less risk averse than person B, without any indication of the magnitude of difference between them. Using rank orderings it is possible to predict in advance that one person will make a more risk-averse choice than another person, but it is not possible to predict specific choices. It also fails to give us a complete ordering of individuals, since it can only make comparative statements about those whose acceptance sets are neat subsets of another's. Relative psychometric scales are an example of this sort of rank ordering.

A related cardinal measure is given by the *certainty equivalent* of a risky gamble. This is the amount that will make the decision maker indifferent to accepting the gamble or not, and can be elicited directly in choice tasks without subscribing to any underlying decision model. The decision maker's certainty equivalent should be equal to his valuation of the gamble.¹ If the decision maker is risk averse, then receiving the expected value of any gamble with certainty should be preferred to playing the gamble itself—thus the decision maker will accept a certainty equivalent that is lower than the expected value. Comparisons between individuals may be achieved by finding this point of indifference for different individuals playing the same gamble: the more risk-averse individual will have a lower certainty equivalent. However, note that finding a lower certainty equivalent for a single gamble only says something about the comparative risk attitudes *for that gamble* and does not necessarily translate to general statements about global risk attitudes.

Generation of a more general measure of risk attitude requires more sophisticated use of observed choices to determine an implicit utility function that would have led to these choices.² A comparison of the features of the utility function at various points can then

¹Though a multitude of possible framing effects mean that this may not be true in practice. See the section titled “Elicitation Methods.”

²Given a specific decision-making model, for example, expected utility theory.

reveal meaningful quantitative differences between individuals' risk attitudes. There are two broad methods of fitting utility functions to observed choices. Nonparametric methods make no assumptions about the appropriate underlying utility function from which observed choices are generated. They simply define the shape of the utility function from the choices themselves. Since they impose no structure on the utility function, nonparametric approaches can be the most direct translation of actual choices into the risk attitude of an individual. On the other hand, they do imply the specification of a decision model that is presumed to underlie choices. Usually the implicit assumption is that decision making is governed by the normative model of expected utility theory (EUT) [42,43] in which case risk attitudes are entirely determined by the utility function. It is nonetheless difficult to translate a nonparametric utility function over many choices to a single comparable measure of risk attitude.

Where decision making under uncertainty is assumed to be governed by a more complex mechanism than EUT, the conditions that define risk aversion can become considerably more complex than simple examination of the utility function. For example, if the most commonly utilized behavioral model of decision making, cumulative prospect theory (CPT) [44], is assumed to be the correct underlying model, then comparison of risk aversion requires separate examination of the shape of the value function for gains and losses (including loss aversion), and the shape of the decision weighting function that governs the way the decision makers distort subjective probabilities when making decisions [45]. It is possible to elicit both functions nonparametrically [46,47], but this is complex and means that it is impossible to arrive at a simple comparison of risk aversion.

Parametric methods presume both a specific decision model and a specific form for the functions within this model (e.g., a power utility function). The function that is arrived at then is much "smoother" than nonparametric utility—in effect, parametric methods seek to abstract from the noise inherent in observed behavior to provide a "cleaned" estimate of the underlying utility function that

could have produced them. This enables a set of risky choices to be described in just one parameter (or more for complex models), which can then be more readily compared between individuals. However, the choice of model and utility function determines (possible incorrectly) the resulting measures of risk attitude.

Utility-Based Measures for Comparing Risk Attitude

If we are prepared to assume that the individual's decisions are governed by the underlying model of EUT, then Yaari's definition is identical to the requirement that the more risk averse person has a more concave utility function.³ In addition, under EUT one person having a more concave utility function than another ensures that this person also has a lower certainty equivalent than another for all gambles.

Utilities themselves are not unique—any affine transformation of a utility function describes the preference set equally well—so we cannot directly compare risk attitudes by comparing utilities. However, the most prevalent utility-based formulations of risk attitude are the Pratt–Arrow coefficients of constant absolute risk aversion (CARA) and constant relative risk aversion (CRRA) [48,49], which are measures of curvature at a point on the function, standardized by dividing by the first derivative to ensure that the ratios are in comparable units, and so are comparable across individuals [50].

The absolute coefficient (CARA) is a direct measure of the degree of concavity of the utility function at the point of *initial* wealth:

$$-\frac{u''(w_0)}{u'(w_0)}.$$

The relative coefficient (CRRA) measures the degree of aversion to risks over percentage changes of wealth from the point of initial wealth:

³A utility function that has zero curvature expresses risk neutrality, and a convex utility function expresses risk seeking.

$$-w_0 \frac{u''(w_0)}{u'(w_0)}.$$

These coefficients measure the *local* curvature at this point, and thus only refer to risk attitudes to “small” gambles (those with infinitesimal variance). Although an advantage of a parametric utility function approach is that measured risk attitudes can be generalized to gambles that are not included in the initial measurement set, we need to remain aware that this generalization may not be accurate beyond the range of outcome values spanned by this set of sample gambles. Nonetheless, the local Pratt–Arrow coefficients are the most frequently used measures of risk aversion when comparing risk attitudes.

ELICITATION METHODS

Our measures of risk attitude depend on the elicitation method used to determine them; thus different methods may give different results for the same person. Different methods of eliciting risk attitudes have been shown to result in different risk-seeking or risk-averse classification, as well as different rank orderings across methods [51–53]. Thus when using different elicitation methods, we are implicitly comparing both individuals and methods. When we have comparable risk-aversion estimates (e.g., CRRA estimates) *across* elicitation methods, we can check if these estimates are consistent with each other. Some measurement methods give systematically different results than others.

To clarify the ways in which methods may influence measurements, Fig. 1 depicts the possible measurements from the same individuals (the circles) using different elicitation methods (lines A through F).

Specific problems with comparisons of individual risk attitudes, which can be observed from the graph in Fig. 1, include the following:

- *Distance versus Rank.* While A and B maintain rank ordering of individuals, they differ in the distance between them.

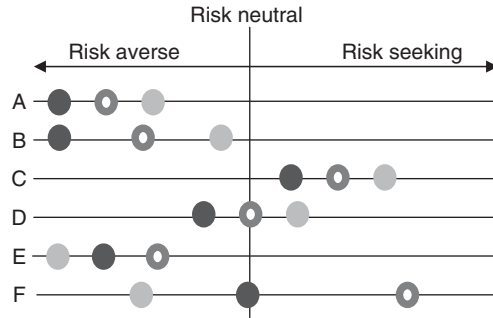


Figure 1. Potential moves in risk-attitude assessment.

- *Rank versus Category.* While A and C maintain rank ordering of individuals, method C reveals risk seeking for all individuals, while A reveals risk aversion.
- *Straddling Risk Neutrality.* Method D reveals risk aversion for one individual, risk neutrality for another, and risk seeking for the last. This means that certainty equivalents vary between positive and negative.
- *Rank versus Rank.* A and E find risk aversion for all individuals, but do not maintain the same rank ordering as A.
- *No Consistency.* Compared to method A, method F results in completely unrelated risk-attitude differences. The rank, dispersion, and risk aversion/seeking have all changed from those in A.

We now review specific methods of eliciting individual risk attitudes, and discuss the strengths and weaknesses of comparing outcomes from each.

Relative Psychometric Methods

Psychometric measurement is the simplest risk-attitude measurement method in terms of elicitation, analysis, and inference. While the use of psychometric techniques is long established in the fields of personality and cognitive psychology, it is relatively new to risk-attitude measurement. Psychometric risk-attitude assessment usually involves posing large sets of simple, usually bipolar,

statements such as “investing in stocks is something I do not do, since it is too risky” [54]. Individuals then indicate agreement along a simple scale of “strongly disagree” to “strongly agree” or similar. Using statistical techniques such as factor analysis, questions that are systematically related are revealed, and represent a potential measurement of the latent trait. More complex models based on item-response theory explicitly model the discrimination of each question, and use weighted scores to generate higher differentiation with fewer items [55].

The psychometric method has a number of advantages: questions are intuitive and quick for individuals to respond to, and do not require numeric or statistical calculations, which avoids confounds such as numeracy [56]. They give a stable measure of risk attitude invariant to changes in outcome expectations [57]. They accord well with self-assessments of relative risk aversion, and correlate in the expected direction with demographic factors such as gender and age [58]. Moreover, most psychometric scales are inherently differentiating across individuals, as the process of determining which questions best represent latent traits leads one to include questions that encourage variation. As a result, psychometric risk-attitude scales give an efficient, reliable, and stable rank ordering of individual risk-attitudes. They score highly on test–retest validity and internal consistency (e.g., Cronbach’s α [59]).

The limitation of psychometric risk-attitude measurement is that individuals are dispersed on a sample-based ordinal scale [60], and thus there is no direct link to quantitative measures of risk aversion. As individuals’ responses do not explicitly map to a utility function, comparisons of individual risk attitudes are restricted to comparisons of rank order. This renders comparisons based on magnitude of difference meaningless, and does not allow the researcher to explicitly categorize the respondent as risk averse or risk seeking.

Initial linkages from psychometric scales to quantitative measure of risk attitude have been encouraging. In a broad and detailed study, psychometric risk attitudes were

found to correlate well with not only standard CRRA survey questions but also CARA estimates of holdings in household portfolios [54]. A carefully controlled longitudinal study revealed that while self-directed investors risk and return expectations changed, their psychometric risk tolerance was remarkably stable [57], and was the best predictor of investment into the stock market over a volatile period when expectations were changing rapidly. In a focused set of graphical risky-choice experiments, individuals with higher psychometric risk tolerance chose riskier investments, even when the outcome distributions were significantly non-normal. Using a CRRA risk attitude to manipulate an experimental design, the researchers provided strong evidence that the psychometric scale was linkable to quantitative measures of risk aversion [61]. Faff *et al.* [13] find that the psychometric risk tolerance measurements are “strongly aligned” to the risk-aversion coefficient implied by lottery choices, and this parallel increases as the stakes of the lottery size increase. Thus, the ease, reliability, and validity of psychometric scales may outweigh the lack of direct mapping to numerical measures of risk aversion.

Experimental Methods and Choice Tasks

Many studies have examined individual risk attitudes using quantitative and experimental techniques, stimuli, and response modes. Importantly, the choice task setup does not necessarily assume any *ex ante* risk attitude, decision theory, or utility functional form. The researcher observes choices and then attempts to explain the choice by fitting a model. Some of the most common elicitation methods are pairwise choice tasks [36,62]; valuation tasks using the Becker *et al.* [28] method or asking for a certainty equivalent or probability equivalent to a risky situation [63]; and ranking tasks [64] or the trade-off method to find points on the utility curve that are equidistant from each other [47]. In many of these studies, statistical analysis is used to estimate parameters of a utility function, such as the CRRA coefficient [41,49,65,66] or the value function in CPT [44], or a nonparametric approach is employed to determine the shape of the utility function.

The benefits of well-designed choice tasks are that they isolate individual risk attitudes and remove potential confounds. The ability to control the interactions between the multivariate factors that can influence risk attitudes enables the researcher to have a much deeper understanding of each in isolation. However, this means that the nature of the interactions can often remain a mystery and reduce the applicability of experimental results to real-world situations.

One of the main disadvantages for experimental choice tasks is that the precision of the elicitation method can often lead to internal inconsistency in the pattern of observed attitudes within individuals. Perhaps the most robust of these is the preference reversal phenomenon [40,67,68], where individuals state their certainty equivalent for two risk situations with the same expected value—one involving a small probability of a large gain (often called the *\$-bet*), and the other involving a large probability of a small win (*P-bet*). Individuals tend to place a higher certainty equivalent (and hence signal a preference) on the *\$-bet*. However, when faced with a direct choice between the *\$-bet* and the *P-bet*, individuals tend to choose the *P-bet* even though this is inconsistent with their previous valuations.

There are also large changes in risk attitude *within individuals* between estimates derived from a first price auction and the BDM (Becker-DeGroot-Marschak) procedure [53]. Isaac and James [53] found that the majority of individuals exhibit more risk seeking in BDM, and the rank ordering of individuals was not preserved. Those that appeared to be the most risk averse in the population under the first price auction tend to be ranked as the most risk seeking in BDM. As a result, comparisons between individuals are difficult to make as they cannot be shown to be consistent even *within* subjects.

Observed Behaviors

Finally, many studies attempt to infer attitude to risk by observing individuals' actions, often in the natural environment in which they make them. This approach

has a number of advantages, most regarding validity. Compared to questionnaire-based assessments, which are usually performed on a demographically restricted sample (i.e., college students), observed behaviors can be assessed in a wide variety of circumstances for more representative samples. Observed behavior studies also benefit from "natural validity," that is, they measure the risk attitude being evidenced in its natural environment. A standard measure of questionnaire and experimental instrument validity is whether it correlates meaningfully with observed behaviors—an easy standard for measures based on observed behaviors to pass.

The downside to observed behaviors is that the reduced control over sample, control variables, and environmental stimuli means that inferences regarding risk attitude may be significantly confounded. As a result, the number of assumptions in observed behavior studies is much larger, the fragility of data analysis is higher, and the ability to make causal inferences is limited. For example, Nosić and Weber [57] found that changes in expectations for stock returns drove changes in the risky allocation proportion, rather than changes in risk aversion. Thus, observed behavior studies may confuse variables that lead to the same outcome—increasing or decreasing apparent risk of the decision—unless confounding variables are controlled for.

A plethora of studies use investment holdings to estimate risk-aversion coefficients [69–75] and compare them across populations. Such studies have also been concerned with the question of whether absolute and relative risk aversion are decreasing, constant, or increasing with wealth levels, with the majority concluding that absolute risk aversion is decreasing [73] and that relative risk aversion is approximately constant or slightly increasing. In addition, many have also used surveys to supplement their understanding of the individuals, resulting in better controlled studies. However, all these studies suffer from the fact that on-line trading accounts most likely represent only a fraction of the individuals' total wealth, and risk-aversion estimates are biased toward a

financially savvy and risk-seeking sample of the public. Much recent activity has used the natural experiment of game shows such as “Jeopardy” and “Deal or No Deal” to estimate risk coefficients [76,77], with the results giving mixed implications for functional forms of CRRA and CARA estimates [78]. The estimated utility function strongly evidenced increasing relative risk aversion (IRRA) [78]. Cohen and Einav structurally estimate risk aversion using car insurance data, and find systematically higher coefficients than much of the existing data [79].

INFERENCES IN COMPARING INDIVIDUALS AND GROUPS

We often want to make general statements such as “person A is more risk averse than person B” or “in general risk aversion increases with age” or even “we must increase a woman’s income by X to make her exhibit risk aversion equivalent to that of an average man.” When comparing risk attitudes in this way, we must ensure the comparison accounts for other factors that may influence the estimated risk attitude. Interestingly, the advice is quite similar for making comparisons in the case of both individuals and groups. Individuals whose circumstances are different may perceive the same objective risk differently, and thus observed behaviors and choices may differ because of the perception of risk, rather than risk attitude [19,80]. An individual’s wealth level, background risk, income risk, and the culture the individual has been raised in all contribute to apparent risky behavior in ways that may bias risk-attitude comparisons. This is not to say that valid comparisons cannot be made, but rather that they must be made carefully, with required individual-level controls.

Control Variables across Individuals and Groups

When making comparisons of risk attitudes, we often ask questions about groups of individuals who differ in some respect. Common questions often involve differences by gender [81], age [75], education level [82], and culture [83]. These variables have been

found to have a systematic effect upon risk-taking behavior across many studies. While the causal mechanism for these differences may vary in direction, it nonetheless means we must account for them when making comparative assessments of individual risk attitude. Whether one should control for these systematic variations depends on the purpose of the comparison.

To make this dependency clear, imagine that you have representative samples from two different groups—group X and group Y. Group X has a significantly higher average risk aversion than group Y according to your measurements. However, group X also has a significantly higher proportion of females than group Y (in the overall population). As there is significant evidence that females are generally more risk averse, the difference in overall measures of risk aversion may be due to differences in gender composition. If the question we are asking is “what is the distribution of risk aversion for an individual belonging to group X?”, then the nonadjusted figure is appropriate—it accurately reflects the probability of observing a given risk attitude conditioned on observing that the individual belongs to group X.

However, if we are seeking to say “belonging to group X is associated with, or causes higher risk aversion,” we are making a different statement that requires adjustment for related factors. We cannot say belonging to group X causes or results in “higher” risk aversion compared to belonging to group Y without controlling for sex. Any grouping variable that is correlated with other factors influencing risk attitudes must be controlled for, before the independent association of risk attitude and grouping can be assessed. Only by controlling for other factors that influence risk attitudes can we say “belonging to group X is associated with higher risk aversion.”

Wealth Level

One of the most common assumptions made in economic models is that individuals’ preferences exhibit CRRA: their attitudes to gambles over identical percentages of wealth remain constant regardless of their wealth levels. This is intuitively plausible—we would expect a millionaire to be more cavalier

about a £ 10 gamble than a pauper; but there is no clear intuition about why a millionaire should be more or less averse to gambling 10% of his total wealth than his companion. Indeed, long-run macrobehavior of the economy would argue that this assumption must be approximately correct. As Campbell and Viceira [50] argue, “per capita consumption and wealth have increased greatly over the past two centuries ... Interest rates and risk premia do not show any evidence of long-term trends in response to this long-term growth; this implies that investors are willing to pay almost the same relative costs to avoid given relative risks as they did when they were much poorer, which is possible only if relative risk aversion is almost independent of wealth.” On the other hand, a number of studies of household portfolios have shown that the share of total household wealth invested in risk assets is increasing with wealth [84–86], which would imply decreasing relative risk aversion (DRRA).⁴

Rabin [66] noted that the risk attitudes that are arrived at by functions fitted to gambles over small amounts result in implausibly high degrees of risk aversion (which are not observed) for choices over large amounts. This casts some doubt on the levels of risk aversion measured through experimental methods using small gambles (though relative comparisons of risk aversion between individuals may still be stable and correct even if the levels are not). Holt and Laury [37] investigated the relationship between real and hypothetical gambles, crossed with small versus large stakes [37]. They found a larger increase in risk aversion along higher stakes in real gambles compared to hypothetical gambles. Similarly, Brinswanger [38] and Kachelmeier and Shehata [39] examined risk aversion over real high stakes corresponding to between one and three months of actual income in rural India and China. They likewise found IRRA as the stake size increased.

⁴One resolution of this apparent conundrum is that less wealthy people face greater relative external (nonfinancial) risks and thus need to compensate for these by reducing the proportion of investible wealth allocated to risky assets.

If individuals do not exhibit CRRA or CARA, then it is much more difficult to compare numerically elicited risk attitudes. Risk aversion evidenced over the same value (\$,£) may vary by background wealth, or the percentage change in wealth, making comparisons between the poor and the rich meaningless.

Culture

Another confounding factor is the individual's culture or background [87]. While risk attitudes, that is, trade-offs of expected risk and return, do not vary tremendously across cultures [80,88], risk perceptions do [89]. Subsumed within the general model is the fact that *perceived* risk is what is relevant when understanding risk attitudes [90,91]. Cultural differences posited to explain these differences in risk perception include the individualism/collectivism dimension [80,92], trust in institutions such as the stock market [83,93], and propensity for regret [83]. These differences mean that while comparisons of risk attitude *within* culture are relatively straightforward, risk attitudes across cultures are less clear. Once individual differences in circumstances are controlled for, the relationship between investment choices and risk attitudes may be the same for all cultures. Thus, controls for relevant individual-level circumstances are necessary for making statements about comparative risk attitude viable.

GUIDELINES FOR DESIGNING STUDIES

As much of the preceding review has shown, comparisons of risk attitudes have many pitfalls deriving from the elicitation method, domain, culture, and measurement precision used. However, this should not deter would-be researchers from investigating differential risk attitudes. In this section, we discuss the guidelines that a researcher can use when designing a study or analyzing data comparing risk attitudes and their determinants.

The ideal risk-attitude measurement study would involve triangulation using repeated measurements across different elicitation methods. First, the individual should

answer enough questions such that the size of the noise (error term) *within elicitation method* can be estimated. This allows us to specify the degree to which differences in estimates are based on consistent differences, rather than noise. Second, individuals should answer questions *across measurement methods*. This allows us to determine the extent to which a given elicitation method may systematically result in higher or lower apparent risk aversion or seeking. The researchers can then decompose the observed estimates into method-specific and individual-specific variance, which allows them to use the *intraclass correlation coefficient*. This coefficient is a measure of the number of choices that are explained by fixed individuality, while allowing task structure and noise to be independently measured [94,95].

Survey and Experimental Designs

When designing a study investigating risk attitudes across individuals, the researchers should first carefully think about exactly what risk attitude they are attempting to measure. There is large extant literature in many domains on how risk is defined, and the researchers should have a clear understanding of the attractions, downsides, and variance of potential outcomes specific to that domain.

Secondly, the researchers should consider what type of analysis they need to perform to answer their research question appropriately. This will often drive the type of measurement scale that must be used to define the risk attitude, such as ordinal or interval. Some risk attitudes are more amenable to quantification (e.g., investing as opposed to social risk), and thus in many domains the measurement scale may be constrained to being ordinal. Given the higher demands that are placed on individuals responding to quantitative and probabilistic questions (used to construct interval and cardinal scales), and the fragility of such measurements, researchers should use them only if magnitudes of differences between individuals are core to the hypothesis.

Finally, the researchers need to choose an elicitation method. The method often codepends on the measurement scale that

can be determined from it, but there are often many ways of obtaining measurements using different methods. All elicitation methods have their own strengths and weaknesses in terms of ease and brevity, stability and validity of responses, and ability to discriminate between noise in responses and significant differences. An ideal elicitation method would combine the positive aspects mentioned above.

Observational Studies

As noted above, the main limitation in observational studies is the potential sample selection, a lack of controls for confounding variables, endogeneity of individual circumstances, and a lack of control over stimuli. When analyzing observational data, there are two key considerations that a researcher must answer. First, on what basis are we attributing the observed choices to differential risk attitudes as opposed to other noncontrolled variables such as education, familiarity, or nonrisk preferences? Second, are measurements valid as a measure of the relevant population, or might they reflect self-selection in riskier or less risky activities?

Recently much progress has been made in mitigating these shortcomings using “natural experiments” [96]. Natural experiments allow the researcher to specifically manipulate the natural circumstances in which a subject is operating, which can reduce sample bias, improve control of stimuli, and even provide strong control variables [97].

REFERENCES

1. Harris J. No two alike: human nature and human individuality. New York: W.W. Norton; 2006.
2. Bouchard T, McGue M. Genetic and environmental influences on human psychological differences. *J Neurobiol* 2003;54:4–45.
3. Cesarini D, *et al.* Genetic variation in financial decision making. *J Finance* 2009. In press.
4. Zyphur M, *et al.* The genetics of economic risk preferences. *J Behav Decis Making* 2009; 22(4):367–377.
5. Levin I, *et al.* Stability of choices in a risky decision-making task: a 3-year longitudinal

- study with children and adults. *J Behav Decis Making* 2006;20(3):241–252.
6. Sahm C. How much does risk tolerance change? 2008.
 7. MacGregor DG, *et al.* Imagery, affect, and financial judgment. *J Psychol Financ Mark* 2000;1:104–110.
 8. de Vries M, Holland RW, Witteman CLM. In the winning mood: affect in the Iowa gambling task. *Judgm Decis Mak* 2008;3(1):42–50.
 9. Hsee C, Rottenstreich Y. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol Sci* 2001;12(3):185–190.
 10. Payne JW, Bettman JR, Schkade DA. Measuring constructed preferences: towards a building code. *J Risk Uncertain* 1999;19(1–3): 243–270.
 11. Slovic P. The construction of preference. *Am Psychol* 1991;50(5):364–371.
 12. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981;211:453–458.
 13. Faff RW, Mulino D, Chai D. On the linkage between financial risk tolerance and risk aversion: evidence from a psychometrically-validated survey versus an online lottery choice experiment (November) 22, 2006. Available at SSRN: Available at <http://ssrn.com/abstract=946679>.
 14. Deck C, *et al.* Measuring risk attitudes controlling for personality traits. Working Paper 0801; 2008.
 15. Kumar A. Who gambles in the Stock Market? *J Finance* 2009;LXIV(4):1889–1933.
 16. Soane E, Chmiel N. Are risk preferences consistent? The influence of decision domain and personality. *Pers Individ Dif* 2005;38(8): 1781–1791.
 17. Nicholson N, *et al.* Personality and domain-specific risk taking. *J Risk Res* 2005;8(2): 157–176.
 18. Blais A-R, Weber EU. A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgm Decis Mak* 2006;1(1):33–47.
 19. Weber EU, Milliman RA. Perceived risk attitudes: relating risk perception to risky choice. *Manage Sci* 1997;43(2):123–144.
 20. Coval J, Moskowitz TJ. Home bias at home: local equity preference in domestic portfolios. *J Finance* 1999;LIV(6):2045–2073.
 21. Graham JR, Harvey CR, Huang H. Investor competence, trading frequency, and home bias. *Manage Sci* 2009;55(7):1094–1106.
 22. Blais A-R, Weber EU. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgm Decis Mak* 2006;1(1):1–14.
 23. Herrnstein RJ, Prelec D. Melioration: a theory of distributed choice. *J Econ Perspect* 1991;5(3):137–156.
 24. Ellsberg D. Risk, ambiguity, and the Savage axioms. *Q J Econ* 1961;75:643–669.
 25. Samuelson P. Risk and uncertainty: a fallacy of large numbers. *Scientia* 1963;98:108–113.
 26. Benartzi S, Thaler RH. Myopic loss aversion and the equity premium puzzle. *Q J Econ* 1995;110(1):73–92.
 27. Braga J, Starmer C. Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environ Resour Econ* 2005;32:55–89.
 28. Becker GM, DeGroot MH, Marschak J. Stochastic models of choice behaviour. *Behav Sci* 1963;8:41–55.
 29. Loomes G, Sugden R. Testing different stochastic specifications of risky choice. *Economica* 1998;65:581–598.
 30. Harless DW, Camerer CF. The predictive utility of generalized expected utility theories. *Econometrica* 1994;62(6):1251–1289.
 31. Luce RD. Individual choice behavior. New York: John Wiley & Sons, Inc.; 1959.
 32. Carbone E, Hey JD. A comparison of the estimates of expected utility and non-expected-utility preference functionals. *Geneva Papers Risk Insur* 1995;20:111–133.
 33. Birnbaum MH, Patton JN, Lott MK. Evidence against rank-dependent utility theories: tests of cumulative independence, interval independence, stochastic dominance, and transitivity. *Organ Behav Hum Decis Processes* 1999;77(1):44–83.
 34. Loomes G, Moffatt PG, Sugden R. A microeconomic test of alternative stochastic theories of risky choice. *J Risk Uncertain* 2002; 24(2):103–130.
 35. Machina MJ. Stochastic choice functions generated from deterministic preferences over lotteries. *Econ J* 1985;95:575–594.
 36. Hey JD, Orme C. Investigating generalizations of expected utility theory using experimental data. *Econometrica* 1994;62(6): 1291–1326.
 37. Holt CA, Laury SK. Risk aversion and incentive effects. *Am Econ Rev* 2002;92(5): 1644–1655.
 38. Binswanger HP. Attitudes towards risk: experimental measures in rural India. *Am J Agric Econ* 1980;62:395–407.

39. Kachelmeier SJ, Shehata M. Examining risk preferences under high monetary incentives: experimental evidence from the People's Republic of China. *Am Econ Rev* 1992;82(5): 1120–1141.
40. Grether DM, Plott CR. Economic theory of choice and the preference reversal phenomenon. *Am Econ Rev* 1979;69:623–638.
41. Yaari ME. Some remarks on measures of risk aversion and on their uses. *J Econ Theory* 1969;1:315–329.
42. von Neuman J, Morgenstern O. The theory of games and economic behaviour. 2nd ed. Princeton (NJ): Princeton University Press; 1947.
43. Savage LJ. The foundations of statistics. New York: Dover; 1954.
44. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5: 297–323.
45. Davies GB, Satchell SE. The behavioural components of risk aversion. *J Math Psychol* 2006;51(1):1–13.
46. Bleichrodt H, Pinto JL. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Manage Sci* 2000;46(11):1485–1496.
47. Abdellaoui M. Parameter-free elicitation of utility of probability weighting functions. *Manage Sci* 2000;46(11):1497–1512.
48. Pratt JW. Risk aversion in the small and in the large. *Econometrica* 1964;32:122–136.
49. Arrow K. Aspects of the theory of risk bearing. Helsinki: Yrjo Jahnssonin Saatio; 1965.
50. Campbell JY, Viceira LM. Strategic asset allocation: portfolio choice for long-term investors. Clarendon lectures in economics. Oxford: Oxford University Press; 2002.
51. Slovic P. Assessment of risk taking behavior. *Psychol Bull* 1964;61:330–333.
52. Hartinger A. Do generalized expected utility theories capture persisting properties of individual decision makers? *Acta Psychol* 1999;102(1):21–42.
53. Isaac RM, James D. Just who are you calling risk averse? *J Risk Uncertain* 2000;20(2): 177–187.
54. Kapteyn A, Teppa F. Subjective measures of risk aversion and portfolio choice. RAND Labor and Population Program, Working Paper Series, 02-03; 2002.
55. Baker FB, Kim S-H. Item response theory: parameter estimation techniques. 2nd ed. Monticello (NY): Marcel Dekker Inc.; 2004.
56. Peters E, *et al.* Numeracy and decision making. *Psychol Sci* 2006;17(5):407–413.
57. Nasic A, Weber M. Determinants of risk taking behavior: the role of risk attitudes, risk perceptions and beliefs. Unpublished; 2007.
58. Hallahan T, Faff RW, McKenzie MD. An empirical investigation of personal financial risk tolerance. *Financ Serv Rev* 2004;13: 57–78.
59. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16(3):297–334.
60. Gob R, McCollin C, Ramalhoto MF. Ordinal methodology in the analysis of likert scales. *Qual Quant* 2007;41(5):601–626.
61. Egan D, Brooks P, Davies GB. Psychometric risk tolerance, exponential utility, and preferences over risky distributions. Working Paper; 2010.
62. Barsky RB, *et al.* Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement study. *Q J Econ* 1997;112(2):537–579.
63. Hershey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: are they equivalent. *Manage Sci* 1985;31(10):1213–1231.
64. McFadden D. The choice theory approach to market research. *Market Sci* 1986;5(4): 275–297.
65. Pratt JW, Zeckhauser R. Proper risk aversion. *Econometrica* 1987;55:143–154.
66. Rabin M. Risk aversion and expected-utility theory: a calibration theorem. *Econometrica* 2000;68(5):1281–1292.
67. Lichtenstein P, Slovic P. Reversals of preference between bids and choices in gambling decisions. *J Exp Psychol* 1971;89:46–55.
68. Tversky A, Slovic P, Kahneman D. The causes of preference reversal. *Am Econ Rev* 1990; 80(1):204–217.
69. Schooley DK, Worden DD. Risk aversion measures: comparing attitudes and asset allocation. *Financ Serv Rev* 1996;5(2):87–99.
70. Huberman G, Dorn D. Preferred risk habit of individual investors. *J Financ Econ* 2010;97(2010):155–173.
71. Dorn D, Huberman G. Talk and action: what individual investors say and what they do. *Rev Finance* 2005;9:437–481.
72. Blume ME, Friend I. The asset structure of individual portfolios and some implications for utility functions. *J Finance* 1975;30(2): 585–603.

73. Friend I, Blume ME. The demand for risky assets. *Am Econ Rev* 1975;65(5):900–922.
74. Calvet LE, Campbell JY, Sodini P. Down or out: assessing the welfare costs of household investment mistakes. *J Polit Econ* 2007; 115(2):707–747.
75. Morin RA, Suarez AF. Risk aversion revisited. *J Finance* 1983;38(4):1201–1216.
76. Metrick A. A natural experiment in “Jeopardy!” *Am Econ Rev* 1995;85(1):240–253.
77. Blavatsky P, Progrebna G. Testing the predictions of decision theories in a natural experiment when half a million is at stake. IEW Working Paper 291; 2006.
78. Post T, *et al.* Deal or no deal? Decision making under risk in a large-payoff game show. *Am Econ Rev* 2008;98(1):38–71.
79. Cohen A, Einav L. Estimating risk preferences from deductible choice. NBER Working Paper W11461; 2005.
80. Weber EU, Hsee CK. Cross-cultural differences in risk perception, but cross-cultural similarities in attitudes towards perceived risk. *Manage Sci* 1998;44(9):1205–1217.
81. Schubert R, *et al.* Financial decision making: are women really more risk-averse? *Am Econ Rev* 1999;89(2):381–385.
82. Haliassos M, Bertaut C. Why do so few hold stocks? *Econ J* 1995;105:1110–1129.
83. Statman M. Countries and culture in behavioral finance. CFA Institute Conference Proceedings Quarterly, 2008.September.
84. Kessler D, Wolf EN. A Comparative Analysis of Household Wealth Patterns in France and the United States. *Review of Income and Wealth* 1991;37:249–266.
85. Guiso L, Jappelli T, Terlizzese D. Income risk, borrowing constraints, and portfolio choice. *Am Econ Rev* 1996;86(1):158–172.
86. Ogaki M, Zhang Q. Decreasing relative risk aversion and tests of risk sharing. *Econometrica* 2001;69(2):515–526.
87. Weber E, Hsee C. Culture and individual judgment and decision making. *Appl Psychol Int Rev* 2000;49(1):32–61.
88. Hsee CK, Weber EU. Cross-national differences in risk preference and lay predictions. *J Behav Decis Making* 1999;12:165–179.
89. Weber E, Hsee C. Cross-cultural differences in risk perception. *Manage Sci* 1998;44(9): 1205–1218.
90. Weber EU, Anderson CJ, Birnbaum MH. A theory of perceived risk and attractiveness. *Organ Behav Hum Decis Processes(Print)* 1992;52(3):492–523.
91. Payne JW. Relation of perceived risk to preferences among gambles. *J Exp Psychol Hum Percept Perform* 1975;104(1):86–94.
92. Hofstede DG. Culture’s consequences: comparing values, behaviors, institutions and organizations across nations. 2nd edn. London: Sage Publications, Inc.; 2003.
93. Guiso L, Sapienza P, Zingales L. Trusting the stock market. *J Finance* 2008;63(6): 2557–2600.
94. Snijders T, Bosker R. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publishers; 1999.
95. Stockard J, O’Brien RM, Peters E. The use of mixed models in a modified Iowa Gambling Task and a prisoner’s dilemma game. *Judgm Decis Mak* 2007;2(1):9–22.
96. Harrison GW, List JA. Field experiments. *J Econ Lit* 2004;42(4):1009–1055.
97. Harrison GW. Field experiments and control. In: Carpenter J, Harrison GW, List JA, editors. *Field experiments in economics*. Greenwich (CT): JAI Press; 2004.

COMPETING RISKS AND LIMITED FAILURE

SANJIB BASU
Division of Statistics,
Northern Illinois
University,
DeKalb, Illinois

INTRODUCTION

In engineering, biomedical, or other studies involving time-to-event outcomes, the units or subjects involved can sometimes experience the time-to-event in different ways. In engineering studies, the event of interest is typically failure of the unit, whereas in biomedical studies the event can be survival or recurrence of disease. The event often results in termination (or restart) of follow-up or inspection process, so that occurrence of the event in one way hinders the occurrence of other types of events. In such cases, the statistical analysis can be performed using the theory of competing risks. Each unit or subject is exposed to many competing risks, one of which causes the failure, unless the unit/subject is right-censored, that is, it does not fail within the follow-up duration. Examples of failure data obtained under competing risks are abundant in reliability and biomedical applications. In engineering applications, the individual causes or risks may signify the risks of failure of individual components or subsystems that comprise an entire system. Occurrence of a system failure is caused by the earliest onset of any of these component failures. In this respect, the framework is that of a system with components connected in a series. Alternatively, the competing risks may signify the multiple modes of failure for a complex unit. In biomedical setting, the competing causes of failure (a failure may indicate death or relapse or other time-to-event outcome) typically refer to various potential

risk factors for a subject (or an animal or a cell) observed in a study. For example, for a breast cancer patient, the competing risks may include breast cancer, other cancers, heart disease, as well as risk from other diseases. The effects of the other competing risks may play an important role in survival studies on slowly progressing diseases.

In some lifetime applications, a fraction of the subjects or units continue to survive one or some of the risks. A classic example of this in industrial applications is the case of infant mortality where some of the units may fail from the risk of infant mortality while others may survive it. The surviving units however, may fail from other risks, such as the risk of wear out. It is seen from an analysis of cancer survival data that an increasing proportion of patients are being cured of cancer because of advances made in treating the same. These patients, however, continue to be exposed to other risks/diseases.

In applications of lifetime data associated with competing causes of failure, an additional complication may arise when the true cause of failure is not exactly identified for a certain subset of the units or subjects. The causes of failure of such items are generally referred to as *masked*. Masking can be complete or only partial when the cause of failure of the unit is narrowed down to a subset but is not exactly identified. In engineering applications, masking often results from an attempt to expedite the process of repair by replacing the entire subset of components responsible for failure instead of further investigation toward identifying the specific cause for failure. In biomedicine, masking indicates lack of or partial knowledge for the cause of the event (death, relapse, etc.), which could be due to lack of complete investigation or partial loss of patient record. In practice, one possibility is to carry out a second stage analysis to uniquely determine the cause; however, statistical inference is possible even when the cause of failure is masked for a subset of systems.

COMPETING RISKS AND LIMITED FAILURE MODELS

The Limited Failure Model

In industrial applications, the occurrence of early or “infant mortality” failures is an important problem. A production lot may contain a few defective units even after a quality control screening. These units with manufacturing or other defects will usually lead to an infant-mortality failure early in their lifetime after they have been operated for some period of time. The nondefective units, on the other hand, are expected to operate unless they are abused in their application and are not expected to fail from the risk of infant mortality. Meeker [1] and Chan and Meeker [2] termed this a *limited failure population*. We note that though the limited failure population is considered in the context of risk from infant mortality, it may apply to other risks of failure as well.

In the context of survival data, limited failure models, which are known as models with surviving or cure fractions, have been considered mostly in the analysis of cancer survival or recurrence. As cancer therapy progresses, the curability of many cancers is becoming a reality. The treatment of cancer has shown substantial progress with an increasing proportion of patients being cured from many types of cancers. The chances of being cured and the survival time since diagnosis are of interest to cancer patients and the medical community alike. From a statistical perspective, when the survival or time-to-recurrence curve for cancer tend to plateau at a value strictly greater than 0, it is taken as an indication of the presence of a proportion of cured patients for whom cancer will not recur.

The probability of being failure-free, also known as the *cure rate* or the *surviving fraction*, is defined as the asymptotic value of the survival function $S(t)$ as $t \rightarrow \infty$, that is,

$$p = \lim_{t \rightarrow \infty} S(t). \quad (1)$$

Here, $S(t) = P(T > t)$ is the survival function for the underlying random variable T . Whenever $p > 0$, the random variable T has a probability mass at infinity (or at an arbitrarily large time point).

Different models have been proposed to analyze time-to-event data with a failure-free fraction. The mixture cure model of Boag [3] (see also Berkson and Gage [4] and Farewell [5], and Maller and Zhou [6]) assumes that a fraction of the units or individuals are failure-free from time 0 and will never experience the risk of interest (infant mortality or cancer). The remaining fraction are exposed to the primary risk (as they have manufacturing or other defects in the case of infant mortality or not-cured in the setting of cancer survival). An alternative cure rate model, known as the *bounded cumulative hazard model*, was proposed in Yakovlev *et al.* [7] and has a growing literature [8,9]. This model defines an asymptote for the cumulative hazard and hence for the survival function.

Boag [3] first proposed the two-component mixture cure model. Berkson and Gage [4] proposed a model that included a background mortality rate and a constant excess mortality for the uncured group. Anscombe [10], Danziger [11], Goldman [12], Greenhouse and Wolfe [13], and Lloyd and Joe [14] considered the mixture cure model under exponential distribution assumption, whereas Bandes and Nadas [15] used the lognormal distribution and Steinhurst [16] used the Weibull distribution. De Angelis *et al.* [17] and Phillips *et al.* [18] simultaneously modeled the cure fraction and the hazard in the uncured group. Verdecchia *et al.* [19] used similar models to estimate and project cancer prevalence. Gamel *et al.* [20], Yu *et al.* [21], and Yu and Tiwari [22] modeled the relative survival from cancer. For an extensive list of reference on mixture cure model, see Tsodikov *et al.* [9].

Meeker [1] provides an example of a limited failure population in an engineering setting that involves life test of 4156 integrated circuits from a pilot production process. In the life test, 28 failures were recorded up to 593 h. The test was continued until 1370 h without another failure. Both engineering judgment and analysis of the data by Meeker indicated that a few more failures would have been observed if the

test on the other 4128 units had been continued.

Competing Risks

Competing-risks failure data and their analysis have an extensive literature and range from studies on survival analysis in biostatistics to applications of reliability in engineering to risk models in actuarial science. Crowder [23] is an excellent resource for statistical theory and analysis of competing risks. The more recent book by Pintilie [24] emphasizes practical application of competing risks theory, mostly in biomedical applications. The 2006 special issue of JSPI [25] contains a wealth of references. In recent engineering applications, Langseth and Lindqvist [26] considered an interesting application of competing risks models in the repairable systems, whereas Bunea and Mazzuchi [27] studied competing failure models in accelerated life testing. Sun and Tiwari [28] analyzed the failure times of small electric appliances that may fail due to two competing risks. Parametric analyses of the competing-risks model were proposed by David and Moeschberger [29], Lagakos [30], and Prentice *et al.* [31]. Cause-specific hazard functions were used in nonparametric estimation by Nelson [32], Aalen [33], and Crowder [23]. Semiparametric methods based on proportional-hazards models were discussed by Kalbfleisch and Prentice [34] and Lawless [35].

The latent failure times approach to competing risks [29], which is also the prevalent approach in most engineering applications, is based on latent failure times $Y_r, r = 1, \dots, R$, corresponding to time-to-failure from competing risks $r = 1, \dots, R$ with associated marginal survival function $S_r(\cdot)$ and marginal hazard $h_r(\cdot)$. Thus, while the units or subjects are exposed to $R \geq 2$ competing risks acting simultaneously, this approach is based on the potential failure time Y_r from risk r that would be observed if the possibility of failure from causes other than r were removed from the unit or subject. The observed lifetime T is then taken to be $T = \min(Y_1, \dots, Y_R)$. If $S(t) = P(T > t)$ is the survival function

for the observed lifetime T , then under the assumption that the potential failure times $Y_1, \dots, Y_R$ from the R competing risks are independent we have $S(t) = \prod_{j=1}^R S_j(t)$. The potential failure time model has been criticized [31,34] since in many cases, especially in biomedical applications, either it is not feasible, or it is physically impossible to remove all other risks from the system. The alternative cause-specific hazard formulation [31,34,36]) is based on the joint distribution of the observed survival time T and cause-of-failure C . In particular, let $S(r, t) = P(C = r, T > t)$ be a “subsurvival” function with $S(t) = \sum_{r=1}^R S(r, t)$ denoting the marginal survival function of T . The corresponding subdensity function is denoted as $f(r, t)$. The cause-specific or subhazard function

$$\begin{aligned} h(r, t) &= \lim_{\delta \rightarrow 0} \frac{P(C = r, T \leq t + \delta | T > t)}{\delta} \\ &= \frac{f(r, t)}{S(t)} \end{aligned} \quad (2)$$

is the instantaneous failure rate from cause r at time t after surviving all risks $1, \dots, R$ up to time t . The cause-specific subhazard functions $h(r, t)$ are thus defined in terms of the joint distribution of the observed time and cause of failure (T, C) . The overall hazard from all risks combined is $h(t) = \sum_{r=1}^R h(r, t)$.

Independence of the potential failure times $Y_1, \dots, Y_R$ implies that failure time from risk r under one set of study conditions in which all R risks are operative is precisely the same as under an altered set of conditions in which all risks except the r th risk have been removed. However, the elimination of certain risks may well alter the hazards from other causes making the independence assumption of questionable validity [30]. If independence is assumed, then the latent and cause-specific approaches lead to identical statistical inference [23,34]. Moreover, for any competing risks model with a joint survivor function $\bar{F}(y_1, \dots, y_R)$ that may imply dependence between $Y_1, \dots, Y_R$, there exists a different joint survivor function in which $Y_1, \dots, Y_R$ are independent such that both models reproduce the same set of sub-survival functions $S(r, t)$. In particular, one

cannot distinguish between the dependent model and the proxy independent model based on observations on (C, T) alone. This is the well-known identifiability problem [23,37,38]. In fact, each such dependent model has a whole class of proxy models (not necessarily independent). Heckman and Honore [39] showed that when there are explanatory variables in the model, identification of the joint survivor function $\bar{F}(y_1, \dots, y_R)$ is possible from the subsurvivor functions within a certain framework. Slud [40] proved a result of similar spirit under a different setup. Identifiability can also be regained with specific parametric forms. Basu and Ghosh [41,42] and Basu and Klein [43] established identifiability of competing risks models for many parametric distributions. Moeschberger and Klein [44] provided a review of analytic approaches for handling competing risks data when the independence assumption is suspect. We note here that parametric forms are common in reliability applications, whereas non- or semiparametric ones are more popular in biomedical applications.

Another approach to competing risks is the relative survival approach, which is sometimes used in biomedical, especially epidemiological, applications. In this approach, the “expected survival” from causes other than the primary risk is estimated from the general population and the relative survival from the primary risk is obtained as the ratio of the overall survival to the expected survival from other causes. Gamel *et al.* [20], Yu *et al.* [21], Yu and Tiwari [22], and Lambert *et al.* [45] considered the relative survival approach. The relative survival is the same as the net survival from cancer if expected survival (from other risks) in the general population is assumed to be the same as the net survival from other causes for the units under study and if the risks of primary risks and other causes are assumed to act independently. This independence assumption is implicit in many works that use the relative survival approach; without this assumption, relative survival can be interpreted only as a ratio [46]. The *independence assumption*, however, is not testable within a fully competing risks framework

(without further structural or parametric assumptions) due to identifiability issues as noted before [23,34,38].

If *independence* holds, it follows that the marginal hazards $h_r(t)$ equals the subhazards $h(r, t)$. The equality of marginal and subhazards is, in fact, weaker than the independence assumption [23] and is often known as the *Makeham assumption* [47].

Competing Risks and Limited Failure

Let $\{(T_i, C_i), i = 1, \dots, n\}$ denote the data from n units or individuals where T_i is the time-to-event and C_i denotes the cause-of-event for the i th unit. Here $C = 1, \dots, R$ denotes the R competing risks and $C = 0$ denotes the primary risk under study. We use $C = 0$ to denote the case when the event is right-censored and assume that the censoring process is noninformative.

The mixture cure competing risks model postulates that fraction p of the units or individuals will be failure free from the primary risk, whereas the remaining $(1 - p)$ fraction will eventually fail from the primary risk if complete follow-up were possible. The population is thus assumed to be a mixture of “risk-free” (from primary risk) and “risk-prone” groups, but the separation between these two groups is latent. In particular, for each unit i , let Q_i be a latent binary indicator, with $Q_i = 0$ and $Q_i = 1$ denoting the risk-prone and risk-free cases, respectively. The joint likelihood of time and cause can then be written as

$$p(C_i = r_i, t_i) = (1 - p)P(C_i = r_i, t_i | Q_i = 0) + pP(C_i = r_i, t_i | Q_i = 1). \quad (3)$$

Basu and Tiwari [48] modeled these two terms corresponding to the latent risk-free and risk-prone groups separately. In particular, in their cause-specific hazard formulation, let $h(C = r, t | Q = q) = h(r, t | q)$ denote the cause-specific hazard from risk r in latent group $Q = q$. When $Q = 1$, that is, in the risk-free group, the units are failure-free from the primary risk of $r = 1$. In particular, there is no hazard from Risk 1 in this group, or $h(C = 1, t | Q = 1) = h(1, t | 1) \equiv 0$ for all $t \geq 0$. The overall hazard in the

Table 1. Hazard Functions of Common Lifetime Distributions

Distribution	$h(t)$	μ, τ	Transformation	$h^0(z)$
Exponential	λ	$\mu = -\log \lambda, \tau = 1$	$z = \log t - \mu$	$\exp(z)$
Weibull	$\lambda \gamma (\lambda t)^{\gamma-1}$	$\mu = -\log \lambda, \tau = \gamma$	$z = \tau(\log t - \mu)$	$\exp(z)$
Lognormal	$\frac{\tau \phi(z)}{1 - \Phi(z)}$	$\mu = \lambda, \tau = \gamma^{-1}$	$z = \tau(\log t - \mu)$	$\frac{\tau \phi(z)}{1 - \Phi(z)}$
Log-Logistic	$\frac{\tau \exp(z)}{1 + \exp(z)}$	$\mu = \lambda, \tau = \gamma^{-1}$	$z = \tau(\log t - \mu)$	$\frac{\tau \exp(z)}{1 + \exp(z)}$
Gumbel	$\frac{\exp(e^{-z}) - 1}{\tau \exp(-z)}$	$\mu = \lambda, \tau = \gamma^{-1}$	$z = \tau(\log t - \mu)$	$\frac{\exp(e^{-z}) - 1}{\tau \exp(-z)}$
Gompertz	$\gamma \exp(\lambda t)$			
Gamma	$\frac{t^{\gamma-1} \exp(-\lambda t)}{\int_t^\infty x^{\gamma-1} \exp(-\lambda x) dx}$			

risk-free and risk-prone groups is given by $h(t|Q=0) = \sum_{r=1}^R h(r, t|Q=0)$ and $h(t|Q=1) = \sum_{r=2}^R h(r, t|Q=1)$, respectively.

The individual $(2R-1)$ cause-specific hazards $\{h(r, t|Q=1), r=2, \dots, R\}$ and $\{h(r, t|Q=0), r=1, \dots, R\}$ for the two latent groups can be modeled using standard hazard models, either semiparametrically as a piecewise constant function or parametrically. Parametric models are more common in engineering applications. A general class of parametric models is given by the location-scale structure

$$h(r, y|q) = \tau_{qr} h_{qr}(\tau_{qr}(y - \mu_{qr})) = \tau_{qr} h_{qr}^0(z), \quad (4)$$

where $z = \tau_{qr}(y - \mu_{qr})$ with some abuse of notations and the base hazards $h_{qr}^0(\cdot)$ are free of parameters. This location-scale family includes a large family of distributions that are traditionally used to model lifetime data (when the failure time is suitably transformed, say, to a log-survival time) and includes the exponential, Weibull, log-normal, log-logistic and Gumbel distributions (Table 1). The Gamma, Gompertz and three-parameter generalized Gamma distributions [21] do not have the location-scale structure; however, they also have tractable parametric forms.

Let θ_{qr} denote the parameters in the sub-hazard $h(r, t|q)$ (for example, $\theta_{qr} = (\mu_{qr}, \tau_{qr})$) and let $\theta_0 = (\theta_{0r}, r=1, \dots, R)$, $\theta_1 = (\theta_{1r}, r=2, \dots, R)$ denote the collection of parameters from the subhazards of all the competing

risks in the two latent groups. Basu and Tiwari [48] considered a Bayesian approach and assumed a joint prior distribution $p(\theta_0, \theta_1, p)$ on the collection of parameters. For example, one can assume prior independence across the latent groups and the risks and assign independent priors for the θ_{qr} parameters, $q=0, 1, r=1, \dots, R$.

On the basis of these general formulation with R competing risks and limited failure from the primary risk ($r=1$), the likelihood of cause and time of event data from n units or individuals $\{t_i, c_i, i=1, \dots, n\}$ can be written as

$$\begin{aligned} & \prod_{i=1}^n \left[(1-p) h(c_i, t_i|Q=0, \theta_{0,c_i})^{I(c_i \neq 0)} \right. \\ & \quad \times \exp \left\{ - \sum_{r=1}^R H(r, t_i|Q=0, \theta_{0r}) \right\} \\ & \quad I(c_i \neq 1) p h(c_i, t_i|Q=1, \theta_{1,c_i})^{I(c_i \neq 0)} \\ & \quad \left. \times \exp \left\{ - \sum_{r=2}^R H(r, t_i|Q=1, \theta_{1r}) \right\} \right]. \quad (5) \end{aligned}$$

We recall here that $c_i = 0$ is used to denote the right-censored cases. Further, if a unit failed due to the primary risk ($c_i = 1$), then it cannot be risk-free from the primary risk and thus the second term in Equation (5) does not apply in this case. Finally, the case of masked causes (when the cause-of-event c_i is not exactly recorded) is not covered

in Equation (5); this case is deferred to the section titled “Masked Causes”.

MASKED CAUSES

The causes of event for some of the units or individuals sometimes are not exactly identified or recorded. This is known as *masking* in competing risks literature. Partial masking refers to the case when the cause is narrowed down to a subset of the risks $\{1, \dots, R\}$ but not exactly identified, whereas complete masking means that the cause can be any one of $\{1, \dots, R\}$. Early works on competing risks with masked causes include Racine-Poon and Hoel [49], who establish a nonparametric estimate of the survival function, while Dinse [50] proposes nonparametric maximum likelihood estimators of prevalence and mortality. Dinse [51] and Kodell and Chen [52] considered bioassays for animal carcinogenicity where masking arises due to the disagreement on the reliability of the cause-of-death information. Several other authors also discuss the problem of missing cause-of-death in carcinogenicity studies [53,54]. Goetghebeur and Ryan [55] and Dewanji [56] construct a log-rank test to assess the difference between survival functions for subgroups of the population under study in the presence of covariates. Goetghebeur and Ryan [57] subsequently generalized the approach to proportional cause-specific hazards regression models. Lu and Tsiatis [58] utilized multiple imputations to generalize this to the case when the baselines are not proportional. Flehinger *et al.* [59,60] consider the analysis of datasets in which there are second stage data. They propose maximum likelihood estimation using a model with nonparametric proportional cause-specific hazards [59] and a model with completely parametric cause-specific hazards [60]. Dewanji and Sengupta [61] developed an Expectation-Maximization (EM) algorithm in the setting of masked but grouped survival data. Lu and Tsiatis [62] compared the two partial likelihood approaches in masked data. Basu *et al.* [63,64] developed Markov chain sampling based on Bayesian analysis for masked data.

Basu and Tiwari [48] considered analysis of unknown causes of death cases in breast cancer patients.

As noted before, the observed data in the absence of masking from n units or individuals are $\{t_i, c_i, i = 1, \dots, n\}$, where t_i is the time and c_i is the cause of event (with $c_i = 0$ denoting right-censoring). When some of the causes-of-events are masked, we instead represent the data as $\{t_i, s_i, i = 1, \dots, n\}$ where s_i denotes a set belonging to the power set of $\{1, \dots, R\}$ (i.e., s_i is a subset of $\{1, \dots, R\}$). For those units for which the cause of event is exactly identified as c_i , the set s_i is a singleton, that is, $s_i = \{c_i\}$. Right-censoring, as before, is denoted as $s_i = \{0\}$. When the cause is not exactly identified, s_i is not a singleton but is a subset of $\{1, \dots, R\}$ (such as $s_i = \{1, 2\}$), the cause is only narrowed down to belong to the set s_i . Guess *et al.* [65] first introduced the s_i notation.

When some of the causes are completely or partially masked, the likelihood of the observed data $\{t_i, s_i, i = 1, \dots, n\}$ can be written as

$$\prod_{i=1}^n \left[\sum_{c \in s_i} q(s_i | c, t_i) \left\{ (1-p) h(c, t_i | Q=0, \theta_{0,c})^{I(0 \notin s_i)} \exp \left(- \sum_{r=1}^R H(r, t_i | Q=0, \theta_{0,r}) \right) \right. \right. \\ \left. \left. I(c \neq 1) p h(c, t_i | Q=1, \theta_{1,c})^{I(0 \notin s_i)} \times \exp \left\{ - \sum_{r=2}^R H(r, t_i | Q=1, \theta_{1,r}) \right\} \right\} \right], \quad (6)$$

where we note that the only way $0 \in s_i$ is if $s_i = \{0\}$, that is, the event is right-censored. Further, the second term in the expression corresponds to the latent risk-free group and hence it does not appear when the failure is due to Risk 1.

The $q(s|c, t)$ in Equation (6) are the masking probabilities and are defined as

$$q(s|c, t) = P(\text{cause is masked in subset } s \mid \text{actual cause } C = c, T = t) \quad (7)$$

with $\sum_{s: c \in s} q(s|c, t) = 1$. If the masking probabilities $q(s|c, t)$ are constant in c , that

is, $q(s|c, t) = q(s|c', t)$ for all risks $c, c' \in$ the masking set s , then the $q(s|c, t)$ term can be ignored from the likelihood for likelihood-based inference. Guess *et al.* [65] described this as the “symmetry” condition, which, in some respect, is similar to the “missing at random” assumption in missing data literature. Similar symmetry condition is assumed in Schäbe’s [66] and Goetghebeur and Ryan [57] semiparametric formulations. Lin and Guess [67] and Gittman *et al.* [68] utilize a proportionality assumption to meet the symmetry condition.

For a unit whose cause-of-event is masked, a factor that is of interest concerns the actual cause of failure among those in the masking set s . This can be inferred based on the posterior probabilities or the diagnostic probabilities [59,69–73]

$$\pi(C = c|s, t) = P(\text{actual failure due to cause } c | \text{cause masked in } s, T = t), \quad (8)$$

which can be obtained from $q(s|c, t)$ and the likelihood model via Bayes rule. Flehinger *et al.* [59,60] consider an interesting case when further data are available from second-stage autopsy on a subset of masked units and propose maximum likelihood estimation using nonparametric proportional hazards [59] and completely parametric hazards [60]. Many authors Craiu and Duchesne [74], Craiu and Reiser [70], Mukhopadhyay [75], and Mukhopadhyay and Basu [76] assume that the masking probabilities are constant over time, that is, $q(s|c, t) = q(s|c)$ for all s, j , and t to achieve parsimony. Craiu and Reiser [70], Craiu and Lee [71], and Craiu and Duchesne [72] used EM-based methods, whereas Craiu and Duchesne [74] compared the performances of EM and Bayesian data augmentation methods in this scenario. Sen *et al.* [77] provided a recent review of competing risks analysis for masked data. Mukhopadhyay [75] obtained maximum likelihood estimation (MLE) and bootstrap estimates of these masking probabilities in a general setting and established consistency and asymptotic normality of the MLEs under suitable regularity conditions. Kuo and Yang [78] developed different models

for the masking probabilities and described Bayesian analysis of these models using Gibbs sampling methods. Mukhopadhyay [76] and Sen *et al.* [79] considered Bayesian models where the masking probabilities were assumed to have Dirichlet distribution prior and used Markov chain sampling for the posterior analysis.

EXAMPLE: COMPETING RISKS WITH LIMITED FAILURE AND MASKED CAUSES

We consider the setting of $R = 2$ competing risks, where a proportion p of the units are risk-free from the first risk and simulate $n = 100$ time-to-event observations. The cause-specific hazards for the two risks are taken to be Weibull with $h(r = 1, t|q = 0) = \lambda_1 \beta_1 t^{\beta_1 - 1}$ and $h(r = 2, t|q = 0) = h(r = 2, t|q = 1) = \lambda_2 \beta_2 t^{\beta_2 - 1}$, $t > 0$. Note that unless the β_j ’s are all equal, these hazards are nonproportional. The units fail at the earliest onset of the event due to either Risk 1 or Risk 2. Further, the causes of failure of a random sample of the units are randomly masked; their causes of failure are only recorded as the masking set $s = \{1, 2\}$. In actual simulation, we used $p = 0.25$ and 10 units had masked causes of failure. We considered a constant hazard (exponential) model for Risk 1 with $\lambda_1 = 0.005$ and $\beta_1 = 1$ and an increasing hazard rate for Risk 2 with $\lambda_2 = 0.0056$ and $\beta_2 = 1.5$. The parameters of the two hazards are matched to have similar median survival times.

In the model we fit, we assume that the cause-specific hazards have Weibull(λ_j, β_j) structure. The β_j ’s are allowed to be unequal, thus resulting in nonproportional hazards. The computations are actually performed in log-time scale as in Table 1, in the location-scale form of the Weibull (smallest extreme value) distribution. We assume independent flat normal prior on the location (μ) parameters and gamma priors on the τ parameters.

The results we report here are based on 20,000 iterations of the Markov chain sampler after a burn-in of 10,000 iterations. Table 2 shows the posterior mean of the risk-free fraction. We note that this estimate is

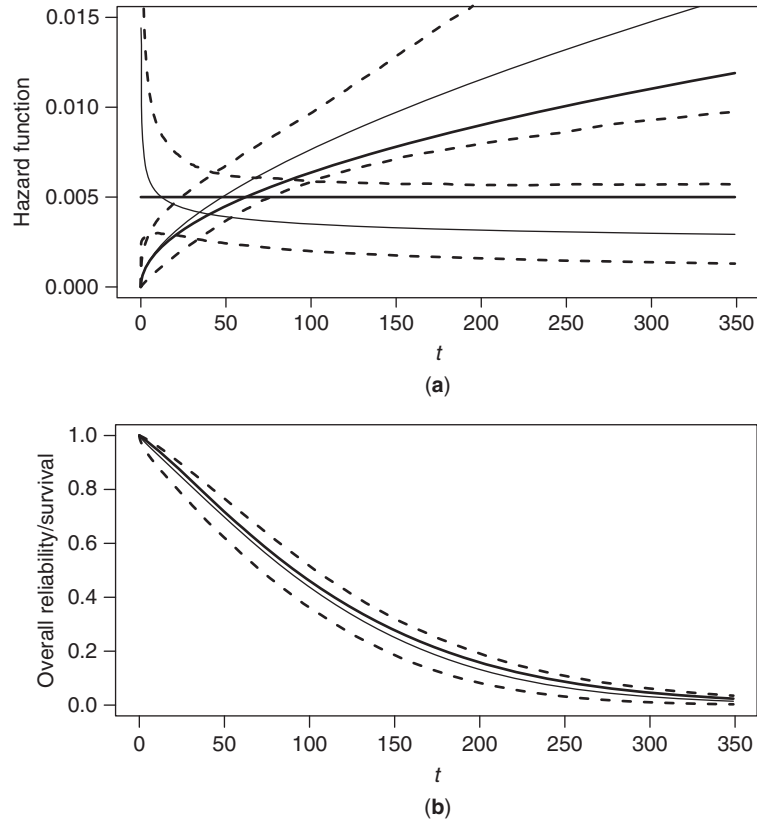


Figure 1. (a) The hazards in the data generating model (dark solid lines), the estimated (posterior mean) cause-specific hazards (solid lines) and 95% pointwise credible bands (dashed line). (b) The overall reliability/survival in the data generating model (dark solid line), the estimated (posterior mean) overall reliability/survival (solid line), and its 95% pointwise credible band.

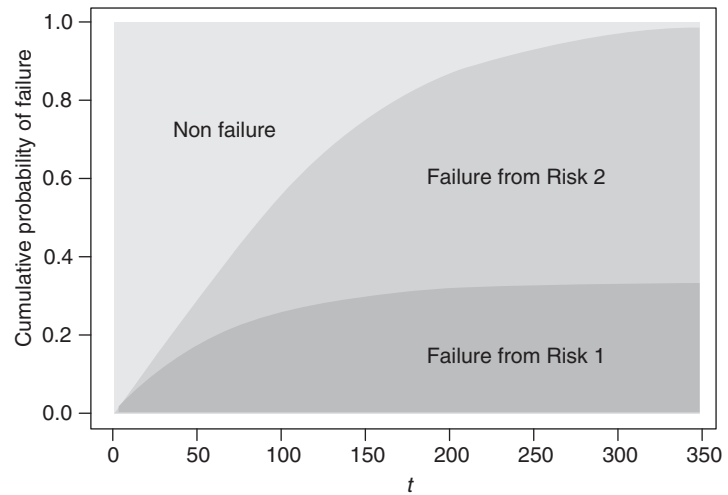


Figure 2. Estimates of cumulative probabilities of failure from the two risks and the probability of survival.

Table 2. Estimate of the Risk-Free (from Risk 1) Fraction

Posterior Mean	95% Credible Interval
0.14	(0.005,0.395)

lower than the $p = 0.25$ of the data generating model. Estimation of the risk-free fraction is known to be a difficult problem [48] and the estimate is known to be sensitive to the modeling assumptions. The 95% credible interval shown in Table 2 does include $p = 0.25$ used in the data generation. Figure 1 plots the posterior means of the cause-specific hazards as well as the associated 95% pointwise credible bands. The constant and increasing shapes of the cause-specific hazards for Risks 1 and 2 in the data generating model are mostly captured by these estimates. The bottom panel shows the estimated overall reliability/survival curve estimate and its 95% credible band; these estimates are, in fact, quite accurate.

In Fig. 2, we show the estimates of the cumulative probability of failure from the two risks as well as the probability of non-failure/survival. The cumulative probability of failure from risk r at time t is defined as $P(C = r, T \leq t)$ whose expression includes evaluation of an integral. We estimate this integral numerically at each time-grid within each iteration of the Markov chain sampler. We note from Fig. 2 and Table 2 that while some units are risk-free from Risk 1, the chance of eventual failure (from either Risk 1 or Risk 2) is 1. This, in fact, is also seen in the overall reliability estimates in Fig. 1. We note from Fig. 2 that Risk 2 becomes the dominant risk because of its increasing hazard rate and because some of the units are risk-free from Risk 1.

In conclusion, we find that even in the presence of competing risks with some masked cause and limited failure from one risk, the overall reliability and cause-specific hazards could be estimated reasonable accurately in this limited study. The cure fraction estimate, however, is not so precise, but this is known to be a difficult estimation problem that is sensitive to modeling assumptions.

REFERENCES

1. Meeker WQ. Limited failure population life tests: application to integrated circuit reliability. *Technometrics* 1987;29:51–65.
2. Chan V, Meeker WQ. A failure-time model for infant-mortality and wearout failure modes. *IEEE Trans Reliab* 1999;48:377–387.
3. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc* 1949;11:15–44.
4. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *J Am Stat Assoc* 1952;47:501–515.
5. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1982;38:257–262.
6. Maller RA, Zhou X. Survival analysis with long-term survivors. New York; Chichester: John Wiley & Sons; 1996.
7. Yakovlev AY, Asselain B, Bardou VJ *et al.* A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. In: Asselain B, Boniface M, Duby C *et al.*, editors. *Biometrie et analyse de donnees spatio-temporelles*. No.12. Rennes: Societe Francaise de Biometrie; 1993. pp. 66–82.
8. Chen M-H, Ibrahim JG, Sinha D. A new Bayesian model for survival data with a surviving fraction. *J Am Stat Assoc* 1999;94:909–919.
9. Tsodikov AD, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: an alternative to two-component mixture models. *J Am Stat Assoc* 2003;98:1063–1078.
10. Anscombe FJ. Estimating a mixed exponential response law. *J Am Stat Assoc* 1961;56:493–502.
11. Danziger L. A life distribution containing immortals unpublished [PhD thesis]. New York University, School of Engineering and Science; 1971.
12. Goldman AI. Survivorship analysis when cure is a possibility: a Monte Carlo study. *Stat Med* 1984;3:153–163.
13. Greenhouse JB, Wolfe RA. A competing risks derivation of a mixture model for the analysis of survival data. *Commun Stat Theory Meth* 1984;13:3133–3154.
14. Lloyd MR, Joe GW. Recidivism comparisons across groups, methods of estimation and tests of significance for recidivism rates and asymptotes. *Eval Q* 1979;3:105–117.

15. Bandes SH, Nadas A. Estimating the Life Distribution of a Population Containing Immortals. Technical Report 22.1345. East Fishkill (NY): IBM, Components Division; 1971.
16. Steinhurst WR. Hypothesis tests for limited failure survival distributions. *Eval Rev* 1981;5:699–711.
17. De Angelis R, Capocaccia R, Hakulinen T, *et al.* Mixture models for cancer survival analysis: application to population-based data with covariates. *Stat Med* 1999;18:441–454.
18. Phillips N, Coldman A, McBride ML. Estimating cancer prevalence using mixture models for cancer survival. *Stat Med* 2002;21:1257–1270.
19. Verdecchia A, DeAngelis G, Capocaccia R. Estimations and projections of cancer prevalence from cancer registry data. *Stat Med* 2002;21:3511–3526.
20. Gamel JW, Weller EA, Wesley MN, *et al.* Parametric cure models of relative and cause-specific survival for grouped survival times. *Comput Methods Programs Biomed* 2000;2:99–110.
21. Yu B, Tiwari R, Cronin KA, *et al.* Cure fraction estimation from the mixture cure models for grouped survival data. *Stat Med* 2004;23(11):1733–1747.
22. Yu B, Tiwari R. Application of EM algorithm to regression model for population-based relative survival data with a cure fraction. *J Data Sci* 2007;5:41–51.
23. Crowder M. Classical competing risks. London: Chapman & Hall/CRC; 2001.
24. Pintilie M. Competing risks : a practical perspective. Hoboken (NJ): John Wiley & Sons; 2006.
25. Deshpande JV, Cooke RM, editors. Competing risks: theory and applications. A special volume of the *J Stat Plann Infer* 2006;136(5):1569–1746.
26. Langseth H, Lindqvist BH. Competing risks for repairable systems: a data study. *J Stat Plann Infer* 2006;136(5):1687–1700.
27. Bunea C, Mazzuchi TA. Competing failure modes in accelerated life testing. *J Stat Plann Infer* 2006;136(5):1608–1620.
28. Sun Y, Tiwari RC. Comparing cumulative incidence functions of a competing-risks model. *IEEE Trans Reliab* 1997;46:247–253.
29. David HA, Moeschberger ML. Volume 39, The theory of competing risks. Griffin's statistical monographs & courses. London: Charles W. Griffin; 1978.
30. Lagakos SW. General right censoring and its impact on the analysis of survival data. *Biometrics* 1979;35:139–156.
31. Prentice RL, Kalbfleisch JD, Peterson AV, *et al.* The analysis of failure time data in the presence of competing risks. *Biometrics* 1978;34:541–554.
32. Nelson W. Hazard plotting for incomplete failure data. *J Qual Technol* 1969;1:27–30.
33. Aalen OO. Nonparametric inference for a family of counting processes. *Ann Stat* 1978;6:701–726.
34. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. New York: John Wiley & Sons; 2002.
35. Lawless JF. Statistical models and methods for lifetime data. 2nd ed. New York: John Wiley & Sons; 2003.
36. Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med* 1993;12:737–751.
37. Cox DR. The analysis of exponentially distributed lifetimes with two types of failures. *J R Stat Soc [Ser B]* 1959;21:411–421.
38. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Natl Acad Sci U S A* 1975;72:20–22.
39. Heckman JJ, Honore BE. The identifiability of the competing risks model. *Biometrika* 1989;77:893–896.
40. Slud E. Nonparametric identifiability of marginal survival distributions in the presence of dependent competing risks and a prognostic covariate. In: Klein JP, Goel PK, editors. *Survival analysis: state of the art*. Boston (MA): Kluwer Academic Publishers; 1992. pp. 355–368.
41. Basu AP, Ghosh JK. Identifiability of the multinormal distribution under competing risks models. *J Multivariate Anal* 1978;8:413–429.
42. Basu AP, Ghosh JK. Identifiability of distributions under competing risks and complementary risks models. *Commun Stat Theory Methods* 1980;9:1515–1525.
43. Basu AP, Klein JP. Some recent results in competing risks theory. In: Crowley J, Johnson RA, editors. *Survival analysis*. California (CA): Hayward; 1982. pp. 216–229.
44. Moeschberger ML, Klein JP. Statistical methods for dependent competing risks. In: *In lifetime data models in reliability and survival analysis*. Jewell NP, Kimber AC, Lee M-LT,

- et al.* editors. Norwell (MA): Kluwer Academic Publishers; 1996. pp. 233–242.
45. Lambert PC, Thompson JR, Weston CL, *et al.* Estimating and modelling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007;8:576–594.
 46. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *J R Stat Soc Ser C Appl Stat* 1987;36:309–317.
 47. Gail M, A review and critique of some models used in competing risk analysis. *Biometrics* 1975;31:209–222.
 48. Basu S, Tiwari R. Breast cancer survival, competing risks and mixture cure model: a Bayesian analysis. *J R Stat Soc Ser A Stat Soc* 2010;173:307–329.
 49. Racine-Poon AH, Hoel DG. Nonparametric estimation of survival function when cause of death is uncertain. *Biometrics* 1984;40:1151–1158.
 50. Dinse GE. Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J Am Stat Assoc* 1986;81:328–336.
 51. Dinse GE. Nonparametric estimates for partially-complete time and type of failure data. *Biometrics* 1982;38:417–431.
 52. Kodell RJ, Chen JJ. Handling cause of death in equivocal cases using the EM algorithm. *Commun Stat Theory Methods* 1987;16:2565–2585.
 53. Lagakos S. Nonparametric estimation of lifetime and disease onset distributions from incomplete observations. *Biometrics* 1982;38:921–932.
 54. Lagakos SW, Louis TA. Use of tumor lethality to interpret tumorigenicity experiments lacking cause-of-death data. *Appl Stat* 1988;37:169–179.
 55. Goetghebeur E, Ryan L. A modified logrank test for competing risks with missing failure type. *Biometrika* 1990;77:207–211.
 56. Dewanji A. A note on the test for competing risks with missing failure type. *Biometrika* 1992;79:855–857.
 57. Goetghebeur E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika* 1995;82:821–833.
 58. Lu K, Tsiatis A. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* 2001;57:1191–1197.
 59. Flehinger BJ, Reiser B, Yashchin E. Survival with competing risks and masked causes of failures. *Biometrika* 1998;85:151–164.
 60. Flehinger BJ, Reiser B, Yashchin E. Parametric modeling for survival with competing risks and masked failure causes. *Lifetime Data Anal* 2002;8:177–203.
 61. Dewanji A, Sengupta D. Estimation of competing risks with general missing pattern in failure types. *Biometrics* 2003;59:1063–1070.
 62. Lu K, Tsiatis A. Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Anal* 2005; 29–40.
 63. Basu S, Basu AP, Mukhopadhyay C. Bayesian analysis of masked system failure data using non-identical Weibull models. *J Stat Plann Infer* 1999;78:255–275.
 64. Basu S, Sen A, Banerjee M. Bayesian analysis of competing risks with partially masked cause-of-failure. *J R Stat Soc Ser C Appl Stat* 2003;52:77–93.
 65. Guess FM, Usher JS, Hodgson TJ. Estimating system and component reliabilities under partial information on cause of failure. *J Stat Plann Infer* 1991;29:75–85.
 66. Schäbe H. Nonparametric estimation of component lifetime based on masked system life test data. *J R Stat Soc B* 1994;56:251–259.
 67. Lin D, Guess FM. System life data analysis with dependent partial knowledge on the exact cause of system failure. *Microelectron Reliab* 1994;34:535–544.
 68. Guttman I, Lin DK, Reiser B, *et al.* Dependent masking and system life data analysis: Bayesian inference for two-component systems. *Lifetime Data Anal* 1995;1:87–100.
 69. Flehinger BJ, Reiser B, Yashchin E. Statistical analysis for masked data. In: Balakrishna N, Rao CR editors. *Advances in reliability*. Vol 20, *Handbook of statistics*. Amsterdam: Elsevier sciences; 2001. pp. 499–522.
 70. Craiu RV, Reiser B. Inference for the dependent competing risks model with masked causes of failure. *Lifetime Data Anal* 2006;12(1):21–33.
 71. Craiu RV, Lee TCM. Model selection for the competing risks model with and without masking. *Technometrics* 2005;47(4):457–467.
 72. Craiu RV, Duchesne T. Inference based on the EM algorithm for the competing risks model with masked causes of failure. *Biometrika* 2004;91(3):543–558.

- 73. Basu S. Inference about the masking probabilities in the competing risks model. *Commun Stat Theory Methods* 2009;38:2677–2690.
- 74. Craiu RV, Duchesne T. Using EM and data augmentation for the competing risks model. In: Gelman A, Meng XL, editors. *Applied Bayesian modeling and causal inference from an incomplete-data perspective*. New York: John Wiley & Sons; 2012.
- 75. Mukhopadhyay C. Maximum likelihood analysis of masked series system lifetime data. *J Stat Plann Infer* 2006;136:803–838.
- 76. Mukhopadhyay C, Basu S. Bayesian analysis of masked series system lifetime data. *Commun Stat Theory Methods* 2007;36:329–348.
- 77. Sen A, Basu S, Bannerjee M. Statistical analysis of life-data with masked cause-of-failure. In: Basu AP, Balakrishnan N editors. *Advances in reliability*. Vol 20, *Handbook of statistics*. Amsterdam: Elsevier Sciences; 2001.
- 78. Kuo L, Yang TY. Bayesian reliability modeling for masked system lifetime data. *Stat Probab Lett* 2000;47:229–241.
- 79. Sen A, Banerjee M, Li Y, *et al.* A Bayesian approach to competing risks analysis with masked cause of death. *Stat Med* 2010;29:1681–1695.

COMPLEMENTARITY PROBLEMS

IGOR V. KONNOV
Department of System Analysis
and Information Technologies,
Kazan University, Kazan,
Russia

BASIC DEFINITIONS

A set K in the n -dimensional Euclidean space $\mathbb{R}^n$ is said to be a cone if $\lambda K \subseteq K$ for any $\lambda \geq 0$. Let K be a convex cone in $\mathbb{R}^n$ and let $F: K \rightarrow \mathbb{R}^n$ be a mapping. Then, we can define the *complementarity problem* (CP for short)

$$\begin{aligned} x^* \in K, F(x^*) \in K^*, \langle x^*, F(x^*) \rangle \\ = \sum_{i=1}^n x_i^* F_i(x^*) = 0, \end{aligned} \quad (1)$$

where $K^* = \{z \in \mathbb{R}^n \mid \langle x, z \rangle \geq 0 \ \forall x \in K\}$ is the conjugate (dual) cone to K . There exist various subclasses and extensions of CP (1), but the following ones are those most investigated. The *standard complementarity problem* corresponds to the case where K is the nonnegative orthant $\mathbb{R}_+^n = \{x \in \mathbb{R}^n \mid x_i \geq 0, \ i = 1, \dots, n\}$, that is, the problem is to find

$$x^* \geq 0, F(x^*) \geq 0, \langle x^*, F(x^*) \rangle = 0 \quad (2)$$

or, equivalently,

$$\begin{aligned} x_i^* \geq 0, F_i(x^*) \geq 0, x_i^* F_i(x^*) = 0 \\ \text{for } i = 1, \dots, n. \end{aligned}$$

Here and in the subsequent discussions, all the inequalities for vectors are interpreted component-wise and 0 denotes the zero vector. Also, the *linear complementarity problem* (LCP) corresponds to the affine case where $F(x) = Ax - b$, A is an $n \times n$ matrix, and b is a fixed element in $\mathbb{R}^n$. Investigation of

CP (LCP) as a separate problem was initiated by the works of Dorn [1] and Cottle [2]; see also Harker and Pang [3], Cottle *et al.* [4], Isac [5], and Facchinei and Pang [6] for more details. CPs have a great number of direct applications in economics, mathematical physics, telecommunications, transportation, and other fields [3–5, 7–9]. For this reason, the theory and methods for CPs are investigated very extensively (see also the surveys by Billups and Murty [10] and Ferris and Kanzow [11] and edited books by Ferris and Pang [12], Fukushima and Qi [13], and Ferris *et al.* [14]). In this article, we intend to describe some of these results obtained by taking CP (2) as a basis. We also discuss their possible extensions.

THEORETICAL BACKGROUND

In this section, we consider CP (2) under the standing assumption that $F: \mathbb{R}_+^n \rightarrow \mathbb{R}^n$ is a continuous mapping. One of the general existence results from the theory of variational inequalities can be specialized for CP (2) as follows.

Theorem 1 [15,16]. *Suppose that there exists a nonempty bounded subset Y of $\mathbb{R}_+^n$ such that for every $x \in \mathbb{R}_+^n \setminus Y$ there is $y \in Y$ with $\langle F(x), x - y \rangle \geq 0$. Then CP (2) has a solution.*

Indeed, the set K in CPs is usually unbounded. Hence we cannot apply the Brouwer fixed point theorem directly [[17], Chapter I, Theorem 4.3] and have to utilize certain coercivity conditions such as that in Theorem 1. However, we can establish existence results by using the concept of the so-called exceptional family of elements.

Definition 1 [18]. A sequence $\{x^k\} \subset \mathbb{R}_+^n$ with $\|x^k\| = k$ is said to be an *exceptional family of elements* for a mapping F if the

following conditions are satisfied:

$$F_i(x^k) \begin{cases} = -\lambda_k x_i^k, & \text{if } x_i^k > 0, \\ \geq 0, & \text{if } x_i^k = 0, \end{cases} \quad \text{where } \lambda_k > 0.$$

Theorem 2 [18]. *If there is no exceptional family of elements for a mapping F , then CP (2) has a solution.*

It is easy to see that the coercivity condition in Theorem 1 implies the absence of an exceptional family of elements for the mapping F . Similarly, we can utilize some other sufficient coercivity conditions that are more suitable for their verification. Further development in this direction is found in Isac *et al.* [19].

We now recall some norm monotonicity concepts for mappings.

Definition 2. Let X be a convex set in $\mathbb{R}^n$. A mapping $Q : X \rightarrow \mathbb{R}^n$ is said to be

- (a) *monotone*, if for each pair of points $x, y \in X$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle \geq 0;$$

- (b) *strictly monotone*, if for each pair of points $x, y \in X, x \neq y$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle > 0;$$

- (c) *strongly monotone* with constant $\tau \geq 0$ if, for each pair of points $x, y \in X$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle \geq \tau \|x - y\|^2.$$

Clearly, we have $(c) \implies (b) \implies (a)$. In the affine case $Q(x) = Ax - b$ is (strictly, strongly) monotone if and only if A is positive (definite) semidefinite matrix. In the differentiable case, the above monotonicity concepts are associated with the Jacobian properties [6, Proposition 2.3.2]. By the streamlined specialization of the known existence and uniqueness results from variational inequalities [9, Chapter 11], we can provide similar results for CP (2).

Proposition 1

- (i) *If F is strictly monotone, then CP (2) has at most one solution.*
- (ii) *If F is strongly monotone, then CP (2) has a unique solution.*

We now show that these results can be strengthened essentially in several directions.

Theorem 3

[20].

- (i) *If F is monotone and there exists a point $\bar{x} \geq 0$ such that $F(\bar{x}) > 0$, then CP (2) has a solution.*
- (ii) *If F is strictly monotone and there exists a point $\bar{x} \geq 0$ such that $F(\bar{x}) \geq 0$, then CP (2) has a unique solution.*

Next, we introduce several weaker concepts of order monotone mappings [5, 21–23].

Definition 3. Let X be a box constrained set in $\mathbb{R}_+^n$. A mapping $Q : X \rightarrow \mathbb{R}^n$ is said to be

- (a) a P_0 -mapping, if for each pair of points $x', x'' \in X, x' \neq x''$, there exists an index i such that $x'_i \neq x''_i$ and

$$(x'_i - x''_i) [Q_i(x') - Q_i(x'')] \geq 0;$$

- (b) a P -mapping, if for each pair of points $x', x'' \in X, x' \neq x''$, it holds that

$$\max_{1 \leq i \leq n} (x'_i - x''_i) [Q_i(x') - Q_i(x'')] > 0;$$

- (c) a *strict P -mapping*, if there exists $\varepsilon > 0$ such that $Q - \varepsilon I$ is a P -mapping.

Clearly, we have $(c) \implies (b) \implies (a)$. Also, it is obvious that each (strictly, strongly) monotone mapping is a (P -, strict P -) P_0 -mapping.

Definition 4. An $n \times n$ matrix A is said to be a P_0 - (P -) matrix, if it has nonnegative (positive) principal minors.

Note that in the affine case $Q(x) = Ax - b$ is a P_0 - (P -) mapping if and only if A is a P_0 - (P -) matrix.

matrix. However, $A = \begin{pmatrix} 2 & 5 \\ 0 & 2 \end{pmatrix}$ is clearly a P -matrix, but it is not even positive semidefinite. Nevertheless, these concepts coincide in the symmetric case, that is, each symmetric P_0 - (P -) matrix is a positive semidefinite (definite) matrix. In the differentiable case, we can utilize the corresponding properties of the Jacobian [4,17,21].

Theorem 4 [20,23,24].

- (i) If F is a P -mapping, then CP (2) has at most one solution.
- (ii) If F is a strict P -mapping, then CP (2) has a unique solution.

So, Theorem 4 strengthens Proposition 1.

Now, we turn to the results exploiting the Z -property (off-diagonal antitonicity) of the mapping F , which appears natural in many applications [4,5,25–29].

Definition 5. Let X be a box constrained set in $\mathbb{R}_+^n$. A mapping $Q : X \rightarrow \mathbb{R}^n$ is said to be

- (a) *isotone (antitone)*, if for each pair of points $x', x'' \in X$ such that $x' \geq x''$, it holds that $F(x') \geq F(x'')$ ($F(x') \leq F(x'')$);
- (b) a *Z -mapping* (or an *off-diagonal antitone mapping*), if, for each pair of points $x', x'' \in X$ such that $x' \geq x''$, it holds that $Q_k(x') \leq Q_k(x'')$ for each k such that $x'_k = x''_k$.

Again, in the affine case $Q(x) = Ax - b$ is an isotone (antitone, Z -) mapping if and only if A has only nonnegative entries (nonpositive entries, nonpositive off-diagonal entries, i.e., it is a Z -matrix). In the differentiable case, we can also utilize the corresponding properties of the Jacobian [4,21,22]. Let us define the auxiliary set

$$W = \{x \in \mathbb{R}^n \mid x \geq \mathbf{0}, F(x) \geq \mathbf{0}\}.$$

For each pair of points $x, y \in \mathbb{R}^n$, we can define their component-wise minimal point (meet) $z = \min\{x, y\}$ as follows:

$$z_i = \min\{x_i, y_i\} \quad \text{for } i = 1, \dots, n.$$

It appears $\min\{x, y\} \in W$ if $x, y \in W$, that is, W is a meet semisublattice if F is Z . Next, we can define the minimal element

$$\min W = \{z \in W \mid z \leq x \quad \forall x \in W\}.$$

One can now derive the special existence result; for more details refer to Yershov [28], Tamir [30], Cottle *et al.* [4], and Isac [5].

Theorem 5. If W is nonempty, $F : \mathbb{R}_+^n \rightarrow \mathbb{R}^n$ is a Z -mapping, then CP (2) has a solution, which coincides with $\min W$.

Note that CP (2) here may admit many solutions. However, additional assumptions, say from Theorem 4 provide uniqueness. Certain results above can be adjusted for CP (1) and further extended to infinite-dimensional spaces; for more details refer to Cottle *et al.* [3], Isac [5], Isac *et al.* [19], and Yao and Chadli [31].

RELATED PROBLEMS

It is known [32] that CP (1) is equivalent to the following *variational inequality* (VI): find $x^* \in K$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0 \quad \forall x \in K. \quad (3)$$

Thus, CPs can in principle be viewed as special classes of VIs where the feasible set is a cone (or a cone segment). These features enable one to essentially enhance many results in the theory and solution methods for CPs in comparison with those for general VIs; see in particular, the section titled “Theoretical Background”. The above property enables one to establish relations of CPs with other general problems.

For instance, let T be a continuous mapping from K into itself. Then the *fixed point problem* which is to find a point $x^* \in K$ such that $x^* = T(x^*)$ coincides with VI (3), where

$$F(x) = x - T(x), \quad (4)$$

and hence with CP (1), (4) [33], [Section 3.1]. Moreover, if T is isotone, then F in Equation(4) is clearly a Z -mapping.

Let us now define the *optimization problem* of finding a point $x^* \in K$ such that

$$\varphi(x^*) \leq \varphi(x) \quad \forall x \in K,$$

or, briefly,

$$\min \rightarrow \{\varphi(x) \mid x \in K\} \quad (5)$$

and suppose that $\varphi : X \rightarrow \mathbb{R}$ is a differentiable function. Then Equation (5) implies VI (3) and CP (1) where $F(x) = \nabla \varphi(x)$, and the reverse implication holds if $\varphi : X \rightarrow \mathbb{R}$ is convex, that is, F is monotone. In this potential case, the Jacobian $\nabla F(x) = \varphi''(x)$ must be symmetric.

Let us consider the *constrained optimization problem*

$$\min \rightarrow \{\varphi_0(u) \mid u \in U\}, \quad (6)$$

where

$$U = \{u \in \mathbb{R}_+^l \mid \varphi_i(u) \leq 0 \quad i = 1, \dots, m\} \quad (7)$$

and $\varphi_i : \mathbb{R}^l \rightarrow \mathbb{R}$, $i = 0, \dots, m$ are convex differentiable functions. Set $n = l + m$ and define the mapping $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ as follows:

$$F(x) = F(u, v) = \begin{pmatrix} \nabla \varphi_0(u) + \sum_{i=1}^m v_i \nabla \varphi_i(u) \\ -\varphi(u) \end{pmatrix}, \quad (8)$$

where $\varphi(u) = (\varphi_1(u), \dots, \varphi_m(u))^T$. The relationships between problem (6)–(7) and CP (2), (8) are known as the *Karush–Kuhn–Tucker optimality conditions*. We need an additional constraint qualification. **(C)** *Either all the functions φ_i , $i = 1, \dots, m$ are affine, or there exists a point $\bar{u} \in \mathbb{R}_+^l$ such that $\varphi_i(\bar{u}) < 0$ for all $i = 1, \dots, m$.*

Proposition 2 [17, Chapter 1, Theorems 3.16 and 3.17.

- (i) *If $x^* = (u^*, v^*)$ solves CP (2), (8) then x^* is a solution to problem (6)–(7).*

- (ii) *If x^* is a solution to problem (6)–(7) and condition (C) holds, then there exists a point $v^* \in \mathbb{R}_+^m$ such that (u^*, v^*) solves CP (2), (8).*

Note that F in Equation (8) is monotone, but its Jacobian $\nabla F(x)$ cannot be symmetric. Also, CP (2), (8) is in fact equivalent to the saddle point problem of the Lagrange function associated with problem (6)–(7):

$$L(u, v) = \varphi_0(u) + \sum_{i=1}^m v_i \varphi_i(u).$$

This is the case for more general convex–concave bi-functions; hence CPs can be used for solving zero-sum two-person games and even noncooperative games; for more details refer to Cottle *et al.* [3], Isac [5], Ferris and Pang [8], Isac *et al.* [19], Facchinei and Pang [6], and Konnov [9].

ALGORITHMS

At first we consider the case of LCP when $F(x) = Ax - b$ in Equation (2). If the matrix A is arbitrary, even this problem can be too hard for a solution. However, if A is positive (semi)definite or possesses P -type properties, there exist a number of the so-called *pivotal algorithms*, which yield a solution in a finite number of iterations [4,34]. These algorithms, which include the most-known Lemke method [35], create a finite sequence of special systems of linear equations with pivotal exchanges of basic and nonbasic variables and maintenance of certain additional conditions. It may also be observed that the above LCP becomes equivalent to the convex quadratic programming problem

$$\min_{x \in \mathbb{R}_+^n} \rightarrow 0.5 \langle Ax, x \rangle - \langle b, x \rangle,$$

if A is symmetric and positive semidefinite. Let us now consider the general case of CP (2) where $F : \mathbb{R}_+^n \rightarrow \mathbb{R}^n$ is a continuous mapping. Observe that pivotal algorithms can be extended to this problem. They involve sequential solution of auxiliary nonlinear equations and require certain

order monotonicity properties [36,37]. However, the *linearization algorithms* based on the sequential solution of LCP problems constitute one of the most popular and investigated classes. At the k th iteration of such an algorithm, given a point $x^k \in \mathbb{R}_+^n$, the next iterate x^{k+1} is defined as a solution of the LCP:

$$x^{k+1} \geq 0, F^k(x^{k+1}) \geq 0, \langle x^{k+1}, F^k(x^{k+1}) \rangle = 0, \quad (9)$$

where $F^k(x) = F(x^k) + C_k(x - x^k)$, C_k is an $n \times n$ matrix. Different choice of C_k leads to various iterative methods. For instance, setting $C_k = \nabla F(x^k)$ ($C_k = 0.5(\nabla F(x^k) + \nabla F(x^k)^\top)$, respectively) gives the Newton (respectively, symmetric Newton) method, while setting $C_k = C$ with a symmetric and positive definite matrix C gives the projection method. Besides, one can take C_k as a part or an approximation of $\nabla F(x^k)$, thus obtaining in particular linearized Jacobi, successive overrelaxation, Levenberg–Marquardt, quasi-Newton methods, and so on. Concerning the convergence of the linearization algorithms we note that approximation of $\nabla F(x^k)$ usually yields rather rapid convergence which is however local, the Newton method attaining the highest quadratic convergence rate.

Theorem 6 [38], Corollary 2.6. *Suppose that x^* solves CP (2), F is continuously differentiable, and $\nabla F(x^*)$ is positive definite. Then, there exists a neighborhood U of x^* such that the sequence $\{x^k\}$ generated by the Newton method is well defined and converges to x^* if $x^0 \in U$. Moreover, if ∇F is Lipschitz continuous in a neighborhood U' of x^* , then $\{x^k\}$ converges quadratically to x^* , that is, there is a constant τ such that $\|x^{k+1} - x^*\| \leq \tau \|x^k - x^*\|^2$.*

Observe that the above positive definiteness of $\nabla F(x^*)$ can be replaced by a weaker regularity condition. The known drawback of the Newton type methods is their very restrictive assumptions for providing global convergence. The simplest projection

method corresponds to Equation (9) with $F^k(x) = F(x^k) + \lambda_k^{-1}I(x - x^k)$ where I is an $n \times n$ unit matrix and $\lambda_k > 0$ and is equivalent to the iterate $x^{k+1} = [x^k - \lambda_k F(x^k)]_+$, where $[a]_+$ denotes the projection of a onto $\mathbb{R}_+^n$. Its global convergence also requires strengthened monotonicity properties and knowledge of the corresponding constants. By utilizing suitable extrapolation procedures, one can provide convergence without such restrictive assumptions [39,40].

Another approach consists in replacing the initial CP (2) by an optimization problem via introduction of some artificial *merit function*, which admits algorithms with line search procedures. For instance, the regularized merit function suggested by Fukushima [41] is defined as follows:

$$\begin{aligned} \varphi_\lambda(x) &= \max_{y \in \mathbb{R}_+^n} \left\{ \langle F(x), x - y \rangle - (2\lambda)^{-1} \|x - y\|^2 \right\} \\ &= \langle F(x), x \rangle + (2\lambda)^{-1} \left\{ \|x - \lambda F(x)\|_+^2 - \|x\|^2 \right\}, \quad \lambda > 0. \end{aligned}$$

Let CP (2) be solvable. Then φ_λ is nonnegative on $\mathbb{R}_+^n$ and attains its minimal (zero) value, that is, CP (2) becomes equivalent to

$$\min_{x \in \mathbb{R}_+^n} \varphi_\lambda(x). \quad (10)$$

Moreover, if $\nabla F(x)$ is a P -matrix for all $x \geq 0$, then the CP

$$x^* \geq 0, \quad \nabla \varphi_\lambda(x^*) \geq 0, \quad \langle \nabla \varphi_\lambda(x^*), x^* \rangle = 0$$

becomes equivalent to both CP (2) and the optimization problem (10). Therefore, one can apply any gradient descent algorithm to Equation (10) in order to find a solution for CP (2). In addition, if $\nabla F(x)$ is a positive definite matrix for all $x \geq 0$, then one can apply the descent algorithms with line search with respect to φ_λ , where solutions of the auxiliary problem (9) are used for computing the descent direction at x^k , including Newton and projection iterates [6,9,42].

There exist a number of merit functions, which can possess some additional nice properties. It is possible to define a merit function

ψ on the whole space, that is, ψ is nonnegative on $\mathbb{R}^n$ and attains its minimal (zero) value; hence CP (2) becomes equivalent to the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} \psi(x). \quad (11)$$

For instance, the so-called implicit Lagrangian [43]

$$\psi_{\text{MS}}(x) = \langle F(x), x \rangle + (2\lambda)^{-1} \left(\| [x - \lambda F(x)]_+ \|^2 - \|x\|^2 + \| [F(x) - \lambda x]_+ \|^2 - \|F(x)\|^2 \right),$$

where $\lambda > 1$ meets these conditions. Observe that $\psi_{\text{MS}}(x) = \varphi_\lambda(x) - \varphi_{1/\lambda}(x)$, that is, it coincides with the so-called *D-gap function* [6,42,44]. Most of such functions of ψ can be created within the *equation reduction approach*. In fact, let a function $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfy the conditions

$$\phi(\alpha, \beta) = 0 \implies \alpha \geq 0, \beta \geq 0, \alpha\beta = 0.$$

Then, ϕ is called an *NCP-function*. If one defines the mapping $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $\Psi_i(x) = \phi(x_i, F_i(x))$ for $i = 1, \dots, n$, then CP (2) becomes equivalent to the nonlinear equation

$$\Psi(x^*) = \mathbf{0}, \quad (12)$$

and, also to problem (11) where

$$\psi(x) = 0.5 \|\Psi(x)\|^2.$$

Clearly, $\psi_{\text{MS}}(x)$ corresponds here to

$$\phi_{\text{MS}}(\alpha, \beta) = \alpha\beta + (2\lambda)^{-1} \left\{ [\alpha - \lambda\beta]_+^2 - \alpha^2 + [\beta - \lambda\alpha]_+^2 - \beta^2 \right\}$$

with $[\gamma]_+ = \max\{\gamma, 0\}$. Besides, one can take

- (i) the min function [3] $\phi_{\min}(\alpha, \beta) = \min\{\alpha, \beta\}$,
- (ii) the Fischer–Burmeister function [45] $\phi_{\text{FB}}(\alpha, \beta) = \sqrt{\alpha^2 + \beta^2} - \alpha - \beta$, and many others [6].

Despite the possible nonsmoothness of the mapping Ψ , Newton type methods can also be used to find a solution to Equation (12) (or Equation (2)). Namely, the next iterate x^{k+1} is defined as a solution of the equation

$$F(x^k) + H_k(x - x^k) = \mathbf{0},$$

where H_k is an analog (or approximation) of the Jacobian, for instance, an element of the Clarke generalized Jacobian $\partial^\uparrow F(x^k)$ [46]. Similarly, a line search procedure with respect to ψ is used to provide global convergence for these methods. Nevertheless, computation of the matrix H_k can meet certain difficulties in the nonsmooth case, whereas smooth mappings Ψ usually have singular Jacobian matrices at degenerate solutions (i.e., when there exists an index k such that both $x_k^* = 0$ and $F_k(x^*) = 0$ at a solution x^*), which prevents the direct application of the Newton method and leads to slow convergence. In order to overcome these difficulties, one can utilize the *continuation approach*, which yields in particular, smooth approximate mappings. Following this approach, one replaces CP (2) by a parameterized problem of the form

$$\begin{aligned} x_i(\mu) &\geq 0, F_i(x(\mu)) \geq 0, x_i(\mu)F_i(x(\mu)) = \mu \\ \text{for } i &= 1, \dots, n, \end{aligned} \quad (13)$$

where $\mu > 0$. Under certain regularity assumptions, it is possible to show that the trajectory $x(\mu)$ will approximate a solution x^* as $\mu \rightarrow 0$. At the same time, this idea can be implemented via replacing ϕ (or Ψ) with its smooth approximation. For instance, utilizing the smoothing Fischer–Burmeister function $\phi_{\text{FB}}^\mu(\alpha, \beta) = \sqrt{\alpha^2 + \beta^2 + \mu} - \alpha - \beta$, one can define the approximate mapping Ψ^μ by $\Psi_i^\mu(x) = \phi_{\text{FB}}^\mu(x_i, F_i(x))$ for $i = 1, \dots, n$, which is smooth if F is so, and then replace Equation (12) by

$$\Psi^\mu(z(\mu)) = \mathbf{0}.$$

Note that $z(\mu)$ satisfies Equation (13), and moreover, $z_i(\mu) > 0$ and $F_i(z(\mu)) > 0$ for $i = 1, \dots, n$. Therefore, it is also treated as an *interior point method* where the solutions $z(\mu)$ form the so-called central path. In order to

solve CP (2), one can thus solve Equation (12) within a given accuracy ε by the suitable Newton method

$$\Psi^\mu(x^k) + \nabla \Psi^\mu(x^k)(x^{k+1} - x^k) = \mathbf{0},$$

(with inserting line search with respect to $\psi^\mu(x) = 0.5\|\Psi^\mu(x)\|^2$ if necessary) and drive μ and ε to 0. The single-level scheme

$$\Psi^{\mu_k}(x^k) + \nabla \Psi^{\mu_k}(x^k)(x^{k+1} - x^k) = \mathbf{0},$$

with simultaneous changes in x^k and $\mu_k \rightarrow 0$ may even appear more efficient here. Within this approach there exist a great number of algorithms, basic and smoothing functions [6,12,13]. Now we turn to the methods that are based on exploiting the Z -property of the mapping F . As discussed above, features of Z -mappings enable us to suggest efficient tools for finding a solution for CP (2). In fact, the nonlinear *coordinate relaxation* (e.g., Jacobi and Gauss–Seidel) algorithms can be then applied to find solutions [4,5,20,21,30,38]. Now, we present a splitting algorithm that utilizes a Jacobi type iteration. It follows Konnov [47] and simplifies similar algorithms of Pang and Chan [38] and Konnov [48].

Let us consider CP (2) where

$$F(x) = G(x) + H(x),$$

$G : \mathbb{R}_+^n \rightarrow \mathbb{R}^n$ is a continuous Z -mapping, and $H : \mathbb{R}_+^n \rightarrow \mathbb{R}^n$ is a continuous antitone mapping.

Algorithm (JS). Choose a point $x^0 \in W$. At the k th iteration, $k = 0, 1, \dots$, we have a point $x^k \in W$. For each separate index $i = 1, \dots, n$, we determine a number x_i^{k+1} as follows: if $G_i(x_{-i}^k, 0) + H_i(x^k) \geq 0$, then set $x_i^{k+1} = 0$; otherwise find $x_i^{k+1} \in [0, x_i^k]$ such that $G_i(x_{-i}^k, x_i^{k+1}) + H_i(x^k) = 0$.

Theorem 7. Suppose that W is nonempty. Then Algorithm (JS) is well defined and generates a sequence $\{x^k\}$ converging to a solution x^* of CP (2).

The Jacobi basic iteration can be replaced by the Gauss–Seidel iteration with similar convergence properties.

Most existence and uniqueness results and convergence properties for solution methods are guaranteed under strengthened (order) monotonicity properties. If they are too restrictive, certain *regularization* techniques can be applied. For instance, the classical Tikhonov–Browder regularization leads to the following auxiliary problem:

$$x^\varepsilon \geq \mathbf{0}, F(x^\varepsilon) + \varepsilon x^\varepsilon \geq \mathbf{0}, \langle F(x^\varepsilon) + \varepsilon x^\varepsilon, x^\varepsilon \rangle = 0, \quad (14)$$

where $\varepsilon > 0$. In the case when F is monotone and continuous, it is known [49] that each problem (14) has a unique solution and that the sequence of solutions $\{x^\varepsilon\}$ converges to the minimal norm solution x_n^* of CP (2) as $\varepsilon \rightarrow 0$. Note that the cost mapping $F^\varepsilon = F + \varepsilon I$ in Equation (14) is then strongly monotone. The problem is to maintain at least some part of these nice convergence properties in the case when F is order monotone, say, satisfies the P_0 condition. Then, F^ε is strict P and each problem (14) has a unique solution [4,24,50]. However, even weakened convergence properties of regularization schemes will require certain additional conditions. For instance, this is the case if the solution set of CP (2) is nonempty and bounded [6,24]. Another approach to justifying regularization methods for nonmonotone CPs consists in utilizing coercivity conditions [51,52]. For instance, if the assumptions of Theorem 1 are fulfilled, then each CP (14) has a solution, each sequence $\{x^\varepsilon\}$ has limit points, and if $\{\varepsilon_k\} \searrow 0$ all these limit points are solutions of CP (2) [52,53]. The proximal point method [54,55] also consists in replacing the initial problem with a sequence of auxiliary ones, but its regularization parameter may be in principle fixed and the perturbed mapping depends on the previous iteration point. At the same time, its convergence requires certain (generalized) monotonicity properties [56,57].

EXTENSIONS

Now we turn to some extensions. For instance, replacing the nonnegativity of

variables in Equation (2) with the two-sided constraints or allowing for possible unrestricted variables, we obtain the *mixed complementarity problem* (MCP): Find $x^* \in D$ such that

$$F_i(x^*) \begin{cases} \geq 0 & \text{if } x_i^* = a_i, \\ = 0 & \text{if } x_i^* \in (a_i, b_i), \\ \leq 0 & \text{if } x_i^* = b_i, \end{cases} \quad \text{for } i = 1, \dots, n, \quad (15)$$

where

$$D = \{x \in \mathbb{R}^n \mid -\infty \leq a_i \leq x_i \leq b_i \leq +\infty \\ i = 1, \dots, n\} \quad (16)$$

is a box-constrained (rectangle) set or a cone segment with respect to the ordering defined by $\mathbb{R}_+^n$. Similarly, we can utilize cone segment restrictions in CP (1) instead of the cone K and obtain its extension. Again, MCP (15)–(16) is also equivalent to VI: Find $x^* \in D$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0 \quad \forall x \in D, \quad (17)$$

compare Equation (3). We note that the previous results obtained for CP (2) can be extended to this box-constrained VI (17) (or to MCP (15)–(16)) after proper modifications [3,8,6,9,29,53,57].

Further extensions may consist in replacing the single-valued cost mapping with the multivalued mapping. So, let K be a convex cone in $\mathbb{R}^n$ and let $\mathbf{F} : K \rightarrow \Pi(\mathbb{R}^n)$ be a multivalued mapping, where $\Pi(S)$ denotes the family of all subsets of a set S . Then, we can define the *generalized complementarity problem* (GCP):

$$x^* \in K, \exists f^* \in \mathbf{F}(x^*), f^* \in K^*, \langle f^*, x^* \rangle = 0, \quad (18)$$

compare Equation (1). In case $K = \mathbb{R}_+^n$, Equation (18) reduces to the standard GCP:

$$x^* \geq 0, \exists f^* \in \mathbf{F}(x^*), f^* \geq 0, \langle f^*, x^* \rangle = 0, \quad (19)$$

compare (2). In turn, GCP (18) is equivalent to the *generalized variational inequality*

problem (GVI), which is to find an element $x^* \in K$ such that

$$\exists f^* \in \mathbf{F}(x^*), \langle f^*, y - x^* \rangle \geq 0 \quad \forall y \in K, \quad (20)$$

so that GCP (19) coincides with GVI (20) where $K = \mathbb{R}_+^n$. The multivalued extensions of the box-constrained VI (17) and MCP (15)–(16) are defined similarly. The necessity to investigate and solve such problems stems from the fact that many applied models with complementarity conditions involve just multivalued mappings [5,8,17,26,29,40,50,52].

The basic existence and uniqueness results from the section titled “Theoretical Background” contained in Theorems 1 and 2 are transformed directly for GCP (19). This is the case for the norm monotonicity and P -type assumptions after their proper extensions [5,23,40,50,52]. However, the streamlined extension of the Z -mapping concept meets certain difficulties and analogs of Theorem 5 are established for some subclasses of multivalued mappings [29,58]. In view of the equivalence results, iterative solution methods devised for GVI (20) are applicable for GCPs (18) and (19) [40,59]. We also notice that the results described in the section titled “Algorithms” for the regularization and proximal point methods remain true in the multivalued case [52,57]. Again, after proper modifications the splitting and coordinate relaxation (e.g., Jacobi and Gauss–Seidel) algorithms can be applied to GCP (19), but their convergence is established for some subclasses of multivalued Z -type mappings [29,47,48,58,60,61]. In addition, we observe that the D -gap function approach enables one to convert GCP (19) whose cost mapping is the sum of a monotone (or P -type) single-valued differentiable mapping and a multivalued diagonal monotone mapping into an unconstrained differentiable optimization problem and to develop efficient descent methods [62,23].

REFERENCES

1. Dorn WS. Self-dual quadratic programs. SIAM J Appl Math 1961;9:51–54.

2. Cottle RW. Nonlinear programs with positively bounded Jacobians. *J SIAM* 1966; 14:147–158.
3. Harker PT, Pang J-S. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Math Program* 1990;48:161–220.
4. Cottle RW, Pang J-S, Stone RE. The linear complementarity problem. Boston (MA): Academic Press; 1992.
5. Isac G. Complementarity problems. Berlin: Springer; 1992.
6. Facchinei F, Pang J-S. Volumes I and II, Finite-dimensional variational inequalities and complementarity problems. Berlin: Springer; 2003.
7. Berschanskii YM, Meerov MV. Theory and solution methods for complementarity problems. *Automat Rem Control* 1983;44:5–31.
8. Ferris MC, Pang J-S. Engineering and economic applications of complementarity problems. *SIAM Rev* 1997;39:669–713.
9. Konnov IV. Equilibrium models and variational inequalities. Amsterdam: Elsevier; 2007.
10. Billups SC, Murty KG. Complementarity problems. *J Comput Appl Math* 2000;124: 303–318.
11. Ferris MC, Kanzow C. Complementarity and related problems: a survey. In: Pardalos PM, Resende MGC, editors. *Handbook of applied optimization*. New York: Oxford University Press; 2002. pp. 514–530.
12. Ferris MC, Pang J-S, editors. *Complementarity and variational problems: state of the art*. Philadelphia (PA): SIAM; 1997.
13. Fukushima M, Qi L, editors. *Reformulation - nonsmooth, piecewise smooth, semismooth and smoothing methods*. Dordrecht: Kluwer Academic Publishers; 1998.
14. Ferris MC, Mangasarian OL, Pang J-S, editors. *Complementarity: applications, algorithms and extensions*. Dordrecht: Kluwer Academic Publishers; 2001.
15. Eaves BC. On the basic theorem of complementarity. *Math Program* 1971;1:68–75.
16. Karamardian S. The complementarity problem. *Math Program* 1972;2:107–129.
17. Nikaido H. *Convex structures and economic theory*. New York: Academic Press; 1968.
18. Smith TE. A solution condition for complementarity problems: with an application to spatial price equilibrium. *Appl Math Comput* 1984;15:61–69.
19. Isac G, Bulavski VA, Kalashnikov VV. *Complementarity, equilibrium, efficiency and economics*. Dordrecht: Kluwer Academic Publishers; 2002.
20. More JJ. Classes of functions and feasibility conditions in nonlinear complementarity problems. *Math Program* 1974;6:327–338.
21. Ortega JM, Rheinboldt WC. *Iterative solution of nonlinear equations in several variables*. New York: Academic Press; 1970.
22. More JJ, Rheinboldt WC. *P- and S-functions and related classes of n -dimensional nonlinear mappings*. *Linear Algebra Appl* 1973;6:45–68.
23. Konnov IV. Properties of gap functions for mixed variational inequalities. *Siberian J Numer Math* 2000;3:259–270.
24. Facchinei F, Kanzow C. Beyond monotonicity in regularization methods for nonlinear complementarity problems. *SIAM J Control Optim* 1999;37:1150–1161.
25. Malishevskii AV. Models of joint operation of many purposeful elements, II. *Automat Rem Control* 1972;33:2010–2028.
26. Polterovich VM, Spivak VA. Gross substitutable mappings in the theory of economic equilibrium. *Itogi Nauki i Tekhniki Ser Sovrem Probl Mat* 1982;19: 111–154 (in Russian).
27. Opoitsev VI. *Nonlinear system statics*. Moscow: Nauka; 1986 (in Russian).
28. Yershov EB. Theory of beaks and input-output modeling. Preprint WP2/2002/03. Moscow: State University - High Economics School; 2002. (in Russian).
29. Konnov IV. An extension of the Jacobi algorithm for multi-valued mixed complementarity problems. *Optimization* 2007;56:399–416.
30. Tamir A. Minimality and complementarity problems associated with Z-functions and M-functions. *Math Program* 1974;7:17–31.
31. Yao J-C, Chadli O. Pseudomonotone complementarity problems and variational inequalities. In: Hadjisavvas N, Komlósi S, Schaible S, editors. *Handbook of generalized convexity and generalized monotonicity*. New York: Springer; 2005. pp. 501–558.
32. Karamardian S. Generalized complementarity problems. *J Optim Theory Appl* 1971;8:161–167.

33. Kinderlehrer D, Stampacchia G. An introduction to variational inequalities and their applications. New York: Academic Press; 1980.
34. Murty KG. Linear complementarity, linear and nonlinear programming. Berlin: Heldermann; 1988.
35. Lemke CE. Bimatrix equilibrium points and mathematical programming. *Manage Sci* 1965;11:681–689.
36. Habetler GJ, Kostreva MM. On a direct algorithm for nonlinear complementarity problems. *SIAM J Control Optim* 1978;16:504–511.
37. Kostreva MM. Block pivot methods for solving the complementarity problem. *Linear Algebra Appl* 1978;21:207–215.
38. Pang J-S, Chan D. Iterative methods for variational and complementarity problems. *Math Program* 1982;24:284–313.
39. Korpelevich GM. Extragradient method for finding saddle points and other problems. *Matecon* 1976;12:747–756.
40. Konnov IV. Combined relaxation methods for variational inequalities. Berlin: Springer; 2001.
41. Fukushima M. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Math Program* 1992;53:99–110.
42. Fukushima M. Merit functions for variational inequality and complementarity problems. In: Di Pillo G, Giannessi F, editors. *Nonlinear optimization and applications*. New York: Plenum Press; 1996. pp. 155–170.
43. Mangasarian OL, Solodov MV. Nonlinear complementarity as unconstrained and constrained minimization. *Math Program* 1993;62:277–298.
44. Peng J-M, Yuan YM. Unconstrained methods for generalized complementarity problems. *J Comput Appl Math* 1997;15:253–264.
45. Fischer A. A special Newton-type optimization method. *Optimization* 1992;24:269–284.
46. Clarke FH. *Optimization and nonsmooth analysis*. New York: John Wiley & Sons; 1983.
47. Konnov IV. A splitting type algorithm for multi-valued complementarity problems. *Optim Lett* 2009;3:573–582.
48. Konnov IV. Methods of coordinate relaxation for multivalued complementarity problems. *Comput Math Math Phys* 2009;49:979–993.
49. Bakushinskii AB, Goncharskii AV. Iterative solution methods for ill-posed problems. Moscow: Nauka; 1989. (in Russian).
50. Konnov IV, Volotskaya EO. Mixed variational inequalities and economic equilibrium problems. *J Appl Math* 2002;2:289–314.
51. Qi HD. Tikhonov regularization for variational inequality problems. *J Optim Theory Appl* 1999;102:193–201.
52. Konnov IV, Ali MSS, Mazurkevich EO. Regularization of nonmonotone variational inequalities. *Appl Math Optim* 2006;53:311–330.
53. Konnov IV. On the convergence of a regularization method for variational inequalities. *Comput Math Math Phys* 2006;46:541–547.
54. Martinet B. Regularization d'inéquations variationnelles par approximations successives. *Rev Franc Inform Rech Opér* 1970;4:154–159.
55. Rockafellar RT. Monotone operators and the proximal point algorithm. *SIAM J Control Optim* 1976;14:877–898.
56. Yamashita N, Imai J, Fukushima M. The proximal point algorithm for the P0 complementarity problem. In: Ferris MC, Mangasarian, OL, Pang, J-S, editors. *Complementarity: applications, algorithms, and extensions*. Dordrecht: Kluwer Academic Publishers; 2001. pp. 361–379.
57. Allevi E, Gnudi A, Konnov IV. The proximal point method for nonmonotone variational inequalities. *Math Methods Oper Res* 2006;63:553–565.
58. Konnov IV, Kostenko TA. Multivalued mixed complementarity problem. *Russ Math (Iz VUZ)* 2004;48(12):26–33.
59. Gol'shtein EG, Tret'yakov NV. *Augmented Lagrange functions*. Moscow: Nauka; 1989. (English translation in John Wiley & Sons, New York, 1996.)
60. Lapin AV. Domain decomposition and parallel solution of free boundary problems. *Proc Lobachevsky Math Center* 2001;13:90–126.
61. Konnov IV. An extension of the Jacobi algorithm for the complementarity problem in the presence of multivalence. *Comput Math Math Phys* 2005;45:1127–1132.
62. Konnov IV. On a class of D-gap functions for mixed variational inequalities. *Russ Math (Iz VUZ)* 1999;43(12):60–64.

COMPUTATION AND DYNAMIC PROGRAMMING

HUSEYIN TOPALOGLU
School of Operations Research and
Information Engineering, Cornell
University, Ithaca, New York

Dynamic programming is a powerful tool for solving sequential decision-making problems that take place under uncertainty. One appeal of dynamic programming is that it provides a structured approach for computing the value function, which assesses the cost implications of being in different states. The value function can ultimately be used to construct an optimal policy to control the evolution of the system over time. However, the practical use of dynamic programming as a computational tool has traditionally been limited. In many applications, the number of possible states can be so large that it becomes intractable to compute the value function for every possible value of the state. This is especially the case when the state is itself a high dimensional vector and the number of possible values for the state grows exponentially with the number of dimensions of the vector. This difficulty is compounded by the fact that computing the value function requires taking expectations and it may be difficult to estimate or store the transition probability matrices that are involved in these expectations. Finally, we need to solve optimization problems to find the best action to take in each state and these optimization problems can be intractable when the number of possible actions is large.

In this article, we give an overview of computational dynamic programming approaches that are directed toward addressing the difficulties described above. Our presentation begins with the value and policy iteration algorithms. These algorithms are perhaps the most standard approaches for solving dynamic programs and they provide a sound starting point for the subsequent development, but they quickly

lose their tractability when the number of states or actions is large. Following these standard approaches, we turn our attention to simulation-based methods, such as the temporal difference learning and Q-learning algorithms, that avoid dealing with transition probability matrices explicitly when computing expectations. We cover variants of these methods that alleviate the difficulty associated with storing the value function by using parameterized approximation architectures. The idea behind these variants is to use a succinctly parameterized representation of the value function and tune the parameters of this representation by using simulated trajectories of the system. Another approach for approximating the value function is based on solving a large linear program to tune the parameters of a parameterized approximation architecture. The appealing aspect of the linear programming approach is that it naturally provides a lower bound on the value function. There are methods that use a combination of regression and simulation to construct value function approximations. In particular, we cover the approximate policy iteration algorithm that uses regression in conjunction with simulated cost trajectories of the system to estimate the discounted cost incurred by a policy. Following this, we describe rollout policies that build on an arbitrary policy and improve the performance of this policy with reasonable amount of work. Finally, we cover state aggregation, which deals with the large number of states by partitioning them into a number of subsets and assuming that the value function takes a constant value over each subset. Given the introductory nature of this article and the limited space, our coverage of computational dynamic programming approaches is not exhaustive. In our conclusions, we point out other approaches and extensions, such as linear and piecewise linear value function approximations and Lagrangian relaxation-based decomposition methods.

There are excellent books on approximate dynamic programming that focus on

computational aspects of dynamic programming. Bertsekas and Tsitsiklis [1] lay out the connections of dynamic programming with the stochastic approximation theory. They give convergence results for some computational dynamic programming methods by viewing these methods as stochastic approximation algorithms. Sutton and Barto [2] provide a perspective from the computer science angle and document a large body of work done by the authors on reinforcement learning. Si *et al.* [3] cover computational dynamic programming methods with significant emphasis on applications from a variety of engineering disciplines. Powell [4] focuses on dynamic programs where the state and the action are high dimensional vectors. Such dynamic programs pose major challenges as one can neither enumerate over all possible values of the state to compute the value function nor enumerate overall possible actions to decide which action to take. The author presents methods that can use mathematical programming tools to decide which action to take and introduces a post decision state variable that cleverly bypasses the necessity to compute expectations explicitly. The development in Powell [4] is unique in the sense that it simultaneously addresses the computational difficulties associated with the number of states, the number of actions, and the necessity to compute expectations. Bertsekas [5] provides a variety of computational dynamic programming tools. The tools in that book chapter deal with the size of the state space by using parameterized representations of the value function and avoid computing expectations by using simulated trajectories of the system. Many value function approximation approaches are rooted in standard stochastic approximation methods. Comprehensive overviews of the stochastic approximation theory can be found in Bertsekas and Tsitsiklis [1], Kushner and Clark [6], Benveniste *et al.* [7], and Kushner and Yin [8]. In this paragraph, we have broadly reviewed the literature on computational dynamic programming and value function approximation methods, but throughout the article, we point to the

relevant literature in detail at the end of each section.

The rest of the article is organized as follows. In the section titled “Notation and Problem Setup,” we formulate a Markov decision problem, give a characterization of the optimal policy, and briefly describe the value and policy iteration algorithms for computing the optimal policy. In the section titled “Policy Evaluation with Monte Carlo Simulation,” we give simulation-based approaches for computing the discounted cost incurred by a policy. We describe a simple procedure for iteratively updating an approximation to the discounted cost. This updating procedure is based on a standard stochastic approximation iteration and we use the updating procedure to motivate and describe the temporal difference learning method. In the section titled “Model-Free Learning,” we cover the Q-learning algorithm. An important feature of the Q-learning algorithm is that it avoids dealing with the transition probability matrices when deciding which action to take. This feature becomes useful when we do not have a precise model of the system that describes how the states visited by the system evolve over time. In the section titled “Linear Programming Approach,” we demonstrate how we can use a large scale linear program to fit an approximation to the value function. This linear program often has too many constraints to be solved directly and we describe computational methods to solve the linear program. In the section titled “Approximate Policy Iteration,” we describe an approximate version of the policy iteration algorithm that uses regression to estimate the discounted cost incurred by a policy. In the section titled “Rollout Policies,” we explain rollout policies and show that a rollout policy always improves the performance of the policy from which it is derived. In the section titled “State Aggregation,” we show how to use state aggregation to partition the set of states into a number of subsets and assume that the value function is constant over each subset. In the section titled “Conclusions,” we conclude by describing some other methods and possible extensions that can further enhance the computational appeal of dynamic programming.

NOTATION AND PROBLEM SETUP

In this article, we are interested in infinite horizon, discounted cost Markov decision problems with finite sets of states and actions, which we respectively denote by $\mathcal{S}$ and $\mathcal{U}$. If the system is in state i and we use action u , then the system moves to state j with probability $p_{ij}(u)$ and we incur a cost of $g(i, u, j)$, where $|g(i, u, j)| < \infty$. The costs in the future time periods are discounted by a factor $\alpha \in [0, 1)$ per time period. We use $\mathcal{U}(i)$ to denote the set of admissible actions when the system is in state i . We assume that $\mathcal{S} = \{1, 2, \dots, n\}$ so that there are n states. For brevity, we restrict our attention to infinite horizon, discounted cost Markov decision problems, but it is possible to extend the algorithms in this article to finite horizon problems or average cost criterion.

A Markovian deterministic policy π is a mapping from $\mathcal{S}$ to $\mathcal{U}$ that describes which action to take for each possible state. As a result, the states visited by the system under policy π evolve according to a Markov chain with the transition probability matrix $P^\pi = \{p_{ij}(\pi(i)) : i, j \in \mathcal{S}\}$. Letting $\{i_0^\pi, i_1^\pi, \dots\}$ be the states visited by this Markov chain, if we start in state i and use policy π , then the discounted cost that we incur can be written as

$$\begin{aligned} J^\pi(i) &= \lim_{T \rightarrow \infty} \mathbb{E} \left\{ \sum_{t=0}^T \alpha^t g(i_t^\pi, \pi(i_t^\pi), i_{t+1}^\pi) \mid i_0^\pi = i \right\}. \end{aligned}$$

Using Π to denote the set of Markovian deterministic policies, the optimal policy π^* satisfies $J^{\pi^*}(i) = \min_{\pi \in \Pi} J^\pi(i)$ for all $i \in \mathcal{S}$, giving the minimum discounted cost starting from each state. This policy can be obtained by solving the optimality equation

$$J(i) = \min_{u \in \mathcal{U}(i)} \left\{ \sum_{j \in \mathcal{S}} p_{ij}(u) [g(i, u, j) + \alpha J(j)] \right\} \quad (1)$$

for $\{J(i) : i \in \mathcal{S}\}$ and letting $\pi^*(i)$ be the optimal solution to the optimization problem on the right side above. If we let $J^* = \{J^*(i) : i \in \mathcal{S}\}$ be a solution to the optimality equation (1), then $J^*(i)$ corresponds to the optimal

discounted cost when we start in state i . We refer to $J^* \in \mathbb{R}^n$ as the “value function” and J^* can be interpreted as a mapping from $\mathcal{S}$ to $\mathbb{R}$, giving the optimal discounted cost when we start in a particular state.

The optimality equation (1) characterizes the optimal policy, but it does not provide an algorithmic tool for actually computing the optimal policy. We turn our attention to standard algorithmic tools that can be used to solve the optimality equation (1).

Value Iteration Algorithm

Throughout the rest of this article, it is useful to interpret the value function J^* not only as a mapping from $\mathcal{S}$ to $\mathbb{R}$, but also as an n dimensional vector whose i th component $J^*(i)$ gives the optimal discounted cost when we start in state i . For $J \in \mathbb{R}^n$, we define the nonlinear operator T on $\mathbb{R}^n$ as

$$[TJ](i) = \min_{u \in \mathcal{U}(i)} \left\{ \sum_{j \in \mathcal{S}} p_{ij}(u) [g(i, u, j) + \alpha J(j)] \right\}, \quad (2)$$

where $[TJ](i)$ denotes the i th component of the vector TJ . In this case, the optimality equation (1) can succinctly be written as $J = TJ$ and the value function J^* is a fixed point of the operator T . The idea behind the value iteration algorithm is to find a fixed point of the operator T by starting from an initial vector $J^1 \in \mathbb{R}^n$ and successively applying the operator T . In particular, the value iteration algorithm generates a sequence of vectors $\{J^k\}_k$ with $J^{k+1} = TJ^k$.

It is possible to show that the operator T is a contraction mapping on $\mathbb{R}^n$ so that it has a unique fixed point and the sequence of vectors $\{J^k\}_k$ with $J^{k+1} = TJ^k$ converge to the unique fixed point of the operator T . Since the value function J^* is a fixed point of the operator T , the value iteration algorithm indeed converges to J^* .

Policy Iteration Algorithm

The sequence of vectors $\{J^k\}_k$ generated by the value iteration algorithm do not necessarily correspond to the discounted cost $J^\pi = \{J^\pi(i) : i \in \mathcal{S}\}$ incurred by some policy π . In

contrast, the policy iteration algorithm generates a sequence of vectors that correspond to the discounted costs incurred by different policies. To describe the policy iteration algorithm, for $J \in \mathbb{R}^n$ and $\pi \in \Pi$, we define the linear operator T_π on $\mathbb{R}^n$ as

$$[T_\pi J](i) = \sum_{j \in S} p_{ij}(\pi(i)) [g(i, \pi(i), j) + \alpha J(j)]. \quad (3)$$

In this case, it is possible to show that the discounted cost J^π incurred by policy π is a fixed point of the operator T_π satisfying $J^\pi = T_\pi J^\pi$. Since the operator T_π is linear, we can find a fixed point of this operator by solving a system of linear equations. In particular, if we let P^π be the transition probability matrix $\{p_{ij}(\pi(i)) : i, j \in S\}$, $g^\pi(i) = \sum_{j \in S} p_{ij}(\pi(i)) g(i, \pi(i), j)$ be the expected cost that we incur when we are in state i and follow policy π and $g^\pi = \{g^\pi(i) : i \in S\}$ be the vector of expected costs, then $T_\pi J$ in Equation (3) can be written in vector notation as

$$T_\pi J = g^\pi + \alpha P^\pi J.$$

Therefore, we can find a J^π that satisfies $J^\pi = T_\pi J^\pi$ by solving the system of linear equations $J^\pi = g^\pi + \alpha P^\pi J^\pi$, which has the solution $J^\pi = (I - \alpha P^\pi)^{-1} g^\pi$, where I denotes the $n \times n$ identity matrix and the superscript -1 denotes the matrix inverse. The policy iteration algorithm computes the optimal policy by starting from an initial policy π^1 and generating a sequence of policies $\{\pi^k\}_k$ through the following two steps.

Step 1 [Policy Evaluation]. Compute the discounted cost J^{π^k} incurred by policy π^k , possibly by solving the system of linear equations $J^{\pi^k} = g^{\pi^k} + \alpha P^{\pi^k} J^{\pi^k}$.

Step 2 [Policy improvement]. For all $i \in S$, let policy π^{k+1} be defined such that we have

$$\begin{aligned} \pi^{k+1}(i) &= \operatorname{argmin}_{u \in \mathcal{U}(i)} \left\{ \sum_{j \in S} p_{ij}(u) [g(i, u, j) + \alpha J^{\pi^k}(j)] \right\}. \end{aligned} \quad (4)$$

Noting the definition of the operator T , policy π^{k+1} that is obtained in Step 2 satisfies $TJ^{\pi^k} = T_{\pi^{k+1}}J^{\pi^k}$. The algorithm stops when the policies at two successive iterations satisfy $\pi^k = \pi^{k+1}$. If we have $\pi^k = \pi^{k+1}$, then we obtain $TJ^{\pi^k} = T_{\pi^{k+1}}J^{\pi^k} = T_{\pi^k}J^{\pi^k} = J^{\pi^k}$ so that J^{π^k} is a fixed point of the operator T , which implies that J^{π^k} is the solution to the optimality equation (1). In the last chain of equalities, the first equality follows by the definition of the policy iteration algorithm, the second equality follows by the fact that $\pi^{k+1} = \pi^k$ at termination, and the last equality follows by the fact that the discounted cost incurred by policy π^k is the fixed point of the operator T_{π^k} .

The value and policy iteration algorithms are perhaps the most standard approaches for solving Markov decision problems, but they quickly lose their computational tractability. If the number of states is large, then computing and storing J^k in the value iteration algorithm and J^{π^k} in the policy iteration algorithm becomes difficult. Furthermore, if the number of actions is large, then solving the optimization problems in Equations (2) and (4) becomes intractable. Finally, solving these optimization problems requires computing expectations according to certain transition probability matrices and leaving the computation of such expectations aside, it can be very costly to even estimate and store the transition probability matrices. Throughout this article, we try to resolve these difficulties by giving algorithms that tend to be computationally more appealing than the value and policy iteration algorithms.

The books by Puterman [9] and Bertsekas [10] are modern and complete references on the theory of Markov decision processes. The notation that we use in this article follows that in Bertsekas [10] closely. Early analysis of the value iteration algorithm is attributed to Shapley [11] and Blackwell [12], whereas the policy iteration algorithm dates back to Bellman [13] and Howard [14]. Bellman [13] coins the term *curse of dimensionality* to refer to the difficulty associated with the large number of states, especially when the state is itself a high dimensional vector. Powell [4]

observes that not only the number of states, but also the number of admissible actions and the necessity to compute expectations can create computational difficulties.

There are a number of extensions for the value and policy iteration algorithms. It is possible to characterize the convergence rate of the value iteration algorithm and use this result to estimate the number of iterations to compute the value function with a certain precision. There is a Gauss Seidel variant of the value iteration algorithm that applies the operator T in an asynchronous manner. Modified policy iteration algorithm evaluates a policy only approximately in the policy evaluation step. There are methods to eliminate the suboptimal actions and such methods ensure that the operator T can be applied faster. Furthermore, the sequence of vectors generated by the value iteration algorithm converge to the value function only in the limit, but the action elimination methods may allow identifying the optimal policy without waiting for the value iteration algorithm to converge to the value function. We focus on problems with finite state and action spaces under the discounted cost criterion, but extensions to more general state and action spaces under average cost and undiscounted cost criteria are possible. Bertsekas [5], Puterman [9], Bertsekas [10,15], and Sennott [16] are good references to explore the extensions that we mention in this paragraph.

POLICY EVALUATION WITH MONTE CARLO SIMULATION

In many cases, it is necessary to compute the discounted cost J^π incurred by some policy π . Computing the discounted cost incurred by a policy requires solving a system of linear equations of the form $J^\pi = g^\pi + \alpha P^\pi J^\pi$ and solving this system may be difficult, especially when the dimensions of the matrices are too large or when we simply do not have the data to estimate the transition probability matrix P^π . In these situations, we may want to use simulation to estimate the discounted cost incurred by a particular policy.

Assume that we generate a sequence of states $\{i_0, i_1, \dots\}$ by following policy π , which

is the policy that we want to evaluate. The initial state i_0 is chosen arbitrarily and the state transitions are sampled from the transition probability matrix P^π associated with policy π . Noting the discussion in the paragraph above, we may not have access to the full transition probability matrix P^π , but the sequence of states $\{i_0, i_1, \dots\}$ may be generated either by experimenting with the system in real time or by making use of the data related to the evolution of the system in the past. Since the sequence of states $\{i_0, i_1, \dots\}$ come from the transition probability matrix P^π , the discounted cost J^π incurred by policy π satisfies $J^\pi(i_k) = \mathbb{E}\{g(i_k, \pi(i_k), i_{k+1}) + \alpha J^\pi(i_{k+1})\}$. In this case, we can keep a vector $J \in \mathbb{R}^n$ as an approximation to J^π and update one component of this vector after every state transition by

$$J(i_k) \leftarrow (1 - \gamma)J(i_k) + \gamma[g(i_k, \pi(i_k), i_{k+1}) + \alpha J(i_{k+1})], \quad (5)$$

where $\gamma \in [0, 1]$ is a step size parameter. The idea behind the updating procedure above is that $J(i_k)$ on the right side of Equation (5) is our estimate of $J^\pi(i_k)$ just before the k th state transition. Noting that $J^\pi(i_k) = \mathbb{E}\{g(i_k, \pi(i_k), i_{k+1}) + \alpha J^\pi(i_{k+1})\}$, if J and J^π are close to each other, then $g(i_k, \pi(i_k), i_{k+1}) + \alpha J(i_{k+1})$ can be interpreted as a noisy estimate $J^\pi(i_k)$ that we obtain during the k th state transition. Therefore, the updating procedure in Equation (5) combines the current estimate and the new noisy estimate by putting the weights $1 - \gamma$ and γ on them. This kind of an updating procedure is motivated by the standard stochastic approximation theory and if the step size parameter γ converges to zero at an appropriate rate, then it is possible to show that the vector J converges to J^π with probability 1 as long as every state is sampled infinitely often.

The simulation-based approach outlined above does not require storing of the transition probability matrix, but it still requires storing of the n dimensional vector J , which can be problematic when the number of states is large. One approach to alleviate this storage requirement is to use a parameterized approximation architecture. In particular, we

approximate $J^\pi(i)$ with

$$\tilde{J}(i, r) = \sum_{\ell=1}^L r_\ell \phi_\ell(i), \quad (6)$$

where $\{\phi_\ell(\cdot) : \ell = 1, \dots, L\}$ are fixed functions specified by the model builder and $r = \{r_\ell : \ell = 1, \dots, L\}$ are adjustable parameters. We can view $\{\phi_\ell(i) : \ell = 1, \dots, L\}$ as the features of state i that are combined in a linear fashion to form an approximation to $J^\pi(i)$. For this reason, it is common to refer to $\{\phi_\ell(\cdot) : \ell = 1, \dots, L\}$ as the *feature functions*. The goal is to tune the adjustable parameters r so that $\tilde{J}(\cdot, r)$ is a good approximation to J^π . There are a number of algorithms to tune the adjustable parameters. One simulation-based approach iteratively updates the vector $r = \{r_\ell : \ell = 1, \dots, L\} \in \mathbb{R}^L$ after every state transition by

$$\begin{aligned} r &\leftarrow r + \gamma [g(i_k, \pi(i_k), i_{k+1}) \\ &\quad + \alpha \tilde{J}(i_{k+1}, r) - \tilde{J}(i_k, r)] \nabla_r \tilde{J}(i_k, r) \\ &= r + \gamma [g(i_k, \pi(i_k), i_{k+1}) + \alpha \tilde{J}(i_{k+1}, r) \\ &\quad - \tilde{J}(i_k, r)] \phi(i_k) \end{aligned} \quad (7)$$

where we use $\nabla_r \tilde{J}(i_k, r)$ to denote the gradient of $\tilde{J}(i_k, r)$ with respect to r and $\phi(i_k)$ to denote the L dimensional vector $\{\phi_\ell(i_k) : \ell = 1, \dots, L\}$. The equality above simply follows by the definition of $\tilde{J}(i_k, r)$ given in Equation (6).

Bertsekas and Tsitsiklis [1] justify the updating procedure in Equation (7) as a stochastic gradient iteration to minimize the squared deviation between the discounted cost approximation $\tilde{J}(\cdot, r)$ and the simulated cost trajectory of the system. They are able to provide convergence results for this updating procedure, but as indicated by Powell [4], one should be careful about the choice of the step size parameter γ to obtain desirable empirical convergence behavior. Another way to build intuition into the updating procedure in Equation (7) is to consider the case where $L = n$ and $\phi_\ell(i) = 1$ whenever $\ell = i$ and $\phi_\ell(i) = 0$ otherwise, which implies that we have $\tilde{J}(i, r) = r_i$ for all $i \in \{1, \dots, n\}$. This situation corresponds to the case where the number of feature

functions is as large as the number of states and the discounted cost approximation $\tilde{J}(\cdot, r)$ in Equation (6) can exactly capture the actual discounted cost J^π . In this case, $\phi(i_k)$ becomes the n dimensional unit vector with a one in the i_k th component and the updating procedure in Equation (7) becomes $r_{i_k} \leftarrow r_{i_k} + \gamma [g(i_k, \pi(i_k), i_{k+1}) + \alpha r_{i_{k+1}} - r_{i_k}]$. We observe that the last updating procedure can be written as $r_{i_k} \leftarrow (1 - \gamma) r_{i_k} + \gamma [g(i_k, \pi(i_k), i_{k+1}) + \alpha r_{i_{k+1}}]$, which is identical to Equation (5). Therefore, if the number of feature functions is as large as the number of states, then the updating procedures in Equations (5) and (7) become equivalent.

We can build on the ideas above to construct more sophisticated tools for simulation-based policy evaluation. The preceding development is based on the fact that J^π satisfies $J^\pi(i_k) = \mathbb{E}[g(i_k, \pi(i_k), i_{k+1}) + \alpha J^\pi(i_{k+1})]$, but it is also possible to show that J^π satisfies

$$\begin{aligned} J^\pi(i_k) &= \mathbb{E} \left\{ \frac{\sum_{l=k}^{\tau-1} (\lambda \alpha)^{l-k} [g(i_l, \pi(i_l), i_{l+1})] \right. \\ &\quad \left. + \alpha J^\pi(i_{l+1}) - J^\pi(i_l) \right\} \\ &\quad + J^\pi(i_k), \end{aligned}$$

for any stopping time τ and any $\lambda \in (0, 1]$. The identity above can be seen by noting that the expectation of the expression in the square brackets is zero, but a more rigorous derivation can be found in Bertsekas and Tsitsiklis [1]. This identity immediately motivates the following stochastic approximation approach to keep and update an approximation $J \in \mathbb{R}^n$ to the discounted cost J^π . We generate a sequence of states $\{i_0, i_1, \dots, i_\tau\}$ by following policy π until the stopping time τ . Once the entire simulation is over, letting $d_l = g(i_l, \pi(i_l), i_{l+1}) + \alpha J(i_{l+1}) - J(i_l)$ for notational brevity, we update the approximation $J \in \mathbb{R}^n$ by

$$J(i_k) \leftarrow J(i_k) + \gamma \sum_{l=k}^{\tau-1} (\lambda \alpha)^{l-k} d_l, \quad (8)$$

for all $\{i_k : k = 0, \dots, \tau - 1\}$. After updating J , we generate another sequence of states until the stopping time τ and continue in a similar fashion. When $\lambda = 1$ and τ deterministically

takes the value 1, the updating procedures in Equations (5) and (8) become equivalent.

The quantity d_l is referred to as the *temporal difference* and the updating procedures similar to that in Equation (8) are called *temporal difference learning*. For different values of $\lambda \in [0, 1]$, we obtain a whole class of algorithms and this class of algorithms is commonly denoted as TD(λ). The approach that we describe above works in batch mode in the sense that it updates the approximation for a number of states after each simulation trajectory is over and it can be viewed as an offline version. There are also online versions of TD(λ) that update the approximation after every state transition. If the number of states is large, then carrying out the updating procedure in Equation (8) for TD(λ) can be difficult as it requires storing of the n dimensional vector J . It turns out that one can use a parameterized approximation architecture similar to that in Equation (6) to alleviate the storage problem. In this case, we update the L dimensional vector r of adjustable parameters by using

$$r \leftarrow r + \gamma \sum_{k=0}^{\tau-1} \nabla_r \tilde{J}(i_k, r) \sum_{l=k}^{\tau-1} (\lambda \alpha)^{l-k} d_l, \quad (9)$$

where the temporal difference above is defined as $d_l = g(i_l, \pi(i_l), i_{l+1}) + \alpha \tilde{J}(i_{l+1}, r) - \tilde{J}(i_l, r)$. Similar to the updating procedure in Equation (7), we can build intuition for the updating procedure in Equation (9) by considering the case where $L = n$ and $\phi_\ell(i) = 1$ whenever $\ell = i$ and $\phi_\ell(i) = 0$ otherwise. In this case, it is not difficult to see that the updating procedures in Equations (8) and (9) become equivalent.

Temporal difference learning has its origins in Sutton [17,18]. The presentation in this section is based on Bertsekas and Tsitsiklis [1], where the authors give convergence results for numerous versions of TD(λ), including online and offline versions with different values for λ , applied to infinite horizon discounted cost and stochastic shortest path problems. The book authored by Sutton [2] is another complete reference on temporal difference learning. Tsitsiklis and Van Roy [19] analyze temporal difference learning with parameterized approximation architectures.

MODEL-FREE LEARNING

Assuming that we have a good approximation to the value function, to be able to choose an action by using this value function approximation, we need to replace the vector J in the right side of the optimality equation (1) with the value function approximation and solve the resulting optimization problem. This approach requires computing an expectation that involves the transition probabilities $\{p_{ij}(u) : i, j \in S, u \in \mathcal{U}\}$. We can try to estimate this expectation by using simulation, but it is natural to ask whether we can come up with a method that bypasses estimating expectations explicitly. This is precisely the goal of model-free learning, or more specifically the Q-learning algorithm.

The Q-learning algorithm is based on an alternative representation of the optimality equation (1). If we let $Q(i, u) = \sum_{j \in S} p_{ij}(u) [g(i, u, j) + \alpha J(j)]$, then Equation (1) implies that $J(i) = \min_{u \in \mathcal{U}(i)} Q(i, u)$ and we can write the optimality equation (1) as

$$Q(i, u) = \sum_{j \in S} p_{ij}(u) [g(i, u, j) + \alpha \min_{v \in \mathcal{U}(j)} Q(j, v)]. \quad (10)$$

In this case, if $Q^* = \{Q^*(i, u) : i \in S, u \in \mathcal{U}(i)\}$ is a solution to the optimality equation above, then it is optimal to take the action $\arg \min_{u \in \mathcal{U}(i)} Q^*(i, u)$ when the system is in state i . One interpretation of $Q^*(i, u)$ is that it is the optimal discounted cost given that the system starts in state i and the first action is u . The fundamental idea behind the Q-learning algorithm is to solve the optimality equation (10) by using a stochastic approximation iteration. The algorithm generates a sequence of states and actions $\{i_0, u_0, i_1, u_1, \dots\}$ such that $u_k \in \mathcal{U}(i_k)$ for all $k = 0, 1, \dots$. It keeps an approximation $Q \in \mathbb{R}^{n \times |\mathcal{U}|}$ to Q^* and updates this approximation after every state transition by

$$Q(i_k, u_k) \leftarrow (1 - \gamma) Q(i_k, u_k) + \gamma [g(i_k, u_k, s_k) + \alpha \min_{v \in \mathcal{U}(s_k)} Q(s_k, v)], \quad (11)$$

where the successor state s_k of i_k is sampled according to the probabilities $\{p_{i_k j}(u_k) : j \in S\}$. The rationale behind the

updating procedure above is similar to that in Equation (5) in the sense that $Q(i_k, u_k)$ on the right side of Equation (11) is our current estimate of $Q^*(i_k, u_k)$ just before the k th state transition and the expression in the square brackets is our new noisy estimate of $Q^*(i_k, u_k)$. If every state action pair is sampled infinitely often in the trajectory of the system and the step size parameter γ satisfies certain conditions, then it can be shown that the approximation kept by the Q-learning algorithm converges to the solution to the optimality equation (10).

For the convergence result to hold, the sequence of states and actions $\{i_0, u_0, i_1, u_1, \dots\}$ can be sampled in an arbitrary manner as long as every state action pair is sampled infinitely often. However, the Q-learning algorithm is often used to control the real system as it evolves in real time. In such situations, given that the system is currently in state i_k , it is customary to choose $u_k = \operatorname{argmin}_{u \in \mathcal{U}(i_k)} Q(i_k, u)$, where Q is the current approximation to Q^* . The hope is that if the approximation Q is close to Q^* , then the action u_k is the optimal action when the system is in state i_k . After implementing the action u_k , the system moves to some state i_{k+1} and it is also customary to choose the successor state s_k in the updating procedure in Equation (11) as i_{k+1} . The reason behind this choice is that after implementing the action u_k in state i_k , the system naturally moves to state i_{k+1} according to the transition probabilities $\{p_{i_k, j}(u_k) : j \in S\}$ and choosing $s_k = i_{k+1}$ ensures that the successor state s_k is also chosen according to these transition probabilities. To ensure that every state action pair is sampled infinitely often, with a small probability, the action u_k is randomly chosen among the admissible actions instead of choosing $u_k = \operatorname{argmin}_{u \in \mathcal{U}(i_k)} Q(i_k, u)$. Similarly, with a small probability, the system is forced to move to a random state. This small probability is referred to as the *exploration probability*.

An important advantage of the Q-learning algorithm is that once we construct a good approximation Q , if the system is in state i , then we can simply take the action $\operatorname{argmin}_{u \in \mathcal{U}(i)} Q(i, u)$. In this way, we avoid dealing with transition probability matrices

when choosing an action. However, the Q-learning algorithm still requires storing of an $n \times |\mathcal{U}|$ dimensional approximation, which can be quite large in practical applications. There is a commonly used, albeit a heuristic, variant of the Q-learning algorithm that uses a parameterized approximation architecture similar to that in Equation (6). Letting $\tilde{Q}(i, u, r)$ be an approximation to $Q^*(i, u)$ parameterized by the L dimensional vector r of adjustable parameters, this variant uses the updating procedure

$$r \leftarrow r + \gamma [g(i_k, u_k, s_k) + \alpha \min_{v \in \mathcal{U}(s_k)} \tilde{Q}(s_k, v, r) - \tilde{Q}(i_k, u_k, r)] \nabla_r \tilde{Q}(i_k, u_k, r) \quad (12)$$

to tune the adjustable parameters. Bertsekas and Tsitsiklis [1] provide a heuristic justification for the updating procedure in Equation (12). In Equation (11), if we have $g(i_k, u_k, s_k) + \alpha \min_{v \in \mathcal{U}(s_k)} Q(s_k, v) \geq Q(i_k, u_k)$, then we increase the value of $Q(i_k, u_k)$. Similarly, in Equation (12), if we have $g(i_k, u_k, s_k) + \alpha \min_{v \in \mathcal{U}(s_k)} \tilde{Q}(s_k, v, r) \geq \tilde{Q}(i_k, u_k, r)$, then we would like to increase the value of $\tilde{Q}(i_k, u_k, r)$, but we can change the value of $\tilde{Q}(i_k, u_k, r)$ only by changing the adjustable parameters r . In Equation (12), if we have $g(i_k, u_k, s_k) + \alpha \min_{v \in \mathcal{U}(s_k)} \tilde{Q}(s_k, v, r) \geq \tilde{Q}(i_k, u_k, r)$ and the ℓ th component of $\nabla_r \tilde{Q}(i_k, u_k, r)$ is nonnegative so that $\tilde{Q}(i_k, u_k, r)$ is an increasing function of r_ℓ , then we increase the ℓ th component of r . The hope is that this changes the ℓ th component of r in the right direction so that $\tilde{Q}(i_k, u_k, r)$ also increases. Of course, this is not guaranteed in general since all of the components of r change simultaneously in the updating procedure in Equation (11). As a result, the updating procedure largely remains a heuristic.

The Q-learning algorithm was proposed by Watkins [20] and Watkins and Dayan [21]. Sutton and Barto [2], Si *et al.* [3] and Barto *et al.* [22] give comprehensive overviews of the research revolving around the Q-learning algorithm. Bertsekas and Tsitsiklis [1] and Tsitsiklis [23] show that the Q-learning algorithm fits within the general framework of stochastic approximation methods and provide convergence results by building on and extending the

standard stochastic approximation theory. The updating procedure in Equation (11) has convergence properties, but that in Equation (12) with a parameterized approximation architecture is a heuristic. Bertsekas and Tsitsiklis [1] indicate that the updating procedure in Equation (12) has convergence properties when the approximation architecture corresponds to state aggregation, where the entire set of state action pairs is partitioned into L subsets $\{\mathcal{X}_\ell : \ell = 1, \dots, L\}$ and the feature functions in the approximation architecture $\hat{Q}(i, u, r) = \sum_{\ell=1}^L r_\ell \phi_\ell(i, u)$ satisfy $\phi_\ell(i, u) = 1$ whenever $(i, u) \in \mathcal{X}_\ell$ and $\phi_\ell(i, u) = 0$ otherwise. Furthermore, they note that the updating procedure in Equation (12) is also convergent for certain optimal stopping problems. Tsitsiklis and Van Roy [24] apply the Q-learning algorithm to optimal stopping problems arising from the option pricing setting. Kunnumkal and Topaloglu [25,26] use projections within the Q-learning algorithm to exploit the known structural properties of the value function so as to improve the empirical convergence rate. The projections used by Kunnumkal and Topaloglu [25] are with respect to the L -2 norm and they can be computed as solutions to least squares regression problems. The authors exploit the fact that solutions to least squares regression problems can be computed fairly easily and one can even come up with explicit expressions for the same.

LINEAR PROGRAMMING APPROACH

Letting $\{\theta(i) : i \in \mathcal{S}\}$ be an arbitrary set of strictly positive scalars that add up to $1 - \alpha$, the solution to the optimality equation (1) can be found by solving the linear program

$$\max \sum_{i \in \mathcal{S}} \theta(i) J(i) \quad (13)$$

$$\begin{aligned} \text{subject to } J(i) - \sum_{j \in \mathcal{S}} \alpha p_{ij}(u) J(j) \\ \leq \sum_{j \in \mathcal{S}} p_{ij}(u) g(i, u, j) \\ i \in \mathcal{S}, \tilde{u} \in \mathcal{U}(i), \end{aligned} \quad (14)$$

where the decision variables are $J = \{J(i) : i \in \mathcal{S}\}$. To see this, we note that a feasible solution J to the linear program satisfies $J(i) \leq \min_{u \in \mathcal{U}(i)} \{\sum_{j \in \mathcal{S}} p_{ij}(u) [g(i, u, j) + \alpha J(j)]\}$ for all $i \in \mathcal{S}$ so that we have $J \leq TJ$. It is a standard result in the Markov decision process literature that the operator T is monotone in the sense that if $J \leq J'$, then $TJ \leq TJ'$. Therefore, we obtain $J \leq TJ \leq T^2 J$ for any feasible solution J , where the last inequality follows by noting that $J \leq TJ$ and applying the operator T on both sides of the inequality. Continuing in this fashion, we obtain $J \leq T^k J$ for any $k = 0, 1, \dots$. In this case, letting J^* be the solution to the optimality equation (1), since the value iteration algorithm implies that $\lim_{k \rightarrow \infty} T^k J = J^*$, we obtain $J \leq J^*$ for any feasible solution J to the linear program, which yields $\sum_{i \in \mathcal{S}} \theta(i) J(i) \leq \sum_{i \in \mathcal{S}} \theta(i) J^*(i)$. Therefore, $\sum_{i \in \mathcal{S}} \theta(i) J^*(i)$ is an upper bound on the optimal objective value of the linear program. We can check that J^* is a feasible solution to the linear program, in which case, J^* should also be the optimal solution. Associating the dual multipliers $\{x(i, u) : i \in \mathcal{S}, \tilde{u} \in \mathcal{U}(i)\}$ with the constraints, the dual of the linear program is

$$\min \sum_{i \in \mathcal{S}} \sum_{u \in \mathcal{U}(i)} \sum_{j \in \mathcal{S}} p_{ij}(u) g(i, u, j) x(i, u) \quad (15)$$

$$\begin{aligned} \text{subject to } \sum_{u \in \mathcal{U}(i)} x(i, u) - \sum_{j \in \mathcal{S}} \sum_{u \in \mathcal{U}(j)} \alpha p_{ji}(u) x(j, u) &= \theta(i) \\ i &\in \mathcal{S} \end{aligned} \quad (16)$$

$$x(i, u) \geq 0 \quad i \in \mathcal{S}, \tilde{u} \in \mathcal{U}(i). \quad (17)$$

In problem (15)–(17), the decision variable $x(i, u)$ is interpreted as the stationary probability that we are in state i and take action u . In particular, if we add the first set of constraints over all $i \in \mathcal{S}$ and note that $\sum_{i \in \mathcal{S}} p_{ji}(u) = 1$ and $\sum_{i \in \mathcal{S}} \theta(i) = 1 - \alpha$, we obtain $\sum_{i \in \mathcal{S}} \sum_{u \in \mathcal{U}(i)} x(i, u) = 1$ so that $x = \{x(i, u) : i \in \mathcal{S}, \tilde{u} \in \mathcal{U}(i)\}$ can indeed be interpreted as probabilities. If we let (J^*, x^*) be an optimal primal dual solution pair to problems (13)–(14) and (15)–(17), then by the discussion above, J^* is the solution to the optimality equation (1). Furthermore, if

$x^*(i, u) > 0$, then we obtain

$$\begin{aligned} \min_{v \in \mathcal{U}(i)} \left\{ \sum_{j \in S} p_{ij}(v) [g(i, v, j) + \alpha J^*(j)] \right\} \\ = J^*(i) = \sum_{j \in S} p_{ij}(u) [g(i, u, j) + \alpha J^*(j)], \end{aligned}$$

where the second equality follows from the complementary slackness condition for the constraint in problem (13)–(14) that corresponds to $x(i, u)$ and the fact that $x^*(i, u) > 0$. Therefore, if we have $x^*(i, u) > 0$ in the optimal solution to the problem (15)–(17), then it is optimal to take action u whenever the system is in state i .

Problem (13)–(14) has n decision variables and $n \times |\mathcal{U}|$ constraints, which can both be very large for practical applications. One way to overcome the difficulty associated with the large number of decision variables is to use a parameterized approximation architecture as in Equation (6). To choose the adjustable parameters $\{r_\ell : \ell = 1, \dots, L\}$ in the approximation architecture, we plug $\sum_{\ell=1}^L r_\ell \phi_\ell(i)$ for $J(i)$ in problem (13)–(14) and solve the linear program

$$\begin{aligned} \max \quad & \sum_{i \in S} \sum_{\ell=1}^L \theta(i) r_\ell \phi_\ell(i) \quad (18) \\ \text{subject to} \quad & \sum_{\ell=1}^L r_\ell \phi_\ell(i) \\ & - \sum_{j \in S} \sum_{\ell=1}^L \alpha p_{ij}(u) r_\ell \phi_\ell(j) \\ & \leq \sum_{j \in S} p_{ij}(u) g(i, u, j) \\ & i \in S, \forall u \in \mathcal{U}(i), \quad (19) \end{aligned}$$

where the decision variables are $r = \{r_\ell : \ell = 1, \dots, L\}$. The number of decision variables in problem (18)–(19) are L , which can be manageable, but the number of constraints is still $n \times |\mathcal{U}|$. In certain cases, it may be possible to solve the dual of problem (18)–(19) by using column generation, but this requires a lot of structure in the column generation subproblems. One other approach is to randomly sample some state action pairs and include

the constraints only for the sampled state action pairs.

An important feature of problem (18)–(19) is that it naturally provides a lower bound on the optimal discounted cost. In particular, if r^* is an optimal solution to the problem (18)–(19), then letting $\tilde{J}(i, r^*) = \sum_{\ell=1}^L r_\ell^* \phi_\ell(i)$, we observe that $\{\tilde{J}(i, r^*) : i \in S\}$ is a feasible solution to the problem (13)–(14), in which case, the discussion at the beginning of this section implies that $\tilde{J}(\cdot, r^*) \leq J^*$. Such lower bounds on the optimal discounted cost can be useful when we try to get a feel for the optimality gap of a heuristic policy. On the other hand, an undesirable aspect of the approximation strategy in problem (18)–(19) is that the quality of the approximation that we obtain from this problem can depend on the choice of $\theta = \{\theta(i) : i \in S\}$. We emphasize that this is in contrast to problem (13)–(14), which computes the optimal discounted cost for an arbitrary choice of θ . The paper of de Farias and Weber [27] investigates the important question of how to choose θ . However, work is needed in this area. Letting $\tilde{J}(\cdot, r) = \sum_{\ell=1}^L r_\ell \phi_\ell(\cdot)$ and using $\|\cdot\|_\theta$ to denote the θ weighted L_1 norm on $\mathbb{R}^n$, de Farias and Van Roy [28] show that the problem (18)–(19) computes an r that minimizes the error $\|J^* - \tilde{J}(\cdot, r)\|_\theta$ subject to the constraint that $\tilde{J}(\cdot, r) \leq T\tilde{J}(\cdot, r)$. This result suggests that we should use larger values of $\theta(i)$ for the states at which we want to approximate the value function more accurately. These states may correspond to those that we visit frequently or those that have serious cost implications. For this reason, $\{\theta(i) : i \in S\}$ are referred to as the *state relevance weights*. One common idea for choosing the state relevance weights is to simulate the trajectory of a reasonable policy and choose the state relevance weights as the frequency with which we visit different states. There are also some approximation guarantees for problem (18)–(19). Letting r^* be an optimal solution to this problem, de Farias and Van Roy [28] show that

$$\|J^* - \tilde{J}(\cdot, r^*)\|_\theta \leq \frac{2}{1 - \alpha} \min_r \|J^* - \tilde{J}(\cdot, r)\|_\infty. \quad (20)$$

The left side above is the error in the value function approximation provided by problem (18)–(19), whereas the right side above is the smallest error possible in the value function approximation if we were allowed to use any value for r . The right side can be viewed as the power of the parameterized approximation architecture in Equation (6). Therefore, Equation (20) shows that if our approximation architecture is powerful, then we expect problem (18)–(19) to return a good value function approximation.

The linear programming approach for constructing value function approximations was proposed by Schweitzer and Seidmann [29] and it has seen revived interest with the work of de Farias and Van Roy [28]. Noting that the right side of Equation (20) measures the distance between J^* and $\tilde{J}(\cdot, r)$ according to the max norm and it can be quite large in practice, de Farias and Van Roy [28] provide refinements on this approximation guarantee. Since the number of constraints in problem (18)–(19) can be large, de Farias and Van Roy [30] study the effectiveness of randomly sampling some of the constraints to approximately solve this problem. Adelman and Mersereau [31] compare the linear programming approach to other value function approximation methods. Desai *et al.* [32] observe that relaxing constraints (19) may actually improve the quality of the approximation and they propose relaxing these constraints subject to a budget on the total relaxation amount. The work by de Farias and Van Roy [33,34] extends the linear programming approach to average cost criterion. Adelman [35] and Veatch and Walker [36] give examples in the revenue management and queueing control settings, where the column generation subproblem for the dual of problem (18)–(19) becomes tractable.

APPROXIMATE POLICY ITERATION

In practice, both steps of the policy iteration algorithm in the section titled “Notation and Problem Setup” can be problematic. In the policy evaluation step, we need to compute the discounted cost incurred by a policy π , potentially by solving the system of linear

equations $J = T_\pi J$ for $J \in \mathbb{R}^n$. In the policy improvement step, we need to find a policy π that satisfies $TJ = T_\pi J$ for a given vector $J \in \mathbb{R}^n$. Carrying out these two steps exactly is almost never tractable, as it requires dealing with large dimensional matrices.

It is possible to use regression and simulation methods, along with parameterized approximation architectures as in Equation (6), to approximately carry out the policy iteration algorithm. In the approximate version of the policy evaluation step, letting π^k be the policy that we need to evaluate, we generate a sequence of states $\{i_0, i_1, \dots\}$ by following policy π^k . If we let $C_l = \sum_{t=l}^{\infty} \alpha^{t-l} g(i_t, \pi^k(i_t), i_{t+1})$ for all $l = 0, 1, \dots$, then C_l provides a sample of the discounted cost incurred by policy π^k given that we start in state i_l . Therefore, letting r^k be the solution to the regression problem

$$\begin{aligned} \min_r \left\{ \sum_{l=0}^{\infty} [\tilde{J}(i_l, r) - C_l]^2 \right\} \\ = \min_r \left\{ \sum_{l=0}^{\infty} \left[\sum_{\ell=1}^L r_\ell \phi_\ell(i_l) - C_l \right]^2 \right\}, \quad (21) \end{aligned}$$

we can use $\tilde{J}(\cdot, r^k)$ as an approximation to the discounted cost J^{π^k} incurred by policy π^k . In the policy improvement step, on the other hand, we need to find a policy π^{k+1} such that $\pi^{k+1}(i)$ is the optimal solution to the problem $\min_{u \in \mathcal{U}(i)} \{ \sum_{j \in \mathcal{S}} p_{ij}(u) [g(i, u, j) + \alpha J^{\pi^k}(j)] \}$ for all $i \in \mathcal{S}$. In the approximate version of the policy improvement step, we let policy π^{k+1} be such that $\pi^{k+1}(i)$ is the optimal solution to the problem $\min_{u \in \mathcal{U}(i)} \{ \sum_{j \in \mathcal{S}} p_{ij}(u) [g(i, u, j) + \alpha \tilde{J}(j, r^k)] \}$ for all $i \in \mathcal{S}$. If the expectation inside the curly brackets cannot be computed exactly, then we resort to simulation. We note that J^{π^k} used in the policy improvement step of the policy iteration algorithm corresponds to the discounted cost incurred by policy π^k , whereas $\tilde{J}(\cdot, r^k)$ used in the policy improvement step of the approximate policy iteration algorithm does not necessarily correspond to the discounted cost incurred by policy π^k . It is also worthwhile to point out that since we need the action taken by policy π^k when generating the sequence of states $\{i_0, i_1, \dots\}$,

we do not have to compute the action taken by policy π^k in every possible state. We can simply compute the action taken by policy π^k as needed when generating the sequence of states $\{i_0, i_1, \dots\}$.

There is some theoretical justification for the approximate policy iteration algorithm. We observe that there are two potential sources of error in the algorithm. First, the regression in the policy evaluation step may not accurately capture the discounted cost incurred by policy π^k . We assume that this error is bounded by ϵ in the sense that $|\tilde{J}(i, r^k) - J^{\pi^k}(i)| \leq \epsilon$ for all $i \in \mathcal{S}$ and $k = 0, 1, \dots$. Second, we may not exactly compute the action taken by policy π^{k+1} in the policy improvement step. This is especially the case when we use simulation to estimate the expectations. We assume that this error is bounded by δ in the sense that $|[T_{\pi^{k+1}}\tilde{J}(\cdot, r^k)](i) - [T\tilde{J}(\cdot, r^k)](i)| \leq \delta$ for all $i \in \mathcal{S}$ and $k = 0, 1, \dots$, where we use $[T_{\pi^{k+1}}\tilde{J}(\cdot, r^k)](i)$ and $[T\tilde{J}(\cdot, r^k)](i)$ to denote the i th components of the vectors $T_{\pi^{k+1}}\tilde{J}(\cdot, r^k)$ and $T\tilde{J}(\cdot, r^k)$, respectively. Bertsekas and Tsitsiklis [1] show that the approximate policy iteration algorithm generates a sequence of policies whose optimality gaps are bounded by ϵ and δ . In particular, the sequence of policies $\{\pi^k\}_k$ generated by the approximate policy iteration algorithm satisfy $\limsup_{k \rightarrow \infty} \|J^{\pi^k} - J^*\|_\infty \leq (\delta + 2\alpha\epsilon)/(1 - \alpha)$. This result provides some justification for the approximate policy iteration algorithm, but it is clearly difficult to measure ϵ and δ . If we use simulation to estimate the expectations, then there is always a small probability of incurring a large simulation error and coming up with a uniform bound on the simulation error may also not be possible.

A practical issue that requires clarification within the context of the approximate policy iteration algorithm is that the computation of $\{C_l : l = 0, 1, \dots\}$ and the regression problem in Equation (21) involve infinite sums, which cannot be computed in practice. To address this difficulty, for large finite integers N and M with $M < N$, we can generate a sequence of N states $\{i_0, i_1, \dots, i_N\}$ and compute the discounted cost starting from each one of the states $\{i_0, i_1, \dots, i_{N-M}\}$ until

we reach the final state i_N . This amounts to letting $C_l = \sum_{t=l}^{N-1} \alpha^{t-l} g(i_t, \pi^k(i_t), i_{t+1})$ for $l = 0, 1, \dots, N - M$ and we can use these sampled discounted costs in the regression problem. It is straightforward to check that if we have $l \in \{0, 1, \dots, N - M\}$, then the two sums $\sum_{t=l}^{N-1} \alpha^{t-l} g(i_t, \pi^k(i_t), i_{t+1})$ and $\sum_{t=l}^{\infty} \alpha^{t-l} g(i_t, \pi^k(i_t), i_{t+1})$ differ by at most $\alpha^M \bar{G}/(1 - \alpha)$, where we let $\bar{G} = \max_{i,j \in \mathcal{S}, u \in \mathcal{U}(i)} |g(i, u, j)|$. Therefore, truncating the sequence of states may not create too much error as long as M is large.

Bertsekas and Tsitsiklis [1] describe an approximate version of the value iteration algorithm that also uses regression. At any iteration k , this algorithm keeps a vector of adjustable parameters r^k that characterize a value function approximation through the parameterized approximation architecture $\tilde{J}(\cdot, r^k)$ as in Equation (6). We apply the operator T on $\tilde{J}(\cdot, r^k)$ to compute $\{[T\tilde{J}(\cdot, r^k)](i) : i \in \tilde{\mathcal{S}}\}$ for only a small set of representative states $\tilde{\mathcal{S}} \subseteq \mathcal{S}$. The adjustable parameters r^{k+1} at the next iteration are given by the optimal solution to the regression problem $\min_r \sum_{i \in \tilde{\mathcal{S}}} [\tilde{J}(i, r) - [T\tilde{J}(\cdot, r^k)](i)]^2$. Bertsekas and Tsitsiklis [1] give performance bounds for this approximate value iteration algorithm. They indicate that one way of selecting the set of representative states $\tilde{\mathcal{S}}$ is to simulate the trajectory of a reasonable policy and focus on the states visited in the simulated trajectory. Noting the definition of the operator T , computing $[T\tilde{J}(\cdot, r^k)](i)$ requires taking an expectation and one can try to approximate this expectation by using simulation, especially when we do not have the data to estimate the transition probability matrices. Approximate value and policy iteration algorithms that we describe in this section are due to Bertsekas and Tsitsiklis [1] and Bertsekas [10]. Tsitsiklis and Van Roy [37] also give an analysis for the approximate value iteration algorithm.

ROLLOUT POLICIES

The idea behind rollout policies is to build on a given policy and improve the performance

of this policy with reasonable amount of work. For a given policy π with discounted cost J^π , we define policy π' such that $\pi'(i)$ is the optimal solution to the problem $\min_{u \in \mathcal{U}(i)} \{\sum_{j \in \mathcal{S}} p_{ij}(u)[g(i, u, j) + \alpha J^\pi(j)]\}$ for all $i \in \mathcal{S}$. In this case, we can use elementary properties of the policy iteration algorithm to show that policy π' is at least as good as policy π . In other words, the discounted cost $J^{\pi'}$ incurred by policy π' is no larger than the discounted cost J^π incurred by policy π . To see this result, we observe that the definition of policy π' implies that

$$\begin{aligned} & \sum_{j \in \mathcal{S}} p_{ij}(\pi'(i))[g(i, \pi'(i), j) + \alpha J^\pi(j)] \\ & \leq \sum_{j \in \mathcal{S}} p_{ij}(\pi(i))[g(i, \pi(i), j) + \alpha J^\pi(j)] = J^\pi(i), \end{aligned}$$

for all $i \in \mathcal{S}$, where the equality follows by the fact that J^π is the fixed point of the operator T_π . We write the inequality above as $T_{\pi'} J^\pi \leq J^\pi$. It is a standard result in the Markov decision process literature that the operator $T_{\pi'}$ is monotone and applying this operator on both sides of the last inequality, we obtain $T_{\pi'}^2 J^\pi \leq T_{\pi'} J^\pi \leq J^\pi$. Continuing in this fashion, we have $T_{\pi'}^k J^\pi \leq J^\pi$ for all $k = 0, 1, \dots$. By using the value iteration algorithm under the assumption that the only possible policy is π' , we have $\lim_{k \rightarrow \infty} T_{\pi'}^k J^\pi = J^{\pi'}$. Therefore, the last inequality implies that $J^{\pi'} \leq J^\pi$.

The biggest hurdle in finding the action chosen by policy π' is the necessity to know J^π . If policy π has a simple structure, then it may be possible to compute J^π exactly, but in general, we need to estimate J^π by using simulation. In particular, given that the system is in state i , to be able to find the action chosen by policy π' , we try each action $u \in \mathcal{U}(i)$ one by one. For each action u , we simulate the transition of the system from state i to the next state. From that point on, we follow the actions chosen by policy π . Accumulating the discounted cost incurred along the way yields a sample of $\sum_{j \in \mathcal{S}} p_{ij}(u)[g(i, u, j) + \alpha J^\pi(j)]$. By simulating multiple trajectories, we can use a sample average to estimate

$\sum_{j \in \mathcal{S}} p_{ij}(u)[g(i, u, j) + \alpha J^\pi(j)]$. Once we estimate this quantity for all $u \in \mathcal{U}(i)$, the action that yields the smallest value for $\sum_{j \in \mathcal{S}} p_{ij}(u)[g(i, u, j) + \alpha J^\pi(j)]$ is the action chosen by policy π' when the system is in state i . This approach is naturally subject to simulation error, but it often performs well in practice and it can significantly improve the performance of the original policy π . It is also worthwhile to emphasize that the original policy π can be an arbitrary policy, including policies driven by value function approximations or heuristics.

There are numerous applications of rollout policies, where one can significantly improve the performance of the original policy. Tesauro and Galperin [38] use rollout policies to strengthen several strategies for playing backgammon. Bertsekas *et al.* [39] view certain combinatorial optimization problems as sequential decision processes and improve the performance of several heuristics by using rollout policies. Bertsekas and Castanon [40] provide applications on stochastic scheduling problems and Secomandi [41] focuses on vehicle routing problems. Yan *et al.* [42] generate strategies for playing solitaire by using rollout policies.

STATE AGGREGATION

An intuitive idea to deal with the large number of states is to use state aggregation, where we partition the state space into a number of subsets and assume that the value function is constant over each partition. To this end, we use $\{\mathcal{X}_\ell : \ell = 1, \dots, L\}$ to denote the partition of the states such that $\cup_{\ell=1}^L \mathcal{X}_\ell = \mathcal{S}$ and $\mathcal{X}_\ell \cap \mathcal{X}_{\ell'} = \emptyset$ for all $\ell \neq \ell'$. We assign the value r_ℓ to the value function over the partition $\mathcal{X}_\ell$. The interesting question is whether we can develop an algorithm that finds a good set of values for $\{r_\ell : \ell = 1, \dots, L\}$. Using $\mathbf{1}(\cdot)$ to denote the indicator function, one approach is to generate a sequence of states $\{i_0, i_1, \dots\}$ so that if we have $i_k \in \mathcal{X}_\ell$, then we update r_ℓ by using the stochastic approximation iteration

$$\begin{aligned}
r_\ell &\leftarrow (1 - \gamma)r_\ell \\
&+ \gamma \min_{u \in \mathcal{U}(i_k)} \sum_{j \in S} p_{ij}(u) \\
&\left[g(i_k, u, j) + \alpha \sum_{l=1}^L \mathbf{1}(j \in \mathcal{X}_l) r_l \right] \quad (22)
\end{aligned}$$

immediately after observing state i_k . One can develop convergence results for the updating procedure in Equation (22) under fairly general assumptions on how the sequence of states $\{i_0, i_1, \dots\}$ are sampled. For example, Bertsekas and Tsitsiklis [1] sample a subset $\mathcal{X}_\ell$ uniformly over $\{\mathcal{X}_\ell : \ell = 1, \dots, L\}$. Given that the subset $\mathcal{X}_\ell$ is sampled, they sample $i_k \in \mathcal{X}_\ell$ according to the probabilities $\{q_\ell(i) : i \in \mathcal{X}_\ell\}$. The probabilities $\{q_\ell(i) : i \in \mathcal{X}_\ell\}$ can be arbitrary except for the fact that they add up to one.

We can expect state aggregation to work well as long as the value function is relatively constant within each of the subsets $\{\mathcal{X}_\ell : \ell = 1, \dots, L\}$. We let $\epsilon_\ell = \max_{i, j \in \mathcal{X}_\ell} |J^*(i) - J^*(j)|$ to capture how much the value function fluctuates within the subset $\mathcal{X}_\ell$. In this case, letting $\epsilon = \max_{\ell=1, \dots, L} \epsilon_\ell$, Tsitsiklis [1] show that the updating procedure in Equation (22) converges to a point $\hat{r} = \{\hat{r}_\ell : \ell = 1, \dots, L\}$ that satisfies $|J^*(i) - \hat{r}_\ell| \leq \epsilon/(1 - \alpha)$ for all $i \in \mathcal{X}_\ell$ and $\ell = 1, \dots, L$. Therefore, we estimate the value function with an approximation error of $\epsilon/(1 - \alpha)$. We note that the limiting point of the updating procedure can depend on the choice of the probabilities $\{q_\ell(i) : i \in \mathcal{X}_\ell\}$, but the choice of these probabilities ultimately does not affect the convergence result and the fact that the limiting point provides an approximation error of $\epsilon/(1 - \alpha)$.

We note that a parameterized approximation architecture such as that in Equation (6) is actually adequate to represent state aggregation. In particular, we can define the feature functions $\{\phi_\ell(\cdot) : \ell = 1, \dots, L\}$ such that $\phi_\ell(i) = 1$ whenever $i \in \mathcal{X}_\ell$ and $\phi_\ell(i) = 0$ otherwise. In this case, the adjustable parameter r_ℓ in the parameterized approximation architecture $\sum_{\ell=1}^L r_\ell \phi_\ell(\cdot)$ captures the value assigned to the value function over

the partition $\mathcal{X}_\ell$. Bean *et al.* [43] analyze state aggregation for shortest path problems. Tsitsiklis and Van Roy [37] and Van Roy [44] provide convergence results for updating procedures similar to that in Equation (22). Singh *et al.* [45] pursue the idea of soft state aggregation, where each state belongs to a subset $\mathcal{X}_\ell$ with a particular probability.

CONCLUSIONS

In this article, we have described a number of approaches for alleviating the computational difficulties associated with dynamic programming. Methods such as the temporal difference learning and Q-learning algorithm avoid transition probability matrices by using simulated trajectories of the system to estimate expectations. There are variants of these methods that use parameterized approximation architectures and the goal of these variants is to address the difficulty associated with storing the value function. The linear programming approach solves a large linear program to fit a parameterized approximation architecture to the value function. Approximate policy iteration uses a combination of regression and simulation in an effort to obtain a sequence of policies that improve on each other. Rollout policies start with an arbitrary policy and improve the performance of this policy with relatively small amount of work. State aggregation builds on the intuitive notion of partitioning the state space into a number of subsets and assuming that the value function is constant over each partition.

Owing to the introductory nature of this article and the limited space, our coverage of computational dynamic programming approaches has not been exhaustive. One important point is that if we have a good approximation $\tilde{J}(\cdot, r)$ to the value function, then to be able to make the decisions by using this value function approximation, we need to plug the value function approximation in the right side of the optimality equation (1) and solve the resulting optimization problem. If the number of admissible actions in each state is small, then we can solve the

resulting optimization problem by checking the objective function value provided by each action one by one, but this approach becomes problematic when the number of admissible actions is large. In problems where the action is itself a high dimensional vector, it may be possible to choose the feature functions $\{\phi_\ell(\cdot) : \ell = 1, \dots, L\}$ such that the optimization problem in question decomposes by each component of the action vector. Another way to deal with the resulting optimization problem is to choose the feature functions such that the value function approximation $\tilde{J}(\cdot, r) = \sum_{\ell=1}^L r_\ell \phi_\ell(\cdot)$ ends up having a special structure and the resulting optimization problem can be solved by using standard optimization tools. For example, if the state and the action take values in the Euclidean space, then we may use linear or piecewise linear functions of the state as the feature functions and it may be possible to solve the resulting optimization problem by using linear or integer programming tools. Godfrey and Powell [46,47], Topaloglu and Powell [48], Schenk and Klabjan [49], and Simao *et al.* [50] use linear and piecewise linear value function approximations in numerous dynamic programs that arise from the freight transportation setting. By using such value function approximations, they are able to deal with states and actions that are themselves vectors with hundreds of dimensions.

If the state and the action take values in the Euclidean space, then another useful approach is to decompose the Markov decision problem in such a way that one can obtain approximations to the value function by concentrating on each component of the state separately. Adelman and Mersereau [31] use the term *weakly coupled dynamic program* to refer to dynamic programs that would decompose if a few linking constraints did not couple the decisions acting on different components of the vector valued state. They use Lagrangian relaxation to relax these complicating constraints. They propose methods to choose a good set of Lagrange multipliers and show that their Lagrangian relaxation idea provides lower bounds on the value function. Karmarkar [51], Cheung and Powell [52], Castanon [53] and Topaloglu

[54] apply Lagrangian relaxation to dynamic programs arising from the inventory allocation, fleet management, sensor management, and revenue management settings.

Our development in this article assumes that the feature functions $\{\phi_\ell(\cdot) : \ell = 1, \dots, L\}$ are fixed and $\{\phi_\ell(i) : \ell = 1, \dots, L\}$ are able to capture the important features of state i from the perspective of estimating the discounted cost starting from this state. The construction of the feature functions is traditionally left to the model builder and this task may require quite a bit of insight into the problem at hand and a considerable amount of trial and error. There is some recent work on developing automated algorithms to construct the feature functions and this is an area that can be beneficial to all of the methods that we described in this article. Klabjan and Adelman [55] use a special class of piecewise linear functions as possible candidates for the feature functions. They show that it is possible to approximate any function with arbitrary precision by using enough number of functions from their special class. Veatch [56] considers control problems in the queueing network setting and describes an approach for iteratively adding feature functions from a predefined set.

REFERENCES

1. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming. Belmont (MA): Athena Scientific; 1996.
2. Sutton RS, Barto AG. Reinforcement learning. Cambridge (MA): The MIT Press; 1998.
3. Si J, Barto AG, Powell WB, *et al.*, editors. Handbook of learning and approximate dynamic programming. Piscataway (NJ): Wiley-Interscience; 2004.
4. Powell WB. Approximate dynamic programming: solving the curses of dimensionality. Hoboken (NJ): John Wiley & Sons, Inc.; 2007.
5. Bertsekas D. Volume II, Dynamic programming and optimal control, Approximate dynamic programming. 3rd ed. Technical report. Cambridge (MA): MIT; 2010. Chapter 6 Available at <http://web.mit.edu/dimitrib/www/dpchapter.pdf>.
6. Kushner HJ, Clark DS. Stochastic approximation methods for constrained and unconstrained systems. Berlin: Springer; 1978.

7. Benveniste A, Metivier M, Priouret P. Adaptive algorithms and stochastic approximations. New York: Springer; 1991.
8. Kushner HJ, Yin GG. Stochastic approximation and recursive algorithms and applications. New York: Springer; 2003.
9. Puterman ML. Markov decision processes. New York: John Wiley & Sons, Inc.; 1994.
10. Bertsekas DP. Dynamic programming and optimal control. Belmont (MA): Athena Scientific; 2001.
11. Shapley L. Stochastic games. *Proc Natl Acad Sci U S A* 1953;39:1095–1100.
12. Blackwell D. Discounted dynamic programming. *Ann Math Stat* 1965;36:226–235.
13. Bellman R. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
14. Howard RA. Dynamic programming and Markov processes. Cambridge (MA): MIT Press; 1960.
15. Bertsekas D, Shreve S. Stochastic optimal control: the discrete time case. New York: Academic Press; 1978.
16. Sennott LI. Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Oper Res* 1989;37:626–633.
17. Sutton RS. Temporal credit assignment in reinforcement learning [PhD thesis]. Amherst (MA): University of Massachusetts; 1984.
18. Sutton RS. Learning to predict by the methods of temporal differences. *Mach Learn* 1988;3:9–44.
19. Tsitsiklis J, Van Roy B. An analysis of temporal-difference learning with function approximation. *IEEE Trans Automat Control* 1997;42:674–690.
20. Watkins CJCH. Learning from Delayed Rewards [PhD thesis]. Cambridge, England: Cambridge University; 1989.
21. Watkins CJCH, Dayan P. *Q*-learning. *Mach Learn* 1992;8:279–292.
22. Barto AG, Bradtke SJ, Singh SP. Learning to act using real-time dynamic programming. *Artif Intell* 1995;72:81–138.
23. Tsitsiklis JN. Asynchronous stochastic approximation and *Q*-learning. *Mach Learn* 1994;16:185–202.
24. Tsitsiklis J, Van Roy B. Regression methods for pricing complex American-style options. *IEEE Trans Neural Networks* 2001;12(4):694–703.
25. Kunnumkal S, Topaloglu H. Exploiting the structural properties of the underlying Markov decision problem in the *Q*-learning algorithm. *INFORMS J Comput* 2008;20(2):288–301.
26. Kunnumkal S, Topaloglu H. Stochastic approximation algorithms and max-norm “projections”. *ACM Trans Model Comput Simul* 2009. In press.
27. de Farias DP, Weber T. Choosing the cost vector of the linear programming approach to approximate dynamic programming. *Proceedings of the 47th IEEE Conference on Decision and Control; Cancun, Mexico; 2008*. pp. 67–72.
28. de Farias DP, Van Roy B. The linear programming approach to approximate dynamic programming. *Oper Res* 2003b;51(6):850–865.
29. Schweitzer P, Seidmann A. Generalized polynomial approximations in Markovian decision processes. *J Math Anal Appl* 1985;110:568–582.
30. de Farias DP, Van Roy B. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math Oper Res* 2004;29(3):462–478.
31. Adelman D, Mersereau AJ. Relaxations of weakly coupled stochastic dynamic programs. *Oper Res* 2008;56(3):712–727.
32. Desai VV, Farias VF, Moallemi CC. Approximate dynamic programming via a smoothed linear program. Technical report. Cambridge (MA): MIT, Sloan School of Management; 2009.
33. de Farias DP, Van Roy B. Approximate linear programming for average-cost dynamic programming. Volume 15, *Advances in neural information processing systems*. Cambridge (MA): MIT Press; 2003a.
34. de Farias DP, Van Roy B. A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Math Oper Res* 2006;31(3):597–620.
35. Adelman D. Dynamic bid-prices in revenue management. *Oper Res* 2007;55(4):647–661.
36. Veatch MH, Walker N. Approximate linear programming for network control: Column generation and subproblems. Technical report. Wenham (MA): Gordon College, Department of Mathematics; 2008.
37. Tsitsiklis JN, Van Roy B. Feature-based methods for large scale dynamic programming. *Mach Learn* 1996;22:59–94.

38. Tesauro G, Galperin G. On-line policy improvement using Monte-Carlo search. Volume 9, *Advances in neural information processing*. Cambridge (MA): MIT Press; 1996. pp. 1068–1074.
39. Bertsekas D, Tsitsiklis J, Wu C. Rollout algorithms for combinatorial optimization. *J Heuristics* 1997;3(3):245–262.
40. Bertsekas D, Castanon D. Rollout algorithms for stochastic scheduling problems. *J Heuristics* 1999;5:89–108.
41. Secomandi N. A rollout policy for the vehicle routing problem with stochastic demands. *Oper Res* 2001;49(5):796–802.
42. Yan X, Diaconis P, Rusmevichientong P, *et al.* *Solitaire: man versus machine*. Volume 17, *Advances in neural information processing systems*. Cambridge (MA): MIT Press; 2005. pp. 1553–1560.
43. Bean J, Birge J, Smith R. Aggregation in dynamic programming. *Oper Res* 1987;35:215–220.
44. Van Roy B. Performance loss bounds for approximate value iteration with state aggregation. *Math Oper Res* 2006; 31(2):234–244.
45. Singh SP, Jaakkola T, Jordan MI. Reinforcement learning with soft state aggregation. Volume 7, *Advances in neural information processing systems*. Cambridge (MA): MIT Press; 1995. pp. 361–368.
46. Godfrey GA, Powell WB. An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: single period travel times. *Transp Sci* 2002a; 36(1):21–39.
47. Godfrey GA, Powell WB. An adaptive, dynamic programming algorithm for stochastic resource allocation problems II: multi-period travel times. *Transp Sci* 2002b;36(1):40–54.
48. Topaloglu H, Powell WB. Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. *INFORMS J Comput* 2006;18(1):31–42.
49. Schenk L, Klabjan D. Intra market optimization for express package carriers. *Transp Sci* 2008;42:530–545.
50. Simao HP, Day J, George AP, *et al.* An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transp Sci* 2009;43(2):178–197.
51. Karmarkar US. The multiperiod multi-location inventory problems. *Oper Res* 1981;29:215–228.
52. Cheung RK, Powell WB. An algorithm for multistage dynamic networks with random arc capacities, with an application to dynamic fleet management. *Oper Res* 1996;44(6):951–963.
53. Castanon DA. Approximate dynamic programming for sensor management. *Proceedings of the 36th Conference on Decision & Control*. San Diego (CA); 1997.
54. Topaloglu H. Using Lagrangian relaxation to compute capacity-dependent bid-prices in network revenue management. *Oper Res* 2009;57(3):637–649.
55. Klabjan D, Adelman D. An infinite-dimensional linear programming algorithm for deterministic semi-Markov decision processes on Borel spaces. *Math Oper Res* 2007;32(3):528–550.
56. Veatch M. Adaptive simulation/LP methods for queueing network control. *INFORMS Conference*. San Diego (CA); 2009.

COMPUTATIONAL BIOLOGY AND BIOINFORMATICS: APPLICATIONS IN OPERATIONS RESEARCH

DAVE GOULET
ALLEN HOLDER
Rose-Hulman Institute of
Technology - Mathematics
Terre Haute, Terre Haute, IN

THE INTERPLAY BETWEEN OPERATIONS RESEARCH AND THE LIFE SCIENCES

Operations research (OR) was a young discipline in the 1950s when it had its first brush with the biological sciences. For instance, initial computer simulations of evolution were conducted by Barricelli in the middle 1950s [1], and in the 1960s, Bremermann described how genetic algorithms could be used to solve optimization problems [2]. A more modern example of how the biological sciences have impacted our ability to heuristically search for optimality is that of ant colony optimization, which was developed in the 1990s [3]. OR's overarching goal to improve, and if possible optimize, the decision-making process has benefited from mimicking natural processes as nature itself seems to seek optimality.

While OR has adopted biological principles for its own advancement, the intrinsic disciplinary overlap suggests that the biological sciences might similarly benefit from OR. Indeed, the innate optimal quality of nature, together with the fact that OR has built a myriad of mathematical and computational methods to compute optimal quantities, has promoted the application of OR to problems in the biological sciences. Biological applications of OR tend to differ from OR applications in business and industry because the biological entities are not independent agents that can make decisions. Even so, the natural trends of many biological entities

are often toward optimal states that can be appropriately modeled with traditional OR techniques.

The application of OR to biology blossomed with the advent of the fields of computational biology and bioinformatics in the 1990s. Numerous biological investigations were moving from wet lab research to computer models that could simulate a natural process. The computational setting promised a vast increase in the speed of our experimental ability, and hence, the number of (simulated) experiments could be far larger than what would have been possible in a laboratory. Simulated results could then be used to identify which experiments should be confirmed by more costly wet lab research.

This article samples some of the OR applications in the biological sciences. The presentation is arranged relative to biological scale, starting with problems in biochemistry and moving toward legacy studies of entire populations. Other surveys of OR applications in biology are divided by the type of OR [4]. Each section below begins with a succinct introduction describing the biological relevance and ends with OR examples. Further details are found in the citations.

PROTEINS

Proteins are the functional workhorses of life. For example, proteins form the contractile elements in muscle tissue as well as the walls of the ion channels in the neurons controlling those muscles. Mammary acinar cells create the large amounts of protein in breast milk while simultaneously responding to protein signaling molecules in their environment. Proteins on the surfaces of some viruses act as syringes for injecting DNA into host cells. Other proteins span the cell walls of bacteria and combine to form molecular motors. Protein tethers and anchors pull chromosomes apart during cell division, and synthetic proteins can mark other proteins on

the surface of cancerous cells so that cytotoxins, made of protein, can kill those cancerous cells without harming healthy ones. There are proteins that allow immune cells to recognize and respond to pathogens, and there are proteins (prions) that are themselves inanimate pathogens.

Biological Introduction

Proteins in their simplest form are bonded chains of amino acid molecules. These polypeptide chains can be composed of tens of thousands of the many known types of amino acids (20 in humans), allowing staggering combinatorial complexity. The amino acid sequence of a peptide chain is known as the protein's *primary structure*.

The importance of proteins and their amino acid building blocks is illustrated by the theory that all life on the Earth started with the formation of amino acids. In 1953 Miller, under the supervision of Urey, placed water, methane, ammonia, and hydrogen into a vessel and applied heat and electricity [5]. Trace amounts of amino acids were identified 2 weeks later. In 1969, the Murchison meteorite struck Australia, and when examined, it was found to contain amino acids. More recently, amino acids were discovered in lunar soil from the Apollo missions [6]. It is tempting to interpret the genesis and apparent ubiquity of amino acids as the fascinating origins of life on the Earth; however, the primary structure only initiates the protein story.

Amino acids within a single peptide chain can form weak bonds with others nearby, leading to *secondary structures* such as α helices, β sheets, and β barrels. Secondary structures create relatively rigid and geometrically fixed regions on an otherwise flexible chain. Because these structures may form in multiple locations along the chain, combinatorial complexity beyond that of the primary sequence is possible.

Stronger bonds between distant amino acids can form either spontaneously or with the assistance of chaperon proteins. These bonds cause proteins to fold and become compact and globular, with some amino acids being internalized. This *tertiary structure* is the basis of protein function. Folding can,

in theory, be modeled by simulating molecular dynamics or quantum mechanics. However, such simulations are complicated by the action of small molecules and other proteins, which may interfere or assist with folding. A stable fold is assumed to represent a locally optimal energetic state, but other local optima may exist, and a protein may oscillate among multiple folded states. These temporal dynamics augment the spatial combinatorial complexity.

Quaternary structures are formed by multiple polypeptide chains, further increasing complexity and potential functionality. An example is that of the potassium ion channel embedded in a neuron's cell membrane. The ion channel is formed by four total protein subunits of two types, each of which can be in an open or closed state.

OR Applications and Proteins

Arguably the biggest outstanding problem in computational biology is the simulation of protein folding. The ability to accurately predict tertiary structure from primary sequence, and hence, the ability to infer functionality from amino acid chains, would revolutionize much of biology, medicine, and health care. However, this proverbial "holy grail" of computational biology has eluded our scientific and computational abilities. From an OR perspective, protein folding is a difficult nonlinear, global optimization problem that minimizes free energy. The free energy model is based on the pioneering work of Christian Anfinsen [7], which leads to the thermodynamic hypothesis stating that a protein's native state is uniquely determined by its primary sequence. There are several introductions to protein folding, see Refs [8–10] as examples.

Each amino acid is uniquely determined by a side-chain of atoms, and the torsion angle of the bond that connects a side-chain to the protein is called a *rotamer*. Only a few choices for each rotamer are possible owing to energy restrictions. The possible angles at each site are cataloged in a library. Determining side-chain conformations to minimize energy is called the *rotamer assignment problem*, which has traditionally been a nonlinear

combinatorial problem [11] (see Ref. [12] for related semidefinite optimization).

Another problem in which OR can help is the searching, comparing, and cataloging of proteins whose structures are stored in a database. The protein data bank, <http://www.rcsb.org>, collects and stores protein structures, and as of 2013, the database was nearing 100,000 structures. Traditional comparisons among proteins were made by aligning amino acid sequences, but the goal of identifying functional similarity is better addressed by aligning three dimensional folds. However, conducting all pairwise comparisons requires efficient algorithms and clever modeling. Several combinatorial approaches in the early 2000s were suggested, see Refs [13] and [14]. Such combinatorial approaches required days of computing to complete the pairwise comparisons of small, 40 protein databases, and their ability to identify known protein families was imperfect. A few years later, several groups realized how to use dynamic programming to efficiently and accurately align larger databases [15–17], and databases with 100s of proteins can now be solved in seconds with perfect accuracy. The increased efficacy has further led to stochastic studies [18].

METABOLISMS

High throughput biological experiments have produced an immense amount of information about cellular processes, and whole-cell models based on this data are becoming possible. The grand goal is to build computational models that couple gene expression and protein interaction with the metabolism. This three tiered approach to whole-cell modeling mirrors the central dogma of molecular biology, which asserts a one-way flow of information from DNA to mRNA to protein, and while imperfect, the central dogma remains the prevailing framework with which to approach cellular biology. The first tier is a network that explains the co-expression of the genes encoded by the DNA, and the second tier is a network that explains pairwise relationships between the resulting proteins. The last tier is the metabolism, which is the collection

of biochemical reactions that supports life. Linking the three networks so that they correctly imitate the regulatory and reactive mechanisms of a cell is a research question on which OR can have an impact. Indeed, standard OR tools have already established themselves substantively in the study of metabolisms as discussed below.

Biological Introduction

A cell's metabolism is the net flux of all biochemicals in and out of the cell. Cells intake carbohydrates, proteins, lipids, and many small molecules and ions. These species act as building materials and fuel for the cell as it grows, repairs itself, recycles molecules, replicates its genome, and exports materials to its environment.

The engines driving these microscopic metabolic factories are the mitochondria. These organelles convert the energy held in the bonds of nutrients into readily usable energy that is stored in the bonds of the molecule adenosine triphosphate (ATP). ATP is transported intra- and intercellularly, and it fuels the metabolism by sequentially breaking off phosphate groups to extract energy, which transforms ATP into diphosphate and monophosphate forms [ADP (adenosine diphosphate) and AMP (adenosine monophosphate)]. An aerobic metabolism produces ATP from carbohydrates using available oxygen. Energy production in animal cells is primarily aerobic. Anaerobic production of ATP is possible, although less efficient. The primary mechanism of ATP production in yeast is anaerobic, and *Escherichia coli* cells can survive in either an aerobic or anaerobic state.

Different cell types emphasize different metabolic processes and products. Muscle cells create relatively large amounts of ATP to fuel the contraction cycle in muscle fibers. Brain cells create and recycle surface receptors, ion channels, and neurotransmitters. Cells of the liver and gall bladder emphasize the production of enzymes used for digestion. Mammary cells secrete milk lipids and proteins. Amoeba secrete cyclic AMP in order to communicate with neighboring amoeba. Penicillium fungi produce antibiotics to fend off bacterial opportunists, and cells of the

hair follicles intake sulfurous compounds to use in cross-linking proteins to produce thin but durable strands of hair.

Although many components of the cellular metabolism are well characterized, the timing and quantification of a metabolism, especially for ensembles of cells, are poorly understood. A collection of cells can be viewed as a black box. In some well-controlled experiments, all inputs to and outputs from this box can be measured. Broad conclusions can be drawn about the metabolism of a typical cell, but this averaging to quantify a typical cell is deceptive. Intercellular interactions are homogenized over the ensemble, as though a single cell in isolation would be capable of accomplishing all metabolic feats on its own.

OR and Metabolic Models

Mathematical programming has proven itself to be a trusted computational tool in the study of whole-cell metabolisms. Metabolic models are constructed by listing the chemical reactions of the cell to create a stoichiometric matrix, S . The linear system $Sv = 0$ holds if the metabolism is assumed to have evolved to a steady state, where v_j is the flux of reaction j . The resulting system is underdetermined, and an objective function is introduced to help identify reasonable metabolic states. Typically, the objective is the growth rate, although others have been suggested and studied [19]. Allowing $g(v)$ to be an appropriate objective, the flux balance analysis (FBA) model is

$$\max\{g(v) : Sv = 0, L \leq v \leq U\},$$

where L and U are variable bounds on the fluxes.

A common use of FBA is to predict lethal gene knockouts. Removing a gene's expression can remove some of the resulting proteins, which can subsequently halt reactions. In cells such as *E. coli*, the map between gene knockouts and reactions is well understood, and an FBA model can replicate the gene knockout by setting the appropriate fluxes to zero. If the optimal value diminishes sufficiently, then the gene knockout is lethal and the gene is said to be essential. FBA

correctly predicts gene essentiality with over 90% accuracy [20].

FBA has been studied and extended in many ways. A quadratic model that minimizes metabolic adjustment is a common adaptation to predict gene essentiality [21]. Extreme pathways have been studied as basic optimal solutions [22], and the central metabolism can be identified if FBA's innate degeneracy is handled carefully [23]. Robust extensions to accommodate stochastic modeling parameters have also been investigated [24]. Lastly, recent directives suggest amalgamations that meld metabolic models with gene expression to include external factors such as temperature stress [25].

MICROTUBULES

Cells require infrastructure to house their metabolic outcomes much like factories require infrastructure to produce their goods. The functional scaffolding of many cells is formed by microtubules, which are rigid proteinaceous tubes similar in diameter to carbon nanowires and nanotubes. Microtubules endow cells with rigidity, transport organelles, provide cellular locomotion, regulate cell growth and geometry, and separate chromosomes during cell division.

Cells come in numerous shapes, sizes, and forms. For example, a huge ostrich egg is a single cell, as is the nucleus-free red blood cell. Microtubule interactions are responsible for the cellular structure in each case. The regulation of microtubule interactions varies depending on cell type, and computational models that simulate the growth and demise of microtubule structures aid our understanding of the complex structural dynamics.

Biological Introduction

At the start of mitotic cell division in animals, a cell's centrosome is duplicated and the resulting pair separates. The centrosomes migrate to opposite poles of the cell. A centrosome is a hub from which microtubule spokes emanate, and as cell division proceeds, tubules emanating from a centrosome lengthen and attach to chromosome pairs

aligned along the cell's midline. Because the centrosomes are anchored to the plasma membrane, the chromosomal microtubules can contract and pull chromosome pairs toward the poles. Once the centrosomes are symmetrically separated, cell division proceeds, leaving each daughter cell with a full genetic complement and a single centrosome.

Microtubules apply force by altering their length, which varies according to a dynamically unstable biochemical process. Tubulin heterodimers bind sequentially to form a hollow tube with 13 dimers per helical revolution. Because the helix is composed of tubulin heterodimers (a bonded pair of two types of tubulin monomers), the two exposed ends of the microtubule present different monomers. This polarization of the microtubule results in different binding affinities for free tubulin on each end, with the + end binding more readily than the - end. The unstable nature of microtubules allows rapid changes in length.

Microtubules grow by the addition of tubulin dimers and decay by their removal. Guanine triphosphate (GTP) biases the reaction governing the addition of tubulin to the helix, making attachment more probable. The tendency for lengthening and shortening varies with the amount of GTP. Even if GTP concentrations maintain a microtubule's length, a dynamic equilibrium is present in which tubulin dimers are constantly added and removed. In general, microtubules are always undergoing a process called *treadmilling*, meaning that tubulin dimers are continually being added and/or dropped.

Microtubules self-organize into microtubule arrays during plant cell development. These cortical microtubules (CMT) exhibit polarized $\pm$ structure, but this polarization does not result in - ends gathering at a common center. Instead, numerous local interactions between a large population of nearly identical CMTs influence dynamic organization and assembly. CMTs in plants have been observed to assemble into astral structures, bundles, and parallel cross-linked sheets. These dynamic CMT arrays act as scaffolds, directing the placement of structural fibers necessary to the formation

of a new cell wall. Movement and placement of microtubules are governed by interactions within the array as well as by treadmilling. Arrays with tubules transverse to the elongation axis are correlated to continued elongation, while arrays with longitudinal or oblique orientation correlate to the cessation of elongation.

OR and Microtubules

A three-dimensional discrete event simulation of CMT organization is developed in Refs [26] and [27]. This model probabilistically assumes that the + end of the microtubule grows, shrinks, or pauses depending on the length of time that it is in one of these three states. The transition from one state to another is modeled with an exponential distribution, and once a transition has been triggered, the + end enters the next state. The - end is modeled similarly but without the pause state. The rate of growth or decay is sampled from a normal distribution. All probabilistic parameters are tuned to mimic experimental observations.

The outcome of a random interaction between two CMTs is decided by the angle of intersection. If an intersecting CMT forms an angle of less than 40° with another CMT (called the *barrier* CMT of the interaction), then the intersecting CMT aligns with the barrier CMT. If the angle is at least 40° , then 30% of the time the intersecting CMT begins to shorten. The remaining 70% of the time the intersecting CMT passes through the barrier CMT. Intersecting the cell wall always forces the microtubule to enter a shortening phase.

Discrete event simulations are commonly used in OR to study stochastic problems, and it is well known that simple, random decision rules can lead to complex dynamics. In the case of CMT organization, a well-tuned simulation reproduces the complex observable phenomena found in cells. Moreover, the model accurately predicts mutant behavior caused by mutations in the MOR1 and FRA2 genes. The trust built by verifying the computational model's efficacy against experimental results then adds credence to computational queries about how a cell's structure might be affected by altering the interactions.

EPIDEMICS

Infectious diseases have plagued humankind throughout history, and modeling their spread is a stalwart in mathematical and computational biology. Standard OR techniques have not been commonly used, although (stochastic) differential equation models and statistical analyses are routine. Indeed, the 2005 review article [28] states that “No humanitarian emergencies (epidemics, famine, war and genocide) were addressed in the OR/MS related journals.” However, some OR work is emerging, and the magnified public safety importance associated with the spread of disease suggests that OR should be considered as the research community investigates new models. In the following section, we introduce the biology of the spread of disease and then mention where OR has made an appearance.

Biological Introduction

Infectious diseases spread from individual to individual via a transmission vector. Disease vectors include viruses, bacteria, fungi, protozoa, and proteins known as *prions*. The spread of disease can occur from one cell to another, as is the case with replicating viruses in multicellular hosts, or between individuals across vast geographical scales, as is the case with the human immunodeficiency virus (HIV) and the influenza pandemic of 1918.

Transmission vectors can occur horizontally among individuals of the same species, vertically from mother to child during gestation or birth, and from one individual to another of a different species or subspecies. The Norwalk virus exemplifies the former as it famously victimized masses of cruise ship patrons, causing symptoms akin to severe food poisoning [29]. Herpes and HIV are easily transmitted vertically. The bird and swine strains of influenza transmit back and forth between humans and other species, mutating each time [30]. The vector may be detrimental to both species or harmless to one, as is the case with malarial parasites in mosquito serum.

The initiation of spreading pathogens is termed an *outbreak*. An *epidemic* is characterized by affected individuals increasing in number and broadening in geographical locale. During a *pandemic*, the extent of infection reaches a global scale. Entities such as the Center for Disease Control and the World Health Organization analyze quantitative models to aid in detecting and containing outbreak events.

The nature of the pathogen effects its mode of transmission within and between populations. Rhinovirus (the common cold) may spread via handshake, whereas *E. coli* bacteria can wait for hosts on a doorknob or a head of lettuce. A viable fungal spore can blow into an area on the wind or be revived from fossilized amber after millions of years. Giardia protozoa are commonly drunk from wells or ponds after being expelled through the gastrointestinal tract of another species. Prions, an inanimate vector made solely of proteins, with no genetic material, can be spread among species by consuming the brains of the infected.

The spread of an infectious disease may have consequences ranging from undetectable to lethal. Warts, caused by some types of HPV (human papillomavirus), are inconvenient. Common colds may seem innocuous, but they are responsible for roughly \$40 billion of lost productivity and medical expenses each year [31]. Crippling conditions are possible from more severe infections such as polio and tuberculosis or from benign infections such as rubella in immunodeficient individuals. Some pathogens even gain control of the infected host and alter behavior [32].

OR Applications and Epidemics

The first overlap between traditional OR and the modeling of epidemics seems to be the combination of differential equations, queuing theory, and policy analysis to study the spread of smallpox [33, 34]. The intersection with OR is the use of a queue to track the individuals awaiting vaccination as a policy of distribution is employed. From an operational perspective, the model can be used to assess containment policies and to guide implementation. While the use of stochastic

processes to study epidemics has a long history, the model in Ref. [33] and the expertise of the authors is a clear bridge between OR and epidemiology. Other works have followed [35–37]. The review article [37] suggests many research opportunities.

POPULATION RECONSTRUCTION

Mapping the generational progressions of species helps us identify the evolutionary aspects that underlie numerous areas of biology. However, our ability to genetically track whole populations is recent, and in most cases, we have what is essentially a brief snap-shot of a long, long genetic evolution. Inferring previous populations from current genetic information importantly informs evolutionary processes, and OR models have found success in doing so.

Biological Introduction

Population genetic studies provide information about migrations, inheritance, mutation rates, evolution, and speciation. If the DNA sequences of a parent and its offspring are compared so as to locate small genetic alterations, for example, single nucleotide polymorphisms (SNPs or snips), then information about mutation rates from one generation to the next can be gleaned. These parent/offspring mutation rates help explain the long-term evolution of the species. However, sequence comparison between family members is only possible if the family structure is known. In wild populations, observing these familial relationships is often impossible, and the genetic information needs to be analyzed to infer family relationships.

The persistence of a population may be influenced by its ability to relocate, which, for example, enables the population to adapt to changing resources and habitat. Some prehistoric human populations were known to have migrated small distances annually along shorelines to ensure adequate marine food sources. By contrast, the arctic tern boasts the longest known migratory patterns, flying over 2 million kilometers during its 30-year life span. Plant seeds and pollen may also be transported large

distances. Air and sea currents are believed to have relocated many plant species across the Atlantic Ocean. While the geographical range occupied by a species can sometimes be observed, the migratory and dispersal mechanisms by which this range is obtained may remain hidden. Reconstructing familial relationships, possibly many generations back, enables our understanding of how populations evolve. Population reconstruction has facilitated some seed and pollen dispersal studies. Indeed, the reconstruction of family relationships in oak trees made clear that offspring could emerge at surprisingly long distances from the parent [38].

The health of a population is partially determined by its DNA sequence. Knowing the primary sequence of DNA reveals much of an organism's ability to produce proteins. Transcription of particular DNA sequences leads to the expression of particular proteins that enable cells to perform particular functions. However, knowledge of the genome and proteome does not reveal which proteins will be expressed in which cells and at what time. Illuminating the roles of RNA and regulatory proteins has led to deeper understanding of the specificity and timing of gene expression.

Further complicating the gene expression mechanism is the *epigenome*. Portions of DNA bind to *histones*, molecules that cause DNA to coil and cluster into untranscribable *chromatin*. Other molecules can alter the binding properties of histones, remodeling the chromatin and allowing transcription to occur. The quantity and variety of histone-regulating molecules is influenced by an organism's diet, environment, stress levels, and inheritance. Studies of the combined effect of inheritance and diet have revealed that nutrient intake by parents or grandparents can influence epigenetic factors in the offspring. This epigenetic memory is known to have led to the emergence of schizophrenia in offspring conceived during famine times [39]. Reconstructing populations allows deeper analysis of the modes of acquisition and inheritance of epigenetic factors.

OR Applications and Population Reconstruction

Genes are sequences of DNA that code for specific traits, and locations on the genome that distinguish individuals within a taxa are called *SNPs*. A *haplotype* is a collection of SNPs, and a *genotype* is a coupled pair of haplotypes, each of which is donated by the individual's parents. The problem of inferring haplotypes from genotypes is called *haplotyping*. Each SNP of a haplotype can have one of two states, and if the SNPs of the two sides of a genotype agree, then the genotype is homozygous at the location of the SNP. Otherwise the genotype is heterozygous at that location.

In the presence of heterozygous SNPs, the parental donations are unclear, and the problem of inferring haplotypes requires an inference rule. A classic inference rule is that of pure parsimony, which asks to identify a smallest collection of haplotypes that can combine to explain the current generation's genotypes. Another inference rule is that of perfect phylogeny, which requires that the selected haplotypes comply with the structure of a tree to ensure that each haplotype is a copy of one of the parental haplotypes. Phylogenetic design is itself a problem with substantial overlap with OR [40–42].

Both inference rules naturally lead to combinatorial optimization problems that have received attention in the OR community, see Ref. [43] as a review. A related problem is to directly identify siblings and half-siblings from genotypes, which again naturally leads to combinatorial problems that include Mendelian laws of inheritance [44]. These combinatorial methods have become accurate enough to identify flaws in existing datasets [45].

CONCLUSION

Biology and its related disciplines in medicine and health have had an impressive and impactful history. However, whereas common motivation once rested on observation and classification, many of the advances of the last few decades have been fueled by a bonanza of information. The result

is a science that stands on its ability to infer new experiments and outcomes based on our new collective of knowledge. The combination of science, data, modeling, computation, and mathematics is now at the forefront of biological research, and the application of OR is natural and productive. That said, the synergy between OR and biology is challenged by a gulf of differences in language, application, and education. The promise of joint benefit to both biology and OR demands the combined expertise of both disciplines. In the authors' experience, operations researchers, mathematicians, computer scientists, and engineers will find welcoming colleagues in the life sciences. The possibility of intriguing joint research is nearly ensured independent of the particular biological focus, indeed, OR applications already exists over a wide range of biological scale as demonstrated by this brief synopsis.

REFERENCES

1. Barricelli N. Symbiogenetic evolution processes realized by artificial methods. *Methods* 1957;9(35-36):143–182.
2. Bremermann H. Optimization through evolution and recombination. In: Yovits M, Jacobi G, Goldstine G, editors. *Self-organizing systems*. Washington (DC): Spartan Books; 1962. p 93–106.
3. Caro G, Dorigo M. Antnet: distributed stigmergetic control for communications networks. *J Artif Intell Res* 1998;9:317–365.
4. Greenberg H, Holder A. Computational biology. In: Gass S, Fu M, editors. *Encyclopedia of Operations Research and Management Science*. Springer US; 2013. p 225–238. ISBN: 978-1-4419-1137-7.
5. Miller S. A production of amino acids under possible primitive earth conditions. *Science* 1953;117(3046):528–529.
6. Elsila J, Callahan M, Glavin D, *et al.* Distribution of amino acids in lunar regolith. Conference Paper JSC-CN-30317; NASA; The Woodlands (TX); May 2014.
7. Anfinsen C. Principles that govern the folding of protein chains. *Science* 1973;181(4096):223–230.
8. Ben-Naim A. The protein folding problem and its solutions. Volume 32. Singapore: World Scientific; 2013.

9. Friesner R. In: Prigogine I, Rice S, editors. *Advances in chemical physics, computational methods for protein folding*. Volume: 120, *Advances in chemical physics*. New York: John Wiley & Sons; 2004. ISBN: 9780471209553.
10. Merz K, LeGrand S, editors. *The protein problem and tertiary structure prediction*. Boston (MA): Birkhäuser; 1994. ISBN: 978-1-4684-6833-5.
11. Kingsford C, Chazelle B, Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* 2005;21(7):1028–1039.
12. Burkowski F, Cheung Y, Wolkowicz H. Efficient use of semidefinite programming for selection of rotamers in protein conformations. *INFORMS J Comput* 2013;26:748–766.
13. Caprara A, Carr R, Istrail S, *et al.* 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap. *J Comput Biol* 2004;11(1):27–52.
14. Di Lena P, Fariselli P, Margara L, *et al.* Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics* 2010;26(18):2250–2258.
15. Bonnel N, Marteau P. LNA: fast protein structural comparison using a Laplacian characterization of tertiary structure. *IEEE/ACM Trans Comput Biol Bioinform (TCBB)* 2012;9(5):1451–1458.
16. Kifer I, Nussinov R, Wolfson H. GOSSIP: a method for fast and accurate global alignment of protein structures. *Bioinformatics* 2011;27(7):925–932.
17. Shibberu Y, Holder A. A spectral approach to protein structure alignment. *IEEE/ACM Trans Comput Biol Bioinform* 2011;8(4):867–875.
18. Holder A, Simon J, Strauser J, *et al.* Dynamic programming used to align protein structures with a spectrum is robust. *Biology* 2013;2(4):1296–1310.
19. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 2007;3:119.
20. Orth J, Conrad T, Na J, *et al.* A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Mol Syst Biol* 2011;7:535.
21. Segrè D, Vitkup D, Church G. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 2002;99(23):15112–15117. 11
22. Papin J, Price N, Palsson B. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res* 2002;12:1889–1900.
23. Almaas E, Oltvai Z, Barabasi A. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* 2005;1(7):e68.
24. Almaas E, Gruber E, Holder A, *et al.* Robust analysis of metabolic pathways. *INFORMS J Comput* 2014. Submitted.
25. Navid A, Almaas E. Genome-level transcription data of *Yersinia pestis* analyzed with a new metabolic constraint-based approach. *BMC Syst Biol* 2012;6(1):150.
26. Can Eren E, Gautam N, Dixit R. Computer simulation and mathematical models of the noncentrosomal plant cortical microtubule cytoskeleton. *Cytoskeleton* 2012;69:144–154.
27. Can Eren E, Ram D, Gautam N. A three-dimensional computer simulation model reveals the mechanisms for self-organization of plant cortical microtubules into oblique arrays. *Mol Biol* 2010;21:2674–2684.
28. Altay N, Green W III. OR/MS research in disaster operations management. *Eur J Oper Res* 2006;175:475–493.
29. Centers for Disease Control and Prevention. Outbreak updates for international cruise ships. Available at <http://www.cdc.gov/nceh/vsp/surv/gilist.htm>. Accessed 2015 Sept 28.
30. Ma W, Lager K, Vincent A, *et al.* The role of swine in the generation of novel influenza viruses. *Zoonoses Public Health* 2009;56:326–337.
31. Fendick A, Monto A, Nightengale B, *et al.* The economic burden of non-influenza-related viral respiratory tract infection in the united states. *Arch Intern Med* 2003;163(4):487–494.
32. Henne D, Johnson S. Zombie fire ant workers: behavior controlled by decapitating fly Parasitoids. *Insectes Sociaux* 2007;54(2):150–153.
33. Kaplan E, Craft D, Wein L. Analyzing bioterror response logistics: the case of smallpox. *Math Biosci* 2003;185:33–72.
34. Kaplan E, Wein L. Decision making for bioterror preparedness: examples from smallpox vaccination policy. In: Brandeau M, Sainfort F, Pierskalla W, editors. *Operations research and health care*. Volume 70: *International Series in Operations Research & Management Science*. US: Springer; 2004. p 519–536.
35. Basu S, Galvani A. The transmission and control of XDR TB in South Africa:

- an operations research and mathematical modelling approach. *Epidemiol Infect* 2008;136(12):1585–1598.
36. Trapman P, Bootsma M. A useful relationship between epidemiology and Queueing theory: the distribution of the number of infectives at the moment of the first detection. *Math Biosci* 2009;219(1):15–22.
 37. Zhang J, Mason J, Denton B, *et al.* Applications of operations research to the prevention, detection, and treatment of disease. In: Gass S, Fu M, editors. *Encyclopedia of Operations Research and Management Science*. Springer US; 2013. ISBN: 978-1-4419-1137-7.
 38. Dow B, Ashely M. Microsatellite analysis of seed dispersal and parentage of saplings in bur oak. *Mol Ecol* 1996;5:615–627.
 39. Heijmans B, Tobi E, Stein A, *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* 2008;105(44):17046–17049.
 40. Bafna V, Halldorsson B, Schwartz R, *et al.* Haplotypes and informative SNP selection algorithms: don't block out information. *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology*; Berlin, Germany: ACM; 2003. p 19–27.
 41. Catanzaro D, Ravi R, Schwartz R. A mixed integer linear programming model to reconstruct phylogenies from single nucleotide polymorphism Haplotypes under the maximum parsimony criterion. *Algorithms Mol Biol* 2013;8(1):3.
 42. Sridhar S, Lam F, Blleloch G, *et al.* Direct maximum parsimony phylogeny reconstruction from genotype data. *BMC Bioinformatics* 2007;8(1):472.
 43. Gusfield D, Orzack S. Haplotype inference. In: Aluru S, editor. *Handbook of computational molecular biology*. Boca Raton (FL): CRC Press; 2006. p 181–1825.
 44. Chaovalitwongse W, Chou C, Berger-Wolf T, *et al.* New optimization model and algorithm for sibling reconstruction from genetic markers. *INFORMS J Comput* 2009;22(2):1–15.
 45. Chou C, Liang Z, Chaovalitwongse W, *et al.* Column-generation framework of non-linear similarity model for reconstructing sibling groups. *INFORMS J Comput* 2015;27(1):35–47.

COMPUTATIONAL METHODS FOR CTMCs

TUĞRUL DAYAR
Department of Computer
Engineering, Bilkent University,
Ankara, Turkey

WILLIAM J. STEWART
Department of Computer Science,
North Carolina State University,
Raleigh, North Carolina

In a *continuous-time Markov chain* (CTMC), a change of state may occur at any point in time. We say that a stochastic process $\{X(t), t \geq 0\}$ is a CTMC if for integers (states) i, j, k and for all time instants s, t, v with $t \geq 0, s \geq 0$, and $0 \leq v \leq s$, we have

$$\begin{aligned} \text{Prob}\{X(s+t) = k | X(s) = j, X(v) = i\} \\ = \text{Prob}\{X(s+t) = k | X(s) = j\}. \end{aligned}$$

Notice from this definition that not only is the sequence of previously visited states irrelevant to the future evolution of the chain, but so too is the amount of time already spent in the current state.

If the CTMC is *nonhomogeneous*, we write

$$p_{ij}(s, t) = \text{Prob}\{X(t) = j | X(s) = i\},$$

where $X(t)$ denotes the state of the Markov chain at time $t \geq s$. On the other hand, when the CTMC is *homogeneous*, these transition probabilities depend on the difference $\tau = t - s$ rather than on the actual values of s and t . In this case, we simplify the notation by writing

$$\begin{aligned} p_{ij}(\tau) &= \text{Prob}\{X(s+\tau) = j | X(s) = i\} \\ &\text{for all } s \geq 0. \end{aligned}$$

This denotes the probability of being in state j after an interval of length τ , given that the current state is state i . It depends on the length τ but not on s , the specific moment at which this time interval begins.

Whereas a discrete-time Markov chain (DTMC) is represented by its matrix of *transition probabilities*, $P(n)$ at time step n , a CTMC is represented by its matrix of *transition rates*, $Q(t)$ at time t . In this article, we restrict our attention to homogeneous CTMCs where the elements of the transition-rate matrix, now written as Q , are given by

$$\begin{aligned} q_{ij} &= \lim_{\Delta t \rightarrow 0} \left(\frac{p_{ij}(\Delta t)}{\Delta t} \right), \quad i \neq j; \\ q_{jj} &= \lim_{\Delta t \rightarrow 0} \left(\frac{p_{jj}(\Delta t) - 1}{\Delta t} \right). \end{aligned}$$

It is apparent from these equations that the diagonal element in each row of Q is equal to the negated sum of the off-diagonal elements in that row and hence the sum of all elements in any row of Q must be zero.

Let $\pi_i(t) = \text{Prob}\{X(t) = i\}$ be the probability that a CTMC is in state i at time t . Then the probability that the CTMC is in state i at time $t + \Delta t$, correct to terms of order $o(\Delta t)$, must be equal to the probability that it is in state i at time t and it does not change state in the period $[t, t + \Delta t)$, plus the probability that it is in some state $k \neq i$ at time t and moves to state i in the interval Δt , that is,

$$\begin{aligned} \pi_i(t + \Delta t) &= \pi_i(t) \left(1 - \sum_{\text{all } j \neq i} q_{ij} \Delta t \right) \\ &\quad + \left(\sum_{\text{all } k \neq i} q_{ki} \pi_k(t) \right) \Delta t + o(\Delta t). \end{aligned}$$

Since $q_{ii} = -\sum_{\text{all } j \neq i} q_{ij}$, we may write

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \left(\frac{\pi_i(t + \Delta t) - \pi_i(t)}{\Delta t} \right) \\ = \lim_{\Delta t \rightarrow 0} \left(\sum_{\text{all } k} q_{ki} \pi_k(t) + o(\Delta t) / \Delta t \right), \end{aligned}$$

that is,

$$\frac{d\pi_i(t)}{dt} = \sum_{\text{all } k} q_{ki} \pi_k(t).$$

In matrix notation, this gives

$$\frac{d\pi(t)}{dt} = \pi(t)Q.$$

It follows that the solution $\pi(t)$ is given by

$$\pi(t) = \pi(0)e^{Qt} = \pi(0) \left(I + \sum_{k=1}^{\infty} (Qt)^k / k! \right),$$

where e^{Qt} is the *matrix exponential* defined by

$$e^{Qt} = \sum_{k=0}^{\infty} (Qt)^k / k!.$$

It is the computation of this vector $\pi(t)$, the *transient distribution* of the CTMC, that is our primary objective in this article. However, we first make some observations concerning the computation of stationary distributions of CTMCs.

STATIONARY DISTRIBUTIONS OF CTMCs

We define an invariant vector of a homogeneous CTMC with transition-rate matrix Q as any nonzero vector z for which $zQ = 0$. If this system of equations has a solution z which is a probability vector ($z_i \geq 0$ for all i and $\|z\|_1 = 1$), then z is a stationary distribution. If replacement of one of the equations by a normalizing equation causes the coefficient matrix to become nonsingular, then the stationary distribution is unique. This is the case when the CTMC is irreducible and finite.

Furthermore, in a homogeneous CTMC, the i th element of the vector $\pi(t)$ is the probability that the chain is in state i at time t and we have just seen that these state probabilities are governed by the system of differential equations

$$\frac{d\pi(t)}{dt} = \pi(t)Q.$$

If the evolution of the CTMC is such that there arrives a point in time at which the rate of change of the probability distribution vector $\pi(t)$ is zero, then the left-hand side of this equation, $d\pi(t)/dt$, is identically equal to zero. In this case, the system has reached a

limiting distribution, which is written simply as π in order to show that it no longer depends on time t .

When the limiting distribution exists, when all its components are strictly positive, and when it is independent of the initial probability vector $\pi(0)$, then it is unique and is called the *steady-state distribution*, also referred to as the *equilibrium* or *long-run probability vector*; its i th element π_i is the probability of being in state i at statistical equilibrium. For a finite, irreducible CTMC, the limiting distribution always exists and is identical to the stationary distribution of the chain. The steady-state distribution may be obtained by solving the system of linear equations

$$\pi Q = 0 \text{ with } \pi u = 1, \quad (1)$$

where u is a column vector containing 1s. With a DTMC, we saw that the steady-state distribution is obtained from

$$\pi P = \pi \text{ with } \pi u = 1, \quad (2)$$

where P is its stochastic transition probability matrix. Observe that both these equations may be put into the same form. We may write the second, $\pi P = \pi$, as $\pi(P - I) = 0$ thereby putting it into the form of Equation (1). Observe that $(P - I)$ has all the properties of a transition-rate matrix, namely, the off-diagonal elements are nonnegative; row sums are equal to zero and diagonal elements are equal to the negated sum of off-diagonal row elements. On the other hand, we may discretize a CTMC. Writing Equation (1) as

$$\pi(Q\Delta t + I) = \pi \text{ with } \pi u = 1, \quad (3)$$

puts it into the form of Equation (2). In this discretized Markov chain, transitions take place at intervals Δt , Δt being chosen sufficiently small that the probability of two transitions taking place in time Δt is negligible, that is, of order $o(\Delta t)$. One possibility is to take

$$\Delta t \leq \frac{1}{\max_i |q_{i,i}|}.$$

In this case, the matrix $(Q\Delta t + I)$ is stochastic and the stationary probability vector π of

the CTMC obtained from $\pi Q = 0$ is identical to that of the discretized chain, obtained from $\pi(Q\Delta t + I) = \pi$. It now follows that numerical methods designed to compute the stationary distribution of DTMCs may be used to compute the stationary distributions of CTMCs. Since this was the topic of the previous article on computational methods for DTMCs, we shall not pursue this topic further. We simply add a cautionary note: transient solutions of the discretized chain represented by the transition probability matrix $Q\Delta t + I$, are not the same as those of the continuous-time chain, represented by the transition-rate matrix Q , although both do have the same stationary distribution.

Before terminating this section, we consider the problem of computing moments of first passage times from a state to a set of states in CTMCs. Without loss of generality, let us assume that the state space is partitioned as $S = \mathcal{F} \cup \mathcal{T}$, $\mathcal{F} \cap \mathcal{T} = \emptyset$, yielding the block partitioning

$$Q = \begin{pmatrix} Q_{\mathcal{F},\mathcal{F}} & Q_{\mathcal{F},\mathcal{T}} \\ Q_{\mathcal{T},\mathcal{F}} & Q_{\mathcal{T},\mathcal{T}} \end{pmatrix}.$$

As for DTMCs, it is possible to compute moments of first passage times from states in $\mathcal{F}$ to $\mathcal{T}$ using a recurrence. However, the recurrence is much simpler in this case:

$$\begin{aligned} -Q_{\mathcal{F},\mathcal{F}} m^{(i+1)} &= (i+1)m^{(i)} \text{ for } i \geq 0 \text{ with } m^{(0)} \\ &= u. \end{aligned} \quad (4)$$

One needs to factorize $-Q_{\mathcal{F},\mathcal{F}}$ once, compute $m^{(i+1)}$ using $m^{(i)}$ and the factors of $-Q_{\mathcal{F},\mathcal{F}}$. The factorization can be performed using Grassmann-Taksar-Heyman (GTH) as for DTMCs, but this time on the coefficient matrix

$$\begin{pmatrix} -Q_{\mathcal{F},\mathcal{F}} & -Q_{\mathcal{F},\mathcal{T}}u \\ w^T & -w^T u \end{pmatrix}$$

(see article **Computational Methods for DTMCs**).

TRANSIENT DISTRIBUTIONS OF CTMCs

Let $\pi_i(t)$ be the probability that a CTMC having transition-rate matrix Q is in state

i at time t . Then, the *transient probability vector* of the CTMC at time t is given by

$$\pi(t) = \pi(0)e^{Qt}.$$

Transient solution methods can be classified into two, based on computing e^{Qt} explicitly or implicitly. The uniformization method and Krylov subspace methods can follow either approach. Decompositional methods and matrix powering follow the former approach, whereas ordinary differential equation (ODE) solvers follow the latter approach. In the next section we start with uniformization, a very simple method suited for this purpose.

Uniformization

Consider the discretized stochastic transition probability matrix

$$P = I + \frac{1}{\Gamma}Q$$

corresponding to the continuous-time process governed by Q that is embedded in a Poisson process of rate $\Gamma = \max_i |q_{i,i}|$. Then the *uniformization equation* [1] is given by

$$e^{Qt} = P(t) = e^{-\Gamma t} \sum_{k=0}^{\infty} \frac{(\Gamma t)^k}{k!} P^k.$$

The right-hand side of this equation can be terminated, say, at $k = K$, and used to compute an approximation to e^{Qt} as in

$$\tilde{\pi}(t) = \pi(0) \sum_{k=0}^K \frac{(\Gamma t)^k}{k!} P^k.$$

Assuming that the order of P is n and it has nz nonzeros, the method is simple to program with two vectors of length n ; it executes about $K(n + nz)$ flops, and has a known truncation error given by

$$\sum_{k=0}^K \frac{(\Gamma t)^k}{k!} \geq \frac{1 - \epsilon}{e^{-\Gamma t}} \Rightarrow \|\pi(t) - \tilde{\pi}(t)\|_{\infty} \leq \epsilon.$$

Its disadvantage is the need to partition t into time steps and use a smaller ϵ for each subinterval when Γt is large.

Next, we consider a group of methods based on decomposing (i.e., factorizing) the coefficient matrix.

Matrix Decomposition

If Q is *nondefective* (meaning, its number of linearly independent eigenvectors is less than its number of eigenvalues), we may write $SQS^{-1} = \Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$, implying

$$e^{Qt} = S^{-1}e^{\Lambda t}S,$$

where $e^{\Lambda t} = \text{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t}\}$. Hence

$$\pi(t) = \pi(0)S^{-1}e^{\Lambda t}S = \sum_{i=1}^n \alpha_i e^{\lambda_i t} s_i,$$

where the linear coefficients α_i are to be computed from $\alpha S = \pi(0)$ and s_i is the i th row of S . Consequently, the transient solution may be computed very efficiently at any time instant.

When Q is (nearly) defective, one can reduce Q to (quasi-)triangular form $OQO^T = W$ using an orthogonal matrix O (through the *Schur decomposition*). Then

$$e^{Qt} = O^T e^{Wt} O.$$

If the QR algorithm of Francis [2,3] is used to compute the Schur decomposition, we have

$$RW = \Lambda R,$$

where R is a triangular matrix, Λ is a diagonal matrix with the eigenvalues, and V is a matrix of left-hand eigenvectors ($V = RO$). Since $Q = O^T R^{-1} \Lambda RO$, we have $\pi(t) = \pi(0)O^T R^{-1} e^{\Lambda t} RO$. Then one can solve $\alpha R = \pi(0)O^T$ for α and compute $\pi(t) = \alpha e^{\Lambda t} V$.

Another group of methods is based on finding a suitable power of two that speeds up the computation.

Matrix Powering

Assuming that we seek $\pi(t)$, let m be an integer and $t_0 \neq 0$ for which $t = 2^m t_0$. Now $t_j = 2t_{j-1}$ for $j \geq 1$, and this implies

$$e^{Qt_j} = P(t_j) = P(t_{j-1})P(t_{j-1}).$$

The idea is to compute $P(t_0)$, then to square it m times to obtain $P(t_m)$, and finally to premultiply $P(t_m)$ with $\pi(0)$.

The uniformization method can be used to compute $P(t_0)$ with a Horner-type scheme when n is small. The recommended value for $m = \log_2(t/t_0)$ due to Marie [4], is

$$m = \lfloor \log_2 [4(\zeta + 3)\Gamma] \rfloor$$

where ζ is the maximum number of nonzero elements in any row of P . The advantage of this method is that the transient solution is also available at intermediate times $t_0, 2t_0, \dots, 2^{m-1}t_0$. Its disadvantages are that the computational cost is proportional to $O(n^3)$ flops, and there is rounding error buildup when $m \gg 1$.

Another way to compute $P(t_0)$ is to use *diagonal Padé approximants* since they are more stable and of higher order than non-diagonal Padé approximants for the same amount of computation. The (p, p) diagonal Padé approximant to e^X , where $X = Qt_0$, is the unique rational function

$$R_{p,p}(X) = \frac{N_{p,p}(X)}{N_{p,p}(-X)}$$

which matches the Taylor series expansion of e^X through terms to the power $2p$. Its coefficients are determined by solving

$$N_{p,p}(X) = \sum_{j=0}^p c_j X^j, \text{ where } c_0 = 1, \\ c_j = c_{j-1} \frac{p+1-j}{j(2p+1-j)}.$$

Efficient Horner-type implementations are possible. Even values of p are better than odd values and $p = 6$ is generally satisfactory [5].

Ordinary Differential Equation Solvers

There are numerous possibilities for applying ODE procedures to compute transient solutions of Markov chains. An immediate advantage of such an approach is that, unlike uniformization and matrix scaling and powering, ODE solution methods are directly applicable to *nonhomogeneous* Markov

chains. Given a first-order differential equation $y' = f(t, y)$ and an initial condition $y(t_0) = y_0$, a solution is a differentiable function $y(t)$ such that

$$y(t_0) = y_0, \frac{d}{dt}y(t) = f(t, y(t)). \quad (5)$$

In the context of Markov chains, the solution $y(t)$ is the row vector $\pi(t)$ and the function $f(t, y(t))$ is simply $\pi(t)Q$.

Numerical procedures to compute the solution of Equation (2) attempt to follow a unique solution curve from its value at an initially specified point to its value at some other prescribed point τ . This usually involves a discretization procedure on the interval $[0, \tau]$ and the computation of approximations to the solution at the intermediate points. Given a discrete set of points (or *mesh*) $\{0 = t_0, t_1, t_2, \dots, t_\eta = \tau\}$ in $[0, \tau]$, we denote the exact solution of the differential Equation (2) at time t_i by $y(t_i)$. The *step size* or *panel width* at step i is defined as $h_i = t_i - t_{i-1}$. A numerical method generates a sequence $\{y_1, y_2, \dots, y_\eta\}$ such that y_i is an approximation to $y(t_i)$. The reader should be careful not to confuse y_i with $y(t_i)$. We use y_i for $i \neq 0$ to denote a *computed approximation* to the *exact value* $y(t_i)$. For $i = 0$, the initial condition gives $y_0 = y(t_0)$.

In computing y_{i+1} , a method may incorporate the values of previously computed approximations y_j for $j = 0, 1, \dots, i$, or even previous approximations to $y(t_{i+1})$. A method that uses only (t_i, y_i) to compute y_{i+1} is said to be an *explicit single-step method*. It is said to be a *multistep method* if it uses approximations at several previous steps to compute its new approximation. A method is said to be *implicit* if computation of y_{i+1} requires an approximation to $y(t_{i+1})$; otherwise, it is said to be *explicit*. We present two elementary methods to illustrate different possibilities.

If the solution is continuous and differentiable, then in a small neighborhood of the point (t, y_i) we can approximate the solution curve by its tangent y'_i at (t_i, y_i) and thereby move from (t_i, y_i) to the next point (t_{i+1}, y_{i+1}) . This method is called the *forward Euler method* (FEM). It is equivalent to a Taylor series expansion of order 1. In the

context of Markov chains, it becomes

$$\pi_{(i+1)} = \pi_{(i)} + h_{i+1}\pi_{(i)}Q. \quad (6)$$

Note that $\pi_{(i)}$ is the *state vector* of probabilities at time t_i . We use this notation, rather than π_i , so as not to confuse the i th component of the vector with the entire vector at time t_i . Thus, moving from one time step to the next is accomplished by a scalar–matrix product and a vector–matrix product.

The *modified Euler method* incorporates the average of the slopes at both points under the assumption that this will provide a better average approximation of the slope over the entire panel $[t_i, t_{i+1}]$. The formula is given by

$$y_{i+1} = y_i + h_{i+1} \frac{f(t_i, y_i) + f(t_{i+1}, y_{i+1})}{2}. \quad (7)$$

This is also referred to as the *trapezoid rule*. When applied to Markov chains, we have

$$\pi_{(i+1)} = \pi_{(i)} + \frac{h_{i+1}}{2} (\pi_{(i)}Q + \pi_{(i+1)}Q),$$

that is,

$$\pi_{(i+1)} \left(I - \frac{h_{i+1}}{2} Q \right) = \pi_{(i)} \left(I + \frac{h_{i+1}}{2} Q \right), \quad (8)$$

which requires in addition to the operations needed by the explicit Euler method, the solution of a system of equations at each step (plus a scalar–matrix product). These additional computations per step are offset to a certain extent by the better accuracy achieved with the trapezoid rule. If the size of the Markov chain is small and the step size is kept constant, the matrix

$$\left(I + \frac{h_{i+1}}{2} Q \right) \left(I - \frac{h_{i+1}}{2} Q \right)^{-1}$$

may be computed at the outset before beginning the stepping process, so that the computation per step required by the modified Euler method becomes identical to that of the explicit Euler. If the Markov chain is large, then the inverse should not be formed explicitly. Instead, an *LU* decomposition should be computed and the inverse replaced with

a backward and forward substitution process with U and L respectively. When the step size is not kept constant, each different value of h used requires that a system of linear equations be solved. Depending on the size and sparsity pattern of Q , the work required by the trapezoid rule to compute the solution to a specified precision may or may not be less than that required by explicit Euler. The trade-off is between an implicit method requiring more computation per step but fewer steps and an explicit method requiring more steps but less work per step! We chose these two simple methods just to illustrate the approach used to compute transient distributions of Markov chains. However, it is more common to use more efficient but more complex methods such as the single-step Runge–Kutta formula and multistep *backward differentiation formulae* (BDF) which we now briefly describe.

Runge–Kutta Methods. A Runge–Kutta algorithm of order p provides an accuracy comparable to a Taylor series algorithm of order p , but without the need to determine and evaluate the derivatives $f', f'', \dots, f^{(p-1)}$, requiring instead the evaluation of $f(t, y)$ at selected points. The derivation of an order p Runge–Kutta method is obtained from a comparison with the terms through h^p in the Taylor series method for the first step, that is, the computation of y_1 from the initial condition (t_0, y_0) . The most widely used Runge–Kutta methods are of order 4. When the standard explicit fourth-order Runge–Kutta method is applied to the Chapman–Kolmogorov equations $\pi' = \pi Q$, the sequence of operations to be performed to move from $\pi_{(i)}$ to the next time step $\pi_{(i+1)}$ is as follows:

$$\pi_{(i+1)} = \pi_{(i)} + h(k_1 + 2k_2 + 2k_3 + k_4)/6,$$

where

$$\begin{aligned} k_1 &= \pi_{(i)} Q, \\ k_2 &= (\pi_{(i)} + hk_1/2) Q, \\ k_3 &= (\pi_{(i)} + hk_2/2) Q, \\ k_4 &= (\pi_{(i)} + hk_3) Q. \end{aligned}$$

Multistep BDF Methods. An important characterization of initial-value ODEs has yet to be discussed: that of *stiffness*. Explicit methods have enormous difficulty solving stiff ODEs, indeed to such an extent that one definition of stiff equations is “problems for which explicit methods don’t work” [6]! Many factors contribute to stiffness, including the eigenvalues of the Jacobian $\partial f/\partial y$ and the length of the interval of integration. The problems stem from the fact that the solutions of stiff systems of differential equations contain rapidly decaying transient terms.

The so-called BDF (*BDF*) methods [7] are a class of multistep methods that are widely used for stiff systems. Only implicit versions are in current use, since low-order explicit versions correspond to the explicit Euler method ($k = 1$) and the midpoint rule ($k = 2$), and higher-order explicit versions ($k \geq 3$) are not stable. Implicit versions are constructed by generating an interpolating polynomial $z(t)$ through the points (t_j, y_j) for $j = i - k + 1, \dots, i + 1$. However, now y_{i+1} is determined so that $z(t)$ satisfies the differential equation at t_{i+1} , that is, in such a way that $z'(t_{i+1}) = f(t_{i+1}, y_{i+1})$. It is usual to express the interpolating polynomial in terms of backward differences

$$\nabla^0 f_i = f_i, \quad \nabla^{j+1} f_i = \nabla^j f_i - \nabla^j f_{i-1}.$$

With this notation, the general formula for implicit BDF methods is

$$\sum_{j=1}^k \frac{1}{j} \nabla^j y_{i+1} = h f_{i+1}$$

and the first three rules are

$$\begin{aligned} k = 1 : y_{i+1} - y_i &= h f_{i+1} \text{ (implicit Euler),} \\ k = 2 : 3y_{i+1}/2 - 2y_i + y_{i-1}/2 &= h f_{i+1}, \\ k = 3 : 11y_{i+1}/6 - 3y_i + 3y_{i-1}/2 - y_{i-2}/3 &= h f_{i+1}. \end{aligned}$$

The BDF formulae are known to be stable for $k \leq 6$ and to be unstable for other values of k . When applied to Markov chains, the implicit ($k = 2$) BDF formula is

$$3\pi_{(i+1)}/2 - 2\pi_{(i)} + \pi_{(i-1)}/2 = h\pi_{(i+1)} Q,$$

so that the system of equations to be solved at each step is

$$\pi_{(i+1)}(3I/2 - hQ) = 2\pi_{(i)} - \pi_{(i-1)}/2. \quad (9)$$

The solutions at two prior points $\pi_{(i-1)}$ and $\pi_{(i)}$ are used in the computation of $\pi_{(i+1)}$. To get the procedure started, the initial starting point $\pi_{(0)}$ is used to generate $\pi_{(1)}$ using the trapezoid rule, and then both $\pi_{(0)}$ and $\pi_{(1)}$ are used to generate $\pi_{(2)}$, using Equation (9).

EXAMPLE

Let us consider the CTMC whose transition-rate matrix is given by

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

and investigate the outcome of applying the methods of matrix decomposition and explicit Runge–Kutta of order 4 with the initial probability vector $\pi(0) = (10)$ for time $t = 1$.

The generator matrix has the decomposition $Q = S^{-1}\Lambda S$, where

$$S = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}.$$

Since $\pi(t) = \pi(0)S^{-1}e^{\Lambda t}S$, we have $\pi(1) = (0.5677, 0.4323)$ in four decimal digits of precision.

As for the Runge–Kutta method of order 4 with the step size $h = 0.25$, for four steps we obtain in four decimal digits of precision $\pi_1 = (0.8034, 0.1966)$, $\pi_2 = (0.6481, 0.3519)$, $\pi_3 = (0.6117, 0.3883)$, and $\pi_4 = (0.5678, 0.4322)$.

REFERENCES

1. Sidje RB, Stewart WJ. A numerical study of large sparse matrix exponentials arising in Markov chains. *Comput Stat Data Anal* 1999;29,(3):345–368. DOI: 10.1016/S0167-9473(98)00062-0.
2. Francis JGF. The QR transformation: a unitary analogue to the LR transformation – Part 1. *Comput J* 1961;4,(3):265–271. DOI: 10.1093/comjnl/4.3.265.
3. Francis JGF. The QR transformation: a unitary analogue to the LR transformation – Part 2. *Comput J* 1962;4,(4):332–345. DOI: 10.1093/comjnl/4.4.332.
4. Marie RA. Transient numerical solutions of stiff Markov chains. 20th International Symposium on Automotive Technology. Florence, Italy; 1989.
5. Philippe B, Sidje RB. Transient solutions of Markov processes by Krylov subspaces. In: Stewart WJ, editor. *Proceedings of the 2nd International Workshop on the Numerical Solution of Markov chains*. Norwell (MA): Kluwer Academic Publishers; 1995. pp. 95–119.
6. Hairer E, Wanner G. Solving ordinary differential equations II: stiff and differential-algebraic problems. *Series in Computational Mathematics*. Berlin: Springer-Verlag; 1991.
7. Gear CW. Numerical initial value problems in ordinary differential equations. Englewood Cliffs (NJ): Prentice-Hall; 1971.

FURTHER READING

Stewart WJ. *Introduction to the numerical solution of Markov chains*. Princeton (NJ): Princeton University Press; 1994.

COMPUTATIONAL METHODS FOR DTMCs

TUĞRUL DAYAR

Department of Computer Engineering,
Bilkent University, Ankara, Turkey

WILLIAM J. STEWART

Department of Computer Science, North
Carolina State University, Raleigh,
North Carolina

A *discrete-time Markov chain* (DTMC) $\{X_n, n = 0, 1, 2, \dots\}$ is a stochastic process, such that for all natural numbers n and all states $i, j, k, l, \dots$,

$$\text{Prob}\{X_{n+1} = j | X_n = i, X_{n-1} = k, \dots, X_0 = l\} \\ = \text{Prob}\{X_{n+1} = j | X_n = i\}.$$

Thus the state in which the system finds itself at time step $n + 1$ depends only on where it is at time step n . The conditional probabilities, $\text{Prob}\{X_{n+1} = j | X_n = i\}$, are called the *single-step transition probabilities*, or just the transition probabilities, of the DTMC. They give the conditional probability of making a transition from state i to state j when the time parameter increases from n to $n + 1$. They are denoted by

$$p_{ij}(n) = \text{Prob}\{X_{n+1} = j | X_n = i\}.$$

The matrix $P(n)$, formed by placing $p_{i,j}(n)$ in row i and column j , for all i and j , is called the *transition probability matrix*. It follows that a DTMC is completely characterized by its transition probability matrix $P(n)$ and a probability vector $\pi(0)$ whose i th element gives the probability of the chain beginning at time step 0 in state i . The DTMC is said to be (*time*) *homogeneous* if for all states i and j ,

$$\text{Prob}\{X_{n+1} = j | X_n = i\} = \text{Prob}\{X_{n+m+1} = j | \\ X_{n+m} = i\},$$

for $n = 0, 1, 2, \dots$ and $m \geq 0$. For a homogeneous DTMC,

$$p_{ij} = \text{Prob}\{X_1 = j | X_0 = i\} = \text{Prob}\{X_2 = j | \\ X_1 = i\} = \text{Prob}\{X_3 = j | X_2 = i\} = \dots$$

and we can replace $P(n)$ with P , since transitions no longer depend on n . In this article, we consider only homogeneous DTMCs.

After n steps, the probability distribution of a DTMC with transition probability matrix P and initial probability distribution $\pi(0)$ is given by

$$\pi(n) = \pi(n-1)P = \pi(n-2)P^2 = \dots = \pi(0)P^n.$$

These distributions are generally referred to as *transient* distributions and their computation rarely poses major problems. The procedure consists of repeatedly multiplying the probability distribution vector obtained at step $(k-1)$ with the stochastic transition probability matrix to obtain the probability distribution at step k , for $k = 1, 2, \dots, n$. If n is large and the number of states in the DTMC is small (not exceeding several hundreds), then some savings in computation time can be obtained by successively squaring the transition probability matrix j times, where j is the largest integer such that $2^j \leq n$. This gives the matrix P^{2^j} which can now be multiplied by P (and powers of P) until the value P^n is obtained. The distribution at time step n is now found from $\pi(n) = \pi(0)P^n$. If the matrix P is sparse, this sparsity is lost in the computation of P^n , so this approach is not appropriate for large sparse DTMCs. Also, if a time trajectory of a statistic of the distribution is needed, this approach may be less than satisfactory, because only distributions at computed values of P^{2^j} will be available. One final point worth noting is that for large values of n , it may be beneficial to maintain a check on successive distributions since they may converge to some desired computational accuracy, prior to step n . Any additional vector-matrix multiplications after this point will not alter the distribution.

If $\lim_{n \rightarrow \infty} P^n$ exists, then $\lim_{n \rightarrow \infty} \pi(n) = \pi(0) \lim_{n \rightarrow \infty} P^n$ exists and is called the

limiting distribution. Not all DTMCs possess a limiting distribution: a periodic DTMC (upon exiting any state, a return to that state can occur only in some multiple of $c > 1$ steps) does not possess a limiting distribution. A limiting distribution which converges, independent of the initial starting distribution $\pi(0)$, to a vector whose components are strictly positive and sum to 1, is called a *steady-state* distribution. Thus, a steady-state distribution is the unique vector π that satisfies

$$\begin{aligned}\pi &= \pi(0) \lim_{n \rightarrow \infty} P^{(n)} = \pi(0) \lim_{n \rightarrow \infty} P^{(n+1)} \\ &= \left[\pi(0) \lim_{n \rightarrow \infty} P^{(n)} \right] P = \pi P,\end{aligned}$$

that is, $\pi = \pi P$. Any probability vector z (the components are probabilities and sum to 1) such that $z = zP$ is called a *stationary* distribution. Some DTMCs may contain multiple stationary distributions while others may have none. When a steady-state distribution π exists, it is the unique stationary distribution. Unless explicitly stated otherwise, in this article we consider *ergodic* DTMCs which have a (unique) steady-state distribution and, for the most part, our objective is to compute this distribution from the defining relationship, $\pi = \pi P$.

We distinguish between solution methods that are *iterative* and solution methods that are *direct*. Iterative methods begin with an initial approximation (or guess) to the solution vector and proceed to modify this approximation in such a way that, at each step of iteration, it becomes closer and closer to the true solution. On the other hand, a direct method attempts to go straight to the final solution. A certain number of well defined steps must be taken, at the end of which the solution has been computed.

Iterative methods of one type or another are by far the most commonly used methods for solving DTMCs. There are several important reasons for this choice. First, an examination of the standard iterative methods shows that the only operation in which the matrices are involved is their multiplication with one or more vectors or with preconditioners—operation which leaves the transition matrices unaltered. Thus, compact

storage schemes, which minimize the amount of memory required to store the matrix and which, in addition, are well suited to matrix multiplication, may be conveniently implemented. Since the matrices involved are usually large and very sparse, the savings made by such schemes can be considerable. With direct equation-solving methods, the elimination of one nonzero element of the matrix during the reduction phase often results in the creation of several nonzero elements in positions which previously contained zeroes. This is called *fill-in*, and not only does it make the organization of a compact storage scheme more difficult, since provision must be made for the deletion and the insertion of elements, but, in addition, the amount of fill-in can often be so extensive that available memory can be exhausted.

Iterative methods have other advantages. Use may be made of good initial approximations to the solution vector, and this is especially beneficial when a series of related experiments are being conducted. Also, an iterative process may be halted once a pre-specified tolerance criterion has been satisfied, and this may be relatively lax. In contrast, a direct method must continue until the final specified operation has been carried out. And lastly, with iterative methods, the matrix is never altered and hence the build-up of rounding error is, to all intents and purposes, nonexistent. For these reasons, iterative methods have traditionally been preferred to direct methods. However, iterative methods have a major disadvantage in that often they require a very long time to converge to the desired solution. Direct methods have the advantage that an upper bound on the time required to obtain the solution may be determined before the computation is initiated. More important, for certain classes of problems, direct methods can result in a much more accurate answer being obtained in less time. Since iterative methods will in general require less memory than direct methods, these latter can only be recommended if they obtain the solution in less time.

The most suitable candidates for solution by direct methods are DTMCs whose transition matrices are small, of the order of a few

hundred, or when they are *banded*, that is, the only nonzero elements of the coefficient matrix are not too far from the diagonal. In this latter case, it means that an ordering can be imposed on the states so that no single-step transition from state i will take it to states numbered greater than $i + \delta$ or less than $i - \delta$. All fill-in will occur within a distance δ of the diagonal, and the amount of computation per step is proportional to δ^2 . We begin with direct methods based on Gaussian elimination.

DIRECT METHODS FOR SOLVING DTMCs

Gaussian elimination (GE) is a direct method for solving systems of linear equations which are usually written in the form $Ax = b$, where A is an $(n \times n)$ coefficient matrix, x the unknown column vector of length n which is to be found, and b the right-hand side column vector of length n . In our case, we seek to determine π from the system of equations $\pi = \pi P$, that is, from

$$\pi(P - I) = 0, \quad (1)$$

where I is the identity matrix. In the standard equation-solving terminology, $A = P^T - I$, $x = \pi^T$, and $b = 0$. The system of equations (Eq. 1) has a solution other than the trivial solution ($\pi_i = 0$, for all i) if and only if the coefficient matrix is singular. Since the determinant of a matrix is equal to the product of its eigenvalues and since P possesses an eigenvalue equal to 1, the singularity of A and hence the existence of a nontrivial solution follows. In passing, we remark that in the DTMC context, A is a negated *M-matrix*.

GE is composed of two phases, a *reduction* phase during which the coefficient matrix is brought to upper triangular form, and a *backsubstitution* phase which generates the solution from the reduced coefficient matrix. The first (reduction) phase is the computationally expensive part of the algorithm, having an $O(n^3)$ operation count when the matrix is full. The backsubstitution phase has order $O(n^2)$. The first step in GE is to use one of the equations to eliminate one of the unknowns in the other $(n - 1)$ equations.

This is accomplished by adding a multiple of one row to the other rows; the particular multiple is chosen to zero out the coefficient of the unknown to be eliminated. The particular equation chosen is called the *pivotal equation* and the diagonal element in this equation is called the *pivot*. The equations that are modified are said to be *reduced*. The operations of multiplying one row by a scalar and adding or subtracting two rows are *elementary operations*; they leave the system of equations invariant. When the first step of GE is finished, one equation involves all n unknowns while the other $(n - 1)$ equations involve only $(n - 1)$ unknowns. These $(n - 1)$ equations may be treated independent of the first. They constitute a system of $(n - 1)$ linear equations in $(n - 1)$ unknowns, which we may now solve using GE and thus the process continues. Initially, $1 - p_{i,i} = \sum_{j \neq i} p_{i,j}$ and this property is maintained during the reduction phase. Therefore, at each step of the reduction, the pivotal elements are the largest in each column, and the multipliers do not exceed 1. Explicit pivoting, which is used in general systems of linear equations to ensure that multipliers do not exceed 1, is generally not needed for solving DTMC problems.

At the end of the reduction phase, we would hope to end up with one equation in one unknown, two equations in two unknowns, and so on, up to n equations in n unknowns. However, A has rank $(n - 1)$, since we assume that the DTMC is ergodic, and the last equation in the reduced system must evaluate to $0 = 0$. Another way to look at this situation is to observe that the system of equations $Ax = 0$ does not tell the whole story. We also know that $e^T x = 1$, where e is a column vector containing n 1's. The n equations of $Ax = 0$ provide only $(n - 1)$ linearly independent equations, but together with $e^T x = 1$, we have a complete basis set. For example, it is possible to replace the last equation of the original system with $e^T x = 1$, which eliminates the need for any further normalization. In this case the coefficient matrix becomes nonsingular, the right-hand side becomes nonzero, and a unique solution is computed. Of course, it is not necessary to replace the last equation of the system by this normalization equation. Indeed, any

equation could be replaced. However, this is generally undesirable, for it will entail more numerical computation. For example, if the first equation is replaced, the first row of the coefficient matrix will contain all 1's and the right-hand side will be $e_1 = (1, 0, \dots, 0)^T$. The first consequence of this is that during the reduction phase, the entire sequence of elementary row operations must be performed on the right-hand side vector, e_1 , whereas if the last equation is replaced, the right-hand side is unaffected by the elementary row operations. The second and more damaging consequence is that substantial fill-in will occur since a multiple of the first row, which contains all 1's, will be added to higher numbered rows and a cascading effect will undoubtedly occur in all subsequent reduction steps. An equally viable alternative to replacing the last equation with $e^T x = 1$ is to set the last component x_n to 1 and leave a final normalization until later.

Once the reduction phase of GE is finished, we proceed to the backsubstitution phase, during which the values of the unknowns are computed. Given one equation in one unknown (the last equation in the reduced system), the value of that unknown is easily computed. Since we now know x_n , the value of x_{n-1} can be found from the penultimate equation in the reduced system. The process of back substitution continues in this manner until all the unknowns have been evaluated.

Connection to LU Factorization

When the coefficient matrix in a system of linear equations $Ax = b$ can be written as the product of a lower triangular matrix L and an upper triangular matrix U , then

$$Ax = LUx = b$$

and the solution can be found by first solving (by forward substitution) $Lz = b$ for an intermediate vector z and then solving (by back substitution) $Ux = z$ for the solution x . The upper triangular matrix obtained by the reduction phase of GE provides an upper triangular matrix U to go along with a lower triangular matrix L whose diagonal elements are all equal to 1 and whose subdiagonal

elements are the multipliers with a minus sign in front of them.

In the DTMC context, the system of equations is homogeneous and the coefficient matrix is singular,

$$(P^T - I)x = (LU)x = 0.$$

If we now set $Ux = z$ and attempt to solve $Lz = 0$, we find that, since L is nonsingular (it is a triangular matrix whose diagonal elements are all equal to 1), we must have $z = 0$. This means that we may proceed directly to the back substitution on $Ux = z = 0$ with $u_{nn} = 0$. It is evident that we may assign any nonzero value to x_n , say $x_n = \eta$, and then determine, by simple back substitution, the remaining elements of the vector x in terms of η . Normalizing the solution obtained from solving $Ux = 0$ yields the desired unique solution vector π . This approach is referred to as an *LU decomposition* or *LU factorization*.

Grassmann–Taksar–Heyman Idea

It is appropriate at this point to mention a version of GE that has attributes that appear to make it even more stable than the usual version. This procedure is commonly referred to as the *Grassmann–Taksar–Heyman* (GTH) algorithm [1,2]. In the GTH idea, the diagonal elements are obtained by summing off-diagonal elements rather than performing subtractions: it is known that subtractions can sometimes lead to loss of significance in numerical computations. These subtractions occur in forming the diagonal elements during the reduction process. Happily, it turns out that at the end of each reduction step, the unreduced portion of the matrix is the transpose of a submatrix with nonnegative off-diagonal elements and the diagonal elements can be formed as the negated sums of these off-diagonal elements. This has a probabilistic interpretation based on the restriction of the DTMC to a reduced set of states and it is in this context that the GTH algorithm is generally developed. It also means that the diagonal elements may be formed by adding off-diagonal elements and placing a minus sign in front of this sum instead of performing a single subtraction.

Moments of First Passage Times

Before terminating this section on direct methods for finding steady-state distributions, we consider the problem of computing moments of first passage times from a state to a set of states in DTMCs. Without loss of generality, let us assume that the state space of P is partitioned as $\mathcal{S} = \mathcal{F} \cup \mathcal{T}$, $\mathcal{F} \cap \mathcal{T} = \emptyset$, yielding the block partitioning

$$P = \begin{pmatrix} P_{\mathcal{F},\mathcal{F}} & P_{\mathcal{F},\mathcal{T}} \\ P_{\mathcal{T},\mathcal{F}} & P_{\mathcal{T},\mathcal{T}} \end{pmatrix}$$

Now, let $\mathcal{F}$ have $n_{\mathcal{F}}$ states and $\mathcal{T}$ have $n_{\mathcal{T}}$ states so that $n = n_{\mathcal{F}} + n_{\mathcal{T}}$.

In Ref. 3, an algorithm is given to compute *moments of first passage times* from states in $\mathcal{F}$ to $\mathcal{T}$ using the recurrence

$$Fm^{(i+1)} = \sum_{j=0}^i (-1)^{i-j} \binom{i+1}{j} m^{(j)} \quad \text{for } i \geq 0 \text{ with } m^{(0)} = e, \quad (2)$$

where $F = I - P_{\mathcal{F},\mathcal{F}}$, $m_f^{(j)}$ is the j th moment of first passage time from state $f \in \mathcal{F}$ to $\mathcal{T}$, and $\binom{i+1}{j}$ denotes the *binomial coefficient* $(i+1) \text{ choose } j$. The integer coefficients of the linear combination on the right-hand side of Equation (2) may easily be computed by observing the alternating signs and using the identity $\binom{i+1}{j} = \binom{i}{j-1} + \binom{i}{j}$.

In order to carry out the moment computation accurately, a nonhomogeneous but consistent linear system $Ax = b$, where

$$B = \begin{pmatrix} I - P_{\mathcal{F},\mathcal{F}} & -P_{\mathcal{F},\mathcal{T}}e \\ w^T & -w^T e \end{pmatrix}, \quad x = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} c \\ v \end{pmatrix},$$

is considered. Observe that B is a singular M-matrix with zero row sums and has rank $n_{\mathcal{F}}$. Hence, $F = I - P_{\mathcal{F},\mathcal{F}}$ can be LU factored in $(n_{\mathcal{F}} - 1)$ reduction steps using the GTH idea, thus making the values of the vector w and the scalar v immaterial. Then y can be obtained from $LUy = c$ through forward and back substitutions. Note that both L and U need to be stored since Equation (2) will be solved for multiple right-hand sides.

The reason behind using the GTH idea is to improve the accuracy in the computed factors when F is close to being singular. Other than providing better roundoff properties, this modification has the additional advantage of being carried out completely in row-wise sparse format with delayed row updates. Otherwise, one could employ GE to factorize F , but in no way should it be inverted and passed to the right-hand side in Equation (2).

ITERATIVE METHODS FOR SOLVING DTMCs

Up until this section, direct methods which execute for a predetermined number of steps are discussed. For large problems, direct methods become inefficient, because they introduce new nonzero elements during factorization. The class of methods introduced here are iterative in that they begin from some initial approximation and compute a new approximation at each iteration while maintaining the nonzero structure of the coefficient matrix, with the expectation that the approximation converges to the steady-state vector. Since the number of iterations required to reach a predetermined accuracy cannot be forecast inexpensively other than in some special cases, rules of thumb regarding the performance of iterative methods were investigated in a sequence of papers [4–7].

There are different classes of iterative methods. The taxonomy considered here consists of those methods that are based on splitting the coefficient matrix, decomposing the coefficient matrix, and using Krylov subspaces. At each iteration, each of these class of methods performs some number of matrix-vector multiplications and in certain cases the direct/iterative solution of linear systems, which are sometimes smaller than the original system of equations.

Methods Based on Splittings

Methods based on splittings are stationary iterative methods in which the singular coefficient matrix $A = P^T - I$ of the homogeneous linear system $Ax = 0$ to be solved for $x = \pi^T$ is written as the difference of two matrices,

the first being nonsingular. They include the *power* (POWER) method, *(block) Jacobi over-relaxation* ((B)JOR), and *(block) successive overrelaxation* ((B)SOR). We remark that the POWER method has been used extensively to compute the eigenvector corresponding to the dominant eigenvalue of a matrix, and can also be viewed as a relaxation of the Richardson type.

Let us consider the block partitioning of A and x given, respectively, by

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,K} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K,1} & A_{K,2} & \cdots & A_{K,K} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix}$$

for some number of blocks $K \in \{1, 2, \dots, n\}$, and write A as the sum of three terms as in

$$A = D_A - L_A - U_A,$$

where D_A is the block diagonal of A , $-L_A$ is its strictly block lower triangular part, and $-U_A$ is its strictly block upper triangular part. In other words,

$$D_A = \begin{pmatrix} A_{1,1} & 0 & \cdots & 0 \\ 0 & A_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{K,K} \end{pmatrix},$$

$$L_A = - \begin{pmatrix} 0 & 0 & \cdots & 0 \\ A_{2,1} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ A_{K,1} & \cdots & A_{K,K-1} & 0 \end{pmatrix},$$

$$U_A = - \begin{pmatrix} 0 & A_{1,2} & \cdots & A_{1,K} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & A_{K-1,K} \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

If A is irreducible, each of the terms D_A , U_A , and L_A is nonpositive, and $a_{s,s} < 0$ for all $s \in \mathcal{S}$, implying the existence of D_A^{-1} and $(D_A - L_A)^{-1}$.

Now, let A be split as

$$A = M_A - N_A,$$

and let us define the iteration matrices

$$T_A = M_A^{-1}N_A$$

of the POWER, (B)JOR, and (B)SOR methods to compute the sequence of approximations

$$x^{(m+1)} = T_A x^{(m)} \quad \text{for } m = 0, 1, \dots$$

The particular splittings corresponding to the three methods are

$$\begin{aligned} M_A^{\text{POWER}} &= -I, \quad N_A^{\text{POWER}} = -(I + A), \\ M_A^{(\text{B})\text{JOR}} &= D_A/\omega, \quad N_A^{(\text{B})\text{JOR}} \\ &= (1 - \omega)D_A/\omega + L_A + U_A, \\ M_A^{(\text{B})\text{SOR}} &= D_A/\omega - L_A, \quad N_A^{(\text{B})\text{SOR}} \\ &= (1 - \omega)D_A/\omega + U_A, \end{aligned}$$

where $\omega \in (0, 2)$ is the relaxation parameter of (B)JOR and (B)SOR. The BJOR and BSOR splittings reduce to block Jacobi (BJ) and block Gauss–Seidel (BGS) splittings for $\omega = 1$, and to point JOR and point SOR for $K = n$ (i.e., when the number of blocks is equal to the number of unknowns). Hence, the iteration matrices corresponding to the three splittings are

$$\begin{aligned} T_A^{\text{POWER}} &= I + A, \\ T_A^{(\text{B})\text{JOR}} &= (1 - \omega)I + \omega D_A^{-1}(L_A + U_A) \quad \text{and} \\ T_A^{(\text{B})\text{SOR}} &= (D_A/\omega - L_A)^{-1}((1 - \omega)D_A/\omega + U_A). \end{aligned}$$

Observe that the relation $T_A x = x$ holds (i.e., x is a fixed point of T_A) for the POWER, (B)JOR, and (B)SOR methods.

If $\omega \in (0, 1)$, then the iteration matrix T_A is nonnegative, irreducible, primitive, and has a spectral radius and an eigenvalue of 1; hence, convergence is guaranteed. For a sequence of converging approximations, one needs to ensure for the initial approximation that $x^{(0)} \notin \text{Range}(I - T_A)$; otherwise, there will be no improvement.

It is well known that the asymptotic rate of convergence of the methods under consideration depends on the magnitude of the subdominant eigenvalue of T_A . The smaller its value, the faster the method

converges. Regarding nonasymptotic convergence behavior, it can be said that there will be improvement in the current approximate solution vector $x^{(m)}$ compared to the initial approximate solution vector $x^{(0)}$ if T_A^m is positive or has a (an almost) positive row or column [8, p. 1041]. The coefficient matrix A is almost always sparse and the iteration matrices associated with the POWER and JOR methods have the same off-diagonal nonzero structure as that of A . Hence, compared to POWER and JOR, SOR has a higher chance of satisfying the conditions for improvement, since its iteration matrix is likely to have a larger number of nonzeros. In general, block iterative methods require more computation per iteration than point iterative methods based on splittings, but this is offset by a faster rate of convergence.

Regarding implementation, it is useful to consider the formulation of the POWER method in

$$x^{(m+1)} = P^T x^{(m)} \quad \text{for } m = 0, 1, \dots,$$

whereas those of (B)JOR and (B)SOR methods, respectively, in

$$\begin{aligned} A_{k,k}x_k^{(m+1)} &= (1 - \omega)A_{k,k}x_k^{(m)} \\ &\quad - \omega \left(\sum_{j=1}^{k-1} A_{k,j}x_j^{(m)} + \sum_{j=k+1}^K A_{k,j}x_j^{(m)} \right) \\ &\quad \text{for } k = 1, 2, \dots, K \end{aligned}$$

and

$$\begin{aligned} A_{k,k}x_k^{(m+1)} &= (1 - \omega)A_{k,k}x_k^{(m)} \\ &\quad - \omega \left(\sum_{j=1}^{k-1} A_{k,j}x_j^{(m+1)} + \sum_{j=k+1}^K A_{k,j}x_j^{(m)} \right) \\ &\quad \text{for } k = 1, 2, \dots, K \end{aligned}$$

for $m = 0, 1, \dots$. Note that the POWER method amounts to a matrix-vector multiplication, whereas (B)JOR and (B)SOR amount to (sub)matrix-(sub)vector multiplications, some vector operations, and, when $K \neq n$, solutions of nonhomogenous linear systems with nonsingular coefficient matrices $A_{k,k}$ for $k = 1, 2, \dots, K$.

When information about a suitable starting point is not known, the initial approximation is chosen as the uniform distribution. The solution vector may be normalized every few iterations to limit the effect of underflow and overflow, and to a certain extent control any irregular convergence behavior. The convergence criteria to be used in these iterative solvers could be, for instance,

$$\text{stop if } m \geq \text{maxit} \text{ or } \|r^{(m)}\|_\infty \leq \text{stop_tol}.$$

Here m is the iteration number, $r^{(m)}$ is the residual vector at iteration m , maxit is the maximum number of iterations the method is permitted to perform, and stop_tol is the user-specified stopping tolerance, which should be less than 1 and greater than machine epsilon. Setting stop_tol to a particular value means that the entries of A (our right-hand side is 0) are considered to have errors in the range $\pm \text{stop_tol} \|A\|$. Note that the computation of $r^{(m)}$ requires a matrix-vector multiplication and should not be performed at each iteration; otherwise, the work done per iteration would double.

In passing to iterative methods based on decomposing the coefficient matrix, we remark that in this section we considered *forward* versions of iterative methods based on splittings. In other words, in a given iteration the first unknown to be updated corresponded to the first state of the underlying DTMC, the next unknown to the second state, and so on. However, it is also possible to update the unknowns in reverse order, starting from the last one and moving toward the first state. This approach is referred to as *backward*. Although it would not make any difference on the magnitude of the subdominant eigenvalue of T_A for POWER and (B)JOR methods, the magnitude of the subdominant eigenvalue of T_A for (B)SOR changes when the states of the underlying DTMC are *renumbered*. This situation gives rise to the idea of using forward and backward versions of (B)SOR simultaneously, and thus to the concept of *symmetric* B(SOR). The effect of renumbering the states of a DTMC (i.e., *symmetrically permuting* A) on convergence is briefly investigated for the GS iteration in Ref. 9 with the result that the

states of the underlying DTMC can always be renumbered so as to ensure convergence.

Decompositional Methods

In the class of decompositional iterative methods for DTMCs, we primarily consider *iterative aggregation-disaggregation* (IAD), which is based on partitioning the state space $\mathcal{S}$ of the DTMC into mutually exclusive subsets $\mathcal{S}_k$ for $k = 1, 2, \dots, K$ (i.e., $\cup_{k=1}^K \mathcal{S}_k = \mathcal{S}$ and $\mathcal{S}_k \cap \mathcal{S}_l = \emptyset$ for $k \neq l$). Thus, P can be symmetrically permuted and partitioned into $(K \times K)$ blocks as in

$$P = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,K} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ P_{K,1} & P_{K,2} & \cdots & P_{K,K} \end{pmatrix}$$

at the outset of IAD. At each iteration, an aggregated matrix is computed and solved, and this step is preceded and/or succeeded by some number of iterations using a method based on splittings. IAD methods are intimately related to *multigrid* methods [10] and are especially useful when dealing with ill-conditioned systems. There are also *Schwarz* methods, which can be considered as a generalization of block iterative methods based on splittings in which the partitioning of the state space $\mathcal{S}$ into K subsets has overlaps (i.e., $\cup_{k=1}^K \mathcal{S}_k = \mathcal{S}$ and $\mathcal{S}_k \cap \mathcal{S}_l \neq \emptyset$ for $k \neq l$). Their additive versions [11] become BJOR and their multiplicative versions become BSOR when the overlaps are removed. These tend to accelerate the convergence of the corresponding block iterative methods, the amount of acceleration depending on the amount of overlap [12].

Iterative Aggregation-Disaggregation. IAD methods, such as the Koury-McAllister-Stewart (KMS) method [13], are suitable for DTMCs that can be symmetrically permuted to the block form

$$P = \text{diag}(P_{1,1}, P_{2,2}, \dots, P_{K,K}) + E,$$

when $\|E\|_\infty$, the *degree of coupling*, is relatively small compared to 1. In other words, for the specified partitioning, there are strong

interactions among the states of the disjoint subsets but weak interactions among the subsets themselves. Such DTMCs are called *nearly completely decomposable* (NCD) and are known to be ill conditioned due to the existence of $(K - 1)$ eigenvalues close to 1. The smaller the degree of coupling, the closer these other eigenvalues to 1, and the more difficult it is to solve the problem with iterative methods based on splittings.

Once again, letting $A = P^T - I$ and partitioning $x = \pi^T$ conformally with P , the transpose of the $(K \times K)$ matrix H whose (i, j) th element is given by

$$h_{i,j} = e^T A_{i,j} \phi_j,$$

where

$$\phi_j = x_j / \|x_j\|_1$$

is referred to as the *exactly aggregated matrix* corresponding to A . Note that one needs to know the exact solution to be able to compute the exactly aggregated matrix, which suggests approximate aggregation, hence iteration.

Now, let $\xi = (\xi_1, \xi_2, \dots, \xi_K)^T$ be the unique positive vector satisfying $H\xi = 0$, $\sum_{i=1}^K \xi_i = 1$. Then the IAD method with a BGS *disaggregation* step (IADBGS) [5] is equivalent to the iterative formula

$$x^{(m+1)} = (D_A - L_A)^{-1} U_A \Sigma^{(m)} x^{(m)} \quad \text{for } m = 0, 1, \dots,$$

where

$$\begin{aligned} h_{i,j}^{(m)} &= e^T A_{i,j} \phi_j^{(m)}, \quad \phi_j^{(m)} = x_j^{(m)} / \|x_j^{(m)}\|_1, \\ H^{(m)} \xi^{(m)} &= 0, \quad \xi^{(m)} > 0, \quad \sum_{i=1}^K \xi_i^{(m)} = 1, \\ \Sigma^{(m)} &= \text{diag} \left(\xi_1^{(m)} I / \|x_1^{(m)}\|, \xi_2^{(m)} I / \|x_2^{(m)}\|, \right. \\ &\quad \left. \dots, \xi_K^{(m)} I / \|x_K^{(m)}\| \right). \end{aligned}$$

Writing this in detail, one obtains

$$\begin{aligned} A_{i,i} x_i^{(m+1)} &= - \sum_{j=1}^{i-1} A_{i,j} x_j^{(m+1)} - \sum_{j=i+1}^K A_{i,j} x_j^{(m+1)} \\ &\quad \text{for } i = 1, 2, \dots, K, \end{aligned}$$

where

$$z^{(m+1)} = \begin{pmatrix} \xi_1^{(m)} \left(\phi_1^{(m)} \right)^T, \xi_2^{(m)} \left(\phi_2^{(m)} \right)^T, \\ \dots, \xi_K^{(m)} \left(\phi_K^{(m)} \right)^T \end{pmatrix}^T.$$

A similar formula may be written for BJ, and hence, for BJOR and BSOR disaggregation steps.

For NCD MCs the convergence of the IADBGs method will be rapid. Its asymptotic convergence analysis reveals that the error in the approximate solution is reduced by a factor of order $\|E\|_\infty$ at each iteration if the aggregated matrix and each diagonal block are solved exactly [14], for instance, with GE. When GE suffers from instability on aggregated matrices in the presence of rounding errors, one can use GTH [15]. The study of local and global convergence of IAD for arbitrary block partitionings (with the possibility of using iterative solution methods for the aggregation step) appears in Refs 16–18.

Krylov Subspace Projection Methods

Krylov subspace projection methods are non stationary iterative methods in which approximate solutions satisfying various constraints are extracted from small dimensional subspaces. The methods differ among each other in the way subspaces are selected and solution approximations extracted from them [4]. Being iterative, their basic operation is matrix-vector multiplication. Compared to other classes of iterative methods, they require a larger number of work vectors of size equal to the size of the state space, n . But, more importantly, they need to be used with preconditioners to result in effective solvers. The main idea behind preconditioning is to transform the linear system so that the difference between the dominant and the subdominant eigenvalues of the preconditioned coefficient matrix is larger than what it used to be in the original system [19].

A *projection step* is formally defined with a subspace $\mathcal{K}$ of dimension m from which the approximation is to be selected and another subspace $\mathcal{L}$ (of the same size m) that is used to set the constraints necessary to extract

the new approximate solution vector from $\mathcal{K}$. For practical purposes, one should have $m \ll n$. Now, let $V = [v_1, v_2, \dots, v_m]$ and $W = [\omega_1, \omega_2, \dots, \omega_m]$, respectively, be the bases of $\mathcal{K}$ and $\mathcal{L}$. Then taking the linear system $Ax = 0$ view into consideration, it is possible to express the approximate solution as $\tilde{x} = Vy$, where y is a vector of length m . This provides m degrees of freedom, and in order to extract a unique y the residual vector $-A\tilde{x}$ is required to be orthogonal to $\mathcal{L}$; that is,

$$-AVy \perp \omega_i, \quad i = 1, 2, \dots, m,$$

in matrix form,

$$-W^T AVy = 0.$$

If $x^{(0)}$ is an initial approximate solution to the system, then $x^{(0)}$ may be adjusted by a vector δ such that $(x^{(0)} + \delta)$ is a solution, that is, $A(x^{(0)} + \delta) = 0$. If we let $r_0 = -Ax^{(0)}$, then

$$\begin{aligned} A(x^{(0)} + \delta) = 0 &\Rightarrow Ax^{(0)} + A\delta = 0 \\ &\Rightarrow A\delta = -Ax^{(0)} = r_0; \end{aligned}$$

and hence, the projection step is applied to the system $A\delta = r_0$ to compute the unknown vector δ .

Projection methods are classified into two main groups. The first is when the *Krylov subspace* $\mathcal{K}$ is taken as $\mathcal{K} = \mathcal{L} = \text{span}\{r_0, Ar_0, \dots, A^{m-1}r_0\}$ and $V = W$ is an orthogonal basis of $\mathcal{K}$. This represents the class of *orthogonal projection methods* (also known as *Galerkin projection methods*). In this group of methods, each iteration aims to minimize the A-norm of the error vector $(x - \tilde{x})$ in the subspace $\mathcal{K}$. The second group of projection methods is when $\mathcal{L} = A\mathcal{K} = \text{span}\{Ar_0, A^2r_0, \dots, A^mr_0\}$ (and hence $W = AV$). Each iteration of this kind of methods aims to minimize the 2-norm of the residual vector. This explains why the latter class of methods are referred to as *minimal residual methods*.

Popular Methods. The most commonly used Krylov subspace methods are *generalized minimum residual* (GMRES), *biconjugate gradient* (BCG), *conjugate gradient squared* (CGS), *biconjugate gradient*

stabilized (BCGStab), and *quasi-minimal residual* (QMR). Of these, GMRES is a minimal residual method based on the Arnoldi procedure for orthogonalizing the Krylov subspace and storing its basis in a Hessenberg matrix. It is normally used with restarts to control storage requirements by forming the new approximation only after each restart. A nice by-product of GMRES is that the residual norm is available at each iteration without having to compute the approximate solution. BCG is an orthogonal projection method, and it takes an advantage over GMRES by reducing the storage demand. This is achieved by replacing the orthogonal sequence of residuals (formed by GMRES to build the basis of the Krylov subspace) by two mutually orthogonal sequences of residual vectors. However, the convergence behavior of BCG is quite irregular and it also requires a matrix-vector multiplication with the transpose of A . To increase the effectiveness of BCG, variants such as CGS and BCGStab are proposed. The rate of convergence of CGS is generally twice that of BCG. However, this is not always the case since a reduced residual vector may not be reduced any further. This explains the highly irregular convergence behavior of CGS. Moreover, rounding errors are very likely to occur in CGS as corrections to the approximate solution may be very large, and hence the finally computed solution may not be very accurate. CGS is almost as timewise expensive as BCG, but does not involve computations with A^T . On the other hand, BiCGStab is developed so that it is as fast as CGS while avoiding the often irregular convergence behavior of the latter. BiCGStab requires slightly more computations per iteration than BCG and CGS. Finally, QMR attempts to overcome the problems of irregular convergence behavior in BCG. It uses a least squares approach similar to that followed in GMRES, but with a bi-orthogonal basis for the constructed Krylov subspace, and hence, the name. To avoid breakdowns, QMR uses look-ahead techniques, which makes it more robust than BCG. Although it requires a matrix-vector multiplication with the transpose of A , a *transpose-free* version TFQMR is also available.

Preconditioners. The idea behind preconditioning is to accelerate the convergence process of an iterative method by redistributing the eigenvalues of the coefficient matrix so that the difference between the dominant and the subdominant eigenvalue becomes larger without changing the solution vector. Therefore, the need for a preconditioner becomes vital when dealing with ill-conditioned systems.

Now, consider the singular system of linear equations in $Ax = 0$, which can be transformed into the *right-preconditioned* system

$$AM^{-1}(Mx) = 0,$$

or into the *(left-)preconditioned* system

$$M^{-1}Ax = 0,$$

where the *preconditioner matrix* M (also called *preconditioner*) has the property that it is a cheap approximation of A . The more M resembles A , the faster the projection method converges. In fact, the methods of (B)JOR and (B)SOR are preconditioned POWER iterations in which the preconditioning matrices are, respectively, $M_{(B)JOR}$ and $M_{(B)SOR}$, and the POWER method can be viewed as a projection method where subspace information is kept only in a single dimension.

In the case of right-preconditioning, the system $AM^{-1}y = b$ is solved for the unknown $y = Mx$, and the final solution x is obtained through $x = M^{-1}y$. To use right-preconditioning, M should also be chosen so that $M^{-1}v$ is cheap to compute for any arbitrary vector v . In the left-preconditioning case, the system is solved on the basis of imposing the necessary stopping criteria on the *preconditioned residual vector* $r = -M^{-1}Ax$. The matrix M^{-1} need not be formed explicitly, and the preconditioned residual may be computed by solving the system $Mr = -Ax$. Therefore, the preconditioner M should be chosen so that solving any linear system of the form $Mv = u$ for any vector v is cheap.

Various types of preconditioners are considered for MCs [4,19]. Their efficiency is highly dependent on the system to be solved, and it is quite difficult to forecast

which preconditioner is the best for a given problem. In general, incomplete LU (ILU) preconditioners based on dropping nonzero elements smaller than a particular threshold value except those along the diagonal provide effective preconditioners for projection methods in a sequential setting. Recall that A is a negated irreducible singular M-matrix, and it is shown that ILU factorizations (in exact arithmetic) exist for such matrices [20]. When these preconditioners are coupled especially with BiCGStab [7], they provide strong solvers. However, one can also opt for inverse preconditioners, thereby reducing the preconditioning step to a matrix-vector multiplication [21,22], which is quite useful in a parallel setting.

REFERENCES

1. Grassmann WK, Taksar MI, Heyman DP. Regenerative analysis and steady state distributions for Markov chains. *Oper Res* 1985;33(5):1107–1116. DOI: 10.1287/opre.33.5.1107.
2. Sheskin TJ. A Markov partitioning algorithm for computing steady-state probabilities. *Oper Res* 1985;33(1):228–235. DOI: 10.1287/opre.33.1.228.
3. Dayar T, Akar N. Computing moments of first passage times to a subset of states in Markov chains. *SIAM J Matrix Anal Appl* 2005;27(2):396–412. DOI: 10.1137/S0895479804442462.
4. Philippe B, Saad Y, Stewart WJ. Numerical methods in Markov chain modelling. *Oper Res* 1992;40(6):1156–1179. DOI: 10.1287/opre.40.6.1156.
5. Stewart WJ, Wu W. Numerical experiments with iteration and aggregation for Markov chains. *ORSA J Comput* 1992;4(3):336–350. DOI: 10.1287/ijoc.4.3.336.
6. Migallón V, Penadés J, Szyld DB. Block two-stage methods for singular systems and Markov chains. *Numer Linear Algebr* 1996;3(5):413–426. DOI: 10.1002/1099-1506.
7. Dayar T, Stewart WJ. Comparison of partitioning techniques for two-level iterative solvers on large sparse Markov chains. *SIAM J Sci Comput* 2000;21(5):1691–1705. DOI: 10.1137/S1064827598338159.
8. Buchholz P, Dayar T. On the convergence of a class of multilevel methods for large sparse Markov chains. *SIAM J Matrix Anal Appl* 2007;29(3):1025–1049. DOI: 10.1137/060651161.
9. Dayar T. State space orderings for Gauss-Seidel in Markov chains revisited. *SIAM J Sci Comput* 1998;19(1):148–154. DOI: 10.1137/S1064827596303612.
10. Krieger U. Numerical solution of large finite Markov chains by algebraic multigrid techniques. In: Stewart WJ, editor. *Computations with Markov Chains*. Boston (MA): Kluwer; 1995. pp. 403–424.
11. Bru R, Pedroche P, Szyld DB. Additive Schwarz iterations for Markov chains. *SIAM J Matrix Anal Appl* 2005;27(2):445–458. DOI: 10.1137/040616541.
12. Marek I, Szyld DB. Algebraic Schwarz methods for the numerical solution of Markov chains. *Linear Algebra Appl* 2004;386:67–81. DOI: 10.1016/j.laa.2003.12.046.
13. Koury JR, McAllister DF, Stewart WJ. Iterative methods for computing stationary distributions of nearly completely decomposable Markov chains. *SIAM J Algebra Discr Methods* 1984;5(2):164–186. DOI: 10.1137/0605019.
14. Stewart GW, Stewart WJ, McAllister DF. A two-stage iteration for solving nearly completely decomposable Markov chains. In: Golub GH, Greenbaum A, Luskin M, editors. *The IMA volumes in mathematics and its applications 60: recent advances in iterative methods*. New York: Springer-Verlag; 1994. pp. 201–216.
15. Dayar T, Stewart WJ. On the effects of using the Grassmann–Taksar–Heyman method in iterative aggregation–disaggregation. *SIAM J Sci Comput* 1996;17(1):287–303. DOI: 10.1137/0917021.
16. Marek I, Szyld DB. Local convergence of the (exact and inexact) iterative aggregation method for linear systems and Markov operators. *Numer Math* 1994;69(1):61–82. DOI: 10.1007/s002110050080.
17. Marek I, Mayer P. Convergence analysis of an iterative aggregation/disaggregation method for computing stationary probability vectors of stochastic matrices. *Numer Linear Algebr* 1998;5(4):253–274. DOI: 10.1002/1099-1506.
18. Marek I, Pultarova I. A note on local and global convergence analysis of iterative aggregation–disaggregation methods. *Linear Algebra Appl* 2006;413(2–3):327–341. DOI: doi:10.1016/j.laa.2005.08.001.

19. Saad Y. Preconditioned Krylov subspace methods for the numerical solution of Markov chains. In: Stewart WJ, editor. *Computations with Markov Chains*. Boston (MA): Kluwer; 1995. pp. 49–64.
20. Buoni JJ. Incomplete factorization of singular M–matrices. *SIAM J Algebra Discr Methods* 1986;7(2):193–198. DOI: 10.1137/0607023.
21. Benzi M, Tuma M. A parallel solver for large-scale Markov chains. *Appl Numer Math* 2002;41(1):135–153. DOI: 10.1016/S0168-9274(01)00116-7.
22. Benzi M, Ucar B. Product preconditioning for Markov chain problems. In: Langville AN, Stewart WJ, editors. *Proceedings of the 2006*

Markov Anniversary Meeting. Raleigh (NC): Boson Books; 2006. pp. 239–256.

FURTHER READING

- Berman A, Plemmons RJ. *Nonnegative matrices in the mathematical sciences*. Philadelphia (PA): SIAM Press; 1994.
- Senata E. *Non–negative Matrices and Markov Chains*. New York: Springer; 1981.
- Stewart WJ. *Introduction to the numerical solution of Markov chains*. Princeton (NJ): Princeton University Press; 1994.

COMPUTATIONAL POOL: AN OR—OPTIMIZATION POINT OF VIEW

JEAN-PIERRE DUSSAULT

JEAN-FRANÇOIS LANDRY

Department d'Informatique, Sherbrooke,
Université de Sherbrooke, Québec, Canada

PHILIPPE MAHEY

LIMOS, Clermont Université and CNRS,
Clermont-Ferrand, France

INTRODUCTION

The interest in computational pool is motivated by a variety of factors. The mechanics of pool has long served as a focus of interest for physicists, an interest which has extended naturally to the realm of computer simulation. Realistic and efficient simulators have led to the proliferation of computer pool games. These games include significant computer graphics components, often including human avatar competitors with unique personalities, as well as an element of artificial intelligence to simulate strategic play. There are numerous one-person games available, of differing levels of physical realism.

Though in the context of this article the primary objective is the creation of a computer pool player, we wish to extend our research beyond the scope of billiards. We hope that by researching the best way of making a perfect player we can develop new AI approaches, and possibly contribute to other problems of that nature.

There have been three instances of pool Olympiads where a rival computational pool player competed. The competitions attracted participants from the artificial intelligence community, a pool player representing a challenge for the agent approach. Yet, only 8-ball tournaments were held, and the champions used a Monte Carlo search tree strategy [1–4]. In this article, we develop a complementary approach based on OR techniques,

simulation, dynamic programming (DP), and numerical optimization.

Although the problem we wish to solve is deterministic in nature, subject to the laws of physics, it is so easily influenced by many small factors that it can actually be seen as stochastic. This makes it very hard to create a perfect player, because even if he never misses, the outcome of the game is never predefined.

In this article we will explore the technical challenges of computational pool from an OR point of view. The section titled “The Cue Sports” describes the elements of the cue sports from a human perspective, including some discussion of the aspects of advanced human play. The section titled “Overview of Computational Pool” introduces the several OR aspects to be detailed later on.

The section titled “Modeling Billiards Physics” presents the requirements to produce a simulation model, and the section titled “Modeling Billiards Players” addresses the optimization of a billiards player—this is work in progress.

Future trends are discussed in the section titled “Perspective.”

THE CUE SPORTS

There are many types of cue sports, and many different terminologies and rules that vary with geographical location. For example, the most popular North American pool game is 8-ball, in which there are seven “low” (solid) balls (1–7), seven “high” (striped) balls (9–15) and the 8-ball. The objective is to be the first to sink all balls in one group, and then the 8-ball. Alternately, snooker is probably the most popular cue sport in England, which is played on a larger table and consists of 15 red balls, and 6 of various colors. The balls are smaller than their North American counterparts, and the objective is to accumulate points by alternately sinking a red and then a colored ball, each having a certain value. In France,

a game called *carom* is played, where only three balls are used on a pocket-less table, and the player must attempt to contact both object balls with the cue ball. In the three cushions (“trois bandes” in French) version of carom, the cue ball is required to rebound off three rails before kissing the object ball.

Pool Terminology

The basic elements common to all of these games are as follows. For a much more complete glossary of cue sport terms, see the Wikipedia [5, p. 3] from which we extract the terms relevant to our computational pool study.

- *Cue*. Also known as *cue stick*. A stick, usually around 55–60 in. in length with a tip made of a material such as leather and sometimes with a joint in the middle, which is used to propel billiard balls.
- *Cue Ball*. Also known as *cueball*. The ball in nearly any cue sport, typically white in color, that a player strikes with a cue stick. Sometimes referred to as the *white ball*, *whitey* or *the rock*.
- *Object Ball*. Depending on context it is
 1. any ball that may be legally struck by the cue ball (i.e., any ball-on);
 2. any ball other than the cue ball.
- *Table*. It is the flat playing surface. The table size varies, depending on the game played, but is always rectangular, being twice as long as it is wide. In most games, the table has a pocket in each of the four corners, and one at the centers of each length.
 1. *Bed*: The table is covered in a textured felt (or baize) material, usually of green color, to add a frictional damping effect to the shots.
 2. *Rail*: A rubberized edge running along the inner boundary of the table, to accommodate rebounds following the collision of a ball (cushions in British English).
 3. *Table State*: the position of all balls at rest on the table, ready for the next shot.
- *Shot*. The use of the cue to perform or attempt to perform a particular motion of balls on the table, such as to pocket (pot) an object ball, to achieve a successful carom (cannon), or to play a safety. Different classes of shot type include the following:
 1. *Direct Shot*: A shot where the cue ball hits an object ball, which then reaches a pocket.
 2. *Kick Shot*: A shot in which the cue ball is driven to one or more rails before reaching its intended target—usually an object ball. Often shortened to “kick.”
 3. *Bank Shot*: A shot in which an object ball is driven to one or more rails prior to being pocketed (or in some contexts, prior to reaching its intended target; not necessarily a pocket). Sometimes “bank” is conflated to refer to kick shots as well, and in the United Kingdom it is often called a double.
 4. *Combination Shot*: This is also known as combination or combo. Any shot in which the cue ball contacts an object ball, which in turn hits one or more additional object balls (which in turn may hit yet further object balls) to send the last-hit object ball to an intended place, usually a pocket. In the United Kingdom this is often referred to as a plant.
 5. *Safety Shot*:
 - An intentional defensive shot, the most common goal of which is to leave the opponent either no plausible shot at all, or at least a difficult one.
 - A shot that is called aloud as part of a game’s rules; once invoked, a safety usually allows the player to pocket his or her own object ball without having to shoot again, for strategic purposes. In games such as seven-ball, in which any shot that does not result in a pocketed ball is a foul under some rules, a called safety allows the player to

miss without a foul resulting. A well-played safety may result in a snooker (no direct shot available).

- *Spin*. Rotational motion applied to a ball, especially to the cue ball by the tip of the cue, although if the cue ball is itself rotating it will impart (opposite) spin (in a lesser amount) to a contacted object ball. Types of spin include follow or top spin, bottom or back spin (also known as draw or screw), and left and right side spin, all with widely differing and vital effects. Collectively they are often referred to in American English as “English.”

Specificity of Game Variants

To excel in any pool variant, one has to thoroughly understand the physics, accurately control the cue ball, and plan several shots in advance. However, each variant has its specificity, and the relative weight of those game aspects slightly differs from one variant to another. We examine here four among the most popular games and their specifics.

Games Using Numbered Balls. On the American continent, most pool tables use numbered balls (1–15) in a variety of games. The “official” tables measure $4.5' \times 9'$, but smaller models ($4' \times 8'$ or even $3.5' \times 7'$) are quite common. Hereafter, we briefly discuss three popular game variants using such equipment.

Eight Ball. This is by far the most popular game played on the American continent. However, some of its popularity stems from its use in bars, on coin-operated tables, and the rules of the game reflect this situation.

It is the variant where chance plays a major role, and this is for three main reasons:

- no call is made on the break shot, so any ball pocketed by luck on the break is valid;
- on a foul play, pocketed balls remain in the pockets;
- both players do not play on the same set of balls.

Although not the primary variant in high level human competitions, this has been the variant in the first three computational pool tournaments.

Planning is limited to eight shots before the win. On the other hand, the presence of opponent’s balls may complicate the plan.

Nine Ball. This variant is played with nine balls, and the specifics include the obligation to hit the lowest numbered ball on the table. Thus, planning is somewhat simpler here because of this restriction on which ball is first contacted by the cue ball.

Straight Pool. This is played on similar tables as eight balls pool, but here, any ball has to be called including on the break shot, and balls sunk on a foul are pulled out of the pocket and respotted on the table. Champions achieve runs of several hundred consecutive successful shots. It is clear that this variant requires both clever planning and extreme accuracy. High risk shots involving several rebounds on the rails or with other balls are rare, the players often preferring to resort to defensive shots than to risk losing its turn leaving the opponent with an easy table.

Snooker. Snooker is the most popular variant in the United Kingdom. It is played on a large ($6' \times 12'$) table using 15 red balls and 6 colored balls in addition to the cue ball. The large dimension of the table and the use of slightly smaller balls call for high accuracy. Here again difficult bank, kick or combination shots are often avoided by the frequent use of defensive shots. The name snooker itself refers to the situation where a player has no legal direct shot available.

Carambol. Carom is played on a pocketless table using only three balls, slightly larger than numbered balls though. The game consists in having the cue ball hit the other two balls. In high level competitions, the three cushions variant is often preferred, in which the cue ball has to rebound on three cushions before hitting the second of the other two balls. Obviously, this three cushions variant is extremely demanding in geometric skills,

and the sole difficulty of a single shot almost precludes any form of planning.

In opposition to most pocket billiards variants, the physics model of the collision with the rail is of primary importance here, as is the accurate control of the cue ball. On the other hand, the strategic planning aspect is significantly less important.

OVERVIEW OF COMPUTATIONAL POOL

In this section, we present the three basic building blocks used in a computational billiards player. As we discussed in the section titled “Specificity of Game Variants,” different variants of the game require different balance and specificity in the building blocks.

Physics Simulation

The simulator is required for two aspects in computational pool.

- The first one is to act as a common table and referee in order for the (computer) players to compete. In an ideal world, the simulated game would be indistinguishable from its counterpart as played on a real table. This involves meticulous modeling of the impacts between the cue, balls, rails, and (whenever balls jump) the table. Moreover, the rolling motion of the balls on the cloth must be addressed faithfully.
- The second aspect is a tool on which the AI players rely in order to predict the outcome of a given shot. Here, speed–accuracy trade-offs are expected.

In the first three editions of the computer pool tournaments, the same simulator was used for both uses, some random noise being added on the shots played on the “common table.” As the situation evolves, it is likely that referee simulators will get more and more complex and faithful, yielding high computational costs perhaps no more compatible with the requirements of computational players under limited resources. Ultimately, the games will be played on a real table by robots, and the simulators used by the players will be an integral part of the player’s strategy.

Control of the Cue Ball

When observing a champion at some variant of pocket billiards, two aspects are striking: he/she has mostly easy-looking shots to play, but whenever he/she faces a difficult shot, he/she nevertheless sinks the ball. The first aspect is less spectacular, but is a consequence of his/her accurate repositioning after the impact on the object ball.

Strategic Planning

Another aspect is that the champion takes enlightened decisions with respect to the order in which he/she will sink the balls. In computational pool, this aspect is addressed using some dynamic model of the game aspect.

Among the most evident strategic decisions is the choice to do a so-called *defensive shot*. Instead of aiming to sink some object ball, the player strives to leave the cue ball in such a position that his/her opponent is left with a very difficult situation.

MODELING BILLIARDS PHYSICS

In this section we first summarize the mathematical models that have been developed over the years, going back to Coriolis in 1835. Next we discuss two computer implementations.

Mathematical Models

All variants of billiards and pool games involve a cue, hard spheres, a bed which is a flat surface covered with a cloth, and rail cushions. Spheres slide and roll on the bed, impact on each other as well as on the rails, and the cue gives the initial impact to the cue ball. Thus, we need to model

- cue–cue ball impacts;
- moving balls; mostly, sliding and rolling on the bed, but sometimes “jumping” off the bed for some tiny amount of time, and even sometimes jumping off the table;
- collisions between moving hard spheres, mostly elastic collisions between rigid bodies;

- collisions between hard spheres and cushions on the rails, partly inelastic collisions between a rigid body and a not-so-rigid one (the rail).

All of this is quite delicate to accurately simulate but since the moving bodies are identical spheres, and the bed is planar, the physics equations are rather simple.

The interest in the mathematics of billiards goes back to at least Coriolis [6], and has attracted recent interest as well. In particular, Alciatore, in addition to his book [7], maintains a web site [8] where he posts more mathematically oriented developments including fine-tuning of physical models required to accurately predict the experiments he conducts using high speed video cameras.

Rolling Motion—2D Simplifications. An important observation yields computationally tractable prediction for balls motion on the table. The trajectory of the balls is described by a piecewise quadratic parametric equation. This allows simplification of the collision detection to simultaneous quadratic equations, for which closed-form formulæ are available [9], (Private communication: Sénéchal D. Mouvement d'une boule de billard entre les collisions. 1999) [10]. Moreover, extension to 3D simulation may still be described by piecewise parametric quadratic equations. Therefore, in principle, it is possible to implement a discrete event simulator without using numerical integration. Moreover, since the underlying parametric equations are of the simple piecewise quadratic type, an improved implementation will provide sensitivities and derivatives that are most useful for the optimization aspects to be discussed in the section titled "Optimization Approach."

Ball–Ball Collisions. The balls are almost rigid bodies, and their collision is almost frictionless. The behavior of the object ball and cue ball after the collision are described by the so-called 30° and 90° rules. Actually, slight nonelasticity, small friction, speed, and spin all slightly modify the rules. In Alciatore [8], one can find numerous technical notes providing the fine points of billiards physics.

Ball–Rail Collisions. Those collisions are much more complicated. In most pocket billiards variants though, kick or bank shots are not that frequent, so a simple approach is justified. In three cushions, as the name of the game suggests, those collisions are of prime importance. Larger balls used in carom variants and high energy ball–rail collisions imply that the cushions undergo important deformations, which can no longer be faithfully computed using simple geometric models alone. For instance, in Classic Billiard [11], (Fray J-M. personal communication), a one-person carom game, shots were performed and captured with a camera and in the actual game, interpolation is used to adjust the simulator to replicate the same result. This usage of adjustable variables was necessary as the direct implementation of the physics of the collisions as described in Coriolis [6] yielded different and inaccurate results when compared to real life.

Existing Physically Documented Simulators

While there are many one-person pool playing games, there are few efforts disclosing the actual equations or physical models underlying their simulator; we describe hereafter, two softwares available under open source licenses.

PoolFiz–FastFiz. The first three instances of the pool Olympiads used the PoolFiz simulator [9,10,12], lately replaced by the reimplementations FastFiz [13]. This simulator implements the simplified 2D rolling equations. The actual movement of balls has a spatial component (how the balls move on the table) and a rotational component (how the balls spin). The spin may be represented as rotational velocities around the three main axes, $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$, $\mathbf{e}_3$ pointing upward and perpendicular to the table's surface. The present implementation neglects any spin in the $\mathbf{e}_3$ axis. The simulator was made available through a common interface to the competitors, who used it to compute their shots. When used as the referee table, some random noise was added to the shot parameters.

In the simulator, the following parameters describe a shot:

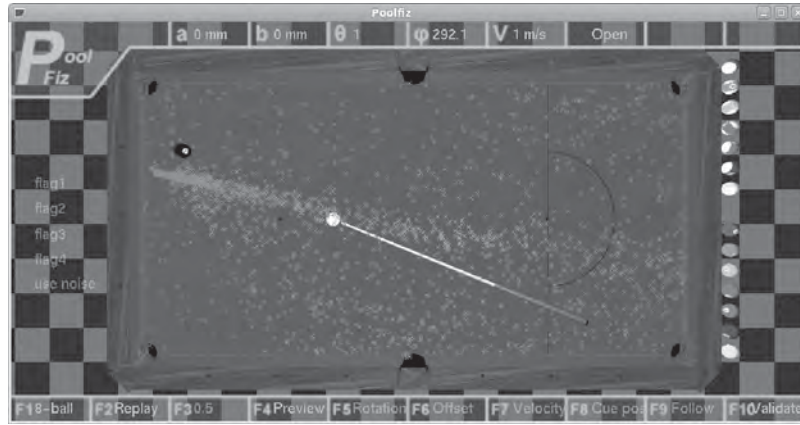
- a and b representing the side-spin and top-spin with respect to the cue ball center;
- θ the elevation of the queue stick;
- ϕ the orientation of the queue stick on the table;
- v , the initial speed given to the cue ball.

Billiards. The billiards—techné project [14] adopts a different approach; namely, it uses the ODE open source dynamics engine.

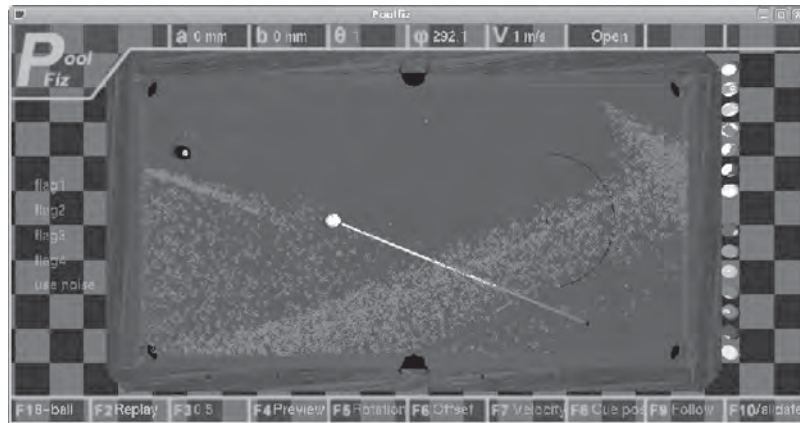
From Smith [15], ODE is an open source, high performance library for simulating rigid body dynamics. It is fully featured, stable, mature, and platform-independent with an

easy-to-use C/C++ API. It has advanced joint types and integrated collision detection with friction. ODE is useful for simulating vehicles, objects in virtual reality environments, and virtual creatures. It is currently used in many computer games, 3D authoring tools, and simulation tools.

As such, physical objects (cue, balls, table bed, table rails, etc.) are described using the ODE interface, and simulated as accurately as ODE allows. No advantage is taken from the simplified nature of perfect spheres rolling on a planar surface. While still in progress, the physics has 3D realism, and care has been taken to validate the



(a)



(b)

Figure 1. Reachable regions. For a simple shot in the upper right corner, we observe the reachable regions for (a) billiards and (b) PoolFiz.

results with high speed camera observations to accurately model the cue ball deflection when using English, an aspect absent from PoolFiz. The approach using ODE has some drawback though: the billiard simulator cannot produce noiseless results since it is built using a fairly sophisticated general purpose physical simulator in which Monte-Carlo techniques are used for improved reliability. There are ways to reduce the random effects but it is impossible, it seems, to completely eliminate randomness.

Simulator Comparison. We illustrate in Fig. 1 the differences in repositioning possibilities of the two simulators. Probably neither is realistic enough to predict repositioning on a real table and further validation is certainly required. Important differences come from the fact that PoolFiz does not take side spin into account on ball–rail collisions, which explains the large area unattainable. Billiards, as compared to PoolFiz, results in less influence of speed on the cue ball deflection after impact with the object ball.

MODELING BILLIARDS PLAYERS

As discussed above, some kind of simulation of the physics is required for the computational player to be able to predict the outcome of a given shot. In this section, we assume such a simulation model is available in the form of a black box. The more accurate the simulator, the more the accurate predictions and planning of the AI player. In the first Olympiads, all the players shared the same simulator, which also acted as table referee. When playing on a robot, or when using a highly accurate simulator, it is expected that different players will use different simulators for their planning.

Two aspects mainly require our attention. A sequence of shots has to be established and thereafter, their execution has to be computed. By a shot, we mean a somewhat generalized shot including repositioning of the cue ball. In the following subsections, we first address the planification of the sequence of shots as modeled by a DP formulation. We also discuss some solution strategies for this

admittedly abstract DP formulation. Actual shots required by the DP planner are then computed using a local optimization model. We cannot avoid the temptation to report our preliminary experience with respect to the local optimization.

Before going further, we must warn that the following modeling assumes the 8-ball game setting. As the model is rather abstract, we expect no fundamental problem in extending this to other pool variants, but of course several minute details will have to be adjusted for other forms of the game.

A Dynamic Programming Formulation

Modeling billiard games is a difficult task and very few known frameworks fit correctly the specific characteristics of these games which are as follows: continuous state and control spaces; actions taken at discrete unknown instants; a specific turn-taking game rule which allows a player to play several times as long as he/she keeps the hand; and stochastic perturbations of moves. Archibald *et al.* [16] have proposed a stochastic game framework valid for finite or repetitive versions of the billiard games and they studied the existence of Nash equilibria associated with stationary Markov strategies. We will first focus on the stochastic control problem, mixing discrete and continuous aspects before introducing the game theoretic ingredients in the model.

Stochastic DP has been introduced to model dynamic discrete processes where random events occur after decisions are applied at each stage of the process. The state space X specifies the position of each ball on the table. It is continuous here but may be discretized to locate balls on the table. The control space $\mathcal{U}$ corresponds to the value of the cue parameters which are adjusted at the beginning of the shot. Each transition is the result of a nonlinear constrained optimization problem that represents the dynamic trajectory of the balls on the table under random perturbations. As seen below, one tries here to minimize the distance of the cue ball from a specified location while maximizing the probability of sinking the object ball. The overall objective function is to maximize the reward of the player (which

is positive when the object ball is sunk in the pocket). The way to model that situation is to add a Boolean variable y_t at each stage which is one when the object ball is sunk. Observe that when $y_t = 0$, the hand is lost and the opponent will inherit the current table. Thus, in difficult situations, a defensive step will try to minimize the opponent's reward. Note also that as long as expectations are used in the stochastic case, a value function can be estimated by backward calculations, starting at an ideal state to pocket the last ball.

Let $\pi = (u_0, \dots, u_{T-1})$ be a feasible policy (series of actions for the states $(s_1, \dots, s_T)$), $P[s' | s, u]$ the probability of going from state s to state s' while taking action u , $R(s)$ the reward for reaching state s . The stochastic program can be stated in the following way (using here a finite-length game):

Maximize

$$\begin{aligned} & \gamma P[y_T = 1 | (u_1, \dots, u_{T-1}; \omega_1, \dots, \omega_{T-1})] \\ & + \sum_t E[R_t(s_t, y_t, \omega_t)] \\ & (s_{t+1}, y_{t+1}) = \operatorname{argmin}_{s(\tau) \in X, u_t \in \mathcal{U}} \\ & \times E[F(s_t, u_t, s(\tau), \bar{s}_t, \omega_t)] \\ & s_t \in X, y_t \in \{0, 1\}, \end{aligned} \quad (1)$$

where $\omega_t \in \Omega_t$ is a random variable with normal distribution and γ is an harmonization factor. The transition step associated with the internal minimization subproblem will be detailed below. It relies on a target position for the cue ball corresponding to state $\bar{s}_t$. The trajectory is represented by $s(\tau)$ where $\tau \in [0, \tau_f]$ is the continuous-time of the motion of the balls, $s_t = s(0)$, $s_{t+1} = s(\tau_f)$.

Let $V^\pi(s)$ now be the optimal expectation of the final reward when applying the sub-policy $(u_t, \dots, u_{T-1})$ from state s_t through the final state.

The equation

$$V^\pi(s) = R(s) + \gamma \sum_{s'} P(s' | s, \pi(s)) V^\pi(s')$$

will be at its optimum if it satisfies the condition

$$V^*(s) = R(s) + \max_{\pi(s)} \sum_{s'} P(s' | s, \pi(s)) V^*(s').$$

Thus, the subpolicy $\pi^* = \{u_t^*, u_{t+1}^*, \dots, u_{T-1}^*\}$ will be optimal for this problem.

For billiards, as stated in Archibald and Shoham [16], we are faced with a continuous-state and continuous-action domain, which introduce further complications.

We find ourselves evaluating an integral instead of a sum

$$V(s) = R(s) + \max_{\pi} \gamma \int_{s'} P(s' | s, \pi(s)) V(s') ds'$$

The optimal strategy of discrete dynamic programs is commonly solved by performing *backward computations* from the final state back to the current state. We will discuss that strategy below, after saying a few words about the underlying game theoretic issues.

Game Issues

Let us consider first, the simpler situation of a finite game (such as 8-ball). It is not easy to analyze the general strategy which interprets a defensive shot as the one which will *leave a difficult table to the adversary, thus expecting to take back the hand later to finish the game*. This means indeed that the final profit associated with $P[y_T = 1]$ is still valid. Nevertheless, we should here pull from a possible table at stage t to an unknown table at $t + q$ where q is the number of estimated shots left to the opponent. We refer to Archibald *et al.* [17] for an interesting AI point of view of the strategic issues.

Solution Strategies for the DP Model

The general problem of finding $V^\pi(s)$ for a deterministic problem may be quite complex by itself, and often becomes intractable as more variables come into play. It is even worst for a problem with a continuous state and action space. However, in our case, we can benefit from the knowledge we have of the game we are solving and extract important information to narrow our search to a limited number of action–state space combinations.

Assumption 1. $y_t = 1, t = 1, \dots, T$.

Assumption 1 corresponds to the so-called *win-off-the-break* winning sequence. The optimal value is hence the expected sum of remaining rewards (for each shot) to get

through the game. The way to cope with the fact that the success of the game is now deterministic is to add a *technical reward* to an easier shot, or equivalently, to a better position for the cue ball on a given table.

Assumption 2. The cue and object balls are the only balls expected to move.

With Assumption 2, we get a direct estimate of the future positions of all future object balls. The player strategy corresponds to the “optimistic planner” described in Archibald *et al.* [17].

Observation. Later, we will naturally relax Assumption 2 to authorize collisions with other balls (and with the rails). Then, a new question will arise: should we consider only side collisions due to the trajectory of the cue and object balls, or should we include different shots aiming at repositioning the cue ball so that a collision happens with another ball (or a cluster of dangerously close balls)? That latter situation is now under study. The skill model to compute the optimal trajectory of the balls will here be much more intricate as it will involve elastic collisions and singularities [18].

Back to the stochastic dynamic program, assuming Assumptions 1 and 2, we will now proceed backward to estimate the optimal value of a given table. We may start with the last table which can be modeled as the current table where we want to sink, not the current object ball, but the one which should be sunk at the last shot. To compute all $V^\pi(s_{T-1})$, we must compute the pocket angle for each of the remaining balls and then perform an inverse step to position the cue ball in all feasible states at $t = T - 1$. The optimization black box will be very useful as it already works with an estimate $\bar{s}_t$ of the desired repositioning state.

Observation. The curse of dimensionality will probably make these computations cumbersome and time-consuming if T is large, but we can begin testing increasingly complex situations setting $T = 2, 3, \dots$ which can be easily adapted from the general case.

Optimization Approach. Here, we describe briefly the minimization subproblem associated with the transition step at stage t as defined in model 1.

Suppose we are using the black box simulator described earlier, with the action variable u_t being the finite-dimensional vector of the five parameters a, b, θ, ϕ , and v . The trajectories $s(\tau)$ of the balls moving on the table at a given shot depend thus on the initial state $s(0) = s_t$ and on u_t , the latter being perturbed by the random variable ω_t .

To compute an optimal action u_t , we minimize a least-square function associated with the desired reposition $\bar{s}_t$ of the cue ball on the table and the desired sinking of the object ball. The objective function can thus be modeled in the following way:

$$\begin{aligned} F(s(\tau), u_t, \omega_t) &= \|s(\tau_f) - \bar{s}_t\|^2 \\ \text{subject to } s(\tau) &= \psi(s_t, u_t, \omega_t) \\ y &= 1 \text{ if } s_{obj}(\tau) = \bar{p} \text{ for } \tau \in [0, \tau_f], \end{aligned}$$

where ψ is the dynamic transition function associated with the simulator, s_{obj} is the component of s associated with the object ball, and $\bar{p}$ is the component of $\bar{s}_t$ associated with the pocket. The solution of the minimization subproblem will then give $s_{t+1} = s(\tau_f)$ with the corresponding value of y updating y_{t+1} . In the stochastic case, the function is replaced by an expectation value computed from the known distribution of the noise on the action variables.

The detailed formulation of the minimization problem and its resolution by an adaptation of Powell’s derivative-free Quasi-Newton method [19] can be found in Landry and Dusault [20].

Value of Billiards. In a general matter, the difficulty of executing a shot on a billiard table is defined by two aspects: (i) the distance between the cue ball, object ball and pocket, and (ii) the angle at which the cue ball will hit the object ball. Other factors, such as the closeness of the cue ball to the rails or other balls partially blocking the path of the shot, also come into play. However, we can still get a fair estimate of a shot difficulty by using the first two aspects described. A real and precise

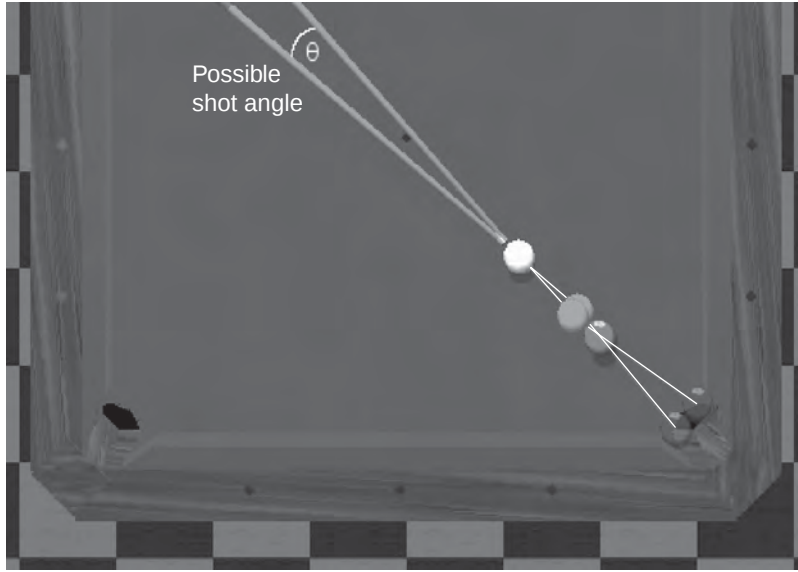


Figure 2. Shot difficulty: The distance of the cue ball to the object ball, of the object ball to the pocket, the cut angle, and the angle with the pocket all influence the difficulty of a shot, proportional to the admissible error with respect to the “perfect” shot that will sink the object ball right in the center of the pocket.

difficulty coefficient is almost impossible to compute since it will not only rely on the percentage of success of the computed shot, but also on the impact of this shot for the cue ball repositioning.

As shown in Fig. 2, both the angle and distance factor result in a calculable angle θ for the shot difficulty. A shot aimed at the center of this angle should normally result in the best success percentage for this shot.

Once we have computed this coefficient, we can use it to evaluate any given table state, simply by taking the maximum value of all the available shots for a given cue ball position. If we now divide our billiard table into a 50×100 grid, we can take the value of $\max_{i=0}^n \theta_i$ at each of these points and come up with the best repositioning targets for the next shots by removing each time the ball we would be aiming for (since we assume it will be pocketed).

However, this first and seemingly logical evaluation of a billiard table state is not necessarily applicable to all variants of the game, and does not for the moment take into

account defensive shots. It does however, provide an estimation of table state values for offensive play.

Results. To show the effectiveness of the optimization model discussed in this article, we ran some tests on randomly generated table states and computed possible shots using local and global optimization libraries. We show here only two of the best performers we tested for the sake of brevity. BOBYQA was selected since it was the fastest of the approaches tested and DIRECT-L was the most successful with the best solutions found. Both approaches were implemented using the NLOpt library [21]. BOBYQA, a local optimization method by James and Powell [19] was tested with five different starting points, distributed over the possible shot parameters. DIRECT-L, a global optimization method by Jones *et al.* [22], was tested with only one starting point.

A total of 1000 random table states were generated and each possible shots on these tables were explored, thus coming close to a total of 20,000 various shots which

Table 1. Results of 19930 Different Shots Computed with the Local Optimization Method BOBYQA and Global Optimization Method DIRECT-L

	Success (%)	Average time (s)	Average time success shots (s)	Average Distance (m)
<i>DIRECT-L</i>				
Direct shots (4310)	97.21	0.37	0.38	0.15
Kick shots (6950)	95.59	0.78	0.78	0.19
Bank shots (7520)	72.0	0.85	0.99	0.59
Combination shots (1150)	85.13	2.02	2.09	0.4 m
<i>BOBYQA</i>				
Direct shots (4310)	96.10	0.08	0.07	0.34
Kick shots (6950)	88.21	0.13	0.12	0.43
Bank shots (7520)	35.45	0.12	0.16	0.68
Combination shots (1150)	46.00	0.31	0.34	0.62

Average times (sec) are displayed for all shots, as well as for successful shots. The average distance (m) is the distance from the repositioning target.

were computed using both libraries. Each of these shots represents a combination of an object ball, an aimed pocket, a cue ball, and a repositioning target. As can be seen in Table 1, types of shots included direct, kick, bank, and combine shots.

The local optimization method, even though launched five times with various starting points, gets a slightly lower success percentage for bank and combine shots. The DIRECT-L, however, being a more global approach, achieves better results but at a higher cost in calls to the simulator. What these results actually show us is that we can obtain satisfying results to compute single shots with very few calculations. The trade-off between the solution’s quality and the computation time needs to be carefully studied to achieve the best compromise, but this will also vary with respect to the variant of game played. It should also be noted that in case a solution is not found for a given shot, it is usually because it is not possible because of other balls or rails restricting the queue motion. Thus, a global method may be able to find a solution by doing a spectacular shot but one that is not very reliable when played on a stochastic system. Further testing is needed to adapt the search criteria to the type of game played, and the level of noise present in the simulator.

PERSPECTIVE

We described a new application field of computational billiard and proposed a method of approaching the problem from an OR perspective. The field is really emerging, and much is to be expected in the near future. However, the ultimate goal to have a computer robot challenge human champions still appears quite remote. We did not discuss the robotic aspect of this challenge, but it should be clear that good progress has been made on the computational aspect, and that much remains to be done in this new field.

From an OR-specific point of view, we have described a stochastic DP framework to model an optimal strategy of a generic billiard player in the presence of noise. The key contribution is an improved estimation of the table value at some stage that takes into consideration the expectation of the difficulty to complete a sequence of winning shots and not only the current one. The present modeling is coherent with the conclusions of Archibald *et al.* [17] who state that “strategic skill” is decisive whenever “the execution skill” is imperfect.

On the other hand, we have presented an improved player simulator based on an accurate optimization tool to drive the cue ball trajectory toward the desired spot with the “best” table value. Numerical simulations on repeated launching of 8-ball games were

performed with relative success, including indirect shots like bank and combine shots, confirming the robustness of our player.

Current work aims at improving the player skill to cope with more realistic situations including the precise modeling of elastic collisions with banks and the breaking of clusters of balls.

Finally, the introduction of defensive strategies to take into consideration the game theoretical aspects will be the direction of further studies. It is indeed evident that the development of future competitions between virtual players controlling robots acting on a real table will progressively increase the level of noise effects, thus making more crucial the improvement of the “strategic skill” of the players.

Acknowledgments

We are grateful to François Gaumond and Benoit Hamelin for their careful reading of the manuscript and useful comments and François Gaumond’s work on the Billiards–PoolFiz interface. We also wish to thank the referees for their constructive suggestions.

REFERENCES

1. Greenspan M. UofA wins the pool tournament. *Int Comput Gaming Assoc J* 2005;28(3): 191–193.
2. Greenspan M. PickPocket wins pool tournament. *Int Comput Gaming Assoc J* 2006;29(3): 153–156.
3. Smith M. PickPocket: a computer billiards shark. *Artif Intell* 2007;171(16-17): 1069–1091.
4. Archibald C, Altman A, Shoham Y. Analysis of a winning computational billiards player. *IJCAI’09: Proceedings of the 21st International Joint Conference on Artificial Intelligence*; San Francisco (CA): Morgan Kaufmann Publishers Inc; 2009. pp. 1377–1382.
5. Glossary of cue sports terms. [www. Wikipedia.com](http://www.Wikipedia.com). Accessed 2009.
6. Coriolis G-G. *Théorie mathématique des effets du jeu de billard*. Paris: J. Gabay; 1835.
7. David A. *The illustrated principles of pool and billiards*. New York: Sterling Publishing; 2004.
8. Alciatore DG. Pool and billiards physics resources. Available at <http://billiards.colostate.edu/physics>. Accessed 2009.
9. Leckie W, Greenspan M. Pool physics simulation by event prediction 1: motion transitions. *Int Comput Gaming Assoc J* 2005;28(4): 214–222.
10. Leckie W, Greenspan M. Pool physics simulation by event prediction 2: collisions. *Int Comput Gaming Assoc J* 2006;29(1): 24–31.
11. Classic billiards. Canal + multimédia.
12. Leckie W, Greenspan M. An event-based pool physics simulator. 11th International Conference on Advances in Computer Games. *Lecture Notes on Computer Science*. No. 4250. Heidelberg: Springer; 2006. pp. 247–262.
13. FastFiz. Available at <http://www.stanford.edu/group/billiards/FastFiz-0.1.tar.gz>. Accessed 2009.
14. Papavasiliou D. Billiards. Available at <http://www.nongnu.org/billiards>. Accessed 2009.
15. Smith R. ODE. Available at <http://www.ode.org/>. Accessed 2009.
16. Archibald C, Shoham Y. Modelling billiards games. In: Decker KS, Sichman JS, Sierra C, Castelfranchi C, editors. *Proceedings of the 8th International Conference on Autonomous Agents and Multiagents Systems (AAMAS 2009)*; Budapest: 2009. pp. 193–199.
17. Archibald C, Altman A, Shoham Y. Success, strategy and skill: an experimental study. *Proceedings of the 9th International Conference on Autonomous Agents and Multiagents Systems (AAMAS 2010)*; Toronto: 2010.
18. Miller BM, Bentsman J. Optimal control problems in hybrid systems with active singularities. *Nonlinear Anal* 2006;65:999–1017.
19. James M, Powell D. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report no. NA06. Cambridge: Department of Applied Mathematics and Theoretical Physics; 2009.
20. Landry J-F, Dussault J-P. AI optimization of a billiard player. *J Intell Robotic Syst* 2007; 50(4):399–417.
21. Johnson SG. The NLOpt nonlinear-optimization package. Available at <http://ab-initio.mit.edu/nlopt>. Accessed 2009.
22. Jones DR, Perttunen CD, Stuckman BE. Lipschitzian optimization without the Lipschitz constant. *J Optim Theor Appl* 1993;79(1): 157–181.

CONCEPTS OF NETWORK RELIABILITY

CHARLES J. COLBOURN
SCIDSE, Arizona State
University, Tempe,
Arizona

THE BASIC MODEL

The likelihood that a network supports an acceptable level of operation is a key metric in network design and network performance. To quantify this, one must develop a precise model of the network, the causes of failures, and the meaning of network operation. Then, depending upon the specification of each, a quantitative measure of the ability of the network to operate is developed.

In this short overview, basic concepts are developed but no attempt is made to capture the wide variety of definitions, algorithms, and theories that have been developed. In general, for more in-depth expositions, see Refs 1–3 for network models, Refs 4–6 for computational complexity results, Refs 1, 7, 8 for Monte Carlo methods, Refs 9–11 for Boolean methods, and Refs 1, 12 for combinatorial bounds using complexes. A limited number of additional references are used in the text to follow.

In order to motivate the topic, let us consider a simple network design example. Seven nodes are to be connected by bidirectional communication links. Although six links suffice to provide at least one path connecting each pair of nodes, one or more of the links chosen may fail. If only six links are constructed, then a single failure results in at least one pair of nodes having no operating communication path. Of course, if network cost is not a substantial concern, we could simply construct all $21 = \binom{7}{2}$ possible links, and then no failure of fewer than six links can cause two nodes to be unable to communicate.

Network reliability is concerned with balancing the desire to limit network construction cost and the need for continued operation in the event of component failure. In our toy example, suppose that we have determined that any 10 links can be constructed. Figure 1 shows three possible ways to select the links.

The basic question is: Which one is “better”? The answer naturally depends on what we expect the network to do. If our goal is to ensure that node 2 can reach node 3, the failure of a single link cannot prevent this. In the first network, there is no way for two links to fail, so no path remains from node 2 to node 3. However, in the second, there is exactly one way. On this basis, we might prefer the first network to the second. If, on the other hand, our goal is to ensure every node can communicate with every other node, again no single link failure can prevent this. However, in the first network there are four ways to remove a pair of links to disrupt some connection, while in the second there are only three. On this basis we might prefer the second network to the first. But for both operations, we might prefer the third choice among these.

Naturally, arguments of this type are both imprecise and overly simple. A genuine application would require a substantial amount of information not provided in this toy example. Here, our objective is to make the concepts involved, including the model, the network operation, and the measures of network reliability, precise. With those in hand, one can choose a reliable network design from a huge variety of candidates typically available, and evaluate how likely it is to support the specified operation.

Network Model

Network reliability has been concerned primarily with networks for digital communication. A network begins as a collection of *nodes*, each representing a site that originates, terminates, or forwards network traffic. Although a wide variety of technologies

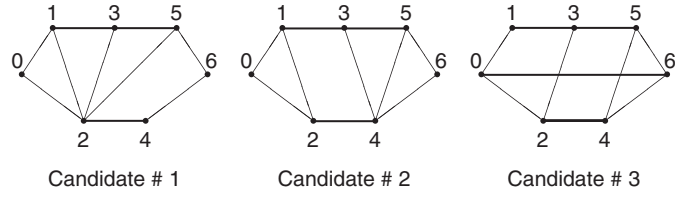


Figure 1. Three networks with seven nodes and 10 links.

are available to support communication, at the most basic level each provides a *link* that permits communication among a specified set of nodes. Links may be *unidirectional*, supporting traffic from one node to others but not supporting return traffic; or *bidirectional*, supporting traffic in both directions between nodes. Links may be *point to point*, supporting traffic between exactly two nodes; *multipoint*, supporting traffic among a collection consisting of some but not all nodes; or *broadcast*, permitting communication among all nodes. Traditionally, network reliability has been concerned with point-to-point links; this is used for illustration here, but the concepts apply more generally.

A *graph* is a set V of elements called *vertices*, and a collection E of pairs of vertices. Each $e = \{x, y\} \in E$ is an *edge* of the graph. When nodes of a network are represented as vertices, a bidirectional link can be represented as an edge. In this way, graphs serve as very simple models of networks. In a similar way, a unidirectional link can be represented as an ordered pair of vertices, forming a *directed edge* or *arc*. When E consists solely of arcs, $G = (V, E)$ is a *directed graph* or *digraph*. When E contains both undirected and directed edges, it is a *mixed graph*.

Graphs, digraphs, and mixed graphs all serve to model various types of point-to-point networks. However, these alone are insufficient to model network operation and network failure in general.

Node and Link Characteristics

In assessing the operational status of a network, requirements and characteristics for nodes and links may play a central role. For example, links may have a capacity, delay, installation cost, transmission cost, distance, or the like; nodes may have an installation

cost, capacity, rate, and the like. Basic graph models are often augmented by equipping each vertex and each edge with measures of these and other characteristics; to do so, the graph model (V, E) of a network is extended by adjoining functions to specify the relevant characteristics of each vertex and each edge. These characteristics may be a deterministic, fixed property of each node or link, or may be stochastic in nature. The nature of these in turn impacts the notion of network operation.

Network Operation

In the reliability context, the ultimate goal of network modeling is to determine the ability of the network to provide an acceptable level of service. There are numerous ways to quantify network service, and to set thresholds to determine a level at which it is acceptable. For example, for a specification of point-to-point traffic requirements, acceptable service may mean that every pair of nodes can simultaneously be allocated communication paths so that all pairs can communicate an amount at least as large as is dictated by the traffic requirements.

In practical settings, acceptable network service may be a complex function of node and link characteristics. Nevertheless, studies of network reliability have focused primarily on very simple metrics based on connectivity. A requirement for point-to-point traffic to be carried from a node s to a node t is that there is a sequence of nodes starting with s and ending with t , each connected to the next by an edge. This is a *path* in the graph. The mere existence of a path from s to t may not suffice to deliver an acceptable level of service, but the absence of such a path almost surely prevents it. A more complex model that accounts for many characteristics of each node and link on the path, and also its interactions

with other paths, may be much more accurate than one that simply considers the presence or absence of a path. Nevertheless, it is almost certainly much harder to analyze in a useful manner. In the trade-off between the accuracy of the model and the difficulty of analyzing it, most network reliability models err on the side of keeping the model quite simple by focusing just on connectivity. While this may paint an overly optimistic picture of the network's ability to support specific network functions if used too generally, it does address the question of whether the network structure itself can be expected to provide the connections needed; whether these connections are large enough, fast enough, and cheap enough is another matter.

With this in mind, three basic operations are most often considered for bidirectional links. In the *all-terminal* operation, every pair of nodes must be connected; in other words, there is a subset of the edges that are operating which, for each pair of vertices, provides a path between them. Equivalently, one requires that the operating edges contain a *spanning tree* of the graph, which is a connected subgraph in which the number of edges is precisely one less than the number of vertices. In the *two-terminal* operation, two specified nodes s and t must be connected by a path of operating edges. Intermediate between these is the *k-terminal* operation, in which the nodes of some specified set $K \subseteq V$ of size k all lie in the same connected component induced by the operating edges. Variants of these operations can be defined for digraphs and mixed graphs as well.

Causes of Failure

Network design is concerned with provisioning a network with sufficient resources to support its operation. However, failure or degradation of components may result in the inability to support a level of operation that would be possible in the absence of such failures. What causes nodes and links to fail? At one extreme, an intelligent and powerful adversary may destroy certain components selectively in order to disrupt the network. Given sufficient power and resources, such an adversary can destroy the entire network,

and there is no design to overcome this. However, some network structures require more resources on the part of the adversary than do others. Design to make adversarial attack unduly costly is the province of *deterministic network reliability* or *network vulnerability*. At the other extreme, failures may result from random, independent events such as component wear-out. Such events affect nodes and links independently of one another, and their probability of occurrence within a specified time frame can sometimes be estimated from historical data concerning similar nodes or links. In this case, the metric of interest is the probability that a network can support a specified operation when its nodes and links fail independently with known probabilities; this is most frequently called the *network reliability* problem, although *probabilistic* or *stochastic network reliability* might be more accurate.

Neither extreme typifies the behavior of most networks. Indeed, failures often arise from environmental causes, such as power failures or traffic congestion, that disrupt links and nodes in a small geographic region. While such failure events can also be assigned an estimated probability of occurrence, they do not act independently. Rather, they induce a statistical dependence affecting the joint probabilities of link and node failures. This leads to *network reliability with dependent failures*.

Despite the many ways that such dependences arise in the real-world, the focus has been on network vulnerability and network reliability. The reasons for this are manifold. While historical data can often serve to estimate link and node failure probabilities prior to designing a network, dependences are by their very nature a product of the environment in which the network is deployed. Determining the possible failure causes, understanding their dependences, and estimating the likelihoods of each and of simultaneous occurrences of them are each challenging problems. Consequently, one assumes that the worst possible dependences may arise that lead to network vulnerability and deterministic measures, or one assumes that the dependences are relatively inconsequential and that a sufficiently accurate

model is obtained by assuming statistical independence.

When neither assumption is reasonable, *fault-tree analysis* [13] and *binary decision diagrams* [14] can be used to model the dependences. Both require a characterization of causes of failures, which is often not readily available.

Network States and Probabilities

Once a particular network operation is selected, and components are identified whose operation or failure impacts the network operation, a *state* of the network is an assignment of “operational” or “failed” to each of the components. Each state is itself either able to support the desired network operation or is not able to, and is *operational* or *failed* accordingly. Fundamentally, the concern of network reliability is to assess the likelihood that the network is in a state that is operational either at some point in time or during an interval of time in which the operation is to be conducted. Naturally, this depends critically on the causes of component failure. However, when components fail as a result of random, independent events, one can determine the probability that a network is in a specific state S . This is done by employing the probability, for each component subject to failure, that the component has the status indicated in state S . Then, as a consequence of independence, one multiplies all of these component probabilities to determine the probability that the network is in state S .

For this reason, reliability in the context of random, independent failure events has been a primary focus of research in the area, and we shall consider only that situation hereafter.

Network Reliability, Resilience, and Performability

For a given network operation, the *network reliability* can be defined precisely as the sum, over all operational network states, of the probability that the network is in that state. This can be determined by a *complete state enumeration* in which all states are listed; failed states are removed from the listing; probabilities of occurrence are

computed for each operational state; and these probabilities are summed to yield the reliability. In particular, when the operation chosen is all-terminal, k -terminal, or two-terminal, this yields the standard definitions of *all-terminal*, *k-terminal*, or *two-terminal reliability*.

Many related notions have been considered, but we mention only three. The models treated here concern components with two states and network operation with two states. Treating components and networks with more than two states yields *multistate reliability*, to which many of the concepts discussed here readily extend. When components each have two states but the network has possibly many, in addition to the probability of occurrence of each state, one can employ a *performance metric* to specify to what degree the network state supports the network operation. Typically, one chooses a numerical performance metric for which larger values indicate “better” performance. Then, the *expected performance*, or *performability*, of the network is the sum, over all states, of the product of the performance metric of that state with the probability of being in that state. (See Ref. 15, for example.) Standard reliability measures are recovered by specifying that the performance metric itself takes on only value 0 for failed, and 1 for operational.

A third related notion concerns network design to support many network operations. For example, on a network $G = (V, E)$, each choice of $\{x, y\} \subseteq V$ underlies a different two-terminal network operation. Yet, one may wish to support *all* two-terminal operations and not just one. In this case, one can consider the average reliability taken over all network operations of interest, with each selected uniformly at random. In this context, *two-terminal resilience* is the average two-terminal reliability, and *k-terminal resilience* is the average k -terminal reliability [16]. While this appears to involve many reliability measures, one for each operation of interest, it is easily recast as a performability measure as follows. Suppose that there are μ types of network operation. Then for state S , let $\omega(S)$ be the number of these operations supported in state S . Define a

performance metric for state S as $\omega(S)/\mu$. Then the resulting performability measure is exactly the average reliability over the chosen network operations.

In this manner, many reliability, performability, and resilience measures can all be viewed in a single, consistent framework.

STATE-BASED MODELS

Now we proceed more formally, focusing on reliability measures. Let $e_1, \dots, e_m$ be a set of components, which we take to be edges of a graph G . Let $p_1, \dots, p_m$ be probabilities, so that p_i is the probability that e_i is operational. A state of the network is specified as a binary vector of length m , $\Phi = (\phi_1, \dots, \phi_m)$, where ϕ_i is 1 when e_i is operational and 0 when not. Equivalently, a state is a subset of $\{1, \dots, m\}$ that contains the indices of the operating components; subsets and their characteristic vectors are used interchangeably to denote states. The set of all states is therefore just the powerset of $\{1, \dots, m\}$, that is, $2^{\{1, \dots, m\}}$. Assuming independence, the probability $\Pr[\Phi]$ that the network is in state Φ is given by $\prod_{i=1}^m (p_i \phi_i + (1 - p_i)(1 - \phi_i))$.

Any (two-state) network operation can be represented as a subset of the set of all states, $\mathcal{P} \subseteq 2^{\{1, \dots, m\}}$. Members of $\mathcal{P}$ are *pathsets*. Then the reliability $\text{Rel}(G, \mathcal{P})$ is $\sum_{\Phi \in \mathcal{P}} \Pr[\Phi]$. Of course, one could equally well define reliability in terms of failed states. Letting $\bar{\mathcal{P}} = 2^{\{1, \dots, m\}} \setminus \mathcal{P}$, one has that $\text{Rel}(G, \mathcal{P}) = 1 - \sum_{\Phi \in \bar{\mathcal{P}}} \Pr[\Phi]$. One can also define $\mathcal{C} = \{X : X = \{1, \dots, m\} \setminus Y, Y \in \bar{\mathcal{P}}\}$. A subset of $\{1, \dots, m\}$ appears in $\mathcal{C}$ exactly when the *failure* of all edges in the subset leads to a failed state of the network; hence members of $\mathcal{C}$ are *cutsets*.

Although these state-based definitions of reliability provide a precise definition, the obvious algorithm to compute them involves the consideration of 2^m states, and is infeasible for large networks. It is therefore natural to ask whether more sophisticated methods can provide algorithms to compute reliability in time bounded by a function whose growth rate is less than exponential in m . We treat the (mostly negative) results next, and later address methods for circumventing the lack of efficient algorithms.

Computational Complexity

Computational complexity addresses the resources such as computational time or space needed to solve specific problems. Although a specific problem is given, we are concerned with methods that solve *each* instance of the problem, no matter how large, in a finite amount of time. The *size* of an instance is usually specified to be the amount of space needed to represent the instance using a finite alphabet. Let $T(n)$ represent the largest amount of time taken for an algorithm to solve an instance of size n (so that we are concerned with *worst case complexity*). Often, $T(n)$ is a complicated function whose exact values are not known. Nevertheless, we can make a clear distinction between functions that grow faster than any polynomial function of n and those that are bounded by some fixed polynomial. Algorithms whose running times are bounded by a polynomial are *efficient* or *polynomial-time* algorithms, and much of the design of algorithms is concerned with finding efficient algorithms for specific problems.

We face two difficult issues in exploring reliability. First, what is the size of the input? To specify an n -vertex, m -edge graph, it suffices to employ $O(m \log n)$ bits. However, to specify the network operation of concern, we must decide whether to represent it explicitly (for example, by providing $\mathcal{P}$) or implicitly (for example, by saying that for two vertices s and t the state contains an s, t -path). If we adopt an explicit representation, the size of the input alone can be $O(2^m)$, and hence the input is by itself exponential in the size of the graph, and any hope for an algorithm whose running time is efficient (in the size of the network) vanishes. Therefore, essentially all effort to date concerns reliability measures for which the description of the network operation requires $O(m \log n)$ space. We make this restriction here, but note that in the process we still consider all-terminal, k -terminal, and two-terminal reliability.

The second issue that we face is how to proceed when the best current efforts have failed to produce an efficient algorithm. Naturally, we would like to establish that none exists. However, known techniques for establishing lower bounds on the complexity of

problems are much weaker in general than those for establishing upper bounds. While a single algorithm suffices to establish an upper bound, one must establish that none of the possible algorithms is efficient in order to establish the desired lower bound. Nevertheless, machinery is well developed for providing compelling evidence that a problem admits no efficient algorithm, through the theory of NP-hardness and #P-hardness. The class #P contains counting problems, while NP contains decision (“yes”/“no”) problems; all #P-hard problems are NP-hard [17]. NP-hardness is usually taken to indicate that no efficient algorithm exists for the problem.

For network reliability, many problems have been shown to be #P-hard, including all-terminal, two-terminal, and k -terminal reliability. Although these results effectively close off the possibility of certain types of algorithms, other avenues open. While the complexity results indicate that there are *some* hard instances of these reliability problems, it does not preclude the possibility that an algorithm exists that solves “most” instances efficiently, which with luck may include most instances arising in practice. Moreover, two exponential time algorithms can exhibit quite different execution times on the same instance, and therefore practical improvements on complete state enumeration are of significant concern.

It is also sensible to relax the requirements of the problem. While network reliability is a precise numerical quantity (once the component probabilities and network operation are specified), numerous assumptions and simplifications led to its definition. Therefore, one might consider obtaining a point estimate of the reliability, along with confidence intervals, using a Monte Carlo strategy. For certain environments, the probabilistic guarantee obtained for the reliability may be insufficient. Nevertheless, in these cases obtaining absolute lower and upper bounds may suffice to ascertain whether a particular network design meets, or does not meet, application requirements. Finally, one might characterize the networks that arise in practice; it may be possible that they share a network structure that admits an efficient algorithm, despite

the expectation that arbitrary network structures do not. Each of these avenues has been pursued extensively; in the remainder, we provide a brief outline of each.

Pivotal Decomposition or Factoring

Suppose that $\mathcal{P}$ defines a set of pathsets, and that e is a component (edge) of the network in question. Define the *deletion* $\mathcal{P} \setminus e$ to be the set $\{P : e \notin P \in \mathcal{P}\}$, that is, the collection of all pathsets that do not contain e . Define the *contraction* $\mathcal{P}/e$ to be the set $\{P \setminus \{e\} : e \in P \in \mathcal{P}\}$, that is, the collection of all pathsets that do contain e , with e then removed. For a graph G with an edge e , $G \setminus e$ is the result of deleting edge e from G , while G/e is the result of identifying the two endnodes of e and then removing the edge e . With this notation and assuming independence, it follows that

$$\begin{aligned} \text{Rel}(G, \mathcal{P}) &= (1 - p_e) \cdot \text{Rel}(G \setminus e, \mathcal{P} \setminus e) \\ &\quad + p_e \cdot \text{Rel}(G/e, \mathcal{P}/e). \end{aligned}$$

This is the *factoring formula*, sometimes called *pivotal decomposition* in which e is the *pivotal element*. When carried through until no edge remains, this effectively reproduces complete state enumeration. However, it forms the basis for many useful algorithms. The primary reason is that, while G may have a complex structure, after certain deletions and contractions the resulting graph may have a simpler structure whose reliability can be calculated by more direct methods. When this occurs, there is no need to apply factoring further to the intermediate graph. In a similar vein, although we may not be able to apply reliability-preserving transformations to simplify the graph itself, deletions and/or contractions may lead to the possibility of applying them at intermediate stages.

Perhaps the greatest benefit of factoring is that, except for independence, no assumptions are required about the graph or the model of network operation.

Transformations and Reductions

All methods discussed are quite sensitive to the size of the instance, that is, to the number

of edges. Hence techniques to remove edges when possible are of critical concern. The easiest case is when an edge e is *irrelevant*, in that no pathset contains e ; in this case, $\text{Rel}(G, \mathcal{P}) = \text{Rel}(G \setminus e, \mathcal{P} \setminus e)$. For all-terminal, k -terminal, and two-terminal reliability, determining when an edge is irrelevant is straightforward; however, for general notions of network operation, this determination can itself be NP-hard. When an edge e appears in every pathset, e is *mandatory*, and $\text{Rel}(G, \mathcal{P}) = p_e \cdot \text{Rel}(G/e, \mathcal{P}/e)$.

Suppose now that every edge is neither irrelevant nor mandatory. Two edges e_1 and e_2 are *complements* if, for every $P \in \mathcal{P}$, either $\{e_1, e_2\} \in P$ or P contains neither. A reliability-preserving reduction is then to contract e_2 , and replace the operation probability of e_1 by $p_{e_1}p_{e_2}$. A specific case for graphs arises when e_1 and e_2 share a node that is not a target; then this is a *series reduction*. Two edges e_1 and e_2 are *substitutes* when for every $P \in 2^{\{3, \dots, m\}}$, $P \cup \{e_1\} \in \mathcal{P}$ if and only if $P \cup \{e_2\} \in \mathcal{P}$. A reliability-preserving reduction is then to delete e_2 , and replace the operation probability of e_1 by $1 - (1 - p_{e_1})(1 - p_{e_2})$. A specific case for graphs arises when e_1 and e_2 both connect the same two nodes; then this is a *parallel reduction*.

Essentially, all methods reduce the graph by treating irrelevant and mandatory edges, and by series and parallel reductions. However, a wide variety of other reductions have been examined as well.

Approximation: Monte Carlo Methods

Despite simplifying reductions, often the number of states is too large to enumerate. When this occurs, an estimate of the reliability can be obtained by selecting states at random, and determining whether that state is operational. This forms the basis for *crude Monte Carlo* algorithms. Indeed, when in each trial a network state is chosen by selecting each component e to operate with probability p_e , an unbiased estimate of the reliability is simply the ratio of the number of trials leading to operating network states to the number of trials performed. Crude Monte Carlo can be effective for small networks, particularly when the network operation of concern does not exhibit structure that can

be exploited to limit the sampling needed. More sophisticated Monte Carlo techniques employ the structure of operational and failed states to reduce the number of samples. The goal of a Monte Carlo method is to produce an unbiased estimator of the reliability, and to produce useful statistical confidence intervals on this estimate within a “small” number of trials.

Most Probable States

Of particular practical concern are networks in which every edge has an operation probability close to 1. In this setting, crude Monte Carlo tends to select operational network states repeatedly, only very rarely encountering a failed network state. Consequently, numerous trials are typically needed to produce useful confidence intervals. Indeed, the same network state is often generated repeatedly. To avoid this, one could sample the network states without replacement, by maintaining a library of states already encountered. An advantage is that one can establish not just a confidence interval, but also *absolute* lower and upper bounds on the reliability, by accumulating the probabilities of the failed and operational states encountered thus far. But the disadvantage is that there is no particular structure in the states so far encountered, because they have been randomly generated.

Instead, *most probable state* methods abandon the search for a point estimate and confidence interval, and instead enumerate states in nonincreasing order of probability of occurrence [18]. The key is to maintain a data structure from which the next most probable state to be generated can be easily determined without explicitly examining states already generated. Improvements are possible when the operational status of a class of network states can be determined from the operational state of a single state.

PATH- AND CUT-BASED METHODS

When the network operation is arbitrary, every method essentially must either enumerate all states (perhaps in a certain order),

or sample from all states, because the operational status of one network state cannot determine the operational status of another. In many practical environments, however, this is not the case at all! Consider a set of pathsets $\mathcal{P}$, and a particular pathset $P \in \mathcal{P}$. When $e \notin P$, can it happen that $P \cup \{e\} \notin \mathcal{P}$? If this occurs, it indicates that, when the components of P operate and the remaining components fail, the network operates; however, when the state of e changes from failed to operating, the state of the network changes from operating to failed. In all-, k -, and two-terminal reliability, this cannot occur.

Coherence of $\mathcal{P}$ is the property that whenever $P \in \mathcal{P}$ and $P \subseteq P'$, we find that $P' \in \mathcal{P}$. Equivalently, if $\mathcal{C}$ is the set of cutsets, then whenever $C \in \mathcal{C}$ and $C \subseteq C'$, we find that $C' \in \mathcal{C}$. Not all natural reliability problems are coherent. For example, if in a graph we take the vertices as the components, and the network operation is that “all operating vertices are connected,” then when only one vertex operates the network is operational, but when two operate, the network is operational only when the two are connected by an edge. Nevertheless, the focus of efforts on reliability has been on coherent problems.

For coherent problems, rather than considering all states, all pathsets, or all cutsets, one can consider much smaller sets of states. A *minpath* is a pathset $P \in \mathcal{P}$ for which there is no $e \in P$ with $P \setminus \{e\} \in \mathcal{P}$. The notation $\mathcal{MP}$ is used for the set of all minpaths. Similarly, a *mincut* is a cutset $C \in \mathcal{C}$ for which there is no $e \in C$ with $C \setminus \{e\} \in \mathcal{C}$. The notation $\mathcal{MC}$ is used for the set of all mincuts. Knowing $\mathcal{MP}$, one can recover $\mathcal{P}$ by taking all supersets of the minpaths; the same holds for cutsets. For example, with all-terminal reliability a minpath is a spanning tree, and all connected subgraphs can be found by adding zero or more edges to a spanning tree.

Boolean Techniques

Let $e_1, \dots, e_m$ be a set of components and $x_1, \dots, x_m$ be the events that they operate. For a particular minpath $P \in \mathcal{MP}$, the event that all components of P operate is $\bigwedge_{e_i \in P} x_i$. Then the event that the network operates is just $\bigvee_{P \in \mathcal{MP}} \bigwedge_{e_i \in P} x_i$. By the same token, for a particular mincut $C \in \mathcal{MC}$, the event that all

components of C fail is $\bigwedge_{e_i \in C} \bar{x}_i$, and the event of network operation is $\bigvee_{C \in \mathcal{MC}} \bigwedge_{e_i \in C} \bar{x}_i$.

While each provides a simple Boolean expression for reliability in terms of minpaths or mincuts, the computational difficulty arises from the fact that, for two minpaths P_1 and P_2 , the events $\bigwedge_{e_i \in P_1} x_i$ and $\bigwedge_{e_i \in P_2} x_i$ are independent only when P_1 and P_2 share no components. Express the event that at least one of the pathsets P_1 and P_2 operate by $(\bigwedge_{e_i \in P_1} x_i) \vee ((\bigwedge_{e_i \in P_1} \bar{x}_i) \wedge (\bigwedge_{e_i \in P_2} x_i))$. The two events in the disjunction are mutually exclusive, or *disjoint*. Hence the probability of either occurring is just the sum of the probabilities of each occurring individually. Extending this to many minpaths (or mincuts), network reliability can be written as a disjunction of disjoint events. In the process, however, expressions of the form $E = (\bigwedge_{e_i \in P_j} x_i) \wedge \bigwedge_{\ell=1}^{j-1} (\bigwedge_{e_i \in P_\ell} \bar{x}_i)$ now arise in which the events $\bigwedge_{e_i \in P_j} x_i$ and $\bigwedge_{e_i \in P_s} \bar{x}_i$ are neither disjoint nor independent. Nevertheless, we can expand E so that every term is a *product*, that is, a conjunction of basic variables and their negations, and all products are disjoint. Then logical simplification can remove terms that contain $x_i \bar{x}_i$, for example, and amalgamate two terms of the form $x_i \wedge T$ and $\bar{x}_i \wedge T$ to form the term T . The result is a disjunction (“sum”) of terms, each of which is a conjunction (“product”), and every two of the products are disjoint. This is a *sum-of-disjoint-products* form.

In general, the goal of Boolean methods is to find a “short” Boolean expression for the event of network operation, whose numerical probability of occurrence can be calculated directly. Many different sum-of-disjoint-product forms exist, and effort typically focuses on selecting one that can be both easily produced and easily evaluated.

Inclusion–Exclusion

When $\mathcal{MP} = \{P_1, \dots, P_h\}$ is the set of minpaths, we have seen that the event of network operation is $\bigvee_{P \in \mathcal{MP}} \bigwedge_{e_i \in P} x_i$. Reliability is the probability that this event occurs, that is, $\Pr[\bigvee_{P \in \mathcal{MP}} \bigwedge_{e_i \in P} x_i]$. In words, this simply states that at least one minpath is operating. For a specific minpath P_j , its probability of operation $\Pr[P_j] = \prod_{e_i \in P_j} p_i$.

Consider computing $\sum_{j=1}^h \Pr[P_j]$. Certainly, every event in which a minpath operates is accounted for in this sum. Unfortunately, the event in which P_j and P_ℓ both operate is included twice in the sum. To attempt to correct this, we calculate the probability that both operate, $\Pr[P_j \wedge P_\ell] = \prod_{e_i \in P_j \cup P_\ell} p_i$, and for all distinct choices of j and ℓ subtract this correction from the sum. However, now when three minpaths P_j , P_k , and P_ℓ all operate, the event was included three times in the initial sum (once each for j , k , and ℓ) and then subtracted three times in the correction (once each for $\{j, k\}$, $\{j, \ell\}$, and $\{k, \ell\}$), and a further addition is needed.

Carrying this to its logical conclusion, for a subset S of $\{1, \dots, h\}$, let $\mathcal{MP}_S$ be the set of minpaths indexed by S , let π_S be the set $\cup_{j \in S} P_j$, and compute $\Pr[\pi_S] = \prod_{e_i \in \pi_S} p_i$. Repeatedly applying inclusion/exclusion as above, we obtain $\Pr[\bigvee_{P \in \mathcal{MP}} \bigwedge_{e_i \in P} x_i] = \sum_{S \subseteq \{1, \dots, h\}} (-1)^{|S|+1} \Pr[\pi_S]$. Naturally, it is not desirable to expand this expression out completely, and hence one typically attempts to simplify this expression both logically and arithmetically. One very important feature of inclusion/exclusion methods is that, if we consider only those subsets S that have a fixed maximum size m , we obtain an absolute bound on the reliability. Indeed, when m is odd, we obtain an upper bound, and when m is even we obtain a lower bound. Treating ever larger values of m causes these bounds to converge to the true reliability.

Reliability Polynomials

Whether Boolean methods, inclusion–exclusion, or state-based methods are used, the sets of states, minpaths, or mincuts are typically examined explicitly. As a result, often a further simplifying assumption is made. Suppose that every edge operates with the same probability p . Then, every state in which exactly i edges operate and $m - i$ fail arises with the same probability $p^i(1 - p)^{m-i}$ as every other. This suggests that we may not need to examine each operating state, but instead compute the number N_i of operating states with i edges. If we can count operating states in this way, the reliability is given by $\sum_{i=0}^m N_i p^i (1 - p)^{m-i}$. This is one form of the *reliability polynomial*.

Unless we can compute or bound the coefficients $\{N_i\}$ without essentially listing all operating network states, this is merely a device for simplifying the expression, not one for accelerating its computation. To accomplish the latter, one exploits combinatorial structure. When $\mathcal{P}$ is the set of all pathsets, let $\mathcal{F} = \{e_1, \dots, e_m\} \setminus P : P \in \mathcal{P}$. Then $\mathcal{F}$ contains all complements of pathsets. For a coherent reliability problem, $\mathcal{F}$ is closed under taking subsets; that is, it is a *complex* or *hereditary family of sets*. Now, $F_i = N_{m-i}$ is the number of pathset complements that contain exactly i edges, and hence the reliability is $\sum_{i=0}^m F_i p^{m-i} (1 - p)^i$.

An easy example serves to illustrate the use of combinatorial structure. Intuitively, for a coherent system, when more edges operate, the network is more likely to operate. Indeed, an old theorem of Sperner [19] states that $\frac{F_i}{\binom{m}{i}} \geq \frac{F_{i+1}}{\binom{m}{i+1}}$. Simplifying, we obtain that $F_i \geq \frac{m-i}{i+1} F_{i+1}$ or that $F_{i+1} \leq \frac{i+1}{m-i} F_i$. While these inequalities do not determine the coefficients $(F_0, \dots, F_m)$, they do constrain them significantly.

In this vein, effort has focused on using additional structure of reliability complexes, and on the determination of specific coefficients. For the former, Kruskal [20] and Katona [21] establish the tightest inequality for F_i and F_{i+1} in a general complex. Most reliability problems do not lead to arbitrary complexes, and hence better inequalities have been developed for polyhedral complexes, shellable complexes, and matroidal complexes. Indeed, for all-terminal reliability, $\mathcal{F}$ is a representation of the *cographic matroid* of the graph, which permits the use of a substantial body of knowledge concerning matroids [22].

Effort to determine coefficients has been directed at efficient algorithms, so that upper and lower bounds on reliability measures can be efficiently calculated.

RESTRICTED CLASSES OF NETWORKS

A final important concept is the treatment of reliability problems by restricting the structure of the network itself, rather than by restricting the network operation or

relaxing the requirements on what is to be computed. The prototypical example here is given by the series-parallel networks, those networks that can be reduced to a single edge by deletion of irrelevant edges, contraction of mandatory edges, and series and parallel reductions. These transformations underlie a linear-time algorithm for computing two-terminal reliability, and easy variants of them extend efficient algorithms to k -terminal reliability and k -resilience.

Despite these successes with series-parallel networks, success with other classes arising in practice has been limited. The natural generalization to graphs of bounded treewidth leads to more general, efficient algorithms [23] but in general networks that are planar and/or have bounded maximum degree remain $\#P$ -hard [4]. Nevertheless, approximation and bounding techniques may exploit such additional structure effectively, as for example in the use of Delta-Wye transformations in the estimation of reliability for planar networks [24].

Finally, improved methods for restricted classes of graphs can be used to limit the number of factoring steps in pivotal decomposition, to simplify portions of a Boolean expression or avoid the need for certain terms in an inclusion-exclusion expansion, or to further constrain the values of coefficients in reliability polynomials. Indeed, the application of the many methods outlined here together often provides a more thorough analysis than any of the individual methods provides in isolation.

REFERENCES

1. Ball MO, Colbourn CJ, Provan JS. Network reliability. In: Ball MO, Monma CL, Magnanti TL, editors. *Network models*. Amsterdam: Elsevier Science; 1995. pp. 673–762.
2. Colbourn CJ. Reliability issues in telecommunications network planning. In: Sanso B, Soriano P, editors. *Telecommunications network planning*. Amsterdam: Kluwer; 1999. pp. 135–146.
3. Van Slyke RM, Frank H. Network reliability analysis: Part I. *Networks* 1972;1:279–290.
4. Jaeger F, Vertigan D, Welsh DJA. On the computational complexity of the Jones and Tutte polynomials. *Math Proc Camb Phil Soc* 1990;108:35–53.
5. Karger DR. A randomized fully polynomial time approximation scheme for the all terminal network reliability problem. *SIAM J Comput* 1999;29:492–514.
6. Provan JS, Ball MO. The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM J Comput* 1983;12:777–788.
7. Fishman GS. A comparison of four Monte Carlo methods for estimating the probability of (s,t) -connectedness. *IEEE Trans Reliab* 1986;R-35:145–155.
8. Gertsbakh IB, Shpungin Y. *Models of network reliability: analysis, combinatorics, and Monte Carlo*. Boca Raton (FL): CRC Press; 2009.
9. Colbourn CJ. Boolean aspects of network reliability. In: Hammer PL, Crama Y, editors. *Boolean models and methods*. In press.
10. Provan JS. Boolean decomposition schemes and the complexity of reliability computations. In: Roberts FS, Hwang FK, Monma CL, editors. *Reliability of computer and communications networks*. Providence (RI): AMS/ACM; 1991. pp. 213–228.
11. Shier DR. *Network reliability and algebraic structures*. New York: Oxford University Press; 1991.
12. Colbourn CJ. *The combinatorics of network reliability*. New York: Oxford University Press; 1987.
13. Barlow RE, Fussell JB, Singpurwalla ND. *Reliability and fault tree analysis*. Philadelphia (PA): SIAM; 1975.
14. Akers SB. Binary decision diagrams. *IEEE Trans Comput* 1978;C-27:509–516.
15. Meyer JF. Performability: a retrospective and some pointers to the future. *Perform Eval* 1992;14:139–156.
16. Farley TR, Colbourn CJ. Multiterminal network connectedness on series-parallel networks. *Disc Math Algor Appl* 2009;1:253–265.
17. Valiant LG. The complexity of enumeration and reliability problems. *SIAM J Comput* 1979;8:410–421.
18. Yang C-L, Kubat P. Efficient computation of most probable states for communication networks with multimode components. *IEEE Trans Commun* 1989;COM-37:535–538.
19. Sperner E. Über einen kombinatorischen Satz von Macaulay und seine Anwendung auf die Theorie der Polynomideale. *Abh Math Semin Univ Hamburg* 1930;7:149–163.

20. Kruskal JB. The number of simplices in a complex. In: Bellman R, editor. Mathematical optimization techniques. Berkeley (CA): University of California Press; 1963. pp. 251–278.
21. Katona G. A theorem of finite sets. In: Erdős P, Katona G, editors. Theory of graphs. Budapest: Akademia Kiadó; 1966. pp. 187–207.
22. Brown JI, Colbourn CJ, Nowakowski RJ. Chip firing and all-terminal network reliability bounds. *Disc Optim* 2009;6:436–445.
23. Wolle T. A framework for network reliability problems on graphs of bounded treewidth. *Lect Notes Comput Sci* 2002;2518:401–420.
24. Feo TA, Provan JS. Delta-wye transformations and the efficient reduction of two-terminal planar graphs. *Oper Res* 1993;41: 572–582.

CONCEPTUAL MODELING FOR SIMULATION

STEWART ROBINSON
Warwick Business School,
University of Warwick,
Coventry, UK

In broad terms, conceptual modeling is the process of abstracting a model from the real world. The modeler is presented with a problem situation that is amenable to simulation modeling, based on which he/she has to determine what aspects of the real world to include, and exclude, from the model, and at what level of detail to model each aspect. These decisions should generally be a joint agreement between the modeler and the problem owners, that is, the stakeholders who require the model to aid decision making.

This article provides an overview of the field of conceptual modeling for simulation. The article starts by describing what is meant by conceptual modeling after which various definitions of the term *conceptual model* are explored. Approaches to developing conceptual models are then discussed, which covers the requirements of a conceptual model, principles of modeling, conceptual modeling frameworks, and methods of model simplification. Finally, ideas for future research in the field of simulation conceptual modeling are provided. Throughout the article the focus is on conceptual modeling for discrete-event simulation [1], although the concepts are also useful to other forms of simulation modeling and modeling more generally.

WHAT IS CONCEPTUAL MODELING?

To understand conceptual modeling it is useful to set it within the wider context of the modeling process for simulation. Figure 1 shows the key artifacts of conceptual modeling [note: an “artifact” is something “made by human workmanship” (Chambers Twentieth

Century Dictionary)]. The “cloud” represents the real world (current or future) within which the problem situation resides; this is the problem that is the basis for the simulation study. The four rectangles represent specific artifacts of the (conceptual) modeling process. These are as follows:

- *System Description*. A description of the problem situation and the system in which the problem situation resides.
- *Conceptual Model*. “The conceptual model is a non-software-specific description of the computer simulation model (that will be, is, or has been developed), describing the objectives, inputs, outputs, content, assumptions, and simplifications of the model.” (2, p. 283).
- *Model Design*. The design of the constructs for the computer model (data, components, model execution, etc.) [3].
- *Computer Model*. A software-specific representation of the conceptual model.

These artifacts are quite separate. This is not to say that they are always explicitly expressed, with the exception of the computer model. For instance, the system description, conceptual model, and model design may not be (fully) documented and can remain within the minds of the modeler and the problem owners. It is, of course, good modeling practice to document each of these artifacts.

The model design and computer model are not strictly part of conceptual modeling, but they do embody the conceptual model within the design and code of the model. These artifacts are included in Fig. 1 for completeness. Our main interest here is in the system description and conceptual model that make up the process of conceptual modeling; as represented by the shape with a dashed outline in Fig. 1. It should also be noted that the system description and conceptual model are independent of any specific software. It is only the model design and computer model

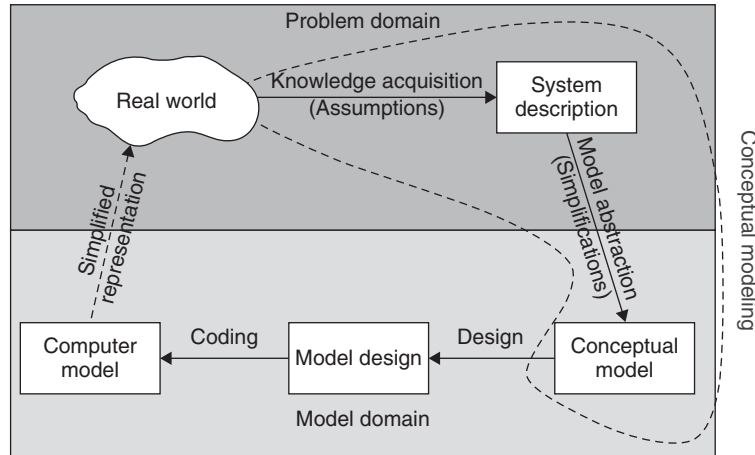


Figure 1. Artifacts of conceptual modeling. [Source: adapted from Kotiadis and Robinson (4, p. 952)].

that are specific to the software chosen for the implementation of the conceptual model.

It is important to recognize the distinction between the system description and the conceptual model. The system description relates to the problem domain, that is, it describes the problem and the real world within which the problem resides. The conceptual model belongs to the model domain in that it describes those parts of the system description that are included in the simulation model and at what level of detail to model. The author's experience is that these two artifacts are often confused and seen as indistinct.

The definitions here are close to those used by Zeigler [5]. He describes five concepts:

- *The Real System.*
- *The Experimental Frame.* The limited set of circumstances under which the real system is observed or experimented with (i.e., specific input–output behaviors).
- *The Base Model.* A model that is capable of accounting for all the input–output behavior of the real system (it cannot be fully known).
- *The Lumped Model.* A simplified model with components lumped together and interactions simplified (it can be fully known and is valid within a chosen experimental frame).

- *The Computer.* The computational device for implementing the lumped model.

Although the correspondence is not exact, the system description in Fig. 1 roughly corresponds to the base model, the difference being that the base model cannot be fully known while the system description embodies all that is known about the real system; it is a subset of the base model. Meanwhile, the conceptual model corresponds to the experimental frame and lumped model together. It is the objectives of the model that define the input–output behaviors (experimental frame) of interest and the content of the model that embodies the simplifications defined by the lumped model; note that objectives and content are central to the definition of a conceptual model. Zeigler is also very clear that the model and the computer (program) are quite distinct. In the same way, Fig. 1 shows that the conceptual model is separate from the model design and computer model.

The arrows in Fig. 1 represent the flow of information; for instance, information about the real-world feeds into the system description. The processes that drive the flow of information are described as knowledge acquisition, model abstraction, design, and coding. The arrows are not specifically representative of the ordering of the steps

within the modeling process, which we know are highly iterative [6,7]. In other words, a modeler may return to any of the four processes at any point in a simulation study, although there is some sense of ordering in that information from one artifact is required to feed the next artifact.

The dashed arrow shows that there is a correspondence between the computer model and the real world. The degree of correspondence depends on the degree to which the model contains assumptions that are correct, the simplifications maintain the accuracy of the model, and the computer code is free of errors. Because the model is developed for a specific purpose, the correspondence with the real world relates only to that specific purpose. In other words, the model is not a general model of the real world, but a simplified representation developed for a specific purpose. The issue of whether the level of correspondence between the model and the real world is sufficient is an issue of verification and validation [6,8–10]. Both conceptual modeling and validation are concerned with developing a simulation of sufficient accuracy for the purpose of the problem being addressed. As a result, there is a strong relationship between the two topics, conceptual modeling being concerned with developing an appropriate model and validation being concerned with whether the model is appropriate.

Knowledge Acquisition and Model Abstraction in Conceptual Modeling

Our attention now turns to concentrate on the specific processes that lead to the development of a conceptual model: knowledge acquisition and model abstraction. The system description is obtained through knowledge acquisition. Knowledge and information about the real world is acquired from subject matter experts (or domain experts) and observations. The conceptual model is obtained through model abstraction. The modeler and problem owners jointly agree on what parts of the system description to model and at what level of detail. We now explore knowledge acquisition and model abstraction in some more detail.

Knowledge Acquisition. Because the real world is not fully known or knowable, the system description is only a partial representation of the real world. There are limits to the knowledge about the real world because of the following points:

- The real world has not been observed in all possible states. If the system exists, it will not have been in every state possible and so cannot have been observed in every state. In many cases, the system does not yet exist; the only state in which it has been observed may be a design drawing.
- Observations about the real world are subject to error. Errors frequently occur in data collected from real systems, particularly where humans are recording the information.
- Observations about the real world are incomplete. Observers will not have been able to record every aspect of the state of the system, and often such information is very limited.
- Observations are subject to observer perceptions. Different observers may interpret events in a system differently. Hence, there may be multiple accounts of the same phenomenon.

Further to this, the nature of the problem situation implies that there are a limited set of modeling objectives. It is desirable to develop a model that addresses only those objectives, rather than a general model of the system. Among other benefits (the section titled “Model Abstraction”), this saves time and reduces data requirements. Hence, the system description (and conceptual model) need focus only on the parts of the real world that are relevant to the problem situation and the modeling objectives.

Because the real world is not fully known or knowable, *assumptions* must be made concerning the real world. “*Assumptions* are made when there are either uncertainties or beliefs about the real world being modelled” (2, p. 283). In general, assumptions are made by the problem owners in consultation with the modeler.

It is good practice to document assumption and assess them for the confidence that can be placed in them and their likely impact on the performance of the real system. Critical assumptions (low confidence, high impact) can be assessed later with the model by performing sensitivity analysis.

Model Abstraction. Model abstraction is important because it is not desirable to model all that is known about the real world, even that which is relevant to the problem situation and modeling objectives. The benefits of simpler models are well documented [11–16]:

- simple models can be developed faster;
- simple models are more flexible;
- simple models require less data;
- simple models run faster;
- the results are easier to interpret since the structure of the model is better understood.

Through abstraction the conceptual model becomes a partial representation of the system description. This is achieved by reducing the scope of the conceptual model from that of the system description and/or by reducing the level of detail in the conceptual model from that of the system description. Both of these imply a process of *simplification*. “*Simplifications* are incorporated in the model to enable more rapid model development and use, and to improve transparency” of the model (2, p. 283). The process of simplification should focus on maintaining sufficient accuracy for addressing the problem situation/modeling objectives.

In general, simplifications are made by the modeler in consultation with the problem owners. It is a good practice to document all simplifications and to assess them for their likely impact on the accuracy of the model. It is unlikely, of course, that high impact simplifications would be appropriate as this implies a model which deliberately embodies significant inaccuracies.

Summary

In understanding conceptual modeling it is important to recognize two distinct elements.

First, within the problem domain is the need to acquire knowledge about the real world and to derive a system description. This process entails making assumptions. Second, within the model domain is the need to abstract a conceptual model from the system description. Within this process, simplifications are made.

As discussed above, the conceptual model may, or may not, be formally expressed and documented. What is certain is that the modeler always carries out the conceptual modeling process in a simulation study; in other words, the modeler makes decisions about the scope and level of detail of the model. Whether this is made explicit (it is documented in some fashion) depends on the level of formality in the modeling process.

A final point is that although we might consider there to theoretically be a right conceptual model for a specific simulation study, it is extremely unlikely that this model could be identified in practice. We might propose that the “right conceptual model” is the model that generates exactly the required level of accuracy with the minimum modeling effort. In practice, we can think only in terms of conceptual models that are better and worse, that is, more or less likely to provide the required level of accuracy with a reasonable amount of effort. The aim of a modeler should be to identify the best conceptual model possible given the constraints of knowledge, data, resource, and time.

DIFFERING OPINIONS ABOUT THE DEFINITION OF A CONCEPTUAL MODEL

Above, the definition for a conceptual model used is “the conceptual model is a non-software-specific description of the computer simulation model (that will be, is, or has been developed), describing the objectives, inputs, outputs, content, assumptions, and simplifications of the model” (2, p. 283). This is by no means an agreed definition. It is hard to find specific definitions for the term *conceptual model* in the simulation and modeling literature, although some authors do discuss the term more generally. Much of this discussion relates to military simulation modeling,

which by nature tends to entail large-scale (and sometimes distributed) models.

Pace defines a conceptual model as “a simulation developer’s way of translating modeling requirements . . . into a detailed design framework . . . , from which the software that will make up the simulation can be built” (17, p. 1). In short, the conceptual model defines what is to be represented and how it is to be represented in the simulation. Pace sees conceptual modeling as being quite narrow in scope, viewing objectives and requirements definition as precursors to the process of conceptual modeling. The conceptual model is largely independent of software design and implementation decisions. Pace [18] identifies the information provided by a conceptual model as consisting of assumptions, algorithms, characteristics, relationships, and data.

Lacy *et al.* [19] further this discussion reporting on a meeting of the Defence Modeling and Simulation Office (DMSO) to try and reach a consensus on the definition of a conceptual model. The paper describes a plethora of views, but concludes by identifying two types of conceptual model: a *domain-oriented* model that provides a detailed representation of the problem domain (similar to the system description in Fig. 1) and a *design-oriented* model that describes in detail the requirements of the model (similar to the conceptual model in Fig. 1). The latter is used to design the model code. Meanwhile, Haddix [20] points out that there is some confusion over whether the conceptual model is an artifact of the user or the designer. This may, to some extent, be clarified by adopting the two definitions of Lacy *et al.* above.

Borah offers the following definition “a simulation conceptual model is an abstraction from either the existing or a notional physical world that serves as a frame of reference for further simulation development by documenting simulation-independent views of important entities and their key actions and interactions. A simulation conceptual model describes what the simulation will represent, the assumptions limiting those representations, and other capabilities needed to satisfy the stakeholder’s requirements.

It bridges between these requirements, and simulation design.” (21, p. 3).

Meanwhile, Balci [22] discusses the role of a conceptual model in model reuse. In this context, he sees a conceptual model as a repository of knowledge about a particular problem domain that can be reused by a community of interest for developing a series of models.

The key areas of agreement in these discussions and definitions are that the conceptual model in some way describes the simulation model and that it is independent of the model code or software.

One specific area where the definition of a conceptual model used in this article differs from others is with the inclusion of the objectives of the model as part of the conceptual model. Others would consider the model objectives to be distinct from the conceptual model. Whereas the author agrees that the task of setting the modeling objectives can be separated from the task of defining the model, the artifacts of modeling objectives and conceptual model are not separate; the conceptual model cannot exist without the objectives of that model. In other words, there is a difference between defining an artifact and the task(s) that lead to the creation of that artifact. Our attention now turns to considering the task of developing a conceptual model.

DEVELOPING CONCEPTUAL MODELS

The process of developing a conceptual model is one of determining the content of the simulation model, what to include in and exclude from a model. This can be split into two components, the scope and level of detail (2, p. 283):

- *The Scope of the Model.* the model boundary or the breadth of the real system that is to be included in the model.
- *The Level of Detail.* the detail to be included for each component in the model’s scope (in effect, the level of abstraction).

The question is how does a simulation modeler decide what the scope and level of

detail of a model should be? This question is not straightforward especially as conceptual modeling is very much more an art than a science. Indeed, Henriksen [23] identifies the need for creativity in the conceptual modeling process. Schmeiser (24, p. 40) points out that “while abstracting a model from the real world is very much an art, with many ways to err as well as to be correct, analysis of the model is more of a science, and therefore easier, both to teach and to do.” The need for creativity does not, however, excuse the need for guidelines on how to model [25]. Ferguson *et al.* (26, p. 24), writing about software development, point out that in “most professions, competent work requires the disciplined use of established practices. It is not a matter of creativity versus discipline, but one of bringing discipline to the work so creativity can happen.”

So what guidance is available to aid the conceptual modeling process? In sum, the guidance can be placed into four categories: requirements of a conceptual model, principles of modeling, conceptual modeling frameworks, and methods of model simplification. Each of these is now discussed in turn.

Requirements of a Conceptual Model

The key requirement is to avoid an overly complex model. As per Occam’s razor, *plurality should not be posited without necessity* (William of Occam; quoted from Ref. 27), the conceptual modeling mantra is one of developing the simplest model possible to meet the objectives of the simulation study. Figure 2 aims to demonstrate the reasons for this. This

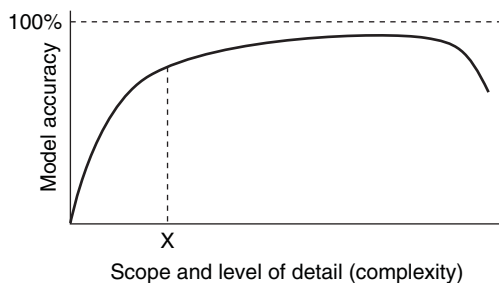


Figure 2. Simulation model complexity and accuracy. [Source: Taken from Robinson (2, p. 287).]

shows a sketch based on experience in which increasing levels of complexity (scope and level of detail) lead to diminishing returns in terms of model accuracy. The model will never be 100% accurate, as this is impossible to achieve; all models are simplifications. Eventually, it is argued that increased complexity leads to reduced model accuracy. This is because neither knowledge nor the data are available to support this level of complexity. For instance, it is not possible to model every nuance of human interaction in a service system, during the evacuation of a building, or in combat.

This desire for the simplest model possible is not to negate the need for complex models; instead, it is an appeal to avoid unnecessary complexity. There are situations where a complex model is required.

Developing the theme of the requirements of a conceptual model further, there are some discussions on more specific criteria for a “good” model. These are summarized in Table 1. Robinson [2] suggests that such criteria are useful for guiding modelers, and the problem owners, when thinking about the scope and level of detail for a model.

While exhortations to develop the simplest model possible and requirements of conceptual models might be useful, neither provides specific guidance on how to determine the conceptual model for a specific simulation study. Modeling principles and framework, and methods of simplification, go at least some way to helping in this respect.

Principles of Modeling

Providing a set of guiding principles for modeling is one approach to advising simulation modelers on how to develop (conceptual) models. For instance, Pidd [34] describes six principles of modeling:

- model simple, think complicated;
- be parsimonious, start small and add;
- divide and conquer, avoid megamodels;
- use metaphors, analogies, and similarities;
- do not fall in love with data;
- modeling may feel like muddling through.

Table 1. Requirements of a Conceptual Model

Pritsker [28]	Henriksen [29]	Nance [30]	Willemain [31]	Brooks and Tobias [32]	van der Zee and van der Vorst [33]	Robinson [2]
Valid	Fidelity	Model	Validity	Model describes behavior	Completeness	Validity
Understandable	Execution speed	correctness	Aptness for	of interest	Transparency	Credibility
Extendible	Ease of	Testability	client's problem	Accuracy of the model's		Utility
Timely	modification	Adaptability	Value to client	results		Feasibility
	Elegance	Reusability	Usability	Probability of containing		
		Maintainability	Feasibility	errors		
				Validity		
				Strength of theoretical		
				basis of model		
				Ease of understanding		
				Portability and ease with		
				which model can be		
				combined with others		
				Time and cost to build		
				model		
				Time and cost to run		
				model		
				Time and cost to analyze		
				results		
				Hardware requirements		

The central theme is one of aiming for simple models through evolutionary development. Others have produced similar sets of principles (or guidelines); for instance, Morris [35], Musselman [36], Powell [37], Pritsker [38], and Law [1].

These principles provide some useful guidance for those developing conceptual models. It is useful to encourage modelers to start with small models and to gradually add scope and detail. What such principles do not do, however, is to guide a modeler through the conceptual modeling process. When should more detail be added? When should elaboration stop? There is a difference between giving some general principles and guiding someone through a process.

Conceptual Modeling Frameworks

A modeling framework goes beyond the idea of guiding principles by providing a specific set of steps that guide a modeler through the development of a conceptual model. There have been some attempts to provide such frameworks going back to Shannon [39] who describes four steps: specification of the model's purpose; specification of the model's components; specification of the parameters and variables associated with the components; and specification of the relationships between the components, parameters, and variables.

Both Nance and Pace have devised frameworks that relate primarily to the development of large-scale models in the military domain. Nance [30] outlines the conical methodology. This is an object-oriented, hierarchical specification language that develops the model definition (scope) top-down and the model specification (level of detail) bottom-up. A series of modeling steps are outlined. Balci and Nance [40] focus specifically on a procedure for problem formulation. Meanwhile, Arthur and Nance [41] identify the potential to adopt software requirements engineering (SRE) approaches for simulation model development. They also note that there is little evidence of SRE actually being adopted by simulation modelers, which is surprising given the close correspondence between the two fields. (The author contends that the key difference between SRE and

simulation conceptual modeling is the need for abstraction in simulation.)

Pace [17,18] explores a four-stage approach to conceptual model development, similar to that of Shannon: collect authoritative information on the problem domain; identify entities and processes that need to be represented; identify simulation elements; and identify relationships between the simulation elements. He also identifies six criteria for determining which elements to include in the conceptual model. These criteria focus on the correspondence between real-world items and simulation objects [18].

Outside of the military domain there is quite limited work on conceptual modeling frameworks. Brooks and Tobias [42] briefly propose a framework for conceptual modeling, but go no further in expanding upon the idea. Papers by Guru and Savory [43] and van der Zee and van der Vorst [33] propose conceptual modeling frameworks in some more detail. Guru and Savory propose a set of modeling templates (tables) useful for modeling physical security systems. Meanwhile, van der Zee and van der Vorst propose a framework for supply chain simulation. Both are aimed at an object-oriented implementation of the computer-based simulation model. Meanwhile, Kotiadis [44] looks to the ideas of soft operational research, and specifically soft systems methodology (SSM) [45], for aiding the conceptual modeling process. She uses SSM to help understand a complex health-care system and then derives the simulation conceptual model from the SSM "purposeful activity model."

Robinson [46] proposes a conceptual modeling framework that guides a modeler from identification of the modeling objectives through to determining the scope and level of detail of a model. Birta and Arbez [47] have devised the ABCmod conceptual modeling framework. This provides a detailed procedure for identifying the components and relationships for a discrete-event simulation model.

Such frameworks certainly appear to have potential for aiding the development of conceptual models, but they cannot be said to be fully developed or in common use. One interesting issue is whether frameworks are

best aimed at a specific domain (e.g., supply chain) or whether generic frameworks can be devised.

Methods of Model Simplification

Simplification entails removing scope and detail from a model or representing components more simply while maintaining a sufficient level of accuracy. In Zeigler's [5] terms this could be described as further lumping of the lumped model. In many respects this is the opposite of the start small and add principle of Pidd [27].

There are quite a number of discussions on simplification, both in the simulation and the wider modeling context. Morris [35] identifies some methods for simplifying models: making variables into constants, eliminating variables, using linear relations,

strengthening the assumptions and restrictions, and reducing randomness. Ward [12] provides a similar list of ideas for simplification. Meanwhile, Courtois [48] identifies criteria for the successful decomposition of models in engineering and science.

For simulation modeling, Zeigler [5] suggests four methods of simplification: dropping unimportant components of the model, using random variables to depict parts of the model, coarsening the range of variables in the model, and grouping components of the model. Yin and Zhou [49] build upon Zeigler's ideas, discussing six simplification techniques and presenting a case study.

Innis and Rexstad [11] enter into a detailed discussion about how an existing model might be simplified. They provide a list of 17 such methods, although they do not claim that these are exhaustive. Robinson

Table 2. Research Themes for Conceptual Modeling^a

The Problem/Modeling Objectives Domain	The Model Domain
Use of 'Soft OR' as a basis for determining a simulation conceptual model	Identifying dimensions for determining the performance of a conceptual model
How best to work with subject matter experts in forming a conceptual model	Comparing different models in the same problem domain
How to organize and structure the knowledge gained during conceptual modeling	Studying expert modelers to understand how they form conceptual models
Alternative sources of contextual data/information for conceptual modeling, including paper, interview, and electronic sources	How software engineering techniques might aid simulation conceptual modeling
Developing curricula to include conceptual modeling in university and industry courses on simulation	Adopting/developing appropriate model representation methods
	Exploring methods of model simplification
	Identifying, adapting, and developing conceptual modeling frameworks
	Refining models through agreement between the modeler and stakeholders—"convergent design"
	Exploring the creative aspects of modeling
	Understanding the organizational diffusion and acceptance of models
	Investigating the impact of other modeling tasks on the conceptual model (iteration in the simulation life cycle)
	Understanding the effect of throw away models versus models with longevity—for example, the time spent on conceptual modeling, documentation, and organizational diffusion

^aSource: Taken from Robinson (51, p. 151).

[50] also lists some methods for simplifying simulation models.

Such ideas are useful for simplifying an existing (conceptual) model, but they do not guide the modeler over how to bring a model into existence. Model simplification acts primarily as a redesign tool and not a design tool.

FUTURE RESEARCH IN CONCEPTUAL MODELING

Work on conceptual modeling has been sporadic over the past 40 or more years. At present there are a number of groups with a specific interest in the topic, for instance, the NATO Research and Technology Organisation Activity on Conceptual Modelling for Modelling and Simulation (www.rta.nato.int/Activities.aspx#), the stream of papers at the Simulation Interoperability Workshops (www.sisostds.org), and the conceptual modelling group formed around the UK Simulation Workshops (conceptualmodeling.info). Indeed, conceptual modeling is currently seen as an important topic in developing the field of simulation.

At a meeting in 2006, the latter group identified a set of key research themes for conceptual modeling. These are reported in Robinson [51] and summarized in Table 2. The themes were split between those belonging to the problem domain and the model domain (Fig. 1). Such a list may prove useful for those considering research in this area.

CONCLUSION

This article provides an overview of the field of simulation conceptual modeling. It explores the meaning and process of conceptual modeling, and provides some ideas on areas for future research. Conceptual modeling is certainly a field that is ripe for further research and a field that is much in need of further development. It is also a field that is difficult to work in because conceptual modeling, by nature, is more of an art than a science.

Acknowledgments

Sections of this article are reproduced from Robinson [2] and Kotiadis and Robinson [4].

REFERENCES

1. Law AM. Simulation modeling and analysis. 4th ed. New York: McGraw-Hill; 2007.
2. Robinson S. Conceptual modelling for simulation part I: definition and requirements. *J Oper Res Soc* 2008;59(3):278–290.
3. Fishwick PA. Simulation model design and execution: building digital worlds. Upper Saddle River (NJ): Prentice-Hall; 1995.
4. Kotiadis K, Robinson S. Conceptual modelling: knowledge acquisition and model abstraction. Proceedings of the 2008 Winter Simulation Conference; Miami. 2008. pp. 951–958.
5. Zeigler BP. Theory of modeling and simulation. New York: Wiley; 1976.
6. Balci O. Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Ann Oper Res* 1994;53:121–173.
7. Willemain TR. Model formulation: what experts think about and when. *Oper Res* 1995;43(6):916–932.
8. Landry M, Malouin JL, Oral M. Model validation in operations research. *Eur J Oper Res* 1983;14(3):207–220.
9. Robinson S. Simulation verification, validation and confidence: a tutorial. *Trans Soc Comput Simul Int* 1999;16(2):63–69.
10. Sargent RG. Verification and validation of simulation models. Proceedings of the 2008 Winter Simulation Conference; Miami. 2008. pp. 157–169.
11. Innis G, Rexstad E. Simulation model simplification techniques. *Simulation* 1983; 41(1):7–15.
12. Ward SC. Arguments for constructively simple models. *J Oper Res Soc* 1989;40(2):141–153.
13. Salt J. Simulation should be easy and fun. Proceedings of the 1993 Winter Simulation Conference; Los Angeles. 1993. pp. 1–5.
14. Chwif L, Barretto MRP, Paul RJ. On simulation model complexity. Proceedings of the 2000 Winter Simulation Conference; Orlando. 2000. pp. 449–455.
15. Lucas TW, McGunnigle JE. When is model complexity too much? Illustrating the benefits of simple models with Hughes' salvo equations. *Nav Res Log* 2003;50:197–217.

16. Thomas A, Charpentier P. Reducing simulation models for scheduling manufacturing facilities. *Eur J Oper Res* 2005; 161(1):111–125.
17. Pace DK. Development and documentation of a simulation conceptual model. Proceedings of the 1999 Fall Simulation Interoperability Workshop; 1999. Available at www.sisostds.org. Accessed 2009 Jan.
18. Pace DK. Simulation conceptual model development. Proceedings of the 2000 Spring Simulation Interoperability Workshop; 2000. Available at www.sisostds.org. Accessed 2009 Jan.
19. Lacy LW, Randolph W, Harris B, *et al.* Developing a consensus perspective on conceptual models for simulation systems. Proceedings of the 2001 Spring Simulation Interoperability Workshop; 2001. Available at www.sisostds.org. Accessed 2009 Jan.
20. Haddix F. Conceptual modeling revisited: a developmental model approach for modeling and simulation. Proceedings of the 2001 Fall Simulation Interoperability Workshop; 2001. Available at www.sisostds.org. Accessed 2009 Jan.
21. Borah J. Simulation conceptual modeling (SCM) study group final report (SISO-REF-017-2006). Simulation Interoperability Standards Organization (SISO); 2006. Available at www.sisostds.org. Accessed 2009 Jan.
22. Balci O. Accomplishing reuse with a simulation conceptual model. Proceedings of the 2008 Winter Simulation Conference; Miami. 2008. pp. 958–965.
23. Henriksen JO. Alternative modeling perspectives: finding the creative spark. Proceedings of the 1989 Winter Simulation Conference; Washington DC. 1989. pp. 648–652.
24. Schmeiser BW. Some myths and common errors in simulation experiments. Proceedings of the 2001 Winter Simulation Conference; Washington DC. 2001. pp. 39–46.
25. Evans JR. Creativity in MS/OR: improving problem solving through creative thinking. *Interfaces* 1992;22(2):87–91.
26. Ferguson P, Humphrey WS, Khajenoori S, *et al.* Results of applying the personal software process. *Computer* 1997;5:24–31.
27. Pidd M. Tools for thinking: modelling in management science. 2nd ed. Chichester: Wiley; 2003.
28. Pritsker AAB. Model evolution: a rotary table case history. Proceedings of the 1986 Winter Simulation Conference; Washington DC. 1986. pp. 703–707.
29. Henriksen JO. One system, several perspectives, many models. Proceedings of the 1988 Winter Simulation Conference; San Diego. 1988. pp. 352–356.
30. Nance RE. The conical methodology and the evolution of simulation model development. *Ann Oper Res* 1994;53:1–45.
31. Willemain TR. Insights on modeling from a dozen experts. *Oper Res* 1994;42(2): 213–222.
32. Brooks RJ, Tobias AM. Choosing the best model: level of detail, complexity and model performance. *Math Comput Model* 1996; 24(4):1–14.
33. van der Zee DJ, van der Vorst JGAJ. A modeling framework for supply chain simulation: opportunities for improved decision making. *Decis Sci* 2005;36(1):65–95.
34. Pidd M. Just modeling through: a rough guide to modeling. *Interfaces* 1999;29(2):118–132.
35. Morris WT. On the art of modeling. *Manage Sci* 1967;13(12):B707–B717.
36. Musselman KJ. Conducting a successful simulation project. Proceedings of the 1992 Winter Simulation Conference; Washington DC. 1992. pp. 115–121.
37. Powell SG. Six key modeling heuristics. *Interfaces* 1995;25(4):114–125.
38. Pritsker AAB. Principles of simulation modeling. In: Banks J, editor. *Handbook of simulation*. New York: Wiley; 1998. pp. 31–51.
39. Shannon RE. *Systems simulation: the art and science*. Englewood Cliffs (NJ): Prentice-Hall; 1975.
40. Balci O, Nance RE. Formulated problem verification as an explicit requirement of model credibility. *Simulation* 1985;45(2):76–86.
41. Arthur JD, Nance RE. Investigating the use of software requirements engineering techniques in simulation modelling. *J Simul* 2007;1(3):159–174.
42. Brooks RJ, Tobias AM. A framework for choosing the best model structure in mathematical and computer modelling. Proceedings of the 6th Annual Conference AI, Simulation, and Planning in High Autonomy Systems; San Diego. 1996. pp. 53–60.
43. Guru A, Savory P. A template-based conceptual modeling infrastructure for simulation of physical security systems. Proceedings of the 2004 Winter Simulation Conference; Washington DC. 2004. pp. 866–873.
44. Kotiadis K. Using soft systems methodology to determine the simulation study objectives. *J Simul* 2007;1(3):215–222.

45. Checkland PB. Systems thinking, systems practice. Chichester: Wiley; 1981.
46. Robinson S. Conceptual modelling for simulation part II: a framework for conceptual modelling. *J Oper Res Soc* 2008;59(3):291–304.
47. Birta LG, Arbez G. Modelling and simulation: exploring dynamic system behaviour. New York: Springer; 2007.
48. Courtois PJ. On time and space decomposition of complex structures. *Commun ACM* 1985;28(6):590–603.
49. Yin HY, Zhou ZN. Simplification techniques of simulation models. *Proceedings of Beijing International Conference on System Simulation and Scientific Computing*; Beijing. 1989. pp. 782–786.
50. Robinson S. Simulation projects: building the right conceptual model. *Ind Eng Chem* 1994;26(9):34–36.
51. Robinson S. Editorial: the future's bright the future's...conceptual modelling for simulation!. *J Simul* 2007;1(3):149–152.

CONDITION-BASED MAINTENANCE UNDER MARKOVIAN DETERIORATION

BORA ÇEKYAY

SÜLEYMAN ÖZEKICI

Department of Industrial Engineering,
Koç University, Istanbul, Turkey

INTRODUCTION

Maintenance actions are vital for companies to increase reliability and availability of the production system and to decrease production costs. At the same time, Bevilacqua and Braglia [1] state that maintenance may require extensive expenditure, which may vary from 15 to 70% of the total production cost depending on the industry. For instance, the total amount of money spent for maintenance is more than \$200 billion in the United States every year as observed by Chu *et al.* [2]. Moreover, a significant portion of total work force in a company is employed in maintenance departments; Waeyenbergh and Pintelon [3] estimate that this is up to 30% or more in chemical process industries. These observations indicate that optimizing the obvious trade-off between maintenance costs and productivity will have a very significant impact on the total cost. This is why it is not surprising that an extensive body of literature on optimal maintenance has accumulated in the last 50 years. The review papers [4–13] survey hundreds of papers in chronological order on optimal maintenance problems.

In general, there are two types of maintenance considered in the literature: corrective maintenance (CM) and preventive maintenance (PM). CM involves actions performed after a failure to restore the system to a better condition. Sim and Endrenyi [14] define PM as the actions performed regularly at preselected times (not necessarily identical) to reduce or eliminate the accumulated

deterioration. PM can be further classified into two main classes: time-based preventive maintenance (TBPM) and condition-based maintenance (CBM). System lifetime is considered as a random variable in TBPM and its distribution is determined by statistical analysis. Then, optimal PM actions are planned according to a mathematical model developed using the failure distribution of the system and related maintenance costs. On the other hand, Wang *et al.* [15] state that maintenance decisions are made according to the actual state or condition of the system under CBM policies. The state of the system may take either discrete values such as real or intrinsic age (e.g., number of flights for planes) or predefined deterioration levels, or continuous values such as temperature, vibration, cumulative wear, and so on. In the former one, multistate Markov decision processes are generally used to determine the states at which a preventive replacement decision is optimal to minimize a cost criteria. In the latter one, the system is generally subject to continuous wear or deterioration. The general purpose of the models developed to investigate such systems is to find an optimal threshold above which a preventive replacement decision is optimal. In this article, we focus on CBM models under Markovian deterioration, where the state of the system can be classified into one of a finite number of states. For more information on CBM models with continuous deterioration, we refer the interested reader to Refs 2 and 15–25 and the references cited in these papers.

CBM models are further classified by the time points at which the state of the system is observed by the decision maker. There are three types of inspection policies applied in CBM literature: continuous inspection policy (CIP), periodic inspection policy (PIP), and sequential inspection policy (SIP). Systems are always monitored or the state of the system is always known by the decision maker under CIP. On the other hand, the condition of the system under PIP or SIP is known only at some discrete time points. The main

difference between PIP and SIP is that in PIP, the system is inspected and its state is observed at equal time intervals, but these intervals do not have to be equal in SIP. If we apply PIP or SIP, the time points at which the system will be inspected must be determined carefully since more frequent inspections will increase the related inspection cost and less frequent inspections will decrease the ability to maintain the survival of the system. We will also review some important optimal inspection and replacement models for which a control-limit policy is optimal.

Although they may not be optimal even under very intuitive conditions (as we will illustrate by an example later), control-limit policies are studied extensively in the literature. The significance of control-limit policies is that they are very easy to understand and implement. Similarly, there are many CBM policies defined using several thresholds in the literature. These models are important because they are also easy to apply and they can reasonably describe the deterioration-maintenance process of some systems in real-world applications. We will also review some of the important papers that analyze maintenance policies defined via several thresholds.

In this article, we review papers where the state of the system at any time can be determined perfectly by observations. However, in recent years, there is growing interest in partially observed systems where the actual state of the system is unknown and can be estimated (imperfectly) based on observed conditions. Partially observed Markov decision processes are generally used to analyze such models. We refer the interested reader to Ghasemi *et al.* [26] and Maillart [27] for recent advances in this research area. Moreover, parameter estimation of the partially observed models is also important to apply these models to real-life situations by utilizing on-line information about the state of the system. Lin and Makis [28] propose a recursive maximum likelihood algorithm which is applicable to systems that deteriorate according to a Markov process.

In the second section, the general assumptions and notation used in CBM models with Markovian deterioration are given. The third and fourth sections are on

optimal replacement, and optimal inspection/replacement models, respectively where optimality of control-limit policies is established. We conclude the article with fifth section, by summarizing some other important maintenance models where policies are described by a few thresholds.

GENERAL ASSUMPTIONS

We shall concentrate specifically on optimal policies for CBM models with Markovian deterioration. In the most generic sense, these models satisfy the following main assumptions.

1. The system can be observed to be in one of the $M + 1$ states from the set $F = \{0, 1, \dots, M - 1, M\}$, where state 0 represents a brand-new system, states $1, 2, \dots, M - 1$ represent intermediate deterioration levels in ascending order, and M denotes system failure.
2. Transitions among the deterioration levels at successive decision times follow an increasing Markov chain with an upper-triangular transition probability matrix $P = [P_{ab}]$, where P_{ab} is the probability that the next deterioration level of the system will be b given that current level is a for every $a, b \in F$.
3. Holding time in each level is a random variable with parameters which may or may not depend on the deterioration level. Let t_a be the holding time in state $a \in F$ with mean $\bar{t}_a$. We suppose that t_a is exponentially distributed for every a so that the deterioration process is a Markov process. Otherwise, it is a semi-Markov process.
4. Replacement durations may be negligible or random. Let r_a denote the replacement duration with mean $\bar{r}_a$ if the replacement decision is given when the system is in state $a \in F$. There is always a replacement cost c_a if the system is replaced when its deterioration level is $a \in F$.
5. The system may be inspected continuously, periodically, or sequentially.

6. A state occupancy cost h_a may be incurred when the system is occupying level $a \in F$. When the system fails, a failure cost K is incurred.
7. At each decision epoch, there are two alternative decisions: the system will be replaced ($s_a = 1$) or not replaced ($s_a = 0$), respectively if the system is in state $a \in F$ at that decision epoch. The replacement action is assumed to be perfect so that the system state is restored to 0 after the replacement.
8. The objective of the problem is to minimize the expected total discounted cost or the average cost per time.

These types of maintenance problems have been studied in the literature since the 1960s. In general, there are two types of research papers where the first type defines a mathematical model of the maintenance problem and obtains a policy (usually a control-limit policy), which solves the problem optimally. The second type investigates the problems using a given simplified policy that is not necessarily optimal and finds the optimum parameters of this class of policies that minimize a cost function. In the following two sections, we will review papers of the first type for replacement and inspection/replacement models respectively. The final section focuses on papers of the second type.

OPTIMAL REPLACEMENT MODELS

One of the earliest and basic cases where the deterioration process is described by a Markov chain is analyzed by Derman [29]. It is assumed that the system is inspected at equally spaced points in time and the system is classified into one of the deterioration levels after each inspection. The holding times are not considered and the decision model is formulated based on the deterioration levels of the system observed at each inspection time. It is assumed that the successive levels follow a Markov chain whose transition matrix is monotone so that the cumulative matrix

$$\bar{P}_{ab} = \sum_{k=b}^M P_{ak} \quad (1)$$

is nondecreasing in a for every $b \in F$. The other assumptions of this model are that the replacement duration is equal to one inspection interval, replacement costs do not depend on the deterioration level ($c_a = c$ for every a and some c), and there is no state occupancy cost ($h_a = 0$). The dynamic programming equation is

$$v(a) = \begin{cases} c + K + \alpha \sum_{b=0}^M P_{0b} v(b), & \text{if } a = M, \\ \min \left\{ \alpha \sum_{b=0}^M P_{ab} v(b), c + \alpha \sum_{b=0}^M P_{0b} v(b) \right\}, & \text{if } a \neq M, \end{cases} \quad (2)$$

where $0 \leq \alpha < 1$ is the periodic discount factor and $v(a)$ is the total expected discounted cost. In our discussions, we will present details primarily on $v(a)$ with the understanding that the average cost can be obtained in a similar fashion. It is proven that the optimal policy has a control-limit structure. In particular, there exists $a^* \in F$ such that

$$s_a^* = \begin{cases} 1, & \text{if } a \geq a^*, \\ 0, & \text{if } a < a^*, \end{cases} \quad (3)$$

for every $a \in F$ where s^* denotes the optimal policy. The same model with state occupancy costs incurred each time that the system is inspected is analyzed by Kolesar [30]. The dynamic programming equation now becomes

$$v(a) = \begin{cases} c + h_M + \alpha \sum_{b=0}^M P_{0b} v(b), & \text{if } a = M, \\ \min \left\{ h_a + \alpha \sum_{b=0}^M P_{ab} v(b), c + h_a + \alpha \sum_{b=0}^M P_{0b} v(b) \right\}, & \text{if } a \neq M, \end{cases} \quad (4)$$

and the optimal policies minimizing both total expected discounted cost and average cost are of the control-limit type if h_a and $\bar{P}_{ab}$ are nondecreasing in a .

Kawai *et al.* [31] consider another model where state occupancy costs are not paid during replacement, and all costs are state dependent. The dynamic programming equation is

$$v(a) = \min \left\{ h_a + \alpha \sum_{b=0}^M P_{ab} v(b), c_a + \alpha v(0) \right\}, \quad (5)$$

and it is shown that the optimal policy still has a control-limit structure even when c_a is increasing in a provided that $h_a, h_a - c_a$, and $\bar{P}_{ab}$ are increasing in a .

A generalization of the model in Kolesar [30] is analyzed by Wood [32] by considering the case where the replacement action may fail with a probability $1 - p$ and the occupancy costs are not paid during replacement. The standard recursion for this model can be formulated as

$$v(a) = \begin{cases} c + K + \alpha p v_\alpha(0) + \alpha(1 - p) v_\alpha(M), & \text{if } a = M, \\ \min \left\{ h_a + \alpha \sum_{b=0}^M P_{ab} v_\alpha(b), c + \alpha p v_\alpha(0) + \alpha(1 - p) v_\alpha(a) \right\}, & \text{if } a \neq M. \end{cases} \quad (6)$$

For this model, the optimality of a control-limit policy minimizing the total expected discounted cost and the average cost is proven under the same assumptions used by Kolesar [30]. The model where the occupancy costs are paid during replacement has the dynamic programming equation

$$v(a) = \begin{cases} c + K + h_M + \alpha p v(0) + \alpha(1 - p) v(M), & \text{if } a = M, \\ \min \left\{ h_a + \alpha \sum_{b=0}^M P_{ab} v(b), c + h_a + \alpha p v(0) + \alpha(1 - p) v(a) \right\}, & \text{if } a \neq M. \end{cases} \quad (7)$$

It is shown that the control-limit rule may not be optimal for this case by a counterexample. In the same paper, a constantly monitored system is also investigated with

the assumptions that the replacement duration and holding times are exponentially distributed, and replacement decisions are allowed only when a transition occurs in the deterioration process of the system. It is assumed that the holding times depend on the deterioration level with rate λ_a for level a , but the replacement duration does not with a constant rate λ . The analysis is done by applying uniformization techniques by which a continuous-time Markov decision process can be converted into an equivalent discrete-time Markov decision process. Wood [32] concludes that the optimal policy is control-limit type for both total expected discounted cost and average cost criteria provided that h_a and $\sum_{k=b}^M P_{ak} \lambda_k$ are nondecreasing in a , and occupancy costs are not paid during replacement. It is also proven that the same result holds for the case where the occupancy costs are also paid during replacement if the replacement duration is stochastically smaller than the holding time in each state ($\lambda \geq \lambda_a$). Özekici and Günlük [33] propose some sufficient conditions, which make the lifetime of a system with Markovian deterioration increasing failure rate on average (IFRA), and also show that these conditions imply the optimality of a control-limit policy if the replacement cost does not depend on the deterioration level of the system.

An interesting extension of these models is obtained when failure parameters are modulated by a Markovian mission process. Many missions performed by a device are composed of different stages or phases. Since each phase has different properties, the system may deteriorate in a phase more than it does during another phase. For instance, if you consider a flight of a plane as a mission, you can divide this mission into three basic phases: take-off, cruise, and landing. It is well-known that the degradation of a plane during take-off and landing is much more than it is during cruise. These type of examples stimulate the models where the failure parameters of the system are dependent on a mission process. They are often called *phased-mission systems* or *mission-based systems* in the literature. The optimal maintenance problem of such a system inspected constantly is analyzed by Çekyay

and Özekici [34]. The mission process is a Markov process on some finite state space E with a transition rate vector μ and transition probability matrix Q . Suppose that the deterioration process of the system follows a Markov process with transition rate vector λ_i and upper-triangular transition probability matrix P_i during phase $i \in E$. Let c_i and K_i be the replacement cost and failure cost, respectively during phase i . Note that the replacement cost does not depend on the deterioration level. It is further assumed that if the system fails during a phase, it will perform the same phase after the replacement since it is not yet completed. A state occupancy cost $h_i(a)$ is incurred during phase i if the deterioration level of the system is a . Replacements are instantaneous, a replacement decision can be given only when a change occurs in the mission process or in the deterioration level, and the state occupancy costs are not paid during replacement. Let $v(i, a)$ be the optimal expected total discounted cost if the initial phase is i and the initial deterioration level of the system is a for every $i \in E$ and $a \in F$. The dynamic programming equation is

$$v(i, a) = \min_{s \in A_a} \{r_s(i, a) + \Gamma_s v(i, a)\}, \quad (8)$$

where $A_0 = \{0\}$, $A_M = \{1\}$, $r_1(i, M) = K_i + h_i(0)$, $A_a = \{0, 1\}$ for $a = 1, 2, \dots, M-1$, $r_0(i, a) = h_i(a)$, $r_1(i, a) = h_i(0) + c_i$,

$$\begin{aligned} \Gamma_0 v(i, a) &= \frac{\mu(i) + \lambda_i(a)}{\mu(i) + \lambda_i(a) + \alpha} \\ &\times \left[\frac{\mu(i)}{\mu(i) + \lambda_i(a)} \sum_{j \in E} Q(i, j) v(j, a) \right. \\ &\left. + \frac{\lambda_i(a)}{\mu(i) + \lambda_i(a)} \sum_{b=0}^M P_i(a, b) v(i, b) \right], \end{aligned} \quad (9)$$

for $a = 0, 1, \dots, M-1$, and

$$\Gamma_1 v(i, a) = \frac{\mu(i) + \lambda_i(0)}{\mu(i) + \lambda_i(0) + \alpha}$$

$$\begin{aligned} &\times \left[\frac{\mu(i)}{\mu(i) + \lambda_i(0)} \sum_{j \in E} Q(i, j) v(j, 0) \right. \\ &\left. + \frac{\lambda_i(0)}{\mu(i) + \lambda_i(0)} \sum_{b=0}^M P_i(0, b) v(i, b) \right], \end{aligned} \quad (10)$$

for $a = 1, \dots, M$. The following characterization is obtained using uniformization techniques.

Theorem 1. *Suppose that s^* is the optimal replacement policy. Then, there exists a_i^* for every $i \in E$ such that*

$$s^*(i, a) = \begin{cases} 1, & \text{if } a \geq a_i^*, \\ 0, & \text{if } a < a_i^*, \end{cases}$$

provided that

- (a) P_i is a monotone matrix for every i ,
- (b) $\lambda_i(a)$ is nondecreasing in a for every i ,
- (c) $c_i \leq K_i$ for every i ,
- (d) $h_i(a)$ is nondecreasing in a for every i ,
- (e) there exists a constant c such that $\sup_{i \in E, a \in F \setminus \{M\}} \{\mu(i) + \lambda_i(a)\} \leq c$.

Theorem 1 indicates that the optimal policy has a control-limit structure where the control limits depend on the phases of the mission. The preventive replacement cost c_i is independent of the deterioration level of the system. In real-life applications, this cost may be increasing in the deterioration level since the salvage value of the system decreases as the system deteriorates. Let $c_i(a)$ be the preventive replacement cost of the system with deterioration level a during phase i . Note that the optimal policy can now be obtained by solving Equations (8)–(10) where $r_1(i, a) = h_i(0) + c_i(a)$ for all $a = 0, 1, \dots, M-1$. Suppose further that $c_i(a)$ is increasing in a for every phase as expected. It is clear that this modification is very reasonable and our intuition may indicate that $v(i, a)$ will be increasing in a and the optimal policy will be of the control-limit type. However, the following counterexample shows that the optimal policy may not have a control-limit structure even when $v(i, a)$ is increasing in a .

Example 1. Suppose that $M = 6$ and the system performs a mission with three phases. The transition probability matrix and the transition rates of the mission process are

$$Q = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0.5 & 0 \end{bmatrix} \quad (11)$$

and

$$\mu = [5 \quad 8 \quad 4]. \quad (12)$$

The transition probability matrices of the deterioration process are

$$P_1 = \begin{bmatrix} 0 & 0.1 & 0.2 & 0.2 & 0.3 & 0.1 & 0.1 \\ 0 & 0 & 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \\ 0 & 0 & 0 & 0.2 & 0.23 & 0.22 & 0.35 \\ 0 & 0 & 0 & 0 & 0.3 & 0.3 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} 0 & 0.1 & 0.2 & 0.2 & 0.2 & 0.1 & 0.2 \\ 0 & 0 & 0.05 & 0.13 & 0.22 & 0.25 & 0.35 \\ 0 & 0 & 0 & 0.17 & 0.22 & 0.23 & 0.38 \\ 0 & 0 & 0 & 0 & 0.24 & 0.32 & 0.44 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (13)$$

$$P_3 = \begin{bmatrix} 0 & 0.1 & 0.1 & 0.1 & 0.15 & 0.2 & 0.35 \\ 0 & 0 & 0.05 & 0.08 & 0.22 & 0.25 & 0.4 \\ 0 & 0 & 0 & 0.05 & 0.24 & 0.26 & 0.45 \\ 0 & 0 & 0 & 0 & 0.18 & 0.32 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.45 & 0.55 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (14)$$

$$v^* = \begin{bmatrix} 312.123 & 337.123 & 352.123 & 372.123 & 526.516 & 562.123 & 587.123 \\ 321.461 & 351.461 & 366.461 & 401.461 & 456.461 & 501.461 & 571.461 \\ 314.990 & 357.354 & 400.434 & 444.990 & 449.990 & 454.990 & 724.990 \end{bmatrix},$$

where the rows and the columns represent the phases and the deterioration levels respectively. It is clear that for phase 1, even when the cost function is increasing, the optimal policy is not the control-limit type.

In all of the papers discussed so far, the holding times are either negligible or

and the related transition rates are

$$\begin{aligned} \lambda_1 &= [1 \quad 2 \quad 3 \quad 3.1 \quad 10 \quad 50] \\ \lambda_2 &= [2 \quad 3 \quad 5 \quad 80 \quad 100 \quad 120] \\ \lambda_3 &= [1 \quad 1.5 \quad 2.5 \quad 40 \quad 45 \quad 50]. \end{aligned}$$

We show that if c depends on the deterioration level of the system, then the optimal policy does not have to be control-limit. Suppose that

$$c = \begin{bmatrix} 15 & 25 & 40 & 60 & 250 & 250 \\ 15 & 30 & 45 & 80 & 135 & 180 \\ 15 & 50 & 95 & 130 & 135 & 140 \end{bmatrix}$$

and

$$h = \begin{bmatrix} 10 & 12 & 14 & 16 & 18 & 20 \\ 5 & 8 & 10 & 12 & 14 & 16 \\ 2 & 3 & 4 & 5 & 6 & 6 \end{bmatrix},$$

where the rows and the columns represent the phases and the deterioration levels respectively, and

$$K = [275 \quad 250 \quad 410],$$

where the columns represent the phases and $\alpha = 0.75$. Then, the optimal replacement policy and optimal costs are

$$s^* = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and

exponentially distributed. However, the optimality of a control-limit policy may be obtained for the models with different (even more general) holding time distributions satisfying some monotonicity properties. A model where replacement and holding times are discrete random variables and the system is monitored continuously is analyzed

by Kao [35]. It is assumed that a replacement decision can only be given after a transition of the deterioration level of the system. The author proves that the optimal policy which minimizes the total discounted expected cost has control-limit structure provided that h_a is nondecreasing in a , $\bar{t}_a$ is nonincreasing in a , P is monotone, and the cost and time for replacement are independent of the deterioration level. In this model, the deterioration process is actually a semi-Markov process where sojourn times are discrete random variables. Following this line of research, So [36] used semi-Markov decision processes to analyze the model where the replacement duration has a general distribution, which is independent of the deterioration level ($\bar{r}_a = \bar{r}$); the holding times are independent and identically distributed random variables, where $\bar{t}_a$ is nonincreasing in a ; h_a and c_a are nondecreasing in a ; and $\sum_{b=a}^M P_{ab}c_b - c_a$ is nondecreasing in a for $1 \leq a < M$. The last assumption might be strict, but it is shown that this condition can be easily verified in some important special cases. It is also assumed that a fixed charge β is incurred, when the system is occupying level 0, to do parametric analysis on β . The optimality of a control-limit policy minimizing average cost is proven under some monotonicity assumptions for every β , in particular, for $\beta = 0$. The author also extends this result to the case where replacement durations are dependent on the deterioration level under the assumptions that $\sum_{b=a}^M P_{ab}c_b - c_a$ is nondecreasing in a for $1 \leq a < M$, $\bar{r}_a$ is nondecreasing in a , $h_a\bar{t}_a$ is nondecreasing in a , and $\sum_{b=a}^M P_{ab}\bar{r}_b - \bar{r}_a + \bar{t}_a$ is nonincreasing in a .

Another study using a semi-Markov process with continuous sojourn times to model the deterioration process of a system is presented by Lam and Yeh [37]. In this model, the holding time in level a has a general distribution F_a with hazard rate function f_a ; state occupancy costs, replacement costs and times are state dependent, and from level a , the deterioration process will make a direct transition either to level $a + 1$ with probability p_a or level M with probability $1 - p_a$. It is assumed that the system is monitored continuously and a decision is given when the system enters a new deterioration level.

When the system enters level a , the decision maker takes a decision to replace the system t_a units of time later if it remains in level a . If $t_a = 0$, then the system is replaced as soon as it enters level a and if $t_a = +\infty$, then the system will not be replaced as long as it stays in level a . The model has the following monotonicity assumptions.

1. The state occupancy cost rate, the replacement cost rate, the expected replacement time, the marginal replacement cost, and the marginal replacement time increase as the system deteriorates,
2. F_a is an increasing failure rate distribution for every a and $f_a(t)$ increases in a for every t , and
3. p_a is nondecreasing in a .

Under these conditions, there exist h^* and k^* with $0 \leq h^* \leq k^* \leq M$, such that

$$t_a^* = \begin{cases} +\infty, & \text{if } a < h^*, \\ s_a, & \text{if } h^* \leq a < k^*, \\ 0, & \text{if } k^* \leq a \leq M, \end{cases}$$

where t_a^* is the optimal decision in level a , and $0 < s_{h^*} \leq s_{h^*+1} \leq \dots \leq s_{k^*-1} < +\infty$. In other words, the system is replaced immediately as soon as it enters one of the states $\{k^*, k^* + 1, \dots, M\}$, and it is never replaced in states $\{0, 1, \dots, h^* - 1\}$. However, in any state $a \in \{h^*, h^* + 1, \dots, k^*\}$, it is replaced after s_a units of time in that state.

In recent years, advances in sophisticated sensor technology has made it possible to continuously observe the real-time condition of a system. This creates an opportunity to design more accurate and timely maintenance policies. The first challenge in using this technology is to find a measure that represents the state of the system based on several data obtained by several sensors built in the system. Chen *et al.* [38] proposed a health index to represent the condition of a system that deteriorates according to a nonstationary Markov chain. They analyze optimal maintenance times that minimize the expected cost per unit time and expected downtime per unit time.

OPTIMAL INSPECTION/REPLACEMENT MODELS

Another interesting research problem involves optimal inspection and replacement where the state of the system can only be observed via inspections performed at selected times. A fixed cost is incurred whenever the system is inspected and then, either a replacement occurs or the time until the next inspection is determined. Such a problem with negligible replacement and inspection times is analyzed by Ohnishi *et al.* [39] who consider a system with Markovian deterioration. It is assumed that holding times are state-dependent exponential random variables; state occupancy and replacement costs are dependent on the deterioration level and a direct transition can occur only to state $a + 1$ or state M . Under some monotonicity assumptions on costs and transition rates, it is shown that the optimal policy minimizing the average cost has a control-limit structure and the optimal time interval between successive inspections becomes shorter as the deterioration level of the system increases. Similar results are obtained by Lam and Yeh [40] for an identical model where replacement and inspection times are not negligible. It is clear that in real-world applications, numerical procedures are necessary to find optimal policies even if it has a control-limit structure. Iterative algorithms are derived for the optimal inspection and replacement problem under different maintenance strategies by Lam and Yeh [40]. These maintenance strategies include failure replacement, age replacement, sequential inspection, periodic inspection, and continuous inspection. These algorithms are valid for a model where the deterioration process is a continuous-time Markov process and from state a , a direct transition can occur only to state $a + 1$ or state M . Numerical procedures for a more general model are proposed by Yeh [41]. In this model, the deterioration process is a semi-Markov process where the holding time in each level follows a general distribution that depends on the level and from level a , direct transitions to levels $a + 1, a + 2, \dots, M - 1, M$ are allowed. Iterative algorithms minimizing the average

cost rate are provided to derive the optimal state-dependent and state-age-dependent inspection/replacement policies. In a state-age-dependent policy, once the state of the system is identified, the maintenance decision is made according to the deterioration level of the system and the time spent in the current state. However, if we apply the state-dependent policy, each maintenance action is determined only according to the deterioration level of the system no matter how long the system has been in that state.

Maillart and Zheltova [42] analyze a system, whose state can only be identified by inspection, using partially observed Markov decision processes. At each decision epoch, the decision maker has three options: replacement, inspection, and do nothing. If the system experiences inspection, the actual state of the system (information level) can be determined perfectly. If the system is not inspected, then the decision will be based on the information available in the previous decision epoch. The authors show the existence of a control-limit rule for this model under some monotonicity assumptions. The optimal decision is to replace the system if the available information level at a decision epoch is over a threshold that depends on the information level.

OPTIMAL MAINTENANCE USING THRESHOLDS

Besides the papers, which investigate the optimality of control-limit policies, there is also abundant literature that focus on a given special class of policies where maintenance decisions are made within that class. This type of model can be useful, especially, when a special policy reasonably describes the deterioration-maintenance process of the system. For instance, an optimal PM model suitable for (not limited to) especially coal pulverizers, circuit breakers, and transformers is proposed by Sim and Endrenyi [14]. In this model, the system is subject to two types of failure: Poisson failures and deterioration failures. The deterioration process is an increasing Markov process where the holding times are exponentially

distributed with a constant rate and $P_{ij} = 1$ for $j = i + 1$. The times to Poisson failures are exponentially distributed with a constant rate independent of the deterioration level of the system. The system is removed from operation periodically for preventive minimal maintenance, which restores the deterioration level to the previous level (i.e., from level i to level $i - 1$ if the deterioration level is i when the PM starts). The duration between two successive minimal PM actions has an Erlang- r distribution with mean $1/\lambda_m$. If a failure occurs, the system is restored to level 0 and the time to repair depends on the type of the failure. The steady-state equations for this model and a simple algorithm to find the steady-state probabilities when $r = 1$ are proposed. The authors also analyze the optimal PM problem to minimize unavailability with respect to λ_m . An extension of this model where $r = 1$ is investigated by Sim and Endrenyi [43]. This paper considers the systems that can be restored to “as good as new” status preventively. It is assumed that the first $s - 1$ PM actions are minimal (i.e., the system is restored to the previous deterioration level), then the following PM is major maintenance where the system is restored to level 0. This system is also subject to Poisson failures; but after a Poisson failure, the system experiences a minimal repair, which is exponentially distributed (i.e., the system is restored to the operable state it was in just before the failure). After a deterioration failure, the system is again overhauled to state 0. A recursive algorithm is proposed to find steady-state probabilities, and closed-form expressions for steady-state probabilities in the case where $s \rightarrow +\infty$ are given. The optimal values of λ_m , which minimize unavailability and average cost, respectively are also discussed.

This line of research is also followed by Chen and Trivedi [44] who consider a model where the holding times are dependent on the deterioration level and each inspection takes an exponentially distributed amount of time. It is assumed that the system is inspected after a random period which is exponentially distributed with rate λ_{in} . The applied PM policy can be summarized by the two thresholds

(g, b) as follows: if the observed deterioration level is $i \leq g$ after an inspection, then no maintenance occurs. If the system deterioration level is i with $g < i \leq b$, then the system is restored to level $i - 1$ by minimal maintenance. The system experiences major maintenance when the deterioration level is found to be in i with $b < i \leq M - 1$, by which the system is restored to level 0. If a deterioration failure occurs, the system is overhauled to level 0. Moreover, the system is subject to Poisson failures after which a minimal repair is performed, which restores the system to the level it was in just before the failure. For this model, the authors give closed-form expressions for steady-state probabilities, steady-state availability, and mean time to failure (MTTF). They also numerically analyze the optimal inspection intervals (λ_{in}) minimizing unavailability and average cost respectively, and maximizing MTTF under a target availability constraint. The optimality of such a threshold policy is shown by Chen and Trivedi [45] who analyze numerical examples for the case where the deterioration rate at each level is the same. A similar model where $g = 0$ is analyzed by Amari and McLaughlin [46]. Closed-form expressions for steady-state probabilities and availability are presented and algorithms to solve three optimization problems maximizing the system availability are given. These problems are formulated to find optimal λ_{in} for a given value of b , to find optimal b for a given value of λ_{in} , and to find optimal values of b and λ_{in} .

Minimal repair action may also be applied at any deterioration level of the system. Moustafa *et al.* [47] analyze a model where each holding time follows a general distribution and, at each state transition, one of three possible actions can be chosen: do nothing, minimal maintenance, and replacement. To derive the optimal policy minimizing the expected long-run cost rate, two different approaches are followed. In the first approach, a control-limit policy with two thresholds is determined. The second approach uses the conventional policy iteration algorithm. By numerical examples, it is shown that the optimal policy may not be control-limit and minimal maintenance may not be optimal in any state when the cost and

the time of minimal maintenance increase relative to the cost and the time of replacement respectively. A similar problem for software maintenance is analyzed by Krishnan *et al.* [48] assuming that the state of a software follows a Markov chain with a monotone transition matrix. The decision maker has three options at each decision epoch: do nothing, minor update, and major upgrade (rewrite of the software). Although they show that there exists a state beyond which major upgrade is always optimal, the optimal policy does not necessarily have a control-limit structure. Several situations specific to the software environment are investigated, such as long rewrite periods (more than one decision epoch), major upgrades due to technological advances, and partial rewrite of the software. They also analyze the effect of these situations on major upgrade decisions under some monotonicity assumptions on the cost functions.

Minimal maintenance restores the system to the previous deterioration level in all previously cited papers. An extension of this idea can be repair, by which the system can be restored to any better deterioration level. For example, a rather general policy $R_{ij}(T, N, \alpha)$ for a continuous-time Markovian deteriorating system is proposed and analyzed by Chiang and Yuan [49]. Under this policy, the system is inspected at times $T, 2T, 3T, \dots$ to identify the current deterioration level a of the system. Let m be the number of repairs already undertaken until the inspection time. The maintenance decision will be do nothing if $a \leq i - 1$, or $i \leq a \leq j - 1$ and $m = N$. The system is repaired to a better state if $i \leq a \leq j - 1$ and $m < N$. The next deterioration level of the system after the repair will be determined by the probability matrix α (i.e., the system will be restored to state r with probability α_{ar}). The maintenance decision will be replacement if $j \leq a \leq M$. An algorithm is also proposed to derive the optimal values of i, j , and T for given N and α .

REFERENCES

1. Bevilacqua M, Braglia M. The analytic hierarchy process applied to maintenance strategy selection. *Reliab Eng Syst Saf* 2000;70:71–83.
2. Chu C, Proth J-M, Wolff P. Predictive maintenance: the one-unit replacement model. *Int J Prod Econ* 1998;54:285–295.
3. Waeyenbergh G, Pintelon L. A framework for maintenance concept development. *Int J Prod Econ* 2002;77:299–313.
4. McCall JJ. Maintenance policies for stochastically failing equipment: a survey. *Manage Sci* 1965;11:493–524.
5. Barlow RE, Proschan F. Statistical theory of reliability and life testing: probability models. New York: Holt, Rinehart and Winston; 1975.
6. Sherif YS, Smith ML. Optimal maintenance models for systems subject to failure: a review. *Nav Res Logist Q* 1981;28:47–74.
7. Jardine AKS, Buzacott JA. Equipment reliability and maintenance. *Eur J Oper Res* 1985;19:285–296.
8. Valdez-Flores C, Feldman RM. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Nav Res Logist* 1989;36:419–446.
9. Cho DI, Parlar M. A survey of maintenance models for multi-unit systems. *Eur J Oper Res* 1991;51:1–23.
10. Dekker R. Applications of maintenance optimization models: a review and analysis. *Reliab Eng Syst Saf* 1996;51:229–240.
11. Dekker R, Wildeman RE, Van der Duyn Schouten FA. A review of multi-component maintenance models with economic dependence. *Math Methods Oper Res* 1997;45:411–435.
12. Wang H. A survey of maintenance policies of deteriorating systems. *Eur J Oper Res* 2002;139:469–489.
13. Nicolai RP, Dekker R. Optimal maintenance of multi-component systems: a review. In: Murthy DNP, Kobbachy KAH, editors. *Complex system maintenance handbook*. Springer series in reliability engineering. Berlin: Springer; 2008. pp. 263–286.
14. Sim SH, Endrenyi J. Optimal preventive maintenance with repair. *IEEE Trans Reliab* 1988;37:92–96.
15. Wang L, Chu J, Mao W. A condition-based replacement and spare provisioning policy for deteriorating systems with uncertain deterioration to failure. *Eur J Oper Res* 2009;194:184–205.
16. Park KS. Optimal continuous-wear limit replacement under periodic inspections. *IEEE Trans Reliab* 1988;37:97–102.
17. Park KS. Optimal wear-limit replacement with wear-dependent failures. *IEEE Trans Reliab* 1988;37:293–294.

18. Pellegrin C. Choice of a periodic on-condition maintenance policy. *Int J Prod Res* 1992; 30:1153–1173.
19. Barbera F, Schneider H, Kelle P. A condition based maintenance model with exponential failures and fixed inspection intervals. *J Oper Res Soc* 1996;47:1037–1045.
20. Barbera F, Schneider H, Watson E. A condition based maintenance model for a two-unit series system. *Eur J Oper Res* 1999;116: 281–290.
21. Grall A, Bérenguer C, Dieulle L. A condition-based maintenance policy for stochastically deteriorating systems. *Reliab Eng Syst Saf* 2002;76:167–180.
22. Grall A, Bérenguer C, Dieulle L, *et al.* Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE Trans Reliab* 2002;51:141–150.
23. Dieulle L, Bérenguer C, Grall A, *et al.* Sequential condition-based maintenance scheduling for a deteriorating system. *Eur J Oper Res* 2003;150:451–461.
24. Wang W. Modelling condition monitoring intervals: a hybrid of simulation and analytical approaches. *J Oper Res Soc* 2003;54: 273–282.
25. Castanier B, Grall A, Bérenguer C. A condition-based maintenance policy with non-periodic inspections for a two-unit series system. *Reliab Eng Syst Saf* 2005; 87:109–120.
26. Ghasemi A, Yacout S, Ouali MS. Optimal condition based maintenance with imperfect information and the proportional hazards model. *Int J Prod Res* 2007;45:989–1012.
27. Maillart LM. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Trans* 2006;38:463–475.
28. Lin D, Makis V. On-line parameter estimation for a partially observable system subject to random failure. *Nav Res Logist* 2006;53: 477–483.
29. Derman C. Finite state Markovian decision processes. London: Academic Press; 1970.
30. Kolesar P. Minimum cost replacement under Markovian deterioration. *Manage Sci* 1966;12: 684–706.
31. Kawai H, Koyanagi J, Ohnishi M. Optimal maintenance problems for Markovian deteriorating systems. In: Osaki S, editor. *Stochastic models in reliability and maintenance*. Berlin: Springer; 2002. pp. 193–218.
32. Wood AP. Optimal maintenance policies for constantly monitored systems. *Nav Res Logist* 1988;35:461–471.
33. Özekici S, Günlük NO. Maintenance of a device with age-dependent exponential failures. *Nav Res Logist* 1992;39:699–714.
34. Çekyay B, Özekici S. Optimal maintenance of systems with Markovian mission and deterioration. Technical report. Istanbul, Turkey: Koç University, Department of Industrial Engineering; 2009.
35. Kao EPC. Optimal replacement rules when changes of state are semi-Markovian. *Oper Res* 1973;21:1231–1249.
36. So KC. Optimality of control limit policies in replacement models. *Nav Res Logist* 1992;39:685–697.
37. Lam CT, Yeh RH. Optimal replacement policies for multistate deteriorating systems. *Nav Res Logist* 1994;41:303–315.
38. Chen A, Wu GS. Real-time health prognosis and dynamic preventive maintenance policy for equipment under aging Markovian deterioration. *Int J Prod Res* 2007;45:3351–3379.
39. Ohnishi M, Kawai H, Mine H. An optimal inspection and replacement policy for a deteriorating system. *J Appl Probab* 1986;23: 973–988.
40. Lam CT, Yeh RH. Optimal maintenance policies for deteriorating systems under various maintenance strategies. *IEEE Trans Reliab* 1994;43:423–430.
41. Yeh RH. Optimal inspection and replacement policies for multi-state deteriorating systems. *Eur J Oper Res* 1996;96:248–259.
42. Maillart LM, Zheltova L. Structured maintenance policies on interior sample paths. *Nav Res Logist* 2007;54:645–655.
43. Sim SH, Endrenyi J. A failure-repair model with minimal and major maintenance. *IEEE Trans Reliab* 1993;42:134–140.
44. Chen D, Trivedi KS. Closed-form analytical results for condition-based maintenance. *Reliab Eng Syst Saf* 2002;76:43–51.
45. Chen D, Trivedi KS. Optimization for condition-based maintenance with semi-Markov decision process. *Reliab Eng Syst Saf* 2005;90:25–29.
46. Amari SV, McLaughlin L. Optimal design of a condition-based maintenance model. *Reliability and Maintainability, 2004 Annual Reliability and Maintainability Symposium*. 2004. pp. 528–533.

47. Moustafa MS, Abdel Maksoudb EY, Sadekb S. Optimal major and minimal maintenance policies for deteriorating systems. *Reliab Eng Syst Saf* 2004;83:363–368.
48. Krishnan MS, Mukhopadhyay T, Kriebel CH. A decision model for software maintenance. *Inf Syst Res* 2004;15:396–412.
49. Chiang JH, Yuan J. Optimal maintenance policy for a Markovian system under periodic inspection. *Reliab Eng Syst Saf* 2001;71:165–172.

CONIC OPTIMIZATION SOFTWARE

IMRE PÓLIK
Optimization Solver Developer
SAS Institute, Inc., Cary,
North Carolina

PROBLEM DESCRIPTION

Conic optimization solvers target problems of the form

$$\begin{aligned} \min c^T x & \quad \max b^T y \\ Ax = b & \quad A^T y + s = c \\ x \in \mathcal{K} & \quad s \in \mathcal{K}^*, \end{aligned} \quad (1)$$

where $b, y \in \mathbb{R}^m$, $c, x, s \in \mathbb{R}^N$, $A \in \mathbb{R}^{m \times N}$, $\mathcal{K}, \mathcal{K}^* \subset \mathbb{R}^N$, and $\mathcal{K}^*$ is the dual of $\mathcal{K}$. In general, solvers exist if $\mathcal{K}$ is one of, or the product of, a few copies of the following cones:

Nonnegative Orthant: the set of nonnegative vectors, $\mathbb{R}_+^n$.

Lorentz Cone: the set $\mathbb{L}_{n+1} = \{(u_0, u) \in \mathbb{R}_+ \times \mathbb{R}^n : u_0 \geq \|u\|\}$, also called the *quadratic* or *ice-cream* cone.

Rotated Lorentz Cone: the set $\mathbb{L}_{n+1}^r = \{(u_0, u_1, u) \in \mathbb{R}_+ \times \mathbb{R}^n : u_0 u_1 \geq \|u\|^2, u_0 \geq 0\}$.

Positive Semidefinite Cone: the cone $\mathbb{PS}^{n \times n}$ of $n \times n$ real symmetric positive semidefinite matrices.

Complex Hermitian Cone: the cone $n \times n$ complex Hermitian positive semidefinite matrices.

The dimensions of the cones forming the product can be arbitrary. For semidefinite optimization, we need to introduce some special notation, since the variables are

matrices:

$$\begin{aligned} \min C \bullet X & \quad \max b^T y \\ A^{(i)} \bullet X = b_i, i = 1, \dots, m & \quad \sum_{i=1}^m A^{(i)} y_i + S = C \\ X \in \mathbb{PS}^{n \times n} & \quad S \in \mathbb{PS}^{n \times n}, \end{aligned} \quad (2)$$

where $X, S, C, A^{(i)} \in \mathbb{R}^{n \times n}$ and $b, y \in \mathbb{R}^m$. For symmetric matrices U and V , the quantity $U \bullet V = \text{Tr } UV$ is a scalar product defined on symmetric matrices, and is identical to the sum of the componentwise products of the matrix elements.

Very often the matrices $A^{(i)}$ are sparse or have some other special structure, which can be exploited in the algorithms. For general results on the theory and algorithms of conic optimization see Wolkowicz *et al.* [1] and Alizadeh and Goldfarb [2] and the references therein.

ALGORITHMS

Interior-point methods (IPMs) are practically the only choice for semidefinite optimization; most of the existing general-purpose solvers fall into this category. PENSDP [3], a modified version of PENNON is the only general-purpose semidefinite programming solver using a different approach. The implementation of IPMs for conic optimization is more complicated than that for linear optimization, see Borchers [4], Sturm [5], and Toh *et al.* [6] for more details.

General Features and Capabilities

For generic, dense, unstructured semidefinite optimization problems, the following limits apply on a current PC. The number of linear equalities can be $m \leq 10,000$, using a few matrix variables each with dimensions $n_i \leq 5000$. Larger problems can be solved if the problem has some special structure (see

the section titled “Exploiting the Structure of the Problems”). One of the most serious limitations of IPMs for semidefinite programming is that the Newton system tends to be dense even if the original problem is very sparse. This is due to the symmetric scaling introduced in forming the Newton system, and thus it limits the problem size for all methods that actually form the Newton system.

Computational Cost

The computation cost of one iteration of IPMs for the semidefinite programming problem (2) is $\mathcal{O}(m^3 + mn^3 + m^2n^2)$. Depending on the dimensions of the cones, any of the three terms can dominate this expression, since m can be as large as n^2 . If the problem is sparse, then the second and third terms can usually be improved. Alternatively, one can decide not to form the Newton system explicitly, and try to use an iterative method to solve the Newton system without forming it. For second-order cone programming, the cost per iteration for a problem with K cones each of dimension n and m linear equalities is $\mathcal{O}(m^3 + m^2n + Kn^2)$.

Initialization

IPMs require a strictly feasible solution to start the algorithm. However, if the initial solution is far from the central path, then the algorithm will progress very slowly, thus the standard way of initializing is not practical. There are two techniques that are used widely to circumvent this:

Self-Dual Embedding. This technique embeds the original problem in a larger problem, which will have a strictly feasible starting solution on the central path. From the optimal solution of the larger problem one can recover an optimal solution of the original problem, or detect infeasibility [7–12].

Infeasible IPMs. This method [6,13,14] starts with an infeasible solution and works toward improving feasibility and optimality simultaneously.

INPUT FORMATS

Text-Based Formats

One of the earliest input formats for general symmetric cone programming is the one introduced in SDPpack. This solver is now obsolete, but there are converters from the SDPpack format to SeDuMi’s binary format, for example. The sparse SDPA format is a simplified version of the SDPpack format and has become the standard format for general semidefinite programming problems. It is supported by most major solvers and can be easily converted to any other format.

Binary Formats

Solvers that are implemented in Matlab use a standard Matlab data file as their input, but the actual structure of the data is different between solvers. Also, Python-based solvers can store their data in a pickled file. These formats are special to the given solver and are not portable between different solvers.

Modeling Language Support

Unfortunately, commercial modeling languages do not support SDP or cone programming in general, thus limit their use in the commercial sector. Second-order conic optimization is in a slightly better situation, since it is easily formulated, but there are only very few specialized solvers available. Of course, the best way to formulate these problems would be an indirect one: the user should not even know that the problem is solved using conic programming. Robust optimization fits this scheme very well.

There are two widely used open-source modeling languages that support conic optimization: Yalmip (<http://control.ee.ethz.ch/~joloef/yalmip.php>) [15] and CVX (<http://www.stanford.edu/boyd/cvx>) [16,17]. Both these packages are written in Matlab and are interfaced to a variety of conic optimization packages.

COIN Optimization Services

The optimization services (OS) interface in COIN-OR has been extended [18] to model general conic optimization problems. Driven

by applications and practical problems, the format provides ways to express problems in the most natural way, preserving the important structure of the problem.

There are some key differences between problem instance representation in COIN-OR and modeling languages. Modeling languages tend to separate problem data from problem structure, which prevents exploiting the structure lying within the data. In contrast, COIN-OR keeps the data and the structure together.

CURRENT SOFTWARE PACKAGES

In the following we give an overview of the existing cone optimization solvers in alphabetical order.

CPLEX

License: Commercial
Developer: IBM
Capabilities: LP, SOCP
Algorithm: Interior-point method
Special features: Mixed-integer SOCP
Website: <http://www.ibm.com>
Reference: [19]

CPLEX solves second-order conic problems by treating them as special (nonconvex) quadratically constrained optimization problems. It is also one of the few solvers that can handle mixed-integer conic problems.

CSDP

License: Open source
Developer: Brian Borchers
Capabilities: LP, SDP
Algorithm: Interior-point method
Special features: Callable library, Interface to R
Website: <http://projects.coin-or.org/Csdp>
Reference: [4,20]

Parallel versions in both shared [21] and distributed [22] memory configurations are available. CSDP is one of the most advanced solvers, it exploits sparsity in the data matrices $A^{(i)}$.

CVXOPT

License: Open source
Developer: Joachim Dahl, Lieven Vandenbergh
Capabilities: LP, SOCP, SDP
Algorithm: Interior-point method
Special features: Exploits triangular, band, and sparse matrices
Website: <http://abel.ee.ucla.edu/cvxopt/>
Reference: [23]

A fairly new package written in Python, with interfaces to BLAS, LAPACK, sparse matrix libraries, and NumPy. The interfaces can also be used independent of the solver. The package is related to the modeling interface CVXMOD.

DSDP

License: Free
Developer: Steve Benson, Yinyu Ye, and Xiong Zhang
Capabilities: LP, SDP
Algorithm: Interior-point method
Special features: Dual scaling
Website: <http://www.mcs.anl.gov/hs/software/DSDP>
Reference: [24–26]

DSDP is using a dual scaling algorithm and is very efficient when the dual slack variable is sparse. It is the only interior-point implementation that is not using a primal–dual approach.

LMIlab

License: Commercial
Developer: The Mathworks
Capabilities: SDP, LMI
Algorithm: Interior-point method
Special features: Exploiting a Kronecker structure
Website: <http://mathworks.com>
Reference: [27,28]

A Matlab toolbox, later renamed as Robust Control Toolbox, implements a projective algorithm of Nesterov and Nemirovski. This is the only solver that can exploit a Kronecker structure in the problems, and is thus very efficient for optimal control problems, which is its intended use.

LOQO

License: Commercial
 Developer: Robert Vanderbei
 Capabilities: LP, SOCP
 Algorithm: Interior-point method
 Special features: Nonlinear programming
 Website: <http://www.princeton.edu/rvdb/loqo>
 Reference: [29]

LOQO does solve second-order conic optimization problems, but it uses a different approach. It handles the constraint $x_1 - \|x_{2:n}\|_2 \geq 0$ as a general nonlinear constraint, with some extra care taken due to the nondifferentiability of this form. In a similar way, other nonlinear programming solvers can solve SOCO problems at least in principle.

MOSEK

License: Commercial
 Developer: Erling Andersen, Mosek ApS
 Capabilities: LP, SOCP
 Algorithm: Interior-point method
 Special features: Mixed-integer SOCP, nonlinear programming
 Website: <http://www.mosek.com>
 Reference: [30]

MOSEK is a commercial solver for second-order cone and nonlinear problems. It also implements methods to solve mixed-integer problems.

PENSDP

License: Commercial
 Developer: Michal Kočvara
 Capabilities: SDP
 Algorithm: Augmented Lagrangian method
 Special features: Nonlinear SDP
 Website: <http://www.penopt.com/pensdp.html>
 Reference: [3]

A modified version of the PENNON nonlinear programming solver. It is very efficient even for large, sparse problems. This is the only truly general-purpose sonic solver not implementing an IPM.

SDPA

License: Open source
 Developers: Katsuki Fujisawa, Mituhiro Fukuda, Yoshiaki Futakata, Kazuhiro Kobayashi, Masakazu Kojima, Kazuhide Nakata, Maho Nakata, and Makoto Yamashita
 Capabilities: LP, SDP
 Algorithm: Interior-point method
 Special features: Callable library, distributed/shared-memory parallel version, arbitrary precision, sparsity/decomposition
 Website: <http://sdpa.indsys.chuo-u.ac.jp/sdpa/software.html>
 Reference: [31–34]

This is the one of the most advanced packages for semidefinite optimization developed by a large research group in Japan. Efficient parallel versions [35,36] are released for both shared and distributed memory environments. SDPA also implements techniques to decompose sparse problems using matrix completion [37–39] methods. A Matlab interface [40] is also available.

SDPlr

License: Open source
 Developer: Samuel Burer, Renato D.C. Monteiro, and Changhui Choi
 Capabilities: LP, SDP
 Algorithm: Interior-point method
 Special features: Special handling of low-rank coefficient matrices
 Website: <http://dollar.biz.uiowa.edu/sburer/>
 Reference: [41–43]

SDPlr puts special emphasis on exploiting the special low-rank structure of the coefficient matrices and variables. It is very efficient on some special problems, arising from, for example, combinatorics.

SDPT3

License: Open source
 Developer: Kim Toh, Michael J. Todd, and Reha Tütüncü
 Capabilities: LP, SOCP, SDP
 Algorithm: Interior-point method
 Special features: Logarithmic objective function, mixed second-order, and semidefinite constraints
 Website: <http://www.math.nus.edu.sg/~matttohkc/sdpt3.html>
 Reference: [14,44,45]

SDPT3 is written in Matlab and C and can handle problems with mixed nonnegative, second-order, and semidefinite variables. It is also one of the few solvers that can handle a logarithmic term ($\log x$, $\log(x_0 - \|x\|)$ or $\log \det(X)$ depending on the cone) in the objective function simply by internally manipulating the central path parameter in the IPM.

SeDuMi

License: GPL
 Developer: Jos Sturm (1998–2003) and Imre Pólik (since 2005)
 Capabilities: LP, SOCP, SDP

Algorithm: Interior-point method
 Special features: Mixed second-order and semidefinite constraints
 Website: <http://sedumi.ie.lehigh.edu>
 Reference: [46,47]

SeDuMi is one of the first widely successful conic optimization packages. It is written in Matlab and C. Its success is due to its simple user interface and its numerical accuracy and robustness. It is one of the few solvers that use a self-dual embedding instead of an infeasible scheme to initialize the problem. It implements most of the techniques discussed in Ref. 5.

Following the tragic death of Jos Sturm in 2003, the development has been continued by Imre Pólik.

SMCP

License: Open source
 Developers: M. S. Andersen, J. Dahl, and L. Vandenberghe
 Capabilities: LP, SDP
 Algorithm: Interior-point method
 Special features: Special handling of sparse problem data
 Website: <http://abel.ee.ucla.edu/smcp>
 Reference: [48]

It is a new research code, currently in experimental stage. It exploits the sparsity of the problem using chordal graphs and decomposition techniques, and is written in Python and C.

OBSOLETE PACKAGES

Here is a list of some packages that are no longer developed and maintained.

SBmethod

Written in C++ by Christoph Helmberg; it implements a spectral bundle algorithm [49,50] for semidefinite optimization. Not very efficient as a general-purpose solver. Last updated in 2004.

SDPHA

A Matlab package for SDP by F. A. Potra, R. Sheng, and N. Brixius. No longer available.

SDPpack

Launched in 1997, SDPpack was the first package to solve mixed SOCP and SDP problems. It was written by Farid Alizadeh, Jean-Pierre A. Haeberly, Madhu V. Nayakkankuppam, Michael L. Overton, and Stefan Schmieta. It introduced an input format combining both sparse and dense storage schemes. The SDPA format is a simpler version of this, using only sparse data and SDP constraints. It is the only package that uses the AHO scaling technique, developed by three of the authors. Available from <http://cs.nyu.edu/overton/software/sdppack.html>.

Sdpsol

Developed by Wu and Boyd [51,52], it was solving SDPs and determinant maximization problems. It was last updated in 1996. Available from http://www.stanford.edu/boyd/old_software/SDPSOL.html.

SOCP.C

A simple SOCP solver [53] written by Miguel Sousa Lobo, Lieven Vandenberge, Stephen Boyd, and Herve Lebret; it was last updated in 1997. Available from http://www.stanford.edu/boyd/old_software/SOCP.html. This is the only open-source SOCP solver written in C.

SP

SP (<http://www.ee.ucla.edu/vandenbe/sp.html>), released in 1994 by Lieven Vandenberge, was the first freely distributed SDP package. It used a Matlab/C implementation of Nesterov and Todd's primal-dual potential reduction method.

CURRENT RESEARCH TRENDS AND FUTURE DIRECTIONS

It is commonly agreed that for unstructured or fully dense semidefinite optimization problems, IPMs have reached their limits. In what

follows, we give a brief overview of current efforts and research directions.

Parallelization

IPMs are very good candidates for parallelization, as they take very few iterations, while the cost per iteration is typically very high. There are a number of ways to achieve good parallel performance. As most of the operations in an iteration can be written using standard BLAS/LAPACK calls, parallelization is fairly easy for dense data on a shared-memory architecture. Another approach is to use OpenMP to automatically parallelize loops in the code. This approach is used by CSDP [4].

SDPA and CSDP have also been extended to distributed memory clusters [22,32]. These approaches distribute the computation of the Newton system.

Exploiting the Structure of the Problems

Current research focuses on uncovering and exploiting the special structure of the problem to reduce the memory requirements and speed up the computations; see de Klerk [54] for a survey. The following techniques are known.

Chordal Decomposition. These methods [34,36–39] exploit the fact that the dual slack variable S inherits the sparsity structure of the coefficient matrices. By enforcing positive semidefiniteness only on a small number of carefully chosen submatrices of a large sparse matrix, one can make sure that the original matrix can be completed to be positive semidefinite. This can lead to a dramatic reduction of the problem size, but, in general, it may take a long time to find a good decomposition. Also, the decomposed problem may take a longer time to solve, especially if a lot of submatrices are used. On the other hand, the reduced problem can be solved with a standard SDP solver.

SDPA and SMCP are the two solvers that implement this technique in their preprocessing routines. The developers of SDPA have also released a stand-alone converter useable with any compatible solver.

Permutations and Group Symmetry. Even if the problem is fully dense, the structure induced by the physical problem can be exploited to reduce the size of the matrices. In particular, invariance under permutations has been shown to have a dramatic impact. See de Klerk and Sotirov [55] and Bai *et al.* [56] for applications in quadratic assignment and truss-topology design problems.

There is currently no solver that would try to uncover symmetry in the problem. In fact, without knowing anything about the background of a particular problem instance, it is very hard to detect this kind of structure. It is left to the modeler to use the more efficient representation. On the other hand, the reduced problem does not require any special SDP solvers.

There are no known software tools in this area yet.

Special Linear Operators. These techniques exploit the special structure of the linear operator $\mathcal{A}$. In general $\mathcal{A}$ maps an $n \times n$ matrix to an m -dimensional vector, thus its size is mn^2 and it takes $\mathcal{O}(mn^2)$ operations to apply it to an $n \times n$ matrix. There are some cases when the structure of the operator allows to save in both the storage and computational cost. A few examples are as follows:

Kronecker Products. If $\mathcal{A}(X) = AXB$ in problem (2), then $\mathcal{A}$ is mapping an $n \times n$ matrix to another $n \times n$ matrix. In general, such an operator is represented by the Kronecker product $B^T \otimes A$, which is an $n^2 \times n^2$ matrix. This only allows problems with n at most 100. Moreover, not only we can save in storage, but this special structure can also be exploited inside the IPM to obtain an $\mathcal{O}(n^4)$ cost per iteration instead of the standard $\mathcal{O}(n^6)$. This structure is very common in optimal control problems [57].

Low-Rank Coefficients. If the coefficient matrices A_i are of low rank, then $A_i \bullet X$ can be simplified quite a bit. For example, if $A_i = aa^T$, then $A_i \bullet X = a^T Aa$, and also this structure can be exploited inside the IPM to

speed up the computation of the Newton system. The solvers LMilab [27], SDPlr [41], and modified versions of CSDP [22] can take advantage of this format.

This technique faces a few challenges. First, it is not immediately clear how structures like this can be expressed in the existing input formats. Some tried to modify the SDPA format to accommodate low-rank coefficient matrices [22]. Modeling systems CVX and Yalmip can express these problems, but they convert them into standard SDP form before passing them on to the solvers. The extension of COIN-OR [18] to cone programming gives both the modeler and the solver access to this structure.

The second problem is that one needs to modify the solvers to be able to exploit these kinds of structures. LMilab was created with this goal in mind and it remains a very efficient solver for problems in control theory as part of the Robust Control Toolbox in Matlab.

Graph Problems. As combinatorial optimization and graph theory is a fruitful area of application for SDP, it is not surprising that problems defined over a graph have a very special structure. In fact, it is possible to rewrite the internal computations of an IPM in terms of the graph data. SDPlr is currently the only solver that can exploit this structure. This article is closely related to the following articles in the encyclopaedia are *Interior-Point Linear Programming Solvers* and *Semidefinite Optimization Applications*.

REFERENCES

1. Wolkowicz H, Saigal R, Vandenberghe L, editors. Handbook of semidefinite programming: theory, algorithms, and applications. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000.
2. Alizadeh F, Goldfarb D. Second-order cone programming. Math Program Ser B 2002;95:3–51.
3. Kočvara M, Stingl M. PENSDP User's Guide (Version 2.2). PENOPT GbR; 2006.
4. Borchers B. CSDP, A C library for semidefinite programming. Optim Methods Softw 1999;11(1):613–623.

5. Sturm JF. Implementation of interior point methods for mixed semidefinite and second order cone optimization problems. *Optim Methods Softw* 2002;17(6):1105–1154.
6. Toh KC, Tütüncü RH, Todd MJ. On the implementation of SDPT3 (version 3.1) – a Matlab software package for semidefinite-quadratic-linear programming. *Proceedings of the IEEE Conference on Computer-aided Control System Design*. Taipei, Taiwan: 2004.
7. Jansen B, Roos C, Terlaky T. The theory of linear programming: Skew symmetric self-dual problems and the central path. *Optimization* 1994;29:225–233.
8. Roos C, Terlaky T, Vial J-Ph. *Theory and algorithms for linear optimization: an interior approach*. 2nd ed. New York: Springer; 2006.
9. Ye Y, Todd MJ, Mizuno S. An $O(\sqrt{n}L)$ -iteration homogeneous and self-dual linear programming algorithm. *Math Oper Res* 1994;19:53–67.
10. Freund RM. On the behavior of the homogeneous self-dual model for conic convex optimization. *Math Program* 2006;106(3):527–545.
11. de Klerk E, Roos C, Terlaky T. Infeasible-start semidefinite programming algorithms via self-dual embeddings. In: Pardalos PM, Wolkowicz H, editors. *Volume 18, Topics in semidefinite and interior point methods*, Fields Institute Communications. Providence (RI): AMS; 1998. pp. 215–236.
12. Luo Z-Q, Sturm JF, Zhang S. Conic linear programming and self-dual embedding. *Optim Methods Softw* 2000;14:169–218.
13. Wright SJ. *Primal-dual interior-point methods*. Philadelphia (PA): SIAM; 1997.
14. Tütüncü RH, Toh KC, Todd MJ. Solving semidefinite-quadratic-linear programs using SDPT3. *Math Program Ser B* 2003;95:189–217.
15. Löfberg J. YALMIP: a toolbox for modeling and optimization in MATLAB. *Proceedings of the 2004 IEEE International Symposium on Computer Aided Control Systems Design*. Taipei, Taiwan: 2004. pp. 284–289.
16. Grant M, Boyd S. CVX: Matlab software for disciplined convex programming. Web page and software; 2008. Available at <http://stanford.edu/boyd/cvx>.
17. Grant M, Boyd S. Graph implementations for nonsmooth convex programs. In: Blondel V, Boyd S, Kimura H, editors. *Recent advances in learning and control (a tribute to M. Vidyasagar)*. London: Springer; 2008.
18. Gassmann H, Ma J, Martin K, *et al.* Extending COIN-OR to model conic optimization problems; 2010. In press.
19. Bixby RE. Solving real-world linear programs: a decade and more of progress. *Oper Res* 2002;50(1):3–15.
20. Borchers B. CSDP 2.3 user's guide. *Optim Methods Softw* 1999;11(1):597–611.
21. Borchers B, Young JG. Implementation of a primal-dual method for SDP on a shared memory parallel architecture. *Comput Optim Appl* 2007;37(3):355–369.
22. Ivanov ID, de Klerk E. Parallel implementation of a semidefinite programming solver based on CSDP in a distributed memory cluster. *CentER Discussion Paper* 2007-20. The Netherlands: Tilburg University; 2007.
23. Dahl J, Vandenberghe L. CVXOPT: a python package for convex optimization. *Proceedings of the European Conference on Operational Research*. Reykjavik, Iceland: 2006.
24. Benson SJ, Ye Y. DSDP5: software for semidefinite programming. Technical Report ANL/MCS-P1289-0905. Argonne (IL): Mathematics and Computer Science Division, Argonne National Laboratory; 2005. Submitted to *ACM Transactions on Mathematical Software*.
25. Benson SJ, Ye Y, Zhang X. Solving large-scale sparse semidefinite programs for combinatorial optimization. *SIAM J Optim* 2000;10(2):443–461.
26. Benson SJ, Ye Y. DSDP5 user guide — software for semidefinite programming. Technical Report ANL/MCS-TM-277. Argonne (IL): Mathematics and Computer Science Division, Argonne National Laboratory; 2005. Available at <http://www.mcs.anl.gov/benson/dsdp>.
27. Nemirovski A, Gahinet P. The projective method for solving linear matrix inequalities. *Proceedings of the American Control Conference*. Baltimore (MD): 1994. pp. 840–844.
28. Nesterov YE, Nemirovski A. *Volume 13, Interior-point polynomial algorithms in convex programming*, SIAM studies in applied mathematics. Philadelphia (PA): SIAM Publications; 1994.
29. Vanderbei RJ. *LOQO User's Guide – Version 4.05*. Princeton (NJ): Princeton University, School of Engineering and Applied Science, Department of Operations Research and Financial Engineering; 2006.
30. Andersen ED, Jensen B, Sandvik R, *et al.* The improvements in MOSEK version 5. Technical

- report 1-2007, MOSEK ApS, Fruebjergvej 3 Box 16, 2100 Copenhagen, Denmark; 2007.
31. Fujisawa K, Kojima M, Nakata K, *et al.* SDPA (SemiDefinite Programming Algorithm) user's manual — version 6.2.0. Research Report B-308. Japan; Department of Mathematics and Computer Science, Tokyo Institute of Technology; 1995. Revised in 2004.
 32. Yamashita M, Fujisawa K, Kojima M. Implementation and evaluation of SDPA 6.0 (SemiDefinite Programming Algorithm 6.0). *Optim Methods Softw* 2003;18:491–505.
 33. Fujisawa K, Fukuda M, Kojima M, *et al.* Numerical evaluation of the SDPA (SemiDefinite Programming Algorithm). In: Frenk H, Roos K, Terlaky T, *et al.*, editors. High performance optimization. Dordrecht, The Netherlands: Kluwer Academic Press; 2000. pp. 267–301.
 34. Fujisawa K, Kojima M, Nakata K. Exploiting sparsity in primal-dual interior-point methods for semidefinite programming. *Math Program* 1997;79:235–253.
 35. Yamashita M, Fujisawa K, Kojima M. SDPARA : SemiDefinite Programming Algorithm parAllel version. *Parallel Comput* 2003;29:1053–1067.
 36. Nakata K, Yamashita M, Fujisawa K, *et al.* A parallel primal-dual interior-point method for semidefinite programs using positive definite matrix completion. *Parallel Comput* 2006;32:24–43.
 37. Fujisawa K, Fukuda M, Kojima M, *et al.* SDPA-C (SemiDefinite Programming Algorithm – Completion method) User's Manual — Version 6.10. Research Report B-409. Japan: Department of Mathematics and Computer Science, Tokyo Institute of Technology; 2004.
 38. Fukuda M, Kojima M, Murota K, *et al.* Exploiting sparsity in semidefinite programming via matrix completion I: General framework. *SIAM J Optim* 2001;11:647–674.
 39. Nakata K, Fujisawa K, Fukuda M, *et al.* Exploiting sparsity in semidefinite programming via matrix completion II: implementation and numerical results. *Math Program* 2003;95:303–327.
 40. Fujisawaa K, Futakata Y, Kojima M, *et al.* SDPA-M (SemiDefinite Programming Algorithm in MATLAB) User's manual — version 6.2.0. Research Report B-359. Japan: Department of Mathematics and Computer Science, Tokyo Institute of Technology; 2000. Revised in 2005.
 41. Burer S, Monteiro RDC. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math Program Ser B* 2003;95(2):329–357.
 42. Burer S, Monteiro RDC. Local minima and convergence in low-rank semidefinite programming. *Math Program* 2005; 103(3):427–444.
 43. Burer S, Choi C. Computational enhancements in low-rank semidefinite programming. *Optim Methods Softw* 2006;21(3):493–512.
 44. Toh KC, Todd MJ, Tutuncu RH. SDPT3 – a Matlab software package for semidefinite programming. *Optim Methods Softw* 1999;11:545–581.
 45. Tutuncu RH, Toh KC, Todd MJ. Solving semidefinite-quadratic-linear programs using SDPT3. *Math Program* 2003;95:189–217.
 46. Sturm JF. Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. *Optim Methods Softw* 1999;11–12:625–653.
 47. Sturm JF. Primal-dual interior point approach to semidefinite programming. In: Frenk JBG, Roos C, Terlaky T, *et al.*, editors. High performance optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999.
 48. Andersen MS, Dahl J, Vandenberghe L. Implementation of nonsymmetric interior-point methods for linear optimization over sparse matrix cones. Technical report, 2009. Submitted to Mathematical Programming Computation, 2010, DOI: 10.1007/s12532-010-0016-2.
 49. Helmberg C, Kiwiel KC. A spectral bundle method with bounds. *Math Program* 2002;93(2):173–194.
 50. Helmberg C, Rendl F. A spectral bundle method for semidefinite programming. *SIAM J Optim* 2000;10(3):673–696.
 51. Wu S-P, Boyd S. Sdpsol: a parser/solver for semidefinite programs with matrix structure. In: El Ghaoui L, Niculescu S-I, editors. Recent advances in LMI Methods for Control. Boston (MA): SIAM; 2000. chapter 4. pp. 79–91.
 52. Wu S-P, Boyd S. Design and implementation of a parser/solver for SDPs with matrix structure. Proceedings of the 1996 IEEE International Symposium on Computer-Aided Control System Design (CACSD). Dearborn (MI): 1996. pp. 240–245.
 53. Lobo M, Vandenberghe L, Boyd S, *et al.* Second-order cone programming. Available as part of the SOCP package. 1997.

54. de Klerk E. Exploiting special structure in semidefinite programming: a survey of theory and applications. *Eur J Oper Res* 2010;201(1):1–10.
55. de Klerk E, Sotirov R. Exploiting group symmetry in semidefinite programming relaxations of the quadratic assignment problem. *Math Program* 2010;122(2):225–246.
56. Bai YQ, de Klerk E, Pasechnik DV, *et al.* Exploiting group symmetry in truss topology optimization. *Optim Eng* 2009;10(3):331–349.
57. Boyd S, El Ghaoui L, Feron E, *et al.* Volume 15, Linear matrix inequalities and system and control theory, SIAM studies in applied mathematics. Philadelphia (PA): SIAM; 1994.

CONSERVATION LAWS AND RELATED APPLICATIONS

JOSÉ NIÑO-MORA

Department of Statistics, Carlos III
University of Madrid, Madrid,
Spain

THE PRINCIPLE OF WORK CONSERVATION

Consider a multiclass G/G/1 queue, where a single server attends a finite number n of customer classes labeled by $k \in N \triangleq \{1, \dots, n\}$. Class k customers arrive to the system at rate λ_k with identically distributed interarrival times, while their service requirements are drawn from a distribution having finite mean $1/\mu_k$. No independence assumptions are made on interarrival and service time processes. The stability condition requiring that the system's utilization factor or traffic intensity be less than unity, $\rho \triangleq \sum_{k \in N} \rho_k < 1$, is assumed to hold, where $\rho_k \triangleq \lambda_k/\mu_k$ is the traffic intensity for class k .

In order to determine how to operate such a system, a scheduling policy or service discipline (see *Queueing Disciplines*) must be adopted, which prescribes the priority order in which customers are to be selected for service. Suppose that the system is operated under a scheduling policy that is *work-conserving*, that is, it neither allows the server to sit idle while there are customers waiting, nor does it affect the total service requirement or the arrival time of any customer. Consider the *unfinished work* (also called the *work-in-system* or *workload*) process $U(t)$, which keeps track of the total remaining service time due to customers in the system at each time $t \geq 0$. The evolution over time of such a process can be described as follows: at each arrival epoch, the process has an upward jump of magnitude equal to the arriving customer's service requirement; while the system is nonempty, the process decreases linearly at unit rate; and,

upon hitting zero, it remains there until a customer arrives. The *principle of work conservation*, which follows immediately from the above, states that the unfinished work process $U(t)$ is sample-path invariant under work-conserving scheduling policies.

Further, classic results on the G/G/1 (see *The G/G/1 Queue*) queue ensure existence and uniqueness of a stationary or steady-state distribution for the process $U(t)$, which will have a finite mean $\bar{U}$, provided the service time distribution of each class k has a finite second moment $m_k^{(2)}$. The latter assumption will be adopted in what follows.

KLEINROCK'S CONSERVATION LAW FOR MEAN WAITING TIMES

Consider now a multiclass M/G/1 queue (see *The M/G/1 Queue*) where customer arrival streams are Poisson and all interarrival and service times are mutually independent. Attention is further restricted to the class Π of *admissible* policies that are (i) *work-conserving*; (ii) *nonanticipative* or history-dependent, that is, decisions at any given time can only be based on current or past system information, but not on future information such as future arrival times or remaining service requirements of customers who have not yet departed from the system; (iii) *ergodic*, that is, they induce on system processes, such as the number of customers in system for each class, steady-state distributions with finite means that match the corresponding sample-path averages; and (iv) either *nonpreemptive*, that is, the server allocates its full service capacity to customers one at a time and without interruptions; or (iv) in the multiclass M/M/1 case, *preemptive* of the *preemptive-resume* type, that is the server can temporarily abandon the customer it was serving to attend another customer, continuing later service of the preempted customer at the point it was left off.

Drawing on the principle of work conservation for the unfinished work process and on

Little's law (see **Little's Law and Related Results**) [1], Kleinrock first established in Refs 2 and 3 that, if such a system is operated under an admissible scheduling policy $\pi \in \Pi$, the vector $\bar{\mathbf{W}}^\pi = (\bar{W}_k^\pi)_{k \in N}$ of attained steady-state mean waiting times for the customer classes satisfies a linear equality constraint of the form

$$\sum_{k \in N} \rho_k \bar{W}_k^\pi = g(N),$$

where $g(N) \triangleq \bar{U} - \bar{V} = \rho \bar{V} / (1 - \rho)$ is a constant and

$$\bar{V} \triangleq \frac{1}{2} \sum_{k \in N} \lambda_k m_k^{(2)} = \sum_{k \in N} \rho_k (m_k^{(2)} \mu_k / 2)$$

is the steady-state mean residual time of the customer found in service by a random arrival. Such an invariance relation was termed by Kleinrock a (work) *conservation law*.

An early application of such a conservation law to the analysis of a multiclass M/M/1 queue with delay-dependent priorities is given in Ref. 4. In Ref. 5, pp. 125–126 the conservation law is used to obtain the optimal admissible scheduling policy for operating a multiclass M/G/1 queue as above, where class k customers incur linear holding costs at rate c_k per unit time and the objective is to minimize the steady-state mean cost rate. For such a system, Cox and Smith [6, pp. 84–85] first showed by an interchange argument that among static nonpreemptive priority policies—which assign a fixed priority ranking to the customer classes, as in Ref. 7, an *index policy* is optimal: the so-called *$c\mu$ -rule*, which awards higher priority to classes with larger values of the priority index $c_k \mu_k$. The optimality of the $c\mu$ rule in the class of work-conserving nonpreemptive stationary policies, which make decisions based on the current state, was established in Ref. 8.

The scope of the conservation law is extended in Ref. 9 to a nonpreemptive multiclass G/G/1 queue without independence assumptions on interarrival and service times, with $g(N) \triangleq \bar{U} - \bar{V}$ and $\bar{V}$ as above. Such early results are reviewed in Ref. 5, Section 3.4.

It must be emphasized that, as pointed out in Ref. 10, p. 175, although the nonanticipativity assumption on admissible scheduling policies is not explicitly mentioned in such early works, it is, however, critical to the validity of the above conservation laws.

Extensions of the conservation law are presented in Ref. 11, pp. 431–435 to multiclass multiserver G/G/m queues (see **The G/G/s Queue**), where for each class arrival times form a renewal process and service times are i.i.d., having the same distribution for all classes if $m \geq 2$. Different formulations of the law are given under nonpreemptive and under preemptive-resume policies.

While the above works have focused on continuous-time queueing systems where customers arrive one at a time, conservation laws for multiclass queues in discrete time with batch arrivals are presented in Refs 12 and 13.

WORK CONSERVATION LAWS FOR CONDITIONAL MEAN WAITING TIMES

For queueing systems that are operated under scheduling policies that are of processor-sharing type [5, Chapter 4] (i.e., the server's capacity may be simultaneously shared by multiple customers at given rates), which may also be preemptive (e.g., the foreground-background (FB) or least-attended service (LAS) rule [14]) or anticipative (e.g., the shortest remaining processing time (SRPT) rule [15] or the shortest job first (SJF) rule [16]), an appropriate measure of performance is the steady-state conditional mean response time $\bar{T}^\pi(x)$ for a customer with total service requirement x , as well as the conditional mean waiting time $\bar{W}^\pi(x) \triangleq \bar{T}^\pi(x) - x$.

For a single-class M/G/1 queue operated under any work-conserving processor-sharing nonanticipative policy π , Kleinrock *et al.* [17] first showed that the conditional mean waiting times $\bar{W}^\pi(x)$ obey a conservation law of the form

$$\int_0^\infty \bar{W}^\pi(x) (1 - B(x)) dx = g,$$

where $B(x)$ is the distribution function of service times and $g \triangleq (1/2)\rho m^{(2)}/(1 - \rho)$.

Such a result is used in Refs 17 and 18 to obtain tight bounds, via mathematical programming, on the performance of processor-sharing scheduling policies.

The scope of such a conservation law for processor-sharing systems is extended in Ref. 19 to G/G/1 queues under work-conserving anticipative policies, incorporating the mean conditional time $\bar{T}^\pi(u, x)$ that a customer with total service time x spends in the system to obtain $u \leq x$ units of service.

Conservation laws for conditional mean waiting times for multiserver G/G/m queues are obtained in Ref. 11, pp. 425–427 under nonanticipative processor-sharing policies, which, if $m \geq 2$, are further required to not distinguish between customers on the basis of their service time characteristics.

Kleinrock's conservation law for conditional mean waiting times has been used in Ref. 20 to analyze a processor-sharing queue with batch arrivals, and the results are applied to analyze a multiclass version of multilevel processor-sharing type.

Work conservation laws for a multiclass queue under discriminatory processor-sharing scheduling [21] are obtained in Ref. 22. The literature in the area is reviewed in Ref. 23, which presents a unifying version of the work conservation law.

ACHIEVABLE PERFORMANCE CHARACTERIZATION AND OPTIMIZATION VIA CONSERVATION LAWS

Kleinrock's conservation law implies that the performance vectors $\bar{\mathbf{W}}^\pi$ of mean waiting times which are achieved under admissible scheduling policies π lie on a hyperplane in $\mathbb{R}^n$. Such a geometric viewpoint leads to the following question: What are the constraints characterizing the (achievable) *performance region* $R = \{\bar{\mathbf{W}}^\pi : \pi \in \Pi\}$, which is spanned by achievable performance vectors $\bar{\mathbf{W}}^\pi$ as the scheduling policy π ranges over all admissible policies. Such an issue was first resolved in Ref. 24 for a preemptive multiclass M/M/1 queue under a modified class of admissible policies having a regenerative structure (see **Markov Regenerative Processes**), in which server allocation decisions are required to

depend on the current busy period's history. The performance region under the resulting class of admissible policies is shown to be the convex polytope (i.e., bounded polyhedron) determined by Kleinrock's conservation law (see **Basic Polyhedral Theory**), along with a family of linear inequality constraints of the form

$$\sum_{k \in S} \rho_k \bar{W}_k^\pi \geq g(S),$$

where $S \subset N$ ranges over all nonempty subsets of customer classes and $g(S)$ is a positive constant that depends on S .

The above inequality for subset S holds with equality under admissible policies that give higher priority to customers whose classes belong to S , which leads to the result that the vertices of the performance polytope are the performance vectors achieved by the $n!$ static priority rules. These results are used to synthesize policies that attain given achievable performance vectors.

Similar results are obtained for multiclass nonpreemptive M/G/1 queues in Ref. 10, Chapter 6. These works further address the synthesis problem, either via mixing policies, which at the start of each busy period draw at random from a certain distribution a static priority rule to be used during the period, or via other more convenient (in terms of reduced variances) complete parameterized families of scheduling policies [25].

The performance region is similarly characterized in Ref. 26 for an exponential multiclass finite-source queue, where the performance vector consists of the utilization factors for the customer classes.

In Ref. 27 an analogous approach is deployed to address the problem of characterizing the class of achievable conditional delay functions $\bar{W}^\pi(x)$ for a nonpreemptive M/G/1 queue, under admissible policies that can make use of exact knowledge of customer processing times.

In Ref. 28 sufficient conditions are given under which the performance region of a nonpreemptive M/G/m queue is a polytope of the type identified in Ref. 24; see also the article titled **The M/G/s Queue** in this encyclopedia. Similar results are obtained for preemptive multiclass G/G/m queues in

Ref. 29; see also the article titled *The G/G/s Queue* in this encyclopedia. The required conditions include verification that the right-hand side set function $g(S)$ is supermodular [30]. Under such conditions, the achievable performance polytope is the base of a polymatroid, a much studied polyhedron in polyhedral combinatorics (see *Basic Polyhedral Theory*) introduced in Ref. 31, which possesses strong structural and algorithmic properties and explains the optimality of the greedy algorithm in a wide range of resource allocation problems. The papers [28,29] further exploit such a polyhedral structure to obtain optimal policies for separable convex performance objectives.

Such early results on achievable performance characterization are unified and new results are obtained via the axiomatic framework of so-called *strong (work) conservation laws* introduced in Ref. 32, which gives simpler sufficient conditions, ensuring that the performance region is the base of a polymatroid, with the ensuing algorithmic consequences for performance optimization. Most notably, such conditions do not require *a priori* verification of supermodularity for the right-hand side functions $g(S)$ in the conservation laws. Instead, it suffices to establish that, for any subset S of classes, the above inequality holds, and equality is achieved by any static priority policy that gives higher priority to customers whose classes belong to S .

The polymatroidal structure of the performance region under conservation laws allows deployment of powerful polyhedral combinatorics methods for analyzing algorithmic performance problems, such as determining whether a given performance vector is achievable [33].

The need in some applications to use adaptive policies, which may be nonergodic or for which ergodicity is hard to establish, motivates Georgiadis and Viniotis [34] to extend conservation laws to a multiclass G/G/1 queue under a broader class of admissible policies, which neither assume a regenerative structure nor convergence to steady state. For the multiclass M/G/1 case, extensions of the conservation laws are given in terms of limits of linear combinations of

sample averages. Such results are used to design lexicographically optimal adaptive policies for M/G/1 queues in Ref. 35.

CONSERVATION LAWS AND CONSTRAINED PERFORMANCE OPTIMIZATION

The results on polymatroidal structure of the achievable performance region via conservation laws can be exploited to obtain tractable scheduling policies that optimize linear performance objectives under linear performance constraints.

In Ref. 36, a multiclass M/G/1 queue is used to model a telecommunication system where customer classes correspond to heterogeneous traffic classes, which can be of interactive or noninteractive type. The problem of finding a scheduling policy that minimizes a weighted mean delay objective for noninteractive classes is addressed, subject to constraints that require the mean delays for interactive classes to not exceed given thresholds. Under appropriate conditions, a tractable optimal policy is obtained using linear programming (LP) methods that exploit the polymatroidal formulation obtained via conservation laws. An optimal policy is obtained that partitions the classes into several groups and uses a static priority rule among the groups. Within each group, an appropriate randomization among priority rules is employed.

Such an approach is extended in Ref. 37 to constrained optimization in a multiclass Jackson queueing network (see *The G/G/s Queue*) populated by interactive and noninteractive traffic classes, for which the required conservation laws are established.

The problem of optimal scheduling of a multiclass M/G/1 queueing model of a flexible manufacturing station subject to upper bounds on average delays for job classes, relative to a nonlinear cost objective, is addressed in Ref. 38. Drawing on the polymatroidal achievable performance results, an optimal policy is constructed having a similar structure as in the two previous models discussed above.

In Ref. 39 conservation laws are used to address the problem of joint design of a static load balancing rule and a dynamic local scheduling policy for a queueing model of a distributed computer system. Each of a set of host computers generates a collection of dedicated traffic classes, which can only be processed at the host and are subject to upper bound constraints on their average response times. In addition, hosts generate a generic traffic class, which can be allocated to be processed at any host. For a given static allocation of generic jobs to hosts, each host behaves as a multiclass M/G/1 queue subject to average performance constraints on its dedicated classes. The polymatroidal characterization of the corresponding performance region due to satisfaction of conservation laws is exploited to obtain a tractable policy that minimizes the average response time of generic jobs, subject to given average response bounds on dedicated jobs.

A related although distinct approach is deployed in Ref. 40 to address the constrained optimization of a discrete-time preemptive multiclass single-server queue with geometric service time distributions, where policies are not required to be ergodic. Using Lagrangian methods, the performance region is (implicitly) characterized in terms of that achieved under the class of time-sharing policies, which do not use randomization and may be nonregenerative. Further discussion regarding how to obtain an optimal time-sharing policy that satisfies linear performance constraints follows.

GENERALIZED CONSERVATION LAWS AND OPTIMIZATION FOR KLIMOV'S AND BANDIT MODELS

Klimov's model [41,42] is an extension of the nonpreemptive multiclass M/G/1 queue, which incorporates instantaneous Bernoulli feedback between customer classes (see **Klimov's Model**). Thus, when a class k customer completes service, it is fed back to the system as a class l customer with probability p_{kl} , and leaves the system with probability p_{k0} . In his ground-breaking paper [41], Klimov addressed the problem of

finding an optimal scheduling policy relative to a linear average holding cost objective. He established the optimality of a static priority rule, where priority ranking is determined by an index attached to each class, with larger index values awarding higher priority. For such a purpose, he used flow balance arguments to obtain an LP formulation of the optimal scheduling problem, whose analysis yields an efficient index algorithm along with an optimality proof for the resulting index rule.

An exact polyhedral characterization of the performance region for Klimov's model under ergodic policies, obtained via the principle of work conservation, was first presented in Ref. 43. The linear constraints characterizing the region of achievable mean delays $\bar{\mathbf{W}}^\pi$ consist of an equality constraint of the form

$$\sum_{k \in N} \rho_k^N \bar{W}_k^\pi = g(N),$$

along with a family of inequality constraints of the form

$$\sum_{k \in S} \rho_k^S \bar{W}_k^\pi \geq g(S)$$

for nonempty subsets $S \subset N$ of customer classes, where the ρ_k^N and ρ_k^S are positive workload coefficients. The resulting polytope is no longer the base of a polymatroid, but that of a new type of polytope, a so-called *extended* polymatroid, which also possesses strong structural and algorithmic properties, as elucidated in Ref. 44. Further extensions of conservation laws under nonergodic policies are introduced in Ref. 43 and further developed in Ref. 45, where they are used to obtain adaptive policies for optimizing nonlinear performance objectives in the Klimov model.

The axiomatic framework of strong conservation laws for multiclass queues without feedback in Ref. 32 is extended in Ref. 46 to the framework of generalized conservation laws, which have the form indicated above for the Klimov model. Further, in Ref. 46 such a framework is deployed to obtain the polyhedral performance region of the more general branching bandit model [47,48], both

under average and discounted performance measures, including as a special case the classic multiarmed bandit problem [49]. A new proof is thus obtained via conservation laws and polyhedral LP methods for the optimality of index policies in such classic problems.

The problem of finding an optimal scheduling policy for the branching bandit problem under the average criterion under linear performance constraints, and hence for the Klimov model with such constraints as a special case, is addressed in Ref. 50 via LP methods, exploiting the extension to branching bandits of the LP formulation originally introduced in Ref. 41. Unlike the LP formulation given in Ref. 46, which has $2^n - 1$ constraints on n variables, that in Refs 41 and 50 has $O(n^2)$ constraints on n^2 variables, which renders it more amenable to direct computational approaches.

In Refs 51 and 52 the properties of generalized conservation laws and extended polymatroids are further reviewed and developed, including connections with submodularity and the efficient algorithmic treatment of side constraints. The reader is referred to Ref. 53, Chapter 11 for an accessible textbook account of such methods.

PSEUDOCONSERVATION LAWS FOR MULTICLASS QUEUES WITH SETUP TIMES

In some applications of multiclass queues, such as in flexible manufacturing systems or in computer communication networks, the assumption that the server can switch instantaneously from one class to another is not realistic. This motivates incorporation of random setup times, which are incurred before the server starts to service a class of customers after leaving another class. A major area of application in communication networks is that of polling systems. See, for example, the review paper [54].

A policy π for dynamic server allocation in a multiclass queue with setup times must specify: (i) a service discipline for determining when the server should leave the class it is currently allocated to (e.g., exhaustive service, gated service, etc.); and (ii) a routing rule, which specifies the class to visit

next after leaving a class (e.g., cyclic routing, routing according to a polling table, etc.). Most research in the area has focused on the performance analysis of particular policies, such as cyclic routing with exhaustive service [55]. Early work on optimal dynamic control of polling systems is surveyed in Ref. 56.

For such systems, the principle of work conservation does not apply, since the time intervals spent by the server in switching from one class to another represents a creation of work. Yet, researchers have identified in a variety of such systems so-called pseudoconservation laws. For a multiclass M/G/1 queue as in the section titled “Kleinrock’s Conservation Law for Mean Waiting Times”, which further incorporates random setup times for each class, a pseudoconservation law is an invariance identity of the form

$$\sum_{k \in N} \rho_k \bar{W}_k^\pi = g(N) + \bar{Y}^\pi,$$

where $\bar{Y}^\pi$ is a policy-dependent term related to the expected amount of work in the system at a random time during setups. Such expressions have been used to provide qualitative insight into system behavior, and in approximate performance analyses of complex systems.

Pseudoconservation laws were first identified in Refs 57–60, under cyclic routing policies where each class is serviced according to a certain type of service rule. In Ref. 61 such results are extended to cyclic routing and mixed service (i.e., possibly a different rule is used for each class) policies, and the connection is established between such laws and the stochastic work decomposition results for queues with vacations (cf. Refs 62 and 63). The latter state that, under certain conditions, the steady-state amount of work in a system with service interruptions can be decomposed as the sum of two independent random variables: the amount of work in a corresponding system without interruptions, and the amount of work-in-system at an arbitrary epoch in a switching interval.

Pseudoconservation laws for cyclic multiclass discrete-time queues with setup

times are given in Ref. 64, and for systems where routing is specified by a noncyclic polling table in Ref. 65. Such early work is surveyed in Ref. 66. In Ref. 67, corresponding results are obtained in the setting of a version of Klimov's model, which incorporates setup times under cyclic policies. Pseudoconservation laws are obtained under static priority rules in Refs 68 and 69, and under more general priority rules in Ref. 70 for systems with batch arrivals. In Ref. 71, general pseudoconservation laws are obtained for Klimov's model with setup times under arbitrary dynamic policies, by reformulating linear flow balance constraints on performance measures in terms of workloads. Such laws are used to obtain a polyhedral LP relaxation of the performance region yielding bounds on optimal performance, determined by constraints of the form

$$\sum_{k \in S} \rho_k^S \bar{W}_k^\pi \geq f(S),$$

where $S \subseteq N$ ranges over subsets of customer classes and $f(S)$ is a positive constant that depends on S .

CONSERVATION LAWS AND PERFORMANCE BOUNDS FOR MULTICLASS QUEUEING NETWORKS

Multiclass multistation queueing networks (see **Multiclass Queueing Network Models**) are powerful models for a wide variety of real systems, most notably computer communication and flexible manufacturing networks, where multiple traffic classes vie for access to system resources (e.g., servers' attention). The performance of such systems, such as the vector of average response times for each class, can be significantly affected by the choice of control policy adopted for dynamic resource allocation (e.g., scheduling, routing). Yet, the exact performance analysis of even relatively simple policies that discriminate among customer classes appears both analytically and computationally intractable. As a result, simulation has been the primary tool for investigating such system models. This motivates the interest of obtaining tractable bounds on

achievable performance, which can be used to assess the degree of suboptimality of proposed policies (see **Performance Bounds in Queueing Networks, The G/G/s Queue**).

A relaxed LP formulation for scheduling problems on such networks, which extends Klimov's original LP formulation [41] and furnishes tractable bounds on optimal performance, was introduced in Ref. 72. A different approach to obtain such bounds for Markovian multiclass queueing networks based on LP relaxations is proposed in Ref. 73, based on Lyapunov function ideas, along with stronger bounds based on nonlinear convex and semidefinite programming constraints. Similar LP relaxations are proposed in Ref. 74, drawing on an analysis of quadratic functions of workload. The bounds are shown to yield good approximations in some networks. The LP equality constraints in both papers are derived in Ref. 75 via flow balance identities by using Palm calculus (cf. Ref. 76). The latter paper further obtains linear and convex nonlinear and semidefinite programming relaxations for Markovian multiclass queueing networks with random setup times.

A polymatroidal LP relaxation of the performance region is used to analyze the version of Klimov's model with multiple exponential servers in parallel in Ref. 77. The LP constraints are obtained in terms of approximate work conservation laws, which emerge by reformulating flow balance identities. The LP relaxation is used to obtain closed-form bounds on the suboptimality gap of the parallel version of Klimov's priority-index rule, which are tight enough to establish the asymptotic optimality of such a rule in the heavy-traffic regime.

The approach of addressing performance optimization problems in stochastic systems via exact or relaxed mathematical programming formulations of their performance regions has been termed the *achievable region approach*. An accessible presentation along with further examples of deployment are given in Ref. 78, followed by discussions by several researchers.

While most work on conservation laws has focused on steady-state expectations, it is of interest in some applications to exploit the sample-path nature of the principle of work conservation. This has been done in Ref. 79, both for certain examples of multiclass queueing systems and for some stochastic fluid models of such systems.

The analysis of a fluid model of the Klimov network via sample-path work conservation and the achievable region approach is carried out in Ref. 80, where it is shown that a priority-index scheduling rule is optimal for linear performance objectives. The achievable region formulation of the fluid Klimov network is employed to obtain a tractable scheduling control policy for a linear performance objective under side constraints in Ref. 81.

A number of papers have identified and exploited work conservation laws to analyze and optimize the performance of more specific complex queueing models arising in communication networks. Some examples of relevant work in the area are Refs 82–94.

PARTIAL CONSERVATION LAWS AND INDEXABILITY FOR RESTLESS BANDIT MODELS

Multiarmed restless bandits are an extension of classic (nonrestless) multiarmed bandits, concerning the optimal dynamic allocation of effort to multiple stochastic projects, modeled as binary-action (active/passive) Markov decision processes. In the restless model, projects can change state under passive dynamics, while in the classic model the states of passive projects remain frozen. Many dynamic resource allocation problems in queueing systems can be readily formulated as multiarmed restless bandit problems.

In Ref. 95, Whittle introduced the model and proposed a tractable heuristic priority-index rule via a Lagrangian relaxation approach, which further yields a bound on optimal performance, as the problem is generally intractable. Unlike its counterpart for classic bandits, however, such an index only exists for the limited class of the

so-called indexable restless bandits. In Ref. 96 the asymptotic optimality of the resulting index policy is established under certain conditions. In Ref. 97, a hierarchy of LP relaxations is introduced using flow balance equations, which yield tighter performance bounds at increasing computational expense.

The issues of finding simple sufficient conditions for indexability and index computation were first addressed in Ref. 98 via the introduction of so-called partial (work) conservation laws, which extend the generalized conservation laws framework in Ref. 46. Such laws require the satisfaction by performance measures x_k^π of an equality constraint

$$\sum_{k \in N} \rho_k^N x_k^\pi = g(N),$$

as well as of inequality constraints

$$\sum_{k \in S} \rho_k^S x_k^\pi \geq g(S),$$

where the class subsets S are only required to range over a typically restricted family $\mathcal{F}$ of customer classes satisfying appropriate connectivity requirements. Such a framework is designed to exploit special structure and insight on the model at hand, as the family $\mathcal{F}$ must be guessed in advance. Satisfaction of partial conservation laws yields a relaxed LP formulation of the achievable performance region, which can be used to obtain optimal policies for a limited range of linear performance objectives. Such conservation laws, which are applied to discounted and long-run average expected occupation measures of single restless bandits, allow to establish the existence of index-type solutions, which are consistent with a prespecified structure on policies (e.g., threshold policies), and further yield an efficient adaptive-greedy index-computing algorithm, by exploiting the underlying polyhedral LP formulation. Such a framework of partial conservation laws is further developed in Ref. 99, where applications are given to dynamic admission control and routing to parallel queues, and in Ref. 100, which presents an application to scheduling a multiclass make-to-order/make-to-stock M/G/1 queue under convex

stock holding and backorder costs. See also Refs 101–104. These and other applications of restless bandit indexation and partial conservation laws are reviewed in Ref. 105.

REFERENCES

1. Whitt W. A review of $L = \lambda W$ and extensions. *Queueing Syst* 1991;9(3):235–268.
2. Kleinrock L. *Communication nets: stochastic message flow and delay*. New York: McGraw-Hill; 1964. Reprinted by New York: Dover Publications, Inc.; 1972.
3. Kleinrock L. A conservation law for a wide class of queueing disciplines. *Naval Res Logist Quart* 1965;12(2):181–192.
4. Kleinrock L, Finkelstein RP. Time dependent priority queues. *Oper Res* 1967;15(1):104–116.
5. Kleinrock L. *Queueing systems. Volume II, Computer applications*. New York: John Wiley & Sons, Inc.; 1976.
6. Cox DR, Smith WL. *Queues*. London: Methuen; 1961.
7. Cobham A. Priority assignment in waiting line problems. *Oper Res* 1954;2(1):70–76.
8. Fife DW. Scheduling with random arrivals and linear loss functions. *Manag Sci* 1965;11(3):429–437.
9. Schrage L. An alternative proof of a conservation law for the queue G/G/1. *Oper Res* 1970;18(1):185–187.
10. Gelenbe E, Mitrani I. *Analysis and synthesis of computer systems*. New York: Academic Press; 1980.
11. Heyman DP, Sobel MJ. *Stochastic models in operations research. Volume I, Stochastic processes and operating characteristics*. New York: McGraw-Hill; 1982. Reprinted by Mineola (NY): Dover Publications Inc.; 2004.
12. Takahashi Y, Hashida O. Delay analysis of discrete-time priority queue with structured inputs. *Queueing Syst* 1991;8(2):149–163.
13. Bisdikian C. A note on the conservation law for queues with batch arrivals. *IEEE Trans Commun* 1993;41(6):832–835.
14. Nuyens M, Wierman A. The foreground-background queue: a survey. *Perform Eval* 2008;65(3–4):286–307.
15. Schrage LE, Miller LW. The queue M/G/1 with the shortest remaining processing time discipline. *Oper Res* 1966;14:670–684.
16. Phipps TE Jr. Machine repair as a priority waiting-line problem. *Oper Res* 1956;4(1):76–85.
17. Kleinrock L, Muntz RR, Hsu J. Tight bounds on the average response time for time-shared computer systems. *Information Processing 71 (Proceedings IFIP Congress, Ljubljana, 1971), Volume 1, Foundations and Systems*. Amsterdam: North-Holland Publishing Co.; 1972. pp. 124–133.
18. Kleinrock L, Nilsson A. On optimal scheduling algorithms for time-shared systems. *J Assoc Comput Mach* 1981;28(3):477–486.
19. O'Donovan TM. Distribution of attained and residual service in general queueing systems. *Oper Res* 1974;22(3):570–575.
20. Avrachenkov K, Ayesta U, Brown P. Batch arrival processor-sharing with application to multi-level processor-sharing scheduling. *Queueing Syst* 2005;50(4):459–480.
21. Altman E, Avrachenkov K, Ayesta U. A survey on discriminatory processor sharing. *Queueing Syst* 2006;53(12):53–63.
22. Avrachenkov K, Ayesta U, Brown P, *et al.* Discriminatory processor sharing revisited. *Proceedings of the 24th Annual Joint Conference IEEE Computer and Communications Societies (INFOCOM 2005)*. Los Alamitos (CA): IEEE; 2005. pp. 784–795.
23. Ayesta U. A unifying conservation law for single-server queues. *J Appl Probab* 2007;44(4):1078–1087.
24. Coffman EG Jr, Mitrani I. A characterization of waiting time performance realizable by single-server queues. *Oper Res* 1980;28(3, part 2):810–821.
25. Mitrani I, Hine JH. Complete parameterized families of job scheduling strategies. *Acta Informat* 1977;8(1):61–73.
26. Kameda H. Realizable performance vectors of a finite-source queue. *Oper Res* 1984;32(6):1358–1367.
27. Mitrani I. On the delay functions achievable by non-preemptive scheduling strategies in m/g/1 queues. In: Dempster MAH, Lenstra JK, Rinnooy Kan AHG, editors. *Deterministic and stochastic scheduling*. Dordrecht, The Netherlands: D. Reidel; 1982. pp. 399–404.
28. Federgruen A, Groenevelt H. M/G/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules. *Manag Sci* 1988;34(9):1121–1138.
29. Federgruen A, Groenevelt H. Characterization and optimization of achievable performance in general queueing systems. *Oper Res* 1988;36(5):733–741.

30. Fujishige S. Submodular functions and optimization. Volume 58, Annals of discrete mathematics. 2nd ed. Amsterdam: Elsevier; 2005.
31. Edmonds J. Matroids and the greedy algorithm. Math Program 1971;1(1):127–136.
32. Shanthikumar JG, Yao DD. Multiclass queueing systems: polymatroidal structure and optimal scheduling control. Oper Res 1992;40, Suppl 2:S293–S299.
33. Itoko T, Iwata S. Computational geometric approach to submodular function minimization for multiclass queueing systems. In: Fischetti M, Williamson DP, editors. Volume 4513, Integer Programming and Combinatorial Optimization (IPCO) 2007, LNCS. Berlin: Springer; 2007. pp. 267–279.
34. Georgiadis L, Viniotis I. On the conservation law and the performance space of single server systems. Oper Res 1994;42(2):372–379.
35. Bhattacharya PP, Georgiadis L, Tsoucas P, *et al.* Adaptive lexicographic optimization in multi-class M/GI/1 queues. Math Oper Res 1993;18(3):705–740.
36. Ross KW, Chen B. Optimal scheduling of interactive and noninteractive traffic in telecommunication systems. IEEE Trans Automat Contr 1988;33(3):261–267.
37. Ross KW, Yao DD. Optimal dynamic scheduling in Jackson networks. IEEE Trans Automat Contr 1989;34(1):47–53.
38. Yao DD, Shanthikumar JG. Optimal scheduling control of a flexible machine. IEEE Trans Robot Automat 1990;6(6):706–712.
39. Ross KW, Yao DD. Optimal load balancing and scheduling in a distributed computer system. J Assoc Comput Mach 1991;38(3):676–690.
40. Altman E, Schwartz A. Optimal priority assignment: a time sharing approach. IEEE Trans Automat Control 1989;34(10):1098–1102.
41. Klimov GP. Time-sharing service systems. I. Theor Probab Appl 1974;19(3):532–551.
42. Klimov GP. Time-sharing service systems. II. Theory Probab Appl 1978;23(2):314–321.
43. Tsoucas P. The region of achievable performance in a model of Klimov. Technical Report RC-16543. Yorktown Heights (NY): IBM Research Division, T. J. Watson Research Center; 1991.
44. Bhattacharya PP, Georgiadis L, Tsoucas P. Extended polymatroids: properties and optimization. In: Balas E, Cornuejols G, Puleyblank WR, editors. Proceedings of the 2nd Conference Integer Programming and Combinatorial Optimization (IPCO II). Pittsburgh: Mathematical Programming Society, Carnegie Mellon University; 1992. pp. 298–315.
45. Bhattacharya PP, Georgiadis L, Tsoucas P. Problems of adaptive optimization in multiclass M/GI/1 queues with Bernoulli feedback. Math Oper Res 1995;20(2):355–380.
46. Bertsimas D, Niño-Mora J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. Math Oper Res 1996;21(2):257–306.
47. Meilijson I, Weiss G. Multiple feedback at a single-server station. Stoch Process Appl 1977;5(2):195–205.
48. Weiss G. Branching bandit processes. Probab Eng Inf Sci 1988;2:269–278.
49. Gittins JC. Multi-armed bandit allocation indices. Chichester: Wiley; 1989.
50. Bertsimas D, Paschalidis IC, Tsitsiklis JN. Branching bandits and Klimov's problem: achievable region and side constraints. IEEE Trans Automat Control 1995;40(12):2063–2075.
51. Yao DD, Zhang L. Dynamic scheduling of a class of stochastic systems: extended polymatroid, side constraints, and optimality. Proceedings 36th IEEE Conference Decision Control. New York: IEEE; 1997. pp. 1191–1196.
52. Yao DD, Zhang L. Stochastic scheduling via polymatroid optimization. In: Yin GG, Zhang Q, editors. Volume 33, Mathematics of stochastic manufacturing systems, Lecture Notes in Applied Mathematics. Providence (RI): AMS; 1997. pp. 333–364.
53. Chen H, Yao DD. Fundamentals of queueing networks: performance, asymptotics, and optimization. New York: Springer; 2001.
54. Levy H, Sidi M. Polling systems: applications, modelling and optimization. IEEE Trans Commun 1990;38(10):1750–1760.
55. Takagi H. Analysis of Polling systems. Cambridge (MA): MIT Press; 1986.
56. Yechiali U. Optimal dynamic control of polling systems. In: Cohen JW, Pack CD, editors. Proceedings 13th International Teletraffic Congress. Amsterdam, The

- Netherlands: North-Holland Publishing Co.; 1991. pp. 205–218.
57. Watson KS. Performance evaluation of cyclic service strategies: a survey. In: Gelenbe E, editor. *Performance '84*. Amsterdam: North-Holland Publishing Co.; 1984. pp. 521–533.
 58. Ferguson MJ, Aminetazh YJ. Exact results for nonsymmetric token ring systems. *IEEE Trans Commun* 1985;COM-33(3): 223–231.
 59. Fuhrmann SW. Symmetric queues served in cyclic order. *Oper Res Lett* 1985;4(3): 139–144.
 60. Boxma OJ. Models of two queues: a few new views. In: Boxma OJ, Cohen JW, Tijms HC, editors. *Teletraffic analysis and computer performance evaluation*. Amsterdam, The Netherlands: North-Holland Publishing Co.; 1986. pp. 75–98.
 61. Boxma OJ, Groenendijk WP. Pseudoconservation laws in cyclic-service systems. *J Appl Probab* 1987;24(4):949–964.
 62. Fuhrmann SW, Cooper RB. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Oper Res* 1985;33(5): 1117–1129.
 63. Doshi BT. Queueing systems with vacations—a survey. *Queueing Syst* 1986; 1(1):29–66.
 64. Boxma OJ, Groenendijk WP. Waiting times in discrete-time cyclic-service systems. *IEEE Trans Commun* 1988;36(2):164–170.
 65. Boxma OJ, Groenendijk WP, Weststrate JA. A pseudoconservation law for service systems with a polling table. *IEEE Trans Commun* 1990;38(10):1865–1870.
 66. Boxma OJ. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Syst* 1989;5(1-3):185–214.
 67. Sidi M, Levy H, Fuhrmann SW. A queueing network with a single cyclically roving server. *Queueing Syst* 1992;11(12):121–144.
 68. Fournier L, Rosberg Z. Expected waiting times in polling systems under priority disciplines. *Queueing Syst* 1991;9(4):419–439.
 69. Shimogawa S, Takahashi Y. A note on the pseudo-conservation law for a multi-queue with local priority. *Queueing Syst* 1992; 11(1-2):145–151.
 70. Katayama T. A note on conservation laws for a multi-class service queueing system with setup times. *Queueing Syst* 1992;11(3): 299–306.
 71. Bertsimas D, Niño-Mora J. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part I, the single-station case. *Math Oper Res* 1999;24(2):306–330.
 72. Rosberg Z. Process scheduling in a computer system. *IEEE Trans Comput* 1985;34(7): 633–645.
 73. Bertsimas D, Paschalidis I, Tsitsiklis J. Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann Appl Probab* 1994;4(1):43–75.
 74. Kumar S, Kumar PR. Performance bounds for queueing networks and scheduling policies. *IEEE Trans Automat Contr* 1994; 39(8):1600–1611.
 75. Bertsimas D, Niño-Mora J. Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part II, the multi-station case. *Math Oper Res* 1999;24(2): 331–361.
 76. Baccelli F, Brémaud P. *Elements of queueing theory: palm martingale calculus and stochastic recurrences*. 2nd ed. Berlin: Springer; 2003.
 77. Glazebrook KD, Niño-Mora J. Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance. *Oper Res* 2001;49(4):609–623.
 78. Dacre M, Glazebrook K, Niño-Mora J. The achievable region approach to the optimal control of stochastic systems. *J R Stat Soc Ser B Stat Methodol* 1999;61(4):747–791. With discussion.
 79. Green TC, Stidham S Jr. Sample-path conservation laws, with applications to scheduling queues and fluid systems. *Queueing Syst* 2000;36(1-3):175–199.
 80. Bäuerle N, Stidham S. Conservation laws for single-server fluid networks. *Queueing Syst* 2001;38(2):185–194.
 81. Lu Y, Yao DD. Optimal control of a fluid network with side constraints. *IEEE Trans Automat Contr* 2003;48(10): 1865–1870.
 82. Wong JW, Sauvé JP, Field JA. A study of fairness in packet-switching networks. *IEEE Trans Commun* 1982;COM-30(2): 346–353.
 83. Clare LP, Rubin I. Performance boundaries for prioritized multiplexing systems. *IEEE Trans Inf Theor* 1987;33(3):329–340.

84. Sumita S, Ozawa T. Achievability of performance objectives in ATM switching nodes. In: Hasegawa T, Takagi H, Takahashi Y, editors. *Proceedings of the International Seminar on Performance of Distributed and Parallel Systems*. Amsterdam: Elsevier (North-Holland); 1989. pp. 45–56.
85. Jeon YH, Viniotis I. Achievability of combined GOS requirements in broadband networks. *Proceedings IEEE International Conference Communications '93*. New York: IEEE; 1993. pp. 192–196.
86. Jeon YH, Viniotis I. Achievable loss probabilities and buffer allocation policies in ATM nodes with correlated arrivals. *Proceedings IEEE International Conference Communications '93*. New York: IEEE; 1993. pp. 365–369.
87. Yoo M, Qiao C, Dixit S. QoS performance of optical burst switching in IP-over-WDM networks. *IEEE J Sel Areas Commun* 2000;18(10):2062–2071.
88. Lui JCS, Wang XQ. An admission control algorithm for providing quality-of-service guarantee for individual connection in a video-on-demand system. 5th IEEE Symposium Computers and Communications (ISCC 2000). Los Alamitos (CA): IEEE; 2000. pp. 456–461.
89. Marbukh V. A framework for performance evaluation and optimization of an emerging multimedia DS-CDMA network. *Proceedings 4th ACM International Workshop Wireless Mobile Multimedia*. New York: ACM; 2001. pp. 55–64.
90. Lui JCS, Wang XQ. Providing QoS guarantee for individual video stream via stochastic admission control. In: Goto K, Hasegawa T, Takagi H, *et al.*, editors. *Performance and QoS of next generation networking*. London: Springer; 2001. pp. 263–279.
91. Gelenbe E, Srinivasan V, Seshadri S, *et al.* Optimal policies for ATM cell scheduling and rejection. *Telecommun Syst* 2001; 18(4):331–358.
92. Yang L, Jiang Y, Jiang S. A probabilistic preemptive scheme for providing service differentiation in OBS networks. *Proceedings IEEE GLOBECOM '03*. New York: IEEE; 2003. pp. 2689–2693.
93. Lu X, Mark BL. Performance modeling of optical-burst switching with fiber delay lines. *IEEE Trans Commun* 2004;52(12): 2175–2183.
94. Tan CW, Gurusamy M, Lui JCS. Achieving proportional loss differentiation using probabilistic preemptive burst segmentation in optical burst switching WDM networks. *Proceedings IEEE Globecom '04*. New York: IEEE; 2004. pp. 1754–1758.
95. Whittle P. Restless bandits: activity allocation in a changing world. In: Gani J, editor. Volume 25A, *A celebration of applied probability*, *Journal of Applied Probability*. Sheffield: Applied Probability Trust; 1988. pp. 287–298.
96. Weber RR, Weiss G. On an index policy for restless bandits. *J Appl Probab* 1990; 27(3):637–648.
97. Bertsimas D, Niño-Mora J. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Oper Res* 2000;48(1):80–90.
98. Niño-Mora J. Restless bandits, partial conservation laws and indexability. *Adv Appl Probab* 2001;33(1):76–98.
99. Niño-Mora J. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Math Program* 2002;93(3):361–413.
100. Niño-Mora J. Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock M/G/1 queues. *Math Oper Res* 2006;31(1):50–84.
101. Niño-Mora J. Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue. *Queueing Syst* 2006;54(4):281–312.
102. Niño-Mora J. Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging. Volume 4465, *Proceedings 1st Euro-FGI Conference Network Control and Optimization (NET-COOP 2007, Avignon, France, LNCS)*. Berlin: Springer; 2007. pp. 138–149.
103. Niño-Mora J. A faster index algorithm and a computational study for bandits with switching costs. *INFORMS J Comput* 2008;20(2):255–269.
104. Cao J, Nyberg C. Linear programming relaxations and marginal productivity index policies for the buffer sharing problem. *Queueing Syst* 2008;60(3–4):247–269.
105. Niño-Mora J. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP* 2007;15(2):161–198. followed by six discussions by Adan IJBF, Boxma OJ,

Altman E, Hernández-Lerma O, Weber R, Whittle P, Yao DD.

FURTHER READINGS

Baccelli P, Brémaud P. Elements of queueing theory: Palm martingale calculus and stochastic recurrences. 2nd ed. Berlin: Springer; 2003.

Chen H, Yao DD. Fundamentals of queueing networks: performance, asymptotics, and optimization. New York: Springer; 2001. Chapter 11.

Niño-Mora J. Stochastic scheduling. In: Floudas CA, Pardalos PM, editors. Encyclopedia of optimization. 2nd ed. New York: Springer; 2008. pp. 3818–3824.

Stidham S Jr. Analysis, design, and control of queueing systems. *Oper Res* 2002;50(1): 197–216.

Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice Hall; 1989.

CONSTRAINT PROGRAMMING LINKS WITH MATH PROGRAMMING

MICHELA MILANO
DEIS, University of Bologna,
Bologna, Italy

INTRODUCTION

Constraint programming (CP) [1–4] is a powerful paradigm for solving combinatorial decision and optimization problems. The first CP languages were based on logic programming. Constraint logic programming has been proposed as a general scheme [5] that can be instantiated on many different sorts, namely, real numbers, integer, pseudo-Boolean, sets, trees, and so on. The most successful instantiation of the scheme is the one on finite domain that has been used to solve combinatorial problems. In the rest of the article, we will focus on constraint programming on finite domains (CP(FD)), also referred to as *constraint programming*.

In CP, problems are modeled using variables ranging on a finite domain of integers and constraints over subsets of these variables. The CP solving methodology interleaves constraint propagation and search. Constraint propagation is an inference mechanism based on domain filtering. Each constraint embeds a filtering algorithm that removes values that cannot appear in any consistent solution. Since variables are involved in several constraints, domain updates are propagated to the other constraints whose filtering algorithms are triggered and possibly remove other domain values in an iterative manner. Unfortunately, removing all inconsistent values has the same complexity of solving the original problem; for this reason, propagation may be incomplete: all removed values are certainly inconsistent, but not all values left are guaranteed to be consistent. Therefore, search is needed to explore the remaining part of the search tree. At each node, propagation

is again triggered until either a solution is found or a failure is reached.

The solution methods of CP and mathematical programming (MP) have complementary strengths. Recent research shows that these strengths can be profitably combined in hybrid algorithms. On one hand CP is very effective for the so called *feasibility reasoning*. Constraint propagation enables to avoid searching in infeasible regions of the search space. In addition, being constraint propagation a mechanism that is based on constraint interaction, adding side constraints to a model in general improves the performance of a constraint solver. On the other hand, MP is effective for the so called *optimality reasoning*. The use of relaxation and the definition of cutting planes enable to avoid searching in suboptimal regions of the search space. In addition, many structured problems have been deeply analyzed in operations research (OR) and specific cutting planes have been devised to tighten the problem relaxation.

Therefore, CP and MP have complementary strength and can be conveniently merged to exploit the advantages of both [6–9]. In this survey we adopt a CP perspective and we focus mainly on the CP literature. Therefore, MP techniques assist the CP solving process, for example, filtering and bounding. However, there is a different perspective: recently some notable contributions from the OR literature propose that CP can contribute in a MP algorithm, as happens in SCIP (solving constraint integer programs) [10] or that CP and MP can contribute to an equal basis to the solving process as happens in SIMPL [11], but these approaches are out of the scope of this article.

From a CP perspective, one very effective technique for combining CP and MP is to use a CP solver as the master solver and integrate MP algorithms within CP global constraints. Global constraints are very powerful constructs that compactly represent a set of primitive constraints. Global constraints are not only syntactic sugar, but also embed

filtering algorithms that reason globally and enable powerful filtering. There are several ways to integrate filtering algorithms coming from MP into global constraints: either using graph-based algorithms or dynamic programming for achieving arc-consistency, or embedding relaxations into the constraint or providing a linearization of the constraint that can be solved by a linear programming solver as happens in MP Branch and Bound.

Another way to combine the strength of CP and MP is to decompose the problem and to use the most effective solver on each component. Two notable examples of this integration are logic based Benders decomposition and CP based column generation. In both cases, depending on the specific structure of the component derived from decomposition, the most suitable solver is used. For example, side constraints are more effectively faced with CP tools, while for structured problems the most efficient technique is MP.

Search strategies have been widely studied in CP. Information coming from relaxation have been used for improving the search by driving it toward promising regions in terms of costs.

Nowadays, many commercial and academic tools have been developed for designing hybrid algorithms. We will provide in this article, a short description of the ones developed in the CP community.

CONSTRAINT PROGRAMMING

CP on finite domains has been recognized as a suitable modeling and solving tool to solve combinatorial problems. In this section, we provide some basic notions on modeling aspects, constraint propagation, and search and optimization.

The CP modeling and solving activity is highly influenced by the Artificial Intelligence area on Constraint Satisfaction Problems, CSPs (see the book by Tsang [12]). A CSP is a triple $\langle V, D, C \rangle$ where V is a set of variables $X_1, \dots, X_n$, D is a set of finite domains $D(X_1), \dots, D(X_n)$ representing the possible values that variables can assume, and C is a set of constraints $C_1, \dots, C_k$. Each constraint involves a set of variables $V' \subseteq$

V and defines a subset of the Cartesian products of the corresponding domains containing feasible value tuples. Therefore, constraints limit the values that variables can simultaneously assume.

A solution to a CSP is an assignment of values to variables which is consistent with constraints. If we are interested in the optimal solution instead of a feasible one, an objective function can be imposed on problem variables.

There exists a one to one mapping between CSP concepts and CP on finite domains syntactic structures. Thus, CP benefits from and extends results achieved for CSPs.

Modeling a Problem

A CP model contains variables ranging on a finite domain of values and a set of constraints involving those variables. Values can be objects of arbitrary type, but in general these languages manage finite domains of integers. Given a variable X its domain $D(X)$ denotes values that variables can assume. For example, in a scheduling application, if *Start* is a variable representing the starting time of an activity, its domain can be the schedule horizon, for example, $Start :: [1..100]$. This (unary) constraint states that variable *Start* can assume one of the integer values between 1 and 100. For another example, if a given variable *Pos* represents the position in a sequence of a given object, and positions available for the object are only 3, 5, and 9, we can express this constraint with a domain variable $Pos :: [3, 5, 9]$.

A constraint C on a set of variables $\{X_1, \dots, X_n\}$ is defined as a subset of the Cartesian product of their domains, that is, $C \subseteq D(X_1) \times D(X_2) \times \dots \times D(X_k)$. A tuple $(d_1, d_2, \dots, d_k) \in C$ is said to be a solution of C . A value $d \in D(X_i)$ is said to be inconsistent with respect to (w.r.t.) C if it does not belong to any solution of C .

Constraints can be either mathematical or symbolic constraints. Mathematical constraints are the usual relations among integer variables (i.e., $=$, $\neq$, $\leq$, $<$, $\geq$, $>$). An example of mathematical constraint is the following: if two activities i and j characterized by starting times S_i and S_j and durations D_i

and D_j are linked by a precedence constraint stating that activity i should be executed before activity j , the following constraint can be imposed, $S_i + D_i \leq S_j$. Symbolic constraints, called also *global constraints*, are more expressive and powerful constraints embedding constraint-dependent filtering algorithms. A typical global constraint is the *alldifferent* ($[X_1, \dots, X_n]$) [13], available in most CP solvers like CHIP [14], ECLⁱPS^e [15], and ILOG Solver [16]. Declaratively, the constraint *alldifferent* ($[X_1, \dots, X_n]$), holds if and only if all variables are assigned to a different value. Thus, it is declaratively equivalent to a set of $n * (n - 1)/2$ binary constraints. However, it is more compact allowing more concise models and it embeds a specialized filtering algorithm, which we will discuss later in the article.

An extension of the *alldifferent* is the *global cardinality constraint* [17]

$$gcc([X_1, \dots, X_n], [v_1, \dots, v_m], \\ [l_1, \dots, l_m], [u_1, \dots, u_m]),$$

which holds if and only if the number of variables in $[X_1, \dots, X_n]$ which assume value v_i is within l_i and u_i .

The objective function in a CP program is represented by a domain variable. For example, if we have to minimize the makespan in a scheduling problem (i.e., its total duration), we have to minimize the ending time of the last activity. Therefore, we have to minimize $Z = \max_i \{S_i + D_i\}$. As another example, if we have a cost associated to each variable–value assignment and we have to minimize the overall cost, in the model we introduce a domain variable C_i representing the cost of the assignment of each problem variable. The objective function is $Z = \sum_i C_i$.

Constraint Programming Solving

The CP solution process interleaves constraint propagation and tree search. The search process enumerates all possible variable–value assignments, until we find a solution or we prove that none exists. To reduce the exponential number of variable–value pairs in the search tree,

domain filtering and constraint propagation are applied at each node of the search tree. Domain filtering operates on individual constraints and removes provably inconsistent domain values. Since variables are involved in several constraints, domain updates are propagated to the other constraints whose filtering algorithms are triggered and possibly remove other domain values.

Filtering algorithms for being effective should be efficient and incremental as they are applied at each node of the search tree. Also they should try to remove as many inconsistent values as possible. If they remove all inconsistent values w.r.t. a constraint C , C is made *domain consistent*. Domain consistency is also called *hyper-arc consistency* or *generalized arc consistency*. However, checking all variable–value pairs for consistency could be extremely computationally expensive.

For this reason, symbolic constraints in general embed propagation algorithms which exploit the semantics of the constraint itself [13,18,19]. Consider, for example, the symbolic constraint *alldifferent* ($[X_1, \dots, X_n]$). It declaratively holds if all variables are assigned to a different value and it is equivalent to a set of binary inequality constraints connecting each pair of values in the list $[X_1, \dots, X_n]$. However, we can perform more global and informed reasoning on the set of variables. In Ref. 13, a filtering algorithm based on graph matching has been defined achieving hyper-arc consistency. It builds a bipartite graph whose nodes are divided into two sets: variable nodes and value nodes. There is an arc between a variable X and a value node v if $v \in D(X)$. The algorithm removes values that do not belong to any feasible matching that covers all variable nodes. We here provide an intuitive example on how the algorithm prunes the search space by reasoning globally. Suppose for example that we have an *alldifferent* constraint among three variables $[X_1, X_2, X_3]$ whose domains are $D(X_1) = D(X_2) = [1, 2]$ and $D(X_3) = [1 \dots 10]$. While a set of binary inequality constraints cannot infer any value removal, the *alldifferent* constraint can reason globally on the cardinality of the sets of variables and values. We have two variables, that is, X_1 and X_2 , whose common domains

$D(X_1) = D(X_2) = [1, 2]$ contain exactly two values. Thus, values 1 and 2 are *reserved* for variables X_1 and X_2 (no matter which value is assigned to which variable), and are no longer feasible for variable X_3 whose domain is reduced to $[3...10]$.

At the end of the constraint propagation process, we have three possible scenarios: (i) a domain becomes empty and a failure occurs; (ii) a solution is found, that is, all variables are assigned to one value; (iii) some domains contain more than one value. In this third case, since constraint propagation is not complete, we need a search strategy in order to explore the remaining search tree.

The way the search space is explored greatly influences the performances of the overall constraint solving process. An important point to be clarified is that during the search, constraints are always taken into account and propagated in order to prune, as much as possible, the search space. In fact, a variable instantiation triggers all the filtering algorithms of the constraints involving that variable and a propagation process starts again. All CP languages have high level predicates enabling the definition of search heuristics (branching methods). This leads to the development, within the CP community of sophisticated branching methods for many type of problems allowing to solve them effectively.

Optimization

In some applications we are not looking for a feasible solution, but for an optimal one w.r.t. some objective function f defined on problem variables. With no loss of generality, we restrict our discussion to minimization problems. CP systems usually implement a naive form of Branch and Bound algorithm to find an optimal solution. The idea is to solve a set of decision (feasibility) problems (i.e., a feasible solution is found if it exists), leading to successively better solutions. In particular, each time a feasible solution z^* is found (whose associated cost is $f(z^*)$), a constraint $f(x) < f(z^*)$, where x is any feasible solution, is added to each subproblem in the remaining search tree. The purpose of the added constraint, called *upper bounding constraint*, is to remove portions of the search

space which cannot lead to better solutions than the best one found so far. The problem with this approach is twofold: (i) CP does not rely on sophisticated algorithms for computing lower and upper bounds for the objective function, but derives their values starting from the variable domains; (ii) in general, the link between the objective function and the problem decision variables is quite poor and does not produce effective domain filtering.

Therefore, recently some efforts have been performed in order to embed in CP some techniques that take into account some form of *optimality reasoning*. In the rest of this article, we will focus on the description of the integration of MP techniques into CP.

INTEGRATION THROUGH GLOBAL CONSTRAINTS

One of the most important aspects of CP languages is the extensive use of global constraints. Global constraints, also called *symbolic constraints*, are n -ary constraints that represent suitable abstractions embedding sophisticated, semantic-aware filtering algorithms.

Global constraints are a very effective mean for integration [20]. The reason is twofold: first, global constraints represent structured subproblems for which efficient algorithms can be devised; second, the reasoning performed by global constraints is local; therefore, they can be developed as independent components interacting through propagation.

MP-Based Filtering

Some constraints represent a polynomial problem. In this case, hyper-arc consistency can be achieved in polynomial time with a special purpose algorithm. For example, one of the most widely used and successful global constraints is the *alldifferent* constraint: it is defined on a set of variables $[X_1, \dots, X_n]$ ranging on domains $[D(X_1), \dots, D(X_n)]$ and it holds if and only if all variables are assigned to different values. The linear model of the constraint can be obtained via a mapping between CP and integer programming variables, first suggested by Rodosek *et al.*

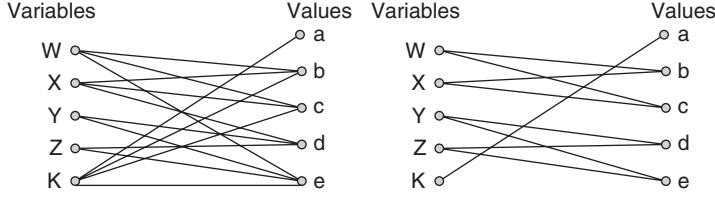


Figure 1. The *alldifferent* constraint filtering.

[21]. We have a set of decision variables x_{ij} that take the value 1 if the CP variable X_i is assigned to value j , 0 otherwise. Since each CP variable can be assigned to a single value in its domain, we have the constraints

$$\sum_j x_{ij} \leq 1 \quad \forall i.$$

Since all variables should assume different values, we have

$$\sum_i x_{ij} = 1 \quad \forall j.$$

The integrality constraint is part of this model. However, the coefficient matrix is totally unimodular meaning that linear programming provides an integer solution. Clearly, the addition of side constraints would break this structure. However, the *alldifferent* constraint is treated locally in CP, thus exploiting the unimodularity property. For this reason the *alldifferent* constraint embeds a polynomial-time filtering algorithm achieving domain consistency.

Many global constraints exploit graph theory results and algorithms for pruning inconsistent values. The filtering algorithm for the *alldifferent* constraint achieves hyper-arc consistency [13] by applying a maximum matching algorithm on a bipartite graph $G = (X \cup V, E)$, called *value graph* [22] built as follows: X is a set of nodes representing variables involved in the constraint, V is a set of nodes representing values contained in the union of the domains, and E represents the set of edges, $(X_i, v) \in E$ if and only if $v \in D(X_i)$. Hyper-arc consistency of an *alldifferent* constraint is established by computing a maximum matching M which covers X and identifying all edges that belong to a maximum matching. Consider for example five variables with the corresponding domain $W ::$

$[b, c, e]$, $X :: [b, c, d]$, $Y :: [e, d]$, $Z :: [e, d]$, and $K :: [a, b, c, e]$. The value graph is depicted in the left side of Fig. 1. The filtering algorithm removes all values that do not belong to a matching covering the set of variables leading to the deletion represented in the right part of Fig. 1. The complexity of computing a maximum matching (or prove there is none) is $O(m\sqrt{n})$ where m is the number of edges and n the number of variables. The complexity of finding all edges belonging to a maximum matching and thus also achieving hyper-arc consistency can be done in $O(m)$ [23].

Another constraint exploiting graph theory algorithms for achieving hyper-arc consistency is a generalization of the *alldifferent* constraint: the global cardinality constraint, *gcc*. The *gcc* [17] limits the cardinality of each value assumed by a set of variables.

Consider the example taken from Ref. 17 where we have seven variables representing employees (peter, paul, mary, john, bob, mike, and julia) and five values representing shifts. We know that for the shifts M and D we need at least one and at most two people, for shift N exactly one person is required and for shifts B and O we need at least zero and at most two people. The syntax of the constraint is the following:

$gcc([peter, paul, mary, john, bob, mike, julia],$
 $[M, D, N, B, O], [1, 1, 1, 0, 0], [2, 2, 1, 2, 2]).$

The value graph of this constraint is depicted in the leftmost part of Fig. 2. Intuitively, the filtering algorithm reasons as follows. Variables peter, paul, mary, and john have only two domain values in their domains, namely, M and D. These four variables should be assigned, no matter how, to these values. Since the sum of the upper bound values for these two values is 4, they are not available for any other variable.

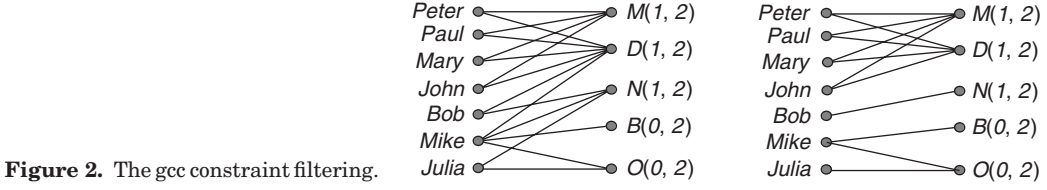


Figure 2. The gcc constraint filtering.

Therefore, they are deleted from the domain of bob and mike. After this deletion bob has a single value in its domain, namely, N whose upper bound is 1. Therefore, it is no longer available for variables mike and julia. The filtered domains are depicted on the rightmost side of Fig. 2.

Again, graph theory is used for achieving hyper-arc consistency and for pruning all infeasible values. Let us start with the value graph (bipartite graph) presented for the *alldifferent* constraint. We have two sets of nodes: variable nodes and value nodes. Here we add a source node s and a sink node t . In addition arcs are associated with a demand and a capacity and are divided in three sets:

- those starting from s and ending in a value node. Their demand is $l(v)$ and their capacity is $u(v)$;
- those connecting a value and a variable node (that are the same for the *alldifferent* constraint). Their demand is 0 and their capacity is 1;
- those starting from a variable node and ending in t . Their demand is 0 and their capacity is 1.

The graph for the above example is depicted in Fig. 3.

A solution of *gcc* corresponds to a maximum flow of value n from s to t in the above

mentioned graph, where n is the number of variables involved. An arc connecting variable x_i with value v can be part of a solution if it carries positive flow in the maximum flow solution, or it carries zero flow but has a reduced cost of zero. Reduced costs can be rapidly computed for identifying strongly connected components of the residual graph. The maximum flow can be found in $O(nm)$, where m is the number of edges in the graph. Again hyper-arc consistency can be achieved in $O(m + n)$. In addition, flow algorithms are incremental and each time a value is removed from a domain of a variable, the filtering algorithm should not start from scratch. Another more complex example of filtering algorithm based on flow for the sequence constraint can be found in Ref. 24.

Beside graph theory, dynamic programming has also been embedded into global constraints. It has been used for the knapsack constraint, or subset-sum constraint [25].

Beldiceanu's *global constraints catalogue* [26] lists some 350 constraints which are explicitly described in terms of graph properties and/or automata.

MP-Based Filtering on Costs. Graph-based algorithms can be also used to implement the so called *cost-based filtering*. Let us consider again an *alldifferent* constraint but now variables have an associated cost. The resulting constraint is called *minweightalldiff*

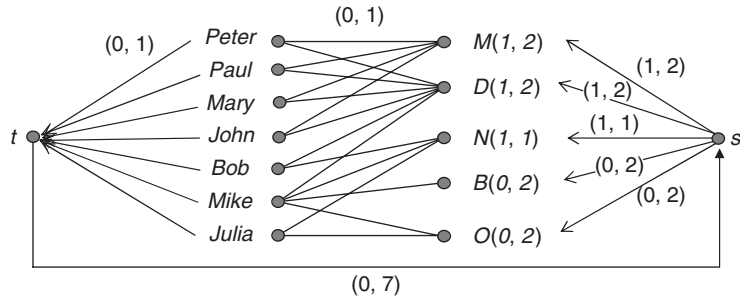


Figure 3. The gcc graph for computing the maximum flow.

[27]. Let $X_1, \dots, X_n$, be the problem decision variables and be Z the cost variable. Let $D(X_1), \dots, D(X_n)$ be the domains associated to variables X and $D(Z)$ the domain of Z . We have either to minimize the cost variable during the solution process or we have an upper bound on it (corresponding to the domain upper bound $D_{\max}(Z)$). Now let w_{ij} for $i = 1, \dots, n$ and all $j \in \bigcup D(X_i)$ be constants representing costs/weights. Then *minweightalldiff*($X_1, \dots, X_n, Z, w$) holds if and only if $X_i \neq X_k$ for all $i, k \in 1, \dots, n$ and $\sum_{i, X_i=j} w_{ij} \leq Z$.

This constraint has a linear formulation equal to the *alldifferent* with in addition the following objective function:

$$\min \sum_i \sum_j x_{ij} w_{ij}.$$

The coefficient matrix is again totally unimodular. Therefore, the problem is polynomial. For this reason achieving hyper-arc consistency is doable and the algorithm is again based on graph theory, and specifically on the minimum cost flow. The graph has the same structure as the one built for the *gcc*. We have variable nodes, value nodes, a source s , and a sink t . Each arc has an associated capacity equal to 1 and a cost. The cost is 0 for those arcs that start from the source and end in variable nodes and for those arcs that start from value nodes and end in the sink. The cost is instead c_{ij} on those arcs that start from variable nodes and end in value nodes. The constraint is consistent if

- a feasible flow f from s to t exists, its value is n corresponding to the number of variables and $\text{cost}(f) \leq D_{\max}(Z)$; and
- the minimum cost $s - t$ flow f has a value n and $\text{cost}(f) \geq D_{\min}(Z)$.

In the same way, also for the global cardinality constraints with costs a polynomial-time algorithm based on minimum cost flow [19] has been proposed for achieving hyper-arc consistency.

Similarly to constraints reasoning on optimization, recently soft-global constraints have been developed exploiting the same techniques, but considering the cost of the

violation to be minimized, see Ref. 28 for a survey.

Non-polynomial Structures

Hyper-arc consistency can be achieved in polynomial time only if the constraint represents a polynomial problem. In some cases, global constraints represent NP-hard problems. Therefore, the constraint embeds a filtering algorithm that enforces a weaker notion of consistency w.r.t. hyper-arc consistency.

As an example, let us consider the cumulative constraint widely used in scheduling problems whose parameters are: a list of variables $[S_1, \dots, S_n]$ representing the starting time of all activities sharing the resource, their duration $[D_1, \dots, D_n]$, the resource consumption for each activity $[R_1, \dots, R_n]$, and the available resource capacity C

$$\text{cumulative}([S_1, \dots, S_n], [D_1, \dots, D_n], [R_1, \dots, R_n], C).$$

Many filtering algorithms have been implemented for this constraint, none achieving hyper-arc consistency of course [29]. A successful algorithm, introduced by the OR community is the edge finder [30], and embedded as filtering algorithm in CP global constraints by Baptiste *et al.* [31].

An example is depicted in Fig. 4 where three activities S1, S2, and S3 must be scheduled on a single capacity machine. Their duration is respectively, 6, 4, and 3, their earliest start times are respectively, 0, 1, and 1 and their latest completion times are respectively, 17, 11, and 12. If activity S1 starts as soon as possible, before S2 and S3, its earliest completion time is 6. The sum of the durations of S2 and S3 is 7 while the difference between the maximum latest completion time of S2 and S3 and the earliest completion time of S1 is 6. Therefore, we can conclude that both S2 and S3 should precede S1, whose earliest start time becomes 8.

The edge finder has been extended to cope with non unary resources. Also variants of the cumulative constraint for preemptive activities, for unary and cumulative resources are presented in Ref. 29.

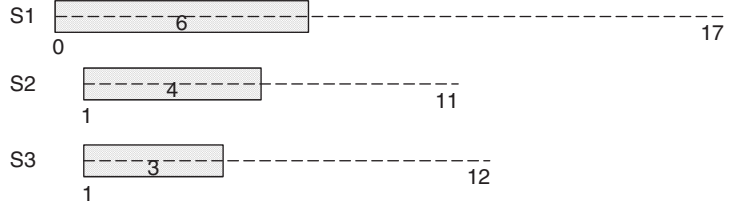


Figure 4. Edge finder.

Constraint-Specific Relaxations

When the constraint represents an NP-hard problem, variables have an associated cost and Z is the cost variable on which we have an upper bound, only an approximation of hyper-arc consistency can be achieved. Therefore, a possibility is to embed in the constraint a relaxation of the constraint itself as proposed in Ref. 32.

Whatever relaxation is used, the unique requirement concerns its output. The relaxation should return three results as follows:

- the optimal solution of the relaxed problem x^* ;
- the optimal solution objective function value LB . This value represents a lower bound the cost variable Z ;
- a gradient function $grad(X_i, j)$ measuring the variable–value assignment cost.

These pieces of information are exploited by the CP solver not only for filtering purposes, but also for guiding the search toward promising branches (in terms of costs), see the section titled “Search and Optimization.”

An example of the use of these pieces of information is the following: we use the traditional mapping between the CP variables involved in the constraints and those involved in the relaxation. We have a binary variable x_{ij} corresponding to each variable–values pair in CP. In particular, $x_{ij} = 1$ if $X_i = j$ in CP, while $x_{ij} = 0$ if $X_i \neq j$ in CP. Constraints $\sum_{j \in D(X_i)} x_{ij} = 1 \forall i$ are part of the mapping and impose that exactly one value should be assigned to each CP variable.

In addition, if a variable–value assignment in CP has a cost, then the obvious link between the cost variable of the CP model Z and the lower bound value LB computed by solving the relaxation is $LB \leq Z$. If LB is greater or equal than the upper bound of the

domain of Z , a failure occurs since the domain of Z becomes empty.

Given this mapping, the variable x_{ij} in the relaxation corresponds to the value j in the domain of the CP variable X_i . Thus, reduced costs $\bar{c}_{ij}$ provide an optimistic evaluation on the cost of CP variable domain values, $grad(X_i, j) = \bar{c}_{ij}$. More formally, for each domain value j of each variable X_i , we can compute a lower bound on the subproblem generated if value j is assigned to X_i as

$$LB_{X_i=j} = LB + grad(X_i, j).$$

If $LB_{X_i=j}$ is greater or equal to the upper bound of the domain of Z , j can be deleted from the domain of X_i .

This filtering algorithm performs a real back-propagation from Z to X_i . Such a technique is known in MP as *variable fixing*. However, variable fixing in the MP context does not trigger in general an iterative process. On the contrary, in CP, filtering variable domains triggers constraints imposed on the modified variables.

An important point which should be stressed is that this filtering algorithm is general and can be applied whenever the relaxation is able to provide these pieces of information. The filtering algorithm is independent on the structure of the relaxation.

SEARCH AND OPTIMIZATION

CP search is similar to MP branch and bound. In fact, it explores a search tree whose leaves are either solutions or inconsistent nodes. In a MP search tree, each node has at most two descendants, but in a CP search tree a node may have many descendants. This is a key difference between CP and MP search.

Second, in a MP search tree a child of a node differs from its parent just in one

(tighter) bound on one variable. In a CP search tree the child typically has one variable instantiated to a fixed value. This form of search is called *labeling*. Nevertheless, other forms of CP search may be used instead of labeling, as any arbitrary constraint can be imposed at each branch. For example, in a scheduling application the added constraint may enforce that one task cannot be scheduled until another has finished.

In general a search tree in CP is explored in depth first, with some notable variants. Credit search [33] is an incomplete method where the number of nondeterministic choices is limited a priori. Another widely used search strategy is the limited discrepancy search (LDS). LDS was first introduced by Harvey and Ginsberg [34]. The idea is that if the heuristic is accurate, it will make the wrong decision only few times. Note that the term *heuristic* has a different meaning w.r.t. the one used in OR. Technically, a heuristic is a function that takes a state as input and outputs a value for that state, often as a guess of how far away that state is from the solution. Thus, at each node of the search tree, the heuristic is supposed to provide the *good* choice (corresponding to the leftmost branch) among possible alternative branches. Any other choice would be *bad* and is called a *discrepancy*. In LDS, one tries to find first the solution with as few discrepancies as possible. LDS builds a search tree in the following way: the first solution explored is that suggested by the heuristic. Then solutions that follow the heuristic for every variable but one are explored: these solutions are that of discrepancy equal to one. Then, solutions at discrepancy equal to two are explored, and so on.

A variant of LDS for dealing with objective functions proposed in the OR community is local branching [35]. The idea is to explore a given discrepancy around a reference solution (neighborhood) and change reference solution if case a better one is encountered in the neighborhood. Discrepancy constraints, also called *local branching constraints* are linear inequalities used to model the neighborhood and to improve its bound. Local branching has been implemented and

properly extended to be smoothly integrated into the CP machinery in Ref. 36.

Branching Heuristics

During search, two fundamental choices highly influence the efficiency of the solution process in CP: (i) the selection of the variable to instantiate, that is, the *variable selection heuristics* and (ii) the selection of the value to assign, that is, the *value selection heuristics*.

These heuristics can be either problem independent or problem specific. An example of problem independent variable selection heuristics is first fail, assigning first the variables that has the smallest domain, as it is the most difficult variable to instantiate. Another recent and successful general purpose search strategy for CP is based on the concept of the impact of a variable [37]. The impact measures the importance of a variable for the reduction of the search space. Impacts are learnt during search. If combined with restart, impact based search outperforms almost all general purpose heuristics.

On the other hand problem dependent heuristics can be designed. An example is the following: in scheduling problems if we have to minimize the makespan a good choice is to select first the task with the smallest earliest start time.

Although we have distinguished variable and value ordering heuristics, there is an important class of heuristics that address variable and value orderings simultaneously. These are relaxation heuristics, which are based on the solution of a relaxed problem at each node of the search tree. If the (optimal) solution to the subproblem is also a feasible solution for the original problem, we have reached a leaf node of the search tree. If, on the other hand, some of the original problem constraints are violated, we choose a variable assignment that conflicts with the constraint. We then assign a different value to the variable in order to satisfy the constraint. A notable example of relaxation based heuristic is the one performed by MP branch and bound. Moreover the constraints omitted from the relaxed problem can be used actively to guide the search heuristics. In CP an example of search heuristic to handle the

relaxed constraint violation is called *unimodular probing* [38].

Using (Linear) Relaxations during Search

The use of MP techniques in CP has shown to be very effective in practice not only when the relaxation is embedded within a CP global constraint, but also when a linear solver is used as a global constraint and triggered at each node of the CP search tree as happens in MP branch and bound. These techniques were proposed first in Ref. 39 and extended by Refalo [40].

For this purpose the problem model should be stated in the CP solver and also passed to the linear programming solver via a mapping. For being effective the interaction between the two solvers should last during the overall computation. Therefore, a communication link between the CP constraints and the linear solver should be established. For this purpose, quite often the linear solver is considered as a software component similar to other constraints, interacting with them and exchanging information. Three pieces of information should be passed back and forth: variable fixing, bound reduction, and the optimal solution of the relaxation. Each time the CP propagation infers a new bound, it should be passed as a linear constraint to the linear solver which takes it into account in next recomputations. Also, variable fixing from the LP solver are passed to the CP solver as traditional propagations. The solution of the relaxation, that is, the lower bound can be used as described in the section titled “Constraint-Specific Relaxations.”

Clearly this procedure is more effective if the bound is tight. Therefore, [40] proposed to tighten the relaxation with cutting planes added during search and coming from CP global constraints. Cutting planes can be derived from global constraints representing subproblems and added to the linear programming model so as to tighten the overall relaxation. It is interesting to note that cutting planes can be derived after domain reduction from the CP side, so as to maintain the tightness of the relaxation during the whole solution process. Refalo [40] provided many examples and shown that it is very effective in practice.

The two approaches, CP and LP, not only complement each other, but also support each other. The linear relaxation yields a bound on the cost variable, which enables additional propagation to occur. Propagation, conversely, tightens variable bounds which in turn tighten the linear relaxation. This again yields a tighter bound on the cost variable which enables more propagation and so on.

Other relaxations have been integrated into a CP solver. Lagrangian Relaxation is a very important example [41–43].

In addition, the CP literature reports some other attempts of dealing with the objective function during search. One possibility is to exploit cost based information coming from global constraints and use the solution of the relaxation to guide search, as shown in the section titled “Constraint-Specific Relaxations.” Examples of research efforts in this direction are Refs 32 and 44. Also, additive bounding has been used for tightening the bound and taking into account the structure of the search tree into the constraints [36,45].

DECOMPOSITION

An alternative way to explore the search space is to decompose the problem at hand into multiple subproblems and use the solver which is most suited for each part. Clearly the solvers should interact and exchange information. Two notable examples of decomposition are the logic-based Benders decomposition and the CP based column generation.

Logic-Based Benders Decomposition

Benders decomposition [46] has been studied in the 1960s and is an effective method for solving a variety of structured problems. It is particularly suited for those problems where fixing a number of *hard* variables makes the problem simpler. Instead of blindly trying tentative values for the *hard* variables we can solve a master problem (which takes into account constraints on *hard* variables). After fixing *hard* variables, a subproblem can be solved, taking into account the remaining variables. The process iterates and converges to the optimal solution. The iteration between the master and the subproblem is regulated

by the so called *Benders cuts*. In OR, the subproblem should be a linear program. Hooker and Ottosson [47] propose to remove this restriction, defining the logic-based Benders decomposition framework. In this setting, the subproblem can be expressed as a constraint satisfaction problem.

An interesting and successful application of logic-based Benders decomposition is scheduling with alternative resources, where each activity can run on a set of parallel machines of different speeds and costs. As an example, taken from Ref. 47, let us consider a scheduling problem where we want to minimize the cost of the schedule. Allocation is modeled and solved with integer linear programming, while the scheduling part is solved through CP. In this case, a set of single machine scheduling problems is solved while in other articles, for example, in Benini *et al.* [48], the scheduling problem is considered overall. Also, other examples with different objective functions are presented in Refs 49–51.

Column Generation

Column generation is a very different technique for improving on non-optimal solutions. It is a very useful approach for problems for which an integer-linear model would involve too many variables—in some cases the number of variables grows exponentially with the size of the problem. The purpose is to enable an optimal solution to be found, and proven optimal, without ever considering more than a small proportion of these variables—certainly only a number that grows polynomially with the problem size.

The CP environment can encapsulate column generation so that it is possible just to plug in the master problem constraints and the subproblem algorithm.

The functionality is similar to that offered by column generation libraries and packages such as Abacus [52] and BC-Opt [53].

Column generation in combination with CP was first introduced in Ref. 54 and has been used for a range of applications including time-tabling, crew scheduling, vehicle routing [55–58].

All these optimization problems may be decomposed in a natural way: they may be viewed as selecting a subset of individual patterns within a huge pool of possible and weighted patterns. The selected combination is the one with the lowest cost to fulfill some given global requirements. In the CP-based column generation framework the master subproblem is solved by traditional MP techniques, while the pricing subproblem is solved by CP. The main benefits of using CP are the expressiveness of its modeling language and the flexibility of its solvers for dealing with complex side constraints.

TOOLS ENABLING INTEGRATION

Many tools have been developed in the CP community that enable or support the integration of CP with integer programming techniques. Clearly, for space limitations it is not possible to be exhaustive in this article, but we list the main tools and their main features. For a survey, see Ref. 59.

ECLⁱPS^e [15] is a Prolog-based system whose aim is to serve as a platform for integrating Logic Programming extensions, the most important being constraint logic programming. ECLⁱPS^e enables to use different constraint solvers in combination: for instance they can share variables and/or constraints. The *eplex* library gives access to linear programming and mixed integer programming solvers, while the *ic* library implements the finite domain constraint solver. Column generation, Benders decomposition, and local search can be implemented in ECLⁱPS^e.

G12 [60] is a solver platform for solving large scale combinatorial problems. Three languages represent the core design of the platform: Zinc, Cadmium, and Mercury. Zinc is a modeling language that is independent from the underlying solver. Cadmium enables the mapping between Zinc models and the underlying solvers. Finally Mercury is a language for building extensible and hybridizable solvers.

OPL Development Studio is designed to support ILOG Cplex and ILOG CP Optimizer. OPL (Optimization Programming Language)

[61] allows to develop single models in either technology. In addition, OPL develops models that use either or both technologies. In addition CP Optimizer provides a new scheduling language together with an automatic, robust and efficient solution strategy, easily accessible to mathematic programmers.

On the other way round, in the Integer Programming community, a number of solvers have integrated CP concepts, like BARON [62], SIMPL [11], and SCIP [63].

CONCLUSION

The emerging research field of the integration of OR techniques in CP is extremely promising and motivating since there are many open issues and challenges to be addressed and studied. The first challenge concerns the user support for the problem solving process. The aim is to provide the user with an abstract conceptual model and let the solver choose the best algorithm to solve it. This is too simple and unrealistic of course, but as a first step, research in this direction should be done in the identification of a mapping between problem structures and a corresponding efficient algorithm.

A second challenge concerns the solution of problems whose data are partially unknown or ill-defined. The problem solving process here should focus not only on searching for the optimal solution, which is often meaningless, but also on its robustness to the unknown parameter changing and to external events. If the problem is defined on probabilistic data, stochastic optimization should be taken into account. For this purpose, many approaches have been studied in the field of OR. The CP community has recently faced stochastic problems: in Ref. 64 stochastic CP is introduced formally and the concept of solution is replaced with the one of *policy*. This work has been extended in Ref. 65 where an algorithm based on the concept of scenarios is proposed.

A third challenge concerns over-constrained problems, that is, those problems that have no solution. These problems are quite common in an industrial setting where solutions represent compromises and violate

some problem constraints. At the state of the art some approaches have been devised to face these problems: (i) rank constraints into classes of importance, (ii) add a penalty to constraints, and (iii) count the violation.

A fourth challenge concerns the improvement of each aspect of the CP solving process: propagation could be improved to cope with constraints representing NP-hard problems, while tree search is more and more often combined with local search and metaheuristics.

REFERENCES

1. Apt KR. Principles of constraint programming. Cambridge, UK: Cambridge University Press; 2003.
2. Bockmayr A, Hooker JN. Constraint programming. In: Aardal K, Nemhauser G, Weismantel R, editors. Handbook of discrete optimization. Amsterdam: Elsevier; 2005. pp. 559–600.
3. Dechter R. Constraint processing. San Francisco (CA): Morgan Kaufmann; 2003.
4. Marriott K, Stuckey PJ. Programming with constraints: an introduction. Amsterdam, The Netherlands: MIT Press; 1998.
5. Jaffar J, Maher M. Constraint logic programming: a survey. *J Logic Program* 1994;19 & 20:503–581.
6. Hooker J. Operations research methods in constraint programming. In: Rossi F, van Beek P, Walsh T, editors. Handbook of constraint programming. Amsterdam: Elsevier; 2006. pp. 527–570.
7. Hooker J. Integrated methods for optimization. New York: Springer; 2007.
8. Milano M. Constraint and integer programming - toward a unifying methodology. Dordrecht: Kluwer Academic Publisher; 2004.
9. Milano M, Van Hentenryck P. Hybrid optimization - the ten years of CPAIOR. Springer optimization and its applications, SOIA 45. Springer. In press.
10. Achterberg T. SCIP: solving constraint integer programs. Mathematical programming computation. Berlin: Springer; 2009. DOI: 10.1007/s12532-008-0001-1, Published online: 20 January 2009.
11. Aron ID, Hooker JN, Yunes TH. SIMPL: a system for integrating optimization techniques. Proceedings of the International Conference on the Integration of AI and OR techniques

- in Constraint Programming - CPAIOR. Nice France: 2004. pp. 21–36.
12. Tsang E. Foundations of Constraint Satisfaction. San Diego (CA): Academic Press; 1993.
 13. Régim JC. A filtering algorithm for constraints of difference in CSPs. In: Hayes-Roth B, Korf R, editors. Proceedings of the 12th National Conference on Artificial Intelligence - AAAI94. Seattle (WA): AAAI Press; 1994. pp. 362–367.
 14. Dincbas M, Van Hentenryck P, Simonis H, *et al.* The constraint logic programming language CHIP. In Proceedings International Conference on 5th Generation Computer Systems; 1988; Tokyo, Japan.
 15. Apt KR, Wallace M. Constraint logic programming using ECLiPSe. Cambridge, UK: Cambridge University Press; 2006.
 16. ILOG Optimization Team. Concert technology. 2003.
 17. Régim J-C. Generalized arc consistency for global cardinality constraint. AAAI/IAAI, Volume 1. Portland (OR): 1996. pp. 209–215.
 18. Beldiceanu N, Contejean E. Introducing global constraints in CHIP. Math Comput Model 1994;20(12):97–123.
 19. Régim J-C. Arc consistency for global cardinality constraints with costs. Proceedings of the International Conference on Principles and Practice of Constraint Programming CP 1999. Alexandria (VA): 1999. pp. 390–404.
 20. Milano M, Ottosson G, Refalo P, *et al.* The role of integer programming techniques in constraint programming's global constraints. INFORMS J Comput 2002;14(4):387–402.
 21. Rodosek R, Wallace M, Hajian M. A new approach to integrating Mixed Integer Programming and Constraint Logic Programming. Recent Advances in Combinatorial Optimization. Ann Oper Res 1997;86:63–87.
 22. Laurière J. A language and a program for stating and solving combinatorial problems. Artif Intell 1978;10:29–127.
 23. van Hoeve WJ. The alldifferent constraint: a systematic overview. 2006.
 24. Brand S, Narodytska N, Quimper C-G, *et al.* Encodings of the sequence constraint. Proceedings of the International Conference on Principles and Practice of Constraint Programming CP. Providence (RI): 2007. pp. 210–224.
 25. Trick M. A dynamic programming approach for consistency and propagation for knapsack constraints. Ann Oper Res 2003;118(1–4): 73–84.
 26. Beldiceanu N, Carlsson M, Rampon J-X. Global constraint catalog. SICS Technical Report T2005:08. Available at <http://www.emn.fr/z-info/sdemasse/gccat/>. Accessed 2005.
 27. Caseau Y, Laburthe F. Solving various weighted matching problems with constraints. In: Smolka G, editor. Principle and practice of constraint programming - CP97, LNCS 1330. Berlin, Heidelberg: Springer; 1997. pp. 17–31.
 28. van Hoeve W, Le Pape C. Over-constrained problems. In: Milano M, Van Hentenryck P, editors. Hybrid optimization - Ten years of CPAIOR. Springer; 2010. In press.
 29. Baptiste P, Pape CL, Nuijten W. Constraint-based scheduling. Dordrecht, the Netherlands: Kluwer Academic Publisher; 2003.
 30. Carlier J, Pinson E. An algorithm for solving job shop scheduling. Manage Sci 1995; 35:164–176.
 31. Baptiste P, Le Pape C, Nuijten W. Efficient operations research algorithms in constraint-based scheduling. 1st Joint Workshop on Artificial Intelligence and Operational Research. Cambridge (MA): 1995.
 32. Focacci F, Lodi A, Milano M. Cost-based domain filtering. In: Jaffar J, editor. Volume 1713, Proceedings of the International Conference on Principles and Practice of Constraint Programming CP'99, LNCS. London, UK: Springer; 1999. pp. 189–203.
 33. Beldiceanu N, Bourreau E, Chan P, *et al.* Partial search strategy in CHIP. Proceedings of the 2nd International Conference on Meta-Heuristics. Alexandria (VA): 1997.
 34. Harvey WD, Ginsberg ML. Limited discrepancy search. In: Mellish CS, editor. Volume 1, Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95). Montreal, Canada: 1995. pp. 607–615.
 35. Fischetti M, Lodi A. Repairing MIP infeasibility through local branching. Comput Oper Res 2008;35(5):1436–1445.
 36. Kiziltan Z, Lodi A, Milano M, *et al.* Cp-based local branching. Proceedings of the International Conference on Principles and Practice of Constraint Programming CP. Providence (RI): 2007. pp. 847–855.
 37. Refalo P. Impact-based search strategies for constraint programming. In: Wallace M, editor. Principle and practice of constraint programming - CP2004, LNCS 3258. Berlin Heidelberg: Springer; 2004. pp. 557–571.
 38. El Sakkout H, Wallace M. Probe backtrack search for minimal perturbation in dynamic

- scheduling. *Constraints* 2000;5(4):359–388.
39. Beringer H, Backer BD. Combinatorial problem solving in constraint logic programming with cooperating solvers. In: Beierle C, Plümer L, editors. *Logic programming: formal methods and practical applications*. Amsterdam, The Netherlands: North Holland Publishing Co.; 1995. pp. 245–272.
 40. Refalo P. Linear formulation of constraint programming models and hybrid solvers. In: Dechter R, editor. *Volume 1894, Proceedings of the International Conference on Principles and Practice of Constraint Programming - CP 2000*, LNCS. Berlin: Springer; 2000.
 41. Ouaja W, Richards B. Hybrid Lagrangian relaxation for bandwidth-constrained routing: knapsack decomposition. *Proceedings of 2005 ACM symposium on Applied computing*. Santa Fe (NM): 2005. pp. 383–387.
 42. Sellmann M, Fahle T. Constraint programming based Lagrangian relaxation for the automatic recording problem. *Ann Oper Res* 2003;118(1-4):17–33.
 43. Sellmann M. Theoretical foundations of cp-based Lagrangian relaxation. *Proceedings of the International Conference on Principles and Practice of Constraint Programming CP*. Toronto, Canada: 2004. pp. 634–647.
 44. Milano M, van Hoeve WJ. Reduced cost-based ranking for generating promising subproblems. *Proceedings of the International Conference on Principles and Practice of Constraint Programming CP*. Ithaca (NY): 2002. pp. 1–16.
 45. Lodi A, Milano M, Rousseau L-M. Discrepancy-based additive bounding procedures. *INFORMS J Comput* 2006;18(4): 480–493.
 46. Benders JF. Partitioning procedures for solving mixed-variables programming problems. *Numer Math* 1962;4:238–252.
 47. Hooker JN, Ottosson G. Logic-based Benders decomposition. *Math Program* 2003; 96:33–60.
 48. Benini L, Bertozzi D, Guerri A, *et al.* Allocation, scheduling and voltage scaling on energy aware MPSoCs. *Proceedings of the International Conference on the Integration of AI and OR techniques in Constraint Programming - CPAIOR*. Cork Ireland: 2006. pp. 44–58.
 49. Hooker JN. A hybrid method for planning and scheduling. In: Wallace M, editor. *Volume 3258, Proceedings of the International Conference on Principles and Practice of Constraint Programming - CP 2004*, LNCS. Berlin: Springer; 2004. pp. 305–316.
 50. Hooker JN. Planning and scheduling to minimize tardiness. In: Van Beek P, editor. *Volume 3709, Proceedings of the International Conference on Principles and Practice of Constraint Programming - CP 2005*, LNCS. Berlin: Springer; 2005. pp. 314–327.
 51. Grossmann IE, Jain V. Algorithms for hybrid MILP/CP models for a class of optimization problems. *INFORMS J Comput* 2001;13:258–276.
 52. Junger M, Thienel S. The ABACUS system for branch-and-cut-and-price algorithms in integer programming and combinatorial optimization. *Softw Pract Exp* 2000;30:1325–1352.
 53. Cordier C, Marchand H, Laundry R, *et al.* BC-Opt: a branch-and-cut code for mixed integer programs. *Math Program* 1999;86(2): 335–354.
 54. Junker U, Karisch SE, Kohl N, *et al.* A framework for constraint programming based column generation. *Proceedings of the International Conference on Principles and Practice of Constraint Programming (CP1999)*. Alexandria (VA): 1999. pp. 261–274.
 55. Fahle T, Junker U, Karisch SE, *et al.* Constraint programming based column generation for crew assignment. *J Heuristics* 2002; 8(1):59–81.
 56. Yunes T, de Souza CC. Hybrid column generation approaches for urban transit crew management problems. *Transp Sci* 2005; 39(2):273–288.
 57. Demassey S, Pesant G, Rousseau L-M. Constraint programming based column generation for employee timetabling. In: Bartak R, Milano M, editors. *Volume 3524, Proceedings of the International Conference on Integration of AI and OR techniques in Constraint Programming for Combinatorial Optimization Problems - CPAIOR*, LNCS. Berlin: Springer; 2005. pp. 140–154.
 58. Demassey S, Pesant G, Rousseau L-M. Constraint programming based column generation for employee timetabling. *Proceedings of the International Conference on the Integration of AI and OR techniques in Constraint Programming CPAIOR 2005*. Prague, Czech Republic: 2005. pp. 140–154.
 59. Yunes T. Software tools supporting integration. In: Milano M, Van Hentenryck P, editors. *Hybrid optimization - ten years of CPAIOR*. Springer; 2010. In press.

60. Wallace M. G12 - towards the separation of problem modelling and problem solving. In: van Hoeve WJ, Hooker JN, editors. Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, 6th International Conference, CPAIOR 2009. Berlin: Springer; 2009. pp. 8–10.
61. Van Hentenryck P. The OPL optimization programming language. Cambridge (MA): MIP Press; 1999.
62. Tawarmalani M, Sahinidis NV. Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. *Math Program* 2004;99(3):563–591.
63. Achterberg T, Berthold T, Koch T, *et al.* Constraint integer programming: a new approach to integrate CP and MIP. Proceedings of the International Conference on the Integration of AI and OR techniques in Constraint Programming CPAIOR 2008. Paris, France: 2008. pp. 6–20.
64. Walsh T. Stochastic constraint programming. In: van Harmelen F, editor. Proceedings of the European Conference on Artificial Intelligence, ECAI. Lyon France: IOS Press; 2002.
65. Tarim A, Manandhar S, Walsh T. Stochastic constraint programming: a scenario-based approach. *Constraints* 2006;11:53–80.

CONSTRAINT QUALIFICATIONS

MIKHAIL V. SOLODOV
 IMPA–Instituto de Matemática
 Pura e Aplicada, Estrada Dona
 Castorina, Rio de Janeiro,
 Brazil, South America

Roughly speaking, (first-order) constraint qualifications (CQs) are properties of the analytical description of a set which ensure that the structure of the set around a given feasible point can be constructively captured by (first-order) approximations of the constraint functions defining the set. If it is so, the consequences are far reaching. CQs are essential for deriving primal and primal-dual characterizations of solutions of optimization and variational problems, duality relations, sensitivity and stability analysis, and convergence and rate of convergence of computational methods for solving optimization and variational problems.

Let D be any set in $\mathbf{R}^n$. An appropriate object to describe the geometry of D around a feasible point $\bar{x} \in D$ is the *tangent (or contingent) cone*

$$\mathcal{T}_D(\bar{x}) = \left\{ d \in \mathbf{R}^n \mid \begin{array}{l} \exists \{t_k\} \subset \mathbf{R}_+ \setminus \{0\}, \{t_k\} \rightarrow 0, \\ \exists \{d^k\} \subset \mathbf{R}^n, \{d^k\} \rightarrow d, \\ \text{such that } \bar{x} + t_k d^k \in D \\ \text{for all } k \end{array} \right\}.$$

The tangent cone includes all the feasible directions, if there are any, as well as “almost-feasible” ones in the stated sense. Suppose further that D is defined by a finite number of equality and inequality constraints, as is common in applications:

$$D = \{x \in \mathbf{R}^n \mid h(x) = 0, g(x) \leq 0\}, \quad (1)$$

where $h: \mathbf{R}^n \rightarrow \mathbf{R}^l$ and $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$ are given functions, which we assume to be continuously differentiable in the region of interest. Then CQs can be thought of as conditions imposed on the functions h , g , and/or their derivatives at or around the

point $\bar{x}$, that guarantee that the tangent cone $\mathcal{T}_D(\bar{x})$ has an explicit algebraic representation in terms of the derivatives of the constraint functions. This is crucial for developing optimality conditions in optimization, since whenever a point $\bar{x}$ is a local solution of the problem

$$\text{minimize } f(x) \quad \text{subject to } x \in D, \quad (2)$$

where $f: \mathbf{R}^n \rightarrow \mathbf{R}$ is differentiable at $\bar{x}$, it holds that

$$\langle f'(\bar{x}), x - \bar{x} \rangle \geq 0 \quad \forall x \in \bar{x} + \mathcal{T}_D(\bar{x}). \quad (3)$$

Or equivalently,

$$-f'(\bar{x}) \in (\mathcal{T}_D(\bar{x}))^\circ, \quad (4)$$

where $K^\circ = \{z \in \mathbf{R}^n \mid \langle z, y \rangle \leq 0 \forall y \in K\}$ stands for the dual (negative polar) cone of a cone K in $\mathbf{R}^n$. Constraint qualifications allow explicit characterization of the tangent cone in condition (3) and of its dual in condition (4), which makes these abstract optimality conditions tractable.

For the same reasons, CQs are important for solving the more general variational problems of the form

$$\begin{aligned} &\text{find } \bar{x} \in D \text{ such that } \langle F(\bar{x}), x - \bar{x} \rangle \geq 0 \\ &\forall x \in \bar{x} + \mathcal{T}_D(\bar{x}), \end{aligned} \quad (5)$$

or equivalently,

$$-F(\bar{x}) \in (\mathcal{T}_D(\bar{x}))^\circ, \quad (6)$$

where $F: \mathbf{R}^n \rightarrow \mathbf{R}^n$. When for some function $f: \mathbf{R}^n \rightarrow \mathbf{R}$, it holds that $F(x) = f'(x)$, $x \in \mathbf{R}^n$, then problem (5) represents the primal optimality conditions for the optimization problem (2), while condition (6) leads to the primal-dual optimality conditions. But in the variational setting F need not be integrable in general. When the feasible set D is convex, problem (5) becomes equivalent to

the classical *variational inequality*

$$\begin{aligned} &\text{find } \bar{x} \in D \text{ such that } \langle F(\bar{x}), x - \bar{x} \rangle \geq 0 \\ &\forall x \in D. \end{aligned}$$

Derivation of tractable first-order primal and primal-dual necessary optimality conditions via computing the tangent cone in condition (3) and its dual in condition (4) is perhaps the most important role of constraint qualifications. The term *CQ* was coined in Ref. 1. Alternatively, the term *regularity* is also sometimes used in the literature to refer to (some of the) CQs. CQs also appear in second-order necessary optimality conditions. And, as already mentioned, they play an important role in deriving duality relations, sensitivity/stability analysis, error bound estimates, and convergence and rate of convergence of computational methods.

TANGENT CONE AND FIRST-ORDER PRIMAL-DUAL OPTIMALITY CONDITIONS

Consider a set D defined by a finite number of equality and inequality constraints, as in Equation (1). Let $\bar{x} \in D$ be any feasible point. If $g_i(\bar{x}) < 0$ for some $i \in \{1, \dots, m\}$, by continuity $g_i(x) < 0$ for all $x \in \mathbf{R}^n$ close to $\bar{x}$, and it is clear that such constraints (*inactive* at $\bar{x}$) do not influence the geometry of the set D around the point $\bar{x}$. Let

$$A(\bar{x}) = \{i = 1, \dots, m \mid g_i(\bar{x}) = 0\},$$

be the set of inequality constraints *active* at $\bar{x}$. All the equality constraints are, of course, active at any feasible point. Consider now the cone of directions obtained by linearizing all the constraints active at $\bar{x}$:

$$\begin{aligned} \mathcal{L}_D(\bar{x}) &= \{d \in \mathbf{R}^n \mid h'(\bar{x})d = 0, \\ &\quad \langle g'_i(\bar{x}), d \rangle \leq 0 \quad \forall i \in A(\bar{x})\}, \end{aligned}$$

which is an intuitively natural candidate to represent directions tangent to D at the point $\bar{x}$. It is easy to see that $\mathcal{T}_D(\bar{x}) \subset \mathcal{L}_D(\bar{x})$ always. The fundamental question is when in fact it holds that

$$\mathcal{T}_D(\bar{x}) = \mathcal{L}_D(\bar{x}). \quad (7)$$

When the latter is the case, applying a theorem of the alternatives [2] to compute the dual of $\mathcal{L}_D(\bar{x})$, we have that

$$\begin{aligned} (\mathcal{T}_D(\bar{x}))^\circ &= \{z \in \mathbf{R}^n \mid z = \sum_{i=1}^l \lambda_i h'_i(\bar{x}) + \sum_{i \in A(\bar{x})} \mu_i g'_i(\bar{x}), \\ &\quad \lambda \in \mathbf{R}^l, \mu_i \geq 0 \quad \forall i \in A(\bar{x})\}. \end{aligned} \quad (8)$$

Then the characterization (6) of solutions of the variational problem (5) immediately translates into the following: there exists $(\lambda, \mu) \in \mathbf{R}^l \times \mathbf{R}^m$ such that

$$\begin{aligned} -F(\bar{x}) &= \sum_{i=1}^l \lambda_i h'_i(\bar{x}) + \sum_{i=1}^m \mu_i g'_i(\bar{x}), \\ h(\bar{x}) &= 0, \\ g(\bar{x}) &\leq 0, \quad \mu \geq 0, \quad \mu_i = 0 \quad \forall i \notin A(\bar{x}). \end{aligned} \quad (9)$$

In the case of $F(x) = f'(x)$, $x \in \mathbf{R}^n$, corresponding to the optimization problem, relations (9) are the *Karush–Kuhn–Tucker optimality (KKT) conditions*

$$\begin{aligned} \frac{\partial L}{\partial x}(\bar{x}, \lambda, \mu) &= 0, \\ h(\bar{x}) &= 0, \\ g(\bar{x}) &\leq 0, \quad \mu \geq 0, \quad \mu_i = 0 \quad \forall i \notin A(\bar{x}), \end{aligned} \quad (10)$$

where

$$\begin{aligned} L &: \mathbf{R}^n \times \mathbf{R}^l \times \mathbf{R}^m \rightarrow \mathbf{R}, \\ L(x, \lambda, \mu) &= f(x) + \langle \lambda, h(x) \rangle + \langle \mu, g(x) \rangle \end{aligned}$$

is the Lagrangian of the problem.

The key question, therefore, is when equality (7) or more generally equality (8) are guaranteed to hold. Obviously, equality (7) is sufficient for equality (8) but not necessary. Equality (7) is called *Abadie CQ* [3], and equality (8) is called *Guignard CQ* [4]. Guignard CQ is in a sense the weakest CQ that ensures that KKT relations (10) are necessary optimality conditions for the associated optimization problem [5]. It should be emphasized though that neither Abadie CQ nor Guignard CQ is verifiable directly in general, since they require the knowledge of the tangent cone or its dual. They are more akin to the desired properties we would like to ensure than CQs as such.

Before proceeding with the presentation of CQs, we make one final observation. CQs and the desired equality (7) depend not only on the geometry of the set D , but also on its analytic representation, that is, on the choice of the constraint functions h and g in Equation (1). For example, consider the set $D = \{0\}$ that has the unique feasible point $\bar{x} = 0$, so that $\mathcal{T}_D(\bar{x}) = \{0\}$. If this set is represented by the equality constraint with $h(x) = x$, then in Equation (1), we get $\mathcal{L}_D(\bar{x}) = \{0\}$ which gives the correct tangent cone, that is, equality (7) holds. If, on the other hand, the set is represented by the equality constraint with $h(x) = \|x\|^2$, then in Equation (1), we get $\mathcal{L}_D(\bar{x}) = \mathbf{R}^n$ and (7) is no longer valid.

Probably the most obvious CQ that guarantees equality (7) is *linearity* of constraints around the point in question. Specifically, if there exists a neighborhood U of $\bar{x}$ such that

$$h \text{ and } g_i \forall i \in A(\bar{x}) \text{ are affine in } U,$$

then equality (7) holds. Moreover, in that case the tangent cone is the set of feasible directions at $\bar{x}$.

The linear independence constraint qualification (LICQ) consists of saying that the gradients of equality and active inequality constraints are linearly independent at $\bar{x}$:

$$\begin{aligned} &\text{the set } \{h'_i(\bar{x}), i = 1, \dots, l\} \cup \{g'_i(\bar{x}), i \in A(\bar{x})\} \\ &\text{is linearly independent.} \end{aligned} \quad (11)$$

Apart from the characterization (7) of the tangent cone, LICQ further implies that the multiplier (λ, μ) satisfying primal-dual relations (9) is actually unique.

The *Mangasarian–Fromovitz constraint qualification* (MFCQ, [6]) assumes that

$$\begin{aligned} &\text{the set } \{h'_i(\bar{x}), i = 1, \dots, l\} \text{ is linearly} \\ &\text{independent and } \exists d \in \mathbf{R}^n \text{ such that} \\ &h'(\bar{x})d = 0, \langle g'_i(\bar{x}), d \rangle < 0 \forall i \in A(\bar{x}). \end{aligned} \quad (12)$$

Applying a theorem of the alternatives [2], the equivalent dual form of MFCQ states that

zero is the unique solution of

the linear system

$$\begin{aligned} &\sum_{i=1}^l \lambda_i h'_i(\bar{x}) + \sum_{i \in A(\bar{x})} \mu_i g'_i(\bar{x}) = 0, \\ &\mu_i \geq 0 \quad \forall i \in A(\bar{x}). \end{aligned} \quad (13)$$

MFCQ (12) also implies the characterization (7) of the tangent cone, while being evidently weaker than LICQ (11). Also, using the dual form (13), it is not difficult to see that at solutions of the variational problem (5), MFCQ is equivalent to the property of the multiplier set of (λ, μ) satisfying the primal-dual relations (9) being nonempty and bounded [7]. If $\bar{x}$ satisfies the primal-dual relations (9) with some (λ, μ) , then the stronger property of uniqueness of Lagrange multipliers is called the *strict Mangasarian–Fromovitz constraint qualification* (SMFCQ, [8]). MFCQ is stable in the following sense. If MFCQ holds at $\bar{x} \in D$ then there exists a neighborhood U of $\bar{x}$ such that MFCQ holds at each $x \in D \cap U$.

Slater constraint qualification consists of the following assumptions:

$$\begin{aligned} &h \text{ is an affine function, each } g_i \text{ is a} \\ &\text{convex function, and } \exists \hat{x} \in \mathbf{R}^n \text{ such that} \\ &h(\hat{x}) = 0, g_i(\hat{x}) < 0 \quad \forall i \in \{1, \dots, m\}. \end{aligned}$$

If g is convex differentiable and no equations appear, then Slater CQ is equivalent to MFCQ (12) holding at every point $\bar{x} \in D$ [9,10].

In the case when the description (1) of the set D does not contain inequality constraints LICQ, MFCQ, and SMFCQ all reduce to the classical *regularity* condition

$$\text{rank } h'(\bar{x}) = l.$$

The fact that under this assumption $\mathcal{T}_D(\bar{x})$ is the tangent subspace $\ker h'(\bar{x})$ is a consequence of the Lyusternik–Graves Theorem [11,12]; see also Ref. 13. Furthermore, KKT relations (10) for optimization reduce in this case to the classical Lagrange optimality conditions.

The *constant rank constraint qualification* (CRCQ, [14]) holds at $\bar{x} \in D$ if there exists a

neighborhood U of $\bar{x}$ such that

for every pair of index sets $I \subset \{1, \dots, l\}$
and $J \subset A(\bar{x})$ the set
 $\{h'_i(x), i \in I\} \cup \{g'_i(x), i \in J\}$
has the same rank for all $x \in D \cap U$.

(14)

In condition (14) the rank in question depends on the choice of I and J but not on the point $x \in D \cap U$. Clearly, LICQ (11) implies CRCQ (14). Linearity of constraints also implies CRCQ. Under CRCQ it holds that the tangent cone $\mathcal{T}_D(\bar{x})$ has the form (7) [14]. CRCQ is neither weaker nor stronger than MFCQ (12). Thus, nothing can be said about the multiplier set in KKT relations (9), except that it is nonempty. Note also that unlike MFCQ, if CRCQ holds at $\bar{x} \in D$, it will continue to hold if any of the equality constraints $h_i(x) = 0$ were to be replaced by the two inequalities $h_i(x) \leq 0$ and $-h_i(x) \leq 0$. MFCQ and CRCQ are related, however, in the following sense: it can be shown that under CRCQ, there exists an alternative representation of the feasible set for which MFCQ holds [15].

The *relaxed constant rank constraint qualification* (rCRCQ, [16]) holds at $\bar{x} \in D$ if there exists a neighborhood U of $\bar{x}$ such that

for every index set $J \subset A(\bar{x})$ the set
 $\{h'_i(x), i = 1, \dots, l\} \cup \{g'_i(x), i \in J\}$
has the same rank for all $x \in U$.

(15)

It is clear that CRCQ (14) implies rCRCQ (15). It can be seen from the following example [16] that the reverse implication is not valid: $D = \{x \in \mathbf{R}^2 \mid x_1 - x_2 = 0, -x_1 \leq 0, -x_1 - x_2^2 \leq 0\}$, $\bar{x} = 0$. It holds that rCRCQ is still sufficient for the tangent cone $\mathcal{T}_D(\bar{x})$ to have the desired form (7) [16]. When there are no inequality constraints, rCRCQ reduces to the *weak constant rank condition* introduced in Ref. 17.

The point $\bar{x} \in D$ satisfies the *constant positive linear dependence* condition (CPLD, [18]), if there exists a neighborhood U of $\bar{x}$ such

that

whenever for some index sets $I \subset \{1, \dots, l\}$
and $J \subset A(\bar{x})$ the system
 $\sum_{i \in I} \lambda_i h'_i(\bar{x}) + \sum_{i \in J} \mu_i g'_i(\bar{x}) = 0,$
 $\mu_i \geq 0 \forall i \in J$
has a nonzero solution, the set
 $\{h'_i(x), i \in I\} \cup \{g'_i(x), i \in J\}$
is linearly dependent for all $x \in U$.

(16)

Comparing the dual form (13) of MFCQ with condition (16), it is immediate that MFCQ implies CPLD but not vice versa. It can be seen that CPLD is also weaker than CRCQ [19]. Nevertheless, CPLD still guarantees that the tangent cone has the desired representation (7); this follows from the results in Refs 19, 20; and also from the error bound in Ref. 16, also see the section titled “Error Bounds and Metric Regularity.” Hence, KKT relations (10) are necessary optimality conditions, which had also been shown in Ref. 19. Finally, CPLD is neither weaker nor stronger than rCRCQ, as can be seen from the following example [16]: $D = \{x \in \mathbf{R}^2 \mid x_2 = 0, x_1 - x_2^2 \leq 0, -x_1 - x_2^2 \leq 0\}$, $\bar{x} = 0$.

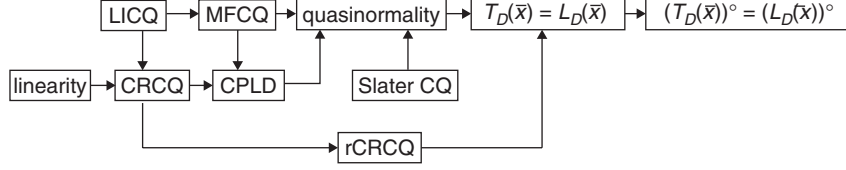
The point $\bar{x} \in D$ is said to be *quasinormal* [20, 22] if

there exist no nonzero $(\lambda, \mu) \in \mathbf{R}^l \times \mathbf{R}_+^m$
and no sequence $\{x^k\} \rightarrow \bar{x}$ such that

$$\sum_{i=1}^l \lambda_i h'_i(\bar{x}) + \sum_{i=1}^m \mu_i g'_i(\bar{x}) = 0, \text{ and} \\ \text{for all } k, \lambda_i h'_i(x^k) > 0 \text{ for all } i \text{ with } \lambda_i \neq 0, \\ \mu_i g'_i(x^k) > 0 \text{ for all } i \text{ with } \mu_i \neq 0. \quad (17)$$

Quasinormality implies that the tangent cone has the desired representation (7) [21]. Quasinormality is implied by CPLD [19].

The main relationships discussed above can be summarized as follows:



When comparing different CQs, it should also be kept in mind that conditions like CPLD (16) and rCRCQ (15) depend not only on the properties of the problem data at the point $\bar{x}$ but also on their properties in some neighborhood of this point. They require more information and, unless some stronger sufficient conditions hold, they are usually much more difficult to verify directly than, say, the classical MFCQ (12).

The more general format of constraints is given by

$$D = \{x \in \mathbf{R}^n \mid S(x) \in Q\}, \quad (18)$$

where Q is a subset of $\mathbf{R}^s$ and $S : \mathbf{R}^n \rightarrow \mathbf{R}^s$. The set (1) defined by a finite number of equality and inequality constraints is clearly a special case of Equation (18) given by $s = l + m$, $S(x) = (h(x), g(x))$, and $Q = \{0\} \times (-\mathbf{R}_+^m)$. When Q is a closed convex set, the Robinson CQ [10] holds at $\bar{x} \in D$ if

$$0 \in \text{int} \{S(\bar{x}) + \text{im } S'(\bar{x}) - Q\}. \quad (19)$$

Robinson CQ ensures that

$$\mathcal{T}_D(\bar{x}) = \{d \in \mathbf{R}^n \mid S'(\bar{x})d \in \mathcal{T}_Q(S(\bar{x}))\}. \quad (20)$$

Furthermore, if Q is a closed convex cone and $\bar{x}$ is a local solution of the optimization problem (2) then there exists $v \in \mathbf{R}^s$ such that

$$\begin{aligned} f'(\bar{x}) + (S'(\bar{x}))^\top v &= 0, \quad S(\bar{x}) \in Q, \quad v \in Q^\circ, \\ \langle v, S(\bar{x}) \rangle &= 0. \end{aligned} \quad (21)$$

In the case of equality and inequality constraints (1), Robinson CQ (19) reduces to MFCQ (12), characterization of the tangent cone (20) reduces to equality (7), and optimality conditions (21) reduce to the KKT relations (10).

SECOND-ORDER NECESSARY OPTIMALITY CONDITIONS

CQs are also important for deriving second-order necessary optimality conditions. But not all CQs discussed in the section titled “Tangent Cone and First-Order Primal-Dual Optimality Conditions” are suitable for this purpose.

In this section the problem data is assumed to be twice continuously differentiable. Let $\bar{x} \in D$ be a local minimizer of f in D . We shall denote by $\mathcal{M}(\bar{x})$ the set of Lagrange multipliers at $\bar{x}$, that is, all $(\lambda, \mu) \in \mathbf{R}^l \times \mathbf{R}^m$ satisfying the KKT relations (10). Denote by

$$\begin{aligned} \mathcal{C}(\bar{x}) = \{d \in \mathbf{R}^n \mid & \langle f'(\bar{x}), d \rangle \leq 0, \quad h'(\bar{x})d = 0, \\ & \langle g'_i(\bar{x}), d \rangle \leq 0 \quad \forall i \in A(\bar{x})\}, \end{aligned}$$

the *critical cone* of the optimization problem at $\bar{x} \in D$. Then under MFCQ (12) it holds that

$$\begin{aligned} \forall d \in \mathcal{C}(\bar{x}) \quad \exists (\lambda, \mu) \in \mathcal{M}(\bar{x}) \\ \text{such that} \quad \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \lambda, \mu)d, d \right\rangle \geq 0. \end{aligned} \quad (22)$$

It is known [23] that when SMFCQ does not hold (i.e., $\mathcal{M}(\bar{x})$ is not a singleton) the stronger version

$$\begin{aligned} \exists (\bar{\lambda}, \bar{\mu}) \in \mathcal{M}(\bar{x}) \quad \text{such that} \\ \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\lambda}, \bar{\mu})d, d \right\rangle \geq 0 \quad \forall d \in \mathcal{C}(\bar{x}) \end{aligned} \quad (23)$$

is not a necessary optimality condition in general. At the same time, under CRCQ (14) even a stronger property than Relation (23) is valid [24]:

$$\begin{aligned} \forall (\lambda, \mu) \in \mathcal{M}(\bar{x}) \quad \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \lambda, \mu)d, d \right\rangle \geq 0 \\ \forall d \in \mathcal{C}(\bar{x}). \end{aligned}$$

Weaker forms of second-order conditions make use of the subspace

$$\begin{aligned} \mathcal{C}^+(\bar{x}) = \{ & d \in \mathbf{R}^n \mid h'(\bar{x})d = 0, \\ & \langle g'_i(\bar{x}), d \rangle = 0 \forall i \in A(\bar{x}) \}, \end{aligned}$$

which is in general smaller than the critical cone $\mathcal{C}(\bar{x})$. The two cones coincide when the *strict complementarity* condition $\bar{\mu}_i > 0 \forall i \in A(\bar{x})$ holds. The following result had been established in Ref. 24. If $\bar{x}$ is a local minimizer satisfying KKT conditions (i.e., $\mathcal{M}(\bar{x}) \neq \emptyset$) and the weak constant rank condition holds, that is, there is a neighborhood U of $\bar{x}$ such that

the set $\{h'_i(x), i = 1, \dots, l\} \cup \{g'_i(x), i \in A(\bar{x})\}$
has the same rank for all $x \in U$,

then

$$\begin{aligned} \forall (\lambda, \mu) \in \mathcal{M}(\bar{x}) \quad & \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \lambda, \mu) d, d \right\rangle \geq 0 \\ \forall d \in \mathcal{C}^+(\bar{x}). \end{aligned} \quad (24)$$

In particular, any CQ that ensures that $\mathcal{M}(\bar{x}) \neq \emptyset$ (see the section titled “Tangent Cone and First-Order Primal-Dual Optimality Conditions”) in combination with the weak constant rank condition, guarantees that condition (24) holds at a local minimizer $\bar{x}$. For example, condition (24) holds under rCRCQ (15), as the latter implies both $\mathcal{M}(\bar{x}) \neq \emptyset$ and the weak constant rank condition.

ERROR BOUNDS AND METRIC REGULARITY

Describing the local structure of a set D defined by a finite number of equality and inequality constraints (1) is also closely related to the so-called *error bounds*, which are estimates of the distance from a given point to the set D in terms of computable quantities measuring violation of its constraints. Specifically, one would like to know when there exist a neighborhood U of $\bar{x} \in D$ and a constant $c > 0$ such that

$$\text{dist}(x, D) \leq c(\|h(x)\| + \|\max\{0, g(x)\}\|) \quad \forall x \in U. \quad (25)$$

Indeed, if the condition (25) is valid then the consequences for the characterization of the tangent cone as in equality (7) are immediate. It is enough to observe that for $d \in \mathcal{L}_D(\bar{x})$ and $t > 0$ small enough it holds that $h(\bar{x} + td) = h(\bar{x}) + th'(\bar{x})d + o(t) = o(t)$, $g_i(\bar{x} + td) < 0$ for all $i \notin A(\bar{x})$ and $g_i(\bar{x} + td) = g_i(\bar{x}) + t\langle g'_i(\bar{x}), d \rangle + o(t) \leq o(t)$ for all $i \in A(\bar{x})$. Then condition (25) immediately implies that $\text{dist}(x + td, D) = o(t)$ so that $d \in \mathcal{T}_D(\bar{x})$.

In the case of linear constraints the error bound (25) (with $U = \mathbf{R}^n$) is the classical Hoffman’s Lemma [25]. More generally, condition (25) is valid under rCRCQ (15) [16,26]. The error bound is also valid assuming MFCQ (12) [10] or, more generally, CPLD (16) [16].

A stronger property than the error bound (25) is that of *metric regularity* [9,27]. Consider the right-hand side perturbation of the set D , that is,

$$\begin{aligned} D(p, q) = \{x \in \mathbf{R}^n \mid & h(x) = p, g(x) \leq q\}, \\ p \in \mathbf{R}^l, q \in \mathbf{R}^m. \end{aligned} \quad (26)$$

Let $\bar{x} \in D(0, 0)$. Then the system in Equation (26) is metrically regular at $(\bar{x}, 0, 0)$ if there exist a neighborhood V of $(\bar{x}, 0, 0)$ and a constant $C > 0$ such that

$$\begin{aligned} \text{dist}(x, D(p, q)) & \leq C(\|h(x) - p\| \\ & + \|\max\{0, g(x) - q\}\|) \quad \forall (x, p, q) \in V. \end{aligned}$$

For smooth constraint systems, metric regularity holds if, and only if, MFCQ (12) holds for the unperturbed set D defined in Equation (1) [10,28]. This is an important stability property that highlights the special role of MFCQ among all the other CQs.

Robinson CQ (19) for the more general smooth constraints (18) is also equivalent to metric regularity in the following sense. Defining the perturbed set

$$D(p) = \{x \in \mathbf{R}^n \mid S(x) + p \in Q\}, \quad p \in \mathbf{R}^s,$$

with $\bar{x} \in D(0)$, metric regularity holds at $(\bar{x}, 0)$ if there exist a neighborhood V of $(\bar{x}, 0)$ and a constant $C > 0$ such that

$$\text{dist}(x, D(p)) \leq C \text{dist}(S(x) + p, Q) \quad \forall (x, p) \in V.$$

SECOND-ORDER REGULARITY

CQs discussed until now were based on at most first-order information about the constraint functions. Sometimes, in particular when classical CQs are violated, second derivatives need to be employed.

One line of analysis is concerned with deriving second-order necessary optimality conditions of the type (22) under assumptions weaker than MFCQ (12). For a feasible set defined by Equation (1), the *second-order regularity condition* holds at $\bar{x} \in D$ in a direction $d \in \mathbf{R}^n$ if

$$\begin{aligned} & \text{the set } \{h'_i(\bar{x}), i = 1, \dots, l\} \text{ is linearly} \\ & \text{independent and } \exists \xi \in \mathbf{R}^n \text{ such that} \\ & h'(\bar{x})\xi + h''(\bar{x})[d, d] = 0, \\ & \langle g'_i(\bar{x}), \xi \rangle + \langle g''_i(\bar{x})d, d \rangle < 0 \quad \forall i \in A(\bar{x}). \end{aligned} \quad (27)$$

(see Ref. 29 for a different but equivalent statement of this property). It can be easily seen that condition (27) holds automatically in any direction $d \in \mathbf{R}^n$ provided MFCQ (12) holds at $\bar{x}$. But condition (27) may hold for a given $d \in \mathbf{R}^n$ when MFCQ is violated. When second-order regularity condition holds at $\bar{x}$ in a direction $d \in \mathcal{L}_D(\bar{x})$, one can constructively characterize the so-called *second-order tangent set in a direction d*, defined by

$$\mathcal{T}_2(\bar{x}, d) = \left\{ \xi \in \mathbf{R}^n \left| \begin{array}{l} \exists \{t_k\} \subset \mathbf{R}_+ \setminus \{0\}, \{t_k\} \rightarrow 0, \\ \exists \{\xi^k\} \subset \mathbf{R}^n, \{\xi^k\} \rightarrow \xi, \\ \text{such that } \bar{x} + t_k d + \frac{1}{2} t_k^2 \xi^k \in D \\ \text{for all } k \end{array} \right. \right\}.$$

Furthermore, for $d \in \mathcal{C}(\bar{x})$, this characterization allows to establish the second-order necessary optimality condition of the form

$$\begin{aligned} & \exists (\lambda, \mu) \in \mathcal{M}(\bar{x}) \quad \text{such that} \\ & \left\langle \frac{\partial^2 L}{\partial x^2}(\bar{x}, \lambda, \mu)d, d \right\rangle \geq 0. \end{aligned} \quad (28)$$

Note that the latter subsumes that $\mathcal{M}(\bar{x})$ is nonempty, which means that the KKT relations (10) hold as well.

Another important concept of second-order regularity was developed in Ref. 30,31 for the case when there are no inequality constraints, under the name of *2-regularity*.

In somewhat different terms, we say that this condition holds at $\bar{x} \in D$ in a direction $d \in \mathbf{R}^n$ if

$$\text{im } h'(\bar{x}) + h''(\bar{x})[d, \ker h'(\bar{x})] = \mathbf{R}^l.$$

The counterpart of this concept for problems with no equality constraints and the related theory were developed in Ref. 32. An extension of these works to the case of equality and inequality constraints, as in Equation (1), was derived in Ref. 33. Specifically, according to Arutyunov *et al.* [33], 2-regularity holds at $\bar{x} \in D$ in a direction $d \in \mathbf{R}^n$ if

$$\begin{aligned} & \text{im } h'(\bar{x}) + h''(\bar{x})[d, \mathcal{L}_D(\bar{x})] = \mathbf{R}^l \text{ and} \\ & \exists \xi^1 \in \mathbf{R}^n, \exists \xi^2 \in \mathcal{L}_D(\bar{x}) \text{ such that} \\ & h'(\bar{x})\xi^1 + h''(\bar{x})[d, \xi^2] = 0, \\ & \langle g'_i(\bar{x}), \xi^1 \rangle + \langle g''_i(\bar{x})d, \xi^2 \rangle < 0 \quad \forall i \in A(\bar{x}). \end{aligned} \quad (29)$$

As in the case of second-order regularity (27), 2-regularity (29) holds automatically in any direction $d \in \mathbf{R}^n$ provided MFCQ (12) holds at $\bar{x}$. But condition (29) may hold for a given $d \in \mathbf{R}^n$ when MFCQ is violated. The role of 2-regularity is to characterize those directions $d \in \mathcal{L}_D(\bar{x})$ that actually belong to $\mathcal{T}_D(\bar{x})$ in the cases when equality (7) does not hold. See Ref. 33 for the detailed exposition of the related theory of first- and second-order necessary optimality conditions, and in particular, for combinations of 2-regularity with second-order regularity and its relevant extensions. It is important to emphasize that, unlike second-order regularity, 2-regularity does not imply that relations (10) are necessary for optimality: $\mathcal{M}(\bar{x})$ can be empty, and the first- and second-order necessary optimality conditions that can be established under 2-regularity are generally weaker than relations (10) and (28), respectively.

Finally, in Ref. 34 the 2-regularity theory was extended to the general constraints, as in Equation (18). The corresponding condition of 2-regularity at $\bar{x} \in D$ in a direction $d \in \mathbf{R}^n$ has the form

$$\begin{aligned} & 0 \in \text{int } \{S(\bar{x}) + \text{im } S'(\bar{x}) \\ & \quad + S''(\bar{x})[d, (S'(\bar{x}))^{-1}(Q - S(\bar{x}))] - Q\}. \end{aligned}$$

It is clear that this condition is implied by Robinson CQ (19).

REFERENCES

1. Kuhn HW, Tucker AW. Nonlinear programming. In: Neyman J, editor. Proceedings of second berkeley symposium of mathematical statistics and probability. Berkeley, (CA): University of California Press; 1950. pp. 481–492.
2. Mangasarian OL. Nonlinear programming. New York: McGraw Hill; 1969.
3. Abadie J. On the Kuhn-Tucker theorem. In: Abadie J, editor. Nonlinear programming. Amsterdam: North-Holland Publishing Co.; 1967.
4. Guignard M. Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space. *SIAM J Contr* 1969;7:232–241.
5. Gould FJ, Tolle JW. A necessary and sufficient qualification for constrained optimization. *SIAM J Appl Math* 1971;20:164–172.
6. Mangasarian OL, Fromovitz S. The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *J Math Anal Appl* 1967;17:37–47.
7. Gauvin J. A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Math Program*. 1977;12:136–138.
8. Kyparisis J. On uniqueness of Kuhn–Tucker multipliers in nonlinear programming. *Math Program* 1985;32:242–246.
9. Robinson SM. Regularity and stability for convex multivalued functions. *Math Oper Res* 1976;1:130–143.
10. Robinson SM. Stability theorems for systems of inequalities. Part II: differentiable nonlinear systems. *SIAM J Numer Anal* 1976;13:497–513.
11. Lyusternik LA. On the conditional extrema of functionals. *Math Sb* 1934;41:390–401.
12. Graves LM. Some mapping theorems. *Duke Math J* 1950;17:111–114.
13. Dontchev AL. The Graves theorem revisited. *J Convex Anal* 1996;3:45–53.
14. Janin R. Direction derivative of the marginal function in nonlinear programming. *Math Program Stud* 1984;21:127–138.
15. Lu S. Implications of the constant rank constraint qualification. *Math Program* 2009. Published online at DOI: 10.1007/s10107-009-0288-3.
16. Minchenko L, Stakhovski S. On relaxed constant rank regularity condition in mathematical programming. *Optimization* 2010. In press.
17. Penot JP. A new constraint qualification condition. *J Optim Theor Appl* 1986;48:459–468.
18. Qi L, Wei Z. On the constant positive linear independence condition and its application to SQP methods. *SIAM J Optim* 2000;10:963–981.
19. Andreani R, Martínez JM, Schuverdt ML. On the relation between the constant positive linear dependence condition and quasinormality constraint qualification. *J Optim Theor Appl* 2005;125:473–485.
20. Bertsekas DP, Ozdaglar AE. Pseudonormality and a lagrange multiplier theory for constrained optimization. *J Optim Theor Appl* 2002;114:287–343.
21. Hestenes MR. Optimization theory: the finite dimensional case. New York: Wiley; 1975.
22. Bertsekas DP, Ozdaglar AE. The relation between pseudonormality and quasiregularity in constrained optimization. *Optim Meth Software* 2004;19:493–506.
23. Arutyunov AV. Perturbations of extremum problems with constraints and necessary optimality conditions. *J Sov Math* 1991;54:1342–1400.
24. Andreani R, Echagüe CE, Schuverdt ML. Constant rank condition and second-order constraint qualification. *Optimization* 2010.
25. Hoffman AJ. On approximate solutions of systems of linear inequalities. *J Res Natl Bur Stand* 1952;49:263–265.
26. Lu S. Relations between the constant rank and the relaxed constant rank constraint qualifications. *Optimization* 2009. In press. Available at <http://www.unc.edu/~shulu/>
27. Borwein JM. Stability and regular points of inequality systems. *J Optim Theor Appl* 1986;48:9–52.
28. Cominetti R. Metric regularity, tangent sets and second-order optimality conditions. *Appl Math Optim* 1990;21:265–287.
29. Ben-Tal JF. Second-order and related extremality conditions in nonlinear programming. *J Optim Theor Appl* 1980;31:143–165.
30. Tretyakov AA. Necessary and sufficient conditions for optimality of p -th order. *USSR Comput Math Math Phys* 1984;24:123–127.
31. Avakov ER. Extremum conditions for smooth problems with equality-type constraints. *USSR Comput Math Math Phys* 1985;25:24–32.
32. Izmailov AF, Solodov MV. Optimality conditions for irregular inequality-constrained problems. *SIAM J Contr Optim* 2001;40:1280–1295.

33. Arutyunov AV, Avakov ER, Izmailov AF. Necessary conditions for an extremum in a mathematical programming problem. *Proc Steklov Inst Math* 2007;256:2–25.
34. Arutyunov AV, Avakov ER, Izmailov AF. Necessary optimality conditions for constrained optimization problems under relaxed constrained qualifications. *Math Program* 2008;114:37–68.

FURTHER READING

For extensions to infinite-dimensional spaces and nonsmooth data, applications in duality and sensitivity, see the following references:

- Bonnans JF, Shapiro A. *Perturbation Analysis of Optimization Problems*. Springer–Verlag, New York. 2000.
- Facchinei F, Pang JS. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer–Verlag, New York. 2003.
- Klatte D, Kummer B. *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods, and Applications*. Kluwer Academic Publishers, Dordrecht. 2002.
- Rockafellar RT, Wets JB. *Variational Analysis*. Springer–Verlag, New York. 1997.
- Mordukhovich BS. *Variational Analysis and Differentiation*. Springer–Verlag, New York. 2006.

CONTINUOUS OPTIMIZATION BY VARIABLE NEIGHBORHOOD SEARCH

JACK BRIMBERG
The Royal Military College of
Canada, Kingston, Canada
and GERAD, Montreal,
Canada

PIERRE HANSEN
HEC and GERAD, Montreal,
Canada

NENAD MLADENović
LAMIH, University of
Valenciennes, France
and Mathematical Institute,
SANU, Belgrade, Serbia

INTRODUCTION

A deterministic optimization problem may be formulated as

$$\min\{f(x)|x \in X, X \subseteq S\}, \quad (1)$$

where S, X, x , and f , respectively, denote the solution space, *feasible set*, a *feasible solution*, and a real-valued *objective function*. If S is a finite but large set, a *combinatorial optimization* problem is defined. If $S = \mathbb{R}^n$, we refer to continuous optimization. A solution $x^* \in X$ is *optimal* if $f(x^*) \leq f(x), \forall x \in X$. An *exact algorithm* for problem (1), if one exists, finds an optimal solution x^* , together with the proof of its optimality, or shows that there is no feasible solution, that is, $X = \emptyset$. Moreover, in practice, the time needed to do so should be finite (and not too long).

Metaheuristics are general frameworks to build heuristics for combinatorial and global optimization problems. Variable neighborhood search (VNS) [1–3] is a metaheuristic that systematically exploits the idea of neighborhood change, both in descent to local minima and in escape from the valleys that contain them.

The basic schemes of VNS and its extensions are simple and require few parameters. Therefore, in addition to providing very good solutions, often in simpler ways than other methods, VNS gives insight into the reasons for such a performance, which, in turn, can lead to more efficient and sophisticated implementations.

We first give preliminary results that are needed as a background to the article. Next, we review the basic rules of VNS (in the section titled “VNS–Basic Schemes”) and some of its recent extensions. In the section titled “Continuous VNS,” a description of VNS-based heuristics for continuous global optimization is given, as well as some computational results.

PRELIMINARIES

VNS is local search type metaheuristic. In this section, we will first give rules of classical local search, or steepest descent. A motivation for developing the global optimization technique VNS comes from the Variable metric method. This is an extension of steepest descent, developed for solving convex (local) optimization problems. Its rules are explained in the section titled “Variable Metric Method.” The section titled “Fixed Neighborhood Search” then gives rules of the simple local search type metaheuristic, known as *iterated local search* (ILS) or fixed neighborhood search (FNS).

Local Search

A *local search* heuristic consists in choosing an initial solution x , finding a direction of descent from x , within a neighborhood $N(x)$, and moving to the minimum of $f(x)$ within $N(x)$ in the same direction. If there is no direction of descent, the heuristic stops; otherwise, it is iterated. Usually, the steepest direction of descent, also referred to as *best improvement*, is used. This set of rules is summarized in Algorithm 1, where

```

Function LocalSearch( $x$ )
1 repeat
2    $x' \leftarrow x$ 
3    $x \leftarrow \arg \min_{y \in N(x')} f(y)$ 
   until ( $f(x) \geq f(x')$ )
4  $x \leftarrow x'$ 

```

Algorithm 1. Local search (steepest descent) heuristic.

we assume that an initial solution x is given. The output consists of a local minimum, also denoted by x , and its value.

Observe that a neighborhood structure $N(x)$ is defined for all $x \in X$. In discrete optimization problems, it usually consists of all vectors obtained from x by some simple modification, for example, in the case of 0–1 optimization, complementing one or two components of a 0–1 vector. Then, at each step, the neighborhood $N(x)$ of x is explored completely. As this may be time-consuming, an alternative is to use the *first descent* heuristic. Vectors $x_i \in N(x)$ are then enumerated systematically and a move is made as soon as a first direction of descent is found.

Variable Metric Method

The variable metric method for solving unconstrained continuous optimization problems has been suggested in Ref. 4 and independently in Ref. 5. The idea is to change the metric (and thus the neighborhood) at each iteration such that the search direction (steepest descent with respect to the current metric) adapts better to the local shape of the function. In the first iteration, a Euclidean unit ball in the n -dimensional solution space is used and the steepest descent (antigradient) direction found. At subsequent iterations, ellipsoidal balls are used and the steepest direction of descent is obtained with respect to a new metric resulting from a linear transformation. The purpose of such changes is to build up, iteratively, a good approximation to the inverse of the Hessian matrix A^{-1} of f , that is, to construct a sequence of matrices H_i with the property,

$$\lim_{i \rightarrow \infty} H_i = A^{-1}.$$

In the convex quadratic programming case, the limit is achieved after n iterations instead of an infinity of them. In this way, the so-called Newton search direction is obtained. The advantages are that (i) it is not necessary to find the inverse of the Hessian (which requires $O(n^3)$ operations) at each iteration and (ii) the second-order information is not needed.

Therefore, even in solving convex programs, a change of metric (and change of the neighborhoods induced by the metric) may produce efficient algorithms. The idea of neighborhood change for solving NP-hard problems may lead to even greater benefits.

Iterated Local Search

ILS [6], also known as *FNS* [7], is a step in between classical local search and variable metric on one side and VNS on the another. Instead of generating initial solutions completely at random, the next starting point for local search in ILS is a randomly generated solution taken from the vicinity of the best one found so far (incumbent solution). With this simple modification of the MLS, two advantages are obtained: (i) improved effectiveness: some of the solution attributes with good values in the incumbent are kept; (ii) improved efficiency: the next local search will have fewer iterations, as the incumbent will be in a deeper valley of the solution space. In order to find a perturbed solution, one needs to define a neighborhood structure $\mathcal{N}(x)$ differently than $N(x)$, the one used in the local search. The perturbed solution x' will belong to $\mathcal{N}(x)$. The ILS procedure is given in Algorithm 2.

Some extensions of the ILS Algorithm 2 are given in Ref. 8. Here the sharp acceptance criterion $f(x') \ll f_{\text{best}}$ is replaced with a more general one, allowing moves to solutions of worse quality than the incumbent x . Such deteriorating moves are also typical in the Tabu search method [9].

VNS–BASIC SCHEMES

VNS uses several neighborhoods in the search. Let us denote with $\mathcal{N}_k$, ($k = 1, \dots, k_{\text{max}}$), a finite set of preselected neighborhood structures, and with


```

Function ILS ( $x, t_{max}$ )
1  $f_{best} \leftarrow 10^{20}$ 
2 repeat
3   Generate perturb point  $x' \in \mathcal{N}(x)$  at random
4    $x' \leftarrow \text{LocalSearch}(x')$  /* Local search */
5   if  $f(x') < f_{best}$  then
6      $x \leftarrow x'; f_{best} = f(x)$ 
7    $t \leftarrow \text{CpuTime}()$ 
until  $t > t_{max}$ 

```

Algorithm 2. Steps of iterated local search.

$\mathcal{N}_k(x)$ the set of solutions in the k th neighborhood of x . We will also use the notation $N_\ell, \ell = 1, \dots, \ell_{\max}$, when describing local descent. Neighborhoods $\mathcal{N}_k$ or N_ℓ may be induced from one or more metric (or quasi-metric) functions introduced into the solution space S . An *optimal solution* x_{opt} (or global minimum) is a feasible solution where a minimum of problem (1) is reached. We call $x' \in X$ a *local minimum* of problem (1) with respect to $\mathcal{N}_k$, if there is no solution $x \in \mathcal{N}_k(x') \subseteq X$ such that $f(x) < f(x')$.

In order to solve problem (1) using several neighborhoods, the three observations listed below may be applied in a deterministic manner, a stochastic manner, or a combination of both. Indeed, VNS heavily relies upon these three observations:

- A local minimum with respect to one neighborhood structure is not necessarily a local minimum for another neighborhood structure.
- A global minimum is a local minimum with respect to all possible neighborhood structures.
- For many problems, local minima with respect to one or several neighborhoods are relatively close to each other.

We first give in Algorithm 3 the steps of the neighborhood change function that will be used later, and that is common to many VNS variants.

Function `NeighborhoodChange` (x, x', k) compares the new value $f(x')$ obtained from neighborhood $\mathcal{N}_k$ or N_l with the incumbent value $f(x)$ (line 1). If an improvement is

```

Function NeighborhoodChange ( $x, x', k$ )
1 if  $f(x') < f(x)$  then
2    $x \leftarrow x'; k \leftarrow 1$  /* Make a move */
  else
3    $k \leftarrow k + 1$  /* Next neighborhood */

```

Algorithm 3. Neighborhood change (or “Move or not”) function.

obtained, neighborhood counter k is returned to its initial value and the new incumbent updated (line 2). Otherwise, the next neighborhood is considered (line 3).

Variable Neighborhood Descent (VND)

The *variable neighborhood descent* (VND) method is obtained when the different specified neighborhoods are all examined in a deterministic way with a given order. Its steps are presented in Algorithm 4. In the descriptions of all algorithms that follow, we assume that an initial solution x is given. Note that the final solution should be a local minimum with respect to all $\ell_{\max}$ neighborhoods; hence, the chances to reach a global optimum are larger when using VND as compared with other heuristics that typically apply a single neighborhood structure. Note also that the order of neighborhoods that are used in a descent could be important, that is, a different order of the same set of neighborhoods may lead to different final results. Beside this *sequential* order of neighborhood structures in VND, one can develop a *nested* strategy. Assume, for example, that $\ell_{\max} = 3$. Then a possible nested strategy is to perform

```

Function VND ( $x, \ell_{max}$ )
1  $\ell \leftarrow 1$ 
2 repeat
3    $x' \leftarrow \arg \min_{y \in N_\ell(x)} f(y)$  /* Find the best solution in  $N_\ell(x)$  */
4   NeighborChange ( $x, x', \ell$ ) /* Change neighborhood */
until  $\ell = \ell_{max}$ 

```

Algorithm 4. Steps of the basic VND.

VND for the first two neighborhoods, in each point x' that belongs to the third ($x' \in N_3(x)$). Such an approach is successfully applied, for example, in Refs 10 and 11.

Reduced VNS

The *Reduced* VNS (RVNS) method is implemented when random points are selected from $\mathcal{N}_k(x)$ and no descent is made. Rather, the values of these new points are compared with that of the incumbent and updating takes place in case of improvement. We assume that a stopping condition has been chosen among various possibilities, for example, the maximum cpu time allowed t_{max} , or the maximum number of iterations between two improvements. To simplify the description of the algorithms, we always use t_{max} . Therefore, RVNS requires two parameters: t_{max} and k_{max} . Its steps are presented in Algorithm 5. With the function *Shake* represented in line 4, we generate a point x' at random from the k th neighborhood of x , that is, $x' \in \mathcal{N}_k(x)$.

RVNS is a useful method for very large instances where local search may be costly. It has been observed that the best value for

the parameter k_{max} is often 2. In addition, the maximum number of iterations between two improvements is usually used as a stopping condition. RVNS is akin to a Monte-Carlo method, but is more systematic. See, for example, [12] where the results for a continuous min–max problem obtained by RVNS are 30% better than those of a Monte-Carlo method. When applied to the p -median problem, RVNS gives solutions comparable to the *fast interchange* heuristic of [13] while being 20–40 times faster [14].

Basic VNS

The BVNS method [1] combines deterministic and stochastic changes of neighborhood. Its steps are given in Algorithm 6 (also see Figure 1).

Often successive neighborhoods $\mathcal{N}_k$ will be nested. Observe that point x' is generated at random in step 4 in order to avoid cycling, as might occur if a deterministic rule were applied. In step 5, a first improvement local search is usually adopted. However, it can be replaced with best improvement (Algorithm 1).

General VNS

Note that the Local search step 5 may also be replaced by VND (Algorithm 4). Using this general VNS (VNS/VND) approach has led to the most successful applications reported (see, recent surveys [3, 15]). The steps of GVNS are given in Algorithm 7.

CONTINUOUS VNS

We assume that $f : R^n \rightarrow R$ is a continuous function in Equation (1). No further assumptions are made on f . In particular, f does not need to be convex or smooth.

```

Function RVNS ( $x, k_{max}, t_{max}$ )
1 repeat
2    $k \leftarrow 1$ 
3   repeat
4      $x' \leftarrow \text{Shake}(x, k)$ 
5     NeighborChange ( $x, x', k$ )
   until  $k = k_{max}$ 
6    $t \leftarrow \text{CpuTime}()$ 
until  $t > t_{max}$ 

```

Algorithm 5. Steps of the reduced VNS.

```

Function VNS ( $x, k_{max}, t_{max}$ )
1 repeat
2    $k \leftarrow 1$ 
3   repeat
4      $x' \leftarrow \text{Shake}(x, k)$            /* Shaking */
5      $x'' \leftarrow \text{LocalSearch}(x')$     /* Local search */
6      $\text{NeighborhoodChange}(x, x'', k)$  /* Change neighborhood */
7   until  $k = k_{max}$ 
8    $t \leftarrow \text{CpuTime}()$ 
9 until  $t > t_{max}$ 

```

Algorithm 6. Steps of the basic VNS.

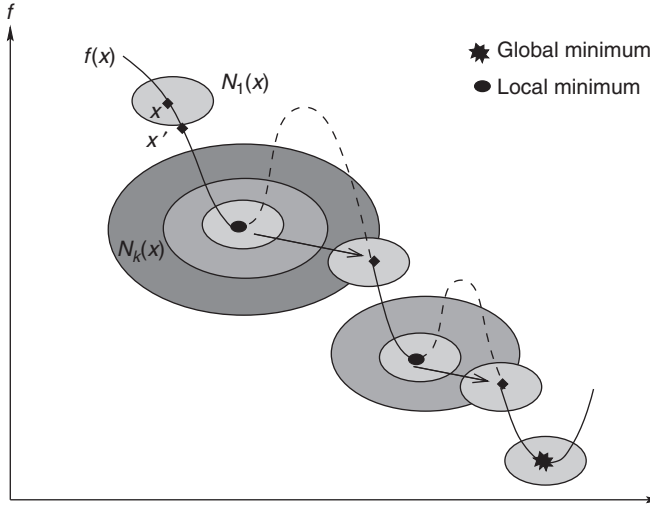


Figure 1. Basic VNS.

```

Function GVNS ( $x, \ell_{max}, k_{max}, t_{max}$ )
1 repeat
2    $k \leftarrow 1$ 
3   repeat
4      $x' \leftarrow \text{Shake}(x, k)$ 
5      $x'' \leftarrow \text{VND}(x', \ell_{max})$ 
6      $\text{NeighborhoodChange}(x, x'', k)$ 
7   until  $k = k_{max}$ 
8    $t \leftarrow \text{CpuTime}()$ 
9 until  $t > t_{max}$ 

```

Algorithm 7. Steps of the general VNS.

Glob-VNS

Several VNS-based methods for solving continuous (un)constrained optimization problems have been proposed in the literature.

For example, radar polyphase code design is considered in Mladenović *et al.* [12], while Audet *et al.* [16] solve the bilinear (pooling) problem; Dražić *et al.* [17] suggest a VNS-based heuristic called **Glob-VNS** for solving general unconstrained optimization problems. Its pseudo-code is given in Algorithm 8.

Beside t_{max} and k_{max} , the choice of the local search solver may be seen as an additional parameter of **Glob-VNS**. The user can use one out of six well-known unconstrained methods offered in the software. Three of them are first-order methods (steepest descent, Fletcher–Reeves, variable metric) and three are direct search methods (Nelder–Mead, Hook–Jeves, Rosenbrock). In the Shaking step, there are two aspects that need to be considered. One is the set of

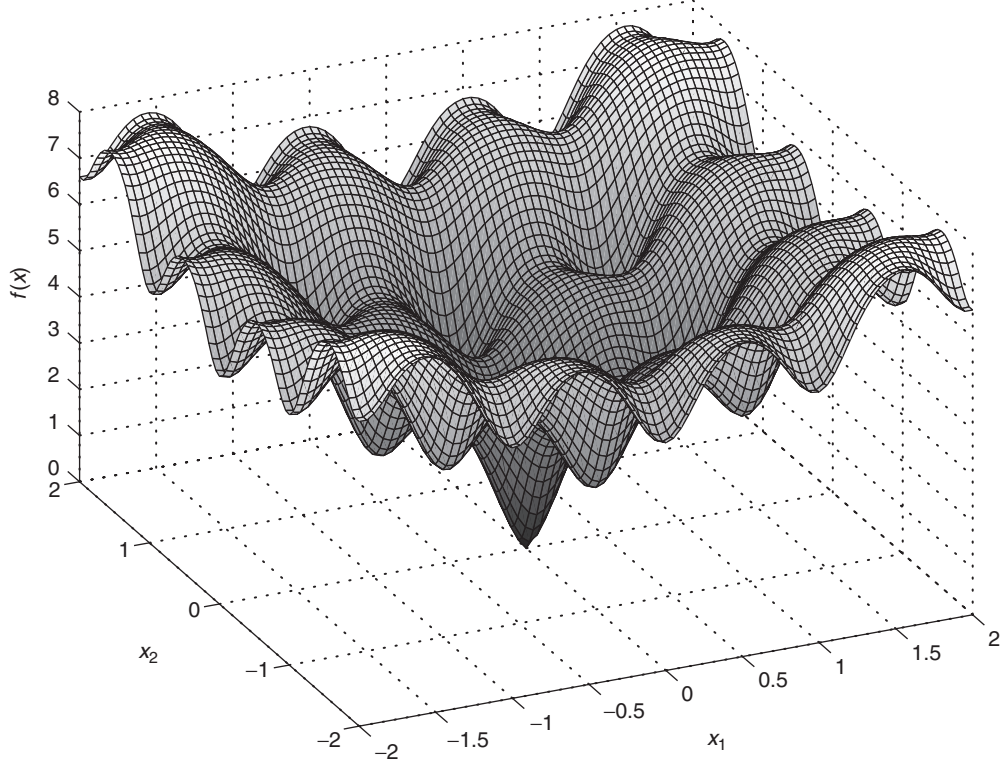


Figure 2. Ackley test function (AC2) over $[-2, 2] \times [-2, 2]$.

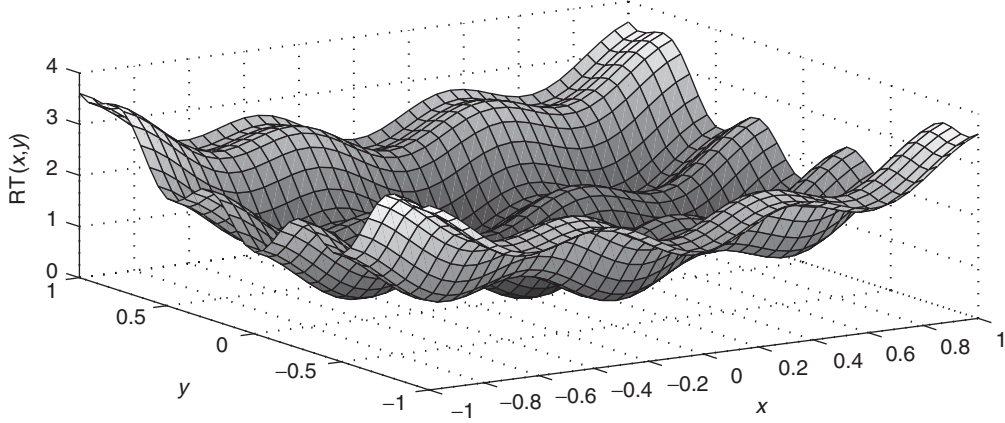


Figure 3. Rastrigin test function (RA2) over $[-1, 1] \times [-1, 1]$.

neighborhood structures $\mathcal{N}_k$ (induced from some metric such as the ℓ_p norm), while the other concerns how to take a random point from $\mathcal{N}_k(x)$. The second question relates

to the distribution $\mathcal{D}_j$ used. Neighborhood structures $\mathcal{N}_k$ are usually induced from the ℓ_∞ norm. Radii around the incumbent are found automatically, taking into account the

Function Glob-VNS (x, k_{max}, t_{max})

```

1 Select the set of neighborhood structures  $\mathcal{N}_k, k = 1, \dots, k_{max}$ 
2 Select the array of distribution types  $\mathcal{D}_j, j = 1, \dots, r$  and an initial point  $x \in X$ 
3  $f^* \leftarrow f(x), t \leftarrow 0$ 
4 while  $t < t_{max}$  do
5    $k \leftarrow 1$  // Initialize neighborhood counter  $k$ 
6   repeat
7     for all distribution types  $\mathcal{D}_j$  do
8        $x' \leftarrow \text{Shake}(x, k)$  // Get  $x' \in \mathcal{N}_k(x)$  at random with  $\mathcal{D}_j$  distribution
9        $x'' \leftarrow \text{LocalSearch}(x')$  // Apply Local Search to get a local minimum  $x''$ 
10      if  $f(x'') < f^*$  then
11         $x \leftarrow x'', f^* \leftarrow f(x''), \text{go to } 5$ 
12       $k \leftarrow k + 1$ 
13    until  $k = k_{max}$ 
14   $t \leftarrow \text{CpuTime}()$ 

```

Algorithm 8. Glob-VNS algorithm for unconstrained optimization.

box constraints: each coordinate is divided in k_{max} intervals from both sides of incumbent, forming k_{max} concentric hypercubes.

Gauss-VNS

Toskari and Guner [18] propose VNS variants that do not use metric distances to define neighborhoods; Mladenovic *et al.* [19] extend their unconstrained solver GLOB-VNS to the constrained case, using an exterior penalty function method; Audet *et al.* [16] and Bierlaire *et al.* [20] hybridize VNS with trust

region methods; Carrizosa *et al.* [21] modify GLOB-VNS by the so-called Gaussian VNS (Gauss-VNS), whose steps are given in Algorithm 9.

The paradigm of neighborhoods is generalized so that problems with unbounded domains can be addressed. The idea is to replace the class $\{\mathcal{N}_k(x)\}_{1 \leq k \leq k_{max}}$ of neighborhoods of point x by a class of *probability distributions* $\{\mathcal{P}_k(x)\}_{1 \leq k \leq k_{max}}$. The next random point in the shaking step is generated using the probability distribution $\mathcal{P}_k(x)$.

Function Gauss-VNS (x, k_{max}, t_{max})

```

1 Select the set of covariance matrices  $\Sigma_k, k = 1, \dots, k_{max}$ 
2 Select an initial point  $x \in X$ 
3  $f^* \leftarrow f(x), t \leftarrow 0$ 
4 while  $t < t_{max}$  do
5    $k \leftarrow 1$ 
6   repeat
7      $x' \leftarrow \text{Shake}(x, k)$  // Get  $x'$  from a Gaussian distribution with
8       // mean  $x$  and covariance matrix  $\Sigma_k$ 
9      $x'' \leftarrow \text{LocalSearch}(x')$  // Apply Local Search to get a local minimum  $x''$ 
10     $k \leftarrow k + 1$ 
11    if  $f(x'') < f^*$  then
12       $x \leftarrow x'', f^* \leftarrow f(x''), k \leftarrow 1$ 
13    until  $k = k_{max}$ 
14   $t \leftarrow \text{CpuTime}()$ 

```

Algorithm 9. Gauss-VNS algorithm for unconstrained optimization.

A comparison of two versions of VNS is conducted in Ref. 21 on the Ackley (ACn) function (see Figure 2 for $n = 2$):

$$\begin{aligned} f(x) &= 20 + e - 20 \exp \left(-0.2 \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \right) \\ &\quad - \exp \left(\frac{1}{n} \sum_{i=1}^n \cos(2\pi x_i) \right) \\ &- 15 \leq x_i \leq 30, i = 1, \dots, n, f_{\min} = 0, \end{aligned}$$

and Rastrigin function (RAn) (see Figure 3 for $n = 2$):

$$\begin{aligned} f(x) &= 10n + \sum_{i=1}^n (x_i^2 - 10 \cos(2\pi x_i)) \\ &- 5.12 \leq x_i \leq 5.12, i = 1, \dots, n, f_{\min} = 0. \end{aligned}$$

Results are reported in Table 1. Both methods used steepest descent local search (SD) and $k_{\max} = 10$ or 15. Global minima are found by both methods and for all test instances, but with a different number of function evaluations (or computer effort (ce)). The % deviation is calculated in the usual way: $(ce_{\text{glob}} - ce_{\text{gauss}})/ce_{\text{gauss}} \cdot 100\%$.

It appears that the Gauss-VNS method was able to find optimal solutions with much less effort than Glob-VNS, while the opposite is true for the Rastrigin function.

VNS with Modified Nelder–Mead as a Local Search

The weak point of both Glob-VNS and Gauss-VNS is the fact that the choice of a local search procedure is left to the user (steps 9 and 8 in Glob-VNS and Gauss-VNS, respectively). In order to make the VNS method more user friendly, a restarted modified Nelder–Mead (RMNM) method [22, 23] is suggested in Ref. 24 as the local search for any type of optimization problem. The basic ideas of RMNM are to (i) perform shrink move one vertex at a time, instead of moving all vertices in the direction of best; (ii) take parameter values $\alpha, \beta, \gamma, \delta$ at random from a given interval, that is, the next simplex vertex is obtained by

$$x_{\text{new}} = (1 + g) \cdot \bar{x} - g \cdot x_{n+1}, \quad (2)$$

where $\bar{x}$ is the centroid of the simplex $X = \{x_1, \dots, x_n, x_{n+1}\} \setminus \{x_{n+1}\}$, x_{n+1} a simplex vertex with the worst objective function value, and $g = \alpha, \alpha \cdot \beta, \alpha \cdot \gamma$ or $-\gamma$, expressing *reflection*, *expansion*, *inside contraction*, and *outside contraction*, respectively [23]; (iii) once the volume of the simplex is less than an arbitrarily small number ε , the procedure is restarted with the initial value of the simplex side length *size* unless no improvement has been reached from the previous big iteration. The steps of the RMNM used in the local search routine of Glob-VNS are given in Algorithm 10.

In step 2, the initial simplex $X = \{x_1, \dots, x_{n+1}\}$ is generated around a given point $x_1 \in R^n$, and then its vertices are ordered according to their quality ($f(x_1) \leq \dots \leq f(x_{n+1})$). If a Nelder–Mead iteration is unsuccessful, that is, if there is no point on the line that connects x_{n+1} and centroid $\bar{x}$ with better objective function value, we perform the shrink step only for one vertex x_j of the simplex X . Otherwise, the simplex is updated: the worst point is swapped with the new one, and all vertices of X are again ranked according to their objective function values. Once a local minimum is reached (a big iteration ends), the procedure is restarted with the same simplex size as the initial one, but now centered around the new local minimum. This restart procedure allows RMNM to attempt an escape from the current local minimum.

Nondifferentiable Test Instances

This user-friendly VNS-based method is compared in Ref. 24 with a Glob-VNS that uses standard NM as the local search. Ten nondifferentiable functions in high-dimensional space taken from Haarala [25] are examined as follows:

LND1. Generalization of MAXQ

$$f(x) = \max_{1 \leq i \leq n} x_i^2, f_{\text{opt}} = 0.$$

LND2. Generalization of MXHILB

$$f(x) = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n \frac{x_j}{i+j-1} \right|, f_{\text{opt}} = 0.$$

Table 1. Comparison of Glob-VNS and Gauss-VNS on Ackley and Rastrigin Functions

Function	n	Local Minimum	$k_{\max}$	Computer Effort (ce)		Deviation (%)
				Glob-VNS	Gauss-VNS	
AC10	10	SD	10	188670	50149	276.22
AC20	20	SD	10	433194	158412	173.46
AC30	30	SD	10	909918	304825	198.51
AC40	40	SD	10	1577138	528718	198.30
AC50	50	SD	10	4791075	1143721	318.90
AC60	60	SD	10	7820247	2315178	237.78
AC70	70	SD	10	36641634	4255533	761.04
AC80	80	SD	10	212944367	17180658	1139.44
Average						412.96
RA10	10	SD	15	52471	85589	-38.69
RA20	20	SD	15	213597	287075	-25.60
RA30	30	SD	15	366950	599635	-38.80
RA40	40	SD	15	697160	1115923	-37.53
RA50	50	SD	15	1334842	1504701	-11.29
RA100	100	SD	15	5388075	6248753	-13.77
RA150	150	SD	15	11007093	13678014	-19.53
RA200	200	SD	15	24026456	31639001	-24.06
Average						-26.16

```

function RMNM( $f, n, x^*, size$ )
1  $x_1 \leftarrow \text{InitialPoint}; j \leftarrow 0$ 
  repeat
2    $X \leftarrow \text{InitialSimplex}(x_1, size); X^* \leftarrow X; x^* \leftarrow x_1$ 
3   while Stopping condition is not met do
4     if not MMIteration( $X$ ) then
5        $X \leftarrow X^*; j \leftarrow j + 1$ 
6        $X \leftarrow \text{Shrink}(X, j)$ 
7       if  $j = n$  then
8          $X^* \leftarrow X; j \leftarrow 0$ 
9       else
10         $X^* \leftarrow X; j \leftarrow 0$ 
  until  $f(x_1) < f(x^*)$ 
return  $x^*$ 

```

Algorithm 10. Restarted Modified NM (RMNM).

LND3. Chained LQ

$$f(x) = \sum_{i=1}^{n-1} \max\{-x_i - x_{i+1}, -x_i - x_{i+1} + (x_i^2 + x_{i+1}^2 - 1)\}, f_{\text{opt}} = -(n-1)\sqrt{2}.$$

LND4. Chained CB3 I

$$f(x) = \sum_{i=1}^{n-1} \max\{x_i^4 + x_{i+1}^2, (2 - x_i)^2 + (2 - x_{i+1})^2, 2e^{-x_i + x_{i+1}}\}, f_{\text{opt}} = 2(n-1).$$

LND5. Chained CB3 II

$$f(x) = \max \left\{ \sum_{i=1}^{n-1} (x_i^4 + x_{i+1}^2), \sum_{i=1}^{n-1} ((2 - x_i)^2 + (2 - x_{i+1})^2), \sum_{i=1}^{n-1} (2e^{-x_i + x_{i+1}}) \right\}, f_{\text{opt}} = 2(n-1).$$

LND6. Number of active faces

$$f(x) = \max_{1 \leq i \leq n} \left\{ g \left(-\sum_{j=1}^n x_j \right), g(x_i) \right\}, g(y) = \ln(|y| + 1), f_{\text{opt}} = 0.$$

LND7. Nonsmooth generalization of Brown function 2

$$f(x) = \sum_{i=1}^{n-1} (|x_i|^{x_{i+1}^2+1} + |x_{i+1}|^{x_i^2+1}), f_{\text{opt}} = 0.$$

LND8. Chained Mifflin 2

$$f(x) = \sum_{i=1}^{n-1} (-x_i + 2(x_i^2 + x_{i+1}^2 - 1) + 1.75|x_i^2 + x_{i+1}^2 - 1|)$$

f_{opt} varies, $f_{\text{opt}} = -34.795$ for n
 $= 50$, $f_{\text{opt}} = -140.86$ for $n = 200$.

Table 2. Large Nondifferentiable Test Functions

Function		VNS + NM			VNS + RMNM		
n	f_{opt}	ce	succ.	f_{min}	ce	succ.	f_{min}
LND1. Generalization of MAXQ							
10	0	5657	10		5601	10	
20	0	28,351	10		26,637	10	
30	0	67,566	10		68,105	10	
40	0	120,672	10		125,107	10	
50	0	253,598	10		243,742	10	
LND2. Generalization of MXHILB							
10	0	62,219	10		10,873	10	
20	0	5,611,105	0	0.000185	63,118	10	
30	0	2,627,389	0	0.000488	167,127	10	
40	0	1,563,528	0	0.000709	747,442	8	0.000101
50	0	995,644	0	0.000989	1,005,813	1	0.000151
LND3. Chained LQ							
10	-12.7279	9296	10		5277	10	
20	-26.8701	122,393	10		29,266	10	
30	-41.0122	550,353	10		62,813	10	
40	-55.1543	4,861,354	5	-55.1426	384,261	10	
50	-69.2965	2,819,445	8	-69.2823	154,489	10	
LND4. Chained CB3 I							
10	18	8601	10		8292	10	
20	38	36,317	10		31,394	10	
30	58	51,788	10		51,699	10	
40	78	97,106	10		93,124	10	
50	98	201,912	10		143,845	10	
LND5. Chained CB3 II							
10	18	8912	10		9775	10	
20	38	28,893	10		48,761	10	
30	58	65,678	10		86,455	10	
40	78	82,356	10		133,536	10	
50	98	107,374	10		145,274	10	

Table 2. (*continued*)

Function		VNS + NM			VNS + RMNM		
n	f_{opt}	ce	succ.	f_{min}	ce	succ.	f_{min}
LND6. Number of active faces							
10	0	199,318	10		7776	10	
20	0	12,991,251	0	0.000218	145,763	10	
30	0	8,069,203	0	0.001117	8,189,895	1	0.000139
40	0	5,486,154	0	0.00220	5,799,417	0	0.000254
50	0	3,878,521	0	0.002894	4,088,098	0	0.000403
LND7. Nonsmooth generalization of Brown function 2							
10	0	34,591	10		8498	10	
20	0	6,869,657	1	0.000182	43,204	10	
30	0	4,413,961	0	0.000691	79,623	10	
40	0	3,089,010	0	0.00146	226,122	10	
50	0	2,344,423	0	0.00299	999,753	9	0.0001
LND8. Chained Mifflin 2							
10	-6.5146	30,960	10		20,413	10	
20	-13.5831	787,015	10		195,133	10	
30	-20.6535	2,595,076	9	-20.6493	1,071,285	10	
40	-27.7243	4,619,462	7	-27.718	2,442,460	9	-27.7186
50	-34.795	4,166,819	2	-34.7809	3,179,337	6	-34.7871
LND9. Chained crescent I							
10	0	14,831	10		7626	10	
20	0	81,819	10		35,149	10	
30	0	130,844	10		68,264	10	
40	0	306,036	10		73,833	10	
50	0	290,204	10		90,509	10	
LND10. Chained crescent II							
10	0	148,252	10		90,998	10	
20	0	7,240,978	10		443,301	10	
30	0	8,651,335	1	0.00059	3,280,270	10	
40	0	6,472,405	0	0.00163	6,492,246	2	0.00023
50	0	4,443,330	0	0.0155	3,928,254	5	0.00508

LND9. Chained crescent I

$$f(x) = \max \left\{ \sum_{i=1}^{n-1} (x_i^2 + (x_{i+1} - 1)^2 + x_{i+1} - 1), \right. \\ \left. \times \sum_{i=1}^{n-1} (-x_i^2 - (x_{i+1} - 1)^2 + x_{i+1} + 1) \right\}, f_{\text{opt}} = 0.$$

LND10. Chained crescent II

$$f(x) = \sum_{i=1}^{n-1} \max \{ x_i^2 + (x_{i+1} - 1)^2 + x_{i+1} - 1, \\ -x_i^2 - (x_{i+1} - 1)^2 + x_{i+1} + 1 \}, f_{\text{opt}} = 0.$$

These test functions are examined in different dimensional spaces with $n = 10, 20, 30, 40$, and 50 . The results are summarized in Table 2. The first two columns contain the dimension and the global optimum value of the test function. The results for each VNS variant are then presented in three successive columns. The first column (ce) contains the average number of overall function evaluations from 10 test runs. The second column (succ.) contains the number of runs where the global optimum is found with respect to desired precision

$$\frac{f_{\text{min}} - f_{\text{opt}}}{|f_{\text{min}}| + |f_{\text{opt}}| + 1} < \varepsilon_1, \quad \varepsilon_1 = 10^{-4}.$$

Local optima are detected if

$$\frac{f_{\max} - f_{\min}}{|f_{\max}| + |f_{\min}| + 1} < \varepsilon_2, \quad \varepsilon_2 = 10^{-5},$$

where $f_{\max}$ and $f_{\min}$ are the maximum and minimum function values, respectively, at the simplex vertices. If the optimal value is not found in all test runs, column ($f_{\min}$) contains the average solution value from 10 runs; otherwise this box is empty.

The results from Table 2 indicate that the more reliable method is VNS + RMNM. This method could not find an optimal solution within 10 restarts in only 2 out of 50 instances. For all 500 runs (50 instances $\times$ 10 restarts), it failed to reach optimality only 59 times. Even in these cases, the solutions found were very close to the global optima (within 10^{-3} precision).

CONCLUSIONS

Basic variants of the global optimization method, VNS, are first briefly discussed in this article. Then some recent VNS-based algorithms for solving continuous unconstrained optimization problems are given. They are computationally compared among themselves on standard differentiable and nondifferentiable test instances from the literature. The VNS procedures are found to be efficient and effective.

REFERENCES

1. Mladenović N, Hansen P. Variable neighborhood search. *Comput Oper Res* 1997;24:1097–1100.
2. Hansen P, Mladenović N, Moreno Pérez JA. Variable neighborhood search. *Eur J Oper Res* 2008;191(3):593–595.
3. Hansen P, Mladenović N, Moreno-Pérez JA. Variable neighborhood search: methods and applications. *Ann OR* 2010;175:367–407.
4. Davidon WC. Variable metric algorithm for minimization. Argonne National Laboratory Report ANL-5990; South Lemont, USA; 1959.
5. Fletcher R, Powell MJD. Rapidly convergent descent method for minimization. *Comput J* 1963;6:163–168.
6. Johnson DS, McGeoch LA. The traveling salesman problem: a case study in local optimization. In: Aarts EHL, Lenstra JK, editors. *Local search in combinatorial optimization*. New York: John Wiley & Sons; 1995. p 215–310.
7. Brimberg J, Hansen P, Mladenović N, et al., Improvements and comparison of heuristics for solving the multisource Weber problem. *Oper Res* 2000;48(3):444–460.
8. Lourenco H, Martin O, Stutzle T. Iterated local search: framework and applications. In: Gendreau M, Potvin JY, editors. *Handbook of Metaheuristics*. 2nd ed. Volume 146, International Series in Operations Research & Management Sciences. London: Kluwer; 2010. p 363–397.
9. Glover F, Laguna M. *Tabu search*. Dordrecht: Springer; 1998.
10. Mladenović N, Urošević D, Hanafi S. Variable neighborhood search for the traveling deliveryman problem. *4-OR* 2012. DOI: 10.1007/s10288-012-0212-1.
11. Mladenovic N, Urošević D, Hanafi S, et al., A general variable neighborhood search for the one-commodity pickup-and-delivery travelling salesman problem. *Eur J Oper Res* 2012;220:270–285.
12. Mladenović N, Petrović J, Kovačević-Vujčić V, et al., Solving spread spectrum radar polyphase code design problem by tabu search and variable neighborhood search. *Eur J Oper Res* 2003;151:389–399.
13. Whitaker R. A fast algorithm for the greedy interchange of large-scale clustering and median location problems. *INFOR* 1983;21:95–108.
14. Hansen P, Mladenović N, Pérez-Brito D. Variable neighborhood decomposition search. *J Heuristics* 2001;7(4):335–350.
15. Hansen P, Mladenovic N, Brimberg J, et al., Variable neighbourhood search. In: Gendreau M, Potvin JY, editors. *Handbook of Metaheuristics*. 2nd ed. Volume 146, International Series in Operations Research & Management Sciences. New-York: Kluwer; 2010. p 61–86.
16. Audet C, Brimberg J, Hansen P, et al., Pooling problem: alternate formulation and solution methods. *Manage Sci* 2004;50:761–776.
17. Dražić M, Kovacevic-Vujčić V, Cangalović M, et al., GLOB - a new VNS-based software for global optimization. In: Liberti L, Maculan N, editors. *Global optimization: from theory to implementation*. London: Springer; 2006. p 135–144.

18. Toksari AD, Güner E. Solving the unconstrained optimization problem by a variable neighborhood search. *J Math Anal Appl* 2007;328(2):1178–1187.
19. Mladenović N, Dražić M, Kovačević-Vujčić V, et al., General variable neighborhood search for the continuous optimization. *Eur J Oper Res* 2008;191(3):753–770.
20. Bierlaire M, Thémans M, Zufferey N. A heuristic for nonlinear global optimization. *INFORMS J Comput* 2010;22(1):59–70.
21. Carrizosa E, Dražić M, Dražić Z, et al., Gaussian variable neighborhood search for continuous optimization. *Comput Oper Res* 2012;39:2206–2213.
22. Zhao Q, Urošević D, Mladenović N, et al., A restarted and modified simplex search for unconstrained optimization. *Comput Oper Res* 2009;36:3263–3271.
23. Zhao Q, Mladenović N, Urošević D. A parametric simplex search for unconstrained optimization problem. *Trans Adv Res* 2012;8:22–27.
24. Dražić M, Dražić Z, Mladenović N, et al., Continuous VNS with modified Nelder-Mead for non-differentiable optimization. *IMA J Management Math*, 2014. DOI: 10.1093/iman/dpu012 (First published online: June 2, 2014)
25. Haarala M. Large-Scale Nonsmooth Optimization: variable metric bundle method with limited memory. Jyväskylä: University of Jyväskylä; 2004. p 96–98.

CONTINUOUS-TIME CONTROL UNDER STOCHASTIC UNCERTAINTY

KURT MARTI

Federal Armed Forces University
Munich, Aero-Space Engineering
and Technology, Munich,
Germany

STOCHASTIC CONTROL SYSTEMS

Optimal control and regulator problems arise in many concrete engineering applications (mechanical, electrical, thermodynamical, chemical, etc.) [1–5]. The basic control system (input–output system) is mathematically represented [5,6] by a system of first-order random differential equations:

$$\dot{z}(t) = g(t, \omega, z(t), u(t)), t_0 \leq t \leq t_f, \omega \in \Omega \quad (1a)$$

$$z(t_0) = z_0(\omega). \quad (1b)$$

Here, ω is the basic random element taking values in a probability space $(\Omega, \mathcal{A}, P)$, and describing the present random variations of model parameters and initial conditions, or the influence of noise terms. The probability space $(\Omega, \mathcal{A}, P)$ consists of the sample space or set of elementary events Ω , the σ -algebra $\mathcal{A}$ of events and the probability measure P . The plant state vector $z = z(t, \omega)$ is an m -vector involving measurable quantities like displacements, stresses, voltage, current, pressure, concentrations, and so on, $z_0(\omega)$ is the random initial state. The plant control $u(t)$ is a deterministic or stochastic n -vector denoting system inputs like external forces or moments, voltage, and so on. Furthermore, $\dot{z}$ denotes the derivative with respect to the time t . We assume that $u = u(t)$ is chosen such that $u(\cdot) \in U$, where U is a suitable linear subspace of the space $PC_0^n[t_0, t_f]$ of piecewise continuous functions $u(\cdot) : [t_0, t_f] \rightarrow \mathbb{R}^n$, normed by the maximum

norm. If $u = u(t, \omega), t_0 \leq t \leq t_f$, is a random input function, then correspondingly we suppose that $u(\cdot, \omega) \in U$ a.s. (almost sure) (with probability 1).

Let Z be a linear space of functions $z(\cdot) : [t_0, t_f] \rightarrow \mathbb{R}^m$ containing the set $PC_1^m[t_0, t_f]$ of possible trajectories of the plant; hence, the set of piecewise differentiable functions on the interval $[t_0, t_f]$, the space Z is also normed by the maximum norm. $D(\subset U)$ denotes the convex set of admissible controls $u(\cdot)$, defined for example, by some box constraints. Using the available information $\mathcal{A}_t$ up to a certain time t , the problem is then to find an optimal control $u^* = u^*(t)$ being most insensitive with respect to random parameter variations. This can be obtained by minimizing the (conditional) expected costs arising along the trajectory $z = z(t)$ and/or at the terminal state $z_f = z(t_f)$ subject to the plant differential equations (1a) and (1b) and the required control and state constraints. Optimal controls being most insensitive with respect to random parameter variations are also called *robust* optimal controls. Such controls can be obtained by stochastic optimization methods [7].

Since feedback control laws can be approximated very efficiently [3] by means of open-loop feedback (OLF) control laws, see the section titled “Control Laws”, for practical applications we may confine to the computation of deterministic stochastic optimal open-loop (OL) controls $u = u(t; t_b), t_b \leq t \leq t_f$, on arbitrary remaining time intervals $[t_b, t_f]$ of $[t_0, t_f]$. Here, $u(\cdot; t_b)$ is stochastic optimal with respect to the information $\mathcal{A}_{t_b}$ available at time t_b .

Random Differential Equations

Since in many technical applications the random input into the underlying dynamic equation, the control system, respectively, is not given via an additive white noise term, but by means of possibly time-dependent random parameters, in the following,

solutions of random differential equations, hence, solutions of *differential equations with random parameters*, are defined in the trajectory-wise sense [8].

In case of a *discrete* or *discretized* probability distribution of the random elements, model parameters, respectively, that is,

$$\begin{aligned}\Omega &= \{\omega_1, \omega_2, \dots, \omega_\varrho\}, P(\omega = \omega_j) = \alpha_j > 0, \\ j &= 1, \dots, \varrho, \sum_{j=1}^{\varrho} \alpha_j = 1,\end{aligned}$$

we can *redefine* Equations (1a) and (1b) by

$$\dot{z}(t) = g(t, z(t), u(t)), t_0 \leq t \leq t_f, \quad (1c')$$

$$z(0) = z_0 \quad (1d')$$

with the vectors and vector functions

$$\begin{aligned}z(t) &:= (z(t, \omega_j))_{j=1, \dots, \varrho}, \\ z_0 &:= (z_0(\omega_j))_{j=1, \dots, \varrho} \\ g(t, z, u) &:= (g(t, \omega_j, z(j), u))_{j=1, \dots, \varrho}, \\ z &:= (z(j))_{j=1, \dots, \varrho} \in \mathbb{R}^{q\varrho}.\end{aligned}$$

Hence, in this case, Equations (1c) and (1d) represent again an *ordinary* system of first order differential equations for the $q\varrho m$ unknown functions

$$z_{ij} = z_i(t, \omega_j), i = 1, \dots, m, j = 1, \dots, \varrho.$$

Results on the existence and uniqueness of solutions of the systems (1a), (1b) and (1c), (1d) and their dependence on inputs ξ can be found in Ref. 9.

Also in the general case we consider a *parametric approach* (solution in the point-wise sense). This means that for each random element $\omega \in \Omega$, Equations (1a) and (1b) is interpreted as a system of ordinary first-order differential equations with the initial values $z_0(\omega)$. Hence, we assume that to each deterministic control $u(\cdot) \in U$ and each random element $\omega \in \Omega$ there exists a unique measurable solution

$$z(t, \omega) = z^u(t, \omega) = z(t, \omega, u(\cdot)), t_0 \leq t \leq t_f, \quad (2a)$$

denoted, with the *solution or input-output operator* S , also by

$$z(\cdot, \omega) = S(\omega, u(\cdot))(\cdot), z(\cdot, \omega) \in PC_1^m[t_0, t_f], \quad (2b)$$

of the integral equation

$$\begin{aligned}z(t) &= z_0(\omega) + \int_{t_0}^t g(s, \omega, z(s), u(s)) ds, \\ t_0 &\leq t \leq t_f;\end{aligned} \quad (2c)$$

moreover, it is assumed that $(t, \omega) \rightarrow S(\omega, u(\cdot))(t)$ is measurable. Obviously, the integral Equation (2c) is the equivalent integral version of the initial value problems (1a) and (1b).

Parametric Representation of the Random Differential Equation. In order to justify the above assumptions, let the random input into the control system, as for example, a time-varying disturbance of the system, be described by an r -dimensional stochastic process $\theta = \theta(t, \omega)$, such that the sample functions $\theta(\cdot, \omega)$ are continuous with probability one. Furthermore, let

$$\tilde{g} : [t_0, t_f] \times \mathbb{R}^r \times \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^m$$

be a continuous m -vector function having continuous Jacobians $D_\theta \tilde{g}, D_z \tilde{g}, D_u \tilde{g}$ with respect to θ, z, u . Now consider the *parametric case* that the function g of the process differential equations (1a) and (1b) is given by

$$g(t, \omega, z, u) := \tilde{g}(t, \theta(t, \omega), z, u),$$

$(t, \omega, z, u) \in [t_0, t_f] \times \Omega \times \mathbb{R}^m \times \mathbb{R}^n$. Furthermore, put $U = C^m[t_0, t_f], Z = C_0^m[t_0, t_f]$, and $\Theta = C_0^r[t_0, t_f]$, where $C_0^v[t_0, t_f]$ denotes the space of all continuous functions of $[t_0, t_f]$ into $\mathbb{R}^v$ normed by the supremum norm $\|\cdot\|_\infty$. By our assumption we have $\theta(\cdot, \omega) \in \Theta$ a.s. Define

$$\Xi = \mathbb{R}^m \times \Theta \times U;$$

Ξ is the space of *model and environmental parameters* of the dynamic system. Hence, Ξ may also be interpreted as the *total space*

of inputs $\xi = (z_0, \theta(\cdot), u(\cdot))$ into the dynamic equation, consisting of the random initial state z_0 , the random input function $\theta = \theta(t, \omega)$ and the control input $u = u(t)$. Now, let the mapping $\tau : \Xi \times Z \rightarrow Z$ related to the plant equations (1a) and (1b) or (2c) be defined by

$$\begin{aligned} \tau(\xi, z(\cdot))(t) \\ = z(t) - \left(z_0 + \int_{t_0}^t \tilde{g}(s, \theta(s), z(s), u(s)) ds \right), \\ t_0 \leq t \leq t_f. \end{aligned} \quad (2d)$$

Obviously, the initial value problems (1a) and (1b) or its integral form (2c) can be represented by the operator equation

$$\tau(\xi, z(\cdot)) = 0. \quad (2e)$$

Operators of the type (2d) are well studied [10]. It is known that τ is *continuously Fréchet (F)-differentiable* [9,10]. Note that the F-differential is just a generalization of the derivatives (Jacobians) of mappings between finite-dimensional spaces to mappings between arbitrary-normed spaces. Thus, the F-derivative $D\tau$ of τ at a certain point $(\bar{\xi}, \bar{z}(\cdot))$ is given by

$$\begin{aligned} \left(D\tau(\bar{\xi}, \bar{z}(\cdot)) \cdot (\xi, z(\cdot)) \right)(t) = z(t) \\ - \left(z_0 + \int_{t_0}^t D_z \tilde{g}(s, \bar{\theta}(s), \bar{z}(s), \bar{u}(s)) z(s) ds \right. \\ + \int_{t_0}^t D_\theta \tilde{g}(s, \bar{\theta}(s), \bar{z}(s), \bar{u}(s)) \theta(s) ds \\ + \left. \int_{t_0}^t D_u \tilde{g}(s, \bar{\theta}(s), \bar{z}(s), \bar{u}(s)) u(s) ds \right), \\ t_0 \leq t \leq t_f, \end{aligned} \quad (2f)$$

where $\bar{\xi} = (\bar{z}_0, \bar{\theta}(\cdot), \bar{u}(\cdot))$ and $\xi = (z_0, \theta(\cdot), u(\cdot))$. Obviously, the F-derivative $D\tau$ of τ , here described in the form of a directional derivative of τ at $(\bar{\xi}, \bar{z}(\cdot))$ in the direction of $(\xi, z(\cdot))$, is represented by means of the Jacobians $D_z \tilde{g}$, $D_\theta \tilde{g}$, $D_u \tilde{g}$ of $\tilde{g}$. Especially, for the F-derivative of τ with respect to $z(\cdot)$ we

find

$$\begin{aligned} \left(D_z \tau(\bar{\xi}, \bar{z}(\cdot)) \cdot z(\cdot) \right)(t) \\ = z(t) - \int_{t_0}^t D_z \tilde{g}(s, \bar{\theta}(s), \bar{z}(s), \bar{u}(s)) z(s) ds, \\ t_0 \leq t \leq t_f. \end{aligned}$$

The related integral equation

$$D_z \tau(\bar{\xi}, \bar{z}(\cdot)) \cdot z(\cdot) = y(\cdot), y(\cdot) \in Z,$$

is a linear Volterra integral equation. By our assumptions this equation has a unique solution $z(\cdot) \in Z$ [11]. Therefore, $D_z \tau(\bar{\xi}, \bar{z}(\cdot))$ is a linear, continuous one-to-one map from Z onto Z . Hence, its inverse exists. This property enables then the application of the implicit function theorem [9,10] to the equation $\tau(\xi, S(\xi)) = 0$, where the following assumptions are still needed.

Assumption 1. Let $(\bar{\xi}, \bar{z}(\cdot)) \in \Xi \times Z$ be a pair of a reference parameter vector and a reference trajectory of Equations (1a) and (1b) such that $\tau(\bar{\xi}, \bar{z}(\cdot)) = 0$. Hence, $\bar{z}(\cdot)$ is the solution of

$$\dot{z}(t) = \tilde{g}(t, \bar{\theta}(t), z(t), \bar{u}(t)), t_0 \leq t \leq t_f, \quad (3a)$$

$$z(t_0) = \bar{z}_0, \quad (3b)$$

where $\bar{\xi} = (\bar{z}_0, \bar{\theta}(\cdot), \bar{u}(\cdot))$.

This assumption means that for the given reference or nominal system input $\bar{\xi}$, the basic system differential equations (1a) and (1b) has the solution $\bar{z} = \bar{z}(t)$. On the basis of this and the former assumptions on $\tilde{g}$, results on the existence and uniqueness of solutions of the system of differential equations for all inputs ξ in a certain neighborhood of $\bar{\xi}$ follow now from the fixed point theorem and the method of successive approximation (Picard–Lindelöf) [9]:

Lemma 1 [Existence and uniqueness of solutions of the system differential equation]. Under Assumption 1 there exists an open neighborhood $V(\bar{\xi})$ of $\bar{\xi}$, and a unique

continuous mapping $S : V(\bar{\xi}) \rightarrow Z$ such that (a) $S(\bar{\xi}) = \bar{z}(\cdot)$; (b) $\tau(\xi, S(\xi)) = 0$ for each $\xi \in V(\bar{\xi})$, that is, $z(t) := S(\xi)(t)$, $t_0 \leq t \leq t_f$, is the unique solution of

$$z(t) = z_0 + \int_{t_0}^t \tilde{g}(s, \theta(s), z(s), u(s)) ds, \quad t_0 \leq t \leq t_f, \quad (3c)$$

where $\xi = (z_0, \theta(\cdot), u(\cdot))$.

Application of the implicit function theorem [9] to the Equation (2e) yields then this result.

Lemma 2 [Differentiation of trajectories with respect to system inputs]. Under the assumptions mentioned in the above lemma, the solution $z(t) = S(\xi)(t)$ of the system differential equations (1a) and (1b) is continuously F -differentiable on $V(\bar{\xi})$, and the partial F -derivative $\zeta(\cdot) = \zeta_{u,h} = D_u S(\xi)h(\cdot)$ with respect to the control $u(\cdot)$ and in the direction of a function $h(\cdot) \in U$ satisfies the integral equation

$$\begin{aligned} \zeta(t) &= \int_{t_0}^t D_z \tilde{g}(s, \theta(s), S(\xi)(s), u(s)) \zeta(s) ds \\ &= \int_{t_0}^t D_u \tilde{g}(s, \theta(s), S(\xi)(s), u(s)) h(s) ds, \\ t_0 \leq t \leq t_f, \end{aligned} \quad (3d)$$

where $\xi = (z_0, \theta(\cdot), u(\cdot))$.

Remark 1. Taking the time derivative of Equation (3d) shows that this integral equation is equivalent to the so-called *perturbation equation* needed in the section titled “Computation of Directional Derivatives”.

Objective Function

In order to obtain optimal controls $u = u(t)$, $t_0 \leq t \leq t_f$, being *robust*, that is, most insensitive with respect to stochastic variations of the model and environmental parameters of the process, the cost criterion

$F = F(u(\cdot))$ related to the controlled process $z = z(t, \omega)$ is defined by the expectation

$$F(u(\cdot)) := Ef(\omega, S(\omega, u(\cdot)), u(\cdot)). \quad (4a)$$

Here, $E = E(\cdot | \mathcal{A}_{t_0})$, denotes the conditional expectation given the information $\mathcal{A}_{t_0}$ about the control process up to the considered starting time point t_0 . Moreover, $f = f(\omega, z(\cdot), u(\cdot))$ denotes the stochastic total costs arising along the trajectory $z = z(t, \omega)$ and at the terminal point $z_f = z(t_f, \omega)$ [1,5]. Hence,

$$\begin{aligned} f(\omega, z(\cdot), u(\cdot)) &:= \int_{t_0}^{t_f} L(t, \omega, z(t), u(t)) dt \\ &+ G(t_f, \omega, z(t_f)), \end{aligned} \quad (4b)$$

$z(\cdot) \in Z, u(\cdot) \in U$. Here,

$$\begin{aligned} L : [t_0, t_f] \times \Omega \times \mathbb{R}^m \times \mathbb{R}^n &\rightarrow \mathbb{R}, \\ G : [t_0, t_f] \times \Omega \times \mathbb{R}^m &\rightarrow \mathbb{R} \end{aligned}$$

are given measurable cost functions. We suppose that $L(t, \omega, \cdot, \cdot)$ and $G(t, \omega, \cdot)$ are convex functions for each $(t, \omega) \in [t_0, t_f] \times \Omega$, having continuous partial derivatives $\nabla_z L(\cdot, \omega, \cdot, \cdot)$, $\nabla_u L(\cdot, \omega, \cdot, \cdot)$, $\nabla_z G(\cdot, \omega, \cdot)$. Note that in this case

$$\begin{aligned} (z(\cdot), u(\cdot)) &\rightarrow \int_{t_0}^{t_f} L(t, \omega, z(t), u(t)) dt \\ &+ G(t_f, \omega, z(t_f)) \end{aligned} \quad (4c)$$

is a convex function on $Z \times U$ for each $\omega \in \Omega$. Moreover, assume that the expectation $F(u(\cdot))$ exists and is finite for each admissible control $u(\cdot) \in D$.

Example 1. In the case of active structural control or for optimal regulator design of robots, cf. Refs 2,4, and 12, the total cost function f is given by defining the individual cost functions L and G as follows:

$$L(t, \omega, z, u) := z^T Q(\omega)z + u^T R(\omega)u \quad (4d)$$

$$G(t_f, \omega, z) := G(z). \quad (4e)$$

Here, Q and R , respectively, are certain positive (semi)definite $m \times m, n \times n$ matrices which may depend also on (t, ω) . Moreover, the terminal cost function G depends then only on z . For example, in case of end point control, the cost function $G(z)$ is given by

$$G(z) = (z - z_f)^T G_f (z - z_f), \quad (4f)$$

with a certain desired, possibly random terminal point z_f and a positive (semi)definite weight matrix G_f .

In this article, we now present methods for the approximate solution for the following minimum expected cost problems: Find a control $u(\cdot)$ solving

$$\min F(u(\cdot)) \text{ s.t. } u(\cdot) \in D. \quad (5)$$

Problem (5) is equivalent to $(E = E(\cdot | \mathcal{A}_{t_0}))$ [13,14],

$$\begin{aligned} \min E \left(\int_{t_0}^{t_f} L(t, \omega, z(t), u(t)) dt \right. \\ \left. + G(t_f, \omega, z(t_f)) | \mathcal{A}_{t_0} \right) \end{aligned} \quad (6a)$$

s.t.

$$\dot{z}(t) = g(t, \omega, z(t), u(t)), t_0 \leq t \leq t_f, \text{ a.s.} \quad (6b)$$

$$z(t_0, \omega) = z_0(\omega), \text{ a.s.} \quad (6c)$$

$$u(\cdot) \in D. \quad (6d)$$

In order to get *robust* optimal controls/regulators, random variations of the model, environmental and applied load parameters, as well as random variations of initial values are incorporated into the optimal control design process by means of stochastic optimization methods. Hence, we use the following notion.

Definition 1. An optimal solution $u^*(\cdot)$ of the optimal control problem (5) or (6a)–(6d) is called a *stochastic optimal control*.

Control Laws

In optimal control of dynamic systems the following types of *control laws* are mostly considered:

1. *Open-Loop Control.* Here, the control function $u = u(t)$ is a *deterministic* function depending only on the given (*a priori*) information $\mathcal{A}_{t_0} \subset \mathcal{A}$ about the system, the model parameters, the initial values, respectively, available at the starting or reference time point t_0 . Hence, for the optimal selection of optimal (OL) controls

$$u(t) = u(t; (t_0, \mathcal{A}_{t_0})), t \geq t_0, \quad (7a)$$

we get optimal control problems of type (5).

2. *Closed-Loop Control.* In this case, the control function $u = u(t)$ is a **stochastic** function

$$u = u(t, \omega) = u(t, \mathcal{A}_t), t \geq t_0 \quad (7b)$$

depending on time t and the total information $\mathcal{A}_t$ on the system available up to time t . Especially $\mathcal{A}_t$ may contain information about the state $z(t) = z(t, \omega)$ up to time t .

3. *Open-Loop Feedback Control.* In this combination of OL and CL control, at each time point $t_b \geq t_0$, first the optimal OL control function for the remaining time interval $t_b \leq t \leq t_b$ is computed, hence,

$$u_{[t_b, t_f]}(t) = u(t; (t_b, \mathcal{A}_{t_b})), t \geq t_b. \quad (7c)$$

Evaluating then the obtained optimal OL control $u_{[t_b, t_f]}(\cdot)$ only at $t_b = t$, the OLF control is defined by

$$u(t, \omega) := u_{[t, t_f]}(t) = u(t; (t, \mathcal{A}_t)), t \geq t_0, \quad (7d)$$

that is, the OL control $u_{[t, t_f]}(s)$, $s \geq t$, for the remaining time interval $[t, t_f]$ is used only for $s = t$. Stochastic optimal (OLF) controls are therefore obtained



Figure 1. Remaining time interval.

by solving again control problems of the type (5). Note that OLF control is the basic tool in the so-called *model predictive control* [3], which is often applied in practice.

Remark 2. In case of OLF control, the initial time point t_0 in the expected total cost function is replaced by the intermediate starting time point t_b of the remaining time intervals $[t_b, t_f]$, see Fig. 1.

Discrete-Time Controllers. In practice, fast, light, accurate, versatile, and economical control laws are often implemented in digital form. Thus, the design of digital or discrete-time controllers is an important problem for practical applications. As discussed in Ref. 15, there are two important design approaches for digital control: direct discrete-time controller design, which is not considered in the present work, and controller redesign of continuous-time controllers as discussed in this article. This second method is very appealing since any technique for continuous-time control can be used to construct an appropriate continuous-time controller, which is then discretized and implemented on a corresponding control device.

CONVEX APPROXIMATION

First, we observe that problem (5) is in general a nonconvex optimization problem [10]. Since for convex (deterministic) optimization problems there is a well-established theory, we consider the approximation of the original problem (5) by suitable convex problems. Initially, we describe below the basic steps of this procedure [30]. Moreover, the construction of descent directions for F at a reference

control and the derivation of necessary optimality conditions for optimal solutions of the control problem (5) are considered.

Let $v(\cdot) \in D$ be an arbitrary, but fixed admissible initial or reference control and assume, see Lemma 1.

Assumption 2. $S(\omega, \cdot)$ is F -differentiable at $v(\cdot)$ for each $\omega \in \Omega$.

Denoting the F -derivative of $S(\omega, \cdot)$ at $v(\cdot)$ by $DS(\omega, v(\cdot))$, the expected total cost function $F = F(u(\cdot))$ is replaced now, see Fig. 2, by

$$F_{v(\cdot)}(u(\cdot)) := Ef(\omega, S(\omega, v(\cdot)) + DS(\omega, v(\cdot)) \times (u(\cdot) - v(\cdot)), u(\cdot)), \quad (8)$$

where $u(\cdot) \in U$, assuming that $F_{v(\cdot)}(u(\cdot)) \in \mathbb{R}$ for all pairs $(u(\cdot), v(\cdot)) \in D \times D$.

Then, problem (5) is approximated [13] by

$$\min F_{v(\cdot)}(u(\cdot)) \text{ s.t. } u(\cdot) \in D. \quad ((5)_{v(\cdot)})$$

Next, a decisive property of this problem is shown:

Lemma 3. *Problem $(5)_{v(\cdot)}$ is a convex optimization problem.*

Proof. According to the section titled “Objective Function”, function (4c) is convex. The assertion now follows from the linearity of the F -differential.

Remark 3. The approximate $F_{v(\cdot)}$ of F is obtained from Equations (4a) and (4b) by means of linearization of the input–output map $S = S(\omega, u(\cdot))$ with respect to the control $u(\cdot)$ at the nominal or reference control $v(\cdot)$. Hence, the convex approximation $F_{v(\cdot)}$ of F follows by *inner linearization* of the control problem with respect to the control $u(\cdot)$ at $v(\cdot)$.

Let $\phi'_+(x; y)$ denote the directional derivative at $x \in X$ and in direction $y \in X$ of a convex

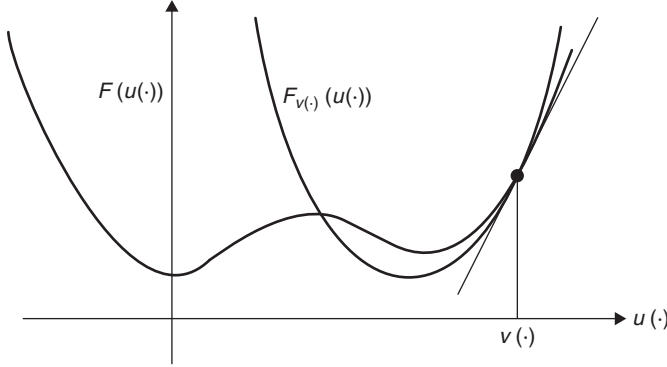


Figure 2. Convex approximation.

function $\phi : X \rightarrow \mathbb{R}$, where X is a linear space [16]. With this definition, from Lemma 3 we may infer that for all $u(\cdot), v(\cdot) \in D$ and $h(\cdot) \in U$ the directional derivative of $F_{v(\cdot)}$ is given by

$$\begin{aligned} F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ = Ef'_+(\omega, S(\omega, v(\cdot)) + DS(\omega, v(\cdot))(u(\cdot) - v(\cdot)), \\ u(\cdot); DS(\omega, v(\cdot))h(\cdot), h(\cdot)). \end{aligned} \quad (9)$$

A solution $\bar{u}(\cdot) \in D$ of Equation (5)_{v(·)} is then characterized by

$$F'_{v(\cdot)+}(\bar{u}(\cdot); u(\cdot) - \bar{u}(\cdot)) \geq 0 \text{ for all } u(\cdot) \in D. \quad (10)$$

For each $v(\cdot) \in D$, let $M(v(\cdot))$ be the set of solutions of problem (5)_{v(·)}.

Assumption 3. $M(v(\cdot)) \neq \emptyset$ for each $v(\cdot) \in D$.

The relation between our original problem (5) and the family of its approximates (5)_{v(·)}, $v(\cdot) \in D$, is described by the following conditions.

Theorem 1. Optimality conditions. Suppose that

$$F'_+(\bar{u}(\cdot); h(\cdot)) = F'_{v(\cdot)+}(\bar{u}(\cdot); h(\cdot)) \quad (11)$$

for each $v(\cdot) \in D$ and $h(\cdot) \in U$. Then

- (a) If $\bar{u}(\cdot)$ is an optimal control, then $\bar{u}(\cdot) \in M(\bar{u}(\cdot))$, that is, $\bar{u}(\cdot)$ is a solution of (5) _{$\bar{u}(\cdot)$} .
- (b) If problem (5) is convex, then $\bar{u}(\cdot)$ is an optimal control if and only if $\bar{u}(\cdot) \in M(\bar{u}(\cdot))$.

Sufficient conditions for the important Condition (11) can be found [14] as follows. By Equations (4a) and (4b), and (9) we have

$$\begin{aligned} F'_{v(\cdot)+}(v(\cdot); h(\cdot)) \\ = Ef'_+(\omega, S(\omega, v(\cdot)), v(\cdot); DS(\omega, v(\cdot))h(\cdot), h(\cdot)) \\ = E \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(f(\omega, S(\omega, v(\cdot) + \varepsilon h(\cdot)), v(\cdot) + \varepsilon h(\cdot)) \right. \\ \left. - f(\omega, S(\omega, v(\cdot)), v(\cdot)) \right). \end{aligned} \quad (12)$$

If the function $f = f(\omega, z(\cdot), u(\cdot))$ is (F-) differentiable with respect to $(z(\cdot), u(\cdot))$, the above Equation (12) can also be obtained by the chain rule [9,10]. On the basis of Equation (12), sufficient conditions for Equation (11) can then be derived as follows.

Lemma 4. Condition (11) in Theorem 1 holds if and only if the expectation operator “E” and the limit process “ $\lim_{\varepsilon \downarrow 0}$ ” in Equation (12) may be interchanged.

The interchangeability needed in Lemma 4 holds, for example, if $\sup \left\{ \|DS(\omega, v(\cdot) + \right.$

$\varepsilon h(\cdot) \| : 0 \leq \varepsilon \leq 1 \}$ is bounded with probability one, and the convex function $f(\omega, \cdot, \cdot)$ satisfies a Lipschitz condition

$$\begin{aligned} & \left| f(\omega, z(\cdot), u(\cdot)) - f(\omega, \bar{z}(\cdot), \bar{u}(\cdot)) \right| \\ & \leq \gamma(\omega) \left\| (z(\cdot), u(\cdot)) - (\bar{z}(\cdot), \bar{u}(\cdot)) \right\|_{\infty} \end{aligned}$$

on a set $Q \subset Z \times U$ containing all vectors $(S(\omega, v(\cdot) + \varepsilon h(\cdot)), v(\cdot) + \varepsilon h(\cdot))$, $0 \leq \varepsilon \leq 1$, where $E\gamma(\omega) < +\infty$. Further conditions were derived in Ref. 14.

Admissible descent directions can be constructed by applying this result.

Lemma 5.

(a) If $\bar{u}(\cdot) \notin M(\bar{u}(\cdot))$ for a control $\bar{u}(\cdot) \in D$, then

$$\begin{aligned} F_{\bar{u}(\cdot)}(u(\cdot)) & < F_{\bar{u}(\cdot)}(\bar{u}(\cdot)) \\ & = F(\bar{u}(\cdot)) \text{ for each } u(\cdot) \in M(\bar{u}(\cdot)) \end{aligned}$$

(b) Let the controls $u(\cdot), v(\cdot) \in D$ be related such that

$$F_{u(\cdot)}(v(\cdot)) < F_{u(\cdot)}(u(\cdot)).$$

Provided that Equation (11) holds for the pair $(u(\cdot), h(\cdot))$, $h(\cdot) = v(\cdot) - u(\cdot)$, then $h(\cdot)$ is an admissible direction of decrease for F at $u(\cdot)$, that is, we have $F(u(\cdot) + \varepsilon h(\cdot)) < F(u(\cdot))$ and $u(\cdot) + \varepsilon h(\cdot) \in D$ on a suitable interval $0 < \varepsilon < \bar{\varepsilon}$.

Proof. See Refs 13 and 14.

Points, controls, and so on, fulfilling at least the necessary optimality conditions of a certain optimization problem are candidates for optimal solutions. Hence, due to the relations shown in Theorem 1, the following notion is used [10,17]:

Definition 2. A control $u(\cdot) \in D$ fulfilling the necessary optimality condition $u(\cdot) \in M(u(\cdot))$ is called a stationary control.

Remark 4. Under the rather weak assumptions in Theorem 1, an optimal control is also stationary, and in the case of a convex problem (5) the two concepts coincide. Hence, *stationary controls are candidates for stochastic optimal controls*. As an appropriate approximate for a stochastic optimal control, we will determine therefore stationary controls [10,18]. Note that stationary admissible decisions may be constructed numerically by means of conditional gradient-type algorithms [17,19].

In order to apply now, the optimality conditions of Theorem 1, hence, to construct stationary controls, according to condition (10) we need the directional derivative $F'_{v(\cdot)+}$ of the convex approximate $F_{v(\cdot)}$ of the expected total cost function F of the control problem.

COMPUTATION OF DIRECTIONAL DERIVATIVES

According to Definition 2 of a stationary control and characterization (10) of an optimal solution of Equation (5)_{v(·)}, we first have to determine the directional derivative $F'_{v(\cdot)+}$. On the basis of the justification in the section titled “Stochastic Control Systems”, we assume again that the solution $z(t, \omega) = S(\omega, u(\cdot))(t)$ of Equation (2c) is measurable in (t, ω) for each $u(\cdot) \in D$, $u(\cdot) \rightarrow S(\omega, u(\cdot))$ is continuously differentiable on D for each $\omega \in \Omega$. Furthermore, we suppose that the F -differential $\zeta(t, \omega) = (D_u S(\omega, u(\cdot))h(\cdot))(t)$, $h(\cdot) \in U$, is measurable in (t, ω) , and is given according to Equations (3a)–(3d) by the *linear perturbation equation*

$$\begin{aligned} \dot{\zeta}(t) & = A(t, \omega, u(\cdot))\zeta(t) + B(t, \omega, u(\cdot))h(t), \\ t_0 & \leq t \leq t_f, \omega \in \Omega \end{aligned} \quad (13a)$$

$$\zeta(t_0) = 0 \quad (13b)$$

with the Jacobians

$$A(t, \omega, u(\cdot)) := D_z g(t, \omega, z^u(t, \omega), u(t)) \quad (13c)$$

$$B(t, \omega, u(\cdot)) := D_u g(t, \omega, z^u(t, \omega), u(t)) \quad (13d)$$

and $z^u(t, \omega)$ defined by

$$z^u(t, \omega) := S(\omega, u(\cdot))(t), \quad (13e)$$

$(t, \omega, u(\cdot)) \in [t_0, t_f] \times \Omega \times U$. The solution $\zeta = \zeta(t, \omega)$ of Equations (13a) and (13b) is also denoted by

$$\begin{aligned} \zeta(t, \omega) &= \zeta_{u,h}(t, \omega) \\ &:= \left(D_u S(\omega, u(\cdot)) h(\cdot) \right)(t), h(\cdot) \in U. \end{aligned} \quad (13f)$$

This means that the approximate $(5)_{v(\cdot)}$ of Equation (5) has the following explicit form:

$$\begin{aligned} \min E \Big(& \int_{t_0}^{t_f} L(t, \omega, z^v(t, \omega) + \zeta(t, \omega), u(t)) dt + \\ & + G(t_f, \omega, z^v(t_f, \omega) + \zeta(t_f, \omega)) \Big) \end{aligned} \quad (14a)$$

s.t.

$$\begin{aligned} \dot{\zeta}(t, \omega) &= A(t, \omega, v(\cdot)) \zeta(t, \omega) \\ &+ B(t, \omega, v(\cdot)) (u(t) - v(t)) \text{ a.s.} \end{aligned} \quad (14b)$$

$$\zeta(t_0, \omega) = 0 \text{ a.s.} \quad (14c)$$

$$u(\cdot) \in D. \quad (14d)$$

With the convexity assumptions in the section titled “Objective Function”, Lemma 3 yields that problem (14a)–(14d) is a convex stochastic control problem, with a *linear* dynamic equation. For the subsequent analysis of the stochastic control problem, we need a representation of the directional derivative $F'_{v(\cdot)+}(u(\cdot); h(\cdot))$ by a scalar product of the increment function $h(\cdot)$ with a certain deterministic vector function $q = q(t)$.

From representation (9) of the directional derivative $F'_{v(\cdot)+}$ of the convex approximate $F_{v(\cdot)}$ of F and definition (4b) of $f = f(\omega, z(\cdot), u(\cdot))$, with Equation (13f) we

obtain

$$\begin{aligned} & F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ &= E \left(\int_{t_0}^{t_f} \left(\nabla_z L(t, \omega, z^v(t, \omega) + \zeta_{v,u-v}(t, \omega), u(t)) \right)^T \right. \\ &\quad \times \zeta_{v,h}(t, \omega) \\ &\quad + \nabla_u L(t, \omega, z^v(t, \omega) + \zeta_{v,u-v}(t, \omega), u(t))^T h(t) dt \\ &\quad + \nabla_z G(t_f, \omega, z^v(t_f, \omega) + \zeta_{v,u-v}(t_f, \omega))^T \\ &\quad \left. \times \zeta_{v,h}(t_f, \omega) \right). \end{aligned} \quad (15)$$

Defining the gradients

$$\begin{aligned} a(t, \omega, v(\cdot), u(\cdot)) &:= \nabla_z L(t, \omega, z^v(t, \omega) + \zeta_{v,u-v}(t, \omega), u(t)) \end{aligned} \quad (16a)$$

$$\begin{aligned} b(t, \omega, v(\cdot), u(\cdot)) &:= \nabla_u L(t, \omega, z^v(t, \omega) + \zeta_{v,u-v}(t, \omega), u(t)) \end{aligned} \quad (16b)$$

$$\begin{aligned} c(t_f, \omega, v(\cdot), u(\cdot)) &:= \nabla_z G(t_f, \omega, z^v(t_f, \omega) + \zeta_{v,u-v}(t_f, \omega)) \end{aligned} \quad (16c)$$

the directional derivative $F'_{v(\cdot)+}$ can be represented by

$$\begin{aligned} & F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ &= E \left(\int_{t_0}^{t_f} \left(a(t, \omega, v(\cdot), u(\cdot)) \right)^T \zeta_{v,h}(t, \omega) \right. \\ &\quad + b(t, \omega, v(\cdot), u(\cdot))^T h(t) dt \\ &\quad \left. + c(t_f, \omega, v(\cdot), u(\cdot))^T \zeta_{v,h}(t_f, \omega) \right). \end{aligned} \quad (17a)$$

Integrating Equation (13a) over $[t_0, t_f]$, for each ω we get

$$\begin{aligned} \zeta_{v,h}(t_f, \omega) &= \int_{t_0}^{t_f} \dot{\zeta}_{v,h}(t, \omega) dt \\ &= \int_{t_0}^{t_f} \left(A(t, \omega, v(\cdot)) \zeta_{v,h}(t, \omega) \right. \\ &\quad \left. + B(t, \omega, v(\cdot)) h(t) \right) dt. \end{aligned} \quad (17b)$$

Putting Equation (17b) into Equation (17a), we find

$$\begin{aligned} & F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ &= E \left(\int_{t_0}^{t_f} \tilde{a}(t, \omega, v(\cdot), u(\cdot))^T \zeta_{v,h}(t, \omega) dt \right. \\ & \quad \left. + \int_{t_0}^{t_f} \tilde{b}(t, \omega, v(\cdot), u(\cdot))^T h(t) dt \right), \end{aligned} \quad (17c)$$

where

$$\begin{aligned} & \tilde{a}(t, \omega, v(\cdot), u(\cdot)) \\ &:= a(t, \omega, v(\cdot), u(\cdot)) + A(t, \omega, v(\cdot))^T \\ & \quad \times c(t_f, \omega, v(\cdot), u(\cdot)), \end{aligned} \quad (17d)$$

$$\begin{aligned} & \tilde{b}(t, \omega, v(\cdot), u(\cdot)) \\ &:= b(t, \omega, v(\cdot), u(\cdot)) + B(t, \omega, v(\cdot))^T \\ & \quad \times c(t_f, \omega, v(\cdot), u(\cdot)). \end{aligned} \quad (17e)$$

In order to transform the first integral in Equation (17c) into the form of the second integral in Equation (17c), we introduce the m -vector function

$$\lambda = \lambda^{v,u}(t, \omega)$$

defined by the following integral equation depending on the random parameter ω :

$$\begin{aligned} & \lambda(t) - A(t, \omega, v(\cdot))^T \int_t^{t_f} \lambda(s) ds \\ &= \tilde{a}(t, \omega, v(\cdot), u(\cdot)). \end{aligned} \quad (18)$$

Under the present assumptions, this *Volterra integral equation* has [11] a unique measurable solution $(t, \omega) \rightarrow \lambda^{v,u}(t, \omega)$. By means of Equation (18) we obtain

$$\begin{aligned} & \int_{t_0}^{t_f} \tilde{a}(t, \omega, v(\cdot), u(\cdot))^T \zeta_{v,h}(t, \omega) dt \\ &= \int_{t_0}^{t_f} \lambda(s)^T \left(\zeta_{v,h}(s, \omega) - \int_{t_0}^s A(t, \omega, v(\cdot)) \right. \\ & \quad \left. \times \zeta_{v,h}(t, \omega) dt \right) ds. \end{aligned} \quad (19a)$$

Now, using again the perturbation equations (13a) and (13b), from Equation (19a) we get

$$\begin{aligned} & \int_{t_0}^{t_f} \tilde{a}(t, \omega, v(\cdot), u(\cdot))^T \zeta_{v,h}(t, \omega) dt \\ &= \int_{t_0}^{t_f} \left(B(t, \omega, v(\cdot))^T \int_t^{t_f} \lambda(s) ds \right)^T h(t) dt. \end{aligned} \quad (19b)$$

Inserting Equation (19b) into Equation (17c), we have

$$\begin{aligned} & F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ &= E \left(\int_{t_0}^{t_f} \left(B(t, \omega, v(\cdot))^T \int_t^{t_f} \lambda(s) ds \right. \right. \\ & \quad \left. \left. + \tilde{b}(t, \omega, v(\cdot), u(\cdot)) \right)^T h(t) dt \right). \end{aligned} \quad (20)$$

By means of Equation (17d), the integral equation (18) may be written as

$$\begin{aligned} & \lambda(t) - A(t, \omega, v(\cdot))^T \left(c(t_f, \omega, v(\cdot), u(\cdot)) + \int_t^{t_f} \lambda(s) ds \right) \\ &= a(t, \omega, v(\cdot), u(\cdot)). \end{aligned} \quad (21)$$

According to Equation (21), we now define the m -vector function

$$\begin{aligned} y &= y^{v,u}(t, \omega) := c(t_f, \omega, v(\cdot), u(\cdot)) \\ & \quad + \int_t^{t_f} \lambda_{v,u}(s, \omega) ds. \end{aligned} \quad (22)$$

Obviously, $y = y^{v,u}(t, \omega)$ solves the system of differential equations

$$\begin{aligned} & \dot{y}(t) = -A(t, \omega, v(\cdot))^T y(t) - a(t, \omega, v(\cdot), u(\cdot)), \\ & t_0 \leq t \leq t_f, \end{aligned} \quad (23a)$$

$$y(t_f) = c(t_f, \omega, v(\cdot), u(\cdot)). \quad (23b)$$

Systems (23a) and (23b) are called the *adjoint differential equations* related to the perturbation equations (13a) and (13b).

Using Equation (22), from Equations (17e) and (20) we now obtain

$$\begin{aligned} F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ = E \left(\int_{t_0}^{t_f} \left(B(t, \omega, v(\cdot))^T y^{v,u}(t, \omega) \right. \right. \\ \left. \left. + b(t, \omega, v(\cdot), u(\cdot)) \right)^T h(t) dt \right). \end{aligned}$$

Summarizing the above transformations, we have the following result.

Theorem 2. *Suppose that the adjoint differential equation (23a), (23b) has a measurable solution $(\omega, t) \rightarrow y^{v,u}(t, \omega)$. Then,*

$$\begin{aligned} F'_{v(\cdot)+}(u(\cdot); h(\cdot)) \\ = \int_{t_0}^{t_f} E \left(B(t, \omega, v(\cdot))^T y^{v,u}(t, \omega) \right. \\ \left. + b(t, \omega, v(\cdot), u(\cdot)) \right)^T h(t) dt. \end{aligned} \quad (24)$$

Note that $F'_{v(\cdot)+}(u(\cdot); \cdot)$ is also the Gâteaux differential of $F_{v(\cdot)}$ at $u(\cdot)$ [10].

Note that in practice, the domain D of admissible controls is often given by

$$D = \{u(\cdot) \in U : u(t) \in D_t, t_0 \leq t \leq t_f\}, \quad (25)$$

where $D_t \subset \mathbb{R}^n$ is a given convex subset of $\mathbb{R}^n$ for each time t , $t_0 \leq t \leq t_f$.

STATIONARY CONTROLS

We now consider the construction of *stationary controls* of problem (6a)–(6d), hence, controls $\bar{u}(\cdot) \in D$ fulfill at least the necessary optimality condition $\bar{u}(\cdot) \in M(\bar{u}(\cdot))$, see Definition 2. Thus, stationary controls are candidates for stochastic optimal controls. For nonlinear control or optimization problems, finding stationary controls or stationary points is often the only possibility to get a useful approximate optimal solution.

Starting again with formula (24), by means of Equation (11), for an element $v(\cdot) \in D$ we have

$$\begin{aligned} F'_+(v(\cdot); h(\cdot)) &= F'_{v(\cdot)+}(v(\cdot); h(\cdot)) \\ &= \int_{t_0}^{t_f} E \left(B(t, \omega, v(\cdot))^T y^v(t, \omega) \right. \\ &\quad \left. + b(t, \omega, v(\cdot), v(\cdot)) \right)^T h(t) dt, \end{aligned} \quad (26a)$$

where

$$y^v(t, \omega) := y^{v,v}(t, \omega) \quad (26b)$$

fulfills, cf. Equations (23a) and (23b), the adjoint differential equation

$$\begin{aligned} \dot{y}(t, \omega) &= -A(t, \omega, v(\cdot))^T y(t, \omega) \\ &\quad - \nabla_z L(t, \omega, z^v(t, \omega), v(t)) \end{aligned} \quad (27a)$$

$$y(t_f, \omega) = \nabla_z G(t_f, \omega, z^v(t_f, \omega)); \quad (27b)$$

moreover, cf. Equations (13c) and (13d),

$$A(t, \omega, v(\cdot)) = D_z g(t, \omega, z^v(t, \omega), v(t)) \quad (27c)$$

$$B(t, \omega, v(\cdot)) = D_u g(t, \omega, z^v(t, \omega), v(t)) \quad (27d)$$

and

$$b(t, \omega, v(\cdot), v(\cdot)) = \nabla_u L(t, \omega, z^v(t, \omega), v(t)), \quad (27e)$$

where $z^v = z^v(t, \omega)$ solves the dynamic equation

$$\dot{z}(t, \omega) = g(t, \omega, z(t, \omega), v(t)), t_0 \leq t \leq t_f, \quad (27f)$$

$$z(t_0, \omega) = z_0(\omega). \quad (27g)$$

Adapting now the *Hamilton–Jacobi theory* for deterministic optimal control problems [6,20], to optimal control problems under stochastic uncertainty, we now

introduce the *stochastic Hamiltonian*:

$$H(t, \omega, z, y, u) := L(t, \omega, z, u) + y^T g(t, \omega, z, u) \quad (28a)$$

related to the basic control problems (6a)–(6d). From Equations (26a) and (26b), and Equations (27a)–(27g), for the directional derivative F'_+ we get then the representation

$$\begin{aligned} F'_{v(\cdot)+} (v(\cdot); h(\cdot)) &= F'_+ (v(\cdot); h(\cdot)) \\ &= \int_{t_0}^{t_f} E \nabla_u H(t, \omega, z^v(t, \omega), y^v(t, \omega), v(t))^T h(t) dt. \end{aligned} \quad (28b)$$

According to Definition 2, a stationary control of problem (5) is an element $\bar{u}(\cdot) \in D$ such that $\bar{u}(\cdot)$ is an optimal solution of the approximate problem (5) $_{\bar{u}(\cdot)}$. Using the characterization (10), under condition (11), a stationary control can now be characterized by

$$F'_+ (\bar{u}(\cdot); u(\cdot) - \bar{u}(\cdot)) \geq 0 \text{ for all } u(\cdot) \in D.$$

Thus, because of representation (28b), for stationary controls $\bar{u}(\cdot) \in D$ we have the characterization

$$\begin{aligned} &\int_{t_0}^{t_f} E \nabla_u H(t, \omega, z^{\bar{u}}(t, \omega), y^{\bar{u}}(t, \omega), \bar{u}(t))^T \\ &\quad \times (u(t) - \bar{u}(t)) dt \geq 0, u(\cdot) \in D. \end{aligned} \quad (29)$$

In order to show that condition (29) is related to an optimality principle, for given $w(\cdot) \in D$ we introduce the function

$$\begin{aligned} \bar{H}^w(u(\cdot)) &:= \\ &\int_{t_0}^{t_f} EH(t, \omega, z^w(t, \omega), y^w(t, \omega), u(t)) dt, \end{aligned} \quad (30a)$$

and we consider the variational problem

$$\min \bar{H}^w(u(\cdot)) \text{ s.t. } u(\cdot) \in D. \quad (30b)$$

If directional differentiation and integration/expectation in Equation (30a) may be

interchanged, which is assumed in the following, then Equation (29) is a necessary condition for

$$\bar{u}(\cdot) \in \operatorname{argmin}_{u(\cdot) \in D} \bar{H}^w(u(\cdot)). \quad (31)$$

Thus, corresponding to *Pontryagin's minimum principle* [6,10], we have this result.

Theorem 3. *Optimality principle. Suppose that a control $\bar{u}(\cdot) \in D$ fulfills problem (31). Then $\bar{u}(\cdot)$ is a stationary control of Equation (5).*

Canonical System of Differential Equations for Stationary Controls

Assume again that the feasible domain D is given by Equation (25). In order to solve the optimization problems (30a) and (30b), we next replace $z^{\bar{u}}(\cdot, \omega), y^{\bar{u}}(\cdot, \omega)$ by measurable functions $\zeta(\cdot), \eta(\cdot)$ on $(\Omega, \mathcal{A}, P)$. Then, minimizing under the integral sign in Equation (30a), we obtain the finite-dimensional *stochastic optimization problem*

$$\min EH(t, \omega, \zeta(\omega), \eta(\omega), u) \text{ s.t. } u \in D_t \quad (P)_{\zeta, \eta}^t$$

for each $t, t_0 \leq t \leq t_f$. For arbitrary measurable functions $\zeta(\cdot), \eta(\cdot)$ on $(\Omega, \mathcal{A}, P)$, such that the expectation in $(P)_{\zeta, \eta}^t$ exists and is finite, let

$$\tilde{u}^* = \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot)), t_0 \leq t \leq t_f,$$

denote a solution of Equation $(P)_{\zeta, \eta}^t$. As a basic tool in the *Hamilton–Jacobi theory* [6,20], we use this important notion.

Definition 3. The function $\tilde{u}^* = \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot)), t_0 \leq t \leq t_f$, is called a **H-minimal control** of problem (6a)–(6d).

For a given H -minimal control $\tilde{u}^* = \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot))$, which can be interpreted as a *stochastic optimal control law*, we consider the related *canonical (Hamiltonian) two-point boundary value problem with*

random parameters:

$$\dot{z}(t, \omega) = g\left(t, \omega, z(t, \omega), \tilde{u}^*\left(t, z(t, \cdot), y(t, \cdot)\right)\right), \quad t_0 \leq t \leq t_f \quad (32a)$$

$$z(t_0, \omega) = z_0(\omega) \quad (32b)$$

$$\begin{aligned} \dot{y}(t, \omega) = & -D_z g\left(t, \omega, z(t, \omega), \tilde{u}^*\left(t, z(t, \cdot), y(t, \cdot)\right)\right)^T y(t, \omega) \\ & -\nabla_z L\left(t, \omega, z(t, \omega), \tilde{u}^*\left(t, z(t, \cdot), y(t, \cdot)\right)\right), \quad t_0 \leq t \leq t_f \end{aligned} \quad (32c)$$

$$y(t_f, \omega) = \nabla_z G\left(t_f, \omega, z(t_f, \omega)\right). \quad (32d)$$

Remark 5. In case of a discrete distribution $\Omega = \{\omega_1, \dots, \omega_\varrho\}$, $P(\omega = \omega_j) = \alpha_j, j = 1, \dots, \varrho$, for the H -minimal control we have

$$\tilde{u}^* = \tilde{u}^*\left(t, z(t, \omega_1), \dots, z(t, \omega_\varrho), y(t, \omega_1), \dots, y(t, \omega_\varrho)\right).$$

Thus, system (32a)–(32d) is then an ordinary two-point boundary value problems for the 2ϱ unknown functions $z = z(t, \omega_j), y = y(t, \omega_j), j = 1, \dots, \varrho$.

Let $(z^*, y^*) = (z^*(t, \omega), y^*(t, \omega)), t_0 \leq t \leq t_f, \omega \in (\Omega, \mathcal{A}, P)$, denote the unique measurable solution of system (32a)–(32d) and define the control

$$u^*(t) := \tilde{u}^*\left(t, z^*(t, \cdot), y^*(t, \cdot)\right), t_0 \leq t \leq t_f. \quad (33)$$

Owing to Equations (13e) and (26b), and (27a) and (27b), for all $\omega \in (\Omega, \mathcal{A}, P)$ we have

$$z^*(t, \omega) = z^{u^*}(t, \omega), t_0 \leq t \leq t_f, \quad (34a)$$

$$y^*(t, \omega) = y^{u^*}(t, \omega), t_0 \leq t \leq t_f \quad (34b)$$

and therefore

$$u^*(t) = \tilde{u}^*\left(t, z^{u^*}(t, \cdot), y^{u^*}(t, \cdot)\right), t_0 \leq t \leq t_f. \quad (34c)$$

Assuming that

$$u^*(\cdot) \in U \text{ (and therefore } u^*(\cdot) \in D), \quad (35)$$

we get this result.

Theorem 4. Suppose that the canonical system (32a)–(32d) has a unique measurable solution $(z^*, y^*) = (z^*(t, \omega), y^*(t, \omega))$, and define $u^*(\cdot)$ by Equation (33) with a H -minimal control $\tilde{u}^* = \tilde{u}^*(t, \zeta, \eta)$. If $u^*(\cdot) \in U$, then $u^*(\cdot)$ is a stationary control; hence, it fulfills the necessary optimality conditions of the optimal control problem (6a)–(6d).

Proof. According to the construction of (z^*, y^*, u^*) , the control $u^*(\cdot) \in D$ minimizes $\overline{H}^{u^*}(u(\cdot))$ on D . Hence,

$$u^*(\cdot) \in \operatorname{argmin}_{u(\cdot) \in D} \overline{H}^{u^*}(u(\cdot)).$$

Theorem 3 yields then that $u^*(\cdot)$ is a stationary control.

STOCHASTIC PROGRAMMING METHODS

Besides the methods for the calculation of expectations and/or solutions of (finite-dimensional) stochastic programming problems considered in this article, that is, discretization and Taylor expansion methods, there are further well-known methods, for example, based on finite-difference approximation and stochastic quasigradient methods, stochastic linearization procedures, see Ref. 21. However, in the present case not a single stochastic program has to be solved once, but H -minimal controls $\tilde{u}^*$, hence, control laws (functions). Thus, at each time instant t and for certain random variables $\zeta(\omega), \eta(\omega)$, stochastic optimal control laws must be determined (i) as solutions, see the section titled “Canonical System of Differential Equations for Stationary Controls”, of

$$\min EH\left(t, \omega, \zeta(\omega), \eta(\omega), u\right) \text{ s.t. } u \in D_t,$$

and (ii) inserted into the Hamiltonian two-point boundary value problem. Consequently, the solution method for finding $\tilde{u}^*$ does not allow many iterations as needed, for example, in case of stochastic gradient methods, based on *stochastic gradients*

$$\widehat{\nabla_u H}^{(k)} := \nabla_u H(t, \omega^{(k)}, \zeta(\omega^{(k)}), \eta(\omega^{(k)}), u^{(k)})$$

at certain iteration points $u^{(k)}$. Here, ∇_u denotes the gradient with respect to u , and $\omega^{(k)}, k = 1, 2, \dots$, is a series of stochastically independent realizations of the random element ω . Moreover, in order to take into account the constraint $u \in D_t$, projection-type stochastic gradient-type methods must be applied. In addition, it is known that the convergence rate of stochastic gradient procedures may be very small. For a more detailed consideration of stochastic approximation, especially stochastic gradient methods, including hybrid methods, see Ref. 7.

On the other hand, using discretization techniques as well as Taylor expansion methods, the control problem under stochastic uncertainty is reduced to a deterministic control problem involving approximate trajectories corresponding to the applied approximation method. Of course, this presupposes that a moderate number of appropriate realizations of the random element ω , of the random parameter vector $\theta(\omega)$, respectively, can be found for the discretization method; the moments of the random elements, random parameters of third and higher order are small in case of the Taylor expansion method.

Stochastic Programs and Recourse Cost Models

The method developed in the sections titled “Computation of Directional Derivatives” and “Stationary Controls” is considered now in the following case.

Let the function g of the dynamic equation (1a) be affine-linear with respect to the control input u , and assume that the cost function L along the trajectory is separable (additive) with respect to state and control,

hence,

$$g(t, \omega, z, u) = \hat{g}(t, \omega, z) + \hat{B}(t, \omega, z)u \quad (36a)$$

$$L(t, \omega, z, u) = \hat{L}(t, \omega, z) + Q(t, \omega, u). \quad (36b)$$

The stochastic optimization problem $(P)_{\xi, \eta}^t$ is then equivalent to the *stochastic program* [7, 14]

$$\min_{u \in D_t} E \left(\left(\hat{B}(t, \omega, \zeta(\omega))^T \eta(\omega) \right)^T u + Q(t, \omega, u) \right). \quad (37)$$

The important properties of this problem and its solutions can be obtained in the following situations:

- (a) If $Q = Q(t, \omega, u)$ is (strictly) convex with respect to u , then Equation (37) is a (strictly) convex program.
- (b) If $B(t, \omega, z)$ does not depend on z , on (ω, z) , respectively, then

$$\begin{aligned} \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot)) &= \tilde{u}^*(t, \eta(\cdot)), \\ \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot)) &= \tilde{u}^*(t, E\eta(\omega)). \end{aligned}$$

- (c) If $Q(t, \omega, u) = \frac{1}{2}u^T u$ and $D_t = \mathbb{R}^n$, then

$$\tilde{u}^*(t, \zeta(\cdot), \eta(\cdot)) = -E\hat{B}(t, \omega, \zeta(\omega))^T \eta(\omega).$$

- (d) If the cost function $Q = Q(t, \omega, u)$ is strictly convex in u for each $(t, \omega) \in [t_0, t_f] \times \Omega$ a.s., and the admissible sets D_t , $t_0 \leq t \leq t_f$, are compact, then the program (37) has a unique optimal solution. Hence, the H -minimal control $\tilde{u}^* = \tilde{u}^*(t, \zeta(\cdot), \eta(\cdot))$ is determined uniquely.
- (e) For a discrete probability distribution $\Omega = \{\omega_1, \dots, \omega_\varrho\}$, $P(\omega = \omega_j) = \alpha_j, j = 1, \dots, \varrho, \sum_{j=1}^{\varrho} \alpha_j = 1$, Equation (37) is reduced to the ordinary deterministic parameter optimization problem

$$\begin{aligned} \min_{u \in D_t} \sum_{j=1}^{\varrho} \alpha_j \left(\hat{B}(t, \omega_j, \zeta(\omega_j))^T \eta(\omega_j) \right)^T u \\ + \sum_{j=1}^{\varrho} \alpha_j Q(t, \omega_j, u) \end{aligned} \quad (37')$$

depending on the parameters $(\omega_j, \zeta(\omega_j), \eta(\omega_j))$ and probabilities α_j , $j = 1, \dots, \varrho$.

- (f) *Recourse cost models* Stochastic programs of the above type (37) are well studied, especially for *recourse cost models* [14, 22] related to compensation of violations of stochastic control constraints, hence,

$$Q(t, \omega, u) := p_t(T(t, \omega)u - b(t, \omega)). \quad (38a)$$

Here, $b : [t_0, t_f] \times \Omega \rightarrow \mathbb{R}^q$, $T : [t_0, t_f] \times \Omega \rightarrow \mathbb{R}^{q \times n}$ are vector-, matrix-valued measurable functions, and $p_t : \mathbb{R}^q \rightarrow \mathbb{R}$ is a convex or sublinear function [14]. We may then interpret $Q(t, \omega, u)$ as the loss resulting from compensating the violation $\delta = T(t, \omega)u - b(t, \omega) \neq 0, \neq 0$, respectively, of stochastic linear equality or inequality control constraints

$$T(t, \omega)u = b(t, \omega), T(t, \omega)u \leq b(t, \omega), \\ t_0 \leq t \leq t_f,$$

after revealing the uncertain parameters. One of great importance is the case that $b(t, \cdot)$, $T(t, \cdot)$ are stable distributed stochastic processes [23]. The objective function of Equation (37) has then the form

$$\phi(u) = \bar{c}(t)^T u + E p_t(T(t, \omega)u - b(t, \omega)), \\ u \in \mathbb{R}^n, \quad (38b)$$

where $\bar{c}(t)$ is defined by

$$\bar{c}(t) = E \hat{B}(t, \omega, \gamma(\omega))^T \eta(\omega). \quad (38c)$$

On the basis of the *stochastic dominance* concept, for problems with stable distributed processes $(T(t, \omega), b(t, \omega))$ several important properties, such as the construction of feasible descent directions, of stochastic programs having an objective function of the type (38b) are shown in Ref. 14.

Computation of Expectations by means of Taylor Expansion

Corresponding to the assumptions in the section titled “Parametric Representation of the Random Differential Equation”, we suppose here the following parametric representation of the functions under consideration:

$$g(t, w, z, u) = \tilde{g}(t, \theta, z, u) \quad (39a)$$

$$z_0(\omega) = \tilde{z}_0(\theta) \quad (39b)$$

$$L(t, \omega, z, u) = \tilde{L}(t, \theta, z, u) \quad (39c)$$

$$G(t, \omega, z) = \tilde{G}(t, \theta, z) \quad (39d)$$

with a time-independent r -vector

$$\theta = \theta(\omega), \omega \in (\Omega, \mathcal{A}, P), \quad (39e)$$

of random model and environmental parameters including initial values. Furthermore, we assume that $\tilde{g}, \tilde{z}_0, \tilde{L}, \tilde{G}$ are sufficiently smooth functions of the indicated variables. For simplification of notation we omit the symbol “ $\sim$ ”.

Again, for simplification, the conditional expectation $E(\dots | \mathcal{A}_{t_0})$, given the information $\mathcal{A}_{t_0}$ up to the considered starting or reference time t_0 is denoted by “ E ”. Thus, let

$$\bar{\theta} := E\theta(\omega) = E(\theta(\omega) | \mathcal{A}_{t_0}) \quad (40a)$$

denote the conditional expectation of the random vector $\theta(\omega)$ given the information $\mathcal{A}_{t_0}$ at time point t_0 . Taking into account the properties of the solution

$$z = z(t, \theta) = S(z_0(\theta), \theta, u(\cdot))(t), t \geq t_0, \quad (40b)$$

of the dynamic equations (3a)–(3d), see Lemma 1 and Lemma 2, the expectations arising in the objective function (6a) can be computed approximately by means of Taylor expansion with respect to θ at $\bar{\theta}$.

According to Equations (39a)–(39e) we have the parametric representation of $z(t, \cdot), y(t, \cdot)$:

$$z(t, \omega) = z(t, \theta(\omega)), y = y(t, \theta(\omega)). \quad (41)$$

We consider then the Taylor expansions

$$\begin{aligned} z(t, \omega) &= z(t, \theta(\omega)) \\ &= z(t, \bar{\theta}) + z_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}) + \dots \end{aligned} \quad (42a)$$

$$\begin{aligned} y(t, \omega) &= y(t, \theta(\omega)) \\ &= y(t, \bar{\theta}) + y_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}) + \dots \end{aligned} \quad (42b)$$

Deleting second- and higher-order terms, we have to determine the functions

$$z = z(t, \bar{\theta}), y = y(t, \bar{\theta}), t_0 \leq t \leq t_f \quad (43a)$$

$$z_\theta = z_\theta(t, \bar{\theta}), y_\theta = y_\theta(t, \bar{\theta}), t_0 \leq t \leq t_f, \quad (43b)$$

with the Jacobians $z_\theta(t, \bar{\theta}) := D_\theta z(t, \bar{\theta})$ and $y_\theta(t, \bar{\theta}) := D_\theta y(t, \bar{\theta})$.

Using this expansion, the stochastic optimization problem $(P)_{\zeta, \eta}^t$ with $\zeta := z(t, \theta(\cdot))$, $\eta := y(t, \theta(\cdot))$ can be approximated by

$$\begin{aligned} \min_{u \in D_t} EH \left(t, \theta(\omega), z(t, \bar{\theta}) + z_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}), \right. \\ \left. y(t, \bar{\theta}) + y_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}), u \right). \end{aligned} \quad (44)$$

Thus, the resulting approximate H -minimal control

$$\tilde{u}^* = \tilde{u}^*(t, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), y(t, \bar{\theta}), y_\theta(t, \bar{\theta})) \quad (45)$$

depends on the unknown vector and matrix functions stated in Equations (43a) and (43b).

Describing random parameter variations by means of a certain finite-dimensional random parameter vector $\theta = \theta(\omega)$, a two-point boundary value problem for the unknown functions stated in Equations (43a) and (43b) can be then obtained from Equations (32a)–(32d), as is described in more detail in the following, by (i) putting $\theta = \bar{\theta}$ and by (ii) differentiating with respect to θ and then putting $\theta = \bar{\theta}$.

Computation of Trajectories and Sensitivities for $\theta = \bar{\theta}$.

1. Trajectory $z = z(t, \omega) = z(t, \theta(\omega))$.

According to the first-order Taylor approximation (42a) of $z = z(t, \theta)$

at $\theta = \bar{\theta}$, we still have to compute the trajectory $t \rightarrow z(t, \bar{\theta}), t \geq t_0$, related to the mean parameter vector $\theta = \bar{\theta}$ and the sensitivities $t \rightarrow \frac{\partial z}{\partial \theta_i}(t, \bar{\theta}), i = 1, \dots, r, t \geq t_0$, of the state $z = z(t, \theta)$ with respect to the parameters $\theta_i, i = 1, \dots, r$, at $\theta = \bar{\theta}$. On the basis of Equations (3a) and (3b), in case of Equations (39a)–(39d) for $z = z(t, \bar{\theta})$, we have the system of differential equations

$$\dot{z}(t, \bar{\theta}) = g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) t \geq t_0, \quad (46a)$$

$$z(t_0, \bar{\theta}) = z_0(\bar{\theta}). \quad (46b)$$

Moreover, assuming that differentiation with respect to $\theta_i, i = 1, \dots, r$, and integration with respect to time t can be interchanged, see Lemma 1, from Equation (3c) we obtain the following system of linear perturbation differential equations for the Jacobian $z_\theta(t, \bar{\theta}) = \left(\frac{\partial z}{\partial \theta_1}(t, \bar{\theta}), \frac{\partial z}{\partial \theta_2}(t, \bar{\theta}), \dots, \frac{\partial z}{\partial \theta_r}(t, \bar{\theta}) \right), t \geq t_0$:

$$\begin{aligned} \frac{d}{dt} (z_\theta(t, \bar{\theta})) &= D_z g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) z_\theta(t, \bar{\theta}) \\ &\quad + D_\theta g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)), \\ t &\geq t_0, \end{aligned} \quad (47a)$$

$$z_\theta(t_0, \bar{\theta}) = z_{0\theta}(\bar{\theta}). \quad (47b)$$

Note. Equations (47a) and (47b) are closely related to the perturbation equation (13a), (13b) for representing the derivative z'' of z with respect to the control u .

2. Adjoint trajectory $y = y(t, \omega) = y(t, \theta(\omega))$.

The adjoint trajectory $y = y(t, \theta)$ is defined by the system of linear differential equation systems (27a) and (27b) with $v(\cdot) = u(\cdot)$. Hence, using the parameter representations (39a)–(39e), for $y = y(t, \theta)$, we have the

terminal value problem

$$\begin{aligned} \dot{y}(t, \theta) = & -D_z g(t, \theta, z(t, \theta), u(t))^T y(t, \theta) \\ & - \nabla_z L(t, \theta, z(t, \theta), u(t)) \end{aligned} \quad (48a)$$

$$y(t_f, \theta) = \nabla_z G(t_f, \theta, z(t_f, \theta)). \quad (48b)$$

Putting $\theta := \bar{\theta}$, from Equations (48a) and (48b) we obtain a terminal value problem for $y = y(t, \bar{\theta})$. Furthermore, corresponding to the derivation of a system of first-order ordinary differential equations for the Jacobian $z_\theta(t, \bar{\theta})$, a system of first-order differential equations for the Jacobian $y_\theta(t, \bar{\theta})$ can be obtained, by differentiation of Equations (48a) and (48b) with respect to θ at $\theta = \bar{\theta}$.

Remark 6. The differentiation of Equations (48a) and (48b) with respect to θ is especially simple if the Jacobian $D_z g = D_z g(t, u(t))$ of g with respect to z is independent of $(\theta, z(t, \theta))$ (see the section titled “Optimal Regulator Design for Dynamic Systems Under Stochastic Uncertainty”).

Deterministic Substitute Optimal Control Problem

Instead of solving the optimal control problem (6a)–(6d) under stochastic uncertainty, using the stochastic Hamiltonian H as described in the sections titled “Computation of Directional Derivatives” and “Stationary Controls”, we can replace this stochastic control problem *directly* by an *approximate deterministic substitute control problem* by computation of the expectation in the objective function (6a) using Taylor expansion with respect to the random parameter vector $\theta = \theta(\omega)$ at $\bar{\theta}$. For this purpose we need the (approximate) Hessians of the cost functions along the trajectory, at the terminal point, respectively,

$$\theta \rightarrow L(t, \theta, z(t, \theta), u), \theta \rightarrow G(t, \theta, z(t, \theta))$$

at $\theta = \bar{\theta}$.

Retaining only first-order derivatives of $z = z(t, \theta)$ with respect to θ , the approximate Hessian Q_L of $\theta \rightarrow L(t, \theta, z(t, \theta), u)$ at $\theta = \bar{\theta}$ is given [7], by

$$\begin{aligned} Q_L(t, \bar{\theta}, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), u(t)) \\ := \nabla_\theta^2 L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) \\ + z_\theta(t, \bar{\theta})^T \nabla_{\theta z} L(t, \bar{\theta}, z(t, \bar{\theta}), u(t))^T \\ + \nabla_{\theta z} L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) z_\theta(t, \bar{\theta}) \\ + z_\theta(t, \bar{\theta})^T \nabla_z^2 L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) z_\theta(t, \bar{\theta}). \end{aligned} \quad (49)$$

Here, $\nabla_\theta L, \nabla_z L$ denotes the gradient of L with respect to θ, z , respectively, z_θ is the Jacobian of $z = z(t, \theta)$ with respect to θ , and $\nabla_\theta^2 L, \nabla_z^2 L$, respectively, denotes the Hessian of L with respect to θ, z .

Thus, for the expected costs $EL = E(L|\mathcal{A}_{t_0})$, with Equation (49) we obtain the expansion

$$\begin{aligned} EL(t, \theta(\omega), z(t, \theta(\omega)), u(t)) \\ = L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) + \frac{1}{2} E(\theta(\omega) - \bar{\theta})^T \\ \times Q_L(t, \bar{\theta}, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), u(t)) \\ \times (\theta(\omega) - \bar{\theta}) + \dots \\ = L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) \\ + \frac{1}{2} \text{tr} Q_L(t, \bar{\theta}, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), u(t)) \\ \times \text{cov}(\theta(\cdot)) + \dots \end{aligned} \quad (50)$$

In order to evaluate formula (50), besides the distributional parameters $\bar{\theta}$ and $\text{cov}(\theta(\cdot)) = E(\theta(\omega) - \bar{\theta})(\theta(\omega) - \bar{\theta})^T$, we need the functions $z = z(t, \bar{\theta})$ and $z_\theta = z_\theta(t, \bar{\theta})$. However, $z = z(t, \bar{\theta}), t \geq t_0$, is the solution of the dynamic equations (3a)–(3c) for $\theta = \bar{\theta}$. Moreover, a system of differential equations for the derivatives $\frac{\partial z_k}{\partial \theta_i} = \frac{\partial z_k}{\partial \theta_i}(t, \bar{\theta}), t \geq t_0$,

is obtained as described in the section titled “Computation of Trajectories and Sensitivities for $\theta = \bar{\theta}$ ”.

Defining the approximate Hessian Q_G of $\theta \rightarrow G(t, \theta, z(t, \theta))$ at $\theta = \bar{\theta}$ as in case of Q_L , see Equation (49), for $EG = E(G|A_{t_0})$ we get the expansion

$$\begin{aligned} EG(t_f, \theta(\omega), z(t_f, \theta(\omega))) &= G(t_f, \bar{\theta}, z(t_f, \bar{\theta})) \\ &+ \frac{1}{2} \text{tr} Q_G(t_f, \bar{\theta}, z(t_f, \bar{\theta}), z_{\theta}(t_f, \bar{\theta})) \text{cov}(\theta(\cdot)) + \dots \end{aligned} \quad (51)$$

For the stochastic control problem (6a)–(6d) we then obtain the following approximation.

Theorem 5. *Suppose that differentiation with respect to the parameters $\theta_i, i = 1, \dots, r$, and integration with respect to time t can be interchanged in Equation (3c). Retaining only first-order derivatives of $z = z(t, \theta)$ with respect to θ , problem (6a)–(6d) can be approximated by the ordinary deterministic optimal control problem:*

$$\begin{aligned} \min \int_{t_0}^{t_f} &\left\{ L(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) \right. \\ &+ \frac{1}{2} \text{tr} Q_L(t, \bar{\theta}, z(t, \bar{\theta}), z_{\theta}(t, \bar{\theta}), u(t)) \text{cov}(\theta(\cdot)) \Big\} dt \\ &+ G(t_f, \bar{\theta}, z(t_f, \bar{\theta})) \\ &+ \frac{1}{2} \text{tr} Q_G(t_f, \bar{\theta}, z(t_f, \bar{\theta}), z_{\theta}(t_f, \bar{\theta})) \text{cov}(\theta(\cdot)) \end{aligned} \quad (52a)$$

s.t.

$$\dot{z}(t, \bar{\theta}) = g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)), t \geq t_0, \quad (52b)$$

$$z(t_0, \bar{\theta}) = z_0(\bar{\theta}) \quad (52c)$$

$$\begin{aligned} \frac{d}{dt}(z_{\theta}(t, \bar{\theta})) &= D_z g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)) z_{\theta}(t, \bar{\theta}) \\ &+ D_{\theta} g(t, \bar{\theta}, z(t, \bar{\theta}), u(t)), t \geq t_0, \end{aligned} \quad (52d)$$

$$z_{\theta}(t_0, \bar{\theta}) = z_{0\theta}(\bar{\theta}) \quad (52e)$$

$$u(\cdot) \in D. \quad (52f)$$

Note. According to previous remarks, we have

$$\begin{aligned} \bar{\theta} = \bar{\theta}^{(t_0)} &:= E(\theta(\omega) | A_{t_0}) \\ \text{cov}(\theta(\cdot)) &= \text{cov}^{(t_0)}(\theta(\cdot)) \\ &:= E\left((\theta(\omega) - \bar{\theta}^{(t_0)}) (\theta(\omega) - \bar{\theta}^{(t_0)})^T | A_{t_0}\right). \end{aligned}$$

Remark 7. The deterministic control problem (52a)–(52f) can be solved, of course, also by applying the Hamilton–Jacobi theory to this deterministic problem. Comparing then the corresponding two-point boundary value problem related to Equations (52a)–(52f) with the two-point boundary value problem resulting from the two-point boundary problem (32a)–(32d) with random parameters after applying Taylor expansion methods with respect to θ at $\bar{\theta}$, as described in the sections titled “Computation of Expectations by means of Taylor Expansion” and “Computation of Trajectories and Sensitivities for $\theta = \bar{\theta}$ ”, a close relationship of these two approaches is revealed.

OPTIMAL REGULATOR DESIGN FOR DYNAMIC SYSTEMS UNDER STOCHASTIC UNCERTAINTY

On the basis of a prescribed *optimal reference trajectory* of a dynamic system under stochastic uncertainty, the problem is to keep the system on this trajectory as exactly as possible by means of control corrections obtained by applying a stochastic optimal feedback control law $\varphi = \varphi(t, \Delta q, \Delta \dot{q})$, hence, a PD controller, where $\Delta q, \Delta \dot{q}$ denote the tracking error and its time derivative [12,24,25].

Taylor expansion of the feedback function φ at $\Delta q = 0, \Delta \dot{q} = 0$, retaining only linear terms, and using the additional condition $\varphi(t, 0, 0) = 0$ we get the approximation:

$$\begin{aligned} \varphi(t, \Delta q(t), \Delta \dot{q}(t)) &\approx D_{\Delta q} \varphi(t, 0, 0) \Delta q(t) \\ &+ D_{\Delta \dot{q}} \varphi(t, 0, 0) \Delta \dot{q}(t). \end{aligned} \quad (53)$$

Linearizing the dynamic equation of the underlying dynamic system around the reference trajectory, its first and second time derivative as well as around the expectation of the stochastic dynamic parameters $p_D(\omega)$, and using the linearized control law stated in Equation (53), for the tracking error Δq we get the following system of linear differential equations

$$\begin{aligned} &Y(t) \Delta p_D(\omega) + K(t) \Delta q(t) \\ &+ D(t) \Delta \dot{q}(t) + M(t) \Delta \ddot{q}(t) \\ &= D_{\Delta q} \varphi(t, 0, 0) \Delta q(t) + D_{\Delta \dot{q}} \varphi(t, 0, 0) \Delta \dot{q}(t), \\ &t_0 \leq t \leq t_f, \end{aligned} \quad (54a)$$

with certain deterministic matrices $Y(t), K(t), D(t), M(t)$ depending on the optimal reference trajectory [12]. Moreover, $\Delta p_D(\omega)$ denotes the deviation of the vector of dynamic parameters from its expectation. The above equation can be rewritten as

$$\begin{aligned} &M(t) \Delta \ddot{q}(t) + (K(t) - D_{\Delta q} \varphi(t, 0, 0)) \Delta q(t) \\ &+ (D(t) - D_{\Delta \dot{q}} \varphi(t, 0, 0)) \Delta \dot{q}(t), \\ &= -Y(t) \Delta p_D(\omega). \end{aligned} \quad (54b)$$

In many applications M is always regular. Hence, defining

$$K_p(t) := M(t)^{-1} (K(t) - D_{\Delta q} \varphi(t, 0, 0)), \quad (54c)$$

$$K_d(t) := M(t)^{-1} (D(t) - D_{\Delta \dot{q}} \varphi(t, 0, 0)), \quad (54d)$$

from Equations (54b)–(54d) we get

$$\begin{aligned} \Delta \ddot{q}(t) &= -K_p(t) \Delta q(t) - K_d(t) \Delta \dot{q}(t) \\ &- M(t)^{-1} Y(t) \Delta p_D(\omega). \end{aligned} \quad (55a)$$

By introducing the state variable

$$\Delta z(t) := \begin{pmatrix} \Delta q(t) \\ \Delta \dot{q}(t) \end{pmatrix}, \quad (55b)$$

the second order differential equation (55a) will be transformed into a system of first-order differential equations

$$\begin{aligned} \Delta \dot{z}(t, \omega) &= \begin{pmatrix} \Delta \dot{q} \\ -K_p(t) \Delta q(t) - K_d(t) \Delta \dot{q}(t) \\ -M(t)^{-1} Y(t) \Delta p_D(\omega) \end{pmatrix} \\ &= \begin{pmatrix} 0 & I \\ -K_p(t) & -K_d(t) \end{pmatrix} \Delta z(t, \omega) \\ &+ \begin{pmatrix} 0 \\ -M(t)^{-1} Y(t) \Delta p_D(\omega) \end{pmatrix}. \end{aligned} \quad (55c)$$

The above Equation (55c) can also be represented by

$$\Delta \dot{z}(t, \omega) = A(V(t)) \Delta z(t, \omega) + a(t, \Delta p_D(\omega)), \quad (56a)$$

where

$$A(V(t)) := \begin{pmatrix} 0 & I \\ -K_p(t) & -K_d(t) \end{pmatrix} \quad (56b)$$

$$a(t, \Delta p_D) := \begin{pmatrix} 0 \\ -M^{(0)}(t)^{-1} Y^{(0)}(t) \end{pmatrix} \Delta p_D(\omega), \quad (56c)$$

and

$$V(t) := (K_p(t), K_d(t)). \quad (56d)$$

The performance of the feedback control is given by means of the quadratic cost functions (4d) and (4e)

$$L(t, \omega, \Delta z, \varphi) := \Delta z^T Q(\omega) \Delta z + \varphi^T R(\omega) \varphi, \quad (57a)$$

where the feedback law φ is approximated (PD controller) according to Equation (53). Moreover, in the present application we have a zero terminal cost function

$$G(z) := 0. \quad (57b)$$

According to Equations (54c) and (54d), there is a 1-1 relationship between the

unknown derivatives $D_{\Delta q}\varphi(t, 0, 0), D_{\Delta \dot{q}}\varphi(t, 0, 0)$ of the feedback control law φ and the regulator parameters contained in the matrices $K_p(t), K_d(t)$. Hence, in case of regulator optimization, the unknown input function $u = u(t)$ is the *the matrix (56d) of unknown, time-dependent regulator parameters*, thus,

$$“u(t)” = V(t) := (K_p(t), K_d(t)). \quad (58a)$$

With Equations (54c) and (54d), the performance function of the regulator optimization problem is then represented, cf. Equations (57a) and (57b), by

$$L(t, \omega, \Delta z, \varphi) = \Delta z^T \tilde{Q}(t, V(t)) \Delta z, \quad (58b)$$

where

$$\tilde{Q}(t, V(t)) := Q + B(t, V(t))^T R B(t, V(t)), \quad (58c)$$

$$B(t, V(t)) := (K(t), D(t)) - M(t)V(t). \quad (58d)$$

Remark 8. Note that in the present case $u, q, \dot{q}, z, \ddot{q}, \theta$ are replaced by the standard notation $V, \Delta q, \Delta \dot{q}, \Delta z, \Delta \ddot{q}, \Delta p_D$ in regulator theory.

According to the approximate computation of expectations developed in the section titled “Computation of Expectations by means of Taylor Expansion”, cf. Equations (42a) and (42b); (43a and (43b); (44); and (45), for finding now a H -minimal control

$$\tilde{V}^* = \tilde{V}^*(t, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), y(t, \bar{\theta}), y_\theta(t, \bar{\theta})) \quad (59)$$

we have to solve the minimization problem

$$\min_{V \in D_t} EH \left(t, \theta(\omega), z(t, \bar{\theta}) + z_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}), y(t, \bar{\theta}) + y_\theta(t, \bar{\theta})(\theta(\omega) - \bar{\theta}), V \right). \quad (60)$$

Here, the Hamilton function reads, cf. Equations (56a)–(56d); Equations (57a) and (57b),

$$H(t, \Delta p_D, \Delta z, y, V) := \Delta z^T \tilde{Q}(t, V(t)) \Delta z + y^T (A(V(t)) \Delta z + a(t, \Delta p_D)), \quad (61a)$$

where

$$\theta(\omega) := \Delta p_D = p_D(\omega) - \bar{p}_D \quad (61b)$$

denotes the deviation of the vector $p_D(\omega)$ of dynamic parameters from its expectation $\bar{p}_D$. According to the linear-quadratic structure of H in the present case, the expectation in Equation (60) can then be determined explicitly in terms of the expectation $\bar{\theta}$ and covariance matrix $cov(\theta(\cdot))$ of the random parameter vector $\theta(\omega) := \Delta p_D(\omega)$. Moreover, depending on the concrete regulator problem, the solution of problem (60) at each time point t is determined (partly) analytically or by using numerical optimization procedures.

Thus, for finding the functions

$$t \rightarrow (z(t, \bar{\theta}), y(t, \bar{\theta}), z_\theta(t, \bar{\theta}), y_\theta(t, \bar{\theta})), t_0 \leq t \leq t_f,$$

needed in the Taylor expansions (42a) and (42b); (43a) and (43b), we have, see sections titled “Stationary Controls” and “Stochastic Programming Methods” and Remark 6, the following two-point boundary value problem:

Theorem 6. Let $\theta(\omega) = \Delta p_D(\omega)$ denote the random parameter vector of the regulator optimization problem. Assuming the existence of the derivatives under consideration, for the functions $z = z(t, \bar{\theta}), y = y(t, \bar{\theta})$ and $z_\theta = z_\theta(t, \bar{\theta}), y_\theta = y_\theta(t, \bar{\theta})$ we have the two-point boundary value problem

$$\dot{z}(t, \bar{\theta}) = A(\tilde{V}^*)z(t, \bar{\theta}) + a(t, \bar{\theta}), t_0 \leq t \leq t_f, \quad (62a)$$

$$z(t_0, \bar{\theta}) = z_0(\bar{\theta}) \quad (62b)$$

$$\dot{y}(t, \bar{\theta}) = -A(\tilde{V}^*)^T y(t, \bar{\theta}) - 2\tilde{Q}(t, \tilde{V}^*)z(t, \bar{\theta}) \quad (62c)$$

$$y(t_f, \bar{\theta}) = 0 \quad (62d)$$

$$\dot{z}_\theta(t, \bar{\theta}) = A(\tilde{V}^*)z_\theta(t, \bar{\theta}) + a_\theta(t, \bar{\theta}) \quad (62e)$$

$$z_\theta(t_0, \bar{\theta}) = z_{0\theta}(\bar{\theta}) \quad (62f)$$

$$\dot{y}_\theta(t, \bar{\theta}) = -A(\tilde{V}^*)^T y_\theta(t, \bar{\theta}) - 2\tilde{Q}(t, \tilde{V}^*)z_\theta(t, \bar{\theta}) \quad (62g)$$

$$y_\theta(t_f, \bar{\theta}) = 0. \quad (62h)$$

where, the H -minimal control $\tilde{V}^* = \tilde{V}^*(t, z(t, \bar{\theta}), z_\theta(t, \bar{\theta}), y(t, \bar{\theta}), y_\theta(t, \bar{\theta}))$ results from solving problem (60).

An optimal matrix $V^* = V^*(t)$ of regulator parameters is then given by

$$V^*(t) = \tilde{V}^*(t, z^*(t, \bar{\theta}), z_\theta^*(t, \bar{\theta}), y^*(t, \bar{\theta}), y_\theta^*(t, \bar{\theta})), \quad (63)$$

where $(z^*(t, \bar{\theta}), z_\theta^*(t, \bar{\theta}), y^*(t, \bar{\theta}), y_\theta^*(t, \bar{\theta}))$ is a solution of system (62a)–(62h).

Manutec r3

The method for finding optimal controls being insensitive with respect to random parameter variations presented here is applied now to the stochastic optimal feedback control of a manipulator (Manutec r3) with three rotational links (three degrees of freedom) [26].

Since for $\theta = \bar{\theta}$ or $\Delta p_D = \overline{\Delta p_D} = 0$, hence, if the vector of dynamic parameters is known, we also have

$$\Delta z(t, 0) = 0, y(t, 0) = 0, \quad (64a)$$

the Taylor expansion of $\Delta z(t, \omega), y(t, \omega)$ is reduced, cf. Equations (42a) and (42b); (56a)–(56c); and (62a)–(62h), to

$$\Delta z(t, \omega) \approx \Delta z_{\Delta p_D}(t, 0) \Delta p_D(\omega), \quad (64b)$$

$$y(t, \omega) \approx y_{\Delta p_D}(t, 0) \Delta p_D(\omega). \quad (64c)$$

Assuming that the payload mass $p_D(\omega)$ is a scalar stochastic variable, the expected Hamiltonian EH is given by, cf. Equation (61a),

$$\begin{aligned} EH(t, \Delta p_D, \Delta z, y, V) \\ = \sigma_{p_D}^2 \left(\Delta z_{\Delta p_D}(t, 0)^T \tilde{Q}(t, V(t)) \Delta z_{\Delta p_D}(t, 0) \right. \\ \left. + y_{\Delta p_D}(t, 0)^T A(V(t)) \Delta z_{\Delta p_D}(t, 0) \right. \\ \left. + y_{\Delta p_D}(t, 0)^T \begin{pmatrix} 0 \\ -M(t)^{-1} Y(t) \end{pmatrix} \right), \quad (65) \end{aligned}$$

where $\sigma_{p_D}^2$ denotes the variance of the scalar payload. As mentioned above, the minimization of Equation (65) subject to the

constraint

$$V \in D_t \quad (66)$$

can be done analytically in certain cases. Assuming, in the present example, that there are no constraints for the regulator parameters contained in V , the minimization of the expected Hamiltonian EH is carried out numerically for given function values $(\Delta z_{\Delta p_D}(t, 0), y_{\Delta p_D}(t, 0))$ at each time t .

With zero terminal costs, hence $G = 0$, because of Equations (64a)–(64c), in the present case, the two-point boundary value problem (62a)–(62h) is then reduced to

$$\begin{aligned} \Delta \dot{z}_{\Delta p_D}(t, 0) &= A(\tilde{V}^*) \Delta z_{\Delta p_D}(t, 0) + a_{\Delta p_D}(t, 0), \\ t_0 &\leq t \leq t_f, \end{aligned} \quad (67a)$$

$$\begin{aligned} \dot{y}_{\Delta p_D}(t, 0) &= -A(\tilde{V}^*)^T y_{\Delta p_D}(t, 0) - 2\tilde{Q}(t, \tilde{V}^*) \\ &\times \Delta z_{\Delta p_D}(t, 0), \quad t_0 \leq t \leq t_f, \end{aligned} \quad (67b)$$

$$\Delta z_{\Delta p_D}(t_0, 0) = 0 \quad (67c)$$

$$y_{\Delta p_D}(t_f, 0) = 0, \quad (67d)$$

where $\tilde{V}^*$ is obtained from the numerical minimization of the approximate expected Hamiltonian $EH(t, \Delta p_D, \Delta z, y, V)$, cf. Equation (65).

For the above described two-point boundary value problem we have then the following 12 state variables x_i , $i = 1, \dots, 12$:

$$\begin{aligned} x_1(t) &= \frac{\partial \Delta q_1}{\partial \Delta p_D}(t, 0), & x_7(t) &= \frac{\partial y_1}{\partial \Delta p_D}(t, 0) \\ x_2(t) &= \frac{\partial \Delta q_2}{\partial \Delta p_D}(t, 0), & x_8(t) &= \frac{\partial y_2}{\partial \Delta p_D}(t, 0) \\ x_3(t) &= \frac{\partial \Delta q_3}{\partial \Delta p_D}(t, 0), & x_9(t) &= \frac{\partial y_3}{\partial \Delta p_D}(t, 0) \\ x_4(t) &= \frac{\partial \dot{\Delta q}_1}{\partial \Delta p_D}(t, 0), & x_{10}(t) &= \frac{\partial y_4}{\partial \Delta p_D}(t, 0) \\ x_5(t) &= \frac{\partial \dot{\Delta q}_2}{\partial \Delta p_D}(t, 0), & x_{11}(t) &= \frac{\partial y_5}{\partial \Delta p_D}(t, 0) \\ x_6(t) &= \frac{\partial \dot{\Delta q}_3}{\partial \Delta p_D}(t, 0), & x_{12}(t) &= \frac{\partial y_6}{\partial \Delta p_D}(t, 0). \end{aligned}$$

Moreover, at the starting time point $t_0 = 0$ we have the initial conditions

$$x_1(0) = x_2(0) = \dots = x_6(0) = 0,$$

and the terminal conditions read:

$$x_7(t_f) = x_8(t_f) = \dots = x_{12}(t_f) = 0.$$

For simplification, we suppose now that the matrix $V(t)$ of regulator parameter functions has *diagonal* submatrices $K_p(t), K_d(t)$, cf. Equation (58a). Hence, the functions to be optimally determined are

$$\begin{aligned} v_1(t) &:= [K_p]_{11}(t) & v_2(t) &:= [K_p]_{22}(t) \\ v_3(t) &:= [K_p]_{33}(t) \end{aligned} \quad (68)$$

$$\begin{aligned} v_4(t) &:= [K_d]_{11}(t) & v_5(t) &:= [K_d]_{22}(t) \\ v_6(t) &:= [K_d]_{33}(t). \end{aligned} \quad (69)$$

In addition, we define

$$c_{ij} = [\tilde{Q}(t, V(t))]_{ij}. \quad (70)$$

Note that the needed representation of $\tilde{V}$ as a function of $t, \Delta z, y$ is obtained by minimizing the approximate expected Hamiltonian (65) with respect to the above mentioned elements $v_i = v_i(t)$, $i = 1, \dots, 6$.

Differential equation (67a) can then be represented by

$$\begin{aligned} \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \\ \dot{x}_6 \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -v_1 & 0 & 0 & -v_4 & 0 & 0 \\ 0 & -v_2 & 0 & 0 & -v_5 & 0 \\ 0 & 0 & -v_3 & 0 & 0 & -v_6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} \\ &+ \begin{pmatrix} 0 \\ 0 \\ 0 \\ -[(M(t))^{-1} Y(t)]_1 \\ -[(M(t))^{-1} Y(t)]_2 \\ -[(M(t))^{-1} Y(t)]_3 \end{pmatrix}, \end{aligned} \quad (71)$$

and Equation (67b) reads:

$$\begin{aligned} \begin{pmatrix} \dot{x}_7 \\ \dot{x}_8 \\ \dot{x}_9 \\ \dot{x}_{10} \\ \dot{x}_{11} \\ \dot{x}_{12} \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 & v_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & v_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & v_3 \\ -1 & 0 & 0 & v_4 & 0 & 0 \\ 0 & -1 & 0 & 0 & v_5 & 0 \\ 0 & 0 & -1 & 0 & 0 & v_6 \end{pmatrix} \begin{pmatrix} x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \end{pmatrix} \\ &- 2 \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56} \\ c_{61} & c_{62} & c_{63} & c_{64} & c_{65} & c_{66} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}. \end{aligned} \quad (72)$$

Having representations (71) and (72), the two-point boundary value problems (67a)–(67c) is solved numerically by means of the multiple-shooting method BVPSOL (Boundary Value Problem Solver) [27]. According to Equation (53), the optimal feedback function $\varphi^* = \varphi^*(t, \Delta_q(t), \Delta_d(t))$ for the robot is then obtained, cf. Equations (54c) and (54d), from

$$D_{\Delta q} \varphi(t, 0, 0) = K(t) - M(t)K_p^*(t) \quad (73a)$$

$$D_{\Delta d} \varphi(t, 0, 0) = D(t) - M(t)K_d^*(t), \quad (73b)$$

where $V^*(t) = (K_p^*(t), K_d^*(t))$ results from the insertion of the solution of system (67a)–(67d) into the approximate H -minimal control $\tilde{V}^*$ obtained from minimizing the expected Hamiltonian (65) subject to $V(t)$.

In order to evaluate the tracking properties of the robust optimal feedback control law obtained by the above described method, we apply the obtained stochastic optimal regulator to 11 actual trajectories of the manipulator generated by 11 realizations of the random payload mass assumed to be $N(0,5)$ -normal distributed. We assume that these realizations lie in the confidence interval related to the confidence number α . Events having a probability smaller than $\frac{1-\alpha}{2}$, larger than $\frac{1+\alpha}{2}$, respectively are not taken into account. For $\alpha = 0.9$, the remaining range is then partitioned by equidistant points. The resulting realizations are given in the following table.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Probability of x	0.05	0.14	0.23	0.32	0.41	0.5
$p_D = \Phi^{-1}(x)$	1.7755	4.5985	6.3055	7.6615	8.8625	10
	Case 7	Case 8	Case 9	Case 10	Case 11	
Probability of x	0.59	0.68	0.77	0.86	0.95	
$p_D = \Phi^{-1}(x)$	11.1375	12.3385	13.6945	15.4015	18.2245	

The optimal feed forward control and the reference trajectory for a point-to-point trajectory optimization problem from the initial point

$$\begin{pmatrix} 0, 0 \\ 1, 5 \\ 0, 5 \end{pmatrix} \text{ to the terminal point } \begin{pmatrix} -2, 8 \\ 0, 2 \\ 0, 6 \end{pmatrix}$$

were computed by means of the stochastic robot optimization software (*OSTP*) [28]. The trajectories were computed by *SIMPACK* [29], a multibody simulation software package.

In the following Figs 3–8 the simulation results for the first link are presented. Here, $q_1 = q_1(t)$ denotes the rotational angle in the first link of the *Manutec r3*, and $u_1 = u_1(t)$ is the total input into the motor driving the first link.

In comparison with the behavior of standard PD controllers, the stochastic optimal feedback controller shows a much better performance in tracking the given reference trajectory in configuration space of the manipulator [25].

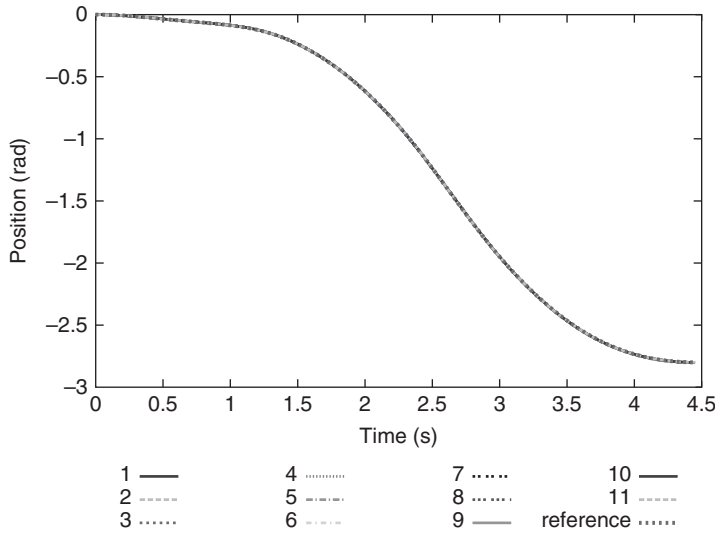


Figure 3. $q_1(t)$, $t_0 \leq t \leq t_f^{(0)}$.

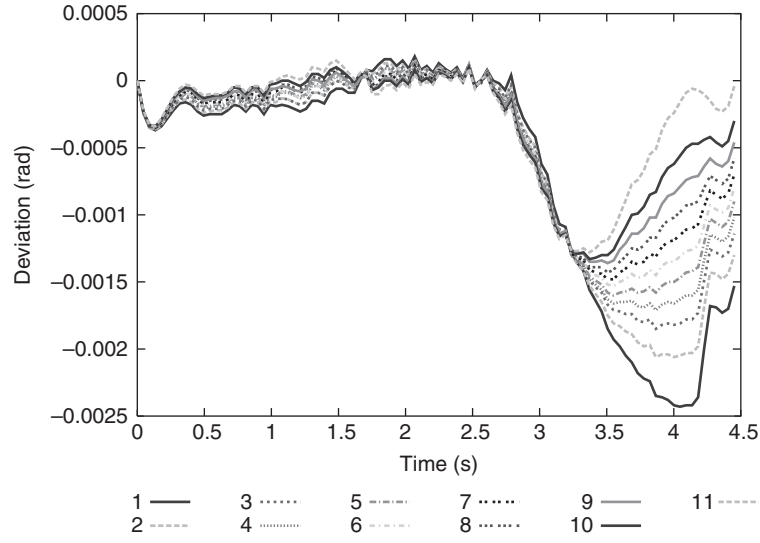


Figure 4. $\Delta q_1(t)$, $t_0 \leq t \leq t_f^{(0)}$.

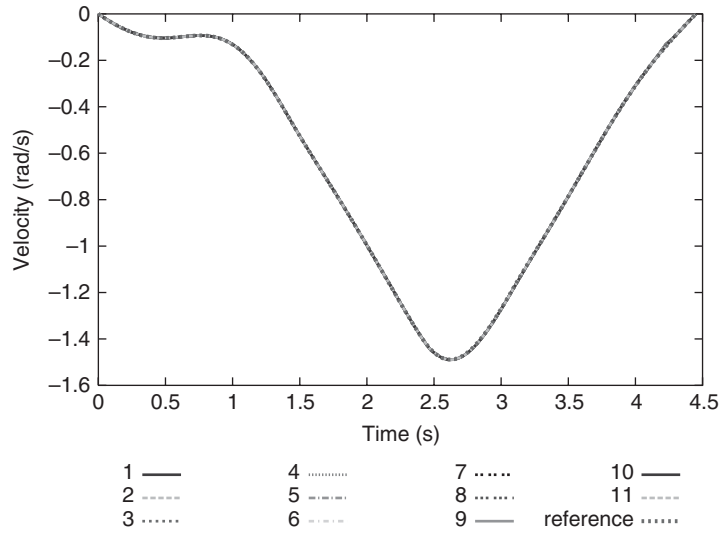


Figure 5. $\dot{q}_1(t)$, $t_0 \leq t \leq t_f^{(0)}$.

CONCLUSION

Approximation and numerical solution procedures have been presented for continuous-time optimal control problems with a convex cost criterion, a nonlinear random process differential equation, a random initial state, and deterministic or stochastic controls. Using

a parametric approach, the solution of the underlying dynamic equation represented by a system of first-order differential equations with random parameters has been considered first. Since closed-loop (CL) (feedback) controls can be approximated in practice very efficiently by means of open-loop feedback (OLF) controls, for practical applications it

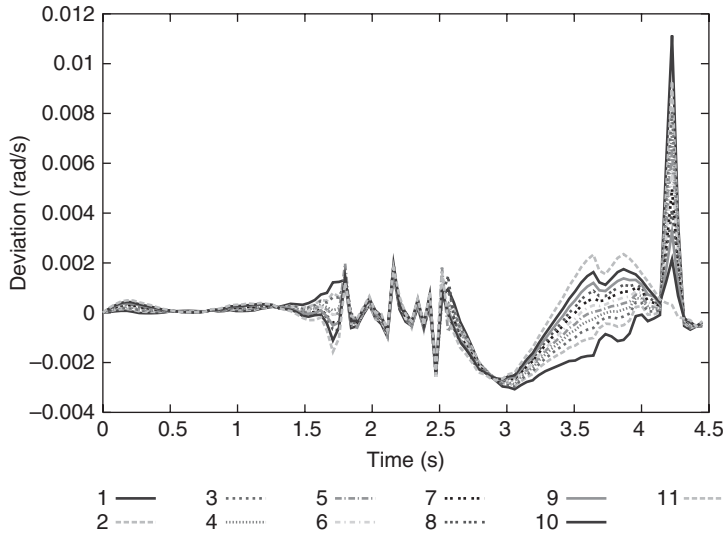


Figure 6. $\Delta \dot{q}_1(t)$, $t_0 \leq t \leq t_f^{(0)}$.

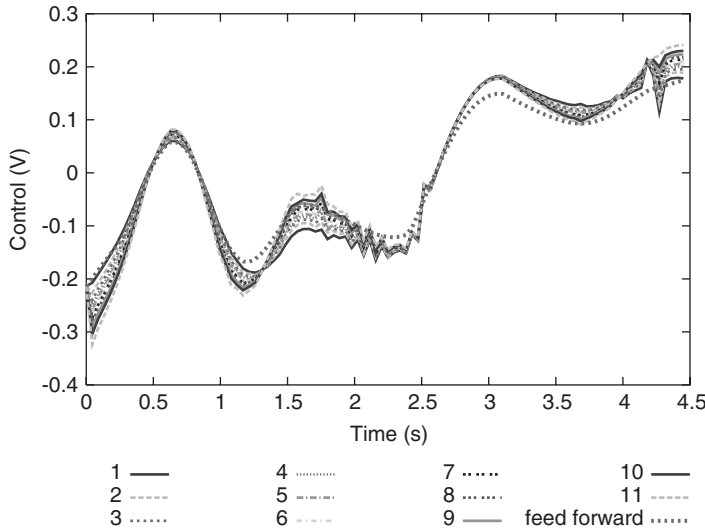


Figure 7. $u_1(t)$, $t_0 \leq t \leq t_f^{(0)}$.

is sufficient to have efficient tools for finding optimal open-loop (OL) controls on arbitrary time intervals being most insensitive with respect to random parameter variations. Thus, random parameter variations have been incorporated into the optimal control design by using stochastic optimization methods for the minimization of the (conditional) expected total costs arising along the trajectory and at the terminal point.

For finding the stochastic optimal OL controls, first convex approximations of the resulting control problem under stochastic uncertainty are obtained by an “inner” linearization of the control problem, that is, by a linearization of the basic dynamic equation. Using then the necessary and sufficient optimality conditions for the convex approximation of the original problem, stochastic optimal control laws, so-called H -minimal

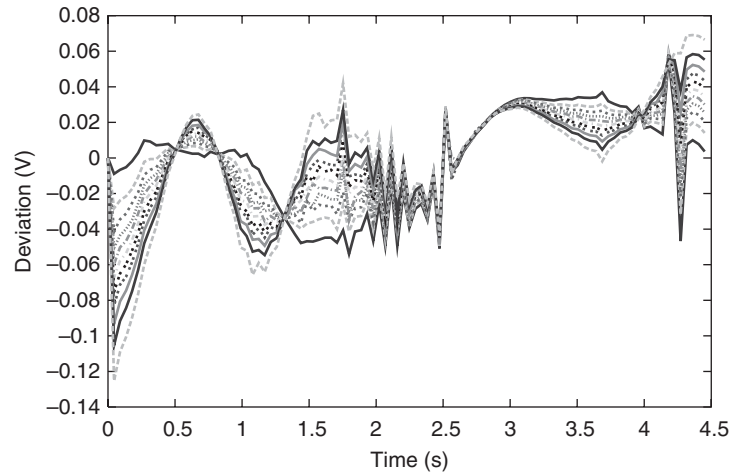


Figure 8. $\varphi_1(t, \Delta q(t), \Delta \dot{q}(t))$, $t_0 \leq t \leq t_f^{(0)}$.

controls, may be determined by solving a finite-dimensional stochastic optimization problem for the minimization of the expected Hamiltonian of the control problem subject to the remaining deterministic control constraints. Having the H -minimal controls, based on a Hamilton–Jacobi approach, stochastic optimal OL controls, may be then obtained by solving the related canonical or Hamiltonian system of first-order differential equations, which is a two-point boundary value problem with random parameters. Finally, having stochastic optimal OL controls for each remaining time interval, a stochastic optimal OLF control law follows.

Successful operation of the present method was shown for the optimal regulator design of robots under stochastic uncertainty.

Acknowledgment

The numerical results presented in this article have been obtained by Dipl. Math. Michael Schacher, Institute for Mathematics and Computer Applications, Federal Armed Forces University, Munich.

REFERENCES

1. Aström KJ. Introduction to stochastic control theory. New York: Academic Press; 1970.
2. Block C. Aktive Minderung personeninduzierter Schwingungen an weit gespannten Strukturen im Bauwesen. Fortschrittberichte VDI, Reihe 11, Schwingungstechnik, Nr. 336. Düsseldorf: VDI-Verlag GmbH; 2008.
3. Richalet J, Rault A, Testud JL, Papon J. Model predictive heuristic control: Applications to industrial processes. *Automatica* 1978;14:413–428.
4. Soong TT. Active structural control: Theory and practice. New York: Longman Scientific and Technical, John Wiley; 1990.
5. Stengel RF. Stochastic optimal control—Theory and application. New York: John Wiley; 1986.
6. Kalman RE, Falb PL, Arbib MA. Topics in mathematical system theory. New York: McGraw-Hill Book Company; 1969.
7. Marti K. Stochastic optimization problems. 2nd ed. Berlin–Heidelberg: Springer; 2008.
8. Bunke H. Gewöhnliche Differentialgleichungen mit zufälligen Parametern. Berlin: Akademie-Verlag; 1972.
9. Dieudonné J. Foundations of modern analysis. New York: Academic Press; 1969.
10. Luenberger DG. Optimization by vector space methods. New York–London: Wiley; 1969.
11. Smirnov WI. Lehrgang der Höheren Mathematik. Berlin: Deutscher Verlag der Wissenschaft; 1966.
12. Marti K. Adaptive optimal stochastic trajectory planning and control (AOSTPC) for robots. In: Marti K, Ermoliev Y, Pflug G,

- editors. Dynamic stochastic optimization. Berlin–Heidelberg: Springer; 2004.155–206.
13. Marti K. Convex approximations of stochastic optimization problems. *Meth Oper Res* 1975;20:66–76.
14. Marti K. Approximationen stochastischer Optimierungsprobleme. Königstein/Ts.: Hain; 1979.
15. Lewis FL. Applied optimal control and estimation: Digital design and implementation. Englewood Cliffs, NJ: Prentice Hall; 1992.
16. Köthe G. Topological vector spaces, Volume 159. Grundlehren der mathematischen Wissenschaften. New York: Springer; 1969.
17. Demyanov VF, Rubinov AM. Approximate methods in optimization problems. New York: American Elsevier Comp., Inc.; 1970.
18. Alt W. Nichtlineare optimierung. Braunschweig/Wiesbaden: Vieweg; 2002.
19. Kantorovich LV, Akilov KP. Functional analysis in normed spaces. New York: Pergamon; 1964.
20. Whittle P. Optimal control—basics and beyond. Chichester: John Wiley; 1996.
21. Spall JC. Introduction to stochastic search and optimization: Estimation, simulation and control. Hoboken, NJ: John Wiley; 2003.
22. Ermoliev Yu, Wets RJ-B. Numerical techniques for stochastic optimization. Berlin: Springer; 1988.
23. Press SJ. Applied multivariate analysis. New York: Holt, Rinehart and Winston, Inc.; 1972.
24. Marti K. Stochastic optimization methods in robots adaptive control of robots. In: Grötschel M, *et al.*, editors. Online optimization of large scale systems. Berlin–Heidelberg–New York: Springer; 2001.545–577.
25. Schacher M, Marti K. Stochastic optimal control of robots under stochastic uncertainty: The stochastic Hamiltonian. Lisbon, Portugal: ISSMO/APMTAC. WCSMO-8, 2009, ISBN 978-989-20-1554-5, paper 1565.
26. Türk S. Dynamische Robotermodelle am Beispiel des Manutec R3, Technischer bericht. Oberpfaffenhofen: DFVLR, Institut für Dynamik der Flugsysteme; 1988.
27. Deuffhard P. Newton methods for nonlinear problems. Berlin: Springer; 2004.
28. Aurnhammer A, Marti K, Schacher M. OSTP—Fortran-Program for optimal stochastic trajectory planning for robots, Version 0.5. Neubiberg/Munich: Federal Armed Forces University Munich, Institute for Mathematics and Computer Applications; 2009.
29. Kick W, Möller G, Rulka W. SIMPACK—ein Programm zur Simulation von Mehrkörpersystemen. Düsseldorf: VDI Verlag; 1989.
30. Marti K. Approximate solutions of stochastic control problems by means of convex approximations. In: Topping BHV, Papadrakakis M, editors. Proceedings of the 9th Int. Conference on computational structures technology (CST08). Stirlingshire, UK: Civil-Comp Press; 2008, paper No. 52.

CONTINUOUS-TIME MARTINGALES

DENIS SAURE

University of Pittsburgh, Department
of Industrial Engineering,
Pittsburgh, Pennsylvania

Modern martingale theory, introduced in large part by Doob [1], is a central component of the general theory of stochastic processes. Continuous-time martingales, the continuous-time analog of discrete-time martingales, are a fundamental tool in areas such as queueing, reliability, and stochastic calculus [2], to name but a few.

The classical example of a continuous-time martingale is the standard Brownian motion (a continuous-time version of a symmetric random walk; see **Brownian Motion**).

In fact, one of the main results in martingale theory, the Martingale representation theorem, asserts that many continuous-time martingales can be written as stochastic integrals with respect to the standard Brownian motion. In the sequel, we first define the concept of continuous-time martingale, followed by classical examples of such processes. Finally, we briefly present some of the main results in continuous-time martingale theory.

Definition. Consider a stochastic process $\{X(t) : t \geq 0\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A collection $\{\mathcal{F}_t : t \geq 0\}$ of sub- σ -fields of $\mathcal{F}$ with the property that $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$ is called a *filtration*.

Here, $\mathcal{F}_t$ represents all of the information available at time $t \geq 0$. We say $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to the filtration $\{\mathcal{F}_t : t \geq 0\}$, if it satisfies the following three conditions:

- (i) $X(t)$ is $\mathcal{F}_t$ -measurable, for all $t \geq 0$,
- (ii) $\mathbb{E}[|X(t)|] < \infty$, for all $t \geq 0$, and
- (iii) $\mathbb{E}[X(t)|\mathcal{F}_s] = X(s)$, for all $0 \leq s \leq t$.

While conditions (i) and (ii) simply impose some regularity, condition (iii) essentially states that the best “prediction” about the state of the process at time t considering the information available at time $s \leq t$ is precisely its state at time s , that is, $X(s)$.

In many situations of interest, the filtration to be considered is that generated by the process itself, that is, $\mathcal{F}_t^X := \sigma\{X(s) : s \leq t\}$ (or by its *augmented* counterpart).

Remark. Note that, in general, martingales do not need to be indexed by $t \in \mathbb{R}_+$ and the definition applies to the case where $t \in T$ for some ordered set T .

In particular, by setting $T = \mathbb{N}$, we recover the definition of discrete-time martingales.

EXAMPLES

Example 1: Compensated Poisson Process. Let $\{N(t) : t \geq 0\}$ denote a Poisson process with a homogeneous rate $\lambda > 0$ equipped with the filtration generated by itself (i.e., $\{\mathcal{F}_t^N : t \geq 0\}$). Define $X(t) := N(t) - \lambda t$, for all $t \geq 0$. Note that $X(t)$ is $\mathcal{F}_t^N$ -measurable, that $\mathbb{E}[|X(t)|] \leq 2\lambda t$, for all $t \geq 0$, and that

$$\begin{aligned} \mathbb{E}[X(t)|\mathcal{F}_s^N] &= \mathbb{E}[N(t) - N(s) + N(s) - \lambda t|\mathcal{F}_s^N] \\ &= \mathbb{E}[N(t) - N(s)|\mathcal{F}_s^N] + \mathbb{E}[N(s) - \lambda t|\mathcal{F}_s^N] \\ &= \lambda(t - s) + N(s) - \lambda s = X(s), \end{aligned}$$

for all $0 \leq s \leq t$, where we have used the independent- and stationary-increments properties of the Poisson process. We conclude that the process $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to $\{\mathcal{F}_t^N : t \geq 0\}$.

Example 2: Standard Brownian Motion. The standard Brownian motion is the classical example of a continuous-time martingale. Consider the *standard Brownian filtration* $\mathcal{F}_t^B := \sigma\{B(s) : s \leq t\}$, where

$\{B(t) : t \geq 0\}$ denotes a standard Brownian motion.

The fact that $B(t)$ is $\mathcal{F}_t^B$ -measurable is evident; by Jensen's inequality we have $\mathbb{E}[|B(t)|]^2 \leq \mathbb{E}[B(t)^2] = t < \infty$ for all t , and, if $s < t$, then

$$\begin{aligned}\mathbb{E}[B(t)|\mathcal{F}_s^B] &= \mathbb{E}[B(t) - B(s)|\mathcal{F}_s^B] + \mathbb{E}[B(s)|\mathcal{F}_s^B] \\ &= 0 + B(s),\end{aligned}$$

where we have used the independent- and stationary-increments properties of the standard Brownian motion, and the fact that $B(s)$ is $\mathcal{F}_s^B$ -measurable. Thus, we have that the standard Brownian motion is a continuous-time martingale with respect to the standard Brownian filtration.

Example 3: A Second Brownian Martingale. Consider the process $\{X(t) : t \geq 0\}$, with $X(t) := B(t)^2 - t$ for all $t \geq 0$, where $\{B(t) : t \geq 0\}$ denotes a standard Brownian motion, and consider the standard Brownian filtration. $X(t)$ is trivially $\mathcal{F}_t^B$ -measurable; we can check that $\mathbb{E}[|X(t)|] \leq \mathbb{E}[B(t)^2] + t = 2t < \infty$ for all $t \geq 0$, and that

$$\begin{aligned}\mathbb{E}[X(t)|\mathcal{F}_s^B] &= \mathbb{E}[(B(t) - B(s))^2 - t|\mathcal{F}_s^B] \\ &= \mathbb{E}[(B(t) - B(s))^2|\mathcal{F}_s^B] + 2\mathbb{E}[(B(t) - B(s))B(s)|\mathcal{F}_s^B] + \mathbb{E}[B(s)^2 - t|\mathcal{F}_s^B] \\ &= t - s + 0 + B(s)^2 - t = X(s),\end{aligned}$$

for all $0 \leq s \leq t$.

We conclude that $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to the standard Brownian filtration.

Example 4: Exponential Martingale for Brownian Motion. The following process, known as the *exponential martingale for Brownian motion*, has played a key role in modeling the evolution of asset prices in finance. For $\theta > 0$, set $X(t) := \exp(\theta B(t) - \theta^2 t/2)$ for all $t \geq 0$, where $\{B(t) : t \geq 0\}$ represents a standard Brownian motion.

Note that $\mathbb{E}[|X(t)|] = \mathbb{E}[\exp(\theta^2 t/2 - \theta^2 t/2)] = 1$, for all $t \geq 0$, and that

$$\begin{aligned}\mathbb{E}[X(t)|\mathcal{F}_s^B] &= X(s)\mathbb{E}[\exp(\theta(B(t) - B(s)) - \theta^2(t-s)/2)|\mathcal{F}_s^B] \\ &= X(s),\end{aligned}$$

where we have used the independent- and stationary-increments properties of the standard Brownian motion.

We conclude that $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to the standard Brownian filtration.

Example 5: Doob's Martingale. Let Y be a real-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[|Y|] < \infty$.

For a given filtration $\{\mathcal{F}_t : t \geq 0\}$, define $X(t) := \mathbb{E}[Y|\mathcal{F}_t]$ for $t \geq 0$. We can show that $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to $\{\mathcal{F}_t : t \geq 0\}$.

Indeed, we have that $X(t)$ is trivially $\mathcal{F}_t$ -measurable, that

$$\mathbb{E}[|X(t)|] = \mathbb{E}[\mathbb{E}[|Y|]|\mathcal{F}_t] = \mathbb{E}[|Y|] < \infty,$$

and that, for $s \leq t$,

$$\mathbb{E}[X(t)|\mathcal{F}_s] = \mathbb{E}[\mathbb{E}[Y|\mathcal{F}_t]|\mathcal{F}_s] = \mathbb{E}[Y|\mathcal{F}_s] = X(s),$$

by the tower property of conditional expectation and the fact that $\mathcal{F}_s \subset \mathcal{F}_t$.

SOME IMPORTANT RESULTS

Here, we present some of main results for continuous-time martingales. The optional stopping theorem and the Martingale convergence theorem are covered in a separate article (see **Theory of Martingales**).

Doob's Maximal Inequality

The following result (e.g., Ref. 3, Theorem 4.2) bounds the probability that a martingale exceeds certain threshold on a given time interval.

Suppose $\{X(t) : t \geq 0\}$ is a continuous-time martingale with respect to some filtration, and that $X(t) \geq 0$ almost surely (a.s.) for all $t \geq 0$. Then, for any $\lambda > 0$ and $p > 1$ we have that

$$\lambda^p \mathbb{P} \left\{ \sup_{\{s: 0 \leq s \leq t\}} X(s) > \lambda \right\} \leq \mathbb{E}[X(t)^p].$$

The previous result is also valid in the case where $\{X(t) : t \geq 0\}$ is a non-negative sub-martingale (see **Theory of Martingales**).

Martingale Representation Theorem

The martingale representation theorem (e.g., Ref. 3, Theorem 12.3) is one of the main results in Martingale theory.

Roughly speaking, the theorem asserts that, under some regularity conditions, *any* continuous-time martingale $\{X(t) : t \geq 0\}$ with respect to the standard Brownian filtration can be written as a stochastic (Itô) integral, and that such a representation is unique.

In particular, we have that

$$X(\omega, t) = \int_0^t f(\omega, s) dB(\omega, s) \quad \text{for all} \\ 0 \leq t \leq T, \omega \in \Omega, \quad (1)$$

for a unique $\mathcal{F}^B$ -adapted process f (i.e., $f(t)$ is $\mathcal{F}_t^B$ -measurable for all $0 \leq t \leq T$) such that

$$\int_0^T f(t)^2 dt < \infty \quad a.s.$$

provided that $\mathbb{E}[X(T)^2] < \infty$.

Moreover, the theorem asserts that any such process $\{f(t) : t \geq 0\}$ defines a continuous-time martingale for $t \in [0, T]$ via (1). This result is central to the theory of stochastic integration. In finance, the martingale representation theorem is fundamental in establishing the existence of hedging strategies.

Girsanov Theorem

Let $\{f(t) : 0 \leq t \leq T\}$ be an $\mathcal{F}^B$ -adapted process such that $\int_0^T f(t) dB(t) < \infty$ a.s., and define

$$X(t) := \exp \left\{ -\frac{1}{2} \int_0^t f(s)^2 ds + \int_0^t f(s) dB(s) \right\} \\ 0 \leq t \leq T.$$

If $\mathbb{E}[X(T)] = 1$, then $\{X(t) : 0 \leq t \leq T\}$ is a continuous-time martingale with respect to $\{\mathcal{F}_t^B : t \geq 0\}$, and the process defined by

$$B'(t) := B(t) - \int_0^t f(s) ds \quad 0 \leq t \leq T$$

is a standard Brownian motion with respect to $\{\mathcal{F}_t^B : t \geq 0\}$ in the probability space $(\Omega, \mathcal{F}_T^B, Q)$, where Q is defined

as $Q(A) := \mathbb{E}[\mathbf{1}\{A\}X(T)]$, $A \in \mathcal{F}_T^B$, and $\mathbf{1}\{\cdot\}$ denotes the indicator function.

The first part of the result generalizes Example 4. The second part of the result is of critical importance in finance, as it facilitates pricing financial instruments under alternative measures, where the value of underlying assets are continuous-time martingales.

Additional Convergence Results

A special class of martingales are the uniformly integrable [4] martingales, which can guarantee to converge with probability one to a terminal integrable random variable X_∞ ; moreover, convergence is also extended to first moments. For square integrable martingales, a subset of the former, convergence is also extended to second moments, and Doob's maximal inequality can be extended (see, e.g., Ref. 2) to show that

$$\mathbb{E} \left[\sup_{\{s: 0 \leq s \leq t\}} X(s)^2 \right] \leq 4 \sup_{\{s: 0 \leq s \leq t\}} \mathbb{E}[X(s)^2] \leq 4 \mathbb{E}[X_\infty^2].$$

APPLICATIONS AND FURTHER READING

Applications

Continuous-time martingales are used to analyze phenomena in many application areas. In finance, martingale theory is used in pricing of finance derivatives; see Chapter 14 in Ref. 3 for a review of applications in this area. In queueing, martingales are used to support asymptotic approximations to queueing models; see Ref. 5 for a review of the application of martingales in this subject.

Martingales have also been applied to statistical inference (e.g., filtering problems [6], Chapter 1), which arise naturally in many areas of interest.

In Operations Research and Management Science (excluding applications to finance), martingales have been applied to analyze problems in conflict resolution [7], revenue management [8,9], supply chain management [10], forecasting [11,12], queueing [13], to name but a few.

Further Reading

Rigorous treatment of continuous-time martingales is technically complex. This article provides only a brief introduction to some of the main results: an accessible (but still rigorous) treatment of the subject can be found in Refs [14] and [15]. See also Refs [16] and [17] for an introduction to martingale concepts. We refer the reader to Ref. 2 for a detailed and rigorous treatment of the subject. References [18] and [19] cover continuous-time martingales from a pure stochastic-process perspective, and Ref. 3 does the same while presenting the fundamentals of stochastic calculus, together with a comprehensive review of its applications to finance. Reference [20] provides an alternative treatment, oriented toward the study of point process dynamics.

REFERENCES

1. Doob JL. Stochastic processes. New York: John Wiley & Sons, Inc.; 1953.
2. Karatzas I., Shreve SE. Brownian motion and stochastic calculus. 2nd ed., New York: Springer-Verlag; 1991.
3. Steele JM. Stochastic Calculus and Financial Applications. Stochastic modelling and applied probability. New York: Springer-Verlag; 2010.
4. Resnick S. A probability path. Boston (MA): Birkhäuser; 1999.
5. Pang G., Talreja R., Whitt W. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probab Surv* 2007;4:193–267.
6. Øksendal B. Stochastic differential equations: an introduction with applications. 6th ed. 2003, corr 5th printing 2010, Berlin Heidelberg New York: Springer-Verlag; 2010.
7. Watson RK. Technical note: an application of martingale methods to conflict models. *Oper Res* 1976;24(2):380–382.
8. Xu X., Hopp WJ. Technical note: price trends in a dynamic pricing model with heterogeneous customers: a martingale perspective. *Oper Res* 2009;57(5):1298–1302.
9. Feng Y., Gallego G. Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Manage Sci* 1995;41(8):1371–1391.
10. Iida T., Zipkin P. Competition and cooperation in a two-stage supply chain with demand forecasts. *Oper Res* 2010;58(5):1350–1363.
11. Beril Toktay L., Wein LM. Analysis of a forecasting-production-inventory system with stationary demand. *Manage Sci* 2001;47(9):1268–1281.
12. Oh S., Özer Ö. Mechanism design for capacity planning under dynamic evolutions of asymmetric demand forecasts. *Manage Sci* 2012: pp. 1–21, ISSN 0025-1909 (print), ISSN 1526-5501 (to be published online).
13. Glynn PW., Melamed B., Whitt W. Estimating customer and time averages. *Oper Res* 1993;41(2):400–408.
14. Williams D. Probability with Martingales. Cambridge, UK: Cambridge University Press; 1991.
15. Karlin S., Taylor H. A First Course in Stochastic Processes. 2nd ed. United States: Academic Press; 1975.
16. Pinsky A., Karlin S. An introduction to stochastic models. 4th ed. United States: Academic Press; 2011.
17. Ross S. Stochastic processes. 2nd ed. New York: John Wiley & Sons; 1996.
18. Revuz D., Yor M. Continuous martingales and brownian motion. 3rd ed. Springer; 2004.
19. Rogers LCG., Williams D. Diffusions, Markov Processes and Martingales. Volume 1 Foundations, Cambridge Mathematical Library. Cambridge, UK: Cambridge University Press; 2000.
20. Brémaud P. Point processes and queues: martingale dynamics. Springer series in statistics. New York: Springer-Verlag; 1981.

CONTRIBUTIONS TO SOFTWARE RELIABILITY WITH OR APPLICATIONS

P.K. KAPUR
RAVI KUMAR
ANSHU GUPTA
Department of Operational
Research, University of Delhi,
Delhi, India

Computer software is the single most important technology in the human world, and an excellent example of the *law of unintended consequences*, as it paves the way for the creation of the new (e.g., genetic engineering), extension of the existing (e.g., telecommunications), and demise of the older (e.g., the printing industry) technologies. Software plays a dual role. In the first role, it is a product that delivers computing potential embodied by computer hardware (e.g., mathematical software), while in the second role it is a vehicle for delivering a product (e.g., software inside the cellular phones). Today almost all businesses, industries, services, government agencies, and even our day-to-day activities are directly or indirectly affected by computing systems. For example, air traffic control, nuclear reactors, patient monitoring system in hospitals, mechanical and safety control in automobiles, on-line ticketing, industrial process control, global networking, information storing, sharing, and so on, are some of the diverse applications. Improvements in hardware performance, reflective changes in computing architectures, gigantic increase in storage capacity, and a huge variety of exotic input and output options have further increased the demand for software in the automation of complex systems and its use as a solving tool for many complex problems. Size and design complexities of the software have also increased many folds and the trend will certainly continue in future. For instance, the National Aeronautics and Space Administration's (NASA) space

shuttle flies with approximately 0.5 million lines of software code on board and 3.5 million lines of code in ground control and processing [1–3].

The computer revolution has benefited society and increased the global productivity many folds, but a major threat of this revolution is that the world has become critically dependent on them for the proper functioning and timing of all its activities. The penalty of a failure may range from inconvenience in social life to economic damage to loss of life. A total breakdown of the system functioning is observed in many cases until the fault is repaired, and even after restoration to the normal state, sometime it takes up huge time, effort, and resources to make up the losses. There are several instances when software failures have caused severe loss to the economy and to human lives. A few examples are the crash of a Boeing 727 of Mexicana airlines (1986), the overdose given to cancer patients by a massive Therac-25 radiation machine in Marietta (1985 and 1986), blackouts in the northeastern United States during the month of August, 2003, and so on [3]. The ability to design, test, and maintain software has grown reasonably but many further improvements are, however, still desired. The software development process (SDP) is a challenging task for the developers. Certainly, the most important SDP problem even now is building it to customer demands so that it is *more reliable*, *built faster*, and *built cheaper* (in general order of importance). Success in meeting these demands affects the market share and profitability of the developers. These conflicting demands have resulted in risk and intense pressure, and therefore created a strong need for a practice that can help to have a tight control over the SDP and develop software to the needs of the software market.

In the early 1970s, the software reliability engineering (SRE) discipline emerged to establish and use sound engineering principles in order to economically obtain software that were not only reliable but also worked

efficiently on real machines, bringing the SDP under the engineering umbrella. The immediate concern of SRE was aimed at scheduling and systematizing the SDP to control the various stages using its tools, methods, and processes to engineer quality software. SRE broadly focuses on quantitatively characterizing the standardized six quality characteristics defined by the International Organization for Standard/International Electrotechnical Commission (ISO/IEC): functionality, usability, reliability, efficiency, maintainability, and portability. *Software Reliability* (SR) is considered the key quality characteristic—a *must-be quality*—as it quantifies software failures—the most unwanted events, making SR a major concern to the developers and users alike. Unlike hardware, any software developed anew is put on test throughout the SDP to satisfy software quality even though it is built on the most recent and productive SRE tools and techniques. The achieved quality level by means of testing has no meaning unless it is measured quantitatively to build confidence in its level. Besides this, many decisions such as release time, resource allocation in testing, as well as those related to the post release (post delivery maintenance cost and warranty) can be made more accurately only if a quantitative measure of quality is known. All this needs measurement technologies to assess the software quality. Software reliability growth models (SRGMs) based on stochastic and statistical principles are the most important and most successful tools to assess software reliability quantitatively.

Studies in SR modeling started as early as in the 1960s. The issues-related reliability measurement arose even at the time of birth of computing systems, but because of high cost and low software and hardware productivity, the concepts of SR were at the infancy stage. Most of studies were concerned with productivity and quality of hardware systems. Haugk *et al.* [4] presented some experimental results concerning the testing of a switching system in which software was an essential part, but there was no direct approach. The first paper on SR modeling appears to have been published in 1967 by Hudson [5]. He viewed software

development as a birth-and-death process in which fault generation corresponded to birth and fault correction to death and obtained a Weibull distribution of intervals between failures. Jelinski and Moranda's work [6] is recognized as the second major step. They assumed a hazard rate for failures that was piecewise constant and proportional to the remaining fault number. The hazard rate changed at each fault correction by a constant amount, but was constant between corrections. They further proposed two variants known as the *JM Geometric model* and the *Moranda Geometric Poisson model* [7]. The 1970s was the most significant in the SR study, as some of the models [6–13] proposed in this decade later formed the basis for further research in the field and found many practical applications.

Shooman [8] viewed the hazard rate as being determined by the rate at which execution of the program results in the remaining faults being passed and the remaining faults are assumed to depend on the number of faults corrected. The correction rate was assumed to be related to the project personnel profile in time proposing several fault correction profiles: the choice depending on the particular project. Schneidewind [9,10] investigated different reliability functions—exponential, normal, gamma, and Weibull—for estimating SR and suggested the choice of the best. He also suggested that the time lag between failure detection (FD) and fault correction (FC) be determined from actual data and used to correct the timescale in forecasts. In early 1975, Musa [11] proposed an execution time model of SR. This model was later studied and applied on many real applications and yielded good results for many. Musa considered execution time in two respects: the operating time of a product delivered to the field and the cumulative execution time that had occurred during test phases of the SDP as well as during post delivery maintenance. The hazard rate was assumed to be constant with respect to operating time but a function of the faults remaining and hence the cumulative execution time. The fault removal rate (FRR) was proportional

to the fault detection rate (FDR), making consideration of debugging personnel profiles unnecessary. A calendar time component was also developed that related execution time to calendar time. Littlewood and Verrall [12] worked out an SRGM based on Bayesian approach. The authors modeled the hazard rate as a random process for the failures experienced.

Goel and Okumoto [13] proposed an SRGM (called the *GO model*) that describes the failure detection as a nonhomogeneous Poisson process (NHPP) assuming that the hazard rate is proportional to the remaining fault number. This paper was a pioneering attempt in SR modeling. Later researchers proposed many SRGMs that describe the FD and/or FC process by NHPP following the basic assumption of the GO model and the research is still continuing. Starting from the late 1960s and approximately in the past 40 years, a vast literature of SR models based on NHPP has been developed. These models have been widely studied and applied on real software projects. Considering the dynamic software testing and debugging (ST&D) process, many new concepts of SR modeling have been studied and incorporated. In this article, we will give an overview of many SRGMs and discuss a few unified approaches for developing the SRGM. In the later sections of the article, we will also give an overview of a few operations research (OR) applications of SR modeling.

Notation

$m(t)$:	expected number of failure/removal by time t , with $m(0) = 0$
$m_f(t), m_{if}(t)$:	expected number of software (or module i /type i) failure by time t ; $m_f(0) = 0$
$m_r(t), m_{ir}(t)$:	expected number of software (or module i /type i) fault removals by time t ; $m_r(0) = 0$
$m_i(t), m_{ii}(t)$:	expected number of software fault (or module i /type i) isolations by time t ; $m_{ii}(0) = 0$
$a(a_i)$:	initial error content in software (or module i /type i) before software testing

$b(t), b_i(t)$:	time dependent FDR/FRR per remaining faults
$b(b_i)$:	constant FDR/FRR per remaining faults in software (module i /type i) $0 < b < 1$
β, c :	constant parameter
$W(t), W_i$:	expected cumulative amount of test effort by time t with $w(t)$ as its density function
τ :	change point

SRGM: AN OVERVIEW AND UNIFIED MODELING APPROACH

A number of instances have demonstrated that execution time is superior to the calendar time [14]. However, since the quantities in terms of calendar time component are more meaningful to engineers and managers, most of the models were developed on the calendar time component. Most of the earlier models assumed that, whenever an attempt is made to remove a detected fault, it is removed perfectly and no new faults are generated (perfect debugging environment). The earliest model in this category was the GO model [13]. It is based on the following simple assumptions:

1. Failure Observation phenomenon is modeled by NHPP.
2. Failures are observed during execution caused by the remaining faults in the software.
3. An immediate effort is made to find the cause of detected failures and the isolated faults are removed prior to future test occasions.
4. All faults in the software are mutually independent.
5. The debugging process is perfect and no new fault is introduced during debugging.

The above assumptions can be described mathematically with the following differential equation:

$$d/dt(m(t)) = b(a - m(t)), \quad (1)$$

which under the initial condition

$$m(0) = 0 \quad \text{yields} \quad m(t) = a(1 - e(-bt)). \quad (2)$$

The model is known as the exponential NHPP model, as it describes an exponential failure curve, and has been applied to a variety of testing environments in practice as it provides good estimation and prediction of reliability. Following the GO model, other exponential SRGMs have been proposed by Ohba [15] and Yamada and Osaki [16]. Ohba assumed that the software consists of a number of independent modules, whereas Yamada and Osaki assumed that there are two types of errors in the software. These models describe the failure phenomenon for each module/error type by the GO model with different parameters, and the mean value function (MVF) is the sum of the MVFs for each module/error type. The early models were exponential, accounting the uniform operational profiles, which may be unrealistic in many real testing profiles. For nonuniform profiles, many researchers later attempted to develop models describing S-shaped failure curves, attributing the S-shapedness to the distinctive explanations.

First, Yamada *et al.* [17] refined the GO model describing testing as a two-stage process, namely FD and FC, and assuming that fault removal follows failure detection with a time lag

$$\begin{aligned} d/dt(m_f(t)) &= b(a - m_f(t)) \quad \text{and} \\ d/dt(m_r(t)) &= b(m_f(t) - m_r(t)). \end{aligned} \quad (3)$$

Solution of the above is

$$\begin{aligned} m_f(t) &= a(1 - e(-bt)) \quad \text{and} \\ m_r(t) &= a(1 - (1 + bt)e(-bt)). \end{aligned} \quad (4)$$

S-shaped SRGMs [18–20] have similar mathematical forms but are developed under a different set of assumptions. An important characteristic of most S-shaped models is that these models can describe both exponential and S-shaped growth curves depending on the parameter values and therefore termed as flexible models. SRGM, which uses calendar time or execution time as the

unit of FD and/or FC period, either assumes a constant or does not explicitly consider the testing effort and its effectiveness. The reliability growth during testing is highly related to the amount of resources (test efforts) spent on detecting and correcting latent faults. A testing effort function describes the distribution of testing resources (CPU time, manpower, number of executed test cases, etc.) during the testing period. Many researchers [21–26] have proposed SRGMs describing the relationship among the testing time, testing effort, and the number of software faults detected. The time-dependent behavior of test-effort expenditures has been described by exponential, Rayleigh, Weibull, logistic, or generalized logistic functions in the literature. Following the GO model, the first NHPP-based SRGM with respect to test effort intensity $w(t)$ is as follows from Ref. 27:

$$d/dt(m(t))/w(t) = b(a - m(t)). \quad (5)$$

The MVF of the model is given as

$$m(t) = a(1 - e(-bW(t))). \quad (6)$$

In applications, first, the best testing effort function is fitted to the observed test effort data and then the parameters of the SRGM for the FD and/or FC process are determined with reference to the estimated test effort data. Later, the effect of testing efforts were studied on many other existing models.

One aspect of major importance to the SR modeling study is testing efficiency. In practice, testing efficiency is usually imperfect. During debugging, the debuggers can remove a fault perfectly; it may be repaired imperfectly, which causes failure again [called *imperfect fault removal* (IFR)], or the original fault is removed perfectly but a new one may be generated (called *fault generation* (FG)), producing some other kind of failure when checked on the same input on removal of the original fault. IFR causes more failures but the fault content remains the same; however, due to FG, the number of failures increases because of increased fault content. The concept of testing efficiency was first introduced by Goel and Okumoto [28] in the JM model.

Kapur and Garg [29] assumed that the FRR per remaining faults is reduced by imperfect debugging in the GO model, that is,

$$\begin{aligned} d/dt(m_r(t)) &= pb[a - m_r(t)] \quad \text{and} \\ d/dt(m_f(t)) &= b[a - pm_f(t)]. \end{aligned} \quad (7)$$

Solution of Equation (7) under the initial conditions $m_r(0) = 0$ and $m_f(0) = 0$ is

$$\begin{aligned} m_r(t) &= a(1 - e^{-bpt}) \quad \text{and} \\ m_f(t) &= (a/p)(1 - e^{-bpt}). \end{aligned} \quad (8)$$

If $p = 1$, the model reduces to the GO model with $m_r(t) = m_f(t)$.

Obha and Chou [30] proposed the first NHPP SRGM incorporating the effect of FG, defined as

$$\begin{aligned} d/dt(m(t)) &= b(t)[a(t) - m(t)] \quad \text{with} \\ b(t) &= b, a(t) = a + \alpha m(t). \end{aligned} \quad (9)$$

Hence the MVF of the SRGM is

$$m(t) = (a/1 - \alpha)(1 - e(b(1 - \alpha)t)). \quad (10)$$

Yamada *et al.* [31] proposed an exponential and a linear function for the FG ($a(t) = ae^{\alpha t}$ and $a(t) = a(1 + \alpha t)$), respectively. Pham *et al.* [32] incorporated the effect of learning on ST&D teams in imperfect debugging models. Zang *et al.* [33] integrated the two types, modeling it on the number of failures experienced. However, the FG rate is expected to be proportional to the FC rates, and hence later Kapur *et al.* [34] comprehensively integrated the two types of imperfect debugging, clearly illustrating the concept of imperfect debugging.

A few other interesting and important aspects of ST&D process studied and incorporated by the several researchers are complexity of faults, measuring reliability in operational phase, fault dependency and debugging (FD&D) time lag, distributed development environment (DDE) modeling, change point analysis, etc. Some researchers tried to study each aspect individually, while others have integrated distinct aspects into single model.

Faults may require different amounts of testing effort and strategy for removal. Yamada and Osaki [16] modified the GO model, defining faults in two categories and assuming different FDR/FRR. The MVF of their SRGM is

$$m(t) = \sum_{i=1}^2 m_i(t) = a \sum_{i=1}^2 p_i(1 - e^{-b_i t}). \quad (11)$$

Here, subscript i defines the fault type and it is expected that $0 < b_2 < b_1 < 1$. Kareer *et al.* [35] modified this model assuming a time-dependent FRR for both types of faults. The FC phenomenon of simple faults is described by an exponential SRGM, while an S-shaped SRGM is used for hard faults. Kapur *et al.* [36] introduced a generalized Erlang SRGM, classifying the faults as simple, hard, and complex. It is assumed that the time delay between the FD&FC represents the complexity of the faults. The more complex the fault, the more the time delay. The model has been extended to n types of faults. For simple faults, it is assumed that the removal follows immediately after detection and the GO model is used to describe their MVF. FD&FC is modeled as a two-stage process for hard faults, and delayed S-shaped SRGM [17] is used for their MVF. FD&FC for complex faults is modeled as a three-stage process (Erlang growth curve), namely, observation, isolation, and removal, and hence the model is formulated as

$$\begin{aligned} dm_{3f}(t)/dt &= b_3[a_3 - m_{3f}(t)] \\ dm_{3i}(t)/dt &= b_3[m_{3f}(t) - m_{3i}(t)] \\ dm_{3r}(t)/dt &= b_3[m_{3i}(t) - m_{3r}(t)] \\ m_3(t) &= m_{3r}(t) = a_3(1 - (1 + b_3 t \\ &\quad + (b_3^2 t^2 / 2))e^{-b_3 t}). \end{aligned} \quad (12)$$

Failure curves for hard and complex faults behave similar to the simple faults in the steady-state, and MVF of the SRGM describing the joint effect of the different types of faults is

$$m(t) = m_1(t) + m_2(t) + m_3(t), a_1 + a_2 + a_3 = a. \quad (13)$$

Such a modeling enables the management to effectively plan and control the testing strategy to tackle each type of fault. Some SRGMs also integrated the effect of learning of the ST&D teams, change point, and FD&D time lag in the Erlang Model [37,38].

The models discussed so far have been formulated under a common assumption that the FDR/FRR is a constant/continuous function of time over the entire testing period. In practice, it can change with time as the testing progresses. For example, consider that the testing of software is started initially by a testing team and after two days of testing the management changes the team constitution by including a new member possessing better technical skills in order to speed up the testing. This reconstitution can change the behavior of the testing process completely. Such changes are often seen in the software development. There are many factors that can change during the testing process. These include the testing environment, testing strategy, complexity and size of the functions, defect density, skill, motivation and constitution of the ST&D teams, and so on. The idea can be analyzed using the *change point concept*, and the time point where the change is observed is termed as the *change point*. The idea is that the testing period is divided into time intervals, and it is assumed that during a particular interval the testing strategy and environment are more or less similar but are slightly different from the other intervals. The FDR/FRR is either assumed to be constant or a function of testing time during each interval but varies from the other intervals. Zhao [39] started the concept of change point and introduced it in hardware reliability. Some important contributions in this area are given in Refs 38,40–43. Chang [41] introduced this concept into the GO model with a single change point defining FDR as

$$b(t) = \begin{cases} b_1 & 0 \leq t \leq \tau; \\ b_2 & t > \tau \end{cases}. \quad (14)$$

The position of the change point can be judged from the actual failure data curve, using some software package such as change point analyzer, or it can be treated as an

unknown model parameter. The MVF for the SRGM is given as

$$m(t) = \begin{cases} [a(1 - e^{-b_1 t})] & 0 \leq t \leq \tau; \\ [a(1 - e^{-b_1 \tau - b_2(t-\tau)})] & t > \tau. \end{cases} \quad (15)$$

Shyur [42] defined the GO model with change point under imperfect debugging environment. The model is extended to analyze the fault complexity. Further, Huang [40] reanalyzed the Chen model with respect to generalized logistic testing effort function. Kapur *et al.* [43] studied the change point concept in the Kapur and Garg [20] model with respect to time and testing efforts. Kapur *et al.* [38] have also integrated change point and fault complexity with DDE. They have also carried out testing effort control problems on some of these models.

To handle the complexities of SDP, developers tend to develop and maintain the software under DDE. Here, the SDP is realized by mapping the full set of system requirements across the various subsystems, which are later integrated to make the complete software. Development of each subsystem is distributed to various teams who develop their modules independently. Some of the modules are build on the existing software/modules which are modified and extended to engineer the new release, while some of them are wholly new components. The role of SRGM is still important in controlling and managing the development of quality software. The first attempt in SR modeling under DDE was due to Yamada *et al.*, [44] who assumed that the software consists of n used and m newly developed components. Failure phenomenon for the used component was described by GO model, while the new components used the S-shaped [17] model. The MVF for the software system is the sum of the MVFs of its entire components, that is,

$$m(t) = a \left[\sum_{i=1}^n p_i (1 - e^{-b_i t}) + \sum_{j=1}^m p_{n+j} \times \left\{ 1 - \{1 + b_{n+j} t\} e^{-b_{n+j} t} \right\} \right]. \quad (16)$$

Kapur *et al.* [45] extended the analysis of DDE and formulated flexible testing coverage and test-effort-based SRGM for DDE.

A conventional assumption of many SRGMs is that the detected faults are immediately removed. ST&D is a time-consuming and expensive process. The time to remove faults depends on their complexity, skills of the ST&D team, the software development environment, and so on. In the general, fault removal may take a longer time after detection, with more removals in the later phases of testing. Therefore, the time delay in the FC process after the FD process cannot be ignored. Besides this, Ohba [18] and Kapur and Garg [20] visualized that software faults are of two types: mutually independent (removed on the detection of a failure) and mutually dependent (gets removed during the removal of some leading faults). Mutually dependent faults can be removed if the faults leading to them are removed. Schneidewind [10] first modeled the fault correction process using a constant delayed fault detection process. Later Xie and Xhao [46] extended this by assuming a time-dependent delay function. Delayed S-shaped model [17] and fault complexity models also consider the time lag between FD and FC. Kapur and Younes [47] analyzed the SR considering the FD&D time lag. Huang and Lin [48] described FD by the GO model, and assuming $\phi(t)$ to be the delay factor, the FC process was given by

$$m_r(t) = m_f(t - \phi(t)) = a(1 - e^{-b(t - \phi(t))}). \quad (17)$$

Various conventional models are derived using different forms of $\phi(t)$. For example, if $\phi(t) = 0$, we obtain the GO model; for $\phi(t) = (1/b)\ln(1 + bt)$ and $\phi(t) = (1/b)\ln((1 + \beta)e^{-bt}/1 + \beta e^{-bt})$ the SRGMs in Refs 17 and 18 respectively, are obtained. Singh [49] incorporated the learning phenomenon using a power function for FDR/FRR in Ref. 14. Xie *et al.* [50] and Wu *et al.* [51] proposed the SRGM claiming that fault-correction SRGM can be defined for the existing fault detection SRGM using different forms of delay functions. With

intensity $\lambda_f(t)$, $m_f(t)$ is given as

$$m_f(t) = \int_0^t \lambda_f(s) ds \quad (18)$$

and with a delay function $\Delta(t)$, MVF of the removal process is defined as

$$m_r(t) = \int_0^t \lambda_r(s) ds = \int_0^t E[\lambda_f(s - \Delta(s))] ds. \quad (19)$$

Assuming $\lambda_f(t)$ as in the GO model, various SRGMs were obtained assuming constant, time-dependent, and random correction times.

Besides modeling for the testing phase, a few researchers have extended their SRGMs to the operational phase to compute the remaining fault number at release time and field failure rate to track the system performance. It is stated that, however, close we claim the test environment is to the operational environment, it cannot be a mirror image. Along with this, removals in this phase can be non instantaneous or deferred because of several reasons. Kapur *et al.* [52] claimed that for SR analysis in operational phase, software must be categorized in two types: *project type* (special purpose) or *product type* (general-purpose). During the testing phase, the software type does not affect the reliability growth. However, the type of software also influences the reliability growth during the operational phase. Tracking the system performance is not difficult for the project-type software since it used on a few readily monitored systems under almost constant usage rate. However, it is difficult for the product-type software due to the wide distribution of the software where the usage typically builds to several thousand independent systems. Kenny [53] argued that the number of failures in the operational phase is strongly influenced by the number of users and proposed a model that includes a factor for the usage rate during the operational phase. Under the assumption that the average number of failures is a function of the number of encountered defects, which is a function of the number of instructions executed up to

the testing time t , the model is

$$\begin{aligned} dm/dt &= (dm/dx)(dx/de)(de/dt) \\ &= b_1^* b_2 r^* b_3 t^k. \end{aligned} \quad (20)$$

Equation (20) yields

$$\begin{aligned} m(t) &= a \left(1 - e \left(-b \left(t^{k+1}/(k+1) \right) \right) \right), \\ b &= b_1 b_2 b_3, \end{aligned} \quad (21)$$

where a is the number of faults at the time of release of the software and r is the remaining number of faults during the field use. This model uses a power function to describe the software usage, which grows as the number of software users increases, particularly for product-type software. In the marketing literature, power function is seldom used for describing the user growth over time. Kapur *et al.* [54] redefined the Kenny's model using the Bass [55] model of innovation diffusion to describe the user growth. They further modified their model under imperfect debugging environment and linking appropriate usage function to both types of software [52].

Apart from all the models discussed above, there are several other SR modeling interests that have gained popularity in recent years, although they have quite older origins such as stochastic differential equations based models, reliability estimation with neural networks, discrete SRGM, and so on. The existing research NHPP models are formulated considering diverse ST&D environments and have been applied successfully to various testing environments but not in the general case: that is, a particular SRGM cannot be applied in general, as the physical interpretation of the ST&D is not general. A solution to this problem is to develop unified modeling (UM) approaches. For any particular application, one needs to study several models and then decide the most appropriate ones. The selected models are then compared on the basis of the results obtained, and a model is selected. Alternatively, following a unification approach several SRGMs can be obtained from a single approach, giving an insightful investigation for these models without making many distinctive assumptions. It can make the task of model

selection and application much simpler. Establishment of a unification methodology is one of the recent topics of study in this area.

Some researchers have worked on formulating generalized approaches. Work in this area started with Shanthikumar [56], who proposed a generalized birth process model to describe both the JM and GO models. Langberg and Singpurwalla [57] showed that several SR models can be viewed from the Bayesian point of view. Miller [58] extended their idea such that the MVF in NHPP models can be characterized by the pdf of fault detection time. Haung *et al.* [59] discussed a unified scheme for discrete time NHPP models by applying the concept of means. In 2009, Kapur *et al.* [60,61] proposed two approaches—first based on infinite server queueing theory and second characterizing MVF from failure distribution considering testing efficiency—that are major contributions in the field. Several existing and new SRGMs are obtainable from these approaches.

UNIFIED MODELING APPROACH

Unified Scheme Based on the Concept of Infinite Server Queue

Assuming immediate repair action, FD&FC can be viewed as an infinite server queue (ISQ). Inoue and Yamada [62] first applied this concept to delayed S-shaped SRGM [17] and, by considering the time distribution of the fault isolation (successive to detection) process, obtained several NHPP models. Here, no consideration is made for the FC phenomenon. The approach in Ref. 60 is based on the basic assumptions of an SRGM defining the three levels of fault complexity [36]. An observed failure forms an arrival process. The number of failures observed at the end of testing period is equivalent to number of customers in the $M^*/G/\infty$ queueing system, where the arrival process (M^*) is an NHPP and service time has a general distribution (G). The MVF for failures process is $m_f(t)$ [$m_{fi}(t)$ for type i fault $i = 1, 2, 3$ (complex, hard, and simple)] and the intensity function $\lambda(t)$ ($\lambda_i(t)$ for type i fault $i = 1, 2, 3$).

For Complex Faults. For these faults, fault isolation (FI) follows detection, which is followed by removal. Fault isolation and removal times are assumed to be independent, with a common distribution $F_1(t)$ and $G_1(t)$, respectively. Let the counting processes $\{X_1(t), t \geq 0\}$, $\{R_1(t), t \geq 0\}$, $\{N_1(t), t \geq 0\}$ represent the cumulative number of failures, faults isolated, and faults removed, respectively, up to time t .

For Hard Faults. It is assumed that fault isolation follows immediately after failure observation with no delay. Fault detection and removal times are independent common distribution $G_2(t)$ and unit function, respectively. Let the counting processes $\{X_2(t), t \geq 0\}$, $\{N_2(t), t \geq 0\}$ represents cumulative observed software failure number and the faults isolated/removed, respectively, up to time t .

For Simple Faults. It is assumed that the faults are removed as soon as they are observed. Let $\{X_3(t), t \geq 0\}$, $\{N_3(t), t \geq 0\}$ be the counting processes that represent the cumulative number of software failure observed and removed, respectively, up to time t .

Assuming that the test began at time $t = 0$, the distribution of $N_i(t)$; $i = 1, 2, 3$ is given by

$$\Pr\{N_i(t) = n\} = \sum_{j=0}^{\infty} \Pr\{N_i(t) = n | X_i(t) = j\} \times \left((m_{fi}(t))^j e^{-m_{fi}(t)} / j! \right). \quad (22)$$

If j failures are observed, then the probability that n faults are removed is given as

$$\begin{aligned} \Pr\{N_i(t) = n | X_i(t) = j\} \\ = \binom{j}{n} (p_i(t))^n (1 - p_i(t))^{j-n}; \\ i = 1, 2, 3, \end{aligned} \quad (23)$$

where $p_i(t)$ is the probability that an arbitrary fault is removed by time t , which is defined using Stieltjes convolution, and concept of the

conditional distribution of arrival times as

$$\begin{aligned} p_1(t) = \int_0^t \int_0^{t-u} G_1(t-u-v) dF_1(u) \\ \times (dm_{fi}(v)/m_{fi}(t)) \end{aligned} \quad (24)$$

$$p_2(t) = \int_0^t G_2(t-u) (dm_{f2}(v)/m_{f2}(t)) \quad (25)$$

$$p_3(t) = \int_0^t 1(t-u) (dm_{f3}(v)/m_{f3}(t)), \quad (26)$$

where $1(\cdot)$ is the unit function. Hence,

$\Pr\{N_1(t) = n\} = m_1(t)^n e^{-m_1(t)} / n!$, where

$$\begin{aligned} m_1(t) = \int_0^t \int_0^{t-u} G_1(t-u-v) dF_1(u) \\ \times dm_{fi}(v) \end{aligned} \quad (27)$$

$\Pr\{N_2(t) = n\} = (m_2(t))^n (e^{-m_2(t)} / n!)$, where

$$m_2(t) = \int_0^t G_2(t-u) dm_{f2}(u), \quad (28)$$

$\Pr\{N_3(t) = n\} = (m_3(t))^n (e^{-m_3(t)} / n!)$, where

$$m_3(t) = \int_0^t 1(t-u) dm_{f3}(u) = m_{f3}(t). \quad (29)$$

Equations (27), (28), and (29) imply $N_1(t), N_2(t), N_3(t) \sim \text{NHPP}$ with MVF $m_1(t), m_2(t), m_3(t)$, respectively. Knowing the MVF $m_{f1}(\cdot), m_{f2}(\cdot), m_{f3}(\cdot)$ and distributions $F_1(\cdot), G_1(\cdot)$, and $G_2(\cdot)$, the MVF of one-, two-, and three-stage FD&FC processes for various existing SRGMs can be derived. Now, if $\{N(t), t \geq 0\}$ is the counting processes that represent the cumulative number of software fault removal up to time t , then $N(t)$ is derived as $N(t) = N_1(t) + N_2(t) + N_3(t)$, that is,

$$\begin{aligned} N(t) = ((m_1(t) + m_2(t) + m_3(t))^n / n!) \\ \times e^{-[m_1(t) + m_2(t) + m_3(t)]}. \end{aligned} \quad (30)$$

Hence

$$m(t) = m_1(t) + m_2(t) + m_3(t). \quad (31)$$

Obtaining SRGM from the Unified Model based on Infinite Queues. Assuming $m_{fi}(t) = \alpha_i(1 - e(-b_i t))$, $i = 1, 2, 3$, $F_1(t) = 1 - e(-b_1 t)$ and $G_i(t) = 1 - e(-b_i t)$, $i = 1, 2$, $m_i(t)$, $i = 1, 2, 3$ describe the MVF of Erlang SRGM [36] for complex, hard, and simple faults. Using Equation (31), the MVF of the total FC process of this model is obtained. Similarly, we can describe various existing and new SRGMs from this UM approach [60].

Unification Describing MVF from Failure Distribution under Imperfect Debugging

Under this approach [61], the intensity function is expressed as

$$\lambda(t) = m'(t) = (a(t) - m(t)) p(F'(t)/1 - F(t)), \quad (32)$$

where the total number of faults present at any testing time, say $a(t)$, is a function of time and can be expressed as a linear function of the expected number of faults detected by time t , that is,

$$a(t) = a + \alpha m(t),$$

hence

$$m(t) = (a/1 - \alpha) \left[1 - (1 - F(t))^{p(1-\alpha)} \right]. \quad (33)$$

Equation (33) assumes immediate removal of faults on failure observation.

To incorporate the concept of the FD&FC process with a delay, the above scheme is further generalized with the two distribution functions $F(t)$ (for FD) and $G(t)$ (for FC). The failure intensity function is now expressed as

$$\lambda(t) = m'(t) = ((f \times g)(t)/(1 - (F \otimes G)(t))) \times p(a + \alpha m(t) - m(t)). \quad (34)$$

Equation (34) implies

$$m(t) = (a/1 - \alpha) \left[1 - (1 - (F \otimes G)(t))^{p(1-\alpha)} \right]. \quad (35)$$

Using the generalized schemes (Eqs 33 and 35), the MVF of several existing and new SRGMs can be obtained from different distributions $F(\cdot)$ and $G(\cdot)$.

Obtaining SRGM from the Unified Model Describing MVF by Failure Distribution under Imperfect Debugging. If we assume $F(t) = 1 - e^{-bt}$, then Equation (33) yields the [63] SRGM defining two types of imperfect debugging:

$$m(t) = (a/1 - \alpha)[1 - e(-bp(1 - \alpha)t)]. \quad (36)$$

If we assume $F(t) = 1 - e^{-bt}$ and $G(t) = 1 - e^{-bt}$, the Yamada delayed S-shaped model is obtained under two types of imperfect debugging, that is,

$$m(t) = (a/1 - \alpha) \left[1 - ((1 + bt)e(-bt))^{p(1-\alpha)} \right]. \quad (37)$$

Similarly, for distinct distribution functions $F(t)$ and $G(t)$, a wide range of models were obtained.

Kapur *et al.* [63] have also shown that the above two unified approaches, as well as that due to Xie *et al.* [50] based on FD&FC processes, are theoretically equivalent. Kapur *et al.* [64–66] have also proposed various other unification schemes in order to unify the study of other SRGMs based on distinct assumptions such as unified scheme for SDE-based SR modeling with respect to time and test effort and unified scheme for change point concept models.

Starting from the beginning, we have described various SRGMs. SRGMs find many practical applications. For example, they are used to develop test cases, schedule status, and resource optimization; to count the number of faults remaining in the software; and to estimate and predict the reliability of the software during testing and in the operational environment. Application of a model for any particular application involves first selecting some appropriate models. For any general testing environment, selection of an appropriate model is a difficult task and requires considerable effort. No one has succeeded in identifying *a priori* the characteristics of software that will ensure that a particular model can be trusted for reliability predictions [67]. Most of the tools and techniques use *trend exhibited by the data* criterion for model selection. Among the tools that rank models is the AT&T software reliability engineering toolkit. This

tool can be used only for a few software reliability growth models. Asad *et al.* [67] discussed some criteria to choose the best model; however, their findings still cannot be considered general. Application of neural network-based approaches and unification schemes provide some solution to this problem. However, a study that can justify the selection of an appropriate model for any software project is still to be done. Once some representative set of models is selected, the next step involves estimating the unknown parameter of the models from the collected failure data. Method of maximum likelihood estimate (MLE) and nonlinear least squares are the two most important and widely used estimation techniques. The third step for the application is to establish the best model applicable to the situation under consideration. This step involves two parts: The first one is comparing the estimation results of the selected models. Various criteria are used to measure the goodness of fit such as mean square error (MSE), bias, coefficient of multiple determination (R^2), Chi-square test, root mean square prediction error (RMSPE), and so on. The second part involves establishing the predictive power of the models by estimating the parameters on the truncated data and predicting it over the remaining. Relative prediction error (RPE) is used to compare these results. Most of the NHPP models are nonlinear in nature and it is difficult to manually obtain the unknown model parameters. Statistical software packages such as SPSS, SAS, Mathematica, and so on. can be used to easily obtain the model parameters with high accuracy levels (details of actual failure datasets and estimation in Refs 1–3). In the next section, we describe some OR applications of these models.

USING SRGM FOR OPTIMIZATION

SRGMs are used at many decision making points in SDP, such as software release time (SRT) determination, resource allocation (RA) and control, optimum component selection for fault tolerant software, and so on. Here, we will highlight mainly two types of OR applications, namely, SRT determination and RA, for unit testing of modular software.

Software Release Time Determination

However important achieving high reliability of software may be, no software can be tested indefinitely and made fault free. Software user requirements are often conflicting with the developer requirements, as the former want faster deliveries, low-cost, and quality products, whereas the latter want to minimize their cost and maximize profit margins meeting competitive requirements. A crucial decision problem that arises here is to know when to stop testing and release the software to the user, which is known as the SRT problem. Delayed release results in penalties, revenue and goodwill loss, and obsolescence for the developers, while a premature release may cost high maintenance and goodwill loss cost during field use. The wealth of research in the area is impressive and confirms its continuing importance as a research topic. The SRT problem became the prime area of concern among researchers around 1980. Most of the problems discussed in the literature have the objective to minimize cost of ST&D during testing and operational (T&O) phases or maximize reliability subject to budgetary constraint and/or the minimum desired reliability level to be achieved by the release time. Models that minimize the number of remaining faults or the failure intensity also fall under this category.

The first attempt in determining SRT seems to have been made by Okumoto and Goel [68], who considered unconstrained problems using the GO model to describe the FD and/or FC processes. The cost function is defined for FI&FC during T&O phases (C_1 and C_2) and cost of testing per unit time (C_3), given as

$$C(T) = C_1 m(T) + C_2 (m(T_1) - m(T)) + C_3 T, \quad (38)$$

where T is the release time. The optimal SRT is determined by minimizing the cost function using the principles of calculus. The authors have also determined SRT such that the reliability $R(x|T)$ reaches a desired level (R_0) irrespective of the cost.

A significant contribution was made by Yamada and Osaki [69], who carried a trade-off between cost and reliability. Two

problems were discussed on the basis of two exponential and one flexible SRGM [13,16,17], the first minimizing the cost subject to reliability aspiration and the second maximizing reliability subject to cost not exceeding a predefined budget C_B , that is,

Case 1. Minimize $C(T)$ subject to $R(x|T) \geq R_0$.

Case 2. Maximize $R(x|T)$ subject to $C(T) \leq C_B$.

Kapur and Garg [70] introduced the concept of releasing the software at the scheduled delivery time for which a penalty cost $C_p(t)$ in $(0, t]$ due to delay is added to the cost function (38), that is,

$$C(T) = C_1 m(T) + C_2 (m(T_L) - m(T)) + C_3 T + \int_0^T C_p(T-t) dG(t). \quad (39)$$

The SRT problem of cost minimization subject to reliability aspiration was studied on three SRGMs [13,16,17]. Kapur and Garg [71] proposed the use of a test-effort-dependent SRGM to release problems, as resources decrease during the testing life-cycle so that the cumulative testing effort approaches a finite limit. Optimal release policies are discussed by maximizing the expected gain simultaneously achieving a given level of failure intensity. Mathematically, the SRT problem with gain maximization is stated as

$$\begin{aligned} \max G(T) &= (C_2 - C_1)m(T) - C_3 W(T) \\ \text{subject to } \lambda(T) &\leq \lambda_0. \end{aligned} \quad (40)$$

Yun and Bhai [72] proposed that it is more appropriate to assume random software life-cycle accounting to alternative competitive product and better releases by the developers. They defined the unconstrained problem of maximizing the total average profit with random life-cycle solving numerically using the bisection method again for exponential and flexible SRGM. Kapur *et al.* [73] reconsidered random life-cycle by defining policies on the basis of cost minimization subject to failure

intensity aspiration. Assuming $h(t)$, $H(t)$, and $r(t)$ to be the pdf, cdf, and hazard rate of the software life-cycle length, the expected total cost is defined as

$$\begin{aligned} C(T) &= C_1 \int_0^T m(t)h(t) dt + C_3 \int_0^T th(t) dt \\ &\quad + C_1 \int_T^\infty m(T)h(t) dt + C_3 T \bar{H}(T) \\ &\quad + C_2 \int_T^\infty (m(t) - m(T))h(t) dt. \end{aligned} \quad (41)$$

In the early 1990s, Kapur and Garg [74] made an initial attempt in determining SRT based on imperfect debugging SRGM by minimizing cost subject to reliability aspiration constraint. The cost function was modified to include separate cost of fixings due to perfect and imperfect debugging during T&O phases. They defined the cost function as

$$\begin{aligned} C(T) &= (C_1 p + C'_1(1-p))m_f(T) \\ &\quad + (C_2 p + C'_2(1-p))(m_f(\infty) - m_f(T)) \\ &\quad + C_3 T. \end{aligned} \quad (42)$$

Kapur and Garg [20] modified the cost model (Eq. 38) to incorporate the cost of removal of dependent faults ($\hat{C}_1, \hat{C}_2$) with the cost of independent faults and formulated the SRT problem based on the cost-reliability criterion. The modified cost model is

$$\begin{aligned} C(T) &= C_1 m_f(T) + \hat{C}_1 (m_f(T) - m_r(T)) \\ &\quad + C_2 (m_f(T_L) - m_f(T)) + \hat{C}_2 (m_r(T_1) \\ &\quad - m_r(T) - (m_f(T_1) - m_f(T))) + C_3 T, \end{aligned} \quad (43)$$

where T_L is the software life-cycle.

Kapur *et al.* [75] proposed a test-effort-based discrete exponential model and for the first time discussed release policies minimizing cost with failure intensity aspiration constraint for discrete SRGM. Further, Kapur *et al.* [76] brought the bi criterion release time optimization into this area carrying a trade-off between cost minimization and reliability maximization subject to budget and reliability aspiration constraints for both

exponential and flexible SRGM. The problem is stated as

$$\begin{aligned}
& \text{Maximize} && R(x/T) \text{ (or } \log R(x/T)); \\
& \text{Minimize} && C(T) \text{ (or } \bar{C}(T)) \\
& \text{s.t.} && C(T) \leq C_B \text{ (or } \bar{C}(T) \leq 1); \\
& && R(x/t) \geq R_0 \quad T \geq 0; \\
& && 0 < R_0 < 1; \quad \bar{C}_T = C_T/C_B.
\end{aligned}$$

The problem is transformed to single objective using scalarization with the parameter $\lambda = \lambda_i, i = 1, 2; \lambda_1 + \lambda_2 = 1$. Such a release policy gives enough flexibility to the developers based on their priority in respect of reliability and cost components. For safety-critical projects, a higher weight ($\lambda_1 > \lambda_2$) may be attached to the reliability objective, whereas for product type more weight may be attached to the cost objective ($\lambda_2 > \lambda_1$).

Pham [77] proposed an SRGM for three-level fault complexity considering fault generation and studied release policies on this model (minimizing cost subject to reliability aspiration or maximizing reliability subject to cost constraint) incorporating penalty cost and random life-cycle in the cost function along with traditional costs. The cost model is given as

$$\begin{aligned}
C(T) = & \sum_{i=1}^3 C_{i1} m_i(T) \\
& + \int_T^\infty \left(\sum_{i=1}^3 C_{i2} (m_i(t) - m_i(T)) \right) g(t) dt \\
& + C_3 T + I(T - T_d) C_p (T - T_d), \quad (44)
\end{aligned}$$

where i denote the fault complexity level in costs C_1 and C_2 , $g(t)$ is the pdf of the life-cycle length, and $I(t)$ is an indicator function. A major contribution was made by Pham and Zhang [78], who modified the cost function incorporating warranty, risk, and correction time costs and carried unconstrained cost minimization to compute SRT. The modified function is defined as

$$\begin{aligned}
C(T) = & C_1 m(T) \mu_y + C_2 \mu_w (m(T + T_w) \\
& - m(T)) + C_3 T^\alpha + C_4 (1 - R(x|T)), \quad (45)
\end{aligned}$$

where μ_y and μ_w are the expected time to remove faults upto time T and during the warranty period, respectively, the per-unit testing time cost is taken as a power function of time T to consider the higher increases in the beginning and slow down later, and the last term denotes the risk cost.

Xie [79] studied the release policies for IFR SRGM [29] though he referred to it as *error generation SRGM* [30]. The author proposed that the cost of testing C_3 is a function of perfect debugging probability p because, if this probability is to be increased, more resources are required, which will result in an increase in C_3 . The modified cost model is given as

$$\begin{aligned}
C(T) = & C_1 m(T) + C_2 (m(\infty) - m(T)) \\
& + (C_3 / (1 - p)) T \quad (46)
\end{aligned}$$

Pham and Zang [80] studied release policies by minimizing the cost subject to reliability requirement for a test-coverage-based SRGM assuming that a risk cost is associated with uncovered code $(1 - C(T))$ remaining at the release time which is to be paid to each user (D in number). The cost model was defined as

$$C(T) = C_1 m(T) \mu_y + C_3 T + C_3 D (1 - C(T)). \quad (47)$$

Kapur *et al.* [81] modified the cost model in Ref. 79 by incorporating a separate cost of fixing due to perfect and imperfect fault debugging during T&O phases. The cost model is

$$\begin{aligned}
C(T, p) = & (C_1 p + C'_1 (1 - p)) m_f(T) \\
& + (C_2 p + C'_2 (1 - p)) (m_f(\infty) - m_f(T)) \\
& + C_3 T / (1 - p). \quad (48)
\end{aligned}$$

They also integrated the effect of two types of imperfect debugging in an exponential SRGM and modified the cost model such that the testing cost per unit testing time is taken as a function of both the probability of perfect debugging p and error generation α . Release policies minimizing cost under the reliability aspiration were considered. Gupta *et al.* [82] have also formulated and solved these policies in

discrete-times. Apart from these, many other RT studies have been described in the literature, most of them being formulated on distinct SRGMs but the major concepts remaining the same as discussed above.

Resource Allocation for Unit Testing of Modular Software

Software testing process involves three successive stages: module, integration, and system. Since the amount of available test resources is usually limited, maximum faults are removed in unit testing, which consumes almost 40–50% of total test resources, and it is very important to spend them judiciously and distribute optimally among the various modules so as to attain highest possible reliability by the release. Resource allocation problems have been studied extensively in the literature. Modules are considered independent during unit testing. Firstly, Ohetera and Yamada [83] considered two problems based on the GO model. One minimizes the mean number of remaining faults in M independent modules under the resource limitation constraint, that is,

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^M v_i a_i e^{-b_i W_i} \\ \text{s.t.} \quad & \sum_{i=1}^M W_i = W, \quad W_i \geq 0. \end{aligned} \quad (49)$$

The second one minimizes the total resource expenditure assuming remaining fault number at the termination of module testing is Z , and is formulated as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^M W_i \\ \text{s.t.} \quad & \sum_{i=1}^M v_i a_i e^{-b_i W_i} = Z, \quad W_i \geq 0. \end{aligned} \quad (50)$$

Here, v_i is the weight assigned to the i th module and test effort is not taken as a function of time since testing is continued for a finite time period. Yamada *et al.* [84] further

incorporated SR aspiration in the problem. Lagrange's technique was used to solve the problems and numerical illustrations were provided.

Leung [85] reconsidered problems in Ref. 83 by dividing the total testing time into K testing periods. In each period, failure data is recorded and at the end of period model parameters, remaining fault number are re - estimated, and then the amount of resources for the next time period for each module is determined. Their problem can be stated as follows:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^M v_i a_{ij} e^{-b_{ij} W_{ij}} \\ \text{s.t.} \quad & \sum_{i=1}^M W_{ij} = W_j, \quad W_{ij} \geq 0. \end{aligned} \quad (51)$$

The constraints ensure that the total resources allocated to all modules in j th period is W_j , where

$$W_j = ((T_{j+1} - T_j)/(T_{K+1} - T_1))W.$$

The author also considered the problem in Equation (50) in the same manner and used the Lagrange method to solve the problems.

Xie and Yang [86] discussed the allocation of the total testing time to each software module using the Lagrange technique such that operational failure intensity is minimum using the GO model formulated as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^M \lambda_i(T_i) \\ \text{s.t.} \quad & \sum_{i=1}^M T_i \leq T, \quad T_i \geq 0, \end{aligned} \quad (52)$$

where T_i is the testing time for module i and T is the total testing time. They also conducted a sensitivity analysis to determine which of the model parameters affects the system most.

Lyu and Moorsel [87] considered software consisting of M modules performing N applications. They defined the problem by minimizing the total testing time of modules such

that the sum of failure rates (λ_i) of the components (δ_i) in each application is less than a desired level ($\delta_j, j = 1, 2, \dots, N$). The problem is stated as

$$\begin{aligned} &\text{minimize } \sum_{i=1}^M T_i \quad \text{s.t.} \quad \sum_{i=1}^M \sigma_{j,i} \cdot \lambda_i \leq \delta_j; \\ &\delta_j, T_i, \lambda_i \geq 0; \quad i = 1, 2, \dots, M, \quad j = 1, 2, \dots, N \\ &\sigma_{j,i} = 1 \text{ if } A_j \text{ uses module } i; \quad 0 \text{ otherwise.} \end{aligned} \quad (53)$$

They also stated and solved the problem alternatively by minimizing the total failure intensity of modules such that the total testing time assigned to each application gets at most a specified amount (d_j), that is,

$$\begin{aligned} &\text{minimize } \sum_{i=1}^M \lambda_i \quad \text{s.t.} \quad \sum_{i=1}^M \sigma_{j,i} \cdot T_i \leq d_j; \\ &d_j, T_i, \lambda_i \geq 0; \quad i = 1, \dots, M, \quad j = 1, \dots, N. \end{aligned}$$

Huang *et al.* [88] formulated and solved three problems assigning weights to each modules based on Kuhn–Tucker conditions. The first minimizes the remaining number of faults subject to budget and reliability aspiration constraint; the second and third minimize the total resources and cost, respectively, subject to desired reliability and remaining fault level constraints. The problems are illustrated on a GO model with test efforts assuming a generalized logistic test effort function. All the previous problems have considered Weibull-type test effort functions. The cost model considered is

$$\begin{aligned} C(T) &= C_1 m(T) + C_2 (m(\infty) - m(T)) \\ &+ C_3 \int_0^T w(x) dx. \end{aligned}$$

Haung *et al.* [89] performed sensitivity analysis on the principal parameters of the models. Lo [90] performed sensitivity analysis [83] to determine which of the model parameters affect the system most.

Rajan and Misra [91] first formulated an unconstrained problem and determined the optimal number of tests needed to test each module in unit testing. Later, they also

included the budget constraint and solved the problem. All these allocation problems have been discussed on an exponential SRGM. Kapur *et al.* [92,93] and Jha *et al.* [94] studied various resource allocation problems maximizing the number of faults removed from each module under constraint on budget and/or management aspirations on reliability. They illustrated their problems by proposing a flexible SRGM incorporating test effort considering fault dependency, fault dependency and imperfect debugging, and testing coverage and debugging time lag. The problem is stated as

$$\begin{aligned} &\text{minimize } \sum_{i=1}^M m_i \\ &\text{s.t.} \quad \sum_{i=1}^M W_i = C_B, \quad W_i \geq 0. \end{aligned} \quad (54)$$

They have discussed dynamic, mathematical, and goal programming approaches to yield solutions of such classes of optimization problems.

The existing literature on SRGM well as optimization based on SRGM is very vast, all of which cannot be described here for the want of space.

REFERENCES

1. Kapur Kapur PK, Garg RB, Kumar S. Contributions to hardware and software reliability. Singapore: World Scientific; 1999. pp. 85–152.
2. Lyu MR. Handbook of software reliability engineering. McGraw-Hill; 1996.
3. Pham H. System software reliability. Rel Eng Series. London: Springer; 2006.
4. Haugk G, Tsiang G, Zimmerman L. System testing of the no. 1 electronic switching system. Bell Syst Tech J 1964; 9: 2575–2592.
5. Hudson GR. Program errors as a birth and death process. Syst Dev Corp Rep 1967;SP-3011(4):1131–1143.
6. Jelinski Z, Moranda P. Software reliability research. In: Freiberger W, editor. Statistical computer performance evaluation. New York: Academic Press; 1972. pp. 465–484.
7. Moranda P. Predictions of software reliability during debugging. Proceedings Annual

- Reliability and Maintainability Symposium. Washington (DC); 1975. pp. 327–332.
8. Shooman M. Probabilistic models for software reliability prediction. In: Freidberger W, editor. Statistical computer performance evaluation. New York: Academic Press; 1972. pp. 485–502.
 9. Schneidewind NF. An approach to software in reliability prediction and quality control. Fall Joint Computer Conference, AFIPS Conference Proceedings. Montvale (NJ): AFIPS Press. 1972. pp. 837–847.
 10. Schneidewind NF. Analysis of error processes in computer software. *Sigplan Not* 1975;10:337–346.
 11. Musa JD. A theory of software reliability and its application. *IEEE Trans Softw Eng* 1975;SE 1:312–327.
 12. Littlewood B, Verrall JL. A Bayesian reliability growth model for computer software. *Appl Stat* 1973;22:332–346.
 13. Goel AL, Okumoto K. Time dependent error detection rate model for software reliability and other performance measures. *IEEE Trans Reliab* 1979;R-28(3):206–211.
 14. Musa JD, Okumoto K. A logarithmic Poisson execution time model for software reliability measurement. *Proceedings of 7th International Conference on Software Engineering; Volume 7. Orlando; 1984. pp. 230–237.*
 15. Ohba M. Software reliability analysis models. *IBM J Res Dev* 1984;28:428–443.
 16. Yamada S, Osaki S. Software reliability growth modeling: models and applications. *IEEE Trans Softw Eng* 1985;11:1431–1437.
 17. Yamada S, Ohba M, Osaki S. S-shaped software reliability growth modeling for software error detection. *IEEE Trans Reliab* 1983;R-32(5):475–484.
 18. Ohba M. Inflection S-shaped software reliability growth models. In: Osaki S, Hatoyama Y, editors. *Stochastic models in reliability theory*. Berlin: Springer; 1984. pp. 44–164.
 19. Bittanti S, Bolzern P, Pedrotti E, *et al.* A flexible modeling approach for software reliability growth. In: Goos G, Harmanis J, editors. *Software reliability modeling and identification*. Berlin: Springer; 1988. pp. 101–140.
 20. Kapur PK, Garg RB. A software reliability growth model for an error removal phenomenon. *Softw Eng J* 1992;7:291–294.
 21. Putnam LH. A general empirical solution to the macro software sizing and estimating problem. *IEEE Trans Softw Eng* 1978;4:345–367.
 22. Huang CY, Kuo SY, Lyu MR. An assessment of testing-effort dependent software reliability growth models. *IEEE Trans Reliab* 2007;56(2):198–211.
 23. Huang CY, Kuo SY. Analysis of incorporating logistic testing effort function into software reliability modeling. *IEEE Trans Reliab* 2002;51(3):261–270.
 24. Bokhari MU, Ahmad N. Analysis of software reliability growth models: the case of log-logistic test-effort function. *Proceedings of the 7th IASTED International Conference on Modeling and Simulation; Volume 7. Montreal, Quebec, Canada; 2006. pp. 540–545.*
 25. Kapur PK, Goswami DN, Gupta A. A software reliability growth model with testing effort dependent learning function for distributed systems. *Int J Reliab Qual Saf Eng* 2004;11(4):365–377.
 26. Kuo SY, Huang CY, Lyu MR. Framework for modeling software reliability, using various testing-efforts and fault-detection rates. *IEEE Trans Reliab* 2001;50(3):310–320.
 27. Yamada S, Ohtera H, Narihisa H. Software reliability growth models with testing-effort. *IEEE Trans Reliab* 1986;R-35:19–23.
 28. Goel AL, Okumoto K. Bayesian software prediction models, Volume 1: an imperfect debugging model for reliability and other quantitative measures of software systems. Rome (NY): Rome Air Development Center; 1978. Rep. RADC-TR-78-155.
 29. Kapur PK, Garg RB. A software reliability growth model under imperfect debugging. *RAIRO* 1990;24:295–305.
 30. Ohba M, Chou XM. Imperfect debugging effect software reliability growth. *Proceedings of 11th International Conference of Software Engineering; 1989 May 15–18 Pittsburg (PA): IEEE Computer Science Society/ACM Press; 1989. pp. 237–244. ISBN 0-8186-1941-4.*
 31. Yamada S, Tokunou K, Osaki S. Imperfect debugging models with fault introduction rate for software reliability assessment. *Int J Syst Sci* 1992;23(12):2253–2264.
 32. Pham H, Nordmann L, Zang X. A general imperfect software debugging model with s-shaped fault detection rate. *IEEE Trans Reliab* 1999;48(2):169–175.
 33. Zang X, Teng X, Pham H. Considering fault removal efficiency in software reliability assessment. *IEEE Trans Syst Man Cybern A Syst Hum* 2003;33(1):114–120.
 34. Kapur PK, Kumar D, Gupta A, *et al.* On how to model software reliability growth in the

- presence of imperfect debugging and fault generation. Proceedings of 2nd International Conference on Reliability and Safety Engineering, INCREASE; 2006 Dec 18–20; Chennai, India. Volume 2. 2006. pp. 261–268.
35. Kareer N, Kapur PK, Grover PS. An S-shaped software reliability growth model with two types of errors. *Microelectron Reliab* 1990;30(6):1085–1090.
 36. Kapur PK, Younes S, Agarwala S. Generalized Erlang software reliability growth model. *ASOR Bull* 1995;35(2):273–278.
 37. Kapur PK, Kumar A, Yadav K, *et al.* Incorporating errors of different severity and change point in software reliability growth modeling. In: Kapur PK, Verma AK, editors. *Quality, Reliability and Infocom Technology*. India: MacMillan India Ltd.; 2007. pp. 605–618.
 38. Kapur PK, Singh VB, BasirZadeh M. Considering errors of different severity in software reliability growth modeling using fault dependency and debugging time lag functions. In: Verma AK, Kapur PK, Ghadge SG, editors. *Advances in performance and safety of complex system*. India: MacMillan India Ltd.; 2008. pp. 839–849.
 39. Zhao M. Change-Point problems in software and hardware reliability. *Commun Stat Theor Meth* 1993;22(3):757–768.
 40. Huang CY. Performance analysis of software reliability growth models with test efforts and change-point. *J Syst Softw* 2005;76:181–194.
 41. Chang IP. An analysis of software reliability with change-point models. Taiwan: National Science Council; 1997. NSC 85-2121-M031-003.
 42. Shyur HJ. A Stochastic software reliability model with imperfect debugging and change-point. *J Syst Softw* 2003;66:135–141.
 43. Kapur PK, Gupta A, Shatnawi O, *et al.* Testing effort control using flexible software reliability growth model with change point. *IJPE* 2006;2:245–262.
 44. Yamada S, Kimura M, Tanaka H, *et al.* Software reliability measurement and assessment with stochastic differential equations. *IEICE Trans Fundam* 1994;E&A(1):109–115.
 45. Kapur PK, Goswami DN, Kumar A. Flexible testing effort dependent software reliability growth modeling with testing coverage for distributed systems. In: Varde PV, Srividya A, Sanyasi Rao VVS, *et al.*, editors. *Reliability, safety and hazard (advances in risk-informed technology)*. New Delhi: Narosa Publications Pt. Ltd.; 2005. pp. 175–182.
 46. Xie M, Zhao M. The Schneidewind software reliability model revisited. Proceedings of the 3rd International Symposium on Software Reliability Engineering: Research Triangle Park (NC): 1992. pp. 184–192.
 47. Kapur PK, Younes S. Software reliability growth model with error dependency. *Microelectron Reliab* 1995;35(2):273–278.
 48. Huang CY, Lin CT. Software reliability analysis by considering fault dependency and debugging time lag. *IEEE Trans Reliab* 2006;35(3):436–449.
 49. Singh VB, Yadav K, Kapur R, *et al.* Considering fault dependency concept with debugging time lag in software reliability growth modeling using a power function of testing time. *Int J Autom Comput* 2007;4(4):359–368.
 50. Xie M, Hu QP, Wu YP, *et al.* A study of the modeling and analysis of software fault-detection and fault-correction processes. *Qual Reliab Eng Int* 2007;23:459–470.
 51. Wu YP, Hu QP, Xie M, *et al.* Modeling and analysis of software fault detection and correction process by considering time dependency. *IEEE Trans Reliab* 2007;56(4):629–642.
 52. Kapur PK, Gupta A, Jha PC. Reliability analysis of project and product type software in operational phase incorporating the effect of fault removal efficiency. *Int J Reliab Qual Saf Eng* 2007;14(3):219–240.
 53. Kenny GQ. Estimating defects in a commercial software during operational use. *IEEE Trans Reliab* 1993;42(1):107–115.
 54. Kapur PK, Jha PC, Goswami DN, *et al.* General framework for modeling software reliability growth during testing and operational use. In: Kapur PK, editor. *Recent developments in quality, Reliability and IT*. Delhi: IMH publisher; 2003. pp. 187–197.
 55. Bass FM. A new product growth model for consumer durables. *Manag Sci* 1969;15:215–227.
 56. Shanthikumar JG. Software reliability models: a review. *Microelectron Reliab* 1983;23:903–943.
 57. Langberg N, Singpurwalla ND. Unification of some software reliability models. *SIAM J Comput* 1985;6:781–790.
 58. Miller DR. Exponential order statistic models of software reliability growth. *IEEE Trans Softw Eng* 1986;SE-12:S12–S24.
 59. Huang CY, Lyu MR, Kuo SY. A unified scheme of some non-homogeneous Poisson process models for software reliability estimation. *IEEE Trans Softw Eng* 2003;29:261–269.

60. Kapur PK, Anand S, Inoue S, *et al.* A unified approach for developing software reliability growth model using infinite server queuing model. Working paper. 2009.
61. Kapur PK, Pham H, Anand S, *et al.* A unified approach for developing software reliability growth models in the presence of imperfect debugging and error generation. IEEE Trans Softw Reliab 2009. In press.
62. Inoue S, Yamada S. A software reliability growth modeling based on infinite server queuing theory. Proceedings of the 9th ISSAT International Conference on Reliability and Quality in Design; Honolulu, Hawaii: 2003. Volume 9. pp. 305–309.
63. Kapur PK, Aggarwal AG, Anand S. A new insight into software reliability growth modeling. Int J Performability Eng 2009;5(3):267–274.
64. Kapur PK, Kumar J, Kumar R. A unified modeling framework incorporating change point for measuring reliability growth during software testing. OPSEARCH 2008; 45(4):317–334.
65. Kapur PK, Basirzadeh M, Inoue S, *et al.* Stochastic differential equation based SRGM for errors of different severity with testing-effort. Communicated; 2009.
66. Kapur PK, Anand S, Yadav K, *et al.* A Unified Scheme for Developing Software Reliability Growth Models Using Stochastic Differential Equation. Communicated; 2009.
67. Asad CA, Muhammad UI, Muhammad RJ. An approach for software reliability model selection. Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04); 2004 Sep 28–30; Hong Kong. 2004. pp. 534–539. ISBN: 0-7695-2209-2.
68. Goel AL, Okumoto K. When to stop testing and start using software? ACM SIGMETRICS Perform Eval Rev 1980;10(1):131–138.
69. Yamada S, Osaki S. Optimal software release policies with simultaneous cost and reliability requirements. Eur J Oper Res 1987;31:46–51.
70. Kapur PK, Garg RB. Cost-reliability optimum release policies for software system under penalty cost. Int J Syst Sci 1989;20: 2547–2562.
71. Kapur PK, Garg RB. Optimal release policies for software systems with testing effort. Int J Syst Sci 1991;22(9):1563–1571.
72. Yun WY, Bai DS. Optimum software release policy with random life-cycle. IEEE Trans Reliab 1990;39(2):167–170.
73. Kapur PK, Garg RB, Bhalla VK. Release policies with random software life-cycle and penalty cost. Microelectron Reliab 1993;33(1):7–12.
74. Kapur PK, Garg RB. Optimal software release policies for software reliability growth models under imperfect debugging. RAIRO: OR 1990;24:295–305.
75. Kapur PK, Xie M, Garg RB, *et al.* A discrete software reliability growth model with testing effort. Conference Proceedings, Ist International Conference on Software Testing, Reliability and Quality Assurance; Volume 1; 1994. pp. 16–20.
76. Kapur PK, Agarwal S, Garg RB. Bi-criterion release policy for exponential software reliability growth models. Rech oper/O.R. 1994;28:165–180.
77. Pham H. A software cost model with imperfect debugging, random life-cycle and penalty cost. Int J Syst Sci 1996;27:455–463.
78. Pham H, Zang X. A software cost model with warranty and risk costs. IEEE Trans Comput 1999;48(1):71–75.
79. Xie M, Yang B. A study of the effect of imperfect debugging on software development cost. IEEE Trans Softw Eng 2003;29(5):471–473.
80. Pham H, Zang X. NHPP software reliability and cost models with testing coverage. Eur J Oper Res 2003;145:443–454.
81. Kapur PK, Gupta D, Gupta A, *et al.* Effect of introduction of faults and imperfect debugging on release time. J Ratio Math 2008;18:62–90.
82. Gupta D, Kapur R, Jha PC. Bicriterion release policy for a discrete software reliability growth model with imperfect fault debugging and fault generation. Commun Dependab Qual Manag Int J 2007;10(3):5–31.
83. Ohetera H, Yamada S. Optimal allocation and control problems for software testing resources. IEEE Trans Reliab 1990;39(2):171–176.
84. Yamada S, Ichimori T, Nishiwaki M. Optimal allocation policies for testing-resource based on a software reliability growth model. Math Comput Model 1995;22:295–301.
85. Leung YW. Dynamic resource allocation for software-module testing. J Syst Softw 1997;37(2):129–139.
86. Xie M, Yang B. Optimal testing-time allocation for modular systems. Int J Qual Reliab Manag 2001;18(8):854–863.
87. Lyu MR, Moorsel APA. Optimal allocation of test resources for software reliability growth

- modeling in software development. IEEE Trans Reliab 2002;51(2):183–192.
88. Huang CY, Lo HJ, Kuo SY, *et al.* Optimal allocation of testing resources for modular software systems. Proceedings of the 13th International System on Software Reliability Engineering (ISSRE'02); Annapolis (MD). Volume 13. 2002. pp. 129–138.
89. Huang CY, Lo HJ. Optimal resource allocation for cost and reliability of modular software systems in the testing phase. J Syst Softw 2006;79:653–664.
90. Lo JH. An algorithm to allocate the testing-effort expenditures based on sensitive analysis method for software module systems. TENCON, IEEE Reg 2005;10:1–6.
91. Rajan R, Misra RB. Optimal testing resource allocation models for modular software. Proceedings of Annual Reliability and Maintainability Symposium RAMS'06-2006; 2006 Jan 23-26; Newport Beach (CA). 2006. pp. 104–109.
92. Kapur PK, Jha PC, Bardhan AK. Dynamic programming approach to testing resource allocation problem for modular software. Ratio Math J Appl Math 2003;14:27–40.
93. Kapur PK, Jha PC, Bardhan AK. Optimal allocation of testing resource for a modular software. Asia Pac J Oper Res 2004;21(3):333–354.
94. Jha PC, Gupta D, Anand S, *et al.* An imperfect debugging software reliability growth model using lag function with testing coverage and related allocation of testing effort problem. Commun Dependab Qual Manag Int J 2006;9(4):148–165.

CONTROL VARIATES

LOO HAY LEE

SHAO JIJUN

Department of Industrial and
Systems Engineering, National
University of Singapore

Control variates (CV), or *regression sampling*, is a widely used variance reduction technique (VRT) in Monte Carlo simulation. This method attempts to take advantage of the correlation between random variables to obtain a variance reduction. Depending on the actual problem and the parameter of interest, this correlation used in different CV techniques may be internally available during the course of the original simulation or externally generated using common random numbers (CRN) [1]. CV techniques are popular because of the ease of implementation, the availability of controls, and the straight intuition of the underlying theory [2].

BACKGROUND

When studying or analyzing a complicated system, or sometimes modeling a real-world problem, Monte Carlo simulation is one of the most commonly used methods. Owing to the randomness nature of Monte Carlo methods, the output from such simulation often contains a large amount of uncertainty. This uncertainty will affect the accuracy of the results when we want to draw conclusions about the underlying characteristics of the parameter of interest from the simulation model. Various techniques have been developed to tackle this problem. However, most of them would require many simulation replications, thus making the experiment costly and time consuming. In contrast, VRTs are often able to achieve higher statistical efficiency by reducing the variance of the output random variables without introducing substantial computing cost and bias in its expectation.

Unlike CRN, which is another common VRT often employed to compare two different system configurations, control variates are effective in analyzing a single system performance. Furthermore, CV techniques do not rely on the sign of the correlation; whereas CRN and antithetic variate (AV) both require the correlation being a particular sign (see **Common Random Numbers** and **Antithetic Variates**). As Nelson pointed out in his article “control variate remedies,” the CV method has several advantages comparing to other methods of variance reduction: “it is applicable in single or multiple response simulations; it does not require altering the simulation run; and any stochastic simulation contains potential control variates.” [3]

CLASSICAL CONTROL VARIATE TECHNIQUES

The CV technique is well established as a useful tool in general stochastic simulation experiments. It is applicable in any stochastic simulation, and applying CV technique to estimate one parameter would not conflict with estimating other parameters. [3]

A lot of work related to the CV technique can be found in the literature, ranging from estimating means, to estimating other performance measures in the model. The basic idea of control variates is to take advantage of correlation between certain random variables and to use variables with known properties to adjust, or “control,” the estimator of interest. The following example is an illustration of how a simple CV technique can be applied in a typical queueing model:

Example 1. Consider an M/M/1 queue, let X be an output random variable representing the average delay in the queue for the first 100 customers and we are interested in X 's expectation $\mu = E(X)$. We can define another random variable Y to be the average service time of the first 99 customers in the same queueing model. In a typical simulation experiment, the expected value of Y would be known to us since it is generated

from some input distributions. Furthermore, it is reasonable to argue that longer than average service times would lead to a longer than average waiting time and vice versa. In other words, if Y is larger than its known expectation, then the value of X would also be greater than its true expectation. It gives us some intuition to adjust the value of the estimator of $E(X)$ downwards, if we find Y is above its expectation. In this case, the variable Y could be used as a control variate for the estimator of $E(X)$, since we can adjust, or “control,” the value of the estimator according to the value of Y and its expectation. [1]

In the above example, the control variate Y is positively correlated with the output variable X whose expectation is of interest. However, the success of control variate techniques does not depend on the sign of the correlation. If we define variable Y as the average interarrival time for the first 100 customers, we will find that Y is negatively correlated with X , that is, when Y is above its expectation, the estimator of X should be smaller than its actual value. As a result, we could still use Y as a control variate for the estimator of $E(X)$, that is, we will adjust the value for the estimator of $E(X)$, upward when the value of Y is smaller than its expectation, and vice versa.

For this queueing problem, if n replications of simulations are performed, we could use $\bar{X}(n)$ to denote the sample average of variable X and $\bar{Y}(n)$ as the sample average of control variate Y obtained from the n simulation runs. Let the expected value of Y be denoted by v , which is known to us, that is, $v = E(Y)$. If $\bar{Y}(n)$ differs from its mean value v , we can adjust the value of the estimator X by using the following formula:

$$X_c = \bar{X}(n) - a [\bar{Y}(n) - v], \quad (1)$$

where X_c is the new estimator for X using Y as a control variate and a is the control coefficient. From the derivation below, we can conclude that the new estimator X_c is an

unbiased estimator of X .

$$\begin{aligned} E(X_c) &= E[\bar{X}(n)] - E\{a [\bar{Y}(n) - v]\} \\ &= E[\bar{X}(n)] - a \{E[\bar{Y}(n)] - v\} \\ &= E(X) = \mu. \end{aligned} \quad (2)$$

Its variance is given by

$$\text{Var}(X_c) = \text{Var}(X) + a^2 \text{Var}(Y) - 2a \text{Cov}(X, Y). \quad (3)$$

From Equation (3), we can see that the variance of the new estimator X_c is smaller than the variance of the original estimator X , if $a^2 \text{Var}(Y) < 2a \text{Cov}(X, Y)$. Hence if X and Y are positively correlated, there will be a variance reduction when the value of coefficient a is positive but smaller than $2\rho(X, Y) \frac{\sigma(X)}{\sigma(Y)}$; if X and Y are negatively correlated, the variance reduction exists when the value of a is negative and $|a| < |2\rho(X, Y) \frac{\sigma(X)}{\sigma(Y)}|$. The optimal choice of a can be found by finding the derivative of the right-hand side of Equation (3) and setting it to zero. The optimal value of coefficient a , denoted by a^* is given by

$$a^* = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} = \rho(X, Y) \frac{\sigma(X)}{\sigma(Y)}. \quad (4)$$

The variance of the optimal control variate estimator, denoted by X_c^* , can be expressed in the following equation:

$$\begin{aligned} \text{Var}(X_c^*) &= \text{Var}(X) - \frac{[\text{Cov}(X, Y)]^2}{\text{Var}(Y)} \\ &= [1 - \rho^2(X, Y)] \text{Var}(X). \end{aligned} \quad (5)$$

The maximum reduction in variance is thus $\rho^2(X, Y) \text{Var}(X)$. If $\rho \rightarrow \pm 1$, which means that X and Y are almost perfectly correlated, the variance of the control variate estimator would become very close to zero, indicating that we can get a very precise estimator for μ .

However, to compute the optimal coefficient a^* would require the value of the variance of the control variate Y and the covariance between X and Y . In the actual application, we may be able to find the variance of Y , since it is an input variable that is generated by us (in some other examples, we may not

be able to obtain this variance too), but it is nearly impossible for us to know the exact value of the covariance between X and Y . Thus, to use control variate techniques, we need to find a way to estimate these important parameters. Many methods have been proposed to estimate α^* , and one common method is to replace $\text{Var}(Y)$ and $\text{Cov}(X, Y)$ with their sample estimators. Thus, the covariance $\text{Cov}(X, Y)$ can be estimated by

$$\hat{C}_{XY}(n) = \frac{\sum_{j=1}^n [X_j - \bar{X}(n)] [Y_j - \bar{Y}(n)]}{n-1}, \quad (6)$$

where X_j and Y_j denote the individual response observed in simulation run j for variable X and Y , and $\hat{C}_{XY}(n)$ is the sample covariance estimated from n simulation replications. The variance of Y , $\text{Var}(Y)$ can also be estimated by the sample variance $S_Y^2(n)$

$$S_Y^2(n) = \frac{\sum_{j=1}^n [Y_j - \bar{Y}(n)]^2}{n-1}. \quad (7)$$

Using these two sample estimators of variance and covariance, we are able to estimate the optimal coefficient α^* by

$$\hat{\alpha}^*(n) = \frac{\hat{C}_{XY}(n)}{S_Y^2(n)}. \quad (8)$$

And the corresponding control variate estimator of μ is then given by $\hat{X}_C^*$

$$\hat{X}_C^* = \bar{X}(n) - \hat{\alpha}^*(n) [\bar{Y}(n) - v]. \quad (9)$$

Another property of the estimator $\hat{\alpha}^*(n)$ is that $\hat{\alpha}^*(n)$ is equal to the least square estimator of the regression coefficient α in the following simple linear regression equation:

$$X = \alpha Y + \beta + \varepsilon, \quad (10)$$

where ε denotes the stochastic disturbance around the regression line and α and β are the regression coefficients.

By definition, coefficient α minimizes the term $\sum_{i=1}^n (X_i - \tilde{X}_i)^2$, where X_i is the observed response from the i th simulation

replication and $\tilde{X}_i$ is the value predicted from the regression equation. Thus, it is equal to minimizing $\sum_{i=1}^n (X_i - \alpha Y_i - \beta)^2$ [4].

Since $\hat{\alpha}^*(n)$ is calculated from the same samples as the sample mean $\bar{Y}(n)$, they are dependent with each other. As a result, this new estimator $\hat{X}_C^*$ would possibly be biased, and the amount of variance reduction might be reduced as well.

To deal with this problem, there are a number of methods proposed by various researchers to find an unbiased control variate estimator. Tocher [5] mentioned in his book that by splitting all n observations (X_i, Y_i) into two groups with each group consisting of half of the observations, he is able to construct an unbiased estimator. From these two groups of data, two different estimators of the coefficient α , $\hat{\alpha}_1^*(\frac{n}{2})$ and $\hat{\alpha}_2^*(\frac{n}{2})$ can be obtained. It is then possible to construct two control variate estimators by combining $\hat{\alpha}_1^*(\frac{n}{2})$ with the second group of data and $\hat{\alpha}_2^*(\frac{n}{2})$ with the first group of data. An unbiased estimator can be obtained by averaging these two new control variate estimators.

Tocher also derived that the variance of the unbiased estimator is two times of the variance of the optimal estimator using α^* derived earlier. However, there is a remedy for this variance increase, that is, by utilizing generalized splitting. This method would require to divide the n pairs of observations into J groups ($2 < J \leq n$), and the estimator of the coefficient of group j is derived using all observations except the observations belonging to group j . Using the same method as in the two group case, a more effective variance reduction is obtained from this new estimator.

Another widely used bias reduction technique is Jackknifing technique. Lavenberg *et al.* [6] discussed how to use Jackknifing technique to estimate the confidence interval for some closed queueing networks.

GENERALIZATION ON BASIC CV TECHNIQUES

The discussion presented above mainly focuses on using a single control variate. We

shall then discuss some generalization of this basic CV technique.

In a simulation experiment, it is common that we are able to find not only a single “control variable”, but several other variables suitable of being used as control variates. As we discussed in the previous example, both average customer interarrival time and customer service time can be used as control variates. Under such a case, we could then formulate a kind of “multiple” control variate estimator using m different control variates $Y_i, i = 1, 2, \dots, m$, as follows:

$$X_{cm} = \bar{X}(n) - \sum_{i=1}^m a_i (\bar{Y}_i(n) - v_i), \quad (11)$$

where $\bar{Y}_i(n)$ is the sample average of the simulation response for the i th control variate, and v_i is the known expectation for Y_i . All a_i 's are the coefficients which need to be estimated. It will be more convenient to consider control variates with zero-expectation, and so Kleijnen [4] redefine the control variates Z_i for each Y_i as follows:

$$Z_i = \bar{Y}_i(n) - v_i, \quad i = 1, 2, \dots, m. \quad (12)$$

It is clear that $E(Z_i) = 0$, and the control variate estimator defined in Equation (11) is then

$$X_{cm} = \bar{X}(n) - \sum_{i=1}^m a_i Z_i. \quad (13)$$

If the control variates $Z_i (i = 1, 2, \dots, m)$ are all independent with each other, which will be the case if we assume that all Y_i s are input variables, the variance of the new estimator X_{cm} is then

$$\begin{aligned} \text{Var}(X_{cm}) &= \text{Var}[\bar{X}(n)] - \sum_{i=1}^m a_i^2 \text{Var}(Z_i) \\ &\quad - 2 \sum_{i=1}^m a_i \text{Cov}(X, Z_i). \end{aligned} \quad (14)$$

Taking partial derivatives of the right-hand side of Equation (14) with respect to each of the a_i s and equating them to zero, we can

find the optimal values for a_i s. The optimal values are

$$a_i^* = \frac{\text{Cov}(X, Z_i)}{\text{Var}(Z_i)}. \quad (15)$$

Using the fact that $E(Z_i) = 0$, we can have a simple estimator of the optimal control coefficient

$$\hat{a}_i^* = \frac{\sum_{j=1}^n X_j Z_{ij}}{\sum_{j=1}^n Z_{ij}^2}, \quad \text{where } i = 1, 2, \dots, m. \quad (16)$$

In Equation (16), X_j is the observation from j th simulation run, and Z_{ij} is the value calculated using $Y_{ij} - v_i$, while Y_{ij} is the observation of variable Y_i in j th run.

We can also show that this estimator is equivalent to the least square estimators of the regression coefficients of the following equation:

$$X_j = \sum_{i=1}^m a_i Z_{ij} + \varepsilon_j, \quad \text{where } j = 1, 2, \dots, n. \quad (17)$$

Lavenberg and Welch [7], as well as Nelson [3], both discussed the case where correlation between control variates $Z_i (i = 1, 2, \dots, m)$ are allowed. Under the assumption that the random output X and the m control variates $Z_i (i = 1, 2, \dots, m)$ follow multivariate normal distribution, it turns out that the optimal variance minimizing coefficients are identical to least-square estimates of the coefficients in a linear-regression model. This is the reason why CV is often referred to as *regression sampling*.

In the queueing model in Example 1, the control variate Y is defined as the average service time of the first 99 customers. In a typical simulation experiment, since the service time of each customer is an input variable and is generated from some given distribution, we can also define the service time of these 99 individual customers as 99 i.i.d random variables. In this way, the control variate Y can be seen as a function (simple average) of these 99 i.i.d input random variables. Similarly, we can also define Y as some

other more complicated function of the individual input variables which are generated from some known input distribution [4]. We define the new control variate Z , as follows:

$$Z = g[Y(1), Y(2), \dots, Y(m)], \quad (18)$$

where g is a function of the m individual input variables $Y(1), Y(2), \dots, Y(m)$ in a single simulation run. For example, in a single server queue similar to the one described above, $Y(1)$ is the service time for the first customer, $Y(2)$ is the service time for the second customer, and $Y(m)$ is the service time for the m th customer. These m random variables are used in the simulation as inputs and are generated using a known distribution. The function g is a general function which could include higher order terms of $Y(i)$ s, some ratio terms, or even terms which requires maximizing or minimizing of some function of the variable $Y(i)$ s [8]. And the number of inputs, m , can also be stochastic. Take a single server queue as an example, if we simulate for a given amount of time, say, t , then because the number of customers in a given time period is random, the number of service times generated for the customers in a single simulation run is also random.

The new variable Z can be seen as a function of the m responses for a stochastic input variable Y . We need to impose an additional condition that the expected value of Z can be obtained. Thus,

$$E(Z) = E\{g[Y(1), Y(2), \dots, Y(m)]\} = \xi. \quad (19)$$

And we can get the new control variate estimator X_{cg} by

$$X_{cg} = \bar{X}(n) - Z + E(Z) = \bar{X}(n) - g[Y(1), Y(2), \dots, Y(m)] + \xi. \quad (20)$$

The expected value and variance of X_{cg} is given by

$$\begin{aligned} E(X_{cg}) &= E[\bar{X}(n)] - E(Z) + \xi \\ &= E[\bar{X}(n)] - \xi + \xi = E(X) \end{aligned} \quad (21)$$

$$\text{Var}(X_{cg}) = \text{Var}[\bar{X}(n)] + \text{Var}(Z) - 2\text{Cov}(X, Z). \quad (22)$$

From Equation (22) we can see that different from all previously discussed control variates, in this case, the control variate Z should be constructed such that it would have a “strong” positive correlation with the parameter of interest X . Considering a general definition where the number of responses, m , is stochastic, Z can be defined as follows:

$$\begin{aligned} Z &= g[Y(1), Y(2), \dots, Y(m)] \\ &= b_1 + b_2 \sum_{i=1}^m Y(i) + b_3 \left[\sum_{i=1}^m Y(i) \right]^2, \end{aligned} \quad (23)$$

Then the control variate estimator using the above defined Z can be written as

$$\begin{aligned} X_{cg2} &= \bar{X}(n) \\ &\quad - \left\{ b_1 + b_2 \sum_{i=1}^m Y(i) + b_3 \left[\sum_{i=1}^m Y(i) \right]^2 \right\} \\ &\quad + E \left\{ b_1 + b_2 \sum_{i=1}^m Y(i) + b_3 \left[\sum_{i=1}^m Y(i) \right]^2 \right\} \\ &= \bar{X}(n) - b_2 \left\{ \sum_{i=1}^m Y(i) - E \left[\sum_{i=1}^m Y(i) \right] \right\} \\ &\quad - b_3 \left\{ \left[\sum_{i=1}^m Y(i) \right]^2 - E \left\{ \left[\sum_{i=1}^m Y(i) \right]^2 \right\} \right\}. \end{aligned} \quad (24)$$

Since

$$E \left[\sum_{i=1}^m Y(i) \right] = vE(m) \quad (25)$$

$$E \left\{ \left[\sum_{i=1}^m Y(i) \right]^2 \right\} = v^2 E(m^2) + \sigma^2 E(m), \quad (26)$$

where $v = E[Y(i)]$ and $\sigma^2 = \text{Var}[Y(i)]$. Since all $Y(i)$ s are generated from some given input distribution, v and σ^2 can be obtained easily. Then, if we can calculate the expected value of the number of responses, we are able to use Equation (24). Otherwise, if we are unable to obtain these expectations, we can replace X_{cg2}

by the following control variate estimator:

$$X_{c_{g3}} = \bar{X}(n) - b_2 \left[\sum_{i=1}^m Y(i) - vm \right] - b_3 \left\{ \left[\sum_{i=1}^m Y(i) \right]^2 - (v^2 m^2 + \sigma^2 m) \right\}. \quad (27)$$

We can see that Equation (27) is an unbiased estimator. Thus, it is suitable for us to run the simulation, obtain the value of v and m at the end of the experiment, and then use Equation (27) as a control variate estimator to estimate the parameter of interest [4].

SELECTING THE BEST CONTROL VARIATES

In most simulation experiments, the optimal control variate coefficient can only be estimated. While we have discussed several ways to estimate the coefficient in the above section, we should note that such estimation may introduce extra variance, thus offsetting the variance reduction obtained from using control variates, especially when using multiple control variates. In previous discussions, we can see that the number of control variates available for a given simulation experiment may be quite large. Simply using all the available control variates may increase the overall cost, and more importantly, reduce the variance reduction we want to achieve. Thus, it is intuitive that we may want to only select a subset of these available control variates to maximize the variance reduction and avoid biasness in the estimators. Bauer and Wilson [9], in their paper, proposed control-variate selection criteria, which minimize the mean-square confidence-region volume.

Let us define the following terms to be used in this section. We use $\mathbf{X} = [x_1, x_2, \dots, x_p]'$ to denote a column vector representing p responses of interest generated on the simulation experiment, with $\mu_X = E(\mathbf{X})$ to be estimated. The pool of m control variates is denoted by another column vector $\mathbf{Y} = [Y_1, Y_2, \dots, Y_m]'$ with m columns and known mean $\mu_Y = E(\mathbf{Y})$. Thus, if we use a $q \times m$ matrix α to denote the control variate coefficient, the

controlled response can be written as

$$\mathbf{X}_c = \mathbf{X} - \alpha (\mathbf{Y} - \mu_Y). \quad (28)$$

This control variate estimator is an unbiased estimator of the response $\mathbf{X}$, and its variance can be minimized with optimal value of α , given by

$$\alpha^* = \Sigma_{XY} \Sigma_Y^{-1}, \quad (29)$$

where Σ_{XY} represents the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$, while Σ_Y denotes the covariance matrix of $\mathbf{Y}$. As it is often the case that we do not know the covariance between $\mathbf{X}$ and $\mathbf{Y}$, this optimal value α^* needs to be estimated by

$$\hat{\alpha}^* = \mathbf{S}_{XY} \mathbf{S}_Y^{-1}. \quad (30)$$

In Bauer and Wilson's paper, a control variate selection criterion is proposed based on the principle of minimizing the mean square volume of the confidence region. In order to present this criterion clearly, additional notation is defined. We define a nonnegative integer r to denote the number of control variate candidates currently under consideration, and $u(m, r) \equiv \frac{m!}{r!(m-r)!}$ to represent the total number of distinct control variate subsets of size r . We define $I(h, r)$, with $h = 1, 2, \dots, u(m, r)$, to denote the h th distinct subset of size r . On the j th run of the simulation model, we use $\mathbf{Y}_{(j)}(h, r) \equiv (Y_{(j)i_1}, Y_{(j)i_2}, \dots, Y_{(j)i_r})'$, where $\{i_1, i_2, \dots, i_r\} = I(h, r)$ to denote the r -dimensional vector of controls corresponding to the index set $I(h, r)$. Similarly, we define $\hat{\Sigma}_{X|Y}(h, r)$ to be the estimator of the conditional covariance matrix of $\mathbf{X}$ given $\mathbf{Y}$, when the control variate vector $\mathbf{Y}(h, r)$ defined by $I(h, r)$ is used. Then the best selection is obtained by minimizing the following expression:

$$\min_{\substack{0 \leq r \leq m \\ 1 \leq h \leq u(m, r)}} \left\{ \frac{|\hat{\Sigma}_{X|Y}(h, r)| \Gamma\left(\frac{n}{2}\right)}{\left[\left(\frac{p}{2}\right) \Gamma\left(\frac{p}{2}\right)\right]^2 \Gamma\left[\frac{n-2p}{2}\right]} \times \left[\frac{2\pi p F_{1-\alpha}(p, n-r-p)}{n(n-r-p)} \right]^p \times \prod_{i=1}^p \left(\frac{n-r-1}{n-r-2i} \right) \right\}, \quad (31)$$

where $F_{1-\alpha}(p, n - r - p)$ denotes the quantile of order $1 - \alpha$ for a F distribution with p and $n - r - p$ degrees of freedom.

As this optimization problem is not easy to solve when the number of available controls become very large, Bauer and Wilson [9] suggested using the method in a stepwise fashion to select some “good” controls instead of finding the optimal ones. For a more detailed explanation and derivation, please refer to Bauer and Wilson’s paper.

ADVANCED NEW DEVELOPMENT IN CV TECHNIQUES

More recently, many new methods of applying CV techniques in different problems have been proposed by researchers. Nelson [3] reviewed some popular control variate techniques in the literature. By considering small sample statistical properties, he conducted systematic analytical and empirical evaluation on these techniques. Szechtman [2] presented a comprehensive review of classical control variate techniques, focusing on both theory and application.

Glynn and Whitt [10] showed that, the asymptotic efficiency of any nonlinear control variates is the same as that of a linear control variate scheme under mild conditions. Porta Nova and Wilson [11] studied methods to develop control variates for multiresponse simulation metamodels. In their study, both the input and output of the metamodels could be multidimensional. As a result, the control variates can be applied to multipopulation, multiresponse simulation experiments. Nelson and Hsu [12] introduced control-variate models of CRN that permit exact statistical inference, specifically for comparing more than two systems. These models explain the effect of CRN via a linear regression of the simulation output on control variates that are functions of the simulation inputs. Rubinstein and Marcus [13] analytically derived the optimal size of the control variate vector under specific assumptions on the covariance matrix. To effectively measure the efficiency of the control variate technique, Venkatraman and Wilson [14] studied ways to measure the efficiency of the control variate technique effectively. They proposed a simplified

formulation of the variance ratio and a loss factor to evaluate the trade-offs in selecting control variates for a simulation experiment.

The effectiveness of the CV technique largely depends on the covariance between the control variates and the parameter of interest. To increase this covariance and thus, to improve the performance of the control variate, different forms of parametric transformations including polynomial transformations, the Box–Cox family of power transformations, and the inverse-normal probability transformation could be used on the linear control variates discussed earlier. Besides these classical approaches, Swain [15] considered a form of nonparametric transformations using natural cubic splines which provide a virtually unrestricted source of transformations, with considerable flexibility at the expense of additional computational effort. A nonlinear regression problem was analyzed to illustrate this transformation methodology.

Classical control variate techniques require that users know the means of the controls. Borogovas and Vakili [16] relaxed this requirement and allow the means to be estimated. They analyzed the properties of controlled estimators, and proposed a class of controls in a parametric estimation setting. Wieland and Schmeiser [17] investigated a case where the variance of the control variate is known, but the covariance between the control variate and the parameter of interest is unknown. A mixed estimator, which uses an estimate of the correlation of the performance measure and the control variate is evaluated. Their results indicated that the mixed estimator had most potential benefit when no information on the correlation is available, especially when sample sizes are small. Kim and Henderson [18] studied the case of finite-horizon simulation when nonlinearly parameterized control variables are available. They developed two adaptive control variate methods based on stochastic approximation and sample average approximation. They also provided conditions under which the adaptive estimators are consistent. Henderson and Glynn [19] discussed control variate application in Markov process simulation. They proposed a new approach

to construct a martingale as an internal control variate. Maire [20] presented a variance reduction method for Monte Carlo integration based on an iterated computation of L2 approximations using control variates. In his results, Monte Carlo estimates with increased convergence rate are obtained for mono-dimensional integrals of regular functions using only sample values.

APPLICATIONS OF CV TECHNIQUES

Control variate techniques can be easily applied to queueing systems, truck dock systems and many other simulation systems [4]. Quite recently, a number of control variate applications in stock and option pricing are proposed by different researchers. Glasserman [21] gave many examples on the application of control variates in various option pricing problems. Pellizzari [22] proposed a new control variate technique and a general Monte Carlo procedure. This technique could be used on a variety of multifactor options. Bolia and Juneja [23] presented a zero-variance or “perfect” control variate to price American options. They also explained how function approximation may be used to approximate this perfect control variate. Ehrlichman and Henderson [24] also developed a class of control variates for the American option pricing problem using multivariate adaptive regression splines.

These applications have shown the effectiveness of control variate techniques in various engineering and business problems. They also demonstrated the potential for the widespread use of control variate techniques.

REFERENCES

1. Law AM, Kelton WD. Simulation modeling and analysis. McGraw-Hill Higher Education; 2000.
2. Szechtman R. Control Variates Techniques for Monte Carlo Simulation. Proceedings of the 2003 Winter Simulation Conference. New Orleans, LA; 2003.144–149.
3. Nelson BL. Control variate remedies. *Oper Res* 1990;38(6):974–992.
4. Kleijnen JP. Statistical techniques in simulation Part 1. New York: Marcel Dekker; 1974.
5. Tocher KD. The art of simulation. London: The English Universities Press; 1963.
6. Lavenberg SS, Moeller TL, Welch PD. Statistical results on control variables with application to queueing network simulation. *Oper Res* 1982;30(1):182–202.
7. Lavenberg SS, Welch PD. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulation. *Manage Sci* 1981;27(3):322–335.
8. Cheng RC, Feast GM. Control variables with known mean and variance. *J Oper Res Soc* 1980;31(1):51–56.
9. Bauer KW, Wilson JR. Control-variate selection criteria. *Nav Res Logist* 1992;39:307–321.
10. Glynn PW, Whitt W. Indirect estimation via $L = \lambda W$. *Oper Res* 1989;37(1):82–103.
11. Porta Nova AM, Wilson JR. Estimation of multiresponse simulation metamodels using control variates. *Manage Sci* 1989;35(11):1316–1333.
12. Nelson BL, Hsu JC. Control-variate models of common random numbers for multiple comparisons with the best. *Manage Sci* 1993;39(8):989–1001.
13. Rubinstein RY, Marcus R. Efficiency of multivariate control variates in Monte Carlo simulation. *Oper Res* 1985;33(3):661–677.
14. Venkatraman S, Wilson JR. The efficiency of control variates in multiresponse simulation. *Oper Res Lett* 1986;5(1):37–42.
15. Swain JJ. Augmenting Linear Control Variates Using Transformations. Proceedings of the 1989 Winter Simulation Conference. Washington (DC); 1989.455–458.
16. Borogovac T, Vakili P. Control Variate Technique: A Constructive Approach. Proceedings of the 2008 Winter Simulation Conference. Miami, FL; 2008.320–327.
17. Wieland JR, Schmeiser BW. Derivative Estimation with Known Control Variate Variances. Proceedings of the 2007 Winter Simulation Conference. Washington (DC); 2007.560–567.
18. Kim S, Henderson SG. Adaptive control variates for finite-horizon simulation. *Math Oper Res* 2007;32(3):508–527.
19. Henderson SG, Glynn PW. Approximating martingales for variance reduction in Markov process simulation. *Math Oper Res* 2002;27(2):253–271.

20. Maire S. Reducing variance using iterated control variates. *J Stat Comput Simulat* 2002; 73(1):1–29.
21. Glasserman P. Monte Carlo methods in financial engineering. New York: Springer; 2004.
22. Pellizzari P. Efficient Monte Carlo pricing of European options using mean value control variates. *Decis Econ Finance* 2001;24: 107–126.
23. Bolia N, Juneja S. Function-approximation-based perfect control variates for pricing American options. *Proceedings of the 2005 Winter Simulation Conference*. Orlando, FL; 2005;1876–1883.
24. Ehrlichman SM, Henderson SG. American options from MARS. *Proceedings of the 2006 Winter Simulation Conference*. Monterey, CA; 2006.719–726.
25. Fishman GS. Using control variates to estimate distribution functions. *Proceedings of the 1986 Winter Simulation Conference*. Washington, DC; 1986;335–336.

COOPERATIVE GAME THEORY WITH NONTRANSFERABLE UTILITY

JUNGSUK OH
College of Business Administration,
Seoul National University,
Seoul, Korea

INTRODUCTION

The description and analysis of cooperative games become substantially simplified when one considers the case of transferable utility (TU), which is equivalent to the case of games with side payments. On the other hand, the situation in which a nontransferable utility (NTU) assumption is appropriate and detailed information regarding individual payoff configuration under various coalitional structure is available calls for an NTU modeling. Consider the following classic example by Roth [1].

Example 1 [1]. Consider a three-player game where each player acting alone yields payoff of zero for everyone, cooperation between Player 1 and 2 yields payoff configuration of $(1/2, 1/2, 0)$, cooperation between Player 1 and 3 yields $(p, 0, 1 - p)$, that between Player 2 and 3 yields $(0, p, 1 - p)$, and the grand coalition yields any convex combination of $(1/2, 1/2, 0)$, $(p, 0, 1 - p)$, and $(0, p, 1 - p)$. The coalitional form representation for this game under TU assumption would yield, along with the player set $N = \{1, 2, 3\}$, the following worth specifications:

$$\begin{aligned}v(i) &= 0 \quad i = 1, 2, 3 \\v(12) &= v(23) = v(13) = 1 \\v(N) &= 1.\end{aligned}$$

For the TU correspondence of this game, the core is empty and the Shapley (TU) value would be $(1/3, 1/3, 1/3)$. On the other

hand, it is apparent that an imposition of a NTU assumption for this game would make the TU solutions inappropriate due to the dependence of the payoff configuration on the parameter p , which is likely to yield an asymmetric payoff distribution if $p \neq 1/2$.

Consider an ordinary (TU) coalitional function $v(S)$ that assigns a real number (worth) to a coalition S . This single number of worth signifies the total payoff that members of S in aggregate can be entitled to. Furthermore, TU coalitional function form does not specify any mechanism of redistributing this sum among coalitional members. This means that the TU coalitional form game assumes that the worth of a coalition can be distributed without any restriction. These assumptions of ease of aggregation and redistribution of utilities for members of coalition have been subject to criticism. In particular, the following criticisms were raised by Aumann and Peleg [2]. First, the TU game construction assumes that there is a common medium of exchange through which members of a coalition can make side payments, which, in turn, transfer utility. Although money is a universal means of exchange in almost all economic contexts, use of it as means of aggregation and redistribution within a coalition may raise subtle issues or may not even be viable, depending on the context of the coalition formation.

Second, the TU game assumes that side payments are implementable both from the physical and legal perspectives. The physical means of side payments are usually assumed to be money so that, even in a global economy setting, a use of numeraire might do the job. However, in many industrial settings, side payments are subject to legal penalty, which limits the applicability of TU game instruments.

Third, it is assumed by the TU game that utility is unrestrictedly transferable, which, in turn, implies that each member's utility for medium of exchange is a linear function of the amount of medium. This assumption is equivalent to saying that the utility function

of a coalition member i is of a separable form

$$u_i(x_1, x_2, \dots, x_k) = v_i(x_1, x_2, \dots, x_{k-1}) + \delta_i x_k, \quad (1)$$

where good k represents the medium of exchange. With this setup, each member's utility function can be normalized by the factor δ_i , which makes it possible that transfers of the k th good entails a linear transfer of utility (i.e., the k th good serves as a medium of exchange). Considering this assumption of unrestricted utility transfer, the applicability of the TU game narrows down even further to those von Neumann–Morgenstern utility functions that are linear and increasing in the medium of exchange, that is, the quasi-linear utility function. Furthermore, as Varian [3] describes, even quasi-linear preferences allow for utility transfers only up to a certain range, which might further limit the applicability of TU games.

For these reasons, much efforts have been put forth to develop parallel cooperative game theories pertaining to contexts in which side payments are not possible, or, equivalently, utilities are nontransferable.

LINK TO NASH BARGAINING SOLUTION

Two axiomatic solution concepts of the cooperative game theory, the Shapley value [4] and the Nash bargaining solution (NBS) [5,6] have been the foundations for the development of the cooperative game theory. Unlike the Shapley value, which is a solution concept for a TU setting, the NBS is a solution applicable to an NTU game (and thereby, is also applicable to the TU game setting). Recognizing its nondependence on the TU assumption, the NBS quickly established itself as a status quo against which many subsequent works on approaches and results of solution concepts have been compared. Particularly, as the setting for the NBS is simple and elegant; there has been much contemplation and efforts on extending its settings, which shed further insights on the solution and, consequently, enrich the NTU cooperative game theory.

The static model of Nash contains a few elements. Let X be the feasible payoff space

and $u_i(x)$, ($i = 1, 2, x \in X$), the utility function for each player. (S, d) is the Nash bargaining problem where the feasible set $S = \{s_1, s_2\}$ is a convex, compact, and comprehensive set and $d = (d_1, d_2) \in S$ is the threat (or disagreement) point. Then, the classic NBS is a unique element in S that maximizes $(s_1 - d_1) \cdot (s_2 - d_2)$. Among various extensions of the two-person pure bargaining setting of Nash, a few are worth noticing. One obvious direction of extension is the development of the theory for the n -person setting, which is discussed in the subsequent subsection on values of NTU game.

Other extensions have taken place in directions that are based on contemplations on each element of Nash bargaining assumptions. Interpretation of cooperative game solutions as expected utility of playing a game was one such approach. After establishing the interpretation of the Shapley value for a TU game as a cardinal utility function [7,8], a similar attempt was made by Roth [9] in which the n -person bargaining problem with explicit consideration of players' risk attitudes was presented. Here, it was shown that the NBS is a solution corresponding to the risk neutral preference case. This treatment of the risk attitude can be viewed as a manipulation of the information contained in the feasible set S .

Binmore, Rubinstein and Wolinsky [10] provided an alternative way to extend the Nash bargaining problem via manipulation of the threat points d . Their framework was applied to two situations corresponding to different incentives to reach an agreement: players' impatience and fear of breakdown of the negotiation. They referred to the solution to their extended model time-preference Nash solution. Certainly, it is an appropriate description of the environment in which the bargaining takes place in an almost deterministic setting and, thus, losses associated with delays in reaching the agreement play a more important role than uncertainty. They also revealed that the limiting equilibrium outcome corresponds to the original NBS, that is, the positions corresponding to the maximum value of products of incremental gains of each player.

There are other extension works of the NBS. For example, Schmitz [11] considered a setting without the threat point. Thompson [12] and Thompson and Lensberg [13] incorporated the possibility that the set of players might change.

EFFECTIVITY CRITERIA AND COALITIONAL FORM NTU GAME

One of the important purposes that a coalitional form game serves is to predict prospects of various coalitions of a cooperative strategic-form game. Of course, given a strategic-form game, a specification of a coalitional function is a key task in identifying a coalitional form correspondence. As von Neumann and Morgenstern [14] suggested, maxmin criteria can be utilized to derive the coalitional function for a general nonzero sum game. Such an approach is evaluated to be “natural” in that the solution is both the maximum amount the player can guarantee for itself and the minimum amount to which the player can be driven by the opponent [15]. Sharing a similar spirit, two notions of “effectiveness” for the case without side payment were introduced in Aumann and Peleg [2] and Aumann [16]. These effectiveness concepts were products of an initial attempt to suggest ways to specify the worth of a coalition for the NTU strategic form game. Specifically, Aumann and Peleg [2] stated that

Intuitively, a coalition S is effective for a payoff vector x if the members of S , by joining forces, can play so that each player i in S receives at least x_i .

On the basis of this generalization of the maxmin approach of von Neumann and Morgenstern, the following definitions were proposed.

Definition 1. A coalition S is α -effective for a payoff vector x if there is a strategy for S such that for each strategy used by $N - S$, each member i of S receives at least x_i .

Definition 2. A coalition S is β -effective for a payoff vector x if for each strategy used by

$N - S$, there is a strategy for S such that each member i of S receives at least x_i .

An α -effective payoff is the amount that S can guarantee for its members regardless of actions of $N - S$, whereas a β -effective payoff is the amount that, through appropriate reaction to each action of $N - S$, S can guarantee for its members. Let $V_\alpha(S)$ and $V_\beta(S)$ be set of α -effective and β -effective payoffs for S respectively. Then, obviously, $V_\alpha(S) \subset V_\beta(S)$ for an NTU game. Moreover, for TU games, $V_\alpha(S) = V_\beta(S)$. The correspondence $V_\alpha(\cdot)$ and $V_\beta(\cdot)$ can be utilized to specify the coalitional form game of cooperative strategic-form game, and, hence, is referred to as α -effective and β -effective coalitional form of the strategic-form game. Although α -effectiveness possesses more intuitive appeal due to its invariant nature, β -effectiveness is more significant when used for the purpose of relating the coalitional function to the cooperative equilibria of a repeatedly played strategic-form game [16].

Example 2 [16]. Let $N = \{1, 2, 3\}$ and $S = \{1, 2\}$. Consider the following strategic form game.

Player		3
1,2	1, -1, arbitrary number	0,0, arbitrary number
	0,0, arbitrary number	-1, 1, arbitrary number

In reflection of the maxmin approach, one can ask the following question. What is the payoff guarantee, (x_1, x_2) , that S can make to its members? It is apparent for this example that, for a payoff vector of $(0, 0)$, S is β -effective but not α -effective.

Whichever way of specifying the coalitional function is utilized, the definition of an NTU coalitional form game is given via the coalitional function that assigns a feasible set for each coalition instead of a single number of worth as in TU coalitional form game.

Definition 3. An NTU game is denoted as (N, V) where N is the set of players and V is a coalitional function which assigns to each coalition $S \subset N$ the feasible payoff vectors of S , $V(S) \subset \mathbb{R}^S$. The feasible payoff vector, $V(S)$, is nonempty for $S \neq \emptyset$, closed, convex, and comprehensive (i.e., $x' \leq x \in V(S) \Rightarrow x' \in V(S)$).

In Definition 3, $V(N)$ is the set of all “feasible” or “attainable” payoff vectors. It is clearly visible from this definition of NTU game that a TU game is a special case of an NTU game. Each TU game (N, v) is equivalent to an NTU game (N, V_v) , where $V_v(S) = \{x \in \mathbb{R}^S : \sum_{i \in S} x_i \leq v(S)\}$ and $v(\cdot)$ is the worth function of (N, v) . Here, the game structure in which one can specify a single, real-number worth that bounds the sum of all members of a coalition enables the specification of a TU game.

Example 3 [Continued; [1]]. A coalitional form NTU game specification of the game in Example 1 would yield the following:

$$\begin{aligned} V(i) &= \{x \in \mathbb{R} : x_i \leq 0\} \quad i = 1, 2, 3 \\ V(1, 2) &= \{x \in \mathbb{R}^2 : (x_1, x_2) \leq (1/2, 1/2)\} \\ V(1, 3) &= \{x \in \mathbb{R}^2 : (x_1, x_3) \leq (p, 1-p)\} \\ V(2, 3) &= \{x \in \mathbb{R}^2 : (x_2, x_3) \leq (p, 1-p)\} \\ V(N) &= \left\{ \begin{array}{l} x \in \mathbb{R}^N : x \leq y \text{ for some } y \text{ in the} \\ \text{convex hull of} \\ (1/2, 1/2, 0), (p, 0, 1-p), \\ (0, p, 1-p) \end{array} \right\}. \end{aligned}$$

CORE OF THE NTU GAME

As one of the most natural solutions to the cooperative game, the core for the TU game, as explicitly defined by Gillies [17], quickly established itself as one of the main solution concepts of the cooperative game theory. One weakness of the core that this intuitive solution might not universally exist has prompted much research efforts to identify the conditions of its existence, which introduced important characterization concepts

and derivative solutions such as balancedness and the ε -core. Parallel efforts have been put forth on the core for the NTU games. First, Aumann [16] provided a mathematical framework for a parallel development of the theory of the core for the NTU game where formal definitions of a coalitional form NTU game and dominance, as well as suggestions for transforming a given normal form game into a coalitional form NTU game using α - and β -effectiveness concepts, which were originally proposed in Aumann and Peleg [2].

The following definition is useful in defining the core for an NTU game.

Definition 4. A coalition S can improve upon a payoff vector y if there is an $x \in V(S)$ such that $x_i > y_i$ for $\forall i \in S$.

Definition 5. The core $C(N, V)$ of an NTU game (N, V) is the set of payoff vectors in $V(N)$ that cannot be improved upon by any coalition. Formally

$$C(N, V) = V(N) \setminus \bigcup_{S \subset N} \text{int } V(S)$$

Example 4. Let $N = \{1, 2, 3\}$ and $0 \leq p < 1$. Let V be given by

$$\begin{aligned} V(i) &= \{x \in \mathbb{R} : x_i \leq p\} \quad i = 1, 2, 3 \\ V(i, j) &= \left\{ x \in \mathbb{R}^2 : x_i \leq 1, \text{ and } x_j \leq 1 \right\}, \text{ for } \forall i, \\ &\quad j \in N, i \neq j \\ V(N) &= \left\{ x \in \mathbb{R}^N : x_1 + x_2 + x_3 \leq 2 + p \right\}. \end{aligned}$$

Then, $C(N, V) = \{(1, 1, p), (1, p, 1), (p, 1, 1)\}$.

Scarf [18] developed an important generalization of the balanced game and nonemptiness condition for the core of an NTU game. Unlike the TU game counterpart in which nonemptiness and balancedness are both mutually necessary and sufficient, Scarf's theorem proves that balancedness is only a sufficient condition for nonemptiness of the core. Subsequent works on deriving necessary conditions for nonemptiness of the core including Billera [19] followed. Billera provided a refined concept of balancedness called π -balancedness, which is proved to be a necessary and sufficient condition of

nonemptiness of the core for hyperplane NTU games only. Recently, Predtetchinski and Herings [20] provided an extended criterion called π -balancedness, which is a necessary and sufficient condition for the core of an NTU game to be nonempty.

VALUES OF THE NTU GAME

Just as in the TU game context, various “solution” concepts applicable to NTU game context have been proposed and compared. For the two-person pure bargaining problem, NBS has long been established to be the central solution concept for the purpose of prediction of the outcome of the game. On the other hand, the axiomatic approach to the solution based on the notion of “fairness” resulted in the Shapley value. The class of solutions that belongs to the values of the NTU game shares the following characteristics, which allows one to perceive them as generalization of both the NBS as well as the Shapley value [21,22]. First, it is equivalent to the NBS when applied to the pure bargaining problem. Second, it is equivalent to the Shapley value when applied to the TU game. Third, the value component for an individual is proportional to the payoff of a player (scale covariance requirement).

Two NTU value concepts that share the above characteristics have been much discussed and compared: The Harsanyi NTU value [23,24] and the Shapley NTU value [25]. Both are generalizations of the NBS for the two-person pure bargaining problem to the n -person NTU game and, simultaneously, it generalizes the Shapley value for TU game to the NTU case. The Harsanyi NTU value arose as the first explicit solution concept as a value for the NTU game. The initial work of Harsanyi [23] had a shortcoming that it did not guarantee individual rationality, which was remedied in the subsequent work [24]. The Shapley NTU value is referred to as the λ -transfer value due to its construction process. Specifically, the NTU Shapley value identifies a TU game correspondence by allowing utility transfer at the rate λ . λ is referred to as the *weight vector* or *comparison vector*. Then, it uncovers the NTU value

for the original NTU game from the Shapley value for the derived TU game. Depending on the choice of the weight vector, the Shapley NTU value might not be unique [1].

In this section, the definition of the Shapley NTU value and a few numerical examples related to values of the NTU game are presented.

Definition 6. Given an NTU game (N, V) and a nonnegative weight vector $\lambda \in \mathbb{R}_+^N, \lambda \neq 0$, define a TU game (N, v_λ) by allowing transfers of utilities at the rates λ , that is, $v_\lambda(S) = \max \{ \sum_{i \in S} \lambda_i x_i : x \in V(S) \}$ for $\forall S \subset N$. A payoff vector x is the Shapley NTU value if $x \in V(N)$ and there exists a nonnegative vector $\lambda \in \mathbb{R}_+^N, \lambda \neq 0$ such that $\lambda_i x_i = \phi_i(N, v_\lambda)$, where $\phi_i(N, v_\lambda)$ is the TU Shapley value for (N, v_λ) .

The following two famous numerical examples are utilized in order to illustrate the construction processes of the Shapley NTU value.

Example 5 [22]. Consider the NTU game (N, V) with three players $N = \{1, 2, 3\}$ and its coalitional function specification is

$$\begin{aligned} V(i) &= \{x \in \mathbb{R} : x_i \leq 0\}, \quad i = 1, 2, 3, \\ V(1, 2) &= \{x \in \mathbb{R}^2 : x_1 + x_2 \leq 36, x_1 + 2x_2 \leq 36\}, \\ V(1, 3) &= \{x \in \mathbb{R}^2 : x_1 + x_3 \leq 0\}, \\ V(2, 3) &= \{x \in \mathbb{R}^2 : x_2 + x_3 \leq 0\}, \\ V(N) &= \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 \leq 36\}. \end{aligned}$$

Except for the coalitional function for $\{1, 2\}$, the (N, V) specification is similar to that of a TU game. For example, if one assigns a value of 36 as a worth of $\{1, 2\}$ instead of the above specification of $V(1, 2)$, (N, V) could conform to a TU game definition. In order to find the Shapley NTU value for (N, V) , one needs to first find the TU game correspondence via allowance of utility transfer at the rate λ . The TU game correspondence, (N, v_λ) , is obtained by specifying its worth: $v_\lambda(S) = \max \{ \sum_{i \in S} \lambda_i x_i : x_i \in V(S) \}$. In practice, Roth [1] suggested considering the

original NTU game with side payment for TU correspondence, which is equivalent to selecting the weight vector $\lambda = (1, 1, 1)$. Then, if the Shapley value for this TU game is feasible for the original game, it can also be viewed as the value for the original game in the spirit of Nash's axiom of independence of irrelevant alternatives.

For this game, (N, v_λ) with $\lambda = (1, 1, 1)$ is obtained by assigning worth v_λ for each coalition. Specifically,

$$v_\lambda(1, 2) = v_\lambda(1, 2, 3) = 36, \\ v_\lambda(i) = v_\lambda(1, 3) = v_\lambda(2, 3) = 0, \quad i = 1, 2, 3.$$

The Shapley TU value for (N, v_λ) can now be calculated to be $(18, 18, 0)$, which is also the Shapley NTU value since it is feasible for (N, V) .

The Shapley NTU value and the Harsanyi NTU value have been often compared against each other. The key difference lies in the order of placing emphasis upon equity and efficiency during the solution construction process. In the construction of the Shapley NTU value, it first determines the worth of the TU game correspondence for each subcoalition during which the efficiency of the subcoalition is emphasized. Then, it imposes the equity restriction by applying the Shapley TU value criteria. On the other hand, the Harsanyi NTU value utilizes the λ -egalitarian solution first, which restricts the specification of the potential of each subcoalition in comparison to the Shapley NTU value. McLean [27] summarized this difference as follows: "The Shapley NTU value emphasizes coalitional utilitarianism while the Harsanyi NTU value emphasizes coalitional equity."

Example 6 [Continued; [1]]. Suppose $p \in [0, 1/2]$ for the game in Example 1. Then, Roth [1] argues that $(1/2, 1/2, 0)$ is the only sensible solution if players are rational utility maximizers since $(1/2, 1/2, 0)$ is strictly preferred by $\{1, 2\}$ and Player 3 cannot contribute to them in achieving this payoff. On the other hand, the Harsanyi NTU value is $(\frac{1}{2} - \frac{p}{3}, \frac{1}{2} - \frac{p}{3}, \frac{2p}{3})$ and the Shapley NTU value is $(1/3, 1/3, 1/3)$ with $\lambda = (1, 1, 1)$ for both. Notice that the Shapley NTU value

coincides with that for the TU counterpart of the same game. The Shapley NTU value does not capture the asymmetric payoff configuration of this NTU game since its process of deriving the TU game correspondence by λ -transfer allowance yields a symmetric game. This feature of the Shapley NTU value is better highlighted by the axiomatization work of Hart [21]. Aumann [28] rebutted the claimed inadequacy of the Shapley NTU value in this example by arguing that there are circumstances in which the solution of $(1/3, 1/3, 1/3)$ might be justified. Specifically, there might be a suspicion between Player 1 and 2 in forming a coalition by themselves since a breakaway by either one of them and forming a coalition with Player 3 would leave that player with nothing. This argument is compelling if p is close to $1/2$, that is, one might roughly interpret p as the likelihood of coalition breakdown. Of course, this argument violates the spirit of group rationality since a coalition between Player 3 and either Player 1 or 2 yields weakly less payoff for Player 1 and 2. However, a fear of coalition breakdown might work in favor of Player 3 and it might be the ground for a positive share for Player 3. Notice that the Harsanyi NTU value captures the rebut rationale of Aumann in that, payoffs for Player 1 and 2 decrease, whereas that for Player 3 increases with higher p .

AXIOMATIZATION APPROACH

An axiomatic approach to a solution concept has a general merit that an axiomatic system for a solution consisting of intuitive and rationally acceptable sets of axioms can solidify its position as a natural solution. This has been the case for prevalent solution concepts such as the NBS, the core, and the Shapley value. Once its position establishes and its applications grow, examples that demonstrate counterintuitive results of the solution are, in many cases, viewed as anomalies instead of evidences that reject the validity and usefulness of the solution. The aforementioned argument between Roth [1] and Aumann [28] is one such example. Other examples of counterintuitive results include Shapley [29] in

which the core of a symmetric market game yields a result where sellers are entitled to all profit and Maschler [30] in which the bargaining set solution predicts better than the core in recognizing the benefit of cooperation in a market game. Although these and other examples are suggestive of ill-behaving potentials of these solutions, they have not threatened their central position in the theory of game.

In addition, axiomatization approach often shed important insights into solution concepts, particularly when axioms for various solutions are compared. Such benefits are highlighted in Peleg [31] in which an axiomatization of the core of NTU game was presented.

First, by obtaining axioms for the core, we single out important properties of solutions that determine the most stable solution in game theory. Second, we may compare the system of axioms for the core with systems of other solutions, and thereby obtain more information on solutions whose definition is not simple or 'natural'.

In other words, axiomatization of a game solution can identify essential and invariant values common to several solutions and, furthermore, can enhance the understanding of the nature of solutions whose definition is not intuitive. For these reasons, many solution concepts of the NTU game, due to their generally less apparent intuitive appeal compared to their TU counterparts, have been axiomatized. A representative of such efforts includes the following: Axiomatization of Core by Peleg [32], Shapley NTU value by Aumann [28], Harsanyi NTU value by Hart [21], Maschler–Owen NTU value by de Clippel *et al.* [33] and Hart [34,35]. More recently, Peleg and Sudholter [36] provided uniform and coherent axiomatizations for most of the main solutions of the cooperative games.

A few of the insights generated by these axiomatization endeavors might be worth mentioning. Close works between Aumann [28] and Hart [21] revealed that axioms that sufficiently define Shapley NTU value and Harsanyi NTU value are almost identical. Both are characterized by common axioms

such as efficiency, scale covariance, conditional additivity, independence of irrelevant alternatives, and unanimity. Hart [21] further identified the source of difference between Shapley NTU value and Harsanyi NTU value, which was attributed to the fact that, when determining the payoff for an intermediate coalition S , the Harsanyi solution takes into account subcoalitions of S , whereas the Shapley solution considers only the grand coalition. On the other hand, Peleg [32] provided the following axioms for the core: nonemptiness, individual rationality, the reduced game property (RGP), the converse reduced game property (CRGP). Such axiomatization of the core enabled a linkage between the core and other solutions such as the Nash bargaining and prekernel.

REFERENCES

1. Roth AE. Values for games without side payments: some difficulties with current concepts. *Econometrica* 1980;48:457–465.
2. Aumann RJ, Peleg B. von Neumann-Morgenstern solutions to cooperative games without side payments. *Bull Am math Soc* 1960;66:173–179.
3. Varian HR. *Microeconomic analysis*. 3rd edn. New York: Norton; 1992.
4. Shapley AE. Utility functions for simple games. *Ann Math Study* 1953;28:307–317.
5. Nash J. The bargaining problem. *Econometrica* 18 1950;18(2):155–162.
6. Nash J. Two-person cooperative games. *Econometrica* 1953;21:128–140.
7. Shapley LS. A value for n -person games, in contributions to the theory of games II (*Annals of Mathematics Studies*, 28). In: Kuhn HW, Tucker AW. Princeton: Princeton University Press; 1953. pp. 307–317.
8. Roth AE. The Shapley value as a von Neumann Morgenstern utility. *Econometrica* 1977;45:657–664.
9. Roth AE. The Nash solution and the utility of bargaining. *Econometrica* 1978;46:587–594.
10. Binmore K, Rubinstein A, Wolinsky A. The Nash bargaining solution in economic modelling. *RAND J Econ* 1986;17(2):176–188.
11. Schmitz N. Two person bargaining without threats: A review note. *Proc Oper Res Symp, Aachen* 1977;29:517–533.

12. Thomson W. Axiomatic theory of bargaining with a variable population: a survey of recent results. In: Roth AE, editors. *Game theoretic models of bargaining*. Cambridge (NY): Cambridge University Press; 1985; pp. 233–258.
13. Thomson W, Lensberg T. Axiomatic theory of bargaining with a variable number of agents. Cambridge (NY): Cambridge University Press; 1989.
14. Morgenstern O, von Neumann J. *Theory of games and economic behavior*. Princeton: Princeton University Press; 1944.
15. Weber RJ. Games in coalitional form. *Handbook Game Theor Econ Appl* 1994;2:1285–1303.
16. Aumann RJ. The core of a cooperative game without side payments. *Trans Am Math Soc* 1961;98:539–552.
17. Gillies DB. Some theorems on n-person games. Ph.D. Thesis, Princeton University, 1953.
18. Scarf HE. The core of an n person game. *Econometrica* 1967;35:50–69.
19. Billera LJ. Some theorems on the core of an n-person game without side payments. *SIAM J Appl Math* 1970;18:567–579.
20. Predtetchinski A, Herings PJ. A necessary and sufficient condition for the non-emptiness of the core of a non-transferable utility game. *J Econ Theor* 2004;116:84–92.
21. Hart S. An axiomatization of Harsanyi's non-transferable utility solution. *Econometrica* 1985;53(6):1295–1313.
22. Hart S. A comparison of non-transferable utility values. *Theor Decis* 2004;56:35–46.
23. Harsanyi JC. A bargaining model for the cooperative n-person game, in contributions to the theory of games. 4th edn. In: Tucker AW, Luce RD. Princeton: Princeton University Press; 1959. pp. 325–355.
24. Harsanyi JC. A simplified bargaining model for the n-person cooperative game. *Int Econ Rev* 1963;4:194–220.
25. Shapley LS. Utility comparison and the theory of games, in *La Decision*. Paris: Editions du CNRS; 1969. pp. 251–263.
26. Owen G. Values of games without side payments. *Int J Game Theor* 1972;1:94–109.
27. McLean RP. Values of non-transferable utility games. *Handbook Game Theor Econ Appl* 2002;3:2077–2120.
28. Aumann RJ. On the non-transferable utility value: a comment on the Roth-Shafer examples. *Econometrica* 1985;53:667–677.
29. Shapley LS. Solutions of compound simple games. Dresher M, Shapley LS, Tuckers AW, editors. Princeton: Princeton University Press; 1959. pp. 267–305.
30. Maschler M. An advantage of the bargaining set over the core. *J Econ Theory* 1976;13:184–192.
31. Peleg B. Axiomatizations of the Core. In: Aumann RJ, Hart S, editors. *Handbook of game theory with economic applications*, 1992. Vol. 1. pp. 397–412.
32. Peleg B. An axiomatization of the core of cooperative games without side payments. *J Math Econ* 1985;14:203–214.
33. deClippel G, Peters H, Zank H. Axiomatizing the Harsanyi solution, the symmetric egalitarian solution, and the consistent shapley solution for NTU-Games, mimeo.. Springer; 2002.
34. Hart S., On prize games. In: Megiddo N, editor. *Essays in game theory*. Berlin: Springer-Verlag; 1994. pp. 111–121.
35. Hart S. An axiomatization of the consistent non-transferable utility value, center for rationality DP-337, mimeo. The Hebrew University of Jerusalem; 2003.
36. Peleg B, Sudholter P. *Introduction to the theory of cooperative games*. 2nd edn. Berlin (NY): Springer; 2007.
37. Shafer WJ. On the existence and interpretation of value allocation. *Econometrica* 1980;48: 467–476.

COOPERATIVE GAMES WITH TRANSFERABLE UTILITY

JENNIFER WILSON
The New School for Liberal
Arts, Eugene Lang College,
New York, New York

The theory of cooperative games was first introduced by von Neumann and Morgenstern who devoted the second half of their seminal book *Theory of Games and Economic Behavior* [1] to the theory of n -person games. As they noted, whenever a game has more than two players, the players' interests cannot be directly opposed. Thus, two or more players may find that they benefit by correlating their strategies. Since the underlying assumption behind noncooperative game theory is that players chose their actions without consultation, this leads to a new concept of a game in which players have the ability to communicate and make binding agreements. Under these assumptions, a number of new questions emerge, namely, which coalitions form and what role do these coalitions play in the players' strategies and in the outcome of the game. The following example illustrates these ideas.

Example 1 [Three-Person Game [2]]. Consider the zero-sum game presented in Table 1.

If the players have no opportunity to cooperate, they are likely to play according to their Nash equilibrium strategies (see *Nash Equilibrium (Pure and Mixed)*). There are two pure-strategy Nash equilibria in which players 1, 2, and 3 choose actions A, B , and A respectively or actions A, A , and B respectively. These yield the two outcomes $(-4, 1, 3)$ and $(-2, -1, 3)$. If two of the players can coordinate their strategies however, they will have a choice of four correlated strategies, each yielding different payoffs depending on the action of the remaining player. For instance, if players 2 and 3 form a coalition,

adding their payoffs together yields the two-person zero-sum game given in Table 2.

In this game, the principle of iterated dominance suggests that player 1 will choose A while the coalition of players 2 and 3 will choose action BA yielding payoffs of $(-4, 4)$. Similarly, if players 1 and 2 form a coalition, it results in another two-person zero-sum game against player 3 and has a unique minimax payoff using probabilistic strategies (see *Saddlepoints and von Neumann Minimax Theorem*), as does the game of players 1 and 3 against player 2. The payoffs of each of these games is given below, along with a breakdown of each coalition's utility among its members, assuming that members use their equilibrium strategies.

$$v(1) = -4 \quad v(23) = 4$$

utility for each player: $(-4, 1, 3)$

$$v(2) = -1.5 \quad v(13) = 1.5$$

utility for each player: $(-0.58, -1.5, 2.08)$

$$v(3) = -4.40 \quad v(12) = 4.40$$

utility for each player: $(-0.64, 5.04, -4.40)$

Based on the values that each coalition attains, it is clear that it is better to be a member of a coalition than to be left out. Players 1 and 2 would prefer to form a coalition with each other while player 3 would prefer to join player 2. However, if we examine each player's utility within the coalitions, we see that the situation is not that simple. Player 1 would do slightly better forming a coalition with player 3 who, knowing this, would prefer to form a coalition with player 1 over remaining alone. But what if player 2 could induce player 1 to stay with her by offering a *side payment* in which she transfers some of her utility to player 1? Player 3 might respond by making a counteroffer, which player 2 might offer to top. If a bidding war ensued, player 2 would win as she has enough utility to outbid player 3 while still making it advantageous for players 1

Table 1. Example 1: Three-Person Game

		Player 2					
		A			B		
Player 1	A	1	-2	1	-4	1	3
	B	3	1	-4	-5	10	-5

Player 3 = A

		Player 2					
		A			B		
Player 1	A	-2	-1	3	2	-4	2
	B	-6	12	-6	3	-1	-2

Player 3 = B

Table 2. Example 1: Player 1 versus Players 2 and 3

		Players 2 and 3							
		AA		AB		BA		BB	
Player 1	A	1	-1	-2	2	-4	4	2	-2
	B	3	-3	-6	6	-5	5	3	-3

and 2 to stay together rather than to defect individually to join player 3.

Considerations such as these led von Neumann and Morgenstern to consider coalitions and the values they can achieve as the fundamental objects on which to base a theory of cooperative games. The purpose of cooperative game theory as they conceived it, was to use the information embedded in the coalitional values to determine appropriate payoffs for each player, thereby defining the *solution* of a cooperative game. Since the publication of their book, cooperative game theory has focused primarily on defining and axiomatizing different solution concepts (see **The Shapley Value and Related Solution Concepts**). Other approaches to cooperative game theory include developing specific bargaining protocols and determining their equilibrium outcomes [3], and solving the general n -person bargaining problem.

Underlying the classical formation of cooperative game theory, are a number of assumptions. The first is that the players can communicate with one another and make binding agreements. The second is that they can make unlimited side payments to each other. As in the analysis of Example 1, this in effect allows them to redistribute the value of a coalition among the players in the coalition in any way they wish. Games in which this is allowed are called games with *transferable utility* or TU games. In TU games, it is reasonable to define the

value of a coalition in terms of a single number, with the understanding that the value corresponds to a whole set of possible payoffs among members of the coalition. Games in which players are not allowed to make side payments are called games with *nontransferable utility* or NTU games. In NTU games, the value of a coalition is defined explicitly as a set of allowable allocations to members of the coalition. In this article, we will focus on TU games. (NTU games are discussed in **Cooperative Game Theory with Nontransferable Utility**.) There are many excellent introductions to cooperative game theory including Refs 2, 4–8, as well as [1]. Our treatment adapts material from all of them. For additional resources, see the section titled “Extensions of Cooperative Game Theory” at the end of this article.

COOPERATIVE AND STRATEGIC FORM GAMES

A cooperative game is defined in terms of its *characteristic function* v which assigns a real number or value to each coalition of players in the game. We will assume throughout that the number of players is finite.

Definition 1. Let $N = \{1, 2, \dots, n\}$ be a set of players. A cooperative game is a function $v : 2^N \rightarrow \mathbb{R}$ such that $v(\emptyset) = 0$.

Note that when the number of players in a coalition is small, we will use notation such

as $v(i)$ to denote $v(\{i\})$ or $v(ij)$ to denote $v(\{i, j\})$, and so on.

From the example in the previous section, it is easy to see how a constant-sum strategic game can be converted into a cooperative game. For each coalition S , consider the two-person constant-sum game between S and $N \setminus S$. Assume that the utility of each coalition is equal to the sum of the utilities of the players in the coalition. Then let $v(S)$ be the value given to S when the two coalitions play their minimax strategies.

This definition of v captures the essence of the underlying constant-sum game since a coalition's interests are always in opposition to the interests of its complement. As the players in the complementary coalition $N \setminus S$ seek to maximize their own payoff, they automatically diminish the payoffs to S . Thus $v(S)$ represents the value that coalition S can reasonably expect for itself: a true "safety" value. Moreover, it is clear that the characteristic function will satisfy the following properties:

- constant-sum property: $v(S) + v(N \setminus S) = v(N)$ for all $S \subset N$,
- superadditivity: $v(S) + v(T) \leq v(S \cup T)$ for all $S, T \subset N$ such that $S \cap T = \emptyset$.

The first property stems from the definition of the value of a zero-sum game. The second property follows from the observation that if the players in S can guarantee themselves collectively at least $v(S)$ and the players in T can guarantee themselves at least $v(T)$, then by playing those strategies, the players in $S \cup T$ together can guarantee themselves at least $v(S) + v(T)$; by coordinating their strategies, they may in fact do better.

When the strategic game is not constant-sum, the relationship between the strategic game and its characteristic function is not as clear. The players' interests may not be directly opposed, and the highest utility of $N \setminus S$ may not correspond to the worst outcome for S . Nevertheless, von Neumann and Morgenstern suggested that the characteristic function for a non constant-sum game be derived in a manner analogous to that of a constant-sum game. Formally, this is achieved by the addition of a fictitious player

who has no strategic role but whose utilities are equal to the negative of the sum of the other players' utilities. This transforms the non constant-sum game into a zero-sum game. When the characteristic function is defined as above, each coalition S of original players is then given the value that the coalition receives by playing its minimax strategy in the zero-sum game based on the sum of the coalition members' utilities.

Example 2 [The Single Commodity Market [1]].

A seller (player 1) possesses an object that she values at c_1 . Players 2 and 3 are potential buyers who value the object at c_2 and c_3 respectively, where $0 \leq c_1 \leq c_2 \leq c_3$. Viewed as a strategic game, player 1 has the option of offering the object to player 2 or 3 and naming a price. The buyers have the option of suggesting their own price to player 1. The object changes hands only if player 1's offer to one of the buyers is equal to that buyer's offer; otherwise the object remains with player 1. Assume that the players' utilities, if no transaction takes place, are $(u_1, u_2, u_3) = (c_1, 0, 0)$, and if the object is sold to player i for price x , then $u_1 = x$, $u_i = c_i - x$ and the utility of the remaining player is unchanged.

What are the safety values of each coalition? Player 1 can guarantee herself at least c_1 (but no more) by offering an unacceptable price to the other two players. Players 2 and 3 are guaranteed no more than their current utility value of 0, either individually or together in a coalition. If players 1 and 2 form a coalition, they can guarantee themselves c_2 by agreeing on a price x and hence netting a total utility of $x + (c_2 - x) = c_2$. Similarly, the coalition of players 1 and 3 is guaranteed a value of c_3 , as is the coalition of all three players together. Hence, the characteristic function for this game is

$$\begin{aligned} v(1) &= c_1, & v(2) &= v(3) = v(23) = 0 \\ v(12) &= c_2, & v(13) &= v(123) = c_3. \end{aligned}$$

In this example, the characteristic function does not satisfy the constant-sum property but is still superadditive. In fact, any cooperative game whose characteristic function is derived from a strategic form game

in this manner will be superadditive by the previous argument.

The reverse implication is also true: any characteristic function v satisfying superadditivity can be interpreted as the characteristic function of a strategic game. Moreover if v is constant-sum, the corresponding strategic game will also be constant-sum [1]. To see this, suppose v is a superadditive cooperative game on a set of players N . We construct an explicit strategic game G having v as its characteristic function. Let the action of each player i in the game G be defined by picking a subset $S_i \subset N$ to which that player belongs. Once these are fixed, call a subset S a *ring* if $\{j : S_j = S\} = S$, meaning that each player in S chooses S as their subset. Consider the set of rings $\{\tilde{S}_1, \dots, \tilde{S}_m\}$. By definition, they are pairwise disjoint. Thus if $\tilde{S}_0$ is the set of players not in a ring, then $\{\tilde{S}_0, \tilde{S}_1, \dots, \tilde{S}_m\}$ forms a partition of N .

Now, let the utility functions for each player i in the game G be given by

$$u_i(S_1, \dots, S_n) = \begin{cases} \frac{1}{|\tilde{S}_k|} v(\tilde{S}_k) & \text{if } i \in \tilde{S}_k, k \neq 0, \\ v(i) & \text{if } i \in \tilde{S}_0. \end{cases}$$

This completes the definition of G . Let $\tilde{v}$ be the characteristic function of this game. We claim $\tilde{v} = v$. To see this, let S be a coalition and consider the value of $\tilde{v}(S)$. Now the members of S can guarantee that S will be a ring by each choosing S as their subset. If they do so, each member of S will receive $u_i = \frac{1}{|S|} v(S)$; collectively, they will receive $v(S)$. Thus the coalition S can guarantee itself $v(S)$, and hence $\tilde{v}(S) \geq v(S)$. Likewise, by a similar argument, the players in $N \setminus S$ can guarantee themselves $v(N \setminus S)$ by ensuring that $N \setminus S$ is a ring. Thus at worst, S must be a union of rings $S'_1, S'_2, \dots, S'_p$ and singletons S'_0 . It follows as before, that each $i \in S'_k$ receives $u_i = \frac{1}{|S'_k|} v(S'_k)$; together the members of S'_k receive $v(S'_k)$. Further, each $i \in S'_0$ receives $v(i)$; together they receive $\sum_{i \in S'_0} v(i)$. Since these sets are disjoint and v is superadditive, this implies $\tilde{v}(S) = \sum_{j=1}^p v(S'_j) + \sum_{i \in S'_0} v(i) \leq v(S)$. Hence, $\tilde{v}(S) = v(S)$. If in addition v is constant-sum, then so will be the strategic game.

The correspondence between strategic games and cooperative games is not one-to-one. Any characteristic function that is not superadditive cannot come from a strategic game, if the value of a coalition is defined as above by the coalition's safety value. Also, a characteristic function may correspond to many strategic games.

UTILITY THEORY AND MARKET GAMES

The ability to make side payments in cooperative TU game is often stated without specific reference to how or in what units the side payments are made. But in order for side payments to be meaningful: (i) there must be a single commodity (such as money) that can be transferred between players; (ii) the players must have an unlimited supply of it; and (iii) the players' rate of exchange of the commodity must be equal. Additionally, a player's utility function must reflect the outcome of the game as well as the money received or given. Aumann [9] proved that these assumptions imply that each player's utility u must be separable and increasing in money. This means that if P is the set of outcomes in the underlying strategic game, then u must be a function on $P \times \mathbb{R}$ of the form

$$u(p, \chi) = w(p) + c\chi,$$

where w depends only on p , $\chi \in \mathbb{R}$ represents the commodity being exchanged and $c \geq 0$. Now in general, a utility function can only be specified up to a positive linear transformation, so that if u represents the utility function of a player then so does $w = \alpha u + \beta$ for all real numbers $\alpha > 0$ and β . Thus, by possibly rescaling each player's utility function, we can assume that $c = 1$ for every player. Hence players can exchange money in a one-to-one way and the total amount of money in a game remains constant after side payments. If the value of $v(S)$ is expressed in units of money, its redistribution will reflect side payments as well as transfers of players' utilities.

These restrictions on players' utility functions are appropriate for many games including market and cost allocation games when all the players in the game have comparable

resources. But there are situations in which a player's utility is not linear in money. The utility of a wealthy individual may grow more slowly in money than a person with fewer resources. Additionally, there are many situations in life when side payments are either restricted or expressly forbidden, as when, for instance, anti-kickback laws are involved. In these cases, NTU game theory must be applied.

The relationship between utility and money in TU games is made explicit in the class of market games introduced by Shapley and Shubik [10].

Example 3 [Market Games]. A market for $N = \{1, 2, \dots, n\}$ traders and m commodities is defined by a quadruple $(N, \mathbb{R}_+^m, A, U)$ where $\mathbb{R}_+^m$ is the commodity space, $A = \{\mathbf{a}_i \in \mathbb{R}_+^m : i \in N\}$ is a set of initial allocations of the commodities among the traders, and $U = \{u_i : i \in N\}$ is a set of continuous and concave utility functions. Assume that u_i can be written as

$$u_i(\mathbf{x}, \chi) = w(\mathbf{x}) + c_i \chi,$$

where each w_i is continuous and concave on $\mathbf{x} \in \mathbb{R}_+^m$, $c_i \geq 0$, and $\chi \in \mathbb{R}$ is money. By possibly rescaling and translating the utility function (as described above), we can assume that $c_i = 1$ for all players i and that initially the players have no money.

Given a subset $\emptyset \neq S \subset N$, a *feasible S-allocation* is a set of allocations $\mathbf{x}_S = \{\mathbf{x}_i \in \mathbb{R}_+^m : i \in S\}$ such that $\sum_{i \in S} \mathbf{x}_i = \sum_{i \in S} \mathbf{a}_i$. Let X^S be the set of feasible S-allocations. (In practice, this allocation is accompanied by a sequence of side payments χ_i to player i such that $\sum_{i \in S} \chi_i = 0$. The total utility to members of S from such an allocation will then be $\sum_{i \in S} u_i(\mathbf{x}_i, \chi) = \sum_{i \in S} w_i(\mathbf{x}_i) + \sum_{i \in S} \chi_i = \sum_{i \in S} w_i(\mathbf{x}_i)$.)

Finally, for each S , let

$$v(S) = \max_{\mathbf{x} \in X^S} \left\{ \sum_{i \in S} w_i(\mathbf{x}_i) \right\}.$$

Example 4 [The Single Commodity Market Revisited]. The single commodity market with three players is a market game

with $m = 1$ and an initial allocation of $a_1 = 1$ and $a_2 = a_3 = 0$. The players' utilities are given by

$$u_i(x, \chi) = c_i x + \chi.$$

Consider the coalition $S = \{1, 2\}$, for example. Its feasible allocations will have the form $x_1 = x$ and $x_2 = 1 - x$ for some $0 \leq x \leq 1$. Since player 2 values the object more than player 1, the allocation resulting in maximum utility for the coalition S corresponds to $x = 0$. Hence $v(S) = c_1 \times 0 + c_2 \times 1 = c_2$. The other values of the characteristic function are derived in similar fashion.

Another of the best-known market games is the glove game [11].

Example 5 [Glove Game]. Let $N = N_1 \cup N_2$ be the union of two nonempty disjoint subsets. Suppose the number of commodities $m = 2$, and the initial allocations $A = \{\mathbf{a}_i \in \mathbb{R}_+^2 : i \in N\}$ are $\mathbf{a}_j = (1, 0)$ for each $j \in N_1$ and $\mathbf{a}_k = (0, 1)$ for each $k \in N_2$. Let the utility function w_i for player i be given by $w_i = \min\{x_1, x_2\}$ where (x_1, x_2) is any initial or subsequent allocation. Then $(N, \mathbb{R}_+^2, A, U)$ is a market and it is easily seen that

$$v(S) = \min\{|S \cap N_1|, |S \cap N_2|\}$$

for all $S \subset N$. This game is called the *glove game* for its analogy to a market for pairs of gloves in which initially some players have a left-handed glove, the remainder have a right-handed glove, and the value of any coalition S is equal to the total number of pairs of gloves $v(S)$ among its members.

PROPERTIES OF COOPERATIVE GAMES

As noted in the section titled "Cooperative and Strategic Form Games," cooperative games that are derived from strategic form games will be superadditive. But there are games of interest that are not superadditive—in particular those modeling situations where antitrust laws prevent players from creating coalitions that are too large. In such cases, the characteristic function will reflect the fact that is not always advantageous to form coalitions.

We will assume, unless otherwise stated, that v is superadditive. Note that in this case, $\sum_{i \in N} v(i) \leq v(N)$. If $\sum_{i \in N} v(i) = v(N)$, then the game is additive: players gain nothing by forming coalitions, and hence should be allocated the value $v(i)$.

Definition 2. A cooperative game v is called *inessential* if $\sum_{i \in N} v(i) = v(N)$. A cooperative game v is *essential* if $\sum_{i \in N} v(i) < v(N)$.

Definition 3. A cooperative game v is *monotonic* if $v(S) \leq v(T)$ for every $S \subset T$.

Many solution concepts in cooperative game theory are built around a player's *marginal contribution* to a coalition S . If $i \notin S$, this is given by $v(S \cup i) - v(S)$. If v is monotonic, a player's marginal contributions will always be nonnegative.

Definition 4. A cooperative game v is *convex* if $v(S) + v(T) \leq v(S \cup T) + v(S \cap T)$ for all $S, T \subset N$.

Convexity is equivalent to the following statement [6]:

$$\begin{aligned} \text{If } S \subsetneq T \text{ and } i \notin T, \text{ then} \\ v(S \cup i) - v(S) \leq v(T \cup i) - v(T). \end{aligned}$$

In other words, the larger the coalition, the larger the marginal contribution of potential members to the coalition.

It is clear that convexity implies superadditivity. Superadditivity however does not imply convexity.

Example 6. Let v be the cooperative game on three players given by

$$\begin{aligned} v(1) = 1 \quad v(2) = v(3) = v(23) = 0 \\ v(12) = 4 \quad v(13) = 2 \quad v(123) = 4. \end{aligned}$$

Then v is an essential superadditive game but v is not convex since $v(12) + v(13) > v(123) + v(1)$.

Superadditivity also does not imply monotonicity unless v is *positive* (its range

restricted to nonnegative numbers); nor does monotonicity imply either superadditivity or convexity. The following two examples demonstrate why these implications fail.

Example 7. Let v be the cooperative game on two players given by

$$v(1) = -2 \quad v(2) = 1 \quad v(12) = 0.$$

Then v is an essential superadditive game but v is not monotonic since $v(2) > v(12)$.

Example 8. Let v be the cooperative game on three players given by

$$\begin{aligned} v(1) = v(2) = v(3) = 2 \\ v(12) = v(13) = v(23) = 3 \quad v(123) = 4. \end{aligned}$$

Then v is a monotonic game but v is not superadditive since $v(1) + v(23) > v(123)$.

Definition 5. A cooperative game v is *symmetric* if $v(S)$ depends only on the size $|S|$ of S for all $S \subset N$.

Definition 6. A cooperative game is *simple* if it is monotonic and if $v(S) = 0$ or 1 for all $S \subset N$.

One of the most useful concepts for classifying cooperative games is that of strategic equivalence. Strategic equivalence arises because, as with strategic games, cooperative games may have different definitions yet yield the same strategic considerations.

Definition 7. Let N be a set of players. Two cooperative games v and w on N are called *strategically equivalent* if there exist constants $a > 0$ and $\{b_i : i \in N\}$ such that for all $S \subset N$,

$$w(S) = av(S) + \sum_{i \in S} b_i.$$

Suppose v is a (superadditive) cooperative game corresponding to a strategic game G . Since the utility functions of the players in G are defined only up to positive linear transformation, a reasonable question to ask is how

v is affected by a change in those utility functions. Since side payments are permitted, the only transformations of the utility functions allowed are those of the form

$$\tilde{u}_i(p, \chi) = a[u_i(p, \chi)] + b_i = aw(p) + a\chi + b_i$$

for all outcomes p and money χ , where $a > 0$ is the same for all players i . We can interpret this transformation by saying each player receives an additional b_i and that the unit of exchange has been scaled up by a factor of a . Summing these over a coalition S yields a new characteristic function $\tilde{v}$ such that

$$\begin{aligned} \tilde{v}(S) &= \sum_{i \in S} \tilde{u}_i(\mathbf{x}, \chi) = a \sum_{i \in S} u_i(\mathbf{x}, \chi) \\ &\quad + \sum_{i \in S} b_i = av(S) + \sum_{i \in S} b_i. \end{aligned}$$

Thus, the new cooperative game $\tilde{v}$ is strategically equivalent to v .

Strategic equivalence defines an equivalence relationship on cooperative games. Two members of each equivalence class of particular note are the *(0)-normalization* and the *(0,1)-normalization* of a characteristic function v . The *(0)-normalization*, given by $w_1(S) = v(S) - \sum_{i \in S} v(i)$, satisfies $w_1(i) = 0$ for each i . The *(0,1)-normalization*, given by $w_2(S) = \frac{v(S) - \sum_{i \in S} v(i)}{v(N) - \sum_{i \in N} v(i)}$, also satisfies $w_2(N) = 1$. Both convexity and superadditivity are preserved under strategic equivalence, however monotonicity or lack of monotonicity is not [6]. In fact the *(0,1)-normalization* of the game in Example 7 is the monotonic game $v(1) = v(2) = 0$ and $v(12) = 1$. A more useful property that is preserved under strategic equivalence is *zero-monotonicity*. A cooperative game v is zero-monotonic if its *(0,1)-normalization* is monotonic.

The set of cooperative games on a player set N is endowed with the usual addition and scalar multiplication of the characteristic functions, and hence forms a vector space. It has a basis given by the set of unanimity games v_T , defined as follows:

Definition 8. Given a subset $\emptyset \subset T \subset N$, the unanimity game v_T is given by

$$v_T(S) = \begin{cases} 1, & \text{if } T \subset S, \\ 0, & \text{else.} \end{cases}$$

In fact, each cooperative game v can be written as $v = \sum_{T \subset N} c_T v_T$, where $c_T = \sum_{\{T' : T' \subset T\}} (-1)^{|T|-|T'|} v(T')$ [4].

Finally, there are several oft-used methods to combine cooperative games into new games.

Definition 9. The *dual* of a cooperative game v is the cooperative game v' , where $v'(S) = v(N) - v(N \setminus S)$ for all $S \subset N$.

If $v(S)$ represents the value that S can guarantee itself, $v'(S)$ represents the portion of the total value of the game $v(N)$ that cannot be claimed by $N \setminus S$. Clearly, a game is self-dual if and only if it is constant-sum.

Definition 10. Let $N' \subset N$ be a subset of players. The *subgame* v' corresponding to N' is given by $v'(S') = v(S')$ for all $S' \subset N'$.

Definition 11. Let $N_1, N_2, \dots, N_m$ be disjoint sets of players, and let $w_1, w_2, \dots, w_m$ be simple games with player sets $N_1, N_2, \dots, N_m$ respectively. Let v be a positive game on the player set $N_0 = \{1, 2, \dots, m\}$. Then the *composition* of v with the w_i is the cooperative game u on the player set $N = N_1 \cup N_2 \cup \dots \cup N_m$ given by

$$u(S) = v(\{j : w_j(S \cap N_j) = 1\})$$

for all $S \subset N$. The composition is denoted by $u = v[w_1, \dots, w_n]$.

COST GAMES

A cost game is one where the players have some common goal (such as gaining access to a public service or building infrastructure) and may obtain some benefit from sharing costs. Given a set of players N and a subset $S \subset N$, let $c(S)$ denote the costs associated with all the members of S receiving the public benefit. Then the associated cost-savings game is given by the characteristic function v where

$$v(S) = \sum_{i \in S} c(i) - c(S).$$

Table 3. Cost and Cost-Savings Functions for Examples 9 and 10

Subsets (S)	Tennessee Valley Authority		Minimum Cost Spanning Tree	
	Cost [$c(S)$]	Savings [$v(S)$]	Cost [$c(S)$]	Savings [$v(S)$]
$\emptyset$	0	0	0	0
A	163,520	0	20	0
B	140,826	0	40	0
C	250,096	0	30	0
AB	301,607	2,739	40	20
AC	378,821	34,795	40	10
BC	367,370	23,552	40	30
ABC	412,584	141,858	50	40

The function v represents the *savings* each coalition gains through collaborating on costs.

Example 9 [Tennessee Valley Authority [12]]. In the 1930s, the Tennessee Valley Authority proposed a project to build a series of dams and reservoirs along the Tennessee River to improve navigation, control flooding, and generate hydroelectric power. Because the costs involved in attaining the project's goals were interdependent, estimations were made for the costs of each possible combination of goals. Let A, B, and C stand for the goals of improved navigation, flood control, and hydroelectric power respectively; these act as the players in a cost allocation game. The associated cost and cost-savings functions are listed in the first three columns of Table 3.

In any cost-savings game, $v(i) = 0$ for all individual players i . The characteristic function v will be superadditive if and only if c is *subadditive* ($c(S) + c(T) \geq c(S \cup T)$ for all disjoint S and T). It will also be convex if and only if c is *concave* ($c(S) + c(T) \geq c(S \cup T) + c(S \cap T)$) [8].

Example 10 [Minimum Cost Spanning Tree [8]]. Three towns A, B, and C wish to receive electrical power from a power company O. The costs of connecting each town to the power company are given by 20, 40, and 30 respectively. Also, the cost of connecting A to B is 20, connecting A to C is 20, and connecting B to C is 10. These facts can be modeled by a graph with vertices A, B, C, and

O and edges weighted with the cost between corresponding vertices. The cost function of a coalition is found by determining the minimal cost tree spanning O and the towns in the coalition. Towns 2 and 3, for instance, can be connected in a number of ways, the cheapest of which uses the edges OC and CB for a cost of $30 + 10 = 40$. The cost and cost-savings functions are listed in the last two columns of Table 3.

For more on cost allocation games, see **Power Indices**. A good survey of cost allocation can be found in Ref. 13 and an analysis of minimum cost spanning trees can be found in Ref. 14.

SIMPLE GAMES

Simple games are most often used to model voting situations in which the object of the game is to approve a bill or decision. Coalitions S such that $v(S) = 1$ are called *winning* coalitions; the remainder are *losing* coalitions. For a simple game not to be superadditive, there must be two disjoint coalitions S and T such that $v(S) + v(T) > v(S \cup T)$. This can only happen if $v(S) = v(T) = 1$. A game such as this is called *improper* and is clearly undesirable from a legislative point of view since two separate coalitions could pass conflicting resolutions. Simple games for which this does not happen are called *proper*. A simple game is proper if and only if it is superadditive.

A simple game is *strong* if for every coalition S , either S or $N \setminus S$ is winning. A simple game that is both proper and strong is

constant-sum and hence self-dual. Such a game is called *decisive*.

The marginal contributions of a player in a simple game can take on only the values 0 or 1. A player $i \notin S$ such that $v(S \cup i) - v(S) = 1$ is called *critical* for that coalition since inclusion of the player is required for the coalition to be winning. A player i that is critical for no coalition S is called a *dummy*; a player i that is critical for all coalitions S is called a *dictator*. A coalition S for which each member is critical is called a *minimal winning coalition*. Since simple games are monotonic, the set of minimal winning coalitions completely defines a simple game.

The dual of a simple game is also a simple game. The winning coalitions of v' are exactly the losing coalitions of v . Hence v' may be interpreted as the associated “blocking game.”

There are a number of well-known families of simple games. The n -person *unanimity game* is the simple v game such that $v(S) = 0$ except when $S = N$. The n -person *majority game* is the symmetric game v such that $v(S) = 1$ if and only if $|S| \geq \lfloor n/2 \rfloor + 1$.

A simple game is a weighted voting game if each of the players can be assigned a nonnegative weight such that the winning coalitions are precisely those coalitions in which the total weight of the players in the coalition is equal to or greater than some predetermined quota. A majority game is a weighted voting game in which each player is given a weight of one and the quota is equal to $\lfloor n/2 \rfloor + 1$. Weighted voting games in which the quota is more than half the total weight of all the players will always be proper (although not necessarily strong).

Example 11 [United Nations Security Council]. The United Nations Security Council is made up of 5 permanent members and 10 rotating members. A coalition is winning if and only if it contains all the permanent members and has at least 9 members in total. This scenario can be represented by a weighted voting game in which the weight of each permanent member is 7, the weight of each rotating member is 1 and the quota is 39.

Simple games can be combined in several ways. Suppose v and w are simple games on the same player set N . Then $v \otimes w$ and $v \oplus w$ are the simple games on N given by

$$v \otimes w(S) = \begin{cases} 1, & \text{if } v(S) = 1 \text{ and } w(S) = 1; \\ 0, & \text{else;} \end{cases}$$

$$v \oplus w(S) = \begin{cases} 1, & \text{if } v(S) = 1 \text{ or } w(S) = 1; \\ 0, & \text{else.} \end{cases}$$

Coalitions in $v \otimes w$ are winning if and only if they are winning in both v and w ; coalitions in $v \oplus w$ are winning if and only if they are winning in at least one of v and w . It is clear from the formulation that

$$v \otimes w = u_1[v, w] \quad \text{and} \quad v \oplus w = u_2[v, w],$$

where u_1 and u_2 are the two-player games $u_1(S) = 1$ if and only if $S = N$, and $u_2(S) = 1$ if and only if $S \neq \emptyset$ respectively. These definitions can be extended in the obvious way to $v_1 \otimes v_2 \otimes \cdots \otimes v_m$ and $v_1 \oplus v_2 \oplus \cdots \oplus v_m$.

For more about simple games, see **Power Indices**. Good introductions to simple games can be found in Refs 15 and 16.

SOLUTION CONCEPTS FOR COOPERATIVE GAMES

The primary objective in traditional cooperative game theory is to determine an appropriate allocation or set of allocations of utility among a game's players based on the characteristic function. Assuming that the game is superadditive and essential, two reasonable restrictions immediately present themselves. The first is that each player receives in the allocation at least as much as the player could receive on his/her own. This is called the *principle of individual rationality*. The second is that the sum of the allocations be equal to the value of the grand coalition. This is called *efficiency* or the *principle of group rationality*. It is also an application of the Pareto principle that if an allocation among the players adds up to less than $v(N)$, it is desirable to look for an increase in allocation that does not harm any player and is strictly better for at least one player. von Neumann and

Morgenstern called any allocation meeting these requirements an *imputation* [1].

Definition 12. The vector $\mathbf{x} \in \mathbb{R}^n$ is an *imputation* if it satisfies $x_i \geq v(i)$ and $\sum_i x_i = v(N)$.

Most solution concepts focus on the set of imputations and attempt to restrict the set to a single imputation or subset of imputations based on the perceived “stability” of the allocation to threats from coalitions wanting to break away, or on some externally determined principles of fairness.

The simplest solution concept is the *core* $C(v) = \{\mathbf{x} \in \mathbb{R}^n \mid \sum_{i \in S} x_i \geq v(S) \text{ for all } S \subset N\}$. An imputation will be in the core if no coalition has an incentive to break away from the grand coalition.

A closely related concept is that of *dominance*.

Definition 13. An imputation $\mathbf{y}$ is said to *dominate* an imputation $\mathbf{x}$ if there is some coalition S such that $y_i > x_i$ for all $i \in S$ and $\sum_{i \in S} y_i \leq v(S)$.

Informally, $\mathbf{y}$ dominates $\mathbf{x}$ if it is better for all members of S and it is feasible; that is, the values of y_i for all i in S sum to at most $v(S)$. The *dominance core* $DC(v)$ is the set of undominated imputations.

Note that if $\mathbf{x}$ is dominated by $\mathbf{y}$ via the coalition S , then $\sum_{i \in S} x_i < \sum_{i \in S} y_i \leq v(S)$. Hence, $\mathbf{x}$ is not in the core. Alternatively, if $\mathbf{x}$ is not in the core then it is possible to find an imputation $\mathbf{y}$ that dominates $\mathbf{x}$. Suppose $\sum_{i \in S} x_i < v(S)$ for some coalition S . Define $\mathbf{y} \in \mathbb{R}^n$ by

$$y_i = \begin{cases} x_i + \frac{1}{S}[v(S) - \sum_{i \in S} x_i], & \text{if } i \in S; \\ v(i) + \frac{1}{N \setminus S}[v(N) - v(S) - \sum_{i \in N \setminus S} v(i)], & \text{if } i \notin S. \end{cases}$$

It is easy to see that $\mathbf{y}$ dominates $\mathbf{x}$ on S and that $\mathbf{y}$ is an imputation (if v is superadditive). Hence $C(v) = DC(v)$ for all superadditive v .

Example 12 [The Single Commodity Market Revisited]. In this example, an imputation is any (x_1, x_2, x_3) such that

$$\begin{aligned} x_1 &\geq c_1, \quad x_2 \geq 0, \quad x_3 \geq 0, \quad \text{and} \\ x_1 + x_2 + x_3 &= c_3. \end{aligned}$$

An imputation is in the core if it additionally satisfies

$$\begin{aligned} x_1 + x_2 &\geq c_2, \quad x_1 + x_3 \geq c_3, \quad \text{and} \\ x_2 + x_3 &\geq 0. \end{aligned}$$

Hence, the core consists of all allocations of the form $(x, 0, c_3 - x)$ where $c_2 \leq x \leq c_3$.

Unfortunately, the core of some games is empty.

Example 13 [Majority Rule]. Let v be the majority game on three players. So

$$\begin{aligned} v(1) &= v(2) = v(3) = 0, \\ v(12) &= v(13) = v(23) = v(123) = 1. \end{aligned}$$

For $\mathbf{x}$ to be in the core, it must satisfy

$$\begin{aligned} x_i &\geq 0, \quad x_i + x_j \geq 1, \\ \text{and} \quad x_1 + x_2 + x_3 &= 1. \end{aligned}$$

Clearly, there is no solution to this problem.

As in the example above, many cooperative games have empty cores, and hence no imputations that are undominated. von Neumann and Morgenstern [1] proposed the concept of a stable set as a solution. A set of imputations is stable if no imputation in the stable set dominates another in the set, and if every imputation outside the stable set is dominated by an imputation inside the stable set. For more on cores and stable sets. Other solutions attempt to minimize the incentive for a coalition or individual to break away from the proposed allocation(s). These include the nucleolus, the kernel, and the related bargaining set.

The best-known solution concept for cooperative games is the Shapley value [17]. The Shapley value gives to each player a weighted average of the player’s marginal contributions to each coalition; that is, a sum of terms of the form $v(S \cup i) - v(S)$. One of

the strengths of the Shapley value is that it is the unique solution concept satisfying several intuitive “fairness” criteria. It also arises as the solution to a two-person bargaining problem, and is open to a number of other useful interpretations. Applied to simple games, it is one common measure of a player’s *power* in a voting situation (see **The Shapley Value and Related Solution Concepts** and **Power Indices**).

ALTERNATIVES TO THE CHARACTERISTIC FUNCTION

One of the drawbacks to using the characteristic function to describe cooperation between players in a strategic game is that it fails to make visible the details of the players’ strategies in the original game. The following well-known example is often used to illustrate the limitations of the characteristic function in fully describing the dynamics of the players’ interactions, in particular the use of threats as bargaining positions.

Example 14 [Employer–Worker [7]]. A worker can threaten to stop working for her employer, thereby preventing the employer from benefiting from her labors. Assume the actions and payoffs of the players are as described in Table 4.

Viewed as a strategic game, it is clear that the worker’s ability to threaten the employer is limited by the severity of the penalty against herself if she does so. Indeed the Nash equilibrium of this game is for the worker to retract and the employer to resist, yielding a payoff of (0, 10).

The characteristic function of the associated cooperative game is

$$v(W) = 0, \quad v(E) = 5, \quad \text{and} \quad v(WE) = 10.$$

Table 4. Example 14: Worker versus Employer

		Employer			
		Resist		Accede	
Worker	Insist	−1000	0	5	5
	Retract	0	10	0	10

In this formulation, the worker is disadvantaged, but the much stronger position of the employer is not obvious.

Shubik [7] called strategic games that translate well into cooperative game *c-games*. Constant-sum games, in which a coalition’s interests are directly opposed to that of the coalition’s complement are *c-games*. But many non constant-sum games are not. In Example 14, for instance, it is not in the best interests of the worker to insist, hence the employer’s equilibrium payoff in the strategic game is higher than the value of $v(E)$ which is based on the assumption that the worker will choose her strategy so as to minimize her employer’s payoff. But the worker is unlikely to act in this way because she can do better by retracting. Similarly, for general non constant-sum games, it may often be in the best interests of the players to cooperate. Among non constant-sum games that are represented well by their characteristic function, Shubik identified the games of consent or games of orthogonal coalitions. In these games, players’ payoffs are determined only by the coalitions they belong to; the strategies of the complementary coalition do not affect their value. Several classes of market games fall into this category [10,11].

For other types of non constant-sum games, Shubik suggested adapting a concept from the bargaining literature to define a new characteristic function called the *extended characteristic function* h . Bargaining theory has a well-developed history beginning in 1950 when Nash introduced the two-person bargaining problem for TU games and proposed a unique solution now called the *Nash bargaining solution* [18]. The solution was later extended to NTU games by Selten [19]. Harsanyi [20] remarked that these ideas could be adapted to determine the value of a coalition in cooperative game theory by supposing the coalitions S and $N \setminus S$ to be engaged in a two-person bargaining game.

This idea uses the fact that a two-person non constant-sum game in strategic form G can be written as the sum

$$G = \frac{1}{2}[G^+ + G^-]$$

where G^+ is the symmetric game in which the players receive identical payoffs equal to the sum of the original individual payoffs and G^- is the zero-sum game in which the players receive the difference of their individual payoffs. The bargaining solution to this problem is for the players to receive the amounts

$$h(1) = \frac{1}{2} [\text{val}G^+ + \text{val}G^-] \quad \text{and} \\ h(2) = \frac{1}{2} [\text{val}G^+ - \text{val}G^-]$$

respectively, where $\text{val}G^+$ is the maximum utility that the players can receive in the game G^+ and $\text{val}G^-$ is the value of the game G^- when the players play their minimax strategies. We illustrate with Example 15.

Example 15 [Employer–Worker Revisited [7]]. The composite games G^+ and G^- are given in Table 5.

Game G^+ has a maximum value of 10 and G^- has a saddlepoint of -10 . This gives $h(W) = 0$ and $h(E) = 10$, which more closely mirrors the weakness of the worker’s position.

We illustrate another use of the extended characteristic function in the bankruptcy game [21]. As discussed by O’Neil, an early version of the bankruptcy game can be found in a passage in the Babylonian Talmud in which Rabbi Abraham Ibn Ezra illustrates how Jewish laws of inheritance can be applied to divide the estate of the biblical figure of Jacob. According to Rabbi Ibn Ezra, after his father’s death, Jacob’s eldest son Reuben produced a deed stating that Jacob had left his entire estate to him. His second son Simeon produced a similar deed stating that Jacob had left half his estate to him. Similarly his sons Levi and Judah claimed one-third and one-fourth of the estate respectively. Using the principles

described by Rabbi Ibn Ezra, this problem can be modeled as follows.

Example 16 [The Bankruptcy Game [21]]. A man dies leaving an estate valued at 120. His sons, (numbered 1, 2, 3, and 4), claim portions valued at 120, 60, 40, and 30 respectively. Assume that the estate is infinitely divisible, that each son must claim a specific section of the estate, and that any portion of the estate that is claimed by more than one son is divided evenly between the claimants.

To determine the value of the extended characteristic function for a coalition, say, $S = \{1, 3\}$, note that together Rueben and Levi can claim the entire estate worth 120. The complementary sons, Simeon and Judah, will claim nonoverlapping pieces totaling $60 + 30 = 90$. Thus the claim of Simeon and Judah will inevitably overlap with that of Levi (and Rueben). Assume the overlap with Levi has length x . Then three sons lay claim to a portion of the estate of length x , two sons lay claim to a portion of length $90 - x$, and Reuben has the only claim to a portion of length 30. Thus, Reuben and Levi will get $(2/3)x + (1/2)(90 - x) + 30 = 75 + (1/6)x$ and Simeon and Judah will get $(1/3)x + (1/2)(90 - x) = 45 - (1/6)x$, resulting in a difference of $30 + (1/3)x$. Reuben and Levi will seek to maximize x while Simeon and Judah will seek to minimize x . The coalitions’ optimal strategies, therefore, are to chose their claims randomly, resulting in an expected overlap of $x = (90 \cdot 40)/120 = 30$ and an expected difference of $30 + (1/3)30 = 40$. Thus $h(13) = (1/2)[120 + 40] = 80$ and $h(24) = (1/2)[120 - 40] = 40$. The other values can be computed similarly.

The values obtained by the extended characteristic function can be contrasted with those of v . The value $v(13)$ is based

Table 5. Example 15: Worker versus Employer Revisited

		Employer	
		Resist	Accede
Worker	Insist	−1000	10
	Retract	10	10

Symmetric Game G^+

		Employer	
		Resist	Accede
Worker	Insist	−1000	0
	Retract	−10	−10

Zero-Sum Game G^-

on the assumption that Simon and Judah act to minimize Reuben and Levi's claim. This means that they will choose their claims so that the overlap x is 0, yielding a value of $v(13) = 75 + (1/6)0 = 75$. Alternatively, if Reuben and Levi act to minimize Simon and Judah claim, they will ensure that the overlap x is 40 leading to $v(24) = 45 - (1/6)40 = 105/3$.

It can easily be seen that $h \geq v$ for all strategic games, and that when the strategic game is constant-sum, the two functions are equal [6].

NONTRANSFERABLE UTILITY COOPERATIVE GAMES

The theory of nontransferable cooperative games was introduced by Aumann and Peleg [22] to model situations when the assumptions of TU do not hold. In these circumstances, the allocations accessible to members of a coalition cannot be expressed as a single number but as a set of payoffs to each coalition member that summarize all the possible allocations available to the coalition. There are two standard ways of defining these sets, each using different underlying concepts of a coalition's "safety" value(s), usually referred to as *alpha* theory and *beta* theory. To distinguish between these two definitions of $v(S)$, we use the notation v for alpha theory and $\tilde{v}$ for beta theory.

In alpha theory, $v(S)$ is defined as the set of payoffs that the players can guarantee themselves regardless of what strategies are used by the players in $N \setminus S$. In other words, a payoff $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ is in $v(S)$ if the members of S have a set of strategies $(\delta_j)_{j \in S}$ such that for all possible strategies $(\delta_k)_{k \in N \setminus S}$ of $N \setminus S$, $u_i((\delta_j)_{j \in S}, (\delta_k)_{k \in N \setminus S}) \geq x_i$ for each player $i \in S$ where u_i is the utility function of player i .

Alternatively, in beta theory, $\tilde{v}(S)$ is defined as the set of payoffs that the opposing coalition cannot prevent. That is, a payoff $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ is in $\tilde{v}(S)$ if for every possible set of strategies $(\delta_k)_{k \in N \setminus S}$ of $N \setminus S$ there is a set of strategies $(\delta_j)_{j \in S}$ of S such

that $u_i((\delta_j)_{j \in S}, (\delta_k)_{k \in N \setminus S}) \geq x_i$ for each player $i \in S$.

Since the requirements for alpha theory are stronger than those for beta theory, $v(S) \subset \tilde{v}(S)$. If the game is TU (and is finite), the two notions coincide. However, for general NTU games, this will not be the case [4].

Assume that the set of strategies is compact and the utility functions of the players are continuous and concave. Then, regardless of which definition is used, the characteristic function (now simply denoted v) will satisfy the following properties [8]:

- For each $\emptyset \neq S \subset N$, $v(S)$ is a nonempty closed convex subset of $\mathbb{R}^n$.
- If $\mathbf{x} \in v(S)$ and $y_i \leq x_i$ for all $i \in S$, then $\mathbf{y} \in v(S)$.
- If $S \cap T = \emptyset$, then $v(S) \cap v(T) \subset v(S \cup T)$.
- For each $S \subset N$, there exists $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x}_i \leq y_i$ for all $\mathbf{x}_i \in S$.

The second condition is referred to as comprehensiveness and states that if the members of S can obtain a set of payoffs $\mathbf{x}$, then they can also obtain less. The last condition guarantees that $v(N)$ is not too big. These properties are used to motivate the following definition.

Definition 14. A cooperative game with non-transferable utility is a triple (N, v, H) where N is a finite player set, H is a convex, compact polyhedral subset of $\mathbb{R}^n$, and v is a function that assigns to each subset $S \subset N$ a subset $v(S) \in \mathbb{R}^n$ satisfying the following conditions:

- for each $\emptyset \neq S \subset N$, $v(S)$ is a nonempty closed convex subset of $\mathbb{R}^n$
- $v(S)$ is comprehensive
- $\mathbf{x} \in v(N)$ if and only if $\mathbf{x} \leq \mathbf{y}$ for some $\mathbf{y} \in H$.

Many of the properties of TU cooperative games can be translated, after suitable adjustment, into the NTU setting as well as the solution concepts (see **Cooperative Game Theory with Nontransferable Utility**).

EXTENSIONS OF COOPERATIVE GAME THEORY

In many economic applications, the number of players is so large that it is convenient to think of the players as either countably infinite or lying on some continuum I , usually identified with the interval $[0, 1]$. In the case of the continuum, the set of coalitions corresponds to a σ -algebra β of subsets of I , and the characteristic function satisfies $v : \beta \rightarrow \mathbb{R}$ where $v(\emptyset) = 0$. Differences with the finite theory emerge almost immediately. For instance, when considering the $(0, 1)$ -normalization of v , it is not appropriate to require $v(i) = 0$ because this is satisfied by, for instance, Lebesgue measure. But Lebesgue measure is additive, and this leads to a v that is not essential. Instead, the $(0, 1)$ -normalization of v is defined by the following:

- $v([0, 1]) = 1$
- $v(S) \geq \alpha$ for all $S \in \beta$
- if α is a measure such that $\alpha \leq v$ then $\alpha \equiv 0$.

The largest measure α satisfying $\alpha \leq v$ is given by $\alpha(S) = \inf \sum v(S_i)$ where the infimum is taken over all sequences of sets $S_i \in \beta$ such that $S \subset \cup S_i$. This measure is used to define the set of imputations of v when evaluating the value of a coalition. Thus, imputations are given by signed measures σ such that

$$\sigma([0, 1]) = v([0, 1]) \quad \text{and} \quad \sigma(S) \geq \alpha(S).$$

An *atom* of v is a coalition $S \in \beta$ such that $v(S) > 0$, but for all disjoint subsets $T_1 \cup T_2 = S$ either $v(T_1) = 0$ or $v(T_2) = 0$. A game v is *nonatomic* if v has no atoms. An extensive theory of nonatomic cooperative game theory has been developed for instance in Ref. 23.

Other extensions of cooperative game theory generalize the idea of a characteristic function or the way in which players participate in a coalition. Games in *partition function form* assume that the value a coalition can attain for itself depends on the coalitional structure of the remaining players. Let π be a partition of N , and let $S \in \pi$.

Then a game in partition function form is given by a real-valued function $w(S, \pi)$ which assigns to each coalition S the value it can guarantee when the players are divided into the partition π . A characteristic function for this game can be defined by

$$v(S) = \min_{\pi: S \in \pi} w(S, \pi).$$

As with NTU games, many of the properties and solution concepts of cooperative games can be extended to games in partition function form [24].

More recent extensions of cooperative game theory allow players more than two levels of participation in a coalition (either in or out of the coalition). Fuzzy games, introduced in Ref. 25, are used to model situations when players' participation in a coalition can best be described by a number $s \in [0, 1]$ where $s = 0$ corresponds to not participating at all and $s = 1$ corresponds to full participation. If participation in a coalition is interpreted as committing a quantity of resources to a cooperative project, then the intermediate levels $s \in (0, 1)$ can be interpreted as committing only a fraction of a player's full resources. A fuzzy coalition $\mathbf{s}$ in a game with N players is a vector $\mathbf{s} = (s_1, s_2, \dots, s_n) \in [0, 1]^n$ indicating each player's degree of participation. Note that every traditional coalition S is associated with the fuzzy coalition $e^S \in F^N \subset F^N$, where $e_i^S = 1$ if $i \in S$, and $e_i^S = 0$ otherwise. Let F^N stand for the set of fuzzy coalitions on N . A cooperative fuzzy game on N is a function $v : F^N \rightarrow \mathbb{R}$ such that $v(e^\emptyset) = 0$.

Similarly, there are situations when a player's ability to participate in a coalition is best described through a sequence of discrete steps. Multichoice games, introduced in Ref. 26, assume that each player i has an action space $M_i = \{0, 1, 2, \dots, m_i\}$ in which 0 corresponds to not participating and m_i corresponds to full participation. A multichoice coalition is an n -tuple $(x_1, x_2, \dots, x_n)$ where $x_i \in M_i$ for each player i . The coalitions $(0, 0, \dots, 0)$ and $(m_1, m_2, \dots, m_n)$ correspond to the empty set and the grand coalition respectively. Let $M = \prod_i M_i$ stand for the set of multichoice coalitions on N . Then, a

cooperative multichoice game on N is a function $v : M \rightarrow \mathbb{R}$ such that $v(0, 0, \dots, 0) = 0$. For good introductions to cooperative fuzzy games and cooperative multichoice games and their solutions, see Ref. 27.

REFERENCES

1. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton University Press; 1944.
2. Rapoport A. N -person game theory. Ann Arbor (MI): University of Michigan Press; 1970.
3. Ray D. A game theoretic perspective on coalition formation. Oxford: Oxford University Press; 2007.
4. Owen G. Game theory. New York: Academic Press; 1982.
5. Straffin P. Game theory and strategy. Washington (DC): Mathematical Association of America; 2004.
6. Weber RJ. Games in coalitional form. In: Aumann RJ, Hart S, editors. Handbook of game theory II. Amsterdam: Elsevier; 1994. pp. 1285–1303.
7. Shubik M. Game theory in the social sciences. Cambridge (MA): Massachusetts Institute of Technology; 1982.
8. Peleg B, Sudholter P. Introduction to the theory of cooperative games. Boston (MA): Kluwer Academic Publishers; 2003.
9. Aumann RJ. Linearity of unrestrictedly transferable utilities. Nav Res Log Q 1960; 7:281–284.
10. Shapley LS, Shubik M. On market games. J Econ Theor 1969;1:9–25.
11. Shapley LS, Shubik M. Pure competition, coalitional power, and fair division. Int Econ Rev 1969;10(3):337–362.
12. Young HP. Cost allocation. In: Aumann RJ, Hart S, editors. Handbook of game theory II. Amsterdam: Elsevier; 1994. pp. 1193–1235.
13. Young HP. Cost allocation. In: Young HP, editor. Volume 33, Fair allocation; proceedings of symposia in applied mathematics. Providence (RI): American Mathematical Society; 1985. pp. 69–94.
14. Granot D, Huberman G. Minimum cost spanning tree games. Math Program 1984; 21:1–18.
15. Taylor AD. Mathematics and politics. New York: Springer; 1995.
16. Felsenthal D, Machover M. The measurement of voting power. Cheltenham: Edward Edgar; 1998.
17. Shapley LS. A value for n -person Games. In: Kuhn HW, Tucker AW, editors. Volume 28, Annals of mathematical studies. Princeton (NJ): Princeton University Press; 1953. pp. 307–317.
18. Nash JF. The bargaining problem. Econometrica 1950;18:155–162.
19. Selten R. Valuation of n -person games. In: Dresher M, Shapley LS, Tucker AW, editors. Advances in game theory. Princeton (NJ): Princeton University Press; 1964. pp. 577–626.
20. Harsanyi JC. A bargaining model for the cooperative n -person game. In: Tucker AW, Luce RD, editors. Volume 4, Contributions to the theory of games. Princeton (NJ): Princeton University Press; 1959. pp. 324–356.
21. O'Neil B. A problem of rights arbitration from the Talmud. Math Soc Sci 1982;2: 345–371.
22. Aumann RJ, Peleg B. von Neumann–Morgenstern solutions to cooperative games without side payments. Bull Am Math Soc 1960;66:173–179.
23. Aumann RJ, Shapley LS. Values of non-atomic games. Princeton (NJ): Princeton University Press; 1974.
24. Thrall RM, Lucas WF. n -person games in partition function form. Nav Res Log Q 1963; 10:281–298.
25. Aubin JP. Cooperative fuzzy games. Math Oper Res 1981;6:1–13.
26. Hsiao CR, Raghaven TES. Shapley value for multi-choice cooperative games I. Games Econ Behav 1993;5:240–256.
27. Branzei R, Dimitrov D, Tijs S. Models in cooperative game theory: crisp, fuzzy and multi-choice games. Berlin: Springer; 2005.

FURTHER READING

- Davis M, Maschler M. The kernel of a cooperative game. Nav Res Log Q 1965;12:223–259. DOI: 10.1002/nav.3800120303.
- Driessen T. Cooperative games, solutions and applications. Kluwer: Dordrecht; 1988.
- Luce HD, Raiffa R. Games and decisions. New York: John Wiley & Sons, Inc.; 1957.
- Maschler M, Peleg B, Shapley LS. Geometric properties of the kernel, nucleolus, and

- related solution concepts. *Math Oper Res* 1979;4:303–338. DOI: 10.1287/moor.4.4.303.
- Myerson R. *Game theory: analysis of conflict*. Cambridge (MA): Harvard University Press; 1991.
- Osborne M, Rubinstein A. *A course in game theory*. Cambridge (MA): MIT Press; 1004.
- Schmeidler D. The nucleolus of a characteristic function game. *SIAM J Appl Math* 1969;17:1163–1170. DOI: 10.1137/0117107.

COORDINATION OF PRODUCTION AND DELIVERY IN SUPPLY CHAIN SCHEDULING

ZHI-LONG CHEN

Robert H. Smith School of
Business, University of
Maryland, College Park,
Maryland

Various types of coordination in a supply chain have been studied in the literature. Examples include coordination of location, inventory, and transportation decisions [1], coordination of scheduling decisions across multiple stages of a supply chain covered in article titled ***Supply Chain Scheduling: Origins and Application to Sequencing, Batching and Lot Sizing*** in this volume, coordination of capacity allocation decisions covered in article titled ***Capacity Allocation in Supply Chain Scheduling*** in this volume, and coordination of pricing, production, and inventory decisions [2]. The coordination of production and distribution decisions are discussed in this article. Production and distribution operations can be decoupled if there is a sufficient amount of inventory between them. Some companies manage these two functions independently with little or no coordination. However, this leads to increased holding costs and longer lead times of products through the supply chain. Fierce competition in today's global market and heightened expectations of customers have forced companies to invest aggressively to reduce inventory levels across the supply chain on one hand and be more responsive to customers on the other. Reduced inventory results in closer linkages between production and distribution functions. Consequently, companies can achieve cost savings and improve customer service by coordinating production and distribution decisions.

In the past two decades, a substantial amount of research has been done on various integrated production–distribution models at the *strategic* and *tactical* planning levels

(involving decisions such as location, network design, capacity, and inventory policies (see ***Multiechelon Multiproduct Inventory Management***). Many survey articles on such models have appeared in the literature, see for example, Refs 3–5. Due to the fact that production and distribution stages at a planning level are often linked by an intermediate inventory stage, almost all of the models reviewed in those survey articles involve inventory decisions in addition to production and distribution decisions.

On the contrary, the research on coordinated production–distribution models at the *detailed scheduling* level is fairly recent; the majority of the work in this area was done in the last several years. However, this area is growing very rapidly, due to the fact that an increasing number of problems encountered in many practical settings can be modeled as a coordinated production and delivery scheduling (CPDS) problem. In a typical CPDS model, it is necessary to optimize detailed order-by-order scheduling of production and finished order delivery jointly by taking into account relevant revenues, costs, and customer service levels at the individual order level. A typical CPDS problem involves both machine scheduling (for order processing) and vehicle scheduling and routing (for order delivery). Although pure machine scheduling problems and pure vehicle routing problems have been studied extensively in the past several decades [6,7], most of the work on CPDS problems was published or done in the last decade, and a large percentage of the work was published in the last five years. Whereas many other areas of supply chain management (SCM) have been well studied, the problem area of CPDS is an exciting new area that clearly needs much more research.

This article is organized as follows: in the section titled “Application Examples” some major application examples documented in the literature are reviewed, especially those that can be modeled as a CPDS problem. In the next section titled “Problem Definition,”

the necessary notation is introduced and the common structure of CPDS problems is described. In the section on classification, existing CPDS models are classified into five major classes based on some major structural characteristics of the models and an overview is given of the existing models studied in the literature. In the section titled “Solution Methods” are discussed, in brief, the major solution techniques developed in the literature for solving these models. In the final section, several problem areas that deserve more research are pointed out. Please note that the contents of this article are mainly drawn from a recent survey paper of the author [8]; interested readers are requested to refer to that paper for more details.

APPLICATION EXAMPLES

CPDS models are often motivated by applications involving make-to-order or time-sensitive (perishable, seasonal) products where finished orders have to be delivered to customers immediately or shortly after production. Consequently, there is little or no finished product inventory in the supply chain such that production and outbound distribution must be scheduled jointly in order to achieve a desired on-time delivery performance at minimum total cost. Since the planning horizon is typically short and speedy delivery of finished orders is often critical, strategic or tactical planning models are not applicable to such applications. Instead, they require detailed scheduling models that integrate production and outbound distribution of finished orders. Examples of such applications include the following:

- *Custom-Made Electronics Products Assembly and Delivery under the Assemble-to-Order Business Model* [9,10]. In a typical assemble-to-order environment for electronics products such as PCs, customers configure their orders and assembly of components does not start until customer orders are received. To be competitive, companies often promise a very short delivery lead time (e.g., 3–5 business days), which means that assembled computers have to be delivered shortly after the assembly. Hence the assembly and delivery operations are closely linked in time.
- *Preparation and Delivery of Food in Food Catering Services* [11]. To keep food fresh, food preparation can take place only after orders are known; hence, no inventory is involved and completed orders are delivered within a few hours. Thus, production operations (i.e., food preparation) and distribution operations (i.e., delivery of completed orders to customers) are linked together directly without any intermediate step.
- *Production and Distribution of Fashion Apparel and Some Toys* [12,13]. Such products typically have a large variety and a short life cycle, and are sold in a very short selling season. Because of high demand uncertainty of the products, retailers typically do not place orders until reliable market information is available shortly before a selling season. On the other hand, because there are significant markdowns for unsold products at the end of the selling season, the manufacturer runs a high risk if it starts production early, before it receives orders from the retailers. As a result, the manufacturer does not start production until orders from the retailers have been placed shortly before the selling season. Due to the fact that there is only a limited amount of production time available, to deliver the orders to the retailers as soon as possible at a low cost, the manufacturer has to schedule the production and distribution operations in a coordinated and efficient manner.
- *Production and Delivery of Perishable Products such as Ready-Mix Concrete Paste* [14] and *Industrial Adhesive Materials* [15]. Those products have a very short lifespan. After the necessary raw materials (cement, sand, gravel, water, etc.) are mixed, the resulting concrete paste solidifies and hence becomes useless within an hour. Similarly, after production, some adhesive

materials used for plywood panels will lose their strength within seven days. So in practice, it is often required that those products be delivered to the customer site immediately after the production. Hence, the production and distribution operations are linked without any idle time between them.

- *Printing and Distribution of Newspapers* [16]. Printing of a newspaper typically starts at midnight and the printed newspaper has to be delivered to customer sites before 6:00 a.m.; to save transportation cost, orders going to different destinations are normally consolidated for delivery. However, given the short time span, each delivery truck cannot wait too long at the printing facility and may have to leave before a full truckload can be formed.
- *Mail Processing and Distribution* [17]. Mails from different origins are first sorted in a mail-processing facility, according to destination. After sorting, they are delivered by trucks to different destinations (local postal offices or other mail-processing facilities). Given the time sensitive nature of mails, there is little, if any, idle time between the sorting and delivery operations.

The products and services in the last three examples are *inherently* time sensitive. For such products and services, we *have to* consider production and outbound delivery scheduling jointly if a satisfactory performance is expected. Many other products are not inherently time sensitive, but become time sensitive under certain business environments. Fierce competition in today's global market and heightened expectations of customers have forced companies to invest aggressively to reduce inventory levels across the supply chain on one hand and be more responsive to customers on the other. As a result, an increasing number of companies now adopt make-to-order (also known as assemble-to-order, build-to-order) business models which often make the otherwise time insensitive product time-sensitive. As in the first three examples described above, products in a make-to-order environment

are often custom-made and delivered to customers within a very short lead time directly from the factory, without the intermediate step of finished product inventory. The close linkage between production and outbound distribution in many make-to-order environments makes the joint scheduling of these two operations imperative.

In addition to the close production and delivery linkage, another common characteristic seen in the production and delivery of many time-sensitive products is that orders are typically distinct and have to be processed individually and the production of an order can start only after the order is received. This is evident in most make-to-order cases where orders are custom-made and there are often such a large number of configurations or product types available for customers to choose from that it is impractical to process an order before it is received. In the case of ready-mix concrete paste and industrial adhesive materials, different orders may require different recipes and specifications and hence, have to be prepared individually. In the case of newspapers, orders destined for different customer regions often require different local contents and hence there are setups between two different orders when printing. This discrete nature of orders necessitates the detailed scheduling level consideration of order processing and delivery because treating the orders in an aggregated manner will not accurately capture the dynamics of the operations.

PROBLEM DEFINITION

First, a general CPDS problem is described and associated parameters and notation are defined, followed by a description of the common structure and characteristics shared by most CPDS problems.

A General Problem

A general CPDS problem can be described as follows. At the beginning of the planning horizon, a manufacturer which has one or more plants receives a set of n orders $N = \{1, 2, \dots, n\}$ from k ($k \geq 1$) customers $K = \{1, 2, \dots, k\}$ which may be geographically

separated. Let $N_i \subset N$ be the subset of orders destined for customer i , and $n_i = |N_i|$ for $i \in K$, where $N = N_1 \cup \dots \cup N_k$, and $n = n_1 + \dots + n_k$. The manufacturer needs to first process these orders and then deliver the finished orders individually or in batches to the customers, using $v \geq 1$ delivery vehicles (e.g., vans, trucks, and air flights). Each order $j \in N$ needs to be processed by one or multiple machines depending on the specific machine configuration involved, and has a known processing time p_j if only one machine is needed, or p_{ij} on the i th machine if multiple machines are needed. In addition, each order j may be associated with a number of parameters, including, among others, a delivery due date d_j , a size q_j which is the units of capacity needed to carry order j by a delivery vehicle, a revenue R_j if order j is processed and delivered to its customer at a desired time (e.g., before or at the deadline, within the time window), and an importance weight w_j .

The delivery vehicles may be homogeneous (i.e., have the same delivery capacity, driving speed, and cost rates) or heterogeneous (i.e., one or more parameters are vehicle dependent). The number of available vehicles may be limited or infinite, and vehicle capacity may be limited (i.e., can only carry a subset of orders in a trip) or infinite (i.e., can carry any number of orders in one trip). There are generally nonzero transportation times and nonzero variable transportation costs for traveling between different locations. There may be a fixed transportation cost associated with each shipment or each vehicle.

A *schedule* (i.e., *solution*) to a given CPDS problem consists of a *production schedule* which specifies when and where each order is processed and a *delivery schedule* which specifies how many shipments are used, the departure time and traveling route of each shipment, which orders are in each shipment, and when each order is delivered. In a given schedule, a number of measurements associated with each order can be defined, such as those given below:

C_j the *completion time* of order j , which is the time when order j 's processing is completed at the plant

S_j the *departure time* of order j , which is the time when the shipment that carries order j departs from the plant
 D_j the *delivery time* of order j , which is the time when order j is received by its customer
 T_j $\max\{0, D_j - d_j\}$, the *delivery tardiness* of order j

The objective of the problem is to optimize one or a combination of some commonly used *time-based*, *cost-based*, and *revenue-based* performance measures. Examples of such measures are given below:

Examples of Time-Based Performance Measures

- Maximum delivery time of orders, that is $\max\{D_j | j \in N\}$.
- Total (weighted) delivery time of orders, that is $D_1 + \dots + D_n$ (or $w_1 D_1 + \dots + w_n D_n$).
- Total (weighted) delivery tardiness of orders, that is $T_1 + \dots + T_n$ (or $w_1 T_1 + \dots + w_n T_n$).

Examples of Cost-Based Performance Measures

- Total trip-based transportation cost (which is the sum of the costs incurred by all individual delivery trips where the cost of a delivery trip may consist of a fixed cost and a variable cost determined by the route of the shipment, for example, total distance, independent of the total size or number of orders in the shipment).
- Total production cost of the orders (applicable when this cost varies with the production schedule of the orders).

Example of Revenue-Based Performance Measure

- Total revenue of the orders processed and delivered, that is, $\sum_{j \in A} R_j$, where A is the set of processed and delivered orders (applicable when not every order can be processed or delivered in a timely manner due to production or transportation capacity limit).

Problem Structure

Most CPDS models can be precisely defined by specifying the following four components: (i) machine configuration, (ii) delivery characteristics, (iii) number of customers, and (iv) objective function.

Machine Configuration. Most existing CPDS models involve a single production plant. In all the single-plant CPDS models, order processing at the plant follows one of the four configurations described below:

- *Single-machine* configuration where all the orders are processed by a single machine.
- *Parallel-machine* configuration where there are m identical parallel machines such that each order needs to be processed only by one of the machines.
- *Flowshop* configuration where there are m machines and each order needs to be processed by all the machines sequentially.
- *Bundling* configuration where each order consists of m independent tasks to be processed independently by m dedicated machines at the plant and the m tasks of an order are bundled together for delivery once they are all completed.

In a few existing models, there are multiple plants located at different locations such that each order needs to be processed by only one of the plants, and only the orders processed at the same plant can be delivered together. In these models, assignment of orders to plants may also be a decision in addition to production and delivery scheduling.

Delivery Characteristics. Most existing models involve homogeneous delivery vehicles, with the following vehicle characteristics (number and capacity of delivery vehicles), and delivery method.

Vehicle Characteristics: Number of Vehicles Available

- A single delivery vehicle available, that is $v = 1$.

- Multiple and finite number of vehicles available, that is $2 \leq v < n$.
- A sufficient number of vehicles available such that vehicle availability is not a constraint, that is $v \geq n$.

Vehicle Characteristics: Vehicle Capacity

- Each shipment can only deliver one order.
- Each shipment can deliver multiple orders.
- Each shipment can deliver any number of orders.
- Each shipment can deliver at most Q units of total order size (in this case, each order j occupies a generally different size of q_j units of the shipment capacity).

Delivery or Shipping Methods Allowed

- *Individual and Immediate Delivery.* Each order is shipped individually and immediately after it completes processing (i.e., $S_j = C_j$ and $D_j = C_j + t_j$ for $j \in N$, where t_j is the transportation time from the plant to order j 's destination).
- *Batch Delivery by Direct Shipping.* Only orders going to the same customer can be delivered together in the same shipment.
- *Batch Delivery with Routing.* Orders going to different customers can be delivered together in the same shipment (where vehicle routing is a part of the decision).
- *Shipping with Fixed Delivery Departure Dates.* Each vehicle has a fixed delivery departure date exactly at which the vehicle must depart from the plant; this is prespecified as a given parameter (this implies that a vehicle, if used, can cover one shipment only).

Many of the cases listed can be easily justified from a practical point of view, whereas some other cases may require explanations. The case with a sufficient number of vehicles is applicable when the delivery is handled

by a third-party logistics (3PL) provider which typically owns a large number of vehicles. The individual and immediate delivery method and the batch delivery by direct shipping method are commonly used in applications of time-sensitive products, whereas a batch delivery with routing method can be used to save transportation costs. Shipping with fixed delivery departure dates is applicable when delivery is done by a 3PL service provider which picks up customer orders at fixed times in a day (e.g., 10:00 a.m., 3:00 p.m.) which cannot be changed by the user of the service.

In most existing models with capacitated vehicles, each order is assumed to have an equal size (i.e., $q_1 = \dots = q_n$) such that the vehicle capacity is simply the maximum number of orders that can be delivered together. For such models, the vehicle capacity is such that each trip can carry either one order or multiple orders. In a handful of existing models, orders are assumed to have generally different sizes. For these models, the vehicle capacity is such that each trip can carry up to a total order size of Q units.

Number of Customers. There are one or more customers, and it is assumed without loss of generality that each customer has a distinct location so that under certain circumstances only orders destined for the same customer can be consolidated for delivery.

Objective Function. Most existing CPDS models consider one or a combination of two of the following three performance measures: customer service level, total cost, and total revenue. These measures are represented, respectively, by the time-based, cost-based, and revenue-based functions defined in the section titled “A General Problem.” The objective function is to be minimized if it is a time-based function, a cost-based function, or the sum of these two, and is to be maximized if it is a revenue-based function.

The recent survey paper [8] uses a five-field representation scheme to represent each CPDS problem precisely. For interested readers, please read the survey paper for further details.

CLASSIFICATION AND OVERVIEW OF EXISTING MODELS

Existing CPDS models can be grouped into the following five classes, according to the delivery method. The models within each class share not only major structural characteristics but also some solution properties. This enables us to discuss major commonalities of the models in each class and major differences between the models in different classes:

- models with individual and immediate delivery;
- models with batch delivery to a single customer by direct shipping method;
- models with batch delivery to multiple customers by direct shipping method;
- models with batch delivery to multiple customers by routing method;
- models with fixed delivery departure dates.

A detailed review of each class of models is given in the survey paper by Chen [8]. An overview of the existing literature is given here. Reference 18 is the first paper on a CPDS problem: the problem studied by Potts involves a single machine, multiple customers, the individual and immediate delivery method with a sufficient number of delivery vehicles, and the objective of minimizing the maximum delivery time of the orders. Most of the CPDS papers with the individual and immediate delivery method are inspired by Potts [18]. Reference 19 is the first paper that considers batch delivery and transportation costs. Many of the CPDS models with batch delivery by direct shipping method that consider transportation costs are inspired by this paper. Reference 20, which considers a variety of CPDS problems with batch delivery by direct shipping method, is the first paper that addresses transportation capacity issues. This paper has inspired many researchers to study models with transportation capacity. Since 2001, many papers have appeared, many more than that in the period from 1980 to 2000. Rapidly growing interest in the CPDS area

is believed to be motivated mainly by the following facts: (i) historically, SCM research has been almost exclusively focused on planning level decisions; (ii) an increasing number of applications (as described in the first section) now require coordinated decisions at a detailed scheduling level; and (iii) CPDS models fill in the gap between the traditional SCM research and new applications.

In addition to the delivery method and number of vehicles in a given CPDS problem, several other parameters can have a major impact on complexity and solution structure. They include number of customers (single or multiple), order sizes (equal or general), and vehicle types (homogeneous or heterogeneous). Based on these parameters, the five classes of CPDS models described earlier can be segregated into several subclasses. Table 1 gives an overview of all the models and a representative set of existing literature. Based on the current state of the literature, the following observations can be made:

- Problems with batch delivery to multiple customers and problems with fixed delivery departure dates have been studied more recently and much less extensively than problems with individual and immediate delivery and problems with batch delivery to a single customer. This is mainly because the earlier two classes of problems are more difficult to solve than the latter ones.
- Only a handful of papers consider problems with heterogeneous vehicles. In such problems, vehicles may have different speed, capacity, or/and transportation cost rates. Clearly, problems with heterogeneous vehicles are more difficult to solve than problems with homogeneous vehicles.
- Only a handful of papers study problems with multiple plants (which are generally separated geographically). In such a problem, order assignment to plants may also be a part of the decision. Clearly, problems with multiple plants are more difficult to solve than problems with a single plant.
- The majority of the existing literature studies problems where all the orders have an equal size. In such problems, the capacity of a vehicle is simply the maximum number of orders that can be carried in a shipment. This avoids the bin-packing decision that is present in the case with general order sizes. Since the bin-packing problem alone is intractable (e.g., [36]), problems with general order sizes are more difficult to solve than problems with equal order sizes.

Almost all the existing CPDS models are deterministic where all the problem parameters are known in advance. This may reflect the fact that decisions at the detailed scheduling level often involve a short planning horizon such that most information about orders is known at the beginning of the horizon. Hence, it may be sufficient to use a deterministic model.

Reference 12 is the only paper that considers models where production costs are a part of the objective function. In their models, there are multiple plants located in different countries such that the processing cost of an order depends on which plant it is assigned to, and hence it is relevant to consider the total production cost of the orders. However, in most existing CPDS problems, there is a single plant with a single machine or several identical parallel machines. Furthermore, the planning horizon is typically short such that there is little or no learning effect which is often seen in longer-term production. Consequently, the production cost of an order is fixed independent of where and when it is processed.

SOLUTION METHODS

In the CPDS literature, a small subset of problems are solved in polynomial time, mainly by dynamic programming algorithms. There are a number of problems with unknown complexity, which need to be resolved in future research. Fast heuristics based on simple rules are developed in the literature for many of the proven NP-hard

Table 1. Overview of Existing CPDS Models

	Problems by Delivery Method	Number of Customers	Order Sizes	Vehicles		Representative Literature
				Type	Number	
8	Problems with individual and immediate delivery	Multiple	Equal	Homogeneous	Sufficient	18,21–23
					Limited	14
				Heterogeneous	Sufficient	24
	Problems with batch delivery to a single customer	Single	Equal	Homogeneous	Sufficient	11,12,19,25,26
					Limited	20,26–28
			General	Homogeneous	Sufficient	29
					Limited	30,31
	Problems with batch delivery to multiple customers: direct shipping method	Multiple	Equal	Homogeneous	Sufficient	13,25,32
					Limited	9
				Heterogeneous	Sufficient	33
	Problems with batch delivery to multiple customers: routing method	Multiple	Equal	Homogeneous	Sufficient	11
					Limited	9,34
			General	Homogeneous	Sufficient	16
					Limited	15
		Multiple	Equal	Heterogeneous	Sufficient	10
			General	Heterogeneous	Limited	17,35

problems. Worst-case or/and asymptotic performance are analyzed for some of the heuristics. There are also branch-and-bound or polynomial time approximation algorithms developed for a handful of the NP-hard problems. Interested readers can find more details in the literature [8]. For fundamentals of dynamic programming and branch-and-bound algorithms, see articles titled ***Computation and Dynamic Programming*** and ***Branch-and-Bound Algorithms*** in this volume, respectively.

DIRECTIONS FOR FUTURE RESEARCH

There are a number of important CPDS problem areas for which very few results exist, and yet these problems are commonly encountered in practice. Such problems clearly deserve more research. Four such problem areas are described below:

- *Problems with Fixed Delivery Departure Dates.* Over 70% of the companies worldwide now rely on 3PL providers for their daily distribution and other logistic needs. Many 3PL service providers including major ones such as UPS, FedEx, and DHL have daily fixed package pickup times. CPDS problems in such a setting involve fixed vehicle departure dates. Only a handful of papers have studied CPDS problems with fixed delivery departure dates. Such problems tend to involve many complex issues and are generally NP-hard. One can design fast heuristics for such problems and analyze their performance.
- *Multicustomer Batch Delivery Problems with Routing Method.* In practice, the majority of deliveries by trucks involve multiple destinations, and delivery consolidation across multiple stops is commonly used to save transportation costs. Vehicle routing is an important part of the decision in such applications. Only a handful of papers have studied CPDS problems involving routing decisions. All the problems studied involve either a single vehicle or an infinite number of vehicles. It will be worthwhile to study problems with multiple but a limited number of vehicles. While vehicle routing problems with time windows have been well studied in the vehicle routing literature, few papers have considered CPDS problems with both routing decisions and time-window constraints. However, due to their practicality, such problems clearly deserve more research.
- *Problems with Heterogeneous Vehicles.* The following class of problems with heterogeneous vehicles has many applications and deserves more research. In these problems, there are several kinds of vehicles which all have the same capacity but different delivery speeds and cost rates. They can be used to model applications where delivery is carried out by a 3PL provider which offers multiple delivery modes, each with a different delivery speed and a different rate. For example, many package delivery services offer overnight, one-day, or two-day delivery, and the cost rate decreases as the delivery lead time increases. As in most 3PL applications, in such problems, there is an infinite number of vehicles available. References 24 and 10 are the only two papers that have studied such problems. One can investigate extensions from Wang and Lee's problems with other objective functions and other machine configurations.
- *Online Problems with Dynamic Order Arrivals.* In many practical make-to-order environments, orders arrive dynamically and precise information about the future orders may not be known before they are received. Therefore, the CPDS problem in such environments is often dynamic and contains some uncertainty. However, almost all the existing CPDS problems studied in the literature are static and deterministic. It is true that a static and deterministic off-line model can be used repeatedly to get a dynamic solution for such applications if the model considers all the orders that have arrived so far and is implemented on a rolling horizon basis. However, such an approach can

be very time consuming and difficult to use in a real-time fashion. An alternative and more practical approach would be on-line algorithms based on simple rules which can be implemented in real time. It will also be interesting to study performance (e.g., competitive ratio) of such algorithms theoretically.

REFERENCES

1. Uster H, Keskin BB, Cetinkaya S. Integrated warehouse location and inventory decisions in a three-tier distribution system. *IIE Trans* 2008;40:718–732.
2. Yano CA, Gilbert SM. Coordinated pricing and production/procurement decisions: a review. In: Chakravarty A, Eliashberg J, editors. *Managing business interfaces: marketing, engineering, and manufacturing perspectives*. Boston (MA): Kluwer Academic Publishers; 2004.
3. Sarmiento AM, Nagi R. A review of integrated analysis of production–distribution systems. *IIE Trans* 1999;31:1061–1074.
4. Bilgen B, Ozkarahan I. Strategic tactical and operational production–distribution models: a review. *Int J Technol Manag* 2004;28:151–171.
5. Chen Z-L. Integrated production and distribution operations: taxonomy, models, and review. In: Simchi-Levi D, Wu SD, Shen Z-J, editors. *Handbook of quantitative supply chain analysis: modeling in the e-business era*. Boston (MA): Kluwer Academic Publishers; 2004.
6. Pinedo M. *Scheduling theory, algorithms, and systems*. 2nd ed. Upper Saddle River (NJ): Prentice Hall; 2002.
7. Ball MO, Magnanti TL, Monma CL, *et al.* *Network routing*. Volume 8, *Handbooks in operations research and management science*. Amsterdam: North-Holland; 1995.
8. Chen Z-L. Integrated production and outbound distribution scheduling: review and extensions. *Oper Res* 2010;58:130–148.
9. Li C-L, Vairaktarakis G, Lee C-Y. Machine scheduling with deliveries to multiple customer locations. *Eur J Oper Res* 2005; 164:39–51.
10. Steckel KE, Zhao X. Production and transportation integration for a make-to-order manufacturing company with a commit-to-delivery business mode. *Manuf Serv Oper Manag* 2007;9:206–224.
11. Chen Z-L, Vairaktarakis GL. Integrated scheduling of production and distribution operations. *Manage Sci* 2005;51:614–628.
12. Chen Z-L, Pundoor G. Order assignment and scheduling in a supply chain. *Oper Res* 2006;54:555–572.
13. Pundoor G, Chen Z-L. Scheduling a production–distribution system to optimize the trade-off between delivery tardiness and total distribution cost. *Nav Res Log* 2005; 52:571–589.
14. Garcia JM, Lozano S. Production and vehicle scheduling for ready-mix operations. *Comput Ind Eng* 2005;46:803–816.
15. Geismar HN, Laporte G, Lei L, *et al.* The integrated production and transportation scheduling problem for a product with a short life span and non-instantaneous transportation time. *INFORMS J Comput* 2008;20: 21–33.
16. Van Buer MG, Woodruff DL, Olson RT. Solving the medium newspaper production/distribution problem. *Eur J Oper Res* 1999;115:237–253.
17. Wang Q, Batta R, Szczerba RJ. Sequencing the processing of incoming mail to match an outbound truck delivery schedule. *Comput Oper Res* 2005;32:1777–1791.
18. Potts CN. Analysis of a heuristic for one machine sequencing with release dates and delivery times. *Oper Res* 1980;28:1436–1441.
19. Cheng TCE, Kahlbacher HG. Scheduling with delivery and earliness penalties. *Asia Pac J Oper Res* 1993;10:145–152.
20. Lee C-Y, Chen Z-L. Machine scheduling with transportation considerations. *J Schedul* 2001;4:3–24.
21. Woeginger GJ. Heuristics for parallel machine scheduling with delivery times. *Acta Inform* 1994;31:503–512.
22. Kaminsky P. The effectiveness of the longest delivery time rule for the flow shop delivery time problem. *Nav Res Log* 2003;50:257–272.
23. Mastrolilli M. Efficient approximation schemes for scheduling problems with release dates and delivery times. *J Schedul* 2003; 6:521–531.
24. Wang H, Lee C-Y. Production and transport logistics scheduling with two transport mode choices. *Nav Res Log* 2005;52:796–809.

25. Hall NG, Potts CN. Supply chain scheduling: batching and delivery. *Oper Res* 2003; 51:566–584.
26. Hall NG, Potts CN. The coordination of scheduling and batch deliveries. *Ann Oper Res* 2005;135:41–64.
27. Soukhal A, Oulamara A, Martineau P. Complexity of flow shop scheduling problems with transportation constraints. *Eur J Oper Res* 2005;161:32–41.
28. Wang X, Cheng TCE. Machine scheduling with an availability constraint and job delivery coordination. *Nav Res Log* 2007;54: 11–20.
29. Chen Z-L, Pundoor G. Integrating order scheduling and packing. *Prod Oper Manage* 2009;18:672–692.
30. Chang Y-C, Lee C-Y. Machine scheduling with job delivery coordination. *Eur J Oper Res* 2004;158:470–487.
31. Zhong W, Dosa G, Tan Z. On the machine scheduling problem with job delivery coordination. *Eur J Oper Res* 2007;182: 1057–1072.
32. Li C.-L., Vairaktarakis G. Coordinating production and distribution of jobs with bundling operations. *IIE Transactions* 2007;39:203–215.
33. Chen B, Lee C-Y. Logistics scheduling with batching and transportation. *Eur J Oper Res* 2008;189:871–876.
34. Armstrong R, Gao S, Lei L. A zero-inventory production and distribution problem with a fixed customer sequence. *Ann Oper Res* 2008;159:395–414.
35. Li KP, Ganesan VK, Sivakumar AI. Scheduling of single stage assembly with air transportation in a consumer electronic supply chain. *Comput Ind Eng* 2006;51: 264–278.
36. Coffman EG Jr, Garey MR, Johnson DS. Approximation algorithms for bin packing: a survey. In: Hochbaum DS. editor. *Approximation algorithms for NP-hard problems*. Boston (MA): PWS Publishing; 1997. pp. 46–93.

COST-EFFECTIVENESS ANALYSIS, HEALTH-CARE POLICY, AND OPERATIONS RESEARCH MODELS

GREGORY S. ZARIC
Ivey School of Business,
University of
Western Ontario, London,
Canada

INTRODUCTION

Cost-effectiveness analysis (CEA) provides a formal process to evaluate the trade-offs between the incremental costs and the incremental benefits associated with adopting a new medical technology. Throughout this article, we shall use the terms “medical technology” and “intervention” to refer to a drug, device, policy, procedure, or anything else that can improve the health of individuals or populations. Starting in the early 1990s with Ontario and Australia, many public sector drug plans have formally required that CEA be used as part of their decision-making process regarding funding and reimbursement decisions for new drugs. This requirement has become common enough that it is now referred to as *the fourth hurdle* [1] (the other three hurdles for drug approval are safety, efficacy, and manufacturing quality).

CEA often involves complex mathematical modeling. This includes decision-analytic models, Markov models, Monte Carlo simulation models, and dynamic compartmental models. In addition, incorporating this information into health policy decisions involves complex organizational processes and structures. The purpose of this article is to provide an overview of CEA and its role in health policy. In the next section, we shall briefly introduce CEA. In the section titled “Models Used in Cost-Effectiveness Analysis” we will describe some of the models commonly used in CEA. In the section titled “Role in Public Policy Decisions”, we shall give a brief

overview of how CEA is used in health-care policy, focusing on Ontario, the United Kingdom, and the United States.

COST-EFFECTIVENESS ANALYSIS

CEA typically involves the construction and interpretation of the incremental cost-effectiveness ratio (ICER). The ICER for a new medical technology is defined as follows:

$$\text{ICER} = \frac{\text{Cost}_{\text{New}} - \text{Cost}_{\text{Old}}}{\text{Health}_{\text{New}} - \text{Health}_{\text{Old}}} = \frac{\Delta C}{\Delta H}.$$

In the equation above, Cost_{New} and Cost_{Old} are the average future lifetime costs for a cohort receiving the new and old technologies, respectively; $\text{Health}_{\text{New}}$ and $\text{Health}_{\text{Old}}$ are the average future lifetime health experienced by a cohort receiving the new and old technologies, respectively. Thus, the ICER gives the incremental cost (ΔC) associated with switching to the new technology per unit of incremental health benefit (ΔH) associated with switching to the new technology. The definition of “Old” is not uniformly agreed upon, with some arguing that it should be the best possible standard of care (not including the new technology) and others arguing that it should reflect a mix of all treatment options currently used. It is, however, generally agreed that “Old” should not mean “no treatment” unless there really is no other alternative since this comparison could produce overly optimistic ICER estimates. This technique can be easily extended to comparisons among multiple interventions. In that case, the ICER is calculated on the basis of a comparison of each intervention relative to the next best alternative.

In some instances, health is measured in disease-specific units (e.g., change in viral load, symptom-free days gained, and change in a disease severity index). However, the preferred measure of benefit in CEA is quality-adjusted life years (QALYs) of survival gained [2]. QALYs are life years that have been scaled between zero (representing

death) and one (representing perfect health) according to the utility associated with the health state. QALYs reflect the notion that years of life lived in some health states are less desirable than years of life in other states. For example, one estimate places the quality of life adjustment for asymptomatic HIV at 0.80 and for major AIDS-defining illness at 0.42 [3]. If quality of life adjustments are not available, which is often the case for conditions involving pediatrics or mental health, then life years of survival are the preferred alternative.

There are three main types of tools for estimating the quality of life associated with a health state: utility-based measures, such as the standard gamble and time trade-off; rating scales, such as the visual analog scale; and multiattribute scores, such as the Health Utilities Index (HUI), EuroQoL, SF-36 and SF-12, as well as numerous disease-specific quality of life measures. The utility-based measures are generally preferred on theoretical grounds, but the multiattribute instruments, which are short survey questionnaires, are much easier to use in research studies.

In the equation for the ICER, the “Cost” and “Health” terms are defined as future lifetime costs and health, respectively. These terms are calculated as the integral of the time spent in each health state until death, weighted either by the cost or the quality of life of each health state for “Costs” and “Health,” respectively. It is common practice to discount future costs and health, with discount rates of 3–5% being suggested by some cost-effectiveness guidelines [4].

If health outcomes are measured in life years or QALYs, then the ICERs for interventions that treat different diseases are expressed in common units like “cost per life-year gained” or “cost per QALY gained.” This allows one to compare the ICERs for interventions that treat different conditions. If health is measured in disease-specific outcomes, then it may be impossible to compare across different conditions. Finally, a comment on terminology: the term “incremental cost utility ratio” (ICUR) is occasionally used to distinguish an ICER in which QALYs are

the measure of health outcome from one in which some other measure is used.

Interpreting Cost-Effectiveness Ratios

Stating that something is or is not cost effective involves making a value judgment about whether the costs of a new technology are justified by the benefits. Some insight about ICER values can be gained by plotting them in the cost-effectiveness (C-E) plane (Fig. 1). In the C-E plane, incremental costs are plotted along the vertical axis and incremental health benefits are plotted along the horizontal axis (some authors present the C-E plane with the axes reversed). A common interpretation is that technologies in the lower right quadrant are cost saving and should always be adopted. They produce positive health benefits at negative costs. Technologies in the upper left quadrant are more expensive and less effective than the existing technology. These are said to be “dominated” and should never be adopted. Technologies in the upper right quadrant are more expensive and more effective than the existing technology and will have positive ICER values. Determining whether they are cost effective involves a value judgment. Technologies in the bottom left quadrant also have positive ICER values and involve value judgments. However, adopting them involves a difficult policy decision because it would mean switching to something that is both less expensive and less effective than the current standard of care.

Two methods that can assist with making the value judgments required with positive ICER values are comparing them versus a fixed willingness-to-pay (WTP) threshold and listing them in a league table. In the United States, a fixed WTP threshold of \$50,000/QALY gained is often used or implied [5]. The \$50,000/QALY threshold is thought to be based on the cost effectiveness of renal dialysis for end stage liver disease, as assessed in the 1970s [6]. There is currently a spirited debate about whether this figure is appropriate, with some arguing that it may be too high [7], others arguing that it is too low and should be inflation-adjusted to reflect current prices [8], and others arguing that threshold interpretations should not be used at all [9].

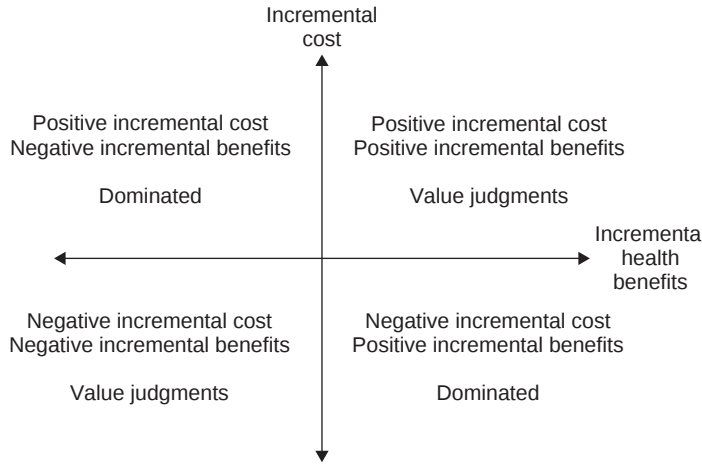


Figure 1. The cost-effectiveness plane.

In Canada, a range WTP threshold values has been suggested, with values less than \$20,000/QALY gained being considered to have strong evidence for adoption, those above \$100,000/QALY gained having weak evidence for adoption, and those in between requiring judgment calls [10]. A review of UK National Health Service (NHS) decisions found an implied threshold of £30,000/QALY [11], a value that has also been reported in the British press [12]. For developing countries, a benchmark of three times GDP per capita has been suggested as a definition of “cost effective” [13].

The second method of interpreting ICERS involves a comparison of ICER values in a league table [e.g., Ref. 14; a sample league table is shown in Table 1]. This method may help with decision making by providing an inferred WTP for other similar technologies.

Justification for Cost-Effectiveness Analysis

The use of CEA in policy decisions is often justified by considering marginal expenditures. A policy maker with only \$1 left in her budget may be confronted with a “menu” of n interventions having ICER values $r_1, r_2, \dots, r_n$. Inverting the ratios gives the incremental health benefits purchased per dollar spent. It is clear that the policy maker would purchase the most health-care benefits for that \$1 (the most “bang for the buck”) by investing in the program with the smallest ICER. (It is also clear that this comparison is impossible

to make if the n different programs use n different measures of health benefits.)

An alternate justification of using CEA is based on its relationship to an optimization problem. Suppose that a planner with a budget B can invest in n different health-care programs. Let x_i be the number of units of program i , let h_i be the incremental health benefit per unit of program i purchased, and let c_i be the cost per unit of program i , $i = 1, \dots, n$. Let U_i , $i = 1, \dots, n$, be an upper limit on the number of units of program i that can be purchased. We index the programs so that the ratios c_i/h_i (the cost-effectiveness ratios of each program) are increasing in i . Assuming that the costs and benefits scale linearly and that the programs do not interact with each other, the investment problem can be written as the following linear program (LP):

$$\begin{aligned} \max \quad & \sum_{i=1}^n h_i x_i \\ \text{s.t.} \quad & \sum_{i=1}^n c_i x_i \leq B, \\ & 0 \leq x_i \leq U_i, \quad i = 1, \dots, n. \end{aligned}$$

This is a “knapsack” LP and it is straightforward to show that the optimal solution is found by investing as much as possible in the set of programs with the lowest ICERs; investing an intermediate amount in one program; and investing nothing in the programs

Table 1. League Table with Some Recent Cost-Effectiveness Estimates

Intervention	Country	ICER Estimate	Source
Letrozole versus tamoxifen as initial adjuvant therapy in postmenopausal women	Canada	\$23,662/QALY gained	15
Colorectal cancer screening in renal transplant recipients	USA	\$22,309 per life-year gained	16
Influenza vaccination in working-age cancer patients	USA	\$224/QALY gained	17
Open versus laparoscopic living-donor nephrectomy	USA	\$1,693,733 per QALY	18
Prevention strategies for gynecologic cancers in Lynch syndrome	USA	\$194,650/QALY gained versus prophylactic surgery at age 40	19
Hepatitis B vaccination in low income countries	The Gambia	US\$28 per disability-adjusted life-year (DALY) averted	20
Cotrimoxazole prophylaxis in HIV-infected children	Zambia	US\$72/LY gained US\$94/QALY gained US\$53/DALY averted	21

with the highest ICERs. More formally, define an index k such that $\sum_{i=1}^{k-1} U_i < B$ and $\sum_{i=1}^k U_i \geq B$. Then $x_i = U_i$, $i = 1, \dots, k-1$; $x_k = B - \sum_{i=1}^{k-1} U_i$; and $x_i = 0$, $i = k+1, \dots, n$. When written this way, the value c_k/h_k can be interpreted as the threshold WTP. Several authors have discussed similar mathematical programming formulations [22].

Note that the LP formulation described above represents a simplified version of the problem in which none of the interventions are dominated. A more general formulation in which there are dominated interventions that must be removed from consideration before solving has been presented elsewhere [23]. The authors show that selecting interventions in increasing order of cost-effectiveness ratios remains optimal in this more general setting.

MODELS USED IN COST-EFFECTIVENESS ANALYSIS

Cost-effectiveness estimates are commonly derived from two sources: clinical trials

and mathematical models. To obtain cost-effectiveness estimates from clinical trials, patients are randomized to different trial groups as in a normal clinical trial. Costs would then be tracked along with health outcomes, and this data used in constructing cost-effectiveness estimates (see Ref. 24 for details).

Although the method just described can be used to generate ICER estimates from trials, many (perhaps most) ICER estimates are derived from mathematical models. Models are used for several reasons: trials typically take place over a short time horizon, like weeks or months, whereas a cost-effectiveness estimate requires an estimate of the impact of the intervention on lifetime health and health-care costs; trials typically evaluate the impact of a technology on a single parameter, whereas a cost-effectiveness estimate may require understanding the impact of changes in multiple parameters, with data derived from several trials; it is uncommon to gather cost information as part of RCTs; some CEAs involve comparing multiple different interventions which, by definition, would

require synthesis of evidence from multiple trials; and trial data is not always available. In the remainder of this section, we present several types of models commonly used in CEA.

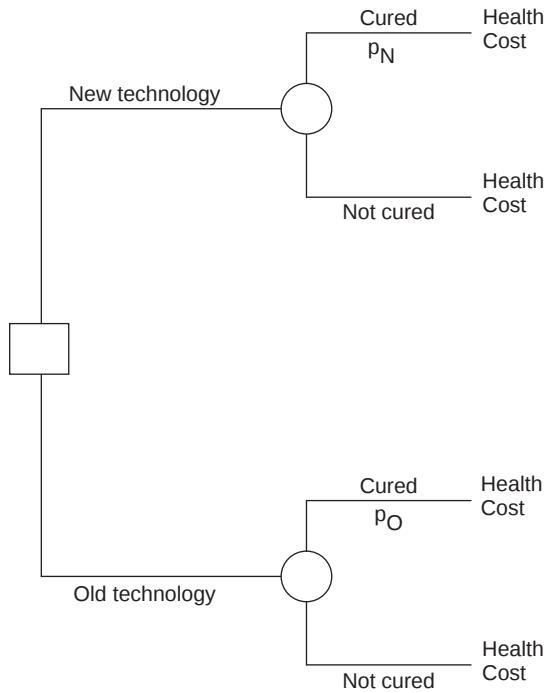
Decision Trees

Decision-tree models are used to illustrate a sequence of uncertain events, including survival, infection status, time of detection, adverse events, and anything else relevant in the calculation of the ICER, following a decision about whether to use a new technology. Figure 2 shows a sample tree for a simple case of two outcomes, “cured” and “not cured.” In this example, the differences between the technologies are the probability of the patient being cured and the cost of the technology itself. More realistic models would likely indicate differences in many other factors resulting in a more detailed

tree structure. Tree models are often used to evaluate technologies that involve a short, one-time application of the technology. This includes screening programs, such as screening newborns for genetic diseases [25]; diagnostic strategies [26]; and treatment of short-term acute infections [27].

Markov Models

Markov models are commonly used to model disease progression for long-term chronic conditions in which patients progress through a number of different health states. Figure 3 depicts a representative model. These models have been used to analyze hepatitis [28], diabetes [29], rheumatoid arthritis [30], and many other diseases. These models assume the Markovian property that the probability of transitioning to a new health state depends only on the current state and not the entire history. If the Markovian property is



$$\text{ICER} = \frac{E[\text{Cost}_{\text{New}}] - E[\text{Cost}_{\text{Old}}]}{E[\text{Health}_{\text{New}}] - E[\text{Health}_{\text{Old}}]}$$

Figure 2. A representative decision tree.

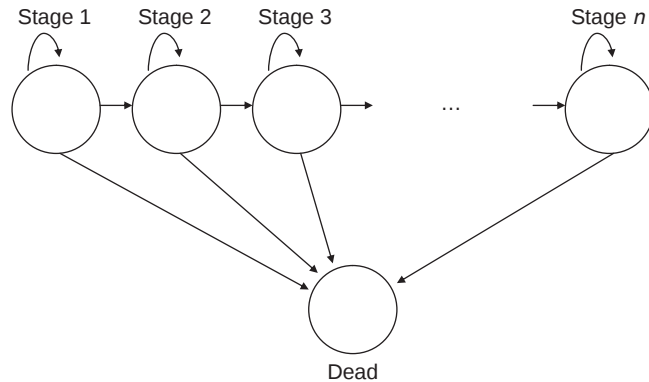


Figure 3. A representative Markov model.

not satisfied, then there are several modeling tricks that can be used to create a Markov model. However, the resulting model may contain a very large number of states.

Markov models often include “dead” as an absorbing state. The other states can be based on symptoms, disease status (e.g., tumor stage in a cancer model, viral load in an HIV model), or other factors. For long time horizons, the probability transition matrix is likely nonhomogeneous because death rates vary with age. The transition probabilities at each time step can be estimated from life tables and adjusted using odds ratios and relative risks derived from clinical trials.

It is common to use Markov models to estimate the ICER for a cohort of individuals rather than to simulate the experience of large numbers of individuals. Under this approach, a cohort will be initially distributed across model compartments (with a special case being all individuals initially in the same compartment). The cohort will be followed either for a long period of time (like 100 years) or until all members of the cohort are dead. When used in this way, Markov models are deterministic: the same cohort will always produce the same expected cost and the same expected health.

Markov models are often used in conjunction with decision-tree models. In this approach, a tree will be used to model a short-term sequence of events and Markov models are used to estimate the expected health and cost components at the ends of the tree [31].

Markov Decision Process Models

In Markov decision processes, a patient’s health status is modeled as a Markov process. A key difference compared to Markov models is that, at each time step, patients make an optimal decision about their treatment which can have the effect of moving them to a different health state, whereas in Markov models patients transition between health states based only on the state transition matrix. These models have been used to model optimal decisions regarding statin therapy treatment strategies [32], HIV treatment [33], and acceptance or rejection decisions for patients waiting for liver transplants [34].

Monte Carlo Simulation Models

If the Markovian property is not satisfied or if the level of complexity exceeds that which can be easily incorporated into a Markov model, then Monte Carlo simulation models are often used. For example, it may be important to model drug resistance in a model of HIV because drug resistance may have an impact on treatment success and hence disease progression. Since drug resistance is based on treatment history, the Markovian property would be violated.

There are two general types of Monte Carlo simulation models: first order and second order. First order models, also known as *microsimulation models*, simulate the lives of individual patients and therefore capture variability between patients. Second order models capture parameter uncertainty

and are typically used to conduct stochastic sensitivity analysis (described in more detail below).

Two well-known Monte Carlo models are the cost effectiveness of preventing AIDS complications (CEPAC) model [35] and the Archimedes model [36,37]. The CEPAC model is a detailed model of HIV progression which simulates at the level of individual patients. For each patient, the model tracks CD4 count, viral load, and history of opportunistic infections, and uses this information to predict disease progression and the occurrence of opportunistic infections. The model is commonly used to estimate the cost effectiveness of HIV treatment options. The Archimedes model simulates organs (e.g., heart, lungs, and kidneys) and biological processes, and then uses this simulated information to simulate the lives of individual patients. The Archimedes model could be used to estimate the cost effectiveness of a new drug by incorporating information about the impact of that drug on various metabolic processes.

As can be imagined, Monte Carlo models are computationally intensive and variability between patients may be high. In the case of the CEPAC model, it is common to simulate 2–5 million patients per scenario [38]. In the case of the Archimedes model, simulation takes place over a distributed network [39].

Compartmental Infectious Disease Models

A special type of compartmental model is often used to estimate the cost effectiveness of technologies where a significant benefit is the prevention of a communicable disease. In these models, the population is

segmented into a number of discrete groups (compartments), which are typically defined by health or behavior. One set of health states represents uninfected individuals and another set represents infected individuals.

A distinguishing feature of this type of model is that the rate of acquiring new infections is proportional to the product of the number of infected individuals and the number of susceptible individuals. This corresponds to a random mixing assumption. If S and I represent the total numbers of susceptible and infected individuals, respectively, then for each susceptible individual the probability that a new encounter, selected randomly, is with an infected individual is $I/(S + I)$. Encounters between susceptible and infected individuals are the only types of encounters that can result in infection transmission. The total number of this type of encounter is $S \times I/(S + I)$. Since the number of individuals in each compartment is constantly changing (i.e., S and I are functions of time), the rate of acquiring new infections is also constantly changing. In models with multiple risk groups, it is possible to incorporate preferential mixing [40,41]. However, within a given type of encounter (e.g., injection drug users with heterosexual nondrug users) the mixing is still random.

Figure 4 shows a representative four-compartment model where the population is divided into groups based on risk behavior (high risk vs low risk) and infection (infected vs not infected). The movement of the population through the set of compartments is indicated by the arrows in Fig. 4. This movement is formally described by a system of

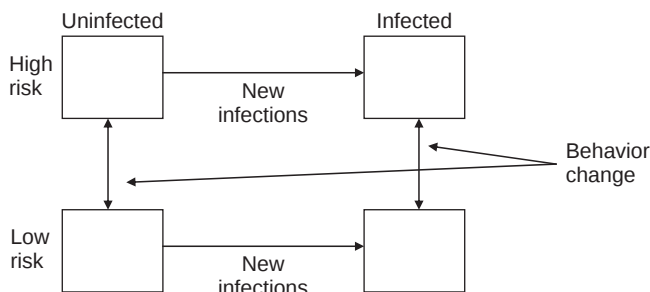


Figure 4. A representative compartmental model for simulating a contagious disease.

nonlinear differential equations constructed using standard principles [42].

These models have been used to evaluate the cost effectiveness of a number of interventions related to HIV including expanding methadone treatment [40] and expanding access to antiretroviral treatment [43]. They have also been used to evaluate vaccination strategies for numerous diseases including those for hepatitis A [44], HPV [45], and varicella [46].

Sensitivity Analysis

CEA models are subjected to extensive sensitivity analysis including one-way, multiway and probabilistic analyses. In one-way and multiway sensitivity analyses, one or more parameters are varied within reasonable ranges and the ICER value is reported as a function of these parameters. A variant of this approach is to report whether the ICER exceeds a WTP threshold for any value with the range. There are many ways to define the range for this type of sensitivity analysis: if a single study or trial is used to estimate a parameter value, then the confidence interval associated with that trial could be used; if multiple trials are considered, then the range of estimates across all trials might be used; if a formal meta-analysis of several trials has been conducted, then the confidence interval associated with that analysis could be used; and in the absence of good trial data, expert opinion could be used.

Many CEA models contain a large number of parameters, making it difficult to describe the simultaneous sensitivity of a model to all parameters in a multiway analysis. Probabilistic sensitivity analysis (also called second order Monte Carlo simulation) has emerged as a way to handle this situation. In probabilistic sensitivity analysis, a prior distribution is defined for each model parameter. The prior distributions may be informed using the same techniques used to determine reasonable ranges for one-way sensitivity analysis. Probabilistic sensitivity analysis is carried out by treating each model parameter as a random variable, drawing a complete set of parameter values, calculating the incremental cost and the incremental effectiveness given the set of parameter values,

and repeating a large number of times. Outcomes for each set of parameter values can be plotted on the C-E plane to give a sense of dispersion in C-E space. This type of sensitivity analysis is frequently used with decision-tree models, Markov models, and compartmental models, but not with Monte Carlo models because of computational intensity.

The cost-effectiveness acceptability curve (CEAC) can be used to interpret the results of probabilistic sensitivity analysis. A representative diagram is shown in Fig. 5. Figure 5a shows sample output plotted in the C-E plane, along with three threshold WTP values, and Fig. 5b shows the CEAC generated from this output. The CEAC is a graph with the WTP threshold along the x -axis and the probability that the intervention is cost effective along the y -axis. It thus shows, for each possible value of the societal WTP threshold, the probability that the intervention should be adopted for funding. Note that the CEAC need not converge to zero (for small WTP) or one (for high WTP) if some of the simulated results fall outside of the northeast quadrant of the C-E plane, as is shown in Fig. 5 and noted elsewhere [47]. From the perspective of the analyst, an advantage of this method is that it does not require any assumptions to be made about a specific WTP value. However, the policy maker who interprets the analysis is not spared from this requirement.

A relatively new tool for sensitivity analysis is value of information analysis [48]. Value of information analysis is a framework similar to the concept of “expected value of perfect information.” In this approach, two ICER values are compared: one derived from a probabilistic sensitivity analysis of the model, and a second, also from a probabilistic sensitivity analysis, but assuming that one or more parameters describing the effectiveness of the new technology are known with certainty. The comparison highlights for decision makers whether they should expect there to be any benefit from additional research [49–51]. If there is no benefit to additional research, then a clear recommendation can be made even in the presence of significant uncertainty.

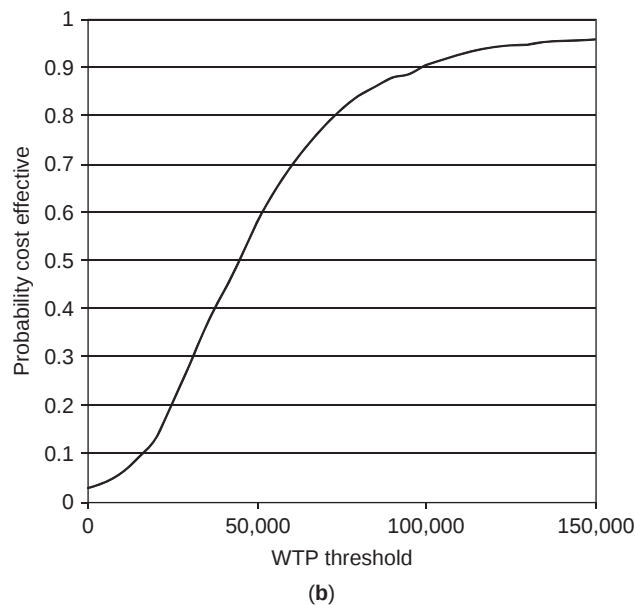
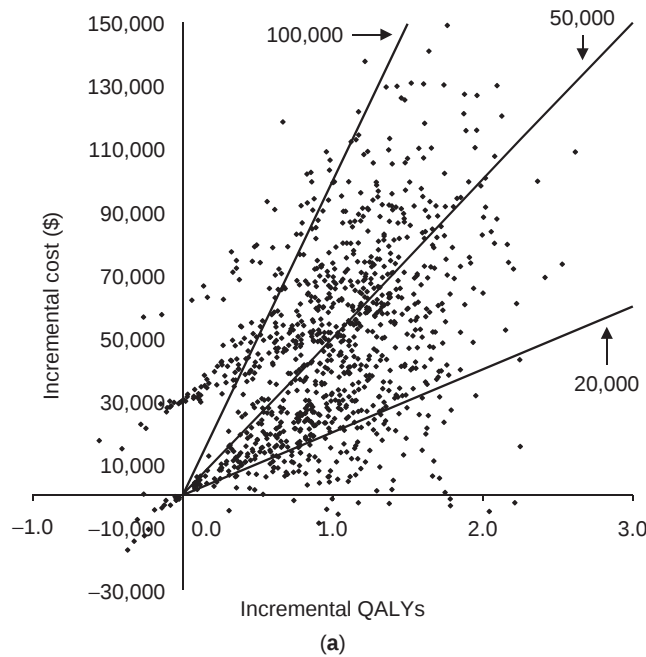


Figure 5. Probabilistic sensitivity analysis and a cost-effectiveness acceptability curve. (a) Sample output from a probabilistic sensitivity analysis showing results plotted in the upper right quadrant of the C-E plane. The three lines correspond to WTP thresholds of \$20,000, \$50,000, and \$100,000. (b) The resulting cost-effectiveness acceptability curve.

There are some difficulties in applying probabilistic sensitivity analysis, and consequently, in conducting value of information analysis. One problem is choosing an appropriate prior probability distribution. This can be particularly difficult if the parameter

estimates were obtained from secondary sources which only presented summary statistics such as a mean and standard deviation. A second problem is that correlations between model parameters are almost never reported. Thus, many analysts assume a

joint distribution in which all parameters are independent, which is unlikely to be true.

ROLE IN PUBLIC POLICY DECISIONS

In many countries, CEA has been formally integrated into health-care policy decision making, and, in particular, in pharmaceutical policy making. A common application is in assisting with formulary decision making. A formulary is a list of drugs that will be reimbursed by an insurance provider. Formulary listing is very important for drug manufacturers because, if a drug is not listed, then patients have a financial incentive to substitute for a similar product that is listed and hence reimbursed by their drug plan. For instance, there are many statins, ACE inhibitors, and migraine medications (to name just a few examples) on the market but not all are approved by all insurance plans. In addition to listing the drugs, a formulary may contain “Limited Use” conditions which are conditions under which each drug can be used. Common types of limited use conditions include restriction to a single use for a product that has many potential uses and restriction of use to cases where a patient has previously experienced failure or toxicity with another drug.

Ontario and Australia are generally recognized as the first jurisdictions to have formally incorporated CEA into their processes for approving new drugs [1,52]. Guidelines for evaluating pharmaceutical products were issued in Ontario in 1991 and in Australia in 1993 [52]. In the remainder of this section, we give a brief overview of the integration of CEA with pharmaceutical policy in Ontario, the United Kingdom, and the United States.

Ontario, Canada

In Ontario, the Ontario Drug Benefits (ODB) Plan is a public drug insurance plan for prescription drugs for people over the age of 65 as well as for low income and special needs populations. Prescription drugs for people not in these groups are paid for privately, either as an out-of-pocket expense or from

a private drug plan. In addition to ODB, Cancer Care Ontario maintains a formulary for cancer drugs and hospitals maintain their own formularies for drugs used by inpatients. Private insurance plans also maintain their own formularies.

A simplified description of the process for a drug to get listed on the formulary in Ontario is as follows:

1. The drug must receive regulatory approval from Health Canada, which evaluates safety and efficacy.
2. The manufacturer must receive pricing approval from the Patented Medicine Prices Review Board (PMPRB) which reviews the “factory gate price” (the price charged by the manufacturer to pharmacies, hospitals, and wholesalers) “to ensure that they are not excessive.”
Following step 2, drugs can be marketed and sold in Canada. However, this does not guarantee that public or private insurance plans will reimburse their use.
For a drug to be added to the ODB formulary there are four additional steps:
3. The drug is reviewed by the Common Drug Review (CDR). CDR looks for evidence of effectiveness (performance in real-world settings as opposed to clinical trial conditions) and cost effectiveness relative to current accepted treatments.
4. The Canadian Expert Drug Advisory Committee (CEDAC) considers the CDR review and makes a recommendation for the provincial formularies in one of three categories: fund, do not fund, or fund with conditions.
5. The CDR review and CEDAC recommendations are sent to the Ontario’s Ministry of Health and Long-Term Care Committee to Evaluate Drugs (CED). In addition to evidence of effectiveness and cost effectiveness, CED requires an estimate of budget impact for the formulary if the drug is approved.
6. The CED makes a recommendation to the executive director, who then makes

a decision about adding the drug to the ODB formulary. The executive director may engage in negotiations with the drug manufacturer regarding price or risk-sharing agreements before adding the drug to the formulary.

The process outlined above is used for new drugs as well as for new uses of existing drugs. Cancer drugs follow a similar process with some exceptions. There is no fixed or generally adhered to process for hospital formulary decision making, and the processes range from ad hoc to very formal depending on the institution.

How are cost-effectiveness models incorporated into this process? The drug manufacturer submits a CEA in the form of a published article from a peer-reviewed journal or an unpublished report. The CEA is formally reviewed by CDR in step 3 and by CED in step 5. CDR requires that the manufacturer submit an electronic copy of the model (e.g., the spreadsheet model or a model prepared in a speciality package like TreeAge) to allow the reviewers to conduct their own sensitivity analysis or change base case parameter estimates if necessary.

What information do the decision makers and reviewers look for in the models? In Ontario, there is a template to assist in the review of the economic models; versions of this template have been published in academic journals [53]. Much of the template addresses important modeling issues, such as establishing model validity, conducting sensitivity analysis, and conducting subgroup analysis, as well as questions on the importance of quality of life assumptions. In addition to technical model questions listed above, the template asks “Does the cost of the therapy justify the clinical and/or quality of life outcomes relative to comparators?” This question forces the reviewers to make a value judgment (in effect, a policy recommendation) based on the model results. One can argue at length about the pros and cons of league tables, fixed WTP thresholds and other decision-making heuristics. However, each newly approved drug represents a policy decision that must be made.

Is there evidence that the models are used in decision making? CEDAC publishes short memos on-line outlining their recommendations [54]. Recommendations for or against funding can be made for a number of reasons, but considerations related to CEA are often cited. For example, for varenicline (Champix/Chantix), the CEDAC recommendation in favor of adoption included the following statement about cost effectiveness: “While this evaluation was based on a number of assumptions, the Committee felt that varenicline was cost effective...” [55]. In the case of zoledronic acid (Alclasta and Zometa), CEDAC recommended against formulary listing. In their recommendation, they stated “...zoledronic acid was not found to be cost effective when compared to oral bisphosphonate agents.” [56]

The United Kingdom

In the United Kingdom, the National Institute for Clinical Excellence (NICE) was established in 1999 to provide guidance to the NHS regarding medical technologies. As with the evaluation process in Canada, recommendations by NICE often include CEA evidence. Readers seeking a detailed description of health technology assessment in the United Kingdom are referred elsewhere [57,58]. Here, we highlight two key differences between the process used by NICE and the process described above for drugs in Canada.

- The formulary in Canada is known as a *positive list* in which drugs must be approved in order to be listed. The United Kingdom maintains a “negative list” which lists drugs that cannot be used. If a drug has not been reviewed by NICE, then there are no restrictions on its use and funding decisions are made by the regional primary care trusts.
- In Canada, the CEA models are supplied by the drug manufacturers. The reviewers assess the evidence provided by the manufacturers. In the United Kingdom, CEAs are commissioned from

an academic medical center. The models are produced independently of the drug manufacturers but may include data provided by the manufacturers such as unpublished data from registries or clinical trials.

Is there evidence that CEA is used as part of the decision-making process? NICE recommendations frequently contain extensive details about the cost-effectiveness modeling that was done to support their recommendation. There has been academic work in the form of case studies to assess the impact of CEA on NICE decisions [59,60]. Additionally, decisions by NICE are often reported in the press. In 2008, the *Times* reported that NICE had recommended against funding four cancer drugs—sunitinib (Sutent), bevacizumab (Avastin), sorafenib (Nexavar), and temsirolimu (Torisel) [12]. The article quotes the public health director of NICE as saying “Although these treatments are clinically effective, regrettably the cost to the NHS is such that they are not a cost-effective use of resources.” [12] However, in February 2009, NICE reversed its recommendation for sunitinib after the manufacturer and the NHS developed a risk-sharing agreement to reduce the cost in patients who do not respond to treatment [61]. In 2007, the BBC reported that “Early in 2007, the drugs watchdog NICE said the drug Alimta was not cost effective.” [62] However, public policy decisions involve more than considerations of technical efficiency, and the same article later states “But groups and sufferers in the North East successfully lobbied to get that decision overturned and the drug is now available.” [62]

The United States

It is easy to find examples of CEA influencing policy in the United Kingdom, Canada, and other places where public insurers have issued guidelines about the use of CEA. Examples in the United States are less obvious, and many have speculated about why this is the case. Luce [63] suggests five categories of reasons: skepticism with methods, potential lack of training by decision makers, legal problems, trust

problems, and political issues related to the perceived relationship between “cost effectiveness” and “rationing.” Grosse *et al.* report that CEAs helped change the recommendations regarding the frequency of cervical cancer screening, although the change in recommendations may not have led to a change in practice [64]. CEA has been cited in guidelines issued by the US Centers for Disease Control and Prevention (CDC). For example, a recent update to the CDC guidelines on HIV screening [65] refers to three cost-effectiveness analyses [38,66,67]. Recent CDC guidelines on hepatitis B identification and management [68] also refer to a CEA [28]. Clinical practice guidelines issued by various medical associations occasionally refer to CEAs. However, these are guidelines and not binding policy mandates.

In 2009, as part of the US Recovery Act, \$1.1 billion was allocated to comparative effectiveness research (CER), to be distributed through three government funding agencies [69]. The interpretation of CER and the allocation of these funds has generated some controversy [70,71].

LIMITATIONS

There are limitations to CEA. The technique may not do a good job at capturing preferences for equity by decision makers. Kaplan and Merson [72] and Zaric and Brandeau [73] have developed models to investigate equity-efficiency trade-offs in the allocation of HIV prevention resources. Some have argued that reliance on CEA might bias funding away from expensive drugs that treat very rare conditions [74]. Such decisions may not reflect societal values. For example, some enzyme replacement therapies cost more than \$300,000 per year, but if they are not funded patients may have few alternatives. The experience in the United Kingdom has illustrated the importance of political considerations in funding decisions. And, as mentioned earlier, some decision makers may be resistant to using CEA, for a variety of reasons [63].

CONCLUSIONS

In this article, we briefly discussed CEA, some of the mathematical models that are used in CEA, and how CEA is used in health-care policy decisions in Canada, the United Kingdom and the United States. Similar processes are used elsewhere, including Australia, New Zealand [57], and much of Europe [75]. It has recently been reported that recommendations made by NICE in the United Kingdom influence decisions in several other countries [76].

The growing acceptance of these models by policy makers combined with the rapid pace of development of new medical technologies will likely create increased demand for this type of research. In addition, advances in computing power and simulation theory may drive the ability to develop increasingly sophisticated large-scale simulation models. For operations researchers, this field represents a great opportunity to become engaged in an important, growing body of research that involves sophisticated models developed by interdisciplinary research teams, and that has the potential for real-world impact.

REFERENCES

1. Taylor RS, Drummond MF, Salkeld G, *et al.* Inclusion of cost effectiveness in licensing requirements of new drugs: the fourth hurdle. *BMJ* 2004;329(7472):972–975.
2. Kaplan RM. Utility assessment for estimating quality-adjusted life years. In: Sloan FA, editor. *Valuing health care: costs, benefits and effectiveness of pharmaceuticals and other medical technologies*. Cambridge: Cambridge University Press; 1995. pp. 31–60.
3. Bayoumi AM, Redelmeier DA. Economic methods for measuring the quality of life associated with HIV infection. *Qual Life Res* 1999;8(6):471–480.
4. Berger ML, Binglefors K, Hedblom EC, *et al.* editors. *Health care cost, quality and outcomes: ISPOR book of terms*. Lawrenceville (NJ): International Society for Pharmacoeconomics and Outcomes Research; 2003.
5. Owens DK. Interpretation of cost-effectiveness analyses. *J Gen Intern Med* 1998;13(10):716–717.
6. Grosse SD. Assessing cost-effectiveness in healthcare: history of the \$50,000 per QALY threshold. *Expert Rev Pharmacoecon Outcomes Res* 2008;8(2):165–178.
7. King JT, Tsevat J, Lave JR, *et al.* Willingness to pay for a quality-adjusted life year: implications for societal health care resource allocation. *Med Decis Making* 2005;25(6):667–677.
8. Ubel PA, Hirth RA, Chernew ME, *et al.* What is the price of life and why doesn't it increase at the rate of inflation? *Arch Intern Med* 2003;163(14):1637–1641.
9. Gafni A, Birch S. Incremental cost-effectiveness ratios (ICERs): the silence of the lambda. *Soc Sci Med* 2006;62(9):2091–2100.
10. Laupacis A, Feeny D, Detsky AS, *et al.* How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *CMAJ* 1992;146(4):473–481.
11. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Econ* 2004;13(5):437–452.
12. Hawkes N. 35,000-a-year kidney cancer drugs too costly for NHS. *TimesOnline*. London; 2008 Aug 7.
13. Sevilla JD, editor. *Macroeconomics and health: investing in health for economic development*. Report of the Commission on macroeconomics and health (CMH) of the World Health Organization. Geneva: World Health Organization; 2001.
14. Tengs TO, Adams ME, Pliskin JS, *et al.* Five-hundred life-saving interventions and their cost-effectiveness. *Risk Anal* 1995;15(3):369–390.
15. Delea TE, El-Ouagari K, Karnon J, *et al.* Cost-effectiveness of letrozole versus tamoxifen as initial adjuvant therapy in postmenopausal women with hormone-receptor positive early breast cancer from a Canadian perspective. *Breast Cancer Res Treat* 2008;108(3):375–387.
16. Wong G, Howard K, Craig JC, *et al.* Cost-effectiveness of colorectal cancer screening in renal transplant recipients. *Transplantation* 2008;85(4):532–541.
17. Avritscher EB, Cooksley CD, Geraci JM, *et al.* Cost-effectiveness of influenza vaccination in working-age cancer patients. *Cancer* 2007;109(11):2357–2364.

18. Hamidi V, Andersen MH, Oyen O, *et al.* Cost effectiveness of open versus laparoscopic living-donor nephrectomy. *Transplantation* 2009;87(6):831–838.
19. Kwon JS, Sun CC, Peterson SK, *et al.* Cost-effectiveness analysis of prevention strategies for gynecologic cancers in Lynch syndrome. *Cancer* 2008;113(2):326–335.
20. Kim SY, Salomon JA, Goldie SJ. Economic evaluation of hepatitis B vaccination in low-income countries: using cost-effectiveness affordability curves. *Bull World Health Organ* 2007;85(11):833–842.
21. Ryan M, Griffin S, Chitah B, *et al.* The cost-effectiveness of cotrimoxazole prophylaxis in HIV-infected children in Zambia. *Aids* 2008;22(6):749–757.
22. Stinnett AA, Paltiel AD. Mathematical programming for the efficient allocation of health care resources. *J Health Econ* 1996;15(5):641–653.
23. Laska EM, Meisner M, Siegel C, *et al.* Ratio-based and net benefit-based approaches to health care resource allocation: proofs of optimality and equivalence. *Health Econ* 1999;8(2):171–174.
24. Ramsey S, Willke R, Briggs A, *et al.* Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. *Value Health* 2005;8(5):521–533.
25. Cipriano LE, Rupar CA, Zaric GS. The cost-effectiveness of expanding newborn screening for up to 21 inherited metabolic disorders using tandem mass spectrometry: results from a decision-analytic model. *Value Health* 2007;10(2):83–97.
26. Shen B, Shermock KM, Fazio VW, *et al.* A cost-effectiveness analysis of diagnostic strategies for symptomatic patients with ileal pouch anal anastomosis. *Am J Gastroenterol* 2003;98(11):2460–2467.
27. Martin M, Quilici S, File T, *et al.* Cost-effectiveness of empirical prescribing of antimicrobials in community-acquired pneumonia in three countries in the presence of resistance. *J Antimicrob Chemother* 2007;59(5):977–989.
28. Hutton DW, Tan D, So SK, *et al.* Cost-effectiveness of screening and vaccinating Asian and Pacific Islander adults for hepatitis B. *Ann Intern Med* 2007;147(7):460–469.
29. Coyle D, Rodby R, Soroka S, *et al.* Cost-effectiveness of irbesartan 300 mg given early versus late in patients with hypertension and a history of type 2 diabetes and renal disease: a Canadian perspective. *Clin Ther* 2007;29(7):1508–1523.
30. Spalding JR, Hay J. Cost effectiveness of tumour necrosis factor-alpha inhibitors as first-line agents in rheumatoid arthritis. *Pharmacoeconomics* 2006;24(12):1221–1232.
31. Kondo M, Hoshi SL, Ishiguro H, *et al.* Economic evaluation of 21-gene reverse transcriptase-polymerase chain reaction assay in lymph-node-negative, estrogen-receptor-positive, early-stage breast cancer in Japan. *Breast Cancer Res Treat* 2008;112(1):175–187.
32. Denton BT, Kurt M, Shah ND, *et al.* Optimizing the start time of statin therapy for patients with diabetes. *Med Decis Making* 2009;29(3):351–367.
33. Shechter SM, Bailey MD, Schaefer AJ, *et al.* The optimal time to initiate HIV therapy under ordered health states. *Oper Res* 2008;56(1):20–33.
34. Sandikci B, Maillart LM, Schaefer AJ, *et al.* Estimating the patient's price of privacy in liver transplantation. *Oper Res* 2008;56(6):1393–1410.
35. Paltiel AD, Scharfstein JA, Seage GR 3rd, *et al.* A Monte Carlo simulation of advanced HIV disease: application to prevention of CMV infection. *Med Decis Making* 1998;18(2 Suppl):S93–S105.
36. Eddy D. Bringing health economic modeling to the 21st century. *Value Health* 2006;9(3):168–178.
37. Schlessinger L, Eddy DM. Archimedes: a new model for simulating health care systems—the mathematical formulation. *J Biomed Inform* 2002;35(1):37–50.
38. Paltiel AD, Weinstein MC, Kimmel AD, *et al.* Expanded screening for HIV in the United States—an analysis of cost-effectiveness. *N Engl J Med* 2005;352(6):586–595.
39. Eddy DM, Schlessinger L, Kahn R. Clinical outcomes and cost-effectiveness of strategies for managing people at high risk for diabetes. *Ann Intern Med* 2005;143(4):251–264.
40. Zaric GS, Barnett PG, Brandeau ML. HIV transmission and the cost-effectiveness of methadone maintenance. *Am J Public Health* 2000;90(7):1100–1111.
41. Bayoumi AM, Zaric GS. The cost-effectiveness of Vancouver's supervised injection facility. *CMAJ* 2008;179(11):1143–1151.

42. Bailey NTJ. The mathematical theory of infectious diseases and its applications. 2nd ed. London: Griffin; 1975. pp. 473.
43. Long EF, Brandeau ML, Galvin CM, *et al.* Effectiveness and cost-effectiveness of strategies to expand antiretroviral therapy in St. Petersburg, Russia. *Aids* 2006;20(17):2207–2215.
44. Bauch CT, Anonychuk AM, Pham BZ, *et al.* Cost-utility of universal hepatitis A vaccination in Canada. *Vaccine* 2007;25(51):8536–8548.
45. Elbasha EH, Dasbach EJ, Insinga RP. Model for assessing human papillomavirus vaccination strategies. *Emerg Infect Dis* 2007;13(1):28–41.
46. Lenne X, Diez Domingo J, Gil A, *et al.* Economic evaluation of varicella vaccination in Spain: results from a dynamic model. *Vaccine* 2006;24(47–48):6980–6989.
47. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves—facts, fallacies and frequently asked questions. *Health Econ* 2004;13(5):405–415.
48. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999;18(3):341–364.
49. Bojke L, Claxton K, Sculpher MJ, *et al.* Identifying research priorities: the value of information associated with repeat screening for age-related macular degeneration. *Med Decis Making* 2008;28(1):33–43.
50. Claxton KP, Sculpher MJ. Using value of information analysis to prioritise health research: some lessons from recent UK experience. *Pharmacoeconomics* 2006;24(11):1055–1068.
51. Claxton K, Cohen JT, Neumann PJ. When is evidence sufficient? *Health Aff (Millwood)* 2005;24(1):93–101.
52. Glennie JL, Torrance GW, Baladi JF, *et al.* The revised Canadian guidelines for the economic evaluation of pharmaceuticals. *Pharmacoeconomics* 1999;15(5):459–468.
53. Detsky AS. Guidelines for economic analysis of pharmaceutical products: a draft document for Ontario and Canada. *Pharmacoeconomics* 1993;3(5):354–361.
54. Canadian Agency for Drugs and Technologies in Health. CADTH: recommendations and status of drug submissions. Available at <http://www.cadth.ca/index.php/en/cdr/recommendations>. Accessed 2008 Oct 8.
55. Canadian Agency for Drugs and Technologies in Health. CEDAC final recommendation and reasons for recommendation. Varenicline tartrate. Available at http://www.cadth.ca/media/cdr/complete/cdr_complete_Champix_August-16-07.pdf. Accessed 2008 Oct 31.
56. Canadian Agency for Drugs and Technologies in Health. CEDAC final recommendation and reasons for recommendation. Zoledronic acid. Available at http://www.cadth.ca/media/cdr/complete/cdr_complete_Aclasta_June-25-2008_e.pdf. Accessed 2008 Oct 31.
57. Morgan SG, McMahon M, Mitton C, *et al.* Centralized drug review processes in Australia, Canada, New Zealand, and the United Kingdom. *Health Aff (Millwood)* 2006;25(2):337–347.
58. Raftery JP. Paying for costly pharmaceuticals: regulation of new drugs in Australia, England and New Zealand. *Med J Aust* 2008;188(1):26–28.
59. Bryan S, Williams I, McIver S. Seeing the NICE side of cost-effectiveness analysis: a qualitative investigation of the use of CEA in NICE technology appraisals. *Health Econ* 2007;16(2):179–193.
60. Williams I, Bryan S, McIver S. How should cost-effectiveness analysis be used in health technology coverage decisions? Evidence from the National Institute for Health and Clinical Excellence approach. *J Health Serv Res Policy* 2007;12(2):73–79.
61. Boseley S. U-turn as Nice approves NHS kidney cancer drug. *The Guardian*. London; 2009 Feb 4.
62. BBC News. BBC—Tees—Places—The deadly legacy of a proud industrial heritage; 2007. Available at http://www.bbc.co.uk/tees/content/articles/2007/10/03/asbestos_feature.shtml. Accessed 2008 Nov 4.
63. Luce BR. What will it take to make cost-effectiveness analysis acceptable in the United States? *Med Care* 2005;43(7 Suppl):44–48.
64. Grosse SD, Teutsch SM, Haddix AC. Lessons from cost-effectiveness research for United States public health policy. *Ann Rev Public Health* 2007;28:365–391.
65. Branson BM, Handsfield HH, Lampe MA, *et al.* Revised recommendations for HIV testing of adults, adolescents, and pregnant women in health-care settings. *MMWR Recomm Rep* 2006;55(RR-14):1–17. quiz CE1-4.
66. Sanders GD, Bayoumi AM, Sundaram V, *et al.* Cost-effectiveness of screening for HIV in the

- era of highly active antiretroviral therapy. *N Engl J Med* 2005;352(6):570–585.
67. Walensky RP, Weinstein MC, Kimmel AD, *et al.* Routine human immunodeficiency virus testing: an economic evaluation of current guidelines. *Am J Med* 2005;118(3):292–300.
 68. Weinbaum CM, Mast EE, Ward JW. Recommendations for identification and public health management of persons with chronic hepatitis B virus infection. *Hepatology* 2009;49(5 Suppl):S35–S44.
 69. U.S. Department of Health and Human Services. Comparative effectiveness research funding. Available at <http://www.hhs.gov/recovery/programs/ceer/index.html>. Accessed 2009 July 8.
 70. Avorn J. Debate about funding comparative-effectiveness research. *N Engl J Med* 2009; 360(19):1927–1929.
 71. Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? *N Engl J Med* 2009; 360(19):1925–1927.
 72. Kaplan EH, Merson MH. Allocating HIV-prevention resources: balancing efficiency and equity. *Am J Public Health* 2002;92(12):1905–1907.
 73. Zaric GS, Brandeau ML. A little planning goes a long way: multi-level allocation of HIV prevention resources. *Med Decis Making* 2007;27(1):71–81.
 74. Drummond M, Evans B, LeLorier J, *et al.* Evidence and values: requirements for public reimbursement of drugs for rare diseases—a case study in oncology. *Can J Clin Pharmacol* 2009;16(2):e273–e281. [Discussion e282–e284].
 75. Tilson L, Barry M. European pharmaceutical pricing and reimbursement strategies. Dublin: National Centre for Pharmacoeconomics; 2005. p. 166.
 76. Harris G. British balance benefit vs. cost of latest drugs. *New York Times*. New York; 2008 Dec 3.

COVER INEQUALITIES

KONSTANTINOS KAPARIS
School of Mathematics, University
of Southampton, Southampton,
UK

ADAM N. LETCHFORD
Department of Management
Science, Lancaster University,
Lancaster, UK

Many problems arising in OR/MS can be formulated as *mixed-integer linear programs* (MILPs): see the article titled **Formulating Good MILP Models** in this encyclopedia. If one wishes to solve a class of MILPs to proven optimality, or near optimality, it is often useful to have a class of *cutting planes* available. Cutting planes are linear inequalities that are satisfied by all feasible solutions to the MILP, but may be violated by solutions to its continuous relaxation (see **Branch and Cut**).

This article is concerned with one particular class of cutting planes—the *cover inequalities*—along with some variants of them. Cover inequalities were originally defined in the context of the *0–1 Knapsack problem* (0–1 KP), but can be applied to any *0–1 linear program* (0–1 LP). Moreover, the idea underlying the cover inequalities has been extended to yield cutting planes for other knapsack-type problems and other kinds of MILPs.

The article is structured as follows. In the section titled “Theory,” we recall the theory underlying the inequalities. In the section titled “Algorithmic Aspects,” we discuss some algorithmic questions that must be addressed if one wishes to actually use cover inequalities in a cutting-plane algorithm. In the section titled “Applications,” we list some of the problems to which cover inequalities have been applied. Finally, at the end of the article, we give some pointers for further reading.

THEORY

Knapsack Polytopes

As mentioned above, the cover inequalities were first presented in the context of the 0–1 KP. The 0–1 KP takes the following form:

$$\max \left\{ p^T x : w^T x \leq c, x \in \{0, 1\}^n \right\}.$$

Here, $p \in \mathbb{Z}_+^n$ is the vector of *profits*, $w \in \mathbb{Z}_+^n$ is the vector of *weights* and $c \in \mathbb{Z}_+$ is the *knapsack capacity*.

The convex hull of feasible solutions to a 0–1 KP, that is, the set

$$\text{conv} \left\{ x \in \{0, 1\}^n : w^T x \leq c \right\},$$

is called a *knapsack polytope*. We will let $KP(w, c)$ denote this polytope. In theory, the strongest possible cutting planes, for a given w and c , are those that define *facets* of $KP(w, c)$ (see **Basic Polyhedral Theory**).

Cover and Extended Cover Inequalities

Now, let N denote $\{1, \dots, n\}$. A set $C \subset N$ is called a *cover* if it satisfies $\sum_{j \in C} w_j > c$. Since it is impossible to set all of the variables in C to 1 simultaneously while maintaining feasibility, the linear inequality $\sum_{j \in C} x_j \leq |C| - 1$ is valid for $KP(w, c)$. This is the *cover inequality* (CI).

The CIs were introduced independently by Balas [1], Hammer *et al.* [2], and Wolsey [3]. These authors observed that the strongest CIs are obtained when the cover C is *minimal*, in the sense that no proper subset of C is also a cover.

A simple way to strengthen the CIs was given by Balas [1]. Compute $w^* := \max_{j \in C} w_j$ and define the *extension* of C as $E(C) := C \cup \{j \in N \setminus C : w_j \geq w^*\}$. The *extended cover inequality* (ECI) $\sum_{j \in E(C)} x_j \leq |C| - 1$ is then easily seen to be valid for $KP(w, c)$.

Example. Suppose that $n = 8$ and that the knapsack constraint takes the form $4x_1 + 6x_2 + 6x_3 + 6x_4 + 6x_5 + 7x_6 + 8x_7 + 15x_8 \leq 22$. The set $\{1, 2, 6, 7\}$ forms a cover, since $4 + 6 + 7 + 8 = 25 > 22$. Moreover, it is a minimal cover. The corresponding CI takes the form $x_1 + x_2 + x_6 + x_7 \leq 3$. Moreover, we have $w^* = 8$ and $E(C) = \{1, 2, 6, 7, 8\}$, and therefore, the corresponding ECI takes the form $x_1 + x_2 + x_6 + x_7 + x_8 \leq 3$.

Although the ECIs dominate the CIs, they are still not guaranteed to define facets of $KP(w, c)$, as explained in the next section.

Lifted Cover Inequalities

Balas [1] and Wolsey [3] showed that, given any minimal cover C , there exists at least one facet-defining *lifted cover inequality* (LCI) of the form

$$\sum_{j \in C} x_j + \sum_{j \in N \setminus C} \alpha_j x_j \leq |C| - 1, \quad (1)$$

where $\alpha_j \geq 0$ for all $j \in N \setminus C$. Moreover, each such LCI dominates the ECI.

Example (continued). From the given cover $\{1, 2, 6, 7\}$, one can derive the following four LCIs:

$$\begin{aligned} x_1 + x_2 + x_3 + x_6 + x_7 + 2x_8 &\leq 3 \\ x_1 + x_2 + x_4 + x_6 + x_7 + 2x_8 &\leq 3 \\ x_1 + x_2 + x_5 + x_6 + x_7 + 2x_8 &\leq 3 \\ x_1 + x_2 + \frac{1}{2}x_3 + \frac{1}{2}x_4 + \frac{1}{2}x_5 + x_6 \\ &\quad + x_7 + 2x_8 \leq 3. \end{aligned}$$

Each of these LCIs clearly dominates the ECI, and all of them can be shown to define facets of $KP(w, c)$.

Note that a facet-defining LCI can have fractional coefficients.

Van Roy and Wolsey [4] noted that one can derive more general LCIs of the form:

$$\begin{aligned} \sum_{j \in C \setminus D} x_j + \sum_{j \in N \setminus C} \alpha_j x_j + \sum_{j \in D} \beta_j x_j &\leq |C \setminus D| \\ &\quad + \sum_{j \in D} \beta_j - 1, \end{aligned} \quad (2)$$

where C is a cover and $D \subset C$. We will call the inequalities (2) “general” LCIs, and we will follow Gu *et al.* [5] in referring to the inequalities (1) as “simple” LCIs.

Example (continued). Assume as before that the knapsack constraint takes the form $4x_1 + 6x_2 + 6x_3 + 6x_4 + 6x_5 + 7x_6 + 8x_7 + 15x_8 \leq 22$. The inequality $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + 2x_7 + 3x_8 \leq 4$ can be shown to define a facet of $KP(w, c)$. It can be derived as a general LCI by setting $C = \{1, 2, 8\}$ and $D = \{8\}$.

We remark that the general LCIs with $|D| = 1$ dominate the so-called “ $(1, k)$ -configuration” inequalities of Padberg [6].

ALGORITHMIC ASPECTS

Computing Lifting Coefficients

A natural question is how to actually compute the coefficients α_j and β_j . The process of computing these coefficients is called *lifting*. One normally distinguishes between *sequential* lifting, in which the coefficients are computed one at a time, and *simultaneous* lifting, in which they are computed “all at once,” so to speak. (See **Lifting Techniques for Mixed Integer Programming** for more details on lifting.)

We will follow Gu *et al.* [5] in referring to the computation of the α_j and β_j as *up-lifting* and *down-lifting*, respectively.

The case that has received most attention is *sequential up-lifting*, which leads to simple LCIs with integral coefficients. Balas [1] and Wolsey [3] both pointed out that different lifting sequences can yield different simple LCIs. (This is how we generate the three simple LCIs with integral coefficients in the example in the section titled “Lifted Cover Inequalities.”) They also showed that, for a given lifting sequence, the up-lifting coefficients can be computed by solving a series of auxiliary 0–1 KPs.

Since solving 0–1 KPs is rather time-consuming, researchers looked for faster and simpler methods. Balas and Zemel [7] showed how to compute, in $O(n + |C| \log |C|)$

time, tight lower and upper bounds on the up-lifting coefficients, which are independent of the lifting sequence. The upper bounds were later improved by Gu *et al.* [8]. Crowder *et al.* [9] proposed to compute *approximate* up-lifting coefficients, by solving the continuous relaxation of the auxiliary 0–1 KPs, and rounding the answers down to integers. Finally, Zemel [10] pointed out that one can compute exact up-lifting coefficients for all variables in $O(n|C|)$ time, by dynamic programming.

The example given in the section titled “Lifted Cover Inequalities” shows that there exist simple LCIs with fractional coefficients, which cannot be obtained with sequential up-lifting. To derive such simple LCIs, one must perform simultaneous up-lifting. Unfortunately, this is not easy. Indeed, Hartvigsen and Zemel [11] showed that even *recognizing* simultaneously-lifted simple LCIs is $\mathcal{NP}$ -hard. On the other hand, there exist a couple of simple and fast procedures for *approximate* simultaneous up-lifting [8,12].

As for general LCIs, Gu *et al.* [5] claim that sequential lifting can be performed in $O(n^3|C|)$ time, provided that the lifting sequence satisfies a certain condition. An $O(nc)$ lifting algorithm, which imposes no conditions on the lifting sequence, is described in our article [13]. A simpler and faster alternative is to solve the continuous relaxation of the auxiliary 0–1 LPs, just as Crowder *et al.* [9] did for simple LCIs.

Separation Algorithms

In order to use a class of valid inequalities as cutting planes, one needs a *separation algorithm*, that is, a routine for generating the inequalities when they are violated (see **Branch and Cut**). Unfortunately, the separation problems for CIs, simple LCIs, and general LCIs have all been shown to be $\mathcal{NP}$ -hard [14–16], and the same seems likely to be true for ECIs. Nevertheless, some useful exact and heuristic separation algorithms are known.

We begin with CIs. Crowder *et al.* [9] noted that CI separation is equivalent to the

following knapsack-type problem:

$$\min \left\{ \sum_{j \in N} (1 - x_j^*) y_j : w^T y > c, y \in \{0, 1\}^n \right\}.$$

Here, y_j is a 0–1 variable, taking the value 1 if and only j is to be inserted into the cover C . It is not hard to show that this problem can be solved by dynamic programming in $O(nc)$ time, which is acceptable if c is not too large. Crowder *et al.* [9] suggested solving it heuristically instead, by inserting items into C in nondecreasing order of $(1 - x_j^*)/w_j$ until a cover is obtained. This heuristic takes $O(n \log n)$ time.

As for ECIs, Gabrel and Minoux [17] showed that the separation problem can be reduced to a sequence of knapsack-type problems. In our article [13], we presented a simplified version of their algorithm, which runs in $O(n^2c)$ time, along with a dynamic programming algorithm that runs in only $O(nc)$ time. We also presented an effective heuristic, that runs in $O(n^2)$ time. All of our algorithms can produce more than one violated ECI per call.

As for simple LCIs, Crowder *et al.* [9] suggested using their heuristic to obtain a CI, and then performing up-lifting to strengthen it. They pointed out that it pays to up-lift the variables with positive x^* -value first, before up-lifting the variables with zero x^* -value.

Several articles address the separation of general LCIs [4,5,13,18]. Van Roy and Wolsey [4] proposed to include only one item into D —namely, the item in C with the largest x^* -value—and leave down-lifting to the end. Hoffman and Padberg [18] proposed to set $D = \{j \in C : x_j^* = 1\}$, and then lift in the following order: up-lift the variables in $N \setminus C$ with positive x^* -value, down-lift the variables in D , then up-lift the variables in $N \setminus C$ with zero x^* -value. Gu *et al.* [5] used the same scheme, but used a different heuristic to construct C : they simply inserted items into C in nonincreasing order of x^* -value, until a cover was obtained. In our article [13], we used a similar scheme, but generated the initial covers by running our ECI heuristic.

A comparison of several different separation algorithms for CIs, ECIs, simple LCIs, and general LCIs, can be found in Ref. 13.

APPLICATIONS

Application to Combinatorial Optimization Problems

Many important combinatorial optimization problems can be relaxed, in a natural way, so that they decompose into one or more 0–1 KPs. If one solves such a problem with an LP-based solution algorithm, such as branch-and-cut, then one can use CIs, ECIs, or LCIs as cutting planes.

For the sake of brevity, we mention just a few problems to which this approach has been applied:

- the *multiple KP* [14],
- the *multidimensional KP* [13,17,19,20],
- the *generalized assignment problem* [21,22],
- the *capacitated facility location problem* [23,24],
- the *fixed-charge transportation problem* [25],
- the *capacitated network design problem* [26],
- the *prize-collecting traveling salesman problem* [27,28],
- the *orienteeing problem* [29],
- the *resource-constrained project scheduling problem* [30].

Application to General 0–1 Linear Programs

As mentioned in the introduction, valid inequalities for 0–1 knapsack polytopes can in fact be used as cutting planes for general 0–1 LPs. The idea is to consider each individual constraint of the 0–1 LP as a knapsack constraint (complementing variables if necessary to obtain nonnegative coefficients), and derive inequalities that are valid for the associated knapsack polytope. Provided the constraint matrix of the 0–1 LP is sparse, the inequalities generated in this way can be expected to be useful.

To our knowledge, Crowder *et al.* [9] were the first to test this idea computationally. They designed what is now called a *cut-and-branch* algorithm, in which cutting planes are added only at the root node of

Table 1. Percentage Gap Closed by Inequalities on MIPLIB Instances

Name	CI	ECI	sLCI	gLCI	All
P0033	63.55	71.93	71.93	80.62	87.42
P0040	76.82	100.00	100.00	100.00	100.00
P0201	33.78	33.78	33.78	33.78	33.78
P0282	94.27	94.35	94.35	96.21	98.59
P0291	95.23	95.23	95.23	96.40	99.43
P0548	67.67	67.68	67.68	67.71	84.34
P2756	86.26	86.26	86.26	86.26	86.36
bm23	5.56	15.82	15.82	16.11	20.08
lseu	39.87	61.36	61.36	66.20	76.09
mod008	3.75	17.59	17.59	18.24	89.23
pipex	16.68	27.96	27.96	73.34	86.55
sentoy	16.58	21.57	21.57	22.72	30.97
Average	50.00	57.80	57.80	63.13	74.49

the branch-and-bound tree. The cutting planes used were simple LCIs. Hoffman and Padberg [18] extended this approach in two ways, by using full branch-and-cut (in which cutting planes can be added at any node), and using general LCIs in place of simple LCIs. Further experiments with this scheme were performed by Gu *et al.* [5].

CIs, ECIs, and LCIs turn out to be remarkably effective when applied to large, sparse 0–1 LPs. Table 1, adapted from Ref [13], shows the percentage of the integrality gap that is closed by various inequalities, when applied to twelve 0–1 LPs taken from the MIPLIB [31]. The columns labeled “sLCI” and “gLCI” correspond to simple and general LCIs, respectively. The column labeled “All” was obtained by calling an exact (and time-consuming) separation algorithm for general facets of knapsack polytopes. The CIs and ECIs were separated exactly, whereas the LCIs were separated heuristically. In all cases, apart from the “All” column, the running times were negligible compared to the time taken to solve the instances to proven optimality.

Observe that even the (theoretically weak) CIs close half of the integrality gap on average, and the other inequalities close even more. This good performance is typical, though only when the constraint matrix is sparse [13]. Observe also that the simple LCIs gave exactly the same results as the ECIs. Indeed, we observed that up-lifting

coefficients greater than 1 occurred very infrequently. This suggests that, if one does not wish to implement down-lifting, then one might as well use ECIs.

FOR FURTHER READING

There exist hard instances of the 0–1 KP that require an exponential number of nodes when solved by branch-and-bound, even if all simple LCIs are included in the LP relaxation [15,32].

LCIs are not the only facet-defining inequalities for 0–1 knapsack polytopes. For other such inequalities, see Refs 33–35.

Separation algorithms have been devised for the 0–1 knapsack polytope itself, rather than for specific classes of inequalities [13,36–38].

Cover inequalities have been derived for more complex KPs, with general integer variables, continuous variables, and so on [20,39–46].

An important related class of inequalities are the *lifted flow cover* inequalities [47–49], which are valid for fixed-charge problems.

REFERENCES

1. Balas E. Facets of the knapsack polytope. *Math Program* 1975;8:146–164.
2. Hammer PL, Johnson EL, Peled UN. Facets of regular 0-1 polytopes. *Math Program* 1975; 8:179–206.
3. Wolsey LA. Faces for linear inequalities in 0-1 variables. *Math Program* 1975;8:165–178.
4. Van Roy TJ, Wolsey LA. Solving mixed integer programming problems using automatic reformulation. *Oper Res* 1987;35:45–57.
5. Gu Z, Nemhauser GL, Savelsbergh MWP. Lifted cover inequalities for 0-1 linear programs: computation. *INFORMS J Comput* 1998;10:427–437.
6. Padberg MW. $(1, k)$ -configurations and facets for packing problems. *Math Program* 1980;18:94–99.
7. Balas E, Zemel E. Facets of the knapsack polytope from minimal covers. *SIAM J Appl Math* 1978;34:119–148.
8. Gu Z, Nemhauser GL, Savelsbergh MWP. Sequence independent lifting in mixed integer programming. *J Comb Opt* 2000;4: 109–129.
9. Crowder H, Johnson E, Padberg MW. Solving large-scale 0-1 linear programming programs. *Oper Res* 1983;31:803–834.
10. Zemel E. Easily computable facets of the knapsack polytope. *Math Oper Res* 1989;14: 760–765.
11. Hartvigsen D, Zemel E. The complexity of lifted inequalities for the knapsack problem. *Discr Appl Math* 1992;39:113–123.
12. Easton T, Hooker K. Simultaneously lifting sets of binary variables into cover inequalities for knapsack polytopes. *Discr Opt* 2008;5:254–261.
13. Kaparis K, Letchford AN. Separation algorithms for 0-1 knapsack polytopes. *Math Program* 2009. DOI: 10.1007/s10107-010-0359-5
14. Ferreira CE, Martin A, Weismantel R. Solving multiple knapsack problems by cutting planes. *SIAM J Opt* 1996;6:858–877.
15. Gu Z, Nemhauser GL, Savelsbergh MWP. Lifted cover inequalities for 0-1 integer programs: complexity. *INFORMS J Comput* 1999;11:117–123.
16. Klabjan D, Nemhauser GL, Tovey C. The complexity of cover inequality separation. *Oper Res Lett* 1998;23:35–40.
17. Gabrel V, Minoux M. A scheme for exact separation of extended cover inequalities and application to multidimensional knapsack problems. *Oper Res Lett* 2002;30:252–264.
18. Hoffman KL, Padberg MW. Improving LP-representations of zero-one linear programs for branch-and-cut. *ORSA J Comput* 1991;3:121–134.
19. Bektas T, Oguz O. On separating cover inequalities for the multidimensional knapsack problem. *Comput Oper Res* 2007;34: 1771–1776.
20. Kaparis K, Letchford AN. Local and global lifted cover inequalities for the multidimensional knapsack problem. *Eur J Oper Res* 2008;186:91–103.
21. Cattrysse D, Degraeve Z, Tistaert J. Solving the generalised assignment problem using polyhedral results. *Eur J Oper Res* 1996;108:618–628.
22. Gottlieb ES, Rao MR. The generalized assignment problem: valid inequalities and facets. *Math Program* 1990;46:31–52.
23. Aardal K. Capacitated facility location: separation algorithms and computational experience. *Math Program* 1998;81:149–175.

24. Aardal K, Pochet Y, Wolsey LA. Capacitated facility location: valid inequalities and facets. *Math Oper Res* 1995;20:562–582.
25. Nemhauser GL, Wolsey LA *Integer and combinatorial optimization*. New York: Wiley; 1988.
26. Chouman M, Crainic TG, Gendron B. A cutting plane algorithm for multicommodity capacitated fixed-charge network design. Technical Report. Montreal: CIRRELT; 2009.
27. Balas E. The prize collecting traveling salesman problem. *Networks* 1989;19:621–636.
28. Bérubé J-F, Gendreau M, Potvin J-Y. A branch-and-cut algorithm for the undirected prize collecting traveling salesman problem. *Networks* 2009;54:56–67.
29. Fischetti M, Salazar-Gonzalez JJ. Solving the orienteering problem through branch-and-cut. *INFORMS J Comput* 1998;10:133–148.
30. Hardin JR, Nemhauser GL, Savelsbergh MWP. Strong valid inequalities for the resource-constrained scheduling problem with uniform resource constraints. *Discr Opt* 2008;5:19–35.
31. Achterberg T, Koch T, Martin A. MIPLIB 2003. *Oper Res Lett* 2006;34:361–372.
32. Hunsaker B, Tovey C. Simple lifted cover inequalities and hard knapsack problems. *Discr Opt* 2005;2:219–228.
33. Atamtürk A. Cover and pack inequalities for (mixed) integer programming. *Ann Oper Res* 2005;139:21–38.
34. Weismantel R. Hilbert bases and the facets of special knapsack polytopes. *Math Oper Res* 1996;21:886–904.
35. Weismantel R. On the 0-1 knapsack polytope. *Math Program* 1997;77:49–68.
36. Avella P, Boccia M, Vasilyev I. A computational study of exact knapsack separation for the generalized assignment problem. *Comput Optim Appl* 2010;45:543–555.
37. Boyd EA. A pseudo-polynomial network flow formulation for exact knapsack separation. *Networks* 1992;22:503–514.
38. Boyd EA. Generating Fenchel cutting planes for knapsack polyhedra. *SIAM J Opt* 1993;3:734–750.
39. Boyd EA. Polyhedral results for the precedence-constrained knapsack problem. *Discr Appl Math* 1993;41:185–201.
40. Ceria S, Cordier C, Marchand H, *et al.* Cutting planes for integer programs with general integer variables. *Math Program* 1998;81:201–214.
41. De Farias IR, Johnson EL, Nemhauser GL. Facets of the complementarity knapsack polytope. *Math Oper Res* 2002;27:210–226.
42. Fernández E, Jørnsten K. Partial cover and complete cover inequalities. *Oper Res Lett* 1994;15:19–33.
43. Glover F, Sherali HD. Second-order cover inequalities. *Math Program* 2008;114:207–234.
44. Louveaux Q, Weismantel R. Polyhedral properties for the intersection of two knapsacks. *Math Program* 2008;113:15–38.
45. Marchand H, Wolsey LA. The 0-1 knapsack problem with a single continuous variable. *Math Program* 1999;85:15–33.
46. Nemhauser GL, Vance PH. Lifted cover facets of the 0-1 knapsack polytope with GUB constraints. *Oper Res Lett* 1994;16:255–263.
47. Gu Z, Nemhauser GL, Savelsbergh MWP. Lifted flow cover inequalities for mixed 0-1 integer programs. *Math Program* 1999;85:439–467.
48. Padberg MW, Van Roy TJ, Wolsey LA. Valid linear inequalities for fixed charge problems. *Oper Res* 1985;33:842–861.
49. Van Roy TJ, Wolsey LA. Valid inequalities for mixed 0-1 programs. *Discr Appl Math* 1986;14:199–213.

CREDIT RISK ASSESSMENT

ARCADY NOVOSYOLOV
Institute of Mathematics, Siberian
Federal University,
Krasnoyarsk, Russia

Credit risk arises when and where any two parties go into a borrowing and lending agreement, so that one party (lender) agrees to lend another party (borrower) a fixed amount of money for a fixed period of time (till maturity). The borrower returns the lender the borrowed money according to a prespecified schedule; usually the returned amount is greater than the borrowed one, the extra payment is called *interest*, it acts as a credit price.

BOND

The simplest form of credit agreement is a bond. Zero-coupon bond represents the obligation of the bond issuer (the party that borrows money) to return the nominal of the bond to the bond holder (the party that lends money) at maturity. Usually bonds also assume regular, say annual, coupon payments, the size of which is a fixed percentage of the nominal bond value.

The price p of the bond with nominal (face) value A that expires in T years and pays annual coupon amount a starting a year from now is equal to its present value calculated with annual interest rate r in mind

$$p = f(r) = \frac{a}{1+r} + \frac{a}{(1+r)^2} + \cdots + \frac{a}{(1+r)^{T-1}} + \frac{a+A}{(1+r)^T}.$$

This formula is also used another way round: given the market price p of a bond, calculate the interest rate y such that $f(y) = p$; this interest rate y is called *yield to maturity*.

More generally, if a debt tool promises payments A_i at times $t_i, i = 1, 2, \dots, n$, then its

present value $f_t(r)$ at time $t < \min\{t_1, \dots, t_n\}$ given annual interest rate r is calculated by

$$f_t(r) = \frac{A_1}{(1+r)^{(t_1-t)}} + \cdots + \frac{A_n}{(1+r)^{(t_n-t)}} = \sum_{i=1}^n \frac{A_i}{(1+r)^{(t_i-t)}}. \quad (1)$$

To illustrate basic concepts, consider an example of a 5-year coupon bond with face value \$300 and annual coupon payment 3% or \$9. This bond generates the cash flow shown in the Table 1; weights row corresponds to Macaulay duration, see Equation (7) below.

Given the interest rate $r = 0.05$ (or 5%), the initial price of this bond is equal to

$$f_0(0.05) = \frac{9}{1.05} + \frac{9}{1.05^2} + \frac{9}{1.05^3} + \frac{9}{1.05^4} + \frac{309}{1.05^5} = \$274.02.$$

Half a year from now the price of this bond, assuming the same interest rate 5%, would raise to

$$f_{1/2}(0.05) = \frac{9}{1.05^{1/2}} + \frac{9}{1.05^{3/2}} + \frac{9}{1.05^{5/2}} + \frac{9}{1.05^{7/2}} + \frac{309}{1.05^{9/2}} = 280.79.$$

This is how the investment in bonds works. The bond value increases by \$6.77 in half a year. However, this is true only if interest rate remains unchanged. Actually, there is an interest rate risk, which means that the bond value may change if interest rate changes. Indeed, if interest rate raises to $r = 0.06$ (6%) in half a year, the new bond price becomes equal to

$$f_{1/2}(0.06) = \frac{9}{1.06^{1/2}} + \frac{9}{1.06^{3/2}} + \frac{9}{1.06^{5/2}} + \frac{9}{1.06^{7/2}} + \frac{309}{1.06^{9/2}} = \$269.84,$$

Table 1. Example Cash Flow

Year Number (i)	1	2	3	4	5
Time (t_i)	1	2	3	4	5
Payment (\$)	9	9	9	9	309
Weights [$w_{i,1/2}(0.05)$]	0.031	0.030	0.028	0.027	0.884

so that the investment in this case brings loss instead of gain, and the bond value is by \$10.95 less than expected. If, on the contrary, the interest rate falls to $r = 0.04$ (4%) in half a year, then the new bond price becomes

$$f_{1/2}(0.04) = \frac{9}{1.04^{1/2}} + \frac{9}{1.04^{3/2}} + \frac{9}{1.04^{5/2}} + \frac{9}{1.04^{7/2}} + \frac{309}{1.04^{9/2}} = \$292.32,$$

so that the investment brings gain, which is by $\$292.32 - \$280.79 = \$11.53$ greater than expected. The changing interest rate would cause deviation of future bond price from the expected one. This deviation may be quantified by derivatives of bond price function (Eq. 1). Thus, the derivatives can serve risk indicators for the bond value.

DURATION AND CONVEXITY

Minus derivative of f with respect to interest rate r is called [1] *modified dollar duration*.

$$D_t(r) = -f'_t(r) = \sum_{i=1}^n \frac{A_i(t_i - t)}{(1+r)^{t_i+1-t}}. \quad (2)$$

Modified dollar duration $D_t(r)$ indicates how large the change $\Delta f_t(r)$ of the bond price resulting from a small change Δr of interest rate would be

$$\Delta f_t(r) \approx -D_t(r)\Delta r. \quad (3)$$

In the above example, the modified dollar duration in half a year equals to

$$D_{1/2}(0.05) = \frac{9}{1.05^{1/2}} + \frac{9}{1.05^{3/2}} + \frac{9}{1.05^{5/2}} + \frac{9}{1.05^{7/2}} + \frac{309}{1.05^{9/2}} = \$1123.6,$$

thus, changing interest rate by 0.01 (1%) would cause a change in the bond value by

approximately $\$1123.6 \times 0.01 \approx \11.24 . In particular, raising the interest rate to 0.06 provides approximate price by

$$\begin{aligned} \tilde{f}_{1/2}(0.06) &= f_{1/2}(0.05) - D_{1/2}(0.05) \times 0.01 \\ &= \$280.79 - \$1123.6 \times 0.01 \\ &= \$269.55. \end{aligned} \quad (4)$$

Dividing the expression for $D_t(r)$ by bond price $f_t(r)$, we obtain *modified duration* [2]

$$B_t(r) = -\frac{f'_t(r)}{f_t(r)} = -\frac{1}{f_t(r)} \sum_{i=1}^n \frac{A_i(t_i - t)}{(1+r)^{t_i+1-t}}. \quad (5)$$

The modified duration indicates how large the relative change $\Delta f_t(r)/f_t(r)$ of the bond price resulting from a small change Δr of interest rate would be

$$\frac{\Delta f_t(r)}{f_t(r)} \approx -B_t(r)\Delta r. \quad (6)$$

In the above example, the modified duration

$$B_{1/2}(0.05) = \frac{D_{1/2}(0.05)}{f_{1/2}(0.05)} \approx 4.$$

This means that raising the interest rate by 1% would cause a drop of the bond value by about 4%. Indeed, making calculations for $t = 1/2$ and raising interest rate from 0.05 to 0.06, we get

$$\frac{f_{1/2}(0.06) - f_{1/2}(0.05)}{f_{1/2}(0.05)} = -0.039,$$

which is close to the predicted drop in value by 4%.

Multiplying $B_t(r)$ by $1+r$, we get the *Macaulay duration* (also known as *unmodified duration*, see Ref. 2)

$$M_t(r) = -(1+r) \frac{f'_t(r)}{f_t(r)} = \sum_{i=1}^n w_{it}(r)(t_i - t), \quad (7)$$

where

$$w_{it}(r) = \frac{A_i}{(1+r)^{t_i-t} f_t(r)}, \quad i = 1, \dots, n.$$

We have

$$w_1 \geq 0, \dots, w_n \geq 0; w_1 + \dots + w_n = 1,$$

so Equation (7) measures how long, on average, the investor waits before receiving a payment, which justifies the name “duration.” Weights values for the above example with $t = 1/2$ and $r = 0.05$ are presented in the last row of Table 1, and the corresponding value of the Macaulay duration is equal to $M_{1/2}(0.05) = 4.202$ years. Macaulay duration is convenient within the continuous compounding model [1].

Duration parameters provide linear approximations of the sort of Equations (3) and (6). To get better approximation, one can use quadratic approximation. Define *modified convexity* $C_t(r)$ as the second derivative of the bond price, divided by the bond price [2]

$$C_t(r) = \frac{f_t''(r)}{f_t(r)}. \quad (8)$$

Then the relative change of the bond price may be approximated by

$$\frac{\Delta f_t(r)}{f_t(r)} \approx -B_t(r)\Delta r + \frac{1}{2}C_t(r)(\Delta r)^2. \quad (9)$$

In the example of Table 1, set $t = 1/2, r = 0.05$. We have $B_{1/2}(0.05) = 4$ and $C_{1/2}(0.05) = 20.57$, so raising the interest rate to $r = 0.06$ provides the approximation

$$\begin{aligned} \tilde{f}_t(0.06) &= \$280.79 \times \left(1 - 4 \times 0.01 + \frac{1}{2} \right. \\ &\quad \left. \times 20.57 \times 0.0001 \right) = \$269.843 \end{aligned}$$

with error \$0.003, which is much better than linear approximation (Eq. 4) with error \$0.29.

DEFAULT RISK

Duration and convexity are used to measure the sensitivity of bond value to change of interest rate. These measures indicate *interest rate risk*. Another source of bond risk is the possibility of default, when a borrower cannot meet his/her obligation according to the debt contract (bond). This risk may be

described by the following two parameters: probability of default p during a period of 1 year, and recovery rate R ; the latter being the debt portion that is reimbursed to the lender, in case of default.

Some debt tools are regarded as *risk-free*, say, US government bonds. The yield to maturity of such bonds is called *risk-free interest rate* r . Risky (corporate and municipal) bonds should provide greater yield y , the difference $s = y - r$ is called *spread* of the corresponding bond, and may be regarded as price or premium for taking the risk of holding this risky bond. Let face value of a bond be A , then expected loss in case of default would be $A(1 - R)p$, and premium for risk taking is As . Assuming *risk-neutral* lender, who would require premium being equal to the expected loss, we have the equation $A(1 - R)p = As$, which gives for the *risk-neutral* probability of default

$$p = \frac{s}{1 - R}. \quad (10)$$

Consider an example. Let risk-free rate r be 3%, and a risky bond with face value A and maturity 1 year has yield to maturity 5%. This means that the spread equals 2% or 200 basis points, where basis point equals 0.0001 or 0.01%. Assuming recovery rate R to be known and equal to 50%, we get the risk-neutral default probability $p = 0.02/(1 - 0.5) = 0.04$.

Rating agencies such as S&P, Fitch, or Moody's, assign ratings to corporate debt instruments and calculate historical default probabilities for specific ratings. Examples of such ratings as well as values of corresponding default probabilities may be found at agencies sites [3–5].

Calculating default probability plays the central role in assessing default risk. The goal may be achieved using a number of models; we will consider structural models and reduced-form models.

STRUCTURAL MODELS

Structural models use information about company's equity prices, which are available for publicly traded companies. The first

model of the class belongs to Merton [6], [7, Chapter 22]. We will describe the Merton model following Ref. 7.

Suppose that a company's debt is represented by a single bond with face value D and maturity t . Denote V_0 and V_t , the values of the company's assets today and at time t , respectively, and let E_0 and E_t stand for values of equity today and at time t .

In case $V_t < D$, the company cannot make the debt repayment and should default; the equity value E_t equals 0 in this case. If, on the other hand, $V_t \geq D$, the company should repay the debt, and the equity equals $E_t = V_t - D$. In short,

$$E_t = \max(V_t - D, 0) = (V_t - D)_+, \quad (11)$$

which means that equity is a call option on assets V with strike price D and maturity t . Assuming that assets V follow a geometric Brownian motion as in Black–Scholes model [7, Chapter 13], we get the equity today value in the form

$$E_0 = V_0 N(d_1) - DE^{-rt} N(d_2), \quad (12)$$

where N stands for the standard normal cumulative distribution function,

$$d_1 = \frac{\log(V_0/D) + (r + \sigma_V^2/2)t}{\sigma_V \sqrt{t}},$$

$$d_2 = d_1 - \sigma_V \sqrt{t}, \quad (13)$$

and σ_V denotes asset volatility, which is assumed constant in this model. The risk-neutral probability of default in this model is equal to

$$p = N(-d_2). \quad (14)$$

Its calculation requires knowing the values of V_0 and σ_V , which are not directly observable in the market. However, from Ito's lemma [7], we get the equation

$$\sigma_E E_0 = N(d_1) \sigma_V V_0, \quad (15)$$

where σ_E stands for equity volatility, which may be extracted from option prices on the company stocks. Solving the two nonlinear

Equations (12) and (15), we get the values of V_0 and σ_V . Now, the probability of default is obtained by applying Equations (13) and (14).

Consider an example. Let $E_0 = 5$, $D = 10$, $\sigma_E = 0.3$, $r = 0.03$, and $t = 2$. Solving Equations (12) and (15), gives $V_0 = 14.41$ and $\sigma_V = 0.1043$. From Equation (13), we have $d_2 = 2.8121$, and, finally, $p = 0.0025$.

PRACTICAL EXAMPLE OF DEFAULT CORRELATION

Now consider a practical example of how underestimation of default correlation might cause financial instability and even a crisis. Indeed such underestimation was one of the reasons which led to the recent 2008 financial turmoil.

The chain that led to mortgage crisis is rather complicated, so we would explain the things using simpler yet similar framework of collateralized debt obligations (CDO).

Suppose a bank holds a portfolio of n bonds of different companies similar in all respects. The bonds possess the same credit quality, in particular, the same default probability p , and all have the face amount of 1 and the same maturity 1 year. For simplicity, we assume zero recovery rate. The bank wishes to transfer (a part of) the default risk to third parties in exchange for (a part of) interest rate payments. To do so, the bank creates the derivative assets called *tranches*. The number of tranches may be different, let it be three in our example, we will call them the worst, the intermediate, and the best quality tranches.

The whole system works as follows. The worst quality tranche accumulates first n_1 defaults. This means that if the actual number m of defaults till maturity does not exceed n_1 , then the holders of the intermediate and the best tranches get full repayment of their debt, while the holder of the worst tranche gets repayment for $n_1 - m$ bonds. The intermediate tranche accumulates the next n_2 defaults. This means that if the actual number of defaults m satisfies $n_1 < m \leq n_1 + n_2$, then the holder of the worst tranche gets no repayments on his/her debt, the holder of the best tranche get full repayment, and

the holder of the intermediate tranche gets repayment on $n_1 + n_2 - m$ bonds. The best tranche accumulates the rest $n_3 = n - n_1 - n_2$ defaults. This means that if the actual number m of defaults exceeds $n_1 + n_2$, then the holders of the worst and the intermediate tranches get no repayments, and the holder of the best tranche gets repayment for $m - n_1 - n_2$ bonds.

The benefit for tranches holders consists of getting interest rate payments and debt repayments conditioned according to the specified scheme. The difference of quality if reflected in the different interest rate flows for tranches. Detailed interest rate calculation is a matter of CDO pricing and is rather complicated. However, it is clear that greater default probability for a tranche should be compensated by greater interest rate payments.

In our example, we will show how wrong specification of default correlation can

dramatically change the probabilities of losses for tranches, especially for the best one. This is exactly what was going on in the mortgage-backed securities market, where tranches rated as AAA securities brought huge losses during 2008.

Now, let $n = 100, n_1 = 15; n_2 = 10, n_3 = 75$, and $p = 0.1$. Assume pairwise default correlation being the same for all pairs of bonds, and it being denoted as ρ . Denote A, B , and C events linked to the random number of defaults m , specifically: $A = \{m \leq n_1\}, B = \{n_1 < m \leq n_1 + n_2\}, C = \{n_1 + n_2 < m \leq n\}$. Monte Carlo simulation of this CDO with 10,000 trials gives the following probabilities for the events A, B, C for a number of ρ values in Table 2:

Judging by the probabilities, we can see that assuming zero correlation, we can assign the best tranche the AAA rating quality by the S&P rating system. However, if $\rho = 0.1$, its rating jumps down to B, which is outside the investment grade. Further correlation increasing moves the tranche even deeper into “junk bonds” area. Figures 1, 2 and 3 illustrate the distributions of defaults number for different values of ρ .

In a few words, the subprime mortgage crisis may be explained as follows: wrong model assessment allowed assigning too high credit rating to the best tranches, which provoked high demand from conservative investors, which gave rise to a flow of poor quality mortgages, and the whole system became a bubble.

Table 2. Probabilities of Default for Tranches

ρ	$P(A)$	$P(B)$	$P(C)$
0	0.9587	0.0413	0
0.1	0.8234	0.1495	0.0271
0.2	0.7839	0.1472	0.0689
0.3	0.7863	0.1218	0.0919
0.4	0.7788	0.1078	0.1134
0.5	0.7823	0.0848	0.1329
0.6	0.7943	0.0708	0.1349

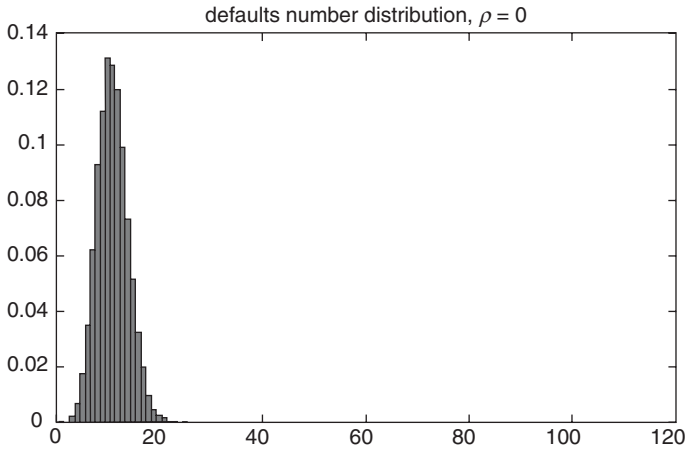


Figure 1. Distribution of number of defaults in example when correlation $\rho = 0$.

Figure 2. Distribution of number of defaults in example when correlation $\rho = 0.2$.

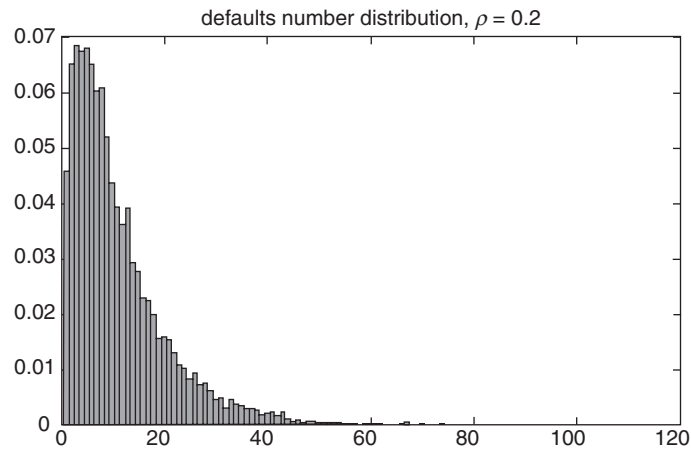
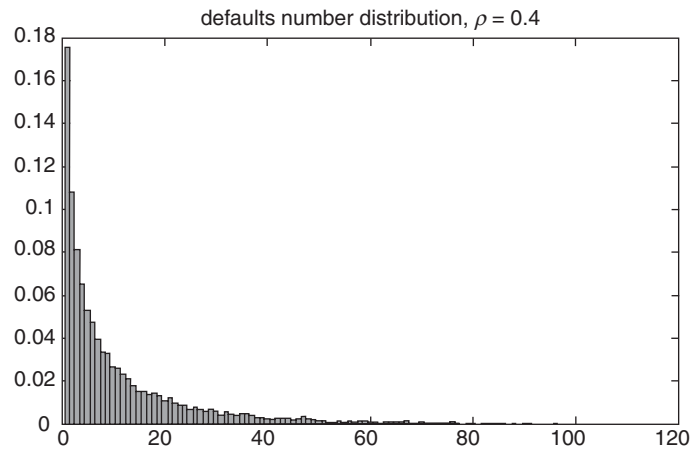


Figure 3. Distribution of number of defaults in example when correlation $\rho = 0.4$.



FURTHER READING

More on duration and convexity of bonds may be found in Ref. 8 (Chapter 13). Calculation of default probability for coupon bearing bonds is illustrated by examples in Ref. 7 (Chapter 22). Empirical analysis of several structural models is presented in Ref. 9. For more practical examples of credit turmoil, the reader is referred to Ref. 10 (Chapter 1).

REFERENCES

1. Panjer H, *et al.* Financial economics: with applications to investments, insurance and pensions. Canada: Society of Actuaries; 1998. p. 669.
2. Golub BW, Tilman LM. Risk management. Approaches for fixed income markets. New York: Wiley; 2000. p. 312.
3. S&P site. 2009. Available at <http://www.standardandpoors.com/>.
4. Moody's site. 2009. Available at <http://www.moodys.com/>.
5. Fitch ratings site. 2009. <http://www.fitchratings.com/>.
6. Merton RC. On the pricing of corporate debt: the risk structure of interest rates. *J Finance* 1974;29(2):449–470.
7. Hull JC. Options, futures and other derivatives. Princeton (NJ): Prentice Hall; 2006. p. 789.
8. Wilmott P. Paul Wilmott on quantitative finance. Boston (MA): John Wiley & Sons; 2006.

9. Eom YH, Helwege J, Huang J. Structural models of corporate bond pricing: an empirical analysis. *Rev Financ Stud* 2004;17(2): 499–544.
10. Abrahams CR, Zhang M. Credit risk assessment: the new lending system for borrowers, lenders, and investors. New York: Wiley; 2009. p. 320.

CREDIT RISK

TIM LEUNG
IEOR Department, Columbia
University, New York, NY USA
RONNIE SIRCAR
ORFE Department, Princeton
University, Princeton, NJ USA

INTRODUCTION

Credit derivatives are financial securities whose payouts are contingent on one or many *default events*, such as the bankruptcy of a firm, nonrepayment of a loan, or missing a mortgage payment. These instruments enormously grew in popularity in the 1990s and 2000s, and obviously declined a lot with the financial crisis of 2008. However, credit markets are still of central importance, as struggling economies enhance default risk of firms and countries. Two recent credit-linked calamities are the impact of potential default by Greece on many credit default swaps (CDSs) written by systemically large financial firms, and the \$5+ billion loss by J.P. Morgan in 2012 from large bets on the CDX credit index.

The main task is to model a random default time when a firm or obligor cannot meet a payment. This may come from a diffusion model of asset values hitting a fixed debt level, in which case we can, in some sense, see the default event coming, or it may be modeled as the jump time of some exogenous Poisson-type process (which does not have a direct economic interpretation), in which case the default comes as a surprise. Valuing multiname credit derivatives involves modeling accurately correlation of defaults across firms, and evaluating basket portfolios affected by many sources of credit risk.

STRUCTURAL MODELS

The structural approach to credit risk involves modeling the default event through

the firm's asset and debt values. For example, Merton [1] assumes that default occurs if the firm's asset value S on date T are below a prespecified debt level D . The firm's creditors hold a bond promising payment of \$1 on expiration date T , unless the firm defaults, so the bond payoff is

$$h(S_T) = \mathbf{1}_{\{S_T > D\}}.$$

Assuming that the cash flows represented by S are perfectly hedgeable (e.g., by purchasing the firm's stock) and that S is a geometric Brownian motion, we can evaluate the Black-Scholes value of the defaultable bond.

Merton Model

In the Merton [1] model, the asset value S is modeled by the geometric Brownian motion:

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \quad (1)$$

where W is a Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Here, $\mathbf{P}$ is the historical or subjective measure. In this case, under the unique risk-neutral pricing measure $\mathbf{P}^*$, the asset price is explicitly given by

$$S_t = S_0 \exp \left(\left(r - \frac{1}{2} \sigma^2 \right) t + \sigma W_t^* \right), \quad (2)$$

where W^* is a $\mathbf{P}^*$ -Brownian motion, and r is the short rate (assumed constant).

Assuming that default occurs if $S_T < D$ for some threshold value D , the price at time t of a defaultable bond is simply the price of an European digital option, which pays 1 if S_T exceeds the threshold and 0 otherwise. It is explicitly given by $u^d(t, S_t)$ where

$$\begin{aligned} u^d(t, S) &= \mathbf{E}^* \left\{ e^{-r(T-t)} \mathbf{1}_{\{S_T > D\}} \mid S_t = S \right\} \quad (3) \\ &= e^{-r(T-t)} \Phi(d_2), \quad (4) \end{aligned}$$

where $\mathbf{E}^*$ is the expectation under measure $\mathbf{P}^*$, and $\Phi(\cdot)$ is the normal cumulative distribution function, and

$$d_2 = \frac{\log(S/D) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}. \quad (5)$$

The formula (3) often motivates the *distance to default* definition

$$\frac{\log(S/D)}{\sigma},$$

which increases as the firm value S is far from the debt level D but decreases with increasing volatility σ .

Black–Cox Model

In the Black and Cox [2] generalization, the default time is modeled by the first time the asset value S hits a lower threshold D , namely,

$$\tau_t := \inf\{s \geq t \mid S_s \leq D\}. \quad (6)$$

Then, the value of a defaultable bond, given no default has occurred at time $t \leq T$, is given by

$$\begin{aligned} u(t, T) &= \mathbf{E}^* \left\{ e^{-r(T-t)} \mathbf{1}_{\{\tau_t > T\}} \mid S_t = S \right\} \\ &= \mathbf{E}^* \left\{ e^{-r(T-t)} \mathbf{1}_{\{\inf_{t \leq s \leq T} S_s > D\}} \mid S_t = S \right\}. \end{aligned} \quad (7)$$

If the asset price has reached D before time t , then $u(t, T) = 0$.

This defaultable zero-coupon bond is in fact a *binary down-and-out barrier option* where the barrier level and the strike price coincide. This can be computed using the distribution of the minimum of a (nonstandard) Brownian motion, obtained from the *reflection principle*, after a Girsanov change of measure. (See Ref. 3 for details of this calculation). Alternatively, the *method of images* [4] leads to a formula for $u(t, S)$ in terms of the Black–Scholes price u^d of the corresponding digital option:

$$u(t, S) = u^d(t, S) - \left(\frac{S}{D}\right)^{1-\frac{2r}{\sigma^2}} u^d\left(t, \frac{D^2}{S}\right), \quad (8)$$

where u^d is the digital pricing function given in Equation (3).

Spread Curves in the Black–Cox Model

The *yield spread* $Y(0, T)$ at time zero is defined by

$$e^{-Y(0, T)T} = \frac{P^B(0, T)}{P(0, T)}, \quad (9)$$

where $P(0, T)$ is the default-free zero-coupon bond price. Under the Black–Cox model with constant interest rate r , we have $P(0, T) = e^{-rT}$ and $P^B(0, T) = u(0, S)$ from Equation (7). This leads to the yield spread formula

$$Y(0, T) = -\frac{1}{T} \log \left(\Phi(d_2^+(T)) - \left(\frac{S}{D}\right)^{1-\frac{2r}{\sigma^2}} \Phi(d_2^-(T)) \right), \quad (10)$$

where we denote

$$d_2^\pm(T) = \frac{\pm \log\left(\frac{S}{D}\right) + \left(r - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}. \quad (11)$$

As discussed in Ref. 5, the challenge for theoretical pricing models is to raise the average predicted spread relative to crude models such as the constant volatility model presented in this section, without overstating the risks associated with volatility or leverage. To this end, several approaches have been proposed, including the introduction of jumps [6–8], stochastic interest rate [9], stochastic volatility [10], or imperfect information [11,12]. On the basis of a comprehensive empirical analysis, Eom *et al.* [5] find that structural models proposed in the literature [9,13] still have difficulties in predicting realistic credit spreads at short maturities.

As is well documented in the literature, in this first passage model, the likelihood of default is essentially 0 for short maturities even for highly levered firms, corresponding to S/D close to 1, as illustrated in Figure 1. In fact, for any continuous diffusion process S , the default time τ_t is a *predictable* stopping time, in the sense that it can be *announced* by an increasing sequence of stopping times. For instance, one can consider the sequence $(\tau_t^{(n)})$ defined by $\tau_t^{(n)} := \inf\{s \geq t, S_s \leq D + 1/n\}$. This feature will be one of the motivations to move to intensity-based models in which default is modeled as a surprise.

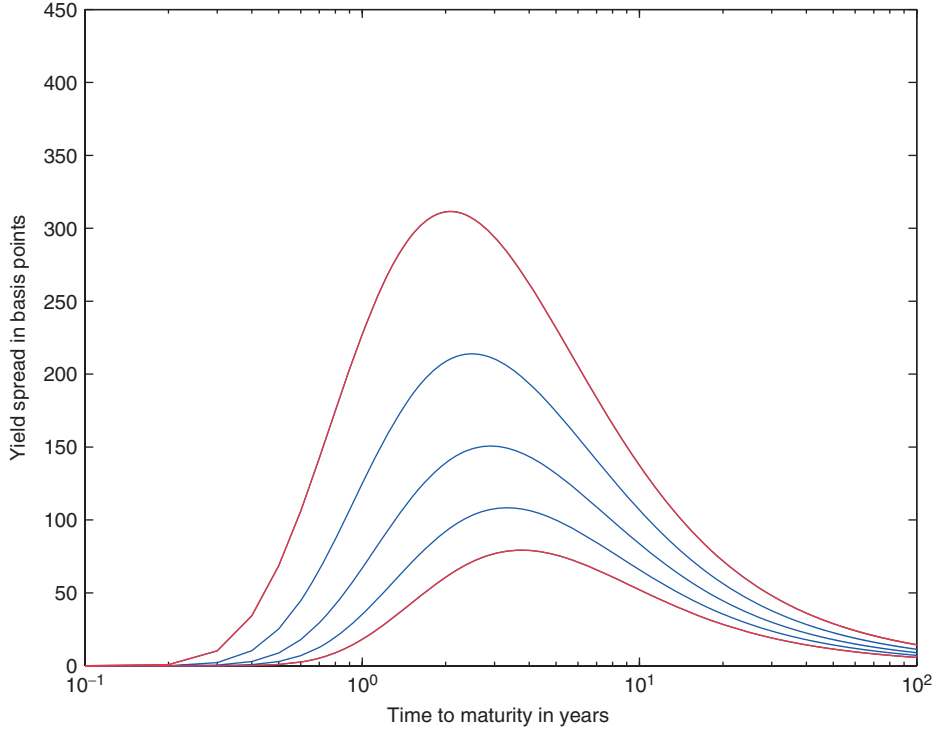


Figure 1. This figure shows the sensitivity of the yield spread to the leverage level. The volatility level is set to 10% and the interest rate is 6%. The curves increase with the decreasing ratios S/D : (1.3, 1.275, 1.25, 1.225, 1.2).

INTENSITY-BASED MODELS

The last observation serves as a motivation for intensity-based (or reduced-form) models in which default occurs at the jump of an exogenous Poisson-type process. In contrast to structural models, as the intensity-based default time is *totally inaccessible*, and cannot be predicted by any predictable stopping times [14]. The element of surprise makes intensity-based models capable of explaining nonzero credit spreads even at short maturities.

Doubly Stochastic or Cox Process

The default times are generated by a two-step randomization process described by a stochastic intensity process, λ , and Poisson process, N . We start with a probability space $(\Omega, \mathcal{G}, \mathbf{P})$ and a filtration $(\mathcal{G}_t)$. Let (H_t) be a counting process with jump times $\tau_1, \tau_2, \dots$, satisfying the nonexplosive

condition: $\lim_{n \rightarrow \infty} \tau_n = +\infty$. Then, (λ_t) is called the *intensity of H* if:

1. λ is a nonnegative predictable process;
2. $\mathbf{E}\{\int_0^t \lambda_s ds\} < \infty$ for all $t > 0$;
3. the process $M_t = H_t - \int_0^t \lambda_s ds$, $t \geq 0$, is a martingale.

Now, suppose that $(\mathcal{F}_t)$ is a subfiltration, that is, $\mathcal{F}_t \subset \mathcal{G}_t$ for all t . Then H is called a *doubly stochastic process* driven by $(\mathcal{F}_t)$ (or a $(\mathcal{F}_t)$ -conditional Poisson process, or a Poisson process conditioned by $(\mathcal{F}_t)$) if λ_t is $\mathcal{F}_t$ -predictable and, for $t > s$, conditional on $\mathcal{G}_s \vee \mathcal{F}_t$, the increment $H_t - H_s$ is Poisson distributed with parameter $\int_s^t \lambda_u du$.

In practice, the filtration $(\mathcal{F}_t)$ will be generated by some factor processes $(\mathbf{Y}_t)$ and $\lambda_t = f(\mathbf{Y}_t)$, for some (bounded nonnegative) function f . The jump times can be generated in the following way. Let $Z_0, Z_1, \dots$ be *i.i.d.* exponentially distributed random variables

with mean 1. Define $\tau_0 = 0$. Then the jump times are given by

$$\tau_n = \inf \left\{ t \geq \tau_{n-1} : \int_{\tau_{n-1}}^t \lambda_s ds \geq Z_n \right\}. \quad (12)$$

From these, we define the counting process (H_t) and the filtration $(\mathcal{H}_t)$ generated by it. Finally, the full filtration is $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}_t$. This construction forms the basis of a simulation method in which we successively simulate unit exponential random variables and the path of λ , while we register a jump in H according to the condition in (11).

This also reveals that, as the standard Poisson process N has unit exponential times between jumps, H can be viewed as a time-changed Poisson process. In other words, $H_t = N_{\Lambda_t}$, where $\Lambda_t = \int_0^t \lambda_s ds$.

Connection with Interest Rate Derivatives

One reason for the rapid adoption of intensity-based models is that various standard calculations central to pricing defaultable securities are similar to interest-rate derivative calculations, so that effective models in fixed income, especially those with closed-form solutions, can be adapted for credit derivatives.

To see this, consider a defaultable bond, which pays \$1 on date T as long as the Cox process H has not jumped by then. As usual, we will denote by τ this jump time:

$$\tau = \inf\{t > 0 \mid H_t > 0\}.$$

The arbitrage-free price of the defaultable bond is given by the expectation of its discounted payoff under the market-chosen risk-neutral probability $\mathbf{P}^*$, which we assume to be given. If the interest rate is zero, we have

$$\begin{aligned} \mathbf{E}^*\{\mathbf{1}_{\{\tau > T\}}\} &= \mathbf{E}^*\{\mathbf{E}^*\{\mathbf{1}_{\{\tau > T\}} \mid \mathcal{F}_T\}\} \\ &= \mathbf{E}^*\left\{e^{-\int_0^T \lambda_s ds}\right\}, \end{aligned} \quad (13)$$

where we have used the property that, conditioned on the information $\mathcal{F}_T$ (in which case, the path $(\lambda_s)_{0 \leq s \leq T}$ is given), H is an inhomogeneous Poisson process with intensity given by that path. Therefore, we reduce to computing Equation (12), which is

similar to pricing a bond with λ playing the role of a stochastic interest rate.

CIR Default Intensity

Consider a default intensity of the form $\lambda(Y_t)$, where Y is a CIR (Cox-Ingersoll-Ross or Feller or square-root) process:

$$dY_t = a(\bar{y} - Y_t)dt + \sigma\sqrt{Y_t}dW_t^*. \quad (14)$$

This is a mean-reverting process (with long-run mean-level $\bar{y} > 0$) that stays positive with probability 1 as long as $\sigma^2 < 2a\bar{y}$. It was introduced as a model of short rates in Ref. 15. Properties of this process, including the aforementioned condition for it to stay positive, were studied by Feller [16].

To compute the conditional default probability $u(t, y) = \mathbf{P}^*\{\tau > T \mid Y_t = y\}$, we look for a solution that is exponentially affine in y :

$$u(t, y) = A(T - t)e^{-B(T-t)y},$$

for functions A and B to be determined. Substituting this *ansatz* into the PDE and comparing terms with and without y , we obtain two ODEs for A and B with the initial conditions $A(0) = 1$ and $B(0) = 0$. The ODEs can be solved explicitly (see Chapter 7, [17]), yielding

$$\begin{aligned} A(s) &= \left(\frac{2\theta e^{(\theta+a)s/2}}{(\theta+a)(e^{\theta s} - 1) + 2\theta} \right)^{2a\bar{y}/\sigma^2} \\ B(s) &= \frac{2(e^{\theta s} - 1)}{(\theta+a)(e^{\theta s} - 1) + 2\theta} \\ \theta &= \sqrt{a^2 + 2\sigma^2}. \end{aligned} \quad (15)$$

The yield spread (for constant short rate) is given by

$$\begin{aligned} Y_{\text{sp}}(0, T) &= -\frac{1}{T} \log u(0, y) \\ &= -\frac{1}{T} (\log A(T) - B(T)y). \end{aligned}$$

It is easy to show that

$$\begin{aligned} B(T) &\sim \frac{2}{\theta + a}, \quad \text{and} \\ A(T) &\sim \left(\frac{2\theta}{\theta + 1} e^{(a-\theta)T/2} \right)^{2a\bar{y}/\sigma^2}, \end{aligned}$$

as $T \rightarrow \infty$. Therefore, the long-term spread converges to a positive constant:

$$Y_{\text{sp}}(0, T) \rightarrow \frac{a\bar{y}}{\sigma^2}(\theta - a) = \frac{a\bar{y}}{\sigma^2}(\sqrt{a^2 + 2\sigma^2} - a).$$

Similarly, in the short maturity limit, as $T \downarrow 0$, we have $Y_{\text{sp}}(0, T) \rightarrow y > 0$.

Credit Default Swaps

A CDS contract is a derivative that provides insurance against the default of a reference entity, which is usually a corporation or a sovereign (i.e., a national government). In a single-name CDS, the *protection buyer*, that is, the counterparty that receives a payoff if the reference entity defaults, pays a periodic premium to the *protection seller*. In return, the protection seller has to compensate the protection buyer in the case where the reference entity defaults before a predetermined maturity time. The premiums that the protection buyer pays to the protection seller are usually paid quarterly or semiannually until the maturity of the CDS contract, or until the time of the default event, whichever comes first. CDSs are among the most liquid single-name credit derivatives and provide a good indication of the market's view toward the credit risk of the underlying firm.

To fix ideas, suppose that the CDS is written on a bond issued by the reference entity. In the case where the reference entity defaults before the maturity of the CDS contract, we assume that the protection seller compensates the protection buyer with a cash settlement. In particular, the protection seller will make a cash payment to the protection buyer equal to the difference of the notional amount of the bond and its postdefault market value, which is usually determined by polling several dealers. Here, we assume that in default the bond recovers $1 - q$ of its face value—which is known as *recovery of face value*—and the protection seller provides the remaining proportion q of the face value to the protection buyer.

While there are other types of settlement used in practice, we will not go into details here. For a detailed specification on the repayment methods at default, more

information on what constitutes a default event, usual practices for settlements in the financial industry, and other legal issues we refer to Chapter 8 [18].

The pricing of a CDS amounts to determining the *CDS spread*, which determines the amount paid by the protection buyer to the protection seller on each payment date. Note that the CDS is *unfunded* in the sense that entering into the contract, as the protection seller does not require depositing the potential insurance payments upfront: an investor can receive the premium (insurance premium) without having to tie up the possible insurance payout. As a result, they have been popular with hedge funds in that they provide access to returns from high yields for bearing credit risk but with little capital constraints.

MULTINAME CREDIT DERIVATIVES

Multiname derivatives, especially collateralized debt obligations (CDOs), have played a significant role in the 2008 financial crisis. It is a major challenge in credit risk to value exotic credit derivatives, whose payoffs are contingent on defaults of many (sometimes hundreds) different firms. These defaults are certainly not independent, as firms may be linked through contracts and transactions and share common macroeconomic conditions.

Collateralized Debt Obligations (CDOs)

CDOs have been attractive because of the high yields they offer relative to poorly performing stocks and bonds in recent years. They also allow certain pension funds, mutual funds, and endowments, which are prohibited from investing directly in lower-rated debt, to acquire higher yields from CDO tranches, which are typically given higher credit ratings than some of the components. See Refs 19–21 for a concise introduction.

A CDO has two parties, the protection buyer and the protection seller (the tranche holder). The former pays to the latter a yield or tranche spread (the insurance premium) on the remaining amount that the protection

seller is still responsible for. As losses reach his tranche, the protection provider pays those claims to the protection buyer.

As an example, the CDX index is an average yield spread for a standardized basket of 125 US investment grade firms. In a CDO on the CDX index, each firm has the same notional, \$80,000, meaning the default of each firm results in a loss of at most \$80,000, but usually less because there is some recovery, to be covered by the tranche holder. The total notional is then $\$80,000 \times 125 = \10 million. Its equity tranche covers the first 3% of losses, the first mezzanine tranche the next 4%, the second mezzanine tranche the next 3%, the third mezzanine the next 5%, and the senior tranche the next 15%. These reward the tranche holder with yields such as 1500, 400, 100, 50, 15 basis points on their remaining notional.

In general, the number of firms underlying the CDO is denoted by N and the total notional by M . Each tranche is characterized by its lower and upper attachment points, denoted K_L and K_U , respectively. For example, for CDOs on the CDX, we have

Tranche	K_L	K_U
Equity	0%	3%
Mezzanine 1	3%	7%
Mezzanine 2	7%	10%
Mezzanine 3	10%	15%
Senior	15%	30%

The corresponding standardized index portfolio for European firms is called the *iTraxx*, and CDOs on that have slightly different attachment points.

The focus of modeling in the credit derivatives industry has been on *correlation* between default times. This is partly due to the industry's early adoption of the one-factor Gaussian copula model as industry standard and the practice (up till recently) of analyzing through implied correlation. This revealed that traded prices of senior tranches could only be realized through these

models with an implausibly high correlation parameter, the so-called *correlation smile*. After the simultaneous downgrades of Ford and General Motors in May 2005, the standard copula model often could not fit the data.

Multiname Intensity Models

The pricing of multiname credit derivatives requires (i) realistic modeling of the firms' default times and the correlation between them and (ii) efficient computational methods for computing the portfolio loss distribution. Dynamic factor models, a widely used class of pricing models in the literature, can be computationally tractable despite the large dimension of the pricing problem, thus satisfying issue (ii), but to have any hope of calibrating market data, such as CDO spreads, numerically intense versions of these models have to be implemented.

The most widely used framework for single-name default modeling is the intensity-based setup, in which the arrival of default of a firm is described by the first jump of a *Cox process* (or doubly stochastic Poisson process). Extensions of this dynamic framework to multiname credit derivatives were proposed by Duffie and Gârleanu [22], and more recently by Mortensen [23] in the class of common factor models.

These are part of the *bottom-up* framework of multiname credit derivatives pricing models, as the distribution of cumulative loss is modeled through the individual firms' default times and losses. Within this framework, a popular approach is to assume that a firm's probability of default depends on two components: firm-specific (or idiosyncratic), and market-wide (or systematic). In addition, the conditional default events of any two firms in the portfolio, given the systematic factors, are independent. This class of factor models has been used for at least a decade, with the technical reports [24–26], where the notions of systematic and idiosyncratic credit risks were introduced.

Suppose there are N firms whose default times are τ_i , $i = 1, \dots, N$. We also have a fixed finite time horizon T (the maturity of the CDO), and assume a constant recovery rate $\delta \in [0, 1)$. Each default time is the first

jump of a Cox process with intensity process $(\lambda_t^{(i)})$, which is decomposed into the common factor way as

$$\lambda_t^{(i)} = X_t^{(i)} + a_i Y_t$$

Here, the a_i are nonnegative constants and the $X^{(i)}$ and Y are independent *affine jump diffusions* (AJD) suitably restricted to remain positive, so that the intensity remains positive. They are defined through the SDEs [27]:

$$\begin{aligned} dX_t^{(i)} &= \kappa_i(\theta_i - X_t^{(i)})dt + \sigma_i\sqrt{X_t^{(i)}}dW_t^{(i)} \\ &\quad + dJ_t^{(i)} \end{aligned} \quad (16)$$

$$\begin{aligned} dY_t &= \kappa_Y(\theta_Y - Y_t)dt + \sigma_Y\sqrt{Y_t}dW_t^{(Y)} \\ &\quad + dJ_t^{(Y)}, \end{aligned} \quad (17)$$

where the processes $(W^{(Y)}, W^{(1)}, \dots, W^{(N)})$ are independent Brownian motions and $(J^{(Y)}, J^{(1)}, \dots, J^{(N)})$ are independent pure jump processes with constant intensities $(\ell_Y, \ell_1, \dots, \ell_N)$ and jump sizes independent of the jump times that are exponentially distributed with means $(\mu_Y, \mu_1, \dots, \mu_N)$.

This is written in the shorthand

$$X^{(i)} \sim \text{AJD}(X_0^{(i)}, \kappa_i, \theta_i, \sigma_i, \ell_i, \mu_i), \quad (18)$$

$$Y \sim \text{AJD}(Y_0, \kappa_Y, \theta_Y, \sigma_Y, \ell_Y, \mu_Y). \quad (19)$$

We assume the usual restrictions: $\sigma_i^2 \leq 2\kappa_i\theta_i$ and $\sigma_Y^2 \leq 2\kappa_Y\theta_Y$. According to (16)–(17), along with the coefficients a_i , the total number of parameters in the model is $6 + 7N$, which is on the order of 800 for a standard CDO of 125 firms. Obviously to calibrate five or six tranche spreads at three or four different maturities, there is a massive over-specification problem.

Mortensen [23] discusses calibration, but with a very reduced form of the model, with a lot of homogeneous parameters. Of course, this is necessary because of the overfitting problem discussed previously. Similar reduction is assumed in Ref. 27. A more thorough test of this type of full bottom-up model is clearly needed, and perhaps the model should be reduced in a more data-driven manner than with an ad hoc selection [28].

Top-Down Models

While bottom-up models have the good feature of allowing for consistency with single-name yield spreads, but they can be computationally intractable when computing CDO spreads unless constructed in a fairly specific manner. In addition, they may be considered better than static copula models, but may have too many parameters to fit a few tranches and maturities. In contrast, the computational problem is trivialized if we pass on to top-down models in which the loss process is modeled directly. The trade-offs of the top-down approach are as follows:

1. No consistency with the constituent single-name spreads. The “down” in top-down is essentially redundant, as we do not know how to go from the macroscopic description of the loss process L_t down to each single-name default timing and loss distribution.
2. No direct way to create individual loss processes consistently for different but overlapping baskets consistent. One can potentially model the loss process for *all* firms associated with all contracts in question, but the difficulty lies in projecting to overlapping subsets for different multiname contracts.

On the other hand, the top-down approach is used all the time when we price S&P 500 index options by modeling the S&P 500 index directly, without worrying about its relation to the component stocks. One of the motivation practitioners has promoting top-down models, as effective descriptors of (at least index) CDO tranche spreads is to encourage a derivative market in options on tranches.

The top-down approach is relatively recent in credit risk. A starting point is [29], which discusses strategies for trying go “down” to single names. Models based on self-exciting Hawkes processes are proposed and analyzed in Ref. 30. Models based on Markov chains are proposed in Refs 31 and 32. An approach related to the HJM framework for interest-rate derivatives is proposed in Refs 33 and 34; see Ref. 35 for details.

Self-Affecting Models

There is a strong case made in Ref. 30, which default events cause credit spreads to rise across the board, and so jumps in the loss process should directly affect the distribution of the next default time, for example, by directly causing a spike in the intensity of the loss process. The bottom-up approach we studied previously had no such “event feedback,” as it was based on doubly stochastic default times. People argue that without event feedback, bottom-up models need “too much” correlation to accurately capture tranche spreads, though of course “too much” is a highly subjective issue.

A simple static model containing event feedback was proposed in Ref. 19. Let Z_i denote the indicator random variable of the i th firm: $Z_i = \mathbf{1}_{\{\tau_i \leq t\}}$, which is clearly a Bernoulli variable with 1 signifying default. The Z_i are given by

$$Z_i = X_i + (1 - X_i) \left(1 - \prod_{j \neq i} (1 - X_j Y_{ji}) \right),$$

where the X_i and Y_{ji} are independent Bernoulli random variables with success probabilities p and q , respectively. The X_i can be thought of as firm i ’s intrinsic default indicator. Alternatively, firm i may default because another firm j defaults and its connection Y_{ji} to it kicks in too. Thus it is “infected” by firm j . The pictures in the Davis-Lo paper show the range of loss distributions that can be achieved with this simple two-parameter family of models, so the infection mechanism is quite powerful.

In the Hawkes process model of Errais *et al.* [30], the default counting process (N_t) admits a stochastic intensity of the form:

$$d\lambda_t = \kappa(\lambda_\infty - \lambda_t) dt + \delta dN_t,$$

so each jump in N induces a jump of size δ in λ . Note however, owing to the macroscopic blurring of identity, any default causes a jump in λ , no matter how insignificant or unconnected that firm is to other members of the portfolio.

Arnsdorff and Halperin [31] argue that the intensity of the loss process should decrease, as the number of defaults increases, because there are less firms left. In other words, an imminent default is more likely when there are more firms in the basket. Hence, they propose an intensity process of the form

$$\lambda_t = f_{N_t}(t)(n - N_t),$$

where n is the total number of firms in the pool and the f_j are deterministic versions of time (to be calibrated). A more general model incorporating level-dependent intensity, Brownian fluctuations, self-excitation is

$$\lambda_t = F(t, N_t)Y_t,$$

where $F(t, j)$ is decreasing in j and the stochastic factor Y follows a jump-diffusion:

$$dY_t = (\dots) dt + (\dots) dW_t + (\dots) dN_t.$$

Lopatin and Misirpashaev [32] proposed a model the intensity

$$d\lambda_t = \kappa(\rho(t, L_t) - \lambda_t) dt + \sigma \sqrt{\lambda_t} dW_t,$$

where $\rho(t, L)$, a function of time and loss, is to be calibrated. This idea is analogous to volatility surface estimation (Dupire’s method, etc.) in equity derivatives.

FURTHER ISSUES IN CREDIT RISK

One recent development in credit risk focuses on modeling counterparty risk and the evaluation of credit valuation adjustment (CVA). This is highlighted in the recent financial crisis, where two-thirds of losses were attributed to CVA losses from counterparty risk, while only one-third were due to actual defaults of the underlying entities. This also raised the related issue of wrong way risk, whereby the underlying contract and the default of the counterparty are correlated in the worst possible way for the investor. The Basel Committee on Banking Supervision (BCBS) has proposed a new regulatory framework—Basel III—to increase bank capital requirements for counterparty credit

risk. Recent studies on counterparty risk for CDS and other derivatives include Refs 36 and 37.

An alternative to the standard risk-neutral pricing mechanism is the utility-based approach that incorporates market frictions, such as hedging constraints, as well as the investor's risk preferences. This methodology has been recently applied to manage credit risk for single-name and multiname credit portfolios; see Refs 38–40.

Credit risk has very far-reaching impacts on the financial market, not only within the United States but also around the globe. After the 2008 financial crisis, credit risk has become a crucial part of *systemic risk* and has motivated a host of recent research. For a collection of recent works and details on the subject, we refer to Ref. 41.

REFERENCES

1. Merton R. On the pricing of corporate debt: the risk structure of interest rates. *J Finance* 1974;29:449–470.
2. Black F, Cox JC. Valuing corporate securities: some effects of bond indenture provisions. *J Finance* 1976;31:351–367.
3. Musiela M, Rutkowski M. Martingale methods in financial modeling, Volume 36, Applications of mathematics. Berlin Heidelberg: Springer-Verlag; 1997.
4. Wilmott P, Howison S, Dewynne J. Mathematics of financial derivatives: a student introduction. Cambridge, UK: Cambridge University Press; 1996.
5. Eom Y, Helwege J, Huang JZ. Structural models of corporate bond pricing: an empirical analysis. *Rev Financ Stud* 2004;17(2):499–544.
6. Zhou C. The term structure of credit spreads with jump risk. *J Bank Finance* 2001;25:2015–2040.
7. Boyarchenko S. Endogenous default under Lévy processes, Working paper. University of Texas at Austin; 2004.
8. Hilberink B, Rogers C. Optimal capital structure and endogenous default. *Finance Stoch* 2002;6:237–263.
9. Longstaff F, Schwartz E. Valuing risky debt: a new approach. *J Finance* 1995;50:789–821.
10. Fouque J-P, Sircar R, Sølna K. Stochastic volatility effects on defaultable bonds. *Appl Math Finance* 2006;13(3):215–244.
11. Duffie D, Lando D. The term structure of credit spreads with incomplete accounting information. *Econometrica* 2001;69:633–664.
12. Duffie D, Singleton K. Credit risk. Princeton, New Jersey, USA: Princeton University Press; 2003.
13. Collin-Dufresne P, Goldstein R. Do credit spreads reflect stationary leverage ratios? Reconciling structural and reduced form frameworks. *J Finance* 2001;56:1929–1958.
14. Giesecke K. Credit risk modeling and valuation: an introduction. In: Shimko D, editor. Volume 2, Credit risk: models and management. London, UK: RISK Books; 2004.
15. Cox J, Ingersoll J, Ross S. A theory of the term structure of interest rates. *Econometrica* 1985;53(2):385–407.
16. Feller W. Two singular diffusion problems. *Ann Math* 1951;54:173–182.
17. Duffie D. Dynamic asset pricing theory. 3rd ed. Princeton, New Jersey, USA: Princeton University Press; 2001.
18. Lando D. Credit risk modeling: theory and applications, Princeton series in finance. Princeton, New Jersey, USA: Princeton University Press; 2004.
19. Davis MHA, Lo V. Modelling default correlation in bonds portfolios. *Quant Finance* 2001;1(4):382–387.
20. Elizalde A. Credit risk models IV: Understanding and pricing CDOs; 2006. Working Papers no. 608, CEMFI, Madrid, Spain. Available at: http://EconPapers.repec.org/RePEc:cmf:wpaper:wp2006_0608.
21. Mortensen A. Essays on the pricing of bonds and credit derivatives [PhD thesis]. Copenhagen Business School; 2005.
22. Duffie D, Gârleanu N. Risk and valuation of collateralized debt obligations. *Financ Anal J* 2001;57(1):41–59.
23. Mortensen A. Semi-analytical valuation of basket credit derivatives in intensity-based models. *J Deriv* 2006;13(4):8–26.
24. Vasicek O. Probability of loss on a loan portfolio. Technical report no, 999-0000-056, San Francisco, California USA: KMV; 1987.
25. Credit Suisse First Boston. CreditRisk⁺: a credit risk management framework. Technical report. London UK: Credit Suisse First Boston; 1997.

26. Gupton G, Finger C, Bhatia M. CreditMetrics. Technical report. New York USA: RiskMetrics Group; 1997.
27. Eckner A. Computational techniques for basic affine models of portfolio credit risk. *J Comput Finance* 2007;1363–97.
28. Papageorgiou E, Sircar R. Multiscale intensity models and name grouping for valuation of multi-name credit derivatives. *Appl Math Finance* 2009;16(4):353–383.
29. Giesecke K, Goldberg L. A top down approach to multi-name credit. *Oper Res* 2011;59(2):283–300.
30. Errais E, Giesecke K, Goldberg L. Affine point processes and portfolio credit risk. *SIAM J Financ Math* 2010;1:642–665.
31. Arnsdorff M, Halperin I. BSLP: Markovian bivariate spread-loss model for portfolio credit derivatives. *J Comput Finance* 2008;12:77–100.
32. Lopatin A, Misirpashaev T. Two-dimensional Markovian model for dynamics of aggregate credit loss. In: Fomby T, Fouque J-P, Solna K, editors. *Advances in econometrics*, Volume 22. Bingley, UK: Elsevier Science; 2008, p 243–274.
33. Sidenius J, Piterbarg V, Andersen L. A new framework for dynamic credit portfolio loss modelling. *Int J Theor Appl Finance* 2008;11(2):163–197.
34. Schönbucher P. Portfolio losses and the term structure of loss transition rates: A new methodology for the pricing of portfolio credit derivatives. Technical report no. 264 Zurich, Switzerland: ETH Zurich; 2006.
35. Carmona R. HJM: a unified approach to dynamic models for fixed income, credit and equity markets. *Paris-princeton lectures in mathematical finance*, Volume 1919, *Lecture notes in mathematics*. New York: Springer-Verlag; 2007. p 3–45.
36. Brigo D, Chourdakis K. Counterparty risk for credit default swaps: Impact of spread volatility and default correlation. *Int J Theor Appl Finance* 2009;12(7):1007–1026.
37. Hull J, White A. CVA and wrong way risk. *Financ Anal J* 2012, 68 (5).
38. Leung T, Sircar R, Zariphopoulou T. Credit derivatives and risk aversion. In: Fomby T, Fouque J-P, Solna K, editors. *Advances in econometrics*, Volume 22. Bingley, UK: Elsevier Science; 2008. p 275–291.
39. Jaimungal S, Sigloch G. Incorporating risk and ambiguity aversion into a hybrid model of default. *Math Finance* 2012;22(1):57–81.
40. Sircar R, Zariphopoulou T. Utility valuation of multiname credit derivatives and application to CDOs. *Quant Finance* 2010;10(2):195–208.
41. Fouque J-P, Langsam J, editors. *Handbook of systemic risk*. Cambridge, United Kingdom: Cambridge University Press; 2012.

CROATIAN OPERATIONAL RESEARCH SOCIETY

ZORAN BABIĆ
University of Split, Faculty of
Economics, Split, Croatia

LUKA NERALIĆ
University of Zagreb, Faculty of
Economics and Business, Zagreb,
Croatia

The war against Croatia, which started in 1991, helped the formation of the national OR society of Croatia. Also, organization of the First OR Conference KOI '91 in December 1991 was important in establishing the Croatian Operational Research Society (CRORS). Two groups of OR people, one from Zagreb (Lj. Martić, V. Čerić, M. Mauher, L. Neralić, I. Gjenero, T. Vujković, I. Jemrić, V. Bahovec, Ž. Požgaj, B. Šego, V. Vojvodić-Rosenzweig, D. Kalpić, and others) and the other from Split (Z. Babić and his group), were active in organizing the conference and establishing CRORS, together with colleagues from Varaždin (T. Hunjak and others,) and Rijeka (M. Marinović and others). Some of the Croatian OR people from abroad (S. Zlobec, J. Skorin-Kapov, D. Skorin-Kapov, and others) and colleagues from Slovenia (S. Indihar, L. Zadnik Stirn, M. Bastič, V. Rupnik, L. Arih, and others) also lent their support in establishing CRORS.

In September 1991 a group of OR people from Croatia, because of the war against Croatia, decided not to take part in the Yugoslav Symposium SYM-OP-IS '91 on OR and to withdraw the papers submitted for presentation at the symposium. After that, due to the initiative taken by the colleagues from Split, the OR people from the Faculty of Economics, University of Zagreb, started to organize the First OR Conference KOI '91 in Zagreb. Members of the program committee were V. Čerić, I. Gjenero, Lj. Martić, L. Neralić, and T. Vujković, and members of the organizing committee were

I. Jemrić, V. Bahovec, Ž. Požgaj, B. Šego, and V. Vojvodić-Rosenzweig. In this manner, OR people from Croatia and Slovenia were given the opportunity to present their papers and recent research results. At the end of October 1991, a call for papers was sent and the First OR Conference KOI '91 was held on December 21, 1991 at the Faculty of Economics, Zagreb. There were more than 60 participants at the conference. They were from Zagreb, Split, Rijeka, Varaždin, Ljubljana, and Maribor. In the proceedings of the conference, edited by Lj. Martić and L. Neralić, 28 scientific OR papers were published. A round table was organized at the conference and one of the conclusions was that the OR Society of Croatia should be established. On March 21, 1992 CRORS was established at the meeting held at the Faculty of Economics Zagreb. There were about 60 participants: mathematicians, economists and engineers from Croatia. At the meeting they accepted the documents necessary for the formation of the society and an executive committee was elected. The first president was L. Neralić, vice president V. Vojvodić-Rosenzweig, secretary K. Šorić, and treasurer I. Jemrić.

The formation of the CRORS was primarily an academia-led initiative. Later, some people from industry and other sectors joined the Society.

In the early years, one of the activities was organizing lectures on theoretical and applied OR topics for members of CRORS and members of the Seminary for Mathematical Programming and Game Theory at the Department of Mathematics, University of Zagreb. The other activity was organizing the national OR conferences, which were also open to participants from abroad, every year. In this manner the second (1992), third (1993), fourth (1994), and fifth (1995) OR conferences were organized. In 1996, the Sixth International OR Conference was organized. At that conference it was agreed with the Slovenian OR group, which was

co-organizer of OR Symposiums every year, that international meetings will be organized every two years.

In 1994, CRORS became the 43rd member of IFORS and a member of EURO.

In 1996 in Prague, at the meeting of representatives of the Austrian, Croatian, Czech, and Slovak OR Societies, it was accepted that CRORS joins these societies (including the Hungarian OR Society) in editing the Central European Journal for Operations Research and Economics (CEJORE). Later the journal was named the Central European Journal for Operations Research (CEJOR) and up to now, CRORS supports it as one of the publishing societies.

In April 24–26 1997, CRORS organized the 20th meeting of the EURO Working Group on Financial Modeling at the Interuniversity Center in Dubrovnik, Croatia, with about 50 participants. The organizers of the meeting were L. Neralić, M. Mauher, J. Skorin-Kapov, D. Skorin-Kapov, and Z. Sikora with the help of Jaap Spronk.

The Sixth International Conference on Parametric Optimization and Related Topics PARAOPT VI was organized by CRORS at the Hotel Excelsior in Dubrovnik, Croatia, in October 4–8, 1999. The main organizers of the conference were J. Guddat, H. Th. Jongen, R. Hirabayashi, S. Zlobec, and L. Neralić (Chairman of the Organizing Committee). There were more than 60 participants from Europe, United States, and Canada. Among them there was a group of graduate students from Croatia. The invited speakers for plenary lectures were H. Maurer, B. Mordukovich, P. Kall, A. P. Wierbicky, and L. M. Seiford. For the tutorial lectures the invited speakers were S. Zlobec, J. Guddat, H. Th. Jongen, R. E. Wendell, H. Peters, M. Kress, and J. Dupačova. Some presented papers were published in the special issue of the journal "Optimization" (Vol. 49, No.4, 2001) with guest editors J. Guddat, H. Th. Jongen, L. Neralić, and S. Zlobec.

CRORS organized the 21st Conference on the European Chapter on Combinatorial Optimization ECCO XXI at the Faculty of Economics in Dubrovnik, Croatia (May 29–31, 2008). At the conference there were more than 100

registered participants from all over the world. The members of the scientific committee were V. Boljunčić (co-chair), J. Skorin-Kapov (co-chair), S. Martello, J. Blazewicz, Van-Dat Cung, A. Hertz, L. Neralić, D. Skorin-Kapov, and P. Thoth, and the members of the organizing committee were K. Šorić (Chair), R. Manger, and V. Boljunčić. Plenary speakers were S. Martello, C. Roucairol, M. Laguna, D. Skorin-Kapov, A. Levin, and J. Edmonds.

Now there are about a hundred active members in CRORS who are mainly professors and assistants from the Croatian faculties, and most of them belong to the department of economics. Most of them are mathematicians by their formal education, but there are also a number of people who are economists or engineers. There are still not many active members from industry, banking, insurance, and other sectors: this seems to be a problem that is not easy to solve, but steps will be taken to change this. Nowadays the Faculty of Economics at Zagreb does not have the two year Master's degree study in OR as it used to. According to the Bologna agreement on education, there is a one-year postgraduate program on "OR and Optimization" for university specialists in OR and optimization. In the last two years, there were not enough students interested in it but in the academic year 2009/2010 the first group of six students started this study. There is no Ph.D. study in OR in Croatia and it is not known when and how it will be established. CRORS expects to get support from the Ministry of Science, Education, and Sport of the Republic of Croatia to continue to be among the OR societies that contribute to publishing CEJOR.

The first president of CRORS, as stated before, was Professor Luka Neralić from the Faculty of Economics and Business, Zagreb. After him, the president was Professor Hunjak from the Faculty of Organization and Informatics, Varaždin, who was succeeded by Professor Kristina Šorić again from the Faculty of Economics and Business, Zagreb. The last two presidents were from the coastal part of the country. The first of the two was Professor Valter Boljunčić from the youngest Croatian university, Juraj Dobrila University

of Pula, from the Department of Economics and Tourism “Dr. Mijo Mirković.” Now the president is again from a town located on the

Adriatic coast: Professor Zoran Babić from the Faculty of Economics in Split.

CROSS-ENTROPY METHOD

DIRK P. KROESE
School of Mathematics and
Physics, The University of
Queensland, Brisbane,
Australia

The *cross-entropy (CE) method* is a recent generic Monte Carlo technique for solving complicated simulation and optimization problems. The approach was introduced by Rubinstein in [1,2], extending his earlier work on variance minimization methods for rare-event probability estimation [3].

The CE method can be applied to two types of problems:

1. *Estimation.* Estimate $\ell = \mathbb{E}[H(\mathbf{X})]$, where $\mathbf{X}$ is a random variable or vector taking values in some set $\mathcal{X}$ and H is function on $\mathcal{X}$. An important special case is the estimation of a probability $\ell = \mathbb{P}(S(\mathbf{X}) \geq \gamma)$, where S is another function on $\mathcal{X}$.
2. *Optimization.* Optimize (i.e., maximize or minimize) $S(\mathbf{x})$ over all $\mathbf{x} \in \mathcal{X}$, where S is some objective function on $\mathcal{X}$. S can be either a known or a *noisy* function. In the latter case, the objective function needs to be estimated, for example, via simulation.

In the estimation setting, the CE method can be viewed as an adaptive *importance sampling* procedure that uses the *CE or Kullback–Leibler divergence* as a measure of closeness between two sampling distributions. In the optimization setting, the optimization problem is first translated into a rare-event estimation problem, and then the CE method for estimation is used as an adaptive algorithm to locate the optimum.

An easy tutorial on the CE method is given in [4]. A more comprehensive treatment can be found in [5]; see also [6], Chapter 8. The CE method homepage can be found at www.cemethod.org.

The CE method has been successfully applied to a diverse range of estimation and optimization problems, including buffer allocation [7], queueing models of telecommunication systems [8,9], optimal control of HIV/AIDS spread [10,11], signal detection [12], combinatorial auctions [13], DNA sequence alignment [14,15], scheduling and vehicle routing [16–21], neural and reinforcement learning [22–26], project management [27], rare-event simulation with light- and heavy-tail distributions [28–31], and clustering analysis [32–34]. Applications to classical combinatorial optimization problems including the max-cut, traveling salesman, and Hamiltonian cycle problems are given in [3] and [35–37]. Various CE estimation and noisy optimization problems for reliability systems and network design can be found in [38–45]. Parallel implementations of the CE method are discussed in [46] and [47], and recent generalizations and advances are explored in [48].

THE CROSS-ENTROPY METHOD FOR ESTIMATION

Consider the estimation of

$$\ell = \mathbb{E}_f H(\mathbf{X}) = \int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \quad (1)$$

where H is the *sample performance function* and f is the probability density of the random variable (vector) $\mathbf{X}$. (For notational convenience it is assumed that $\mathbf{X}$ is a continuous random variable; if $\mathbf{X}$ is a discrete random variable, simply replace the integral in (1) by a sum.) Let g be another probability density such that for all $\mathbf{x}$, $g(\mathbf{x}) = 0$ implies that $H(\mathbf{x})f(\mathbf{x}) = 0$. Using the probability density g , we can represent ℓ as

$$\ell = \int H(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = \mathbb{E}_g \left[H(\mathbf{X}) \frac{f(\mathbf{X})}{g(\mathbf{X})} \right]. \quad (2)$$

Consequently, if $\mathbf{X}_1, \dots, \mathbf{X}_N$ are independent random vectors, each with probability density g , then

$$\hat{\ell} = \frac{1}{N} \sum_{k=1}^N H(\mathbf{X}_k) \frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)} \quad (3)$$

is an unbiased estimator of ℓ . Such an estimator is called an *importance sampling estimator*. The optimal importance sampling probability density is given by $g^*(\mathbf{x}) \propto |H(\mathbf{x})|f(\mathbf{x})$ [6, p. 132], which in general is difficult to obtain. The idea of the CE method is to choose the importance sampling density g in a specified class of densities such that the *CE* or *Kullback–Leibler divergence* between the optimal importance sampling density g^* and g is minimal. The Kullback–Leibler divergence between two probability densities g and h is given by

$$\begin{aligned} \mathcal{D}(g, h) &= \mathbb{E}_g \left[\ln \frac{g(\mathbf{X})}{h(\mathbf{X})} \right] = \int g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{h(\mathbf{x})} d\mathbf{x} \\ &= \int g(\mathbf{x}) \ln g(\mathbf{x}) d\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (4)$$

In most cases of interest, the sample performance function H is nonnegative, and the “nominal” probability density f is parameterized by a finite-dimensional vector $\mathbf{u}$; that is, $f(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})$. It is then customary to choose the importance sampling probability density g in the *same* family of probability densities; thus, $g(\mathbf{x}) = f(\mathbf{x}; \mathbf{v})$ for some *reference parameter* $\mathbf{v}$. The CE minimization procedure then reduces to finding an optimal reference parameter vector, $\mathbf{v}^*$. This $\mathbf{v}^*$ turns out to be the solution to the maximization problem $\max_{\mathbf{v}} \int H(\mathbf{x}) f(\mathbf{x}; \mathbf{u}) \ln f(\mathbf{x}; \mathbf{v}) d\mathbf{x}$, which in turn can be estimated via simulation by solving with respect to $\mathbf{v}$, the stochastic counterpart program

$$\max_{\mathbf{v}} \frac{1}{N} \sum_{k=1}^N H(\mathbf{X}_k) \frac{f(\mathbf{X}_k; \mathbf{u})}{f(\mathbf{X}_k; \mathbf{v})} \ln f(\mathbf{X}_k; \mathbf{v}), \quad (5)$$

where $\mathbf{X}_1, \dots, \mathbf{X}_N$ is a random sample from $f(\cdot; \mathbf{w})$, for any reference parameter $\mathbf{w}$. The maximization (5) can often be solved *analytically*, in particular, when the class of

sampling distributions forms an *exponential family* [6, pp. 319–320]. Indeed, analytical updating formulas can be found whenever explicit expressions for the *maximal likelihood estimators* of the parameters can be found [4, p. 36].

Often $\ell = \mathbb{P}(S(\mathbf{X}) \geq \gamma)$, for some performance function S and level γ , in which case $H(\mathbf{x})$ takes the form of an *indicator function*: $H(\mathbf{x}) = I_{\{S(\mathbf{x}) \geq \gamma\}}$; that is, $H(\mathbf{x}) = 1$ if $S(\mathbf{x}) \geq \gamma$, and 0 otherwise. A complication in solving (Eq. 5) occurs when ℓ is a *rare-event probability*; that is, a very small probability (say less than 10^{-5}). Then, for moderate sample size N most or all of the values $H(\mathbf{X}_k)$ in (5) are zero, and the maximization problem becomes useless. In that case a *multilevel* CE procedure is used, where a sequence of reference parameters and levels is constructed with the goal that the first converges to $\mathbf{v}^*$ and the second to γ . This leads to the following algorithm [6, p. 238].

Algorithm 1 [CE Algorithm for Rare-Event Estimation].

1. Define $\hat{\mathbf{v}}_0 = \mathbf{u}$. Let $N^e = \lceil \rho N \rceil$. Set $t = 1$ (iteration counter).
2. Generate a random sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ according to the probability density $f(\cdot; \hat{\mathbf{v}}_{t-1})$. Calculate the performances $S(\mathbf{X}_i)$ for all i , and order them from smallest to largest, $S_{(1)} \leq \dots \leq S_{(N)}$. Let $\hat{\gamma}_t$ be the sample $(1 - \rho)$ -quantile of performances; that is, $\hat{\gamma}_t = S_{(N - N^e + 1)}$. If $\hat{\gamma}_t > \gamma$, reset $\hat{\gamma}_t$ to γ .
3. Use the **same** sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ to solve the stochastic program (5), with $\mathbf{w} = \hat{\mathbf{v}}_{t-1}$. Denote the solution by $\hat{\mathbf{v}}_t$.
4. If $\hat{\gamma}_t < \gamma$, set $t = t + 1$ and reiterate from Step 2; otherwise, proceed with Step 5.
5. Let T be the final iteration counter. Generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_{N_1}$ according to the probability density $f(\cdot; \hat{\mathbf{v}}_T)$ and estimate ℓ via importance sampling, as in (3).

Apart from specifying the family of sampling probability densities, the initial vector $\hat{\mathbf{v}}_0$, the sample size N , and the rarity

parameter ρ (typically between 0.01 and 0.1), the algorithm is completely self-tuning. The sample size N for determining a good reference parameter can usually be chosen to be much smaller than the sample size N_1 for the final importance sampling estimation, say $N = 1000$ versus $N_1 = 100,000$. Under certain technical conditions, the deterministic version of Algorithm 1 is guaranteed to terminate (reach level γ) provided that ρ is chosen small enough [5, Section 3.5].

THE CROSS-ENTROPY METHOD FOR OPTIMIZATION

Let $\mathcal{X}$ be an arbitrary set of *states* and let S be a real-valued performance function on $\mathcal{X}$. Suppose we wish to find the maximum of S over $\mathcal{X}$, and the corresponding state $\mathbf{x}^*$ at which this maximum is attained (assuming for simplicity that there is only one such state). Denote the maximum by γ^* , we thus have

$$S(\mathbf{x}^*) = \gamma^* = \max_{\mathbf{x} \in \mathcal{X}} S(\mathbf{x}). \quad (6)$$

This setting includes many types of optimization problems: discrete (combinatorial), continuous, mixed, and constrained problems. Moreover, if one is interested in minimizing rather than maximizing S , one can simply maximize $-S$.

Now associate with the above problem, the estimation of the probability $\ell = \mathbb{P}(S(\mathbf{X}) \geq \gamma)$, where $\mathbf{X}$ has some probability density $f(\mathbf{x}; \mathbf{u})$ on $\mathcal{X}$ (e.g., corresponding to the uniform distribution on $\mathcal{X}$) and γ is some level. If γ is chosen close to the unknown γ^* , then ℓ is, typically, a rare-event probability, and the CE approach of the previous section can be used to find an importance sampling distribution close to the theoretically optimal importance sampling density, which concentrates all its mass on the point $\mathbf{x}^*$. Sampling from such a distribution thus produces optimal or near-optimal states. A main difference with the CE method for rare-event simulation is that in the optimization setting the final level $\gamma = \gamma^*$ is not known in advance. The CE method for optimization produces a sequence of levels $\{\hat{\gamma}_t\}$ and reference parameters $\{\hat{\mathbf{v}}_t\}$ such that

the former tends to the optimal γ^* and the latter to the optimal reference vector $\mathbf{v}^*$ corresponding to the point mass at $\mathbf{x}^*$ [6, p. 251].

Algorithm 2 [CE Algorithm for Optimization].

1. Choose an initial parameter vector $\hat{\mathbf{v}}_0$. Let $N^e = \lceil \rho N \rceil$. Set $t = 1$ (level counter).
2. Generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ from the probability density $f(\cdot; \hat{\mathbf{v}}_{t-1})$. Calculate the performances $S(\mathbf{X}_i)$ for all i , and order them from smallest to largest, $S_{(1)} \leq \dots \leq S_{(N)}$. Let $\hat{\gamma}_t$ be the sample $(1 - \rho)$ -quantile of performances; that is, $\hat{\gamma}_t = S_{(N - N^e + 1)}$.
3. Use the **same** sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ and solve the stochastic program

$$\max_{\mathbf{v}} N^{-1} \sum_{k=1}^N I_{\{S(\mathbf{X}_k) \geq \hat{\gamma}_t\}} \ln f(\mathbf{X}_k; \mathbf{v}). \quad (7)$$

Denote the solution by $\hat{\mathbf{v}}_t$.

4. If the stopping criterion is met, stop; otherwise, set $t = t + 1$, and return to Step 2.

To run the algorithm, one needs to provide the class of sampling probability densities, the initial vector $\hat{\mathbf{v}}_0$, the sample size N , the rarity parameter ρ , and the stopping criterion. Any CE algorithm for optimization, thus involves the following two main iterative phases:

1. *Generate* a random sample of objects in the search space $\mathcal{X}$ (trajectories, vectors, etc.) according to a specified probability distribution.
2. *Update* the parameters of that distribution, based on the N^e best performing samples (the so-called *elite samples*), using CE minimization.

Apart from the fact that Step 3 in Algorithm 1 is missing in Algorithm 2, another main difference between the two algorithms is that the *likelihood ratio* term

$f(\mathbf{X}_k; \mathbf{u})/f(\mathbf{X}_k; \hat{\mathbf{v}}_{t-1})$ in Equation (5) is missing in Equation (7).

Often a smoothed updating rule is used, in which the parameter vector $\hat{\mathbf{v}}_t$ is taken as

$$\hat{\mathbf{v}}_t = \alpha \tilde{\mathbf{v}}_t + (1 - \alpha) \hat{\mathbf{v}}_{t-1}, \quad (8)$$

where $\tilde{\mathbf{v}}_t$ is the solution to Equation (7) and $0 \leq \alpha \leq 1$ is a smoothing parameter. Many other modifications can be found in [5,6] and [49], and in the list of references. When there are two or more optimal solutions, the CE algorithm typically “fluctuates” between the solutions before focusing in on one of the solutions. The effect that smoothing has on convergence is discussed in detail in [50]. In particular, it is shown that with appropriate smoothing the CE method converges and finds the optimal solution with probability arbitrarily close to 1. Necessary conditions and sufficient conditions under which the optimal solution is generated eventually with probability 1 are also given. Other convergence results, including a proof of convergence along the lines of convergence for simulated annealing can be found in [51].

Combinatorial Optimization

When the state space $\mathcal{X}$ is finite, the optimization problem (6) is often referred to as a *discrete or combinatorial optimization* problem. For example, $\mathcal{X}$ could be the space of combinatorial objects such as binary vectors, trees, paths through graphs, and so on. To apply the CE method, one needs to specify first a convenient parameterized random mechanism to generate objects in $\mathcal{X}$. For example, when $\mathcal{X}$ is the set of binary vectors of length n , an easy generation mechanism is to draw each component independently from a Bernoulli distribution; that is, $\mathbf{X} = (X_1, \dots, X_n) \sim \text{Ber}(\mathbf{p})$, where $\mathbf{p} = (p_1, \dots, p_n)$. Given an elite sample set $\mathcal{E}$, the updating formula is then [4, p. 56]

$$\hat{p}_i = \frac{\sum_{\mathbf{x} \in \mathcal{E}} X_i}{N_e}, \quad i = 1, \dots, n. \quad (9)$$

That is, the updated success probability for the i th component is the mean of the i th components of the vectors in the elite set.

A possible stopping rule for combinatorial optimization problems is to stop when the overall best objective value does not change over a number of iterations. Alternatively, one could stop when the sampling distribution has “degenerated” enough. In particular, in the Bernoulli case (9) one could stop when all $\{p_i\}$ are less than some distance ϵ away from either 0 or 1.

Continuous Optimization

It is also possible to apply the CE algorithm to continuous optimization problems; in particular, when $\mathcal{X} = \mathbb{R}^n$. The sampling distribution on $\mathbb{R}^n$ can be quite arbitrary, and does not need to be related to the function that is being optimized. However, the generation of a random vector $\mathbf{X} = (X_1, \dots, X_n) \in \mathbb{R}^n$ is usually established by drawing the coordinates independently from some two-parameter distribution. In most applications, a normal (Gaussian) distribution is employed for each component. Thus, the sampling distribution for $\mathbf{X}$ is characterized by a vector of means $\boldsymbol{\mu}$ and a vector of standard deviations $\boldsymbol{\sigma}$. At each iteration of the CE algorithm, these parameter vectors are updated simply as the vectors of sample means and sample standard deviations of the elements in the elite set [49]. During the course of the algorithm, the sequence of mean vectors ideally tends to the maximizer $\mathbf{x}^*$, while the vector of standard deviations tend to the zero vector. In short, one should obtain a degenerated probability density with all mass concentrated in the vicinity of the point $\mathbf{x}^*$. A possible stopping criterion is to stop when all standard deviations are smaller than some ϵ .

Constrained Optimization

Constrained optimization problems can be put in the framework (6) by taking $\mathcal{X}$ a (nonlinear) region defined by some system of inequalities:

$$G_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, L. \quad (10)$$

To solve the program (6) with constraints (10), two approaches can be adopted. The first approach uses *acceptance–rejection*: generate a random vector $\mathbf{X}$ from, for example, a

multivariate normal distribution with independent components, and accept or reject it depending on whether the sample falls in $\mathcal{X}$ or not. Alternatively, one could sample directly from a truncated distribution (for example, a truncated normal distribution) or combine such a method with acceptance–rejection. Once a fixed number of such vectors has been accepted, the parameters of the normal distribution can be updated in exactly the same way as in the unconstrained case—simply via the sample mean and standard deviation of the elite samples. A drawback of this method is that a large number of samples could be rejected before a feasible sample is found.

The second approach is the *penalty approach*. Here, the idea is to modify the objective function as follows:

$$\tilde{S}(\mathbf{x}) = S(\mathbf{x}) + \sum_{i=1}^L P_i(\mathbf{x}), \quad (11)$$

where the $\{P_i\}$ are *penalty functions*. Specifically, the i th penalty function P_i (corresponding to the i th constraint) is defined as

$$P_i(\mathbf{x}) = H_i \max(G_i(\mathbf{x}), 0) \quad (12)$$

and $H_i > 0$ measures the importance (cost) of the i th penalty.

Thus, by reducing the constrained problem (6) and (10) to an unconstrained one ((6) with $\tilde{S}$ instead of S), one can again apply Algorithm 2. Further details on constrained multiextremal optimization with the CE method may be found in Ref. 49.

Noisy Optimization

Noisy (or stochastic) optimization problems—in which the objective function is corrupted with noise—arise in many contexts, for example, in stochastic scheduling and stochastic shortest/longest path problems, and simulation-based optimization [52]. The CE method can be easily modified to deal with noisy optimization problems. Consider the maximization problem (6) and assume that the performance function is noisy. In particular, suppose that $S(\mathbf{x}) = \mathbb{E}\hat{S}(\mathbf{x})$ is not available, but that a sample value $\hat{S}(\mathbf{x})$

(unbiased estimate of $\mathbb{E}\hat{S}(\mathbf{x})$) is available, for example, via simulation. The principal modification of the Algorithm 2 is to replace $S(\mathbf{x})$ by $\hat{S}(\mathbf{x})$. In addition, one may need to increase the sample size in order to reduce the effect of the noise. Although various applications indicate the usefulness of the CE approach for noisy optimization [7,40–42], little is still known regarding theoretical convergence results in the noisy case. Spall [53, Section A.0.0] discusses various divergence results for general types of stochastic methods. A possible stopping criterion is to stop when the sampling distribution has degenerated enough. Another possibility is to stop the stochastic process when $\{\hat{y}_t\}$ has reached stationarity [5, p. 207].

REFERENCES

1. Rubinstein RY. The cross-entropy method for combinatorial and continuous optimization. *Methodol Comput Appl Probab* 1999;2:127–190.
2. Rubinstein RY. Combinatorial optimization, cross-entropy, ants and rare events. In: Uryasev S, Pardalos PM, editors. *Stochastic optimization: algorithms and applications*. Dordrecht: Kluwer; 2001. pp. 304–358.
3. Rubinstein RY. Optimization of computer simulation models with rare events. *Eur J Oper Res* 1997;99:89–112.
4. de Boer PT, Kroese DP, Mannor S, *et al.* A tutorial on the cross-entropy method. *Ann Oper Res* 2005;134(1):19–67.
5. Rubinstein RY, Kroese DP. *The cross-entropy method: a unified approach to combinatorial optimization, Monte Carlo simulation and machine learning*. New York: Springer; 2004.
6. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo method*. 2nd ed. John Wiley & Sons, Inc.: New York; 2007.
7. Alon G, Kroese DP, Raviv T, *et al.* Application of the cross-entropy method to the buffer allocation problem in a simulation-based environment. *Ann Oper Res* 2005;134:137–151.
8. de Boer PT. *Analysis and efficient simulation of queueing models of telecommunication systems* [PhD thesis]. University of Twente; 2000.
9. de Boer PT, Kroese DP, Rubinstein RY. A fast cross-entropy method for estimating buffer

- overflows in queueing networks. *Manag Sci* 2004;50(7):883–895.
10. Sani A. Stochastic modelling and intervention of the spread of HIV/AIDS [PhD thesis]. Brisbane: The University of Queensland; 2009.
 11. Sani A, Kroese DP. Controlling the number of HIV infectives in a mobile population. *Math Biosci* 2008;213:103–112.
 12. Liu Z, Doucet A, Singh SS. The cross-entropy method for blind multiuser detection. *IEEE International Symposium on information theory*. Chicago, Piscataway; 2004.
 13. Chan JCC, Kroese DP. Randomized methods for solving the winner determination problem in combinatorial auctions. *Proceedings of the 2008 Winter Simulation Conference*. Miami (FL); 2008.
 14. Keith J, Kroese DP. Sequence alignment by rare event simulation. *Proceedings of the 2002 Winter simulation conference*. San Diego (CA); 2002. pp. 320–327.
 15. Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* 2007;23(13):1607–1615.
 16. Bendavid I, Golany B. Setting gates for activities in the stochastic project scheduling problem through the cross entropy methodology. *Ann Oper Res* 2009. In press.
 17. Caserta M, Quinonez-Ricob E. A cross entropy-Lagrangian hybrid algorithm for the multi-item capacitated lot-sizing problem with setup times. *Comput Oper Res* 2009;36(2):530–548.
 18. Chepuri K, Homem-de-Mello T. Solving the vehicle routing problem with stochastic demands using the cross entropy method. *Ann Oper Res* 2005;134(1):153–181.
 19. Helvik BE, Wittner O. Using the cross-entropy method to guide/govern mobile agent's path finding in networks. In: Pierre S, Glitho R, editors. *Mobile agents for telecommunication applications: third international workshop, MATA 2001*; Montreal. New York: Springer; 2001. pp. 255–268.
 20. Ianovsky E, Kreimer J. An optimal routing policy for unmanned aerial vehicles (analytical and cross-entropy simulation approach). *Ann Oper Res* 2009. In press.
 21. Wang J, Gao X, Shi J, *et al.* Double unmanned aerial vehicle's path planning for scout via cross-entropy method. Volume 2, 8th ACIS International conference on software engineering, artificial intelligence, networking, and parallel/distributed computing. 2007. pp. 632–635.
 22. Lorincza A, Palotaia Z, Szirtesb G. Spike-based cross-entropy method for reconstruction. *Neurocomputing* 2008;71(16–18): 3635–3639.
 23. Mannor S, Rubinstein RY, Gat Y. The cross-entropy method for fast policy search. *The 20th International Conference on machine learning (ICML-2003)*. Washington (DC); 2003.
 24. Menache I, Mannor S, Shimkin N. Basis function adaption in temporal difference reinforcement learning. *Ann Oper Res* 2005;134(1): 215–238.
 25. Unveren A, Acan A. Multi-objective optimization with cross entropy method: Stochastic learning with clustered pareto fronts. *IEEE Congress on Evolutionary Computation (CEC 2007)*; 2007. pp. 3065–3071.
 26. Wu Y, Fyfe C. Topology perserving mappings using cross entropy adaptation. *7th WSEAS international conference on artificial intelligence, knowledge engineering and data bases*. Cambridge; 2008.
 27. Cohen I, Golany B, Shtub A. Managing stochastic finite capacity multi-project systems through the cross-entropy method. *Ann Oper Res* 2005;134(1):183–199.
 28. Asmussen S, Kroese DP, Rubinstein RY. Heavy tails, importance sampling and cross-entropy. *Stoch Mod* 2005;21(1):57–76.
 29. Chan JCC, Kroese DP. Rare-event probability estimation with conditional Monte Carlo. *Ann Oper Res* 2009.
 30. Homem-de-Mello T. A study on the cross-entropy method for rare event probability estimation. *INFORMS J Comput* 2007;19(3):381–394.
 31. Kroese DP, Rubinstein RY. The transform likelihood ratio method for rare event simulation with heavy tails. *Queueing Syst* 2004;46:317–351.
 32. Botev ZI, Kroese DP. Global likelihood optimization via the cross-entropy method, with an application to mixture models. In: Ingalls RG, Rossetti MD, Smith JS, *et al.*, editors. *Proceedings of the 2004 Winter simulation conference*. Washington (DC): The Society for Computer Simulation; 2004. pp. 529–535.
 33. Boubouzoula A, Paris S, Ouladsinea M. Application of the cross entropy method to the GLVQ algorithm. *Pattern Recognit* 2008;41(10):3173–3178.

34. Kroese DP, Rubinstein RY, Taimre T. Application of the cross-entropy method to clustering and vector quantization. *J Glob Optim* 2007;37:137–157.
35. Eshragh A, Filar JA, Haythorpe M. A hybrid simulation-optimization algorithm for the Hamiltonian cycle problem. *Ann Oper Res* 2009. In press.
36. Rubinstein RY. Combinatorial optimization via cross-entropy. In: Gass S, Harris C, editors. *Encyclopedia of operations research and management sciences*. Dordrecht: Kluwer Academic Publishers; 2001. pp. 102–106.
37. Rubinstein RY. The cross-entropy method and rare-events for maximal cut and bipartition problems. *ACM Trans Model Comput Simulation* 2002;12(1):27–53.
38. Caserta M, Cabo-Nodar M. A cross entropy based algorithm for reliability problems. *J Heurist* 2009;15(5):479–501.
39. Hui K-P, Bean N, Kraetzl M, *et al.* The cross-entropy method for network reliability estimation. *Ann Oper Res* 2005;134:101–118.
40. Kroese DP, Hui K-P. Chapter 3: Applications of the cross-entropy method in reliability. *Computational intelligence in reliability engineering*. New York: Springer; 2006.
41. Kroese DP, Nariyai S, Hui K-P. Network reliability optimization via the cross-entropy method. *IEEE Trans Reliab* 2007;56(2):275–287.
42. Nariyai S. Cross-entropy methods in telecommunication systems [PhD thesis]. Brisbane: The University of Queensland; 2009.
43. Nariyai S, Kroese DP. On the design of multi-type networks via the cross-entropy method. *Proceedings of the Fifth International Workshop on the design of reliable communication networks (DRCN)*. Ischia, Italy; 2005. pp. 109–114.
44. Nariyai S, Kroese DP, Hui K-P. Designing an optimal network using the cross-entropy method. *Intelligent data engineering and automated learning. Lecture notes in computer science*. New York: Springer; 2005. pp. 228–233.
45. Ridder A. Importance sampling simulations of Markovian reliability systems using cross-entropy. *Ann Oper Res* 2005;134(1):119–136.
46. Evans GE. Parallel and sequential Monte Carlo methods with applications [PhD thesis]. Brisbane: The University of Queensland; 2009.
47. Keith JM, Evans GE, Kroese DP. Parallel cross-entropy optimization. *Proceedings of the 2007 Winter Simulation Conference*. Washington (DC); 2007. pp. 2196–2202.
48. Taimre T. Advances in cross-entropy methods [PhD thesis]. Brisbane: The University of Queensland; 2009.
49. Kroese DP, Porotsky S, Rubinstein RY. The cross-entropy method for continuous multi-extremal optimization. *Methodol Comput Appl Probab* 2006;8:383–407.
50. Costa A, Owen J, Kroese DP. Convergence properties of the cross-entropy method for discrete optimization. *Oper Res Lett* 2007; 35(5):573–580.
51. Margolin L. On the convergence of the cross-entropy method. *Ann Oper Res* 2005;134(1): 201–214.
52. Rubinstein RY, Melamed B. *Modern simulation and modeling*. New York: John Wiley & Sons Inc.; 1998.
53. Spall JC. *Introduction to stochastic search and optimization: estimation, simulation, and control*. New York: John Wiley & Sons, Inc.; 2003.

CTMCs WITH COSTS AND REWARDS

NATARAJAN GAUTAM
Department of Industrial and
Systems Engineering, Texas
A&M University, College
Station, Texas

Consider a system that can be modeled as a CTMC (continuous-time Markov chain) $\{X(t), t \geq 0\}$ such that $X(t)$ is the state of the system at time t . Note that $X(t)$ could possibly be a vector but for notational convenience we sometimes write $P\{X(t) = j\}$ with the understanding that j in those cases is also a vector. Let S be the state space of this CTMC; that is, the set of all possible values of $X(t)$ for all t . We use the notation Q to denote the infinitesimal generator matrix of the CTMC with negative values along the diagonals such that each row sums to zero. For terminologies and notation used for CTMCs see **Definition and Examples of Continuous-Time Markov Chains and Asymptotic Behavior of Continuous-Time Markov Chains** in this encyclopedia. The reader is also encouraged to consult books by Kulkarni [1] and Ross [2].

For CTMCs with costs and rewards, an additional piece of notation is necessary. Define C_i as the cost incurred by the system per unit time it spends in state i , for all $i \in S$. Notice that if C_i is negative it is equivalent to the system getting a reward (as opposed to cost) of $-C_i$ per unit time it spends in state i . Here, we present only the analysis of systems over an infinite horizon. In particular, we consider two infinite horizon cases: average costs and discounted costs. Each case has its own pros and cons that are subsequently discussed along with some examples. Then we conclude by presenting applications to performance analysis.

AVERAGE COSTS CASE

The objective here is to obtain an expression for the long-run average cost per unit time

incurred by the system. For this, we need some additional assumptions and notation. Recall that we consider a CTMC $\{X(t), t \geq 0\}$ with state space S , infinitesimal generator matrix Q , and costs C_i per unit time the system is in state i for all $i \in S$. We assume that this CTMC is irreducible and positive recurrent (together it is also known as *ergodic*) with steady-state probabilities p_j , that is,

$$p_j = \lim_{t \rightarrow \infty} P\{X(t) = j\}.$$

The p_j values can be computed by solving for $pQ = 0$ and $\sum_{i \in S} p_i = 1$, where p is a row vector of p_j values.

It is important to notice that for ergodic CTMCs, the steady-state probabilities p_j have two other meanings. First, p_j is also the long-run fraction of time the system is in state j , that is,

$$p_j = \lim_{T \rightarrow \infty} \frac{\int_0^T I(X(t) = j) dt}{T}, \quad (1)$$

where $I(\cdot)$ is an indicator function such that $I(X(t) = j) = 1$ if $X(t)$ is j at time t and $I(X(t) = j) = 0$ otherwise. Another definition of p_j is that it is the stationary probability that the CTMC is in state j . That means if the initial state of the CTMC is selected according to p_j , that is, $P\{X(0) = j\} = p_j$, then at any time t the CTMC would be in state j with probability p_j , that is, $P\{X(t) = j\} = p_j$ for all $t \geq 0$. Thus the system is considered to be stationary at all $t \geq 0$. Now we are in a position to develop an expression for the long-run average cost per unit time for this system.

Let C be the long-run average cost incurred by the system per unit time. By definition, since C is time averaged, it can be written as

$$C = \lim_{T \rightarrow \infty} \frac{\int_0^T C_{X(t)} dt}{T}.$$

We can rewrite the right-hand side of the above expression using a sum to get

$$C = \lim_{T \rightarrow \infty} \frac{\int_0^T \sum_{i \in S} C_i I(X(t) = i) dt}{T}.$$

Since the CTMC is positive recurrent, we are allowed to write the above expression as

$$C = \sum_{i \in S} \lim_{T \rightarrow \infty} \frac{\int_0^T C_i I(X(t) = i) dt}{T}.$$

Using the definition of p_i in Equation (1) and substituting, we get

$$C = \sum_{i \in S} C_i p_i. \quad (2)$$

DISCOUNTED COSTS CASE

Here, we consider the case where the costs incurred by the system are discounted at rate α . By definition, this means that if the system incurs a cost \$ \$ \$d\$ at time \$t\$, its present value at time 0 is \$de^{-\alpha t}\$. In other words, it is as though a cost of \$de^{-\alpha t}\$ is incurred at time 0. We call \$\alpha\$ as the discount factor and it is positive in sign. Having defined the discounted cost, we are ready to develop an expression for the expected total discounted cost over an infinite horizon. Let \$D_c\$ be the total discounted cost (note this is a random variable) over an infinite horizon incurred by the system. Therefore, we have by definition

$$D_c = \int_0^\infty C_{X(t)} e^{-\alpha t} dt.$$

Our objective is to obtain \$E[D_c]\$. For this, we require two notations. Let \$a_i\$ be the initial probability that the CTMC is in state \$i\$, that is,

$$a_i = P\{X(0) = i\}.$$

Notice that if the initial state of the CTMC is known, say \$j\$, then \$a_j = 1\$ and all other \$a_i\$'s would be 0. Also let \$\bar{d}_i\$ denote the expected

total discounted cost incurred over the infinite horizon starting in state \$i\$, that is,

$$d_i = E[D_c \mid X(0) = i].$$

Thus, we have

$$\begin{aligned} E[D_c] &= \sum_{i \in S} E[D_c \mid X(0) = i] P\{X(0) = i\} \\ &= \sum_{i \in S} d_i a_i. \end{aligned}$$

Note that \$E[D_c]\$ is dependent on the initial conditions via the \$a_i\$ values. So, we need to be given \$a_i\$; assuming we have that, next we show how to compute \$d_i\$ for all \$i \in S\$. On the basis of the definition we have

$$\begin{aligned} d_i &= E[D_c \mid X(0) = i] \\ &= E \left[\int_0^\infty C_{X(t)} e^{-\alpha t} dt \mid X(0) = i \right] \\ &= \int_0^\infty e^{-\alpha t} \sum_{j \in S} C_j P\{X(t) = j \mid X(0) = i\} dt \end{aligned}$$

provided \$C_j\$ is bounded. Using the column vector \$d\$ with elements \$d_i\$ such that \$d = [d_i]\$, we can write down a matrix relation

$$d = \left[\int_0^\infty e^{-\alpha t} P(t) dt \right] \bar{C},$$

where \$\bar{C}\$ is a column vector of \$C_j\$ values and \$P(t)\$ is a matrix of \$P\{X(t) = j \mid X(0) = i\}\$ for all \$i \in S\$ and \$j \in S\$. Using the Laplace–Stieltjes transform (LST) notation for \$P(t)\$, we get

$$\tilde{P}(s) = \int_0^\infty e^{-st} P(t) dt.$$

Thus,

$$d = \tilde{P}(\alpha) \bar{C}.$$

However, since \$P(t)\$ satisfies \$\frac{dP(t)}{dt} = P(t)Q\$, by taking the LST, we get \$s\tilde{P}(s) - P(0) = \tilde{P}(s)Q\$. Using the fact that \$P(0) = I\$ and solving for \$\tilde{P}(s)\$, we get

$$\tilde{P}(s) = (sI - Q)^{-1}.$$

Therefore, we have

$$d = (\alpha I - Q)^{-1} \bar{C}.$$

Notice that the matrix $\alpha I - Q$ is invertible for any $\alpha > 0$. Thus, we can compute d_i , and hence $E[D_c]$.

DISCUSSION AND EXAMPLES

It is worthwhile to contrast the discounted costs against the average costs. In the discounted cost case, the CTMC does not have to be ergodic (it could be reducible or irreducible, transient or recurrent). However, the average cost results presented require the CTMC to be ergodic, although there are ways to deal with some nonergodic cases. One of the demerits of the discounted cost case is that the results depend on the initial conditions, unlike the average cost case. Also, the discount factor in many practical situations are hard to determine. Next, we present one example for discount cost and one for average cost.

Example 1. Consider a machine that toggles between being under working condition and under repair. The machine stays working for an exponential amount of time with mean $1/\gamma$ h and then breaks down. A repair person arrives instantaneously and spends an exponential amount of time with mean $1/\beta$ h to repair the machine. A revenue of \$ r per hour is obtained when the machine is working and the repair person charges \$ b per hour.

Let $X(t)$ be the state of the machine at time t . Therefore, if $X(t) = 0$, then the machine is under repair at time t . Also, if $X(t) = 1$, the machine is up at time t . The state space $S = \{0, 1\}$. The generator matrix is

$$Q = \begin{bmatrix} -\beta & \beta \\ \gamma & -\gamma \end{bmatrix}.$$

Consider the discount cost case with discount factor α and given initial probabilities $a_0 = P\{X(0) = 0\}$ and $a_1 = P\{X(0) = 1\} = 1 - a_0$. In fact, we could assume that at time 0 the machine could be up or down (i.e., a_0 is 0 or 1, respectively). We seek to obtain an expression for $E[D_c]$, the expected total discounted cost over an infinite horizon incurred

by the system in terms of α , a_0 , a_1 , b , r , β , and γ .

Using the discounted cost results, we have

$$\begin{aligned} E[D_c] &= E[D_c | X(0) = 0]P\{X(0) = 0\} \\ &\quad + E[D_c | X(0) = 1]P\{X(0) = 1\} \\ &= d_0 a_0 + d_1 a_1, \end{aligned}$$

where

$$\begin{aligned} \begin{bmatrix} d_0 \\ d_1 \end{bmatrix} &= (\alpha I - Q)^{-1} \begin{bmatrix} b \\ -r \end{bmatrix} \\ &= \begin{bmatrix} \frac{b(\alpha + \gamma) - r\beta}{\alpha(\alpha + \gamma + \beta)} \\ \frac{b\gamma - r(\alpha + \beta)}{\alpha(\alpha + \gamma + \beta)} \end{bmatrix}. \end{aligned}$$

Notice that the notations used in the discounted costs derivation d and $\bar{C}$ are $d = [d_0 \ d_1]'$ and $\bar{C} = [b \ -r]'$. The expected total discounted cost over an infinite horizon incurred by the system is therefore

$$\begin{aligned} E[D_c] &= d_0 a_0 + d_1 a_1 \\ &= \frac{b\gamma - r\beta + (ba_0 - ra_1)\alpha}{\alpha(\alpha + \gamma + \beta)}. \end{aligned}$$

Example 2. Consider two independent and identical machines each with up times and repair times as described in Example 1. There is only one repair person who repairs the machines in the order they failed.

For this example, we seek to obtain the long-run average cost per hour incurred by the system. Recall that this is denoted by C which can be obtained using Equation (2). Let $X(t)$ be the number of working machines at time t . The state space $S = \{0, 1, 2\}$. The generator matrix is

$$Q = \begin{bmatrix} -\beta & \beta & 0 \\ \gamma & -\beta - \gamma & \beta \\ 0 & 2\gamma & -2\gamma \end{bmatrix}.$$

The steady-state probabilities p_0 , p_1 , and p_2 can be obtained by solving for $[p_0 \ p_1 \ p_2]$ $Q = [0 \ 0 \ 0]$ and $p_0 + p_1 + p_2 = 1$ as $p_0 = \frac{2\gamma^2}{2\gamma^2 + 2\gamma\beta + \beta^2}$, $p_1 = \frac{2\gamma\beta}{2\gamma^2 + 2\gamma\beta + \beta^2}$, and $p_2 =$

$\frac{\beta^2}{2\gamma^2 + 2\gamma\beta + \beta^2}$. Also, notice that $C_0 = b$, $C_1 = b - r$, and $C_2 = -2r$. Therefore, the long-run average cost per hour incurred by the system using Equation (2) is

$$C = \frac{2(\gamma b - \beta r)(\gamma + \beta)}{2\gamma^2 + 2\gamma\beta + \beta^2}.$$

PERFORMANCE EVALUATION

Here, we consider only the average costs case. An important realization to make is that the costs C_i do not have to be dollar costs. They are just the rates at which something is incurred when the source is in state i or even any time-averaged measure that depends on the sojourn time in state i . This is extremely useful in performance analysis and evaluation of systems. We illustrate this using examples.

Example 3. Consider a telephone switch that can handle at most N calls simultaneously. Assume that calls arrive according to a Poisson process with parameter λ to the switch. Any call arriving when there are N other calls in progress receives a busy signal (and hence rejected). Each accepted call lasts for an exponential amount of time with mean $1/\mu$ amount of time.

Let $X(t)$ be the number of on-going calls in the switch at time t . Clearly, the state space is $S = \{0, 1, \dots, N\}$. The infinitesimal generator matrix is

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots & 0 \\ \mu & -(\lambda + \mu) & \lambda & 0 & \dots & 0 \\ 0 & 2\mu & -(\lambda + 2\mu) & \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & N\mu & -N\mu \end{bmatrix}.$$

Now, let $(p_0, p_1, \dots, p_N)$ be the solution to $[p_0 \ p_1 \ \dots \ p_N]Q = [0 \ 0 \ \dots \ 0]$ and $p_0 + p_1 + \dots + p_N = 1$. We do not write down the expression for p_i , but assume that this can easily be done by the reader.

Say we are interested in finding out the rate of call rejection. Then, we have $C_N = \lambda$

(since in state N calls are rejected at rate λ per unit time) and $C_i = 0$ for $i = 0, 1, \dots, N - 1$ (since no calls are rejected when there are less than N in system). Therefore, the long-run average number of calls rejected per unit time is $\sum_{i=0}^N C_i p_i = \lambda p_N$. We can also obtain the long-run time-averaged number of lines utilized as $\sum_{i=0}^N i p_i$.

Example 4. The unslotted aloha protocol for multiaccess communication works as follows. Consider a system where messages arrive according to a Poisson process with parameter λ . As soon as a message arrives, it attempts transmission. The message transmission times are exponentially distributed with mean $1/\mu$ units of time. If no other message tries to transmit during the transmission time of this message, the transmission is successful. If any other message tries to transmit during this transmission, a collision results and all transmissions are terminated instantly. All messages involved in a collision are called *backlogged* and are forced to retransmit. All backlogged messages wait for an exponential amount of time (with mean $1/\theta$) before starting retransmission.

Let $X(t)$ denote the number of backlogged messages at time t and $Y(t)$ be a binary variable that denotes whether or not a message is under transmission at time t . Then, the process $\{(X(t), Y(t)), t \geq 0\}$ is a CTMC. Let $p = (p_{00}, p_{01}, p_{10}, p_{11}, p_{20}, p_{21}, \dots)$ be the steady-state probabilities, that is, the solution to $pQ = 0$ and $\sum_{i,j} p_{ij} = 1$. Say, we are interested in obtaining the system throughput, that is, the average number of successful transmissions per unit time. The rate of successful transmission is μ whenever $Y(t)$ is 1 and 0 when $Y(t)$ is 0. Also, when $Y(t)$ is 1, it is not necessary that a transmission would occur, that happens only with probability $\mu/(\mu + i\theta + \lambda)$ when $X(t)$ is i . Therefore, the throughput can be computed as $\sum_{i=0}^{\infty} p_{i1} \mu^2 / (\mu + i\theta + \lambda)$. Interestingly, there is a much simpler way to obtain the throughput. If the system is stable, then in steady state the throughput must indeed be λ , using a conservation of flow argument. To obtain the time-averaged

number of backlogged messages, we can compute $\sum_{i=0}^{\infty} i(p_{i0} + p_{i1})$.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. In: Texts in Statistical Science Series. London: Chapman and Hall, Ltd.; 1995.
2. Ross SM. Introduction to probability models. San Diego (CA): Academic Press; 2003.

CUSTOMER RELATIONSHIP MANAGEMENT: MAXIMIZING CUSTOMER LIFETIME VALUE

MICHAEL LEWIS
Olin Business School Washington,
University in St. Louis,
St. Louis, Missouri

Customer relationship management (CRM) has been defined [1] as the “process of carefully managing detailed information about customers and all customer touch points to maximize customer loyalty.” Successful CRM requires a combination of information technology that enables monitoring of individual customers, appropriate models for predicting customer behavior and a managerial philosophy that views customer relationship as a valuable economic asset [2]. CRM applications are popular in industries such as retailing, financial services, travel, and professional sports. However, while the CRM concept and the associated systems are now common in a wide range of industries, the reality is that an inordinate percentage of CRM implementations are viewed as failures [3]. A main cause of these failures is that firms often struggle with the process of converting customer information into actionable marketing policies.

Perhaps the ultimate level of CRM is individual level or one-to-one marketing [4] where firms create customized marketing policies that maximize the economic value of each customer. Implementation of true individual level marketing that specifies marketing tactics at the level of individual customers in a manner that maximizes profitability requires sophisticated statistical and optimization techniques. Recent review articles have highlighted the need for academic researchers to develop techniques for creating individual level marketing policies that maximize customer value [5,6]. The purpose of this article is to highlight

several key concepts and metrics related to the economic value of customer assets and to discuss how management science techniques may be employed to create marketing policies that maximize the value of customer relationship assets.

CUSTOMER RELATIONSHIP MANAGEMENT

The last 15 years have seen the development of a significant body of academic literature focused on CRM. Early work focused on the development of metrics that focused on the financial value of customer assets [7,8]. This work acknowledged that customers are the ultimate source of a firm’s profits and highlighted the importance of factors such as retention rates and acquisition costs. The academic community has responded with a significant amount of research related to models that predict customer acquisition, retention, and cross-selling activity [9].

For example, customer acquisition has been modeled using binary models [10,11] or by using diffusion modeling techniques [12]. The binary logit models are notable in that they treat customer acquisition as stochastic and as a function of marketing activities. Customer retention models have been formulated in both contractual settings and in categories where purchases tend to occur at irregular intervals. For example, Bolton [13] models customer duration using a hazard modeling approach while Schmittlein *et al.* [14] develop a negative binomial distribution that can be used to predict whether a customer is still alive. In addition to acquisition and retention, customer profit ability can also be influenced by the level of a customer’s purchasing and the number of categories shopped in. Representative models of cross selling have been developed by Li *et al.* [15] for sequential purchases of financial product and Kamakura *et al.* [16] in the context of pharmaceuticals.

Perhaps the greatest limitation of these existing models is in terms of their dynamic

structures and the role of marketing actions. For instance, in terms of customer acquisition researchers have focused on the relationship between acquisition discounts and customer lifetimes [17–19]. These models begin to relate long-term behavior to initial marketing actions but tend to lack full dynamic structures. Several researchers [5,6] have noted this type of limitation and called for the development of models that create optimal individual level marketing policies across multiple campaigns [20] or over the entire customer life cycle.

CUSTOMER ECONOMICS

Two particularly important and closely related CRM profitability metrics are customer lifetime value (CLV) and customer equity [7]. Gupta and Lehmann [2] define CLV as “the present value of all current future profits generated from a customer over the life of his or her business with a firm.” CLV is therefore a measure of the profitability of individual customers. In contrast, customer equity is a measure of the value of the firm’s overall portfolio of customers. Rust *et al.* [21] define customer equity as “the total of the discounted lifetime values summed over all the firm’s current and potential customers.”

The preceding definition of customer equity with its emphasis on the discounted profitability of the firm’s present and future customers highlights that a firm’s value is the sum of the economic value of its customer relationships. Gupta and Lehmann [2] emphasize this point and discuss how customer equity can be used for valuation of firms and for merger and acquisition analyses. Other researchers have considered how customer equity considerations can be used to guide marketing decision making. Blattberg and Deighton [7] discuss the use of customer equity calculations to balance customer acquisition and retention spending levels and Rust *et al.* [21] provide models for linking marketing investments in advertising, loyalty programs, or service quality to customer equity.

Customer Lifetime Value

While customer equity links the overall value of the firm to its relationship assets, CLV operates at a more micro level and can be used at the level of the individual customer. CLV is an established concept in the database marketing industry [22] and is growing in relevance in the broader marketing community [23]. A detailed summary of the CLV literature is provided by Jain and Singh [23]. Berger and Nasr [8] give the basic structural model of CLV as

$$CLV = \sum_{t=0}^T \frac{\alpha^t(R_t - C_t)}{(1+d)^t}, \quad (1)$$

where t indexes time period, R_t is the revenue from the customer in period t , C_t is the total cost in period t , T is the total number of time periods used in the analysis, d is a single-period discount factor, and α is a retention rate. Berger and Nasr [8] also discuss procedures for adapting the calculations to account for variations in purchase frequency and revenue patterns. It should be noted, however, that the variations in purchase timing and revenue are incorporated by assuming specific deterministic patterns. These basic formulations suffer from two key limitations. First, the expressions do not account for the differences between individual customers. Second, the formulas lack mechanisms for evaluating alternative policies since retention and revenue rates are not treated as functions of marketing decisions.

Dwyer [22] makes a significant contribution by introducing the concept of a migration model. In his approach, retention rates are based on historical rates for different transaction history measures. Specifically, Dwyer computes CLV using purchasing rates that vary based on the time since a customer’s last purchase, L . This measure of time since last purchase is similar to the direct marketing concept of recency [24]. If a customer buys in a given period, at the next purchase occasion the expected purchase probability for the customer is the probability observed for customers who have purchased in the last period ($L_t = 1$). If a purchase is not made, the customer’s recency measure (time

lapsed) is incremented by one ($L_{t+1} = L_t + 1$). In this case, the purchase probability at $t + 1$ would be the historical rate for customers with a recency score of $L_t + 1$. In a migration model, calculation of CLV requires a tree-like structure to account for the cumulative effects of these probabilities.

Figure 1 shows a two-period representation of a typical migration diagram for a customer starting in the highest recency category ($L_1 = 1$). Each node in the network represents a buying decision. If a purchase is made, the customer's time lapsed reverts to one and if a purchase is not made the time lapsed increases by one. The expected first period revenue for a customer who bought in the last period ($L_1 = 1$) is,

$$\text{Revenue}_1 = \Pr(\text{buy}|L_1 = 1) * P + \{1 - \Pr(\text{buy}|L_1 = 1)\} * 0, \quad (2)$$

where P is the offered price and $\Pr(\text{Buy}|L)$ is the probability of purchase conditional to the customer being in state L . In the second period, the expected revenue must consider the probabilistic nature of period 1. The following expression accounts for the probabilities that the consumer is in either state $L = 1$ or $L = 2$ during period 2.

$$\begin{aligned} \text{Revenue}_2 = & \Pr(\text{buy}|L_1 = 1) * [\Pr(\text{buy}|L_2 = 1) \\ & * P + \{1 - \Pr(\text{buy}|L_2 = 1)\} * 0] \\ & + \{1 - \Pr(\text{buy}|L_1 = 1)\} \\ & * [\Pr(\text{buy}|L_2 = 2) * P \\ & + \{1 - \Pr(\text{buy}|L_2 = 2)\} * 0]. \end{aligned} \quad (3)$$

A primary benefit of the migration model is that revenues are based on transaction-history-based retention rates. This is an important distinction because it acknowledges the value of the information contained in observed customer behavior, and CLV estimates are computed as a function of transaction history measures. This approach allows firms to understand the relative value of different segments within the overall portfolio of customers. A difficulty with the migration model is that the calculations become cumbersome when a long time horizon is used.

Understanding the differences in CLV is useful as it allows firms to concentrate resources on their most valuable customers. A popular application of CLV is as a segmentation variable [25]. In this type of application the firm forecasts the long-term value of each member of the customer base and this estimate of future profits is used to determine customer service levels and marketing tactics. Zeithaml *et al.* [25] noted that firms such as FedEx, Bank of America, and The Limited all vary in service levels based on customer profitability estimates. In general, the notion of using CLV as a segmentation variable involves allocating resources based on the economic value of individual customers or segments.

However, while CLV estimates can be used to allocate marketing resources based on differences in forecasted customer value the models discussed thus far are not designed to guide the development of marketing policy. While it may be intuitive that firms should devote more marketing resources to high CLV customers, the CLV

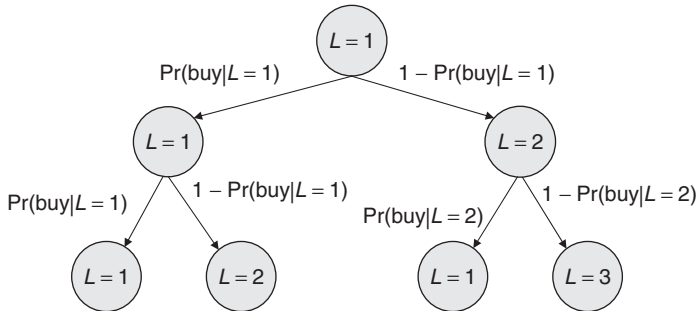


Figure 1. Customer migration.

calculations discussed in this section are not designed to guide the degree to which marketing investments should vary based on CLV. In fact, it could be counterargued that high CLV customers already exhibit significant loyalty and that firms should invest in less loyal segments of customers. The fundamental issue is that the CLV calculations discussed thus far do not model CLV as a function of marketing policy.

MAXIMIZING CUSTOMER LIFETIME VALUE

A limitation with using CLV as a segmentation variable or as an input to a customer equity calculation is that marketing activity is not explicitly considered. In particular, in the preceding equations the revenues, costs and retention rate are all treated as fixed quantities rather than as functions of marketing. This is problematic as key building blocks such as revenue, R_t , are directly a function of the marketer's decisions regarding price. Likewise, the retention rate is also likely to be a function of pricing. Rather than treating CLV as a static measure, a more marketing-oriented approach to CRM would view CLV as an objective function to be maximized through the selection of marketing tactics and strategies. This view that the goal of a firm's CRM strategy should be to maximize each individual's profitability over time has been noted by multiple researchers [26–28].

Marketing and Lifetime Value

The first step in relating marketing decisions to CLV is to acknowledge that the key elements of the CLV equations are dependent on marketing tactics. In Equation (4) below, Equation (2) is rewritten to make explicit that retention rates, revenues and costs may all be functions of marketing activity, M .

$$CLV = \sum_{t=0}^T (\alpha(M_t)^t) \times (R_t(M_t) - C_t(M_t)). \quad (4)$$

Writing the equation for CLV as a function of marketing actions alters how CLV is viewed. Rather than as a static quantity

that may be used to segment customers, CLV is now an objective to be maximized by selecting marketing actions. For example, pricing may increase per period revenues but simultaneously decrease retention. The proper pricing strategy therefore involves balancing short-term revenue with long-term retention.

In addition to including marketing actions in the CLV calculation, it is also useful to include observable customer data. As noted, Dwyer's [22] migration model uses recency or time since last purchase to better account for differences in retention rates across customers. Other salient transaction history variables include a customer's past frequency of buying and the monetary value of past transactions. These three measures are collectively referred to as *RFM* (*recency, frequency, and monetary value*) and are frequently used in database marketing applications [24].

Furthermore, as transaction history measures may be reflective of underlying customer preferences it may be useful to select marketing actions based on past customer behavior. Specifically, marketing tactics may have differential effects across customer segments. This is salient as it suggests that current behavior may be determined by interactions between past buying activity and marketing actions. For instance, there is some belief that long-term customers may be less price sensitive than newer customers [29]. Equation (5) makes explicit that customer behavior may be a function of a set of state variables, X_t , that describe a customer at time t .

$$CLV = \sum_{t=0}^T (\alpha(M_t, X_t)^t) \times (R_t(M_t, X_t) - C_t(M_t, X_t)). \quad (5)$$

Maximizing Customer Lifetime Value

The goal in CRM is to determine the marketing activities that maximize each customer's lifetime value. The multiperiod time horizon of the problem, combined with the probabilistic nature of consumer behavior, means that the development of optimal

CRM marketing policies requires some form of dynamic optimization. The expression in Equation (5) provides an objective function for CRM applications.

This section briefly describes how CRM can be formulated as a dynamic programming problem. The emphasis will be on the general form of the problem rather than on solution techniques (see the sections titled “Markov Decision Processes and Dynamic Programming” in this encyclopedia). For the general case, customers are classified into one of $x_t \in X$ possible states based on transaction history at each point in time, t . The firm is assumed to be able to select a marketing action or controls, m_t , from a set of feasible actions, M . Defining the single-period profit function as $\pi(x_t, m_t)$, the single-period discount factor as d , and the decision-horizon length as T , the firm’s CRM objective function may be written as $\max_{m_t \in M} E \left\{ \sum_{t=0}^T d^t * \pi(x_t, m_t) \right\}$. This objective function is the discounted sum of single-period rewards over a T period horizon. In this expression, the single-period profit, π , is equivalent to the revenue minus cost term, $(R_t - C_t)$, in the previous CLV expressions.

The customer management goal is to determine the mapping of marketing policies to observed customer states $m(x_t)$ that maximizes CLV. For example, in the context of financial or subscription services this could entail determining the life cycle pricing strategy that optimally balances customer acquisition and retention [28]. In the case of direct marketing, the marketing problem might be to determine the mailing frequency that balances revenues with mailing costs [30].

More generally, the firm’s optimization problem is to select the sequence of marketing actions that maximizes profitability over some time horizon. The solution to the dynamic optimization problem is defined in terms of the value function, denoted as V . By definition, the value function is the greatest feasible expected payoff from time t forward, given that the customer is in state x_t at time t . The value function is equivalent to the maximum expected CLV based on optimal

selection of marketing policies:

$$CLV(x_0) = \max_{m_t \in M} E \left\{ \sum_{t=0}^T \alpha^t * \pi(x_t, m(x_t)) \right\}. \quad (6)$$

Thus far, the objective function does not include a direct statement of how customers evolve over time. In Equation (6) optimal customer value (CLV) is rewritten to make the relationship dynamics more explicit. In this equation $\Pr(x_{t+1}|x_t, m(x_t))$ is the probability a customer in state x_t transitions to state x_{t+1} , given the marketing actions specified by $m(x_t)$:

$$\begin{aligned} CLV(x_t, m(x_t)) &= E[\pi(x_t, m(x_t))] \\ &+ \sum_{\tau=t+1}^T \sum_{x' \in X} \alpha^\tau \\ &* \pi(x'_\tau, m(x_\tau)) \\ &* \Pr(x'_{\tau+1}|x_\tau, m(x_\tau)). \end{aligned} \quad (7)$$

The first term on the right side is the expected one-period profit from applying the selected marketing tactics in the current time period. The term involving the double summations represents the sum of the expected profits for future periods from $t+1$ to T . This term includes the transition probabilities through the inner summation over the customer’s potential future states (x'). The outer summation is over future time periods.

The term for the current period profits, $E[\pi(x_t, m_t)]$, is an expectation as customers’ decisions are stochastic from the firm’s perspective [31]. Purchasing probabilities also define the state-to-state transition rates, $\Pr(x_{t+1}|x_t, m_t)$ if the state variables include transaction history measures such as recency or frequency. Customer buying decisions are embedded in both the single-period profit and in the state-to-state transition probabilities.

Customer Evolution

Given that from the firm’s perspective consumers behave nondeterministically, the dynamic optimization problem is to identify the strategy that optimally controls the probabilities of transitions from one

transaction history state to another. In other words, the managerial problem is to influence the evolution of customer relationships so that customers inhabit profitable states. If the state variables are transaction history measures such as recency, frequency, and monetary value, then transition probabilities can be directly related to consumer choices. There is an extensive marketing literature devoted to modeling consumer choices (see Leeflang *et al.* [32] for a review). For example, Gupta [33] provides a model of incidence, expenditure amount, and brand choice. This type of model can generate the probabilities of future customer transaction history states as a function of the current state and marketing activities.

When the goal of marketing becomes an effort to control transaction history profiles special care needs to be taken when modeling consumer behavior. A particularly salient element of consumer decision making models is state dependence or purchase feedback effects. Roy *et al.* [34] define structural state dependence as “the influence of observed past experience with a brand on current choice probabilities.” This type of state dependence is relevant to CRM as positive purchase feedback effects provide a means for developing behavioral loyalty. If product usage increases future propensity to buy then marketing activities such as promotions can be viewed as investments in loyalty. Previous work has identified purchase feedback effects in many categories [35–37]; Seethuraman 2004. In the CRM literature, the notion of developing habit persistence through purchase feedback effects is the key element in models that have modeled customer relationships as Markov chains [38,39].

In addition to state dependence, the marketing literature has also devoted significant attention to identifying unobserved preference heterogeneity [40–42]. Unobserved preference heterogeneity refers to differences in underlying preferences across individuals that are not directly observable by the researcher. Accounting for underlying differences in preferences is critical in CRM models. The issue is that persistent choice or behavioral loyalty is often explainable by either positive purchase feedback effects

or underlying preference heterogeneity [37]. Failing to account for preference heterogeneity can be problematic as it can lead to an overestimation of the effect of past buying (spurious state dependence). Computing optimal policies using a model with spurious state dependence will result in a recommendation of overly aggressive and, therefore, suboptimal promotional policies. Given the emphasis on measures of past loyalty to define customer states in most CRM systems it is critical to account for unobserved preferences.

The marketing literature includes a variety of approaches for estimating unobserved preference heterogeneity. Kamakura and Russell [41] illustrate a method for estimating the latent segments while Rossi *et al.* [42] describe a variety of models that use hierarchical structures to estimate individual level preferences. While the estimation of individual level parameters is intuitively appealing, in practice it may be difficult to generate optimal policies for individual customers. In addition to estimating a model that generates individual level response parameters, the optimization model would also need to be executed for each individual customer. In contrast, segment level models of consumer behavior only require solving a dynamic programming model for a small number of customer segments. An important question is whether the benefits of generating individual level marketing policies justify the required additional effort.

Dynamic Programming

The definition of an appropriate objective function, marketing decision variables, and a model of consumer decision are the necessary components to treat CRM as a dynamic programming model. In terms of implementation, there are significant computational challenges. Consumer behavior is complex; the state space needed to adequately model consumer decision making can quickly become enormous. For example, a consumer state space that includes 10 levels of recency, 10 levels of frequency, and 10 levels of monetary value would involve 1000 states. If we added demographics such as sex and five levels of age, the state space

would grow to 10,000 states. If we also considered past marketing actions such as two alternative channels of customer acquisition and whether the consumer was acquired via a discount, the space would involve 40,000 states. Bellman [43] coined the term *the curse of dimensionality* to describe this exponential growth of problem size as a function of the problem's dimensions. While powerful computing systems and algorithmic advances [44,45] have made it possible to solve relatively simple customer management problems, the use of dynamic programming in CRM is still in its infancy (see the section titled "Large-scale MDPs" in this encyclopedia).

Dynamic programming models have been used primarily in the context of catalog mailing decisions [30,38,46]. In these models the firm attempts to determine the ideal mailing schedule as a function of customer transaction history. Simester *et al.* [30] find that it may be beneficial to send catalogs to customers who may not be likely to purchase in the current period. The implication is that sending catalogs not only spurs immediate purchases but may also have long-term benefits similar to advertising. Lewis [28] used a dynamic programming model to determine optimal life cycle pricing schedules for newspaper subscribers. This study found that the optimal pricing policy involves using a sequence of decreasing discounts as customer tenure increases. The studies by Simester *et al.* and Lewis illustrate a salient element of dynamic programming models of CRM. In both cases the prescribed marketing policies involve investments in customer relationships that would be neglected in myopic models.

Khan *et al.* [47] use a dynamic programming model to determine the optimal sequence and types of promotions to offer to grocery buyers. This study is executed using data from an online grocer that offers a variety of promotions including free shipping, discount coupons, and a loyalty program. In terms of findings, Khan *et al.* find that free shipping is the most effective promotion for customers who tend to purchase large amounts while coupons that offer linear discounts are more effective with customers

who usually purchase small amounts. Free shipping is also found to be the most effective promotion for reacquiring lapsed customers.

In addition, there have been a number of other recent studies that, while not explicitly stated as dynamic programming models, have developed marketing policies designed to maximize multiperiod customer revenue based on observed individual level customer data. Zhang and Krishnamurthi [48] create three-week price discount schedules for category level promotions to optimize a brand's profitability. Rust and Verhoef [27] optimize the mailings of relationship building communications. Venkatesan *et al.* [49] use Bayesian decision theory to optimally select customers based on CLV projections.

Consumer Behavior Considerations

The use of dynamic programming models to determine CLV and marketing policies highlights the marketing side of CRM. The output of these models is a set of policies that map marketing tactics to observed customer transaction history measures. This allows marketing decisions to be directly linked to the data the firm possesses about each individual customer. From a data availability and implementation standpoint this is a necessary approach. However, it should be noted that standard transaction history data may be of limited value in predicting customer behavior.

Consumers are obviously complex and it will always be a significant challenge to capture sufficient data and to develop models that account for the vagaries of consumer decision making. For example, there may be a link between customer acquisition and long-term customer retention [10]. Anderson and Simester [19] and Lewis [17] find that new customers acquired under different pricing promotions exhibit different long-term behaviors. Similarly, it may be useful to consider past marketing stimuli, customer transaction history, and interactions between marketing and past consumer decisions when analyzing customer reacquisition [18,50]. Service failures might be another area where dynamic programming models could be employed. Previous studies of the link between retention and service

failures have yielded mixed evidence [51]. In one notable study, Bolton [13] finds a positive link between unreported service failures and retention. Bolton suggests that this counterintuitive result may occur because consumers who do not report service failures may be inherently more tolerant. Including measures of service failures or complaints in a customer state space would be useful for guiding service recovery efforts. In many respects, incorporating information on service failures within a CRM optimization model would represent an effort to quantify the ill will effects that are often considered in standard newsvendor models [52]; see also article titled *Newsvendor Models*.

From a more conceptual perspective, it may be useful to define customer “states” in a more flexible manner and to consider data that may not be typically available in customer data warehouses. Advertising exposures, previous service quality [13], past promotions, and a myriad of other interactions between the customer and the brand may influence consumer loyalty. The crucial point is that it may be useful to augment customer databases with additional information that may be useful for guiding marketing policy development. For example, it may be useful to collect information on the communications that occur between brands and consumers. The development of relationships between brands and consumers is a critical but under-researched topic in CRM [53,54]. The level of the psychological or emotional connections between brands and consumers is clearly a salient element of a customer’s “state.” An open question is on how relationships between brands and consumers are best developed. This type of brand development problem could be cast as a dynamic optimization problem where the objective is to determine the optimal longitudinal communications strategy. Aaker *et al.* [55] present a field study that examines how brand personality and brand failures collectively impact relationship development. They find that differences in brand personality influence the resiliency of the consumer–brand relationship. While this study is not presented along the lines of an optimization framework that strives

to map the right marketing actions to customer states, the underlying concepts are similar in that the goal is to understand how brand personality messages and brand transgressions impact relationship strength.

In addition, many of the factors that determine customer loyalty may be unobservable. For example, the marketing literature includes models that have incorporated consumer learning and expectations [56–58]. Expectations of future quality, pricing, or catalog mailings are potentially very important. The existence of expectations highlights that consumers are active, learning entities that may themselves adopt a long-term perspective in their marketing relationships. The role of consumer expectations has long been a focal topic in services research. Specifically, the services literature has been concerned with how customer satisfaction translates into customer retention. For instance, a question has been raised regarding whether very high levels of satisfaction (“Delight”) increase future loyalty or merely raise expectations to unsustainable levels [59,60]. This type of question could conceivably be answered by adding satisfaction or other measures of service quality to a customer state space.

FUTURE DIRECTIONS

The idea of using optimization techniques to maximize the value of customers is intuitive and has been contemplated since at least the 1950s [61]. However, the computational challenges associated with dynamic programming and the costs of maintaining large-scale detailed customer databases have left these types of techniques infeasible until only recently. Many technical challenges still remain. For example, the notion of one-to-one marketing popularized by Peppers and Rogers [4] suggests that marketing policies should be determined at the level of individual customers. However, while one-to-one marketing is an intuitively appealing concept, there is little research that has studied the relative benefits and costs of one-to-one marketing programs.

Khan *et al.* [47] examine the benefits of customizing marketing policies at various levels. They use a hierarchical Bayesian approach [42] to estimate individual level response to prices and a variety of promotional instruments. The customer-specific parameters are then used to determine individual level optimal marketing policies. In addition to creating individual level policies, they also measure the benefits of determining marketing policies at the individual, segment, and overall population levels. They find that the majority of the benefits of customization can be achieved by considering transaction history measures and segment level differences. This is a salient finding because the computational challenges associated with determining individual level demand models and generating individual level marketing policies remain difficult. This is particularly true in environments characterized by large, complicated customer state spaces and in organizations that have large numbers of customers. Khan *et al.*'s findings suggest that the majority of the possible benefits from CRM can be achieved by optimizing marketing at the segment level. However, much additional research is needed on this topic and it should be noted that massive customization of marketing efforts may also have significant cost effects [62]; see also article titled **Mass Customization**.

Furthermore, while CRM systems now frequently provide the means to customize marketing policies to increasingly fine segments, there is reason to be concerned about potential backlash from customers who receive less advantageous offers. Customization of prices based on observable behavior or inferences about customer type are potentially controversial [63,64]. While basing pricing on transaction history measures is generally accepted in some contexts such as discounts to new customers or when airlines charge higher prices to business customers, other price discrimination techniques may harm customer relationships. For example, Coca Cola earned negative publicity for testing vending machines that varied prices based on the weather [65]

and Amazon.com came under fire in 2000 when consumers learned they were paying different prices for the same DVDs [66]. Because of adverse publicity Coca Cola chose not to implement the temperature-based pricing technology and Amazon.com used refunds and apologies to placate consumers who were charged higher prices in the dynamic pricing experiment. From a modeling standpoint, the key point is that it may be useful to incorporate information about the treatment of other customers into predictive models.

The computational challenges and the possibility of consumer backlash highlight the need for additional research. The incorporation of increasingly sophisticated models of consumer behavior is a particularly important avenue for future research. For example, just as firms can benefit from adopting a dynamic orientation so too can customers. If customers learn to anticipate firm actions then the firm's optimization models must also consider strategic behavior on the part of consumers [11,46]. Optimization in circumstances where consumers are strategic and forward looking is even more computationally burdensome as the problem requires the solution of a dynamic game between the firm and consumers [46].

REFERENCES

1. Keller P, Keller KL. A framework for marketing management. 4th ed. Upper Saddle River (NJ): Pearson Prentice Hall; 2009.
2. Gupta S, Lehmann D. Managing customers as investments: the strategic value of customers in the long run. Philadelphia (PA): Wharton School Press; 2005.
3. Rigby D, Reichheld F, Scheffer P. Avoid the four perils of CRM. *Harv Bus Rev* 2002; 80(2):101–109.
4. Peppers D, Rogers M. The one to one future. New York: Doubleday; 1993.
5. Rust R, Chung TS. Marketing models of service and relationships. *Mark Sci* 2006; 25(6):560–580.
6. Kumar V, Venkatesan R, Bohling T, *et al.* The power of CLV: managing customer lifetime value at IBM. *Mark Sci* 2008;27(4):585–599.

7. Blattberg R, Deighton J. Manage marketing by the customer equity. *Harv Bus Rev* 1996;74:136–144.
8. Berger P, Nasr N. Customer lifetime value: marketing models and applications. *J Interact Mark* 1998;12(1):17–30.
9. Gupta S, Zeithaml V. Customer metrics and their impact on financial performance. *Mark Sci* 2006;25(6):718–739.
10. Thomas J. A methodology for linking customer acquisition to customer retention. *J Mark Res* 2001;38(2):262–268.
11. Lewis M. The influence of loyalty programs and short-term promotions on customer retention. *J Mark Res* 2004;41(3):281.
12. Gupta S, Lehmann D, Stuart J. Valuing customers. *J Mark Res* 2004;41(1):1–17.
13. Bolton R. A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction. *Mark Sci* 1998;17(1):45–65.
14. Schmittlein D., Bemmaor A., Morrison D. Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Market Sci* 1985;4(3):255–266.
15. Li S, Sun B, Wilcox R. Cross-selling sequentially ordered products: an application to consumer banking. *J Mark Res* 2005;42(2):233–239.
16. Kamakura W, Kossar B, Wedel M. Identifying innovators for the cross-selling of new products. *Manage Sci* 2004;50(8):1120–1133.
17. Lewis M. Customer acquisition promotions and customer asset value. *J Mark Res* 2006;43:195–204.
18. Thomas J, Blattberg R, Fox E. Recapturing lost customers. *J Mark Res* 2004;41(1):31–45.
19. Anderson E, Simester D. Long-run effects of promotional depth on new versus established customers: three field studies. *Mark Sci* 2004;23(1):4–21.
20. Blattberg R, Kim B, Neslin S. Database marketing. New York: Springer; 2008.
21. Rust R, Lemon K, Zeithaml V. Return on marketing: using customer equity to focus marketing strategy. *J Mark* 2004;68(1):109–127.
22. Dwyer F. Customer lifetime value to support marketing decision making. *J Dir Mark* 1989;8(2):73–81.
23. Jain D, Singh S. Customer lifetime value research in marketing: a review and future directions. *J Interact Mark* 2002;16(2):34–46.
24. Hughes A. Strategic database marketing. 2nd ed. New York: McGraw-Hill; 2000.
25. Zeithaml V, Rust R, Lemon K. The customer pyramid: creating and serving profitable customers. *Calif Manage Rev* 2001;43(4):118–142.
26. Reinartz W, Kumar V. The impact of customer relationship characteristics on profitable lifetime duration. *J Mark* 2003;67(1):77–99.
27. Rust R, Verhoef P. Optimizing the marketing interventions mix in intermediate-term CRM. *Market Sci* 2005;24(3):477–489.
28. Lewis M. Research note: a dynamic programming approach to customer relationship pricing. *Manage Sci* 2005;51(6):986–994.
29. Reichheld F, Teal T. The loyalty effect. Boston (MA): Harvard Business School Press; 1996.
30. Simester D, Sun P, Tsitsiklis J. Dynamic catalog mailing policies. *Manage Sci* 2006;52(5):683–696.
31. Gaudagni P, Little J. A logit model of brand choice calibrated on scanner data. *Mark Sci* 1983;2:203–238.
32. Leeflang P, Wittink D, Wedel M, *et al.* Building models for marketing decisions. Boston (MA): Kluwer Academic Publishers; 2000.
33. Gupta S. Impact of sales promotions on when, what and how much to buy. *J Mark Res* 1988;25(4):342–355.
34. Roy R., Chintagunta P. Haldar S. A framework for investigating habits: the hand of the past, and heterogeneity in dynamic brand choice. *Market Sci* 1996;15(3):280–299.
35. Gedenk K, Neslin S. The role of retail promotion in determining future brand loyalty: its effect on purchase event feedback. *J Retail* 1999;75(4):433–459.
36. Chang K, Siddarth S, Weinberg C. The impact of heterogeneity in purchase timing and price responsiveness on estimates of sticker shock effects. *Mark Sci* 1999;18(3):178–192.
37. Keane M. Modeling heterogeneity and state dependence in consumer choice behavior. *J Bus Econ Stud* 1997;15(3):310–327.
38. Bitran GR, Mondschein SV. Mailing decisions in the catalog sales industry. *Manage Sci* 1996;42(9):1364–1381.
39. Pfeifer P, Carraway R. Modeling customer relationships as Markov chains. *J Interact Mark* 2000;14(4):43–55.
40. Heckman J. Heterogeneity and state dependence. In: Rosen S, editor. *Studies in labor markets*. Cambridge (MA): National Bureau of Economic Research; 1981. pp. 91–139.

41. Kamakura W, Russell GJ. A probabilistic choice model for market segmentation and elasticity structure. *J Mark Res* 1989;26:379–390.
42. Rossi P, Allenby G, McCulloch R. Bayesian statistics and marketing. West Sussex, England: Wiley; 2005.
43. Bellman R. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
44. Rust J. Using randomization to break the curse of dimensionality. *Econometrica* 1997; 65(3):487–516.
45. Keane M, Wolpin K. The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence. *Rev Econ Stat* 1994;76:648–672.
46. Gonul F, Shi MZ. Optimal mailing of catalogs: a methodology using estimable structural dynamic programming models. *Manage Sci* 1998;44(9):1249–1262.
47. Khan R, Lewis M, Sing V. Dynamic customer management and the value of one-to-one marketing. *Mark Sci* 2009.
48. Zhang J, Krishnamurthi L. Customizing promotions in online stores. *Mark Sci* 2004;23(4):561–578.
49. Venkatesan R, Kumar V, Bohling T. Optimal customer relationship management using Bayesian decision theory: an application for customer selection. *J Mark Res* 2007;XLIV:579–594.
50. Lewis M. Incorporating strategic consumer behavior into customer valuation. *J Mark - (Special issue on customer relationship management)* 2005;69(3):230–238.
51. Gustafsson A, Johnson M, Ross I. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *J Mark* 2005;69(4):210–218.
52. Petruzzi N, Dada M. Newsvendor models. *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken (NJ): Wiley; 2010. In press.
53. Fornier S. Consumers and their brands: developing relationship theory in consumer research. *J Consum Res* 1998;24(4): 343–373.
54. Aaker J. Dimensions of brand personality. *J Mark Res* 1997;34(3):347–356.
55. Aaker J, Fournier S, Brasel A. When good brands do bad. *J Consum Res* 2004; 31(1):458–469.
56. Erdem T, Keane M. Decision-making under uncertainty: capturing dynamic brand choice processes in turbulent consumer markets. *Mark Sci* 1996;15(1):1.
57. Gonul F, Srinivasan K. Estimating the impact of consumer expectations of coupons on purchase behavior: a dynamic structural model. *Mark Sci* 1996;15(3):262.
58. Boulding W, Kalra A, Staelin R, *et al.* A dynamic process model of service quality: from expectations to behavioral intentions. *J Mark Res* 1993;30:7–27.
59. Oliver R, Rust R, Varki S. Customer delight: foundations, findings, and managerial insight. *J Retail* 1997;73:311–336.
60. Rust R, Oliver R. Should we delight the customer. *J Acad Mark Sci* 2000;28(10):86–94.
61. Howard R. Comments on the origin and application of Markov decision processes. *Oper Res* 2002;50(1):100–102.
62. Parlakturk A, Swaminathan J. Mass customization. *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken (NJ): Wiley; 2010. In press.
63. Kahneman D, Knetsch JL, Thaler R. Fairness as a constraint on profit seeking entitlements in the market. *Am Econ Rev* 1986;76(4):728–741.
64. Feinberg F, Krishna A, John Zhang Z. Do we care what others get? A behaviorist approach to targeted promotions. *J Mark Res* 2002;39(3):277–291.
65. Egan C. Vending-machine technology matures, offering branded foods, convenience. *Wall St J (East Ed)* 2001:B13.
66. Hamilton D. E-commerce special report: the price isn't right. *Wall St J (East Ed)* 2001:R.8.

CUSTOMIZED PRICE RESPONSES TO BID OPPORTUNITIES IN COMPETITIVE MARKETS

MARK FERGUSON
College of Management, Georgia
Institute of Technology,
Atlanta, Georgia

INTRODUCTION

Price optimization applications are typically associated with the pricing of consumer goods or the optimal design of auctions. There are a large percentage of firms, however, who face pricing decisions in a business-to-business setting where a customer requests bids from a small set of competing firms and the firms vying for the customer's business respond with a single price quote for the product or service. When the total annual sales to the firm requesting the bid does not justify a dedicated sales person on behalf of the firm responding to the bid, many firms start using analytical models to provide customized pricing recommendations on what price to offer for the business being bid upon. Customized pricing models provide a probability of winning for every possible price response, allowing a firm to balance a decreasing margin with an increasing win probability needed in a price optimization model. Examples of firms using customized pricing models include the united parcel service (UPS), when responding to bids for their small-to-medium size customers [1]; BlueLinx, when responding to requests for products from construction companies [2]; Marriot, when responding to group sales requests for their properties [3]; and Siemens, for pricing of industrial automation controls [4]. The models have also been used in the business-to-consumer market in the financial services industry, when banks determine what interest rate to offer when responding to request for mortgages, credit cards, and auto-loans [5,6]. The financial

improvement from using customized pricing models can be significant. UPS reported an increase in profits of over \$100 million per year by optimizing their price offerings using customized pricing models [7].

The value of customized pricing models comes from the ability to determine significant price segments based on a firm's historical bid history. Price segments are defined as sets of transactions, classified by customer, product, and transaction attributes, which exhibit similar price sensitivities. Customer attributes may include customer location, size of the market the customer is in, type of business the customer is in, the way the customer uses the product, customer purchase frequency, customer size, and customer purchasing sophistication. Product attributes may include product type, life-cycle stage, and the degree of commoditization. Transaction attributes may include order size, other products on the order, channel, specific competitor, when the order is placed, and what the urgency is of the bidder. In addition, some models assume knowledge of the historical and current bid-price of competing firms participating in the bid opportunity.

A common characteristic of environments where firms may employ customized pricing models is when the bidder with the lowest price does not always win the bid. Thus, markets are characterized by product differentiation where a given firm may command a positive price-premium over its competitors, depending upon the particular customer requesting the bid response. Sometimes even bids from the same customer may contain some inherent amount of uncertainty in the bid winning probability due to the bid-requesting firm randomly allocating its business to different competitors to ensure a competitive market for future bids. Because of these practices, a firm will never be able to remove all uncertainty from the bid-price response process and must work with probabilistic models.

A second common characteristic of an appropriate environment is when the size

of the bid opportunities is not large enough to justify a sales person dedicated to each customer. Instead, a single sales person may respond to multiple bid opportunities from a variety of potential customers each day. The most common alternatives to using customized pricing models is either to charge a fixed price to all customers or to have a sales agent respond to each separate bid opportunity with a customized price. Charging a fixed price leads to missed opportunities to price discrimination between different customer segments—a practice that has been well publicized for significantly increasing a firm's profit in many different industries. The other alternative, relying on a sales agent to respond to multiple bid opportunities, is also problematic. Theoretically, the sales agent should have knowledge of the market, based on a history of former bid responses with the customer requesting the bid, allowing the sales agent to respond with a customized price that optimizes this inherent trade-off between decreasing margins, due to lowering the price, and increasing probabilities of winning the bid. In reality, sales agents often do not make good trade-off decisions in these situations because of a lack of historical knowledge, the inability to process this historical knowledge into probability distributions, or misaligned incentives [8]. The judicious use of customized pricing models allows firms to capture historical bid information, analyze it, and present nonbiased price recommendations for future bidding opportunities. If there is additional information available regarding the bidding opportunity that cannot be captured in the model, the model's recommended price may serve as one of the many possible inputs to the person responsible for making the bid-response decision. Additional background and information on where customized pricing models are best applied can be found in Ref. 9.

RELATION TO TRADITIONAL PRICE OPTIMIZATION

The practice of price optimization is well known for applications where the demand

for a product over a certain period of time is correlated with the price charged for the product over the same time period. In such a situation, a firm can measure the demand at different price points for a product and use this data to estimate a price-response function. This traditional application to price optimization works well in environments where there are a large number of potential customers who may each buy a small quantity of the product (e.g., consumer retail stores) or where a small number of potential customers may purchase large quantities of a product but spread their purchases over many suppliers (e.g., commodity spot markets such as grain, steel, or oil). It is less helpful, however, in winner-take-all bidding situations that are common in business-to-business transactions, where the customer commits to purchasing a given quantity of goods or services and solicits bids from a set of firms capable of providing it. In these situations, the decision a providing firm is concerned about is not if (or how much) the customer will buy, but rather, will the customer buy from a firm as opposed to one of its competitors. Thus, the provider firm does not face a decision of what price to place on a product to attract demand, but instead, what price to quote to this particular customer for this particular bid opportunity (customized pricing).

Figure 1 shows a historical demand plot for a customized pricing scenario. Notice that the wins (indicated by values of 1) are more common for the lower prices, while the losses (indicated by values of 0) are more common for the higher prices.

The data captured in Fig. 1 is the historical win/loss data for a firm over 150 past bid opportunities for the same product. Judging from the data, it appears that the average price offered has been around \$10 per unit but the firm has historically priced it as low as \$8, and as high as a little over \$12. It also appears that a lower than average price does not guarantee a win, nor does a higher than average price guarantee a loss—there is some uncertainty in the demand. Therefore, we focus on maximizing the firm's *expected profit*, consisting of the total profit, assuming the firm wins the bid (margin multiplied by the quantity,



Figure 1. Historical demand data for customized pricing scenario.

Q) times the probability of winning a bid for a given price, which we label $\rho(P) \in [0, 1]$. Thus, the firm's expected profit is

$$\pi(P) = (P - C)Q\rho(P).$$

ESTIMATING THE PROBABILITY FUNCTION

Once $\rho(P)$ is estimated, the actual price optimization part for the problem is straight forward and differs little from the traditional linear demand problem. The difficulty, of course, lies in estimating $\rho(P)$. Before we discuss some various methods of doing so, we define some properties that any estimated function should possess. First, the function should decrease monotonically as the price increases. Second, the function should be bounded by 0 and 1 (Fig. 2).

While there are several models that provide an S-shaped probability function, the most common model used in practice is the Logit model. This model is described in Refs 7 and 9, and is compared against a competing model in Ref. 10. In its simplest form, the Logit model is represented by

$$\rho(P) = \frac{e^{\alpha + \beta P}}{1 + e^{\alpha + \beta P}}, \quad (1)$$

where α and β are parameters that must be estimated to fit the historical win/loss data. The parameter estimation is performed by minimizing the squared errors of the residuals or by using maximum-likelihood estimates. Before estimating the parameter values of the models, however, it is important to divide the historical win/loss data set into two segments; one for estimating the parameter values and the other for measuring

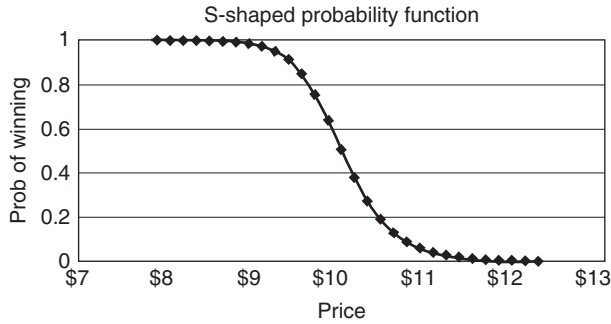


Figure 2. Fitted S-shaped probability function to win/loss data.



Figure 3. Marginal deal contribution versus unit price.

the fit. Similar to time series forecasting models, measuring the goodness-of-fit on the same data that the parameter values were estimated on may result in a misleadingly close fit as compared to testing the model on a holdout sample. Phillips [9] describes how each estimation method is applied to the Logit function. In general, if we define each historical bid opportunity with the subscript i (with W_i representing the indicator variable, $1 = \text{win}$ and $0 = \text{loss}$, for each bid i and P_i representing the firm's price response for bid opportunity i), the most common estimation method is to choose parameter values that maximize the likelihood of the observed values

$$\begin{aligned} \text{Max}_{\alpha, \beta} \quad & \sum_i \ln[\rho(P_i | \alpha, \beta) W_i \\ & + [1 - \rho(P_i | \alpha, \beta)](1 - W_i)]. \end{aligned}$$

Price Optimization for Customized Pricing

We now see how customized pricing probability models are used in price optimization. For the following discussion, we use the objective of maximizing expected profits. However, other strategic or operational objectives can be easily accommodated such as increasing market share or including constraints on capacity, inventory, price, or margin. In addition, the firm is assumed to be risk-neutral in this example. A risk-averse firm may prefer to optimize expected revenue with a concave utility function so as to mitigate the chances of bad individual outcomes. The price optimization problem for bid opportunity i (with



Figure 4. Probability of win versus unit price.

Q_i now representing the quantity that is requested in bid i) is

$$\text{Max}_{P_i} \quad \pi(P_i) = \rho(P_i) \times (P_i - C) \times Q_i. \quad (2)$$

Note that the margin $(P_i - C)$ is strictly increasing in price (Fig. 3) but the probability of winning the bid is strictly decreasing in price (Fig. 4). Therefore, the expected profit is often a unimodal function as shown in Fig. 5.

Determining the optimal price involves finding a global maximum for the expected profit. Since the profit function is unimodal, any search-based optimization algorithm can be used to solve for the price that maximizes profit, including the solver package available in MS Excel. The profit-maximizing price can also be found by solving for the price where the elasticity of the expected profit function is equal to the inverse of the marginal contribution ratio.



Figure 5. Expected profit versus unit price.

SEGMENTING CUSTOMERS BASED ON HISTORICAL PRICE BEHAVIOR

The basic Logit model described in Equation (1) only includes the price as a predictor of the probability of winning a bid. This model is appropriate if there are no discernable differences in the price sensitivity of a firm's customer set; every customer has the same probability, for every bid opportunity, of accepting the firm's bid at a given price. In this situation, the customized price optimization (Eq. 2) will recommend the same optimal price for every future bid opportunity. In practice, this is rarely the case—if it was, then why would the firm be applying customized pricing to begin with? A more common scenario is where a firm's sales force has historically set different prices for each bid opportunity based on certain characteristics of the bid or of the customer requesting the bid. The characteristics driving the different prices, which may even differ by salesperson within a firm, may include such things as the size of the customer (annual revenue), the length of time a firm has been a customer, the quantity requested by the bid, or the number of firms offered to submit to this particular bidding opportunity. Thus, a bid opportunity from a small customer who has had a long-term relationship with the firm and typically involves only two additional firms in the bidding process may receive a higher price quote than a large customer with little sales history who includes at least five competitive quotes in every bid opportunity. In an interesting

example of customized pricing in the banking industry, Kadet [6] describes how some banks place consumers into pricing segments and quote customized interest rates each time a potential customer inquires about a loan. It is exactly this case, when customers can be segmented on the basis of their price sensitivities, that customized pricing models provide the largest benefit.

The Logit model (1) can be expanded to include segmentation variables. To show how, we use an example from an application of customized pricing at a major credit bureau, which is a real company that we call Alpha Company because of the sensitive nature of the data set. Alpha sells credit scores of individuals to businesses that extend credit to their customers such as auto dealerships and jewelry stores. If an individual purchases his or her credit score, they pay a list price of around \$10. Businesses that purchase multiple credit scores annually often send out bids to the three largest credit bureaus, promising a minimum volume of score requests per year in exchange for a discounted price. Alpha is typically included in these bid opportunities so they have a substantial historical database that includes each past bid opportunity, the size of the minimum quantity promised in each contract, and the length of time in months that the business requesting the bid has been a customer of Alpha. We denote the quantity of each bid by Q and the length of the relationship with Alpha in months by M . In addition, we define δ as the coefficient for the quantity

parameter and λ as the coefficient for the length of the relationship. Including these new variables and coefficients into Equation (1) results in

$$\rho(P|Q, M) = \frac{e^{\alpha + \beta P + \delta Q + \lambda M}}{1 + e^{\alpha + \beta P + \delta Q + \lambda M}}. \quad (3)$$

The estimation of the coefficients ($\alpha, \beta, \delta, \lambda$) is performed through a maximum likelihood fit. However, it still remains to determine which segmentation variables should be included in the model. Just because a firm has historical data for a segmentation variable does not mean that it should be included in the model. Next, we explain how to determine whether a segmentation variable should be included.

Many different approaches to segment data exist. The number and type of segments can be determined in advance (*a priori*), such as asking the sales team what customer characteristics they use when determining the price to respond to a bid. While this knowledge (perhaps based on years of experience) should not be discounted, it should, however, be statistically tested. There are many cases where a firm implementing a customized pricing model finds, when doing so, that many of the characteristics they have historically used to segment customers are not statistically significant based on the historical sales data. Thus, it is also useful to determine (or confirm) customer segments based on data analyses (*post hoc*). Some methods for determining customer segments include nonoverlapping and overlapping clustering methods, classification and regression trees, and Expectation Maximization algorithms. A detailed analysis of these methods is beyond the scope of this article, but we refer the reader to Ref. 11 for a detailed overview. However, the easiest and most popular method is probably to run a logistic regression [12,13]. This technique is available in most statistical software packages and, similar to linear regression, a statistical significance test can be applied to the predictor variables. Running a logistic regression on Alpha's historical win/loss data results in the output shown in Table 1.

Observing the p -values in the regression output table shows that only the constant, the

Table 1. Output from Logistic Regression on Alpha's Historical Win/Loss Data

Regression Coefficients	Coefficient	p -Value
Constant	-0.270631513	0.0476
Quantity	8.33914E-06	0.6596
Price	-0.228872395	0.0094
Active months	0.362277116	<0.0001

price, and the active months are significant at the 95% level (p -value less than 0.05). Thus, the quantity variable is dropped from the model and a second regression is run using only the significant variables. The estimated coefficients of the variables from the second regression are shown in Table 2.

Substituting in the estimated coefficient values into Equation (3) leaves

$$\rho(P|M) = \frac{e^{-0.21 - 0.24P + 0.36M}}{1 + e^{-0.21 - 0.24P + 0.36M}}. \quad (4)$$

The fact that quantity was not a significant segmentation variable was a surprise to the sales team at Alpha. Prior to this study, they felt that the quantity requested in a bid was a better indicator of the price sensitivity of the customer than the length of time the customer had been doing business with Alpha. There may be other significant segmentation variables that were not included in the model, such as geographic location, and company size. Moving forward, Alpha plans on collecting additional information on each bidding opportunity so that other possible segmentation variables can be tested. In general, the number of customer attributes (segments) that can be accurately estimated depends on the amount of historical bid information available. If extensive historical data is available, greater degrees of segmentation

Table 2. Output from Logistic Regression after Removing Quantity

Regression Coefficients	Coefficient	p -Value
Constant	-0.205193446	0.0401
Price	-0.244049279	0.0028
Active months	0.363146316	<0.0001

can be achieved without compromising the accuracy and robustness of the statistical estimation of the parameter values.

While building statistically significant probability models is important, what firms really care about is improvements in profits. In the next section, we show how to test the performance of a customized pricing model.

MEASURING PERFORMANCE

We alluded to the point earlier that the historical win/loss data should be divided into an estimation set and a holdout sample set. A holdout set is critical for obtaining a realistic measure of the model's performance; it is misleading to measure the performance on the same data that was used to estimate the model's coefficients. There are two performance metrics available: percent improvement in profits over unoptimized actual profits and percent improvement in profits over unoptimized expected profits. To understand the difference between the two performance metrics, consider the following bid opportunity (Table 3) from Alpha's historical win/loss data:

The above is obtained by applying the Logit model from Equation (4) to the bid opportunity and optimizing results in an optimal price of \$5.15 for this particular bid opportunity. Substituting in the original bid price of \$3.92 into Equation (4) results in the probability of winning for the unoptimized bid of 0.24 or 24%. Substituting the optimal price results in a probability of winning for the optimized bid = 0.19.

Since Alpha's marginal cost of providing a credit score is essentially 0, the actual profit from this bid opportunity is = (original unit price – marginal cost) $\times$ order size $\times$ win/loss indicator variable = $(\$3.92 - 0) \times 240 \times 1 = \940.80 . If the original bid had resulted in a loss, the actual profit

would be 0. The original bid expected profit = (original unit price – marginal cost) $\times$ order size $\times$ probability of win at the original bid price = $(\$3.92 - 0) \times 240 \times 0.24 = \225.79 . Note that the expected profit is always smaller than the actual profit when the bid was won, and is always larger when the bid was lost. The optimized bid expected profit = (optimized price – marginal cost) $\times$ order size $\times$ probability of win at the optimized bid price = $(\$5.15 - 0) \times 240 \times 0.19 = \235.60 . These calculations can then be used to calculate the two performance measures for this bid opportunity as follows:

- Percent improvement in optimized bid expected profits over unoptimized bid actual profits = $(\$235.60 - \$940.80)/\$940.80 = -0.75 = -75\%$
- Percent improvement in optimized bid expected profits over unoptimized bid expected profits = $(\$235.60 - \$225.79)/\$225.79 = 0.04 = 4\%$.

The calculations above should be performed for every bid opportunity in the holdout set and the sum of each measure (over each bid opportunity in the holdout set) can then be used as a measure of performance for the model. For this data set (holdout sample equal to 160 historical bids), the total percent improvement in actual profits was 137% and the total percent improvement in expected profits was 112%.

IMPLEMENTING A CUSTOMIZED PRICING OPTIMIZATION PACKAGE

Thus far, we have focused on the technical aspects of customized pricing optimization. What is clearly evident, however, from the presentations by individuals whose firms

Table 3. Example: Historical Bid Opportunity

Bid	Win	Order Size	Active Months	Original Bid	Optimal Bid	Probability of Win at Original Bid	Probability of Win at Optimized Bid
451	Y	240	0	\$3.92	\$5.15	0.24	0.19

have implemented customized pricing models, is that the most difficult part of an implementation is not the proof of the value of the system; nor is it building the models and performing the price optimizations. Instead, the most quoted difficulty involves the acceptance of the system's price recommendations by the existing sales team [1,2,4,14]. The removal of some decision-making authority away from individuals to a more automated system is problematic in any environment; however, it is particularly difficult for pricing since the sales team may feel that the ability to generate customized price quotes is a large part of the value they bring to the firm and their incentive system (often a commission based on total revenue) may not match the profit maximization objective of a customized pricing optimization system.

Given the almost certain cultural issues and expected resistance to a system implementation, are there best practices that may be applied to help mitigate these problems? It is clear that there are some practices that seem to help, based on the same panel discussions referenced above. The most frequently suggested technique is to treat the customized price optimizer as a decision support tool for the sales team rather than a system that will automatically set prices for all future bid opportunities. No automated pricing system will ever completely replace the ability of humans to factor in extenuating circumstances or information that is not captured in the historical sales data. Sales personnel often overemphasize the value of human judgment so there needs to be some incentive to follow the recommendations of the pricing models. One such incentive that has worked for several firms is to track the frequency that a sales person sets a price within the recommended range from the pricing model and then publish these results along with the monthly profits made by each sales person. If, as expected, the sales personnel ranked the highest for pricing within the recommended price ranges are also ranked the highest for monthly profits, then the rest of the sales team will copy this practice to improve their own performance.

CONCLUSIONS

To summarize, the following steps are involved in building a customized pricing optimization model:

1. Start with a historical data set of the firm's previous bid opportunities for the product of interest. This data set should include both wins and losses along with the price submitted for each bid opportunity and any other segmentation data available on the customer or bid. The historical data set should be randomly divided into two distinct sets: the first for estimating the model parameter values and the second for performance evaluation (the holdout data).
2. A win/loss probability model, such as the Logit model, should be developed that includes coefficients for any segmentation variables.
3. Using the estimation set from the historical data, the parameter values for the probability model should be estimated using maximum likelihood estimators. This can be done by running a logistic regression if a Logit model is used for the probability model. The output from the regression will identify the segmentation variables that are significant.
4. After selecting the win/loss probability model that provides the best fit for the holdout sample data, use this model to optimize the bid prices for all the bids in the holdout set from the historic data.
5. Percent improvements over expected profits and over actual profits can then be calculated using the holdout data to measure the model's performance.

While customized pricing models hold great potential for substantially increasing profits, any firm considering adopting price models for customized pricing needs to be aware of its limitations. The models behind customized pricing assume that the bid opportunities are exogenous and are not affected by the bid responses suggested

through the optimization. In reality, a firm's pricing strategy may have a significant impact on customer retention, especially if the optimization model recommends consistently pricing higher than the competition for a particular customer class. Also, the optimization models do not assume any strategic response from the firm's competitors. Instead, they assume that the actions of competitors will stay the same as they were during the time period covered by the historical data set. In reality, competitors may react to a firm's new pricing strategy causing the historical bid opportunity data to be unrepresentative of future bid price responses. To help detect these possibilities, mechanisms should be put in place to monitor and evaluate the performance of the models over time. If competitors change their bid-pricing behavior due to the implementation of a customized pricing solution, more involved models using concepts from game theory should be employed.

REFERENCES

1. Kniple J. Director of pricing strategy and solutions at UPS. In: Panelist in non-traditional industries, Georgia Tech 2nd annual workshop on Price Optimization and Revenue Management; Atlanta (GA). 2006 May 18th.
2. Dudziak B. Senior manager in the planning and analysis group of BlueLinx. In: Panelist in non-traditional industries, Georgia Tech 2nd annual workshop on Price Optimization and Revenue Management; Atlanta (GA). 2006 May 18th.
3. Hormby S, Morrison J. Marriot International, Instructors for workshop "Is Bigger Really Better? Bulk Pricing and Negotiated Deals." In: Georgia Tech and Revenue Analytics 4th \$annual conference on Price Optimization and Revenue Management; Atlanta (GA). 2008 Nov 11th.
4. Stucke G. Manager of global and national agreements at siemens energy and automation. In: Panelist in Bulk Sales and Negotiated Deals—Winning at the Highest Margins, Georgia Tech and Revenue Analytics 4th \$annual conference on Price Optimization and Revenue Management; 2008 Nov 11th.
5. Phillips R. Pricing optimization in consumer credit. In: Presentation at the 2005 INFORMS Annual Meeting, San Francisco (CA); 2005a.
6. Kadet A. Price profiling. *Smart Money* 2008 Apr 28: 81–85.
7. Boyd D, Gordon M, Anderson J, *et al.* Manugistics. Target Pricing System. Patent US 6963854 B1. 2005.
8. Garrow L, Ferguson M, Keskinocak P, *et al.* Expert opinions: current pricing and revenue management practice across U.S. industries. *J Revenue Pricing Manage* 2006;5(3):248–250.
9. Phillips R. Pricing & revenue optimization. Stanford (CA): Stanford University Press; 2005b.
10. Agrawal V, Ferguson M. Bid-response models for customized pricing. *J Revenue Pricing Manage* 2007;6(3):212–228.
11. Wedel M, Kamakura W. Market segmentation: conceptual and methodological foundations, International series in quantitative marketing. Norwell (MA): Kluwer Academic Publishers; 1998.
12. Hosmer D, Lemeshow S. Applied logistics regression, Wiley series in probability & statistics. Honoken (NJ): John Wiley & Sons, Inc; 1989.
13. Kutner M, Nachtsheim C, Neter J. Applied linear regression models. 4th ed. New York: McGraw-Hill; 2004.
14. Preuer G. Director of pricing at cooper industries. In: Panelist in Bulk Sales and Negotiated Deals—Winning at the Highest Margins, Georgia Tech and Revenue Analytics 4th \$annual conference on Price Optimization and Revenue Management; Atlanta (GA). 2008 Nov 11th.

CUSUM CHARTS FOR MULTIVARIATE MONITORING AND FORECASTING

KWOK-LEUNG TSUI

YAJUN MEI

School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

WEI JIANG

Department of Industrial
Engineering and Logistics
Management, Hong Kong
University of Science and
Technology, Hong Kong, P.R.
China

INTRODUCTION

The cumulative sum (CUSUM) statistic, which was first introduced by Page [1], is an unweighted sum of all past observations.

$$S_t = \sum_{i=1}^t (X_i - \mu_0),$$

where X_i is the observation at time i and μ_0 is the targeted mean of the process. The initial purpose of the CUSUM statistic is to detect changes of the mean level in the sequence of X_t . Intuitively, if the mean of the process is larger than μ_0 , the mean path of $\{S_t\}$ directs upward when plotted along time. If the process mean level is smaller than μ_0 , the sequence of $\{S_t\}$ will show a downward trend. Page [1] proposed to issue an alarm at time t when $S_t - \min_{1 \leq i \leq t} S_i > h$ or $\max_{1 \leq i \leq t} S_i - S_t > h$, where h is a positive threshold value to control the false alarm rate.

Since its first introduction in quality control, the CUSUM chart and associated algorithms have found tremendous applications and proved successful in detecting process changes. Basseville and Nikiforov [2] and Hawkins and Olwell [3] discussed a

few engineering areas including quality control in manufacturing and production improvement, condition monitoring of inertial navigation systems, seismic signal processing and segmentations of signals, as well as vibration monitoring of mechanical systems. In manufacturing systems, by assuming two states of a production process, in control and out of control, production decisions have to be reached sequentially to decide if the process has entered the out-of-control state. Similarly, in navigation systems, using sensor information and the motion equations to estimate the coordinates and the velocities of moving objects is critical for object tracking and control. In signal and image processing applications, seismic data processing and signal segmentations are usually the first processing step in determining *in situ*, the geographical coordinates and boundaries. This article will briefly discuss several other applications in disease surveillance, sensor networks, and forecasting applications. Within these diversified applications, as a statistical tool, the CUSUM statistic is of great interest in three common problems, (i) detect abrupt process changes, (ii) estimate the effect of changes, and (iii) track and forecast dynamic systems. The goal of the decision rules is generally the quickest detection of the disorders with as few false alarms as possible.

CUSUM STATISTICS, CHANGE-POINT PROBLEM, AND SEQUENTIAL TESTING

Page's CUSUM statistic stems on Wald's [4] sequential probability ratio tests of whether the process mean has changed to some specific value μ_1 . In general, assume that X_t has probability density $f(x, \theta)$, where θ is the parameter of the probability distribution. To test if the parameter has changed from θ_0 to θ_1 at time i given a sequence of observations $\{X_t\}$, the generalized likelihood ratio (GLR)

statistic can be developed as

$$W_t = \max_{1 \leq i \leq t} \Lambda_i^t,$$

where $\Lambda_i^t = \prod_{k=i}^t f(X_k, \theta_1)/f(X_k, \theta_0)$, which triggers an alarm when $W_t > h$. When X_i is normally distributed with mean μ and variance 1, the GLR test is equivalent to the CUSUM statistic S_t by searching the change-point time when maximizing the GLR statistic [1]. Barnard [5] also developed a V-mask to judge crossing the boundary. Kemp [6] further showed that the use of V-mask on a CUSUM chart is equivalent to monitoring the two CUSUM

$$\begin{aligned} W_t^+ &= \max(W_t^+ + X_t - \mu_0 - k, 0), \\ W_t^- &= \max(W_t^- - X_t + \mu_0 - k, 0), \end{aligned}$$

with a horizontal limit h , where k is a positive reference value. The reference value k is usually set at half of the mean shift magnitude. The threshold value h is generally determined to maintain a prespecified in-control average run length (ARL). Statistically, ARL is the expected number of observations before an alarm is triggered. A larger in-control ARL and a smaller out-of-control ARL are preferred in practice. In general, when X_t follows distributions other than Gaussian, the above CUSUM forms may be adjusted.

The widespread applications of the CUSUM statistic is mainly due to its optimality for detecting a prespecified mean shift, either small or large [7–9]. Theoretically, when X_t follows a standard normal distribution, an optimal CUSUM chart can be obtained for detecting a mean shift of size δ in X_t , if the reference value $k = \delta/2$. In practice, the CUSUM chart may perform poorly when actual mean shift is different from $2k$. When the shift size is unknown, Siegmund [10] proposed the GLR statistic optimized not only with reference to the change-point time, but also with reference to the shift magnitude. Unlike the CUSUM statistic, this GLR statistic involves heavy computations and is hard to implement in practice.

CUSUM AND STATISTICAL PROCESS CONTROL (SPC)

When the shift magnitude is unknown in industrial practice, several modifications and simplifications of the CUSUM charts have been proposed to achieve good performance over a range of mean shifts. Several authors suggest running multiple CUSUM charts simultaneously, each optimized for a different size mean shift [11–14]). Hawkins *et al.* [15] and Lai [16] investigated a change-point model by assuming a prior distribution of the shift size. Sparks [13] proposed an adaptive cumulative sum (ACUSUM) procedure by dynamically adjusting the reference value of the conventional CUSUM chart for detecting a range of mean shifts. Luo *et al.* [17] also developed an ACUSUM chart with variable sampling intervals.

Another challenge facing industrial applications of the CUSUM schemes is the assumption of constant parameter shifts. Linear drift is a common type of mean shift that has been investigated widely. Other types may include residuals from time series models [18]. To track dynamic parameter changes, weighting cumulative sums is a natural generalization of the traditional CUSUM charts for different purposes. Yashchin [19] proposed two different weighted cumulative sum (WCUSUM) charts for monitoring process locations when sample sizes vary over time and when the process mean is subject to linear drifts rather than step changes. Cumulative score (Cuscore) statistic proposed by Box and Ramirez [20] is another example of WCUSUMs for detecting feared fault signals in a white noise process. Although the Cuscore statistic is a generalization of the traditional CUSUM statistic, the former critically depends on the starting point of the change, whereas the latter does not. Owing to this reason, it often requires the assumption that the number of the suspected periods at which a change may occur is small or known *a priori*, so that a Cuscore chart can be started at any such suspected period. Shu *et al.* [21] proposed a WCUSUM control chart for monitoring dynamic mean shifts from time series models.

Another important application of the CUSUM charts in industrial process control is to detect process variance changes, which, however, has not been popularly discussed [22]. Logarithm of sample variance has been traditionally suggested for CUSUM monitoring procedures due to the property that it is approximately normally distributed with a constant variance [23]. Recently, research has shown that the CUSUM chart based on the sample variance itself is more efficient than that based on the logarithm of the sample variance [24].

MULTIVARIATE CUSUM (MCUSUM)

It is natural to extend the univariate CUSUM charts to monitoring multivariate data. Woodall and Ncube [25] defined a multiple CUSUM method under multivariate normal distributions. The idea is similar to considering the maximum over the CUSUM statistic of individual responses. Similar to the T^2 statistics [26], Crosier [27] proposed a multivariate cumulative sum (MCUSUM) method as charting the T^2 statistic over the CUSUM statistic of individual responses. Pignatiello and Runger [28] proposed two MCUSUM statistics, one of which is defined as the CUSUM statistic of the univariate T^2 statistic from the multivariate vector of responses. While some simulations have been conducted for evaluation purposes, it is not clear which CUSUM statistic has better performance in general. Jiang and Tsui [29] related the T^2 chart, M-chart, and regression-adjusted chart [30] and showed that the performance of a multivariate control chart may depend on correlation of the variables as well as the direction and magnitude of the process shift. Although the discussion was limited to Shewhart-type control charts, the conclusions can be generalized to MCUSUM charts. They further proposed a hybrid method that combines the T^2 chart and regression-adjusted chart for better performance.

Under various applications, different CUSUM statistics have been proposed for monitoring multistreams (regions) data. Tarakovsky *et al.* [31] studied a simpler

CUSUM statistic similar to the one proposed by Sonesson [32]. Their idea is to first compute the temporal CUSUM statistic for each region, then to define the overall test statistic as the maximum over the individual CUSUM statistic over the multiple streams. Raubertas [33] developed another CUSUM statistic: Instead of computing the temporal CUSUM statistic for each region, he proposed to compute the simultaneous CUSUM statistic for the center and four (or fewer) adjacent regions for each region center, and then define the overall testing statistic as the maximum over these simultaneous CUSUM statistics.

MCUSUM AND DISEASE SURVEILLANCE

Scan statistic is a special class of MCUSUM statistic for spatiotemporal surveillance. The objective for disease surveillance is to detect and signal outbreaks as early as possible, which may occur at certain regions and last for certain amount of time. Kulldorff *et al.* [34] and Kulldorff [35] proposed retrospective and prospective scan statistics, respectively, based on the basic theory for the spatial case explained above. The spatiotemporal scan statistic approach can be regarded as an extension of the spatial method. The cycle scanning the area of interest is attributed a “height” so that a scan cylinder is created. The height corresponds to the time window that is scanned. The cylinder height can vary just as the cycle radius with one end fixed to the most recent time for the prospective scan case, and both ends free for the retrospective case. The time window also needs to lie within certain constraints.

Woodall *et al.* [36] reviewed and examined various scan statistics used for spatiotemporal surveillance. They pointed out that this method can only work if a cluster contains at least two incidence counts because otherwise the likelihood function could be maximized by decreasing the size of the cycle while centered on a single data point. They also stated that for the prospective scan case, only very few performance evaluations have been undertaken; and they proposed performance measures to compare this method to

statistical process control (SPC) methods. In addition, a robust and detailed evaluation of the spatiotemporal scan statistic seems to be hard to obtain, given the current state of research.

By extending the scan statistics in Kulldorff [35], Sonesson [32] proposed a CUSUM method for timely detection of emerging clusters of diseases. On the basis of a general likelihood function, he defined a general CUSUM statistic as the maximum over all individual CUSUM likelihoods over all possible subsets of regions (variables). He showed that the CUSUM likelihood statistic conditional on a given center region with a fixed radius (or nearest neighbor) can be written in the traditional CUSUM form with a recursive formula. He also showed that Kulldorff's spatiotemporal scan statistic method could be interpreted as a conditional likelihood approach given the total number of incidents and given that the parameters are known. Simulation studies were conducted to compare Sonesson's CUSUM with Kulldorff's scan statistic with known parameters, and the performance of the CUSUM was found to be superior. This result is not surprising as the CUSUM statistic has been shown to be optimal under certain conditions.

In disease surveillance, the SPC-based spatiotemporal surveillance problem has been studied by just a few researchers. Rogerson [37] presented some control chart-based work on the spatiotemporal regional case. He extended the retrospective method of Tango [38] to a prospective application using a CUSUM method. Rogerson *et al.* [39] considered the spatiotemporal problem for which the counts in the subregions are correlated at each particular time, with the correlation decreasing as the distance between the subregions increases. They compared the performance of the use of multiple CUSUM charts [25] for each region against an MCUSUM method [27].

Other applications of CUSUM methods in health care include risk-adjusted monitoring schemes for clinical practices and surgical performance evaluation [40,41]. Owing to the heterogeneity of the baseline patient risk, the surgeon's performance may be masked by the heterogeneity, which makes it slow

to respond to performance changes. Grigg and Farewell [42] provide an overview of risk-adjusted charts and present a comparison based on an empirical case from a single data set. Tsui *et al.* [43] provided a comprehensive review of applications of SPC in health care, public health, and syndromic surveillance.

MCUSUM AND SENSOR NETWORK APPLICATIONS

The CUSUM statistic/procedure has been successfully applied in Mei [44] to tackle the decentralized quickest change detection problems in sensor networks, which have many important applications, including multisensor data fusion, mobile and wireless communication, surveillance systems, and distributed detection.

Specifically, in a widely used configuration of sensor network systems, at each time step, each of a set of local sensors receives an observation and then sends a sensor message to a central processor, called the *fusion center*, which makes a final system-wise decision when observations are stopped. In order to reduce cost and increase reliability, it is required that the sensor messages are "quantized" in the sense of belonging to a finite alphabet (perhaps binary). This limitation is dictated in practice by the need for data compression and limitations of channel bandwidth. In decentralized quickest change detection problems, it is assumed that at some unknown time, the distributions of the sensor observations change abruptly and simultaneously at all sensors. The goal is for the fusion center to utilize the "quantized" sensor messages to detect the change as soon as possible over all possible protocols for generating sensor messages and over all possible decision rules at the fusion center, under a restriction on the frequency of false alarms.

One novel decentralized scheme developed in Ref. 44 is for each local sensor to calculate the local CUSUM statistic at each time step, and then send binary "sensor decisions" $\{0,1\}$ to the fusion center, depending on whether the local CUSUM statistic exceeds

the local threshold. The fusion center then combines all local sensor decisions using a “consensus” rule, that is, it stops and decides that a change has occurred as soon as all local CUSUM statistics exceed their respective local thresholds. A surprising theoretical result in Ref. 44 is that with suitably chosen local thresholds, this decentralized scheme has the same (first-order) asymptotic performance as the optimal centralized scheme that has access to all the sensor observations, although it may perform poorly in some practical situations because of slow asymptotic convergence. It is interesting to note that in the definition of the proposed decentralized scheme, one cannot replace the CUSUM statistic by other well-known change-point detection statistics such as exponentially weighted moving average (EWMA) or Shiryaev–Roberts statistic [otherwise the asymptotic optimality properties will no longer hold, see Remark 4 on p. 2674 of Mei [44]]. Further research is needed to find fairly simple decentralized schemes that are not only asymptotically optimal but also have good performance for practical values.

CUSUM AND FORECASTING APPLICATIONS

One of the major issues in forecasting is the necessity of defining the correct moment for intervening in forecasting procedures that have begun to produce systematic biases. In principle, interventions are called for when the size of forecasting discrepancies increases due to a change in the reality of the forecasting environment. It is thus straightforward to extend the use of control charts for monitoring forecast errors. Trigg [45] first provided a rigorous treatment in monitoring forecast errors. CUSUM forecast monitoring schemes have been employed since 1959, following the procedures of Harrison and Davies [46]. Gardner [47] compared the CUSUM chart for residuals and other smoothed-error forecast monitoring schemes. West and Harrison [48] developed a Bayesian framework that applies the CUSUM scheme to track forecasting errors in predicting complex time series models.

In forecasting economic time series, ignoring structural breaks, which occur in the time

series, significantly reduces the accuracy of the forecast. Brown *et al.* [49] proposed several CUSUM tests based on recursive residuals and squares to identify a single structural break in long memory processes. Inclán and Tiao [50] used the CUSUM for multiple structural break detection. Pesaran and Timmermann [51] proposed reverse CUSUM for detecting the most recent break. Pang and Ting [52] further extended the reverse CUSUM and proposed a data selection method for time series prediction. Although different alternative models are assumed, the CUSUM statistic is obtained based on the likelihood ratio test in principle.

DISCUSSIONS

CUSUM charts and methods have been recognized to be very powerful in process monitoring and forecasting. They have been successfully applied to many applications including manufacturing, health surveillance, and sensor networking. Much research has been done in both theoretical and application domains. In theory, there remains a big challenge in optimality aspects of various developed charting procedures in both univariate and multivariate cases. In applications, there is a great potential for developing additional CUSUM methods for complex monitoring problems, such as spatiotemporal surveillance of infectious diseases [43] and activity monitoring of business processes [53].

REFERENCES

1. Page ES. Continuous inspection scheme. *Biometrika* 1954;41(1/2):100–115.
2. Basseville M, Nikiforov IV. Detection of abrupt changes: theory and application. Englewood Cliffs (NJ): Prentice-Hall; 1993.
3. Hawkins DM, Olwell DH. Cumulative sum charts and charting for quality improvement. New York: Springer Verlag; 1998.
4. Wald A. Sequential analysis. New York: Wiley; 1947.
5. Barnard GA. Control charts and stochastic processes. *J R Stat Soc [Ser A]* 1959;21:239–271.

6. Kemp KW. The average run length of the cumulative sum control chart when a 'V' mask is used. *J R Stat Soc [Ser A]* 1961;23:149–153.
7. Lorden G. Procedures for reacting to a change in distribution. *Ann Math Stat* 1971;42:1897–1908.
8. Pollak M. Optimal detection of a change in distribution. *Ann Stat* 1985;13:206–227.
9. Moustakides GV. Optimal stopping times for detecting changes in distributions. *Ann Stat* 1986;14:1379–1387.
10. Siegmund D. Sequential analysis: tests and confidence intervals. New York: Springer-Verlag; 1985.
11. Lorden G, Eisenberger I. Detection of failure rate increases. *Technometrics* 1973; 15:167–175.
12. Lucas JM. Combined Shewhart-CUSUM quality control schemes. *J Qual Technol* 1982;14:51–59.
13. Sparks RS. CUSUM charts for signalling varying location shifts. *J Qual Technol* 2000;32:157–171.
14. Zhao Y, Tsung F, Wang Z. Dual CUSUM control schemes for detecting a range of mean shifts. *IIE Trans* 2005;37:1047–1057.
15. Hawkins DM, Qiu P, Kang CW. The change point model for statistical process control. *J Qual Technol* 2003;35:355–365.
16. Lai TL. Sequential analysis: some classical problems and new challenges. *Stat Sinica* 2001;11:303–350.
17. Luo Y, Li Z, Wang Z. Adaptive CUSUM control chart with variable sampling intervals. *Comput Stat Data Anal* 2009;53:2693–2701.
18. Alwan LC, Roberts HV. Time series modeling for statistical process control. *J Bus Econ Stat* 1988;6:87–95.
19. Yashchin E. Weighted CUSUM techniques. *Technometrics* 1989;31:321–338.
20. Box GEP, Ramirez JG. Cumulative score charts. *Qual Reliab Eng Int* 1992;8:17–27.
21. Shu LJ, Jiang W, Tsui K-L. A weighted CUSUM chart for detecting patterned mean shifts. *J Qual Technol* 2008;40(2):194–213.
22. Page ES. Controlling the standard deviation by CUSUM and warning lines. *Technometrics* 1963;5:307–315.
23. Chang TC, Gan FF. A cumulative sum control chart for monitoring process variance. *J Qual Technol* 1995;27:109–119.
24. Acosta-Mejia CA, Pignatiello JJ, Rao BV Jr. A comparison of control charting procedures for monitoring process dispersion. *IIE Trans* 1999;31:569–579.
25. Woodall WH, Ncube MM. Multivariate CUSUM quality control procedures. *Technometrics* 1985;30(3):291–303.
26. Hotelling H. Multivariate quality control. In: Eisenhart C, Hastay M, Wallis W, editors. *Techniques of statistical analysis*. New York: McGraw-Hill; 1947.
27. Crosier RB. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* 1988;30(3):291–303.
28. Pignatiello JJ, Runger GC Jr. Comparisons of multivariate CUSUM charts. *J Qual Technol* 1990;22(3):173–186.
29. Jiang W, Tsui K-L. A theoretical framework and efficiency study of multivariate control charts. *IIE Trans* 2008;40(7):650–663.
30. Hawkins DM. Multivariate quality control using regression-adjusted variables. *Technometrics* 1991;33:61–75.
31. Tarakovsky AG, Veeravalli VV. Change-point detection in multichannel and distributed systems with applications. In: Mukhopadhyay N, Datta S, Chattopadhyay S, editors. *Applications of sequential methodologies*. New York: Marcel Dekker, Inc.; 2004. pp.331–363.
32. Sonesson C. A CUSUM framework for detection of space-time disease clusters using scan statistics. *Stat Med* 2007;26(26):4770–4789.
33. Raubertas RF. An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Stat Med* 1989;8:267–271.
34. Kulldorff M, Athas WF, Feuer EJ, *et al.* Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *Am J Public Health* 1998;88(9):1377–1380.
35. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *J R Stat Soc [Ser A]* 2001;164: 61–72.
36. Woodall WH, Marshall JB, Joner MD, *et al.* On the use and evaluation of prospective scan methods for health-related surveillance. *J R Stat Soc [Ser A]* 2008;171:223–237.
37. Rogerson PA. Surveillance systems for monitoring the development of spatial patterns. *Stat Med* 1997;16(18):2081–2093.
38. Tango T. A class of tests for detecting general and focused clustering of rare diseases. *Stat Med* 1995;14(21–22):2323–2334.
39. Rogerson PA, Yamada I. Monitoring change in spatial patterns of disease:

- comparing univariate and multivariate cumulative sum approaches. *Stat Med* 2004;23(14):2195–2214.
40. Steiner SH, Cook RJ, Farewell VT, *et al.* Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000;1:441–452.
 41. Spiegelhalter DJ, Grigg OA, Kinsman R, *et al.* Sequential probability ratio tests (SPRTs) for monitoring risk-adjusted outcomes. *Int J Qual Health Care* 2003;15:1–7.
 42. Grigg O, Farewell V. An overview of risk-adjusted charts. *J R Stat Soc [Ser A]* 2004;167(3):523–539.
 43. Tsui K-L, Goldsman D, Chiu W, *et al.* A review of healthcare, public health, and syndromic surveillance. *Qual Eng* 2008;20(4):435–450.
 44. Mei Y. Information bounds and quickest change detection in decentralized decision systems. *IEEE Trans Inf Theory* 2005;51:2669–2681.
 45. Trigg DW. Monitoring a forecasting system. *Oper Res Q* 1964;15:271–274.
 46. Harrison PJ, Davies O. The use of cumulative sum (CUSUM) techniques for the control of product demand. *Oper Res* 1964;12:325–333.
 47. Gardner ES. CUSUM vs smoothed-error forecast monitoring schemes—some simulation results. *J Oper Res Soc* 1985;36:43–47.
 48. West M, Harrison J. Bayesian forecasting and dynamic models. 2nd ed. New York: Springer; 1997.
 49. Brown RL, Durbin J, Evans JM. Techniques for testing the constancy of regression relationships over time. *J R Stat Soc [Ser A]* 1975;37:149–163.
 50. Inclán C, Tiao GC. Use of cumulative sums of squares for retrospective detection of changes of variance. *J Am Stat Assoc* 1994;89:913–923.
 51. Pesaran H, Timmermann A. Market timing and return prediction under model instability. *J Emp Finance* 2002;9:495–510.
 52. Pang KP, Ting KM. Improving time series prediction by data selection. *International Conference on Computational Intelligence for Modeling, Control & Automation*; 2003 Feb 12–14; Vienna. pp. 803–813.
 53. Jiang W, Au T, Tsui K-L. A statistical process control approach for business activity monitoring. *IIE Trans* 2007;39:235–249.

CZECH SOCIETY FOR OPERATIONS RESEARCH

PETR FIALA
Department of Econometrics,
University of Economics,
Praha, Prague, Czech
Republic

The Czech Society for Operations Research (CSOR) was registered at the Home Office of the Czech Republic on October 10, 1993 as a profession society seated in Prague.

The objectives of the Society are:

- to maintain the development of the Operations Research (OR) as a branch;
- to maintain the education and the professional activities specializing in the field of OR;
- to organize and coordinate the cooperation with partner societies abroad.

The membership of the Society consists of private persons (individual membership) and corporations (collective membership). The membership is conditioned by professional qualification. The membership fee is annually approved by the Council of the Society. Reduced fee is available for the elderly and students.

The CSOR is very small society with approximately 40 members. Council of the CSOR consists of the president, two vice presidents, and six members of the Council. Presidents are elected for the period of two-years.

PAST PRESIDENTS

1993–1995 Miroslav Maňas, University of Economics, Prague
1995–1997 Karel Zimmermann, Charles University, Prague
1997–1999 Jaroslav Ramík, Silesian University, Opava

1999–2001 Jiří Beck, University of West Bohemia, Pilsen

2001–2003 Jan Pelikán, University of Economics, Prague

2003–2005 Jana Hančlová, Technical University of Ostrava, Ostrava

2005–2007 Josef Jablonský University of Economics, Prague

2007–2009 Petr Fiala, University of Economics, Prague

CSOR ACTIVITIES

Main activities of the CSOR are annual conferences, seminars, and consultations with firms. The annual conferences, Mathematical Methods in Economics, are organized by the Czech Society for Operations Research and the Czech Econometric Society; mathematical methods in economics is a traditional meeting of professionals from universities and business who are interested in the theory and applications of OR and econometrics. The conferences take place in various Czech cities in the first half of September. The conferences are regularly attended by more than 100 participants from Czech Republic and other countries. The main themes of the conference include

- Operational Research
- Econometrics
- Quantitative Management Methods
- Simulations
- Mathematical Modeling
- Process Optimization

In the year 2009 an informal seminar for the Society members on results and problems in education, science, and practice of OR was organized for the first time. The activity received a very positive reception by members. The informal seminars should be a tradition in the forthcoming years, and the Doctoral Dissertation Award for OR students is presented at this conference.

The CSOR is a member of IFORS and EURO. Representatives of the CSOR attend regularly the councils of these associations. The CSOR cooperates by editing of the quarterly published *Central European Journal of Operations Research*. Members of the Society are granted the subscription of the journal. The Society has strong contacts with the Slovak Society for Operations Research. There are very often visits between the Societies and members of both societies frequently participate in each other's conferences. The CSOR cooperates closely with the Austrian Society as well.

The main international activity organized by the Czech Society for Operations Research was the 22nd European Conference on Operational Research EURO XXII in 2007. The conference was organized by the CSOR in cooperation with the University of Economics, Prague. The conference was held in the campus of the University of Economics which is situated very close to the Prague historical center. The conference was very successful, attended by more than 2000 participants from all over the world.

DAMAGE, STRESS, DEGRADATION, SHOCK

WALTRAUD KAHLE
Department of Mathematics,
Otto-von-Guericke University,
Magdeburg, Germany

INTRODUCTION

Investigating the lifetime of products or biological organisms that are affected by degradation, damage, stress, or shocks it is often necessary to consider the development of a stochastic process which models ageing, wearing, damage accumulation, and other changes of the state. In the following, we will call such processes *degradation models*.

The applicability of a purely lifetime-based reliability analysis is limited due to several reasons. Collecting a sample of lifetime data is often expensive. In applications from engineering, observed lifetimes are frequently very long. Furthermore, there are many products for which breakdown causes serious difficulties. When designing the mathematical model of a degradation process, the following steps have to be taken:

1. Starting from basic technical considerations, a suitable mathematical model for the degradation measure process $Z(t)$ has to be defined. For instance, the hypothesis of additive degradation accumulation may be used. In this case, we may select the class of homogeneous processes with independent increments as a suitable model class. Another possibility is the description of degradations by shock models.
2. By applying the level exceeding theory, it is possible to find the type of lifetime distribution, that is, the distribution of the time at which the stochastic process firstly exceeds the degradation level.

3. The parameters of the selected model have to be estimated by means of samples related to the degradation measure $Z(t)$.
4. Estimations of the parameters of lifetime distribution can be obtained from the estimated parameters of the model $Z(t)$, if the limit level of degradation is known.

As example, one can model the degradation $Z(t)$ by the linear path model

$$Z(t) = t/A,$$

where A is a positive random variable [1,2]. A generalization and statistical analysis by nonparametric methods of estimation from joint linear degradation and failure time data with partial renewals and competing failure modes is given in Ref. 3.

Another widely used model is the gamma degradation process: a continuous-time stochastic process with independent and gamma-distributed increments. A detailed description and many further references are given in Refs 4–6 (see also ***Accelerated Life Models***).

In the next section, we describe in more detail continuous degradation models based on a Wiener process with drift.

Another widely used class of models are *shock models* [7]. The standard assumptions in shock models are that the failure is related either to the cumulative effect of a large number of shocks or that failure is caused by a shock that exceeds a certain critical level, or a mixture of both of them. Asymptotic properties of a wide class of models are given in Ref. 8. In the last section, we consider a general *parametric* shock model.

CONTINUOUS DEGRADATION MODELS

The choice of the mathematical model for these degradation processes is based on the assumption of an additive accumulation of

degradation with constant wear intensity. Regarding every degradation increment as an additive superposition of a large number of small effects, we can assume the degradation process to be normally distributed. Therefore, the degradation measure $Z(t)$ can be described by the following model [9]:

$$Z(t) = z_0 + \sigma W(t - t_0) + \mu \cdot (t - t_0), \quad t \geq t_0 \quad (1)$$

with

z_0 = constant initial degradation ($z_0 \in \mathbb{R}$),
 t_0 = beginning of the degradation ($t_0 \in \mathbb{R}$),
 μ = drift parameter ($\mu \in \mathbb{R}$),
 σ = variance parameter ($\sigma > 0$),
 $W(t)$ = standard Wiener process on $[0, \infty)$.

We suppose that a failure of a product will occur if the degradation process arrives at a certain boundary degradation level, which, in general, is unknown. For a given boundary level h , the lifetime T_h of the product is then determined as the instant at which the degradation process $Z(t)$ exceeds the level h for the first time:

$$T_h = \inf\{t \geq t_0 : Z(t) \geq h\}. \quad (2)$$

It is well known that for $z_0 < h$ the lifetime T_h follows an inverse Gaussian distribution with the Lebesgue density function

$$f_{T_h}(t) = \frac{h - z_0}{\sqrt{2\pi\sigma^2(t - t_0)^3}} \times \exp\left(-\frac{(h - z_0 - \mu(t - t_0))^2}{2\sigma^2(t - t_0)}\right) I_{[t > t_0]}, \quad (3)$$

where $I_{\{\cdot\}}$ denotes an indicator variable, assuming the value 1 if the relation in brackets is true, and the value 0 otherwise. Obviously, P^{T_h} is the Dirac measure ε_{t_0} concentrated at t_0 if $z_0 \geq h$. To estimate the parameters of this model, we can have several kinds of observations:

1. The increments of the degradation process are observed.

In this case, the likelihood function of the observation is a product of Gaussian

densities. It is possible to estimate the parameters of the degradation process μ , σ^2 , z_0 , and t_0 , but it is impossible to estimate the degradation level h . Note that neither the same distance between observation points nor observation of all realizations at one and the same time points are required. Parameter estimations can be found if the observations are taken at individual time points in each realization.

2. The failure times are observed.

In this case, we get the classical likelihood function from a sample of inverse Gaussian distributed random variables. From this likelihood function, we can estimate the parameters μ , σ^2 , and t_0 . The degradation reserve $h - z_0$ must be given in advance.

3. The sample includes both process increments and failure times.

Assume that the sample consists of N failure times and $\sum_{i=1}^n m_i$ increments, being mutually independent. Then, the likelihood function is the product of the likelihood functions mentioned below. In case the failure times and process increments belong to the same process, the likelihood function is more complicated. In this case, the density of a process increment is a conditional density under the assumption that the process does not cross the degradation level in the actual time interval. If the sample includes both process increments and failure times, then it is possible to estimate all the parameters of the model.

These models and their parameter estimation are described in Kahle [9] and Kahle and Lehmann [10]. A similar model and its application in medicine is described in Doksum and Normand [11]. Several generalizations of this simple model were given. It is possible to include measurement errors [12], or to transform the time scale [13]. It is also possible to consider bivariate models [14].

The advantages of using the Wiener process and its generalizations for describing the degradation process are its simple form (at least for the univariate Wiener process), and

secondly, a statistical analysis can be carried out for observations at any discrete time points. But these models have also disadvantages: It is possible that the degradation is decreasing in any interval and this is difficult to interpret in practical applications. The second disadvantage is that these models become very complicated if a nonlinear drift is assumed. But for many products, we can expect an increasing degradation which becomes faster over time.

PARAMETRIC SHOCK MODELS

Introduction

This model deals with a degradation process (Z_t) whose paths are monotonically increasing step functions. We make use of marked point processes $\Phi = ((T_n, X_n))_{n \geq 1}$. A mathematical foundation is given by Last and Brandt [15] and Anderson *et al.* [16]. The cumulative process (Z_t) is assumed to be generated by a doubly stochastic Poisson process (T_n) . This process was introduced by Cox [17]. It was proposed as a model in risk theory by Cramér [18]. Grandell [19] gave a detailed discussion of doubly stochastic Poisson processes and their impact on risk theory. Further applications may be found in reliability theory, medicine, and queueing theory [16, 20, 21].

There is a stream of papers concerning shock models. Applications were given—among others—by Sobczyk [22] and Esary *et al.* [23]. Further papers dealing with shock models are Refs 24–33. In the present paper, we assume cumulated intensities of (T_n) given as a random multiple Y of a deterministic function η . Frequently, these stochastic intensities are more realistic than the deterministic intensity of a Poisson process. Each realization of the random variable Y leads to an other individual intensity of (T_n) which could be a result of individual environmental conditions or of the individual frailty of the item.

Our aim is to describe suitable models for degradation accumulation, to calculate the corresponding cumulative degradation at time t , to consider a first passage time, and to determine its distribution.

Modeling Degradation by Marked Point Processes

Let $[\Omega, \mathcal{F}, P]$ be a fixed probability space and let $\{\mathcal{F}_t\}$ be a filtration in $\mathcal{F}$. The random variable T_n ($n \geq 1$) is the time of the n th shock. We suppose

$$\begin{aligned} T_n < T_{n+1} & \quad \text{if } T_n < \infty \quad \text{and} \\ T_n = T_{n+1} & = \infty \quad \text{otherwise.} \end{aligned}$$

The size of the n th increment of the cumulative degradation process $(Z(t))_{t \geq 0}$, is given by a nonnegative random variable X_n ($n \geq 1$). Thus,

$$Z(t) = \sum_{n=1}^{\infty} I(T_n \leq t) \cdot X_n$$

describes the total amount of degradation at time t . The sequence $\Phi = ((T_n, X_n))$ is called a *marked point process*, and $\Phi(t)$ is defined as the random variable representing the number of events that occurred up to time t . Frequently, it is of interest to discuss the first passage problem that the process (Z_t) exceeds a prespecified constant threshold level $h > 0$ for the first time. It is also possible to regard a (random) state of first X_0 at time $T_0 := 0$. Then the corresponding first passage time Z^h is given as

$$Z^h = \inf \left\{ t : \sum_{n=0}^{\infty} I(T_n \leq t) \cdot X_n \geq h \right\}. \quad (4)$$

Let us mention that Z^h coincides with some T_m for $m \in \mathbf{N}$.

We consider the filtration $\{\mathcal{F}_t\} = \{\mathcal{F}_t^{\Phi}\} \vee \sigma(Y)$, where $\{\mathcal{F}_t^{\Phi}\}$ describes the internal history of Φ , and Y (generating $\mathcal{F}_0$) is a nonnegative random variable with finite expectation. The cumulated $(P, \mathcal{F}_t)$ -stochastic intensity $\bar{\nu}(t)$ of (T_n) is assumed to be given by $\bar{\nu}(t) = Y \cdot \eta(t)$, where $\eta(t)$ is a deterministic function with derivative $\xi(t) \geq 0$. Hence, given the outcome $Y = y$ the random variable $\Phi(t)$ is Poisson distributed with mean $y \cdot \eta(t)$. And the unconditional distribution of $\Phi(t)$ is given by

$$\begin{aligned} P(\Phi(t) = k) &= E \left[\frac{[Y\eta(t)]^k}{k!} \exp(-Y\eta(t)) \right], \\ k &= 0, 1, \dots \end{aligned} \quad (5)$$

The sequence T_n is called a $(P, \mathcal{F}_t)$ -doubly stochastic Poisson process. Both, the deterministic function η and the distribution of Y can be specified. For instance, if $\eta(t) = t$ then T_n is called a *mixed Poisson process*, and if $P(Y = y_0) = 1$ the sequence (T_n) represents a *non homogeneous Poisson process* with the deterministic intensity function $y_0 \cdot \xi(t)$.

The following types belong to the most common models for η :

1. Weibull type: $\eta(t) = t^{\alpha+1}$ ($\alpha > -1$)
2. log-linear type: $\eta(t) = t \cdot e^{\alpha t^\gamma}$ ($\alpha \geq 0$, $\gamma \geq 0$)
3. logistic type: $\eta(t) = t \cdot [1 + \ln(1 + \alpha t^\gamma)]$ ($\alpha \geq 0$, $\gamma > -1$).

The random variable Y is a frailty variable. The notion of frailty provides a convenient way to introduce random effects, association, and unobserved heterogeneity into models for survival data. In its simplest form, a frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals. Y can be specified too.

Let us now consider a marking of the sequence T_n [15]. At every time point T_n , a shock causes a random degradation. We describe the degradation increment at T_n by the mark X_n . Let $(\mathbf{R}^+, \mathcal{B}^+)$ be the space of marks and let G be a stochastic kernel from $(\mathbf{R}^+, \mathcal{B}^+)$ to $(\mathbf{R}^+, \mathcal{B}^+)$. Then $\Phi = (T_n, X_n)$ is said to be a *position-dependent G-marking* of T_n if $X_1, X_2, \dots$ are conditionally independent given T_n where for all $B \in \mathcal{B}^+$ and $n \in \mathbf{N}$ the following relation holds true:

$$\begin{aligned} P(X_n \in B | (T_n)) &= G(T_n, B) \\ P - a.s. \quad \text{on } \{T_n < \infty\}. \end{aligned} \quad (6)$$

Moreover, we assume that each mark X_n and Y are conditionally independent given (T_n) , that is, $P(X_n \in B | (T_n), Y) = P(X_n \in B | (T_n))$. With a position-dependent marking, it is possible to describe a degradation model where the increment at the n th shock depends on the time T_n of the n th shock.

For example, let U_n be a sequence of nonnegative iid random variables with the density f_U and let $\delta \in \mathbf{R}$. We assume that

the sequence U_n is independent of T_n . The sequence X_n is defined by $X_n = U_n \cdot e^{\delta T_n}$. That means, we get degradation increments which tend to be increasing ($\delta > 0$) or decreasing ($\delta < 0$). The stochastic kernel G is given by

$$\begin{aligned} G(t, B) &= \int_B f_U(x \cdot e^{-\delta t}) \cdot e^{-\delta t} dx, \\ B &\in \mathcal{B}^+, \quad t \geq 0. \end{aligned}$$

For $\delta = 0$, we have the special case of independent marking. In this case, G defines a probability measure which is independent on the time t .

Characteristics of the Degradation Process

In the section titled “Modeling Degradation by Marked Point Processes,” we introduced a position-dependent marking of a $(P, \mathcal{F}_t)$ -doubly stochastic Poisson process where the filtration $\{\mathcal{F}_t\}$ is given by $\mathcal{F}_t = \mathcal{F}_t^\Phi \vee \sigma(Y)$.

Then from Ref. 15, $\Phi = (T_n, X_n)$ has the $(P, \mathcal{F}_t)$ -stochastic intensity kernel

$$\lambda(t, B) = Y \xi(t) G(t, B), \quad B \in \mathcal{B}^+,$$

where

$$\begin{aligned} \lambda(t+, B) &= \lim_{s \rightarrow 0+} \frac{1}{s} \cdot E \left[\sum_{n=1}^{\infty} I(t < T_n \leq t + s, \right. \\ &\quad \left. X_n \in B) \middle| \mathcal{F}_t \right], \quad B \in \mathcal{B}^+. \end{aligned}$$

But usually, Y cannot be observed such that the actual available information does not coincide with the filtration $\{\mathcal{F}_t\}$. In some cases, as for applying the maximum likelihood theory, we have to consider the intensity kernels with respect to the observed filtration. For example, the $(P, \mathcal{F}_t^\Phi)$ -stochastic intensity kernel of Φ has the representation

$$\begin{aligned} \tilde{\lambda}(t, B) &= E[\lambda(t, B) | \mathcal{F}_{t-}^\Phi] \\ &= \xi(t) G(t, B) E[Y | \mathcal{F}_{t-}^\Phi], \quad B \in \mathcal{B}^+, \end{aligned}$$

with

$$E[Y | \mathcal{F}_{t-}^\Phi] = \frac{\int_0^\infty y^{\Phi(t-)+1} e^{-y \eta(t)} F_Y(dy)}{\int_0^\infty y^{\Phi(t-)} e^{-y \eta(t)} F_Y(dy)},$$

where F_Y denotes the distribution function of Y .

We want to present some characteristics of the counting process $(\Phi(t))$, the sequence of marks (X_n) , the degradation process (Z_t) , as well as the first passage time (Z^h) .

The Counting Process $\Phi(t)$. Frequently, the expected number of shocks up to any time t and other moments of $\Phi(t)$ are of interest. According to Equation (5), we can express the k th ordinary and central moments of $\Phi(t)$ as linear combinations of moments of Y multiplied by powers of the deterministic function $\eta(t)$.

Actually, let $S(k, u)$ denote the Stirling numbers of the second kind where $S(k, u)$ can be recursively determined

$$S(k, u) = S(k-1, u-1) + u \cdot S(k-1, u), \\ 1 \leq u \leq k, \quad (S(0, 0) := 1, S(0, u) = 0).$$

We make use of

$$n^k = \sum_{u=1}^k S(k, u) n(n-1) \times \cdots \times (n-u+1)$$

and we consider the factorial moments of a Poisson-distributed random variable with mean $y \eta(t)$. Some elementary calculations yield

$$E[\Phi(t)^k] = \int_0^\infty \sum_{n=0}^\infty n^k \frac{[y \eta(t)]^n}{n!} e^{-y \eta(t)} dF_Y(y) \\ = \sum_{u=1}^k S(k, u) \cdot E[Y^u] \cdot \eta(t)^u. \quad (7)$$

In particular, we find

$$\begin{aligned} E[\Phi(t)] &= E[Y] \eta(t) \\ \mu_2(\Phi(t)) &= \text{Var}(\Phi(t)) = E[Y] \eta(t) \\ &\quad + \text{Var}(Y) \eta(t)^2 \\ \mu_3(\Phi(t)) &= E[Y] \eta(t) + 3 \text{Var}(Y) \eta(t)^2 \\ &\quad + \mu_3(Y) \eta(t)^3 \\ \mu_4(\Phi(t)) &= E[Y] \eta(t) + (4 \text{Var}(Y) + 3E[Y^2]) \\ &\quad \times \eta(t)^2 + (6E[Y^3] - 12E[Y^2]E[Y] \\ &\quad + 6E[Y]^3) \eta(t)^3 + \mu_4(Y) \eta(t)^4, \quad (8) \end{aligned}$$

where $\mu_k(\cdot)$ denotes the k th central moment of a random variable.

The Sequence (X_n) and the Degradation Process (Z_t) . In the considered case of a position-dependent marking the size of the n th degradation increment depends on the time T_n of the n th shock. Using Equation (6), the distribution of the n th mark X_n is given by

$$P(X_n \in B) = E[G(T_n, B)] \\ = \int_0^\infty G(t, B) f_{T_n}(t) dt, \quad (9)$$

where the density f_{T_n} of T_n has the representation

$$f_{T_n}(t) = \frac{d}{dt} P(\Phi(t) \geq n) \\ = n \cdot \frac{\xi(t)}{\eta(t)} \cdot P(\Phi(t) = n), \quad n \geq 1.$$

As we can see from Equation (9), it is often difficult or strenuous to determine the distribution and the expectation of X_n , respectively. Frequently, we are more interested in the expected cumulative degradation at any time t . In order to simplify notations, we deal with continuous marks X_n . Let $g(T_n, x)$ denote the conditional density of X_n where $g(t, x)$ is given by $g(t, x) = \frac{\partial}{\partial x} G(t, [0, x])$.

Theorem 1. Let $\Phi = ((T_n, X_n))$ be a position-dependent marking of a doubly

stochastic Poisson process with the stochastic intensity kernel

$$\lambda(t, B) = Y \cdot \xi(t) G(t, B), \quad B \in \mathcal{B}^+.$$

Then the expectation of the cumulative degradation at time t is given as

$$E[Z(t)] = E[Y] \int_0^t \xi(s) \cdot \int_0^\infty x \cdot g(s, x) dx ds, \quad t \in \mathbf{R}^+. \quad (10)$$

Sketch of the proof.

By the definition of an intensity, it may be shown that

$$E \left[\sum_{n=1}^{\Phi(t)} \psi(T_n) \right] = E[Y] \int_0^t \psi(s) \xi(s) ds, \quad t \in \mathbf{R}^+$$

for any measurable function ψ . Further, we can express the expectation $E[Z(t)]$ as

$$E[Z(t)] = E \left[\sum_{n=1}^{\Phi(t)} \int_0^\infty x g(T_n, x) dx \right],$$

such that Equation (10) immediately follows.

Let us mention that in the special case of an independent marking, we get the well-known result $E[Z(t)] = E[\Phi(t)] \cdot E[X_1]$. Generally, the expectation $E[Z(t)]$ depends on both, the shape of the deterministic functions η , ξ , and the kernel G .

The Cumulative Degradation at Time t and the First Passage Time Z^h . Another problem is to find the distribution of the cumulative degradation $Z(t)$ at time t . It is easy to see that the distribution can be calculated by

$$P(Z(t) \leq u) = I(0 \leq u)P(\Phi(t) = 0) + \sum_{n=1}^{\infty} P(X_1 + \dots + X_n \leq u, \Phi(t) = n).$$

In the case of independent marking, we get the more simple equation

$$P(Z(t) \leq u) = I(0 \leq u)P(\Phi(t) = 0) + \sum_{n=1}^{\infty} G^{*(n)}(u) \cdot P(\Phi(t) = n), \quad (11)$$

where G is the distribution function of any mark and $G^{*(n)}$ represents the n th convolution of G . The calculation of this distribution is difficult even in the case of independent marking. Methods for calculating the degradation distribution for special distributions of marks can be found in Ref. 21. Explicit expressions for this distribution are known mostly for simple models.

We defined the first passage time Z^h for any state of first X_0 at time $T_0 := 0$ and a random level $h - X_0$ respectively, that is,

$$Z^h = \inf \left\{ t : \sum_{n=0}^{\infty} I(T_n \leq t) \cdot X_n \geq h \right\} = \inf \left\{ t : Z(t) \geq h - X_0 \right\}.$$

If X_0 is deterministic and if it is possible to calculate the degradation distribution then with that, the distribution of the first passage time is also given. Here, we use the relation

$$P(Z^h > t) = P(Z(t) < h - X_0). \quad (12)$$

Special examples as well as parameter estimates for these models can be found in Refs 34–36.

REFERENCES

1. Meeker WQ, Escobar L. Statistical methods for reliability data. New York: Wiley; 1988.
2. Lu CJ, Meeker WQ. Using degradation measures to estimate a time-to-failure distribution. Technometrics 1993;35:161–174.
3. Bagdonavicius V, Bikelis A, Kazakevicius V, et al. Analysis of joint multiple failure mode and linear degradation data with renewals. J Stat Plan Inference 2007;137:2191–2207.
4. van Noortwijk JM. A survey of the application of gamma processes in maintenance. Reliab Eng Syst Saf 2009;94:2–21.

5. Bagdonavicius V, Nikulin M. Estimation in degradation models with explanatory variables. *Lifetime Data Anal* 2001;7:85–103.
6. Voinov VG, Nikulin MS. Unbiased estimators and their applications. Volume 2, Multivariate case. Dordrecht: Kluwer Academic Publishers; 1996.
7. Sumita U, Shanthikumar JG. A class of correlated cumulative shock models. *Adv Appl Probab* 1985;17:347–366.
8. Gut A, Husler J. Realistic variation of shock models. *Stat Probab Lett* 1999;74:187–204.
9. Kahle W. Simultaneous confidence regions for the parameters of damage processes. *Stat Papers* 1994;35:27–41.
10. Kahle W, Lehmann A. Parameter estimation in damage processes: dependent observations of damage increments and first passage time. In: Kahle W, *et al.*, editors. *Advances in stochastic models for reliability, quality and safety*. Boston (MA): Birkhauser; 1998. pp. 139–152.
11. Doksum KA, Normand ST. Models for degradation processes and event times based on gaussian processes. In: Jewell NP, *et al.*, editors. *Lifetime data: models in reliability and survival analysis*. Dordrecht: Kluwer Academic Publishers; 1996. pp. 85–91.
12. Whitmore GA. Estimating degradation by a Wiener diffusion process subject to measurement error. *Lifetime Data Anal* 1995;1:307–319.
13. Whitmore GA, Schenkelberg F. Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Anal* 1997;3:27–45.
14. Whitmore GA, Crowder MJ, Lawless JF. Failure inference from a marker process based on a bivariate Wiener model. *Lifetime Data Anal* 1998;4:229–251.
15. Last G, Brandt A. Marked point processes on the real line. New York: Springer; 1995.
16. Anderson P, Borgan O, Gill R, *et al.* Statistical models based on counting processes. New York: Springer; 1993.
17. Cox DR. Some statistical methods connected with series of events. *J R Stat Soc [ser B]* 1955;17:129–164.
18. Cramer H. On streams of random events. *Skand Aktuar Tidskr Suppl* 1969;85:13–23.
19. Grandell J. Aspects of risk theory. New York: Springer; 1991.
20. Bremaud P. Point processes and queues. New York: Springer; 1981.
21. Grandell J. Mixed Poisson processes. London: Chapman & Hall; 1997.
22. Sobczyk K. Stochastic models for fatigue damage of materials. *Adv Appl Probab* 1987;19:652–673.
23. Esary JD, Marshall AW, Proshan F. Shock models and wear processes. *Ann Probab* 1973;1:627–649.
24. Feng W, Adachi K, Kowada M. Optimal replacement under additive damage in a Poisson random environment. *Commun Stat Stoch Mod* 1994;10:679–700.
25. Shaked M. Wear and damage processes from shock models in reliability theory. *Reliability theory and models*. New York: Academic Press; 1984. pp. 43–64.
26. Aven T, Jensen U. Stochastic models in reliability. New York: Springer; 1998.
27. Wendt H. Parameterschätzungen für eine Klasse doppelt stochastischer Poisson Prozesse bei unterschiedlichen Beobachtungsinformationen [PhD Thesis]. 1999.
28. Alsmeyer G, Gut A. Limit theorems for stopped functionals of Markov renewal processes. *Ann Inst Stat Math* 1999;51:369–382.
29. Anderson KK. Limit theorems for general shock models with infinite mean intershock times. *J Appl Probab* 1987;24:449–456.
30. Anderson KK. A note on cumulative shock models. *J Appl Probab* 1988;25:220–223.
31. Gut A. Cumulative shock models. *Adv Appl Probab* 1990;22:504–507.
32. Gut A. Mixed shock models. *Bernoulli* 2001;7:541–555.
33. Gut A, Husler J. Extreme shock models. *Extremes* 1999;2:93–305.
34. Kahle W, Wendt H. Statistical analysis of some parametric degradation models. In: Nikulin M, Commenges D, Huber C, editors. *Probability, statistics and modelling in public health*. New York: Springer Science + Business Media; 2006. pp. 266–279.
35. Wendt H, Kahle W. On parameter estimation for a position-dependent marking of a doubly stochastic poisson process. In: Nikulin MS, *et al.* editors. *Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life*. Birkhauser book series. Statistics for Industry and Technology; 2004. pp. 473–468.
36. Wendt H, Kahle W. On a cumulative damage process and resulting first passage times. *Appl Stoch Mod Bus Ind* 2004;20:17–26.

DANTZIG–WOLFE DECOMPOSITION

WILLIAM CHUNG
Department of Management
Sciences, City University of
Hong Kong, Hong Kong

DANTZIG–WOLFE DECOMPOSITION METHOD

Consider the following linear programming (LP) problem:

Block-Angular Linear Programming (BALP)

$$\begin{aligned} \text{Min}_{x_1, \dots, x_m} \quad & c_1^T x_1 + c_2^T x_2 + \dots + c_m^T x_m \\ \text{s.t.} \quad & L_1 x_1 + L_2 x_2 + \dots + L_m x_m \geq h \quad (\beta) \\ & B_1 x_1 \geq b_1 \quad (\pi_{S1}) \\ & B_2 x_2 \geq b_2 \quad (\pi_{S2}) \\ & \vdots \\ & B_m x_m \geq b_m \quad (\pi_{Sm}) \\ & x_j \in R_+^{n_j} \text{ for } j = 1, \dots, m, \end{aligned}$$

where all vectors are column vectors and the superscript T indicates the transpose of a vector or matrix. The vectors c_j, h , and b_j and the matrices L_j and B_j have appropriate dimensions. $\sum_{j=1}^m L_j x_j \geq h$ represents the *linking constraints* (also known as *coupling constraints* or *global constraints*), and $B_j x_j \geq b_j$ is the *block constraints* (or *local constraints*). β and π_{S_j} are the dual price vectors of the linking constraints and the block constraints, respectively. The linking constraints involving variables from different blocks drastically complicate the solution of the problem and prevent its solution by blocks. In order to improve computational efficiency by exploiting the sparsity and the special structure of the BALP, Dantzig and Wolfe [1,2] devised the first decomposition

method to decompose the large-scale BALP into m small subproblems (SPs) and a so-called *master problem*. Applying the corresponding decomposition algorithm, the SP and the master problem are to be solved iteratively until the solution to the BALP is found.

The main idea of the Dantzig–Wolfe (DW) decomposition method is to relax the linking constraints from the BALP and consider them at a superordinate level in the master problem. The resulting problem can be decomposed into m small SPs and is often solved more efficiently than BALP. It is worth mentioning that if $m = 1$ and the resulting problem has a special structure such as a network structure, the computation of the SP solution might be very efficient.

Minkowski's Representation Theorem

Define the polyhedron of local constraints by $S_j = \{x_j | x_j \geq 0, B_j x_j \geq b_j\}$. Let $X_j^{\bar{P}_j} = (x_j^1, x_j^2, \dots, x_j^{\bar{P}_j})$ and $X_{rj}^{\bar{R}_j} = (r_j^1, r_j^2, \dots, r_j^{\bar{R}_j})$ denote the matrix of all the extreme points and extreme rays in S_j , respectively. Hence, the superscripts $\bar{P}_j$ and $\bar{R}_j$ indicate the number of extreme points and extreme rays in S_j , respectively. That is, $X_j^{\bar{P}_j}$ and $X_{rj}^{\bar{R}_j}$ consist of $\bar{P}_j$ and $\bar{R}_j$ columns, respectively. Through Minkowski's Representation Theorem [3], we can show that any point $\hat{x}_j$ in S_j may be expressed as a convex combination of its extreme points plus a nonnegative combination of the extreme rays. That is,

$$\hat{x}_j = X_j^{\bar{P}_j} \lambda_j^{\bar{P}_j} + X_{rj}^{\bar{R}_j} \delta_j^{\bar{R}_j}, \text{ with}$$

$$\sum_{p=1}^{\bar{P}_j} \hat{\lambda}_j^p = 1, \hat{\lambda}_j^p \geq 0, \hat{\delta}_j^r \geq 0,$$

$$p = 1, \dots, \bar{P}_j, r = 1, \dots, \bar{R}_j,$$

where $\lambda_j^{\bar{P}_j} = (\hat{\lambda}_j^1, \dots, \hat{\lambda}_j^{\bar{P}_j})^T$ and $\delta_j^{\bar{R}_j} = (\hat{\delta}_j^1, \dots, \hat{\delta}_j^{\bar{R}_j})^T$.

Reformulation

Using the Minkowski's Representation Theorem, the BALP can be reformulated into the following equivalent problem, the so-called *full master problem*:

$$\begin{aligned}
& \text{Min}_{\substack{\bar{P}_j, \bar{R}_j \\ \lambda_j^j, \delta_j^j}} \sum_{j=1}^m c_j^T \left(X_j^{\bar{P}_j} \lambda_j^{\bar{P}_j} + X_{r_j}^{\bar{R}_j} \delta_j^{\bar{R}_j} \right) \\
& \text{s.t.} \sum_{j=1}^m L_j \left(X_j^{\bar{P}_j} \lambda_j^{\bar{P}_j} + X_{r_j}^{\bar{R}_j} \delta_j^{\bar{R}_j} \right) \geq h \quad (\beta) \\
& e^{\bar{P}_j^T} \lambda_j^{\bar{P}_j} = 1 \quad (\pi_j) \\
& \text{for } j = 1, \dots, m \\
& \text{all } \lambda_j^{\bar{P}_j} \geq 0, \delta_j^{\bar{R}_j} \geq 0,
\end{aligned}$$

where $e^{\bar{P}_j} = (1, 1, \dots, 1)^T$, the unit $\bar{P}_j$ -vector, and π_j is the dual price value corresponding to the j th convexity constraint $e^{\bar{P}_j^T} \lambda_j^{\bar{P}_j} = 1$.

The Decomposition Algorithm

It is not necessary to write the complete specification of the *full master problem* in advance because the *full master problem* could have a great many variables. Rather, the columns of the *full master problem* should be generated as they are needed. Following is a discussion of the mechanism for generating these columns.

Consider iteration k of the algorithm and suppose a subset of the extreme points, P_j columns, and the extreme rays, R_j columns, has been generated. Obviously, $P_j + R_j < \bar{P}_j + \bar{R}_j$. Thus, the matrices of the extreme points and rays become $X_j^{P_j}$ and $X_{r_j}^{R_j}$, respectively. With these matrices, what is called a *restricted master program* (RMP) is solved, representing an approximation to the original BALP (or the full master problem):

$$\begin{aligned}
& [\text{RMP}^k] \\
& \text{Min}_{\substack{P_j, R_j \\ \lambda_j^j, \delta_j^j}} \sum_{j=1}^m c_j^T \left(X_j^{P_j} \lambda_j^{P_j} + X_{r_j}^{R_j} \delta_j^{R_j} \right)
\end{aligned}$$

$$\begin{aligned}
& \text{s.t.} \sum_{j=1}^m L_j \left(X_j^{P_j} \lambda_j^{P_j} + X_{r_j}^{R_j} \delta_j^{R_j} \right) \geq h \quad (\beta^k) \\
& e^{P_j^T} \lambda_j^{P_j} = 1 \quad (\pi_j^k) \\
& \text{for } j = 1, \dots, m \\
& \text{all } \lambda_j^{P_j} \geq 0, \delta_j^{R_j} \geq 0,
\end{aligned}$$

where β^k and π_j^k are the dual prices of the corresponding constraints at iteration k .

Note that the adjective “*restricted*” refers to the fact that the RMP ^{k} does not contain all extreme points and rays in S_j , but only the subset of extreme points and rays.

Consider the dual problem of the RMP ^{k} , we have

$$\begin{aligned}
& \text{Max}_{\beta^k, \pi_j^k} \beta^k h + \pi_1^k + \pi_2^k + \dots + \pi_j^k \\
& \text{s.t.} \quad \beta^k L_j X_j^{P_j} + \pi_j^k \leq c_j^T X_j^{P_j} \\
& \quad \text{for } j = 1, 2, \dots, m \\
& \quad \beta^k L_j X_{r_j}^{R_j} \leq c_j^T X_{r_j}^{R_j} \\
& \quad \text{for } j = 1, 2, \dots, m \\
& \quad \beta^k \geq 0,
\end{aligned}$$

Assume that we find the optimal solutions for the RMP ^{k} , and, β^k and π_j^k are the associated optimal solution to the dual problem. It is easily shown that if $\beta^k L_j x_j + \pi_j^k \leq c_j^T x_j$ and $\beta^k L_j r_j \leq c_j^T r_j$ for all x_j and r_j in S_j respectively, then the convex combinations of $X_j^{P_j}$ and $X_{r_j}^{R_j}$ are optimal in the original problem, BALP. It also implies that no other negative reduced costs column exists for improving the RMP ^{k} . However, if there exists a vector x_j^{k+1} (or r_j^{k+1}) in S_j for any j , such that

$$\begin{aligned}
& \beta^k L_j x_j^{k+1} + \pi_j^k > c_j^T x_j^{k+1} \\
& \beta^k L_j r_j^{k+1} > c_j^T r_j^{k+1}; \quad \text{i.e.,} \\
& c_j^T x_j^{k+1} - \beta^k L_j x_j^{k+1} - \pi_j^k < 0 \\
& c_j^T r_j^{k+1} - \beta^k L_j r_j^{k+1} < 0,
\end{aligned}$$

then adding this new point to the RMP^k may result in an improved solution. To determine whether such x_j^{k+1} (or r_j^{k+1}) exists in S_j , for each j we need to solve the SP.

$$\text{SUB}_j^k : \text{Min}_{x_j \in S_j} (c_j^T - \beta^{k-1T} L_j) x_j.$$

Let $x_{S_j}^k$ (or $r_{S_j}^k$ if SUB_j^k is unbounded) be the solution of the SUB_j^k . If SUB_j^k is unbounded, an extreme ray, $r_{S_j}^k$, with $c_j^T r_{S_j}^k - \beta^{k-1T} (L_j r_{S_j}^k) < 0$ is to be found as a dual extreme ray. For any j , if $c_j^T x_{S_j}^k - \beta^{k-1T} (L_j x_{S_j}^k) - \pi_j^{k-1} < 0$ (or $c_j^T r_{S_j}^k - \beta^{k-1T} (L_j r_{S_j}^k) < 0$ if SUB_j^k is unbounded), then the corresponding $x_{S_j}^k$ (or $r_{S_j}^k$) can be a new *proposal* column x_j^{k+1} (or r_j^{k+1}) to be added to the matrices $X_j^{P_j}$ (or $X_{r_j}^{R_j}$) in the “next” RMP^k , for an improved solution. Therefore, the columns of the full master problem are generated as they are needed. For details, the reader is directed to the article on column generation (see **Column Generation**).

Now we state the *DW decomposition algorithm*: A “null matrix” is a “matrix” with no columns.

Step 0. Set $k = 0$. Choose $\beta^0 = 0$ and $P_j = R_j = 0$. Set X_j^0 and $X_{r_j}^0$ to the null matrix, $j = 1, 2, \dots, m$.

Step 1. Increment $k \leftarrow k + 1$. Solve all the SUB_j^k .

Step 1.1. If $k = 1$ and any of the SUB_j^k is infeasible, then STOP: the original problem is infeasible; otherwise, go to Step 2.

Step 1.2. If $k > 1$ and for any j ,

$$\begin{aligned} & c_j^T x_{S_j}^k - \beta^{k-1T} (L_j x_{S_j}^k) - \pi_j^{k-1} < 0 \\ & \text{(or } c_j^T r_{S_j}^k - \beta^{k-1T} (L_j r_{S_j}^k) < 0 \text{ if } \text{SUB}_j^k \\ & \text{is unbounded),} \end{aligned}$$

go to Step 2; otherwise, STOP: the solution of RMP^{k-1} is an optimal solution.

Step 2. $P_j \leftarrow P_j + 1$ and place the solution $x_{S_j}^k$ in the matrices $X_j^{P_j} = [X_j^{P_j-1}, x_{S_j}^k]$ (or $R_j \leftarrow R_j + 1$ and place $r_{S_j}^k$ in $X_{r_j}^{R_j} = [X_{r_j}^{R_j-1}, r_{S_j}^k]$ if SUB_j^k is unbounded); and go to Step 3.

Step 3. Solve RMP^k . Record β^k, π_j^k . Go back to Step 1.

Initialization

The decomposition algorithm can be started with an initial set of proposals that can be found in Step 0 and Step 1.1. If for any of the SPs, SUB_j^1 is infeasible, then the original problem is infeasible. Otherwise, we can use these optimal values $x_{S_j}^1$ (or rays $r_{S_j}^1$) to generate an initial set of proposals to be used in the RMP [2,4].

However, the initial proposals may violate the linking constraints in the RMP^k . We can start solving the RMP^k using the same device employed for the general LP problem, called the *Phase One*. We formulate the following Phase-One RMP^k by introducing artificial variables and minimizing them.

$$\begin{aligned} & \text{Min}_{x_a} e^{\dim x_a^T} x_a \\ & \text{s.t. } \sum_{j=1}^m L_j (X_j^{P_j} \lambda_j^{P_j} + X_{r_j}^{R_j} \delta_j^{R_j}) + x_a \geq h \quad (\beta^k) \\ & e_j^{P_j^T} \lambda_j^{P_j} = 1 \quad (\pi_j^k) \quad \text{for } j = 1, 2, \dots, m \\ & \lambda_j^{P_j} \geq 0; \delta_j^{R_j} \geq 0 \quad \text{for } j = 1, 2, \dots, m \\ & x_a \geq 0. \end{aligned}$$

We also have the corresponding *Phase-One* $\text{Sub}_j^k : \text{Min}_{x_j \in S_j} (-\beta^{kT} L_j) x_j$. After using the DW decomposition algorithm to solve the Phase-One problem, we either have (i) $e^{\dim x_a^T} x_a > 0$, indicating that the original problem has no feasible solution, or (ii) $e^{\dim x_a^T} x_a = 0$, indicating that we have a set of proposals for which there is a convex combination that is feasible in the linking constraints, to start the decomposition.

On the other hand, we may use the “big M” method: artificial variables with large cost

coefficients (M) are added in each linking constraint of the RMP^k , like the one in *Phase-One RMP^k* . The corresponding objective function becomes $\sum_{j=1}^m c_j^T (X_j^{P_j} \lambda_j^{P_j} + X_{r_j}^{R_j} \delta_j^{R_j}) + M \times x_a$, and the constraint $x_a \geq 0$ is to be added.

Finite Convergence

The decomposition algorithm terminates in a finite number of iterations, yielding a solution of the original problem (assuming it is a non-degeneracy case). At each iteration, the SUB_j^k provides extreme points $x_{S_j}^k$ (or extreme rays $r_{S_j}^k$) as proposals to the RMP^k . Since the SP is a LP model that consists of a finite number of extreme points and rays, the algorithm converges in a finite number of iterations. Moreover, the stopping condition indicates that there is no new proposal (extreme point or extreme ray) improving the original problem. See Ref. 2 for details.

THE MYTH OF COMPUTATIONAL EFFICIENCY

Apart from its theoretical elegance, it was believed that the DW decomposition algorithm held great promise for large-structured LPs, where for instance, the SP may be further decomposed into much smaller SPs according to its special structure, similar to the block-angular one. However, early attempts at implementation and application were generally not encouraging.

Beale *et al.* [5] reported the first practical application of DW decomposition to an oil field–operation problem. The algorithm was implemented as an internal extension to the commercial LP/90/94 mathematical programming software. The software had a theoretical capacity of 100 blocks, 100,000 constraints in total, 1000 constraints in any block, and an unlimited number of variables. Mainly, the DW method did not perform better than the rival simplex method.

A number of practical experiences were summarized in Ref. 6. Based on the selected studies' documents, for instance [4,7–9], generally, the greater the number of proposals used to initialize the algorithm, the faster the solution can be found. However, the DW

method is time consuming and converges slowly toward an optimal solution. Hence, the DW method's primary advantage may be when it is the only algorithmic choice for a problem that is too large to be solved in one piece.

Long-Tail Convergence

Dirickx and Jennergren [6] also summarized that the algorithm is prone to rapid initial convergence but “tails off” upon approaching optimality, that is, while a near optimal solution is usually approached considerably fast, only little progress per iteration is made close to the optimum. Consequently, time-savings from the rapid initial convergence tend to be offset by the number of the tail-off cycles required. This phenomenon is called the *tailing-off effect* or *long-tail convergence*.

There is an intuitive assessment of this phenomenon; however, a theoretical understanding has been only partly achieved. Adler and Ülkücü [10] used combinatorial arguments to explain this tail-off behavior. They used the concept of diameter of polytopes, which is the maximal number of affinely independent points contained in a polytope. Obviously, the diameter is a monotone increasing function of the number of extreme points. Hence, they argued that the number of simplex steps might be proportional to the diameter of the polytope defining the feasible region. Since it is well known that the decomposition algorithm is essentially the simplex method applied to the MP (master program), it is expected to have many more cycles than simplex iterations required for the original problem.

Nazareth [11] used a set of LP problem examples consisting of a few constraints and variables to identify the numerical difficulties that can generally occur when applying the DW method. Nazareth showed that even when stable techniques such as a stable refactorization of the corresponding basis matrix are employed, the columns of the computed MP could differ substantially from those of the true one. Moreover, it is shown that errors in the extreme points (or rays) defining the columns of the computed master can be correlated with errors in solutions obtained from this computed master, thus perturbing the

optimal solution. Hence, this optimal solution must be relatively insensitive to perturbations of the initial data. Later, Nazareth [12] noted that because of the changes of primal and dual variables are applied iteratively, a certain amount of “cleaning up” is to be expected. Kim and Nazareth [13] discussed that finding an appropriate convex combination of existing proposals might be held back by the possible “complex combinatorial structure” of the faces of the polyhedron defined by the SPs.

While the question of numerical stability was first discussed by Nazareth [11], although not in the context of convergence, Ho [14] explained that the implicit error bounds might imply a tolerance for convergence below which further apparent improvements should be considered as noise. As the exact error analysis for this kind of algorithm is impossible, the most promising approach for further investigating this question is to conduct controlled experiments using various degrees of numerical accuracy in the arithmetic and comparing the resulting convergence behavior. If inaccuracy causes (at least in part) long convergence or nonconvergence, then specific techniques for improvement may be designed. In Ref. 14, certain procedures based on the method of iterative refinement are proposed. However, as pointed out earlier, what the decomposition approach needs is a way to clean-up the data periodically, that is, an analogy to basis reversion for the revised simplex method.

However, Lübbecke and Desrosiers [15] remarked that tailing-off also occurs when columns are computed exactly (e.g., by use of combinatorial algorithms).

Swoveland [16] suggested a modification of the column-generation operation in the DW decomposition. Instead of the usual procedure of solving one or more SPs at each iteration, it is shown how the SPs may be solved parametrically to minimize the immediate improvement in the objective’s value in the MP, rather than to minimize the reduced cost of the entering column. It is hoped that in some cases, the use of the suggested procedure may improve the slow rates of convergence common in decomposition algorithms.

Column Dropping

One may consider that column dropping from the MP would improve computational efficiency. Dantzig [17] described three strategies for deciding how long proposals should persist in the MP: retain all proposals; delete all nonbasic proposals immediately (apart from those just generated); or keep all proposals until a certain proposal capacity is reached, and then purge the proposal with the worst reduced cost. The question of when to delete proposals from the MP is not entirely clear. Beale *et al.* [5] chose the first option. The best reported result for a problem found that the DW method performed marginally better than the rival simplex method. For *nonlinear programming (NLP)*, Murphy [18] provided two conditions under which columns may be dropped from the restricted master. However, the computational results of [19] showed virtually no difference in the measures of computational efficiency when all columns are retained and when only basic columns are retained.

BOUNDS ON THE OPTIMAL SOLUTION VALUE

The tail-off convergence behavior indicates that the number of iterations needed to solve the problem by the DW method could be much larger than the number required by an application of the ordinary simplex method. Hence, it may be desirable to terminate before optimality is reached. Indeed, providing bounds on the optimal solution value is one of the attractive features of the DW method.

Let z^* be the optimal solution value of the original problem. Suppose that at iteration k , one or more of the SUB_j^k have unbounded optimal solutions, then no bound on z^* can be obtained. However, if all SUB_j^k have bounded optimal solutions, we can have $z_M^k = \beta^{kT} h + \sum_{j=1}^m \pi_j^k$ to be the optimal solution value of the “restricted” MP. Obviously, for the minimization problem, $z^* \leq z_M^k$. On the other hand, from the results of the SUB_j^k ’s dual solution, we have $\pi_{Sj}^{kT} B_j + \beta^{k-1T} L_j \leq c_j^T$

which provides the dual feasible solution $(\pi_{S_j}^k, \beta^{k-1})$ to the original BALP. Hence, we have $\beta^{k-1^T} h + \sum_{j=1}^m \pi_{S_j}^{kT} b_j \leq z^*$. Let $z_S^k = \sum_{j=1}^m \pi_{S_j}^{kT} b_j - \pi_j^{k-1} = \sum_{j=1}^m c_j^T x_{S_j}^k - \beta^{k-1^T} (L_j x_{S_j}^k) - \pi_j^{k-1}$, and we get $z_M^k + z_S^k \leq z^* \leq z_M^k$, which may be used to construct a simple stopping rule: terminate the algorithm if the gap between bounds is smaller than the preset tolerance constant.

ADVANCED IMPLEMENTATION FOR COMPUTATIONAL EFFICIENCY

One may consider that a more advanced implementation may enhance the computational efficiency of the DW method. However, the results of Refs 20 and 21 resubstantiated that it was unlikely for the DW method to be significantly more efficient than the simplex approach, whenever the latter can produce a solution at an acceptable cost. Hence, the DW method should be used as a modeling approach, similar to its application in Ref. 22, rather than simply as a solution tool.

Ho [23] provided more computational results on a diverse collection of problems from real applications in order to clarify the tailing-off convergence phenomenon. A *relative primal-dual gap* of 0.1% (or smaller) was used for all cases as a stopping criterion, and it was observed that more iterations were expected if a much more stringent tolerance was used. However, most test problems required a surprisingly small number of iterations. Therefore, there is a question of whether these relatively low numbers of iterations observed were simply a result of suboptimality. In fact, there may not be any simple answer because of numerical error propagation. Ho also reported that the underlying numerical accuracy might not even guarantee a resolution of 0.1% in many situations. Moreover, such a tolerance is usually considered acceptable in practice.

Ho *et al.* [24] reported the results of their computational experience of the DW method implemented in *parallel* computers using the following decomposition strategies.

- A1. Start with RMP idle. Solve SUB_j, $j = 1, \dots, m$. SUB_j becomes idle when done. When all SUB_j are done, solve RMP with all S_j idle. Repeat until optimal.
- A2. With RMP idle, solve SUB_j, $j = 1, \dots, m$ until the first SUB_j is done with a new proposal. Solve RMP with all SUB_j idle. Repeat until optimal.
- A3. With RMP idle, solve SUB_j, $j = 1, \dots, m$ until the first proposal generation procedure is triggered by the first feasible solution $x_{S_j}^k$ to the SUB_j that satisfies $(c_j^T - \beta^{k-1^T} L_j) x_{S_j}^k - \pi_j^{k-1} < 0$. The corresponding proposal is generated [20]. Solve RMP with all SUB_j idle. Repeat until optimal.
- A4. With RMP idle, solve SUB_j, $j = 1, \dots, m$ until the first proposal. Continue SUB_j if not done while solving RMP. Whenever RMP is done, check for new proposals. Resolve. Whenever SUB_j is done, check for new prices from RMP. Resolve. Repeat until optimal.

While A1 optimizes all SPs, it was not necessarily true that it would converge fastest. The computational results showed that A2 increased the utilization of the processors, and the total run time was significantly inferior to the standard version (A1) in most of the cases tested. These results implied that A3 would be even less interesting. Hence, the only remaining option was A4. It was reported that the performance of A4 in keeping all the processors busy was improved.

RECENT COMPUTATIONAL EXPERIENCE

Gnanendran and Ho [25] investigated the strategies for improving efficiency in distributed DW decomposition by better balancing the load between the MP and SP processors. They reported computational experience on an Intel iPSC/2 hypercube multiprocessor with test problems having dimensions up to about 30,000 rows, 87,000 columns, and 200 coupling constraints. It was found that the parallel efficiency

of the distributed approach was critically dependent on the duration of the inherently serial master phase relative to that of the bottleneck SP. That is, the efficiency is bound by a factor that depends on the ratio of the duration of the master phase to the duration of the bottleneck SP. This ratio, in turn, is dependent on the number and “tightness” of the coupling constraints. Unlike speedup, which can often be improved by scaling the problem size in step with the number of processors, efficiency improvements must come about by better load balancing, such as through the partial overlapping of the MP and SP processes.

Rosen and Maier [26] implemented the DW method for a large-scale BALP in parallel, which was computationally tested on both a shared-memory vector multiprocessor (CRAY-2) and a local-memory hypercube (NCUBE/seven) with 64 processors. Computational results for problems with as many as 24,000 rows and 74,000 columns (1024 blocks and 1.4 million nonzero elements) were presented. A problem of this size was solved on the NCUBE in less than four minutes and the CRAY-2 in 37 s. Sundarraj *et al.* [27] applied DECUBE based on DECOMP to handle a power systems operations problem, which is a BALP with up to 37,200 rows, 120,000 columns, and 30 SPs. DECUBE written in Fortran employed the subroutines of DECOMP as building blocks which was a serial implementation of the DW method [20].

Tebboth [28] developed a parallel implementation of the DW method to solve problems up to 83,500 rows, 83,700 columns, and 622,000 nonzero elements. One instance of the problems was solved in 63% of the time taken by a commercial implementation of the simplex method.

The most recent application can be found in Ref. 29. With the planning horizon of a refinery ranging from 2–300 days, the problem consisted of up to 1 million constraints, 2.5 million variables, and 59 million nonzeros in the constraint matrix. Large instances became computationally challenging for generic state-of-the-art LP solvers, such as CPLEX. The computational results showed that the interior point algorithm and the block coordinate-descent

with a particular order of the variables were more effective than the DW method.

PRICE-DIRECTIVE DECOMPOSITION

From an economic viewpoint, the DW decomposition algorithm of the profit maximization problems can be interpreted as a process of price-directive coordination of the SPs, which are allowed to optimize their own objectives while vying for the shared resources. The MP plays the role of a coordinator who sets the prices (dual prices of the linking constraints), β^{k-1} , on the shared resources, h . Given such price levels, each SP seeks to optimize its own objective while “paying” for the consumption of shared resources. This process is the SP SUB $_j^k$, whose solution gives rise to a proposal (proposed consumption) $L_j x_j$. If the corresponding value of a stopping condition is positive, it signals to the coordinator that the proposal can be used to improve on the system-wide objective. After collecting all such profitable proposals from this round of coordination, the coordinator optimizes the system-wide objective over all available proposals. By taking convex combinations of proposals from each SP and requiring such combinations to satisfy the balance of shared resources, the coordinator can maintain system-wide feasibility while improving on the objective. This is the process involved in the “next” MP.

Optimal Allocation Model

Considering each sequence of MPs and SPs as a cycle in the coordinating process, equilibrium prices β^* on the shared resources will be obtained after a finite number of k iterations. At such prices, no SP will be able to generate a profitable proposal. The optimal activity levels for each SP are given by the convex combination of proposed solutions: $X_j^k \lambda_j^k$. This illustrates the well-known fact that decentralization cannot be accomplished by price direction alone. Generally, given only β^* , it is impossible for an SP to recover the optimal solution $X_j^k \lambda_j^k$ by itself. Nonetheless, having the SP activities dictated by the coordinator according to $X_j^k \lambda_j^k$ obviously detracts from the spirit of decentralized planning. Fortunately,

a better alternative exists. From the optimal restricted MP, the coordinator can compute the optimal allocation of the shared resources to each SP as $h_j^* = L_j X_j^k \lambda_j^k$. Given this allocation, each SP can then proceed to determine its own activities by solving the *Optimal Allocation Model* [30]:

$$\begin{aligned} & \text{SUB}_j^{k*} \\ & \text{Min}_{x_j} c_j^T x_j \\ & \text{s.t. } L_j x_j \geq h_j^* = h - \sum_{i \neq j} L_i X_i^k \lambda_i^k \\ & \quad B_j x_j \geq b \\ & \quad x_j \geq 0. \end{aligned}$$

This way, the SPs are accorded more procedural autonomy.

Optimal Degree of Decentralization

Since SPs can be aggregated, there is a choice in the degree of decentralization. Given a maximum of m SPs as in the aforementioned development, it is possible to decompose with anywhere from 1 to m SPs. Let us examine first the extreme case of using a single SP. There is a single convexity constraint in the MP. Moreover, the SP can still be decomposed and solved SP by SP as before, because their constraints do not interact. The only difference is that once the SP activities x_j are determined, an aggregate proposal $\sum_{j=1}^m L_j X_j^k$ is formed and submitted to the MP. This apparent decomposability of the aggregate SP may lead one to speculate that the formulations with 1 or m convexity constraints are equivalent. Extending the same argument to any other grouping of SPs implies that the number of convexity constraints is a direct measure of the degree of decentralization.

We can now address the question of what effect the degree of decentralization has on the process of price-directive resource allocation. Early literature [31,32, Chapter 3] was inconclusive mainly because of the emphasis on algorithmic performance and the lack of tools and models for significant empirical results.

Ho and Loute [30] reported some empirical results related to this issue. They focused

on the number of cycles required by the DW method as a measure of the amount of coordination necessary to achieve price-directive resource allocation. The results supported the general observation that the amount of coordination decreases with the increasing degree of decentralization, that is, the optimal degree of decentralization is the maximum number of SPs.

A Variant DW Method

Balas [33] developed an interesting infeasibility-pricing approach to apply the DW method by which the SPs were first solved. The infeasibility of the resulting solution for the linking program was used to generate a price-adjustment process (much like a market mechanism). This gradually corrected the objective functions of the SPs, until either some optimal solutions to the latter yielded a feasible solution to the MP (then the problem is solved), or evidence was obtained that the problem had no feasible solution.

ORGANIZATIONAL STUDIES

Following the economic viewpoint of price-directed coordination, the decomposition of mathematical programs have been developed extensively as models for organizational design and resource allocation in decentralized organizations [34–39]. Generally, we can conceptualize the allocation problem as a multilevel managerial coordination procedure, involving local (divisional) and global (organization-wide) resources, in which informational autonomy is to be maintained. That is, a coordination of resource usage by relatively autonomous divisions is to be effected in these models without any one agent in the organization accumulating complete knowledge of the technical resource transformations, payoffs or cost coefficients, and detailed plans that divisions utilize in converting resources to useful ends.

Empirical Investigation

Beside the above conceptual models, Christensen and Obel [40] conducted a simulation study about the coordination and

decentralized planning mechanism based on the DW decomposition algorithm. Almost at the same time, Moore [41] reported a laboratory empirical investigation into the implementation problems of applying the decomposition of mathematical programs in decentralized organizations. An example of the results was that high decision-time pressure would not worsen decision making.

Transfer Pricing

On the other hand, Burton *et al.* [42] studied the transfer pricing (or price coordination) in the context of the decentralized decision making process in the firm. They developed the analogy of the resource allocation versus transfer pricing through decomposition methods. Later, Burton and Obel [43] investigated the efficacy of price-and/or budget-planning approaches. They showed that for a changing environment, historical information (i.e., a past solution) was generally inferior to a combination of less precise information on current transfer prices, reasonable rules of thumb, and capacity constraints. However, more current information on prices and budgets yielded better plans. Further, they found that the transfer pricing yielded more nearly optimal plans in the first few iterations than a combination of prices and budgets, which in turn, was more efficient than budgets alone.

STAIRCASE LINEAR PROGRAMMING

Staircase-structured linear programs can naturally be found in the study of multi-period [44] and multistage problems, and can be formulated with the following basic structure.

$$\begin{aligned}
 & \text{Min}_{x_j \in R_j} c_1^T x_1 + c_2^T x_2 + \cdots + c_{m-1}^T x_{m-1} + c_m^T x_m \\
 & \text{s.t. } A_1 x_1 \geq b_1 \\
 & \quad B_1 x_1 + A_2 x_2 \geq b_2 \\
 & \quad \quad B_2 x_2 + A_3 x_3 \geq b_3 \\
 & \quad \quad \quad \vdots \\
 & \quad \quad \quad B_{m-1} x_{m-1} + A_m x_m \geq b_m \\
 & \quad x_j \in R_+^{n_j}.
 \end{aligned}$$

According to the DW terminology for block-angular systems, a multistage linear program is defined with linking variables that connect consecutive stages. Optimality conditions for the composite problem are partitioned into local and linking conditions. Hence, *nested decomposition* techniques apply the basic DW algorithm to a sequence of SPs, which were discussed in the last section of Ref. 1. That is, when the DW decomposition scheme is applied with the first stage as the master, the SP is also a multistage linear program with one stage less. The same decomposition is then applied to the SP, giving rise to a nested decomposition scheme in which each stage acts as an MP for the following stage and an SP for the preceding. Optimizing a single-stage problem implies that the corresponding “local” optimality conditions are satisfied. [45–48] further developed the DW algorithm for staircase LPs with more discussions on bounds, initialization, convergence, Phase-Two, and computational experience.

In spite of these efforts, the nested decomposition techniques inherited the computational inefficiency of the DW method.

NONLINEAR PROGRAMMING PROBLEMS

Since this chapter is used specifically to describe the DW method of LP, other related topics of the DW method are given briefly.

Concave Programs/NLP

Dantzig and Wolfe [2] mentioned that if the cost function $c_j^T x_j$ was replaced by the convex function $f_j(x_j)$, the decomposition algorithm would be unchanged, except that the SPs become $\text{Min}_{x_j \in S_j} [f_j(x_j) - \beta^k L_j x_j]$. In fact, the DW method for concave programs can be viewed as an inner linearization of the *RMP* followed by a relaxation (the SP). Holloway [49] presented an approach that allows selective inner linearization of functions and utilizes a generalized pricing problem. It allows considerable flexibility in the choice of functions to be approximated using inner linearization. Using the generalized pricing problem guarantees that strict improvement in the value of the objective function is

achieved at each iteration and provides a termination criterion that is both necessary and sufficient for optimality. A specialization of this procedure results in an algorithm that is a direct extension of the DW decomposition algorithm. Other extensions and modifications of the DW method in the context of non-LP can be found in Refs 50 and 51 for nonconvex programming, and in Ref. 52 for providing an additional column and cut in the MP.

DW and Simplicial Decomposition

Von Hohenbalken [53] presented a simplicial decomposition method for non-LP problems and mentioned that the simplicial decomposition method belongs to a class of decomposition methods in another sense, namely, those descending from the DW decomposition principle for linear programs. Patriksson [54, Chapter 9] brought the simplicial decomposition method closer to the DW method by dualizing linking constraints in the SP with prices provided by the MP. Their works showed the close relationship between the DW method and the simplicial decomposition method in LP and NLP problems. (see the section titled “Network flows” in this encyclopedia).

Mixed-Integer Program, Branch-and-Bound, Branch-and-Price

Holm and Tind [55] presented a method on how to generalize the DW method to integer programming. Regardless of the choice between the two main solution methods, branch-and-bound or cutting-plane, the price functions generated by these two methods are similar, both being polyhedral and concave. These price functions are used to evaluate the consumption of central resources and to penalize any attempts to violate the requirement that the result must be an integer. Wentges [56] developed a so-called *weighted DW decomposition* procedure to improve the quality of the dual prices of the linking constraints for linear mixed-integer programming. Barnhart *et al.* [57] discussed a branch-and-price method, in which the DW decomposition is used for bounding within a branch-and-bound method for specially structured

integer programs. Columns are generated for each node (of a branch-and-bound tree) at which an LP relaxation is solved.

After this generalized paper, a series of MIP applications has been reported.

DW for Variational Inequality Problems

Chung *et al.* [58] presented a DW decomposition algorithm for a class of equilibrium models, which integrate a price-dependent demand component with an LP supply component. This method allows for further decomposition of the supply side (e.g., by region or commodity). Existing demand–supply decomposition methods for economic equilibrium models, based on the cobweb algorithm, may fail to converge. They showed that the DW method converges in a finite number of iterations. Later, Chung *et al.* [59] developed a general procedure to extend any decomposition method for optimization models to multiregional equilibrium models and showed how to apply the procedure with the DW decomposition method. Fuller and Chung [60] modified and generalized the work of Chung *et al.* [58,59] to produce an extension of DW decomposition to a class of variational inequality problems (see the article on **Variational Inequalities** in this encyclopedia).

REFERENCES

1. Dantzig GB, Wolfe F. Decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
2. Dantzig GB, Wolfe F. The decomposition algorithm for linear programs. *Econometrica* 1961;29:767–778.
3. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. New York: John Wiley & Sons, Inc.; 1988.
4. Chvatal V. *Linear programming*. New York: Freeman; 1983.
5. Beale EML, Hughes PAB, Small RE. Experiences in using a decomposition program. *Comput J* 1965;8:13–18.
6. Dirickx YMI, Jennergren LP. *Systems analysis by multilevel methods*. New York: Wiley; 1979.
7. Kutcher GP. On the decomposing price-endogenous models. In: Goreux LM, Manne

- AS, editors. Multi-level panning: case studies in Mexico. Amsterdam: North-Holland; 1973. pp. 499–519.
8. Tcheng TH. Scheduling of a large forestry-cutting problem by linear programming decomposition [Ph.D. dissertation]. University of Iowa, 1966.
 9. Williams HP, Redwood AC. A structured programming model in the food industry. *Oper Res Q* 1974;25:517–527.
 10. Adler I, Ülkücü A. On the number of iterations in Dantzig-Wolfe decomposition. In: Himmelblau DM, editor. Decomposition of large scale problems. Amsterdam: North-Holland; 1973. pp. 181–187.
 11. Nazareth L. Numerical behavior of LP algorithms based upon the decomposition principle. *Linear Algebra Appl* 1984;57:181–189.
 12. Nazareth JL. Computer solution of linear programs. Oxford: Oxford University Press; 1987.
 13. Kim K, Nazareth JL. The decomposition principle and algorithms for linear programming. *Linear Algebra Appl* 1991;152:119–133.
 14. Ho JK. Convergence behavior of decomposition algorithms for linear programming. *Oper Res Lett* 1984;3:91–94.
 15. Lübbecke ME, Desrosiers J. Selected topics in column generation. *Oper Res* 2005;53:1007–1023.
 16. Swoveland C. A note on column generation in Dantzig-Wolfe decomposition algorithms. *Math Prog* 1974;6:365–370.
 17. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
 18. Murphy FH. Column dropping procedures for the generalized programming algorithm. *Manag Sci* 1973;19:1310–1321.
 19. O'Neill RP. Column dropping in the Dantzig-Wolfe convex programming algorithm: computational experience. *Oper Res* 1977;25:148–155.
 20. Ho JK, Louie E. An advanced implementation of the Dantzig-Wolfe decomposition algorithm for linear programming. *Math Prog* 1981;20:303–326.
 21. Ho JK, Louie E. Computational experience with advanced implementation of decomposition algorithms for linear programming. *Math Prog* 1983;27:282–290.
 22. Costa AM, Jennergen LP. Trade and development in the world economy: methodological features of project DYNAMICO. *J Pol Model* 1982;4:3–22.
 23. Ho JK. Recent advances in the decomposition approach to linear programming. *Math Prog Stud* 1987;31:119–128.
 24. Ho JK, Lee TC, Sundarraj RP. Decomposition of linear programs using parallel computation. *Math Prog* 1988;42:391–405.
 25. Gnanendran SK, Ho JK. Load balancing in the parallel optimization of block-angular linear programs. *Math Prog* 1993;62:41–67.
 26. Rosen JB, Maier RS. Parallel solution of large-scale, block-angular linear programs. *Math Prog* 1993;58:201–228.
 27. Sundarraj RP, Gnanendran SK, Ho JK. Distributed price-directive decomposition applications in power systems operations. *IEEE Trans Power Syst* 1995;10:1350–1360.
 28. Tebbboth JR. A computational study of Dantzig-Wolfe decomposition [Ph.D. Thesis]. University of Buckingham, 2001.
 29. Alabi A, Castro J. Dantzig-Wolfe and block coordinate-descent decomposition in large-scale integrated refinery-planning. *Comput Oper Res* 2009;36:2472–2483.
 30. Ho JK, Louie E. On the degree of decentralization in price-directive resource allocation. *Informatica* 1996;7:337–348.
 31. Madsen OBG. The connection between decomposition algorithms and optimal degree of decomposition. In: Himmelblau DM, editor. Decomposition of large scale problems. Amsterdam: North-Holland; 1973. pp. 241–250.
 32. Lasdon LS. Optimization theory for large systems. New York: MacMillan; 1970.
 33. Balas E. An infeasibility-pricing decomposition method for linear programs. *Oper Res* 1966;14:847–873.
 34. Baumol WJ, Fabian T. Decomposition, pricing for decentralization and external economies. *Manag Sci* 1964;11:1–32.
 35. Atkins D. Managerial decentralization and decomposition in mathematical programming. *Oper Res Q* 1974;25:615–624.
 36. Kydland F. Hierarchical decomposition in linear economic models. *Manag Sci* 1975;21:1029–1039.
 37. Sweeney DJ, Winkofsky EP, Roy P, *et al.* Composition vs. decomposition: two approaches to modeling organizational decision processes. *Manag Sci* 1978;24:1491–1499.
 38. Vakhutinsky IY, Dudkin LM, Ryvkin AA. Iterative aggregation—a new approach to the

- solution of large-scale problems. *Econometrica* 1979;47:821–841.
39. van de Panne C. Decentralization for multidivision enterprises. *Oper Res* 1991; 39:786–797.
 40. Christensen J, Obel B. Simulation of decentralized planning in two Danish organizations using linear programming decomposition. *Manag Sci* 1978;24:1658–1667.
 41. Moore JH. Effects of alternate information structures in a decomposed organization: a laboratory experiment. *Manag Sci* 1979;25:485–497.
 42. Burton RM, Damon WW, Loughridge DW. The economics of decomposition: resource allocation vs transfer pricing. *Decis Sci* 1974;5:297–310.
 43. Burton RM, Obel B. The efficiency of the price, budget, and mixed approaches under varying a priori information levels for decentralized planning. *Manag Sci* 1980;26:401–417.
 44. Glassey CR. Dynamic linear programs for production scheduling. *Oper Res* 1971;19:45–56.
 45. Glassey CR. Nested decomposition and multistage linear programs. *Manag Sci* 1973;20:282–292.
 46. Ho JK, Manne AS. Nested decomposition for dynamic models. *Math Prog* 1974;6:121–140.
 47. Jackson PL, Lynch DF. Revised dantzig-wolfe decomposition for staircase-structured linear programs. *Math Prog* 1987;39:157–179.
 48. Ruszczyński A. Parallel decomposition of multistage stochastic programming problems. *Math Prog* 1993;58:201–228.
 49. Holloway CA. A generalized approach to Dantzig-Wolfe decomposition for concave programs. *Oper Res* 1973;21:210–220.
 50. Nemhauser GL, Widhelm WB. A modified linear program for columnar methods in mathematical programming. *Oper Res* 1971;19:1051–1060.
 51. Thach PT, Konno H. A generalized Dantzig-Wolfe decomposition principle for a class of nonconvex programming problems. *Math Prog* 1993;62:239–260.
 52. Holmberg K, Jönsten K. A simple modification of Dantzig-Wolfe decomposition. *Optimization* 1995;34:129–145.
 53. von Hohenbalken B. Simplicial decomposition in nonlinear programming algorithms. *Math Prog* 1977;13:49–68.
 54. Patriksson M. Volume 23, Nonlinear programming and variational inequality problems: a united approach, Applied optimization series. Dordrecht: Kluwer Academic Publishers; 1999.
 55. Holm S, Tind J. A unified approach for price directive decomposition procedures in integer programming. *Discrete Appl Math* 1988;20:205–219.
 56. Wentges P. Weighted Dantzig-Wolfe decomposition for linear mixed-integer programming. *Int Trans Oper Res* 1997;4:151–162.
 57. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-Price: Column Generation for Solving Huge Integer Programs. *Oper Res* 1998;46:316–329.
 58. Chung W, Fuller JD, Wu YJ. A new demand-supply decomposition method for a class of economic equilibrium models. *Comput Econ* 2003;21:231–243.
 59. Chung W, Fuller JD, Wu YJ. A new decomposition method for multi-regional economic equilibrium models. *Oper Res* 2006;54: 643–655.
 60. Fuller JD, Chung W. Dantzig-Wolfe decomposition of variational inequalities. *Comput Econ* 2005;25:303–326.

DATA CLASSIFICATION AND PREDICTION

ERIC YANG
Center for Engineering in
Medicine, Massachusetts
General Hospital,
Charlestown, Massachusetts

IOANNIS P. ANDROULAKIS
Biomedical Engineering
Department, Rutgers
University, Piscataway,
New Jersey

INTRODUCTION

With the advent of the computer, there has been a corresponding rise in our ability to acquire and store data [1,2]. In many cases, the challenge lies not in the collection of the data, but in making intelligent inferences about the data. For instance, in biology, the use of gene expression microarrays have allowed researchers to measure the expression level of thousands of genes at once. But these gene expression measurements are not useful unless it is possible to identify the important genes that characterize a given phenomenon [3]. Therefore, with this rise in the amount of data that is available for us, there is the commensurate challenge in converting it into usable information. The methods used to process this information fall under the general rubric of “data mining” [4].

The overall goal of data mining, and hence the approaches that will be covered in this article, is to process the data such that underlying relationships between either the data points or the measurements can be identified. These relationships essentially define hypotheses that can be tested with more targeted experiments [5,6]. Specifically, this short article will focus primarily upon one specific problem under the data mining umbrella, namely, classification. In classification, one seeks to utilize a small training

set, that is data that has been previously labeled as belonging to a certain class, in order to classify data whose class is unknown, that is to predict the class the data belongs to.

CLASSIFICATION APPROACHES

The problems of data classification is essentially one of dimensionality reduction [7]. In such a case, we seek to find a mapping f , which allows us to map $x \in R^m \rightarrow R^{m'}$ s.t. $m \gg m'$. In the case of data classification, the reduction is performed such that the data is reduced to a single dimension $m' = 1$, the class. This dimension then becomes the classification. For the purpose of consistency, we will define the data as a matrix X of dimension $m \times n$. In this representation, each row represents an individual data point, and each column represents an individual descriptor in the data. Therefore, each data point $x \in X \subseteq R^n$ can be treated as a point in n -dimensional space. Appropriate preprocessing of the data may be required to transform the data into this form. Although in many cases this transformation can be trivial, in other cases transforming the data into a “workable” form is a highly nontrivial task. The goal therefore is to estimate an explicit, or implicit, function that maps points of the feature vector from the input space $X \subseteq R^n$, to an output space C , given a finite sample. The concept of the finite sample is important because, in general, what we are given is a finite representative subset of the original space (training set) and we wish to make predictions on new elements of the set (testing set). The data mining tasks can thus be defined based on the nature of the mapping C and the extent to which the training set is characterized. If the predicted quantity is a categorical value, and if we know the value that corresponds to each element of the training set, then the question becomes how to identify the mapping that connects the feature vector and the corresponding categorical value (class). This problem is formally defined as the classification problem

(supervised learning). If the class assignment is not known and we seek to (i) identify whether a small, yet unknown, number of classes exist; (ii) define the mapping assigning the features to classes, then we have a clustering problem (unsupervised learning). A related problem associated with superfluous information in the feature vector is the so-called feature selection problem. This is a problem closely related to overfitting in regression. Having a minimal number of features, this leads to simpler models, better generalization, and easier interpretation. One of the fundamental issues in data mining is therefore, to identify the least number of features and subset of the original set of features that best address the two problems previously defined. The concept of parsimony (Occam's razor) is often invoked to bias the search [8]: never do with more what can be done with fewer. Although numerous methods exist for addressing these problems, they will not be reviewed here. Extensive articles of classification were recently presented in Refs 9 and 10.

Classification, and for that matter most of the data mining tasks, are fundamentally optimization problems. Mathematical programming methodologies formalize the problem definition and make use of recent advances in optimization theory and applications for the efficient solution of the corresponding formulations. In fact, mathematical programming approaches, particularly linear programming, have long been used in data mining tasks. The pioneering work of Mangasarian in Refs 11 and 12 demonstrated how to formulate the problem of constructing planes to separate linearly separable sets of points. In this summary, we will follow the formalism put forth in Ref. 13, since it presented one of the most comprehensive approaches to this problem. One of the major advantages of a formulation based on mathematical programming is the ease in incorporating explicit problem-specific constraints. This will be discussed in greater detail later in this summary.

As discussed earlier, the main goal in classification is to predict a categorical variable (class) based on the values of the feature vector. The general families of

methods for addressing this problem include the following [9]:

1. estimation of the conditional probability of observing class C , given the feature vector x ;
2. analysis of various proximity metrics and evaluation of the decision of class assignment based on proximity;
3. recursive input space partitioning to maximize a score of class purity (tree-based methods).

The three broad families which we will be covering are decision trees [14], support vector machines (SVMs) [15], and Bayes and nearest neighbor classifiers [16]. Decision trees represent an extension of the orthogonal partition such that rather than utilizing global partition planes, it is possible to utilize local partition planes. SVMs represent the use of arbitrary partition hyperplanes rather than orthogonal planes, and Bayesian classifiers utilize probabilistic measures rather than hard partitions.

Decision Trees

Decision trees represent a recursive partitioning of the feature space. Rather than utilizing a global partitioning of the feature space, a decision tree has a hierarchy of partitions. In lieu of a plane cutting through the entire data space, a decision tree partitions the space into a series of hyperrectangles [17]. The use of a recursive partitioning provides a method for one to take into account the inherent relationship between different features. By not treating the features as independent, it is possible to improve the accuracy of such a classifier. Aside from the increase in accuracy, one of the primary advantages of a decision tree over the other methods is the fact that they summarize the feature space in a concise and humanly interpretable manner.

A common algorithm for creating decision trees is C4.5 [18]. The algorithm first iterates through all of the different features and identifies the partition which most enriches a given class, as determined via a normalized information gain function [18], which

is given in Equation (1). The normalized gain function is essentially the application of entropy to determine the order that the partitions have imposed upon the training set. In Equation (1), a represents the attributes in a given feature, c represents the classes, and p represents the probability that a value lies within a given class. This information gain is then normalized via the denominator where S represents the splits at each level, and S_i represents how many objects are in each leaf node after a split. After the first partition has been created, the data set is recursively processed until the data set has been correctly classified. After the initial training set has been processed, the decision tree can then be used to classify unknown data points.

$$G = \frac{\sum_{i=1}^{a \in A+1} \sum_{j=1}^c -\frac{1}{p_j} \log\left(\frac{1}{p_j}\right) - \sum_{i=1}^a \sum_{j=1}^c -\frac{1}{p_j} \log\left(\frac{1}{p_j}\right)}{\sum_{i=1}^B -\frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}} \quad (1)$$

Support Vector Machines

The two-class classification problem can be formulated as the search of a function that assigns a given input vector x into two disjoint point sets A and B . The data are represented in the form of matrices. Assuming that the set A has m elements and the set B has k elements, then $A \in R^{m \times n}$, $B \in R^{k \times n}$, describe the two sets, respectively. The discrimination is based on the derivation of the hyperplane:

$$P = \{x|x \in R^n, x^T \omega = \gamma\}$$

with normal distance from the origin $\frac{|\gamma|}{\|\omega\|_2}$. The optimization problem then becomes determination of ω and γ such that the separating hyperplane P defines two open half spaces containing mostly points in A and B , respectively:

$$\begin{aligned} &\{x|x \in R^n, x^T \omega < \gamma\} \\ &\{x|x \in R^n, x^T \omega > \gamma\} \end{aligned}$$

Unless A and B are disjoint, the separation can only be satisfied within some error. Minimization of the average violation provides a possible approximation of the separating hyperplane [13]:

$$\min_{\omega, \gamma} \frac{1}{m} \|(-A\omega + e\gamma + e)_+\|_1 + \frac{1}{k} \|(-B\omega + e\gamma + e)_+\|_1$$

In Ref. 13, a number of linear programming reformulations are discussed; these explore the properties and structure of the optimization problem. In particular, an effective robust linear programming (RLP) reformulation was suggested, making possible the solution of large-scale problems:

$$\begin{aligned} \min_{\omega, \gamma, y, z} & \frac{e^T y}{m} + \frac{e^T z}{k} \\ \text{s.t.} & -A\omega + e\gamma + e \leq y \\ & -B\omega + e\gamma + e \leq z \\ & y, z \geq 0 \end{aligned}$$

In Ref. 19, it was further demonstrated how the aforementioned formulation can be applied repeatedly to produce complex space partitions similar to those obtained by the application of standard decision tree methods such as C4.5 [20] and CART [21].

SVMs were originally designed for binary classification. Extension to multiclass problems is still an open research area [22]. The earliest multiclass implementation is the one against all [23] by constructing k SVM models, where k is the number of classes. The i th SVM classifies the examples of class i against all the other samples in all other classes. Another alternative builds one classifier against another [24] by building $k(k-1)/2$ models where each is trained on data from two classes. The emphasis of current research is on novel methods for generating all the decision functions through the solution of a single, but much larger, optimization problem [22].

Bayes and Nearest Neighbor Classifiers

Probability-based methods are normally classification methods that utilize Bayes rule to calculate the probability that a given

data point will belong to one class versus another [25]. Thus, for every data point, we essentially calculate the probability $P(C_i|x_1, x_2, \dots, x_n)$, which is the conditional probability of a data point belonging to class C_i given its feature set x . In the most general implementation of Bayes rule, one implicitly assumes that each of the features can depend upon each other in terms of the overall classification. In this formulation, the use of the standard Bayesian rule can be thought of as a probabilistic implementation of an SVM, in which rather than having a separating hyperplane, one is placing different N-D probability distribution functions in the feature space. By looking at the intersection of the different probability distribution functions, one can calculate the probability by which a data point belongs within a certain class.

However, while the use of Bayes rule represents the most general form of probabilistic classification techniques, the most widely used technique is naïve Bayes, in which the Bayesian relationship $P(C_i|x_1, x_2, \dots, x_n)$, is approximated by $\forall i, \prod_{j=1}^n P(C_i|x_j)$ [26]. This simplification is based upon the assumption that the individual features are statistically independent; that is, that there are no relationships between the values of one feature versus another. The use of this second formulation is akin to the reduction of the SVM into the simpler partitioning scheme with orthogonal partition planes. The training of probability-based classifiers is a straightforward application of Bayes rule. Thus, in the general case, one will essentially calculate $P(C_i|x_1, x_2, \dots, x_n)$ via Bayes rule, and in the simpler case of naïve Bayes, the probability calculation will be performed via the assumption that the individual features are independent.

Nearest neighbor approaches are nonlinear methods for assigning classes. However, rather than calculating a global classifier that partitions the space, nearest neighbor methods utilize a voting method. This method is exemplified by the k-nearest neighbor (KNN) algorithm [27]. In a KNN classifier, the KNN is taken and a simple voting process takes place to determine whether the unknown point belongs to a certain class. For instance,

in a 3-NN classifier, if two of the nearest points indicate that the class of the unknown point is C_i , and a third class is indicated as C_k , the assigned class will be C_j .

The primary benefit of KNN classifiers is the general simplicity of the algorithm as well as its ability to approximate arbitrary nonlinear partition functions between the different classes. However, it suffers from the disadvantage that the classification process is computationally expensive in comparison to the other methods. The classification of a single point requires calculation of the distance between that point and all of the points in the training set, as well as determining the k-smallest distances. This is in contrast to the other methods, that after a relatively expensive training process, the classification process can usually be implemented as a computationally efficient lookup table.

BROAD CHALLENGES

While we have presented a survey on the general basis of most algorithms that currently exist, data mining is still an active area of research. There are three primary unaddressed issues that drive current research.

1. *Scalability and the Curse of Dimensionality.* With the advances in technology, the ability to gather a wide range of information has increased, and thus the feature space that describes a single data point increases as well. However, with a linear increase in the number of features that describe the data, there is a corresponding exponential in the algorithmic complexity. Thus, while Moore's law suggests the same exponential doubling of computational power roughly every 12–18 months, hardware power is not the only answer. It is imperative that as databases get larger, there be a commensurate improvement in the algorithms that conduct classification, clustering, and feature selection.
2. *Noise and Infrequent Events.* Noise and uncertainty in the data are given features. Therefore, the question is how to

design algorithms that functions well in an environment with a significant amount of noise, or at least develop robust methods that quantify the effect of noise upon the results of these algorithms. Furthermore, while it is customary to look for repeated, that is frequent, events to derive rules underpinning a systemic response, it is often the case that infrequent events are the ones of particular interest. Robustly discriminating between an underlying rare event and outliers therefore becomes a critical issue.

3. *Interpretation and Visualization.* The ultimate goal of any data mining, and thus classification as well, is the understanding of the data and developing actionable strategies based on the conclusions. We therefore need to improve not only the interpretation of the derived models but also the knowledge delivery methods based on the derived models. Optimization and mathematical programming needs to provide not just the optimal solution, but also some way of interpreting the implications of the results.

REFERENCES

1. Grossman R. Data mining for scientific and engineering applications. Dordrecht, Boston (MA): Kluwer Academic Press; 2001. p. 605.
2. Pardalos PM, Boginski VL, Vazacopoulos A. Data mining in biomedicine. Springer optimization and its applications. New York: Springer; 2007. pp. xvii, 579.
3. Lei L, Jiong Y, Anthony KHT. Data mining techniques for microarray datasets. Proceedings of the 21st International Conference on Data Engineering. Tokyo, Japan: IEEE Computer Society; 2005.
4. Hand DJ, Mannila H, Smyth P. Principles of data mining. Volume 32, Adaptive computation and machine learning. Cambridge (MA): MIT Press; 2001. xxxii, 546 p.
5. Nakai K, Vert JP. Genome informatics for data-driven biology. Genome Biol 2002;3(4):reports4010.1–reports4010.3.
6. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Rep 2002;3(8):702.
7. Hayes AW. Principles and methods of toxicology. 5th ed. Boca Raton (FL): CRC Press/Taylor & Francis Group; 2008. pp. xxiii, 2270.
8. Blumer A, *et al.* Occam razor. Inf Process Lett 1987;24(6):377–380.
9. Hand DJ, Mannila H, Smyth P. Principles of data mining. Cambridge (MA): The MIT Press; 2001.
10. Grossman RL, *et al.* Data mining for scientific and engineering applications. Boston (MA): Kluwer Academic Publishers; 2001.
11. Mangasarian OL. Linear and nonlinear separation of patterns by linear programming. Oper Res 1965;13:444–452.
12. Mangasarian OL. Multi-surface method of pattern separation. IEEE Trans Inform Theor 1968;IT-14:801–807.
13. Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: formulations and challenges. Inform J Comput 1999;11(3):217–238.
14. Ruggieri S. Efficient C4.5. IEEE Trans Knowl Data Eng 2002;14(2):438–444.
15. Abe S. Support vector machines for pattern classification. Advances in pattern recognition. London: Springer; 2005. xiv, 343 p.
16. Manning CD, Schütze H. Foundations of statistical natural language processing. Cambridge (MA): MIT Press; 1999. pp. xxxvii, 680.
17. Rokach L, Maimon OZ. Data mining with decision trees : theory and applications. Series in machine perception and artificial intelligence. Singapore, Hackensack (NJ): World Scientific; 2008. pp. xviii, 244.
18. Quinlan JR. C4.5 : programs for machine learning. The Morgan Kaufmann series in machine learning. San Mateo (CA): Morgan Kaufmann Publishers; 1993. p. 302.
19. Mangasarian OL, Street WN, Wolberg WH. Breast-cancer diagnosis and prognosis via linear-programming. Oper Res 1995;43(4):570–577.
20. Quinlan R. C4.5 Programs for machine learning. San Mateo (CA): Morgan Kaufmann Publishers; 1993.
21. Breiman L, *et al.* Classification and regression trees. Boca Raton (FL): Chapman & Hall; 1984.
22. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. IEEE Trans Neural Networks 2002;13(2):415–425.

23. Scholkopf B, Burges C, Vapnik V. Extracting support data for a given task. In: First International Conference on Knowledge Discovery and Data Mining. Menlo Park (CA): AAAI Press; 1995.
24. Ulrich HGK. Pairwise classification and support vector machines. In: Advances in Kernel methods: support vector learning. Cambridge (MA): MIT Press; 1999. pp. 255–268.
25. Castillo E, Gutiérrez JM, Hadi AS. Expert systems and probabilistic network models. Monographs in computer science. New York: Springer; 1997. pp. xiv, 605.
26. Speed TP. Statistical analysis of gene expression microarray data. Interdisciplinary statistics. Boca Raton (FL): Chapman & Hall/CRC Press; 2003. pp. xiii, 222; p. 24 of plates.
27. Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973. pp. xvii, 482.

DATA MINING IN CONSTRUCTION BIDDING POLICY

TREFOR WILLIAMS
Professor of Civil Engineering,
Center for Advanced
Infrastructure and
Transportation, Rutgers
University, Piscataway, New
Jersey

Many construction projects are selected using competitive bidding. This is a reverse auction where the lowest bidder wins the project. It is often suspected that winners of construction projects are subject to a “bidder’s curse” because the lowest bid may include an error or omission.

It is useful to examine the nature of construction bids and to study the relationship of the original low bid with the completed project cost. It is possible to hypothesize that construction bids occur in repeatable patterns and that relationships exist among the submitted bids that can allow a prediction of completed project costs.

It is generally accepted that the ultimate cost of a complex project is difficult to predict and subject to variations caused by many different factors [1]. Skamris and Flyvberg [2] have reported that for large construction projects in Denmark, the construction costs are often underestimated. They believe that overoptimism in the initial planning stages of a construction project leads to decisions that misallocate funds. Models that can provide predictions of the magnitude of the project cost would allow owner organizations to plan and budget better for the actual costs of construction projects.

Prices often escalate above the low bid price particularly on new, complex projects. The bid price developed by a contractor for a construction project will include the total of the net estimate with appropriate additions for overhead, profit, and risk margin [3]. Subjective judgment is often used when calculating a bid price. Many other factors affect the

price of the project during construction. These factors include the quality and constructability of the design, management techniques employed by the contractor, location of the project, and macroeconomic conditions.

Change orders are the primary way in which costs escalate during construction projects. Change orders are negotiated contracts that modify the original project requirements. Although the low bidder on a construction project is initially responsible for constructing the project for the original low bid price, changes in the form of change orders inevitably increase the cost of complex and innovative projects. Typical reasons for a change order are differing site conditions, errors or omissions in the project plans, and design changes initiated by the owner.

The primary research problem is to determine if the bidding patterns reflect the uncertainties about a project’s completed cost in repeatable ways. The purpose of this analysis is to study if the individual bids for a project combine into useful patterns that indicate a project’s likelihood to have significant problems, and change orders that would cause extensive cost overruns from the original bid amount. An owner organization could use knowledge of a project’s likelihood for cost overruns at the time of the bid opening to determine if the project should proceed or be reexamined to identify problems. This article will explore how rules can be generated using association and classification from a database of project bidding data.

BIDDING PATTERNS

Various scenarios can be postulated for patterns in submitted bids. A project that is well designed and easily understood by contractors could be expected to have all the bids clustered together, and that project would be constructed with little increase in cost from the original low bid amount. Various strategies and patterns have been identified in the construction literature. In

the early 1990s, a study was undertaken to evaluate construction cost overruns and identify the factors that significantly impact construction cost overruns based on data from over 400 construction projects administered by the Washington State Department of Transportation during the fiscal years 1985 and 1989 [4]. Many factors that had the strongest association with construction cost overruns were identified. It was found that cost overruns increased with the size of the project as well as the number of bidders and with the increased dispersion of the various bids submitted per project. In other words, increased cost overruns appeared to occur on larger projects, projects with greater risks, projects with more complexity, and projects with larger numbers of bidders.

Bidders often bid in systematic ways that may be discernible as bid patterns. For example, a bidder seeing a poor design may bid a low amount, expecting to make money from the change orders that will be issued [5]. Crowley and Hancher [6] have identified three types of bidders. Bidders can be classified as phantom, mistaken, and fair based upon their bidding strategies. Phantom bidders consistently submit bids that are lower than the other bids received on a project. Mistaken bidders submit unusually low bids due to an error or omission. Fair bidders consistently target the market value with their bids. The variability between the bids of fair bidders is caused by minor differences in judgment or assumptions. Of course, the low bidder is contractually obligated to complete the project at the low bid price, but change orders are often issued during construction and increase the project cost.

PURPOSE AND BENEFITS OF ANALYZING BIDDING COSTS AND POLICIES

In this article, the focus is on using data mining techniques to identify patterns in highway projects bidding data that may provide a prediction of the magnitude of a project's cost overrun during construction. Possibly, a project's potential for cost overruns is related to the pattern of the submitted bids. It can be assumed that

bidders for a large construction project have carefully examined the plans and specifications before submitting their bids. Possibly bidders behave in repeatable patterns depending on the characteristics of the project they are bidding for. This knowledge could be exploited by owners leasing projects to determine if the bids received are suitable and as a way of planning for contingency funds to cover a project's construction costs.

BACKGROUND

Various models have been proposed for predicting completed project cost. In particular, multiple regression and neural networks have been applied to develop predictive models of construction costs. Wilmot and Cheng [7] have described the development of a model to estimate overall future highway construction costs in Louisiana. In the model, future construction costs are described in terms of predicted index values based on forecasts of the price of labor, materials, and equipment. Additional factors included in the model are the expected contract characteristics and contract environment. The model produces an index that projects future aggregate costs for groups of projects. The model does not predict the future costs of individual projects.

Emsley *et al.* [8] have developed a neural network model to predict total construction costs in the form of cost per square meter and log of cost per square meter based on input data from nearly 300 building projects. They developed several neural network models that included inputs about the project such as the duration, building function, required mechanical installations, and the requirements for piling. They concluded that the neural network models compared favorably with regression models.

Trost and Oberlander [9] used factor analysis and multivariate regression to develop an estimate score procedure that allows project managers to score an estimate, and then predict its accuracy based on the estimate score. The statistical analysis identified five significant factors: basic process design, team experience and cost information, time allowed to prepare the

estimate, site requirements, and the bidding and labor climate. They produced a regression equation to predict the required bid contingency using these five factors.

DATA MINING: ASSOCIATION AND CLASSIFICATION

Data mining is the process of automatically discovering useful information in large data repositories. Two common techniques for data mining are association and classification. The output of an association analysis is a set of association rules that show those items in the database that form frequently occurring sets [10], whereas a classification algorithm is applied to the bidding data to generate rules that can predict a project's anticipated level of cost overrun above the original low bid amount. Both algorithms automatically identify rules from databases. Association algorithms identify the best-performing rules in the database. Some of the cases in the database may not be included in the generated rule set. Classification algorithms generate a set of rules that will classify all of the items in the database. In this article, data was used to generate example classification and association analyses from highway projects conducted by the Texas Department of Transportation from 1995 to 2000. The database included data from 2765 competitively bid highway projects. For each of these projects, the lowest responsible bidder was selected to perform the work. This analysis considers only the bids of the responsible bidders; rejected bids are not considered.

Software Used for Analysis

The Weka software was used to classify the bidding data and to generate association rules [11]. The Weka software is a collection of machine learning algorithms and data preprocessing tools that have been developed at the University of Waikato in New Zealand. The system is written in Java and will run on most computer platforms. The Weka system is an open-source software and can be obtained from the Weka software web page at <http://www.cs.waikato.ac.nz/~ml/weka/>.

Classification Algorithms Employed

Classification can be defined as the task of assigning objects to one of several predefined categories [10]. Here, the object is a description of the bids submitted for a project and the predefined categories are the level of the completed project cost overrun. Several different classification algorithms were provided in the Weka system. The best-performing algorithm for predicting large project cost overruns was found to be the Ridor algorithm. The Ridor classification algorithm was applied to the Texas highway project bidding ratio data. The Ridor (Ripple Down Rules) algorithm is a machine learning technique that automatically generates rules from a data set [12]. The Ridor algorithm was initially developed to maintain rules for one of the first medical expert systems, and is now routinely used in chemical pathology laboratories to provide interpretive comments to assist doctors in using medical laboratory reports [13]. The Ridor algorithm learns rules with exceptions by generating a default rule. The default or top-level rule is the class of the output that occurs most frequently. Then the algorithm uses incremental reduced-error pruning to find exceptions with the smallest error rate, finding the best exceptions for each exception and iterating [11]. Thus it performs a treelike expansion of exceptions; the exceptions are a set of rules that predict cases other than the default.

Association Algorithms Employed

Several different association algorithms are provided in the Weka system. Through experimentation, two algorithms were found to find interesting relationships in the bidding ratio databases. The best-performing algorithms for predicting the level of the cost increase observed during construction was found to be the Predictive Apriori algorithm and the Tertius algorithm. Using the Texas data, rules were generated using both of these algorithms.

When discussing the generation of association rules two terms, *support* and *confidence*, are important to understand. The support is the number of instances in a database that a rule predicts correctly. The confidence is

the proportion of examples covered by the premise that are also covered by the consequence of a rule. In constructing a method to find association rules it is desirable to find rules that are not too broad (rules that would have broad support yet low confidence) and rules that are too narrow (rules that have high confidence yet apply to only a few cases). The Predictive Apriori algorithm combines confidence and support into a single measure called *predictive accuracy* to generate association rules that provide "...a trade-off between confidence and support which maximize the chance of correct predictions on unseen data" [14]. The Predictive Apriori algorithm implemented in Weka generates the n best association rules, with n selected by the user.

Tertius is an algorithm developed by Flach and Lachiche [15]. The Tertius algorithm uses a heuristic measure, confirmation, to induce rules that are interesting and unusual. Tertius employs a best-first search algorithm that uses an optimistic estimate of the best confirmation of possible refinements of a rule to prune the search.

BIDDING RATIOS

The inputs to the models were ratios that would be available at the time of the bid opening that could potentially describe a pattern in the submitted bids. Williams [16] calculated five ratios that describe the nature of the submitted bids. These ratios are a way of representing the relationships between bids for a project; they are dimensionless and are not dependent on the project magnitude. The rationale for the use of the bid ratios is that ratios describing the "signature" of the bids for a project can give clues about the project's likelihood to experience cost increases during construction. Potentially, bids that are closely bunched together or contain extreme outliers may give clues about the completed project cost. The ratios include the second lowest bid ratio, the mean bid ratio, the maximum bid ratio, and the coefficient of variation of the submitted bids. The formulas used for the calculation of the ratios are given below. A ratio was calculated to compare the second

lowest bid with the low bid amount. This ratio determines if the low bidder and next lowest bidder basically agree about the project cost. The ratio is given in Equation (1).

$$\begin{aligned} &\text{Second Lowest Bid Ratio} \\ &= \frac{\text{Second Lowest Bid} - \text{Low Bid}}{\text{Low Bid}}. \end{aligned} \quad (1)$$

Another ratio measures the difference between the low bid and the mean bid and is given by Equation (2). The purpose is to measure how far the low bid is from the "consensus" bid.

$$\text{Mean Bid Ratio} = \frac{\text{Mean Bid} - \text{Low Bid}}{\text{Low Bid}}. \quad (2)$$

The mean bid ratio may indicate the degree of clustering of the bids. If the ratio of the low bid to the mean bid is large, it probably indicates that the low bidder has submitted an artificially low bid or a project where there is little agreement about costs. Equation (3) shows a median bid ratio that was also calculated as an alternative measure of the low bids deviation from the group consensus.

$$\text{Median Bid Ratio} = \frac{\text{Median Bid} - \text{Low Bid}}{\text{Low Bid}}. \quad (3)$$

A ratio is also calculated relating the maximum bid to the low bid. This ratio indicates the spread of the submitted bids, and indicates if there is significant variation in the range of values of the submitted bids. It is also an indication of the existence of an extremely high bid. The maximum bid ratio is shown in Equation (4).

$$\begin{aligned} &\text{Maximum Bid Ratio} \\ &= \frac{\text{Maximum Bid} - \text{Low Bid}}{\text{Low Bid}}. \end{aligned} \quad (4)$$

As a way of measuring the agreement between bidders, the coefficient of variation

was calculated. The coefficient of variation is given by Equation (5).

$$\text{Coefficient of Variation} = \sigma/\bar{x}. \quad (5)$$

Where σ = the standard deviation of the bids submitted for a project, and $\bar{x}$ is the mean of the submitted bids. The coefficient of variation is a measure of the spread of the submitted bids and is a measure of the agreement between the bidders.

STATISTICAL RELATIONSHIP BETWEEN RATIO LEVELS AND COST OVERRUNS

Williams *et al.* [17] have found that there is a statistically significant link between the level of the ratios and the completed project cost. Their study was conducted using this highway-bidding data from Texas. There was found to be statistically significant difference in the value of the bidding ratios between projects completed at a cost near the low bid amount and projects where the completed project cost differed significantly from the low bid amount. Higher values of the ratios are observed for projects completed with significant deviations from the mean. Projects completed near the original low bid amount tend to have lower values of the ratios. It was also noted that the elevated ratio values seem to occur for projects that have large cost increases and for projects that are completed for significantly less than the original bid amount. It was also found that due to the noise in the bidding data, it was difficult to construct regression or neural network models that could exploit knowledge of the bidding ratios to predict the magnitude of a project's likely cost increase during construction. Therefore, data mining techniques that are able to generate association and classification rules from the data are a useful area to explore, so as to determine if they can provide useful predictions.

PREPARING THE BIDDING DATA FOR CLASSIFICATION AND ASSOCIATION

The focus of the analysis is to attempt to identify in advance, projects that will have

a large cost increase. A large percentage of the projects are completed for near or less than the low bid amount (1712 projects out of the total number of 2765 projects in the database). Each record in the database consisted of the data for one project. Each project had a value defined for the five ratios described above, and for the magnitude of the low bid amount. These could take on the nominal values of "low", "medium", or "high." The ranges for each variable were determined by the values that placed one-third of the projects at each level. Each project also had its cost overrun recorded as "Near", "Overrun", or "BigOverrun." A "Near" project was completed for less than or up to a cost overrun of 5%. An "Overrun" project was completed for a cost overrun of between 5% and 10%. A "BigOverrun" project was one in which the completed project cost exceeded the low bid amount by more than 10%. There were 444 projects in the Overrun category and 609 projects in the BigOverrun category.

Table 1 shows the relationships in the Texas highway between the level of three of the bidding ratios and the weighted average percentage difference between the original low bid amount and the completed project cost. The projects in each grouping were weighted according to their low bid magnitude. In Table 1, it can be observed that elevated levels of the ratio values are associated with projects that tend to have a larger percentage difference between the completed project cost and the low bid. For example, projects where all of the ratio values are low tend to have a lower cost overrun than projects where all of the ratio levels are high. However, it can also be seen that there are many complex relationships between the levels of the three ratios and the project's ultimate cost increase. In addition, examination of individual projects in the database indicates that there are many exceptions to general trends. There are some projects in the database that have large cost increases but low ratio values. These many exceptions make it difficult to produce accurate classification rules and association rules. The variability in the data will be reflected in the results shown in Table 1.

Table 1. Weighted Average Percentage Difference between Completed Project Cost and Low Bid for all Texas Projects

Coefficient Index Value	Second Lowest Bid Index Value	Maximum Bid Index Value	Number of Projects	Weighted Average Percentage Difference
High	Low	Medium	53	13.94
Medium	Medium	Low	31	9.08
High	High	High	421	7.94
Medium	Low	Low	49	7.28
High	High	Medium	51	6.87
Low	Low	High	28	6.85
Medium	Medium	High	54	6.61
High	Medium	High	187	6.52
High	Medium	Medium	35	6.45
Medium	High	Medium	224	6.20
Low	Low	Low	380	5.96
Low	High	Low	55	5.85
Low	Low	Medium	32	5.74
Medium	High	Low	41	5.59
Medium	High	High	56	5.15
High	Low	High	173	4.88
Medium	Medium	Medium	231	4.62
Low	Medium	Medium	55	4.34
Low	Medium	Low	328	4.27
Low	High	Medium	33	3.58

GENERATING CLASSIFICATION RULES

Model Inputs and Generated Rules

A rule set was constructed using the Ridor algorithm. Using the Weka software, the bidding ratio data were first preprocessed to identify the best ratios to be used to generate classification rules. Using the Weka data preprocessor, an attribute selection algorithm was used that considered the predictive value of each attribute individually, along with the degree of redundancy among them. The attribute selection method employed a greedy hill-climbing search method with backtracking to select the best attributes [11]. Using the attribute selection method employed in Weka indicated that the second lowest bid ratio, the mean ratio, the coefficient of variation, and the magnitude of the low bid were the best predictive indications. Therefore, the maximum bid ratio and the median bid ratio were not considered in the generation of rules.

Figure 1 shows the rules that were generated. Six rules were generated in total. The antecedents for each rule, the left hand

side of the rule, are nominal values of the level of the coefficient of variation, the low bid ratio, and/or the second lowest bid ratio. These variables can take the value of low, medium or high. The right hand side of the rule is the prediction, giving the level of the project cost increase as “Near”, “Overrun”, or “Big Overrun.”

The major class of the output attribute is a prediction of the project overrun to be near the original bid amount. By default, this becomes the top-level rule. An examination of the rules shows that the exception rules generated are instances where the antecedents in the rules, the ratio levels, are of a high or medium level and the project’s final cost is predicted to be a “BigOverrun” or an “Overrun.” This result confirms that elevated levels of the ratios are generally associated with increased cost overruns. For the exception rules in Fig. 1, the numbers in parentheses indicate the total number of training cases covered by the rule and the second number is total number of cases that would be classified incorrectly by the rule.

The first exception rule indicates that all projects with a median bid ratio level of

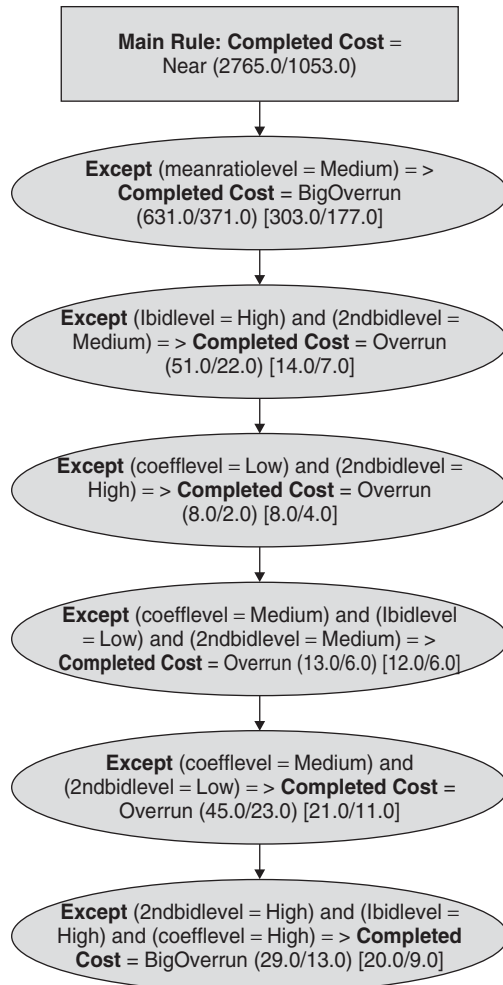


Figure 1. Classification rules generated using the Ridor algorithm.

medium will have a large overrun. However, examination of the output indicates that the rule is incorrect for about 58% of the test cases. A better rule is the final exception rule generated where if three ratios are simultaneously high, then a large cost overrun is predicted. This rule is correct for 11 out of 20 test cases.

Performance of the Classification Rules

To test the performance of the rules, a test set was used which included 33% of the data selected at random. The other 67% of the data were used to generate the rules. The

Table 2. Confusion Matrix for Classification Rule Predictions

Actual Cost Overrun	Predicted Cost Overrun			
	Near	Overrun	Big Overrun	Total
Near	307	21	267	595
Overrun	81	6	69	156
Big Overrun	87	3	100	190
Total	475	30	436	941

numbers in brackets in Fig. 1 at the end of the exception rules indicate the cases in the test set covered by the rule and the second number is the number of cases incorrectly classified by the rule. The success of the generated rules in providing predictions is given by the confusion matrix in Table 2. In the confusion matrix, there is a row and column for each predicted level of the project cost overrun.

The confusion matrix indicates that the generated rules produced correct predictions for 43.89% of the projects. However, the model classified 100 out of 190 projects (53%) with large cost overruns correctly. Projects that were completed “Near” the original low bid amount were predicted correctly 65% of the time. It was found to be difficult to classify the “Overrun” category because it has a smaller number of cases than the “Near” or “BigOverrun” categories.

ASSOCIATION RULES GENERATED

Rules were generated using both the Tertius and Predictive Apriori association algorithms. Using the Weka software, the bidding ratio data were first preprocessed to identify the best ratios to use so as to generate classification rules. Using the Weka program, an attribute selection algorithm was used that considered the predictive value of each attribute individually, along with the degree of redundancy among them. The attribute selection method employed a greedy, hill-climbing search method with backtracking to select the best attributes [11]. Using the attribute selection method employed in Weka indicated that the second lowest bid ratio, the mean ratio, the coefficient of variation, and



Figure 2. Predictive Apriori best association rules for different levels of cost overrun.

the magnitude of the low bid were the best predictive indications. Therefore, the maximum bid ratio and the median bid ratio were not considered in the generation of rules.

Predictive Apriori Rules

Figure 2 shows some of the best rules from the computer output from the Predictive Apriori algorithm run. The one hundred best rules were generated. The best rules for the prediction of each of the three levels of cost overrun are shown in the figure. The underlined italic number is the ranking of the rule as generated by the computer software. The rule premise (the left-hand portion of the rule) is made up of the level of

various combinations of the bidding ratios. The number after the premise of each rule is the number of cases in the database that displayed these combinations of the bidding ratios. Each ratio could have the value of low, medium or high. The consequence of the rule is the level of cost overrun. This is the portion of the rule to the right of the arrow. The number after the outcome of the rule is the number of cases in the database that are correctly classified. The predictive accuracy is shown at the end of each rule. This number is a combination of two metrics: the confidence in the rule, and the support for the rule. The confidence in a rule is the proportion of examples covered by the premise

1. /* 0.058923 0.030018 */2ndbidlevel = Medium and meanratiolevel = Low and coefflevel = Low ==> cost_overrun = Near
2. /* 0.053895 0.088608 */meanratiolevel = Low and coefflevel = Low ==> cost_overrun = Near
3. /* 0.052951 0.033635 */2ndbidlevel = Medium and meanratiolevel = Low ==> cost_overrun = Near
4. /* 0.050094 0.114086 */meanratiolevel = Low ==> cost_overrun = Near
5. /* 0.047197 0.007957 */2ndbidlevel = Low and meanratiolevel = High and lbidlevel = Low ==> cost_overrun = Near
6. /* 0.045437 0.049548 */meanratiolevel = Low and lbidlevel = High ==> cost_overrun = Near
7. /* 0.045263 0.041230 */meanratiolevel = Low and coefflevel = Low and lbidlevel = High ==> cost_overrun = Near
8. /* 0.043946 0.248463 */coefflevel = High ==> cost_overrun = bigOverrun
9. /* 0.043763 0.122966 */2ndbidlevel = High and coefflevel = High ==> cost_overrun = bigOverrun
10. /* 0.043681 0.047378 */2ndbidlevel = Medium and coefflevel = Low ==> cost_overrun = Near
11. /* 0.042628 0.134901 */2ndbidlevel = High and meanratiolevel = High ==> cost_overrun = bigOverrun
12. /* 0.042186 0.013382 */2ndbidlevel = Low and meanratiolevel = High ==> cost_overrun = Near
13. /* 0.041989 0.116817 */coefflevel = Low ==> cost_overrun = Near
14. /* 0.040961 0.010850 */2ndbidlevel = Low and meanratiolevel = High and coefflevel = High ==> cost_overrun = Near
15. /* 0.039287 0.249910 */2ndbidlevel = High ==> cost_overrun = bigOverrun

Figure 3. Best rules generated by Tertius algorithm.

that are also covered by the consequence of the rule. The coverage is the number of instances that are correctly predicted. Highly accurate predictions have a predictive accuracy approaching one. The output using the Predictive Apriori algorithm shows that it is possible to develop highly accurate rules to predict when the project cost overrun will be near to the original low bid amount. However, it was more difficult to produce accurate rules that can predict both, overruns and large cost overruns. Four out of the five best rules for predicting a BigOverrun had a high level of the coefficient of variation ratio.

Rules Generated by the Tertius Algorithm

Figure 3 shows the top 15 ranked rules generated by the Tertius algorithm. This algorithm appears to be more successful in identifying rules related to large project cost overruns. In the output shown in the figure, the two numbers provided for each rule are the value of the degree of confirmation calculated for

the rule, and the relative frequency of counterinstances. The rules are ranked in order of the value of the confirmation heuristic. When this second number, the relative frequency of counterinstances, is zero it would indicate that all of the cases predicted by the body of the rule were correct. Examination of the output indicates that rules containing low levels of the ratio values were generated that predicted a completed project cost near the original low bid amount. Several rules were generated that predicted large cost overruns when ratio levels were high; however, these rules had a much higher frequency of counterinstances.

CONCLUSIONS

This exploration of the use of the Ridor classification algorithm showed some potential for producing rules that can predict construction project cost overruns. In particular, the rules are able to predict projects with a greater than 10% cost overrun with a success rate of

53%. However, examination of the bidding data indicates that the existence of some contradictory cases in the test set make it difficult for the Ridor algorithm to produce rules that are highly accurate.

The use of association algorithms also showed some potential for producing rules that can indicate the level of cost increase a construction project is likely to have. The analysis has shown that elevated levels of the bidding ratios are more likely to be associated with projects that have high cost overruns. Both methods of producing association rules were able to automatically produce rules that indicated projects completed near the low bid amount with a high level of accuracy. A disadvantage of the association rules was their inability to predict an outcome for all possible combinations of input variables.

The research indicates that analysis of the submitted bids and the patterns they may contain are a fertile area for further research. An enhanced understanding of the bidding strategies used by construction contractors can allow owners to better understand when projects may be prone to cost overruns during construction. Possibilities for future research include the exploration of ways to "cleanse" the data-set to remove anomalous cases or cases that are outliers, consideration of other classification methods and algorithms, and methods of combining rule-based prediction with subjective expert input.

REFERENCES

1. de Neufville R. Applied systems analysis. New York: McGraw-Hill; 1990.
2. Skamris M, Flyvbjerg B. Accuracy of traffic forecasts and cost estimates on large transportation projects. *Transp Res Rec Transp Plann Appl* 1996;1518:65–69.
3. Smith AJ. Estimating, tendering and bidding for construction. Houndsmill: Macmillan Press; 1995.
4. Hinze J, Selstead G, Maloney JP. Cost overruns on State of Washington construction contracts. *Transp Res Rec* 1992;1351:87–93.
5. Halpin D, Woodhead R. Construction management. New York: Wiley; 1998.
6. Crowley L, Hancher D. Evaluation of competitive bids. *J Constr Eng Manage* 1995;121(2):238–245.
7. Wilmot CG, Cheng G. Estimating future highway construction costs. *J Constr Eng Manage* 2003;129(3):272–279.
8. Trost SM, Oberlander GD. Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *J Constr Eng Manage* 2003;129(2):198–204.
9. Emsley MW, Lowe DJ, Duff AR, *et al.* Data modeling and the application of a neural network approach to the prediction of total construction costs. *Constr Manage Econ* 2002;20(6):465–472.
10. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Boston (MA): Addison-Wesley; 2006.
11. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco (CA): Morgan Kaufmann Publishers; 2005.
12. Gaines BR, Compton P. Induction of ripple-down rules applied to modeling large databases. *J Intell Inf Syst* 1995;5(3): 211–228.
13. Compton P, Peters L, Edwards G, *et al.* Experience with ripple-down rules. *Knowl Based Syst* 2006;19:356–362.
14. Scheffer T. Finding association rules that trade support optimally against confidence. *Intell Data Anal* 2005;9:381–395.
15. Flach P, Lachiche N. Confirmation-guided discovery of first-order rules with tertius. *Mach Learn* 2001;42:61–95.
16. Williams T. Bidding ratios to predict highway project costs. *Eng Constr Arch Manage* 2005;12(1):38–51.
17. Williams T, Lakshminarayanan S, Sackrowitz H. Analyzing bidding statistics to predict completed project cost. In: Proceedings of the International Conference on Computing in Civil Engineering [CD-ROM]. ASCE; Cancun, Mexico 2005.

DECISION ANALYSIS AND COUNTERTERRORISM

JAMIE D. LLOYD
JASON R. W. MERRICK
Statistical Sciences and Operations
Research, Virginia Commonwealth
University, Richmond, Virginia

COUNTERTERRORISM ANALYSIS USING DECISION TREES AND PROBABILITY ASSESSMENT

From its early inception, the Department of Homeland Security (DHS) has been concerned about attacks on commercial airlines using surface-to-air missiles, called *Man Portable Air Defense System* (MANPADS). Such missiles can be guided by an automated infrared detection system that targets the heat of the airplane's exhausts or by laser target identification from a person on the ground. In 2004, DHS awarded research and development funding for protective measures against MANPADS attacks. The researchers sought to modify protective measures currently used on military aircraft. von Winterfeldt and O'Sullivan [1] perform a decision analysis to determine whether such directed infrared countermeasures (DIRCMs) are cost effective.

Figure 1 illustrates a decision tree for the MANPADS decision. The sequence of events that can lead to a downed airliner is a MANPADS attack, a failed interdiction, a hit on the plane, and a fatal crash. If no countermeasures are installed then the probabilities of each event conditioned on the occurrence of the previous events are denoted as p , $(1 - q)$, h , and r . If countermeasures are installed, then the probabilities are modified by parameters representing the effectiveness of the countermeasures. Thus, the probabilities of each event conditioned on the occurrence of previous events and countermeasures are denoted as $(1 - d)*p$, $(1 - f)*(1 - q)$, $(1 - e)*h$, and $(1 - g)*r$. This implies that $(1 - d)$ is the proportional reduction in the probability of a MANPADS-based attempt, $(1 - f)$ is the reduction in the

probability that interdiction fails, $(1 - e)$ is the reduction in probability that the missile hits the plane, and $(1 - g)$ is the reduction in probability that the plane crashes if countermeasures are installed. The costs are also parameterized. Table 1 shows a range and base estimate for each probability parameter and the costs associated with each outcome.

At base parameter values, the optimal decision is to install countermeasures with an expected cost of \$15 billion in comparison to \$19 billion for no countermeasures. This parameterized decision tree is effective in capturing the structure of the decision problem with a simple model, and allows the significant uncertainty about the parameters of the problem to be tested. The question then turns to the parameters that are driving the optimal decision. von Winterfeldt and O'Sullivan [1] describe the creation of a spreadsheet decision-support tool that allows decision makers to change the parameters within the ranges in Table 1 and to test the resulting optimal decision. Varying parameters at the decision makers' discretion can be useful to allow them to see the potential changes in the optimal decision, but it can also be useful to perform sensitivity on all the parameters to see how and when they change over specified ranges. We will not reproduce the tornado diagram here for brevity, but we can describe key findings. The largest effect on expected cost comes from the swing from the minimum to maximum value of EL (economic loss/fatal crash). With increase in EL from the base value, the decision remains to install countermeasures, but as EL decreases the optimal decision switches to no countermeasures. The switch is calculated to occur at \$74.3 billion. The same pattern repeats for the next five most influential parameters: the cost of countermeasures, the probability of an attack attempt in the next 10 years with no countermeasures, the probability of a crash given a missile hits with no countermeasures, the percent loss in passenger volume given a missile hit and a safe landing, and the cost of a false alarm.

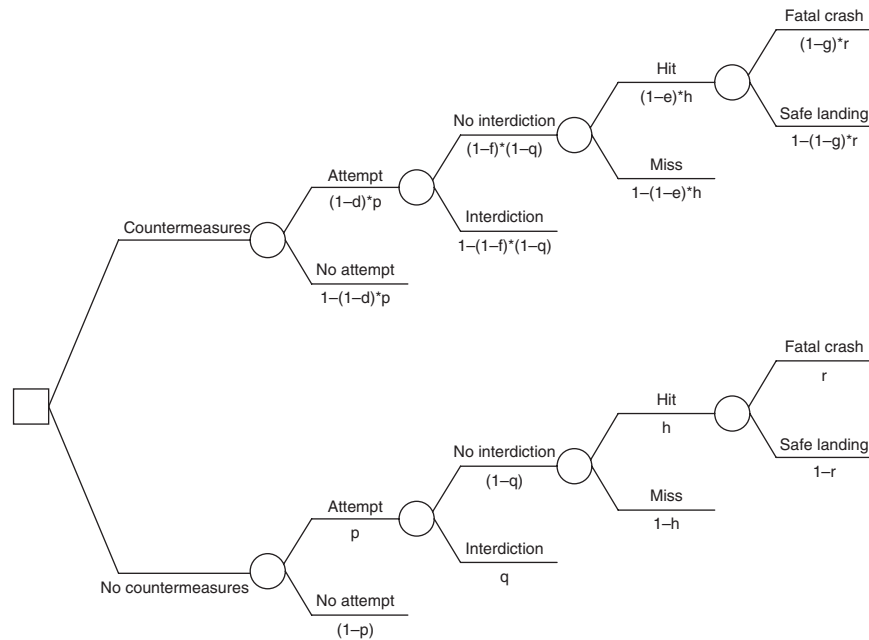
Bakir [2] repeated the parameterized decision tree approach for decisions about containers entering the US by truck over

Table 1. Decision Tree Inputs—Probabilities and Effectiveness Ranges [1]

Parameter	Min	Base	Max
<i>Probabilities</i>			
p—attack attempted in 10 years	0	0.5	1
q—interdiction/attempt	0	0	0.25
h—hit/attack	0.5	0.8	1
r—crash/hit	0	0.25	0.5
<i>Countermeasure effects</i>			
d—deterrence effect	0	0.5	1
f—interdiction effect	0	0	0.25
e—diversion/destruction effect	0.5	0.8	1
g—crash reduction effect	0	0	1
<i>Consequences</i>			
LL—fatalities/crash	0	200	400
CP—cost of plane (millions)	0	200	500
EL—economic loss/fatal crash (billions)	0	100	500
a—percent loss/hit and safe landing (%)	0	25	50
b—percent loss/miss (%)	0	10	25
FA—number of false alarms/year	0	10	20
CC—cost of countermeasures (billions)	5	10	50

the Mexican border. Three decisions are considered, improving transportation security, installing screening equipment on the Mexican side of the border, and whether to screen on the US side of the border with advanced

scintillator portals (ASPs) or keep existing radiation portal monitors (RPMs). The chain of events that lead to a successful attack is first an attempt to smuggle a nuclear threat, failure to detect the threat prior to reaching

**Figure 1.** Decision tree for MANPADS (adapted from [1]).

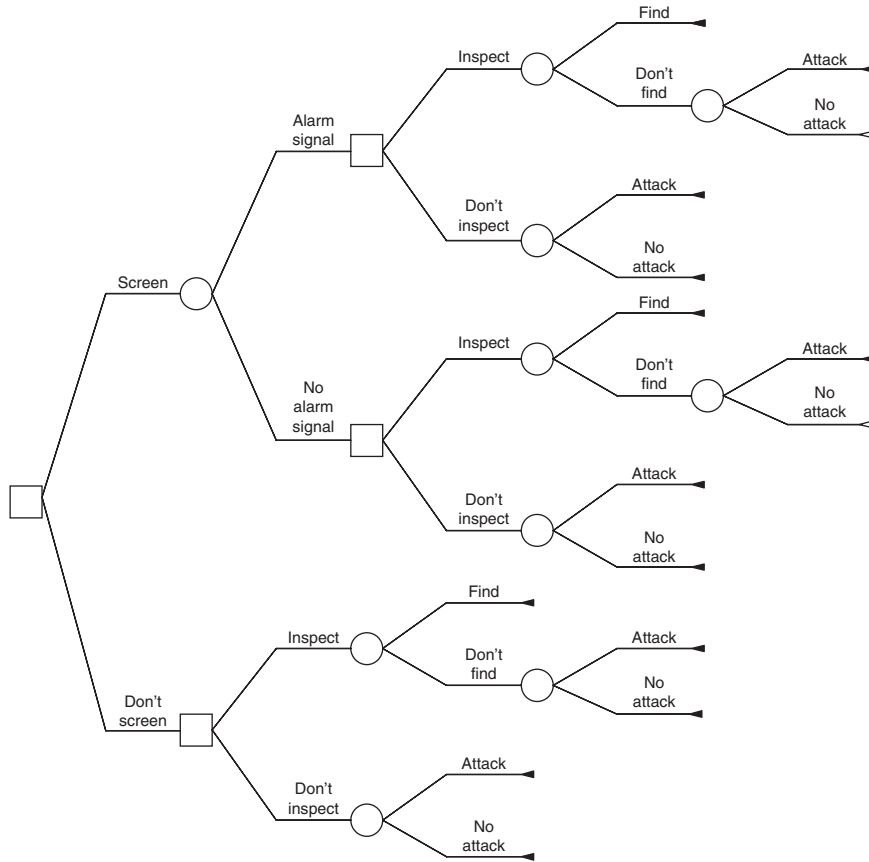


Figure 2. A decision tree for the container-screening problem with one screening.

the border, failure to detect on the Mexican side of the border, failure to detect on the US side, the terrorist deciding not to detonate at the border, failure to interdict within the US, and lastly the success of detonation at the intended target. Within Bakir's analysis is the assumption that if an alarm sounds when screening the container, then the container is sent for secondary screening and inspection. Bakir [2] provides a wealth of probability and cost estimates and performs a comprehensive sensitivity analysis on each parameter. For instance, we would have to screen 10 billion containers to detect four true threats. However, we would also detect 249,999,996 naturally occurring radioactive material (NORM) sources (regular goods that give off radiation, such as ceramic tile and irradiated iron). Inspecting containers

because of NORM sources is costly and delays the cargo. Bakir [2] concludes that US Customs and Border Protection (CBP) should not implement new ASP technology.

Merrick and McLay [3] extend Bakir's analysis, examining specifically whether screening should be performed at all. Figure 2 shows a decision tree for this decision that includes whether to inspect the container if an alarm sounds. Given Bakir's base parameter values, the answer is no. The value of information from screening is not sufficient to justify the cost of screening and the cost of false alarms from NORM sources. However, such an analysis ignores the deterrence effect of screening. Bakir parameterizes this effect and assumes that the probability of an attempt reduces by an estimated factor if additional security is

added. Merrick and McLay [3] find that the probability of an attempt if containers are not screened must be at least 1.6 times the probability of an attempt if they are screened to make screening economically worthwhile. The question is how to determine the effect of screening on terrorist decisions. This is more the prevue of game theory, but methods such as those in Parnell *et al.* [4] are potential solutions and are discussed in the section titled “Counterterrorism and the Intelligent Adversary.”

COUNTERTERRORISM ANALYSIS USING VALUES AND OBJECTIVES

Keeney [5] proposes that decision makers consider values and objectives in the analysis of counterterrorism decisions. Keeney developed 14 attributes for evaluating the outcomes of such decisions from the defender’s perspective. Six attributes are costs, namely

- direct and indirect costs to individuals
- direct and indirect costs to businesses
- direct and indirect costs to the government.

Three other attributes are counts of the effect of the attack on individuals, namely:

- the number of deaths caused by the attack;
- the number of victims disabled by the attack;
- the number of jobs lost because of the attack.

These measure the effect of the attack on the victims, both directly and indirectly. The five other attributes are associated with the effects on continuing terrorist operations and US society, including

- the effect on recruitment of future terrorists;
- the number of dollars flowing to support terrorism;
- the level of political support for US actions to counterterrorism;

- the number of US citizens experiencing limits to their freedom;
- the number of citizens experiencing fear and despair.

Keeney [5] also illustrates a value model of terrorist preferences developed with subject matter experts at Lawrence Livermore Laboratory. Keeney lists three terrorist objectives for a decision related to terrorist attempts to enrich plutonium for a radioactive dispersal device.

1. Maximize the amount of plutonium acquired.
2. Maximize the purity of the plutonium.
3. Minimize the radiation danger to the terrorist.

The experts determine the following attributes:

- X_1 plutonium extracted, measured from 10 to 2500 grams (gPu)
- X_2 purity of stolen material, measured from 0 to 333 grams of uranium per liter of material (gU/L)
- X_3 radiation dose, measured from 0 to 100 Rads/h.

Keeney [5] creates the utility function $u(x_1, x_2, x_3)$, where x_i is a specific level of attribute $X_i, i = 1, 2, 3$. For example, $u(150, 27, 5)$, which denotes the utility of the terrorist obtaining 150 g of plutonium with purity 27 gU/L that gives off 5 Rads/h. Keeney establishes the appropriateness of the additive or multiplicative utility function by determining if independence assumptions are upheld [6]. Holding two attributes at a constant level and varying the third, one can elicit single attribute utility function curves. Keeney elicits example utility function using the lottery indifference technique. Weights for either the additive or the multiplicative utility function are then elicited using other indifference judgments.

Leung *et al.* [7] consider bridges as vulnerable targets for terrorist attacks and rank the most likely targets. Leung *et al.* uses a risk filtering, ranking, and management (RFRM) method for their counterterrorism

analysis which builds on hierarchical holographic modeling (HHM) to identify risks. The risks are then ordered to help the decision maker prioritize. Risk management is then used to determine potential plans of action to protect the bridges most at risk. Other applications of such multiple objective methods include ranking various types of critical infrastructures to protect [8,9]. Buckshaw *et al.* [10] implement a similar structure when evaluating alternatives designed for the protection of critical information systems. Feng and Keller [11] used a multiple objective decision analysis approach to assess potential distribution plans in a specified region of potassium iodide to counter the release of radioactive iodine due to a terrorist attack or an accident.

Keeney and von Winterfeldt [12] analyze statements made by Al Qaeda to determine the objectives implied. They determine five strategic objectives.

1. Inspire and incite Islamic movements and the Muslim masses of the world to attack the enemies of Islam
2. Expel western powers from the Middle East
3. Destroy Israel
4. Establish Islamic religious authority in the Middle East
5. Extend Islamic authority and religion into new areas of the world.

Keeney and von Winterfeldt then define two groups of fundamental objective; one group related to growth of the movement, and one related to military outcomes. The growth related objectives are

- G1. Main status as the primary Middle Eastern force to be reckoned with,
- G2. Created homegrown terror cells in the US and Western Europe,
- G3. Win the hearts and minds of the Muslim population,
- G4. Maximize financial contributions from supporters,
- G5. Recruit new followers.

The military objectives are

- M1. Attack US targets,
- M2. Cause economic loss for the US,
- M3. Expel the Americans from Iraq,
- M4. Kill large numbers of infidels.

An understanding of Al Qaeda's objectives can enrich existing decision analyses. For instance, Merrick and McLay [3] consider a two-objective value function for the container-screening decision that includes elements of objectives G3, G4, and G5 in the growth area, and M1, M2, and M4 in the military area. They find that depending on the value function used, the optimal container-screening decision can change significantly from that obtained with a single cost objective.

COUNTERTERRORISM AND THE INTELLIGENT ADVERSARY

Game theory has also been widely applied in counterterrorism to model the decisions of the terrorist adversary, including general strategy and defense modeling [13–17], and specific applications, such as protecting power transmission systems [18] and commercial airlines [19]. In this section, we will concentrate on adversarial methods that use decision analysis as their primary technique.

Paté-Cornell and Guikema [20] created a model for setting priorities among threats and countermeasures that combines aspects of probabilistic risk analysis, decision analysis, and game theory. Figure 3 shows the influence diagrams used. The influence diagram in Fig. 3a is the terrorist decision, and can be repeated for multiple potential terrorist groups. The authors solve the terrorist influence diagram to discover the utility of each potential terrorist alternative. These utilities are then transformed into probabilities of terrorist behavior for use in the US influence diagram in Fig. 3b. As one of the earliest combinations of decision analysis and game theory in the post 9/11 emphasis on counterterrorism, the approach includes a sequential consideration of the terrorist decision and then the counterterrorist. However, the authors also point out that such thinking

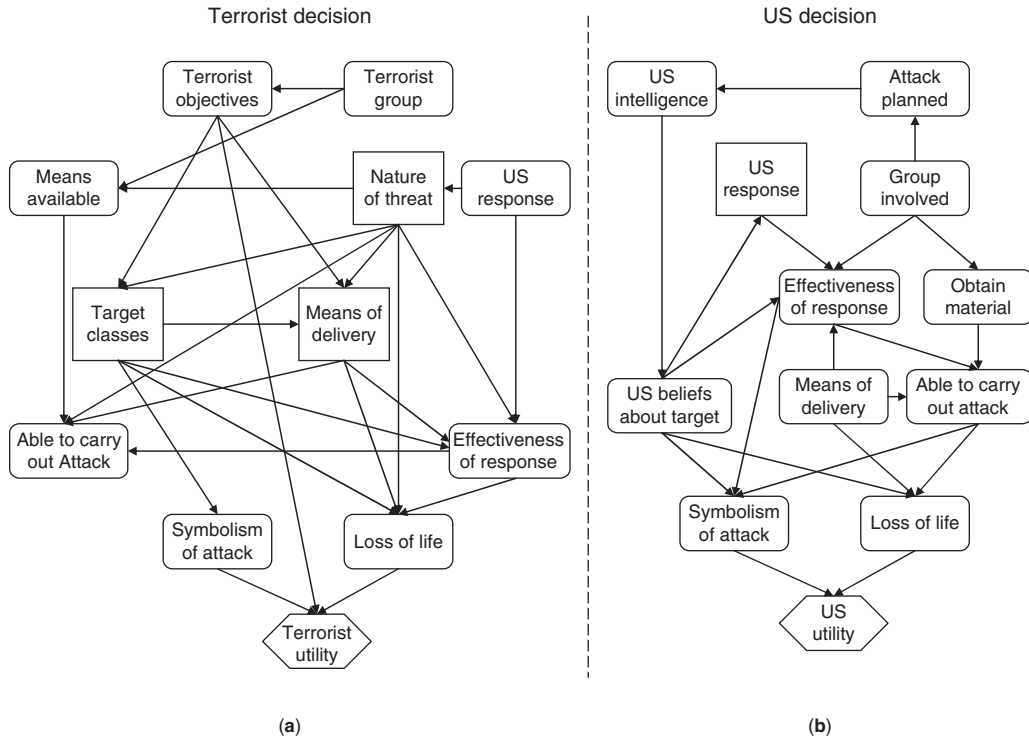


Figure 3. Influence diagrams for both terrorist and US [20].

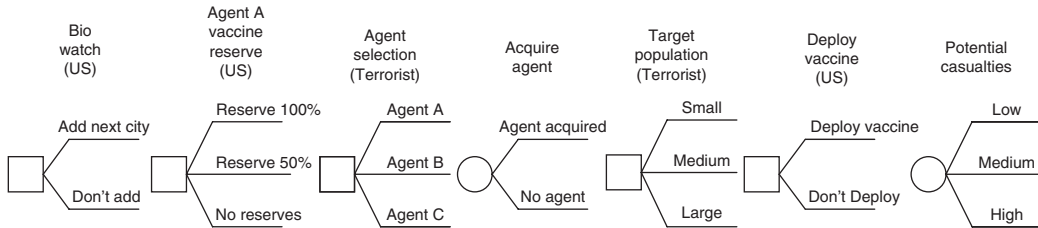


Figure 4. Decisions trees including decisions for both terrorist and US [4].

can be iterated between the two adversarial decisions.

Parnell *et al.* [4] consider the problem of defending against bioterrorism. They consider two decisions for the counterterrorist, specifically whether to add a city to the biowatch program, and whether to store and deploy vaccines for use after a bioterrorism attack. The terrorist decision is what kind of biological agent to use in an attack. These decisions are sequential and can, thus, be represented in a single decision tree or influence diagram. The decision tree is shown in Fig. 4.

To simplify solution of the decision tree, Parnell *et al.* assume that the terrorist objectives are diametrically opposed to the defender's. That is, if we aim to minimize the expected total cost of defense and attacks, then they aim to maximize the same measure. Solving Fig. 4 then follows the standard backward induction method for solving decisions trees, where one starts at the right (end) of the tree, takes an expected value at any probability node, takes the "best" expected value at any decision node, and folds the tree backwards. However, at a terrorist decision node "best"

is interpreted as the maximum, and at a US decision node “best” is interpreted as the minimum. It should be noted that this also assumes that the terrorists share our same degree of belief, or probabilities, about the likelihood of random outcomes in the decision tree. Paté-Cornell and Guikema [20] make no such assumptions about sharing probabilities and opposite objectives, but their solution method is less developed.

Insua *et al.* [21] propose an approach called *adversarial risk analysis* (ARA) to analyze counterterrorism decisions. ARA is a combination of game theory and Bayesian decision analysis. In its simplest form, ARA can mirror Parnell *et al.* [4]. However, there are two additional facets. First, ARA considers the state of knowledge about terrorist beliefs and preferences. For instance, if we believe that the probability that terrorists can obtain a particular biological agent is 0.05, we do not have to assume that they also believe this is the probability. Instead, we can place a prior distribution on their probability, iteratively sample their probability, and solve their decision to obtain a probability distribution over their actions. Second, ARA allows for simultaneous decisions in the parlance of game theory, not just sequential decisions like Fig. 4. The term simultaneous refers to a terrorist decision and a US decision where neither knows what the other has chosen when they make their decision. As simultaneous decisions are the purview of game theory, we shall leave discussion of such methods to that section of the encyclopedia.

DECISION ANALYSIS CONCLUSIONS

Decision analysis can be used in a variety of ways to help the decision maker understand the decision in its full context and to accurately assess alternatives and trade-offs among the fundamental objectives. We have discussed approaches to counterterrorism decisions that use decision trees and influence diagrams with single attributes, multiple objectives and multiattribute utility functions, and sequential game theory formulations including adversarial decisions. These formulations lay the groundwork for

substantial research to come that should prove beneficial not just for counterterrorist decisions.

REFERENCES

1. von Winterfeldt D, O’Sullivan TM. Should we protect commercial airplanes against surface-to-air missile attacks by terrorists *Decis Anal* 2006;3(2):63–75.
2. Bakir NO. A decision tree model for evaluating countermeasures to secure cargo at United States southwestern ports of entry. *Decis Anal* 2008;5(4):230–248.
3. Merrick JRW, McLay LA. Is screening cargo containers for smuggled nuclear threats worthwhile? *Decis Anal* 2010;7(2):155–171.
4. Parnell GS, Smith CM, Moxley FI. Intelligent adversary risk analysis: a bioterrorism risk management model. *Risk Anal* 2009;30(1):32–48.
5. Keeney RL. Modeling values for anti-terrorism analysis. *Risk Anal* 2007;27(3):585–596.
6. Keeney RL, Raiffa. H. *Decisions with multiple objectives*. New York: Wiley; 1976.
7. Leung M, Lambert JH, Mosenthal A. A risk-based approach to setting priorities in protecting bridges against terrorist attacks. *Risk Anal* 2004;24(4):963–984.
8. Haimes YY, Kaplan S, Lambert JH. Risk filtering, ranking, and management framework using hierarchical holographic modeling. *Risk Anal* 2002;22(2):383–397.
9. Apostolakis GE, Lemon DM. A screening methodology for the identification and ranking of infrastructure vulnerabilities due to terrorism. *Risk Anal* 2005;25(2):361–376.
10. Buckshaw DL, Parnell GS, Unkenholz WL, *et al.* Mission-oriented risk and design analysis of critical information systems. *Mil Oper Res* 2005;10(2):1938.
11. Feng T, Keller LR. A multiple-objective decision analysis for terrorism protection: potassium iodide distribution in nuclear incidents. *Decis Anal* 2006;3(2):76–93.
12. Keeney GL, von Winterfeldt D. Identifying and structuring the objectives of terrorists. CREATE technical report. Los Angeles (CA): University of Southern California; 2009. Available at <http://create.usc.edu/publications/KeeneyReport.pdf>.
13. Kunreuther H, Heal G. Interdependent security. *J Risk Uncertain* 2003;26(2–3):231–249.

14. Zhuang J, Bier VM. Balancing terrorism and natural disasters—defensive strategy with endogenous attacker effort. *Oper Res* 2007; 55(5):976–991.
15. Bier VM, Oliveros S, Samuelson L. Choosing what to protect: strategic defensive allocation against an unknown attacker. *J Public Econ Theory* 2007;9(4):563–587.
16. Zhuang J, Bier VM, Gupta A. Subsidies in interdependent security with heterogeneous discount rates. *Eng Econ* 2007;52(1): 1–19.
17. Bier VM. Game-theoretic and reliability methods in counter-terrorism and security. In: Wilson A, Limnios N, Keller-McNulty S, *et al.*, editors. *Mathematical and statistical methods in reliability, series on quality, reliability and engineering statistics*. Singapore: World Scientific; 2005. pp. 17–28.
18. Bier VM, Gratz ER, Haphuriwat NJ, *et al.* Methodology for identifying near-optimal interdiction strategies for a power transmission system. *Reliab Eng Syst Saf* 2007;92: 1155–1161.
19. Heal G, Kunreuther H. IDS models of airline security. *J Conflict Resolut* 2005;49(2): 202–217.
20. Paté-Cornell E, Guikema S. Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures. *Mil Oper Res* 2002;7(4):5–20.
21. Insua DR, Rios J, Banks D. Adversarial risk analysis. *J Am Stat Assoc* 2009;104(486): 841–854.

DECISION MAKING UNDER PRESSURE AND CONSTRAINTS: BOUNDED RATIONALITY

ALEX KIRLIK

SVEN BERTEL

Human Factors Division and Beckman
Institute for Advanced Science and
Technology, University of Illinois at
Urbana-Champaign, Urbana, Illinois

Concepts of rationality have evolved over centuries [1], and models for both augmenting human rationality, such as the many techniques found in this encyclopedia, and for describing and understanding unaided human rationality, continue to be revised [2]. In this light, our understanding of the “bounds” of human rationality is ever changing, and as such, the meaning of the term “bounded rationality” has both a historical lineage and a rapidly growing set of contemporary associations. The aim of this article is to both situate the concept of bounded rationality in a historical perspective and to trace the many ways in which the original concept has been, and continues to be, developed, elaborated, and revised through psychological research.

ORIGINS

The concept of bounded rationality is typically credited to Simon [3] who noted that, under the constraints and time pressure of much of everyday life, people are incapable of making decisions according to the types of normative procedures and models described previously in this section. Simon [4] put the matter bluntly:

“The SEU model assumes the decision maker contemplates, in one comprehensive view, everything that lies before him. He understands the range of alternative choices open to him, not only at the moment but the whole panorama of the future. He understands the consequences of each of the available

choice strategies, at least up to the point of being able to assign a joint probability distribution to future states of the world. He has reconciled or balanced all his conflicting partial values and synthesized them into a single utility function that orders, by his preferences for them, all these future states of the world. When these assumptions are stated explicitly, it becomes obvious that SEU theory has never been applied, and never can be applied—with or without the largest computers—in the real world.”

Simon goes on to explain that purported “applications” of SEU (subjective expected utility) theory always involve either substituting a highly abstracted problem for the actual problem in its full complexity, or else a microproblem that has been neatly carved out of a more complex reality.

Satisficing and Simon’s Scissors

Two concepts are central to Simon’s original conceptualization of bounded rationality, which he considered to be the “behavioral alternative” to normative theory such as SEU. The first is “satisficing.” Whether it is bees searching for nectar, trout searching for insects or people searching for a restaurant, Simon observed that biological creatures do not optimize but instead tend to select the first decision option that exceeds a specified aspiration level. Although some contemporary psychological research conducted in the name of bounded rationality nevertheless holds on to one or another notion of constrained optimization, it is important to note that Simon himself explicitly rejected this view: “bounded rationality is not the study of optimization in relation to task environments” [5].

As will be discussed in the following, those following in Simon’s footsteps have fractured into two general camps on whether or not bounded rationality should be understood in optimizing terms or instead by appeal to some other metaphors or mechanisms. At this point, suffice to say that Simon’s original

formulation of bounded rationality made no appeal to evolutionary, biological or cognitive mechanisms that optimize in the sense of the term used elsewhere throughout this *Encyclopedia*.

A second core concept in Simon's original conceptualization is the metaphor of a pair of scissors to describe the relationship between a decision maker and a decision environment. A deeper understanding of bounded rationality can emerge only by research that explicitly examines both "the structure of task environments and the computational capabilities of the actor" in terms of "a scissors [with] two blades" [6]. Just as the cutting power of a pair of scissors arises in the mesh of the two blades, the decision competence of an actor can be understood only by examining the degree to which an adaptive mesh exists between the actor's decision strategy and the structure of a decision problem.

Simon introduced this metaphor to hammer home the notion that what is rational and what is not is not, *per se*, a property of the decision maker, nor a decision technique. Instead, rationality must be understood to be a relational property of the actor–environment system. In some environments (such as a casino), using a decision strategy of blind guessing may be said to be completely rational in this light. As is the case with the concept of satisficing, Simon's ecologically oriented scissors metaphor for understanding bounded rationality has continued to be influential. However, this influence has played out in different ways in the study of cognition, judgment, and decision making by various psychological schools. While the alternatives will be discussed in the following, here it is sufficient to observe that all those performing research today in the name of bounded rationality make some type of appeal to the notion that the essence of rationality lies in the degree of fit between cognition and the structure of a decision or task environment.

WHY IS BOUNDED RATIONALITY RELEVANT TO OR AND MS?

Before moving on to describe how the original conceptualization of bounded rationality

has played out in contemporary psychological research, it is worthwhile to address the question of why the student of operations research and management science should care about the psychology of judgment and decision making. Aren't the many methods, models, and techniques described elsewhere in this *Encyclopedia* a testament to the bounded (i.e., limited) nature of unaided human cognition, and thus to be considered as replacements for human cognition whenever possible? Or perhaps even more generally, even if a formal technique for decision making is unavailable, should not a decision maker always strive to be as analytical (i.e., as unbounded) in thinking as possible?

Here, we are not referring to the trivial observation that most people believe that some decisions are still best left in the hands of the expert or the wise, or that none of us break out the full resources of decision analysis or related analytical techniques for every mundane problem. (Though some cases should make us pause: Gigerenzer (7, p. 26) relates the case of Harry Markowitz, winner of the 1990 Nobel prize for his research on optimal asset allocation. Despite inventing an optimization technique for maximizing return while minimizing risk, Markowitz himself used the widely used heuristic of allocating $1/N$ of his assets to each of his N investments to manage his own retirement portfolio—Gigerenzer goes on to describe a study in which none of a dozen asset allocation algorithms, including Markowitz's, was able to outperform the $1/N$ heuristic in seven stock allocation problems).

The reasons that we *do* raise this question are two-fold. First, an enormous body of research has been conducted in the guise of bounded rationality demonstrating often stark departures of the outcomes of intuitive human judgment and decision making with the prescriptions of rational models (see ***Paradoxes and Violations of Normative Decision Theory***). Much of this research, which is often taken to paint a rather dismal picture of unaided human cognition, must be considered in light of the fact that it typically does not truly pit intuition versus analysis on a level playing field.

Rather, such research typically compares the judgments and decisions of unaided cognition against the outcome of an analytical procedure that has been flawlessly executed by a human who is an expert at using that procedure. Only rarely, such as in the case of Hammond *et al.* [8], have studies been conducted to *directly* compare a human's use of an intuitive strategy for performing a task with that *same human's* use of an analytical procedure for performing the task. Hammond and his coworkers found that, in their study of professional highway engineers (judging highway capacity, safety, and aesthetics) that performance was best when the mode of cognition induced (e.g., analysis versus intuition) was most consistent with the nature of the various judgment tasks (i.e., in some cases intuition outperformed analysis). Results such as these should make us pause when making any blanket statements about recommending that all decision makers should try to be as analytical as possible, in all task situations.

Second, and due largely to the recent failures of the US financial system, much serious thought is now being given to the wisdom of turning over near-complete responsibility for the real-time evaluation of risk and even trading authority to computer-based models resulting from financial engineering, economics, and related research. Bernstein (9, p. 6) has summed up the matter as follows:

The story that I have to tell is marked all the way through with a persistent tension between those who assert that the best decisions are based on quantification and numbers, determined by patterns of the past, and those who base their decisions on more subjective degrees of belief about the uncertain future. This is a controversy that has never been resolved.

Lacking any resolution, those involved with developing and prescribing the use of analytical techniques such as those presented in this encyclopedia should be mindful of the trade-offs one incurs when determining whether particular judgments or decisions are best made by analytical models, by knowledgeable humans, or perhaps by joint human–technology systems [10].

To take just one example, intuitive judgment often yields estimates that are imprecise, yet often relatively accurate in the sense of being centered around the correct answer, while the errors of analysis tend to be much fewer in number but potentially sizable in magnitude [8]. This observation, originally due to mid-twentieth-century psychologist Egon Brunswik (11, Fig. 22) emerged from Brunswik's pioneering investigations into the relative strengths and weaknesses of highly intuitive processes such as perception and the more analytical processes we typically describe as "thinking." This phenomenon is already familiar to anyone who has graded an exam on which a student using a calculator (an analytical procedure) provided a probability as an answer with a value less than 0 or greater than 1. It is difficult to conceive how any student with an appreciation of the meaning of probability could arrive at such an answer using intuitive, as opposed to analytical, cognition. Trade-offs between intuition and analysis are especially important when prescription involves training or implementing decision-support technology in operational contexts, where decision makers are unlikely to possess the expertise needed to understand the rationale underlying, and any limitations of, analytical decision strategies they may be trained to use, or assumptions and approximations associated with the decision algorithms embodied in computational aids.

CONTEMPORARY RESEARCH ON BOUNDED RATIONALITY

As mentioned previously, Simon's original conceptualization of bounded rationality has been cited as being central to a wide variety of contemporary research programs on judgment and decision making. Indeed, many of these programs are topics of entire articles in this section (see *Paradoxes and Violations of Normative Decision Theory*; *Fast and Frugal Heuristics*; and *The Naturalistic Decision Making Perspective*). Each of these articles provides a definitive statement of each approach and overview of research in

each respective area, so we will not deal with these areas in depth here.

Instead, experience has taught us that students of bounded rationality, whether they be from psychology or OR/MS, seeking an introduction to this behavioral research often experience significant difficulty in coming to understand how each of the many contemporary approaches to bounded rationality relate. To take just one notable example, at a very general level one can see vast differences across these approaches and research programs in the overall evaluation of the competence of human judgment and decision making, ranging from highly pessimistic to largely optimistic. As such, the purpose of the remainder of this article will be to provide a set of distinctions that will hopefully organize this large, and often apparently conflicting, literature.

Multiple Conceptions of Bounded Rationality: Some Distinctions

In the following sections, we make three distinctions to clarify the ways in which bounded rationality research has been extended since Simon's original work.

Judgment versus Decision Making. Although differences of opinion exist over whether it is useful to make a sharp distinction between bounded rationality in judgment as opposed to bounded rationality in decision making [12,13], we have found such a distinction to be of heuristic value. A decision is an irrevocable selection of a course of action from among a set of alternatives. In contrast, a judgment is an estimation (or prediction, inference, diagnosis, etc.) of a past, present, or future state, event, or value of a variable. In this light, a weather forecaster *judges* that a hurricane will make landfall at a specified location, and on this basis, along with a consideration of various costs and benefits, a public official may *decide* to order an evacuation on the basis of this judgment.

This distinction is useful for organizing the literature on bounded rationality because clear cases exist where a research program focuses solely on judgment (with little to say about values or utilities) and another

research program focuses instead on decision making. In the latter case, to focus on how competing alternatives are selected, the (often) significant amount of information about alternatives, consequences, and their likelihoods, and so forth are simply provided to the decision maker, when in reality, this information must be obtained by the decision maker using prior judgment. Clear cases of such research programs involve much of the material in the related articles mentioned above, in which experimental participants are presented with lotteries consisting of multiple options from which the participant is asked to select. In naturalistic decision-making research, in contrast, the tasks being studied nearly always involve studies of humans, often professionals, making various judgments (e.g., situation assessments) followed by the selection of an irrevocable course of action (e.g., decision making).

This being said, no heuristic is perfect, and clear cases exist where research is indeed focused on choices and decisions, but where the cognitive mechanisms involved are either primarily or even solely judgmental in nature. Judgment researchers understand tasks to be comprised of a set of probabilistic cues, or fallible indicators [11], and a criterion to be estimated on the basis of those cues (as opposed to a decision analyst or researcher who understands a task to have the content and format of a decision tree). For example, a typical judgment task would involve estimating the prospective success (the criterion) of a graduate student on the basis of cues such as undergraduate GPA and standardized test scores that are known to have only limited predictive validity.

In most of the literature on humans' abilities to make such judgments, the gold standard for the assessment of judgment performance is the best linear-additive model relating cues to the criterion as determined by linear regression [14,15]. In such judgment models, the β weights associated with each cue are interpreted to indicate the relative weight or importance the judge places on that cue. High levels of judgment performance are associated with cue weightings that mirror the objective or "ecological" validities of the cues in the task ecology,

although Dawes [16] has demonstrated that “improper” (i.e., equally weighted rather than ecologically accurate) linear models are often sufficient to describe human judgment, once the relevant cues are known.

Some of the difficulty involved in maintaining an adequate distinction between judgment and decision making, and thus in understanding how these literatures are related, almost certainly arises due to the fact that in many cases the gold (prescriptive) standard for the assessment of decision, rather than judgment, performance also takes the form of a linear-additive model. Assuming certain conditions, such as mutual utility independence, are met, normative models in multiattribute decision tasks [e.g., MAUT (multi-attribute utility theory)] also take the form of a linear-additive model. But in contrast to the case in judgment, MAUT decision weights represent utilities, and unlike in the case of judgment, these utilities need have no environmental anchor or correct value (instead, it is sufficient merely that certain consistency conditions apply).

As such, the fact that linear-additive models are commonly used to describe both the selection of an action (decision making) and rendering an estimation of an external criterion (judgment) contributes to the difficulty in maintaining a clear distinction between research done on judgment and decision making from a bounded rationality perspective. Models of judgment, that is, of cue-criterion inference, are often sufficient to model decision making or choice in tasks where utility functions do not play an important role and instead, the task boils down to selecting an option solely based on external cue values (e.g., the price and size of a flat panel TV).

Bounded Rationality versus Constrained Optimization. As mentioned previously, Simon himself rejected the idea that bounded rationality referred to cognitively constrained optimization with respect to the structure of a decision environment. Despite this view, it is fair to say that the lion’s share of research performed today in the name of bounded rationality does appeal to a notion of constrained optimization, or

optimization under constraints [17]. For example, Kahneman and Tversky’s Prospect Theory [18] is commonly understood to exist as an illustration of bounded rationality in that it makes provisions (such as framing effects and the lack of a veridical estimation of probabilities) for human limitations in being able to live up to the Olympian ideal of SEU (which Simon originally meant to reject, not merely revise, in his own conception of bounded rationality). However, despite these behavioral provisions, Prospect Theory maintains the general notion that a human decision maker will still seek to optimize to the extent the bounds of cognition (e.g., biases and heuristics, resource limitations) allow.

Some researchers holding true to Simon’s original concept reject the view that any theory of judgment and decision making that appeals to notions of cognitively constrained optimization or maximization should be considered to be a theory of bounded rationality [19] (see *Fast and Frugal Heuristics* and *The Naturalistic Decision Making Perspective*). Despite these views, research programs that construe bounded rationality in terms of constrained optimization to the structure of decision or task environments have flourished in recent decades. Perhaps the most influential has been John Anderson’s research program on “rational analysis” [20]. Characterizing the cognitive theories of the time as a “random set of postulates let loose on the world” and as “bizarre and implausible,” Anderson sought to ground cognitive theory more firmly in the structure of the environment, and the demands the environment makes on cognition. Rational analysis is “an explanation of human behavior based on the assumption that it is optimized somehow to the structure of the environment” (21, p. 471).

Although clearly an approach to modeling cognition as constrained optimization, Anderson [20] nevertheless argued that rational analysis was not incompatible with Simon’s concept of satisficing. However, Simon [22,23] criticized rational analysis for placing too much emphasis on the analysis of the decision environment and not enough emphasis on the knowledge and strategies of the

human decision maker. Whether in response to Simon or not, after developing rational analysis Anderson then turned his attention to complementing rational analysis with a theory of a “cognitive architecture” capable of embodying a scientist’s assumptions of both the declarative (know that) and procedural (know how) knowledge that human cognition brings to a task. This cognitive architecture, adaptive control of thought-rational (ACT-R) [24], has been quite influential in cognitive science and has been used to model a variety of cognitive tasks including decision making, although Belavkin [25] discusses problems associated with the fact that ACT-R does so using expected value maximization, and suggests alternative formulations (also see Ref. 26. Oaksford and Chater [27] provide a discussion of how rational analysis might be brought in line with Simon’s original conception of bounded rationality by integrating aspects of ACT-R and Gigerenzer’s program of fast and frugal heuristics (see *Fast and Frugal Heuristics*). Byrne *et al.* [28] provide one concrete example of this type of integration in their research modeling the decision strategies of experienced airline pilots using ACT-R to implement a set of fast and frugal heuristics.

Coherence versus Correspondence Aspects of Rationality. A third distinction useful for understanding the diverse ways in which the original conceptualization of bounded rationality has been developed in judgment and decision making research concerns Kenneth Hammond’s distinction between coherence and correspondence as alternative benchmarks for assessing rationality [29–31]. Hammond notes that, prior to the pioneering research of Tversky and Kahneman [32], the dominant criterion for assessing the quality of behavior in psychological research was correspondence competence. Correspondence competence refers to whether human judgments (perceptions, assessments, predictions, etc.) of aspects of the external world are empirically accurate. The previous discussion of judgment (as opposed to decision) tasks provides a clear instance of correspondence competence: for example, on the basis of the available information or

cues, does a person render an estimate of the value of an external criterion (say, the distance of a mountain) that is empirically accurate? How accurate is it?

Hammond notes that, in their influential heuristics and biases research program [32], Tversky and Kahneman made a subtle but ultimately extremely important shift of reference in evaluating human behavior away from correspondence competence even while appealing to such competence for explanatory purposes. Tversky and Kahneman (32, p. 1124) wrote:

The subjective assessment of probability resembles the subjective assessment of physical quantities such as distance or size. These judgments are all based on data of limited validity, which are processed according to heuristic rules. For example, the apparent distance of an object is determined in part by its clarity. The more sharply the object is seen, the closer it appears to be. This rule has some validity, because in any given scene the more distant objects are seen less sharply than nearer objects. However, the reliance on this rule leads to systematic errors in the estimation of distance. Specifically, distances are often overestimated when visibility is poor because the contours of objects are blurred. On the other hand, distances are often underestimated when visibility is good because the objects are seen sharply. Thus, the reliance on clarity as an indication of distance leads to common biases. Such biases are also found in the intuitive judgment of probability.

Concerning the final statement in this passage, Hammond (29, p. 30) observes:

But “such biases” are not found in the intuitive judgment of probability. The authors’ error lies in failing to recognize that the “errors” in the two circumstances are very different. In intuitive judgments of the distance of an object, an error occurs when a person’s subjective judgment of distance does not agree with an objective measure of that distance. In the intuitive estimate of a probability, an error occurs when the estimate does not agree with the calculation of a probability from a mathematical formula.

By shifting the criterion against which human judgments are measured from external, empirical ones to those resulting from a

mathematical formula, Hammond notes that Tversky and Kahneman shifted the yardstick by which rationality is measured from empirical accuracy to the ability to be internally coherent in one's reasoning. Why is this distinction important?

The point was made earlier that different research programs and traditions in psychology all claiming to follow, in one way or another, in the footsteps of Simon's bounded rationality have come to quite different evaluations about the competence of human judgment and decision making. Hammond's coherence versus correspondence distinction is useful in organizing the apparently conflicting empirical findings in this research.

As a rough cut, research that assesses bounded rationality from a coherence perspective, against yardsticks of probability theory and SEU theory (both theories solely of how to be internally consistent, *not* empirically accurate) often tend to find human cognition frail and wanting. In contrast, research that assesses bounded rationality from the perspective of empirical accuracy (see as in *Fast and Frugal Heuristics* and *The Naturalistic Decision Making Perspective*), generally paint a much more optimistic picture of bounded rationality. If Simon was correct that no decision maker can truly apply the normative theories created to ensure internal coherence as a criterion for rational behavior, perhaps we should not be surprised that, when it comes to assessing people instead on their empirical success, that is, on their correspondence competence, they fare much better indeed.

CONCLUSION

The origins of bounded rationality in the pioneering research of Herbert A. Simon have been presented, and the influence of these original ideas have been traced forward to examine how they have influenced a wide variety of contemporary lines of psychological research on both the capabilities and mechanisms of human judgment and decision making. Perhaps not surprisingly, the quality of behavior produced by bounded rationality appears to be higher in tasks for which the

concept was invented (decision making under pressure and the realistic constraints of life) than in tasks for which our normative models are restricted to achieving internally coherent thought, rather than empirically adaptive behavior.

REFERENCES

1. Daston LJ. Classical probability in the enlightenment. Princeton (NJ): Princeton University Press; 1988.
2. Gigerenzer G, Selten R, editors. Bounded rationality: the adaptive toolbox. Cambridge (MA): MIT Press; 2001. pp. 1–12.
3. Simon HA. Models of man. New York: John Wiley and Sons, Inc.; 1957.
4. Simon HA. Reason in human affairs. Stanford (CA): Stanford University Press; 1983.
5. Simon HA. Models of my life. New York: Basic Books; 1991.
6. Simon HA. Invariants of human behavior. *Ann Rev Psychol* 1990;41:1–19.
7. Gigerenzer G. Gut feelings: the intelligence of the unconscious. New York: Viking; 2007.
8. Hammond KR, Hamm RM, Grassia J, *et al.* Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Trans Syst Man Cybern* 1987;17:753–770.
9. Bernstein PL. Against the gods: the remarkable story of risk. New York: Wiley; 1998.
10. Kirlik A. Adaptive perspectives on human-technology interaction. New York: Oxford University Press; 2006.
11. Brunswik E. Perception and the representative design of psychological experiments. Berkeley (CA): University of California Press; 1956.
12. Connolly T, Arkes HR, Hammond KR, editors. Judgment and decision making: an interdisciplinary reader. 2nd ed. Cambridge: Cambridge University Press; 2000. pp. 1–11.
13. Goldstein WM, Hogarth RM. Research in judgment and decision making: currents, connections, and controversies. New York: Cambridge University Press; 1997.
14. Dawes RM, Corrigan B. Linear models in decision making. *Psychol Bull* 1974;81(2):95–106.
15. Karelaia N, Hogarth RM. Determinants of linear judgment: a meta-analysis of lens studies. *Psychol Bull* 2008;134:404–426.

16. Dawes RM. The robust beauty of improper linear models in decision making. *Am Psychol* 1979;34:571–582.
17. Goodie AS, Ortman A, Davis JN, *et al.* Demons versus heuristics in artificial intelligence, behavioral ecology, and economics. In: Gigerenzer G, Todd PM, ABC Research Group, editors. New York: Oxford University Press; 1999. pp. 327–356.
18. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47:263–291.
19. Gigerenzer G, Todd PM. Fast and frugal heuristics: the adaptive toolbox. In: Gigerenzer G, Todd PM, the ABC Research Group, editors. New York: Oxford University Press; 1999. pp. 3–34.
20. Anderson JR. The adaptive character of thought. Hillsdale (NJ): Erlbaum; 1990.
21. Anderson JR. Is human cognition adaptive? *Behav Brain Sci* 1991;14(3):471–485.
22. Simon HA. Cognitive architectures and rational analysis: comment. In: Van Lehn K, editor. Architectures of intelligence: the 22nd Carnegie Mellon symposium on intelligence. Hillsdale (NJ): Erlbaum; 1991.
23. Simon HA. What is an explanation of behavior? *Psychol Sci* 1992;3:150–161.
24. Anderson JR, Bothell D, Byrne MD, *et al.* An integrated theory of the mind. *Psychol Rev* 2004;111:1036–1060.
25. Belavkin RV. Towards a theory of decision-making without paradoxes. Proceedings of Seventh International Conference on Cognitive Modeling. 2006. pp. 38–43.
26. Gray WD, Schoelles MJ, Sims CR. Adapting to the task environment: explorations in expected value. *Cogn Syst Res* 2005;6: 27–40.
27. Oaksford M, Chater N. Rational models of cognition. New York: Oxford University Press; 1998.
28. Byrne MD, Kirlik A, Fick CS. Kilograms matter: rational analysis, ecological rationality, and closed-loop modeling of interactive cognition and behavior. In: Kirlik A, editor. Adaptive perspectives on human-technology interaction. New York: Oxford University Press; 2006. pp. 267–286.
29. Hammond KR. Judgments under stress. New York: Oxford University Press; 2000.
30. Hammond KR. Coherence and correspondence theories in judgment and decision making. In: Connolly T, Arkes HR, Hammond KR, editors. Judgment and decision making: an interdisciplinary reader. 2nd ed. Cambridge: Cambridge University Press; 2000. pp. 53–65.
31. Hammond KR. Beyond rationality: the search for wisdom in a troubled time. New York: Oxford University Press; 2007.
32. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185:1124–1131.

DECISION MAKING WITH PARTIAL PROBABILISTIC OR PREFERENCE INFORMATION

LUIS V. MONTIEL
J. ERIC BICKEL
The University of Texas, Texas,
USA

INTRODUCTION

Mathematical models often require the use of epistemological probabilities to account for possible states of the world. In general, we could think of such probabilities as an expression of our beliefs, or as a consequence of reasoning based on specific information. In any case, the complete characterization of epistemological probabilities is difficult, and in some cases is beyond the reach of the decision maker (DM). An analogous problem can be observed in the elicitation of utility and value functions [1], where the specification of the parameters is subject to the DM's attitude. In this article, we present a general procedure to solve decision problems under uncertainty when the epistemological probabilities or utility functions are incompletely specified.

The problem of interest to the DM is to find a unique optimal alternative among all possible choices given the underspecified probabilities [2, 3] or utilities [4–8]. However, such under specification generates scenarios where two or more alternatives are potentially optimal. Hence, the challenge is to determine up to what degree an alternative is optimal.

The methods presented by Yu [9], Sarin [10], White *et al.* [11], Moskowitz *et al.* [12] and Greco *et al.* [8] excluded dominated alternatives by means of solving a series of optimization problems. Others such as [4, 5] and [6] use similar optimization formulations to extract information from the set of possible weights. However, the final outcome of all of these methods is often a set of potentially

optimal alternatives, any of which can be optimal under some particular scenario. Hence, these methods can fail to recognize the single best alternative within the set, leaving the DM with no clear route of action.

In contrast to the previous methods, Refs 3, 7 and 13 proposed approaches in which each alternative is evaluated according to its region of optimality. These methods rely on the analysis of the volume of the relative interior of the set of all possible probabilities (or weights) and provide information about the level of optimality of each alternative according to the portion of the volume for which each alternative is the best, second best, and so on. Hence, it is possible to provide guidance to the DM based on the fact that two potentially optimal alternatives can have different degrees of optimality.

In particular, this article reviews the joint distribution simulation procedure (JDSIM) proposed by Montiel and Bickel [13]. This procedure explores the volume of the relative interior of any arbitrarily constrained polytope without the need to compute the limits of an n -dimensional integral. Hence, it is possible to use the concepts developed by Charnetski [3], and Lahdelma and Salminen [7] beyond the trivial case of the unit simplex, and apply scores of optimality in highly constrained problems in large dimensional subspaces, which was computationally intractable with previous techniques.

The remainder of this article is organized as follows. The section titled “The Truth Set” presents a review of JDSIM basic concepts to encode partial information and develops the constraints and the sampling methodology to evaluate multiple alternatives. The section titled “Alternatives and Truth Sets” provides an overview of the tools to analyze different alternatives when there are neither dominant nor dominated alternatives. The section titled “Example of Index and Regions

of Optimality” presents an example to illustrate the use of JDSIM, and the section titled “Conclusions” concludes and describes possible lines of future research.

THE TRUTH SET

As described by Howard [14], a decision requires three main elements: a set of alternatives ($\mathbf{J}$), a preference rule often defined by a utility function $u(\cdot)$, and information, which is the source of the epistemological probabilities. We will concentrate on the epistemological probabilities. However, as shown by Charnetski [3], Charnetski and Soland [15], and Lahdelma and Salminen [7], the problem of incompletely specified utility functions is mathematically equivalent. For example, Refs 16 and 1 presented applications for joint probability distributions and multi-attribute utility functions, respectively.

In decisions with partial but incomplete information, it is necessary to examine the set of all possible realizations, which we named the “truth set” ($\mathbb{T}$). In the case of epistemological probabilities, that set contains all joint distributions that match the partial beliefs. When $\mathbb{T}$ is defined by linear constraints, that is, the beliefs can be expressed using linear inequalities, the truth set becomes closed, continuous, convex, and

bounded [17]. More precisely, the truth set is defined as: $\mathbb{T} = \{\mathbf{p} | \mathbf{A}\mathbf{p} \leq \mathbf{b}, \mathbf{p} \geq 0\}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix of constraints, $\mathbf{b} \in \mathbb{R}^m$ is a vector of beliefs, and $\mathbf{p} \in \mathbb{R}^n$ is a joint distribution with n joint elements and subject to m beliefs.

Truth Sets with Partial Information

A simple example will illustrate the effects of partial information. Let V_1 and V_2 be two binary random variables with marginal probabilities $P(V_1 = 1) = p$ and $P(V_2 = 1) = q$. If p and q are unknown, the joint distribution

$$P(V_1 = 1, V_2 = 1) = P_1,$$

$$P(V_1 = 1, V_2 = 0) = P_2,$$

$$P(V_1 = 0, V_2 = 1) = P_3,$$

$$P(V_1 = 0, V_2 = 0) = 1 - P_1 - P_2 - P_3,$$

can be expressed as a non-negative vector (P_1, P_2, P_3) whose elements sum to less than one. Figure 1 shows the truth set when there are no expressed beliefs about V_1 or V_2 . Figure 1a and b present the front and back views of $\mathbb{T}$, where every gray point represents a full joint distribution in the interior of a three-dimensional unit simplex.

As new information is assessed, the truth set is reduced, providing a better understanding of the behavior of the uncertainties. For example, assume that we believe that

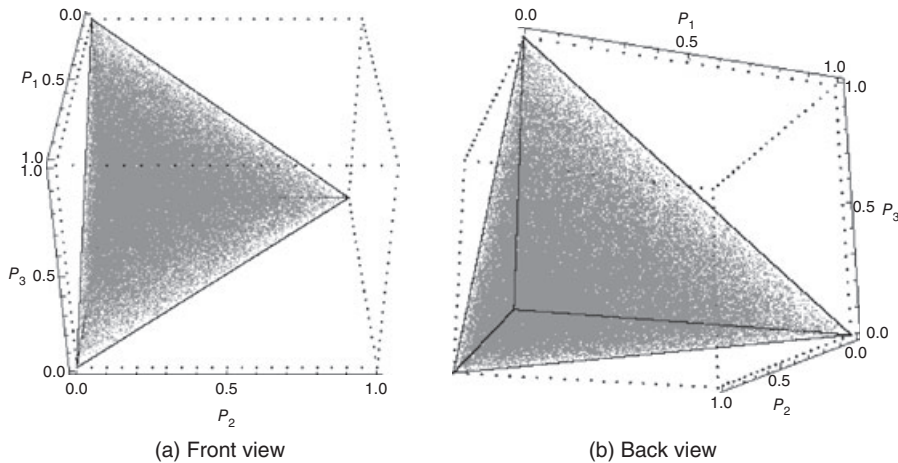


Figure 1. Truth set for two binary variables with unknown information.

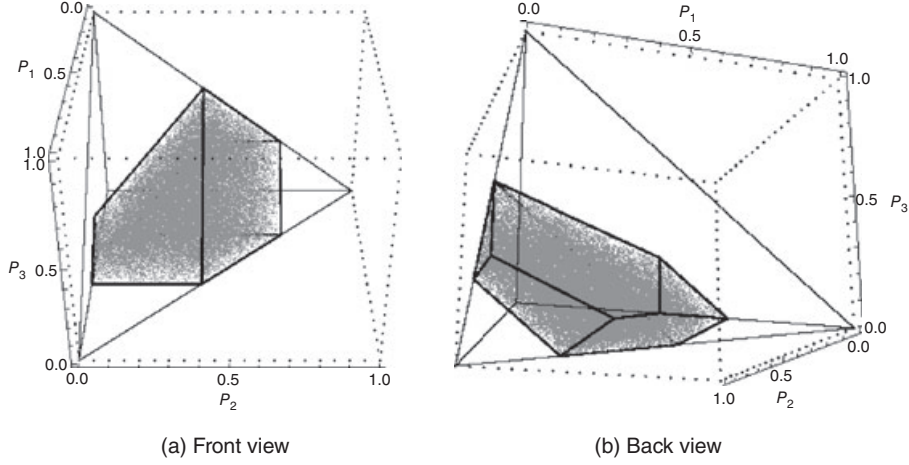


Figure 2. Truth set for two binary variables with beliefs $p \geq 0.4$ and $0.6 \geq q \geq 0.3$.

$p \geq 0.4$ and $0.6 \geq q \geq 0.3$ for $p = P_1 + P_2$ and $q = P_1 + P_3$. Then, the set of all possible joint distributions is reduced to a new set $\mathbb{T}'$ as shown in Fig. 2, where Fig. 2a and b present the front and back views of $\mathbb{T}'$, which are oriented to facilitate comparison to Fig. 1a and b.

These figures show that the available information is insufficient to specify a unique joint distribution. However, the new constraints reduce the space of possible joint distributions, and possibly reduce the number of potentially optimal alternatives.

Constructing the Truth Set

The truth set can encode any beliefs that are expressed in terms of linear inequalities. A modified version of JDSIM can also include nonlinear inequalities under the condition that $\mathbb{T}$ does not violate the convexity assumption. However, the additional flexibility generates a reduction on the speed of the algorithm. Moreover, most of the reliable information that can be collected from the DM can be expressed in linear inequalities (marginal and pairwise probabilities, moments, preferences, etc.), making the use of nonlinear inequalities an uncommon event. In the following, we present some of the assessment constraints that can be included in $\mathbb{T}$ and refer the reader

to [18], and [13, 19] for a broader overview on assessment constraints.

To simplify the notation, we use the dot notation as follows. We order the random variables as $(V_1, V_2, \dots, V_N)$ for a total of n possible joint events. For each joint event, let $P(V_1 = v_1, V_2 = v_2, \dots, V_N = v_n) = p_{v_1, v_2, \dots, v_n}$, and replace the realization v_i with a “.” whenever we take the sum over all possible outcomes of V_i . For example, if $N = 3$, then $p_{v_1, \cdot, \cdot}$ is defined as in Equation (1).

$$p_{v_1, \cdot, \cdot} = \sum_{i,j} p_{v_1, i, j}, \quad \forall V_2 = i, V_3 = j. \quad (1)$$

Hence, $p_{v_1, \cdot, \cdot}$ represents the marginal probability $P(V_1 = v_1)$, and $p_{v_1, \cdot, v_3}$ represents the pairwise probability $P(V_1 = v_1, V_3 = v_3)$.

Matching Necessary Conditions. To match the necessary conditions, the joint probability mass function (pmf) must sum to one and each probability must be non-negative. We represent these constraints with Equation (2a) and (2b), where the set $\mathbf{V} = \{(i, j, \dots, k) \mid V_1 = i, V_2 = j, \dots, V_N = k\}$ is the Cartesian product of the outcomes of the random variables V_1 to V_N .

$$\sum_{(i, j, \dots, k) \in \mathbf{V}} p_{i, j, \dots, k} = 1, \quad (2a)$$

$$p_{i, j, \dots, k} \geq 0, \quad \forall (i, j, \dots, k) \in \mathbf{V}. \quad (2b)$$

Equation (2a) and (2b) give the necessary and sufficient conditions for $\mathbf{p}$ to be a pmf and are required in all cases. Equation (2a) reduces the dimension of the polytope $\mathbb{T}$ from n to $n - 1$, and Equation (2b) limits $\mathbb{T}$ to the positive orthant.

Matching Marginal and Pairwise Probabilities. Assessments regarding marginal and pairwise belief probabilities can be implemented as Equation (3a) and (3b), respectively. These equations can be extended to cover three-way, four-way, or higher-order joint probability assessments.

$$\sum_{(v_1, i, j, \dots, k) \in \mathbf{V}} p_{v_1, i, j, \dots, k} \geq P(V_1 = v_1), \quad (3a)$$

$$\sum_{(v_1, i, v_3, \dots, k) \in \mathbf{V}} p_{v_1, i, v_3, \dots, k} \geq P(V_1 = v_1, V_3 = v_3), \quad (3b)$$

Matching Moments. When the marginal probabilities are unknown and the random variables take numerical values, it is possible to match beliefs about the moments. For example, if V_1 takes values in $\{v_1^1, v_1^2, \dots, v_1^k\}$ where $v_1^i \in \mathbb{R}$, we can use Equation (4) as follows.

$$\sum_{i=1}^k (v_1^i)^z \cdot p_{v_1^i, \dots, \dots} \geq E[(V_1)^z], \quad (4)$$

where $p_{v_1^i, \dots, \dots}$ is a summation over all joint event probabilities with $V_1 = v_1^i$.

Equation (4) matches the z th moment of V_1 . For $z = 1$, we can match the expected value, and for $z = 2$ we can match the second raw moment. $z = 0$ is simply a restatement of Equation (2a) and the requirement that the probabilities sum to one. Because the outcomes $\{v_1^1, v_1^2, \dots, v_1^k\}$ are known, the constraint is linear in the joint probabilities $p_{v_1, j, \dots, k}$.

Sampling the Truth Set

After characterizing $\mathbb{T} = \{\mathbf{p} | \mathbf{A}\mathbf{p} \leq \mathbf{b}, \mathbf{p} \geq 0\}$ using Equations (2), (3), (4), or any other linear constraint that expresses probabilistic beliefs [17, 18, 20], JDSIM proceeds to create a collection of joint distributions uniformly spread over the volume of $\mathbb{T}$. To sample the interior of the truth set, we use a Markov Chain Monte Carlo procedure named Hit and Run (HR), which was proposed by Smith [21]. This procedure is the fastest-known algorithm to sample the interior of an arbitrary polytope [22]. HR provides a sampling method that allowed us to test problems considerably larger [17] than those considered by Charnetski [3], and Lahdelma and Salminen [7].

HR creates a random grid of joint distributions over $\mathbb{T}$. Hence, each subset of $\mathbb{T}$ of equal volume has approximately the same number of distributions. The algorithm is briefly described below and illustrated in Figure 3. Montiel and Bickel [13] provided more detail regarding this procedure.

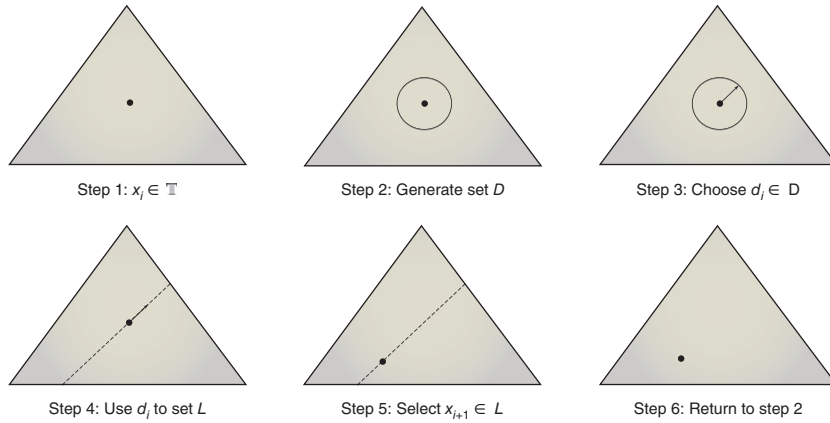


Figure 3. Illustration of the sampling procedure in two dimensions.

- Step 1: Set $i = 0$ and select an arbitrary point $\mathbf{p}_i \in \mathbb{T}$.
- Step 2: Generate a set $D \subseteq \mathbb{R}^n$ of directions.
- Step 3: Choose a random direction $\mathbf{d}_i$ uniformly distributed over D .
- Step 4: Find the line set $L = \mathbb{T} \cap \{\mathbf{p} | \mathbf{p} = \mathbf{p}_i + \lambda \mathbf{d}_i, \lambda \text{ a real scalar}\}$.
- Step 5: Generate a random point uniformly distributed over $\mathbf{p}_{i+1} \in L$.
- Step 6: If $i = N$, stop. Otherwise, set $i = i + 1$ and return to Step 2.

ALTERNATIVES AND TRUTH SETS

Dominance and Potential Optimality

When the characterization of the joint distribution is unique and the set of possible alternatives ($\mathbf{J}$) is finite, standard procedure is to find the alternative that maximizes the expected utility, for some utility function $u(\mathbf{x}^j)$ over a vector of attributes $\mathbf{x}^j, \forall j \in \mathbf{J}$. However, in the case of incomplete or partial information, the number of joint distributions is infinite. This generates three possible relevant scenarios related to an alternative $i \in \mathbf{J}$ with respect to $\mathbb{T}$.

Scenario one: alternative i dominates all other alternatives. Then:

$$\forall \mathbf{p} \in \mathbb{T}, \text{ and } \forall j \in \mathbf{J}, \text{ such that} \\ E_{\mathbf{p}}[u(\mathbf{x}^i)] \geq E_{\mathbf{p}}[u(\mathbf{x}^j)], \quad (5)$$

with at least one strict inequality. In this scenario, the DM has no ambiguity in his/her optimal selection, there is only one alternative to be implemented.

Scenario two: alternative i is dominated by at least one other alternative. Then:

$$\forall \mathbf{p} \in \mathbb{T}, \exists j \in \mathbf{J} \text{ such that} \\ E_{\mathbf{p}}[u(\mathbf{x}^j)] \geq E_{\mathbf{p}}[u(\mathbf{x}^i)], \quad (6)$$

with at least one strict inequality. In this scenario, the DM can remove alternative i and reduce the optimal selection problem. This action does not eliminate the ambiguity on the optimality of the rest of the alternatives, it only reduces the size of the problem.

Scenario three: alternative i is a potential optimal alternative. Then:

$$\forall j \in \mathbf{J}, \exists \mathbf{p} \in \mathbb{T} \text{ such that } E_{\mathbf{p}}[u(\mathbf{x}^i)] \geq E_{\mathbf{p}}[u(\mathbf{x}^j)], \quad (7)$$

with at least one strict inequality. This scenario is common in decision problems and provides neither a clear optimal choice nor a reduction of the size of the problem. In this case, we need to apply a procedure to select the best alternative among all potential optimal alternatives.

With the exception of [3] and [7], previous methodologies exploit scenarios one and two by searching for dominant and dominated alternatives by solving different optimization models. However, to study the potential optimal alternatives (third scenario) requires a more detailed understanding of its relation to the truth set.

Index and Regions of Optimality

Basic decision analysis theory shows that for every joint distribution in $\mathbb{T}$, there exists an alternative $i \in \mathbf{J}$ that maximizes $E_{\mathbf{p}}[u(\mathbf{x}^i)] \forall j \in \mathbf{J}$. Hence, we want to know for which subset of $\mathbb{T}$ a potentially optimal alternative is, in fact, optimal. To this purpose, define the set $\mathbb{T}^i = \{\mathbf{p} \in \mathbb{T} | E_{\mathbf{p}}[u(\mathbf{x}^i)] \geq E_{\mathbf{p}}[u(\mathbf{x}^j)] \forall j \in \mathbf{J}\}$ as the region of optimality for alternative i .

The region of optimality provides new relevant information about alternative i . Consider the possible scenarios of alternative i . If i is dominant, then $\mathbb{T}^i = \mathbb{T}$. If i is dominated, then $\mathbb{T}^i = \emptyset$. And if i is potentially optimal, then $\mathbb{T}^i \subset \mathbb{T}$. In the first two cases, the region of optimality provides the same information as would the methodologies cited above. However, in the last scenario, our procedure provides new information about the fraction of the truth set for which an alternative is optimal.

By taking the ratio of the volume of the region of optimality to the volume of the truth set, we can define the index of optimality for alternative i as $I^i = \frac{\text{Vol}(\mathbb{T}^i)}{\text{Vol}(\mathbb{T})}$. The range of the index of optimality is $[0, 1]$, where zero indicates that the alternative is dominated, one indicates that the alternative dominates all others, and values in $(0, 1)$ provide the fraction of $\mathbb{T}$ for which alternative i is dominant.

Hyperplanes and Regions of Optimality

After creating a collection of joint distributions from $\mathbb{T}$, we can cluster all joint distributions for which alternative i is optimal. These clusters are approximations of $\mathbb{T}^i$, and under linearity conditions can be separated by equations of the form $\mathbf{a}^T \mathbf{p}^i \leq \mathbf{a}^T \mathbf{p}^j$ for some $\mathbf{a}^T$, where $\mathbf{p}^i \in \mathbb{T}^i$, $\mathbf{p}^j \in \mathbb{T}^j$, and $i \neq j$.

The vector $\mathbf{a}^T$ represents a hyperplane that separates the distributions in $\mathbb{T}^i$ from the distributions in $\mathbb{T}^j$. In general, the vector $\mathbf{a}^T$ is not intuitive or easy to assess, but it provides a guide for choosing cuts that the DM may find easier to understand. As a result, it can be used to establish questions that, if answered by the DM, can help to differentiate among potentially optimal alternatives with similar indices of optimality. Hyperplanes and regions of optimality provide an important tool in the analysis of potentially optimal alternatives.

EXAMPLE OF INDEX AND REGIONS OF OPTIMALITY

This example assumes that we need to select one alternative among 12 possible options (a_i), and each alternative will return a payoff depending on the state of the world (s_j) as depicted in Table 1. In a typical decision problem, we will assess a probability for each state of the world and choose the alternative with the highest expected value. However, a complete assessment of the joint probabilities is not always possible, and in those situations we need to rely on methods that exploit partial information.

If we assume that there is no information about the uncertainties, then any joint

distribution could be considered valid. However, a better approach is to ask not which distribution should we use to solve the problem, but which alternative is optimal under the largest number of possible discrete distributions, that is, which alternative has the largest index of optimality.

Table 2 shows that, given no information, there are no dominant alternatives ($I^i = 1$) nor dominated alternatives ($I^i = 0$). Hence, any attempt to reduce the problem or to find an optimal alternative will be unsuccessful. In this scenario, a decision needs to be based on the criterion proposed by Charnetski [3], where the best alternative is the one with the largest region of optimality (i.e., index of optimality), which corresponds to alternative a_4 .

Eliciting new information from the DM might cause some of the joint distributions to become inconsistent. This reduction of the truth set generates new regions of optimality that facilitate the selection of a single alternative. For example, assume that as in the section titled “Truth Sets with Partial Information”, we know that $p \geq 0.4$, and that $0.6 \geq q \geq 0.3$. Then, the new indices of optimality are as shown in Table 3.

The index of optimality has revealed that about 72% of the truth set is dominated by two alternatives: a_4 and a_{12} . At this point, if a decision needs to be made, we know that a_4 dominates a_{12} over a larger region of $\mathbb{T}$, which makes it a more suitable choice given what we know.

Figure 4 shows that given the new information, the regions of optimality for a_1 (black), a_2 (dark gray), and a_3 (mid-dark gray) are outside of $\mathbb{T}$ (Bold-dark-solid lines) and can be excluded from the set of

Table 1. Payoffs for Four Possible States of the World under 12 Possible Alternatives

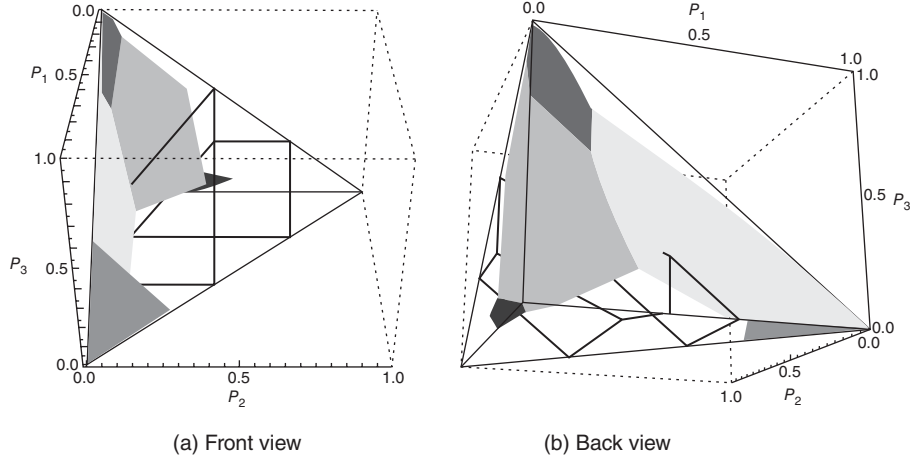
States	Alternatives											
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
s_{11}	22	50	36	49	50	48	49	49	50	30	34	43
s_{10}	48	46	22	48	25	19	50	4	20	49	44	50
s_{01}	36	47	50	49	-4	38	49	37	47	46	48	30
s_{00}	50	20	45	32	47	48	1	49	42	47	50	49

Table 2. Index of Optimality for Truth Sets with No Information

a_1	0.0021	a_4	0.3471	a_7	0.0561	a_{10}	0.0566
a_2	0.0064	a_5	0.0053	a_8	0.0238	a_{11}	0.2308
a_3	0.0068	a_6	0.0046	a_9	0.0585	a_{12}	0.2017

Table 3. Index of Optimality with Partial Information ($p \geq 0.4$, $0.6 \geq q \geq 0.3$)

a_1	0.0	a_4	0.4279	a_7	0.0901	a_{10}	0.0722
a_2	0.0	a_5	0.0049	a_8	0.0253	a_{11}	0.0682
a_3	0.0	a_6	0.0111	a_9	0.0093	a_{12}	0.2910

**Figure 4.** Truth set for alternatives a_1 (black), a_2 (dark gray), a_3 (mid-dark gray), a_9 (light gray), and a_{11} (mid-light gray) assuming no information. The black straight lines describe the truth set, given the additional information.

potentially optimal alternatives. Moreover, alternatives a_9 (light gray) and a_{11} (mid-light gray) have regions of optimality that have been dramatically reduced.

Given these results, a question of interest for the DM is: Which new information could be helpful to increase or reduce to zero the index of optimality of a_4 ? By choosing the vector $\mathbf{a}^T$ that separates $\mathbf{a}^T \mathbf{p}^i \leq \mathbf{a}^T \mathbf{p}^j$ for all $\mathbf{p}^i \in \mathbb{T}^4$ and all $\mathbf{p}^j \in \mathbb{T}^{12}$, we can find a question that increases or reduces the index of optimality of one of the alternatives. For example, Equation (8) shows the analytical description of the hyperplane that separates the collection of distributions sampled from $\mathbb{T}^4$ and $\mathbb{T}^{12}$.

$$0.3052 \cdot P_1 + 0.2032 \cdot P_2 + 0.4915 \cdot P_3 = 0.2286. \quad (8)$$

We illustrate the regions of optimality of a_4 in light gray and a_{12} in dark gray, separated by a hyperplane represented by a black speckled shade in Figure 5.

Equation 8 might seem unintuitive to the DM. However, it is possible to find approximations to $\mathbf{a}^T$ that are easy to interpret. For example, we can approximate Equation 8 by Equation 9:

$$\frac{1}{3} \cdot P_1 + \frac{1}{3} \cdot P_2 + \frac{1}{3} \cdot P_3 = 0.2234, \quad (9)$$

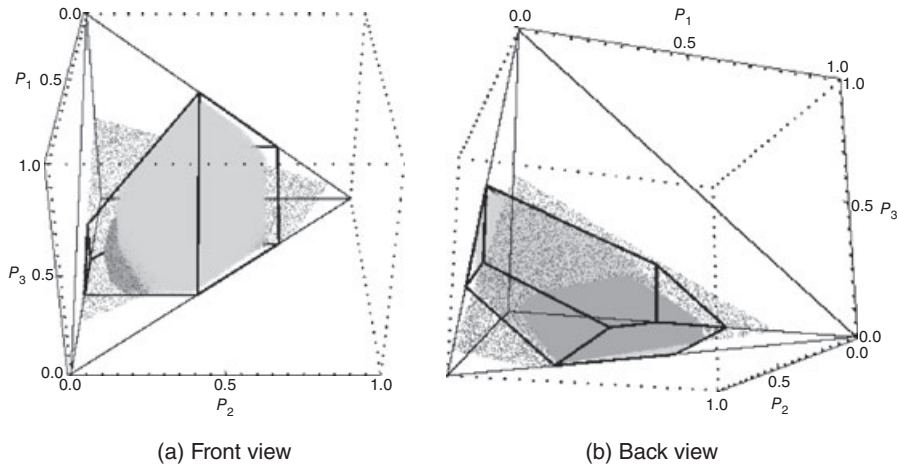


Figure 5. Truth set for alternatives a_4 (light gray) and a_{12} (dark gray), considering the additional information, and the hyperplane that separate both regions (black speckled shade). The black straight lines describe the truth set, given additional information.

which can be interpreted as $P(V_1 = 0, V_2 = 0) = 0.3299$. Hence, if we can assess whether $P(V_1 = 0, V_2 = 0) \geq 0.3299$ or $P(V_1 = 0, V_2 = 0) \leq 0.3299$, we could eliminate alternative a_4 ($I_{a_4} = 0$) or increase its index of optimality to $I_{a_4} = 0.535$. This procedure can be repeated until the DM is comfortable choosing the alternative that is optimal for most of the valid joint distributions, even if there is no dominant alternative.

CONCLUSIONS

The analysis of decisions where alternatives are neither dominant nor dominated requires a deep understanding of the interactions among regions of optimality. In this article, we explained how JDSIM can be used to explore and understand such interactions when partial information is available.

JDSIM can be applied to address incompletely specified joint distributions or to analyze the parameters of multi-linear utility functions when the DM is able to provide only partial assessments. This sampling procedure is applicable to high-dimensional complex problems and is a simple and efficient technique.

Future research needs to address the process in which elicitation information is gathered, in particular, an implementation of

JDSIM that can use sampled collections to optimize the sequence of questions to be asked will represent an important advance for practitioners.

REFERENCES

1. Montiel LV, Bickel JE. A generalized sampling approach for multi-linear utility functions given partial preference information. In review at *Decis Anal* 2014;11(3):147–170.
2. Fishburn PC. Analysis of decisions with incomplete knowledge of probabilities. *Oper Res* 1965;3(2):217–237.
3. Charnetski JR. Bayesian decision making with ordinal information. *Oper Res* 1977;25(5):889–892.
4. Hazen GB. Partial information, dominance, and potential optimality in multiattribute utility theory. *Oper Res* 1986;32(1):296–310.
5. Insua DR, French S. A framework for sensitivity analysis in discrete multi-objective decision-making. *Eur J Oper Res* 1991;54(2):176–190.
6. Salo A, Hämäläinen RP. Preference programming through approximate ratio comparisons. *Eur J Oper Res* 1995;82(3):458–475.
7. Lahdelma R, Salminen P. Smaa-2: stochastic multicriteria analysis for group decision making. *Oper Res* 2001;49(3):444–454.
8. Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: multiple criteria ranking

- using a set of additive value functions. *Eur J Oper Res* 2008;191(2):416–436.
9. Yu PL. Introduction to domination structures in multicriteria decision problems. In: Cochrane J, Zeleny M, editors. *Multiple criteria decision making*. Columbia (SC): University of South Carolina Press; 1974.
 10. Sarin RK. Interactive evaluation and bound procedure for selecting multi-attributed alternatives. In: Starr MK, Zeleny M, editors. *TIMS Studies Management Sciences*, Volume 6. Amsterdam: North-Holland Publishing Company; 1977. p 211–224.
 11. White CC III, Sage AP, Scherer WT. Decision support with partially identified parameters. Charlottesville (VA): Department of Engineering Science and Systems, University of Virginia; 1981.
 12. Moskowitz H, Preckel PV, Yang A. Decision analysis with incomplete utility and probability information. *Oper Res* 1993;41(5):864–879.
 13. Montiel LV, Bickel JE. Generating a random collection of discrete joint probability distributions subject to partial information. *Methodol Comput Appl Probab* 2013;15(4):951–967.
 14. Howard RA. The foundations of decision analysis. In: Edwards W, Miles RF Jr., von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. p 32–56.
 15. Charnetski JR, Soland RM. Multiple-attribute decision making with partial information: the comparative hypervolume criterion. *Nav Res Logist* 1978;25(2):279–288.
 16. Montiel LV, Bickel JE. Approximating joint probability distributions given partial information. *Decis Anal* 2013;10(1):26–41.
 17. Montiel LV. Approximations, simulation, and accuracy of multivariate discrete probability distributions in decision analysis [PhD thesis]. Austin (TX): The University of Texas at Austin; 2012.
 18. Ghosh S, Henderson G. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Oper Res* 2002;50(5):820–834.
 19. Montiel LV, Bickel JE. A simulation-based approach to decision making with partial information. *Decis Anal* 2012;9(4):329–347.
 20. Lowell DG. Sensitivity to relevance in decision analysis [PhD thesis]. Stanford (CA): Engineering Economic Systems, Stanford University; 1994.
 21. Smith RL. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper Res* 1984;32(6):1296–1308.
 22. Lovasz L. Hit-and-run mixes fast. *Math. Program* 1998;86(3):443–461.

DECISION PROBLEMS AND APPLICATIONS OF OPERATIONS RESEARCH AT MARINE CONTAINER TERMINALS

ANNE GOODCHILD

WENJUAN ZHAO

ERICA WYGONIK

Department of Civil and Environmental
Engineering, University of
Washington, Seattle, Washington

INTRODUCTION

The freight transportation system has played a significant role in boosting the growth of international trade and economic prosperity. Among the freight transportation modes, maritime transportation accounts for two-thirds of world trade in metric tonnes [1]. In particular, container transportation has become the predominant mode of nonbulk world cargo traffic, accounting for over 90% of the world's nonbulk deep-sea cargo [2]. Some routes, especially between economically strong and stable countries, are containerized at close to 100% [3].

Containers were first used for international transport of marine freight in the mid-1950s. Using containers has a number of advantages over conventional bulk shipping, which includes less need for product packaging; more protection of goods against weather, pilferage, and damage; and higher productivity of the entire goods movement chain [4]. They allow for faster handling due to increased automation and reduced labor requirements as their contents do not need to be unpacked at each point of transfer [4]. These benefits are realized because to some extent container dimensions have been standardized. Containers are measured by TEU (twenty-foot-equivalent unit), which refers to the volume held by one container with a length of 20 ft. The container with a length of 40 ft or 45 ft represents two TEUs. Domestic containers in the United States are wider

and longer, measuring 53 ft in length. Owing to these differences in container size and relative pricing, many goods imported into the United States are still transferred from marine containers to the larger US domestic containers at US ports.

Container terminals are important intermodal interfaces between maritime and inland transportation. Their main function is to facilitate the transfer of containers between marine vessels and inland modes such as rail and truck transportation. In addition, some container terminals serve ship-to-ship transfers or transshipments. Container terminals are large, interdependent, and complex systems, with many processes and activities involved in container movements. Figure 1 illustrates a container terminal from Poland and shows the ship berths, waterside container cranes, (called *quay cranes* [QC]), and container storage areas. The smaller orange cranes within the container stacks (called *yard cranes*) are used to retrieve and store containers and serve inland transportation vehicles.

Growing global trade and container traffic has put pressure on container terminal operations and inland freight distribution [5]. Consequently, older container terminals are facing increasing congestion problems since they are generally located in urban areas with little available land for physical expansion or inland transport network improvement. Meanwhile, newly constructed container terminals increase the competition between ports. In addition, container terminals have been under pressure to reduce the air quality impacts of their operations.

These issues have attracted the attention of the academic community, which has used operations research techniques, simulation models, and analytical models to improve terminal operations. Terminals rely on many operational models and analyses to support their improvements and allow for increased cargo volumes. This article provides a brief introduction to container terminal operations, but emphasizes the decision



Figure 1. Baltic container terminal, Gdynia, Poland (http://portaid.com/html/bct_photo.html).

problems and applications of operations research in the marine terminal setting. A few formulations and solution algorithms are also presented. This introductory article does not examine the complexity of these formulations or their objectives.

AN OVERVIEW OF CONTAINER TERMINAL OPERATIONS

Container terminals can be viewed as open systems of container flow with two external interfaces [4]. The two interfaces are the landside, where containers arrive and depart by trucks or trains, and the quayside, where containers are unloaded from and loaded onto

ships. Containers are temporarily stored in the container yard facilitating the decoupling of quayside and landside operations [4]. Figure 2 is a schematic of a container terminal illustrating its system components.

Import containers flow through the terminal in several steps. When a ship arrives at the terminal, it is allocated to a berth, and quay cranes are assigned to move the import containers off the ship. Next, the unloaded containers are transported from quay cranes to the storage yard. For export containers upon arrival, the terminal transportation vehicle drops off the container directly, or a piece of yard handling equipment facilitates taking the container off the vehicle, and stores it on the yard. The import containers

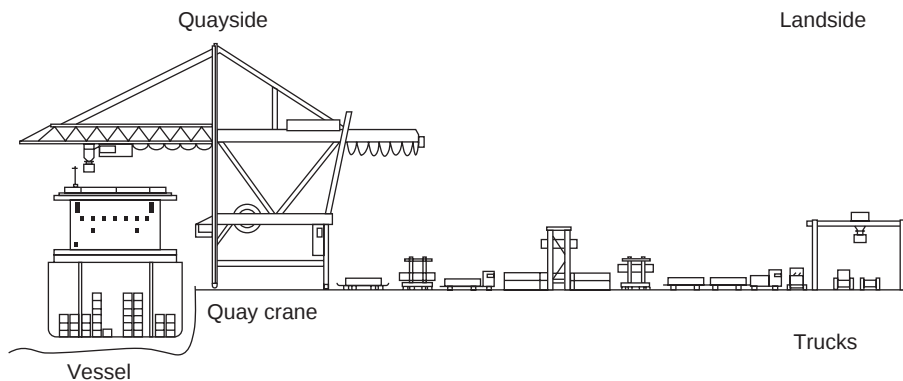


Figure 2. A schematic of container terminal [4].

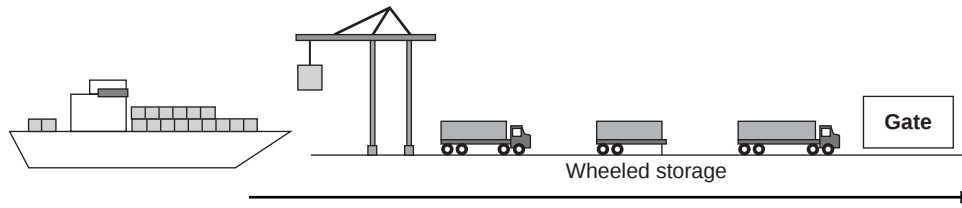


Figure 3. Wheeled operation at container terminals [7].

may be stored for some length of time and then are retrieved from the yard and transferred to other modes such as trucks, trains, or barges to depart from the terminal [6]. The export containers flow in a reverse direction—they arrive by other modes, such as trucks or trains; are stored in the yard; and depart by container ships. At some terminals, storage time in the terminal has been reduced by improved coordination in marine vessel and truck or rail scheduling; however, in most terminals containers are still stored temporarily.

Containers are either stored on a chassis or on the ground at container yards. In chassis or wheeled storage, the container is stored with the chassis as a unit (Fig. 3).

This method of storage allows for easy and fast transfer of containers and reduces labor requirements, but it requires a larger terminal area to store the same number of containers when compared to grounded storage. Wheeled storage is used in US terminals with large terminal areas and at relatively high labor costs. With grounded storage, containers are stored directly on the ground, without a chassis, and may be stacked on top of each other. Containers are moved in and out of stacks by yard handling equipment (Fig. 4). Ground stacking reduces the terminal area requirements, but requires more handling effort and increases transaction time for a truck to retrieve a container. This storage method is popular at European



Figure 4. Ground stacking of containers at a container yard (<http://ports.co.za/admin/large/image-303.jpg>).



Figure 5. Super-PostPanamax cranes in Port of Rotterdam (http://upload.wikimedia.org/wikipedia/commons/0/0b/Portainer_%28gantry_crane%29.jpg).

and Japanese ports, the Port of Hong Kong, and US terminals where land is scarce. Grounded storage methods increase the complexity of terminal operations because containers are not independent—accessing one container may require moving others.

Terminal Equipment

The material handling equipment at container terminals can be classified into two types: cranes and transporters.

The two types of cranes used are quay cranes (Fig. 5) that load and unload ships, and yard cranes (Fig. 4) that stack and handle containers on the yard. The two main components of a crane are a supporting framework that can traverse the length of a quay or yard, and a moving platform called a *spreader*. The spreader can be lowered down to access a container and it locks on to the container's four locking points. Cranes normally lift a single container at a time. However, some cranes are able to pick up four or more TEUs.

There are two types of quay cranes namely, single-trolley and dual-trolley

cranes. The trolley travels along the arm of a crane to lift the containers with its spreader. Dual-trolley cranes are expected to produce twice as many moves per hour as conventional cranes, though neither achieves theoretical capacity in practice. Steenken *et al.* [4] noted “the technical performance of a quay crane is in the range of 50–60 moves/h, while in operation is in the range of 22–30 moves/h.” In Singapore, robotics has allowed a single operator to operate many cranes, but this is not yet common.

Three types of yard cranes are used in container terminals: rubber-tired gantry cranes (RTGs, Fig. 4); rail-mounted gantry cranes (RMGs); and overhead bridge cranes (OBC). An RTG is a mobile gantry crane, which runs on rubber-tires and capable of straddling at least 6 rows and stacking 4–7 containers high [8]. An RMG is less flexible and runs on fixed tracks. The largest RMG is able to straddle 12 to 20 rows and can stack up to eight high [8]. An OBC is a type of crane with the trolley traveling along a horizontal beam that is mounted on two parallel and elevated



Figure 6. Straddle carrier (http://en.wikipedia.org/wiki/File:Containerlift_straddle_carrier.jpg).

runways. The operation of the OBC is limited to the area between the runways [7]. The technical performance of gantry cranes is approximately 20 moves/h.

Transporters are used for moving containers within the storage yard and between the storage yard and the marine vessel. The two types of transporters used are passive transporters that must be loaded or unloaded by other equipment and transporters that are capable of lifting containers by themselves. Passive transporter vehicle types include yard trailers, multiload yard trailers, and automated guided vehicles (AGV). An AGV is driven by an automatic control system instead of a driver and follows guided paths electronically stored in the memory of the control system. Transporters with lifting abilities include straddle carriers (SC, Fig. 6); reach stackers; and forklifts. SCs have been widely used and are able to stack up to four containers. Automated SCs have also been developed, and are successfully being used at some terminals such as the Patrick container terminal in Brisbane.

While a considerable literature exists regarding performance measurement and

comparing marine terminal performance [9], the remainder of this article focuses on decision problems and applications.

CONTAINER TERMINAL DECISION PROBLEMS AND THE APPLICATION OF OPERATIONS RESEARCH MODELS

Decision problems in container terminals can be classified into three types: design problems, operational planning problems, and real-time control problems [10]. Design decisions are made in the initial planning stages with regard to terminal configuration including determining the number and type of material handling equipment, the terminal layout (or placement of equipment), and the degree of automation of handling equipment. An important related issue is the estimation of various performance measures of the intended terminal configuration to compare and evaluate different terminal designs [9]. Operational planning is performed after the terminal has been designed but before operation actually happens, and it is performed to allocate resources for handling activities to maximize the efficiency of terminal operations. The key resources include berths, quay cranes, yard cranes, and the terminal area. Real-time control problems match handling tasks with the required resources and develop detailed operational schedules and plans for day-to-day operating decisions.

The following sections describe important decisions that are supported by operations research techniques, and must be made to address the design, operational, and real-time control problems faced by marine terminals as they grow and improve. We build upon four previous articles that have summarized or described operations research work in marine terminals, namely, Vis and de Koster [6], Steenken *et al.* [4], Kim [11], and Stahlbock and Vos [12].

Quayside Decision Problems

This section addresses the three types of decision problems as they relate to quayside, or waterside operations of a container terminal. In terms of design problems, this

section considers the type of quay cranes and transporters to be used in a terminal. Then, the section examines the use of operations research models to address relevant operational planning problems, including the berth allocation, quay crane allocation, discharge and load sequencing of containers, and determination of number of transporters needed. Finally, transporter allocation, scheduling, and routing problems, which are optimized in real time during terminal operations, are discussed.

Determination of Quayside Equipment Types. Terminal design is constrained by the amount of available shoreline with good marine and land access, and is dependent on the characteristics of the port, the degree of automation desired, traffic volume, relevant regulations, and the costs and benefits of the design. Choosing specific types of equipment depends on the technical characteristics of the equipment. For example, automated handling equipment requires a large initial investment but has lower operating costs, while manned equipment is less expensive initially but requires significant labor, contributing to higher operating costs. This choice should be made by performing a feasibility and economic analysis on various types of equipment [13]. The equipment choice also depends on the design of the area in which the equipment operates. Vis and Harika [13] used simulation to examine the effect of using automated lifting vehicles (ALV) and AGV on the unloading times of a ship. Their work only considered the container unloading process. They found using AGVs or ALVs does not impact the unloading time of a ship. However, 38% more AGVs are required than if ALVs are used to minimize the ship unloading time. They concluded, by observing only purchasing costs of equipment, ALVs are the less expensive option than AGVs since fewer are required. Their sensitivity analysis indicated the layout of the terminal and the technical aspects of equipment (quay crane and vehicle) impact the number of vehicles required and thus the choice of equipment.

Berth Allocation, Quay Crane Allocation, and Scheduling. Berth allocation is the

assignment of quay space and service time to vessels that have to be unloaded and loaded at a terminal. After the berth plan is generated, quay cranes need to be allocated to ships and the ship bays to perform all required transshipments. This process is referred to as *quay crane allocation*. The third step, called *quay crane scheduling*, is to determine the sequence and time schedule of discharging and loading operations that the assigned quay cranes will perform for a ship. The three decisions can be made in a sequential fashion and they are interrelated. For example, the number of quay cranes assigned to a vessel affects the berth duration of the vessel. However, because of the complexity involved in ship operations planning, most academic researchers decompose the problem into these three steps. An overview of berth allocation and quay crane scheduling problems is given in Bierwirth and Meisel [14].

The berth allocation problem (BAP) assigns a berthing position and a berthing time to each vessel, to optimize a given objective function [14]. A graphical representation of a berth schedule for five vessels is shown in Fig. 7a. The vertical axis represents positions along the quay, while the horizontal axis represents time schedule. The BAP solution can be represented by rectangles placed between two axes. These rectangles must not overlap with each other for the berth plan to be feasible. Several constraints are involved in berth allocation. Spatial constraints restrict the feasible berthing positions of vessels according to the layout of the quay into berths. Some studies consider a discrete layout, where the quay is partitioned into a number of sections, called *berths*. Only one vessel can be served at each single berth at a time. Other studies use a continuous layout, in which the quay is not partitioned and vessels can berth at any point within the boundaries of the quay [14]. Berth planning is more complicated for a continuous layout than for a discrete layout, but the quay space can be better utilized. Other constraints should also be considered. For example, the vessels must be berthed at positions with sufficient water depth. The general goal of berth planning is to provide fast and reliable services of vessels.

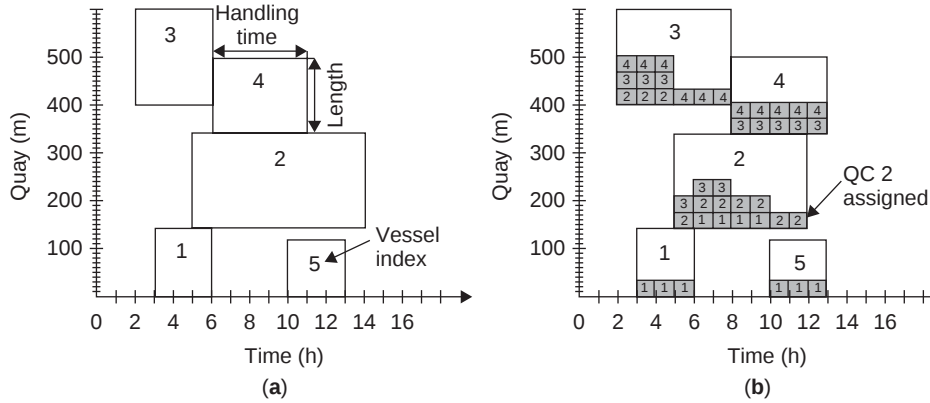


Figure 7. Space-Time representation of a berth plan (a), assignment of cranes to vessels (b) [14].

Common objectives include minimizing the sum of the waiting and handling time of vessels, minimizing the maximum amount of quay space used at any time, minimizing the workload of terminal resources, and minimizing the number of vessels rejected for service at a terminal.

The dynamic berth assignment problem in the public berth system was studied in Imai *et al.* [15]. This dynamic problem took into account ships that have already arrived and those ships that have not arrived at the time of planning but will arrive at some time later during the planning horizon. The problem was formulated as a mixed integer programming model. The model objective is to minimize the sum of waiting time for berth availability in addition to the handling time spent at the berth, assuming the ship handling time depends on the berth it is assigned and the berths are discrete. A heuristic procedure employing the subgradient optimization procedure was developed based on the Lagrangian relaxation of the original problem. Computational experiments showed that the proposed algorithm is adaptable to real-world problems. The dynamic berth allocation problem with a continuous quay space was also explored in Imai *et al.* [16], and solved using a heuristic algorithm extended from the method developed for discrete BAP [15]. The continuous BAP was addressed by Kim and Moon [17] using a different solution approach. The objective

function of their research problem is to minimize the costs from the delayed departures of vessels and the additional handling costs from deviations of the berthing position from the best location. They suggested a simulated annealing algorithm based on the properties of the optimal solution that is able to find near-optimal solutions.

Quay crane allocation has three specific requirements: a set of specific cranes must be assigned to a vessel, the number of assigned cranes is unchangeable throughout the service process, and a minimum number of cranes serving the vessel throughout the whole process can be prescribed [14]. Figure 7b shows an example of allocating cranes to vessels for fulfilling all the required container transfers. In practice, quay crane allocation is not a difficult problem and has received little attention on its own in academic research. Owing to its impact on ship handling time; however, crane allocation has been studied as a problem integrated with berth allocation. Imai *et al.* [18] addressed the simultaneous berth and crane allocation problem at a multiuser container terminal. This work assumed that berth space is discrete, ship handling does not begin until the required number of cranes is available, and a crane cannot move from one berth to another via other berths if they are engaged in ship handling. A mathematical formulation for this integrated problem was developed to minimize the total ship service time. A genetic algorithm-based

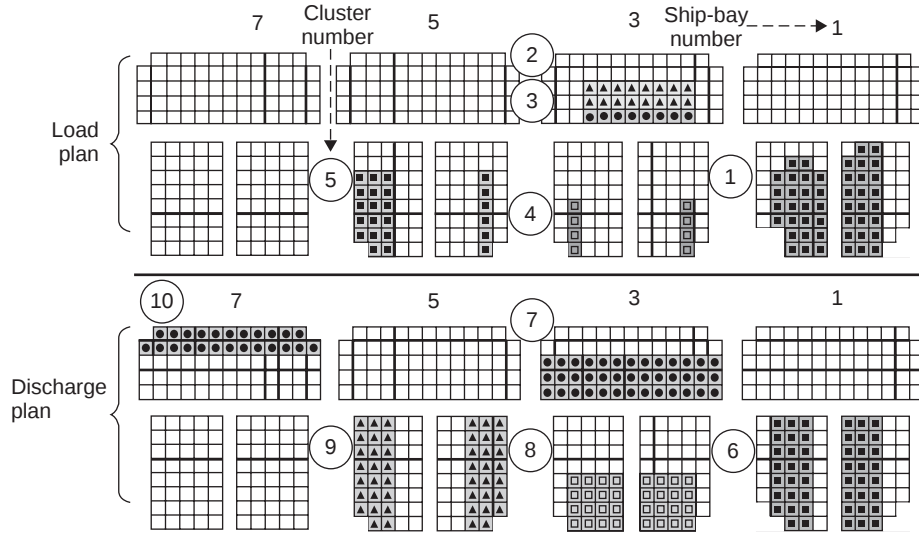


Figure 8. A partial example of ship stowage plan [19].

(GA) heuristic was developed to find an approximate solution for this problem. The two decision making procedures, BAP and quay crane allocation problem (CAP), are constructed one-by-one in an iteration of the entire GA solution procedure. In the BAP phase, the ship-to-berth allocation is determined; given that allocation, in the CAP phase the feasibility of ship service schedule is checked, and the cranes are scheduled by being assigned to ship handling tasks in sequence across all berths. The experimental results showed that the proposed heuristic is able to solve this terminal operation problem sufficiently.

The scheduling goal of the quay crane (QC) is to determine the sequence of discharging and loading operations a QC will perform to minimize the completion time. The two constraints important in this problem are that the container discharging operation must precede the loading operation if performed at the same ship-bay, and the quay cranes cannot cross over each other since they are on the same track. This problem is similar to a minimum makespan scheduling problem with parallel identical machines and precedence constraints, but involves more constraints. The ship stowage plan is required as the input information, and tasks to be

scheduled on a QC can be defined based on bay areas, single bays, container groups, or individual containers. A task defined based on bay areas consists of all the container loading and unloading operations within a certain area of the bays. A task defined based on single bays consists of all the operations in a bay. A task defined based on groups refers to a group of containers stored in adjacent slots of a bay with the same attribute, for example, those containers with the same destination. Lastly, the task defined based on individual containers refers to the operation of a single container [14]. An example of a ship stowage plan used to construct the crane schedule is provided in Fig. 8, which consists of four cross-sectional views, each corresponding to a ship-bay and labeled with odd numbers from 1 to 7. This figure shows four groups of containers to be discharged from the ship bays, and four groups of containers to be loaded into the ship. In Kim and Park [19] the tasks are defined based on container groups (Fig. 8); and several constraints are considered in addition to those mentioned above: tasks on a deck must be performed before tasks in the hold of the same ship-bay when a discharging operation is performed in a ship-bay; the loading operation in a hold must precede the loading operation

on the deck of the same ship-bay; and certain pairs of tasks cannot be performed simultaneously when their locations are too close to each other to avoid crane interference. They formulated the quay crane scheduling problem as a mixed integer programming model and proposed a branch and bound method to obtain the optimal solution, and also suggested a greedy randomized adaptive search procedure to overcome the computational difficulty of the branch and bound method. Quay crane scheduling problem was also addressed by Lee *et al.* [20] using GA. Different from Kim and Park [19], they defined the tasks based on the ship bays.

There has been research to consider the relaxation of the assumption that the discharging operations must occur prior to loading operations. This research identifies a sequence of containers to be handled assuming that loading and discharging operations occur simultaneously. This operation is referred to as *double cycling*. Goodchild and Daganzo [21] formulated this problem as a job scheduling problem with precedence constraints and compared the performance of a greedy algorithm to the optimal solution and found that the results were insensitive to the operating strategy. In another work Goodchild and Daganzo [22] used queueing analysis to examine the implications of double cycling for other aspects of terminal operations, for example, the impact on the need for yard cranes and storage areas.

Operational Planning and Real-Time Decisions Related to Transportation Equipment. One of the operational planning problems related to transportation equipment is in determining the number of transporters required to support ship loading and unloading operations. A sufficient fleet of transporters are needed to avoid quay crane waiting during the ship unloading and loading process, but too many transporters will cause yard congestion and increase cost. This transportation fleet sizing problem at container terminals has been explored by Kang *et al.* [23]. The container unloading process was examined and decomposed into the following four stages:

1. movement of the empty yard truck from yard to berth;
2. truck loading process at the berth;
3. movement of a loaded truck from the berth to yard; and
4. truck unloading process at the yard.

By considering yard trucks as the customers of the queue system, a cyclic queue model was developed to evaluate the steady-state performance of the terminal fleet, and such performance is employed to determine the optimal fleet size which minimizes the terminal cost (operating cost of all handling equipment) and satisfies the terminal throughput requirements. This cyclic queue model can yield the optimum fleet size for long-term operation. To facilitate terminal operators managing fleet in real time, they also proposed a Markovian decision process (MDP) model by considering the stochastic nature of unloading operations. This MDP model could provide dynamic operational policies for fleet management.

The real-time control problem allocates containers to vehicles and routes and is referred to as the *yard vehicle dispatching problem* and *routing problem*. Figure 9 shows an example of dispatching and routing vehicles to transport containers between berths and storage areas. This example includes three quay cranes to be served and nine container stack points in the yard. Two tours are generated to connect cranes and stack points. For vehicle types that are unable to lift containers by themselves, direct transfer of containers to and from quay cranes is required, and there is a delay if a queue forms in front of the quay crane, or when no vehicle is available and the quay crane must wait to release a container. The objective of such problems is to minimize the delay time of quay cranes and the travel time of vehicles. Yard vehicle dispatching and routing are interrelated and have been studied as an integrated problem by some researchers. Nishimura *et al.* [24] addressed the yard assignment and routing problem for both the single-trailer and multitrailer trucks. To improve the productivity of the terminal, a dynamic assignment method was applied,

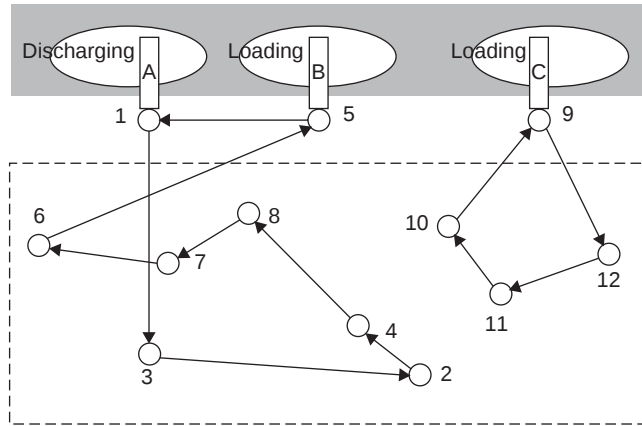


Figure 9. An example of vehicle dispatching and routing decision [24].

which pools trailers to serve a set of quay cranes instead of assigning each trailer to a specific quay crane. Mathematical formulations were developed to minimize the total travel distance of trucks. The single-trailer problem resembles the single-depot vehicle routing problem with multiple tours from the depot. That problem was reduced to the classical assignment problem whose optimal solution can be readily found. However, the multitrailer version of the problem does not have an efficient exact solution procedure, and the GA was used to find a near-optimal solution. Their research demonstrated that dynamic assignment is superior to the static assignment with regard to the total distance traveled by trailers. It also reduces the fleet size, with 15% saving in fleet cost.

Yard dispatching and routing have also been studied as two independent problems. Kim and Bae [25] examined how to dispatch AGVs by utilizing information about locations and time of future delivery tasks. It was assumed that any AGV is not dedicated to one quay crane. The vehicle scheduling issue was not discussed by ignoring the queueing time of AGVs under yard cranes and the congestion among AGVs on guide paths. The problem was formulated as a mixed integer programming model for assigning optimal delivery tasks to AGVs. Its objective function is to minimize delays in operation of quay cranes and travel time of AGVs, with a higher priority given to quay crane delay. This problem is similar

to a parallel machine scheduling problem with sequence-dependent setup times and precedence constraints among jobs. Since it is nondeterministic polynomial-time (NP-hard), a heuristic algorithm was developed to solve this problem. Their simulation results demonstrated the proposed heuristic algorithm outperforms conventional dispatching rules such as assigning based on the shortest travel time, the shortest distance, or the earliest due date.

Yard Decision Problems

Import and export containers can be temporarily stored in the yard before being loaded onto ships or inland transportation vehicles (Fig. 4). In general, containers are stacked on the ground in several tiers and the entire storage space is divided into rectangular areas called *blocks*. Each block consists of many parallel bays, each bay is composed of several stacks, and each stack stores several containers. An example of a block-and-bay configuration is provided in Fig. 10, with a yard crane employed to perform container delivery and retrieval operations. The maximum number of tiers is limited by the lifting ability of stacking equipment, either straddle carriers or yard cranes, and the structural integrity of the containers themselves. The storage of different types of containers is generally segregated. For example, import containers, export containers, refrigerated containers, and empty containers, each generally have

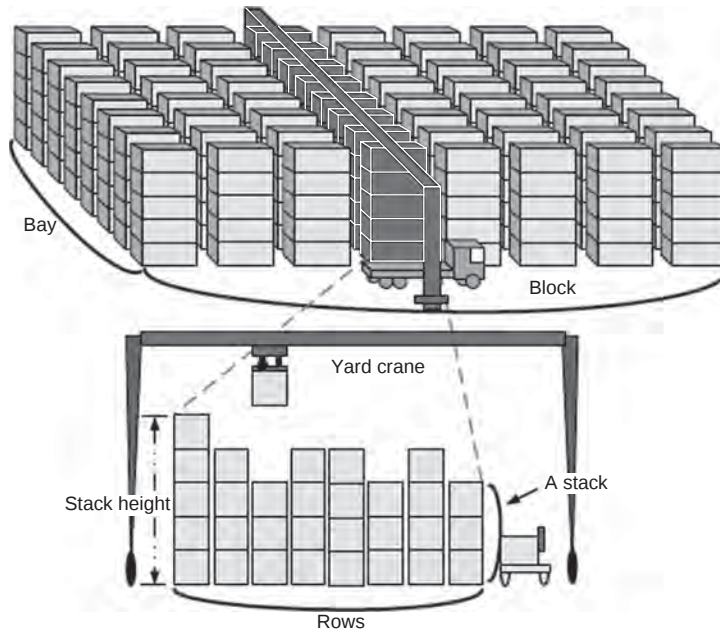


Figure 10. Container block, bay configuration, and yard crane positioning [26,27].

a designated area. The transfer points of containers are located at the two ends of each container block for a perpendicular layout of blocks (popular at European ports), or a separate lane is dedicated for container transfer beside the block for a parallel layout of blocks (popular at US ports). Since containers are stacked, any container above the desired container must be moved. This process is called *container rehandling*. It is unproductive work and occupies a large amount of total handling time. The number of container rehandles is one metric to measure the operational efficiency of yard operation.

Yard stacking equipment handles containers from both the quayside and landside. They perform two types of operations: picking up import containers delivered by transporters from quayside or export containers delivered by drayage trucks from landside and stacking them into blocks; and retrieving export containers from storage to load onto transporters (for ship loading) or retrieving import containers for loading onto drayage trucks.

Yard design problems include determining yard layout, container block configuration, and the type and number of required

handling equipment, such as yard cranes and straddle carriers. An important operational planning problem concerning the container yard includes storage space allocation for arriving containers. Real-time decisions include assigning tasks to yard handling equipment. This section describes those problems in detail and discusses the application of operations research techniques in addressing them.

Yard Design Problem. The organization of the container yard is crucial for achieving an efficient interface with other processes in the terminal. The type and number of storage and retrieval equipment used in the yard are critical design decisions to ensure yard operations are not the bottleneck of the terminal system. Straddle carriers and yard cranes are the two most popular types of equipment used for storing and retrieving containers. One of the main differences between straddle carriers and yard cranes is span width. A straddle carrier can travel over only a single row of containers, whereas yard cranes can span six or more rows of containers. As a result, straddle carriers need to travel to the end of a row to move to another row while yard

cranes can change rows at any location in the block. The selection between manned straddle carriers and automated stacking cranes (ASC) was discussed in Vis [28]. She performed a simulation study to compare the performance of straddle carriers and ASCs for handling a fixed number of storage and retrieval requests from both the seaside and landside of the terminal. The yard configuration with transfer points located at two block ends was considered, and the total travel time (which includes full and empty travel times, hoisting times, and rehandling times) was used as the performance measure. A sensitivity analysis was also performed to examine the effect of operational aspects (size of workload and arrival patterns of containers) and block layout on the performance of container storage and retrieval equipment. She concluded that if blocks are configured with a width of six containers, the automated stacking crane outperforms the straddle carrier. However, if the block width is increased beyond six to up to nine containers, which will probably reach the maximum span for an ASC, the performance of the SC matches the performance of the ASC.

Kim and Kim [29] discussed how to determine the optimal amount of storage space and the optimal number of yard cranes for handling import containers. The trade-off between the yard space and number of yard cranes was analyzed by considering the associated cost. A terminal operator's cost model was developed to minimize the operational cost of the terminal, yard space cost, and capital cost of yard cranes without considering the level of service for the customers. To evaluate integrated system success, another cost model was developed to minimize the waiting cost of outside trucks as well as the total cost for terminal operation. Solution procedures were suggested to find optimal solutions.

Storage Space Allocation. The method of allocating storage spaces for arriving containers has a significant impact on the turnaround time of ships and drayage trucks because the locations of containers affect the efficiency of delivery and retrieval operations [11]. Common objectives of the space allocation problem are to maximize

storage utilization, minimize rehandles, minimize travel distance, and maximize yard equipment utilization.

The storage space allocation problem was studied in Zhang *et al.* [30] for a complex situation in which inbound, outbound, and transit containers are mixed in the yard. The problem was decomposed into two levels with an overall objective of minimizing the vessel berthing time and maximizing the quay crane throughput rate, and solved using a rolling-horizon approach. The first level determines the total number of containers to be placed in each block to balance the workload of yard cranes and quay cranes. The second level determines the number of containers associated with each vessel to be allocated to each block to minimize the total traveling distance for moving containers between their storage area and ships.

Kim and Kim [31,32] address the storage space allocation problem for import containers. This problem utilized a segregation strategy for allocating containers to empty slots by considering the container's duration of stay at the bottom of stack, and its objective is to minimize the expected total number of rehandles. Different cases were analyzed where the arrival rate of import containers is constant, cyclic, and dynamic. Mathematical models for each case were formulated using stack height and allocation space as decision variables. Problems were solved to optimality based on the Lagrangian relaxation technique.

After a container bay or block with empty slots is already assigned to a specific container group, the next stage of storage space allocation is to determine where to stack an arriving container among a group of available slots. This problem considers the configuration of the container bay to minimize the expected number of container rehandles that occur during the loading operations of a containership or a drayage truck. Kim *et al.* [33] examined the problem of determining the storage location for arriving export containers. They assumed the distribution of container weights is known, and the heavier containers are loaded onto a ship before lighter containers, so rehandles will be incurred if lighter containers are stacked on top of heavier containers at the

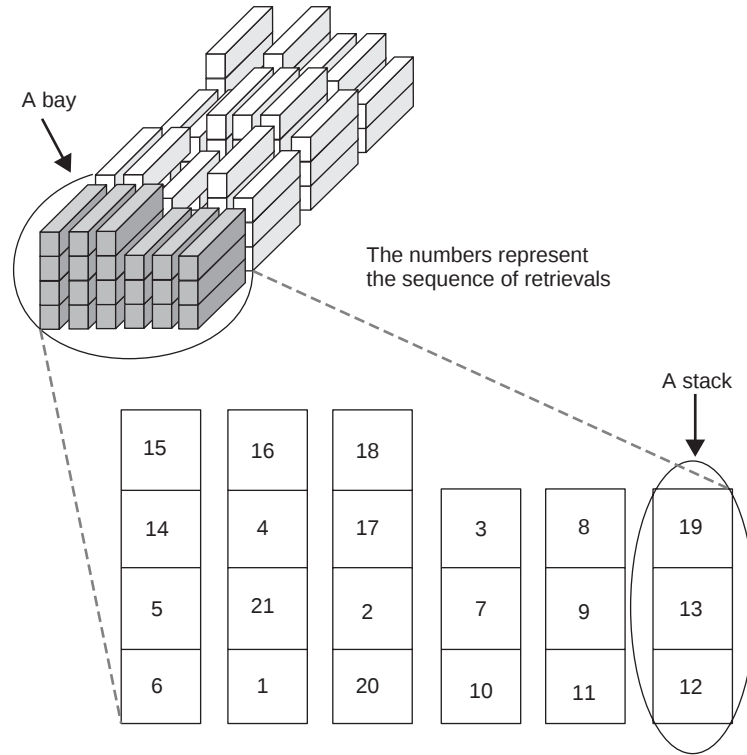


Figure 11. An illustration of a yard-bay of blocks with a fixed sequence of pickup [34].

container yard. A dynamic programming model was formulated to determine the optimal storage location, and a decision tree was developed from the set of optimal solutions to support real-time decisions. Kim and Hong [34] studied how to minimize the total number of container rehandles during the pickup operation on the yard by carefully determining the storage location of rehandled containers. They assumed the initial configuration of the container bay is known and the pickup priorities among containers in the bay are also given (Fig. 11). They suggested a branch and bound algorithm to obtain the optimal solution, and then proposed a heuristic rule that estimates the number of future rehandles for the bay to shorten the computational time and find a near-optimal solution.

Task Assignment for Yard Handling Equipment. Two hierarchical problems exist in task assignment decisions: the equipment

deployment problem and the task scheduling problem. Equipment deployment involves the allocation of a certain group of equipment to a specific type of task. Task scheduling includes the assignment of tasks to each piece of equipment and the sequencing of assigned tasks to be carried out [11].

The objective of yard equipment deployment problem is to balance the workload among blocks and has been examined in Cheung *et al.* [35] and Zhang *et al.* [36]. Since block workload varies across different storage areas and time, the handling equipment should be efficiently used to ensure all the required tasks are finished. Zhang *et al.* [36] addressed the dynamic crane deployment problem to find the time and routes of crane movements among blocks so that the total delayed workload in the yard is minimized. The problem was formulated as a mixed integer programming model and solved by Lagrangean relaxation.

Task scheduling for yard handling equipment is performed in real time. It must consider the priorities among tasks to sequence tasks, and also the interference between yard cranes or straddle carriers to estimate the task completion time [11]. The objective of the task scheduling problem is to minimize the completion time of all the tasks (the makespan). An alternate objective is to minimize the sum of transporter waiting time on the yard. The single straddle carrier routing problem during the loading operation of export containers was examined by Kim and Kim [31,32]. The objective is to minimize the total traveling distance of a straddle carrier, and the main constraint considered is that the loading schedule must satisfy the requirements of the work schedule of the quay crane. This routing problem is similar to the network optimization problem, and a two-stage solution procedure was developed. The first stage determines the number of containers to be picked up at each yard-bay during a subtour, and the second stage determines the visiting sequence of yard-bays by the straddle carrier.

The problem of scheduling multiple yard cranes in a yard zone was examined in Ng [37]. It was assumed that the expected arrival time of yard trailers is given, and the objective is to minimize the sum of yard trailer waiting times in the zone. The inter-crane interference constraint was considered, and the problem was formulated as an integer programming model. Since the problem is NP-complete, a two-phase scheduling heuristic was developed to solve the problem. In the first phase, the multiple-crane scheduling problem is simplified and decomposed into m independent single crane scheduling problems; in the second phase, a job reassignment procedure is used to improve the schedule obtained from the first phase. Computational experiments showed that the heuristic can find effective solutions, which is on average 7.3% above their lower bounds.

Landside Decision Problems

In landside operations at container terminals, containers depart and arrive by trains or drayage trucks. For rail operations, trains

have fixed tracks and transportation equipment is needed to move containers from the storage yard to rail facilities. For drayage truck operations, when trucks arrive at the terminal gate, trucks are directed to container transfer points within the terminal, where the container storage and retrieval equipment unloads a container from truck, or transfers a container from the block to the truck. Large container terminals serve thousands of trucks a day. Owing to the dynamic nature of truck arrivals and high truck volume, truck congestion at terminal gates and in container yards is a problem faced by many container terminals around the world.

This section focuses on the road transport mode (drayage trucking) and describes the design of the terminal's landside connection, the operational planning problem of landside transport, and real-time assignment and scheduling of tasks to serve outside trucks.

Design of the Terminal's Landside Connection. The terminal's landside connection is an important interface between the container terminal operations and on-dock rail and drayage truck operations. One of the crucial decisions related to landside design is whether or not to decouple landside operation from the storage yard operation and allocate a separate area dedicated for container transfers to and from trains and drayage trucks, or to allow these vehicles to physically enter the terminal.

In general, on-dock rail operations require a dedicated zone with a container buffer area alongside the tracks. Container handling equipment such as gantry cranes are deployed to unload and load containers onto trains and terminal transportation equipment such as yard trucks and straddle carriers are needed to transport containers between the storage yard and train loading area.

The majority of container terminals around the world do not setup a separate area for truck operation. Instead drayage trucks drive directly into the storage yard for container transfers. However, some automated container terminals have also been redesigning their terminal layout to

improve the terminal capacity, including setting up a dedicated truck operation area.

One example is Patrick Corporation's container terminal at Port Botany in Sydney, which has changed its terminal layout and designed a semiautomated container exchange facility for container transfers to and from trains and drayage trucks. This change is intended to reduce space requirements and increase the container exchange capacity [38]. A simplified layout of Port Botany is provided in Fig. 12. This exchange facility is equipped with rubber mounted gantry cranes, which perform the container loading and unloading operation for trucks and trains. Straddle carriers transport containers between the exchange area and main storage yard. A detailed layout of this exchange facility is shown in Fig. 13 and includes a Gantry-Road interface (GRI), where the drayage trucks arrive and are served by gantry cranes; a Gantry-Rail interface (GRAI), where the trains are loaded and unloaded; an intermediate stacking area (ISA), where the arrival containers are temporarily stacked before being moved to the storage yard, and the booked import containers are placed for quick transfer by cranes onto trucks or trains; and a Gantry-Straddle interface (GSI), where straddle carriers place import containers or pick up export containers [38].

Landside Transport Planning. After the landside layout is determined, a plan must be made for how to serve drayage trucks. Another relevant issue is the management of empty containers. Empty containers may be returned to the terminal after being emptied at the customers' locations, and are moved from the terminal to export customers' locations for loading when demanded. This practice generates unproductive empty truck trips, and exacerbates terminal congestion. By repositioning empty container depots to inland areas, these empty truck trips may be reduced. Although the empty container repositioning problem extends beyond the physical boundary of container terminal, it is included in this section since it closely relates to landside operations.

Drayage Truck Service Problem. The main objective of this problem is to improve the level of service for drayage trucks and the landside throughput. Different strategies can be used to deal with this problem. Terminals can control the number of trucks entering the container terminal to avoid yard congestion, prioritize certain types of trucks to, sequence trucks for container delivery and receiving operations to minimize the truck delay, or carefully determine the storage location of containers to minimize total number of rehandles and reduce truck transaction time. The last two methods fall into the category of real time control decisions and is discussed in the section titled "Real-Time Control of Landside Operation."

Huynh and Walton [39] discussed how to regulate truck arrivals at container terminals. They determined the maximum number of trucks a terminal can accept into a yard zone within each time window, while satisfying the resource and operational constraints and maintaining a specified target truck transaction time. Their methodology is a combination of mathematical formulation and computer simulation, which identifies a solution beneficial for both the terminal operator and truckers.

Prioritizing the service for containers with a higher priority was discussed in Holguín and Walton [40]. They considered a group of priority systems—locating high-priority containers on special hatches, storing them on chassis, or using automatic equipment identification devices at gates—and assessed the impacts on different users based on computer simulation. They concluded the implementation of priority service significantly improves the performance of high-priority containers without overly penalizing the level of service for low-priority containers or significantly affecting the terminal's operating costs.

Empty Container and Chassis Repositioning Problem. The problems of managing empty containers and container chassis have also been addressed with operations research techniques. Crainic *et al.* [41] developed a framework for minimizing the cost of empty container allocation and distribution using

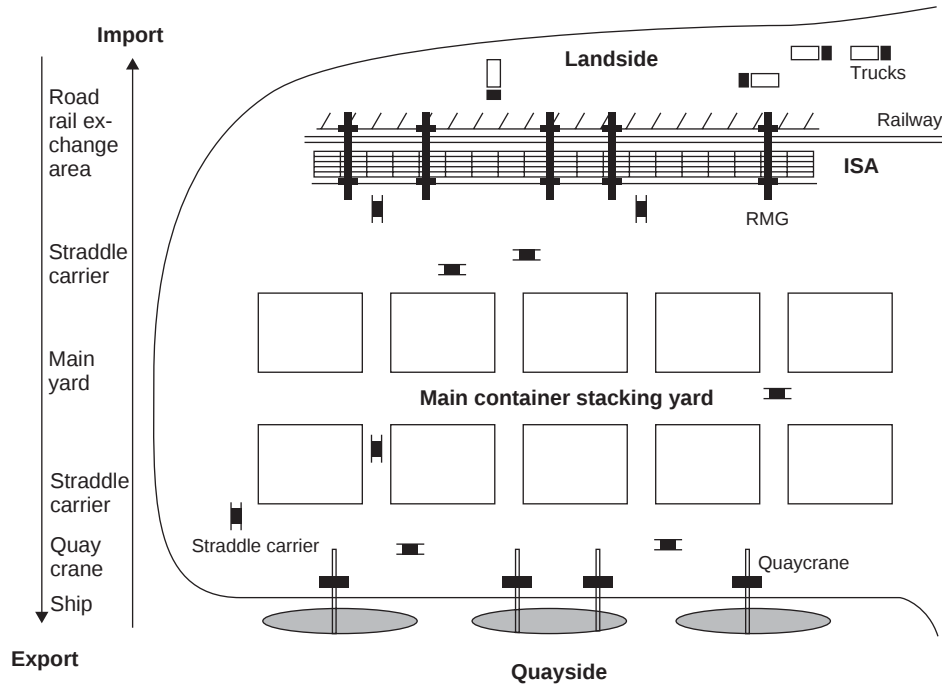


Figure 12. The layout of Port Botany.

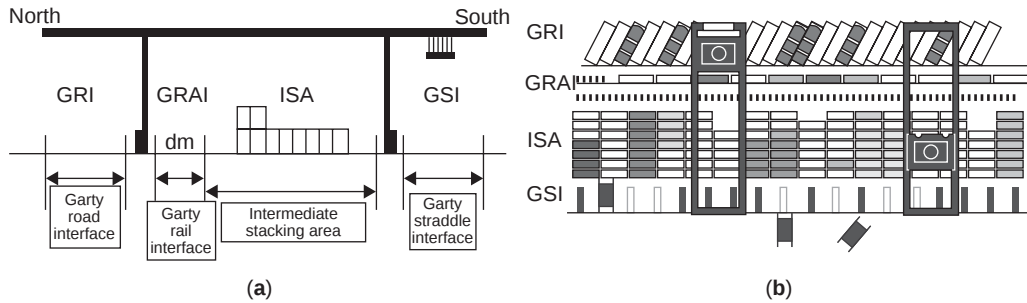


Figure 13. Layout of the container exchange area at Port Botany. (a) elevation and (b) ground plan.

optimization schemes. This work provides a foundation for research into the problem of empty containers—to move goods in a container; an empty container must be closely available in both time and space. When trade is balanced, the empty container can be loaded proximate to the destination of its full contents. When trade is imbalanced, significant repositioning of empty containers is required. In an ideal world, once a container

is emptied it would be moved directly to the closest shipper who requires one. However, due to the complex ownership, insurance, and information-sharing issues, empty containers are often inefficiently repositioned. Crainic *et al.* [41] developed a single-commodity, deterministic optimization model and found a multicommodity, stochastic model to be too detailed to be computationally tractable for large networks.

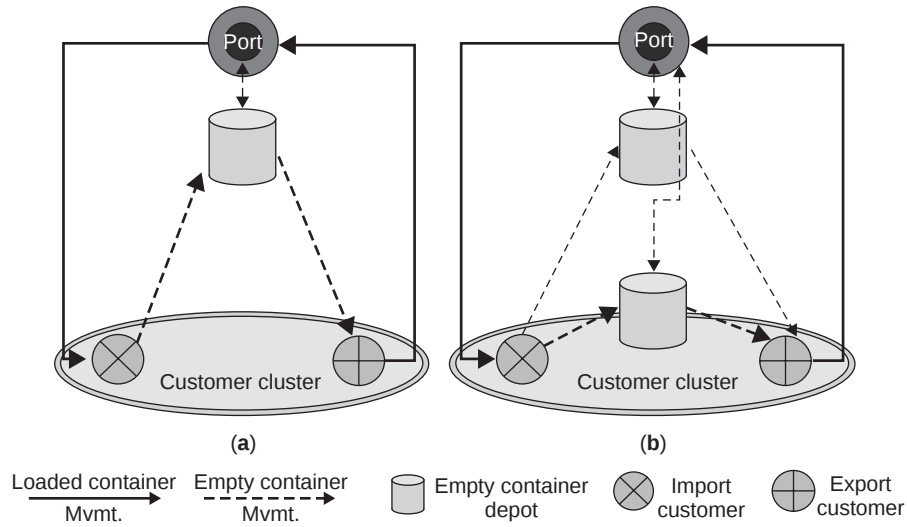


Figure 14. Container depot repositioning [42].

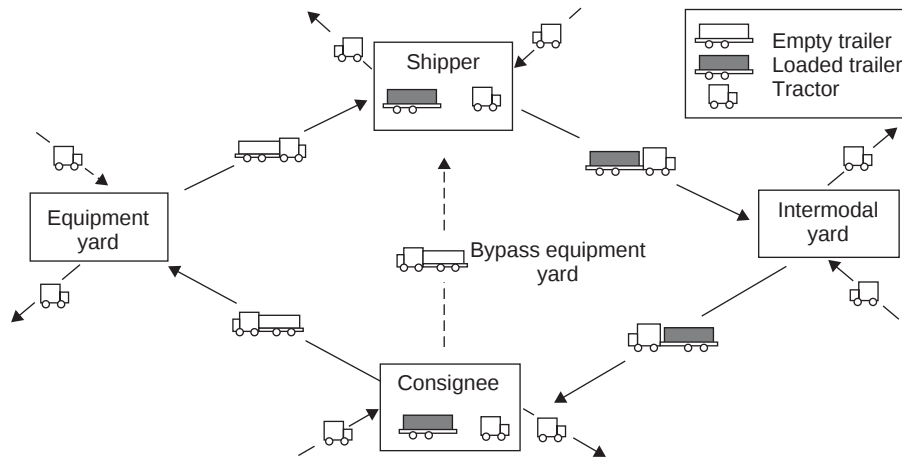


Figure 15. The problem of empty container movement [43].

Boile [42] looked for ways to reduce empty container movements through repositioning of empty depots, using optimization to identify locations that minimize travel, reducing the number of truck miles traveled (Fig. 14). Smilowitz [43] merges network modeling and optimization, assigning not just vehicles to routes, but also tasks to vehicles, considering both loaded and empty containers in an effort to reduce overall travel in a sample of movements in the

Chicago area. In doing so she can identify assignments that minimize vehicle miles traveled (Fig. 15).

Holguin-Veras and Thorson [44] examined the flow of empty containers using probability models, expanding on earlier work to include both a primary and secondary stop before a truck returns to its depot.

Real-Time Control of Landside Operation. Real-time control problems in landside

operation include determining the sequencing of drayage trucks for container transfer operations in the yard, deciding the storage location of containers following delivery and receiving operations, and scheduling the container transfers in real time to serve drayage trucks. The objectives of this problem include minimizing truck delays, rehandling work, and traveling distance of container handling equipment.

Sequencing trucks for container delivery and receiving operations is discussed in Kim *et al.* [45]. They assumed each truck has its own due time for the container transfer service by a yard crane, and their research objective was to minimize the total delay cost. First, a static sequencing problem was examined in which all the truck arrivals are known in advance, and a dynamic programming model is suggested. Later, a dynamic sequencing problem was considered in which new trucks arrive continuously, and a learning-based method for deriving decision rules is suggested and compared to several heuristic rules by simulation study. They found the learning-based rule outperformed the other sequencing rules including a first-come-first-served rule and a shortest-processing-time rule.

Research to determine the optimal storage location of rehandled containers to minimize the number of container rehandles can be found in Kim and Hong [34] and Zhao and Goodchild [26]. Zhao and Goodchild [26] examined how to reduce container rehandles by using drayage truck arrival information. A heuristic rule was used to find the optimal storage location of rehandled containers based on truck arrival information, and computer simulation was developed to evaluate how the truck information quality and container bay configuration affect the amount of rehandling work. Simulation results demonstrated any amount of truck information is useful for reducing number of container rehandles, and that updating information in real time significantly lowers information requirements. The impact of truck arrival information on yard crane productivity and truck transaction time was further investigated in Zhao and Goodchild [27].

Optimizing operations of a container exchange area (described in the section titled “Design of the Terminal’s Landside Connection”) at Port Botany in Sydney was discussed in Froyland *et al.* [38]. This exchange facility is a complex subsystem of the terminal, where both import and export container flows are handled simultaneously. The control decisions to be made includes scheduling the RMG movements, allocating container positions in the ISA, and allocating the delivery locations for trucks and straddle carriers. Multiple objectives were considered including moving all containers while satisfying time window constraints, ISA capacity limit, and target truck waiting time; minimizing the maximum number of straddle carriers required for moving containers between the main storage yard and the exchange area; and balancing the RMGs utilization. This problem does not address train operations and assumed all the data is known in advance for the entire month. Owing to the complexity of the decision making process, this problem was decomposed into three sequential subproblems and solved using a three-stage algorithm. The first subproblem is to assign all container moves at an hourly level and is formulated as an integer programming model. The second subproblem is to assign import containers to positions in the gantry-straddle carrier interface within each hour and is solved by integer programs. The third subproblem is to assign trucks to GRI positions and export containers to GSI positions, and schedule the precise RMG operations within each hour. Since the exact arrival times of trucks within every hour are unknown, an online algorithm was developed to solve the third subproblem. Computational results demonstrated that the proposed approach can find near-optimal solutions.

SUMMARY

This article introduced container terminal operations and a variety of decision problems that exist in operating their different subsystems (quayside, container yard, and

landside). The operations research methods used to address these problems are demonstrated through identifying examples in the academic literature. Depending on the length of planning horizon, the problems were classified into design problems, operational planning problems, and real-time control problems.

Owing to the complexity of container terminal operations, in almost all cases the many interrelated decisions are decomposed into hierarchical problems to simplify the decision process, and the solution of higher level problems are used as the constraints and inputs for optimizing lower level decisions. The majority of research in this field addressed only single types of material handling equipment or a detailed subproblem in the series of hierarchical problems, while a few studies considered an integrated system and optimized the operation of various equipments simultaneously. Most of the work in this field utilized computer simulation and operations research techniques such as queueing theory and stochastic models, network optimization, and metaheuristic optimization algorithms.

While these tools have had significant impact on terminal operations, future research will advance the state of these models by continuing to integrate subcomponents of the system, building more sophisticated operational research models that sufficiently address terminal complexity. It is a ripe area for future research and the application of operations research techniques.

REFERENCES

1. UNCTAD. Review of MARITIME transport. Geneva: United Nations Publications; 2001.
2. Ebeling CE. Evolution of a box. *Invent Technol* 2009;23(4):8–9.
3. Muller G. Intermodal freight transportation. 3rd ed. Westport (CN): Eno Foundation for Transportation; 1995.
4. Steenken D, Vos S, Stahlbock R. Container terminal operation and operations research: a classification and literature review. *OR Spectr* 2004;26(1):3–49.
5. Rodrigue JP, Slack B, Notteboom T. Port terminals. The geography of transport systems. New York: Routledge; 2009. Chapter 4.
6. Vis IFA, de Koster R. Transshipment of containers at a container terminal: an overview. *Eur J Oper Res* 2003;147(1):1–16.
7. Ioannou PA, Kosmatopoulos EB, Julia H, *et al.* Cargo handling technologies final report. 2000. Available at http://www.usc.edu/dept/ee/catt/2002/jula/Marine/FinalReport_CCDoTT_97.pdf. Accessed 2010.
8. Chinnery K. The Future Seems Rail-Mounted. *Port Strategy*. 2005. Available at http://www.portstrategy.com/features101/port-operations/cargo-handling/rail-mounted-gantry-cranes/the_future_seems_rail-mounted. Accessed 2010.
9. Huynh N. Analysis of container dwell time on marine terminal throughput and rehandling productivity. *J Int Logist Trade* 2008;6(2):69–89.
10. Günther HO, Kim KH. Container terminals and automated transport systems. Boca Raton (FL): Springer; 2005.
11. Kim KH. Operational issues in modern container terminals. *Intelligent freight transportation*. Boca Raton (FL): CRC Press; 2008. Chapter 4.
12. Stahlbock R, Vos S. Operations research at container terminals: a literature update. *OR Spectr* 2008;30(1):1–52.
13. Vis IFA, Harika I. Comparison of vehicle types at an automated container terminal. *OR Spectr* 2004;26(1):117–143.
14. Bierwirth C, Meisel F. A survey of berth allocation and quay crane scheduling problems in container terminals. *Eur J Oper Res* 2010;202:615–627.
15. Imai A, Nishimura E, Papadimitriou S. The dynamic berth allocation problem for a container port. *Transp Res Part B* 2001;35:401–417.
16. Imai A, Sun X, Nishimura E, *et al.* Berth allocation in a container port: using a continuous location space approach. *Transp Res Part B* 2005;39(3):199–221.
17. Kim KH, Moon KC. Berth scheduling by simulated annealing. *Transp Res Part B* 2003;37:541–560.
18. Imai A, Chen HC, Nishimura E, *et al.* The simultaneous berth and quay crane allocation problem. *Transp Res Part E* 2008;44:900–920.

19. Kim KH, Park YM. A crane scheduling method for port container terminals. *Eur J Oper Res* 2004;156:752–768.
20. Lee D-H, Wang HQ, Miao L. Quay crane scheduling with noninterference constraints in port container terminals. *Transp Res Part E* 2008;44:124–135.
21. Goodchild A, Daganzo C. Double-cycling strategies for container ships and their effect on ship loading and unloading operations. *Transp Sci* 2006;40(4):473–483.
22. Goodchild A, Daganzo C. Crane double-cycling in container ports: planning methods and evaluation. *Transp Res Part B* 2007;41:875–891.
23. Kang S, Medina JC, Ouyang YF. Optimal operations of transportation fleet for unloading activities at container ports. *Transp Res Part B* 2008;42:970–984.
24. Nishimura E, Akio Imai A, Papadimitriou S. Yard trailer routing at a maritime container terminal. *Transp Res Part E* 2005;41:53–76.
25. Kim KH, Bae JW. A look-ahead dispatching method for automated guided vehicles in automated port container terminals. *Transp Sci* 2004;38(2):224–234.
26. Zhao W, Goodchild AV. The impact of truck arrival information on container terminal rehandling. *Transp Res Part E* 2010;46:327–343.
27. Zhao W, Goodchild AV. The impact of truck arrival information on system efficiency at container terminals. *J Transp Res Board* 2010. In press.
28. Vis IFA. A comparative analysis of storage and retrieval equipment at a container terminal. *Int J Prod Econ* 2006;103:680–693.
29. Kim KH, Kim HB. The optimal sizing of the storage space and handling facilities for import containers. *Transp Res Part B* 2002;36:821–835.
30. Zhang C, Liu J, Wan Y, *et al.* Storage space allocation in container terminals. *Transp Res Part B* 2003;37:883–903.
31. Kim KH, Kim HB. Segregating space allocation models for container inventories in port container terminals. *Int J Prod Econ* 1999;59:415–423.
32. Kim KH, Kim KY. Routing straddle carriers for the loading operation of containers using a beam search algorithm. *Comput Ind Eng* 1999;36:109–136.
33. Kim KH, Park YM, Ryu K-R. Deriving decision rules to locate export containers in container yards. *Eur J Oper Res* 2000;124:89–101.
34. Kim KH, Hong G-P. A heuristic rule for relocating blocks. *Comput Oper Res* 2006;33:940–954.
35. Cheung RK, Li C-L, Lin W. Interblock crane deployment in container terminals. *Transp Sci* 2002;36(1):79–93.
36. Zhang C, Wan Y-W, Liu J, *et al.* Dynamic crane deployment in container storage yards. *Transp Res Part B* 2002;36:537–555.
37. Ng WC. Crane scheduling in container yards with inter-crane interference. *Eur J Oper Res* 2005;164:64–78.
38. Froyland G, Koch T, Megow N, *et al.* Optimizing the landside operation of a container terminal. *OR Spectr* 2008;30:53–75.
39. Huynh N, Walton CM. Robust scheduling of truck arrivals at marine container terminals. *J Transp Eng* 2008;134(8):347–353.
40. Holguín VJ, Walton CM. Implementation of priority systems for containers at marine intermodal terminals. *Transp Res Rec* 1997;1602:57–64.
41. Crainic TG, Gendreau M, Dejax P. Dynamic and stochastic models for the allocation of empty containers. *Oper Res* 1993;41(1):102–126.
42. Boile M, Theofanis S, Baveja A, *et al.* Regional repositioning of empty containers: Case for inland depots. *Transp Res Rec* 2008;2066:31–40.
43. Smilowitz K. Multi-resource routing with flexible tasks: An application in drayage operations. *IIE TRANSACTIONS* 2006;38(7):577–590.
44. Holguin-Veras J, Thorson E. Modeling commercial vehicle empty trips with a first order trip chain model. *Transp Res Part B* 2003;37(2):129–148.
45. Kim KH, Lee KM, Hwang H. Sequencing delivery and receiving operations for yard cranes in port container terminals. *Int J Prod Econ* 2003;84(3):283–292.

DECISION RULE PREFERENCE MODEL

SALVATORE GRECO
University of Portsmouth,
Portsmouth, UK
BENEDETTO MATARAZZO
University of Catania, Catania,
Italy
ROMAN SŁOWIŃSKI
Poznań University of Technology,
Poznań, Poland

INTRODUCTION

Decision rule model is a preference model for multiple criteria decision aiding (MCDA) (see [1] for a survey), constructed using the dominance-based rough set approach (DRSA) proposed by Greco *et al.* [2]. According to [3], MCDA aims at giving the decision maker (DM) a recommendation in terms of the best actions (*choice*), in terms of the assignment of actions to predefined and preference-ordered classes (*classification*, called also *sorting*), or of the ranking of actions from the best to the worst (*ranking*). Until the end of the last century, in the practice of MCDA, two major preference models were used: the multiple attribute utility theory (MAUT; see [4] and [5]) and the outranking approach (see [6] and [7]). As the preference models are main drivers of decision aiding, they are also called *decision models*. Building these models requires from the DM a specific preference information that is more or less explicitly related to their parameters, like substitution rates for a value function of MAUT, or discrimination thresholds and importance weights of criteria for outranking methods. Direct elicitation of this information requires a great cognitive effort on the side of the DM. Even if one can reduce this cognitive effort by acquiring indirectly these parameters through a training set

of decision examples, like in some ordinal regression methods [8–10], MAUT and outranking methods are often processing such indirect preference information in a way that is not clear for the DM, such that (s)he cannot see what are the exact relations between the provided information and the final recommendation. Consequently, very often, the decision model is perceived by the DM as a *black box* whose result has to be accepted because the analyst's authority guarantees that the result is "right". In this context, the aspiration of the DM to find good reasons for some decision is frustrated and rises the need for a more transparent methodology in which the relation between the original preference information and the final recommendation is clearly shown. Such a transparent methodology has been called *glass box* [11]. Its typical representative is using a training set of decision examples as the input preference information provided by the DM, and it is expressing the decision model in terms of a set of "if..., then ..." decision rules induced from the input preference information. Examples of such rules are the following:

- "if maximum speed of car x is at least 175 km/h, and its price is at most \$ 12,000, then car x is comprehensively at least medium,"
- "if car x is at least weakly preferred to car y with respect to acceleration, and the price of car x is no more than slightly worse than that of car y , then car x is at least as good as car y ".

From one side, the decision rules are explicitly related to the original information, and from the other side, they give understandable justifications for recommended decisions.

The rules induced from the input preference information provided in terms of

exemplary decisions represent a decision model, which is transparent for the DM and enables his/her understanding of the reasons of his/her past decisions. The acceptance of the rules by the DM justifies, in turn, their use for future decisions.

The induction of rules from examples is a typical approach of artificial intelligence. In this context, rough set theory [12–15] proved to be a useful tool for analysis of vague description of decision situations [16,17]. This description is given as a data table, where rows correspond to objects (actions) and columns to condition and decision attributes. Each row of the data table is a decision example. The aim of rough set analysis is the explanation of the dependence between the values of some decision attributes, playing the role of “dependent variables”, by means of the values of some condition attributes, playing the role of “independent variables”. For example, in a diagnostic context, data about the presence of some diseases are given by decision attributes, while data about symptoms are given by condition attributes. An important advantage of the rough set approach is that it can deal with partly inconsistent examples, that is, objects indiscernible by condition attributes but discernible by decision attributes (e.g., cases where the presence of different diseases is associated with the presence of the same symptoms). Moreover, it provides useful information about the role of particular attributes and their subsets and prepares the ground for representation of knowledge hidden in the data by means of “if..., then ...” decision rules relating values of some condition attributes with values of decision attributes (e.g., “if symptoms A and B are present, then there is disease X ”).

For a long time, however, the use of the rough set approach and, in general, of data mining techniques, has been restricted to classification problems where the preference order of evaluations is not considered. Typical examples of such problems come from medical diagnostics. In this context, symptom A is not better or worse than symptom B , or disease X is not preferable to disease Y . It is, thus, sufficient to consider A as different

from B and X as different from Y . There are, however, situations, where discernibility is not sufficient to handle all relevant information. Consider, for example, two firms, α and β , evaluated for assessment of bankruptcy risk by a set of criteria including the “debt ratio” (total debt/total assets). If firm α has a low value of the debt ratio while firm β has a high value of the debt ratio, then, within data mining and classical rough set theory, α is different (discernible) from β with respect to the considered attribute (debt ratio). However, from the viewpoint of preference analysis and, say, bankruptcy risk evaluation, the debt ratio of α is not simply different from the debt ratio of β but, clearly, the former is better than the latter.

The basic principle of the classical rough set approach (CRSA) and data mining techniques is the following *indiscernibility principle*: if x is indiscernible with y , that is, x has the same characteristics as y , then x should belong to the same class as y ; if not, there is an inconsistency between x and y . According to this principle, if a patient has symptom A and disease X while another patient has symptom B and disease Y , there is not any inconsistency and one can draw a simple conclusion that symptom A is associated with disease X , while symptom B is associated with disease Y .

In multiple criteria decision analysis, indiscernibility principle is not sufficient to convey all important semantics of the available information. Consider again the above two firms: α having a low value of debt ratio and β having a high value of the debt ratio. Suppose that evaluations of these firms on other attributes (profitability indices, quality of managers, market competitive situation, etc.) are equal. Suppose, moreover, that firm α has been assigned by a DM to a class of higher risk than firm β . According to the indiscernibility principle, one can say that α and β are discernible, and it follows that low debt ratio is associated with high risk while high debt ratio is associated with low risk. This is contradictory, of course. The reason is that, within multiple criteria decision analysis, the indiscernibility principle has to be replaced by the following *dominance principle*: if x dominates y , that is, x is at least

as good as y with respect to all considered criteria (condition attributes), then x should belong to a class not worse than the class of y (class assignment by decision attribute); if not, there is an inconsistency between x and y . Applying the dominance principle to α and β , one can state an inconsistency between their debt ratio and the risk of their bankruptcy, which leads to a paradoxical conclusion that the lower the debt ratio is, the higher will be the risk of bankruptcy.

For this reason, Greco *et al.* [2, 18, 19–21] have proposed an extension of rough set theory based on the dominance principle, which permits to deal with MCDA problems. This innovation is mainly based on substitution of the indiscernibility relation by a dominance relation in the rough approximation of decision classes. An important consequence of this fact is a possibility of inferring from exemplary decisions the preference model in terms of decision rules being logical statements of the type “if..., then...”. The separation of certain and doubtful knowledge about the DM’s preferences is done by distinction of different kinds of decision rules, depending whether they are induced from examples consistent with the dominance principle or from examples inconsistent with the dominance principle. The latter rules are very important because they represent situations of hesitation in the DM’s expression of preferences.

This is to say that, using a properly modified technique of artificial intelligence, one can construct a preference model in form of decision rules, by inducing them from exemplary decisions. Such a preference model has some interesting properties: it is expressed in a natural language without using any complex analytical formulation, its interpretation is immediate and does not depend on technical parameters, it is often using only a subset of the considered attributes in each rule, and, finally, it can represent situations of hesitation, typical for a real expression of preferences. The decision rule preference model is, therefore, a new approach to MCDA, which becomes a strong alternative for MAUT and outranking methods.

Dominance-Based Rough Set Approach (DRSA)

Criteria and the class assignment considered within DRSA correspond to the *condition attributes* and the *decision attribute*, respectively, in the CRSA [12, 13, 16]. Different from the CRSA, in DRSA, we are considering criteria that are condition attributes with value sets ordered according to decreasing or increasing preference.

Let us consider the set of n criteria $F = \{g_1, \dots, g_n\}$, the set of their indices $I = \{1, \dots, n\}$, and a finite universe of objects (actions, solutions, and alternatives) U such that, without loss of generality, $g_i : U \rightarrow \mathbf{R}$ for each $i = 1, \dots, n$, and, for all objects $x, y \in U$, $g_i(x) \geq g_i(y)$ means that “ x is at least as good as y with respect to criterion i ”, which is denoted as $x \succsim_i y$ (observe that in case of a criterion g_i that, like the price of a car, is decreasing with respect to the preferences, it is enough to multiply its values by -1 to get again a criterion increasing with respect to the preferences). Therefore, we suppose that $\succsim_i$ is a complete preorder, that is, a strongly complete and transitive binary relation, defined on U on the basis of evaluations $g_i(\cdot)$. Note that, in the context of multiobjective optimization, g_i corresponds to an objective function. Furthermore, we assume that there is a decision attribute d that makes a partition of U into a finite number of decision classes, called *classification*, $\mathbf{Cl} = \{\text{Cl}_1, \dots, \text{Cl}_m\}$, such that each $x \in U$ belongs to one and only one class Cl_t , $t = 1, \dots, m$. We suppose that the classes are preference ordered, that is, for all $r, s = 1, \dots, m$, such that $r > s$, the objects from Cl_r are preferred to the objects from Cl_s . More formally, if $\succsim$ is a *comprehensive weak preference relation* on U , that is, if for all $x, y \in U$, $x \succsim y$ reads “ x is at least as good as y ”, then we suppose

$$[x \in \text{Cl}_r, y \in \text{Cl}_s, r > s] \Rightarrow x \succ y,$$

where $x \succ y$ means $x \succsim y$ and *not* $y \succsim x$. The above assumptions are typical for consideration of a multiple criteria classification problem (also called *ordinal classification with monotonicity constraints*).

In DRSA, the explanation of the assignment of objects to preference-ordered decision classes is made on the basis of their evaluation with respect to a subset of criteria $P \subseteq I$. This explanation is called *approximation* of decision classes with respect to P . Of course, the most commonly considered case is $P = I$. In order to take into account the order of decision classes, in DRSA, the classes are not considered one by one, but, instead, unions of classes are approximated: *upward union* from class Cl_t to class Cl_m , denoted by $Cl_t^{\geq}$, and *downward union* from class Cl_t to class Cl_1 , denoted by $Cl_t^{\leq}$, that is:

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s, \quad Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s, \quad t = 1, \dots, m.$$

The statement $x \in Cl_t^{\geq}$ reads “ x belongs to at least class Cl_t ”, while $x \in Cl_t^{\leq}$ reads “ x belongs to at most class Cl_t ”. Let us remark that $Cl_1^{\geq} = Cl_m^{\leq} = U$, $Cl_m^{\geq} = Cl_m$, and $Cl_1^{\leq} = Cl_1$. Furthermore, for $t = 2, \dots, m$, we have

$$Cl_{t-1}^{\leq} = U - Cl_t^{\geq} \quad \text{and} \quad Cl_t^{\geq} = U - Cl_{t-1}^{\leq}.$$

For example, in multiple criteria classification of cars, the considered cars are the objects evaluated on such criteria as maximum speed, acceleration, price, and fuel consumption, and they are to be assigned to one of three classes of overall evaluation: “bad”, “medium”, and “good”. Then, the upward unions are: $Cl_{\text{medium}}^{\geq}$, that is the set of all the cars classified at least medium (i.e., the set of cars classified as medium or good), and $Cl_{\text{good}}^{\geq}$, that is the set of all the cars classified at least good (i.e., the set of cars classified as good), while the downward unions are: $Cl_{\text{medium}}^{\leq}$, that is the set of all the cars classified at most medium (i.e., the set of cars classified as medium or bad), and $Cl_{\text{bad}}^{\leq}$, that is the set of all the cars classified at most bad (i.e., the set of cars classified as bad). Notice that, formally, also $Cl_{\text{bad}}^{\geq}$ is an upward union and $Cl_{\text{good}}^{\leq}$ is a downward union; however, as bad and good are extreme classes, these two unions boil down to the whole universe U .

The key idea of the rough set approach is explanation (approximation) of knowledge generated by the decision attributes, by *granules of knowledge* generated by condition attributes.

In DRSA, where condition attributes are criteria and decision classes are preference ordered, the knowledge to be explained is the assignment of objects to *upward* and *downward unions of classes* and the granules of knowledge are sets of objects contained in *dominance cones* defined in the space of evaluation criteria.

We say that object x *dominates* object y with respect to $P \subseteq I$ (shortly, x *P-dominates* y), denoted by $x D_P y$, if for every criterion $i \in P$, $g_i(x) \geq g_i(y)$. The relation of *P-dominance* is reflexive and transitive, that is, it is a partial preorder.

Given a set of criteria $P \subseteq I$ and $x \in U$, the granules of knowledge used for approximation in DRSA are:

- a set of objects dominating x , called *P-dominating set*,
 $D_P^+(x) = \{y \in U : y D_P x\}$,
- a set of objects dominated by x , called *P-dominated set*,
 $D_P^-(x) = \{y \in U : x D_P y\}$.

In terms of multiple criteria evaluations, the above definitions correspond to:

$$D_P^+(x) = \{y \in U : g_i(y) \geq g_i(x), \text{ for all } i \in P\},$$

$$D_P^-(x) = \{y \in U : g_i(y) \leq g_i(x), \text{ for all } i \in P\}.$$

Let us recall that the dominance principle requires that an object x dominating object y with respect to considered criteria (i.e., x having evaluations at least as good as y on all considered criteria) should also dominate y on the decision (i.e., x should be assigned to at least as good decision class as y). This principle is the only objective principle that is widely agreed upon in the multiple criteria comparisons of objects.

Given $P \subseteq I$, the inclusion of an object $x \in U$ to the upward union of classes $Cl_t^{\geq}$, $t = 2, \dots, m$, is *inconsistent with the dominance principle* if one of the following conditions holds:

- x belongs to class Cl_t or better but it is *P-dominated* by an object y belonging to a class worse than Cl_t , i.e., $x \in Cl_t^{\geq}$ but $D_P^+(x) \cap Cl_{t-1}^{\leq} \neq \emptyset$,

- x belongs to a worse class than Cl_t but it P -dominates an object y belonging to class Cl_t or better, i.e., $x \notin Cl_t^{\geq}$ but $D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset$.

If, given a set of criteria $P \subseteq I$, the inclusion of $x \in U$ to $Cl_t^{\geq}$, where $t = 2, \dots, m$, is inconsistent with the dominance principle, we say that x belongs to $Cl_t^{\geq}$ *with some ambiguity*. Thus, x belongs to $Cl_t^{\geq}$ *without any ambiguity* with respect to $P \subseteq I$, if $x \in Cl_t^{\geq}$, and there is no inconsistency with the dominance principle. This means that all objects P -dominating x belong to $Cl_t^{\geq}$, i.e., $D_P^+(x) \subseteq Cl_t^{\geq}$.

Furthermore, object x *possibly belongs to* $Cl_t^{\geq}$ with respect to $P \subseteq I$ if one of the following conditions holds:

- according to decision attribute d , x belongs to $Cl_t^{\geq}$,
- according to decision attribute d , x does not belong to $Cl_t^{\geq}$, but it is inconsistent in the sense of the dominance principle with an object y belonging to $Cl_t^{\geq}$.

In terms of ambiguity, x possibly belongs to $Cl_t^{\geq}$ with respect to $P \subseteq I$, if x belongs to $Cl_t^{\geq}$ with or without any ambiguity. Owing to the reflexivity of the dominance relation D_P , the above conditions can be summarized as follows: x *possibly belongs to* class Cl_t or better, with respect to $P \subseteq I$, if among the objects P -dominated by x there is an object y belonging to class Cl_t or better, that is, $D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset$.

The P -lower approximation of $Cl_t^{\geq}$, denoted by $\underline{P}(Cl_t^{\geq})$, and the P -upper approximation of $Cl_t^{\geq}$, denoted by $\overline{P}(Cl_t^{\geq})$, are defined as follows ($t = 2, \dots, m$):

$$\begin{aligned}\underline{P}(Cl_t^{\geq}) &= \{x \in U : D_P^+(x) \subseteq Cl_t^{\geq}\}, \\ \overline{P}(Cl_t^{\geq}) &= \{x \in U : D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset\}.\end{aligned}$$

Analogously, one can define the P -lower approximation and the P -upper approximation of $Cl_t^{\leq}$, as follows ($t = 1, \dots, m-1$):

$$\begin{aligned}\underline{P}(Cl_t^{\leq}) &= \{x \in U : D_P^-(x) \subseteq Cl_t^{\leq}\}, \\ \overline{P}(Cl_t^{\leq}) &= \{x \in U : D_P^+(x) \cap Cl_t^{\leq} \neq \emptyset\}.\end{aligned}$$

The P -boundaries of $Cl_t^{\geq}$ and $Cl_t^{\leq}$, denoted by $Bn_P(Cl_t^{\geq})$ ($t = 2, \dots, m$) and $Bn_P(Cl_t^{\leq})$ ($t = 1, \dots, m-1$), respectively, are defined as follows:

$$\begin{aligned}Bn_P(Cl_t^{\geq}) &= \overline{P}(Cl_t^{\geq}) - \underline{P}(Cl_t^{\geq}), \\ Bn_P(Cl_t^{\leq}) &= \overline{P}(Cl_t^{\leq}) - \underline{P}(Cl_t^{\leq}).\end{aligned}$$

For every $P \subseteq I$, the objects that are consistent in the sense of the dominance principle with all upward and downward unions of classes are P -correctly classified. For every $P \subseteq I$, the *quality of classification* $\mathbf{Cl}$ by a set of criteria $P \subseteq I$ is defined as the ratio of the number of objects P -consistent with the dominance principle and the number of all the objects in U . Since the P -consistent objects are those that do not belong to any P -boundary $Bn_P(Cl_t^{\geq})$ or $Bn_P(Cl_t^{\leq})$, and $Bn_P(Cl_t^{\geq}) = Bn_P(Cl_{t-1}^{\leq})$ for $t = 2, \dots, m$, the quality of classification $\mathbf{Cl}$ by a set of criteria $P \subseteq I$ can be written as

$$\begin{aligned}\gamma_P(\mathbf{Cl}) &= \frac{\left| U - \left(\bigcup_{t \in \{2, \dots, m\}} Bn_P(Cl_t^{\geq}) \right) \right|}{|U|} \\ &= \frac{\left| U - \left(\bigcup_{t \in \{1, \dots, m-1\}} Bn_P(Cl_t^{\leq}) \right) \right|}{|U|}.\end{aligned}$$

$\gamma_P(\mathbf{Cl})$ can be seen as a degree of consistency of the classification examples, where P is the set of criteria and $\mathbf{Cl}$ is the considered classification.

Each minimal (in the sense of inclusion) subset $P \subseteq I$, such that $\gamma_P(\mathbf{Cl}) = \gamma_I(\mathbf{Cl})$, is called a *reduct* of classification $\mathbf{Cl}$ and is denoted by $\text{RED}_{\mathbf{Cl}}$. Let us remark that, for a given set of classification examples, one can have more than one reduct. The intersection of all reducts is called the *core* and is denoted by $\text{CORE}_{\mathbf{Cl}}$. Criteria in $\text{CORE}_{\mathbf{Cl}}$ cannot be removed from consideration without deteriorating the quality of classification $\mathbf{Cl}$. This means that, in set I , there are three categories of criteria:

- *indispensable* criteria included in the core,
- *exchangeable* criteria included in some reducts but not in the core,

- *redundant* criteria, neither indispensable nor exchangeable, and, thus, not included in any reduct.

The dominance-based rough approximations of upward and downward unions of decision classes can serve to induce a generalized description of classification decisions in terms of “if . . . , then . . . ” decision rules. For a given upward or downward union of classes, $Cl_t^{\geq}$ or $Cl_s^{\leq}$, the decision rules induced under a hypothesis that objects belonging to $\underline{P}(Cl_t^{\geq})$ or $\underline{P}(Cl_s^{\leq})$ are positive examples (i.e., objects that have to be matched by the induced decision rules), and all the others are negative (i.e., objects that have to be not matched by the induced decision rules), suggest a *certain* assignment to “class Cl_t or better” or to “class Cl_s or worse”, respectively. On the other hand, the decision rules induced under a hypothesis that objects belonging to $\overline{P}(Cl_t^{\geq})$ or $\overline{P}(Cl_s^{\leq})$ are positive examples, and all the others are negative, suggest a *possible* assignment to “class Cl_t or better”, or to “class Cl_s or worse”, respectively. Finally, the decision rules induced under a hypothesis that objects belonging to the intersection $\overline{P}(Cl_s^{\leq}) \cap \overline{P}(Cl_t^{\geq})$ are positive examples, and all the others are negative, suggest an assignment to some classes between Cl_s and Cl_t ($s < t$). These rules are matching inconsistent objects $x \in U$, which cannot be assigned without doubts to classes Cl_r , $s < r < t$, because $x \notin \underline{P}(Cl_r^{\geq})$ and $x \notin \underline{P}(Cl_r^{\leq})$ for all r such that $s < r < t$.

Given the preference information in terms of classification examples, it is meaningful to consider the following five types of decision rules:

1. *certain* $D_{\geq}$ -decision rules, providing lower profiles (i.e., sets of minimal values for considered criteria) of objects belonging to $\underline{P}(Cl_t^{\geq})$, $P = \{i_1, \dots, i_p\} \subseteq I$: if $g_{i_1}(x) \geq r_{i_1}$ and . . . and $g_{i_p}(x) \geq r_{i_p}$, then $x \in Cl_t^{\geq}$, $t = 2, \dots, m$, $r_{i_1}, \dots, r_{i_p} \in \mathbf{R}$; for example, this is the case of a rule “if maximum speed of car x is at least 175 km/h, and its price (to be minimized) is at most \$ 12,000, then car x is comprehensively at least medium”;
2. *possible* $D_{\geq}$ -decision rules, providing lower profiles of objects belonging to $\overline{P}(Cl_t^{\geq})$, $P = \{i_1, \dots, i_p\} \subseteq I$: if $g_{i_1}(x) \geq r_{i_1}$ and . . . and $g_{i_p}(x) \geq r_{i_p}$, then x possibly belongs to $Cl_t^{\geq}$, $t = 2, \dots, m$, $r_{i_1}, \dots, r_{i_p} \in \mathbf{R}$; for example, this is the case of a rule “if maximum speed of car x is at least 130 km/h, and its fuel consumption (to be minimized) is at most 7 l/100 km, then car x is possibly comprehensively at least medium”;
3. *certain* $D_{\leq}$ -decision rules, providing upper profiles (i.e., sets of maximal values for considered criteria) of objects belonging to $\underline{P}(Cl_t^{\leq})$, $P = \{i_1, \dots, i_p\} \subseteq I$: if $g_{i_1}(x) \leq r_{i_1}$ and . . . and $g_{i_p}(x) \leq r_{i_p}$, then $x \in Cl_t^{\leq}$, $t = 1, \dots, m-1$, $r_{i_1}, \dots, r_{i_p} \in \mathbf{R}$; for example, this is the case of a rule “if maximum speed of car x is at most 140 km/h, and its price is at least \$ 18,000, then car x is comprehensively at most medium”;
4. *possible* $D_{\leq}$ -decision rules, providing upper profiles of objects belonging to $\overline{P}(Cl_t^{\leq})$, $P = \{i_1, \dots, i_p\} \subseteq I$: if $g_{i_1}(x) \leq r_{i_1}$ and . . . and $g_{i_p}(x) \leq r_{i_p}$, then x possibly belongs to $Cl_t^{\leq}$, $t = 1, \dots, m-1$, $r_{i_1}, \dots, r_{i_p} \in \mathbf{R}$; for example, this is the case of a rule “if maximum speed of car x is at most 140 km/h, and its fuel consumption is at least 6 l/100 km, then car x is possibly comprehensively at most medium”;
5. *approximate* $D_{\geq\leq}$ -decision rules, providing simultaneously lower and upper profiles of objects belonging to $Cl_s \cup Cl_{s+1} \cup \dots \cup Cl_t$, without possibility of discerning to which class: if $g_{i_1}(x) \geq r_{i_1}$ and . . . and $g_{i_k}(x) \geq r_{i_k}$ and $g_{i_{k+1}}(x) \leq r_{i_{k+1}}$ and . . . and $g_{i_p}(x) \leq r_{i_p}$, then $x \in Cl_s \cup Cl_{s+1} \cup \dots \cup Cl_t$, $\{i_1, \dots, i_p\} \subseteq I$, $s, t \in \{1, \dots, m\}$, $s < t$, $r_{i_1}, \dots, r_{i_p} \in \mathbf{R}$; for example, this is the case of a rule “if maximum speed of car x is at least 140 km/h, and its fuel consumption is at least 6 l/100 km, then car x is comprehensively medium or good”.

The rules of type 1 and 3 represent certain knowledge extracted from the data table, while the rules of type 2 and 4 represent

possible knowledge. Rules of type 5 represent doubtful knowledge. Algorithms for induction of decision rules from rough approximations of upward and downward unions of decision classes have been described in [22, 23].

It is well known that the computational complexity of the problem of searching for reducts, as well as the problem of generating minimal-cover or all decision rules, is NP-complete, even if the attributes are not ordered (CRSA) [14]. Consideration of ordered attributes (criteria) does not increase the computational complexity [22]. For this reason, for large scale problems, efficient heuristics have been proposed, which perform pretty good. In case of very big data, sampling and bagging techniques are recommended. It is also worth mentioning that there is an advantage in consideration of ordered instead of unordered attributes: in case of ordered attributes, the granules used to build approximations are dominance cones in the evaluation space, with the origins in particular objects; however, in case of unordered attributes, the granules are limited hypercubes in the evaluation space, resulting from discretization of this space. As discretization is always arbitrary to some extent, and it has an impact on the results, it is good to note that DRSA does not require discretization.

More detailed surveys and tutorials on DRSA can be found in [21, 24–26].

An Illustrative Practical Example

Similar to our previous tutorials (e.g., [26–28]), to illustrate the application of DRSA to a real-world multicriteria classification problem, we will use a part of the data provided by a Greek industrial bank ETEVA, which finances industrial and commercial firms in Greece [29]. A sample composed of 39 firms has been chosen for the study in co-operation with the ETEVA's financial manager. The manager has classified the selected firms into three classes of bankruptcy risk. The sorting decision is represented by decision attribute d , making a trichotomic partition of the 39 firms:

$d=A$ means “acceptable”,
 $d=U$ means “uncertain”,
 $d=NA$ means “nonacceptable”.

The partition is denoted by $\mathbf{CI} = \{Cl_A, Cl_U, Cl_{NA}\}$ and, obviously, class Cl_A is better than Cl_U , which is better than Cl_{NA} .

The firms were evaluated using the following 12 criteria ($\uparrow$ means *preference increasing with value* and $\downarrow$ means *preference decreasing with value*):

- g_1 = earnings before interests and taxes/total assets, $\uparrow$
- g_2 = net income/net worth, $\uparrow$
- g_3 = total liabilities/total assets, $\downarrow$
- g_4 = total liabilities/cash flow, $\downarrow$
- g_5 = interest expenses/sales, $\downarrow$
- g_6 = general and administrative expense/sales, $\downarrow$
- g_7 = managers' work experience, $\uparrow$ (very low=1, low=2, medium=3, high=4, very high=5)
- g_8 = firm's market niche/position, $\uparrow$ (bad=1, rather bad=2, medium=3, good=4, very good=5)
- g_9 = technical structure—facilities, $\uparrow$ (bad=1, rather bad=2, medium=3, good=4, very good=5)
- g_{10} = organization—personnel, $\uparrow$ (bad=1, rather bad=2, medium=3, good=4, very good=5)
- g_{11} = special competitive advantage of firms, $\uparrow$ (low=1, medium=2, high=3, very high=4)
- g_{12} = market flexibility, $\uparrow$ (very low=1, low=2, medium=3, high=4, very high=5).

The first six criteria are cardinal (financial ratios) and the last six are ordinal. The data table is presented in Table 1.

The main questions to be answered by the knowledge discovery process were the following:

- Is the information contained in Table 1 consistent?
- What are the reducts of criteria ensuring the same quality of approximation

Table 1. Financial Data Matrix

Firm	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	d (class)
F1	16.4	14.5	59.82	2.5	7.5	5.2	5	3	5	4	2	4	A
F2	35.8	67.0	64.92	1.7	2.1	4.5	5	4	5	5	4	5	A
F3	20.6	61.75	75.71	3.6	3.6	8.0	5	3	5	5	3	5	A
F4	11.5	17.1	57.1	3.8	4.2	3.7	5	2	5	4	3	4	A
F5	22.4	25.1	49.8	2.1	5.0	7.9	5	3	5	5	3	5	A
F6	23.9	34.5	48.9	1.7	2.5	8.0	5	3	4	4	3	4	A
F7	29.9	44.0	57.8	1.8	1.7	2.5	5	4	4	5	3	5	A
F8	8.7	5.4	27.4	3.3	4.5	4.5	5	2	4	4	1	4	A
F9	25.7	29.7	46.8	1.7	4.6	3.7	4	2	4	3	1	3	A
F10	21.2	24.6	64.8	3.7	3.6	8.0	4	2	4	4	1	4	A
F11	18.32	31.6	69.3	4.4	2.8	3.0	4	3	4	4	3	4	A
F12	20.7	19.3	19.7	0.7	2.2	4.0	4	2	4	4	1	3	A
F13	9.9	3.5	53.1	4.5	8.5	5.3	4	2	4	4	1	4	A
F14	10.4	9.3	80.9	9.4	1.4	4.1	4	2	4	4	3	3	A
F15	17.7	19.8	52.8	3.2	7.9	6.1	4	4	4	4	2	5	A
F16	14.8	15.9	27.94	1.3	5.4	1.8	4	2	4	3	2	3	A
F17	16.0	14.7	53.5	3.9	6.8	3.8	4	4	4	4	2	4	A
F18	11.7	10.01	42.1	3.9	12.2	4.3	5	2	4	2	1	3	A
F19	11.0	4.2	60.8	5.8	6.2	4.8	4	2	4	4	2	4	A
F20	15.5	8.5	56.2	6.5	5.5	1.8	4	2	4	4	2	4	A
F21	13.2	9.1	74.1	11.21	6.4	5.0	2	2	4	4	2	3	U
F22	9.1	4.1	44.8	4.2	3.3	10.4	3	4	4	4	3	4	U
F23	12.9	1.9	65.02	6.9	14.01	7.5	4	3	3	2	1	2	U
F24	5.9	-27.7	77.4	-32.2	16.6	12.7	3	2	4	4	2	3	U
F25	16.9	12.4	60.1	5.2	5.6	5.6	3	2	4	4	2	3	U
F26	16.7	13.1	73.5	7.1	11.9	4.1	2	2	4	4	2	3	U
F27	14.6	9.7	59.5	5.8	6.7	5.6	2	2	4	4	2	4	U
F28	5.1	4.9	28.9	4.3	2.5	46.0	2	2	3	3	1	2	U
F29	24.4	22.3	32.8	1.4	3.3	5.0	2	3	4	4	2	3	U
F30	29.7	8.6	41.8	1.6	5.2	6.4	2	3	4	4	2	3	U
F31	7.3	-64.5	67.5	-2.2	30.1	8.7	3	3	4	4	2	3	NA
F32	23.7	31.9	63.6	3.5	12.1	10.2	3	2	3	4	1	3	NA
F33	18.9	13.5	74.5	10.0	12.0	8.4	3	3	3	4	3	4	NA
F34	13.9	3.3	78.7	25.5	14.7	10.1	2	2	3	4	3	4	NA
F35	-13.3	-31.1	63.0	-10.0	21.2	23.1	2	1	4	3	1	2	NA
F36	6.2	-3.2	46.1	5.1	4.8	10.5	2	1	3	3	2	3	NA
F37	4.8	-3.3	71.9	34.6	8.6	11.6	2	2	4	4	2	3	NA
F38	0.1	-9.6	42.5	-20.0	12.9	12.4	1	1	4	3	1	3	NA
F39	13.6	9.1	76.0	11.4	17.1	10.3	1	1	2	1	1	2	NA

of the multicriteria classification as the whole set of criteria?

- What decision rules can be extracted from Table 1?
- What are the minimal sets of decision rules?

We will answer these questions using DRSA. The **first result** from this approach is a discovery that the financial data matrix

is **consistent** for the complete set of criteria C . Therefore, the C -lower and C -upper approximations of $Cl_{NA}^{\leq}, Cl_U^{\leq}$ and $Cl_U^{\geq}, Cl_A^{\geq}$ are, respectively, the same. In other words, the quality of approximation of all upward and downward unions of classes, as well as the quality of classification, is equal to 1.

The **second discovery** is a set of 18 **reducts** of criteria ensuring the same quality of classification as the whole set of 12 criteria:

$$\begin{aligned}
\text{RED}_{\text{CI}}^1 &= \{g_1, g_4, g_5, g_7\}, & \text{RED}_{\text{CI}}^2 &= \{g_2, g_4, g_5, g_7\}, \\
\text{RED}_{\text{CI}}^3 &= \{g_3, g_4, g_6, g_7\}, & \text{RED}_{\text{CI}}^4 &= \{g_4, g_5, g_6, g_7\}, \\
\text{RED}_{\text{CI}}^5 &= \{g_4, g_5, g_7, g_8\}, & \text{RED}_{\text{CI}}^6 &= \{g_2, g_3, g_7, g_9\}, \\
\text{RED}_{\text{CI}}^7 &= \{g_1, g_3, g_4, g_7, g_9\}, & \text{RED}_{\text{CI}}^8 &= \{g_1, g_5, g_7, g_9\}, \\
\text{RED}_{\text{CI}}^9 &= \{g_2, g_5, g_7, g_9\}, & \text{RED}_{\text{CI}}^{10} &= \{g_4, g_5, g_7, g_9\}, \\
\text{RED}_{\text{CI}}^{11} &= \{g_5, g_6, g_7, g_9\}, & \text{RED}_{\text{CI}}^{12} &= \{g_4, g_5, g_7, g_{10}\}, \\
\text{RED}_{\text{CI}}^{13} &= \{g_1, g_3, g_4, g_7, g_{11}\}, & \text{RED}_{\text{CI}}^{14} &= \{g_2, g_3, g_4, g_7, g_{11}\}, \\
\text{RED}_{\text{CI}}^{15} &= \{g_4, g_5, g_6, g_{12}\}, & \text{RED}_{\text{CI}}^{16} &= \{g_1, g_3, g_5, g_6, g_9, g_{12}\}, \\
\text{RED}_{\text{CI}}^{17} &= \{g_3, g_4, g_6, g_{11}, g_{12}\}, & \text{RED}_{\text{CI}}^{18} &= \{g_1, g_2, g_3, g_6, g_9, g_{11}, g_{12}\}.
\end{aligned}$$

All the eighteen subsets of criteria are equally good and sufficient for the perfect approximation of the classification performed by ETEVA's financial manager on the 39 firms. The core of CI is empty ($\text{CORE}_{\text{CI}} = \emptyset$), which means that no criterion is indispensable for the approximation. Moreover, all the criteria are exchangeable, and no criterion is redundant.

The **third discovery** is the set of *all* decision rules. We obtained 74 rules describing $\text{Cl}_{\text{NA}}^{\leq}$, 51 rules describing $\text{Cl}_{\text{U}}^{\leq}$, 75 rules describing $\text{Cl}_{\text{U}}^{\geq}$, and 79 rules describing $\text{Cl}_{\text{A}}^{\geq}$.

The **fourth discovery** is the finding of *minimal sets* of decision rules. Several minimal sets were found. One of them is shown below. The number in parenthesis indicates the number of objects (firms) that support the corresponding rule, that is, the rule strength:

1. **if** $g_3(x) \geq 67.5$ **and** $g_4(x) \geq -2.2$ **and** $g_6(x) \geq 8.7$, **then** $x \in \text{Cl}_{\text{NA}}^{\leq}$, (4),
2. **if** $g_2(x) \leq 3.3$ **and** $g_7(x) \leq 2$, **then** $x \in \text{Cl}_{\text{NA}}^{\leq}$, (5),
3. **if** $g_3(x) \geq 63.6$ **and** $g_7(x) \leq 3$ **and** $g_9(x) \leq 3$, **then** $x \in \text{Cl}_{\text{NA}}^{\leq}$, (4),
4. **if** $g_2(x) \leq 12.4$ **and** $g_6(x) \geq 5.6$, **then** $x \in \text{Cl}_{\text{U}}^{\leq}$, (14),
5. **if** $g_7(x) \leq 3$, **then** $x \in \text{Cl}_{\text{U}}^{\leq}$, (18),
6. **if** $g_2(x) \geq 3.5$ **and** $g_5(x) \leq 8.5$, **then** $x \in \text{Cl}_{\text{U}}^{\geq}$, (26),
7. **if** $g_7(x) \geq 4$, **then** $x \in \text{Cl}_{\text{U}}^{\geq}$, (21),
8. **if** $g_1(x) \geq 8.7$ **and** $g_9(x) \geq 4$, **then** $x \in \text{Cl}_{\text{U}}^{\geq}$, (27),
9. **if** $g_2(x) \geq 3.5$ **and** $g_7(x) \geq 4$, **then** $x \in \text{Cl}_{\text{A}}^{\geq}$, (20).

As the minimal set of rules is complete and composed of $\text{D}_{\geq}$ -decision rules and $\text{D}_{\leq}$ -decision rules only, application of these rules to the 39 firms will result in their exact reclassification to classes of risk.

Minimal sets of decision rules represent the most concise and nonredundant knowledge representations. The above minimal set of 9 decision rules uses 8 criteria and 18 elementary conditions, that is, 3.85% of descriptors from the financial data matrix.

DRSA FOR MULTIPLE-CRITERIA CHOICE AND RANKING

DRSA can also be applied to multiple criteria choice and ranking problems. However, there is a basic difference between classification problems from one side, and choice and ranking from the other side. To give a didactic example, consider a set of companies A for evaluation of a risk of failure, taking into account the debt ratio criterion. To assess the risk of failure of company x , we will not compare the debt ratio of x with the debt ratio of all the other companies from A . The comparison will be made with respect to a fixed risk threshold on the debt ratio criterion. Indeed, the debt ratio of x can be the highest of all companies from A , and nevertheless, x can be classified as a low risk company if its debt ratio is below the fixed risk threshold. Consider, in turn, the situation, in which we must choose the lowest risk company from A or we want to rank these companies from the less risky to the most risky one. In this situation, the comparison of the debt ratio of x with a

fixed risk threshold is not useful, and instead, a pairwise comparison of the debt ratio of x with the debt ratio of all other companies in A is relevant for the choice or ranking. Thus, in general, while classification is based on absolute evaluation of objects (e.g., comparison of the debt ratio with the fixed risk threshold), choice and ranking refer to relative evaluation, by means of pairwise comparisons of objects (e.g., comparisons of the debt ratios for pairs of companies).

The necessity of pairwise comparisons of objects in multiple criteria choice and ranking problems requires some further extensions of DRSA. Simply speaking, in this context, we are interested in the approximation of a binary relation (corresponding to a comprehensive preference relation) using other binary relations (corresponding to marginal preference relations on particular criteria) for pairs of objects. In the above example, we would approximate the binary relation “from the viewpoint of the risk of failure, company x is comprehensively preferred to company y ”, using binary relations on the debt ratio criterion, like “the debt ratio of x is *much better* than that of y ” or “the debt ratio of x is *weakly better* than that of y ”, and so on.

Technically, the modifications of DRSA necessary to approach the problems of choice and ranking are twofold:

1. *Pairwise comparison table* (PCT) is considered instead of the simple data table [19]: PCT is a data table whose rows represent pairs of objects for which multiple criteria evaluations and a comprehensive preference relation are known;
2. *dominance principle* is considered for comparisons of pairs of objects instead of simple objects: if object x is preferred to y at least as strongly as w is preferred to z on all the considered criteria, then x must be comprehensively preferred to y at least as strongly as w is comprehensively preferred to z .

The application of DRSA to the choice or ranking problems proceeds as follows. First, the DM gives examples of comparison of pairs

of some reference objects, well known to the DM. This can be a complete ranking from the best to the worst of a limited number of such objects. From this set of examples, a preference model in terms of “*if . . . , then . . .*” decision rules is induced. These rules are applied to a larger set of objects. In result, one gets a preference relation in this set of objects. A proper exploitation of this preference relation leads to a final recommendation for the decision problem at hand. For more details about application of DRSA to the choice and ranking problems, see [30].

OTHER DEVELOPMENTS FOR DRSA

Other interesting fields of research and applications involving DRSA are discussed further.

DRSA with a Joint Consideration of Dominance, Indiscernibility, and Similarity Relations

Very often, in data tables describing realistic decision problems, there are data referring to a preference order (criteria), for which a dominance relation has to be considered, together with data not referring to any specific preference order (attributes). The latter can be data referring to nominal (symbolic) description, for which an indiscernibility relation has to be considered, and data referring to quantitative (numeric) description, for which a similarity relation has to be considered in the set of objects. Let us remark that indiscernibility is the binary relation considered within the CRSA, while an extension of rough sets to the similarity relation has been proposed by Słowiński and Vanderpooten [31, 32] (for a fuzzy extension of this approach see [33, 34]). Greco *et al.* [20, 35] proposed an extension of DRSA to deal with data table where preference, indiscernibility, and similarity are to be considered jointly.

DRSA and Interval Orders

Owing to imprecise measurement, random variation of some parameters, unstable perception, or incomplete definition of both evaluation scales of criteria and decision classes, some evaluations on particular

criteria and/or some assignments to decision classes considered in a data table may not be univocal. DRSA adapted to this situation has been proposed in [36]. Its basic idea consists in approximation of an interval order of the comprehensive evaluation by means of interval orders on particular criteria; the key concept of this approximation is a specially defined dominance relation.

Fuzzy DRSA—Rough Approximations by Means of Fuzzy Dominance Relations

The concept of dominance can be refined by introducing gradedness through the use of fuzzy sets in the sense of semantics expressing preferences for pairs of objects (for a detailed presentation of fuzzy preferences see [37]). The gradedness introduced by the use of fuzzy sets refines the classic crisp preference structures. Such a reasoning about a smooth transition from truth to falsity of an inclusion relation can be applied to the dominance relation considered in the rough approximations. In [18,38,39] and in [40], DRSA was extended by using fuzzy dominance relation in two different ways. These extensions of the rough approximation into the fuzzy context maintain the same desirable properties of the crisp rough approximation of preference-ordered decision classes. They follow the traditional way of using fuzzy logical connectives in definitions of lower and upper approximations. In fact, however, there is no rule for the choice of the “right” connective; so, this choice is always arbitrary to a certain extent. For this reason, in [41–43], a new fuzzy rough approximation was proposed. It avoids the use of fuzzy connectives, such as T -norm, T -conorm, and fuzzy implication, which extend “and”, “or”, and “if . . . , then . . .” operators within fuzzy logic (see, e.g., [37]), at the price of introducing a certain degree of subjectivity related to the choice of one or another of their functional forms. The proposed approach solves this problem because it is based on the ordinal properties of fuzzy membership functions only. It employs DRSA to cope in a natural way with the monotonic relationship between grades of membership to the concepts (evaluations and classes) on condition and decision sides of objects’ description.

DRSA with Missing Values—Multiple Criteria Classification Problem with Missing Values

In practical applications, the data table is often incomplete because some data are missing. An extension of DRSA enabling the analysis of incomplete data tables has been proposed in [44] and [45]. In this extension, it is assumed that the dominance relation between two objects is a directional statement, where a subject object is compared to a referent object having no missing values on considered criteria. The advantage of this approach is that the rules induced from the rough approximations defined using the extended dominance relation are *robust*, that is, each rule is supported by at least one object with no missing value on the criteria represented in the condition part of the rule. To better understand this feature, let us mention that, in another approach proposed to deal with missing values [46, 47], an object having a missing value is replaced by a set of objects obtained by putting each possible evaluation in the place of the missing value.

DRSA extended to deal with missing values maintains all good characteristics of the dominance-based rough set approach and boils down to the latter when there are no missing values. This approach can also be used to analyze a data table in which dominance, similarity, and indiscernibility must be considered jointly with respect to criteria and attributes.

DRSA for Hierarchical Structure of Attributes and Criteria

In many real life situations, the process of decision-making can be decomposed into sub-problems; this decomposition may either follow from a natural hierarchical structure of the evaluation or from a need of simplification of a complex decision problem. These situations are referred to hierarchical decision problems. The structure of a hierarchical decision problem has the form of a tree whose nodes are attributes and criteria describing the objects. In [48], hierarchical decision problems are considered where the decision is made in a finite number of steps because of a hierarchical structure

of regular attributes and criteria. The proposed methodology is based on decision rule preference model induced from examples of hierarchical decisions made by the DM on a set of reference objects. To deal with inconsistencies appearing in decision examples, DRSA has been adapted to hierarchical classification problems. In these problems, the main difficulty consists in propagation of inconsistencies along the tree, that is, taking into account at each node of the tree the inconsistent information coming from lower level nodes. In the proposed methodology, the inconsistencies are propagated from the bottom to the top of the tree in the form of subsets of possible attribute values instead of single values. In the case of hierarchical criteria, these subsets are intervals of possible criterion values. Subsets of possible values may also appear in leafs of the tree, that is, in evaluations of objects by the lowest level attributes and criteria. To deal with multiple values of attributes in description of objects, the CRSA has been adapted adequately. Interval evaluations of objects on particular criteria can be handled by DRSA extended to interval orders [48].

DRSA AND OPERATIONS RESEARCH PROBLEMS

DRSA is also a useful instrument in the toolbox of operations research (OR). DRSA has been applied to the following OR problems:

1. interactive multiobjective optimization (IMO-DRSA) [11];
2. interactive evolutionary multiobjective optimization under risk and uncertainty [49];
3. decision under uncertainty and time preference [50].

DRSA to Interactive Multiobjective Optimization (IMO-DRSA)

IMO-DRSA [11] permits to deal with many optimization problems considered within OR (ranging from inventory management to scheduling, passing through portfolio management) in a way that is very much oriented toward interaction with the users. In fact,

in IMO-DRSA, a sample of representative solutions to a multiobjective optimization problem is presented to the DM who is asked to indicate a subset of relatively “good” solutions in the sample. Applying DRSA to the sample of representative solutions classified into “good” and “others” by the DM, a set of decision rules is induced in the form: “if objective $f_{i_1}(x) \geq \alpha_{i_1}$ and ... objective $f_{i_p}(x) \geq \alpha_{i_p}$, then x is a *good* solution”. The DM selects the rule that in his/her opinion is the most representative of his/her preferences, and the constraints coming from that rule are joined to the set of constraints imposed on the Pareto optimal set, in order to focus on a part interesting from the point of view of DM’s preferences in the next iteration. For example, if the DM selects the rule “if objective $f_{i_1}(x) \geq \alpha_{i_1}$ and ... objective $f_{i_p}(x) \geq \alpha_{i_p}$, then x is a *good* solution”, then the constraints $f_{i_1}(x) \geq \alpha_{i_1}$ and ... $f_{i_p}(x) \geq \alpha_{i_p}$ are joined to the set of constraints of the multiobjective optimization problem, such that the new set of constraints implies a Pareto optimal set being the subset of the original Pareto optimal set. This subset satisfies the requirements of the selected rule, so that it is composed of solutions that are at this stage considered as relatively good by the DM. The procedure continues iteratively until the DM is satisfied with one solution from the current sample—this is the most preferred solution.

DRSA to Interactive Evolutionary Multiobjective Optimization (EMO-DRSA)

Very often, real life optimization problems are so complex that exact methods fail to find an optimal solution. In these cases, some heuristics are to be applied. Within multiobjective optimization, EMO appeared to be particularly efficient; see, for example, [51, 52]. The underlying reasoning behind the EMO search of an approximation of the Pareto optimal frontier is that, in the absence of any preference information, all Pareto optimal solutions have to be considered equivalent. On the other hand, if the DM (alternatively called *user*) is involved in the multiobjective optimization process, then the preference information provided

by the DM can be used to focus the search on the most preferred part of the Pareto optimal frontier. This idea stands behind IMO methods proposed long time before EMO has emerged. Recently, it became clear that merging the IMO and EMO methodologies should be beneficial for the multiobjective optimization process [53]. Several approaches have been presented in this context; see, for example, [51, 54–61]. The methodology of interactive EMO based on DRSA [49, 62], involves application of decision rules in EMO, which are induced from easily elicited preference information by DRSA, proposing two general schemes, called *DRSA-EMO* and *DRSA-EMO-PCT*. This results in focusing the search of the Pareto optimal frontier on the most preferred region. More specifically, DRSA is used for structuring preference information obtained through interaction with the user, and then, a set of decision rules representing user's preferences is induced from this information. These rules are used to rank solutions in the current population of EMO, which has an impact on the selection and crossover. Within interactive EMO, one can also apply DRSA for decision under uncertainty. This permits to take into account robustness concerns in the multiobjective optimization. In fact, two methods of robust optimization methods combining DRSA and interactive EMO have been proposed: DARWIN [49] (dominance-based rough set approach to handling robust winning solutions in interactive multiobjective optimization) and DARWIN-PCT [62] (DARWIN using pairwise comparison tables). DARWIN and DARWIN-PCT can be considered as two specific instances of *DRSA-EMO* and *DRSA-EMO-PCT*, respectively.

DRSA to Decision Under Uncertainty and Time Preference

DRSA can also be applied to preference modeling for decision under uncertainty with consequences distributed over time, using the idea of time-stochastic dominance, that is, putting together the concept of time dominance and stochastic dominance [50]. Preference information provided by the DM is a set of decision examples specifying the quality of

some chosen acts, that is, assigning these acts to preference-ordered classes. The resulting preference model expressed in terms of “*if... then...*” decision rules is much more intelligible than any utility function. Moreover, it permits to handle inconsistent preference information. Let us observe that the approach handles additive and nonadditive probability distribution and even a qualitative ordinal probability. Furthermore, in case, the elements of sets of possible probability values and of time epochs were very numerous (like in real life applications in which very often they are infinite), it would be enough to consider a subset of the most significant probability values (e.g., 0, 0.1, 0.2, ..., 0.9, 1) and a subset of the most significant epochs (e.g., each month). Applying DRSA to decision under uncertainty and time preference, we get decision rules of the type:

“if the cumulated outcome at t_1 is at least 50 with a probability of 0.5, and the cumulated outcome at t_2 is at least 300 with a probability of 0.7, then act a is (at least) good”,

or

“if the cumulated outcome at t_1 is at most 100 with a probability of 0.25, and the cumulated outcome at t_2 is at most 150 with a probability of 0.8, then act a is at most medium”.

This method can be extended on the case of pairwise comparisons [63], obtaining rules whose syntax is:

“if the difference between the cumulated outcome of act a and act b is not smaller than 50 at t_1 with a probability of 0.6 and not smaller than 300 at t_2 with a probability of 0.4, then act a is at least weakly preferred to act b ”.

The above methodology can be very useful for dealing with many OR problems where uncertainty of outcomes and their distribution over the time play a fundamental role, such as portfolio selection, scheduling with time–resource interactions, and inventory management. Indeed, putting the decision rules produced by this methodology together with IMO-DRSA and DRSA applied to EMO provides an important tool for dealing with even more OR problems. Examples of recent applications of this methodology to typical OR problems, which are inventory

control and portfolio selection, can be found in [64] and [65], respectively.

Acknowledgment

The third author wishes to acknowledge financial support from the Polish National Science Centre, Grant no. DEC-2011/01/B/ST6/07318.

REFERENCES

1. Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis, state-of-the-art surveys. New York: Springer; 2005.
2. Greco S, Matarazzo B, Słowiński R. Rough sets theory for multicriteria decision analysis. *Eur J Oper Res* 2001;129:1–47.
3. Roy B. Decision science or decision aid science? *Eur J Oper Res* 1993;66:184–203.
4. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: John Wiley and Sons; 1976.
5. Dyer JS. MAUT-multiattribute utility theory. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state-of-the-art surveys. Chapter 7. New York: Springer; 2005. p 265–295.
6. Roy B. The outranking approach and the foundations of ELECTRE methods. *Theory Decis* 1991;31:49–73.
7. Figueira J, Mousseau V, Roy B. ELECTRE methods. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state-of-the-art surveys. Chapter 14. New York: Springer; 2005. p 133–162.
8. Jacquet-Lagrèze E, Siskos Y. Assessing a set of additive utility functions for multicriteria decision making: the UTA method. *Eur J Oper Res* 1982;10:151–164.
9. Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *Eur J Oper Res* 2008;191:415–435.
10. Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. *Eur J Oper Res* 2009;195:460–48.
11. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach to interactive multiobjective optimization. In: Branke J, Deb K, Miettinen K, *et al.*, editors. Multiobjective optimization: interactive and evolutionary approaches. LNCS 5252, State-of-the-Art Surveys. Chapter 5. Berlin: Springer; 2008. p 121–155.
12. Pawlak Z. Rough sets. *Int J Inf Comput Sci* 1982;11:341–356.
13. Pawlak Z. Rough sets: theoretical aspects of reasoning about data. Dordrecht: Kluwer; 1991.
14. Polkowski L. Rough sets: mathematical foundations. Heidelberg: Physica; 2002.
15. Słowiński R, editor. Intelligent decision support. Handbook of applications and advances of the rough sets theory. Dordrecht: Kluwer; 1992.
16. Pawlak Z, Słowiński R. Rough set approach to multi-attribute decision analysis. *Eur J Oper Res* 1994;72:443–459.
17. Słowiński R. Rough set learning of preferential attitude in multi-criteria decision making. In: Komorowski J, Ras Z, editors. In: Methodologies for intelligent systems. LNAI 689. Berlin: Springer; 1993. p 642–651.
18. Greco S, Matarazzo B, Słowiński R. The use of rough sets and fuzzy sets in MCDM. In: Gal T, Stewart T, Hanne T, editors. Advances in multiple criteria decision making. Dordrecht: Kluwer; 1998. p 14.1–14.59.
19. Greco S, Matarazzo B, Słowiński R. Rough approximation of a preference relation by dominance relations. *Eur J Oper Res* 1999;117:63–83.
20. Greco S, Matarazzo B, Słowiński R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. *Eur J Oper Res* 2002;138:247–259.
21. Greco S, Matarazzo B, Słowiński R. Decision rule approach. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state-of-the-art surveys. Chapter 13. New York: Springer; 2005. p 507–562.
22. Błaszczyński J, Słowiński R, Szelaż J. Sequential covering rule induction algorithm for variable consistency rough set approaches. *Inf Sci* 2011;181:987–1002.
23. Greco S, Matarazzo B, Słowiński R, Stefanowski J. An algorithm for induction of decision rules consistent with dominance principle. In: Ziarko W, Yao Y, editors. Rough sets and current trends in computing. LNAI 2005. Berlin: Springer; 2001. p 304–313.
24. Słowiński R, Greco S, Matarazzo B. Rough sets in decision making. In: Meyers RA, editor.

- Encyclopedia of complexity and systems science. New York: Springer; 2009. p 7753–7786.
25. Słowiński R, Greco S, Matarazzo B. Rough set and rule-based multicriteria decision aiding. *Pesqui Oper* 2012;32:213–269.
 26. Słowiński R, Greco S, Matarazzo B. Rough-set-based decision support. In: Burke EK, Kendall G, editors. *Search methodologies: introductory tutorials in optimization and decision support techniques*. 2nd ed. Chapter 19. New York: Springer; 2014. p 557–609.
 27. Greco S, Matarazzo B, Słowiński R. A new rough set approach to evaluation of bankruptcy risk. In: Zopounidis C, editor. *Operational tools in the management of financial risks*. Dordrecht: Kluwer; 1998. p 121–136.
 28. Greco S, Matarazzo B, Słowiński R. Multicriteria classification. In: Kloesgen W, Zytkow J, editors. *Handbook of data mining and knowledge discovery*. Chapter 16.1.9. New York: Oxford University Press; 2002. p 318–328.
 29. Słowiński R, Zopounidis C. Application of the rough set approach to evaluation of bankruptcy risk. *Intell Syst Account Finance Manage* 1995;4:27–41.
 30. (a) Szeląg M, Greco S, Słowiński R. Rule-based approach to multicriteria ranking. In: Doumpos M, Grigoroudis E, editors. *Multicriteria decision aid and artificial intelligence: links, theory and applications*. Chapter 6. London: Wiley-Blackwell; 2013. p 127–160.
(b) Szeląg M, Greco S, Słowiński R. Variable consistency dominance-based rough set approach to preference learning in multicriteria ranking. *Information Sciences* 2014; 277:525–552.
 31. Słowiński R, Vanderpooten D. Similarity relation as a basis for rough approximations. In: Wang PP, editor. *Advances in machine intelligence & soft-computing*. Volume IV. Durham (NC): Duke University Press; 1997. p 17–33.
 32. Słowiński R, Vanderpooten D. A generalized definition of rough approximations based on similarity. *IEEE Trans Data Knowl Eng* 2000;12:331–336.
 33. Greco S, Matarazzo B, Słowiński R. Fuzzy similarity relation as a basis for rough approximations. In: Polkowski L, Skowron A, editors. *Rough sets and current trends in computing*. Berlin: Springer; 1998. p 283–289.
 34. Greco S, Matarazzo B, Słowiński R. Rough set processing of vague information using fuzzy similarity relations. In: Calude CS, Paun G, editors. *Finite versus infinite – contributions to an eternal dilemma*. London: Springer; 2000. p 149–173.
 35. Greco S, Matarazzo B, Słowiński R. A new rough set approach to multicriteria and multiattribute classification. In: Polkowski L, Skowron A, editors. *Rough sets and current trends in computing*. Berlin: Springer; 1998. p 60–67.
 36. Dembczyński KS, Greco S, Słowiński R. Rough set approach to multiple criteria classification with imprecise evaluations and assignments. *Eur J Oper Res* 2009;198:626–636.
 37. Fodor J, Roubens M. *Fuzzy preference modelling and multicriteria decision support*. Dordrecht: Kluwer; 1994.
 38. Greco S, Matarazzo B, Słowiński R. Fuzzy dominance-based rough set approach. In: Masulli F, Parenti R, Pasi G, editors. *Advances in fuzzy systems and intelligent technologies*. Maastricht (NL): Shaker Publishing; 2000.
 39. Greco S, Matarazzo B, Słowiński R. Fuzzy extension of the rough set approach to multicriteria and multiattribute sorting. In: Fodor J, De Baets B, Perny P, editors. *Preferences and decisions under incomplete knowledge*. Heidelberg: Physica; 2000. p 131–151.
 40. Greco S, Inuiguchi M, Słowiński R. Dominance-based rough set approach using possibility and necessity measures. In: Alpigini JJ, Peters JF, Skowron A, *et al.*, editors. *Rough sets and current trends in computing*. Berlin: Springer; 2002. p 85–92.
 41. Greco S, Inuiguchi M, Słowiński R. A new proposal for fuzzy rough approximations and gradual decision rule representation. *Transactions on rough sets II*. LNCS 3135. Berlin: Springer; 2004. p 319–342.
 42. Greco S, Inuiguchi M, Słowiński R. Fuzzy rough sets and multiple-premise gradual decision rules. *Int J Approx Reason* 2006;41:179–211.
 43. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach as a proper way of handling graduality in rough set theory. *Transactions on rough sets VII*. LNCS 4400. Berlin: Springer; 2007. p 36–52.
 44. Greco S, Matarazzo B, Słowiński R. Handling missing values in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zhong N, Skowron A, Ohsuga S, editors. *New directions in rough sets, data mining and granular-soft computing*. Berlin: Springer; 1999. p 146–157.

45. Greco S, Matarazzo B, Słowiński R. Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zanakakis SH, Doukidis G, Zopounidis C, editors. Decision making: recent developments and worldwide applications. Dordrecht: Kluwer; 2000. p 295–316.
46. Kryszkiewicz M. Rough set approach to incomplete information systems. *Inf Sci* 1998;112:39–49.
47. Kryszkiewicz M. Rules in incomplete information systems. *Inf Sci* 1999;113:271–292.
48. (a) Dembczyński KS, Greco S, Słowiński R. Methodology of rough-set-based classification and sorting with hierarchical structure of attributes and criteria. *Control Cybern* 2002;31:891–920. (b) Dembczyński K, Greco S, Słowiński R. Rough set approach to multiple criteria classification with imprecise evaluations and assignments. *European J. Operational Research* 2009;198:626–636.
49. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach to interactive evolutionary multiobjective optimization. In: Greco S, Marques Pereira RA, Squillante M, *et al.*, editors. Preferences and decisions: models and applications. Studies in fuzziness 257. Berlin: Springer; 2010. p 225–260.
50. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach to decision under uncertainty and time preference. *Ann Oper Res* 2010;176:41–75.
51. Coello Coello CA, Van Veldhuizen DA, Lamont GB. Evolutionary algorithms for solving multi-objective problems. New York: Kluwer; 2002.
52. Deb K. Multi-objective optimization using evolutionary algorithms. Chichester: John Wiley and Sons; 2001.
53. Branke J, Deb K, Miettinen K, *et al.*, editors. Multiobjective optimization: interactive and evolutionary approaches. LNCS 5252, state-of-the-art surveys. Berlin: Springer; 2008.
54. Fonseca CM, Fleming PJ. Genetic algorithms for multiobjective optimization: formulation, discussion, and generalization. Proceedings of the 5th International Conference on Genetic Algorithms; 1993. p 416–423.
55. Deb K, Sundar J, Rao N, *et al.* Reference point based multi-objective optimization using evolutionary algorithms. *Int J Comput Intell Res* 2006;2(3):273–286.
56. Deb K, Chaudhuri S. I-MODE: an interactive multi-objective optimization and decision-making using evolutionary methods. *Appl Soft Comput* 2010;10:496–511.
57. Branke J, Kausler T, Schmeck H. Guidance in evolutionary multi-objective optimization. *Adv Eng Softw* 2001;32:499–507.
58. Greenwood GW, Hu XS, D'Ambrosio JG. Fitness functions for multiple objective optimization problems: combining preferences with Pareto rankings. In: Belew RK, Vose MD, editors. Foundations of genetic algorithms. San Mateo, CA: Morgan Kaufmann; 1997. p 437–455.
59. Jaskiewicz A. Interactive multiobjective optimization with the Pareto memetic algorithm. *Found Comput Decis Sci* 2007;32:15–32.
60. Phelps S, Köksalan M. An interactive evolutionary metaheuristic for multiobjective combinatorial optimization. *Manage Sci* 2003;49(12):1726–1738.
61. Branke J, Greco S, Słowiński R, *et al.* Interactive evolutionary multiobjective optimization using robust ordinal regression. In: Ehrgott M, Fonseca CM, Gandibleux X, *et al.*, editors. Evolutionary multi-criterion optimization (EMO'09). LNCS 5467. Berlin: Springer; 2009. p 554–568.
62. Greco S, Matarazzo B, Słowiński R. Interactive evolutionary multiobjective optimization using dominance-based rough set approach. Proceedings of WCCI 2010, IEEE World Congress on Computational Intelligence; 2010 July 18–23; Barcelona, Spain; p 3026–3033.
63. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach to preference learning from pairwise comparisons in case of decision under uncertainty. In: Hüllermeier E, Kruse R, Hoffmann F, editors. Computational intelligence for knowledge-based systems design, 13th International Conference on Information Processing and Management of Uncertainty. IPMU 2010. LNCS 6178. Berlin: Springer; 2010. p 584–594.
64. Greco S, Matarazzo B, Słowiński R, Vaccarella G. Inventory control using interactive multi-objective optimization guided by dominance-based rough set approach; 2012. Submitted.
65. Greco S, Matarazzo B, Słowiński R. Beyond markowitz with multiple criteria decision aiding. *J Bus Econ* 2013;83(1):29–60.

DECISION-THEORETIC FOUNDATIONS OF SIMULATION OPTIMIZATION

PETER I. FRAZIER
School of Operations Research
and Information Engineering,
Cornell University, Ithaca,
New York

Simulation is an invaluable tool in operations research (OR) and management science (MS). We may use it whenever we wish to manipulate a system to see what would happen, but actually performing the manipulation in reality would be too expensive, too risky, or simply impossible. Instead of manipulating the real system, we can build a simulation of the system and manipulate that instead.

In OR and MS, we often focus on stochastic simulation, and, in particular, on stochastic discrete-event simulation [1]. Important examples of systems simulated in this way include supply chains, queueing systems, manufacturing systems, and financial markets. The mathematical techniques that we will describe here may also be applied to simulations of systems of interest in engineering, medicine, and the natural sciences, for example, simulations of the physical and chemical processes in a combustion engine.

Often, our goal in building a simulation is to optimize the real system. To do so, we may test a variety of possible actions using the simulation, find the one that performs best in simulation, and implement it in the real system. For example, we may simulate inventory policies in a supply chain to minimize expected costs, or we may simulate methods for repositioning ambulances in a city to maximize the number of calls answered in a timely manner. Finding the inputs to a simulation that maximize some aspect of its output is called *simulation optimization*.

A related goal is *calibration* of a simulated system. In calibration, we search for parameter settings that cause the simulation to behave as similarly as possible to the real

system. For example, we may search for a parametric model of credit risk in a financial market that causes predicted prices of liquidly traded securities to match market prices, or we may search for a representation of consumer behavior that cause simulated sales data to match what is observed in practice. Because we are searching for simulation inputs that cause its output to match a target as well as possible, calibration is also a simulation optimization problem.

In general, any simulation optimization problem may be written as

$$\max_{x \in \mathcal{X}} \mathbb{E}[f(x, W)], \quad (1)$$

where f is a function implemented by the logic of the simulation (often highly complicated), W is random variable (often highly multivariate) generated by the random number generator inside the simulation, and $\mathcal{X}$ is the set of possible inputs to the simulation. In this article, we will consider two cases. In the first case, called *ranking and selection* (R&S), $\mathcal{X}$ a finite set. In the second case, called *Bayesian Global Optimization* (BGO) $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $x \mapsto \mathbb{E}[f(x, W)]$ is a continuous function.

Simulation optimization differs from other types of optimization in that it does not rely on an analytic expression for $\mathbb{E}[f(x, W)]$. This makes simulation optimization more flexible than other types of optimization, but also more challenging. In most simulations, f is very complicated, and W is highly multivariate, and so an analytic expression is simply not available. Instead, the only way to estimate $\mathbb{E}[f(x, W)]$ is to run the simulator many times with the same input x , but different randomly generated values of W , say $W_1, \dots, W_n$, and average the outputs $f(x, W_1), \dots, f(x, W_n)$. Accurately estimating $\mathbb{E}[f(x, W)]$ for even a single value of x may require many simulations and a great deal of computer time.

The goal of the field of simulation optimization is to design algorithms that will solve Equation (1) as quickly and accurately

as possible. Simulations of highly complex systems often require significant amounts of time to generate samples of $f(x, W)$, and so an algorithm that solves the problem with fewer samples can save a significant amount of time. In situations with large search spaces $\mathcal{X}$ or very time-consuming simulation, an efficient algorithm may make a solution to Equation (1) available, where otherwise a solution would be impossible to find. For example, if a large discrete-event simulation takes 30 min of computer time to generate a single sample, then an algorithm that solves the problem to reasonable accuracy with 100 samples could be used in practice, while an algorithm that requires 10^6 samples would not (the first requires 50 h of computer time, while the second requires more than 50 years).

At its heart, an algorithm is a sequence of decisions—decisions about which points x to sample with the simulation (the “allocation or measurement decisions”), a decision of when to stop sampling (the “stopping decision”), and at the end, a decision of which point x^* to declare as an approximate solution (the “selection decision”). Thus, decision theory provides a framework in which to assess and compare the quality of different algorithms, and ultimately to guide us toward an optimal algorithm. In particular, since simulation optimization is concerned with the collection of information, it may be understood as a problem in the value of information [2,3].

A number of reviews have been written that consider decision theory applied to the decisions made in simulation, and many of these consider simulation optimization. Many of these reviews [4,5], introduce decision theory to readers already familiar with simulation, while Merrick [6] introduces problems in simulation to readers already familiar with decision theory. Also, we do not consider here the larger field of simulation optimization separated from its decision-theoretic aspects. For such broader discussion, see Refs 7–10.

In this article, we begin by describing the R&S problem in a Bayesian context. This problem is canonical, being simple enough to allow a thorough analysis, while at the same time exemplifying the issues at the

core of all simulation optimization problems. We then discuss a worst-case analysis of R&S called the *indifference-zone formulation*, which is historically more prominent than the Bayesian formulation. We conclude by discussing the BGO problem, in which we optimize a continuous function over a continuous domain and must deal with an infinite number of alternatives.

BAYESIAN RANKING AND SELECTION

The decision-theoretic understanding of simulation optimization is best introduced in the context of the R&S problem. In this problem, a collection of distinct “alternatives,” numbered 1 through k , comprises our set $\mathcal{X}$ of possible inputs to the simulation. We might think of the alternatives as possible configurations of a manufacturing process, or as inventory policies that might be used in a supply chain.

Associated with each alternative x is a sampling distribution, which is the distribution of $f(x, W)$. In R&S, our goal is to maximize $\mathbb{E}[f(x, W)]$, that is, to find the alternative whose sampling distribution has the largest mean. We often assume that the sampling distributions are normal, which is justified by supposing that, when samples are not normal, we average many together to obtain approximately normal batch averages. When the sampling distribution is normal, it is completely determined by its mean and variance. Let θ_x and λ_x be the mean and variance, respectively, of the sampling distribution for alternative x . Let $\theta = (\theta_1, \dots, \theta_k)$ and $\lambda = (\lambda_1, \dots, \lambda_k)$. These values are unknown *a priori*, and simulation is the tool through which we may learn about them.

Whenever we choose to simulate an alternative x , we observe a sample from this sampling distribution. In R&S, we often assume that samples are independent over time and across the alternatives. Let n index the number of samples observed, and let x_n be the alternative that is observed at time n . We may choose x_n based on previous samples, and it is this choice or decision, along with the stopping decision, that is the main focus within R&S. For each $n \geq 1$ we observe a

value y_n that results from sampling alternative x_n . It has a distribution

$$y_n \sim \text{Normal}(\theta_{x_n}, \lambda_{x_n}) \mid x_n, \theta, \lambda,$$

and is independent of all previous allocation decisions and samples, that is, of x_m and y_m , $m < n$.

By taking more samples, we learn more about the sampling distributions and we improve our ability to find the best one. However, taking more samples consumes more time. In some formulations of the R&S problem, the amount of time that we may spend sampling is constrained. In others, we control how much time we spend sampling, but sampling incurs an explicit cost that we must balance against the value of the information obtained. In this section we will suppose an explicit cost $c > 0$ for each sample taken.

Let τ be the number of samples taken. As with the decision of which alternatives to sample, the decision to stop taking samples may be made adaptively, based on the samples observed so far. After we stop sampling, we select an alternative for implementation based on the accumulated evidence. Call the alternative selected x^* . Our selection decision may depend on $x_1, \dots, x_\tau$ and $y_1, \dots, y_\tau$. As stated, we wish to select the alternative with the largest sampling mean, $\arg \max_x \theta_x$. When we make a mistake and we fail to select the one with the largest θ_x , we suffer a loss $L(x^*, \theta)$, where L is our loss function. Common loss functions are the linear loss function $L(x^*, \theta) = \max_x \theta_x - \theta_{x^*}$, and the 0–1 loss function $L(x^*, \theta) = \mathbf{1}_{\{x^* \neq \arg \max_x \theta_x\}}$. One could also consider a loss function depending on both θ and λ , although commonly this is not considered. In addition to this loss associated with an incorrect selection, we also pay the cost of sampling, which is $c\tau$. Thus, the total loss is $L(x^*, \theta) + c\tau$.

Let π be a generic set of rules for making the decisions τ , $(x_n)_{n \leq \tau}$, and x^* . We call this set of rules a *policy*. It must obey the previously stated rules governing the information upon which decisions may be based. These rules require that the decision $\mathbf{1}_{\{\tau=n\}}$ to stop and the decision x_{n+1} of which alternative to sample next (when not stopping) depend only upon $x_{1:n} = (x_1, \dots, x_n)$ and $y_{1:n} = (y_1, \dots, y_n)$.

They also require that the selection decision x^* depend only upon $x_{1:\tau}$ and $y_{1:\tau}$. Given a policy π obeying these rules, and fixed values θ and λ , the expected total loss suffered is

$$R^\pi(\theta, \lambda) = \mathbb{E}^\pi [L(x^*, \theta) + c\tau \mid \theta, \lambda],$$

where $\mathbb{E}^\pi$ indicates that when the expectation is taken, the values τ , $(x_n)_{n \leq \tau}$, and x^* are chosen according to the policy π .

If we knew the true θ and λ , then we would prefer the policy π with the smallest value of $R^\pi(\theta, \lambda)$. However, we do not know these values at all. If we did, we could simply choose $x^* \in \arg \max_x \theta_x$ directly without sampling at all, and achieve 0 loss every time. Since we do not know θ and λ , we instead need a method for combining performance across multiple values of θ and λ to establish a preference order over policies.

In the Bayesian formulation, we place a prior probability distribution on the unknown quantities, θ and λ . In some cases, this is justified because the simulation optimization problem at hand really is drawn at random from some larger set of simulation optimization problems with which we have a great deal of experience and whose distribution we can characterize. One such case might be the calibration of a financial model of credit risk based on current market prices for liquidly traded credit derivatives. The model must be calibrated each day to a new set of market prices, and because the calibration is performed repeatedly, we may estimate the overall distribution of truths θ and λ in the simulation optimization problems encountered, and use this as our prior. Much more frequently, however, the prior probability distribution encodes our subjective belief about θ and λ . This is in keeping with Bayesian methods as used throughout decision theory [11]. If one has very little information concerning θ and λ , it is also possible to specify a noninformative prior distribution.

Given a prior distribution, the Bayes risk (expected total loss under the prior) of a policy π is

$$\mathbb{E} [R^\pi(\theta, \lambda)] = \mathbb{E}^\pi [L(x^*, \theta) + c\tau],$$

where the first expectation is over the truth θ , λ as drawn from the prior, and the second expectation is over both θ , λ , and the intrinsic randomness introduced by the sampling noise. This first expectation lacks a π in the superscript because it is taken only over the prior on θ , λ , which does not depend upon the policy (however, note that $R^\pi(\theta, \lambda)$ does depend on π), while the second expectation has a π in the superscript because the distribution of x^* and τ depends upon the policy.

If a noninformative prior distribution is specified, then the expectation $\mathbb{E}^\pi [L(x^*, \theta) + c\tau]$ may not exist. In such situations, one may perform a first stage of measurements in which each alternative is measured an equal number of times to produce a proper posterior belief. One may then take this resulting belief and use it as the prior under which we formulate the decision-theoretic problem.

Given this prior, whether it is the original prior or the belief resulting from an earlier noninformative prior and a first stage of measurements, we value a policy according to its Bayes risk. One policy is better than the other if it has smaller Bayes risk, and an optimal policy π is one that attains minimal Bayes risk

$$\inf_{\pi \in \Pi} \mathbb{E}^\pi [L(x^*, \theta) + c\tau]. \quad (2)$$

Here, Π is the set of policies considered, and may be the set of all policies, or may be a strict subset, for example, the set of two-stage policies described in the section titled “Allocation and Stopping Rules.” Finding a policy π that attains this infimum is one of the ultimate goals in Bayesian R&S.

Beginning with a prior probability distribution, sampling results in a posterior distribution that combines the prior with evidence from sampling. This posterior may be calculated using Bayes’ rule. Given a set of samples and the resulting posterior distribution, the conditional expected loss of a given selection decision x^* at time τ is

$$\mathbb{E} [L(x^*, \theta) | x_{1:\tau}, y_{1:\tau}].$$

One can show that the best implementation decision x^* is any from the set $\arg \min_x \mathbb{E} [L(x, \theta) | x_{1:\tau}, y_{1:\tau}]$. For the linear loss function, this is the one with the largest posterior mean, but for other loss functions, including the 0–1 loss function, this is not necessarily true [12]. Nevertheless, given a posterior distribution obtained by a given policy, choosing x^* is relatively straightforward.

If we fix the rule for choosing x^* to the optimal method, we may rewrite Equation (2) as

$$\inf_{\pi \in \Pi} \mathbb{E}^\pi \left[\min_{x^*} \mathbb{E} [L(x^*, \theta) | x_{1:\tau}, y_{1:\tau}] + c\tau \right]. \quad (3)$$

Within this expression, $\min_{x^*} \mathbb{E} [L(x^*, \theta) | x_{1:\tau}, y_{1:\tau}]$ is the best that we can do given the information we have collected by time τ , $c\tau$ is the cost we pay for this information, and $\mathbb{E}^\pi [\min_{x^*} \mathbb{E} [L(x^*, \theta) | x_{1:\tau}, y_{1:\tau}] + c\tau]$ is the overall expected net loss incurred by sampling according to the policy π . By subtracting this from the value of taking no samples at all, one obtains $\min_{x'} \mathbb{E} [L(x', \theta)] - \mathbb{E}^\pi [\min_{x^*} \mathbb{E} [L(x^*, \theta) | x_{1:\tau}, y_{1:\tau}] + c\tau]$, which is the net value of the information collected by π .

The challenge in designing algorithms for Bayesian R&S is choosing rules for making allocation and stopping decisions whose expected loss is as close to this minimal value (Eq. 3) as possible. In theory, such rules may be found using dynamic programming (see Refs 13 and 14 for dynamic-programming formulations of R&S), but the amount of computational power that is actually required to find them far exceeds what is currently available. Instead, a number of heuristic rules have been suggested in the literature. We discuss these after introducing the most commonly used prior distributions.

Choice of Prior Probability Distribution

Bayesian R&S is most convenient if the prior resides within a conjugate family for the sampling distribution. The most commonly considered case is

$$\begin{aligned} \theta_x | \lambda_x &\sim \text{Normal}(\mu_{0x}, \lambda_x/t_{0x}), \\ \lambda_x^{-1} &\sim \text{Gamma}(a_{0x}, b_{0x}), \end{aligned}$$

which is parameterized by μ_{0x} , t_{0x} , a_{0x} , b_{0x} , $x = 1, \dots, k$. With this prior and a sequence of samples, the posterior is

$$\begin{aligned}\theta_x \mid \lambda_x, x_{1:n}, y_{1:n} &\sim \text{Normal}(\mu_{nx}, \lambda_x/t_{nx}), \\ \lambda_x^{-1} \mid x_{1:n}, y_{1:n} &\sim \text{Gamma}(a_{nx}, b_{nx}),\end{aligned}$$

where the parameters for the unmeasured alternatives remain unchanged from n to $n + 1$, and the parameters for the measured alternative, $x = x_{n+1}$, can be calculated recursively as

$$\begin{aligned}\mu_{n+1,x} &= (t_{nx}\mu_{nx} + y_{n+1}) / (t_{nx} + 1), \\ t_{n+1,x} &= t_{nx} + 1, \\ b_{n+1,x} &= (y_{n+1} - \mu_{nx})^2 t_{nx} / (2t_{nx} + 2), \\ a_{n+1,x} &= a_{nx} + \frac{1}{2}.\end{aligned}$$

Details, derivations, and the limiting non-informative case may be found in Ref. 15, Sections 9.6 and 10.2.

If λ is known, then the prior-posterior analysis simplifies. The priors and posteriors, for $n \geq 0$, are $\theta_x \sim \text{Normal}(\mu_{nx}, \lambda_x/t_{nx})$, and the parameters of the posterior are updated recursively for $x = x_{n+1}$ as $\mu_{n+1,x} = (t_{nx}\mu_{nx} + y_{n+1}) / (t_{nx} + 1)$, and $t_{n+1,x} = t_{nx} + 1$.

In both cases, as we continue to sample an alternative x , our posterior belief about its sampling mean concentrates around its true sampling mean θ_x . The mean of our belief, μ_{nx} , converges to θ_x , and the variance of our belief shrinks to 0.

Allocation and Stopping Rules

In the framework discussed so far, the decisions made at time $n + 1$ (the measurement decisions x_{n+1} and the decision of whether to stop or not, that is, whether to set $\tau = n + 1$) may be made based upon *all* of the information $x_{1:n}, y_{1:n}$ available.

In some situations, it may be useful to restrict the set of information upon which a decision may be based. This is done for two reasons. The first is to run simulations in parallel—if we have multiple processors, we would like to start several simulations at once without waiting for the results of the

last simulation to decide which alternative to simulate next. The second reason is simplicity of analysis. Complex adaptive policies are usually more difficult to analyze and develop than those with more restrictive information dependencies.

Policies are classified according to the complexity of their information dependence. The simplest policies are those in which the alternatives to sample, $x_{1:\tau}$ are chosen up front. In this case, the x_n and τ are deterministic quantities. These are called *one-shot*, *single-stage*, or *deterministic* policies. One can often obtain better performance than with one-shot policies by taking a fixed number of samples from each alternative in what is called a *first stage*, and then allocating a second stage of measurements based on the results. These are called *two-stage policies*. Three-stage and multistage policies are defined similarly. Finally, a policy that uses the results from all previous samples to make each measurement decision is called *fully sequential*. As we increase the number of stages and decrease the number of measurements per stage down to the limit established by the number of parallel processors, we allow greater efficiency but must contend with more complicated policies.

In the special case of $k = 2$ alternatives, the optimal one-shot policy was found in Ref. 16. This one-shot policy was shown to be optimal among the larger class of fully sequential policies with fixed measurement horizons in Ref. 14. If the prior variances and measurement variances are equal, this optimal policy is the one that allocates an equal number of measurements to each alternative. For more than two alternatives, however, finding optimal single-stage allocation policies is nontrivial, and finding optimal multistage and fully sequential policies is even more difficult.

We briefly review heuristic policies developed within the Bayesian decision-theoretic framework. Many earlier papers considered allocation rules in isolation, without considering the stopping rule. Using a decision-theoretic analysis and an asymptotic approximation, Chick and Inoue [17] introduce the LL and 0–1 allocation rules (for the linear loss and 0–1 loss functions

respectively), in both one-shot, two-stage, and multistage forms. In a different line of research motivated by Bayesian methods, a number of allocation rules have been developed under the collective name of optimal computing budget allocation (OCBA) in Refs 18–20 and other works. A fully sequential allocation rule derived using a one-step heuristic in a decision-theoretic dynamic-programming framework with known sampling variance was introduced in Ref. 16 and analyzed more fully in Ref. 14, where it was noted that the policy was a larger collection of myopic methods called *knowledge-gradient* (KG) methods. This allocation rule was extended to handle unknown sampling variance in Ref. 21 and given the name LL(1). A variety of adaptive stopping rules derived within a Bayesian framework were introduced in Refs 22–24, and Branke *et al.* [22] showed that good adaptive stopping rules perform much better than does stopping at a fixed time. Chick and Gans [13] considered both allocation and stopping rules for a natural extension of the R&S problem that includes discounting in time of the implementation payoff. A large empirical study of many of these policies was conducted in Ref. 25.

OTHER FORMULATIONS

We have focused on R&S for pedagogical reasons—it is simple enough to be easily and thoroughly explained while still exhibiting the core characteristics shared by all simulation optimization problems. We have further focused on Bayesian R&S because modern decision theory tends to be more interested toward Bayesian methods. Despite this focus on Bayesian R&S, there are other rigorous and useful formulations of the simulation optimization problem. Here we briefly describe two of them, the indifference-zone (IZ) formulation of the R&S problem, and BGO.

Indifference-Zone Ranking and Selection

The historical weight of the R&S literature employs a worst-case analysis known as the *indifference-zone* (IZ) formulation. Although

worst-case analyses have well-known drawbacks that have been highlighted within the decision-theoretic literature [11, Section 5.5], the IZ formulation represents an important approach within R&S.

In the IZ formulation of R&S, the decision maker’s goal is to implement the best alternative with a probability that exceeds a lower bound whenever the true configuration falls within a particular set of configurations. In detail, the IZ formulation begins by defining our loss function as

$$L(x^*, \theta) = \mathbf{1}_{\{x^* \in \arg \max_x \theta\}},$$

and the conditional risk of a policy is

$$R^\pi(\theta, \lambda) = \mathbb{P}^\pi \{x^* \in \arg \max_x \theta_x \mid \theta, \lambda\}.$$

This is also known as the *probability of incorrect selection*. The quantity $\text{PCS}^\pi(\theta, \lambda) = 1 - R^\pi(\theta, \lambda)$ is known as the *probability of correct selection*.

Then, rather than valuing a policy π according to its average-case performance under a prior, we specify a fixed parameter $\delta > 0$ and consider the policy’s worst-case performance over the set of configurations $\text{PZ}(\delta)$ defined by

$$\text{PZ}(\delta) = \left\{ \theta \in \mathbb{R}^k : \theta_{[1]} - \theta_{[2]} \geq \delta \right\},$$

where $[1], \dots, [k]$ are a permutation of the indices $1, \dots, k$ so that $\theta_{[1]} \geq \dots \geq \theta_{[k]}$. This set of configurations is called the *preference zone*. In addition to δ , we also specify a level $\alpha \in (0, 1 - 1/k)$, often 0.01 or 0.05, and require that π satisfies

$$\inf_{\theta \in \text{PZ}(\delta), \lambda \in \mathbb{R}_+^k} \text{PCS}^\pi(\theta, \lambda) \geq 1 - \alpha.$$

If a policy satisfies this, then we say that it satisfies the IZ guarantee for α and δ .

Whenever a configuration falls within the preference zone $\text{PZ}(\delta)$, the difference between best and all other alternatives is at least δ , and the decision maker is said to prefer this best alternative. Whenever the configuration falls outside the preference zone, the difference between the best and second best alternatives is small enough that the decision

maker is said to be indifferent to selecting the best. The IZ guarantee then guarantees that the decision maker selects the best with probability at least $1 - \alpha$, unless the configuration is such that he/she is indifferent. A more recent and stringent version of this formulation proposed in Ref. 26 additionally requires that the probability of selecting a “good” alternative (one whose sampling mean is within δ of the best) exceeds $1 - \alpha$ over all configurations $\theta \in \mathbb{R}^k$.

A large literature exists on the IZ formulation of the R&S problem. This literature develops policies that satisfy this IZ guarantee while at the same time taking samples as few as possible. This literature dates to the 1950s and the seminal works of Bechhofer [27,28]. For a more recent monograph, see Ref. 29, and for a current survey of the literature, see Refs 30 and 31.

Relative to Bayesian methods, the theoretical guarantee that IZ procedures offer is often quite appealing. This guarantee comes at a price, however, because IZ methods tend to be quite conservative, often taking many more samples than is strictly required to meet the guarantee for most configurations. This conservatism has two sources. First is the need to protect against extremely unfavorable configurations, even if these are seldom met in practice. This type of conservatism is intrinsic to the worst-case analysis used by the IZ formulation. Second, the existing procedures often over-sample, even under the most unfavorable configurations. This is because current methods for proving the IZ guarantee rely on inequalities that are often quite loose, especially for large numbers of alternatives. Recent efforts including those of Nelson *et al.* [32], Kim and Nelson [33], and Goldsman *et al.* [34] have greatly reduced but not eliminated this second type of conservatism. Thus, if one desires a well-understood statistical guarantee on the performance of the R&S procedure used, and is willing to spend some extra time on sampling, then IZ methods are appropriate. If instead, one desires good performance on average, and is willing to use a procedure whose performance is not as well characterized theoretically, then Bayesian methods are appropriate. For further discussion and extensive numerical

comparisons between Bayesian and IZ procedures, see Ref. 25.

Bayesian Global Optimization

R&S makes no assumptions about the relationship between different alternatives. In that problem, learning the value of one alternative tells us nothing about the value of other alternatives. This is a consequence of the independence of the alternatives’ true sampling means under the prior. In many applications of simulation optimization, however, the alternatives *are* related to each other. For example, in the classic problem of finding the best (s, S) inventory policy [1], two inventory policies with similar values for s and S usually perform similarly.

One can incorporate such information about the relationship between alternatives into the prior. One may replace the independent prior used in R&S with a prior under which the values of the alternatives are correlated. Conceptually, this is an easy extension of the Bayesian formulation of the R&S problem, but it presents many numerical and algorithmic challenges.

When the number of alternatives is finite, as it would be with the inventory policy example, one can use a multivariate normal prior. This case is considered in Ref. 35. When the space of alternatives is a multidimensional continuous space, as it would be when optimizing several continuous parameters, and the true sampling mean is a continuous function of these parameters, one can place a Gaussian process prior [36] on the unknown sampling mean. The resulting problem is known as the *Bayesian Global Optimization problem* (BGO).

In BGO, we have an infinite number of alternatives, but we have a characterization of their relationships to each other that allows efficient search for an optimum. Much of the literature on BGO assumes that the simulation is deterministic, but some of the literature considers stochastic simulation. Deterministic simulation is considered in Refs 37–43, and stochastic simulation in Refs 44 and 45. For broader issues in the use of Bayesian methods for inference on continuous functions produced as output from a simulator, see Refs 46–48.

A broad literature exists on global optimization beyond that, which is based on decision theory. Much of this literature assumes that function evaluations are free from noise—see the recent survey by Floudas and Gounaris [49] and also the recent book by Zhigljavsky and Žilinskas [50], both of which discuss noise-free BGO methods in this broader context. For a discussion of methods designed to handle noise in function evaluations see the surveys [7–10] mentioned in the introduction. Relative to other methods of global optimization, BGO requires more computational effort per measurement to decide which points to sample, but tends to find a good solution with fewer measurements. This is discussed and demonstrated numerically on a few example problems in Ref. 45. Thus, BGO tends to be more useful for optimizing complex long-running simulations, while other techniques are preferable for short-running simulations.

CONCLUSION

Simulation optimization is a difficult and important problem that is ubiquitous in OR, MS, and other fields. Decision theory offers us a way to understand how decisions should be made within simulation optimization algorithms. Better algorithms and deeper understanding of existing algorithms have resulted from the application of decision theory to these problems.

At the same time, many exciting challenges remain. Computing the optimal policy for all but the simplest problems, especially in sequential settings, remains practically impossible. In addition, existing work on simulation optimization has only begun to consider many aspects of the way that simulation optimization problems present themselves in practice, including aspects like risk aversion, the relationship between alternatives, steady-state simulations, common random numbers, and the difference between the simulation's output and the behavior of the real system. By understanding the decision-theoretic foundations of simulation optimization, we may more easily understand the proper response to these novel aspects

of the problem, and more easily design the algorithms of the future.

REFERENCES

1. Law AM, Kelton WD. Simulation modeling and analysis. 3rd ed. New York: McGraw-Hill; 2000.
2. Howard R. Information value theory. *IEEE Trans Syst Sci Cybern* 1966;2(1):22–26.
3. Howard R. Decision analysis: practice and promise. *Manage Sci* 1988;34(6):679–695.
4. Chick S. Bayesian methods: Bayesian methods for simulation. *Proceedings of the 32nd Conference on Winter Simulation*. Orlando (FL): 2000. pp. 109–118.
5. Chick S. Bayesian ideas and discrete event simulation: why, what and how. *Proceedings of the 37th Conference on Winter Simulation*. Winter Simulation Conference. Monterey (CA): 2006. pp. 96–105.
6. Merrick J. Bayesian simulation and decision analysis: an expository survey. *Decis Anal* 2009;6:222–238.
7. Andradóttir S. Simulation optimization. *Handbook of simulation: principles, methodology, advances, applications, and practice*. New York: Wiley-Interscience; 1998. pp. 307–333.
8. Swisher J, Hyden P, Jacobson S, *et al.* A survey of simulation optimization techniques and procedures. *Simulation Conference Proceedings*, 2000. Winter, 1. Orlando (FL); 2000.
9. Fu M. Optimization for simulation: theory vs. practice. *INFORMS J Comput* 2002;14(3):192–215.
10. Spall J. Introduction to stochastic search and optimization: estimation, simulation, and control. Hoboken (NJ): John Wiley & Sons, Inc.; 2003.
11. Berger JO. Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer; 1985.
12. Berger J, Deely J. A bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology. *J Am Stat Assoc* 1988;83(402):364–373.
13. Chick S, Gans N. Economic analysis of simulation selection problems. *Manage Sci* 2009;55:421–437.
14. Frazier P, Powell WB, Dayanik S. A knowledge gradient policy for sequential information collection. *SIAM J Control Optim* 2008;47(5):2410–2439.

15. DeGroot MH. Optimal Statistical Decisions. New York: John Wiley and Sons; 1970.
16. Gupta S, Miescke K. Bayesian look ahead one-stage sampling allocations for selection of the best population. *J Stat Plan Inference* 1996;54(2):229–244.
17. Chick S, Inoue K. New two-stage and sequential procedures for selecting the best simulated system. *Oper Res* 2001;49(5):732–743.
18. Chen C. A lower bound for the correct subset-selection probability and its application to discrete-event system simulations. *IEEE Trans Automat Control* 1996;41:1227–1231.
19. Chen C, Lin J, Yücesan E, *et al.* Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dyn Syst* 2000;10(3):251–270.
20. He D, Chick S, Chen C. Opportunity cost and ooba selection procedures in ordinal optimization for a fixed number of alternative systems. *IEEE Trans Syst Man Cybern C Appl Rev* 2007;37(5):951–961.
21. Chick S, Branke J, Schmidt C. Sequential sampling to myopically maximize the expected value of information. *INFORMS J Comput* 2010;22:71–80.
22. Branke J, Chick S, Schmidt C. New developments in ranking and selection: an empirical comparison of the three main approaches. In: Kuhl M, Steiger N, Argstrong F, *et al.*, editors. *Proceedings 2005 Winter Simulation Conference*. Piscataway (NJ): IEEE, Inc.; 2005. pp. 708–717.
23. Frazier P, Powell W. The knowledge-gradient stopping rule for ranking and selection. *Winter Simulation Conference Proceedings*. Miami (FL); 2008.
24. Chick S, Frazier P. The conjunction of the knowledge gradient and economic approach to simulation selection. *Winter Simulation Conference Proceedings*. Austin (TX); 2009.
25. Branke J, Chick S, Schmidt C. Selecting a selection procedure. *Manage Sci* 2007;53(12):1916–1932.
26. Nelson B, Banerjee S. Selecting a good system: procedures and inference. *IIE Trans* 2001;33(3):149–166.
27. Bechhofer R. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann Math Stat* 1954;25(1):16–39.
28. Bechhofer R, Kiefer J, Sobel M. Sequential identification and ranking procedures. Chicago: University of Chicago Press; 1968.
29. Bechhofer R, Santner T, Goldsman D. Design and analysis of experiments for statistical selection, screening and multiple comparisons. New York: John Wiley & Sons, Inc.; 1995.
30. Kim S, Nelson B. Selecting the best system. *Handbook in operations research and management science: simulation*. Amsterdam: Elsevier; 2006.
31. Kim S, Nelson B. Recent advances in ranking and selection. *Proceedings of the 39th conference on winter simulation: 40 years! the best is yet to come*. Washington (DC): IEEE Press; 2007. pp. 162–172.
32. Nelson B, Swann J, Goldsman D, *et al.* Simple procedures for selecting the best simulated system when the number of alternatives is large. *Oper Res* 2001;49(6):950–963.
33. Kim S-H, Nelson BL. A fully sequential procedure for indifference-zone selection in simulation. *ACM Trans Model Comput Simul* 2001;11(3):251–273.
34. Goldsman D, Kim S, Marshall W, *et al.* Ranking and selection for steady-state simulation: procedures and perspectives. *INFORMS J Comput* 2002;14(1):2–19.
35. Frazier P, Powell WB, Dayanik S. The knowledge gradient policy for correlated normal beliefs. *INFORMS J Comput* 2009;21(4):599–613.
36. Rasmussen C, Williams C. Gaussian processes for machine learning. Cambridge (MA): MIT Press; 2006.
37. Žilinskas A. Single-step Bayesian search method for an extremum of functions of a single variable. *Cybern Syst Anal* 1975;11(1):160–166.
38. Mockus J. Bayesian approach to global optimization: theory and applications. Dordrecht: Kluwer Academic; 1989.
39. Žilinskas A. A review of statistical models for global optimization. *J Glob Optim* 1992;2(2):145–153.
40. Schonlau M. Computer experiments and global optimization [PhD thesis]. University of Waterloo; 1997.
41. Jones D, Schonlau M, Welch W. Efficient global optimization of expensive black-box Functions. *J Glob Optim* 1998;13(4):455–492.
42. Schonlau M, Welch W, Jones D. New developments and applications in experimental design. Volume 34, *Global versus local search in constrained optimization of computer models*. Seattle (WA): Institute of Mathematical Statistics; 1998. pp. 11–25.

43. Sasena M. Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations [PhD thesis]. University of Michigan; 2002.
44. Williams B, Santner T, Notz W. Sequential design of computer experiments to minimize integrated response functions. *Stat Sin* 2000;10:1133–1152.
45. Huang D, Allen T, Notz W, *et al.* Global optimization of stochastic black-box systems via sequential kriging meta-models. *J Glob Optim* 2006;34(3):441–466.
46. Kennedy M, O'Hagan A. Bayesian calibration of computer models. *J R Stat Soc Ser B Stat Methodol* 2001;63(3):425–464.
47. Santner T, Williams BW, Notz W. The design and analysis of computer experiments. New York: Springer; 2003.
48. O'Hagan A. Bayesian analysis of computer code outputs: a tutorial. *Reliab Eng Syst Saf* 2006;91(10-11):1290–1300.
49. Floudas C, Gounaris C. A review of recent advances in global optimization. *J Glob Optim* 2008;45:3–38.
50. Zhigljavsky A, Žilinskas A. Stochastic global optimization. New York: Springer. 2008.

DECOMPOSITION ALGORITHMS FOR TWO-STAGE RECURSE PROBLEMS

JÁNOS MAYER

Institute for Operations Research,
University of Zurich, Zurich,
Switzerland

In this article decomposition methods for two-stage linear recourse problems with a finite discrete distribution are discussed. The following algorithms are presented: L-shaped decomposition, regularized decomposition, and stochastic decomposition. Variants for recourse-constrained problems and special cases including simple recourse with a random technology matrix are also considered. In connection with stochastic decomposition, the scope of which is not restricted to finite discrete distributions, algorithms for the continuous-distribution case are discussed briefly with references to the literature.

PROBLEM FORMULATION

We present decomposition algorithms for two-stage linear recourse problems with a finite discrete distribution of the random model parameters. Concerning the general case, which involves also continuous distributions, we have included a brief discussion in the section titled "Stochastic Decomposition" with references to the literature.

The basic problem formulation is the following:

$$\left. \begin{array}{l} \min_x \quad c^T x + \sum_{s=1}^S p_s Q_s(x) \\ Ax = b \\ x \geq 0, \end{array} \right\} \quad (1)$$

where $x \in \mathbb{R}^n$ is the vector of first-stage decision variables, A is an $(m \times n)$ matrix, and b and c are vectors with compatible dimensions. For the sake of simplicity of presentation, we assume that $B := \{x \mid Ax = b, x \geq 0\}$ is a nonempty and bounded set.

The second term in the objective function represents the expected recourse costs with $p_s > 0$ representing the scenario probabilities, $s = 1, \dots, S$, $\sum_{s=1}^S p_s = 1$. The *recourse function* $Q_s(x)$ stands for the recourse costs in scenario s , defined via the optimal objective value of *recourse subproblems* as follows:

$$Q_s(x) := \min_y \left. \begin{array}{l} (q^s)^T y \\ W^s y = h^s - T^s x \\ y \geq 0, \end{array} \right\} \quad (2)$$

where $y \in \mathbb{R}^{n_2}$ is the vector of recourse variables, W^s is an $(m_2 \times n_2)$ matrix, and the matrix T^s and the vectors h^s and q^s have compatible dimensions. W^s is called the *recourse matrix* and T^s is called the *technology matrix*. If W^s varies with s , then we have *random recourse* whereas *fixed recourse* means that $W^s \equiv W$, $\forall s$ holds, that is, the recourse matrix has no stochastic entries in this case. A model with fixed recourse is called a *complete recourse* model, if subproblem (2) has feasible solutions for any $x \in \mathbb{R}^n$ and for all $s = 1, \dots, S$. A special case of complete recourse is *simple recourse*, meaning that $n_2 = 2m_2$ and $W = (I, -I)$ hold with I standing for an $(m_2 \times m_2)$ unit matrix. A fixed-recourse model has *relatively complete recourse* if subproblem (2) is feasible for any $x \in B$ and for all s .

We also need the dual of the recourse subproblem, which reads

$$Q_s(x) := \max_u \left. \begin{array}{l} (h^s - T^s x)^T u \\ (W^s)^T u \leq q^s. \end{array} \right\} \quad (3)$$

It may be noted that the feasible domain of subproblem (3) does not depend on x . Therefore, if for some scenario s the dual problem is infeasible, this immediately implies that for any x the recourse subproblem (2) is either infeasible or otherwise the objective function is unbounded from below, and consequently our two-stage recourse problem (1) has no optimal solution. Therefore we assume in the sequel that subproblem (3) is feasible for all $s = 1, \dots, S$.

From the duality theory of linear programming (LP) [1], it follows under this assumption that for any $x \in \mathbb{R}^n$ and for any $s \in \{1, \dots, S\}$, the recourse subproblem (2) is either infeasible or otherwise it has an optimal solution.

Denoting the set of those x vectors for which the recourse subproblem (2) has feasible solutions by $K_s \subset \mathbb{R}^n$, it turns out that $Q_s(x)$ is a piecewise linear convex function, also called a *polyhedral function* defined over the polyhedral set K_s . Setting $K := \bigcap_s K_s$, it follows that the objective function of subproblem (1) is a finite-valued piecewise linear convex function over K . Adding the requirement $x \in K$, called *induced constraint*, to the constraint set of subproblem (1), the resulting problem can be classified as a convex nonsmooth optimization problem, with a polyhedral objective function. In fact, this is the starting point for a class of optimization algorithms for two-stage recourse problems. It may be noted that K is not explicitly given along with the dataset of the problem. Consequently, in the absence of relatively complete recourse, algorithms have to work with appropriate approximations of the set K .

It is easy to see that by introducing the auxiliary variables $y^1, \dots, y^S$, our two-stage recourse problem (1) can be equivalently formulated as an LP problem as follows:

$$\left. \begin{array}{ll} \min_{x, y^1, \dots, y^S} & c^T x + p_1(q^1)^T y^1 \dots + p_S(q^S)^T y^S \\ & \left. \begin{array}{ll} Ax & = b \\ T^1 x & + W^1 y^1 = h^1 \\ \vdots & \vdots \\ T^S x & + W^S y^S = h^S \\ x & \geq 0 \end{array} \right\} \\ & y^1 \geq 0 \dots y^S \geq 0. \end{array} \right\} \quad (4)$$

Having the above LP formulation, it is tempting to conclude that two-stage recourse problems with a finite discrete distribution can just be solved by utilizing general-purpose LP solvers applied directly to the LP problem (4). Current developments in LP-solver technology along with developments in computer hardware seem to justify this view. To see the weak point in this

reasoning, first note that the size of the constraint matrix of the LP problem (4) is $((m + S \cdot m_2) \times (n + S \cdot n_2))$. Considering a problem with altogether r stochastically independent random entries each of them having k realizations results in $S = k^r$ joint realizations, which grows exponentially with r . Thus, the combinatorial explosion quickly leads to practical problems that are numerically intractable with general-purpose solvers. For a recent experimental study that highlights the shortcomings of the above-mentioned naïve view, see Zverovich *et al.* [2].

Consequently, special-purpose algorithms that utilize the structure of the problem are needed. Considering the structure of the LP problem (4), we observe that the constraint matrix is a sparse matrix with a *dual block-angular* structure. In fact the dual problem of the LP problem (4) has a block-angular structure fitting into the scheme of the Dantzig–Wolfe decomposition [3] of LP and historically this was the first proposal for solving two-stage recourse problems with a finite discrete distribution provided by Dantzig and Madansky [4] in 1961. Concerning numerically efficient methods, the breakthrough came in 1969 with the L-shaped decomposition method of Van Slyke and Wets [5]. The basic idea is to apply the decomposition method of Benders [6] to the specially structured problem (4). The L-shaped method was the basis for the development of several further decomposition algorithms, like the regularized decomposition method of Ruszczyński [7] or stochastic decomposition algorithms of Higle and Sen [8]. Detailed presentations of these methods can be found in the following books: Birge and Louveaux [9, Chapter III, Sections 5 and 6], Kall and Mayer [10, Chapter 1, Sections 2.6–2.8 and Chapter 4, Section 4.2], Kall and Wallace [11 Sections 1.6.4 and 3.1–3.3], Mayer [12 Sections 1.2–1.3 and 3.1], and Prékopa [13 Chapter 12]. For the theoretical background see also Shapiro *et al.* [14] and the handbook edited by Ruszczyński and Shapiro [15]. For survey articles, see Wets [16] and Ruszczyński [17,18].

THE L-SHAPED METHOD

We give a short description of the method below under the simplifying assumption of *fixed recourse* and assuming that $q^s = q$, $\forall s$ holds. On this basis the algorithm for the general case can be derived in a straightforward manner. Under our assumptions the polyhedral feasible domain of the dual recourse subproblem (3) does not depend on s ; let us denote the finite set of vertices by $\mathcal{U}$ and the finite set of extremal rays of this polyhedral set by $\mathcal{V}$. From the duality theory of LP (see also the lemma of Farkas), it follows that the set $K_s \subset \mathbb{R}^n$ for which the s th recourse subproblem (2) has feasible solutions has the representation $\{x \mid v^T(h^s - T^s x) \leq 0, \forall v \in \mathcal{V}\}$ and for $x \in K_s$ the recourse function can be written as $Q_s(x) = \max_{u \in \mathcal{U}} u^T(h^s - T^s x)$. Using these observations, the two-stage recourse problem (1) can be written equivalently as

$$\min_{x, w} \left\{ \begin{array}{l} c^T x + w \\ \sum_{s=1}^S p_s u_s^T (h^s - T^s x) - w \leq 0, (u_1, \dots, u_S) \in \tilde{\mathcal{U}} \\ v^T (h^s - T^s x) \leq 0, v \in \mathcal{V} \\ x \in \mathcal{B} \end{array} \right\} \quad (5)$$

where $\tilde{\mathcal{U}}$ is the set of all S -tuples of vertices. Problem (5) is called the *full master problem* in the *aggregate form*. Since $\tilde{\mathcal{U}}$ and $\mathcal{V}$ are finite sets, this is an LP problem containing a huge amount of constraints, in general. The main idea of the L-shaped method of Van Slyke and Wets [5] is to generate a sequence of outer approximations to the feasible set via *constraint generation*. The constraints generated corresponding to the first type of constraints in the full master problem (5) will be called *optimality cuts*, while those of the second type will be termed as *feasibility cuts*. Let us denote by $\tilde{\mathcal{U}}^v$ the set of S -tuples of vertices corresponding to optimality cuts and denote by $\mathcal{V}^v$ the set of extremal rays corresponding to feasibility cuts, generated up to iteration $v \geq 1$. At $v = 0$ both sets are empty and $w = 0$ is fixed. The *relaxed master problem* at iteration v is the LP problem

$$\min_{x, w} \left\{ \begin{array}{l} c^T x + w \\ \sum_{s=1}^S p_s u_s^T (h^s - T^s x) - w \leq 0, (u_1, \dots, u_S) \in \tilde{\mathcal{U}}^v \\ v^T (h^s - T^s x) \leq 0, v \in \mathcal{V}^v \\ x \in \mathcal{B} \end{array} \right\} \quad (6)$$

At each iteration v , the following steps are carried out:

- S1. Solve the current relaxed master problem (6) leading to an optimal solution (x^v, w^v) . Because of the outer approximation, the optimal objective value $\underline{F}^v$ is a lower bound of the optimal objective value of the two-stage problem (1).
- S2. For $s = 1, \dots, S$ in turn solve the dual recourse subproblems (3) with $x = x^v$, by utilizing a simplex-type method. If for some s the objective function turns out to be unbounded on the feasible domain, add a feasibility cut to the relaxed master problem and go to the next iteration. Otherwise, that is, if optimal solutions were obtained for all $s = 1, \dots, S$, add an optimality cut, compute the expected value of the recourse function $Q(x^v) := \sum_{s=1}^S p_s Q_s(x^v)$, and continue with the next step. In this case, $\bar{F}^v := c^T x^v + Q(x^v)$ is an upper bound on the optimal objective value.
- S3. Check for optimality: if $Q(x^v) - w^v \leq 0$ then STOP. x^v is an optimal solution of the two-stage recourse problem (1) (in practical implementations the optimality check is carried out by employing a stopping tolerance $\varepsilon > 0$, of course). Otherwise continue with the next iteration.

It can be shown that the algorithm terminates in a finite number of iterations. The reason is that none of the feasibility cuts may reappear and if an optimality cut reappears then x^v is already optimal.

It may be noted that an optimality cut is only added to the relaxed master problem if the recourse subproblems had optimal solutions for all scenarios s . An alternative equivalent formulation of the two-stage recourse

problem (1) is the following LP problem:

$$\min_{x,w} \left. \begin{array}{l} c^T x + \sum_{s=1}^S p_s w_s \\ u^T(h^s - T^s x) - w_s \leq 0, u \in \mathcal{U} \\ v^T(h^s - T^s x) \leq 0, v \in \mathcal{V} \\ x \in \mathcal{B}. \end{array} \right\} \quad (7)$$

Problem (7) is called the *full master problem* in the *disaggregate form*. Applying the analogous constraint generation scheme as before leads to the *multicut* L-shaped method of Birge and Louveaux [19]. The relaxed master problem at iteration ν now reads as follows:

$$\min_{x,w} \left. \begin{array}{l} c^T x + \sum_{s=1}^S p_s w_s \\ u^T(h^s - T^s x) - w_s \leq 0, u \in \mathcal{U}^\nu \\ v^T(h^s - T^s x) \leq 0, v \in \mathcal{V}^\nu \\ x \in \mathcal{B}, \end{array} \right\} \quad (8)$$

where $\mathcal{U}^\nu$ consists of the vertices and $\mathcal{V}^\nu$ is the set of extremal rays that appeared up to iteration ν in the cuts. The difference in the algorithm is in Step 2, which for the multicut version runs as follows:

- S2. For $s = 1, \dots, S$ in turn solve the recourse subproblems (3) with $x = x^\nu$, by utilizing a simplex-type method. If for some s the objective function turns out to be unbounded on the feasible domain, add a feasibility cut to the relaxed master problem; otherwise add an optimality cut. If optimal solutions were obtained for all $s = 1, \dots, S$ then compute the expected value of the recourse function as before and continue with the next step. Otherwise continue with Step 1 of the next iteration.

Consequently, in the multicut version S cuts are added to the relaxed master problem in each of the iterations. This leads to a rapid growth of the number of constraints in the relaxed master problem.

In Step 2 of the L-shaped algorithm, S subproblems have to be solved for generating the cuts, which may become a computational bottleneck for large S . Either the dual

recourse subproblem (3) is solved directly or the recourse subproblem (2) is solved and the dual solution is derived from this. In the latter case, assuming $W^s = W$ and $q^s = q$, $\forall s$, we have to do with a sequence of LP problems that only differ in their right-hand sides. Algorithms for the solution of such sequences of LP problems have been proposed by Haugland and Wallace [20] and Gassmann and Wallace [21].

Step 2 also lends itself to parallel computing, and the subproblems can be solved in a parallel fashion by employing several processors. In fact, such variants have been developed by Linderoth and Wright [22,23].

Abaffy and Allevi [24] present a modification of the L-shaped method with aggregate cuts for fixed-recourse problems. Utilizing the special structure, they propose an algorithm with the ABS technique [25] applied to the LP problems arising in the decomposition algorithm, which leads to a substantial reduction in the overall number of arithmetic operations. It may be noted here that the acronym ABS is composed of the initials of the names of the authors Abaffy, Broyden and Spedicato, who have invented this class of methods.

For large S , the decomposition methods may become quite slow. Zhao [26] proposes to apply decomposition methods that are based on interior point (IP) techniques. Both the master problem and subproblems are smoothed via logarithmic barrier functions. The application of the IP technique basically means employing the very fast Newton method for the solution of the smoothed problems and results in polynomial-time algorithms; see Refs Boyd and Vandenberghe [27], Nesterov and Nemirovsky [28], Vanderbei [1] and Ye [29]. Berkelaar *et al.* [30] have developed an IP algorithm for solving two-stage recourse problems with random recourse. The algorithm is based on a homogeneous self-dual method where the decomposition is applied in the direction-finding subproblem of the algorithm.

In applications it may happen that the subproblems (2) are themselves large-scale LP problems; Thus, it makes sense to take approximate solutions for generating the optimality cuts, obtained, for example, by a primal-dual IP method. As long as the appr-

imate solution is dual feasible, this will generate valid cuts. Zakeri *et al.* [31] propose a convergent modification of the L-shaped method based on ε -subgradients and cuts, which are inexact in the above sense.

Inexact cuts form the basis of the level-decomposition method of Fábíán and Szóke [32], too. On the basis of a general algorithm for convex programming, in their method feasibility and optimality features are taken into account simultaneously. The probability distribution is approximated by coarser discrete distributions with increasing approximation accuracy across iterations, leading to inexact cuts in the decomposition method.

Because of similarities with the decomposition methods we also mention basis factorization methods utilizing the special structure of the LP problem (4) in the simplex method, when applied to the solution of the LP problem (4). The first method of this type has been proposed by Strazicky [33]; for further algorithms in this Class, see Birge and Louveaux [19, Section 5.6].

Takriti and Ahmed [34] consider two-stage recourse problems with suitably chosen variability measures in the objective and propose a variant of the L-shaped method for the solution of such problems.

REGULARIZED DECOMPOSITION

In the sequence of iterations of the L-shaped method, the number of constraints is steadily growing in the relaxed master problem. This growth is especially rapid in the variant with disaggregate cuts where S new cuts are added in each iteration. In the L-shaped method, there is no reliable strategy for dropping redundant cuts, without destroying the finite termination property of the algorithm. Consequently, with increasing number of iterations the resulting large number of constraints may cause numerical difficulties in the relaxed master problem. In addition to this, the relaxed master problem becomes increasingly ill-conditioned because of almost parallel cuts added especially in the vicinity of the optimal solution.

Another unfavorable property of the L-shaped algorithm is the following: even if the starting point x^0 is already close to the

optimal solution of the two-stage recourse problem (1), after adding cuts in Step 2 the subsequent points x^k , $k \geq 1$ may be far away from the optimum and it may take a large number of iterations till the iterates come back to the neighborhood of the optimal solution.

For overcoming these difficulties, Ruszczyński [7] proposed to apply regularization to the relaxed master problem in the disaggregate form leading to the following objective function in the relaxed master problem (8):

$$c^T x + \alpha \|x - z^v\|^2 + \sum_{s=1}^S p_s w_s,$$

which now contains a regularizing term $\alpha \|x - z^v\|^2$, where z^v is the current *candidate solution*, α is a scaling factor, and $\|\cdot\|$ is the Euclidean norm. α is kept zero till the first candidate solution is found; afterward it is either set to a constant $\alpha > 0$ or it is used as a control parameter varying across iterations in a finite interval $[\alpha_1, \alpha_2]$ with $\alpha_1 > 0$. The first candidate solution will be the solution of the relaxed master problem after an optimality cut has been added for the first time. The idea is that the current candidate solution is kept over iterations, $x^{v+1} := x^v$, as long as the approximation to the objective function is not good enough in the vicinity of it. The criterion for changing the candidate to $z^v := x^v$ in an iteration v without feasibility cuts is

$$\sum_{s=1}^S p_s (u_s^v)^T (h^s - T^s x^v) \leq \sum_{s=1}^S p_s w_s^v,$$

where $(u_1^v, \dots, u_S^v)$ are the vertices corresponding to the optimality cuts added at iteration v . Besides this, further rules for changing the candidate solution have also been suggested, based on approximations; see Ruszczyński [7] and Ruszczyński and Święt anowski [35].

It may be noted that the regularized relaxed master problem is now a quadratic programming problem of the constrained least squares type. Reference [7] also contains a proposal for solving it, based on an active-set strategy.

An outstandingly important feature of the regularized decomposition method is that it

allows for cut-dropping. After having solved the current regularized relaxed master problem, all cuts having zero Lagrange multipliers at the optimal solution can be dropped. Thus, the number of cuts in the regularized relaxed master problems to solve is at most $n + 2S$ throughout the iterations. It can be shown that the regularized decomposition method terminates in a finite number of iterations.

Alternative regularization techniques are based on trust region constraints around the solution from the previous iteration. Linderoth and Wright [22,23] propose to add a trust-region constraint to the relaxed master problem having the form $\|x - x^{v-1}\|_\infty = \max_i |x_i - x_i^{v-1}| < \Delta_v$. The choice of the ∞ -norm allows for an equivalent reformulation of the relaxed master problem as an LP problem. In fact, the trust-region constraint can be written equivalently as

$$\begin{aligned} |x_i - x_i^{v-1}| &< \Delta_v, \quad \forall i \\ \Leftrightarrow -\Delta_v &< x_i - x_i^{v-1} < \Delta_v, \quad \forall i, \end{aligned}$$

resulting in linear constraints. The regularized version is embedded into an L -shaped method for parallel computation.

STOCHASTIC DECOMPOSITION

In Step 2 of the L -shaped method, S LP problems have to be solved. For S really large this step will be impossible to carry out. Having, for instance, a recourse problem with 10 stochastically independent random entries, each of them having 10 realizations, the number of joint realizations S will be $S = 10^{10}$ for which Step 2 becomes impossible to carry out. Under such circumstances stochastic decomposition techniques can be used. For these methods fixed recourse and $q^s = q$ is assumed.

Infanger [36,37] proposes an algorithm where, in each iteration v of the L -shaped method, in Step 2 a separate sample is drawn from the distribution. On the basis of this sample, a Monte Carlo approximation of the cuts is computed employing variance reduction via importance sampling. The sample size is increased in each of the

iterations. As in the L -shaped method, upper and lower bounds are utilized for the stopping rule. The difference is that the bounds are in this case random variables themselves and the stopping rule involves Student's t -test for checking whether the difference of these bounds is less than the stopping tolerance, on a given significance level. Since the algorithm involves sampling within the L -shaped method, it belongs to the class of *internal* sampling algorithms (also called *interior* sampling methods).

The basic representant of the class of internal sampling algorithms is the stochastic decomposition method (SD) of Higle and Sen [8,38]. While in Infanger's method a separate whole sample is drawn at each of the iterations, in the SD method just one sample point is drawn per iteration. Thus a single sample is being built as the iterations progress.

For discussing the SD algorithm, we assume complete recourse. We also assume that $Q_s(x) \geq 0$ holds for all $x \in \mathcal{B}$ and for all s ; for a weaker assumption, see Higle and Sen [38]. In the SD method the sample is drawn in a step-by-step manner and in each of these steps an appropriately modified iteration step of the L -shaped method is carried out. At iteration v the next sample element (h_v, T_v) is drawn and in Step 2 of the L -shaped algorithm a single recourse subproblem (3) is solved corresponding to $s = v$. Thus, during the course of the algorithm a single sample is being built that has sample size v at iteration v . The next optimality cut is computed on the basis of the following inequality:

$$\frac{1}{v} \sum_{s=1}^v u_s^T (h^s - T^s x) \leq \frac{1}{v} \sum_{s=1}^v Q_s(x),$$

where $((h_1, T_1), \dots, (h_v, T_v))$ is the sample drawn up to iteration v and $u_s \in \mathcal{U} \forall s$ holds. On the right-hand side of this inequality stands the actual Monte Carlo approximation of the expected recourse function, based on the current sample, while the left-hand side represents the optimality cut in the aggregate form to be added in this iteration. It may be noted that the inequality is valid for any $u_s \in \mathcal{U}$. The idea

is to store the optimal dual vertices up to iteration $\nu - 1$ while u_ν comes from the solution of the recourse subproblem. For each $s = 1, \dots, \nu - 1$ a dual vertex corresponding to the highest $u_s^T(h^s - T^s x)$ value for $x = x^\nu$ is chosen. The principle behind this procedure is to generate a cut as tight as possible.

The cuts generated in the previous iterations, corresponding to Monte Carlo approximations from previous steps, may become invalid in the current iteration. Consequently, those cuts need to be updated. Let $\nu \geq 2$ and $\kappa < \nu$. The basis for an update is the following inequality

$$\begin{aligned} \frac{1}{\nu} \sum_{s=1}^{\kappa} u_s^T(h^s - T^s x) &\leq \frac{1}{\nu} \sum_{s=1}^{\kappa} Q_s(x) \\ &\leq \frac{1}{\nu} \sum_{s=1}^{\nu} Q_s(x), \end{aligned}$$

which holds for any $u_s \in \mathcal{U}$ because of LP duality and because of the assumption $Q_s(x) \geq 0$, for all $x \in \mathcal{B}$. At the earlier iteration κ a cut involving the function

$$\frac{1}{\kappa} \sum_{s=1}^{\kappa} u_{s(\kappa)}^T(h^s - T^s x),$$

has been generated, with $u_{s(\kappa)} \in \mathcal{U}$, $s = 1, \dots, \kappa$, being the dual vertices chosen at iteration κ . Owing to the inequality above, replacing this cut function by

$$\frac{1}{\nu} \sum_{s=1}^{\kappa} u_{s(\kappa)}^T(h^s - T^s x),$$

results in a valid cut at iteration $\nu > \kappa$. It may be noted that with increasing ν the earlier cut, based on a smaller sample size κ , becomes gradually phased out. It is easy to see that this transformation can be carried out as follows: at each of the iterations $\nu \geq 2$, all of the previous cuts are multiplied by $\frac{\nu-1}{\nu}$.

The main problem concerning stochastic algorithms, like the SD method outlined so far, is to identify appropriate stopping rules. The authors employ *incumbent solutions* for this reason. Initially the incumbent solution is just the solution of the expected value problem. At each iteration it is checked whether

the current optimal objective function value of the relaxed master problem is sufficiently lower than the approximate objective value at the incumbent solution. If yes, the current solution x^ν of the relaxed master problem becomes the new incumbent solution. The cut corresponding to the incumbent solution is updated at each iteration. On the basis of a subsequence of iterations where the incumbent changes, several stopping rules are proposed by the authors; see Higle and Sen [38].

The method as outlined above is based on aggregate cuts. An SD method employing disaggregate cuts has been developed by Higle *et al.* [39]. As in the deterministic L -shaped method, the accumulation of redundant cuts induces numerical difficulties in the relaxed master problem. To avoid this, Higle and Sen [40] and Yakowitz [41] developed a regularized SD method that allows for dropping redundant cuts in a safe manner, similar to the deterministic regularized decomposition method.

The monograph of Higle and Sen [38] contains a detailed description of the SD method and its variants discussed above. Sen *et al.* [42] present an overview on recent developments concerning the SD algorithm. It may be noted that, by its very nature, SD is also applicable for two-stage recourse problems with continuous distributions.

For providing information to the interested reader, we briefly discuss two algorithmic approaches for the case with continuous distributions, although they do not belong to the class of decomposition methods.

Sample average approximation (SAA) methods are *external* sampling methods. A single sample is drawn and considering it as a special discrete distribution the resulting problem is solved, for example, by applying the L -shaped method. Important related problems include the choice of a proper sample size and validation analysis. For a detailed presentation, see Shapiro [43] and the references therein.

Successive discrete approximation methods construct a sequence of discrete approximations to the distribution. The idea is to partition the support of the distribution into a finite number of subsets, and to

construct an approximate discrete distribution by taking the conditional expectations and conditional probabilities corresponding to this partition. The resulting two-stage recourse problems with discrete distributions are solved and based on the results, a refinement of the partition is carried out for the next (refined) approximation. An outstanding feature of these algorithms is that at each iteration deterministic lower and upper bounds are provided for the optimal objective value of the original problem. For details, see Kall and Stoyan [44], Frauendorfer and Kall [45], and Frauendorfer [46]. For an introduction and implementation of these methods, see also Kall and Mayer [10] and Mayer [12].

SPECIAL CASES AND RECOURSE CONSTRAINTS

An important special case of the general two-stage recourse problem is simple recourse. Let us first assume that in the simple recourse problem only the right-hand side is stochastic, meaning that $q^s = q$ and $T^s = T$ hold for all s . For this case, very efficient solution algorithms are available, as discussed in Birge and Louveaux [19], Kall and Wallace [11], and Wets [47,16]. The basic observation is that in this case the recourse function becomes separable in the variables $\chi := Tx$ and an equivalent LP can be formulated that has a much smaller size than the LP (4). This LP can be solved very efficiently; consequently, in this case there is no need to apply decomposition methods. We emphasize that having a deterministic technology matrix T is essential for obtaining the small-scale equivalent LP.

Considering simple recourse problems with a *random technology matrix*, the simplifications outlined above do not carry over to this case. Klein Haneveld and Van der Vlerk [48] observed that, for such problems, solving the recourse subproblems in Step 2 of the L -shaped method essentially reduces to just checking signs of the components of the vector $h^s - T^s x^v$, which yields a great speedup in the overall method. They apply the algorithm for stochastic optimization problems with integrated chance constraints. Künzi-Bay

and Mayer [49] applied this technique for solving stochastic optimization problems involving the minimization of Conditional Value-at-Risk (CVaR:), which can be reformulated as a simple recourse problem with a random technology matrix. The authors report on a speedup of several *orders of magnitude* in the computing time when it is compared with solving the equivalent LP by general-purpose commercial LP solvers.

Another practically important special case is network recourse, which means that the recourse subproblems are network-flow problems. A numerically efficient variant of the L -shaped method has been proposed by Wallace [50] for such problems.

Stochastic optimization problems with a *recourse constraint* have been introduced and first studied by Higle and Sen [51]. Such problems have the form

$$\left. \begin{array}{l} \min_x \quad c^T x \\ d^T x + \sum_{s=1}^S p_s Q_s(x) \leq \gamma \\ x \in \mathcal{B}, \end{array} \right\}$$

with γ prescribed. Yakowitz [52] developed an algorithm based on exact penalty functions for this problem class. Kall and Mayer propose in Chapter 4, Section 4.2 of Ref. 10 an application of the L -shaped method for the solution of such problems. In Section 4.4 of the same chapter a numerically efficient L -shaped algorithm is presented for problems with a CVaR constraint, as a special case.

REFERENCES

1. Vanderbei RJ. Linear programming: foundations and extensions. 3rd ed. New York: Springer; 2007.
2. Zverovich V, Fábíán C, Ellison F, *et al.* A computational study of a solver system for processing two-stage stochastic linear programming problems. Technical Report 2009 Nr. 3. Stochastic Programming E-Print Series (SPEPS), 2009.
3. Dantzig GB, Wolfe P. The decomposition principle for linear programs. Oper Res 1960;8: 101–111.
4. Dantzig GB, Madansky A. On the solution of two-stage linear programs under uncertainty.

- In: Neyman IJ, editor. Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1961. pp. 165–176.
5. Van Slyke R, Wets RJ-B. *L-shaped linear programs with applications to optimal control and stochastic linear programs*. SIAM J Appl Math 1969;17:638–663.
6. Benders JF. Partitioning procedures for solving mixed-variables programming problems. Num Math 1962;4:238–252. Re-published in Comput Manag Sci 2005;2:3–19.
7. Ruszczyński A. A regularized decomposition method for minimizing a sum of polyhedral functions. Math Program 1986;35:309–333.
8. Higle JL, Sen S. Stochastic decomposition: an algorithm for two stage linear programs with recourse. Math Oper Res 1991;16:650–669.
9. Birge JR, Louveaux F. Introduction to stochastic programming. Springer Series in Operations Research. Berlin/Heidelberg: Springer; 1997.
10. Kall P, Mayer J. Stochastic linear programming. Models, theory and computation. New York: Springer; 2005.
11. Kall P, Wallace SW. Stochastic programming. Chichester: John Wiley & Sons; 1994.
12. Mayer J. Stochastic linear programming algorithms: a comparison based on a model management system. Amsterdam: Gordon and Breach Science Publishers; 1998.
13. Dréková A. Stochastic programming. Dordrecht: Kluwer Academic Publishers; 1995.
14. Shapiro A, Dentcheva D, Ruszczyński A. Lectures on stochastic programming: modeling and theory. Philadelphia (PA): SIAM; 2009.
15. Ruszczyński A, Shapiro A, editors. Stochastic programming. Handbooks in operations research and management science. Amsterdam: Elsevier; 2003.
16. Wets RJ-B. Large scale linear programming techniques in stochastic programming. In: Ermoliev YuB, Wets RJ-B, editors. Numerical techniques for stochastic optimization. Berlin: Springer; 1988. pp. 65–93.
17. Ruszczyński A. Some advances in decomposition methods for stochastic linear programming. Ann Oper Res 1999;85:153–172.
18. Ruszczyński A. Decomposition methods. In: Ruszczyński A, Shapiro A, editors. Stochastic programming. Handbooks in operations research and management science. Amsterdam: Elsevier; 2003. pp. 141–211.
19. Birge JR, Louveaux FV. A multicut algorithm for two stage linear programs. Eur J Oper Res 1988;34:384–392.
20. Haugland D, Wallace SW. Solving many linear programs that differ only in the righthand side. Eur J Oper Res 1988;37:318–324.
21. Gassmann HI, Wallace SW. Solving linear programs with multiple right-hand-sides: pricing and ordering schemes. Ann Oper Res 1996;64:237–259.
22. Linderoth J, Wright S. Decomposition algorithms for stochastic programming on a computational grid. Comput Optim Appl 2003;24: 207–250.
23. Linderoth J, Wright S. Computational grids for stochastic programming. In: Wallace SW, Ziemba WT, editors. Applications of stochastic programming, MPS–SIAM Series in Optimization. Philadelphia: SIAM and MPS; 2005. pp. 61–77.
24. Abaffy J, Allevi E. A modified L-shaped method. J Optim Theor Appl 2004;123:255–270.
25. Abaffy J, Spedicato E. ABS projections algorithms: mathematical techniques for linear and nonlinear algebraic equations. Chichester, England: Ellis Horwood; 1989.
26. Zhao G. A log-barrier method with Benders decomposition for solving two-stage stochastic programs. Math Program 2001;90:507–536.
27. Boyd S, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004.
28. Nesterov Y, Nemirovsky A. Interior point polynomial methods in convex programming: theory and algorithms. Philadelphia (PA): SIAM; 1993.
29. Ye Y. Interior point algorithms-theory and analysis. New York: John Wiley & Sons; 1997.
30. Berkelaar A, Dert C, Oldenkamp B, *et al.* A primal-dual decomposition-based interior point approach to two-stage stochastic linear programming. Oper Res 2002;50:904–915.
31. Zakeri G, Philpott AB, Ryan DM. Inexact cuts in Benders decomposition. SIAM J Optim 2000;10:643–657.
32. Fábíán CsI, Szóke Z. Solving two-stage stochastic programming problems with level decomposition. Comput Manag Sci 2007;4: 313–353.
33. Strazicky B. Some results concerning an algorithm for the discrete recourse problem. In: Dempster MAH, editor. Stochastic programming. New York: Academic Press; 1980. pp. 263–271.

34. Takriti S, Ahmed S. On robust optimization of two-stage systems. *Math Program* 2004;99: 109–126.
35. Ruszczyński A, Świątanowski A. Accelerating the regularized decomposition method for two stage stochastic linear programming problems. *Eur J Oper Res* 1997;101:328–342.
36. Infanger G. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Ann Oper Res* 1992;39:69–95.
37. Infanger G. Planning under uncertainty. Solving large-scale stochastic linear programs. Danvers (MA): Boyd & Fraser; 1994.
38. Higle JL, Sen S. Stochastic decomposition. A statistical method for large scale stochastic linear programming. Dordrecht: Kluwer Academic Publishers; 1996.
39. Higle JL, Lowe WW, Odio R. Conditional stochastic decomposition: an algorithmic interface for optimization and simulation. *Oper Res* 1994;42:311–322.
40. Higle JL, Sen S. Finite master programs in regularized stochastic decomposition. *Math Program* 1994;67:143–168.
41. Yakowitz DS. A regularized stochastic decomposition algorithm for two-stage stochastic linear programs. *Comput Optim Appl* 1994;3: 59–81.
42. Sen S, Zhou Z, Huang K. Enhancements of two-stage stochastic decomposition. *Comput Oper Res* 2009;36:2434–2439.
43. Shapiro A. Monte Carlo sampling methods. In: Ruszczyński A, Shapiro A, editors. Stochastic programming. Handbooks in operations research and management science. Amsterdam: Elsevier; 2003. pp. 353–425.
44. Kall P, Stoyan D. Solving stochastic programming problems with recourse including error bounds. *Math Operationsforsch Stat, Ser Opt* 1982;13:431–447.
45. Frauendorfer K, Kall P. A solution method for SLP recourse problems with arbitrary multivariate distributions—the independent case. *Prob Contr Inf Theor* 1988;17:177–205.
46. Frauendorfer K. Solving SLP recourse problems with arbitrary multivariate distributions—the dependent case. *Math Oper Res* 1988;13:377–394.
47. Wets RJ-B. Solving stochastic programs with simple recourse. *Stochastics* 1983;10:219–242.
48. Klein Haneveld WK, van der Vlerk MH. Integrated chance constraints: reduced forms and an algorithm. *Comput Manag Sci* 2006;3: 245–269. Originally appeared as Research Report 02A33. SOM, University of Groningen; 2002.
49. Kunzi-Bay A, Mayer J. Computational aspects of minimizing conditional value-at-risk. *Comput Manag Sci* 2006;3:3–27.
50. Wallace SW. Solving stochastic programs with network recourse. *Networks* 1986;16: 295–317.
51. Higle JL, Sen S. Statistical approximations for recourse constrained stochastic programs. *Ann Oper Res* 1995;56:157–175.
52. Yakowitz DS. An exact penalty algorithm for recourse-constrained stochastic linear programs. *Appl Math Comput* 1992;49: 39–62.

DECOMPOSITION METHODS FOR INTEGER PROGRAMMING

TED K. RALPHS

Department of Industrial and Systems
Engineering, Lehigh University,
Bethlehem, Pennsylvania

MATTHEW V. GALATI

Advanced Analytics - Operations R & D,
SAS Institute, Chesterbrook,
Pennsylvania

INTRODUCTION

Decomposition methods are techniques for exploiting the tractable substructures of an integer program in order to obtain improved solution procedures. In particular, the goal is to derive improved methods of bounding the optimal solution value, which can then be used to drive a branch-and-bound algorithm (see **Branch-and-Bound Algorithms**). Such methods are the preferred solution approaches for a wide range of important models arising in practice and have also been the basis for solution approaches for many well-known combinatorial problems. A small sample of the problems for which decomposition approaches have been proposed in the literature includes the multicommodity flow problem [1], the cutting stock problem [2,3], the traveling salesman problem [4], the generalized assignment problem (GAP) [5,6], the bin packing problem [7], the axial assignment problem [8], the Steiner tree problem [9], the single-machine scheduling problem [10], the graph coloring problem [11], and the capacitated vehicle routing problem [12–15].

To expose the desired substructure, a common approach is to relax a set of “complicating constraints.” This is the approach taken by the Dantzig–Wolfe Decomposition (DWD), Lagrangian relaxation, and cutting-plane methods (see **Dantzig–Wolfe Decomposition; Lagrangian Optimization for LP**). Substructure can also be

exposed by fixing the values of a set of variables, that is, considering restrictions of the original problem. This is the approach taken by Benders’ decomposition (see **Benders Decomposition**). This article reviews decomposition methodologies based on relaxation of constraints and examines how they are used to solve mixed integer linear programs (ILPs). For a broader overview, including variable restriction methods, see Vanderbeck and Wolsey [16].

To simplify the exposition, we consider only pure ILPs with finite upper and lower bounds on all variables, so that the set of feasible solutions is finite. The framework can easily be extended to more general settings. For the remainder of the article, we consider an ILP instance whose feasible set is the integer vectors contained in the polyhedron $Q = \{x \in \mathbb{R}^n \mid Ax \geq b\}$, where $A \in \mathbb{Q}^{m \times n}$ is the constraint matrix and $b \in \mathbb{Q}^m$ is the right-hand side vector. Let $\mathcal{F} = Q \cap \mathbb{Z}^n$ be the set of all *feasible solutions* to the ILP and let the polyhedron $\mathcal{P}$ be the convex hull of $\mathcal{F}$. In terms of this notation, the ILP is to determine

$$\begin{aligned} z_{\text{IP}} &= \min_{x \in \mathcal{F}} \{c^\top x\} = \min_{x \in \mathcal{P}} \{c^\top x\} \\ &= \min_{x \in \mathbb{Z}^n} \{c^\top x \mid Ax \geq b\}, \end{aligned} \quad (\text{IP})$$

where $c \in \mathbb{Q}^n$. By convention, $z_{\text{IP}} = \infty$ if $\mathcal{F} = \emptyset$ (the problem is infeasible).

THE PRINCIPLE OF DECOMPOSITION

Computing bounds is an essential element of the branch-and-bound algorithm, which is the most effective and most commonly used method for solving such mathematical programs. Bounds are often computed by solving a *bounding subproblem* that is a tractable relaxation of the original problem. The most commonly used bounding subproblem is the linear programming (LP) relaxation. The LP relaxation is often too weak to be effective in solving difficult ILPs. Decomposition

techniques based on constraint relaxation can be used to obtain improved bounding subproblems.

To apply the principle of constraint decomposition, we consider the relaxation of (IP) defined by

$$\begin{aligned} \min_{x \in \mathcal{F}'} \{c^\top x\} &= \min_{x \in \mathcal{P}'} \{c^\top x\} \\ &= \min_{x \in \mathbb{Z}^n} \{c^\top x \mid A'x \geq b'\}, \end{aligned} \quad (\text{R})$$

where $\mathcal{F} \subset \mathcal{F}' = \{x \in \mathbb{Z}^n \mid A'x \geq b'\}$ for some $A' \in \mathbb{Q}^{m' \times n}, b' \in \mathbb{Q}^{m'}$, and $\mathcal{P}'$ is the convex hull of $\mathcal{F}'$. As usual, we assume that there exist practical¹ algorithms for optimizing over and/or separating from $\mathcal{P}'$. With respect to $\mathcal{F}'$, let $[A'', b''] \in \mathbb{Q}^{m'' \times n''}$ be a set of additional *side constraints* needed to describe $\mathcal{F}$; that is, $[A'', b'']$ is such that $\mathcal{F} = \{x \in \mathbb{Z}^n \mid A'x \geq b', A''x \geq b''\}$. We denote by $\mathcal{Q}'$ the polyhedron described by the inequalities $[A', b']$ and by $\mathcal{Q}''$ the polyhedron described by the inequalities $[A'', b'']$. Hence, the initial LP relaxation is the linear program defined by $\mathcal{Q} = \mathcal{Q}' \cap \mathcal{Q}''$, and the *LP bound* is given by

$$\begin{aligned} z_{\text{LP}} &= \min_{x \in \mathcal{Q}} \{c^\top x\} \\ &= \min_{x \in \mathbb{R}^n} \{c^\top x \mid A'x \geq b', A''x \geq b''\}. \end{aligned} \quad (\text{LP})$$

Note that $[A', b']$ and $[A'', b'']$ are often a partition of the rows of $[A, b]$ into a set of “nice constraints” and a set of “complicating constraints,” but this is not a strict requirement.

Optimization and/or separation over $\mathcal{P}'$ may be more tractable than over $\mathcal{P}$ for a number of reasons. First, it may simply be that $\mathcal{P}'$ has a known structure for which we have well-developed techniques. This first case arises frequently when the original problem is formed by adding side constraints to a well-studied base problem. Second, the resulting relaxation may decompose because of identifiable block-diagonal structure of the matrix A' . In this case, optimization over

$\mathcal{P}'$ can be accomplished by optimizing over each of these blocks independently (possibly in parallel). This second case arises when the constraints of A'' are constraints that link a collection of otherwise independent subsystems (see Example 2). A particular case in which decomposition is often applied very effectively is when A' consists of a set of identical blocks. In this case, the optimization problem over $\mathcal{P}'$ reduces to optimization over just one of the blocks. This can be seen as a way of eliminating symmetry present in the original problem and is one of the primary reasons why decomposition methods work well on certain classes of combinatorial problems.

DECOMPOSITION TECHNIQUES

In this section, we discuss three techniques for computing bounds on z_{IP} , which are based on the principle of decomposition. In the section titled “Relationship of the Techniques,” we show that these methods are very closely related.

Lagrangian Relaxation

For a given vector of *dual multipliers* $u \in \mathbb{R}_+^{m''}$, the *Lagrangian relaxation* [17–20] of (IP) is given by

$$z_{\text{LR}}(u) = \min_{s \in \mathcal{F}'} \{(c^\top - u^\top A'')s + u^\top b''\}. \quad (\text{LR})$$

It is then easily shown that $z_{\text{LR}}(u)$ is a lower bound on z_{IP} . The problem

$$z_{\text{LD}} = \max_{u \in \mathbb{R}_+^{m''}} \{z_{\text{LR}}(u)\} \quad (\text{LD})$$

of maximizing this bound over all choices of dual multipliers is a dual to (IP) called the *Lagrangian dual* (LD) and also provides a lower bound z_{LD} , which we call the *LD bound*. A vector of multipliers $\hat{u}$ that yield the largest bound is called *optimal (dual) multipliers*.

The Lagrangian function $z_{\text{LR}}(u)$ is piecewise linear concave in u and hence can be maximized using the subgradient algorithm to obtain z_{LD} . Alternatively, it can also be

¹The term *practical* is not defined rigorously, but denotes an algorithm with a “reasonable” average-case running time.

solved by rewriting it as the equivalent linear program

$$z_{LD} = \max_{\alpha \in \mathbb{R}, u \in \mathbb{R}_+^{m'}} \left\{ \alpha + u^\top b'' \mid \alpha \leq (c^\top - u^\top A'') s \right. \\ \left. \forall s \in \mathcal{F}' \right\}. \quad (\text{LDLP})$$

Of course, this linear program has a large number of constraints and so must be solved by means of a cut generation algorithm (known as *Kelley's cutting-plane algorithm* [21]). In either approach to solving the LD, most of the computational effort goes into evaluating $z_{LR}(u)$ (called the *Lagrangian subproblem*) for a given sequence of dual multipliers u . This is an optimization problem over $\mathcal{P}'$ and can be solved effectively by assumption. Both these approaches are described in detail in Nemhauser and Wolsey [22].

Dantzig–Wolfe Decomposition

The approach of DWD [23] is to reformulate (IP) by implicitly requiring the solution to be a member of $\mathcal{F}'$, while explicitly enforcing the inequalities $[A'', b'']$. Relaxing the integrality constraints on the variables from the original formulation, we obtain the linear program

$$z_{DW} = \min_{\lambda \in \mathbb{R}_+^{\mathcal{F}'}} \left\{ c^\top \left(\sum_{s \in \mathcal{F}'} s \lambda_s \right) \mid \right. \\ \left. A'' \left(\sum_{s \in \mathcal{F}'} s \lambda_s \right) \geq b'', \right. \\ \left. \sum_{s \in \mathcal{F}'} \lambda_s = 1 \right\}. \quad (\text{DWLP})$$

We call this LP the *Dantzig–Wolfe linear program* (DWLP), though it is frequently referred to as the *master problem* in the literature. Although the number of columns in this linear program is $|\mathcal{F}'|$, it can be solved by column generation². In each iteration of such a method, a restricted version of (DWLP), commonly referred to as the *restricted master*

problem (RMP), is solved to obtain optimal primal and dual solutions. The column generation subproblem is to generate a new column that has negative reduced cost with respect to the dual solution $\hat{u}$ of the RMP (if one exists). This is equivalent to evaluating $z_{LR}(\hat{u})$, as in (LR).

Because the DWLP is solved by column generation, methods that use a DWLP to obtain bounds within a branch-bound-bound algorithm are sometimes referred to generically as *column generation methods* [24,25]; see also the article **Column Generation**. In fact, most column generation algorithms for solving integer programs, even those for which the column generation derives directly from a “natural” formulation, can be seen as arising from the application of DWD to an underlying “compact” model, even if that compact model is not always first formulated explicitly [26]. Branch-and-bound approaches in which column generation is used to obtain a bound in each node are known as *branch-and-price* algorithms [6,27,28]; see also the article **Branch-Price-and-Cut Algorithms**.

It is easy to verify that (DWLP) is an LP dual of (LDLP), which immediately shows that $z_{DW} = z_{LD}$; see Parker and Rardin [29] for a detailed treatment of this fact. Hence, z_{DW} is a valid lower bound on z_{IP} that we call the *DW bound*. Note that the dual variables in (DWLP) correspond to the dual multipliers from (LR). Furthermore, if we combine the members of $\mathcal{F}'$ using $\hat{\lambda}$, we obtain an optimal solution

$$\hat{x}_{DW} = \sum_{s \in \mathcal{F}'} s \hat{\lambda}_s \quad (\text{MAP})$$

to the DWLP, which we call the *optimal fractional solution* corresponding to $\hat{\lambda}$. Note that (MAP) provides a mapping of each point in the feasible set of the Dantzig–Wolfe reformulation to a unique member of $\mathcal{P}' \cap \mathcal{Q}''$. This mapping will be extensively used in the methods discussed in the section titled “Branch, Price, and Cut.” Using this mapping, we can see that $z_{DW} = c^\top \hat{x}_{DW}$. Since $\hat{x}_{DW}$ must lie within $\mathcal{P}' \subseteq \mathcal{Q}'$ and also within $\mathcal{Q}''$, this shows that $z_{DW} \geq z_{LP}$.

²Note that we can equivalently replace $\mathcal{F}'$ with the extreme points of $\mathcal{P}'$.

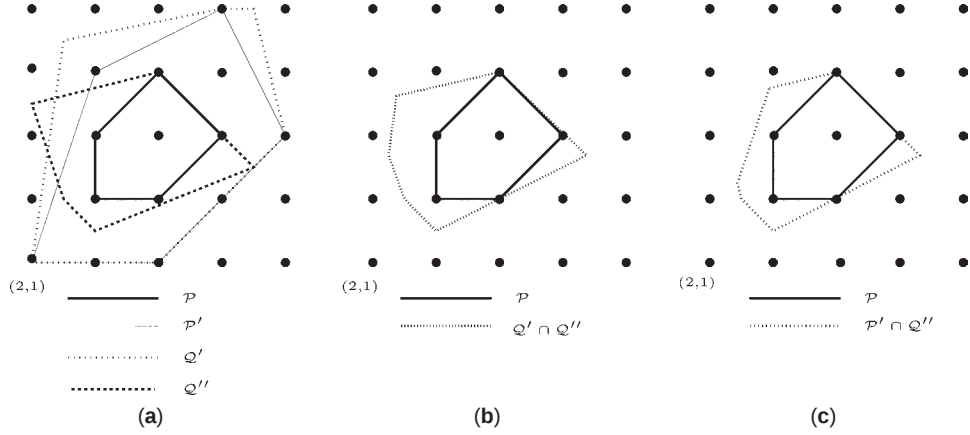


Figure 1. Polyhedra (Example 1: SILP).

Cutting-Plane Method

Although the cutting-plane method [30–33] is not usually thought of as a decomposition method, it can in fact be viewed as such. Under our assumption that the separation problem associated with $\mathcal{P}'$ can effectively be solved, a cutting-plane method in which inequalities describing $\mathcal{P}'$ (i.e., the facet-defining inequalities) are generated dynamically can be implemented. By separating the solutions to a series of augmented LP relaxations from $\mathcal{P}'$, as in the usual cutting-plane approach, the initial LP relaxation can be iteratively augmented to obtain the *CP bound*

$$z_{CP} = \min_{x \in \mathcal{P}'} \{c^\top x \mid A''x \geq b''\}. \quad (\text{CPLP})$$

Note that $\hat{x}_{DW}$, as defined in (MAP), is an optimal solution to this augmented linear program. Hence, the CP bound is equal to both the DW bound and the LD bound. We refer to the augmented linear program above as the *cutting-plane linear programming* (CPLP).

Although we have described the cutting-plane method here as being based on our ability to solve the separation problem exactly over a single polyhedron $\mathcal{P}'$ containing $\mathcal{P}$, one of the advantages of the cutting-plane method over the first two

decomposition approaches is that it is possible, in practice, to simultaneously generate valid inequalities from multiple decompositions. This can further improve the bound beyond what can be achieved with either of the other two methods alone. Furthermore, it is also possible to obtain a valid bound without solving the separation problem *exactly*; that is, the procedure does not have to guarantee to generate a valid inequality violated by the solution to the current LP relaxation whenever one exists. More details on these relationships between the methods are presented in the next section. In the section titled “Advanced Methods,” we discuss how to integrate cutting-plane methods with other decomposition techniques.

RELATIONSHIP OF THE TECHNIQUES

The following well-known result showed that the three approaches just described are simply three different algorithms for computing the same quantity.

Theorem 1 [Geoffrion 34]. $z_{IP} \geq c^\top \hat{x}_{DW} = z_{LD} = z_{DW} = z_{CP} = \min \{c^\top x \mid \mathcal{P}' \cap \mathcal{Q}''\} \geq z_{LP}$.

We refer to the quantity $z_D = \max\{c^\top x \mid \mathcal{P}' \cap \mathcal{Q}''\}$ as the *decomposition bound*. The proof of this result follows directly from the

discussion of the section titled “Decomposition Techniques.” Example 1 illustrates the application of the decomposition principle on a simple two-variable ILP and Fig. 1 demonstrates computation of the bound graphically.

Example 1 [SILP]. Consider the following ILP in two variables:

$$\begin{aligned}
 & \min x_1 \\
 & \text{s.t. } 7x_1 - x_2 \geq 13, & (1) \\
 & \quad x_2 \geq 1, & (2) \\
 & \quad -x_1 + x_2 \geq -3, & (3) \\
 & \quad -4x_1 - x_2 \geq -27, & (4) \\
 & \quad -x_2 \geq -5, & (5) \\
 & \quad 0.2x_1 - x_2 \geq -4, & (6) \\
 & \quad -x_1 - x_2 \geq -8, & (7) \\
 & \quad -0.4x_1 + x_2 \geq 0.3, & (8) \\
 & \quad x_1 + x_2 \geq 4.5, & (9) \\
 & \quad 3x_1 + x_2 \geq 9.5, & (10) \\
 & \quad 0.25x_1 - x_2 \geq -3, & (11) \\
 & \quad x \in \mathbb{Z}^2. & (12)
 \end{aligned}$$

In this example, we let

$$\begin{aligned}
 \mathcal{P} &= \text{conv} \left\{ x \in \mathbb{R}^2 \mid x \text{ satisfies (1) -- (12)} \right\}, \\
 \mathcal{Q}' &= \left\{ x \in \mathbb{R}^2 \mid x \text{ satisfies (1) -- (6)} \right\}, \\
 \mathcal{Q}'' &= \left\{ x \in \mathbb{R}^2 \mid x \text{ satisfies (7) -- (11)} \right\}, \text{ and} \\
 \mathcal{P}' &= \text{conv} \left(\mathcal{Q}' \cap \mathbb{Z}^2 \right).
 \end{aligned}$$

In Fig. 1a, we show the associated polyhedra, where the set of feasible solutions $\mathcal{F} = \mathcal{Q}' \cap \mathcal{Q}'' \cap \mathbb{Z}^2 = \mathcal{P}' \cap \mathcal{Q}'' \cap \mathbb{Z}^2$ and $\mathcal{P} = \text{conv}(\mathcal{F})$. Figure 1b depicts the continuous approximation $\mathcal{Q}' \cap \mathcal{Q}''$, while Fig. 1c shows the improved approximation $\mathcal{P}' \cap \mathcal{Q}''$. For the objective function in this example, optimization over $\mathcal{P}' \cap \mathcal{Q}''$ leads to an improvement over the LP bound obtained by optimization over $\mathcal{Q}$.

It is easy to see that we can only have $z_D > z_{LP}$ when $\mathcal{P}' \subset \mathcal{Q}'$. This is formalized in the following corollary of Theorem 1.

Corollary 1. *If $\mathcal{Q}'$ is an integral polyhedron, then $z_D = z_{LP}$.*

It is important to point out that this result does not depend on the objective function—if $\mathcal{Q}'$ is integral, then $z_D = z_{LP}$ for *any* objective function vector. On the other hand, the converse is not true. If $\mathcal{Q}'$ is not integral, we might still have $z_D = z_{LP}$ for *certain* objective functions. In the following example, we illustrate the result of the corollary.

Example 2 [GAP]. The GAP is that of finding a minimum cost assignment of n tasks to m machines such that each task is assigned to precisely one machine subject to capacity restrictions on the machines. With each possible assignment, we associate a binary variable x_{ij} , which, if set to one, indicates that machine i is assigned to task j . For ease of notation, let us define two index sets $M = \{1, \dots, m\}$ and $N = \{1, \dots, n\}$. Then an ILP formulation of GAP is as follows:

$$\begin{aligned}
 & \min \sum_{i \in M} \sum_{j \in N} c_{ij} x_{ij}, \\
 & \sum_{j \in N} w_{ij} x_{ij} \leq b_i \quad \forall i \in M, & (13) \\
 & \sum_{i \in M} x_{ij} = 1 \quad \forall j \in N, & (14) \\
 & x_{ij} \in \{0, 1\} \quad \forall i, j \in M \times N.
 \end{aligned}$$

In this formulation, Equation (14) ensures that each task is assigned to exactly one machine. Inequalities (13) ensure that for each machine, the capacity restrictions are met.

One possible decomposition of GAP is to let the relaxation be defined by the assignment constraints as follows:

$$\mathcal{Q}' = \{x_{ij} \in \mathbb{R}^+ \mid \forall i, j \in M \times N \mid x \text{ satisfies (14)}\}$$

and

$$\mathcal{Q}'' = \{x_{ij} \in \mathbb{R}^+ \forall i, j \in M \times N \mid x \text{ satisfies (13)}\}.$$

For this decomposition, the polytope $\mathcal{Q}'$ is integral, that is, $\mathcal{Q}' = \text{conv}(\mathcal{Q}' \cap \{0, 1\}^{M \times N})$. Hence, according to Corollary 1, the decomposition bound z_D is equal to that of the LP relaxation. If we instead choose a relaxation defined by the capacity constraints, we get the following:

$$\mathcal{Q}' = \{x_{ij} \in \mathbb{R}^+ \forall i, j \in M \times N \mid x \text{ satisfies (13)}\}$$

and

$$\mathcal{Q}'' = \{x_{ij} \in \mathbb{R}^+ \forall i, j \in M \times N \mid x \text{ satisfies (14)}\}.$$

In this case, the relaxation is a set of knapsack problems. These knapsack polyhedra do not have the integrality property, so the decomposition bound *may* be better than the LP bound. This is also an example of a decomposable subproblem—the optimization over $\mathcal{P}'$ reduces to a set of independent knapsack problems, one for each machine, which can be solved effectively in practice.

It should now be clear that these three methods are very closely related. In each of the three methods, we are given a polyhedron $\mathcal{P}$ over which we would like to optimize, along with two polyhedra that contain $\mathcal{P}$, the polyhedron $\mathcal{Q}''$, which has a small description and can be represented explicitly, and the polyhedron $\mathcal{P}'$, with a much larger description that must be represented implicitly. The conceptual difference between DWD, Lagrangian relaxation, and the cutting plane method is that both DWD and Lagrangian relaxation utilize an *inner representation* of $\mathcal{P}'$, generated dynamically by solving the corresponding optimization problem, whereas the cutting-plane method relies on an *outer representation* of $\mathcal{P}'$, generated dynamically by solving the corresponding separation problem. In this framework, most dynamic methods for solving ILPs can be viewed as decomposition methods.

In fact, there are even deeper connections between decomposition methods based on outer approximation and the more traditional methods based on inner representation

that serve to shed further light on the relationship. Recall that after solving (DWLP), we can obtain the optimal fractional solution $\hat{x}_{DW}$ corresponding to the optimal solution $\hat{\lambda}$ using the mapping (MAP). Note that $\hat{x}_{DW}$ is a convex combination of members of $\mathcal{F}'$. In particular, it is a combination of the members of the set $\{s \in \mathcal{F}' \mid \hat{\lambda}_s > 0\}$, which we will refer to as the *decomposition* of $\hat{x}_{DW}$ arising from $\hat{\lambda}$. It can be shown that $\hat{x}_{DW}$ must always be contained in a proper face of $\mathcal{P}'$ that we can characterize precisely as follows. Consider the set

$$S = \{s \in \mathcal{F}' \mid (c^\top - \hat{u}^\top A'')s = (c^\top - \hat{u}^\top A'')\hat{s}\}, \quad (15)$$

where $\hat{s}$ is some member of the decomposition (and hence corresponds to a basic variable in an optimal solution to (DWLP)). The following theorem states that the set S must contain all members of the decomposition. The proof is straightforward and can be found in Ralphs and Galati [35].

Theorem 2. $\{s \in \mathcal{F}' \mid \hat{\lambda}_s > 0\} \subseteq S$.

In fact, we can show something stronger. The set S is comprised exactly of those members of $\mathcal{F}'$ corresponding to columns of the (DWLP) with reduced cost zero, so we have the following theorem. The proof is again straightforward and can be found in Ralphs and Galati [35].

Theorem 3. $\text{conv}(S)$ is a face of $\mathcal{P}'$ and contains $\hat{x}_{DW}$.

An important consequence of this result is the following corollary, which states that the face of optimal solutions to (CPLP) is contained in $\text{conv}(S) \cap \mathcal{Q}''$.

Corollary 2. If F is the face of optimal solutions to (CPLP), then $F \subseteq \text{conv}(S) \cap \mathcal{Q}''$.

Hence, the convex hull of the decomposition is a subset of $\text{conv}(S)$ that contains $\hat{x}_{DW}$ and can be thought of as a surrogate for the face of optimal solutions to (CPLP).

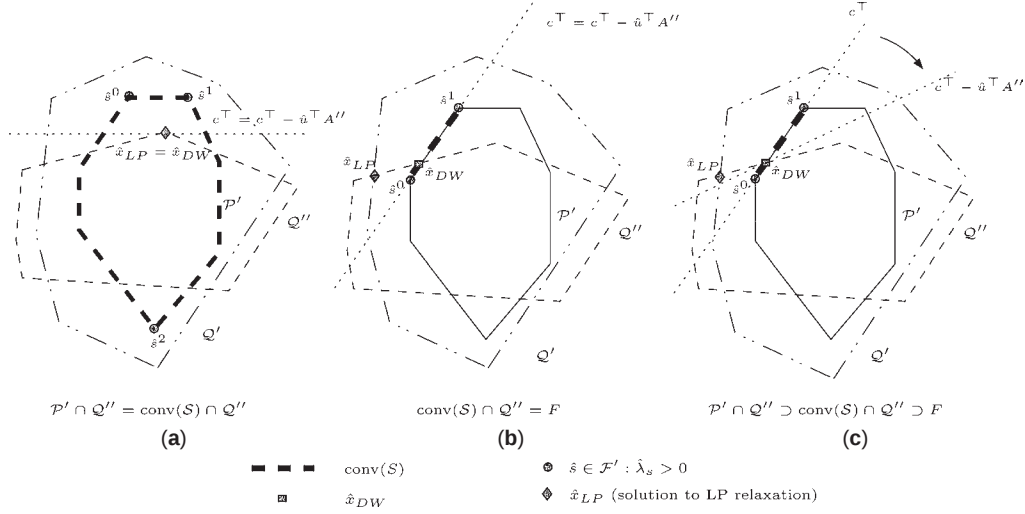


Figure 2. Illustration of bounds for different cost vectors.

As noted earlier, $\text{conv}(S)$ is typically a proper face of $\mathcal{P}'$. It is possible, however, for $\hat{x}_{DW}$ to be an inner point of $\mathcal{P}'$. In this case, illustrated graphically in Fig. 2a, $z_{DW} = z_{LP}$ and DWD does not improve the bound. All columns of (DWLP) have reduced cost zero and any extremal member of $\mathcal{F}'$ could be made a member of the decomposition. Hence, the effectiveness of the procedure is reduced and the chosen relaxation may be too weak. A necessary condition for an optimal fractional solution to be an inner point of $\mathcal{P}'$ is that the dual value of the convexity constraint in an optimal solution to the (DWLP) be zero. This result is further examined in the next section.

A second case of potential interest is when $F = \text{conv}(S) \cap \mathcal{Q}''$, illustrated graphically in Fig. 2b. This condition can be detected by examining the objective function values of the members of the decomposition. If they are all equal, then any member of the decomposition that is contained in $\mathcal{Q}''$ (if one exists) must be optimal for the original ILP, since it is feasible and has objective function value equal to z_D . In this case, all constraints of (DWLP) *other than* the convexity constraint must have dual value zero, since removing them does not change the optimal solution value. The more typical case, in which F is

a proper subset of $\text{conv}(S) \cap \mathcal{Q}''$, is shown in Fig. 2c.

ADVANCED METHODS

As previously mentioned, the cutting-plane method has a number of advantages in practice over traditional decomposition approaches. Within the framework presented here, the generation of valid inequalities can be interpreted as a dynamic tightening of either $\mathcal{Q}''$ or $\mathcal{P}'$. We will focus on the former interpretation and discuss two methods that employ either (DWLP) or (LD) as the bounding subproblem, but also integrate generation of valid inequalities.

We first consider the basic steps of the cutting-plane method, assuming the problem at hand is feasible and bounded. To begin, we construct an initial LP relaxation and solve it. We then perform the separation step, attempting to generate inequalities valid for $\mathcal{P}$ but violated by the resulting solution. If successful, we add these inequalities to the LP relaxation, hoping to improve the bound. This procedure is iterated until no more violated inequalities are found. The most important step is that of separation, which results in a tightening of $\mathcal{Q}''$ and (hopefully) an improved bound.

Decomposition with Cut Generation

1. Construct the initial bounding subproblem P^0 and set $i \leftarrow 0$.

$$z_{CP} = \min_{x \in \mathcal{P}'} \{c^\top x \mid A''x \geq b''\}$$

$$z_{LR} = \max_{u \in \mathbb{R}_+^m} \min_{x \in \mathcal{P}'} \{(c^\top - u^\top A'')x + u^\top b''\}$$

$$z_{DW} = \min_{\lambda \in \mathbb{R}_+^s} \{c^\top (\sum_{s \in \mathcal{F}'} s\lambda_s) \mid A''(\sum_{s \in \mathcal{F}'} s\lambda_s) \geq b'', \sum_{s \in \mathcal{F}'} \lambda_s = 1\}$$

2. Solve P^i to obtain a valid lower bound z^i .
3. Generate a set of *improving inequalities* $[D^i, d^i]$ valid for $\mathcal{P}$.
4. If valid inequalities were found in Step 3, form the bounding subproblem P^{i+1} by setting $[A'', b''] \leftarrow [\frac{A''}{D^i}, \frac{b''}{d^i}]$. Then, set $i \leftarrow i + 1$ and go to Step 2.
5. If no valid inequalities were found in Step 3, then output z^i .

Figure 3. Basic outline of the dynamic decomposition method.

In principle, this procedure can be generalized by replacing (CPLP) with either (DWLP) or (LD) as the bounding subproblem. The steps of this generalized method are shown in Fig. 3. The important step is Step 3, that of generating a set of *improving inequalities*; that is, inequalities valid for $\mathcal{P}$ that when added to the description of the $\mathcal{P}'$ result in an increase in the computed bound. Aside from Step 3, the method is straightforward. Steps 1 and 2 are performed as in a traditional decomposition framework. Step 4 is accomplished by simply adding the newly generated inequalities to the list $[A'', b'']$ and reforming the appropriate bounding subproblem.

It is important to point out that the generalized methods described in Fig. 3 require that we can solve the subproblem (LR) for *any* objective function. In some cases, the algorithm used for solving the subproblem may be efficient only when the objective function has certain structure. In such cases, it may only be possible to integrate generation of classes of valid inequalities whose addition does not destroy the required structure of the subproblem.

In the cutting-plane method, techniques for generating valid inequalities typically measure the potential effectiveness of adding a given valid inequality by the degree to which it violates $\hat{x}_{DW}$, since violation of $\hat{x}_{DW}$ is a necessary condition for an inequality to be improving. It is not clear, however, what

criteria should be used to measure potential effectiveness if the bounding subproblem is either (DWLP) or (LD). In the next two sections, we discuss these two cases in more detail.

Branch, Price, and Cut

We first consider using (DWLP) as the bounding subproblem in the procedure of Fig. 3, a procedure we call *price and cut*. When employed as the bounding procedure in a branch-and-bound framework, the overall technique is called *branch, price, and cut* and has been studied by a number of authors [10, 36–38]. Simultaneous generation of columns and valid inequalities is difficult in general because the addition of valid inequalities may destroy the structure of the column generation subproblem (for a discussion, see van den Akker *et al.* [10]). Having solved (DWLP), however, we can easily recover an optimal solution to (CPLP) using (MAP) and try to generate improving inequalities using standard separation methods for $\mathcal{P}$. Because the generation of these valid inequalities takes place in the space of the original formulation, it does not destroy the structure of the column generation subproblem (with the possible exception of the objective function, as discussed above). Hence, the approach enables dynamic generation of valid inequalities while still retaining the bound improvement

and other advantages yielded by a DWD. It is important to note, however, that the case of block decompositions with identical subproblems presents unique challenges to the implementation of price and cut, since collapsing the identical subproblems into a single subproblem means that the mapping (MAP) cannot be used to provide a unique optimal fractional solution; see Vanderbeck [28] for a discussion.

The same bound improvement obtained in price and cut could be realized in the cutting-plane method by adding the generated inequalities directly to (CPLP). Despite the apparent symmetry, there is one fundamental difference between the cutting-plane method and price and cut. An optimal solution to (DWLP) provides a decomposition of $\hat{x}_{\text{DW}}$ (MAP) into a convex combination of members of $\mathcal{F}'$. In particular, the weight on each member of $\mathcal{F}'$ in the combination is given by the optimal decomposition.

An important observation is that an inequality can be improving only if it is violated by at least one member of the decomposition. This follows easily from the fact that the degree of violation of $\hat{x}_{\text{DW}}$ is a convex combination of the degrees of violation of each member of the decomposition. In some cases, it is much easier to separate members of $\mathcal{F}'$ from $\mathcal{P}$ than to separate arbitrary real vectors. In fact, it is easy to find polyhedra for which the problem of separating an arbitrary real vector is difficult, but the problem of separating a solution to a given combinatorial relaxation is easy. This notion has been discussed in the literature in several contexts [14,39,40] and leads to a separation technique that can be embedded within price and cut [35]. The idea is to replace direct separation of the fractional point (which may be difficult) with separation of members of the decomposition, which act as surrogates. Given a decomposition, the efficiency of this procedure depends both on the cardinality of the decomposition (which must be less than or equal to $\dim(\mathcal{P}') + 1$) and the time required to separate members of $\mathcal{P}'$ from $\mathcal{P}$.

Branch, Relax, and Cut

We now move on to discuss the dynamic generation of valid inequalities when the bounding subproblem is (LD). When employed in a branch-and-bound framework, the overall method is called *branch, relax, and cut* and has also previously been studied by several authors (see Lucena [39] for a survey). Solving the LD as the linear program (LDLP) is equivalent to solving (DWLP), so, in this case, the methods just discussed apply directly without modification. We assume from here on that the LD is solved in the form (LD), using subgradient optimization.

Suppose we have solved (LD) to obtain $\hat{u}$, a vector of optimal multipliers and $\hat{s} = \operatorname{argmin} z_{\text{LR}}(\hat{u})$. As before, let $\hat{\lambda}$ be an optimal decomposition and $\hat{x}_{\text{DW}} = \sum_{s \in \mathcal{F}'} \hat{\lambda}_s \hat{s}$ be an optimal fractional solution. The main consequence of using subgradient optimization to solve (LD) is that we do not obtain the primal solution information available to us with both the cutting-plane method and price and cut. This means there is no obvious way of constructing an optimal fractional solution or verifying that any generated valid inequalities would be violated by such a solution. We can, however, attempt to separate the solution to the Lagrangian relaxation (LR), which is a member of $\mathcal{F}'$, from $\mathcal{P}$. Note that again, we are taking advantage of our ability to separate members of $\mathcal{F}'$ from $\mathcal{P}$ effectively. If successful, we immediately “dualize” this new constraint by adding it to $[A'', b'']$. This has the effect of introducing a new dual multiplier and slightly perturbing the objective function used to solve the Lagrangian relaxation.

As with both the previously discussed methods, the difficulty with relax and cut is that the valid inequalities generated by separating $\hat{s}$ from $\mathcal{P}$ may not be improving, as Guignard first observed in her work [41]. As we have already noted, we cannot verify that generated inequalities are violated by an optimal fractional solution, which is the usually employed necessary condition for an inequality to be improving. To deepen our understanding of the potential effectiveness of the valid inequalities generated during relax and cut, we now further examine the relationship between $\hat{s}$ and $\hat{x}_{\text{DW}}$.

From the reformulation of the LD as the linear program (LDLP), we can see that each constraint binding at an optimal solution corresponds to an alternative optimal solution to the Lagrangian subproblem with multipliers $\hat{u}$. The binding constraints of (LDLP) correspond to variables of (DWLP) with reduced cost zero, so it follows immediately that the set of all alternative solutions to the Lagrangian subproblem with multipliers $\hat{u}$ is the set S from Equation (15).

Because $\hat{x}_{\text{DW}}$ is both an optimal solution to (CPLP) and is contained in $\text{conv}(S)$, it also follows that

$$c^\top \hat{x}_{\text{DW}} = (c^\top - \hat{u}^\top A'') \hat{x}_{\text{DW}} + \hat{u}^\top b''.$$

In other words, the penalty term in the objective function of the Lagrangian subproblem (LR) serves to rotate the original objective function so that it becomes parallel to the face $\text{conv}(S)$, while the constant term $\hat{u}^\top b''$ ensures that $\hat{x}_{\text{DW}}$ has the same cost with respect to both the original and the Lagrangian objective function. This is shown in Fig. 2c.

The fundamental connection between relax and cut and price and cut is that solving (DWLP) produces a *set* of alternative optimal solutions to the LD, at least one of which must be violated by a given improving inequality. This yields a verifiable necessary condition for a generated inequality to be improving. Relax and cut, on the other hand, produces only one member of this set, though possibly at a much lower computational cost. Even if improving inequalities exist, it is possible that none of them are violated by $\hat{s}$, especially if $\hat{s}$ has a small weight in the optimal decomposition. As in price and cut, when $\hat{x}_{\text{DW}}$ is an inner point of $\mathcal{P}'$, the decomposition does not improve the bound and all members of $\mathcal{F}'$ are alternative optimal solutions to the Lagrangian subproblem. This situation is depicted in Fig. 2a. In this case, separating $\hat{s}$ is unlikely to yield an improving inequality.

CONCLUSION

Decomposition algorithms capitalize on our knowledge of identifiable substructure in

a given integer programming model. This knowledge could be in the form of either a method of optimizing over or separating from the convex hull of solutions to a given relaxation. In the former case, an inner representation of the convex hull can be implicitly used in either a DWD or Lagrangian relaxation framework. In the latter case, an outer representation can be used in a cutting-plane framework. The use of advanced techniques that integrate generation of valid inequalities with the use of either DWD or Lagrangian relaxation is an attempt to achieve the “best of both worlds.” Implementations of these advanced algorithms are becoming more common, and this trend is sure to continue as more software tools become available.

REFERENCES

1. Ford LR, Fulkerson DR. A suggested computation for maximal multicommodity network flows. *Manage Sci* 1958;5:97–101.
2. Gilmore PC, Gomory RE. A linear programming approach to the cutting stock problem. *Oper Res* 1961;9:849–859.
3. Vance PH, Barnhart C, Johnson EL, *et al.* Solving binary cutting stock problems by column generation and branch and bound. *Comput Optim Appl* 1994;3:111–130.
4. Held M, Karp RM. The traveling salesman problem and minimum spanning trees. *Oper Res* 1970;18:1138–1162.
5. Guignard M, Rosenwein M. An improved dual-based algorithm for the generalized assignment problem. *Oper Res* 1989;37:658–663.
6. Savelsbergh M. A branch-and-price algorithm for the generalized assignment problem. *Oper Res* 1997;45(6):831–841.
7. Vanderbeck F. Computational study of a column generation algorithm for bin packing and cutting stock problems. *Math Program* 1999; 86(3):565–594.
8. Balas E, Saltzman M. Facets of the three-index assignment polytope. *Discrete Appl Math* 1989;23:201–229.
9. Lucena A. Steiner problem in graphs: Lagrangean relaxation and cutting planes. *COAL Bull* 1992;28:2–8.
10. van den Akker J, Hurkens C, Savelsbergh M. Time-indexed formulations for machine scheduling problems: Column generation. *INFORMS J Comput* 2000;12:111–124.

11. Mehrotra A, Trick M. A column generation approach to graph coloring. *INFORMS J Comput* 1996;8(4):344–354.
12. Agarwal Y, Mathur K, Salkin HM. A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks* 1989;19:731–749.
13. Fisher M. Optimal solution of vehicle routing problems using minimum k-trees. *Oper Res* 1994;42(4):626–642.
14. Ralphs T, Kopman L, Pulleyblank W, *et al.* On the capacitated vehicle routing problem. *Math Program* 2003;94:343–359.
15. Fukasawa R, Longo H, Lysgaard J, *et al.* Robust branch-and-cut-and-price for the capacitated vehicle routing problem. *Math Program* 2006;106:491–511.
16. Vanderbeck F, Wolsey L. Reformulation and decomposition of integer programs. In: Jünger M, Liebling ThM, Naddef D, *et al.*, editors. 50 years of integer programming 1958–2008. New York: Springer; 2010; pp. 431–498.
17. Fisher M. The Lagrangian relaxation method for solving integer programming problems. *Manage Sci* 1981;27:1–18.
18. Beasley J. Lagrangean relaxation. In: Reeves C, editor. Modern Heuristic techniques for combinatorial optimization. New York: John Wiley & Son; 1993.
19. Neame PJ. Nonsmooth dual methods in integer programming [PhD thesis]. University of Melbourne; 1999.
20. Lemaréchal C. Lagrangean relaxation. In: Jünger M, Naddef D, editors. Computational combinatorial optimization. New York: Springer; 2001. pp. 112–156.
21. Kelley JE. The cutting plane method for solving convex programs. *J SIAM* 1960;8: 703–712.
22. Nemhauser G, Wolsey L. Integer and combinatorial optimization. New York: John Wiley & Sons; 1988.
23. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;8(1): 101–111.
24. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch and price: column generation for solving huge integer programs. *Oper Res* 1998;46:316–329.
25. Lübbecke M, Desrosier J. Selected topics in column generation. *Oper Res* 2005;53(6): 1007–1023.
26. Villeneuve D, Desrosiers J, Lübbecke M, *et al.* On compact formulations for integer program solved by column generation. *Ann Oper Res* 2005;139:375–388.
27. Ryan D, Foster B. An integer programming approach to scheduling. In: Wren A, editor. Computer scheduling of public transport urban passenger vehicle and crew scheduling. Amsterdam: North-Holland Publishing Co.; 1981. pp. 269–280.
28. Vanderbeck F. Branching in branch-and-price: a generic scheme. *Math Program* 2010. To appear, available at <http://www.springerlink.com/content/2572975068528215/>.
29. Parker R, Rardin R. Discrete optimization. San Diego (CA): Academic Press, Inc.; 1988.
30. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Mon* 1958;64:275–278.
31. Dantzig GB, Fulkerson DR, Johnson SM. Solution of a large scale traveling salesman problem. *Oper Res* 1954;2:393–410.
32. Padberg MW, Rinaldi G. A branch and cut algorithm for the solution of large scale traveling salesman problems. *SIAM Rev* 1991;33: 60–100.
33. Hoffman K, Padberg M. LP-based combinatorial problem solving. *Ann Oper Res* 1985;4: 145–194.
34. Geoffrion A. Lagrangian relaxation for integer programming. *Math Program Study* 1974;2: 82–114.
35. Ralphs TK, Galati MV. Decomposition and dynamic cut generation in integer linear programming. *Math Program* 2006;106(2): 261–285.
36. Vanderbeck F. Lot-sizing with start-up times. *Manage Sci* 1998;44:1409–1425.
37. Kohl N, Desrosiers J, Madsen O, *et al.* 2-path cuts for the vehicle routing problem with time windows. *Transp Sci* 1999;33:101–116.
38. Barnhart C, Hane CA, Vance PH. Using branch-and-price-and-cut to solve origin-destination integer multi-commodity flow problems. *Oper Res* 2000;48:318–326.
39. Lucena A. Non delayed relax-and-cut algorithms. *Ann Oper Res* 2005;140(1):375–410.
40. Balas E, Saltzman M. An algorithm for the three-index assignment problem. *Oper Res* 1991;39:150–161.
41. Guignard M. Efficient cuts in Lagrangean “relax-and-cut” schemes. *Eur J Oper Res* 1998;105:216–223.
42. Galati MV. Decomposition in integer programming [PhD thesis]. Lehigh University; 2009.

DEFINING OBJECTIVES AND CRITERIA FOR DECISION PROBLEMS

JASON R. W. MERRICK
Statistical Sciences and
Operations Research,
Virginia Commonwealth
University, Richmond,
Virginia

INTRODUCTION

Clemen [1] defines an objective as “a specific thing that you want to achieve.” We write an objective in terms of the noun describing the quality of an alternative that we value and a verb indicating the direction in which we would like to improve this quality. For instance, in making many business decisions, we value the monetary profits that an alternative may lead to and we wish to maximize these profits. Thus, profit is the quality of an alternative and maximize is the verb indicating our direction of preference. We may also wish to maximize the market share for our products, minimize impacts on the environment, and maximize customer satisfaction.

An objective is a reason to value one alternative over another. Without values, all alternatives are equivalent and there is no decision. As values are more fundamental to a decision than alternatives, Keeney [2] suggests that we should consider values before we consider available alternatives. Considering our values gives us the ability to conceptualize the structure of a decision by considering the objectives a decision maker seeks to achieve. We can then consider whether such objectives are a means to an end or fundamental to the current decision.

However, in the end we wish to achieve our objectives by taking the best action or choosing the best alternative. To find the best alternative, we must consider the widest range of alternatives and discriminate between them

in the most comprehensive and thoughtful way. We want to have as wide a range of focus as possible to expand our search for the best alternative [3]. Keller and Ho [4] point out that we may use heuristics and biases to simplify the task of generating alternatives. Using the availability heuristic, we may simply recall the alternatives most readily available in memory. Alternatively, we may use the representativeness heuristic and recall a list of alternatives that are representative of similar decisions. Each of these heuristics narrows the focus of our consideration to just those alternatives readily available in memory. Jungermann *et al.* [5] show that the best way to expand our focus is to have the largest set of objectives possible. Gettys *et al.* [6] show that the set of objectives should be as complete and hierarchical as possible if we are to generate high utility actions or alternatives. Leon [7] claims five advantages to using a comprehensive set of objectives in choosing amongst alternatives.

1. Alternatives with more innovative characteristics are considered.
2. The range of alternatives considered becomes wider.
3. The future consequences of decisions are taken into account.
4. Alternatives are integrated, which at first glance would not be considered.
5. More desirable consequences are considered.

The purpose of this article is to examine how well decision makers can generate a comprehensive set of objectives and to offer methods to improve the generation process.

We start by introducing a framework for thinking about and classifying objectives in the section titled “The Decision Context.” We then examine the problems faced in the generation of objectives and examine methods for lessening these problems in the section titled “What is Good Set of Objectives?”

THE DECISION CONTEXT

Keeney uses the concept of a *decision context* to explain the decisions that people make. A decision context consists of a decision maker's set of alternatives and the objectives that the decision maker is attempting to achieve when choosing. The objectives can be categorized as means, fundamental, and strategic. A *means objective* is one way to achieve another objective. A *fundamental objective* is an objective that governs a decision maker's choice in a particular decision context. A *strategic objective* is one that reflects the long term goals of a decision maker's organizational setting; many decisions in an organization may affect performance on a strategic objective. Keeney [8] states that misunderstanding the decision context is a common mistake in making trade-offs between alternatives.

We do not wish to measure achievement on an objective that is just one means to achieve what we value. Nor do we wish to measure achievement on an objective that is affected by decisions other than this one. Thus, we evaluate each alternative by how well they achieve the decision maker's fundamental objectives. However, that is not to say that means objectives and strategic objectives are not useful. Means objectives give us different means to achieving our fundamental objectives, so they are useful in developing alternatives that we may not otherwise think of. Strategic objectives are useful to make sure that each decision problem is aimed at our overall values.

Our objectives will also change depending on the decision context that we consider. Consider choosing a car to buy. Two fundamental objectives might be minimizing purchase cost and minimizing monthly fuel costs. However, if we broaden the decision frame to how to get to work, then we extend the alternatives to include different cars and different routes, as well as public transportation and carpooling options. Now minimizing monthly commuting costs becomes fundamental for the broader decision context.

It is important to understand whose values and perspectives should be considered. Keeney [8] uses the example of a public

utility company and two major objectives: minimize electricity costs to ratepayers and maximize returns to company shareholders. If faced with alternatives that lead to a low cost and a low return or a high cost and a high return, it is important to understand whose perspective is being considered. The ratepayers would prefer the former, the company shareholders would prefer the latter, and the public utility commission would prefer a trade-off between the two.

Keeney notes that it is also important to understand the timeline of the decision. Suppose you are deciding between two job offers. Two important objectives would be to maximize salary and minimize hours worked per week. Would you choose a job with a high salary that required 80 h of work each week or one with a lower salary that only required 50 h a week? Timeline is important in this decision. Many people would choose the high salary job with long work hours if this lasted for a limited number of years or months, like residency for a doctor or early associate work for lawyers and management consultants. However, this choice would be less desirable if there was no end to the long work hours.

WHAT IS A GOOD SET OF OBJECTIVES?

Keeney [9] proposes five criteria for a good set of objectives.

1. Completeness
2. Nonredundancy
3. Conciseness
4. Specificity
5. Understandability

The set of objectives must be complete in the sense that all possible angles of the problem are considered. If we omit an important objective, then we may choose the wrong alternative, or at least not foresee all the consequences of our chosen alternative. For example, consider buying a car and considering purchase cost, ongoing costs, visual appeal, reliability, drivability, and cargo space, but not comfort.

We wish to avoid redundant objectives, as this will lead to double counting the benefits

of an alternative. For example, in our car example we should not include maximize miles per gallon and minimize monthly fuel costs as objectives, as these quantities overlap. However, a car could have a high miles per gallon rating, but require an expensive grade of fuel. In this case, we could use maximize miles per gallon and minimize price per gallon of fuel required. Or, we could simplify to minimize monthly fuel costs.

A set of objectives must be concise. A set that is too large can lead to “paralysis by analysis,” as we are bogged down by considering the many objectives. Imagine you are hungry and are deciding where to eat dinner; considering a hundred objectives would probably be more frustrating than helpful.

We must be able to measure how well each alternative performs on achieving each objective or at least compare the alternatives under each objective. This means that our objectives must be specific and understandable to remove ambiguity in their measurement. Some examples of objectives that are hard to measure are maximize aesthetic appeal or minimize pain level. Research has been performed that has led to reliable measures of even these objectives by specific definitions, but the level of effort required should be commensurate with the importance of the decision being made. Developing attributes to measure objectives will be discussed at more length in the article titled ***Problem Structuring for Multicriteria Decision Analysis Interventions*** in this encyclopedia.

The set of objectives should also be decomposable so that we may consider each part separately, simplifying the judgment task and leading to a better appraisal of the alternatives. This criterion requires that the performance of an alternative on each objective can be judged independently of all other objectives. For instance, in choosing a hotel room, can you consider comfort and size of the room separately, or we must define our objectives as the comfort of the furniture and the size of the room to decompose the pieces of this assessment. This concept will be discussed at length in the article titled ***Problem Structuring for Multicriteria Decision Analysis Interventions*** in this encyclopedia.

PROBLEMS WITH GENERATING OBJECTIVES

We wish to determine the decision maker’s objectives for a given decision context. The objectives are generated through interviews with decision makers and stakeholders. However, decision makers are not generally good at stating the objectives that they are trying to achieve. Bond *et al.* [10] note that decision makers attempt to simplify their environment and avoid the actual complexity of the decisions they make. This means that the objectives they offer are cued by “an incomplete mental representation of the situation” [10].

Bond *et al.* performed three studies to test the ability of decision makers to generate objectives. In each case, the experimenters first asked the subjects to generate a list of objectives for a specified decision context. The subjects were then given a pregenerated list of objectives that was considered complete. The experimenters asked the subjects to check the objectives that they saw as important and then to indicate which of the list were objectives they had developed before seeing it. The objectives that they had checked off but did not develop before seeing the list were called the *recognized* objectives. The objectives that they checked off and had developed before seeing the list were called the *self-generated* objectives. Lastly, they were asked to assess the importance of all the objectives that they had checked off, both recognized and generated.

In the first experiment, MBA students were asked to consider the choice of a program to join. They completed the exercise in one sitting, ending with a ranking of the objectives. In all, the study obtained 62 usable responses. The second experiment was identical to the first in all but one aspect. The ranking of objectives was asked one week later to reduce any bias toward the self-generated objectives. Seventy usable responses were obtained. For the third experiment, the decision context changed to the choice of a summer internship. The MBA students completed the same tasks, but the experimenters gave them one week to generate their own list of objectives. The ranking of objectives was also replaced with

a rating on a scale from 1 (not important) to 10 (extremely important). Twenty-eight responses were obtained.

For the first experiment, the subjects generated 7.4 objectives on average, while 7.6 were recognized. For the second experiment, the averages were 5.9 self-generated objectives and 7.7 recognized. For the third experiment, an average of 6.8 objectives were self-generated and 7.9 were recognized. The mean number of recognized objectives is statistically significantly higher than the mean number of self-generated objectives in each experiment.

In the first experiment, 71% of the subjects overlooked at least one of their top five objectives and 79% in the second. Interestingly, 35% of subjects did not self-generate their most important objective in the second experiment compared to 10% in the first experiment. In the third experiment, the average importance rating for self-generated objectives (7.4) was significantly higher than the average importance for the recognized objectives (6.8). Thus, while a significant number of objectives were not self-generated, the subjects did self-generate objectives that are more important. However, the difference in the importance rating is not that large in a practical sense.

In a final real-world case study, Bond *et al.* [10] reanalyzed the results of a professional elicitation performed for Seagate Software [11]. The professional facilitator worked with 12 individual decision makers in the company to generate their list of objectives, and then the results were combined into a complete value hierarchy containing 39 specific objectives grouped into 8 categories. On average, the subjects generated at least one specific objective in 5.16 of the 8 categories. They generated an average of 14.17 specific objectives of the 39, which is 36%. Only one subject generated more than half (20 or more). Thus, these three experiments and one case study show that decision makers will not generate a complete list of objectives on their own.

Bond *et al.* [12] reanalyzed the experimental results from their earlier study [10]. The master lists of objectives for each of the three experiments were categorized into four sets of related objectives. The number

of generated objectives and the number of recognized objectives were counted by group for each subject. The authors concluded that the number of self-generated objectives was significantly less than the total number of checked objectives in each category, so the subjects were not thinking with sufficient depth in any category. However, the authors also found that the category in which each subject generated the most objectives had significantly more than would be expected if the generation was independent of category. The category for which each subject generated the least objectives also had significantly less than would be expected. Thus, the subjects were not thinking with sufficient breadth, concentrating on a few categories and ignoring others. This implies an anchoring effect within a subset of the possible categories.

Leon [7] describes a similar study, but instead compares a decision maker's ability to generate objectives when they focus on specific alternatives or when they focus on values and objectives. The subjects of the generation study were 28 psychology students with some training in decision analysis. The experimenter asked the subjects to generate objectives to use in choosing amongst advanced elective courses. A randomly assigned 14 students were asked to generate objectives to help in the decision while considering a predefined list of elective courses. These were the alternative-focused group. The remaining students were asked to generate objectives using a set of value-focused questions adapted from Keeney [2] and were called the *value-focused* group. Each subject was then asked whether each objective they generated was a specific objective or whether it was a general objective made up of lower level specific objectives.

The value-focused group generated a total of 28 different objectives, while the alternative-focused group generated 22. Leon [7] then developed representative value hierarchies for each group consisting of the average number of general objectives and the average number of specific objectives, with the actual objectives chosen by ranking the number of decision makers in the group that generated them. Thus, the value-focused representative hierarchy contained the top

three most prevalent general objectives and the top nine most prevalent specific objectives in that group. The alternative-focused representative hierarchy contained one overall general group and the top five most prevalent specific objectives in that group. Thus, the value-focused group developed a larger set of objectives and organized them hierarchically. Qualitative consideration of the hierarchies revealed that all the issues covered in the alternative-focused representative hierarchy were covered in the value-focused representative hierarchy, but not vice versa. Thus, the value-focused hierarchy was more complete.

In a follow-on study, Leon asked a different set of 30 decision makers to assess the two representative hierarchies from the first study. The second set of subjects assessed which of the two hierarchies better met five criteria for a useful hierarchy. In each case, the results were statistically significantly in favor of the value-focused hierarchy. The 30 decision makers were also asked to rate the two hierarchies in terms of completeness, ability to operationalize, conciseness, and understandability. The results were statistically significantly in favor of the value-focused hierarchy in all but conciseness, where the results were not significant in either direction.

In summary, decision makers are not good at generating a large and complete set of objectives. They get anchored in a few categories of objectives, not thinking with sufficient breadth. They also do not think with sufficient depth in any category. However, they are better if they focus on values than on alternatives. The methods discussed in the next section are aimed at broadening the consideration leading to a more complete set of objectives. Leon [7] and Bond *et al.* [10] both show that combining the objectives generated by a number of individuals leads to the most complete set. Thus, we do not hold individual interviews with individual decision makers, but discuss their decisions as a group. This allows the various decision makers to increase each other's perspective of the situation and broaden each other's view. This also leads to a more complete set of objectives.

METHODS FOR GENERATING OBJECTIVES

Keeney [13] offers the following list of devices to use in generating objectives:

1. *A Wish List.* What do you want? What do you value? What should you want?
2. *Alternatives.* What is a perfect alternative, a terrible alternative, some reasonable alternative? What is good or bad about each?
3. *Problems and Shortcomings.* What is wrong or right with your organization? What needs fixing?
4. *Consequences.* What has occurred that was good or bad? What might occur that you care about?
5. *Goals, Constraints, and Guidelines.* What are your aspirations? What limitations are placed upon you?
6. *Different Perspectives.* What would your competitor or your constituency be concerned about? At some time in the future, what would concern you?
7. *Strategic Objectives.* What are your ultimate objectives? What are your values that are fundamental?
8. *Generic Objectives.* What objectives do you have for your customers, your employees, your shareholders, and yourself? What environmental, social, economic, or health and safety objectives are important?
9. *Structuring Objectives.* Follow means–ends relationships: why is that objective important, how can you achieve it? Use specification: what do you mean by this objective?
10. *Quantifying Objectives.* How would you measure achievement of this objective? Why is objective A three times as important as objective B?

Let us consider a facilitation performed by the author to see the benefits of following this list [14]. The goal of the facilitation was to understand the objectives of decision makers in an oil shipping company as they related to safety performance. We will overview the techniques that we used in our interviews.

1. *A Wish List.* We asked a group of crew members from various vessels what they look for when they join a new crew; if they were to describe a perfectly operating vessel and crew, what would it look like. The crew described a perfect vessel and crew, and we examined the factors that they saw exemplified such a vessel. We also examined what exemplified an unsafe vessel and crew.
2. *Alternatives.* We asked the crew members about vessels they had worked on and previous crews they had worked with; what was good or bad about them? The crew members told good and bad stories about crews they had worked with and what factors affected their performance.
3. *Problems and Shortcomings.* We talked through various problems they encountered and what could be done to avoid them.
4. *Consequences.* We asked the crew what accidents or near misses had occurred on vessels they served on and what could have helped avoid these events.
5. *Goals, Constraints, and Guidelines.* We examined safety procedures in the company and safety regulations, both federal and international, to see what objectives they set minimum standards for.
6. *Different Perspectives.* We included bridge crew, deck crew, and engineering crew in our interviews. We included senior and junior personnel. We also included vessel inspectors and shore side managers.
7. *Strategic Objectives.* This includes working through means to achieving reduction in accidents, near misses, human errors, mechanical failures, and other precursor events.
8. *Generic Objectives.* We considered basic considerations of the company's mission statement and other generic policies.
9. *Structuring Objectives.* We considered each generated objective and what it was a means to achieving. We then followed the chains of means-ends

relationships to generate additional objectives. Training arose frequently in the discussions, but training is a means to achieving improved performance, so we examined different training programs and what they sought to achieve.

10. *Quantifying Objectives.* Here we considered how one could measure each objective to further specify what each meant. We also considered the metrics they had used and the data they collected to see what objectives they measured.

Leon [7] offers a list of questions to help decision makers discern their objectives. The relationship between the questions and the list of devices above is apparent on inspection.

1. What do you look for in a ____?
2. What is really important when you think about using and enjoying ____?
3. What do you want to obtain?
4. What need is not covered now that would be covered?
5. Once acquired, what are the desirable consequences of its use and enjoyment?
6. What are the undesirable consequences?
7. What are the restrictions on its acquisition?
8. If there were no restrictions, which alternative would you choose? What characteristic makes it the one you would choose?
9. Which alternative would you definitely not choose and why?
10. How does the use of a help you in terms of your objectives or fundamental values in life?
11. When you think of the choice, what aspects do you think should be taken into account?
12. When you think of several alternatives, which differences are the most obvious?
13. From the point of view of users, if you had to carry out a study to their satisfaction, which aspects would you take into account?

Interviews with decision makers are not the only way to develop objectives. Parnell *et al.* [15] calls decision maker and stakeholder interviews the “Platinum Standard” technique. Other approaches would be to base the development on documentation approved by the decision makers on vision, policy, or strategy, “Gold Standard,” or to use interviews with people who will inform or assist the actual decision maker, “Silver Standard.” The same devices can be used while examining the literature for objectives or interviewing people who provide decision support, but do not make the actual decision.

Bond *et al.* [12] test techniques for improving the breadth and depth of objectives generation. In their first experiment, subjects were asked to generate a list of objectives for writing their dissertation. They were then asked to expand on their list in a second session and given category cues to attempt to broaden their consideration. In the second experiment, subjects were asked to generate objectives for choosing an MBA program and given category cues and example objectives in each category in a single session. In the third experiment, subjects were asked to generate objectives for choosing an internship. They were then asked to expand on their list in a second session by adding a specified number of additional objectives. Thus, the first and third experiments had two sessions, the first unaided and the second using different interventions. The intervention was used straightaway in the second experiment. In the first two experiments the intervention was category cues, while in the third experiment the intervention was asking for specific numbers of additional objectives.

The results of the three experiments revealed several interesting conclusions. First, providing example objectives in the second experiment anchored the subjects and hindered their generation of objectives. Secondly, providing the category cues straightaway was ineffective, but it was effective when used in a second session asking for an expanded list. Thirdly, asking for a specific number of additional objectives was effective in the third experiment, whereas just asking for more is not. Fourthly, stating that subjects are generally able to add more

objectives sets an expectation of performance that the subjects strive to achieve. As subjects in these experiments are generally able to generate 30–50% of the master list, asking for as many again as the first session appears appropriate.

AN EXAMPLE FROM ENVIRONMENTAL MANAGEMENT

A set of objectives was developed for the Upham Brook Watershed to demonstrate how this technique can be applied to a complex environmental management problem [16,17]. The interdisciplinary team, composed of faculty members and graduate students from Virginia Commonwealth University assisted by a hydrologist from the U.S. Geological Survey (USGS), a water quality regulator from DEQ, a community organizer, and a wetlands specialist, all participated in the development of the model. The group met every two weeks for eight months, each session lasting 2h. The author of this article facilitated the sessions and supplied the understanding of the techniques used. While the group had significant substantive expertise, the members of the team were not stakeholders. Thus, this is a silver standard example in the terminology of Parnell *et al.* [15].

The first step in the process is to define the overall objective for the project. The overall objective selected was “Maximize the quality of the Upham Brook Watershed.” The decision context in this case is necessarily broad. The second step in the process is to establish a set of objectives that further defines what is meant by the overall objective.

Each member of the interdisciplinary group identified and wrote down on post-it notes 10–15 action verbs and nouns that define objectives for improving the quality of the watershed. Similar objectives were grouped together to form an affinity diagram and to determine the objectives of the group (Fig. 1). Keeney’s 10 devices from the previous section were used sequentially through the session to assist each individual in generating as broad a set of objectives as possible.

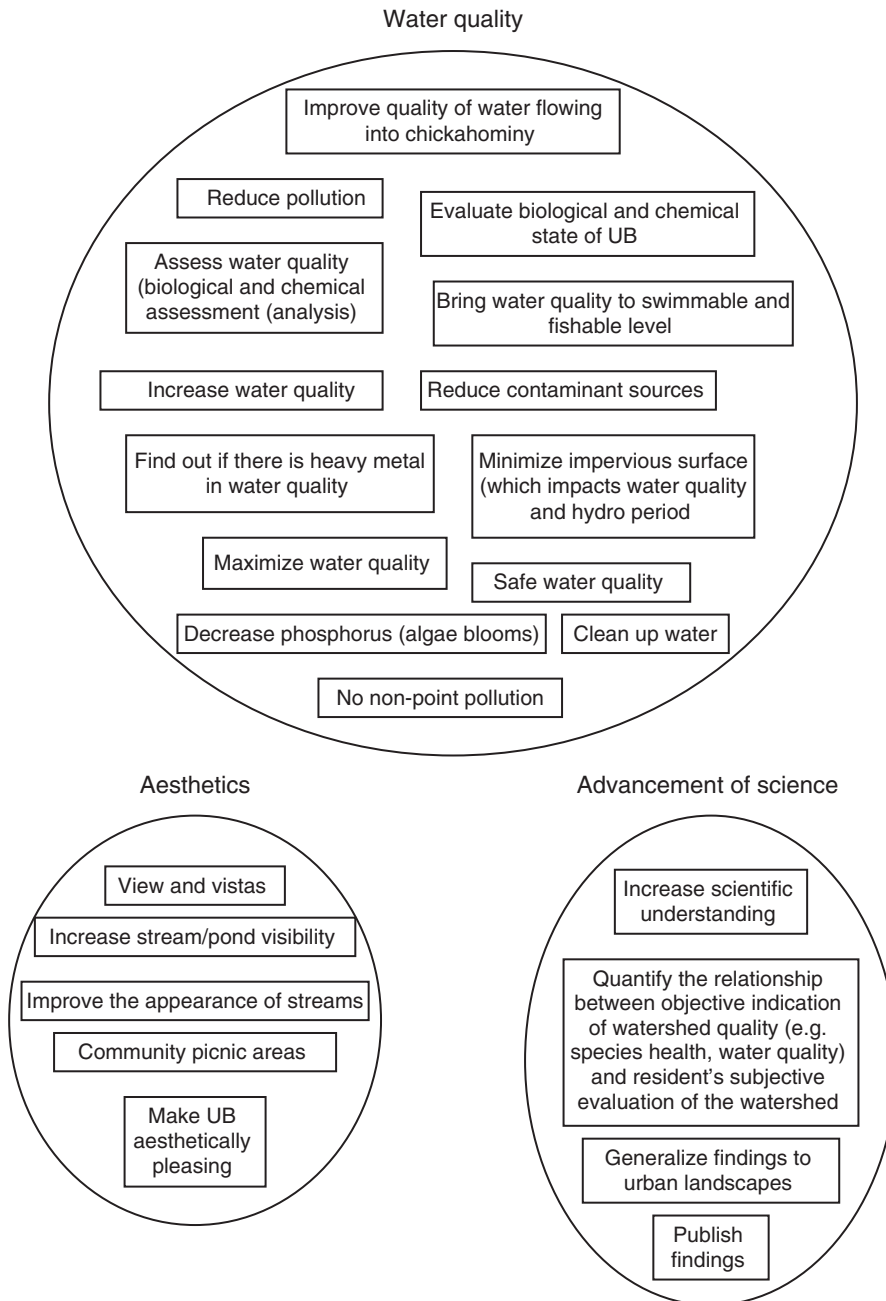


Figure 1. A sample of the initial groups of objectives.

As the interdisciplinary team viewed the objectives that were developed, it noted that some of the objectives were means objectives that lead to the fulfillment of other fundamental objectives [18,19]. An

example is the action to “increase citizen participation in watershed improvement.” Increased participation does not directly improve the quality of the watershed, even though specific activities like cleaning

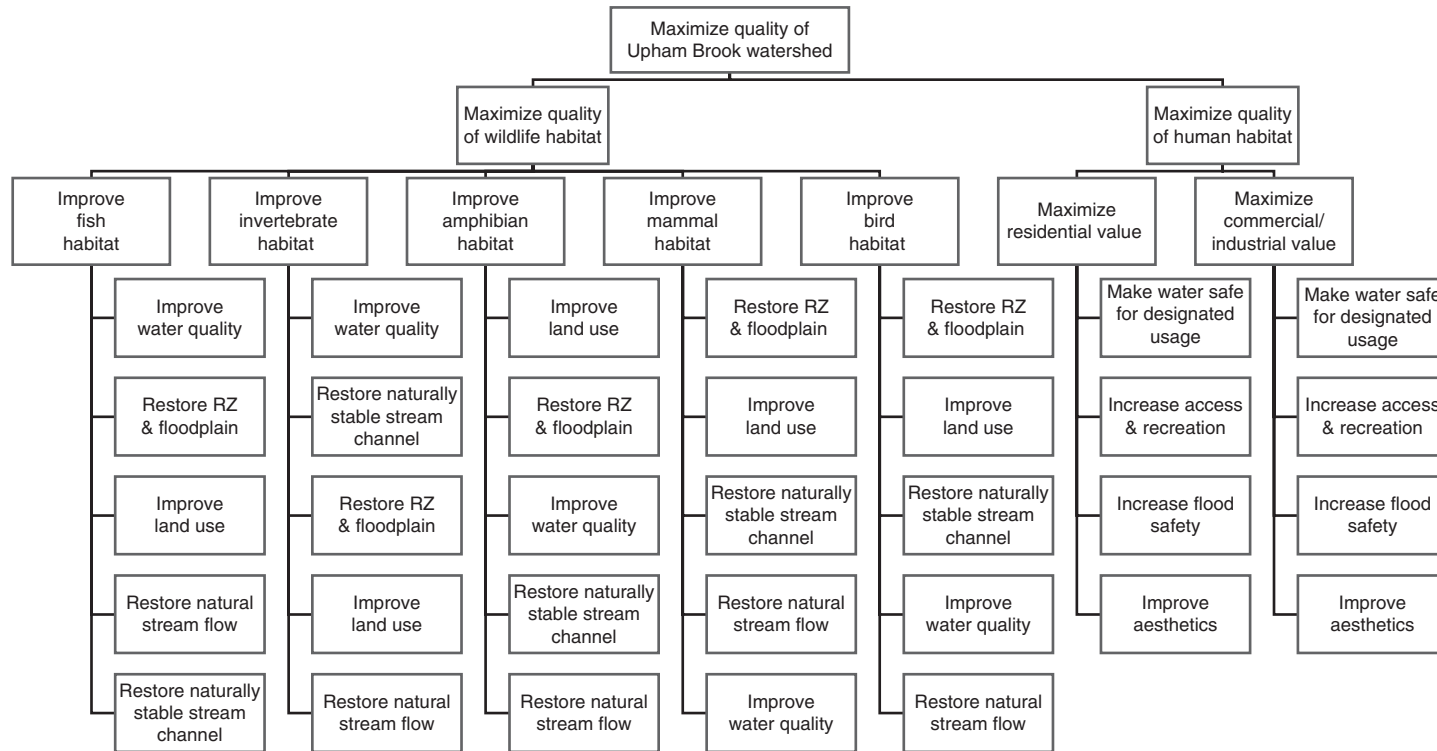


Figure 2. Value hierarchy for the Upham Brook Watershed MODA (multiple objective decision analysis) model.

up the banks and removing blockages from streams can improve water quality. Such objectives were removed from the value hierarchy but retained for use in identifying decision alternatives later in the process. The fundamental objectives were then organized into a value hierarchy [19].

To do this, the group sorted through the fundamental objectives to remove repetition and then took each group of fundamental objectives in the affinity diagram and determined which higher level objective the group represented. The set of higher level objectives define the overall objective. It should be noted that additional levels can be used in the value hierarchy, but were not needed in our application.

The members of the group were surprised by the level of thinking involved in this process and the frequent challenges to their own mental models as a result of the diverse backgrounds of the other members. The development of the value hierarchy took nine 2-h sessions. The hierarchy went through several stages of evolution as the group defined what they meant by the quality of the watershed. After a revision of the initial hierarchy, the final value hierarchy was formed (Fig. 2). The value hierarchy may be viewed as the interdisciplinary team's definition of watershed quality [1,19].

The hierarchy divides the quality of a watershed into two objectives: maximizing the quality of wildlife habitat and maximizing the quality of human habitat. The wildlife habitat is broken down into five objectives, one for each class of wildlife species that could reside within the Upham Book Watershed's area. The same five lowest level objectives are listed below each species class. The interdisciplinary team believed that the five objectives would have a different importance for each class of species and possibly different single-dimensional value functions. The human habitat is divided into two objectives: maximize residential value and maximize commercial/industrial value. These two objectives are then broken down into four lowest level objectives that were appropriate for both objectives though with differing importance.

CONCLUSIONS

It is evident that people do not generate complete and comprehensive lists of objectives to use in making better decisions. The results from Bond *et al.* [10] suggest that groups of decision makers and stakeholders should be used to generate the broadest and most comprehensive list of objectives possible. However, the results from Bond *et al.* [12] show that we must use caution in designing the group process. We do not wish objectives generated by other group members to anchor their fellow stakeholders. The 10 devices from Keeney [13] and the questions from Leon [7] can be used to assist in stretching the breadth and depth of the thought process, but group members should be asked to generate objectives independently at first. Further, a second phase asking for additional information should be employed before any group interaction is allowed. In the second phase, stakeholders should be told to generate at least as many new objectives as in the first phase. Category cues can also be employed based on objectives generated by the whole group in the first phase and on relevant vision, policy, or strategy literature. Finally, after two phases of objective generation, the group can work together to remove redundancy and to determine a final list. Furthermore, Leon [7] shows conclusively that the stakeholders should be asked to think about values and objectives throughout the process rather than being anchored to specific alternatives.

REFERENCES

1. Clemen RT. Making hard decisions. Belmont (CA): Duxbury; 1996.
2. Keeney RL. Value-focused thinking. Cambridge (MA): Harvard Univ. Press; 1992.
3. Volkema RJ. Problem formulation in planning and design. *Manage Sci* 1983;29:639–651.
4. Keller LR, Ho JL. Decision problem structuring: generating options. *IEEE Trans Syst Man Cybern* 1988;18:715–728.
5. Jungermann H, Von Ulardt L, Hausmann L. The role of the good for generating actions. In: Humphreys P, Svenson O, Vari A, editors.

- Analyzing and aiding decision processes. Amsterdam: North-Holland Publishing Co.; 1983. pp. 223–236.
6. Gettys CF, Pliske RM, Manning C, *et al.* An evaluation of human act generation performance. *Organ Behav Hum Decis Processes* 1987;39:23–51.
 7. Leon OG. Value-focused thinking versus alternative-focused thinking: effects on generation of objectives. *Organ Behav Hum Decis Processes* 1999;80(3):213–227.
 8. Keeney RL. Common mistakes in making value trade-offs. *Oper Res* 2002;50(6): 935–945.
 9. Keeney RL. Developing objectives and attributes. In: *Advances in decision analysis: from foundations to applications*. Cambridge, UK: Cambridge University Press. pp. 104–128.
 10. Bond SD, Carlson KA, Keeney RL. Generating objectives: can decision makers articulate what they want? *Manag Sci* 2008;54(1): 56–70.
 11. Keeney RL. Developing of a foundation for strategy at Seagate software. *Interfaces* 1999; 29(6):4–15.
 12. Bond SD, Carlson KA, Keeney RL. Improving the generation of decision objectives. *Decis Anal* 2010. In press. DOI: 10.1287/deca.1100.0172.
 13. Keeney RL. Creativity in decision making with value-focused thinking. *Sloan Manag Rev* 1994;35(4):33–41.
 14. Merrick JRW, Grabowski M, Ayyalasomayajula P, *et al.* 2005a Understanding organizational safety using value-focused thinking. *Risk Anal* 2005;25(4):1029–1041.
 15. Parnell G, Conley H, Jackson J, *et al.* Foundations 2025: a framework for evaluating future air and space forces. *Manage Sci* 1998;44(10):1336–1350.
 16. Merrick JRW, Garcia MW. Using value-focused thinking to improve watersheds. *J Am Plann Assoc* 2004;70(3):313–327.
 17. Merrick JRW, Parnell GS, Barnett J, *et al.* 2005b A multi-objective decision analysis of stakeholder values to identify watershed improvement needs. *Decis Anal* 2006;2(1):44–57.
 18. Keeney RL, Raiffa H. *Decisions with multiple objectives: preferences and value tradeoffs*. New York: Wiley; 1976. (Reprinted by Cambridge University Press; 1993)
 19. Kirkwood CW. *Strategic decision making, multiobjective decision analysis with spreadsheets*. Belmont (CA): Duxbury Press; 1997.

DEFINITION AND EXAMPLES OF CONTINUOUS-TIME MARKOV CHAINS

EVIN UZUN

Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

DEFINITIONS

Consider a continuous-time stochastic process $\{X(t), t \geq 0\}$ with a countable state space S . Such a process is said to satisfy the Markov property if the conditional distribution of the future state at time $t + s$, given the present state at s and all states prior to s , depends only on the present state, that is, for all $i, j \in S$ and $s, t \geq 0$,

$$\begin{aligned} P(X(t + s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s) \\ = P(X(t + s) = j | X(s) = i). \end{aligned} \quad (1)$$

Then, $\{X(t), t \geq 0\}$ is called a *continuous-time Markov chain* (CTMC) if it has the Markov property. If, in addition, $P(X(t + s) = j | X(s) = i)$ is independent of s , then the CTMC is said to be *time-homogeneous*. For the rest of this article, we assume that all CTMCs are time-homogeneous. For a time-homogeneous CTMC $\{X(t), t \geq 0\}$, let

$$p_{ij}(t) = P(X(t) = j | X(0) = i), \forall i, j \in S, t \geq 0,$$

be the transition probabilities and

$$P(t) = [p_{ij}(t)], \forall i, j \in S, t \geq 0,$$

be the transition probability matrix (TPM). Next, let $a_i = P(X(0) = i)$ for all $i \in S$, and

$$\mathbf{a} = [a_i], i \in S,$$

be a row vector representing the initial distribution. A CTMC is completely described by

its initial distribution $\mathbf{a}$ and the TPMs $P(t)$ for each $t \geq 0$, see Theorem 6.2 in Ref. 1.

From the Markov property, it follows that the amount of time a CTMC spends in a state i before making a transition into a different state is exponentially distributed with some rate v_i such that $0 \leq v_i \leq \infty$. A state i for which $v_i = \infty$ is called an *instantaneous state* since when entered it is instantaneously left. (If $v_i = 0$, the state i is called an *absorbing state* since when entered it is never left.) A CTMC with such states is called a *nonregular CTMC*. (An example for a nonregular CTMC is the one having $p_{i,i+1} = 1$ and $v_i = i^2$.) A CTMC is said to be regular if the number of transitions at any finite length of time is finite with probability 1. For the rest of this article, we assume that all CTMCs are regular. Then, we can alternatively define a CTMC as follows (A CTMC, defined as below, has the Markov property and is time-homogeneous; see Theorem 6.1 in Ref. 1):

Definition 1. The stochastic process $\{X(t), t \geq 0\}$ is called a *CTMC* if it has a countable state-space S , and the sequence $\{X_0, (X_n, Y_n), n \geq 1\}$ satisfies

$$\begin{aligned} P(X_{n+1} = j, Y_{n+1} > y | X_n = i, Y_n, X_{n-1}, \\ Y_{n-1}, \dots, X_1, Y_1, X_0) \\ = P(X_1 = j, Y_1 > y | X_0 = i) \\ = p_{ij} e^{-v_i y}, i, j \in S, n \geq 0, \end{aligned} \quad (2)$$

where $X_n = i$ denotes that the process is at state i after the n th jump and Y_n is the amount of time spent at state i (which is exponentially distributed with rate v_i), and $P = [p_{ij}]_{i,j \in S}$ is a stochastic matrix with $p_{ii} = 0$ and $0 \leq v_i < \infty$ for all $i \in S$.

Condition (2) implies that Y_1 is exponentially distributed with rate v_i given $X_0 = i$ independent of the history of the process. Moreover, independent of the history of the process upto time t and the time spent at state i , when

the process leaves state i , it will next enter state j with some probability $p_{i,j}$ such that $\sum_{i \in S} p_{i,j} = 1$ and $p_{i,i} = 0$.

Condition (2) also implies that $\{X_n, n \geq 0\}$ is a discrete-time Markov chain (DTMC) on state space S with TPM P . This DTMC is called the *embedded DTMC* corresponding to the CTMC. Hence, Definition 1 tells us that a CTMC is a stochastic process that moves from state to state in accordance with a DTMC and the amount of time it spends in each state before proceeding to another is exponentially distributed.

In Definition 1, P and $v_i, i \in S$, completely characterize a CTMC. (This follows from the time-homogeneity and regularity assumptions; see Ref. 2.) Next, we define a matrix which is sufficient to characterize a CTMC. Let $q_{i,j}$ be defined by $q_{i,j} = v_i p_{i,j}$ for all $i \neq j$. Since v_i is the rate at which the process leaves state i and $p_{i,j}$ is the probability that it then goes to state j , it follows that $q_{i,j}$ is the rate that the process makes a transition into state j when in state i . Hence, $q_{i,j}$ is called the *transition rate* from i to j . The (infinitesimal) generator matrix of a CTMC is defined as

$$Q = [q_{i,j}], i, j \in S,$$

where $q_{i,i} = -\sum_{k \in S, k \neq i} q_{i,k}, i \in S$. Note that the row sums of Q are zero. Moreover, given the generator matrix Q , we can derive P as follows: Let

$$q_i = \sum_{k \in S, k \neq i} q_{i,k} = -q_{i,i}, i \in S.$$

Note that q_i is the rate that the process jumps out of state i . If $q_i = 0$, then state i is an absorbing state, so we define $p_{i,i} = 1$, denoting that the process will stay in state i with probability 1. If $q_i > 0$, we define

$$p_{i,j} = q_{i,j}/q_i, \forall i, j \in S, j \neq i.$$

Note that $v_i = q_i$ for all $i \in S$. Hence, using Q , we generate P and v_i , for all $i \in S$, which implies that a CTMC is characterized by Q .

Finally, the behavior of a CTMC can be visualized using transition rate diagrams. A rate diagram is a directed graph with a vertex for each state $i \in S$, and an edge from i to j if $q_{i,j} > 0$. In the next section, we will see examples of continuous-time Markov chains, and illustrate some of them with transition rate diagrams.

EXAMPLES

Example 1 [Poisson Process]. Consider a Poisson process with rate λ . First of all, the state space is $\{0, 1, 2, \dots\}$, so this stochastic process has a countable state space. Furthermore, the amount of time this process stays at any state is exponentially distributed with rate λ , and the Markov's property defined in Equation (1) holds due to the memoryless property of the exponential distribution. Hence, the Poisson process is a CTMC with $q_i = q_{i,i+1} = \lambda$ and $p_{i,i+1} = 1$. We can write its generator matrix as follows:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots & 0 \\ 0 & -\lambda & \lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & \lambda & \dots & 0 \\ \vdots & & & \ddots & & \end{bmatrix}$$

The transition rate diagram for a Poisson process is given in Fig. 1.

Example 2 [Birth and Death Process]. Consider a CTMC with a state space $\{0, 1, 2, \dots\}$ such that the process at state i can only go to either state $i + 1$ or state $i - 1$. Such a process is called a *birth and death* process. We can think of the state of a birth and death process as the size of a population, with the state increasing by one whenever a birth occurs, and the state decreasing by one whenever a death occurs. Let

$$q_{i,i+1} = \lambda_i \text{ and } q_{i,i-1} = \mu_i.$$

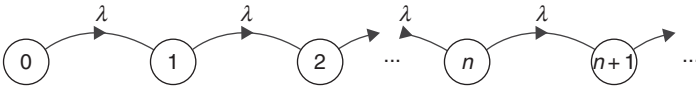


Figure 1. Transition rate diagram for a Poisson process.

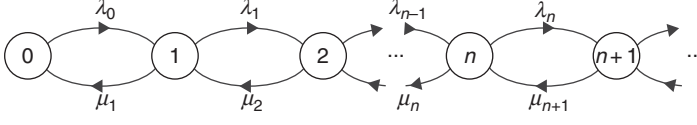


Figure 2. Transition rate diagram for a birth and death process.

Then, $\{\lambda_i, i \geq 0\}$ are called the *birth rates* and $\{\mu_i, i \geq 1\}$ are called the *death rates*. Therefore, in a birth and death process with i people in the system, the time until the next birth and death are exponentially distributed with rates λ_i and μ_i , respectively. We can write the generator matrix as follows:

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots & 0 \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \mu_2 & -\lambda_2 - \mu_2 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & & \end{bmatrix}$$

Moreover, we have $q_i = \lambda_i + \mu_i$, and hence

$$\begin{aligned} p_{i,i+1} &= \frac{\lambda_i}{\lambda_i + \mu_i} \text{ and } p_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i} \\ &= 1 - p_{i,i+1}. \end{aligned}$$

Furthermore, if $\mu_i = 0$ for all $i \geq 0$, the process is called a *pure birth* process. For instance, the Poisson process in Example 1 is a pure birth process. Similarly, if $\lambda_i = 0$ for all $i \geq 0$, the process is called a *pure death* process.

The transition rate diagram for a birth and death process is illustrated in Fig. 2.

Example 3 [M/M/1 queue]. Consider a system where customers that need a particular service arrive in accordance with a Poisson process with rate λ . The service is provided by a single server and service times are exponentially distributed with a mean of $1/\mu$. This system is called an *M/M/1 queueing system* and the process denoting the number of customers in the queue is a birth and death process with a birth rate λ for all $i \geq 0$ and a death rate μ for all $i \geq 1$. We can write the generator matrix as follows:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots & 0 \\ \mu & -\lambda - \mu & \lambda & 0 & 0 & \cdots & 0 \\ 0 & \mu & -\lambda - \mu & \lambda & 0 & \cdots & 0 \\ \vdots & & & \ddots & & & \end{bmatrix}$$

Example 4 [M/M/1 Queue with Reneging]. Consider the M/M/1 queueing system in Example 3. Assume, in addition, that a customer leaves the system if she waits longer than her maximum waiting time in the queue, which is an exponentially distributed random variable with a mean of $1/\gamma$. Then, the process denoting the number of customers in the queue is a birth and death process with a birth rate λ for all $i \geq 0$ and a death rate $\mu + i\gamma$ for all $i \geq 1$. We can write the generator matrix as follows:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots & 0 \\ \mu + \gamma & -\lambda - \mu - \gamma & \lambda & 0 & 0 & \cdots & 0 \\ 0 & \mu + 2\gamma & -\lambda - \mu - 2\gamma & \lambda & 0 & \cdots & 0 \\ \vdots & & & \ddots & & & \end{bmatrix}$$

Example 5 [An Epidemic Model].

This example is from Ref. 3, and it considers a population of $m \geq 1$ people with one infected person and $m - 1$ susceptible people at time zero. In any time interval h , any given infected person infects any susceptible person with probability $h\alpha + o(h)$, and once infected, a person stays infected forever. Then, the process denoting the number of infected people in the system is a pure birth process with state space $S = \{1, \dots, m\}$ and birth rate $\lambda_i = i(m - i)\alpha$ for all $i \in S$. For example, for $m = 5$, we can write the generator matrix as follows:

$$Q = \begin{bmatrix} -4\alpha & 4\alpha & 0 & 0 & 0 \\ 0 & -6\alpha & 6\alpha & 0 & 0 \\ 0 & 0 & -6\alpha & 6\alpha & 0 \\ 0 & 0 & 0 & -4\alpha & 4\alpha \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Example 6 [An Epidemic Model with Recovery]. Consider the epidemic model in Example 5. Assume, in addition, that infected people recover from the infection after some time, but after the recovery, they can be infected again. The recovery

time for each infected person is exponentially distributed with a mean of $1/\mu$. Then, the process denoting the number of infected people in the system is a birth and death process with state space $S = \{0, 1, \dots, m\}$. Considering the transition rates, for all $i \in S$, the birth and death rates are given by $\lambda_i = i(m-i)\alpha$ and $\mu_i = i\mu$, respectively. For example, for $m = 5$, we can write the generator matrix as follows:

$Q =$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \mu & -4\alpha - \mu & 4\alpha & 0 & 0 & 0 \\ 0 & 2\mu & -6\alpha - 2\mu & 6\alpha & 0 & 0 \\ 0 & 0 & 3\mu & -6\alpha - 3\mu & 6\alpha & 0 \\ 0 & 0 & 0 & 4\mu & -4\alpha - 4\mu & 4\alpha \\ 0 & 0 & 0 & 0 & 5\mu & -5\mu \end{bmatrix}$$

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. New York (NY): Chapman and Hall; 1995.
2. Cinlar E. Introduction to stochastic processes. Englewood Cliffs, NJ: Prentice Hall; 2002.
3. Ross SM. Stochastic processes. 2nd ed. New York (NY): John Wiley and Sons, Inc.; 1996.

DEFINITION AND EXAMPLES OF DTMCs

NATARAJAN GAUTAM
Department of Industrial and Systems
Engineering, Texas A&M University,
College Station, Texas

A stochastic process describes the state of a system and its random evolution over time. Broadly speaking, stochastic processes are of four types depending on whether the system is observed continuously or at discrete time points, and whether the observations or states are discrete quantities or continuous, as follows:

1. *Discrete State and Discrete Observation.* The number of cereal boxes on a grocery store shelf observed at the beginning of a day.
2. *Discrete State and Continuous Observation.* The number of e-mails in someone's *InBox* observed at all times.
3. *Continuous State and Discrete Observation.* The amount of water in a reservoir observed immediately after a rain shower.
4. *Continuous State and Continuous Observation.* The temperature inside a car observed all the time.

A discrete-time Markov chain (DTMC) is a special type of stochastic process belonging to the first category; that is, system states are discrete and system is observed at discrete times. All DTMCs are discrete state and discrete observation stochastic processes but for a stochastic process to be a DTMC, some other conditions need to be satisfied which are provided in the next section. However, it is crucial to point out that the observations do not have to be periodic (although in most cases they would be), such as every day or every hour. It is also critical to realize that the notion of "time" is not rigid; for example, the observations can be done over

space (but that is certainly atypical). In summary, DTMCs can be applied in fairly general settings to situations commonly found in real life. They have gained significant attention in the queueing [1] and inventory [2] literature, and have been used extensively in the analysis of production systems, computer-communication systems, transportation systems, biological systems, and so on.

We first describe the definition and notations used in modeling a system as a DTMC. This is the focus of the section titled "Definition and Notation." Then, in the section titled "Modeling a System as a DTMC," we provide a methodical technique to model a given system using a DTMC. Finally, in the section titled "Examples," we describe several examples to illustrate the modeling framework as well as to motivate the wide variety of applications where DTMCs are used. In addition, there are several texts [3,4] and websites (<http://www.sosmath.com/matrix/markov/markov.html>) that provide additional examples that would enhance the understanding of concepts included in this article.

DEFINITION AND NOTATION

Consider a system that is randomly evolving in time. Let X_n be the state of the system at the n th observation. In essence, X_n describes the system based on the n th observation. For example, X_n could be (i) the status of a machine (up or down) at the beginning of the n th hour; (ii) the number of repairs left to be completed in a shop at the end of the n th day; (iii) the cell phone company a person is with when he/she received his/her n th phone call; (iv) the number of boys and the number of girls in a classroom at the beginning of the n th class period; (iv) the number of gas stations at the n th exit on a highway.

It is crucial to understand that although all the above examples are discrete states and discrete observations, it is not necessary

that the stochastic process $\{X_0, X_1, X_2, \dots\}$ is a DTMC. For that, some additional properties are necessary. However, notice from the examples that the states can be one-dimensional or multidimensional, they can take finite or infinite values, they could be over time or space, they do not have to be equally spaced (periodic), and they do not have to be numerical values. We define the *state space* S as the set of all possible values of X_n for all n . In the previous paragraph, in example (i) we have $S = \{Up, Down\}$, in example (ii) we have $S = \{0, 1, 2, \dots\}$, and in example (iv) we have $S = \{(0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0), \dots\}$. In summary, we require X_n and S to be discrete valued (or also called *countable*).

The next question to ask is when can a stochastic process $\{X_0, X_1, X_2, \dots\}$ be called a DTMC? A stochastic process $\{X_0, X_1, X_2, \dots\}$, which we henceforth denote as $\{X_n, n \geq 0\}$, is called a DTMC if it satisfies the *Markov property*. In words, Markov property states that if the current state of the system is known, the future states are independent of the past. In other words, to predict (probabilistically) the states in the future, all one needs to know is the present state and nothing from the past. For example, if X_n denotes the amount of inventory of a product at the beginning of the n th day (say today), then it is conceivable that to predict the inventory at the beginning of the next day and the day after, all one would need is today's inventory. Yesterday's inventory or how you got today's inventory is not necessary.

Mathematically, Markov property is defined as

$$\begin{aligned} P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \\ X_{n-2} = i_{n-2}, \dots, X_0 = i_0\} \\ = P\{X_{n+1} = j \mid X_n = i\}. \end{aligned}$$

In other words, once X_n is known, we can predict (probabilistically) $X_{n+1}, X_{n+2}, \dots$ without needing to know $X_{n-1}, X_{n-2}, \dots$. In summary, a stochastic process $\{X_n, n \geq 0\}$ defined on a countable state space S is called a DTMC if it satisfies the Markov property. This is the technical definition of a DTMC. Now that we have defined a DTMC, in the next

section we explain how a system can be modeled as a DTMC so that the system can be analyzed.

MODELING A SYSTEM AS A DTMC

Before presenting a systematic approach to modeling a system as a DTMC, we provide some more assumptions and notation. Besides the Markov property, there is another property that needs to be satisfied in order to be able to effectively analyze a DTMC. A DTMC is called *time homogeneous* if the probability of transition from a state to another state does not vary with time. In other words, for a DTMC with state space S to be time homogeneous, for every $i \in S$ and $j \in S$,

$$P\{X_{n+1} = j \mid X_n = i\} = P\{X_1 = j \mid X_0 = i\}.$$

We denote the above expression as p_{ij} , the one-step transition probability of going from state i to j . Therefore, for a time-homogeneous DTMC, for $i \in S$ and $j \in S$,

$$p_{ij} = P\{X_{n+1} = j \mid X_n = i\}.$$

Notice that p_{ij} is not a function of n which is essentially because the DTMC is time homogeneous.

Using the transition probabilities p_{ij} for every $i \in S$ and $j \in S$, we can build a square matrix $P = [p_{ij}]$. The matrix P is called *transition probability matrix*. The rows of the matrix correspond to the given current state, while the columns correspond to the next state. The sum of the elements of each row adds to 1 because given that the DTMC is in state i (for some $i \in S$), the next state ought to be one of the states in S ; hence, for every $i \in S$,

$$\sum_{j \in S} p_{ij} = 1.$$

Oftentimes, a transition diagram is used to represent the transition probabilities, as opposed to the matrix P . A transition diagram is a directed graph with nodes corresponding to the states (thereby the set of nodes is indeed S), arcs corresponding to possible one-step transitions, and arc values being

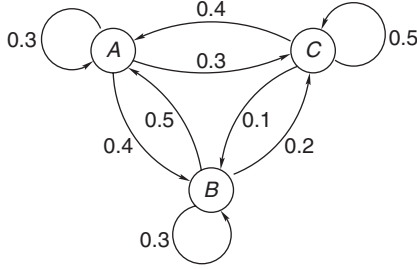


Figure 1. Transition Diagram for Jill's Switching Patterns

transition probabilities. To draw a transition diagram, for every $i \in S$ and $j \in S$, draw an arc from node i to node j if $p_{ij} > 0$. Now we illustrate an example of X_n , S , the transition probability matrix P , and transition diagram.

Consider three long-distance telephone companies A , B , and C . Every time a sale is announced, users switch from one company to another. Let X_n denote the long-distance company with which a particular user named Jill is just before the n th sale announcement. We have $S = \{A, B, C\}$. On the basis of the customers' switching in the past, it is estimated that Jill's switching patterns follow the following transition probability matrix:

$$P = \begin{matrix} & \begin{matrix} A & B & C \end{matrix} \\ \begin{matrix} A \\ B \\ C \end{matrix} & \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{matrix}.$$

For example, if Jill is with B before a sale is announced, she would switch to A with probability 0.5 and C with probability 0.2 or stay with B with probability 0.3, as evident from the second row in P . Now, converting the above P matrix into a transition diagram, we get the picture as shown in Fig. 1.

Having described all the definitions, notations, and requirements, we are now ready to model a system as a DTMC. Five steps are involved in systematically modeling a system as a DTMC: (i) define X_n ; (ii) write down S ; (iii) verify that the Markov property is satisfied; (iv) verify that the DTMC is time homogeneous; and (v) obtain P or draw the transition diagram.

Next, we present some examples where systems are modeled as DTMCs using the five steps described above. Note that although modeling a system as a DTMC is considered an art, the key scientific aspect is that X_n must be chosen carefully so that S is as small as possible and steps (iii)–(v) can be performed. If not, one must consider revising X_n . In particular, one of the most important steps in modeling is to choose X_n so that Markov property is satisfied. One approach is to define X_n in such a way that it carries all the past information needed to forecast the future states, or that the future state could be written in terms of only the present state and some other factors that are independent of the past and present states. We illustrate this in some of the examples in the next section.

EXAMPLES

In the examples that follow, the objective is to model the system described as a DTMC by stating X_n and S as well as obtaining P . The order of states in the rows and columns of the P matrices is identical to those in the order of the corresponding S .

Example 1. A company uses two forecasting tools for making demand predictions. Tool i is effective with probability p_i (for $i = 1, 2$). If the n th prediction uses tool i and it is observed to be effective, then the $(n + 1)$ th prediction is also done using the same tool; if it is observed to be ineffective, then the $(n + 1)$ th prediction is made using the other tool. This is sometimes known as “play the winner rule.”

To model this system as a DTMC, we let X_n to be the tool used for the n th prediction. The state space $S = \{1, 2\}$. It can be verified that Markov property is satisfied because the tool used for the $n + 1$ th prediction depends on only what was used for the n th prediction and the outcome of the n th prediction and not the history. Further, p_i for $i = 1, 2$ remains constant over time; hence, the DTMC is time homogeneous. Clearly, from the description,

$$P = \begin{bmatrix} p_1 & 1 - p_1 \\ 1 - p_2 & p_2 \end{bmatrix}.$$

Example 2. Probe vehicles are sent through two similar routes from source A to destination B to determine the travel times. If a route was congested when a probe vehicle is sent, then it will be congested with probability q when the next probe vehicle needs to be sent. Likewise, if a route was not congested when a probe vehicle is sent, it will not be congested with probability p when the next probe vehicle is sent. The routes behave independent of each other and since they are similar we assume that p and q do not vary between the routes.

Let X_n be the number of uncongested routes when the n th set of probe vehicles are sent. The state space $S = \{0, 1, 2\}$. It can be verified that Markov property is satisfied because the congestion states of each route depends only on whether they were congested during the previous probe and not the history. Further, p and q remain constant over time; hence, the DTMC is time homogeneous. From the description, it is possible to show that

$$P = \begin{bmatrix} q^2 & 2q(1-q) & (1-q)^2 \\ q(1-p) & pq + (1-p)(1-q) & p(1-q) \\ (1-p)^2 & 2p(1-p) & p^2 \end{bmatrix}.$$

It may be worthwhile to describe some of the above p_{ij} values. In particular, $p_{02} = (1 - q)^2$ since the probability of going from both routes being congested to no route congested is if both congested routes become uncongested, each happening with probability $1 - q$. Likewise $p_{10} = q(1 - p)$ because the probability that the congested route remains congested is q and the probability that the uncongested route becomes congested is $1 - p$. In a simi-

lar manner, all the p_{ij} values can be obtained using a similar logic.

Example 3. Consider the following weather forecasting model: if today is sunny and it is the n th day of the current sunny spell, then it will be sunny tomorrow with probability p_n regardless of what happened before the current sunny spell started. Likewise, if today is rainy and it is the n th day of the current rainy spell, then it will be rainy tomorrow with probability q_n regardless of what happened before the current rainy spell started.

Oftentimes, one is tempted to say that the state of the system is the type of day (sunny or rainy). However, one of the concerns is that Markov property would not be satisfied since to predict the next state one needs to know the history to figure out how long the spell has been. Therefore, one needs to carry the spell information too in the state. In this light, let X_n be a two-dimensional state denoting the type of day on the n th day and how many days the current spell has lasted. Letting s and r to denote rainy and sunny respectively, the state space S is $S = \{(s, 1), (r, 1), (s, 2), (r, 2), (s, 3), (r, 3), \dots\}$. It can easily be verified that Markov property is satisfied. The DTMC is time homogeneous, although one must not mistake n in p_n or q_n to be the n in X_n . Consider that the system is in state (s, i) during a day. Then, from the description, on day $n + 1$ the system would be in state $(s, i + 1)$ with probability p_i if it does not rain and $(r, 1)$ with probability $1 - p_i$ if it does rain. A similar argument can be made for (r, i) . Therefore, the P matrix in the order the states are represented in S is

[illegible]

Example 4. Many biological processes especially at the cellular level are a result of polymerization and depolymerization of organic compounds. A polymer in a cell is made up of N monomers that are attached back to back with one end of the chain anchored to the cell and the other end grows or shrinks. For example, a polymer $M_0-M-M-M$ is made of four monomers and M_0 indicates it is anchored. In each time unit with probability p , a new monomer joins the growing end, and with probability q the nonanchored end monomer leaves the polymer. For example, the $M_0-M-M-M$ chain in the next time unit becomes $M_0-M-M-M-M$

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & 0 & \dots \\ q & 1-p-q & p & 0 & 0 & \dots \\ 0 & q & 1-p-q & p & 0 & \dots \\ 0 & 0 & q & 1-p-q & p & \dots \\ 0 & 0 & 0 & q & 1-p-q & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Example 5. A machine produces two items per day. The probability that an item is nondefective is p . Successive items are independent. Defective items are thrown away instantly. The demand is one item per day which occurs at the end of a day. Any demand that cannot be satisfied immediately is lost.

Let X_n be the number of items in storage at the beginning of the n th day (before production and demand of that day). Since demand takes place after production on a given day, we have, if $X_n > 0$,

$$X_{n+1} = \begin{cases} X_n + 1 & \text{w.p. } p^2 \\ X_n & \text{w.p. } 2p(1-p) \\ X_n - 1 & \text{w.p. } (1-p)^2 \end{cases}$$

and, if $X_n = 0$,

$$X_{n+1} = \begin{cases} X_n + 1 & \text{w.p. } p^2 \\ X_n & \text{w.p. } 1-p^2 \end{cases}$$

Since X_{n+1} depends only on X_n , this is a DTMC with state space $S = \{0, 1, 2, \dots\}$. The transition probabilities p_{ij} for all $i \geq 0$ and

with probability p , M_0-M-M with probability q , and stays as $M_0-M-M-M$ with probability $1-p-q$.

We consider a single polymer and let X_n denote the length of the polymer (in terms of number of monomers) at the beginning of the n th time unit. Therefore, the state space is $S = \{1, 2, 3, 4, \dots\}$. Since one side is anchored, we assume that the last monomer would never break away. Verify that Markov property is satisfied and $\{X_n, n \geq 0\}$ is time homogeneous. The P matrix in the order the states are represented in S is

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ p & 0 & 0 & 0 & \dots \\ 1-p-q & p & 0 & 0 & \dots \\ q & 1-p-q & p & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$j \geq 0$ are given by

$$p_{ij} = \begin{cases} p^2, & \text{if } j = i + 1, \\ 2p(1-p), & \text{if } j = i \text{ and } i > 0, \\ 1-p^2, & \text{if } j = i = 0, \\ (1-p)^2, & \text{if } j = i - 1, \\ 0, & \text{otherwise.} \end{cases}$$

Example 6. Consider a time division multiplexer from which packets are transmitted at times 0, 1, 2, and so on. Packets arriving between time n and $n+1$ have to wait until time $n+1$ is transmitted. However, at most, one packet can be transmitted at a time. Let Y_n be the number of packets that arrive during time n to $n+1$. Assume that $a_i = P\{Y_n = i\}$ and that there is infinite room in the waiting space.

Let X_n be the number of packets awaiting transmission just before time n . Clearly, we have the state space $S = \{0, 1, 2, \dots\}$. Define the transition probabilities p_{ij} as (for $i > 0$)

$$\begin{aligned} p_{ij} &= P\{X_{n+1} = j \mid X_n = i\} \\ &= a_{j-i+1}. \end{aligned}$$

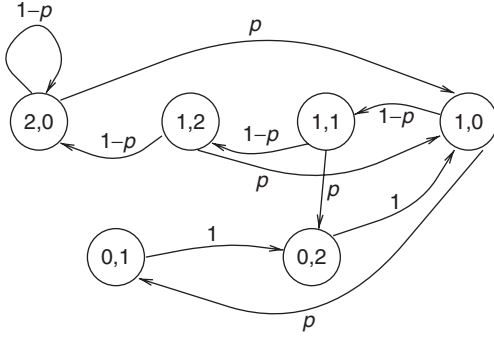


Figure 2. Transition Diagram for Manufacturing System in Example 7

Similarly,

$$p_{0j} = P\{X_{n+1} = j \mid X_n = 0\} = P\{Y_n = j\} = a_j.$$

Verify that Markov property is satisfied and $\{X_n, n \geq 0\}$ is time homogeneous. Therefore, the transition probability matrix $P = [p_{ij}]$ is

$$P = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ a_0 & a_1 & a_2 & a_3 & \dots \\ & a_0 & a_1 & a_2 & \dots \\ & & a_0 & a_1 & \dots \\ & & & a_0 & \dots \end{bmatrix}.$$

Example 7. Consider a manufacturing system of two identical machines that are in working condition; one is in use and the other is in standby. The probability that a machine that is in use fails during an hour is p (assume that a machine in standby mode does not fail). The system is observed once every hour. It takes just a little less than 3 h to repair a failed machine. Only one machine can be repaired at a time and repairs are according to first-come first-serve basis.

Let X_n be the number of machines in working condition at the beginning of time unit n . Let Y_n be the number of time units of repair completed at the beginning of time unit n . Notice that X_n and Y_n are chosen such that Markov property is satisfied and $\{(X_n, Y_n), n \geq 0\}$ is time homogeneous. Therefore, $\{(X_n, Y_n), n \geq 0\}$ is a DTMC with the transition diagram shown in Fig. 2.

REFERENCES

1. Gross D, Harris CM. Fundamentals of queueing theory. 3rd ed. New York: John Wiley and Sons, Inc.; 1998.
2. Zipkin PH. Foundations of inventory management. New York (NY): McGraw Hill and Company Inc.; 2000.
3. Kulkarni VG. Modeling and analysis of stochastic systems. In: Texts in Statistical Science Series. London: Chapman and Hall, Ltd.; 1995.
4. Ross SM. Introduction to probability models. San Diego (CA): Academic Press; 2003.

DEFINITION AND EXAMPLES OF RENEWAL PROCESSES

LIQIANG LIU
EURANDOM, Eindhoven
University of Technology,
Eindhoven, Netherlands

This article introduces a class of stochastic processes called *renewal processes*, occasionally referred to as *recurrent processes*, which play several important roles in the grand scheme of stochastic processes.

First, they help us relax distributional assumptions in Markov chains, namely, the geometric distributions for the DTMCs (discrete-time Markov chains), and the exponential distributions for the CTMCs (continuous-time Markov chain). Renewal processes provide us with important tools (see **Limit Theorems for Renewal Processes**) to deal with general distributions.

Second, renewal processes provide a unifying theoretical framework (renewal theory) for studying the limiting behavior of specialized stochastic processes. For example, we shall see that renewal processes appear as embedded processes in the DTMCs and the CTMCs, and thus convergence results about these processes can be obtained within the framework of renewal theory.

Third, renewal processes lead to important generalizations such as semi-Markov processes (see **Semi-Markov Processes**) and regenerative processes (see **Markov Regenerative Processes**). With the capacity to handle general cost and reward structures, renewal processes and their generalizations are effective and prevailing stochastic models in operations research and management science.

DEFINITIONS

Consider a certain event repetitively occurring at time $S_1, S_2, \dots$. Assume

$$0 \leq S_1 \leq S_2 \leq \dots$$

Thus S_n is the time of the n th occurrence of the event. Define

$$X_n = S_{n+1} - S_n, \quad n \geq 1.$$

Thus $X_1, X_2, \dots$ is a sequence of waiting times between two successive occurrences of the event. Let $S_0 = 0$ by convention. Next, we define

$$N(t) = \sup\{n \geq 0 : S_n \leq t\}, \quad t \geq 0.$$

Thus the process $\{N(t), t \geq 0\}$ counts the number of occurrences of the event up to time t . A typical sample path of this process is shown in Fig. 1. The process is a continuous-time process with piecewise constant right-continuous sample paths. With these notations we are ready to define a renewal process.

Definition 1 [Renewal sequence, renewal process]. The sequence $S_1, S_2, \dots$ is called the *renewal sequence* and the process $\{N(t), t \geq 0\}$ is called the *renewal process* if

1. $S_1, X_1, X_2, \dots$ are mutually independent nonnegative random variables,
2. $X_1, X_2, \dots$ have a common distribution F and $F(0) < 1$.

Definition 2 [Pure and delayed renewal process]. A renewal process is pure if S_1 is a random variable with a distribution $G = F$, and is delayed if $G \neq F$. A pure renewal process is also known as an *ordinary renewal process*.

Definition 3 [Recurrent and transient renewal process]. A renewal process is recurrent if F is a proper distribution, or transient if F is defective. Here F is proper means that $F(\infty) = \lim_{t \rightarrow \infty} F(t) = 1$. If $F(\infty) < 1$, the distribution is defective. It means that there is a positive probability that the random variable is unbounded. This implies that almost surely $X_n = \infty$ for

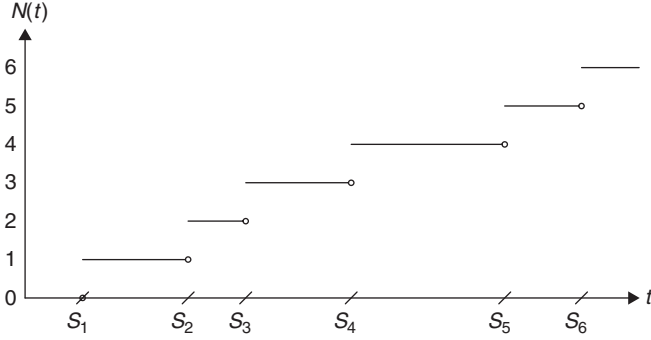


Figure 1. Sample path of a counting process.

some $n = K$, so that the repetition eventually terminates after K times. Recurrent renewal process is also known as *persistent renewal process*. Transient renewal process is also known as *terminating renewal process*.

The distribution F is also known as *inter-renewal, waiting, holding or recurrence-time distribution*. We assume $F(0) < 1$ to exclude the trivial case, $S_n - S_1 = 0$ for all $n > 1$ with probability 1. Note that allowing $X_n = 0$ with a positive probability implies that the event may occur multiple and almost surely finite times simultaneously. Some authors [1–3] define renewal process to be the process $\{S_n, n \geq 1\}$. This is a mathematically equivalent definition in a sense that

$$N(t) \leq n \text{ if and only if } S_{n+1} > t,$$

$$N(t) \geq n \text{ if and only if } S_n \leq t.$$

In this case $\{N(t), t \geq 0\}$ and $\{S_n, n \geq 1\}$ are called it renewal counting process and *renewal point process*, respectively, for clearness [4].

EXAMPLES

Example 1 [Coin tossing]. Consider tossing a coin in a row and the event: the outcome is head. Suppose for each toss, the probability of getting a head is p . Let S_n be the trial at which the n th head is obtained and the next head is obtained after X_n tosses. Then the sequence $S_1, S_2, \dots$ is a renewal sequence with recurrence time of geometric distribution,

$$\begin{aligned} \mathbb{P}\{S_1 = k\} &= \mathbb{P}\{X_n = k\} = (1-p)^{k-1}p, \\ k, n &= 1, 2, \dots \end{aligned}$$

Example 2 [Poisson process]. A Poisson process with mean arrival rate λ is a renewal process with F being the exponential distribution with mean $1/\lambda$. Thus, a renewal process can be seen as a direct generalization of the Poisson process with respect to a relaxation of the exponential assumption in the waiting-time distribution.

Example 3 [Renewal processes in Markov chains]. Let $\{Y(t), t \geq 0\}$ be a CTMC on $\{1, 2, \dots\}$. Let $S_n, n \geq 1$ be the time of the n th occurrence of the event: the state transits from some $i > 0$ into 0. Let $N(t)$ count the number of such transitions up to time t . Since $X_n = S_{n+1} - S_n$ is the time between two successive transitions into state 0, the Markov property of the CTMC implies that $X_1, X_2, \dots$ is a sequence of iid (independent and identically distributed) positive random variables. Hence $\{N(t), t \geq 0\}$ is a renewal process. If $Y(0) = 0$ with probability 1, then this renewal process is pure. Similarly, we can also identify renewal processes in a DTMC.

Example 4 [Machine maintenance]. Suppose a manufacturing process requires continuous use of a machine. Suppose we follow the age-replacement policy: replace the machine in use when it fails or when it reaches an age of T years. This policy is useful in reducing the number of on-job failures, which tend to be more costly. If we schedule our replacement more frequently (smaller T),

then there would be less on-job failures. Let L be the lifetime of a new machine. Assume replacement of the machine is completed instantaneously. Let $S_n, n \geq 1$ be the time of the n th replacement. This is a renewal sequence since $X_1, X_2, \dots$ are iid copies of the random variable $X = \min(T, L)$.

One may also consider a subsequence such as the epochs when a failed machine is replaced. Let S_{I_n} be the time of the n th failure. Then $I_1, I_2, \dots$ is a renewal sequence in the “Coin tossing” example above (Example 1), with $p = \mathbb{P}\{T \geq L\}$. By Definition 1, it is easy to check that a subsequence of a renewal sequence is also renewal if the subindex sequence is renewal. Thus the on-job replacement times, $S_{I_1}, S_{I_2}, \dots$ is a renewal sequence. Similarly, the planned replacement times also form a renewal sequence.

The renewal processes illustrated in this example are pure if we begin with a new machine at time 0.

Example 5 [Renewal processes in queueing models]. Consider a $G/G/1$ queue. We can easily identify the arrival epochs $S_1, S_2, \dots$ form a renewal sequence, since by definition (see *The G/G/1 Queue*) the interarrival times in a $G/G/1$ queue are nonnegative iid random variables. Hence $\{N(t), t \geq 0\}$, which counts the number of arrivals up to time t is a renewal process. We say the arrival process is renewal. Here, we see an example of renewal process being an explicit assumption in the queueing model.

Next, let us consider event a : a customer arrives at an empty queue, which occurs chronologically at time $A_1, A_2, \dots$. Intuitively, after each occurrence of such an event, the queueing system restarts from scratch in the sense that the evolution of the queue is a replica of its evolution after A_1 . Thus $A_1, A_2, \dots$ is a renewal sequence.

Last, let us consider another event d : the queue becomes empty after a customer departs, and check whether the epochs form a renewal sequence. Let D_n be the n th epoch when event d occurs. Clearly, event a and event d occur alternatively. Suppose the system is empty at time 0. Then we get a sequence $A_1, D_1, A_2, D_2, \dots$ if we

order the occurrences of both events. Define $B_n = D_n - A_n$ and $I_n = A_{n+1} - D_n$, $n \geq 1$. The random variable B_n is the n th busy period and I_n is the n th idle period (we do not count the period from 0 to A_1 as the first idle period). Since the system “starts from scratch” after each occurrence of event a , the sequence $B_1, B_2, \dots$, the sequence $I_1, I_2, \dots$, and the sequence $(B_1, I_1), (B_2, I_2), \dots$ are all iid. (The last bivariate random variable sequence is called it alternating renewal sequence. See *Alternating Renewal Processes*) However, for $G/G/1$ queue, $D_1, D_2, \dots$ is not necessarily a renewal sequence since $D_{n+1} - D_n = I_n + B_{n+1}$ and I_n may depend on B_n . If the distribution of interarrival time is exponential, then, due to the memoryless property, $B_1, I_1, B_2, I_2, \dots$ is a sequence of mutually independent random variables. Hence for $M/G/1$ queues, $D_1, D_2, \dots$ is a renewal sequence.

REMARKS

The intrinsic property that defines the renewal process is the existence of a sequence of nonnegative iid random variables, that is, the recurrence times. In layman’s terms, the event recurs in a way that no matter when it occurs or what happened before that, the waiting time until the next occurrence is always the same, of course, in a stochastic sense. For this salient feature, renewal process earns its important footing in the study of stochastic processes. The probability law of a stochastic process may be too intricate to analyze completely but the existence of an embedding renewal sequence helps us understand essential features of the process, prove limit theorems, and so on. The study of renewal processes, called *renewal theory*, contributes greatly to the simplification and unification of many theories. For further coverage of this topic, the reader is encouraged to see the excellent monographs by Feller [1] and Cox [2], or a fine expository article by Smith [5].

Lastly, it should be remarked that the definition of renewal process is not restricted to the temporal interpretation of t . For example, S_n can also be thought of as the location of a

set of particles on a line and $N(t)$ counts the number of particles in the segment $[0, t]$.

REFERENCES

1. Feller W. An introduction to probability theory and its applications. New York: John Wiley & Sons; 1966.
2. Cox D. Renewal theory. London: Methuen; 1962.
3. Asmussen S. Applied probability and queues. New York: Springer; 2003.
4. Rolski T, Schmidli H, Schmidt V, *et al.* Stochastic processes for finance and insurance. New York: Wiley; 1999.
5. Smith W. Renewal theory and its ramifications. J R Stat Soc Ser B (Methodological) 1958;20(2):243–302.

DEGENERACY AND VARIABLE ENTERING/EXITING RULES

ISTVÁN MAROS
Department of Computing,
Imperial College London,
London, UK

Department of Computer
Science and Systems
Technology, University of
Pannonia, Veszprém,
Hungary

The simplex method (SM) of linear programming (LP) is an iterative algorithm. The feasible set of an LP problem, if not empty, is a convex polyhedron. Assume it is nonempty. Theory has shown that an optimal solution to a problem, if exists, is achieved at a vertex of this set. A vertex of the polyhedron algebraically corresponds to a basic solution. The SM moves from a basic solution (basis) to a neighboring basic solution along a connecting edge such that the objective value in the new vertex is better (not worse) in terms of the objective value. Thus, an iteration of the SM consists of two steps, finding an improving direction and then identifying the neighboring vertex (finding the steplength) in that direction.

If the solution in the neighboring basis is always better then no basis can be repeated and the algorithm terminates in a *finite number of iterations* either at an optimal solution or detects that the solution is unbounded.

Neighboring bases differ from each other in just one column. So, an iteration consists of determining an “incoming” (entering) nonbasic column and an “outgoing” (exiting) basic column.

Finiteness of the SM can be in danger if the steplength of an iteration is zero (nonimproving iteration). This can happen in the presence of degeneracy and can lead to a sequence of nonimproving iterations and a return to the first basis in the sequence. If it happens the sequence is repeated infinitely resulting in cycling and the SM does not terminate.

This situation was early recognized when the proof of finiteness of the SM was constructed. First the proof was completed under the “nondegeneracy assumption,” that is, all bases were assumed nondegenerate. Later, theoretically sound rules [1] were introduced regarding how to choose the entering and exiting variables during iterations such that finiteness is guaranteed.

In the sequel we give a definition of degeneracy and investigate the phenomenon both from theoretical and computational point of views.

PROBLEM STATEMENT, DEFINITIONS

What is degeneracy? How does it come about? How does it cause problems, if at all? How can it disappear? To answer these questions we consider the following general LP problem:

$$\min z = \sum_{j=1}^n c_j x_j$$

subject to

$$L_i \leq \sum_{j=1}^n a_j^i x_j \leq U_i, \quad i = 1, \dots, m$$

(general constraints)

$$\ell_j \leq x_j \leq u_j, \quad j = 1, \dots, n$$

(individual bounds).

After suitable transformations and the introduction of a new variable for each general constraint, we obtain

$$z_i + \sum_{j=1}^n a_j^i x_j = b_i, \quad i = 1, \dots, m, \quad (1)$$

where z_i 's are called *logical variables* while x_j 's are *structural variables*. Equation (1) can be written in matrix form

$$\mathbf{I}z + \mathbf{A}x = \mathbf{b},$$

or after redefining variables and the matrix of constraints

$$\mathbf{Ax} = \mathbf{b}. \quad (2)$$

Individual bounds can be translated such that the finite lower bounds are all zero. In vector form they can be written as $\ell \leq \mathbf{x} \leq \mathbf{u}$. Components of the lower bound vector belonging to free variables are $-\infty$. Details of bringing a general LP to this form can be found in Maros [2].

As matrix $\mathbf{A}$ of Equation (2) is of full row rank, there are m linearly independent columns of $\mathbf{A}$ that form a basis $\mathbf{B}$. Without restriction of generality, it can be assumed that they are the first m columns of $\mathbf{A}$, that is, $\mathbf{A} = [\mathbf{B} \ \mathbf{R}]$, where $\mathbf{R}$ denotes the “remaining” part of the matrix. With this notation the LP problem can be written in a partitioned form

$$\begin{aligned} \min \quad & z = \mathbf{c}_B^T \mathbf{x}_B + \mathbf{c}_R^T \mathbf{x}_R \\ \text{subject to} \quad & \mathbf{B}\mathbf{x}_B + \mathbf{R}\mathbf{x}_R = \mathbf{b} \end{aligned} \quad (3)$$

$$\ell_B \leq \mathbf{x}_B \leq \mathbf{u}_B \quad (4)$$

$$\ell_R \leq \mathbf{x}_R \leq \mathbf{u}_R. \quad (5)$$

Expressing $\mathbf{x}_B$ from Equation (3)

$$\mathbf{x}_B = \mathbf{B}^{-1}(\mathbf{b} - \mathbf{R}\mathbf{x}_R). \quad (6)$$

Definition 1. A basis $\mathbf{B}$ (basic solution $\mathbf{x}_B$) is *feasible* if $\mathbf{x}_B$ satisfies Equations (3)–(5).

Definition 2. A basis (basic solution) is *degenerate* if at least one of the basic variables has a value equal to its lower or upper bound.

An immediate consequence of this definition is that a free basic variable ($-\infty \leq x_{Bi} \leq +\infty$) never causes degeneracy as it has no individual bound, while a fixed basic variable ($\ell_{Bi} = x_{Bi} = u_{Bi}$, e.g., z_i of an equality constraint) always makes the basis degenerate.

Let $\mathcal{D}$ denote the set of basic positions that contribute to the degeneracy of the basis. We can say that the basic solution is nondegenerate if $\mathcal{D} = \emptyset$. Otherwise, it is degenerate. The *degree of degeneracy* is the number of degenerate positions: $|\mathcal{D}|$.

RATIO TEST IN THE SIMPLEX METHOD

If the current basis is not optimal the simplex method chooses a nonbasic variable that violates its optimality condition to enter the basis as it has the potential of improving the objective function. The outgoing (leaving, or exiting) variable must be chosen in such a way that the other basic variables remain feasible after the basis change. This is achieved by the application of the ratio test as described by Dantzig [3] for the standard LP problem (all constraints are equalities, all variables are nonnegative, $x_j \geq 0, \forall j$).

Here, we consider a more general version of the ratio test [2] that takes into account finite lower and/or upper bounds on variables.

As the incoming variable, say x_q , improves the objective function proportionally to its displacement t , the purpose is to move it away from its current nonbasic value (lower or upper bound) as far as possible. This movement is stopped either by a basic variable that would go infeasible or the individual upper bound u_q of x_q itself.

In Equation (6) the values of basic variables are uniquely determined by the values of nonbasic variables. As some nonbasic variables can be at upper bound

$$\mathbf{x}_B = \mathbf{B}^{-1} \left(\mathbf{b} - \sum_{k \in \mathcal{U}} u_k \mathbf{a}_k \right), \quad (7)$$

where $\mathbf{a}_k$ is column k of $\mathbf{A}$ in Equation (2) and $\mathcal{U}$ is the index set of nonbasic variables at upper bound

$$\mathcal{U} = \{k : k \in \mathcal{R}, x_k = u_k\}.$$

Introducing notation

$$\mathbf{b}_U = \mathbf{b} - \sum_{k \in \mathcal{U}} u_k \mathbf{a}_k,$$

Equation (7) can be written as

$$\mathbf{x}_B = \mathbf{B}^{-1} \mathbf{b}_U, \text{ or } \mathbf{B}\mathbf{x}_B = \mathbf{b}_U. \quad (8)$$

If the improving variable changes from x_q to $x_q + t$ with $t > 0$, then basic variables also

change their values to maintain the main equations (Eq. 8)

$$\begin{aligned} \mathbf{B}\mathbf{x}_B(t) + t\mathbf{a}_q &= \mathbf{b}_U, \text{ or} \\ \mathbf{x}_B(t) &= \mathbf{B}^{-1}\mathbf{b}_U - t\mathbf{B}^{-1}\mathbf{a}_q. \end{aligned} \quad (9)$$

Using notations $\alpha_q = \mathbf{B}^{-1}\mathbf{a}_q$, $\beta = \mathbf{B}^{-1}\mathbf{b}_U$, and $\beta(t) = \mathbf{x}_B(t)$, the second equation of Equation (9) becomes

$$\beta(t) = \beta - t\alpha_q. \quad (10)$$

The i th component of Equation (10) is

$$\beta_i(t) = \beta_i - t\alpha_{iq}.$$

Let λ_i and φ_i denote the lower and upper bounds of the i th basic variable, respectively. The feasibility of the i th basic variable is maintained as long as the displacement t of the incoming variable is such that

$$\lambda_i \leq \beta_i - t\alpha_{iq} \leq \varphi_i. \quad (11)$$

It is straightforward to see that Equation (11) is satisfied, if t is the minimum of θ^+ and θ^- where

$$\theta^+ = \min_{\alpha_{iq} > 0} \{t_i\} = \min_{\alpha_{iq} > 0} \left\{ \frac{\beta_i - \lambda_i}{\alpha_{iq}} \right\} \quad (\text{type-A ratio}) \quad (12)$$

and

$$\theta^- = \min_{\alpha_{iq} < 0} \{t_i\} = \min_{\alpha_{iq} < 0} \left\{ \frac{\beta_i - \varphi_i}{\alpha_{iq}} \right\} \quad (\text{type-B ratio}). \quad (13)$$

However, to determine the final length of displacement (steplength) the individual bound of the selected incoming variable must also be taken into account. It results in the actual steplength of

$$\theta = \min\{\theta^+, \theta^-, u_q\}. \quad (14)$$

If the minimum is attained by a type-A ratio the outgoing variable becomes nonbasic at its lower bound, if the minimum is defined by a type-B ratio the outgoing variable exits at its upper bound. If $\theta = u_q$ there is no basis

change, x_q is moved to its opposite bound (bound flip). This is the *generalized ratio test*.

The development is similar if $t < 0$ is the improving displacement. In either case, the value of the objective function changes as

$$\hat{z} = z + \theta d_q, \quad (15)$$

where d_q is the reduced cost of the incoming variable (on the basis of which it was selected as an improving candidate).

If the basis is degenerate, a basic variable is at its lower or upper bound. In this case the numerator of the ratio test in Equation (12) or (13) can be zero if α_{iq} is of the “wrong” sign making the overall steplength $\theta = 0$. It also means that the incoming variable will remain at its bound maintaining the degeneracy of the corresponding basic position, unless a free variable enters at zero level. Thus, according to Equation (15), the value of the objective function remains unchanged. (Note, if the iteration is a bound flip the objective always improves.) Then it can happen that the ratio test of the new iteration also gives $\theta = 0$. It means, while the basis keeps changing the objective does not improve. After a number of such iterations the algorithm may return to an already visited basis. If the selection rule for the *entering variable* remains unchanged the algorithm will go through the same sequence of nonimproving steps indefinitely. This phenomenon is called *cycling*. There is also a possibility that a long sequence of zero steplength iterations may end in an improving one. This is called *stalling*.

Cycling prevents the simplex method from termination, while stalling “just” slows down the progress. From theoretical point of view, it is the possibility of cycling that has to be eliminated by the introduction of proper entering/exiting rules for a basis change. This will ensure the theoretical finiteness of the algorithm.

A basic solution can become degenerate in two ways. The trivial case is when the ever present unit matrix of logical variables is chosen as the starting basis and the $\mathbf{b}$ vector has some values (usually zeros) that are equal to one of the bounds of the corresponding logical variables. In this case the basis

will be degenerate. The other possibility is that a basis becomes degenerate as a result of an iteration. Namely, if the minimum in Equation (14) is not unique then we choose one of the positions in the tie and the corresponding basic variable leaves the basis. However, all other basic variables that are in the tie go to their bounds making the new basis degenerate.

There are two things to be noted about degeneracy.

1. Degeneracy does not always cause a nonimproving iteration. If the degenerate positions do not take part in the ratio test (if $\alpha_{iq} = 0$ or of the “wrong” sign) the steplength can be different from zero.
2. Degeneracy can disappear. During improving iterations (as above) basic values of the degenerate positions also get transformed. If the corresponding $\alpha_{iq} \neq 0$ the basic variable moves away from its boundary and can take a nondegenerate value.

ANTIDEGENERACY VERSIONS OF THE SIMPLEX METHOD

The possibility of cycling in both the primal and the dual simplex algorithm was early recognized and small examples were created [4,5] where cycling did occur. Even in 2004 a new example of cycling was published [6].

Cycling in the simplex method is possible only if degeneracy is present. At a degenerate basic solution (vertex) the number of active constraints is greater than the minimum necessary. It is illustrated in Fig. 1, where constraints a, c , and d are active in vertex D . Thus, D can be represented as the intersection of any two of these constraints. Therefore, the representation is not unique. Typically, a degenerate iteration just moves from one representation to another without improving the objective function. It has been shown [7,8] that the problem of exiting a degenerate vertex is as hard as the general linear programming problem.

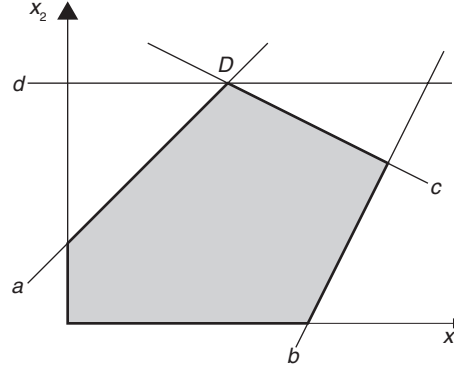


Figure 1. A two dimensional convex polyhedron with degeneracy at vertex D .

Perturbation and Lexicographic Methods

Cycling cannot occur if the objective is improved in every iteration. It can “artificially” be guaranteed if, for instance, a controlled perturbation is applied to a degenerate basis. Dantzig [3] used this idea and developed the *perturbation method* for the simplex method which theoretically ensures that no basis is repeated. The proposed algorithm turned out to be fully equivalent to the *lexicographic simplex method* of Dantzig *et al.* [1]. Both algorithms require considerable amount of extra computations. Here, we just present the outline of the methods. For a complete discussion the reader is referred to the indicated sources.

The key feature of the methods is the rule how to choose the exiting variable in case the minimum in the ratio test is not unique. To break the tie rows of the inverse of the basis need to be used in the following way. For simplicity of the presentation (but without loss of generality), we assume all variables in the problem are nonnegative, $x_j \geq 0, \forall j$. Now the ratio test reduces to the evaluation of “ $\beta_i/\alpha_{iq}, \alpha_{iq} > 0$ ” type quantities. Let the inverse of the basis be denoted by $\mathbf{V} = [v_{ij}]$.

θ of the ratio test is defined by

$$\theta = \frac{\beta_p}{\alpha_{pq}} = \min_{\alpha_{iq} > 0} \left\{ \frac{\beta_i}{\alpha_{iq}} \right\}. \quad (16)$$

If the minimum ratio is attained for more than one i , index p (the basic position of the outgoing variable) is not uniquely defined.

Assume there are more $i = p_1, p_2, \dots$ indices that satisfy Equation (16). Choose from these i -s one to be the position of the exiting variable p such that

$$\frac{v_{p1}}{\alpha_{pq}} = \min_{i=p_1, p_2, \dots} \left\{ \frac{v_{iq}}{\alpha_{iq}} \right\}. \quad (17)$$

If p is still not unique a further ratio test is performed involving those positions that were tied in Equation (17). This recursive procedure is repeated until the actual ratio is uniquely defined. It will determine the exiting variables. The depth of recursion is no more than m as shown in Dantzig *et al.* [1] and Dantzig [3].

The heavy computational price that must be paid for both methods is the use of rows $p_1, p_2, \dots$ of the inverse of the basis. In the revised simplex method (which is the basis of all simplex implementations) is used the inverse is not explicitly available. As such, the required rows have to be determined by $\mathbf{e}_i^T \mathbf{B}^{-1}$ (often called *BTRAN*) type operations for the $i = p_1, p_2, \dots$ indices.

Bland's Rule

Bland has proposed and proved [9] an interesting finite version of the simplex method. The beauty of his method is its simplicity. It is commonly referred to as *Bland's rule*.

For the standard LP (minimization) the rule says the following. If a basis is not optimal then choose the variable x_q having the lowest index among improving candidates, that is,

$$q = \min\{j : d_j < 0\}.$$

In column q choose the outgoing basic variable on the basis of the ratio test. If there is a tie at the minimum select the one with the lowest index. In other words, the basic position of the outgoing variable is determined by

$$\begin{aligned} p &= \min \left\{ k : \alpha_{kq} > 0, \text{ and } \frac{\beta_k}{\alpha_{kq}} \right. \\ &= \left. \min \left\{ \frac{\beta_i}{\alpha_{iq}} : \alpha_{iq} > 0 \right\} \right\}. \end{aligned}$$

The selection of both the entering and leaving variables is the simplest possible.

As a result, the computational work per iteration is the smallest among all strategies. Unfortunately, Bland's rule does not perform well in practice. It usually requires many (often three to five times) more iterations than any other selection method. Therefore, it is not practical to use it. However, its theoretical importance remains undisputed.

Wolfe's "ad hoc" Method

There is a method that is theoretically sound (guarantees finite termination) and computationally attractive. It is Wolfe's *ad hoc method* [10]. In the following description it is assumed, again, that all variables of the problem are nonnegative and $\mathbf{B}$ is a feasible basis with $\boldsymbol{\beta} = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. The degeneracy set of basic variables is defined as

$$\mathcal{D} = \{i : \beta_i = 0\}.$$

If the basis is not optimal and the improving variable is x_q the solution is unbounded if x_q is not blocked by any of the basic variables, that is, $\alpha_{iq} \leq 0, \forall i$. This direction is called a *direction of recession*. If such a direction, restricted to the degenerate positions, can be found a nondegenerate iteration can be made. This observation is the key motivation of the *ad hoc* method. In other words, a direction of recession is sought for $\mathcal{D}$.

The algorithm uses the notion of *depth of degeneracy*. At the outset every position in $\mathcal{D}$ has a depth of zero. Every time a position with the largest depth undergoes a "virtual perturbation" the degeneracy deepens. On the other hand, when a direction of recession is found for the subproblem consisting of the deepest positions, degeneracy shallows. The only thing that needs to be noted of a position in $\mathcal{D}$ is its depth which will be denoted by $g_i, i \in \mathcal{D}$.

Initialization: Set $g_i = 0, i \in \mathcal{D}$. Assuming that an improving candidate x_q has been selected and α_q is available, one iteration of the algorithm can be described by the following steps.

- Step 1. For each i with $\beta_i = 0$, replace β_i by a small positive random number δ_i . Note, after random choice of δ_i the perturbed values remain feasible.

Replace g_i by $g_i + 1$ to signify that degeneracy has deepened.

- Step 2. Let $G = \max_i \{g_i\}$. Perform a ratio test with positions for which $g_i = G$. If a pivot row p is found with $\theta = \beta_p / \alpha_{pq}$ go to Step 3, otherwise, as all $\alpha_{iq} \leq 0, i \in \mathcal{D}$, go to Step 4.
- Step 3. Apply a special update. The outgoing variable leaves the basis at its exact bound. The basic variables in $\mathcal{D}$ are updated such that for any position with $g_i < G$ everything remains unchanged. For positions with $g_i = G$ replace β_i by $\beta_i - \theta \alpha_{iq}$ for $i \neq p$ and β_p by θ . Keep $g_p = G$. Basic variables outside $\mathcal{D}$ remain unchanged as the current pivot step corresponds to a degenerate iteration for the original problem. Choose a new incoming column and return to Step 1. If none can be chosen the current basis is optimal. Solution needs to be recomputed with restored bounds for variables in $\mathcal{D}$.
- Step 4. As a pivot row was not found, we have a direction of recession for the subproblem consisting of rows with $g_i = G$, consequently, the degeneracy has shallowed. If $G > 0$, for each i with $g_i = G$ replace β_i by zero and g_i by $g_i - 1$. Return to Step 2. Otherwise, if $G = 0$ the problem is unbounded.

From the above description it can be seen that degeneracy can deepen if the minimum ratio is not unique in the ratio test. However, as the iterations are always nondegenerate, the number of elements in the deeper level of degeneracy will be at least one less. It can go down no deeper than m where just one element can be present. The purpose of the random choice of the perturbation is to reduce the chance of deepening.

An important advantage of the method is that it does not introduce infeasibilities. Therefore, when a real nondegenerate step is made the solution remains feasible. Further advantages are the moderate increase in the

computational effort and the simplicity of implementation.

The *ad hoc* method performs very well in practice. The depth of degeneracy hardly ever goes beyond 2 if the perturbations are properly selected. Therefore, the amount of extra computations remains very moderate.

Heuristics

Interestingly, cycling hardly ever occurs in practice. At the same time, stalling is a relatively frequent phenomenon unless some strategy is used to shorten it. Ideas that help prevent cycling are usually beneficial also in case of stalling.

Examples of cycling use a predefined column selection rule. As there are many column selection (pricing) strategies [2, pp. 184–203] one can try to use different pricing in subsequent iterations. Depending on what combination of strategies is used it can work quite efficiently. However, there are more elaborate techniques that directly target degeneracy. Some of them suggest how to select incoming variables so that the chance of finding nondegenerate pivots increases [2, pp. 232–234], others deal with row selection [11]–[13] and use some form of numerical perturbation of degenerate positions. Most recently, “bound shifting” [2, pp. 243–244] has emerged as a simple but effective method to get through degenerate bases.

There are also techniques that aim at improving the efficiency of the simplex method while, as a side effect, they are quite powerful in making improving iterations even in the presence of degeneracy. Just to mention a few, the “steepest edge” type column selection strategies [11,14] try to select columns that promise the largest improvement of the objective function. Advanced pivot row selection methods in phase 1 of the primal simplex [15], [2, pp. 203–227], and in both phases of the dual [2, pp. 270–291; 16] have the capability to bypass degenerate pivots during row selection.

FURTHER READINGS

The literature has actively discussed various issues of degeneracy since 1952. Interested readers may find in Charnes [17], Greenberg

[18], Greenberg [19], Gal and Geue [20], and Gass and Vinjamuri [21] further details of degeneracy in the simplex method.

Experience has shown that LP relaxations of network and combinatorial optimization problems can be quite degenerate. In this category some excellent readings are Cunningham [22], Ryan and Osborne [23], and Goldfarb *et al.* [24].

REFERENCES

1. Dantzig GB, Orden A, Wolfe P. A generalized simplex method for minimizing a linear form under linear inequality constraints. *Pac J Math* 1955;5(2):183–195.
2. Maros I. Computational techniques of the simplex method. Boston (MA): Kluwer Academic Publishers; 2003. 325. pp.
3. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963. 625. pp.
4. Hoffman AJ. Cycling in the simplex method. Report No. 2974. Washington (DC): National Bureau of Standards; 1953.
5. Beale EML. Cycling in the dual simplex algorithm. *Nav Res Logist Q* 1955;2:269–276.
6. Hall JAJ, McKinnon KIM. The simplest example where the simplex method cycles and conditions where EXPAND fails to prevent cycling. *Math Program* 2004;100:133–150.
7. Megiddo N. A note on degeneracy in linear programming. *Math Program* 1986;35:365–367.
8. Gal T. Degeneracy in mathematical programming and degeneracy graphs. *ORiON* 1990;6:3–36.
9. Bland RG. New finite pivot rule for the simplex method. *Math Oper Res* 1977;2:103–107.
10. Wolfe Ph. A technique for resolving degeneracy in linear programming. *SIAM J Appl Math* 1963;11:205–211.
11. Harris PMJ. Pivot selection method of the devex LP code. *Math Program* 1973;5:1–28.
12. Benichou M, Gautier JM, Hentges G, *et al.* The efficient solution of large-scale linear programming problems. *Math Program* 1977;13:280–322.
13. Gill PE, Murray W, Saunders MA, *et al.* A practical anti-cycling procedure for linearly constrained optimization. *Math Program* 1989;45:437–474.
14. Forrest JJH, Goldfarb D. Steepest edge simplex algorithms for linear programming. *Math Program* 1992;57(3):341–374.
15. Maros I. A general Phase-I method in linear programming. *Eur J Oper Res* 1986;23(1):64–77.
16. Maros I. A piecewise linear dual procedure in mixed integer programming. In: Gianesi F, Komlosi S, Rapsak T, editors. *New trends in mathematical programming*. Dordrecht: Kluwer Academic Publishers; 1998. pp. 159–170.
17. Charnes A. Optimality and degeneracy in linear programming. *Econometrica* 1952;20:160–170.
18. Greenberg HJ. Pivot selection tactics. In: Greenberg HJ, editor. *Design and implementation of optimization software*. Alphen aan den Rijn: Sijthoff and Nordhoff; 1978.
19. Greenberg HJ. An analysis of degeneracy. *Nav Res Log Q* 1986;33:635–655.
20. Gal T, Geue F. A new pivoting rule for solving various degeneracy problems. *Oper Res Lett* 1992;11:23–32.
21. Gass SI, Vinjamuri S. Cycling in linear programming problems. *Comput Oper Res* 2004;31:303–311.
22. Cunningham WH. Theoretical properties of the network simplex method. *Math Oper Res* 1979;4(2):196–208.
23. Ryan DM, Osborne MR. On the solution of highly degenerate linear programmes. *Math Program* 1988;41:385–392.
24. Goldfarb D, Hao J, Kai S-R. Anti-stalling pivot rules for the network simplex algorithm. *Networks* 1990;20:79–91.

DELAYED RENEWAL PROCESSES

NILAY TANIK ARGON
Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

Let $\{X_n, n \geq 1\}$ be a sequence of independent nonnegative random variables such that $F(x) = \Pr\{X_1 \leq x\}$, for $x \geq 0$, and $G(x) = \Pr\{X_n \leq x\}$, for $x \geq 0$ and $n \geq 2$. Let also $S_0 = 0$ and $S_n = \sum_{i=1}^n X_i$, for $n \geq 1$, and define

$$N_D(t) = \sup\{n \geq 0 : S_n \leq t\}.$$

The stochastic process $\{N_D(t), t \geq 0\}$ is called a *delayed renewal process*. Hence, a delayed renewal process is essentially a renewal process except that the first interrenewal time has a distribution different from the others.

Example 1. Let $\{Z(t), t \geq 0\}$ be a continuous-time Markov chain with state space $S = \{0, 1, 2, \dots\}$. Let $N_D(t)$ be the number of visits to state $i \in S$ during $(0, t]$ starting from state $j \in S$ at time zero. If $i = j$, then $\{N_D(t), t \geq 0\}$ is a standard renewal process; otherwise, it is a delayed renewal process.

We will see in the following that delayed renewal processes possess almost all the properties of standard renewal processes.

TRANSIENT BEHAVIOR

To avoid trivialities, assume that $G(0) < 1$. We can then obtain the following result, which shows that $N_D(t)$ is a nondefective (or proper) random variable for any given $t \geq 0$.

Theorem 1. $N_D(t) < \infty$ with probability one for all $t \geq 0$.

We next study the distribution of $N_D(t)$ for some given $t \geq 0$. First let G_k be the k -fold convolution of G with itself, where

$$G_0(t) = 1$$

$$G_k(t) = \int_0^t G(t-u) dG_{k-1}(u), \text{ for } k \geq 1.$$

Theorem 2. Let $F * G_k(t)$ be the convolution of F and G_k , that is,

$$F * G_k(t) = \int_0^t F(t-u) dG_k(u), \text{ for } k \geq 0.$$

Then, for any given $t \geq 0$, we have

$$\Pr\{N_D(t) = 0\} = \Pr\{S_1 > t\} = 1 - F(t)$$

$$\Pr\{N_D(t) = k\} = \Pr\{S_k \leq t\} - \Pr\{S_{k+1} \leq t\}$$

$$= F * G_{k-1}(t) - F * G_k(t),$$

$$\text{for } k \geq 1.$$

It is also of interest to obtain an expression for the expected value of $N_D(t)$ for a given $t \geq 0$. Let $M_D(t) = \mathbb{E}[N_D(t)]$ be the renewal function of the delayed renewal process. Let also $M(t) = \mathbb{E}[N(t)]$ be the renewal function of a standard renewal process $\{N(t), t \geq 0\}$ with interrenewal distribution G . Conditioning on X_1 , we get

$$\mathbb{E}[N_D(t)|X_1 = x] = \begin{cases} 0 & \text{if } x > t \\ 1 + M(t-x) & \text{if } x \leq t. \end{cases}$$

Then, unconditioning, we get

$$M_D(t) = \int_0^\infty \mathbb{E}[N_D(t)|X_1 = x] dF(x)$$

$$= \int_0^t (1 + M(t-x)) dF(x)$$

$$= F(t) + \int_0^t M(t-x) dF(x). \quad (1)$$

Taking Laplace–Stieltjes transform (LST) of both sides of this equation yields

$$\tilde{M}_D(s) = \tilde{F}(s) + \tilde{M}(s)\tilde{F}(s),$$

where $\tilde{M}_D$, $\tilde{F}$, and $\tilde{M}$ denote the LST of M_D , F , and M , respectively. Now, using the fact that $\tilde{M}(s) = \tilde{G}(s)/(1 - \tilde{G}(s))$ (e.g., Section 8.3 in Kulkarni [1]), we get

$$\tilde{M}_D(s) = \frac{\tilde{F}(s)}{1 - \tilde{G}(s)}, \quad (2)$$

where $\tilde{G}$ denotes the LST of G . This expression can be useful in computing the renewal function of a delayed renewal process. However, unlike its counterpart for a standard renewal process, it does not uniquely characterize the underlying delayed renewal process.

LIMITING BEHAVIOR

Assume that $F(\infty) = 1$ and $G(\infty) = 1$. This assumption ensures that $\lim_{t \rightarrow \infty} N_D(t) = \infty$ with probability one. The following theorem provides the rate at which $N_D(t)$ goes to infinity as t becomes large. Let τ be the mean of X_n for all $n \geq 2$.

Theorem 3. *With probability one,*

$$\frac{N_D(t)}{t} \rightarrow \frac{1}{\tau} \text{ as } t \rightarrow \infty \text{ for } \tau < \infty.$$

The next theorem shows that the renewal function $M_D(t)$ goes to infinity (as $t \rightarrow \infty$) at the same rate as $N_D(t)$ does.

Theorem 4.

$$\frac{M_D(t)}{t} \rightarrow \frac{1}{\tau} \text{ as } t \rightarrow \infty \text{ for } \tau < \infty.$$

When the renewal argument is applied to a delayed renewal process, the outcome is generally an integral equation that takes the following form:

$$H_D(t) = C(t) + \int_0^t H(t-u) dF(u), \quad (3)$$

where $C(t)$ is a known function, $H(t)$ is a function associated with the standard renewal process with interrenewal distribution G , and $H_D(t)$ is the function to be determined. In the previous section, we have already seen one use of such an equation. In particular, Equation (1), which is an integral equation that satisfies the form given by Equation (3), yielded an expression for the LST of the renewal function (Eq. 2). The following theorem provides a result on the limiting behavior of $H_D(t)$.

Theorem 5. *Suppose that $H_D(t)$ is a function that satisfies Equation (3). If $C(t)$ and $H(t)$ are bounded functions with finite limits for large t , then $\lim_{t \rightarrow \infty} H_D(t) = \lim_{t \rightarrow \infty} C(t) + \lim_{t \rightarrow \infty} H(t)$.*

Theorem 5 can be useful for the asymptotic analysis of a delayed renewal process because $C(t)$ is generally a function with a known finite limit. Furthermore, in most cases, $H(t)$ is a function associated with a standard renewal process, and hence one could apply the key renewal theorem for standard renewal processes to obtain its limit. The following example demonstrates the use of Theorem 5.

Example 2. Let $B_D(t) = S_{N_D(t)+1} - t$ be the *remaining life* for the delayed renewal process at time $t \geq 0$. We will use Theorem 5 to show that the limiting distribution for $B_D(t)$ is the same as the limiting distribution for $B(t)$, which is the remaining life for a standard renewal process $\{N(t), t \geq 0\}$ with interrenewal time distribution G . First, let $H_D(t) = \Pr\{B_D(t) > x\}$ and $H(t) = \Pr\{B(t) > x\}$ for some $x > 0$ and $t \geq 0$. Conditioning on X_1 , we get

$$\Pr\{B_D(t) > x \mid X_1 = u\} = \begin{cases} H(t-u) & \text{if } 0 \leq u \leq t \\ 0 & \text{if } t < u \leq x+t \\ 1 & \text{if } u > x+t. \end{cases}$$

Unconditioning, we get

$$H_D(t) = \int_0^\infty \Pr\{B_D(t) > x \mid X_1 = u\} dF(u)$$

$$\begin{aligned}
 &= \int_0^t H(t-u) dF(u) + \int_{x+t}^{\infty} dF(u) \\
 &= 1 - F(x+t) + \int_0^t H(t-u) dF(u). \quad (4)
 \end{aligned}$$

Note that Equation (4) is of the form given in Equation (3) with $C(t) = 1 - F(x+t)$. Hence, using the fact that $\lim_{t \rightarrow \infty} C(t) = 0$ together with Theorem 5 yields

$$\begin{aligned}
 \lim_{t \rightarrow \infty} H_D(t) &= \lim_{t \rightarrow \infty} H(t) \\
 &= \frac{1}{\tau} \int_x^{\infty} (1 - G(u)) du. \quad (5)
 \end{aligned}$$

See for example, Section 8.6 of Kulkarni [1] for a proof of Equation (5).

Remark 1. As expected, the asymptotic results above do not contain the first interrenewal time distribution F .

EQUILIBRIUM RENEWAL PROCESSES

In this section, we consider a delayed renewal process for which F is the limiting distribution of excess life (Eq. 5) for a standard renewal process with interrenewal distribution G . To be more specific, suppose that

$$F(x) = \frac{1}{\tau} \int_0^x (1 - G(u)) du, \text{ for } x > 0. \quad (6)$$

We call this special delayed renewal process an *equilibrium renewal process* and denote it by $\{N_e(t), t \geq 0\}$. We first study the renewal function $M_e(t) = \mathbb{E}[N_e(t)]$. Using Equation (6), we find the LST of F as

$$\tilde{F}(s) = \frac{1 - \tilde{G}(s)}{\tau s}.$$

Now plugging this into Equation (2) gives $\tilde{M}_e(s) = 1/(\tau s)$, where $\tilde{M}_e$ is the LST of M_e . Inverting $\tilde{M}_e(s)$, we obtain the result given below.

Theorem 6. $M_e(t) = t/\tau$ for all $t \geq 0$.

Theorem 6 implies that $M_e(t)/t = 1/\tau$ for all $t \geq 0$, which is only asymptotically true for

other delayed and standard renewal processes, as in Theorem 4. The next theorem provides the distribution of the excess life $B_e(t)$ at time $t \geq 0$ for the equilibrium renewal process $\{N_e(t), t \geq 0\}$.

Theorem 7. For all $t \geq 0$ and $x > 0$,

$$\Pr\{B_e(t) \leq x\} = F(x) = \frac{1}{\tau} \int_x^{\infty} (1 - G(u)) du.$$

Thus, the excess life distribution at any given time $t \geq 0$ for an equilibrium renewal process is the same as the limiting distribution of excess life in a delayed or standard renewal process with the same interrenewal time distribution G . This leads to the following result, which explains why we call $\{N_e(t), t \geq 0\}$ an equilibrium renewal process.

Theorem 8. $\{N_e(t), t \geq 0\}$ has stationary increments.

Example 3. Suppose that $G(x) = 1 - e^{-\lambda x}$, for $x \geq 0$, that is the interrenewal times after the first renewal are exponentially distributed with rate λ . Suppose also that F satisfies Equation (6) and hence the resulting delayed renewal process is an equilibrium renewal process. In this case, it is easy to see that F and G are the same, and hence the equilibrium renewal process is indeed a stationary Poisson process.

FURTHER READING

The treatment of delayed renewal processes in this article follows Kulkarni [1] fairly closely. For omitted proofs and further examples, the interested reader is referred to Çinlar [2], Kulkarni [1], Ross [3], and references therein.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman and Hall; 1995.
2. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
3. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1996.

DEMAND RESPONSIVE TRANSPORTATION

SOPHIE N. PARRAGH
KARL F. DOERNER
RICHARD F. HARTL
Department of Business
Administration, University
of Vienna, Vienna, Austria

Demand responsive transportation systems are flexible (public) transportation services in which vehicle routes and schedules are generated on the basis of user requests. Demand responsive transportation also arises in private systems such as in on-demand aircraft services [1,2]. Existing public demand responsive systems are mostly dedicated to the transportation of the disabled, elderly, or ill [3]. In order to increase the efficiency of these systems, OR models and optimization techniques have been developed. In this article we focus on those passenger transportation problems where user requests consist of paired pickup and delivery locations, also known as *dial-a-ride problems* (DARPs). They fall under the broad umbrella of vehicle routing problems with pickup and delivery [4]. Classical pickup and delivery problems (PDPs) deal with the transportation of goods while DARPs deal with passenger transportation. Hence, DARPs focus on quality of service related issues. Following another classification scheme [5], DARPs belong to the one-to-one problem class. DARPs have received considerable attention in the literature. The first publications in the field of passenger transportation date back to the late 1960s and early 1970s [6–9]. A rather recent survey on DARPs was written by Cordeau and Laporte [10].

Following Cordeau [11], the basic DARP is modeled on a complete directed graph $G = (V, A)$, where V is the set of all vertices and A the set of all arcs. For each arc (i, j) a nonnegative travel cost c_{ij} and a

nonnegative travel time t_{ij} are considered. A total of n customer requests, each consisting of a pickup and delivery vertex pair $(i, n+i)$, are to be served by a set K of m identical vehicles with a capacity of C ; P denotes the set of pickup vertices and D the set of delivery vertices. All vehicles are stationed at a central depot denoted by 0 (origin depot) and $2n+1$ (end depot). Furthermore, every route has to be completed within a maximum route duration limit T . At every pickup vertex a certain number of passengers ($q_i > 0$) waits to be transported. The demand at every delivery vertex is equal to $q_{n+i} = -q_i$. The pickup vertex (origin) and/or the delivery vertex (destination) are associated with time windows $[e_i, l_i]$. In the case where different types of requests are considered, outbound requests, that is, from home to an appointment, have a tight time window only at the destination; and inbound requests, that is, from an appointment back home, have a tight time window only at the origin. The service time for loading or unloading operations at each vertex is denoted by d_i . Beginning of service at each vertex i by vehicle k , denoted as B_i^k , and the load when leaving vertex i with vehicle k , denoted as Q_i^k , are to be determined. The ride time of a client is calculated by $L_i^k = B_{n+i}^k - (B_i^k + d_i)$. In order to limit user inconvenience caused by long detours, a maximum user ride time $\bar{L}$ has to be respected. Finally, the binary variable x_{ij}^k is used to indicate whether arc (i, j) is traversed by vehicle k . The basic DARP can thus be formulated as the following 3-index program (based on Cordeau [11]):

$$\sum_{i \in V} \sum_{j \in V} \sum_{k \in K} c_{ij} x_{ij}^k, \quad (1)$$

subject to,

$$\sum_{i \in V} \sum_{k \in K} x_{ij}^k = 1 \quad \forall j \in P, \quad (2)$$

$$\sum_{j \in V} x_{ij}^k - \sum_{j \in V} x_{n+i,j}^k = 0 \quad \forall i \in P, k \in K, \quad (3)$$

$$\sum_{j \in V} x_{0j}^k = 1 \quad \forall k \in K, \quad (4)$$

$$\sum_{j \in V} x_{ji}^k - \sum_{j \in V} x_{ij}^k = 0 \quad \forall i \in P \cup D, k \in K, \quad (5)$$

$$\sum_{i \in V} x_{i,2n+1}^k = 1 \quad \forall k \in K, \quad (6)$$

$$x_{ij}^k = 1 \Rightarrow B_j^k \geq B_i^k + d_i + t_{ij} \quad \forall i \in V, j \in V, k \in K, \quad (7)$$

$$x_{ij}^k = 1 \Rightarrow Q_j^k \geq Q_i^k + q_j \quad \forall i \in V, j \in V, k \in K, \quad (8)$$

$$L_i^k = B_{n+i}^k - (B_i^k + d_i) \quad \forall i \in P, k \in K, \quad (9)$$

$$B_{2n+1}^k - B_0^k \leq T \quad \forall k \in K, \quad (10)$$

$$e_i \leq B_i^k \leq l_i \quad \forall i \in V, k \in K, \quad (11)$$

$$t_{i,n+i} \leq L_i^k \leq \bar{L} \quad \forall i \in P, k \in K, \quad (12)$$

$$\max\{0, q_i\} \leq Q_i^k \leq \min\{C, C + q_i\} \quad \forall i \in V, k \in K, \quad (13)$$

$$x_{ij}^k \in \{0, 1\} \quad \forall i \in V, j \in V, k \in K. \quad (14)$$

The objective function (1) minimizes total routing costs. Constraints (2–6) ensure that each request is served exactly once; each origin is visited by the same vehicle as its destination; and each vehicle's route starts at the origin depot and ends at the destination depot. Load propagation and consistency with respect to time variables is guaranteed by inequalities (7) and (8). User ride times are set by equalities (9). Maximum route duration restrictions, time window compliance, and maximum ride time limits are enforced by constraints (10–12); constraints (12) (together with (9)) also ensure that each request's pickup location is visited before its delivery location. Capacity limits are imposed by Equation (13).

A number of different versions of this standard DARP have been considered in the literature. Owing to strict space limitations we do not claim that the following overview of real-world aspects and solution methods in the field of DARPs is exhaustive. We, however, hope to provide a useful selection of articles, covering all important aspects.

REAL-WORLD ASPECTS

Most publications consider a static variant of the DARP (all transportation requests are known in advance). Taking the above stated 3-index formulation as a starting point, a number of variations exist. Many authors do not consider a maximum user ride time limit; instead time windows are placed on both the pickup and the delivery point, which yields an implicit (and, in most cases, more restrictive) ride time limit [12–14]. The according problem can be defined by removing all ride time related constraints given by Equations (9) and (12), and by introducing the following additional set of constraints on the beginning of service variables:

$$B_i^k \leq B_{n+i}^k \quad \forall i \in P, k \in K. \quad (15)$$

It ensures that user i 's pickup location is visited before his/her delivery location [previously enforced by Equations (9) and (12)].

In addition, different vehicles may travel at different speeds and travel costs may vary according to weight, size, and equipment. These two aspects are easily accommodated in the 3-index model. Homogeneous travel times and costs are simply replaced by vehicle dependent ones (c_{ij}^k, t_{ij}^k). However, vehicles may also differ with respect to the number of persons that can be transported. This aspect can be handled by adding a superscript k to the capacity limit C^k to distinguish the different vehicles. Finally, in many real-life applications, especially in the transportation of the handicapped or the ill, different transportation modes have to be considered. At least three such modes can be distinguished: seated, on a bed, or in a wheelchair. It leads to the introduction of a set of transportation modes R and the distinction of different capacities for each mode: C^r . Then, each user demands a place in the according mode q_i^r . In combination with several vehicle types, each vehicle k disposes of $C^{r,k}$ space for transportation mode r . Toth and Vigo [15] distinguished between seated passengers and passengers in wheelchairs; Beaudry *et al.* [16] provided for three types of passengers (seated, on a bed, and in a wheelchair); Parragh *et al.* [17] considered four modes of transportation

($R = \{0, 1, 2, 3\}$; 0 = staff seat, 1 = patient seat, 2 = bed, 3 = wheelchair). An accompanying person may also be present. They should use staff seats, may, however, also sit on patient seats or the bed; a seated patient should use the patient seat, may, however, also sit on the bed. Introducing variables $Q_i^{r,k}$, giving the load of resource r on vehicle k when leaving vertex i , these so-called upgrading options are modeled by the following inequalities (replacing Equations (8) and (13) in the above formulation):

$$x_{ij}^k = 1 \Rightarrow Q_j^{r,k} \geq Q_i^{r,k} + q_j^r \quad \forall i, j \in V, k \in K, r \in R, \quad (16)$$

$$\sum_{r'=r}^2 Q_i^{r',k} \leq \sum_{r'=r}^2 C^{r',k} \quad \forall i \in V, k \in K, r \in R \setminus \{3\}, \quad (17)$$

$$Q_i^{3,k} \leq C^{3,k} \quad \forall i \in V, k \in K. \quad (18)$$

Finally, different real-world situations also call for different objective functions. Many authors only consider user inconvenience in terms of constraints, that is, time windows, maximum (excess) user ride times, or maximum waiting times, while minimizing a cost related objective. Others incorporate the user inconvenience aspect into the objective function, either instead of or besides a cost related term. An objective function incorporating the difference between the actual and the shortest possible or direct ride times is considered by Bodin and Sexton [18] (in addition to the difference between actual and desired drop-off times), by Diana and Dessouky [19] (in addition to the total distance traveled and vehicle idle times), and by Jørgensen *et al.* [20] (in addition to route duration, total ride time, waiting time, time window, and other time related violations).

EXACT SOLUTION APPROACHES

Models (1–14), leaving aside route duration constraint (10), can be reformulated in more efficient ways. Ropke *et al.* [21] proposed two 2-index formulations for the PDP and the

DARP. Both exploit the fact that a homogeneous set of vehicles is considered, that is, time and load related conditions do not depend on which vehicle travels along which arc. All vehicle routes are thus defined on a single set of arc variables x_{ij} . The first formulation explicitly formulates all time- and load related constraints using time B_i and load Q_i variables (beginning of service constraints (7) also take care of subtour elimination; ride time constraints (9) and (12) take care of precedence). In the second formulation, these constraints are reformulated in the form of precedence constraints [22],

$$\sum_{i,j \in S} x_{ij} \leq |S| - 2 \quad S \in \mathcal{S}, \quad (19)$$

capacity constraints,

$$\sum_{i,j \in S} x_{ij} \leq |S| - \max\{1, \lceil |q(S)|/C \rceil\} \quad S \subseteq P \cup D, |S| \geq 2, \quad (20)$$

and infeasible path constraints,

$$\sum_{(i,j) \in A(F)} x_{ij} \leq |A(F)| - 1 \quad F \in \mathcal{F}, \quad (21)$$

where S denotes a vertex set; $\mathcal{S}$ denotes the set of all vertex subsets $S \subseteq V$ such that $0 \in S$, $2n+1 \notin S$, and $i \notin S$, $n+i \in S$ for at least one request i ; $q(S) = \sum_{i \in S} q_i$, $\mathcal{F}$ denotes the set of all time-infeasible paths, and $A(F)$ all arcs that are part of the infeasible path F . The advantage of the second reformulation is the reduced number of variables. Its major drawback consists in a loss of flexibility: time or capacity related issues can no longer be easily integrated into the objective function. Ropke *et al.* [21] used each 2-index formulation as the basis for a branch-and-cut algorithm, separating several additional families of valid inequalities. The latter proves to work better. Both outperform the branch-and-cut algorithm based on the 3-index formulation [11]. The largest instance solved to optimality in Cordeau [11] comprises 36 requests; Ropke *et al.* [21] solved instances with up to 96 requests. On the basis of these findings, Parragh [23] developed a 3-index and a 2-index formulation for the DARP with

heterogeneous users and vehicles, including upgrading (see above). The introduction of artificial start and end depots (one per vehicle) makes a 2-index formulation possible. Let $D_o = \{2n + 1, \dots, 2n + m\}$ denote the set of vehicle start depots and $D_d = \{2n + m + 1, \dots, 2n + 2m\}$ the set of vehicle end depots. The correct pairing of two depots $\{2n + k, 2n + m + k\}$ is ensured by

$$\sum_{i \in U} \sum_{j \notin U} x_{ij} \geq 1 \quad \forall U \in \mathcal{U}. \quad (22)$$

Here $\mathcal{U}$ is defined as the set of all vertex subsets $U \subseteq V$, such that for exactly one $k \in K$ the origin depot $2n + k \in U$ and the destination depot $2n + m + k \notin U$, while for all other $l \in K \setminus \{k\}$ origin depots $2n + l \notin U$ and destination depots $2n + m + l \in U$. To handle the heterogeneous capacity constraints, let $\mathcal{H}$ denote the set of infeasible paths $H = \{j_1, \dots, j_h\}$ with respect to load violations, such that $j_1 \in D_o$. A constraint similar to Equation (21) is used. Also in this case, the reformulation as a 2-index model leads to a more efficient branch-and-cut algorithm. The largest instance solved to optimality comprises 40 requests.

Other research in the field of exact methods for the DARP predominantly belongs to the dynamic programming domain and deals with single-vehicle versions of the DARP. The dynamic programming algorithm of Psaraftis [24] takes care of service quality by means of maximal position shift constraints compared to a first-come first-serve visiting policy; it is also applied to the dynamic case. A modified version of this algorithm, considering time windows, applies forward instead of backward recursion [25]. In a further forward dynamic programming algorithm [26], possible states are reduced by eliminating those that are incompatible with respect to vehicle capacity, precedence, and time window restrictions.

HEURISTIC SOLUTION APPROACHES

In a heuristic context, once the visiting order of pickup and delivery locations has been decided upon, the scheduling subproblem,

setting the beginning of service at each node, has to be solved. Especially in the context of maximum route duration and maximum user ride time constraints, its solution is not as straightforward as one may assume. The route duration part has already been investigated in the context of the vehicle routing problem with time windows: Savelsbergh [27] introduced the notion of forward time slack. It gives the amount of time by which the beginning of service at the origin depot can be shifted forward, yielding a schedule of minimum duration. On the basis of the forward time slack, Cordeau and Laporte [28] developed an eight-step evaluation scheme. It calculates a forward time slack for each pickup location part of the route to be evaluated. It is thus able to check whether or not a given node sequence can be visited in a feasible way with respect to time window, route duration, and ride time constraints. However, it does not necessarily yield the minimum average ride time for a given route. This issue is addressed in Parragh *et al.* [29]; the scheme of Parragh *et al.* [29] yields shorter ride times, but, in some cases, a route is considered as infeasible although a feasible schedule would exist if ride time increases were allowed. Also other authors have addressed the scheduling issue. Hunsaker and Savelsbergh [30], for example, proposed a fast feasibility check for the DARP, dealing with time window, waiting time, as well as ride time restrictions; it is also more restrictive in terms of feasibility than the evaluation scheme of Cordeau and Laporte [28].

Over the past decades, a large number of heuristic algorithms have been proposed for DARPs. Heuristic methods in the field of vehicle routing classify into classical heuristics and metaheuristics [31]. Classical heuristic methods comprise two-phase heuristics, construction heuristics, and improvement methods. The following overview of heuristic solution methods follows this classification scheme. We distinguish between heuristics for static and heuristics for dynamic and/or stochastic variants.

Heuristics for Static Variants

The term *static* indicates that all requests are known in advance of the planning. This

assumption is made by all research works summarized in the following.

Two-Phase Heuristics. One of the first heuristic solution procedures for a static multivehicle DARP was proposed by Cullen *et al.* [32]. They developed an interactive algorithm that follows the cluster-first route-second approach. It is based on a set partitioning formulation solved by means of column generation. The location-allocation subproblem is only solved approximately. Other cluster-first route-second approaches for different versions of the DARP are due to authors [18,33–36]. A classical cluster-first route-second algorithm was proposed by Borndörfer *et al.* [13]. Here, the clustering as well as the routing problem are modeled as set partitioning problems. The clustering problem can be solved optimally while the routing subproblems are solved approximately by a branch-and-bound algorithm (only a subset of all possible tours is used). An optimization-based miniclustering algorithm is presented in Ioachim *et al.* [37]. It uses column generation to obtain miniclusters and an enhanced initialization procedure to decrease processing times. Also the case of multiple depots is considered. The idea of miniclusters has been employed by a number of authors [12,38–40].

Construction Heuristics. A heuristic routing and scheduling algorithm for the single-vehicle DARP using Benders decomposition is described in Sexton and Bodin [41,42]. The scheduling problem can be solved optimally; the routing problem is solved with a heuristic algorithm. A minimum spanning tree heuristic for the single-vehicle case is described in Psaraftis [43]. Jaw *et al.* [44] proposed a sequential insertion procedure for the multivehicle case: customers are ordered according to increasing earliest time for pickup and are inserted using the cheapest feasible insertion criterion. Other construction heuristics are described by authors [45–48]. A regret insertion algorithm is due to Diana and Dessouky [19]. Regret insertion is also subject to analysis in a study by Diana [49] in order to determine why the performance of this heuristic is superior to that of other insertion rules. In

addition, beam search was used to solve the multivehicle DARP [50].

Construction-Improvement Heuristics. For the single-vehicle DARP, adaptations of 2-opt and 3-opt heuristics [51] and a 2-opt improvement scheme switching between optimizing and sacrificing phases [52] were developed. The construction-improvement concept was also used to solve multivehicle DARPs [14,53,54]. In the improvement phase operators such as trip insertion, exchange, and move are employed.

Metaheuristics. Among the most popular metaheuristic concepts applied to the DARPs are simulated annealing [55,56], tabu search [28,57,58], and genetic algorithms [20,59,60]. Toth and Vigo [15] described a further local search-based metaheuristic, a tabu thresholding algorithm, employing the neighborhoods defined in Toth and Vigo [53]. In Parragh *et al.* [61] a variable neighborhood search algorithm is described and applied to different versions of the DARP. It is compared to the tabu search algorithm of Cordeau and Laporte [28] and the genetic algorithm of Jørgensen *et al.* [20]. Similar variable neighborhood search algorithms are employed in Parragh [23] and Parragh *et al.* [17] for heterogeneous variants of the DARP and in combination with path relinking to a biobjective version [29].

Heuristics for Dynamic and Stochastic Variants

The term *dynamic* indicates that routing is done in real time, that is, new requests come in during the day and have to be scheduled into existing routes; *stochastic* versions integrate knowledge about the future in terms of probability functions into the planning.

Two Phase-Heuristics. In a cluster-first route-second framework, considering only the most urgent requests in a dynamic context, Colorni and Righini [62] solved the routing subproblems to optimality with a branch-and-bound algorithm. Also Caramia *et al.* [63] iteratively solved the single-vehicle subproblems to optimality, using a dynamic programming algorithm. Another cluster-first route-second approach, applying

dynamic programming in the routing phase, assigns new transportation requests to clusters according to least cost insertion [64]. Teodorovic and Radivojevic [65] proposed a two-stage fuzzy logic-based heuristic algorithm also following the cluster-first route-second idea. The first approximate reasoning heuristic decides which vehicle a new request is assigned to. The second heuristic handles the adjustment of the vehicle's route.

Construction Heuristics. Daganzo [66] analyzed three different insertion strategies: visiting the closest stop next; visiting the closest origin or the closest destination in alternating order; and inserting delivery locations only after a fixed number of users have been picked up. A multiobjective approach is followed in Madsen *et al.* [67], discussing an insertion-based algorithm called REBUS. The objectives considered are the total driving time, the number of vehicles, the total waiting time, the deviation from promised service times as well as cost. Fu [68] developed a parallel insertion algorithm for a stochastic version of the DARP, considering time-dependent travel times. The same author proposed a simulation environment to test future solution methods [69].

Construction-Improvement Heuristics. Coslovich *et al.* [70] proposed a two-phase insertion heuristic. A simple insertion procedure allows for quick answers with respect to inclusion or rejection of a new customer. The initial solution is improved by means of local search using 2-opt arc swaps. A dynamic DARP in a stochastic environment is considered in Xiang *et al.* [71], taking into account travel time fluctuations, absent customers, vehicle breakdowns, cancellation of requests, traffic jams, and so on. To solve this complex problem situation the heuristic of Xiang *et al.* [14] is adapted to the dynamic case. Horn [72] provided a software environment for fleet scheduling and dispatching of demand responsive services. Incoming requests are inserted into existing routes according to least cost insertion. A steepest descent improvement phase is run periodically. The system was tested in the modeling framework LITRES-2 [73].

Metaheuristics. Attanasio *et al.* [74] adapted the tabu search of Cordeau and Laporte [28] to the dynamic DARP by means of parallelization; and in Beaudry *et al.* [16] it is used to solve a dynamic DARP with additional real-world constraints. In Hanne *et al.* [75] a dynamic in-hospital version is solved with an evolutionary algorithm. A tabu search for a probabilistic DARP is discussed and compared to a construction-improvement heuristic in Ho and Haugland [76]. Berbeglia *et al.* [77] introduced a hybrid algorithm for the dynamic DARP, combining tabu search with constraint programming. Finally, a scenario-based variable neighborhood search algorithm, a multiple plan approach and a multi-scenario approach, both integrating a variable neighborhood search algorithm, for the dynamic stochastic DARP are discussed in Schilde *et al.* [78].

SUMMARY

State-of-the-art exact algorithms developed for the static DARP solve some instances with up to 96 requests [21]. However, the size of a test instance is not a really meaningful measure as tightly constrained instances are easier to solve than those with less tight constraints. In case of heuristic and metaheuristic methods, large parts of the proposed solution procedures are motivated by real-world problem situations. They differ regarding problem type (single and multidepot, homogeneous and heterogeneous fleet), constraints, and objective(s). Most of the solution algorithms for the dynamic DARP are based on repeated calls to static solution routines. In both cases, almost all solution methodologies were either tested on different data sets or, in the case where the same data set was used, varying, problem-specific (real-world) assumptions were made. Therefore, in either case, it is not possible to specify the best method developed so far. An overview of available test data sets can be found in Parragh *et al.* [4] and Parragh [79].

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under grant #L510-N13

and the Austrian Research Promotion Agency (FFG) under grant #826151 within the program IV2Splus. This support is gratefully acknowledged.

REFERENCES

1. Espinoza D, Garcia R, Goycoolea M, *et al.* Per-seat, on-demand air transportation part I: problem description and an integer multi-commodity flow model. *Transp Sci* 2008;42: 263–278.
2. Espinoza D, Garcia R, Goycoolea M, *et al.* Per-seat, on-demand air transportation part II: parallel local search. *Transp Sci* 2008; 42:279–291.
3. Diana M, Dessouky MM, Xia N. A model for the fleet sizing of demand responsive transportation services with time windows. *Transp Res B Meth* 2006;40:651–666.
4. Parragh SN, Doerner KF, Hartl RF. A survey on pickup and delivery problems. Part II: transportation between pickup and delivery locations. *J Betriebswirtschaft* 2008;58:81–117.
5. Berbeglia G, Cordeau J-F, Gribkovskaia I, *et al.* Static pickup and delivery problems: a classification scheme and survey. *TOP* 2007; 15:1–31.
6. Rebibo KK. A computer controlled dial-a-ride system. traffic control and transportation systems. *Proceedings of 2nd IFAC/IFIP/IFORS Symposium Monte Carlo*; 1974 Sept. Amsterdam: North-Holland Publishing Co.; 1974.
7. Wilson H, Weissberg H. Advanced dial-a-ride algorithms research project: final report. Technical Report R76-20. Cambridge (MA): Department of Civil Engineering, MIT; 1967.
8. Wilson H, Sussman J, Wang H, *et al.* Scheduling algorithms for dial-a-ride system. Technical Report USL TR-70-13. Cambridge (MA): Urban Systems Laboratory, MIT; 1971.
9. Wilson H, Colvin N. Computer control of the Rochester dial-a-ride system. Technical Report R-77-31. Cambridge (MA): Department of Civil Engineering, MIT; 1977.
10. Cordeau J-F, Laporte G. The dial-a-ride problem: models and algorithms. *Ann Oper Res* 2007;153:29–46.
11. Cordeau J-F. A branch-and-cut algorithm for the dial-a-ride problem. *Oper Res* 2006; 54:573–586.
12. Desrosiers J, Dumas Y, Soumis F, *et al.* An algorithm for mini-clustering in handicapped transport. Technical Report G-91-02. Montréal, Canada: HEC; 1991.
13. Borndörfer R, Grötschel M, Klostermeier F, *et al.* Telebus Berlin: vehicle scheduling in a dial-a-ride system. In: Wilson N, editor. *Proceedings of the 7th International Workshop on Computer-Aided Transit Scheduling*. Berlin: Springer; 1999. pp. 391–422.
14. Xiang Z, Chu C, Chen H. A fast heuristic for solving a large-scale static dial-a-ride problem under complex constraints. *Eur J Oper Res* 2006;174:1117–1139.
15. Toth P, Vigo D. Heuristic algorithms for the handicapped persons transportation problem. *Transp Sci* 1997;31:60–71.
16. Beaudry A, Laporte G, Melo T, *et al.* Dynamic transportation of patients to hospitals. *OR Spectr* 2009;32:77–107.
17. Parragh SN, Cordeau J-F, Doerner KF, *et al.* Models and algorithms for the heterogeneous dial-a-ride problem with driver related constraints. Technical Report no. CIRRELT-2010-13. University of Vienna, Faculty of Business, Economics and Statistics; 2009.
18. Bodin LD, Sexton T. The multi-vehicle subscriber dial-a-ride problem. *TIMS Stud Manage Sci* 1986;22:73–86.
19. Diana M, Dessouky MM. A new regret insertion heuristic for solving large-scale dial-a-ride problems with time windows. *Transp Res B Meth* 2004;38:539–557.
20. Jørgensen RM, Larsen J, Bergvinsdottir KB. Solving the dial-a-ride problem using genetic algorithms. *J Oper Res Soc* 2007;58: 1321–1331.
21. Ropke S, Cordeau J-F, Laporte G. Models and branch-and-cut algorithms for pickup and delivery problems with time windows. *Networks* 2007;49:258–272.
22. Ruland KS, Rodin EY. The pickup and delivery problem: faces and branch-and-cut algorithm. *Comput Math Appl* 1997;33:1–13.
23. Parragh SN. Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem. *Transp Res C Emer* 2010. DOI: 10.1016/j.trc.2010.06.002.
24. Psaraftis HN. A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transp Sci* 1980;14:130–154.
25. Psaraftis HN. An exact algorithm for the single vehicle many to many dial-a-ride problem with time windows. *Transp Sci* 1983;17: 351–357.

26. Desrosiers J, Dumas Y, Soumis F. A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. *Am J Math Manage Sci* 1986; 6:301–325.
27. Savelsbergh MWP. The vehicle routing problem with time windows: minimizing route duration. *ORSA J Comput* 1992;4:146–154.
28. Cordeau J-F, Laporte G. A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transp Res B Meth* 2003;37:579–594.
29. Parragh SN, Doerner KF, Gandibleux X, *et al.* A heuristic two-phase solution method for the multi-objective dial-a-ride problem. *Networks* 2009;54:227–242.
30. Hunsaker B, Savelsbergh MWP. Efficient feasibility testing for dial-a-ride problems. *Oper Res Lett* 2002;30:169–173.
31. Semet F, Laporte G. Classical heuristics for the capacitated VRP. In: Toth P, Vigo D, editors. *The vehicle routing problem*. Volume 9, SIAM monographs on discrete mathematics and applications. Philadelphia (PA): SIAM; 2002. pp. 109–128.
32. Cullen F, Jarvis J, Ratliff D. Set partitioning based heuristics for interactive routing. *Networks* 1981;11:125–143.
33. Stein DM. An asymptotic probabilistic analysis of a routing problem. *Math Oper Res* 1978;3:89–101.
34. Stein DM. Scheduling dial-a-ride transportation systems. *Transp Sci* 1978;12:232–249.
35. Wolfer Calvo R, Colorni A. An effective and fast heuristic for the dial-a-ride problem. *4OR* 2007;5:61–73.
36. Psaraftis HN. Scheduling large-scale advance-request dial-a-ride systems. *Am J Math Manage Sci* 1986;6:327–367.
37. Ioachim I, Desrosiers J, Dumas Y, *et al.* A request clustering algorithm for door-to-door handicapped transportation. *Transp Sci* 1995;29:63–78.
38. Desrosiers J, Dumas Y, Soumis F. The multiple vehicle dial-a-ride-problem. In: Daduna JR, Wren A, editors *Computer-aided transit scheduling*. Volume 308, *Lecture notes in economics and mathematical systems*. Berlin: Springer; 1988. pp. 15–27.
39. Dumas Y, Desrosiers J, Soumis F. Large scale multi-vehicle dial-a-ride problems. Technical Report G-89-30. Montréal, Canada: HEC; 1989.
40. Karabuk S. A nested decomposition approach for solving the paratransit vehicle scheduling problem. *Transp Res B Meth* 2009; 43:448–465.
41. Sexton T, Bodin LD. Optimizing single vehicle many-to-many operations with desired delivery times: I. Scheduling. *Transp Sci* 1985;19:378–410.
42. Sexton T, Bodin LD. Optimizing single vehicle many-to-many operations with desired delivery times: II. Routing. *Transp Sci* 1985;19:411–435.
43. Psaraftis HN. Analysis of an $O(n^2)$ heuristic for the single vehicle many-to-many Euclidean dial-a-ride problem. *Transp Res B Meth* 1983;17:133–145.
44. Jaw J, Odoni AR, Psaraftis HN, *et al.* A heuristic algorithm for the multi-vehicle advance-request dial-a-ride problem with time windows. *Transp Res B Meth* 1986;20:243–257.
45. Roy S, Rousseau JM, Lapalme G, *et al.* Routing and scheduling for the transportation of disabled persons : The tests. Technical Report TP 5598E. Montréal, Canada: CRT; 1985.
46. Roy S, Rousseau JM, Lapalme G, *et al.* Routing and scheduling for the transportation of disabled persons : the algorithm. Technical Report TP 5596E. Montréal, Canada: CRT; 1985.
47. Alfa AS. Scheduling of vehicles for transportation of elderly. *Transp Plan Tech* 1986;11:203–212.
48. Kikuchi S, Rhee J. Scheduling algorithms for demand-responsive transportation system. *J Transp Eng* 1989;115:630–645.
49. Diana M. Innovative systems for the transportation disadvantaged: towards more efficient and operationally usable planning tools. *Transp Plan Tech* 2004;27:315–331.
50. Potvin J-Y, Rousseau J-M. Constrained-directed search for the advance request dial-a-ride problem with service quality constraints. In: Balci O, Shrada R, Zenios ZA, editors. *Computer science and operations research: new developments in their interfaces*. Oxford: Pergamon Press; 1992. pp. 457–574.
51. Psaraftis HN. k-interchange procedures for local search in a precedence-constrained routing problem. *Eur J Oper Res* 1983;13: 391–402.
52. Healy P, Moll R. A new extension of local search applied to the dial-a-ride problem. *Eur J Oper Res* 1995;83:83–104.
53. Toth P, Vigo D. Fast local search algorithms for the handicapped persons transportation problem. In: Osman IH, Kelly JP,

- editors. *Metaheuristics: theory and applications*. Boston (MA): Kluwer Academic Publishers; 1996. pp. 677–690.
54. Wong KI, Bell MGH. Solution of the dial-a-ride problem with multi-dimensional capacity constraints. *Int Trans Oper Res* 2006;13:195–208.
 55. Colorni A, Dorigo M, Maffioli F, *et al.* Heuristics from nature for hard combinatorial optimization problems. *Int Trans Oper Res* 1996;3:1–21.
 56. Baugh JW, Krishna G, Kakivaya R, *et al.* Intractability of the dial-a-ride problem and a multiobjective solution using simulated annealing. *Eng Optim* 1998;30:91–123.
 57. Aldaihani M, Dessouky MM. Hybrid scheduling methods for paratransit operations. *Comput Ind Eng* 2003;45:75–96.
 58. Melachrinoudis E, Ilhan AB, Min H. A dial-a-ride problem for client transportation in a health-care organization. *Comput Oper Res* 2007;34:742–759.
 59. Uchimura K, Saitoh T, Takahashi H. The dial-a-ride problem in a public transit system. *Electron Commun Jpn* 1999;82:30–38.
 60. Rekiek B, Delchambre A, Saleh HA. Handicapped person transportation: an application of the grouping genetic algorithm. *Eng Appl Artif Intell* 2006;19:511–520.
 61. Parragh SN, Doerner KF, Hartl RF. Variable neighborhood search for the dial-a-ride problem. *Comput Oper Res* 2010;37:1129–1138.
 62. Colorni A, Righini G. Modeling and optimizing dynamic dial-a-ride problems. *Int Trans Oper Res* 2001;8:155–166.
 63. Caramia M, Italiano GF, Oriolo G, *et al.* Routing a fleet of vehicles for dynamic combined pickup and deliveries services. In: Chamoni P, Leisten R, Martin A, *et al.* editors. *Operations Research Proceedings 2001*. Berlin, Germany: Springer; 2002. pp. 3–8.
 64. Dial R. Autonomous dial-a-ride transit introductory overview. *Transp Res C Emer* 1995;3:261–275.
 65. Teodorovic D, Radivojevic G. A fuzzy logic approach to dynamic dial-a-ride problem. *Fuzzy Set Sys* 2000;116:23–33.
 66. Daganzo CF. An approximate analytic model of many-to-many demand responsive transportation systems. *Transp Res* 1978;12:325–333.
 67. Madsen OBG, Ravn HF, Rygaard JM. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Ann Oper Res* 1995;60:193–208.
 68. Fu L. Scheduling dial-a-ride paratransit under time varying, stochastic congestion. *Transp Res B Meth* 2002;36:485–506.
 69. Fu L. A simulation model for evaluating advanced dial-a-ride paratransit systems. *Transp Res A Pol* 2002;36:291–307.
 70. Coslovich L, Pesenti R, Ukovich W. A two-phase insertion technique of unexpected customers for a dynamic dial-a-ride problem. *Eur J Oper Res* 2006;175:1605–1615.
 71. Xiang Z, Chu C, Chen H. The study of a dynamic dial-a-ride problem under time dependent stochastic environments. *Eur J Oper Res* 2008;185:534–551.
 72. Horn MET. Fleet scheduling and dispatching for demand-responsive passenger services. *Transp Res C Emer* 2002;10:35–63.
 73. Horn MET. Multi-modal and demand-responsive passenger transport systems: a modelling framework with embedded control systems. *Transp Res A Pol* 2002;36:167–188.
 74. Attanasio A, Cordeau J-F, Ghiani G, *et al.* Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. *Parallel Comput* 2004;30:377–387.
 75. Hanne T, Melo T, Nickel S. Bringing robustness to patient flow management through optimized patient transports in hospitals. *Interfaces* 2009;39:241–255.
 76. Ho SC, Haugland D. Local search heuristics for the probabilistic dial-a-ride problem. *OR Spectr* 2009. DOI: 10.1007/s00291-009-0175-6.
 77. Berbeglia G, Cordeau J-F, Laporte G. A hybrid tabu search and constraint programming algorithm for the dynamic dial-a-ride problem. Technical Report no. CIRRELT-2010-14. Montréal, Canada: HEC; 2010.
 78. Schilde M, Doerner KF, Hartl RF. Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports. Technical Report. University of Vienna; 2010.
 79. Parragh SN. Ambulance routing problems with rich constraints and multiple objectives [PhD thesis]. University of Vienna, Faculty of Business, Economics and Statistics; 2009.

DESCRIBING DECISION PROBLEMS BY DECISION TREES

ROBERT F. BORDLEY
General Motors Research
and Development Labs,
Warren, Michigan

Industrial and Operations
Engineering Department,
University of Michigan, Ann
Arbor, Michigan

OPERATIONAL AND STRATEGIC DECISIONS

For operational decisions, Deming's four-step process (plan/do/check/act) involves

1. checking where actual system behavior differs from desired behavior,
2. analyzing the causes of these discrepancies,
3. planning corrective actions, and
4. implementing corrective action.

This is commonly represented as the Deming cycle (Fig. 1).

The Deming cycle is a scientifically informed trial and error process focused on continuous improvement. But when there are large time lags (which creates uncertainty about which interventions are helpful), it is useful to think through alternatives, think about future uncertainties, and think carefully about long term values. These time lags and uncertainties are especially common in strategic contexts. In these contexts, decision analysis (with decision trees) can be superior to the simpler continuous improvement process. But in other contexts where rapid learning is possible, the decision analysis process can be overly cumbersome. Thus both processes have their role in individual decision making (Fig. 2).

Like continuous improvement, decision analysis can be viewed as having four phases:

1. identifying the key deciders with the resources to make or respond to change;

2. having deciders articulate desires for improvements in the system and analyze these desires to identify the underlying values of the deciders;
3. analyzing how different interventions could impact how well the system performs (as measured by decider values) given all the relevant uncertainties; and
4. choosing (and committing to) a plan for intervening in the system.

With some reinterpretation, we can similarly represent decision analysis as a cycle (Fig. 3).

This cycle is useful in highlighting the four key elements of a decision:

1. *Deciders*. Creating a list of deciders that is too small can dramatically limit the decisions that one can feasibly implement and, all too frequently, lead the decision to be blocked or revised by stakeholders who are left out of the decision. Creating a list of deciders that is too large can make the decision-making process too cumbersome. Ideally the list of deciders, as well as the list of issues that are considered within the scope of decision, reflects the proper balance to responsibly, but efficiently, make a decision.
2. *Values*. Explicit enumeration of all the relevant values, including monetary values, safety issues, and impact on the quality of life of all the deciders is critical in guiding our decision. Ignoring critical values or failing to acknowledge trade-offs between certain values (e.g., safety vs. income) can clearly weaken the impact of the decision.
3. *Uncertain States*. Much of decision analysis is focused on the analytical evaluation of the consequences of various choices in various uncertain states. It requires the application of good logic, which—because of the complexity of probability theory—is

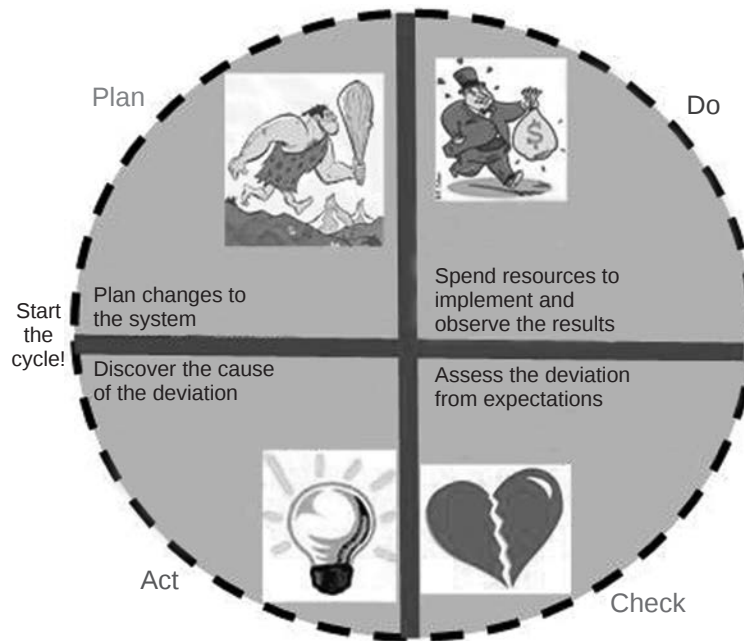


Figure 1. Deming cycle.

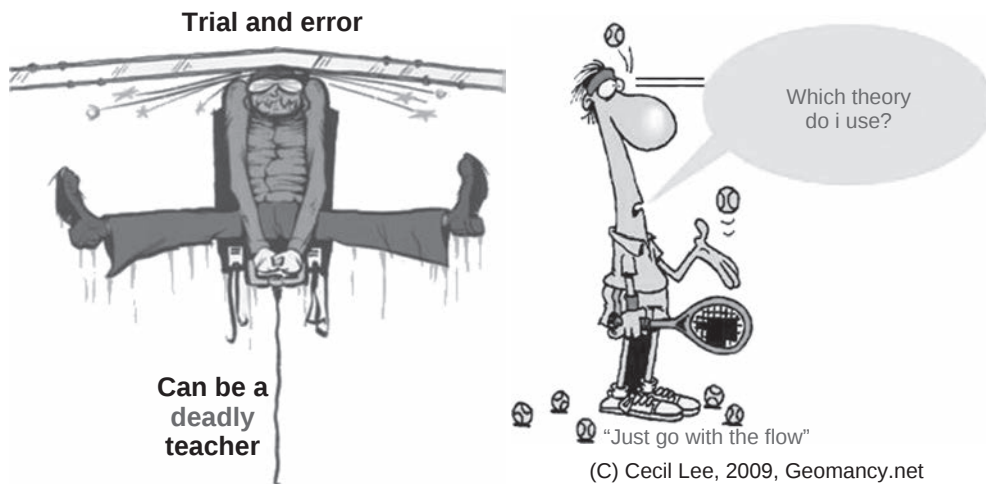


Figure 2. Perils of both continuous improvement and decision analysis.

often easy to violate. It also requires the collection of quality data—which can be especially onerous in real applications. Fortunately, decision analysis contains mechanisms, often

referred to as *value of information* analysis, which helped identify which data must be gathered and which data need not be gathered.

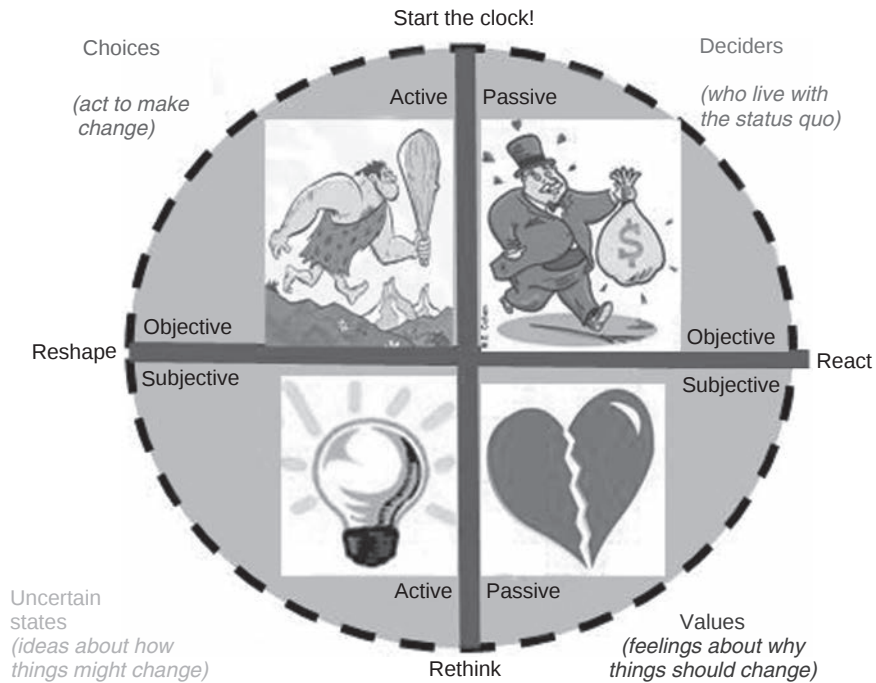


Figure 3. The decision cycle.

4. *Choices.* Choices are irreversible (and, as the caveman symbolizes, often violent) commitments to change the status quo. The deciders should brainstorm as many choices as possible. At the same time, the Archimedean axiom of decision theory requires that the decider screen out any choice which potentially leads to outcomes that are “infinitely worse” than the outcomes produced by other decisions. Thus the decider must screen out choices that are morally or legally unacceptable.

To ensure that deciders remember these four elements, diamonds are used to symbolize deciders with resources, hearts to symbolize values, spades to symbolize states of uncertainty, and clubs to symbolize choice. Hence any decider familiar with a deck of cards can remember the four elements.

Since failing to make good decisions is, in certain contexts, legally equivalent to negligence, most decision making has a moral dimension. As a result, the characteristics

of good decision making (including Matheson’s six characteristics of effective decision making) are decision “virtues” and are aligned with the classic virtues of justice, temperance, prudence, and courage (Fig. 4).

PROBLEM FRAMING

To ensure that all values relevant to the decision are considered, we draw a means/ends diagram that begins with each value and asks “why is that value important?” Repeatedly answering this question eventually leads us to a fundamental value which can subsume many of the other values that might have been voiced as important. For example, one person might say that money is important for the following reasons:

Making Money

is important because

I can live well;

This is important because

I enjoy physical pleasures.

Another person might say that making money is important for different reasons:

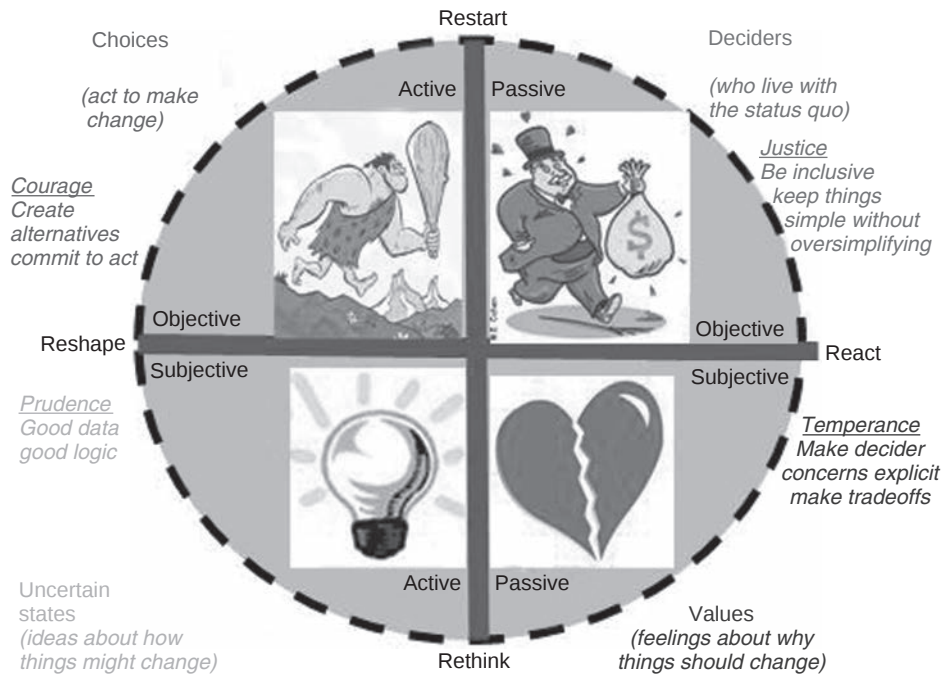


Figure 4. The decision clock.

Making Money

is important because

I can support the environment movement;

This is important because

I want future generations to have a clean world;

This is important because

I value the future of the human race.

Hence the same apparent interest in making money is based on potentially fundamentally different values. The means/ends diagram can also highlight other values that we consider important but failed to articulate. Once values have been identified, we define scales for measuring performance on each value.

We then assess the relevant states and decisions using an influence diagram. An influence diagram begins with a fundamental value and asks the question:

What are the small number of first-level factors which must be specified in order to know how well we performed on the fundamental value?

Once this question is answered, we then focus on each first-level factor and identify the small number of second-level factors required to specify performance on the first-level factor. We continue iteratively until the influence diagram has the requisite detail. In many cases, the analyst may have to completely scrap the influence diagram and draft a different influence diagram that makes different distinctions. *While tedious, making good distinctions is critical to making good decisions.* Since influence diagrams can become very detailed, contemporary influence diagram software like Analytica introduced modularity to allow the influence diagram to decompose into submodels. Analytica also introduced

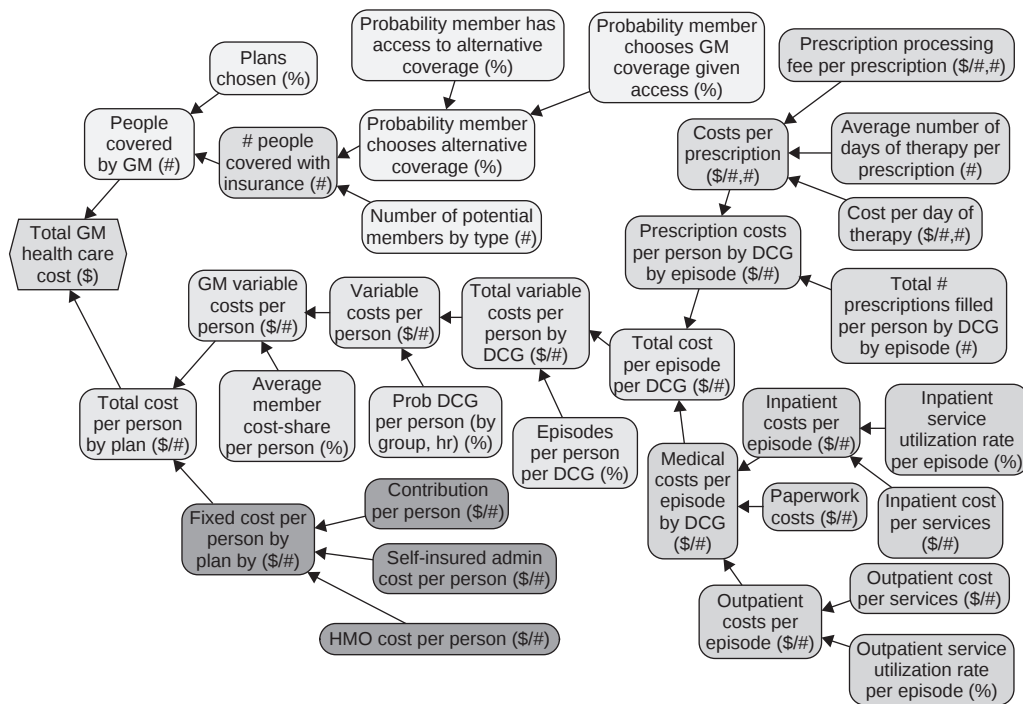


Figure 5. Drivers of health cost.

intelligent arrays that allow the analyst to quickly change the number of dimensions associated with each variable. The factors identified in an influence diagram can then be characterized as either controllable decisions or uncontrollable states. We only focus on those uncontrollable states whose value is unknown to us (uncertainties).

Figure 5 gives an example of an influence diagram focused on health-care costs. We first write health-care costs as a function of total cost per person by the health-care plan multiplied by the number of people covered by the health-care plan. We then write the total costs per person by plan as

a function of the variable costs per person and the fixed cost per person per plan. And we continue expanding in this way until our influence diagram has allowed us to identify possible actions and critical uncertainties in understanding our variable of interest, that is, total health-care costs.

In some contexts, the familiar arc and node representation of influence diagrams—which have sometimes been criticized as “spaghetti diagram”—is confusing. When the influence diagram can be decomposed in a treelike fashion, it is possible to represent it using the following outline representation:

Health Care Costs

People Covered

Plans Chosen

People Covered with Insurance

Number of Potential Members by Type

Probability Member Chooses Alternative Coverage

Probability Member has Access to Alternative Coverage

Probability Member Chooses Alternative Coverage Given Member Access

Total Cost per Person by Plan
 Fixed Cost per Person by Plan
 HMO Cost per Person
 Self-Insured Administration cost per person
 Contribution per Person
 Corporate Variable Costs per Person
 Average Member Cost-Share
 Variable Costs per Person
 Disease Category Probability per Person
 Variable Costs per Disease Category
 Episodes per Person by Disease Category
 Costs per Episode by Disease Category
 Drug Costs per Episode
 No. of Prescriptions per Episode
 Cost per Prescription
 Medical Costs per Episode
 In-patient Medical Costs per Episode
 Out-patient Medical Costs per Episode

The choice between these two representations clearly depends upon the audience. The second representation is less useful when there is one factor that affects several branches of the tree that expands out from the health-care cost node. But the author found the second representation useful with an audience more comfortable with an outline format.

The next critical step is assigning values to the end nodes (or to the lowest level categories in the outline format.) These values may be numbers, matrices, or multidimensional arrays. Once this is done, the relationship between the value assigned to a node and the values assigned to a node with arrows leading into that node (or the value assigned to a higher level category as a function of the values assigned to the lower level categories) must be quantified. This allows one to specify the topmost level (health-care costs) as a function of the lowest level costs. Note that the outline format is readily quantifiable in an MS Excel worksheet.

After quantifying the influence diagram, we specify how much each variable (whether it is a controllable variable or an uncontrollable variable) can be varied. We then typically conduct a Pareto analysis on the variables to identify which variables materially impact the topmost level variable (Fig. 6).

If the variable is an uncontrollable variable—which varies because of chance—then we may seek to gather more information on that variable. If the variable is an uncontrollable variable—which varies because the individual controlling that variable was not involved in our decision making—we may want to consider adding the individual to our group of deciders (the decision board). Alternatively, we may simply choose to treat that uncontrollable variable as an uncertainty.

We then eliminate all the variables that are less important so that we can restrict our focus to those variables, uncontrollable and controllable, that really matter. Thus the influence diagram helps us to determine the deciders, the relevant decisions, and the relevant states of uncertainties that will be included in our decision tree.

DECISION ANALYSIS: THE SIMPLEST CASE

The next step is to assign a time to each decision and uncertainty. The time for an uncertainty is the earliest time at which the outcome of that uncertainty is known to the deciders (or their representative). The time for a decision is the latest time at which that decision can be made without penalty. Thus, suppose you need to make a decision about whether to invest in some startup company.

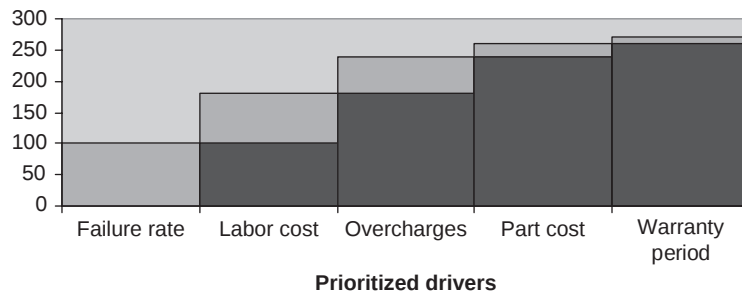


Figure 6. Cumulative Pareto chart (fake).

After you invest, you learn a few months later whether the company is viable or goes bankrupt. A few months after learning that the company is viable, you then learn how much profit the company makes. Thus the timeline is

Invest → Company Success/Failure?
→ Company Profitability

Note that the attractiveness of this investment opportunity would generally be quite different from the attractiveness of an opportunity where you learned about whether the company would go bankrupt before investing. In this case, the timeline would be

Company Success/Failure → Invest
→ Company Profitability

Note that only the sequence in which decisions are made and uncertainties are resolved is relevant in this ordering. There is no explicit consideration, at this point, of the time lag between investing and company success/failure. Instead, the time lag is implicitly considered in a later stage of evaluating the tree. If there is uncertainty in the time lag, then it is explicitly considered by introducing an uncertainty, “time until a relevant uncertainty is resolved” and treating that uncertainty like any other uncertainty. Thus the standard approach to treating decision trees differs from stochastic decision trees which attempt to represent the uncertainty in these time lags explicitly in the structuring of the tree.

The next step is listing the possible outcomes for each uncertainty and the possible outcomes of each decision. Thus there might be two outcomes for company success/failure, namely success or failure, and there might be two outcomes for company profitability, namely high or low. Finally, there might be two possible choices for each decision, “Invest” or “Don’t Invest.” Since there may be many ways in which the outcomes of an uncertainty could be decomposed, some care is required. Too detailed an elaboration of the possible outcomes of an uncertainty could make the analysis more difficult than necessary. Too limited an elaboration of the possible outcomes of an uncertainty might aggregate two outcomes which have very different characteristics.

This could be done through the visual tree where an uncertainty is represented by a rectangle, with the relative size of the rectangle representing the relative probability of that decision. Thus we might represent the investment problem as in Fig. 7.

In this representation, we represent our choice between either investing or not investing with two arrows, one going to the alternative, *Invest*, and the other going to the alternative, *Don’t invest*. The “Invest” alternative then leads to either “Success” or “Failure,” with the relative sizes of the rectangles indicating that success is more likely than failure. Finally, the “Success” rectangle leads to the “High payoff” and “Low payoff” rectangles. Since these two rectangles have roughly comparable sizes, the chance of high payoff given success is roughly the same as the chance of low payoff given success.

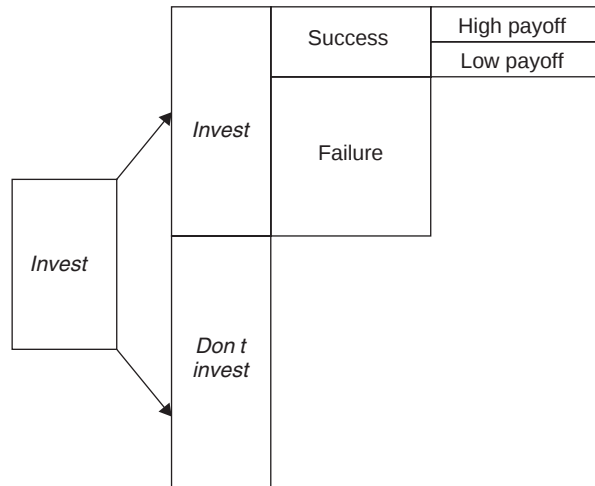


Figure 7. Visual decision tree.

We now need to evaluate the tree. While we constructed the tree beginning with the present and moving forward in time, we evaluate the tree by going to the last consequences and moving backward in time. More specifically, we evaluate the tree by first attaching values to the end consequences of the decision tree.

In order to do so, we first consider the simple case in which we are only interested in whether or not we achieve some consequence (e.g., enough money to retire comfortably.) In this case, the decision tree can be expanded to include the ultimate outcome of interest (Fig. 8).

In color theory, the value of a color is the amount of lightness in that color, with white representing the maximum value and black representing the minimum value. Following this convention, we now shade our endmost nodes by shading the “Retire” rectangle “white” and the “Don’t” rectangle black (Fig. 9).

To see why this is useful, note that the natural value to attach to the “Don’t invest” rectangle is the probability of being able to retire given no investment. The magnitude of this probability is just the color we get by mixing the white in the “Retire” rectangle with the “black” in the “Don’t” rectangle in proportion to the size of these two rectangles. (We can literally think of this as the result of dumping two small cans of white

and black paint to create a single can of an intermediate color.) We can repeat this mixing operation to shade the “Failure” rectangle and the “High payoff” and “Low payoff” rectangles. This gives Fig 10.

We next need to color the “Invest” rectangle. The “Invest” node has a small chance of leading to success, which has an even chance of leading to high payoff (which is almost white indicating that retirement is almost certain given this payoff) and an even chance of leading to low payoff (which is darker indicating a lower chance of achieving retirement given this payoff.) The color of the “Invest” rectangle should be the overall chance of achieving retirement given investment. This can be easily calculated from the adjacent two rectangles of “Success” and “Failure” by mixing their colors (in proportion to the size of the rectangles) as previously discussed. This leads to Fig. 11.

We now face a choice between investing and not investing. The “Invest?” rectangle differs fundamentally from the other rectangles in representing a decision. While the outcome high payoff or low payoff given success was determined by chance, the outcome of the “Invest” rectangle will be determined by choice. Hence we choose the rectangle of highest value, which in this case is not to invest. That rectangle has a dark gray color (while the alternative of investing has a nearly black color).

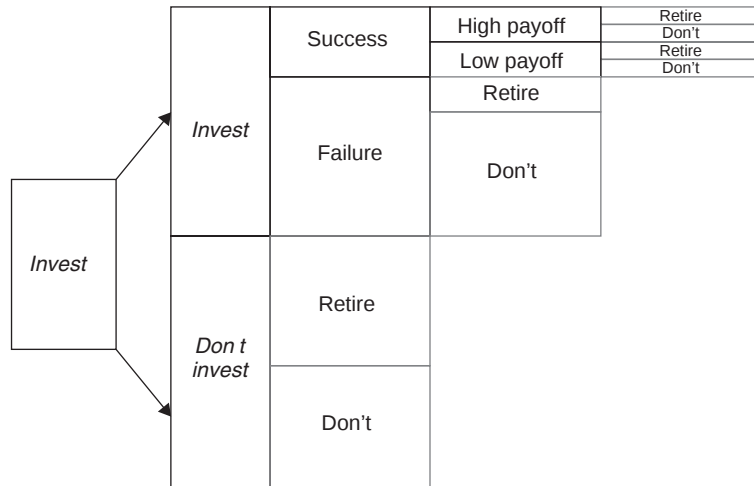


Figure 8. Visual decision tree with ultimate outcome.

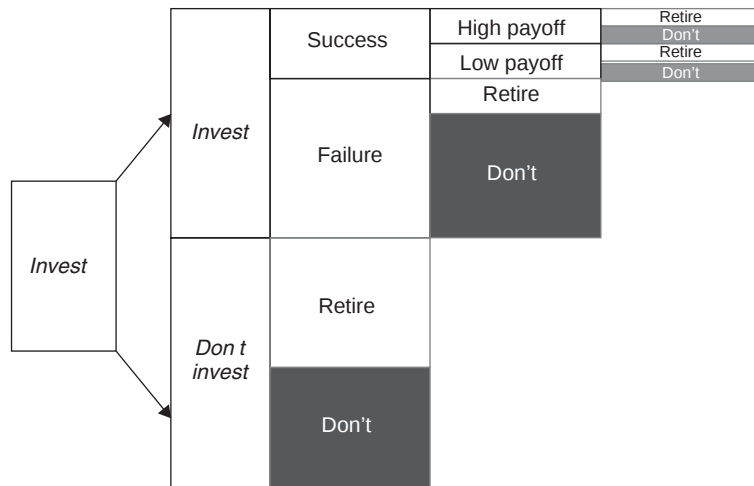


Figure 9. Valuing the end-nodes.

We can now erase the rectangles referring to this ultimate target to get the tree (Fig. 12).

In this problem, we valued the end rectangles in terms of an ultimate target that is either achieved or not achieved. There may also be cases in which there is an ultimate target but—within the

time frame of relevance—we will never eliminate all uncertainty about whether the target is achieved. For example, a physician subscribing to the Hippocratic oath may want to ensure that the life which the patient enjoys after treatment is clearly better than the life which an untreated

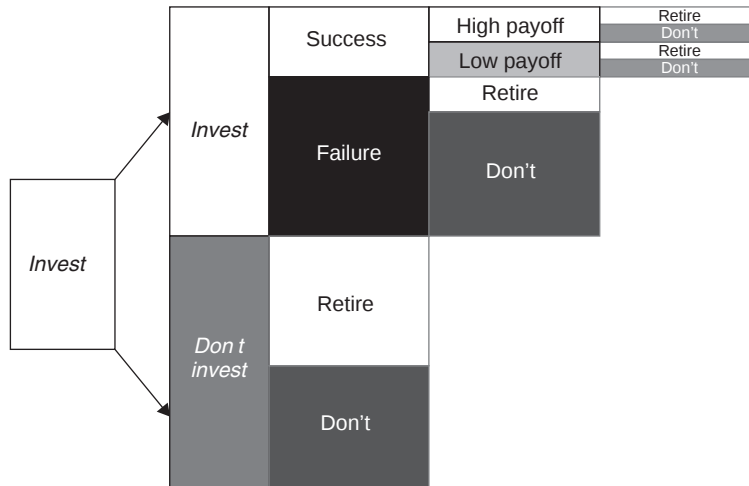


Figure 10. Valuing the nodes preceding the end nodes.

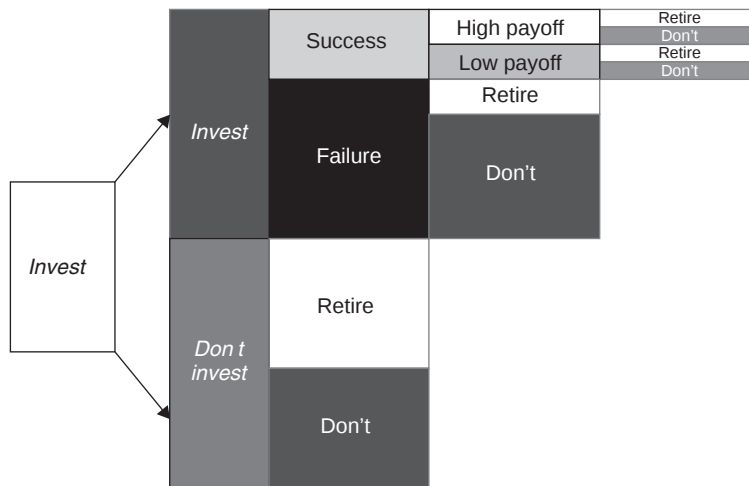


Figure 11. Valuing the nodes preceding the nodes in Figure 10.

patient would enjoy. Since the physician clearly cannot foresee the patient's future life, the requirements that the physician must achieve—in order to achieve minimal improvement in the patient's future life—are uncertain. But extending the previous approach to allow for uncertainty about the requirements needed to achieve the physician's target is straightforward.

DECISION ANALYSIS: THE GENERAL CASE

There are, of course, many problems for which there are no well-defined targets. In these cases, decision analysis first identifies the best possible outcome, for example, high payoff, and the worst possible outcome, for example, failure, and assigns them both some value, say, 0.9 and 0.1. (In practice, they are assigned values of

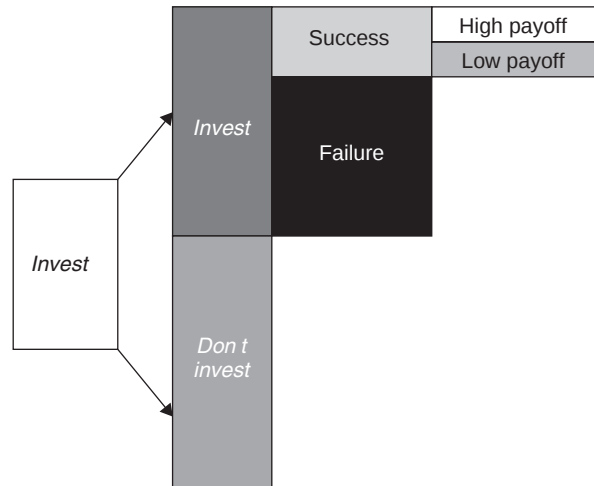


Figure 12. Valuing the remaining nodes.

zero and 1). To attach a value to any intermediate outcome, we ask the decider to imagine an experiment in which he/she must choose between the intermediate outcome and a lottery with some probability p of leading to the best possible outcome (and leading to the worst possible outcome otherwise). Decision analysis assumes there is some value of p for which the decider is indifferent.

To visualize this experiment, we imagine that the intermediate outcome, which we call risk, can lead to either the best possible outcome with this probability (and leads to the worst possible outcome.) As Fig. 13 specifies, this shows that the color we attach to this intermediate outcome is the color we gain by proportionately mixing the colors in the two adjacent rectangles.

This hypothetical experiment is what must be introduced when we cannot assume

that each outcome is ultimately only valued on the basis of whether it helps the decider achieve some fixed target.

Thus the problem of maximizing the probability of achieving the target (or the uncertain target) in the previous section is generalized to the problem of maximizing the expected value of the pseudoprobability (or preference probability) where the preference probability is associated with hypothetical gambles. Since any preference probability can always be interpreted as the probability of achieving some uncertain target (which might be purely hypothetical), one could view the above assessment procedure as a procedure for assessing the probability distribution of a hypothetical, uncertain target. Whether one should use the language of utility or the language of uncertain targets will depend upon which language is more meaningful to the client. In some cases, the

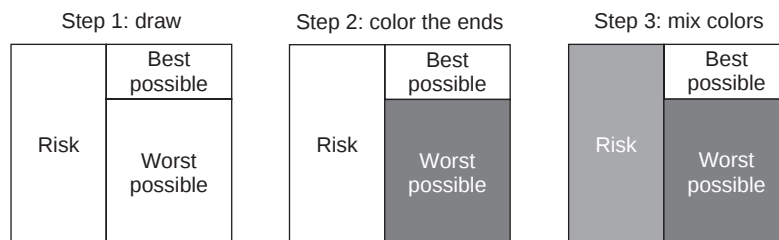


Figure 13. Computing expected utility visually.

notion of utility as some kind of measure of “happiness” will make utility the more natural notion. In other cases where clients are familiar with constraints and target, the language of uncertain targets will be more natural. Whatever the interpretation, this allows us to shade the end rectangles as shown in the tree in Fig. 14.

As previously noted, we proceed backwards in assigning values (or shadings to the end consequences). We assign a color to the success rectangle by mixing the shadings of the adjacent rectangles, high payoff and

low payoff. We then assign a shading to the “Invest” rectangle by mixing the shadings of the newly shaded “Success” rectangle and the “Failure” rectangle (Fig. 15).

As before, we make the decision to invest by choosing the brand of higher value (or lighter color). Thus a decision tree essentially replaces the complex uncertainty associated with any series of decisions by a two-outcome lottery (leading to either the best possible outcome or the worst possible outcome.) Since it is clear how to choose among such two-outcome lotteries (i.e., we always choose that

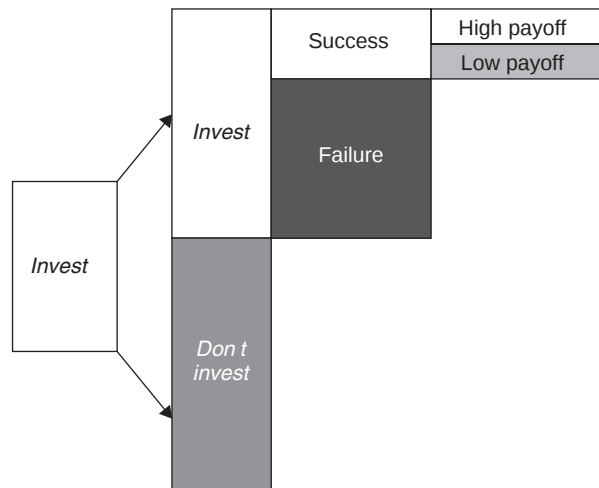


Figure 14. The visual decision tree with no ultimate outcome.

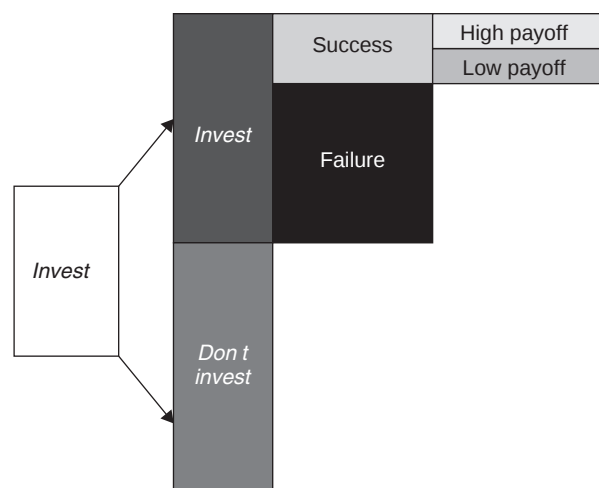


Figure 15. Valuing the nodes in the tree.

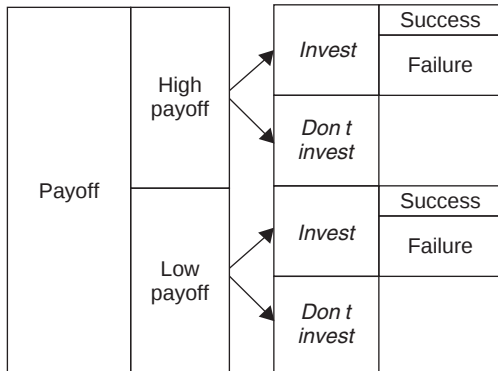


Figure 16. The visual tree with perfect information.

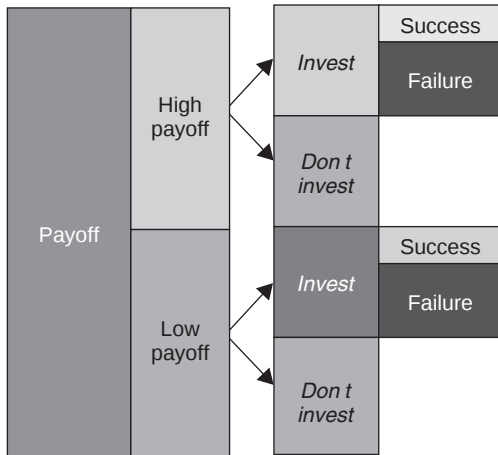


Figure 17. Valuing the visual tree with perfect information.

lottery with the highest probability of yielding the best outcome), this specifies a defensible procedure for decision making.

The order in which we need to make these decisions and resolve these uncertainties remains critical. For example, if we knew whether a successful investment would be high payoff before we made our decision, we would draw the tree as in Fig. 16, which, upon being appropriately shaded would have the form as in Fig. 17.

The fact that the color attached to the “Payoff” rectangle in this case is lighter than the color attached to the “Payoff” rectangle (when we decided before we knew the

outcome of a success investment) indicates that it is better to have this information earlier.

THE NUMERICAL DECISION TREE

The visual decision tree displays all the key elements of decision analysis. But since it uses colors and shapes, it is inherently less precise than a numerical representation since a number can be represented to an arbitrary degree of accuracy. In many problems, the precision implied by a number is misleading and hence the visual representation may be appropriate. But in dealing with low probability, high consequence risks, precision is essential. In such cases, the visual tree (Fig. 18) would be replaced by the numerical decision tree (Fig. 19).

In the numerical decision tree, all shapes have comparable sizes and colors, with numbers under the arcs indicating the probabilities and numbers at the end of each branches (or above the nodes) representing values. In reading this decision tree, we distinguish between decisions and uncertainties by representing decisions with rectangles and uncertainties with ovals. Above the arcs emanating from a decision, we list the decision alternative associated with that arc. Above the arcs emanating from an uncertainty (e.g., outcome), we list the possible outcomes (success, failure) of that uncertainty. Below the arc labeled success, we list the probability of the “success” event occurring—conditioned on all decisions and uncertainties preceding the outcome node occurring. Thus if we move to the arc labeled “Hi”, the probability of 0.5 is conditioned on a successful result from the outcome node and a “Yes” decision to invest. The joint probability of the branch associated with a successful outcome and a high payoff is the product of the probabilities of the arcs making up that brand. Thus the joint probability of a high payoff from a successful outcome given a “Yes” decision to invest is $(0.7) \times (0.5) = 0.35$.

The numbers at the end of each branch (e.g., the 0.4 at the end of the branch emanating from a “No” decision on investing) are

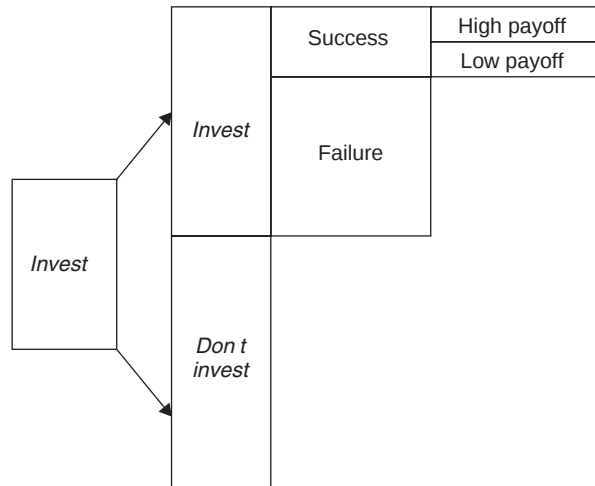


Figure 18. Comparison of the visual tree and the numerical tree.

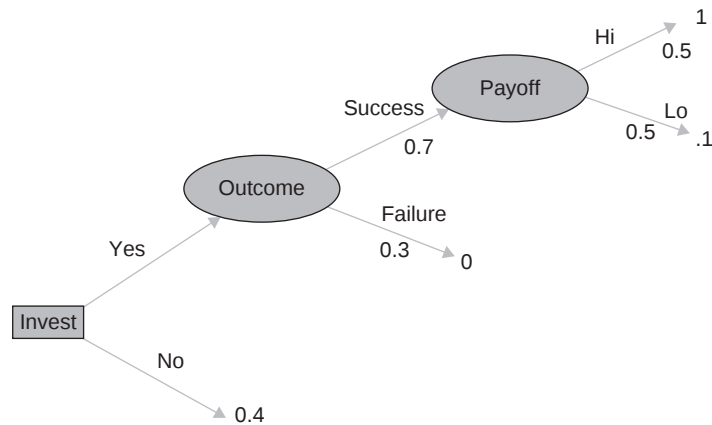


Figure 19. The numerical decision tree.

the values attached with all the events and decisions along that branch occurring. Thus if we do not invest, then we are indifferent between the outcome of this branch and a lottery offering us a 40% chance of the best possible outcome (and giving us the worst possible outcome otherwise). The value of 1 appearing at the end of the branch associated with a “Yes” on investing; a successful outcome and a high payoff implies that this outcome is the best possible outcome. The value of 0 appearing at the end of the branch associated with a “Yes” on investing, and a failed outcome implies that this outcome is the worst possible outcome. The 0.1 appearing at the end of the branch associated with a “Yes” on investing, a successful outcome, and

a low payoff indicates an outcome less desirable than not investing. (The 0.1 indicates that we are indifferent between that outcome and a 10% chance of the best possible outcome—with the worst possible outcome occurring otherwise.)

After assigning numbers at the end of the branches, we again move backward in the tree and assign values to the adjacent nodes. Instead of coloring the “Payoff” node by mixing colors from “end rectangles,” we compute $0.5 \times 0.1 + 0.5 \times 1 = 0.55$ and place that number above the “Payoff” node. Instead of coloring the outcome node by mixing colors from a “Failure” rectangle and a “Payoff” rectangle, we compute $0.7 \times 0.55 + 0.3(0) = 0.385$ and place that

number above the outcome node. Finally instead of coloring the “Invest” rectangle with the lighter of the two colors of the adjacent rectangles, we write the largest of the two values from the adjacent rectangles, that is, the maximum of 0.385 and 0.4 above the “Invest” rectangle. The value associated with the optimal decision is then 0.4. Hence this opportunity is equivalent to a 40% chance of getting the best possible outcome (investing, succeeding, and getting the high payoff) and a 60% chance of getting the worst possible outcome (investing and failing.)

This more analytical representation can itself be replaced by a more general representation when uncertainties are continuous. In this case, we rewrite the tree in collapsed form as

$$\begin{aligned} \text{Max(Invest)} &\rightarrow E[\text{Outcome}|\text{Invest}] \\ &\rightarrow E[\text{Payoff}|\text{Invest, Outcome}] \end{aligned}$$

where $E[\text{Outcome}|\text{Invest}]$ indicates that we compute the probability-weighted average (or expected value) of our payoff over the possible outcomes given we invest and given we don't invest and where Max(Invest) indicates that we compute the maximum of the value given invest and given we don't invest.

Note that when we learn the payoff of a successful outcome before investing, the tree must be written as

$$\begin{aligned} E[\text{Payoff}] &\rightarrow \text{Max(Invest}|\text{Payoff)} \\ &\rightarrow E[\text{Outcome}|\text{Invest, Payoff}] \end{aligned}$$

Since the individual first learns about the payoff, he/she then makes the investment decision and then learns the outcome of the investment.

In the visual decision tree, we represent the value attached to end points by shades of gray. The shade of gray corresponds to the probability which would make one indifferent between the outcome colored gray and a gamble which has that probability of leading to the best possible outcome (colored white) and which leads to the worst possible outcome (colored black). In the conventional decision tree, we represent the value attached

to end points by numbers. But as with the visual tree, these numbers correspond to the probabilities which would make one indifferent between the current outcome and a gamble which has that probability of leading to the best possible outcome and which leads to the worst possible outcome otherwise. As previously noted, these numbers are called *utilities*.

In many applications, individuals prefer to work with cash values rather than the more abstract notion of utility. Once we ascertain the utility of an opportunity (i.e., the value of Max(Invest)), this can be computed by simply finding that cash value which has the same utility as Max(Invest) from the tree:

$$\begin{aligned} \text{Max(Invest)} &\rightarrow E[\text{Outcome}|\text{Invest}] \\ &\rightarrow E[\text{Payoff}|\text{Invest, Outcome}] \end{aligned}$$

If we are asked to compute the value of information on the payoff, then we would first solve the revised tree which assumes we know the payoff before making a decision:

$$\begin{aligned} E[\text{Payoff}] &\rightarrow \text{Max(Invest}|\text{Payoff)} \\ &\rightarrow E[\text{Outcome}|\text{Invest, Payoff}] \end{aligned}$$

We would then find the cash value which has the same utility as $E[\text{Payoff}]$. Finally, we compute the value of information on payoff as the difference between the cash value of $E[\text{Payoff}]$ and the cash value of Max(Invest) . This is the amount by which having information on the payoff before we make the decision has increased the cash value of the opportunity.

While decision trees represent all quantities in terms of lotteries giving rise to either the best outcome or the worst outcome, it is common to try to approximate them with decision trees that work in terms of deterministic quantities, for example, cash values. (In target-oriented utility theory, this corresponds to assuming that the individual wishes his wealth to exceed a uniformly distributed random variable.) In this approximation, we replace all utility values by cash values. Thus we begin by assigning cash values to the end points of the tree (Fig. 20).

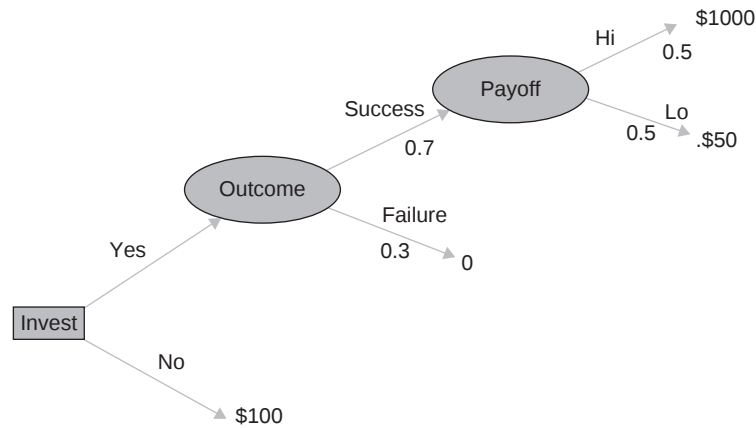


Figure 20. Expected value with the numerical tree.

We then do the same calculations using cash values as we would with utility values.

Thus the value assigned to the “Payoff” node is the probability-weighted average of \$50 and \$1000, or \$525. The value attached to the outcome node is the probability-weighted average of \$525 and 0, or \$367.50. The value attached to the “Invest” node is the maximum of \$367.50 and 0, or \$367.50. Note that in this tree we choose to invest, whereas in the tree with utilities we chose not to invest. This highlights the fact that the expected cash value rule is only a heuristic which is sometimes a good approximation for the result that would be achieved using the utility rule. One way of measuring the error in the approximation is by examining the difference between the cash value associated with $\text{Max}(\text{Invest})$ from the tree with utilities and the cash value associated with the tree using expected cash values. This is called the *risk premium*. An individual for whom the risk premium is zero, that is, for whom the expected cash value rule always gives the same results as the exact expected utility rule, is said to be risk neutral. In this case, utility is proportional to the cash value.

The expected cash value rule is also an excellent rule for decision trees applied to evaluating stock market derivatives (as well as for real options not directly connected to the stock market decisions). In these cases,

the probability associated with a given uncertainty (described by a security in the stock market) must be replaced by a risk-neutral probability. This risk-neutral probability is defined in terms of the values attached to derivatives and the underlying securities and has no relationship to the actual chance of the event occurring. Nonetheless, with these risk-neutral probabilities, maximization of the expected cash values can be fully justified.

CONCLUSIONS

The decision tree is an excellent tool for reducing complicated problems involving considerable uncertainty to more transparent choices between simpler gambles. Our review of decision trees differed from traditional reviews of decision trees in three ways:

1. *Explicit Comparison with Continuous Improvement.* Since decision analysis is designed for addressing strategic (and not operational) decisions, this article goes into more detail on Deming’s process as an operational alternative.
2. *Targets as the Simplest Case.* Traditional courses on decision analysis begin by teaching the rule of maximizing expected cash value, then telling students the rule is flawed, and then

teaching expected utility maximization. As an alternative, this article begins with maximizing the chance of achieving a target—a decision criterion already widely understood—and then introduces the expected utility criterion to describe cases where targets are not well defined. This also makes it easier to describe the rationality assumptions underlying decision theory (if desired). The expected value criterion is introduced last as a critical special case.

3. *Visual versus Numerical Representation.* Many clients cannot handle numbers or believe that using numbers requires precision. In response, this article begins with visual decision tree representation (where great precision is not possible) and transitions to numerical representations which allows for great precision.

Hence this article provides a rigorous but user-friendly perspective on decision trees.

DESCRIPTION OF THE FRENCH OPERATIONAL RESEARCH AND DECISION-AID SOCIETY: *SOCIÉTÉ FRANÇAISE DE RECHERCHE OPÉRATIONNELLE ET D'AIDE À LA DÉCISION* (ROADEF)

OLIVIER HUDRY
École nationale supérieure
des télécommunications,
Paris, France

PRESENTATION OF THE ROADEF

Formed in 1998, the *Société française de recherche opérationnelle et d'aide à la décision* (ROADEF; <http://www.roadef.org>) is the French society of operational research (OR) and decision-aid. It is a member of the *Association of European Operational Research Societies* (EURO), itself within the *International Federation of Operational Research Societies* (IFORS). The ROADEF aims to promote, in France, all the activities related to the OR and the decision-aid in all possible fields such as teaching, research, industries, civil and military applications, and so on. In this respect, the ROADEF carries on the work of the *Association française pour la cybernétique économique et technique* (AFCET). Founded in 1968 by the merging of four previous scientific associations, the AFCET was the main association dealing with computer science and information theory during the 1970s and the 1980s. The AFCET gathered several groups; one of them was devoted to OR.

In 1998, while the AFCET disappeared, 20 founder members created the ROADEF.

- Some of them were individuals such as Denis Bouyssou (ESSEC), Jacques Carlier (Université technologique de Compiègne), Philippe Chrétienne (Université Paris VI), Stéphane Dauzère-Pérès (École des

Mines de Nantes), Dominique Fortin (INRIA), Xavier Gandibleux (Université de Valenciennes), Nelson Maculan (Universidade Federal do Rio de Janeiro), Ali Ridha Mahjoub (Université Clermont-Ferrand II), Gérard Plateau (Université de Paris-Nord), Jean-Charles Pomerol (Université Paris VI), Marie-Claude Portmann (Institut national polytechnique de Lorraine), Christian Proust (Université de Tours), Catherine Roucairol (Université de Versailles), and Bernard Roy (Université de Paris-Dauphine).

- The others were legal (industrial or academic) entities such as Bouygues, Euro-Décision Technologies, Gaz de France, LAMIH/ROAD (Université de Valenciennes), LAMSADE (Université de Paris-Dauphine), and SNCF.

After a period of about 10 years, the ROADEF has gathered between 300 and 400 members (390 in 2009), from universities or industrial companies. Members can be researchers, engineers, Ph. D students (64 in 2009), retired persons (3 in 2009), institutions or companies. We should also mention two honorary members—Bernard Roy and Marie-Claude Portmann—for their great involvement in OR or in the life of the ROADEF.

THE SUCCESSIVE BOARDS OF THE ROADEF

Marie-Claude Portmann was the first president of the ROADEF. More precisely, the board of the ROADEF has six members, elected by the other members of the ROADEF for 2 years, with at most two mandates: one president (P), three vice presidents (VP), one secretary (S), and one treasurer (T). From the birth of the ROADEF, the boards were composed of the following:

- 1998–1999: Marie-Claude Portmann (P), Alix Munier (S), Laura Wynter (T),

Jean-Robert Leroy (VP), Xavier Gandibleux (VP), and Alexis Tsoukias (VP);

- 2000–2001: Denis Bouyssou (P), Alix Munier (S), Ariane Partouche (T), Stéphane Dauzère-Pérès (VP), Xavier Gandibleux (VP), and Jean-Robert Leroy (VP);
- 2002–2003: Arnaud Fréville (P), Marie-Christine Costa (S), Éric Gourdin (T), Philippe Baptiste (VP), Van-Dat Cung (VP), and Anass Nagih (VP);
- 2004–2005: Marie-Christine Costa (P), Jean-Charles Billaut (S), David de Almeida (T), Christian Artigues (VP), Safia Kedad-Sidhoum (VP), and Éric Sanlaville (VP);
- 2006–2007: Jean-Charles Billaut (P), Clarisse Dhaenens-Flipo (S), David de Almeida (T), Safia Kedad-Sidhoum (VP), Éric Sanlaville (VP), and Mohamed Ali Aloulou (VP);
- 2008–2009: Olivier Hudry (P), Clarisse Dhaenens-Flipo (S), Denis Montaut (T), Christelle Guéret-Jussien (VP), Francis Sourd (VP), and Mohamed Ali Aloulou (VP).
- The board for 2010 and 2011 will be Francis Sourd (P), Nadia Brauner (S), Denis Montaut (T), Christelle Guéret-Jussien (VP), Nathalie Sauer (VP), and François Vanderbeck (VP). Moreover, Éric Pinson will be in charge of the communication and promotion of OR in the view of the different partners of the ROADEF such as authorities, industrial companies, universities, research laboratories, and so on.

ANNUAL NATIONAL MEETINGS OF THE ROADEF

The ROADEF organizes a national meeting every year, in different places such as Paris (1998, 2002), Autrans (1999), Nantes (2000), Avignon (2003), Tours (2005), Lille (2006), Grenoble (2007), Clermont-Ferrand (2008), Nancy (2009), and Toulouse (2010). These meetings gather about 300 participants (for

instance, there were 320 participants in Nancy with 198 communications). A book of selected contributions involving students as authors or coauthors is published by the organizers and distributed to all the participants, as well as a booklet of all the abstracts. Since 2002, a selection of articles has been published after the meeting as a special issue of a journal, *Mathematics and Social Sciences* in 2002 (with I. Charon and O. Hudry as the invited editors), *RAIRO-Operations Research* for the meetings from 2002 to 2008 (with the following invited editors: I. Charon and O. Hudry in 2002; C. Artigues, D. Feillet, and P. Michelon in 2003; J.-C. Billaut and E. Néron in 2005; L. Brotcorne, C. Dhaenens, S. Hanafi, and T. El-Ghazali in 2006; C. Artigues, V.-D. Cung, G. Finke, and B. Penz in 2007; P. Mahey and A. Quilliot in 2008), the *European Journal of Industrial Engineering* for the meeting of 2009 (with A. Oulamara and M.-C. Portmann as the invited editors).

WORKING GROUPS OF THE ROADEF

The ROADEF organizes thematic workshops also, devoted to some specific subjects. They are often, though not always, linked to the activities of the working groups of the ROADEF. Indeed, the ROADEF supports several working groups, more or less linked to the ROADEF, including some regional groups. In 2009, these groups are devoted to the following subjects:

- combinatorial optimization (in collaboration with the European Chapter on Combinatorial Optimization—ECCO);
- complex systems and distributed decisions (in collaboration with the working group Modelisation, Analysis, Management of Dynamic Systems of the French National Centre of Scientific Research—CNRS);
- constraints and OR (in collaboration with the French Association for Programming by Constraints);
- knapsack and optimization;
- mathematics of optimization and decisions (in collaboration with the

working group Mathematics of Optimization and Decisions of the French Society of Industrial and Applied Mathematics—SMAI);

- multicriteria decision aiding (in collaboration with the EURO working group Multicriteria Decision Aiding);
- multiobjective mathematical programming (in collaboration with the working group Information–Interaction–Intelligence of the CNRS);
- OR in the Loire region;
- OR in the Paris region;
- OR in the north-east (gathering teams of the universities of Liège, Luxembourg, Metz, Nancy, Sarrebruck, and Trier);
- optimization in networks (in collaboration with the working group OR of the CNRS);
- parallel processing in OR (in collaboration with the EURO working group Parallel Processing in OR);
- polyhedron and combinatorial optimization (in collaboration with the working group OR of the CNRS);
- scheduling (in collaboration with the working group Modelisation, Analysis, Management of Dynamic Systems, and with the working group OR of the CNRS);
- supply chain and logistics (in collaboration with the working group OR of the CNRS);
- theory and applications of metaheuristics (in collaboration with the working group Modelisation, Analysis, Management of Dynamic Systems, with the working group Mathematical Computer Science of the CNRS, and with the European Chapter on Metaheuristics).

PUBLICATIONS INVOLVING THE ROADEF

The ROADEF is involved in the publication of two international journals such as *4OR*, a *Quarterly Journal of Operations Research*

(<http://homepages.vub.ac.be/~fouror/>) and *RAIRO-Operations Research* (<http://www.rairo-ro.org/>).

The journal *4OR* was founded in 2002 by the OR societies of Belgium (ORBEL/Sogesci-B.V.W.B; <http://www.orbel.be>), France (ROADEF), and Italy (AIRO; <http://www.airo.org/>). Its editors-in-chief were D. Bouys-sou (replaced by P. Baptiste in 2009), S. Martello (still editor-in-chief), and F. Plastria (replaced by T. Marchant in 2010). Its creation was accompanied by the disappearance of the former journals of the Belgian OR society (*JORBEL*) and of the Italian OR society (*Ricerca Operativa*).

The journal *RAIRO-Operations Research* was founded in the 1960s. It is devoted to high-level pure and applied research on all aspects of OR. Its current editor-in-chief is P. Mahey, assisted by five associate editors (A. Billionnet, A. Brandwajn, P. Chrétienne, M. Gendreau, and C. Gonzaga). The ROADEF and the SMAI are the scientific advisors of the journal.

Besides these two journals, the ROADEF also publishes a newsletter and a bulletin. The newsletter, edited electronically by one of the vice-presidents (C. Dhaenens-Flipo in 2009, N. Brauner in 2010), provides brief announcements as news about the association, new books, new conferences, and so on. It typically contains about three pages and is sent to the ROADEF's members, every second month. The bulletin, edited by another vice president (C. Guéret-Jussien in 2009, as well as in 2010) aims at providing a framework for discussions and exchanges inside the OR community, with two issues per year. Each issue typically contains between 20 and 30 pages, usually with an editorial describing the OR activities of an academic or industrial team, an invited scientific article (for instance, in December 2009, this article was written by G. Cornuéjols, on the occasion of his Dantzig prize), reports of some events, scientific or educational notes, as well as various contributions dealing with the activities of the ROADEF's working groups, with the conferences on OR, and more generally with any activity or event related to the ROADEF or OR.

ROBERT FAURE PRIZE

The ROADEF also sponsors the *Robert Faure Prize*, which existed even before the creation of the ROADEF (the first awards go back to 1993). Robert Faure was a professor of OR at the *Conservatoire national des arts et métiers* in Paris, where he held the chair of OR, a unique one in France. This prize is awarded every three years to three (sometimes four) young researchers for their original contributions to OR, especially for work combining theoretical developments and applications [see <http://www.roadef.org/content/roadef/rfaure.htm> for the list of the laureates; several of them became members of the board of the ROADEF: É. Pinson (1993), S. Dautère-Pérès (1996), P. Baptiste (1999), N. Sauer (1999), N. Brauner (2009), and F. Sourd (2003)].

CHALLENGES ORGANIZED BY THE ROADEF

Among the other achievements of the ROADEF, we should mention the “challenge”. The challenge is based on a subject proposed by an industrial or by a public or private institution. For instance, the subject of the 2010 issue is defined by *Électricité de France*; it deals with a large-scale production management problem with varied constraints. Then, different teams study the problem and design methods to solve the problem. There are two categories for the teams: junior and senior. The proposed methods are tested on benchmarks, which allow ranking of the teams and choosing the

winners according to the category (see the website of the ROADEF for the list of the winners). For the first time, in 2010, there will be a prize dedicated to parallel computing methods. Thanks to the success of the previous six challenges (in 1999, 2001, 2003, 2005, 2007, and 2009), the 2010 challenge is jointly organized with EURO. The finalists will be revealed during the ROADEF meeting at Toulouse, while the results will be announced and the winners will be awarded during EURO XXIV, in Lisbon.

The challenge is open to everyone, and particularly to young researchers, excluding people professionally involved with the industrial partners. The aims of the ROADEF’s challenges are multiple: They allow the industrial partners to witness recent developments in the field of OR. In the junior category, young researchers have the opportunity to face a complex optimization problem occurring in industry with its manifold constraints; the challenge gives them an opportunity to explore the requirements and difficulties encountered in industrial applications. In the senior category, the challenges allow qualified researchers to share their knowledge and know-how on practical problems. Moreover, the challenges help to establish or maintain a permanent partnership between industrials and researchers on industrial size projects requiring scientific and practical abilities.

Van-Dat Cung was the organizer of the first five challenges, between 1999 and 2007. He has been replaced by Murat Afsar, Christian Artigues, Éric Bourreau, and Olivier Briant for the 2009 challenge.

DESCRIPTIVE MODELS OF DECISION MAKING

MANEL BAUCELLS

Department of Economics and Business,
Universitat Pompeu Fabra,
Barcelona, Spain

KONSTANTINOS V. KATSIKOPOULOS

Center for Adaptive Behavior and
Cognition, Max Planck Institute for
Human Development, Berlin,
Germany

WHY BEHAVIORAL DECISION THEORY?

Decision theory builds on the premise that subjects seek and organize the information available to make a decision in a rational way. Rational here mean “according to a set of appealing principles.” The implications of and relationships between these principles are the area of study of decision theory. The objective of the field of decision theory and analysis is to propose decision rules and processes that might be used in practice to produce more informed and rational decisions. For instance, rational preferences have the property that decision makers are always able to produce consistent rankings of all alternatives. Thus, they are able to associate a utility score to each consequence, and choose the alternative yielding the highest utility. In the face of uncertainty, rational decision makers will calculate the average utility weighted by the subjective probabilities attached to the possible consequences. They will then choose the alternative that yields the highest expected utility.

In the rational model, the evaluation of the consequences and the estimation of probabilities are invariant to the way the consequences and the events are described, or the method employed to elicit these utilities and probabilities. Further, rational preferences are supposed to be dynamically consistent: the partial resolution of uncertainty or the

passage of time does not change the desirability of the residual alternatives. Moreover, factors such as the inclusion or not of past costs should be irrelevant for the decision at hand; only the future consequences matter. Finally, the quality of a decision is to be judged using the information available at the moment of the decision, and a bad outcome might not necessarily be associated to a bad decision. These principles are supposed to be *normative*, because any rational decision maker would follow them.

Common sense indicates that, for complex decisions, subjects might be incapable of producing rational decisions. Ultimately, we have limited capacity to process information. Complexity may arise due to the multiplicity of alternatives, uncertainties, or objectives. Precisely in complex problems is where the role of decision analysis is preeminent. In complex problems, the decision analyst will suggest to separate the problem in its basic elements (objectives, alternatives, uncertainties, consequences), elicit the utilities and probabilities, and then combine these elements to manufacture a decision that mimics that of a rational decision maker. Decision trees to clarify the interplay between uncertainties and alternatives, Monte Carlo techniques to simulate multiplicity of risks, or optimization methods to handle multiplicity of alternatives are part of the toolbox of decision analysis. Hence, decision theory has a *prescriptive* value because it helps decision makers create rational decisions in complex problems.

However, one fundamental question needs to be addressed before this program can be successfully implemented: do people want to behave rationally? The answer to this question seems so obvious, that for a long time the question was not even formulated. The assumption was that people would behave rationally in simple decisions, and the challenge was to extend this rationality to complex decisions. But what if subjects do not want to behave rationally even in simple problems? Then, these subjects will not

seek the aid of decision analysis for complex decisions, as they would feel that acting rationally runs against their preferences.

Behavioral decision theory (BDT) seeks to understand better the irrationalities of subjects. One clear objective of this program is to help subjects convince themselves that acting rationally is better [1,2]. Another objective is to propose *descriptive* utility models that are parsimonious and as predictive as possible of people's behavior. These models can be helpful in the fields of economics and the social sciences. For example, often decisions are not games against nature, but against a quasi-rational agent. We want to understand this quasi-irrational agent in order to set more accurate probabilities in the chance nodes. BDT is considered as one of the three pillars of decision analysis (the other two being the normative and the prescriptive pillars).

Often BDT uncovers irrationalities. But on occasions it may broaden our views on rationality and help the normative branch of decision analysis. For instance, descriptive deviations from rationality may reveal that the "real" consequences subjects care for are not what the naïve rational model proposes. For instance, habit formation models [3] together with lending and borrowing constraints justifies the rationality of preference for increasing sequences of income [4] (increasing consumption is the optimal pattern in a model with habit formation).

Finally, in nonmonetary decisions under certainty, the only rule of rationality is to predict as accurately as possible the experienced utility of each consequence [5]. BDT is also interested in developing predictive models of experienced utility. These models might outperform the predictions of subjects, and therefore gain immediate prescriptive status.

Rather than providing an extensive survey of the field, the aim of this article is to introduce the main findings of BDT. We utilize a very simple setup to illustrate what separates behavioral preferences from rational preferences. Specifically, we examine choices over prospects of the following form: receiving outcome x with a probability p at time t , and zero otherwise. The vector (x, p, t) will denote such a typical prospect. In the section

titled "Decisions under Risk and over Time: The Rational Model," we present the preferences of the rational decision maker, while in section titled "Decisions under Risk and over Time: The Behavioral Model," we contrast it with the preferences of the modal subject that we observe in experiments. In these two sections the focus is on how decision makers should or would treat probability and time. There, the probability p is objectively given, and we set x to some simple object, such as amount of money, or some quantity of a consumption good. Unless stated otherwise, x will be a gain or positive reward.

In the section titled "Estimating Subjective Probability" we consider the subjective estimation of p . Rather than risk (given objective probabilities), subjects face uncertainty. We show multiple biases in the estimation of probabilities.

Finally, in the section titled "Facing Trade-Offs," where we focus on multi-attribute consequences, x will be a vector encoding a potentially rich set of attributes of the consequence. This section also illustrates the different approaches to the behavioral study of decision making: rather than a behavioral utility model of *preferences*, presented in the first part, the approach is to model *process* that people follow to make complex decisions. While the approach in the sections "Decisions under Risk and over Time: The Rational Model" and "Decisions under Risk and over Time: The Behavioral Model" considers the trade-offs subjects might be making, the approach in the section titled "Facing Trade-Offs" considers procedures that subjects might follow precisely to avoid making trade-offs!

DECISIONS UNDER RISK AND OVER TIME: THE RATIONAL MODEL

How would a rational decision maker decide in simple problems involving risk and time? To begin with, such a decision maker would hold a preference over pairs of prospects that is complete, transitive, and (arguably) continuous. This condition holds if and only if there is a continuous function $V(x, p, t)$ that scores prospects according to preference. Additional conditions on preferences

will then impose constraints on the shape of V . One such constraint is monotonicity, which makes V increasing in x and p , and decreasing in t (impatience). What other constraints can we impose on V in the name of rationality?

V is Linear in Probability

Assume we start flipping a fair coin. Contemplate the prospect of gaining \$1000 if the coin lands on heads the next *five* times, and zero otherwise. How would you compare this offer with the prospect of gaining \$500 if the coin lands on heads the next *four* times, and zero otherwise? In other words, Choice #1 is as follows:

	Prospect A	Prospect B
#1	(\$1000, $1/2^5$, now)	(\$500, $1/2^4$, now)

The expected value of both outcomes is the same. Let us suppose our preference for risk leads us to choose A as we see the rewards of \$1000 more appealing. So far, there is nothing wrong with our preference. As we start flipping the coin, the original prospects in #1 change. For instance, if the coin lands tails the first time, then the residual prospects A_T and B_T are both zero. But, what if the coin lands heads the next four times? Then, the residual prospects are

	Prospect A_{4H}	Prospect B_{4H}
#2	(\$1000, $1/2$, now)	(\$500, for sure, now)

Given a choice between \$500 for sure and a 50% chance of \$1000, we may prefer B_{4H} , the sure thing. Somehow, as the coin lands on heads, we have shifted our preference toward the \$500 reward (B_{4H} in #2), eventually reversing our original preference for the \$1000 reward (A in #1). Is this rational? Arguably, this preference reversal is not rational because it is clear from the beginning that either reward is available only if

the first four flips land on heads. Why should we change our mind if we observe four flips that land on heads? This reversal implies that, from a distance, we pay more attention to outcome size. When the possibilities are closer, then we become more sensitive to probabilities. This cannot be a good way to make a decision, because it shows the lack of clarity in the objective one is trying to reach. It is better to choose either A or B and stay with the choice.

Ruling out this type of reversals is the essence of the independence axiom, proposed by Von-Neumann and Morgenstern [6] and broadly discussed in the literature. In a general setup, this reversal will not occur if and only if the function V is linear in the probability p ,¹ that is, $V(x, p, t) = pg(x, t)$, for some function $g(\cdot)$.

V is Exponential in Time

Consider Choice #3, involving a sure gain of \$1000 in 5 years from now, or a sure gain of \$500 in four years from now.

	Prospect A_{4H}	Prospect B_{4H}
#3	(\$1000, for sure, 5 years)	(\$500, for sure, 4 years)

We are patient, and prefer A. Assume the rewards are fixed to some calendar date. This means that, as time passes, the prospects in

¹The “only if” statement holds in the case of multiple-outcome lotteries, or in the case of simple lotteries if we add a *double cancellation* condition [7] that makes V a multiplicative separable function of p and x , that is, $V(x, p) = w(p)v(x)$. The no reversal condition implies $v(y)/v(x) = w(p)/w(q) = w(\lambda p)/w(\lambda q)$. Fixing q and letting $w(\lambda q)/w(q)$ be a function $g(\lambda)$ of λ , we have $w(\lambda p) = g(\lambda)w(p)$. The unique solution to this Pexider equation is $w(p) = bp^c$ [8]. But in the same way that the multiple-outcome utility is unique up to affine transformation, in the simple lottery case $\ln(V)$ is unique up to affine transformation, which means that we can normalize so that $w(p) = p$.

#3 change and get closer. In fact, 4 years later, the residual prospects are

	Prospect A_{4Y}	Prospect B_{4Y}
#4	(\$1000, for sure, 1 year)	(\$500, for sure, now)

Suppose we now become impatient, and prefer B_{4Y} .² Is the choice pattern A in #3 and B_{4Y} in #4 rational? Arguably not, because it is clear from the beginning that time will pass, and why should we change our mind as time passes? Preference reversals due to lack of stationarity are expected to be frequent, as the passage of time affects all our decisions. The condition that rules out these preference reversals is called the *stationarity axiom*. In a general setup, this type of reversal is ruled out if and only if the function V is exponential in the delay t . To recapitulate, the rational model will look as follows:

$$V(x, p, t) = pe^{-rt}v(x) = \frac{v(x)}{e^{\ln(1/p)+rt}} \quad (1)$$

where $v(x)$ is any increasing function. Note the unorthodox presentation; we write $p = 1/e^z$, where $z = \ln(1/p)$ is a measure of *risk distance*. Suppose the coin we used in Choice #1 had probability $1/e$ of landing heads. A reward that depends on the coin landing heads in the first z flips is at a risk distance of z . Zero distance corresponds to $p = 1$; distance one corresponds to $p = 1/e = 37\%$; distance two to $p = 1/e^2 = 14\%$; distance 3 to $p = 1/e^3 = 5\%$; and infinite distance to $p = 0$.

The parameter r is known as the *discount rate*. The discount rate reflects time preference and impatience: to which extent a larger later reward shall be preferred to a smaller sooner reward. However, in Equation (1) the parameter r has a second, very different interpretation, namely, *the trade-off rate*

²In experimental studies, we rarely make subjects wait for four years and ask them the question. We present Choice #4 now, and assume that the preference for the immediate reward would hold both for now and four years.

between risk distance and time distance. For instance, delaying the reception of a reward by one period is equivalent to a probability reduction of e^r . Or, multiplying the probability of receiving the reward by a factor of λ is equivalent to delaying its reception by $\ln(1^\lambda)/r$ time units. When interpreted this second way, r will be called the *probabilistic discount rate*.

No Reference Dependence

Consider the function v in Equation (1). Does rationality imposes further restrictions on v ? Recall Choice #2 between (\$1000, $1/2$, now) and (\$500, for sure, now), and our preference for the sure \$500. Suppose that we want to meet the target of making a profit of \$1000. This could be because we want to recover a recent loss of \$1000, or match the earnings of some friend to avoid envy. Relative to our target, a gain of \$1000 is neutral, and receiving zero becomes a loss of \$1000. Choice #2 is then perceived as

	Prospect $A_{4H-\$1000}$	Prospect $B_{4H-\$1000}$
#5	(-\$1000, $1/2$, now)	(-\$500, for sure, now)

Suppose we choose $A_{4H-\$1000}$. Is the pattern B_{4H} in #2 and $A_{4H-\$1000}$ in #5 rational? It is clearly irrational, because the alternatives in #2 and #5 are actually the same! All that changes is the description and perception of the same consequences. The change of perception is the consequence of comparing the same consequence to a different reference point. Changing the reference point is equivalent to adding or subtracting a common quantity from all the consequences. This change might be quite common, for example, by including past or future committed costs, or by adding existing balances of money or asset values, or by expressing the consequence relative to some target. Imposing that preferences should not change based on these reference changes leads to two possibilities:

1. The classical proposal is to set $v(x)$ equal to $u(x + a) - u(a)$, where u is a

stable utility function defined over final wealth positions and a is the wealth prior to the decision. This step will undo the effect of changes in references because the correct calculation of a provides an absolute reference. This yields the expected utility model, the benchmark of rationality:

$$V(x, p, t) = \frac{u(x + a) - u(a)}{e^{\ln(1/p) + rt}}. \quad (2)$$

2. In prescriptive applications, neither u nor a are known, and it is often difficult or artificial to elicit them. Therefore, a more practical possibility is to adopt a functional form for v that is invariant to common additive changes in the payoffs. As it is well known, only the linear form, $v(x) = x$, or the exponential form, $v(x) = 1 - e^{-x/\rho}$, will satisfy this invariance condition.³ The linear and the exponential model also satisfies other desirable properties such as the one switch property of Bell [9], and are mathematically very tractable for applications of expected utility.⁴ Taking this second possibility, the rational model becomes

$$V(x, p, t) = \frac{1 - e^{-x/\rho}}{e^{\ln(1/p) + rt}}. \quad (3)$$

Adopting the second approach, a rational decision maker is characterized by two preference parameters, ρ and r . The additional constraint of risk aversion implies that ρ is strictly positive. The limiting case of ρ equals to infinity correspond to risk neutral preferences, $v(x) = x$.

³If the consequences are evaluated multiplicatively with respect to some reference, say individual income over average income, then the only utility function that is invariant to multiplicative changes in references is the power form, $v(x) = x^\alpha$.

⁴Exponential utility combined with a normal probability distributions yields a mean-variance framework for risk analysis, important in finance. If utility is exponential, then the problem of timing of resolution of uncertainty has an explicit recursive solution [10].

The extension of the rational model to multiple outcomes leads to expected utility: V is the sum-product of utilities and probabilities. In the case of receipts of cash flows over time, the rational approach is to first take the net present values of these cash flows, and then calculate the expected utility of the net present value [11]. In the case of receipts consumption flows, one possible approach is to first take the utility of this consumption, and then discount the utility [12]. However, it might be appropriate to modify the discounted utility model to account for habit formation [3] or satiation [13].

DECISIONS UNDER RISK AND OVER TIME: THE BEHAVIORAL MODEL

Are subjects rational? For a long time it was assumed that, for relatively simple problems, most subjects were to decide rationally. But years of laboratory experiments and field data demonstrate that most subjects exhibit the type of reversals that we just classified as irrational [14–16]. But if Equation (2) or (3) does not describe subjects' preferences, what is the form that V takes?⁵ We answer this question using the mainstream behavioral

⁵One may even question if subjects preferences can be described by a function V . Lichtenstein and Slovic [17] show that, varying the elicitation method, subjects violate the most basic tenant of rationality: if A is strictly preferred to B , then B cannot be strictly preferred to A . Here is the example: most subjects prefer the prospect (\$6, 80%, now) to (\$50, 10%, now). This is their answer in a *choice* question. But when asked for the certainty equivalents (the sure amount that would make them indifferent to the prospect), they report that the certainty equivalent for (\$50, 10%, now) is higher than the certainty equivalent for (\$6, 80%, now). This is their answer in a *matching* question. Given this difference, most researchers in behavioral decision making adopt choice questions to elicit preferences. Matching is replaced by the bisection method, which proposes a sequence of choice questions that ultimately make the two prospects close to indifferent. All evidence we present in this article is obtained using choice questions.

models in decisions under risk and over time [18,19].

Sensitivity to Risk and Time Distance

In the rational model subjects are supposed to possess constant sensitivity to risk and time distance. That is, adding an additional

flip of the coin in Choice #1 or a common delay in Choice #3 should not alter the relative desirability of the prospects. Do subjects exhibit constant sensitivity to risk and time distance? Consider the following experimental evidence:

	Prospect A	Prospect B	Response	N
#6	(€ 9, for sure, now)	(€ 12, with 80%, now)	58% vs 42%	142
#7	(€ 9, with 10%, now)	(€ 12, with 8%, now)	22% vs 78%	65
#8	(100 fl, for sure, now)	(110 fl, for sure, 4 weeks)	82% vs 18%	60
#9	(100 fl, for sure, 26 weeks)	(110 fl, for sure, 30 weeks)	37% vs 63%	60

Let us compare Choices #6 and #7. Choice #6 captures the trade-off between selecting a larger outcome but with lower probability. Choice #7 is constructed from Choice #6, after multiplying all probabilities by a common factor. If preferences were linear in probabilities, then the multiplication of probabilities by a common factor should not change. However, most subjects switch, a reversal known as the *common ratio effect*. Let us compare Choices #8 and #9. Choice #8 shows the trade-off of receiving a larger outcome but later in time. Choice #9 is constructed from Choice #8, after adding a common delay to both prospects. Most subjects switch and want the later larger outcome, a reversal known as the *common difference effect*.

The common ratio and the common difference effects reveal that subjects exhibit

diminishing sensitivity to risk and time distance, respectively. That is, they notice the difference between $z = 0$ (100%) and $z = 0.2$ (80%) in Choice #6 more than the difference between $z = 2.3$ (10%) and $z = 2.3 + 0.2$ (8%) in Choice #7. And they feel the difference between now and one month (#8) more than the difference between 26 and 30 weeks (#9).

An obvious modification of the rational model is to introduce a concave function, $d(z)$, which we could call the *psychological distance function*, or *pdf*. The *pdf* captures the diminishing sensitivity toward risk and time distance. Is the psychological distance function for risk the same as the one for time? To answer this, compare Choices #6 with #10 and #8 with #11.

	Prospect A	Prospect B	Response	N
#6	(€ 9, for sure, now)	(€ 12, with 80%, now)	58% vs 42%	142
#10	(€ 9, for sure, 3 months)	(€ 12, with 80%, 3 months)	43% vs 57%	221
#8	(100 fl, for sure, now)	(110 fl, for sure, 4 weeks)	82% vs 18%	60
#11	(100 fl, with 50%, now)	(110 fl, with 50%, 4 weeks)	39% vs 61%	100

Choice #10 is constructed from #6, after adding a common delay. What we observe is that this common delay produces the same reversal toward the larger outcome, “as if”

time were probability. Similarly, Choice #11 is constructed from #8, but this time after multiplying probabilities by a common factor. This common factor produces the same

reversal toward the larger outcome, “as if” probability were time.⁶

Following Ref. 20, we propose that risk distance can be exchanged with time distance at a rate r . We call r the *probabilistic discount rate*, because this is how fast time makes the outcome disappear as if it were a probability. Formally, we propose the following behavioral model:

$$V(x, p, t) = \frac{v(x)}{e^{d(\ln(1/p) + rt)}}, \quad (4)$$

where d is a concave function. The more concave the d is, the more we depart from

the rational benchmark. The concavity of d is intuitive. In the same way that subjects exhibit diminishing sensitivity to outcomes (a windfall of \$10,000 feels better than an additional \$10,000 added to a windfall of \$100,000), subjects also experience diminishing sensitivity to risk and time distance. A delay of hour feels longer than a delay of one hour added to a four hour delay; and moving from certainty to a probability of 50% feels stronger than a reduction of the probability by 50% when the base probability is 10%, therefore leading to a 5% probability.

Is the probabilistic discount rate r constant? Consider the following two choices.

	Prospect A	Prospect B	Response	N
#12	(€ 5, with 90%, now)	(€ 5, for sure, 1 month)	57% vs 43%	79
#13	(€ 100, with 90%, now)	(€ 100, for sure, 1 month)	19% vs. 81%	79

Choice #12 shows the trade-off of receiving an outcome with higher probability but later. In this case, the modal preference is to choose the sooner but less certain prospect. Choice #13 is constructed from Choice #12, but with a larger outcome. For larger outcomes, most subjects switch and want the later but more certain prospect. We call this preference reversal the *subendurance*.

The subendurance implies that the trade-off between delay and probability is outcome dependent. People are more enduring as the outcome increases. To account for the subendurance, one needs to make the probabilistic discount rate dependent on the magnitude of the outcome, or

$$V(x, p, t) = \frac{v(x)}{e^{d(\ln(1/p) + r(x)t)}}, \quad (5)$$

where $r(x)$ is a decreasing function of the outcomes. Hence, for small outcomes, time runs faster than for large outcomes.⁷ In summary, the core behavioral findings of the risk and time domain can be encoded in a *value function*, $v(x)$, a psychological distance function, $d(\ln(1/p) + r(x)t)$, and a probabilistic discount rate function, $r(x)$, that regulates the trade-off between risk and time.

Historically, diminishing sensitivity to risk distance has been incorporated by means

⁶Interestingly, reversals #6–#10 and #8–#11 rule any of the following forms: $V(x, p, t) = f(t)w(p)v(x)$, or $V(x, p, t) = w(p)v(x, t)$, or $V(x, p, t) = f(t)v(x, p)$. To see this, note that in #1 the time is the same in both prospects. Hence, in comparing both evaluations the term $f(t)$ would drop, and the preference be driven by the values of p and x . But in #10 time also drops, and the tasks have the same values of p and x , hence no preference reversal should be expected. Similarly, the term $w(p)$ would disappear from the utility comparison in both #8 and #11.

⁷Preliminary experimental evidence suggests that if time is measured in months, then $r(x)$ could be written as R/x , with $R = \$1.5$ [21]. Thus, for an outcome of \$15, time fades at a monthly rate of 10%, for \$150 at a monthly rate of 1%, and for \$1500 at a monthly rate of 0.1%. Demographical variables influence the probabilistic discount rate: it decreases with age at a rate of 5% per year, and it is 35% higher for females; the passage of one time unit carries more “uncertainty” for the young and for females [22]. The probabilistic discount rate influences the discount rates (the marginal rate of substitution between x and t in V). Therefore, if r decreases with age, we expect a higher willingness to wait for a larger outcome as we grow older.

of a nonlinear probability weighting function (*pwf*), $w(p)$, by Kahneman and Tversky [15], and diminishing sensitivity to time distance by means of a time discounting function (*tdf*), $f(t)$, exhibiting decreasing impatience. The translation between these classical behavioral models and our framework is given by the following equality:

$$\begin{aligned} V(x, p, t) &= e^{-d(\ln(1/p) + rt)} v(x) = w(pe^{-rt}) v(x) \\ &= f(t + \ln(1/p)^{1/r}) v(x). \end{aligned} \quad (6)$$

A concave *pdf*, a subproportional *pwf*, or a subadditive *tdf* are all equivalent. If in addition the *pdf* crosses the identity line from above, then the *pwf* is inverse-S-shaped and crosses the identity line from above. Of course, a *pdf* equal to the identity, a *pwf* equal to the identity, and an exponential *tdf* (constant impatience) are all equivalent.

Sensitivity to Gains and Losses

What is the shape of the value function one encounters in experiments? First, subjects do not have a stable utility function u over absolute consequences that determine $v(x)$. Rather, what is stable is a reference-dependent value function. That is, a given gain is perceived the same regardless of the base level of wealth. Decision makers would evaluate the prospect of a gain of \$10,000 when their income is \$90,000 the same as a gain of \$10,000 when their income is \$190,000. In both cases, we assume that the reference point has adjusted to the habitual income level. Hence, the absolute income does not significantly affect the preference over the prospect of a gain of \$10,000, because the habitual income level will be factored away from consideration.⁸

Reference dependence would not matter if v were approximately exponential, as the exponential shape makes subject's preferences invariant to reference effects. An exponential value function would imply diminishing sensitivity to gains, and

increasing sensitivity to losses. But this is far from what is usually found. Instead, subjects exhibit diminishing sensitivity to both gains and losses. In other words, v is S-shaped. Moreover, subjects have more sensitivity to losses than to gains of equal magnitude, that is, v is steeper for losses than for equivalent gains [15] (Figure 1).⁹

Empirical Estimation of the Descriptive Model

Various authors have attempted to empirically determine the prevailing shape for v and w (or d). The frequent form that has been adopted is $v(x) = x^\alpha 1_{x \geq 0} - \lambda(-x)^\alpha 1_{x < 0}$ and $d(z) = \delta z^\gamma$.¹⁰ This power specification for the *pwf* yields the Prelec form [27], $w(p) = e^{-\delta(\ln 1/p)^\gamma}$, and the Ebert and Prelec form [28] for the time discounting function, $f(t) = e^{-\delta(rt)^\gamma}$. The model allows for sign-dependent sensitivities to risk and time distance. Hence, we use δ^+ and δ^- , depending on the sign of the outcome (the exponents α and γ for gains and losses are found to be similar, and we make them the same.) If rational decision makers were characterized by the pair (ρ, r) , their irrational counterparts could be characterized by five parameters $(\alpha, \lambda, \gamma, \delta^+, \delta^-)$. Plausible values for these parameters based on literature averages are [29]

$$\begin{aligned} (\alpha = 0.85, \lambda = 2.25, \gamma = 0.65, \delta^+ = 1, \\ \delta^- = 0.87). \end{aligned}$$

In this framework, individual or demographic differences in risk attitudes boil

⁸Strong reference dependence, together with a reference that adapts to past income or that of the others, explains the happiness treadmill [23].

⁹The combination of reference dependence and loss aversion explains why, out of the following two equivalent proposals, the second has a stronger impact [2, p. 37]: (i) if you use energy conservation methods, you will save \$350 per year; or (ii) if you do not use energy conservation methods, you will lose \$350 per year.

¹⁰A more satisfactory behavioral specification could be the expo-power value function $v(x) = (1 - \exp(-\beta x^\alpha))/\beta$ as it contains both the power and exponential as particular cases, and it might be empirically more accurate [21, 24]. Regarding the *pwf*, popular specifications such as that of Tversky and Kahneman [25] or Goldstein and Einhorn [26] forms produce unnatural *pdfs*. Hence, we stay with Prelec's [27] form.

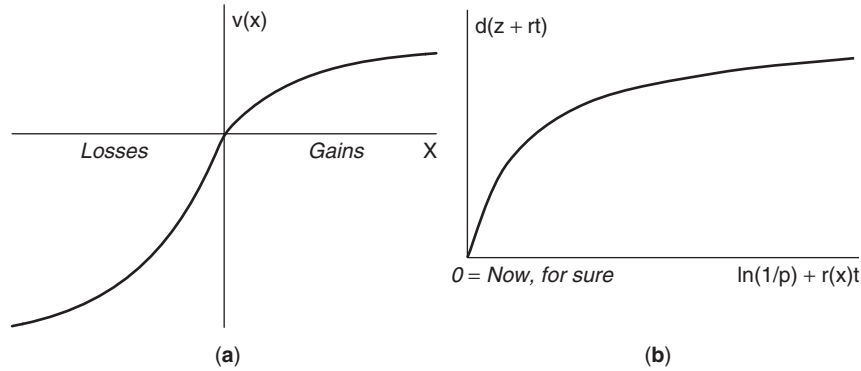


Figure 1. The behavioral model is characterized by three functions. The *value function* (a) describes diminishing sensitivity to gains and losses, including loss aversion. The *physiological distance function* (b) describes diminishing sensitivity to risk and time distance. The *probabilistic discount rate* (not drawn) describes the exchange rate between time and probability (which may depend on the magnitude of the outcome and demographical variables).

down to making these parameters dependent on say age, gender, and education. Interestingly, income has little influence on the parameters, confirming the validity of the reference-dependent model. Females often exhibit more risk aversion than males. In a rational framework, this implies a more concave utility function. But in a behavioral framework, risk attitudes are a complex composite of curvature of the value function and of the *pwf*. Booij *et al.* [29] confirm previous findings that gender differences in risk attitudes are not found on the curvature parameters, but on the parameter of loss aversion (λ is higher on average by 0.25 in females).

Older people seem to value money more linearly, with a 50 year age difference being associated with a power that is 0.15 higher. This effect works to reduce risk aversion, but it is countered by more nonlinear weighting of probabilities (γ in older people lower by 0.3).¹¹ Education, defined as having a higher vocational or academic education, does not affect utility curvature, nor is it associated with a more linear weighting of probabilities. However, education is associated with a

lower degree of loss aversion (λ in the highly educated is lower by 0.32), which suggest that the reduction in risk aversion with years of schooling that is often observed [32] stems mainly from lower λ .

Emotions, for instance, are naturally incorporated in this model. Hsee and Rottenstreich [33] show that in high arousal emotional states the curvature of both v and d increase (α and γ decrease with arousal). Thus, stressed individual are more risk averse for gains, and more risk seeking for losses of moderate probability [34]. An obviously exciting research area is to connect these functional specifications with findings from neuroimaging experiments [35].

ESTIMATING SUBJECTIVE PROBABILITY

Going back to the work of Kahneman, Tversky and their colleagues in the 1970s, there have been many empirical demonstrations of peoples' difficulties in estimating the quantities needed for a rational decision. In the previous section we have focused on the behavioral replacement of the utility function, and assuming that the delay and the probability were objectively provided. Here, our concern is the estimation of the probability when this parameter is to be estimated subjectively.

There are numerous biases that question the accuracy in the estimation of

¹¹The total effect of these estimates depends on the prospects under study. For prospects that entail a small probability of a large gain one may find risk aversion to decrease with age, while in those that do not, increasing risk aversion is more likely, which is what is usually found [30–32].

probabilities. Most of these biases affect the process of forming quantity judgments in general, and some are specific to probabilities. Subjects use a number of heuristics such as *similarity/representativeness*, *availability*, or *partition dependence* to come up with probability estimates. These heuristics may perform relatively well in natural environments, but their simplicity necessarily produces systematic mistakes. One example is the representative heuristic, which judges the probability of an element to belong to set by its similarity to a representative sample. For instance, subjects are given the description of a woman called Linda. Given the description, they think it is more likely that Linda is “a bank teller involved in feminist movement” rather than “a bank teller.” But the second set is larger than the first, so this likelihood ordering is absurd. The availability heuristic (judging probabilities by the facility with which our mind produces examples) leads us to give higher probability to events that are easier to imagine, or appear more often in the news [36]. Finally, when subjects are asked to assign probabilities to an event that has been partitioned in n cases, they tend to give probabilities of $1/n$ to each partition [37].

It appears to be difficult for people to apply Bayes theorem and correctly modify the probabilities based on a new piece of evidence. More specifically, laypeople and also professionals such as medical doctors and judges seem to be putting too little weight on the prior probability (this phenomenon has been called *base rate neglect*; [38]). This explains why, for example, doctors and patients often believe that the probability of suffering from a disease such as an HIV infection is almost one if one tests positive. While the sensitivity and specificity of medical tests is often close to one, the *a priori* probability (the base rate) of suffering from a disease, such as HIV infection, is very low. Thus the accurate, when taking into account base rates, value of the probability is medium. Note, however, that base rate neglect can be alleviated by the use of formats that represent probabilistic information so as to facilitate computation, such as “natural frequencies” [39].

Probability judgments can be made more accurate by combining the estimates of different sources. Subjects often fail to appreciate the power of this averaging principle, and rather fall in the trap of *chasing the expert* [40].

Another empirical phenomenon that refers to peoples’ probabilistic thinking is *overconfidence*. Various phenomena have been labeled overconfidence and there are many different definitions [41]. Sometimes overconfidence is measured by the miss-calibration of subjective probabilities, which equals the discrepancy between subjective probabilities and objective probabilities. For example, if one believes that the probability of a fair coin landing heads equals $1/3$, then their miss-calibration would equal $1/2 - 1/3 = 1/6$. Other measures include overestimation. That is, if a student who took a 10-item quiz believes that he answered five of the questions correctly when, in fact, he got only three correct, then he has overestimated his score by two items.

It was originally thought that overconfidence is very widely spread. There is no doubt that people are overconfident under some conditions. But during the last 20 years or so, researchers have questioned the empirical adequacy of the claim that most people are overconfident most of the time. Problems have been identified in the way researchers select items, how data are analyzed, and other problems including overconfidence as a reflection of regression-to-the-mean-like artifacts [41]. There are, however, no comprehensive reviews available of the evidence for overconfidence across the different phenomena labeled as overconfidence.

We can finally ask what is the normative status of these biases. For instance, is overconfidence always bad. Or is it? If people are overconfident, it might be that they are overconfident for a reason. In other words, overconfidence may confer adaptive advantages by, for example, increasing resolve, increasing the ability to bluff opponents, and winning net payoffs from risky opportunities despite occasional failures. Again, systematic mistakes reflect the use of simple processes to handle complex problems. If we were constrained to use simple processes, then the

heuristics most people use might be efficient, as they provide quick and relatively accurate answers most of the time.¹²

FACING TRADE-OFFS

After it was empirically demonstrated [43] and acknowledged that human behavior does not always conform to the ideal of unbounded rationality, different theorists took different approaches to building a behaviorally sound decision theory. One approach, that is highlighted in the first part of this entry, is to see how the expected utility benchmark can be modified so that it accounts for the major empirical phenomena. One may note that this approach differs from that of Herbert Simon. The construct of an expected utility function does not feature prominently in Simon's work. An alternative approach to a utility-based theory of rationality is the study of "how actual people make decisions without calculating utilities and probabilities" [44, p. 8]. Rather than forcing subjects into a

choice paradigm constructed from artificial comparisons between pairs of prospects, the focus is to the study the actual process subjects follow to reach a decision. This latter approach is the one we adopt in this section, as applied to complex decisions among multiattributed alternatives.

You are a member of the search committee for a postdoctoral research fellowship. The secretary hands to you the packet with the applicants' cover letters, vitas, reprints, and recommendations. You discover that there are more than a dozen applicants and there is simply not enough time to read all material before the committee meets tomorrow. And, yet, you must have good reasons for the short list you propose. What do you do?

The Rational Model: Multiattribute Utility Theory

Let us discuss first what it is that supposedly you should do. The standard theory of multiattribute choice is Keeney and Raiffa utility theory [45]: Assume that the value x_{ij} of option X_i on attribute j can be mapped onto a utility $u_j(x_{ij})$. For example, a candidate may have zero refereed publications, and thus also zero utility from publications. Under some technical assumptions, for example, of preferential independence it can be proven that there exist scalar weights w_j so that the total utility of option X_i equals $\sum_j w_j u_j(x_{ij})$. The weights can be interpreted as quantifying how utility due to one attribute can be traded off against utility due to other attributes. According to multiattribute utility theory, the decision maker must choose an option with maximum total utility. The rational model can justify a number of specifications, such as the multiplicative form if there are interactions between the attributes.

The field of prescriptive decision analysis is concerned with finding ways of restructuring the decision problem so that people can provide better estimates, and, more generally, correctly apply utility theory in situations with multiple attributes, under both certainty and uncertainty [46,47].

¹²Evolutionary arguments are often invoked to support that rationality is supposed to be the rule. If evolution has endowed us with such perfect organs for perception and movement, why should our cognitive abilities be systematically biased. In a recent paper, Livnata and Pippenger [42] give an evolutionary argument that support "systematic mistakes" as more adaptive than perfection. The key argument hinges on assuming a bound on the processing capacity of the brain. They model the brain as an informational processing network with fixed number of components. If the capacity is fixed, then an informational processing network that handles a larger set of inputs with systematic mistakes is superior to an informational processing network that handles a smaller set of inputs with no mistakes. This result suggests that departures from rationality in humans are explainable by natural selection, as evolution operates under constraints. If the information processing capacity is upper bounded, there is a trade off between the exactitude of the response and the size of the environment the processor can handle. A processor that gives approximately correct answers to a broad domain of problems is superior to a processor that gives exact answers to a narrow domain of problems.

The Behavioral Model: Noncompensatory Heuristics

This is a clear theory but it is not always so clear if people can apply it in practice. To apply it, subjects must measure x_{ij} , estimate their utility function $u_j(x_{ij})$, and weight w_j for each attribute j . For instance, the estimates of attribute weights turn out to be influenced by variables such as the range of the attribute values that are theoretically irrelevant, or the base attribute chosen to elicit the trade-offs with the rest of attributes. And, professionals, such as urban fire-ground commanders, often cannot or refuse to provide utilities (for more examples, see Katsikopoulos and Fasolo [48]). A field called *naturalistic decision making* is especially critical of the applicability of utility theory to real-world decision problems [49].

Let us say that you will not apply multi-attribute utility theory. Here is something you could do instead: You first eliminate all candidates who have no refereed journal publications. From the pool of remaining candidates, you further eliminate all of those who have at least one recommendation that is lukewarm or negative.

Tversky [50] proposed elimination by aspects (EBA) as a model of how people choose among multiattribute options. In EBA, the attributes are ordered and considered one at a time. In the example above, the first attribute is the number of publications and the second attribute is the number of nonpositive recommendations. When one attribute is considered, every option that has a value on the attribute that is inferior to a predetermined aspiration level is eliminated (unless all the alternatives are unacceptable on that attribute). In our example, the aspiration level on the number of publications is one and the aspiration level on the number of nonpositive recommendations is zero. The heuristic terminates when there is just one alternative left, or all attributes have been considered. In the latter case, we can choose randomly from the set of noneliminated options.

EBA implements a simple way of facing trade-offs among attributes: by, in fact, not making them. In the example above, excellent recommendations or a great cover letter

do not compensate for lack of publications. Models of multiattribute choice under certainty that do not make trade-offs among attributes are called *noncompensatory*, and often involve a *lexicographic* of elimination of alternatives by sorted criteria.

There are a number of ways by which people could be ordering attributes. A simple assumption is that attributes are ordered randomly (this is the case for EBA). It is, however, more psychologically plausible that the order of attributes depends on the population from which pairs of options are sampled. In the take-the-best (TTB) heuristic [51], attributes are ordered by their validity. Given a pair of randomly sampled options X_1 and X_2 such that X_1 has the higher “true utility” for a decision maker, the validity of attribute j is the conditional probability $Pr[x_{1j} = 1, x_{2j} = 0 | x_{1j} \neq x_{2j}]$. It turns out that the validity of an attribute is equal to a positive linear transformation of a correlation measure between the attribute and the true validity. This correlation measure is naïve in the sense that it ignores the interdependencies between attributes.

EBA, and other noncompensatory models, capture the intuition that people often make decisions by using a limited amount of resources, such as time, information, and computation. For example, in EBA, the decision maker does not look up all available information because s/he does not use all attribute values of all options. In our example, the recommendations attribute is not used for some of the candidates (those with zero publications). The cover letter attribute is not used for any of the candidates. In sum, a shortlist of candidates is chosen quickly, based on a few pieces of information. Note also that EBA is not computationally demanding as no sophisticated operations are applied to the pieces of information that are actually used. Unlike utility theory, attribute values are neither weighted nor aggregated. They are compared to a fixed threshold.

Simon [52,53] is widely credited as being the first who forcefully claimed that people use limited time, information, and computation when making decisions. The resources available for making decisions can be limited because of the way the decision problem is

structured in the real world. For example, the time you have for reviewing applications may be short because you have many other things to do as well. Often the computations needed for implementing a compensatory model that makes trade-offs are beyond the capacity of humans or machines, as in deciding your next move in a game of chess. Finally, the decision maker may prefer a simpler decision process in order to have a better understanding of how a decision was reached and be able to justify it to others.

In psychology, Tversky, Kahneman, and their colleagues [43,54,55] provided extensive evidence for Simon's claim. They popularized the term *heuristics* for simple rules of thumb that people use for making decisions. In a review of 45 studies on tracing human decision processes, Ford *et al.* [56] concluded that "noncompensatory strategies were the dominant mode used." Gigerenzer and his colleagues [44,51,57] documented the use of noncompensatory heuristics in a variety of fields such as biology, engineering, and law.

In sum, there are good reasons for believing that people do make decisions by using simple noncompensatory heuristics. The analytical study of these heuristics has been facilitated by the use of mathematical models.

Despite their naiveté lexicographic heuristics such as EBA and TTB can, under some conditions, lead to decisions that increase total utility in comparison to benchmarks such as linear regression. Payne *et al.* [58] and Gigerenzer and Todd [51] demonstrated this by computer simulations. A difference between the two studies was that in the former the true utility of options was assumed to equal the value of a linear combination of attribute values, while in the latter the true utility was exogenously given. Further analyses have uncovered mathematical reasons for the success of lexicographic heuristics [59], provided the attributes are sorted correctly before applying sequential elimination of alternatives. For example, the heuristics are optimal if and only if attribute validities are very far apart [60], or there exists a cumulatively dominating option. Cumulative dominance provides a structural justification for the reasonable performance of a broad class of lexicographic heuristics [61].

CONCLUSIONS

The extension of the behavioral model to domains with multiple risky outcomes and/or multiple time periods is still an open problem. For multiple-outcome lotteries that resolve now, cumulative prospect theory is the prevailing extension [25]. However, empirical evidence is mixed [62,63] and replacements such as the TAX model proposed by Birnbaum [64] may outperform cumulative prospect theory.

In the time domain, the extension to multiple outcomes received at different time moments is still an open problem. With multiple outcomes received over time, one needs to account for possible changes in reference points. It has been shown that subjects miscalculate these changes, thinking that future reference points will be closer to current reference points. In reality, reference points may change more than anticipated (*adaptation bias*). Adaptation bias is a particular instance of the more general *projection bias* [65]. In complex decisions, the behavioral model also needs to specify the type of simplifications subjects may perform to simplify the problem; one such simplification being *narrow choice bracketing* [66]. Projection bias and narrow bracketing introduce the distinction between predicted/decision utility and realized/experienced utility.

Diminishing sensitivity to risk distance accommodates anomalies of the classical model such as the Allais paradox [14], the coexistence of gambling and insurance, betting on long-shots at horse races [67], and the avoidance of probabilistic insurance [68]. In the time domain, evidence for diminishing sensitivity to time distance is abundant [69–71], and models incorporating this effect are now part of economic analysis [72], and are known under the name of *diminishing impatience* of hyperbolic discounting models. Persistent phenomena such as repeated procrastination are explained by decreasing impatience [73]. Loss aversion explains the endowment effect, and excessive risk aversion in risks involving potential gains and losses [74]. For risk with only losses, diminishing sensitivity yields to chains of bad decisions known as *escalation of commitment*.

The subendurance explains the attractiveness of consumption loans: the payment is made smaller through monthly installments and delayed in time. The model predicts that the attractiveness further increases if the payments are made smaller by increasing the frequency of payments (say weekly instead of monthly). The subendurance indirectly explains the observation that discount rates depend on the magnitude of the outcomes, known as the *magnitude effect* [16].

The approach of noncompensatory heuristics has also been recently used to describe how people handle trade-offs under uncertainty. For choices between gambles, it has been shown analytically that the so-called priority heuristic implies major violations of expected utility theory such as the common-consequence effect, the common ratio effects, the reflection effect, and the fourfold pattern of risk attitudes [75].

REFERENCES

1. Ariely D. Predictably irrational: the hidden forces that shape our decisions. New York: Harper Collins; 2008.
2. Thaler RH, Sunstein CR. Nudge: improving decisions about health, wealth, and happiness. New York: Penguin; 2008.
3. Wathieu L. Habits and the anomalies in intertemporal choice. *Manage Sci* 1997;43(11): 1552–1563.
4. Loewenstein G, Prelec D. Preferences for sequences of outcomes. *Psychol Rev* 1993;100(1):91–108.
5. Kahneman D. Experienced utility and objective happiness: a moment-based approach. In: Kahneman D, Tversky A, editors. *Choices, values, and frames*. Cambridge: Cambridge University Press and the Russell Sage Foundation; 2000. pp. 673–692.
6. Von-Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton: Princeton University Press; 1944.
7. Krantz D, Luce D, Suppes P, *et al.* Volume I, *Foundations of measurement*. New York: Academic Press; 1971.
8. Aczél J. *Lectures on functional equations and their applications*. New York: Academic Press; 1966.
9. Bell DE. One-switch utility functions and a measure of risk. *Manage Sci* 1988;34(12): 1416–1424.
10. Smith J. Evaluating income streams: a decision analysis approach. *Manage Sci* 1998;44(12):1690–1708.
11. Baucells M, Sarin R. Evaluating time streams of income: discounting what? *Theory Decis* 2007;63(2):95–120.
12. Koopmans TC. Stationary ordinal utility and impatience. *Econometrica* 1960;28(2): 287–309.
13. Baucells M, Sarin R. Satiation in discounted utility. *Oper Res* 2007;55(1):170–181.
14. Allais M. Le comportement de l'homme rationnel devant le risque: critique le postulats et axiomes de l'école américaine. *Econometrica* 1953;21(4):503–546.
15. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–292.
16. Frederick S, Loewenstein G, O'Donoghue T. Time discounting and time preference: a critical review. *J Econ Lit* 2002;40(2):351.
17. Lichtenstein S, Slovic P. Reversals of preference between bids and choices in gambling decisions. *J Exp Psychol* 1971;89(1): 46–55.
18. Starmer C. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *J Econ Lit* 2000;38:332–382.
19. Prelec D, Loewenstein G. Decision-making over time and under uncertainty - a common approach. *Manage Sci* 1991;37(7):770–786.
20. Baucells M, Heukamp FH. Probability and time tradeoff. IESE Business School; 2007. Working Paper.
21. Baucells M, Heukamp FH. Common ratio using delay. *Theory Decis* 2010;68:149–158.
22. Baucells M, Heukamp FH, Villasís A. Trading-off probability and time: experimental evidence. IESE Business School; 2009. Working paper.
23. Easterlin RA. Explaining happiness. *Proc Natl Acad Sci U S A* 2003;100(19): 11176–11186.
24. Laury S, Holt C. Further reflections on prospect theory. 2000. Working paper.
25. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992; 5:297–323.

26. Goldstein WM, Einhorn HJ. Expression theory and the preference reversal phenomena. *Psychol Rev* 1987;94(2):236–254.
27. Prelec D. The probability weighting function. *Econometrica* 1998;66(3):497–527.
28. Ebert JEJ, Prelec D. The fragility of time: time-insensitivity and valuation of the near and far future attention to the future on far and near future valuations. *Manage Sci* 2007;53(9):1423–1438.
29. Booi AS, van Praag B, van de Kuilen G. A parametric analysis of prospect theory's functionals for the general population. *Theory Decis* 2010;68(1–2):115–148.
30. Donkers B, Melenberg B, Van Soest A. Estimating risk attitudes using lotteries; a large sample approach. *J Risk Uncertain* 2001;22:165–195.
31. Halek M, Eisenhauer JG. Demography of risk aversion. *J Risk Insur* 2001;68:1–34.
32. Hartog J, Ferrer-i Carbonell A, Jonker N. Linking measured risk aversion to individual characteristics. *Kyklos* 2002;55:3–26.
33. Hsee CK, Rottenstreich Y. Music, pandas, and muggers: on the affective psychology of value. *J Exp Psychol* 2004;133(1):23–30.
34. Porcelli AJ, Delgado MR. Acute stress modulates risk taking in financial decision making. *Psychol Sci* 2009;20(3):278–283.
35. Glimcher PW. *Neuroeconomics: decision making and the brain*. San Diego: Elsevier Science and Technology Books; 2008.
36. Tversky A, Kahneman D. Availability: a heuristic for judging frequency and probability. *Cognit Psychol* 1973;5(2):207–232.
37. Fox CR, Clemen RT. Subjective probability assessment in decision analysis: partition dependence and bias toward the ignorance prior. *Manage Sci* 2005;51:1417–1432.
38. Bar-Hillel M. The base rate fallacy in probability judgments. *Acta Psychol* 1980;44:221–233.
39. Hoffrage U, Lindsey S, Hertwig R, *et al.* Communicating statistical information. *Science* 2000;209:2261–2262.
40. Larrick RP, Soll JB. Intuitions about combining opinions: misappreciation of the averaging principle. *Manage Sci* 2006;52(1):111–127.
41. Juslin P, Winman A, Olsson H. Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychol Rev* 2000;107:384–396.
42. Livnata A, Pippenger N. Communicating statistical information. *J Theor Biol* 2008;250:410–423.
43. Kahneman D, Slovic P, Tversky A, editors. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, 1982.
44. Gigerenzer G, Selten R. *Bounded rationality: the adaptive toolbox*. Cambridge: MIT Press; 2001.
45. Keeney RL, Raiffa H. *Decisions with multiple objectives: preference and value tradeoffs*. New York: John Wiley & Sons; 1976.
46. von Winterfeldt D, Edwards W. *Decision analysis and behavioral research*. New York: Cambridge University Press; 1986.
47. Edwards W, Fasolo B. Decision technology. *Ann Rev Psychol* 2001;52(1):581–606.
48. Katsikopoulos KV, Fasolo B. New tools for decision analysts. *IEEE Trans Syst, Man, Cybern A: Syst Hum* 2006;36(5):960–967.
49. Klein GA, Orasanu J, Calderwood R, *et al.* editors. *Decision-making in action: models and methods*. Norwood (NJ): Alex; 1993.
50. Tversky A. Elimination by aspects: a theory of choice. *Psychol Rev* 1972;79(4):281–299.
51. Gigerenzer G, Todd PM, the ABC Research Group. *Simple heuristics that make us smart*. New York: Oxford University Press; 1999.
52. Simon HA. A behavioral model of rational choice. *Q J Econ* 1955;69(1):99–118.
53. Simon HA. Rational choice and the structure of the environment. *Psychol Rev* 1956;63:129–138.
54. Kahneman D, Tversky A, editors. *Choices, values and frames*. Cambridge: Cambridge University Press and the Russell Sage Foundation; 2000.
55. Gilovich Th, Griffin D, Kahneman D, editors. *Heuristics and biases: the psychology of intuitive judgment*. Cambridge: Cambridge University Press; 2002.
56. Ford JK, Schmitt N, Schechtman SL, *et al.* Process tracing methods: contributions, problems, and neglected research questions. *Organ Behav Hum Decis Process* 1989;43(1):75–117.
57. Gigerenzer G, Engel E. *Heuristics and the law*. Cambridge: MIT Press; 2001.
58. Payne JW, Bettman JR, Johnson EJ. *The adaptive decision maker*. New York: Cambridge University Press; 1993.

59. Hogarth RM, Karelaia N. Heuristic and linear models of judgment: matching rules and environments. *Psychol Rev* 2007;114(3):733–758.
60. Katsikopoulos KV, Martignon L. Naïve heuristics for paired comparisons: some results on their relative accuracy. *J Math Psychol* 2006;50:488–494.
61. Baucells M, Carrasco JA, Hogarth RM. Cumulative dominance and heuristic performance in binary multiattribute choice. *Math Social Sci* 2008;56(5):1289–1304.
62. Wu G, Zhang J, Abdellaoui M. Testing prospect theories using tradeoff consistency. *J Risk Uncertain* 2005;30:107–131.
63. Wu G, Markle A. An empirical test of gain-loss separability in prospect theory. *Manage Sci* 2008;54:1322–1335.
64. Birnbaum MH. Three new tests of independence that differentiate models of risky decision making. *Manage Sci* 2005;51(9):1346–1358.
65. Loewenstein G, O'Donoghue T, Rabin M. Projection bias in predicting future utility. *Q J Econ* 2003;118(3):1209–1248.
66. Read D, Loewenstein G, Rabin M. Choice bracketing. *J Risk Uncertain* 1999;19(1-3):171–197.
67. Jullien B, Salanié B. Estimating preferences under risk: the case of racetrack bettors. *J Polit Econ* 2000;108(3):503–530.
68. Wakker PP, Thaler RH, Tversky A. Probabilistic insurance. *J Risk Uncertain* 1997;15:7–28.
69. Thaler RH. Some empirical-evidence on dynamic inconsistency. *Econ Lett* 1981;8(3):201–207.
70. Loewenstein G, Prelec D. Anomalies in intertemporal choice - evidence and an interpretation. *Q J Econ* 1992;107(2):573–597.
71. Benzion U, Rapoport A, Yagil J. Discount rates inferred from decisions—an experimental-study. *Manage Sci* 1989;35(3):270–284.
72. Laibson D. Golden eggs and hyperbolic discounting. *Q J Econ* 1997;112(2):443–477.
73. Akerlof GA. Procrastination and obedience. *Am Econ Rev* 1991;81(2):1–19.
74. Benartzi S, Thaler RH. Myopic loss aversion and the equity premium puzzle. *Q J Econ* 1995;110(1):73–92.
75. Katsikopoulos KV, Gigerenzer G. One-reason decision-making: modeling violations of expected utility theory. *J Risk Uncertain* 2008;37(1):35–56.

DESCRIPTIVE MODELS OF PERCEIVED RISK

YITONG WANG

L. ROBIN KELLER

The Paul Merage School of Business,
University of California Irvine, Irvine,
California

JAY SIMON

Defense Resources Management Institute,
Naval Postgraduate School, Monterey,
California

Risk plays a central role in decision making. Accordingly, risk has been a popular research topic for more than four decades. Finding a generic definition of risk is hard, since this term is used in many areas such as economics, political science, management science, and medical research. However, one thing in common is that risk is always related to both the negative outcomes and uncertainty. In addition, we know that risk is normally subjective and constructed by a human's perception process. But indeed how do people perceive risks? Is there any model capable of describing this procedure and predicting people's perceived risk? In this article, we focus on empirical studies of perceived risk.

In operations research/management science as well as in psychological research, much effort has been spent on defining the subjective perception of risk. The central topic is how we can define or predict people's perceived risk of a risky option based on the characteristics of that option. The basic structure has been simply depicted in Fig. 1 below.

In Fig. 1, vector C represents an option's characteristics related to its perceived risk. The characteristics can be objectively known or can be based on subjective judgments, and the perceived risk is a real function of these characteristics. Thus, the main research objective is to find what C includes and what the functional form $f(\cdot)$ is.

On the basis of different contexts that risk perception models can be applied to, we roughly divide this topic into two subcategories. One focuses on perceived risk of monetary options, such as risky investments and gambles. The other focuses on societal risks such as natural disasters, terrorist attacks, and new technologies. These subcategories are not mutually exclusive. However, they do have many differences. For instance, monetary options can be explicitly interpreted as gambles with numerical payoffs and probabilities, but societal risks normally are hard or impossible to be transformed in that way. As a result, in the first category, C usually contains objective values from the option rather than subjective values from people's judgments, which are used in the second one.

In the next section, we review the risk perception studies of monetary risks. The section titled "Perception of Societal Risks" mainly describes the research on societal risks in a psychometric framework. The section titled "Further Developments of Risk Models" briefly introduces some developments of perceived risk. The last section concludes the article and gives some possible future research directions.

PERCEPTION OF MONETARY RISKS

Broadly speaking, monetary risks include all options whose consequences and chances of these consequences can be described in a numeric form. For example, a hit and run accident can be regarded as an event with monetary risks (i.e., a gamble) if the driver has a specific subjective probability (e.g., 50%) that there are witnesses who can recognize his/her plate number and will report to the police. He/She knows that the consequences should be a specific monetary loss (e.g., \$500). Monetary risks represent a large group of events in which people focus on the monetary consequences. In the remainder

of this section, we basically deal with abstract gambles. However, keep in mind that those gambles are just parsimonious representations of many events in real life.

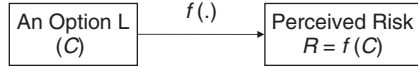


Figure 1. Basic structure of perceived risk models.

The objective of empirical research on perceived risk models is to discover and verify C and $f(\cdot)$. In most studies, they are first developed on the basis of a series of assumptions or axioms (see Jia *et al.* [1] and article titled *Axiomatic Models of Perceived Risk* for details), then they are tested by experiments. There are two possible ways to test the proposed model: (i) to test the validity of its assumptions or (ii) to test the power of this model using people's holistic judgments; either way verifies the model. In the following parts of this section, we do not differentiate between the two ways to test models but focus on the descriptive tests of different models.

Experiments on Models Using Moments

Just as the name implies, the moments models take C as moments of the distribution of outcomes. These experiments were conducted in the late 1960s–1980s. Coombs and his colleagues [2–5] used the moments of the distribution of payoffs such as mean, variance, and skewness to predict people's risk perceptions. Their work was motivated by Coombs' portfolio theory [6,7], which stated that people face a trade-off between the expected value and perceived risk when choosing among risky options. For example, in one of Coombs and Huang's papers [2], they formulated the gambles into the form of $[a + b, 50\%; b - a]^c$ (i.e., a gamble gives you a 50% chance to win " $a + b$ " and a 50% chance to win " $b - a$," and this gamble will be repeated " c " times), which has a mean factor " b ," a dispersion factor " a " and a repeat time factor " c ." Varying the value of " a ," " b ," and " c ," they asked subjects to rank the perceived risk of several sets of gambles. Their data supported the moments model. However, later Barron [8] found contrary results in his experiments.

From a more rigorous modeling approach, Pollatsek and Tversky [9] and Rotar and Sholomitsky [10] provided an axiomatic system that supported their risk theory. Their axioms were also tested by experiments [3,4]. Generally speaking, the moments models are not totally successful as descriptive models since the empirical tests give mixed results. However, using distributional variables to predict people's perception could be suitable in the sense that it is very intuitive. Believing that, Lopes [11] used complex gambles (i.e., gambles with more than three outcomes) as stimuli to test distributional models and stated that "the data support the distributional model of risk . . . , and they show that people's judgments of possible risks are similar functionally to judgments of distributional inequality."

Luce's Assumptions and Related Experimental Results

Luce [12,13] made assumptions to define several alternative measures of perceived risk. Please refer to Refs 12–16 and the article titled *Axiomatic Models of Perceived Risk* in this encyclopedia for the details of the derived measurements and their extensions. Here we only describe the two key assumptions. Specifically, the first assumption claimed that if all outcomes of a gamble are multiplied by a constant number, the risk either increases additively or multiplicatively. The second assumption claimed that there might be two ways to transfer the random variable (i.e., a gamble) into a single number (i.e., the perceived risk): (i) to aggregate some transformation of the random variable or (ii) to aggregate some transformation of the density function of the random variable. Notably, this is different from the moments models since it derives measures of risk by assuming some specific characteristics of people's risk judgments. Luce's work [12,13] only provided several possible models. Then, there were several studies testing both the assumptions and people's holistic judgments of his model. For example, Keller *et al.* [14] empirically tested the axioms proposed by Luce. In their study, they asked subjects to rank or directly compare several groups of well-defined gambles.

They observed “remarkable consistency in the risk judgments of the US and German subjects.” They also found many other interesting results: first, adding a positive number to all payoffs decreased the risk of the gamble; second, when the probability of loss was high, it had relatively more influence on people’s risk perception, whereas when the probability of loss was low, the amount of loss had more influence; third, the skewed gambles were more risky than corresponding symmetric gambles. Almost at the same time, Weber [17] tested the validity of Luce’s assumptions and got the result that two-thirds of the data supports the additivity assumption. On the basis of empirical results [17,18], Luce and Weber [15] axiomatized a risk perception model called *conjoint expected risk* (CER), which states that the perceived risk of a gamble is a weighted function of five dimensions: probability of gain, probability of loss, probability of status quo, expected values of gain and loss. Weber [18] ran several experiments testing the holistic validity of the CER model and found that (i) the CER model could predict people’s risk perception for risky gambles relatively well and (ii) the CER model was not very powerful to deal with two-outcome gambles. Later, Weber and Bottom [19] empirically tested the adequacy of the axioms of the CER model, and found support for transitivity (if A is more risky than B , and B is more risky than C , then A is more risky than C) and monotonicity (if A is more risky than B , then for any C , $p \cdot A + (1 - p) \cdot C$ is more risky than $p \cdot B + (1 - p) \cdot C$). In addition, they found that the conjoint structure had more power on gambles with negative outcomes than on gambles with positive outcomes.

Two-Attribute Models for Perceived Risk

By decomposing a lottery (X) into its mean ($\bar{X}$) and its standard risk ($X' = X - \bar{X}$), Jia *et al.* [1,20] (see also article titled ***Axiomatic Models of Perceived Risk*** in this encyclopedia) used a two-attribute model to represent people’s preference such as value and risk. The model was based on expected utility theory [21]. Since the general form of this model is very flexible, by varying the functional forms of each part, this two-attribute model may be

transformed into many existing models such as Pollatsek and Tversky’s risk model [9] and Bell’s disappointment model [22]. It also included possible different effects between positive payoffs and negative payoffs, which were emphasized in prospect theory [23]. For a detailed description, please refer to the article titled ***Axiomatic Models of Perceived Risk*** in this encyclopedia. The fundamental assumptions (axioms) of this model were also tested, and “the data indicate that the participants’ responses were generally consistent with the key assumptions of risk-value models” [24].

In this section, we reviewed several studies on perceived risk of financial gambles. As mentioned earlier, this type of gamble is an abstract representation of a large group of real-life events. Thus, they are important contributions to behavioral decision research as well as public policy making. Though so far no functional form $f(\cdot)$ has been proved to be universally valid even in very simple situations, researchers did find several patterns that are generally consistent among groups. They are (also see ***Axiomatic Models of Perceived Risk***):

1. When the variability (range, variance) of gambles increases, the perceived risk increases [2,25].
2. When the expected loss increases, the perceived risk increases [2,25].
3. When a constant amount is subtracted from the positive outcome, the perceived risk increases [14].
4. When all outcomes of a zero-mean lottery are multiplied by a positive number, the perceived risk increases [26].
5. When a zero-mean gamble is repeated many times, the perceived risk increases [26].
6. A gamble repeated fewer times with a high expected loss is more risky than a gamble repeated more times with lower expected loss [25].

We believe it is unlikely that a mathematically simple form can capture all aspects of perceived risk in realistic situations, but we do believe that every step further in the

research on perceived risk will shed some light on how people really make decisions and what kind of information or framing can lead people to better decisions under risk.

PERCEPTION OF SOCIETAL RISKS

Is taking a flight more risky than driving a car? Is nuclear power more risky than smoking? Probably most people will say yes in both cases. However, in terms of the monetary risks that we introduced in the last section, their answers are wrong, since there is no evidence that the latter ones have lower probabilities of losses or lower potential losses to the human beings. But why do people think in that way and are they right? As we said at the beginning of this article, people have different interpretations of risks [27]. Experts normally define the risk of a certain event by its consequences (e.g., annual fatalities) and the probability of those consequences. A lay person's judgment of perceived risk is related to additional characteristics of that event. Thus, when we try to predict the general public's perception and response to some risky events, especially those events with rare and delayed consequences, the tools we used in monetary risks do not suffice. We need to employ broader factors that relate to people's risk judgments. In this section, we focus on a special type of risks, which are sometimes hard to convert to pure numeric presentations. They are defined as societal risks.

Consider Fig. 1 again. People's perception of any kind of risk should come from the characteristics of this risk (C). It can be objective or subjective. In the domain of financial risks, we use objective values to describe people's risk perception. However, in the domain of societal risks, we use both objective and subjective characteristics. Slovic *et al.* [27–29] developed a psychometric paradigm for societal risks that can be used to understand different risks in a multidimensional manner. They used psychological scaling and multivariate analysis methods to give a quantitative representation to risks. The dimensions vary among different studies. For

example, in Ref. 28, the authors employed nine dimensions such as voluntariness, immediacy of effect, known to exposed, known to science, controllability, newness, chronic-catastrophic, dread-common, and severity of consequences. The explanations of these terms [28, p. 133] are as follows: “(i) Voluntariness: Do people get into these situations voluntarily? (ii) Immediacy of effect: To what extent is the risk of death immediate—or is death likely to occur at some later time? (iii) Known to exposed: To what extent are the risks known precisely by the persons who are exposed to those risks? (iv) Known to science: To what extent are the risks known to science? (v) Controllability: If a person is exposed to the risk of each activity or technology, to what extent can he/she, by personal skill or diligence, avoid death while engaging in the activity? (vi) Newness: Are these risks new, novel ones or old, familiar ones? (vii) Chronic-catastrophic: Is this a risk that kills people one at a time, or a risk that kills large numbers of people at once? (viii) Dread-common: Is this a risk that people have learned to live with and can think about reasonably calmly, or is it one that people have great dread for—on the level of a gut reaction? (ix) Severity of consequences: When the risk from the activity is realized in the form of a mishap or illness, how likely is it that the consequence will be fatal?” On these nine scales, people made judgments about their current perceived level of each dimension, which helped researchers to further analyze different risks with respect to their characteristics. In their study, they also found that these factors were not independent. Peters and Slovic [30] found that all the above dimensions could be distilled into two primary factors: dread and risk of unknown, which explained above 90% of the variance.

Numerous studies have been conducted using the psychometric paradigm. For example, Feng *et al.* [31] examined patterns of risk perceptions and decisions when people are facing consumer product-caused quality risks. They focused on the contexts of contaminated pet food and lead-painted toys. They evaluated these two risks and positioned them into a two-factor space diagram

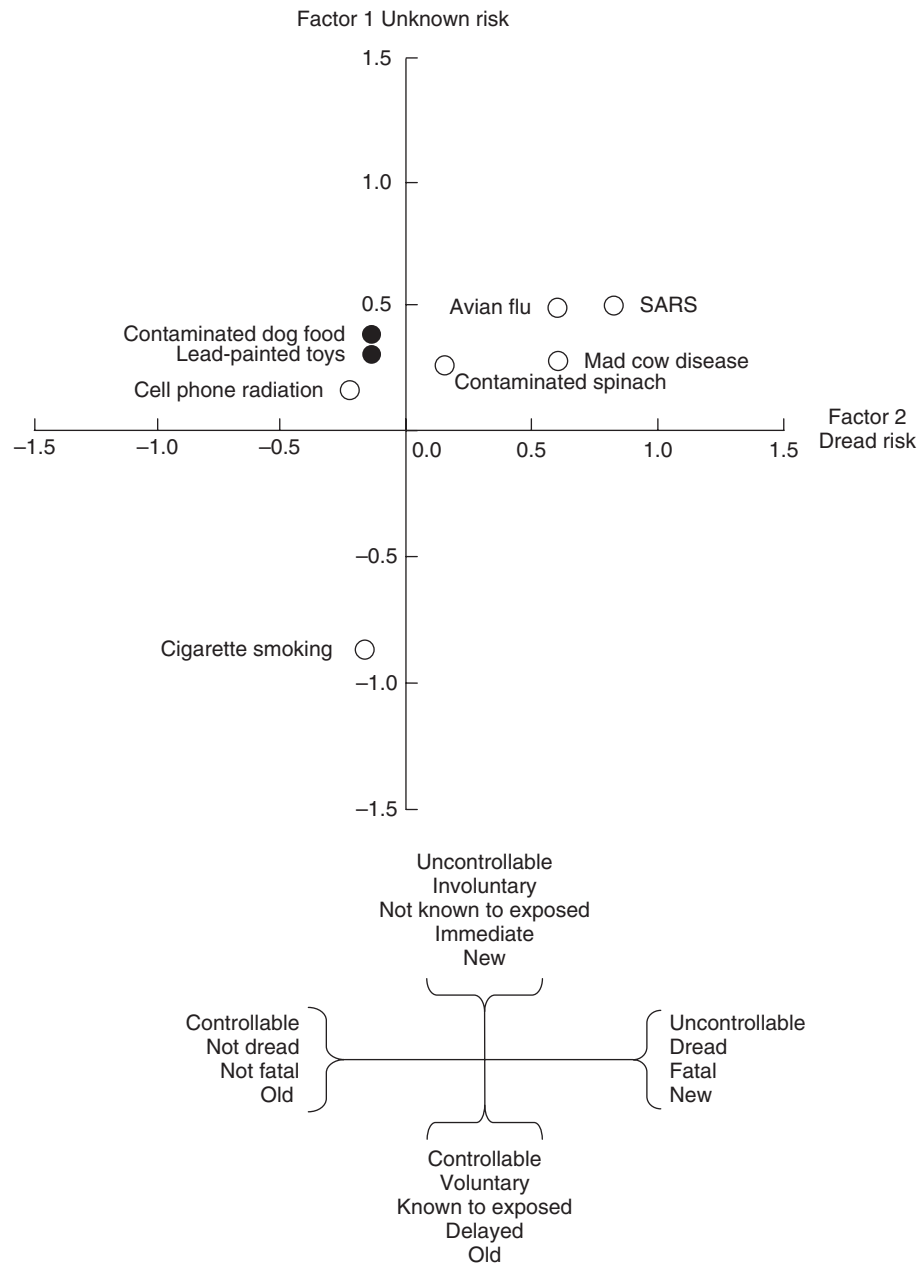


Figure 2. Location of two risks within the two-factor space. (From Feng *et al.* [31], with authors' permission).

(as shown in Fig. 2) of various societal risks. Comparing with other risks, they found that these two risks were similar in the sense that they were very closely positioned in the diagram. As stated in Ref. 27, the locations of risks in the two-factor space can largely predict people's attitudes toward the risks and also can help government and authorities to deliver the right information in the sense that it can help the general public estimate the risks correctly.

FURTHER DEVELOPMENTS OF RISK MODELS

Risk-As-Feelings Hypothesis

If we compare the studies in the last two sections, we notice a fundamental difference. The first group of studies tried to use objective factors of risks to evaluate the perceived risk. In other words, these models judged the risk based on an individual's anticipated monetary outcomes. For example, consider a lottery that gives a person a 50% chance to win \$100 and a 50% chance to lose \$85. The perceived risk of this lottery, in the first group of models, basically comes from the projections of loss and gain and their chances to occur. However, in the psychometric models, researchers used different psychological dimensions to predict a person's perceived risk. They include not only anticipated outcomes after their decisions but also feelings before they make decisions. In a highly cited paper, Loewenstein *et al.* [32] proposed a risk-as-feelings hypothesis and defined the former consideration as *anticipated emotions* and the latter consideration as *anticipatory emotions*. The risk-as-feelings hypothesis emphasizes the role of the psychological affect experienced at the moment of the decision making, shows that the emotional reactions and the cognitive judgments are often different and also states that, when emotional reactions and cognitive reactions conflict, the emotional side normally drives an individual's behavior. In their article, they also showed several determinants that could change people's emotional responses, such as mental vividness, time interval, and evolutionary makeup. The risk-as-feelings hypothesis has been used in various fields including

decision theory, psychology, and marketing. For example, Weber [33] employed the risk-as-feelings hypothesis to explain why global warming does not scare people. She explained that the affect played an important role in risk perception and people would take action only if they had a vivid personal experience. However, since global warming is not a very salient event, and involves a great temporal distance, people tend to underestimate the risk associated with it.

A Hybrid Model

Holtgrave and Weber [34] compared two models of risk perception: the CER model and the psychometric model. They used both financial and health risk stimuli, and the CER model provided a better fit if both models were used separately. However, when combining the two models together, they found that the hybrid model could obtain the best fit.

Other than comparing the power between the two models, this research also demonstrates the relationship between them. It proves that Slovic's psychometric model still has an explanatory power even after controlling for the effect of probability and outcomes. In addition, this can be regarded as being in favor of the risk-as-feelings hypothesis.

Cross-Cultural Study of Perceived Risk

The previously described studies focus on constructing a generic model of human risk perception. However, how people perceive risks is not necessarily the same across cultures. If different cultures have different perceptions of the same risk, knowing the difference can help them when there are conflicts between them [35]. For example, a Chinese company and an American company may be both better off if they can reach a settlement that both parties find less risky and in which both get higher payoffs.

Weber and Hsee [35,36] investigated subjects from China, United States, Germany, and Poland by asking them to give both the buying prices and perceived risk of risky financial lotteries. They found that the Chinese participants were more risk-seeking than American participants in a traditional

expected utility framework. However, after some further analysis, they found that the difference came from the cross-cultural difference in risk perception. That is, the Chinese subjects and American subjects are similarly risk-averse but their risk perceptions of the same lottery are different. Furthermore, Weber *et al.* [37] compared the American and Chinese proverbs of risk and risk-taking to give concrete proof of the cultural explanation of the different risk perceptions. Their result is consistent with the conclusion of the cushion hypothesis in Ref. 35: the collectivist Chinese culture gives each individual more of a cushion against financial risks since collectivism usually cushions in-group people to face financial risks and deal with the possible negative outcomes together.

Different cultures also teach people to select different risks for attention. Douglas and Wildavsky [38] divided cultures into five types such as hierarchical, individualist, egalitarian, fatalist, and hermitic. In each culture, people selectively attend to some specific categories of risks and choose to ignore others. In this way, the perceived risk of an event will vary across different types of cultures.

CONCLUSIONS AND FUTURE RESEARCH

For more than 40 years, researchers of operations research/management science, economics, and psychology have paid much attention to perceived risk. In this article, we have only reviewed part of the literature. Generally speaking, in our article, the research can be categorized into two parts.

The first part focuses on monetary risks or any risks whose information can be distilled into payoffs and their probabilities. The objective is to find a parsimonious functional form to describe and predict people's perceived risk of simple gambles. However, human perception is a complex process and normally is hard to describe in a simple and good way. As a result, there is still a long way to go. Thus, several directions still need to be developed further: (i) a parsimonious way to

model people's risk perception; (ii) risk perception and decision making under risk are two well-developed research areas, but little work tries to link them together, so it would be interesting to consider how to incorporate perceived risk models into decision making under risk; and (iii) risk perception with a time dimension.

The second part focuses on psychological research of societal risks such as Slovic's psychometric models. The objective is to find a way to decompose people's processes for perception of risks. By analyzing an individual's feelings about risks in many dimensions, researchers can categorize risks and find appropriate ways to handle risks accordingly. We believe several directions need more investigation: (i) connecting societal risk research with risk communication research to find more effective ways of risk communication; (ii) development of a direct connection between locations of risks in the two-factor space and public policies; (iii) a way to help the general public perceive risks correctly; and (iv) investigating antecedents and consequences of societal risks.

Perceived risk research is a fruitful area. This article has only reviewed part of this work. Specifically, we have only focused on the empirical studies. For those who want to know more about the axiomatic research on perceived risk, please refer to the literature review part of Ref. 1 and the article titled *Axiomatic Models of Perceived Risk* in this encyclopedia.

REFERENCES

1. Jia J, Dyer JS, Butler JC. Measures of perceived risk. *Manage Sci* 1999;45(4):519–532.
2. Coombs CH, Huang LC. Polynomial psychophysics of risk. *J Math Psychol* 1970;7(2): 317–338.
3. Coombs CH, Bowen JN. Test of VE-theories of risk and the effect of the central limit theorem. *Acta Psychol* 1971;35(1):15–28.
4. Coombs CH, Bowen JN. Additivity of risk in portfolios. *Percept Psychophys* 1971;10: 43–46.
5. Coombs CH, Lehner PE. Evaluation of two alternative models for a theory of risk I: are

- moments of distributions useful in assessing risk? *J Exp Psychol Hum Percept Perform* 1981;7(5):1110–1123.
6. Coombs CH. Portfolio theory: a theory of risky decision making. La decision. Paris: Centre National de la Recherche Scientifique; 1969.
 7. Coombs CH. Portfolio theory and the measurement of risk. In: Kaplan MF, Schwartz S, editors. *Human judgment and decision processes*. New York: Academic Press; 1975. pp. 63–85.
 8. Barron FH. Polynomial psychophysics of risk for selected business faculty. *Acta Psychol* 1976;40(2):127–137.
 9. Pollatsek A, Tversky A. A theory of risk. *J Math Psychol* 1970;7(3):540–553.
 10. Rotar VI, Sholomitsky AG. On the pollatsek-tversky theorem on risk. *J Math Psychol* 1994;38(3):322–334.
 11. Lopes L. Risk and distributional inequality. *J Exp Psychol Hum Percept Perform* 1984;10(4):465–485.
 12. Luce RD. Several possible measures of risk. *Theory Decis* 1980;12(3):217–228.
 13. Luce RD. Correction to 'Several possible measures of risk'. *Theory Decis* 1981;13(4):381.
 14. Keller LR, Sarin RK, Weber M. Empirical investigation of some properties of the perceived risk of gambles. *Organ Behav Hum Decis Processes* 1986;38(1):114–130.
 15. Luce RD, Weber EU. An axiomatic theory of conjoint, expected risk. *J Math Psychol* 1986;30(2):188–205.
 16. Sarin RK. Some extensions of Luce's measures of risk. *Theory Decis* 1987;22(2):125–141.
 17. Weber EU. Combine and conquer: a joint application of conjoint and functional approaches to the problem of risk measurement. *J Exp Psychol Hum Percept Perform* 1984;10(2):179–194.
 18. Weber EU. A descriptive measure of risk. *Acta Psychol* 1988;69(2):185–203.
 19. Weber EU, Bottom WP. An empirical evaluation of the transitivity, monotonicity, accounting, and conjoint axioms for perceived risk. *Org Behav Hum Decis Processes* 1990;45(2):253–275.
 20. Jia J, Dyer JS. A standard measure of risk and risk-value models. *Manage Sci* 1996;42:1961–1705.
 21. von Neumann J, Morgenstern O. *Theory of games and economic behavior*. 2nd ed. 1947. Princeton (NJ): Princeton University Press; 1944.
 22. Bell D. Disappointment in decision making under uncertainty. *Oper Res* 1985;33(1):1–27.
 23. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
 24. Butler JC, Dyer JS, Jia J. An empirical investigation of the assumptions of risk-value models. *J Risk Uncertain* 2005;30(2):133–156.
 25. Coombs CH, Huang LC. Tests of a portfolio theory of risk preference. *J Exp Psychol* 1970;83(1):23–29.
 26. Coombs CH, Meyer DE. Risk-preference in coin-toss games. *J Math Psychol* 1969;6:514–527.
 27. Slovic P. Perception of risk. *Science* 1987;236:280–285.
 28. Fischhoff B, Slovic P, Lichtenstein S, *et al.* How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sci* 1978;9:127–152.
 29. Slovic P, Fischhoff B, Lichtenstein S. Behavioral decision theory perspectives on risk and safety. *Acta Psychol* 1984;56:183–203.
 30. Peters E, Slovic P. The role of affect and worldviews as orienting dispositions in the perception and acceptance of nuclear power. *J Appl Soc Psychol* 1996;26:1427–1453.
 31. Feng T, Keller LR, Wang L, Wang Y. Product quality risk perceptions and decisions: contaminated pet food and lead-painted toys. *Risk Analysis*, forth coming.
 32. Loewenstein GF, Weber EU, Hsee CK, *et al.* Risk as feelings. *Psychol Bull* 2001;127(2):267–286.
 33. Weber EU. Experience-based and description-based perceptions of long-term risk: why global warming does not scare us (yet). *Clim Change* 2006;77:103–120.
 34. Holtgrave DR, Weber EU. Dimensions of risk perception for financial and health risks. *Risk Anal* 1993;13(5):553–558.
 35. Weber EU, Hsee CK. Cross-cultural differences in risk perception but similar attitudes toward perceived risk. *Manage Sci* 1998;44(9):1205–1217.
 36. Hsee CK, Weber EU. Cross-national differences in risk preference and lay predictions. *J Behav Decis Making* 1999;12:165–179.
 37. Weber EU, Hsee CK, Sokolowska J. What folklore tells us about risk and risk taking:

- cross-cultural comparisons of american, german, and chinese proverbs. *Org Behav Hum Decis Processes* 1998;75(2):170–186.
38. Douglas M, Wildavsky A. Risk and culture: an essay on the selection of environmental and technological dangers. Berkeley (CA): University of California Press; 1982.

DESIGN AND CONTROL PRINCIPLES OF FLEXIBLE WORKFORCE IN MANUFACTURING SYSTEMS

SEYED M. R. IRAVANI

Department of Industrial Engineering and
Management Science, Northwestern
University, Evanston, Illinois

INTRODUCTION

Advances in information technology and movement toward a global market that intensifies global competition have radically altered manufacturing systems over the past two decades. Under pressure to continually increase product variety, decrease price, and improve responsiveness to customers, firms have increasingly moved toward flexibility in their operations. Flexibility in manufacturing systems has been defined as the system's ability to adapt to changes in its environment [1]. Flexibility can be achieved through a variety of mechanisms, such as enhanced use of information systems, improved logistics, delayed product differentiation, small batch sizes, flexible (multipurpose) equipment, and flexible (cross-trained) workforce.

A cross-trained worker (also called an *agile worker*) is a worker who is trained to perform more than one type of task. Cross-training increases workers' utilizations by allowing them to use their idle time to perform other tasks in the system, which in turn increases the system's productivity. Cross-training also has other advantages, including (i) improving communication among workers, since they need to better coordinate tasks; (ii) effective problem solving, since more workers work on each particular task; (iii) improving worker satisfaction, due to greater compensation; and (iv) better ergonomic effects of task variety, due to less boredom and less repetitive tasks. Some examples of successful applications of utilizing flexible workforce in manufacturing systems are the module fabrication and assembly lines (cells) in John

Deere's planter factory, IBM's production line for unpopulated printed circuit boards, and the order picking operations at Gap Inc. [2].

In this article we focus only on the operational benefits of cross-training production/assembly workers in improving the performance of manufacturing systems, and we show how the performance of such systems can be improved through effective design and control of a flexible (cross-trained) workforce.¹

OPERATIONAL BENEFITS OF FLEXIBLE WORKERS

By "operational benefits of a flexible workforce" we mean the benefits that a cross-trained workforce provides in improving the operational performance of production systems, such as increasing throughput and reducing work-in-progress (WIP) inventory and cycle time. One obvious benefit of a flexible workforce is its ability to mitigate the effects of unpredictable disruptions such as worker absence. When more than one worker is trained for a given task, the absence of one worker will not have a significant impact on performance—in contrast to situations in which only one worker is trained for that task.

In this section we discuss the other two main features of a cross-trained workforce that are drivers of the operational benefits, namely (i) capacity balancing, and (ii) variability buffering.

Capacity Balancing

When a production line is not balanced in the sense that average processing times at all stations in the line are not equal, then some workers will have occasional

¹We would like to emphasize that the objective of this article is by no means to present a complete review of the literature on manufacturing systems with flexible workforce. An extensive literature on systems with cross-trained workforce exists, which we cannot discuss here due to limited space.

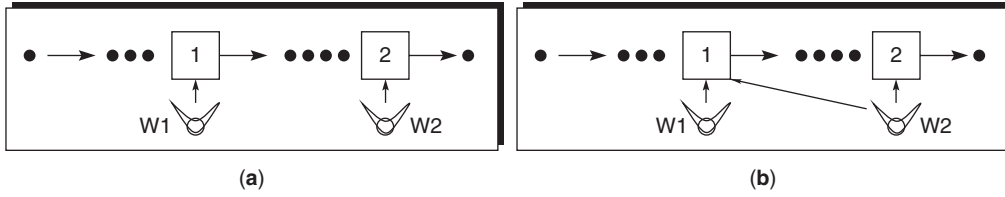


Figure 1. Two-station lines with two workers: (a) no cross-training, (b) Worker W2 is cross-trained.

idling. Cross-training enables workers to divide their effort among stations in the line, thus improving their utilization. Known as *capacity balancing*, this is the main driving force behind improving system performance through worker cross-training.

As an example, consider a two-station line in Figure 1a in which Worker W1 (Worker W2) is assigned to Station 1 (to Station 2). Processing a job at Station 1 takes 1 min (i.e., Station 1 has the capacity of producing 60 jobs an hour). After being processed at Station 1, the job is sent to Station 2. The processing time at Station 2 is 0.6 min, resulting in a capacity of 100 jobs an hour for that station. Assume deterministic processing times at each station, and suppose that ample jobs are available at the buffer of Station 1, but the buffer at Station 2 has a maximum limit of five jobs.

If workers are not cross-trained, then the maximum throughput of the line is $TH = 60$ jobs an hour, that is, the rate of the bottleneck Station 1. Note also that at the maximum capacity of 60 jobs an hour, Worker W2 at Station 2 is always idle for 0.4 min after processing a job at Station 2. One way to improve the performance of this line is *line balancing*. Line balancing requires redesigning the tasks at Stations 1 and 2 such that the processing time at Stations 1 and 2 become equal [3]. This is, however, often not possible, since not all tasks are easily divisible into smaller subtasks.

Another approach to improving performance, as shown in Figure 1b, is to cross-train Worker W2 at Station 2 to also perform the task at Station 1 (assuming that Worker W2 can work at Station 1 without interfering with Worker W1). This will increase the

throughput of the line significantly. Specifically, suppose that the line starts with five jobs at the buffer of Station 2. After serving those five jobs at Station 2 (which takes $5 \times 0.6 = 3$ min), Worker W2 switches to Station 1. At that time, the buffer of Station 2 has 3 jobs. Worker W2 processes one job at Station 1 (which takes 1 min) and then switches back to Station 2. At that time there will be five jobs at the buffer of Station 2, and this cycle repeats itself. Thus, in every 4-min cycle, five jobs are processed at Station 2 (by Worker W2), and five jobs at Station 1 (four by Worker W1 and one by Worker W2). Hence, the output rates of both stations are balanced at the rate of 75 jobs an hour. This results in a line throughput of $TH = 75$ jobs an hour, which is a 25% $(= (75 - 60)/60)$ increase in throughput due to cross-training.

As this example shows, capacity balancing, achieved in this case by cross-training Worker W2, can result in a significant improvement. The Japanese word *shojinka*² denotes this practice in the Toyota Production System. Shojinka achieves flexibility by changing the number of cross-trained workers in the line, and obtains greater productivity through capacity balancing.

Note that balancing capacity in our two-station line with two workers is simple. As the number of workers and the number of stations increases, capacity balancing naturally becomes more difficult. Nevertheless, simple Linear Programming (LP) models can be developed for balancing capacity in larger lines. For details see Refs 4–6 and references therein.

²*Shojinka* is a combination of *sho* (reduce) + *jin* (worker) + *ka* (to change)

Variability Buffering

As opposed to our simple example in which processing times were deterministic, in the real world there is always variability in the processing times in each station (due to such factors as changes in product mix, equipment failures, worker behavior, or lack of raw material). Cross-training can improve a system's performance not only through capacity balancing, but also by mitigating the negative impact of variability in the line.

Consider our example of a two-station line with no cross-training, and assume that the average processing time at both stations is 1 min. Suppose also that, the processing time at Station 2 is exactly 1 min (i.e., deterministic), but the processing time at Station 1 is uniformly distributed between 0.5 and 1.5 min.

Note that in this line with no cross-training, the capacity of the line is already balanced, since each station has an average capacity of producing 60 jobs an hour. If processing times at both stations were deterministic, then both workers had no idle time and the throughput of the line could reach its maximum of 60 jobs an hour. In this case there is no need for cross-training. When there is variability in processing times, however, the throughput of this line is less than 60 jobs an hour. This is because workers will have occasional idling, due to the variability in processing times at Station 1. Worker W1 can become idle due to the blocking of Station 1 (which occurs when the buffer of Station 2 is full), and Worker W2 becomes idle due to the starvation of Station 2 (which occurs when the buffer of Station 2 is empty). Cross-training workers will decrease workers' idle times and increase the utilization of each workstation, since instead of idling, each worker can work in the other station. This will in turn increase the throughput of the line. This benefit of cross-training that diminishes the impact of variability is called *variability buffering*.

WORKER CONTROL PRINCIPLES

The capacity balancing and variability buffering benefits of a cross-trained workforce cannot be fully utilized without a proper worker control policy. A *worker control policy* determines how each worker should allocate his or her time to the different tasks for which he or she is cross-trained. Specifically, in systems with cross-trained workers, a worker control policy is a series of rules that specifies (i) how long each worker should continue working at his/her current station, and (ii) which station she/he should switch to next.

As we will discuss in this article, the optimal worker control policies that take the greatest advantage of capacity balancing and variability buffering are often very complex. A desirable worker control policy is one that performs almost as well as the optimal policy, but it is also easy to implement and prevents worker confusion. In this article we present some easy to implement but effective worker control policies in systems with cross-trained workers. Our main focus is on production/assembly *serial lines*, which are also called *flow lines*. In an N -station flow line all jobs must be processed at station $i + 1$, after they are processed at station i ($i = 1, 2, \dots, N - 1$).

Regardless of the nature of the manufacturing environment, there are two common ways to release jobs into production lines: (i) push protocol, and (ii) pull protocol.

PUSH LINES WITH FLEXIBLE WORKERS

In a *push* line, new jobs are pushed into the production system according to a predetermined schedule. In push lines there is usually no control on the number of WIP in the line. Jobs are released into the first station in the line; in stable line, the throughput of the line will be the same as the job release rate [assuming no yield loss, [7]]. The objective in these lines is therefore to run the line so that a desired throughput is achieved with minimum WIP. For a given throughput (i.e., job release rate), this is equivalent to minimizing the cycle time.

One important consideration in push lines is the stability of the line. If the job release

rate is larger than the capacity of the line, then the WIP in the line will increase to infinity. A *stable* push line is defined as a line in which the average WIP in the line is finite.

Single-Worker Push Lines

Production lines with a single server are common in manufacturing systems. They are often known as one-worker manufacturing cells, in which one worker is fully cross-trained to perform all tasks in the line.

Consider a two-station push line in which one fully cross-trained worker is in charge of both stations; suppose also that the switchover times between stations is negligible (i.e., is zero). Switchover time D_{ij} between Station i and Station j is the time elapsed from when the worker leaves Station i to when the worker can start processing a job at Station j . Switchover time D_{ij} includes the worker's walking time from Station i to Station j , and the setup time for the machine (or station) j , which is required before processing of a job can start at Station j . Switchover times are negligible in lines where stations are close to each other and they do not require setup; or if they do, the setup times are negligible compared to processing times.

We model processing times at Station i as a random variable S_i with mean of $E[S_i]$ and we consider a case in which jobs are released randomly to Station 1 according to a Poisson process with rate λ jobs per unit time. The line would be a stable line if job release rate λ is less than the maximum capacity of the line, which is $1/(E[S_1] + E[S_2])$. Stability condition $\lambda < 1/(E[S_1] + E[S_2])$ guarantees that the average WIP in the line will not grow to infinity. The question we investigate here is as follows: for a given job release rate λ in a stable system, what is the worker control policy that minimizes the WIP and cycle time in this line?

The answer is quite simple. The optimal worker control policy is a pick-and-run policy (PR) [8], as follows:

Pick-and-Run (PR) Policy. Under a PR policy, a worker finishes one job by processing it at all stations in the line

consecutively, before starting to process the next job at Station 1.

If one considers holding cost h_i for a unit of WIP per unit time at Station i , this policy has been shown to also minimize the total expected holding cost in the system, if $h_2 \geq h_1$ [9]. Note that condition $h_2 \geq h_1$ holds in practice; since operations at Station 1 add value to jobs, the WIP at Station 2 is therefore more valuable than that at Station 1. It has been shown that the PR policy also minimizes WIP and cycle times in one-worker push lines with more than two stations and zero switchover times.

Single-Worker Lines with Switchover Times and Costs. When the switchover time D_{ij} and/or the switchover cost K_{ij} associated with worker switches from station i to j are not zero,³ the optimal worker control policy becomes more complex. For example, consider the above two-station push line with one fully cross-trained worker. If switchover times D_{12} and/or D_{21} are not zero, then the simple PR policy does not always minimize the WIP or cycle time in the line.

For this two-station line, the optimal worker control policy that minimizes the expected holding and switching costs if $h_2 \geq h_1$ is a simple greedy and exhaustive policy at Station 2, but it has a very complex structure at Station 1 [9]. Note that when $h_1 = h_2$ and $K_{ij} = 0$, this policy also minimizes the WIP and cycle time in the line.

Greedy and Exhaustive Policies. When a worker arrives at a station, under a greedy policy the server never idles in that station, and under an exhaustive policy the worker processes jobs in that station until the station becomes empty, whereupon the worker switches to another station.

Thus, under the optimal worker control policy, when the worker switches to station

³See Ref. 10 for examples of systems with switchover costs.

2, regardless of the WIP accumulation (due to job arrival) at Station 1, the worker processes all jobs at Station 2 until that Station becomes empty; only then does the worker switch to Station 1. In contrast to the exhaustive policy at Station 2, which is independent of the WIP levels at Station 1 and 2, the optimal worker control policy at Station 1 has a very complex dynamic structure that depends on the WIP levels at both stations. This policy is not easy to implement in practice. The following triple threshold (TT) heuristic policy is, however, easier to implement than the optimal policy, and has also been shown to result in a cost very close to that obtained under the optimal policy.

According to the TT policy, when the worker switches to Station 1, she starts processing jobs in that station only if the number of jobs in that station is at least M_w ; otherwise, she waits there until the number of jobs reaches that limit. When the worker does start processing jobs at Station 1, she does not switch to Station 2 until either M_n ($M_n \geq M_w$) jobs are completed without interruption, or Station 1 becomes empty and there are at least M_e jobs at Station 2 ($M_w \leq M_e \leq M_n$), whichever occurs first.

The stability condition in the push line under the TT policy is $\rho + \eta/M_n < 1$, where $\rho = \lambda(E[S_1] + E[S_2])$, and $\eta = \lambda(E[D_{12}] + E[D_{21}])$.

For systems with both switchover times and switchover costs, there exists no close-form solution for the best values for M_n , M_w , and M_e that minimize the total expected holding and switching costs. These values can be obtained through search algorithms using either a complex queueing model [9] or a computer simulation. However, for systems with only switchover costs (and zero switchover times), an efficient and robust policy can be easily obtained. Zero switchover times represent situations where the workstations are close to each other (i.e., worker walking times are negligible), and setup times are reduced significantly by performing the majority of setup operations off-line [11,12]. Thus, lines with zero switchover times but nonzero switchover costs represent situations where the workers' switching times (due to on-line machine

setup times) have been reduced to insignificance in comparison with processing times; however, there exist switching costs that represent costs associated with off-line setup operations.

For our two-station push line with switchover costs (and zero switchover times) and $h_2 \geq h_1$, the optimal worker control policy is still greedy and exhaustive at Station 2. However, the following simple heuristic policy, called *limited policy with start-up batch* (LPSB), has been shown to be very close to optimal and easy to calculate.

LPSB. According to LPSB, when a worker arrives at a station, she/he remains idle until the number of jobs in that station reaches M_s . The worker then starts processing jobs until either that station becomes empty or at most M ($M > M_s$) jobs are processed, whereupon she/he switches to another station.

The key aspect of this heuristic is that the best values for its parameters M and M_s can be easily obtained using only the first moments of stochastic processing and inter-arrival times. For example, the best value for M —namely M^* —is an integer that satisfies

$$\begin{aligned} & \frac{M^*(M^* - 1)}{2} \\ & \leq \frac{K_{12} + K_{21}}{h_2(E[S_1] + E[S_2]) + h_1(E[S_1] + 1/\lambda)} \\ & \leq \frac{M^*(M^* - 1)}{2}. \end{aligned}$$

The best value for M_s can also be obtained through a simple algorithm that uses only average job processing times $E[S_i]$ and job release rate λ [10].

Note that the main difference between the optimal worker control policy in lines with switchover times (and/or costs) and in lines without either is that in lines with switchovers, the worker does not switch to another station until a batch of jobs is processed. For example, under the LPSB policy, the worker processes a batch of size m ($M_s \leq m \leq M$) before switching to another station.

The optimal worker control policies in lines with switchover times and/or costs become very complex as the number of stations in the line increases. Those policies dictate complex batching policies, which depend on the WIP levels in each station in the line. For worker control policies in single-worker N -station lines with switchover times and costs [13–16].

Single-Worker Push Lines with Automated Machines. All push lines presented so far consist of stations in which processing a task requires the constant presence of a worker during the entire operation. We refer to these as *worker-based production environments*. Such environments are usually found in systems in which processing a job requires simple tools or nonautomated machinery. In production systems with automated machinery, however, processing a job on a machine typically requires four basic operations in the following sequence: (i) loading, (ii) setup, (iii) machine processing, and (iv) unloading. In general, the worker must be present for loading, setup, and unloading, but not during processing, which is automated. We refer to production environments with this type of equipment as *machine-based environments*. Although manufacturing cells are usually machine-based environments, most of the literature on cellular manufacturing with cross-trained operators formulate these systems as worker-based environments by assuming (explicitly or implicitly) that the worker must supervise or control the machine during processing. Reviews of this research can be found in Refs 17 and 18.

Machine-based environments present more opportunities for worker cross-training, since workers will be idle during the automatic processing times. On the other hand, worker control policies in worker-based environments are not applicable to machine-based environments. For instance, consider a single-worker two-station push line with an automatic machine at Station 1. Suppose that the operation on the automatic machine requires a loading (done by the worker) and an automatic processing (that

does not require any worker presence).⁴ As is true in practice, suppose that a worker cannot load a new job on the machine when the machine is processing a job.

We showed that in one-worker two-station push lines without automatic machines (for systems with or without switchover times or costs), the optimal worker control policy requires that after a worker switches to Station 2, he does not switch back to Station 1 (regardless of the WIP level at Station 1) until all jobs at Station 2 are processed (i.e., a greedy and exhaustive policy). With an automated machine at Station 1, however, this policy is no longer optimal.

For a system with Poisson job release and exponentially distributed task times (e.g., loading time at Station 1, and processing times at Stations 1 and 2), Hopp *et al.* [19] showed that the optimal control policy has the following properties: (i) When the machine is automatically processing a job at Station 1, the worker must switch to station 2 and work there, if Station 2 is not empty. (ii) When the machine is not processing a job, the decision on whether to work at Station 2 or to load a new job on the automated machine at Station 1 has a complex monotone structure that depends on the WIP levels at both Stations 1 and 2. When the automated machine is at station 2, however, the optimal worker control policy becomes simple: it is a fixed-priority policy, which always gives priority to loading the machine at Station 2. Only when the machine at Station 2 is processing a job, or if Station 2 is empty does the worker work at Station 1.

The optimal worker control policy becomes more complex in longer lines with more than one automated station. This makes the optimal worker control policy difficult to implement in practice. Thus, the following two-worker control policies are preferred in Hopp *et al.* [19] as easy to implement policies⁵:

⁴Unloading times are insignificant and setup is considered as a part of loading operations.

⁵For the stability condition under fixed-priority and cyclic policy, see Ref. 19.

Fixed-Priority Policy. The fixed-priority policy is a static priority policy under which the worker attends stations in a fixed order. The order specifies stations from the highest priority to the lowest. When the worker arrives at a station, she/he works at that station until she/he becomes idle (due to the station becoming empty or the automated machine being loaded, if the station is an automated station); she/he then switches to the next lower-priority station. If a job becomes available at a higher-priority station (e.g., when an automated machine finishes processing and requires unloading), the worker interrupts her/his work at the current station and immediately switches back to the higher-priority station.

Cyclic Policy. Under a cyclic policy, the worker attends workstations in a cyclic fashion. When the worker arrives at an automated station, she/he waits for the machine to finish processing the current job (if it is not yet finished), then unloads the processed job, loads the new job onto the machine, switches the machine on, and then carries the previously completed job to the next station. If the worker arrives at a manual station, she/he processes the job at that station. If the worker arrives at a station with no jobs, she/he immediately moves on to the next station in her cycle.

In a simulation study performed by Hopp *et al.* [19] of one-worker three-station push lines with one, two, or three automated stations, in which the automated machines required loading, automatic processing, and unloading operations to complete a job, it was observed that the performance of the best fixed-priority policy is always better than the performance of the best cyclic policy. Also, it has been shown that when the best fixed-priority policy is adopted, automating a station at downstream is more effective than upstream automation. In contrast, when the best cyclic policy is adopted, upstream automation is more effective than downstream automation. If one station is to

be automated in a single-worker push line, then placing automation downstream and using the best fixed-priority policy results in the lowest WIP and cycle time in the line, all other things (e.g., costs) being equal.

Multiple-Worker Push lines

One of the features of shojinka is its flexibility in adjusting the throughput of a line by changing the number of workers in the line. When demand is high, the number of workers is increased, resulting in lines with multiple (cross-trained) workers. When more than one worker can work at a given station, this creates two different kinds of environments for us to consider: (i) a *collaborative environment* in which workers can work together as team on one job, so the job can be finished faster; and (ii) a *noncollaborative environment*, in which workers cannot collaborate, although they can work independently on different jobs in a station. In this section we present some design and control principles for both collaborative and noncollaborative environments in multiple-worker push lines

The collaborative environments that we discuss in this article are those in which collaboration is *additive*. Specifically, when w workers work together on a single job, then the job completion has a rate of $\mu_1 + \mu_2 + \dots + \mu_w$, where μ_j is the job processing rate of Worker j if she/he processes the job by herself/himself.

Collaborative Environments. To start, consider a simple two-station line with two fully cross-trained workers who can work at both stations. Suppose that the processing times at Stations 1 and 2 are exponentially distributed with rates μ_1 and μ_2 , respectively. Jobs are released randomly to Station 1 according to a Poisson process with rate λ . There are holding costs h_i per job per unit time for jobs at station i , but the switchover times and costs are zero. Also, workers can preempt (i.e., interrupt) processing a job at one station to work on another job at a different station.

Consider a collaborative environment in which Workers W1 and W2 can work on a single job at station i ; the average job processing time is therefore shortened by half,

that is, the processing rate in that station is doubled to $2\mu_i$. In this case, the optimal worker allocation policy has a very simple structure [20]. If $\mu_1(h_1 - h_2) \geq \mu_2 h_2$, then the optimal worker control policy is exhaustive at Station 1. An exhaustive policy at Station 1 assigns both workers to Station 1 until it becomes empty; both workers then switch to Station 2. While working on Station 2, if a job becomes available at Station 1, both workers interrupt their work at Station 2 and switch to Station 1 immediately. On the other hand, if $\mu_1(h_1 - h_2) \leq \mu_2 h_2$, then the optimal policy is exhaustive at Station 2; this means that both workers work together on a single job at Station 1, and then switch to Station 2 to finish the processing of that job. They then switch back to Station 1 and start working together on a new job. This policy is also called an expedite policy for a collaborative environment:

Expedite (EX) Policy. Under an expedite policy all workers work on a single job at all consecutive stations until the job is processed at the last station. All workers then start working on a new job at Station 1.

Note that when $h_1 = h_2$, the objective of minimizing the expected holding cost becomes equivalent to minimizing the expected WIP in the line. On the other hand, when $h_1 = h_2$, the optimal worker control policy in two-station Markovian lines⁶ with two workers is the expedite policy (since $0 = \mu_1(h_1 - h_2) \leq \mu_2 h_2 = \mu_2$). In fact, Van Oyen *et al.* [8] show that while minimizing WIP or cycle time is the goal in a collaborative environment, the expedite policy is also optimal in longer lines with N stations ($N \geq 2$) and W workers ($W \geq 2$) and with generally distributed job processing times, as long as the switchover times and costs are zero.

⁶Markovian lines are lines in which event times (e.g., processing times, job interarrival times) are exponentially distributed.

Noncollaborative Environments. In noncollaborative environments, however, things are different.

Consider the above two-station line in a noncollaborative environment. It has been shown that condition $\mu_1(h_1 - h_2) \geq \mu_2 h_2$ is only a sufficient condition for the optimality of the exhaustive policy at Station 1; but when $\mu_1(h_1 - h_2) \leq \mu_2 h_2$, there is no guarantee for the optimality of the exhaustive policy at Station 2 [20]. For these cases, one can use a push/pull heuristic worker control policy. According to a push/pull policy, one worker is assigned to each of the two stations. Each worker works in his/her station until there is no work available there, whereupon the worker switches to the other station and works there until a job becomes available at his/her original station. A push/pull heuristic is shown to perform well in cases where the sufficient condition for the optimality of an exhaustive policy at Station 1 does not hold.

The optimal worker control policy in longer lines with more than two workers is even more complex. For an N -station line with W fully cross-trained workers and zero switchover times and costs, it has been shown that the PR policy, while not necessarily optimal, is an effective worker control policy in noncollaborative environment [8]. Under PR policy each worker works sequentially (and independent of other workers) in consecutive stations 1 to N .

Lines with Switchover Times. In lines with nonzero switchover times and/or costs, the PR or EX policies may not be optimal. As shown in one-worker lines, the existence of a setup time (or cost) calls for worker control policies that impose lower and/or upper bounds on the number of jobs (i.e., batch size) processed in a station. Furthermore, like push lines with stochastic arrivals and/or job processing times and nonzero switchover times, the stability of the line depends on the worker control policy (e.g., stability under a TT policy depends on M_n). Thus, the key issues in lines with switchover times are (i) finding what the maximum capacity (throughput) of the line is, and (ii) determining the worker control policies that can achieve the maximum capacity with lower WIP levels.

Andradóttir *et al.* [4] investigate these questions in an N -station push line with switchover times and with W cross-trained workers such that each worker is cross-trained to work on at least one station, and each station has at least one worker who is trained to work there. Job interarrivals and process times as well as worker switchover times are stochastic. Using a deterministic LP model for the flow maximization in the line (i.e., capacity balancing), they obtain a tight upper bound on the maximum capacity (i.e., throughput) of the line, as well as the fraction of time each worker needs to spend at each station to achieve that capacity. They also show that the amount of cross-training needed for achieving maximal capacity is $M + W - 1$ skills (out of a total of MK skills which when all workers are fully cross-trained).

Consider these multiple-worker push lines with switchover times in a nonpreemptive environment in which once a worker starts processing a job, it cannot be stopped until the job is completed. The following *generalized round-robin* (GRR) policy is shown to achieve a capacity (i.e., throughput) arbitrarily close to the maximum capacity.

GRR Policy for Noncollaborative Environments. Under a GRR policy, each Worker j visits the set of stations $\mathcal{V}_j$ in cyclic order. When the worker arrives at station $k \in \mathcal{V}_j$, the worker starts processing jobs at that station until $M_j^{(k)}$ jobs are served, or the worker becomes idle, whichever occurs first; the worker then switches to the next station in $\mathcal{V}_j$ (i.e., an LPSB with $M = M_j^{(k)}$ and $M_s = 0$). If Worker j has no job to process (because either all stations in set $\mathcal{V}_j$ are empty or all available jobs in that set are being served), then she/he idles.

For collaborative environments, the above GRR policy is replaced with the following policy that uses time-based thresholds rather than WIP-based thresholds for workers' movements.

GRR Policy for Collaborative Environments. Under a GRR policy, each

worker j visits the set of stations $\mathcal{U}_j$ in cyclic order. When the worker arrives at station $k \in \mathcal{U}_j$, she/he works at that station for $T_j^{(k)}$ units of time. If Station k empties before this occurs, then the worker switches to the next station in $\mathcal{U}_j$. If the worker is still working on a job when this occurs, the worker switches to the next station after completing that job. If Worker j has no job to process (because either all stations in set $\mathcal{U}_j$ are empty or all available jobs in that set are being served), then she/he idles.

For the cases of both collaborative and noncollaborative environments, Andradóttir *et al.* [4] developed an algorithm that obtains the parameters of the GRR policies (i.e., $\mathcal{V}_j$, $M_j^{(k)}$, $\mathcal{U}_j$, and $T_j^{(k)}$ for all $j = 1, 2, \dots, W$) that guarantees achieving a desired throughput.

Although the GRR policies are designed to achieve a particular throughput, they have enough flexibility to be designed for secondary considerations as well, such as lowering the WIP in the line. For systems with both high variability in the service and/or interarrival times and small values for switchover times, smaller values of WIP may be attained by lowering parameters $M_j^{(k)}$ and $T_j^{(k)}$, as appropriate.

PULL LINES WITH FLEXIBLE WORKERS

In contrast to a push protocol under which jobs are released to the line independent of the WIP level in the line, the release of new jobs under a *pull* protocol depends on the status (i.e., WIP levels) in the line. For example, under a CONWIP (CONstant Work-In-Process) pull protocol, a new job is released to the beginning of the line only when a job has been completed and has left the end of the line [7]. In pull systems, limited WIP (which can be enforced using limited buffers in the line or by other means) may cause blocking or starvation of a station in the line, which in turn reduces the throughput of the line. Thus, as opposed to push lines in which the throughput is given (i.e., it is equal to the job release rate in a stable line) and the objective is to minimize WIP, in pull lines

the maximum WIP is given and the objective is to design and control the line so that the throughput of the line is maximized.

CONWIP Lines with Flexible Workers

Consider an N -station pull line with W fully cross-trained workers and no limit on the buffer sizes at each station, and suppose that the line is run under CONWIP pull protocol. It has been shown that, if the processing times of the jobs in each station are exponentially distributed, then the EX policy maximizes the throughput of the line in *collaborative* environments with fully cross-trained workers. Of course, since all workers under the EX policy work together on a single job, the constant WIP level of one is enough for the CONWIP line to reach its maximum throughput. In *noncollaborative* environments, however, the PR policy has been shown to be optimal, when the minimum constant WIP level required to reach the maximum throughput is equal to W , that is, equal to the number of workers in the line [8].

One of the common cross-training practices is to cross-train workers assigned to nonbottleneck stations so that (if needed) they can also work in bottleneck station(s). This often results in some (but not all) of the workers being cross-trained for just some of the stations, that is, becoming *partially* cross-trained workers. The worker control policy for lines with fully cross-trained workers may not be effective or even feasible in lines with partially cross-trained workers.

As an example, consider a simple two-station CONWIP line with two workers and zero switchover times in which the average processing time at Station 1 is larger than that at Station 2; Station 1 is therefore the bottleneck. Suppose that Worker W1 is assigned to Station 1, and Worker W2 to Station 2. Worker W2, however, is also cross-trained to work at Station 1. The worker control policy, of course, only concerns Worker W2, since Worker W1 is not cross-trained. Gel *et al.* [21] show that under the optimal control policy, Worker W1 will never idle as long as there is at least one job at Station 1, and Worker W2 will never idle as long as there is at least one job at Station 2. Also, Worker W2 will not switch to

Station 1 as long as there is at least one job at Station 2. Worker W2 may, however, idle at Station 2 (and not switch to Station 1) when Station 2 is empty. On the other hand, if job processing times at Stations 1 and 2 are exponentially distributed, then it becomes optimal that Worker W2 switch to Station 1 immediately after Station 2 becomes empty.

The reverse of the above policy is also true, namely, when Station 2 is the bottleneck and only Worker W1 in Station 1 is cross-trained to work at the bottleneck Station 2. These results point to the principle that a cross-trained worker should give priority to processing the task(s) that he/she is uniquely qualified for before helping other workers out. This is called the *fixed-before-shared principle* in Gel *et al.* [21].

CONWIP Lines with Automated Machines

As shown in section titled “Single-Worker Push Lines with Automated Machines,” automating a station in a push line makes the worker control policy complex. It is interesting to study how automating a work station in a CONWIP line impacts the worker control policy in that line. Consider, for example, a three-station CONWIP line with one fully cross-trained worker, and assume that, without loss of generality, Station 1 is automated. Specifically, operations at Station 1 require loading a job on the automated machine, and automatic processing of the job by the machine (which does not require the presence of the worker). The processing at Stations 2 and 3 are manual and are performed by a single worker. Also, suppose that the worker can interrupt her tasks and switch to another station, if necessary (i.e., preemption is allowed).

Assuming exponentially distributed times for loading, as well as for both automatic and manual processing, Hopp *et al.* [22] show that the optimal worker control policy in this three-station line is a *fixed-priority policy*, which assigns the highest priority to the automated station and the second highest priority to the station that feeds the automated station. Specifically, under the optimal worker control policy, the worker first checks Station 1 (the automated station), and loads it if there is a job and the machine is not

automatically processing. She/he then checks Station 3, which directly feeds Station 1, and works there if a job is available; otherwise, she/he processes a job at Station 2, if there is any. The worker idles only when none of the above conditions is satisfied. Comparing this policy with the optimal control policy in a one-worker push line, we can easily see that the control policy in the CONWIP line is much simpler and thus easier to implement in practice.

By examining a fixed-priority policy and a cyclic policy in a simulation study of three- and five-station single-worker lines with one or more automated stations, Hopp *et al.* [22] show that, as opposed to push lines, which benefit from putting automation in downstream, the location of automation does not have a significant impact on the performance of a CONWIP line. They also identified other benefits of single-worker CONWIP lines over their push counterparts, such as efficiency, robustness, and design flexibility.

Finite Buffer Pull Lines with Flexible Workers

Now consider an N -station line with fully cross-trained workers in which, the WIP in the line is limited not by using a CONWIP protocol, but by limiting the size of the buffer at each station. Specifically, suppose that the buffer (downstream) of Station i in the N -station line can hold a maximum of B_i jobs. Also, suppose that there are always ample jobs available at the buffer of Station 1 (so Station 1 never starves), and the completed jobs at Station N depart the system immediately (so Station N is never blocked). Let the average processing time at each station be $1/\mu_i = 1$ (i.e., the line is balanced), but suppose that workers have different speeds. Specifically, if Worker j works at Station i , the average time to finish the job at that station is $1/\mu_{ij}$. A worker at Station i is blocked, if after completion of a job at Station i , she/he finds the upstream buffer at Station $i + 1$ full (i.e., production blocking).

Note that, in general, the limited buffer sizes in the line and the variability in processing times result in nonzero probability that stations are blocked or starved, which in turn results in lower throughput. Thus, the worker control policy in these lines plays an

important role in maximizing the throughput of the line. It has been shown that in two-station Markovian lines in collaborative environments with two fully cross-trained workers, the optimal worker control policy is to assign both workers to Station 1 when Station 2 is starved, and to assign both workers to Station 2 when Station 1 is blocked. Otherwise, the optimal control policy is to assign Worker W1 to Station 1 and Worker W2 to Station 2 if $\mu_{11} \mu_{22} \geq \mu_{21} \mu_{12}$, or Worker W1 to Station 2 and Worker W2 to Station 1 if $\mu_{11} \mu_{22} \leq \mu_{21} \mu_{12}$ [23].

For longer lines with $N \geq 3$ stations and the same number of fully cross-trained workers as stations ($N = W$), one extension of the above policy is as follows.

Primary Assignment. Assign Worker i to Station $j_i \in \{1, 2, \dots, N\}$ such that exactly one worker is assigned to each station and $\prod_{i=1}^M \mu_{ij_i}$ is maximized.

Contingency Plan. At any time when Station $j \in \{1, 2, \dots, M\}$ is *blocked*, the server assigned to Station j will be working at the nearest nonblocked downstream Station k ($k > j$). At any time when Station j is *starved* (but not blocked), the server assigned to that station will be working at the nearest nonstarved upstream station k ($k < j$).

If all workers have the same speed at each station (i.e., $\mu_{ij} = \mu_i$ for all $j = 1, 2, \dots, N$), and processing times are exponentially distributed, the EX policy is optimal.⁷ Note that the Ex policy achieves the maximum throughput with the minimum WIP level in the line, that is, one. However, when one or more workers are faster than others in some stations (i.e., some workers are specialized on some tasks in the line), then the Ex policy is not optimal. It has been shown that

⁷Recall that when all workers have the same speed in each station of the line, the Ex policy maximizes the throughput of the line in a collaborative environment under a CONWIP protocol with only one unit of WIP. Since in a line with finite buffers each station has the capacity of holding at least one job, the Ex policy will also be optimal in those lines.

in those cases, the above heuristic achieves higher throughput than the Ex policy. This was observed through a simulation study of three- and five-station lines [23].

Bucket Brigade Lines

Bucket brigade lines are N -station lines with W fully cross-trained workers, where $W < N$. These lines are motivated by the TSS (Toyota Sewn Products System), where there are more stations than workers. In TSS lines, workers are cross-trained to do all tasks within a zone of consecutive stations, and possibly for the entire line.

The worker control policy in bucket brigade lines, which is for lines in non-collaborative environments, is based on numbering workers from 1 to W according to their sequence in the line (in the direction of product flow). Each worker then follows the following forward and backward rules.

Forward Rule. Each worker should process a job in successive workstations, where at any station, the worker with a higher index has priority. If the job is taken over by the worker's successor, then the worker should relinquish the job and begin to follow the backward rule. If the worker is the last worker in the line (i.e., Worker W) and completes the job at the last station, she/he must relinquish the job and begin to follow the backward rule.

Backward Rule. The worker should walk back and take over the job of his/her predecessor. If the worker is the first worker in the line, the worker should walk back and pick up the raw material to start a new job at Station 1. The worker should then follow the forward rule.

Note that under the bucket brigade control policy, Worker i is always between worker $i - 1$ and $i + 1$ in the line. Even if Worker i is faster than Worker $i + 1$, she/he never passes Worker $i + 1$, because the bucket brigade policy requires that her job be handed off to Worker $i + 1$. Furthermore, in bucket brigade lines, there is always a constant WIP level

of W jobs in the line, since each worker is working on only one job at any time.

Bucket brigade lines have been shown to work effectively in environments in which (i) walking and handoff times are insignificant, (ii) each worker $i = 1, 2, \dots, W$ is characterized by a distinct, constant work velocity, and (iii) the nominal work content of the product is a constant (i.e., deterministic) and the work content is spread continuously and uniformly along the flow line. It has been shown that in such an environment, the maximum throughput of the line is achieved by arranging the workers in increasing order of their work speed, that is, by making the slowest worker be first in the line, and the fastest be last. Furthermore, under the slowest-to-fastest sequencing scheme, the line balances itself (see Ref. 24 for details). A bucket brigade line is called *balanced* if each worker repeats the same interval of work content on successive jobs. A balanced line produces at a steady rate and each worker (although fully cross-trained) ends up working on only a subset of the work content. Further studies on bucket brigade lines have shown that these lines also perform relatively well in some cases where processing times are stochastic and work content is discrete along the line [25,26; and references therein].

However, when the variability in process times is not low, and when the differences in workers' speeds are not significant, a bucket brigade control policy performs poorly. This is because, under these circumstances, substantial idle time occurs whenever workers catch up with their downstream coworkers, but are not allowed to pass them [27]. Nevertheless, bucket brigade lines have been shown to be very successful in order picking environments such as warehouses.

Finite Buffer Lines with Partially Cross-Trained Workers

As pointed out, cross-training all workers for all stations in the line is not always practical or economical. Therefore, many lines exist in practice whose workers are partially cross-trained to work only in their specified working zone. For each Worker i , a working zone Z_i is a set of consecutive stations in

the line for which the worker is cross-trained (or allowed) to work $Z_i \subseteq \{1, 2, \dots, N\}$. Note that working zones of workers may overlap, which means that there will be stations in the line for which more than one worker is responsible.

McClain *et al.* [27] performed an extensive simulation study of lines in noncollaborative environments with partially cross-trained workers, different worker speeds, different buffer sizes, different station-to-worker ratios, and different process time variability levels. The study investigates the impact on the throughput of the line on the sizes of the worker zones, the sequencing of the workers (slowest-to-fastest or fastest-to-slowest), and worker control policies.

For a wide range of lines with partially cross-trained workers and finite buffers, McClain *et al.* [27] show that the following WIP-based worker control policy is an effective one.

Half-Full Buffer. If the buffer between two stations i and $i + 1$ is below half-empty, workers work at Station i . If it is above half-full, workers work at station $i + 1$.

They also show that increasing the machine-to-worker ratio improves efficiency by giving workers more opportunity to keep busy without interfering with each other. (See Ref. 27 and references therein for more details.)

Implementation

Although there might be several worker control policies that could be implemented in a production line, one should be careful when choosing among them. The decision as to which worker control policy to adopt in practice depends on the task itself and the characteristics of the environment. These characteristics include training efficiency, switching efficiency, preemption efficiency, and collaboration efficiency [2].

Training efficiency represents how effectively workers can be cross-trained for different tasks. Factors that affect training efficiency are the cost of training, the difficulty of the tasks, and the impact of cross-training on quality. *Switching*

efficiency corresponds to how effectively a cross-trained worker can switch from one task to another. Examples of environments with low switching efficiency are lines with long walking times between work stations, and/or long setup times between two tasks. *Preemption efficiency* is concerned with how effectively one worker can take over the job of the original worker, while the original worker is still processing that job. Preemption efficiency is affected by such factors as breaking the rhythm of the workers, the need for information exchange when an incomplete job is handed off to another worker, and the possibility of quality problems due to a handoff. Finally, *collaboration efficiency* corresponds to how collaboration among workers in teams can improve performance, such as by increasing throughput or improving quality. For example, it is critical to know whether two workers who work as a team can finish a task more than twice as fast as when each works independently, or whether they can produce a better quality product as a team.

All of these environmental factors should be considered in choosing a worker control policy. For example, for environments in which training efficiency is high and switching efficiency is low, it would be appropriate to choose XP and PR policies. This is because, in order to use XP and PR policies, all workers must be cross-trained for all tasks and all workers must switch among all stations. Moreover, for the XP policy in particular, the collaboration efficiency should also be high. On the other hand, choosing a bucket brigade policy would be appropriate in an environment in which the two most critical environmental factors are training efficiency and preemption, and especially in environments where work content is continuous (e.g., order picking). And GRR policies are best suited for environments where switching efficiency is high; in this case, the collaboration efficiency must also be high if jobs are processed in teams.

DESIGN OF CROSS-TRAINING STRUCTURES

In the previous sections, we focused primarily on the worker *control* principles for some

given cross-training structures in push and pull lines. A higher strategic level decision is how to *design* worker cross-training structures (i.e., how to decide which worker should be cross-trained for which task) to achieve maximum performance.

As mentioned in the introduction, cross-training can significantly improve a system's performance through capacity balancing. Thus, the first property of an effective cross-training structure is that it should balance capacity. As described in section titled "Capacity Balancing", cross-training structures that can balance capacity can be obtained using simple LP models (see [4–6], and references therein).

Often, more than one cross-training structure can achieve capacity balancing. The decision as to which one to choose then depends on both monetary considerations such as costs of cross-training and environmental factors such as the maximum number of workers who can work at a station or on a job at any one time. For example, a structure in which all workers are cross-trained for all tasks can balance all pull and all stable push lines. However, it has a very high worker training cost; moreover, it is usually not feasible in practice for all workers to work at one station.

If more than one structure can balance capacity and satisfy system constraints, then the decision as to which structure to choose depends on the effectiveness of those structures in buffering variability. For example, consider a CONWIP line with four stations and four workers, and suppose that the average processing times at Stations 1, 2, 3, and 4, are 1, 0.6, 1, and 0.6 min, respectively. Figure 2 shows two possible cross-training structures that can balance capacity in this line. In Fig. 2a, both Workers W1 and W2 are cross-trained to work at Stations 1 and 2, while both workers W3 and W4 are cross-trained to work at Stations 3 and 4. In Fig. 2b, however, Worker W1 is cross-trained for Stations 1 and 2, Worker W2 for Stations 2 and 3, Worker W3 for Stations 3 and 4, and Worker W4 for Stations 4 and 1.

The structures in Fig. 2 are similar in many aspects. In both structures each worker is trained to work at two stations (i.e., cost of

cross-training is the same in both structures), and each station is attended by a maximum of two workers. Also, both structures balance the capacity of the line.⁸ The question is therefore which of these two structures provides better performance against variability, and thus results in a higher throughput for the line.

Using simulation for lines with different WIP and variability levels, [28] show that the structure in the first case of Fig. 2b results in a throughput that can be significantly higher than that in 2a. This example shows that cross-training structures that balance capacity and have similar properties (e.g., the same total number of skills, the same number of skills per worker) may result in significantly different performance. This implies that the variability buffering feature of cross-training structures should be considered in designing cross-training structures.

If used under the appropriate worker control policy, the full cross-training structure has the most flexible capacity balancing and the most effective variability buffering features. As mentioned, however, the bad news is that full cross-training may not always be practical or economical. The good news, on the other hand, is that in some situations, some partial cross-training structures exist, which can achieve a performance close to that of the full cross-training structure. One of these structures is the structure in Fig. 2b which is called a *two-skill chain* structure. The concept of a chain structure was first introduced by Jordan and Graves [28] for assigning products to factories. In the context of worker cross-training, the two-skill chaining structure in an *N*-Station line is defined as follows.

Two-Skill Chaining. Under a two-skill chain, the number of workers and the number of work stations are the

⁸Note that processing times at Stations 1 and 2 are the same as those in our two-station line example in section titled "Capacity Balancing." The same is true for Stations 3 and 4. Thus, capacity balancing in Fig. 2 is analogous to that for the two-station line in section titled "Capacity Balancing."

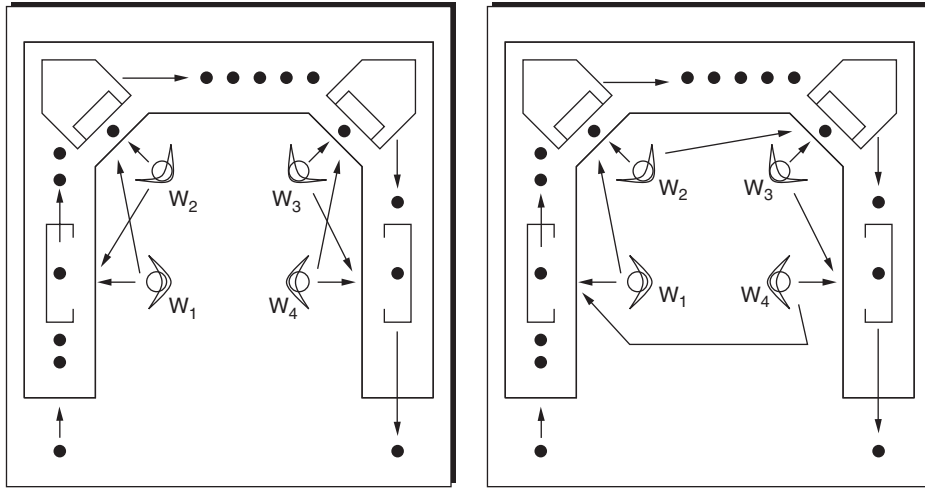


Figure 2. Two cross-training structures for a U-shaped four-station CONWIP line with four workers.

same, and each Worker i is cross-trained to work in Stations i and $i + 1$ ($i = 1, 2, \dots, N - 1$), while Worker N is cross-trained to work in Stations N and 1. This creates a chain of skills (work zones) that overlaps by only one station.⁹

After this skill chaining principle was introduced, many researchers studied the properties of skill chaining structures in supply chain design, call center agent cross-training, repairman cross-training, and worker cross-training [5,6,29–33], among others. These studies have led researchers to the following general conclusions: (i) *efficiency*: skill chaining cross-training structures capture the majority of the benefit of full cross-training structures, (ii) *robustness*: skill chaining structures are robust with respect to changes in the environment (e.g., a temporary change in workload, or changes in variability) and in worker control policy.

Note that under a full cross-training structure, all workers can directly help each other and, if needed, all workers can even work at

one station. The efficiency and robustness of skill chaining structures are due to the fact that, while only a maximum of two workers can work at each station, nevertheless all workers can also directly or *indirectly* help each other when needed [6].

In conclusion, the question of which cross-training structure is the most effective among all capacity balancing structures is not easy to answer. This is because the underlying queueing model of systems with cross-trained workers is a complex queueing system in which, not only jobs, but also servers move among queues. Thus, a computer simulation is often needed to evaluate and compare the performances of different cross-training structures. Iravani *et al.* [6], however, introduce the concept of *structural flexibility* and develop *structural flexibility metric* that can be used to choose the most effective structure among all capacity balancing structures. This metric is obtained by simply solving a deterministic max-flow problem established based on the bipartite graph of cross-training structures. This heuristic method is then shown to be very effective in choosing the most effective cross-training structures in serial and parallel job processing systems such as production lines and call centers. Iravani *et al.* [34] further extend this metric to include

⁹For a more general definition of a D -skill chain structure [6].

cases where resources (e.g., workers) have different speeds.

CONCLUSION

The design and control of manufacturing systems with a flexible (cross-trained) workforce are significantly more difficult than the design and control of traditional manufacturing systems. This article has presented design and control principles for serial production lines with flexible workers. It has shown how cross-training can improve performance through capacity balancing and variability buffering. It has also described how different manufacturing practices (e.g., push and pull), automation, operations characteristics (e.g., setup times and costs), and task nature (collaborative or noncollaborative) impact the decision as to how each worker should allocate her/his time among stations for which she/he is cross-trained. Although the optimal worker control policies are usually too complex to be used in practice, cost-effective heuristic policies, can often be found, which are simple and thus easy to implement.

The design of worker cross-training structures (i.e., deciding which worker should be cross-trained for which task) is also a difficult task. This is because manufacturing systems with cross-trained workers are complex queueing networks in which, not only jobs, but even servers move among queues. An efficient design must not only balance capacity, but also satisfy physical and environmental constraints. While full cross-training structures can effectively balance capacity and buffer variability, they are often neither economical nor practical in lines with large number of stations. In contrast, chain cross-training structures, which require less worker training, have been shown to capture the majority of the benefit of full cross-training structures. They are also reasonably robust with respect to changes in the environment.

REFERENCES

1. Sethi AK, Sethi SP. Flexibility in manufacturing: a survey. *Int J Flex Manuf Syst* 1990;2:289–328.
2. Hopp WJ, Van Oyen MP. Agile workforce evaluation: a framework for cross-training and coordination. *IIE Trans* 2004;36:919–940.
3. Nahmias S, Production and Operations Analysis. Sixth Edition McGraw–Hill/Irwin, 2009;
4. Andradóttir S, Ayhan H, Down DG. Dynamic server allocation for queueing networks with flexible servers. *Oper Res* 2003;51:952–968.
5. Hopp WJ, Tekin E, Van Oyen MP. Benefits of skill chaining in production lines with cross-trained workers. *Manage Sci* 2004;50:83–98.
6. Iravani SMR, Van Oyen MP, Sims KT. Structural flexibility: a new perspective on the design of manufacturing and service operations. *Manage Sci* 2005a;51:151–166.
7. Hopp WJ, Spearman ML. *Factory physics*. 3rd ed. Burr Ridge (IL): McGraw-Hill; 2008.
8. Van Oyen MP, Gel EGS, Hopp WJ. Performance opportunity of workforce agility in collaborative and noncollaborative work systems. *IIE Trans* 2001;33:761–777.
9. Iravani SMR, Posner MJM, Buzacott JA. A two-stage tandem queue attended by a moving server with holding and switching costs. *Queueing Syst* 1997;26:203–228.
10. Iravani SMR, Posner MJM, Buzacott JA. A robust policy for serial agile production systems. *Nav Res Log* 2005b;52:58–73.
11. Monden Y. *Toyota production systems: practical approach to operations management*. Norcross (GA): Industrial Engineering and Management Press; 1983.
12. Ohno T. *Toyota production systems: beyond large-scale production*. Cambridge (MA): Productivity Press; 1988.
13. Duenyas I, Gupta D, Lennon TM. Control of a single server tandem queueing system with setups. *Oper Res* 1998;46:218–230.
14. Van Oyen MP, Teneketzis D. Optimal stochastic scheduling of forest networks with switching penalties. *Adv Appl Probab* 1994;26:474–497.
15. Iravani SMR, Posner MJM, Buzacott JA. Operations and shipment scheduling of a batch on a flexible machine. *Oper Res* 2003;51:585–601.
16. Iravani SMR, Teo CP. Asymptotically optimal schedules for single-server flowshop problems with setup costs and times. *Oper Res Lett* 2005;33:421–430.
17. Treleven M. A review of the dual resource constraint system research. *IIE Trans* 1989;21:279–287.

18. Askin RG, Strada S. A survey of cellular manufacturing practices. In: Irani S, editor. Handbook of cellular manufacturing systems. New York: John Wiley & Sons, Inc.; 1999.
19. Hopp WJ, Iravani SMR, Shou B. Serial agile production systems with automation. *Oper Res* 2005;53:852–866.
20. Ahn HS, Duenyas I, Lewis ME. The optimal control of a two-stage tandem queueing system with flexible servers. *Probab Eng Inf Sci* 2002;16:453–469.
21. Gel ES, Hopp WJ, Van Oyen MP. Hierarchical cross-training in work-in-process-constrained systems. *IIE Trans* 2007;39:125–143.
22. Hopp WJ, Iravani SMR, Shou B, *et al.* Design and control of agile automated CONWIP production lines. *Nav Res Log* 2009;59:42–56. In press.
23. Andradóttir S, Ayhan H, Down DG. Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Manage Sci* 2001;47:1421–1439.
24. Bartholdi JJ III, Eisenstein DD. A production line that balances itself. *Oper Res* 1996;44:21–34.
25. Bartholdi JJ III, Eisenstein DD. Using bucket brigades to migrate from craft manufacturing to assembly lines. *Manuf Serv Oper Manage* 2005;17:121–129.
26. Lim Y.F, Yang K.K. Maximizing the throughput of Bucket Brigades on Discrete Workstation. *Production and Operations Management* 2009;18:48–59.
27. McClain JO, Schultz KL, Thomas LJ. Management of worksharing systems. *Manuf Serv Oper Manage* 2000;2:49–67.
28. Jordan WJ, Graves SC. Principles on the benefits of manufacturing process flexibility. *Manage Sci* 1995;41:577–594.
29. Graves SC, Tomlin BT. Process flexibility in supply chains. *Manage Sci* 2003;49:907–919.
30. Jordan WC, Inman RR, Blumenfeld DE. Chained cross-training of workers for robust performance. *IIE Trans* 2004;36:953.967.
31. Iravani SMR, Krishnamurthy V. Workforce agility in repair and maintenance environments. *Manuf Serv Oper Manage* 2007;9:168–184.
32. Gurumurthi S, Benjaafar S. Modeling and analysis of flexible queueing systems. *Nav Res Log* 2004;51:755–782.
33. Aksin OZ, Karaesmen F. Characterizing the performance of process flexibility structures. *Oper Res Lett* 2007;35:477–484.
34. Iravani SMR, Kolfal B, Van Oyen MP. Capability flexibility: a decision support methodology for parallel service and manufacturing systems with flexible servers. Working paper. Evanston (IL): Department of Industrial Engineering and Management Sciences. Northwestern University; 2009.

DESIGN CONSIDERATIONS FOR SUPPLY CHAIN TRACKING SYSTEMS

HONGYAN DAI

Hong Kong University of Science and
Technology, Clearwater Bay, Kowloon,
Hongkong

MITCHELL M. TSENG

Hong Kong University of Science and
Technology, Clearwater Bay, Kowloon,
Hongkong

MARIO M. MONSREAL

Zaragoza Logistics Center and the Center
for Latinamerican Logistics Innovation,
Avda, Zaragoza, Spain

Knowledge about products and their sub-assemblies regarding location, status, physical integrity, environment conditions and identity in the supply chain at any given time is of great economic value to stakeholders. For instance, according to the US Department of Agriculture's Food Safety and Inspection Service [1], on May 14, 2010, Sampco, Inc. recalled 87,000 pounds of beef products that may contain the animal drug Ivermectin; the next day, Montclair Meat Co., Inc. recalled 53,000 pounds of ground beef products that may have been contaminated with *Escherichia coli* O157 : H7, a deadly bacterium. Every year, the US Consumer Product Safety Commission has notified several hundred recalls in the United States, which have incurred extraordinary financial consequences. In order to track product production and distribution, so that recalls can be controlled or targeted, it is necessary to manage supply chains with tracking systems [2]. Such systems provide visibility, traceability and associated information throughout different stages of a supply chain. The rationale behind an effective tracking system includes efficient operations, product safety and security, obsolescence and shrinkage reduction, responsiveness to the market and recovery of items, which makes a supply chain tracking system imperative. However, there are several challenges to design an effective tracking system.

First of all, the continuing momentum towards globalization and a multi-layer outsourcing policy stretches supply chains across geographic and cultural boundaries, leading to longer chains [3]. As more parties in different locations are involved in one supply chain, there is a surge in the complexity of coordination and operation. Tracking the entire supply chain and designing the tracking system in a cost-effective way become increasingly difficult to achieve.

Secondly, there are various identification, communication and IT technologies available in the market, which are evolving rapidly. On top of this, some technologies have different standards to fit into a specific industry sector and operation environment. It is natural to infer that it will be a tough decision to pick up the right technology at the right level following the proper standard in order to better support the purpose of a tracking system.

Thirdly, a supply chain tracking system usually involves different IT systems. How to handle the information flow to align different interests in order to achieve information efficiency and security remains a challenge. The decision may depend on the industry sector, operation environment, and product characteristics, further complicated by the cost constraint. The answers to questions like when and where to post and retrieve information and how to facilitate the information exchange among multi-parties are still unclear.

In summary, various considerations in terms of technologies, standards, operations and strategies should be taken into account to design an effective supply chain. However, there are no complete design principles available at this moment. Some decision makers may obtain only partial guidelines, lose track of some inconspicuous but critical issues, and finally fail to fulfill the original goals. Others may invest exorbitant amounts in a tracking system with extra capacity, to resolve their problems. Therefore, it is the intent of this article to provide some feasible and practical guidelines to design an effective supply

chain tracking system in an economical way.

DESIGN CONSIDERATION OVERVIEW

Object identification (ID), data capturing, and exchange are the three main layers required to operate a tracking system, which could be illustrated by the EPCglobal architecture framework [4] in Fig. 1.

We elaborate the detail design considerations from the corresponding three aspects, ID system, reader infrastructure, and IT system in the section titled “Design Consideration for ID System,” “Design Considerations for reader infrastructure,” and “Design Considerations for IT system,” specified in Table 1. ID system considerations try to understand in what conditions to apply which ID technology and to which level in order to fit into a specific application environment. Reader infrastructure considerations relate to how to select readers to effectively capture information and communicate with the backend IT system and services. IT system considerations refer to how to exchange data efficiently.

DESIGN CONSIDERATIONS FOR ID SYSTEM

ID system considerations have two perspectives here, the first being how to choose a certain ID technology, including how to choose from (RFID) radio frequency identification and barcode, which are the two typical ID technologies for radio frequency (RF) and optical categories, respectively; and how to choose among various RFID tags. The second is what should be taken into consideration when applying the ID technology to a certain coding or packaging level.

ID Technology

RFID versus Barcode. On the basis of reports from inLogic Inc. and Atlas RFID Solutions [5,6], a brief comparison of the performance of RFID and barcode is summarized in Table 2. Basically, RFID is preferred when (i) the tag reading is required to be efficient, such as large reading speed and

range, non-line-of-sight and multi-reading; (ii) the tag is required to be rewritable, reusable or anti-counterfeiting; (iii) the operation environment is harsh such as requiring an ID label to be resistant to wear and tear, and to rough handling and inclement weather [7]; (iv) the labor cost is significant, as RFID can automate some operations and save labor costs [8]; (v) real-time information is required, as RFID enables continuous monitoring. If the tag performance could be compromised and cost is the critical consideration, barcode is competent. The choice under a specific application environment depends on the explicit estimation of the ID technology's value. It turns out to be the balance between more benefits due to RFID or less investment of barcode [9,10].

Choice Among Different RFID Tags. Different application areas and environments require RFID tags with different features and cost implications. Several important aspects are listed as follows:

1. *Operating Frequency.* There are basically three categories of RFID operating frequency: low frequency (LF), high frequency (HF), and ultra high frequency (UHF), with frequency less than 135 KHz, 13.5 MHz and 860–930 MHz, respectively [11]. From LF to UHF, the tag read range and rate increase, but the ability to read near metal or wet surface decreases. If the read range requirements are less than 1 m and do not require fast read speed, LF or HF may be an option; otherwise UHF is preferred. The application areas for LF and HF could be access control, payment ID, libraries, laundries, apparel; and that for UHF could be supply chain management, baggage handling, and electronic toll collection [11].
2. *Battery.* According to whether the device is battery powered or not, the RFID tag can be categorized as passive (no battery), semi-passive (battery assisted backscatter), and active (powered transmitter). Basically, if the application does not require a large read range (< 5 m) and long tag

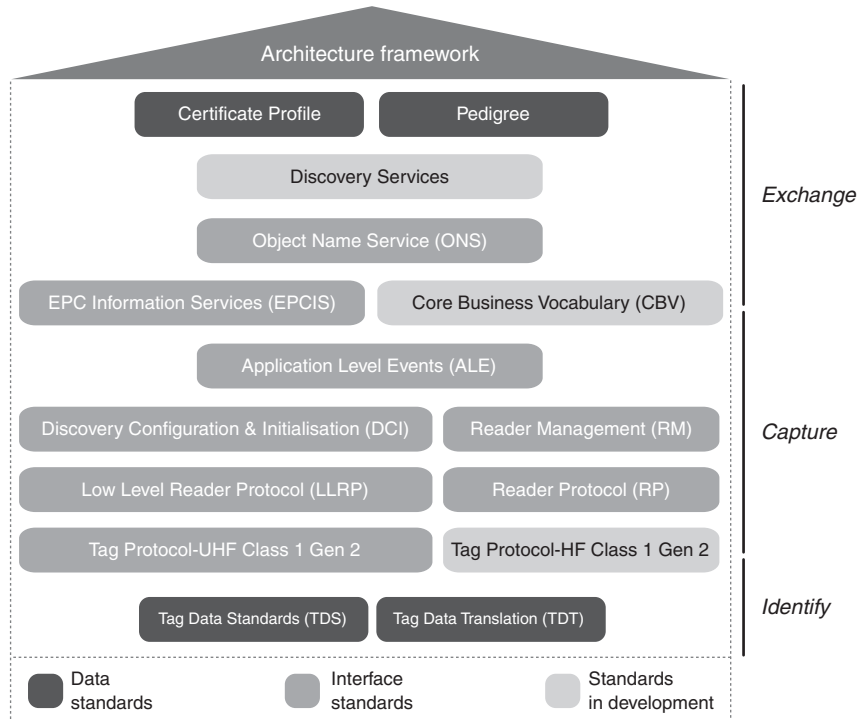


Figure 1. EPCglobal standards architecture framework [4].

life time, passive tags are preferred. When choosing power assisted tags to achieve longer read range (10–100 m), several concerns deserve attention. Firstly, those tags are usually 7 to 10 times more expensive than passives ones due to the battery cost [12]. Secondly, recharging batteries may introduce operational complexity. For example, GreenOrbs is a long-term wireless sensor network system with more than 1000 nodes in the Tianmu forest in Zhejiang, China, for forest surveillance, fire risk evaluation, and

so on [13]. The battery recharge in a kilo-scale forest significantly challenges the system maintenance. Thirdly, the RF proliferation of the active transponders and battery discharge incur environmental hazards [12]. Therefore, it is important to develop the energy harvest technology and low-cost RFID circuit designs in order to fill the gaps of cost and energy consumption for power assisted tags.

3. *Environment.* It is important to select the tags with features that are aligned

Table 1. Overview of Design Considerations

ID System	Reader Infrastructure	IT System
Section: ID technology	Section: Regional operating frequency allocation	Section: Information content
Section: ID coding level	Section: Simple versus smart	Section: Information distribution
Section: ID attaching level	Section: Reader networking	Section: Standardization

with the requirements of the use case [11]. If the operation environment is harsh, RFID tags need some special packaging; for instance, a plastic cover may be added to prevent erosion [7]. As metallic surfaces reflect RF signals, RFID tags will detune near metal resulting in lower read rate. Nagi *et al.* [14] suggests to have as few metallic materials as possible around tags and readers. If this cannot be avoided, metal mount tags with special RF engineering and encapsulation technology may be required [11], which costs up to US\$16.

Combination of Optical and RF Technologies. It is preferred that there is only one ID technology in one system for the ease of operation and management. In the following two situations, we may need to combine optical and RF technologies (i) when it is unaffordable to attach one RFID tag to each item; (ii) when the item-level RFID reading performance declines as it may incur tag reading collision as more tags are being read simultaneously. In the meantime, the possibility of inner tags (there are some obstacles between the tags and the antennas/readers) increases due to packaging limitation. Under the two situations, common practice is that each item is attached with one optical tag and the RFID tag is attached to a certain level of packaging, such as the case, carton, or container. For example, Walmart mandates its suppliers to apply RFID technologies in order to track pallets and cases, and to attach one barcode to each item for identification [15].

ID Coding Level/ID Attaching Level

ID coding level consideration relates to the concept of traceability precision in the supply chain, defined as the smallest unit with a unique ID, which could be item, batch or SKU, and so on. This determines how narrowly one can identify the grouping of material, components or other information associated with products, and logistics [16]. ID attaching level consideration refers to at which product packaging level the ID tags should be placed. ID attaching level has a tight relationship with product logistical and packaging hierarchy [17], from container, pallet, carton, case down to individual item and subpart. According to the traceability standard of GS1 (the leading organization dedicated to the design and implementation of global standards and solutions), each traceable unit should carry a unique RFID. The specific standards for different levels of coding and attaching is specified in Fig. 2 [18].

Lower ID coding (such as item) or attaching level (such as subpart) usually suggests that more detailed information can be recorded, but at the same time it incurs more cost. The basic rule for proper levels is that ID or attaching is implemented at the highest level that is sufficient to achieve the required benefit. For example, the extraordinary financial consequences of the recall crises of Mattel in 2007 are partly due to the improper ID coding level. All the products in the same SKU share the same ID code; therefore, all the products have to be removed from the supply chain as the qualified and defective products cannot be differentiated although only a few batches of toy are defective [19]. According to this fact, on the one hand, if every item has a unique ID code, Mattel may have a chance to largely

Table 2. Performance Comparison of RFID and Barcode

	RFID	Barcode
Line of sight	Not required	Required
Read range	< 100 m	< several feet
Read rate	Multiple (> 100)	One at a time
Read/write capability	Read/write	Read only
Durability	High	Low
Cost	Up to several dollars	Up to several cents

	Precision of the identification			
Unique (serialized)	Shipment identification number (SIN)	SSCC	GTIN + serial number SGTIN	GTIN + serial number SGTIN
	Not applicable	Not applicable	GTIN + batch/lot number	GTIN + batch/lot number
	Not applicable	Not applicable	GTIN	GTIN
Specific (batch)				
Generic				
	Shipment	Logistic units	Trade item not crossing the point of sale	Trade item crossing the POS, consumer unit
	Level in the logistical hierarchy			

Figure 2. GS1 unique ID standards [18].

reduce excess recalls. On the other hand, item-level ID coding does not help much in this case as most of the contamination occurs in batches, and leads to additional cost, which is difficult to be justified as the toys are up to several dollars per unit. Therefore, in this case, batch-level ID coding is the proper course of action.

Moreover, products with higher value or risk sensitivity should have lower ID coding and attaching levels [9]. This consideration is supported by current industry practice. TAGSYS, the global leader in item-level RFID infrastructure, offers solutions that focus on pharmaceuticals, fashion apparel, quality watches, jewelry, especially in the luxury market, as the item-level investments can only be afforded by those high-end products [20].

DESIGN CONSIDERATIONS FOR READER INFRASTRUCTURE

We highlight three considerations in this section to help select the right readers in order to capture information for a specific application in a cost effective way.

Regional Operating Frequency Allocation

UHF RFID tags and readers have different operating frequencies allocated by different regions, that is, US 902–928 MHz, Europe 866–869 MHz, and Japan 952–954 MHz. Therefore, for global applications, tag design for wide frequency range is crucial and usually the RF engineering has to compromise the tag performance to achieve this [11]. At the same time, multi-frequency readers should be considered for compatibility. Since being unaware of this issue could be extremely expensive for global business, various organizations for standardization and solution providers have been putting efforts to install awareness to practitioners. For example, in February 2009, GSI Hong Kong held “EPC/RFID Technology Live Test in Global Logistics” to enhance users’ understanding of the technical know-how and distribute the implementation experience of the pioneers in the international logistics business [21].

Simple versus Smart

The Intermec report [22] notes that when the RFID application requires capability of local intelligence and peripheral device control, based on the data collected from the tag,

smart readers are the clear choice. When the tag placement, volume, and speed is unpredictable, readers combined with multiple antennas and data filter algorithms could help to increase read rate and real-time local decisions.

If the primary function is to collect and transfer tag data to the backend IT system for further management, simple readers are more cost effective, especially when (i) a local IT system is already available, (ii) the orientation and location of tags are fixed, and (iii) the tag traffic is not busy that not so many tags being read at the same time.

Reader Networking

Existing RFID readers, especially for passive tags, suffer from the performance limitation in the sense of short read range, high power consumption as well as slow and unreliable read rate. For passive RFID applications, if the requirements for the read volume, range, speed, and rate are relatively high, it is important to equip readers with capability to communicate with other readers via concentrators or middleware [11]. This reader network as well as the associated MAC and system level protocol and algorithm designs could help to increase the overall read performance. At the same time, it could help to reduce the total reader amount compared with applying isolated single readers for achieving the same read performance, thereby presenting an opportunity to reduce the overall system costs.

DESIGN CONSIDERATIONS FOR IT SYSTEM

In this section, we focus on information collection, distribution and standardization by considering business characteristics in order to facilitate data retrieving and sharing.

Information Content

It is not true that gathering more information is always satisfactory. Redundant information can sidetrack the decision makers and bring complexity in operation and management. Moreover, excess information consumes more memory and cost. On the

other hand, we need to guarantee the completeness of the information set and gather all the necessary information on record to achieve information availability, defined as the ability of the systems responsible for delivering, storing and processing the information to be accessed when needed. Therefore, the precondition for a cost-effective tracking system is to make sure what kind of information should be gathered and shared, and to what extent [8,23,24]. Thomas *et al.* [8], Steele [16], Jansen-Vallers *et al.* [25], Bertrand *et al.* [26], Miller [27], and Verdenius [28] present required data categories in detail, mostly for the food industry. We try to provide some general guidance on how to select the right set of data to a proper level for a certain party in a supply chain.

Business Interest. Different parties in a supply chain may have different interests in the sort of data they want to gather. For example, manufacturers are interested in raw material and production information, while retailers pay more attention to the logistics, retailing, and product quality information. Some data can be collected within its own business, some needs to be retrieved from others. For instance, retailers can collect sales information within their four walls; however, they are also interested in product logistics and quality information, which needs to turn to the logistics providers and product inspection and quarantine departments. The union of all the parties' interested information constitutes the complete data set, requiring the cooperation of the entire supply chain.

Risk Sensitivity/Information Traffic. When the data is accumulating as time goes on, the required database memory will sharply increase. More importantly, the data retrieval speed and efficiency are likely to deteriorate when the database overspreads. If the database memory is limited, two types of data as follows have higher priority to be included in the data set: (i) the vital data whose absence may result in significant loss in profit or reputation, (ii) certain data that is of interest to various parties and is retrieved frequently with high information traffic.

Information Distribution

The circulated information in the supply chain can be stored in the tag or in the back database by linking the information to the tag ID. Depending on different information sharing modes: distributed, centralized [29], or hybrid [8], the data stored in the database could be in the local IT system of each party or in the central database, if there is one. The information in the central database may come from the local IT systems or from other institutes for the integrated statistical data. The consideration of information distribution refers to locating the proper place to store the respective kind of data. Different industry requirement, competition environment, government regulations, and industry specifications have different impacts on the information distribution strategy, which requires intensive understanding of the business context of different industries.

For example, the European Union, United States, and Japan mandate the traceability of the food industry [30–32], such as requiring a manufacturing date be recorded. This piece of information is also frequently retrieved by logistics providers, wholesalers, retailers and end customers. Therefore, it is advisable to store the manufacturing date in the central database. However, the detailed information about the raw material supply will not be accessed so frequently and the supply network is sensitive to the competitors, suggesting that only the authorized partners can access. In this sense, storing the supplier information in the local database is preferred. For example, in the wine industry, the temperature during transportation is critical to wine quality. Temperature sensors embedded in the RFID tags are usually applied in this cold chain management to

monitor the temperature change [33]. It will be more convenient for the logistics providers, the wine agents, the retailers, and even the end consumers to update or check the temperature information in the RFID tags, rather than linking to the back database for operation. Generalized from the industry observations, guidance for information distribution is summarized in Table 3.

Standardization

To ensure the tracking system works efficiently among all the involved parties with different operations, coding methods, and local IT systems, it is important to make sure everyone speaks the same language and holds the same understanding. We list several key aspects regarding standardization as follows.

Standardize the Data Semantic and Format.

For every single piece of information in the supply chain, different existing systems may have different explanations and formats. For example, “production date” may stand for the off-production line date for manufacturers, or ex-factory date for wholesalers. Different interpretations will not only cause confusion but also block traceability. Therefore, the definition of the acceptable meaning of each piece of information should be based on actual needs and ground on succinct and clear thinking. In addition, inconsistent data format may incur inefficiencies and confusions as well. For example, one product produced on 10 February 2009, can be recorded as 10/2/2009 or 2/10/2009 or 2009-10-02. Therefore, to enable information sharing across different systems, both the data semantics and format should be standardized.

Table 3. Information Distribution Guidance

Data Characteristics	Storage Location		
	Tag	Local Database	Central Database
Large volume		✓	✓
Frequent access	✓		✓
Enforced release			✓
Voluntary sharing	✓	✓	✓
Business sensitive	✓	✓	

Standardize the Interface between Different IT Systems. As the information may be stored in different IT systems, local or central, standardizing the interface among these IT systems enables data exchange and facilitates automatic information enquiry [29].

Information Exchange Standard. Practitioners should be aware that there are different information exchange standards available, such as ISO 18000, EPCglobal or uID [14,34]. No matter which one the supply chain chooses, all the parties should join the same one.

A common software specification for real-time events is EPC Information Services (EPCIS), which encompasses both interfaces for data exchange and specifications of the data itself [35]. It provides a precise definition of all types of EPCIS data and events: static data including class-level and instance-level, and transactional data including instance, quantity and business transaction observations.

DISCUSSIONS

In this article, we present detailed design considerations regarding an ID system, reader infrastructure, and IT system to guide the decision makers to design an effective supply chain tracking system in order to achieve desired benefits with minimum cost. Some of the design considerations aim to bring out hidden factors that the supply chain may be unaware of. Some end up being trade-off decisions. The key to the cost effectiveness is to notice various design concerns and try to balance them out. Some other issues, which are not discussed in detail but are also important for the success of a supply chain tracking system, are listed below.

Design Consideration Integration. The eight design considerations target different aspects of the supply chain tracking system, and the design decisions mainly depend on three factors: characteristics of supply chain, product, and information. Supply chain characteristics relate to different industry sectors, operation environments,

and processes. Product characteristics refer to different product materials, values, and sensitivities to environment. Information characteristics suggest different levels of information volume, risk and business sensitivities. If the design results based on those factors conflict, a trade-off is required. The final decision rests with budget constraint, the input from the decision makers, and the strategy of the supply chain.

Investment Estimation. When it comes to the investment of a supply chain tracking system, the costs which first come to mind are (i) hardware costs including tag, antenna, reader, sensors, printer/encoder, (ii) software costs including middleware and supporting IT system; and (iii) system implementation, integration, and maintenance costs. Most of those are tangible and are one-time fixed investments.

The challenges of the cost estimation lie in two aspects. First, some of the above-mentioned costs may be underestimated. For example, RFID middleware has not reached the “plug-and-play” stage in the current market for most applications. Therefore, some initial adopters need to spend considerable effort and money on integrating the RFID system with their existing IT system and business process.

Secondly, rather than underestimation, there are some hidden costs that are more likely to be ignored and become a hazard for investment [25]. After implementing the tracking system, daily operation and management for each party may need some changes to fit into the new system. At the same time, the operators and even the managers need some training to better understand the new system in order to take full advantage of it. All these changes incur cost, and some are exorbitant, depending on the change extent. During the system implementation and operation, the supply chain may face various complex problems and has to turn to experts from consulting companies as these problems are usually beyond their expertise. In addition, subscribing to a service incurs a service provider fee. For example, some small to medium size companies cannot afford to design their own information

exchange platform, and have to pay to share a commercial exchange platform with others.

Close-Loop or Open-Loop Implementation. For barcode systems, the current practice is open-loop implementations. Customers can see one barcode for every product they purchase. When it comes to RFID, the lowest tag price in the current market is around 10 cents, which is a luxury for most commodities for item-level or even carton-level attaching. In this case, close-loop implementation may be a feasible alternative as the tags are reused. One example could be inventory management for shrinkage reduction or efficient management. One RFID tag is attached to one keeping unit, and the related information is recorded when the unit enters the warehouse and is unattached when it leaves. This closed-loop system recycles the RFID tags, and therefore largely cuts down the tag cost. However, it incurs the cost of additional tag handing and information recording for every replenishment.

Human Factors. To ensure the success of a tracking system, human factors are also critical. Firs, Ngai *et al.* [14] notes that to secure the senior management's support and understand their requirements and needs is the first step towards success as they can commit resources, personnel and time to the project. Secondly, a supply chain tracking system is not an internal system; every party needs to collaborate with others to gain system effectiveness. Retaining effective coordination with other partners paves the way to success.

Acknowledgment

This research is supported by Hong Kong Research Grants Council (RGC CERG HKUST 620308 and 620609).

REFERENCES

1. U.S. Department of Agriculture. Food Safety and Inspection Service [Internet]. c2010. Available at http://www.fsis.usda.gov/Fsis_Recalls/index.asp. Accessed 2010 May 15, cited 2010 Jun 2.
2. Resende-Filho MA, Buhr BL. Economics of traceability for mitigation of food recall costs [Internet]. 2007. Available at <http://ssrn.com/abstract=995335>. Accessed 2010 Jun 2.
3. Dai HY, Tseng MM. The RFID impacts on lead time compression in a serial manufacturing system. CD-ROM Proceedings of the 3rd World Conference on Production and Operations Management; Tokyo: 2008. pp. 1029–1039.
4. GS1 EPCglobal [Internet]. EPCglobal standards overview. c2010. Available at <http://www.epcglobalinc.org/standards>. Accessed 2010 Jun 2.
5. inLogic Inc. [Internet]. RFID vs. barcodes comparison. c2009. Available at http://www.inlogic.com/rfid/rfid_vs_barcode.aspx. Accessed 2010 Jun 2.
6. Atlas RFID Solutions [Internet]. RFID vs. barcode. c2009. Available at <http://www.atlasrfidsolutions.com/rfid-vs-barcode.asp>. Accessed 2010 Jun 2.
7. Opara LU. Traceability in agriculture and food supply chain: a review of basic concepts, technological implications, and future prospects. *Food Agric Environ* 2003;1(1):101–106.
8. Thomas K, Katerina P, Georgios D. RFID-enabled traceability in the food supply chain. *Ind Manage Data Syst* 2007;107(2):183–200.
9. Dai HY, Tseng MM, Zipkin PH. Design of traceability systems for product recall [Internet]. 2010. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1619172. Accessed 2010 Jun 2.
10. Dai HY, Tseng MM. The impacts of RFID implementation on reducing inventory inaccuracy in a multi-stage supply chain [Internet]. 2009. Available at <http://ssrn.com/abstract=1387024>. Accessed Apr cited 2010 Jun 2.
11. Lewis S. A basic introduction to RFID technology and its use in the supply chain. Laran RFID. 2004.
12. ODIN technologies. RFID tag pricing guide. Technical report. 2009. Available at www.odintechnologies.com/rfid-tag-pricing-guide.
13. GreenOrbs [Internet]. 2010. Available at <http://www.greenorbs.org/>. Updated 2010 May 16, cited 2010 Jun 2.
14. Ngai EWT, Cheng TCE, Lai KH, *et al.* Development of an RFID-based traceability system: experiences and lessons learned from an aircraft engineering company. *Prod Oper Manage* 2007;16(5):554–568.
15. RFID Journal LLC [Internet]. Wal-Mart

- details RFID requirement. c2002-2010. Available at <http://www.rfidjournal.com/article/view/642>. Accessed 2010 Jun 2.
16. Steele DCA. Structure for lot-tracing design. *Prod Inventory Manage J* 1995;36(1):53–59.
 17. Lee D, Park J. RFID-based traceability in the supply chain. *Ind Manage Data Syst* 2008; 108(6):713–725.
 18. GS1. The GS1 traceability standard: what you need to know. Technical report. 2007. Available at www.gs1.org/.../traceability/GS1_traceability_what_you_need_to_know.pdf.
 19. Hau LL, Mitchell MM, David H. Unsafe for children: Mattel's toy recalls and supply chain management. Stanford graduate school of business, case no. GS63; 2008.
 20. TAGSYS. [Internet]. Our markets. c2009-2010. Available at <http://www.tagsysrfid.com/>. Accessed 2010 Jun 2.
 21. GS1 EPCglobal [Internet]. EPC/RFID technology live test in global logistics. c2010. Available at http://www.epcglobalinc.org/about/events/EPC_RFID_Technology_Live_Test_in_Global_Logistics/. Accessed 2010 Jun 2.
 22. Intermec. Intermec RFID readers meeting the scalable RFID challenge. Technical report. 2009. Available at http://www.intermec.com/public-files/brochures/en/RFIDReader_brochure_web.pdf.
 23. Bevilacqua M, Ciarapica FE, Giacchetta G. Business process reengineering of a supply chain and a traceability system: A case study. *J Food Eng* 2009;93(1):13–22.
 24. Manikas I, Manos B. Design of an integrated supply chain model for supporting traceability of dairy products. *Int J Dairy Technol* 2008;62(1):126–138.
 25. Jansen-Vullers MH, van Dorp CA, Beulens AJM. Managing traceability information in manufacture. *Int J Inf Manage* 2003;23(5):395–413.
 26. Bertrand JWM, Wortmann JC, Wijngaard J. Production control: a structural and design oriented approach. *Int J Prod Econ* 1990;27(2):183.
 27. Miller D. Food product traceability: new challenges, new solutions. *Food Technol* 2009; 63(1):32–36.
 28. Verdenius F. In: Smith I, Furness A, editors. Improving traceability in food processing and distribution. Cambridge: Woodhead Publishing; 2006. pp. 26–51.
 29. Kärkkäinen M, Ala-Risku T, Främling K. Efficient tracking for short-term multi-company networks. *Int J Phys Distrib Log Manage* 2004;34(7):545–564.
 30. Regulation (EC) no. 178/2002 of the European Parliament and of the Council. 2002 Jan 28.
 31. Regulation 21CFR820, U.S. Food and Drug Administration. 2004.
 32. Committee on the guidelines for introduction of food traceability systems. Guidelines for introduction of food traceability systems. Tokyo: Japan Frozen Foods Inspection Association; 2003.
 33. RFID Journal LLC [Internet]. Cold chain heats up RFID adoption. c2002-2010. Available at <http://www.rfidjournal.com/article/purchase/3243>. Accessed 2010 Jun 2.
 34. Martínez-Sala AS, Egea-López E, García-Sánchez F, *et al.* Tracking of returnable packaging and transport units with active RFID in the grocery supply chain. *Comput Ind* 2009;60(3):161–171.
 35. Armenio F, Barthel H, Dietrich P, *et al.* The EPCglobal architecture framework version 1.3. EPCglobal. 2009 Mar.

DESIGN FOR MANUFACTURING AND ASSEMBLY

BABAR M. KORAISHY
KRISTIN L. WOOD
ANDREW TILSTRA
Mechanical Engineering
Department, University of
Texas at Austin

Historically, an enterprise would consider two basic functional entities for creating a product or system idea and evolving it into an actual physical realization: design and manufacturing. Design activities were typically initiated with a detailed identification of requirements and specifications. A range of concepts were then developed around these requirements. The selected concept(s) was further refined, and an overall spatial layout was developed in the embodiment phase. This layout of the design would then proceed into the detailed design phase where materials are selected, geometry (dimensions, shape, tolerances, and surface finish) is specified, different components are iteratively modeled for performance, and final drawings are developed which are ready to be handed over to manufacturing.

Once the manufacturing department received the design, they would typically study it, define material and processing requirements, develop process plans, develop tooling and fixtures as required, test and optimize the process setup, and initiate production. The entire process from design to manufacturing was, historically, sequential in nature.

Traditionally, design and manufacturing departments worked in relative isolation from each other, and, after evolving a design to engineering drawings, the “design” would be thrown “over the wall” to the manufacturing department as shown in Fig. 1. The manufacturing department would then be left to comprehend the design after all primary design choices had been made. If manufacturing found a problem in fabricating certain aspects of the design, the design would be handed back to the design team, and a lengthy redesign process would be initiated.

This serial process led to considerable inefficiencies, time delays, and high costs [1,2].

Efforts to alleviate and overcome this “serial” approach in engineering activities led to the practice of “concurrent engineering” (CE). CE examines a product development cycle as shown in Fig. 2, and emphasizes the need to implement important considerations of downstream processes, such as manufacturing and testing (and service, disposal, recycling, disassembly, and refurbishing) during the design phase. CE encourages parallel design activities, which not only leads to refinements in the design but also reduces the “time to market” of the resulting product or system. This parallel or concurrent approach to design is achieved by developing cross-functional teams, where the design team continually interacts with area experts such as manufacturing, testing, and so on [3,4].

Such an interaction with manufacturing has further evolved into an area known as *Design for manufacturing and assembly* (DFMA). Design for manufacturing (DFM) and design for assembly (DFA), or collectively known *DFMA*, applies generally across an enterprise’s approach for developing a product or system, including both business and engineering practices. For the purposes of this article, we will consider an engineering design focus of DFMA as entailing the guidelines, tools, and techniques to improve the manufacturability and assemblability of products or systems.

DFA, as a collective set of techniques, is used to study a design and simplify the part interfaces, with the intention of making components easier to assemble. An effective level of assemblability is achieved by refining components in a design such that redundant parts are removed, and such that the remaining parts are evolved to the point where separate parts have clear functional purposes. Effective assemblability of a product or system directly translates into a reduction of assembly time and hence a reduction in assembly costs.

DFM is the study and advancement of a design where each part is examined and refined from a process choice, part feature, and tooling standpoint. DFM techniques

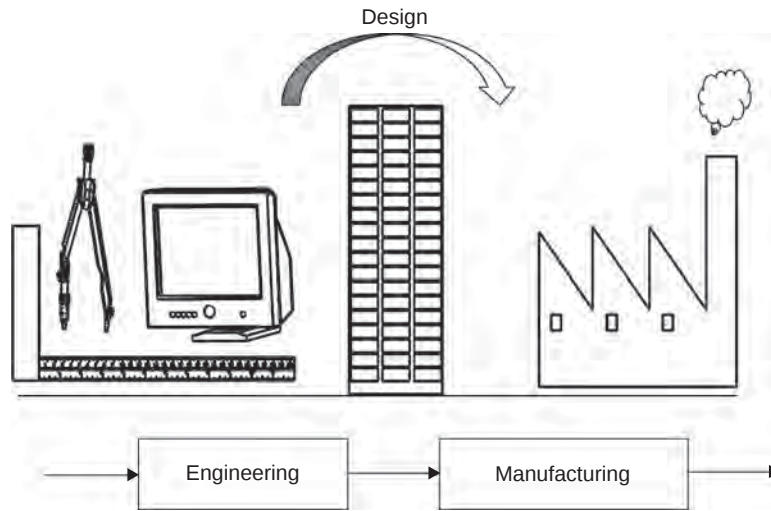


Figure 1. “Over the wall” approach to design and manufacturing.

are intended to evolve individual parts to minimize the number of manufacturing processes and process steps, while creating design features of the individual part (within constraints of their intended function) so that they are less expensive to fabricate.

Assembly techniques have received considerable emphasis over recent years, as the potential for significant improvements in product assembly time and associated profit margins have helped spread DFA interest into business entities [5]. Assembly processes typically account for between 40 and 60% of the overall production time for a product or system [6]. As a result, industry has focused on assembly processes through studies of time and motion and division of work, as well as the use of robotics and automation.

Particular advancements in DFA emerged from many national and international research efforts. Boothroyd and colleagues, for example, published an original product DFA manual in 1982 [1,7,8]. This manual is regarded as the pioneering work of DFA techniques. It is composed of a formalized, systematic approach that includes selection of an assembly method—manual, high-speed-automatic, or robotic assembly, followed by an analysis of the assemblability of a design using a DFA worksheet. The worksheet (or associated software product)

takes into account factors such as handling time, geometries, insertion time, and theoretical minimum number of parts to provide for an evaluation of the “design efficiency” of products or systems. Besides Boothroyd and colleagues, many others have contributed to the field [9–17].

In recent years, DFM has also drawn considerable attention. Industries have invested a great deal of effort and resources into new manufacturing techniques, such as gas-assisted injection molding, powder metallurgy, and laser beam machining [18]. DFM software packages such as computer-aided process planning (CAPP) have provided engineers with tools during the various design stages [17–22].

DESIGN FOR MANUFACTURING AND ASSEMBLY CONSTRUCTS: OVERVIEW AND MOTIVATION

DFMA techniques are tremendously important for any product or system design activity. Such techniques result in significant cost savings, which are achieved by reducing part count, simplifying the part interfaces so that components are easier to assemble, and creating parts that are efficient to manufacture. By refining the design of individual parts, potential manufacturing flaws or features

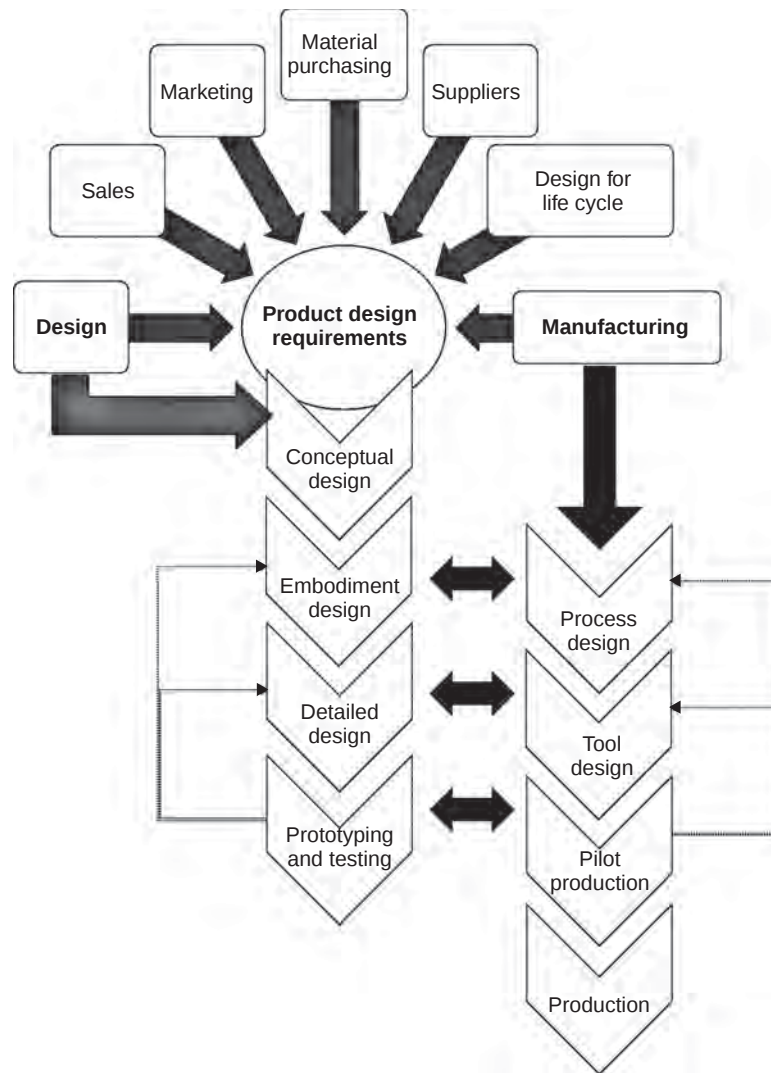


Figure 2. Product development cycle.

that cannot be economically manufactured are removed or replaced. A subsequent benefit of this activity is that part quality is also improved. Another important benefit is that, since area experts in manufacturing are intimately involved with DFMA activities, manufacturing planning can be accomplished in parallel with the design activity, which affects the overall cycle time of a project where less time is spent in redesign and iteration.

For the purposes of this article, we consider a two-part construct to DFMA. The first construct considers DFA and DFM guidelines, where a guideline is a heuristic, rule, or generally proven directive for simplifying a component for manufacturability and assemblability. Designers and design teams may apply guidelines to concepts and existing products or systems for systematic improvement. The second construct is the application of force flow analysis. Force flow

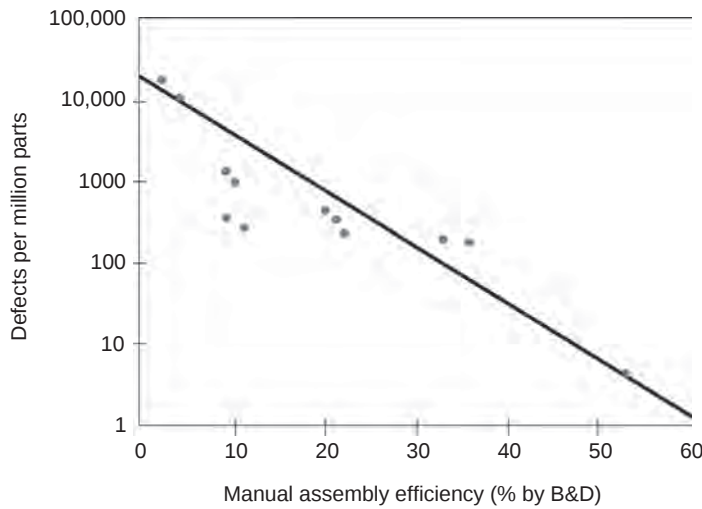


Figure 3. Exemplar plot showing reliability improvements when applying DFA.

analysis considers the systematic mapping of applied forces or efforts to a concept, existing product, or existing system to identify the degree of freedom and relative motion between components due to the flow of applied forces. Through this analysis, insights may be obtained as to the functionality of components and the identification of possibilities for direct combination of components for product or system simplification.

Historical Example

In the late 1980s, Motorola introduced a new vehicle-mounted remote radio for base-remote communication. Their product at the time, the MX Converta Com, had 217 distinct parts requiring over 2700 s of assembly time. Motorola redesigned the product using DFM and assembly methods, and reduced the part count to 97 with 1350 s of required assembly time, an 87% reduction in direct cost. The number of screw fasteners was reduced from 72 to 0 [23]. Similar results were achieved for other Motorola product lines. Figure 3 shows an exemplar plot of number of defects as a function of Manual Assembly Efficiency following the Boothroyd and Dewhurst method [1]. This plot effectively illustrates the improvement in reliability that is possible using DFA techniques.

DESIGN FOR MANUFACTURING AND ASSEMBLY APPLICATION

Design Process

It is important to very briefly introduce the product/system design and development process, so as to enable further discussion on how and at what stage in the design process DFMA techniques may be applied. An engineering design process is the set of engineering activities that are employed to transform an idea, a concept, or a business goal into a tangible physical product that meets defined customer needs and requirements. Besides other activities, the two functional activities that are relevant to the present discussion are design and manufacturing.

A typical design process can be classified into three main types as follows:

1. Original design
2. Redesign
 - a. Adaptive design
 - b. Variant design
3. Product/System benchmarking

Original design refers to those design problems where solutions result in the creation of new marketplaces or the significant expansion or redefinition of current marketplaces. Redesign, on the other hand, entails

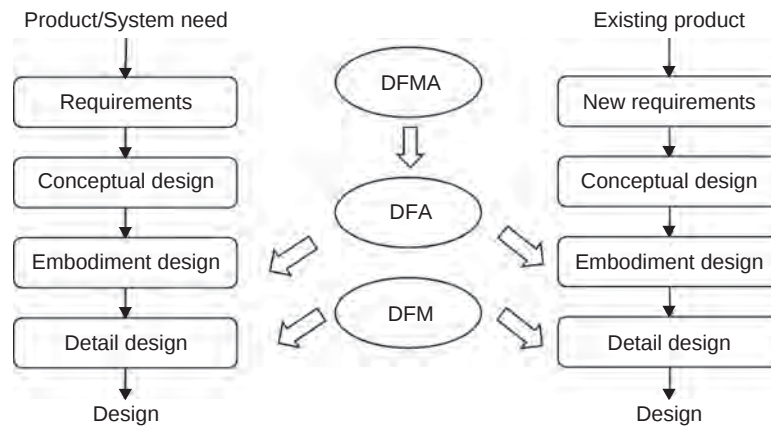


Figure 4. Application of DFA and DFM in the original and redesign process.

the evolution of a current instantiation of a product or system. Redesign involves two subtypes. In adaptive redesign, the essence of an existing design is adapted through the addition of new or enhanced functionality or the implementation of new technologies to replace current functionality. Variant redesign involves taking an existing design and evolving it parametrically, such as the scaling or reorganizing components to satisfy new requirements, or basic changes in geometry or material properties to improve product performance. A benchmarking design activity is actually not a creative activity but is typically employed to study an existing design, and quantify its performance, cost, and so on relative to competitors' designs or analogous designs.

Design for Manufacturing and Assembly Application in the Design Process

DFA techniques are typically applied before DFM techniques because DFA activities typically lead to the elimination of redundant parts, and/or lead to the significant modification of the features and functions of others. In the overall design processes (original and redesign) shown in Fig. 4, DFA is employed at the "embodiment design" stage. Embodiment design is the stage when a final selected concept is transformed into a physically realizable design with spatial form. It need not be the final form (as developed in detailed

design), but components, their functions, and layout are defined in the embodiment design stage.

Once a design reaches the detailed design stage, the physical form of every part is decided, performance modeled, materials and manufacturing processes selected, and all necessary engineering tasks are completed. This stage is the point where many design for manufacturing techniques are typically applied to the design to further refine it and eliminate any potential problems which might surface downstream in manufacturing or testing.

BASIC DESIGN FOR MANUFACTURING AND ASSEMBLY TECHNIQUES

Design for Assembly Guidelines

Figure 5 shows a basic view of a product assembly process. The different stages defined in the assembly process are part storage, part handling, insertion and position, and finally joining and fastening. DFA guidelines are basic heuristics and rules that aid the designer or design team in evaluating a design so that opportunities for improvement can be identified. A typical application of the guidelines, once a design concept(s) has been selected and preliminarily embodied, is for a designer or design teams to study the design

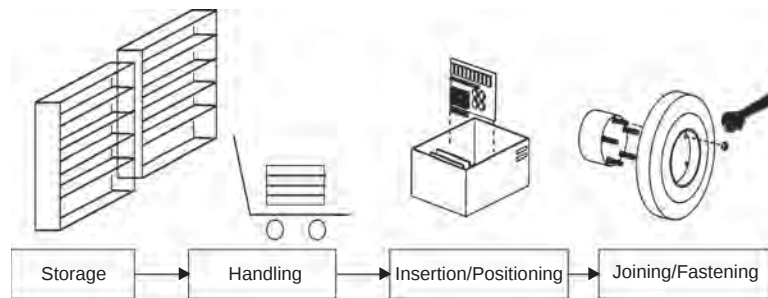


Figure 5. Assembly process.

as per every guideline, one by one, and modify the design until it satisfies the guidelines. It should be noted that all guidelines may be difficult to satisfy simultaneously depending on the desired functionality, potential manufacturing processes, and geometric layout of a product or system. Because of

this difficulty, as many guidelines should be applied as possible, resulting in alternative component layouts and geometries that are evaluated by the designer or design team.

Table 1 shows a list of DFA design guidelines [2,7,12,20,24–26]. These guidelines are

Table 1. DFA Design Guidelines

<i>System guidelines</i>	
1	Standardize components and interfaces to reduce part variety
2	Modularize multiple parts into single subassemblies
3	Assemble in open space, not in confined spaces. If possible, never bury Important components
4	Reduce part count by incorporating multiple functions into single part (function sharing)
5	Maximize part symmetry
6	Create features on parts or components to identify how to orient them for insertion
<i>Storage and handling guidelines</i>	
7	Prevent nesting of parts
8	Eliminate tangly parts
9	Color code parts that are different but shaped similarly
10	Provide features so that parts can be gripped properly
<i>Insertion and positioning guidelines</i>	
11	Provide orienting features on nonsymmetries
12	Insert new parts into an assembly from above (top-down assembly)
13	Insert parts or components from the same direction, or minimize the number of assembly directions. If possible, never require the assembly to be turned over
14	Provide alignment features on parts or components
15	Design in geometric or weight polar properties if a component or subassembly is nonsymmetric
16	Design mating features for easy insertion of components
<i>Joining and fastening guidelines</i>	
17	Eliminate fasteners
18	Use permanent or semi permanent fastening methods (such as welding, soldering, shrink fits) for joints that will require minimal disassembly
19	Allow proper spacing for fastening tools
20	Provide flats for uniform fastening and fastening ease
21	Place fasteners away from obstructions
22	Deep channels should be sufficiently wide to provide access to fastening tools. The use of no channels is preferred

organized as per the assembly stage to which they apply.

System Guidelines. The set of system guidelines shown in Table 1 apply to the overall product or system design and not to any specific stage of assembly. The first guideline indicates that, by reducing part variety, one can not only reduce assembly time but also reduce the logistics and inventory of maintaining a multitude of parts in stock. A good example for this particular guideline would be fasteners.

Breaking down a design into modules, as recommended in the second guideline, makes the assembly process more manageable and also influences quality control, as subassemblies can be individually tested rather than the entire assembly be tested for defects.

The third guideline recommends that one should not design a product where assembly would require one to reach into tight spaces with assembly tools. A product should be designed such that, during assembly, joints and fastening positions are easily accessible.

The fourth guideline urges the designer or design team to reevaluate the design and reduce the part count. There are a few different methods to do this, one of which is the force flow analysis, which is discussed in the subsequent sections.

The fifth guideline suggests that the designer should maximize part symmetry in a design so that part orientation becomes redundant. If it is not possible to create or fabricate a symmetric part, then the asymmetry should be made as obvious as possible, so that the correct orientation is obvious during assembly, and the assembler does not waste significant time in discovering the correct orientation.

The last system guideline suggests that having self-locating features designed in individual parts facilitates the assembly process, by eliminating the time spent in aligning parts. Exemplars illustrations of this and other DFA system guidelines are shown in Table 2.

Handling and Storage Guidelines. The next category of DFA guidelines is handling and

storage guidelines (Table 3). The first guideline recommends that features be built into the part that will prevent nesting. Nesting means that, when parts that are stacked on top of each other, there exists the potential that the parts will be difficult to assemble at nesting interfaces or may bind at these interfaces. Such nesting hinders their handling and storage and adds to assembly time.

The second guideline suggests that parts be designed such that they do not tangle or stick together. This hazard is especially apparent for wire harnesses, coil springs, and small components. Untangling and separating components can waste significant amounts of time and may lead to defects.

The third guideline suggests color-coding parts that are of similar shape and size so that they are easily identifiable by their color and less time is spent in identifying them during the assembly operation.

The fourth handling and storage guideline suggests that parts which are difficult to grip should have features built into them so that it becomes easier to handle them.

Insertion and Positioning Guidelines. The next category of DFA guidelines is insertion and position guidelines (Table 4). The first insertion and positioning guideline suggests that orienting features be built into nonsymmetrical parts so that the inherent asymmetries in the part can be aligned for mating. Often such features are not automatically designed when asymmetrical parts are designed, and existence of such features reduces assembly time.

The second guideline recommends that parts be designed such that top-down assembly is possible. In the case of a bottom-up assembly, the assembler would be fighting gravity when positioning and aligning parts, which increases the workload and assembly time. Alternatively, in top-down assembly, gravity will aid in part positioning and alignment. Designing a stable base part onto which other parts are attached from the top is recommended.

The third guideline is linked to the previous guideline. There exist many cases where

Table 2. DFA System Guidelines—Illustrations

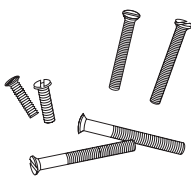
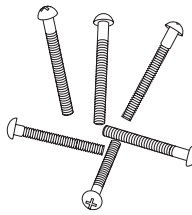
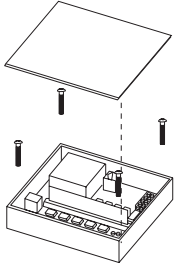
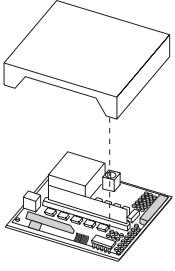
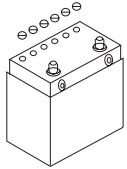
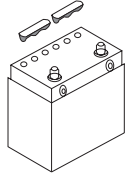
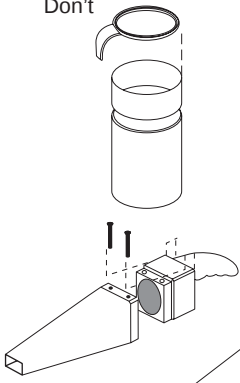
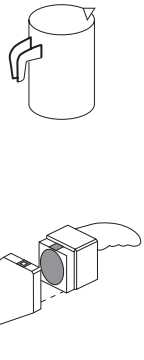
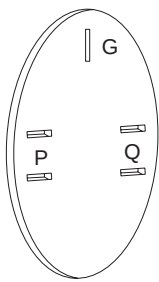
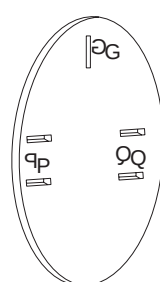
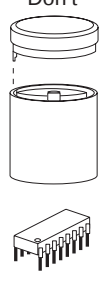
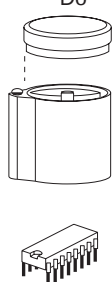
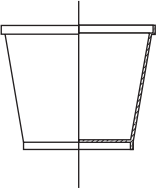
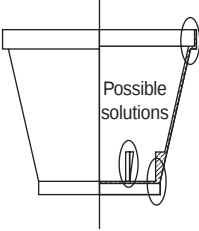
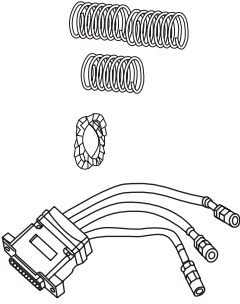
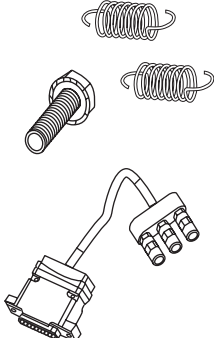
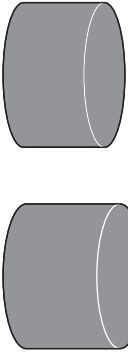
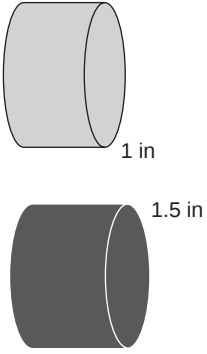
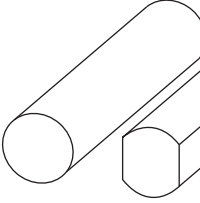
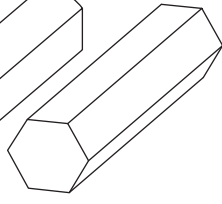
Don't  Do  Standardize to reduce part variety (a)	Don't  Do  Design open enclosures to permit assembly in open space, not in confined spaces. Never bury important components (b)
Don't  Do  Modularize multiple parts into single subassemblies (c)	Don't  Do  Part count can be reduced by implementing multiple functions into single parts (d)
Don't  Do  Maximize part symmetry (e)	Don't  Do  Parts should easily indicate orientation for insertion (f)

Table 3. DFA Handling and Storage Guidelines—Illustrations

Don't  Do  (a) Prevent nesting of parts	Don't  Do  (b) Eliminate tangly parts
Don't  Do  (c) Color code parts that are different but shaped similarly	Don't  Do  (d) Provide features so that parts can be gripped properly

a top-down assembly approach is not possible or preferred. In such cases, it is recommended to have as few part insertion directions as possible, because for every new direction, either the assembler has to reposition himself or herself, or the product assembly has to be reoriented. A primary direction for assembly and, at most, one more reorientation are recommended to complete a product or system assembly. Continual reorientations significantly add to assembly time, especially with respect to the cognitive requirements placed on an assembler.

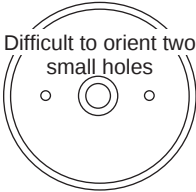
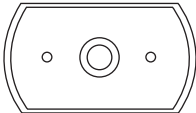
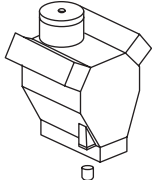
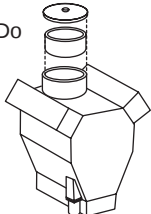
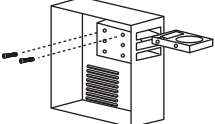
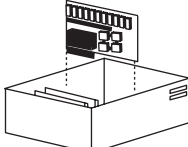
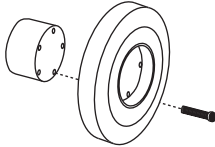
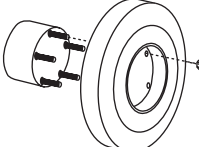
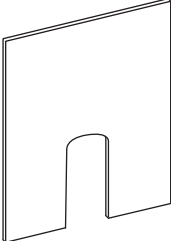
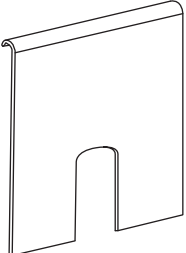
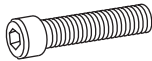
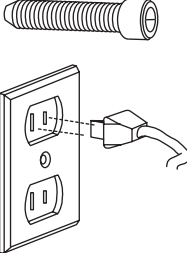
The fourth guideline indicates that alignment features be built into a product or

system design. This guideline seeks, upon insertion of part components into an assembly, that there exists automatic alignment of components, without the need for any measurements or the need for a component to be temporarily held in a position while it is fastened.

The fifth guideline suggests that asymmetrical parts be designed such that they can be easily oriented, especially those designed for automatic assembly where weight polar properties can aid in orientation during assembly.

The last insertion and positioning guideline recommends building in features such as

Table 4. DFA Insertion and Positions Guidelines—Illustrations

Don't  Do  Provide orienting features on non-symmetries (a)	Don't  Do  Insert new parts into assembly from above (b)
Don't  Do  Insert from the same direction, or very few. Never require the assembly to be turned over (c)	Don't  Do  Provide alignment features (d)
Don't  Do  For automated assembly, design in weight polar properties across non-symmetries (e)	Don't  Do  Design the mating features for easy insertion (f)

chamfers and fillets to make them tolerant to any variation in the assembly process. Such features will guide the part into its mating condition during insertion and will overcome manufacturing process variations.

Joining and Fastening Guidelines. The final category of DFA guidelines is joining and fastening guidelines (Table 5). Joining or fastening is typically the last stage in an assembly process. Once a part has been inserted into an assembly and positioned in the required location, it must be fixed to the remainder of the assembly. This process requires access for tools or sister components so that the part can be secured. The guidelines provided in this section identify steps a designer or design team can take to facilitate the joining and fastening stage in the overall assembly process in an efficient and cost-effective manner.

The first guideline suggests the removal or reduction of as many fasteners as possible (without compromising the structural integrity of the assembled part). Fasteners such as screws, nuts, and bolts require a certain time to be properly assembled, and if they can be replaced with quick-insert-type joints, quick connects, or architecture alignment and securing forces, then time savings can be reached, as well as a reduction in overall part count.

The second guideline recommends that components that will not be frequently disassembled can be potentially be joined via a permanent or semipermanent joining method, such as welding, soldering, adhesives, riveting, and so on. These joints are typically less expensive to install, but the cost of removal is much higher (potentially impacting disassembly, recycling, repair, refurbishing, and reuse).

The third guideline suggests providing adequate spacing between fasteners so that there is adequate space for a tool to access the fastener and the material integrity is not compromised by leaving inadequate material between joints.

The fourth guideline suggests that fasteners should not be fastened against angled surfaces. If angled parts are to be joined, then the designer or design team should design

parts that have machined “flats,” where, for example, a bolt head can sit and secure the assembled parts.

The fifth and sixth guideline are closely related. The fifth guideline requires that fasteners be placed in positions where they can be easily accessed for removal, without any special tools. The sixth guideline recommends that if fasteners are to be placed in pockets or channels, they must be wide enough to provide access to tools.

Design for Manufacturing Guidelines

DFM guidelines, much like their DFA counterparts, provide heuristics and rules for improving the manufacturability of components for products and systems. Tables 6–10 provide representative and key guidelines for DFM. These guidelines are dependent on particular manufacturing processes. The guidelines shown in the following tables cover typical and highly applied manufacturing processes in industry. Specific manufacturing references are available for specialized manufacturing processes or other manufacturing processes beyond those represented by Tables 6–10.

DFMA Force Flow Analysis

Building upon the DFA and DFM guidelines, the next basic technique for DFMA is known as *force flow analysis* (or more generally as *effort flow analysis*). Force flow analysis encourages evolution in components of products or systems from the mechanical energy domain by identifying component combination opportunities that are achievable using rigid body and/or compliant mechanisms [20,27–29]. Force flow analysis uses a force flow diagram to represent the transfer of force/effort through product or system components. The force flow diagram is a semantic network composed of nodes and links that are described using the fundamentals of graph theory [30,31]. The nodes of the diagram represent the components of the product or system, while the links represent the interfaces between the components.

Whenever possible, force flow diagrams are laid out in such a manner that the general topology of the product is depicted in

Table 5. DFA Joining and Fastening Guidelines—Illustrations

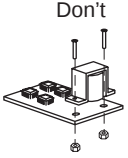
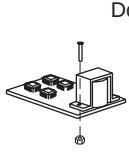
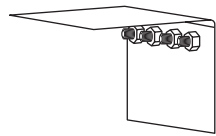
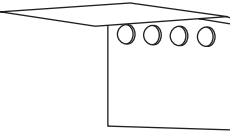
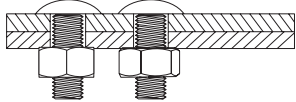
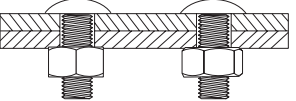
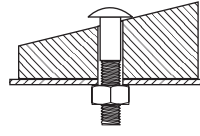
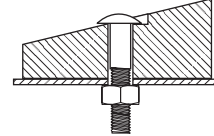
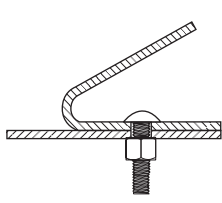
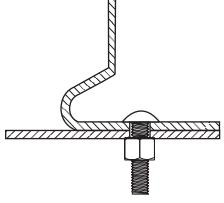
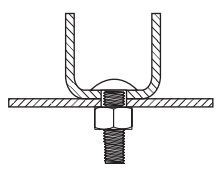
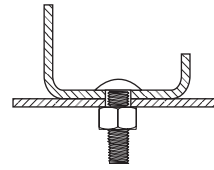
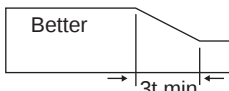
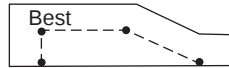
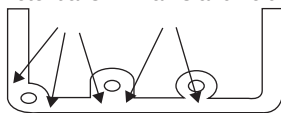
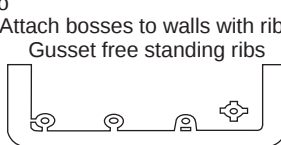
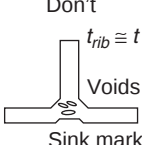
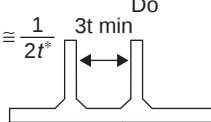
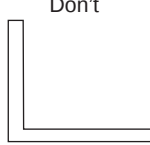
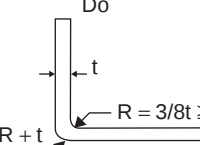
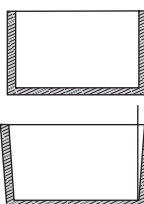
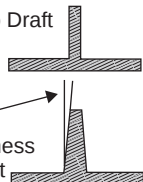
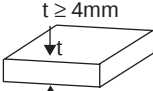
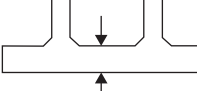
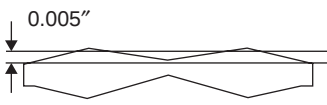
Don't  Do  (a) Eliminate fasteners	Don't  Do  Nut and screw replaced by rivets Use permanent or semi permanent fastening methods (such as welding, soldering, rivets) for joints that will require minimal disassembly (b)
Don't  Do  Proper spacing insures allowance for a fastening tool (c)	Don't  Do  Providing flats for uniform fastening and fastening ease (d)
Don't  Do  Place fasteners away from obstructions (e)	Don't  Do  Deep channels should be sufficiently wide to provide access to fastening tools. No channel is best (f)

Table 6. DFM Guidelines—Mold Processes

Don't  Stepped thickness transition Do  Better  Best Make all transitions smooth and avoid changes in thickness when possible (a)	Don't  Potential sink marks and voids Do  Attach bosses to walls with ribs Gusset free standing ribs (b) Keep section thickness uniform around bosses															
Don't  Voids Sink marks Do  $t_{rib} \equiv \frac{1}{2t}$ 3t min Keep rib thickness less than 60% of the part thickness to prevent voids and sinks (c)	Don't  Sharp corner Do  $R = \frac{3}{8}t \geq 0.06$ Avoid sharp corners, they produce stress concentrations and obstruct material flow (d)															
Don't  No Draft Do  2° min Always provide a draft angle for easier mold removal (e)	Don't  $t < 4\text{mm}$ Do  Use ribs instead $0.065'' \leq t \leq 0.5''$ Minimize section thickness, cooling time is proportional to the square of the thickness of the parts(s), and reducing the cooling time directly reduces costs (f)															
General Tolerances (mm) <table><tr><th>From</th><th>To</th><th>±</th></tr><tr><td>0</td><td>25</td><td>±0.5</td></tr><tr><td>25</td><td>125</td><td>±0.8</td></tr><tr><td>125</td><td>300</td><td>±1.0</td></tr><tr><td>300</td><td></td><td>±1.5</td></tr></table> Use standard dimensional variation capability: Do no tolerance (g)	From	To	±	0	25	±0.5	25	125	±0.8	125	300	±1.0	300		±1.5	 0.005" Use standard thickness variation capability; do not over tolerance (h)
From	To	±														
0	25	±0.5														
25	125	±0.8														
125	300	±1.0														
300		±1.5														

the topology of the diagram. As an example, a force flow diagram for a Staple Remover product (Fig. 6) is shown in Fig. 7. Notice in the figure that finger forces enter, for the operation of removing a staple from a set of papers, the top and bottom pad and flow

through the components of the product to work on the staple (work piece) and to store energy in the torsion spring for returning the stapler remover jaws after the operation is complete. The main benefit of modeling a product using a force flow diagram is the

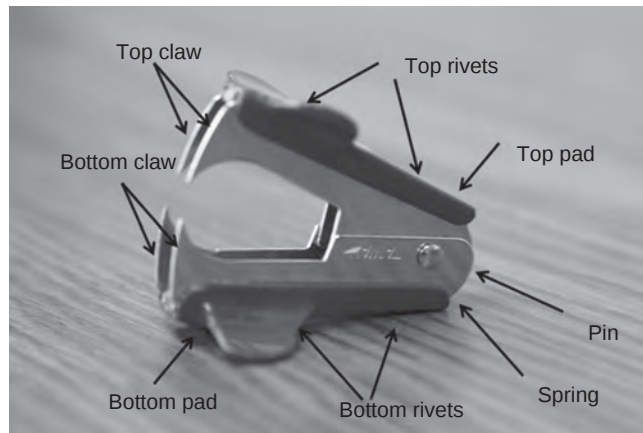


Figure 6. Staple remover product, 10 components/parts.

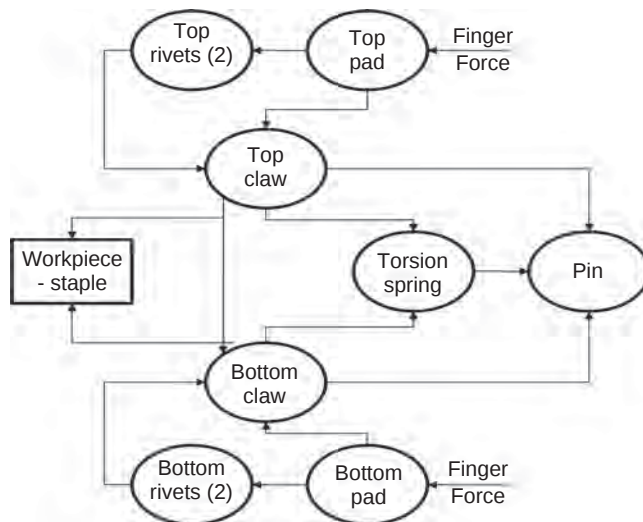


Figure 7. Force flow diagram of staple remover product (input finger forces).

identification of possible component combination opportunities. These opportunities are made more obvious when the relative motion at the interfaces between connected components is characterized. Characterizing the relative motion entails labeling the links between components. The links model the interaction between components, and are the conduit for force flow (force or torque transfer in the mechanical domain). Labels are added to the links to identify relative motion characteristics across the link. The labels on links between connected components are defined as follows:

“N”: no relative motion between components.

“R”: relative motion at the interface and between other component regions.

Figure 8 shows the labeled links for the force flow diagram of Fig. 7. The “R” label is used to denote an interface where “relative motion occurs both between the extents of connected components as well as at the physical interface between the components (i.e., the nodes).” Once characterization of relative motion and link labeling is completed, the following guideline applies:

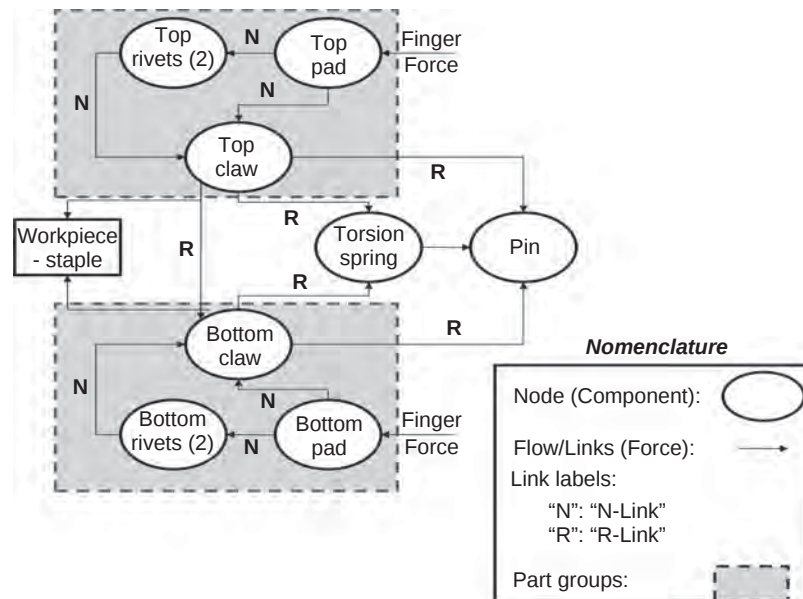


Figure 8. Force flow diagram of staple remover product with link labels and part groups.

Force flow guideline: Groups of components where no relative motion exists (no R-Links connecting members within the group) are candidates for combination if not prohibited by material or assembly/disassembly issues.

Combination between components or component groups that are connected by R-Links may be possible, but will require more complex redesign. This guideline suggests that parts that do not experience relative motion are direct candidates for combination and that parts that do experience relative motion may still be candidates for combination but will require compliance or other creative steps to obtain equivalent functionality.

The next step in constructing an effort flow diagram is to identify groups of components connected by non-relative-motion links (N-Links). These groups of components are the starting point for further investigation of component combination. They suggest direct combinations through the force flow analysis, providing opportunities that may not have been intuitive or otherwise considered by a designer or design team. Figure 8 shows the groupings of components that may be

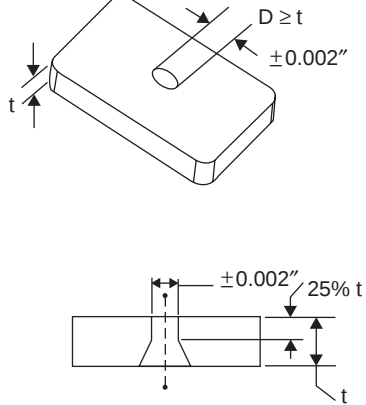
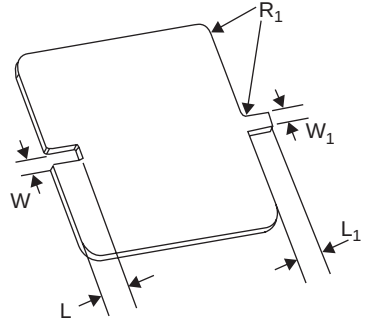
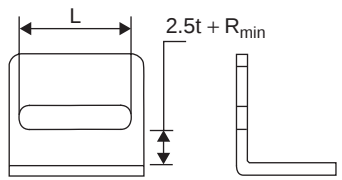
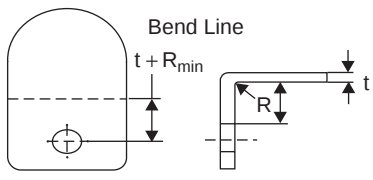
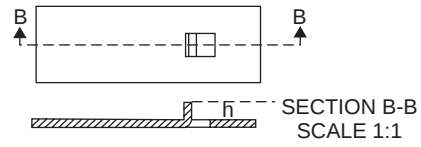
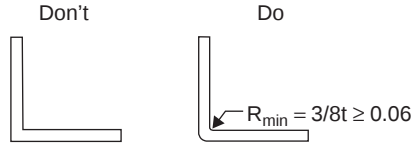
considered directly. In particular, due to the symmetry of the product, the pads, rivets, and jaws on the top and bottom of the product, these parts may be considered for combination, such as through molding, sheet metal, or casting manufacturing processes.

Summary: Force Flow Analysis Process.

Based on this description of force flow analysis and force flow diagrams, the DFMA process for this technique maybe summarized as follows:

1. Choose a concept or existing product/system to model.
2. Identify the operation(s) of the concept, product, or system to model.
3. Identify the input or internal forces and torques for the chosen operation(s).
4. Create a force flow diagram of the concept, product, or system, starting with the input and internal forces and torques, and following these forces and torques through until they terminate and/or create reaction forces.

Table 7. DFM Guidelines—Sheet Metal Forming Processes

 (a)	 (b)
 (c) Position openings away from bends	 (d) Position holes away from bends
 (e) Shear and form operations should have a minimum height (h) of 2½ the blank thickness	 (f) Avoid sharp corners, or the material will tear

- Assess each flow between components, and label the flow links with “N” or “R.”
- Identify all groups of components connected by non-relative-motion links (N-Links) with no intermediate R-Links.
- Create concepts to combine the components in each component group while maintaining the functionality of the groups. Consider alternative manufacturing processes for the new component group concepts.
- (Optional) Consider component combinations across R-Links. For example, where relative motion exists between components (but not at the interfaces), consider compliant structures to achieve the component functionality while combining components. Examples of this situation are when

Table 8. DFM Guidelines—Sheet Metal Forming Processes II

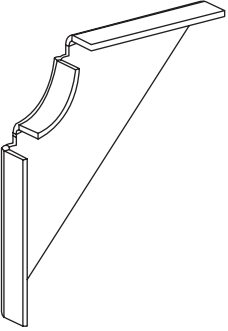
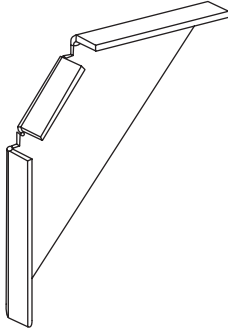
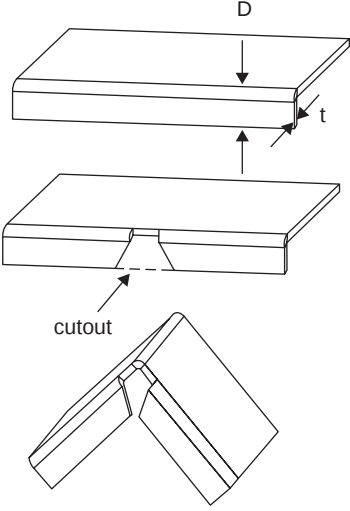
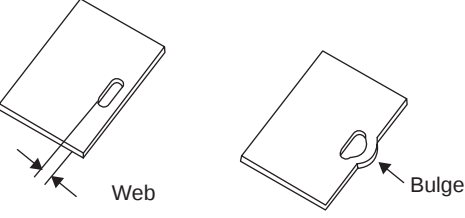
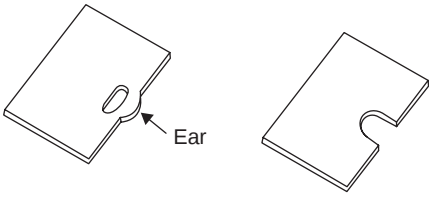
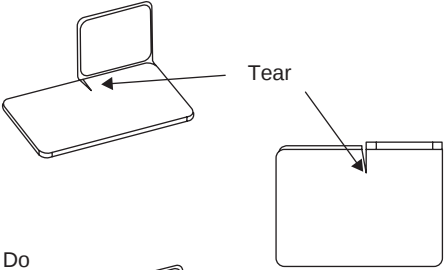
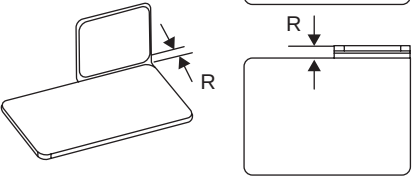
Don't  Do  Use separated straight flanges where possible (a)	 If $D \geq 2t$, A cutout is needed to bend flange (b)
Don't  Do  A narrow web will cause bulging. Provide an ear in the blank or include the hole as a notch (c)	Don't  Do  Offset bends (d)

Table 9. DFM Guidelines—Casting Processes

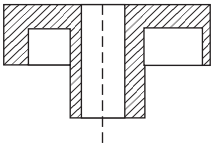
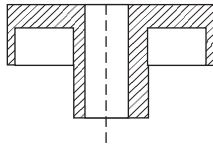
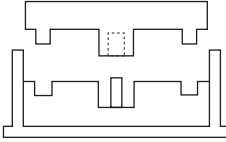
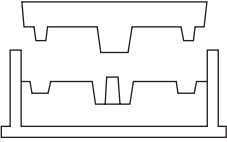
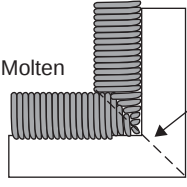
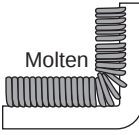
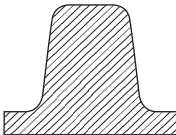
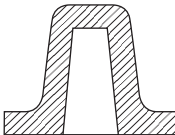
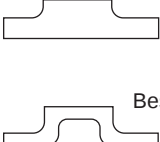
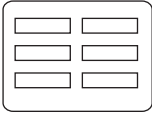
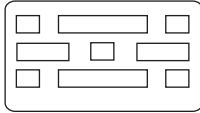
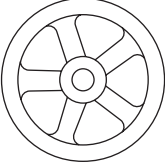
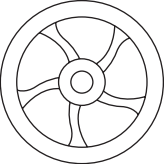
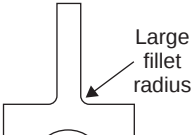
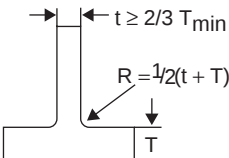
Don't  Do  (a) Maintain section thickness uniform	Don't  Do  (b) Provide a draft angle to help removal from mold
Don't  Do  (c) Avoid sharp corners	Don't  Do  (d) Design bosses with uniform thickness
Don't  Do  (e) Avoid abrupt changes in section thickness	Don't  Do  (f) Stagger ribs to prevent hot spots
Don't  Do  (g) Restraining a part of the casting produces tensile stresses that can result in hot tears	Don't  Do  (h) Non-uniform section thickness caused hot spots that causes shrinkage defects. Design T-junctions to prevent hot spots

Table 10. DFM Guidelines—Machining Processes

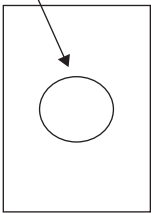
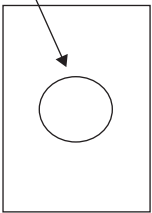
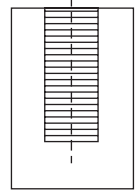
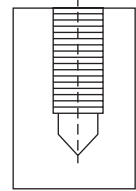
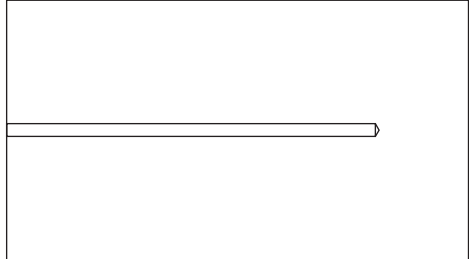
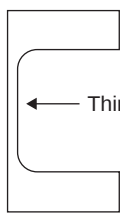
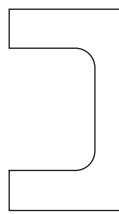
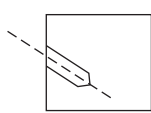
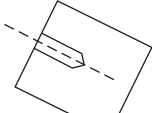
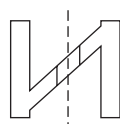
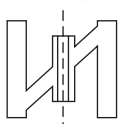
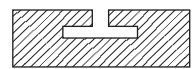
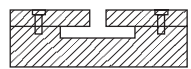
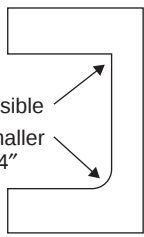
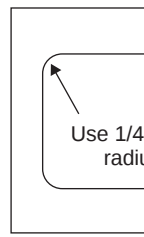
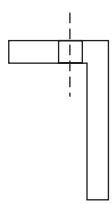
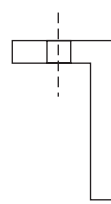
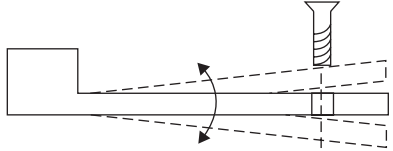
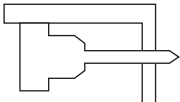
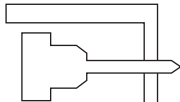
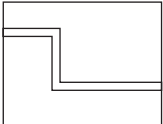
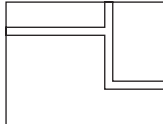
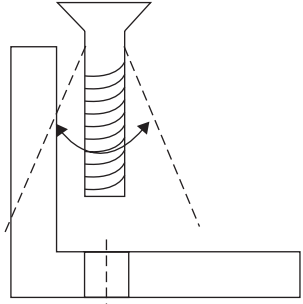
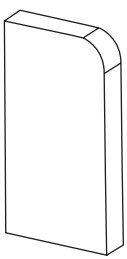
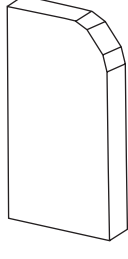
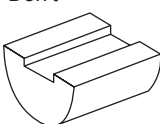
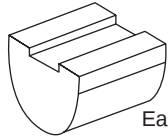
Don't $D = 0.627''$  Do $D = 0.625''$  (a) Use standard dimensions	Don't  Do  Design holes to the shape of tool. If hole is to be tapped, provide space for it (b)
 (c) Avoid long narrow holes	Don't  Thin wall Do  Avoid thin walls that break when machining (d)
Don't  Do  Don't  Do  (e) Avoid drilling inclined faces	Don't  Do  Don't  Do  Do not design impossible to machine hollows or overhangs (f)
Don't  Impossible Radius smaller than 1/4" Do  Use 1/4–1/2" radius Design for reasonable internal pocket radii (g)	Don't  Do  (h) Place holes away from corners and edges

Table 10. (Continued)

 (i) Design for reasonable internal pocket radii	Don't  Do  (j) Provide access for tools
Don't  Do  (k) Holes can't change direction	 (l) Deep pockets also cause vibration of the tool
Don't  Do  (m) Avoid outside rounds, which are difficult unless CNC-machined	Don't  Difficult to fixture Do  Easier to hold (n) Design parts that are easy to fixture

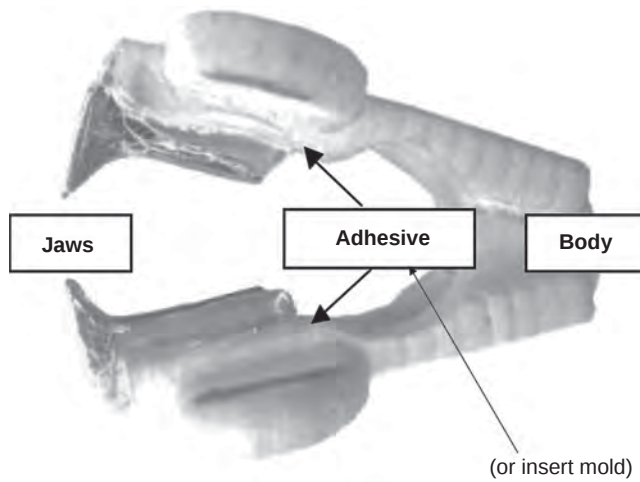


Figure 9. Evolved stapler remover with three parts (insert mold or adhesive jaws w/compliant joint).

springs are used, such as the torsion spring in the stapler remover product (Figs 7 and 8).

and a compliant (living) hinge is created at the right region of the product to replace the torsion spring and pin.

Figures 6–9 show the application of the force flow analysis to the stapler remover product. Figure 9 shows a developed concept which reduces the original 10 components of the product to three components. An insert mold is used to create the pad-jaw structures,

PRODUCT APPLICATION: CORDLESS SCREWDRIVERS

To demonstrate the utility of DFMA techniques discussed in this article, an industrial

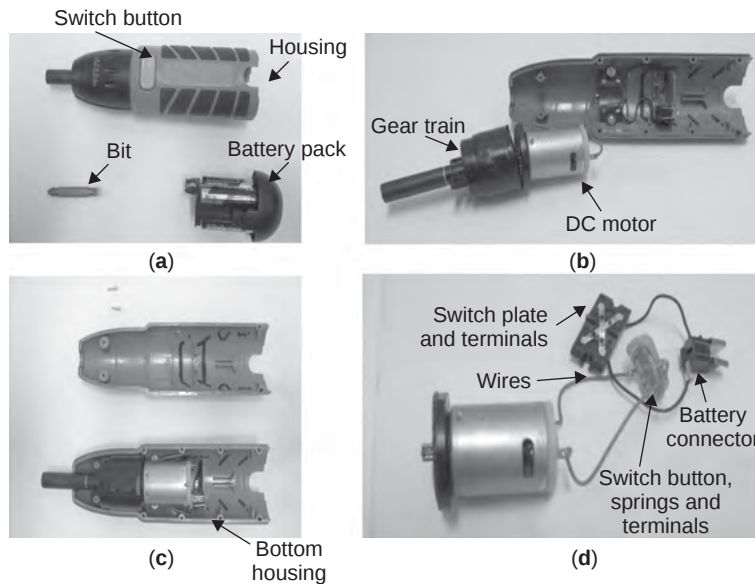


Figure 10. Industrial cordless screwdriver product, illustrating different stages of disassembly.

product, a cordless screwdriver, is selected and analyzed. The product operates with four AA batteries and accepts screwdriver bits. Figure 10 shows the disassembly of the product. Figure 10a shows the product with its AA battery pack and screwdriver bit removed. In Fig. 10b the housing is opened up by the removal of two screws and the internal components of the driver are visible. The key components easily identified here are the DC motor and the necessary planetary gear train which reduces the rotational speed and increases the torque. Figure 10c shows additional components, which are shown in greater detail in Fig. 10d. These components include a battery connector, the direction switch, and a battery terminal plate.

Subtract-and-Operate Procedure (SOP). A bill-of-materials (BOM) for the cordless screwdriver product is shown in Table 11. A subtract-and-operate (SOP) procedure was applied to the product, and the effect of removal of each part is identified in the table. This process simply considers the systematical removal and replacement of a component of the product to deduce the functionality of each component and its affect of the product's mechanical degrees of freedom [20,28]. As shown in the table, there are no obviously redundant parts. The product is well designed at this level, and it is apparent that the designers kept manufacturing and assembly in mind while designing the product, especially in terms of part count (such as with fasteners).

Force Flow Analysis. Force flow diagrams represent an invaluable tool when a designer or a design team seeks to identify opportunities for part assimilation and the overall simplification of the design. A force flow analysis was performed on the product and the results are shown in Fig. 11.

Three groups separated by R-links are identified. By analyzing Group 1, one can see that there are multiple energy/effort flows, electrical energy as well as mechanical energy. The battery plugs into a battery connector, and current is conducted via terminals into wires (see BOM, Table 11) and into the switch plate. When the direction

switch button of this product is pressed, the DC motor is energized. Potentially, these components, as shown in Fig. 11, could be combined. In the current product, Group 1 contains 17 parts.

Group 2 contains the DC motor and the motor spur (pinion) gear. These components are already a basic module without fasteners, and they can typically ship together as one unit from an original equipment manufacturer.

Group 3 includes three parts: the shaft coming from the gear train, the chuck that accepts bits, and the bit itself. Potentially these three components could be combined, but then the product would be limited to one single bit, and not accept other bits for different screw heads. This combination would limit functionality and would be infeasible (except for perhaps the subcombination of the shaft and chuck).

Of the three components groups and redesign avenues explored through force flow analysis, the first group appears to be the most promising. A number of concepts may be developed for this component group at this stage. However, we will expedite and systematically explore concepts by using DFMA guidelines.

Application of DFMA Guidelines

Device Function. The components identified for redesign are shown in Fig. 12. The first region of the figure shows how the parts are assembled in the current product, and the bottom region of the figure shows the completed assembly (excluding other components not under study).

On the basis of Fig. 12, the way the system works is that the battery pack plugs in from the back, and upon insertion into the product, the battery pack terminals make contact with a battery connector jack (as shown in Fig. 13c). The metal battery connector terminals have wires soldered onto them that are connected to the switch plate. The switch plate's function is to reverse the direction of current flow, depending on which side of the switch the user presses. When the direction switch button is pressed, metal switch terminals make contact with the corresponding switch plate terminals on the switch plate,

Table 11. Bill-of-Materials (BOM) and Effect Table—Cordless Screwdriver Product

	Module/ Part #	Qty	Mfg. Process	Material	Effect of Removal
1	Battery holder	1	Injection molded	Plastic	Batteries are not held in place against the terminals.
2	Series terminal clip	2	Stamped and bent	Steel	Does not operate. Batteries not connected.
3	Battery connector	1	Injection molded	Plastic	Doesn't operate. Plug terminals are not pressed against battery clip.
4	Battery connector terminals	2	Stamped and bent	Steel	Does not operate. Switch and motor are not connected to batteries.
5	Series terminal plate	1	Stamped and bent	Steel	Does not operate. Batteries not connected.
6	AA Battery	4	Standard part	Various	Does not operate. No source of electrical energy.
7	Bit	1	Machined	Steel	Cannot transmit torque to bit.
8	Wire, Switch to Motor	2	Standard part	Various	Electrical energy does not reach motor.
9	Direction Switch Button	1	Injection molded	Plastic	User cannot input control to product. Electrical energy does not reach motor.
10	Switch plate	1	Injection molded	Plastic	Switch is not held in place. Electrical energy does not reach motor.
11	Switch springs	2	Twisted	Metal	Switch does not operate
12	Switch terminals	2	Stamped and bent	Metal	Electrical energy does not reach motor.
13	Switch plate terminals	2	Stamped and bent	Metal	Electrical energy does not reach motor.
14	Wire, Battery terminal to switch	2	Standard part	Various	Electrical energy does not reach motor.
15	Battery Clip	1	Standard part	Various	Does not operate. Batteries are not connected.
16	Bottom Housing	1	Injection molded	Plastic	Product can operate, but battery pack will fall out.
17	Back Housing	1	Injection molded	Plastic	Parts fall out of place
18	Front Housing	1	Injection molded	Plastic	Parts fall out of place
19	Screws	2	Standard part	Metal	Housing comes apart and parts fall out.
20	Pin	2	Standard part	Metal	Operates, lower parts of housing will not stay aligned.
21	Motor	1	Standard part	Various	Electrical energy is not converted to torque
22	Screws, motor plate	1	Standard part	Metal	Motor is free to spin within product housing

(continued overleaf)

Table 11. (Continued)

	Module/ Part #	Qty	Mfg. Process	Material	Effect of Removal
23	Motor Plate	1	Injection molded	Plastic	Gears fall out of transmission. Motor stator spins within product.
24	Planetary Gears, set1	3	Molded	Plastic	Product does not operate. Torque is not transmitted.
25	Planetary Carrier and sun gear	1	Molded	Plastic	Product does not operate. Torque is not transmitted.
26	Planetary Gears, set2	3	Molded	Plastic	Product does not operate. Torque is not transmitted.
27	Brake Coupler	1	Molded	Plastic	Torque is not transmitted to shaft
28	Ring Gear	1	Molded	Plastic	Transmission gears are not held in place.
29	Coupler	1	Machined	Metal	Torque is not transmitted to shaft
30	Brake Drum	1	Machined	Metal	Product operates but cannot be used manually.
31	Brake Roller	2	Machined	Steel	Product operates but cannot be used manually.
32	Shaft	1	Machined	Steel	Torque is not transmitted to chuck
33	Spring Clip	1	Standard part	Metal	Operates but shaft may pull out of transmission.
34	Washer	1	Standard part	Steel	Minimal effect
35	Chuck	1	Machined	Metal	Cannot be used with standard hex bits.
Total	52				

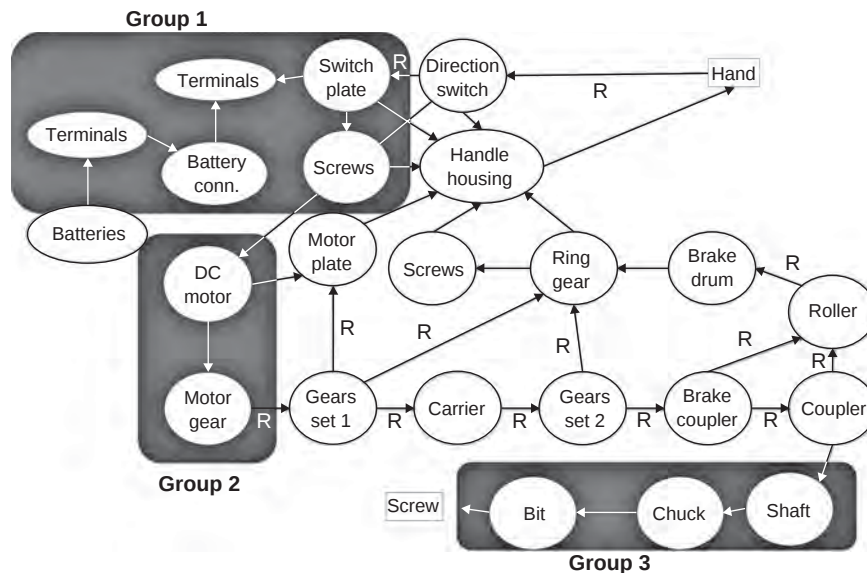


Figure 11. Force flow analysis of a cordless screwdriver product.

and the circuit is completed. Current is then carried through wires, which are connected to the switch at one end and to the DC motor at the other end. Figure 13 shows the three modules under discussion. The direction switch button has five parts. The switch plate, Fig. 13b, has six parts, and the battery connector is made up of three parts. Besides these modules and parts, four wires are also used to connect between these parts and the motor.

DFA Guidelines. DFA guidelines were employed in the redesign on the subassemblies shown in Fig. 13. The first guideline employed, as shown in Table 12, is a system guideline that suggests that multiple parts be incorporated into a single subassembly, and the second suggests that part count be reduced by incorporating multiple functions in a single part. With these suggestions in mind, the following concept (Fig. 14) was conceived.

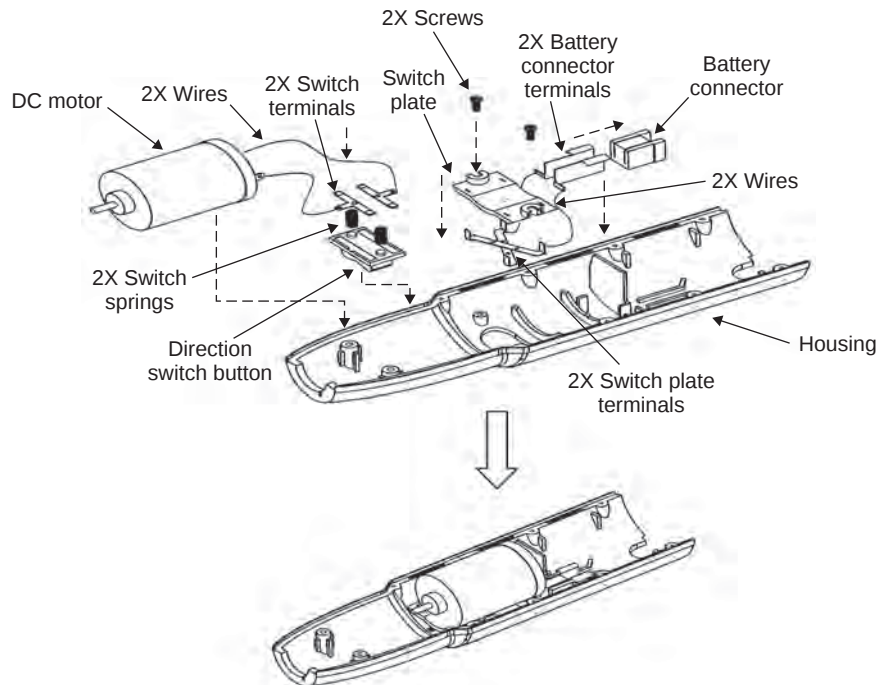


Figure 12. Cordless screwdriver product, isolating components for redesign.

Table 12. DFA Guidelines Applied to the Cordless Screwdriver Product

1	Modularize multiple parts into single subassemblies
2	Reduce part count by incorporating multiple functions into single part
3	Eliminate tangly parts
4	Provide features so that parts can be gripped properly
5	Provide alignment features
6	Design the mating features for easy insertion
7	Eliminate fasteners
8	Use permanent or semi permanent fastening methods (such as welding, soldering, shrink fits) for joints that will require minimal disassembly

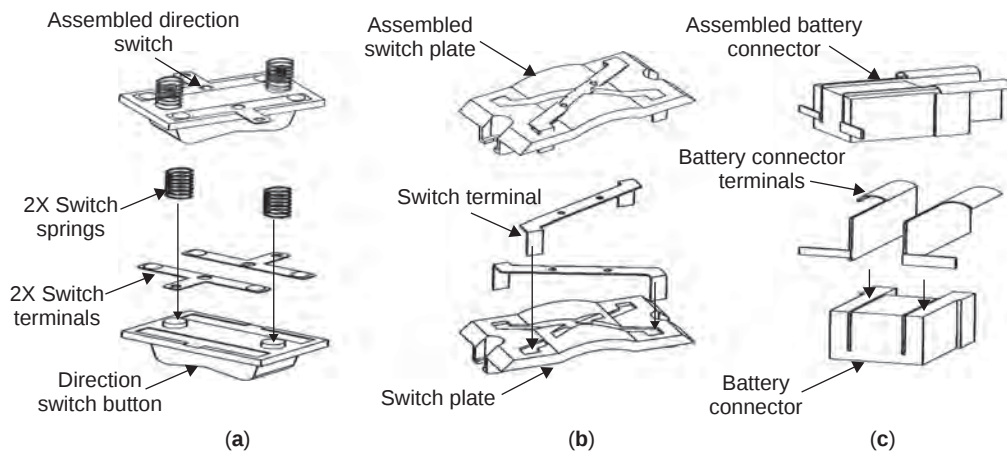


Figure 13. Cordless screwdriver product, showing particular components from Fig. 12.

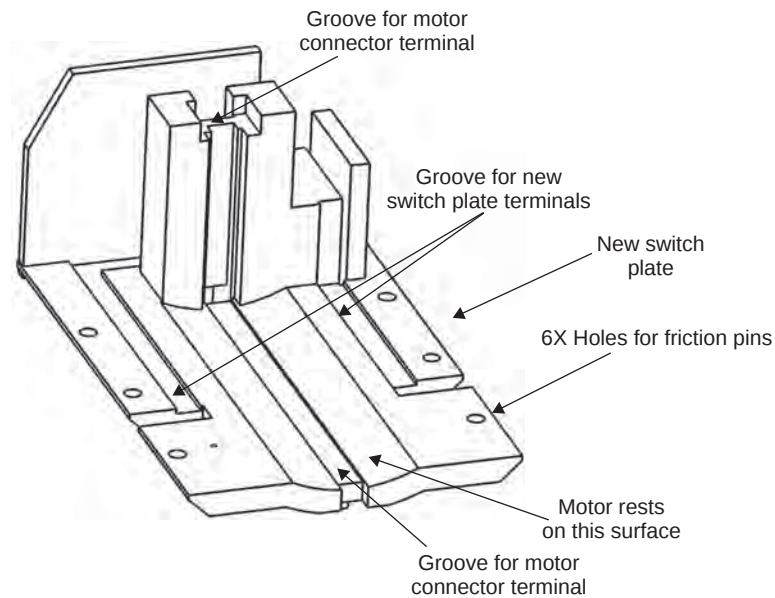


Figure 14. New component concept for cordless screwdriver product.

This new part concept combines the functionality of the switch plate and the battery connector jack from the original cordless screwdriver (Fig. 13b and c) into one part.

Using the third and fourth guidelines in Table 12, the springs in the switch assembly (Fig. 13a) were eliminated. The springs in the original product return the switch button

back to its original configuration when the user is finished pressing it. This function was incorporated into another multifunction part shown in Fig. 15. These metal tabs are bent so that they act as a spring (flat or leaf spring), simultaneously act as part of the direction switching circuit, and act as connector terminals to the motor (making obsolete the need for wires).

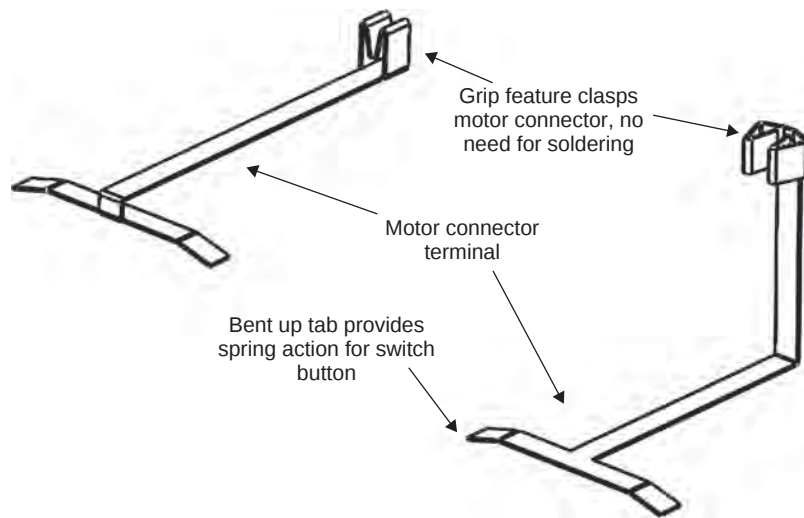


Figure 15. Additional new component concepts for cordless screwdriver product.

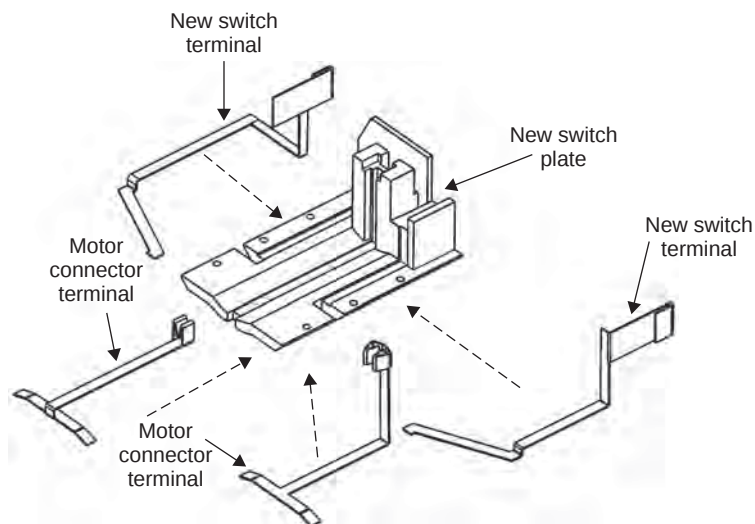


Figure 16. Overall component concepts for cordless screwdriver product.

Alignment features were provided as suggested in guideline 5 from Table 12 by providing grooves where these metal tabs will connect and be joined using permanent fastening methods as suggested in guideline 8 from the same table. The fastening method chosen for this application is a heat-staking process. The overall design, integrating the new component concepts, is shown in Fig. 16.

Grooves built into the design provides features for easy insertion, and two screws were eliminated by using friction pins as shown in Fig. 17. The two other tabs act as the battery connector jack terminals as well as the direction switching circuit. This arrangement eliminates the need for wires and soldering. The total part count of the assembly has been reduced from 17 to 5 by this design.

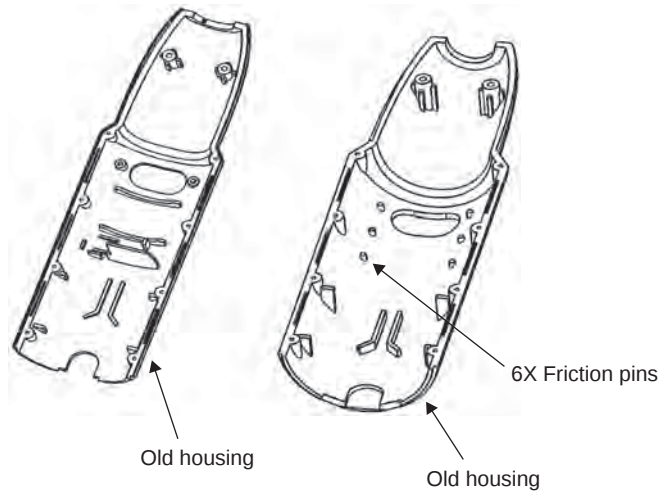


Figure 17. New housing concepts based on DFM guidelines as applied to cordless screwdriver product.

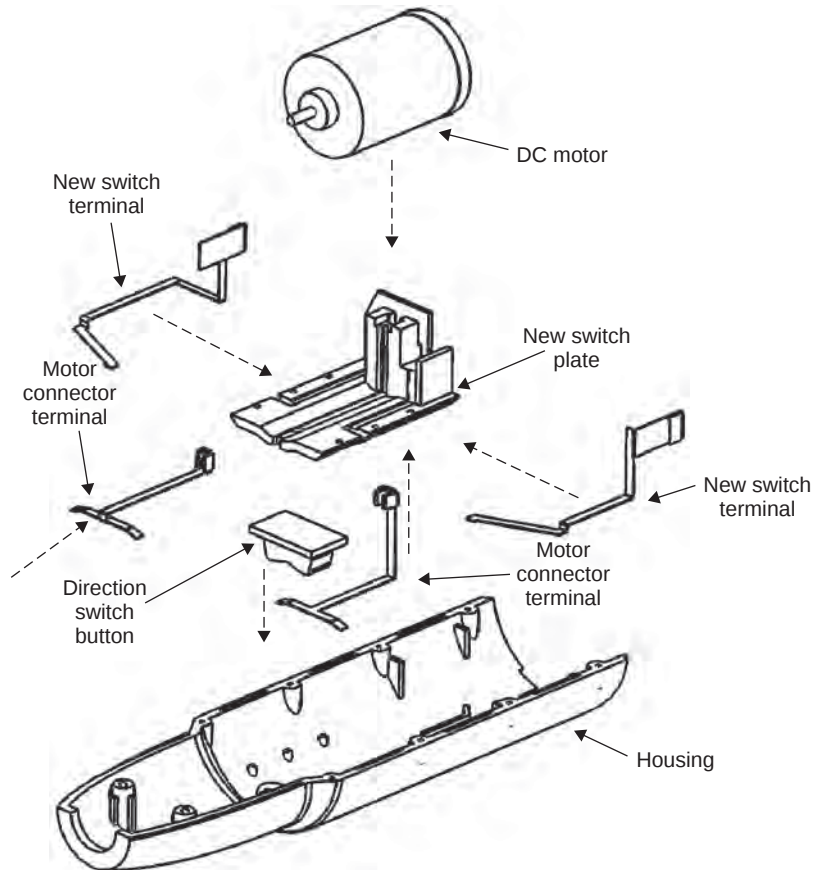


Figure 18. Overall design of new components for cordless screwdriver product based on the application of DFMA techniques.

DFM Guidelines. DFM guidelines for injection-molded parts were also employed when desinging the new component concepts. Draft angles were provided for easier mold removal, and sharp corners were minimized and free-standing ribs were used, where possible. The new design also utilized the removal of extraneous ribs, as shown in Fig. 17. The overall assembly of the new design is shown in Fig. 18.

CONCLUSIONS AND GOLDEN NUGGETS

In this article, we explored the contemporary business and engineering topic known as *DFMA*. At an enterprise level, *DFMA* is pervasive and reaches across organizational structures. In our exploration here, we have scratched the proverbial “tip of the iceberg” of *DFMA*; however, a set of guidelines and techniques have been constructed for effectively transforming design concepts and existing products and systems to efficient designs that have great potential for competitive advantage through cost savings, increased reliability, and innovative component structures. The basic techniques presented in this chapter do not replace the designer in creating efficient designs in terms of manufacturability and assemblability: instead, novel insights are provided that enable designers and design teams to identify new and innovative component concepts that would possibly not be explored through *ad hoc* approaches.

Acknowledgments

The authors would like to acknowledge support in the development of this article as provided by an MURI grant by the Office of Naval Research (ONR) and the Cullen Endowed Professorship in Engineering, The University of Texas at Austin. Any opinions, findings, or conclusions found in this article are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

1. Boothroyd G, Dewhurst P. Product design for assembly. New York: McGraw-Hill; 1989.
2. Pugh S. Total design. New York: Addison-Wesley; 1991.
3. Huang G. Design for X. London: Chapman & Hall; 1996.
4. Herrmann J, *et al.* New directions in design for manufacturing. DETC'04; 2004 Sep 28–Oct 2; Salt Lake City (UT): ASME; 2004.
5. Tatikonda MV. Design for assembly: a critical methodology for product reengineering and new product development. *Prod Invent Manage J* 1994;35(1):31–38.
6. Andreasen MM. Design for assembly. UK: IFS LTD; 1983.
7. Boothroyd G, Dewhurst P, Knight W. Product design for manufacture and assembly. 2nd ed. New York: CRC; 2002.
8. Boothroyd G, Poli C, Murck L. Automatic assembly. New York: Marcel Dekker; 1982.
9. Ashley S. Cutting costs and time with *DFMA*. *Mech Eng* 1995 (March);3:74–77.
10. Chow WW. Cost reduction in product design. New York: Van Nostrand Reinhold Company; 1978.
11. Defazio TL, Whitney DE. Simplified generation of mechanical assembly sequences. *IEEE J Robot Autom* 1988;RA-3(6):640–656.
12. Dieter GE. Engineering design: a materials and processing approach. New York: McGraw Hill; 1991.
13. Ishii K, Juengel C, Eubanks CF. Design for product variety: key to product line structuring. *Proceedings of ASME DETC*; Boston (MA); 1995.
14. Kim GJ, Bekey GA. Combining assembly planning with redesign: an approach for more effective DFA. *Proceedings of IEEE Conference on Robotics and Automations*; Atlanta (GA); 1993. p. 319.
15. Lee S, Yi C. Force-based reasoning in assembly planning. Volume DE-73, *ASME flexible assembly systems*. 1994. pp. 165–172.
16. Pine JB, Victor B, Boynton AC. Making mass customization work. *Harv Bus Rev* 1993;9:108–119.
17. Poli C, Graves R, Groppetti R. Rating products for ease of assembly. *Mach Des* 1986;8:79–83.
18. Kalpakijia S. Manufacturing engineering and technology. New York: Addison-Wesley Publishing Co.; 1992.
19. Ettlie JE, Bridges WP. Methods that work for integrating design and manufacturing. *Managing the design-manufacturing process*. New York: McGraw-Hill Book Company; 1990.

20. Otto K, Wood KL. Product design: techniques in reverse engineering and new product development. New Jersey: Prentice Hall; 2001.
21. Ullman D. The mechanical design process. New York: McGraw-Hill Science/Engineering/Math; 1992.
22. Ulrich KT, Eppinger SD. Product design and development. New York: McGraw-Hill; 1994.
23. Branan B. Six sigma quality and DFA. DFMA Insight 1991;2:1–3.
24. Kusiak A. Engineering design: products, processes, and systems. Orlando (FL): Academic Press, Inc.; 1999.
25. Pahl G, Beitz W, Wallace KM, Blessing L, Bauert F. In: Wallace KM, Blessing L, Bauert F. Engineering design: a systematic approach. 2nd enlarged and updated ed. London: Springer; 1977.
26. Ullman D. The mechanical design process. 2nd ed. New York: McGraw-Hill Science/Engineering/Math; 2002.
27. Greer J, Jensen D, Wood K. Effort flow analysis: a methodology for directed product evolution. Des Stud 2004;25(2):193–214.
28. Lefever D, Wood KL. A reverse engineering and redesign methodology. Res Eng Des 1998;10(4):226–243.
29. Otto K, Wood KL. A reverse engineering and redesign methodology. Res Eng Des 1998;10(4):226–243.
30. Harary F. Graph theory. Reading (MA): Addison-Wesley; 1969.
31. Tsai L. Mechanism design: enumeration of kinematic structures according to function. Boca Raton (FL); CRC; 2000.

DESIGN FOR NETWORK RESILIENCY

MICHAEL TORTORELLA
Rutgers Center for Operations
Research, Rutgers University,
Piscataway, New Jersey

FOREWORD

Rationale

In the beginning there was the social network. From the earliest days of *Australopithecus* and *homo erectus*, individuals formed complex webs of social interactions that we today conceptualize as networks. It is no exaggeration to say that these networks enabled the species to survive and evolve into today's humans who are capable of such conceptualization. Until roughly the Renaissance, mostly that is all there was, when the advent of sailing ships and intrepid sailors enabled goods to be moved greater distances. This, and further advances during the nineteenth century, begat the supply chain network, although, again, no one used that name at the time. The invention of the telephone in 1876 begat the telecommunications network. The postal system, the electric power grid, and commodity networks such as oil and gas transport systems brought forth a new way of life. Indeed, civilization as we know it would be impossible without networks.

In the twentieth century, mathematicians began to study networks as abstract objects to do a better job of creating and improving them. Our position in this paper is that networks are not ends in themselves; they exist to perform certain functions. Society is interested in the continued performance of these functions at an adequate degree and at reasonable cost. The concept of *resiliency* is introduced to enable discussion of the degree to which the functions performed by a network continue to be performed when various disruptions occur. The purposes of this article

are to describe a mathematical framework for discussing network resiliency in quantitative terms and to list principles and practices that can be employed during the design of a network to promote its greater resiliency.

Scope

The material presented in this article is generic in the sense that it applies to all types of networks and is not restricted to a particular application. We consider networks that transport continuum materials such as oil and gas, and networks that transport discrete physical objects such as letters and packages as well as notional objects such as data packets in telecommunications. Social networks are also included in that they exist to share something among the individuals comprising the nodes of the network: information, or status, or power relationships.

We reserve for future consideration further development of an alternate property characterizing resilient networks, namely that they return quickly to a normal functioning state after a disruption. Again, a matter of degree is involved, and it will be necessary to develop quantitative models to treat this phenomenon adequately.

Background

The etymology of "resiliency" is the Latin *resilio*, which means "to leap or spring back." The informal use of the word stems from this origin. "Network resiliency" began to be used as a phrase because people wanted a way to express how well a network may continue to perform its intended functions when the inevitable degradation of operating conditions occurs, or how quickly it returns to normal operation after a disruption. Infrastructure degradation includes failures and/or loss of capacity of the network's constituent elements (switches, routers, valves, transport systems, etc.) that may be due to reliability problems (hardware, software, operations) in the physical equipment (pipes, wires, towers, circuit cards, trucks, buildings, etc.) mak-

ing up the network elements. These failures, in turn, may be due to “random” manifestation of broken or deteriorated components, or due to natural disasters such as earthquakes and fires, or due to deliberate attacks. The study of the effects of such failures on the network as a whole gives rise to the mathematical theory of network reliability [1,2], which usually considers figures of merit like two-terminal connectivity and all-terminal connectivity as indicators of the purposes of the network. The literature on the mathematics of network reliability is extensive.

In prior studies, “network resiliency” was used to convey a sense of continued satisfactory fulfillment of some purpose of the network in the face of failures and disruptions and/or the rapid return of the network to normal operation after a disruption. Specific purposes studied have included communication among node-pairs [3], survivability [4], the ability to provide alternate communication paths [5], reachability [6], economic damage [7], and disconnection of the network [8,9]. Occasionally, the term is used without specific definition, assuming that resiliency would be equated with “good” properties implicitly known to the reader, as in Ref. 10. In these cases, some specific measure of how well the network continues to operate, that is, provides its intended service(s), serves as the proximate interpretation of network resiliency. The generalization to delivery functions (see the section titled “Delivery Function”) and delivery importance (see the section titled “Delivery Importance”) is natural.

Resiliency and Reliability Contrasted.

Before proceeding further, some differences between resiliency and reliability are worth examining. The primary difference is that resiliency applies to the functions or services performed by the network, whereas reliability is usually used to describe the failure and repair behavior of network elements. One of the most important *delivery functions* (see the section titled “Delivery Function”) we use in defining network resiliency is the service reliability of the services or applications that run on the network (see Refs 11 and 12 for further clarification on service reliability),

but using service reliability as a delivery function is not the only reasonable approach to network resiliency characterization. In this article, we see that figures of merit and metrics for network resiliency depend on the particular delivery function associated with the network. The examples in the section titled “Three Examples of Delivery Functions” are specializations of the general theory presented in the sections titled “Delivery Function and Delivery Importance” and “Network Resiliency.”

In addition, network resiliency resembles reliability importance [13,14] more than plain reliability. In essence, network resiliency is an attempt to describe how the reliability of services that run on the network changes when certain parts of the network experience reductions in capacity (possibly stemming from equipment failures, natural disasters, deliberate attacks, etc.). Therefore, reliability is an important primitive concept in the construction of network resiliency measures, and such measures incorporate the change in reliability (broadly interpreted) as network conditions change.

Resiliency and Survivability Contrasted.

The related notion of survivability has been studied primarily in the telecommunications context [15]. The earliest approaches to survivability were mainly concerned with the provision of geographically diverse alternate routes so that traffic that would normally be carried on facilities that could be lost due to natural disaster or deliberate attack would continue to reach its destination [16]. While the amount of traffic that is successfully rerouted and carried is a key indicator of network resiliency, survivability studies did not necessarily explicitly incorporate measures of how much traffic would be successfully rerouted. Later studies began to incorporate such considerations [17], and thereby resemble more comprehensive resiliency studies.

Some care is required in interpreting the literature regarding reliability, survivability, and resiliency because these terms are sometimes used interchangeably with no warning.

Resiliency and Vulnerability Contrasted. In a sense, network resiliency is a more complete description of network operation under degraded conditions than network vulnerability. For example, certain network elements may be vulnerable to damage to a greater degree than other network elements, but if the consequences of damage to the former are minor in terms of, say, added congestion or other disruptions of the flow in the network, and the service consequences of this flow disruption are also minor, then the network may still be resilient even though vulnerable. General considerations aside, network vulnerability would have to be defined precisely, and in quantitative terms, in order to make valid comparisons. Any such definition has not yet received wide acceptance.

Synopsis

The approach taken in this article goes beyond that of the usual network reliability theory. To capture the notion that the network exists to perform a desired function, we define a *delivery function* associated with a network that tells what the network is being used for. The precise definition is given in the section titled “Delivery Function,” some examples of delivery functions here serve to provide a flavor of what this means. In the most general case, we may consider a delivery function to be the reliability of a service or application [11,12] to a user in the sense that is used in telecommunications networks: for example, a user pays an internet service provider (ISP) for access to the ISP’s network so that the user may acquire page views from the world wide web, upload files, and so on, and the reliability of those services as seen by the user/purchaser of the service is the delivery function. For a commodity network, such as an oil or gas transport network, the delivery function may be the quantity of commodity delivered to specified destination nodes over a stated period of time. In other cases, the delivery function can be as simple as the two-terminal connectivity in an uncapacitated network. The key idea is that the delivery function captures the essence of what the network is being used for.

Delivery importance of a subset of the network is then defined in terms of the change in the value of the delivery function when conditions in that subset change (e.g., increase or decrease in capacity or removal from service). A derivative comes into play here, and so delivery importance represents a generalization of the idea of reliability importance [13,14] to this broader context. *Network resiliency* is defined as the degree to which the network is sensitive to changes in its infrastructure: a network is resilient if there are “few” subsets that have “high” delivery importance. We then discuss three examples of particular delivery functions in increasing order of complication: connectivity, reliability, and flow in a stochastic network. We give examples of each and discuss computational issues.

Some design principles promoting greater network resiliency are reviewed in the section titled “Design for Network Resiliency Principles and Practices.” The two principles discussed are overprovisioning and network control. A straightforward network resiliency optimization problem is introduced in the section titled “Network Resiliency Optimization,” and the network interdiction problem is discussed in the section titled “Network Interdiction” as a converse to the network resiliency problem.

DELIVERY FUNCTION AND DELIVERY IMPORTANCE

Delivery Function

To enable discussion of network resiliency in quantitative terms, we associate with a network a function Ψ that we call the *delivery function*. The intent is that this function will express quantitatively the purpose of the network or the service that the network is intended to provide. It may be a function, such as connectivity, that is already familiar in network reliability theory, or it may be a function, such as service reliability, that has not previously been considered in this context. The delivery function associates a real number or a vector with the elements of the network.

A network $(\mathcal{H}, \Psi)$ with delivery function is an ordered pair consisting of a graph $\mathcal{H}$ and associated delivery function Ψ . $\mathcal{H} = (\mathcal{N}, \mathcal{L})$ where $\mathcal{N}$ is a set of elements, called *nodes*, and $\mathcal{L}$ is a subset of $\mathcal{N} \times \mathcal{N}$. We denote the number of nodes in the graph by $|\mathcal{N}| = N$. If, for $i_1 \neq i_2$ belonging to $\mathcal{N}$, $(i_1, i_2) \in \mathcal{L}$, then (i_1, i_2) is a link; if $i_2 = i_1$, then (i_1, i_1) is another way of writing the node i_1 . $\mathcal{N} \cup \mathcal{L}$ is the set of network elements and Ψ maps $\mathcal{N} \cup \mathcal{L}$ into the real numbers or a vector space, depending on the range appropriate to the particular case studied.

Delivery Importance

This section discusses delivery importance, the influence of a network element, or a set of network elements, on the delivery function. The concept of delivery importance is introduced to enable us to see how changes in the network element(s) are reflected in the delivery function, that is, how well the network continues to perform the functions expected of it when there are changes in the network infrastructure. We first consider delivery importance in deterministic capacitated networks, both continuum and discrete. Delivery importance is defined for uncapacitated networks by considering the presence or absence of nodes and/or links in the graph, and assigning probabilities in the case of a stochastic uncapacitated network; see the section titled “Delivery Importance in Deterministic Uncapacitated Networks.” We then extend the definitions to stochastic capacitated networks, both continuum and discrete.

Delivery Importance in Deterministic Capacitated Networks. The *capacity* of a network element is a nonnegative real-valued function $c : \mathcal{N} \cup \mathcal{L} \rightarrow \mathbf{R}^+$. The capacity may represent the presence or absence of a network element (in which case the range of the capacity function is $\{0, 1\}$), the probability that a network element is in the graph, or the *reliability* of the network element (in which case the range is $[0, 1]$), or it may represent what is commonly understood as capacity in a flow network [18]. Capacity may be constant (static network capacity) or a function of time (dynamic network capacity). We also

consider capacitated networks in which the capacities are random quantities (stochastic capacitated networks) as being distinct from deterministic capacitated networks in which the capacities are not random.

We define the *capacity matrix* $C(\mathcal{H})$ to be the matrix whose (i, j) entry c_{ij} is the capacity of the link (i, j) if (i, j) is a link ($i \neq j$), 0 if (i, j) is not a link ($i \neq j$), and the capacity of the node i if $j = i$. In capacitated networks, we consider the domain of the network’s delivery function to be the space Γ of capacity matrices $\mathbf{R}^{N \times N}$. Therefore, for example, a scalar-valued delivery function maps a capacity matrix $C(\mathcal{H})$ into $\mathbf{R}^+$. It is also possible to consider vector-valued delivery functions. These are important for applications, but the extension of the theory given here to vector-valued delivery functions is straightforward (Eq. 8); therefore, a full discussion of vector-valued delivery functions is not undertaken here. A generic capacitated network with associated delivery function may be denoted by $(\mathcal{H}, C, \Psi)$.

Delivery Importance in Deterministic Capacitated Continuum Networks. In a continuum network, the range of the capacity function is the nonnegative real numbers, and a flow in the network [18] may take on nonnegative real values as well (subject to the constraint that the value of the flow at any network element not exceed the capacity of that network element). Continuum network models describe networks that deliver continuum materials such as oil, gas, water, and electricity or social networks in which the weight on a link (its “capacity”) is the strength of the relationship between the two nodes it connects. In a continuum network, we assume that the delivery function is a continuously differentiable function of the capacities except possibly on a set of Lebesgue measure zero. That is, for each network element $(i, j) \in \mathcal{N} \cup \mathcal{L}$, the derivative $\partial \Psi / \partial c_{ij}$ exists, except possibly on a set of Lebesgue measure zero in the space Γ of capacity matrices, and is continuous as a function of the capacity wherever it exists. It is possible that this derivative may be zero; if so, the network element is “unimportant” to the delivery function. This is an analogue

of the notion of coherence [19] in the mathematical theory of reliability.

The main idea of delivery importance is to see how the delivery function changes when the network capacity changes incrementally. That is, we would like to compare $\Psi(C_0 + hA)$ to $\Psi(C_0)$, where C_0 is a nominal capacity matrix, or study value of the capacity matrix, A is a matrix representing a “direction” in which capacity increases (or decreases) form the basis of the study, and h is a real number, typically close to zero. The basic delivery importance equation for the single network element (i,j) is, for fixed $a \neq 0$,

$$\Psi(c_{ij} + ha) - \Psi(c_{ij}) = \frac{\partial \Psi}{\partial c_{ij}} \Big|_{C_0} ha + o(h), \quad (1)$$

where the remainder term satisfies $o(h)/h \rightarrow 0$ as $h \rightarrow 0^+$, provided Ψ is differentiable at $(C_0)_{ij}$. There is no loss of generality in taking a to be 1 or -1 . $a = 1$ (respectively $a = -1$) indicates the delivery importance when the capacity of network element (i,j) increases (respectively decreases). The one-term Taylor series expansion (Eq. 1) motivates the following definition.

Definition. For a differentiable scalar delivery function, the *delivery importance at C_0 in the direction a ($a = \pm 1$)* of a network element $(i,j) \in \mathcal{N} \cup \mathcal{L}$, denoted by $\Omega_a(i,j; C_0)$, is defined as the derivative of Ψ with respect to the capacity c_{ij} of the network element (i,j) , evaluated at the capacity matrix C_0 :

$$\Omega_a(i,j; C_0) = a \frac{\partial \Psi}{\partial c_{ij}} \Big|_{C_0}, \quad (2)$$

when it exists.

Combining this with Equation (1) yields $\Psi(c_{ij} + ha) - \Psi(c_{ij}) = \Omega_a(i,j; C_0)h + o(h)$ for $a = \pm 1$ and $h \rightarrow 0^+$.

Example 1. Consider the simple directed, capacitated network consisting of two links in series, shown in Fig. 1.

The three nodes each are considered to have infinite capacity and the capacities of

the two links are x and y , respectively. The capacity matrix for this network is

$$C = \begin{pmatrix} \infty & x & 0 \\ 0 & \infty & y \\ 0 & 0 & \infty \end{pmatrix}. \quad (3)$$

The delivery function for this network is defined as the maximum flow from node 1 to node 3, which in this case is $\Psi(C) = \min\{x, y\} = 1/2(|x + y| - |x - y|)$, $x, y \geq 0$. Here, Ψ is differentiable except on the line $x = y$. We have, for $x \neq y$,

$$\begin{aligned} D_1 \Psi(x, y) &= I\{x < y\} \text{ and } D_2 \Psi(x, y) \\ &= I\{x > y\}, \end{aligned} \quad (4)$$

so the delivery importance in the positive direction of the link (1, 2) is $\Omega_1(1, 2; C) = 1$ if $x < y$ and 0 if $x > y$, and that of link (2, 3) is 0 if $x > y$ and 1 if $x < y$. Define the nominal capacity matrix

$$C_0 = \begin{pmatrix} \infty & c_{12} & 0 \\ 0 & \infty & c_{23} \\ 0 & 0 & \infty \end{pmatrix}. \quad (5)$$

For this nominal capacity matrix, $\Omega_1(1, 2; C_0) = 1$ if $c_{12} < c_{23}$ and $= 0$ if $c_{12} > c_{23}$, and $\Omega_1(2, 3; C_0) = 0$ if $c_{12} < c_{23}$ and $= 1$ if $c_{12} > c_{23}$.

When $x = y$, the derivatives do not exist. Define

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (6)$$

Then we return to the basic idea of delivery importance to write, for $h > 0$,

$$\begin{aligned} &\frac{1}{h} [\Psi(C_0 + hA) - \Psi(C_0)] \\ &= \frac{1}{h} [\min\{c_{12} + h, c_{12}\} - c_{12}] = 0. \end{aligned} \quad (7)$$

For equal initial link capacities, the delivery importance of link (1, 2) is zero when its



Figure 1.

capacity increases and -1 when its capacity decreases (use $-A$ in place of A); a similar analysis shows that the same is true for link $(2, 3)$. Thus, when the two links have the same capacity, they are of equal delivery importance. If the initial link capacities are equal, making one of them larger has no effect on the delivery function.

For a vector delivery function $\Psi = (\Psi_1, \dots, \Psi_v)$, the delivery importance of a network element $(i, j) \in \mathcal{N} \cup \mathcal{L}$ is the vector

$$\left(\frac{\partial \Psi_1}{\partial c_{ij}}, \dots, \frac{\partial \Psi_v}{\partial c_{ij}} \right), \quad (8)$$

when it exists. While vector delivery functions are important for many applications, the fact that the delivery importance is computed element by element, as in Equation (8), means that the delivery importance of a vector delivery function can be evaluated element by element, so we do not deal with vector delivery functions further in this article.

Because $\Psi: \Gamma \rightarrow \mathbf{R}$, $\Psi'(C)$ is a linear map from Γ into $\mathbf{R}$ for every C [20], and is therefore represented by a vector. To explicitly display this vector, it is convenient to order the network's incidence matrix as a vector by writing out its rows one after another as a row vector of length N^2 . Then the vector representing $\Psi'(C)$ is the vector of partial derivatives

$$\left(\frac{\partial \Psi}{\partial c_{11}}, \dots, \frac{\partial \Psi}{\partial c_{1N}}, \frac{\partial \Psi}{\partial c_{21}}, \dots, \frac{\partial \Psi}{\partial c_{2N}}, \right. \\ \left. \frac{\partial \Psi}{\partial c_{31}}, \dots, \frac{\partial \Psi}{\partial c_{NN}} \right),$$

where the (i, j) entry is zero if $(i, j) \notin \mathcal{L}$. Thus, we may write $\Omega_a(i, j; C_0)$ from Equation (2) as $\Omega_a(i, j; C_0) = \langle \Psi'(C_0), \mathbf{1}_{(i,j)} \rangle$ where $\mathbf{1}_{(i,j)}$ is a vector containing 1 in the (i, j) position and zeros elsewhere, and $\langle \cdot, \cdot \rangle$ represents the ordinary inner product in $\mathbf{R}^{N^2}$.

We may also consider the delivery importance of larger subsets of $\mathcal{N} \cup \mathcal{L}$. Let $\mathcal{M} \subset \mathcal{N} \cup \mathcal{L}$. Define $\mathbf{1}_{\mathcal{M}}$ to be the matrix whose (i, j) entry is given by $I\{(i, j) \in \mathcal{M}\}$, written as a row vector (see preceding paragraph) by concatenating the rows as a row vector of length N^2 . That is, $\mathbf{1}_{\mathcal{M}}$ contains a 1 in every position

corresponding to a network element belonging to $\mathcal{M}$ and a zero everywhere else. The direction $\mathbf{1}_{\mathcal{M}}$ represents increasing capacity of each of the network elements in $\mathcal{M}$. Then the delivery importance of $\mathcal{M}$ in the positive direction at C is defined as the derivative of the delivery function at C in the direction $\mathbf{1}_{\mathcal{M}}$, that is,

$$\Omega_+(\mathcal{M}; C) = \langle \Psi'(C), \mathbf{1}_{\mathcal{M}} \rangle. \quad (9)$$

The delivery importance of the empty set $\mathcal{M} = \emptyset$ is clearly zero.

Under suitable regularity conditions, the delivery importance may be computed by the Fréchet differential [21]

$$\langle \Psi'(C), \mathbf{1}_{\mathcal{M}} \rangle = \lim_{h \rightarrow 0^+} \frac{1}{h} [\Psi(C + h\mathbf{1}_{\mathcal{M}}) - \Psi(C)]. \quad (10)$$

We may consider delivery importance with any arrangement of signs in the matrix $\mathbf{1}_{\mathcal{M}}$, allowing for any combination of increases and decreases in network element capacity to be studied. In the definition of delivery importance, we have used the positive sign everywhere primarily for consistency with the classical definition of reliability importance [13]. For network resiliency (see the section titled "Network Resiliency"), we are primarily interested in what happens when capacities decrease, so in the definition of network resiliency we use delivery importance in the negative direction $-\mathbf{1}_{\mathcal{M}}$. It is also reasonable for some applications to consider delivery importance in other directions, which can be represented by matrices that contain $\pm I\{(i, j) \in \mathcal{M}\}$, where the 1 and -1 are chosen to represent growth or diminishment, respectively, of the capacity of the network element (i, j) . In this case, letting $A_{\mathcal{M}}$ stand for a matrix of this kind that contains $\pm I\{(i, j) \in \mathcal{M}\}$, we represent the delivery importance in the direction $A_{\mathcal{M}}$ by $\Omega_A(\mathcal{M}; C) = \langle \Psi'(C), A_{\mathcal{M}} \rangle$; in case all $-$ signs are chosen, we write $\Omega_-(\mathcal{M}; C) = \langle \Psi'(C), -\mathbf{1}_{\mathcal{M}} \rangle$.

Theorem 1. *For an almost everywhere differentiable scalar delivery function, the delivery importance in the positive direction at C_0 of a set $\mathcal{M}$ of network elements is equal*

to the sum of the partial derivatives of Ψ with respect to the capacities of the network elements in $\mathcal{M}$, evaluated at the nominal capacity matrix C_0

$$\Omega_+(\mathcal{M}; C_0) = \sum_{(i,j) \in \mathcal{M}} \left. \frac{\partial \Psi}{\partial c_{ij}} \right|_{C_0}, \quad (11)$$

when it exists.

Proof. This is a restatement of the inner product expression $\Omega_+(\mathcal{M}, C_0) = \langle \Psi'(C_0), \mathbf{1}_{\mathcal{M}} \rangle$.

Clearly, when $\mathcal{M}$ consists of a single element (i, j) , $\Omega_+(\mathcal{M}; C)$ reduces to $\Omega_1(i, j; C)$. Again, this definition is motivated by a one-term Taylor series expansion

$$\begin{aligned} \Psi(C_0 + hA) - \Psi(C_0) \\ = \sum_{(i,j) \in \mathcal{M}} h_{(i,j)} \left. \frac{\partial \Psi}{\partial c_{ij}} \right|_{C_0} a_{ij} + o(|h|), \end{aligned} \quad (12)$$

where h is now an $|\mathcal{M}|$ -vector; Equation (12) is derived by taking all elements of h to be equal (which amounts to modeling the postulate that the capacities of all network elements in $\mathcal{M}$ are changed at the same rate).

Example 1 (continued). We compute the delivery importance of the two links (1, 2) and (2, 3) together. In this example, this is the delivery importance of the entire network with respect to itself. We have

$$\Omega_+(\mathcal{H}; C_0) = I\{x < y\} + I\{y < x\} = I\{c_{12} \neq c_{23}\}, \quad (13)$$

which is 1 as long as $c_{12} \neq c_{23}$. Again when $c_{12} = c_{23}$, Equation (11) is not defined and we use the $|\mathcal{M}|$ -dimensional analog of Equation (1) instead, with $\mathbf{1}_{\mathcal{H}}$ as the 3×3 matrix containing 1 in the (1, 2) and (2, 3) positions and

zeros elsewhere

$$\begin{aligned} \Omega_+(\mathcal{H}; C_0) \\ = \lim_{h \rightarrow 0^+} \frac{1}{h} [\Psi(C_0 + h\mathbf{1}_{\mathcal{H}}) - \Psi(C_0)] \\ = \lim_{h \rightarrow 0^+} \frac{1}{h} [\min\{c_{12} + h, c_{12} + h\} - c_{12}] = 1. \end{aligned} \quad (14)$$

In this example, the delivery importance of the entire network with respect to itself is 1 regardless what C_0 might be.

Delivery Importance in Deterministic Capacitated Discrete Networks. In a capacitated discrete network, the capacity matrix has integer entries and the flows also take on only integer values. Derivatives of the delivery function are not defined in the classical sense. There are (at least) two alternatives. First, one may approximate the discrete network by a sequence of continuum networks for which the delivery importance is defined as in the section titled “Delivery Importance in Deterministic Capacitated Continuum Networks,” in this article, and then define the delivery importance for the discrete network as the limit of the delivery importance values in the approximating sequence of continuum networks, if the limit exists. The second alternative is to define delivery importance in a purely discrete fashion, as follows. Let $(\mathcal{H}, C, \Psi)$ be a deterministic capacitated discrete network with associated delivery function Ψ . The delivery importance at C of a network element (i, j) in the positive direction is given by

$$\Omega_+(i, j; C) = \Psi(C + \mathbf{1}_{ij}) - \Psi(C), \quad (15)$$

where $\mathbf{1}_{ij}$ is a matrix with a 1 in the i, j position and zeroes everywhere else. Similarly, for the delivery importance in the positive direction of a subset $\mathcal{M}$ of $\mathcal{N} \cup \mathcal{L}$, replace in Equation (15) the $\mathbf{1}_{ij}$ matrix by $\mathbf{1}_{\mathcal{M}}$. Unless otherwise noted, in this article we shall use the second alternative.

Delivery Importance in Stochastic Capacitated Networks. In a stochastic capacitated network, the network element capacities are

random variables. Consequently, the delivery function (we will assume it is measurable), delivery importance, and network resiliency in such a network are also random variables. As such, figures of merit for delivery importance and network resiliency may be constructed from moments and percentiles of these random variables. Thus, we speak of, for example, expected delivery importance as

$$\begin{aligned} E\Omega_+(\mathcal{M}; C) &= E\langle \Psi'(C), \mathbf{1}_{\mathcal{M}} \rangle \\ &= \int \langle \Psi'(C(\omega)), \mathbf{1}_{\mathcal{M}} \rangle P(d\omega), \end{aligned} \quad (16)$$

where the integration is over the sample space of the capacity random variables. In a similar fashion, we may employ the variance of delivery importance, the median delivery importance, the ninetieth percentile delivery importance, and so on.

Static Networks. In a static stochastic capacitated network, the capacities are fixed random quantities. These may represent the reliability of the network elements (in which case the range of the capacity function is $[0, 1]$ and the capacity represents the probability that the element is present in $\mathcal{H}$; see Example 2 for an illustration of this case), or the load-carrying capacity of the network element in a model in which the capacity is indeterminate for any reason (e.g., lack of precise information on the part of the network manager about the status of recently added elements to the network). It is required that the joint distribution of the network element capacities be known.

Dynamic Networks. In a dynamic stochastic capacitated network, the capacities are represented by stochastic processes with time as their parameter space. These networks model realistic networks having network element capacities that change at random because of, say, failures and repairs to parts of the network element. For example, the capacity of a network element in a deterministic capacitated network could be modeled with a continuous-time, discrete-state semi-Markov process. For dynamic stochastic capacitated networks,

the figures of merit for delivery importance, like those described in the section “Delivery Importance in Stochastic Capacitated Networks,” become functions of time as well.

Delivery Importance in Uncapacitated Networks

Delivery Importance in Deterministic Uncapacitated Networks In an uncapacitated network, proper functioning of the network is affected by only the presence or absence of a node or link, so “capacities” zero or one are assigned to each element of the network. The delivery function in such a network is usually one of the connectivity functions: 2-terminal, k -terminal, or all-terminal. Delivery importance in the positive direction, as defined in the section titled “Delivery Importance in Deterministic Capacitated Discrete Networks,” makes sense only for entries in the nominal capacity matrix where there are zeros. Similarly, delivery importance in the negative direction makes sense only for entries in the nominal capacity matrix where there are ones. For applications, probably the most useful delivery importance measure is delivery importance in the negative direction starting from a fully operational network, as in

$$\Omega_-(\mathcal{M}; H) = \Psi(H - \mathbf{1}_{\mathcal{M}}) - \Psi(H), \quad (17)$$

where H is the full incidence matrix (i.e., including the nodes) of $\mathcal{H}$. As usual, a vector delivery function is handled term by term.

Delivery Importance in Stochastic Uncapacitated Networks In a stochastic uncapacitated network, the basic setup is as in the section titled “Delivery Importance in Deterministic Uncapacitated Networks” with the additional feature that network elements are each assigned a number representing the probability that the element is in the graph or not. In case these probabilities are fixed, the network is static in the sense of the section titled “Static Networks”; in case they vary over time in some known fashion, the network is dynamic in the sense of the section titled “Dynamic Networks”. Moments, percentiles, and so on of the delivery function

and various delivery importance values can then be computed from these probabilities.

Example 1 (continued). Consider again the network shown in Fig. 1, now with unreliable links. We set $p_{ij} = EI\{(i,j) \in \mathcal{H}\} = P\{(i,j) \in \mathcal{H}\}$, $i, j = 1, 2, 3$. For economy's sake, we assume $p_{11} = p_{22} = p_{33} = 1$, and $p_{13} = p_{31} = 0$ because the network does not contain a link from node 1 to node 3. $p_{21} = p_{32} = 0$, because the network is directed, and we assume that the remaining probabilities are neither zero nor one. Then the expected value of the delivery function is $E\Psi = p_{12} + p_{23} - p_{12}p_{23}$. The expected delivery importance in the negative direction of the links are $E\Omega_-(1, 2; H) = E\Omega_-(2, 3; H) = p_{12}p_{23} - p_{12} - p_{23}$.

NETWORK RESILIENCY

Deterministic Networks

The first essential idea of network resiliency is that a resilient network is one that has few subsets that have large delivery importance, or conversely most subsets have small delivery importance. This expresses the idea that removal of, or decrease in capacity of, most network elements has little effect on the delivery function of the network if the network is resilient. The remainder of this section is devoted to formalizing this idea in the context of the framework we have created for delivery functions and delivery importance.

In the most general formulation, we define network resiliency to be the proportion of subsets of the network whose absolute delivery importance values in the negative direction do not exceed a given value. Formally, this is

$$\rho(\mathcal{H}, C_0; x) = \frac{1}{2^{|\mathcal{N} \cup \mathcal{L}|}} \sum_{\mathcal{M} \subset \mathcal{N} \cup \mathcal{L}} I\{|\Omega_-(\mathcal{M}; C_0)| \leq x\},$$

$$x \geq 0, \quad (18)$$

a nondecreasing function. A resilient network is one for which this function remains close to zero, only increasing sharply to 1 for large values of x .

The computation of $\rho(\mathcal{H}, C_0; x)$ by Equation (18) can be daunting for even modest-sized networks, and in practical cases there is usually little reason to consider very large subsets of $\mathcal{N} \cup \mathcal{L}$ unless events that disturb all but a small number of network elements are common.¹ Therefore, there is interest in restricted network resiliency figures of merit based on only certain subsets of $\mathcal{N} \cup \mathcal{L}$, especially if study of only a particular subset is of interest or if it is possible to identify in advance subsets that have less influence on the delivery function. It is useful to denote restricted network resiliency, say with respect to a subset Z of $\mathcal{N} \cup \mathcal{L}$, by $\rho(\mathcal{H}, Z, C_0; x)$. Then we would have

$$\rho(\mathcal{H}, Z, C_0; x) = \frac{1}{|Z|} \sum_{\mathcal{M} \subset Z} I\{|\Omega_-(\mathcal{M}; C_0)| \leq x\},$$

$$x \geq 0. \quad (19)$$

For instance, we may be interested in the resiliency of the network when disruptions to only single network elements are considered; we call this the 1-resiliency of the network and compute it with Z equal to the set of all single-element subsets of $\mathcal{N} \cup \mathcal{L}$. More generally, we may consider the k -resiliency of the network

$$\rho_k(\mathcal{H}, C_0; x) = \left(\binom{|\mathcal{N} \cup \mathcal{L}|}{k} \right)^{-1} \sum_{\mathcal{M} \subset Z_k} I\{|\Omega_-(\mathcal{M}; C_0)| \leq x\},$$

$$x \geq 0, \quad (20)$$

where Z_k represents the set of all subsets of $\mathcal{N} \cup \mathcal{L}$ that contain k elements.

In applications, these expressions may be difficult to work with because they are functions, rather than scalars, so comparisons between different networks are not always possible.² Also, significant computation might be required to evaluate them. Thus, there is interest in simpler network

¹Certain kinds of cyber attacks can disturb large subsets of a network at once.

²A partial order, similar to stochastic order, can be developed on the basis of these expressions.

resiliency characterizations; the section titled “Figures of Merit for Network Resiliency” is devoted to some explorations of these ideas.

Stochastic Networks

In a stochastic network, $\Omega_-(\mathcal{M}, C_0)$ is a random variable, and the expected value of Equation (18) contains its distribution:

$$\begin{aligned} E\rho(\mathcal{H}, C_0; x) &= \frac{1}{2^{|\mathcal{N} \cup \mathcal{L}|}} \sum_{\mathcal{M} \subset \mathcal{N} \cup \mathcal{L}} P\{|\Omega_-(\mathcal{M}; C_0)| \leq x\}, \\ x &\geq 0. \end{aligned} \quad (21)$$

Similar expressions for the expected values of Equations (19) and (20) may be developed.

FIGURES OF MERIT FOR NETWORK RESILIENCY

Deterministic Networks

We use the concepts of delivery function and delivery importance of (sets of) network elements to define network resiliency in quantitative terms. Note that because delivery importance is keyed to a specific delivery function and nominal capacity matrix, network resiliency is specifically defined only with respect to the chosen delivery function and nominal capacity matrix.

Definition. For a deterministic capacitated network $(\mathcal{H}, C, \Psi)$ with associated delivery function, *scalar network resiliency* $\rho^*(\mathcal{H}; C_0)$ at a nominal capacity matrix C_0 is defined as the largest of the absolute delivery importance values in the negative direction (capacities decrease) across all subsets of the elements (nodes and links) in the network:

$$\rho^*(\mathcal{H}; C_0) = \max_{\mathcal{M} \subset \mathcal{N} \cup \mathcal{L}} |\Omega_-(\mathcal{M}; C_0)| \quad (22)$$

The largest delivery importance value indicates the subset of $\mathcal{N} \cup \mathcal{L}$ to which the delivery function is most sensitive or that has the largest “lever effect” on the delivery function. Note that smaller is better in

Equation (22) because we want a resilient network to be one that is insensitive to changes in its network elements, that is, one that makes the right-hand side of Equation (22) small. That is, we say a network is resilient if $\rho^*(\mathcal{H}, C_0)$ is small. Computation of network resiliency for a scalar delivery function by this definition requires only $|\mathcal{N} \cup \mathcal{L}|$ delivery importance computations and at most $|\mathcal{N} \cup \mathcal{L}| \cdot 2^{|\mathcal{N} \cup \mathcal{L}|}$ additions because the delivery importance for a subset $\mathcal{M}$ of $\mathcal{N} \cup \mathcal{L}$ is obtained by summing the delivery importance values of each of the individual network elements contained in $\mathcal{M}$; see Theorem 1.

Clearly, the same kinds of restricted scalar network resiliency values as, for example, in Equations (19) and (20) are possible here as well, unless otherwise specified, further discussion of scalar network resiliency in this article refers to the full definition (22).

Scalar network resiliency is not an absolute scale quantity but it induces a partial order on the set of networks with associated delivery functions. Therefore, it is mainly useful for making comparisons in a set of networks rather than for associating a meaningful number with a fixed network.

Stochastic Networks

In a stochastic network, scalar network resiliency (Eq. 22) is a random variable, and may be a function of time also if the network is dynamic. As before, figures of merit for network resiliency are constructed on the basis of moments and percentiles of the network resiliency random variable. If the network is dynamic, these figures of merit will be functions of time as well.

METRICS FOR NETWORK RESILIENCY

A figure of merit, in the sense used in the section “Figures of Merit for Network Resiliency,” represents an “ideal” description of some phenomenon in a model. A figure of merit is computed from assumptions and a model structure and represents a quantitative description of some inherent quality of the phenomenon under study. For

example, “maintainability” is the inherent ability of a system to be repaired and restored to service when maintenance is conducted by personnel using specified skill levels and prescribed procedures and resources [22]. Many figures of merit are associated with maintainability, such as maintenance downtime and logistic delay. These may be developed from a model of the system under study (e.g., an alternating renewal process) together with assumptions about location of spares, skill of repair personnel, and so on. When the system is in operation, data may be collected relating to these figures of merit, and estimators of the population value of the figure of merit may be constructed from these data. These estimators are called *metrics*. For example, during system operation we may record the amount of time the system is out of service for maintenance reasons (preventive, corrective, opportunistic, etc.) and these data may be used to estimate the mean and variance of the maintenance downtime across a population of similar systems. For further discussion of figures of merit and metrics, see [11].

In the case of network resiliency, we have introduced several figures of merit in the section titled “Figures of Merit for Network Resiliency.” Metrics may be constructed for these from data relating to delivery importance. For instance, we may record the change in the delivery function between two specified network nodes when there is a failure of one element somewhere in the network. Aggregation of these data would permit estimation of the 1-resiliency of the network. Clearly, many details have been omitted from this description, and several questions would need to be settled before satisfactory metrics for network resiliency could be developed. We reserve detailed discussion of metrics for network resiliency for further treatment elsewhere.

THREE EXAMPLES OF DELIVERY FUNCTIONS

Connectivity

Consider a network $(\mathcal{H}; C)$ in which all capacities are either zero or one (i.e., $c_{ij} = I\{(i, j) \in \mathcal{L}\}$). Associated delivery functions consistent

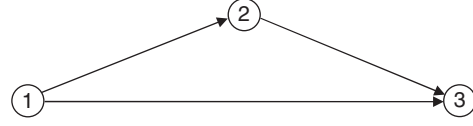


Figure 2.

with this example are 2-terminal, k -terminal ($2 < k < N$), or all-terminal connectivity of $(\mathcal{H}; C)$. We define the connectivity of some subset of nodes in the network to be 1 if there are paths connecting all the nodes in the subset to one another (i.e., every node in the subset can be reached from other node in the subset), and 0 otherwise. Note that some authors define connectivity to be the probability that paths connecting the nodes exist when the probability that a network element is in the graph is given; this is the expected connectivity in our formulation and is studied in the next section. Computation of all-terminal connectivity in the latter sense (i.e., all-terminal reliability) using a reduction algorithm is reviewed in Ref. 23.

The following simple example conveys the essential points of network resiliency involving connectivity.

Example 2. Consider the simple three-terminal directed network H shown in Fig. 2.

The full incidence matrix of $\mathcal{H}$ is

$$H = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad (23)$$

and the delivery function Ψ for this example is the two-terminal connectivity for terminals 1 and 3. That is, $\Psi = 1$ if there is a path connecting node 1 to node 3 and is zero otherwise. The capacity of element $(i, j) \in \mathcal{H}$ is $c_{ij} = I\{(i, j) \in \mathcal{H}\}$, $i, j = 1, 2, 3$. For economy's sake, let us ignore the contribution of the nodes; it is clear enough how to incorporate their contribution should it be desirable. Then

$$\begin{aligned} \Psi &= I\{(1, 3) \in \mathcal{H}\} \cup [(1, 2) \in \mathcal{H}] \cap [(2, 3) \in \mathcal{H}] \\ &= c_{13} + c_{12}c_{23} - c_{12}c_{23}c_{13}. \end{aligned} \quad (24)$$

The delivery importance of link (1,3) in the negative direction is

$$\begin{aligned}
\Omega_-((1,3); C) &= \Psi(C - \mathbf{1}_{(1,3)}) - \Psi(C) \\
&= c_{12}c_{23} - [c_{13} + c_{12}c_{23} - c_{12}c_{23}c_{13}] \\
&= c_{12}c_{23}c_{13} - c_{13}. \tag{25}
\end{aligned}$$

Table 1 gives the delivery importance values for each nonempty subset of $\mathcal{H}$.

Choose the network's incidence matrix H (Eq. 23) as a nominal capacity matrix. This represents a situation in which all the network's elements are initially in an operating condition. Then, with respect to this nominal capacity matrix, the delivery importance values from Equation (17) are $\Omega_-(\mathcal{M}, H) = 0$ for $\mathcal{M} = \{(1,3)\}, \{(2,3)\}, \{(1,2)\}$, and $\{(1,2) \cup (2,3)\}$, and $\Omega_-(\mathcal{M}, H) = -1$ for $\mathcal{M} = \{(1,3) \cup (2,3)\}, \{(1,3) \cup (1,2)\}$, and $\mathcal{H}$. The network resiliency function (18) at the nominal capacity matrix H is a right-continuous step function with jumps at 0 and 1; the height of the jump at zero is 5/8 and the height of the jump at 1 is 3/8.

Reliability

Consider a network $(\mathcal{H}; C)$ in which, instead of all capacities being zero or one (i.e., $c_{ij} = I\{(i,j) \in \mathcal{L}\}$ as before), we postulate that $P\{c_{ij} = 1\} = p_{ij}$ for all $(i,j) \in \mathcal{L}$. Such a model represents a network in which network elements are unreliable, that is, may fail to function in the network with some given probability. Suitable delivery functions in this case include 2-terminal, k -terminal, and all-terminal reliability, defined as the probability that there are working paths

Table 1. Delivery Importance Values for Example 2

Subset of $\mathcal{H}$	Delivery Importance
$\{(1,3)\}$	$c_{12}c_{23}c_{13} - c_{13}$
$\{(2,3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$
$\{(1,2)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$
$\{(1,2) \cup (2,3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$
$\{(1,3) \cup (2,3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13} = -\Psi(\mathcal{H})$
$\{(1,3) \cup (1,2)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13}$
$\mathcal{H}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13}$

connecting a given subset of the nodes of the network (two nodes, k nodes, or all nodes, as appropriate). We continue with Example 2 to illustrate network resiliency in this case.

Example 2 (continued). The capacity of element $P\{c_{ij} = 1\} = p_{ij}$ for all $(i,j) \in \mathcal{L}$. We assume the network elements are mutually stochastically independent. For economy's sake, we again ignore the contribution of the nodes. Table 2 gives the delivery importance and expected delivery importance values for each nonempty subset of $\mathcal{H}$.

As before, when the nominal capacity matrix is H , the delivery importance values from Equation (17) are $\Omega_-(\mathcal{M}, H) = 0$ for $\mathcal{M} = \{(1,3)\}, \{(2,3)\}, \{(1,2)\}$, and $\{(1,2) \cup (2,3)\}$, and $\Omega_-(\mathcal{M}, H) = -1$ for $\mathcal{M} = \{(1,3) \cup (2,3)\}, \{(1,3) \cup (1,2)\}$, and $\mathcal{H}$. The network resiliency function is zero with probability $p_\infty = P[\{(1,3) \notin \mathcal{H}\} \cup \{(2,3) \notin \mathcal{H}\} \cup \{(1,2) \notin \mathcal{H}\} \cup \{(\{(1,2) \notin \mathcal{H}\} \cap \{(2,3) \notin \mathcal{H}\})\}]$ and is -1 with probability $1 - p_\infty$. The expected network resiliency is $p_\infty - 1$.

Stochastic Flow

Consider a continuum network that delivers, for example, oil from node A to node B. The node and link capacities in this network are modeled as continuous-time Gaussian processes $\{X_{ij}(t) : t \geq 0\}$ having almost surely continuously differentiable sample paths [25, Section 4.3], indicating the change in capacity of the network elements from time to time due to failures and repairs, degradation, corrosion, or similar phenomena. The delivery function is the volume of oil delivered from node A to node B over a stated interval of time $[0, T]$, $T < \infty$. The capacity of a network element is the maximum flow (volume per unit time) permitted in that network element. Set $X(t) = (X_{ij}(t) : i, j = 1, \dots, N)$ and $C_0 = X(0)$. Let $\varphi(t)$ denote the flow in the network at time t assuming a maximum-flow scenario [18], and $\varphi_{AB}(t)$ denote the flow from node A to node B under this scenario. Then the delivery function consistent with this model is

$$\Psi(X; T) = \int_0^T \varphi_{AB}(t) dt. \tag{26}$$

Then the delivery importance of a subset $\mathcal{M}$ of network elements is given by the random variable

$$\Omega_-(\mathcal{M}, X) = - \sum_{(i,j) \in \mathcal{M}} \int_0^T \frac{\partial \varphi_{AB}}{\partial x_{ij}}(t) dt. \quad (27)$$

If, as is usually reasonable to assume, the flow φ_{AB} is a nondecreasing function of the network element capacities, then $\Omega_-(\mathcal{M}, X) \leq 0$, and the scalar network resiliency with respect to the given delivery function is

$$\rho^*(\mathcal{H}; X) = \max_{\mathcal{M} \subset \mathcal{N} \cup \mathcal{L}} \sum_{(i,j) \in \mathcal{M}} \int_0^T \frac{\partial \varphi_{AB}}{\partial x_{ij}}(t) dt. \quad (28)$$

Assuming the processes $\{X_{ij}(t)\}$ are mutually stochastically independent, we obtain

$$P\{\Psi(X; T) > V\} = \int_0^\infty \cdots \int_0^\infty P \left\{ \int_0^T \varphi_{AB}(t) dt > V | X_{ij}(t) = x_{ij} \right\} dp \{X_{ij}(t) \leq x_{ij}\} \quad (29)$$

and then the methods of Ramirez-Marquez *et al.* [25, 26] may be used to calculate the conditional probabilities in Equation (29), where V is a desired volume of oil.

NETWORK RESILIENCY OPTIMIZATION

In creation of real networks, it is in many cases clearly of interest to be able to optimize network resiliency. The framework outlined above enables the formulation of an optimization problem for network resiliency in a capacitated network when a cost function for provisioning specified node and link capacity

is known. $\gamma(C)$ denotes the cost of deploying a network whose capacity matrix is C .

Solving the following optimization problem leads to a network designed for maximum scalar resiliency:

$$\text{Maximize } \rho^*(\mathcal{H}; C) \text{ subject to } \gamma(C) \leq \gamma_0,$$

where γ_0 is the allowable maximum cost. The design variables that may be manipulated to complete the optimization include the network topology and the capacity matrix.

DESIGN FOR NETWORK RESILIENCY PRINCIPLES AND PRACTICES

To develop a good catalog of design for network resiliency principles and practices requires a extensive experience in solving network resiliency optimization problems and applying the solutions in real networks. Because this is a rather new area, not much such experience exists. Nonetheless, some general principles can be stated by appealing to properties of the delivery function, graph structure, and so on. This section discusses some specific actions that can be taken by network designers to increase network resiliency. We evaluate the efficacy of design actions by determining whether the scalar network resiliency (see the section titled “Figures of Merit for Network Resiliency”) decreases (smaller is better) as a result of the action. In a capacitated network, it is reasonable to assume that the delivery function is a nondecreasing function of the capacity matrix (i.e., an increase in one or more entries in the capacity matrix causes the delivery function to increase or remain the same).

Table 2. Expected Delivery Importance Values for Example 2

Subset of $\mathcal{H}$	Delivery Importance	Expected Delivery Importance
$\{(1, 3)\}$	$c_{12}c_{23}c_{13} - c_{13}$	$p_{12}p_{23}p_{13} - p_{13}$
$\{(2, 3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$	$p_{12}p_{23}p_{13} - p_{12}p_{23}$
$\{(1, 2)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$	$p_{12}p_{23}p_{13} - p_{12}p_{23}$
$\{(1, 2) \cup (2, 3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23}$	$p_{12}p_{23}p_{13} - p_{12}p_{23}$
$\{(1, 3) \cup (2, 3)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13} = -\Psi(\mathcal{H})$	$p_{12}p_{23}p_{13} - p_{12}p_{23} - p_{13}$
$\{(1, 3) \cup (1, 2)\}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13}$	$p_{12}p_{23}p_{13} - p_{12}p_{23} - p_{13}$
$\mathcal{H}$	$c_{12}c_{23}c_{13} - c_{12}c_{23} - c_{13}$	$p_{12}p_{23}p_{13} - p_{12}p_{23} - p_{13}$

Systematic Overprovisioning

In a capacitated network, one of the most common ways of promoting greater network resiliency is by overprovisioning, that is, making the capacities of the network elements larger than would be required to just meet stated service reliability objectives. For example, in early telecommunications networks, certain trunk groups (collections of circuits connecting two switching offices) were designed to block no more than 1% of the offered traffic during the busy hour [27] (i.e., no more than 1% of the arriving calls will find all trunks already occupied). For an offered load of 100 Erlangs (one Erlang is one call arrival per mean holding time; for example, if the mean holding time of a call is 6 min, one Erlang of traffic constitutes a Poisson process whose arrival rate is 10 calls per hour), the number of trunks required to meet this goal is 117. To overprovision, a larger number of trunks, say 144, would be installed, and then, with the same 100 Erlang load, the blocking probability would be reduced to 6.7×10^{-6} . This same idea can be replicated throughout a capacitated network; while the computation of delivery importance and network resiliency might be more complicated, the basic idea of overprovisioning is the same as it is in the simple circuit-switched telecommunications trunk group example.

If the delivery function is a nondecreasing function of the capacity matrix, then overprovisioning (i.e., providing more capacity on a route than is required to just meet a stated service objective for that route) leads to increased scalar network resiliency with respect to a nominal capacity matrix representing exact provisioning because all delivery importance values are nonnegative. Whether overprovisioning is an efficient means of promoting network resiliency is another matter that is beyond the scope of this article.

Failure Decoupling

When a network's associated delivery function is nondecreasing, the larger the number of network elements that experience malfunction (or decrease in capacity), the larger

the decrease in the value of the delivery function. It follows that network resiliency will be improved if there can be employed schemes to prevent failures in network elements from cascading or propagating to other network elements. We refer to such schemes broadly as *failure decoupling*. Failure to properly decouple network elements may lead to widespread network element outages and greatly impaired delivery functions. For example, in January 1990, the AT&T Signaling System 7 (SS7) transaction management data network suffered an extensive outage because a software failure in one of the SS7 switching nodes led to a situation in which nodes connected to the failed node went out of service because they received improper status information from the failed node. The failures cascaded in less than 10 min to over 100 SS7 nodes, resulting in a 9-h outage that affected 60,000 customers and cost AT&T approximately \$60 million in lost revenue [28]. Had flawed status information not been passed on to neighboring nodes, the scope of the outage would have been greatly reduced.

Measures that can be taken as part of network design to minimize the possibility of propagating a network element failure to other network elements include implementation of a supervisory network monitoring system such as that employed in undersea cable telecommunications systems [29], implementation of a "heartbeat" status monitoring and communication system [30], and implementation of an alarm capability to notify network operations managers of network element failure in real time.

The latter scheme points toward the role of network management and control in containing the propagation of network element failures. As many large network operations implement a centralized network monitoring and intervention capability (e.g., the AT&T Global Network Operations Center [31]), we discuss this idea in more detail below.

Reactive Control. Reactive control refers to the actions that are taken by network managers (human or automatic) in response to a disruption in the network. A disruption is by definition something that interferes with the satisfactory functioning of one or more

network elements, leading to a degradation in the performance of the network's functions, and (with the assumption about monotonicity of the delivery function given in the section titled "Design for Network Resiliency Principles and Practices") is reflected in a decrease in the value of the network's delivery function. Disruptions may be routine and manageable, or they may be sudden and unexpected.

Routine disruptions are those that can be anticipated and planned for in advance of their occurrence. For example, the chronic process of failures and repairs that takes place in all network elements can be thought of as routine disruption because network managers know that these will occur even if it is not possible to predict exactly when and where they will occur. In a sense, they can be predicted statistically; indeed, about 3% of the routers in the Internet are out of service at any time because of routine failures. Consequently, managers can anticipate these failures and put in place operational tactics that mitigate their effects. Indeed, the TCP/IP network protocol is specifically designed to automatically route around failed routers and/or links. This, together with the hub-and-spoke topology of the Internet, allows us to conjecture that the resiliency (Eq. 22) of the Internet has a logistic shape: it remains small for increasing numbers of failed routers and/or links until a threshold is reached, and then it increases abruptly as the number of failed routers and/or links continues to increase. An example of an automatic reactive control is given by the AT&T FAS-TAR [32] system created to enable automatic rerouting around failed DS3 transport links in the public switched telephone network.

Sudden and unpredictable disruptions are due to events like natural disasters and deliberate attacks. These are usually large in scope and extended in time. In circuit-switched telecommunications networks, long experience helped codify a set of network management principles that were employed to mitigate the effects of losing large portions of a network due to, for example, earthquakes, floods, and other natural disasters. That technology had advanced to a point that it was encapsulated in an expert system [33].

Smart Networks. There has been a recent trend toward exploiting information technology for improved management of network operations. The ability to acquire and archive large quantities of data makes it possible to obtain a richer picture of network operations and to run analyses in real time that allow more intelligent decisions to be made regarding management of a network for continued satisfactory operation of a delivery function when changes to the network occur. Two approaches have been common in recent literature: a top-down approach and a bottom-up approach. We briefly discuss each of these in turn.

Macro-Smarts. A macro-smart network is one that uses global, or network-wide, data to inform network managers about current status and potential evolution of network operation. Internet protocol (IP) telecommunications networks encapsulate this sort of network management by the way routing tables change in response to failed links and routers. Updating routing tables in (close to) real time allows an IP network's open-shortest-path-first (OSPF) routing scheme to automatically bypass failed routers and links. That is, a router in an IP network will never even try to send a packet over a failed link or to a failed router when its routing tables are up to date with current information on the location of failures in the network.

Micro-Smarts. A micro-smart network is one that uses local data to inform network managers about current status and potential evolution of network operation. For example, electric power networks are beginning to use so-called *smart metering* to enable the network operator to control the flow of power to specific end users at specific times so that focused overloads, such as might occur because of heavy air-conditioning demands on hot summer days, can be avoided for the benefit of a larger number of users who will not experience a blackout that might otherwise occur due to the overload. This is an example of an active intervention that is possible because information that was not previously available at the end user level of

detail is now made visible by smart metering. It is anticipated that the smart grid market will reach \$200 billion by 2015 [34].

NETWORK INTERDICTION

Network interdiction can be considered as the study of methods for making a network *less* resilient. Network interdiction studies aim to find the network element, or set of network elements, that maximally disrupts the operation of the network. In the language introduced here, maximal disruption to the operation of the network means the largest decrease in the network's delivery function. The set $\mathcal{M}_0$ of $\mathcal{N} \cup \mathcal{L}$ that realizes the scalar network resiliency value, $\rho^*(\mathcal{H}; C) = |\Omega_-(\mathcal{M}_0; C)|$, is the subset of network elements whose removal from the network decreases the delivery function (when the capacity matrix is C) more than the removal of any other subset and therefore is the set sought by network interdiction.

Network interdiction studies began in a military context; for some examples see Refs 36 and 37. Other network interdiction studies include Refs 37–40.

CONCLUSION

This article presents a quantitative framework for studying network resiliency. The concept of network resiliency has gained increased importance recently as executives and managers realize the possibilities for disruption of critical network infrastructures because of not only natural disasters but also malicious actors and deliberate attacks. We introduce the concepts of a network with associated delivery function and of delivery importance to enable quantitative characterization of network resiliency. We discuss the application to discrete and continuum capacitated and uncapacitated networks, both deterministic and stochastic. Three examples of delivery functions are provided, and we briefly summarize how to solve the network interdiction using the proposed network resiliency framework.

REFERENCES

1. Burr SA, editor. The mathematics of networks. Volume 26, The Proceedings of Symposia in Applied Mathematics. Providence (RI): American Mathematical Society; 1982.
2. Colbourn C. The combinatorics of network reliability. New York: Oxford University Press; 1987.
3. Colbourn C. Network resilience. SIAM J Algebra Discr Methods 1987;8(3):404–409.
4. Touvet F, Harle D. Network resilience in multilayer networks: a critical review and open issues, Lecture Notes in Computer Science No. 2093. New York: Springer; 2001. pp. 829–837.
5. Bu T, Norden S, Woo T. Trading resiliency for security: model and algorithms. In: Proceedings of the 12th IEEE Conference on Network Protocols. 2004. pp. 218–227.
6. Dolev D, Jamin S, Mokryn O, *et al.* Internet resiliency to attacks and failures under BGP policy routing. Comput Netw 2006;50(16):3183–3196.
7. Sohn J, Tschangho JK, Hewings GJD, *et al.* Retrofit priority of transport network links under an earthquake. J Urban Plann Dev 2003;129(4):195–210.
8. Najjar W, Gaudiot J-L. Network resilience: a measure of network fault tolerance. IEEE Trans Comput 1990;39(2):174–181.
9. Najjar W, Srimani PK. Network resilience of star graphs: a comparative analysis. In: Proceedings of the 19th Annual Conference on Computer Science. 1991. pp. 349–357.
10. Lang JP, Drake J. Mesh network resiliency using GMPLS. Proc IEEE 2002;90(9):1559–1568.
11. Tortorella M. Service reliability theory and engineering, I: foundations. Qual Technol Quant Manage 2005;2(1):1–16.
12. Tortorella M. Service reliability theory and engineering, II: models and examples. Qual Technol Quant Manage 2005;2(1):17–37.
13. Birnbaum ZW, Esary JD, Saunders SC. Multicomponent systems and their reliability. Technometrics 1961;3:55–77.
14. Ramirez-Marquez JE, Rocco CM, Gebre DA, *et al.* New insights on multi-state component criticality and importance. Reliab Eng Syst Saf 2006;91(8):894–904.
15. Medhi D. A unified approach to network survivability for teletraffic networks: models, algorithms, and analysis. IEEE Trans Commun 1994;42(2/3/4):534–548.

16. Frank H. Survivability analysis of command and control communication networks. *IEEE Trans Commun* 1974;22(5):589–605.
17. Liu Y, Trivedi KS. A general framework for network survivability quantification. In: *Proceedings of the 12th GI/ITG Conference on Measuring, Modelling and Evaluation Of Computer and Communication Systems (MMB) together with 3rd Polish-German Tele-traffic Symposium*. 2004.
18. Ford LR, Fulkerson E. *Flows in networks*. Princeton (NJ): Princeton University Press; 1962.
19. Barlow RE, Proschan F. *Mathematical theory of reliability*. New York: John Wiley & Sons, Inc.; 1965.
20. Dieudonne J. *Foundations of modern analysis*. New York: Academic Press; 1968.
21. Tapia RA. The differentiation and integration of nonlinear operators. In: Rall LB, editor. *Nonlinear functional analysis and applications*. University of Wisconsin Mathematics Research Center Publication No. 26. New York: Academic Press; 1971.
22. Blanchard BS, Verma D, Peterson EL. *Main-tainability: a key to effective serviceability and maintenance management*. New York: John Wiley & Sons, Inc.; 1995.
23. Wood RK. Factoring algorithms for computing K-terminal network reliability. *IEEE Trans Reliab* 1986;R-35(3):269–278.
24. Cramer H, Leadbetter MR. *Stationary and related stochastic processes: sample function properties and their applications*. New York: John Wiley & Sons, Inc.; 1967.
25. Ramirez-Marquez JE, Gebre B. A Classification Tree Based Approach for the Development of Minimal Cut and Path Vectors of a Network. *IEEE Trans Reliab* 2007;56(3):474–487.
26. Ramirez-Marquez J, Coit D, Tortorella M. A generalized multi-state based path vector approach for multi-state two-terminal reliability. *IIE Trans Qual Reliab Eng* 2006;38(6):477–488.
27. Rey RF, editor. *Engineering and operations in the bell system*. 2nd ed. Murray Hill (NJ): AT&T Bell Laboratories; 1983.
28. http://www.informit.com/library/content.aspx?b=Signaling_System_No_7&seqNum=19 Accessed 2010.
29. Runge PK. Undersea lightwave systems. *AT&T Tech J* 1992;71(1):5–13.
30. Nguyen HT, Griffiths M, LaPierre S. *Systems and Methods for an Infrastructure Centralized Heartbeat*. US Patent 7,120,688. 2006.
31. <http://www.corp.att.com/history/nethistory/management.html>. Accessed 2010.
32. Chao C-W, Fuoco G, Kropfl D. FASTAR™ platform gives the network a competitive edge. *AT&T Tech J* 1994;73(4):69–81.
33. Cronk RN, Callahan PH, Bernstein L. Rule-based expert systems for network management and operations: an introduction. *IEEE Netw* 1988;2(5):7–21.
34. http://telephonyonline.com/residential_services/news/smart-grid-market-1229/. Accessed 2010.
35. McMasters AW, Mustin TM. Optimal interdiction of a supply network. *Nav Res Log Q* 1970;17(3):261–268.
36. Golden B. A problem in network interdiction. *Nav Res Log Q* 1978;25(4):711–713.
37. Ramirez-Marquez JE, Rocco CM. Stochastic network interdiction optimization via capacitated network reliability modeling and probabilistic solution discovery. *Reliab Eng Syst Saf* 2009;94(5):913–921.
38. Israeli E, Wood RK. Shortest-path network interdiction. *Networks* 2002;40(2):97–111.
39. Lim C, Smith JC. Algorithms for discrete and continuous multicommodity flow network interdiction problems. *IIE Trans* 2007;39(1):15–26.
40. Cormican KJ, Morton DP, Wood RK. Stochastic network interdiction. *Oper Res* 1998;46(2):184–197.

DETERMINISTIC DYNAMIC PROGRAMMING (DP) MODELS

DAVID K. SMITH
Mathematical Sciences in
SECaM, University of Exeter,
Devon, UK

In deterministic dynamic programming (DP) models, the transition between states following a decision is completely predictable. Knowledge of the triple “stage, state, decision” allows the modeler to know the subsequent state and the single-stage cost associated with the decision and transition. This knowledge makes the recurrence relation straightforward, as the cost of a decision is the appropriate combination of single-stage cost and the cost of the subsequent optimal policy. The optimal policy for the current state and stage can then be found by finding the best cost, taken over all possible decisions.

As a result, deterministic DP models may appear easy to solve. And in general they are. Deterministic models are normally simpler than models of a similar size where there are stochastic aspects. This does not make the models irrelevant or trivial. As has already been discussed, DP models of all kinds are not “black boxes.” There are many varieties of deterministic DP problems.

It is not possible to describe here every area of application of deterministic DP. Annually, many hundreds of research papers describe new areas, or refinements of old ones. Frequently these are based on the DP premise that a given problem can be broken into a sequence of similar, smaller versions of itself, using the principle of recursion. Computer scientists have recognized this premise in several problems of search and storage of strings of information. A search in a long string may identify a crucial point in that string, and that point defines two shorter (but essentially similar) strings to be searched. The problem known as the *longest common subsequence problem* compares

two such strings and tries to find those parts that are most like each other [1]. The problem has been used to store the changes to computer files (instead of storing two copies of a large file, store one copy and use DP to record the changes) and computer displays. Biologists studying DNA use DP to match subsequences of DNA from different sources, to find similarities between different organisms. Economists and meteorologists have used DP to match time series in order to identify periods of time with similar conditions, whether concerned with the stock market or the weather. Many areas of pattern recognition use DP to identify similarities [2].

As will be described below, DP principles are used in some everyday electronics. Matching of sequences of information by DP is part of some voice recognition systems, which try to match the inflexion of a caller saying (e.g.) a single digit with recorded versions of people speaking each of the 10 digits [3].

Also within the realm of computer science is the problem of matrix multiplication. To multiply two matrices together it is necessary to perform multiplications and additions for every cell in the final matrix. The larger this is, the greater the amount of work. Multiplying three or more matrices will produce intermediate answers. Using DP, one can decide on the order of calculating these intermediate results, so as to minimize the number of numerical operations that are needed [4].

The classic “traveling salesperson problem,” which requires the identification of a constrained optimal route through several nodes (cities) can be formulated as a DP problem. The size of practical examples of this problem means that the formulation is a tool for searching for the solution, and not a solution method in itself.

The formulations of these problems are characterized by the method of calculating the single-stage costs, the form of the transition, and the nature of the recurrence relation. In the sections that follow, some well-known examples of the varieties of these forms will be described. In some cases,

the DP model is useful for thinking about the problem (as mentioned in the preceding paragraph about the traveling salesperson problem) and there are alternative computational techniques available. Most of the examples below were developed by Bellman and can be found in his books, for example [5].

THE STAGECOACH PROBLEM (SHORTEST PATH)

Over the last few years, the use of satellite navigation systems and route-finding computer programs (or web sites) has increased immensely. The logic behind such systems is based on a basic DP model. This model is based on a simple piece of reasoning, although in practice the implementation may be complex. Suppose the traveler wishes to find the best (meaning least distance) route from place A to place B. Then the journey will start with a journey to place C, different from A. Place C could be one of several possible places, $C_1, C_2, C_3, \dots, C_r$. These can be regarded as a set of places C_i , where i varies from 1 to r . The traveler knows the distance from A to each of these places. Bellman's principles show that following the selection of a place C_i , the traveler must follow the best route from C_i to B. So the distance of interest for the traveler from A to B via C_1 is the distance from A to C_1 added to the length of the best route from C_1 to B. Similarly for each of the possible places $C_1, C_2, C_3, \dots, C_r$. Given these r possible routes, and their distances, the traveler's policy is to choose the shortest route. This assumes that the traveler has already calculated the length of the best route from each of the possible places $C_1, C_2, C_3, \dots, C_r$. This can be done in the same way, reasoning that after C_i there will be a next place and this can be selected to minimize the distance from C_i to B. That journey will require one fewer stage than the journey from A, and so the calculation will be simpler. Repeatedly considering the next place will eventually lead the traveler to consider places for which the best place to go to next is place B, the chosen destination.

Different applications will need different sets of places. The options for a motorist,

a cyclist, or a walker will be road junctions. Users of public transport may consider places where it is possible to change between forms of public transport. For a long journey, there may be thousands of possible places to consider.

So, in words, the value of distance from A to B via C_i equals distance from A to C_i plus optimal distance from C_i to B. The optimal distance from A to B is found by taking the minimum over all possible C_i of (distance from A to B via C_i).

Considering this in DP terms, the states and decisions have been identified, along with the recurrence relation in words. Stages have not yet been identified. One way of identifying stages is to define place A as stage 1, and make place B as stage M, where M is some large number, say 999. All the possible places C_i become states for stage 2, all the places that can be reached from any of the C_i become states of stage 3, and so on. The list of states is a list of those places that might be visited on a journey from A to B. Of course, place B is in the list; it may appear before stage M, but in that case the distance from B to B is 0—the optimal decision when state B is reached is to stay put.

The formal statement of the mathematical model is as follows:

$$f_n(s) = \min_x (d(s, x) + f_{(n+1)}(x)),$$

where s represents the state at stage n , the variable x identifies the possible states that can be considered at stage $(n + 1)$ (with the decision being made to go to x) and $d(s, x)$ is the distance from s to x , the single-stage cost. The value of the function $f_n(s)$ represents the shortest distance from s to the destination B. The “boundary conditions” are that any distance from B to itself has value 0. To calculate the value of $f_n(s)$ for other states, one evaluates the function for those places where there is a direct link from s to B, then uses these values to find the function value for places where there is one intermediate stop, and so on back to place A and $f_1(A)$.

This model is often called a *stagecoach model* because of the explanatory model devised by Harvey Wagner in the 1960s. He imagined that the decision maker was

a traveler in the nineteenth century, in the United States of America. The traveler needed to make a journey from one state to another, some distance away. To make this journey, the traveler would go by stagecoach from one state to an adjacent one. In each state, it would be necessary to decide which state to visit next. Instead of describing the single-stage cost as the distance, Wagner imagined that the traveler bought some form of travel insurance and the cost of this was the value used in $d(s, x)$ in the expression above. The combination of the insurances for each decision yielded an insurance policy, which, of course, could be described as the optimal policy.

It is worth noting that all deterministic DP problems can be rearranged to be like the stagecoach problem. Because a decision is deterministic, it can be represented as the transition between two states, with an associated cost. In spite of this, it is normal to consider variations as separate ways of formulating problems, as explained in the later part of this article.

THE KNAPSACK PROBLEM

Another classic deterministic DP problem is the knapsack problem. This model is concerned with maximizing an objective, usually considered to be the value of the “knapsack.” The decision is about which items to pack, from a large set of items, which are too numerous and bulky for all to be packed in the limited space (The name “knapsack” has stuck with this problem, even though the model has wider application.). Suppose that there are N items that can be packed in the limited space. Each item i has a value V_i and weight W_i , and the need is to select whether or not to pack each item and not exceed a maximum total weight, MAXW. The total value of the load is the sum of the values of the items that are packed.

DP imagines that the decisions about what to pack can be made sequentially. At the beginning, no decisions have been made and there is a possible weight MAXW available. Then the decision is made whether or not to pack item number 1. After that, there will

be either weight MAXW or weight MAXW - W_1 available for packing. The decision about item 1 has happened, and the problem progresses to a decision about item 2. It will be necessary to consider what to do with each possible value of the remaining weight available. Once this has been done, the focus moves to item number 3 and so on. Considered as a DP model, the stages will be the number of decisions that have been made, starting from 0 and ending at N . The state of the formulation will be the weight that is available (in the hypothetical knapsack) for packing. The single-stage cost is the value of the decision, which is 0 if an item is not selected, and the worth of the item if it is selected. The objective function will be the total value of a knapsack packed optimally with the remaining items, given the weight available.

This leads to a recurrence relation

$$f_n(w) = \max(0 + f_{n+1}(w), V_n + f_{n+1}(w - W_n)),$$

(provided that w is at least as large as W_n) where w is the weight available after n items have been considered. The two expressions on the right-hand side represent the two possible situations that could occur after the decision. The first corresponds to not packing item n , the second to packing it, thereby gaining the value and losing the capacity. The value of $f_{N+1}(w)$ is 0 for any value of w , showing that there is no value in any space that is left when $N + 1$ decisions have been made.

As an example, suppose there are three items, with weights 5, 4, and 2 units, and MAXW is 6. The three items have values 25, 18, and 8 respectively.

Then $f_4(w) = 0$ for all w .

From the recurrence relation, $f_3(w) = \max(f_4(w), 8 + f_4(w - 2))$ and so $f_3(0) = 0$, $f_3(1) = 0$, and $f_3(w) = 8$ when $w \leq 2$. This means, obviously, that when w is 2 or more, the best thing to do is to pack one item of type 3.

Again, $f_2(w) = \max(f_3(w), 18 + f_3(w - 4))$ and so $f_2(0) = f_2(1) = 0$, $f_2(2) = f_2(3) = 8$, $f_2(4) = f_2(5) = 18$, and $f_2(6) = 26$. This means that when there is insufficient space for an item of type 2, the best thing is to ignore it

and follow the optimal policy for the next stage; when there is space for an item of type 2, then pack it, and if there is enough space left over, then pack an item of type 3.

Finally, $f_1(6) = \max(f_2(6), 25 + f_2(1)) = \max(26, 25 + 0) = 26$ corresponding to packing items 2 and 3, even though these were not the most valuable items, nor did they give the largest value per unit weight.

In this case, the optimal decision was described with each stage of the problem, and in a small problem this will be sufficient. Generally, the record of optimal decisions and costs can be kept together as two tables. Each will have a column for every stage and a row for every discrete value of space that could be achieved. The entry in one will be the value of f , the other the optimal decision [The optimal decision could also be recovered from the table of optimal values following the logic that the recurrence $f_n(w) = \max(0 + f_{n+1}(w), V_n + f_{n+1}(w - W_n))$ means that $f_n(w)$ is either the same as $f_{n+1}(w)$ corresponding to not packing item n , or equal to $V_n + f_{n+1}(w - W_n)$ which means that item n is packed.].

This example was simple and straightforward, and readers will have worked out the best way of packing this small knapsack without needing so many calculations. However, when there are tens or hundreds of items, and a significant number of them need to be packed, the dynamic programming approach is worthwhile.

The simple model can be extended in several ways. It may be possible to pack multiple copies of an item in the knapsack. The value may be proportional to the number of copies packed, or there may be some other relationship between the value of a set of items and the number of items in the set. Economists have used the knapsack model when the value follows the law of diminishing returns; packing one spare set of clothes will have a value, but packing a second spare set will have a reduced value. (This phenomenon has been noted in treating the placing of advertising budgets as knapsack problems. Doubling the expenditure on advertising in a particular medium does not double the expected impact.) Another extension is when there are other constraints on the knapsack besides the

weight; the container to be packed into has two (or more) dimensions, such as weight and volume, and each item takes up some of each. Then the functions f will have one parameter for each dimension, each representing the amount of that resource dimension still available.

PRODUCTION MODELS

In many industrial processes, a machine can be used to produce a variety of products. By varying the fillings, different types of chocolate bars can be made on the same production line. By changing the quantity of the ingredients, different sizes of bars can be made on the same line. This is obviously of benefit to the producer, as it means that an expensive machine can be used efficiently. However, the saving in capital and maintenance costs compared with having several production machines is balanced by the expense and time of changing the nature of the item being made. So long production runs are desirable, but these mean that items have to be stored until they are needed.

One of the earliest production processes to be studied using a DP model was that for scheduling production for several future time periods, when the demand for the item was known exactly. Here, the time period is a month, but this is arbitrary. At the start of each month, a decision has to be made about whether to make any items in that month—and if so, how many—or not. Zero production implies that there are enough items in stock to supply all the orders for the month. Zero production has no production cost that is linked to the problem, while producing any number of items incurs the cost of starting a production run and the cost of making those items.

The time periods provide an obvious way of defining the stages. One can imagine a manager inspecting the inventory at the start of each month, counting how many items there are, and deciding whether or not to produce any more. For simplicity, one assumes that the items that are needed in 1 month can be produced in that month. If there are items in stock, then they will be used to meet that

month's demands, and any surplus will be held until the following month, and be subject to a charge for storage. So, in this model, the state is identified with the number of items found at the start of the month (or stage), and the decision is the number to produce, including the option to produce nothing. The single-stage cost will be the storage cost for the items that need to be held to the following month, plus the cost of production if it is appropriate. Letting h_n be the cost of storage per item from month n to month $n + 1$, demand in month n be D_n , and $A_n + C_n x$ be the production cost for x items when x is not 0, one has:

$$f_n(s) = \min(h_n(s - D_n) + f_{n+1}(s - D_n), \\ A_n + C_n x + h_n(s - D_n + x) \\ + f_{n+1}(s - D_n + x))$$

as the cost of the optimal policy from the start of month n , with s items in store. In the expression, the minimum is taken over all values of x such that $(s - D_n + x)$ is not less than 0 and x should not be larger than the amount that has still to be produced.

(Alternatively, one can combine the two alternatives into one expression, by defining a cost function $TC(n, x, s)$, which is equal to the cost of producing x items in month n plus the cost of holding $(s - D_n + x)$ to the next month. Then the recurrence becomes

$$f_n(s) = \min(TC(n, x, s) + f_{n+1}(s - D_n + x)),$$

taken over the same values of x as before.)

At the end of the planning period, at the end of N months, any value $f_{N+1}(s)$ is 0.

There may appear to be a great many calculations, but the problem is surprisingly simple. If there are no items in stock, then the decision is limited to a small number of alternative values. The production quantity should be enough for the current month, or the current month and the whole of one or more subsequent months. If there are items in stock, the rule just stated ensures that there will be enough for the current month. This result can be proved formally, but it is enough to follow this informal argument. Suppose it were ever optimal to produce items

in a month, which began with items in stock at the start. Now suppose that there had been one item less at the start of month i ; the cost of producing an extra item will be C_i . If this is less than the total cost associated with having one less item that was made in an earlier month and held in stock, then it is better to have fewer items in stock. The argument can be continued until there were no items in stock at the start of month i . If C_i is bigger than the total cost, then the number in stock should be increased to be at least as large as the demand for month i , which will not only reduce the cost per item, but also remove the fixed cost A_i of starting a production run in month i .

LINE BREAKING IN TeX

TeX and LaTeX are widely used computerized document preparation systems, which process text materials into an appropriate format for printing. The systems are used by many scientists because the systems were designed to produce attractive layouts for mathematical content. An important part of such a system is the ability to break a paragraph into lines. Many people use a word processor with the requirement that the text is justified (the left- and right-hand sides of the page appear as a straight line). When this happens, the space between words is increased on some lines, and if the appropriate option has been selected, some words may be hyphenated. Sometimes, if the lines are short, or the words are long, there will be many large spaces on the page, resulting in an ugly page.

In the development of TeX (LaTeX is similar) the creators decided to use a more sophisticated method of "Hyphenation and Justification" than the usual word processor algorithm, which lays out one line at a time, without referring to any other part of the paragraph. Their work is reported in [6] (and also in Chapter 14 of the TeXbook [7]) and it is based on a deterministic DP model. The input to the model is an entire paragraph; within this, all the possible places (break points) where a line of text could be broken are then identified. These will be spaces

between words and permitted hyphenation points within words. A paragraph begins and the model examines each possible decision for the break point at the end of the first line. The decision leads to a new state, described by the place where the next line will start and whether or not there is a hyphen. The process is repeated for each successive line, and the lines determine the stages of the DP formulation. The decision also generates a cost, which is a measure of how ugly the line will look by itself together with the previous line. The paper [6] discusses how to measure ugliness in terms of extra space between words, reduction of space between words, the need to hyphenate words, and whether or not the previous line ended with a hyphen. The ugliness of a paragraph is then found by combining the uglinesses of the lines in it.

Unlike most word processors, the algorithm in TeX does not break any lines in a paragraph until the whole paragraph is complete, and then determines the optimal (minimizing ugliness or maximizing attractiveness) layout for the whole paragraph.

This DP model is a shortest path problem in a slight disguise. The value given to a paragraph's appearance is found by combining the values for each line, and each line is determined by the states (break points) at either end. As an example, the paragraph below shows the break points as hyphens in a short paragraph, which was to fit into a small column, so that the first break appeared in the first "construction." The optimal layout for this paragraph is shown below.

In the United States, vast amounts of construction waste are produced every year. Construction waste accounts for a significant portion of the municipal waste stream of the United States.

In the United States, vast amounts of construction waste are produced every year. Construction waste accounts for a significant portion of the municipal waste stream of the United States.

JOB SCHEDULING

Many people and organizations have problems of scheduling work. DP is useful for solving some types of problems, which involve

schedules. The idea behind the use of DP in scheduling is that the optimal schedule for a set of jobs is made up of the optimal schedule for a subset of jobs followed by the optimal schedule for the complement of that subset given the later starting time(s) that results from having done some of the jobs. And thus one has the seeds of a DP formulation, because the problem has been broken into smaller parts in sequence. Consider a simple example of finding the best schedule for five jobs on one machine with the objective of minimizing the mean tardiness (or the total tardiness). Tardiness is measured as 0 if the job is completed early and if the job is late, it is the number of time units (using minutes here) late that the job is finished after its deadline.

Suppose that the jobs have deadlines as shown in the table below, and that the jobs take the different processing times listed.

Job	1	2	3	4	5
Processing time (min) $[p_j]$	24	8	24	31	18
Deadline (min from now) $[d_j]$	53	29	39	50	69

To consider this problem using DP, one needs to define the essential features of any such formulation. Therefore define the states of the system, a series of stages, an objective function, and a recurrence relation that connects the successive states and stages. The stages are comparatively obvious. The person responsible for devising the schedule has to decide which jobs to schedule next, and the decisions are effectively taken as each job finishes. Therefore, there are five stages corresponding to the need to schedule one, two, three, four, and five jobs respectively. At each stage there will be a number of possible states, corresponding to the possible sets of jobs that have been completed by that stage. The decision that answers the question "What next?", follows from considering the list of jobs that are waiting to be scheduled. The objective of the model is to minimize the mean tardiness, and as this is simply a multiple of the total tardiness, it is a straightforward transformation to the

problem to make minimizing the total as the objective.

When the scheduler decides on a job to be scheduled next, this decision fixes the tardiness for that job because the set of jobs which have already been finished has determined the time when the decision is being taken, the duration of the job being scheduled determines when it will finish, and the due date for the job will determine the contribution of that job to the total tardiness function. And thus it is possible to write down a recurrence relationship.

Suppose that the set of jobs to be scheduled is a set S , and that the decision is considered once all the jobs in a subset J of S have been completed. The time now will be $T(J) = \sum_j p_j$. The stage n will be the number of jobs already scheduled, that is, the total number of jobs that are in J . Take a job i as the next job to be completed. It will finish at time $T(J) + p_i$. If this is less than d_i , there is no tardiness. If not, and if job i is completed too late, it will have a tardiness $T(J) + p_i - d_i$. Mathematically, the tardiness

is written as $\max(0, T(J) + p_i - d_i)$. Then the total tardiness which follows from scheduling one of the jobs not in J , job i say, will be

$$\max(0, T(J) + p_i - d_i) + f_{n+1}(J + \{i\}).$$

All that is needed to find $f(J, T)$ is to look at all possible values of i , calculate the tardiness contribution, and select the job i , which gives the smallest value. So, one has

$$f_n(J) = \min[\max(0, T + p_i - d_i) + f_{n+1}(J + \{i\}, T + p_i)]$$

taken over all i that are not in J .

Here the tardiness gives the single-stage “cost” and this is followed by the optimal policy for the set of jobs that will have been scheduled at the time when the next decision is made. It is useful to record the time ($T(J)$) as well as the set (J) for ease of calculation. To illustrate this formulation, suppose stage 1 has been calculated to be as shown in table below.

The Set of Jobs, J	{2,3,4,5}	{1,3,4,5}	{1,2,4,5}	{1,2,3,5}	{1,2,3,4}
Job not in J	1	2	3	4	5
$T(J)$	81	97	81	74	87
$F_1(J)$	52	76	66	55	36

Then the calculation for f_2 will be based on the possible sets J with three finished jobs; one such will be $J = \{1, 3, 4\}$ and this will have $T = 79 (= 24 + 24 + 31)$. If job 2 is done next, it will finish at time 87, and be 58 min tardy, and the table above shows that the contribution from scheduling job 5 last will be 36 min. So, scheduling job 2 next contributes $58 + 36 = 94$ min. Alternatively, scheduling job 5 next contributes $28 + 76 = 104$ min. There are 10 possible sets of 3 jobs at stage 2 to consider; 10 at stage 3; 5 at stage 4; and 1 (no jobs scheduled) at stage 5. The best schedule is 2-3-1-5-4.

This scheduling problem is one for which the curse of dimensionality applies. The number of states that have to be examined

is 2 raised to the power of the number of jobs in the set S . So with 5 jobs in S , there are 32 states to be considered. With twice that number, there would be 1024 states. Even so, these numbers are considerably less than the number of schedules possible, which would be 120 (for 5 jobs) and 3,628,800 (for 10 jobs).

REFERENCES

1. Bergroth L, Hakonen H, Raita T. A survey of longest common subsequence algorithms. SPIRE'00. Washington, DC: IEEE Computer Society; 2000. pp. 39–48.
2. Warwick J, Phelps B. An application of dynamic programming to pattern recognition. J Oper Res Soc 1986;37:35–40.

3. Furtuna TF. Dynamic programming algorithms in speech recognition. *Rev Inform Econ* 2008;2:94–98.
4. Dorn F. Dynamic programming and fast matrix multiplication. Volume 416, *Lecture notes in computer science*. New York (NY): Springer-Verlag; 2006. pp. 280–291.
5. Bellman RE *Dynamic programming* New York (NY): Dover Publications; 2003.
6. Knuth DE, Plass MF. Breaking paragraphs into lines. *Software Pract Ex* 1981;11:1119–1184.
7. Knuth DE. *The TeXbook*. Reading (MA): Addison-Wesley; 1984 (and later editions).

DETERMINISTIC GLOBAL OPTIMIZATION

PANOS M. PARDALOS
Department of Industrial and
Systems Engineering and
Biomedical Engineering,
University of Florida, Center for
Applied Optimization,
Gainesville, Florida

QIPENG P. ZHENG
Department of Industrial and
Management Systems
Engineering, West Virginia
University, Morgantown, West
Virginia

ASHWIN ARULSELVAN
DIMAP, Warwick Business School,
University of Warwick,
Coventry, UK

INTRODUCTION

Global optimization is a very active and exciting branch of the optimization tree. It mainly deals with problems that involve nonconvexity either in the objective function or in the constraints. The nonconvexity of the functions makes the problem intrinsically hard due to the presence of multiple local optima and it is also the reason for it to be NP-hard. Global optimization is a very general optimization area that involves both continuous and discrete variables. As will be seen in the article, most discrete optimization problems can be transformed to a continuous problem by adding some nonconvex constraints, for example, the complementarity constraints. Global optimization includes a variety of problems that arise in many applications such as economics, biology, medicine, management, sciences, engineering, etc. Also numerous algorithms were proposed and studied extensively to solve these problems. This is an area with abundant scope for research for new problems,

applications, and algorithms. In this survey, we focus on deterministic computational methods in global optimization.

This article adopts a problem-based organization, where we focus on a specific type of problem in each section and several available deterministic computational methods are discussed for each problem. In the section titled “Lipschitz Optimization,” we show the Lipschitz optimization problem and solution methods. In the section titled “DC Optimization,” we discuss DC optimization, its optimality conditions, and a few solution algorithms. The section titled “Monotone Optimization” discusses monotone optimization. In the section titled “Quadratic Programming,” we discuss quadratic programming and optimality conditions. Also several algorithms for both equality-constrained and inequality-constrained quadratic programs are discussed. In the section titled “General Concave Minimization,” we discuss general concave optimization. In the section titled “Linear Complementarity Problem, and Mathematical Programs with Complementarity Constraints,” we focus on linear complementarity problems and briefly mention mathematical programs with complementarity constraints. In the section titled “Discrete Optimization Problems,” we discuss the relationship between discrete optimization and continuous optimization. A few transformations from discrete to continuous optimization are also shown. And, the final section concludes the article.

LIPSCHITZ OPTIMIZATION

In Lipschitz optimization problem (LOP), we deal with real valued objective and constraint functions that are Lipschitz continuous over a compact set with known Lipschitz constants. Consider the following LOP:

$$\text{Minimize } f_0(x) \quad (1)$$

$$\text{s.t. } x \in P, \quad (2)$$

$$f_i(x) \leq 0, \quad i = 1, \dots, m, \quad (3)$$

where $f_i(x)$, $i = 0, \dots, m$ are Lipschitz continuous on P , such that they satisfy

$$|f_i(x_1) - f_i(x_2)| \leq L_i \|x_1 - x_2\|,$$

for all $x_1, x_2 \in P$, with the Lipschitz constant $L_i = L_j(f_i, P)$, $i = 0, \dots, m$. Let us define the set C as

$$C = \{x \in P : f_i(x) \leq 0, \quad i = 1, \dots, m\},$$

which we assume is compact.

Algorithm:

(Initialization)

Step 1 : $m_X \leftarrow \frac{(a+b)}{2}$, $Q \leftarrow \{a, b, m_X\}$, $\alpha = \min\{f(x) : x \in Q\}$ (α is an upper bound on the optimal solution)

Step 2 : Find $x^1 \in Q$, such that $\alpha = f(x^1)$

Step 3 : Set $\mu(X) \leftarrow \max\{\max\{f(a), f(b)\} - L\|b - a\|, f(m(X)) - L\frac{\|b-a\|}{2}\}$ (lower bound on the optimal value)

Step 4 : Set $k \leftarrow 1$, $\mathcal{M} \leftarrow X$, $\mu \leftarrow \mu(X)$

(Iteration k)

Step 5 : If $\mu = \alpha$ return x^k as the optimal solution

Step 6 : Find $b_j - a_j = \max\{b_i - a_i, \quad i = 1 \dots n\}$

Step 7 : Set $a^1 \leftarrow a$, $b^1 \leftarrow \left\{b_1, \dots, \frac{(b_j+a_j)}{2}, \dots, b_n\right\}^T$

Step 8 : Set $a^2 \leftarrow \left\{a_1, \dots, \frac{(b_j+a_j)}{2}, \dots, a_n\right\}^T$, $b^2 \leftarrow b$

Step 9 : Set $R_1 = \{x : a^1 \leq x \leq b^1\}$, $R_2 = \{x : a^2 \leq x \leq b^2\}$

Step 10: Set $m(R_1) = \frac{(a^1+b^1)}{2}$, $m(R_2) = \frac{(a^2+b^2)}{2}$

Step 11: Set $\mu(R_1) = \max\left\{\mu(X), \max\{f(a^1), f(b^1)\} - L\|b^1 - a^1\|, f(m(R_1)) - L\frac{\|b^1-a^1\|}{2}\right\}$

Step 12: Set $\mu(R_2) = \max\left\{\mu(X), \max\{f(a^2), f(b^2)\} - L\|b^2 - a^2\|, f(m(R_2)) - L\frac{\|b^2-a^2\|}{2}\right\}$

Step 13: Set $Q \leftarrow \{x^k, b^1, m(R_1), a^2, m(R_2)\}$, $\alpha \leftarrow \min\{f(x) : x \in Q\}$

Step 14: Set $k \leftarrow k + 1$, $\mathcal{M} \leftarrow \mathcal{M} \setminus X \cup \{R_1, R_2\}$, $D \leftarrow \{X \in \mathcal{M} : \mu(X) \leq \alpha\}$

Step 15: Set $\mu = \begin{cases} \min\{\mu(X) : X \in D\} & \text{if } D \neq \emptyset \\ \alpha & \text{if } D = \emptyset \end{cases}$

Step 16: Choose $X \in D$ with $\mu(X) = \mu$ and update a, b such that $X = \{x : a \leq x \leq b\}$

Step 17: Find $x^k \in Q$, such that $\alpha = f(x^k)$

Step 18: Goto step 5

Partition Strategies

We first discuss a branch and bound method for an LOP, where the feasible set is a rectangle $\{x \in \mathcal{R}^n : a \leq x \leq b, a < b \in \mathcal{R}^n\}$. Consider the following LOP with rectangular feasible region.

$$\text{Maximize } f(x) \quad (4)$$

$$\text{s.t. } x \in X = \{x \in \mathcal{R}^n : a \leq x \leq b, \quad a < b \in \mathcal{R}^n\}, \quad (5)$$

where $a = \{a_1, a_2, \dots, a_n\}^T$, $b = \{b_1, b_2, \dots, b_n\}^T \in \mathcal{R}^n$, $a < b$, and $f : X \rightarrow \mathcal{R}$ with the Lipschitz constant $L = L(f, X)$. The algorithm bisects the rectangles at the midpoint of the longest edge.

The convergence of the algorithm is a consequence of the following lemma [1].

Lemma 1. *Successive bisections of n -rectangles at the midpoint of the one of its longest edges is exhaustive.*

We discuss a partition method in which the feasible region is an n -dimensional polytope X to get the following global optimization problem:

$$\text{Minimize } f(x) \quad (6)$$

$$\text{s.t. } x \in X, \quad (7)$$

where $f : X \rightarrow \mathcal{R}$ is a Lipschitz function with the Lipschitz constant $L = L(f, X)$ and X is an n -dimensional polytope. Let $S_0 \supset X$ be an n -simplex that tightly approximates X . We employ the following partitioning strategy to solve the above problem.

We partition S_0 into n -simplices S with known vertex set $V(S)$, such that $S_0 = \bigcup_{i=0}^n S_i$, $S_i \cap S_j = \emptyset$, for $i \neq j$. For each simplex S , we determine a lower bound

$$\mu(S) \leq \min \{f(x) : x \in S \cup D\}.$$

We determine a nonempty set $Q \subset D$ and determine the upper bound on the optimal value $\alpha = \min\{f(x) : x \in Q\}$. After deleting the simplices S with $\mu(S) \geq \alpha$ or $S \cap D = \emptyset$, we proceed to the next iteration. For details of the implementation of the lower bound and upper bound computation, we refer the reader to Horst *et al.* [1] and Pintér [2]. There are several techniques to estimate the lower bound.

There are several other partition algorithms that are available in the literature. We refer the reader to Hansen and Walster [3], Neumaier [4], Ratschek and Rokne [5], and Kearfott [6] for a detailed list of all of these and details about their implementation.

Covering Method

Numerous algorithms have been proposed to solve the Lipschitz optimization problem $\max\{f(x) : x \in X, a \leq x \leq b\}$ to an ϵ optimal value. We try to find $x_{\text{opt}} \in X$, $f_0(x_{\text{opt}}) \leq f_0^* + \epsilon$, where f_0^* is the global optimal value and

$f_0(x_{\text{opt}})$ is the ϵ -optimal value and x_{opt} is the corresponding point. An algorithm is ϵ -convergent if it solves this above problem in finite number of iterations.

The following theorem is established from the Lipschitz property of the objective function f .

Theorem 1 [7]. *There is no algorithm with finite convergence for the optimization problem, where we maximize a Lipschitz function with known Lipschitz constant over a hyper-rectangular region, such that it attains the global optimal solution x^* and prove the optimality in finite number of iterations unless f satisfies the following:*

$$\begin{aligned} \exists \delta > 0 \text{ such that } f(x) &= f(x^*) - L|x - x^*|, \\ \forall x &\in [x - \delta, x + \delta] \cap [a, b]. \end{aligned} \quad (8)$$

The implication of the above theorem is that around the neighborhood of x^* there is a sawtooth with slopes $-L$ and L to its left and right, respectively.

Piyavskii's algorithm maximizes a univariate Lipschitz's function over a hyper rectangular feasible region. The function f is upper bounded by a piecewise linear function F . If we could find the function F and a point x_{opt} such that $f(x_{\text{opt}})$ is close to the optimal value of f^* and the function F with maximum value as low as possible, we could provide an ϵ -optimal solution to our Lipschitz optimization problem. This method is called *covering method*, since it constructs a sequence of upper envelope functions $F_k(x)$, where k denotes the k th iteration, that satisfy

$$\begin{aligned} F_k(x) &\geq f(x), \quad \forall x \in [a, b], \quad k = 1, 2, \dots, \\ F_k(x) &\geq F_{k+1}(x), \quad \forall x \in [a, b], \quad k = 1, 2, \dots, \\ F_k(x_j) &= f(x_j), \quad j = 1, 2, \dots, n, \end{aligned}$$

where $\{x_j : j = 1, 2, \dots, n\}$ is a sequence of feasible solutions and the upper envelope function $F_k(x)$ is defined in the following theorem.

Theorem 2 [8,9]. *The one-step optimal upper bounding function F_k for f on the interval $[a, b]$ is*

$$F_k(x) = \min_{i=1,2,\dots,k} g_i^P(x), \quad (9)$$

where $g_i^P(x) = f(x_i) + L|x - x_i|$.

This is because of the property of Lipschitz function, $|f(x) - f(x_i)| \leq L|x - x_i|$, $\forall x \in [a, b]$, and then $g_i^P(x)$ is greater or equal to $f(x)$ for all i and $x \in [a, b]$. Hence, the minimum of all $g_i^P(x)$'s is greater than $f(x)$, which means an upper envelope of $f(x)$.

The shape of this upper bounding function is a sawtooth. One consequence of the

above theorem is that within a subinterval $[c, d] \subset [a, b]$, the best upper bound of the function f depends only on the values of f at the evaluation points within the subinterval $[c, d]$, but the best bound of $[c, d]$ could be dependent on points outside $[c, d]$. The peak point of the sawtooth cover F_k between two consecutive evaluation points x_i and x_{i+1} in the interval $[a, b]$ is given by $\frac{x_i + x_{i+1}}{2} + \frac{f(x_{i+1}) - f(x_i)}{2L}$.

Algorithm (Piyavskii)

Initialization

Step 1 : $k \leftarrow 1, x_k \leftarrow \frac{a+b}{2}$

Step 2 : $x_{\text{opt}} \leftarrow x_k$

Step 3 : $F_k(x_{\text{opt}}) \leftarrow f(x_{\text{opt}}) + L \frac{(b-a)}{2}$

Step 4 : $F_k(x) \leftarrow f(x_k) + L|x - x_k|$

Iteration

Step 5 : **If** $F_k(x_{\text{opt}}) - f(x_{\text{opt}}) \leq \epsilon$ **Then** return x_k as optimal solution

Step 6 : $x_{k+1} \leftarrow \arg \max_{x \in [a, b]} F_k(x)$

Step 7 : **If** $f(x_{k+1}) > f(x_{\text{opt}})$ **Then** $x_{\text{opt}} \leftarrow x_{k+1}$

Step 8 : $F_{k+1}(x) \leftarrow \min_{i=1,2,\dots,k+1} F_{k+1}(x)$

Step 9 : $F(x_{\text{opt}}) \leftarrow \max_{x \in [a, b]} F_{k+1}(x)$

Step 10: Goto Step 5

The peak point of a sawtooth is an upper bound of the function f and at each iteration the largest upper bound or highest peak is identified and chosen as the evaluation point. In other words, the sawtooth with the highest peak point is broken down at each iteration. This similarity between branch and bound method and Piyavskii's algorithm was observed by Horst and Tuy [8]. The proof of convergence of Piyavskii's algorithm could be found in Piyavskii [9], Hansen and Jaumard [10], and Pintér [2]. Also this algorithm can be easily generalized to higher dimension problems, but the problems become much harder to solve.

DC OPTIMIZATION

The objective functions, or the constraint set, that are not convex of many optimization problems could be written as the difference of two convex functions (DC functions). In

this way, we try to overcome the inherent difficulty in the optimization problems due to the nonconvexity and develop efficient methods to solve them.

A function $f : X \rightarrow \mathcal{R}$, where X is convex, is a DC function on X , if $\forall x \in X$ we have

$$f(x) = g(x) - h(x),$$

where g, h are convex functions on X .

An optimization problem is a DC optimization problem if it could be written in the following form:

$$\text{Minimize } f_0(x) \quad (10)$$

$$\text{s.t. } x \in X \subset \mathcal{R}^n, \quad (11)$$

$$f_i(x) \leq 0, \quad i = 1, \dots, m. \quad (12)$$

Also, $f_i(x)$, ($i = 0 \dots m$) are DC functions and X is a compact convex set. Any DC problem could be reduced to a canonical DC (CDC) problem, where a linear function is

minimized over the intersection of a compact convex set, X , and an open convex set. The open convex set could be represented by the reverse convex constraint,

$$h(x) \geq 0,$$

where $h : \mathcal{R}^n \rightarrow \mathcal{R}$ is a convex function.

Optimality Conditions

We provide the optimality conditions for the DC optimization problems without the proofs, as the emphasis of the article is on computational methods and the proofs are available in Horst *et al.* [1] and Tuy [11]. Let us consider the DC optimization problem

$$\omega^* = \min g(x) - h(x) : x \in X, \quad (13)$$

where g and h are convex functions and X is a compact convex set.

Theorem 3. $x^* \in X$ is an optimal solution to Equation (13) if and only if there is a t^* such that

$$0 = \inf \{ -h(x) + t : x \in X, \\ g(x) - t \leq g(x^*) - t^* \}.$$

The geometric optimality conditions could be stated as follows.

Theorem 4. $x^* \in X$ is an optimal solution to Equation (13) if and only if

$$\text{epi } \bar{g} | X \subset \text{epi } h | X,$$

where $\bar{g}(x) = g(x) - (g(x^*) - h(x^*))$.

For any function f , we define $\text{epi } f | X = \{(x, t) \in \mathcal{R}^n \times \mathcal{R} : x \in X, f(x) \leq t\}$. Another way of stating the above theorem is as follows.

Theorem 5. $x^* \in X$ is an optimal solution to Equation (13) if and only if

$$\partial X_\epsilon h(x^*) \subset \partial X_\epsilon g(x^*), \quad \forall \epsilon \geq 0.$$

For any function f , we define $\partial X_\epsilon f(\bar{x}) = \{s : f(\bar{x}) - f(x) + s^T(x - \bar{x}) - \epsilon \leq 0, \forall x \in X\}$.

Branch and Bound

We discuss the branch and bound method for the following DC optimization problem in this section.

$$\text{Minimize } f_0(x) \quad (14)$$

$$\text{s.t. } f_i(x) \leq 0, i = 1, \dots, m, \quad (15)$$

$$x \in X \subset \mathcal{R}^n, \quad (16)$$

where $f_i(x) = g_i(x) - h_i(x)$, $i = 0, \dots, m$, are DC functions and X is a polytope. Let F be the feasible set of the above global optimization problem.

Initialize. Construct a simplex $S^1 \supseteq X$ and compute $l(S^1)$, the lower bound for f_0 over the set $S^1 \cap F$. Determine a set $M^1 \subset S^1 \cap F$. Compute the upper bound $g^1 = \{\text{Minimize } f_0(x) : x \in M^1\}$. Choose a point x^1 such that $f(x^1) = g^1$. Set, $l^1 = l(S^1)$, $R^1 = \{S^1\}$ and $k = 1$.

Iteration k . If $l^k = g^k$, then return $x^* = x^k$ as the optimal solution, else break down S^k into r disjoint simplices S_i^k , $i = 1 \dots r$, such that $S^k = \cup_{i=1}^r S_i^k$. For each $i = 1 \dots r$, compute lower bounds $l(S_i^k) \geq l^k$ for f_0 on $S_i^k \cap F$. Let $M_i^k \subset S_i^k$ and set $M^{k+1} = M^k \cup \{M_i^k, i = 1 \dots r\}$. Compute $g^{k+1} = \text{Minimize } f_0(x) : x \in M^{k+1}$, and choose $x^{k+1} \in M^{k+1}$, such that $f(x^{k+1}) = g^{k+1}$. Set $R^{k+1} = \{R^k \setminus S^k \cup S_i^k, i = 1 \dots r\}$. Delete all $S \in R^{k+1}$, where $l(S) \geq g^{k+1}$ and R^{k+1} be the remaining set. If $R^{k+1} = \emptyset$ then set $l^{k+1} = g^{k+1}$, else set $l^{k+1} = \{\min l(S) : \forall S \in R^{k+1}\}$, $S^{k+1} = \{\text{argmin } l(S) : \forall S \in R^{k+1}\}$, $k = k + 1$ and go to iteration k .

We refer the reader to Tuy [11] for the details of the implementation of simplex breakdown, computation of lower bound and upper bound in the above algorithm.

Outer Approximations

We present the sketch of an outer approximation method for solving the CDC optimization problem. The complete algorithm could be referred in Tuy [11,12] and Tuy and Horst

[13]. Consider the following CDC problem:

$$\text{Minimize } f(x) \quad (17)$$

$$\text{s.t. } x \in X \quad (18)$$

$$h(x) \geq 0, \quad (19)$$

where f is a linear function, X is a compact convex set, and h is a convex function. Let us define the set C as

$$C = \{x : h(x) \leq 0\}.$$

Let $\text{int } C$ refer to the interior of the set C . The regularity condition for the above CDC problem is stated as

$$\inf\{f(x) : x \in X \setminus \text{int } C\} = \inf\{f(x) : x \in X \setminus C\}. \quad (20)$$

Set $D(\gamma) = \{x \in X : f(x) \leq \gamma\}$. Under the above regularity assumption, solving the CDC problem is equivalent to finding:

$$\gamma^* = \inf \{\gamma : D(\gamma) \setminus C \neq \emptyset\}.$$

We begin the algorithm with a known feasible solution $\bar{x}$ and set $\gamma = f(\bar{x})$. As in many outer approximation, we determine a polytope P^1 that outer approximates the set $D(\gamma)$. Let V be the vertex set of the polytope P^1 . Now, we seek a point $z^1 \in P^1 \setminus C$ or prove that $P^1 \subset C$. This could be accomplished by checking whether $\{\max h(x) : x \in V\} \leq 0$. If $P^1 \subset C$, then we terminate the iteration. Otherwise, we improve z^1 to $x^1 \in \partial X \cap \partial C$, the intersection of the boundaries of X and C . If $x^1 \notin X$ and hence infeasible, we can separate z^1 from the set $X \supset D(\gamma)$ with a hyperplane $l(x) \leq 0$ and set $P^2 = P^1 \cap \{x : l(x) \leq 0\}$ and restart the iteration. If $x^1 \in X$, and we have $f(x^1) = \gamma^1 \leq \gamma$, then we have $f(x) \leq \gamma^1$ strictly separating z^1 from $D(\gamma^1)$. We set $P^2 = P^1 \cap \{x : f(x) \leq \gamma^1\}$ and restart the iteration. We either terminate at a polytope $P^k \subset C$, where the current γ satisfies $D(\gamma) \subset C$, as $P^k \supset D(\gamma)$ or we generate a sequence of polytopes $P^1 \supset P^2 \dots$, such that $P^k \supset D(\gamma^k)$ and $z^k \in P^k \setminus C$. The terminating γ value is the optimal value of the CDC.

Polyhedral Annexation Method (PA)

Polyhedral annexation method (PA) method is also referred to as the *dual method* or *inner approximation method*. Let E^* be the polar of a set E defined by

$$E^* = \{y \in \mathcal{R}^n : y^t x \leq 1, x \in E\}.$$

Now the dual feasibility problem is based on the fact that

$$D(\gamma) \subset C \Leftrightarrow C^* \subset D(\gamma)^*.$$

The dual methods can overcome the dimensionality problems that might arise in the primal methods. In the PA method, we generate a sequence of polyhedrons such that, $P_1 \subset P_2 \dots \subset C$. If a polyhedron $P_k \supset D(\gamma)$, then the current value of γ is the optimal value. If we find a point $x \in D(\gamma) \subset C$, we can find a better feasible solution and continue the procedure after updating the incumbent solution. The procedure is similar to the outer approximation method that we discussed, but for the dual. If S_k is the polar of P_k , then $S_k \supset C^*$. This could be inferred from the definition of the polar. So we generate a sequence $S_1 \supset S_2 \dots C^*$ and we stop when we find a $S_k \subset D(\gamma)^*$ or if a point $y \in C^* \setminus D(\gamma)^*$ is found. More details about the implementation of the algorithm could be found in Tuy [11,14].

Covering Method

Similar to Piyavskii's algorithm [9], while solving a maximization problem, $\max_{x \in X} f(x)$, some authors [15–18] proposed covering methods for a function with bounded second-order derivative, that is,

$$f(x) \leq f(y) + [\nabla f(y)]^T(x - y) + K\|x - y\|^2, \\ x, y \in X,$$

where X is the feasible region. By this assumption, the upper envelope function can be constructed as follows:

$$F_k(x) = \min_{i=1, \dots, k} g_i^B(x),$$

where $g_i^B(x) = f(x_i) + [\nabla f(x_i)]^T(x - x_i) + K\|x - x_i\|^2$. Baritomp and Cutler [16] generalized this to the following:

$$g_i^B(x) = f(x_i) + [\nabla f(x_i)]^T(x - x_i) + (x - x_i)^T H(x - x_i),$$

with a positive definite H . Sometimes, solving the maximization problem of the upper envelope function, $F_k(x)$, over the feasible region is as hard as solving the original problem. Blanquero and Carrizosa [19] proposed a new upper envelope function for the DC functions and showed that the above approach by Baritomba and Cutler [16] is just a special case of applying DC decomposition on $f(x)$. Suppose function $f(x)$ is a DC function, and then it can be decomposed as follows:

$$f(x) = f_1(x) + f_2(x), \quad \forall x \in X,$$

where $f_1(x)$ is a convex function and $f_2(x)$ is a concave function. By taking a first-order approximation of $f_2(x)$ at y , we can have

$$f(x) \leq f_1(x) + f_2(y) + \xi_2^T(y)(x - y),$$

where $\xi_2(y)$ is a subgradient of $f_2(x)$ at y . Then we can define that

$$g_i(x) = f_1(x) + f_2(x_i) + \xi_2^T(x_i)(x - x_i),$$

and obtain a new upper envelope function. It is interesting to note that if we decompose a function as

$$f(x) = x^T H x + [f(x) - x^T H x],$$

which is a DC decomposition, we will get the same upper envelope function as discussed by Baritomba and Cutler [16].

MONOTONE OPTIMIZATION

Nonconvex programs with special structures such as monotone functions and d.m. (difference of monotone) functions have been

given a lot of attentions in the recent past. Monotonic optimization arises in many engineering and economics fields and many nonconvex problems such as multiplicative programming, indefinite quadratic programming, and Lipschitz optimization [20]. For the interested readers, a number of recent surveys and works [20–23] are available on monotone optimization.

A monotone optimization problem could be stated as

$$\min f(x) : x \in G \cap H,$$

where f is monotone nondecreasing function, which is monotonely nondecreasing on every ray in the orthant (i.e., $f(y) \geq f(x)$ whenever $x \leq y$), and $G = \{x \in \mathbb{R}^n : g_i(x) \leq 0, i = 1 \dots l, x \in [a, b]\}$ and $H = \{x \in \mathbb{R}^n : h_i(x) \geq 0, i = 1 \dots k, x \in [a, b]\}$, where g_i and h_i are monotone nondecreasing lower and upper semicontinuous functions, respectively. G is a compact *normal set* defined by the property $x \leq y \in G \Rightarrow x \in G$ and H is a closed *reverse normal set* defined by the property $x \geq y \in H \Rightarrow x \in H$. A detailed theory on normal sets and lower semicontinuous functions could be found in Tuy [20]. Now we proceed to describe a popular algorithm, polyblock outer approximation, for solving d.m. problems. The idea behind the algorithm is to approximate the normal sets with *polyblocks* in $[a, b] \in \mathbb{R}_+^n$, defined as the union of intersection of finite set of boxes $[a, z] \subset [a, b], z \in T$ and the set T is called the *vertex set* of the polyblock. We define a vertex v^k as *proper* if it is not dominated by any other vertex in T , that is, $v^k \notin [a, v'], \forall v' \in T$. All other elements of T are *improper*. We refer the reader to Tuy [20] and Rubinov *et al.* [21] for more details about polyblocks and their properties. The idea behind the algorithm is to construct a series of polyblocks approximating $G \cap H \subset P_k \subset \dots \subset P_2 \subset P_1$.

Algorithm (Polyblock outer approximation)

Initialization

Step 1: Select a tolerance $\epsilon \geq 0$, set $k = 0$

Step 2: Take $P_k = [a, b], T_k = b$

Step 3: Let the current best feasible solution be $\bar{x}^k$ and the corresponding objective value (CBV) = $f(\bar{x}^k)$ (if no feasible solution is found, then CBV = ∞)

Iteration

Step 4: Let L_k^1 be the set of all vertices $v \in T_k \notin H$ and L_k^2 be the set of all vertices with $f(v) \leq CBV + \epsilon$, let $\tilde{T}^k = T_k \setminus L_k^1 \cup L_k^2$

Step 5: If $\tilde{T}^k = \emptyset$, then terminate with $\bar{x}^k$ as the current best feasible solution if $CBV > -\infty$, else problem is infeasible

Step 6: if $\tilde{T}^k \neq \emptyset$ compute $v^k = \arg \max\{f(v) : v \in \tilde{T}^k\}$ and compute $x^k = \pi_G(v^k) = \lambda v^k$, $\lambda = \max\{\alpha > 0 : \alpha v^k \in G\}$, if $x^k \in G$, then we are done and return the solution v^k is an optimal solution

Step 7: Let T_{k+1} be the set obtained from $\{\tilde{T}^k \setminus v^k\} \cup \{v^k - (v_i^k - x_i^k)e^i, i = 1 \dots n\}$ after removing the improper vertices, set $k = k + 1$ and goto step 4

Many enhancements of this algorithm and competing algorithms have been developed in the recent years [22,23].

QUADRATIC PROGRAMMING

Quadratic programming (QP) is a very important type of continuous optimization problem by itself. It arises in many applications, such as portfolio theory, linear regression, VLSI design, optimal control, and optimal power flow. Also, it acts as the subproblem of many important methods, such as sequential quadratic programming and augmented Lagrangian methods. QP problems are optimization problems with quadratic objective function and linear constraints as follows:

$$\begin{aligned} \text{Minimize} \quad & f(x) = \frac{1}{2}x^T Qx + c^T x \\ \text{s.t.} \quad & Ax \leq b, \end{aligned} \quad (21)$$

where Q is a real symmetric $n \times n$ matrix, and A is $m \times n$ matrix composed of row vectors $a_i, i = 1, \dots, m$, and c and b are $n \times 1$ and $m \times 1$ column vectors, respectively. The characteristics of the above problem (21) heavily depend on the structure of Q . If Q is positive semidefinite, the QP problem is convex and its KKT point is a global optimal solution. If Q is negative semidefinite, then $f(x)$ is concave and it is a quadratic concave programming problem. If Q has both positive and negative eigenvalues, it is called *indefinite QP*.

For the convex QP, there exist many solution methods among which some are polynomial algorithms. However, we focus here on nonconvex QP, which has either negative semidefinite or indefinite Q matrix. No polynomial algorithm is known for the nonconvex QP problems since these problems

are NP-hard. For concave QP, if a global optimum exists, then a global optimal solution exists among extreme points. For indefinite QP, the global optimal solution is not necessarily attained at an extreme point, but it is surely at the boundary of the polyhedron.

Optimality Conditions

Since most of the QP algorithms are dealing with the KKT conditions of the problem, we show here the KKT (first-order optimality) conditions as follows:

$$Ax \leq b, \quad (22)$$

$$Qx + c + A^T \lambda = 0 \quad \text{and} \quad \lambda \geq 0, \quad (23)$$

$$(Ax - b)^T \lambda = 0, \quad (24)$$

where λ is the dual vector (Lagrangian multipliers). Any point satisfying the above conditions, (22)–(24), is called a *KKT (first-order critical) point*. Inequality (22) is always referred to as the *primal feasibility*, and Equation (23) is the *dual feasibility* and Equation (24) is the *complementary slackness*. Since $\lambda \geq 0$ and $Ax \leq b$, Equation (24) is equivalent to $(a_i x - b_i) \cdot \lambda_i = 0, i = 1, \dots, m$. Moreover, a KKT point is called *strictly complementary* if and only if $(a_i x - b_i)$ and λ_i do not equal to 0 simultaneously.

Many algorithms try to find a weak first-order critical point, because finding the global optimal solution in polynomial time is very difficult. The active set at x is composed of all indices at which the linear constraints are binding, that is,

$$I(x) = \{i \mid a_i x = b_i\}.$$

A second-order critical point is a KKT point satisfying the following condition:

$$s^T Q s \geq 0, \quad \forall s \in \mathcal{S}(x), \quad (25)$$

where

$$\mathcal{S}(x) = \left\{ s \mid \begin{aligned} &a_i s = 0, \quad \forall i \in I(x) \cap \{i \mid \lambda_i > 0\}, \quad \text{and} \\ &a_i s \geq 0, \quad \forall i \in I(x) \cap \{i \mid \lambda_i = 0\} \end{aligned} \right\}.$$

If strict inequality holds in Equation (25) for $s \neq 0$, x is referred to as *strong second-order critical point*.

To define a weak second-order critical point, let

$$\mathcal{S}'(x) = \{s \mid a_i s = 0, \quad \forall i \in I(x)\}.$$

Then a weak second-order point is a KKT point such that $s^T Q s \geq 0, \quad \forall s \in \mathcal{S}'(x)$. Weak second-order critical points are easier to obtain than (strong) second-order critical ones. By the definitions above, a weak second-order critical point is also second-order critical if it is strictly complementary.

Different techniques are used to solve QP problems, according to different constraints' structures of QP, mainly equality-constrained and inequality-constrained QP.

Equality-Constrained QP

The equality-constrained QP is as follows:

$$\begin{aligned} \text{Minimize} \quad & f(x) = \frac{1}{2} x^T Q x + c^T x \\ \text{s.t.} \quad & Ax = b. \end{aligned} \quad (26)$$

To obtain a KKT point x^* , we need to solve the following linear equation:

$$\begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}. \quad (27)$$

By rearranging and substitution, $x^* = x + s$ (x is an estimate of x^* and s is the difference), we can rewrite Equation (27) as

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} -s^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} d \\ g \end{bmatrix}, \quad (28)$$

where $s = x^* - x$, $d = c + Qx$, and $g = Ax - b$.

Instead of solving linear equations (27), we solve equations (28) where the matrix is

symmetric, making it easier to handle. The symmetric matrix in Equation (28) is referred to as *KKT matrix* usually denoted by K , which is always indefinite.

There are mainly two branches of methods to solve Equation (28), direct solution methods and iterative solution methods.

In direct solution methods, full KKT system factorization is the most basic one. We can perform the Gauss elimination method to solve it, which does not take advantage of symmetry. A better way is to exploit symmetric indefinite factorization, which is very effective in a number of problems. When Q is of full rank and can be easily inverted, we can also use the Schur-complement method. By rearranging the terms of Equation (28), we can get

$$\begin{aligned} A Q^{-1} A^T \lambda^* &= A Q^{-1} d - g, \\ s &= Q^{-1} A^T \lambda^* - d. \end{aligned}$$

Q , however, is not always nonsingular. In the singular cases, null space methods are usually used. Interested readers may refer to Nocedal and Wright [24].

For very large systems or with parallel computation, the iterative method is a better choice. Conjugate gradient (CG) method is the most basic and prominent method of this type. CG can be applied directly to the reduced system of QP (26). In the reduced system, x is decomposed to two parts as follows:

$$x = Yx_1 + Zx_2,$$

where Z is the $n \times (n - m)$ null space matrix of A , Y is a matrix such that $[Y|Z]$ is nonsingular. The problem is reduced to solving the following linear equations by CG method,

$$Z^T Q Z x_2 = -Z^T Q Y (A Y)^{-1} - Z^T c. \quad (29)$$

To improve the convergence rate, usually preconditioning is introduced to the CG method. The CG method works with the null space matrix Z at each step, which makes its performance greatly affected by the choice of Z . The projected CG method will not use Z explicitly, which can overcome the poor choice of Z . Also, preconditioning can be applied to the projected CG method to enhance convergence.

Inequality-Constrained QP

Inequality-constrained QPs are more general than the equality-constrained ones because of the linear inequality constraints. However, this makes the problem harder and we cannot just solve a linear system to get the optimal solution most of the time. The most basic but time-consuming method is enumerative method, which is listed here to give some basic ideas. Active-set method usually is a directed enumerative method, which is good for small and medium-size QPs. For large-scale QPs, generally interior point method is a better choice. However, if we solve a sequence of related problems, active-set method performs better because of its properties that are explained below.

Enumerative Method. For the nonconvex QPs, the optimal solutions occur on the boundary of the feasible region. At the optimal solution x , by the notion of active set $A(x)$, the KKT conditions reduce to the following linear equations:

$$\begin{aligned} Qx + A^T\lambda &= -c, \\ a_ix &= b_i, \quad i \in A(x), \\ \lambda_i &= 0, \quad i \in \{1, 2, \dots, m\} \setminus A(x). \end{aligned}$$

Of course, we can enumerate all the possible choice of active set $A(x)$ and solve the above linear system and check if they are in the feasible region. If we are not lucky, this at most requires $2^m(m+n)^3$ arithmetic operations, because there are 2^m choices of $A(x)$ and many $\mathcal{O}((m+n)^3)$ algorithms for the above linear system. One way to avoid total enumeration is to use extreme point ranking and linear underestimation [25].

Active-Set Method. A smarter way of enumeration is the active-set method. First, we discuss the inertia of a symmetric matrix since we are dealing with nonconvex QPs. The inertia of symmetric matrix is a three-dimension vector,

$$\text{inertia}(K) = (n_+, n_-, n_0),$$

where n_+ , n_- , and n_0 indicate the numbers of positive, negative, and zero eigenvalues, respectively. This method starts with

a boundary point, and solves a sequence of subproblems, each of which is to find a descent direction and manipulates the active constraints. At step k , the subproblem is to solve a equality-constrained QP, given a fixed boundary point, x_k , as follows:

$$\begin{aligned} \text{Minimize} \quad & f(x_k + s) = \frac{1}{2}s^T Qs + \left(c + Qx_k^T\right)s \\ & + \left(c^T x_k + \frac{1}{2}x_k^T Qx_k\right) \quad (30) \\ \text{s.t.} \quad & A_{\mathcal{I}}(x_k + s) = b_{\mathcal{I}}, \end{aligned}$$

where $A_{\mathcal{I}}$ is a matrix that is composed of linearly independent rows drawn from the active set $\mathcal{I}$. Also the constraint is equivalent to $A_{\mathcal{I}}s = 0$, since $A_{\mathcal{I}}x_k = b_{\mathcal{I}}$. When $s \neq 0$, let α be as follows:

$$\alpha = \min_{i \notin \mathcal{I}} \left\{ \frac{b_i - a_i x_k}{a_i s} \right\}. \quad (31)$$

If $\alpha \geq 1$ and there are no new constraints satisfying $b_i = a_i x_k$, then there is no new constraint becoming active. Otherwise, we should add a new constraint with index i^* corresponding to the minimizer of Equation (31). When $s = 0$, we also need to know the value of dual variable, λ , of the equality-constrained QP subproblem (30). If $\lambda \geq 0$, we stop with the current solution. Otherwise, we drop constraint j^* such that λ_j is the largest for all j in the working set. Inertia controlling is introduced in the nonconvex case. After it was first proposed by Fletcher [26], a lot of variants have been studied. The key point of inertia controlling is that at each step we always select a working set that can control the inertia of the reduced Hessian, $Z^T QZ$ in Equation (29), such that $n_- \leq 1$. Interested reader can refer to the survey paper by Gill *et al.* [27].

Interior Point Method. Interior point method is such a method that traverses a trajectory inside of the feasible region to approach the optimal solution. It is a very efficient method for QP. Instead of considering the format of QP (21), we treat QP in an alternative formulation

$$\begin{aligned}
\text{Minimize} \quad & f(x) = \frac{1}{2}x^T Qx + c^T x \\
\text{s.t.} \quad & Ax = b, \\
& x \geq 0.
\end{aligned} \tag{32}$$

Then add a barrier function in the objective function to eliminate the nonnegativity constraints and make the above QP (32) an equality-constrained nonlinear problem,

$$\begin{aligned}
\text{Minimize} \quad & f(x) = \frac{1}{2}x^T Qx + c^T x + \mu B(x) \\
\text{s.t.} \quad & Ax = b.
\end{aligned} \tag{33}$$

where $\mu > 0$ and usually we use the logarithmic barrier function,

$$B(x) = -\sum_{i=1}^n \ln x_i.$$

The KKT conditions to the barrier problem is as follows:

$$Qx + c - z + A^T \lambda = 0, \tag{34}$$

$$Ax = b, \tag{35}$$

$$x_i z_i = \mu, \quad \forall i = 1, 2, \dots, n, \tag{36}$$

where $z = (z_1, z_2, \dots, z_n)$ and $z_i = \frac{\mu}{x_i}$. Conditions (34)–(36) will coincide with the original KKT conditions (22)–(24) as μ approaches 0. We then use the symbols, x_μ , λ_μ , and z_μ , since they are considered a function of μ in the above system. The trajectory of $(x_\mu, \lambda_\mu, z_\mu)$, as μ approaches 0, is called the *central path*. At each step, there are two main tasks. One is to find an interior point that is close enough to the central path. The other one is to decrease the barrier parameter μ . To calculate a new interior point, two methods are used: line-search-based methods [24] and trust-region-based methods [28]. This is called the primal dual path following method. Also, there exist a variety of interior point methods. In practice, the predictor-corrector primal dual interior point method is very popular. More details about interior point methods can be found in Ye [29].

In general, the active-set methods need a lot of steps to converge, but each step

is comparatively easy. In contrast, interior point methods take fewer steps to converge, but each step is more expensive. In most cases of large-scale problems, usually interior point methods outperform the active set ones. However, with “warm start,” the active-set methods may only need few cheap steps to yield the solution.

GENERAL CONCAVE MINIMIZATION

In this section, we focus on minimizing general concave function $f(x)$ over a polyhedron D . Usually these kinds of problems are referred to as *general concave minimization problems*. Sometimes, the feasible region also takes the form of general compact convex set, which is commonly defined by constraints as $g_i(x) \leq 0$, where $g_i(x)$ is convex.

General concave minimization has a variety of applications. In economics and supply chain management, commodity purchase cost is usually a concave function of purchase quantity, as in the fixed charge, economies of scale assumptions. Hence the objective function is often a concave function.

Sometimes, integer programming problem,

$$\begin{aligned}
\text{Minimize} \quad & f(x) \\
\text{s.t.} \quad & x \in C \cap \{0, 1\}^n,
\end{aligned}$$

is transferred to a concave minimization problem,

$$\begin{aligned}
\text{Minimize} \quad & f(x) + \mu x^T (e - x) \\
\text{s.t.} \quad & x \in C \cap E,
\end{aligned}$$

where C is compact, $E = \{x | 0 \leq x \leq (1, \dots, 1)^T\}$, μ is a sufficiently large number that makes the second term in the objective function dominant and its Hessian is negative semidefinite.

Bilinear programming problem with $Q_{m \times n}$,

$$\begin{aligned}
\text{Minimize} \quad & p^T x + x^T Q y^T + q^T y \\
\text{s.t.} \quad & x \in X, y \in Y,
\end{aligned}$$

is also equivalent to a concave programming problem with piecewise linear objective function because there is no interaction between x and y in the constraints.

In applied mathematics, economics, decision sciences, and engineering, the complementary conditions plays a very important role. It takes the forms as follows:

$$g(x) \geq 0, h(x) \geq 0, g(x)^T h(x) = 0.$$

Finding a solution to the above conditions can be converted to an equivalent concave minimization problem. If $g(\cdot)$ and $h(\cdot)$ are concave on a convex set D , then, whenever it has a solution, it is equivalent to the following concave minimization problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^n \min [g_i(x), h_i(x)] \\ \text{s.t.} \quad & g(x) \geq 0, h(x) \geq 0 \\ & x \in D, \end{aligned}$$

where $\sum_{i=1}^n \min [g_i(x), h_i(x)]$ is concave function because of the concavity of $g(\cdot)$ and $h(\cdot)$, and the fact that $\min[a, b] + \min[c, d] \geq \min[a + c, b + d]$.

By using the concavity of $\ln(\cdot)$ function, multiplicative programs can also be transformed into general concave minimization problems. More and detailed applications can be found in Floudas and Pardalos [30,31]. Not only general concave minimization has many applications, but it also has many solution algorithms. In the following sections, we mainly talk about three important algorithms for general concave minimization: cutting plane algorithm, branch and bound method, and outer and inner approximation algorithm. Interested readers may also refer to the survey paper by Pardalos and Rosen [32].

Cutting Plane Algorithm

A point y is called the γ -extension of x in the direction $d \in \mathcal{R}^n \setminus \{0\}$ with respect to concave function $f(\cdot)$ if $y = x + \theta d$ and $\theta = \min\{\theta_1, \sup\{t | f(x + td) \geq \gamma\}\}$ where θ_1 is a sufficiently large number and γ is a lower bound of $f(x)$.

Let v be a nondegenerate vertex such that $f(v) \geq \gamma$, and $d_1, d_2, \dots, d_n$ be n linear independent edges emanating from v . Suppose that γ is the smallest function value among all function values at feasible points known so far, and $y_1, y_2, \dots, y_n$ are the γ -extensions of vertex $v \in D$ in direction $d_1, d_2, \dots, d_n$ with respect to function $f(\cdot)$. Because $f(\cdot)$ is a concave function, every point x in the simplex $S(v, y_1, y_2, \dots, y_n)$ satisfies that $f(x) \geq \gamma$.

There exists a unique hyperplane containing $y_1, y_2, \dots, y_n$, because $d_1, d_2, \dots, d_n$ are linearly independent. The hyperplane can be defined as

$$H = \{x : e^T Y^{-1} x = 1 + e^T Y^{-1} v\},$$

where $e = (1, 1, \dots, 1)^T$, $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)^T$, and $Y = ((y_1 - v), (y_2 - v), \dots, (y_n - v))$. Let $H_+ = \{x : e^T Y^{-1} x \geq 1 + e^T Y^{-1} v\}$ and $H_- = \{x : e^T Y^{-1} x \leq 1 + e^T Y^{-1} v\}$ be two half spaces generated by hyperplane H . By the argument in last paragraph, we know that $f(x) \geq \gamma$ for all $x \in D \cap H_-$. Usually, the hyperplane H is called a *concavity cut*.

The cutting plane algorithm adds a concavity cutting plane at each step until no feasible region can be cut off furthermore. First, set active region $P = D$, and find a vertex v such that $f(v) \geq \gamma$, where γ is an upper bound of the problem. At each iteration, construct a concavity cut H with respect to v , P , and γ . Check if $P \subset H_-$. If true, stop the algorithm. Otherwise, set $P \leftarrow P \cap H_+$. With the new active feasible region, γ, v are recalculated again. Then go to the next iteration. The basic idea is simple, but there are a variety of methods to recalculate γ, v and obtain the concavity cuts based on special properties of different problems.

Branch and Bound Algorithm

Branch and bound algorithm is a widely used technique in not only optimization literature, but also practical optimization software. It is a very powerful method for concave minimization. Usually branch and bound algorithm is finite when only an ϵ -optimal solution is needed. Also convergence speed always depends on ϵ . Many finite exact branch and

bound methods have been proposed since the 1980s[33–35]. Two important parts that differentiate the branch and bound methods from each other are how to partition the space and how to calculate an underestimate of the objective function. Usually there are three ways of partitioning: simplicial, conical, and rectangular subdivisions. Details of partitioning can be found in Horst *et al.* [1]. By taking advantage of the concavity of the objective function, there are many ways to underestimate the objective function value.

Starting with an arbitrary feasible point with upper bound UB, the simplicial algorithm for concave optimization first constructs a simplex, $S_0 \supset D$ and sets the unsolved candidate simplex collection $\mathcal{P} = \{S_0\}$ and solved candidate simplex collection $\mathcal{M} = \{S_0\}$. At each iteration, solve the following problem for all simplices $S(v_0, v_1, \dots, v_n) \in \mathcal{P}$,

$$\begin{aligned} \mu(S) = \text{Minimize} \quad & \sum_{i=0}^n \lambda_i f(v_i) \\ \text{s.t.} \quad & \sum_{i=0}^n \lambda_i v_i \in D, \\ & \sum_{i=0}^n \lambda_i = 1, \quad \lambda_i \geq 0. \end{aligned} \quad (37)$$

Problem (37) yields a lower bound $\mu(S)$ of function $f(\cdot)$ on $S \cap D$, since $f(\cdot)$ is concave. Then set $\text{UB} \leftarrow \min\{\text{UB}, \mu(S), \forall S \in \mathcal{P}\}$, and delete those simplices S from $\mathcal{M}$ such that $\mu(S) \geq \text{UB} - \epsilon$. Select a simplex S^* from $\mathcal{M}$, usually the one which yields the current UB, and subdivide it to at most $n + 1$ simplices by using a point $x^* \in S^*$. Let $\mathcal{P}^*$ be the collection of the new simplices. Update $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{P}^* \setminus S^*$, and $\mathcal{P} = \mathcal{P}^*$. When $\mathcal{M} = \emptyset$, stop the algorithm. UB is an ϵ -optimal value. The choice of x^* determines the way the simplex is subdivided. Bisection subdivides the simplex into two parts by using the middle point of its edge. The bisection methods with a positive ϵ has already been proved to be finite. But it does not terminate in finite steps when ϵ equals to 0.

Some modifications will make the above algorithm finite even when $\epsilon = 0$. To ensure

finite convergence, some concavity cuts are added at each iteration. Instead of using D in Equation (37) at every iteration, always use a smaller feasible region D_i , which is constructed by adding new concavity cuts obtained in the last iteration. The finite convergence is achieved only if the global optimum of $f(\cdot)$ is only attained at vertices of D . More details about the new algorithm combining branch and bound and cutting plane and finite convergence conditions can be found in Locatelli and Thoai [36]. Also, other than branch and bound, there are more general methods based on interval analysis, and we refer the readers to other articles [3, 37–45].

Outer and Inner Approximation Algorithms

Outer and inner approximation algorithms are two inherently correlated methods for concave minimization. Both algorithms keep refining the searching region to approximate the feasible region D . Starting with a superset of the feasible region D , the outer approximation shrinks the searching region at each step by adding a new constraint violated by the current optimal vertex. The newly added constraint cuts off some infeasible vertices and introduce new vertices. Usually, vertex enumeration algorithm is used to update the vertex set. However, the inner approximation algorithm works the other way around. It starts with a subset of D or some set whose intersection with D is not empty, and keeps expanding the searching area at each step by adding a new vertex. Both of these two algorithms converge finitely. They can be applied to problem with general convex constraints. For more details of these algorithms, please refer to Horst *et al.* [1].

LINEAR COMPLEMENTARITY PROBLEM, AND MATHEMATICAL PROGRAMS WITH COMPLEMENTARITY CONSTRAINTS

Complementarity appears in a lot of areas such as mathematics, economics, science, engineering, and management. Here, we discuss two important types of problems associated with complementarity, the linear

complementary problem and mathematical programs with complementarity conditions.

Linear Complementarity Problem

Linear complementarity problem (LCP) is a very basic and important class of problems in mathematical optimization area. It is to find a solution z^* to the following system:

$$(Mz + q)^T z = 0, \quad (38)$$

$$Mz + q \geq 0, \quad (39)$$

$$z \geq 0, \quad (40)$$

where M is an $n \times n$ parameter matrix, q is an n -dimensional parameter vector, and z is the variable vector. Usually the complementarity problem, (38)–(40), is denoted by $LCP(q, M)$. LCP unites linear programming, QP, and bimatrix game. The KKT conditions of a linear programming or QP problem is equivalent to an LCP. The KKT conditions of LP $\min\{c^T x | Ax \geq b, x \geq 0\}$ are $(Ax - b)^T y = 0$, $(A^T y - c)^T x = 0$, $A^T y \leq c$, $y \geq 0$, and the primal feasible constraints, where y is the dual variable. By rearranging the terms, the KKT conditions are as follows:

$$\begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c \\ -b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0, \quad (41)$$

$$\begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \geq 0, \quad (42)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} \geq 0. \quad (43)$$

It is an LCP where $M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix}$, $q = \begin{bmatrix} c \\ -b \end{bmatrix}$, and $z = \begin{bmatrix} x \\ y \end{bmatrix}$. Similarly, the KKT conditions of QP (21) with additional constraint $x \geq 0$ are also equivalent to the LCP $\left(\begin{bmatrix} c \\ b \end{bmatrix}, \begin{bmatrix} Q & A^T \\ -A & 0 \end{bmatrix} \right)$.

Let A and B be two $m \times n$ matrices, and e_m be a column vector of m 1's. For every solution, (x', y') , to LCP $\left(\begin{bmatrix} -e_m \\ -e_n \end{bmatrix}, \begin{bmatrix} 0 & A \\ B^T & 0 \end{bmatrix} \right)$, there exists a corresponding Nash equilibrium point, (x^*, y^*) , of the bimatrix game (two persons game) $\Gamma(A, B)$, and vice versa. The relationship is as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \frac{x^*}{(x^*)^T B y^*} \\ \frac{y^*}{(x^*)^T A y^*} \end{bmatrix},$$

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{bmatrix} \frac{x'}{e_m^T x'} \\ \frac{y'}{e_n^T y'} \end{bmatrix}.$$

For more details of the relationship, the interested reader may refer to Murty [46] and Cottle *et al.* [47].

LCP can be solved sometimes as a QP problem,

$$\begin{aligned} &\text{Minimize} && (Mz + q)^T z \\ &\text{s.t.} && Mz + q \geq 0, \\ &&& z \geq 0, \end{aligned} \quad (44)$$

because the objective function is always non-negative with respect to its constraints. With this transformation, we can apply any algorithm discussed before for QP.

However, there exist more specific methods to LCPs, such as Lemke's complementary pivot method and Cottle and Dantzig's principal pivot method [46, 47]. To introduce the pivot algorithms, a little change is needed to the original LCP formulation (38)–(40),

$$wI - Mz = q, \quad (45)$$

$$w^T z = 0, \quad (46)$$

$$z, w \geq 0, \quad (47)$$

where I is an $n \times n$ identity matrix. LCP actually tries to find a complementary cone in $C(M)$ in which vector p lies. Every complementary cone is a positive cone, $Pos(a_1, a_2, \dots, a_n) = \{y | y = \sum_{i=1}^n \alpha_i a_i\}$, formed by n vectors that are drawn from n complementary vector pairs (I_j, M_j) , which are the j th column vectors of matrix I and M , respectively.

Lemke's complementary pivot algorithm works in a way very similar to the simplex algorithm. It starts with an almost complementary feasible basic solution in which only one complementary vector pair is not selected in the basis. Then the algorithm keeps changing the basis by the pivot method. But the entering variable at each step is

always the complementary variable of the dropping variable in the previous step. Usually lexicographic rules are applied to prevent cycling. This algorithm cannot guarantee a feasible solution when the LCP is actually feasible. And the worst possible computation complexity is 2^n . Although it faces these difficulties, this algorithm is widely applied and has a lot of variants.

The generic principal pivot algorithm does not introduce any new artificial variable and only pivots on a variable whose complement is nonpositive. At each step the entering variable is always the complement of the dropping variable. But in the process of this algorithm, the signs of variables may change multiple times. This algorithm is most useful for LCP with M being a P -matrix, a square matrix with every principal minor greater than 0. Cottle and Dantzig's principal pivot algorithm works with major cycles in which there are many substeps. Starting with a complementary basic vector of variables, every major cycle yields an improved complementary basic vector of which at least one more variable is made nonnegative. All the substeps work with almost complementary basic vectors. Cottle and Dantzig's principal pivot algorithm works well for LCP's with M being either a P -matrix or a PSD matrix. According to different special structures of M in LCP, there exist many polynomial bounded algorithms. We refer the reader to Murty [46] and Cottle *et al.* [47] for more details.

LCP is only a special case of the family of complementarity problems (CPs). Nonlinear complementarity problem (NCP) is also a very popular topic, which includes a nonlinear transformation instead of the affine transformation. For a nonlinear mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$, $\text{NCP}(F)$ is to find a vector x such that

$$\begin{aligned} x^T F(x) &= 0, \\ F(x) &\geq 0, \quad x \geq 0. \end{aligned}$$

In addition, the mixed linear and nonlinear complementarity problems also draw a lot of attention, both being special cases of the generalized complementarity problems. A broader topic is the variational inequality

(VI), which is usually denoted by $\text{VI}(K, F)$, where $K \subseteq \mathbb{R}^n$, $F : K \mapsto \mathbb{R}^n$. CP and VI comprise a very important mathematical programming branch and include a lot of applications. More comprehensive discussion can be found in Facchinei and Pang [48].

Mathematical Program with Complementarity Constraints

Mathematical program with complementarity constraints, usually abbreviated as MPCC, is a very important and useful type of problem that appears in literature. MPCC, as its name suggests, is an optimization problem where some of its constraints are formulated as complementarity problems. Generally, it has the following form:

$$\begin{aligned} \text{Minimize} \quad & f(x) \\ \text{s.t.} \quad & F_i(x)G_i(x) = 0, \quad i = 1, \dots, m, \\ & F_i(x) \geq 0, \quad i = 1, \dots, m, \\ & G_i(x) \geq 0, \quad i = 1, \dots, m, \\ & g(x) \leq 0, \\ & h(x) = 0. \end{aligned}$$

MPCC is a very difficult problem that involves nonconvexity and nonsmoothness. However, many useful algorithms have been used to solve MPCC, such as branch and bound method, complementary pivot method, active-set method, sequential quadratic programming method, and interior point method [48–50].

Branch and bound method is widely used for MPCC. Its idea is very simple but powerful. Because complementarity constraints are intrinsically combinatorial, MPCC can be solved by using enumeration methods. Total enumeration always takes a lot of time. A better way to avoid this is to branch and bound. At the initial node, all the complementarity constraints are dropped. At any other node of the binary branch and bound tree, one more complementarity constraint is added in such a way that one branch emanating from the node includes an additional constraint $F_i(x) = 0$, the other with $G_i(x) = 0$. Then a lower bound is computed with additional

constraints as above, but without any complementarity constraint. The upper bound can be achieved by solving the current relaxed problem by arbitrarily choosing any of the complementarity, $F_i(x)$ or $G_i(x)$, to be zero for all the undetermined complementarity constraints. The difference when compared with standard branch and bound method is that both lower bound and upper bound are generated at each node. Interested reader may refer to Scheel and Scholtes [50] for more details about the properties of MPCC.

Also bilevel programming problems are often reformulated as MPCC because of the KKT conditions of the inner optimization problem. Interested readers may want to refer to Dempe [51] and Colson *et al.* [52] for more comprehensive details. Including integer variables into these problems would make them much harder. Interested readers may refer to DeNegre and Ralphs [53], Moore and Bard [54], Wen and Yang [55] for some algorithms to solve the integer bilevel programming problems. Bilevel programming belongs to hierarchical programming problems, where there is a hierarchy of decisions. A variety of theories, problems, algorithms, and applications of this type can be found in Migdalas *et al.* [56]. On the other hand, MPCC is a special case of mathematical programs with equilibrium constraints (MPEC), where complementarity conditions are replaced by their extension, namely, variational inequalities. For more details about MPEC, the interested reader may refer to Luo *et al.* [49].

DISCRETE OPTIMIZATION PROBLEMS

Many 0-1 integer programming problems could be reduced to continuous optimization problems. This reduction permits us to employ the techniques in global optimization to these problems. Consider the following 0-1 integer programming problem:

$$\text{Minimize } c^T x \quad (48)$$

$$\text{s.t. } Ax \leq b, \quad (49)$$

$$x_i \in \{0, 1\}, i = 1, \dots, n. \quad (50)$$

The above problem is equivalent to the concave minimization problem:

$$\text{Minimize } c^T x + \mu x^T(e - x) \quad (51)$$

$$\text{s.t. } Ax \leq b, \quad (52)$$

$$0 \leq x \leq e. \quad (53)$$

For a sufficiently large value of μ , at optimality $x^T(e - x) = 0$. Similar reduction for 0-1 integer programming problems are possible. We discuss such formulations for some combinatorial optimization problem namely, the clique problem, the satisfiability problem, and quadratic assignment problem in this section.

Satisfiability Problem (SAT)

The satisfiability (SAT) problem could be stated as follows: Given a set of literals $L = \{x_1, x_2, \dots, x_m\}$ and a collection of clauses $C = \{c_1, c_2, \dots, c_n\}$ over L , is there a truth assignment of L that simultaneously satisfies all clauses of C ? The problem could be expressed in the conjunctive normal form (CNF)

$$\bigwedge_{i=1}^n c_i = \bigwedge_{i=1}^n \left(\bigvee_{j=1}^{m_i} l_{ij} \right), \quad (54)$$

where $l_{ij} \in \{x_i, \bar{x}_i\}$ is the literal j in clause i . Given a CNF, we define a new function $f: \mathcal{R}^m \rightarrow \mathcal{R}$ and provide the following global optimization problem that is equivalent to SAT:

$$\text{Minimize } f(y) \quad (55)$$

$$\text{s.t. } y \in \mathcal{R}^m, \quad (56)$$

where

$$f(y) = \sum_{i=1}^n \prod_{j=1}^{m_i} q_{ij}(y_j), \quad (57)$$

$$q_{ij}(y_j) = \begin{cases} (y_j - 1)^{2p} & \text{if } x_j \text{ is in } c_i, \\ (y_j + 1)^{2p} & \text{if } \bar{x}_j \text{ is in } c_i, \\ 1 & \text{otherwise,} \end{cases} \quad (58)$$

where p is a positive integer and $y_j = 2x_j - 1$.

Theorem 6. *There is a satisfiable truth assignment to the CNF if and only if $f(y) = 0$ for the corresponding $y \in \{-1, 1\}^m$.*

The discrete SAT problem is transformed to the above continuous global optimization problem and there is equivalence between the optimal solution of the global optimization problem and the truth assignment for the satisfiability problem. A more comprehensive survey on satisfiability problem could be found in Du and Pardalos [57]. Another article on global optimization [58] also discusses the SAT problem.

Quadratic Assignment Problem

In a quadratic assignment problem, we need to assign n facilities to n locations. We are given two $n \times n$ matrices A and B that provides the flow of material between facilities and distance between the locations, respectively. We denote the elements of matrices A and B as $a_{ij}, i, j = 1, 2, \dots, n$ and $b_{ij}, i, j = 1, 2, \dots, n$, respectively. The minimum cost assignment problem could be formulated as follows:

$$\text{Minimize} \quad \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ik} b_{jl} x_{ij} x_{kl} \quad (59)$$

$$\text{s.t.} \quad \sum_{i=1}^n x_{ij} = 1, j = 1, 2, \dots, n, \quad (60)$$

$$\sum_{j=1}^n x_{ij} = 1, i = 1, 2, \dots, n, \quad (61)$$

$$x \in \{0, 1\}^n, \quad (62)$$

where x_{ij} is the binary variable indicating whether facility i has been assigned to location j or not. The above 0-1 quadratic integer programming problem could be reduced to the following concave programming problem:

$$\text{Minimize} \quad x^T Q x \quad (63)$$

$$\text{s.t.} \quad \sum_{i=1}^n x_{ij} = 1, j = 1, 2, \dots, n, \quad (64)$$

$$\sum_{j=1}^n x_{ij} = 1, i = 1, 2, \dots, n, \quad (65)$$

$$x \geq 0, \quad (66)$$

where $Q = S - \alpha I, \alpha > \|S\|_\infty$ and S is the $n^2 \times n^2$ matrix with the entries $a_{ij} b_{kl}$. The rows of S are defined by i and j fixed and the columns with k and l fixed. The problem was well studied by Lawler [59] and the reader may refer this for a more comprehensive review on the quadratic assignment problem.

Maximum Clique Problem

The maximum clique problem is to determine a subgraph of maximum size in graph $G = (V, E)$, where G is a simple undirected graph and V and E are its corresponding vertex and edge set. Let A_G be the adjacency matrix of the graph G . Then the maximum clique problem could be formulated as follows:

$$\text{Minimize} \quad f_G(x) = \sum_{(ij) \in E} x_i x_j = x^T A_G x \quad (67)$$

$$\text{s.t.} \quad x \in X, \quad (68)$$

where $X = \{x : \sum_i x_i = 1, x_i \geq 0, i = 1, 2, \dots, n\}$. The following theorem provides the equivalence between the optimal solution of the above concave programming problem and the maximum clique in graph G .

Theorem 7. *If $\alpha = \max\{f_G(x) : x \in X\}$ then the maximum clique C in graph G is of size $\frac{1}{(1-2\alpha)}$. This is attained with $x_i = \frac{1}{k}$ if $i \in C$ and $x_i = 0$ if $i \notin C$.*

The maximum independent set problem, which seeks the maximum empty subgraph within a given graph, could be modeled as follows:

$$\text{Maximize} \quad x^T A x \quad (69)$$

$$\text{s.t.} \quad x \in \{0, 1\}^n, \quad (70)$$

where the binary variable x_i indicating the presence of vertex i in the optimal solution and matrix $A = I_{n \times n} - A_G$.

The maximum cut problem in a graph $G(V, E)$ with a weight function on the edge set $w : E \rightarrow \mathbb{R}$, seeks a partition of the vertex set V into S and $\bar{S}$ such that the cut set formed by the partition is maximum ($\max cut(S, \bar{S})$). This could be modeled as a quadratic $\{-1, 1\}$ -variable problem as follows.

$$\text{Maximize} \quad \frac{1}{2} \sum_{i,j \in V} w_{ij} (1 - x_i x_j) \quad (71)$$

$$\text{s.t.} \quad x \in \{-1, 1\}^n, \quad (72)$$

where x_i takes the value -1 if the vertex i is in set S and value 1 if it is in set $\bar{S}$ in the optimal solution.

CONCLUSIONS

Global optimization methods are usually NP-hard problems. There are exact and heuristic methods available to solve these problems. In this article, we discussed some deterministic methods to solve various global optimization methods. More specifically, we discussed about QP, DC, optimization and relations between global and discrete optimization problems. We provided some theoretical results relating them to specific problems and the algorithmic methods that were developed based on the theoretical results. This article focuses on deterministic and exact methods in global optimization and for stochastic and metaheuristic based methods we refer the readers to the following articles [60–65].

REFERENCES

1. Horst R, Pardalos PM, Thoai NV. Introduction to global optimization. 2nd ed. The Netherlands: Kluwer Academic Publishers; 2000.
2. Pinter JD. Advanced software products for nonlinear systems optimization. Halifax (VA): Pinter Consulting Services; 2004.
3. Hansen E, Walster GW. Global optimization using interval analysis. 2nd ed. New York: Marcel Dekker; 2004.
4. Neumaier A. Interval methods for systems of equations, Cambridge: Cambridge University Press; 1990.
5. Ratschek H, Rokne JG. Interval methods. handbook of global optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1995. pp. 751–828.
6. Kearfott RB. Rigorous global search: continuous problems. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1996.
7. Hansen P, Jaumard B, Lu S-H. An analytical approach to global optimization. Math Program 1991;16:334–350.
8. Horst R, Tuy H. Global optimization: deterministic approaches. Berlin: Springer; 1996.
9. Piyavskii SA. An algorithm for finding absolute extremum of a function. USSR Comput Math Math Phys 1972;12:57–67.
10. Hansen P, Jaumard B. Lipschitz optimization. Handbook of global optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1995. pp. 751–828.
11. Tuy H. Dc optimization: theory, methods and algorithms. Handbook of global optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1994.
12. Tuy H. An outer approximation method for concave minimization. Acta Math Vietnam 1983;8:3–34.
13. Tuy H, Horst R. Convergence and restart in branch and bound algorithms for global optimization: application in concave minimization and D.C. optimization problems. Math Program 1988;42:161–183.
14. Tuy H. Polyhedral annexation, dualization and dimension reduction technique in global optimization. J Glob Optim 1991;1:229–244.
15. Baritomba W. Customizing methods for global optimization—a geometric viewpoint. J Glob Optim 1993;3:193–212.
16. Baritomba W, Cutler A. Accelerations for global optimization covering methods using second derivatives. J Glob Optim 1994;4:329–341.
17. Breiman L, Cutler A. A deterministic algorithm for global optimization. Math Program 1993;58:179–199.
18. Cutler A. Optimization methods in statistics [PhD thesis]. Berkeley: University of California; 1988.
19. Blanquero R, Carrizosa E. On covering methods for D.C. optimization. J Glob Optim 2000;18:265–274.
20. Tuy H. Monotonic optimization: problems and solution approaches. SIAM J Optim 2000;11(2):464–494.

21. Rubinov A, Tuy H, Mays H. An algorithm for monotonic global optimization problems. *Optimization* 2001;49:205–221.
22. Sun X, Li J, Li D. Convexification and monotone optimization. *Continuous optimization current trends and modern applications*. New York: Springer; 2005. pp. 275–291.
23. Tuy H, Al-Khayyal F, Thach P. *Monotonic optimization: branch and cut methods*. Essays and surveys in global optimization. New York: Springer; 2005. pp. 39–78.
24. Nocedal J, Wright SJ. *Numerical optimization*. 2nd ed. New York: Springer; 2006.
25. Murty KG. Solving the fixed charge problem by ranking the extreme points. *Oper Res* 1968;16:268–279.
26. Fletcher R. A general quadratic programming algorithm. *J Inst Math Appl* 1971;7: 76–91.
27. Gill PE, Murray W, Saunders MA, Wright MH. Inertia-controlling methods for general quadratic programming. *SIAM Rev* 1991;33(1):1–36.
28. Gould NIM, Toint PL. Numerical methods for large-scale non-convex quadratic programming. In: Siddiqi AH, Kocvara M, editors. *Trends in industrial and applied mathematics*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002. pp. 149–179.
29. Ye Y. *Interior-point algorithm: theory and analysis*. New York: John Wiley & Sons; 1997.
30. Floudas CA, Pardalos PM. *Recent advances in global optimization*. Princeton: Princeton University Press; 1992.
31. Floudas CA, Pardalos PM. *State of the art in global optimization: computational methods and applications*. The Netherlands: Kluwer Academic Publishers; 1996.
32. Pardalos PM, Rosen JB. Methods for global concave minimization: a bibliographic survey. *SIAM Rev* 1986;28(3):367–379.
33. Horst R. A general class of branch-and-bound methods in global optimization with some new approaches for concave minimization. *J Optim Theory Appl* 1986;51(2):271–291.
34. Horst R. A note on the convergence of an algorithm for nonconvex programming problems. *Math Program* 1980;19:237–238.
35. Horst R. A new branch-and-bound approach for concave minimization problems. Volume 41, *Lecturer Notes in Computer Science*. Berlin/Hidelberg: Springer; 1976. pp. 330–337.
36. Locatelli M, Thoai NV. Finite exact branch-and-bound algorithms for concave minimization over polytopes. *J Glob Optim* 2000;18: 107–128.
37. Casado LG, García I, Csendes T. A new multisection technique in interval methods for global optimization. *Computing* 2000;65: 263–269.
38. Csallner AE, Csendes T, Markót MCs. Multisection in interval branch-and-bound methods for global optimization I. Theoretical results. *J Glob Optim* 2000;16:371–392.
39. Csallner AE, Csendes T. On the convergence speed of interval methods for global optimization. *Comput Math Appl* 1996;31: 173–178.
40. Csendes T. New subinterval selection criteria for interval global optimization. *J Glob Optim* 2001;19:307–327.
41. Csendes T, Ratz D. Subdivision direction selection in interval methods for global optimization. *SIAM J Numer Anal* 1997;34: 922–938.
42. Markót MCs, Csendes T, Csallner AE. Multisection in interval branch-and-bound methods for global optimization. II. Numerical tests. *J Glob Optim* 2000;16:219–228.
43. Moore RE. *Interval analysis*. Englewood Cliffs (NJ): Prentice-Hall; 1966.
44. Ratschek H, Rokne J. *New computer methods for global optimization*. Chichester: Ellis Horwood; 1988.
45. Ratschek H, Voller RL. What can interval analysis do for global optimization? *J Glob Optim* 1991;1:111–130.
46. Murty KG. *Linear complementarity, linear and nonlinear programming*. Berlin: Helderman-Verlag; 1988.
47. Cottle RW, Pang JS, Stone RE. *The linear complementarity problem*. Boston: Academic Press; 1992.
48. Facchinei F, Pang JS. Volumes I and II, *Finite-dimensional variational inequalities and complementarity problems*. New York: Springer; 2003.
49. Luo ZQ, Pang JS, Ralph D. *Mathematical programs with equilibrium constraints*. Cambridge (NY): Cambridge University Press; 1996.
50. Scheel H, Scholtes S. Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. *Math Oper Res* 2000;25(1):1–22.

51. Dempe S. Foundations of bilevel programming. The Netherlands: Kluwer Academic Publishers; 2002.
52. Colson B, Marcotte P, Savard G. Bilevel programming: a survey. *4OR* 2005;3:87–107.
53. DeNegre ST, Ralphs TK. A branch-and-cut algorithm for integer bilevel linear programs. In: Chinneck JW, *et al.*, editors. Operations research and cyber-infrastructure, Operations research/Computer science interfaces series 47. New York: Springer; 2009.
54. Moore JT, Bard JF. The mixed integer linear bilevel programming problem. *Oper Res* 1990;38:5.
55. Wen UP, Yang YH. Algorithms for solving the mixed integer two-level linear programming problem. *Comput Oper Res* 1990;17(2): 133–142.
56. Migdalas A, Pardalos PM, Varbrand P. Multilevel optimization algorithms and applications. The Netherlands: Kluwer Academic Publishers; 1998.
57. Du DZ, Gu J, Pardalos PM. Satisfiability problem: theory and applications. Volume 35, DIMACS series. USA: American Mathematical Society; 1997.
58. Pardalos PM, Chinchuluun A. Some recent developments in deterministic global optimization. *Oper Res Int J* 2005;4:3–28.
59. Lawler EL. The quadratic assignment problem: a brief review. *Combinatorial programming methods and applications*. Dordrecht: Holland Publishing; 1975. pp. 351–360.
60. Arts E, Lenstra JK. Local search in combinatorial optimization. Chichester: John Wiley & Sons; 1997.
61. Glover F, Kochenberger GA. Handbook of metaheuristics. Norwell (MA): Kluwer Academic Publishers; 2003.
62. Pardalos PM, Romeijn HE, Tuy H. Recent developments and trends in global optimization. *J Comput Appl Math* 2000;124: 209–228.
63. Reeves CR. Modern heuristic techniques for combinatorial problems. New York: John Wiley & Sons; 1993.
64. Sait SM, Youssef H. Iterative computer algorithms with applications in engineering. Los Alamitos: IEEE Computer Society; 1999.
65. Thomas W. Global optimization algorithms—theory and application. Online e-Book. 2008. Available at <http://www.it-weise.de/projects/book.pdf>.

DIFFERENT FORMATS FOR THE COMMUNICATION OF RISKS: VERBAL, NUMERICAL, AND GRAPHICAL FORMATS

DANIELLE TIMMERMANS
JURRIAAN OUDHOFF
Department of Public and
Occupational Health, EMGO
Institute for Health and Care
Research, VU University
Medical Center, Amsterdam,
The Netherlands

INTRODUCTION

In our daily life we encounter many risks. In order to make informed decisions an adequate understanding of the risks is necessary, whether these are financial risks for investment decisions, health risks, or treatment risks such as surgery. Unfortunately, people have great difficulty in adequately understanding risks. Besides limited cognitive skills, this may also be due to a suboptimal presentation of the risks.

Risk is often defined as the probability that a negative event will happen, or the probability of a hazard occurring that will cause a harmful effect. Risk is measured uncertainty or a quantification of uncertainty. The quantitative nature and the uncertainty that risks convey both add to the difficulty in understanding, as they make the information inherently abstract. Information about the quantitative aspect of risks can be presented in many different ways (Table 1). Risks can be described in verbal terms (“a high risk”), by means of numerical estimates (“a risk of 7%”), or they can be represented graphically in diagrams or charts. These different ways of presenting risk information affect how people perceive risks and can influence the decisions they make. Some risk representations may be better suited to inform people about a single risk, whereas others may provide more clarity when

complex trade-offs need to be made. In this article, we will discuss the characteristics of different formats for risk communication and the evidence about their suitability for risk communication in different situations. We will focus on verbal terms for risks, the use of numerical formats, and the effects of graphical risk representations. We will illustrate this with examples from the domain of health risks. We end with some general recommendations for risk communication.

USING VERBAL TERMS TO EXPRESS PROBABILITIES

In daily life, verbal terms are probably the most commonly used mode for communicating information about risks. Physicians often express the probability of their diagnosis or the risk of treatment in verbal probability terms, for instance, by stating: “Your headache is *probably* caused by stress; it is *very unlikely* that it indicates a tumor and it will *almost certainly* get better in a week.” At first sight, this seems a clear way of communicating probability information as the terms have a very intuitive feel and meaning attached to them. Although verbal terms appear very meaningful, it is often not clear how they should be interpreted exactly. Many studies show that people differ greatly in the interpretation and use of verbal labels [1–5]. These differences do not just reflect differences in knowledge or experience, as they have been found to exist between experienced physicians when judging cases within their own domain [1,6]. There is no consensus, for example, on where the numerical threshold should be drawn between risks that are classified as “high” and “low.” Timmermans *et al.* [1] found that terms such as “extremely rare” had a wide range of numerical interpretations, that is, from 0.001% to 10%. Similarly, the term “very likely” was associated with probabilities ranging from 40% to 100% (Fig. 1). Using verbal terms to express uncertainty can thus result in ambiguous messages.

Table 1. Common Formats for Communicating Risks

<i>Verbal probability terms</i>
“a large chance”
“highly likely”
<i>Numerical terms</i>
Percentages (%)
Frequencies
1 out of X
X out of 100(0)
Odds
Probabilities
Relative risk
Number needed to treat
<i>Graphical formats</i>
Bar charts
Population diagrams
Miscellaneous (mixtures)

As a risk format, verbal probability terms reflect an interpretation of the risk size and not a factual statement. Individuals differ in how they interpret risks and which verbal term best reflects a certain risk. Moreover, several aspects have been found to affect the interpretation of verbal terms. Verbal risk labels are attributed a higher probability when the outcome it refers to is perceived to be relatively more common or when the consequences are more severe [8,9]. For instance, the probability conveyed by the

phrase “It is likely that you will get influenza this year” is seen as involving a higher quantitative risk than the phrase “It is likely that you will get pneumonia this year”. Mapes [9] found that less serious “rare” side effects of antihistamines were deemed more likely than more serious “rare” side effects of beta-blockers. Verbal probability terms thus reflect a personal interpretation of the perceived absolute probability, the relative probability, and the severity of the potential outcome that the risk term refers to. Verbal terms can thereby be very rich in meaning, but this meaning is multidimensional, and recipients may not always share the same interpretation as the person giving the information [10–12]. Difficulties may particularly arise when individuals interpret a verbal risk term as reflecting the severity of, for example, a medical condition instead of the probability. In this regard, Gurmankin *et al.* [13,14] found that many patients thought that the physician attempted to downplay cancer risks when they were described in verbal terms. In an attempt to “correct” this feeling, the patients subsequently perceived the risks as higher than they really were. Thus, in addition to large interindividual differences in the numerical interpretation of verbal terms of uncertainty, there are also large intersituational differences. The observed ambiguity of verbal terms often

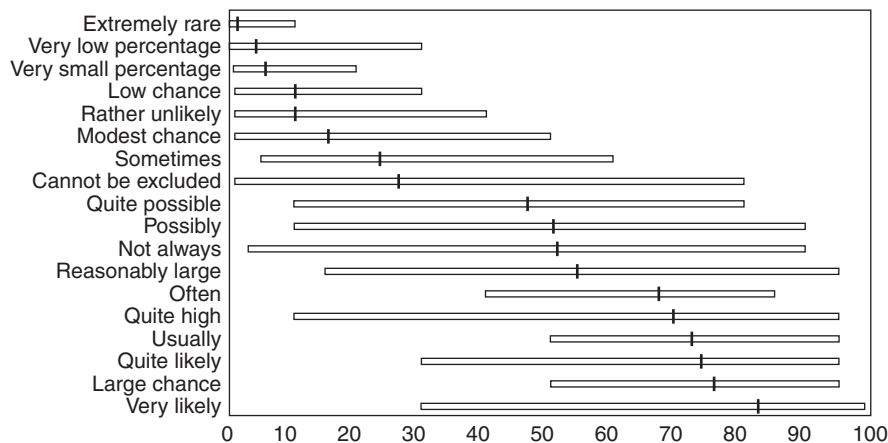


Figure 1. The numerical interpretation of verbal terms (used in patient cases) of internists, surgeons, residents in internal medicine and surgery, and interns. The vertical lines represent the means and the horizontal lines depict the range of estimates [7].

lead to miscommunications, suggesting that it would be better if verbal expressions are replaced by more precise and clearer numerical terms.

The appealing simplicity and common use of verbal probability terms has led to several initiatives exploring the standardized use of verbal risk terms to express specific numerical probabilities. The EU has, for instance, proposed using a list of verbal risk descriptors that each have a predefined and limited range of associated probabilities they are to convey [15,16]. In this way, it was expected that the rich meaning of verbal terms could be retained, while the large range in interpretations could be restricted. Such standardizations, however, have repeatedly proven to be impractical, because the range of numerical values is too variable and sensitive to the specific situation. Subsequently, the use of standardized verbal labels probably creates more confusion than it resolves.

Whereas their ambiguity may limit the usefulness of verbal risk terms as a universal standard for risk communication, their rich meaning can be practical when providing guidance in interpretation is considered more important than providing clarity over the exact risk size. Sometimes epidemiological data may be absent or too unspecific to provide exact risk estimates in a situation. Less accurate, but meaningful, verbal descriptions may then be most appropriate. In other situations, verbal risk terms may be useful as they allow physicians to incorporate persuasive meanings when patients clearly need them. Physicians may, for instance, want to reassure patients who seem unduly worried, or to convince patients about the riskiness of certain choices. Verbal expressions of probability may then give a direction to how a patient or consumer should interpret the risks [17]. In each situation, the apparent elegance of a compact verbal risk term must be traded off against a more precise, but perhaps less familiar numerical expression.

In conclusion, on the one hand, using verbal terms instead of numbers leads to a loss of information, because the exact meaning is not clear. On the other hand, by using verbal expressions additional meaning beyond probability is communicated. This might explain

Table 2. Pros and Cons of Verbal Risk Formats

Verbal Risk Format	
Pros	Cons
Seemingly simple and clear	Prone to miscommunication
Part of everyday language	Large variation in interpretation between individuals and between situations
Allows for affective message	Less trustworthy
May give additional meaning and guidance for interpretation	Imprecise and vague
Intuitive appeal	

the finding that the majority of people prefer to receive probabilistic or uncertain information in numerical terms, but like to use verbal terms when communicating this information themselves [18]. The speaker intends to communicate more information than the probability of an event, which might be more than the receiver expects. See Table 2 for a list of pros and cons of using verbal terms for risks.

USING NUMBERS TO EXPRESS PROBABILITIES

The quantitative nature of risk information makes numerical risk formats a logical tool for risk communication. Using numbers allows one to be specific, exact, and factual about the absolute and relative size of risks. In many situations where risks are communicated, individuals have the option to choose between two or more alternatives. These alternatives can, for instance, include different types of treatment or changing one's behavior versus maintaining a status quo. As the risks and benefits of the alternatives need to be weighed against each other, the degree to which these risks differ needs to be clarified. Within the realm of evidence-based medicine and informed decision making, such factuality and precision are appealing features of numerical risk formats. In

addition, when asked, many persons indicate that they trust numerical information more compared to verbal terms when they are informed about risks [13,14]. Unfortunately, laypeople often have great difficulty understanding and using numerical estimates of risk [19].

Although, the communication of risks in numbers appears to contribute to a more transparent and neutral communication process, one study shows that the interpretation of numerical risks is not straightforward [20]. Many people have a limited intuitive feel for numerical information. This may be particularly problematic if the risks are unfamiliar, as is commonly the case in the medical domain. There are various numerical formats that can be used to convey risk information. The most commonly used formats use a standardized scale like percentages (0–100%) or probabilities (0–1). Other formats like relative frequencies (x out of 100 or 1 out of x) and odds (x to 1) do not have a standardized scale.

Communicating Absolute Risks: Odds, Probabilities, Percentages and Frequencies

The literature on risk communication about the effects of different risk formats focuses primarily on the use of percentages and frequencies. The few studies [21] which discussed odds show that the use of odds for risk communication to professionals or the public does not seem ideal.

Risk perceptions may be significantly different, depending on the numerical format used. A general characteristic to explain differences in the interpretation of numerical risk formats is the degree to which the format reflects the occurrence of concrete and imaginable events. Communicating a risk as a true probability statement (e.g., “a chance of 0.10” or “a chance of 10% that event Y happens”) leads to poorer comprehension than when the risk statement explicitly states which sample is at risk (e.g., “out of a 100 patients who undergo procedure X, event Y will happen to 10” [22,23]. A reason is that in the latter statement the reference class, that is the population at risk, is clarified. Further, several studies have observed that a format that expresses risks as frequencies

yields higher perceived risks than when the risk is expressed as a percentage. Slovic *et al.* [24], for instance, found that if the risk of a psychiatric patient committing an act of violence was presented as a frequency (e.g., “20 out of 100 patients similar to this one commit an act of violence”) about twice as many psychiatrists (40% vs. 20%) refused to release the patient from care as opposed to when the risk was presented as a percentage (i.e., “a 20% chance”). Although this is supported by some studies [25–28], other studies failed to find a difference [13,14,21,29,30], thus indicating that the effect may not always be profound. Overall, frequency formats appear to at least have the potential of eliciting stronger effects on risk perception than percentages when single risks are communicated. The common hypothesized explanation for these findings is that frequencies are better at evoking an image of the risk since they refer to concrete numbers of events, whereas percentages primarily represent a more abstract and normalized proportion on a continuous scale. These explanations reflect that frequencies may be perceived as more meaningful, which is an important property in view of effective risk communication.

The explanation that the use of concrete numbers in risk communication stimulates intuitive risk perception may also explain why probabilities proper are hardly ever used in risk communication. As probabilities are standardized on a scale of 0–1, they do not directly refer to real events. Similarly, the beneficial effect of using real numbers would advise against the use of numbers with decimals when communicating risks. This may lend further support for the use of frequencies, as they allow using real numbers when risks are lower than 1%.

While it is difficult for many people to understand single risks, conditional risks are even harder to understand. Conditional risks are difficult to comprehend because they are risks that depend on another outcome. For example, when a rather young woman with a low prior risk on breast cancer tests positive on a mammogram, she still has a low chance of having breast cancer. The posterior probability of having cancer with a positive test result depends on the accuracy of the test,

as well as on the prior probability. This is easier to understand if the probabilities are presented in natural frequencies [22,31–33]. Because such sampling of frequencies preserves base rates, the predictive value is clear without the need for further calculations. This method of using one sample to illustrate the diverse possibilities is considered to reflect how people observe risks in daily life and this mode of risk communication has subsequently been dubbed using “natural frequencies.”

Communicating Absolute Risks: Different Formats for Frequencies

While using frequencies may generally be more concrete, they comprise a diverse range of formats that do not necessarily have the same effects. A first distinction between different frequency formats is that between using a fixed numerator (1 out of x) and using a fixed denominator (e.g., x out of 100 or x out of 1000). Studies that looked at differences between these two formats reported contrasting findings. In line with the explanation that risks appear higher when it is easier to imagine them, that is, when they are more concrete, a few studies report that risks are perceived as higher when frequency formats use smaller denominators [34,35]. These findings suggest that, because it is easier to visualize a group of 10 persons of whom 1 is ill than it is to imagine a group of 1000 persons of whom 100 are ill, the former risk appears higher. Other studies, however, found that risks are perceived as higher when larger denominators are used, for example, 10 in 100 is perceived as larger than 1 in 10 [36]. People might be focusing mainly on the nominator *or* the denominator without assessing the size of the risk. While it thus may be important for risk perception to use a numerical format that refers to concrete events, more complex situations would account for people’s tendency to process information in the simplest and fastest way.

Other Numerical Formats: Relative Risk, Number Needed to Treat

As described earlier, comparing risks for, for example, treatment decisions requires

a presentation of absolute risks that allows for easy and intuitive comparisons (e.g., frequencies with the same denominator). The size of the difference between the risks is often explicitly communicated in order to clarify its importance for the decision. Typically, this difference can be presented as a relative risk or as an absolute risk reduction. The latter can be presented as just the absolute difference in risks (for example: “a risk reduction of 2%”), or as its reciprocal via the “number needed to treat” (for example: 50 patients need to be treated to prevent 1 case). In a large systematic review, Covey [37] found strong evidence that differences between risks are generally considered to be greater when they are expressed as relative risks instead of absolute differences between incidence rates, and that subsequently more patients agreed to treatment. This effect was particularly strong when the word “relative” was not explicitly mentioned when the risk reduction was communicated. For example, a 30% reduction of the risk of developing heart disease is perceived as much larger than an absolute reduction from 3% to 2%. So while relative risk reductions can be persuasive, they also easily create misperceptions about the true effects of a treatment or change in lifestyle, which may be at odds with informed decision making [38]. It is therefore commonly advised to present relative risk reductions only when the absolute risks are also clear. This clarifies the perceived effectiveness of treatment.

As for presenting absolute risks reductions, the “number needed to treat (or harm)” (NNT; i.e., the number of patients that need to be treated for one to benefit) has been advocated by some as a good and simple tool for communicating the effectiveness of a treatment [39–42]. However, there is no real evidence that suggests that people’s risk perceptions or decision-making benefits from using the NNT format as opposed to communicating the absolute risk reduction.

In conclusion, the studies on numerical risk formats show that two generic characteristics are important for optimizing risk perceptions and decision making. First, more concrete formats seem to evoke more meaningful risk images, which improves risk

Table 3. Pros and Cons of Numerical Risk Formats

Numerical Risk Formats	Pros	Cons
	Exact and precise Factual	Difficult to understand Not easy to imagine, abstract
	More trustworthy Transparent and neutral	
<i>Percentages</i>	When > 1% clear, in particular when reference class is made explicit Familiar format	Not easy to imagine when < 1%
<i>Frequencies</i>	More concrete (refer to real events/people) Easier for conditional risks	
1 of X	Easier to visualize, more concrete	More difficult for comparing risks
X of 100	Easier for comparing risk	Harder to imagine because larger group
<i>Odds, probabilities</i>	Useful for scientific communication	Difficult to imagine, less concrete Difficult to understand

perception. Therefore, if the size of single risks is clear, it seems advisable to use a frequency format. Lower denominators (e.g., “1 in X ”) may then yield the most concrete risk image, and numbers with decimals should be avoided. Second, in more complex situations, it is particularly important to minimize the complexity of risk information and to present risks so that errors are prevented. This means, for instance, that the risks that are most relevant for the specific decision need to be made explicit and that computations should be avoided. Further, the format needs to accommodate the task that is needed for decision making. If risks need to be compared or added, it is important to use either percentages or frequencies that share the same denominator. When risk estimates concern nested numbers, for example, joint probabilities or conditional risks, it may be particularly relevant to use one confined sample of persons or events to illustrate all possible outcomes, as this reduces cognitive effort. When presenting relative risk reductions, it is advisable to also present the absolute risks. This clarifies the perceived effectiveness of, for example,

treatments. See Table 3 for a listing of the pros and cons of using numerical risk formats.

USING GRAPHICS TO EXPRESS PROBABILITIES

Given the vagueness of verbal terms and the difficulty of numerical formats, alternative formats for risk communication have been explored. The use of visual displays or graphical formats is particularly popular in this respect. In their extensive review, Lipkus and Holland [43] describe three potential benefits of using graphical formats: (i) they allow the illustration of complex causal relations, trends in time, or quantitative part-to-whole proportions; (ii) they can evoke specific mathematical operations such as comparisons without cognitive effort; (iii) they are vivid and grab the attention so that they can be used to underline specific messages. Because of these assumed benefits, graphical formats are often used in daily practice. However, evidence about the degree to which they are really helpful in decision making is not unequivocally strong.

Graphical risk formats typically reflect the quantitative nature of risk information. The most common formats present these quantities either by the relative size on a standardized scale, for instance, as a bar chart, or by absolute and discrete numbers of icons that represent events, for example, in a population diagram. These specific features affect their suitability for conveying different types of information and as a result their effect may differ per decision situation. On a basic level, it has been found that when processing visual information, people are more sensitive to differences in height and areas than to differences in the angles between lines. Feldman-Stewart *et al.* [44], found that vertical bars were superior to pie charts, horizontal bars, or numbers, with respect to the accuracy and time efficiency when people had to compare two risks. This suggests that bar charts or similarly designed visual formats would be suitable for highlighting differences when several alternatives need to be compared. Population diagrams refer to discrete events and may make them, thus, more suitable for presenting information about a limited number of single risks.

Timmermans *et al.* [30] found that illustrating a single risk visually in a population diagram yielded stronger affective reactions to the risk, while comprehension was not affected. Such a stronger affective impact of risk information, possibly reflecting a higher awareness of the risks, has been observed more often. Others [45–47] found that participants behaved more risk averse when risks of accidents were illustrated visually as icons (stick figures, or other symbols to represent the number of affected persons) compared with numerical risk information only. Timmermans *et al.* [1] found that presenting risk in a population diagram made fewer patients opt for surgery (i.e., the more risky option) as opposed to watchful waiting than when risks were presented numerically. Graphical risk information may thus increase the saliency of risk information, especially when it concerns single risks and the format illustrates the risk in a discrete manner.

Apart from grabbing the attention, it is less clear whether visual information aids

decision making in more complex situations. Fagerlin *et al.* [48] did find that the use of population diagrams prevented people from basing decisions on anecdotal information more effectively than using numerical risk formats. However, in the study by Timmermans *et al.* [1] the aversion to surgical risks when a population diagram was shown made more people opt for watchful waiting despite the fact that this was the option with lower life expectancy. Other studies [27,29,49] found no effects of using bar charts or population diagrams on decision making. There is, thus, very limited evidence that suggests that visual risk formats substantially aid decision making.

The translation of quantitative risk information into a visually meaningful representation may look very different depending on the decision situation. This makes it difficult to merge findings from different studies and to assess the true value of visual risk formats. There is limited evidence as to which risk representation and which visual format may be most effective under which circumstances. As mentioned, the discrete nature of population diagrams may, in theory, make them more useful for communicating single risks than for more complex situations. However, their practical use may sometimes be limited. Risks below 1% would require diagrams that describe populations of 1000 or more persons or events, and this may take up a disproportional amount of space in a leaflet. Additionally, when various risks are involved and complex decisions are at hand, population diagrams allow little room for summarizing data and can easily become complicated. It may seem preferable to then use visual displays that can summarize data, such as bar charts, as they allow simplification of complex comparisons. Stacked bar charts can, for instance, represent total risks that consist of different outcomes and illustrate differences between several alternatives more concisely. Waters *et al.* [50] showed that such presentations may yield small benefits in terms of accuracy as compared to presenting only numbers.

In addition to bar charts, other summarizing visual formats are possible, such as

thermometer scales or survival curves or even 3D graphs. Whereas such formats may have appeal and could grab attention, there are few to no studies that have assessed their benefit or usefulness. Some studies did find that fairly complicated graphs, such as survival curves, are not easily understood, even after training [51]. This would limit the use of these complex graphs, as it may induce wrong interpretations.

In conclusion, in terms of the true benefits of graphical formats, the increased awareness of the risk information may be the most important one. In addition, visual displays can look appealing, and many patients indicate that they find them useful and appreciate it when they are used [43]. These aspects and effects may be important, given the tendency of people to ignore risk information in decision making. Visual formats may then be particularly useful when the primary purpose of risk communication is to increase awareness of risks. There is, however, limited evidence until date that suggests that graphical formats substantially improve understanding and aid decision making. See Table 4 for a listing of pros and cons of using graphical formats for risks.

Table 4. Pros and Cons of Graphical Risk Formats

Graphical Formats	
Pros	Cons
Effective in summarizing a lot of information	Difficult to determine which graphical format is best
Useful for clarifying relationships of risks (e.g., time trends)	No evidence for positive effect on understanding or informed decision making, some graphical formats are very difficult
Might reduce cognitive effort in processing risk information	Population diagrams not insightful for very small risks (e.g., 1 out of 1 000)
More vivid and draws one's attention	Might have an inappropriate effect by emphasizing information

GENERAL RECOMMENDATIONS FOR RISK COMMUNICATION

Based on the evidence about the effectiveness of different formats for risk communication, the following recommendations can be given. Risk formats should preferably be numerical (i.e., exact), as simple as possible (i.e., reducing cognitive complexity) and as concrete as possible (i.e., easy to imagine). In general, people prefer precise, numerical risk information, even though this might be difficult to comprehend. Further, it is important to reduce the cognitive complexity of the information because if the information is too complex people resort to heuristic processing, which sometimes leads to suboptimal decisions. It is important that risk information is easy to imagine so that people find the risk applicable to themselves and are more likely, for instance, to change their behavior. The 10 general recommendations for effective risk communication can be summarized as follows: *Be exact, make it concrete, and keep it simple!* Therefore, for the communication of the probability or quantitative aspect of risks, this means:

Be exact

1. *Risks should be quantified if possible.* Providing verbal descriptions only (e.g., “a very big chance”) or using colors (“you belong to the red group”) is not sufficient.
2. *When verbal terms* are used to express probabilities, be as neutral as possible and be aware of its goal when using verbal terms (e.g., to reassure a patient).

Make it concrete

3. *Do not use expressions of probability that do not refer to concrete events.* In other words: do not use percentages smaller than 1%, do not use probabilities (0–1), and do not use frequencies or percentages with decimals.
4. *Use frequencies if possible*, for example, “15 out of 100 people like you will suffer from heart disease within 10 years”, *not* “there is a

chance of 15% that you will suffer from heart disease within 10 years.”

5. *The reference group should be made explicit*, and when more than one risk is communicated the same reference group should be used. It should be clear to which group of people or events the probability refers, for example, “80% of the people with this disease will get better”, *not* “there is a chance of 80% of getting better.”
 6. *Use the smallest possible denominator*, when using frequencies to express single risks, for example, “1 out of 30 people will develop cancer”; *not* “3 out of 100 people will develop cancer”.
 7. *Shorter time frames* are preferred over longer time frames (five-year risk and not lifetime risk). However, this may lead to underestimating longer-term risks.
- Keep it simple
8. *Use the same format* when communicating several risks. Do not mix frequencies and percentages, for example, do *not* say “there is a 5% chance on mortality, and 10 out of 100 people report severe side effects”; instead, say “5 out of 100 people die, and 10 out of 100 people report severe side effects.”
 9. *Keep the denominator constant*, if several risks are given. It is easier for people to compare these risks, for example, “80 out of 100 people will get better, 4 out of 100 will experience side effects”; *not* “80 out of 100 people will get better, 1 out of 25 will experience side effects”.
 10. *The use of graphics* might be a good idea, although there is insufficient evidence that this improves people’s comprehension and decision making. It has been shown to have a larger affective impact than numerical information and may thus be used to motivate people to, for example, change their health behavior.

REFERENCES

1. Timmermans D, Molewijk B, Stiggelbout A, *et al.* Different formats for communicating surgical risks to patients and the effect on choice of treatment. *Patient Educ Couns* 2004;54:255–263.
2. Mazur DJ, Hickam DH. Patients’ interpretations of probability terms. *J Gen Intern Med* 1991;6:237–240.
3. Berry DC, Knapp PR, Raynor T. Is 15 per cent very common? Informing people about the risks of medication side effects. *Int J Pharm Pract* 2002;10:145–151.
4. Karelitz TM, Budescu DV. You say “Probable” and I say “Likely”: improving interpersonal communication with verbal probability phrases. *J Exp Psychol Appl* 2004;10:25–41.
5. Smits T, Hoorens V. How probable is probably? It depends on whom you’re talking about. *J Behav Decis Making* 2005;18:83–96.
6. Timmermans DR, Mileman PA. Lost for words: using verbal terms to express probabilities in oral radiology. *Dentomaxillofac Radiol* 1993;22:171–172.
7. Timmermans DRM. The role of experience and domain of expertise in using numerical and verbal probability terms in medical decisions. *Med Decis Making* 1994;14:146–156.
8. Knapp P, Raynor DK, Berry DC. Comparison of two methods of presenting risk information to patients about the side effects of medicines. *Qual Saf Health Care* 2004;13:176–180.
9. Mapes RE. Verbal and numerical estimates of probability in therapeutic contexts. *Soc Sci Med* 1979;13A:277–282.
10. Sutherland HJ, Lockwood GA, Trichter DL, *et al.* Communicating probabilistic information to cancer patients: is there ‘noise’ on the line? *Soc Sci Med* 1991;32:725–731.
11. France J, Keen C, Bowyer S. Communicating risk to emergency department patients with chest pain. *Emerg Med J* 2008;25:276–278.
12. Wallsten TS, Budescu DV. A review of human linguistic probability processing: General principles and empirical evidence. *Knowl Eng Rev* 1995;10:43–62.
13. Gurmankin AD, Baron J, Armstrong K. Intended message versus message received in hypothetical physician risk communications: exploring the gap. *Risk Anal* 2004;24:1337–1347.
14. Gurmankin AD, Baron J, Armstrong K. The effect of numerical statements of risk on

- trust and comfort with hypothetical physician risk communication. *Med Decis Making* 2004;24:265–271.
15. Berry D, Raynor T, Knapp P, *et al.* Over the counter medicines and the need for immediate action: a further evaluation of European Commission recommended wordings for communicating risk. *Patient Educ Couns* 2004;53:129–134.
 16. Berry DC, Raynor DK, Knapp P. Communicating risk of medication side effects: an empirical evaluation of EU recommended terminology. *Psychol Health Med* 2003;8:251–263.
 17. Teigen KH, Brun W. The directionality of verbal probability expressions: effects on decisions, predictions, and probabilistic reasoning. *Organ Behav Hum Decis Process* 1999;80:155–190.
 18. Wallsten TS, Budescu DV, Zwick R, *et al.* Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bull Psychon Soc* 1993;31:135–138.
 19. Yamagishi K. When a 12.86% mortality is more dangerous than 24.14%: implications for risk communication. *Appl Cogn Psychol* 1997;11:495–506.
 20. Wertz DC, Sorenson JR, Heeren TC. Clients' interpretation of risks provided in genetic counseling. *Am J Hum Genet* 1986;39:253–264.
 21. Goodyear-Smith F, Arroll B, Chan L, *et al.* Patients prefer pictures to numbers to express cardiovascular benefit from treatment. *Ann Fam Med* 2008;6:213–217.
 22. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *Br Med J* 2003;327:741–744.
 23. Gigerenzer G, Hertwig R, van den Broek E, *et al.* "A 30% Chance of Rain Tomorrow": how does the public understand probabilistic weather forecasts? *Risk Anal* 2005;25:623–629.
 24. Slovic P, Monahan J, MacGregor DG. Violence risk assessment and risk communication: the effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law Hum Behav* 2000;24:271–296.
 25. Tan SB, Goh C, Thumboo J, *et al.* Risk perception is affected by modes of risk presentation among Singaporeans. *Ann Acad Med Singapore* 2005;34:184–187.
 26. Fair AKI, Murray PG, Thomas A, *et al.* Using hypothetical data to assess the effect of numerical format and context on the perception of coronary heart disease risk. *Am J Health Promot* 2008;22:291–296.
 27. Waters EA, Weinstein ND, Colditz GA, Emmons KM. Aversion to side effects in preventive medical treatment decisions. *Br J Health Psychol* 2007;12:383–401.
 28. Monahan J, Heilbrum K, Silver E, *et al.* Communicating violence risk: frequency formats, vivid outcomes, and forensic settings. *Int J Forensic Ment Health* 2002;1:121–126.
 29. Zikmund-Fisher BJ, Ubel PA, Smith DM, *et al.* Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient Educ Couns* 2008;73:209–214.
 30. Timmermans DRM, Ockhuysen-Vermeij CF, Henneman L. Presenting health risk information in different formats: the effect on participants' cognitive and emotional evaluation and decisions. *Patient Educ Couns* 2008;73:443–447.
 31. Gigerenzer G. *Reckoning with risk. Learning to live with uncertainty.* London, England: Penguin Books; 2002.
 32. Hoffrage U, Lindsey S, Hertwig R, *et al.* Communicating statistical information. *Science* 2002;290:2261–2262.
 33. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med* 1998;73:538–540.
 34. Zhai G, Suzuki T. Effects of risk representation and scope on willingness to pay for reduced risks: evidence from Tokyo Bay, Japan. *Risk Anal* 2008;28:513–522.
 35. Halpern DF, Blackman S, Salzman B. Using statistical risk information to assess oral contraceptive safety. *Appl Cogn Psychol* 1989;3:251–260.
 36. Denes-Raj V, Epstein S. Conflict between intuitive and rational processing: when people behave against their better judgment. *J Pers Soc Psychol* 1994;66:819–829.
 37. Covey J. A meta-analysis of the effects of presenting treatment benefits in different formats. *Med Decis Making* 2007;27:638–654.
 38. Berry DC, Knapp P, Raynor T. Expressing medicine side effects: assessing the effectiveness of absolute risk, relative risk, and number needed to harm, and the provision of baseline risk information. *Patient Educ Couns* 2006;63:89–96.
 39. Weinstein ND. Unrealistic optimism about susceptibility to health problems. *J Behav Med* 1982;5:441–460.

40. Weinstein ND. Why it won't happen to me: perceptions of risk factors and susceptibility. *Health Psychol* 1984;3:431–457.
41. Weinstein ND. Optimistic biases about personal risks. *Science* 1989;246:1232–1233.
42. Kristiansen IS, Gyrd-Hansen D, Nexoe J, *et al.* Number needed to treat: Easily understood and intuitively meaningful? Theoretical considerations and a randomized trial. *J Clin Epidemiol* 2002;55:888–892.
43. Lipkus IM, Hollands JG. The visual communication of risk. *J Natl Cancer Inst Monogr* 1999;25:149–163.
44. Feldman-Stewart D, Kocovski N, McConnell BA, *et al.* Perception of quantitative information for treatment decisions. *Med Decis Making* 2000;20:228–238.
45. Stone ER, Yates JF, Parker AM. Effects of numerical and graphical displays on professed risk-taking behavior. *J Exp Psychol Appl* 1997;3:243–256.
46. Stone ER, Sieck WR, Bull BE, *et al.* Foreground: background salience: Explaining the effects of graphical displays on risk avoidance. *Organ Behav Hum Decis Process* 2003;90:19–36.
47. Chua HF, Yates JF, Shah P. Risk avoidance: graphs versus numbers. *Mem Cognit* 2006;34:399–410.
48. Fagerlin A, Wang C, Ubel PA. Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? *Med Decis Making* 2005;25:398–405.
49. Edwards A, Thomas R, Williams R, *et al.* Presenting risk information to people with diabetes: evaluating effects and preferences for different formats by a web-based randomised controlled trial. *Patient Educ Couns* 2006;63:336–349.
50. Waters EA, Weinstein ND, Colditz GA, *et al.* Formats for improving risk communication in medical tradeoff decisions. *J Health Commun* 2006;11:167–182.
51. Armstrong K, Schwartz JS, Fitzgerald G, *et al.* Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. *Med Decis Making* 2002;22:76–83.

DIFFERENTIAL GAMES

NGO VAN LONG
Department of Economics,
McGill University,
Montreal, Quebec,
Canada

Differential games (DG) are dynamic games that take place in continuous-time. The theory of differential games is a natural extension of optimal control theory (or its predecessor, the calculus of variations) and is a notable example of cross fertilization between game theory and theory of dynamic optimization. This theory provides a powerful tool of analysis in operations research, management science, and economics. Ref. 1 gives comprehensive exposition of the theory and techniques, including applications in various fields. As a subfield of game theory, the theory of differential games adds to the classical game theory a more structured dynamic dimension and provides well-developed methods of solving an important class of dynamical game-theoretic problems. As pointed out in Ref. 1, this theory owes much to the pioneering work of Isaacs in the 1950s.

BASIC FEATURES OF DIFFERENTIAL GAME

In optimal control theory [2], a single decision maker faces a changing environment represented by a vector of m state variables, denoted by $x(t) \in X$.

He chooses a time path of his vector of control variables $u(t)$ to maximize his payoff which is the integral of a stream of discounted benefits $b(t)$ plus the discounted value of a terminal scrap value $S(x_T, T)$ where T denotes the terminal time. The benefit $b(t)$ is a function of the state variables, the control variables, $b(t) = B(x(t), u(t), t)$. The evolution of the state variables is described by a system of differential equations, with given initial conditions.

Let $r > 0$ be the decision maker's discount rate. The payoff of the decision maker is

$$W = \int_0^T \exp(-rt) B(x(t), u(t), t) dt \\ + \exp(-rT) S(x_T, T).$$

The control vector must belong to a specified set U . The terminal vector x_T may be fixed, or free, or restricted to a specified set X_T .

In a differential game, there are n decision makers who typically do not cooperate with each other. We call them the players of the game. Let $N \equiv \{1, 2, \dots, n\}$ denote the set of players. Player i has a vector of control variables u_i . The benefit function of player i generally depends not only on x and u_i but also on the control variables of other players:

$$b_i(t) = B_i(x(t), u_1(t), \dots, u_n(t), t).$$

The evolution of the state variable x_k for $k = 1, 2, \dots, m$ generally depends on the actions of all the players:

$$\frac{dx_k}{dt} = f_k(x, u, t).$$

The initial condition is $x_k = x_0$.

In vector notation,

$$\frac{dx}{dt} = f(x, u, t) \text{ with initial condition } x(0) = x_0.$$

The payoff of decision maker i is

$$W_i = \int_0^T \exp(-r_i t) B_i(x, u_1, u_2, \dots, u_n, t) dt \\ + \exp(-r_i T) S_i(x_T, T).$$

In a noncooperative differential game, each player i takes the strategies of other players as given, and chooses a time path for his control vector u_i to maximize his payoff. The simplest class of strategies is that of "open-loop strategies," and

the corresponding equilibrium concept is open-loop Nash equilibrium.

OPEN-LOOP STRATEGIES AND OPEN-LOOP NASH EQUILIBRIUM

An open-loop strategy of player i is a function of time alone, denoted by $\phi_i(t)$. If player i chooses $\phi_i(t)$, this function determines the actions that he takes at each time $t \in [0, T]$. Thus,

$$u_i(t) = \phi_i(t).$$

Choosing $\phi_i(t)$ means that player i is committed to a time path of action. For this reason, open-loop strategies are sometimes called “path strategies,” as in Ref. 3.

Let $\phi = (\phi_1, \phi_2, \dots, \phi_n)$ denote the vector of open-loop strategies chosen by the players. Given x_0 and ϕ , assume the differential equation system yields a unique solution path $x^*(t)$. The payoff of player i is then

$$W_i(x_0, \phi) = \int_0^T \exp(-r_i t) B_i(x^*(t), \phi(t), t) dt + \exp(-r_i T) S_i(x_T, T).$$

Given $\phi = (\phi_1, \phi_2, \dots, \phi_n)$, we can represent it as $(\phi_i, \bar{\phi}_{-i})$. Define an open-loop Nash equilibrium as a strategy profile $\bar{\phi}$ such that no player can gain by deviating from it, that is, for all i it holds that

$$W_i(x_0, \bar{\phi}) \geq W_i(x_0, \phi_i, \bar{\phi}_{-i}).$$

for all admissible strategies ϕ_i .

To find an open-loop Nash equilibrium, the usual method is to employ the Maximum Principle to derive the necessary conditions for the optimal time path of action of each player, given the open-loop strategies of other players. Sufficiency conditions must also be verified. Finally, one searches for a fixed point. For relatively simple games, an open-loop Nash equilibrium can be found analytically. In other situations, one must rely on numerical methods [4,5].

Example 1. Competition under sticky price using open-loop strategies. Consider a game between two firms

producing a homogeneous good. Let $u_i(t)$ denote the output of firm i at time t . Its production cost is $C(u_i) = cu_i - \frac{1}{2}(u_i)^2$. The total supply is $S(t) = u_1(t) + u_2(t)$. Given the price $p(t)$ the market demand is $D(t) = a - p(t)$. The price is sticky and does not adjust instantaneously to equate demand with supply. Assume that the rate of change of the price is proportional to the excess demand:

$$\frac{dp}{dt} = \theta(a - p(t) - u_1(t) - u_2(t)).$$

The parameter $\theta > 0$ is the speed of adjustment.

Each firm takes the above law of price adjustments as given, and seeks to maximize the integral of its stream of discounted profit.

Assume that firms use open-loop strategies. Each firm i chooses a time path of its output $u_i(t)$, taking as given the time path of its rival's output. Let $r > 0$ be the rate of discount, common to both firms. Firm i solves the following optimal control problem:

$$J_i = \max \int_0^\infty \exp(-rt) (p(t)u_i(t) - C(u_i(t))) dt.$$

subject to

$$\begin{aligned} \frac{dp}{dt} &= \theta(a - p(t) - u_i(t) - \phi_j(t)), \\ \text{where } p(0) &= p_0 \text{ (given).} \end{aligned}$$

One sets up the Hamiltonian function for firm i , with λ_i as the current-value costate variable:

$$\begin{aligned} H_i &= \exp(-rt) \left[pu_i(t) - cu_i(t) - \frac{1}{2}(u_i(t))^2 \right] \\ &+ \exp(-rt) \lambda_i(t) \theta [a - p(t) - u_i(t) - \phi_j(t)]. \end{aligned}$$

The necessary conditions are

$$\begin{aligned} p - c - u_i - \lambda_i u_i &= 0 \\ \frac{d\lambda_i}{dt} &= r\lambda_i - u_i + \theta\lambda_i \\ \frac{dp}{dt} &= \theta[a - p - u_i - \phi_j]. \end{aligned}$$

The transversality condition is

$$\lim_{t \rightarrow \infty} \exp(-rt) \lambda_i(t) = 0.$$

Since the two firms are identical, consider a symmetric open-loop Nash equilibrium, where $u_i = u_j = u$. Solving the above equations, one obtains the open-loop strategies. It can be shown that along the equilibrium path, the state variable converges to the steady-state value

$$p_{\infty}^{OL} = \frac{a(2\theta + r) + 2c(\theta + r)}{3r + 4\theta}.$$

Notice that if $r \rightarrow 0$ or $\theta \rightarrow \infty$ the steady-state value will be the same as the static Cournot equilibrium price, $p^C = \frac{a+c}{2}$.

Interestingly, in the opposite case where $r \rightarrow \infty$ or $\theta \rightarrow 0$, the steady-state value will be the same as the perfect competition price, where each firm's output equates its marginal cost to the price.

A nice property of open-loop equilibria is that they satisfy a property called *time-consistency*. This means that given an open-loop Nash equilibrium strategy profile $\bar{\phi}$, if at any point of time $s \in (0, T)$ player i reconsiders the game and wants to maximize his payoff starting from time s , he would find that the solution is to use the same strategy, given that all other players have not deviated from their equilibrium strategies. The time-consistency property follows from the fact that Bellman's Principle of Optimality applies to each player's optimization problem.

On the other hand, if by mistake some players have deviated from his planned course of action (or perhaps if there is some unexpected shock in the system) so that the stock size x_s is different from what was anticipated at time zero, then at s the players will in general find that they would be better off by switching to another strategy. We conclude that open-loop Nash equilibria are not robust to "trembling-hand" deviations. One may say that open-loop Nash equilibria are not "subgame perfect." For this reason, many researchers prefer a different equilibrium concept, called *Markov-perfect Nash equilibrium*. Before explaining this concept, we must define the concept of Markovian strategies (sometimes called *closed-loop strategies*).

MARKOVIAN STRATEGIES, MARKOVIAN NASH EQUILIBRIUM, AND MARKOV-PERFECT NASH EQUILIBRIUM

A Markovian strategy of player i is a function ψ_i that determines at any (state, time) pair (x, t) the action that he takes at t :

$$u_i(t) = \psi_i(x(t), t).$$

Markovian strategies are sometimes called *decision-rule strategies* (or *closed-loop strategies*) since under Markovian strategies, the action of player i at time t is conditioned on the currently observed value of the vector of state variables x . In the degenerate case where a Markovian strategy is insensitive to x , that is, it does not contain x as an argument, then it is an open-loop strategy.

Given a profile of Markovian strategies, one for each player, $\psi = (\psi_1, \psi_2, \dots, \psi_n)$ the system of differential equations is

$$\frac{dx}{dt} = f(x(t), \psi(x(t), t))$$

with initial condition $x(0) = x_0$.

Assume that this system has a unique solution, denoted by $x^{**}(t)$. The payoff of player i is

$$J_i(x_0, \psi) = \int_0^{\infty} \exp(-r_i t) B_i(x^{**}(t), \psi(x^{**}(t), t), t) dt + \exp(-r_i T) S_i(x_T^{**}, T).$$

Define a Markovian Nash equilibrium as a strategy profile $\bar{\psi}$ such that no player can gain by deviating from it, that is, for all $i \in N$, it holds that

$$J_i(x_0, \bar{\psi}) \geq J_i(x_0, \psi_i, \bar{\psi}_{-i}),$$

for all admissible strategies ψ_i . Note that this inequality is required to hold only for a given $x(0) = x_0$, the initial condition at time zero. Markovian Nash equilibrium are sometimes called *decision-rule Nash equilibrium*.

Markovian Nash equilibria are time-consistent, for the same reason that makes open-loop Nash equilibria time-consistent. Also, like open-loop Nash equilibria, without further restrictions Markovian Nash

equilibria are not robust to trembling-hand deviations. In other words, although the actions induced by a Markovian Nash equilibrium strategy profile are credible along the equilibrium trajectory x^{**} they are in general not credible when the system is off the equilibrium path. [A simple example is given in Ref. 1 (p. 102.)] For this reason, a stronger property is desirable: subgame perfection, a concept attributed to Selten [6]. A Markovian Nash equilibrium that satisfies the subgame-perfect property is called a *Markov-perfect Nash equilibrium*.

To introduce the idea of subgame perfection in continuous-time, consider any arbitrary (state, date) pair (x, t) , and the system of differential equations

$$\frac{dx(\tau)}{d\tau} = f(x, \psi(x, \tau), \tau) \text{ for } \tau \geq t \text{ and } x(t) = x.$$

Assume this equation yields a unique solution. Define the performance index for player i at the (state, date) pair by

$$V_i(x, t, \psi) = \int_t^T \exp(-r(\tau - t)) B_i(x(\tau), \psi(x(\tau), \tau), \tau) d\tau + \exp(-r(T - t)) S_i(x_T, T).$$

Define a Markov-perfect Nash equilibrium (also called a *feedback Nash equilibrium*) as a strategy profile $\bar{\psi}$ such that, at any (state, date) pair (x, t) , the following inequality holds

$$V_i(x, t, \bar{\psi}) \geq V_i(x, t, \psi_i, \bar{\psi}_{-i}),$$

for all admissible ψ_i .

It is important to stress the requirement that this inequality hold for all possible (state, date) pairs, not just for the initial pair at time zero. To be Markov-perfect, a Markovian Nash equilibrium must satisfy the property that the continuation of the given decision rules constitutes a Nash equilibrium when viewed from any future (date, state) pair. This requirement corresponds to the concept of subgame-perfect Nash equilibrium originally proposed by Selten [6] for games that take place in discrete-time.

To find a Markov-perfect Nash equilibrium, one normally makes use of the Hamilton–Jacobi–Bellman (HJB) equations. Let V_i denote the value function for player i .

The HJB equation for player i is

$$r_i V_i(x, t) - \frac{\partial V_i}{\partial t} = \max_{u_i} \left[B_i(x, u_i \cdot \psi_{-i}, t) + \frac{\partial V_i}{\partial x} f(x, u_i, \psi_{-i}, t) \right],$$

with the boundary condition $V_i(x, T) = S_i(x, T)$.

If T is infinite, the above boundary condition is replaced by

$$\lim_{t \rightarrow \infty} \exp(-r_i t) V_i(x(t), t) = 0.$$

In an important class of infinite-horizon differential games, both the objective function and the transition equation are independent of time. For such problems, called *infinite-horizon autonomous* problems, it is natural to focus on stationary strategies (i.e., strategies that specify the control variables as functions of the state variable only, independent of time). In that case, the value functions are also independent of time.

Example 2. Competition under sticky price using Markovian strategies.

Return to Example 1 above, and modify the assumption about the behavior of firms. Assume that they use decision-rule strategies, so that $u_i(t) = \psi_i(x(t), t)$. Since the problem is linear quadratic, it is natural to conjecture that the value function is quadratic in the state variable. Furthermore, since in this problem time does not enter directly in the benefit function B_i and the transition equation, one expects that the value function is independent of time. Then the HJB equation for firm i is

$$rV_i(p) = \max_{u_i} \left[pu_i - cu_i - \frac{1}{2}(u_i)^2 + \frac{dV_i}{dp} \theta(a - p - u_i - \psi_j(p)) \right].$$

Let $V_i = A_i + B_i p + D_i p^2$ where A_i, B_i , and D_i are to be determined. Focus on the linear Markovian strategies of the form $u_i(t) = \alpha_i + \beta_i p(t)$. Then it can be shown that in equilibrium the firms use the same linear Markovian strategy, with

$$\begin{aligned}
\alpha_i &= \alpha_j = \alpha = (1 - \theta K) > 0 \\
\beta_i &= \beta_j = \beta = \theta E - c < 0 \\
K &\equiv \frac{r + 6\theta - \sqrt{(r + 6\theta)^2 - 12\theta^2}}{6\theta^2} > 0 \\
E &= \frac{c - a\theta K - 2c\theta K}{r + 3\theta + 3K\theta^2}.
\end{aligned}$$

(Here the positive root K is chosen to ensure convergence of p to a steady-state).

Notice that $\alpha > 0$, indicating that the equilibrium decision-rule strategies instruct firms to produce more if they observe a higher price. This suggests that firms are more competitive in the Markov-perfect Nash equilibrium than in the open-loop Nash equilibrium. To verify this, compute the steady-state price under the Markov-perfect Nash equilibrium,

$$p_\infty^M = \frac{a + 2c - 2\theta E}{3 - 3\theta K}.$$

In the limiting case, where $r \rightarrow 0$ or $\theta \rightarrow \infty$, one can see that the Markov-perfect steady-state price is lower than the steady-state price under open-loop Nash equilibrium. The reason is that if each firm expects its rival to produce more when the price is higher, then it has little incentive to restrict its output to raise the price.

It is worth noting that open-loop Nash equilibrium and Markov-perfect Nash equilibrium can be thought of as based on two alternative assumptions about the ability of players to precommit. In an open-loop Nash equilibrium, players commit to a whole time path of actions. In a Markov-perfect Nash equilibrium, players cannot precommit at all. One can argue that in some cases, players may be able to commit to actions in the near futures (e.g., by forward contracts), but not to actions in the distant future [3]. Think of a continuous-time model where a game begins at time 0 and ends at a fixed time T , and there are k periods of equal length δ , where $k\delta = T$. At the beginning of each period, agents can commit to a path of action during that period. The special case where $k = 1$ corresponds to the open-loop formulation, and open-loop Nash equilibrium is the appropriate equilibrium concept. At the

other extreme, where $\delta \rightarrow \infty$ the appropriate equilibrium concept is Markov-perfect Nash equilibrium.

The choice of equilibrium concepts is to some extent dependent on tractability. The relative ease of finding an open-loop Nash equilibrium is one of its attractive features.

STACKELBERG EQUILIBRIUM

There are dynamic-game situations where players do not make their moves simultaneously. Consider a two-player game where Player 1 has priority of moves over Player 2. We call Player 1 the leader and Player 2 the follower. The leader is sometimes called the *Stackelberg leader*, a term borrowed from economics, and the equilibrium concept is called the *Stackelberg equilibrium*.

There are several distinct notions of Stackelberg equilibrium, depending on what strategies are admitted and on the exact nature of the priority of moves.

Open-Loop Stackelberg Equilibrium

When Player 1 can announce at the start of the dynamic game the whole sequence of her moves and is able to commit to it, we say she is the “open-loop Stackelberg leader.” Player 2 takes the time path of action chosen by the leader as given, and chooses his own time path of action in response to maximize his payoff. The open-loop strategy chosen this way by the follower is called his “reaction” to the leader’s path strategy. The leader knows the follower’s reaction to each of her possible time paths. She then chooses her best time path, that is, the one that maximizes her payoff, taking into account the response of the follower.

Technically, a common method of finding the open-loop Stackelberg equilibrium is to treat the follower’s necessary conditions as constraints that the leader faces in formulating her optimization problem. Consider a game where x is the state variable (a scalar) and u_1 and u_2 are the control variables of the leader and the follower respectively. Let λ denote the follower’s costate variable, and express u_2 as a function of λ and x . The equation of motion of the costate variable λ in

the follower's necessary conditions is an additional differential equation that the leader faces. She must therefore treat this costate variable, λ , as a second state variable. The leader then solves an optimal control problem involving two state variables, x and λ . Let μ and γ be her costate variables associated with these state variables. An interesting feature of the leader's optimization problem is that while x_0 is given, λ_0 is not. In general, the whole time path of λ depends on the leader's time path u_1 (which is known to the follower at the start of the game). Therefore λ_0 generically depends on the time path u_1 . When this is the case, we say that λ_0 is controllable [1, Chapter 5]. When λ_0 is controllable, the leader will want to influence it so as to maximize her payoff. This means that her optimal choice of λ_0 , denoted by λ_0^* , must be such that any small deviation of λ_0 from λ_0^* would not improve her payoff. This implies that the leader's initial shadow price $\gamma(0)$ must be zero. Therefore $\gamma(0) = 0$ is a necessary condition for the leader, if λ_0 is controllable.

An undesirable property of open-loop Stackelberg equilibrium is that it is generally time-inconsistent. To see this, recall that $\gamma(0) = 0$ is a necessary condition for the leader. Typically, $\gamma(t)$ is not identically zero for all $t > 0$. Consider some time t_1 such that $\gamma(t_1)$ is different from zero. If at time t_1 the leader is allowed to revise her path strategy, the necessary conditions of her optimization problem at t_1 would imply that she would want the new $\gamma(t_1)$ to be zero. Resetting $\gamma(t_1)$ from a nonzero value to zero implies resetting the whole time path of her control variable $u_1(t)$ for $t \geq t_1$. This establishes the time-inconsistency of open-loop Stackelberg equilibrium.

The time-inconsistency of the leader's initial plan would be known by any smart follower. It follows that the leader's plan is not credible. For this reason, alternative concepts of Stackelberg equilibrium must be found.

Stagewise Stackelberg Equilibrium

We define a stagewise Stackelberg leader is a leader that cannot credibly announce a whole time path of her action. Instead, it is assumed that at each date t , she chooses $u_1(t)$ before

Player 2 chooses his move. Thus the leader only leads at each stage (date), but does not have the overall leadership.

To solve for a stagewise Stackelberg equilibrium, dynamic programming can be used. The solution method is most easily understood in a setting with discrete-time and a finite horizon. In the last period, a static Stackelberg equilibrium is found. The payoff of each player for the last period can be computed as a function of the opening stock of the last period. Working backwards, the equilibrium strategies for the second-last period can be found, and so on.

In continuous-time, the solution method for a stagewise Stackelberg equilibrium is via the use of the HJB equations, which are not solved simultaneously as in the case of the Markov-perfect Nash equilibrium. Instead, the HJB equation for the follower is considered first, to express the control variable of the follower as a function of the leader's control variable and the parameters of the follower's value function. Such a "reaction function" is then fed into the leader's HJB equation, and the necessary conditions for the leader's optimal control are obtained. These are then used to solve for the parameters of the follower's value function [7].

Stagewise Stackelberg equilibria are time-consistent because this equilibrium concept is based on Bellman's Principle of Optimality.

Global Stackelberg Equilibrium with Decision-Rule Strategies

An alternative equilibrium concept for leader-follower games is global Stackelberg leadership, where the leader uses a decision rule that determines her control variable as a function of the state variable. The set of decision rules from which the leader chooses her strategy is normally restricted, for example, her control variable must be a linear affine function of the state variable, where the two parameters are to be chosen optimally.

Example 3. Taxing a polluting monopolist. In this game, the leader is a government that wants to maximize social welfare and the follower is a monopolist that produces an

output which emits pollution as a by-product. The state variable is the stock of pollution, denoted by $X(t)$. The control variable of the follower is the output, $y(t)$, which is equal to the emission rate. The pollution stock evolves according to the differential equation

$$\frac{dX}{dt} = y - \delta X.$$

Here $\delta > 0$ is the rate of decay of the pollution stock. Consumers suffer damages caused by the pollution stock. Their aggregate damage cost is $\frac{\sigma}{2}X^2$, where σ is a positive parameter.

The government's control variable is the tax rate $\theta(t)$ per unit of output. The government's choice of the tax is restricted to a family of linear affine function $\theta(t) = \alpha X(t) + \beta$, where α and β are to be chosen optimally.

Assume production cost is c per unit. The demand curve is given by $p = A - By$ where p is the price, and A and B are positive constants. The profit of the monopolist is $\pi(t) = [p(t) - c - \theta(t)]y(t)$. The monopolist chooses the time path of output to maximize

$$\int_0^\infty \exp(-\rho t)[(A - c)y - By^2 - (\alpha X + \beta)y]dt$$

subject to $\frac{dX}{dt} = y - \delta X$.

Assume $A - c > \beta$ and $\alpha > 0$. The monopolist's optimal control problem has a unique solution. The pollution stock converges to a steady-state

$$X_\infty(\alpha, \beta) = \frac{(A - c - \beta)(\rho + \delta)}{\alpha\rho + 2\alpha\delta + 2\delta B(\rho + \delta)}.$$

The time path of the pollution stock is

$$X(t) = (X_0 - X_\infty)\exp(\mu(\alpha)t) + X_\infty,$$

where

$$\mu(\alpha) \equiv \frac{1}{2}(\rho - \sqrt{\Delta}) < 0$$

and

$$\Delta \equiv (\rho + 2\delta)^2 + \frac{2\alpha(\rho + 2\delta)}{B}.$$

The monopolist's output can be expressed in the feedback form

$$y(X, \alpha, \beta) = \delta X_\infty(\alpha, \beta) + [X - X_\infty(\alpha, \beta)][\delta + \mu(\alpha)].$$

The government's objective is to maximize the integral of discounted welfare $w(t)$ defined as the sum of consumer surplus net of pollution damage costs, producer surplus, and tax revenue

$$w = U(y) - p(y)y - \frac{\sigma}{2}X^2 + \pi + \theta y,$$

where $U(y) \equiv \int_0^y p(y)dy$. After simplification, the government's objective function becomes

$$\max \int_0^\infty \exp(-\rho t)[U(y) - cy - (\sigma/2)X^2]dt.$$

The government chooses the parameters α and β to maximize this objective function, knowing that $y = y(X, \alpha, \beta)$. The solution is

$$\alpha^* = \frac{2\sigma}{\rho + 2\delta} \quad \text{and} \\ \beta^* = \frac{[\sigma - B(\rho + \delta)\delta - (\rho + 2\delta)\alpha^*]X_\infty}{\rho + \delta},$$

where

$$X_\infty = \frac{(A - c)(\rho + \delta)}{\sigma + B\delta(\rho + \delta)}.$$

It can be verified that this solution yields a time path of pollution that is identical to what a benevolent dictator who can directly control y would choose. Therefore the linear tax rule found above achieves the first-best outcome. It follows that the commitment to such a tax rule is credible.

EXTENSIONS OF DIFFERENTIAL GAMES

Non-Markovian Strategies

Markovian strategies are restrictive: players cannot condition their actions on the full history of the game. Non-Markovian strategies allow players to make threats which can help enforce cooperation. A powerful type of non-Markovian strategies is called *trigger strategy*: a player can threaten to switch from a cooperative mode to a punishment mode if

he observes that other players have deviated from a cooperative mode. It is assumed that if a player defects at time t , other players will start to punish him at time $t + \varepsilon$, where ε is a fixed positive time delay. As shown in Ref. 1, if the time delay is small enough and the rate of discount is not too high, trigger strategies can ensure full cooperation.

Stochastic Differential Games

Uncertainty can be incorporated into differential games. Two common ways of modeling uncertainty are (i) allowing a small random disturbance in the evolution of the state variables, often represented by a Wiener process, and (ii) introducing piecewise deterministic processes, where uncertain changes in the environment occur at unknown isolated points of time. Both modeling strategies are discussed in Ref. 1.

REFERENCES

1. Dockner EJ, Jørgensen S, Long NV, Sorger G. Differential games in economics and management science. Cambridge: Cambridge University Press; 2000.
2. Leonard D, Long NV. Optimal control theory and static optimization in economics. Cambridge: Cambridge University Press; 1992.
3. Reinganum JF, Stokey NL. Oligopoly extraction of a common property natural resource: the importance of period of commitment in dynamic games. *Int Econ Review* 1985;26: 161–173.
4. Jørgensen S, Zaccour G. Development in differential game theory and numerical methods: economic and management applications. *Comput Manag Sci* 2007;4(2):159–181.
5. Jørgensen S, Zaccour G. Differential games in marketing. Boston: Kluwer Academic Publishers; 2004.
6. Selten R. Reexamination of perfectness concepts for equilibrium points in extensive form games. *Int J Game Theor* 1975;4(1): 25–55.
7. Başar T, Haurie A, Ricci G. On the dominance of capitalists leadership in a “feedback Stackelberg” solution of a differential game model of capitalism. *J Econ Dynam Contr* 1985;9:101–125.

DIRECT SEARCH METHODS

ROBERT MICHAEL LEWIS
Department of Mathematics, College of
William & Mary, Williamsburg,
Virginia

VIRGINIA TORCZON
Department of Computer Science,
College of William & Mary,
Williamsburg, Virginia

INTRODUCTION

Direct search methods were first formally proposed in the late 1950s and early 1960s [1–3]. They have remained popular with users due to their simplicity and their practical success on a variety of problems. Well-known examples of some of the original direct search methods include the simplex method of Nelder and Mead [4] (not to be confused with the simplex algorithm of linear programming) and the pattern search method of Hooke and Jeeves [2].

It is difficult to give a succinct, precise characterization of direct search methods. The term *direct search* was introduced by Hooke and Jeeves in 1961 [2] in connection with their algorithm for unconstrained optimization:

We use the phrase “direct search” to describe sequential examination of trial solutions involving comparison of each trial solution with the “best” obtained up to that time together with a strategy for determining (as a function of earlier results) what the next trial solution will be. The phrase implies our preference, based on experience, for straightforward search strategies which employ no techniques of classical analysis except where there is a demonstrable advantage in doing so.

From this perspective, direct search involves the comparison of each trial solution with the best previous solution. The stated preference

for “straightforward search strategies which employ no techniques of classical analysis” means that direct search methods do not rely on models or approximations, such as the linear (first-order) model used by steepest descent or the quadratic (second-order) model used by Newton’s method. Instead, direct search methods work *only* with values of the objective computed at various trial points, and for this reason are sometimes called *zeroth-order methods* since no derivatives are involved. Nevertheless, many direct search strategies share important mathematical features with derivative-based techniques such as steepest descent. In particular, a properly designed direct search method will enjoy analytical guarantees similar to those found for steepest descent [e.g., convergence to a Karush–Kuhn–Tucker (KKT) point], even without explicit use of the gradient [5,6].

Direct search methods are most easily understood for the case of unconstrained minimization,

$$\text{minimize } f(x), \quad (1)$$

where the objective $f : \mathbb{R}^n \rightarrow \mathbb{R}$. The discussions in sections titled “Compass Search for Unconstrained Minimization”, “Characteristics of Direct Search Methods”, and “Direct Search Methods for Unconstrained Minimization” focus on this case. Direct search techniques for problems with bound, linear, and nonlinear constraints are discussed in the section titled “Direct Search Methods for More General Optimization”.

COMPASS SEARCH FOR UNCONSTRAINED MINIMIZATION

The basic ideas underlying direct search methods are easily illustrated with *compass search*. Compass search is an exceedingly simple instance of a direct search method, yet one for which there are surprisingly

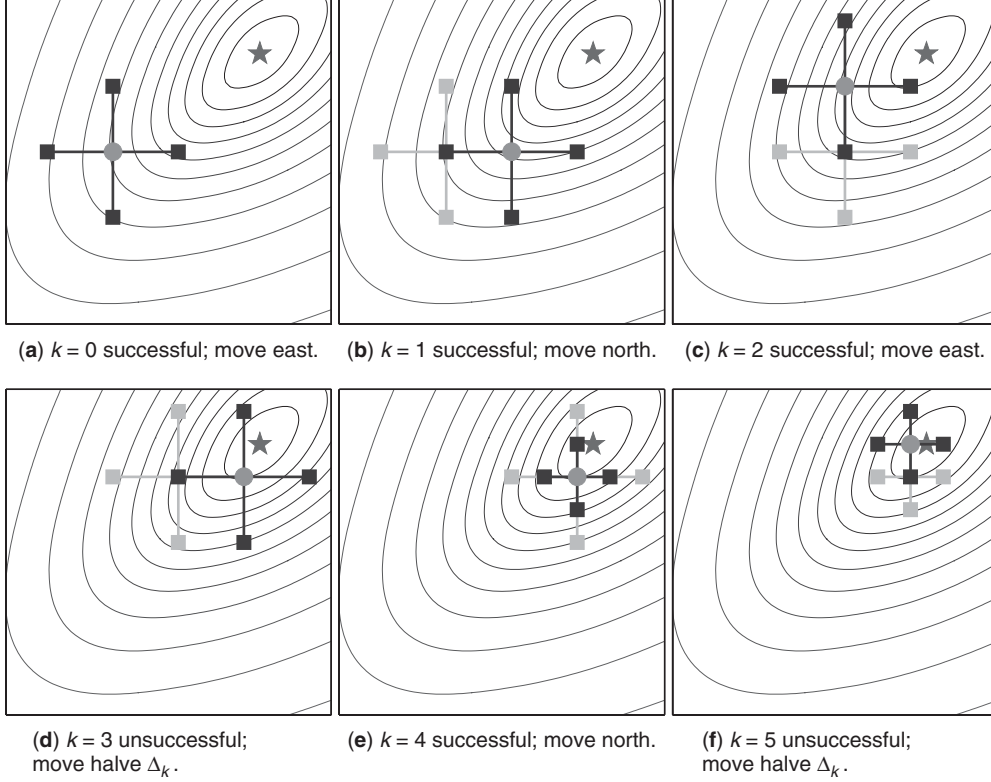


Figure 1. A version of compass search applied to the Broyden tridiagonal function.

strong convergence guarantees.

At the beginning of iteration k of compass search one has a current estimate x_k of the solution of Problem (1) and a parameter Δ_k that determines the length of trial steps. Let $e_1, e_2, \dots, e_n$ denote the standard unit basis vectors for $\mathbb{R}^n$. Compass search then proceeds with a local search around x_k , trying steps of the form $x_k \pm \Delta_k e_i$, $i = 1, \dots, n$, looking for a point that improves the objective. Any such point may be taken as the next iterate x_{k+1} . If no such point is found—and only if no such point is found—the step-length parameter is decreased ($\Delta_{k+1} = \Delta_k/2$ is typical). Note that the search is not required to find a trial point that has the lowest objective value—any trial point that improves on x_k will suffice. The name *compass search* is derived from the two-dimensional situation,

where the trial steps correspond to the cardinal points of the compass (east, west, north, and south). An example of several iterations of compass search applied to the Broyden tridiagonal function [7] is shown in Figure 1, with level curves of the objective shown in the background and the minimizer x_* indicated with a star. The current iterate x_k is indicated with a circle, while the trial points $x_k \pm \Delta_k e_i$, $i = 1, \dots, n$, are shown as squares. For $k > 0$, the configuration from the previous iteration is also included in the background.

An iteration at which Δ_k is reduced is called *unsuccessful*. For k in the subsequence of such iterations, under mild assumptions on the objective, one can show that $\|\nabla f(x_k)\| = O(\Delta_k)$. The argument for this relation is given in the section titled “Generating Set Search” as part of the

discussion of generating set search (GSS) methods for unconstrained minimization since compass search is one instance of a GSS method. In practice, the algorithm is terminated once Δ_k falls below a predetermined threshold, which is justified by the relation $\|\nabla f(x_k)\| = O(\Delta_k)$ at unsuccessful iterations.

CHARACTERISTICS OF DIRECT SEARCH METHODS

Because direct search methods neither compute nor approximate derivatives, they are often described as being “derivative-free.” This description is incomplete, however, since the same can be said of other methods that are very different in nature, such as genetic algorithms.

One distinctive feature of direct search methods in the case of unconstrained optimization is that they do not require numerical function values: the relative rank of objective values is sufficient [1,8]. This makes direct search useful in situations that involve subjective judgments of “better” and “worse,” for instance, the optimization of the perceived quality of sound by systematically varying the settings of a hearing aid [9]. Other distinctive characteristics of direct search methods can be made clearer by contrasting them with other nonlinear optimization methods.

Relation to Methods based on Taylor’s Series

Direct search methods do not rely in an intrinsic way on the construction of models of the objective in Problem (1). Many classical approaches, on the other hand, are explicitly based on the use of a Taylor’s series approximation of the objective as a model of the objective. Steepest descent uses first derivatives and the first-order Taylor polynomials to construct local linear approximations of f , while Newton’s method uses first and second derivatives and the second-order Taylor polynomial to construct local quadratic approximations of f . Quasi-Newton methods use quadratic approximations of the objective in which the Hessian is approximated via secant updates using the first derivative.

Intermediate to methods based on Taylor series and direct search are the methods of Mifflin and May, and more recent approaches that have appeared under the name of *implicit filtering* [10,11]. In Mifflin’s method for unconstrained optimization [12] and May’s method for linearly constrained optimization [13], a finite-difference interval is chosen and Newton’s method is applied with derivatives computed using this interval until no further progress is found. The finite-difference interval is then decreased and the process repeated.

Relation to Other Derivative-Free Methods

There exist other algorithms that are classified as derivative-free but which nonetheless rely on models of the objective function. These models are typically low-order polynomials constructed from values of the objective via regression or interpolation, though models based on radial basis functions [14–17] or kriging [18,19] also have been proposed.

An early example of such a method is that of Powell [20,21], which constructs a quadratic model of the objective along each search direction so that when the method is applied to a convex quadratic, it produces a sequence of conjugate search directions. Glad and Goldstein proposed an algorithm that constructs local quadratic models of the objective via linear least-squares regression [22]. Their approach was motivated by objectives whose values can only be obtained with errors that are large enough to make the use of finite-difference approximation of derivatives for a Taylor series model problematic.

More recently, other model-based derivative-free approaches have appeared under the name of *derivative-free optimization* (DFO). DFO algorithms also construct local quadratic models via interpolation but use incompletely determined quadratic models to reduce computational cost [23–27]. Much of the work on DFO methods that construct models via sampling has focused on the question of the interaction between the quality of the model (particularly with respect to second-order information) and the geometric properties (*poisedness*) of the sample points used to construct the model.

Powell [25] initially proposed using a simplex because a simplex contains the minimum number of points $(n + 1)$ required to capture first-order effects, and thus avoid the often prohibitive cost associated with requiring $O(n^2)$ sample points to build a full quadratic model at each iteration of the search. As the search proceeds, the simplex is monitored to ensure that it does not collapse into a lower-dimensional hyperplane. Powell's more recent work allows users to choose the number of sample points to be used to construct the model for each iteration, ranging from $\min\{2n + 1, n + 6\}$ to $(n + 1)(n + 2)/2$, with care taken to ensure the distribution of the points used to build the model of the objective [28–30]. There are no analytical results guaranteeing first-order convergence. Another approach is to attempt to capture full second-order effects using $O(n^2)$ sample points, starting with a minimal linear model at the first iteration and building up to a full quadratic model as more and more sample points are accumulated during the course of the search [23]. For this approach there exist first-order convergence results in the unconstrained case [31].

Response surface methodology (RSM) is another approach based on polynomial modeling. RSM was originally proposed by Box and Wilson [32] as a way to optimize a mean response function whose measured values contain random errors. Because derivatives of the response function are not available, samples of the response function are taken at designated points in a neighborhood of the current iterate and a local linear or quadratic model of the mean response is constructed via multiple linear regression. (The simplicity of working directly with the results of the sampling, rather than constructing an approximation, led Box to propose the evolutionary optimization (EVOP) direct search method [1] as an alternative for finding optima experimentally.)

In RSM, the objective is stochastic and models are constructed using regression, while in DFO, the objective is deterministic and models are typically constructed via interpolation. DFO and RSM share a common concern with experimental design; monitoring the geometry of the set of sample

points during the course of such algorithms is critical.

DIRECT SEARCH METHODS FOR UNCONSTRAINED MINIMIZATION

Direct search methods for unconstrained optimization can be grouped into two broad categories: *generating set search* and *simplex methods* (again, *not* to be confused with the simplex method for linear programming). The principles behind each of these classes of methods make it a relatively straightforward matter to devise new algorithms and explain the wealth of variations that have appeared over the years.

Generating Set Search

GSS methods [5] include the simple compass search algorithm described previously, the original pattern search method of Hooke and Jeeves [2], methods with adaptive sets of search directions such as Rosenbrock's method [33], and later methods such as parallel multidirectional search [34–36] and generalized pattern search algorithms [6].

GSS methods include, at every iteration, a set of potential search directions that form a positive spanning set for $\mathbb{R}^n$, a feature that makes it possible to prove convergence results for GSS methods that are analogous to similar results for steepest descent. A positive spanning set for $\mathbb{R}^n$ is a set of vectors $\{v_1, v_2, \dots, v_r\}$ such that

$$\begin{aligned} \{v \in \mathbb{R}^n \mid v = \alpha_1 v_1 + \dots + \alpha_r v_r, \alpha_i \\ \geq 0, i = 1, \dots, r\} = \mathbb{R}^n. \end{aligned}$$

A positive spanning set is a set of generators for $\mathbb{R}^n$ viewed as a cone, hence the name *generating set search*. In the case of compass search, the positive spanning set is $\{\pm e_i \mid i = 1, \dots, n\}$, which is a positive basis. A positive basis is a minimal positive spanning set; a positive basis for $\mathbb{R}^n$ can have as few as $n + 1$ or as many as $2n$ elements [37].

An elaboration of compass search is *coordinate search*. A single iteration of coordinate search is described in Algorithm 1. Here, the set of potential search directions

is exponential in size since it consists of the vectors representing all possible combinations of 1, 0, and -1 except the zero vector. This set clearly is a positive spanning set since it includes the set $\{\pm e_i | i = 1, \dots, n\}$, which is a positive basis. But the search does not consider all trial

points defined by the positive spanning set. Rather, it dynamically selects between n and $2n$ trial steps from those defined by the positive spanning set and, in fact, in the worst case (i.e., no improvement on the current iterate x_k), it considers exactly the same set of trial points as compass search.

Algorithm 1. An iteration of coordinate search for unconstrained minimization.

Step 0. Let $x_+ = x_k$.
Step 1. for $i = 1, \dots, n$
 if $(f(x_+ + \Delta_k e_i) < f(x_+))$
 $x_+ = x_+ + \Delta_k e_i$
 else if $(f(x_+ - \Delta_k e_i) < f(x_+))$
 $x_+ = x_+ - \Delta_k e_i$
 end
 end
Step 2. if $x_+ = x_k$ then set $\Delta_{k+1} = \Delta_k/2$ and $x_{k+1} = x_k$; otherwise, set $\Delta_{k+1} = \Delta_k$
 and $x_{k+1} = x_+$.

The pattern search method of Hooke and Jeeves adds speculative *exploratory moves* in an attempt to accelerate coordinate search. To illustrate the exploratory moves, suppose that iteration $k - 1$ was successful, so that $x_k \neq x_{k-1}$. In this case, the following heuristic is applied: since the move from x_{k-1} to x_k produced a decrease in f , immediately try another step, of the same length and along the same direction, from x_k . Such a step is called a *pattern step*, denoted by x_p . Then, regardless of whether $f(x_p) < f(x_k)$, the algorithm proceeds to one iteration of coordinate search around x_p . If none of the exploratory moves leads to an improvement in the objective, the search falls back to an iteration of coordinate search around x_k . Again, the positive spanning set is exponential in size, but the search dynamically selects between $n + 1$ and $4n + 1$ trial points. In the worst case (i.e., no improvement on the current iterate x_k) it, too, ultimately considers the same set of trial steps as compass search.

In addition to including a positive basis for $\mathbb{R}^n$ in the set of search directions, GSS methods feature a step-length control parameter Δ_k that determines the length of steps at each iteration. A crucial feature of GSS methods is that one is not allowed to *decrease* the step-length control parameter until one has determined that no improvement can be found

among the steps along *any* of the directions in a positive basis (which need not be unique) contained in the set of search directions. Such an iteration is called *unsuccessful*. At unsuccessful iterations, under mild assumptions on f , one can show that $\|\nabla f(x_k)\| = O(\Delta_k)$. This is a consequence of the fact that if x_k is not a stationary point, then at least one vector in a positive basis must be within 90° of the direction of steepest descent and thus must be a descent direction—even though the gradient of f at x_k is unknown. For GSS methods one can also show that there will be an infinite number of unsuccessful iterations and so $\Delta_k \rightarrow 0$ as $k \rightarrow \infty$. As a consequence, under standard assumptions one obtains $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$, a stationarity result similar to that for steepest descent. This means that a subsequence of the iterates will converge to a stationary point [5,6]. Moreover, as with steepest descent, one can prove local convergence results: under standard assumptions of second-order sufficiency (local convexity) at a local minimizer x_* , once the sequence of iterates is sufficiently close to x_* the entire sequence is guaranteed to converge to x_* [38]. It is even possible to make statements about the Hessian of the objective at limit points for GSS methods that are stronger than what can be said in the case of steepest descent [39].

Adaptive direction methods attempt to accelerate the search by constructing directions designed to use information about the curvature of the objective obtained during the course of the search. For example, Rosenbrock's method [33] was consciously derived to cope with the peculiar features of Rosenbrock's famous "banana function," the minimizer of which lies inside a narrow, curved valley (see Ref. 3, Figure 3.4). As the algorithm progresses, an orthogonal set of search directions is rotated in order to orient one of the directions in a promising direction of improvement. More general ideas for adaptive direction methods to acquire and use information about the structure of the objective, curvature in particular, have appeared in recent years [40–46].

Simplex Methods

The simplex-based direct search methods try steps along a heuristic direction of steepest descent. Two well-known examples of such algorithms are that of Spendley *et al.* [47] and that of Nelder and Mead [4].

Spendley, Hext, and Himsworth observed that a simplex ($n + 1$ points in $\mathbb{R}^n$ not lying in a proper subspace of $\mathbb{R}^n$) is the minimal number of points sufficient to capture first-order information about the objective (that is, to construct a linear model). In their algorithm, at any iteration there is a set of $n + 1$ points forming a simplex at which the objective values are known. Let $v_0, v_1, \dots, v_n$ be the vertices of the simplex, ordered from best to worst in the sense that $f(v_0) \leq f(v_1) \leq \dots \leq f(v_n)$. A trial point is computed by reflecting the vertex v_n with the highest objective value through the centroid $c = (v_0 + \dots + v_{n-1})/n$

of the opposite face of the simplex. Figure 2a illustrates this reflection. The new simplex that results is accepted if the reflected vertex $v_r = v_n + 2(c - v_n)$ yields a lower value of f than the vertex v_{n-1} with the second highest function value in the original simplex. This acceptance criterion is motivated by the observation that if v_r becomes v_n at the next iteration (i.e., it has the highest function value in the new simplex), the new reflection step will be the former worst vertex.

The Nelder–Mead simplex algorithm is a variation on this basic idea that introduces a simple line search of the form $v_n + \alpha(c - v_n)$, with four possible choices for α . Typical choices are $\alpha \in \{\frac{1}{2}, \frac{3}{2}, 2, 3\}$, as illustrated in Figure 2b. The line search allows the shape of the simplex to deform (for any choice of α other than 2), which is intended to allow the simplex to adapt to the local behavior of the objective.

There is long-standing numerical evidence that algorithms such as the Nelder–Mead simplex algorithm can stall away from stationary points of the objective. Moreover, McKinnon [48] constructed a family of functions in $\mathbb{R}^2$ which demonstrate analytically that the Nelder–Mead simplex algorithm can fail to converge to a stationary point of f . In these examples, repeated deformations cause the simplices to collapse (i.e., some of the interior angles tend to zero) and converge to a straight line that is orthogonal to the direction of steepest descent. The collapse of the simplices means that the search is reduced to a line that does not contain a minimizer.

Interestingly, these examples are very delicate and McKinnon found that floating-point rounding error was enough to prevent the

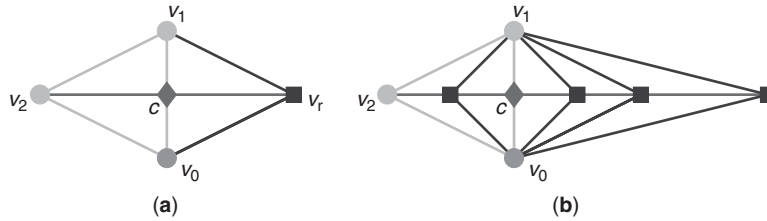


Figure 2. Examples of (a) the single possible trial step for the simplex algorithm of Spendley–Hext–Himsworth, and (b) the four basic trial steps for the Nelder–Mead simplex algorithm.

algorithm from converging to the nonstationary point in these examples. Experimental evidence [35] suggests that simplex algorithms fail in practice because the search direction becomes increasingly orthogonal to the direction of steepest descent. Numerous modifications to the Nelder–Mead simplex algorithm have been proposed [49–53], but many users find that the easiest response when the algorithm stalls is simply to restart with a new simplex.

Finally, the distinction between GSS methods and simplex methods is not sharp. There exist methods that are based on a simplex heuristic for generating promising search directions but which possess the feature of GSS methods of a set of search directions that positively span $\mathbb{R}^n$ [8,35,36].

DIRECT SEARCH METHODS FOR MORE GENERAL OPTIMIZATION

Direct search methods can be applied, with varying degrees of success, to constrained and nonsmooth (Lipschitz) optimization.

Bound Constraints

Direct search can be effective for bound constrained minimization:

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in \Omega = \{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}. \end{aligned} \quad (2)$$

The simplicity of the geometry of the feasible region makes it relatively straightforward to develop specialized direct search methods for bound constrained optimization.

For instance, a simple modification of compass search results in an algorithm that, under standard assumptions, guarantees global convergence to a KKT point of Problem (2) [54]. All that is required is that attention be restricted to trial points that are feasible, that is, trial points in the set $\{x_t \in \Omega \mid x_t = x_k \pm \Delta_k e_i, i = 1, \dots, n\}$.

GSS methods can be adapted to bound constrained minimization, as in the preceding variant of compass search, by including the directions $\{\pm e_i \mid i = 1, \dots, n\}$ as search directions when near the boundary [54,55]. If x

is feasible, let $N(x, \varepsilon)$ be the cone determined by the outward-pointing normals to the constraints that are active within distance ε of x . The polar cone $T(x, \varepsilon) = N^\circ(x, \varepsilon)$ then approximates the feasible region in a neighborhood of x . The set of search directions $\{\pm e_i \mid i = 1, \dots, n\}$ always contains a set of generators for the cone $T(x, \varepsilon)$, once again illustrating the motivation for the name *generating set search*.

GSS methods are not allowed to decrease the step-length parameter Δ_k unless all feasible steps have been tried along all of the directions $\{\pm e_i \mid i = 1, \dots, n\}$ and no improvement in the objective has been found. If P_Ω represents projection onto the feasible region Ω , then under standard assumptions one can show that at unsuccessful steps (those where Δ_k is reduced), $\|P_\Omega(x_k - \nabla f(x_k)) - x_k\| = O(\Delta_k)$. One also has $\Delta_k \rightarrow 0$ as $k \rightarrow \infty$, so $\lim_{k \rightarrow \infty} \|P_\Omega(x_k - \nabla f(x_k)) - x_k\| = 0$. This means that a subsequence of the iterates will converge to a KKT point.

The search directions $\{\pm e_i \mid i = 1, \dots, n\}$ are only needed near the boundary. In the interior of the feasible region one has greater flexibility in the choice of search directions since feasibility with respect to the bound constraints is not an immediate issue. For instance, Keefer’s Simpat algorithm [56] uses the Nelder–Mead simplex algorithm in the interior of the feasible region and the Hooke and Jeeves pattern search algorithm near the boundary.

Simplex-based direct search methods are more problematic for bound constrained optimization since it is difficult to ensure that they search along directions that take into account the proximity of the boundary. In addition, mechanisms to ensure only feasible steps can cause the simplices to collapse into a lower-dimensional subspace near the boundary.

Linear Constraints

GSS methods have also been developed for general linearly constrained minimization:

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in \Psi = \{x \in \mathbb{R}^n \mid Ax \leq b\}. \end{aligned} \quad (3)$$

As with bound constraints, the key is to include search directions that capture the geometry of the feasible region near the current iterate. If x is feasible, let $N(x, \varepsilon)$ be the cone determined by the outward-pointing normals to the constraints that are active within distance ε of x . The polar cone $T(x, \varepsilon) = N^\circ(x, \varepsilon)$ approximates the feasible region in a neighborhood of x . In a GSS method for linearly constrained minimization the set of search directions must always contain a set of generators for the cone $T(x, \varepsilon)$. If $T(x, \varepsilon)$ has a degenerate vertex at the origin (in the sense of linear programming) the calculation of a full set of generators for $T(x, \varepsilon)$ can be an expensive polyhedral geometry calculation [57]. On the other hand, if $T(x, \varepsilon)$ does not have a degenerate vertex at the origin, a straightforward linear algebra calculation yields the generators [57, 58].

As with GSS methods for bound constrained minimization, the algorithm is not allowed to decrease the step-length parameter unless all feasible steps have been tried along at least a minimal set of generators for $T(x, \varepsilon)$ and no improvement in the objective has been found. Again, under mild assumptions, one can show that a suitable measure of stationarity that is zero at KKT points is $O(\Delta_k)$ at unsuccessful iterations. One also has $\Delta_k \rightarrow 0$ as $k \rightarrow \infty$ and so once again, this means that a subsequence of the iterates will converge to a KKT point [58–61].

Overall, the convergence results for GSS methods applied to linearly constrained minimization are comparable to those for gradient projection methods. In addition to convergence to KKT points one can also prove that GSS methods have active set estimation properties comparable to those of gradient projection [60].

Nonsmooth (Lipschitz) Optimization

Because direct search methods do not explicitly rely on derivatives, they are useful for noisy optimization problems in which the objective and constraints may be viewed as smooth functions with small amplitude and high-frequency noise superimposed. Given their derivative-free nature it is natural to consider direct search methods for optimization of nonsmooth but Lipschitz continuous

functions. In many of these problems, the loci of nondifferentiability are more structured and, if one can take advantage of this structure, then it is possible to treat such problems effectively using direct search [62–65]. Unfortunately, in the general case, structured loci of nondifferentiability tend to cause trouble for direct search methods [66]. The descent-based searches conducted by direct search methods frequently lead the search to regions of nondifferentiability. In these regions the set of descent directions is no longer a half-space but a proper convex cone and there may fail to be a search direction inside this cone. As a consequence, direct search methods can stall away from a stationary point and be difficult to restart successfully, even for convex functions (see Ref. 5, Section 6).

General Nonlinear Constraints

Finally, consider the general nonlinear programming problem:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0 \\ &\quad g(x) \leq 0. \end{aligned} \tag{4}$$

Direct search methods for general nonlinear programming can be devised by applying direct search methods for unconstrained or linearly constrained optimization to approaches based on penalization. Examples include quadratic penalization, logarithmic and reciprocal barrier methods, and augmented Lagrangian methods [67–71].

Constraints may also be treated by applying a direct search method to an exact, nonsmooth penalty function. This approach has the potential difficulties associated with applying direct search to nonsmooth optimization, and the algorithm may fail to converge to a stationary point, as discussed in the section titled “Nonsmooth (Lipschitz) Optimization”. One way to ameliorate this difficulty is to conduct multiple simultaneous searches with a mixture of smooth and nonsmooth penalty functions [72, 73]. By sharing information between the searches, one can avoid stalling due to nonsmoothness

by allowing the search to jump past regions of nondifferentiability.

There also exist direct search methods for the case where the derivatives of the nonlinear constraints are available, or where estimation of these derivatives is deemed appropriate. These methods make explicit use of derivatives of the constraints to compute feasible directions at the boundary of the feasible region. The feasible region near the current iterate is approximated by the linearization of the nearby constraints, as in traditional feasible directions methods that rely on derivatives. The important search directions are the generators of the cones polar to the cones generated by the nearby outward-pointing normals, as in the linearly constrained case. Examples of such methods include that of Klingman and Himmelblau [74] and Glass and Cooper [75]. More recently, Lucidi *et al.* [61] presented a feasible directions method that uses linearization to approximate constraints but applies direct search to the objective. A restoration step is used to ensure feasibility, and convergence to KKT points can be shown under mild conditions.

Finally, there are numerous direct search approaches to constrained optimization that apply heuristics to deal with constraints. For instance, in Box's Complex method [76], one of the earliest direct search methods designed to address problems with constraints, the objective is sampled at a broader set of points than in simplex-based methods in an attempt to avoid premature termination. Another approach is to interpret constrained optimization as a bilevel problem of improving feasibility and then optimality. The flexible tolerance method [77,78] alternately attempts to reduce the objective and constraint violation, depending on the extent to which the iterates are infeasible. A similar view motivated the adaptation of the filter method [79] to direct search [80].

STRENGTHS OF DIRECT SEARCH METHODS

Direct search methods for unconstrained and bound minimization are, in general, easy to describe and easy to implement. Methods for

linearly constrained optimization are more complex, but the ideas are familiar ones from linear programming or feasible directions methods.

GSS methods also enjoy theoretical guarantees of convergence to (constrained) stationary points that are similar to those for methods based on steepest descent. This perhaps surprising discovery has renewed interest in direct search methods in the mathematical optimization community [81].

Direct search methods are useful in problems where derivatives of the objectives or constraints are not available or not reliable. This situation can arise, for instance, when the evaluation of the objective or constraints involves computer calculations that exhibit a lack of smoothness due to features such as look-up tables, if-then-else logic, or the use of adaptive or iterative techniques. Problems in which evaluation of the objective involves experimentation (e.g., wind tunnel evaluation of drag [82]) are another place where derivatives are problematic.

Direct search methods can be applied in situations where there is no explicit numerical objective being optimized since they require only the ability to decide whether one set of decision variables is better than another. Such situations arise when the optimization involves subjective comparisons of "better" and "worse," as in the hearing aid tuning application mentioned previously [9]. Multiobjective optimization is another application that can be posed in this way [83].

There are many opportunities for scalable computational parallelism with direct search methods. Two successful parallel implementations of GSS are APPSPACK [84–86] for linearly constrained minimization and its successor HOPSPACK [72,73] for general nonlinearly constrained minimization.

WEAKNESSES OF DIRECT SEARCH METHODS

Direct search methods typically make rapid initial progress toward the solution. In this way they behave like steepest descent. However, convergence to a minimizer will only become apparent as the algorithm reduces the length of the trial steps. This means that

while a direct search algorithm may quickly approach a minimizer, it may be slow to detect this fact. This is the price of not explicitly using gradient information. Similarly, direct search methods may end up taking very small steps in regions where the objectives or constraints are highly nonlinear if no use is made of curvature (i.e., second derivative) information. It should be noted, however, that methods such as steepest descent or Newton's method can suffer from similar problems (if finite differences are used in Newton's method to estimate derivatives).

Direct search methods are also limited with respect to the size of the problems they can solve. They can be successfully applied to problems with a few hundreds of decision variables, but they cannot handle problems with many hundreds or thousands. This curse of dimensionality is due to the difficulty in obtaining sufficiently dense sets of search directions as the dimension increases.

In addition, many of the classical direct search methods, such as the simplex methods of Spendley *et al.* [47] and Nelder–Mead [4], do not enjoy theoretical guarantees of convergence when $n > 1$. For instance, the Nelder–Mead simplex algorithm can be shown to fail in theory [48] and is known to stall sometimes in practice.

REFERENCES

1. Box GEP. Evolutionary operation: a method for increasing industrial productivity. *Appl Stat* 1957;6(2):81–101.
2. Hooke R, Jeeves TA. Direct search solution of numerical and statistical problems. *J Assoc Comput Mach* 1961;8(2):212–229.
3. Lewis RM, Torczon V, Trosset MW. Direct search methods: then and now. *J Comput Appl Math* 2000;124(1–2):191–207.
4. Nelder JA, Mead R. A simplex method for function minimization. *Comput J* 1965;7(4):308–313.
5. Kolda TG, Lewis RM, Torczon V. Optimization by direct search: new perspectives on some classical and modern methods. *SIAM Rev* 2003;45(3):385–482.
6. Torczon V. On the convergence of pattern search algorithms. *SIAM J Optim* 1997;7(1):1–25.
7. Broyden CG. A class of methods for solving nonlinear simultaneous equations. *Math Comput* 1965;19(92):577–593.
8. Lewis RM and Torczon V. Rank ordering and positive bases in pattern search algorithms. Technical Report 96–71. Hampton (VA): Institute for Computer Applications in Science and Engineering, NASA Langley Research Center; 1996.
9. Franck BAM, Dreschler WA, Lyzenga J. Methodological aspects of an adaptive multidirectional search to optimize speech perception using three hearing-aid algorithms. *J Acoust Soc Am* 2004;116(6):3620–3628.
10. Choi TD, Kelley CT. Superlinear convergence and implicit filtering. *SIAM J Optim* 2000;10(4):1149–1162.
11. Gilmore P, Kelley CT. An implicit filtering algorithm for optimization of functions with many local minima. *SIAM J Optim* 1995;5(2):269–285.
12. Mifflin R. A superlinearly convergent algorithm for minimization without evaluating derivatives. *Math Program* 1975;9(1):100–117.
13. May JH. Solving nonlinear programs without using analytic derivatives. *Oper Res* 1979;27(3):457–484.
14. Fang H, Horstemeyer MF. Global response approximation with radial basis functions. *Eng Optim* 2006;38(4):407–424.
15. Jakobsson S, Patriksson M, Rudholm J, *et al.* A method for simulation based optimization using radial basis functions. *Optim Eng* 2009. DOI: 10.1007/s11081-009-9087-1.
16. Ouevray R, Bierlaire M. BOOSTERS: a derivative-free algorithm based on radial basis functions. Technical Report GR-BIE-ARTICLE-2006-001. Switzerland: Ecole Polytechnique Fédérale de Lausanne (EPFL), CH–1015 Lausanne; 2005.
17. Wild SM, Regis RG, Shoemaker CA. ORBIT: optimization by radial basis function interpolation in trust-regions. *SIAM J Sci Comput* 2008;30(6):3197–3219.
18. Booker AJ, Dennis JE Jr, Frank PD, *et al.* A rigorous framework for optimization of expensive functions by surrogates. *Struct Multidiscipl Optim* 1999;17(1):1–13.
19. Torczon V, Trosset MW. Using approximations to accelerate engineering design optimization. Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization; 1998

- Sep 2–4; St. Louis (MO). pp. 738–748. AIAA–1998–4800.
20. Powell MJD. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput J* 1964;7(2):155–162.
 21. Zangwill WI. Minimizing a function without calculating derivatives. *Comput J* 1967;10(3):293–296.
 22. Glad T, Goldstein A. Optimization of functions whose values are subject to small errors. *BIT* 1977;17(2):160–169.
 23. Conn AR, Scheinberg K, Vicente LN. Introduction to derivative-free optimization, MPS/SIAM series on optimization. Philadelphia: SIAM; 2009.
 24. Conn AR, Toint PL. An algorithm using quadratic interpolation for unconstrained derivative free optimization. In: Di Pillo G, Giannessi F, editors. *Nonlinear optimization and applications*. New York: Kluwer Academic/Plenum Publishers; 1996. pp. 27–47. *Proceedings of the International School of Mathematics “G. Stampacchia” 21st Workshop*; 1995 June 13–21; Erice, Italy.
 25. Powell MJD. A direct search optimization method that models the objective and constraint functions by linear interpolation. In: Gomez S, Hennart J-P, editors. *Volume 275, Advances in optimization and numerical analysis. Proceedings of the 6th Workshop on Optimization and Numerical Analysis*, Oaxaca, Mexico, Mathematics and its applications. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1994. pp. 51–67.
 26. Powell MJD. On trust region methods for unconstrained minimization without derivatives. *Math Program* 2003;97(3):605–623.
 27. Powell MJD. A view of algorithms for optimization without derivatives. Technical Report DAMTP 2007/NA03. Cambridge, England: Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Cambridge University; 2007.
 28. Powell MJD. Least Frobenius norm updating of quadratic models that satisfy interpolation conditions. *Math Program* 2004;100(1):183–215.
 29. Powell MJD. The NEWUOA software for unconstrained optimization without derivatives. In: DiPillo G, Roma M, editors. *Volume 83, Large-scale nonlinear optimization, Non-convex optimization and its applications*. New York: Springer; 2006. pp. 255–297.
 30. Powell MJD. Developments of NEWUOA for minimization without derivatives. *IMA J Numer Anal* 2008;28(4):649–664.
 31. Conn AR, Scheinberg K, Toint PL. On the convergence of derivative-free methods for unconstrained optimization. In: Buhmann MD, Iserles A, editors. *Approximation theory and optimization: tributes to M. J. D. Powell*. Cambridge: Cambridge University Press; 1997. pp. 83–108.
 32. George EPB, Wilson KB. On the experimental attainment of optimum conditions. *J Roy Stat Soc Ser B* 1951;XIII(1):1–45.
 33. Rosenbrock HH. An automatic method for finding the greatest or least value of a function. *Comput J* 1960;3(3):175–184.
 34. Dennis JE Jr, Torczon V. Direct search methods on parallel machines. *SIAM J Optim* 1991;1(4):448–474.
 35. Torczon V. Multi-directional search: a direct search algorithm for parallel machines [PhD thesis]. Houston (TX): Department of Mathematical Sciences, Rice University; 1989. Available as Technical Report 90-07. Houston (TX): Department of Computational and Applied Mathematics, Rice University.
 36. Torczon V. On the convergence of the multi-directional search algorithm. *SIAM J Optim* 1991;1(1):123–145.
 37. Davis C. Theory of positive linear dependence. *Am J Math* 1954;76:733–746.
 38. Dolan ED, Lewis RM, Torczon VJ. On the local convergence properties of pattern search. *SIAM J Optim* 2003;14(2):567–583.
 39. Abramson MA. Second-order behavior of pattern search. *SIAM J Optim* 2005;16(2):515–530.
 40. Coope ID, Price CJ. A direct search conjugate directions algorithm for unconstrained minimization. *Aust N Z Ind Appl Math (ANZIAM) J* 2000;42 electronic Part C: C478–498.
 41. Coope ID, Price CJ. On the convergence of grid-based methods for unconstrained optimization. *SIAM J Optim* 2001;11(4):859–869.
 42. Colson B, Toint PL. Exploiting band structure in unconstrained optimization without derivatives. *Optim Eng* 2001;2(4):299–412.
 43. Colson B, Toint PL. Optimizing partially separable functions without derivatives. *Optim Methods Softw* 2005;20(4–5):493–508.
 44. Frimannslund L, Steihaug T. A generating set search method exploiting curvature and

- sparsity. In: Yuan Di, editor. Volume 14, Proceedings of the 9th Meeting of the Nordic Section of the Mathematical Programming Society (Nordic MPS 2004); 2004 Oct 22–23. Norrköping, Sweden: Linköpings Universitet; 2004. pp. 57–71.
45. Frimannslund L, Steihaug T. A generating set search method using curvature information. *Comput Optim Appl* 2007;38(1):105–121.
 46. Price CJ, Toint PL. Exploiting problem structure in pattern search methods for unconstrained optimization. *Optim Methods Softw* 2006;21(3):479–491.
 47. Spendley W, Hext GR, Himsworth FR. Sequential application of simplex designs in optimometric and evolutionary operation. *Technometrics* 1962;4(4):441–461.
 48. McKinnon KIM. Convergence of the Nelder–Mead simplex method to a non-stationary point. *SIAM J Optim* 1998;9(1):148–158.
 49. Kelley CT. Detection and remediation of stagnation in the Nelder–Mead algorithm using a sufficient decrease condition. *SIAM J Optim* 1999;10(1):43–55.
 50. Nazareth L, Tseng P. Gilding the lily: a variant of the Nelder–Mead algorithm based on golden-section search. *Comput Optim Appl* 2002;22(1):133–144.
 51. Price CJ, Coope ID, Byatt D. A convergent variant of the Nelder Mead algorithm. *J Optim Theory Appl* 2002;113(1):5–19.
 52. Shekarforoush H, Berthod M, Zerubia J. Direct search generalized simplex algorithm for optimizing non-linear functions. Technical Report 2535. Sophia-Antipolis, France: INRIA; 1995.
 53. Tseng P. Fortified-descent simplicial search method: a general approach. *SIAM J Optim* 1999;10(1):269–288.
 54. Lewis RM, Torczon V. Pattern search algorithms for bound constrained minimization. *SIAM J Optim* 1999;9(4):1082–1099.
 55. Lucidi S, Sciandrone M. A derivative-free algorithm for bound constrained optimization. *Comput Optim Appl* 2002;21(2):119–142.
 56. Keefer DL. Simpat: self-bounding direct search method for optimization. *Ind Eng Chem Process Des Dev* 1973;12(1):92–99.
 57. Lewis RM, Shepherd A, Torczon V. Implementing generating set search methods for linearly constrained minimization. *SIAM J Sci Comput* 2007;29(6):2507–2530.
 58. Lewis RM, Torczon V. Pattern search methods for linearly constrained minimization. *SIAM J Optim* 2000;10(3):917–941.
 59. Kolda TG, Lewis RM, Torczon V. Stationarity results for generating set search for linearly constrained optimization. *SIAM J Optim* 2006;17(4):943–968.
 60. Lewis RM, Torczon V. Active set identification for linearly constrained minimization without explicit derivatives. *SIAM J Optim* 2009;20(3):1378–1405.
 61. Lucidi S, Sciandrone M, Tseng P. Objective-derivative-free methods for constrained optimization. *Math Program* 2002;92(1):37–59.
 62. Bogani C, Gasparo MG, Papini A. A pattern search method for discrete L_1 -approximation. *J Optim Theory Appl* 2007;134(1):47–59.
 63. Bogani C, Gasparo MG, Papini A. Generalized pattern search methods for a class of non-smooth problems with structure. *J Comput Appl Math* 2009;229:283–293.
 64. Bogani C, Gasparo MG, Papini A. Generating set search methods for piecewise smooth problems. *SIAM J Optim* 2009;20:321–335.
 65. Liuzzi G, Lucidi S, Sciandrone M. A derivative-free algorithm for linearly constrained finite minimax problems. *SIAM J Optim* 2006;16(4):1054–1075.
 66. Gill PE, Murray W, Wright MH. Practical optimization. London: Academic Press; 1981.
 67. Lewis RM, Torczon V. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM J Optim* 2002;12(4):1075–1089.
 68. Kolda TG, Lewis RM, Torczon V. A generating set direct search augmented Lagrangian algorithm for optimization with a combination of general and linear constraints. Technical Report SAND2006-5315. Albuquerque (NM), Livermore (CA): Sandia National Laboratories; 2006.
 69. Lewis RM, Torczon V. A direct search approach to nonlinear programming problems using an augmented Lagrangian method with explicit treatment of the linear constraints. Technical Report WM-CS-2010-01. Williamsburg (VA); Department of Computer Science, College of William & Mary; 2010.
 70. Liuzzi G, Lucidi S. A derivative-free algorithm for inequality constrained nonlinear programming via smoothing of an ℓ_∞ penalty function. *SIAM J Optim* 2009;20(1):1–29.

71. Liuzzi G, Lucidi S, Sciandrone M. Sequential penalty derivative-free methods for nonlinear constrained optimization. *SIAM J Optim* 2010;20(5):2614–2635.
72. Griffin JD, Kolda TG. A parallel, asynchronous method for derivative-free nonlinear programs. In: Iglesias A, Takayama N, editors. Volume 4151/2006, Mathematical software –ICMS 2006, Lecture notes in computer science. Berlin: Springer; 2006. pp. 260–262.
73. Plantenga TD. HOPSPACK 2.0 user manual. Technical Report SAND2009-6265. Albuquerque (NM), Livermore (CA): Sandia National Laboratories; 2009.
74. Klingman WR, Himmelblau DM. Nonlinear programming with the aid of a multiple-gradient summation technique. *J Assoc Comput Mach* 1964;11(4):400–415.
75. Glass H, Cooper L. Sequential search: a method for solving constrained optimization problems. *J Assoc Comput Mach* 1965;12(1):71–82.
76. Box MJ. A new method of constrained optimization and a comparison with other methods. *Comput J* 1965;8(1):42–52.
77. Himmelblau DM. Applied nonlinear programming. New York: McGraw–Hill; 1972.
78. Paviani D, Himmelblau DM. Constrained nonlinear optimization by heuristic programming. *Oper Res* 1969;17:872–882.
79. Fletcher R, Leyffer S. Nonlinear programming without a penalty function. *Math Program Ser A* 2002;91:239–269.
80. Charles A, Dennis JE Jr. A pattern search filter method for nonlinear programming without derivatives. *SIAM J Optim* 2004;14(4):980–1010.
81. Wright MH. Direct search methods: once scorned, now respectable. In: Griffiths DF, Watson GA, editors. Volume 344, Numerical Analysis 1995 (Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis), Pitman research notes in mathematics. Boca Raton (FL): CRC Press; 1996. pp. 191–208.
82. Jacobsen M. Real time drag minimization using redundant control surfaces. *Aerosp Sci Technol* 2006;10(7):574–580.
83. Woods DJ. An interactive approach for solving multi-objective optimization problems [PhD thesis]. Houston (TX): Department of Mathematical Sciences, Rice University; 1985. Also available as TR85-05, Houston (TX): Department of Computational and Applied Mathematics, Rice University.
84. Gray GA, Kolda TG. Algorithm 856: APPSPACK 4.0: asynchronous parallel pattern search for derivative-free optimization. *ACM Trans Math Softw* 2006;32(3):485–507.
85. Griffin JD, Kolda TG, Lewis RM. Asynchronous parallel generating set search for linearly constrained optimization. *SIAM J Sci Comput* 2008;30(4):1892–1924.
86. Hough PD, Kolda TG, Torczon VJ. Asynchronous parallel pattern search for nonlinear optimization. *SIAM J Sci Comput* 2001;23(1):134–156.
87. Lagarias JC, Reeds JA, Wright MH, *et al.* Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM J Optim* 1998;9(1):112–147.

DISCRETE OPTIMIZATION WITH NOISY OBJECTIVE FUNCTION MEASUREMENTS

Stacy D. Hill
The Johns Hopkins University,
Applied Physics Laboratory,
Laurel, Maryland

Discrete optimization with “noisy” objective function measurements—also referred to as *discrete (parameter) stochastic optimization*—is concerned with maximizing (or minimizing) real-valued functions defined on discrete sets, where the values of the objective function are unknown but can be estimated to within some measurement noise.

Some applications that give rise to the problem of optimizing noisy objective functions are: message transmission in communications networks [1–4]; locating a service or resource facility, such as a manufacturing warehouse, supply distribution center, school, or hospital [5,6]; scheduling machine usage in a production plant [7,8]; allocating people to evacuation routes [9]; and, more generally, allocating a finite number of units of a fixed resource to users or activities so as to minimize a given performance measure [10–12].

The salient feature of such optimization problems—which is a major source of difficulty—is the discreteness of the feasible region on which the objective function is defined. In contrast to continuous parameter optimization, objective functions defined on discrete sets lack derivatives that provide information about local optimality and guide the search for the optimum. The optimization of discrete functions relies on comparing differences of the objective function values, which must be estimated when those values contain noise.

This article will describe the general problem of stochastic discrete optimization and discuss characteristics and properties of several main approaches to solving it.

PROBLEM FORMULATION

Preliminaries

Although there is no generally accepted definition of what is meant by a discrete feasible region in optimization, usages of the term imply that the set is at most countable. Cardinality alone, however, does not suffice to describe discrete sets. (For example, rational numbers are at most countable but are not discrete.) In the case of optimization over subsets of real d -dimensional space, the common interpretation of a discrete set is that its points are separated by some fixed positive distance. An essential property of such sets is that functions defined on them lack derivatives. One way to guarantee that a derivative is undefinable is to require that the set have no accumulation points. This observation motivates the following topological definition of discreteness (see, e.g., [13]), which suffices for the discussion in this article and is sufficiently general to cover most problem settings.

A set in a topological space is discrete if it has no accumulation points. Equivalently, a discrete set is one that contains only isolated points. For example, the integers $\mathbb{Z}$ form a discrete set in the reals $\mathbb{R}$, whereas the set of rational numbers do not.

The introduction of a topology, aside from its usefulness in defining discreteness, simplifies the discussion of convergence concepts. In particular, a sequence of points $\{\theta_n\}$ in a discrete set Θ converges to a point θ in Θ if $\theta_n = \theta$ for n sufficiently large. (Take any neighborhood of θ containing no point of Θ other than θ . Such neighborhoods exist because discrete sets contain only isolated points.) Similarly, θ_n converges to a subset E of Θ , consisting of finitely many points, if $\theta_n \in E$ for sufficiently large n .

The Optimization Problem

The discrete stochastic optimization problem has the following general formulation. Let L

be a real-valued *objective function* defined on a discrete, at most countable feasible region Θ —the problem domain. The value $L(\theta)$ is unknown, but it is assumed that, for each $\theta \in \Theta$, there is available a random variable $Y = Y(\theta)$ such that

$$Y = L(\theta) + \varepsilon(\theta), \quad \theta \in \Theta, \quad (1)$$

where $\varepsilon(\theta)$ —measurement or estimation noise—is a random variable with mean zero. Thus, $Y(\theta)$ satisfies $L(\theta) = \mathbb{E}[Y(\theta)]$, for each θ . The problem is to find $\theta^* \in \Theta$ such that

$$L(\theta^*) = \min_{\theta \in \Theta} L(\theta) \quad (2)$$

using only measurements of Y . To avoid trivialities, it is assumed throughout this article that Θ is nonempty and, furthermore, that L is nonconstant on Θ .

The objective function will sometimes be interpreted as measuring some type of loss, such as cost. The problem, then, is to find points in the feasible region that minimizes loss. Note that the problem of minimizing $L(\theta)$ is equivalent to the problem of maximizing $-L(\theta)$. In addition, any method for finding a maximum is, therefore, easily modified to yield a method for finding a minimum. The choice of problem—minimization or maximization—is a matter of convenience and, usually, is dictated by the particular application. The choice here is to view optimization as minimization (of loss) unless otherwise stated.

Let Θ^* denote the set consisting of all points in Θ that solve Equation (2), that is, the set consisting of the points in Θ that *optimize* L . When Θ is finite, Θ^* is nonempty (because it is assumed that Θ is nonempty). When Θ is infinite, it will be assumed that Θ^* is nonempty. In either instance, the minimum value of $L(\theta)$ exists and will be denoted by L^* .

It is tempting to view discrete optimization problems as optimization over $\mathbb{Z}$, as any function defined on a discrete set can be transformed into a function on a subset of integers and minimized there. (For example, if h is a one-to-one correspondence between Θ and a subset J of $\mathbb{Z}$, define the loss function on J to be $\tilde{L}(z) = L(h(z))$ for each z in

J .) However, there may be little or no benefit in doing so. In fact, the transformed problem might eliminate some inherent problem structure (e.g., property of the objective function) that aids in the problem solution. A simple example of this is the discrete resource allocation problem [11]. The problem is to distribute a finite resource to a finite number of users so as to minimize some allocation cost that is assumed to be discretely convex (see, e.g., [14]). If d is the number of users, then the feasible region is a subset of $\mathbb{Z}_+^d$, the set of points in $\mathbb{R}^d$ with nonnegative integer coordinates, where the coordinates represents the amount of resource allocated to each user. There are infinitely mappings that define a one-to-one correspondence between $\mathbb{Z}_+^d$ and $\mathbb{Z}$, but there is no guarantee that any one of them will preserve the convexity of the cost function in the original problem. Thus, the formulation and representation of a problem requires some care. (For discussion of issues associated with problem formulation, see, e.g., [15,16].)

OPTIMIZATION APPROACHES

It is worth noting that deterministic optimization algorithms do not generally apply to noisy objective functions as they rely on the values of $L(\theta)$ in the search for an optimum. Such algorithms assume that Y contains no measurement noise and, therefore, assume—erroneously—that a small value of $Y(\theta)$ necessarily corresponds to a small value of $L(\theta)$. Thus, deterministic methods seek points in the feasible region that minimize $Y(\theta)$ and, therefore, do not apply without some modification.

There is a vast literature on optimizing discrete objective functions with noise. This section highlights aspects of several classes of solution approaches that encompass a broad range of techniques. These include statistical approaches, random search methods, stochastic approximation (SA), and sample path methods. Unless otherwise stated, it will be assumed that Θ is finite. The case in which Θ is infinite will be discussed later.

Statistical Approaches

Statistical optimization methods make decisions about which parameter values are the best typically in terms of hypothesis testing procedures or confidence intervals. The methods include ranking and selection (R&S) and multiple comparisons (e.g., [17–22]) and adaptive allocation or sequential design of experiments (see, e.g., [23–25]). Spall [Chap. 12, Ref. 26] discusses the application of R&S and multiple comparisons to discrete stochastic optimization, with worked examples.

Statistical procedures treat the objective function values as the unknown parameters corresponding to k statistical populations, $1 < k < \infty$. The i th population, $i = 1, 2, \dots, k$, has probability density function $f(y; \theta(i))$ (with respect to some dominating probability measure ν , say) and mean $L_i = L(\theta(i))$. The form of $f(\cdot; \theta(i))$ is assumed known, where $\theta(i)$ is a fixed but unknown parameter. Thus, Θ is the set consisting of the points $\theta(1), \theta(2), \theta(3), \dots, \theta(k)$. The populations are ranked according to the L_i values, where the smaller values determine the better ranks, with L^* being the smallest of the L_i 's. The problem is to construct procedures for identifying the best-ranked population by taking observations of the random variables $Y(\theta(i))$.

Ranking and Selection, Multiple Comparison. These methods make decisions regarding the best in terms of independent observations of Y , where n_i independent identically distributed (i.i.d.) observations are taken of $Y(\theta(i))$, $i = 1, 2, 3, \dots, k$. Inference is based on the sample means of the $Y(\theta(i))$'s.

R&S methods are decision rules for selecting a single population or a subset of populations (of fixed or random size) such that the selection includes the best-ranked population with probability at least equal to P^* . The probability of obtaining a correct selection, P^* , is specified before selection. A variant of such procedures is the selection of a subset that includes the best population, with probability of at least P^* , when the parameters satisfy some other condition (such as the condition that the difference between the best

and the next best value be greater than or equal to some specified value).

Multiple comparison procedures rely on simultaneous confidence intervals to select the best population. An example of this type of procedure is the construction of simultaneous confidence intervals about the differences between the values of the objective function, $L_i - L_j$, where $i \neq j$, such that each interval is of the same confidence level. Confidence intervals about the differences between objective function values facilitate inference about the magnitudes and directions of the differences and, consequently, enable inferences about the best value of $L(\theta)$. Similar to R&S, the confidence intervals are constructed by taking i.i.d. samples of $Y(\theta(i))$, for each i .

Some useful measures of performance of these statistical procedures are the expected size and the expected sum of ranks of the selected subset (for R&S procedures) and the expected width of the confidence intervals (for the multiple comparison procedures). For example, consider the problem of selecting a set that contains the best-ranked population with probability P^* . The size of the selected set is a random variable S such that $1 \leq S \leq k$. Suppose that the decision rule selects population i for inclusion in the set by comparing estimates of the $\hat{L}(\theta)$'s defined to be the sample means of the observed values of $Y(\theta)$. Suppose that i is included if

$$\hat{L}(\theta(i)) \leq \min\{\hat{L}(\theta), \theta \in \Theta\} + d \quad (3)$$

where d is chosen so that the set contains L^* with probability P^* . The expected value of S is given by

$$E(S) = \sum_{i=1}^k P(\text{Selecting the population of rank } i). \quad (4)$$

It is also of interest to consider under what conditions on the L_i values, for example, is the maximum value of $E(S)$ achieved.

Sequential Design. The method of optimization by sequential design (initiated in Ref. [23]) sequentially samples $y_1, y_2, y_3, \dots$ from k populations, where y_i is drawn from population m_i . The problem is to select the populations from which to sample the

observations so as to optimize the expected value of the sum $S_n = y_1 + y_2 + y_3 + \dots + y_n$ as $n \rightarrow \infty$. (In this section, optimization will be taken to be maximization, in order to keep with the literature on this problem. Thus, L^* , here, denotes the maximum value of L .)

The choice of population from which to sample the n th observation y_n is a “decision” φ_n such that the event $\{\varphi_n = i\}$, $1 \leq i \leq k$, “sample the n th observation y_n from population i ,” depends only on $\varphi_1, y_1, \dots, \varphi_{n-1}, y_{n-1}$, $n = 1, 2, 3, \dots$. The sequence of decisions defines a sequential experimental design $\varphi = (\varphi_1, \varphi_2, \varphi_3, \dots)$. If

$$\lim_{n \rightarrow \infty} \frac{E(S_n)}{n} = L^*, \quad (5)$$

then φ is said to be consistent.

Robbins [23] developed consistent sequential experimental designs for the case $k = 2$. Robbins and Lai [24] consider the general case, $k \geq 2$.

The problem of maximizing $E(S_n)$ is equivalent to minimizing the *regret* ([24])

$$R_n = nL^* - E(S_n) = \sum_{i: L(\theta_i) < L^*} (L^* - L_i) E T_n(i), \quad (6)$$

where $T_n(i)$ is the number of times, within the first n observations, that φ samples from population i .

Lai and Robbins [24] call a decision rule *asymptotically efficient* if its regret satisfies $R_n = o(n^\beta)$, for every $\beta > 0$. They show how to construct asymptotically efficient rules that asymptotically satisfy

$$E T_n(i) \sim (\log n) / I(\theta(i); \theta^*) \quad (7)$$

for each $\theta(i)$ such that $L_i < L^*$, where $I(\theta, \lambda)$ denotes the Kullback-Leibler number

$$I(\theta, \lambda) = \int [\log (f(y; \theta) / f(y; \lambda))] f(y; \theta) \, d\nu(y).$$

The interpretation of (7) is that an asymptotically efficient rule takes about $(\log n) / I(\theta(i); \theta^*)$ observations from an inferior population. A consequence of (6) and (7)

is that such rules satisfy

$$R_n \sim \left\{ \sum_{i: L_i < L^*} (L^* - L_i) / I(\theta(i); \theta^*) \right\} \log n, \quad (8)$$

which provides a measure of expected long run performance.

Lai [25] discusses generalizations of the problem studied in Ref. [24] and discusses applications to engineering and economics.

Random Search

Optimization by random or stochastic search is similar in spirit to the method of sequential design. Random search algorithms produce a sequence $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_n$ of estimated solutions to Equation 2 such that the next estimate $\hat{\theta}_{n+1}$ is obtained by searching for candidate points “near” $\hat{\theta}_n$, according to some specific random procedure. The goal is to obtain estimates $\hat{\theta}_n$ that converge, in some sense, to the optimizing set Θ^* .

In general terms, random search iteratively selects a candidate value for updating the current solution estimate. Candidate values are chosen from the set of “neighbors” of the current value. The candidate value is compared to the current solution estimate in some particular manner and, depending on the outcome of the comparison, the candidate value either replaces the current solution estimate (and becomes the new value of the estimate) or is discarded, in which instance, the current value of the estimate is retained. The procedure is repeated to obtain the next estimate and, thereby, generate a sequence of solution estimates.

Search methods differ in several ways: the set of points that are considered neighbors of a given point, the random mechanism for selecting candidate values from among the neighbors of a given point, the procedure for comparing the candidate and current values of the estimates, and the criterion for accepting the candidate as the new solution estimate. The convergence properties of the solution estimates vary depending on the method.

The general steps of random search algorithms are outlined below. For each $\theta \in \Theta$, there is a nonempty set $N(\theta) \subset \Theta \setminus \{\theta\}$, called

the set of *neighbors* of θ , from which solution candidates are chosen when the current solution estimate is θ . A candidate value θ' is chosen from $N(\theta)$ with (conditional) probability $R(\theta, \theta')$. The next step is to decide whether to accept the candidate as the new solution estimate or reject it, in which instance, there is no change and the current estimate is taken to be the new estimate.

The sets of neighbors, it is assumed, are constructed in such a way that all points in Θ are *reachable* from one another. (This assumption ensures that every point in Θ is chosen as a candidate during the search for the optimum.) For a point θ' to be reachable from θ , there must be a finite sequence $\theta_0, \theta_1, \dots, \theta_l$ in Θ , for some integer $l \geq 1$, such that $\theta_0 = \theta, \theta_l = \theta'$ and, for $i = 0, 1, \dots, l-1$, the point θ_{i+1} belongs to the set of neighbors of θ_i . The smallest such l is the *distance* between θ and θ' .

The function $R(\cdot, \cdot)$, called the *transition probability* for Θ , satisfies

$$\sum_{\theta' \in N(\theta)} R(\theta, \theta') = 1, \quad (9)$$

and $R(\theta, \theta') > 0$ if and only if $\theta' \in N(\theta)$. The transition probability for Θ and neighbor sets are assumed to be symmetric, that is, $R(\theta, \theta') = R(\theta', \theta)$ and $\theta' \in N(\theta)$ if and only if $\theta \in N(\theta')$, for all θ, θ' belonging to Θ . It is sometimes convenient (as in Refs. [27–29]) to relax the requirement that R be symmetric and define it in terms of an unnormalized symmetric function R' , where

$$R(\theta, \theta') = \frac{R'(\theta, \theta')}{D(\theta)} \quad (10)$$

and the denominator $D(\theta) = \sum_{\theta' \in N(\theta)} R'(\theta, \theta') > 0$ for each θ in Θ . A simple example of an unnormalized function is $R'(\theta, \theta') = 1$ if and only if θ' belongs $N(\theta)$. The denominator $D(\theta)$ is then equal to the number of points in $N(\theta)$.

The estimates $\hat{\theta}_n$ in general random search algorithms are obtained as follows:

General Random Search Algorithm

Step 1: Pick an initial value $\hat{\theta}_0$ for the estimate. Initialize counter: $n = 0$.

Step 2: Given $\hat{\theta}_n$, choose a candidate new value for the solution $\theta'_n \in N(\hat{\theta}_n)$, with probability $R(\hat{\theta}_n, \theta'_n)$.

Step 3: Using samples of $Y(\theta'_n)$ and possibly $Y(\hat{\theta}_n)$, given θ'_n and $\hat{\theta}_n$, decide whether to accept or reject θ'_n as the new estimate. If accepted, put $\hat{\theta}_{n+1} = \theta'_n$; otherwise $\hat{\theta}_{n+1} = \hat{\theta}_n$.

Step 4: Increment counter, $n = n + 1$ and go to step (2).

Several methods that highlight random search approaches are the stochastic ruler, stochastic comparison, and simulated annealing (SAN) with noise.

Stochastic Ruler. Yan and Mukai [30] convert problem (2) into a maximization problem in which candidate solutions θ' are evaluated by comparing estimates of their loss function values $L(\theta')$ with an absolute scale. Candidate values are accepted if, with high probability, their loss function estimates are small compared with the scale value. The “scale” against which comparisons are made is a uniformly distributed random variable.

Yan and Mukai [30] showed that if U is uniformly distributed over (a, b) , then every solution to the following maximization problem solves the original minimization problem (2):

$$\max P\{Y(\theta) \leq U\}, \quad \theta \in \Theta, \quad (11)$$

provided that (a, b) is sufficiently large. The only assumption here, in addition to Θ being a finite set, is that, for each $\theta \in \Theta$, $Y(\theta)$ is independent of U and has a finite second moment. In addition, if the $Y(\theta)$'s are uniformly bounded random variables, then a and b can be taken to be any pair of upper and lower bounds.

In step (3) of the general random search algorithm, the original stochastic ruler method [30] uses estimates of the probability that $Y(\theta') \leq U$, given $\theta'_n = \theta'$ and current estimate $\hat{\theta}_n = \theta$, in the decision to accept or reject the current estimate. More precisely, given $\hat{\theta}_n = \theta$ and $\theta'_n = \theta'$, define $\hat{\theta}_{n+1}$ so that

$$\hat{\theta}_{n+1} = \begin{cases} \theta', & \text{with probability } p_n(\theta'_n), \\ \theta, & \text{with probability } 1 - p_n(\theta'_n), \end{cases} \quad (12)$$

where

$$p_n(\theta') = (\mathbb{P}\{Y(\theta') \leq U\})^{M_n} \quad (13)$$

and M_n is a fixed positive integer depending on n . The probability p_n is estimated using up to M_n samples of Y .

The sequence $\{\hat{\theta}_n\}$ is a nonstationary, irreducible Markov chain with state space Θ . Yan and Mukai [30] showed that $\hat{\theta}_n$ converges to the set Θ^* in probability, that is,

$$\lim_{n \rightarrow \infty} \mathbb{P}\{\hat{\theta}_n \in \Theta^*\} = 1, \quad (14)$$

if the sequence of positive integers $M_n \rightarrow \infty$ at rate $\log n$. In fact [Thm. 7.1 of Ref. 30],

$$M_n \sim C \log n \quad \text{as } n \rightarrow \infty, \quad (15)$$

where the constant $C > 0$ depends on the distances between points and the comparison probabilities $\mathbb{P}\{Y(\theta) \leq U\}$, $\theta \in \Theta$.

An obvious drawback of the stochastic ruler method is that M_n increases with n , which implies that the estimate of p_n in (13) requires an increasing number of samples of Y with increase in iteration. Alrefaei and Andradottir [27–29] developed modified versions of the stochastic ruler method in which M_n in (13) is constant. (In Ref. [28], the feasible region Θ is allowed to be countably infinite.)

The modified stochastic ruler methods differ from the original method of Yan and Mukai [30] in two key ways. First, the acceptance probability in the modified method is

$$p_n(\theta) = (\mathbb{P}\{Y(\theta) \leq U\})^M \quad (16)$$

where M is a fixed positive integer. Thus, unlike the acceptance probability (13) in the original method, the number of samples of Y required to estimate the probability p_n does not grow with n .

Second, the modified methods introduce an intermediate step in the decision to accept or reject a candidate solution. This decision generates a sequence $\{\tilde{\theta}_n\}$, which is then used to define the estimates $\hat{\theta}_n$ as a subsequence of $\{\tilde{\theta}_n\}$. To be explicit, initialize the algorithm with $\tilde{\theta}_0 \in \Theta$ and suppose that $\tilde{\theta}_n$ has been defined. Given $\tilde{\theta}_n$, choose θ'_n from $N(\tilde{\theta}_n)$ with

probability $R(\tilde{\theta}_n, \theta'_n)$. Given $\theta'_n = \theta'$ and $\tilde{\theta}_n = \tilde{\theta}$, the next value $\tilde{\theta}_{n+1}$ is defined by putting

$$\tilde{\theta}_{n+1} = \begin{cases} \theta', & \text{with probability } p_n(\theta'), \\ \tilde{\theta}, & \text{with probability } 1 - p_n(\theta'). \end{cases} \quad (17)$$

The estimates $\hat{\theta}_n$ are obtained as an embedded Markov chain that is defined by extracting a particular subsequence of $\{\tilde{\theta}_n\}$. Put $\hat{\theta}_0 = \tilde{\theta}_0$. Suppose that $\hat{\theta}_n$ has been defined. The decision whether or not to choose $\tilde{\theta}_{n+1}$ as the next value $\hat{\theta}_{n+1}$ or take $\hat{\theta}_{n+1}$ to be the current value $\hat{\theta}_n$ is based on the number of times that the Markov chain visits the two state values through the first n iterations. To be more precise, let $V_n(\theta)$ denote the number of times the Markov chain $\{\tilde{\theta}_n\}$ “visits” state θ in the first n steps. Thus, $V_n(\theta) = \#\{i : \tilde{\theta}_i = \theta, i = 1, 2, 3, \dots, n\}$. Put

$$\hat{\theta}_{n+1} = \begin{cases} \tilde{\theta}_{n+1}, & \text{if } V_{n+1}(\tilde{\theta}_{n+1})/D(\tilde{\theta}_{n+1}) \\ & > V_{n+1}(\hat{\theta}_n)/D(\hat{\theta}_n), \\ \hat{\theta}_n, & \text{otherwise.} \end{cases} \quad (18)$$

The function D , recall, is defined in Equation 10. State visits V_n are defined in terms of $\{\tilde{\theta}_n\}$ as follows. For $n = 0$, set $V_0(\theta) = 1$ if $\theta = \tilde{\theta}_0$; otherwise, $V_0(\theta) = 0$. Having defined $V_n(\theta)$, for $n \geq 1$, let

$$V_{n+1}(\theta) = \begin{cases} V_n(\theta) + 1, & \text{if } \theta = \tilde{\theta}_{n+1}, \\ V_n(\theta), & \text{otherwise.} \end{cases}$$

Alrefaei and Andradottir [29] considered variants of (18), which differ depending on how V_n is defined. Each of the modified stochastic ruler methods define $\{\hat{\theta}_n\}$ to be some subsequence of $\{\tilde{\theta}_n\}$.

The modified stochastic ruler has stronger convergence properties than the original. The modified method produces a stationary Markov chain such that [29]

$$\mathbb{P}\{\lim_{n \rightarrow \infty} \hat{\theta}_n \in \Theta^*(a, b)\} = 1. \quad (19)$$

In other words, the sequence of estimates in the modified method converge almost surely (a.s.). Wang [31] contains a comprehensive analysis of the rate of convergence of the stochastic ruler method and the stochastic comparison method discussed in the next

section. In addition, see [29] for rate of convergence results.

Stochastic Comparison. Stochastic comparison [32] is similar to the stochastic ruler method in that it replaces the minimization problem [Equation (2)] with a maximization problem that involves comparing random variables. In this method, each $Y(\theta)$ is compared, in probability, with all other $Y(\theta')$, $\theta' \neq \theta$. This comparison probability, denoted $\text{sp}(\theta)$, is called the *sigma-probability* corresponding to θ . Any θ that maximizes the sigma-probability over Θ solves the minimization problem [Equation (2)].

The sigma-probability is defined by

$$\text{sp}(\theta) = \sum_{\theta' \neq \theta} \mathbb{P}\{Y(\theta) < Y(\theta')\}, \quad \theta \in \Theta. \quad (20)$$

Gong *et al.* [32] showed that, if the measurement noises $\varepsilon(\theta)$ in (1) are i.i.d., zero mean, and have symmetric continuous probability density functions for every θ in Θ , then any point θ that maximizes $\text{sp}(\theta)$ over Θ solves the minimization problem [Equation (2)] and conversely.

The implementation steps of stochastic comparison are the same as the stochastic ruler but with a different acceptance probability. Given $\hat{\theta}_n = \theta$ and $\theta'_n = \theta'$, the following acceptance probability replaces (13)

$$p_n(\theta') = (\mathbb{P}\{Y(\theta') < Y(\theta)\})^{M_n}. \quad (21)$$

As in (13), this probability is estimated using up to M_n samples of $Y(\theta')$ and $Y(\theta)$.

Stochastic comparison yields a sequence of estimates $\hat{\theta}_n$ that converge in probability to Θ^* . The proof of convergence ([32]) is along the same lines as that for the stochastic ruler. (See [31] for rate of convergence analysis.)

Simulated Annealing. SAN for objective functions with noise is considered in [Section 8.2.2 of Ref. 26] and Refs. [33–37]. In terms of the general random search algorithm, SAN for noisy loss functions takes the following form. Given $\hat{\theta}_n = \theta$, choose θ'_n from $N(\theta)$ with probability $R(\theta, \theta'_n)$ in step (2). The decision

step (3) is, given $\hat{\theta}_n = \theta$ and $\theta'_n = \theta'$,

$$\hat{\theta}_{n+1} = \begin{cases} \theta', & \text{with probability } p_n(\theta', \theta), \\ \theta, & \text{with probability } 1 - p_n(\theta', \theta). \end{cases} \quad (22)$$

In this instance, the acceptance probability is

$$p_n(\theta', \theta) = \exp \left\{ \frac{-[\hat{L}(\theta') - \hat{L}(\theta)]^+}{T_n} \right\}, \quad (23)$$

where T_n is the usual *temperature* parameter, $\hat{L}(\theta')$ and $\hat{L}(\theta)$ are sample means of $Y(\theta')$'s and $Y(\theta)$'s, respectively, based on samples of sizes, say, k_n and k'_n , respectively, and $\alpha^+ = \max\{\alpha, 0\}$. The acceptance probability (23) is

$$p_n(\theta', \theta) = \exp \left\{ \frac{-[L(\theta') - L(\theta) + w_n]^+}{T_n} \right\}, \quad (24)$$

where w_n is the difference of the noise terms in the estimates of $L(\theta')$ and $L(\theta)$.

The assumptions in Refs. [33–35] on the transition probability $R(\cdot, \cdot)$ and temperature sequence $\{T_n\}$ is that convergence holds for the “undisturbed” SAN sequence, that is, the sequence of estimates $\hat{\theta}_n$ obtained by setting $w_n = 0$ in (24).

SAN algorithms differ depending on assumptions made about the temperature sequence $\{T_n\}$, the noise in (1) [and, hence, in (24)] and the estimates of the objective function values. For example, Gelfand and Mitter [33] showed that, if $T_n \rightarrow 0$ and if the noise w_n in (24) is Gaussian with mean zero noise and standard deviation $\sigma_n = o(T_n)$, then $\hat{\theta}_n$ converges to Θ^* in probability if and only if convergence in probability holds for the “undisturbed” SAN sequence. A consequence of the assumption $\sigma_n = o(T_n)$ is that $k'_n, k_n \rightarrow \infty$, so that the number of samples required to for the estimates $\hat{L}$ in (23) increases with iteration.

Gutjahr and Pflug [35] relaxed the Gaussian assumption in Ref. [33] on w_n by assuming only that the noise distribution is more concentrated or “peaked” about the origin than a Gaussian distribution with mean zero and standard deviation $\sigma_n = O(n^{-\alpha})$, where $\alpha > 1$ and $\sigma_n > 0$.

In other words, Thm. 4.1(ii) of Ref. 35 assumes that

$$P\{|w_n| \geq t\} \leq \int_{-t}^t \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{\frac{-x^2}{2\sigma_n^2}\right\} dx$$

for each $t \geq 0$ (i.e., the tail probabilities of the $N(0, \sigma_n^2)$ distribution dominate those of w_n).

The SAN algorithms by Alrefaei and Andradottir [37] seem to be the most practical in terms of implementation and yield strong convergence. They introduced modified SAN algorithms that relax the assumption that $T_n \rightarrow 0$, requiring only that the temperature T be a (strictly) positive constant. Their algorithms are similar to their embedded Markov chain algorithms for the stochastic ruler in Refs. [27–29]. The modified SAN algorithms in Ref. [37] produce estimates that converge a.s. to the set of loss function optimizers. The rate of convergence of this algorithm is considered in Ref. [36], which also examines the convergence rate for the modified stochastic ruler method.

Stochastic Approximation

SA refers to a class of recursive algorithms for solving two types of problems: finding roots and minima (or maxima) of noisy objective functions, where Θ is a subset of $\mathbb{R}^d$. There is a well-established SA literature for continuous parameter problems, beginning with the seminal works of Robbins and Monro [38] and Kiefer and Wolfowitz [39]. (See Chaps. 4–7 of Ref. 26 or Ref. 40 for surveys of SA methods.)

The general form of SA is

$$\hat{\theta}_{n+1} = \hat{\theta}_n - a_n \hat{g}_n(\hat{\theta}_n). \quad (25)$$

In continuous parameter minimization, the loss function L is assumed to be differentiable and $\hat{g}_n(\theta)$ is an estimate of its gradient $g(\theta)$. For the problem of finding the root of an unknown function, which is the classical SA setting of Robbins and Monro [38], the term $\hat{g}_n(\theta)$ is an estimate of the unknown function and is assumed to be unbiased. In all SA procedures, the *gain sequence* $\{a_n\}$ is a sequence of positive numbers such that $a_n \rightarrow 0$ and

$$\sum_n a_n = \infty. \quad (26)$$

In Robbins-Monro SA, the gain sequence also satisfies

$$\sum_n a_n^2 < \infty. \quad (27)$$

Kiefer-Wolfowitz SA, where the goal is to minimize an unknown function, imposes additional assumptions on the gain sequence. In this latter SA setting, the gradient estimate $\hat{g}_n(\theta)$ is biased and is obtained from noisy estimates of the loss.

If the loss function has a unique minimizer, then minimizing point θ^* is the root of the gradient. The goal, then, is to obtain a sequence $\{\hat{\theta}_n\}$ that converges a.s. or, more generally, in probability to the minimizing value of the loss.

Dupač and Herkenrath [41] considered the problem of finding the root of an unknown vector-valued function $g(\theta)$, defined on a lattice type discrete set Θ in $\mathbb{R}^d$ such as $\mathbb{Z}^d$. (A discrete subset of $\mathbb{R}^d$ is a lattice if it is an infinite set and the distance between pairs of points that are nearest neighbors is bounded by a fixed, positive constant.) The problem is to find a point in Θ that minimizes $|g(\theta)|$ using only noisy measurements of $g(\theta)$. (In the continuous parameter setting, the problem is to find a root of $g(\theta)$.) They apply their root finding procedure to the problem of optimizing real-valued functions defined on a lattice.

The general root-finding approach in Ref. [41] is to transform the discrete problem into a continuous one by interpolating the discrete function to a continuous function defined over $\mathbb{R}^d$ and, then, apply the Robbins-Monro SA procedure to find a root of the interpolated function. The resulting SA sequence $\{\hat{\theta}_n\}$ does not necessarily lie in the lattice Θ . From this sequence, [41] constructs a lattice-valued sequence $\{\tilde{\theta}_n\}$ that has convergence properties similar to the original. In particular, when $\{\hat{\theta}_n\}$ converges to θ^* , so does $\{\tilde{\theta}_n\}$ and in the same mode.

The lattice-valued sequence $\{\tilde{\theta}_n\}$ has a simple form when Θ is the lattice $\mathbb{Z}$. Each $\hat{\theta}_n$ satisfies the inequality $\lfloor \hat{\theta}_n \rfloor \leq \hat{\theta}_n \leq \lceil \hat{\theta}_n \rceil$. Moreover, there are unique scalars $\lambda_1, \lambda_2 \geq 0$ (depending on $\hat{\theta}_n$) such that $\lambda_1 + \lambda_2 = 1$ and $\hat{\theta}_n = \lambda_1 \lfloor \hat{\theta}_n \rfloor + \lambda_2 \lceil \hat{\theta}_n \rceil$. Put

$$\tilde{\theta}_n = \begin{cases} \lfloor \hat{\theta}_n \rfloor & \text{with probability } \lambda_1, \\ \lceil \hat{\theta}_n \rceil & \text{with probability } \lambda_2. \end{cases} \quad (28)$$

In other words, $\tilde{\theta}_n$ is a random “projection” of $\hat{\theta}_n$ onto the lattice Θ .

The interpolations in Ref. [41] are piecewise linear interpolations constructed so that they satisfy the conditions of Robbins-Monro SA and have the same set of solutions as the functions that they interpolate. In the root finding problem, in particular, the estimate $\hat{g}_n(\theta)$ in (25) is defined so that it is an unbiased estimate of the interpolation $\bar{g}$ of the unknown function g . Thus,

$$\mathbb{E}[\hat{g}_n(\hat{\theta}_n) | \hat{\theta}_n] = \bar{g}(\hat{\theta}_n). \quad (29)$$

Dupač and Herkenrath [41] showed that their root finding procedure converges exponentially fast when Θ is a lattice type set in $\mathbb{R}$. In the real discrete setting, their Corollary 1 implies that there is a positive constant C such that

$$\mathbb{P}(|\hat{\theta}_n - \theta^*| > 1) \leq \exp(-Cn). \quad (30)$$

In terms of the lattice-valued projections $\tilde{\theta}_n$ of $\hat{\theta}_n$,

$$\mathbb{P}(\tilde{\theta}_n \neq \theta^*) \leq \exp(-Cn). \quad (31)$$

There are other examples of the use of continuous interpolation in discrete optimization, for example, Gokbayrak and Cassandras [42] (the “surrogate function” method) and Wang and Schmeiser [43]. Both Refs. [42] and [43] assume that the feasible region Θ is a subset of $\mathbb{R}^d$ and both linearly interpolate functions over simplexes. The resulting continuous optimization problem is then solved by means of standard SA methods. Finally, the estimates in the continuous parameter problem are projected onto closest point in the discrete region to obtain a sequence of estimates for the solution of the original discrete problem. Thus, if $\hat{\theta}_n$ is the estimated solution to the continuous problem, then its projection $\tilde{\theta}_n$ onto Θ is taken to be the estimated solution to the original discrete problem.

The linear interpolations in Refs. [42] and [43] are essentially Lovász extensions of a function (Refs. [44] and [45]), which depend on the representation of points in a simplex. The Lovász extension of L on the unit cube in $\mathbb{R}^d$ is obtained by linearly interpolating L piecewise at the vertices of each simplex

in a particular collection of simplexes. Each simplex has $d + 1$ vertices and is defined in terms of a permutation on d integers. If τ is such a permutation (there are $d!$ unique permutations on d integers), then the simplex S_τ defined by τ is the set consisting of all points $x = (x_1, x_2, \dots, x_d)$ in the unit cube such that $x_{\tau(1)} \leq x_{\tau(2)} \leq x_{\tau(3)} \leq \dots \leq x_{\tau(d)}$. Each point in a simplex can be uniquely expressed as a convex combination of the vertices. To be more precise, if the vertices of S_τ are denoted $v_0, v_1, v_2, \dots, v_d$, then, for each x in S_τ , there are $d + 1$ scalars $\lambda_i \geq 0$, $i = 0, 1, 2, \dots, d$, called *weights*, such that $\sum_{i=0}^d \lambda_i = 1$ and

$$x = \sum_{i=0}^d \lambda_i v_i. \quad (32)$$

Such representations are unique. Each point in the unit cube has a representation, as the $d!$ simplexes form a covering for the unit cube. The representation is extended to any point in $\mathbb{Z}^d$ in the obvious manner. The representation (32) leads to a natural extension of functions on $\mathbb{Z}^d$. If f is a function defined on the vertices of the unit cube, then the simplex representation of points leads to the following natural extension:

$$\tilde{f}(x) = \sum_{i=0}^d \lambda_i f(v_i) \quad (33)$$

where x is given by (32).

Gerencsér *et al.* [46,47], Hill *et al.* [48,49], and Hill [50] introduced an optimization approach similar to Refs. [2] and [43], where the discrete region is $\mathbb{Z}^d$. In Refs. [48–50], the objective function is assumed to be integer convex and separable ([14]). The implication of this assumption is that the discrete loss function can be viewed as the restriction of a piecewise linear, separable convex function defined on all of $\mathbb{R}^d$. Thus, the resulting SA algorithm is a Robbins-Monro type procedure for convex optimization Ref. [51] (Sec. 5.6), Ref. [52].

Because piecewise linear functions may fail to be differentiable at some points, subgradients rather than gradients should be

used in the search for the optimum. References [48–50] use efficient estimates of the subgradient to obtain a computationally efficient SA algorithm. The subgradient estimates use the computationally efficient simultaneous perturbation (SP) approximation developed by Spall [53] in his simultaneous perturbation stochastic approximation (SPSA) algorithm for continuous parameter optimization. This type of subgradient estimate, which leads to discrete SPSA (DSPSA), is the key difference between Refs. [48–50] and other works that rely on continuous interpolation of the loss function.

For scalar θ , the SP subgradient estimate is [see Equation 16 of Ref. 50] (making the obvious substitutions for θ , c , and Δ)

$$\hat{g}_n(\hat{\theta}_n) = \frac{\bar{Y}(\hat{\theta}_n + c_n \Delta_n) - \bar{Y}(\hat{\theta}_n - c_n \Delta_n)}{2c_n \Delta_n} \quad (34)$$

where $\bar{Y}$ is a piecewise linear interpolation of Y , $c_n > 0$, and Δ_n is a ± 1 Bernoulli random variable independent of measurement noise. For small enough c_n , this subgradient estimate reduces to

$$\hat{g}_n(\hat{\theta}_n) = \begin{cases} (\bar{Y}(\hat{\theta}_n + 1) - \bar{Y}(\hat{\theta}_n - 1))/2, & \text{when } \hat{\theta}_n \text{ is an integer,} \\ \bar{Y}(\lfloor \hat{\theta}_n \rfloor + 1) - \bar{Y}(\lfloor \hat{\theta}_n \rfloor), & \text{otherwise,} \end{cases} \quad (35)$$

where $\lfloor \theta \rfloor$ is the vector consisting of the floor of each component of θ . The subgradient estimate in (34) has the obvious extension to vector valued θ 's. Replace $\bar{Y}$ by its Lovász extension. Then, (34) becomes

$$\hat{g}_n(\hat{\theta}_n) = \frac{(\bar{Y}(\hat{\theta}_n + c_n \Delta_n) - \bar{Y}(\hat{\theta}_n - c_n \Delta_n))}{2c_n} \Delta_n^{-1} \quad (36)$$

where Δ_n^{-1} is formed by taking inverses component-wise.

It should be noted here that the subgradient estimate (36) differs from that in Refs. [48–50]. The SA algorithm in those works contain an incorrect subgradient estimate, which can lead to the nonconvergence of $\hat{\theta}_n$ to Θ^* that was noted in Ref. [54]. The subgradient estimate in Refs. [48–50] for scalar θ reduces to $\hat{g}(\theta) = Y(\lfloor \hat{\theta} \rfloor) - Y(\lfloor \hat{\theta} \rfloor - 1)$, which

has expected value $L(\lfloor \hat{\theta} \rfloor) - L(\lfloor \hat{\theta} \rfloor - 1)$. At noninteger values of $\hat{\theta}$, the difference $L(\lfloor \hat{\theta} \rfloor) - L(\lfloor \hat{\theta} \rfloor - 1)$ is not a subgradient of the (linear) interpolation of L . The correct subgradient at noninteger $\hat{\theta}$ is given by $L(\lfloor \hat{\theta} \rfloor + 1) - L(\lfloor \hat{\theta} \rfloor)$, as the interpolation of L is differentiable between successive integer points and subgradients reduce to the derivative where the latter exists.

The subgradient approximation (36) requires $d + 1$ evaluations of the loss for each θ . An alternative is to use a randomized Lovász extension to define $\bar{Y}(\theta)$ for arbitrary θ . (The idea is similar to that used in Ref. [41] to project the estimates $\hat{\theta}_n$ onto the lattice.) For each θ in $\mathbb{R}^d$, let $v_0, v_1, v_2, \dots, v_d$ be the vertices and $\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_d$ the weights defined by the Lovász extension. Let $\tilde{\theta} = v_i$ with probability λ_i , and set $\bar{Y}(\theta) = Y(\tilde{\theta})$. Thus, the random variable $\bar{Y}(\theta)$ requires only a single evaluation of Y for each θ , which implies that the subgradient estimate in Equation 36 requires only two evaluations of Y . Note that the randomized extension has a mixture probability distribution given by

$$P\{\bar{Y}(\theta) \in A\} = \sum_{i=0}^d \lambda_i P\{Y(v_i) \in A\}, \quad (37)$$

where A is a measurable set of real numbers. It follows that

$$E[\bar{Y}(\theta)] = \bar{L}(\theta). \quad (38)$$

A variant of Refs. [48] and [50] is the DSPSA algorithm of Wang and Spall [54,55] and Wang [31]. The method of computing the difference estimates in Refs. [31,54,55] differ from those in Refs. [46–50] and applies to more general types of loss functions. Wang and Spall used “middle point” perturbations, where the observations Y of the loss function used to compute difference estimates are taken at $\lfloor \theta \rfloor + (\mathbf{1} \pm \Delta)/2$, and the term in the denominators are Δ_i^{-1} , rather than $2 \Delta_i^{-1}$. The notation $\mathbf{1}$ denotes the d -dimensional vector consisting of all 1's. Wang and Spall [54,55] and Wang [31] also examined the rate of convergence of their procedure. Wang [31] presents a detailed analysis of the convergence properties of the “middle point”

discrete SA procedure and compares its rate of convergence with the stochastic ruler and stochastic comparison methods.

Sample Path Methods

Sample path methods refer to a class of techniques that use sample average estimates of the loss function in the optimization (see, e.g., Refs. [43,56–60]). Thus, the estimates $\hat{\theta}_n$ are derived from a sample average estimate of the loss function, where the optimization is performed by optimizing $\hat{L}$. Thus, any applicable deterministic procedure can be used in the sample path approach.

There are two variants of sample path methods, depending on whether or not the sample path in the estimation is “fixed” or “variable.” To describe these estimates, let ω denote the random effects in the noise in (1). The observation corresponding to this value of ω is $Y(\theta, \omega)$. Let $\omega_1, \omega_2, \omega_3, \dots, \omega_m$ be an i.i.d. sample of ω , where $m \geq 1$ is fixed. This random sample defines a fixed sample path estimate of the loss

$$\hat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m Y(\theta, \omega_i). \quad (39)$$

The sample of ω 's in this estimate of the loss function remain fixed throughout the estimation to obtain $\hat{\theta}_n$. A fixed sample path estimate of the loss yields a fixed sample path estimate of the optimum. This definition assumes that the parameter θ in $Y(\theta, \omega)$ can varied for a fixed value of ω . This is possible, in particular, when the observations Y are obtained by Monte Carlo.

Variable sample path estimates are obtained by letting the random effects vary with each sample of observations and iteration to obtain the estimate $\hat{\theta}_n$. In other words, consider a sample of k_n i.i.d. observations of $\omega_1^n, \omega_2^n, \omega_3^n, \dots, \omega_{k_n}^n$ of ω at each iteration n . These observation define a variable sample path estimate of the loss

$$\hat{L}_{k_n}(\theta) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y(\theta, \omega_i^n). \quad (40)$$

The ω_i^j 's are assumed to be independent of the ω_i^l 's, for $j \neq l$. In addition, in the most general

form of (40), the sampling distribution for ω_i^n depends on the current estimate $\hat{\theta}_n$.

The key distinction between the two types of loss function estimates is that variable sample path estimates vary with iteration index for $\hat{\theta}_n$, whereas fixed sample path estimates do not.

The convergence of sample path methods has been extensively studied (see, e.g., Refs. [56–59]). Reference [59], for example, showed that the optimal value of $\hat{L}_m$, denoted $\hat{L}_m^*$, converges a.s. to the optimal value L^* of loss. It also studied approximate optimal solutions called δ -optimal solutions, $\delta \geq 0$. A point θ is a δ -optimal solution of L if $|L(\theta) - L^*| \leq \delta$. The loss function values of such points are within δ of the optimum L^* . The set of such points is denoted Θ_δ^* . When $\delta = 0$, the set of optimal and δ -optimal solutions coincide. Approximate optimal solutions of sample path estimates are similarly defined. Thus, a δ -optimal solution of $\hat{L}_m$ is any point θ such that $|\hat{L}_m(\theta) - \hat{L}_m^*| \leq \delta$. The set of such approximate optimal solutions is denoted $\Theta_{m,\delta}^*$. Convergence a.s. holds for approximate solutions ([59]): the set Θ_δ^* contains $\Theta_{m,\delta}^*$ a.s. for m sufficiently large.

The rates of convergence for sample path estimates are also known ([59]): for $\delta \geq 0$

$$\limsup_{m \rightarrow \infty} \frac{1}{m} \log [1 - P(\Theta_{m,\delta}^* \subset \Theta_\delta^*)] \leq -C. \quad (41)$$

In other words, the probability that Θ_δ^* contains $\Theta_{m,\delta}^*$ converges to 1 exponentially fast as $m \rightarrow \infty$. See Ref. [59] for performance bounds on sample path estimates. Reference [56] studied the convergence properties of variable sample path methods.

CONCLUSION

There are many approaches for discrete optimization. The particular choice depends on the problem structure, including the assumptions about the feasible region such as its cardinality, the measurement noise, and the loss function itself. For problems in which the feasible region is not large, statistical approaches are a reasonable choice, because those approaches require sampling the loss function at each point in the search space.

Random search methods and sample path methods are alternatives when the search space is large and there is little or no structure in the loss function that can be exploited in the optimization. For the random search methods, the modified stochastic ruler or modified SAN seem to be computationally efficient choices as they do not require an increasing number of function evaluations at each iteration. In situations where the search space is large and there are properties of the loss function that can be used in the search for an optimum, such as convexity, then stochastic approximation methods are available.

ACKNOWLEDGMENT

Sincere thanks are extended to the reviewers and the Topical Editor, Dr. James C. Spall. Their comments and suggestions for improving this manuscript were invaluable during the review process and were greatly appreciated.

REFERENCES

1. Wieselthier JE, Barnhart CM, Ephremides A. Optimal admission control in circuit-switched multihop radio networks. *Proceedings of the 31st Conference on Decision and Control, Tucson, AZ; 1992.* p 1011–1013.
2. Cassandras CG, Julka V. Scheduling policies using marked/phantom slot algorithms. *Queueing Syst Theory Appl* 1995;20:207–254.
3. Krishnamurthy V, Wang X, Yin G. Spreading code optimization and adaptation in cdma via discrete stochastic approximation. *IEEE Trans Infor Theory* 2004;50(9):1927–1949.
4. Mishra V, Bhatnagar S, Hemachandra N. Discrete parameter simulation optimization algorithms with application to admission control with dependent service times. *Proceedings of the 46th Conference on Decision and Control, New Orleans, LA; 2007.*
5. Ermoliev YM, Leonardi G. Some proposals for stochastic facility location models. *Math Model* 1982;3(5):407–420.
6. Ermoliev YM. Facility location problem. In: Ermoliev Y, Wets RJ-B, editors. *Numerical techniques for stochastic optimization*. New York: Springer; 1988. p 413–434.
7. Ho YC, Eyler MA, Chien TT. A gradient technique for general buffer storage design in a production line. *Int J Prod Res* 1979;17(6):557–580.
8. Spinellis D, Papadopoulos C, MacGregor Smith J. Large production line optimization using simulated annealing. *Int J Prod Res* 2000;38(3):509–541.
9. Francis RL. A “uniformity principle” for evacuation route allocation. *J Res Natl Bur Stand* 1981;86(5):509–513.
10. Ibaraki T, Katoh N. *Resource allocation problems*. Cambridge (MA): MIT Press; 1988.
11. Cassandras CG, Dai L, Panayiotou CG. Ordinal optimization for a class of deterministic and stochastic discrete resource allocation problems. *IEEE Trans Auto Contr* 1998;43(7):881–900.
12. Castañón DA, Wohletz JM. Model predictive control for stochastic resource allocation. *IEEE Trans Auto Contr* 2009;54(8):1739–1750.
13. Weisstein EW. Discrete set. Available at <http://mathworld.wolfram.com/DiscreteSet.html>, From MathWorld—A Wolfram Web Resource, Accessed 2014 Jan 24, copyright date (1999–2014).
14. Favati P, Tardella F. Convexity in nonlinear integer programming. *Ric Oper* 1990;53:3–44.
15. Hoffman KL. Combinatorial optimization: current successes and directions for the future. *J Comput Appl Math* 2000;124:341–360.
16. Sherali HD, Driscoll PJ. Evolution and state-of-the-art in integer programming. *J Comput Appl Math* 2000;124:319–340.
17. Gupta SS. On some multiple decision (selection and ranking) rules. *Technometrics* 1965;7(2):225–238.
18. Robbins H, Sobel M, Starr N. A sequential procedure for selecting the largest of k means. *Ann Math Stat* 1968;39(1):88–92.
19. Goldsman D, Nelson BL. Ranking, selection and multiple comparisons in computer simulation. *Proceedings of 1994 Winter Simulation Conference, Lake Buena Vista, FL; 1994.* p 192–198.
20. Hsu JC. *Multiple comparisons theory and methods*. London: Chapman and Hall; 1996.
21. Swisher JR, Jacobson SH, Yücesan E. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: a survey. *ACM Trans Model Comput Simul (TOMACS)* 13(2), 2003. p 134–154.

22. Kim S-H, Nelson BL. Recent advances in ranking and selection. Proceedings of the 2007 Winter Simulation Conference, Washington, DC; 2007. p 162–172.
23. Robbins H. Some aspects of the sequential design of experiments. *Bull Am Math Soc* 1952;55:527–535.
24. Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. *Adv Appl Math* 1985;6:4–22.
25. Lai TL. Sequential analysis: some classical problems and new challenges. *Stat Sin* 2001;11(2):303–350.
26. Spall JC. Introduction to stochastic search and optimization. Hoboken (NJ): John Wiley and Sons; 2003.
27. Alrefaei MH, Andradóttir S. Discrete stochastic optimization via a modification of the stochastic ruler method. Proceedings of 1996 Winter Simulation Conference, Coronado, CA; 1996. p 406–411.
28. Alrefaei MH, Andradóttir S. A modification of the stochastic ruler method for discrete stochastic optimization. *Eur J Oper Res—Theo Meth* 2001;133:160–182.
29. Alrefaei MH, Andradóttir S. Discrete stochastic optimization using variants of the stochastic ruler method. *Naval Res Log* 2005;52:344–360.
30. Yan D, Mukai H. Stochastic discrete optimization. *SIAM J Contr Optim* 1992;30(3): 594–612.
31. Wang Q. Optimization with Discrete Simultaneous Perturbation Stochastic Approximation Using Noisy Loss Function Measurements [PhD dissertation]. The Johns Hopkins University, Department of Applied Mathematics and Statistics; Oct 2013. ArXiv e-prints, arXiv:1311.0042.
32. Gong W-B, Ho YC, Zhai W. Stochastic discrete optimization. *SIAM J Optim* 1999;10(2): 384–404.
33. Gelfand SB, Mitter SK. Simulated annealing with noisy or imprecise energy measurements. *J Optim Theory Appl* 1989;62:49–62.
34. Fox BL, Heine GW. Probabilistic search with overrides. *Ann Appl Prob* 1995;5(4): 1087–1094.
35. Gutjahr WJ, Pflug GCh. Simulated annealing for noisy cost functions. *J Glob Optim* 1996;8:1–13.
36. Andradóttir S. Accelerating the convergence of random search methods for discrete stochastic optimization. *J Assoc Comput Mach* 1999;9(4):349–380.
37. Alrefaei MH, Andradóttir S. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Manage Sci* 1999;45(5):748–764.
38. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22: 400–407.
39. Kiefer J, Wolfowitz J. A stochastic approximation method. *Ann Math Stat* 1952;23: 426–466.
40. Kushner HJ. Stochastic approximation: a survey. *Wiley Interdiscip Rev Comput Stat* 2010;2(1):87–96.
41. Dupač V, Herkenrath U. Stochastic approximation on a discrete set and the multi-armed bandit problem. *Commun Stat Sequen Anal* 1982;1(1):1–25.
42. Gokbayrak CG, Cassandras K. Stochastic discrete optimization using a surrogate problem methodology. Proceedings of the 38th Conference on Decision and Control, Phoenix, AZ; 1999. p 1779–1784.
43. Wang H, Schmeiser BW. Discrete stochastic optimization using linear interpolation. Proceedings of the 2008 Winter Simulation Conference, WSC'08, Winter Simulation Conference; 2008. p 502–508.
44. Lovász L. Submodular functions and convexity. In: Bachem A, Grötschel M, Korte B, editors. *Mathematical Programming the State of the Art*. Berlin: Springer-Verlag; 1983. p 235–257.
45. Marichal J-L, Mathonet P. Approximations of lovász extensions and their induced interaction index. *Discrete Appl Math* 2008;156(1):11–24.
46. Gerencsér L, Hill SD, Vágó Z. Optimization over discrete sets via spsa. Proceedings of the 38th Conference on Decision and Control, Phoenix, AZ; 1999. p 1791–1795.
47. Gerencsér L, Hill SD, Vágó Z. Discrete optimization via spsa. Proceedings of 2001 American Control Conference, Arlington, VA; 2001. p 1503–1504.
48. Hill SD, Gerencsér L, Vágó Z. Stochastic approximation on discrete sets using simultaneous difference approximations. Proceedings of 2004 American Control Conference, Boston, MA; 2004. p 2795–2798.
49. Hill SD, Gerencsér L, Vágó Z. Discrete stochastic approximation via simultaneous difference approximations. Proceedings of 2005 American Control Conference, Portland, OR; 2005. p 307–308.

50. Hill SD. Discrete stochastic approximation with application to resource allocation. Johns Hopkins APL Tech. Dig. 2005;26(1):15–21.
51. Kushner HJ, Yin GG. Stochastic approximation algorithms and applications. New York: Springer; 1997.
52. He Y, Fu MC, Marcus SI. Convergence of simultaneous perturbation stochastic approximation for nondifferentiable optimization. IEEE Trans Automat Contr 2003;48(8): 1459–1463.
53. Spall JC. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans Automat Contr 1992;37(3):332–341.
54. Wang Q, Spall JC. Discrete simultaneous perturbation stochastic approximation on loss function with noisy measurements. Proceedings of 2011 American Control Conference, San Francisco, CA; 2011. p 4520–4525.
55. Wang Q, Spall JC. Rate of convergence analysis of discrete simultaneous perturbation stochastic approximation algorithm. Proceedings of 2013 American Control Conference, Washington, DC; 2013. p 4778–4783.
56. Homem-de-Mello T. Monte carlo methods for discrete stochastic optimization. In: Uryasev S, Pardalos PM, editors. Stochastic optimization: algorithms and applications. Norwell (MA): Kluwer Academic Publishers; 2001. p 97–119.
57. Homem-de-Mello T. Variable-sample methods for stochastic optimization. ACM Trans Model Comput Simul (TOMACS) 2003;13(2): 108–133.
58. Homem-de-Mello T. On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. SIAM J Optim 2008;19(2):524–551.
59. Kleywegt AJ, Shapiro A, Homem-de-Mello T. The sample average approximation method for stochastic discrete optimization. SIAM J Optim 2002;12(2):479–502.
60. Chen H, Schmeiser BW. Stochastic root finding via retrospective approximation. IIE Trans 2001;33(3):259–275.

DISCRETE-TIME MARTINGALES

KYLE Y. LIN

Operations Research Department,
Naval Postgraduate School,
Monterey, California

In the eighteenth century, martingales originally referred to a set of betting strategies that seemingly guarantees a sure win (pages 122–123 in Ref. 1). One such strategy is to instruct the gambler to wage \$1 on the first bet and to double up the wager after each loss, in the hope that he will eventually win a bet so as to win \$1 overall. For instance, if he loses three bets in a row, but wins the fourth bet, then he wins $-1 - 2 - 4 + 8 = \$1$.

In probability theory, a discrete-time martingale is a stochastic process designed to track a gambler's fortune in a sequence of fair bets. The concept of martingales came from Paul Pierre Levy, but it was Joseph Leo Doob (Ref. 2), who developed the theory. An important result of the martingale theory is that there exists no strategy that can guarantee a sure win in a fair game.

DEFINITION

In probability theory, a discrete-time martingale is a sequence of random variables that are motivated by a gambler's fortune in a fair game. Precisely, a discrete-time stochastic process $\{Z_n, n = 1, 2, \dots\}$ is called a *discrete-time martingale* if it satisfies the following two conditions:

$$E[|Z_n|] < \infty, \quad \text{for all } n, \quad (1)$$

and

$$E[Z_{n+1}|Z_1, Z_2, \dots, Z_n] = Z_n. \quad (2)$$

By interpreting Z_n as a gambler's fortune after the n th bet, Equation (2) states that the expected fortune after the $(n + 1)$ th bet is equal to the fortune after the n th bet, for

$n \geq 1$. In other words, the expected payoff of each bet is equal to 0, so the game is fair. The purpose of the technical condition in Equation (1), loosely speaking, is to ensure the mathematical regularity that allows the derivations of many useful results.

A discrete-time martingale tracks a gambler's fortune in a fair game. Taking expectations of Equation (2) results in

$$E[Z_{n+1}] = E[Z_n],$$

and therefore, $E[Z_n] = E[Z_1]$ for all n . An important result in martingale theory is that no strategy can guarantee a sure win in a fair game.

EXAMPLES

This section provides several examples of discrete-time martingales.

Example 1. Suppose that in each bet the gambler will win \$5, win \$2, or lose \$6, with respective probabilities 0.4, 0.2, and 0.4, independent of previous bets. Let Z_n denote the fortune after the n th bet, then $\{Z_n, n \geq 1\}$ is a discrete-time martingale, because

$$\begin{aligned} E[Z_{n+1}|Z_1, Z_2, \dots, Z_n] &= 0.4(Z_n + 5) \\ &+ 0.2(Z_n + 2) + 0.4(Z_n - 6) = Z_n. \end{aligned}$$

In fact, $\{Z_n, n \geq 1\}$ is a discrete-time martingale as long as the expected winning in each bet is \$0.

Although an arbitrary discrete-time stochastic process $\{X_n, n \geq 1\}$ is most likely not a discrete-time martingale, it is often possible to define another stochastic process $\{Z_n, n \geq 1\}$ based on $\{X_n, n \geq 1\}$, such that the new stochastic process is a discrete-time martingale.

Example 2. Let X_n denote the time of the n th arrival in a homogeneous Poisson process

with rate λ . The stochastic process $\{X_n, n \geq 1\}$ is not a discrete-time martingale because

$$E[X_{n+1}|X_1, \dots, X_n] = X_n + \frac{1}{\lambda} \neq X_n.$$

However, let $Z_n = X_n - n/\lambda$, then $\{Z_n, n \geq 1\}$ is a discrete-time martingale because

$$\begin{aligned} E[Z_{n+1}|Z_1, \dots, Z_n] \\ &= E\left[X_{n+1} - \frac{n+1}{\lambda} \middle| X_1, \dots, X_n\right] \\ &= \left(X_n + \frac{1}{\lambda}\right) - \frac{n+1}{\lambda} = X_n - \frac{n}{\lambda} = Z_n, \end{aligned}$$

where the first equality follows because $\{X_1, \dots, X_n\}$ can be deduced from $\{Z_1, \dots, Z_n\}$, so both of them provide the same information.

Example 3 [Likelihood Ratio Test].

Suppose we believe that a population is distributed either according to density function f or density function g . A random sample $X_1, \dots, X_n$ is taken and

$$Z_n = \prod_{i=1}^n \frac{g(X_i)}{f(X_i)}$$

is called the *likelihood ratio test statistic*. If the population is distributed according to density g , then the observations tend to occur at values where g is relatively large, so the test statistic tends to be large. For the same reason, if the population is distributed according to density f , then the test statistic tends to be small. The likelihood ratio test has important applications in statistical decision theory.

If the population is distributed according to density function f , then $\{Z_n, n \geq 1\}$ is a discrete-time martingale. To prove this claim, we compute

$$\begin{aligned} E[Z_{n+1}|Z_1, \dots, Z_n] \\ &= E\left[Z_n \frac{g(X_{n+1})}{f(X_{n+1})} \middle| Z_1, \dots, Z_n\right] \\ &= Z_n \times E\left[\frac{g(X_{n+1})}{f(X_{n+1})} \middle| Z_1, \dots, Z_n\right] \\ &= Z_n \times \int \frac{g(x)}{f(x)} f(x) \, dx = Z_n, \end{aligned}$$

where the third equality follows because X_{n+1} is independent of $X_1, \dots, X_n$, and therefore independent of $Z_1, \dots, Z_n$. The last equality follows because $g(x)$ is a density function, so $\int g(x) \, dx = 1$.

In the previous examples, it is quite straightforward to use Equation (2) directly to show that a stochastic process is a discrete-time martingale, but it is not always the case. Sometimes, it is more convenient to show that the conditional expected value of Z_{n+1} —given not only $Z_1, \dots, Z_n$, but also some additional information—is equal to Z_n (see page 296 in Ref. 3). If we can show that

$$E[Z_{n+1}|Z_1, \dots, Z_n, \mathbf{Y}] = Z_n,$$

then $\{Z_n, n \geq 1\}$ is a discrete-time martingale because

$$\begin{aligned} E[Z_{n+1}|Z_1, \dots, Z_n] \\ &= E[E[Z_{n+1}|Z_1, \dots, Z_n, \mathbf{Y}]|Z_1, \dots, Z_n] \\ &= E[Z_n|Z_1, \dots, Z_n] = Z_n, \end{aligned}$$

where the first equality follows from the identity

$$E[X|Z] = E[E[X|Y, Z]|Z]. \quad (3)$$

Example 4 [Doob Martingale]. When estimating the expected value of a random variable Y based on a sequence of observations $X_1, X_2, \dots$, we can write

$$Z_n = E[Y|X_1, X_2, \dots, X_n]$$

as the estimate after n observations. The sequence of estimates $Z_1, Z_2, \dots$ is a discrete-time martingale, and is named after Joseph Leo Doob. To show that $Z_1, Z_2, \dots$ is a discrete-time martingale, compute the conditional expected value of Z_{n+1} —given not only $Z_1, \dots, Z_n$, but also $X_1, \dots, X_n$ —as follows:

$$\begin{aligned} E[Z_{n+1}|Z_1, \dots, Z_n, X_1, \dots, X_n] \\ &= E[Z_{n+1}|X_1, \dots, X_n] \\ &= E[E[Y|X_1, \dots, X_n, X_{n+1}]|X_1, \dots, X_n] \\ &= E[Y|X_1, \dots, X_n] = Z_n, \end{aligned}$$

where the first equality follows because $Z_1, \dots, Z_n$ can be computed from $X_1, \dots, X_n$, and the third equality follows from Equation (3).

Perhaps two of the most important theorems concerning martingales are the martingale convergence theorem and the optional stopping theorem. Owing to its root in betting strategies, martingale theory has since found many applications in modern finance. Interested readers can refer to Refs 4 and 5 to see how martingale theory applies to financial modeling.

FURTHER READING

In this article, the discrete-time martingale is introduced without a rigorous mathematical definition based on measure theory. Readers can refer to Ref. 3 for a treatment with this approach.

A rigorous definition of discrete-time martingales relies on measure theory. Many textbooks on probability theory, such as Refs 2 and 6–9, introduce martingales with this approach. Reference 10 contains several interesting examples, where one can

cleverly devise a discrete-time martingale to derive elegant solutions for some interesting problems.

REFERENCES

1. Balsara NJ. Money management strategies for futures traders. New York: Wiley; 1992.
2. Doob JL. Stochastic processes. New York: Wiley; 1953.
3. Ross S. Stochastic processes. 2nd ed. New York: Wiley; 1996.
4. Bouleau N, Thomas A. Financial markets and martingales: observations on science and speculation. Germany: Springer; 2004.
5. Musiela M, Rutkowski M. Martingale methods in financial modelling. Germany: Springer; 2005.
6. Billingsley P. Probability and measure. 3rd ed. New York: Wiley; 1979.
7. Durrett R. Probability: theory and examples. 2nd ed. Belmont (CA): Duxbury Press; 1996.
8. Resnick SI. A probability path. Boston (MA): Birkhauser; 1999.
9. Williams D. Probability with martingales. Cambridge: Cambridge University Press; 1991.
10. Ross S, Peko EA. A second course in probability. Boston (MA): Peko books; 2007.

DISCRETIZATION METHODS FOR CONTINUOUS PROBABILITY DISTRIBUTIONS

ROBERT K. HAMMOND
J. ERIC BICKEL
The University of Texas, Texas,
USA

INTRODUCTION

Decision analysis problems typically involve probabilistic modeling of discrete and continuous uncertainties. Examples of discrete uncertainties include the presence of hydrocarbons and whether or not a competitor will enter the market. Examples of continuous uncertainties include the volume of hydrocarbons and market share. Discrete uncertainties are modeled with a probability mass function (pmf) and continuous uncertainties with a probability density function (pdf). In order to facilitate assessment or represent continuous uncertainties in a discrete decision tree, it is common for decision analysts to represent pdfs with properly designed pmfs, a process known as *discretization*.

Specifically, consider a decision model that takes as an input the continuous random variable X with support $S \subseteq \mathbb{R}$. We approximate the pdf $f(x)$, or the cumulative distribution function (cdf) $F(x)$, with a set of values $x_i \in S$, $i = 1, 2, \dots, N$, and associated probabilities $p_i \equiv p(x_i)$.

In most discretization methods, N is equal to three, but five is not uncommon. The values, which can be expressed as fractiles, or percentiles, of $F(x)$, and probabilities are selected such that desired properties of X are preserved. Common properties of interest are the moments of X (e.g., the mean, variance, skewness, and kurtosis). Smith [1] provided the following argument in support of matching moments: In decision and risk analysis, we are often interested in computing the expectation of some value function

(e.g., net present value or utility), $v(x)$, that is a function of the uncertain quantity X . If $v(x)$ can be closely approximated by a differentiable function, it can be written as a Taylor series expansion,

$$v(x) = v(a) + \frac{v'(a)}{1!}(x-a) + \frac{v''(a)}{2!}(x-a)^2 + \frac{v'''(a)}{3!}(x-a)^3 + \dots \quad (1)$$

Taking the expectation $E[v(X)]$ yields a weighted sum of the raw moments of X . For example, if $a = 0$, we have,

$$E[v(X)] = v(0) + \frac{v'(0)}{1!}E[X] + \frac{v''(0)}{2!}E[X^2] + \frac{v'''(0)}{3!}E[X^3] + \dots \quad (2)$$

Thus, a good approximation of $E[v(x)]$ must closely approximate the moments of X . In other words, in order to properly estimate the mean of our output variable, we need to match all the moments of the input variable.

Several discretization methods are in common use. For example, Extended Swanson–Megill (ESM) weights the 10th (P10), 50th (P50), and 90th (P90) percentiles of $F(x)$ by 0.300, 0.400, and 0.300, respectively. It is heavily used within the oil and gas industry [2–5], and has several other example applications in the literature [6–10]. A similar discretization is the “25-50-25” method, which weights the P10, P50, and P90 by 0.250, 0.500, and 0.250, respectively, and is also heavily used in the oil and gas industry [5]. Another method, Extended Pearson–Tukey (EPT), weights the P5, P50, and P95 by 0.185, 0.630, and 0.185. Example applications include Keeney [11] and Brooks and Kirkwood [12]. A flexible method called *bracket mean* (BMn), or equal areas, which uses the conditional means of partitions of a distribution is also commonly used [5]. Bickel et al. [5] and Hammond and Bickel [13] discussed the development and history of these and other methods.

The remainder of this article is organized as follows. The section titled “Discretization Methods” describes several discretization methods. The section “Example” gives example applications of some of these methods. The section “Multivariate Discretization” briefly discusses multivariate discretization, and the section “Discretization in Other Fields” gives an overview of discretization in other fields. The section “Recommendations for Practice” provides guidelines and recommendations for the practical application of these methods. Finally, the last section concludes.

DISCRETIZATION METHODS

Bickel et al. [5] categorized discretization methods into *shortcuts*, which do not require that the full distribution be known, and *distribution-specific*, which do. We first describe several shortcut methods, and then describe three distribution-specific methods.

Shortcuts

Shortcuts, as the name implies, are easy to apply and are a good starting point in probabilistic analysis. Many shortcuts have been proposed. Shortcuts are typically symmetric, that is, for a three-point discretization, $p_1 = p_3$ and $F^{-1}(x_1) = 1 - F^{-1}(x_3)$.

Table 1 summarizes several discretization shortcuts. Of these, EPT generally best preserves a distribution’s mean and variance, or the certain equivalent of a lottery when considering risk aversion [14, 15, 13]. ESM is generally best if the P10, P50, and P90 are used [13]. Although EPT is more accurate, ESM’s less extreme percentiles may be easier for experts to assess, making it more appropriate in some scenarios. MCS, though widely used, is generally inferior to ESM [16].

Hammond and Bickel [13] used the Pearson distribution system [17–19] to construct new three-point discretization shortcuts that are tailored to distributions with different ranges of support. The Pearson system subsumes many common distributions (e.g., normal, beta, uniform, and exponential) and is divided into several types. Three main types

make up the majority of Pearson distributions, the rest being special and limiting cases. These are the generalized beta (Type I), beta prime (Type VI), and the Type IV. The primary distinction between these three types are their ranges of support; the Type I has bounded support, $[a, b]$, the Type VI support is unbounded in one direction, $[a, \infty]$ or $[-\infty, b]$, and the Type IV has unbounded support, $[-\infty, \infty]$. Hammond and Bickel [13] found shortcuts that minimized the error in matching the mean, on average, over a large number of distributions in each of these types. This allows the analyst to tailor a shortcut to a distribution using only knowledge of the support boundedness. One set of shortcuts, the EPT++ methods,¹ were fit by changing the upper and lower discretization points and the probabilities, allowing them to be asymmetric. Another set, the standard percentile (SP) methods, fix the points to be the P10, P50, and P90, as these are commonly used in practice, and fit the probabilities. Table 2 gives the EPT++ shortcuts and Table 3 gives the SP shortcuts. Note that as the EPT++ shortcuts are asymmetric, and were fit to right-skewed distributions, the shortcuts should be reflected when dealing with left-skewed distributions. For example, the shortcut for a right-skewed type IV is (P7, P50, P94, 0.231, 0.551, 0.218), but for a left-skewed type IV, the shortcut would become (P6, P50, P93, 0.218, 0.551, 0.231). These shortcuts are more accurate than those in Table 1, on average, in matching the mean (and often the variance as well) of the distributions to which they were tailored.

Distribution-specific Methods

Distribution-specific methods require specifying the pdf (or the cdf) to be discretized, instead of only three percentiles. The first method we describe, Gaussian Quadrature (GQ), is a form of numerical integration designed to match the moments of a distribution. The other two methods, BMn,

¹The name indicates that these methods are an extension of the Pearson and Tukey [20] procedure. Hammond and Bickel [13] also found symmetric “EPT+” shortcuts.

Table 1. Summary of Shortcut Discretization Methods

Shortcut	Percentile Points	Probability Weights
Extended Pearson–Tukey (EPT)	P5, P50, P95	0.185, 0.630, 0.185
Extended Swanson–Megill (ESM)	P10, P50, P90	0.300, 0.400, 0.300
McNamee–Celona Shortcut (MCS)	P10, P50, P90	0.250, 0.500, 0.250
Zaino–D’Errico “Improved” (ZDI)	P4.2, P50, P95.8	0.167, 0.667, 0.167
Zaino–D’Errico–Tagichi (ZDT)	P11, P50, P89	0.333, 0.333, 0.333
Miller–Rice One-Step (MRO)	P8.5, P50, P91.5	0.248, 0.504, 0.248

Table 2. EPT++ Shortcuts

Distribution Type	Support	Percentile Points	Probability Weights
I-U Beta (EPT1U++)	$[a, b]$	P1, P50, P85	0.216, 0.491, 0.293
I-J Beta (EPT1J++)	$[a, b]$	P2, P50, P94	0.184, 0.615, 0.201
I- $\cap$ Beta (EPT1 $\cap$ ++)	$[a, b]$	P5, P50, P95	0.184, 0.632, 0.184
VI Beta Prime (EPT6++)	$[a, \infty]$	P4, P50, P96	0.164, 0.672, 0.164
IV (EPT4++)	$[-\infty, \infty]$	P7, P50, P94	0.231, 0.551, 0.218

Table 3. SP Shortcuts

Distribution Type	Support	Points	Probability Weights
I-U Beta (SP1U)	$[a, b]$	P10, P50, P90	0.228, 0.544, 0.228
I-J Beta (SP1J)	$[a, b]$	P10, P50, P90	0.273, 0.454, 0.273
I- $\cap$ Beta (SP1 $\cap$)	$[a, b]$	P10, P50, P90	0.296, 0.408, 0.296
VI Beta Prime (SP6)	$[a, \infty]$	P10, P50, P90	0.308, 0.384, 0.308
IV (SP4)	$[-\infty, \infty]$	P10, P50, P90	0.322, 0.356, 0.322

also known as *Equal Areas* [5], and bracket median (BMd), divide the cdf into intervals or brackets (three is common, but not necessary). BMn summarizes each interval with the conditional mean of that interval, whereas BMd uses the conditional median.

Gaussian Quadrature. GQ is a method of numerical integration that exactly matches the largest number of moments for any n -point discretization. Specifically, an n -point GQ will match the 0th through $2n-1$ st moments of the underlying distribution. Gauss [21] introduced the original method, which has been described more recently by Davis and Rabinowitz [22], among others. Miller and Rice [23] and Smith [1] described GQ applied to probability distributions for the purpose of discretization.

Smith [1] provided a procedure for finding an n -point GQ for a distribution F using

orthonormal polynomials.² First, construct the sequence of polynomials that are orthonormal with respect to F , $P_0^*, \dots, P_n^*$. These can be recursively computed:

$$\begin{aligned}
 P_{-1}(x) &= 0; P_0(x) = 1, \\
 P_{i+1}(x) &= (x - E[xP_i^{*2}(x)])P_i^*(x) \\
 &\quad - E[P_i^2(x)]^{1/2}P_{i-1}^*(x) \\
 &\text{for } 1 \leq i \leq n-1 \\
 &\text{where } P_i^*(x) = \frac{P_i(x)}{E[P_i^2(x)]^{1/2}}.
 \end{aligned}$$

E denotes the expectation with respect to F . The n roots of P_n^* , $y_1, \dots, y_n$, are distinct

²Miller and Rice [23] alternatively constructed a linear system of equations to solve for the polynomial coefficients.

and provide the n discretization values. The corresponding probabilities are given by,

$$p_i = \frac{k_n}{k_{n-1}P_n^*(y_i)P_{n-1}^*(y_i)},$$

where k_n and k_{n-1} are the leading coefficients of $P_n^*(x)$ and $P_{n-1}^*(x)$, and $P_n^*(x)$ denotes the derivative of $P_n^*(x)$. See Smith [1, 24] for more thorough discussion and proofs.

While GQ is the most accurate in terms of moment matching, it is also the most complex method to apply. GQs for some common distributions have been tabulated [5], but finding a GQ in general requires software implementation to find the roots y_i .

GQ typically uses rather extreme values (e.g., the P99.9) for some discretization points, in order to match higher moments. A criticism of this method is that expert assessments or historical data are often unreliable for such extreme values. We can address this issue by a modification that in turn reduces the number of moments that can be matched. Rather than taking the n roots of P_n^* as the discretization values, we can specify the values *a priori*, for example, using a set of percentiles. The probabilities matching the 0th through $n-1$ st moments can then be found by solving the linear equations,

$$\sum_{i=1}^n p_i x_i^j = \mu_j, \quad j = 0, \dots, n-1$$

$$p_i \geq 0, i = 1, \dots, n$$

for the p_i , where the x_i 's are given, μ_j is the j th raw moment, and $j = 0$ is the condition that the probabilities sum to one. However, this system is not guaranteed to have a solution for all sets of values x_i . Bickel et al. [5] demonstrated this method using the P10, P50, and P90 for uniform, normal, exponential, and triangular distributions.

Bracket Mean. BMn partitions the distribution support into intervals or brackets and then uses the conditional means of these regions. The method is more adaptable to the underlying distribution than the shortcuts and is also easier to apply than GQ. BMn

perfectly matches the mean of the underlying distribution, but always underestimates the variance and higher even-moments [23]. Hammond and Bickel [13] demonstrated that BMn's error in the variance, for three- and five-point bracket examples, tends to be higher than some three-point discretization shortcuts, such as EPT.

Part of BMn's appeal stems from the fact that it can be graphically explained and applied to a distribution for which the cdf, but not necessarily the functional form, is available [25]. For example, Fig. 1 shows the cdf of a standard-normal distribution. Divide the cumulative scale into three intervals: [0.0, 0.25], [0.25, 0.75], and [0.75, 1.0] (this division is common but arbitrary). To find the conditional mean of the first bracket, which we will use as the discretization point x_1 , find the location of the vertical line that makes the shaded areas A_1 and A_2 equal. The X value at this point is conditional mean. For a proof of this method, see Merkhofer [25] or McNamee and Celona [26]. Repeat for each interval, assigning the probability of an interval to the corresponding point as the discretization. Here, the points -1.24 , 0 , and 1.24 are respectively assigned probabilities 0.25 , 0.50 , and 0.25 .

Bracket Median. BMd uses the conditional *medians* of the partitions of the support. If partitioning on the cumulative probability axis, the medians can be determined simply by finding the appropriate percentiles. In that case, the median is just the percentile midway between the ends of a bracket (i.e., for the bracket [0, 0.25], the median is the 0.125 percentile, or the P12.5). As the conditional distributions are generally skewed, the median and the mean differ and BMd will regularly result in different discretizations than BMn.

Despite BMd's simple application, it typically incurs large errors in the moments. Hammond and Bickel [13] showed that three- and five-point BMd discretizations using equal-sized brackets tend to perform far worse than three- and five-point BMn, and even the less-accurate shortcuts, in matching the mean and variance of a distribution.

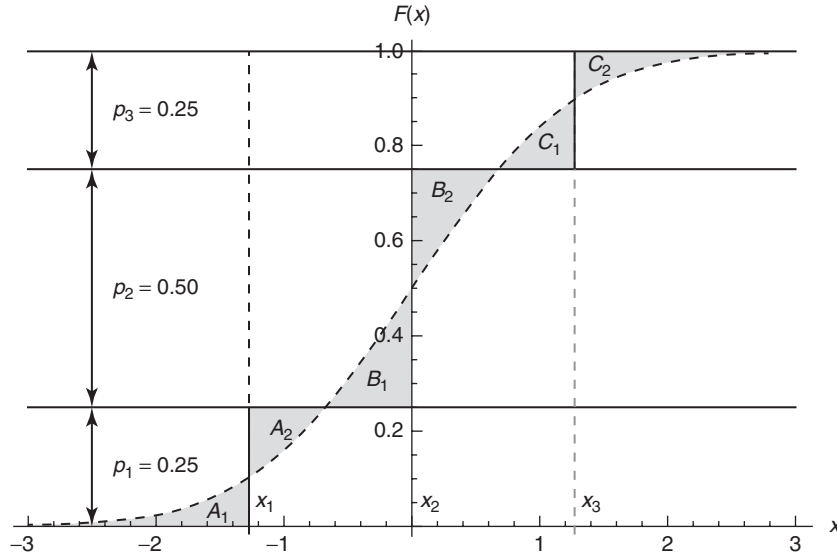


Figure 1. Example application of bracket mean with intervals $[0, 0.25]$, $[0.25, 0.75]$, $[0.75, 1]$.

Table 4. SP6 Values and Probabilities for Lognormal(1, 0.5)

Values	1.4322	2.7183	5.1592
Probabilities	0.3080	0.3840	0.3080
Percentiles (%)	10.00	50.00	90.00

Table 5. SP6 Moment Errors for Lognormal(1, 0.5)

	Mean	Variance	Skewness	Kurtosis
Est. Moments	3.0740	2.2180	0.4276	1.6398
% Error	-0.20%	-17.69%	-75.57%	-81.57%

However, BMd is frequently recommended in decision analysis texts [27–29].

EXAMPLE

In this section, we illustrate the use of SP, EPT++, GQ, and BMn on a lognormal distribution with parameters $\mu = 1$, $\sigma = 0.5$.

SP6

Our first example will demonstrate one of the SP shortcuts. Because the lognormal distribution is unbounded to only one side, we use

the SP6 shortcut. The lognormal cdf and the resulting SP6 discretization cdf are shown in Fig. 2. The values and probabilities used by the discretization are given in Table 4. The SP6 shortcut is similar to ESM, but assigns slightly higher probabilities to the P10 and P90. The SP6 moment estimations and errors for the mean, variance, skewness, and kurtosis of this lognormal distribution are given in Table 5. Note the high errors in the skewness and kurtosis estimations of -75.57% and -81.57% , respectively.

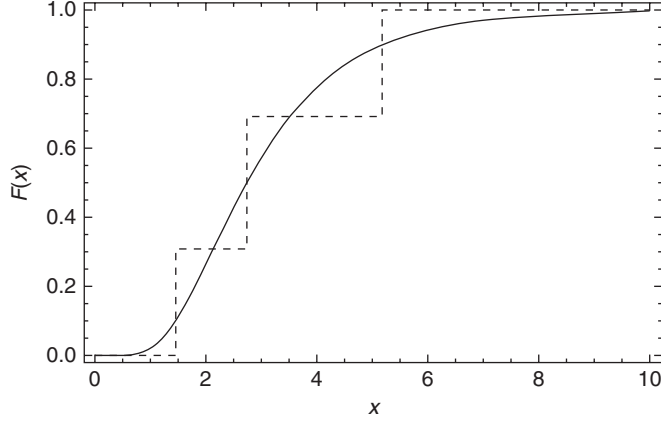


Figure 2. SP6 cdf.

Table 6. EPT6++ Values and Probabilities for Lognormal(1, 0.5)

Values	1.1328	2.7183	6.5231
Probabilities	0.1640	0.6720	0.1640
Percentiles (%)	4.00	50.00	96.00

EPT6++

Now we demonstrate one of the EPT++ shortcuts. Because the lognormal distribution is unbounded to only one side, we use the EPT6++ shortcut. The lognormal cdf and resulting discretization cdf are shown in Fig. 3. The values and probabilities for the discretization are given in Table 6. This shortcut closely matches the mean and variance, with errors of 0.07% and -1.52% , respectively, but significantly underestimates the skewness and kurtosis, with errors of -28.20% and -59.52% , respectively, as seen in Table 7. This poor performance in matching the skewness and kurtosis, also seen in Table 5 for SP6, is typical of shortcut methods, especially for unbounded distributions. The distribution tails, which shortcuts do not capture well, are more important to these higher moments than to the mean and variance. Again comparing Tables 5 and 7, we see that EPT6++ has lower errors in each of the first four moments than SP6, as its more extreme percentiles better capture the tails than SP6. However, EPT6++'s more extreme percentiles may be

more difficult to assess than the P10 and P90 used by SP6.

GQ

Here, we demonstrate three- and five-point GQs, GQ3, and GQ5, respectively. GQ3 matches moments 0 through 5 and GQ5 matches moments 0 through 9, so we do not give moment errors.³ The cdfs are shown in Fig. 4, and the values, probabilities, and equivalent percentiles are given for GQ3 and GQ5 in Tables 8 and 9, respectively. Notice the extreme percentiles used by each method, both making use of the P99.9, or more extreme, percentiles.

The set of orthonormal polynomials for GQ3 in this example are

$$P_{-1}^*(x) = 0$$

$$P_0^*(x) = 1$$

$$P_1^*(x) = -1.876 + 0.609x$$

$$P_2^*(x) = 2.640 - 1.524x + 0.169x^2$$

$$P_3^*(x) = -3.207 + 2.484x - 0.489x^2 + 0.024x^3$$

³Since GQ requires numeric computation, there will generally be negligible numerical errors.

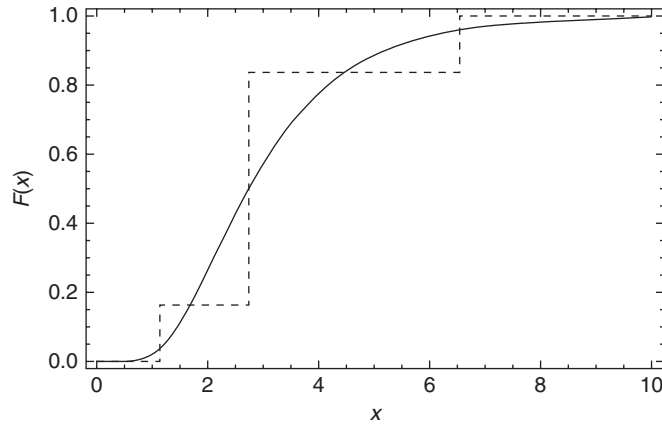


Figure 3. EPT6++ cdf.

Table 7. EPT6++ Moment Errors for Lognormal(1, 0.5)

	Mean	Variance	Skewness	Kurtosis
Est. Moments	3.0822	2.6539	1.2567	3.6017
% Error	0.07%	-1.52%	-28.20%	-59.52%

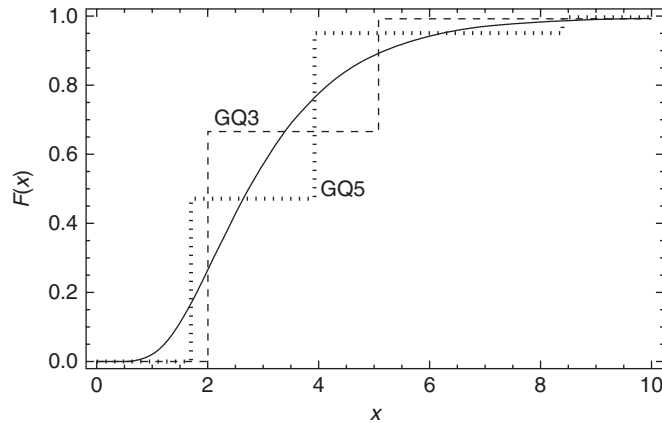


Figure 4. GQ3 and GQ5 cdfs.

BMn

Figure 5 shows the cdfs for a three-point bracket mean discretization (BMn3) using the brackets $[0, 0.25]$, $[0.25, 0.75]$, and $[0.75, 1]$ (alternatively referred to as 25-50-25) and a five-point bracket mean discretization (BMn5) using equal brackets of 0.2. The values and probabilities for BMn3 and BMn5 are given in Tables 10 and 11, respectively.

While any BMn discretization will theoretically match the mean of the distribution, they are often inferior to some shortcuts in matching higher moments. The moment errors for BMn3 and BMn5 in Tables 12 and 13, respectively, are larger than those for EPT6++ in the variance, skewness, and kurtosis, as seen by comparing these to Table 7.

The discretization points are more tightly clustered about the mode of the distribution

Table 8. Three-point GQ of Lognormal(1, 0.5)

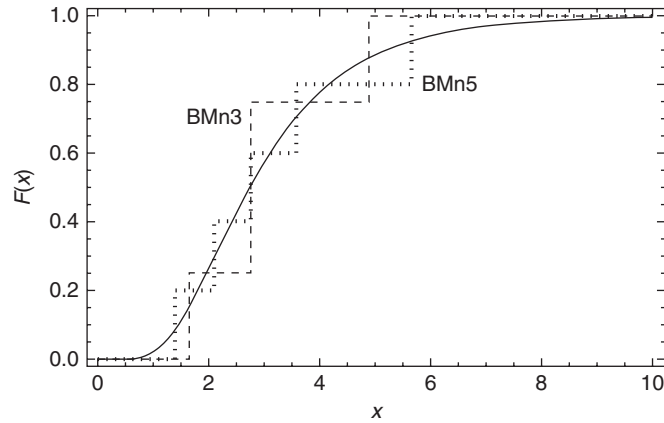
Values	2.0003	5.0784	12.8934
Probabilities	0.6652	0.3285	0.0063
Percentiles (%)	26.98	89.44	99.91

Table 9. Five-point GQ of Lognormal(1, 0.5)

Values	1.6885	3.9121	8.3729	17.9201	41.5203
Probabilities	0.4729	0.4789	0.0477	0.0005	2.6081E-07
Percentiles (%)	17.05	76.67	98.78	99.99	≈ 100.00

Table 10. Three-point BMn Values and Probabilities for Lognormal(1, 0.5) Using Cumulative Brackets [0, 0.25], [0.25, 0.75], and [0.75, 1]

Values	1.4797	2.7670	5.3071
Probabilities	0.2500	0.5000	0.2500
Percentiles (%)	11.19	51.42	90.96

**Figure 5.** BMn3 and BMn5 cdfs.**Table 11. Five-point BMn Values and Probabilities for Lognormal(1, 0.5) Using Equal Cumulative Brackets of 0.2**

Values	1.3839	2.0909	2.7255	3.5591	5.6417
Probabilities	0.2000	0.2000	0.2000	0.2000	0.2000
Percentiles (%)	8.85	29.98	50.21	70.51	92.79

Table 12. BMn3 Moment Errors for Lognormal(1, 0.5)

	Mean	Variance	Skewness	Kurtosis
Est. Moments	3.0802	1.9292	0.6421	2.0939
% Error	0.00%	-28.41%	-63.31%	-76.47%

Table 13. BMn5 Moment Errors for Lognormal(1, 0.5)

	Mean	Variance	Skewness	Kurtosis
Est. Moments	3.0802	2.1545	0.6971	2.2558
% Error	0.00%	-20.05%	-60.17%	-74.65%

than GQ, which, like the shortcuts, fails to capture the distribution tails. Whether BMn or a shortcut is preferred depends on the importance of the different moments.

MULTIVARIATE DISCRETIZATION

Here, we briefly discuss multivariate discretization. Much of the arguments for, and analysis of, univariate discretization methods does not extend to multiple variables. For example, there is no “mean” of a multivariate distribution $f(X_1, \dots, X_N)$ with support S^N without specification of a function $v(x_1, \dots, x_N): S^N \rightarrow \mathbb{R}$.

Keefer and Bodily [14] discussed sequential discretization of multivariate distributions. If we wish to discretize $f(X, Y)$, first discretize X , then discretize the conditional distribution $Y|x_i$. Some examples make use of sequential discretization of individual variables of a multivariate model with dependent components. Clemen and Reilly [30] and Wang and Dyer [31] used copulas to model multivariate distributions, and discretized the resulting conditional distributions.

Alternatively, the uncertainties of a multivariate distribution may be simultaneously discretized. Hoyland and Wallace [32] fit a discretization to multiple uncertainties by solving a (generally nonconvex) optimization problem of minimizing a distance measure of the differences in statistical properties of the continuous and discrete joint distributions. This optimization problem generally does not guarantee an optimal solution, and like GQ, requires software implementation.

DISCRETIZATION IN OTHER FIELDS

Discretization has wide application in fields other than decision analysis. Some of these have different aims than decision analytic discretization, such as summarizing or

partitioning data rather than approximating a continuous probability distribution.

Data Mining and Machine Learning

The data mining and machine learning fields are concerned with classification problems and often discretize real-valued observational data. Kotsiantis and Kanellopoulos [33] reviewed the literature on discretization for machine learning applications. They discussed variations of bracket methods for summarizing and categorizing data statically (considering one uncertainty at a time) and dynamically (considering multiple uncertainties simultaneously to incorporate dependence). Mizianty et al. [34] gave a wealth of references for discretization in classification problems, and compare the performance of 12 Naïve Bayes discretization algorithms.

Stochastic Programming

Multistage stochastic programming is another field that often requires distribution approximations to make the problems tractable. Kaut and Wallace [35] reviewed several methods used in the stochastic programming literature, including moment matching using GQ. Kall et al. [36] discussed discretization using partitions of the distributions’ support (bracket methods), particularly a method corresponding to BMn, which can be used with the Jensen and Edmundson–Madansky inequalities to bound the optimal value [37]. Other methods include Monte Carlo sampling [38] and nonlinear optimization [39, 40]. Discretization is also sometimes called *scenario generation*, as the result is a discrete set of uncertainty realizations.

Histograms

Histograms are commonly used to graphically represent frequency data and are discrete by definition. Histograms represent the frequencies or counts of observations falling

within a particular range (also known as a *bin*) as the height of the bar in the graph. The discretization methods discussed in this article are typically used as simplifying approximations to continuous distributions, whereas histograms are used to represent the distributions implied by empirical data [41]. The parameters for constructing a histogram include the number of bins, bin widths, and orientation of the bins. Two common heuristic rules for selecting the parameters for histograms with homogenous bin widths are Sturges' Rule, which determines the number of bins, and Scott's Rule, which determines the bin width for a given number of bins [41, 42, 43]. The flexibility of histogram derivation rules are limited by the quality and format of the empirical data, such as when the data are collected into bins of unequal width [44].

PERT

The Program Evaluation and Review Technique (PERT) is a popular method for estimating project completion time that began as a US Navy project to develop a quantitative evaluation methodology for project management [45]. PERT methods typically use separate estimation formulae for the mean and variance that don't necessarily form discrete probability distributions. The classical PERT formula for estimating the mean and variance use weighted combinations of the P0, mode, and P100 [46]. Many variations of these formulae have been proposed [47–55].

Finance

Discretization is often used in financial asset pricing models, such as the well-known binomial lattice model [56]. Moment-matching is common, and examples include Tauchen and Hussey [57], who used GQ in asset pricing with dynamic nonlinear models, and Sullivan [58] who used it to value American put options.

RECOMMENDATIONS FOR PRACTICE

Shortcut methods are useful as first approximations, which can be used in sensitivity

analysis to help identify important uncertainties. These uncertainties can then be given more attention when ascertaining the full distribution and using discretization methods such as GQ or BMn with several points. The discretizations should evolve with the analysis and be adapted to the importance of specific uncertainties. The characteristics of the decision ultimately determine the appropriateness of model refinement.

The choice of discretization method generally depends on two factors: the information available and the significance of the uncertainty to the decision(s). GQ is a highly accurate method that matches several moments of the distribution, but requires that these moments exist and be known. It can also be difficult to implement, and as a result might be best reserved for approximating key uncertainties. GQs have been tabulated for some common distribution families (e.g., normal, uniform, and triangular), but use of the method generally requires software [5].

If the full distribution is known, but GQ is not feasible or not desired, there are many potential methods to choose from. BMn, BMd, and GQ all give the analyst the freedom to choose the number of points, and in the case of BMn and BMd, the intervals to use. BMn is generally more accurate than BMd in matching moments using a given number of discrete points and specified intervals. For multi-modal distributions, BMn is far more accurate than shortcut methods, which can be very poor at approximating these distributions [13]. BMd is the simplest of these to automate, as it requires only reading percentiles from the continuous distribution, and may be preferred if calculation time is a concern.

It is often the case in decision analytic applications that probability assessments are expensive or time-consuming, such as resulting from the output of a large simulation model or as the result of careful expert judgment. In these cases, it may only be feasible to assess a few percentiles of the distribution, in which case discretization shortcuts are most valuable. EPT and ZDI are generally good choices, but better performance might be obtained using the EPT+ and

EPT++ shortcuts for their respective distribution types. If the distribution is bounded on both ends, then the analyst might be able to use knowledge of its shape (e.g., whether it is $\cap$ -shaped) to further specify the appropriate approximation. If the analyst is constrained to use the common P10, P50, and P90 percentiles, then choosing one of the SP methods provides distinct improvements over ESM or MCS. However, it should be remembered that shortcuts can incur large errors for even moderately skewed distributions. Shortcuts using more extreme percentiles (e.g., P5 and P95) tend to be more accurate for skewed distributions, but still incur errors. As mentioned above, shortcuts are particularly useful as first approximations in sensitivity analysis, to determine which uncertainties are most important and warrant more careful assessment and discretization.

SUMMARY

We have described the use of several different methods of discretizing continuous probability distributions, each having its own advantages and drawbacks. GQ is the best in terms of matching the moments of a distribution, but is not trivial to apply in practice without software support, and often uses extreme discretization values. Shortcuts, however, are generally tailored *a priori* to a wide range of distributions, and while they usually have some error (which can be large), they are very easy to apply. BMn is a compromise between the two approaches, being relatively easy to apply and reliably matching the mean. However, it does not match higher moments well.

Which method is most appropriate ultimately depends on the situation. Uncertainties to which a decision is very sensitive should be more accurately assessed and discretized during the analysis. Shortcuts are useful for first approximations of distributions used in initial probabilistic analysis. The iterative nature of the decision analysis cycle [59] should be used to refine the discretizations of those uncertainties that are most important, as a part of sensitivity analysis and information gathering.

REFERENCES

1. Smith JE. Moment methods for decision analysis. *Manag Sci* 1993;39(3):340–358.
2. Megill RE. *An introduction to risk analysis*. Tulsa, Oklahoma: PennWell; 1984.
3. Hurst A, Brown GC, Swanson RI. Swanson's 30-40-30 rule. *AAPG Bull* 2000;84(12):1883–1891.
4. Rose PR. Volume 12, *Risk analysis and management of petroleum exploration ventures*. Tulsa, Oklahoma: The American Association of Petroleum Geologists; 2001.
5. Bickel JE, Lake LW, Lehman J. Discretization, simulation, and Swanson's (inaccurate) mean. *SPE Econ Manag* 2011;3(3):128–140.
6. Keefer DL. Facilities evaluation under uncertainty: pricing a refinery. *Interfaces* 1995;25(6):57–66.
7. Bailey DE, Loos JJ, Perry ES, *et al.* A retrospective evaluation of 316(b) mitigation options using a decision analysis framework. *Environ Sci Policy* 2000;3:S25–S36.
8. Stonebraker JS. How Bayer makes decisions to develop new drugs. *Interfaces* 2002;32(6):77–90.
9. Stonebraker JS, Keefer DL. Modeling potential demand for supply-constrained drugs: a new hemophilia drug at Bayer biological products. *Oper Res* 2009;57(1):19–31.
10. Bailey MJ, Snapp J, Yetur S, *et al.* Practice summaries: American airlines uses should-cost modeling to assess the uncertainty of bids for its full-truckload shipment routes. *Interfaces* 2011;41(2):194–196.
11. Keeney RL. An analysis of the portfolio of sites to characterize for selecting a nuclear repository. *Risk Anal* 1987;7(2):195–218.
12. Brooks DG, Kirkwood CW. Decision analysis to select a microcomputer networking strategy: a procedure and a case study. *J Oper Res Soc* 1988;39(1):23–32.
13. Hammond R, Eric Bickel J. Reexamining discrete approximations to continuous distributions. *Decision Anal* 2013a;10(1):6–25.
14. Keefer DL, Bodily SE. Three-point approximation for continuous random variables. *Manag Sci* 1983;29(5):595–609.
15. Keefer DL. Certainty equivalents for three-point discrete-distribution approximations. *Manag Sci* 1994;40(6):760–773.
16. Hammond R, Eric Bickel J. Approximating continuous probability distributions using the 10th, 50th, and 90th percentiles. *Eng Econ* 2013b;58(3):189–208.

17. Pearson K. Contributions to the mathematical theory of evolution. Ii. Skew variation in homogenous material. *Philos Trans R Soc Lond* 1895;186:343–414.
18. Pearson K. Mathematical contributions to the theory of evolution. X. Supplement to a memoir on skew variation. *Philos Trans R Soc Lond* 1901;197:443–459.
19. Pearson K. Mathematical contributions to the theory of evolution. Xix. Second supplement to a memoir on skew variation. *Philos Trans R Soc Lond* 1916;216(429–457):429.
20. Pearson ES, Tukey JW. Approximate means and standard deviations based on distances between percentage points of frequency curves. *Biometrika* 1965;52(3/4):533–546.
21. Gauss CF. *Methodus nova integralium valores per approximationem inveniendi*. Göttingen, Germany: *Carl friedrich gauss werke*. Königlichen Gesellschaft der Wissenschaften; 1866. pp. 163–196.
22. Davis PJ, Rabinowitz P. *Methods of numerical integration*. Orlando, Florida: Academic Press, Inc.; 1984.
23. Miller AC, Rice TR. Discrete approximations of probability distributions. *Manag Sci* 1983;29(3):352–362.
24. Smith JE. *Moment methods for decision analysis*. Engineering-Economic System: Stanford University, Stanford, CA; 1990.
25. Merkhofer MW. *Flexibility and decision analysis*. Engineering-Economic Systems: Stanford University; 1975.
26. McNamee P, Celona J. *Decision analysis with supertree: instructor's manual*. South San Francisco, California: The Scientific Press; 1991.
27. Schlaifer R. *Analysis of decisions under uncertainty*. New York: McGraw-Hill; 1969.
28. Holloway CA. *Decision making under uncertainty: models and choices*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.; 1979.
29. Clemen RT. *Making hard decisions: an introduction to decision analysis*. Boston, Massachusetts: PWS-KENT Publishing Company; 1991.
30. Clemen RT, Reilly T. Correlations and copulas for decision and risk analysis. *Manag Sci* 1999;45(2):208–224.
31. Wang T, Dyer JS. A copulas-based approach to modeling dependence in decision trees. *Oper Res* 2012;60(1):225–242.
32. Hoyland K, Wallace SW. Scenario trees for multistage decision problems. *Manag Sci* 2001;47(2):295–307.
33. Kotsiantis S, Kanellopoulos D. Discretization techniques: a recent survey. *GESTS Int Trans Comp Sci Eng* 2006;32(1):47–58.
34. Mizianty MJ, Kurgan MA, Ogiela MR. Discretization as the enabling technique for the naive bayes and semi-naive bayes based classification. *Knowl Eng Rev* 2010;25(4):421–449.
35. Kaut M, Wallace SW. Evaluation of scenario-generation methods for stochastic programming. *Pac J Optim* 2007;3(2):257–271.
36. Kall P, Ruszczyński A, Frauendorfer K. Approximation techniques in stochastic programming. In: Ermoliev Y, Wets R, editors. *Numerical techniques for stochastic optimization*. Heidelberg, Germany: Springer-Verlag; 1988. pp. 33–64.
37. Huang CC, Ziemba WT, Ben-Tal A. Bounds on the expectation of a convex function of a random variable: with applications to stochastic programming. *Oper Res* 1977;25(2):315–325.
38. Birge JR, Louveaux F. *Introduction to stochastic programming*. New York: Springer-Verlag; 1997.
39. Høyland K, Wallace SW. Generating scenario trees for multistage decision problems. *Manag Sci* 2001;47(2):295–307.
40. Pflug GC. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Math Program* 2001;89(2):251–271.
41. Scott DW. Sturges' and scott's rules. In: Lovric M, editor. *International encyclopedia of statistical science*. Berlin: Springer-Verlag; 2011. pp. 1563–1566.
42. Scott DW. On optimal and data-based histograms. *Biometrika* 1979;66:605–610.
43. Scott DW. Scott's rule. *Wiley Interdiscip Rev Comput Stat* 2010;2:497–502.
44. Scott DW, Scott WR. Smoothed histograms for frequency data on irregular intervals. *Am. Statistician* 2008;62(3):256–261.
45. Malcolm DG, Roseboom JH, Clark CE, et al. Application of a technique for research and development program evaluation. *Oper Res* 1959;7(5):646–669.
46. MacCrimmon KR, Ryavec CA. An analytical study of the pert assumptions. *Oper Res* 1964;12(1):16–37.
47. Moder JJ, Rodgers EG. Judgment estimates of the moments of pert type distributions. *Manag Sci* 1968;15(2):B76–B83.
48. Perry C, Greig ID. Estimating the mean and variance of subjective distributions in

- pert and decision analysis. *Manag Sci* 1975; 21(12):1477–1480.
49. Farnum NR, Stanton LW. Some results concerning the estimation of beta distribution parameters in pert. *J Oper Res Soc* 1987; 38(3):287–290.
 50. Golenko-Ginzburg D. On the distribution of activity time in pert. *J Oper Res Soc* 1988; 39(8):767–771.
 51. Johnson D. The robustness of mean and variance approximations in risk analysis. *J Oper Res Soc* 1998;49(3):253–262.
 52. Lau H-S, Lau AH-L. An improved pert-type formula for standard deviation. *IIE Trans* 1998;30(3):273–275.
 53. Shankar NR, Sireesha V. An approximation for the activity duration distribution, supporting original pert. *Appl Math Sci* 2009; 3(57):2823–2834.
 54. Shankar NR, Rao KSN, Sireesha V. Estimating the mean and variance of activity duration in pert. *Int Math Forum* 2010;5(18): 861–868.
 55. Kim SD, Hammond R, Eric Bickel J. On the accuracy of pert discretizations. *IEEE Trans Eng Manag* 2014;61(2):362–369.
 56. Cox JC, Ross SA, Rubinstein M. Option pricing: a simplified approach. *J Financ Econ* 1979;7(3):229–263.
 57. Tauchen G, Hussey R. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. *Econometrica* 1991;59(2):371–396.
 58. Sullivan MA. Valuing american put options using gaussian quadrature. *Rev Financ Studies* 2000;13(1):75–94.
 59. Matheson JE, Howard RA. An introduction to decision analysis, *Reading in decision analysis*. Menlo Park, California: Decision Analysis Group, SRI International; 1976.

DISJUNCTIVE INEQUALITIES: APPLICATIONS AND EXTENSIONS

PIETRO BELOTTI
Department of Industrial and
Systems Engineering, Lehigh
University, Bethlehem,
Pennsylvania

LEO LIBERTI
LIX, École Polytechnique, France

ANDREA LODI
DEIS, Università di Bologna,
Bologna, Italy

GIACOMO NANNICINI
Tepper School of Business,
Carnegie-Mellon University,
Pittsburgh, Pennsylvania

ANDREA TRAMONTANI
DEIS, Università di Bologna,
Bologna, Italy

INTRODUCTION

A general optimization problem can be expressed in the form

$$\min \{cx : x \in S\}, \quad (1)$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, $c \in \mathbb{R}^n$ is a linear objective function, and $S \subset \mathbb{R}^n$ is the set of feasible solutions of Equation (1). Because S is generally hard to deal with, a possible approach for tackling Equation (1) is to optimize the objective function over a suitable relaxation (i.e., easy to solve) $P \supseteq S$. Let x^* be the optimal solution over P . If $x^* \in S$ the problem is solved. Otherwise, one can derive a valid inequality for S in order to separate x^* from S , that is, an inequality $\alpha x \geq \beta$ satisfied by all the feasible solutions in S and such that $\alpha x^* < \beta$. The addition of the cutting plane $\alpha x \geq \beta$ to the constraints defining P leads to a tighter relaxation $P' = P \cap \{x \in \mathbb{R}^n : \alpha x \geq \beta\}$ and the process can eventually be iterated.

The disjunctive approach to the separation problem, as introduced by Balas [1], considers defining an intermediate set $Q \supseteq S$ not containing x^* and separating x^* from Q . The set Q is obtained by applying to P a valid disjunction D for the set S , such as

$$D := \left\{ x \in \mathbb{R}^n : \bigvee_{h=1}^q D^h x \geq d_0^h \right\}, \quad (2)$$

where $D^h \in \mathbb{R}^{m_h \times n}$, $d_0^h \in \mathbb{R}^{m_h}$ ($h = 1, \dots, q$), and $S \subseteq D$ (i.e., any feasible solution of Equation (1) satisfies at least one of the conditions of D). Thus, the set Q , denoted as the *disjunctive hull* of D , is defined as $Q := \text{conv}(P \cap D)$, and any valid inequality for Q is a disjunctive cutting plane for problem (1). An alternative approach consists in solving all the problems

$$\min \{cx : x \in S, D^h x \geq d_0^h\}, \quad h = 1, \dots, q, \quad (3)$$

as at least one of these yields the optimal solution to Equation (1). In this case, Equation (1) is solved by *branching* instead of *cutting*.

Since the early 1990s, the disjunctive approach has been successfully exploited both in the context of mixed integer linear programs (MILPs) as well as in that of mixed integer nonlinear programs (MINLPs). The article titled **Disjunctive Programming** gives a general overview of disjunctive programming. In the present entry we survey some applications and extensions of disjunctive programming with special emphasis on recent developments. The paper is organized as follows: In the next section we recall the basic ingredients of disjunctive inequalities for MILPs and we report on recent results on this topic. In the section titled “Disjunctive Inequalities in MINLP,” the application of disjunctive constraints both as modeling tools and cutting planes is discussed for MINLPs. Finally, in the section titled “Branching on General Disjunctions,” we

consider the application of disjunctions as branching conditions in enumerative algorithms, as opposed to the cutting approach.

DISJUNCTIVE INEQUALITIES IN MILP

In the special case of an MILP, S is defined as $S = \{x \in \mathbb{R}^n : Ax \geq b, x \geq 0, x_i \in \mathbb{Z} \forall i \in N^I\}$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $N^I \subseteq \{1, \dots, n\}$ is the set of variables constrained to be integers. In this context, the intrinsic difficulty of the problem is due to the integrality restrictions on the variables in N^I . Thus, the relaxation P that is typically considered is the polyhedron associated with S , that is, $P = \{x \in \mathbb{R}^n : Ax \geq b, x \geq 0\}$. Because

$$\begin{aligned} P \cap D &= P \cap \{x \in \mathbb{R}^n : \bigvee_{h=1}^q D^h x \geq d_0^h\} \\ &= \bigcup_{h=1}^q \left(P \cap \{x \in \mathbb{R}^n : D^h x \geq d_0^h\} \right), \end{aligned}$$

then the disjunctive hull Q (for a fixed disjunction D) is defined as a union of polyhedra. More precisely,

$$Q := \text{conv} \left(\bigcup_{h=1}^q P_h \right),$$

with $P_h := P \cap \{x \in \mathbb{R}^n : D^h x \geq d_0^h\}$. In such a case, Balas [1,2] has shown that Q is a polyhedron as well. Even if a full description of Q in the space of the x variables may require an exponential number of constraints, the key result of Balas [1,2] is that Q has a compact representation in a higher-dimensional space. Namely, there exists a polyhedron $\tilde{Q} := \{(x, y) \in \mathbb{R}^{n+p} : Bx + Cy \geq d\}$ whose projection onto the x -space is Q , and $\tilde{Q}$ has $O(qn)$ variables and $O(qm + \sum_{h=1}^q m_h)$ constraints. This implies that separating x^* from Q can be solved by linear programming.

Indeed, each polyhedron P_h ($h = 1, \dots, q$) can be defined as $P_h = \{x \in \mathbb{R}^n : A^h x \geq b^h, x \geq 0\}$, where

$$A^h = \begin{bmatrix} A \\ D^h \end{bmatrix} \text{ and } b^h = \begin{bmatrix} b \\ d_0^h \end{bmatrix}.$$

Another key result of Balas [2] is that all the inequalities $\alpha x \geq \beta$ valid for Q are described by the polyhedral cone

$$\begin{aligned} Q^\# &= \left\{ (\alpha, \beta) \in \mathbb{R}^{n+1} : \alpha \geq u^h A^h, \beta \leq u^h b^h \right. \\ &\quad \left. \text{for some } u^h \geq 0, h = 1, \dots, q \right\}. \end{aligned}$$

Furthermore, for a full-dimensional polyhedron Q there is a one-to-one correspondence among the extreme rays of $Q^\#$ and the facets of Q .

Separating Disjunctive Cuts in MILP

Typically, disjunctive cuts are separated by considering a very special subset of two-term disjunctions (i.e., with $q = 2$), namely, the so-called split disjunctions of the form of Equation

$$\pi x \leq \pi_0 \quad \bigvee \quad \pi x \geq \pi_0 + 1, \quad (4)$$

with $\pi \in \mathbb{Z}^n$, $\pi_0 \in \mathbb{Z}$, $\pi_i = 0 \forall i \notin N^I$. Disjunctive cuts arising from disjunctions of the form (4) are also known as *split cuts* [3] (see **Split Cuts**).

Given a solution x^* of $P \setminus S$, a common approach for separating x^* from S is to consider an *elementary* split disjunction of the form

$$x_i \leq \pi_0 \quad \bigvee \quad x_i \geq \pi_0 + 1, \quad (5)$$

where $x_i \in N^I$, $x_i^* \notin \mathbb{Z}$, $\pi = e_i$ is the i -th unit vector, and $\pi_0 = \lfloor x_i^* \rfloor$. Thus, the disjunctive hull Q is simply defined as the union of the two polyhedra $P_0 = \{x \in P : x_i \leq \pi_0\}$ and $P_1 = \{x \in P : x_i \geq \pi_0 + 1\}$. By Farkas lemma, a most-violated disjunctive cut $\alpha x \geq \beta$ valid for Q , with $(\alpha, \beta) \in Q^\#$, can be found by solving the so-called *cut-generating linear program* (CGLP) that determines the Farkas multipliers (u, u_0, v, v_0) associated with the inequalities defining P_0 and P_1 so as to maximize

¹For 0–1 mixed integer programs, split cuts arising from elementary disjunctions of the form (7) are denoted as *lift-and-project* cuts [4] (see **Lift-and-Project Inequalities**).

the violation of the resulting cut with respect to x^* :

$$\begin{aligned}
 (\text{CGLP}) \quad & \min \quad \alpha x^* - \beta \\
 & \alpha \geq uA - u_0 e_i, \quad \beta \leq ub - u_0 \pi_0, \\
 & \alpha \geq vA + v_0 e_i, \quad \beta \leq vb + v_0(\pi_0 + 1), \\
 & u, v, u_0, v_0 \geq 0.
 \end{aligned} \tag{6}$$

Once a violated cut has been found as a solution of Equation (6), the cut can be easily strengthened *a posteriori* through the Balas and Jeroslow [5] procedure. Such a strengthening can be seen as finding the best split disjunction for a given set of Farkas multipliers, that is, the split disjunction that yields the strongest possible disjunctive cut over the nonnegative orthant $\{x \in \mathbb{R}_+^n\}$.

By construction, any feasible CGLP solution with negative objective function value corresponds to a violated disjunctive cut. However, the feasible CGLP set is a cone and needs to be truncated by means of a suitable normalization condition, so as to produce a bounded linear program (LP) in case a violated cut exists. Different normalization conditions might lead to completely different results in terms of the *strength* of the separated cuts because they heavily affect the choice of the optimal CGLP solution.

Several normalizations have been proposed in the literature, see, e.g., Balas *et al.* [4,6], Ceria and Soares [7], Balas and Perregaard [8], and Fischetti *et al.* [9]. The most natural and simple normalization is

$$u_0 + v_0 = 1, \tag{7}$$

and takes into account only the dual multipliers of the disjunctive constraints. Such a normalization has been considered, for instance, by Balas and Saxena [10] to optimize over the first *split closure*, that is, the polyhedron obtained by adding to P all split cuts which could be derived from the original set of linear constraints (see **Basic Polyhedral Theory** and **Split Cuts**). One of the most widely used (and effective) truncation conditions reads instead

$$eu + ev + u_0 + v_0 = 1, \tag{8}$$

where $e = (1, \dots, 1)$. This latter condition was proposed by Balas [11] and investigated, for instance, by Ceria and Soares [7] and Balas and Perregaard [8,12].

Recently, Fischetti *et al.* [9] analyzed computationally the above two normalization conditions, thus showing that the truncation condition (8) generally outperforms condition (7). Indeed, normalization in Equation (8) naturally enforces the separation of *sparse* and *low rank* inequalities that turn out to be more effective in terms of the percentage of the integrality gap closed.

Unfortunately, any normalization can in some cases produce weak cuts. The projection of the disjunctive cone of Equation (6) onto the polar space (α, β) yields precisely the cone $Q^\#$. As discussed, for a full-dimensional polyhedron Q , there is a one-to-one correspondence among the extreme rays of $Q^\#$ and the facets of Q . However, this correspondence is lost in the cone of Equation (6), that is defined in a lifted space that explicitly involves the Farkas multipliers (u, v, u_0, v_0) . Indeed, there are many extreme rays of the CGLP cone that correspond to dominated cuts, and this property is independent of the normalization used to truncate the cone. Fischetti *et al.* [9] gave a theoretical characterization of *weak rays* of the disjunctive cone of Equation (6) that lead to dominated cuts, and showed how the presence of redundant constraints in the formulation of P_0 and P_1 increases the number of weak rays in Equation (6), thus affecting the quality of the generated disjunctive cuts. It remains an open question how to find a computationally efficient way of dealing with redundant constraints in the separation.

A key step in the separation of disjunctive cuts for MILPs is the work by Balas and Perregaard [12]. Even if disjunctive cuts can be separated by linear programming, solving the CGLP in the lifted space (u, v, u_0, v_0) may be in general, too time consuming. Addressing the normalization of Equation (8), Balas and Perregaard [12] provided a precise correspondence between the bases of the CGLP truncated with Equation (8) and the bases of the original LP $Ax - Is = b, x \geq 0, s \geq 0$. Then, they developed an elegant and efficient way of solving the CGLP by performing pivot

operations in the original tableau involving (x, s) variables only.

In particular, Balas and Perregaard [12] also showed that the *mixed integer gomory* (MIG) cut [13] associated with the row of the basic variable x_i , $i \in N^I$, in the optimal simplex table of the system $Ax - Is = b, x \geq 0, s \geq 0$, corresponds to a basic solution of the CGLP truncated with Equation (8). Thus, solving the CGLP with normalization of Equation (8) can be interpreted as a way of strengthening the MIG cut from the table. Indeed, the basic CGLP solution yielding the MIG cut is precisely an *optimal* solution if the CGLP is truncated with the normalization of Equation (7) [4,8,9,12]. Hence, the strengthening of the MIG cut lies in the better behavior of the normalization of Equation (8) with respect to Equation (7).

The work by Balas and Perregaard [12] has been recently extended by Balas and Bonami [14]. In Ref. 14, the authors considered different normalization conditions of the form

$$\lambda u + \lambda v + u_0 + v_0 = \lambda_0,$$

with $\lambda \in \mathbb{R}_+^m$, $\lambda_0 \in \mathbb{R}_+$, and developed a procedure that iteratively combines the “pivoting mechanism” of Ref. 12 for finding the best set of Farkas multipliers for a fixed disjunction, with the Balas and Jeroslow [5] strengthening of the disjunction for a given set of multipliers. The results presented in Ref. 14 show that disjunctive inequalities can be profitably used to improve on the performance of general purpose branch-and-cut algorithms.

DISJUNCTIVE INEQUALITIES IN MINLP

Disjunctions are essential for modeling several types of nonconvex constraints, including those arising in MINLP. For these problems, S is defined as

$$S = \left\{ x \in \mathbb{R}^n : g_j(x) \leq 0 \ \forall j \in M, x_i \in \mathbb{Z} \ \forall i \in N^I \right\},$$

where, for all $j \in M$, $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ is a multivariate (possibly nonconvex) function. When all of the g_j are affine, we have an MILP, which is therefore a subclass of MINLP.

While the integrality of a subset of variables represents an important class of nonconvex constraints, there exist other nonconvexities in MINLP that can give rise to disjunctions, and that can be used by branch-and-bound algorithms. Below, we describe a few examples of how disjunctions are used to generate disjunctive inequalities for the general MINLP case and for some special cases.

General Nonconvex MINLP

Despite the more general nonlinear setting, most of the literature in MINLP and global optimization consider linear disjunctions of the form (2). MINLP solvers address nonconvex nonlinear constraints by means, for example, of *spatial* disjunctions, which are of the form $x_i \leq \pi_0 \vee x_i \geq \pi_0$.

One of the lower bounding procedures for nonconvex MINLP consists of *reformulating* the set of constraints into a set of nonconvex equality constraints of the form $x_k = f_k(x_1, x_2, \dots, x_{k-1})$, with $f_k : \mathbb{R}^{k-1} \rightarrow \mathbb{R}$ nonlinear [15,16]. A lower bound can then be obtained by solving an LP relaxation of the reformulation, with $\bar{n} \geq n$ variables: $P_{LP} = \{x \in \mathbb{R}^{\bar{n}} : Ax \geq b\}$. P_{LP} is constructed by adding, for each constraint $x_k = f_k(x_1, x_2, \dots, x_{k-1})$, a system of linear constraints $A^k x \geq b^k$ which constitutes a relaxation of the set $\{x \in \mathbb{R}^k : x_k = f_k(x_1, x_2, \dots, x_{k-1})\}$. An example of this step is shown in Fig. 1(a) for the constraint $x_2 = f_2(x_1) = x_1^2$, where the inequalities of $A^k x \geq b^k$ delimit the shaded area.

Let x^* be the optimal solution to P_{LP} . If x^* satisfies integrality constraints and all nonlinear constraints $x_k = f_k(x_1, x_2, \dots, x_{k-1})$, the problem is solved. If it satisfies all integrality constraints while violating at least one of the nonlinear ones, a disjunction is sought that is violated by x^* . The simple disjunction $x_i \leq \pi_0 \vee x_i \geq \pi_0$, although valid for MINLP, is clearly not violated. However, the linear relaxations P'_{LP} and P''_{LP} of $S' = S \cap \{x \in \mathbb{R}^n : x_i \leq \pi_0\}$ and $S'' = S \cap \{x \in \mathbb{R}^n : x_i \geq \pi_0\}$, respectively, depicted in Fig. 1(b), contain new linear constraints that can be used to construct a disjunction of type (2) that is violated by x^* .

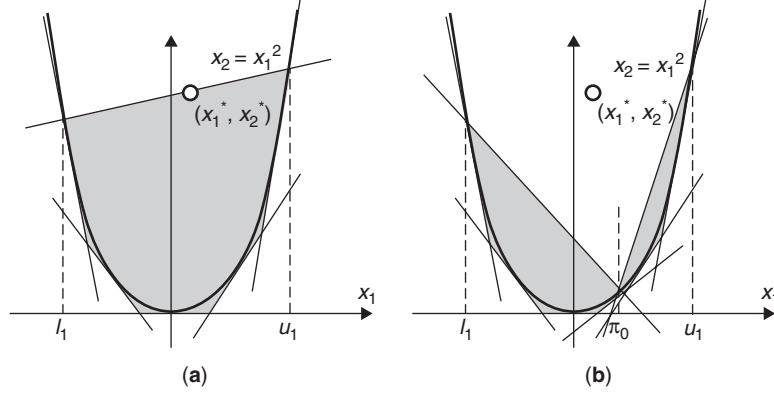


Figure 1. An MINLP disjunction. In (a) an LP relaxation of a nonconvex constraint, the shaded area is the LP relaxation of the constraint $x_2 = f_2(x_1) = (x_1)^2$ with $x_1 \in [\ell_1, u_1]$, whereas (x_1^*, x_2^*) is the value of x_1 and x_2 in the optimum of the LP relaxation. In (b) a simple MINLP disjunction, a disjunction $x_1 \leq \pi_0 \vee x_1 \geq \pi_0$, albeit not violated by (x_1^*, x_2^*) , can be used to generate two LP relaxations (the two smaller shaded areas) which in turn allow to construct a disjunction $D^1x \geq d^1 \vee D^2x \geq d^2$ violated by x^* .

Since a disjunctive inequality for $P'_{LP} \cup P''_{LP}$ is also valid for $S' \cup S''$, a procedure similar to that used in MILP, where the CGLP is solved using the LP relaxation, can be applied [17]. Zhu *et al.* [18] proposed a simple extension to MINLP where the CGLP is solved for disjunctions on binary variables.

Special Classes of MINLP

Specific cutting-plane procedures can be devised for MINLPs with a certain structure. For MINLP with binary variables and whose continuous relaxation is convex, Stubbs and Mehrotra [19] generalized the procedure proposed by Balas *et al.* [4] and described a separation procedure based on a convex optimization problem.

A similar (specialized) procedure has been successfully applied to mixed integer conic programming (MICEP) problems, which can be thought of as MILP amended by a set of conic constraints. The second-order cone and the cone of symmetric semidefinite matrices are among the most important classes of conic constraints in this class. Çezik and Iyengar [20] and later Drewes [21] proposed, for MICEP where all disjunctions are generated from binary variables, an application of the separation procedure for disjunctive cuts to the conic case, whereby inequalities are

obtained by solving a continuous conic optimization problem. Analogously, Frangioni and Gentile [22] describe a special case of disjunctive inequalities called *perspective cuts*. These are obtained from the perspective relaxation of a convex MINLP containing binary variables used to model semicontinuous variables, as discussed by Aktürk *et al.* [23] and Linderoth and Günlük [24].

Mathematical Programs with Equilibrium Constraints (MPEC). These MINLPs contain nonconvex constraints $x^\top y = 0$, with $x \in \mathbb{R}_+^k$, $y \in \mathbb{R}_+^k$. These can be more easily stated as $x_i y_i = 0$ for all $i = 1, 2, \dots, k$, and give rise to simple disjunctions $x_i = 0 \vee y_i = 0$. Júdice *et al.* [25] studied mathematical programming with equilibrium constraints (MPEC) where the only nonlinear constraints are the complementarity constraints. Relaxing the latter yields an LP. Disjunctive cuts are generated from solutions of the LP that violate a complementarity constraint (i.e., $x_i > 0$ and $y_i > 0$) through the observation that both variables are basic and by applying standard disjunctive arguments to the corresponding tableau rows.

Disjunctive cuts have been proposed by Audet *et al.* [26] in the context of linear bilevel programming, where the optimality of a *lower level* subproblem is required for

a solution to be feasible for the *upper level* problem, and is enforced through complementarity constraints.

Quadratically Constrained Quadratic Programming (QCQP). Saxena *et al.* [27,28] proposed two classes of disjunctive cuts for quadratically constrained quadratic programming (QCQP) problems, which contain constraints of the type $x^\top Qx + b^\top x + c \leq 0$, that is, nonconvex because in general $Q \not\succeq 0$. These problems are reformulated as LPs with an extra nonconvex constraint of the form $Y = xx^\top$, where Y is an $n \times n$ matrix of auxiliary variables. Relaxing these constraints to $Y - xx^\top \succeq 0$, thereby allowing solutions such that $xx^\top - Y \succeq 0$, leads to a (convex) semidefinite program, which yields good lower bounds [29,30].

Saxena *et al.* [27,28] used the nonconvex constraint $xx^\top - Y \succeq 0$, which is relaxed in the above relaxation, to obtain disjunctive cuts. Disjunctions are derived from the nonconvex constraint $(v^\top x)^2 \geq v^\top Y v$, where vector v is obtained from the negative eigenvalues of the matrix $\bar{x}\bar{x}^\top - \bar{Y}$, and $(\bar{x}, \bar{Y})$ is a solution of the relaxation. In the second paper [27], this procedure is refined to generate cuts for the nonreformulated problem, thus avoiding the need of the auxiliary matrix variable Y .

Generalized Disjunctive Programming

Generalized disjunctive programming (GDP) is an extension of disjunctive programming [1] originally proposed in Ref. 31 as a modeling framework (with a branch-and-bound based general-purpose solution algorithm) targeting problems in chemical engineering and process synthesis. GDP explicitly formulates conditional constraints via Boolean variables and logic formulæ. This emphasizes the role that logic-based transformations (such as De Morgan’s laws or passing from conjunctive to disjunctive normal forms) have on the formulation. Furthermore, it allows the specification of different lower bounding problems.

The GDP is formulated as follows:

$$\min f(x) + \sum_{k \in K} c_k, \quad (9)$$

$$g(x) \leq 0, \quad (10)$$

$$\forall k \in K \quad \bigvee_{i \in D_k} (Y_{ik} \wedge c_k = \gamma_{ik} \wedge r_{ik}(x) \leq 0), \quad (11)$$

$$\Omega(Y) \wedge x \in X \subseteq \mathbb{R}^n \wedge c \in \mathbb{R}^{|K|}, \quad (12)$$

where K and D_k ($k \in K$) are sets of indices, γ_{ik} is a known parameter for all $k \in K, i \in D_k$, Y are Boolean and x, c are continuous decision variables, Ω is a conjunctive normal form (CNF) logic formula having free variables Y and containing the clauses $\bigvee_{i \in D_k} Y_{ik}$ for all $k \in K$, with $\bigvee$ denoting “exclusive or.” The formula $\Omega(Y)$ acts as a constraint in the sense that it should be true in order for Y to be a feasible solution of Equations (9)–(12).

It is clear that Equations (9)–(12) can be reformulated to an MINLP by replacing $c_k = \gamma_{ik}$ by $Y_{ik}(c_k - \gamma_{ik}) = 0$ and $r_{ik}(x) \leq 0$ by $Y_{ik}r_{ik}(x) \leq 0$, and writing $\Omega(Y)$ as the corresponding Boolean expression in the $\{0, 1\}$ -variables Y . When X is a vector of nonempty intervals (say $[0, U]$), the products involving the Y variables can be further reformulated to “big-M” type constraints, which are convex whenever $r_{ik}(x)$ are convex. Thus, a “big-M” convex relaxation can be easily obtained from Equations (9)–(12) using well-known techniques on the possibly nonconvex functions f and g [32]. The interest of formulations (9)–(12), however, is that it can be used to yield a different convex relaxation [33] which is in general stronger than the “big-M” one, and rests on a convex-hull type reformulation of the disjunctive constraints of Equation (11):

$$\forall k \in K \quad x = \sum_{i \in D_k} v^{ik},$$

$$\forall k \in K \quad c_k = \sum_{i \in D_k} \lambda_{ik} \gamma_{ik},$$

$$\sum_{i \in D_k} \lambda_{ik} = 1,$$

$$\begin{aligned}
\forall k \in K, i \in D_k \quad \lambda_{ik} r_{ik}(v^{ik}/\lambda_{ik}) &\leq 0; \\
\forall k \in K, i \in D_k \quad 0 \leq v^{ik} &\leq \lambda_{ik} U; \\
\forall k \in K, i \in D_k \quad \lambda_{ik} &\in [0, 1].
\end{aligned}$$

BRANCHING ON GENERAL DISJUNCTIONS

Within the context of a branch-and-bound algorithm [34], a key decision that has to be repeatedly taken is how to divide a problem into subproblems. Indeed, branch-and-bound makes an implicit use of the concept of disjunction: whenever we are not able to solve a problem $\mathcal{P}$ of the form of Equation (1), we choose a disjunction D of the feasible region of $\mathcal{P}$ (i.e., a disjunction D of the form of Equation (2) valid for S), and we divide $\mathcal{P}$ into two or more subproblems $\mathcal{P}_1, \dots, \mathcal{P}_q$ corresponding to the q terms of the disjunction; this is equivalent to solving the subproblems of Equation (3). Hence, we are guaranteed that the optimal solution of one of the q subproblems coincides with the optimum of $\mathcal{P}$. In this case, we say that we *branch on the disjunction* D . Clearly, this process can be iterated. The choice of the disjunction D at each branching step determines the effectiveness of solving Equation (1) with this approach, that is, the number of subproblems (Eq. 3) that need to be solved to find (and certify) the optimal solution of Equation (1). Of course, we are interested in finding disjunctions that generate subproblems (Eq. 3) which are as easy to solve as possible.

In MILP, branching is typically done on two-term disjunctions, and in particular, the standard method is to branch on a single variable, that is, an *elementary* disjunction. Let x^* be the solution to the LP relaxation of the current problem $\mathcal{P}$, and let $i \in N^I$ such that $x_i^* \notin \mathbb{Z}$. Then, we branch on the variable x_i by dividing $\mathcal{P}$ into $\mathcal{P}_0$ and $\mathcal{P}_1$, where $\mathcal{P}_0$ is equal to $\mathcal{P}$ with the addition of the constraint $x_i \leq \lfloor x_i^* \rfloor$, and $\mathcal{P}_1$ is equal to $\mathcal{P}$ with the addition of the constraint $x_i \geq \lceil x_i^* \rceil$. Choosing which variable should be branched on at each step is of fundamental importance for the performance of branch-and-bound. We refer to Achterberg *et al.* [35] for a recent survey on this topic.

Even though branching on single variables is the method currently implemented

by solvers for integer programs (both linear and nonlinear), it need not be so. For MILPs, branching can be done on any split disjunction of the form of Equation (6). By integrality, every feasible solution to $\mathcal{P}$ satisfies any split disjunction. In the branching literature, every disjunction which is not elementary is labeled as a *general* disjunction. Several approaches to branch on general disjunctions have been proposed in the MILP literature. These approaches can be ascribed to one of two categories. The first category contains methods that try to identify “thin” directions of the polyhedron P associated with the LP relaxation of the MILP; the second category focuses on improving as much as possible, the LP bound at the children nodes. The concept of thin direction requires the concept of *width* of a full-dimensional polyhedron P along a direction u , which is defined as $\max_{x,y \in P}(ux - uy)$. Thus, for a *pure* integer program associated with P , the *integer width* is defined as

$$\min_{\pi \in \mathbb{Z}^n \setminus \{0\}} \max_{x,y \in P}(\pi x - \pi y).$$

This definition naturally extends to the mixed integer linear case by considering integer directions $\pi \in \mathbb{Z}^n \setminus \{0\}$ with $\pi_j = 0$ for $j \notin N^I$. Mahajan and Ralphs [36] discussed the formulation of optimization models to select both thin directions and split disjunctions that yield maximum bound improvement. Note that both the problems of finding a disjunction with smallest integer width or largest bound improvement are strongly NP-hard [37].

Branching on Thin Directions

Branching on thin directions of the polyhedron P associated with the LP relaxation of the MILP is a method that derives from the work of Lenstra [38] on solving integer programs in fixed dimension in polynomial time [39,40]. The idea is as follows: First, some thin directions of P are computed, using the lattice basis reduction algorithm by Lenstra *et al.* [41]. Then, the space is transformed so that these directions correspond to unit vectors, and the problem is solved by branch-and-bound in the new space. Branching on simple disjunctions in this space translates

back to branching on general disjunctions in the original space. This method has proven successful for some particular instances where standard branch-and-bound fails because of the huge size of the enumeration tree: Aardal *et al.* [42] discussed the solution of the difficult market split instances [43], while Krishnamoorthy and Pataki studied decomposable knapsack problems [44]; see also the study by Mehrotra and Li [45].

Branching for Maximum Bound Improvement

Another line of research which has been pursued is that of selecting a good general disjunction for branching at each node of the branch-and-bound tree, in order to improve as much as possible the bound at the children nodes. Owen and Mehrotra [46] proposed branching on split disjunctions with coefficients in $\{-1, 0, 1\}$ on the integer variables with fractional values at the current node. They generate all possible such disjunctions, and evaluate them using strong branching [35], in order to select the one that gives the largest improvement of the dual bound. Karamanov and Cornuéjols [47] provided a computational study of branching on the split disjunctions that define the mixed integer gomory cuts [13] associated with the optimal simplex tableau of the current branch-and-bound node. Several such disjunctions are generated, using the quality of the underlying cut as a proxy for the strength of the disjunction, and strong branching is used to select one. A similar approach was proposed by Cornuéjols *et al.* [48], but instead of considering the split disjunctions associated with MIG cuts that can be read directly from the optimal simplex table, an improvement step is applied first: the rows of the simplex table are linearly combined with integer coefficients in order to reduce the coefficients of the resulting combination. This step corresponds to tilting the underlying disjunctions so as to produce stronger MIG cuts (see also the reduce-and-split algorithm by Andersen *et al.* [49]).

REFERENCES

1. Balas E. Disjunctive programming. In: Hammer PL, Johnson EL, Korte BH, editors. *Annals of discrete mathematics* 5: discrete optimization. Amsterdam: North-Holland Publishing Co.; 1979; pp. 3–51.
2. Balas E. Disjunctive programming: properties of the convex hull of feasible points. *Discrete Appl Math* 1998;89:3–44.
3. Cook W, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program* 1990;47:155–174.
4. Balas E, Ceria S, Cornuéjols G. A lift-and-project cutting plane algorithm for mixed 0–1 programs. *Math Program* 1993;58:295–324.
5. Balas E, Jeroslow R. Strengthening cuts for mixed integer programs. *Eur J Oper Res* 1980; 4:224–234.
6. Balas E, Ceria S, Cornuéjols G. Mixed 0–1 programming by lift-and-project in a branch-and-cut framework. *Manage Sci* 1996; 42(9):1229–1246.
7. Ceria S, Soares J. Disjunctive cuts for mixed 0–1 programming: duality and lifting. Technical report, GSB. New York: Columbia University; 1997.
8. Balas E, Perregaard M. Lift-and-project for mixed 0–1 programming: recent progress. *Discrete Appl Math* 2002;123:129–154.
9. Fischetti M, Lodi A, Tramontani A. On the separation of disjunctive cuts. *Math Program, Ser A* 2009. DOI: 10.1007/s10107-009-0300-y.
10. Balas E, Saxena A. Optimizing over the split closure. *Math Program, Ser A* 2008; 113(2):219–240.
11. Balas E. A modified lift-and-project procedure. *Math Program* 1997;79:19–31.
12. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0–1 programming. *Math Program* 2003;94(2-3):221–245.
13. Gomory RE. An algorithm for the mixed integer problem. Technical Report RM-2597. Santa Monica (CA): The RAND Corporation; 1960.
14. Balas E, Bonami P. Generating lift-and-project cuts from the LP simplex tableau: open source implementation and testing of new variants. *Math Program Comput* 2009;1:165–199.
15. Smith EMB, Pantelides CC. A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Comput Chem Eng* 1999;23:457–478.
16. Tawarmalani M, Sahinidis NV. Convexification and global optimization in continuous

- and mixed-integer nonlinear programming: theory, algorithms, software and applications, Volume 65, Nonconvex optimization and its applications. Dordrecht: Kluwer Academic Publishers; 2002.
17. Belotti P. Disjunctive cuts for non-convex MINLP. Technical report. Bethlehem (PA): Lehigh University; 2009.
18. Zhu Y, Hu Y, Wu H, *et al.* An improved branch-and-cut algorithm for mixed-integer nonlinear systems optimization problem. *AIChE J* 2008;54(12):3239–3247.
19. Stubbs RA, Mehrotra S. A branch-and-cut method for 0–1 mixed convex programming. *Math Program* 1999;86:515.
20. Çezik MT, Iyengar G. Cuts for mixed 0–1 conic programming. *Math Program, Ser A* 2005;104:179–200.
21. Drewes S. Mixed integer second order cone programming [PhD thesis]. Technische Universität Darmstadt; 2009.
22. Frangioni A, Gentile C. Perspective cuts for a class of convex 0–1 mixed integer programs. *Math Program* 2006;106(2):225–236.
23. Aktürk M, Atamtürk A, Gürel S. A strong conic quadratic reformulation for machine-job assignment with controllable processing times. *Oper Res Lett* 2009;37(3):187–191.
24. Günlük O, Linderoth J. Perspective relaxation of mixed integer nonlinear programs with indicator variables. In: Lodi A, Panconesi A, Rinaldi G, editors. Volume 5035, Proceedings of the 13th Integer Programming and Combinatorial Optimization Conference, Lecture Notes in Computer Science. Berlin: Springer; 2008; pp. 1–16.
25. Júdice JJ, Sherali HD, Ribeiro IM, *et al.* A complementarity-based partitioning and disjunctive cut algorithm for mathematical programming problems with equilibrium constraints. *J Glob Optim* 2006;36:89–114.
26. Audet C, Haddad J, Savard G. Disjunctive cuts for continuous linear bilevel programming. *Optim Lett* 2007;1(3):259–267.
27. Saxena A, Bonami P, Lee J. Convex relaxations of non-convex mixed integer quadratically constrained programs: Projected formulations. *Math Program* 2010. In press.
28. Saxena A, Bonami P, Lee J. Disjunctive cuts for non-convex mixed integer quadratically constrained programs. In: Lodi A, Panconesi A, Rinaldi G, editors. Volume 5035, Proceedings of the 13th Integer Programming and Combinatorial Optimization Conference, Lecture Notes in Computer Science. 2008. pp. 17–33.
29. Helmberg C, Rendl F. Solving quadratic (0,1)-problems by semidefinite programs and cutting planes. *Math Program* 1998;82(3):291–315.
30. Rendl F, Sotirov R. Bounds for the quadratic assignment problem using the bundle method. *Math Program* 2007;109(2-3):505–524.
31. Raman R, Grossmann I. Modelling and computational techniques for logic-based integer programming. *Comput Chem Eng* 1994;18(7):563–578.
32. Belotti P. Couenne: a user's manual. Technical report. Bethlehem (PA): Lehigh University; 2009.
33. Grossmann I, Lee S. Generalized convex disjunctive programming: nonlinear convex hull relaxation. *Comput Optim Appl* 2003;26:83–100.
34. Land AH, Doig AG. An automatic method of solving discrete programming problems. *Econometrica* 1960;28(3):497–520.
35. Achterberg T, Koch T, Martin A. Branching rules revisited. *Oper Res Lett* 2005;33(1):42–54.
36. Mahajan A, Ralphs TK. Experiments with branching using general disjunctions. Proceedings of the 11th INFORMS Computing Society Meeting; Charleston (SC); 2009.
37. Mahajan A, Ralphs TK. On the complexity of branching on general hyperplanes for integer programming. *SIAM J Optim* 2009;20(5):2181–2198.
38. Lenstra HW Jr. Integer programming with a fixed number of variables. *Math Oper Res* 1983;8(4):538–548.
39. Grötschel M, Lovász L, Schrijver A. Progress in combinatorial optimization. In: Geometric methods in combinatorial optimization. Toronto: Academic Press; 1984. pp. 167–183.
40. Lovász L, Scarf HE. The generalized basis reduction algorithm. *Math Oper Res* 1992;17(3):751–764.
41. Lenstra AK, Lenstra HW Jr, Lovász L. Factoring polynomials with rational coefficients. *Math Ann* 1982;4(261):515–534.
42. Aardal K, Bixby RE, Hurkens CAJ, *et al.* Market split and basis reduction: towards a solution of the Cornuéjols–Dawande instances. *INFORMS J Comput* 2000;12(3):192–202.
43. Cornuéjols G, Dawande M. A class of hard small 0-1 programs. In: Bixby RE, Boyd EA, editors. Volume 1412, Proceedings of the 6th IPCO Conference, Lecture Notes in Computer Science. Berlin: Springer; 1998. pp. 284–293.

- 44. Krishnamoorthy B, Pataki G. Column basis reduction and decomposable knapsack problems. *Discrete Optim* 2009;6(3):242–270.
- 45. Mehrotra S Li Z. On generalized branching method for mixed integer programming. Technical report. Evanston (IL): Northwestern University; 2004.
- 46. Owen J, Mehrotra S. Experimental results on using general disjunctions in branch-and-bound for general-integer linear program. *Comput Optim Appl* 2001;20:159–170.
- 47. Karamanov M, Cornuéjols G. Branching on general disjunctions. *Math Program, Ser A* 2009. DOI: 10.1007/s10107-009-0332-3.
- 48. Cornuéjols G, Liberti L, Nannicini G. Improved strategies for branching on general disjunctions. *Math Program, Ser A* 2009. DOI: 10.1007/s10107-009-0333-2.
- 49. Andersen K, Cornuéjols G, Li Y. Reduce-and-split cuts: improving the performance of mixed integer Gomory cuts. *Manag Sci* 2005;51(11):1720–1732.

DISJUNCTIVE PROGRAMMING

EGON BALAS

Tepper School of Business,
Carnegie Mellon University,
Pittsburgh, Pennsylvania

Disjunctive programming is optimization over unions of polyhedra [1–3]. While polyhedra are convex, their unions are obviously not. To put it differently, disjunctive programming is optimization over disjunctive sets. A disjunctive set is a set defined by linear inequalities connected by conjunction ($\wedge$, “and,” juxtaposition), disjunction ($\vee$, “or”), or negation ($\neg$, “complement of”). The name reflects the fact that of the above three operations, it is the disjunction that makes the sets defined in this way nonconvex. Pure and mixed integer programs, in particular pure and mixed 0–1 programs, can be viewed as disjunctive programs but the same is true of a host of other problems; for instance, the linear complementarity problem. The main application of disjunctive programming has, so far, been in 0–1 programming, where it has served as the source of a rich family of cutting planes and in the early 1990s has produced a computationally successful variant known as lift-and-project cuts.

The basic concepts of disjunctive programming can be expressed using linear inequalities $a^i x \geq b_i$ and the halfspaces H_i^+ defined by them as building blocks. Thus, the intersection of a finite number of halfspaces, $P = \cap_{i \in M} H_i^+$, is a polyhedron; while the union of a finite number of halfspaces, $D = \cup_{i \in M} H_i^+$, is an *elementary* disjunctive set. A disjunctive set can be stated in many equivalent forms, of which the two extremes are the *conjunctive* normal form and the *disjunctive* normal form. The first one is a conjunction, or intersection, of elementary disjunctive sets $\cap_{i \in T} D_i$, while the second one is a disjunction, or union, of polyhedra $\cup_{i \in Q} P_i$. Since any disjunctive set can be brought to any of these extreme forms by a finite number of simple operations [4],

disjunctive programming can be defined as optimization over unions of polyhedra.

The two central results of disjunctive programming can be stated succinctly as follows:

1. There is a compact representation of the convex hull of a union of polyhedra in a higher-dimensional space, which in turn can be projected back into the original space. The first step of this operation, that is the higher-dimensional representation, may be viewed as lifting (or extended formulation), while the second step is projection. As a result, one obtains the convex hull in the original space.
2. A large class of disjunctive sets, called *facial*, among them mixed 0–1 programs, can be convexified sequentially; that is, their convex hull can be derived by imposing the elementary disjunctions of their conjunctive normal form one at a time, generating the convex hull of the current set each time.

Beyond these specific results, the basic idea of disjunctive programming that the reformulation of a given problem in a higher dimensional space can often lead to a more advantageous problem structure, has turned out to be a very fruitful one with countless applications in combinatorial optimization [5–7].

ORIGINS

Historically and conceptually, disjunctive programming originates in the work around 1970 on intersection cuts for mixed integer programs [8–10]. Suppose the linear programming relaxation of a mixed integer program has an optimal basic solution expressed in the form

$$x = \bar{x} - \sum_{j \in J} \bar{a}_j s_j, \quad (1)$$

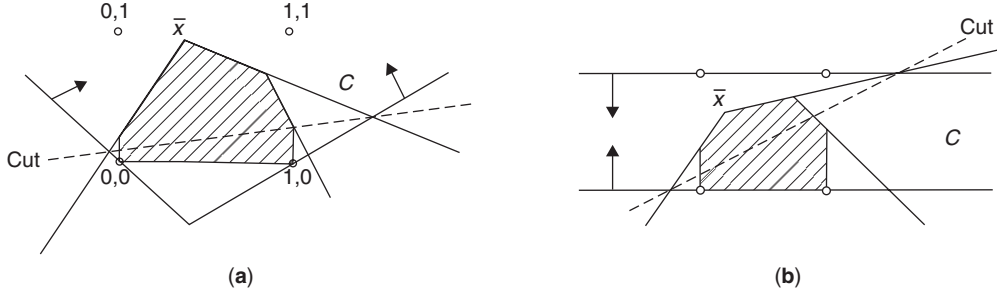


Figure 1. Intersection cut from (a) cone C , (b) strip C .

where $\bar{x}$ is the (fractional) solution in question; $s_j \geq 0, j \in J$, are the nonbasic variables; and we are interested in generating a valid cutting plane (i.e., an inequality violated by $\bar{x}$, but satisfied by all feasible integer points). One way of doing this is by choosing a suitable closed convex set C , namely one whose interior contains $\bar{x}$ but no feasible integer point and intersecting the boundary of C with the n halfplanes $\bar{x} - \bar{a}_j s_j, s_j \geq 0, j \in J$. The hyperplane through the n intersection points $s_j^*, j \in J$, then defines a valid cutting plane, the *intersection cut*. In the space of the s -variables, it has the form $\alpha s \geq 1$, with $\alpha_j = \frac{1}{s_j^*}$, where $s_j^* = \max\{s_j : \bar{x} - \bar{a}_j s_j \in C\}$. Figure 1 (a) and (b) illustrate intersection cuts. In both cases, the convex set C consists of the intersection of two halfspaces, but in (a) the intersection of the halfspaces is a cone, whereas in (b) the halfspaces are defined by hyperplanes parallel to each other and so their intersection is an infinite strip. The intersection cut from this latter set C is the mixed integer Gomory (MIG) cut.

The connection with disjunctive programming becomes clear if we consider an intersection cut from a polyhedral set $C := \{x : d^i x \leq d_0^i, i = 1, \dots, q\}$. Then the condition that the interior of C contain no feasible integer point, is equivalent to the condition that every feasible integer point should satisfy at least one of the weak complements of the inequalities defining C , that is should satisfy the disjunction

$$\bigvee_{i=1}^q (d^i x \geq d_0^i). \quad (2)$$

So an intersection cut from C can be viewed as a disjunctive cut from (2). While these two cuts are the same, the disjunctive point of view opens up new perspectives. Thus, suppose in addition to (2), all feasible solutions have to satisfy $Ax \geq b$. Then one way to proceed would be to generate all valid cutting planes from (2) and append them to $Ax \geq b$, resulting in the constraint set

$$P := \left\{ x \in \mathbb{R}^n : (Ax \geq b) \cap \text{conv} \left(\bigvee_{i=1}^q (d^i x \geq d_0^i) \right) \right\}.$$

But another way to proceed is to introduce $Ax \geq b$ into every term of the disjunction 2, that is replace (2) by

$$\bigvee_{i=1}^q \left(Ax \geq b \right) \quad (3)$$

and take the convex hull of this union of polyhedra

$$\bar{P} := \text{conv} \left(\bigvee_{i=1}^q \left(Ax \geq b \right) \right).$$

Now it is easy to see that $\bar{P} \subseteq P$, but the fact is that except for some special cases [4], $\bar{P} \neq P$, that is $\bar{P}$ is tighter than P .

THE CONVEX HULL OF A UNION OF POLYHEDRA

Disjunctive programming has two characterizations of the (closed) convex hull of a union of polyhedra.

Let

$$F := \bigcup_{i \in Q} P_i, P_i := \{x \in \mathbb{R}^n : A^i x \geq b^i\} \neq \emptyset, \\ \text{with } A_{m_i \times n}^i, i \in Q. \quad (4)$$

Theorem 1 [1,4].

$$\text{cl conv } F \\ = \{x \in \mathbb{R}^n : x - \sum_{i \in Q} y^i = 0, \\ A^i y^i - b^i y_0^i \geq 0, \\ y_0^i \geq 0 \quad i \in Q, \quad (5) \\ \sum_{i \in Q} y_0^i = 1, \\ \text{for some } (y^i, y_0^i) \in \mathbb{R}^{n+1}, i \in Q\}.$$

This theorem gives a polyhedral representation of the convex hull of a union of $q = |Q|$ polyhedra in a higher-dimensional space, one whose dimension is linear in q , the number of polyhedra in the union: it is thus a compact representation. In the form stated above, it assumes that the polyhedra are nonempty. However, the theorem remains valid without this assumption, provided that the following condition is satisfied: if any of the empty polyhedra has directions of infinity (i.e., solutions to the homogeneous system $A^i x \geq 0$), such directions must belong to the sum of the recession cones of the nonempty polyhedra.

In the special case of a disjunction of the form $x_j \in \{0, 1\}$, when $|Q| = 2$ and

$$P_{j0} := \{x \in \mathbb{R}_+^n : Ax \geq b, x_j = 0\}, \\ P_{j1} := \{x \in \mathbb{R}_+^n : Ax \geq b, x_j = 1\},$$

$\text{cl conv}(P_{j0} \cup P_{j1})$ is the set of those $x \in \mathbb{R}^n$ for which there exist vectors $(y, y_0), (z, z_0) \in \mathbb{R}_+^{n+1}$ such that

$$\begin{array}{rcl} x - y & - z & = 0, \\ Ay - by_0 & & \geq 0, \\ -y_j & & = 0, \\ & Az - bz_0 & \geq 0, \\ & z_j - z_0 & = 0, \\ y_0 + z_0 & & = 1, \end{array} \quad (6)$$

a system of quite manageable size.

The second description of the convex hull of a union of polyhedra is obtained by projecting the system (5) onto the subspace of the x -variables.

Theorem 2 [1,2].

$$\text{cl conv } F \\ = \{x \in \mathbb{R}^n : \alpha x \geq \beta \text{ for all } (\alpha, \beta) \in \mathbb{R}^{n+1} \\ \text{such that} \\ \alpha = u^i A^i, \beta \leq u^i b^i, i \in Q, \\ \text{for some } u^i \in \mathbb{R}^{m_i}, u^i \geq 0, i \in Q\}. \quad (7)$$

This theorem can be viewed as a Farkas lemma for disjunctive systems, or unions of polyhedra. For a single polyhedron or system of inequalities $Ax \geq b$, one version of the Farkas lemma asserts that an inequality $\alpha x \geq \beta$ is a consequence of $Ax \geq b$ if and only if there exists a vector $u \geq 0$ such that $\alpha = uA, \beta \leq ub$. Viewed in this light, Theorem 2 asserts that given a disjunctive set or union of polyhedra of the form of (4), an inequality $\alpha x \geq \beta$ is a consequence of (is implied by, is valid for) the disjunctive set (the union of polyhedra) if and only if it is a consequence of (is implied by, is valid for) each consistent member $A^i x \geq b^i$ of the union (here consistent means $P_i \neq \emptyset$).

This second characterization is directly relevant to the task of generating valid cuts for $\text{cl conv } F$, since stating it in a slightly different form, it asserts that $\alpha x \geq \beta$ is valid for F if and only if there exist vectors $u^i \in \mathbb{R}_+^{m_i}, i \in Q$, such that

$$\begin{aligned} \alpha &\geq \sup_{i \in Q} u^i A^i, \\ \beta &\leq \inf_{i \in Q} u^i b^i. \end{aligned} \quad (8)$$

In other words, all valid cuts for F are of the form of (8).

The family of all valid inequalities for F is the reverse polar cone of F , defined as $F^\# := \{(\alpha, \beta) \in \mathbb{R}^{n+1} : \alpha x \geq \beta \text{ for all } x \in F\}$. In particular, from Theorem 2 it follows that

$$F^\# = \{(\alpha, \beta) \in \mathbb{R}^{n+1} : \alpha = u^i A^i, \beta \leq u^i b^i \\ \text{for some } u^i \geq 0, i \in Q\}$$

and we have the following characterization of the facets of $\text{cl conv } F$. Let β be fixed, $\beta \neq 0$, and $F_\beta^\# = \{\alpha \in \mathbb{R}^n : (\alpha, \beta) \in F^\#\}$.

Theorem 3 [1,2]. *Let F be full-dimensional. Then the inequality $\alpha x \geq \beta$ defines a facet of $\text{cl conv } F$ if and only if α is a vertex of $F_\beta^\#$.*

Analogous results are known for the cases when F is less than full dimensional and/or $\beta = 0$.

From these results, it follows that the facets of the convex hull of F , and more generally, all valid inequalities $\alpha x \geq \beta$ for F , can be found by minimizing some properly chosen objective function over the constraint set of $F_\beta^\#$; that is, by solving a linear program of the form

$$\begin{aligned} \min \{g\alpha : \alpha & - u^i A^i = 0 \\ & u^i b^i \geq \beta \\ & u^i \geq 0, \quad i \in Q\}, \end{aligned} \quad (9)$$

for some appropriate vector $g \in \mathbb{R}^n$. It can be shown [1] that the linear program of (9) has a finite minimum if and only if $\lambda g \in \text{cl conv } F$ for some $\lambda > 0$; that is, if and only if g belongs to the cone generated by F ; and should g satisfy this condition, $\bar{\alpha}x \geq \beta$ defines a facet of $\text{cl conv } F$ (where F is assumed to be full dimensional), if and only if $\alpha = \bar{\alpha}$ for every optimal solution (α, u) of (9). This last condition is needed to guarantee that α is a vertex of $F_\beta^\#$.

FACIAL DISJUNCTIVE PROGRAMS

Given a disjunctive program in conjunctive normal form, that is as a conjunction of elementary disjunctive sets, the question arises whether it is possible to generate the convex hull of feasible points by imposing the elementary disjunctions one by one, at each step computing the convex hull of the set defined by the disjunctions imposed in the earlier steps, plus the new one. The answer is negative in general, but positive for an important class of disjunctive programs called *facial*. This class includes (pure or

mixed) 0–1 programs, but not general (pure or mixed) integer programs.

Consider a disjunctive program with constraint set

$$F := \{x \in F_0 : \bigvee_{i \in Q_j} (d^i x \geq d_0^i), j \in S\}, \quad (10)$$

where $F_0 := \{x \in \mathbb{R}^n : Ax \geq b\}$.

Then F is called *facial* if every term $d^i x \geq d_0^i$ of every elementary disjunction indexed by S defines a face of F_0 (in other words, if $F_0 \cap \{x : d^i x > d_0^i\} = \emptyset, i \in Q_j, j \in S$).

Theorem 4 [1,2]. *Let F be as in (10). For an arbitrary ordering of $S := \{1, \dots, t\}$, define recursively*

$$F_j := \text{conv} \left(F_{j-1} \cap \left\{ x : \bigvee_{i \in Q_j} (d^i x \geq d_0^i) \right\} \right), \quad j = 1, \dots, t.$$

Then $F_t = \text{conv } F$.

The most important class of facial disjunctive programs are mixed 0–1 programs and for this case, the above theorem asserts that the convex hull of

$$F := \{x \in \mathbb{R}^n : Ax \geq b, x_j \in \{0, 1\}, j = 1, \dots, p \leq n\}$$

can be obtained in p steps, through a procedure that at each step imposes one new disjunction, $x_j \leq 0 \vee x_j \geq 1$, and generates the convex hull of the resulting two-term disjunctive set

$$\{x \in F_{j-1} : x_j \leq 0\} \cup \{x \in F_{j-1} : x_j \geq 1\}.$$

With this result, disjunctive programming theory has established the most important distinguishing feature of 0–1 programs among integer programs, namely sequential convexifiability. Figure 2 illustrates the fact that general integer programs, even when the integers are bounded between 0 and 2, are not sequentially convexifiable: the convex hull of the integer points in the shaded area of Fig. 2(a) is the line joining them in Fig. 2(d), but sequential convexification ends with the set in Fig. 2(c).

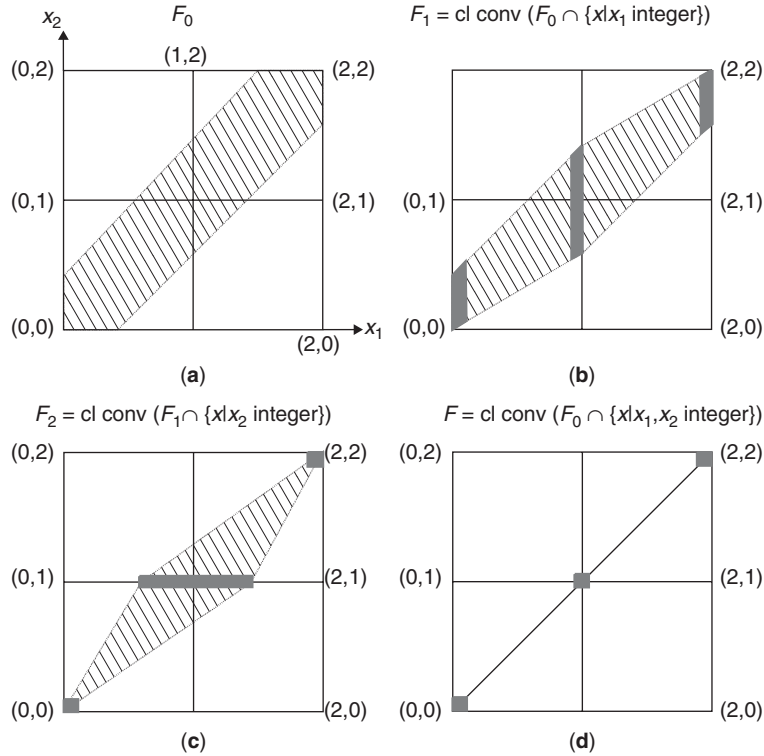


Figure 2. General integer programs are not sequentially convexifiable.

The significance of this property becomes obvious when it comes to classifying valid inequalities. A mixed integer Gomory (MIG) cut is said to be of *Chvátal rank* k if its derivation requires at least k recursive applications of integer rounding. In the same spirit, the *disjunctive rank* of a valid inequality for a disjunctive set is defined as the minimum number of recursive iterations of the abovementioned sequential convexification procedure needed to derive it. It then follows that the disjunctive rank of any valid inequality for a mixed 0–1 program can never exceed the number of 0–1 variables. This is in sharp contrast with the disjunctive rank—or for that matter, with the Chvátal rank—of a valid inequality for a general mixed integer program, for which no such bound exists.

LIFTING AND PROJECTION

A key idea of disjunctive programming is that a disjunctive program is equivalent to

a linear program in a higher-dimensional space (Theorem 1). This makes it possible to generate strong cuts for the original problem by solving a higher-dimensional linear program. Historically, the implementation of the disjunctive programming approach into a practical 0–1 programming algorithm required not only the choice of a specific version of disjunctive cuts, but also a judicious combination of cutting with branching, made possible in turn by the discovery of an efficient procedure for lifting cuts generated in a subspace (for instance, at a node of the search tree) to be valid in the full space (i.e., throughout the search tree). The particular version that turned out to be computationally successful is known as lift-and-project [11].

Suppose we wish to generate strong cuts for a mixed integer 0–1 program, whose linear programming relaxation has the constraint set $Ax \geq b$ (notice that this subsumes $x \geq 0$ and the inequalities $x_j \leq 1$ for the integer-constrained variables) and

the optimal solution $\bar{x}$. Then the linear program (9), restated for the disjunction $x_j = 0 \vee x_j = 1$, with the objective of cutting off $\bar{x}$ by a maximum amount (i.e., generating a “deepest cut”) and with its constraint set normalized, becomes the “cut generating linear program” for x_j :

$$\begin{aligned}
\min \quad & \alpha \bar{x} - \beta \\
\text{s.t.} \quad & \alpha - uA + u_0 e_j \geq 0, \\
& \alpha - vA - v_0 e_j \geq 0, \\
& -\beta + ub = 0, \\
& -\beta + vb + v_0 = 0, \\
& ue + ve + u_0 + v_0 = 1, \\
& u, v \geq 0.
\end{aligned} \tag{CGLP}_j$$

Here, $e = (1, \dots, 1)$ and e_j is the j -th unit vector.

The last equation has the purpose of truncating the cone defined by the rest of the constraint set. Several other normalizations are possible, such as $\beta \in \{1, -1\}$ or $\|\alpha\|_1 \leq 1$, with pros and cons for each one and their choice plays an important role in determining the optimum; different normalizations can yield different optimal solutions, that is different “deepest” cuts, for the same objective function.

Solving $(\text{CGLP})_j$ yields a cut $\alpha x \geq \beta$, where $\beta = ub = vb + v_0$, and

$$\alpha_k = \begin{cases} \max\{ua_k, va_k\} & k \in N \setminus \{j\}, \\ \max\{ua_j - u_0, va_j + v_0\} & k = j, \end{cases} \tag{11}$$

with N the column index set and a_k the k -th column of A .

From Theorem 2, the solution set of $(\text{CGLP})_j$ gives the coefficient vectors (α, β) of all the valid inequalities for the set $\{x \in \mathbb{R}^n : Ax \geq b, x_j \in \{0, 1\}\}$; and from Theorem 3, if we generate all the cuts corresponding to basic solutions of $(\text{CGLP})_j$, we obtain the convex hull of this set.

If we then iterate this procedure for all x_h , $h \neq j$, from Theorem 4 (on sequential convexification) we obtain the convex hull

of $\{x \in \mathbb{R}^n : Ax \geq b, x_h \in \{0, 1\}, h \in N_1\}$, where N_1 is the set of integer-constrained variables.

In a practical algorithm of course one cannot generate all the cuts and the objective is to generate a collection of “strong” cuts and make the best use of them, possibly in combination with a branch and bound procedure. The particular structure of disjunctive cuts of the type (11) has made it possible to avoid a major stumbling block on the way to combine cutting planes with branch-and-bound. A cutting plane derived at a node of the search tree defined by a subset $B_0 \cup B_1$ of the 0–1 variables, where B_0 and B_1 index those variables fixed at 0 and 1, respectively, is valid at that node and its descendants in the tree (where the variables in $B_0 \cup B_1$ remain fixed at their values). Such a cut can, in principle, be made valid at other nodes of the search tree, where the variables in $B_0 \cup B_1$ are no longer fixed, by calculating appropriate values for the coefficients of these variables—a procedure called *lifting*. However, calculating such coefficients is in general a daunting task, which may require the solution of an integer program for every coefficient. One important advantage of the cuts discussed here is that the multipliers u, u_0, v, v_0 obtained along with the cut vector (α, β) by solving $(\text{CGLP})_j$ can be used to calculate by closed form expressions the coefficients α_h of the variables $h \in B_0 \cup B_1$.

While this possibility of calculating efficiently the coefficients of variables absent from a given subproblem (i.e., fixed at certain values) is crucial for making it possible to generate cuts during a branch-and-bound process that are valid throughout the search tree, its significance goes well beyond this aspect. Indeed, most columns of A corresponding to nonbasic components of $\bar{x}$ typically play no role in determining the optimal solution of $(\text{CGLP})_j$ and can therefore be ignored. In other words, the cuts can be generated in a subspace involving only a subset of the variables, and then lifted to the full space. This is the procedure followed in Ref. 12, where the subspace used is that of the variables indexed by some $R \subset N$, such that R includes all the 0–1 variables that are fractional and all the continuous variables

that are positive at the LP optimum. The lifting coefficients for the variables not in the subspace, which are all assumed w.l.o.g. to be at their lower bound, are then given by $\alpha_\ell := \max\{ua_\ell, va_\ell\}$, $h \in N \setminus R$, where u and v are the optimal vectors obtained by solving (CGLP)_j.

The cut $\alpha x \geq \beta$ derived from a disjunction of the form $x_j \in \{0, 1\}$ can be strengthened by using the integrality conditions on variables other than x_j [2,13]. Indeed, if x_k is such a variable, the coefficient $\alpha_k = \max\{ua_k, va_k\}$ can be replaced by

$$\bar{\alpha}_k := \min\{ua_k + u_0 \lceil m_k \rceil, va_k - v_0 \lfloor m_k \rfloor\}, \quad (12)$$

where

$$m_k := \frac{va_k - ua_k}{u_0 + v_0}.$$

This strengthening procedure can also be applied to cuts derived from disjunctions other than $x_j \in \{0, 1\}$, including disjunctions with more than two terms. In the latter case, however, the closed form expression for the values m_k used above has to be replaced by a procedure for calculating those values, whose complexity is linear in the number of terms in the disjunction (see Refs 2 and 13 for details).

The strengthened lift-and-project cuts (12) have been shown to be in 1–1 correspondence with MIG cuts [14]. To be more precise, an MIG cut from the x_k row of the optimal simplex tableau is equivalent to the lift-and-project cut $\alpha x \geq \beta$ corresponding to the basic solution (α, u, v, u_0, v_0) of (CGLP)_k in which $u = v = 0$ and $u_0 = 1/a_{k0}$, $v_0 = 1/(1 - a_{k0})$. The lift-and-project cuts corresponding to other basic solutions of (CGLP)_k are equivalent to MIG cuts from the x_k row of other (possibly infeasible) simplex tableaus. The precise correspondence established in Ref. 14 between bases of the LP and those of the CGLP make it possible to mimic on the LP tableau the solution procedure of the CGLP, without explicitly formulating the latter. From this perspective, finding the “deepest” lift-and-project cut amounts to pivoting in the LP tableau in order to find a tableau (whether feasible or not) whose x_k row yields the “deepest” MIG cut [14,15].

The lift-and-project procedure with its cut-generating linear program can also be derived in an alternative way, by the following procedure:

1. Choose $x_j \in \{0, 1\}$. Multiply $Ax \geq b$ with x_j and $1 - x_j$ to obtain the system $x_j(Ax - b) \geq 0$, $(1 - x_j)(Ax - b) \geq 0$.
2. Linearize the resulting system by substituting y_i for $x_i x_j$, $i = 1, \dots, n$, and x_j for x_j^2 .
3. Project the resulting polyhedron onto the x -space.

It is not hard to see that the outcome of steps 1 and 2 is precisely the system 6; that is, the higher-dimensional representation of $\text{conv}(P_{j0} \cup P_{j1})$ known from disjunctive programming [1]. Furthermore, step 3 results in the projection of (6) onto the x -space, which is precisely the constraint set of CGLP_j; that is, the system whose solutions are the coefficient vectors of all the valid inequalities for $\{x \in \mathbb{R}^n : Ax \geq b, x_j \in \{0, 1\}\}$. It is also clear from the sequential convexification Theorem 3 that iterating this procedure for all $j \in N_1$ yields the coefficient vectors of all the inequalities describing the convex hull of $\{x \in \mathbb{R}^n : Ax \geq b, x_j \in \{0, 1\}, j \in N_1\}$.

The above three-step procedure is a streamlined version of the matrix-cone procedure of Lovász and Schrijver [16]. The latter involves in step 1 multiplication with x_j and $1 - x_j$ for every $j \in N_1$ rather than just one. While obtaining the convex hull of $\{x \in \mathbb{R}^n : Ax \geq b, x_j \in \{0, 1\}, j \in N_1\}$ still involves $|N_1|$ iterations of steps 1, 2, 3, the added computational cost brings a reward: after each iteration, the coefficient matrix of the linearized system must be positive-semidefinite, a condition that tightens the constraint set. A procedure similar to this one (without the semidefiniteness implication) was proposed by Sherali and Adams [17].

SPLIT CUTS AND OTHER DISJUNCTIVE CUTS

A two-term disjunction of the form $x_k \leq 0 \vee x_k \geq 1$ is a special case of a class called *split disjunction*:

$$\pi x \leq \pi_0 \vee \pi x \geq \pi_0 + 1, \quad (13)$$

where (π, π_0) is an arbitrary vector in $\mathbb{Z}^{n+1}$. Clearly, any $x \in \mathbb{Z}^n$ satisfies 13. Cuts derived from split disjunctions are called *split cuts*. The collection of all valid split cuts added to the linear constraint set $F_0 := \{x \in \mathbb{R}^n : Ax \geq b\}$ forms the (elementary, or rank 1) split closure of F_0 , known to be a polyhedron [18]. Iterating this procedure k times gives the rank k split closure. A given cut valid for F_0 has split rank k if it can be derived from F_0 in k , but not fewer than k , iterations. It is known [18] that there is no upper bound on the split rank of inequalities needed to define the convex hull of a mixed integer set. This finding is in sharp contrast to the fact that the split rank of any inequality defining the convex hull of a mixed 0–1 set is at most equal to the number of 0–1 variables (due to the sequential convexifiability of such sets).

A number of interesting connections between intersection cuts, split cuts, lift-and-project cuts, MIG cuts, and other mixed integer programming cuts that can be derived from two-term disjunctions, are discussed in Ref. 19. The disjunctive programming characterization of the convex hull of a union of polyhedra (Theorem 1) has been extended in Ref. 20 to unions of convex sets, and used to generate cuts for mixed integer convex programs. Disjunctive cuts for mixed integer programs with non-convex constraint sets have been derived in Ref. 21.

REFERENCES

1. Balas E. Disjunctive programming: properties of the convex hull of feasible points. *Discrete Appl Math* 1998;89:1–44.
2. Balas E. Disjunctive programming. *Ann Discrete Math* 1979;5:3–51.
3. Jeroslow RG. Cutting plane theory: disjunctive methods. *Ann Discrete Math* 1977;1:293–330.
4. Balas E. Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM J Algebr Discrete Meth* 1985;6:466–486.
5. Balas E, Pulleyblank WR. The perfectly matchable subgraph polytope of an arbitrary graph. *Combinatorica* 1989;9:321–337.
6. Ball M, Liu W, Pulleyblank WR. Two terminal steiner tree polyhedra. In: Cornet B, Tulkens H, editors. *Contributions to operations research and economics—the twentieth anniversary of CORE*. Cambridge (MA): MIT Press; 1989. pp. 251–284.
7. Pochet Y, Wolsey L. *Production planning and mixed integer programming*. New York: Springer; 2006.
8. Balas E. Intersection cuts—a new type of cutting planes for integer programming. *Oper Res* 1971;19:19–39.
9. Glover F. Convexity cuts and cut search. *Oper Res* 1973;21:123–131.
10. Young RC. Hypercylindrically deduced cuts in zero-one integer programs. *Oper Res* 1971;19:1393–1405.
11. Balas E, Ceria S, Cornuéjols G. A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math Program* 1993;58:295–324.
12. Balas E, Ceria S, Cornuéjols G. Mixed 0-1 programming by lift-and-project in a branch-and-cut framework. *Manag Sci* 1996;42:1229–1246.
13. Balas E, Jeroslow RG. Strengthening cuts for mixed integer programs. *Eur J Oper Res* 1980;4:224–234.
14. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0-1 programming. *Math Program B* 2003;94:221–245.
15. Balas E, Bonami P. Generating lift-and-project cuts from the LP simplex tableau: open source implementation and testing of new variants. *Math Program Comput* 2009;1:165–199.
16. Lovasz L, Schrijver A. Cones of matrices and set functions and 0-1 optimization. *SIAM J Optim* 1991;1:166–190.
17. Sherali H, Adams W. A hierarchy of relaxations between the continuous and convex hull representations for 0-1 programming problems. *SIAM J Discrete Math* 1990;3:311–330.
18. Cook W, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program* 1990;47:155–174.
19. Cornuéjols G, Li Y. Elementary closures for mixed integer programs. *Oper Res Lett* 2001;28:1–8.

20. Stubbs RA, Mehrotra S. A branch-and-cut method for 0-1 mixed convex programming. Math Program A 1999;86:515–532.
21. Saxena A, Bonami P, Lee J. Convex relaxations of non-convex mixed integer quadratically constrained problems: extended formulations. Math Program A 2010. Appeared in electronic form in March.

DISTRIBUTED SIMULATION IN ORMS

SIMON J.E. TAYLOR
School of Information Systems,
Computing and Mathematics,
Brunel University, Middlesex, UK

NAVONIL MUSTAFEE
School of Business and Economics,
Swansea University, Swansea,
Wales, UK

INTRODUCTION

In Operational Research and Management Science (ORMS), discrete-event simulation has been used to analyze production and logistics problems in many areas such as commerce, defense, health, manufacturing, and logistics for many years [1]. The first discrete-event simulation languages appeared in the late 1950s. These evolved during the 1960s and 1970s. With the arrival of the IBM PC, the 1980s saw the rise of visual interactive modeling environments that allowed simulation modelers to visually create and simulate discrete-event models. These have matured into the commercial off-the-shelf simulation packages (CSPs) that are very familiar to ORMS practitioners today. They include ArenaTM (www.arenasimulation.com), AnylogicTM (www.xjtek.com), FlexsimTM (www.flexsim.com), Simul8TM (www.simul8.com), WitnessTM (www.lanner.com), etc. and simulation languages such as SLXTM and GPSS/HTM (www.wolverinesoftware.com) [2]. Each has a wide range of functionality including visual model building, simulation run support, animation, optimization, and virtual reality. Some have their own dedicated programming language and all are able to be linked to other COTS (commercial off-the-shelf) software (such as Microsoft Excel). Nearly all CSPs only run under Microsoft WindowsTM. CSPs are typically used by modelers skilled in ORMS.

What is a distributed simulation? Essentially, a distributed simulation is a set of simulations, each running on a separate computer, that interact or *interoperate* with each other to accomplish the goals of the simulation. The typical reasons for using distributed simulation are that a simulation takes a long time to run or that models need to be merged to form a larger simulation (e.g., a manufacturing model and a warehouse model merging to expand the scope of both models in a supply chain) [3–7]. Further, if one wishes to combine one or more models developed in different CSPs, as these tools do not have the ability to save models as in a format compatible with different packages (as one might when saving a Microsoft Word document as text) or to import models from different packages, interoperability is really the only alternative (other than manual conversion!) Here are some ORMS examples of where distributed simulation can be used:

- A supply chain distributes equipment to front line troops. A model is built to represent the different supply centers and transportation links to various battlefronts. Experimentation investigates the reliability of the supply chain under different threat conditions.
- An automotive company is planning to build a new factory. The manufacturing line is modeled and simulated using a CSP. Experimentation investigates how many engines can be produced in one year against different levels of resources (machines, buffers, workers, etc.).
- A regional health authority needs to plan the best way of distributing blood to different hospitals. A model is built using a CSP. Experimentation is carried out to investigate different supply policies against “normal” and emergency supply situations.
- A police authority needs to determine how many officers need to be on patrol and how many need to be in the different

police stations that it manages. A model is built and experiments are carried out to investigate staffing against different scenarios (football matches, terrorist attacks, etc.).

- A bank sells different financial products. When a new product is planned, managers need to determine the resource impact against current financial services. Using existing business process models (in Business Process Modeling Notation or BPMN for example), a new model is built and simulated. Experiments investigate different resource levels in different departments.

To reinforce why distributed simulation may be needed, consider each of the above examples. If models are geographically remote (i.e., in separate places) then distributed simulation is needed to link the models together. If models are large, then distributed simulation is needed to share the processing load over different computers. If the models are written in different packages, then distributed simulation is again needed to link the models together. For example, in the case of the regional health authority, models in different CSPs may exist in different hospitals with different data sources that are confidential and cannot be easily moved. The combined model may also be large. Distributed simulation is therefore needed in this case to *both* link the models and their different CSPs together *and* to share the processing load. Not all of the above examples may need to share the processing. However, all may need distributed simulation as a result of being in separate places and using different packages.

To investigate this topic, we now discuss distributed simulation in ORMS and its technological difficulties. We then present an approach used to simplify the realization of distributed simulations and a case study illustrating its use.

DISTRIBUTED SIMULATION IN ORMS

As outlined above, distributed simulation in ORMS will typically involve the interaction,

or interoperability, of CSP. The set of techniques used to achieve this is called *commercial off-the-shelf simulation package interoperability* (CSPI). The simple act of linking together, or interoperating, two or more CSPs and their models, can be extremely difficult. This is due to time synchronization requirements and the complexity of distributed simulation algorithms and/or software used to create the link. Let us consider briefly both these issues.

The Time Synchronization Problem

Consider two models, a manufacturing system and a warehouse. The manufacturing system passes jobs to the warehouse for storage and the warehouse passes material to the manufacturing system for processing. Each model runs in its “own” CSP running on a separate computer. Let us assume that both simulations begin at 9 a.m. Monday. Both continue to simulate their systems. However, the manufacturing system model is more complex than the warehouse model and takes longer to simulate. After half an hour in real time, the manufacturing simulation is at 3 p.m. Tuesday and the warehouse simulation is at 11 a.m. Thursday. During this time both models have been interacting. As shown in Fig. 1, the effect of this is that if the manufacturing system sends a job to the warehouse, the job arrives in the past for the warehouse model! Figure 2 shows what should occur, that is, the two models are synchronized to ensure that the event is scheduled at the correct simulation time. To prevent this when simulations interact or *interoperate*, a *time synchronization method* is needed.

Time Synchronization Methods

How can we synchronize time? Essentially, each model becomes a *logical process* (LP) or *federate* that interacts with other models by time-stamped messages that represent the interaction of one model with another (say, when an entity leaves one part of a model and arrives at another) [8]. In other words, these time-stamped messages represent events that one model “schedules” for another, that is, a job leaving our manufacturing system is scheduled to arrive at the warehouse

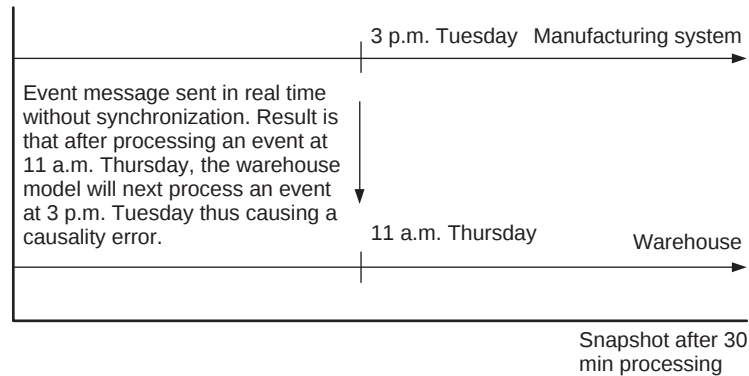


Figure 1. Distributed simulation without synchronization.

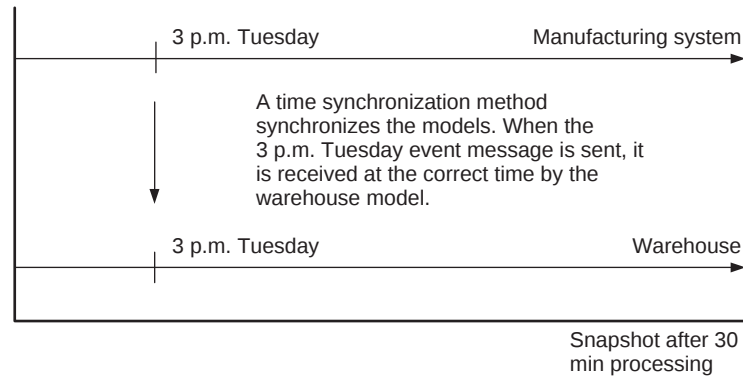


Figure 2. Distributed simulation with synchronization.

at 3 p.m. Tuesday. Time synchronization therefore involves techniques to correctly schedule that event in the destination model's event list. There are two major approaches to achieving this. *Conservative* approaches [9] do not allow any simulation to process an event until it is "safe" to do so; that is, when a simulation wishes to advance to the next event, the time synchronization approach determines if there is an event message that schedules an event before the simulation's next event time. If there is, the simulation executes that new event and repeats the process. *Optimistic* approaches in general are based on *Time Warp* [10], which instead allows simulations to process their events and to periodically check for these time-stamped event messages. If the time-stamp of an event message is in the past of the simulation, the simulation is "rolled back" to the last event before the time of the

new event message. The event message is then processed and the simulation continues. Reviews the many variants of these two approaches can be found in Ref. 11.

More recently, the "practice" of distributed simulation has been effected on a wide scale by attempts to standardize approaches. These standardization efforts took place in the mid to late 1990s and resulted in the IEEE 1516 standard *The High Level Architecture* (HLA), which supports the general needs of distributed simulation [12–15]. This is rooted in the early work by Chandy and Misra and Jefferson but considerably updates it with advances in distributed computing. The HLA defines the overall "rules," the runtime infrastructure (RTI) functions, and the format of the data that are used by a collection of models (federates) running on different computers to interact (such a collection of models is termed a *federation*) [16].

Examples of HLA RTIs include commercial RTIs from MAK (www.mak.com) and PITCH (www.pitch.se), and the service-oriented high-level architecture runtime infrastructure (SOHR) [17]. In HLA terms, a simulated model is a *federate*, and a collection of such models is termed a *federation*. The HLA specification supports different time synchronization approaches and has to some extent simplified the creation of a distributed simulation. However, the full specification runs to several hundred pages and the creation of a distributed simulation is a considerable task. In reality, distributed simulation is a very complex field that utilizes quite complex technologies. It is the purpose of this article to present distributed simulation from an ORMS viewpoint, that is, one that hides much of the complexities of distributed simulation and “exposes” those issues that should only concern ORMS end users.

Figure 3 shows the problem outlined above. A distributed simulation or federation is composed of a set of CSPs and

their models. Each model/CSP represents a federate normally running on its own computer. In a distributed simulation, each model/CSP federate therefore exchanges data directly or via an RTI implemented over a network in a time-synchronized manner (as denoted by the thick double-headed arrow). Federate F1 consists of the model M1 and the CSP1 and federate F2 consists of the model M2 and the simulation package CSP2. Note that A, Q, and R are generic symbols that represent activities, queues, and resources, respectively, and can be mapped onto any corresponding CSP objects (such as workstations, storage bins, etc.) In this case, federate F1 publishes and sends information to the RTI in an agreed format and time-synchronized manner, and federate F2 must subscribe to and receive that information in the same agreed format and time-synchronized manner; that is, both federates must agree on a common representation of data and both must use the RTI in a similar way. To further complicate matters,

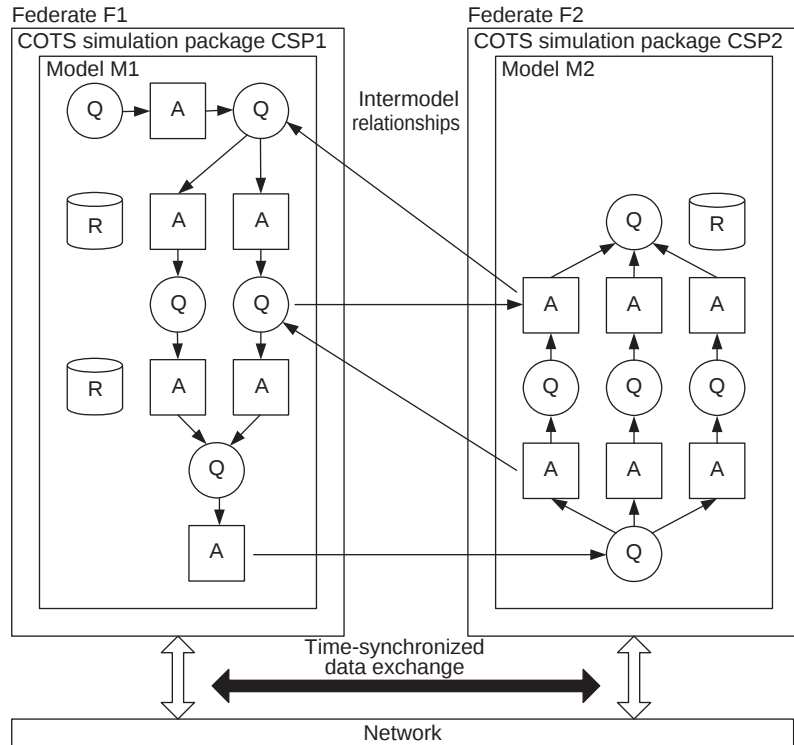


Figure 3. The COTS simulation package interoperability problem.

the “passing” of entities and the sharing of resources require different distributed simulation protocols. In entity passing, the departure of an entity from one model and the arrival of an entity at another can be the same scheduled event in the two models: most distributed simulations represent this as a time-stamped event message sent from one federate to another. The sharing of resources cannot be handled in the same way. For example, when a resource is released or an entity arrives in a queue, a CSP executing the simulation will determine if a workstation can start processing an entity. If resources are shared, each time an appropriate resource changes state, a time-stamped communication protocol is required to inform and update the changes of the shared resource state. Further problems arise when we begin to “dig” further into the subtleties of interoperability. In the next section we consider what has been done to simplify the implementation of ORMS distributed simulations and CSP interoperability.

COTS SIMULATION PACKAGE INTEROPERABILITY

There are several published examples of how distributed simulations have been created with CSPs. The first major work in the area was done by Straßburger [18]. Various strategies, implementations, and applications have been investigated since then. Illustrations of different approaches that have been used to create ORMS distributed simulations with CSPs can be found in Refs 19–37. This work, while contributing to the development of a distributed simulation, presents a problem. There is a variety of different approaches to distributed simulation with CSPs. As argued in Ref. 38, a common approach is needed. To standardize this approach, an international standardization group—Simulation Interoperability Standards Organization’s Commercial Off-The-Shelf Simulation Package Interoperability Product Development Group (CSPI PDG)—has produced a set of draft standards in this area, which is described initially in Ref. 36. The distributed approach used in this article is based on the

standards being developed by the CSPI PDG and can be found in full in Refs 39 and 40.

To simplify the process of generating a distributed simulation, the CSPI PDG has created a standardized set of interoperability reference models (IRMs) or “interoperability design patterns,” which attempt to capture these subtleties. These are effectively a set of simulation patterns or templates that enable modelers, vendors, and solution developers to specify the interoperability problems that must be solved. The IRMs are intended to be used as follows:

- To clearly *identify* the model/CSP interoperability *capabilities* of an *existing* distributed simulation; for example, the distributed supply chain simulation is compliant with IRMs Type A.1, A.2 and B.1;
- To clearly *specify* the model/CSP interoperability *requirements* of a *proposed* distributed simulation; for example, the distributed hospital simulation must be compliant with IRMs Type A.1 and C.1.

An IRM is defined as the simplest representation of a problem within an identified interoperability problem type. Each IRM can be subdivided into different subcategories of problem. As IRMs are usually relevant to the boundary between two or more inter-operating models, models specified in IRMs are as simple as possible to “capture” the interoperability problem and to avoid possible confusion. These simulation models are intended to be representative of real model/CSPs but use a set of “common” model elements that can be mapped onto specific (generalized) CSP elements. Where appropriate, IRMs specify time synchronization requirements and present alternatives. IRMs are intended to be cumulative (i.e., some problems may well consist of several IRMs). Most importantly, IRMs are intended to be understandable by simulation developers, CSP vendors, and technology solution providers.

Interoperability Reference Model Types

There are currently four different types of IRM. These are

Type A: entity transfer
 Type B: shared resource
 Type C: shared event
 Type D: shared data structure.

Briefly, IRM Type A, entity transfer, deals with the requirement of transferring entities between simulation models, such as an entity *Part* leaving one model and arriving at the next. IRM Type B, shared resource, refers to sharing of resources across simulation models. For example, a resource *R* might be common between two models and represents a pool of workers. In this scenario, when a machine in a model attempts to process an entity waiting in its queue it must also have a worker. If a worker is available in *R*, then processing can take place. If not, the work must be suspended until one is available. IRM Type C, shared event, deals with the sharing of events across simulation models. For example, when a variable within a model reaches a given threshold value (a quantity of production, an average machine utilization, etc.), it should be able to signal this fact to all models that have an interest in this fact (to throttle down throughput, route materials via a different path, etc.). IRM Type D, shared data structure, deals with the sharing of variables and data structures across simulation models. Such data structures are semantically different from resources, for example, a bill of materials or a common inventory. We now introduce the most commonly used IRM.

IRM TYPE A: ENTITY TRANSFER

Overview

IRM Type A, entity transfer, represents interoperability problems that can occur when transferring an entity from one model to another. Figure 4 shows an illustrative example of the problem of entity transfer where an entity *e1* leaves activity *A1* in model *M1* at *T1* and arrives at queue *Q2* in model *M2* at *T2*. For example, if *M1* is a car production line and *M2* is a paint shop, then this represents the system where a car leaves a finishing activity in *M1* at *T1* and arrives in a buffer in *M2* at *T2* to await painting.

IRM Type A Subtypes

There are currently three IRM Type A subtypes:

- IRM Type A.1: general entity transfer
- IRM Type A.2: bounded receiving element
- IRM Type A.3: multiple input prioritization.

IRM Type A.1: General Entity Transfer

IRM Type A.1, general entity transfer, represents the case, as described above and shown in Fig. 4, where an entity *e1* leaves activity *A1* in model *M1* at *T1* and arrives at queue *Q2* in model *M2* at *T2* (see above for an example). This IRM is inclusive of cases where

- there are many models and many entity transfers (all transfers are instances of this IRM).

This IRM does not include cases where

- the receiving element is bounded (IRM Type A.2), and
- multiple inputs need to be prioritized (IRM Type A.3).

The IRM Type A.1, general entity transfer, is defined as the transfer of entities from one model to another such that an entity *e1* leaves model *M1* at *T1* from a given place and arrives at model *M2* at *T2* at a given place and $T1 \leq T2$ or $T1 < T2$. The place of departure and arrival will be a queue, workstation, etc. Note that this inequality must be specified.

IRM Type A.2: Bounded Receiving Element

Consider a production line where a machine is just finishing working on a part. If the next element in the production process is a buffer in another model, the part will be transferred from the machine to the buffer. If, however, the next element is *bounded*, for example a buffer with limited space or another machine (i.e., no buffer space), then a check must be performed to see if there is space or the next machine is free. If there is no space, or the

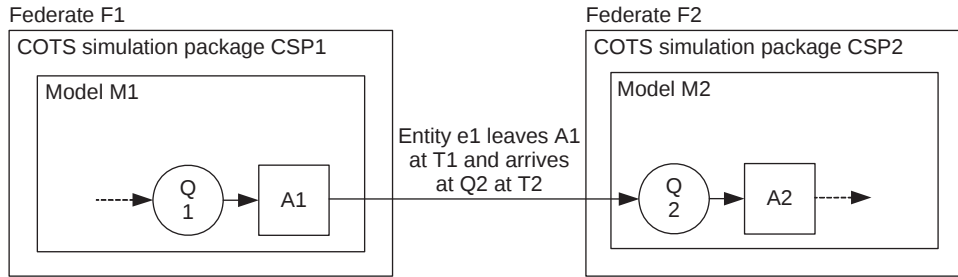


Figure 4. IRM Type A.1: general entity transfer.

next machine is busy, then to correctly simulate the behavior of the production process, the current machine must hold onto the part and *block*, that is, it cannot accept any new parts to process until it becomes unblocked (assuming that the machine can only process one part at a time). The consequences of this are quite subtle. This is the core problem of the IRM Type A.2. Figure 5 shows an illustrative example where an entity e_1 attempts to leave model M1 at T1 from activity A1 and to arrive at model M2 at T2 in bounded queue Q2. If A1 represents a machine, then the following scenario is possible. When A1 finishes work on a part (an entity), it attempts to pass the part to queue Q2. If Q2 has spare capacity, then the part can be transferred. However, if Q2 is full, then A1 cannot release its part and must block. Parts in Q1 must now wait for A1 to become free before they can be machined. Further, when Q2 once again has space, A1 must be notified so that it can release its part and transfer it to Q2. Finally, it is important to note the fact that if A1 is blocked the rest of model M1 still functions as normal, that is, a correct solution to this problem must still allow the rest of the model

to be simulated (rather than just stopping the simulation of M1 until Q2 has unblocked).

This IRM is therefore inclusive of cases where

- the receiving element (queue, workstation, and so on) is bounded.

This IRM does not include cases where

- multiple inputs need to be prioritized (IRM Type A.3).

A solution to this IRM problem must also

- be able to transfer entities (IRM Type A.1).

The IRM Type A.2 is defined as the relationship between an element O in a model M1 and a bounded element Ob in a model M2 such that if an entity e is ready to leave element O at T1 and attempts to arrive at bounded element Ob at T2, then

- if bounded element Ob is empty, the entity e can leave element O at T1 and arrive at Ob at T2; or

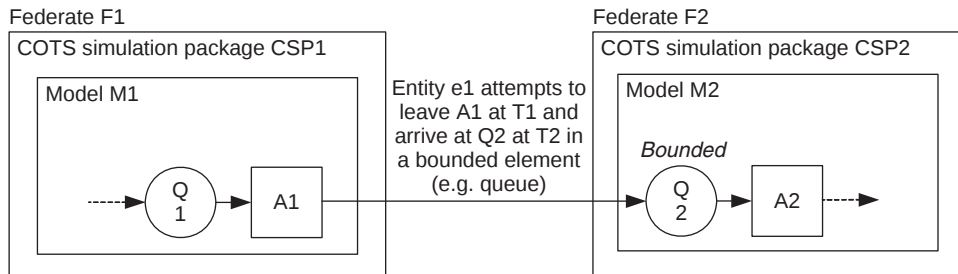


Figure 5. IRM Type A.2: bounded receiving element.

- if bounded element Ob is full, the entity e cannot leave element O at T1; element O may then block if appropriate and must not accept any more entities;
- when bounded element Ob becomes not full at T3, entity e must leave O at T3 and arrive at Ob at T4; element O becomes unblocked and may receive new entities at T3;
- $T1 \leq T2$ and $T3 \leq T4$;
- if element O is blocked, then the simulation of model M1 must continue.

Note:

- In some special cases, element O may represent some real-world process that may not need to be blocked.
- If $T3 < T4$, then it may be possible for bounded element O to become full again during the interval if other inputs to Ob are allowed.

IRM Type A.3: Multiple Input Prioritization

As shown in Fig. 6, the IRM Type A.3, multiple input prioritization, represents the case where a model element such as queue Q1 (or workstation) can receive entities from multiple places. Let us assume that there are two models M2 and M3 which are capable of sending entities to Q1 and that Q1 has a first in, first out (FIFO) queueing discipline. If an entity e1 is sent from M2 at T1 and arrives at Q1 at T2 and an entity e2 is sent from M3 at T3 and arrives at Q1 at T4, then if $T2 < T4$ we would expect the order of entities in Q1 would be e1, e2. A problem arises

when both entities arrive at the same time, that is, when $T2 = T4$. Depending on implementation, the order of entities would either be e1, e2 or e2, e1. In some modeling situations, it is possible to specify the *priority* order if such a conflict arises, for example, it can be specified that model M1 entities will always have a higher priority than model M2 (and therefore require the entity order e1, e2 if $T2 = T4$). Further, it is possible that this priority ordering could be dynamic or specialized.

This IRM is therefore inclusive of cases where

- multiple inputs need to be prioritized.

This IRM does not include cases where

- the receiving element is bounded (IRM Type A.2).

A solution to this IRM problem must also

- be able to transfer entities (IRM Type A.1).

The IRM Type A.3, multiple input prioritization, is defined as the preservation of the priority relationship between a set of models that can send entities to a model with receiving queue Q, such that priority ordering is observed if two or more entities arrive at the same time.

Note:

- The priority rules must be specified.
- Priority rules may change during a simulation if required for the real system being simulated.

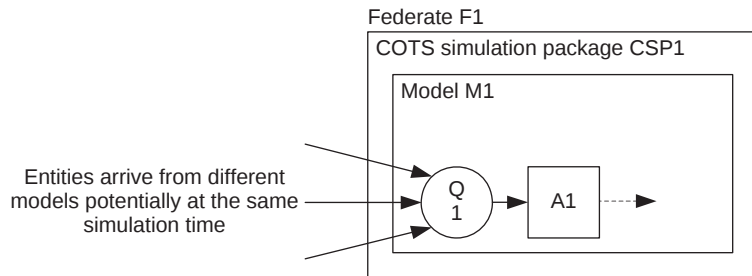


Figure 6. IRM Type A.3: multiple input prioritization.

CASE STUDY

In Ref. 41, experiences with the use of discrete-event simulation to investigate the supply chain of blood from the UK National Blood Service (NBS) Centre to hospitals are described. The CSP used in this work was Simul8™. The purpose of the study was to achieve improved inventory control, ordering, and distribution policies; to bring less wastage and less shortage; and provide better quality service. This was done by reconfiguring the processes and parameters of the system. A problem identified in this work was that, as the system being modeled grew in size and complexity, the time taken to perform one simulation run increased to a point that made the use of simulation infeasible. Figure 7 shows a simplified illustration of this simulation model. For multiple hospitals, Fig. 8 shows an example of the relationships between the NBS supply center and the hospitals it serves, which in the “conventional” approach is simulated on a single computer.

Figure 9 shows the conventional model of Fig. 8 using our distributed approach. The natural “division” in the model is between the hospitals and the NBS Center. The NBS distributed simulation fits the profile of the Type A.1 IRM (entity transfer, $T1 < T2$), that is, dividing the model into federates results in individual, distributed models of the

Southampton NBS PTI and hospitals that interact by exchanging entities. There are no bounded buffers, and queueing discipline is not an issue. The condition $T1 < T2$ holds as entities will always take simulated time to transfer from one model to another (effectively travel time). These models run in separate copies of Simul8. Together, these form a federation that interact by time-stamped messages that represent the interaction of one model part with another (e.g., when an entity leaves one part of a model and arrives at another). Entities are represented using the CSPI entity transfer specification and exchanged using HLA interactions. In this investigation, the interaction between the models/Simul8 and the HLA-DMSO RTI is via an Excel file (as the main data used by each model is held in this file). For example, entities representing orders are written into the file by Simul8 during the execution of hospital models. The HLA-DMSO RTI, augmented by the CSPI standards, then correctly transfers this information to the NBS model by means of HLA interactions. The incoming orders from each hospital are collected into their corresponding queues in the NBS model and the orders are matched with the available stock of blood. The resulting matched units are written into the Excel spreadsheet in the NBS federate. This information is then

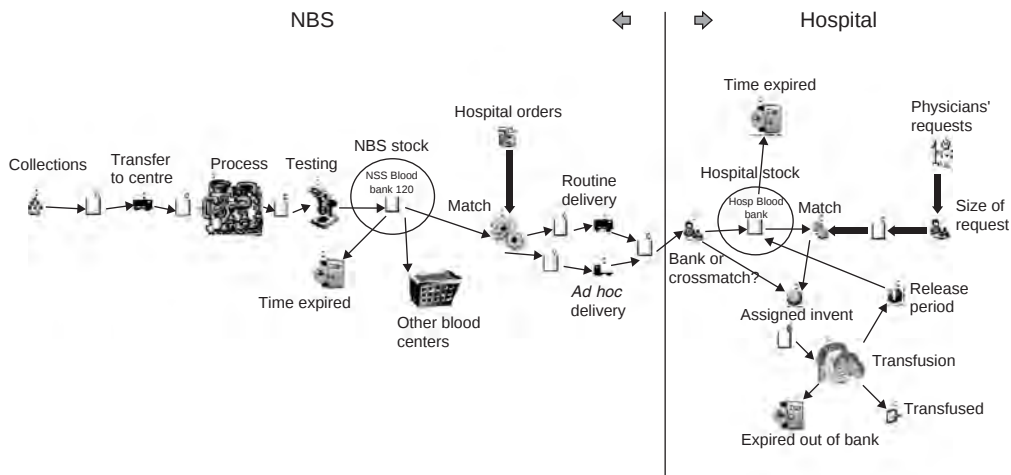


Figure 7. Simplified National Blood Service model.

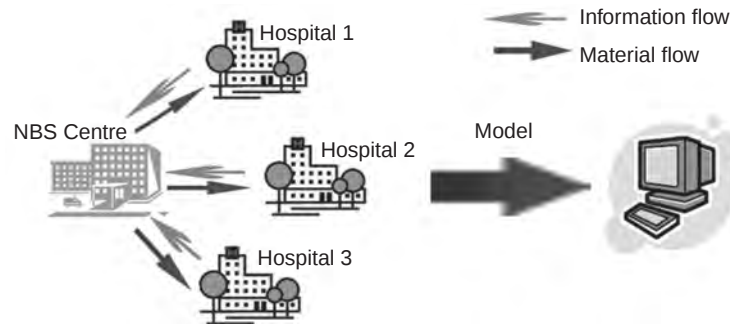


Figure 8. National Blood Service conventional model.

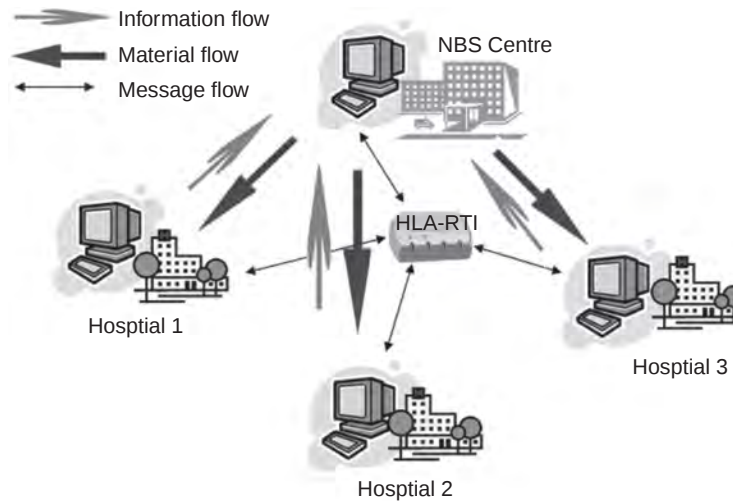


Figure 9. National Blood Service distributed model.

sent to the different hospital models again through HLA interactions. The interactions between the hospitals and the NBS center are sent every 60 min of simulated time, provided orders/delivers exist. Thus, although the discrete-event simulation generates orders and deliveries as the model progresses in time, these are only released at specific time steps. It is to be noted here that this time-stepped information exchange behavior occurs as a result of the blood ordering and delivery policies in place with NBS. The distributed simulation successfully achieved a faster runtime than the standalone counterpart. Further details on this work and the implementation of the distributed simulation can be found in Refs 42 and 43.

FURTHER READING

This article has presented distributed simulation from an ORMS viewpoint. Distributed simulation is a complex area which in some ways is a realm for computer scientists. The case study briefly presented was the result of a close collaboration between ORMS researchers, computer scientists, and a CSP vendor. Previous experience such as this has resulted in the IRMs in this article as an accessible entry point to this fascinating field. It is hoped that this “re-presentation” of this field in terms of ORMS will make the highly useful technique of distributed simulation more available to ORMS practitioners and researchers. In terms of further reading,

readers should refer to the Winter Simulation Conference (www.wintersim.org), which has online, free-to-access papers on a variety of simulation-related topics. There are regular papers on distributed simulation and interoperability. Additionally, readers should also investigate how *Grid Computing* can speed up experimentation and optimization in ORMS in this area, as this is also proving to be another fruitful outcome from the interface between ORMS and computer science [44,45]. Related to this is the development of the SOHR, which is a runtime infrastructure based on Grid services [46]. Readers should also see the study performed in Ref. 47, which reflects on associated areas of simulation technology.

REFERENCES

1. Law AM. Simulation Modeling and Analysis. 4th ed. Boston: McGraw-Hill; 2007.
2. Swain JJ. Simulation Reloaded: Sixth biennial survey of discrete-event software tools. *OR/MS Today* 2003;30(4):46–57.
3. Gan BP, Li L, Jain S, *et al.* Distributed supply chain simulation across enterprise boundaries. Proceedings of the 2000 winter simulation conference. New York: Association for Computing Machinery Press; 2000. pp. 1245–1251.
4. Lendermann P, Gan BP, McGinnis LF. Distributed simulation with incorporated APS procedures for high-fidelity supply chain optimization. Proceedings of the 2001 winter simulation conference. New York: Association for Computing Machinery Press; 2001. pp. 1138–1145.
5. Lendermann P, McGinnis LF, McLean CR, *et al.* Panel: distributed simulation in industry—a real-world necessity or ivory tower fancy? Proceedings of the 2007 winter simulation conference. New York: Association for Computing Machinery Press; 2007. pp. 1053–1062.
6. Mertins K, Rabe M. Inter-enterprise planning of manufacturing systems applying simulation with IPR protection. 5th International Conference on Design of Information Systems for Manufacturing (DIISM). Osaka (Japan); 2002. pp. 149–156.
7. Taylor SJE, Bruzzone A, Fujimoto R, *et al.* Distributed simulation and industry: potentials and pitfalls. Proceedings of the 2002 winter simulation conference. New York: Association for Computing Machinery Press; 2002. pp. 688–694.
8. Fujimoto RM. Distributed simulation systems. Proceedings of the 2003 winter simulation conference. New York: Association for Computing Machinery Press; 2003. pp. 124–134.
9. Chandy KM, Misra J. Distributed simulation: a case study in design and verification of distributed programs. *IEEE Trans Softw Eng* 1978;SE-5(5):440–452.
10. Jefferson DR. Virtual time. *ACM Trans Program Lang Syst* 1985;7(3):404–425.
11. Fujimoto RM. Parallel and distributed simulation systems. New York: John Wiley & Sons, Inc.; 2000.
12. IEEE 1516. IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA). New York: Institute of Electrical and Electronics Engineers; 2000.
13. IEEE 1516.0. IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)—Rules. New York: Institute of Electrical and Electronics Engineers; 2000.
14. IEEE 1516.1. IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Federate Interface Specification. New York: Institute of Electrical and Electronics Engineers; 2000.
15. IEEE 1516.2. IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) - Object Model Template (OMT) Specification. New York: Institute of Electrical and Electronics Engineers; 2000.
16. Kuhl F, Weatherly R, Dahmann J. Creating computer simulation systems: an introduction to the high level architecture. Upper Saddle River (NJ): Prentice Hall PTR; 1999.
17. Pan K, Turner SJ, Cai W, *et al.* Design and performance evaluation of a service oriented HLA RTI on the grid. In: Wang L, Jie W, Chen J, editors. Grid computing: infrastructure, service, and application. Boca Raton (FL): Taylor and Francis; 2008.
18. Straßburger S. Distributed simulation based on the high level architecture in civilian application domains. Ghent: Society for Computer Simulation International; 2001.
19. Borshchev A, Karpov Y, Kharitonov V. Distributed simulation of hybrid systems with AnyLogic and HLA. *Future Gener Comput Syst* 2002;18(6):829–839.
20. Gan BP, Lendermann P, Low MYH, *et al.* Interoperating autosched AP using the high

- level architecture. Proceedings of the 2005 winter simulation conference. New York: Association for Computing Machinery Press; 2005. pp. 394–401.
21. Gan BP, Low MYH, Wang X, *et al.* Using manufacturing process flow for time synchronization in HLA-based simulation. Proceedings of the 9th International Symposium on Distributed Simulation and Real-Time Applications. Montreal, Canada; 2005. pp. 148–157.
 22. Hibino H, Fukuda Y, Yura Y, *et al.* Manufacturing adapter of distributed simulation systems using HLA. Proceedings of the 2002 winter simulation conference. New York: Association for Computing Machinery Press; 2002. pp. 1099–1107.
 23. Lendermann P, Low MYH, Gan BP, *et al.* An integrated and adaptive decision-support framework for high-tech manufacturing and service networks. Proceedings of the 2005 winter simulation conference. New York: Association for Computing Machinery Press; 2005. pp. 2052–2062.
 24. McLean C, Riddick F. The IMS MISSION architecture for distributed manufacturing simulation. Proceedings of the 2000 winter simulation conference. New York: Association for Computing Machinery Press; 2000. pp. 1539–1548.
 25. Mertins K, Rabe M. Inter-enterprise planning of manufacturing systems applying simulation with IPR protection. Proceedings of the 5th International Conference on Design of Information Systems for Manufacturing (DIISM). Osaka, Japan; 2002. pp. 149–156.
 26. Mertins K, Rabe M, Jäkel FW. Neutral template libraries for efficient distributed simulation within a manufacturing system engineering platform. Proceedings of the 2000 winter simulation conference. New York: Association for Computing Machinery Press; 2000. pp. 1549–1557.
 27. Mustafee N, Taylor SJE, Katsaliaki K, *et al.* Distributed simulation with COTS simulation packages: a case study in health care supply chain simulation. Proceedings of the 2006 winter simulation conference. New York: Association for Computing Machinery Press; 2006. pp. 1136–1142.
 28. Mustafee N, Taylor SJE. Investigating distributed simulation with COTS simulation packages: experiences with simul8 and the HLA. Proceedings of the 2006 Operational Research Society Simulation Workshop (SW06). UK; 2006. pp. 33–42.
 29. Rabe M, Jäkel FW. On standardization requirements for distributed simulation. Proceedings production and logistics. Twente: Building the Knowledge Economy; 2003. pp. 399–406.
 30. Rabe M, Jäkel FW. Non military use of HLA within distributed manufacturing scenarios. Proceedings of Simulation und Visualisierung. Magdeburg, Germany; 2001. pp. 141–150.
 31. Revetria R, Blomjous PEJN, Van Houten SPA. An HLA federation for evaluating multi-drop strategies in logistics. Proceedings of the 15th European Simulation Symposium and Exhibition Conference. Delft, The Netherlands; 2003. pp. 450–455.
 32. Strassburger S. The road to COTS-interoperability: from generic HLA-interfaces towards plug-and-play capabilities. Proceedings of the 2006 winter simulation conference. New York: Association for Computing Machinery Press; 2006. pp. 1111–1118.
 33. Strassburger S, Schulze T, Klein U, *et al.* Internet-based Simulation using off-the-shelf Simulation Tools and HLA. Proceedings of the 2003 winter simulation conference. New York: Association for Computing Machinery Press; 1998. pp. 1669–1676.
 34. Taylor SJE, Bohli L, Wang X, *et al.* Investigating distributed simulation at the ford motor company. Proceedings of the 9th International Symposium on Distributed Simulation and Real-Time Applications. Montreal, Canada; 2005. pp. 139–147.
 35. Taylor SJE, Turner SJ, Low MYH, *et al.* Developing interoperability standards for distributed simulation and COTS simulation packages with the CSPI PDG. Proceedings of the 2006 winter simulation conference. New York: Association for Computing Machinery Press; 1101–1110.
 36. Taylor SJE, Wang X, Turner SJ, *et al.* Integrating heterogeneous distributed COTS discrete-event simulation packages: an emerging standards-based approach. IEEE Trans Syst Man Cybern Part A 2006;36(1):109–122.
 37. Wang X, Turner SJ, Low MYH, *et al.* COTS simulation package (CSP) interoperability—a solution to synchronous entity passing. Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation. Singapore; 2006. pp. 201–210.
 38. Taylor SJE. The high level architecture—COTS simulation package interoperability forum (HLA-CSPIF). Proceedings of the

- European simulation interoperability workshop 2003. Florida: Simulation Interoperability Standards Organization, Institute for Simulation and Training; 2003. 03E-SIW-129.
39. Simulation Interoperability Standards Organization. Standard for Commercial-Off-The-Shelf Simulation Package Interoperability Reference Models; 2007. Available at <http://www.sisostds.org>
 40. Taylor SJE, Turner SJ, Strassburger S. Guidelines for commercial off-the-shelf simulation package interoperability. Proceedings of the 2008 winter simulation conference. New York: Association for Computing Machinery Press; 2008. pp. 193–204.
 41. Katsaliaki K, Brailsford SC. Using simulation to improve the blood supply chain. *J Oper Res Soc* 2007;58(2):219–227.
 42. Katsaliaki K, Mustafee N, Taylor SJE, *et al.* Comparing conventional and distributed approaches to simulation in a complex supply-chain health system. *J Oper Res Soc* 2009;60(1):43–51.
 43. Mustafee N, Taylor SJE, Katsaliaki K, *et al.* Facilitating the analysis of a UK national blood service supply chain using distributed simulation. *Simul: Trans Soc Model Simul Int* 2009;82(2):113–128.
 44. Mustafee N, Taylor SJE. Speeding up simulation applications using wingrid. *Virtual Environ Real Time Appl* 2009;21(11):1504–1523.
 45. Mustafee N. A grid computing framework for commercial simulation packages [PhD thesis]. Uxbridge: Brunel University; 2007.
 46. Pan K, Turner SJ, Cai W, *et al.* Design and performance evaluation of a service oriented HLA RTI on the grid. In: Wang L, Jie W, Chen J, editors. *Grid computing: infrastructure, service, and application*. Boca Raton (FL): Taylor and Francis; 2009. pp. 459–482.
 47. Taylor SJE, Robinson S. So where to next? A survey of the future for discrete-event simulation. *J Simul* 2006;1(1):1–6.

DOMINATION PROBLEMS

SIQIAN SHEN

Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

INTRODUCTION

Given a graph $G(V, E)$ with vertex set V and edge set E , we say that a vertex can “dominate” itself as well as all adjacent vertices. A *dominating set* S is a subset of V , such that every vertex in V is dominated by at least one vertex in S . Domination began as a problem on a chessboard [1], where we solve for a minimum number of queens that can be placed on a chessboard so that every square is either occupied by a queen, or can be occupied by one of the queens within a single move. A standard form of the domination seeks a dominating set with minimum cardinality [2,3].

Domination-related problems have received extensive attention after the 1970s because of their great importance in the usage of military, social network, medicine, and many other real-world problems. Over 1200 papers in areas of mathematics, computer science, and operations research have been published based on the study of domination and its related topics, such as independence and covering. See Refs 4 and 5 for a comprehensive discussion of theories and algorithms that have been studied regarding fundamentals and advanced topics of domination.

Notations and Definitions

We first introduce the notation that is commonly used in the domination literature. Assume that $G(V, E)$ is undirected, and let $|\cdot|$ be the cardinality of $(\cdot)$. The number of vertices, denoted as $|V| = n$, is the *order* of G , and the number of edges, denoted as $|E| = m$,

is the *size* of G . We specify the vertex set $V = \{v_1, \dots, v_n\}$ and edge set $E = \{e_1, \dots, e_2\}$. Denote $e = v_i v_j \in E$ as an undirected edge $\{v_i, v_j\} \in E, v_i, v_j \in V, v_i \neq v_j$. If $e = v_i v_j \in E$, we say that v_i and v_j are *adjacent*, and v_i (or v_j) and e are *incident*.

Let $N(v)$ be the *open neighborhood* of vertex $v \in V$, such that $N(v) = \{u \in V | vu \in E\}$, and $N[v] = N(v) \cup \{v\}$ is called the *closed neighborhood* of v . Similarly, for any subset $S \subseteq V$, the open and closed neighborhoods of S are denoted as $N(S)$ and $N[S]$, respectively, where $N(S) = \bigcup_{v \in S} N(v)$ and $N[S] = N(S) \cup S$. The *degree* of vertex v , denoted as $\deg(v)$, is calculated as the number of distinct vertices that are adjacent to v , that is, $\deg(v) = |N(v)|$. We describe the *degree sequence* of graph G as $\{\deg(v_1), \dots, \deg(v_n)\}$, typically in a nonincreasing or a nondecreasing order. We also use $\delta(G)$ and $\Delta(G)$ to denote the *minimum* and the *maximum degrees* of vertices in V , respectively.

For any pair of $u, v \in V$, the *distance* $d(u, v)$ is the shortest path length between u and v in G . In particular, if u and v are disconnected, we have $d(u, v) = +\infty$. We define *ecc*(v) as the *eccentricity* of v , such that $\text{ecc}(v) = \max\{d(v, u) | \forall u \in V\}$. Moreover, given $v \in V$ and $S \subseteq V$, we define the *distance* between v and S as $d(v, S) = \min_{u \in S} \{d(v, u)\}$, and the *eccentricity* of S is $\text{ecc}(S) = \max_{v \in V} \{d(v, S)\}$. For any graph G , we also define the *radius* $\text{rad}(G)$ and the *diameter* $\text{diam}(G)$ such that $\text{rad}(G) = \min_{v \in V} \{\text{ecc}(v)\}$, and $\text{diam}(G) = \max_{v \in V} \{\text{ecc}(v)\}$. By applying triangle inequality, the following result holds.

Theorem 1 [5]. For any connected graph G , $\text{rad}(G) \leq \text{diam}(G) \leq 2\text{rad}(G)$.

We illustrate the notation in Fig. 1, in which we present an unweighted and undirected graph with $n = 8$ vertices and $m = 9$ edges. The open neighborhood of vertex 3 is $N(3) = \{2, 4, 5\}$ with vertex degree $\deg(3) = 3$. The closed neighborhood $N[\{3, 5\}]$ is $\{2, 3, 4, 5, 6, 8\}$, and $d(2, 5)$ is 2. Moreover,

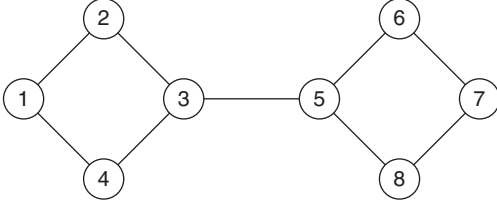


Figure 1. Illustration of notation and definitions.

$ecc(3) = \max\{d(3, u)\} = d(3, 7) = 3$, $rad(G) = \min_{v=1, \dots, 8} \{ecc(v)\} = ecc(3) = 3$, and $diam(G) = \max_{v=1, \dots, 8} \{ecc(v)\} = ecc(1) = ecc(7) = 5$. The minimum and maximum degrees of vertices are $\delta(G) = 2$ and $\Delta(G) = 3$, respectively.

Graph Variations

The characteristics of domination have been explored on graphs with different parameters. Next, we define all classes of graphs that we will use in our later discussion.

- *Split Graphs and Bipartite Graph.* A set S is a *stable set/independent set* if any pair of vertices in S is not adjacent. A set S is a *clique* if every pair of vertices is adjacent to each other. A graph is called a *split graph* if it can be partitioned into a stable set and a clique. A graph is called a *bipartite graph* if it can be partitioned into two stable sets.
- *Intersection Graph.* A graph $G(V, E)$ is called an *intersection graph* for a finite family $\mathcal{F}$ if there is a one-to-one correspondence between $\mathcal{F}$ and V such that two sets in $\mathcal{F}$ have nonempty intersection if and only if their corresponding vertices in V are adjacent. $\mathcal{F}$ is an *intersection model* of G .
 - *Interval Graph.* If $\mathcal{F}$ is given as a family of intervals on a real line, G is called an *interval graph*.
 - *Circular-Arc Graph.* If $\mathcal{F}$ is given as a family of arcs in a circle, G is called a *circular-arc graph*.
- *Chordal Graph.* A *perfect elimination order* is an ordering $V = \{v_1, \dots, v_n\}$ of G such that for every i, j , and k ($i < j < k$), if $v_i v_j \in E, v_i v_k \in E$, then $v_j v_k \in E$. A graph is *chordal* if and only if it admits a perfect elimination order, that is, every cycle of G of length greater than 3 has a “chord”—an edge between two nonconsecutive vertices in the cycle.
- *Strongly Chordal Graph.* A *strong elimination order* is a perfect elimination order satisfying for every i, j, k , and l ($i < j < k < l$) $:= v_i v_k \in E, v_i v_l \in E, v_j v_k \in E$, then $v_j v_l \in E$. A graph is *strongly chordal* if it admits a strong elimination ordering, that is, a chordal graph with every cycle of even length exceeding 5 has an odd chord. Strongly chordal graphs are important, including many interesting classes of graphs such as trees, block graphs, interval graphs, and directed path graphs.
- *Distance-Hereditary Graph.* A graph G is a *distance-hereditary graph* if every pair of vertices has the same distance in every connected induced subgraph containing them.
- *Series-Parallel Graph.* A *two-terminal graph* (TTG) $G(s, t)$ contains two distinguished nodes s and t , representing a *source* and a *sink*, respectively. Given two TTGs $G_1(s_1, t_1)$ and $G_2(s_2, t_2)$, a *series composition* creates a new TTG by merging the sink t_1 and the source s_2 , where s_1 becomes new source and t_2 becomes the new destination node. We define such a series operation by $G = S(G_1, G_2)$. A *parallel composition* creates a new TTG by merging the two sources s_1 and s_2 into the new source node, and the two sinks t_1 and t_2 into the new destination node. The parallel operation is denoted by $G = P(G_1, G_2)$. A *series-parallel graph* (SPG) is a graph that is constructed by a sequence of series and parallel compositions starting from a set of copies of a single edge graph.
- *Directed and Undirected Path Graph.* A *directed path graph* is a directed graph with no directed cycles. *Undirected path graphs* are the vertex intersection graphs of paths in a tree.
- *Permutation Graph.* A *permutation graph* is the intersection graph of a

family of line segments that connect two parallel lines in the Euclidean plane. Equivalently, given a permutation $(\pi_1, \dots, \pi_n)$ of $\{1, \dots, n\}$, a permutation graph has a vertex for each number $1, \dots, n$, and an edge between any two numbers that are in reversed order in the permutation.

The remainder of the paper is organized as follows. In the section titled “The Dominating Set Problem,” we formally state the dominating set problem. We review the algorithmic complexity of domination and domination-related parameters of graphs, and provide some solving strategies. The section titled “The Domination Chain” relates domination with independence and irredundance, and shows properties of the domination chain and its parameters’ complexity. In the section titled “Variants of Domination,” we list variants of domination in graphs, and describe the corresponding bounds and complexity details. The section titled “Applications” states several domination-related applications. Finally, the section titled “Summary and Open Questions” concludes the article and provides some open problems that are remaining in domination for potential future research.

THE DOMINATING SET PROBLEM

Originating from the queens problem, the dominating set problem and many of its variants have been introduced to solve varied practical problems. A general standard form of a dominating set problem is defined by

Definition 1. Given $G(V, E)$, a subset $S \subseteq V$ is called a *dominating set* if every vertex $v \in V$ is either an element of S or adjacent to at least one element of S .

Haynes *et al.* [5] summarize the following equivalent definitions, each of which defines the dominating set problem. A subset $S \subseteq V$ is a dominating set if one of the following is true:

1. $\forall v \in V - S$, there exists $u \in S$ such that $uv \in E$;
2. $\forall v \in V - S$, $d(v, S) \leq 1$;

3. $N[S] = V$;
4. $\forall v \in V - S$, $N(v) \cap S \neq \emptyset$.

Note that if $S \subseteq V$ is a dominating set, any superset of S is also a dominating set. This motivates us to find the *minimal dominating set* S of G , such that no proper subset $S' \subset S$ is a dominating set. To verify whether a given set S is a minimal dominating set, we introduce the definitions of *enclave* and *isolate*, such that

Definition 2. A vertex $v \in S$ is called an *enclave* of S if $N[v] \subseteq S$, and $v \in S$ is an *isolate* if $N(v) \subseteq V - S$.

Ore [3] gives the detailed proof of the following first three theorems for the dominating set problem.

Theorem 2 [3]. A dominating set S is minimal if and only if one of the following two conditions hold for every vertex $u \in S$:

- u is an isolate of S ;
- there exists a vertex $v \in V - S$, such that $N(v) \cap S = \{u\}$.

Theorem 3 [3]. Every connected graph G of order $n \geq 2$ has a dominating set S whose complement set $V - S$ is also a dominating set.

Theorem 4 [3]. If G has no isolated vertices, then for every minimal dominating set S , the complement set $V - S$ is also a dominating set.

Cockayne and Hedetniemi [6] are the first to introduce the concept of domination number, denoted as $\gamma(G)$, and $\gamma(G)$ and $\Gamma(G)$ denote the minimum and the maximum cardinality of a minimal dominating set in G , respectively. A minimum dominating set is also called a γ -set. Letting $\epsilon_F(G)$ to be the maximum number of pendant (leaf) edges in a spanning forest of G , we present the following basic theorem.

Theorem 5 [7]. $\gamma(G) + \epsilon_F(G) = n$.

We illustrate the difference between concepts of the *minimal* and the *minimum*

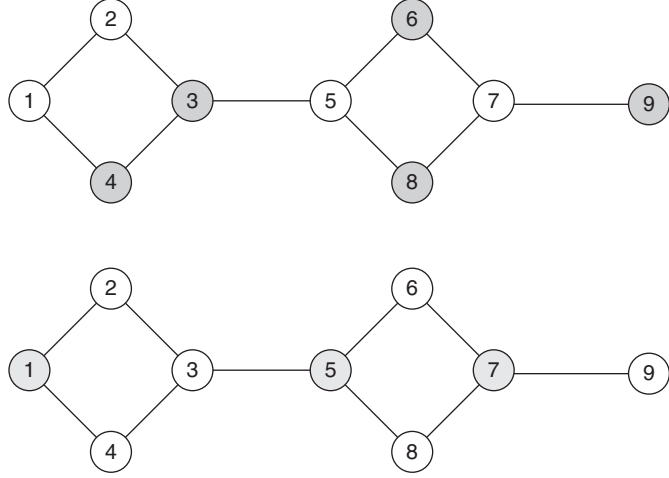


Figure 2. Illustration of minimal and minimum dominating sets.

in Fig. 2, in which both solutions (i.e., the marked subsets of vertices) provide us minimal dominating sets, since the deletion of any vertex will lead to a non-dominated graph. However, the solution in the bottom (the lightly marked one) also gives a minimum dominating set, that is, $\gamma(G) = 3$.

Finding the minimum dominating set in a graph G then becomes one of the standard combinatorial problems that is substantially interesting. A decision version of this problem is stated as “does G have a dominating set of size $\leq k$ for any given positive integer k ?” It has been shown that this problem, called *DOMINATING SET*, is $\mathcal{NP}$ -complete in general graphs, derived from a transformation of 3-SAT [8]. Moreover, the study of *DOMINATING SET* has been extended to problems with modified domination-related statements, or given varied parameters of graphs. In some cases, it is possible to obtain polynomial algorithms. The complexity and algorithm details with respect to each domination-related problem are discussed throughout the remainder of the article.

Bounds for the Domination Number $\gamma(G)$

Since the dominating set problem is $\mathcal{NP}$ -complete for general graphs, many studies seek good-quality bounds for $\gamma(G)$. This problem is tackled from different angles such as in terms of number of vertices, number of edges, degree, and diameter. We briefly review some

of the related results that have been proposed during the past two decades. Note that they are not comprehensive, and only represent a part of the literature.

It is trivial to see that $1 \leq \gamma(G) \leq n$, in which $\gamma(G)$ reaches the lower bound if there exists at least one vertex $v \in V$ such that $\deg(v) = n - 1$, and $\gamma(G)$ achieves the upper bound if all vertices are isolated. Recall that $\delta(G)$ is the minimum degree of vertices in V . If $\delta(G) \geq 1$, that is, no isolated vertex exists in G , we have a direct result in the following:

Theorem 6 [3]. *If G has no isolated vertices such that $\delta(G) \geq 1$, then $\gamma(G) \leq \lfloor n/2 \rfloor$.*

Then, an intuitive question arises as to whether we can tighten the upper bound for $\gamma(G)$ when $\delta(G)$ increases. We show some representative results as follows:

Theorem 7 [9]. *If G is a connected graph with $\delta(G) \geq 3$, then $\gamma(G) \leq \lfloor 3n/8 \rfloor$.*

Theorem 8 [10,11]. *For any graph G with $\delta(G) \geq 7$, we have $\gamma(G) \leq n \left[1 - \delta(G) \left(\frac{1}{\delta(G)+1} \right)^{1+1/\delta} \right]$.*

However, for graph G with $4 \leq \delta(G) \leq 6$, the question is still open [5]. The following theorem calculates bounds for $\gamma(G)$ for almost every graph.

Theorem 9 [12]. *For almost every order- n graph, let $k = \lfloor (\log n - 2 \log \log n + \log \log e) \rfloor$. We have $k + 1 \leq \gamma(G) \leq k + 2$.*

In terms of n and $\Delta(G)$, to obtain a lower and an upper bound for $\gamma(G)$, a trivial result is shown as follows:

Theorem 10 [2,13]. *For any graph G , $\left\lceil \frac{n}{1+\Delta(G)} \right\rceil \leq \gamma(G) \leq n - \Delta(G)$.*

The proof is intuitive, and here we generally describe the idea. For the lower bound, since each vertex can dominate itself and all its adjacent vertices, that is, at most $(1 + \Delta(G))$ vertices, we need at least $\left\lceil \frac{n}{1+\Delta(G)} \right\rceil$ vertices to dominate G . For the upper bound, a feasible dominating solution can be given as a set containing the maximum degree vertex (dominating itself and all its connections) and all other vertices (dominating themselves), and thus, $\gamma(G) \leq n - \Delta(G)$.

Moreover, considering a graph G with a *nonincreasing* degree sequence $(\deg(v_1), \dots, \deg(v_n)) = (d_1, \dots, d_n)$, a tighter lower bound can be given by

Theorem 11 [14]. $\gamma(G) \geq \min\{k : k + (d_1 + d_2 + \dots + d_k) \geq n\}$.

The proof is similar to finding the lower bound in Theorem 10. We sequentially include vertices from the degree sequence such that they dominate themselves as well as their adjacent vertices. We repeat this procedure until all the vertices in V are dominated, and Theorem 11 provides the smallest vertices that we need in the best case such that all the picked vertices are not adjacent to each other. Moreover, we refer readers to Refs 5,15, and 16 for varied upper bound results in terms of $\delta(G)$ and $\Delta(G)$.

Similarly, we give the following bounds for $\gamma(G)$ in terms of n and m .

Theorem 12 [17]. *If a graph G has $\gamma(G) \geq 2$, then $m \leq \lfloor \frac{1}{2}(n - \gamma(G))(n - \gamma(G) + 2) \rfloor$.*

Theorem 13 [18]. *If a graph G has $\gamma(G) \geq 2$ and $\Delta(G) \leq n - \gamma(G) - 1$, then $m \leq \frac{1}{2}(n - \gamma(G))(n - \gamma(G) + 1)$.*

Considering bounds for $\gamma(G)$ in terms of the degree and diameter, we give the following two elementary theorems:

Theorem 14 [5]. *Given a graph G with $\text{diam}(G) = 2$, we have $\gamma(G) \leq \delta(G)$, that is, $N(v)$ dominate G for any vertex $v \in V$.*

Theorem 15 [5]. *If G is a connected graph, we have $\left\lceil \frac{\text{diam}(G)+1}{3} \right\rceil \leq \gamma(G)$.*

Solving Domination Problems

Solving the domination problem on various types of graphs has stimulated great interests in studies of operations research, combinatorial optimization, and computer science. In this section, we particularly review optimization techniques to formulate the domination, such as integer programming (IP). We also examine a polynomial-time algorithm for solving the domination problem on trees.

Combinatorial Optimization and Integer Programming. Let x_i be a binary variable indicating whether vertex $i \in V$ is contained in a minimum dominating set S , such that $x_i = 1$ if $i \in S$, and $x_i = 0$ otherwise. We propose the following IP formulation:

$$\gamma(G) = \min \sum_{i \in V} x_i \quad (1a)$$

$$\text{s.t.} \quad \sum_{i \in N[v]} x_i \geq 1 \quad \forall v \in V \quad (1b)$$

$$x_i \in \{0, 1\} \quad \forall i \in V, \quad (1c)$$

where Equation (1c) causes main computational difficulty in solving Equation (1a). By relaxing $x \in [0, 1]^{|V|}$, a linear programming (LP) relaxation provides optimal solutions for a *fractional domination* problem [19], and a lower bound for the IP model [20]. An alternative approach can also be used to obtain a lower bound value as follows. We define $Y = xx^T$, which is a positive semi-definite (PSD) $|V| \times |V|$ matrix (i.e., $z^T Y z \geq 0$ holds for any vector $z \in \mathbb{R}^n$). Note that $y_{ii} = x_i^2 = x_i$ due to the binary values of x_i for all $i \in V$. We present a semi-definite programming (SDP) model as

$$\gamma(G) \geq \min tr(Y) \quad (2a)$$

$$\text{s.t. } \sum_{i \in N[v]} y_{ii} \geq 1 \quad \forall v \in V \quad (2b)$$

$$Y \succeq 0, \quad (2c)$$

where $tr(Y)$ is the trace of Y , that is, the sum of all diagonal values of matrix Y , and $Y \succeq 0$ denotes that Y is PSD. One can employ general-purpose convex optimization methods for solving Equation (1b).

Complexity of Domination-Related Parameters of Graphs and Algorithms. Recall that DOMINATING SET is $\mathcal{NP}$ -complete for an arbitrary graph, and $\mathcal{NP}$ -complete for bipartite graphs [21,22] (based on a transformation from DOMINATING SET), and chordal graphs (based on a transformation from EXACT COVER BY 3-SETS [8]). However, it is shown to be in the class $\mathcal{P}$ on trees, interval graphs, directed path graphs, and permutation graphs. In particular, we briefly describe a polynomial *labeling* algorithm for solving TREE DOMINATION [23] on an example as follows:

For a tree $T(V, E)$, we define three types of subsets of V as V_1 , V_2 , and V_3 that consist of *free*, *bound*, and *required* vertices, respectively, such that all the “required” vertices are contained in *any* optimal dominating set S of T . Every “bound” vertex in V_2 is dominated by V_3 , and it is either an element of S or is adjacent to at least one vertex in S . “Free” vertices in V_1 need not be dominated, or need not be in S , but still can be used in S to dominate some bound vertices. We require $V = V_1 \cup V_2 \cup V_3$. To compute the domination number $\gamma(T)$, we seek a partition $V = V_1 \cup V_2 \cup V_3$ associated with an optimal domination through a labeling process. The following observations lead to a polynomial algorithm for solving the domination for trees.

Theorem 16 [24]. *Given a tree T whose vertices are labeled “free” (V_1), “bound” (V_2), and “required” (V_3). Let v be an end-vertex of T that is adjacent to a vertex u , and $T - v$ is the tree obtained by deleting v from T . We have*

1. if $v \in V_1$, then $\gamma(T) = \gamma(T - v)$;
2. if $v \in V_2$ and T' is the tree obtained by relabeling u as “required” in $T - v$, then $\gamma(T) = \gamma(T')$;
3. if $v \in V_3$ and $u \in V_3$, then $\gamma(T) = 1 + \gamma(T - v)$;
4. if $v \in V_3$ and $u \notin V_3$, and T' is the tree obtained by relabeling u as “free” in $T - v$, then $\gamma(T) = 1 + \gamma(T')$.

Now, we root T at an arbitrary vertex, say vertex 1, and for every other vertex $t \in T$, we define $P[t]$ as the parent vertex of t ($P[1] = 0$ since the root vertex then has no parent vertex). Without presenting a rigorous proof of the algorithm correctness and complexity (for which we refer the interested readers to Refs 24 and 5), a solution scheme that solves the minimum cardinality dominating set S of T in $O(n)$ time is

1. Label all vertices $1, \dots, N$ “bound.”
2. For decreasing sequence of vertices $t = N, \dots, 2$, check the following cases: (i) if t is “bound”, label $P(t)$ as “required”; (ii) if t is “required” and $P(t)$ is “bound”, then $S = S \cup \{t\}$, and change $P(t)$ ’s label to “free”; (iii) if t is “required” and $P(t)$ is not labeled “bound,” then only update $S = S \cup \{t\}$.
3. Finally, check root vertex 1. If it is labeled “bound” or “required,” then update $S = S \cup \{1\}$; else, vertex 1 is not included in the dominating set S .

THE DOMINATION CHAIN

Since its first appearance in Ref. 25 in 1978, the domination chain has become a motivation for many studies of the domination and its variants. It reflects the relationship between domination, independence, and irredundance in graphs. We start this section by introducing two concepts regarding an arbitrary property of a general set, that is, *hereditary* and *superhereditary*. We say that a property P of set S is hereditary if every proper subset $S' \subset S$ also has P ; on the other hand, P is superhereditary if every proper superset $S'' \supset S$ has this property. For instance, being a dominating set is superhereditary.

Moreover, we say that a set S is a P -set if P holds for S . A P -set is a *maximal* P -set or a *minimal* P -set if every proper superset $S'' \supset S$ or every proper subset $S' \subset S$ does not have the property P , respectively. Thus, for all hereditary properties P , we are interested in finding the related maximal P -set, and for all superhereditary properties P , the results of minimal P -set are more favorable.

Next, we introduce the definitions of independent and irredundant sets, and explore the relations between the dominating set and the two sets, leading to a fruitful discussion of the domination chain.

Independent Sets and Irredundant Sets

Recall that given $G(V, E)$, a set $S \subset V$ is *independent* if no pair of vertices in S is adjacent. $\beta_0(G)$ is called the *independence number*, which is the maximum cardinality of an independent set in G . In Fig. 2, note that the γ -set (the second minimal dominating set) is an independent set, and we call it an *independent dominating set* of G . The *independent domination number* $i(G)$ is the minimum cardinality of an independent dominating set of G .

Note that being independent is a hereditary property, and thus we are interested in finding maximal independent sets in a graph G . An independent set S is maximal if and only if for all vertices $u \in V - S$, there exists a vertex $v \in S$ such that $uv \in E$. Note that this condition is exactly the definition of a dominating set. Thus, we have

Proposition 1 [2]. *A set S is a maximal independent set if and only if it is independent and dominating.*

Therefore, $i(G)$ also represents the minimum cardinality of a maximal independent set of G . Furthermore, we have

Proposition 2 [2]. *For any graph G , every maximal independent set is a minimal dominating set.*

Recall that $\gamma(G)$ is the minimum cardinality of a dominating set, and $\Gamma(G)$ is the maximum cardinality of a minimal dominating set. According to Propositions 1 and 2, we

describe a part of the domination chain as $\gamma(G) \leq i(G) \leq \beta_0(G) \leq \Gamma(G)$.

Before starting the discussion of the irredundant set, we define the private neighbor set of vertex $v \in S$ as $pn[v, S] = N[v] - N[S - \{v\}]$, that is, all vertices that are *only* adjacent to v but not any other vertex in S . The private neighbor set of S is then the collection of vertices in S that have nonempty private neighbor set, that is, $pn[S] = \{v : pn[v, S] \neq \emptyset\}$. A set S is *irredundant* if for every vertex $v \in S$ such that $pn[v, S] \neq \emptyset$, that is, every vertex in S is adjacent to at least one vertex that other vertices in S cannot reach.

Next, we consider the superhereditary property of being dominating, and focus on the minimal dominating sets. A dominating set of S is minimal if and only if for all vertices $v \in S$, $pn[v, S] \neq \emptyset$. Otherwise, we can exclude v from S , and are still able to dominate G . This condition is exactly the definition of an irredundant set, and similarly, the following two propositions hold.

Proposition 3 [25]. *A set S is a minimal dominating set if and only if it is dominating and irredundant.*

Proposition 4 [26]. *For any graph G , every minimal dominating set is a maximal irredundant set of G .*

Let $ir(G)$ denote the minimum cardinality of a maximal irredundant set, and $IR(G)$ be the maximum cardinality of a maximal irredundant set. Here comes the famous inequality as the domination chain:

Theorem 17 [25]. *For any graph G ,*

$$ir(G) \leq \gamma(G) \leq i(G) \leq \beta_0(G) \leq \Gamma(G) \leq IR(G). \quad (3)$$

Properties of the Domination Chain

Some interesting questions arise regarding Theorem 17, including (i) when do the inequalities in Equation (3) hold as equalities? (ii) Does any other similar inequality chain hold like Equation (3)? (iii) Given a graph G , what is a valid assignment for

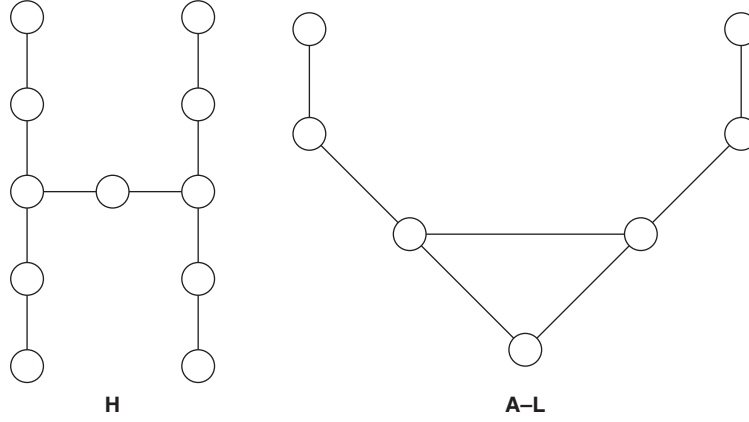


Figure 3. Illustration of H graph and A-L graph.

values of parameters in Equation (3)? Next, we answer the three questions by reviewing some elementary results in literature.

Regarding question (i), two examples of graph G have been created to illustrate the cases when $ir(G) < \gamma(G)$, denoted as H [27] and A-L [28] in Fig. 3, respectively.

Theorem 18 [29]. *If G is a chordal graph containing no H or A-L as subgraphs, then $ir(G) = \gamma(G)$.*

Haynes *et al.* [5] also review many other conditions under which $ir(G) = \gamma(G)$. Moreover, considering $\beta_0(G)$, $\Gamma(G)$, and $IR(G)$, similar results can be derived as

Theorem 19 [30]. *If G is a bipartite graph, a chordal graph, or a graph that has no isolated vertices and $\gamma(G) + IR(G) = n$, then $\beta_0(G) = \Gamma(G) = IR(G)$.*

Next, regarding question (ii), consider one variant as the *edge* domination, independence, and irredundance. $F \subseteq E$ is called an *independent edge set* if no two edges in F share a common vertex. $\beta_1(F)$ and $i'(F)$ denote the maximum and minimum cardinality of maximal independent edge sets, respectively. Also, $F \subseteq E$ is a *dominating set* if all vertices in V are incident to at least one edge in F , and $\gamma'(F)$ and $\Gamma'(F)$ are the minimum and maximum cardinality of minimal dominating edge sets, respectively. Lastly, $F \subseteq E$ is an *irredundant set* if every edge in F has a

private neighbor edge, and $ir'(G)$ and $IR'(G)$ are the minimum and maximum cardinality of maximal irredundant sets. Haynes *et al.* [5] conclude that the same inequality chain holds for all the foregoing parameters.

Theorem 20 [5]. *For any graph G ,*

$$ir'(G) \leq \gamma'(G) \leq i'(G) \leq \beta_1(G) \leq \Gamma'(G) \leq IR'(G). \quad (4)$$

In particular, we have $\gamma'(G) = i'(G)$ for any graph G according to Ref. 28. Finally, considering question (iii), the following result considers the parameter validity of Equation (3).

Theorem 21 [31]. *For any graph G ,*

1. $ir(G) = 1$ implies $\gamma(G) = 1$ and $i(G) = 1$;
2. $\beta_0(G) = 1$ implies $\Gamma(G) = 1$ and $IR(G) = 1$;
3. $\gamma(G) \leq 2ir(G) - 1$.

We refer the interested readers to Ref. 5 for a thorough discussion on other variants of the domination chain and the corresponding results in terms of graph parameters.

Complexity of Parameters of the Domination Chain

Recall that the DOMINATING SET, that is, finding the domination number $\gamma(G)$, is $\mathcal{NP}$ -complete as stated in the section

titled “The Dominating Set Problem.” This section reviews the algorithmic complexity of computing other parameters in the domination chain. First, finding the irredundance number $ir(G)$ (IRREDUNDANT SET) on general graphs is $\mathcal{NP}$ -complete [32], and this result remains true on chordal graphs. However, it is polynomial solvable on any tree structure [33]. Moreover, Garey and Johnson [8] become the first to show that finding the independent domination number $i(G)$ (INDEPENDENT DOMINATING SET) is $\mathcal{NP}$ -complete on general graphs, and even on bipartite graphs [34] or in an edge domination version [35]. Farber [36] provides polynomial algorithms for both domination and independent domination on strongly chordal graphs.

Moreover, although finding the independence number $\beta_0(G)$ (INDEPENDENT SET) belongs to $\mathcal{NP}$ -completeness on general graphs [8], it can be solved in polynomial time when restricted to bipartite graphs, or chordal graphs [5]. Owing to Theorem 19, finding the upper domination number $\Gamma(G)$ and the upper irredundance number $IR(G)$ are also polynomially solvable on bipartite graphs and chordal graphs, but remain $\mathcal{NP}$ -complete on general graphs.

To emphasize, all parameters of Equation (3) are obtainable in polynomial time on trees. Recall that Ref. 24 is the first to give an $O(n)$ algorithm for solving $\gamma(T)$. We refer readers to Refs 33, 37, and 38 for the first linear algorithms for solving $ir(T)$, $i(T)$, and $\beta_0(T)$, respectively. Finally, owing to Theorem 19, $\Gamma(T)$ and $IR(T)$ are then polynomially solvable on trees as well [38].

VARIANTS OF DOMINATION

To solve practical real-world problems, we design domination-related problems by combining domination with another graph property P . For instance, we require domination along with independence or irredundance, of which we discuss the results in the last section. Some other interesting properties are naturally imposed on the dominating set S , including that S (i) has no isolated vertices, (ii) is connected, (iii) includes a complete graph, (iv) forms a perfect matching, and so on. More conditions can also be imposed on

the dominated set $V - S$, or on the dominating method, including multiple domination, strong and weak domination, and so on. Next, we discuss details of the domination with certain graph properties.

Total Dominating Set

Recall that $S \subseteq V$ being a dominating set is equivalent to $N[S] = V$. In some applications, we might be interested in those dominating sets that all elements in S are adjacent to at least one of the other elements in S , that is, no vertex in S is an isolate of S . We define it as a *total dominating set* of V if $N(S) = V$, and the total dominating set of G with the minimum cardinality is denoted as a γ_t -set.

Define the *boundary* of S as $B(S) = \{v \in S : |N[v] \cap S| = 1\}$, and the *open boundary* of S is $OB(S) = \{v \in S : |N(v) \cap S| = 1\}$. The minimal total dominating set can be characterized by

Theorem 22 [5]. *A dominating set S is minimal if and only if $B(S)$ dominates S . Similarly, a total dominating set S is minimal if and only if $OB(S)$ dominates S .*

We propose the following bounds for the total domination number $\gamma_t(G)$.

Theorem 23 [39]. *If G is a connected graph with $n \geq 3$, then $\gamma_t(G) \leq 2n/3$.*

Theorem 24 [39]. *If a graph G has no isolated vertices, then $\gamma_t(G) \leq n - \Delta(G) + 1$. If a graph G is connected and $\Delta(G) < n - 1$, then $\gamma_t(G) \leq n - \Delta(G)$.*

TOTAL DOMINATING SET is $\mathcal{NP}$ -complete on general graphs [40], bipartite graphs [40], and undirected graphs [41]. But it is in the class $\mathcal{P}$ on trees [41], interval graphs [42], and distance-hereditary graphs [43].

Connected Dominating Set

Sampathkumar and Walikar [44] first define a *connected dominating set* S , such that all elements in S are connected. $\gamma_c(G)$ is called the *connected domination number*. It is trivial to see that only connected graphs have a

connected dominating set, and the connected dominating set is a special case of the total dominating set. Thus, $\gamma(G) \leq \gamma_t(G) \leq \gamma_c(G)$ for any connected graph G with $\Delta(G) < n - 1$. According to Theorem 5, we give an upper bound for determining $\gamma_c(G)$ as follows:

Theorem 25 *If G is connected and $n \geq 3$, then $\gamma_c(G) = n - \epsilon_F(G) \leq n - 2$.*

Let $l(T)$ be the number of end-vertices of a tree T . The following corollary is straightforward according to Theorem 24.

Corollary 1 [44]. *If T is a tree with $n > 3$, $\gamma_c(T) = n - l(T)$.*

We propose several upper bounds in terms of the minimum and the maximum degrees $\delta(G)$ and $\Delta(G)$ as follows:

Theorem 26 [44,45]. *For any connected graph G , we have*

1. *if $\delta(G) \geq 3$, then $\gamma_c(G) \leq 3n/4 - 2$;*
2. *if $\delta(G) \geq 4$, then $\gamma_c(G) \leq 3n/5 - 2$;*
3. *$n/(\Delta(G) + 1) \leq \gamma_c(G) \leq 2m - n$, and equality is achieved at the lower bound if and only if $\Delta(G) = n - 1$, and is achieved at the upper bound if and only if G is a path;*
4. *$\gamma_c(G) \leq n - \Delta(G)$.*

Now considering $\gamma(G) \leq \gamma_t(G) \leq \gamma_c(G)$, the following theorem provides one scenario in which the equality is achieved.

Theorem 27 [46]. *If G is the complement of a bipartite graph or a split graph, we have $ir(G) = \gamma(G) = \gamma_t(G) = \gamma_c(G)$.*

A further variation of the connected dominating set is *weakly connected domination*, that is, for a dominating set S , the subgraph $G[S]_w = (N[S], E_w)$ weakly induced by S is connected, where E_w is the set of all edges with at least one incident vertex in S . Recent results of bounds for $\gamma_w(G)$ can be found in Refs 47 and 48. Swaminathan [48] also states necessary and sufficient conditions for

a graph G to have unique minimal weakly connected dominating set.

Finally, we review the algorithmic complexity of CONNECTED DOMINATING SET as follows. It has been shown in the class of $\mathcal{NP}$ -Complete on general graphs [32], bipartite graphs [32], and split graphs (and thus chordal graphs) [46], but in the class of $\mathcal{P}$ for distance-hereditary graphs [49]. It has very close relations with *Steiner tree* problem, and is thus $\mathcal{NP}$ -complete on many graphs. We refer readers to Ref. 50 for a linear-time algorithm for solving a minimum weighted connected dominating set and a minimum vertex-weighted Steiner tree problem on distance-hereditary graphs.

Dominating Cliques

A dominating set is called a *dominating clique* if all vertices of a dominating set S form a complete graph (i.e., every vertex is adjacent to every other vertex in S). The domination clique number is denoted by $\gamma_{cl}(G)$. Since a clique is a special case of connected graphs, we have

$$\gamma(G) \leq \gamma_t(G) \leq \gamma_c(G) \leq \gamma_{cl}(G). \quad (5)$$

In particular, if $\gamma(G) = 1$, then $\gamma_t(G) = \gamma_c(G) = \gamma_{cl}(G) = 1$. Also, the results of $\gamma_{cl}(G)$ are interesting mostly for chordal graphs. Kratsch *et al.* [51] characterize the chordal graphs that have dominating cliques.

Theorem 28 [51]. *Given a chordal graph G , it has a dominating clique if and only if $\text{diam}(G) \leq 3$.*

Theorem 29 [51]. *Given a strongly chordal graph G that has a dominating clique, we have*

1. $\gamma(G) = \gamma_{cl}(G)$;
2. $\gamma(G) = \gamma_t(G) = \gamma_c(G) = \gamma_{cl}(G)$, if $\gamma(G) \geq 2$.

The dominating clique problem is $\mathcal{NP}$ -complete on general graphs, but can be solved in polynomial time on strongly chordal graphs and undirected path graphs [52].

Paired-Dominating Sets

Another type of the domination is defined for the presence of perfect matching. A *paired-dominating set* is a dominating set S such that its induced subgraph has a perfect matching. From a practical perspective, the paired-domination is applicable and important for the following case [53]. Given a graph G , we need to assign guards to protect all vertices. The locations of all guards form a dominating set. If a total dominating set is required, each guard should be protected by at least one of other guards. In a paired-dominating set, we even further require each guard protected by his/her exclusive partner, and vice versa. Hence, the paired-dominating set is a special case of the total dominating set, and we have $\gamma(G) \leq \gamma_t(G) \leq \gamma_p(G)$ where $\gamma_p(G)$ is the paired-domination number. Moreover, we state the following two representative bounds for $\gamma_p(G)$.

Theorem 30 [53]. *For any G having no isolated vertices, $\gamma_p(G) \geq n/\Delta(G)$.*

Theorem 31 [53]. *For any G having no isolated vertices, $\gamma_p(G) \leq 2\gamma(G)$; if $\gamma_p(G) = 2\gamma(G)$, then every γ -set is an i -set.*

The paired-domination problem is $\mathcal{NP}$ -complete on general graphs [53]. Qiao *et al.* [54] describe a linear algorithm for solving paired-domination on tree structures. Polynomial algorithms have also been given on interval graphs, circular-arc graphs [55], and permutation graphs [56].

Multiple Domination

The following sections focus on varied dominations in terms of dominating methods. First, we are interested in *multiple domination* such that given a dominating set S , each vertex in $V - S$ is dominated by at least k vertices in S where k is a given positive integer. The concept was first introduced in Ref. 57, to support the idea of network robustness. That is, deleting a vertex in a normal dominating set might lead to a non-dominated case, while a vertex failure in a multiple domination would still be tolerable for dominating the graph afterward. We

denote the *k-domination number* $\gamma_k(G)$ as the minimum cardinality of an *k-dominating set*, such that every vertex in $V - S$ is dominated by at least k vertices in S . Note that $\gamma_1(G) = \gamma(G)$, and for $1 \leq i \leq j$, if S is a j -dominating set, it is also an i -dominating set, and thus $\gamma_i(G) \leq \gamma_j(G)$.

Before further exploring relations between $\gamma_k(G)$ and $\gamma(G)$, we present the following observation.

Theorem 32 [57]. *Given a γ -set S of G , at least one vertex in $V - S$ is dominated by no more than two vertices in S .*

According to this, if $\Delta(G) \geq 3$ and $k \geq 3$, no k -dominating set for G can be a minimum dominating set, and the following strict inequality holds

Theorem 33 [57]. *If G is a graph with $\Delta(G) \geq 3$ and $k \geq 3$, then $\gamma(G) < \gamma_k(G)$.*

A further variant is the *k-tuple dominating set*, which is obtained by modifying the definition of k -domination. A set S is *k-tuple dominating set* if each vertex in V is dominated by at least k vertices in S , which is different from the requirement by k -domination where only every vertex in $V - S$ needs to be dominated by at least k vertices in S . $\gamma_{\times k}(G)$ is the k -tuple domination number. It follows the definitions that, for any graph G

$$\begin{aligned} \gamma(G) &\leq \gamma_k(G) \leq \gamma_{\times k}(G) \text{ and} \\ \gamma(G) &= \gamma_1(G) = \gamma_{\times 1}(G). \end{aligned} \quad (6)$$

A special case of k -tuple domination is *double domination* when $k = 2$, which has been shown in the class of $\mathcal{P}$ on trees [58]. Liao and Chang [59] further provide a polynomial algorithm for solving the k -tuple domination on strongly chordal graphs, and show that the problem remains $\mathcal{NP}$ -complete on split and bipartite graphs.

Strong and Weak Domination

In 1996, Sampathkumar and Latha [60] first introduced the concepts of strong and weak domination. Given two adjacent vertices u and v , we say that u *strongly dominates* v if

$\deg(u) \geq \deg(v)$, and u weakly dominates v if $\deg(u) \leq \deg(v)$. A set $S \subseteq V$ is a *strong dominating set/weak dominating set* if every vertex in $V - S$ is strongly/weakly dominated by at least one vertex in S . $\gamma_S(G)$ and $\gamma_W(G)$ are the *strong domination number* and *weak domination number*, respectively. Since either strong or weak dominating set is a special case of the dominating set, $\gamma(G) \leq \gamma_S(G)$ and $\gamma(G) \leq \gamma_W(G)$, but no inequality relation can be drawn between $\gamma_S(G)$ and $\gamma_W(G)$. Letting $\bar{G}$ to be the complement graph of G , we propose the strong and weak domination properties in terms of the graph degrees as follows:

Theorem 34 [60]. *For any graph G , the following inequalities hold*

1. $\gamma(G) \leq \gamma_S(G) \leq n - \Delta(G)$;
2. $\gamma(G) \leq \gamma_W(G) \leq n\delta(G)$;
3. $\gamma_S(\bar{G}) + \gamma_W(G) \leq n + 1$.

However, no inequality relation can be drawn between $\gamma_S(G)$, $\gamma_W(G)$ and $i(G)$, $\gamma_c(G)$, and $\alpha_0(G)$ [60]. Last, both STRONG and WEAK DOMINATING SET problems are $\mathcal{NP}$ -complete on general graphs, and even on bipartite graphs. But they are in the class $\mathcal{P}$ when applied to trees [5], and the labeling algorithm is again employed to solve the problem.

In addition to all domination-related problems previously mentioned, we refer readers to Ref. 5 for more variants, including *dominating cycles*, *restrained domination*, *locating domination*, *distance domination*, *directed graph domination*, and so on. Also, some new domination variants have recently been introduced because of their convenience for describing practical problems, for example, domination with exponential decay [61] and rainbow domination [62]. The rainbow domination is $\mathcal{NP}$ -complete on most graphs, including bipartite and split graphs. In a very recent paper, Chang *et al.* [63] provide a polynomial-time algorithm for solving the rainbow domination on tree structures.

APPLICATIONS

The domination has received substantially large attention in the past three decades due

to its wide usage in real-world applications. We next review its variants that serve as theoretical fundamentals for several practical problems.

Domination in Wireless Ad hoc Networks and Approximation Algorithms

A mobile ad hoc network is a wireless and infrastructureless communication network that contains a set of local sensors. A sensor receives data and sends them to others that are within the signal transmission range according to some routing protocols. Williams and Camp [64] comprehensively evaluate and compare performances of many protocols under different network conditions. In general, most protocols feature on-demand routing, and require flooding of route requests by using local broadcasts.

Ad hoc networking-related problems have drawn substantially great interests because of their importance in military, disaster relief, health care industry, social communications, and so on. However, these networks usually feature dynamic topology, resource limitation (e.g., battery and bandwidth), and a lack of infrastructure. Motivated by the broadcast storm problem [65] caused by flooding used in many routing protocols, as well as the broadcast unreliability issue [66], some wireless network construction papers have suggested the virtual backbone as an alternative to a fixed routing infrastructure, which is also a backup when route is temporarily unavailable [67,68].

The virtual backbone is a cluster placed at certain place in the network, and captures all local changes that happen within the range. As we also need to route messages between clusters, a connected dominating set (CDS) becomes a natural alternative to approximate virtual backbone infrastructure. In literature, a wireless network is usually characterized by a unit disk graph [69], consisting of a set of all identical sensors and a set of edges between paired sensors whose distance is within the communication range.

Computing CDS on a unit disk graph is $\mathcal{NP}$ -hard [70], and thus many works have been conducted on finding approximation algorithms with good performance [71,72]. Three types of algorithms exist

for solving CDS: centralized algorithms [73,74], constant-time (fast) localized distributed algorithms [75–77], and linear-time localized distributed algorithms [78,79], as summarized in Ref. 80. We review some representative works in each category as follows.

Guha and Khuller [74] propose two polynomial-time centralized algorithms that have approximation factors of $2H(\Delta) + 2$ and $H(\Delta) + 2$ from optimum, respectively, where H is a harmonic function and Δ is the maximum node degree. However, centralized algorithms are usually difficult to implement due to the hardness of collecting and maintaining the complete graph topology information. On the other hand, distributed localized algorithms analyze the CDS topology in a local neighborhood, and sequentially pass this information through each node in the network. Bharghavan and Das [81] cast the two centralized algorithms of Guha and Khuller [74] into distributed implementations, whose performance, however, is still restricted to the transmission messages' complexities. Wan *et al.* [79] provide a constant-time approximation for solving CDS in ad hoc networks; several other papers on this concept include Refs 70,82–84. They use a similar two-stage solution scheme that constructs a dominating set in the first phase, and connects it to a CDS in the second phase. Gandhi and Parthasarathy [80] propose two constant-time distributed algorithms that converge in $O(\log^2 n \Delta)$ and in $O(\log^2 n)$ times, respectively, where Δ again represents the maximum degree and n is the number of nodes in the network. To the best of our knowledge, they are the first fast localized distributed approximation algorithms for constructing CDS on wireless ad hoc networks considering loss of messages due to wireless interference. Tiwari *et al.* [85] study the problem on a 2D unit disk graph, as well as a 3D ball graph, and employ coloring methods to derive constant-time approximation algorithms for interference-aware broadcast scheduling in multihop wireless sensor networks. A large amount of literature exists on the creation of polynomial-time approximation algorithms for seeking the minimum CDS in wireless ad hoc networks, and we list a few here: [86–92].

Moreover, the discussion has also been extended to a weighted case, in which we assume that the wireless networks contain different types of sensors with varied data transmission cost. The problem is then formulated on unit disk graphs with node weights [93]. We refer the interested readers to Ref. 94 for a recent survey on study of the minimum CDS in wireless networks. Some other domination variants are also employed in the analysis of mobile ad hoc networks. For instance, Swaminathan [48] implements the weakly connected domination (see the section titled “Connected Dominating Set”) that is attractive in practice since the model would further reduce the number of sensors located.

Domination in Power Network Construction and Social Networks

In addition to constructing wireless ad hoc networks, the domination has also been applied to power network construction. Haynes *et al.* [95] describe a *power domination* problem for the usage of monitoring an electric power system and obtaining the information on its voltage and phase angle. The variant is $\mathcal{NP}$ -complete for general graphs, transformed from 3-SAT, but polynomially solvable on trees.

Kelleher [96] and Kelleher and Cozzens [97] apply the domination to social network graphs, in which they study the relationship between individuals of a group in social activities. These studies also provide another angle to tackle the problem of identifying the independent dominating sets. Let a *status* be a set $W \subseteq V$ of vertices, such that for any two vertices $u, v \in W$, $N(u) \cap (V - W) = N(v) \cap (V - W)$, that is, all the vertices in a status dominate the same set of vertices outside the status. We say that two vertices u and v are *structurally equivalent* if either $N(u) = N(v)$ or $N[u] = N[v]$. Moreover, we have

Theorem 35 [96]. *If a dominating set S of a connected graph G of order $n \geq 2$ is a status, then S is an independent dominating set of cardinality two.*

Theorem 36 [97]. *If a dominating set S of a connected graph G is a structurally equivalent set, then S consists of two independent vertices each of which has degree of $n - 1$.*

We refer the interested readers to Refs 96 and 97 for a detailed proof. On the other hand, socialists are interested in finding every status and every maximal structure equivalent set in a social network graph, which can also be solved by using properties of dominating sets in graphs.

Broadcast Domination Problem

Erwin [98] introduced a *broadcast domination* problem, in which integer powers can be assigned to each vertex, and a vertex with positive power “dominates” itself and all vertices that are at a distance of at most the value of its power assignment away. The *Broadcast domination number* $\gamma_b(G)$ is the minimum sum of all power assigned on each vertex in the dominating set that dominates the graph. Blair *et al.* [99] give a polynomial-time algorithm for solving the problem on trees, interval graphs, and series-parallel graphs. They show that for any proper interval graph G , $\gamma_b(G) = \gamma(G)$.

Given an unweighted and undirected graph, Heggernes and Lokshtanov [100] develop the following polynomial algorithm for solving the optimal broadcast domination problem. For all $i \in V$, define $B(i, f_i)$ as a *ball* of vertex i if assigned with integer power f_i , which corresponds to the set of all vertices that are within a distance of f_i from vertex i . A broadcast domination f is *efficient* if $B(i, f_i) \cap B(j, f_j) = \emptyset$ for all $i, j \in V_f, i \neq j$. For an efficient broadcast domination f on G , define a domination graph $G_f = (V_f, E_f)$ where $E_f = \{(i, j) | N(B(i, f_i)) \cap B(j, f_j) \neq \emptyset\}$. They show that if G is an unweighted graph, there always exists an efficient optimal broadcast domination f' on G . This result leads to the proof that an optimal broadcast always exists such that the degree of any vertex in its associated domination graph $G_{f'}$ is no more than 2, and thus, $G_{f'}$ is either a path or a cycle. These results are the foundation for their polynomial algorithm.

Then, Heggernes and Lokshtanov [100] run a *shortest path* algorithm on a directed auxiliary graph G_u which is polynomially constructed from the original graph G , with respect to each vertex u , such that u belongs to a ball corresponding to one of the endpoint in G_u . Furthermore, we extend the discussion to a weighted case, called the *weighted broadcast domination* problem, in which every edge is associated with a distance value, and a positive integer value of f_v covers some vertex $u \in V$ if and only if $f_v \geq d(u, v)$. We again seek a minimum sum of all integer powers located on a subset of vertices that covers the graph. However, the complexity of the weighted broadcast domination problem remains open [101].

SUMMARY AND OPEN QUESTIONS

In this article, we start with the standard dominating set problem, and discuss domination-related problems in graphs. We relate the domination with independence and irredundance, and state the well-known domination inequality chain. Moreover, the bounds and the algorithmic complexity of each domination variant and the domination chain parameters are discussed in detail. Finally, we describe several applications to which the domination is commonly applied.

The complexity issue associated with many domination-related problems still remains open for many different types of graphs. For instance, the complexity of the independent domination, the weighted domination, and the weighted total domination problems remains open in distance-hereditary graphs [43]. The complexity of the weighted broadcast domination problem is also open for general graphs [101].

Acknowledgments

The author is grateful to Dr. J. Cole Smith, and the anonymous referee(s), whose remarkable suggestions led to an improved version of the article. The author is also thankful to Dr. I. Hicks for his great effort in organizing the article.

REFERENCES

1. de Jaenisch CF. Applications de l'analyse mathématique au jeu des échecs. Petrograd 1862.
2. Berge C. Theory of graphs and its applications. London: Wiley; 1962.
3. Ore O. Theory of Graphs. AMS colloquium publications. Volume 38, American Mathematical Society. 3rd ed. Providence (RI): AMS Bookstore; 1967.
4. Haynes TW, Hedetniemi ST, Slater PJ, editors. Domination in Graphs: Advanced Topics. New York (NY): Marcel Dekker, Inc.; 1998.
5. Haynes TW, Hedetniemi ST, Slater PJ. Fundamentals of domination in graphs. New York (NY): Marcel Dekker, Inc.; 1998.
6. Cockayne EJ, Hedetniemi ST. Towards a theory of domination in graphs. *Networks* 1977;7:247–261.
7. Nieminen J. A method to determine a minimal dominating set of a graph G_0 . *IEEE Trans Circuits Syst* 1974;CAS-21:12–14.
8. Garey MR, Johnson DS. Computers and intractability: A guide to the theory of NP-completeness. New York: W. H. Freeman; 1979.
9. Reed B. Paths, stars and the number three. *Comb Probab Comput* 1996;5:277–295.
10. Caro Y, Roditty Y. On the vertex-independence number and star decomposition of graphs. *Ars Combinatoria* 1985;20:167–180.
11. Caro Y, Roditty Y. A note on the k -domination number of a graph. *Int J Math Sci* 1990;13:205–206.
12. Weber K. Domination number for almost every graph. *Rostock Math Kolloq* 1981;16:31–43.
13. Walikar HB, Acharya BD, Sampathkumar E. Recent developments in the theory of domination in graphs. Volume 1, MRI Lecture notes in math. Allahabad, India: Mahta Research Institute; 1979.
14. Slater PJ. Locating dominating sets and locating-dominating sets. In: Alavi Y, Schwenk A, editors. Graph Theory, Combinatorics, and Applications, Proceedings of the 7th Quadrennial International Conference on the Theory and Application of Graphs. New York: John Wiley & Sons, Inc.; 1995. pp. 1073–1079.
15. Flach P, Volkmann L. Estimations for the domination number of a graph. *Discrete Math* 1990;80:145–151.
16. Marcu D. A new upper bound for the domination number of a graph. *Q J Math Oxf Ser* 2 1985;36:221–223.
17. Vizing VG. A bound on the external stability number of a graph. *Dokl Akad Nauk SSSR* 1965;164:729–731.
18. Sanchis/ LA. Maximum number of edges in connected graphs with a given domination number. *Discrete Math* 1991;87:65–72.
19. Rubalcaba RR. Fractional domination, fractional packings, and fractional isomorphisms of graphs [PhD thesis]. Auburn University; 2005.
20. Eldar YC. Comparing between estimation approaches: admissible and dominating linear estimators. *IEEE Trans Signal Process* 2006;54:1689–1702.
21. Booth KS, Johnson JH. Dominating sets in chordal graphs. *SIAM J Comput* 1982;11:191–199.
22. Chang GJ, Nemhauser GL. The k -domination and k -stability problems in sun-free chordal graphs. *SIAM J Algebr Discr Methods* 1984;5:332–345.
23. Mitchell SL, Cockayne EJ, Hedetniemi ST. Linear algorithms on recursive representations of trees. *J Comput Syst Sci* 1979;18(1):76–85.
24. Cockayne EJ, Goodman SE, Hedetniemi ST. A linear algorithm for the domination number of a tree. *Inform Process Lett* 1975;4:41–44.
25. Cockayne EJ, Hedetniemi ST, Miller DJ. Properties of hereditary hypergraphs and middle graphs. *Can Math Bull* 1978;21:461–468.
26. Bollobás B, Cockayne EJ. Graph theoretic parameters concerning domination, independence, and irredundance. *J Graph Theory* 1979;3:241–250.
27. Slater PJ. Irreducible point independence numbers and independence graphs. *Congressus Numerantium* 1974;10:647–660.
28. Allan RB, Laskar RC. On domination and independent domination numbers of a graph. *Discrete Math* 1978;23:73–76.
29. Laskar R, Peters K. Domination and irredundance in graphs. Technical Report No. 429 Clemson University: Department of Mathematical Science; 1983.
30. Cockayne EJ, Favaron O, Payan C, Thomason AG. Contributions to the theory of domination, independence and irredundance in graphs. *Discrete Math* 1981;33:249–258.

31. Cockayne EJ, Mynhardt CM. The sequence of upper and lower domination, independence and irredundance number of a graph. *Discrete Math* 1993;122:89–102.
32. Pfaff J, Laskar R, Hedetniemi ST. NP-completeness of total and connected domination, and irredundance for bipartite graphs. Technical Report No. 428. Clemson (SC): Department of Mathematical Science, Clemson University; 1983.
33. Bern MW, Lawler EL, Wong AL. Why certain subgraph computations require only linear time. *Proceedings of the 26th Annual IEEE Symposium on the Foundations of Computer Science*; Portland, OR; 1985. pp. 117–125.
34. Corneil DG, Perl Y. Clustering and domination in perfect graphs. *Discrete Appl Math* 1984;9:27–39.
35. Yannakakis M, Gavril F. Edge dominating sets in graphs. *SIAM J Appl Math* 1980;38:264–272.
36. Farber M. Domination, independent domination, and duality in strongly chordal graphs. *Discrete Appl Math* 1984;7:115–130.
37. Beyer TA, Proskurowski A, Hedetniemi ST, *et al.* Independent domination in trees. *Congressus Numerantium* 1977;19:321–328.
38. Daykin DE, Ng. CP. Algorithms for generalized stability number of tree graphs. *J Aust Math Soc* 1966;6:89–100.
39. Cockayne EJ, Dawes RM, Hedetniemi ST. Total domination in graphs. *Networks* 1980;10:211–219.
40. Pfaff J, Laskar R, Hedetniemi ST. Linear algorithms for independent domination and total domination in series-parallel graphs. *Congressus Numerantium* 1984;45:71–82.
41. Laskar RC, Pfaff J, Hedetniemi SM, *et al.* On the algorithmic complexity of total domination. *SIAM J Algebra Discr Methods* 1984;5:420–425.
42. Keil JM. Total domination in interval graphs. *Inform Process Lett* 1986;22:171–174.
43. Chang MS, Wu SC, Chang GJ, *et al.* Domination in distance-hereditary graphs. *Discrete Appl Math* 2002;116:103–113.
44. Sampathkumar E, Walikar HB. The connected domination number of a graph. *J Math Phys Sci* 1979;13:607–613.
45. Kleitman DJ, West DB. Spanning trees with many leaves. *SIAM J Discrete Math* 1991;4:99–106.
46. Laskar R, Pfaff J. Domination and irredundance in split graphs. Technical Report No. 430. Department of Mathematical Science, Clemson University; Clemson, SC; 1983.
47. Lemanńska M, Patyk A. Weakly connected domination critical graphs. *Opusc Math* 2008;28(3):325–330.
48. Swaminathan V. Weakly connected domination in graphs. *Electron Notes Discr Math* 2009;33:67–73.
49. D'Atri A, Moscarini M. Distance-hereditary graphs, Steiner trees, and connected domination. *SIAM J Comput* 1988;17:521–538.
50. Yeh HG, Chang GJ. Weighted connected domination and Steiner trees in distance-hereditary graphs. *Discrete Appl Math* 1998;87:245–253.
51. Kratsch D, Damaschke P, Lubiw A. Dominating cliques in chordal graphs. *Discrete Math* 1994;128:269–275.
52. Kratsch D. Finding dominating cliques efficiently, in strongly chordal graphs and undirected path graphs. *Discrete Math* 1990;86:225–238.
53. Haynes TW, Slater PJ. Paired-domination in graphs. *Networks* 1998;32:199–206.
54. Qiao H, Kang LY, Cardei M, *et al.* Paired-domination of trees. *J Glob Optim* 2003;25:43–54.
55. Cheng TCE, Kang LY, Ng CT. Paired-domination on interval and circular-arc graphs. *Discrete Appl Math* 2007;155:2077–2086.
56. Cheng TCE, Kang LY, Shan E. A polynomial-time algorithm for the paired-domination problem on permutation graphs. *Discrete Appl Math* 2009;157:262–271.
57. Fink JF, Jacobson MS. n -domination in graphs. In: Alavi Y, Chartrand G, Lick DR, *et al.* editors. *Graph theory with applications to algorithms and computer science*. Kalamazoo, MI; John Wiley & Sons, Inc.; 1985. pp. 283–300.
58. Liao C, Chang GJ. Algorithmic aspect of k -tuple domination in graphs. *Taiwan J Math* 2002;6:415–420.
59. Liao C, Chang GJ. k -tuple domination in graphs. *Inform Process Lett* 2003;87:45–50.
60. Sampathkumar E, Latha LP. Strong weak domination and domination balance in a graph. *Discrete Math* 1996;161:235–242.
61. Dankelmann P, Day D, Erwin D, *et al.* Domination with exponential decay. *Discrete Math* 2009;309:5877–5883.

62. Bresar B, Henning MA, Rall DF. Paired-domination of cartesian product of graphs and rainbow domination. *Electron Notes Discr Math* 2005;22:233–237.
63. Chang GJ, Wu J, Zhu X. Rainbow domination on trees. *Discrete Appl Math* 2010;158:8–12.
64. Williams B, Camp T. Comparison of broadcasting techniques for mobile ad hoc networks. *Proceedings of the 3rd ACM International Symposium on Mobile ad hoc Networking and Computing*. Lausanne, Switzerland; 2002. pp. 194–205.
65. Ni S-Y, Tseng Y-C, Chen Y-S, *et al.* The broadcast storm problem in a mobile ad hoc network. *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MOBI-COM)*. Seattle (WA); 1999. pp. 151–162.
66. Sinha P, Sivakumar R, Bharghavan V. Enhancing ad hoc routing with dynamic virtual infrastructures. *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, Volume 3. Anchorage, AK; 2001. pp. 1763–1772.
67. Das B, Sivakumar R, Bharghavan V. Routing in ad hoc networks using a virtual backbone. *Proceedings of the 6th International Conference on Computer Communications and Networks (ic3N'97)*. Montreal, Quebec, Canada; 1997. pp. 1–20.
68. Liang B, Haas ZJ. Virtual backbone generation and maintenance in ad hoc network mobility management. *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. Tel-Aviv, Israel; 2000.1293–1302.
69. Clark BN, Colbourn CJ, Johnson DS. Unit disk graphs. *Discrete Math* 1990; 86:165–177.
70. Cardei M, Cheng X, Cheng X, *et al.* Connected domination in multihop ad hoc wireless networks. North Carolina; *Proceedings of the 6th International Conference on Computer Science and Informatics (CS&I'2002)*. 2002.
71. Orecchia L, Panconesi A, Petrioli C, *et al.* Localized techniques for broadcasting in wireless sensor networks. *Proceedings of DIALM-POMC*. 2004. pp. 41–51.
72. Stojmenovic I, Seddigh M, Zunic J. Dominating sets and neighbor elimination-based broadcasting algorithms in wireless networks. *IEEE Trans Parall Distrib Syst* 2002;13(1):14–25.
73. Cheng X, Huang X, Li D, *et al.* A polynomial-time approximation scheme for the minimum-connected dominating set in ad hoc wireless networks. *Networks* 2003; 42:202–208.
74. Guha S, Khuller S. Approximation algorithms for connected dominating sets. *Algorithmica* 1998;20(4):374–387.
75. Dai F, Wu J. An extended localized algorithm for connected dominating set formation in ad hoc wireless networks. *IEEE Trans Parall Distrib Syst* 2004;15(10):908–920.
76. Wu J. Dominating-set-based routing in ad hoc wireless networks. *Handbook of wireless networks and mobile computing*. New York: John Wiley & Sons, Inc.; 2002. pp. 425–450.
77. Wu J. Extended dominating-set-based routing in ad hoc wireless networks with unidirectional links. *IEEE Trans Parall Distrib Syst* 2002;13(9):866–991.
78. Alzoubi KM, Wan P-J, Frieder O. New distributed algorithm for connected dominating set in wireless ad hoc networks. *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35'02)*; Hawaii; 2002.
79. Wan P-J, Alzoubi KM, Frieder O. Distributed construction of connected dominating set in wireless ad hoc networks. *Monet* 2004;9(2):141–149.
80. Gandhi R, Parthasarathy S. Distributed algorithms for connected domination in wireless networks. *J Parall Distrib Comput* 2007;67:848–862.
81. Bharghavan V, Das B. ad hoc routing using minimum connected dominating sets. *Proceedings of the IEEE International Conference on Communications*. Montreal, Quebec, Canada; 1997. pp. 376–380.
82. Funke S, Kesselman A, Meyer U. A simple improved distributed algorithm for minimum CDS in unit disk graphs. *ACM Trans Sens Netw* 2006;2:444–453.
83. Funke S, Kesselman A, Kuhn F, *et al.* Improved approximation algorithms for connected sensor cover. *Wirel Netw* 2007;13: 153–164.
84. Wan P-J, Wang L, Yao F. Two-phased approximation algorithms for minimum cds in wireless ad hoc networks. *Proceedings of the 28th International Conference on Distributed Computing Systems*. Beijing, China; 2008. pp. 337–344.
85. Tiwari R, Dinh TN, Thai MT. On approximation algorithms for interference-aware

- broadcast scheduling in 2d and 3d wireless sensor networks. Volume 5682, Wireless algorithms, systems, and applications, Lecture notes in computer science. Berlin/Heidelberg, Germany: Springer; 2009. pp. 438–448.
86. Chen Z, Qiao C, Xu J, *et al.* A constant approximation algorithm for interference aware broadcast in wireless networks. Proceedings of the IEEE Conference on Computer Communications (INFOCOM). Anchorage, AK; 2007.
 87. Gandhi R, Kim Y-A, Lee S, *et al.* Approximation algorithms for data broadcast in wireless networks. Proceedings of the IEEE Conference on Computer Communications (INFOCOM). Rio de Janeiro, Brazil; 2009.
 88. Kim D, Wu Y, Li Y, *et al.* Constructing minimum connected dominating sets with bounded diameters in wireless networks. IEEE Trans Parall Distrib Syst 2009; 20(2):147–157.
 89. Li D, Liu L, Yang H. Minimum connected r -hop k -dominating set in wireless networks. Discrete Math Algorithms Appl 2009; 1(1):45–57.
 90. Li M, Wan P-J, Yao FF. Tighter approximation bounds for minimum CDS in wireless ad hoc networks. Proceedings of the 20th International Symposium on Algorithms and Computation (ISAAC 2009). Hawaii; 2009. pp. 677–686.
 91. Nieberg T, Hurink J. A PTAS for the minimum dominating set problem in unit disk graphs. Volume 3879, Proceedings of the Third Workshop on Approximation and Online Algorithms, Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer; 2005. pp. 296–306.
 92. Nieberg T, Hurink J, Kern W. Approximation schemes for wireless networks. ACM Trans Algorithms 2008;4(4):1–17.
 93. Zou F, Wang Y, Li X, *et al.* New approximations for minimum-weighted dominating set and minimum-weighted connected dominating set on unit disk graphs. Theor Comput Sci 2010. In press. DOI: 10.1016/j.tcs.2009.06.022.
 94. Wu W, Gao X, Pardalos PM, Du D-Z. Wireless networking, dominating and packing. Optim Lett 2010;4(3):347–358.
 95. Haynes TW, Hedetniemi SM, Hedetniemi ST, *et al.* Domination in graphs applied to electric power networks. SIAM J Discr Math 2002;15(4):519–529.
 96. Kelleher LL. Domination in graphs and its application to social network theory [PhD thesis]. Northeastern University; 1985.
 97. Kelleher LL, Cozzens MB. Dominating sets in social network graphs. Math Soc Sci 1988;16:267–279.
 98. Erwin DJ. Dominating broadcasts in graphs. Bull Inst Comb Appl 2004;42:89–105.
 99. Blair JRS, Heggernes P, Horton S, *et al.* Broadcast domination algorithms for interval graphs, series-parallel graphs, and trees. Congressus Numerantium 2004;169:55–77.
 100. Heggernes P, Lokshtanov D. Optimal broadcast domination in polynomial time. Discrete Math 2006;306:3267–3280.
 101. Lokshtanov D. Broadcast domination [Master's thesis]. University of Bergen; 2007.

DRAMA THEORY

JAMES BRYANT

Professor of Operational Research &
Strategy Sciences, Sheffield
Business School at Sheffield
Hallam University, Sheffield, UK

Drama theory is a theory about how people resolve their differences and was developed to inform strategic interaction: that is, interaction in situations where there is interdependence in the choices made by the individuals involved. It is one of a number of frameworks that developed from game theory in response to the paradoxes that follow from adherence to a strict doctrine of instrumental rationality. Its origins lie in Howard's [1] metagame analysis of which it retains two key features. First, it is a positive approach. That is, it is neither purely formal nor normative: assumptions about the "players" are taken to be empirical statements, testable against actual conditions. Second, it is a nonquantitative approach: numerical utilities, for example, are not used. In addition to these features, drama theory directly addresses the paradoxes of rationality revealed by Howard's original work. In doing so, it takes a different route from alternative derivatives of metagame analysis—the Graph Model [2] and the Theory of Moves [3], for example—since they persist in seeking a "solution" to an interaction within a structure where players' preferences and options are fixed. Drama theory recognizes instead that confrontations are practically resolved by appealing to the wider context within which the "small world" of formal game-theoretic models is set.

Drama theory can be applied to real-life situations involving collaboration or conflict: indeed, the former can too readily morph into the latter. It provides support to people in such situations by offering a language that they can use to represent how those involved say they would like a situation resolved and what they would feel they would have to

do if they cannot reach a solution. From this formulation, the pressures that each party faces can systematically be explored, and hence guidance on the management of the interaction can be developed. There is always a range of ways that the pressures can be handled and, while drama theory tells us about the implications of each of these, it does not attempt to tell people which to choose: that is a matter of personal choice and style. In a competitive situation, these insights may help someone gain advantage. However, often, drama theory is used to help several parties reach an agreement.

Drama theoretic analysis has been carried out of confrontations in international relations, in corporate mergers and alliances, in departmental coworking, and in human relations management. Analyses have been used to create focused role-play training exercises, permitting participants to rehearse new roles or relationships. They have also been used to identify the source of tensions in inter- and intraorganizational relationships and to suggest potential ways of reducing them. Like game theory, the framework is completely general in the way that it distils the essence of a situation and it makes no assumptions about the purpose or the locus of analysis.

"PLAY" IN GAMES AND DRAMA

The word "play" relates to both games and drama. In each arena, it refers to the actions of the collection of parties who drive a situation forward.

Therefore, in a game, several "players" have and make choices as to what they should do. These choices may be constrained by the rules or conventions under which their interaction with others is bound to take place, but the responsibility for making them lies firmly with the players themselves. In deciding what to do they will take account of what other players involved in the situation might simultaneously or subsequently determine to do. Game theory models these

decisions, normally assuming that players seek to attain some objective—perhaps to maximize the benefit for themselves—and that in their interaction they make rational choices that will help them to achieve this aim.

By contrast, situations are “played out” in drama by “characters,” who encountering resistance by, or fearing duplicity in, other characters with whom they are in conflict or with whom they are trying to cooperate, send messages to them about the things that they might or might not do. They communicate to warn or convince the other parties, and thereby transform the shared predicament, so that its prospect represents less of a challenge to their own aspirations. Such strategic communications—strategic in the sense that they are informed by a deliberate intention to change characters’ current stances—precede the eventual deliberation that each character must give (possibly informed by game theory) as to what it shall actually do in the situation. Drama theory models this pre-play communication, where the game to be eventually played is itself negotiated. Using Savage’s [4] distinction between decision making in “small worlds” (where you can “look before you leap”) and in “large worlds” (where you “cross that bridge when you come to it”), game theory is the precise, small-world technology that says how a player should act in a game, whereas drama theory illuminates the large world within which this game-theoretic decision making occurs. They are essential complements to each other.

“Play” and “pre-play” and the corresponding relationship between game theory and drama theory can be clearly illustrated by an initial example involving two characters, Alf and Bet. Suppose that Bet is interested in buying a used car that Alf has for sale. They meet and through discussion get an understanding of the outcome that each is seeking. As often happens in such cases, they disagree on the price. If they cannot agree then Bet fails to win the car and Alf has to find another buyer: neither wants this. Alf is in a quandary as to whether to accept Bet’s low offer, while Bet faces the predicament that Alf may refuse to drop the price he demands. Alf’s hints that Bet should accept his asking

price: if she does not then he is sorry but he will not be able to close the deal. Bet advises Alf to take the amount that she has offered for the car: if he does not then she will look elsewhere. Clearly they face an impasse. Moreover, neither wishes to yield.

The game theoretic approach is to look for an equilibrium: an outcome that no player can better through its own actions, given what others are doing (if anyone thought they could gain advantage by choosing differently, then rationally they would do so). However, in the game being played by Alf and Bet, there is no unique equilibrium. It is actually an example of the game called *Chicken* to which there are two equilibria. This can be readily seen. If Alf stubbornly holds out for a high price, then Bet is better off accepting than walking away from the deal; while if Bet refuses to yield then Alf is better off reducing his price than losing a potential buyer. Either of these outcomes (Alf “wins” or Bet “wins”) could be reached from the impasse but there is no way of telling which. Now had one player been prepared to concede to the other (and the other had known this), then the first could take advantage of the situation, indicate refusal to give way, and so win the game. So there is a risk that each may attempt to force stabilization at its own preferred equilibrium, but in so doing could steer the outcome toward the worst prospect, “no deal.”

Drama theory assumes that the characters communicate before they act. This is not a strong assumption since it includes implicit (e.g., by deeds or attitudes) as well as explicit communication. What characters communicate is their *stands*. There are three elements in a character’s stand:

- *Its position.* which is its proposed solution to the situation. This includes statements about what it intends to do and what it would have others do.
- *Its stated intentions.* what it says it will do, given everyone’s positions and stated intentions.
- *Its expressed doubts.* the doubts that it now has about others’ stated intentions, and the doubts that it would have were any declared position to be achieved.

Exchanges between characters are about these elements, since these tell the other characters what a communication is meant to achieve, and how (i.e., understanding others' stands enables a character to interpret the strategic relevance of what they are saying). Of course, any character may be lying or be disbelieved—drama theory makes no assumptions to the contrary—but this is not relevant to the requirements of the model. Nevertheless, characters' stands are what they try to make credible.

MODELING INTERACTION IN DRAMA THEORY

Return to Alf and Bet. Their stands are modeled in drama theory using a frugal notation originally developed for metagame analysis [1]. The *options board* of Table 1 summarizes all the elements required.

At the left side, the two characters are named and their respective *options*—the things which are for them to decide—are listed. The columns in the main body of the table summarize possible future states of affairs. So in the column headed “Alf’s Position” is his proposal that Bet should agree to the high price he is demanding, while he refuses to accept her low offer: Bet’s contrary position is shown in the next column. Note that both of these require specific actions from each character. The column on the left of these includes the stated intentions of both characters (these can all be written in one column as a character’s intentions just relate to the options about which it has discretion): Alf will refuse to accept Bet’s low offer, and Bet will refuse to agree to Alf’s high price.

Alf and Bet’s expressed doubts are shown by question marks in the table. These are justified as follows. If the characters were in the impasse shown by the leftmost column, then either of them or both together do better not to carry out their stated intentions. Since this is common knowledge (i.e., each knows that the other knows it, etc.), the doubts shown are explained. Incidentally, the modeling does not attempt to distinguish degrees of doubt: either an intention is doubted or it is not.

The options board model used in drama theory does not have an exact counterpart in game theory. While its elements represent the components of a game in the normal form, there is no assumption in the drama theoretic model about whether or not characters’ choices are made simultaneously and independently. Such information is not provided by the options board; nor is the possibility of chance moves.

DILEMMA ANALYSIS

The table also introduces a key feature of drama theoretic modeling: the statement of the characters’ *dilemmas*. Indeed, the analysis of confrontations using drama theory has sometimes been termed *dilemma analysis* because of the centrality of this concept. Dilemmas arise from the attempt by characters to act rationally. There are three generic dilemmas, each named for what it makes it hard for a character to do. Each can face a character *with* another character *over* a particular option. They are as follows:

- *Persuasion Dilemma*. A character faces a persuasion dilemma with another character over an option if it has no doubt that the other character will flout its position on that option.
- *Rejection Dilemma*. A character has a rejection dilemma with another character over an option if the other character doubts that the first character will flout the latter’s position on that option.
- *Trust Dilemma*. A character has a trust dilemma with another character over an option if it doubts whether the other character will support its position on that option.

The persuasion and rejection dilemmas come into play when the characters have incompatible positions; the trust dilemma relates to a situation when they are in agreement. Further, the first two dilemmas may occur in one of two so-called *modes* depending upon whether it is a position or an intention that gives rise to them.

Table 1. Option Board: Car-Purchase Example—Confrontation

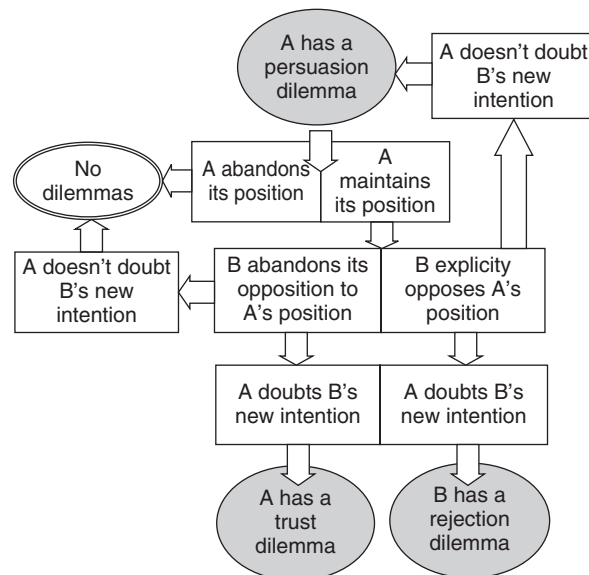
	Stated Intentions	Alf's Position	Bet's Position	Alf's Dilemma(s)	Bet's Dilemma(s)
Alf					
Accept low offer	NOT accept low offer	Not accept low offer	~		
Bet					
Agree to high price	??? Agree to high	??? Agree to high	??? Agree to high	Trust	

Looking again at Table 1, persuasion dilemmas in position mode have been noted against each character in the two rightmost columns. How are these explained? Alf's Per(p) dilemma arises because (as presently assumed and encapsulated in the options board) he has no doubt that Bet will not agree to a high price. This contradicts his position that she should do so. The dilemma facing Bet can be understood in the same way. The Rejection dilemmas facing each character arise because each doubts that if it came to a confrontation the other would stand firm.

The identification of dilemmas does not in itself show how a situation may be resolved, but it is an essential step in this direction. A fundamental proposition of drama theory is that characters will seek to eliminate the

dilemmas that they face. Dilemmas cause discomfort and tension and raise the emotional temperature of an interaction. As tempers rise, characters try to escape from the conceptual cul-de-sac at which they are stuck and they notice possibilities that might help them achieve their ambitions. Alternatively, the frustrations of the impasse may cause them to act in a counter-rational manner (i.e., against their declared preferences) and, for example, either press harder for their own position or resign themselves to acceding to the other party's solution. In every case, though, the "game" is transformed.

Let us consider Alf's treatment of his persuasion dilemma with Bet over her refusal to agree to the high (to her) price for the car. What can he do? Alternatives available to him are shown schematically in Fig. 1.

**Figure 1.** Dealing with a persuasion dilemma.

Drama theory does not prescribe which of these pathways might be followed: it is a matter of personal choice, perhaps reflecting the personality of the character concerned. Suppose that he maintains his position, then he will try to get Bet to abandon her opposition. This might be done, for example, by making clearer to her the advantages of her doing so; or by stressing the potential disadvantages of not doing so. The former might involve pointing out that the car is ready to drive away, comes with some “extras” that she may not have appreciated, or is a bargain compared with other cars of a similar age and condition. The latter might mean reminding her that there are few cars of this type presently on the market at this price and that she will be losing a bargain to pass it over. Imagine that Bet is persuaded by such blandishments. Then Alf’s dilemma disappears, and indeed Bet also loses her persuasion dilemma if the option board becomes that shown in Table 2.

However, note that a new dilemma has appeared. The (not unreasonable) assumption has been made that, although Bet has agreed to the price Alf is asking, he is not yet sure that she will honor this promise. This is depicted in the table by the question marks appearing against Bet’s option, signifying that this decision is doubted by Alf. The tilde (~) in the first row of her position column incidentally means that her position on Alf’s action is now undeclared: she does not say one way or the other. To eliminate this trust dilemma, Alf and Bet will need further exchanges. One commonplace tactic in the present case might be, for instance, for Alf to request, or for Bet to offer a deposit to show her good faith in the bargain. The precise route followed is clearly open to

the characters to determine for themselves; what is significant is that the drama theory provides a rationale for and explanations of these routes.

THE DYNAMICS OF INTERACTION: EPISODIC DEVELOPMENT

It might be thought that this is the end of the affair between Alf and Bet, but this is not so. Recall that it was stated earlier that drama theory and game theory are complementary, the former being relevant to pre-play communication and the latter to action choices in play. Supposing all dilemmas have been eliminated between characters as just outlined for Alf and Bet, then they have still to take the actions upon which they have agreed: Alf to sell the car for the agreed price and Bet to provide her deposit as a precursor to making the purchase at the amount settled between them. But this is the point at which each must act independently to implement these decisions. And as Bet reflects alone (or possibly with a partner, but in any event away from Alf), she may size up the agreement differently and conclude that the car is not such a bargain after all. If so, then she may renege on the deal, possibly even risking losing her deposit in the process. In the same way, Alf may be approached by a third party who makes a better offer than the one he has concluded with Bet, and he may feel like finding an excuse for withdrawing from the settlement with Bet. He might even use game theory to evaluate this possibility, as shown in Table 3.

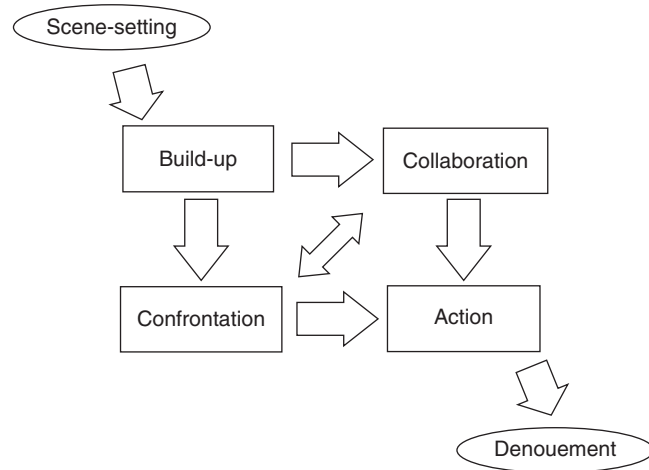
If the payoffs were seen as shown, then Alf is clearly better off selling to the third party, regardless of what Bet decides. However, Bet

Table 2. Option Board: Car-Purchase Example—Cooperation

	Stated Intentions	Alf's Position	Bet's Position	Alf's Dilemma(s)	Bet's Dilemma(s)
Alf					
Accept low offer	??? NOT accept low	NOT accept low offer	Accept low offer	Rej (t)	Per (p)
Bet					
Agree to high price	??? NOT agree to	Agree to high price	NOT agree to high	Per (p)	Rej (t)

Table 3. Option Board: Car-Purchase Example—Alf's Game

In each cell (Payoff to Alf, Payoff to Bet)		Bet's Choices	
		Withdraw	Honor Deal
Alf's Choices	Sell to third party	4,3	3,1
	Sell to Bet	1,2	2,4

**Figure 2.** A “Dramatic” episode.

does not have a corresponding choice “no-matter-what.” If Alf does decide to sell to the third party, then Bet’s best decision is to withdraw (and at least not lose face by foolishly honoring the deal).

The overall process of interaction just described can be depicted as in Fig. 2. Earlier episodes shape the context for interaction within which certain issues must be settled by certain characters. These characters explore each other’s views in a tentative manner and thereby create a common reference frame (they are then able to communicate unambiguously). If the communicated intentions are compatible, then there is potential collaboration: otherwise there is confrontation. As has been shown in the car-selling example, confrontation may turn into collaboration; but of course the reverse is also true. Eventually, parties commit to cooperate or to engage in conflict and take actions that may or may not reflect these declared commitments. The actions lead to the denouement: an irreversibly altered environment. The episodic process shown in Fig. 2 is just one fragment of the larger tree comprising antecedent and possible successor episodes.

Before leaving the example involving Alf and Bet, it is important to draw out another key assumption of drama theory. This is that movements to eliminate dilemmas are accompanied by specific emotions. The corollary is that emotion plays an integral role in enabling characters to adjust their positions, intentions, or doubts. Returning briefly to Alf’s persuasion dilemma: if he takes the encouraging road of cajoling Bet into accepting a higher price because of benefits that she had not previously appreciated, then he should do this with positive emotion (in a friendly, generous way) rather than in a bullying, browbeating manner. Drama theory associates specific types of emotion—positive, negative, or neutral—with the dissolution of dilemmas along different pathways. This aspect of the theory was first explored by Bennett and Howard [5] and has been developed in subsequent work.

FORMAL TREATMENT OF DRAMA THEORY

The following statement of the theory is based on Howard [6]: The central concept in drama

theory is that of a frame. A *frame*, F , is a structure having these elements:

- a set C (called the *Cast*) of characters;
- a set A of actions available to the players in C ;
- a function $O: A \rightarrow C$ from the set of actions to the set of characters (i.e., it tells us who “owns” each action);
- for each character $c \in C$:
 - *Position* p_c . this is a subset of A such that $p_c \cap A(k) \neq \emptyset$ for all $k \in C$: that is, c ’s position intersects with every character’s set of available actions. It is the set of actions that c can countenance;
 - *Intentions* n_c . these are a subset of $A(c)$ such that $n_c \cap A(c) \neq \emptyset$: that is, c ’s intentions intersect with each member of its set of available actions. It is the set of its own actions that c says it is considering;
 - *Doubted Actions* d_c . this is a subset of A such that $d_c \cap A(c) \neq n \cap A(c)$ for all $A(c)$: that is, the actions which c doubts must not coincide with its intentions. It is the set of actions about which c is sceptical of the assertions (to accept or implement) made by others.

The triple (p_c, n_c, d_c) is called the *stance* of the character.

These definitions allow us to define three dilemmas by defining three sets of dilemmatic actions:

- Character c ’s *Persuasion Dilemma*, $Per(c, k)$, with character k corresponds to the set of actions $n_k - d_c - p_c$. This is the set of actions that k says it is considering but that c cannot countenance (they are not in its position) and about which c does not doubt k ’s assertion to implement.
- Character c ’s *Rejection Dilemma*, $Rej(c, k)$, with character k corresponds to the set of actions $n_c \cap d_k - p_k$. This is the set of actions that c says it is considering but that k cannot countenance (they are not in its position) and furthermore

about which k doubts c ’s assertion to implement.

- Character c ’s *Trust Dilemma*, $Tru(c, k)$, with character k corresponds to the set of actions $A_k \cap d_c - n_k - p_c$. This is the set of actions that k claims it has excluded but which c believes it is considering and which c cannot countenance (they are not in its position).

Characters are motivated to eliminate their dilemmas because, unless they do so, they must assume that either their stated intentions are disbelieved or their positions will be flouted. If all dilemmas are eliminated, stated intentions implement every position and every player assumes that its position will be implemented.

The fundamental theorem of drama theory is that if and only if all three dilemma sets are empty for all c, k then:

1. Stated intentions implement every position (i.e., $n \subseteq p$, where $p = \cap p_c$).
2. Intentions assumed by each character implement that character’s position (i.e., $(n - d_c) \cup (d_c - n) \subseteq p_c$ for each c).

There is a corollary. That if all dilemmatic sets are empty, then, provided the set A of actions is nonempty, p is nonempty (i.e., all positions are compatible). That is, a necessary and sufficient condition for a complete trusting agreement is the absence of dilemmas.

During the period of pre-play communication, one frame succeeds another as characters facing dilemmas are motivated to find and communicate emotions and rationalizations that enable them to escape these dilemmas. They are able to do so to the extent both that the then current model permits and to the extent that opportunities for reframing are recognized in the real-world situation.

APPLICATIONS

Drama theory can be applied in a number of ways. Clearly, it provides a basis for analyzing interactions. Such analysis might be used to inform or support one or more of

the parties involved: it would be expected that analysis should help parties handle their interactions more effectively. The results of analysis might also be used to capture the essentials of a situation so as to provide a “snapshot” of its strategic state: this could, for example, be useful as a means of briefing an incoming manager and taking over an ongoing business relationship. However, such a summary might alternatively be used to shape a simulation or role play for the purposes of training or development. Personnel could be exposed in an artificial environment to encounters or group situations that are tightly defined (in terms of the dilemmas that they embody) for which they would be able to test and extend their repertoire of responses. These applications will be examined briefly in turn below.

Drama–Theoretic Analysis

Drama–theoretic analysis has been carried out of situations that are both “real” and fictitious. One example of the former is given by Bryant [7], who studied confrontations in the delivery of UK health services; a contrasting example appears in Howard [8], which despite its necessary fictionalization is based upon actual military operations in a postwar zone. Howard [9] also gave a light-hearted but completely rigorous analysis of the endings of two popular films (“Pulp Fiction” and “Reservoir Dogs”) to demonstrate the subtlety of understanding possible with this form of analysis. Levy *et al.* [10,11] have shown how the theory could be used to help manage global hazards and disasters. A large number of what may best be referred to as *fragments of analysis* have been shared on the dedicated, web-based, drama theory forum at <http://www.dilemmasgalore.com>, which shows how nonexperts are using drama theory to investigate and manage live management challenges. Many of these involve the analysis of interactions between individuals rather than organizations, and can be characterized in organizational terms as relating to human resources issues (e.g., bullying at work). Others have been of topical events. The practice of relying upon news media reports as the basis of the latter has been defended on the grounds that the media

report the stories in which the protagonists see themselves as involved: parties may act so as to try to change this story, but the stories themselves are not “cooked up” by anyone as no one is “in charge.” Of course, that does not mean that the news story should not be augmented by intelligence data, but it usually suffices to model the broad strategic confrontations.

Mediation is a special case that merits further examination. When two (or more) parties are in dispute and a mediator becomes involved in helping them bridge their differences, it might be thought that a drama theory model could be created to assist them. While this is true, it cannot be done in a direct manner. This is because such a model would need to incorporate the very differences and doubts that the parties must overcome and so will not share with each other. However, what a mediator can do is to employ a model, based upon explorations with each party, to identify questions that need to be put to each side if matters are to be resolved. For instance, rather than asking a party if what they are proposing (whether it be a threat or a promise) is credible, the mediator asks the other party if it finds the proposal credible: and if it does not, then the burden is on the first party to make its proposal credible. The analysis specifically highlights the presence of incredible or undeterrable threats and promises and so provides the perfect tool for a mediator.

In many situations, it is appropriate to analyze related interactions taking place at a number of different levels: for example, the transactions between front-line staff and customers, between staff within a department, between functional or product departments, and between organizations within a partnership, may all need to be understood and in some sense aligned. The notion of an enterprise confrontation management system (ECMS) has been proposed [12] so that the above can be better accomplished, representing a civil version of the C2CC system developed for the military [13].

Immersive Drama

Completed analyses provide the foundations for the purposeful construction of simulations

and role-plays. Most directly, the role-players take on the parts of the “real” characters in the analyzed situation. Their briefing documents consist of “their” character’s option boards (though possibly presented in a less spartan format) supplemented by other background information of a more conventional form. The former provide them with an immediate prompt for their interactions with other role-players, since they must seek to resolve their dilemmas, and to do this they have to engage in strategic communication with others. A full discussion of the creation and running of such a developmental role-play is given in Bryant and Darwin [14].

SOFTWARE SUPPORT FOR DRAMA ANALYSIS

Drama analysis for whatever purpose can be a time-consuming and demanding activity. Fortunately, some software tools have now appeared (e.g., Confrontation Manager™ software [15]) to ease this burden. These enable users to represent complex situations with the options board notation, so that a hierarchy of confrontations is captured. The software also identifies the dilemmas facing characters, provides an explanation for them, and suggests ways in which these dilemmas might be eliminated. While Confrontation Manager was produced in collaboration with the defense sector, it is quite general in the way that situations are modeled.

REFERENCES

1. Howard N. Paradoxes of rationality: Theory of metagames and political behavior. Cambridge, MA: MIT Press; 1971.
2. Kilgour DM, Hipel KW, Fang L. The graph model for conflicts. *Automatica* 1987; 23(1): 41–55.
3. Brams SJ. The theory of moves. Cambridge, UK: Cambridge U.P.; 1994.
4. Savage L. The foundations of statistics. New York: Wiley; 1951.
5. Bennett P, Howard N. Rationality, emotion and preference change: drama-theoretic

- models of choice. *Eur J Oper Res* 1996; 92: 603–614.
6. Howard N. Drama theory as a theory of pre-game communication and equilibrium selection. Brighton, UK: ISCO Ltd.; 2007.
7. Bryant JW. Confrontations in health service management: Insights from drama theory. *Eur J Oper Res* 2002; 142: 610–624.
8. Howard N. Confrontation analysis: How to win operations other than war. Washington, DC: CCRP Publications, Department of Defense; 1999.
9. Howard N. Negotiation as drama: how ‘games’ become dramatic. *Int Negot* 1996; 1: 125–152.
10. Levy JK, Hipel KW, Howard N. Advances in drama theory for managing global hazards and disasters. Part I: Theoretical foundation. *Group Decis Negot* 2009; 18(4): 303–316.
11. Levy JK, Hipel KW, Howard N. Advances in drama theory for managing global hazards and disasters. Part II: Coping with global climate change and environmental catastrophe. *Group Decis Negot* 2009; 18(4): 317–334.
12. Bryant J, Howard N. Achieving strategy coherence. In: O’Brien FA, Dyson RG, editors. *Supporting strategy: Frameworks, methods and models*. Chichester, UK: John Wiley & Sons; 2007. pp. 55–86.
13. Stubbs L, Howard N, Tait A. How to model a confrontation—computer support for drama theory. *Proceedings of 1999 Command and Control Research and Technology Symposium*, Naval War College, Newport, RI, 29 June–1 July, 1999.
14. Bryant JW, Darwin J. Exploring inter-organisational relationships in the health service: an immersive drama approach. *Eur J Oper Res* 2004; 152: 655–666.
15. Idea Sciences. Confrontation manager user manual. Washington DC: Idea Sciences; 2005.

FURTHER READING

- Bryant J. The six dilemmas of collaboration: Inter-organisational relationships as drama. Chichester, UK: John Wiley & Sons; 2003.
- Howard N. Drama theory and its relation to game theory. Part 1: Dramatic resolution vs. rational solution & Part 2: Formal model of the resolution process. *Group Decis Negot* 1994; 3: 187–206; 207–235.

DTMCs WITH COSTS AND REWARDS

SUSAN XU

Department of Supply Chain
and Information Systems,
Smeal College of Business,
Penn State University,
University Park,
Pennsylvania

DTMCs WITH COSTS AND REWARDS

In many applications, it is of great relevance to associate a reward, often in monetary terms, to a state occupancy or to a state transition of a discrete-time Markov chain (DTMC). A DTMC with an attached reward structure is called the *DTMC with rewards* [1, Vol. 2]. When the amounts of rewards are negative, we have a DTMC with costs. A DTMC with rewards or costs can be used to study problems such as corporate profits and portfolio performance, but as we will illustrate by examples in the subsequent sections, such models can also be broadly applied to many other types of problems.

Denote $\{X_n, n \geq 0\}$ as a *homogeneous* DTMC with the one-step transition probability matrix $P = \{p_{ij}\}$ and the discrete state space S (see Ref. 2 for treatments of nonhomogeneous Markov and semiMarkov reward processes). Suppose that each occupancy of state i is associated with some random reward R_i with the expected reward $E[R_i] = r_i, i \in S$. The system will receive a sequence of rewards as the Markov chain evolves over time. We assume that rewards are uniformly bounded; that is $|r_i| \leq B < \infty$, for all $i \in S$.

We shall consider several commonly used performance measures in Markov reward models in the subsequent subsections.

Expected Total Reward

In this section, we consider two performance measures in turn. The first performance measure is the expected total reward over a

finite number of transitions, and the second performance measure is the expected total reward the system receives until the process enters a designated (e.g., an absorbing) state.

We denote the expected total reward of a DTMC over a fixed-time interval n , given the process starts in state $X_0 = i$, by $g_i(n)$, which can be expressed as

$$g_i(n) = E \left[\sum_{k=0}^n R_{X_k} | X_0 = i \right]. \quad (1)$$

Denote column vectors $r = \{r_i, i \in S\}$ and $g(n) = \{g_i(n), i \in S\}$, respectively, as the expected occupancy reward vector and the expected total reward (over $n + 1$ periods) vector. Denote also $M(n) = \{m_{ij}(n), i, j \in S\}$, where $m_{ij}(n)$ represents the expected number of times the Markov chain, starting in state $X_0 = i$, visits state j during the first $n + 1$ transitions. It is known that $M(n)$ satisfies the following equation [3]:

$$M(n) = \sum_{k=0}^n P^k. \quad (2)$$

Then, it is intuitively appealing and can indeed be shown formally (Theorem 5.11 of Ref. 4), that

$$g_i(n) = \sum_{j \in S} m_{ij}(n) r_j. \quad (3)$$

In matrix form, the previous expression is written as $g(n) = M(n)r$.

Example 1. A periodic-review inventory system with an (s,S) policy. We consider a periodic-review inventory system with lost sales and instantaneous order replenishment. Suppose weekly demands for an item are independently and identically distributed (iid) random variables with the common distribution

$$P(D = d) = p_d, \quad d = 0, 1, 2, \dots \text{ and } \sum_{d=0}^{\infty} p_d = 1.$$

Suppose the system uses a replenishment policy of the (s, S) type: if at the beginning of week n the stock on hand, X_n , is lower than s , then we place an order to bring the inventory level to S ; otherwise, we do not order. Assume also that the system incurs a holding cost of h per unit at the end of a week, a shortage cost of p per unit of unmet demand, a fixed ordering cost of K for placing an order, and a variable cost of c per unit ordered.

The stochastic process $\{X_n, n \geq 0\}$ forms a Markov chain with the state space $\{0, 1, \dots, S\}$, where

$$X_{n+1} = \begin{cases} \max\{X_n - D_n, 0\}, & \text{if } X_n \geq s, \\ \max\{S - D_n, 0\}, & \text{if } X_n < s, \end{cases}$$

where D_n is the demand in week n . We need to distinguish two cases, $i \geq s$ and $i < s$. In the former case, we obtain the transition probabilities of the Markov chain as

$$p_{ij} = \begin{cases} p_{i-j}, & 1 \leq j \leq i, \\ \sum_{d=i}^{\infty} p_d, & j = 0, \end{cases}$$

and for $i < s$,

$$p_{ij} = \begin{cases} p_{s-j}, & 1 \leq j \leq S, \\ \sum_{d=i}^{\infty} p_d, & j = 0. \end{cases}$$

Let us compute the expected total cost the system incurs over the next 12 weeks, assuming the system holds no stock at time 0; that is, $X_0 = 0$. To proceed, we need to compute $m_{0j}(11)$, the expected number of weeks the system has j units of ending inventory during the next 12 weeks, given that it initially has 0 units on hand, for $j \in S$. This quantity can be obtained from Equation (2), $M(11) = \sum_{k=0}^{11} P^k$. The expected weekly cost when the system starts with i units of stock on hand is given by

$$r_i = \begin{cases} K + c(S - i) + hi + p \sum_{d=S}^{\infty} (d - S)p_d, & i < s, \\ hi + p \sum_{d=i}^{\infty} (d - i)p_d, & i \geq s. \end{cases}$$

From Equation (3), the expected total cost of the inventory system over the next 12 weeks becomes

$$\begin{aligned} g_0(11) &= \sum_{j=0}^S m_{0j}(11) r_j \\ &= \sum_{j=0}^{s-1} m_{0j} \left(K + c(S - j) + hj \right. \\ &\quad \left. + p \sum_{d=S}^{\infty} (d - S)p_d \right) \\ &\quad + \sum_{j=s}^S m_{0j} \left(hj + p \sum_{d=j}^{\infty} (d - j)p_d \right). \end{aligned}$$

Next, we consider the expected total reward the system receives until it enters a final state. In this case, since the time until the process reaches the final state is random, the duration that the process accumulates rewards is also random. Without loss of generality, we designate the final state as 0. Since the problem does not depend on what happens after the final state is reached, we can treat state 0 as an absorbing state. That is, we modify the transition probabilities by setting $p_{00} = 1, p_{0j} = 0$, for $j > 0$, and leaving p_{ij} unchanged for all $i > 0$ and all j .

Let g_{i0} be the expected total reward received by the system, starting in state i until the first entrance of the final state 0, with $v_{00} = 0$. Since g_{i0} is simply the present reward r_i plus the expected total reward from whatever state is entered in the next period, we can obtain g_{i0} by solving simultaneous equations:

$$g_{i0} = r_i + \sum_{j \in S} p_{ij} g_{j0}, \quad i \in S. \quad (4)$$

These simultaneous equations, with $g_{00} = 0$, has a unique solution (Lemma 2, p. 124 of Ref. 5). This method is known as the *first-step analysis*, because we condition on the first step, viz $X_1 = j$, that the system visits.

We demonstrate the use of Equation (4) in the following example.

Example 2. Expected first passage time. One problem of interest in the study of stochastic processes is to compute the expected first passage time from states i to j , which is the expected number of transitions the process takes for its first entrance from state i to state j . In this example, we consider a discrete-time queue that can accommodate at most N customers. Suppose we wish to find the expected total time all customers spent in the system, starting at an integer time when i customers are present and ending at the time when the system first becomes empty. From Little's Law, this total waiting time is the sum of the numbers of customers in the system, summed over the first passage time from state i to state 0. We let $r_i = i$ be the "reward" in state i , and let g_{i0} be the expected total waiting time of all customers until the system becomes empty, starting at a time when there are i customers in the system, with $g_{00} = 0$. Once the transition probability matrix P of this discrete-time queue is specified, we can use Equation (4) to obtain the unique solution of the simultaneous linear equations:

$$g_{i0} = i + \sum_{j=0}^N p_{ij} g_{j0}, \quad i = 1, \dots, N.$$

If the total rewards received in a Markov reward model are infinite as the number of transitions tends to infinity, it becomes appropriate to consider a performance measure such as the *discounted reward* or the *long-run average reward*. We shall consider these two measures in the following two sections.

Discounted Reward

In discounted reward models, rewards received in future periods are discounted by a discount factor $0 < \beta < 1$. Define g_i^β as the expected total β -discounted reward received over an infinite horizon, given $X_0 = i$. Then,

$$g_i^\beta = r_i + E \left[\sum_{k=1}^{\infty} \beta^k (R_{X_k} | X_0 = i) \right].$$

Conditioning on the next state the process enters in period 1 (the first-step analysis), we obtain

$$\begin{aligned} g_i^\beta &= r_i + E \left[\sum_{k=1}^{\infty} \beta^k (R_{X_k} | X_0 = i) \right] \\ &= r_i + \beta \sum_{j \in S} \left(r_j + E \left[\sum_{k=2}^{\infty} \beta^{k-1} (R_{X_k} | X_1 = j) \right] \right) \\ &\quad \times p_{ij} \\ &= r_i + \beta \sum_{j \in S} \left(r_j + E \left[\sum_{k=1}^{\infty} \beta^k (R_{X_k} | X_0 = j) \right] \right) \\ &\quad \times p_{ij} \\ &= r_i + \beta \sum_{j \in S} g_j^\beta p_{ij}. \end{aligned}$$

Note that we have not made any assumptions on transition probability matrix P . This means that the above expression holds for any Markov chain; that is, we do not require the Markov chain to be irreducible and positive recurrent.

Let $g^\beta = \{g_i^\beta, i \in S\}$ be the column vector corresponding to the discounted rewards g_i^β . In matrix form, the previous expression is given by

$$g^\beta = r + \beta P g^\beta, \quad (5)$$

or, equivalently,

$$g^\beta = (I - \beta P)^{-1} r, \quad (6)$$

where I is the identity matrix and matrix $(I - \beta P)$ is invertible if $0 < \beta < 1$.

Example 3. A Vehicle Insurance Problem. A transport firm has an insurance contract for a fleet of vehicles it owns. The premium payment is due at the beginning of each year. There are three possible premium classes with a premium payment of P_j for class j , with $P_3 < P_2 < P_1$. The premium class of the firm in a given year depends on its premium class and claim history during the past year. In case no damage is claimed in the past year and the previous premium is P_j , the next premium is P_{j+1} (with $P_4 = P_3$, by convention); otherwise, the highest premium P_1 is due. Since the insurance contract is for a whole fleet of vehicles, the transport firm has obtained the option to decide only

at the end of the year, whether the accumulated damage during that year should be claimed or not. Under a given claim limit policy $(\alpha_1, \alpha_2, \alpha_3)$, the transport firm claims only damages larger than α_j when the current premium class is j . In case a claim is made, the insurance company compensates the accumulated damage minus an own risk, which amounts to d_j for premium class j . The total damages in successive years are iid random variables with a common cumulative distribution function F and a density function f . The transport firm is interested in determining the expected discounted total payment of premiums under the claim limit policy $(\alpha_1, \alpha_2, \alpha_3)$.

Define X_n as the premium class for the transport firm at the beginning of year n . The stochastic process $\{X_n, n \geq 0\}$ is a DTMC with three possible states $j = 1, 2, 3$. The one-step transition probability matrix of this Markov chain works out as

$$P = \begin{bmatrix} 1 - F(\alpha_1) & F(\alpha_1) & 0 \\ 1 - F(\alpha_2) & 0 & F(\alpha_2) \\ 1 - F(\alpha_3) & 0 & F(\alpha_3) \end{bmatrix}. \quad (7)$$

The expected cost per year associated with premium class j is given by

$$r_j = P_j + \int_0^{\alpha_j} x f(x) dx + \beta d_j (1 - F(\alpha_j)). \quad (8)$$

In the previous expression, we see that the expected annual costs of the firm in premium class j include the annual premium of class j , the expected damage absorbed by the firm by not filing the claims when the accumulated damage is less than α_j , and the deductible when the accumulated damage exceeds the claim limit α_j . The expected deductible given in the last term of the above expression is discounted by β to reflect the timing of the payment. From Equation (5), we can compute the discounted payment of the firm by solving the following simultaneous equations:

$$\begin{aligned} g_1^\beta &= r_1 + \beta \left[(1 - F(\alpha_1)) g_1^\beta + F(\alpha_1) g_2^\beta \right] \\ g_2^\beta &= r_2 + \beta \left[(1 - F(\alpha_2)) g_1^\beta + F(\alpha_2) g_3^\beta \right] \\ g_3^\beta &= r_3 + \beta \left[(1 - F(\alpha_3)) g_1^\beta + F(\alpha_3) g_3^\beta \right]. \end{aligned}$$

Average Reward

The average reward per unit time is a commonly used criterion to measure the reward rate associated with a Markov reward model, which is defined as

$$g_i = \lim_{n \rightarrow \infty} \frac{1}{n+1} E \left[\sum_{k=0}^n R_{X_k} | X_0 = i \right].$$

For expositional simplicity, we only consider irreducible and positive recurrent Markov chains. Again, we assume that rewards are uniformly bounded, $|r_i| \leq B < \infty$. Let the column vector $\pi = \{\pi_i, i \in S\}$ be the stationary distribution of the DTMC, satisfying $\pi = \pi P$ and $\sum_{i \in S} \pi_i = 1$. To compute g_i , we note

$$\begin{aligned} g_i &= \lim_{n \rightarrow \infty} \frac{1}{n+1} E \left[\sum_{k=0}^n R_{X_k} | X_0 = i \right] \\ &= \lim_{n \rightarrow \infty} \frac{1}{n+1} \left[\sum_{k=0}^n \sum_{j \in S} r_j p_{ij}^{(n)} \right] \\ &= \sum_{j \in S} \left[\lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n p_{ij}^{(n)} \right] r_j \\ &= \sum_{j \in S} \pi_j r_j, \end{aligned}$$

where $p_{ij}^{(n)} = P(X_{n-1} = j | X_0 = i)$. The interchange of limits and summations in the third step is justified because $\sum_{i \in S} \pi_i = 1$ and rewards are uniformly bounded. Notice that the average reward g_i is independent of initial state i . Therefore, the average reward per unit time satisfies

$$g = \sum_{j \in S} \pi_j r_j. \quad (9)$$

We illustrate the use of Equation (9) with two examples.

Example 3. (continued). Consider the DTMC model of the vehicle insurance problem as described in Example 3. Suppose the firm now wishes to determine its long-run average annual payment under the claim limit policy $(\alpha_1, \alpha_2, \alpha_3)$. Since the DTMC is irreducible and positive recurrent, we can

compute its stationary distribution by $\pi = \pi P$ and $\sum_{i=1}^3 \pi_i = 1$. The steady-state probabilities are

$$\begin{aligned}\pi_1 &= \frac{1 - G(\alpha_3)}{1 - G(\alpha_3) + G(\alpha_1) - G(\alpha_1)G(\alpha_3) + G(\alpha_1)G(\alpha_2)} \\ \pi_2 &= \frac{(1 - G(\alpha_3))G(\alpha_1)}{1 - G(\alpha_3) + G(\alpha_1) - G(\alpha_1)G(\alpha_3) + G(\alpha_1)G(\alpha_2)} \\ \pi_3 &= \frac{G(\alpha_1)G(\alpha_2)}{1 - G(\alpha_3) + G(\alpha_1) - G(\alpha_1)G(\alpha_3) + G(\alpha_1)G(\alpha_2)}.\end{aligned}$$

Finally, with the cost expression given in Equation (8), we can use Equation (9) to determine the long-run average annual payment of the transport firm.

In many applications, it is also common to associate rewards with state transitions instead of state occupancies. In this case, let r_{ij} denote the expected reward associated with a transition from state i to state j . Then, the expected reward associated with a transition from state i is $r_i = \sum_j p_{ij} r_{ij}$. We illustrate this concept with the following example.

Example 4. Taxicab Fare. Suppose a metropolitan area is partitioned into three zones, with states 1, 2, and 3 representing zones 1, 2, and 3, respectively. Denote $\{X_n, n \geq 0\}$ as a Markov chain, where X_n is the location of a taxicab after n trips. The probability that a customer boarding in state i needs a ride to state j is p_{ij} , and the fare for such a trip is r_{ij} . For this example, the transition matrix and transition reward matrix are given by

$$P = \{p_{ij}\} = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 0 \\ 3/4 & 1/4 & 0 \end{bmatrix} \quad \text{and} \quad \{r_{ij}\} = \begin{bmatrix} 10 & 20 & 40 \\ 20 & 20 & 40 \\ 40 & 40 & - \end{bmatrix},$$

respectively. We are interested in computing the average fare per trip. To proceed, we first note that the expected fare associated with a ride starting from state i is $r_i = \sum_j p_{ij} r_{ij}$. From the above matrices, we obtain $r = (r_1, r_2, r_3) = (20, 25, 40)$. The stationary distribution of the Markov chain can be solved by $\pi = \pi P$ and $\sum_{i=1}^3 \pi_i = 1$, which yields the solution $(\pi_1, \pi_2, \pi_3) = (11/20, 5/20, 4/20)$. Therefore, the average fare per trip is given by

$$g = \sum_{j=1}^3 \pi_j r_j = 1010/40 = \$25.25.$$

For further readings of discrete-time Markov reward models and their applications, the reader is referred to Refs 2, 4–8, among others.

REFERENCES

1. Howard RA. Dynamic probabilistic systems. New York: John Wiley & Sons Inc.; 1971.
2. Janssen J, Manca R. Applied semi-Markov processes. New York (NY): Springer; 2006.
3. Wolff RW. Stochastic modeling and the theory of queue. Englewood Cliffs (NJ): Prentice-Hall; 1989.
4. Kulkarni VG. Modeling, analysis, design and control of stochastic systems. Springer; New York (NY): 1999.
5. Gallager RG. Discrete stochastic processes. Norwell, MA: Kluwer Academic Publishers; 1995.
6. Kulkarni VG. Modeling and analysis of stochastic systems. London (UK): Chapman Hall; 1995.
7. Ross SM. Introduction to probability models. 9th ed. San Diego: Academic Press; 2006.
8. Tijm HC. Stochastic models: an algorithmic approach. New York: John Wiley & Sons Inc.; 1994.

DUAL SIMPLEX

MIHAI BANCUIU
School of Management,
Bucknell University,
Lewisburg, Pennsylvania

Consider the following linear programming (LP) problem P expressed in standard form:

$$\begin{aligned} \text{[P]} \quad & \min \mathbf{c}^T \mathbf{x} \\ & \text{subject to: } \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \geq 0, \end{aligned}$$

where $\mathbf{c}^T = (c_1, c_2, \dots, c_n) \in \mathbf{R}^n$ is an n -dimensional cost vector, $\mathbf{A} = (a_{ij}) \in \mathbf{R}^{m \times n}$ is a matrix of constraint coefficients, $\mathbf{b} = (b_1, b_2, \dots, b_m)^T \in \mathbf{R}^m$ is an m -dimensional right hand side (RHS) vector, and $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbf{R}^n$ is an n -dimensional vector of decision variables. We refer to the total cost $z = \mathbf{c}^T \mathbf{x} : \mathbf{R}^n \mapsto \mathbf{R}$ as the objective value of problem P . Stated in this form, P is called the *primal* problem, and the vector $\mathbf{x}^*$ that minimizes the objective function $\mathbf{c}^T \mathbf{x}^*$, while satisfying the equality condition $\mathbf{Ax}^* = \mathbf{b}$, is called the *optimal solution* to the problem P . Associated with the primal problem P , there exists a dual problem D , which is formulated in the following way:

$$\begin{aligned} \text{[D]} \quad & \max \mathbf{b}^T \mathbf{u} \\ & \text{subject to: } \mathbf{A}^T \mathbf{u} \leq \mathbf{c} \\ & \mathbf{u} \text{ free.} \end{aligned}$$

Here, the vector $\mathbf{u} = (u_1, u_2, \dots, u_m) \in \mathbf{R}^m$ is called the vector of *dual variables*. Notice that formulating the dual problem of D will yield the original primal problem P . The relationship between problems P and D is established by the following two theorems (the proofs of these and any subsequent result presented in this article can be found in any introductory LP textbook (see Refs 1, 2; **LP Duality and KKT Conditions for LP**).

Theorem 1 [Weak Duality]. *If $\mathbf{x}$ is a feasible solution of P , and $\mathbf{u}$ is a feasible solution of D , then $\mathbf{b}^T \mathbf{u} \leq \mathbf{c}^T \mathbf{x}$.*

Theorem 2 [Strong Duality]. *If problem P has an optimal solution, then problem D also has an optimal solution. Moreover, the respective optimal solution values are equal to one another.*

An immediate consequence of weak duality is that if the primal problem P is unbounded from below, then its dual D must be infeasible, and correspondingly, if the dual problem D is unbounded from above, then the primal P must be infeasible. Additionally, if the choice of $\mathbf{x}$ and $\mathbf{u}$ is such that the inequality described in the weak duality theorem becomes an equality, then $\mathbf{x}$ and $\mathbf{u}$ are the optimal solutions to P and D , respectively. The optimality of $\mathbf{x}$ and $\mathbf{u}$ can also be verified by the *complementary slackness* conditions.

Theorem 3 [Complementary Slackness]. *Let $\mathbf{x}$ and $\mathbf{u}$ be feasible solutions to P and D , respectively. Then, $\mathbf{x}$ and $\mathbf{u}$ are optimal solutions for each problem if and only if $\mathbf{u}^T(\mathbf{Ax} - \mathbf{b}) = 0$ and $(\mathbf{c} - \mathbf{A}^T \mathbf{u})^T \mathbf{x} = 0$.*

Theorem 3 implies three things. First, observe that in problem P , the first complementary slackness condition, that is $\mathbf{u}^T(\mathbf{Ax} - \mathbf{b}) = 0$, is automatically satisfied (since P is expressed in standard form). Second, the complementary slackness theorem states that in problem D , if any constraint is not tight, then the corresponding value of the primal variable corresponding to each such constraint is 0. If the constraint is tight, then the corresponding primal value is nonnegative. Finally, the theorem provides an easy way of translating primal solutions into corresponding dual solutions, and vice versa, as well as providing a certificate for whether a proposed solution for P is, in fact, optimal. The following example illustrates this last property.

Example [Complementary Slackness]. Consider the following LP expressed in standard form, together with its dual:

$$\begin{aligned}
 &\min && 3x_1 + 4x_2 - x_3 \\
 &\text{subject to:} && 3x_1 - 4x_2 + 2x_3 = 8 \\
 &&& 4x_1 - 2x_2 + 5x_3 = 22 \\
 &&& x_1, x_2, x_3 \geq 0. \\
 &\max && 8u_1 + 22u_2 \\
 &\text{subject to:} && 3u_1 + 4u_2 \leq 3 \\
 &&& -4u_1 - 2u_2 \leq 4 \\
 &&& 2u_1 + 5u_2 \leq -1.
 \end{aligned}$$

Suppose we are presented with the solution $\mathbf{x} = (0, 0.25, 4.5)$ and wish to establish optimality via complementary slackness. The first complementary slackness condition is automatically satisfied, because the problem is in standard form. Since $x_2 > 0$ and $x_3 > 0$, from the second complementary slackness condition we obtain $-4u_1 - 2u_2 = 4$ and $2u_1 + 5u_2 = -1$, which on solving, gives $\mathbf{u} = (-1.125, 0.25)$. This solution is dual feasible, as it also satisfies the first constraint of the dual problem, and the total cost is -3.5 . This is the same cost of the primal problem; therefore, by the strong duality theorem, $\mathbf{x}$ is an optimal solution to the original LP.

The following theorem establishes the Karush–Kuhn–Tucker (KKT) conditions for an LP problem. For more advanced discussions on the topic, consult the article titled **LP Duality and KKT Conditions for LP** in this encyclopedia.

Theorem 4 [KKT Optimality Conditions for LP]. *Given an LP problem P , the vector $\mathbf{x}$ is an optimal solution to P if and only if there exist vectors $\mathbf{u}$ and $\mathbf{r}$ such that*

1. $\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0$ (primal feasibility),
2. $\mathbf{A}^T \mathbf{u} + \mathbf{r} = \mathbf{c}, \mathbf{r} \geq 0$ (dual feasibility),
3. $\mathbf{r}^T \mathbf{x} = 0$ (complementary slackness).

In Theorem 4, the vector $\mathbf{r}$ is simply the vector of slack/surpluses $\mathbf{c} - \mathbf{A}^T \mathbf{u}$ from the dual problem D . In the primal problem P , the

vector $\mathbf{r}$ is referred to as the vector of reduced costs, and it is usually denoted by $\bar{\mathbf{c}}$.

From a computational perspective, the importance of the theorem is that a computer algorithm designed to solve problem P using an extreme vertex approach on the polyhedron described by the constraints of P , can work in one of two possible ways. The first approach is to make sure that both primal feasibility and complementary slackness are preserved at every iteration of the algorithm, while seeking dual feasibility only at the optimal solution. The second approach is to ensure dual feasibility and complementary slackness during every iteration, while seeking primal feasibility only at the optimal solution. The first approach yielded George Dantzig’s (primal) *simplex algorithm* around 1947 [3–5], which is, arguably, the cornerstone of the entire linear optimization literature. The second method led to the discovery of the *dual simplex algorithm*, in 1954 by Carlton Lemke [6].

While the simplex algorithm, in its both incarnations, has shown remarkably good performance in practice, in that it usually requires $O(m)$ iterations (pivots), there exist instances where an exponential number of iterations may be required until an optimal solution is found (see Ref. 7 for the first description of instances that lead to a performance on the order of $O(2^n)$). Hence, later theoretical efforts are concentrated on a different attack for solving the LP problem. These efforts were successful, first in 1979 when Leonid Khachian [8] discovered the ellipsoidal method for solving the LP problems in polynomial time, and later in 1984 with Narendra Karmarkar’s discovery of the first interior point method [9], which created a very fertile research area in subsequent years (see the section titled “Non-simplex Algorithms for LP” in this encyclopedia for more information). In the following section, we will discuss the theory and the implementation of the dual simplex algorithm.

DUAL SIMPLEX ALGORITHM

Throughout this section, we assume familiarity with basic linear algebra concepts, such

as bases, spaces, and polyhedra, as well as an understanding of the (revised) simplex algorithm, which is described in the article titled *The Simplex Method and Its Complexity* in this encyclopedia.

Assume that the coefficient matrix $\mathbf{A}$ is of full rank $m(m < n)$, and that there exist a basis $\mathbf{B}$ corresponding to the m linearly independent columns of $\mathbf{A}$, and a matrix $\mathbf{N}$ that contains the remaining $n - m$ nonbasic columns of $\mathbf{A}$. We partition the solution vector $\mathbf{x}$ into two components $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b} = (x_1, x_2, \dots, x_m)^T$ and $\mathbf{x}_N = (x_{m+1}, x_{m+2}, \dots, x_n)^T = \mathbf{0}$, called the *basic* and *nonbasic* solutions, respectively. In a similar fashion, we partition the cost vector $\mathbf{c}^T = [\mathbf{c}_B | \mathbf{c}_N]$ into a basic and a nonbasic component, that is, $\mathbf{c}_B$ and $\mathbf{c}_N$, which are the cost coefficients associated with $\mathbf{x}_B$ and $\mathbf{x}_N$, respectively. Suppose, moreover, that the basis $\mathbf{B}$ induces a dual feasible solution $\mathbf{u}$, that is, $\mathbf{A}^T \mathbf{u} \leq \mathbf{c}$ and $\mathbf{u}^T = \mathbf{c}_B^T \mathbf{B}^{-1}$. It is easy to see that the equality constraints of P are satisfied, since

$$\mathbf{A}\mathbf{x} = [\mathbf{B} | \mathbf{N}] \begin{bmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix} = \mathbf{b}$$

and that complementary slackness is also satisfied, since

$$\begin{aligned} (\mathbf{c} - \mathbf{A}^T \mathbf{u})^T \mathbf{x} &= \mathbf{c}^T \mathbf{x} - \mathbf{u}^T \mathbf{A}\mathbf{x} \\ &= \mathbf{c}^T \mathbf{B}^{-1}\mathbf{b} - \mathbf{c}_B^T \mathbf{B}^{-1}\mathbf{b} = 0 \end{aligned}$$

The only missing ingredient for an optimal solution to P is the requirement that $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. If this inequality happens to hold, then $\mathbf{x}$ is optimal for P , since all conditions of Theorem 4 are met. On the other hand, if there exists at least one negative entry in $\mathbf{x}_B$, then we perform a change of basis that eliminates this entry from $\mathbf{B}$, thus reducing the primal infeasibility. The procedure repeats until all entries in $\mathbf{x}_B$ are nonnegative. This is the essence of the dual simplex algorithm.

Naturally, the two questions that arise are about how this change of basis is actually performed, and whether or not the new solution obtained after the basis change is actually an improvement over the old one. The following subsections describe the actual

flow of the dual simplex algorithm, alongside with a discussion on several theoretical and practical implementation challenges.

Summary of the Dual Simplex Method: Description and Geometry

The dual simplex method entails the following steps.

Step 0 (Initialization). Find a dual feasible basis $\mathbf{B}$ such that the associated reduced cost vector $\bar{\mathbf{c}} = \mathbf{c} - \mathbf{u}^T \mathbf{A}$ is nonnegative (in the sense that each vector entry is nonnegative). (We will later provide a quick procedure for identifying such a basis.)

Step 1 (Pricing). If $\mathbf{x}_B = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$, stop; the current solution is optimal. Otherwise, select a pivot row i with the corresponding value of the basic variable $x_i < 0$. If there are multiple candidates, select the one with the minimum value. The variable x_i will leave the basis.

Step 2 (Ratio Test). If all constraint coefficients a_{ij} alongside the i th row are nonnegative ($a_{ij} \geq 0, \forall j$), stop. The dual problem is unbounded, and therefore the primal problem is infeasible. Otherwise, select the pivot column j , by performing the following minimum ratio test:

$$j = \arg \min_{k=1, \dots, m} \left\{ \frac{\bar{c}_k}{a_{ik}} \mid a_{ik} < 0 \right\} \quad (1)$$

The corresponding variable x_j will enter the basis.

Step 3 (Basis Change). Pivot on element a_{ij} and go to step 1.

We will illustrate the dual simplex method by contrasting it with the primal simplex, on a simple example, presented below. We

Table 1. The General Form of a Simplex Tableau

	z	$\mathbf{x}_B$	$\mathbf{x}_N$	RHS
z	1	$\mathbf{0}$	$\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N^T$	$\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$
$\mathbf{x}_B$	$\mathbf{0}$	$\mathbf{I}$	$\mathbf{B}^{-1} \mathbf{N}$	$\mathbf{B}^{-1} \mathbf{b}$

Table 2a. Primal Simplex on Primal Problem

	z	x_1	x_2	x_3	x_4	x_5	x_6	RHS
z	1	$2M - 2$	$M - 1$	$-M$	$-M$	0	0	$5M$
x_5	0	1	1	-1	0	1	0	3
x_6	0	1*	0	0	-1	0	1	2
z	1	0	$M - 1$	$-M$	$M - 2$	0	$-2M + 2$	$M + 4$
x_5	0	0	1*	-1	1	1	-1	1
x_1	0	1	0	0	-1	0	1	2
z	1	0	0	-1	-1	$-M + 1$	$-M + 1$	5
x_2	0	0	1	-1	1	1	-1	1
x_1	0	1	0	0	-1	0	1	2

*Pivoting element.

will use the traditional simplex tableau exposition (Table 1), presented in the following fashion (see Ref. 10 for one of the first descriptions of the simplex tableau, and Ref. 11 for a detailed description of this particular implementation form).

Note that in other treatments of LP [1], the reduced costs are defined the other way around, that is, $\mathbf{c}_B \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N$. Under this interpretation, the arguments presented below still hold, but the signs are reversed—for example, the algorithm would terminate when all reduced costs are nonnegative (in a minimization example), rather than nonpositive, as we assume.

Example [Primal Simplex]. Consider the following LP problem, together with its corresponding dual:

$$\begin{aligned}
 \min \quad & 2x_1 + x_2 \\
 \text{subject to: } & x_1 + x_2 \geq 3 \\
 & x_1 \geq 2 \\
 & x_1, x_2 \geq 0 \\
 \max \quad & 3u_1 + 2u_2 \\
 \text{subject to: } & u_1 + u_2 \leq 2 \\
 & u_1 \leq 1 \\
 & u_1, u_2 \geq 0.
 \end{aligned}$$

The following primal simplex tableaus (Tables 2a and 2b) solve the primal and the dual problems. For the primal problem, we

Table 2b. Primal Simplex on Dual Problem

	z	u_1	u_2	u_3	u_4	RHS
z	1	3	2	0	0	0
u_3	0	1	1	1	0	2
u_4	0	1*	0	0	1	1
z	1	0	2	0	-3	-3
u_3	0	0	1*	1	-1	1
u_1	0	1	0	0	1	1
z	1	0	0	-2	-1	-5
u_2	0	0	1	1	-1	1
u_1	0	1	0	0	1	1

*Pivoting element.

use the big- M method to form an initial basis, rather than the two-phase method, to keep everything in one tableau.

The optimal solution to the primal problem is $\mathbf{x} = (2, 1, 0, 0, 0, 0)^T$; the corresponding optimal dual vector is $\mathbf{u} = (1, 1, 0, 0)^T$. The optimal value of the objective function is $z = 5$. One can also quickly check the complementary slackness conditions and verify that they are met.

We will now examine the way in which the dual simplex algorithm operates. First, we look at the dual tableau (Table 3a), with the starting basis given by the surplus variables x_3 and x_4 (and thus, we multiply both rows by -1 to obtain the basis given by the identity matrix $\mathbf{I}_2$). Notice that this basis is primal infeasible, because of the negative terms in $\mathbf{B}^{-1} \mathbf{b}$.

Table 3a. Initial Basis for Dual Simplex

	z	x_1	x_2	x_3	x_4	RHS
z	1	-2	-1	0	0	0
x_3	0	-1	-1*	1	0	-3
x_4	0	-1	0	0	1	-2

*Pivoting element.

Table 3b. First Pivot Using Dual Simplex

	z	x_1	x_2	x_3	x_4	RHS
z	1	-1	0	-1	0	3
x_2	0	1	1	-1	0	3
x_4	0	-1*	0	0	1	-2

*Pivoting element.

Table 3c. Second Pivot Using Dual Simplex

	z	x_1	x_2	x_3	x_4	RHS
z	1	0	0	-1	-1	5
x_2	0	0	1	-1	1	1
x_1	0	1	0	0	-1	2

We are at step 1 of the dual simplex method. We select x_3 as the leaving variable ($-3 < -2$). According to step 2, the problem is not infeasible (two negative entries on the selected row). The entering variable is x_2 , by the minimum ratio test: $\min\{-2/-1, -1/-1\} = 1$. The new tableau is obtained by subtracting row 1 from row 0, and by dividing row 1 by -1 (Table 3b).

The solution is still primal infeasible. The variable x_4 leaves the basis, and x_1 enters, by the minimum ratio test. Row 2 gets subtracted from row 0, row 2 is added to row 1, and row 2 is divided by -1 to get the new tableau (Table 3c).

Since all terms in $\mathbf{B}^{-1}\mathbf{b}$ are positive and all entries in row 0 are negative, this is the optimal basis. The algorithm terminates, with the same solution as reported by the primal simplex above. Since the dual problem maximizes the objective function, the increase in z from one pivot to the next is natural.

Several observations are in order. First, notice the correspondence between the dual simplex, as exhibited in Tables 3a–3c, and the primal simplex method applied to the

dual problem, as shown in Table 2b. Through simple transposition manipulations, we can see that the two tableaus are equivalent. This is important: from a mathematical point of view, the dual simplex method used to solve a primal problem is equivalent to the primal simplex method applied to the dual problem, as long as the dual is not converted into standard form. (Converting the dual problem into standard form could lead to different tableaus.) Second, from a geometrical perspective, the dual simplex iterates around infeasible primal solutions (but dual feasible!), until a solution that is both primal and dual feasible is found. In our example, Table 3a is equivalent to the primal basic infeasible solution $\mathbf{x} = (0, 0, -3, -2)^T$. By complementary slackness, the associated dual solution is $\mathbf{u} = (0, 0)^T$, which is basic feasible in the dual problem. Figure 1, adapted from a similar example presented in Ref. 1, displays the progression of the dual simplex method, in terms of primal and dual basic feasible solutions.

In Fig. 1, the path A–B–C traces, in both spaces, the pivot sequence of the dual simplex applied to the example problem. In the primal space, bases A and B are infeasible, but at optimality, C is feasible. In the dual space, the algorithm maintains feasibility at each base A, B, and finally C. Naturally, the solution is improved during every iteration.

Implementation Issues

Cycling. While iterating, it is possible that the dual simplex can arrive at a situation where the reduced cost associated with the selected pivot column is zero, a situation first identified in Ref. 12. If this happens, then the reduced costs remain unchanged after pivoting, and therefore the value of the dual objective function ($\mathbf{c}_B^T \mathbf{B}^{-1} \mathbf{b}$) does not change as well. As a result, the dual simplex algorithm may cycle. The situation is avoided by choosing a lexicographic pivoting rule, which avoids cycling. Similar to anticycling rules for the primal simplex method, the rule states that one should choose the lexicographically smallest column, found by dividing all entries in the consideration set (columns with $a_{ik} < 0$) by a_{ik} . If a tie exists between several lexicographically smallest columns, one

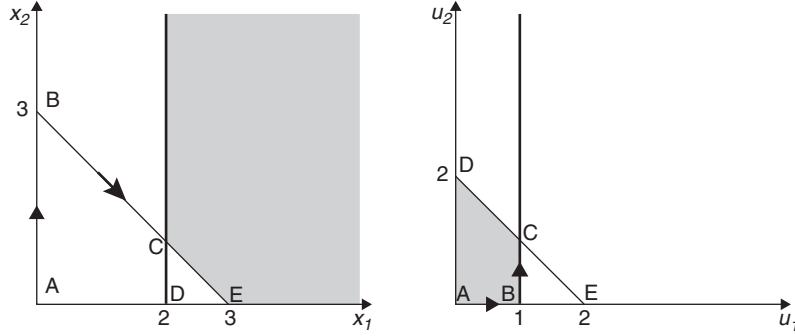


Figure 1. The primal and the dual feasible spaces.

should choose the column with the smallest index. The proof of the finiteness of the simplex algorithm (either in primal or dual form) when using the lexicographic rule can be found in a number of reference texts on LP such as Refs 1, 13, or 14. Besides the lexicographic rule, Bland's rule [15] (choose as entering variable the column with the smallest index that corresponds to a positive reduced cost and break ties among row candidates by favoring the row with the smallest index number) can also be easily adapted from the primal simplex algorithm and implemented in the dual simplex.

Degeneracy. Another issue that could appear while implementing the algorithm is dual degeneracy. Dual degeneracy appears whenever there are more than m reduced costs that are 0; that is, there exists at least a nonbasic variable with reduced cost equal to zero. Since cycling and degeneracy are somewhat related, practical implementations of the dual simplex algorithm try to resolve both issues simultaneously, if possible (see, for e.g., the COIN-OR (COmputational INfrastructure for Operations Research (www.coin-or.org)) project that provides, among other things, open source implementations of simplex algorithms. A good overview of the project can be found in Refs 16 and 17).

Finding an Initial Basic Dual Feasible Solution. As we have mentioned in the outline of the dual simplex algorithm, the algorithm must start with a basic dual

Table 4. Finding a BDFS - Initial Tableau

	z	x_1	x_2	x_3	x_4	x_5	RHS
z	1	0	1	5	-1	0	0
x_1	0	1	2	-1	1	0	4
x_5	0	0	3	4	-1	1	3

Table 5. Finding a BDFS - Adding an Artificial Constraint

	z	x_1	x_2	x_3	x_4	x_5	x_6	RHS
z	1	0	1	5	-1	0	0	0
x_6	0	0	1	1*	1	0	1	M
x_5	0	1	2	-1	1	0	0	4
x_1	0	0	3	4	-1	1	0	3

*Pivoting element.

feasible solution (BDFS), which sometimes is not readily available. Assuming that in the primal problem, the initial basis is given by the first m columns, a dual feasible basis can be constructed by adding the constraint $\sum_{j=m+1}^n x_j \leq M$ to the primal problem. Then, let the leaving variable be the slack measure associated with this constraint, and let the entering variable be the one corresponding to the maximum reduced cost. After the pivot, either the new basis is dual feasible and thus the dual simplex can commence, or the primal problem is optimal, or the primal is unbounded (see Ref. 11 for additional discussions as to why one of these outcomes must happen).

Table 6. Finding a BDFS – Basis Identification

	z	x_1	x_2	x_3	x_4	x_5	x_6	RHS
z	1	0	-4	0	-6	0	-5	$-5M$
x_3	0	0	1	1	1	0	1	M
x_1	0	1	3	0	2	0	1	$M + 4$
x_5	0	0	-1	0	-5	1	-4	$-4M + 3$

Table 7. LP Relaxation of Knapsack Example

	z	x_1	x_2	x_3	x_4	x_5	x_6	x_7	RHS
z	1	2	3	4	0	0	0	0	0
x_4	0	1	2	3	1	0	0	0	4
x_5	0	1	0	0	0	1	0	0	1
x_6	0	0	1	0	0	0	1	0	1
x_7	0	0	0	1	0	0	0	1	1

Example [Finding a BDFS [11], p. 277]. Consider the following tableau (Table 4), which we want to solve using dual simplex.

To get a dual basic feasible solution, we add the artificial constraint $x_2 + x_3 + x_4 \leq M$ to the problem. The slack of this constraint is x_6 , and the resulting tableau is as given in Table 5.

In the next step, x_6 leaves the basis, and x_3 enters (the maximum reduced cost is 5). The new tableau (Table 6) is dual feasible (but not primal feasible—notice the negative value of x_5).

While intuitively easy, this method is superseded in practical implementations by dual phase 1 algorithms. For an extensive review of dual phase 1 methods, see Ref. 18.

Selection Rules

From an implementation perspective, two important decisions that are quintessential to the practical performance of the dual simplex algorithm are the proper choices for the entering variable (also known as the *pricing* step) and the exiting variable (the *minimum ratio test* step). For a relatively long period of time, from the late 1950s until the early 1990s, the dual simplex was relatively ignored in practice, due to the

perception that it would not outperform the primal simplex algorithm. The last important theoretical contribution from the 1950s is Wagner’s adaptation of the dual simplex for bounded variables [19]. However, as the understanding of the structure of LP problems grew, it has become increasingly obvious that there exist classes of programs for which the primal simplex method is better, and other types of problems that tend to be better solved by the dual simplex [20]. Forrest and Goldfarth [21] and Fourer [22] helped in changing the perception about the dual simplex algorithm, by devising clever pricing techniques (steepest edge) and better ratio tests (bound flipping ratio test), that boost the dual simplex performance over its primal counterpart. Separately, Ref. 23 generalized the two-phase primal simplex algorithm to its dual counterpart. An excellent survey of the progress made with respect to the practical implementation of the dual simplex algorithm can be found in Ref. 24. Further breakthroughs emerged with the advance of parallelized implementations of the dual simplex [25].

APPLICATION OF DUAL SIMPLEX: INTEGER PROGRAMMING

One of the most important applications of the dual simplex is in large-scale optimization, particularly large-scale mixed integer programming. Traditionally, if the relaxation of a general mixed integer programming problem does not satisfy the original integrality requirements, then some mixture of variable branching and addition of valid inequalities occurs, until the branch-and-bound/cut/price implementation converges to the optimal solution (assuming it exists). Notice that branching on a fractional variable creates

Table 8. Optimal Tableau for LP Relaxation of Knapsack Example

	z	x_1	x_2	x_3	x_4	x_5	x_6	x_7	RHS
z	1	0	0	0	$-4/3$	$-1/3$	$-1/3$	0	$-19/3$
x_2	0	0	1	0	0	0	1	0	1
x_1	0	1	0	0	0	1	0	0	1
x_7	0	0	0	0	$-1/3$	$1/3$	$2/3$	1	$2/3$
x_3	0	0	0	1	$1/3$	$1/3$	$-2/3$	0	$1/3$

Table 9. Addition of Valid Inequality to LP Relaxation of Knapsack Example

	z	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	RHS
z	1	0	0	0	$-4/3$	$-1/3$	$-1/3$	0	0	$-19/3$
x_2	0	0	1	0	0	0	1	0	0	1
x_1	0	1	0	0	0	1	0	0	0	1
x_7	0	0	0	0	$-1/3$	$1/3$	$2/3$	1	0	$2/3$
x_3	0	0	0	1	$1/3$	$-1/3$	$-2/3$	0	0	$1/3$
x_8	0	0	1	1	0	0	0	0	1	1

two subproblems, each of which has an additional constraint—a bound on the branching variable, which destroys the primal feasibility of the original problem. Similarly, the addition of a valid inequality to the problem breaks primal feasibility. However, in the dual space, the problem remains feasible, since the addition of a row to the primal is equivalent to the addition of a column to the dual. If the added column is nonbasic, the dual basis can be used to continue the optimization via dual simplex, without the need to recompute and refactor a new basis (both of which are time intensive operations). This hot-starting feature makes the dual simplex very attractive for large-scale optimization projects. We illustrate this property with the following knapsack example.

Example [Valid Inequalities and Dual Simplex]. Consider the following knapsack problem:

$$\begin{aligned}
 \max \quad & 2x_1 + 3x_2 + 4x_3 \\
 \text{subject to:} \quad & x_1 + 2x_2 + 3x_3 \leq 4 \\
 & x_1, x_2, x_3 \text{ binary}
 \end{aligned}$$

The initial tableau for the relaxation is as given in Table 7.

After several iterations we obtain the final optimal tableau (Table 8) for the LP relaxation.

The solution is $x_1 = x_2 = 1, x_3 = 1/3$. This solution does not satisfy the binary requirements. On the other hand, notice that x_2 and x_3 cannot be simultaneously equal to 1 in the optimal solution, as that would violate the knapsack constraint. Therefore, the inequality $x_2 + x_3 \leq 1$ is valid for the original knapsack problem, and can be added to the previous tableau yielding Table 9.

Subtracting row 1 and row 4 from row 5, in order to get $\mathbf{I}_5$ as basis, yields a dual feasible solution (notice that row 5 has a negative RHS, rendering the solution primal infeasible). Dual simplex leads to optimality after one pivot (Table 10).

The solution is $x_1 = x_3 = 1, x_2 = 0$. Since this set also satisfies the binary requirements, it must be optimal to the original knapsack. The value of the objective function is $z = 6$.

CONCLUSIONS

Modern advances in LP theory establish the use of the dual simplex algorithm as a powerful optimization tool. While

Table 10. Application of Dual Simplex on LP Relaxation of Knapsack Example

	z	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	RHS
z	1	0	0	0	$-4/3$	$-1/3$	$-1/3$	0	0	$-19/3$
x_2	0	0	1	0	0	0	1	0	0	1
x_1	0	1	0	0	0	1	0	0	0	1
x_7	0	0	0	0	$-1/3$	$1/3$	$2/3$	1	0	$2/3$
x_3	0	0	0	1	$1/3$	$-1/3$	$-2/3$	0	0	$1/3$
x_8	0	0	0	0	$-1/3$	$1/3$	$-1/3^*$	0	1	$-1/3$
z	1	0	0	0	-1	$-2/3$	0	0	-1	-6
x_2	0	0	1	0	-1	1	0	0	3	0
x_1	0	1	0	0	0	1	0	0	0	1
x_7	0	0	0	0	-1	1	0	1	2	0
x_3	0	0	0	1	1	$-1/3$	0	0	-2	1
x_6	0	0	0	0	1	-1	1	0	-3	1

the performance of the dual simplex was originally considered to lag behind that of its most popular primal variant, the dual simplex is widely used today in practice. One of the more popular applications of the algorithm includes large-scale mixed integer programming, where row additions break primal feasibility, but typically, dual feasibility remains intact. The dual simplex is also used as a *de facto* LP solver, since research has identified [20] many classes of problems where the application of the dual algorithm outperforms the performance of its primal counterpart. All these interesting properties of the method, coupled with the fact that linear optimization is a fundamental topic in operations research, ensure that advances in this area will continue to be positioned at the forefront of the discipline.

REFERENCES

1. Bertsimas D, Tsitsiklis J. Introduction to linear optimization. 1st ed. Belmont, MA: Athena Scientific; 1997.
2. Murty KG. Linear programming. 1st ed. New York: Wiley; 1983.
3. Dantzig GB. Programming in a linear structure. Washington, DC: Comptroller, USAF; 1948.
4. Dantzig GB. Programming of interdependent activities II: mathematical model. *Econometrica* 1949;17(3/4):200–211.
5. Dantzig GB. Maximization of a linear function of variables subject to linear inequalities. In: Koopmans TC, editor. Activity analysis of production and allocation. New York: John Wiley and Sons, Inc.; 1951. pp 359–373.
6. Lemke CE. The dual method of solving the linear programming problem. *Nav Res Logist Q* 1954;1:36–47.
7. Klee V, Minty GJ. How good is the simplex algorithm? In: Shisha O, editor. Inequalities—III. New York: Academic Press; 1972. pp 159–175.
8. Khachian LG. A polynomial algorithm in linear programming. *Soviet Math Doklady* 1979;20:191–194.
9. Karmarkar N. A new polynomial-time algorithm for linear programming. *Combinatorica* 1984;4:373–395.
10. Dantzig GB. Linear programming and extensions. 1st ed. Princeton, NJ: Princeton University Press; 1963.
11. Bazaraa MS, Jarvis JJ, Sherali HD. Linear programming and network flows. 2nd ed. New York: Wiley; 1990.
12. Beale EML. Cycling in the dual simplex algorithm. *Nav Res Logist Q* 1955;2(4):269–275.
13. Chvátal V. Linear programming. 1st ed. New York: Freeman Publishers; 1983.
14. Vanderbei RJ. Linear programming: foundations and extensions. 3rd ed. New York: Springer US; 2008.
15. Bland RG. New finite pivoting rules for the simplex method. *Math Oper Res* 1977;2(2):103–107.

16. Lougee-Heimer R, *et al.* The COIN-OR initiative: open-source software accelerates operations research progress. *ORMS Today* 2001;28(5):20–22.
17. Lougee-Heimer R. The common optimization interface for operations research: promoting open-source software in the operations research community. *IBM J Res Dev* 2003;47(1):57–66.
18. Koberstein A, Suhl U. Progress in the dual simplex method for large scale LP problems: practical dual phase 1 algorithms. *Comput Optim Appl* 2007;37(1):49–65.
19. Wagner HM. The dual simplex algorithm for bounded variables. *Nav Res Logist Q* 1958;5:257–261.
20. Bixby RE. Solving real-world linear programs: a decade and more of progress. *Oper Res* 2002;50(2):3–15.
21. Forrest JJ, Goldfarb D. Steepest edge simplex algorithms for linear programming. *Math Program* 1992;57(3):341–374.
22. Fourer R. Notes on the dual simplex method. Draft report, Northwestern University, 1994.
23. Maros I. A generalized dual phase-2 simplex algorithm. *Eur J Oper Res* 2003;149(1):1–16.
24. Koberstein A. Progress in the dual simplex algorithm for solving large scale LP problems: techniques for a fast and stable implementation. *Comput Optim Appl* 2008;41(2):185–204.
25. Bixby RE, Martin A. Parallelizing the dual simplex method. *INFORMS J Comput* 2000;12(1):45–56.

DYNAMIC AUCTIONS

DIRK BERGEMANN
Department of Economics, Yale
University, New Haven,
Connecticut

MAHER SAID
Microsoft Research New
England, Cambridge,
Massachusetts

Olin Business School,
Washington University in
St. Louis, St. Louis, Missouri

Many important economic problems have been studied with the tools of auction theory and mechanism design more generally. Much of the literature, however, studies a static, one-time decision. In many problems of interest, more than a single decision needs to be made; rather, a sequence of decisions needs to be made. These decisions often depend crucially on the dynamic aspects of the environment. For example, in the classical airline revenue management problem, an airline must decide how to price seats on a flight in response to its changing inventory as well as to the evolution of the customer base. A search engine must choose how to allocate its advertising inventory in response to changing search queries and advertiser budgets. Network capacity (either bandwidth or computational resources) needs to be dynamically reallocated in response to the arrival of new computational tasks of varying priority. Private and public procurement agencies must decide how to assign new contracts in response to the changing experiences, competencies, and availabilities of contractors.

The theory of auctions and mechanism design has been extremely successful in tackling a wide variety of problems. Unfortunately, solutions to static problems often do not translate directly to dynamic settings. Consider, for instance, the standard second-price auction for a single object with private values. Buyers are asked to submit bids for

the good, which is then allocated to the highest bidder. This winning bidder then pays the second-highest of the submitted bids. As first demonstrated by Vickrey [1], it is a (weakly) dominant strategy for buyers to bid their true values. Thus, the object is allocated efficiently: the buyer who values the object the most receives it.

In a dynamic setting, however, the second-price auction need not retain its desirable properties. The following example is illustrative. Suppose that there are two objects to be allocated: use of a production facility today and use of that same facility tomorrow. Today, two perfectly patient buyers wish to make use of the facility, the first valuing that use at \$100 and the second valuing it at \$75. A third buyer who will arrive tomorrow values use of the facility at \$50. If a second-price auction is used in each period and the buyers bid truthfully, then the first buyer will win in the first period at a price of \$75, and the second buyer will win in the second period at a price of \$50. On the other hand, suppose that the first buyer instead bids \$60 in the first period and truthfully in the second. She will then lose in the first period, but win in the second and pay a price of \$50, below what she otherwise would have paid. Thus, bidding one's value need not be an optimal strategy in a sequence of second-price auctions.

Notice that in the above example, despite the lack of truthful behavior by the bidders, the outcome is still efficient. As we will discuss later, this need not always be the case. In a more general setting when buyers are uncertain about the arrival time and willingness-to-pay of each of their competitors, Said [2] demonstrates that the second-price sealed-bid auction is often incapable of yielding an efficient outcome, even when bidders are forward-looking and fully rational. Thus, in order to achieve desirable outcomes in dynamic settings, we must move beyond tools that are best suited to static environments.

This is precisely the aim of the recent (and growing) literature on dynamic mechanism

design. We survey this literature, categorizing it by the nature of the dynamics involved. In particular, we examine two strands of literature: one in which the population of agents changes over time, but their private information is fixed; and the other in which the population of agents is fixed, but their private information changes over time. We will further divide these two categories into those papers whose aim is achieving efficient outcomes and those whose aim is achieving revenue-maximizing outcomes.

A DYNAMIC POPULATION WITH FIXED INFORMATION

We begin by exploring the possibility of implementing desirable outcomes when faced with a dynamic population whose private information is fixed. In the course of our discussion, we will examine some of the problems that arise in such settings and how they differ from those found in static, one-shot models. Note that all the models and papers we discuss assume private values.

Maximizing Social Welfare

Efficiency in the Face of Arrival and Departure Dynamics. Parkes and Singh [3] develop a framework for studying sequential allocation problems in a dynamic population setting. In their world, various subsets of agents are present at various times. In particular, each agent has an arrival time and a departure time, and has a utility function for decisions made while she is present. What makes the problem especially difficult is that the decision maker (and mechanism designer) does not know any of this information—the designer must incentivize the agents to reveal that information.

Despite this difficulty, Parkes and Singh are able to construct a mechanism (the “online VCG” mechanism) which generalizes the Vickrey-Clarke-Groves mechanism from static environments and implements the efficient outcome. Their mechanism is a direct revelation mechanism in which buyers report their private information upon arrival to the mechanism. Given these reports, the mechanism carries out an efficient policy.

In order to enable the truthful revelation of private information, the mechanism chooses transfers that align agents’ incentives with those of a benevolent social planner. Similar to how the static VCG mechanism sets transfers equal to the externality that an agent imposes on other agents, the online VCG mechanism sets transfers equal to the *expected* externality that an agent imposes. This expected externality takes into account each agent’s impact on the other agents currently present *as well as* on agents that may arrive in the future.

We should note, however, that *unlike* the static VCG mechanism, this truthful reporting is *not* a dominant strategy, as the calculation of expected externalities depends upon the truthful reporting of all other agents. If, instead, an agent believes that agents in the future are not reporting truthfully, it may be in her best interest to misreport her true type.

Efficient Sequential Assignment with Impatient Buyers. In related work, Gershkov and Moldovanu [4] examine the allocation of a finite set of heterogeneous durable goods to a dynamic population of randomly arriving buyers. Crucially, the buyers in this setting are impatient—they wish to purchase an object immediately upon their arrival on the market. Moreover, these buyers share a common ranking of the various objects. In this setting, objects are durable while buyers are impatient. This implies that the relevant trade-off is between allocating an object to a buyer at the current time and the option of assigning it to a buyer in the future who might value it more. In a complete information setting, the efficient dynamic allocation policy was first characterized by Albright [5]. This policy prescribes an ordered partition of the set of possible valuations at each possible time. Agents with values in the highest element of the partition receive the best available object; agents in the second-highest element receive the second-best remaining object, and so on. These intervals may depend upon the time of an agent’s arrival or on the set of remaining objects.

Gershkov and Moldovanu [4] show that this efficient policy is, in fact, implementable in the presence of incomplete information

about the valuations of arriving buyers. Since buyers are impatient, there is no loss of generality in considering direct revelation mechanisms where each buyer reports her type to the mechanism designer upon her arrival to the market. Each buyer is charged the *expected* externality that he/she imposes on future agents, where expectations are taken with respect to the arrival process and the valuations of future agents. This implies that each agent's net utility is exactly equal to his/her expected marginal contribution to the social welfare, conditional on the information available at the time of her arrival—essentially, the efficient mechanism is a dynamic Vickrey-Clarke-Groves mechanism like the Parkes and Singh [3] online VCG mechanism discussed above.

Efficient Auctions with Dynamic Populations. In a closely related recent work, Said [2,6] examines the allocation of a sequence of indivisible goods to a dynamic population of buyers. In this setting, patient buyers demand a single unit of perishable homogeneous goods (such as computational time on a central computing facility, or capacity in a production line). These buyers arrive at random times to the market, and have private valuations, which are drawn independently and identically from a given distribution.

While a variant of the online VCG mechanism of Parkes and Singh [3] may be used in order to achieve an efficient allocation of objects to buyers, these mechanisms are direct revelation mechanisms, requiring buyers to report their values to the mechanism upon their arrival to the market. In practice, however, direct revelation mechanisms may be difficult to implement or undesirable due to privacy or complexity concerns.

Therefore, *indirect* auction mechanisms are often useful. In standard static settings, the canonical auction for efficiently allocating goods to buyers is the sealed-bid second-price auction. This mechanism is the auction analogue of the Vickrey-Clarke-Groves mechanism, allocating the object to the buyer who submits the highest bid and charging her a price equal to the second-highest bid. Said [2] shows that, when selling a sequence of objects, one in each period, to a stream

of buyers who arrive at random times, the sealed-bid second-price auction is no longer efficient.

In a sequential auction, buyers have an “option value” associated with losing in a particular auction, as losing bidders have the possibility of winning in a future auction. The value of this option to a particular buyer depends on his/her expectations of the prices he/she may have to pay in the future. This is, of course, dependent upon the private information and values of other competitors. Thus, despite the fact that buyers' values for an object do not depend upon the information of their competitors, the dynamics of a sequential auction market *generate interdependence* in (option) values.

This interdependence implies that a standard second-price sealed-bid auction does not reveal sufficient information for the determination of buyers' option values. In contrast, the ascending (or English) auction is a simple *open* auction format that does allow for the gradual revelation of buyers' private information. As buyers drop out of each auction, they indirectly reveal their private information to their competitors, who are then able to condition their current-period bids on this information. Moreover, this process is repeated in every auction in every period. This allows newly arrived buyers to learn about their competitors without being privy to the detailed events of the previous periods, thereby allowing for an efficient outcome.

Maximizing Revenue

Before discussing revenue-maximizing dynamic mechanisms and auctions, it will be helpful to briefly review the (standard) results for static environments. In a static setting, in which all buyers are present simultaneously, Myerson [7] solves the optimal mechanism design problem in the following manner:

1. Incentive compatibility requires that the allocation rule be monotone; that is, buyers who report a higher type must have a greater probability of receiving an object.
2. Given a monotone allocation rule, the corresponding payment rule is pinned

down (up to a constant) via the revenue-equivalence principle. When the goal is maximizing revenues, the constant can be determined via individual rationality constraints.

3. Combining these two observations, the revenue-maximization problem can be transformed to the problem of maximizing “virtual” surplus—the surplus from allocating objects less the information rents that must be paid in order to induce truthful revelation of types.

Thus, instead of maximizing the sum of agents’ values as in the case of efficient mechanism design, Myerson’s insight is that revenue-maximization is equivalent to maximizing the sum of agents’ *virtual* values, where the virtual value of a buyer is given by the difference between his/her true value and the information rents necessary for inducing truth-telling. An optimal allocation rule, therefore, allocates objects to the buyers with the highest nonnegative virtual values—precisely those with the most favorable trade-off between surplus and information rents.

When buyers are *ex ante* symmetric and the distribution of values is such that virtual values are nondecreasing then this optimal allocation rule is monotone in values. Thus, incentive compatibility is achieved “for free” and the optimal auction corresponds to a second-price auction with an appropriately chosen reserve price. If, on the other hand, virtual values are not increasing, then the monotonicity constraint may be satisfied via a procedure termed “ironing” which pools some types of agents and randomizes among them. Thus, in the static setting, increasing virtual values is sufficient for revenue-maximization to correspond to maximizing virtual surplus.

Revenue-Maximization in the Face of Arrival and Departure Dynamics. Returning to the setting of arrival and departure dynamics, Pai and Vohra [8] present what may be thought of as the revenue-maximizing counterpart to the efficient online VCG mechanism of Parkes and Singh [3]. Here, a seller with a fixed, finite supply of a

homogenous good faces a population of potential buyers with unit demand who arrive and depart over the course of a finite time horizon. As in Parkes and Singh, the times at which each agent arrives and departs from the market are her private information, as is her valuation for an object. Agents’ types are assumed to be independent draws from a common distribution. The seller wishes to design an incentive-compatible direct mechanism in order to maximize his revenues.

The challenge then is to extend Myerson’s notion of a virtual value in order to account for the additional privately known arrival and departure times. Moreover, even if such a virtual value is defined, the multidimensional nature of the private information implies that we must also find an appropriate notion of monotonicity in order to guarantee incentive compatibility.

Some care must be used when considering direct revelation mechanisms in this setting with random arrival and departure times. In particular, note that buyers cannot make reports before their arrival or claim to depart after their true departure time (see Green and Laffont [9] for more on restricted mechanisms and the revelation principle). Taking this into account, Pai and Vohra use an appropriate formulation of the revelation principle of Myerson [10] to restrict attention to mechanisms that allocate to buyers, if at all, only in the period of their departure. They use the incentive compatibility constraints to show that the optimal allocation rule must be monotone in valuations, holding entry and exit times fixed. Moreover, incentive compatibility also requires that the allocation rule be monotone in entry and exit times. In particular, a buyer who arrives earlier or departs later should have a greater probability of receiving an object. (Thus, buyers are incentivized to report the greatest possible window of opportunity.)

Making use of the conditions imposed by incentive compatibility, Pai and Vohra [8] are able to arrive at a notion of virtual valuation similar to that of Myerson [7], except that information rents vary across reported arrival and departure times. Thus, in each period, the seller compares the

virtual surplus from allocating an object to a departing agent to the expected virtual surplus of waiting an additional period. A departing agent will receive an object only if the former is greater than the latter. However, unlike the setting of Myerson, increasing virtual values are no longer sufficient for such an allocation rule to be optimal, but instead additional “ironing” may be necessary: the optimal mechanism may need to occasionally withhold an object despite a lower future expected virtual surplus.

Optimal Sequential Assignment with Impatient Buyers. Recall the relationship between Parkes and Singh [3] and Gershkov and Moldovanu [4] discussed earlier. An analogous relationship exists between the work of Pai and Vohra [8] and Gershkov and Moldovanu [11]. This latter work considers the problem of revenue-maximization in a setting with a dynamic population of impatient buyers. As in the paper discussed earlier, a seller has a finite set of heterogeneous durable goods that he wishes to sell to randomly arriving buyers who wish to purchase an object immediately upon their arrival to the market. (In a setting with homogeneous durable goods and patient buyers arriving stochastically, Gallien [12] shows that when the distribution of buyer inter-arrival times has an increasing failure rate, the optimal mechanism allocates goods to buyers only upon their arrival. There, as Board [13] also shows in a related discrete-time setting, the seller may behave as though buyers are impatient.)

As in Pai and Vohra [8], there is a dynamic trade-off that is introduced by the arrival process of buyers. In particular, when considering allocating an object to a particular agent, the seller must consider the opportunity cost of allocating that object to future agents. However, the problem here is (in some ways) simpler as buyers no longer have multidimensional types. Since buyers are impatient, there is no need to consider incentives for revealing arrival or departure times—the only dimension of private information is the private value that buyers have for each object.

Thus, the authors are able to prove that (as in the case of the efficient policy) any incentive compatible policy must take the form of an ordered partition of possible valuations at each point in time. This follows directly from the fact that short-lived agents’ decision problems are static, and so static insights about incentives apply directly. Thus, agents with values in the highest element of the partition receive the best available object; agents in the second-highest element receive the second-best remaining object, and so on.

The next step in the characterization of the revenue-maximizing dynamic mechanism is to determine the revenue generated by any particular incentive compatible policy. The problem faced by the seller upon the arrival of any particular agent can then be translated into a static problem. More specifically, the seller must decide between allocating an object to the newly arrived agent and obtaining the revenue generated by ignoring that agent’s arrival and proceeding with the (given) policy. This is the same as the problem faced by a standard (static) revenue-maximizing monopolist selling a single object, when the cost of that object is given by the salvage value of the inventory (which corresponds to the expected future revenues from the dynamic policy). It is possible, by working backwards from the case of a single-object inventory, to determine the optimal partition of the valuation space for each possible inventory and point in time. The revenue-maximizing prices that correspond to these cutoffs are then derived via revenue-equivalence.

Optimal Auctions with Dynamic Populations. In a complementary setting, Said [2] explores the properties of revenue-maximizing dynamic auctions. Instead of durable goods and impatient buyers, he examines the assignment of a sequence of perishable goods to a population of patient buyers. The buyers arrive to the market at random times and remain until they receive an object. Since their values for the objects are private information, incentives must be provided for information revelation in order to achieve an optimal allocation.

In this setting, the seller and buyers discount the future with a common discount

factor $\delta \in (0, 1)$. This implies that even though objects are assumed to be individual units of a homogeneous good, from the perspective of any individual buyer, they are differentiated products. To make this clear, consider a buyer i with value v_i who is present on the market at some time t . If this buyer receives an object in period t , she receives a payoff of v_i . However, if she anticipates receiving an object in period $t + 1$, then her expected payoff is δv_i . Thus, she does not value the two objects identically.

This alters the appropriate condition for incentive compatibility. Recall that in static settings with single-unit demand, incentive compatibility requires that increasing a buyer's type should increase her probability of receiving an object. In this dynamic setting with single-unit demand, incentive compatibility instead requires that increasing a buyer's type should increase her probability of receiving an object sooner. This characterization of incentive compatibility allows for the derivation of a revenue-equivalence theorem for this setting, which implies that revenue-maximization can be achieved (in the Myersonian tradition) by maximizing the virtual social surplus.

Said [2] shows that applying a variant of the online VCG mechanism discussed earlier to virtual values maximizes revenues. Instead of providing each buyer with a net utility equal to her expected marginal contribution to the social surplus, Said's mechanism leaves each buyer with a net utility equal to her expected marginal contribution to the *virtual* surplus. Moreover, if an indirect mechanism is desired, the earlier discussion on efficient but indirect mechanisms remains relevant—instead of a sequence of second-price sealed-bid auctions with an optimally chosen reserve price, a seller should use a sequence of ascending auctions with that same reserve price.

If objects are durable instead of perishable, the recent work of Board and Skrzypacz [14] provides a characterization of the optimal dynamic auction. As before, the revenue-maximizing direct mechanism applies an efficient mechanism to virtual values. This implies that there is a constant cutoff value below which objects are not

allocated, and the object is allocated to the buyer with the highest valuation exceeding that cutoff. This cutoff is determined by a simple one-period-look-ahead policy which requires the seller to be indifferent between selling to the cutoff type in the current period and waiting an additional period for new buyers to arrive.

With a single durable object to be allocated, these cutoffs are constant in all but the final period, and then drop in that period to the static monopoly price. This implies that buyers either buy immediately upon their arrival or wait for the “fire sale” in the final period. If the seller wishes to use an indirect mechanism in order to implement this optimal policy, Board and Skrzypacz show that a sequence of second-price auctions with deterministically declining reserve prices is optimal. These reserve prices are chosen such that the cutoff type is exactly indifferent between purchasing in the current period and waiting an additional period. The reserve price must be declining so as to compensate this cutoff-type buyer for the losses due to discounting *as well as* for the possibility of new competition arriving in future periods. It follows that the reserve price must fall at a rate faster than the discount rate.

A FIXED POPULATION WITH DYNAMIC INFORMATION

The preceding section examined dynamic auctions and mechanisms for settings in which there is a dynamic population with static, fixed types. A large and important class of problems, however, does not fit within this framework. Consider, for instance, the case of sponsored search, where search engines sell advertising slots alongside “organic” search results. In this setting, advertisers who are interested in a particular keyword may revise their estimates of the value of an advertisement based on their experiences over time. The natural question that arises in such a setting is how to design mechanisms and auctions in order to properly account for the changing private information.

We first consider the case of implementing socially efficient outcomes in environments

with dynamic information. The mechanisms we examine generalize insights from static settings to enable the truthful revelation of changing private information in a dynamic environment. We then turn to the objective of maximizing revenues and discuss the recent work in that area. As in the previous section, we focus on private value environments.

Maximizing Social Welfare

Dynamic Pivot. Bergemann and Välimäki [15] consider the problem of providing incentives for truthtelling and satisfying participation constraints in an efficient mechanism. They consider a general infinite-horizon dynamic model in which participants observe a sequence of private signals over time, and a sequence of decisions must be taken in response to these signals. The distribution of signals may depend on previously observed signals or decisions; however, these signals are independent across agents (conditional on observables).

Bergemann and Välimäki propose the Dynamic Pivot Mechanism. This mechanism is the natural generalization of the static Vickrey-Clarke-Groves mechanism to a dynamic environment. Recall that VCG-like mechanisms incentivize truthtelling by making efficient decisions and choosing transfers such that each agent's payoff is equal to the total social surplus. By varying the transfers to each agent in a way that does not depend on her own report, we obtain an entire set of efficient mechanisms. The "Pivot" mechanism (which is often considered the canonical VCG mechanism) is the member of this class that subtracts from each agent's transfer the surplus that the other agents could have achieved in her absence; that is, it chooses transfers equal to the externalities imposed on the rest of the system so that each agent's net utility is equal to her marginal contribution to the social welfare.

The key insight of Bergemann and Välimäki [15] is that in a setting with dynamic private information, incentives for truthtelling and efficiency may be guaranteed by choosing payments in each period to be equal to the *flow* externalities. More precisely, the externality that an

agent's report in any given period imposes on all other participating agents may be decomposed into a current period effect and a future (expected) effect. By committing to a sequence of transfers equal to the sequence of current-period externalities, the mechanism designer or social planner is able to give each buyer a flow payoff equal to her flow marginal contribution to the social welfare. Therefore, each agent's expected discounted payoff, looking forward from any period, is equal to her *total* marginal contribution to the social welfare (again, starting at that point). This property (which figured prominently in Bergemann and Välimäki [16,17] in the construction of dynamic first-price auctions in a complete information setting) implies that, in each period, all participating agents internalize the impact of their current reports on others, thereby aligning their incentives with those of the efficiency-minded planner.

Cavallo, Parkes, and Singh [18] develop a mechanism similar to the Dynamic Pivot Mechanism in a setting with agents, whose type evolution follows a Markov process. In Cavallo, Parkes, and Singh [19], they extend dynamic VCG to settings in which buyers are "periodically inaccessible" and are unable to make reports (as in the case, for instance, of a network setting where connectivity is sometimes lost). Moreover, Cavallo [20] has shown that the dynamic pivot mechanism is also similar to its static VCG counterpart, in that it achieves greater revenue than any other possible efficient (dynamic) mechanism. Therefore, in settings where the mechanism designer is concerned with both revenue and efficiency, the use of the dynamic pivot mechanism may provide a partial resolution to the conflict among the two.

Efficiency and Budget-Balance. Athey and Segal [21] consider a similar setting to that of Bergemann and Välimäki [15]; however, they are also interested in finding an efficient mechanism that is budget-balanced—that is, a mechanism that requires no external subsidies. In order to do so, Athey and Segal propose the efficient "Team Mechanism." This mechanism induces truthfulness by providing a transfer to each agent, in each

period that is equal to the sum of utilities of all the other agents in that period. Thus, in the same manner as the Dynamic Pivot Mechanism (as well as the static VCG mechanism), such a transfer scheme ensures that each agent is a residual claimant for the total social surplus.

As the Team Mechanism is not budget-balanced, Athey and Segal modify it in a generalization of the classic AGV mechanism of d'Aspremont and Gerard-Varet [22]. The AGV mechanism, in a static environment, permits truthtelling by charging agents transfers equal to the “expected externality,” their reports impose on others. However, the value of these transfers are dependent upon agents’ beliefs, and in a dynamic setting, these beliefs may be evolving over time and may, in fact, be manipulated by other agents. In order to avoid such manipulations, Athey and Segal construct the “Balanced Mechanism,” which takes advantage of the dynamic setting using the *changes* in the expected present value of other agents’ utilities. By doing so, agents continue to internalize the expected externality that they impose upon others. Moreover, these new transfers are not manipulable by other agents, thereby allowing for budget-balance while also implementing efficient outcomes.

One should note that, as with the static AGV mechanism, it may be impossible to both balance the budget and also satisfy participation constraints. Although the Balanced Mechanism satisfies *ex ante* participation constraints (so that all agents wish to participate in the mechanism at the time of contracting), it differs from the Dynamic Pivot Mechanism of Bergemann and Välimäki [15], in that some agents may wish to exit the mechanism after some histories. Athey and Segal [21] show, however, that if the time horizon is infinite and agents are sufficiently patient, then it is possible to satisfy these “periodic” participation constraints.

Maximizing Revenue

Contracting with Dynamic Information. In many environments, buyers learn about their demand over time. In an important contribution, Courty and Li [23] consider optimal

dynamic contracts for such a setting. In their model, buyers have private information about the future distribution of their valuations for the seller’s object. Subsequently, these buyers then privately learn the realization of their value. The monopolist could choose to wait until buyers learn their realized values and then charge the standard monopoly price. Surprisingly, by requiring the buyers to reveal their private information sequentially, the seller is able to extract additional surplus. Instead of “screening” buyers once after all private information is acquired, revenue-maximization requires screening them twice with a set of contingent contracts.

As in a standard static mechanism design problem, revenue-maximization corresponds to allocating to the buyers with the highest virtual values, thereby maximizing virtual surplus. Since buyers’ initial types influence their ultimate value for the good, the concept of a virtual value in this context must take into account this influence. Specifically, the information rents “paid” to buyers in the first period (at the time of contracting) must account for the informativeness of initial types about future types: the more informative the initial type about the future realization, the greater the information rents that must be paid. Since incentive compatibility requires monotonicity in the allocation rule, this implies that distortions away from efficiency are correspondingly increased.

Eső and Szentes [24] consider a closely related setting in which buyers have an initial estimate of their valuation, and the seller is able to control buyers’ acquisition of additional refined information. Specifically, the seller is able to choose whether to allow buyers to receive additional signals that are correlated with their private information. They show that in the optimal revenue-maximizing mechanism, the seller releases all the additional private information that he is able to. Moreover, his revenue in this case is as large as the revenue he would be able to achieve if he could observe this information—quite surprisingly, buyers gain no advantage from this additional information and receive no additional information rents when the seller is able to control

the flow of information. Roughly speaking, by using a more sophisticated dynamic mechanism, the seller pays information rents to buyers in the initial period, in order to incentivize the revelation of their initial private information, but subsequently extracts all remaining potential surplus.

In a particular application, Eső and Szentes [24] show how such a mechanism would operate. Suppose buyers' initial private information is a noisy signal of their value, and that they can then receive additional information that informs them of their actual value. The optimal mechanism in this case is a "handicap auction" in which buyers can purchase an advantage in a second-stage efficient auction. More concretely, each buyer purchases a price premium from the seller (where a smaller premium costs more), and then in the second period (after the revelation of additional information) the buyers engage in a modified second-price sealed-bid auction in which the winner is required to pay her purchased premium in addition to the second-highest bid. The seller is able to extract additional revenue by inducing buyers with higher expected values to purchase smaller handicaps for the second-stage auction.

Dynamic Incentives and Revenue Equivalence. The recent contribution of Pavan, Segal, and Toikka [25] generalizes the results of Myerson [7] and Baron and Besanko [26] to a general dynamic setting. In particular, they characterize incentive compatibility and revenue in multiperiod settings with dynamic private information. The primary challenge in such settings is an element of multidimensionality. Instead of the single dimension of potential misreports possible in a traditional static mechanism design problem, agents may be able to misrepresent multiple "pieces" of information, possibly choosing misreports contingent on her previous private information.

The key step in their analysis is the development of a "dynamic payoff formula" which expresses the derivative of an agent's expected payoff (and hence the seller's revenue) in an incentive compatible mechanism with respect to her private information.

This formula summarizes the local incentive compatibility constraints that a mechanism designer must respect. Pavan, Segal, and Toikka demonstrate the manner in which this dynamic payoff formula relies on incentive compatibility not only in a given period, but also in all future periods; therefore, they are able to capture not just the impact of a change in an agent's private information on her current-period payoffs, but also the influence on future private information.

This machinery yields a dynamic revenue-equivalence theorem: the expected revenue in any dynamic mechanism is determined entirely (up to a constant) by the allocation rule. Moreover, the expected revenue in an incentive compatible mechanism is given exactly by the expected virtual surplus generated by that mechanism, where agents' virtual values (as in Baron and Besanko [26]; Country and Li [23]; and Eső and Szentes [24]) depend upon the informativeness of current-period private information on future private information. Thus, Pavan, Segal, and Toikka [25] provide necessary conditions (and some sufficient conditions) for incentive compatibility and a general methodology for revenue-maximization in dynamic settings.

One limitation of the generality of that work, however, is the difficulty in establishing general sufficient conditions for the incentive compatibility of a mechanism that are easily verified. In a complementary paper, Pavan, Segal, and Toikka [27] use an approach pioneered by Eső and Szentes [24] to represent an agent's future-period types as a function of their first-period type and a random "shock" that is independent of the first-period type. This approach allows the identification (under slightly different assumptions) of conditions for incentive compatibility in infinite-horizon settings.

Despite this, however, sufficient conditions for incentive compatibility often need to be determined on a case-by-case basis for various models. One such determination may be found in Kakade, Lobel, and Nazerzadeh [28], which sets out a model of sponsored search auctions in which advertisers learn about the efficacy of their advertisements over time. Their model combines elements of both private information and statistical

learning. In particular, private information evolves according to a multiarmed bandit process (Deb [29] also considers a similar setting). The authors also impose “separability” conditions on the stochastic process governing the evolution of private information (specifically, that it is a Markov process) along with additive or multiplicative separability of buyers’ utility functions in their initial private information and future private information. These conditions are, in fact, sufficient conditions under which the maximization of virtual surplus by using an efficient algorithm (the Gittins index policy) applied to virtual values maximizes revenue.

In a related setting, Battaglini [30] examines an infinite-horizon model and characterizes the revenue-maximizing long-term contract of a monopolist facing a buyer whose value follows a two-type Markov process. This optimal contract is nonstationary; however, this contract (eventually) converges over time to the efficient supply. A crucial insight (that follows from the work of Pavan, Segal, and Toikka) is that, in an irreducible two-type Markov process, the impact of changing the current-period type vanishes in the long run—the current-period type is almost completely uninformative about types in the distant future. Therefore, distortions away from the efficient contract diminish in the long run, regardless of the degree to which types are persistent.

FURTHER READING

A large body of work has developed in this area. For further reading and examples of dynamic auctions and mechanisms for settings with changing populations and private information, the curious reader may wish to consult, among others, Akan, Ata, and Dana [31]; Aviv and Pazgal [32]; Cavallo [33]; Doepke and Townsend [34]; Gershkov and Moldovanu [35]; Kittsteiner and Moldovanu [36]; Parkes [37]; Satterthwaite and Shneyerov [38]; Shen and Su [39]; Talluri and van Ryzin [40] [Chapter 6]; and Vulcano, van Ryzin, and Maglaras [41].

REFERENCES

1. Vickrey W. Counterspeculation, auctions, and competitive sealed tenders. *J Finance* 1961;16(1):8–37.
2. Said M. Auctions with dynamic populations: efficiency and revenue maximization. *Proceedings of the 1st Conference on Auctions, Market Mechanisms and Their Applications (AMMA 2009)*; 2009; Boston (MA).
3. Parkes DC, Singh S. An MDP-based approach to online mechanism design. *Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS’03)*. Vancouver, Canada: 2003.
4. Gershkov A, Moldovanu B. Efficient sequential assignment with incomplete information. *Games Econ Behav* 2010;68(1):144–154.
5. Albright SC. Optimal sequential assignments with random arrival times. *Manage Sci* 1974;21(1):60–67.
6. Said M. Information revelation and random entry in sequential ascending auctions. *Proceedings of the 9th ACM Conference on Electronic Commerce (EC’08)*; 2008; Chicago (IL).
7. Myerson RB. Optimal auction design. *Math Oper Res* 1981;6(1):58–73.
8. Pai M, Vohra R. Optimal Dynamic Auctions. Unpublished manuscript, Northwestern University; 2009.
9. Green JR, Laffont J-J. Partially verifiable information and mechanism design. *Rev Econ Stud* 1986;53(3):447–456.
10. Myerson RB. Multistage games with communication. *Econometrica* 1986;54(2):323–358.
11. Gershkov A, Moldovanu B. Dynamic revenue maximization with heterogeneous objects: a mechanism design approach. *Am Econ J Microecon* 2009;1(2):168–198.
12. Gallien J. Dynamic mechanism design for online commerce. *Oper Res* 2006;54(2):291–310.
13. Board S. Durable-goods monopoly with varying demand. *Rev Econ Stud* 2008;75(2):391–413.
14. Board S, Skrzypacz A. Optimal dynamic auctions for durable goods: posted prices and fire-sales. Unpublished manuscript, Stanford University; 2010.
15. Bergemann D, Välimäki J. The dynamic pivot mechanism. *Econometrica* 2010;78(2):771–789.
16. Bergemann D, Välimäki J. Dynamic common agency. *J Econ Theory* 2003;111(1):23–48.

17. Bergemann D, Välimäki J. Dynamic price competition. *J Econ Theory* 2006;127(1): 232–263.
18. Cavallo R, Parkes DC, Singh S. Optimal coordinated planning among self-interested agents with private state. Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI'06); 2006; Cambridge (MA). pp. 55–62.
19. Cavallo R, Parkes DC, Singh S. Efficient mechanisms with dynamic populations and dynamic types. Unpublished manuscript, Harvard University; 2009.
20. Cavallo R. Efficiency and redistribution in dynamic mechanism design. Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08); 2008; Chicago (IL).
21. Athey S, Segal I. An efficient dynamic mechanism. Unpublished manuscript, Harvard University; 2007.
22. d'Aspremont C, Gérard-Varet L-A. Incentives and incomplete information. *J Public Econ* 1979;11(1):25–45.
23. Courty P, Li H. Sequential screening. *Rev Econ Stud* 2000;67(4):697–717.
24. Eső P, Szentes B. Optimal information disclosure in auctions and the handicap auction. *Rev Econ Stud* 2007;74(3):705–731.
25. Pavan A, Segal I, Toikka J. Dynamic mechanism design: revenue equivalence, profit maximization and information disclosure. Unpublished manuscript; Northwestern University; 2009.
26. Baron DP, Besanko D. Regulation and information in a continuing relationship. *Inf Econ Policy* 1984;1(3):267–302.
27. Pavan A, Segal I, Toikka J. Infinite-horizon mechanism design: the independent-shock approach. Unpublished manuscript, Northwestern University; 2010.
28. Kakade SM, Lobel I, Nazerzadeh H. An optimal dynamic mechanism for multi-armed bandit processes. Unpublished manuscript, Microsoft Research New England; 2010.
29. Deb R. Optimal contracting of new experience goods. Unpublished manuscript, Yale University; 2008.
30. Battaglini M. Long-term contracting with Markovian consumers. *Am Econ Rev* 2005;95(3):637–658.
31. Akan M, Ata B, Dana J. Revenue management by sequential screening. Unpublished manuscript, Carnegie Mellon University; 2009.
32. Aviv Y, Pazgal A. Optimal pricing of seasonal products in the presence of forward-looking consumers. *Manuf Serv Oper Manage* 2008;10(3):339–359.
33. Cavallo R. Mechanism design for dynamic settings. *ACM SIGecom Exchanges* 2009; 8, (2).
34. Doepke M, Townsend RM. Dynamic mechanism design with hidden income and hidden actions. *J Econ Theory* 2006;126(1):235–285.
35. Gershkov A, Moldovanu B. Learning about the future and dynamic efficiency. *Am Econ Rev* 2009;99(4):1576–1587.
36. Kittsteiner T, Moldovanu B. Priority auctions and queue disciplines that depend on processing time. *Manage Sci* 2005;51(2):236–248.
37. Parkes DC. Online mechanisms. In: Nisan Noam, Roughgarden Tim, Tardos Eva, *et al.*, editors. *Algorithmic game theory*. Cambridge: Cambridge University Press; 2007. Chapter 16. pp. 411–439.
38. Satterthwaite M, Shneyerov A. Dynamic matching, two-sided incomplete information, and participation costs: existence and convergence to perfect competition. *Econometrica* 2007;75(1):155–200.
39. Shen Z-JM, Su X. Customer behavior modeling in revenue management and auctions: a review and new research opportunities. *Prod Oper Manage* 2007;16(6):713–728.
40. Talluri KT, van Ryzin G. The theory and practice of revenue management. New York: Springer; 2005.
41. Vulcano G, van Ryzin G, Maglaras C. Optimal dynamic auctions for revenue management. *Manage Sci* 2002;48(11):1388–1407.

DYNAMIC MODELS FOR ROBUST OPTIMIZATION

RUKEN DÜZGÜN

AURÉLIE THIELE

Department of Industrial and Systems
Engineering, Lehigh University,
Bethlehem, Pennsylvania

TRADITIONAL MODELS OF DYNAMIC OPTIMIZATION

Stochastic Programming

Information in real-life applications is often revealed in stages, forcing the manager to make decisions with only limited knowledge of the data, and to adjust his strategy as he observes the realization of random parameters such as customer demand over time. Dantzig [1] first investigated decision making under uncertainty in the 1950s, pioneering the field now known as stochastic programming (SP). This methodology assumes that parameters are random but obey a known discrete distribution and that the decision maker minimizes (or maximizes) the expected objective value over the possible scenarios. The most frequent setup has two stages, with two groups of decision variables: *here-and-now* (or first-stage) variables, which represent decisions made before the manager can observe the resolution of the uncertainty; and *wait-and-see* (or second-stage) variables, which represent recourse actions. The SP problem is linear but of potentially very large size, motivating the use of structure-specific solution methods. The structure here is that the constraints either connect first-stage decision variables with each other, or first-stage decision variables with second-stage decision variables for a specific scenario, but never second-stage decision variables for different scenarios. Such a structure is said to be L-shaped because of the position of the nonzero elements in the coefficient matrix. The second-stage problem, formulated for

the given first-stage variables, is called the *recourse problem*. It can be shown to be convex and piecewise linear. The main algorithm used to solve this problem relies on generating pieces of the recourse function as needed, which can be interpreted as a delayed constraint generation algorithm. The reader is referred to Birge and Louveaux [2], Kall and Wallace [3], Prékopa [4], and Ruszczyński and Shapiro [5], and the references therein for a wide-ranging treatment of SP.

If recourse decisions are required to be integer (which is called SP with integer recourse), the integrality constraints increase the complexity of the second-stage problem significantly and the master problem becomes very hard to solve. Stochastic integer programs (where some variables are forced to be integer, in the first and/or second stages) are computationally intractable. Algorithms used to solve these problems include variants of the L-shaped method based on Benders decomposition [6–9] and branch-and-bound [10], among others.

Drawbacks. Under the assumption that the stochastic parameters are independently distributed, Dyer and Stougie [11] show that the two-stage SP problems are NP-hard. Furthermore, the number of scenarios grows exponentially with the number of parameters when the latter are independent, creating tractability issues. (For instance, a retailer considering 15 independent products with demand for each taking three possible values would need to generate over 14 million scenarios.) Moreover, it is often difficult to estimate probability distributions accurately. Even when the distributions of random parameters for past time periods can be estimated with a high degree of precision using historical data, these distributions might differ from future ones in unpredictable ways due to changing conditions. The difficulties faced by two-stage SP are compounded in the multistage case, where multiple piecewise linear recourse functions, corresponding to different time

periods, need to be approximated by generating linear pieces as needed. Shapiro and Nemirovski [12] provide an in-depth discussion of the complexity of two-stage and multistage SP problems, and argue that multistage SP problems are, in general, intractable.

Dynamic Programming

Dynamic programming (DP) deals with multistage decision making. The key idea is that the manager behaves optimally at all time periods, which reduces the number of strategies to be considered. This is formulated in mathematical terms through a system of recursive equations known as the *Bellman equations*:

$$\begin{aligned} V_t(S_t) &= \min_{x_t \in X_t} \{C_t(S_t, x_t) + E\{V_{t+1}(S_{t+1}|S_t, x_t)\}\} \\ \forall S_t &\in \mathcal{S}_t, \end{aligned} \quad (1)$$

where $V_t(S_t)$ is the *optimal* value (cost-to-go function) associated with being in state S_t in stage t and S_{t+1} is the state at the next time period. $\mathcal{S}_t$ is the space of possible states at time t . Intuitively, the equations state that the manager will follow the optimal strategy from time $t + 1$ onward, so that the problem at time t reduces to minimizing the current costs plus the optimal costs incurred from the next time period up to the end of the time horizon.

Solving Problem (1) requires the investigation of structural properties of the value functions to guarantee global optimality. DP is most insightful when the optimal policy can be proved to have a certain structure and the decision maker only has to compute the parameters defining that policy (for instance, basestock levels in dynamic inventory management under uncertainty [13]). It is important to note that DP leads to optimal *policies*, rather than optimal numbers, so that the decision maker does not need to resolve his/her problem as time progresses and uncertainty is revealed: he/she simply implements the precomputed policy that corresponds to the state and the time period he/she is in. Because DP relies on the fact that the decision maker will act on new information

at the next stage, it is sometimes called *closed-loop optimization* (new observations are incorporated in a feedback loop), in contrast to open-loop optimization, which generates numbers rather than policies and where the problem needs to be resolved at each time period to capture new information. (An advantage of open-loop optimization is that it is much less computationally demanding.) Readers may refer to Bertsekas [13] and Powell [14] for detailed treatments of DP.

Drawbacks. Value functions need to be stored for each time period and each possible state at that time period, creating severe dimensionality challenges as the size of the state space and/or the time horizon increases. This is known as the *curse of dimensionality*, which makes DP impractical in many applications. Approximate DP attempts to overcome these shortcomings [14,15], but remains hard to implement. In addition, as for SP, it can be difficult to estimate the underlying probability distributions accurately.

In summary, there are two main issues with the traditional methods of dynamic decision making under uncertainty:

1. Probability distributions are difficult to estimate accurately.
2. Dimensionality issues arise even when the distributions of the random parameters are exactly known.

We show in subsequent sections of this article how robust optimization (RO) can address these issues.

POSSIBLE DEFINITIONS FOR ROBUST OPTIMIZATION

RO is another method that addresses data uncertainty. Unlike SP and DP, RO models uncertainty assuming that uncertain parameters belong to a bounded, convex uncertainty set. While SP minimizes expected cost (or maximize expected revenue), RO is a worst-case analysis and minimizes the maximum value of the objective over the uncertainty set. This approach was pioneered by Soyster [16] in the 1970s, although he did not

call it “robust optimization” at the time. He assumed that each uncertain parameter took values in an uncertainty *interval* and proposed an optimization model to generate feasible solutions for the worst case. Because the model led to each parameter being equal to its worst-case value, it was thought to be too conservative for implementation in business; however, the issue of solution feasibility has remained an important topic of research, which saw critical advances in the 1990s.

Scenario-Based Robust Optimization

The expression “robust optimization” became popular in the mid-1990s, when Mulvey *et al.* [17] proposed a scenario-based description of the problem data and penalized the variance of the optimal solution for any realization of the scenarios, given a certain level of risk expressed by the decision maker. This approach does not match Soyster’s and is not in line with later uses of the expression “robust optimization,” which is described below. We mention it here because of its relation to SP as well as to avoid confusion for the reader familiar with Mulvey *et al.* [17] and subsequent approaches called RO.

Consider the LP optimization model

$$\begin{aligned} \min_{x \in \mathbb{R}_1^n, y \in \mathbb{R}_2^m} \quad & c^T x + d^T y \\ \text{s.t.} \quad & Ax = b, \\ & Fx + Gy = h, \\ & x, y \geq 0. \end{aligned} \quad (2)$$

Let $\Omega = (1, 2, \dots, S)$ be the set of scenarios with p_s probability of scenario s , where each scenario $s \in \Omega$ is associated with the vector (d_s, F_s, G_s, h_s) of realizations for the uncertain coefficients. y_s is the second-stage control vector for scenario $s \in \Omega$; also, we introduce a new vector z_s , which measures the extent of the constraint infeasibility under data scenario s . The objective function $\xi = c^T x + d^T y$ becomes a random variable that takes values $\xi_s = c^T x + d_s^T y_s$ with probability p_s for $s \in \Omega$.

The robust counterpart (RC) of Problem (2) using the Mulvey–Vanderbei–Zenios framework can then be formulated as

$$\begin{aligned} \min \quad & \sigma(x, y_1, \dots, y_s) + \omega \cdot \rho(z_1, \dots, z_s) \\ \text{s.t.} \quad & Ax = b, \\ & F_s x + G_s y_s + z_s = h_s, \quad \forall s \in \Omega \\ & x, y_s \geq 0 \quad \forall s \in \Omega. \end{aligned} \quad (3)$$

The $\rho(z_1, \dots, z_s)$ term in the objective function is a feasibility penalty function, whereas $\sigma(x, y_1, \dots, y_s)$ is the optimality robustness term. (The scalar ω captures the decision maker’s trade-off between the two goals of feasibility and optimality.) If the solution is “close” to the optimal one for any realization of the scenario $s \in \Omega$, then it is referred to as *solution-robust*. If it remains “almost” feasible for any scenario realization, then it is referred to as *model-robust*. The authors suggest possible functions in their work [17] to help users define solution-robustness and model-robustness. For instance, for moderate- to high-risk decisions under uncertainty, they advocate the use of the average objective plus a constant (λ) times its variance as an appropriate solution-robustness term:

$$\begin{aligned} \sigma(x, y_1, \dots, y_s) \\ = \sum_{s \in S} p_s \xi_s + \lambda \cdot \sum_{s \in S} p_s \left(\xi_s - \sum_{s' \in S} p_{s'} \xi_{s'} \right)^2 \end{aligned}$$

An example of feasibility penalty function is

$$\rho(z_1, \dots, z_s) = \sum_{s \in S} p_s z_s^T z_s$$

where both positive and negative violations of the control constraints are equally penalized.

Mulvey *et al.* [17] apply this technique in their work to applications such as power capacity expansion, matrix balancing and image reconstruction, air-force airline scheduling, scenario immunization for financial planning, and minimum-weight structural design. Other applications include capacity expansion of telecommunication networks [18], a more in-depth look at power capacity expansion [19], and the portfolio management of callable bonds [20].

Drawbacks. This approach to RO changes the structure of the deterministic model; the robust model is no longer linear. In addition, it suffers from the same dimensionality issues encountered in SP.

Set-Based Robust Optimization

Ben-Tal and Nemirovski [21–23], El-Ghaoui and Lebret [24], and El-Ghaoui *et al.* [25] focus w.l.o.g. on uncertainty in the constraints of mathematical programming problems and define robust solutions as solutions that are feasible for the worst-case value of the parameters within an uncertainty set. They use ellipsoidal uncertainty sets and propose a tractable mathematical reformulation that turned linear programming problems into second-order cone problems, while reducing the conservatism of Soyster’s [16] approach. Furthermore, Ben-Tal and Nemirovski [26] study RO applied to conic-quadratic and semidefinite programming. The reader is referred to Ben-Tal *et al.* [27] for an overview of RO, with an emphasis on ellipsoidal sets. One drawback of that framework is that it increases the complexity of the nominal problem.

Bertsimas and Sim [28,29] and Bertsimas *et al.* [30] investigate the case where the uncertainty set is a polyhedron. Specifically, the uncertainty set consists in range forecasts (confidence intervals) for each parameter and a constraint called a budget-of-uncertainty constraint, which limits the number of coefficients that can take their worst-case value. The approach preserves the degree of complexity of the problem (the RC of a linear problem is linear) and allows the decision maker to control the degree of conservatism of the solution. Bertsimas and Sim [28] also provide a probabilistic guarantee of constraint violation. A drawback of the approach is that it adds auxiliary variables and constraints to the initial formulation. Bertsimas and Thiele [31] survey the RO literature up to 2006, especially for polyhedral uncertainty sets.

Those early works spearheaded significant research efforts on the theory and practice of RO. Bertsimas and Brown [32] provide a methodology to construct uncertainty sets within the framework of RO, using the decision maker’s risk preferences expressed

by a coherent risk measure. RO has been applied to inventory management [33,34], revenue management [35], portfolio management [36,37], telecommunications [38], among others. Notice that while RO was initially developed to address data ambiguity (i.e., uncertain parameters whose value was unknown but constant), the methodology was subsequently applied to random variables with uncertain distributions; in the early 2000s, Bertsimas and Thiele [33] was the first work to model random variables as uncertain parameters and apply RO to the deterministic problem. In the context of multistage decision making, this leads to *open-loop problems*. A drawback is that these problems had to be reoptimized at each time period, as a necessary trade-off to achieve tractable formulations.

In addition, if the knowledge of the probability distributions driving the random variables is imprecise, but the manager knows that the distribution belongs to a family of distributions, then it is possible to implement an RO approach to the uncertain probability density functions themselves. The manager optimizes his/her objective over the worst-case value of the probabilities. This is referred to in the literature as the *minimax* or *min-max* approach (rather than RO) [39–41]. Thiele [42] applies the RO approach using polyhedral uncertainty sets to stochastic optimization problems and provides theoretical insights into the solution. Unfortunately, the RO approach to SP is only as tractable as the underlying SP problem for the nominal probabilities.

Goh and Sim [43] suggest tractable approximations to distributionally RO using piecewise linear decision rules. Ben-Tal *et al.* [44] propose a framework for RO that extends the standard notion of robustness by allowing the decision maker to vary the protection level across the uncertainty set. This captures the fact that at least some partial probabilistic description of the world is available in many applications such as finance. The approach in their work [44] allows for different performance guarantees for different subsets within $\mathcal{P}$, where $\mathcal{P}$ represents the set of possible underlying probability measures for the random variables. For

instance, performance guarantees of a portfolio can then be linked in a smooth way to the performance of the market as a whole. The authors call this new approach the *soft robust approach* and show that it preserves convexity properties of the nominal problem.

Dimensionality issues explain why researchers have focused on the alternative view of RO, which incorporates uncertainty to the *deterministic* formulation of the problem, instead of considering a stochastic description of uncertainty. While the tractability of the RO approach is appealing, the limitations of open-loop policies have incited researchers to investigate dynamics models in RO in more depth. Of particular interest has been the development and analysis of decision rules with a specific structure, which are described in the following section.

ROBUST DYNAMIC OPTIMIZATION

Linear Decision Rules and Adjustable Optimization

Ben-Tal *et al.* [45] extend the scope of RO by introducing the adjustable RO methodology. Consider the uncertain linear programming problem

$$\left\{ \min_x \left\{ c^T x : Ax \leq b \right\} \right\}_{\zeta \equiv [A, b, c] \in \mathcal{Z}}, \quad (4)$$

where $\zeta \equiv [A, b, c]$ lies in the uncertainty set $\mathcal{Z} \subset \mathbb{R}^n \times \mathbb{R}^{m \times n} \times \mathbb{R}^m$, a nonempty compact convex set. (This section uses the same notation as that is used in Ben-Tal *et al.*'s work [45].)

The RC of the uncertain problem (4) is defined as

$$\min_x \left\{ \sup_{\zeta \equiv [A, b, c] \in \mathcal{Z}} (c^T x) : Ax - b \leq 0 \right. \\ \left. \forall \zeta \equiv [A, b, c] \in \mathcal{Z} \right\}.$$

In the traditional RO approach, all the variables represent “here-and-now” decisions, so that the optimal solution x must be chosen to optimize the worst-case

objective over all possible values of ζ in $\mathcal{Z}$. Ben-Tal *et al.* [45] suggest to incorporate “wait-and-see” decisions; that is, decisions that are taken after the uncertainty is realized. In the RO terminology, the variables that correspond to recourse action are called *adjustable*, while the others are called *nonadjustable*. The vector x in Equation (4) is partitioned according to nonadjustable (u) and adjustable (v) variables such that $x = (u^T, v^T)^T$ and Problem (4) becomes

$$\min_{(s, u), v} \left\{ s : c^T \begin{pmatrix} u \\ v \end{pmatrix} \leq s, Uu + Vv \leq b \right\}_{[U, V, b, c] \in \mathcal{Z}},$$

where (s, u) represents the nonadjustable part of the solution. The uncertain LP problem (4) can be rewritten w.l.o.g. (after redefining parameters appropriately) as

$$LP_{\mathcal{Z}} = \left\{ \min_{u, v} c^T u : Uu + Vv \leq b \right\}_{[U, V, b] \in \mathcal{Z}}. \quad (5)$$

The matrix V is called the *recourse matrix*. When V is not subject to uncertainty, Problem (5) is a *fixed-recourse* LP.

Definition 1. The adjustable robust counterpart (ARC) of the uncertain LP problem $LP_{\mathcal{Z}}$ is defined as

$$ARC : \min_u \left\{ c^T u : \forall \zeta = [U, V, b] \in \mathcal{Z} \exists v : \right. \\ \left. Uu + Vv \leq b \right\}. \quad (6)$$

ARC has a larger robust feasible set and hence is more flexible and less conservative than RC. On the other hand, it is a semi-infinite LP problem; it suffers from computational tractability issues and is NP-hard even for simple uncertainty sets. As a remedy, the authors introduce the affinely adjustable robust counterpart (AARC) of an uncertain LP by restricting the adjustable variables to be *affine* functions of the corresponding data. In other words, v is forced to be of the form

$$v = w + W\zeta$$

for some w, W parameters to be determined.

Definition 2. AARC of Equation (5) is defined as the optimization problem:

$$\min_{u,w,W} \left\{ c^T u : Uu + V(w + W\zeta) \leq b \right. \\ \left. \forall \zeta = [U, V, b] \in \mathcal{Z} \right\}. \quad (7)$$

Ben-Tal *et al.* [45] show that if the uncertainty set $\mathcal{Z}$ in a *fixed-recourse* problem $LP_{\mathcal{Z}}$ is computationally tractable (in the sense that a tractable separation oracle exists), then the AARC is also computationally tractable, that is, polynomially solvable. When the uncertainty set is also “well structured” (i.e., given by a list of linear matrix inequalities such as polyhedral, conic-quadratic, or semidefinite representations), then the corresponding AARC is also “well structured” and thus can be solved by linear or semidefinite programming techniques. On the other hand, if the recourse matrix V is subject to uncertainty, the AARC of $LP_{\mathcal{Z}}$ can become computationally intractable. The authors [45] show that in many such cases, AARC has a tight computationally tractable approximation (which is an explicit semidefinite programming problem). Ben-Tal *et al.* [46] apply the AARC heuristic to the two-echelon multiperiod supply chain problem (specifically, a retailer–supplier flexible commitment (RSFC) problem) and derive a single deterministic convex optimization problem that is either a linear or a conic-quadratic problem.

Ordóñez and Zhao [47] identify the conditions on the uncertainty set that would lead to a tractable ARC approach in the case of a robust capacity expansion problem (RCEP) for network flows. They consider the classic network flow problem with additional decision variables representing arc capacity expansions. The capacity expansion problem can be written as

$$z_D(b, c) = \min_{x,y} c^T x \\ \text{s.t. } Nx = b \\ x \leq u + y \\ d^T y \leq q \\ x, y \geq 0, \quad (8)$$

where $c \in \mathbb{R}^m$ are the transportation cost coefficients, $x \in \mathbb{R}^m$ are the arc flow variables, and $y \in \mathbb{R}^m$ are the decision variables representing the new arc capacities. $N \in \mathbb{R}^{n \times m}$ is the node-arc incidence matrix, initial arc capacities are given by $u \in \mathbb{R}^m$ and the demand–supply vector is given by $b \in \mathbb{R}^n$. Expanding the capacity of arc i costs d_i where q is the total budget for investment. The demand b and the travel times c belong to closed, convex and bounded uncertainty sets U_b and U_c , respectively. It is assumed that the y variables are determined prior to the realization of the uncertain data (b, c) and the flow variables x will adapt to the realized data. Then, the ARC of RCEP is defined as

$$z_{\text{ARC}} = \min_{y, \gamma} \gamma \\ \text{s.t. } d^T y \leq q \\ y \geq 0 \\ \forall c \in U_c, b \in U_b \exists x: \begin{cases} Nx = b \\ 0 \leq x \leq u + y \\ c^T x \leq \gamma. \end{cases}$$

Ordóñez and Zhao [47] present three cases (problem with fixed demand, single commodity problem with uncertain demand, and multicommodity problem with uncertain demand), where the RCEP is formulated as a conic problem and solved by interior point methods in polynomial time.

Shapiro and Nemirovski [12] state that the main reason for using linear decision rules is their tractability, but linear decision rules are rarely optimal. In other words, there is no guarantee that the true optimal solution is close to the one given by the linear decision rule and the optimality gap is not known. Linear decision rules may perform poorly or even lead to infeasible instances even in the case of complete recourse. Chen *et al.* [48] give the following example: suppose that the support of $\tilde{z}$ is $\mathcal{W} = \{-\infty, \infty\}$. Then, the nonnegativity constraint on

$$w(\tilde{z}) = w^0 + \sum_{k=1}^N w^k \tilde{z}_k$$

implies that

$$w^k = 0 \quad \forall k \in \{1, \dots, N\},$$

and the decision rule reduces to $w(\tilde{z}) = w^0$, which is independent of the underlying uncertainty and may lead to infeasibility.

Ben-Tal *et al.* [49] extend the AARC approach by relaxing the “uncertain-but-bounded” assumption of RO that characterizes the uncertain parameters. Requiring that all realizations of the uncertain data lie in the uncertainty set may be pessimistic (leading to overly large sets) or infeasible if the random variables have unbounded support. The approach in Ben-Tal *et al.* [49] exhibits controlled performance degradation for “large deviations” in the uncertain data. Assume that the uncertainty set $\mathcal{Z}$ represents the typical range for the uncertain data. When $\zeta \in \mathcal{Z}$, the solution must satisfy the constraints of the problem. When the data ζ falls outside the uncertainty set $\mathcal{Z}$, the violation of the constraints should not exceed a prescribed multiple of the deviation of the data from its normal range. These multiples serve as “global sensitivities” of the constraints. Ben-Tal *et al.* [49] call this extension of AARC the *comprehensive RC* of uncertain linear problems.

Piecewise Constant Rules and Adaptive Optimization

Bertsimas and Caramanis [50] introduce a variable degree of adaptability on the grounds that complete adaptability can be too expensive and assuming the exact realization of uncertainty (which is required to implement AARC above) is overly optimistic. The generic problem the authors consider is

$$\min \left\{ c^T x : Ax \geq b, \forall A \in \mathcal{P} \right\}, \quad (9)$$

where $\mathcal{P}$ is a polytope. The reoptimization formulation is given by

$$\max_{A \in \mathcal{P}} \left\{ \min \left\{ c^T x : Ax \geq b \right\} \right\}. \quad (10)$$

The authors introduce the concept of k -adaptability (Problem (11)), which is formulated as a disjunctive optimization problem with infinitely many constraints. The decision maker selects k solutions and commits to one solution after uncertainty is revealed; at least one of the k solutions must

be feasible regardless of the realization of the uncertainty.

$$\begin{aligned} \min \quad & \max \{ c^T x_1, c^T x_2, \dots, c^T x_k \} \\ \text{s.t.} \quad & [Ax_1 \geq b \text{ or } Ax_2 \geq b \text{ or } \dots \text{ or } Ax_k \geq b] \\ & \forall A \in \mathcal{P}. \end{aligned} \quad (11)$$

Problem (11) is equivalent to the k -partitioning problem (12) where the uncertainty set $\mathcal{P}$ has been partitioned into a finite number k of regions: $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \dots \cup \mathcal{P}_k$. The decision maker learns which region of the partition will contain the realization of the uncertainty before he has to commit to a solution.

$$\min_{\mathcal{P} = \mathcal{P}_1 \cup \dots \cup \mathcal{P}_k} \left[\begin{array}{l} \min \max \{ c^T x_1, c^T x_2, \dots, c^T x_k \} \\ \text{s.t.} \quad Ax_1 \geq b, \forall A \in \mathcal{P}_1 \\ \quad \quad \quad \vdots \\ \quad \quad \quad Ax_k \geq b, \forall A \in \mathcal{P}_k \end{array} \right]. \quad (12)$$

Problem (12) represents an application of finite adaptability to *single-stage optimization*. For two-stage optimization as in the ARC model (6), the problem becomes

$$\min_{\mathcal{P} = \mathcal{P}_1 \cup \dots \cup \mathcal{P}_k} \left[\begin{array}{l} \min : c^T u \\ \text{s.t.} \quad Uu + Bv_1 \geq b, \forall (U, V) \in \mathcal{P}_1 \\ \quad \quad \quad \vdots \\ \quad \quad \quad Uu + Bv_k \geq b, \forall (U, V) \in \mathcal{P}_k \end{array} \right]. \quad (13)$$

Let V_{RO} , V_{ReOpt} , V_{adapt}^k be the optimal objective values of Problems 9, 10, and (12), respectively. We have

$$V_{\text{RO}} \geq V_{\text{adapt}}^k \geq V_{\text{ReOpt}}.$$

Furthermore, with an additional continuity assumption, $\lim_{k \rightarrow \infty} V_{\text{adapt}}^k = V_{\text{ReOpt}}$. Hence, finite adaptability bridges the gap between RO (total lack of adjustability, where all decision variables are here-and-now) and reoptimization (total adjustability, where all decision variables are wait-and-see).

While the k -adaptability problem was formulated above as a disjunctive program,

it can also be formulated as a bilinear optimization program. For $k = 2$, the bilinear optimization problem becomes

$$\begin{aligned} \min \quad & \max\{c^T x_1, c^T x_2\} \\ \text{s.t.} \quad & u_{i,j}[(A^l x_1)_i - b_i] + (1 - u_{i,j})[(A^l x_2)_j - b_j] \geq 0 \\ & \forall 1 \leq i, j \leq m \quad \forall 1 \leq l \leq K \\ & 0 \leq u_{i,j} \leq 1, \quad \forall 1 \leq i, j \leq m. \end{aligned}$$

m is the number of rows of the constraint matrix A and K is the number of extreme points of the uncertainty set $\mathcal{P}$, where $\mathcal{P}$ is defined as the convex hull of its extreme points: $\mathcal{P} = \text{conv}(A^1, \dots, A^K)$. Bertsimas and Caramanis [50] prove that obtaining the optimal partition $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$ in 2-adaptability is NP-hard in general, provide a hierarchy of the levels of adaptability, and propose a heuristic tractable algorithm.

Bertsimas and Goyal [51,52] consider two-stage adaptive optimization problems and investigate the power and limitations of robust solutions and affine policies. In their work [51], it is shown that the RO approach is a good approximation to solving the corresponding two-stage mixed integer stochastic optimization problem to optimality. They compare the optimal cost of the robust problem to the optimal costs of the stochastic problem and the adaptability problem. If the uncertainty set and the probability distribution over the uncertain set are symmetric, and if the second-stage variables are continuous variables, the optimal cost of the robust problem is at most twice the expected cost of the optimal two-stage solution to the stochastic problem. In another work [52], Bertsimas and Goyal show that an affine policy is optimal if the uncertainty set $u \in \mathbb{R}_+^m$ is a convex combination of $m + 1$ affinely independent points in $\mathbb{R}_+^m$ and this optimality result is almost tight. They also prove that if the uncertainty set is a polytope, the worst-case cost occurs at an extreme point of the uncertainty set. On the other hand, for uncertainty sets with $m + 2$ nonzero extreme points, the affine policy is suboptimal. The authors also give a lower and upper bound for the performance of an optimal affine policy compared to an optimal fully adaptable two-stage solution [52].

Other Approaches

Thiele *et al.* [53] develop an RO approach for generic two-stage stochastic problems with uncertainty on the right-hand side. The authors apply a cutting-plane algorithm based on Kelley's algorithm to the robust linear problem with general recourse and test the methodology on a multi-item newsvendor problem and production planning example where the demand is uncertain but must be met.

Chen *et al.* [54] generalize the robust linear optimization frameworks of Ben-Tal and Nemirovski [23] and Bertsimas and Sim [28] by introducing a new uncertainty set that captures the asymmetry of the underlying random variables through the use of new deviation measures (forward and backward deviations). They integrate these deviation measures to the uncertainty set and obtain solutions to chance-constrained problems. Applying the linear decision rule of Ben-Tal *et al.* [45], Chen *et al.* [54] present a tractable RO approach in order to find less conservative feasible solutions for SP with chance constraints. This framework also leads to the computational scalability of multistage stochastic models. Chen *et al.* [48] propose a novel framework to approximate multistage stochastic optimization by introducing two new piecewise linear decision rules. The first one is called *deflected linear decision rule*, which is best suited for SP problems with semicomplete recourse and provides a tighter approximation of the original objective function than the linear decision rule. The second one is called *segregated linear decision rule*, which is best suited for SP problems with general recourse. When combined with the first rule, the *segregated linear decision rule* exhibits better performance than both linear and deflected linear decision rules. Under these new piecewise linear rules, the authors show that the computational tractability of the problems is preserved. Chen and Zhang [55] and See and Sim [56] extend the theory of RO to approximate solutions of multistage problems. See and Sim [57] apply these RO approaches to multiperiod inventory control.

As a novel application, Düzgün and Thiele [58] apply RO to R&D project management

problems where investments are done in stages (such as the phases of a drug trial) and cash flows are uncertain. They describe how to address the modeling challenges raised in the RO framework by the discrete, stochastic outcome of each phase (success or failure) and the computational challenges raised by the use of binary variables to represent project selection.

CONCLUSIONS

Dynamic models of RO have been the focus of significant research efforts, both in terms of theory and applications. Piecewise constant, linear and piecewise linear decision rules have been thoroughly investigated. The use of RO has allowed for the design of an elegant, intuitive, and tractable framework to sequential decision making under uncertainty.

Acknowledgments

Work supported in part by NSF Grant CMMI-0757983 and an IBM Faculty Award.

REFERENCES

1. Dantzig GB. Linear programming under uncertainty. *Manage Sci* 1955;1(3-4): 197–206.
2. Birge JR, Louveaux F. Introduction to stochastic programming. Springer series in operations research. New York: Springer; 1997.
3. Kall P, Wallace SW. Stochastic programming. Wiley-Interscience series in systems and optimization. 2nd ed. Chichester: John Wiley & Sons Ltd; 1994.
4. Prékopa A. Stochastic programming. Dordrecht, Boston: Kluwer Academic Publishers; 1995.
5. Ruszczyński A, Shapiro A. Stochastic programming. Volume 10, Handbooks in operations research and management science. Amsterdam: North-Holland; 2003.
6. Laporte G, Louveaux FV. The integer l-shaped method for stochastic integer programs with complete recourse. *Oper Res Lett* 1993;13:133–142.
7. Carøe CC, Tind J. L-shaped decomposition of two-stage stochastic programs with integer recourse. *Math Program* 1998;83:451–464.
8. Sen S, Sherali HD. Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming. *Math Program* 2006;106(2):203–223.
9. Ntairo L, Sen S. A branch-and-cut algorithm for two-stage stochastic mixed-binary programs with continuous first-stage variables. *Int J Comput Sci Eng* 2007;3(3):232–241.
10. Ahmed S, Tawarmalani M, Sahinidis NV. A finite branch-and-bound algorithm for two-stage stochastic integer programs. *Math Program* 2004;100:355–377.
11. Dyer M, Stougie L. Computational complexity of stochastic programming problems. *Math Program* 2006;106:423–432.
12. Shapiro A, Nemirovski A. On complexity of stochastic programming problems. Volume 99, Continuous optimization. New York: Springer; 2005. pp. 111–146.
13. Bertsekas DP. Volume 1, Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2005.
14. Powell WB. Approximate dynamic programming: solving the curses of dimensionality. Wiley-Interscience series in probability and statistics. Hoboken (NJ): John Wiley & Sons, Inc.; 2007.
15. Bertsekas DP. Volume 2, Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2007.
16. Soyster AL. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper Res* 1973;21:1154–1157.
17. Mulvey JM, Vanderbei RJ, Zenios SA. Robust optimization of large-scale systems. *Oper Res* 1995;43(2):264–281.
18. Laguna M. Applying robust optimization to capacity expansion of one location in telecommunications with demand uncertainty. *Manage Sci* 1998;44(11):101–110.
19. Malcolm SA, Zenios SA. Robust optimization for power systems capacity expansion under uncertainty. *J Oper Res Soc* 1994; 45(9):1040–1049.
20. Vassiadou-Zeniou C, Zenios SA. Robust optimization models for managing callable bond portfolios. *Eur J Oper Res* 1996;91(2): 264–273.
21. Ben-Tal A, Nemirovski A. Robust convex optimization. *Math Oper Res* 1998;23(4): 769–805.
22. Ben-Tal A, Nemirovski A. Robust solutions of uncertain linear programs. *Oper Res Lett* 1999;25:1–13.

23. Ben-Tal A, Nemirovski A. Robust solutions of linear programming problems contaminated with uncertain data. *Math Program* 2000;88:411–424.
24. El-Ghaoui L, Lebret H. Robust solutions to least-square problems to uncertain data. *SIAM J Matrix Anal Appl* 1997;18(4):1035–1064.
25. El-Ghaoui L, Oustry F, Lebret H. Robust solutions to uncertain semidefinite programs. *SIAM J Optim* 1998;9:33–52.
26. Ben-Tal A, Nemirovski A. Robust optimization: methodology and applications. *Math Program* 2002;92:453–480.
27. Ben-Tal A, El-Ghaoui L, Nemirovski A. Robust optimization. Princeton series in applied mathematics. Princeton (NJ): Princeton University Press; 2009.
28. Bertsimas D, Sim M. The price of robustness. *Oper Res* 2004;52(1):35–53.
29. Bertsimas D, Sim M. Robust discrete optimization and network flows. *Math Program B* 2003;98:49–71.
30. Bertsimas D, Pachamanova D, Sim M. Robust linear optimization under general norms. *Oper Res Lett* 2004;32:510–516.
31. Bertsimas D, Thiele A. Robust and data-driven optimization: modern decision-making under uncertainty. *Tutorials Oper Res* 2006;95–122.
32. Bertsimas D, Brown DB. Constructing uncertainty sets for robust linear optimization. *Oper Res* 2009;57(6):1483–1495.
33. Bertsimas D, Thiele A. A robust optimization approach to inventory theory. *Oper Res* 2006;54(1):150–168.
34. Bienstock D, Ozbay N. Computing robust base-stock levels. *Discrete Optim* 2008;5: 389–414.
35. Adida E, Perakis G. A robust optimization approach to dynamic pricing and inventory control with no backorders. *Math Program Ser B* 2006;107(1):97–129.
36. Bertsimas D, Pachamanova D. Robust multi-period portfolio management with transaction costs. *Comput Oper Res* 2008;35(1):3–17.
37. Kawas B, Thiele A. A robust optimization approach to log-robust portfolio management. *OR Spectr* 2010. In press.
38. Ye W, Ordoñez F. Robust optimization models for energy-limited wireless sensor networks under distance uncertainty. *IEEE Trans Wirel Commun* 2008;7(6):2161–2169.
39. Dupačová J. Minimax stochastic programs with nonseparable penalties. Optimization techniques (Proceedings of the 9th IFIP Conference on Optimization Techniques; 1979; Warsaw), Part 1. Volume 22, Lecture notes in control and information sciences. Berlin, Heidelberg: Springer; 1980. pp. 157–163.
40. Shapiro A, Kleywegt A. Minimax analysis of stochastic problems. *Optim Methods Softw* 2002;17:523–542.
41. Shapiro A. Worst-case distribution analysis of stochastic programs. *Math Program Ser B* 2006;107:91–96.
42. Thiele A. Robust stochastic programming with uncertain probabilities. *IMA J Manage Math* 2008;19(3):289–321.
43. Goh J, Sim M. Distributionally robust optimization and its tractable approximations. *Oper Res* 2010. In press.
44. Ben-Tal A, Bertsimas D, Brown DB. A soft robust model for optimization under ambiguity. *Oper Res* 2010. In press.
45. Ben-Tal A, Goryashko A, Guslitzer E, *et al.* Adjustable robust solutions of uncertain linear programs. *Math Program Ser A* 2004;99:351–376.
46. Ben-Tal A, Golany B, Nemirovski A, *et al.* Retailer-supplier flexible commitments contracts: a robust optimization approach. *Manuf Serv Oper Manage* 2005;7(3):248–271.
47. Ordoñez F, Zhao J. Robust capacity expansion of network flows. *Networks* 2007; 50(2):136–145.
48. Chen X, Sim M, Sun P, *et al.* A linear decision-based approximation approach to stochastic programming. *Oper Res* 2008;56(2):344–357.
49. Ben-Tal A, Boyd S, Nemirovski A. Extending scope of robust optimization: comprehensive robust counterparts of uncertain problems. *Math Program Ser B* 2006;107:63–89.
50. Bertsimas D, Caramanis C. Finite adaptability in linear optimization. Technical Report. Cambridge (MA): Massachusetts Institute of Technology; 2005.
51. Bertsimas D, Goyal V. On the power of robust solutions in two-stage stochastic and adaptive optimization problems. *Math Oper Res* 2010;35:284–305.
52. Bertsimas D, Goyal V. On the power and limitations of affine policies in two-stage adaptive optimization. *Math Program* 2010. In press.
53. Thiele A, Terry T, Epelman M. Robust linear optimization with recourse. Technical Report. Lehigh University and University of Michigan; 2009.

- 54. Chen X, Sim M, Sun P. A robust optimization perspective of stochastic programming. *Oper Res* 2007;55(6):1058–1071.
- 55. Chen X, Zhang J. Uncertain linear programs: extended affinely adjustable robust counterparts. *Oper Res* 2009;57(6):1469–1482.
- 56. See W, Sim M. Goal-driven optimization. *Oper Res* 2009;57(2):342–357.
- 57. See C-T, Sim M. Robust approximation to multi-period inventory management. *Oper Res* 2010;58(3):583–594.
- 58. Düzgün R, Thiele A. Robust optimization with multiple ranges: theory and application to R&D project selection. Technical Report. Bethlehem (PA): Lehigh University; 2010.

DYNAMIC PRICING STRATEGIES FOR MULTIPRODUCT REVENUE MANAGEMENT PROBLEMS

COSTIS MAGLARAS
Graduate School of Business,
Columbia University,
New York, New York

INTRODUCTION

We consider a firm that owns a fixed capacity of a certain resource that is consumed in the process of producing or offering multiple products or services, and which must be consumed over a finite time horizon. The firm's problem is to maximize its total expected revenues by dynamically selecting the price of each of its products over time. Revenue management problems of that sort gained interest in the late 1970s in the context of the airline industry, and have since been successfully introduced in a variety of other areas such as hotels, cruise lines, rental cars, and retail. This article considers three variants of the above problem. In the first, the firm offers just one product and is assumed to be a monopolist or to operate in a market with imperfect competition, and thus to have power to influence the demand for each product by varying its price. In this setting, the firm's problem is to choose a dynamic pricing strategy for its product in order to optimize expected revenues. In the second variant, the firm offers many products that consume the capacity of the same scarce resource, and again has pricing power and seeks to optimize its expected revenues by choosing a multidimensional dynamic pricing policy. Finally, in the third variant, prices are assumed to be fixed either by the competition or through a higher-order optimization problem, and the firm's problem is now to choose a dynamic capacity allocation rule that controls when to accept new requests for each of these products. In the sequel, the first two will be referred to as *dynamic pricing*

problems, while the third will be referred to as a *capacity control* problem.

This article illustrates how these three problems can be reduced to a common formulation, thus connecting prior results that have appeared in the literature under a unified framework, and explores some of the consequences of this formulation. Specifically, we show that the multiproduct dynamic pricing and the capacity control problems can be recast within this common framework, and be treated as different instances of a single-product pricing problem for appropriately selected concave revenue functions. Broadly speaking, this is done by decoupling the revenue maximization problems in two parts: first, at each point in time the firm selects an aggregate capacity consumption rate from all products and, second, it computes the vector of demand rates to maximize instantaneous revenues subject to the constraint that all products jointly consume capacity at the aforementioned rate. The latter is akin to the basic microeconomics problem of resource allocation subject to a budget constraint, and gives rise to an appropriate *aggregate revenue rate function* in each case.

Adopting this common formulation, we review some of the key structural results regarding the monotonicity properties of the value function and the associated controls, which were derived in the literature [1–6]. We present the deterministic and continuous (fluid) approximations of the dynamic pricing problems described above, review their solution, the pricing heuristics that can be gleaned from them, and the performance bounds that they offer for the expected revenues of the stochastic and discrete revenue maximization of original interest.

The remainder of this article is structured as follows. This section concludes with a brief literature review. The section titled “Single-Product Dynamic Pricing Problem” formulates the single-product dynamic pricing problem and reviews some known structural results, and subsequently studies the deterministic and continuous (fluid)

analog of the dynamic pricing problem. The section titled “Multiple Products, Single Resource” studies the multiproduct variants described above. The section titled “Dynamic Pricing Network Revenue Management Problems” outlines how the same approach can be extended to a network setting. The section titled “Extensions” offers some concluding remarks, including interesting extensions of the baseline models we study. In terms of style of exposition, we focus on specifying the underlying models and providing sketches and intuition for the type of analysis that one may carry through, and refer the reader to appropriate references for details on lengthy derivations and proofs.

The papers by Elmghraby and Keskinocak [7], Bitran and Caldentey [8], and McGill and van Ryzin [9], and the book by Talluri and van Ryzin [10] provide comprehensive overviews of the areas of dynamic pricing and revenue management. The modeling framework adopted in this paper closely matches that of Gallego and van Ryzin [1,2]. Additional references on the capacity control formulation are Brumelle and McGill [11] and Lautenbacher and Stidhman [4]. The reduction of the multiproduct and capacity control problems to single-product pricing problems and their subsequent solutions are from Maglaras and Meissner [6]. The asymptotic analysis that is briefly reviewed in this article builds on the setup used in Refs 1 and 2 and Cooper [12], and the results reviewed here are adopted from [6]. The idea of efficient controls and that of a notion of an efficient frontier on how to choose the product level pricing and capacity control decisions to achieve a desired capacity absorption rate is from [6]. Related ideas have appeared in slightly different settings in Talluri and van Ryzin [13], in the context of a capacity control problem for a model with customer choice among products, and in Feng and Xiao [14,15], while studying pricing problems with a predetermined set of price points; our presentation of this topic borrows from [6]. The pricing and capacity control heuristics that we discuss have appeared in many references such as Gallego and van Ryzin [1,2], McGill and van Ryzin [9], Feng and Xiao [15], Lin *et al.* [6,16]. The feedback form of the static pricing policy

that is optimal for the deterministic and continuous (fluid) analog of the stochastic and discrete dynamic pricing problem is from [6] and the property of asymptotic optimality of the “resolving” pricing heuristic that is based on that feedback policy is also from [6].

SINGLE-PRODUCT DYNAMIC PRICING PROBLEM

Model

Consider a firm endowed with C units of capacity of a product that is to be sold over a finite horizon τ , and where capacity cannot be replenished up to that time. The salvage value of the remaining capacity at time τ is assumed to be 0. (A constant per-unit salvage value would also result in formulations similar to those developed below.) The firm is either a monopolist or is assumed to operate in a market with imperfect competition, and, in that, has power to influence the demand for each product by varying its menu of prices. Let $p(t)$ denote the price posted at time t . The demand process is assumed to be a nonhomogeneous Poisson process with rate vector λ determined through a demand function $\lambda(p(t))$, where $\lambda : \mathcal{P} \rightarrow \mathcal{L}$, $\mathcal{P} \subseteq \mathbb{R}$ is the set of feasible price vectors, and $\mathcal{L} = \{x \geq 0 : x = \lambda(p), p \in \mathcal{P}\} \subseteq \mathbb{R}_+$ is the set of achievable demand rate vectors. We assume that $\mathcal{L}$ is a convex set. For ease of exposition, the demand function $\lambda(\cdot)$ is assumed to be stationary; an extension to allow for non-stationarities could follow Gallego and van Ryzin [1,2], and Zhao and Zheng [5]. We consider *regular* demand functions that satisfy some additional conditions. In the sequel, x' denotes the transpose of any matrix x , for any real number y , $y^+ := \max(0, y)$, e is the vector of 1's of appropriate dimension and “a.s.” stands for almost surely.

Definition 1. A demand function is said to be regular if it is a continuously differentiable, bounded function, and: (a) $\lambda(p)$ is strictly decreasing in p , (b) $\lim_{p \rightarrow \infty} \lambda(p) = 0$, and (c) the revenue rate $p\lambda(p)$ is bounded for all $p \in \mathcal{P}$ and has a finite maximizer $\bar{p}$.

We assume that there exists a continuous inverse demand function $p(\lambda)$, $p : \mathcal{L} \rightarrow \mathcal{P}$,

which maps an achievable demand rate λ to the corresponding price $p(\lambda)$. This allows us to view the demand rate as the firm's control, and infer the appropriate price using the inverse demand function. The expected revenue rate can be expressed as a function of the vector of demand rates λ as $R(\lambda) := \lambda p(\lambda)$, and is assumed to be continuous, bounded, and strictly concave.

Example 1 [Linear Demand Model].

The demand for the product at price p is given by $\lambda(p) = \Lambda - bp$, where $\Lambda > 0$ is the market potential for the product and $b > 0$ is the price sensitivity parameter. The inverse demand and revenue functions are $p(\lambda) = (\Lambda - \lambda)/b$ and $R(\lambda) = \lambda(\Lambda - \lambda)/b$, respectively.

Example 2 [Logit Demand Model].

$\lambda(p) = \Lambda e^{-bp}/(1 + e^{-bp})$, $p(\lambda) = (1/b)\ln(\Lambda/\lambda - 1)$, and $R(\lambda) = (\lambda/b)\ln(\Lambda/\lambda - 1)$, respectively.

The problem that we address is roughly described as follows: given an initial capacity C , a selling horizon τ , and a demand function that maps a posted price to a corresponding instantaneous demand rate, the firm's goal is to choose a nonanticipating dynamic pricing strategy in order to maximize its total expected revenues.

We adopt a discrete-time formulation, that is, one where time has been discretized in small intervals of length δt , indexed by $t = 1, \dots, T$, such that $\mathbb{P}(\text{one product request in } [0, \delta t]) = \lambda \delta t + o(\delta t)$, $\mathbb{P}(\text{two product requests in } [0, \delta t]) = \lambda^2 (\delta t)^2 + o((\delta t)^2)$, and so on, where $o(x)$ implies that $o(x)/x \rightarrow 0$ as $x \rightarrow 0$. In addition, $T = \tau/\delta t$. With slight abuse of notation, we write λ in place of $\lambda \delta t$, and refer to λ either as the demand or as the buying probability. The random demand in period t , denoted by $\xi(t; \lambda)$, is Bernoulli with probability $\lambda(t) = \lambda(p(t))$, and $\mathbb{P}(\xi(t) = 1) = \lambda(p(t))$ and $\mathbb{P}(\xi(t) = 0) = 1 - \lambda(p(t))$. Treating the demand rate λ as the control variables (prices are inferred via the inverse demand relationship), the discrete-time formulation of the dynamic pricing problem of Gallego

and van Ryzin [1] is

$$\max_{\{\lambda(t), t=1, \dots, T\}} \left\{ \mathbb{E} \left[\sum_{t=1}^T p(\lambda(t)) \xi(t; \lambda) \right] : \sum_{t=1}^T \xi(t; \lambda) \leq C \text{ a.s. and } \lambda(t) \in \mathcal{L} \forall t \right\}. \quad (1)$$

Analysis of the Dynamic Program

Let x denote the number of remaining units of capacity at the beginning of period t , and $V(x, t)$ be the expected revenue-to-go starting at time t with x units of capacity left. Then, the Bellman equation associated with Equation (1) is

$$V(x, t) = \max_{\lambda \in \mathcal{L}} \left\{ \lambda [p(\lambda) + V(x-1, t+1)] + (1-\lambda) V(x, t+1) \right\}, \quad (2)$$

with the boundary conditions

$$V(x, T+1) = 0 \forall x \quad \text{and} \quad V(0, t) = 0 \forall t. \quad (3)$$

Letting $\Delta V(x, t) = V(x, t+1) - V(x-1, t+1)$ denote the marginal value of one unit of capacity as a function of the state (x, t) , Equation (2) can be rewritten as

$$V(x, t) = \max_{\lambda \in \mathcal{L}} \{ R(\lambda) - \lambda \Delta V(x, t) \} + V(x, t+1). \quad (4)$$

The optimal control $\lambda^*(x, t)$ is computed as follows:

$$\lambda^*(x, t) = \arg \max_{\lambda \in \mathcal{L}} \{ R(\lambda) - \lambda \Delta V(x, t) \},$$

where $R(\cdot)$ is a concave increasing revenue function. Using the properties of $R(\cdot)$, one can show that $\lambda^*(x, t)$ is decreasing in $\Delta V(x, t)$, which, using a backwards induction argument in t , gives that $\Delta V(x, t)$ is decreasing in x and t . These monotonicity results are the key structural properties that one can extract from the dynamic program that is summarized below; a proof can be found in [10] (Proposition 5.2, Chapter 4).

Proposition 1 [10, Proposition 5.2 Chapter 4]. *For the problem defined in Equation (1):*

1. $\lambda^*(x, t)$ is decreasing in the marginal value of capacity $\Delta V(x, t)$, and
2. $\Delta V(x, t)$ is decreasing in x and t .

The dynamic program in Equations (2) and (3) admits a closed-form solution for the special case of the exponential demand model, and is fairly easy to solve numerically in other cases. The optimal policy takes the form of a two-dimensional table that specifies a price for each (remaining inventory, time) pair. The optimal price path continuously decreases price between sales, and jumps up at every sales epoch.

The Fluid Model

The “fluid” model has deterministic and continuous dynamics, and in broad terms is obtained by replacing the Poisson demand process with nonhomogeneous rate $\lambda(t)$, where demand requests arrive stochastically over time and require discrete units of capacity, by a deterministic and continuous process where demand for the product arrives continuously at the deterministic rate $\lambda(t)$. The resulting inventory dynamics (in discrete-time) are given by

$$X(t+1) = X(t) - \lambda(t), \quad x(0) = C, \quad X(T) \geq 0.$$

This deterministic analog is a simplification of the discrete and stochastic model we started with in the previous sections. It can be rigorously justified as a limit under a law-large-numbers type of scaling as the potential demand and the capacity grow proportionally large, and, as such, one would expect to provide more useful analysis and policy recommendations in settings where the firm has many units to sell and operates in a market with high demand. For example, one would expect that the discrete and stochastic nature of the pricing problem to be relevant when selling four newly constructed single family homes over the course of 24 weeks, but it may be less relevant when selling 4000

pairs of skis over a similar time duration from, say, October to March.

The fluid model formulation of the dynamic pricing problem is one where the firm selects a demand rate $\lambda(t)$ (or a price) at each time t to

$$\max_{\{\lambda(t), t=1, \dots, T\}} \left\{ \sum_{t=1}^T R(\lambda(t)) dt : \sum_{t=1}^T \lambda(t) dt \leq C \text{ and } \lambda(t) \in \mathcal{L} \forall t \right\}. \quad (5)$$

Optimal Policy for Fluid Pricing Problem.

An important result derived in Gallego and van Ryzin [1] is that a constant price (and thus a constant demand rate) is optimal for Equation (5). Specifically, let $\hat{\lambda} = \operatorname{argmax}\{R(\lambda) : \lambda \in \mathcal{L}\}$ and $\hat{p} = p(\hat{\lambda})$ be the demand rate and price that maximize the revenue rate disregarding any capacity considerations, respectively. Also, let $\lambda^0 = C/T$ be the *run-out* rate that depletes capacity at time T , and $p^0 = p(\lambda^0)$.

Proposition 2 [10, Section 5.2.1.2]. *The optimal solution for Equation (5) denoted by $\bar{\lambda}$ and $\bar{p}$ is given by*

$$\bar{\lambda}(x, t) = \min(\lambda^0, \hat{\lambda}) \quad \text{and} \quad \bar{p} = \max(p^0, \hat{p}) \quad \text{for } t = 1, \dots, T. \quad (6)$$

Intuitively, the firm uses the revenue maximizing price $\hat{p}$ unless this would deplete the capacity too soon, in which case it increases its unit price to p^0 and sell its capacity by time T , while accruing higher total revenues. The proof is simple and exploits the structure of Equation (5) that seeks to maximize a concave function over the capacity constraint; the first-order optimality conditions set the marginal revenue rates at each time t to be equal, which, in turn, is achieved via a static pricing policy.

The static nature of the optimal policy for the fluid control problem is simple and intuitive, but also lacks the capability of corrective action against stochastic fluctuations. This does not arise in the fluid formulation, where the capacity is drained along the optimal deterministic trajectory, but it

is relevant for the stochastic problems of original interest. Alternatively, $\bar{\lambda}$ can also be expressed in the feedback form. Note that the deterministic trajectory of the fluid model is such that $x/(T-t) = C/T$ for all t if $\hat{\lambda} \geq C/T$, and $x/(T-t) = (C - \hat{\lambda}t)/(T-t) \geq C/T$ if $\hat{\lambda} < C/T$. Given this observation, $\bar{\lambda}$ can be expressed as

$$\bar{\lambda}(x, t) = \min\left(\hat{\lambda}, \frac{x}{T-t}\right). \quad (7)$$

An Upper Bound on Achievable Performance. Apart from good policy recommendations, the fluid model offers a useful upper bound on the achievable expected revenue in the underlying stochastic and discrete problem in Equation (1). Specifically, Gallego and van Ryzin [1] showed the following:

Proposition 3 [10, Section 5.2.2.3]. $V(C, 1) \leq (\bar{\lambda}\bar{p})T$.

This result establishes a tractable limit for the best achievable performance that is useful in establishing suboptimality gaps for heuristics that one may wish to use, and to prove asymptotic optimality results of candidate policies that achieve that bound in an appropriate asymptotic sense. As an illustration, we conclude this section with a brief sketch of the setup and nature of such an asymptotic optimality result.

Using k as an index, we consider a sequence of problems with demand model $\lambda^k(\cdot) = k\lambda(\cdot)$ and capacity $C^k = kC$, and let k increase to infinity; hereafter, a superscript k will denote quantities that scale with k . Let N denote a unit rate Poisson process, and recall that the functional strong-law-of-large numbers for the Poisson process asserts that as $k \rightarrow \infty$ and for all $t \geq 0$,

$$\frac{N(kt)}{k} \rightarrow t \quad \text{a.s.} \quad (8)$$

The remaining capacity at time t is denoted by $X^k(t)$ and is given by

$$X^k(t) = C^k - N(A^k(t)), \quad (9)$$

where $N(A^k(t))$ is the cumulative number of sales up to period t and $A^k(t) = \sum_{s=1}^t \lambda^k(s) ds$. Let R^k denote the total cumulative revenue extracted under policy (Eq. 6). The firm uses the constant price vector $\bar{p}$, which induces the demand rates $\lambda^k(\bar{p}) = k\bar{\lambda}$. Under this policy, $A^k(t) = \bar{\lambda}kt$, and the capacity dynamics are $X^k(t) = (C^k - N(A^k(t)))^+$. (The subscript is used to identify the policy.) As $k \rightarrow \infty$

$$\frac{N(A^k(t))}{k} \rightarrow \bar{\lambda}t \quad \text{a.s., } \forall t = 1, \dots, T, \quad (10)$$

from which it follows that as $k \rightarrow \infty$ and for all $t \in [0, T]$ $X^k(t)/k \rightarrow C - \bar{\lambda}t$ a.s. uniformly on compact sets as $k \rightarrow \infty$. Let $\tau^k := \inf\{s \geq 0 : N(\bar{\lambda}ks) \geq C^k\}$ be the random time where the capacity is depleted. Then, $R^k := \bar{p}N(k\bar{\lambda} \min(T, \tau^k)) - \delta$, where δ is a random variable that corrects revenues for the case where $\tau^k < T$; clearly $\delta < \bar{p}$. (We do not delve into an accurate description of δ , since it is asymptotically negligible.) Using Equation (10) and arguing by contradiction, one can easily conclude that $(T - \tau^k)^+ \rightarrow 0$ a.s., as $k \rightarrow \infty$, and get the following:

Proposition 4 [6, Proposition 4]. *Suppose that demand and capacity are scaled according to $\lambda^k(\cdot) = k\lambda(\cdot)$ and $C^k = kC$, and consider the static pricing policy $p^k(x, t) = \bar{p}$ for all x, t and all k . Then, as $k \rightarrow \infty$, $X^k(t)/k \rightarrow C - \bar{p}t$ a.s., uniformly in t , and $\frac{1}{k}R^k \rightarrow \bar{p}\bar{\lambda}T$ a.s.*

Given Proposition 3, we have that $R^k \leq k\bar{p}\bar{\lambda}T$. Applying the bounded convergence theorem gives that $\mathbb{E}R^k/k \rightarrow \sum_{i=1}^n \bar{p}_i \bar{\lambda}_i T$, and establishes the asymptotic optimality of the static pricing heuristic in Equation (6). A similar proof can establish the asymptotic optimality of the feedback rule given in Equation (7), which when applied to the stochastic system of interest yields a dynamic pricing policy; see Maglaras and Meissner [6].

MULTIPLE PRODUCTS, SINGLE RESOURCE

This section studies the multiproduct dynamic pricing and capacity control problems studied by Gallego and van Ryzin [2] and Lee and Hersh [3], respectively, among others.

Problem Formulations

Dynamic Pricing Problem. The basic elements of the problem are similar to those in the section titled “Single Product Dynamic Pricing Problem” with the key difference that the firm is now selling multiple products or services, indexed by $i = 1, \dots, n$ that consume the capacity C . Each product i request requires one unit of capacity. Let $p(t) = [p_1(t), \dots, p_n(t)]$ denote the vector of prices at time t . The demand process is assumed to be an n -dimensional nonhomogeneous Poisson process with rate vector λ determined through a demand function $\lambda(p(t))$, where $\lambda : \mathcal{P} \rightarrow \mathcal{L}$, $\mathcal{P} \subseteq \mathbb{R}^n$ is the set of feasible price vectors, and $\mathcal{L} = \{x \geq 0 : x = \lambda(p), p \in \mathcal{P}\} \subseteq \mathbb{R}_+^n$ is the set of achievable demand rate vectors. We assume that $\mathcal{L}$ is a convex set. As in the section titled “Single Product Dynamic Pricing Problem,” the demand function $\lambda(\cdot)$ is assumed to be stationary. The definition of regular demand functions is as follows:

Definition 2. A demand function is said to be regular if it is a continuously differentiable, bounded function, and, (a) for each product i , $\lambda_i(p)$ is strictly decreasing in p_i , (b) $\lim_{p_i \rightarrow \infty} \lambda_i(p) = 0$ (i.e., consumers have bounded wealth), and (c) the revenue rate $p' \lambda(p) = \sum_{i=1}^n p_i \lambda_i(p)$ is bounded for all $p \in \mathcal{P}$ and has a finite maximizer $\bar{p}$.

We assume that there exists a continuous inverse demand function $p(\lambda), p : \mathcal{L} \rightarrow \mathcal{P}$, that maps an achievable vector of demand rates λ into a corresponding vector of prices $p(\lambda)$. This allows us to view the demand rate vector as the firm’s control, and infer the appropriate prices using the inverse demand function. The expected revenue rate can be expressed as a function of the vector of demand rates λ as $R(\lambda) := \lambda' p(\lambda)$, and is assumed to be continuous, bounded, and strictly concave.

Example 3 [Linear Demand Model]. The demand for product i is given by

$$\lambda_i(p) = \Lambda_i - b_{ii}p_i - \sum_{j \neq i} b_{ij}p_j$$

or (in vector form) $\lambda(p) = \Lambda - Bp$,

where Λ_i is the market potential for product i and b_{ii}, b_{ij} are the price and cross-price sensitivity parameters. The inverse demand and revenue functions are $p(\lambda) = B^{-1}(\Lambda - \lambda)$ and $R(\lambda) = \lambda' B^{-1}(\Lambda - \lambda)$, respectively. Assumption 1 requires that $b_{ii} > 0$ for all i . To ensure that the inverse demand function is well defined and the revenue function is concave, we require that either $b_{ii} > \sum_{j \neq i} |b_{ji}|$ or $b_{ii} > \sum_{j \neq i} |b_{ij}|$ for all i ; both conditions guarantee that B is invertible and that its eigenvalues have positive real parts [17, Theorem 6.1.10].

Example 4 [Multinomial Logit (MNL) Model]. Potential customers arrive with rate Λ and have utilities for each product i given by $v_i - p_i + \xi_i$, where v_i is the deterministic portion that is common to all customers, p_i is the price, and ξ_i is the random term (that differentiates customers) that is drawn from a Gumbel distribution with mean 0 and parameter 1 (the latter is assumed w.l.o.g.), and is IID across products and customers. The no-purchase option has utility $u_0 + \xi_0$, ξ_0 is IID with the ξ_i ’s and $u_0 = 0$. The demand rates for product i is given by

$$\lambda_i(p) = \Lambda \frac{e^{v_i - p_i}}{1 + \sum_j e^{v_j - p_j}}.$$

Again adopting a discrete-time formulation, the random demand vector in each period t , denoted by $\xi(t; \lambda)$, is Bernoulli with probabilities $\lambda(t) = \lambda(p(t))$, and $\mathbb{P}(\xi_i(t) = 1) = \lambda_i(p(t))$ and $\mathbb{P}(\xi_i(t) = 0) = 1 - \lambda_i(p(t))$ for all i . Treating the demand rates λ_i as the control variables (prices are inferred via the inverse demand relationship), the discrete-time formulation of the dynamic pricing problem of Gallego and van Ryzin [2] is

$$\begin{aligned} \max_{\{\lambda(t), t=1, \dots, T\}} & \left\{ \mathbb{E} \left[\sum_{t=1}^T p(\lambda(t))' \xi(t; \lambda) \right] \right. \\ & \left. : \sum_{t=1}^T e' \xi(t; \lambda) \leq C \text{ a.s. and } \lambda(t) \in \mathcal{L} \forall t \right\}. \end{aligned} \quad (11)$$

Capacity Control Problem. The next variant we consider is the one studied by Lee and Hersh [3], where the price vector p and the demand rate vector $\lambda = \lambda(p)$ are fixed, and the firm optimizes over capacity allocation decisions. For this problem and without any loss of generality, we assume that products are labeled such that $p_1 \geq p_2 \geq \dots \geq p_n$. The firm has discretion as to which product requests to accept at any given time. This is modeled through the control $u_i(t)$ that is equal to the probability of accepting a product i request at time t . It is customary to assume that the firm is “opening” or “closing” products, thus considering controls $u_i(\cdot)$ that are 0 or 1, but this need not be imposed as a restriction. The dynamic capacity control problem is the following:

$$\max_{\{u(t), t=1, \dots, T\}} \left\{ \mathbb{E} \left[\sum_{t=1}^T p' \xi(t; u\lambda) \right] : \sum_{t=1}^T e' \xi(t; u\lambda) \leq C \text{ a.s. } 1 \text{ and } u_i(t) \in [0, 1] \forall t \right\}, \quad (12)$$

where $u\lambda$ above denotes the vector with coordinates $u_i \lambda_i$.

The remainder of this section describes how to reduce Equations (11) and (12) into dynamic optimization problems where the control is the (one-dimensional) aggregate capacity consumption rate. The reduced problems can be studied through a unified analysis.

A Common Formulation in Terms of the Aggregate Capacity Consumption

Dynamic Pricing Problem. Let x denote the number of remaining units of capacity at the beginning of period t , and $V(x, t)$ be the expected revenue-to-go starting at time t with x units of capacity left. Then, the Bellman equation associated with Equation (1) is

$$V(x, t) = \max_{\lambda \in \mathcal{L}} \left\{ \sum_{i=1}^n \lambda_i [p_i(\lambda) + V(x-1, t+1)] + (1 - e'\lambda) V(x, t+1) \right\}, \quad (13)$$

with the boundary conditions

$$V(x, T+1) = 0 \quad \forall x \quad \text{and} \quad V(0, t) = 0 \quad \forall t. \quad (14)$$

Letting $\Delta V(x, t) = V(x, t+1) - V(x-1, t+1)$ denote the marginal value of one unit of capacity as a function of the state (x, t) , Equation (13) can be rewritten as

$$\begin{aligned} V(x, t) &= \max_{\lambda \in \mathcal{L}} \left\{ R(\lambda) - \sum_{i=1}^n \lambda_i \Delta V(x, t) \right\} + V(x, t+1) \\ &= \max_{\rho \in \mathcal{R}} \{ R^r(\rho) - \rho \Delta V(x, t) \} + V(x, t+1), \end{aligned} \quad (15)$$

where $\rho := \sum_{i=1}^n \lambda_i$ is the *aggregate* rate of capacity consumption, $\mathcal{R} := \{ \rho : \sum_{i=1}^n \lambda_i = \rho, \lambda \in \mathcal{L} \}$ is the set of achievable capacity consumption rates, and

$$R^r(\rho) := \max_{\lambda} \left\{ R(\lambda) : \sum_{i=1}^n \lambda_i = \rho, \lambda \in \mathcal{L} \right\} \quad (16)$$

is the maximum achievable revenue rate subject to the constraint that all products jointly consume capacity at a rate ρ . Note that Equation (16) is a concave maximization problem over a convex set, and its solution is readily computable, often in closed form. The aggregate revenue function $R^r(\cdot)$ is concave and satisfies the conditions of Assumption 1. The optimal vector of demand rates, denoted by $\lambda^r(\rho)$, is unique and continuous in ρ .

Example 5 [Linear Demand Model]. For $\lambda(p) = \Lambda - Bp$, the associated aggregate revenue function $R(\rho)$ defined through Equation (16) can be expressed as

$$R^r(\rho) = -\alpha_i \rho^2 + \beta_i \rho + \gamma_i \quad \text{for } \rho \in [r_{i-1}, r_i),$$

for $0 = r_0 \leq r_1 \leq r_2 \leq \dots \leq r_I$, and constants $(\alpha_i, \beta_i, \gamma_i)$ and r_i that depend on the model parameters Λ, B, μ , and are such

that $R^r(\rho)$ is continuous, almost everywhere differentiable, and increasing for all $\rho \leq \hat{\rho} := \arg \max_{\rho} R^r(\rho)$. The value of r_{i-1} is that of the smallest capacity consumption rate above which it is optimal to start offering the i most profitable products. The derivation of the constants $(\alpha, \beta, \gamma, r)$ can be found in the Appendix of [18].

Example 6. For the MNL model, straightforward manipulations show that

$$\begin{aligned} R(\rho) &= \rho \ln \left(\sum_j e^{v_j} \right) - \rho \ln(\rho/(\Lambda - \rho)), \\ \lambda_i(\rho) &= \rho \frac{e^{v_i}}{\sum_j e^{v_j}} \quad \text{and} \\ p_i(\rho) &= \ln \left(\sum_j e^{v_j} \right) - \ln(\rho/(\Lambda - \rho)). \end{aligned}$$

Proposition 5. *The dynamic pricing problem (Eq. 11) can be reduced to the dynamic programs (15) and (14) expressed in terms of the aggregate consumption rate. In particular, if $\rho^*(x, t)$ denotes the associated optimal control and $\lambda^*(x, t)$ and $p^*(x, t)$ denote the optimal demand rate and price vectors associated with Equation (11), then, $\lambda^*(x, t) = \lambda^r(\rho^*(x, t))$ and $p^*(x, t) = p(\lambda^r(\rho^*(x, t)))$.*

The Capacity Control Problem. Similarly, the Bellman equation associated with Equation (12) is

$$\begin{aligned} V(x, t) &= \max_{u_i \in [0, 1]} \left\{ \sum_{i=1}^n \lambda_i u_i [p_i + V(x-1, t+1)] \right. \\ &\quad \left. + (1 - u'\lambda) V(x, t+1) \right\}, \end{aligned} \quad (17)$$

with the boundary condition (3), which using the marginal value of capacity ΔV

becomes

$$\begin{aligned} V(x, t) &= \max_{u_i \in [0, 1]} \left\{ \sum_{i=1}^n \lambda_i u_i p_i - u'\lambda \Delta V(x, t) \right\} \\ &\quad + V(x, t+1) \quad (18) \\ &= \max_{0 \leq \rho \leq \sum_{i=1}^n \lambda_i} \{ R^a(\rho) - \rho \Delta V(x, t) \} \\ &\quad + V(x, t+1) \quad (19) \end{aligned}$$

where $\rho = u'\lambda$ and $R^a(\rho) = \max_u \{ \sum_{i=1}^n u_i \lambda_i p_i : u'\lambda = \rho, u_i \in [0, 1] \}$ is the maximum revenue rate when the capacity is consumed at a rate equal to ρ , and $u^a(\rho)$ is the corresponding control.

Proposition 6. *The capacity control problem (12) can be reduced to the dynamic programs (19) and (14) expressed in terms of the aggregate consumption rate ρ . In particular, if $\rho^*(x, t)$ denotes the optimal solution of Equations (19) and (14) and $u^*(x, t)$ denote the optimal policy for Equation (12), then $u^*(x, t) = u^a(\rho^*(x, t))$.*

A similar result was derived by Talluri and van Ryzin [13] for a capacity control problem for a model with customer choice.

A Unified Analysis of the Pricing and Capacity Control Problems

The preceding analysis illustrates that both problems can be reduced to “appropriate” single-product pricing problems, highlighting their common structure and enabling a unified treatment. As a starting observation, for both Equations (15) and (19), the optimal control $\rho^*(x, t)$ is computed from

$$\rho^*(x, t) = \arg \max_{\rho \in \mathcal{R}} \{ R(\rho) - \rho \Delta V(x, t) \},$$

where $R(\cdot)$ is a concave increasing revenue function. Using the properties of $R(\cdot)$, one gets that $\rho^*(x, t)$ is decreasing in $\Delta V(x, t)$, which using a backwards induction argument in t gives that $\Delta V(x, t)$ is decreasing in x and t . These standard results for single-product dynamic pricing problems are summarized below; a proof can be found in [10] (Proposition 5.2 Chapter 4).

Proposition 7 [10, Proposition 5.2 Chapter 4]. *For both problems defined in Equations (1) and (12), we have that:*

1. $\rho^*(x, t)$ is decreasing in the marginal value of capacity $\Delta V(x, t)$, and
2. $\Delta V(x, t)$ is decreasing in x and t .

Structural results for the pricing and capacity allocation policies follow from the properties of R^r, λ^r , and R^a, u^a , respectively. For example, consider the pricing problem for the case where the products are nonsubstitutes, that is, the demand for product i is only a function of the price for that product p_i . In that case, the Lagrangian associated with Equation (16) is $L(\lambda, x, y) = R(\lambda) + x(\rho - \sum_{i=1}^n \lambda_i) - y'\lambda$, with first-order conditions given by $\partial R(\lambda)/\partial \lambda_i = x + y_i$, for some $x \geq 0$ and $y_i \leq 0$ with $y_i = 0$ if $\lambda_i > 0$. It is easy to show that x is decreasing in ρ (i.e., the shadow price for the capacity consumption constraint decreases as ρ increases), and that $\lambda_i^r(\rho)$ is decreasing in x .

Corollary 1. *Consider the problem specified in Equations (2) and (3) and further assume that the products are nonsubstitutes, that is, $\lambda_i(p) = \lambda_i(p_i)$ for all i . Then, $\lambda_i^*(x, t)$ is nondecreasing in $\rho^*(x, t)$ (and nonincreasing in $\Delta V(x, t)$).*

A similar result can be obtained when products are substitutable provided that the demand model satisfies certain conditions analogous to those of the sensitivity matrix B of the linear model described earlier in this section.

For the capacity control problem, it is easy to recover some well-known structural properties of the optimal policy; see Lee and Hersh [3]. Our derivation based on the capacity consumption rate offers new intuition as to why they hold. Specifically, $R^a(\cdot)$ is a knapsack solution for which

$$R^a(\rho) = \min_i c_i + p_i \rho \quad \text{and} \quad u_k^a(\rho) = \min \left(\frac{\left(\rho - \sum_{i < k} \lambda_i \right)^+}{\lambda_k}, 1 \right), \quad (20)$$

where $c_1 = 0$ and $c_i = \sum_{k < i} \lambda_k(p_k - p_i)$, and for any $x \in \mathbb{R}$, $x^+ := \max(x, 0)$, and the optimal control $\rho^*(x, t)$ reduces to the solution to $\max\{\min_i c_i + (p_i - \Delta V(x, t))\rho : 0 \leq \rho \leq \sum_{i=1}^n \lambda_i\}$. Let $i^*(x, t) = \max\{i \geq 1 : p_i \geq \Delta V(x, t)\}$. Then, by inspecting the form of the piecewise linear objective function involved in the calculation of $\rho^*(x, t)$, we get that $\rho^*(x, t) = \sum_{i \leq i^*(x, t)} \lambda_i$. That is, the solution is “bang-bang” in the sense that the form of the optimal control is such that $u_i^*(x, t)$ is 0 if $i > i^*(x, t)$ and 1 if $i \leq i^*(x, t)$. In addition, from Proposition 7 part 1, we see that $i^*(x, t)$ is decreasing in the marginal value of capacity $\Delta V(x, t)$. Therefore,

Corollary 2. *For the capacity control problem (12) or equivalently, Equations (19) and (3), the optimal allocation policy is nested in that $u_i^*(x, t) = 1$ if $i \leq i^*(x, t)$, and $u_i^*(x, t) = 0$ otherwise, and $i^*(x, t)$ is decreasing in the marginal value of capacity $\Delta V(x, t)$.*

Efficient Frontier. The subproblem of computing the optimal revenue subject to a constraint on the aggregate capacity consumption rate specified in Equations (16) and (20) defines an efficient frontier $(\rho, R^r(\rho))$ and $(\rho, R^a(\rho))$ for the dynamic pricing and capacity allocation problems, respectively. As in the context of portfolio optimization, the efficient frontier provides a systematic framework for comparing different policies and highlights the structure of the respective optimal controls. It may also lead to computational improvements if this subproblem can be solved efficiently, preferably in closed form. This is possible in some common demand models such as the linear and the multinomial logit. The preceding discussion is from [6]. The idea of an efficient frontier has also appeared and is discussed in more detail in Feng and Xiao [14,15] and Talluri and van Ryzin [13]. Finally, we note that the structure of the dynamic programs studied in this section has been observed in other papers, such as Lin *et al.* [16] and their study of single-resource capacity control problems where each arrival may request multiple units of capacity, and Vulcano *et al.*

[19] and their analysis of optimal dynamic auctions. The latter involves an analysis of a discrete-time, batch demand analog to the dynamic program studied here.

Solution to the Deterministic Multiproduct Pricing Problem

As before, the “fluid” model has deterministic and continuous dynamics, and is obtained by replacing the discrete stochastic demand process by its rate, which now evolves as a continuous process. The realized instantaneous demand for product i at time t is deterministic and is given by $\lambda_i(t)$. We allow product i requests to consume capacity at a rate of $a_i > 0$ units per unit of demand, and denote the vector $[a_1, \dots, a_n]$ by a . This is a generalization of the model considered thus far that assumed uniform capacity requirements (all equal to 1). With a general capacity requirement vector a , the capacity consumption rate is defined by $\rho = a'\lambda$, and the definitions of R^r and λ^r can be appropriately adjusted to reflect that change. The system dynamics are given by $X(t+1) = X(t) - \sum_{i=1}^n a_i \lambda_i(t)$, $X(0) = C$, together with the boundary condition that $X(T) \geq 0$. The firm selects a demand rate $\lambda_i(t)$ (or a price) at each time t . The fluid formulation of the multiproduct pricing problem is the following:

$$\max_{\{\lambda(t), t=1, \dots, T\}} \left\{ \sum_{t=1}^T R(\lambda(t)) dt : \sum_{t=1}^T a' \lambda(t) dt \leq C \right. \\ \left. \text{and } \lambda(t) \in \mathcal{L} \forall t \right\}. \quad (21)$$

Gallego and van Ryzin [2, Section 4.5] partially extended their single-product results to multiple products, but without providing such a succinct solution as the one presented in the previous section. An alternative approach that exploits the action space reduction described in the section titled “A Common Formulation in Terms of the Aggregate Capacity Consumption” was described in Ref. 6 and is reviewed below. Specifically, recalling the definitions of the aggregate revenue function $R^r(\rho)$ and optimal demand rate vector $\lambda^r(\rho)$ in Equation (16) adjusted for

the fact that $\rho = a'\lambda$, Equation (21) can be rewritten as

$$\max_{\{\rho(t), t=1, \dots, T\}} \left\{ \sum_{t=1}^T R^r(\rho(t)) dt : \sum_{t=1}^T \rho(t) dt \leq C, \rho(t) \in \mathcal{R} \forall t \right\}. \quad (22)$$

Note that Equation (22) is identical to a single-product problem with revenue function R^r , and thus is solvable using the approach described above. Let $\rho^0 := C/T$ and $\hat{\rho} = \arg \max_{\rho} R^r(\rho)$. Then, the optimal solution to Equation (22) is to consume capacity at a constant rate $\bar{\rho}$ given by

$$\bar{\rho}(t) := \min(\hat{\rho}, \rho^0) \quad \forall t, \quad (23)$$

the corresponding vector of demand rates is $\lambda^r(\bar{\rho})$, while the price vector is $p(\lambda^r(\bar{\rho}))$. A direct verification that this solution satisfies the optimality conditions for Equation (21) establishes the following:

Proposition 8. *Let $\bar{\lambda}(\cdot)$ and $\bar{p}(\cdot)$ denote the optimal vectors of demand rates and prices for Equation (21). Then, $\bar{\lambda}, \bar{p}$ are constant over time and are given by $\bar{\lambda}(t) = \lambda^r(\bar{\rho})$ and $\bar{p}(t) = p(\lambda^r(\bar{\rho}))$.*

Asymptotically Optimal Heuristics Extracted From the Deterministic Model

Finally, we discuss three heuristics for the revenue management problems studied in this section. For each of these policies, one could show that they achieve the optimal asymptotic performance in the spirit of the results of the section titled “The Fluid Model”; details of these arguments can be found in [6].

1. *The Static Pricing Heuristic of Gallego and van Ryzin [2].* This policy implements the static prices $\bar{p}$ specified in Proposition 8. (This is the “make-to-order” heuristic in [2].)
2. *A List Price Capacity Control (LPCC) Heuristic.* One way to implement Equation (25) is by introducing capacity control capability on top of the static

prices given in (1). Specifically, our second heuristic is defined as follows:

- a. price according to $\bar{p}$ and label products such that $\bar{p}_1/a_1 \geq \bar{p}_2/a_2 \geq \dots \geq \bar{p}_n/a_n$, and
- b. compute $\bar{\rho}(x, t)$ and use the capacity controls $u_1(x, t) = 1$ if $x > 0$, $u_1(0, t) = 0$, and

$$u_i(x, t) = \begin{cases} 1 & \text{if } \bar{\rho}(x, t) - \sum_{j < i} a_j \bar{\lambda}_j \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i \geq 2. \quad (24)$$

Note that this policy can only reduce the aggregate capacity consumption rate from its nominal value of $\sum_{i=1}^n a_i \bar{\lambda}_i$, but can never increase it. A product is made available only if the fluid solution starting from that state would choose to sell this product in all future time periods, and “closes” the product if the fluid solution would dictate only partial acceptance of the associated demand.

This policy was described in [6]. It is a refinement of the static pricing policy in (1) and the make-to-order heuristic of Gallego and van Ryzin [2]. Other examples of joint pricing and capacity controls can be found in the recent papers by Vulcano *et al.* [19], by Lin *et al.* [16] and Feng and Xiao [15].

3. *A Dynamic Pricing Heuristic.* The solution of the fluid pricing problem of the section titled “Solution to the Deterministic Multiproduct Pricing Problem” can be described in feedback form as

$$\bar{\rho}(x, t) = \min \left(\hat{\rho}, \frac{x}{T-t} \right), \quad (25)$$

where x is the remaining capacity at time t . The third heuristic that we present translates the aggregate control $\bar{\rho}(x, t)$ into product-level rates

(and prices) through

$$\begin{aligned} \lambda(x, t) &= \lambda^r(\bar{\rho}(x, t)) \quad \text{and} \\ p(x, t) &= p(\lambda(x, t)), \end{aligned} \quad (26)$$

where the mapping $\lambda^r(\cdot)$ was the maximizer in Equation (16) and was continuous in ρ . This corresponds to the idea of “resolving” the fluid problem as we step through time, which has also been discussed in the section titled “Single-Product Dynamic Pricing Problem.” This is widely applied in practice, where, however, the resolving occurs at discrete points in time, for example, daily or weekly depending on the application setting. These resolving policies seem to have been first analyzed in Ref. 6. Specifically, the results in [6] show that the idea of resolving applied in the context of the dynamic pricing or the LPCC heuristics is fluid-scale asymptotically optimal, as is the static policy (1); the single-product version of this feedback pricing policy has also been studied by Reiman [20]. These results show that the suboptimal behavior demonstrated by a resolving policy in the negative example studied in Cooper [12] does not persist in systems with large capacity and large demand. Intuitively, our preceding discussion illustrates that “resolving” is nothing but implementing the fluid policy in the feedback form. Numerical experiments documented in many papers [6] demonstrate that such feedback heuristics tend to outperform policies that are “static.”

DYNAMIC PRICING NETWORK REVENUE MANAGEMENT PROBLEMS

Suppose that the firm is operating a network of resources, indexed by $j = 1, \dots, m$, and that each product i request consumes A_{ij} units of resource j capacity. Let $A := [A_{ij}]$ denote the associated capacity consumption matrix, and assume that the initial capacity for each resource j is C_j . Then, the fluid

model formulation of the network dynamic pricing problem is

$$\max_{\{\lambda(t), t=1, \dots, T\}} \left\{ \sum_{t=1}^T R(\lambda(t)) dt : \sum_{t=1}^T A\lambda(t) dt \leq C \right. \\ \left. \text{and } \lambda(t) \in \mathcal{L} \forall t \right\}. \quad (27)$$

As before, this problem can be expressed in terms of ρ , which is defined by $\rho := A\lambda$. Specifically, let

$$R^r(\rho) := \max_{\lambda} \{R(\lambda) : A\lambda = \rho, \lambda \in \mathcal{L}\}, \quad (28)$$

be the maximum achievable revenue rate when resource capacity is consumed at a rate ρ , and $\lambda^r(\rho)$ denote the corresponding vector of optimal demand rates. Then, Equation (27) can be reduced to

$$\max_{\{\rho(t), t=1, \dots, T\}} \left\{ \sum_{t=1}^T R^r(\rho(t)) dt : \sum_{t=1}^T \rho(t) dt \leq C \right. \\ \left. \text{and } \rho(t) \in \mathcal{R} \forall t \right\}. \quad (29)$$

Let $\bar{\rho}$ denote the solution to Equation (29). Then, $\lambda^r(\bar{\rho})$ is the vector of optimal demand rates for Equation (27). This reduction could prove computationally beneficial, since as is often the case the number of products (e.g., the number of fare-class and origin-destination pairs) tends to be greater than the number of resources (e.g., number of flights in a hub-and-spoke network). We refer the reader to Gallego and van Ryzin [2] and Kleywegt [21] for fluid formulations to multiproduct network revenue management problems.

EXTENSIONS

There are several extensions that we chose not to address in this article. Some may be the result of practical considerations. For example, one may want to consider pricing policies where the feasible price grid is discrete, say in \$10 increments. Feng and Xiao [15] have studied this problem. Another

extension would incorporate costs incurred when the price is changed; this problem was studied in Celik *et al.* [22].

Two other important research directions of practical and theoretical interest are the following. The first studies revenue maximization problems for which the seller does not have accurate information about the underlying demand model. In such settings, the seller uses its pricing decisions to simultaneously learn the demand and optimize revenues. The second studies the effect of strategic consumer behavior on revenue maximization practices, such as their intentional waiting for sale periods to make their retail purchases.

REFERENCES

1. Gallego G, van Ryzin G. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Manage Sci* 1994;40(8): 999–1020.
2. Gallego G, van Ryzin G. A multiproduct dynamic pricing problem and its applications to network yield management. *Oper Res* 1997; 45(1):24–41.
3. Lee T, Hersh M. A model for dynamic airline seat inventory control with multiple seat bookings. *Transp Sci* 1993;27(3): 252–265.
4. Lautenbacher C, Stidham S. The underlying Markov Decision Process in the single-leg airline yield-management problem. *Transp Sci* 1999;33(2):136–146.
5. Zhao W, Zheng Y-S. Optimal dynamic pricing for perishable assets with nonhomogeneous demand. *Manage Sci* 2000;46(3): 375–388.
6. Maglaras C, Meissner J. Dynamic pricing strategies for multi-product revenue management problems. *Manuf Serv Oper Manage* 2006;8(2):136–148.
7. Elmaghraby W, Keskinocak P. Dynamic pricing: research overview, current practices and future directions. *Manage Sci* 2003;9(10):1287–1309.
8. Bitran G, Caldentey R. An overview of pricing models for revenue management. *Manuf Serv Oper Manage* 2003;5(3):203–229.
9. McGill J, van Ryzin G. Revenue management: research overview and prospects. *Transp Sci* 1999;33(2):233–256.

10. Talluri K, van Ryzin G. The theory and practice of revenue management. New York; Kluwer Academic Publishers; 2004.
11. Brumelle S, McGill J. Airline seat allocation with multiple nested fare classes. *Oper Res* 1993;41(1):127–137.
12. Cooper W. Asymptotic behavior of an allocation policy for revenue management. *Oper Res* 2002;50(4):720–727.
13. Talluri K, van Ryzin G. Revenue management under a general discrete choice model of consumer behavior. *Manage Sci* 2004;50(1):15–33.
14. Feng Y, Xiao B. A continuous-time yield management model with multiple prices and reversible price changes. *Manage Sci* 2000;46(5):644–657.
15. Feng Y, Xiao B. Integration of pricing and capacity allocation for perishable products. *Eur J Oper Res* 2006;168(1):17–34.
16. Lin GY, Lu Y, Yao DD. The stochastic knapsack revisited: structure, switch-over policies, and dynamic pricing. *Oper Res* 2008;56(4):945–957.
17. Horn RA, Johnson CR. Matrix analysis. Cambridge, UK: Cambridge University Press; 1994.
18. Maglaras C. Revenue management for a multi-class single-server queue via a fluid model analysis. *Oper Res* 2006;54(5):914–932.
19. Vulcano G, van Ryzin G, Maglaras C. Optimal dynamic auctions for revenue management. *Manage Sci* 2002;48(11):1388–1407.
20. Reiman MI. Asymptotically optimal dynamic pricing for a wholesale - retail telecommunications service provider. In: Presentation, 2nd Informs Conference on Revenue Management. New York; 2002 June.
21. Kleywegt AJ. An optimal control problem of dynamic pricing. Working paper. Georgia Institute of Technology; 2001.
22. Celik S, Muharremoglu A, Savin S. Airline seat allocation with multiple nested fare classes. *Oper Res* 2009;57(5):1206–1219.
23. Dai JG, Williams RJ. Existence and uniqueness of semimartingale reflecting brownian motions in convex polyhedrons. *Theory Probab Appl* (in Russian) 1994;40:3–53. To appear in the SIAM translation of the journal with the same name.

DYNAMIC PRICING UNDER CONSUMER REFERENCE-PRICE EFFECTS

HASAN ARSLAN

Sawyer Business School, Suffolk
University, Boston, Massachusetts

SOULAYMANE KACHANI

Department of Industrial Engineering and
Operations Research, Columbia
University, New York City, New York

INTRODUCTION

Consumers have a memory of past prices of products that are repeatedly purchased. As consumers are exposed to prices of products over a period of time, they develop certain price expectations. These price expectations are called *reference prices* [1–3]. They become benchmark prices against which consumers compare the current prices of products before they purchase. Empirical studies on consumer purchasing behavior in the marketing literature indicate that consumers utilize past prices to form reference prices [4,5]. The notion that consumers form reference prices to evaluate product prices is an application of the psychological theory of adaptation-level formation [6]. Reference prices, thus, are used as adaption levels for the observed prices [7].

Many articles have been written in the marketing literature on how reference prices influence consumers' purchasing decisions, how reference prices are formed, and how reference prices can be measured. Winer [8] reviews the theoretical foundations and modeling of the reference-price concept. Kalyanaram and Winer [9] survey the literature and propose empirical generalizations from reference-price research to support the reference-price concept. Briesch *et al.* [10] provide an empirical comparison of different reference-price models in the marketing literature. Mazumdar *et al.* [11] review the marketing literature on reference price in

both the behavioral and modeling areas and identify the unresolved issues.

The concept of reference price asserts that consumers make purchasing decisions for products with repeated purchases based on both observed and perceived prices. If the observed price of a product is below (above) the reference price, consumers perceive gains (losses) and purchasing becomes more (less) attractive [7,12–17]. Price promotions lower reference prices for future purchases. Thus, consumers may not perceive future price promotions as a good deal as the earlier ones. Consumers may perceive increase in prices when price promotions are over. Kalwani and Yim [5] report that the price consumers expected to pay for a product was significantly lower after they observed either more frequent or deeper promotions for the product on the prior purchase occasions. Also, an excessive use of promotions would confuse consumers. Consumers would perceive a return to the regular price as a price increase. Thus, there is an optimal frequency and amount of promotions that does not lower the reference price significantly nor confuses consumers about the regular price [9].

Firms are not able to directly modify how consumers form reference prices. However, they can influence reference prices through dynamic price patterns. Firms can develop more informed dynamic pricing strategies by taking consumer reference-price effects into account [18]. Our objective in this study is to explore the current-state of literature on how dynamic pricing models incorporate consumer reference-price effects in developing more informed dynamic pricing strategies for products that have repeated consumer interactions. We limit our review to studies where the reference-price concept is incorporated into dynamic pricing models. We use the following four-step framework for our review of the literature. First, we describe the effects of reference prices on consumer demand and purchasing decisions. Second, we present the formation process of reference prices. Third, we review the major dynamic

pricing models and some of their extensions in the presence of consumer reference-price effects. Finally, we identify research gaps and provide directions for further research on dynamic pricing area in the presence of consumer reference-price effects.

REFERENCE-PRICE EFFECTS ON CONSUMER DEMAND

The impact of reference price on demand is behaviorally explained at the individual level by the “prospect theory” [19]. The objective of prospect theory is to describe or predict consumer behavior under reference-price effects. Prospect theory posits the following: (i) consumers code the observed price of a product as loss or gain, depending on whether it is greater than or less than the reference price, respectively; (ii) consumers are asymmetric in their response to the observed price such that they are more sensitive to the prospect of a loss than to the prospect of a gain. Reference-price effects are considered to be asymmetric if consumers’ responsiveness to a positive difference between a reference price and a purchase price (i.e., a gain) is not the same as their responsiveness to a negative difference (i.e., a loss).

Symmetric Reference-Price Effects

There are many empirical studies in the marketing field that analyze reference-price effects on consumer demand using panel data. Reference-price effects on consumer demand is estimated by comparing a consumer demand model that contains no reference price with a model that incorporates reference-price term. Winer [7] observes through empirical testing that the discrepancies between reference prices and observed prices strongly affect consumers’ both product and quantity choices. Winer [7] is the first to propose a symmetric “sticker shock” model of reference price. According to his model, consumer utility from purchasing a good not only depends on the observed price but also on the difference between the reference price and the observed price. He assumes that a positive (negative) difference between the reference price and the observed

price increases (decreases) consumer utility from the product. Furthermore, he assumes that consumers’ response to a positive and a negative difference is symmetric. Winer [7] suggests that a firm’s optimal promotion policy over time should consider reference-price effects on consumer choice for frequently purchased products.

Asymmetric Reference-Price Effects

As indicated in Kalyanaram and Winer [9], further studies in the marketing field [12,13,15,20–22] provide empirical support for the following generalization for frequently purchased products: consumers react more strongly to price increases than to price decreases relative to the reference price—the notion that losses loom larger than gains.

Putler [15] is the first to derive the asymmetric reference-price demand model by formally integrating prospect theory into the traditional economic theory of consumer choice, thus providing a stronger theoretical base for the empirical results. His theoretical and empirical analysis indicate that reference prices have significant impact on consumers’ purchasing behavior and consumers react to losses (price increases relative to the reference price) more than gains (price decreases relative to the reference price)—consumers exhibit *loss aversion*. Hardie *et al.* [22] also study the asymmetric reference effects on consumers. They define that consumers perceive gains (losses) when price of a brand is less (greater) than that of the reference brand. They find that losses relative to the reference brand have more impact on consumer purchasing behavior than gains.

However, some studies in the marketing field find that consumers are not always more responsive to a loss than to a gain. Greenleaf [18], for example, finds that consumers may react less strongly to price decreases than price increases for frequently purchased products when consumers are exposed to either infrequent promotions or an irregular pattern of promotions. This result is in contrast to Kalyanaram and Little [12], Kalwani *et al.* [13], and Kalwani and Yim [5].

REFERENCE-PRICE FORMATION PROCESS

Modeling-based reference-price research uses consumer purchase history as the main determinant of reference price. Briesch *et al.* [10] and Mazumdar *et al.* [11] provide a summary of previously used reference-price models. The most commonly used reference-price model in the literature assumes that the reference price $r(t)$ in period t is an exponentially weighted average of prior prices $p(s)$, $s < t$ [12–15,18,23,24]

$$r(t) = \alpha r(t-1) + (1-\alpha)p(t-1). \quad (1)$$

Parameter $0 \leq \alpha \leq 1$ denotes the consumer memory effect of prior prices on reference price. Reference price becomes more dependent on past prices as the value of α increases. The lower the value of α is, the shorter the memory of consumers on prices is. Namely, if $\alpha = 0$, then the reference price equals to the past price. Empirical studies show that parameter α ranges approximately from 0.60 to 0.85. This indicates that prices experienced beyond two or three purchase occasions have negligible impact on reference price [25]. Consumer memory parameter is also an indicator of consumer shopping frequency. Thus, it can be used as a proxy for consumer loyalty [24]. Kopalle *et al.* [26] indicate that regression models can be used to estimate the value of parameter α . Equation (1) is also the most empirically validated reference-price formation process in the literature [7,18,24,26–28]. We illustrate the formation of reference prices in Equation

(1) on an example with consumer memory parameter $\alpha = 0.5$ in Fig. 1.

In continuous time, reference price at time t is modeled as a continuous weighted average of prior prices with an exponentially decaying weighting function [27–29], which can be represented by the following ordinary differential equation

$$\frac{dr(t)}{dt} = \alpha(p(t) - r(t)). \quad (2)$$

Consumers begin making purchasing decisions on a frequently purchased product with an initial reference price, $r(0)$. In all analytical models in the literature, the initial reference price is assumed to be an external parameter. There is no specific study focusing on identifying the drivers of initial reference price for newly introduced products. Thus, estimating initial reference price is an open research question to be considered. As consumers observe posted prices, they update their perceptions of the value of the product. Equivalently, the continuous reference-price formation in Equation (2) can be represented by the following formulation

$$r(t) = e^{-\alpha t} \left(r(0) + \alpha \int_0^t e^{\alpha s} p(s) ds \right). \quad (3)$$

Similarly, the discrete reference-price formation can also be written as a sum of weighted averages of prior prices.

Hardie *et al.* [22] provide an alternative reference-price formulation. They set a reference brand (rather than price) for consumers based on their purchase history. The reference price becomes the current price of

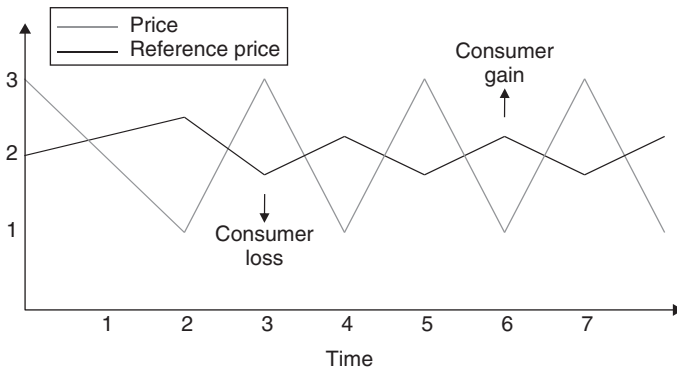


Figure 1. Exponential smoothing reference-price formation process, $\alpha = 0.5$.

the last brand bought—the reference brand. Consumers perceive gains (losses) from a brand when its price is below (above) the price of the reference brand. They also show that reference brand formation is empirically valid in a multibrand environment. Kopalle *et al.* [26] adapt the reference brand formulation in Hardie *et al.* [22] in determining the optimal dynamic pricing policy for a monopolist managing two brands as well as for a duopolist.

DYNAMIC PRICING MODELS UNDER DISCRETE AND CONTINUOUS TIME SETTINGS

The major managerial implications on dynamic pricing in the presence of reference-price effects are provided by Greenleaf [18], Kopalle *et al.* [26], Fibich *et al.* [28], and Popescu and Wu [24]. These studies investigate the impact of reference price on dynamic pricing strategies. The key managerial implication from these major studies is: firms can develop more informed dynamic pricing and price promotion strategies for frequently purchased products by considering how their past pricing patterns impact the current demand, and thus they should consider the impact of reference-price effects in developing their dynamic pricing and promotion policies. We next review the above major studies and some of their extensions that develop analytical models to explore optimal dynamic pricing strategies for firms in the presence of reference-price effects.

Base Model

Greenleaf [18] studies both analytically and empirically how reference-price effects influence promotion profits. He indicates that retailers cannot directly control how consumers form reference prices; however, they can indirectly control reference prices through changing the price pattern that consumers observe over time. The existing literature [7,13,15,17] before Greenleaf [18] focused on describing how consumers form reference prices and testing how reference prices affect consumer demand. Greenleaf [18] is the first to analyze the impact of

reference prices on developing promotion strategies.

Greenleaf [18] considers a single retailer who promotes a single brand over a finite number of time periods. The objective of the retailer is to estimate how the reference-price effects generated by the promotion impact the profits. Thus, the retailer maximizes the net present value of brand profits generated by the pricing decisions. His model does not analyze cross-brand profit effects and does not consider competition among different brands. He assumes the following reference-price formation, which is an extension of the discrete reference formation in Equation (1):

$$r(t) = \alpha r(t-1) + (1-\alpha)p(t-1) + \gamma(t); \quad (4)$$

where $r(t)$ is an error term with expected value of zero. He also assumes that the retailer faces the following demand for the brand in period t

$$d(r(t), p(t)) = q(p(t)) + \delta(r(t) - p(t)) + \varepsilon(t) \text{ if } r(t) > p(t), \quad (5)$$

$$d(r(t), p(t)) = q(p(t)) + r(r(t) - p(t)) + \varepsilon(t) \text{ if } r(t) > p(t), \quad (6)$$

where $q(p(t))$ is the demand in the absence of reference-price effects while the remaining part reflects the reference-price effect on demand and $\varepsilon(t)$ is an error term. He assumes that $q(p(t))$ is linearly decreasing in price $p(t)$. The demand model in Equations (5)–(6) assumes that in the absence of reference-price effects the product demand $q(p(t))$ is linearly decreasing in price; the reference-price effects are linear in $p(t) - r(t)$; reference-price effects are additive; consumers perceive gains when $p(t) < r(t)$; consumers perceive losses when $p(t) > r(t)$; reference-price effects are asymmetric (consumers are loss-seeking or loss-averse or risk-neutral when $\delta > \gamma$ or $\delta < \gamma$ or $\delta = \gamma$, respectively).

His single promotion model explains how reference-price effects impact promotion profits of the retailer. He extends the single period model to multiple periods using dynamic programming to analyze the impact of reference prices on recurring promotion strategies. The recurring promotion policies

prescribe when the retailer should promote and what promotion price the retailer should charge. He analyzes the optimal promotion policies numerically for a brand of peanut butter sold by a supermarket chain. His main findings from this numerical study are the following:

- Gains have higher impact on demand than losses. $\delta > \gamma$ (loss-seeking consumers).
- Repeating cycles of profitable promotions can occur when consumers are loss-seeking. In these promotions, the nonpromotion price is charged at least half of the time. In repeating promotion cycles, promotions are followed by regular prices.
- Repeating cycles of profitable promotions can also occur when consumers are loss-averse or risk-neutral. In these promotions, the retailer charges the regular price less than half of the time and the demand at the regular price is zero. These are called *reverse promotions*.

In reverse promotions, the only purpose of charging the high regular price is to raise consumers' reference prices. This implies that it would be better for the retailer to completely sacrifice the demand during the periods of high price in order to raise reference prices and generate greater demand during the periods of low price. The same result is also derived in Fibich *et al.* [30] in a continuous time setting. This result is very surprising since almost all price promotions in practice involve a price reduction and not a price increase. We have to keep in mind that this result is only valid when losses have an equal or greater impact than gains, which was empirically observed only in Greenleaf [18] and Murthi and Kalyanaram [31]. Most empirical studies in the literature, however, suggest that gains have an equal or greater impact than losses [7,14].

Greenleaf [18] also observes that increasing consumer memory parameter has two opposite impacts on promotion profits. When it is higher, reference prices increase more slowly after a promotion. Therefore, more time must elapse before the reference price is

again high enough to create a large reference-price effect that makes promotions profitable if the impact of gains is higher than that of losses. Since profitable promotions occur less frequently, promotion profits decrease. Promotion profits may also increase with consumer memory parameter. When reference prices change more slowly, the retailer has more control over which precise reference price to choose for starting a new promotion. This allows the retailer to select a reference price closer to the one that maximizes promotion profits. This observation supports the managerial importance of estimating consumer memory parameter in developing more profitable promotion policies. He also observes that irregular promotions often maximize profit from reference-price effects. He thus recommends managers to avoid regular promotion timing so that consumers would not strategize the timing of promotion deals. Optimal pricing policies in Greenleaf [18] are summarized in Table 1. Greenleaf [18] concludes that reference-price effects impact promotion profits. Thus, retailers should not ignore reference-price effects and should develop promotion strategies that allow reference-price effects to increase profits.

Discrete Time Setting

The objective of Kopalle *et al.* [26] is to generalize the results in Greenleaf [18]. They extend Greenleaf's result analytically for a single brand in a monopolistic setting where all customers are homogeneously either loss-averse or loss-seeking. They use this setting as a building block to analyze a monopolist retailer with multiple brands. They first consider a monopolist firm that maximizes the sum of discounted profits over an infinite time horizon for a single brand. They assume that prices are bounded. They also assume that the reference price in each period is an exponentially weighted average of past observed prices as we explain in Equation (1). Following Greenleaf [18], they assume that the demand in period t is $d(r(t), p(t))$ as in Equations (5)–(6), without the error term $\varepsilon(t)$. They assume that $q(p(t))$ is continuous and differentiable. They use a dynamic programming model to determine the optimal

Table 1. Optimal Pricing Policies When Consumers are Loss-Seeking (Loss-Averse)^a

	Loss-Seeking Consumers (Loss-Averse Consumers)					
	Exponential smoothing process			Reference brand process		
	Monopoly	Duopoly	Oligopoly	Monopoly	Duopoly	Oligopoly
Greenleaf [18] (Linear reference-price effects, linear price effects)	Cyclical (Cyclical)	—	—	—	—	—
Kopalle <i>et al.</i> [26] (linear reference-price effects, nonlinear price effects)	Cyclical (Constant)	Cyclical (Constant)	Cyclical (Constant)	Cyclical (Constant)	Constant (Constant)	(Constant)
Popescu and Wu [24] (nonlinear reference-price effects, nonlinear price effects)	Cyclical (Constant)	—	—	—	—	—

^aRefs 18, 24, and 26.

pricing policy for a given value of initial reference price. They derive the following three main results for the monopolist firm:

- If all consumers are loss-seeking $\delta > r$, it is optimal for the monopolist to vary prices over time.
- If all consumers are risk-neutral $\delta = r$, the optimal pricing policy monotonically converges to a constant steady-state price.
- If all consumers are loss-averse $\delta < r$ and the consumer memory parameter $\alpha = 0$, then the optimal pricing policy monotonically converges to a constant steady-state price.

On the basis of the above three main results, they conjecture that the optimal pricing policy for a monopolist firm with a single brand monotonically converges to a constant price when consumers are either risk-neutral or loss-averse. They examine this conjecture numerically using linear and nonlinear $q(p(t))$ functions that have empirical support in the marketing literature. Their numerical findings indicate that the optimal pricing policy maximizing either the sum of discounted profits or average profit per period monotonically converges to a constant steady-state price. They extend their model to a monopolist firm with multiple brands. They assume

that each brand forms its own reference price independently using only its own prior prices. They derive the following three main results for the monopolist firm selling multiple brands:

- If all consumers are loss-seeking $\delta > r$ for all brands, then at least one brand must have a cyclical pricing policy when the sum of discounted profits over an infinite horizon is maximized.
- If all consumers are loss-seeking $\delta < r$ for all brands, then all brands have cyclical pricing policies when the average profit per period is maximized.
- If all consumers are loss-averse $\delta < r$ for all brands, then a constant price for all brands is more profitable than cyclical prices for any brand when the average profit per period is maximized.

Kopalle *et al.* [26] also consider a duopoly case in which two brands compete with each other. Each brand maximizes its own average profit per period. Reference-price effects of each brand are independent from one another. The only interdependency between the two competing brands is the cross-price effect. They derive the following four main results for the duopoly case:

- If each brand decides to maintain cyclical prices, the equilibrium strategy is to offer cyclical prices.
- If consumers are loss-seeking, the equilibrium strategy is to offer cyclical prices.
- If consumers are loss-averse, a constant price policy performs better than a cyclical pricing policy.
- If consumers are loss-seeking toward the first brand, whereas they are loss-averse towards the second brand, in equilibrium the prices of the first brand should cycle while the second brand should maintain a constant price.

Kopalle *et al.* [26] also investigate using simulation on how their results depend on the specific reference-price formation process—exponential smoothing process. They use an alternative reference formation process, an aggregate version of the *reference brand* formation process in Hardie *et al.* [22], to check robustness of their results. Under this alternative reference formation, the reference price is the current price of the last brand purchased (called the *reference brand*) by consumers. They simulate both a monopolistic retailer with two brands and a duopoly case under the reference brand formation process. For the monopolist retailer with two brands, they observe that constant (cyclical) pricing policy is optimal for both brands when consumers are loss-averse (loss-seeking). These observations on pricing policies are the same as those under the exponential smoothing process. They also observe that a constant pricing policy is optimal for both brands when consumers are loss-seeking for one brand and loss-averse for the other brand. This observation is in contrast to the corresponding one for the monopolist retailer with two brands using the exponential smoothing process. For the duopoly case, their numerical findings indicate that a constant pricing policy is optimal for both brands when consumers are all either loss-seeking or loss-averse. They also observe that a constant pricing policy is still optimal when consumers are loss-seeking for one brand and loss-averse for the other. Under the reference brand

formulation, pricing decisions of each brand impact the other through cross-price effect and reference-price effect. Thus, neither brand can increase its demand by executing cyclical pricing policies. Under the exponential smoothing process, however, consumers select a brand in any period independent of their prior choices. Thus, brands can increase their demands by executing cyclical prices.

Finally, Kopalle *et al.* [26] consider heterogeneous consumers through segmenting the market into two: loss-seeking consumers (segment 1) and loss-averse consumers (segment 2). Their goal is to determine what percentage of the market should be segment 1 in order for cyclical pricing policies to be optimal—called as *critical ratio*. They derive a sufficiency condition for the critical ratio in order for cyclical pricing policies to be optimal in a single brand monopoly, a symmetric multibrand monopoly, a symmetric duopoly, and a symmetric oligopoly under an exponential smoothing process of reference-price formation. They also numerically explore the critical ratio for the reference brand formation process. Their analytical and numerical findings conclude that a relatively small value of critical ratio leads to cyclical pricing policies to be optimal in all cases when there is a homogenous market with only loss-seeking consumers. Optimal pricing policies in Kopalle *et al.* [26] are summarized in Table 1.

Continuous Time Setting

Similar to Kopalle *et al.* [26], Fibich *et al.* [28] extend the model in Greenleaf [18] to a competitive pricing setting in a duopoly environment. They develop a new method for the determination of optimal pricing strategies explicitly for firms that experience reference-price effects. They adapt the demand function $d(r(t), p(t))$ in Greenleaf [18] in Equations (5)–(6) for a single product in the continuous environment. They model the reference price in continuous time as a continuous weighted average of past prices with an exponentially decaying weighting function as we outline in detail in Equation (3), which can also be represented through the ordinary differential equation as we explain in Equation (2).

They consider a two-stage method to derive the optimal pricing policy explicitly in the presence of asymmetric reference-price effects. In the first-stage, they consider the symmetric reference-price effects $\delta = r$. Using variational optimization tools, they derive an explicit expression for the optimal pricing strategy for the monopolist. The models in Greenleaf [18] and Kopalle *et al.* [26] use a discrete time approach, thus the resulting dynamic programming models become hard to solve for explicit optimal pricing policies. The continuous-time approach by Fibich *et al.* [28] thus has the advantage over these earlier studies. In the second stage, they consider the asymmetric reference-price effects. They show that the solution for the asymmetric case is equivalent to an appropriately chosen solution for the symmetric case. This result is extremely useful. It provides the optimal pricing strategy for the asymmetric case, for which standard optimization techniques do not apply due to the nonsmoothness of the profit function. One can simply use the solution for the symmetric case to construct a solution for the asymmetric case. In summary, the two-stage method is: (i) determine explicitly the optimal pricing strategy for the symmetric reference-price effects, and (ii) use the solution in (i) to construct the optimal pricing strategy for the asymmetric reference-price effects.

They also show that the initial pricing strategy depends on whether the initial

reference price is higher or lower than the steady-state optimal price when the planning horizon is infinite as depicted in Fig. 2:

- If the initial reference price is lower than the steady-state price, then the penetration price is set lower than the steady-state price. During the introductory stage, the price increases monotonically and approaches to the steady-state price (penetration strategy).
- If the initial reference price is higher than the steady-state price, then the initial price is set higher than the steady-state price. During the introductory stage, the price decreases monotonically and approaches to the steady-state price (skimming strategy).

They show that the optimal price reaches to the steady-state value after the introductory stage for the infinite planning horizon case. They also derive the optimal pricing strategy explicitly for the finite planning horizon case. They show that the introductory stage has either penetration or skimming strategy; the introductory stage is followed by a steady-state pricing strategy; and toward the end of the planning horizon optimal price decreases since there is no need to make long-term reference-price planning. In summary, the optimal pricing strategy has

- a short introductory stage in which the initial reference-price value determines

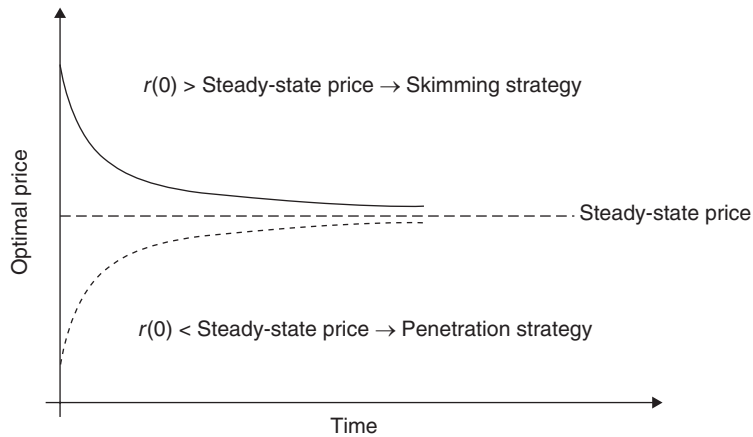


Figure 2. Optimal pricing policies in Fibich *et al.* [28].

- whether there is an initial decrease (skimming) or increase (penetration) in price,
- a steady-state region in which the price is constant when the planning horizon is infinite,
- an additional short final decline stage when the planning horizon is finite.

The above results on the optimal pricing strategy heavily depend on the value of the initial reference price. However, there is no detailed discussion in Fibich *et al.* [28] on how the initial reference price can be estimated. This is an open research question that would be investigated in detail.

Fibich *et al.* [28] extend the model to the oligopolistic Cournot competition among multiple firms in which each firm may have a different unit production cost and decision variables are the firms' production levels. They consider both an open-loop and closed-loop Nash equilibrium. Using the two-stage method (developed for the monopoly case), they derive the open-loop equilibrium strategies explicitly for both the risk-neutral $\delta = r$ case and loss-averse case $\delta \leq r$ when the planning horizon is infinite. They show an interesting result that both the open-loop and closed-loop equilibrium strategies are qualitatively the same as in the monopolistic case, that is, skimming and penetration strategies in the introductory stage and converging to steady-state pricing strategy in the long run. Thus, they conclude that open-loop and closed-loop competition do not add new qualitative features to the model (i.e., they do not lead to cyclical pricing strategies).

To measure the impact of the optimal pricing strategy, they compare the optimal pricing strategy in the presence of reference-price effects with the following three alternative (suboptimal) pricing strategies

- *Pricing Strategy with No Reference-Price Effects.* The firm ignores the reference-price effects in developing the optimal pricing strategy (the optimal price is constant under the no reference-price effects).
- *Myopic Price Strategy.* The firm takes reference-price effect into account, however, follows a myopic approach; the firm develops the optimal pricing strategy that maximizes the instantaneous profits.
- *Optimal EDLP (Every Day Low Price) Strategy.* The firm chooses a constant optimal pricing strategy throughout the planning horizon.

They derive three main results in the comparison. First, the differences in profits among the optimal pricing strategy, optimal pricing strategy with no reference-price effects, and optimal EDLP strategy are very small. Firms, thus, can obtain near optimal profits using EDLP strategy. The analysis in Fibich *et al.* [28] thus provides a strong support for the EDLP strategy. Second, the myopic pricing strategy may result in substantially lower profits compared to the optimal pricing policy with no reference-price effects. Third, the optimal steady-state price is slightly below the optimal price in the absence of reference-price effects. They validate these three results through an empirical study using the data from Greenleaf [18] for a brand of peanut butter sold by a supermarket chain.

EXTENSIONS

Generalized Consumer Demand

Following Greenleaf [18], Popescu and Wu [24] treat the problem of consumers with memory and asymmetric reference-price effects in a discrete time setting using dynamic programming techniques to characterize dynamic pricing strategies for a frequently purchased product. Their objective is to extend the models in Greenleaf [18], Kopalle *et al.* [26], and Fibich *et al.* [28] using very general modeling assumptions on consumer demand. Greenleaf [18], Kopalle *et al.* [26], and Fibich *et al.* [28] use the demand model with piecewise linear reference-price effects in Equations (5) and (6), which limits the applicability of their results. Also, different from Fibich *et al.* [28], they provide structural results on optimal

pricing strategies rather than parametric results. Thus, Popescu and Wu [24] extend and generalize the models used in Greenleaf [18], Kopalle *et al.* [26], and Fibich *et al.* [28].

They consider a discrete-time model in which a monopolist repeatedly interacts with homogenous consumers over an infinite time horizon. In each period, the monopolist decides what price to charge from a bounded positive price interval. They assume a general reference-price dependent demand which is decreasing in price and increasing in the reference price. In each period, consumers update their current reference price through the exponential smoothing reference-price formation process in Equation (1). The objective of the monopolist is to choose prices in each period to maximize its discounted sum of long-term profits (over an infinite planning horizon) for a given value of initial reference price. Popescu and Wu [24] construct a dynamic programming model to analyze the optimal pricing path for the monopolist with the reference price as the state variable. More specifically, their objective is to determine when a steady-state solution exists for this dynamic programming model and to characterize stability and monotonicity properties of the optimal price and reference price paths.

They first analyze the case when consumers are risk-neutral (i.e., the reference-price effects are symmetric). For the risk-neutral case, they characterize the convergence of the optimal prices to a unique steady-state optimal price. They next extend their analysis to explore the transient and the steady-state behavior of the optimal dynamic pricing strategy when consumers are loss-averse as predicted by the prospect theory. They characterize the convergence of the optimal prices to a unique steady-state optimal price. Optimal pricing policies in Popescu and Wu [24] are summarized in Table 1. They derive a list of alternative sufficiency conditions in order for all optimal pricing paths to converge monotonically to the unique steady-state price. With these sufficiency conditions, they show that there are only three optimal pricing strategies: penetration strategy, if initial reference price is low (begin with charging a low

price and keep increasing it over time); skimming strategy, if initial reference price is high (begin with charging a high price and keep decreasing it over time); and constant pricing strategy. They derive the following structural results on the optimal pricing policies:

- An optimal steady-state price exists if consumers are either risk-neutral or loss-averse.
- If consumers are loss-seeking, there is no steady-state pricing strategy—the optimal pricing policy is cyclical. Kopalle *et al.* [26] obtain the same result when reference-price effects are piecewise linear. Greenleaf [18] also shows numerically that in the presence of reference-price effects, the optimal pricing policy for a monopolist can be cyclical.
- The optimal prices converge to a steady-state price after a transient stage.
- Optimal steady-state price decreases with α , the consumer memory parameter.
- As consumers become more loss-averse, the range of optimal steady-state prices becomes wider, that is, increasing pricing flexibility.
- Optimal reference prices follow a monotonic trajectory.
- Optimal prices follow a monotonic trajectory if consumers are very sensitive to reference-price effects or they have a short-term memory of past prices.

They also compare the pricing strategies of two firms for which consumer memory evolves according to the same process starting with the same initial reference price. One firm (myopic firm) ignores the effect of current prices on future demand, thus applies a myopic pricing policy, whereas the other firm (optimizer) considers the impact of current prices on future demand in developing its pricing policy. They show that the myopic firm systematically charges lower prices than the optimizer firm. This result is driven by the assumption that the profit function is supermodular in price and reference price.

By charging higher prices, the optimizer firm increases its future profits through increasing the reference price.

Popescu and Wu [24] enrich the state of the literature on dynamic pricing with reference-price effects by developing a very general nonlinear reference-price-dependent demand model that considers the dynamics of reference-price effect in response to changes in the reference price. Popescu and Wu [24] confirm the results of the steady-state analysis of the optimal pricing strategy in Kopalle *et al.* [26] and Fibich *et al.* [28], but not necessarily monotonicity of the transient pricing policy. Their approach is similar to the two-stage method in Fibich *et al.* [28]. Their technical contributions are characterization of the optimal steady-state prices, and convergence and monotonicity properties of the optimal pricing strategy. Their methodological contribution is to provide a new method (similar to the two-state method of Fibich *et al.* [28]) of proving structural properties for dynamic programming models with kinked rewards.

MULTIPRODUCT DYNAMIC PRICING

Calicchio and Krell [32] find that a surprisingly small number of products (less than four on average) account for half of the information customers use to determine their perceptions of a store's price level. These products are usually the generic and most commonly used household items. Asvanunt and Kachani [33] refer to these as the *core* products. They argue that considering core products has important pricing implications for retailers. They show that a retailer can price its core products competitively in order to improve the customers' perception of their store's price level and to lure more customers to shop at the store. The economics literature models competition through assuming that the demand of a product is a function of both its competitor's price and its own price. Asvanunt and Kachani [33] develop an alternative model (the concept of store reference price) where the retailers compete through customer's perception of their stores' price levels. They first consider the optimal dynamic pricing policy in discrete time

for a monopolist retailer offering multiple products, where the monopolist has the objective of maximizing its revenue over an infinite time horizon. They use the following reference-price formulation

$$r(t) = \alpha r(t-1) + (1-\alpha)\lambda'p(t-1), \quad (7)$$

where λ is a vector of nonnegative weights and $\lambda'p$ as the weighted average price of the retailer's products—the current store's price level. They assume that the demand for a product has the following form

$$d(r(t), p(t)) = q(p(t)) + \lambda(r - \lambda'p), \quad (8)$$

where $q(p(t))$ is linearly decreasing in price $p(t)$.

They derive a closed-form solution of the optimal pricing policy for a monopolist selling a single product. Their approach in discrete time yields the same expression of the steady-state reference price as in Fibich *et al.* [28]. They also prove in discrete time the same monotonicity and convergence results as in Fibich *et al.* [28]. It is interesting to point out that the optimal policy in discrete time is a function of the state that is the current level of the reference price, and is much more practical than the continuous-time solution, which was provided as a function of time in Fibich *et al.* [28]. They extend their results to multiple products and show that the monotonicity and convergence results still hold. They find that for otherwise identical products, the steady-state price of a core product is lower than that of a noncore product.

They next extend their model to a duopoly case in which two competing retailers offer the same two products to consumers. They derive the optimal pricing policy for a retailer when (i) its competitor's price is constant, (ii) its competitor's price is a linear function of the retailer's own price, and (iii) its competitor follows a myopic pricing policy (only maximizing the current revenue). They also illustrate how their framework can be extended for a duopoly in which each retailer offers identical multiple products, and how a similar closed-form heuristic may be obtained.

Dynamic Pricing with Product Introduction Decisions

Arslan *et al.* [34] consider a firm's decision on the product introduction timing and dynamic pricing for successive product generations in the presence of reference-price effects. They analyze the case where a firm introduces multiple generations of a product, for which demand is impacted by the reference price formed by consumers' shopping experience. Their model is based on the dynamic pricing model in Arslan *et al.* [35].

They focus on developing dynamic pricing and product introduction policies for two successive product generations—a base product and a new product. They use the demand functions in Equations (5)–(6), with symmetric reference-price effects. They also assume that reference price for each generation is formed through continuous weighted moving average of past prices with exponentially decaying weighting function as in Kopalle *et al.* [23] and Fibich *et al.* [24]. They consider two scenarios: complete replacement of the old product by the new one, and the case of generations offered concurrently. For the complete replacement case, they define the reference price for the new product recursively through base product's reference price and posted price. For the case of coexisting generations, they consider three cases of how reference prices evolve. In the first case, similar to the reference-price dynamics in complete replacement, reference prices of the concurrent product generations only share a common past, but evolve independently at present. In the second case, they assume that the firm considers a joint reference price for product generations. Each generation contributes to the joint reference price with a relative weight that measures how consumers are impacted by different existing product generations in forming their perceptions. In the third case, concurrent generations form the joint reference price as in the previous case; however, the two generations internally compete with each other. This case is applicable to products that are offered by competing divisions with different revenue objectives within the same firm.

They characterize the product introduction timing and pricing strategies explicitly

for two successive product generations, and propose a quasi-analytic solution method, based on the structure of the memory-dependent problem, for an arbitrary number of generations. They show that the total profit function and its derivatives are quasipolynomial functions in introduction time. They also compute the derivative of the profit function with respect to introduction time, which is subsequently used to determine the optimal introduction time. They show that the sign of this derivative depends on profit contribution margins of the base product and the new product. They observe that, as consumers become sensitive to reference-price effects, optimal introduction time shifts toward the less profitable product. They indicate that optimal introduction time as the root of the quasipolynomial profit function is not available in closed-form; however, they show that it can be calculated as a first-order correction to its corresponding no consumer memory value since an explicit expression for the derivative of the introduction time with respect to reference-price sensitivity is already derived. Their numerical experiments show that such an approximation performs very well. They show that as consumers become more sensitive to the reference-price effects, then firms gain more in profits. They extend their analysis to a duopoly environment in which two firms compete with each other, derive the Nash equilibria for the duopoly game, and illustrate how reference-price effects influence the Nash equilibria.

Dynamic Pricing with Product Innovation Decisions

Arslan *et al.* [36] study the competition among firms that produce and sell short life cycle products to consumers who are segmented in terms of price sensitivity and their purchasing behavior toward new and innovative products. Their objective is to derive insights on how pricing and product innovation strategies impact these firms' market positioning. They assume that the market demand for a newly introduced short life cycle product responds to two main factors: deviation of the product's own price from a marketwide reference price and deviation

of the product's innovation (manner) level from the current market average.

They assume that consumers evaluate attributes of past products to form a reference innovation level that affects their current perceptions on how innovative new products and firms are. They assume that consumers determine the reference product innovation level through a "mental calculation" of the average rate of product improvements or the average rate of new product introductions in the market. According to their model, consumers compare a product's current attributes or a firm's current innovation rate with their formed perception of the product innovation level in the market to determine their perceived gains and losses from purchasing the current product. Consumer demand increases (decreases) as consumers perceive gains (losses).

They assume that the reference product innovation level effects on consumer demand are linear; the reference product innovation level effects are additive; and the impact of perceived gains and losses through reference product innovation level effects on consumer demand is symmetric.

They formulate pricing and product innovation strategies in a differential game model. Their model measures the joint impact of reference prices and reference innovation levels in formulating optimal pricing and innovation strategies for competing firms. They incorporate the inherent diversity in consumer demand by characterizing consumers in terms of their sensitivity to product innovation (fashion) levels and product prices. They also characterize different types of firms in the market in terms of their pricing and innovation strategies. They are the first to consider reference memory effects of product innovation.

They derive the optimal pricing and innovation strategies for firms in a competitive environment. They show that the solution for the duopoly game under open-loop strategies is a linear combination of exponentials. They extend their analysis to an oligopoly game with N competing firms and derive the corresponding open-loop Nash equilibria. They consider two special cases of firms: discounter firms that control their pricing

policy while keeping their product innovation level constant and innovation-oriented firms that control their product innovation policy while executing constant pricing policy. They also analyze specific competition among firms that adhere to different types of pricing and product innovation strategies and derive the corresponding open-loop Nash equilibria in these settings. They derive the optimal pricing policy for a pure discounter firm under both monopoly and duopoly settings. They also derive optimal product innovation policies for pure innovation-oriented firms in a duopoly environment. They analyze the competition between a pure discounter firm and a pure product innovation-oriented firm. They determine optimal pricing and product innovation strategy for a firm that competes with a rival that executes both constant pricing and product innovation policies. They also show that the difference between closed-loop Nash equilibria and open-loop Nash equilibria is small, at least for reasonably low values of demand sensitivity parameters.

They consider a case study in the fast-fashion retail industry, to illustrate how their differential game model is applied in an environment with firms executing different pricing and product innovation strategies to compete. For example, they consider a game between fast-fashion clothing retailer Inditex and its competitor H&M. Inditex has the ability to introduce new products very fast (the average shelf-life of a product is as low as 24 days). Since they replace the products very quickly, they do not have the time to focus on price discounts. Thus, they choose product innovation as their main competitive strategy. H&M, on the other hand, does not have the ability to replace its products as fast as Inditex does. They have longer shelf-life for their products. Therefore, they utilize price discounts as their main marketing strategy. Another competitor type is a large department store that follows a hybrid strategy. It replaces products faster or provides more options to consumers in terms of product variety. At the same time, it also applies heavy price discounting. For such competitors, we develop a symmetric equilibrium and compare it against less elaborate strategies that focus on only product innovation or price

discounting. Their findings in the case study suggest that a hybrid strategy, which puts emphasis on both pricing and innovation, is optimal. Thus, they suggest that any comprehensive decision tool to determine pricing and product development (innovation) strategies should seek to integrate both reference price and reference product innovation level effects. Their conclusions are also confirmed by an extensive empirical study of new products' performance in the manufacturing sector by Carbonell and Rodriguez [37].

Price Negotiation

Huh *et al.* [38] extend the model in Kopalle *et al.* [26] and Asvanunt and Kachani [33] to a setting in which consumers are offered two separate prices. They classify consumers into two classes depending on how they form their reference prices: price-takers and price-hagglers. Price-takers form their reference prices through exponential smoothing reference-price process as in Equation (1). Price-hagglers, on the other hand, consider the price offered to price-takers as their reference price, similar to Hardie *et al.* [22]. They assume that demand from price-takers depends on both the posted price and the reference price, whereas the demand from price-hagglers depends on the price offered to them and how much this price deviates from the price offered to price-takers. Thus, they assume that price-takers do not have memory of past prices; instead they choose the current price offered to price-takers as of their reference price. They assume that demand functions are linear in posted prices and reference-price effects are linear. They also assume that consumers are loss-neutral (i.e., reference-price effects are symmetric).

They construct a dynamic programming model in which a monopolist retailer chooses its optimal pricing policy for price-takers and price-hagglers to maximize its revenues from a single product over an infinite planning horizon. They derive the closed-form solution for the optimal pricing policies for the monopolist. They also prove the monotonicity and convergence results of the optimal pricing policies. They extend their model to a duopoly setting, similar to the one introduced

by Asvanunt and Kachani [33], in which the competition occurs through the reference-price effect and consumers use a weighted average of prices offered by both firms to form their reference price. They show that monotonicity and convergence results hold true for the duopoly setting when one of the retailers follows a suboptimal policy—a constant pricing policy, a markup (down) pricing policy, or a myopic pricing policy (maximizing current period revenue). They provide a numerical method to calculate equilibrium pricing policies when the equilibrium policies are of linear form. They also perform a numerical analysis to measure the value of offering two separate prices to consumers by comparing their numerical results to those in Asvanunt and Kachani [33].

Peak-End Consumer Anchoring

Nasiry and Popescu [39] study the impacts of a new memory-based reference-price formation, based on a peak-end memory model of Fredrickson and Kahneman [40], on dynamic pricing policies with reference-price effects. This new memory-based reference-price model suggests that consumers develop a reference-price based on a weighted average of the lowest and the most recent prices. It is supported by the vast empirical research in the psychology literature [41,42]. It provides a behaviorally compelling alternative to exponential smoothing reference-price formation process in Equation (1), where consumers anchor on a weighted average of all prior prices.

As indicated in Nasiry and Popescu [39], the analysis of the dynamic pricing policies under the new memory-based reference-price formation process is quite difficult due to nonsmoothness in the profit function. Thus, the standard optimization techniques do not work. Nasiry and Popescu [39] develop a novel approach based on a bounding technique to characterize the optimal steady-state pricing policies. They show that a constant pricing policy is optimal for a range of relatively low initial price expectations. Unlike the exponential smoothing reference-price formation process, the optimal steady-state price depends on initial low price expectations. This result suggests that

the impact of initial low price expectations is more lasting than previously perceived in the literature. They also show that when initial price expectations of consumers are outside of the steady-state range, then firms should use price skimming or penetration strategies to gradually manage price perception, which eventually lead prices to converge to the optimal steady-state value in the long run.

The key managerial insight in Nasiry and Popescu [39] is that if a constant pricing strategy, such as EDLP, is engrained in consumer expectations, its long-term effects on prices and profits should be expected and these effects can be lessened through gradually increasing prices to manage consumer expectations.

RESEARCH GAPS AND FUTURE RESEARCH DIRECTIONS

The current studies on dynamic pricing models with reference-price effects are all deterministic and do not incorporate any uncertainty in model parameters. Thus, it would be useful to develop models that explore optimal dynamic pricing strategies when firms' knowledge on consumers' memory parameter and sensitivity to past prices are uncertain. Such study enables us to check how robust the resulting dynamic pricing policies are under information uncertainty and thus provides important insights in practice.

Consumers may learn through their purchasing experience and build their anticipations on when prices will go down and how prices will cycle over time. Such consumers may act strategically toward the dynamic pricing policies. Thus, firms should take this strategic consumer learning behavior into account in establishing their dynamic pricing policies. It would be useful to investigate what characteristics dynamic pricing policies would have when consumers learn from firms' actions, purchase strategically, and have reference-price sensitivities.

The current studies on dynamic pricing models with reference-price effects do not incorporate any capacity constraints on

firms. They all assume that firms have enough capacity to fill all consumer demand. It would be interesting to check how dynamic pricing strategies interact with capacity decisions. It is also useful to investigate whether product stock-outs have any impact on consumers' reference-price formation process. Another relevant research problem would be to build models to coordinate promotion and production decisions within a firm under the influence of reference-price effects and to investigate the value of taking reference-price effects into account in forming optimal promotion and production decisions.

The current research on dynamic pricing policies with reference-price effects focus on how consumer demand (in terms of quantity) is affected by pricing decisions. For service products, pricing strategies impact consumers' reference prices, which may impact how long consumers use the service. Thus, current studies can be extended for service products to explore how the duration of service usage by consumers is impacted by reference-price effects.

The current studies model the reference-price effects on demand through an additive function as in Equations (5)–(6). This modeling approach can be extended by considering multiplicative reference-price effects on demand.

The reference-price effect has been shown to vary across consumers [11]. This creates an opportunity for segmenting consumers based on their reference-price sensitivities. Thus, one can consider how a firm should develop dynamic pricing policies when there are several consumer segments with different reference-price sensitivities.

Current studies assume that reference-price effects for both gains and losses are independent from consumers' sensitivities to price and brand loyalty. It would be useful to develop both analytical and empirical models to assess the robustness of this assumption in developing optimal dynamic pricing and price promotion models.

REFERENCES

1. Ginzberg E. Customary price. *Am Econ Rev* 1936; 26:296.

2. Gabor A, Granger C. Price sensitivity of the consumer. *J Advert Res* 1964;4:40–44.
3. Monroe KB. Buyers' subjective perceptions of price. *J Mark Res* 1973;10:70–80.
4. Della Bita AI, Monroe KB. The influence of adaptation level on selective price perceptions. In: Ward S, Wright P, editors. *Advances in consumer research 1*. Ann Arbor, (MI): Association for Consumer Research; 1974. pp. 359–369.
5. Kalwani MU, Yim CK. Consumer price and promotion expectations: an experimental study. *J Mark Res* 1992;29:90–100.
6. Helson M. *Adaptation level theory*. New York: Harpor and Row; 1964.
7. Winer RS. A reference price model of demand for frequently purchased products. *J Consum Res* 1986;13(2):250–256.
8. Winer RS. Behavioral perspectives on pricing: buyers' subjective perception of price revisited. In: Timothy MD, editor. *Issues in pricing: the theory and research*. Lexington, (MA): Lexington; 1988. pp. 35–37.
9. Kalyanaram G, Winer RS. Empirical generalizations from reference price research. *Market Sci* 1995;14(3):G161–G169.
10. Briesch RA, Krishnamurthi L, Mazumdar T, *et al.* A comparative analysis of reference price models. *J Consum Res* 1997;24(2):202–214.
11. Mazumdar T, Raj SP, Sinha I. Reference price research: review and propositions. *J Market* 2005;69:84–102.
12. Kalyanaram G, Little JDC. An empirical analysis of latitude of price acceptance in consumer package goods. *J Consum Res* 1994;21(3):408–418.
13. Kalwani MU, Heikki JR, Sugita Y. A price expectations model of consumer brand choice. *J Market Res* 1990;27:251–262.
14. Lattin JM, Bucklin RE. Reference effects of price and promotion on brand choice behavior. *J Mark Res* 1989;26:299–310.
15. Putler DS. Incorporating reference-price effects into a theory of consumer choice. *Market Sci* 1992;11:287–309.
16. Winer RS. A price vector model of demand for consumer durables: preliminary developments. *Market Sci* 1985;4(Winter):74–90.
17. Thaler RH. Mental accounting and consumer choice. *Market Sci* 1985;4(3):199–214.
18. Greenleaf EA. The impact of reference-price effects on the profitability of price promotions. *Market Sci* 1995;14:82–104.
19. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
20. Krishnamurthi L, Mazumdar S, Raj P. Asymmetric response to price in consumer brand choice and purchase quantity decisions. *J Consum Res* 1992;19(3):387–400.
21. Mayhem GE, Winer RS. An empirical analysis of internal and external reference prices using scanner data. *J Consum Res* 1992;19:62–70.
22. Hardie BGS, Johnson EJ, Fader PS. Modeling loss aversion and reference dependence effects on brand choice. *Market Sci* 1993;12:378–394.
23. Anderson CK, Rasmussen H, MacDonald L. Competitive pricing with dynamic asymmetric price effects. *Int Trans Oper Res* 2005;12:509–525.
24. Popescu I, Wu Y. Dynamic pricing strategies with reference effects. *Oper Res* 2007;55(3):413–429.
25. Dickson PR, Sawyer AG. The price knowledge and search of supermarket shoppers. *J Market* 1990;54(3):42–53.
26. Kopalle PK, Rao A, Assuncao JL. Asymmetric reference-price effects and dynamic pricing policies. *Market Sci* 1996;15:60–85.
27. Sorger G. Reference price formation and optimal marketing strategies. In: Feichtinger G, editor. *Optimal control theory and economic analysis*. Amsterdam: Elsevier Science, North-Holland Publishing Co.; 1988. pp. 97–120.
28. Fibich G, Gavious A, Lowengart O. Explicit solutions of optimization models and differential games with non-smooth (asymmetric) reference-price effects. *Oper Res* 2003;51(5):721–734.
29. Kopalle PK, Winer RS. A dynamic model of reference price and expected quality. *Market Lett* 1996;7:41–52.
30. Fibich G, Gavious A, Lowengart O. Optimal price promotion in the presence of asymmetric reference price effects. *Manag Decis Econ* 2007;28:569–577.
31. Murthi BPS, Kalyanaram G. An empirical analysis of asymmetry in widths of the region of price insensitivity. *Marketing Science Conference*; Syracuse, NY. 1999.
32. Calicchio N, Krell A. Price promotions in Latin America. *McKinsey Q* 2007. (online issue: April 2007). Available at <https://www.mckinseyquarterly.com/Price-promotions-in-Latin-American-retailing-1973>.

33. Asvanunt A, Kachani S. Multi-product pricing with reference effects in monopoly and duopoly markets. Working Paper Columbia University; 2008.
34. Arslan H, Kachani S, Shmatov K. Joint memory-dependent pricing and product introduction for multiple generations. Working Paper Columbia University; 2009.
35. Arslan H, Kachani S, Shmatov K. Optimal product introduction and life cycle pricing policies for multiple product generations under competition. *J Revenue Pricing Manag* 2009;8(5):438–451.
36. Arslan H, Kachani S, Shmatov, K. Competition in innovation and pricing for short life-cycle products. Working Paper Columbia University; 2009.
37. Carbonell PA, Rodriguez I. The impact of market characteristics and innovation speed on perceptions of positional advantage and new product performance. *Int J Market Res* 2006;23:1–12.
38. Huh WT, Kachani S, Sadighian A. A two-stage multi-period negotiation model with reference price effect. Working Paper Columbia University; 2009.
39. Nasiry J, Popescu I. Dynamic pricing with loss averse consumers and peak-end anchoring. Working Paper INSEAD, 2010.
40. Fredrickson BL, Kahneman D. Duration neglect in retrospective evaluations of affective episodes. *J Pers Soc Psychol* 1993;65(1): 45–55.
41. Fredrickson BL. Extracting meaning from past affective experiences: the importance of peaks, ends and specific emotions. *Cogn Emot* 2000;14(4):577–606.
42. Kahneman D. Evaluation by moments: past and future. In: Kahneman D, Tversky AS, editors. *Choices, values, and frames*. New York: Cambridge University Press and The Russell Sage Foundation; 2000;693–708.

DYNAMIC PROGRAMMING: INTRODUCTORY CONCEPTS

MOSHE SNIEDOVICH
Department of Mathematics and Statistics,
University of Melbourne, Melbourne,
Victoria, Australia

Dynamic programming (DP) is a general-purpose problem-solving methodology based on problem decomposition. Its origins, more accurately the origins of some of its rudimentary ideas, can no doubt be traced back to the days when humans began to *plan*. But it was the applied mathematician Richard Ernest Bellman (1920–1984), who brought together and systematized the simple, intuitive, yet highly effective ideas that underlie this problem-solving methodology, into a coherent theory [1,2]. Details on the history of DP can be found in Bellman's [3] autobiography and in Refs 4 and 5.

To set the scene, call the problem of interest, the *Target Problem*: This is the problem we seek to solve. DP approaches the solution of the *Target Problem* indirectly, by considering a family of “related problems,” called *Modified Problems*. Typically, the *Modified Problems* are generated by varying the values of some parameters of the *Target Problem*. In this framework, the *Target Problem* is equivalent to one of the *Modified Problems* associated with some given values of these parameters.

Conceptually, the family of *Modified Problems* represents the sequence of problems arising from the application of a sequence of *decisions*. That is, following the application of the first decision, the process will progress to a new *state* so that a decision must be made as to what is to be done next. This process either continues indefinitely or terminates upon reaching its *final state*.

The question is then how should the sequence of decisions be determined, given the *goals* and *constraints* stipulated by the *Target Problem*?

THE METARECIPE

Methodologically, the DP approach to problem solving can be formulated in terms of the following metarecipe:

The DP Metarecipe

1. **Define** the *Target Problem*.
2. **Embed** the *Target Problem* in a family of related *Modified Problems* using one or more features of the *Target Problem* as parameters.
3. **Derive** a *Functional equation* relating the solutions of the *Modified Problems* to one another.
4. **Solve** the functional equation.
5. **Recover** a solution for the *Target Problem* from the solution obtained for the functional equation.

The main goal of this discussion is to describe the basic concepts employed by this recipe. The focus is on the first three stages of the recipe, hence on the *modeling* aspects of DP.

THE ART OF DYNAMIC PROGRAMMING

DP's treatment of a problem, especially modeling a problem in DP style, is widely considered to be an “art” [6,7]. This, apparently, is due to the considerable dose of imagination, creativity, insight, and so on that the modeler/analyst must often inject into the formulation of a DP model and the functional equation of DP. In practice, this means that general recipes outlining in broad terms the DP approach to problem solving are of the *easier said than done* type.

The function of the above recipe is to provide a succinct description of DP's general plan of attack; more accurately, the logic driving DP's treatment of a problem. And to illustrate its working, we shall examine DP in the setting of an *optimization problem*.

The reason for limiting this discussion to optimization problems is obvious: Although the DP approach is in fact applicable to a wide variety of problem types, in the field of *operations research*, it is used primarily to model, analyze, and solve *optimization problems*. Therefore, in this discussion the focus is exclusively on this class of problems.

The specific optimization problem to be used to illustrate the modeling aspects of DP will be the following well-known combinatorial optimization problem [8].

Traveling Salesman Problem (TSP). Determine the order in which to visit a prescribed set of cities so as to minimize the total travel time. It is required that each city be visited exactly once. The trip commences and terminates at a given “home” city.

There are several mathematical formulations of this problem. For the purposes of this discussion, it is convenient to state the TSP as follows:

Target Problem for the TSP:

$$t^* := \min_{x_1, \dots, x_k} \left\{ t(h, x_1) + t(x_k, h) + \sum_{j=1}^{k-1} t(x_j, x_{j+1}) \right\} \quad \text{s.t. } \{x_1, \dots, x_k\} = \mathcal{K}, \quad (1)$$

where h denotes the home city, $\mathcal{K} := \{1, 2, \dots, k\}$ denotes the set of cities to be visited, $t(i, j)$ denotes the travel time along the (shortest) direct link from city i to city j , and x_j denotes the j th city on the tour. If there is no direct link from city i to city j , set $t(i, j) = \infty$.

Note that since $\mathcal{K} = \{1, 2, \dots, k\}$, it follows that the constraint $\{x_1, \dots, x_k\} = \mathcal{K}$ in Equation (1) requires a tour $(x_1, \dots, x_k)$ to be a *permutation* of the elements of $\mathcal{K}$. Also, observe that formally we have $x_{k+1} = h$, namely the last decision is “return to the home city.” However, since this decision is fixed in advance, we do not incorporate it explicitly in this model.

To formulate a DP functional equation for this *Target Problem*, let

$$\mathbf{S} := \{(c, C) : C \subset \mathcal{K}, c \in \mathcal{K} \setminus C\} \cup \{(h, \mathcal{K}), (h, \{\})\}, \quad (2)$$

where $\{\}$ denotes the empty set.

This set ($\mathbf{S}$) represents the *state space* of the problem. A state $s \in \mathbf{S}$ is then a pair (c, C) , where c represents the “current city,” namely, the city being visited, and C represents the set of cities yet to be visited. The *initial state* is $s = (h, \mathcal{K})$ and the *terminal state* is $s = (h, \{\})$. As the tour progresses from one city to the next, the state of the process is varied in accordance with the following *transition law*: if while at state $s = (c, C)$ the decision is to go to next city $x \in C$, then the next state of the process will be $s' = (c', C')$, where $c' = x$ and $C' = C \setminus \{x\}$.

Consider now the following family of *Modified Problems*.

Modified Problem $P(c, C)$, $(c, C) \in \mathbf{S}$:

$$f(c, C) = \min_{x_1, \dots, x_m} \left\{ t(c, x_1) + t(x_m, h) + \sum_{n=1}^{m-1} t(x_n, x_{n+1}) \right\}, m = |C| \quad (3)$$

$$\{x_1, \dots, x_m\} = C, \quad (4)$$

where $|C|$ denotes the cardinality of set C .

In this formulation, the parameter $s = (c, C)$ represents the *state* of the decision-making process, observing that for $s = (h, \mathcal{K})$, *Modified Problem $P(s)$* is equivalent to the *Target Problem*, hence $t^* = f(h, \mathcal{K})$ and any optimal solution to *Modified Problem $P(h, \mathcal{K})$* is also optimal with respect to the *Target Problem* given in Equation (1).

Note that to simplify the notation, we do not include the final decision $x_{k+1} = h$ (“return to the home city”) explicitly in this formulation. However, observe that the impact of this decision on the objective function is represented by the term $t(x_m, h)$ in Equation (3).

In words, the task set by *Modified Problem $P(c, C)$* is to find the shortest tour that commences at city c , goes through the cities in

C , and terminates at the home city h , subject to the constraint that each city in C must be visited exactly once.

From a DP point of view, *Modified Problem* $P(c, C)$ requires identifying the best sequence of decisions $(x_1, \dots, x_m)$ (with $x_{m+1} = h$) that transforms the state $s = (c, C)$ to the final state $s' = (h, \{\})$ subject to the above constraint. The term $t(x_m, h)$ in the objective function (3) represents the cost of the transition from the state $(x_m, \{\})$ to the final state $(h, \{\})$.

The DP functional equation for these *Modified Problems* is as follows:

$$f(c, C) = \min_{x \in C} \{t(c, x) + f(x, C \setminus \{x\})\}, \quad (c, C) \in \mathbf{S}, C \neq \{\}, \quad (5)$$

with $f(c, \{\}) := t(c, h)$, $\forall c \in \mathcal{K}$ (representing the transition cost from state $s = (c, \{\})$ to the final state $s' = (h, \{\})$).

The element of “art” comes into play in the DP approach when the statement of the *Target Problem* is translated into a family of *Modified Problems* for which there is a valid DP functional equation. The question of course is what kind of “instruction” can be given in this regard. Consider then Bellman’s [2, p. 82] thoughts on this matter:

We have purposely left the description a little vague, since it is the spirit of the approach that is significant rather than the letter of some rigid formulation. It is extremely important to realize that one can neither axiomatize mathematical formulation nor legislate away ingenuity. In some problems, the state variables and the transformations are forced on us; in others, there is a choice in this matters and the analytic solution stands or falls upon this choice; in still others, the state variables and sometimes the transformations must be artificially constructed. Experience alone, combined with often trial and error, will yield suitable formulations of involved processes.

One need hardly point out that this statement remains relevant to this day, not as a sign of lack of progress, but rather as testimony to the fundamental difficulties encountered in DP modeling.

BASIC CONCEPTS

Experience has shown, though, that when modeling in DP style, it is expedient to treat the *Target Problem* and its affiliated *Modified Problems* as *sequential* or *multistage* decision-making problems [2, 6, 7, 9–11]. The advantage of this approach is of course reflected already in DP’s basic idiom and in the logical progression underlying the derivation of the DP functional equation.

Sequential Decision Models

The theory of DP offers a variety of models to describe the *Target Problem*, *Modified Problems*, and the *functional equation* relating them to one another. For our purposes, however, it will be best to study a simple model, which at the same time will exhibit the main generic characteristics of DP models.

Suppose then that under consideration is a process requiring k decisions $(d_1, d_2, \dots, d_k)$, where k is a given positive integer. In this case, the *Target Problem* under consideration would be as follows:

$$\text{Target Problem: } z^* := \underset{(d_1, \dots, d_k)}{\text{opt}} g(d_1, \dots, d_k), \quad (6)$$

subject to some constraints on the decision variables $(d_1, d_2, \dots, d_k)$, and where g is a real-valued *objective function*.

From the standpoint of DP, this optimization problem is a sequential decision problem induced by a *sequential decision process*. So the fundamental modeling issue here is the following:

Suppose that the values of the first $n < k$ decision variables $(d_1, \dots, d_n)$ have already been determined (not necessarily in accordance with the optimality criterion). *What information about this partial list of decisions is required in order to find the optimal values of the remaining decisions $(d_{n+1}, \dots, d_k)$?*

The information required about $(d_1, \dots, d_n)$ for this purpose is represented by the *state variable* of a DP model, so it is defined accordingly. With this in mind, consider the sequential decision process whose basic ingredients are as follows:

Stages. There are k decision stages, $n = 1, 2, \dots, k$, at each of which a decision d_n is made. The process terminates at the final stage $n = k + 1$, where no decision is made. Let then $\mathcal{N} := \{1, 2, \dots, k\}$ denote the set of decision stages.

There are cases where the stage variable is used *implicitly*. For instance, in the case of the TSP, the state $s = (c, C)$ specifies the stage of the process implicitly, namely, $n = k + 1 - |C|$. In such cases, an explicit reference to n is not required but is often made for convenience and/or clarification.

States. At any stage, the status of the decision-making process is described by a parameter called *state*. Let $S(n)$ denote the set of feasible states at stage n . We shall refer to $S(n)$ as the *state space at stage n* . Let $\mathbf{S}$ denote the set of all the states, namely, set $\mathbf{S} = \cup_{n=1}^{k+1} S(n)$. We shall refer to $\mathbf{S}$ as the *state space*. It is assumed that the initial state of the process, namely s_1 , is given. Let σ denote the initial state, hence let $S(1) := \{\sigma\}$.

Recall that in the context of the TSP, the state of the process is a pair $s = (c, C)$, where c represents the current city and C represents the set of cities to be visited. More formally, the state space is given by Equation (2) with the state $s = (h, \{\})$ being the *terminal state*. In contrast, the state $s = (h, \mathcal{K})$ is the *initial state* associated with the *Target Problem*, meaning that the tour is at the home city ($c = h$) and that all the other cities ($C = \mathcal{K}$) are yet to be visited.

Decisions. The task is to make a sequence of decisions $(d_1, \dots, d_k)$. The decisions available at a given stage depends on the state of the process at that stage. That is, the n th decision, d_n , must be selected from a set whose content depends on the stage and state of the process when the choice is made. Let $D(n, s_n)$ denote the set of feasible decisions associated with state s_n at stage n . Hence, the requirement is that at each stage n , the decision d_n must be selected from the set $D(n, s_n)$.

In the case of the TSP, the decision made at state $s = (c, C)$ is the next city to be visited, hence $D(c, C) = C$, if $C \neq \{\}$ and $D(c, \{\}) = \{h\}$ if $C = \{\}$, recalling that C denotes the set of cities yet to be visited. Note that, as explained above, in this case the stage of

the process is not specified explicitly because its value is determined (implicitly) by the state.

State Transition. The dynamics of the process, insofar as the state is concerned, is governed by a *Markovian* transition function. That is, the next state of the process is determined by the current state and the current decision. Let T denote the (state) transition function, namely, let $T(n, s, d)$ denote the new state resulting at stage $n + 1$ from the decision d made at stage n when the process is in state s . Thus, if the initial state of the process is s_1 and the sequence of decisions $(d_1, \dots, d_k)$ is made, the resulting *state trajectory* $(s_1, s_2, \dots, s_{k+1})$ would be determined by $s_{n+1} = T(n, s_n, d_n)$, $n \in \mathcal{N}$. Note that this means that

$$S(n+1) = \{T(n, s, d) : s \in S(n), d \in D(n, s)\}, \\ n = 1, 2, \dots, k, \quad (7)$$

recalling that $S(1) = \{\sigma\}$, where σ denotes the initial state of the process.

In the case of the TSP, $T((c, C), d) = (d, C \setminus \{d\})$, $d \in D(c, C)$, observing once again that there is no explicit reference to the stage variable.

Objective Function. The decisions are made for the purpose of optimizing some given real-valued function of the decision and state variables, call it g . For simplicity, assume that the objective function is *additive* in that it has the following additive structure.

$$g(s_1, d_1, d_2, \dots, d_k) = \sum_{n=1}^k r(n, s_n, d_n), \quad (8)$$

where r is a real-valued function of (stage, state, decision) triplets. We refer to $r(n, s, d)$ as the *stage return* induced by the (stage, state, decision) triplet (n, s, d) .

In the case of the TSP, $r((c, C), d) = t(c, d)$, namely, the stage return $r((c, C), d)$ represents the travel time from city c to city d along the direct link between these cities.

In the next section, we examine how this model is used to define the *Modified Problems* related to the *Target Problem*.

Modified Problems

To obtain the *Modified Problems* associated with the above model, the values of initial segments of the complete sequence of decisions $(d_1, \dots, d_k)$ are fixed. Those problems that are *equivalent* with regard to the optimal values of the remaining decision variables are then aggregated in the *state* variable. The difference between the *Modified Problems* lies in the choice of the (stage, state) pair. So, in this framework, the *Target Problem* is simply one of the *Modified Problems* associated with a given (stage, state) pair, namely, $(n, s) = (1, \sigma)$, recalling that σ denotes the initial state of the process.

The following formulation captures this intuitive idea and provides a convenient mathematical framework for an examination of the basic concepts of DP.

Modified Problem $P(n, s), n \in \mathcal{N}, s \in S(n)$:

$$f_n(s) := \text{opt}_{d_n, \dots, d_k} \left\{ \sum_{j=n}^k r(j, s_j, d_j) \right\} \quad (9)$$

$$s_n = s \quad (10)$$

$$d_j \in D(j, s_j), j = n, n+1, \dots, k \quad (11)$$

$$s_{j+1} = T(j, s_j, d_j), j = n, n+1, \dots, k. \quad (12)$$

By definition then, $f_n(s)$ denotes the optimal value of the objective function given that the initial stage of the process is $n \in \mathcal{N}$ and the initial state of the process is $s \in S(n)$. The sum of the stage returns is often called the *cost-to-go*. These problems are defined so that the *Target Problem* is equivalent to *Modified Problem* $P(1, \sigma)$.

Let $X^*(n, s)$ denote the set of optimal solutions to *Modified Problem* $P(n, s)$, $n \in \mathcal{N}, s \in S(n)$. Assume that every *Modified Problems* has at least one optimal solution.

We are now ready to sketch out the logical process yielding the DP functional equation for the *Modified Problems* defined above. But before we do this, it is important to point out that the simplicity of the concepts underlying the derivation of the DP functional equation in no way casts doubts on the ability of those concepts to vouchsafe the validity of the equation. Indeed, Bellman himself was aware that such doubts might be raised.

And so, interrupting the discussion in which the derivation of the functional equation was demonstrated for the first time in the book, Bellman suggested the following [[2], p. 8]:

This observation, simple as it is, is the key to all our subsequent mathematical analysis. It is worthwhile for the reader to pause here a moment and make sure that he really agrees with this observation which has the deceptive simplicity of a half-truth.

It is important to keep this suggestion in mind!

Conditional Problems

The notion *Conditional Problems* represents a “what if” ploy aimed at solving the *Modified Problems* surreptitiously. That is, in lieu of asking bluntly the obvious question “what is the optimal value of d_n for *Modified Problem* $P(n, s)$?” which may be rather difficult to work out, we ask the following naive question: “What if we set the value of d_n to some arbitrary element d of $D(n, s)$?” More precisely, “What is the best result that can be obtained in the context of *Modified Problem* $P(n, s)$ if the value of d_n is fixed to some arbitrary $d \in D(n, s)$?”

It turns out that the answer to this question is easily obtainable if we are “allowed” to express the optimal solution to *Modified Problem* $P(n, s)$ in terms of the optimal solutions to modified problems associated with stage $n+1$. Indeed, this takes us closer to formulating the DP functional equation. In short, to obtain the family of *Conditional Problems* associated with the above family of *Modified Problems*, we simply fix the value of decision d_n to some arbitrary value $d \in D(n, s)$. Hence,

Conditional Problem $P(n, s, d), n \in \mathcal{N}, s \in$

$S(n), d \in D(n, s)$:

$$f_n(s, d) := \text{opt}_{d_n, \dots, d_k} \left\{ \sum_{j=n}^k r(j, s_j, d_j) \right\} = r(n, s, d) + \text{opt}_{d_{n+1}, \dots, d_k} \left\{ \sum_{j=n+1}^k r(j, s_j, d_j) \right\} \quad (13)$$

$$s_n = s \quad (14)$$

$$d_n = d \quad (15)$$

$$d_j \in D(j, s_j), \quad j = n+1, \dots, k \quad (16)$$

$$s_{j+1} = T(j, s_j, d_j), \quad j = n, \dots, k. \quad (17)$$

Let $X^*(n, s, d)$ denote the set of optimal solutions to *Conditional Problem* $P(n, s, d)$. Assume that every *Conditional Problem* has at least one optimal solution. Observe that by definition, $f_k(s, d) = r(k, s, d)$, $\forall s \in S(k)$, $d \in D(k, s)$.

Principle of Conditional Optimization

By definition, the only difference between *Modified Problem* $P(n, s)$ and *Conditional Problem* $P(n, s, d)$ is that in the former d_n represents a *decision variable* to be selected from $D(n, s)$, whereas in the latter the value of d_n is set to some *arbitrary value* $d \in D(n, s)$. It follows then that

$$f_n(s) = f_n(s, d), \text{ for some } d \in D(n, s), \quad (18)$$

for all $n \in \mathcal{N}$, $s \in S(n)$. Hence,

Theorem 1 [Principle of Conditional Optimization, 7].

$$f_n(s) = \underset{d \in D(n, s)}{\text{opt}} f_n(s, d) \quad \forall n \in \mathcal{N}, s \in S(n). \quad (19)$$

The significance of this obvious result is that it enables linking *Modified Problem* $P(n, s)$ to *Modified Problem* $P(n+1, T(n, s, d))$ via *Conditional Problem* $P(n, s, d)$. As suggested by the *metarecipe*, this link is the DP functional equation.

Principle of Optimality

This principle was formulated by Bellman [2, p. 83] as follows:

Principle of Optimality. An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

In the context of our analysis, it can be stated formally as follows:

Theorem 2 [Principle of Optimality]. For any triplet (n, s, d) such that $1 \leq n < k$, $s \in S(n)$, $d \in D(n, s)$, the set of optimal solutions to *Conditional Problem* $P(n, s, d)$ is equal to the set of optimal solutions to *Modified Problem* $P(n+1, T(n, s, d))$. That is,

$$X^*(n, s, d) = X^*(n+1, T(n, s, d)), \quad \forall 1 \leq n < k, \\ s \in S(n), d \in D(n, s). \quad (20)$$

The *Principle of Optimality* asserts that any optimal solution to *conditional problem* at (n, s, d) is also optimal for the *Modified Problems* that it generates, namely, *Modified Problem* $P(n+1, T(n, s, d))$, and vice versa. Thus, insofar as the optimal values of the (remaining) decision variables $(d_{n+1}, \dots, d_k)$ are concerned, *Conditional Problem* $P(n, s, d)$ is equivalent to the *Modified Problem* $P(n+1, T(n, s, d))$.

The following is a direct implication of the *Principle of Optimality*.

Theorem 3 [7].

$$f_n(s, d) = r(n, s, d) + f_{n+1}(T(n, s, d)), \\ \forall 1 \leq n < k, s \in S(n), d \in D(n, s). \quad (21)$$

The question is then, what is the best choice for d given (n, s) ? That is, what is the best choice for decision d if we are at stage n observing state s ? The answer is obvious: the best value of d is one that optimizes the right-hand side of Equation (21).

Functional Equation

The following is a direct implication of Theorem 1 and 3.

Theorem 4 [DP Functional Equation, 7].

$$f_n(s) = \underset{d \in D(n, s)}{\text{opt}} \{r(n, s, d) + f_{n+1}(T(n, s, d))\}, \\ \forall n \in \mathcal{N}, s \in S(n), \quad (22)$$

with $f_{k+1}(s) := 0$, $\forall s \in S(k+1)$.

It should be emphasized that Equation (22) is not a definition of $f_n(s)$. That is, $f_n(s)$ is defined in Equation (9). Rather, Theorem 4 spells out an important property of the functions $(f_n : n = 1, 2, \dots, k)$ defined in Equation (9).

Terms such as *functional equation*, *Bellman equation*, *optimality equation*, and *recurrence relation* are used to describe the relation stipulated by Theorem 4.

Policies

In the framework of DP, a *policy* is a *rule* governing the choice of decisions [2,7,11]. DP offers a number of such rules, but of special interest is the rule governing the choice of a decision which depends on the *stage and state* of the process. We refer to such policies as *Markovian policies*. So, in the context of the DP model described in the preceding section, a *Markovian policy* is a rule δ that assigns to each (stage, state) pair (n, s) a feasible decision $d \in D(n, s)$.

An application of a Markovian policy δ to a state s at stage n yields a state trajectory $(s_n, \dots, s_{k+1})$ and a sequence of decisions $(d_n, \dots, d_k)$ such that $s_n = s$, $d_j = \delta(j, s_j)$, and $s_{j+1} = T(j, s_j, d_j)$, $j = n, \dots, k$.

A *DP policy* is a Markovian policy that is a solution to the DP functional equation. In other words, a DP policy is a Markovian policy δ^* such that

$$\delta^*(n, s) \in D^*(n, s), \forall n \in \mathcal{N}, s \in S(n), \quad (23)$$

where $D^*(n, s)$ denotes the set of all the optimal solutions to the functional equation at (n, s) , namely,

$$D^*(n, s) := \{d^* \in D(n, s) : f_n(s, d^*) = f_n(s)\}, \quad n \in \mathcal{N}, s \in S(n), \quad (24)$$

recalling that $f_n(s, d) = r(n, s, d) + f_{n+1}(T(n, s, d))$.

All the DP policies are simultaneously optimum for all the *Modified Problems*, including the one corresponding to the *Target Problem*.

Solution Methods

There is no general-purpose recipe for the solution of DP functional equations. Still, it is possible to distinguish between two approaches that can be outlined broadly as follows:

Direct Methods [2,7,11]. These methods basically determine the value of $f_n(s)$ for each $n \in \mathcal{N}$ and $s \in S(n)$ by evaluating the right-hand side of the DP functional equation. In particular, if there are finitely many stages and states, the following procedure suggests itself.

Initialization: Set $n \leftarrow k + 1$ and $F_{k+1}(s) \leftarrow 0$, $\forall s \in S(k + 1)$.

Stage Iteration: For $n = k, k - 1, \dots, 1$ Do:

State Iteration: for $s \in S(n)$ Do:

$$F_n(s) \leftarrow \text{opt}_{d \in D(n, s)} \{r(n, s, d) + F_{n+1}(T(n, s, d))\} \quad (25)$$

End Do

End Do

There is likewise no “general-purpose” method for solving the optimization problem specified by the right-hand side of Equation (25). The method used is contingent on the structure of the objects involved.

If this optimization problem is solved by enumerating $d \in D(n, s)$, then in addition to the “stage loop” and the “state loop,” there is also an inner “decision loop.” Upon completion, we obtain $F_n(s) = f_n(s)$, $\forall n \in \mathcal{N}, s \in S(n)$.

Successive Approximation Methods [2,7,11]. These are iterative procedures that begin with an initial approximation of either f or a DP policy, and improve on it successively. That is, using the current approximation, the DP functional equation is solved for this approximation, to obtain a new approximation and so on. The procedure terminates when no further improvement is possible.

For instance, consider the following “stage-less” DP functional equation.

$$f(s) = \text{opt}_{d \in D(s)} \{r(s, d) + f(T(s, d))\}, s \in \mathbf{S}, \quad (26)$$

where function f is defined on $\mathbf{S}$ by a DP sequential decision model.

If the initial approximation of the unknown function f is set to $f^{(a)}$, then a basic successive approximation scheme would be as follows:

Initialization: For each $s \in \mathbf{S}$ set $F(s) \leftarrow f^{(a)}(s)$.

Iteration: For each $s \in \mathbf{S}$ set

$$F(s) \leftarrow \text{opt}_{d \in D(s)} \{r(s, d) + F(T(s, d))\} \quad (27)$$

Termination: Stop if there is no improvement in F .
Otherwise continue with the iteration.

An important factor in successive approximation schemes of this type is the order in which the states are enumerated in each iteration. And of course, certain technical “regularity” conditions must hold so as to guarantee the convergence of such successive approximation schemes and to ensure that the solutions obtained by the scheme is the function f defined by the DP model.

One of the most celebrated algorithms in computer science, namely, *Dijkstra’s Algorithm* for the shortest path problem, is a typical DP successive approximation method, albeit not as simple as the basic one outlined above [11,12].

COMPLETE ENUMERATION

One of the stories about Bellman’s work on DP is that one of his early papers was rejected from publication on the grounds that the proposed procedure (for the solution of the DP functional equation) was deemed a *complete enumeration* procedure in disguise. Indeed, this is how this procedure can appear to newcomers to the field.

It is important to stress, therefore, that except for some pathological cases, the solution of the DP functional equation clearly does not amount to a complete enumeration. The factor that is instrumental in “saving” the procedure from this “ignominy” is the *state variable*.

For example, consider the TSP. Recall that in this context a state is a pair $s = (c, C)$,

where c is a city and C is a set of cities. The fact that C represents all the subtours (partial solutions) consisting of $m = |C|$ cities selected from C (without replacement) means that typically a single state $s = (c, C)$ represents many subtours. Hence, the task of enumerating the states of a TSP is much less demanding than the task of enumerating all the subtours associated with the problem.

To illustrate, consider the case where $c = 3$ and $C = \{1, 2, 4, 7, 8, 9\}$. Since C consists of 6 cities, it represents $6! = 720$ partial solutions, namely permutations of the elements of C . This means that the “saving” is substantial, especially in large problems. Note that a TSP consisting of k cities has $k!$ feasible tours, whereas the state space of the DP model has less than $k2^{k+1}$ states, hence if k is large then, $k!$ is much larger than $k2^{k+1}$.

The trouble is, however, that for all its manifest superiority over complete enumeration in the solution of combinatorial optimization problems, there are many cases where the state space of the DP model grows very fast (e.g., exponentially) with the size of the problem—a point leading directly to our next topic.

THE CURSE OF DIMENSIONALITY

This graphic phrase was coined by Bellman [2, p. xii] to describe the difficulty hindering the solution of problems, whose complexity grows inordinately fast (e.g., exponentially) with their “size.”

There are many cases where the DP solution procedure is not affected by this “curse.” But there are many others where, for all the progress in DP, in computing, and so on, the “curse” presents as onerous a problem as it did in the early days of DP.

The fast (e.g., exponential) growth in the complexity of a problem with its “size” often means an inordinate growth in the *state space* of the DP model. Therefore the size of the *state space*, $\mathbf{S}$, is often a good indication of the complexity of DP algorithms.

For example, as pointed out above, in the case of the TSP, a problem of size k requires a state space consisting of more than 2^k states [8]. The situation is similar in many *scheduling problems* where the solution space

consists of *permutations* of a list of elements (jobs, projects) that grows linearly with the size of the problem [7]. Note that in these cases, there are $n!$ feasible solutions. The implication is then that the DP approach is obviously incomparably “more efficient” than *complete enumeration*. But, the fact remains that the DP functional equation would still be prohibitive, if the problem is large.

In short, the sad reality is that in many cases, it is impossible to obtain exact solutions to the DP functional equation due to the computational resources required for this purpose.

Given this state-of-affairs, for many years now the development of effective “heuristic” methods has been the subject of intensive research in DP. The appeal of such heuristic methods is that they are fast and can handle large problems. The price tag attached, though, is that such algorithms do not guarantee optimal solutions for the *Target Problem*. Worse, often such methods do not even provide an indication as to how “good/bad” the solutions they generate are in relation to the exact optimal solutions.

The fundamental difficulties encountered in providing such bounds attest to the insurmountable difficulties presented by the *curse of dimensionality*. So, despite the tremendous progress that has been made in this area over the past fifty years, the *curse of dimensionality* remains unbeatable, a fact that, needless to say, provides grist to the mills of ongoing research into the methods, which aimed at coping with it [13–16].

PERSPECTIVE

The basic concepts examined in the preceding sections can be modified so as to enable the adaptations of the DP model and functional equation discussed above to a variety of more/less complex problems. In this section, we examine briefly a number of such modifications and other aspects of DP.

Generalizations

Nonadditive Objective Functions. The objective function of a DP model need not necessarily be *additive*. While the additive form

is no doubt the most prevalent by far, other objective functions can be easily handled by DP. For instance, consider the case where the objective function has the following generic form:

$$g(s_1, d_1, d_2, \dots, d_k) = r(1, s_1, d_1) \oplus r(2, s_2, d_2) \\ \oplus \dots \oplus r(k, s_k, d_k), \quad (28)$$

where $\oplus$ is an associative binary operator [7,17].

If $\oplus$ is nondecreasing with its right argument, then the *Principle of Optimality* holds, and the DP functional equation takes the familiar form

$$f_n(s) = \underset{d \in D(n,s)}{\text{opt}} \{r(n, s, d) \oplus f_{n+1}(T(n, s, d))\}, \\ \forall n \in \mathcal{N}, s \in S(n), \quad (29)$$

with $f_{k+1}(s) = R_{\oplus}, \forall s \in S(k+1)$, where $R_{\oplus}$ denotes the (right) *identity element* of $\oplus$, namely $a \oplus R_{\oplus} = a, \forall a$ [7,17].

Some of the nonadditive forms of $\oplus$ that are used in DP models are the *multiplicative form* ($\oplus = \times$), the binary *max form* ($a \oplus b = \max\{a, b\}$), the binary *min form* ($a \oplus b = \min\{a, b\}$), and the *right-identity form* ($a \oplus b = b$). The latter plays an important role in Bellman’s conception of DP [7]. Note that the multiplicative form requires the returns to be nonnegative.

It is important to note that if $a \oplus b$ is not increasing with b , then the functional equation (Eq. 29) may not be valid, in which case it may be necessary to expand the state space and modify the DP model [7].

Infinite-Horizon Models. DP models can have infinitely (countably) many decision stages ($k = \infty$). Such models are called *infinite-horizon models*, or *nontruncated models*. Often, such models require some sort of a *discounting* feature to ensure that the total return is bounded. Furthermore, some regularity conditions must hold to guarantee that the total returns induced by Markovian policies are well defined [2,7,17].

Stochastic Models. DP models can be modified to enable the treatment of *stochastic* decision processes where the next state of

the process depends on the current state, the current decision, and some probabilistic events [2,18,19]. This gives rise to *Markovian decision processes*.

From a DP perspective, such models are characterized by the property that the application of a decision at a given (stage, state) pair can generate more than one state. That is, $T(n, s, d)$ is not an element of $S(n+1)$, it is a *subset* of $S(n+1)$. As a result, each time a decision is made a number of diverging subprocesses are initiated. The objective is often to optimize the *expected value* of the total return/cost.

Nonserial Models. These are generalizations of stochastic models in the sense that they consists of a number of subprocesses that interact with one another in a variety of ways such as diverge (as in the case of stochastic processes), converge, feedbackward, and feed-forward [9,17,20,21].

Multiobjective Models. The objective function of a DP model need not necessarily be a real-valued function [22–24]. Thus, it can be a function taking values in $\mathbb{R}^m$, representing m objectives. In this case, the DP functional equation can be used to generate the *pareto* optimal solutions to the *Target Problem*.

Forward Decomposition. The *Modified Problems* defined by Equations (9)–(12) are based on objective functions that represent the remaining part of the decision-making process, that is the “cost-to-go.” Hence, by definition $f_n(s)$ represents the optimal return generated from stage n to the final decision stage, namely stage k . Such decomposition schemes are called *backward* decomposition schemes.

There are cases where it is advantageous to use *forward* decomposition schemes [6,11]. In this framework, *Modified Problem* $P(n, s)$ poses the question: what is the best way of reaching state s at stage n from a given initial state σ at stage 1? The *Target Problem* is equivalent to *Modified Problems* $P(k+1, s)$ for some given final state $s \in S(k+1)$.

The following is the *forward* counterpart of the *backward* DP functional equation given in Equation (22).

$$f_{n+1}(s') = \underset{\substack{s \in S(n) \\ d \in D(n,s) \\ s' = T(n,s,d)}}{\text{opt}} \{f_n(s) + r(n, s, d)\},$$

$$n \in \mathcal{N}, s' \in S(n+1), \quad (30)$$

with $f_1(\sigma) = 0$, where $f_{n+1}(s')$ denotes the optimal return of reaching state s' at stage $n+1$ from the initial state σ at stage 1.

There are of course other generalizations of the DP model outlined above. For instance, the above model can be modified so as to serve as a medium for solving problems that are not “genuine” optimization problems [7] (e.g., constraint satisfaction problems), or problems where some of the sets are fuzzy [25].

Software

As pointed out already, one of the peculiarities of DP is the lack of truly “general-purpose” DP software packages for the modeling, analysis, and solution of DP functional equations. This fact is a direct implication of the rather “abstract” methodological framework put forward by DP.

There are, of course, software packages for special applications of DP and for educational purposes. More on this important aspect of DP can be found in Refs 26–29.

SUMMARY

DP is a general, wide-ranging approach to problem solving which employs the idea of *problem decomposition* as a basic ploy. Traditionally, DP has been cast as a framework expressing a *sequential decision-making process* where the *state variable* is used to aggregate equivalent subproblems into a family of *Modified Problems*. The relationship between the optimal solution to these *Modified Problems* is described by a *functional equation* whose solution provides a *policy* that is optimal with respect to all the *Modified Problems*, including the one representing the *Target Problem*.

Numerous methods are available for the solution of DP functional equations. Nevertheless, there are cases where the equation eludes solution or is subject to the *curse of dimensionality*. The relentless effort to tackle

the computational aspects of DP—including the development of user-friendly, general-purpose DP software—continues.

REFERENCES

1. Bellman R. On the theory of dynamic programming. *Proc Natl Acad Sci U S A* 1952;38: 716–719.
2. Bellman R. *Dynamic programming*. Princeton (NJ): Princeton University Press; 1957.
3. Bellman R. *Eye of the Hurricane: an autobiography*. Singapore: World Scientific; 1984.
4. Dreyfus SE. Richard Bellman on the birth of dynamic programming. *Oper Res* 2002;50(1):48–51.
5. Dreyfus SE. IFORS operational research hall of fame: Richard Ernest Bellman. *Int Trans Oper Res* 2003;10:543–545.
6. Dreyfus SE, Law AM. *The art of dynamic programming*. New York: Academic Press; 1977.
7. Sniedovich M. *Dynamic programming*. New York: Marcel Dekker; 1992.
8. Balas E, Simonetti N. Linear time dynamic programming algorithms for new classes of restricted TSP's: a computational study. *INFORMS J Comput* 2001;13(1):56–75.
9. Nemhauser GL. *Introduction to dynamic programming*. New York: Wiley; 1966.
10. Smith DK. *Dynamic programming: a practical introduction*. Chichester: Ellis Horwood; 1991.
11. Denardo EV. *Dynamic programming*. Mineola (NY): Dover; 2003.
12. Sniedovich M. Dijkstra's algorithm: the dynamic programming connexion. *J Contr Cybern* 2006;35(3):599–620.
13. Bertsekas DP, Tsitsiklis JN. *Neuro-dynamic programming*. Nashua (NH): Athena Scientific; 1996.
14. Sniedovich M, Voß S. The Corridor method: a dynamic programming inspired metaheuristic. *J Contr Cybern* 2006;35(3):551–578.
15. Powell WB. *Approximate dynamic programming: solving the curses of dimensionality*. New York: Wiley; 2007.
16. Hartman JC, Perry TC. Approximating the solution of a dynamic, stochastic multiple knapsack problem. *Contr Cybern* 2006;35(3):535–550.
17. Denardo EV, Mitten LG. Elements of sequential decision processes. *J Ind Eng* 1967; 18:106–112.
18. Bertsekas DP. *Dynamic programming and stochastic control*. New York: Academic Press; 1976.
19. Puterman ML. *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley; 1994.
20. Bertele U, Brioschi F. *Nonserial dynamic programming*. New York: Academic Press; 1972.
21. Esogbue AO, Marks BR. Non-serial dynamic programming—a survey. *Oper Res Q* 1974;25: 253–265.
22. Brown TA, Strauch RE. Dynamic programming in multiplicative lattices. *J Math Anal Appl* 1965;12:364–370.
23. Mitten LG. Preference order dynamic programming. *Manag Sci* 1974;21:43–46.
24. Sobel JM. Ordinal dynamic programming. *Manag Sci* 1975;21:967–975.
25. Bellman R, Zadeh LA. Decision-making in a fuzzy environment. *Manag Sci* 1970; 17:141–164.
26. Raffensperger JF, Richard P. Implementing dynamic programs in spreadsheets. *INFORMS Trans Educ* 2005;5(2):25–46.
27. Sklenar J, Popela P. Generation of prototype dynamic programming solutions. In: *Proceedings of the 12th International Conference on Soft Computing*. 2006. pp. 157–162.
28. Lew A, Mauch H. *Dynamic programming: a computational tool*. New York: Springer; 2007.
29. Jensen PA, Bard JF. *Oper Res Mod Meth*. New York: Wiley; 2008.

DYNAMIC PROGRAMMING VIA LINEAR PROGRAMMING

İSMET ESRA BÜYÜKTAHTAKIN
Systems and Industrial Engineering,
University of Arizona, Tucson,
Arizona

RELATED LITERATURE

A wide variety of problems have been shown to be polynomially solvable via dynamic programming (DP) recursive formulations. The problem is attacked by decomposing it into a sequence of interrelated subproblems defined by a recursive function and starting to solve the smallest subproblem. The problem is then enlarged by finding the current optimal solution from the preceding subproblem until the original problem is solved in its entirety. The main strength of this approach is to offer treatments to different type of problems involving sequential decision making with discrete, nonlinear, and stochastic characteristics.

DP was introduced to solve Markov decision processes by Bellman [1]. The basics of a dynamic program can be described as follows: consider a system being observed over a finite or infinite time horizon divided into periods or stages. At each stage, a decision or an action updates the state to be observed at the next stage, and depending on the state and the decision made, an immediate reward (cost) is observed. The value function represents the expected total reward (cost) from the current stage through the end of the planning horizon, while the functional equation expresses the relationship between the value function at the present stage and the successive stage. Optimal decisions depending on stage and state are determined backwards, iteratively, by maximizing (minimizing) the right-hand side of the functional equation.

The use of linear programming (LP) to solve DP formulations was first introduced by

D'epenoux [2] and Manne [3]. Manne [3] studied an undiscounted Markov decision model with an infinite planning horizon. He specifically analyzed an inventory control problem where the “state” variable represents the initial stock level i and the “decision variable” corresponds to the order quantity j . In the linear program in Manne [3], i and j are introduced as subscripts to the variables x_{ij} . These variables x_{ij} represent the joint probabilities with which the initial stock equals i and the production quantity equals j . The steady-state probabilities of inventory, production, and shortage levels are then derived. The constraints of the LP are the requirements regarding steady-state probabilities, and the objective function is the minimization of the expected cost corresponding to the steady-state probabilities. D'epenoux [2] provided a linear program for the discounted version of the problem in Manne [3] by linearizing the functional equations of the corresponding dynamic program.

Howard [4] combined DP with Markov chain theory to develop Markov decision processes. He also contributed to the solution of infinite horizon problems by developing the policy iteration method as an alternative to the backward induction method of Bellman [1], which is known as *value iteration*. The policy iteration algorithm generates a sequence of stationary policies by evaluating and improving the policies until the optimal policy is obtained. Later, Osaki and Mine [5] formulated a semi-Markovian decision process as an LP and showed that the dual of this LP is equivalent to DP formulation of Howard [4].

Another group of DP and Markov decision processes research arises from the observation that finding the optimal cost function can be cast as a LP problem [3,6–11]. However, the resulting linear program suffers from the curse of dimensionality: as the problem size linearly increases, the size of the state space grows exponentially. This LP has as many decision variables as states in the Markov decision processes, and an

even greater number of constraints. This difficulty is addressed by approximate dynamic programming (ADP), where the DP optimal cost-to-go function of the problem is approximated within the span of some prespecified set of basis functions as introduced by Schweitzer and Seidmann [12]. de Farias and Van Roy analyzed and further developed this approach by proposing the procedure known as *approximate linear programming* (ALP), which reduces the dimensionality of the linear program by utilizing a linear combination of basis functions combined with sampling only a small subset of the constraints [13–15]. de Farias and Van Roy [14] also established strong approximation guarantees for ALP-based approximations assuming knowledge of a “Lyapunov”-like function that must be included in the basis. An extension to the ALP approach that automatically generates the basis functions in the linear framework was developed by Valenti *et al.* [16]. Later, Desai *et al.* [17] studied the ADP via a smoothed linear program and developed an error bound that characterizes the quality of approximations produced by ADP.

Eppen and Martin [18] and Martin *et al.* [19] studied the underlying DP network structure to reformulate difficult integer programs into new models that have better bounds and solve them more quickly. Eppen and Martin [18] provided tighter mixed integer programming formulations for the single and multi-item lot-sizing problems using a variable redefinition approach. They first removed the capacity constraints from the traditional lot-sizing formulation and represented the subproblem with the DP network structure. This shortest path network can be written as an integer linear program (IP), with the arcs corresponding to binary variables and the nodes corresponding to flow balance constraints. Eppen and Martin then related the variables of the traditional model to the new set of variables through a linear transformation. By using this transformation, they inserted the complicating constraints in terms of the new variables into the new formulation. Although this new reformulation had a greater number of variables and constraints, it had a tighter LP relaxation lower bound leading to reduced

solution times. Martin *et al.* [19] formulated polynomially solvable optimization problems as shortest path problems by using DP. They then represented the dynamic program as an LP having a polynomial number of variables and constraints. The extreme points of this LP are represented by the solution vectors of the DP, and the dual of the LP provides the DP formulation. They also showed that with an appropriate change of variables, the LP formulation obtained from the DP provides a polyhedral description of the model considered.

There are a number of studies that utilize DP algorithms to formalize and solve integer linear programming problems [20–25]. In a recent study, Hartman *et al.* [26] introduced a set of DP-based inequalities that can be used to augment the capacitated lot-sizing integer linear programming (CLSP) formulation. These authors utilized iterative solutions of forward DP formulations for CLSP to generate inequalities for an equivalent integer programming formulation. The inequalities capture convex and concave envelopes of intermediate-stage value functions, and can be lifted by examining potential state information at future stages. Lawler and Wood [27] discussed the close relationship between branch and bound and DP, while Morin and Marsten [28] studied branch and bound techniques to reduce storage and computational requirements in discrete DP. In particular, they utilized relaxations and fathoming criteria in branch and bound to eliminate states of DP, which will not lead to optimal policies.

The article is outlined as follows. In the second section, we review the close relationship between LP techniques and DP in stochastic control, while we discuss the utilization of DP driven acyclic decision graphs to describe polyhedral characteristics of discrete optimization problems in the third section. The fourth section concludes the article and offers directions for future research.

LP-DP RELATIONSHIP IN STOCHASTIC CONTROL

In this section, we consider a discrete-time stochastic control problem involving a finite

state space $S = (1, \dots, n)$. Let $\mathcal{U}$ be a finite decision set for all $i \in S$. Given state i , the use of decision $u \in \mathcal{U}$ specifies the transition probability $p_{ij}(u)$ to the next stage j , and a cost $g(i, u)$ is incurred when a decision u is taken in state i . Future costs are discounted by a factor $\alpha \in (0, 1)$. A policy of the stochastic control problem is denoted by $\mu : S \rightarrow \mathcal{U}$. The problem is then to minimize the so-called cost-to-go function $J_\mu : S \rightarrow \mathcal{U}$ over the set of admissible policies $\mathcal{P}$:

$$\min_{\mu \in \mathcal{P}} J_\mu(i_0) = \min_{\mu \in \mathcal{P}} \mathbb{E} \left[\sum_{k=0}^{\infty} \alpha^k g(i_k, \mu(i_k)) \right], \quad (1)$$

where $i_0 \in S$ is an initial state and the expectation $\mathbb{E}$ is taken over the possible future states $\{i_1, i_2, \dots\}$, given i_0 and the policy μ .

The optimal cost associated with the optimal policy μ^* , denoted by $J^* = J_{\mu^*}$, satisfies the Bellman equation:

$$J(i) = \min_{u \in \mathcal{U}} \left[g(i, u) + \alpha \sum_{j \in S} p_{ij}(u) J(j) \right], \quad \forall i \in S, \quad (2)$$

which is called value iteration and is a principal method for calculating the optimal cost function J^* [7]. Once J^* is found by solving Equation (2), the optimal policy μ^* can be computed.

An alternative approach to solving the stochastic control problem described here is to fix the policy μ and then solve the following linear system:

$$J_\mu(i) = g(i, \mu(i)) + \alpha \sum_{j \in S} p_{ij}(\mu(i)) J_\mu(j), \quad \forall i \in S. \quad (3)$$

Solving Equation (3) is called as *policy evaluation* yielding J_μ , the cost-to-go of the fixed policy μ . After computing the policy's cost-to-go function, a better policy can be constructed by performing a policy improvement step. The policy iteration method repeatedly performs policy evaluation followed by policy improvement. This procedure generates a sequence of policies that is guaranteed to converge to the optimal policy μ^* after a

finite number of iterations since the new policy must be strictly better than the previous policy and there is only a finite number of possible policies in total [4].

LP Formulation of DP

Suppose that we use value iteration to generate a sequence of vectors $J_k = (J_k(1), \dots, J_k(n))$ given an initial condition vector $J_0 = (J_0(1), \dots, J_0(n))$. Then the following constraints

$$J(i) \leq g(i, u) + \alpha \sum_{j \in S} p_{ij}(u) J(j), \quad \forall i \in S, u \in \mathcal{U}, \quad (4)$$

form a polyhedron in R^n . The optimal cost vector $J^* = (J^*(1), \dots, J^*(n))$ solves the following problem (in $w_1, \dots, w_n$)

LP1:

$$\max \sum_{i \in S} w_i \quad (5)$$

$$\text{subject to } w_i \leq g(i, u) + \alpha \sum_{j \in S} p_{ij}(u) w_j, \quad \forall i \in S, u \in \mathcal{U}. \quad (6)$$

This is a linear program with n variables and as many as $n \times q$ constraints, where q is the maximum number of elements of the set $\mathcal{U}$. For very large n and q , the linear program can be solved by the use of specialized, large-scale LP algorithms [7, 11].

Dual and Policy Iteration

We now investigate the dual of **LP1** which is shown to be the same as policy iteration [9]. Duality theory of LP [29] asserts that the following dual linear program:

Dual1:

$$\min \sum_{i \in S} \sum_{u \in \mathcal{U}} q(i, u) g(i, u) \quad (7)$$

$$\text{subject to: } \sum_{u \in \mathcal{U}} q(i, u) - \alpha \sum_{j \in S} \sum_{u \in \mathcal{U}} q(j, u) p_{ji}(u) = 1, \quad \forall i \in S, \quad (8)$$

$$q(i, u) \geq 0, \quad \forall i \in S, u \in \mathcal{U}, \quad (9)$$

has the same optimal value as **LP1**. The variables $q(i, u)$, $i \in S$, $u \in \mathcal{U}$, of the dual program can be interpreted as the steady-state probabilities that state i will be visited at the typical transition and control u will then be applied. The constraints of **Dual1** are the constraints that $q(i, u)$ must satisfy in order to be feasible steady-state probabilities. The cost function

$$\sum_{i \in S} \sum_{u \in \mathcal{U}} q(i, u) g(i, u)$$

is the steady-state average cost per transition.

Denardo [9] shows that the feasible bases for **Dual1** are in one-to-one correspondence with the policies. Denardo also proves that the application of the simplex method to the dual program performs the same sequence of pivots as does policy iteration and policy iteration is the same as multiple substitution (block pivoting) in the dual simplex method, as applied to **LP1**.

LP-DP RELATIONSHIP THROUGH ACYCLIC GRAPHS

Many discrete optimization problems that are solvable through DP can be represented by directed acyclic shortest path decision graphs [9,22,30,31]. Given a finite state set S , vertices of the graph correspond to states of S reflecting the transition phases in the solution of the underlying sequential decision process. Arcs (i, j) between vertices $i, j \in S$ represent decisions changing the state i to state $j > i$. Furthermore, each arc is assigned a cost equal to the length of the arc. The problem is then to find the shortest path from an initial state to a goal state. Generally, states are partitioned into stages such that the decision at a stage updates the state at the stage into the state for the next stage.

Once the shortest path formulation of a discrete optimization problem is constructed, the shortest path problem can easily be written as a linear program where variables represent arcs and the flow balance at each vertex is expressed as a constraint. This LP provides the polyhedral characterization of the discrete optimization problem where

every face of the associated polytope contains the incidence vector of a decision path in the DP graph. For any given cost function, one such incidence vector will be the LP optimal solution [19].

One problem with the acyclic shortest path paradigm is that it is inadequate for more complex discrete optimization problems since a typical decision involves composing two or more partial solution elements into a single element. To overcome this difficulty, Martin *et al.* [19] developed a directed acyclic decision hypergraph framework for deriving extended formulations of combinatorial optimization problems from DP algorithms as mentioned in the section titled "Related Literature". The DP algorithms considered in Martin *et al.* [19] search hyperpaths in a directed hypergraph $\mathcal{H} = (S, H)$ on a finite state space S with cardinality $|S| = n$, where the directed hyperarcs in $\mathcal{H}$ are of the form (K, j) with $j \in S$ and $K \subseteq S$. The states are ordered by a numbering function $\sigma : S \rightarrow \{1, 2, \dots, m \triangleq |S|\}$ such that $\sigma(i) < \sigma(j)$ for all $(K, j) \in H$ and $i \in K$. Furthermore, it is assumed that the directed hypergraph is single tailed. The set $S_T \subset S$ of boundary states is the set of all $j \in S$, for which there is a hyperarc $(\emptyset, j) \in H$ and all nonboundary states are called as *intermediate*. Finally, there must be a finite reference set $I(j) \neq \emptyset$ for each state $j \in S$ such that $I(i) \subseteq I(j)$ and $I(i) \cap I(i') = \emptyset$, for all $i, i' \in K$, where $i \neq i'$ is satisfied for all $(K, j) \in H$.

In this setting, Martin *et al.* [19] proved the convex hull of the characteristic vectors of decision paths in the DP hypergraph is described by the following linear program

$$\text{LP2:} \quad \min \sum_{(K,j) \in H} c[K, j] z[K, j] \quad (10)$$

$$\text{subject to:} \quad \sum_{(K,m) \in H} z[K, m] = 1, \quad (11)$$

$$\sum_{(K,\bar{j}) \in H} z[K, \bar{j}] - \sum_{(K,j) \in H \text{ with } \bar{j} \in K} z[K, j] = 0$$

$$\text{for } \sigma(\bar{j}) = 1, \dots, m-1, \quad (12)$$

$$z[K, j] \geq 0 \quad \forall (K, j) \in H, \quad (13)$$

where $z[K, j]$ is a binary (0 – 1) variable that takes value 1 if the hyperarc (K, j) is selected, and 0 otherwise, and $c[K, j]$ represents the cost of selecting the hyperarc (K, j) . The objective function (10) minimizes the cost of choosing a particular path, while constraint (11) ensures that exactly one decision will terminate at the global state m . Constraint (12) enforces flow balance conditions. Constraints (11) and (12) with the nonnegativity constraints (13) are shown to be sufficient to produce a binary solution to the primal linear program **LP2**.

LP2 can be viewed as the dual to the computations of the dynamic program itself. Thus, we provide the dual of the problem above as follows:

Dual2:

$$\max \omega \quad (14)$$

$$\text{subject to: } \omega - \sum_{i \in K} \lambda[i] \leq c[K, m],$$

$$\forall (K, m) \in H, \quad (15)$$

$$\lambda[j] - \sum_{i \in K} \lambda[i] \leq c[K, j],$$

$$\forall (K, j) \in H; j \neq m. \quad (16)$$

Here ω denotes the dual multiplier for constraint (11), and $\lambda[j]$ is the dual variable for row $\sigma(\bar{j})$ of Equation (12). Note that **Dual2** is equivalent to the DP formulation.

CONCLUSIONS AND FUTURE REMARKS

In this article we survey LP approaches for solving DP formulations of hard optimization problems. In particular, we analyze techniques used for casting a dynamic program as a linear program. An LP with few variables and a large number of constraints is often tractable via large-scale LP algorithms such as constraint generation. However, for the problems where DP requires an exponential number of states, the corresponding LP formulation consists of an exponential number of constraints. This difficulty necessitates research on new techniques to reduce the state space of DP by eliminating the states

and decisions that will not lead to the optimal policy.

As each problem requires a specific DP formulation and a program to be solved, LP provides an advantage of easily writing and solving tractable DP models using commercial software. Furthermore LP enables us to use sensitivity analysis for dynamic programs. LP sensitivity analysis and its relation to the states and decisions of the dynamic program could be further investigated. Another direction for research is utilization of LP formulation of DP in order to obtain stronger IP formulations in discrete optimization. This article is closely related to the following articles in the encyclopaedia are **Approximate Dynamic Programming I: Modeling; Approximate Dynamic Programming—II: Algorithms; Computation and Dynamic Programming and Dynamic Programming, Control, and Computation**.

REFERENCES

1. Bellman R. A Markovian decision process. J Math Mech 1957;6:679–684.
2. D’epenoux F. A probabilistic production and inventory problem. Manage Sci 1963;10:98–108. (Translation of an article published in Revue Frangaise de Recherche Operationnelle 14, 1960).
3. Manne AS. Linear programming and sequential decisions. Oper Res 1960;6(3):259–267.
4. Howard RA. Dynamic programming and Markov processes. Cambridge (MA): MIT Press; 1960.
5. Osaki S, Mine H. Linear programming algorithms for semi-Markovian decision processes. J Math Anal Appl 1968;22:356–381.
6. Bertsekas DP. Volume 1, Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2005.
7. Bertsekas DP. Volume 2, Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2007.
8. Denardo EV. On linear programming in a Markov decision problem. Manage Sci 1970;16(5):282–288.
9. Denardo EV. Dynamic Programming: Models and Applications. 2nd ed. Englewood Cliffs (NJ) Prentice-Hall; 2003.

10. Hordijk A, Kallenberg LCM. Linear programming and Markov decision chains. *Manage Sci* 1979;25(4):352–362.
11. Trick MA, Zin SE. A linear programming approach to solving stochastic dynamic programs. 1993. Unpublished.
12. Schweitzer P, Seidmann A. Generalized polynomial approximations in Markovian decision processes. *J Math Anal Appl* 1985;110(2): 568–582.
13. de Farias DP. The linear programming approach to approximate dynamic programming: theory and application [PhD. thesis]. Stanford University; 2002.
14. de Farias DP, Van Roy B. The linear programming approach to approximate dynamic programming. *Oper Res* 2003;51(6):850–865.
15. de Farias DP, Van Roy B. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math Oper Res* 2004;29(3):462–478.
16. Valenti M, Bethke B, How J. Embedding Health Management into Mission Tasking for UAV Teams. American Controls Conference. New York: 2007.
17. Desai VV, Farias VF, Moallemi CC. Approximate dynamic programming via a smoothed approximate linear program. Working paper. 2009.
18. Eppen GD, Martin RK. Solving multi-item capacitated lot sizing problems using variable redefinition. *Oper Res* 1987;35(6):832–848.
19. Martin RK, Rardin RL, Campbell BA. Polyhedral characterization of discrete dynamic programming. *Oper Res* 1990;38(1):127–138.
20. Bellman R. On a routing problem. *Q Appl Math* 1958;16:87–90.
21. Elmaghraby SE. The concept of state in discrete dynamic programming. *J Math Anal Appl* 1970;29(3):523–557.
22. Karp RM, Held M. Finite state processes and dynamic programming. *SIAM J Appl Math* 1967;15:693–718.
23. Nemhauser G. Introduction to dynamic programming. New York (NY): John Wiley and Sons; 1966.
24. Shapiro JF. Dynamic programming algorithms for the integer programming problem-I: the integer programming problem viewed as a knapsack type problem. *Oper Res* 1968;16(1):103–121.
25. Wolsey LA. Generalized dynamic programming methods in integer programming. *Math Program* 1973;4(1):222–232.
26. Hartman JC, Büyüktaktakın IE, Smith JC. Dynamic programming based inequalities for the capacitated lot-sizing problem. *IIE Trans* 2010. In press.
27. Lawler EL, Wood DE. Branch-and-bound methods: a survey. *Oper Res* 1966;14(4): 699–719.
28. Morin TL, Marsten RE. Branch-and-bound strategies for dynamic programming. *Oper Res* 1976;24(4):611–627.
29. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
30. Ibaraki T. Solvable classes of discrete dynamic programming. *J Math Anal Appl* 1973;43: 642–693.
31. Morin TL. Monotonicity and the principle of optimality. *J Math Anal Appl* 1982;86: 665–674.

DYNAMIC PROGRAMMING, CONTROL, AND COMPUTATION

FLOYD B. HANSON
Department of Mathematics,
Statistics, and Computer Science,
University of Illinois at Chicago,
Chicago, Illinois

Department of Mathematics,
University of Chicago, Chicago,
Illinois

DYNAMIC PROGRAMMING IN CONTINUOUS-TIME

Dynamic programming in the application areas of operations research often arise from optimal objectives such as to minimize costs, maximize profits, or maximize the utility of wealth and its consumption. These problems are invariably formulated in an uncertain environment since deterministic problems rarely occur without some sort of random perturbation. Hence, the optimization will necessarily be over an expected or mean value, else an optimum will not be well-posed. Assuming that the perturbing noise is Markovian, then Bellman's [1] principle of optimality fits both stochastic and deterministic environments, with just some extra analysis in the stochastic case. As a result of taking expectation before optimization, the equation of dynamic programming is a deterministic partial integro-differential equation (PIDE). See Hanson [2,3] or Kushner and Dupuis [4] for more details.

Assuming that the objective is a minimal expected cost problem and that the minimum is unique, then the minimum cost [3], starting at any time t in the horizon $[t_0, t_f]$, is

$$v^*(\mathbf{x}, t) \equiv \min_{\mathbf{U}(t, t_f)} \left[\mathbb{E}_{(W, P)(t, t_f)} \right]$$

$$\times \left[V[\mathbf{X}, \mathbf{U}, t_f](\mathbf{x}, t) \middle| \mathbf{X}(t) = \mathbf{x}, \right. \\ \left. \mathbf{U}(t) = \mathbf{u} \right],$$

$$V[\mathbf{X}, \mathbf{U}, t_f](\mathbf{x}_0, t) = \int_t^{t_f} C(\mathbf{X}(s), \mathbf{U}(s), s) ds \\ + Z(\mathbf{X}(t_f), t_f), \quad (1)$$

where the value of the total cost on $(t, t_f]$ is based on cumulative running cost $C(\mathbf{x}, \mathbf{u}, t)$ and final cost $Z(\mathbf{x}, t_f)$ functions. The $\mathbf{X}(t)$ is the state n_x -vector process on state domain $\mathcal{D}_x$, $\mathbf{U}(t)$ is the control n_u -vector process or optimizing variable on control domain $\mathcal{D}_u$. The set $\{\mathbf{W}(t), \mathbf{P}(t)\}$ comprise fundamental Markov diffusion and jump processes. The diffusion component characterizes the central continuous part and the jump part characterizes the large random, instantaneous, discontinuous rare events, since the diffusion part is insufficient for modeling catastrophic risk events. Obviously, the final optimal value is given by $v^*(\mathbf{x}, t_f) = Z(\mathbf{x}, t_f)$, for any $\mathbf{x} \in \mathcal{D}_x$ from setting $t = t_f$ in Equation (1).

The stochastic dynamic constraint is the jump-diffusion stochastic differential equation (SDE),

$$d\mathbf{X}(t) = \mathbf{f}(\mathbf{X}(t), \mathbf{U}(t), t) dt \\ + g(\mathbf{X}(t), \mathbf{U}(t), t) d\mathbf{W}(t) + d\mathbf{J}(t), \\ \mathbf{J}(t) = \int_0^t \int_{\mathcal{D}_q} h(\mathbf{X}(t), \mathbf{U}(t), t, \mathbf{q}) \mathcal{P}(d\mathbf{t}, d\mathbf{q}), \\ \mathbf{P}(t) = \int_0^t \int_{\mathcal{D}_q} \mathcal{P}(d\mathbf{t}, d\mathbf{q}), \quad (2)$$

when $t_0 \leq t \leq t_f$ subject to a given initial state $\mathbf{X}(t_0) = \mathbf{x}_0$, $d\mathbf{W}(t)$ is the n_w -dimensional differential diffusion or Wiener process, $\mathcal{P}(d\mathbf{t}, d\mathbf{q})$ is the compound n_p -dimensional Poisson random measure that introduces the Poisson random jump times for the time infinitesimal measure $d\mathbf{t} = (t, t + dt]$ and underlying

n_p -dimensional random jump amplitude variables in $d\mathbf{q} = (\mathbf{q}, \mathbf{q} + d\mathbf{q})$ when there is a jump. The diffusion process is normally distributed with zero infinitesimal mean $E[d\mathbf{W}(t)] = \mathbf{0}$, but correlation between components is allowed, $\text{Cov}[d\mathbf{W}(t), d\mathbf{W}^\top(t)] = [\rho_{ij}(t)]_{n_w \times n_w} dt$, such that $\rho_{i,i}(t) = 1$. The usual Poisson process is denoted by $\mathbf{P}(t)$, such that it is Poisson distributed with vector jump intensity $\lambda(t)$, infinitesimal moments $E[d\mathbf{P}(t)] = \lambda(t)dt = [\lambda_i(t)]_{n_p \times 1} dt$, $\text{Cov}[d\mathbf{P}(t), d\mathbf{P}^\top(t)] = [\lambda_i(t)\delta_{ij}]_{n_p \times n_p} dt$ and $E[\mathcal{P}_j(d\mathbf{t}, d\mathbf{q})] = \lambda_j(t)dt\phi_{Q_j}(q_j)dq_j$, assuming no correlations between Poisson component processes, since simultaneous jumps are unlikely. The underlying jump amplitude random variables Q_j are IID for each component $j = 1:n_p$ with density $\phi_{Q_j}(q_j)$. Similarly, there are no correlations between Poisson and Wiener processes. The Poisson jumps are by definition instantaneous, so at the instant of a jump, there is zero time for continuous change and hence the discontinuous changes are being calculated independent of continuous changes. At the k th jump of j th Poisson component $P_j(t)$ at random time $T_{j,k}$ with random mark $Q_{j,k}$, $\mathbf{X}(t)$ jumps by the amplitude

$$\begin{aligned} \mathbf{X}(T_{j,k}) - \mathbf{X}(T_{j,k}^-) &= \hat{\mathbf{h}}_j(\mathbf{X}(T_{j,k}^-), \mathbf{U}(T_{j,k}^-), T_{j,k}^-, q_{j,k}) \\ &\equiv \left[h_{ij}(\mathbf{X}(T_{j,k}^-), \mathbf{U}(T_{j,k}^-), \right. \\ &\quad \left. T_{j,k}^-, q_{j,k}) \right]_{n_x \times 1}, \end{aligned}$$

where the function $h(\mathbf{x}, \mathbf{U}, t, \mathbf{q})$ is the $n_x \times n_p$ jump amplitude coefficient array. Like for diffusions, the jump-diffusion approximation consistency conditions require that the increment version of $d\mathbf{X}(t)$ in Equation (2) have conditional infinitesimal mean and variance that are $E[\Delta\mathbf{X}(t)|\mathbf{X}(t)] = O(\Delta t)$ and $\text{Var}[\Delta\mathbf{X}(t)|\mathbf{X}(t)] = O(\Delta t)$. The drift or mean appreciation coefficient function is $\mathbf{f}(\mathbf{x}, \mathbf{U}, t)$ and the diffusion coefficient function is $\mathbf{g}(\mathbf{x}, \mathbf{U}, t)$, both, along with $h(\mathbf{x}, \mathbf{U}, t, \mathbf{q})$, are assumed to be bounded for computational feasibility and commensurate in multiplication with respect to the other variables. Stronger existence results are available assuming Lipschitz continuity, but some important models such as for stochastic volatility are non-Lipschitzian. For pure

deterministic problem, set $g \equiv 0$ and $h \equiv 0$, while for diffusion only noise set $h \equiv 0$.

Assuming that minimization and expectation operators are decomposable into factors, then Bellman's principle of optimality [3] is

$$\begin{aligned} v^*(\mathbf{x}, t) &= \min_{\mathbf{U}(t, t+\delta t)} \left[\begin{aligned} &E_{(\mathbf{W}, \mathbf{P})(t, t+\delta t)} \\ &\times \left[\int_t^{t+\delta t} C(\mathbf{X}(s), \mathbf{U}(s), s) ds \right. \\ &\left. + v^*(\mathbf{X}(t+\delta t), t+\delta t) \mid \mathbf{X}(t) = \mathbf{x}, \right. \\ &\quad \left. \mathbf{U}(t) = \mathbf{u} \right] \end{aligned} \right], \end{aligned} \quad (3)$$

given some small time step δt , leads in the limit to the Hamilton–Jacobi–Bellman (HJB) equation of stochastic dynamic programming (SDP), keeping only δt -terms,

$$\begin{aligned} 0 &= v_t^*(\mathbf{x}, t) + \min_{\mathbf{u}} [\mathcal{H}(\mathbf{x}, \mathbf{u}, t)] \\ &\equiv v_t^*(\mathbf{x}, t) + \mathcal{H}^*(\mathbf{x}, t), \end{aligned} \quad (4)$$

where the *Hamiltonian* functional is given by

$$\begin{aligned} \mathcal{H}(\mathbf{x}, \mathbf{u}, t) &\equiv C(\mathbf{x}, \mathbf{u}, t) + \nabla_{\mathbf{x}}^\top [v^*](\mathbf{x}, t) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \\ &\quad + \frac{1}{2} \text{Tr} [(gRg^\top)(\mathbf{x}, \mathbf{u}, t), \\ &\quad \nabla_{\mathbf{x}} [\nabla_{\mathbf{x}}^\top [v^*]](\mathbf{x}, t)] \\ &\quad + \sum_{j=1}^{n_p} \lambda_j(t) \int_{\mathcal{D}_q} [v^*(\mathbf{x} + \hat{\mathbf{h}}_j \\ &\quad \times (\mathbf{x}, \mathbf{u}, t, q_j), t) - v^*(\mathbf{x}, t)] \\ &\quad \times \phi_{Q_j}(q_j) dq_j, \\ R = R(t) &\equiv [\delta_{ij} + \rho_{ij}(t)(1 - \delta_{ij})]_{n_w \times n_w}, \\ \text{Tr}[A, B] &\equiv \sum_{i=1}^n \sum_{j=1}^n A_{ij} B_{ij}. \end{aligned} \quad (5)$$

Along with appropriate boundary conditions, the function completes the formal specification of the boundary-final value problem, the final value condition, and the placement of the time derivative implying that the dynamic programming problem is backward in time. Thus, the HJB equation (4) is, in general, a functional PIDE, so possesses

properties beyond the usual partial differential equations (PDEs) in that there is global dependence due to the integral properties whereas the partial derivatives impart only local dependence. In addition, the minimization can introduce nonlinear behavior, as in the case of the often assumed quadratic control costs. These and other properties lead to many computational problems beyond the usual linear PDE.

For specification of the minimization of the Hamiltonian in Equation (5), unique minimum and continuous control derivatives are up to second-order such that the Hessian is positive-definite; that is, $\nabla_u[\nabla_u^\top[\mathcal{H}]](\mathbf{x}, \mathbf{u}, t) > 0$ are assumed. The critical condition $\nabla_u[\mathcal{H}](\mathbf{x}, \mathbf{u}, t) = \mathbf{0}$ yields the unconstrained minimal or regular control $\mathbf{u}^{(\text{reg})}(\mathbf{x}, t) = \text{argmin}_{\mathbf{u}}[\mathcal{H}(\mathbf{x}, \mathbf{u}, t)]$, the primary in purely mathematical problems, but a preliminary to the constrained optimal control in applications. Real problems have constraints, so assuming hypercube constraints for simplicity, that the constrained optimal control $\mathbf{u}^*(\mathbf{x}, t)$ satisfies the constraints $U_i^{(\min)} \leq u_i^*(\mathbf{x}, t) \leq U_i^{(\max)}$ for $i = 1:n_u$ components. Thus,

$$\mathbf{u}^*(\mathbf{x}, t) = \left[\min \left[\max \left[u_i^{(\text{reg})}(\mathbf{x}, t), U_i^{(\min)} \right], U_i^{(\max)} \right] \right]_{n_u \times 1}. \quad (6)$$

COMPUTATIONAL REDUCING CANONICAL FORMS

There are some simplifications that can be used to make the computation of the optimal control simpler. An example is to use a linear drift dynamics and quadratic control cost model; that is, a linear quadratic (LQ) model in control only [3], so that

$$\begin{aligned} \mathbf{f}(\mathbf{x}, \mathbf{u}, t) &= \mathbf{f}_0(\mathbf{x}, t) + \mathbf{f}_1(\mathbf{x}, t)\mathbf{u}, \\ g(\mathbf{x}, \mathbf{u}, t) &= g_0(\mathbf{x}, t), h(\mathbf{x}, \mathbf{u}, t, q) = h_0(\mathbf{x}, t, q), \\ C(\mathbf{x}, \mathbf{u}, t) &= c_0(\mathbf{x}, t) + c_1(\mathbf{x}, t)\mathbf{u} + \mathbf{u}^\top c_2(\mathbf{x}, t)\mathbf{u}, \\ \mathcal{H}(\mathbf{x}, \mathbf{u}, t) &= \mathcal{H}_0(\mathbf{x}, t) + \mathcal{H}_1(\mathbf{x}, t)\mathbf{u} \\ &\quad + \frac{1}{2}\mathbf{u}^\top \mathcal{H}_2(\mathbf{x}, t)\mathbf{u}, \end{aligned} \quad (7)$$

where all functions are assumed to be specified and that c_2 is positive definite. Since c_2 and $\mathcal{H}_2$ appear only in quadratic forms, they might as well be symmetric, then the regular optimal control is explicit,

$$\begin{aligned} \mathbf{u}^{(\text{reg})}(\mathbf{x}, t) &= -c_2^{-1}(\mathbf{x}, t)(c_1^\top(\mathbf{x}, t) \\ &\quad + f_1^\top(\mathbf{x}, t)\nabla_x[v^*](\mathbf{x}, t)), \end{aligned} \quad (8)$$

and the optimal control $\mathbf{u}^*(\mathbf{x}, t)$ is still given by the formula in Equation (6).

In the classical LQ case, that is LQ in both control and state spaces, both the Hamiltonian and the optimal value v^* are quadratics, $\mathcal{H}$ in both control and state, while v^* can be shown to be a quadratic in state alone, so that the solution for v^* involves solving ODE final value problems only for coefficient functions of time t (see Refs 3 and 5 for the well-known details).

FINITE DIFFERENCE PDE COMPUTATIONAL METHODS

Finite differences have the advantage of straightforward application, but with a significant difficulty in convergence if the ratio of the time to space meshes is not sufficiently small. The first step is the discretization, where the time $t \in [t_0, t_f]$ is partitioned into N_t discrete backward time nodes and assuming for simplicity that the spatial variable is one-dimensional ($n_x = 1$) for $x \in [x_0, x_f]$,

$$\begin{aligned} T_k &= t_f - (k - 1) \cdot \Delta T \quad \text{for } k = 1:N_t, \\ &\quad \text{with } \Delta T \equiv (t_f - t_0)/(N_t - 1), \\ X_j &= x_0 + (j - 1) \cdot \Delta x \quad \text{for } j = 1:N_x, \\ &\quad \text{with } \Delta X \equiv (x_f - x_0)/(N_x - 1). \end{aligned}$$

Next central finite differences (CFD) are used for second-order accuracy for spatial derivatives included in the Crank–Nicolson implicit (CNI) method along with half-steps in time to obtain a simple second-order accurate form for the time derivative,

$$\begin{aligned}
v^*(X_j, T_k) &\rightarrow V_{j,k}, \\
v_i^*(X_j, T_{k+0.5}) &\rightarrow (V_{j,k+1} - V_{j,k})/(-\Delta T), \\
v_x^*(X_j, T_k) &\rightarrow \Delta DV_{j,k} = 0.5(V_{j+1,k} - V_{j-1,k})/\Delta x, \\
v_{xx}^*(X_j, T_k) &\rightarrow DDV_{j,k} = (V_{j+1,k} - 2V_{j,k} + V_{j-1,k})/(\Delta x)^2, \\
u^{(\text{reg})}(X_j, T_k) &\rightarrow UR_{j,k} = -(C_{1,j,k} + F_{1,j,k} \Delta DV_{j,k})/C_{2,j,k}, \\
u^*(X_j, T_k) &\rightarrow US_{j,k} = u_0(UR_{j,k}) = \min[\max[UMIN, UR_{j,k}], UMAX], \\
v^*(X_j + h_0(X_j, T_k, q), T_k) &\rightarrow VH_{j,k}(q) \& \Lambda_k = \lambda(T_k), \\
F_{i,j,k} = f_i(X_j, T_k) &\& C_{i,j,k} = c_i(X_j, T_k) \& G_{0,j,k} = g_0(X_j, T_k), \\
F_{j,k} = f(X_j, US_{j,k}, T_k) &\& C_{j,k} = c(X_j, US_{j,k}, T_k), \\
UMIN = U^{(\min)} &\& UMAX = U^{(\max)},
\end{aligned} \tag{9}$$

where the LQ assumption has been used for explicitness in the optimal control or otherwise different procedures will be needed [3]. The integral part for the jump amplitude distribution is probably not familiar and requires at least linear interpolation between a postjump position and nearest nodes, so

$$\begin{aligned}
IVH_{j,k} &\equiv \int_{\mathcal{D}_q} VH_{j,k}(q) \phi_Q(q) dq \\
&\simeq \sum_{i=1}^{N_q} w_i \cdot ((1 - \theta_i) V_{j+\ell_i,k} + \theta_i V_{j+\ell_i+1,k}),
\end{aligned} \tag{10}$$

where the $\{j + \ell_i, j + \ell_i + 1\}$ are the nearest-neighbor x -nodes corresponding to the postjump position $X_j + h_0(X_j, T_k, q_i)$ for $i = 1:N_q$ mark nodes, the θ_i are the corresponding linear interpolation weights and the w_i are the numerical integration weights assigned to the mark nodes q_i with respect to the density $\phi_Q(q)$. See Ref. 3 for additional information, procedures, and examples.

The basic CNI step approximates the midpoint in backward time by the average preserving second-order accuracy,

$$\begin{aligned}
V_{j,k+1} &\simeq V_{j,k} + \Delta T \cdot \mathcal{H}_{j,k+0.5} \simeq V_{j,k} \\
&\quad + 0.5 \Delta T \cdot (\mathcal{H}_{j,k} + \mathcal{H}_{j,k+1}), \\
\mathcal{H}_{j,k} &= C_{j,k} + F_{j,k} DV_{j,k} \\
&\quad + 0.5 G_{0,j,k}^2 DDV_{j,k} + \Lambda_k (IVH_{j,k} - V_{j,k}).
\end{aligned} \tag{11}$$

Although Crank–Nicolson is unconditionally stable and has significant algebraic simplifications for linear diffusions problems, the general complexity of the jump-diffusion

problem with quadratic costs here requires iterations to handle the nonlinearities and jump integrals in the optimal value function. An extrapolated predictor–corrector method has been found to be robust for a variety of applications [3]. An abbreviated pseudoalgorithm for the method is as follows:

1. Given constants: $\{N_t, N_x, N_q, N_\gamma, \text{tol}_v, t_f, \text{UMIN}, \text{UMAX}, x_0, x_f\}$;
2. Get $T_k \forall k$; Get $X_j, \forall j$;
3. Given functions: $\{c_0, c_1, c_2, f_0, f_1, g_0, h_0, \lambda_0, u_0, Z\}$;
4. Get $V_{j,1} = Z(X_j, t_f) \forall j$; Get $\Lambda_k = \lambda_0(T_k), \forall k$;
5. For $k = 1:N_t - 1$ do;
6. Extrapolation step: $V_{j,k+0.5}^{(1)} = V_{j,1}$; If $k > 1$, then $V_{j,k+0.5}^{(1)} = 0.5(3V_{j,k} - V_{j,k-1})$, endif; plus $\mathcal{H}_{j,k+0.5}^{(1)} \forall j$, using Equation (11);
7. Prediction step: $V_{j,k+1}^{(2)} = V_{j,k} + \Delta T \cdot \mathcal{H}_{j,k+0.5}^{(1)} \forall j$;
8. Predictor evaluation step: $V_{j,k+0.5}^{(2)} = 0.5(V_{j,k+1}^{(2)} + V_{j,k})$, plus $\mathcal{H}_{j,k+0.5}^{(2)} \forall j$, using Equation (11);
9. For $\gamma = 2:N_\gamma$ do;
10. Correction steps: $V_{j,k+1}^{(\gamma+1)} = V_{j,k} + \Delta T \cdot \mathcal{H}_{j,k+0.5}^{(\gamma)} \forall j$;
11. Corrector convergence check step given sufficient relative tolerance tol_v :

$$\text{If } \max_j [V_{j,k+1}^{(\gamma+1)} - V_{j,k+1}^{(\gamma)}] < \text{tol}_v$$

$$\cdot \max_j [V_{j,k+1}^{(\gamma)}]$$

then $k = k + 1, V_{j,k+1} = V_{j,k+1}^{(\gamma+1)}, \forall j$,
break from loop γ to return to step 6;
endif;

12. Corrector evaluation step: $V_{j,k+0.5}^{(\gamma+1)} = 0.5(V_{j,k+1}^{(\gamma+1)} + V_{j,k})$, plus $\mathcal{H}_{j,k+0.5}^{(\gamma+1)}$, $\forall j$, using Equation (11);
13. end loop γ ;
14. end loop k .

However, more discussion of stability and its effect on convergence is needed. Using the CFD for the first-order derivative in Equation (9) means that diffusion dominance has been tacitly assumed and the time step is selected by a mesh ratio condition, that is,

$$\min_{j,k} [G_{0,j,k}^2 - |F_{j,k}| \Delta x] \geq 0, \quad (12)$$

$$\Delta T < (\Delta x)^2 / \max_{j,k} [G_{0,j,k+0.5}^2].$$

If the jump-diffusion is not diffusion dominated, then it is advisable to use an unwinding finite difference for the first-order difference to correspond to the direction of the drift,

$$DV_{j,k} = (V_{j+s,k} - V_{j,k}) / (s\Delta x),$$

$$s = \text{sgn}(F_{j,k}), \quad \text{sgn}(0) \equiv 1, \quad (13)$$

even though this approximation is only first order in Δx . Second-order forward and backward finite differences of second-order in Δx can be used as they would be used for derivative boundary conditions [3] to preserve second-order accuracy.

SDP can become computationally demanding as the state dimension n_x and the common number of nodes per dimension N_x grow, as it does with PDEs, because the demand per time step grows exponentially with exponent $n_x \ln(N_x)$ and this growth is called *Bellman's Curse of Dimensionality* [3]. Advanced computing techniques like massively parallel and vector processor are needed to treat large dimensions [2].

PROBABILISTIC MARKOV CHAIN APPROXIMATION

The Markov chain approximation (MCA) [4,6] uses the discrete Markov chain process transition probabilities implied by finite differences to generate proper computational time

steps and weak convergence properties, developed by Kushner in the 1970s. This method, as in the previous section, is applicable to deterministic as well as stochastic processes of the Markov type.

Assuming the same optimization model 1 and stochastic dynamics (Eq. 2), but $n_x = 1$ and $n_u = 1$, the Bellman SDP equation (4) with Equation (5) and forward discrete-time value function $V_k(x) \simeq v^*(x, t_k)$ becomes

$$V_{k-1}(x) = V_k(x) + \Delta t_{k-1} \min_u [C_k(x, u) + F_k(x, u)V'_k(x) + \frac{1}{2}G_k^2(x)V''_k(x) + \Lambda_k(\text{IVH}_k(x) - V_k(x))], \quad (14)$$

for $k = 1:N_t - 1$, $\Delta t_{k-1} = t_k - t_{k-1}$, $t_1 = t_0$, $t_{N_t} = t_f$, and $V_1(x) = Z(x, t_f)$, using similar notation of the previous sections. Let ξ_k denote the k th stage of the Markov chain for $k = 1:N_1$ such that $\Delta \xi = \xi_{k+1} - \xi_k = O(\Delta x)$. The Markov chain transition probabilities must satisfy the proper conditions of non-negativity and conservation of probability, as well the jump-diffusion approximation $O(\Delta t_k)$ consistency conditions mentioned in the section titled "Dynamic Programming in Continuous-Time."

For the self or nearest-neighbor transition probabilities associated with the diffusion component processes and the finite difference grid with step ΔX using the central form for $v''_k(x)$ as in Equation (9) and the stabilizing upwinded form for $v'_k(x)$ as in Equation (13) with $s = \text{sgn}(f_k(x, u))$,

$$p_k^{(D)}(X, X|X = x, u_{k-1})$$

$$= 1 - \Delta t_{k-1} \left(g_k^2(x) + \Delta x |f_k(x, u_{k-1})| \right) / (\Delta x)^2,$$

$$p_k^{(D)}(X, X \pm \Delta x |X = x, u_{k-1})$$

$$= \Delta t_{k-1} \left(0.5g_k^2(x) + \Delta X \max[\pm f_k(x, u_{k-1})] \right) / (\Delta x)^2. \quad (15)$$

The global time interpolation mesh condition for diffusive convergence is

$$\Delta t_{k-1} \leq (\Delta x)^2 / \min_{x,u} \left[g_k^2(x) + \Delta x |f_k(x, u_{k-1})| \right]. \quad (16)$$

Provided that this time step is sufficiently small, then the well-known infinitesimal

Poisson transition probabilities, for $j = 0, 1, 2$ or more jumps, are given by

$$p_{j,k}^{(j)} = \begin{cases} 1 - \lambda_k \Delta t_{k-1} + o(\lambda_k \Delta t_{k-1}), & j = 0 \text{ jumps} \\ \lambda_k \Delta t_{k-1} + o(\lambda_k \Delta t_{k-1}), & j = 1 \text{ jump} \\ o(\lambda_k \Delta t_{k-1}), & j \geq 2 \text{ jumps,} \end{cases} = p_{0,k}^{(j)} \delta_{j,0} + p_{1,k}^{(j)} \delta_{j,1} + o(\lambda_k \Delta t_{k-1}), \quad (17)$$

more specifically for the mean jump count $\lambda_k \Delta t_{k-1} \ll 1$ and ignoring the $o(\lambda_k \Delta t_{k-1})$ yields the 0–1 Bernoulli process transition probability terms, $\{p_{0,k}^{(j)} = 1 - p_{1,k}^{(j)}, p_{1,k}^{(j)} = \lambda_k \Delta t_{k-1}\}$, that are consistent with the jump-diffusion approximation. Finally, adding the transition probability for the node-interpolated jump amplitude when there is a jump leading to this MCA dynamic programming computation formulation.

$$V_{k-1}^{(JD)}(x) \simeq \min_u \left[C_k(x, u) \Delta t_{k-1} + p_{0,k}^{(j)} \sum_{\ell=1}^3 p_k^{(D)} \right. \\ \times (x, x + (\ell - 2) \Delta x | x, u) V_k^{(JD)} \\ \left. \times (x + (\ell - 2) \Delta X) + p_{1,k}^{(j)} \widetilde{IVH}_k^{(JD)}(x, u) \right], \quad (18)$$

where $\widetilde{IVH}_k^{(JD)}$ is the node-interpolated version of the integral interpolation in Equation (10). For more details, see Kushner and Dupuis [4].

SUMMARY AND SOME OTHER APPROACHES

This article covers the fairly general computational SDP for Markov stochastic processes, that includes deterministic, diffusion, and jump-diffusion models as subsets. Due to space limitations, it has not been possible to cover some related topics. There have been recent results for convergence rates of computational dynamic programming in diffusion problems, most recently by Song and Yin [7], who also refer to prior convergence

rate work of others. Genuine stochastic difference equation models have not been covered, but Sennott [8] supplies a wealth of discrete applications and other kinds of objective functions. See also the section in Ref. 2 reviewing the discrete version of differential dynamic programming with alternative backward and forward successive approximations.

REFERENCES

1. Bellman RE. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
2. Hanson FB. Computational stochastic dynamic programming. In: Leondes CT, editor. Volume 76, Stochastic digital control system techniques, control and dynamic systems: advances in theory and applications. New York: Academic Press; 1996. pp. 103–162.
3. Hanson FB. Applied stochastic processes and control for jump–diffusions: modeling, analysis and computation, Series in advances in design and control Volume DC13. Philadelphia (PA): SIAM Books; 2007.
4. Kushner HJ, Dupuis PG. Numerical methods for stochastic control problems in continuous time. 2nd ed. New York: Springer; 2001.
5. Anderson BDO, Moore JB. Optimal control: linear quadratic methods. Englewood Cliffs (NJ): Prentice–Hall; 1990.
6. Kushner HJ. Numerical methods for stochastic control problems in continuous time. SIAM J Control Optim 1990;28:999–1048.
7. Song QS, Yin G. Rates of convergence of numerical methods for controlled regime-switching diffusions with stopping times in the costs. SIAM J Control Optim 2009;48:1831–1857.
8. Sennott LI. Stochastic dynamic programming and the control of queueing systems. New York: John Wiley & Sons, Inc.; 1999.

DYNAMIC VEHICLE ROUTING

BARRETT W. THOMAS

Department of Management Sciences,
Tippie College of Business,
University of Iowa, Iowa City, Iowa

This article addresses vehicle routing problems in which some problem information is dynamic. Dynamic information is information that becomes known over the problem horizon, where the problem horizon is the period of time during which customer visits can actually take place. For example, the volume of a customer's demand may be unknown until a vehicle visits the customer, or in some cases, who all of the customers are may be unknown until the vehicle is actually on the road. In many problem variants, there may be some knowledge of the dynamic information in advance, usually in the form of a probability distribution. For instance, while the volume of a customer's demand may be unknown until the customer is visited, its demand can be described with a particular probability distribution whose parameters can be estimated through historical data. Because the scope of routing problems is wide, this discussion is limited to problems in which a set of customers is served by a fleet of vehicles, potentially a fleet of a single vehicle, that may or may not be capacity constrained.

Recent years have seen numerous innovations in dynamic vehicle routing methodologies and applications. In part, this innovation has been driven by the fact that computing power and mobile communication technology have made real-world implementations a reality. Only 15 years ago, for example, it was common for drivers of local pickup and delivery routes to need to seek out pay phones in order to receive information about new requests for service. The result was that dispatchers were limited in the routing changes that could be made as new information became known. Now, with mobile phones ubiquitous, new information and routing changes can be communicated

instantaneously. Further, computing power has reached the point in both speed and availability that even small companies have the ability to employ decision-support technology. Finally, companies can combine the computing power with inexpensive data storage to capture and to exploit historical information in planning for the future.

Examples of the growing use of dynamic vehicle routing can be found in multiple industries and throughout the world. For example, Bieding *et al.* [1] describe a dynamic routing problem that emerges for a German newspaper publisher who needs to supply newspapers to subscribers who, for whatever reason, did not receive a paper via the regular morning delivery. du Plessis [2] uses dynamic routing to improve operations for a South Africa-based autocarrier. Erera *et al.* [3] developed a dynamic routing scheme for an Atlanta-based beverage distributor. Attanasio *et al.* [4] demonstrate how the combination of real-time information, handheld computers, and decision support improve the operations of a London-based courier service. Flatberg *et al.* [5] discuss the case of the local freight pickup and delivery operations of a Norwegian distribution company. Applications of dynamic vehicle routing are not confined to commercial deliveries. An excellent example is the routing of unmanned aerial vehicles in sensor network monitoring [6]. Outside of routing, readers may be interested in examples of dynamic problems in fleet management [7,8], multiperiod routing [9,10], ambulance deployment [11], inventory routing and dispatching [12–14], energy dispatching [15], scheduling [16,17], and military operations [18].

This article is organized around the solution methodologies for dynamic routing problems. Readers interested in an in-depth classification of dynamic vehicle routing problem variants are referred to Ref. 19 with a comprehensive taxonomy of vehicle routing problems found in Ref. 20. Larsen *et al.* [21] provide a review of the recent dynamic vehicle routing literature, particularly focused

on problem variants. The solution methodologies for dynamic vehicle routing problems can be characterized by two properties. First, we characterize the solution methodologies by the restrictions that they place on the structure of the solutions to their respective dynamic routing problems. Second, we can characterize the solutions by the extent to which they use known information.

Because of the complex nature of most dynamic routing problems, even modern computing is limited to only the smallest problems in its ability to find exact solutions. In particular, most exact solutions require identifying what action should be taken for any possible situation that might arise over the problem horizon. This set of actions is known as a *policy*. The number of situations may number in the billions, and for complex problems, the number of possible actions for a given situation could number in the millions. Often, the resulting optimal policy cannot even be stored in memory. Hence, most dynamic routing problems are solved via heuristic solution methods. A common distinction among these heuristics is what restrictions they place on the structure of the solutions. In essence, the computational challenge of computing an optimal policy is diminished by restricting the number of situations and actions that need to be considered. There are three general categories of restrictions: a priori, insertion, and unrestricted.

A *a priori* or fixed route policies rely on the use of a route or routes constructed in advance, before any dynamic information is revealed. Essentially, a priori restrictions simplify the solution of the dynamic problem by transforming it into a static one. To account for the realization of random events, a priori methods rely on recourse rules. For example, customers who do not require service during the problem horizon are skipped when reached on the a priori tour. A priori restrictions require that all possible customers are known in advance. *Insertion* restrictions develop a routing policy by inserting newly arriving service requests into an existing tour. In the case in which some subset of service requests is known in advance of the problem horizon, an initial tour is constructed, into which requests will

be inserted when the vehicle or vehicles is en route. Insertion-restricted methodologies can be described as real time. That is, they are designed to make decisions as the vehicle is en route and to make decisions for only those situations that are actually encountered. As the name implies, *unrestricted* routing policies do not impose, in advance, a particular structure on the final solution. Notably, no customer order is fixed at any point in the problem horizon.

While both the a priori and the insertion restrictions have some limitations in the problem variants to which they can be applied, neither is limited by a problem's objective function. In dynamic vehicle routing, the most common objective functions are either the minimization of some measure of routing cost or the maximization of customer service. The former is a natural extension of the travel cost—minimization objective common in deterministic routing problems. However, the service-maximization objective has become increasingly common. There are multiple reasons for the shifting objective functions. One reason is that the cost structure in dynamic vehicle routing, because it is often a local operation, is likely to have much higher costs associated with labor than with travel [22,23]. Further, these labor costs are essentially fixed. Hence, it is advantageous to spread the labor costs over as many customers as possible. Second, in dynamic environments, minimizing travel costs can lead to suboptimal performance in increasingly important service metrics [24]. This issue is particularly relevant in service businesses subject to long-term customer relationships. The ability to renew the relationships is often directly tied to the ability to meet customer expectations [25]. Failure to meet the expectations can result in the loss of business to competitors.

In this article, we consider two ways in which solution methods can account for problem information. For one, solutions can be constructed without regard to future information. We call such solution methodologies *myopic*. Contrarily, solutions can be constructed so as to account for the fact that the problem information is revealed over

time. We call such solution methodologies *anticipatory*.

The two-dimensional characterization of dynamic routing solution structures results in six categories. The remainder of this article will be devoted to discussing these in more detail. Each section covers one of the three solution structures with subsections for the information use. We discuss, in more detail, a priori strategies in the first section, and insertion-restricted solutions in the second section. The third section discusses unrestricted solution methods. This article concludes with a discussion of emerging challenges in the real-world implementation of dynamic routing solutions.

A PRIORI

An a priori route is one in which all customers are sequenced in advance of the active time horizon. Thus, a priori-restricted solutions are limited to problems in which all potential customers are known in advance. To compensate for realizations of the random events, prespecified recourse rules are used. Recourse rules are rules that specify what is to happen when a problem constraint is violated during the course of the problem horizon. An example is that a vehicle is required to return to the depot to unload if it reaches capacity before the end of the a priori route that it is following.

It is important to note that, while they are used as solutions to dynamic problems, a priori solutions are themselves static. That is, the sequence of customers visits remains the same regardless of what new information is learned over the problem horizon. This feature of the solution structure is what simplifies the solution process. In particular, the solution can be stored as a vector or matrix that is easily manipulated in a local-search procedure.

The quintessential example of an a priori routing problem is the a priori solution to the probabilistic traveling salesman problem (PTSP), first introduced in Ref. 26. In the PTSP, a set of customers must be sequenced keeping in mind that not every customer will need to be visited every day. The problem differs from the deterministic traveling

salesman problem in that there is a known probability for whether or not a customer will require service on a given day. In the case that a customer does not need to be serviced on a particular day, the recourse is to skip the customer when the tour is implemented.

As an example of the PTSP, consider a set of customers $\{1, 2, 3, 4, 5, 6, 7, 8\}$. An a priori solution orders the customers in advance, say $3 \rightarrow 8 \rightarrow 5 \rightarrow 7 \rightarrow 2 \rightarrow 6 \rightarrow 4 \rightarrow 1$. Suppose that on a given day, only customers 2, 4, 6, and 7 need service. In that case, the vehicle would serve the customers in the order $7 \rightarrow 2 \rightarrow 6 \rightarrow 4$, the order prescribed by the a priori solution but with customers 1, 3, 5, and 8 skipped.

In the case of both the literature on myopic and anticipatory a priori restrictions, rarely do authors present the problem as a dynamic problem. Rather, the problem description assumes the a priori solution structure. In some cases, perhaps because of managerial implications or technological limitations at the time of writing, such a restriction may truly describe the available solution set. As noted previously, however, information technology and computing power now allow for the possibility of less restricted solutions. Yet, regardless of whether or not the method was originally developed for a dynamic problem or not, existing a priori solution methods can still play a role in the solution of dynamic problems.

Myopic

A myopic a priori routing solution is one that assumes that all problem information is deterministic. For instance, consider our previous example of the PTSP. A myopic solution to the problem would assume, in determining the a priori tour, that all customers are going to be visited everyday. The problem then becomes the traveling salesman problem, although in implementation a recourse is still required. Jaillet [26] provides an example demonstrating the poor quality that can result from myopic solution to the PTSP. Campbell and Thomas [27] and Chen *et al.* [28] empirically demonstrate the outcome for variants of the PTSP.

The volume of deterministic routing literature is vast. For an introduction, the reader is referred to Refs 29, 30.

Anticipatory

The anticipatory a priori routing literature is voluminous. In addition to Ref. 26, foundational work in a priori routing can be found in Refs 31, 34. A review of the a priori routing literature can be found in Ref. 35. An important feature of the recent literature is that, even with modern computing power and an a priori restriction on solution structure, exact solutions are limited to small problems [36]. Recent work in a priori routing has explored the improvement of solution quality while still maintaining the basic structure of an a priori routing solution. For example, Ak and Erera [37] introduce a new recourse scheme for the stochastic vehicle routing problem that allows for two vehicles to share the work of each other's routes, thus taking advantage of variance pooling. Other recent work considers robust optimization (minimizing worst-case outcomes) in the context of a priori routing [38,39]. This work differs from much of the rest of the literature in which the focus is on minimizing expected costs.

INSERTION

Solution methodologies framed around insertion restrictions have a rich history in the routing literature. Insertion-restricted solutions maintain a route or routes throughout the problem horizon. As new service requests become known, the requests are inserted into one of the existing routes. The nature of the solution structure requires that insertion-restricted solutions be limited to problems in which customer service needs become known during the problem horizon.

As a concrete example, consider a single vehicle that must serve a set of customers $\{1, 2, 3, 4, 5, 6, 7, 8\}$. At some point in the problem horizon, some set of customers, say $\{2, 3, 5, 6, 7\}$, will have requested service. With an insertion-restricted solution, an order of the customers say $3 \rightarrow 5 \rightarrow 7 \rightarrow 2 \rightarrow 6$ would exist. Suppose that the next customer who requests

service is customer 8. In the case of an insertion-restricted solution, the relative order already existing for the customers ($3 \rightarrow 5 \rightarrow 7 \rightarrow 2 \rightarrow 6$) would be maintained, and customer 8 is routed according to some rule that inserts it into the existing order. The result is the tour $3 \rightarrow 8 \rightarrow 5 \rightarrow 7 \rightarrow 2 \rightarrow 6$.

In practice, the advantage of insertion-restricted solution approaches arises from the ability to quickly make decisions in real time. For example, consider minimum cost insertion (identifying the two customers in the tour between whom adding the new request will add the least additional travel distance). Computing the point of insertion can be done quickly and with minimal computation. Another advantage of the insertion technique is that it keeps preexisting routes relatively intact, altering them only to insert a newly arrived customer service requests. The importance of the maintenance of the relative order of customers is discussed further in the section titled "Operational Challenges."

For simplicity, in the following, we will refer to customers who are known before the problem horizon as advance-request customers and those who become known during the problem horizon as immediate-request customers.

Myopic

Not surprisingly, the earliest work using insertion strategies ignored information about potential future requests. Examples of early work can be found in Refs 40 and 41. In both cases, insertion techniques are applied to insert newly arrived customers into preexisting routes, without any consideration of future requests. Larsen *et al.* [42] consider a variety of insertion procedures for dynamic routing and show that, when the objective is to minimize total expected distance traveled, a minimum cost insertion policy performs well. Reviews of myopic insertion and other early dynamic routing literature can be found in Refs 43–45.

Anticipatory

Anticipatory insertion improves upon myopic insertion by including known information

about the future, particularly in the determination of where and when to wait and in the construction of the initial tours of known customers. As an example, Thomas [24] considers the case where the locations of both advance- and immediate-request customers are known in advance. For an insertion-restricted solution with one immediate-request customer, the author identifies exactly when and where the vehicle should wait to maximize the probability of serving the immediate-request customer should it request service. The author then extrapolates this special case to a the general case and shows that, when the percentage of immediate-request customers is 25% or less, the extrapolation works well. As more of the customers become immediate, the waiting heuristic of Ref. 46 performs well.

Just as when and where to wait is important in the performance of an insertion-based solution, so is the initial tour into which immediate requests are to be inserted. Using information about potential future events in the construction of the tour can be advantageous. For instance, Manni [47] explores the value of using information about future requests in the construction of the initial tour. Interestingly, for the problem studied, Manni [47] finds that a good initial tour can be constructed using only the locations of all potential customers, without the need for any probabilistic information. Branke *et al.* [48] use Monte Carlo sampling of possible future events in order to derive initial tours in their application. Sampling techniques have played a prominent role in the methods discussed in the section titled “Anticipatory” under the section titled “Unrestricted Policies”. An example of other work in anticipatory insertion includes Ref. 49.

UNRESTRICTED POLICIES

Unrestricted dynamic routing solutions are those in which no structure is placed on the solution. For instance, consider the example given at the beginning of the previous section. A single vehicle serves a set of customers $\{1, 2, 3, 4, 5, 6, 7, 8\}$, the set $\{2, 3, 5, 6, 7\}$ of which have request service

and have been ordered $3 \rightarrow 5 \rightarrow 7 \rightarrow 2 \rightarrow 6$. Again assume that the next service request is customer 8. For the case in which the solution structure is unrestricted, no regard needs to be given to maintaining some semblance of the current order of the tour. Thus, in choosing to serve customer 8, the new sequence of customers might become $7 \rightarrow 5 \rightarrow 8 \rightarrow 3 \rightarrow 6 \rightarrow 2$. Further, in the case of unrestricted solutions, other than as an algorithmic feature, such as discussed in Ref. 50, there is no need to maintain any order of customers. At each time that a decision needs to be made, only the next decision, in our example the decision to visit customer 7, needs to be known.

Without the restrictions on the solution structure, unrestricted solution methodologies offer the possibility of better objective values than those found with the restricted solution structures. At the same time, solution-unrestricted methods can be applied across all problem classes.

Myopic

In the case of unrestricted solutions, the myopic approaches can be considered reactive in that requests for service are considered only when they occur. Initial work in this area focuses on what are known as *reoptimization* strategies. However, it is important to note that, despite the use of reoptimization in the nomenclature, the updated solution may in fact result from the application of a heuristic [51–55] or a hybrid exact and heuristic approach [56,57]. Regardless, the idea is that, after each new arrival, these reoptimization approaches treat the dynamic problem as a static one. Historically, these myopic approaches followed naturally from the extensive existing literature for static routing problems. However, these myopic, unrestricted methods differ from a priori methods in that the solution is updated, albeit without regard to dynamic information that may become known in the future, when dynamic information is learned. In recent work, Mitrović-Minić *et al.* [58] consider a reoptimization approach in which the objective is modified to implicitly account for future requests. This work builds on Ref. 46,

in which waiting rules were combined with reoptimization to improve performance.

In the past, the heuristic component was often necessary in order to find solutions in real time. In recent work, however, Chen and Xu [59] use an exact column-generation-based approach to update their solution as new service requests arrive. Thus, for some applications at least, it is possible to solve exactly the static problem and provide real-time decision making.

Anticipatory

As in the case of the other anticipatory solution methodologies, anticipatory, unrestricted solution methods incorporate information about future events into the solutions. In this class of solutions, exact solution methods are often limited to very small problems. For example, Thomas and White [60] demonstrate the value of including such information into the solution process, but demonstrates exact solutions for problems with only seven customers.

Similar to the transition from myopic insertion-restricted solutions to the anticipatory version, a natural approach for an anticipatory, unrestricted solution method is to combine the reoptimization of the section titled “Myopic” under the section titled “Unrestricted Policies” with an anticipatory waiting strategy. An example of such an approach is given in Ref. 50. Secomandi and Margot [61] also consider a reoptimization approach, but uses a restricted model of the stochastic problem in order to have a tractable problem that incorporates information about future events.

An interesting feature of many of the unrestricted solution methods is that they often employ of a priori, insertion-structured solutions, or both in order to solve subproblems. In this way, the unrestricted solution method can take advantage of existing research. For example, Erera and Daganzo [62] consider a two-phase real-time solution algorithm in which the first phase implements a traditional a priori scheme. Then, at a given point in time, all of the system information is assessed and a new solution is determined that reflects the known data. References 63 and 64 iteratively apply an

anticipatory a priori routing solution method in a heuristic algorithm known generally as a rollout algorithm [65]. The a priori solution method is applied iteratively at each decision point to update the solution as new information becomes known. Novoa and Storer [66] demonstrate the value of more sophisticated a priori routing than those used in Refs 63 and 64. Because many a priori routing problems need to be solved at every decision epoch, one of the difficulties encountered in the iterative use of a priori routing schemes is that they can be computationally expensive. Goodson *et al.* [67] explore this issue and in particular demonstrates a modification to traditional rollout algorithms, which while still requiring the iterative solution of a priori problems, greatly reduces the number that need to be solved.

In recent years, a class of solution methods has emerged that construct solutions by sampling future events. The intuition is that using sampling to help determine the next decision will allow the solution to better accommodate future events. An example of such an algorithm is given in Ref. 68. In the algorithm proposed in Ref. 68, the authors use sampling to construct a set of potential routes containing existing customers as well as possible future customers. Using a prespecified measure, the algorithm chooses a “best” routing plan from these sampled routes, and the next decision becomes the next customer specified in the chosen routing plan. An interesting feature of the algorithm proposed in Ref. 68 as well as related work such as Refs 69–73 is that the problems resulting from the sampling procedures are deterministic problems, like those that emerge in the approaches discussed in the section titled “Myopic” under the section titled “Unrestricted Policies”. As a result, these sample-based methods can take advantage of the existing literature for deterministic problems.

An emerging solution technique is known as *approximate dynamic programming* (ADP). A key feature of many ADP methods is the use of off-line simulations to estimate parameters of value-function approximations (see Ref. 74 for more on ADP). Theoretically, with the appropriate approximation, such

an ADP algorithm can return an optimal solution but without the constraint of needing to store the entire optimal policy. While there has been limited application of ADP to dynamic vehicle routing, the method has shown great promise in other dynamic decision making applications [7,8,11,12,14,15,18]. While ADP approaches may incorporate a priori- or insertion-restricted solutions such as in Refs 75 and 76, a straightforward ADP implementation, similar to what is found in Ref. 63, would not require knowledge of such solution approaches. Yet, three of the four ADP-based approaches [75–77] of which the author is aware all cite difficulty with the size of the action space that emerges in dynamic vehicle routing problems. Overcoming this challenge will be an important advance as a promising solution methodology is applied to dynamic vehicle routing.

A special class of dynamic routing problems is known as the *dynamic traveling repairman problem* (DTRP) or the *on-line traveling salesman problem* (OTSP). In both cases, a server (vehicle) seeks to serve customers who become known over time. The two differ in their objectives. In the DTRP, the objective is to minimize the expected amount of time that a customer waits after requesting service until being served. The OTSP seeks to minimize the expected routing cost. In contrast to much of the literature discussed previously, the research on the DTRP and OTSP focuses on characterizing asymptotically optimal policies or algorithms with provable asymptotic bounds on the difference between the on-line solution and the full information case (the competitive ratio). A recent survey can be found in Ref. 78. Examples of research in this area include Refs 79, 83.

OPERATIONAL CHALLENGES

As new technology and new methodologies make it possible to solve problems with complex objectives without restricting the structure of the solution, it is important to recognize the role that the structure of the solution plays in operational implementations of solutions. Wong [84] discusses the

fact that, in the package-express industry, it is important that drivers serve the same customers at about the same time every day. This “consistency” has two implications. One, it helps drivers to improve customer service by allowing them to build relationships with customers whom they see regularly. Second, drivers are more efficient the more familiar they are with the area in which they are driving. References 85–88 consider these issues by developing multiple period models that explicitly include consistency or more generally customer service measures. Further, a number of authors discuss that, while not explicitly modeling for customer service, a priori routing solutions can offer some of the desired customer service benefits [35]. At the same time, authors have shown that a priori routes can be competitive with the myopic, unrestricted approaches discussed previously [89,90]. Recently, Ghiani *et al.* [91] showed that an anticipatory, insertion-restricted solution can offer comparable performance to an anticipatory, unrestricted solution found through a state-of-the-art algorithm. More research is necessary to thoroughly understand whether or not the existing restricted solution structures can offer comparable value in terms of customer service or if there is value in developing models that explicitly model for it. The advantage of the restricted solution structures is that we have experience with and methodologies for solving them.

REFERENCES

1. Bieding T, Gortz S, Klose A. On-line routing per mobile phone: a case on subsequent deliveries of newspapers. In: Bertazzi L, Speranza MG, A J, *et al.* editors. Volume 619, Innovations in distribution logistics, Lecture notes in economics and mathematical systems. Berlin: Springer; 2009. pp. 29–51.
2. du Plessis AJ. Determining tactical operations policies for an auto carrier using discrete-event simulation [Master's Thesis]. Department of Industrial Engineering, Stellenbosch University; 2008. Available at <http://etd.sun.ac.za/jspui/handle/10019/2026>. Accessed 2010 Jan 18.
3. Erera AL, Savelsbergh M, Uyar E. Fixed routes with backup vehicles for stochastic

- vehicle routing with time constraints. *Networks* 2009;54(4): 270–283.
4. Attanasio A, Bregman J, Ghiani G, *et al.* Real-time fleet management at ECourier LTD. In: Ziempeks V, Tarantilis CD, Giaglis GM, *et al.* editors. Volume 38, Dynamic fleet management: concepts, systems, algorithms, and case studies, Operations research/computer science interfaces. New York: Springer; 2007; pp. 41–63.
 5. Flatberg T, Hasle G, Kloster O, *et al.* Dynamic and stochastic vehicle routing in practice. In: Ziempeks V, Tarantilis CD, Giaglis GM, *et al.*, editors. Volume 38, Dynamic fleet management: concepts, systems, algorithms, and case studies, Operations research/computer science interfaces. New York: Springer; 2007; pp. 41–63.
 6. Flint M, Fernandez E, Kelton WD. Simulation analysis for UAV search algorithm design using approximate dynamic programming. *Mil Oper Res. Res* 2009;14:41–50.
 7. Simao HP, Day J, George AP, *et al.* An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transp Sci* 2009; 43(2): 178–197.
 8. Topaloglu H, Powell WB. Dynamic-programming approximations for stochastic time-staged integer multicommodity-flow problems. *INFORMS J Comput* 2006; 18(1): 31–41.
 9. Angelelli E, Bianchessi N, Mansini R, *et al.* Short term strategies for a dynamic multi-period routing problem. *Transp Res Part C* 2009;17(2):106–119.
 10. Angelelli E, Speranza MG, Savelsbergh MWP. Competitive analysis for dynamic multi-period uncapacitated routing problems. *Networks* 2007;49(4):308–317.
 11. Maxwell MS, Restrepo M, Henderson SG, *et al.* Approximate dynamic programming for ambulance redeployment. *INFORMS J Comput* 2010;22(2):266–281.
 12. Adelman D. A price-directed approach to stochastic inventory/routing. *Oper Res* 2004; 52(4):499–514.
 13. Giesen R, Mahmassani HS, Jaillet P. Logistics in real-time: inventory-routing operations under stochastic demand. In: Bertazzi L, Speranza MG, van Nunen JAEE, editors. Volume 619, Innovations in Distribution Logistics, Lecture notes in economics and mathematical systems. Berlin: Springer; 2009. pp. 109–148.
 14. Klewegt AJ, Nori VS, Savelsbergh MWP. The stochastic inventory routing problem with direct deliveries. *Transp Sci* 2002; 36(1):94–118.
 15. Nascimento JM, Powell WB. An optimal approximate dynamic programming algorithm for the energy dispatch problem with grid-level storage. Working paper. 2009. Available at <http://www.castlelab.princeton.edu/Papers/Nascimento-Powell-EnergyStorageSept0720092009.pdf>. Accessed 2010 Jan 18.
 16. Bent R, Van Hentenryck P. The value of consensus in online stochastic scheduling. In: Zilberstein S, Koehler J, Koenig S, editors. Proceedings of the 14th International Conference on Automated Planning and Scheduling (ICAPS 2004). Cambridge (MA): American Association for Artificial Intelligence; 2004. pp. 219–226.
 17. Branke J, Mattfeld DC. Anticipation and flexibility in dynamic scheduling. *Int J Prod Res* 2005;43(15):3103–3129.
 18. Wu T, Powell WB, Whisman A. The optimizing-simulator: an illustration using the military airlift problem. *ACM Trans Model Comput Simul* 2009;19(3):1–31.
 19. Larsen A, Madsen OGB, Solomon MM. Classifications of dynamic vehicle routing systems. In: Ziempeks V, Tarantilis CD, Giaglis GM, *et al.*, editors. Volume 38, Dynamic fleet management: concepts, systems, algorithms, and case studies, Operations research | computer science interfaces. New York: Springer; 2007. pp. 19–40.
 20. Eksioglu B, Vural AV, Reisman A. The vehicle routing problem: a taxonomic review. *Comput Ind Eng* 2009;57(4):1472–1483.
 21. Larsen A, Madsen OGB, Solomon MM. Recent developments in dynamic vehicle routing systems. In: Golden B, Raghavan R, Wasil E, editors. Volume 43, The vehicle routing problem: latest advances and challenges, Operations Research/Computer science interfaces. New York: Springer; 2008. pp. 199–218.
 22. American Transportation Research Institute. Research results: an analysis of the operational cost of trucking. Available at http://www.atr-online.org/research/results/economicanalysis/Operational_Costs_OnePager.pdf. Accessed 2010 Jul 14.
 23. The Tioga Group. Chapter 3: truck transportation. Project 99-130: Goods Movement Truck and Rail Study; Jan 2003. Prepared for the Southern California Association of Governments. Available at <http://www.scag.ca.gov/goodsmove/truckrail.htm>. Accessed 2010 Jan 18.

24. Thomas BW. Waiting strategies for anticipating service requests from known customer locations. *Transp Sci* 2007;41(3):319–331.
25. van de Klundert J, Wormer L. ASAP: the after-salesman problem. *Manuf Serv Oper Manage*. In press.
26. Jaillet P. A priori solution of the traveling salesman problem in which a random subset of customers are visited. *Oper Res* 1988;36:929–936.
27. Campbell AM, Thomas BW. The probabilistic traveling salesman problem with deadlines. *Transp Sci* 2008;42(1):1–21.
28. Chen S, Golden B, Wong R, *et al.* Arc-routing models for small-package local routing. *Transp Sci* 2009;43(1):43–55.
29. Crainic TG, Laporte G, editors. *Fleet management and logistics*. Boston (MA): Kluwer Academic Publishers; 1998.
30. Toth P, Vigo D, editors. *The vehicle routing problem, SIAM monographs on discrete mathematics and applications*. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2002.
31. Bertsimas DJ. A vehicle routing problem with stochastic demand. *Oper Res* 1992;40:574–585.
32. Bertsimas DJ, Chervi P, Peterson M. Computational approaches to stochastic vehicle routing problems. *Transp Sci* 1995;29:342–352.
33. Dror M, Trudeau P. Stochastic vehicle routing with modified savings algorithm. *Eur J Oper Res* 1986;23:228–235.
34. Laporte G, Louveaux FV, Mercure H. The vehicle routing problem with stochastic travel times. *Transp Sci* 1992;26:161–170.
35. Campbell AM, Thomas BW. Advances and challenges in a priori routing. In: Golden B, Raghavan R, Wasil E, editors. Volume 43, *The vehicle routing problem: latest advances and challenges*, Operations Research/Computer Science Interfaces. New York: Springer; 2008. pp. 123–142.
36. Christiansen C, Lysgaard J. A branch-and-price algorithm for the capacitated vehicle routing problem with stochastic demands. *Oper Res Lett* 2007;35:773–781.
37. Ak A, Erera AL. A paired-vehicle recourse strategy for the vehicle-routing problem with stochastic demands. *Transp Sci* 2007;41(2):222–237.
38. Morales JC, Erera AL, Savelsbergh MWP. Robust duration-constrained tours for vehicle routing problems with stochastic demands. *Proceedings of the 6th Triennial Symposium on Transportation Analysis, TRISTAN VI*. Phuket; 2007. Available at <http://transp-or2.epfl.ch/tristan/finalProgramBySpeaker.php>.
39. Sungur I, Ordonez F, Dessouky M. A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty. *IIE Trans* 2008;40(5):509–523.
40. Jaw J-J, Odoni AR, Psaraftis HN, *et al.* A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transp Res Part B* 1986;20B:243–257.
41. Madsen OBG, Ravn HF, Rygaard JM. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Ann Oper Res* 1995;60:193–208.
42. Larsen A, Madsen OBG, Solomon M. Partially dynamic vehicle routing - models and algorithms. *J Oper Res Soc* 2002;53:637–646.
43. Powell WB, Jaillet P, Odoni A. Stochastic and dynamic networks and routing. In: Ball MO, Magnanti TL, Monma CL, *et al.* editors. Volume 8, *Network routing, Handbooks in operations research and management science*. Amsterdam: North-Holland Publishing Co.; 1995. pp. 141–295.
44. Psaraftis HN. Dynamic vehicle routing problems. In: Golden BL, Assad AA, editors. *Vehicle routing: methods and studies*. Amsterdam: North-Holland Publishing Co.; 1988. pp. 223–248.
45. Psaraftis HN. Dynamic vehicle routing: status and prospects. *Ann Oper Res* 1995;61:143–164.
46. Mitrović-Minić S, Laporte G. Waiting strategies for the dynamic pickup and delivery problem with time windows. *Transp Res Part B* 2004;38:635–655.
47. Manni E. *Topics in real-time fleet management [PhD thesis]*. Università Della Calabria; 2007.
48. Branke J, Middendorf M, Noeth G, *et al.* Waiting strategies for dynamic vehicle routing. *Transp Sci* 2005;39:298–312.
49. Larsen A, Madsen OBG, Solomon M. The a-priori dynamic traveling salesman problem with time windows. *Transp Sci* 2004;38:459–472.
50. Ichoua S, Gendreau M, Potvin J-Y. Exploiting knowledge about future demands for real-time vehicle dispatching. *Transp Sci* 2006;40(2):211–225.
51. Cheung BKS, Choy KL, Li CL, *et al.* Dynamic routing model and solution methods for fleet

- management with mobile technologies. *Int J Prod Econ* 2008;113:694–705.
52. Gendreau M, Guerten F, Potvin J-Y, *et al.* Parallel tabu search for real-time vehicle routing and dispatching. *Transp Sci* 1999;33:381–390.
 53. Ichoua S, Gendreau M, Potvin J-Y. Diversion issues in real-time vehicle dispatching. *Transp Sci* 2000;34:426–438.
 54. Kilby P, Prosser P, Shaw P. Dynamic VRPs: a study of scenarios. Technical Report APES-06-1998. Glasgow: Department of Computer Science, Strathclyde University; 1998. Available at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.1.6748>.
 55. Li J, Mirchandani P, Borenstein D. Real-time vehicle rerouting problems with time windows. *Eur J Oper Res* 2009;194(3):711–727.
 56. Dial RB. Autonomous dial-a-ride transit introductory overview. *Transp Res Part C* 1995;3:261–275.
 57. Savelsbergh M, Sol M. DRIVE: dynamic routing of independent vehicles. *Oper Res* 1998; 46:474–490.
 58. Mitrović-Minić S, Krishnamurti R, Laporte G. Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows. *Transp Res Part B* 2004;38:669–685.
 59. Chen Z-L, Xu H. Dynamic column generation for dynamic vehicle routing with time windows. *Transp Sci* 2006;40(1):74–88.
 60. Thomas BW, White CC III. Anticipatory route selection. *Transp Sci* 2004;38:473–487.
 61. Secomandi N, Margot F. Reoptimization approaches for the vehicle-routing problem with stochastic demands. *Oper Res* 2009;57(1):214–230.
 62. Erera AL, Daganzo CF. A dynamic scheme for stochastic vehicle routing; 2006. Submitted for publication. Available at http://www2.isye.gatech.edu/people/faculty/Alan_Erera/pubs/route.html. Accessed 2010 Jan 18.
 63. Secomandi N. Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands. *Comput Oper Res* 2000;27:1201–1225.
 64. Secomandi N. A rollout policy for the vehicle routing policy with stochastic demands. *Oper Res* 2001;49:796–802.
 65. Bertsekas D, Tsitsiklis J, Wu C. Rollout algorithms for combinatorial optimization. *J Heuristics* 1997;3(3):245–262.
 66. Novoa C, Storer R. An approximate dynamic programming approach for the vehicle routing problem with stochastic demands. *Eur J Oper Res* 2009;196:509–515.
 67. Goodson J, Ohlmann JW, Thomas BW. New rollout policies for approximate dynamic programming with application to the vehicle routing problem with stochastic demand and duration limits; 2010. Available at <http://myweb.uiowa.edu/bthoa/Research.html>. Accessed 2010 Apr 30.
 68. Bent RW, Van Hentenryck P. Scenario-based planning for partially dynamic vehicle routing with stochastic customers. *Oper Res* 2004;52:977–987.
 69. Campbell AM, Savelsbergh MWP. Decision support for consumer direct grocery initiatives. *Transp Sci* 2005;39(3):313–327.
 70. Ghiani G, Manni E, Quaranta A, *et al.* Anticipatory algorithms for same-day courier dispatching. *Transp Res Part E* 2009;45:96–106.
 71. Hvattum LM, Løkketangen A, Laporte G. Solving a dynamic and stochastic vehicle routing problem with a sample scenario hedging heuristic. *Transp Sci* 2006;40(4):421–438.
 72. Hvattum LM, Løkketangen A, Laporte G. A branch-and-regret heuristic for stochastic and dynamic vehicle routing problems. *Networks* 2007;49(4):330–340.
 73. van Hemert JJ, La Poutré JA. Dynamic routing with fruitful regions: models and evolutionary computation. In: Yao X, Burke E, Lozano JA, *et al.*, editors. Volume 3242/2004, Proceedings of the 8th International Conference on Parallel Problem Solving from Nature, Lecture notes in computer science. Berlin: Springer; 2004. pp. 692–701.
 74. Powell WB. Approximate dynamic programming: solving the curses of dimensionality, Wiley series in probability and statistics. 1st ed. Hoboken (NJ): Wiley-Interscience; 2007.
 75. Goodson J. Dynamic solutions for the multi-vehicle routing problem with stochastic demand and duration limits. Invited Presentation, INFORMS Annual Meeting; 2009 Oct; San Diego (CA).
 76. Meisel S, Suppa U, Mattfeld D. GRASP based approximate dynamic programming for dynamic routing of a vehicle. In: Voss S, Caserta M, editors. Proceedings of MIC 2009: The VIII Metaheuristics International Conference. Hamburg: Springer; 2009.
 77. Proper S, Tadepalli P. Scaling model-based average-reward reinforcement learning for

- product delivery. In: Carbonell JG, Siekmann J, editors. Volume 4212/2006, Machine learning: ECML 2006, Lecture notes in computer science. Berlin: Springer; 2006. pp. 753–742.
78. Jaillet P, Wagner MR. Online vehicle routing problems: A survey. In: Golden B, Raghavan R, Wasil E, editors. Volume 43, The vehicle routing problem: latest advances and challenges, Operations Research/Computer science interfaces. New York: Springer; 2008. pp. 221–238.
 79. Bertsimas DJ, Ryzin GV. A stochastic and dynamic vehicle routing problem in the Euclidean plane. *Oper Res* 1991;39:601–615.
 80. Bertsimas DJ, Ryzin GV. Stochastic and dynamic vehicle routing problem in the euclidean plane with multiple capacitated vehicles. *Oper Res* 1993;41:60–76.
 81. Jaillet P, Wagner MR. Online routing problems: value of advanced information as improved competitive ratios. *Transp Sci* 2006;40(2):200–210.
 82. Jaillet P, Wagner MR. Generalized online routing: new competitive ratios, resource augmentation, and asymptotic analyses. *Oper Res* 2008;56(3):745–757.
 83. Swihart MR, Papastavrou JD. A stochastic and dynamic model for the single-vehicle pick-up and delivery problem. *Eur J Oper Res* 1999;114:447–464.
 84. Wong RT. Vehicle routing for small package delivery and pickup services. In: Golden B, Raghavan R, Wasil E, editors. Volume 43, The vehicle routing problem: latest advances and challenges, Operations Research/Computer science interfaces. Springer; 2008. pp. 123–142.
 85. Groer C, Golden B, Wasil E. The consistent vehicle routing problem. *Manuf Serv Oper Manage* 2009;11(4):630–643.
 86. Ordóñez F, Dessouky M, Sungur I. Better delivery/pick up routes in the presence of uncertainty. *Mettrans Project 06-11*. METTRANS Transportation Center, Department of Industrial and Systems Engineering, University of Southern California; Jan 2008. Available at <http://www.mettrans.org/research/final/06-11%20Final.pdf>. Accessed 2010 Jan 18.
 87. Smilowitz K, Nowak M, Jiang T. Workforce management in periodic delivery operations. Technical Report 09-004. Evanston (IL): Department of Industrial Engineering and Management Sciences, Northwestern University; 2009.
 88. Zhong H, Hall RW, Dessouky M. Territory planning and driver learning in vehicle dispatching. *Transp Sci* 2007;41(1):74–89.
 89. Benton WC, Rosetti MD. The vehicle scheduling problem with intermittent customer demands. *Comput Oper Res* 1992;19:521–531.
 90. Bowler NE, Fink TMA, Ball RC. Characterization of the probabilistic traveling salesman problem. *Phys Rev E* 2003;68:036703.
 91. Ghiani G, Manni E, Thomas BW. A comparison of anticipatory algorithms for the stochastic and dynamic traveling salesman problem; 2010. Submitted for publication. Available at <http://myweb.uiowa.edu/bthoa/Research.html>. Accessed 2010 Jun 19.

EDGEWORTH MARKET GAMES: PRICE-TAKING AND EFFICIENCY

PRABAL ROY CHOWDHURY
Indian Statistical Institute,
Delhi Center, Delhi, India

Consider M apple growers, each with an endowment of a unit of apples, and N bread makers, each with an endowment of b units of bread. In the absence of trade, the apple growers just consume their own apples, and the bread makers just consume their own bread. Most people prefer variety however, and presumably they would all prefer to consume some bread and some apple, rather than just one of these commodities. Edgeworth [1] examined the set of possible outcomes if trade now opens up in this economy. Using the notion of core (formalized later), Edgeworth provided a beautiful resolution of this question.

This question has continued to be one of the central ones in economics, generating a large literature which are often clubbed together under the title of Edgeworth market games. This literature models the trading process in an economy as a game played by utility maximizing agents. The solution concept used can be a cooperative one, or a noncooperative one.

Once technical issues of existence are resolved, the issues of interest are essentially twofold. A key aim of this literature is to provide a strategic foundation for price-taking behavior, a key element of perfect competition. By way of contrast, in the cooperative approach to trading, prices do not play a mediating role since agents can write binding contracts. Similarly, in the Arrow–Debreu–McKenzie (henceforth ADM) framework, while prices do play a mediating role, agents are *assumed* to be price takers. In contrast, the market game approach examines if price-taking emerges in the limit as the economy becomes large.

A related issue, which dates back to Ref. 1, is that of efficiency. Are the equilibrium outcomes Pareto efficient, in the sense there is no other possible outcome that is at least as good for all agents, and strictly better for some? This question is often examined in the context of large economies, where the issue of interest is whether, as the economy becomes large, the set of equilibrium outcomes converge to the set of Walrasian equilibria of the limit economy.

These are both complex issues however, and given the, by now, vast literature on market games, one must pick and choose. The focus of this article will be production economies in a partial equilibrium framework. While this does abstract from some of the intermarket interactions, doing so allows one to focus on the issue of price-taking behavior in a more transparent fashion. In such a framework, the issue of price taking, as well as efficiency reduce to a single one; that is, whether the equilibrium outcomes converge to the competitive one as the number of firms becomes large. Further, in a partial equilibrium framework one can examine these issues without going into measure theoretic complexities that are unavoidable in a general equilibrium framework.

Consequently, while we shall of course discuss other branches of the literature, these will be in lesser details. This in no way suggests that these are any less relevant, or important. Another reason is that there are existing surveys of these branches which do a brilliant job of surveying the literature.

The rest of the article is organized as follows. In the section titled “Exchange Economies: The Core” we discuss some of the key insights in the cooperative market games literature, before moving on, in the rest of the article, to the noncooperative approach. The section titled “Exchange Economies: The Trading Post Approach” discusses the trading post approach to market games, developed in several seminal articles by Shubik and Shapley. In the sections titled

“Production Economies: Cournot Competition Under Partial Equilibrium” and “Price Competition: Partial Equilibrium,” we turn our attention to market games with production, analyzing both the quantity, as well as the price setting framework in a partial equilibrium framework. The section titled “Quantity Competition with Free Entry: General Equilibrium” extends the free entry Cournot approach to a general equilibrium setting while the final section concludes this article.

EXCHANGE ECONOMIES: THE CORE

Consider an exchange economy (i.e., without production), with M apple traders (type 1), each with an endowment of a units of apples, and N bread traders (type 2), each with an endowment of b units of bread. The utility function of agent i is given by $u(a_i, b_i)$, where a_i and b_i denote the consumption of apples and bread respectively by the agent. The utility function is taken to be concave (weaker restrictions would suffice), capturing the fact that there is a love for variety. That the utility functions are taken to be symmetric is entirely for convenience.

Let x_i (respectively y_i) be the amount of apples (respectively bread) consumed by agent i . $(x, y) = (x_1, \dots, x_{M+N}, y_1, \dots, y_{M+N})$ constitutes an *allocation* vector for the economy if $x_i \geq 0$, $y_i \geq 0$, $\sum x_i = Ma$ and $\sum y_i = Nb$. The utility vector corresponding to an allocation will be called an imputation.

Definition. An allocation (x, y) is said to be a *core* allocation if there exists no subset of agents that can improve on it by trading only among themselves, that is all agents in the subset do at least as well compared to (x, y) , and at least one agent does strictly better.

We proceed by solving for the characteristic function of this game. Let $S = \{S_a, S_b\}$ denote a subset of agents consisting of s_a apple traders and s_b bread traders. Let $v(S)$ denote the maximal aggregate utility if agents in the set S trade with each other and

side-payments are allowed for. Thus

$$v(S) = \max_{x_1, \dots, x_{s_a+s_b}, y_1, \dots, y_{s_a+s_b}} \sum_{i=1}^{s_a+s_b} u(x_i, y_i), \quad (1)$$

where $\sum_{i=1}^{s_a+s_b} x_i = s_a a$ and $\sum_{i=1}^{s_a+s_b} y_i = s_b b$.

Since $u(a_i, b_i)$ is concave, $\sum_{i=1}^{s_a+s_b} u_i$ is maximized when the outcome is symmetric. Thus

$$v(S) = (s_a + s_b)u\left(\frac{s_a a}{s_a + s_b}, \frac{s_b b}{s_a + s_b}\right). \quad (2)$$

Remark 1. Note that for a bilateral exchange game, that is the $[1, 1]$ economy, $v(\{1\}) = (a, 0)$, $v(\{2\}) = u(0, b)$ and $v(\{1, 2\}) = 2u(a/2, b/2)$. Thus, the core imputation of this economy is

$$\{(u(a, 0) + \alpha c, u(0, b) + (1 - \alpha)c), 0 \leq \alpha \leq 1\}, \quad (3)$$

where $c = 2u(a/2, b/2) - u(a, 0) - u(0, b)$. The core for this game can be graphically represented using the Edgeworth box diagram and the contract curve in an entirely standard fashion.

$[1, N]$ -Economy

Consider a monopolist agent of type 1 facing many agents of type 2. In such a scenario can the monopolist charge very high prices? The following proposition suggests that the answer is in the affirmative.

Proposition 1. *The imputation $((N + 1)u\left(\frac{a}{N+1}, \frac{Nb}{N+1}\right) - Nu(0, b), u(0, b), \dots, u(0, b))$ lies in the core.*

Note that in this outcome, all bread traders are held down to their utility under autarchy, and all the gains from trade accrue to the monopolistic apple trader. Intuitively, the result follows since no group of agents that exclude the single apple trader can do any better since, given that none of them have any apples, the best that they can do is not to trade at all. This gives the apple trader an advantageous position. Further, while there could be other core allocations,

for N large, it can be shown that *any* core allocation must be arbitrarily close to the monopolistic outcome.

$[N, N]$ -Economy

For $i \notin S$, let $S^i = S \cup \{i\}$.

Definition. The economy is said to exhibit increasing returns to scale (IRS) (respectively decreasing returns to scale (DRS)) if for all S and $i, i \notin S$, $\frac{v(S^i)}{|S^i|} \geq \frac{v(S)}{|S|}$ (respectively $\frac{v(S^i)}{|S^i|} \leq \frac{v(S)}{|S|}$).

It can be shown that whether a core exists or not is intimately connected to whether the economy exhibits IRS, or DRS.

Proposition 2. *For the $[N, N]$ economy, a core exists, if the economy exhibits IRS, but not if it exhibits DRS.*

This is quite intuitive. With IRS, any group that plans to deviate from the grand coalition will have less surplus to distribute among its members, as its size is smaller compared to the grand coalition. Hence deviation from the grand coalition is not very attractive, and a core exists.

We shall not discuss the limit properties of such $[N, N]$ economies, as the economy becomes large. As discussed in the introduction, the issue is whether the set of core outcomes coincide with the Walrasian equilibria of the limit economy to which the $[N, N]$ economy converges. If so, then this implies that the set of Walrasian equilibria (i) has predictive power, and (ii) captures all outcomes arising out of the trading process. While any in-depth discussion of these issues must involve technical issues that are beyond the scope of this article, the results in the literature suggest that such convergence does take place (under the appropriate conditions). It is interesting that for large economies there is a close relation between the notion of core, which is a cooperative notion, and Walrasian equilibria, which is a noncooperative one. We refer the readers to Refs 2–4 for a detailed discussion of these issues.

EXCHANGE ECONOMIES: THE TRADING POST APPROACH

Much of the subsequent literature has, however, followed a noncooperative approach to modeling market games. One reason is that the cooperative approach with its emphasis on binding contracts is conceptually more attractive for small economies where the costs of writing such contracts are likely to be small. Further, the possibility of writing binding contracts imply that the process of price formation is implicit, thus abstracting from one of the key issues of interest. It should be mentioned though that for large economies there is a close connection between the core and Walrasian equilibria, which is (i) noncooperative, and (ii) does require a mediating role for prices.

We then discuss the trading post approach to market games first discussed in Refs 5 and 6. In contrast to the cooperative approach, prices play a central mediating role here. Further, prices emerge out of the strategic Cournotian interaction of the traders, with no need for either an Walrasian auctioneer, or any assumption of price-taking on the part of the agents. *Inter alia*, the trading post approach seeks to supplement the almost institution-free description of economies in the ADM tradition with explicit institutional details.

Consider an exchange economy with l commodities, and a set of consumers characterized by their preferences and initial endowments. There are l trading posts, one for each commodity, and some element called money. This money, may, or may not be fiat money (depending on the model). At each market j , an agent i decides on how much to offer of each commodity, denoted by y_i^j , and a bid of m_i^j for money. Let $y_i^j \leq w_i^j$, where w_i^j is the endowment of the i -th agent of the j -th commodity. Thus, the strategy for every agent consists of the pair (m_i^j, y_i^j) . The corresponding vector for all the agents can be written as (m, y) . In Ref. 5, for example, it is taken that agents have to supply all their endowments for trade, and the only strategic vector is the bid. Further, they impose the condition that $m_i^j \geq 0$, since they

consider a scenario where traders have to pay cash-in-advance.

The price emerges out of the sum total of all bids and offers, with the clearing price in market j being given by

$$p^j = \frac{\sum_i m_i^j}{\sum_i y_i^j}.$$

Note that the clearing rule for each market only depends on the bids and offers in that market. Given the prices, trade in each market can be solved for, with trader i obtaining $m_i^j/p^j - y_i^j$ of commodity j . Thus, unlike the ADM framework, there is no need for an auctioneer. Clearly the rules of the game are well defined (with some appropriate convention if prices are zero).

Next, consider a Cournot–Nash equilibrium of this game. A vector (m, y) is said to be feasible for agent i if the total amount of money pledged is less than equal to money obtained for this agent.

Definition. A feasible bid-offer vector (m, y) is a Cournot–Nash equilibrium if for all i , there does not exist any deviating offer-bid vector that is feasible for agent i , and is strictly preferred by this agent.

Reference 5 shows that under standard assumptions one can use the Kakutani fixed point theorem to prove existence of a Cournot–Nash equilibrium. Moreover, for the Ref. 5 model, an equilibrium with positive trade exists whenever the agents are sufficiently diverse [7]. In general of course a no-trading equilibrium always exists in this class of models.

Much of the subsequent literature has concentrated on the limit properties of Cournot–Nash equilibria, when the finite E_n economy with n agents approximates a continuum Aumann economy E . The literature shows that under the appropriate conditions, the set of Cournot–Nash equilibria of the trading post economy approximates the set of Walrasian equilibria of the limit economy as the economy becomes large.

We refer the readers to Ref. 8 for a detailed discussion of these issues. The Ref. 9 symposium issue on noncooperative foundations of

competitive equilibria bring together some important articles on this and related issues.

Remark 2. There are several related variations of the trading post models, such as Refs 10–13. Reference 14 provides an axiomatic treatment of this class of models, identifying the factors that lead to efficiency. Further, note that while we restrict attention to exchange models, the trading post model has been extended to allow for production [7]. This continues to be an active research area, with recent articles attempting to model issues of information transmission [15], default [16], and so on, using this approach.

PRODUCTION ECONOMIES: COURNOT COMPETITION UNDER PARTIAL EQUILIBRIUM

Continuing our discussion of noncooperative market games, we next turn to economies with production. In particular, we begin by discussing quantity competition under a partial equilibrium framework. This dates back to Cournot [17], who examined a game where two owners of mineral springs simultaneously decide on how much to supply in the market. Clearly, costs are zero. Even so, Cournot shows that the market outcome involves a strictly positive price, so that the equilibrium exhibits market power; that is, price is higher than marginal cost. This is clearly inefficient, even though competition does mitigate some of the inefficiencies associated with monopoly.

Efficiency is concerned with the divergence between the Cournot equilibrium price and the perfectly competitive price. The folk theorem of perfect competition suggests that, as the number of firms increases and/or the firm size becomes relatively small, the market price converges to the competitive one. This section examines the validity of the folk theorem in the quantity competition framework (the case of price competition will be taken up in the next section).

Recall that the Marshallian perfectly competitive model is based on U-shaped average cost functions (i.e., $c(q)/q$), with the competitive price being given by the minimum

point of the average cost function. Given the emphasis on the folk theorem, U-shaped average cost functions will play an important role in the analysis (for quantity, as well price competition). We shall allow for other forms of cost functions though.

Consider a model with n firms, all producing a homogeneous good. Let q_i denote the output of firm i , $Q = \sum_{i=1}^n q_i$ denote aggregate output, and Q_i denote the aggregate output of all firms but the i -th one. Let the inverse demand function be $f(Q)$ and the cost function of all firms be $c(q)$. Thus, the profit function of the i -th firm is given by

$$\pi(q_i, Q_i) = q_i f(Q) - c(q_i). \quad (4)$$

We make the following assumptions on the demand and the cost functions:

- A1.** The inverse demand function $f : [0, \infty) \rightarrow [0, \infty)$ is twice differentiable, intersects the quantity axis at $\hat{Q} > 0$, and is negatively sloped for all Q such that $\hat{Q} > Q > 0$.
- A2.** The cost function $c : [0, \infty) \rightarrow [0, \infty)$ is twice differentiable (except possibly at $q = 0$), increasing in q , and $c(0) = 0$.
- A3.** The Hahn conditions hold, that is $\forall q, Q$, where $0 \leq q \leq Q$, (i) $f'(Q) - c''(q) < 0$, and (ii) $f''(Q) - q_i f'''(Q) < 0$.

Consider the Cournot game where all the firms simultaneously decide on their quantity q_i , where $q_i \in [0, \infty)$. We look for Nash equilibria of this game in pure strategies.

Proposition 3. *Under A1–A3, the Cournot game has an equilibrium in pure strategies.*

Given A1–A3, $\pi(q_1, \dots, q_n)$ is strictly concave in q_i and continuous in $(q_1, \dots, q_n)$. Further, the strategy set for all firms can be effectively restricted to $[0, \hat{Q}]$ so that it is compact. The preceding proposition then follows from the Debreu fixed-point theorem.

While uniqueness is of some interest, it is not central to our concerns, and we just refer the interested reader to Ref. 18 instead. We then concentrate on studying the limit properties of Cournot equilibria. We shall

consider two different approaches, one with exogenous, and the other with free entry.

As the formal analysis (to follow) demonstrates, the analysis depends on the relative efficacy of two forms of competition, actual and potential, in generating efficiency. By *actual competition* we mean the number of active firms in an industry, since these firms exert a direct competitive effect on the other firms in the industry. *Potential competition* refers to the fact that even inactive firms create a competitive pressure in that if the market price is too high, then such firms can enter and have a positive profit.

We shall demonstrate that the strength of actual competition depends on the shape of the average cost function, in particular whether it is increasing, or U-shaped. It turns out that actual competition is sufficient to generate the competitive price as long as the average cost function is increasing, but not if it is U-shaped. When average costs are U-shaped, we find that the competitive price obtains if potential competition is strong, that is there is free entry of firms, and firm size vis-a-vis market size is small.

Exogenous Entry

Under the exogenous entry approach, firm size vis-a-vis market size is taken to be given, and one examines the limiting properties of equilibrium prices as the number of firms, which is exogenously given, taken to infinity. This approach was pioneered by Okuguchi [19] and Ruffin [20], among others.

To begin with, note that if there are fixed costs of production, so that $c(0) > 0$, then the number of active firms cannot be arbitrarily large, since some firms will be making losses otherwise. While this is an important caveat, it is well understood and in the rest of the subsection we therefore concentrate on the case where there are no fixed costs, so that cost functions are continuous at zero.

Proposition 4. *Let A1–A3 hold, the cost function be continuous at $q = 0$ and further assume that for all n a unique, symmetric, and interior equilibrium exists.*

(a) *If the average cost is increasing in q , then for n large the Cournot equilibrium price converges to the competitive price, that is $c'(0)$.*

(b) If the average cost has the classical U-shape, then for n large, the Cournot equilibrium price does not converge to the competitive price.

Proof. We begin by arguing that for n large, the equilibrium price converges to $c'(0)$. Let $q(n)$ denote the unique symmetric equilibrium price. Given that the demand function intersects the quantity axis, $\lim_{n \rightarrow \infty} q(n) = 0$. Thus, from the first order condition and symmetry

$$\begin{aligned} \lim_{n \rightarrow \infty} f(nq(n)) \\ = \lim_{n \rightarrow \infty} [-q(n)f'(nq(n)) + c'(q)] = c'(0). \end{aligned} \quad (5)$$

Finally, note that for increasing average costs, the competitive price is in fact $c'(0)$. Whereas for U-shaped average costs, the competitive price is the minimum point of the average cost curve, which is less than the competitive price.

Proposition 4 has interesting efficiency implications. It shows that while the folk theorem goes through under the exogenous entry approach whenever the average cost is increasing, it fails whenever it is U-shaped. Under the exogenous entry approach therefore, the nature of technology, that is firm size, has an important bearing on the limiting outcome. Let firm size be formalized by the minimum efficient scale. Thus, efficiency obtains when the minimum efficient scale is zero (i.e., average costs are increasing), but not otherwise. Interestingly, a firm size close to zero does not ensure a limit equilibrium price close to zero (since a firm size close to zero is consistent with $c'(0)$ being bounded away from the competitive price).

Why do the competitive outcome obtains even in the absence of free entry (i.e., potential competition) whenever average cost is increasing? Intuitively, with increasing average costs, the minimum efficient scale is zero (recall that the competitive price corresponds to the average cost at the efficient size). Thus as actual competition increases, competition drives the output level of every firm towards the efficient scale, so that price goes towards

the competitive price. With U-shaped average costs however, an increase in average costs drives the output level of the firms away from the minimum efficient scale.

Free Entry

Novshek [21] argues that the notion of perfect competition itself incorporates notions of (i) free entry, and (ii) firm size being small. It is a combination of these two elements that generates the competitive pressure under perfect competition, leading to efficiency. Thus any reasonable oligopoly foundation of perfect competition needs to start with a model that allows for free entry and then examine limiting behavior when firm size becomes small vis-a-vis market size. Consequently, it may be argued that is not very surprising that the folk theorem fails under the exogenous entry approach, which allows for neither of these aspects.

Consider a market with a countably infinite number of firms and an inverse demand function $f(Q)$ satisfy A1. The firms are symmetric and have the cost function $c(q)$, where $c(q)$ satisfies A2 and

A4. The average cost function $AC : (0, \infty) \rightarrow (0, \infty)$ is continuous on $(0, \infty)$, strictly U-shaped, and achieves a unique minimum at $q = 1$. Moreover, there exists q such that $f(q) > AC(q)$.

Note that the cost function need not be continuous at $q = 0$, thus allowing for fixed costs.

We begin by defining the notion of a free-entry Cournot equilibrium.

Definition. $(q_1^*, \dots, q_n^*)$ constitutes a free-entry Nash equilibrium with n active firms if the following two conditions are satisfied:

- (a) *Cournot Equilibrium.* $(q_1^*, \dots, q_n^*)$ constitutes the equilibrium of an n -firm Cournot model without free entry, that is $q_i^* f(Q^*) - c(q_i^*) \geq q_i f(Q_i^* + q_i) - c(q_i)$, for all active firm i and $\forall q_i \geq 0$.
- (b) *Free Entry.* No inactive firm can enter and make a strictly positive profit, i.e. there does not exist $q' > 0$ such that $q' f(Q^* + q') - c(q') > 0$.

We examine the limit properties of free-entry equilibrium as firm size becomes small. We therefore, next turn to the task of formalizing a reduction in firm size. Recall that firm size is captured through the minimum efficient scale. That is, the output level where AC is minimized. We next define a family of cost functions parametrized by α defined as

$$AC_\alpha(q) = AC\left(\frac{q}{\alpha}\right). \quad (6)$$

Note that a firm with a cost function $AC_\alpha(q)$ achieves a minimum average cost at $q = \alpha$ and thus, has a firm size of α . Further, for any such α , the competitive price equals $AC_\alpha(\alpha) = AC(1)$. Finally, let X denote the competitive output, that is $f(X) = AC(1)$.

Proposition 5 below shows that for α small, the equilibrium output is arbitrarily close to the competitive one. Thus the folk theorem holds in this case. Interestingly, the result holds even if there are fixed costs, or if the average cost is U-shaped.

Proposition 5. *Let A1, A2, and A4 hold. For α small, a free-entry Cournot equilibrium exists. Moreover, the equilibrium output lies in $[X - \alpha, X]$.*

Proof. For the existence proof we refer the readers to Ref. 21. Next suppose that the aggregate equilibrium output, $Q^* > X$. But then $f(Q^*) < f(X)$, where recall that $f(X)$ is the minimum average cost. But then all active firms must be making a loss. Next, suppose that $Q^* < X - \alpha$, so that $f(Q^* + \alpha) > f(X)$. But then an inactive firm can enter, supply exactly α and make a positive profit.

Intuitively, the result is driven by the presence of free entry. With free entry, whenever prices are too high, inactive firms can always enter and produce at the minimum efficient scale. Since the price level is high to begin with, even with an increase in the output level, the resultant price is going to be higher than the average cost of the new entrant so that such a firm has an incentive to enter. This consideration provides a bound on how far the equilibrium price can

exceed the competitive price, with the bound approaching the competitive price as the efficient scale becomes smaller.

We note the fact that the inverse demand function is negatively sloped plays an important role here as this ensures that it sends the right signal for entry, ensuring that entry occurs precisely when prices are high. In order to emphasize this point, consider an inverse demand function which is not necessarily negatively sloped. In particular, let there be three aggregate output levels Q, Q', Q'' , such that $Q < Q' < Q''$, that generate a market price of $AC(1)$. Further let the demand function be negatively sloped at Q , and Q'' , but positively sloped at Q' . Mimicking Proposition 4, it is straightforward to show that for α small, no output in $[Q', Q' + \alpha]$ can be sustained as a free-entry equilibrium. This is because at any such output level, an inactive firm will have an incentive to enter and make a positive profit. As we shall see later, the critical condition that ensures efficiency under Cournot competition in a general equilibrium framework is something similar, that is some generalized demand function be negatively sloped.

PRICE COMPETITION: PARTIAL EQUILIBRIUM

Bertrand [22] criticized Cournot, arguing that firms actually compete over prices, and not quantities. While it is known that there are close connections between models of quantity and price competition, examining the limit properties of equilibria under price competition is of interest. Price competition has the further appeal that prices emerge directly out of the strategic interactions of the agents.

Next, we therefore examine market games where firms compete over prices, rather than quantities, with an aim to analyzing the limit properties of such equilibria. As we shall see, there are some differences in results between price and quantity competition. Under price competition we shall examine two different frameworks, the Bertrand–Chamberlin and the Bertrand–Edgeworth framework,

showing that the limit properties of the equilibrium prices are critically dependent on the underlying framework.

Bertrand–Chamberlin Framework

Under this framework, first developed by Chamberlin [23], firms have to supply the whole of the demand coming to them. This may be because of either government regulations, or reputational effects. Reference 18, for example, suggests that in many public utilities like electricity and telephone in the United States, such restrictions are in place. Under the “common carrier” regulation, firms are required to supply all demand at the set prices. Alternatively, under some sealed bid auctions, firms have to supply all demand that come to them in case they win the auction. Such costs are routinely adopted in operations research [24,25].

We now proceed with a description of the model. The analysis in this subsection follows [26]: Let there be n firms with the market demand function $F(p)$ satisfying

- A5.** $F : [0, \infty) \rightarrow [0, \infty)$. Further, $F(p)$ is continuous, intersects the price axes at $\hat{p}$, and is strictly decreasing on $[0, \hat{p})$.

Given the Chamberlin assumption, the residual demand facing firm i , if the announced price vector is $(p_1, p_2, \dots, p_n)$ is

$$D_i(p_1, \dots, p_i, \dots, p_n) = \begin{cases} 0, & \text{if } p_i > p_j, \text{ for some } j, \\ \frac{F(p_i)}{m}, & \text{if } p_i \leq p_j, \forall j, \text{ and } \#(l : p_l = p_i) = m. \end{cases}$$

Note that the market is equally shared by all firms that charge the lowest prices. All other firms have no demand. Thus, the profit of the i -th firm when the average cost function is $AC(q)$ is

$$\begin{aligned} \pi_i(p_1, \dots, p_n) &= (p_i - AC(D_i(p_1, \dots, p_n)))D_i(p_1, \dots, p_n). \end{aligned} \quad (7)$$

We look for pure-strategy Nash equilibria of the game where all firms simultaneously announce their prices.

We begin by considering the case where cost functions are linear. Whenever there

are at least two firms, it turns out that price equals marginal cost equals average costs, so that there are no supernormal profits. Hence the competitive equilibrium obtains even with only two firms—the Bertrand paradox! The paradox arises because intuition suggests that with only a few firms, the market price should be greater than the competitive one (however, see Ref. 27).

Proposition 6. *Let the cost function be $cq_i \forall i$, and A5 hold. The unique pure-strategy equilibrium involves all firms charging the same price c .*

Given the Bertrand paradox we consider strictly convex costs, in particular we allow for U-shaped average cost functions. Let $c(q)$ denote the common cost function of all firms where $c(q)$ satisfies A2. Further, let the average cost function $AC(q)$ satisfy

- A6.** The average cost function $AC : (0, \infty) \rightarrow (0, \infty)$ is continuous on $(0, \infty)$. There exists $\tilde{q}$ such that $AC(q)$ is decreasing for $q \leq \tilde{q}$, and increasing for $q \geq \tilde{q}$. Moreover, there exists p such that $p > AC(F(p))$.

We need some more definitions (these are well defined under our assumptions).

$$b = \lim_{q \rightarrow 0} AC(q).$$

$$c^* = \inf_q AC(q).$$

$d(1)$ is the minimum p such that $AC(f(p)) = p$. A monopoly firm charging some price p , makes a loss whenever $p < d(1)$. Whereas for any $p > d(1)$, there exists some $p' < p$, such that charging p' yields a positive profit to a monopoly firm.

Exogenous Entry. Recall that under the exogenous approach firm size relative to market size is taken as given, and one then examines the effect of an increase in the number of firms on the outcome. One therefore examines the limit equilibrium set which comprises of all prices that can be obtained as the limit of a sequence of equilibrium prices as the number of active firms becomes large.

We thus examine the properties of the set S , where S is the set of all prices p such

that for all sufficiently large n , there is some equilibrium where all firms are active and the equilibrium price is arbitrarily close to p .

Proposition 7. *Let A2, A5, and A6 hold, and let the Chamberlin assumption hold, i.e. firms supply all demand. The limit equilibrium set under exogenous entry $S = [b, d(1)]$ whenever $b \leq d(1)$. Otherwise, S is empty.*

The result is quite intuitive. Consider any price $p < b$. Can such a p be an element of the limit equilibrium set? Then, as the number of active firms increases, the output of all firms becomes small, so that average cost goes to b . But this implies that the firms make losses. Next, consider any price $p > d(1)$. For n large, the profit of the firms goes to zero, whereas by undercutting, the firms get a profit that is bounded away from zero for all n . Finally, consider some $b < p < d(1)$. For n large the average cost is close to b , so that firms make positive profits with $b > p$. Moreover, since $p < d(1)$, undercutting is not profitable.

Thus, while the competitive price is an element of the limit equilibrium set, there are other elements also. Intuitively, multiple equilibria obtain since under the Chamberlin assumption, any under-cutting firm will have to supply the whole of the demand coming to it, which given that cost functions are convex, is very costly. Reference 28 examines a model of Bertrand–Chamberlin competition that allows for limited collusion among the firms. They find that with symmetric firms, the coalition-proof Nash equilibrium converges to the competitive price when the number of firms becomes large.

Comparing with the results under Cournot competition (Proposition 4), with increasing average costs, the folk theorem holds under Cournot competition, but fails under price competition. Whereas for U-shaped average costs, the folk theorem fails under both approaches.

Free Entry. We then examine free-entry equilibrium under price competition, with free entry being modeled as the presence of inactive firms in equilibrium. Under this approach, one solves for the limit equilibrium set T , comprising all prices that

can be sustained as the limit of a sequence of free entry equilibrium prices as firm size becomes small relative to market size. We are thus interested in the set T , where T is the set of all prices p such that if the demand sector is large enough then there is an equilibrium involving both active and inactive firms such that the equilibrium price is arbitrarily close to p . Consider an r -economy, where the demand sector is replicated r -fold.

We need one more piece of notation. Let $d^* = \lim_{r \rightarrow \infty} d(r)$, where $d(r)$ solves $AC(rf(p)) = p$.

In Proposition 8 below we characterize the set T for U-shaped average cost functions.

Proposition 8. *Let A2, A5, and A6 hold, the average cost be U-shaped, and firms supply all demand. Then the limit equilibrium set under free entry $T = [c^*, \min\{b, d^*\}]$.*

While the complete argument can be found in Ref. 26, here we just consider p such that $c^* < p < \min\{b, d^*\}$. Why does such a p belong to the limit equilibrium set? This follows since for any sufficiently large r -economy, undercutting p will lead to a loss. For r large enough, in case of undercutting, the undercutting firm must supply a demand of at least $rf(p)$. For r large, the corresponding average cost will be close to $\min\{b, d^*\}$. Since this is greater than p , the undercutting firm will make a loss.

It is interesting to compare the results with Proposition 6. Note that in the presence of free entry, the limit equilibrium set shifts to the left, suggesting that free entry leads to a fall in prices. Thus free entry does play a role in generating outcomes close to the competitive one. There are other prices however that can be sustained in equilibrium. Intuitively, this is because with the Chamberlin assumption and convex costs, undercutting is very costly, so that even free entry and small firm size is not sufficient to guarantee that the competitive price obtains.

Recall that under Cournot equilibrium, for U-shaped average costs the folk theorem holds under the free-entry approach (Proposition 5). Thus under the free entry approach, the folk theorem fails under

price competition, but goes through under Cournot competition. This shows that under the Chamberlin assumption, free entry is not sufficient to restore efficiency. This is because with convex costs and the Chamberlin assumption, either undercutting, or even matching the existing market price can be very costly for an inactive firm, thus weakening the competitive impact of free entry.

Bertrand–Edgeworth Competition

We then examine a price-setting game where the voluntary trading constraint holds; that is, firms are free to supply less than the demand coming to them. This implies of course that no firm will supply beyond the point such that price equals marginal cost. Such a framework makes sense when there is no cost of turning customers away.

Note that the voluntary trading constraint implies that firms charging higher prices may get a residual demand if the lower priced firms are not meeting the demand coming to them. We shall restrict attention to the cases where the residual demand is either proportional, or parallel.

For the case where $n = 2$, let firm 1 charge $p_1 > p_2$ and firm 2 supply $q_2 < d(p_2)$. Then the residual demand coming to firm 1 is $\max\{0, d(p_1) - q_2\}$ if the demand is parallel, and $d(p_1) \frac{d(p_2) - q_2}{d(p_2)}$ if the demand is proportional. For a description for the general case we refer the readers to Ref. 25.

With convex costs we have nonexistence of equilibria in pure strategies, the well-known Edgeworth paradox [29].

Proposition 9. *Let $n = 2$ and both firms have a strictly convex cost function $C(q)$. Then, no pure-strategy Nash equilibrium exists for either the parallel, or the proportional residual demand function.*

The result is intuitive. For any market price higher than the competitive price, one of the firms will have an incentive to undercut. If, however, both firms charge the competitive price, then increasing prices slightly is profitable, since the price effect dominates the quantity effect which has a second-order impact (as price equals

marginal cost at the competitive price). In fact, Ref. 30 shows that even with linear costs, the Edgeworth paradox obtains if firms produce to stock, that is decide on their prices and quantities simultaneously.

Given the Edgeworth paradox, the literature has pursued several different approaches. Reference 31 examines conditions under which Bertrand–Edgeworth games will have existence.

Capacity-Constrained Linear Costs. One route has been to examine cost functions which are linear, but have a capacity constraint. For such models, Refs 32 and 33 both find that the equilibrium converges in distribution to the competitive price. However, while Ref. 33 shows that for the parallel rationing rule there is convergence in the support as well, for the proportional rationing rule Ref. 32 finds that in general the monopolistic price always belongs to the support of equilibrium prices. For the parallel rationing rule Ref. 34 shows that iterated elimination of dominated strategies yields prices close to the competitive one.

Clearly, both Refs 32 and 33 adopt the exogenous entry approach. Further, both authors examine limiting procedures where the capacity level, which can be taken to be a proxy for firm size, is taken to zero. Thus with exogenous entry, capacity constraints, and a limiting procedure that takes firm size to zero, the folk theorem goes through under price competition, at least in the sense of convergence in distribution. Recalling Proposition 4(a), we find that these results are similar in flavour to that under Cournot competition with increasing average costs. In both cases the point seems to be that whenever the efficient firm scale is zero, actual competition is enough to generate efficiency.

Note however that while we are interpreting these models as one of Bertrand–Edgeworth competition, these might very well be interpreted as one with Bertrand–Chamberlin competition where the Chamberlin constraint takes the weaker form that firms must supply till the maximum of its capacity, and the demand coming to it. This is because with linear costs, if it

is profitable to supply any positive amount, then it must be optimal to supply till capacity.

Grid Pricing

Reference 35 examines a model with grid-pricing and smoothly convex cost functions, where the firms simultaneously decide on both their prices and quantities. Further, in case of a tie in the lowest prices, suppose that beyond a point the residual demand coming to any firm is insensitive to the quantity supply by this firm. Then for any grid size, if the number of firms (n) is large enough, then there is a unique Nash equilibrium where the equilibrium price is within a grid-unit of the competitive price. Moreover, the output levels of individual firms become vanishingly small as n becomes very large, and the aggregate output is close to the competitive one.

The intuition is as follows. Suppose all firms charge the lowest possible price in the grid that is greater than the marginal cost at zero. Then for n large, it is residual demand rather than marginal cost which determines firm supply. In that case price would not equal marginal cost, and firms may have no incentive to increase their price levels.

Interestingly, in case firms tied at the lowest price can manipulate the residual demand coming to them by increasing their output level, the aggregate equilibrium output may be ill-behaved. While the competitive price (modulo the grid) still turns out to be an equilibrium, the limiting value of the aggregate output diverges to infinity if $c'(0) = 0$ and n is taken to infinity. This is because in this case the costs of manipulating residual demand via quantity is small, so that all firms have an incentive to produce too much.

Remark 3. Reference 36 examines a model with decreasing average costs, showing that for small enough grid size the equilibrium price converges to the contestable one when the firms are symmetric. If the firms are asymmetric however, then Ref. 37 shows that the convergence is to the limit-pricing one.

Remark 4. Reference 29 also allows for convex costs, examining the ϵ -Nash

equilibria of Bertrand–Edgeworth equilibria. For parallel rationing rules, Ref. 29 shows that an ϵ -Nash equilibria exists when the market is replicated, and moreover the ϵ -equilibria converge to the competitive one. In Ref. 29, the replication process involves replicating market demand, which involves making firm size relatively small, as well as increasing the number of firms.

Remark 5. An alternate approach is to examine the case when products are differentiated [38]. This again is a vast literature, and for lack of space will not be discussed here.

Thus, the preceding discussion suggests that for Bertrand–Edgeworth competition and the exogenous entry approach, the equilibrium outcome approximates the competitive one in the limit whenever (i) either the cost functions are linear with capacity constraints and the limit procedure involves taking firm size to zero, or (ii) there is grid-pricing and the tie-breaking rule is not too manipulable. For the proportional rationing rule and capacity constraints, the convergence is however in distribution, and not in the support. Further, under grid-pricing and strongly manipulable tie-breaking rules, while the equilibrium price converges to the competitive one, the aggregate equilibrium output may not.

QUANTITY COMPETITION WITH FREE ENTRY: GENERAL EQUILIBRIUM

This section is based on Refs 39 and 40. Proposition 5 earlier shows that the key to obtaining efficiency under partial equilibrium Cournot competition is to allow for free entry. Extending this intuition to a general equilibrium framework, one obtains what Novshek and Sonnenschein call the Cournot–ADM–Marshallian synthesis. This is of importance for several reasons:

1. This explains price-taking behavior, rather than postulate it as under the Arrow–Debreu–McKenzie framework.

2. It allows a role for dynamics via entry and exit of firms.
3. It allows for intermarket interactions, not allowed for in a partial equilibrium framework.
4. It allows us to reexamine the fundamental welfare theorems in a framework that allows for entry and exit of firms, and obtains price-taking behaviour in the limit.

Consider a general equilibrium economy where the firm sector consists of countably infinite number of firms and free entry, with each firm's production possibility being given by two components: the origin (which allows for free entry) and a compact, strictly convex component, which moreover is bounded away from zero. Thus, unlike the ADM framework, the production set is nonconvex. Moreover, it provides a natural analogue of the U-shaped average cost under partial equilibrium (given that the second component is bounded away from zero). Finally, let the second component of the production set be parametrized by α , so that a smaller α naturally implies a scaling down and captures the idea of smaller-sized firms. The central point is that as α goes to zero, the aggregate production set converges to a convex technology.

The consumers are treated as price-takers, with each consumer i owning a fraction θ_i of all firms and having an initial wealth level of w_i . Thus, given a price vector p , each consumer has an income level of $p \cdot w_i + \theta_i p \cdot y$, where y denotes net aggregate production. Let the demand function of this consumer be $D_i(p, p \cdot w_i + \theta_i p \cdot y)$.

Define an inverse demand selection $F(y)$ as a price vector that clears markets given the output and entry decisions

$$\sum_i D_i(p, p \cdot w_i + \theta_i p \cdot y) - \sum_i w_i = y. \quad (8)$$

Note that $F(y)$ is the general equilibrium equivalent of the partial equilibrium inverse demand function. Further, given $F(y)$, it is straightforward to extend the notion of free-entry Cournot equilibrium to the general equilibrium framework for an economy with firm size α , denoted $E(\alpha)$.

References 39 and 40 show that under a condition they call DSD on the generalized inverse demand $F(y)$, analogs of the fundamental welfare theorems extend to this set up. They say that the aggregate inverse demand $F(y)$ satisfies the DSD condition if, for α small, entry by a new firm reduces profits after changes in all prices are accounted for. Intuitively, it is clear why DSD is required. Consider an equilibrium of the limit economy that does not satisfy DSD. Then for any aggregate output close to but less than this output, an inactive firm will have an incentive to enter since, by doing so, it can earn a positive profit.

Thus Refs 39 and 40 provide a synthesis of the Cournot–Marshall approach with the ADM framework. In so doing, they provided a foundation for the price-taking assumption, as well as the perfectly competitive model. They also show that the fundamental welfare theorems go through in this framework, and also enhances these by showing that (i) competition leads to the optimal number of firms, (ii) emphasizes the role of wealth distribution on price signals, and (iii) does not impose convexity of the aggregate production function.

Remark 6. While the analysis in this section (as well as the earlier two sections) focus on the case with nonconvex technologies, Refs 41 and 42 examine an alternative framework where the aggregate production set is assumed to be convex and the number of agents is exogenously given. We refer the readers to Ref. 8 for a discussion of this approach.

CONCLUSION

The market game literature contributes a lot to our understanding of several aspects of multilateral trading. The cooperative literature shows that if binding contracts are not too costly, then, for large markets, trading outcomes converge to the Walrasian equilibrium. The noncooperative approach focuses on the role of price formation. The trading post models confirm the intuition that even

when agents behave noncooperatively, the invisible hand of Adam Smith continues to operate. The partial equilibrium oligopoly models clarify the role of firm interaction in generating efficiency. Under quantity competition, one key insight is that unless average costs are increasing, efficiency, that is convergence to the competitive price, requires not only actual competition, but both free entry as well as firm sizes being small. Under price competition, some new insights emerge. Whether the voluntary trading constraint holds or not, turns out to be of importance. When the voluntary trading constraint holds, we find that the perfectly competitive price obtains in the limit (subject to some caveats). The results however, are quite different in case the firms have to supply all the demand. In this case, even free entry and small firm size is not enough to guarantee the efficient outcome. Finally, the free-entry Cournot model with general equilibrium provides a synthesis of the Marshallian model with the Arrow–Debreu–McKenzie framework. *Inter alia*, it provides a foundation for price-taking behavior, and augments the fundamental welfare theorems with new insights.

Acknowledgments

I am deeply indebted to the area editor Professor Marc Kilgour, and two anonymous referees for very helpful comments.

REFERENCES

1. Edgeworth F. The mathematical economics of professor Amaraso. *Econ J* 1922;32:400–407.
2. Hildenbrand WC. *Equilibria of a large economy*. Princeton (NJ): Princeton University Press; 1974.
3. Hildenbrand W, Kirman AP. *Introduction to equilibrium analysis*. Amsterdam: North-Holland Publishing Company/New York: American Elsevier; 1976.
4. Shubik M. Edgeworth market games. In: Luce RD, Tucker AW, editors. *Contributions to the theory of games IV*. Princeton (NJ): Princeton University Press; 1959. pp. 267–278.
5. Shapley L, Shubik M. Trade using one commodity as a means of payment. *J Polit Econ* 1977;85:927–968.
6. Shubik M. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *West Econ J* 1973;11:24–28.
7. Dubey P, Shubik M. A closed economic system with production and exchange modeled as a game of strategy. *J Math Econ* 1977;4:253–287.
8. Mas-colell A. The Cournotian foundations of Walrasian equilibrium theory: an exposition of recent theory. In: Hildenbrand W, editor. *Advances in economic theory*. Cambridge: Cambridge University Press; 1985.
9. Mas-colell A, editor. Symposium issue on non-cooperative approaches to the theory of perfect competition. *J Econ Theory* 1980;22(2): 279–312.
10. Jaynes J, Okuno M, Schmeidler D. Efficiency in an atomless economy with fiat money. *Int Econ Rev* 1978;19:149–157.
11. Postlewaith A, Schmeidler D. Approximate efficiency of non-Walrasian Nash equilibria. *Econometrica* 1978;46:127–135.
12. Schmeidler M. Walrasian analysis via strategic outcome functions. *Econometrica* 1980;48:1585–1593.
13. Shubik M. *Strategy and market structure*. New York: John Wiley & Sons, Inc.; 1959.
14. Dubey P, Mas-colell A, Shubik M. Efficiency properties of strategic market games: an axiomatic approach. *J Econ Theory* 1980;22:339–362.
15. Dubey P, Geanakoplos J, Shubik M. The revelation of information in strategic market games. *J Math Econ* 1987;16:105–137.
16. Dubey P, Geanakoplos J, Shubik M. Default and punishment in general equilibrium. *Econometrica* 2005;73:1–37.
17. Cournot A. *The mathematical principles of the theory of wealth*. Homewood (IL): Richard D. Irwin, Inc.; 1963.
18. Vives X. *Oligopoly pricing: old ideas and new tools*. Cambridge (MA), London: MIT Press; 1999.
19. Okuguchi K. Quasi-competitiveness and Cournot oligopoly. *Rev Econ Stud* 1973;40: 145–148.
20. Ruffin RJ. Cournot oligopoly and competitive behavior. *Rev Econ Stud* 1971;38:493–502.
21. Novshek W. Cournot equilibrium with free entry. *Rev Econ Stud* 1980;67:473–486.
22. Bertrand J. Book review of *theorie mathematique de la richesse sociale and of recherches sur les principes mathematiques de la theorie des richesses*. *J de Savants* 1883;67:499–508.

23. Chamberlin EH. The theory of monopolistic competition. Cambridge (MA): Harvard University Press; 1933.
24. Dixon H. Bertrand-Edgeworth equilibria when firms avoid turning customers away. *J Ind Econ* 1990;34:131–146.
25. Taha H. Operations research. New Delhi: Prentice Hall; 1982.
26. Novshek W, RoyChowdhury P. Bertrand equilibria with entry: limit results. *Int J Ind Organ* 2003;21:795–808.
27. Dubey P. Price-quantity strategic market games. *Econometrica* 1981;50:111–126.
28. Roy Chowdhury P, Sengupta K. Coalition-proof Bertrand equilibria. *Econ Theory* 2004;24:307–324.
29. Dixon H. Approximate equilibria in a replicated industry. *Rev Econ Stud* 1987;54:47–62.
30. Roy Chowdhury P. Bertrand-Edgeworth duopoly with linear costs: a tale of two paradoxes. *Econ Lett* 2005;88:61–65.
31. Maskin E. The existence of equilibrium with price setting firms. *Am Econ Rev* 1986;76:382–386.
32. Allen B, Hellwig M. Bertrand-Edgeworth equilibria in large markets. *Rev Econ Stud* 1986;53:175–204.
33. Vives X. Rationing rules and Bertrand-Edgeworth equilibria in large markets. *Econ Lett* 1986;21:113–116.
34. Borgers T. Iterated elimination of dominated strategies in a Bertrand-Edgeworth model. *Rev Econ Stud* 1992;59:163–176.
35. Roy Chowdhury P. Bertrand - Edgeworth equilibrium with a large number of firms. *Int J Ind Organ* 2007;26:746–761.
36. Ray Chaudhuri P. The contestable outcome as a Bertrand equilibrium. *Econ Lett* 1996;50:237–242.
37. Roy Chowdhury P. Limit-pricing as Bertrand equilibrium. *Econ Theory* 2002;19:811–822.
38. Vives X. On the efficiency of Bertrand and Cournot equilibria with product differentiation. *J Econ Theory* 1985;36:166–175.
39. Novshek W, Sonnenschein H. Cournot and Walras equilibrium. *J Econ Theory* 1978;19:223–266.
40. Novshek W, Sonnenschein H. General equilibrium with free entry: a synthetic approach to the theory of perfect competition. *J Econ Lit* 1987;15:1281–1308.
41. Gabsewicz JJ, Vial JP. Oligopoly a la cournot in general equilibrium analysis. *J Econ Theory* 1972;4:381–400.
42. Roberts K. The limit points of monopolistic competition. *J Econ Theory* 1980;22:256–278.

EFFECTIVE APPLICATION OF GRASP

PAOLA FESTA

Department of Mathematics
and Applications, University
of Napoli FEDERICO II,
Naples, Italy

MAURICIO G. C. RESENDE

Algorithms and Optimization
Research Department,
AT&T Labs Research,
Florham Park, New Jersey

INTRODUCTION

Many combinatorial optimization problems especially those found in industry and government are either computationally intractable by their nature, or sufficiently large so as to preclude the use of exact algorithms. In such cases, heuristic methods are usually employed to find good, but not necessarily guaranteed optimal, solutions. The effectiveness of these methods depends upon their ability to adapt to a particular realization, avoid entrapment at local optima, and exploit the basic structure of the problem. Building on these notions, various heuristic search techniques have been developed that have demonstrably improved our ability to obtain good solutions to difficult combinatorial optimization problems. The most promising of such techniques include simulated annealing [1], tabu search [2–4], genetic algorithms [5], variable neighborhood search (VNS) [6], and GRASP (greedy randomized adaptive search procedure) [7,8].

A GRASP is a multistart or iterative process [9], in which each GRASP iteration consists of two phases, a construction phase, in which a feasible solution is produced, and a local search phase, in which a local optimum in the neighborhood of the constructed solution is sought. The best overall solution is retained as the result.

In the GRASP construction phase, starting from an empty solution at each iteration

a new element is added to the partial solution in a greedy, randomized, and adaptive fashion. Let C be the set of element candidates to be inserted in the solution under construction. The greedy and random construction is blended either by using greediness to build a restricted candidate list ($RCL \subseteq C$) of the current best ranked elements and randomness to select an element from the list, or by using randomness to build the list and greediness for selection. Candidate elements $e \in C$ are sorted according to their greedy function value $g(e)$. In a *cardinality-based construction*, the RCL is made up by the k top-ranked elements. In a *value-based construction*, the RCL consists of the elements in the set $\{e \in C : g_m \leq g(e) \leq g_m + \alpha(g_M - g_m)\}$, where $g_m = \min\{g(e) : e \in C\}$, $g_M = \max\{g(e) : e \in C\}$, and $\alpha \in [0, 1]$. Since the best value for α is often difficult to determine, it is often assigned a random value for each GRASP iteration.

An especially appealing characteristic of GRASP is the ease with which it can be implemented. Few parameters need to be set and tuned, and therefore development can focus on implementing efficient data structures to assure quick GRASP iterations.

In this article, we briefly review the literature of operations research and computer science applications of GRASP as well as industrial applications. We cover applications in scheduling, routing and transportation, logic and partitioning, graph theory, location and assignment, manufacturing, telecommunications and power systems. The article concludes with a section on applications in biology and related fields.

SCHEDULING

GRASP has been applied to a wide variety of scheduling problems. One of the pioneering works illustrating a GRASP for scheduling is the paper of Bard and Feo [10], which appeared in 1989. This paper describes a method for efficiently sequencing cutting operations associated with the manufacture

of discrete parts. The problem is first modeled as an integer program and then relaxed via Lagrangian relaxation into a min-cut problem on a bipartite network. To obtain lower bounds, a max-flow algorithm is applied and the corresponding solution is input to a GRASP. Feo *et al.* [11] proposed a GRASP for single machine scheduling with sequence dependent setup costs and linear delay penalties. Here, the greedy function of the GRASP construction phase proposed is made up of two components: the switch-over cost and the opportunity cost associated with not inserting a specific job in the next position and instead, inserting it after half of the unscheduled jobs have been scheduled. This greedy function tends to lead to a balance between the natural order and nearest neighbor approaches. The local search uses 2-exchange, insertion exchange, and a combination of the two. Another interesting application of GRASP for solving real-world scheduling problems is due to Xu and Chiu [12], who proposed an effective heuristic procedure for a field technician scheduling problem, where the objective is to assign a set of jobs at different locations with time windows to technicians with different job skills. Several heuristics, including a GRASP, are designed and tested for solving the problem. The greedy choice is to select jobs with the highest unit weight. The local search implements four different moves, among them the 2-exchange and a swap that exchanges an assigned job with another job unassigned under the candidate schedule.

Several hybrid approaches have been also proposed that combine GRASP with other techniques, including tabu search for the school timetabling problem [13], path relinking for the job shop scheduling [14] and branch-and-bound (B&B) for the job shop scheduling [15], and a scheduling problem with nonrelated parallel machines and sequence-dependent setup times [16], where at each GRASP iteration, the greedy criterion adopted to build a feasible solution consists in sorting the jobs in a nondecreasing order using the due date and then assigning each job to the machine that can finish it first. In 2008, two further

interesting hybrid techniques had been proposed. Alvarez-Valdes *et al.* [17], describe a reactive GRASP for a strip-packing problem, while they, in [18], propose several heuristic algorithms based on GRASP and path relinking for project scheduling under partially renewable resources. In this case, the greedy selection is based on a priority rule and the improvement phase consists of two steps. First, it identifies the activities whose completion times must be reduced in order to have a new solution with the shortest makespan and labeling these activities as critical. Second, it moves critical activities in such a way that the resulting sequence is feasible according to precedence and resource constraints. In particular, the authors have designed two types of moves: (i) in a simple move, only a critical activity is moved, leaving the remaining activities unchanged; (ii) in a double move, noncritical activities are moved to make the move of a critical activity possible.

ROUTING AND TRANSPORTATION

GRASP has been applied to several vehicle, aircraft, telecommunications, and inventory routing problems. In 1995, Kontoravdis and Bard [19] proposed a GRASP for minimizing the fleet size of temporarily constrained vehicle routing problems with two types of service. The greedy function of the construction phase takes into account both the overall minimum insertion cost and the penalty cost. Local search is applied to the best solution found every five iterations of the first phase, rather than to each feasible solution. In 1998, a modified and improved version of this GRASP was extended for the inventory routing problem with satellite facilities by Bard *et al.* [20], who proposed a methodology that decomposes the problem over the planning horizon, and then solves daily rather than multiday vehicle routing problems.

In 2002, Carreto and Baker [21] investigated the incorporation of interactive tools into heuristic algorithms and proposed a GRASP interactive approach to the vehicle routing problem with backhauls. A GRASP

has been used in the routes construction and improvement phase. The construction phase is implemented using a clustering heuristic that constructs the routes by clustering the remaining customers according to the vehicles defined by seeds while applying the 3-opt heuristic to reduce the total distance traveled by each vehicle. The greedy function takes into account routes with smallest insertion cost and customers with biggest difference between the smallest and the second smallest insertion costs and smallest number of routes they can traverse.

Corberán *et al.* [22] proposed a pure GRASP for the mixed Chinese postman problem, while a hybrid GRASP with tabu search has been proposed by Li *et al.* [23] for vehicle routing with both time window and limited number of vehicles. To efficiently solve the rural postman problem, de la Peña [24] published, in 2004, an interesting case study to compare several metaheuristics approaches, including a simulated annealing, a GRASP, and a genetic algorithm.

GRASP has been used also to find approximate solutions of problems in air, rail, and intermodal transportation. One among the pioneering papers illustrating these applications appeared in 1989, due to Feo and Bard [25], who studied the problem of locating maintenance stations and developing flight schedules that better meet the cyclical demand for maintenance. The problem has been formulated as a large-scale mixed integer program, that is, a minimum cost, multicommodity flow network with integral constraints, where each airplane represents a separate commodity and each arc has an upper and lower capacity of flow. Since obtaining feasible solutions from the relative LP relaxation is difficult, the authors proposed a GRASP. Another interesting paper is due to Argüello *et al.* [26], who described a GRASP to reconstruct aircraft routings in response to groundings and delays experienced over the course of the day. The objective is to minimize the cost of reassigning aircraft to flights taking into account available resources and other system constraints. The proposed GRASP heuristic includes a neighborhood search

technique that takes as input an initial feasible solution, so that the “classical” GRASP construction phase is omitted. Instead, in the procedure, the neighbors of an incumbent solution are generated and evaluated, and the most desirable are placed on an RCL, from which one is randomly selected and becomes the incumbent. Two types of partial route exchange operations are described. The first type is the exchange of flight sequences with identical end points. In the second type, the sequence of flights being exchanged must have the same origination airport, but the termination airports are swapped.

LOGIC, PARTITIONING, AND GRAPH THEORETIC APPLICATIONS

GRASP has been applied to problems in logic, including SAT, MAX-SAT, and logical clause inference. In 1996, Resende and Feo [27] proposed a GRASP for the satisfiability problem that can be applied to both the weighted and unweighted versions of the maximum satisfiability problem. The adaptive greedy function is a hybrid combination of two functions that seek to maximize the number of yet-unsatisfied clauses that become satisfied after the assignment of each construction iteration, and maximize the number of yet-unassigned literals in yet-unsatisfied clauses that become satisfied if opposite assignments were to be made. The local search flips the assignment of each variable, one at a time, checking if the new truth assignment increases the number of satisfied clauses. In 2006, Festa *et al.* [28] proposed a hybrid GRASP with path relinking for the weighted maximum satisfiability problem (MAX-SAT). The authors designed accurate experiments to determine the effect of path relinking on the convergence of the GRASP.

Partitioning problems have been also treated with GRASP. These include number partitioning [29] and software/hardware partitioning [30]. In particular, in Argüello *et al.* [29], randomized methodologies are described for partitioning a finite set of integers into two disjoint subsets such that

the difference of the sums of the elements in the subsets is minimized. Here, the greedy criterion consists in considering only large elements for differencing. In Pu *et al.* [30], the authors propose a hybrid GRASP with tabu search as local search for software/hardware partitioning within timed automata.

Perhaps the largest class of problems for which GRASP has been applied is graph theory. The wide applicability of GRASP to these problems includes among others the maximum independent set problem, the maximum covering problem, the Steiner tree problem, the max cut problem, the feedback vertex set and arc set problems, graph coloring, and the multicriteria minimum spanning tree problem.

For the maximum independent set, in 1994, Feo *et al.* [31] proposed a GRASP whose greedy function orders admissible vertices with respect to the minimum admissible vertex degree, that is, a vertex with the minimum degree and that is not adjacent to any vertex in the current independent set. The neighborhood definition used in the local search is (2,1)-exchange, where two nonadjacent vertices can be added to the current solution if a single vertex from the solution is removed. Resende [32] describes a GRASP for the maximum covering problem. Here, the greedy function is the total weight of the yet-uncovered demand points that become covered after the selection and the local search procedures uses a 2-exchange neighborhood structure.

Martins *et al.* [33] proposed a GRASP for approximately solving general instances of the Steiner problem in graphs. The construction phase is based on the distance network heuristic. A distance network corresponding to the original graph is built and associated with each edge of the distance network is a weight that takes into account the shortest distances in the original graph. With respect to the new weight distribution, Kruskal's algorithm is used to solve the minimum spanning tree problem and the edges in the minimum spanning tree so computed are replaced by the edges in the corresponding shortest paths in the original graph. The local search is based on insertions and eliminations of nodes to and from the current

solution. A parallel version of a GRASP for the Steiner tree problem has been described in Martins *et al.* [34], where the authors propose a local search that uses a combination of key path-based local search and node-based local search.

For the max cut problem, in 2002, Festa *et al.* [35] designed and tested a GRASP, a VNS, a path relinking heuristic, and new hybrid heuristics that combine GRASP, VNS as GRASP local search, and path relinking.

Pardalos *et al.* [36] designed a GRASP for finding approximate solutions to the feedback vertex set problem on a digraph. Several greedy functions were tested, all of them favoring vertices with high degree, while the local search procedure tries at each iteration to eliminate redundant vertices. Festa *et al.* [37] described a set of ANSI standard Fortran 77 subroutines to find approximate solutions of both the feedback vertex set problem and the feedback arc set problem. The GRASP of Pardalos *et al.* is used to produce the approximate solutions of the feedback set problem, while feedback arc set problems are converted into feedback vertex set problems and then solved.

For graph coloring, in 2001, Laguna and Martí [38] proposed a GRASP, whose construction phase greedy function chooses the vertex having the maximum degree among the uncolored vertices adjacent to at least one colored vertex. The local search combines the two smallest cardinality color classes into one and tries to find a valid color for each violating vertex.

Very recently, for the multicriteria minimum spanning tree problem, in 2008, Arroyo *et al.* [39] proposed a GRASP, whose local search tries to improve the current solution by defining a drop-and-add neighborhood, where the spanning trees are represented by Prufer numbers.

LOCATION AND ASSIGNMENT PROBLEMS

GRASP has been successfully used to find approximate solutions to location and assignment problems.

Already in 1992, Klincewicz [40] proposed two heuristics based on tabu search and

GRASP for the p -hub location problem, whose local search procedure is based on the 2-exchange neighborhood. Later, in 2005, Pérez *et al.* [41] described a hybrid GRASP with the path-relinking algorithm for the capacitated p -hub median problem. In this proposal, the greedy evaluation function has two phases: a location phase and an allocation phase. In the location phase, one element is added at each time to the set of hubs, while the allocation phase consists of an allocation to the nearest hub. As local search, the authors proposed a greedy procedure for location and an allocation to the nearest hub. For the p -median and other location problems, efficient multistart hybrid heuristics that combine elements of several traditional metaheuristics including GRASP have been recently proposed by Resende and Werneck [42,43]. Here, the greedy criterion consists in choosing, at each iteration, one among the most profitable facilities, while the local search procedure is based on swapping facilities.

Starting with the pioneering work of Li *et al.* [44], GRASP has been applied to several hard assignment problems, including quadratic, biquadratic, and multidimensional assignment. For the quadratic assignment problem (QAP), Li *et al.* [44] describe a GRASP, whose construction mechanism defines a greedy function on the assignment interaction cost and whose local search procedure is a 2-assignment exchange. The construction phase of the GRASP proposed by Li *et al.* has been later applied by Ahuja *et al.* [45] to generate the initial population of a genetic algorithm for the QAP. In 1998, Mavridou *et al.* [46] proposed a GRASP for finding approximate solutions of the biquadratic assignment problem. As in the case of GRASP for the QAP, the construction phase has two stages. The first stage simultaneously makes four assignments, selecting the pairs corresponding to the smallest interaction costs, while the second stage makes the remaining assignments, one at time. The greedy function in the second stage selects the assignment corresponding to the minimum interaction cost with respect to the already-made assignments. In the local search phase, 2-exchange is applied to the

permutation constructed in the first phase. An interesting study has been conducted, in 1999, by Fleurent and Glover [47]. In their paper, they showed that the GRASP for QAP of Li *et al.* [44] can be improved upon by using memory strategies, such as learning, intensification, candidate list strategies, and the Proximate Optimally Principle (POP). For the multidimensional assignment problem, Robertson [48], in 2001, proposed four GRASP implementations by combining two construction methods (randomized greedy and randomized max regret) and two local search methods (2-exchange and variable depth exchange). The greedy function of the randomized max regret construction method is a measure of the competition between the two leading cost candidates, while the maximum regret value corresponds to the candidate assignment that has the largest winning margin between itself and its next highest competitor. The variable depth exchange is an extension of the 2-exchange method that allows a more extensive search of the surrounding neighborhood space.

MANUFACTURING

GRASP has been applied to solve several optimization problems arising in manufacturing. The first paper that is an example of this kind of application is due to Feo and Bard [49], who, in 1989, proposed a method for minimizing the sum of tool setup and volume removal times associated with metal cutting operations on a flexible machine. The problem was modeled as an integer program and relaxed into a min-cut problem on a simple network. After obtaining a tentative solution, the authors used a GRASP to identify good feasible points corresponding to alternative process plans. The same authors [50], in 1991, designed an algorithm for the manufacturing equipment selection problem, where the objective is to determine how many of each machine type to purchase and for what fraction of the time each piece of equipment will be configured for a particular type of operation. The problem has been transformed into a mixed integer linear program (MILP) and a depth-first branch and bound algorithm has been used, employing the greedy

randomized set covering heuristic of Feo and Resende [7].

More recently, for constrained two-dimensional nonguillotine cutting problems, Alvarez-Valdes *et al.* [51], in 2005, proposed a GRASP, whose construction mechanism takes the smallest rectangle breaking the ties by the nearest distance to a corner of the stock rectangle. Then, two criteria have been considered to select the piece: (i) the first piece in an ordered list, giving priority to pieces that must be cut; (ii) the piece producing the largest increase in the objective function. In 2008, Monkman *et al.* [52] described a production scheduling heuristic for an electronics manufacturer with sequence-dependent setup costs. The problem studied here involves assignment, sequencing, and time-scheduling steps, and has been modeled as a traveling salesman subset-tour problem. For the sequencing step, the authors proposed a GRASP, whose construction phase uses a cardinality-based RCL and the greedy function takes into account the cost associated with the arcs of the underlying graph. The local search uses two different neighborhoods: a node elimination and a node swap neighborhood.

TELECOMMUNICATIONS

Telecommunications and network design are two fields of quite a large number of applications of GRASP. In the following, only some recent papers will be cited.

Liu *et al.* [53], in 2000, proposed a hybrid GRASP for frequency assignment in mobile radio networks, whose local search is a simulated annealing and whose construction phase uses two greedy functions. The first chooses a vertex from the set of unselected vertices with high saturation degrees, while the second function is used to assign a frequency to the selected vertex. Prais and Ribeiro [54] designed a GRASP for a matrix decomposition problem arising in the context of traffic assignment in communication satellites. The local search phase of the GRASP proposed is based on a new neighborhood, defined by constructive and destructive moves. In 2003, Amaldi *et al.* [55] proposed

a GRASP and a tabu search to efficiently solve the downlink base station (BS) location problem. Here, during the local search, the following moves are considered: removing a BS, installing a new BS, removing an existing BS, and installing a new one (swap). The output GRASP solution is used as initial solution for a tabu search algorithm. In 2008, Commander *et al.* [56] designed and implemented a GRASP for the cooperative communication problem on *ad hoc* networks, where the problem consists in maximizing the amount of connectivity among a set of users, subject to constraints on the maximum distance traveled, as well as restrictions on what types of movement can be performed. The greedy function value of each candidate element is a measure of additional connections created by its insertion in the partial solution under construction. The local search procedure is based on a perturbation function consisting of selecting a wireless agent and rerouting.

APPLICATIONS IN BIOLOGY

It is only in the past few years that it has been shown that a large number of molecular biology problems can be formulated as combinatorial optimization problems, usually computationally intractable so as to employ heuristic methods to find good solutions in a reasonable amount of running time. Brown *et al.* [57] proposed a GRASP for selecting a population subset for use in a high-density genetic mapping project. At each iteration of the construction phase, the authors add to the partial solution, one among the r unchosen population members that most improve the objective function value. The authors investigated very small-sized RCL (i.e., $r = 3$ and $r = 5$) and the local search they implemented removes from the current solution some members and greedily includes other members, until no further improving exchange can be done.

For the phylogeny problem (consisting in finding a phylogeny with the minimum number of evolutionary steps—the so-called parsimony criterion), Andreatta and Ribeiro [58] and later Ribeiro and Vianna [59] proposed

a GRASP, and a VNS and a GRASP, which uses variable neighborhood descent for local search, respectively.

Phylogenetic footprints are short pieces of no-coding DNA sequences in genes that are conserved between evolutionarily distant species. Fried *et al.* [60] showed that solving the footprint sorting problem requires the solution of a minimum weight vertex feedback set problem and for solving the latter, they used the GRASP provided in Festa *et al.* [37].

In 2007, Festa [61] proposed a GRASP to find improved solutions for the far from most string problem, where the objective is to determine a sequence far from most of the sequences in a given input set. In this case, it is intuitive to relate the greedy function to the occurrence of each character in a given position, while to perform the local search phase the 2-exchange algorithm has been used. A solution for the far from most string problem can help, for example, to identify a sequence fragment that distinguishes the pathogens from the host, so that the potential exists to create a drug that harms several but not all pathogens.

REFERENCES

1. Kirkpatrick S. Optimization by simulated annealing: quantitative studies. *J Stat Phys* 1984;34:975–986.
2. Glover F. Tabu search—Part I. *ORSA J Comput* 1989;1:190–206.
3. Glover F. Tabu search—Part II. *ORSA J Comput* 1990;2:4–32.
4. Glover F, Laguna M. Tabu search. Norwell (MA): Kluwer Academic Publishers; 1997.
5. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading (MA): Addison-Wesley; 1989.
6. Hansen P, Mladenović N. An introduction to variable neighborhood search. In: Voss S, Martello S, Osman IH, *et al.*, editors. Meta-heuristics, advances and trends in local search paradigms for optimization. Norwell (MA): Kluwer Academic Publishers; 1998. pp. 433–458.
7. Feo TA, Resende MGC. A probabilistic heuristic for a computationally difficult set covering problem. *Oper Res Lett* 1989;8:67–71.
8. Feo TA, Resende MGC. Greedy randomized adaptive search procedures. *J Glob Optim* 1995;6:109–133.
9. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling-salesman problem. *Oper Res* 1973;21:498–516.
10. Bard JF, Feo TA. Operations sequencing in discrete parts manufacturing. *Manage Sci* 1989;35:249–255.
11. Feo TA, Sarathy K, McGahan J. A GRASP for single machine scheduling with sequence dependent setup costs and linear delay penalties. *Comput Oper Res* 1996;23:881–895.
12. Xu JY, Chiu SY. Effective heuristic procedure for a field technician scheduling problem. *J Heuristics* 2001;7:495–509.
13. Souza MJF, Maculan N, Ochi LS. A GRASP-tabu search algorithm for solving school timetabling problems. Metaheuristics: computer decision-making. Norwell (MA): Kluwer Academic Publishers; 2004. pp. 659–672.
14. Aiex RM, Binato S, Resende MGC. Parallel GRASP with path-relinking for job shop scheduling. *Parallel Comput* 2003;29:393–430.
15. Fernandes S, Lourenço HR. A GRASP and branch-and-bound metaheuristic for the job-shop scheduling. Volume 4446, Evolutionary computation in combinatorial optimization, Lecture notes in computer science. New York (NY): Springer; 2007. pp. 60–71.
16. Rocha PL, Ravetti MG, Mateus GR. The metaheuristic GRASP as an upper bound for a branch and bound algorithm in a scheduling problem with non-related parallel machines and sequence-dependent setup times. Proceedings of the 4th EU/ME Workshop: Design and Evaluation of Advanced Hybrid Meta-Heuristics, Volume 1. Nottingham (UK); 2004. pp. 62–67.
17. Alvarez-Valdes R, Parreño F, Tamarit JM. Reactive GRASP for the strip-packing problem. *Comput Oper Res* 2008;35(4):1065–1083.
18. Alvarez-Valdes R, Crespo E, Tamarit JM, *et al.* GRASP and path relinking for project scheduling under partially renewable resources. *Eur J Oper Res* 2008;189(3):1153–1160.
19. Kontoravdis G, Bard JF. A GRASP for the vehicle routing problem with time windows. *ORSA J Comput* 1995;7:10–23.
20. Bard JF, Huang L, Jaillet P, *et al.* A decomposition approach to the inventory routing

- problem with satellite facilities. *Transp Sci* 1998;32:189–203.
21. Carreto C, Baker B. A GRASP interactive approach to the vehicle routing problem with backhauls. In: Ribeiro CC, Hansen P, editors. *Essays and surveys on metaheuristics*. Norwell (MA): Kluwer Academic Publishers; 2002. pp. 185–200.
 22. Corberán A, Martí R, Sanchis JM. A GRASP heuristic for the mixed Chinese postman problem. *Eur J Oper Res* 2002;142(1):70–80.
 23. Li Z, Guo S, Wang F, *et al.* Improved GRASP with tabu search for vehicle routing with both time window and limited number of vehicles. In: Orchard B, Yang C, Ali M, editors. *Volume 3029, Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2004)*, Lecture notes in computer science. New York (NY): Springer; 2004. pp. 552–561.
 24. de la Peña MGB. Heuristics and metaheuristics approaches used to solve the rural postman problem: a comparative case study. *Proceedings of the Fourth International ICSC Symposium on Engineering of Intelligent Systems (EIS 2004)*. Island of Maderia, Portugal; 2004.
 25. Feo TA, Bard JF. Flight scheduling and maintenance base planning. *Manage Sci* 1989;35:1415–1432.
 26. Argüello MF, Bard JF, Yu G. A GRASP for aircraft routing in response to groundings and delays. *J Comb Optim* 1997;1:211–228.
 27. Resende MGC, Feo TA. A GRASP for satisfiability. In: Johnson DS, Trick MA, editors. *Volume 26, Cliques, coloring, and satisfiability: the second DIMACS implementation challenge, DIMACS series on discrete mathematics and theoretical computer science*. Providence (RI): American Mathematical Society; 1996. pp. 499–520.
 28. Festa P, Pardalos PM, Pitsoulis LS, *et al.* GRASP with path-relinking for the weighted maxsat problem. *ACM J Exp Algorithm* 2006;11:1–16.
 29. Argüello MF, Feo TA, Goldschmidt O. Randomized methods for the number partitioning problem. *Comput Oper Res* 1996;23(2):103–111.
 30. Pu GG, Chong Z, Qiu ZY, *et al.* A hybrid heuristic algorithm for HW-SW partitioning within timed automata. *Volume 4251, Proceedings of Knowledge-based Intelligent Information and Engineering Systems, Lecture notes in artificial intelligence*. New York (NY): Springer; 2006. pp. 459–466.
 31. Feo TA, Resende MGC, Smith SH. A greedy randomized adaptive search procedure for maximum independent set. *Oper Res* 1994;42:860–878.
 32. Resende MGC. Computing approximate solutions of the maximum covering problem using GRASP. *J Heuristics* 1998;4:161–171.
 33. Martins SL, Pardalos PM, Resende MGC, *et al.* Greedy randomized adaptive search procedures for the Steiner problem in graphs. In: Pardalos PM, Rajasekaran S, Rolim J, editors. *Volume 43, Randomization methods in algorithmic design, DIMACS series on discrete mathematics and theoretical computer science*. Providence (RI): American Mathematical Society; 1999. pp. 133–145.
 34. Martins SL, Resende MGC, Ribeiro CC, *et al.* A parallel GRASP for the Steiner tree problem in graphs using a hybrid local search strategy. *J Glob Optim* 2000;17:267–283.
 35. Festa P, Pardalos PM, Resende MGC, *et al.* Randomized heuristics for the MAX-CUT problem. *Optim Methods Softw* 2002;7:1033–1058.
 36. Pardalos PM, Qian T, Resende MGC. A greedy randomized adaptive search procedure for the feedback vertex set problem. *J Comb Optim* 1999;2:399–412.
 37. Festa P, Pardalos PM, Resende MGC. Algorithm 815: FORTRAN subroutines for computing approximate solution to feedback set problems using GRASP. *ACM Trans Math Softw* 2001;27:456–464.
 38. Laguna M, Martí R. A GRASP for coloring sparse graphs. *Comput Optim Appl* 2001;19:165–178.
 39. Arroyo JEC, Vieira PS, Vianna DS. A GRASP algorithm for the multi-criteria minimum spanning tree problem. *Ann Oper Res* 2008;159:125–133.
 40. Klincewicz JG. Avoiding local optima in the p -hub location problem using tabu search and GRASP. *Ann Oper Res* 1992;40:283–302.
 41. Pérez M, Almeida F, Moreno-Vega JM. A hybrid GRASP-path relinking algorithm for the capacitated p -hub median problem. *Volume 3636, Hybrid metaheuristics, Lecture notes in computer science*. New York (NY): Springer; 2005. pp. 142–153.
 42. Resende MGC, Werneck RF. A hybrid heuristic for the p -median problem. *J Heuristics* 2004;10:59–88.

43. Resende MGC, Werneck RF. A fast swap-based local search procedure for location problems. *Ann Oper Res* 2007;150:205–230.
44. Li Y, Pardalos PM, Resende MGC. A greedy randomized adaptive search procedure for the quadratic assignment problem. In: Pardalos PM, Wolkowicz H, editors. Volume 16, Quadratic assignment and related problems, DIMACS Series on Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 1994. pp. 237–261.
45. Ahuja RK, Orlin JB, Tiwari A. A greedy genetic algorithm for the quadratic assignment problem. *Comput Oper Res* 2000;27:917–934.
46. Mavridou T, Pardalos PM, Pitsoulis LS, *et al.* A GRASP for the biquadratic assignment problem. *Eur J Oper Res* 1998;105:613–621.
47. Fleurent C, Glover F. Improved constructive multistart strategies for the quadratic assignment problem using adaptive memory. *INFORMS J Comput* 1999;11:198–204.
48. Robertson AJ. A set of greedy randomized adaptive local search procedure (GRASP) implementations for the multidimensional assignment problem. *Comput Optim Appl* 2001;19:145–164.
49. Feo TA, Bard JF. The cutting path and tool selection problem in computer-aided process planning. *J Manuf Syst* 1989;8:17–26.
50. Bard JF, Feo TA. An algorithm for the manufacturing equipment selection problem. *IIE Trans* 1991;23:83–92.
51. Alvarez-Valdes R, Parreño F, Tamarit JM. A GRASP algorithm for constrained two-dimensional non-guillotine cutting problems. *J Oper Res Soc* 2005;56(4):414–425.
52. Monkman SK, Morrice DJ, Bard JF. A production scheduling heuristic for an electronics manufacturer with sequence-dependent setup costs. *Eur J Oper Res* 2008;187(3):1100–1114.
53. Liu X, Pardalos PM, Rajasekaran S, *et al.* A GRASP for frequency assignment in mobile radio networks. In: Rajasekaran S, Pardalos PM, Hsu F, editors. Volume 52, Mobile networks and computing, DIMACS Series on Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 2000. pp. 195–201.
54. Prais M, Ribeiro CC. Reactive GRASP: an application to a matrix decomposition problem in TDMA traffic assignment. *INFORMS J Comput* 2000;12:164–176.
55. Amaldi E, Capone A, Malucelli F, *et al.* Optimization models and algorithms for downlink umts radio planning. *Proceedings of Wireless Communications and Networking, (WCNC 2003), Volume 2*. New Orleans (LA); 2003. pp. 827–831.
56. Commander C, Festa P, Oliveira CAS, *et al.* A greedy randomized algorithm for the cooperative communication problem on ad hoc networks. In: Pardalos PM, Prokopyev OA, Grundel DA, editors. *Cooperative networks: control and optimization*. Northampton (MA): Edward Elgar Publishing; 2008. pp. 187–207.
57. Brown DG, Vision TJ, Tanksley SD. Selecting mapping: a discrete optimization approach to select a population subset for use in a high-density genetic mapping project. *Genetics* 2000;155:407–420.
58. Andreatta AA, Ribeiro CC. Heuristics for the phylogeny problem. *J Heuristics* 2002;8:429–447.
59. Ribeiro CC, Vianna DS. A GRASP/VND heuristic for the phylogeny problem using a new neighborhood structure. *Int Trans Oper Res* 2005;12:325–338.
60. Fried C, Hordijk W, Prohaska SJ, *et al.* The footprint sorting problem. *J Chem Inform Comput Sci* 2004;44(2):332–338.
61. Festa P. On some optimization problems in molecular biology. *Math Biosci* 2007;207(2):219–234.

EFFECTIVE APPLICATION OF GUIDED LOCAL SEARCH

CHRISTOS VOUDOURIS
EDWARD P. K. TSANG
ABDULLAH ALSHEDDY
School of Computer Science and
Electronic Engineering,
University of Essex, Colchester,
UK

Guided local search (GLS) is a penalty-based approach that sits on top of local search (LS) methods to bring them out of local minima. GLS heavily employs information about the search history and the structure of solutions to distribute the search effort in the various regions of the search space. The procedure of GLS is prescriptive and relatively simple to implement. As it is the case with other metaheuristics, it requires an LS method to be readily available for the problem at hand. This is both, in terms of a way of producing an initial solution either randomly or heuristically and also in terms of a local improvement procedure which can iteratively optimize a given solution.

To assist with implementation of the technique, we provide the detailed pseudocode for

the main GLS algorithm. We also detail the two related methods of FLS and guided fast local search (GFLS) which can improve the running times of LS (in the case of FLS) and integrate tighter LS with GLS (in the case of GFLS). The set of techniques presented constitute the foundations of the GLS approach and they can be applied with little modifications to a diverse range of problems from routing and scheduling to assignment problems and constraint optimization, including large-scale instances and real-world applications. Suggestions and insights on the practical implementation of the methods in software are provided and discussed.

One important element in the effective application of GLS to a specific problem is the definition of features and their costs so that GLS can effectively guide LS and efficiently distribute the search effort. We pay special attention to this subject, providing examples of possible features to use to apply GLS to common classes of problems.

IMPLEMENTING GUIDED LOCAL SEARCH

We start our detailed presentation of the techniques with the main GLS algorithm. The pseudocode for the algorithm is provided below:

```
procedure GuidedLocalSearch( $p, g, \lambda, [I_1, \dots, I_M], [c_1, \dots, c_M], M$ )  
begin  
   $k \leftarrow 0$ ;  
   $s_0 \leftarrow \text{ConstructionMethod}(p)$ ;  
  /* set all penalties to 0 */  
  for  $i \leftarrow 1$  until  $M$  do  
     $p_i \leftarrow 0$ ;  
  /* define the augmented objective function */  
   $h \leftarrow g + \lambda * \sum p_i * I_i$ ;  
  while NOT StoppingCriterion do  
    begin  
       $s_{k+1} \leftarrow \text{ImprovementMethod}(s_k, h)$ ;  
      /* compute the utility of features */  
      for  $i \leftarrow 1$  until  $M$  do  
         $\text{util}_i \leftarrow I_i(s_{k+1}) * c_i / (1 + p_i)$ ;  
      /* penalize features with maximum utility */
```

```

for each  $i$  such that  $util_i$  is maximum do
     $p_i \leftarrow p_i + 1$ ;
     $k \leftarrow k+1$ ;
end
 $s^* \leftarrow$  best solution found with respect to objective function  $g$ ;
return  $s^*$ ;
end

```

where p is the problem, g the objective function, h the augmented objective function, λ the lambda parameter, I_i the indicator function of feature i , c_i the cost of feature i , M the number of features, p_i the penalty of feature i , *ConstructionMethod*(p) the method for constructing a initial solution for problem p , and *ImprovementMethod*(s_k, h) the method for improving solution s_k according to the augmented objective function h .

Two external methods, usually part of a typical LS procedure, are invoked from within the GLS algorithm, namely the *ConstructionMethod* and the *ImprovementMethod*. GLS starts the searching process from an initial solution. This is generated by the *ConstructionMethod*. The solution can be generated either randomly or heuristically, based on some known techniques for constructing solutions for the particular problem. We found that GLS is not very sensitive to the starting solution; provided sufficient time is available to the algorithm to explore the search space of the problem. A method for improving the solution is also required. The *ImprovementMethod* can be a simple LS algorithm or a more sophisticated one such as Variable Neighborhood Search [1], Variable Depth Search [2], Ejection Chains [3] or combinations of LS methods with exact search algorithms [4]. The reader should note here that the *ImprovementMethod* uses the augmented objective function (h) rather than the original one (g) as also shown in the pseudocode above.

It is not essential for the improvement method to generate high quality local minima. Experiments with GLS and various TSP heuristics reported in Ref. 5 have shown that high quality local minima take time to produce, resulting in less intervention by GLS in the overall allocated search time. This may sometimes lead to inferior results compared

to a simpler but more computationally efficient improvement method.

Apart from the construction method and the improvement method mentioned above, which are specific to the problem at hand, the rest of the algorithm presented above is generic and rather straightforward. We provide below suggestions and insights on how to efficiently implement the various key components of the algorithm.

Utility Function

The penalties of features are initialized to zero. At each iteration, the improvement method is called which will return a local minimum. Note here that the solution returned may in fact not be a “true” local minimum (e.g., because of penalties introduced by GLS or limited neighbors considered by the improvement method) but for the sake of simplicity, we will refer to it as one. At this local minimum, GLS will have to select which features to penalize to help the LS to escape. The utility function is defined for that purpose. The utility of each feature exhibited in the returned local minimum is calculated, and those features with maximum utility are penalized by incrementing their penalty values.

The indicator functions I_i that define the features rarely need to be implemented in software code. Looking at the values of the decision variables, one can directly identify the features present in a local minimum. When this is not possible, data structures with constant time deletion/addition operations (e.g. based on double-linked lists) can incrementally maintain the set of features present in the working solution, avoiding the need for an expensive computation to detect features from scratch every time GLS reaches a local minimum.

The selection of features to penalize can also be efficiently implemented by using the same loop for computing the utility formula for features present in the local minimum (the rest can be ignored) and also placing features with maximum utility in a vector. With a second loop, the features with maximum utility contained in this vector have their penalties incremented by one.

λ Parameter

The parameter λ is the only parameter to the GLS method (at least in its basic version) and in general, it is instance dependent and more specifically linked to the values that objective function takes as the search moves within the landscape of the problem. It has been observed that, for several problems, a good value for λ can be calculated by dividing the value of the objective function in a local minimum with the average number of features present. In these problems, λ is dynamically computed after the first local minimum and before penalizing features for the first time. Providing an α parameter, which is relatively instance-independent, λ is calculated by the following formula:

$$\lambda = \alpha \times g(s^*)/Fs^*, \quad (1)$$

where g is the objective function of the problem, s^* is a local minimum, and Fs^* is the number of features present in s^* .

Tuning α can result in λ values, which work for many instances of a problem class. Once tuned, α can be fixed in industrialized

versions of software, resulting in a ready-to-use GLS algorithm for the end user.

Note here that depending on the value of λ (or α), several penalty cycles may be required before a move is executed out of a local minimum. This should not be viewed as an undesirable situation and it should not motivate using high λ (or α) values to quickly escape from local minima. The very uncertainty in the information used in guiding the search (e.g., feature costs) makes necessary the testing of the GLS decisions against the landscape of the problem in a gradual manner. Smaller penalty values tend to help GLS find the “best” escape route out of local minima without unnecessarily distorting the landscape. This approach is illustrated in Fig. 1.

Augmented Objective Function

It is important to mention that although the search in GLS uses the augmented objective function, the program should keep track of the actual objective value in all operations relating to storing the best solution or finding a new best solution. This is not limited to the solutions accepted by LS but also includes those evaluated. The reason is that LS uses the augmented objective function and, hence, a move that generates a new best solution may be missed. To the contrary, keeping track of the value of the augmented objective value (e.g., after adding the penalties) is not necessary since LS methods will be looking only at the differences in the augmented objective value when evaluating LS moves.

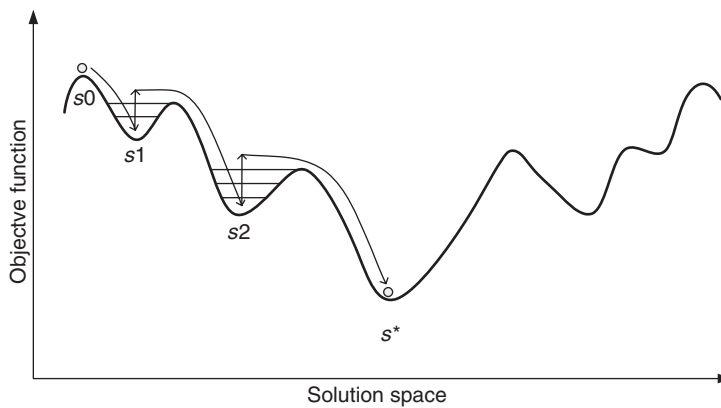


Figure 1. GLS is escaping from a valley (i.e., local minimum) in the landscape by increasing the objective function value for its solutions. This is conducted in a gradual and focused manner so that search is guided toward local minima of higher quality.

However, the move evaluation mechanism needs to be revised to work efficiently with the augmented objective function.

Normally, the move evaluation mechanism does not directly evaluate the objective value of the new solution generated by the move. Instead, it calculates the difference Δg in the objective function. This difference should be combined with the difference in the amount of penalties. This can be easily done and has no significant impact on the time needed to evaluate a move. In particular, we have to take into account only features that change state (being deleted or added). The penalties of the features deleted are summed together. The same is done for the penalties of features added. The change in the amount of penalties due to the move is then simply given by the difference

$$\sum_{\text{over all features } j \text{ added}} p_j - \sum_{\text{over all features } k \text{ deleted}} p_k, \quad (2)$$

which then has to be multiplied by λ and added to Δg .

Stopping Criterion

There are many choices possible for the *StoppingCriterion* of the algorithm. Since GLS is not trapped in local minima, there is no obvious point when to stop the algorithm. Like in other metaheuristics, a measure related to the length of the search process is usually used. A limit on the number of moves performed, the number of moves evaluated, or the CPU time spent by the algorithm are examples of such a

measure. A lower bound can be utilized as a stopping criterion, if it is known, by setting the excess above the lower bound that we require to be achieved. Criteria can also be combined to allow for a more flexible way to stop the GLS method. The idea of threshold value was proposed as a termination condition for every GLS stage in Ref. 6.

IMPLEMENTING FAST LOCAL SEARCH

FLS is designed to reduce the running times of LS algorithms. The idea of FLS is to speed up neighborhood search by exploiting historical information that would help to ignore parts of the neighborhood which are unlikely to yield better solutions. The approach is similar to candidate list strategies as used in Tabu Search [3] and other metaheuristics.

In FLS, the problem's neighborhood is broken down into a number of subneighborhoods and an activation bit is attached to each one of them (bit_i). The idea is to examine subneighborhoods in a given order, searching only those with the activation bit set to one. The neighborhood search process does not restart whenever we find a better solution but it continues with the next subneighborhood in the given order. The initial values of the bits are given as an input to the procedure, which allows the procedure to be focused on certain subneighborhoods and not the whole neighborhood, which may be a large one. The detailed pseudocode for FLS is provided below.

procedure FastLocalSearch($s, h, [bit_1, \dots, bit_L], L$)
begin

```

  while  $\exists j, bit_j = 1$  do
    /* i.e. while active sub-neighborhood exists */
    for  $i \leftarrow 1$  until  $L$  do
      begin
        if  $bit_i = 1$  then
          /* search sub-neighborhood  $i$  */
          begin
            Moves  $\leftarrow$  MovesForSubneighborhood( $i$ );
            for each move  $m$  in Moves do
              begin
                 $s' \leftarrow m(s)$ ;

```

```

/* s' is the result of move m */
if h(s') < h(s) then
/* minimization case is assumed here */
begin
  /* spread activation */
  ActivateSet ← SubneighborhoodsForMove(m);
  for each sub-neighborhood j
  in ActivateSet do
    bitj ← 1;
    s ← s';
  goto ImprovingMoveFound
end
end
biti ← 0; /* no improving move found */
end
ImprovingMoveFound:
  continue;
end;
return s;
end

```

where s is the solution, h the augmented objective function, L the number of subneighborhoods, bit_i the activation bit for subneighborhood i , *MovesForSubneighborhood*(i) the method which returns the set of moves contained in subneighborhood i , and *SubneighborhoodsForMove*(m) the method which returns the set of subneighborhoods to activate when move m is performed.

Two components in FLS procedure need to be defined. They are the breaking-down of the neighborhood into subneighborhoods and the mapping from moves to subneighborhoods for spreading activation, both of which are dependent on the move operator used. In general, the move operator depends on the solution representation. Fortunately, several problems share the same solution representation which is typically based on some well-known simple or composite combinatorial structure (e.g., selection, permutation, partition, composition, path, cyclic path, tree, graph, etc.). This allows us to use the same move operators for many different problems (e.g., 1-Opt, 2-Opt, Swaps, Insertions, etc.). Mapping subneighborhoods to moves (*MovesForSubneighborhood* procedure) can be defined by looking at the implementation of a typical LS procedure for the problem. This implementation, at its core, will usually contain a number of nested for-loops for

generating all possible move combinations. The variable in the outer-most loop in the move generation code can be used to define the subneighborhoods. The moves in each subneighborhood will be those generated by the inner loops for the particular subneighborhood index value at the outer-most loop.

Subneighborhoods can be overlapping. FLS usually examines limited numbers of moves compared to exhaustive neighborhood search methods and therefore duplication of moves is not a problem. Moreover, this can be desirable sometimes to give a greater range to each subneighborhood. Since not all subneighborhoods are active in the same iteration, if there is no overlapping, some of improving moves may be missed. The spreading activation method (*SubneighborhoodsForMove* procedure) returns a set of subneighborhoods to activate after a move is performed. The lower bound for this set is the subneighborhood where the move originated. The upper bound (although not useful) is all the subneighborhoods in the problem.

This last method can be defined by looking at the particular move operator used. Moves will affect part of the solution directly or indirectly while leaving other parts unaffected. If a subneighborhood contains affected parts, then it needs to be activated since an opportunity could arise there for an improving move as a result of the original

move performed. The FLS loop is settling down in a local minimum when all the bits of subneighborhoods turn to zero (i.e., no improving move can be found in any of the subneighborhoods).

FLS in that respect is similar to other LS procedures. The main differences are that the method can be focused to search particular parts of the overall neighborhood and secondly, it works in an opportunistic way, looking at parts of the solution which are likely to contain improving moves rather than the whole solution. In the next section, we look at GFLS, which uses FLS as its improvement method.

IMPLEMENTING GUIDED FAST LOCAL SEARCH

The GFLS is similar to GLS explained previously. All differences relate to the manipulation of activation bits for the purpose of focusing FLS. Initially, all bits are initialized to one. This enables all subneighborhoods to be examined by the first call to FLS. At each subsequent call to FLS, only the subneighborhoods that relate to the features penalized by GLS will be examined. The detailed pseudocode realizing this scheme is provided below:

```
procedure GuidedFastLocalSearch( $p, g, \lambda, [I_1, \dots, I_M], [c_1, \dots, c_M], M, L$ )
begin
   $k \leftarrow 0$ ;
   $s_0 \leftarrow \text{ConstructionMethod}(p)$ ;
  /* set all penalties to 0 */
  for  $i \leftarrow 1$  until  $M$  do
     $p_i \leftarrow 0$ ;
  /* set all sub-neighborhoods to the active state */
  for  $i \leftarrow 1$  until  $L$  do
     $\text{bit}_i \leftarrow 1$ ;
  /* define the augmented objective function */
   $h \leftarrow g + \lambda * \sum p_i * I_i$ ;
  while NOT StoppingCriterion do
    begin
       $s_{k+1} \leftarrow \text{FastLocalSearch}(s_k, h, [\text{bit}_1, \dots, \text{bit}_L], L)$ ;
      /* compute the utility of features */
      for  $i \leftarrow 1$  until  $M$  do
         $\text{util}_i \leftarrow I_i(s_{k+1}) * c_i / (1 + p_i)$ ;
      /* penalize features with maximum utility */
      for each  $i$  such that  $\text{util}_i$  is maximum do
        begin
           $p_i \leftarrow p_i + 1$ ;
          /* activate sub-neighborhoods related
          to penalized feature  $i$  */
          ActivateSet  $\leftarrow \text{SubneighborhoodsForFeature}(i)$ ;
          for each sub-neighborhood  $j$  in ActivateSet do
             $\text{bit}_j \leftarrow 1$ ;
        end
      end
       $k \leftarrow k + 1$ ;
    end
   $s^* \leftarrow$  best solution found with respect to objective function  $g$ ;
  return  $s^*$ ;
end
```

where p is the problem, g the objective function, h the augmented objective function, λ the lambda parameter, I_i the indicator function of feature i , c_i the cost of feature i , M the number of features, p_i the penalty of feature i , L the number of subneighborhoods, bit_i the activation bit for subneighborhood i , $ConstructionMethod(p)$ the method for constructing an initial solution for problem p , $FastLocalSearch(s_k, h, [bit_1, \dots, bit_L], L)$ the FLS method as depicted in the previous section, and $SubneighborhoodsForFeature(i)$ the method which returns the set of subneighborhoods to activate when feature i is penalized.

GFLS introduces the new method *SubneighborhoodsForFeature*, similar in function to the *SubneighborhoodsForMove* method, in the stand-alone FLS procedure. Its task is to identify the subneighborhoods that are related to a penalized feature. The *SubneighborhoodsForFeature* method is usually defined based on an analysis of the move operator. When penalties are increased, the search should focus on moves which remove or have the potential to remove penalized features from the solution. The subneighborhoods, which contain such moves, are prime candidates for activation when FLS restarts after each penalty application cycle.

DEFINING FEATURES AND THEIR COSTS

Features are not necessarily specific to a particular problem and they can be used in several problems of similar structure. Therefore, for each category of applications, we provide examples of features that can be deployed in the context of various problems. The reader may find them helpful and use them in his/her own optimization application. Furthermore, real-world problems, which sometimes incorporate elements from several academic problems, can benefit from using more than one feature-set to guide LS to optimize better all different terms of a complex objective function with several components.

Features for Routing and Scheduling Problems

In routing and scheduling problems, travel (or setup) times are usually a key factor of

cost. This could be expressed in terms of minimizing the travel time required by a vehicle to travel between customers or the setup time for a machine to be utilized from one activity to the next. Problems in this category include the traveling salesman problem (TSP), vehicle routing problem (VRP), and machine scheduling with sequence-dependent setup times.

Travel or setup times are modeled as edges in a path or graph structure commonly used to represent the solution of these problems. An objective function is defined (at least partly) as the sum of lengths for the edges used by the solution. In such cases, edges are ideal GLS features. A solution either contains an edge or not. Furthermore, each edge has a cost equal to its length. We can define a feature for each possible edge and assign a cost to it that equals the edge length. GLS quickly identifies and penalizes long and costly edges guiding LS to high quality solutions which use as much as possible the available short edges.

As an alternative, a slightly different feature cost strategy was proposed in Refs 7 and 8 for the VRP. The authors there consider the relative arc length between two customers according to the rest of customer population (i.e., $d_{ij} / \text{avg}(d_{ij})$) rather than d_{ij} as the feature cost. The idea is that this variation would lead to a more balanced arc selection for penalization as compared to the frequently employed strategy (only d_{ij}).

Features for Assignment and Allocation Problems

In assignment problems, a set of items has to be assigned to another set of items (e.g., airplanes to flights, locations to facilities, people to work etc.). Each assignment of item i to item j usually carries a cost. Assignment problems normally include constraints that must be satisfied (e.g., capacity or compatibility constraints). The assignment of item i to item j can be seen as a solution property which is either present in the solution or not. Since each assignment also carries a cost, this is another good and straightforward example of a feature to be used in a GLS implementation.

Assignment problems can be used to model resource allocation applications. A special but important case in resource allocation is when the resources available are not sufficient to service all requests. Usually, the objective function will contain a sum of costs for the unallocated requests, which is to be minimized. The cost incurred when a request is unallocated will reflect the importance of the request or the revenue lost in the particular scenario.

A possible feature to consider for these problems is whether a request is unallocated. If the request is unallocated, then a cost is incurred in the objective function, which we can use as the feature cost to guide LS. The number of features in a problem is equal to the number of requests that may be left unallocated, one for each request. There may be hard constraints which state that certain requests should always be allocated a resource. There is no need to define a feature for such hard constraints.

In some variations of the problem such as the quadratic assignment problem (QAP), the cost function is more complicated and assignments have an indirect impact on the cost. Even in these cases, we found that the GLS can generate good results by assigning to all features the same feature costs (e.g., equal to one or some other arbitrary value). Although, GLS is not guiding the improvement method to good solutions (since this information is difficult to extract from the objective function), it can still diversify the search because of the penalty memory incorporated and it is capable of producing results comparable to popular heuristic methods.

Features for Constraint Optimization Problems

Several combinatorial optimization problems concern the search for a feasible solution with respect to a set of constraints. In some cases, finding such a solution is impossible, and then the task gets modified into finding a solution that minimizes the number of unsatisfied constraints (relaxations).

Constraint violations may have costs (weights) associated with them, in which case the sum of constraint violation costs is to be minimized. LS usually considers the number of constraint violations (or their weighted

sum) as the objective function even in cases where the goal is to find a solution, which satisfies all the constraints. Constraints by their nature can be easily used as features. They can be modeled by indicator functions and they also incur a cost (i.e., when violated/relaxed), which can be used as their feature cost. Problems which can benefit from this modeling include the constraint satisfaction and partial constraint satisfaction problem, the famous SAT and its MAX-SAT variant, graph coloring, various frequency assignment problems [9], and others.

Features for Continuous Optimization Problems

The potential applications of GLS are not limited to optimization problems of a discrete nature, as shown in previous categories of applications, but can also be applied to difficult continuous optimization problems. In the optimization of real-valued functions, no features can be deduced from the structure of the objective function. In this case, artificial solution features can be created. A simple setting, for instance, is to divide the domains of variables into a number of nonoverlapping and equal-sized intervals. Penalizing a subdomain discourages the LS method from revisiting this area in the search space in future iterations.

Table 1 provides a summary of the applications and features for each category as discussed in this section. In the next section, we provide an example by looking at how GLS can be applied to the TSP.

EXAMPLE: THE TRAVELING SALESMAN PROBLEM

There are many variations of the TSP. In here, we examine the classic symmetric TSP. The problem is defined by N cities and a symmetric distance matrix $D = [d_{ij}]$ which gives the distance between any two cities i and j . The goal in TSP is to find a tour (i.e., closed path), which visits each city exactly once and is of minimum length. A tour can be represented as a cyclic permutation π on the N cities if we interpret $\pi(i)$ to be the city

Table 1. Common Features Used for Each Application Category

Category	Common Features
Routing	Edges in a graph, which model the travel between cities, customers, and so on. The travel time is used as feature cost for each edge
Scheduling	Edges in a graph, which model the sequence of the scheduled processes. The setup time is used as feature cost for each edge
Assignment	Every possible match (i.e. assignment) between items can be a feature, which has its (unique) cost
Allocation	The allocation of each request is a feature. Feature cost can be set to the cost of not allocating a particular request
Constrained optimization	Constraint violation is a common feature; unless constraints vary in their importance, all corresponding features are set to equal cost (e.g., 1)
Continuous optimization	Whether a variable is assigned a value within a certain interval; features are set to equal cost

visited after city $i, i = 1, \dots, N$. The cost of a permutation is defined as

$$g(\pi) = \sum_{i=1}^N d_{i\pi(i)} \quad (3)$$

and gives the cost function of the TSP.

Construction Method

A simple construction method is to generate a random tour. More advanced TSP heuristics can be used if we require a higher quality starting solution to be generated [10]. This is useful in real-time/on-line applications where a good tour may be needed very early in the search process in case the user interrupts the algorithm. If there are no concerns like that, then a random tour generator suffices since the GLS metaheuristic tends to be relatively insensitive to the starting solution and capable of finding high quality solutions even if left to run for a relatively short time.

Improvement Method

Most improvement methods for the TSP are based on the k -Opt moves. Using k -Opt moves, neighboring solutions can be obtained by deleting k edges from the current tour and reconnecting the resulting paths using k new edges. The reader can consider using the simple 2-Opt [11] method, which in addition to its simplicity, is very effective when combined with GLS. With 2-Opt, a neighboring solution is obtained from the current solution

by deleting two edges, reversing one of the resulting paths, and reconnecting the tour. Computing incrementally the change in solution cost by a 2-Opt move is relatively simple. Let us assume that edges $e1, e2$ are removed and edges $e3, e4$ are added with lengths $d1, d2, d3$, and $d4$ respectively. The change in cost is the following:

$$d3 + d4 - d1 - d2. \quad (4)$$

When we discuss the features used in the TSP, we will explain how this evaluation mechanism is revised to account for penalty changes in the augmented objective function.

Guided Local Search

For the TSP, a tour includes a number of edges and the solution cost (tour length) is given by the sum of the lengths of the edges in the tour (3). As mentioned in the section titled “Features for Routing and Scheduling Problems,” edges are ideal features for routing problems such as the TSP. First, a tour either includes an edge or not and second, each edge incurs a cost in the objective function equal to the edge length, as this is given by the distance matrix $D = [d_{ij}]$ of the problem. A set of features can be defined by considering all possible undirected edges $e_{ij}(i = 1 \dots N, j = i+1 \dots N, i \neq j)$ that may appear in a tour with feature costs given by the edge lengths d_{ij} . To each edge e_{ij} connecting cities i and city j is attached a penalty p_{ij}

initially set to zero, which is increased by GLS during search. When implementing the GLS algorithm for the TSP, the edge penalties can be arranged in a symmetric penalty matrix $P = [p_{ij}]$. Penalties have to be combined with the problem's objective function to form the augmented objective function which is minimized by LS. We therefore need to consider the auxiliary distance matrix

$$D' = D + \lambda \cdot P = [d_{ij} + \lambda \cdot p_{ij}], \quad (5)$$

LS must use D' instead of D in move evaluations. GLS modifies P and (through that) D' whenever LS reaches a local minimum.

In order to implement this, we revise the incremental move evaluation formula (Eq. 7) to take into account the edge penalties and also the λ parameter. If $p1, p2, p3$, and $p4$ are the penalties associated with edges $e1, e2, e3$, and $e4$ respectively, the revised version of Equation (7) is as follows:

$$(d3 + d4 - d1 - d2) + \lambda(p3 + p4 - p1 - p2). \quad (6)$$

Similarly, we can implement GLS for higher order k-Opt moves.

The edges penalized in a local minimum are selected according to the utility function which for the TSP takes the following form:

$$Util(tour, e_{ij}) = I_{e_{ij}}(tour) \cdot \frac{d_{ij}}{1 + p_{ij}}, \quad (7)$$

where

$$I_{e_{ij}}(tour) = \begin{cases} 1, & e_{ij} \in tour \\ 0, & e_{ij} \notin tour \end{cases} \quad (8)$$

The only parameter of GLS that requires tuning is the λ parameter. Alternatively, we can tune the a (α) parameter which is defined in “ λ Parameter” and it is relatively instance-independent. For 2-Opt, the following range for a generates high quality solutions for instances in the TSPLib [12].

$$1/8 \leq a \leq 1/2. \quad (9)$$

The reader may refer to Ref. 5 for more details on the experimentation procedure and the full set of results.

Guided Fast Local Search

We can exploit the way LS works on the TSP to partition the neighborhood into subneighborhoods as required by GFLS. Each city in the problem may be seen as defining a subneighborhood, which contains all 2-Opt edge exchanges, removing either one of the edges adjacent to the city. For a problem with N cities, the neighborhood is partitioned into N subneighborhoods, one for each city in the instance.

The subneighborhoods to be activated after a move is executed are those cities that are at the ends of the edges removed or added by the move. Finally, the subneighborhoods activated after penalization are those defined by the cities at the ends of the edge(s) penalized. There is a good chance that these subneighborhoods will include moves that remove one or more of the penalized edges.

SUMMARY AND CONCLUSIONS

It is relatively easy to adapt GLS to the different problems. Although LS is problem dependent, the other structures of GLS and also GFLS are problem independent. Moreover, a rather mechanical, step-by-step procedure can usually be followed when GLS or GFLS is applied to a new problem (i.e., implement an LS procedure, identify features, assign costs, define subneighborhoods, etc.). This approach makes GLS and GFLS easier to use by nonspecialists and software engineers.

This can be aided further by the development of libraries and toolkits which can provide to end users ready-to-use implementations of common metaheuristic methods. As an example of that, GLS and FLS have been incorporated into iOpt, a software toolkit for heuristic search methods [13], and iSchedule [14], which is an extension to iOpt for planning and scheduling applications (e.g., for solving problems such as the VRP [15]).

Examples of computer programs and demos of algorithms can help researchers and practitioners understand and effectively apply the various metaheuristic techniques. On that front, Essex University's website [16] provides experimental software for the

application of GLS and FLS on a variety of other problems, including the maximum channel assignment problem, TSP, and QAP.

REFERENCES

1. Hansen P, Mladenovic N. An introduction to variable neighborhood search. In: Voss S, *et al.* editors. Meta-heuristics, advances and trends in local search paradigms for optimization. Boston (MA): Kluwer; 1999. pp. 433–458.
2. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling salesman problem. *Oper Res* 1973;21:498–516.
3. Glover F, Laguna M. Tabu search. Boston (MA): Kluwer Academic Publishers; 1997.
4. Pesant G, Gendreau M. A constraint programming framework for local search methods. *J Heuristics* 1999;5(3):255–279.
5. Voudouris C, Tsang EPK. Guided local search and its application to the travelling salesman problem. *Eur J Oper Res* 1999;113(2):469–499.
6. Kytöjoki J, Nuortio T, Braysy O, *et al.* An efficient variable neighborhood search heuristic for very large scale vehicle routing problems. *Comput Oper Res* 2007;34(9):2743–2757.
7. Tarantilis CD, Zachariadis EE, Kiranoudis CT. A guided tabu search for the heterogeneous vehicle routing problem. *J Oper Res Soc* 2008;59(12):1659–1673.
8. Tarantilis CD, Zachariadis EE, Kiranoudis CT. A hybrid guided local search for the vehicle-routing problem with intermediate replenishment facilities. *INFORMS J Comput* 2008;20(1):154–168.
9. Murphey RA, Pardalos PM, Resende MGC. In: Du D-Z, Pardalos P, editors. Volume 4, Frequency assignment problems. Handbook of combinatorial optimization. Netherlands: Kluwer Academic Publishers; 1999.
10. Reinelt G. Volume 840, The traveling salesman: computational solutions for TSP applications. Lecture notes in computer science. Berlin: Springer; 1995.
11. Croes A. A method for solving traveling-salesman problems. *Oper Res* 1958;5:791–812.
12. Reinelt G. A traveling salesman problem library. *ORSA J Comput* 1991;3:376–384.
13. Voudouris C, Dorne R, Lesaint D, *et al.* In: Walsh T, editor. Volume 2239, iOpt: a software toolkit for heuristic search methods. Practice of constraint programming - CP 2001. Lecture notes in computer science. Berlin: Springer; 2001. pp. 716–729.
14. Dorne R, Voudouris C, Liret A, *et al.* iSchedule an optimisation tool-kit based on heuristic search to solve BT scheduling problems. *BT Technol J* 2003;21(4):50–58.
15. Dorne R, Mills P, Voudouris C. Solving vehicle routing using iOpt. In: Doerner KF, Gendreau M, Greistorfer P, *et al.*, editors. Volume 39, Metaheuristics: progress in complex systems optimization. Operations research/computer science interfaces series. Hoboken (NJ): Springer; 2007. pp. 389–408.
16. GLS Demos. Available at <http://cswwww.essex.ac.uk/CSP/glsdemo.html>. 2008.

EFFECTIVE APPLICATION OF SIMULATED ANNEALING

JEFFREY W. OHLMANN

Department of Management Sciences,
University of Iowa, Iowa City, Iowa

JON W. VAN LAARHOVEN

Department of Applied Mathematics and
Computational Sciences, University of
Iowa, Iowa City, Iowa

Simulated annealing (SA) is a stochastic local search algorithm in which a probabilistic transition mechanism governs the acceptance or rejection of each move from a current solution x to a candidate solution y . The salient feature of SA is its ability to escape from becoming trapped in local minima by accepting nonimproving solutions. At each iteration, SA accepts the candidate solution with a probability dependent on two factors: (i) the relative cost of the candidate solution versus the current solution and (ii) the value of a control parameter referred to the *temperature*. The temperature parameter derives its name from the algorithm's thermodynamic analogy to the metallurgical process of annealing; as in the physical process of heating material and cooling it to change its properties, the manipulation of the temperature parameter is often critical to the success of SA on an optimization problem. See ***Axiomatic Measures of Risk and Risk-Value Models*** for additional background information on the simulated annealing algorithm.

ALGORITHMIC IMPLEMENTATION

To implement an SA algorithm, the user must define a precise problem representation, a method for generating candidate solutions, and a cooling schedule to control the transition mechanism over the course of the algorithm. In this section, we elaborate on implications of the design of these features.

To aid our exposition, we will rely on Algorithm 1, which illustrates the pseudocode for a general SA algorithm for a minimization problem. Adhering to traditional implementations, Algorithm 1 incorporates a probabilistic mechanism for transitioning between solutions based on the Boltzmann distribution, although the acceptance function in Line 13 can certainly be modified, for example, see the section titled “Orthogonal Stock-Cutting Problem.”

Problem-Specific Design Choices

The first step in applying any solution technique to an optimization problem is clearly delineating the problem, keeping in mind that there may exist multiple ways to represent the problem, and that the choice of representation has implications on solving the problem. By characterizing what constitutes a feasible solution and whether infeasible solutions will be considered along the search path, the problem representation will define the solution space S . Furthermore, to evaluate a solution, the user must specify a cost function $c : S \rightarrow \mathbb{R}$.

SA considers moves from a current solution by generating candidate solutions from a defined neighborhood (Line 5). To facilitate the local search procedure, a neighborhood function $N : S \rightarrow 2^S$ generates candidate solutions for consideration. For $x \in S$, $N(x) \subseteq S$ is called the *neighborhood* of x , typically consisting of solutions that share some structural commonality with x . It is most common to define $N(x)$ as the set of all possible solutions that can be obtained by applying a perturbation operator to x . For example, if $x = [1, 2, 3]$, then with respect to an exchange operation on every pair of elements, $N(x) = \{[2, 1, 3]; [1, 3, 2]; [3, 2, 1]\}$.

To make a local search algorithm such as SA more efficient, it is important to take advantage of the relationship between a current solution x and a neighbor solution

$y \in N(x)$ so that $c(y)$ can be updated from $c(x)$ rather than completely recalculated. The use of appropriate data structures can further facilitate fast manipulations. When evaluation of the cost difference between two solutions is computationally expensive, Tovey [1] advocates the use of a surrogate function to approximate the cost of a neighbor solution at most iterations and occasionally calculating the true solution cost, for example, when neighbor solution is accepted.

SA typically generates a neighbor solution by randomly sampling the neighborhood. An

alternative is to cyclically sample the neighborhood so that a transition to a neighbor solution would not be considered twice until all other neighbor solutions are considered once. However, due to the stochastic nature of SA, the implications of cyclic sampling are difficult to anticipate and it may improve or deteriorate performance. As opposed to randomly generating a neighbor solution, Fox [2] proposes the development of mechanisms that “intelligently” generate neighbors similar to the intensification and diversification protocol of tabu search.

Algorithm 1. Simulated annealing for minimization problems.

```

1  Initialize solution,  $x$ , and best solution found,  $x_{\text{best}} \leftarrow x$ ;
2  Initialize temperature,  $\tau$ ;
3  while Outer termination criteria not satisfied do
4      while Inner termination criteria not satisfied do
5          Generate neighbor solution  $y \in N(x)$ ;
6          if  $c(y) \leq c(x)$  then
7               $x \leftarrow y$ ;
8              if  $c(x) < c(x_{\text{best}})$  then
9                   $x_{\text{best}} \leftarrow x$ ;
10             end
11         else
12             Generate uniform (0,1) random variable  $u$ ;
13             if  $\exp\left(\frac{c(x)-c(y)}{\tau}\right) > u$  then
14                  $x \leftarrow y$ ;
15             end
16         end
17     end
18     Update  $\tau$  via cooling function;
19 end
20 Return  $x_{\text{best}}$ ;

```


The choice of neighborhood can have a significant impact on the effectiveness of a SA algorithm. Determining a neighborhood that facilitates an effective search within SA is largely dependent on the problem domain and may require extensive experimentation as well as problem insight. Ideally, the selected neighborhood should impose a topology that avoids numerous deep local minima that can hinder SA's ability to escape their basins of attraction. Furthermore, neighborhoods that facilitate *reachability*, that is, every solution can be reached from any other through a series of transitions, are theoretically appealing. Observations on the effect of neighborhood size (as measured by $|N(x)|$) are inconclusive and appear to problem-dependent [3–5]. Another possibility is to adjust the neighborhood structure with changes in the temperature parameter [1].

For a constrained optimization problem, it may be difficult to generate neighbor solutions that maintain feasibility with respect to one or more constraints. The primary options for applying a local search algorithm to a constrained optimization problem are (i) ensure the generation of feasible neighbors that obey all constraints, or (ii) allow the consideration of infeasible neighbor solutions with a penalization for infeasibility. There are advantages and drawbacks of both approaches. Depending on the nature of the constraint(s), the process of generating only feasible neighbors may be computationally expensive if one cannot develop a procedure to ensure feasibility. In these cases, infeasible neighbor solutions may be discarded or attempts to repair them may be executed through a predefined protocol. Additionally, requiring feasible solutions may degrade the reachability of a neighborhood, retarding SA's ability to search the solution space. However, maintaining feasibility prevents the search potentially wasting time evaluating a large number of infeasible solutions that may not lead to improved feasible solutions. Allowing the consideration of infeasible solutions simplifies the generation of a neighbor solution, but at the cost of increasing the set of candidate solutions to be considered.

Generic Design Features

We consider the design of a cooling schedule to be generic in the sense that it does not explicitly incorporate characteristics of the problem under consideration. The cooling schedule is fully specified by an initial temperature value, number of iterations at each temperature value, cooling function, and final temperature value. With reference to Algorithm 1, the inner while loop (Line 4) controls the number of iterations at each temperature value while the outer while loop (Line 3), in conjunction with the cooling function (Line 18), controls the number and value of different temperature settings. Thus, the cooling schedule governs the stochastic process, a series of finite homogeneous Markov chains, which describes the search path of a SA algorithm.

Initial Temperature. For relatively large values of temperature (τ), SA is essentially a form of random search, as a transition to virtually any neighbor solution is accepted. To capture the benefits of being able to escape local minima and still satisfactorily explore the basins containing them, temperature is typically initiated at a large value and reduced gradually over time. The ideal initial temperature, τ_0 , may vary across problem types (and even across instances of the same problem type), but typically τ_0 is set to a large value so that the final solution is effectively independent of the initial solution. One method to setting τ_0 is to specify a percentage of proposed uphill transitions, χ_0 , to be accepted at the algorithm's onset. Values of χ_0 typically range from 0.80 to 0.99, although if the SA algorithm is seeded with a “good” initial solution and the user wishes to retain the potential benefit of this starting point, χ_0 (and subsequently τ_0) may be set lower in order to restrain the search. The initial temperature is then given by:

$$\tau_0 = \frac{\bar{d}}{\ln(1/\chi_0)},$$

where $\bar{d}$ is an estimate of the average difference in cost of a current solution and its neighbor. Another simple rule is to set $\tau_0 =$

$d_{\max}$, where $d_{\max}$ is an estimate of the largest difference in evaluation function between any two neighboring solutions. Regardless of how τ_0 is estimated, a preliminary number of transitions can be tracked at the onset of the algorithm to ascertain that the percentage of accepted moves surpasses χ_0 ; if τ_0 does not satisfy this threshold, it can be increased and retested.

Iterations per Temperature Value. Theoretical results suggest that the Markov chain generated by the inner loop's set of iterations should be allowed to converge to its stationary distribution before decrementing the temperature. Unfortunately, the number of iterations necessary to even approximate the stationary distribution is exponential in problem size [6]. In practice, the length of the Markov chain at each temperature is usually related to the size of the neighborhood structure or the size of the problem. An alternate approach determines the length of the k^{th} Markov chain by terminating the inner loop if either a minimum threshold of transitions have been accepted or a maximum limit of trials have been attempted. Regardless of the approach, experimental testing is typically required to obtain an effective setting for a wide range of problem instances.

Cooling Function. After employing the search at a fixed temperature for a number of inner loop iterations, the temperature is modified according to a cooling function. Traditionally, the cooling function monotonically decreases the temperature. One basic cooling function is a geometric decrement, that is, $\tau_{k+1} = \alpha \tau_k$, where $0 < \alpha < 1$, with values typically ranging as low as 0.5 and as high as 0.99. Other cooling functions may be designed so that the arithmetic difference between successive temperature settings is constant (rather than the ratio of successive temperature settings).

Many more sophisticated cooling functions exist in the literature, but it is typically advisable to begin with a simple cooling function before moving onto alternatives out of necessity. Empirical evidence suggests that the particular form of cooling schedule is not necessarily relevant as long as the algorithm

spends the appropriate number of iterations over a "critical" temperature range. Rather than strictly adhering to a monotonic cooling schedule, some applications employ the concept of "reheating" in which temperature is increased via an adaptive mechanism that monitors algorithm features such as the number of iterations since last update in best-found solution or the history of accepted and rejected moves, for example, see the sections titled "Job Shop Scheduling Problem" and "Orthogonal Stock-Cutting Problem."

Final Temperature. As temperature approaches zero, SA becomes a descent method in which only transitions to improving neighbor solutions are accepted resulting in convergence to a local minima. There are possible numerous stopping conditions, but the most effective typically triggered when the algorithm's mobility is sufficiently limited. The simplest outer loop termination criteria is to specify a total number of temperature settings to be considered, thereby implicitly defining a final temperature value in conjunction with the cooling function. Placing a ceiling on the amount of computation time also implies a run-time dependent final temperature. Another popular criterion is to terminate the algorithm when the percentage of accepted moves drops below a threshold, χ_f , for a specified number of iterations, and the best overall solution has not been updated recently.

Parameter Calibration

The need for potentially tedious parameter calibration is often a deterrent to successfully implementing SA. To circumvent *ad hoc* parameter-tuning, several researchers have proposed structured parameter calibration approaches based on statistical design of experiments [7–9]. In the development of the algorithm, it is often helpful to track the run-time performance through metrics such as percentage accepted moves, evaluation function of current solution, and evaluation function of best-found solution.

ALGORITHM APPLICATION AND DEVELOPMENT

The past few decades have provided an overwhelming amount of computational experience about the SA algorithm. SA can consistently find high-quality solutions on a wide array of optimization problems, but often at the high cost of computational effort. The remainder of this article highlights some applications of SA (and variants thereof) selected because they demonstrate various innovations in the use of SA that have proven to be empirically competitive with alternate approaches.

Job Shop Scheduling Problem

Van Laarhoven *et al.* [10] apply SA to the job shop scheduling problem (JSP) in which a set of N jobs, each consisting of a sequence of m operations that require execution on a specified series of m machines, must be completed in as little time as possible. For the JSP, a solution can be represented by a set $x = \{\pi_1, \dots, \pi_m\}$, where π_i is a permutation of jobs on machine i . The cost of a solution x is computed by solving a longest-path problem (in $O(N)$ time) for a corresponding directed graph. Thus, the JSP can be reduced to the problem of finding a set of permutations that minimizes the longest path in an appropriately defined directed graph.

Van Laarhoven *et al.* [10] implement a cooling schedule of the form

$$\tau_{k+1} = \frac{\tau_k}{1 + \tau_k \cdot \frac{\ln(1+\delta)}{3\sigma_k}},$$

where δ is a small positive constant (ranging from 0.001 to 1 in the reported experiments) and σ_k is the standard deviation in cost values of the solutions along the search path for the last temperature setting, τ_k . Furthermore, van Laarhoven *et al.* set the number of iterations at each temperature to be equal to the size of the largest neighborhood (over each solution in S).

Van Laarhoven *et al.* [10] define a neighborhood transition in reference to a current solution x ; their neighborhood involves reversing one edge on the longest path and adapting the

adjacent arcs to ensure the generation of feasible candidate solutions. Each solution's neighborhood contains $O(N)$ neighbors and is asymmetric, that is, $y \in N(x) \not\Rightarrow x \in N(y)$.

Motivated by the observation that the asymmetry of standard neighborhoods for the JSP complicates the convergence properties of SA, Kolonko [11] integrates features from genetic algorithms with an annealing approach. Specifically, Kolonko introduces a small population of independent, parallel SA runs in which the “fitness” (adopting the vernacular of genetic algorithms) of an individual SA run in this population is defined by the cost of its best-found solution. After the SA runs are executed for specified number of iterations, they interact via a crossover operation. The crossover operation is applied to the best-found solutions corresponding to randomly selected pairs of SA runs. Relying on a Gantt chart representation, the crossover combines solutions by taking the first portion of the schedule depicted by the first solution and augmenting it the second portion of the schedule depicted by the second solution (with appropriate modifications to ensure feasibility). Repeated application of the crossover operation generates a pool of offspring solutions, which are then used to seed the next generation of SA runs.

Kolonko [11] applies an adaptive control to modify temperature after the consideration of each neighbor solution. If a move from current solution x to candidate solution y is accepted, then the temperature is decreased by an amount depending on whether $c(y)$ is smaller or larger than $c(x)$. If a move from x to y is rejected, the temperature is increased.

SA consistently produces good solutions to the JSP with a modest amount of implementation effort (testing and tuning) and relatively little insight into the combinatorial structure of the problem. However, SA may require large running times, and the computational results are not typically best-in-class.

Traveling Salesman Problem with Time Windows

Ohlmann and Thomas [12] consider SA integrated with a variable penalty method to

allow infeasible solutions in search of near-optimal solutions for the traveling salesman problem with time windows (TSPTW). In the TSPTW, the objective is to find a minimum-cost tour that visits each of n cities exactly once within a city-specific time window. A solution x (a tour through n cities) is represented by an n -permutation and it is evaluated by computing the cost of traversing the tour, $f(x)$, as well as cumulative violation of the time windows, $p(x)$. The cost of a solution x is evaluated as $c(x) = f(x) + \lambda p(x)$, where λ is a scalar penalty multiplier. For a set value of λ , SA can be applied as in Algorithm 1 (with minor changes to account for the storage of the best *feasible* solution in x_{best}).

Determining the value of λ can be a difficult issue. To address the difficulty of determining the appropriate value of the penalty multiplier, Ohlmann *et al.* [13] examine a variant of SA in which the value of λ is initialized at a small value and increased over the course of the algorithm. Within the context of the traveling salesman problem with time windows, Ohlmann and Thomas [12] demonstrate that increasing the penalty multiplier according to a limited-exponential schedule while decreasing temperature according to a geometric schedule outperforms SA executed at various fixed values of λ . Ohlmann and Thomas utilize a one-shift neighborhood, which removes a randomly selected element of the n -permutation and reinserts before another randomly selected element; they discovered that the choice of this neighborhood is critical to the algorithm's effectiveness. Furthermore, Ohlmann and Thomas obtained competitive solutions (best-in-class for some benchmark problems), although typically at the cost of more computation time.

Vehicle-Routing Problems

Bent and van Hentenryck [14] employ SA in a two-phase approach for the vehicle-routing problem with time windows (VRPTW). For a set of customers with known demands and a fleet of vehicles with identical capacities, the VRPTW asks for a set of routes originating from a depot and visiting each customer exactly once within a specified time window. In collectively satisfying the customers' demands, the vehicles' capacities cannot be

exceeded. The vehicle-routing problem (VRP) solutions can be represented as a set of permutations $x = \{\pi_1, \dots, \pi_m\}$, where π_i is the sequence of customers to be visited by vehicle i .

The VRPTW has a hierarchical objective of minimizing the fleet size, and then subject to the minimum fleet size, minimizing the total travel cost of the routes. Bent and van Hentenryck [14] approach the VRPTW with a two-stage algorithm in which the first stage utilizes SA to minimize the number of vehicles, and the second stage uses a large neighborhood search to minimize the total travel cost.

A novel feature of the SA implementation for the first-stage is the "lexicographic" objective function, which guides the search by rewarding solutions that, in order of decreasing importance, (i) minimize the number of routes, (ii) maximize the sum of the squared route cardinalities, and (iii) minimize the minimal delay of the routes. As there are many solutions with the same number of routes, the second and third criteria are effective tie breakers. Maximizing the sum of the squared route cardinalities favors solutions having some routes with many customers and some routes with very few customers over solutions with routes having similar numbers of customers; this aids the minimization of routes since routes with few customers are typically easier to eliminate. For the third criteria, minimal delay is a routing-specific concept that is used to favor solutions such that the smallest route contains customers that can be relocated with the smallest possible violation of capacity or time window constraints (again facilitating route elimination).

Bent and van Hentenryck [14] utilize a neighborhood consisting of five interroute or intraroute operators and consider only feasible neighbors by computationally confirming satisfaction of capacity and time window constraints. At each iteration of the algorithm, a randomly selected operator is applied to a randomly selected customer to construct all feasible candidate solutions in the neighborhood corresponding to the selected operator and customer. Borrowing the concept of aspiration criterion from tabu search, Bent

and van Hentenryck augment the traditional SA approach by first checking to see if the best candidate solution from the constructed neighborhood is better than the best-found solution. If the best candidate solution is better than the best-found-so-far, the algorithm executes the move to this solution; if not, the SA algorithm proceeds by selecting a candidate solution from the constructed neighborhood (incorporating a biased sampling approach in which the likelihood of a solution being selected depends on its objective function value) and then the traditional SA acceptance protocol is applied (Line 13).

The other SA features are standard: a geometric cooling schedule, iteration limit for the inner loop (Line 4), and the combination of a time limit or minimum temperature for the outer loop (Line 3). This two-stage algorithm has proven to be very robust and produces high-quality results across all benchmark VRPTW problems.

Bräysy *et al.* [15] designed a three-phase heuristic for the fleet size and mix vehicle-routing problem with time windows (FSMVRPTW). In the FSMVRPTW, the decision-maker must determine the fleet size and composition (vehicles are heterogeneous with respect to capacity) in addition to the routing of each of the vehicles to collectively visit a set of n customers exactly once within their respective time windows. The first two phases employ problem-specific routing procedures to obtain a good initial solution and then attempt to minimize the fleet size. The final phase of their three-phase heuristic utilizes a deterministic, which replaces the probabilistic transition mechanism of SA with a deterministic rule to determine acceptance or rejection of a neighbor solution. Deterministic annealing aims to achieve the robust search behavior of SA with the efficiency of a deterministic procedure.

Bräysy *et al.*'s deterministic annealing routine mimics an algorithm known as *threshold accepting*. A threshold accepting algorithm substitutes a threshold T in lieu of temperature τ and replaces the condition in Line 13 with $c(y) - c(x) < T$, where T may be controlled in a manner similar to temperature in a SA algorithm. Bräysy *et al.*

[15] decrement T by ΔT units at each nonimproving iteration, and periodically reset T to escape local minima. We note that when generating T from an exponential distribution with a mean equal to τ , threshold accepting becomes equivalent to SA. Another variant of deterministic annealing is *record-to-record travel*. Record-to-record travel substitutes R in place of τ and replaces the condition in Line 13 with $c(y) < Rc(x)$, where R is a user-controlled parameter slightly greater than 1.

Bräysy *et al.* [15] equip their deterministic annealing approach with a compound neighborhood consisting of four different operators, applied in a specified order. The utilization of these intraroute and interroute operators emphasizes the importance of neighborhood choice in designing an effective local search algorithm. Bräysy *et al.* restrict their search solely to feasible solutions. To limit the computation time, Bräysy *et al.* utilize efficient constraint feasibility checks (for the vehicle capacity constraints and time windows) facilitated by sophisticated data structures. In their feasibility tests, they check the capacity constraint first since it is the quickest to calculate, and only proceed to time feasibility if capacity constraints are obeyed.

In addition to the opportunistic constraint-checking, Bräysy *et al.* [15] restrict the neighborhood search to reduce computational effort. Move operations are not applied to routes or route pairs if previous iterations have revealed that improvement is not possible. In addition, moves that introduce solution characteristics unlikely to appear in a near-optimal solution are avoided. To execute this strategy, Bräysy *et al.* prohibit moves that introduce long travel distances into the solution; they employ a learning approach across restarts of the algorithm to vary the definition of a "long" travel distance.

Finally, Bräysy *et al.* [15] incorporate a frequency-based long-term memory (similar to that of tabu search) to facilitate learning of arcs that frequently compose good solutions. This memory is then used to probabilistically generate initial solutions for restarting the algorithm. In summary, the three-phase approach produces excellent results on benchmark problems.

Orthogonal Stock-Cutting Problem

Burke *et al.* [16] embed SA in a two-stage approach for the orthogonal stock-cutting problem (OSCP). In the OSCP, a set of rectangular objects must be placed within the confines of a stock sheet of fixed width so that they do not overlap. The objective is to minimize the required length of the stock sheet under the restriction that objects can only be rotated in 90° increments of their original orientation.

The first stage of the algorithm invokes the best-fit heuristic [17], a tailored placement algorithm for stock-cutting problems, to pack $n - m$ rectangles (from the total set of n rectangles). The second stage packs the remaining m rectangles (not changing the placement of the first $n - m$ rectangles) using an SA algorithm to govern a bottom-left-fill heuristic (another tailored placement heuristic for stock-cutting problems). The solution for this second stage consists of a permutation of rectangle identification numbers. A solution is evaluated by placing the rectangles into the stock sheet in the order denoted by the permutation according to the protocol of the bottom-left-fill heuristic. A neighbor solution is generated by randomly swapping elements of the permutation as well as a probabilistic rotation of first random rectangle involved in the swap. Acceptance or rejection of neighbor solutions is governed by the SA algorithm; a transition from solution x to a worse neighbor solution y is accepted with a probability of $(1 - (c(y) - c(x))/\tau)$ (replacing the traditional acceptance function of Line 13 in Algorithm 1). If a neighbor solution is accepted, temperature is reduced using a geometric multiplier of 0.999. If a neighbor solution is rejected, temperature is increased using a geometric multiplier of 1.0001.

The two-stage approach achieves best-found solutions on a majority of instances, with most solutions at optimal or a few percentage over optimal. Moderate values of m allow the algorithm to couple the benefit of starting with a good partial solution from the best-fit heuristic with a robust annealing-based search. Burke *et al.* [16] note that they also attempted using tabu search and genetic algorithms for the second stage, but simulated annealing outperformed both of these.

FURTHER READING

Noising methods are an approach related to SA. In a noising method, noise is randomly chosen from an interval, which decreases as the algorithm progresses. This noise is added to the objective value, and solutions are then accepted deterministically. Note that a noising method can be viewed as an annealing approach where the randomization is moved from the acceptance probability to the cost evaluation. Alternative methods of managing the noise introduced into the solution evaluation result in variants such as weight annealing [18] and quantum annealing [19].

Consider the class of discrete stochastic optimization problems (DSOPs) in which the objective function values for solutions cannot be evaluated exactly, but rather must be estimated (perhaps via simulation). In this context, noise is not artificially introduced (as in a noising method), but it is inherent in the evaluation due to sampling error. Alrefaei and Andradóttir [20] describe an annealing algorithm for DSOPs that fixes temperature at a constant value. Branke *et al.* [21] study various “stochastic” annealing algorithms for DSOPs, which view the uncertainty in the objective function as analogous to the probabilistic transition mechanism via temperature in a SA algorithm. In stochastic annealing, the probabilistic transition mechanism is transferred from the acceptance criterion into the objective function itself, where sampling an increasing number of neighbor solutions reduces uncertainty in objective function evaluation, analogous to decreasing temperature.

The widespread application of SA is due to the relative ease of implementing its logic, its general nature and flexibility allowing use with little problem-specific tailoring (at least for alpha versions of an implementation), and its robust performance with respect to solution quality and running time on a wide array of problems. The relative simplicity and implementation ease of SA paired with its ability to frequently obtain competitive solutions often make it a method to consider, either as a stand-alone approach, as part of a multistage approach, or hybridized with other heuristic approaches. The algorithmic

implementation suggestions and applications that this article details are just a small sample of the voluminous body of literature related to SA. Other applications in which SA has made a strong impact include time-tabling [22,23], the football pool problem [24], the design of very large scale integrated circuits [25], and image processing [26]. There also exist several variants of simulated annealing for continuous optimization [27,28] and multiobjective optimization [29].

REFERENCES

1. Tovey C. Simulated simulated annealing. *Am J Math Manage Sci* 1988;8:389–407.
2. Fox B. Integrating and accelerating tabu search, simulated annealing, and genetic algorithms. *Ann Oper Res* 1993;41:47–67.
3. Cheh K, Goldberg J, Askin R. A note on the effect of neighborhood structure in simulated annealing. *Comput Oper Res* 1991;18:537–547.
4. Ogbu F, Smith D. The application of the simulated annealing algorithm to the solution of the $n/m/C_{max}$ flowshop problem. *Comput Oper Res* 1990;17:243–253.
5. Goldstein L, Waterman M. Neighborhood size in the simulated annealing algorithm. *Am J Math Manage Sci* 1988;8:409–423.
6. van Laarhoven P, Aarts E. *Simulated annealing: theory and applications*; Dordrecht, The Netherlands; Kluwer; 1988.
7. Park M, Kim Y. A systematic procedure for setting parameters in simulated annealing algorithms. *Comput Oper Res* 1998;25:207–217.
8. Coy SP, Golden BL, Runger GC, *et al.* Using experimental design to find effective parameter settings for heuristics. *J Heuristics* 2000;7:77–97.
9. Adenso-Diaz B, Laguna M. Fine-tuning of algorithms using fractional experimental designs and local search. *Oper Res* 2006;54:99.
10. van Laarhoven P, Aarts E, Lenstra J. Job shop scheduling by simulated annealing. *Oper Res* 1992;40:113–125.
11. Kolonko M. Some new results on simulated annealing applied to the job shop scheduling problem. *Eur J Oper Res* 1999;113:123–136.
12. Ohlmann J, Thomas B. A compressed-annealing heuristic for the traveling salesman problem with time windows. *INFORMS J Comput* 2007;19:80.
13. Ohlmann J, Bean J, Henderson S. Convergence in probability of compressed annealing. *Math Oper Res* 2004;29:837–860.
14. Bent R, Van Hentenryck P. A two-stage hybrid local search for the vehicle routing problem with time windows. *Transp Sci* 2004;38:515–530.
15. Bräysy O, Dullaert W, Hasle G, *et al.* An effective multirestart deterministic annealing metaheuristic for the fleet size and mix vehicle-routing problem with time windows. *Transp Sci* 2008;42:371–386.
16. Burke E, Kendall G, Whitwell G. A simulated annealing enhancement of the best-fit heuristic for the orthogonal stock-cutting problem. *INFORMS J Comput* 2009;21:505–516.
17. Burke E, Kendall G, Whitwell G. A new placement heuristic for the orthogonal stock-cutting problem. *Oper Res* 2004;52:655–671.
18. Loh K, Golden B, Wasil E. Solving the one-dimensional bin packing problem with a weight annealing heuristic. *Comput Oper Res* 2008;35:2283–2291.
19. Martoňák R, Santoro G, Tosatti E. Quantum annealing of the traveling-salesman problem. *Phys Rev E* 2004;70:57701.
20. Alrefaei M, Andradóttir S. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Manage Sci* 1999;45:748–764.
21. Branke J, Meisel S, Schmidt C. Simulated annealing in the presence of noise. *J Heuristics* 2008;14:627–654.
22. Abramson D. Constructing school timetables using simulated annealing: sequential and parallel algorithms. *Manage Sci* 1991;37:98–113.
23. Zhang D, Liu Y, M'Hallah R, *et al.* A simulated annealing with a new neighborhood structure based algorithm for high school timetabling problems. *Eur J Oper Res* 2010;203:550–558.
24. van Laarhoven P, Aarts E, van Lint J, *et al.* New upper bounds for the football pool problem for 6, 7 and 8 matches. *J Comb Theory Ser A* 1989;52:304–312.
25. Sechen C. *VLSI placement and global routing using simulated annealing*. Dordrecht, The Netherlands; Kluwer Academic Pub; 1988.
26. Carnevali P, Coletti L, Patarnello S. Image processing by simulated annealing. *IBM Journal of Research and Development*. 1985;29:569–579.
27. Dekkers A, Aarts E. Global optimization and simulated annealing. *Math Program* 1991;50:367–393.

28. Romeijn H, Smith R. Simulated annealing for constrained global optimization. *J Glob Optim* 1994;5:101–126.
29. Bandyopadhyay S, Saha S, Maulik U, *et al.* A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Trans Evol Comput* 2008;12:269–283.

EFFICIENT ITERATIVE COMBINATORIAL AUCTIONS

DEBASIS MISHRA
Indian Statistical Institute,
New Delhi, India

INTRODUCTION

Auctions are a popular way of selling goods—some examples include auctioning of flowers in the Netherlands, auctioning of advertising slots on search engines such as Yahoo! and Google, and auctioning of various goods on eBay. Typically, auctions come in two variants (i) sealed-bid, where bidders submit bids only once and (ii) iterative auctions, which are characterized by monotonic price adjustments given the demands of buyers at every price. Iterative auctions are preferred over their sealed-bid counterparts because of price discovery, decentralized nature, valuation privacy, and better revenue and efficiency properties [1].

Traditionally, auction theory in economics looks at the theory of selling a single indivisible good to multiple buyers [2]. But this changed dramatically with the success of spectrum auctions in the United States, where multiple spectrums were allocated to buyers simultaneously. This also inspired many applications of auctions in which multiple objects were sold simultaneously to buyers—auction of airport landing slots, auction of truckloads in the transportation industry, auction of bus routes in London, and auction of industrial procurement [3]. Such auctions where multiple indivisible objects were sold simultaneously are commonly called *combinatorial auctions* because bidders have values on *bundles* of objects as opposed to individual objects.

Because objects may be either complements or substitutes, the necessity to elicit bids on bundles of objects is clear—for instance, if there is a computer and a keyboard for sale, a bidder may have values 10 and 5, respectively, on each of them, but may have a value of 20 on the bundle of computer and keyboard.

With the success of combinatorial auctions in practice, researchers started investigating research issues in this area. Two strands of literature started—(i) computational, which looked at various computational problems associated with combinatorial auctions and (ii) auction theoretic, which looked at designing combinatorial auctions assuming participants are rational selfish agents. This survey tries to touch on some issues that can be described at the interface of these two strands of literature.

This survey is *not about* “optimal auction design” literature, which asks the question how to design auctions for selling multiple objects so that the expected revenue of the auctioneer is maximized. This problem was addressed by Myerson [4] in his seminal paper for the single object case, but its solution for the multiple object case remains illusive till date. As we explain later, this survey is about designing auctions for selling multiple objects that achieve economic efficiency.

A standard model in auction theory to model the sale of a single indivisible object is that of private values, where each bidder (buyer/agent) has a value, which is his private information, but the seller does not know the value of buyers. The seller would like to sell the object to the highest-valued bidder (economic efficiency). But unless the auction procedure is properly designed, bidders may lie about their value to get more profit. To remedy this, auction theory prescribes the second-price auction, where the good is sold to the bidder with the highest bid but charged a price equal to the second-highest bid. It can be shown that bidding the

true value is a dominant strategy for every bidder.¹

However, this auction is impractical for many reasons [5], and is rarely used in practice. However, its iterative variant (sometimes referred to as the *English auction*) is widely used in practice (for example, Sotheby's auctions, auctions on eBay web site). The auction is iterative and maintains a price in each iteration. It asks buyers if they are willing to buy the good at the current price. If more than one buyer is interested in the good at the current price, then the price is raised. Else, the auction terminates with the only interested buyer winning the good at the current price. Because of its transparency in price discovery, this auction is more practical than the second-price auction, though both the auctions achieve the same outcome in equilibrium.

This inspired researchers to think of the generalization of the English auction for the combinatorial auctions setting. The first such paper was in the mid-1980s, by Demange *et al.* [6] on a specific combinatorial auction setting. We investigate some recent work on the general combinatorial auction setting. We will mainly focus on the works of Mishra and Parkes [7], Bikhchandani and Ostroy [8], Ausubel [9], Parkes and Ungar [10], de Vries *et al.* [11], and Ausubel [12].

The reason for surveying this branch of literature is twofold: (i) the fundamentals of this literature are linear and integer programming, and someone with a good operations research background will enjoy this application; (ii) the motivation of this literature is economic, in the sense that the auctions we will talk about have simple equilibrium bidding strategies for bidders to follow. In the process, we will leave out a large literature on iterative auctions, which

did not consider the economic motivations, and were purely inspired by computational and practical considerations.

Critical to the survey is the presence of strategic agents who are payoff maximizers. What we mean by efficient is that in every state of the world efficiency is achieved. This is possible only by considering the appropriate incentive schemes that are discussed. There are many iterative auctions, which achieve efficiency in the presence of truthful bidders (i.e., bidders are assumed to behave truthfully even in the absence of appropriate incentive schemes), for example, the Clock auctions in Ref. 13. However, we do not treat these as efficient auctions in this survey. The readers are referred to the classic book [14] on this topic.

THE MODEL AND BACKGROUND

The abstract model of the survey is defined in this section. We denote $N = \{1, \dots, n\}$ to be the set of n buyers (bidders/agents) and $M = \{1, \dots, m\}$ to be the set of m indivisible goods. Note that the goods in M can be anything: they can be identical (homogeneous) or heterogeneous or both. Let $\mathbb{S} = \{S : S \subseteq M\}$ be the set of all bundles of goods. Every buyer has a valuation function. The valuation function of buyer i is $v_i : \mathbb{S} \rightarrow \mathbb{R}_+$. The valuation function of buyer i is completely known to buyer i but not known to the seller who is selling the goods. We make the following assumptions about the valuation function.

- *Zero Outside Option.* For every $i \in N$, $v_i(\emptyset) = 0$.
- *Bundle Monotonicity (Free Disposal).* For every $i \in N$ and every $S, T \subseteq \mathbb{S}$ with $S \subseteq T$, we have $v_i(S) \leq v_i(T)$.

Such valuation functions are very general. In particular settings, there are succinct ways to represent a valuation function [15]. We give two such examples below. Let $v : \mathbb{S} \rightarrow \mathbb{R}_+$ be a general valuation function, which satisfies zero outside option and bundle monotonicity.

1. *Unit demand.* A valuation function v is a *unit demand* valuation

¹There is a significant literature in auction theory which considers interdependent valuations [2]. In such a model, every bidder is not sure of his valuations and it depends on the private signals of all the bidders. There are results in this model for simple settings, but not much is known for the multiple object auction case. Interested readers may find Ref. 2 useful.

function if for every $S \in \mathbb{S}$ we have $v(S) = \max_{j \in S} v(\{j\})$. In this case, a buyer is interested in at most one good. So, when presented with a bundle of goods, he picks the one with the highest value amongst those goods. For example, consider a potential buyer is choosing to buy a house amongst a set of houses. Such a valuation function can be very succinctly expressed: every valuation vector can only take m values—one for each good.

2. *Homogeneous Goods.* A valuation function v is a *homogeneous goods* valuation function if for every $S, T \in \mathbb{S}$ with $|S| = |T|$, we have $v(S) = v(T)$. In this case, goods are identical. So, only the size of the bundle dictates the value. Here also, the value function can only take m values, and thus can be succinctly represented.

Below, we give an example to illustrate how valuation functions may look like in a combinatorial auction model. Suppose there are three agents $\{1, 2, 3\}$ and two goods $\{1, 2\}$. A possible valuation profile is shown in Table 1.

There are two parts to the problem: (i) an allocation rule (who should get which bundles) (ii) a payment rule (who should pay how much). While there are many allocation rules possible, the one we discuss here (and which is the center of attention in the literature) is the efficient allocation rule. An allocation $X = (X_0, X_1, \dots, X_n)$ is partitioning of the set of goods where $X_i \in \mathbb{S}$ denotes the bundle allocated to bidder i and X_0 is the unallocated set of goods. Let $\mathbb{X}$ be the set of all allocations. Given a profile of valuation functions $v = (v_1, \dots, v_n)$, an allocation X is efficient if

$$\sum_{i=1}^n v_i(X_i) \geq \sum_{i=1}^n v_i(Y_i) \quad \forall Y \in \mathbb{X}.$$

An efficient allocation rule chooses an efficient allocation for every profile of valuations. Computing an efficient allocation for a given valuation profile is a computationally hard problem [16], but some special cases can be solved easily [17].

A payment rule on the other hand decides how much a bidder pays. If a bidder i has valuation function v_i and he is allocated bundle X_i and pays p_i , then his net utility is given by $v_i(X_i) - p_i$. Such a form of utility function is called the *quasilinear* utility function.

The problem is that the seller has no idea about the valuation functions of the buyer. So, he elicits this information from the buyers in various forms. For example, he may ask the buyers, directly, the valuation function. This is referred to as the *direct* mechanism in auction theory. For many purposes, it is without loss of generality to focus on the direct mechanisms. This is known as the *revelation principle*—for a formal definition of revelation principle, the readers are referred to any standard book in microeconomics, for example, Ref. 18.

However, there is an incentive problem here. The buyer may get higher net utility by lying about his valuation function. For this, the payment and allocation rules need to be designed carefully. Fortunately, for the efficient allocation rule, there is a family of payment functions, which can overcome this incentive problem. These are called the *Groves payment rules* [19]. Suppose $v = (v_1, \dots, v_n)$ is the valuation profile of the buyers. Denote by v_{-i} the valuation profile of buyers other than i . Then, payment of buyer i in a Groves payment rule is given by

$$p_i^G(v) = h_i(v_{-i}) - \sum_{j \neq i} v_j(X_j^*),$$

where X^* is an efficient allocation and h_i is any arbitrary function from valuations of other buyers to real numbers. A particular choice of h_i function stands out. Suppose $h_i(v_{-i}) = \max_X \sum_{j \neq i} v_j(X_j)$. Then the payment

Table 1. Valuations in a Combinatorial Auction

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1, 2\}$
v_1	0	8	9	12
v_2	0	6	8	14
v_3	0	6	5	6

function becomes

$$p_i^{\text{vcg}}(v) = \max_X \sum_{j \neq i} v_j(X_j) - \sum_{j \neq i} v_j(X_j^*).$$

This particular payment rule is popularly known as the *pivotal* payment rule or the *Vickrey–Clarke–Groves (VCG)* payment rule [19–21]. The VCG payment rule along with the efficient allocation rule is called the *VCG mechanism*. Note that the net utility of buyer i in the VCG mechanism is given by

$$\begin{aligned} \pi_i^{\text{vcg}}(v) &= v_i(X_i^*) - p_i^{\text{vcg}}(v) \\ &= \max_X \sum_{j \in N} v_j(X_j) - \max_X \sum_{j \neq i} v_j(X_j). \end{aligned}$$

In other words, every buyer gets a net utility equal to his marginal contribution—the increase in total value of all the buyers when he participates in the auction. In the example in Table 1, the efficient allocation X^* is to allocate good 1 to buyer 1 and good 2 to buyer 2. So the externality of buyer 3 is zero. The externality of buyer 1 is $(8 + 6) - 8 = 6$ and that of buyer 2 is $(9 + 6) - 8 = 7$. Hence, the VCG payments of buyers 1, 2, and 3 are 6, 7, and 0 respectively.

The VCG mechanism is central in economic theory due to various reasons [22]. For example, it satisfies the following properties.

1. *Implements Efficiency.* Every buyer has no incentive to lie about their valuations in the VCG mechanism, that is, bidding truthfully is a dominant strategy.
2. *Individual Rationality.* The VCG mechanism satisfies the property that every buyer has nonnegative net utility. To see this, note that $\max_X \sum_{j \in N} v_j(X_j) \geq \max_X \sum_{j \neq i} v_j(X_j)$, and hence, $\pi_i^{\text{vcg}}(v) \geq 0$.
3. *Losers Pay Nothing.* In the VCG mechanism, if a buyer does not win any bundle, then he pays nothing. This follows from the zero outside option and bundle monotonicity assumptions.
4. *Nonnegative Payments.* Every buyer pays nonnegative amount in the VCG mechanism, that is, no buyer

is awarded or given money. This follows from the fact that $\max_X \sum_{j \neq i} v_j(X_j) \geq \sum_{j \neq i} v_j(X_j^*)$.

However, the VCG mechanism suffers from various practical drawbacks also [5, 14, 23]. We list some below.

1. *Computational.* The VCG mechanism requires one to solve $n + 1$ optimization problems. These optimization problems are computationally hard for arbitrary valuation functions [16].
2. *Communication.* The VCG mechanism requires that every buyer communicates his entire valuation function to the seller. The fact that every buyer's valuation function can have an exponential number of bundle values makes this communication difficult.
3. *Price Discovery and Privacy.* Irrespective of the compelling theoretical property of dominant strategy incentive compatibility, one does not expect buyers to reveal their true values in any mechanism. Privacy of valuations is an important aspect for buyers in practice. Further, the natural process of price discovery is absent in the (sealed-bid) VCG mechanism.

As Ausubel and Milgrom [14] point out, there are some economic issues with the VCG mechanism too—losing bidders have incentives to collude, the outcome in the VCG mechanism is not *stable* and so on. However, these are inherent problems in any Groves mechanism (which implements the efficient allocation rule). For example, Rastegeri *et al.* [24] point out that any Groves mechanism will not be *stable* for a large class of valuation functions. Further, whenever these economic issues are not present for any Groves mechanism, the VCG mechanism is the natural choice for use.

It is difficult to come up with an *algorithm* to overcome the computational and communication overheads of the VCG mechanism. However, algorithms can be designed to implement (achieve) the VCG mechanism, which is more transparent (and respects privacy to some extent) than

the VCG mechanism we just described. These algorithms are precisely the *iterative VCG combinatorial auctions*. Since they implement the same allocation (efficient) and the same payment (VCG) as the VCG mechanism they inherit the incentive properties of the VCG mechanism. One needs some restrictions on the algorithm/iterative auction to inherit the incentive properties of the VCG mechanism [7]. We do not discuss this issue in detail here. In brief, we need some activity rules in the auction to ensure that every bidder bids in a *myopically optimal* manner (as we will see later), and this leads to the VCG outcome. With these activity rules in place, it is also optimal (in a global sense) for every bidder to bid in this manner. We will not discuss the incentive issues in these iterative VCG combinatorial auctions, and assume that bidders are always bidding truthfully. But, we keep in mind that truthful bidding is an equilibrium in these auctions under some mild restrictions on bidding activity.

FOUNDATIONS OF ITERATIVE VCG COMBINATORIAL AUCTIONS

We start by laying the foundations of designing iterative VCG combinatorial auctions. Let us first recall that for the single object case, the Vickrey (second-price) auction has a counterpart in the iterative auction world—the English auction. This auction sheds light on how one can generalize this to the multiple object case. Suppose $v_1 \geq v_2 \geq \dots \geq v_n \geq 0$ be the values of n buyers for the single indivisible good. The English auction starts from a low price (say zero) and asks every buyer if they are interested in the object at this price. If more than one buyer is interested in the good, then the price is increased, and the process is repeated till exactly one buyer is interested. At this point, this buyer wins the good at the current price. It can be shown the buyers have no incentive to lie in this auction, that is, buyers should express interest in the good till the price reaches their respective valuations. Further, if the amount of price increase is low enough, the auction terminates at the second-highest valuation, and

hence implements the same outcome as the Vickrey auction. There are other interesting features of this auction.

1. The buyers see a single price (in $\mathbb{R}_+$), which is increased in every iteration, and this makes the auction very transparent.
2. Call a price a *competitive equilibrium* (CE) price if exactly one buyer demands the object at that price. Any price between v_1 and v_2 is a competitive equilibrium price. Indeed, the auction terminates at the lowest such CE price (ignoring issues of ties). Moreover, such an equilibrium price satisfies the property that if we remove any one of the buyers, this remains an equilibrium price of the new economy as well: v_2 is a CE price of an economy without any buyer. As we will see next, that such pricing is pivotal in designing iterative VCG combinatorial auctions.

Complex Pricing

Unfortunately, for general combinatorial auction models, the simple pricing of the English auction for the single object auction model is not enough. One needs richer prices [7]. These richer and complex prices are essential even without strategic considerations, that is, to simply compute efficiency [25]. We will mainly be concerned with four types of prices.

1. *Linear and Anonymous*. In this pricing, a price is maintained for every good, and every buyer sees the same price. The price of a bundle is the sum of prices of goods in that bundle. Formally, a linear and anonymous price p is a vector in $\mathbb{R}_+^m$, where $p(j)$ indicates the price of good $j \in M$ and price of a bundle of goods S is just $\sum_{j \in S} p(j)$.
2. *Linear and Nonanonymous*. In this pricing, a price is maintained for every good and every buyer, that is, every buyer sees a different linear price vector. Formally, a linear and nonanonymous price p is a vector in $\mathbb{R}_+^{m \times n}$, where $p_i(j)$ indicates the price of good $j \in M$

for buyer i , and price of a bundle of goods S for buyer i is just $\sum_{j \in S} p_i(j)$.

3. *Nonlinear and Anonymous.* In this pricing, a price is maintained for every bundle of good and every buyer sees the same price. Formally, a nonlinear and anonymous price p is a vector in $\mathbb{R}_+^{|\mathbb{S}|}$, where $p(S)$ indicates the price of bundle of goods S .
4. *Nonlinear and Nonanonymous.* This is the most complex of pricing schemes. In this pricing, a price is maintained for every bundle of good and every buyer sees a different price vector. Formally, a nonlinear and nonanonymous price p is a vector in $\mathbb{R}_+^{|\mathbb{S}| \times n}$, where $p_i(S)$ indicates the price of bundle of goods S for buyer i .

Such different types of pricing schemes were first discussed in Ref. 8. As we will discuss, most combinatorial auction models require nonlinear and nonanonymous prices for iterative VCG combinatorial auctions. However, there are instances where simpler prices such as linear and anonymous prices are sufficient to design iterative VCG combinatorial auctions.

Competitive Equilibrium Prices

We will define some notions in this section. These notions are defined for nonlinear and nonanonymous prices, but can be defined for other types of prices as well. From now on, we fix a profile of valuation functions $v = (v_1, \dots, v_n)$. For any price $p \in \mathbb{R}^{|\mathbb{S}| \times n}$, define the (*maximum*) *payoff* of buyer i as

$$\pi_i(p) = \max_{S \in \mathbb{S}} [v_i(S) - p_i(S)]$$

and the *demand set* of buyer i as

$$D_i(p) = \{S \in \mathbb{S} : v_i(S) - p_i(S) = \pi_i(p)\}.$$

Similarly, for any price $p \in \mathbb{R}^{|\mathbb{S}| \times n}$, the *revenue* of the seller is

$$\pi^s(p) = \max_{X \in \mathbb{X}} \sum_{i=1}^n p_i(X_i)$$

and the *supply set* of the seller is

$$F(p) = \left\{ X \in \mathbb{X} : \sum_{i=1}^n p_i(X_i) = \pi^s(p) \right\}.$$

Like in any economic equilibrium, the supply and demand should match. Here, this translates to saying that we can pick bundles in the demand sets of buyers in such a manner that it leads to an allocation in the supply set of the seller.

Definition 1. A price vector p and an allocation X form a *competitive equilibrium* if $X_i \in D_i(p)$ for all $i \in N$ and $X \in F(p)$. If (p, X) is a competitive equilibrium then p is called a competitive equilibrium (CE) price vector.

A CE always exists if the price vector is chosen to be nonlinear and nonanonymous: choose X to be an efficient allocation and p_i to be equal to v_i for all $i \in N$. If price vectors are restricted to be linear and anonymous, then a CE exists under a condition called *gross substitutes* [26,27]. The gross substitutes condition requires goods to be substitutes of each other. This condition is satisfied for unit demand valuations and for homogeneous goods valuations if the marginal value of a unit is nonincreasing.

If (p, X) is a CE, then X is an efficient allocation. This fact can be shown by formulating the problem of finding an efficient allocation as a linear program. To do so, define for every $i \in N$ and every $S \in \mathbb{S}$ a binary variable $x_i(S) \in \{0, 1\}$, where $x_i(S) = 1$ indicates that buyer i gets bundle S and $x_i(S) = 0$ indicates that buyer i does not get bundle S . Also, define a binary variable $z(X) \in \{0, 1\}$ for every $X \in \mathbb{X}$, which indicates if allocation X is chosen or not. Then, the problem of finding an efficient allocation can be formulated as follows [8].

$$\begin{aligned} V(N) = \max_{x, z} & \sum_{i=1}^n \sum_{S \in \mathbb{S}} v_i(S) x_i(S) \\ \text{s.t.} & \end{aligned} \tag{1}$$

$$\begin{aligned}
\sum_{S \in \mathbb{S}} x_i(S) &= 1 \quad \forall i \in N \\
\sum_{X \in \mathbb{X}} z(X) &= 1 \\
x_i(S) &= \sum_{X \in \mathbb{X}: X_i = S} z(X) \quad \forall i \in N, \forall S \in \mathbb{S} \\
x_i(S) &\in \{0, 1\} \quad \forall i \in N, \forall S \in \mathbb{S} \\
z(X) &\in \{0, 1\} \quad \forall X \in \mathbb{X}.
\end{aligned}$$

It is easy to see that Equation (1) finds an efficient allocation. The constraints of Equation (1) ensure that an allocation is chosen and the objective function maximizes the total value of buyers over all allocation. As we will show later, there is another simpler formulation possible than formulation (1)—see formulation (5). However, formulation (1) has some nice features as we discuss below.

The linear programming relaxation of Equation (1) admits integer solution [8]. Hence, the problem of finding an efficient allocation reduces to the following linear program.

$$\begin{aligned}
V(N) &= \max_{x, z} \sum_{i=1}^n \sum_{S \in \mathbb{S}} v_i(S) x_i(S) \\
\text{s.t.} \quad & \\
\sum_{S \in \mathbb{S}} x_i(S) &= 1 \quad \forall i \in N \\
\sum_{X \in \mathbb{X}} z(X) &= 1 \\
x_i(S) &= \sum_{X \in \mathbb{X}: X_i = S} z(X) \quad \forall i \in N, \forall S \in \mathbb{S} \\
x_i(S) &\geq 0 \quad \forall i \in N, \forall S \in \mathbb{S} \\
z(X) &\geq 0 \quad \forall X \in \mathbb{X}.
\end{aligned} \tag{2}$$

This linear program has an exponential (in number of goods and buyers) number of variables and constraints. So, in practice, it is not useful. However, it serves as an ideal tool to design an iterative VCG combinatorial auction. For this, it is useful to look at its dual given below.

$$\begin{aligned}
V(N) &= \min_{p, \pi, \pi^s} \sum_{i \in N} \pi_i + \pi^s \\
\text{s.t.} \quad & \\
p_i(S) + \pi_i &\geq v_i(S) \quad \forall i \in N, \forall S \in \mathbb{S} \\
\pi^s - \sum_{i \in N} p_i(X_i) &\geq 0 \quad \forall X \in \mathbb{X}
\end{aligned} \tag{3}$$

The interpretation of the dual variables are clear: p is the price vector, π is the payoff vector of buyers, and π^s is the revenue of the seller. The following result establishes the connection between CE and these formulations.

Theorem 1 [8]. *A pair (p, X) is a competitive equilibrium if and only if X corresponds to the optimal solution of Linear Program (2) and p corresponds to the optimal solution of Linear Program (3).*

The proof of this result involves writing down the complementary slackness conditions and showing their equivalence to the CE conditions. A consequence of this result is that finding an efficient allocation is equivalent to finding a CE price vector. Of course, a CE price vector may be an exponential (in number of goods) sized vector. But the theorem also works in settings where simpler (anonymous and linear) CE price vectors exist. For example, in unit-demand settings, anonymous and linear CE price vectors exist, and finding, there, an efficient allocation is equivalent to finding such a CE price vector.

The foundations of iterative auctions are CE price vectors, in the sense that every iterative auction (appropriately defined) searches for a CE price vector, and finds an efficient allocation. However, the incentive properties of the VCG mechanism is not inherited by every CE price vector. As we will show next, there is a large class of CE prices from which VCG payments can be deduced. However, there are CE price vectors that correspond directly to the VCG payments if a condition on a valuation function is satisfied. We discuss two such conditions. For every $K \subseteq N$, let $V(K)$ denote the objective function value of the optimal solution of Equation (2), when Equation (2) is solved by considering agents in K only.

Definition 2 *Agents are substitutes if $V(N) - V(K) \geq \sum_{j \in N \setminus K} [V(N) - V(N \setminus \{j\})]$ for all $K \subseteq N$. Agents are submodular if $V(T) - V(T \setminus \{j\}) \geq V(K) - V(K \setminus \{j\})$ for all $T \subseteq K \subseteq N$ and for all $j \in N$.*

It is easy to see that if agents are submodular, then agents are substitutes. However, the converse is not true (Ref. 7 contains examples).

We say that a CE price vector is *minimal* if it is a solution to the following linear program (4).

$$\begin{aligned} \Pi^{\max} &= \max_{p, \pi, \pi^s} \sum_{i \in N} \pi_i \\ \text{s.t.} \quad & \\ & \sum_{i \in N} \pi_i + \pi^s = V(N) \\ & p_i(S) + \pi_i \geq v_i(S) \quad \forall i \in N, \forall S \in \mathbb{S} \\ & \pi^s - \sum_{i \in N} p_i(X_i) \geq 0 \quad \forall X \in \mathbb{X} \end{aligned} \quad (4)$$

Formulation (4) finds a CE price vector, which gives the maximum total payoff to agents over all CE price vectors. By maximizing the total payoff to agents, we also minimize their total payments over all CE price vectors. This results in finding the minimal CE price vector.

Theorem 2 [8]. *Agents are substitutes if and only if for every $i \in N$ we have*

$$p_i^{\text{vcg}}(v) = p_i^{\min}(X_i),$$

where $p^{\min}$ is an optimal solution of Linear Program (4) and X corresponds to an optimal solution of Linear Program (2) that is, $p^{\min}$ is a minimal CE price vector and X is an efficient allocation.

In general, the VCG mechanism requires one to solve the linear program (2) $n + 1$ times. The implication of this theorem is that under the agents are substitutes condition, one needs to solve only the linear programs (2) and (4) to find the outcomes of the VCG mechanism. The agents are substitutes condition is satisfied if goods are somewhat substitutes (and not complements). For example,

it is satisfied in the unit demand case and in the homogeneous goods case if the marginal value of a unit is nonincreasing for buyers.

It is worth commenting that there is an integer programming formulation of (2). Let $x_i(S) \in \{0, 1\}$ denote if agent i gets bundle S or not. Then an alternative formulation of finding an efficient allocation is as follows.

$$\begin{aligned} V(N) &= \max_x \sum_{i \in N} \sum_{S \in \mathbb{S}} v_i(S) x_i(S) \\ \text{s.t.} \quad & \\ & \sum_{i \in N} \sum_{S \in \mathbb{S}: j \in S} x_i(S) \leq 1 \quad \forall j \in M \\ & \sum_{S \in \mathbb{S}} x_i(S) = 1 \quad \forall i \in N \\ & x_i(S) \in \{0, 1\} \quad \forall i \in N, \forall S \in \mathbb{S}. \end{aligned} \quad (5)$$

The constraints in Equation (5) ensures that no object is assigned to more than one buyer and every agent is assigned to exactly one bundle (possibly the empty bundle). The objective function maximizes the total value from the allocation specified by the feasibility constraints. Though this is an integer programming formulation of the efficient allocation problem, it involves fewer number of constraints and variables than formulation (2).

Theorem 2 is silent on two accounts with regard to the power of CE price vectors in terms of extracting the VCG mechanism information: (i) it does not say how *informative* nonminimal CE price vectors can be; (ii) it does not say what happens when the agents are substitutes condition does not hold. These two difficulties can be overcome by generalizing the notion of CE prices.

Universal Competitive Equilibrium Prices

We introduce the notion of universal competitive equilibrium (UCE) prices. It is a useful concept to analyze and design iterative VCG auctions. For any nonlinear and nonanonymous price vector p , for every $K \subseteq N$ ($K \neq \emptyset$) let p^K denote the restriction of price vector p to the set of agents K . Define for every $i \in N$, $N_{-i} = N \setminus \{i\}$ and $\mathbb{N} = N \cup_{i \in N} N_{-i}$. An *economy* is defined by a set of agents N , a

set of goods M , and the valuation functions $v \equiv \{v_i\}_{i \in N}$ of agents.

Definition 3. A competitive equilibrium price vector p of economy (N, M, v) is a *universal competitive equilibrium (UCE)* price vector of this economy if p^K is a competitive equilibrium price vector for economy $(K, M, \{v_i\}_{i \in K})$ for every $K \in \mathbb{N}$.

Note that the definition does not require the price vector to be nonlinear and nonanonymous. A UCE price vector can be defined for any type of price vector. Similarly, the result we state relating the UCE price vector to the VCG mechanism is independent of the type of UCE price vector (i.e., the representation of the UCE price vector). A nonlinear and nonanonymous UCE price vector always exists: let $p_i := v_i$ for all $i \in N$.

We give three examples of UCE prices below. We note that anonymous and linear UCE prices are available in the first two examples below.

1. *Single Good.* In the single indivisible good case, there is a unique anonymous UCE price. Take an example with four buyers with values 10, 8, 6, 4 for the good. Then, any price in $[8, 10]$ is an anonymous CE price. However, the price 8 is the unique anonymous UCE price. To see this, note that 8 continues to be a CE price even if we remove any one of the buyers from the economy. As we know, 8 is also the price paid by the highest bidder in the second-price Vickrey auction.
2. *Unit Demand.* In the unit demand valuation case, there is a unique linear and anonymous UCE price vector [28]. Consider an example with two goods and three buyers. The valuations of the buyers on goods are given as follows: $v_1(1) = 5, v_1(2) = 3, v_2(1) = 3, v_2(2) = 4, v_3(1) = 2, v_3(2) = 2$. Since this is a unit demand setting, the valuations of buyers on bundles can be derived from this. It is easy to verify that the unique efficient allocation for this example is buyer 1 gets good 1 and buyer 2 gets

Table 2. Valuations in a Combinatorial Auction

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1, 2\}$
v_1	0	3	0	3
v_2	0	0	6	6
v_3	0	0	2	4

good 2. The unique linear and anonymous UCE price vector is $(2, 2)$. At this price vector, the demand set of buyer 3 includes all possible bundles of goods, including $\emptyset$. The demand set of buyer 1 consists of good 1 and that of buyer 2 consists of good 2. Using these facts, one can easily verify that $(2, 2)$ is a linear and anonymous UCE price vector.

3. *General Valuation.* Consider the example in Table 2 with three buyers and two goods. Consider the following nonlinear and nonanonymous price vector for this example: $p_1(\{1\}) = 2 = p_1(\{1, 2\})$, $p_1(\{2\}) = 0$, $p_2(\{1\}) = 0$, $p_2(\{2\}) = 4 = p_2(\{1, 2\})$, $p_3(\{1\}) = 0$, $p_3(\{2\}) = 2$, $p_3(\{1, 2\}) = 4$. The efficient allocation for this example is buyer 1 gets good 1 and buyer 2 gets good 2. The reader can verify that this is a UCE price vector.

The following important theorem illustrates the relationship between UCE price vectors and VCG payments.

Theorem 3 [7]. Suppose (p, X) is a CE of economy (N, M, v) . The VCG payments of agents in this economy can be computed from (p, X) if and only if p is a UCE price vector of this economy. Moreover, if p is a UCE price vector then for every $i \in N$,

$$p_i^{\text{VCG}}(v) = p_i(X_i) - [\pi^s(p) - \pi^s(p^{N-i})].$$

This result says that UCE prices (with an efficient allocation) are necessary and sufficient for computing VCG payments from CE prices and efficient allocation. Moreover, it says that the bidders need to be given discounts from their CE prices to obtain VCG payments. Note that these discounts are zero if the UCE price vector is anonymous—this

is the case for single good setting and the unit demand setting, where anonymous UCE price vectors are known to exist.

For the example in Table 2, a UCE price vector p is

$$\begin{aligned} p_1(\{1\}) &= 2 = p_1(\{1, 2\}), p_1(\{2\}) = 0 \\ p_2(\{1\}) &= 0, p_2(\{2\}) = p_2(\{1, 2\}) = 4 \\ p_3(\{1\}) &= 0, p_3(\{2\}) = 2, p_3(\{1, 2\}) = 4. \end{aligned}$$

The efficient allocation for this example is to allocate good 1 to buyer 1 and good 2 to buyer 2. We can compute $\pi^s(p) = 6$, $\pi^s(p^{N-1}) = 4$, $\pi^s(p^{N-2}) = 4$, $\pi^s(p^{N-3}) = 6$. Hence, the discounts of buyers 1, 2, and 3 are respectively, 2, 2, and 0. So, the VCG payments of buyers 1, 2, and 3 are $2 - 2 = 0$, $4 - 2 = 2$, and 0 respectively.

The result in Theorem 3 also gives hints on how to achieve the outcome of the VCG mechanism using CE prices—we need to search for UCE prices. Mishra and Parkes [7] give a class of auctions using nonlinear and nonanonymous prices for general valuations of bidders, which search for UCE prices and achieve the VCG outcome. We revisit one of the auctions.

iBundle—EXTEND AND ADJUST (iBEA) AUCTION

We present an iterative auction for general valuations of combinatorial auctions. The auction is an extension of the iBundle auction in Ref. 10 and the auction in Ref. 9. This description follows along the lines of Ref. 7.

In every iteration of the auction, a nonlinear and nonanonymous price vector is maintained. Denote such a price vector in round t of this auction as p^t . Given such a price vector, the building block of the auction is to solve an integer programming problem, commonly referred to as the *winner determination problem* (WDP).

The Winner Determination Problem

In this section, we formalize the WDP. In a WDP, we are given a nonlinear and nonanonymous price vector $p \in \mathbb{R}_+^{n \times |S|}$ and

the demand sets of buyers at this price vector. The objective is to determine an allocation that *maximizes* the revenue of the seller at this price such that every buyer is either assigned a bundle from his demand set or the $\emptyset$ bundle, that is, find an allocation X at price vector p , such that $\pi^s(p) = \sum_{i \in N} p_i(X_i)$ subject to $X_i \in D_i(p) \cup \{\emptyset\}$. This problem can be formulated as an integer program as follows. Let $x_i(S) \in \{0, 1\}$ denote whether agent i is allocated a bundle $S \in D_i(p) \cup \{\emptyset\}$. The following integer program solves the WDP at price vector p .

$$\begin{aligned} \pi^s(p) &= \max_x \sum_{i \in N} \sum_{S \in D_i(p) \cup \{\emptyset\}} p_i(S) x_i(S) \\ \text{s.t.} & \\ \sum_{i \in N} \sum_{S \in D_i(p) \cup \{\emptyset\}} x_i(S) &\leq 1 \quad \forall j \in M \\ \sum_{S \in D_i(p) \cup \{\emptyset\}} x_i(S) &= 1 \quad \forall i \in N \\ x_i(S) &\in \{0, 1\} \quad \forall i \in N, \forall S \in D_i(p) \cup \{\emptyset\}. \end{aligned} \tag{6}$$

The constraints of formulation (6) ensure that every agent is assigned either a bundle from his demand set or the empty bundle, and every good is assigned to at most one agent. The objective function ensures that we choose an allocation that maximizes the revenue to the seller at this price vector. It is worth noting that formulation (6) becomes equivalent to formulation (5) if the price vector is the same as the valuation vector.

An extensive literature exists on the computational aspects of the WDP—a good reference is Ref. 17. We want to note here that the WDP, in general, is NP-hard to solve, but some tractable instances exist [16, 17].

In the auction we discuss next, we will solve instances of WDP for any economy $K \in \mathbb{N}$. We will therefore denote WDP^K to denote the winner determination problem of economy with agents $K \in \mathbb{N}$. We denote the allocation chosen by the WDP^K at price vector p^K as $X^{p,K}$, and call it an *optimal allocation* at p^K . Let $L(p, K) = \{i \in K : X_i^{p,K} = \emptyset \text{ and } \emptyset \notin D_i(p^K)\}$ be the set of agents who do not get any bundle in the WDP^K and their demand sets do not contain the $\emptyset$. We call such agents

unsatisfied agents at price vector p . The revenue of the seller from the WDP of economy with agents K at price vector p will be denoted as $\pi^s(p^K)$ —it is the value of the optimal solution of Equation (6) at price vector p^K for economy with agents K .

Description of the iBEA Auction

For simplicity of description and analysis, we assume that the valuations of agents are integers. However, the ideas of the auction extends to arbitrary real-valued valuation functions. Assuming any real valuation brings additional convergence problems due to bid increments, which require cumbersome analysis techniques. These problems disappear if one assumes integer valuations.

The iBEA auction orders the elements of the set $\mathbb{N}$. Suppose the ordering is N , followed by N_{-1} , followed by $N_{-2}, \dots, N_{-n}$. Denote $N \equiv N_{-0}$.

iBEA AUCTION

- S0 Initiate the round number $t = 0$ and let p^0 be the zero price vector.
- S1 For $i = 0, 1, \dots, n$
 - S1.1 Ask demand sets of buyers at price vector p^t .
 - S1.2 Solve $\text{WDP}^{N_{-i}}$ at price vector p^t (restricted to N_{-i}).
 - S1.3 If $L(p^t, N_{-i}) = \emptyset$ (i.e., no unsatisfied agents), then $i = i + 1$ and repeat from Step S1 with $p^{t+1} = p^t$ and $t = t + 1$.
 - S1.4 If $L(p^t, N_{-i}) \neq \emptyset$, then set $p_j^{t+1}(S) = p_j^t(S) + 1$ for all $j \in L(p^t, N_{-i})$ and for all $S \in D_j(p^t)$. Repeat from Step S1.1 with $t = t + 1$.
- S2 Let X be the allocation chosen by $\text{WDP}^{N_{-0}}$ in the final round when $i = 0$. Let the final price vector in the auction be p . Then, the final allocation of the auction is X and payment of agent j is $p_j(X_j) - [\pi^s(p) - \pi^s(p^{N_{-j}})]$ (discounts from the terminating prices).

The iBEA auction searches for a UCE price vector, and as we will show next, it terminates at a UCE price vector. We illustrate the auction with an example in Table 3.

The last columns indicate seller's revenue in economy N (main economy) and *marginal economies* (economies with agents N_{-1}, N_{-2} , and N_{-3} respectively). Prices in parenthesis for a buyer and a bundle indicate that such a bundle is in the demand set of the respective buyer at that price. We see in Table 3 that iBEA solves the WDP of main and marginal economies in every round and terminates at a UCE price vector. Discounts are then calculated for every bidder to compute the VCG payments.

We say that the iBEA auction *terminates after the main economy* if the auction does not see any price increase for $i > 0$ in Step S1, that is, we reach a UCE price vector after finding a CE price vector of economy with all the agents. The main result is that this auction terminates with a UCE price vector and achieves the VCG outcome. However, under a variety of conditions, it terminates as soon as a CE price vector of main economy (economy with all the agents) is discovered.

Theorem 4. *Suppose agents submit their demand sets truthfully in an iBEA auction. Then, the following statements are true.*

1. *The iBEA auction terminates at a UCE price vector and the outcome is the same as that of the VCG mechanism [7].*
2. *If the agents are substitutes condition holds, then the iBEA auction terminates after the main economy with the VCG outcome [7].*
3. *If the agents are submodular condition holds, then the iBEA auction terminates after the main economy with the VCG outcome, and the discount given to every agent is zero [7,9].*

CONCLUDING REMARKS

1) General Class of Ascending Auctions. Mishra and Parkes [7] discuss a more general class of ascending auctions for general valuation functions. This class contains the iBEA auction and generalizations of auctions in Ref. 11. Results (1) and (2) in Theorem 4 hold for any auction in this general class.

Table 3. Progress of *i*BEA for an Example

Values	Buyer 1			Buyer 2			Buyer 3			Seller revenue in main and marginal economies
	{1}	{2}	{1, 2}	{1}	{2}	{1, 2}	{1}	{2}	{1, 2}	
	3	0	3	0	6	6	0	2	4	
1	(0)	0	(0)	0	(0)	(0)	0	0	(0)	{0,0,0,0}
	WDP of N . WDP selects allocation $\{\{1\}, \{2\}, \emptyset\}$. Buyer {3} is unsatisfied.									
2	(0)	0	(0)	0	(0)	(0)	0	0	(1)	{1,1,1,0}
	WDP of N . WDP selects $\{\emptyset, \emptyset, \{1, 2\}\}$. Buyers {1, 2} are unsatisfied.									
3	(1)	0	(1)	0	(1)	(1)	0	0	(1)	{2,1,1,2}
	WDP of N . WDP selects $\{\{1\}, \{2\}, \emptyset\}$. Buyer {3} is unsatisfied.									
4	(1)	0	(1)	0	(1)	(1)	0	(0)	(2)	{2,2,2,2}
	WDP of N . WDP selects $\{\{1\}, \{2\}, \emptyset\}$. Buyer {3} is unsatisfied.									
5	(1)	0	(1)	0	(1)	(1)	0	(1)	(3)	{3,3,3,2}
	WDP of N . WDP selects $\{\emptyset, \emptyset, \{1, 2\}\}$. Buyers {1, 2} are unsatisfied.									
6	(2)	0	(2)	0	(2)	(2)	(0)	(1)	(3)	{4,3,3,4}
	WDP of N . WDP selects $\{\{1\}, \{2\}, \emptyset\}$. Buyer {3} is unsatisfied.									
7	(2)	0	(2)	0	(2)	(2)	(0)	(2)	(4)	{4,4,4,4}
	CEs of economies with agents N , N_{-2} , and N_{-3} are reached. $\{\{1\}, \{2\}, \emptyset\}$ is an efficient allocation. WDP of N_{-1} . WDP selects $\{\emptyset, \emptyset, \{1, 2\}\}$. Buyer {2} is unsatisfied.									
8	(2)	0	(2)	0	(3)	(3)	(0)	(2)	(4)	{5,4,4,5}
	WDP of N_{-1} . WDP selects $\{\emptyset, \emptyset, \{1, 2\}\}$. Buyer {2} is unsatisfied.									
9	(2)	0	(2)	0	(4)	(4)	(0)	(2)	(4)	{6,4,4,6}
	A UCE price is reached. Final allocation: $\{\{1\}, \{2\}, \emptyset\}$. Final payment: $p_1(\{1\}) = 2 - [6 - 4] = 0$, $p_2(\{2\}) = 4 - [6 - 4] = 2$, $p_3(\emptyset) = 0$.									

Result (3) holds for *i*BEA auction and the auction in Ref. 11. These auctions can be interpreted to be solving the linear program (2) using a primal dual algorithm. This connection is established explicitly in Ref. 11; also see Refs 10 and 29.

2) Experimental Evidence. In a recent paper, Chen and Takeuchi [30] investigate how the *i*BEA auction performs in practice. They conduct experiments and compare its performance against the sealed-bid VCG auction. The experiment uses humans and robots to study these two auction formats. The main finding of this study is that while the sealed-bid VCG auction achieves greater revenue for the seller, the *i*BEA auction achieves greater bidder profit and efficiency. Further, the *i*BEA auction achieves 100% efficiency in significantly large instances than the sealed-bid VCG auction. This substantiates the use of *i*BEA over its sealed-bid counterpart.

3) Simpler Prices. In general, the *i*BEA auction requires complex nonlinear and

nonanonymous prices. Such complex prices are necessary for general valuations, as we have argued earlier. For simple valuation functions, one may do away with such complex prices. We discuss two such valuation functions.

One is the case of unit demand valuations. There exist ascending auctions, which achieve the VCG outcome using anonymous and linear prices [6].

The other is the case of *homogeneous goods with nonincreasing marginal valuations (NIMV)*. In such a case, the goods are homogeneous, but the marginal value of a unit is nonincreasing for every buyer. In this case, an ascending auction with simple prices exists, which achieves the VCG outcome [12]. This auction implicitly maintains nonlinear and nonanonymous prices [31]. These nonlinear and nonanonymous prices find a succinct representation in this case. It is an *open question* if such auctions with simple prices is

possible in any other interesting valuation functions.

4) Descending Auctions. One wonders if the ascending price trajectories can be replaced by descending ones in these auctions. The answer is not trivial as the search for a UCE price vector has to be done carefully, specially if we have auctions, which use simpler prices. Mishra and Parkes [28] demonstrate that the UCE price concept can be used to design descending price auction counterparts of the auctions in Refs 6 and 12 for unit demand and homogeneous good NIMV settings respectively. These descending auctions maintain simple prices and achieve the VCG outcome. With complex nonlinear and nonanonymous prices, the descending auction counterpart of the *i*BEA auction can be found in Ref. 32. However, this auction has the undesirable feature of terminating at “price=value” (i.e., $p_i(S) = v_i(S)$ for all $i \in N$ and for all $S \in \mathbb{S}$) UCE price vector often. It is an *open question* to design descending auction counterpart of the *i*BEA auction for special cases of valuation functions with simpler prices.

Auctions for Complements. Though the *i*BEA achieves the VCG outcome for general valuations, one needs to run the *i*BEA auction even after achieving the CE price vector of the main economy if agents are substitutes condition does not hold. This is an undesirable feature since in the *i*BEA rounds for $i \neq 0$ in Step S1, the demand sets of bidder i is not used in the auction. Further, the bids of bidders in N_{-i} is only used to compute the payment of bidder i . These undesirable features are present if the agents are substitutes condition does not hold.

The agents are substitutes condition almost implies that goods are substitutes and not complements [8,9]. It is an *open question* as to how to design iterative auctions, which avoid this problem of *i*BEA when goods are complements. A conjecture is that it may be an auction, which is a mixture of ascending and descending prices.

Acknowledgments

I am indebted to David Parkes, with whom I did most of my work on combinatorial auctions, for being a wonderful coauthor and an

even more wonderful friend and person. This survey is dedicated to David. I acknowledge support from Planning and Policy Research Unit (PPRU) at ISI, Delhi.

REFERENCES

1. Cramton P. Ascending auctions. *Eur Econ Rev* 1998;42:745–756.
2. Krishna V. Auction theory. San Diego (CA): Academic Press; 2002.
3. Ball M, Bichler M, Cantillon E, *et al.* Part V: applications. In: Cramton P, Shoam Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge: MIT Press; 2006. pp. 505–612.
4. Myerson RB. Optimal auction design. *Math Oper Res* 1981;6:58–73.
5. Rothkopf MH, Teisberg TJ, Kahn EP. Why are Vickrey auctions rare? *J Polit Econ* 1990;98(1):94–109.
6. Demange G, Gale D, Sotomayor M. Multi-item auctions. *J Polit Econ* 1986;94(4):863–872.
7. Mishra D, Parkes DC. Ascending price Vickrey auctions for general valuations. *J Econ Theory* 2007;132(1):335–366.
8. Bikhchandani S, Ostroy J. The package assignment model. *J Econ Theory* 2002;107(2):377–406.
9. Ausubel LM, Milgrom PR. Ascending auctions with package bidding. *Front Theor Econ* 2002;1(2):1–42.
10. Parkes DC, Ungar LH. Iterative combinatorial auctions: theory and practice. In: *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*. Austin: Texas; 2000. pp. 74–81.
11. de Vries S, Schummer J, Vohra RV. On ascending Vickrey auctions for heterogeneous objects. *J Econ Theory* 2007;132(1):95–118.
12. Ausubel LM. An efficient ascending-bid auction for multiple objects. *Am Econ Rev* 2004;94(5):1452–1475.
13. Ausubel LM, Milgrom P. Chapter 5: the clock-proxy auction: a practical combinatorial auction design. In: Cramton P, Shoam Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge: MIT Press; 2006. pp. 115–138.
14. Ausubel LM, Milgrom P. Chapter 1: the lovely but lonely Vickrey auction. In: Cramton P, Shoam Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge: MIT Press; 2006. pp. 17–40.

15. Lehmann B, Lehmann D, Nisan N. Combinatorial auctions with decreasing marginal utilities. *Games Econ Behav* 2006;55(2):270–296.
16. Rothkopf MH, Pekec A, Harstad RM. Computationally manageable combinatorial auctions. *Manage Sci* 1998;44(8):1131–1147.
17. Lehmann D, Muller R, Pekec A, *et al.* Part III: complexity and algorithmic considerations. In: Cramton P, Shoam Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge: MIT Press; 2006. pp. 295–412.
18. Mas-Colell A, Whinston MD, Green JR. *Microeconomic theory*. Oxford: Oxford University Press; 1995.
19. Groves T. Incentives in teams. *Econometrica* 1973;41(4):617–631.
20. Vickrey W. Counterspeculation, auctions, and competitive sealed tenders. *J Finance* 1961;16(1):8–37.
21. Clarke E. Multipart pricing of public goods. *Public Choice* 1971;11(1):17–33.
22. Moulin H. Characterizations of the pivotal mechanism. *J Public Econ* 1986;31:53–78.
23. Rothkopf MH. Thirteen reasons why Vickrey-Clarke-Groves process is not practical. *Oper Res* 2007;55(2):191–197.
24. Rastegeri B, Condon A, Leyton-Brown K. Revenue monotonicity in deterministic, dominant-strategy combinatorial auctions. *Artif Intell J* 2010. In press.
25. Blumrosen L, Nisan N. Informational limitations of ascending combinatorial auctions. *J Econ Theory* 2010;145(3):1203–1223.
26. Crawford VP, Kelso AS. Job matching, coalition formation, and gross substitutes. *Econometrica* 1982;50:1483–1504.
27. Gul F, Stacchetti E. Walrasian equilibrium with gross substitutes. *J Econ Theory* 1999;87(1):95–124.
28. Mishra D, Parkes DC. Multi-item Vickrey Dutch auctions. *Games Econ Behav* 2009;66(1):326–347.
29. Bertsekas DP. *Linear network optimization: algorithms and codes*. Cambridge (MA): MIT Press; 1991.
30. Chen Y, Takeuchi K. Multi-object auctions with package bidding: an experimental comparison of Vickrey and iBEA. *Games Econ Behav* 2010;68:557–579.
31. Bikhchandani S, Ostroy J. Ascending price Vickrey auctions. *Games Econ Behav* 2006;55(2):215–241.
32. Mishra D, Veeramani D. Vickrey-Dutch procurement auction for multiple items. *Eur J Oper Res* 2007;180(2):617–629.

EFFICIENT USE OF MATERIALS AND ENERGY

VALERIE M. THOMAS
School of Industrial and Systems
Engineering, School of Public
Policy, Georgia Institute of
Technology, Atlanta, Georgia

THE LINKAGE OF ENERGY, MATERIALS, AND ECONOMIC DEVELOPMENT

Worldwide, economic development has been closely tied to increases in energy and material consumption; in many countries a strong correlation has been observed between economic growth and growth in energy consumption. Material use has also risen in line with economic growth and energy use. Some developed economies are significantly more efficient than others in terms of gross domestic product (GDP) per unit of energy consumption, as illustrated in Fig. 1. Moreover, a number of economies are showing increased energy efficiency, with energy use rising less fast than economic production, as illustrated in Fig. 2 for the United States.

From a technological perspective, there is considerable potential for increased energy and material efficiency. The conversion of fuels into electricity, motion, light, and other services is markedly inefficient [1], as illustrated in Table 1. Energy end-use efficiency could be considerably improved with better designs of buildings, products and services, and transportation systems. Material use is perhaps even less efficient. On average, Americans dispose of about 2 kg of municipal waste per person per day; Europeans, on average, dispose of about 1.5 kg per person per day. Although industrial waste production is not well measured, the quantity is at least several times greater than municipal waste. With costs of disposal being relatively low, there has been little incentive to increase material efficiency.

ENERGY AND ENVIRONMENTAL IMPACTS FROM A COMPANY PERSPECTIVE

Through the 1980s, energy and environmental policy and research focused on industrial emissions and specific highly polluting activities. Legislation focused on emissions from power plants and large industrial facilities, and on specific types of emissions such as pesticides, heavy metals, oxides of sulfur and nitrogen, and particulate matter. To some extent, because of progress in reducing the most acute emissions and impacts, and to an even greater extent as the comprehensive nature of energy and environmental challenges became more clear, underlying issues of the environmental impacts of products and their entire life-cycle supply chains have come to the fore [2,3]. While existing legislation focuses on specific pollutants from a risk framework, the focus of environmental analysis has shifted to a comprehensive view of product systems.

To make processes, products, and material management more efficient and environmentally benign requires extensive redesign of processes and products [4]. A framework for environmental design requires evaluation of the environmental implications of different choices of materials, product design, production processes, and service format as well as choices regarding recycling, repair, and disposal.

Life Cycle Assessment

Environmental life-cycle assessment (LCA) is the method by which the environmental impacts of different products or activities can be compared on a systematic basis (see ***Closed-Loop Supply Chains: Environmental Impact***). It is the analytical foundation of many environmental design tools. In principle, an LCA evaluates the entire environmental impact of a product through its life cycle, including manufacturing, use, and disposal. Supply chain models that incorporate environmental life-cycle assessment are

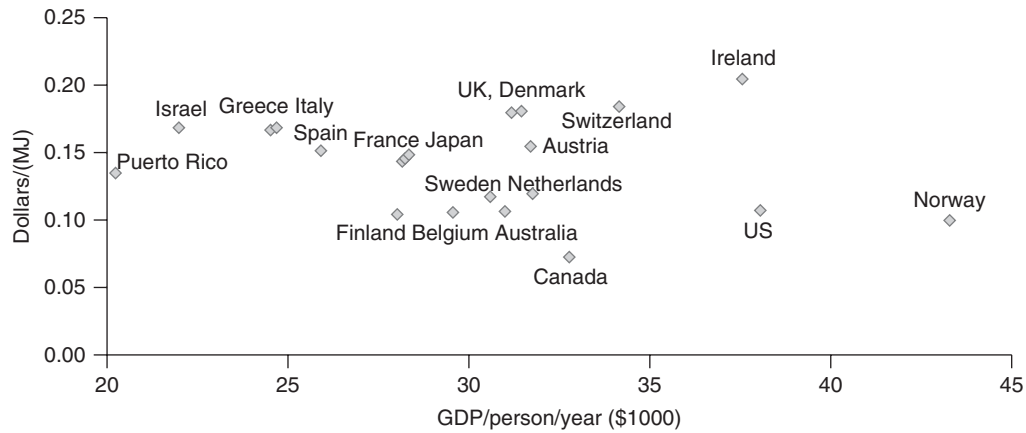


Figure 1. The ratio of gross domestic product (GDP) to total primary energy consumption for a selection of developed countries, shown with respect to GDP per person in each country. Data are for 2006, and GDP is calculated using purchasing power parities. [Source: US Department of Energy, Energy Information Administration, 2008. International Energy Annual 2006].

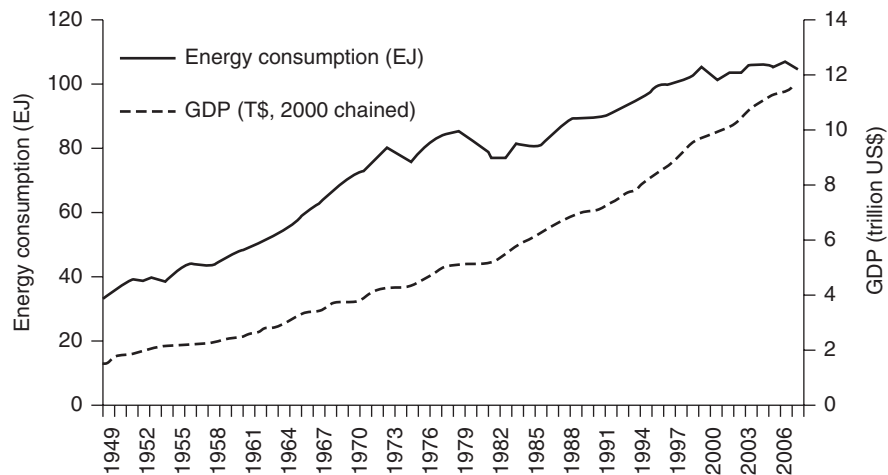


Figure 2. US GDP and total primary energy consumption from 1949 to 2008. [Source: US Department of Energy, Energy Information Administration, 2009. Annual Energy Review 2008].

beginning to provide a basis for evaluating supply chains in terms of a range of environmental impacts as well as cost [5]. Life-cycle assessments of supply chains typically find that industrial and agricultural operations have greater energy and greenhouse gas impact than transportation, although transportation typically is the main use of petroleum in a supply chain. Within the

transportation sector, air transport has considerably greater energy intensity than truck freight, which has significantly greater energy intensity than rail freight, and thus mode shifting from air to truck or from truck to rail can considerably reduce the petroleum consumption of a supply chain.

Policies to limit greenhouse gas emissions could have a significant impact on supply

Table 1. Efficiency of Energy Conversion Technologies.

Conversion Technology	Efficiency (%)	Physical Limitations
Gas and steam turbines	46	Thermodynamics
Internal combustion engines	40	Thermodynamics
Fuel cells	70	Electrode, electrolyte, catalysts
Solar photovoltaics	41	Band gap, crystalline structure
Wind turbines	59	Betz limit
Incandescent lamps	5	Blackbody spectrum
Fluorescent lamps	20	Phosphors, ballasts

Efficiency percentages refer to achieved, predicted, or maximum theoretical limits, and are in some cases subject to change as technology develops.

chain management. Coal and electricity prices are generally more sensitive to the cost of carbon dioxide emissions, and thus a price on greenhouse gas emissions is likely to increase the cost of electricity and coal more substantially than the cost of petroleum. Consequently, supply chain management could begin to favor low-greenhouse gas manufacturing, even if supply chain transport distances increase.

There is considerable scientific understanding of the environmental and health impacts of industrial pollutants, and of a range of industrial and human activities. However, most products have more than one type of environmental impact, and these impacts are usually incommensurate. There are ecosystem impacts, and there are human health impacts. There are varying levels of uncertainty associated with different environmental impacts. There are different types of human health impacts, including acute toxicity, carcinogenicity, and developmental toxicity, that affect different groups of people differently. Some environmental impacts can be compared on a strictly scientific basis, but for many others, a judgment must be made about their relative importance. To integrate across diverse dimensions of environmental

performance, a number of weighted environmental metrics have been developed. For example, a scoring system called *eco-indicator* is designed as a measure of overall environmental impact; human toxicity potential has been developed as a measure of the toxicity of chemical compounds over a range of human health end points [6]. Willingness-to-pay has been developed as an economic measure of environmental health impacts, and quality adjusted life years (QALYs) have been developed as a measure of health outcomes [7]. While a number of software packages provide integration over different environmental impacts, it is also common practice not to aggregate over different types of environmental impacts because the scientific basis for aggregation is not established.

Environmental Management Strategies

While traditional command-and-control environmental policy aims to reduce pollution by setting a cap on emissions or by requiring pollution control devices, some new environmental strategies aim for economically efficient environmental protection. These strategies are related to the efficiency movement in management, with its emphasis on eliminating waste, and could be viewed as part of an overall lean business strategy. Markets for pollutants have been successfully developed to reduce SO₂ emissions in the United States at low cost. Mandatory carbon dioxide markets have been established in the European Union and as of 2009, they were under serious consideration for a national policy in the United States. Other new environmental strategies aim to promote innovation, not simply by controlling emissions but by encouraging the development of processes and product designs that are inherently clean.

It has been suggested that by selling a service rather than a product, firms and consumers could have an incentive to be more efficient with materials and energy. For example, an integrated pest management service might provide crop protection rather than selling pesticides per se. Some efforts to implement this strategy have failed to take hold, with carpet leasing a notable example,

whereas car sharing is seeing commercial success [8].

Another approach is to provide consumers with environmental information about products. Many programs to provide environmental information to consumers have been shown to be ineffective in changing behavior, and thus careful program design is essential [9].

A third approach is to discourage some types of individual or corporate consumption through taxes. Examples might include taxes on fossil fuels or taxes on use of hazardous materials in products. Taxation as an incentive structure has been studied extensively within economics; there is a literature on tax shifting for environmental purposes [10]. Additional approaches include the reuse of products, the recycling of products, efficiency improvements in industry, and a range of voluntary industry/government environmental programs (see *Take-Back Legislation and Its Impact on Closed-Loop Supply Chains; Remanufacturing; Recycling*). The OECD Environment Directorate reports that policy initiatives to promote more sustainable consumption patterns have been only mildly effective [11].

An approach to increasing the material and energy efficiency of manufacturing is industrial symbiosis. In Kalundborg, Denmark, a power plant, a petroleum refinery, a gypsum wallboard manufacturer, a pharmaceutical manufacturer, and the city's heating system all share resources. What would normally be waste products or waste energy is used as input feedstock in other facilities. A range of industrial symbiosis examples have developed worldwide [12].

MATERIAL AND ENERGY EFFICIENCY FROM A GLOBAL PERSPECTIVE

From a more global perspective, the underlying issue is the long-term relationship between technological development, social and economic development, and the environment. Global issues include not only geographically global environmental problems such as climate change and stratospheric ozone depletion but also overall

human health and welfare, the state of the economy, and resource availability. As indicated in Figs 1 and 2, energy use, as well as material use, has a dynamic relationship to economic development, determined not only by specific technological improvements (Table 1), but also by a broader set of economic, social, and policy developments.

Material budgets and cycles are a basic approach to evaluating large-scale environmental impacts and resource issues. A material budget assesses material flow into and out of a single reservoir; a material cycle assesses flows among several reservoirs. Important reservoirs include atmosphere, soil, mines, oceans, built infrastructure, landfills, human tissues, and food chains for humans and other species. The analysis involves not only quantifying use and disposal of materials, but also understanding the mechanisms that control the transport among reservoirs, the mechanisms that relate materials use to impacts, and the mechanisms through which materials use might change.

One aim of material cycle and budget research is to identify recycling opportunities and to understand how they might relate to resource and energy use minimization on a global basis. There have been a number of suggestions that materials could be mined from cities or landfills; the concentrations of metal resources in many waste streams are higher than for virgin resources mined from the earth [13]. Better understanding of the status, composition, and location of wastes, stockpiles, discards, and of the institutional and organizational requirements for access could open a wealth of opportunities for recycling and reclamation.

ENVIRONMENTAL IMPACTS OF CONSUMPTION

Material and energy use will change over time. Since the 1970s, a growing body of research has suggested that greater material efficiency, use of improved materials, and growth of the service economy are contributing to the "dematerialization" of the

economy. A related body of research suggests that expensive, scarce, or environmentally harmful resources may be able to be replaced by resources that are cheap, abundant, and environmentally benign. "Substitution" can be seen in the changes in energy sources that have occurred over the past century. As those sources have shifted from wood and coal to petroleum and natural gas, the average amount of carbon per unit energy produced has fallen, resulting in a "decarbonization" of world energy use [14].

There have been many attempts to alter consumption patterns to reduce environmental impacts. But there is no overall understanding of how far these efforts might go and whether there are fundamental limits or obstacles to the potential to change consumption patterns to reduce environmental impacts [15]. Beyond the short-term changes that might be induced by innovative policy approaches, there is a need for understanding the longer-term drivers of change in consumption patterns [16].

One approach to understanding the potential for changes in materials use is to analyze the metabolisms of cities. Cities are known to be strong attractors of resources and weak dispersers of them. Establishing the metabolic differences between cities in different parts of the world of different sizes, and in different stages of evolution could provide better understanding of the diversity in material use patterns.

There is need for better understanding of "rebound" effects, in which increases in production swamp the product-level environmental gains. Ultimately, the challenge is to understand the extent to which consumption can be decoupled from environmental impacts while raising the quality of life. Integration of economic and physical analyses could provide a comprehensive framework for understanding the relationship of environmental, economic, and technological change.

Resource constraints are complex and interdependent: one resource may be substituted for another, and resource availability depends on the technology of resource extraction [17]. There is a need for more rigorous understanding of the overall potential and

limitations of resource substitution, and of the potential and limitations for market forces and technology to overcome physical depletion of particular resources.

Claims and counterclaims about the need to constrain the consumption of materials and energy have characterized the most heated environmental policy debates. At one pole, it has been suggested that consumption of materials and energy in developed countries needs to be significantly reduced in order to protect and maintain environmental systems. At the other pole, it has been suggested that technological innovation will allow us to reduce environmental impacts without reducing economic growth or use of materials and energy. In some cases, technological innovation can surely obviate the need for constraints on consumption. In other cases, constraints on consumption (through, for example, energy taxes or restrictions on certain chemicals) can in themselves drive innovation. The development of a theoretical foundation for understanding the relative roles of technological innovation and consumption constraints in meeting environmental goals would provide a stronger basis for the long-term development of effective environmental policy [18]. At a broader level, the aim is to understand how the environmental impacts of industrialized systems, including energy and transportation, consumer products, cities, and food production, can be understood, managed, and reduced over the long term.

REFERENCES

1. Hoffert MI, Caldeira K, Benford G. Advanced technology paths to global climate stability: energy for a greenhouse planet. *Science* 2002;298:981–987.
2. Neto JQF, Bloemhof-Ruwaard JM, van Nunen JAEE, *et al.*. Designing and evaluating sustainable logistics networks. *Int J Prod Econ* 2007;111(2):195–208.
3. Guide VDR. Production planning and control for remanufacturing: industry practice and research needs. *J Oper Manage* 2000;18(4):467–483.

4. Anastas PT, Zimmerman JB. Design through the twelve principles of green engineering. *Environ Sci Technol* 2003;37(5):94A–101A.
5. Morrow WR, Griffin WM, Matthews HS. National-level infrastructure and economic effects of switchgrass co-firing with coal in existing power plants for carbon mitigation. *Environ Sci Technol* 2008;24(10):3501–3507.
6. Hertwich EG, Pease W, McKone TE. Evaluating toxic impact assessment methods: what works best? *Environ Sci Technol* 1998;32(2):138A–144A.
7. Kaiser J. How much are human lives and health worth? *Science* 2003;299(5614):1836–1837.
8. Ehrenfeld J. Sustainability by design: a subversive strategy for transforming our consumer culture. New Haven (CT): Yale University Press; 2008.
9. Stern PC. Changing behavior in households and communities: what have we learned? In: Dietz T, Stern PC, editors. *New tools for environmental protection: education, information, and voluntary measures*. Washington (DC): National Academy Press; 2002.
10. Goulder LH. Environmental policy making in a second-best setting. In: Stavins RN, editor. *Economics of the environment: selected readings*. 4th ed. New York: W. W. Norton; 2000.
11. Zacarias-Farah A, Geyer-Allely E. Household consumption patterns in OECD countries: trends and figures. *J Cleaner Prod* 2003;11:819–927.
12. Chertow MR. Uncovering industrial symbiosis. *J Ind Ecol* 2007;11(1):11–30.
13. Allen DT, Behmanesh N. Waste as raw materials. In: Allenby BR, Richards DJ, editors. *The greening of industrial ecosystems*. Washington (DC): National Academy Press; 1994.
14. Nakicenovic N. The liberation of the environment. *Daedalus J Am Acad Arts Sci* 1996;125(3):1–17.
15. von Weizsacker EU, Lovins AB, Lovins LH. *Factor four: doubling wealth, halving resource use*. London: Earthscan Publications; 1997.
16. Hertwich EG. Consumption and industrial ecology. *J Ind Ecol* 2005;9(1/2):1–6.
17. Tilton JE. On borrowed time? Assessing the threat of mineral depletion. Washington (DC): *Resources for the Future*; 2003.
18. Grubler A. *Technology and global change*. Cambridge: Cambridge University Press; 1998.

ELECTRONIC NEGOTIATION SYSTEMS

RUDOLF VETSCHERA
University of Vienna, Vienna,
Austria

SABINE T. KOESZEGI
Vienna University of Technology,
Vienna, Austria

MAREIKE SCHOOP
University of Hohenheim,
Stuttgart, Germany

Negotiations are a particular form of communication and collective decision-making. Like all forms of collective action, negotiations involve different parties, which can be individuals, groups, or organizations. Three characteristics distinguish negotiations from other forms of collective decision-making as group decisions or social choice (voting) procedures:

- There is some conflict of interest between the parties involved.
- An agreement can be reached only by consent of all parties. If parties are not able to reach consensus, one particular outcome (usually the status quo) will take place.
- An agreement is not reached immediately but involves an interactive process in which offers are exchanged until a mutually acceptable solution is found.

Research on electronic systems to support negotiations began in the 1980s as an extension to the concept of decision support systems. Early research on negotiation support systems (NSS), for example, Ref. [1], focused on the multiperson aspect of negotiations and the possibilities of information technology to connect parties across time and space. The latter aspect gained importance with the rise of the internet and

the development of web-based NSS [2]. Nowadays, we can distinguish between communication-, decision-, and document-oriented approaches [3]. Surprisingly, e-mail is often used for business negotiations although it does not offer decision support or advanced communication support [4].

Connecting parties is only one function of NSS. As NSS were often seen as an extension to individual DSS, the potential of NSS to improve decision-making is an important concern. Lim and Benbasat [5] distinguished two main components in the architecture of an NSS, a decision support component and a communication support component, and postulated that these two components would have different impacts on negotiations.

The concept of decision support in negotiations draws attention to another important difference between individual and collective decision problems: in individual decision support, a support system assists a decision-maker to better achieve his or her objectives. Decision support in negotiations can have two different goals: supporting one party to better achieve its own goals vis-à-vis others or to support all parties to achieve higher level goals such as efficiency or fairness. In the latter case, the systems take the role of a neutral third party, and in the former case, they serve as advisors to one party. Modern negotiation theory, in particular negotiation analysis developed by Raiffa [6] and Sebenius [7], emphasizes the fact that these two goals are not necessarily in conflict and that finding win-win solutions that benefit all parties is important. However, most negotiations contain distributive elements in which the interests of parties are in conflict. Therefore, already early concepts of NSS [8] distinguished between support components for individual parties and decision support for a mediator. As negotiations by definition require consensus to reach an agreement, supporting parties to find a mutually acceptable solution, and breaking deadlocks in negotiations, is also an important goal of negotiation support at the collective level.

This decision support is also different from decision support for groups. Negotiations are more competitive than group decisions; support in group decisions aims at guiding group members to a consensus decision, which considers all (or a majority of) opinions present in the group. Thus, feedback and aggregation of preferences are a key concern. In negotiations, there are no “aggregated group preferences”; an agreement needs to satisfy individual interests to the extent that all parties can accept it.

Communication in negotiation follows a double goal [9]. On the one hand, the overall goal is mutual understanding. Without that, no interaction is possible. Thus, the negotiators aim to clarify their statements and to prepare a common ground for their interactions. On the other hand, the negotiators want to achieve their individual goals. To this end, they try to convince the partner to accept the current offer using information, threats, compliments, promises, and so on.

NSS are based on a prescriptive–descriptive perspective. Negotiators aim to make rational choices themselves, for which they need support. They also need to be aware of the fact that the other parties might not always act rationally but could be influenced by decision biases. Thus, game theoretic models that assume perfectly rational opponents are only of limited use for negotiation support. It should also be kept in mind that rational behavior in negotiations does not imply a focus on economic outcomes but has to consider that outcomes of negotiations involve several dimensions.

The process dimension is another characteristic of negotiations. NSS can intervene in different ways into these processes. Vetschera [10] distinguished three levels of process support: (i) *facilitation*, which does not guide the process into a particular direction; (ii) *interactive guidance*, which implements a specific interaction structure; and (iii) *normative guidance*, which provides solutions fulfilling normatively desirable criteria such as efficiency or fairness. Kersten and Lai [11] further differentiated NSS into (i) *passive systems*, (ii) *active facilitation–mediation systems*, and (iii) *proactive intervention–mediation systems*.

Passive systems only facilitate communication of geographically distributed negotiators and do not intervene into their behavior. Active facilitation–mediation systems provide advice (e.g., on offers a negotiator could make) only at the user’s request, whereas proactive systems monitor the process and interfere without explicit requests.

Support of negotiation processes serves different goals, individual goals of the parties, and collective goals characterizing the success of the entire negotiation. Collective goals can be of different nature. One obvious goal is to reach an agreement. If an agreement is reached, its quality can be evaluated on several dimensions. On the one hand, the agreement has to specify solutions for all issues of the negotiation, and these solutions can be evaluated in terms of efficiency and fairness. On the other hand, concluding an agreement is often not the end of the interaction between parties, the agreement has to be implemented via joint activities. The subjective evaluation and satisfaction of parties with the agreement, the behavior of their opponent(s), and the bargaining process have a considerable impact on their future relationship and their willingness to actually implement the agreement once they have left the bargaining table.

These different goals require different types of interventions. Reaching an agreement can be supported by improving the quality of communication between parties; thus, communication support (which can go far beyond providing communication channels) is an important topic in NSS. Methods from individual decision support can be applied and extended to enable negotiators to reach efficient and fair outcomes, whereas behavioral interventions can help parties to overcome deadlocks and improve their relationship in order to better cooperate in the implementation of an agreement.

COMMUNICATION SUPPORT IN NEGOTIATIONS

As mentioned earlier, the first perspective on negotiation support has been a decision-oriented one. However, communication is the

essence of negotiation. Without communication, no negotiation is possible as it is impossible not to communicate [12].

To understand an utterance, its contents and its mode need to be understood. According to the communication theories of Searle [13] and Habermas [14], the propositional content represents the factual level of an utterance, that is, what it is about. The illocutionary force represents the mode or the speaker's intention, that is, how it is to be interpreted, for example, as a promise, as a request, as a statement, or as an apology. Taken together, propositional content and illocutionary force provide the meaning of the utterance.

Understanding needs to take place on all levels of semiotics, that is, the syntactic, the semantic, and the pragmatic levels. On the syntactic level, the messages need to be transferred correctly; on the semantic level, their meaning must be unambiguous; and on the pragmatic level, the sender's intention must be explicit.

Tutzauer [15] argues that we can distinguish between offer communication and non-offer communication. The former consists of all the (mostly quantitative) details of the offers, whereas the latter complements offer communication and is equally important for negotiations. Non-offer communication has a factual function that aims at creating a mutual understanding about facts and positions. Furthermore, it has a procedural function to achieve mutual understanding about negotiation protocols, procedures, correlations between issues, and so on. Finally, its relational function focuses on achieving mutual understanding about the underlying norms, values, rules, and guidelines.

The duality of cooperation and competition is a distinctive characteristic of negotiation communication [9, 16]. Although understanding is a constant goal in negotiation, misunderstandings can be used strategically. To remain vague about a conflict issue can help to test out possibilities. Negotiation communication follows structured approaches that are efficient. However, there are also strategies such as time pressure that explicitly delay responses to the last moment in the hope of achieving better deals.

Finally, no negotiation will succeed without compromises and cooperation by all partners. However, negotiations can also be conducted in a situation where power is unevenly distributed, limiting the freedom of one party.

The medium in electronic negotiations has an effect on the communication processes. It can lead to a de-escalation of conflicts because of the asynchronous manner of interaction, thus generating more trust. As the utterances are made in writing, every utterance is documented and thus traceable, again leading to more trust.

Communication support of electronic negotiations is a complex topic. The Negoisst system [3, 17] is the only NSS offering sophisticated communication support. It is based on the theories of Searle and Habermas and implements a language-action approach [18]. On the syntactic level, a negotiation protocol guides the interaction. No deletion of messages is allowed so that every message is documented and traceable. On the semantic level, Negoisst employs a semistructured communication approach. Every part of the natural text message can be linked to a structured vocabulary so that the meaning (i.e., the semantics) of text is explicated and shared between the negotiators. On the pragmatic level, the mode of utterance (i.e., the illocutionary force) is made explicit using message types. Negotiators are thus supported in transparent and open exchanges that would lead to higher communication quality [9].

Compared to simple e-mail exchanges, such communication-oriented NSS approaches provide the benefit of combining structure and flexibility. The communication support allows for natural language messages and thus a high level of flexibility. At the same time, the controlled vocabulary, the messages types, and the interaction protocol provide the structure necessary to allow for traceability, explicit deductions of commitments, and easy search for (parts of) utterances.

DECISION SUPPORT IN NEGOTIATIONS

Decision-oriented tools in NSS focus on the economic dimension of negotiations and are

based on concepts from economics, decision analysis, and game theory. From an economic perspective, efficiency or *Pareto optimality* of the agreement is the main goal. An agreement is called *efficient* if no other possible agreement exists, in which at least one party would better off, and no party worse off. If two parties negotiate about one single issue (such as buyer and seller about price), then all possible outcomes are Pareto optimal, because one party can only gain if the other party makes a concession. In negotiations about multiple issues, possibilities for Pareto improvements do exist. If each party makes a concession in an issue that is of less importance to itself and obtains a concession in a more important issue, both parties can benefit. Decision support in negotiations is therefore closely linked to multiattribute decision-making.

Efficiency is not the only dimension along which outcomes of negotiations can be evaluated. Assuming that negotiators evaluate outcomes according to a given utility function, it is also possible to measure the fairness of outcomes via the (im)balance of utilities it provides to parties or by comparing it to normative solutions such as the Nash bargaining solution [19]. However, most existing NSS do not provide these calculations.

During a negotiation, offers and counteroffers are exchanged. After receiving an offer, a negotiator has to decide among three alternatives:

1. accept the offer;
2. reject the offer and terminate the negotiation;
3. reject the offer and make a counteroffer.

Decision support in negotiations thus should support negotiators in evaluating offers from the opponent, as well as candidates for own offers, and in the preparation of offers. Therefore, a multiattribute preference model is the basis of the decision support component of an NSS. Most of the widely used NSS such as Inspire [2], Negoisst [17], or SmartSettle [20] use an additive value function for this purpose, but there are also systems employing other preference models such as outranking relations [21].

For a comprehensive review of these different methods, see, for example, Refs 22 and 23. Most systems elicit a negotiator's preferences in a preparation phase before the actual negotiation and then employ the calibrated model to evaluate offers during the negotiation process.

For a negotiator, it is not only important to know the value of the most recent offer but also the history of the entire negotiation. An offer often implies a concession, and many negotiators tend to follow the principle of reciprocity and make concessions of a similar size as they received from the opponent. Therefore, they need to keep track of the opponent's as well as their own previous offers and concessions. Many NSS support this behavior by providing a graphical representation of the negotiation process, in which utility values of offers from both parties are plotted over time. Although the system has information about the utility functions of both parties, most systems only refer to a negotiator's own utility, because preferences are considered private information.

Economic theory and utility-based decision analysis usually assume that preferences are stable over time. This assumption is also required for the approach outlined earlier. However, several problems can arise in actual negotiations. Negotiators might change their preferences in view of arguments made by the other party. In this case, the utility function needs to be elicited again. Furthermore, offers might be incomplete and not specify values for all issues. Some systems (e.g., Negoisst [17]) therefore provide lower and upper bounds for utilities, which allow for the fact that issues not specified in the offer might take on their worst or best values. Finally, negotiators might be insecure about their preferences or reluctant to specify exact values. To overcome this problem, decision models under incomplete information can be incorporated in NSS [24].

Providing utility values of offers and counteroffers enables negotiators to solve the first decision problem outlined earlier: an offer made by the opponent should be accepted if it provides a higher utility than the counteroffer the negotiator is ready to make. Similarly, to decide about terminating

a negotiation, offers are compared to the disagreement outcome. However, calculating utility values provides only limited support for the preparation of a counteroffer. By reviewing the previous concession patterns, a negotiator might determine a target utility level for concessions. Some NSS [25] provide models to determine offers that lead to the desired utility levels. A more comprehensive approach [26] calculates a target utility level using desirable properties of the bargaining process such as concession making, reciprocity, and value creation. Multi-issue offers leading to this target utility level are selected based on their similarity to previous offers from the focal negotiator and the opponent. Given the complexity of these problems, the use of methods from artificial intelligence was also proposed for supporting mediators [27] and to advise individual negotiators via intelligent agents [25, 28].

Decision support methods in NSS deal with the exchange of offers as the substantive dimension of the negotiation process. However, they do not consider the relationship dimension of the negotiation process and the outcome dimensions related to it.

BEHAVIORAL SUPPORT IN NEGOTIATIONS

The interaction of negotiators also affects and determines their relationship to each other. Negotiations therefore can also be framed as a process of establishing, defining, or redefining the relationship between parties. A mere focus on the substantive dimension on negotiations would neglect the “human” (behavioral) dimension of these encounters.

Raiffa [6] has already pointed out the synergy of both, in what he called the art and science of negotiations. While decision analytic approaches constitute the science part, the art of negotiation comprises socioemotional skills, the knowledge about various bargaining and negotiation tactics, and how and when to employ them. This aspect is also closely related to the descriptive side of negotiation analysis.

There are two basic orientations toward conflict resolution. (i) Negotiators who have a *cooperative orientation* view negotiations as a process of joint problem solving. They

believe that a conflict can best be resolved when information about the problem and about one's own needs and motivations are shared. Furthermore, a cooperative orientation includes the willingness to understand the counterpart's motivations and needs and the effort to achieve the objectives of all parties. (ii) Negotiators having a *distributive orientation* view negotiations as a process in which one side wins and the other side loses. They attempt to achieve their own objectives at any cost and take everything that is possible from the negotiation table.

These orientations are reflected in a negotiator's strategy and tactics. The negotiation strategy determines the overall set of activities a negotiator implements, and tactics are the elements of this strategy. Integrative behaviors include information sharing, establishing a positive relationship and trust, concession making, and the like. Distributive behaviors include tactics such as making commitments, using power-related tactics, time pressure, and similar behaviors.

Empirical research has shown that negotiators with competitive orientation reach better own outcomes but at the same time reduce the prospects of reaching an agreement [29]. Integrative behavior such as cooperation and joint problem solving facilitates mutually beneficial solutions and helps the participants to “create joint value” [6] and to “invent options for mutual gain” [30]. However, individual outcomes tend to be lower [31]. It therefore depends on the situation, the conflict, and the dynamics of the process, which approach, strategies and tactics, is best suited to reach the negotiators' objectives.

Objectives are not always tangible and go beyond the substantive level. In some situations, the parties are forced to jointly resolve their conflicting interests. For example, in labor negotiations, law could impose deadlines for settlements. In such cases, a (well-) functioning relationship between the parties is an important outcome per se. However, even when parties have outside options, establishing a positive relationship during negotiations creates a solid basis to implement the agreement [32].

Trust is an important aspect of relationship building in negotiations. There is ample empirical evidence that a trusting relationship between negotiators has positive effects on negotiation processes and outcomes: mutual trust facilitates a problem-solving approach [33], the use of integrative strategies [30], and information sharing [34].

Behavioral negotiation support aims at providing negotiators with the knowledge when to use which strategy and how to build up a trusting relationship. However, providing behavioral advice is challenging, as ideas of “good” negotiation behaviors are disputed. Experienced negotiators have accumulated narrative knowledge, that is, “wisdom” about the art of negotiation, which is difficult to generalize and formalize. As there is no general criterion to guide behavioral support, the challenge is to provide guidance and orientation toward “good” agreements. Therefore, behavioral negotiation support has taken a process perspective and focuses primarily on increasing the prospects of an agreement [35]. Nowadays, only a few systems provide behavioral negotiation support. The first system, negotiator assistant (NA), was designed in the late 1990s as an electronic mediation support system to be used either by a human mediator or as an e-mediator in face-to-face negotiations [26]. Its successor, VienNA [36], was developed as a behavioral advice tool in e-negotiations [37]. Both systems are process focused and designed to direct the behavior of negotiators toward relationship and trust building, flexibility, and fairness [35]. The systems provide their users on request with diagnosis and analysis as well as appropriate advice.

The diagnosis function consists of a forced choice questionnaire to analyze the progress and actual status of the negotiation. The tool assesses how parties evaluate the status quo of the negotiation in five categories: parties, issues, activities, situations, and processes. Some questions prompt branching into new sets of questions. For instance, depicting the process as bargaining (win–lose) or problem-solving (win–win) approach leads to different questions addressing the tactics and strategies used. When all questions have been answered, the program generates

a diagnostic grid, which depicts “flexibility vectors” of the parties. This grid allows to forecast the expected outcome of the negotiation. The dimensions of the flexibility vector, which are reflected in the questions, stem from research on factors that influence negotiating behavior, whereas weights have been derived from effect size calculations in a meta-analysis of bargaining experiments. In addition, the model was cross-validated with nine historical negotiation cases (for more details, see Ref. 38).

The analysis and advice functions search for the sources of impasses and stalemates and provide tailor-made recommendations for each user. The system utilizes an expert database on negotiations and identifies areas of difficulties using the information supplied in the diagnosis phase. The most problematic areas are presented to each user individually. For instance, the system might identify that negotiators follow a win–lose bargaining approach and are not flexible enough to trade-off issues. In this case, the system would suggest a logrolling strategy, in which negotiators trade-off less important for more important issues.

OUTLOOK AND CONCLUSIONS

The different approaches to negotiation support outlined earlier have reached different levels of maturity. Nevertheless, potential for future developments exists in all of them. Decision support is probably the approach with the longest research tradition. Still, several issues remain open in this field, in particular the development of more active support methods to suggest offers. Analytical decision support in its present form operates from a strictly prescriptive perspective, providing users a “correct” solution. Another challenge could be to tailor support strategies more specifically to a user’s individual needs to counteract specific decision biases a user might have (or, in a one-sided approach, to exploit biases of the opponent).

Communication support is more difficult to realize as it deals with nonquantifiable aspects of negotiation. One challenge is to realize a proactive communication support, actively supporting the user in his or her

choice of communicative strategies, in preventing misunderstandings, and communication conflicts.

Behavioral support is a more recent concept, which offers many opportunities for research. One particular aspect of negotiations that so far has been rarely addressed in this context is the role of emotions. Emotions are at the core of conflict interactions, they might hinder negotiators to behave rationally [39]. Emotions emerge from the individual's evaluation of events, are either positive or negative, and predispose individuals to bivalent behavior, that is, either approach or withdrawal [40]. Because of emotional reciprocity and contagion, the expression of emotions leads to a nonconscious and automatic mimicking of counterparts. Emotional stimuli may cause distinct reactions by individuals and create complex patterns in human interactions [41]. Empirical research in negotiations shows that negative emotions increase competitiveness, whereas positive emotions increase cooperation [42]. However, negative affect can also be positive in negotiations: expressions of pain or distress may induce cooperative behavior, and anger and fear may motivate to overcome a crisis [43]. Similarly, positive emotion may negatively affect negotiations by causing biased judgments or expectations. The analysis and management of emotions in negotiation processes has not yet found appropriate consideration in behavioral decision support. It would be necessary to monitor negotiation processes to identify problematic emotional patterns in order to intervene into the process in due time.

The three forms of support we have described earlier address different aspects of the negotiation process. However, their integration and consistency across dimensions is an important issue, which is not yet addressed in existing NSS. Existing systems typically do not provide support in all three dimensions. Integrated support could not only help negotiators to overcome a wider range of obstacles but also address the question of consistency across dimensions. Consistent behavior (e.g., open and cooperative communication about preferences and large concessions) will transmit a coherent

message to the other party and thus could improve the possibility of success. However, inconsistencies could also be used as a deliberate negotiation device. Identifying such (in-) consistencies, both by the supported negotiator and by the opponent, and providing advice how to deal with them, could be an important function of future integrated NSS.

REFERENCES

1. Jarke M. Knowledge sharing and negotiation support in multiperson decision support systems. *Decis Support Syst* 1986;2(1):93–102.
2. Kersten GE, Noronha SJ. WWW-based negotiation support: design, implementation, and use. *Decis Support Syst* 1999;25(2):135–154.
3. Schoop M. Support of complex electronic negotiations. In: Kilgour DM, Eden C, editors. *Handbook of Group Decision and Negotiation*. Dordrecht: Springer; 2010. p 409–423.
4. Schoop M, Köhne F, Staskiewicz D, *et al.* The antecedents of renegotiations in practice - an exploratory analysis. *Group Decis Negot* 2008;17(2):127–139.
5. Lim L-H, Benbasat I. A theoretical perspective of negotiation support systems. *J Manage Inform Syst* 1992;9(3):27–44.
6. Raiffa H. *The art and science of negotiation*. Cambridge (MA): Belknap; 1982.
7. Sebenius JK. Negotiation analysis: a characterization and review. *Manage Sci* 1992;38(1):18–38.
8. Jarke M, Jelassi MT, Shakun MF. *MEDIA-TOR: towards a negotiation support system*. *Eur J Oper Res* 1987;31(3):314–334.
9. Schoop M, Köhne F, Ostertag K. Communication quality in business negotiations. *Group Decis Negot* 2008;19(2):193–209.
10. Vetschera R. Group decision and negotiation support - a methodological survey. *OR Spektrum* 1990;12(2):67–77.
11. Kersten GE, Lai H. Negotiation support and e-negotiation systems: an overview Group. *Decision and Negotiation* 2007;16(6):553–586.
12. Watzlawick P, Beavin JH, Jackson DD. *Pragmatics of human communication*. New York: W. W. Norton; 1967.
13. Searle JR. *Speech acts—an essay in the philosophy of language*. Cambridge: Cambridge University Press; 1969.
14. Habermas J. *Theorie des kommunikativen Handelns*. Frankfurt: Suhrkamp Verlag; 1981.

15. Tutzauer F. The communication of offers in dyadic bargaining. In: Putnam L, Roloff M, editors. *Communication and negotiation*. Newbury Park (CA): Sage; 1992. p 67–82.
16. Duckek K. *Ökonomische Relevanz von Kommunikationsqualität in elektronischen Verhandlungen*. Wiesbaden: Gabler; 2010.
17. Schoop M, Jertila A, List T. Negoisst: a negotiation support system for electronic business-to-business negotiations in e-commerce. *Data Knowl Eng* 2003;47(3):371–401.
18. Schoop M. A language-action approach to electronic negotiations. *Syst Signs Action* 2005;1(1):62–79.
19. Nash JF. The bargaining problem. *Econometrica* 1950;18(2):155–162.
20. Thiessen EH, Loucks DP, Stedinger JW. Computer-assisted negotiations of water resources conflicts. *Group Decis Negot* 1998;7(2):109–129.
21. Wachowicz T. Decision support in software supported negotiations. *J Bus Econ Manage* 2010;11(4):576–597.
22. Belton V, Stewart TJ. *Multiple criteria decision analysis*. Amsterdam: Kluwer; 2001.
23. Pomerol J-C, Barba-Romero S. *Multicriterion decision in management: principles and practice*. Dordrecht: Kluwer; 2000.
24. Sarabando P, Dias LC, Vetschera R. Mediation with incomplete information: approaches to suggest potential agreements. *Group Decis Negot* 2013;22(3):561–597.
25. Chen E, Vahidov R, Kersten GE. Agent-supported negotiations in the e-marketplace. *Int J Electron Bus* 2005;3(1):28–49.
26. Vetschera R, Filzmoser M, Mitterhofer R. An analytical approach to offer generation in concession-based negotiation processes. *Group Decis Negot* 2012. DOI: 10.1007/s10726-012-9329-z.
27. Sycara KP. Negotiation planning: an AI approach. *Eur J Oper Res* 1990;46:216–234.
28. Kersten GE, Lo G. Aspire: an integrated negotiation support system and software agents for e-business negotiation. *Int J Internet Enterp Manage* 2003;1(3):293–315.
29. Pruitt DG. *Negotiation behavior*. New York: Academic Press; 1981.
30. Fisher R, Ury W. *Getting to yes*. Boston (MA): Houghton Mifflin; 1981.
31. Olekalns M, Smith PL. Understanding optimal outcomes: the role of strategy in competitive negotiations. *Human Commun Res* 2000;26(4):527–557.
32. Zand DE. Trust and managerial problem solving. *Adm Sci Q* 1972;17(2):229–239.
33. Butler JK Jr. Behaviors, trust, and goal achievement in a win-win negotiating role play. *Group Organ Manage* 1995;20(4):486–501.
34. Greenhalgh L, Chapman D. Negotiator relationships: construct measurement, and demonstration of their impact on the process and outcomes of negotiation. *Group Decis Negot* 1998;7(6):465–489.
35. Druckman D, Mitterhofer R, Filzmoser M, *et al.* e-mediation and e-negotiation: behavioral decision support or utility maximization? *Group Decis Negot* 2013. DOI: 10.1007/s10726-013-9356-4. Forthcoming.
36. Mitterhofer R, Druckman D, Filzmoser M, *et al.* Integration of behavioral and analytic decision support in electronic negotiations. 45th Annual Hawaii International Conference on System Sciences. Maui Hawaii: IEEE Computer Society Press; 2012.
37. Gettinger H, Dannemann A, Druckman D, *et al.* in Hernández J, Zaraté P, Dargam F, *et al.* Impact of and interaction between behavioral and economic decision support in electronic negotiations. In: Hernández JE, *et al.*, editors. *Collaboration in real environments*. Berlin: Springer; 2012, 151–165 in print.
38. Druckman D, Harris R, Ramberg B. Computer-assisted international negotiation: a tool for research and practice. *Group Decis Negot* 2002;11(3):231–256.
39. Barry B, Oliver RL. Affect in dyadic negotiation: a model and propositions. *Org Behav Human Decis Process* 1996;67(2):127–143.
40. Hatfield E, Cacioppo J, Rapson RL. Emotional contagion. *Curr Dir Psychol Sci* 1993;2(3):96–100.
41. Lerner JS, Keltner D. Beyond valence: toward a model of emotion-specific influences on judgement and choice. *Cognit Emotion* 2000;14(4):473–493.
42. Van Kleef GA, De Dreu CKW, Manstead ASR. The interpersonal effects of emotions in negotiations: a motivated information processing approach. *J Pers Soc Psychol* 2004;87(4):510–528.
43. Friedman RA, Brett C, Brett J, *et al.* The positive and negative effects of anger on dispute resolution: evidence from electronically mediated disputes. *J Appl Psychol* 2004;89(2):369–376.

ELICITING SUBJECTIVE PROBABILITIES FROM INDIVIDUALS AND REDUCING BIASES

BETHANY H. MIHAJLOVITS

JASON R. W. MERRICK

Statistical Sciences and Operations Research,
Virginia Commonwealth University,
Richmond, Virginia

INTRODUCTION

The overall process of arriving at an informed decision usually includes the likelihood of events or the distribution of unknown quantities. Numerous important analyses have relied on expert judgment. North and Stengel [1] used expert judgments to assist the US magnetic fusion energy program in determining the best fusion concepts to fund. Keeney *et al.* [2] examined setting standards for ambient air quality, using expert judgments to develop a risk-assessment model that related adverse health effects to alternative carbon monoxide standards. Winkler and Poses [3] performed assessments with physicians to provide judgments of patient survivability in critical care situations. DeWispelare *et al.* [4] studied the performances of high level nuclear waste geologic repositories over a 10,000-year period with experts assessing changes in the climate over this period. Expert judgments have been used in numerous risk assessment situations [5,6], including numerous maritime risk assessments [7–11].

The use of expert judgment is not without its difficulties. Human beings tend to incorporate conscious and unconscious biases in their predictions of uncertainties. We use “primitive cognitive techniques” to simplify the task of assessing uncertainty, called *heuristics*. Heuristics are easy to use and simplify the task, but result in systematic faults with the resulting judgments, called *biases*. This article focuses on ways to improve the assessments provided. When an expert is asked to

provide a probability of an event’s occurrence, they search their memory for applicable knowledge, combine that knowledge with the information at hand, and attempt to provide the best feasible judgment [12]. Therefore, the judgment provided will depend on what is retrieved from the subject’s memory, what current information is used, and even the order in which the subject thinks about these pieces. Gigerenzer *et al.* [13] point to the literature, suggesting that memory is often “excellent in storing frequency information from various environments and the registering of event occurrences for frequency judgments is a fairly automatic cognitive process requiring very little attention or conscious effort.” This series of cognitive processes that occur without effort in the subject leave much room for error of interpretation of uncertainty and that error can be presented as a biased assessment [14].

The representative heuristic is used when assessing a probability that an object belongs to a specific category. When using representativeness, we evaluate the probability “by the degree to which *A* is representative of *B*, that is, the degree to which *A* resembles *B*” [15]. This heuristic involves making the judgment by comparing the information provided on the subject with a stereotypical member of the category. The more complete the resemblance between the two, the higher is the probability that the subject will be placed into the category. There are several biases that result from the representative heuristic. Kahneman and Tversky [16] describe a representativeness study where students are asked to predict the career area for a fictitious character called Tom W. “Tom W. is of high intelligence, although lacking in true creativity. Tom has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by Hashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and

little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense” [16]. One group of students is asked to predict the proportion of the US population that is employed in each of nine career areas (business administration, computer science, engineering, etc.) to form a base rate. Another group is asked to assess the similarity of Tom W. to people employed in each of these areas by ranking from the most to least similar. A third group is asked to assess the probability that Tom W. is employed in each of these areas. The correlation between the assessed probabilities and the similarity rankings was 0.97, while the correlation between the assessed probabilities and the base rates was -0.65 . Thus, the probability assessments were primarily based on the representativeness of the description for each career area and completely unrelated to the overall base rates of each career in the population.

The availability heuristic can be used when a probability of an event’s occurrence is assessed on the basis of the ease with which one can retrieve similar events from memory [17]. It is often easier to recall instances of large classes than those of less frequent classes. External events and influences can have a substantial impact on the availability of similar incidents, such as the media or particularly emotional events [18]. Kahneman *et al.* [19] suggest that witnessing an accident will have a greater effect on a person’s judgment of the probability of an accident than reading about it in a newspaper. Tversky and Kahneman [17] describe an experiment where subjects were read lists of names of famous men and women and were asked to assess whether there were more men or women on the list. Some subjects were given lists where the men were more famous and they tended to respond that there were more men on the list. Some subjects were given lists where the women were more famous and they tended to respond that there were more women on the list. Thus, the ease of recall affected the judgment. This is also true, for instance, in searching memory and imagining possible events.

An individual will often assess a probability by selecting an initial anchor point

(suggested either by the formulation of the problem or by result of partial computation) and then adjusting the anchor based on his/her knowledge of the specific event [20]. Tversky and Kahneman [15] term this heuristic anchoring and adjusting. The adjustments are usually inadequate since different estimates are obtained when based on different anchors. For instance, if we ask to provide a best estimate of an unknown quantity, such as the median, and then ask for a range, such as the 25th and 75th percentiles, then it is likely that the range will be too small. This bias is called *overconfidence*, as it appears that we are overconfident in our estimates. The best estimate is anchoring our range estimates and we do not adjust sufficiently away from it.

Given the tendency to use such heuristics, we must design both the elicitation question and the elicitation process to assist the experts in providing their best possible assessment. We must also ensure that we are asking for assessments that are sufficiently specific and within the range of the expertise of those asked, a task best achieved with correct decomposition of the problem. In the section titled “What is a Good Probability Judgment?” we discuss the desirable properties of a probability judgment. Elicitation processes are discussed in the section titled “Elicitation Processes for Minimizing Biases.” Decomposition is discussed in the section titled “Elicitation Methods for Minimizing Biases.” In the section titled “Conclusions,” we discuss elicitation methods.

WHAT IS A GOOD PROBABILITY JUDGMENT?

A major question in using probability judgments in decision analysis is the validity of the judgments. In psychometric terms, validity is the relationship of probability judgments to an independent and previously validated scale [12]. Calibration is often used as a measure of validity. A judgment is well calibrated if the events that the expert assigns a probability p to occur $p\%$ of the time. For example, a forecast of a 60% chance of precipitation is well calibrated if precipitation is observed on 60% of the days when

that forecast is made. From a statistical perspective, p must be close to the observed relative frequency of the event, $\hat{p}$. Wallsten and Budescu [12] observe that calibration is not a measure of validity in the strict psychometric sense as p and $\hat{p}$ are not necessarily independent, as they are both based on available data. Keren [21] notes that p is in fact a measure of the expert's belief and therefore does not need to correspond to $\hat{p}$ to be valid, particularly if the expert has not observed the same data as used to estimate $\hat{p}$.

The concept of calibration is related to the overconfidence bias [22]. Specific overconfidence occurs when $p > \hat{p}$ for a given event. General overconfidence occurs when $p > \hat{p}$ for $p > 0.5$ (for Bernoulli events). Most literature refers to general overconfidence, leading to the well-known S-shaped calibration curve and probability judgments that are more extreme than the corresponding relative frequency. Keren [21] also suggests that if the errors in representing the expert's true beliefs are random, then an S-shaped curve can at least partly be explained by regression to the mean (see also Refs 23 and 24). Juslin [25] suggests that the phenomenon could be due to the experimenter's choice of almanac questions, a phenomenon borne out by Murphy and Winkler [26], who find that weather forecasters are well calibrated within their domain expertise. However, Lin and Bier [27] find that most overconfidence is random and not due to the question asked.

Calibration is not sufficient for a good probability judgment. Consider the case of an expert assessing the probability of the sex of a newborn baby. An expert can say 50–50 and be perfectly calibrated as this is the observed relative frequency across all births, but this does not provide any helpful information to the parent (decision maker) [22]. Alternatively, an expert could always get the sex wrong, but with absolute certainty. This expert is perfectly miscalibrated, but useful if you know about the systematic bias and can correct for it. Essentially, an expert that is diagnostic (p is related to $\hat{p}$ in some known way) is all that is truly needed.

Another problem with probability judgments occurs when multiple probability judgments are requested, but the set of judgments are incoherent (do not obey the laws of probability). An example of this is the conjunction fallacy that is caused by the representativeness heuristic. Tversky and Kahneman [28] offered subjects a description of a woman called *Linda* that was representative of a feminist and asked them to rank the following statement:

1. Linda is an active in the feminist movement.
2. Linda is a bank teller.
3. Linda is a bank teller and is active in the feminist movement.

Many subjects ranked the third option above the second, even though this violates the rules of probability.

The calibration, resolution, and coherence of probability judgments can be improved by the way the analyst interacts with the judge or by the analysis of the judgment. Essentially, we can improve the judge or we can improve the judgment. In this article, we concentrate on improving the judgment through the elicitation process, the decomposition of the problem, and the specific elicitation question or method. In the section titled “Conclusions,” we point the reader to other articles on improving the judgment.

ELICITATION PROCESSES FOR MINIMIZING BIASES

The analyst should aim to structure a situation where biases are minimized and the assessments determined actually symbolize the subject's opinions [12]. This section overviews processes presented by Spetzler and Von Holstein [29] and Keeney and Von Winterfeldt [30] that are aimed at eliciting accurate probabilities while minimizing the effects of the biases presented in the previous section.

Spetzler and Von Holstein [29] start with a set of five principles to use as a guideline for encoding uncertain quantities, and violating any one of these principles can lead to

problems in the encoding process. Only those uncertainties that are important to the decision itself are chosen. The quantity is defined as an unambiguous state variable (a state variable being one that has values beyond the control of the decision maker). The quantity is structured carefully with thoughtful consideration to conditionalities. The quantity is defined precisely by performing the clairvoyance test. The quantity is described using an appropriate scale that the subject finds meaningful, so that the subject does not spend time trying to fit his/her answers to the scale versus evaluating the uncertainty. These principles ensure the clarity of the task required of the expert and reduce the cognitive effort by using familiar language, definitions, and scales.

Spetzler and Von Holstein [29] provide an interview process procedure that is divided into five phases. Their process is for eliciting probabilities from a single individual:

1. *Motivating* introduces the subject to the encoding task and explore the possibility of any motivational biases;
2. *Structuring* defines the uncertain quantity to be assessed.
3. *Conditioning* reviews how the subject assesses probabilities so that any potential biases can be avoided.
4. *Encoding* encodes the uncertain quantity.
5. *Verifying* reviews the assessment and tests for subject approval.

Wallsten and Budescu [12] suggest that when encoding an expert's (or nonexpert) opinion, it is imperative that the analyst carefully specifies the class of events in question, the sources of information to be considered, and the potential causes of unreliability in the information (e.g., sampling error). The anchoring and adjusting heuristic can play a key role when/if the analyst accidentally gives undue weight to the earliest information considered. The analyst should review the heuristics with the subject to introduce them to how most of the people process uncertainty and thus attempt to eliminate most of the biases associated with the heuristics.

The suggested route to encode the quantities is to begin by asking for the extreme values before moving to central estimates. This route is suggested to ensure that the biases associated with anchoring and adjusting are eliminated. In the verifying phase, if the subject is not comfortable with the assignments provided, the process may need to be repeated until the subject is confident with his judgments. A typical interview should last 30–90 min, depending on the complexity of the situation. Spetzler and Von Holstein [29] make a final note that the preencoding steps generally take longer than the encoding steps but are the most essential to good probability assignments. In addition, interactive techniques have proven to facilitate judgment by minimizing randomness and eliminate unwanted inconsistencies [12].

Keeney and Von Winterfeldt [30] provide lessons learned from the NUREG 1150 study and a recommended process for eliciting probabilities from experts. The recommended process for eliciting probabilities from multiple experts consists of seven components:

1. *Identification and selection of the issues* emphasizes that, even before the study begins, it is important to recognize all uncertainties that need expert probability elicitation and select the ones that need to go through a formal elicitation procedure.
2. *Identification and selection of the experts* should involve both specialists and analysts. Specialists should be the most knowledgeable in their field of study and the analysts who perform the elicitation should have expertise in probability theory, statistics, and decision analysis.
3. *Discussion and refinement of the issues* is to ensure that the specialists understand the project and what is expected of them and arrive at unambiguous definitions of the events and the uncertainties that will be elicited.
4. *Training for elicitation* is performed by an analyst with an intent to familiarize the specialist with the techniques used in elicitation, provide them with

an assessment practice session, inform them of the potential biases that can occur, and motivate them on their judgment task.

5. *Elicitation* is an interview between the specialist and analyst with a possible generalist assistant who provides project inputs and recording services.
6. *Analysis, aggregation, and resolution of disagreements* is not always feasible to combine the specialist's judgments during the elicitation process, thus the analyses should be provided to the specialist as soon as possible after the interview so that any necessary changes can be made while the interview is still fresh in mind.
7. *Documentation and communication* is performed throughout the entire elicitation procedure. At the very minimum, documentation of elicitation results and reasoning, event definitions, expert identification and selection, the training session, the actual elicitation session, and the aggregation and analysis should be maintained.

The first two components of this process are self-evident, deciding what should be assessed and by whom. The third component, refinement of the issues, should be conducted in a meeting of the specialists and analysts. The in-depth discussion of definitions amongst experts should help eliminate biases associated with the representative heuristic with the elimination of any ambiguity.

After the first meeting is completed, the issues clarified, and the specialists given time to think about the issues at hand, the training for elicitation should be conducted. Wallsten and Budescu [12] determined that the accuracy of the assessments determined by either direct or indirect methods could be increased with just simple and short training procedures. Benson and Onkal [31] demonstrate that training can reduce biases in judgments. Keeney and Von Winterfeldt [30] suggest that the elicitation itself should start with an informal discussion of the specialist's approach to the issue and review of the event definitions in order to

map out the decomposition to identify the necessary judgments. Next, a series of easy questions should be asked to determine the probabilities or probability distributions of the decomposition, followed by harder questions. Consistency checks should be administered once the judged probabilities are obtained. The end points or boundaries should be defined, followed by the fractiles. The determination of end points first helps eliminate any anchoring that might occur on behalf of the specialist. The median judgment should be the first fractile assessed. A graphical depiction should be maintained throughout the process to provide continuous consistency checks and feedback to the experts. Murphy and Winkler [26] show good calibration among weather forecasters who receive continuous feedback on the accuracy of their forecasts, implying that such long-term training can also be beneficial. In experimental testing, González-Vallejo and Bonham [32] demonstrate that providing both positive and negative feedback improves calibration.

The next step in this process is the analysis, aggregation, and resolution of disagreement components. The analyst aggregates across the experts by an averaging process and performs sensitivity analysis. A final group meeting can be held for all specialists and analysts involved to establish any agreements or disagreements on the probabilities elicited or events in question.

Mellers and Locke [33] have additional suggestions for the elicitation procedures to eliminate biases. The first concerns disclosure, a common procedure for constraining self-interest. This suggestion is based on the assumption that when a subject reveals a conflict of interest he or she will provide a less biased assessment. If not, the analyst can discount the assessment to achieve an unbiased decision. Another study, however, suggests that disclosure actually increases the subject's bias, and also increases the analyst's discount, although not enough to correct the initial increase in bias. Mellers and Locke [33] believe that the subjects may feel their self-interests are disclosed and so the decision makers should rely more heavily on their own information without knowing how

much the conflict of interest has affected the provided assessment. In real life, if the decision maker trusts the subject providing the assessment and the decision maker is unsure of his own information, having a provided disclosure may allow the interpretation that the subject's information is more credible.

Another suggestion from Mellers and Locke [33] for the elimination of biases during the elicitation procedure is to apply accountability to the situation, because predecision accountability to an unknown audience can reduce cognitive biases, especially those biases that result from lack of effort of understating one's own judgmental processes. This suggestion can however have a null or detrimental effect if the decision maker hedges his choice by sticking with the status quo. For the last suggestion, Mellers and Locke [33] state that some tasks, such as recall of items and the effects of anchoring and adjusting, showed improved performance with the use of incentives. Incentives can however be harmful if the task is too complicated, and most studies actually showed little effect on performance with the use of incentives.

The elicitation process is just as important as the method used. The process must ensure that all experts are using the same definitions of the events or quantities in question and that the experts understand the biases that can be introduced if they use simple heuristic strategies. Feedback and the chance for revision of initial assessments are also a critical part of any elicitation process.

DECOMPOSITION FOR MINIMIZING BIASES

Moskowitz and Sarin [34] study the assessment of conditional probabilities. They present difficulties associated with causal versus diagnostic relationships between events, as well as positive versus negative relationships. In assessing $p(B|A)$, the relationship is causal if A is perceived as the cause of B and diagnostic if B is perceived as the cause of A . The study found that probability judgments tend to be higher if the relationship is perceived to be causal. This is because individuals perceive a

causal relation as being more informative or representative. In addition, the relationship is positive if the occurrence of A increases the probability of B occurring and negative if it reduces the probability of B occurring. The subjects felt that a positive relation was more informative, so higher probabilities were assessed. Normatively speaking, however, neither the causality/diagnostic relationship nor the positive/negative relationship should influence the conditional probability assignments. These effects are noted in the study reviewed below.

The study involved undergraduate students enrolled in managerial statistics with knowledge of the concept of probability. The subjects assumed the role of a bank lending officer assessing the base rate probabilities of delinquency (D) and good (G) or bad (B) credit. The subjects assessed the conditional probabilities $p(D|B)$, $p(B|D)$, and $p(D|G)$. The subjects first specified their perceptions of the relationship strengths between delinquency and credit rating. They then assessed conditional probabilities for 15 questions partitioned into five groupings and provided their assessments. The subjects were given a joint probability table with the marginal event probabilities included as an assessment aid. They could then reassess several of the previous questions, and finally rate the probability table as an assessment aid on a scale with descriptions as to how it helped.

Each groups questions were designed to test the consistency with respect to the probability axioms that must be satisfied for a conditional probability, say $p(B|A)$: $p(B|A) \leq 1$, $p(B|A) \leq \frac{p(B)}{p(A)}$, $p(B|A) \geq \frac{p(B)+p(A)-1}{p(A)}$, and $p(B|A) \geq 0$. A number of violations were observed. The authors suggested that the subjects ignored the based rate frequency of the events (the base rate bias) and the causal versus diagnostic and positive versus negative relationships. Causal and positive relationships led to more revisions and higher conditional assessments than diagnostic and negative relationships. With the addition of the probability table, the subjects tended to have more confidence in the marginal probability assessments and were more likely to

revise their conditional probability assessments. The results suggested that the use of a probability table drastically improves the consistency, reduces the frequency of the violations of the conditions, and lessens the effects of the causality/diagnostic and positive/negative relationships. However, if the probability table is not used properly, the conditional probabilities can be inconsistent.

Decomposition is the process of breaking down a probability assessment into smaller or manageable chunks. Clemen and Reilly [20] recognize three scenarios where decomposition is appropriate: thinking about how the event in question is related to other events, thinking about what uncertain outcomes could lead to the occurrence of the event in question, and thinking about what uncertain outcomes must happen for the event in question to occur. Further, the best decomposition to use is the one that is easiest to contemplate and that gives the most transparent view of the uncertainty in question. Ravinder *et al.* [35] sought to reduce inconsistency by relying on decomposition rather than direct elicitation. They assessed the impact of the decomposition procedure on the reliability of the assessed probabilities.

Ravinder *et al.* [35] regard decomposition as a useful technique for reducing the complexity of challenging judgment problems. The decomposition approach is able to capture the relationship between the event in question, the target event, and its background events. For example, the survivability of a patient may depend on the disease, the strength of the patient, the complexity of the medical procedure performed, and so on. The distribution of the survivability of the patients could be assessed conditionally on the outcome of the disease they had or even conditioned on multiple events, like the disease the patient had was nonsignificant, the complexity of the surgery was minuscule, and the patient's strength was favorable. The decomposed judgments are later combined using the law of total probability. We denote the probability of the target event as A and assume that it is assessed conditionally upon its background events $B_1, \dots, B_n$. The event B_i can be a single event or a

scenario and the set must form a mutually exclusive and exhaustive partition of the relevant event space ($\sum_{i=1}^n \Pr(B_i) = 1$). The probability of the target event is $\Pr(A) = \sum_{i=1}^n \Pr(A | B_i) \Pr(B_i)$. Evidently, elicitation by decomposition involves more questions than direct assessment as marginal and conditional probabilities must be assessed for each background event.

Ravinder *et al.* [35] draw several conclusions to minimize inconsistency in conditional probability assessments as follows:

- Decomposition can reduce random errors associated with probability encodings.
- Error reduction has an upward limit and assessing additional probabilities will not be necessary to obtain further reduction.
- The marginal probabilities of the conditioning events should be selected to be equal.
- Conditional probabilities that cover an extremely wide range of values should be avoided.
- The component probability assessment errors should be as small as possible and the analyst should try to maintain independence of the component assessments.
- A decomposition approach is conceptually simple, is implemented easily, and apparently has great potential for error reduction.

Andradóttir and Bier [36] also address the trade-off between simplifying judgment tasks by decomposition and adding additional estimates with ensuing cumulative error. They find that large numbers of conditioning events are not necessary, but it is better to edge on the side of too many and not too few to avoid significant drop-offs in performance. It is also important to ensure that the marginal probabilities of conditioning events can be accurately estimated.

Howard [37] notes that information is generally received fragmentally when it comes from an expert or a group of experts, and the daunting task of gathering and coordinating

these fragments can be lessened with the use of knowledge maps. A knowledge map is centered on the concept of influence diagrams and captures the relevant knowledge of an event. An influence diagram represents the actions that a decision maker may take and the information possessed at the time of the decision by way of decision nodes (boxes representing an action to be taken), chance nodes (circles representing an uncertain event), and arrows (showing connectivity between the events). A knowledge map is also a method for decomposing a complex, uncertain event into workable and more easily understood units.

Howard [37] uses the [17] study of Tom W. to demonstrate how knowledge maps can aid in correcting the representative heuristic. This study presents a high school description of Tom as introverted, mechanically minded, lacking in creativity, and well organized to a group of students who are supposed to assess the likelihood that Tom would be enrolled in a specific field of graduate study. Generally, the assessors relate Tom to being a “nerd” and subsequently place him into the field of computer science or engineering without regard to the relative enrollment populations of each field of study. Howard [37] first suggests decomposing this problem. One could assess the likelihood that a student would study each field and the likelihood of a report such as Tom’s given he is in each field of study. Next, the assessors can multiply the two likelihoods and normalize to yield a posterior probability that Tom works in each field.

This first step improves understanding but does not capture everything relevant to the second probability, that a report like Tom’s would describe someone in a given field. Howard believes that additional relevant information can be captured with three considerations:

- Is Tom’s personality now different from that in high school (when the report was completed)?
- Are there valid issues with the report itself?
- Is personality a good indicator of field of study?

Figure 1 shows the corresponding knowledge map capturing these considerations.

Despite the fact that this knowledge map encompasses all the necessary information for determining the likelihood, it does not reflect the notion that an individual might not feel comfortable trying to assess the relevance of the present personality to graduate field because they do not think about characterizing the personalities of graduate students. Yet, assessing the probability distribution on field of study and what is known about the personalities of graduate students in each field is probably more comfortable for most assessors. The redundancy map in Fig. 2 represents this way of thinking. The assessments for the original knowledge map (Fig. 1) can now be determined by using the new assessments from Fig. 2 iteratively until the personality now distributions agree and then performing the Bayesian reversal of the arrow from “graduate field” to “personality now.”

“A knowledge map represents a possible assessment order for the joint distribution on the uncertain quantities (events) it contains.” In some instances, the decision maker is only interested in the joint distribution of a few events [37]. For these instances, constructing a disjoint knowledge map can simplify the assessment. The corresponding example provided by Howard [37] is assessing a probability distribution on the income of the next person to enter the room. To aid in thinking about this particular assessment, the subject might consider the amount of education and the age of the individual. With these additional events comes additional assessments and the subject is only interested in the distribution on income. Thus, the subject would want to draw the corresponding disjoint knowledge map represented in Fig. 3. The subject can now assess a distribution on age and another on education given age. The law of total probability can then provide a distribution for education. As Howard suggests, creating disjoint knowledge maps can reduce the difficult task of assessment to manageable pieces. One problem with disjoint knowledge maps, however, is that they only represent one point in the spectrum and they do

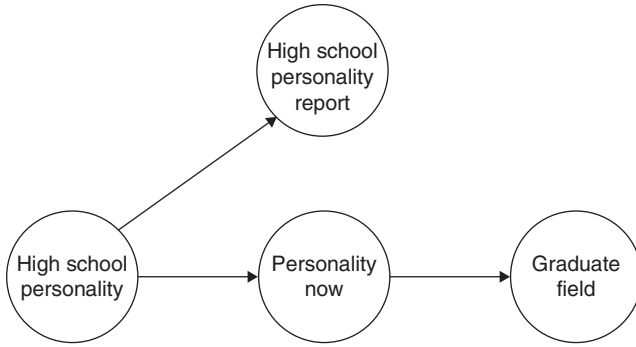


Figure 1. A basic knowledge map for Tom W.

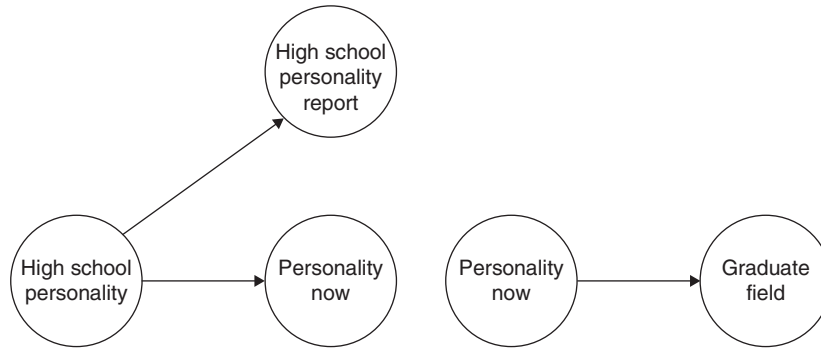


Figure 2. A redundant knowledge map for Tom W.

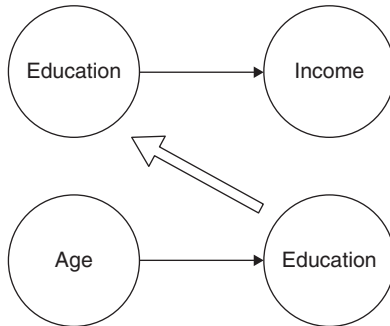


Figure 3. A redundant knowledge map for assessing income.

not provide all the consistency checks and inferences of a complete knowledge map. Yet, overall, knowledge maps allow indirect assessment and the use of redundant and disjoint knowledge maps “can often reconcile the needs of consistency and convenience” [37].

The literature reveals that the events and quantities in question must be suitably

decomposed before elicitation. Conditionalities must also be fully understood by the experts during elicitation. Decomposition can improve probability judgments, but can also introduce inconsistencies. However, these inconsistencies can be reduced. If event A causes event B , then we should elicit $p(B | A)$ rather than $p(A | B)$. Further, if $p(B | A) < p(B)$, then we can elicit $p(\text{not } B | A)$ to ensure a positive relationship. Knowledge maps can also help reduce inconsistencies. Using knowledge maps, a few conditional relationships are elicited at a time and then the expert is informed of inconsistencies and asked to iterate their assessments until coherent conditional probabilities are obtained.

ELICITATION METHODS FOR MINIMIZING BIASES

Spetzler and Von Holstein [29] review methods for eliciting probabilities. Questions

can be answered either directly (by providing numbers) or indirectly (by choosing between alternatives until an indifference point is determined). Direct assessment, or structured questioning, supports data generation with questioning schemes that are designed to trigger elements in the subject's associative memory [38]. There are multiple indirect response techniques to choose from such as odds-based lotteries and reference lotteries. Verbal encoding is the most common direct method studied in the literature, where verbal descriptions are used to characterize events and then quantitative interpretations are added. Benson *et al.* [38] suggest that graphical representation is one of the most used and best-developed aids for assessment, with decision trees being the dominant form of problem representation. However, graphic representation, like other existing aids, has a limiting account of relevance and therefore has a tendency to constrain rather than exploit the variety of reasoning strategies that humans are capable of [38]. The indirect response mode is generally superior to begin encoding, because subjects tend to have trouble in giving direct numerical probability [29]. Additionally, indirect assessment avoids the availability heuristic because subjects do not assess probabilities directly based on memory. We highlight two indirect methods, an odds-based lottery (or betting) technique and a reference lottery (or probability wheel) technique, and one direct method, verbal encoding.

Using the lottery technique, we consider bets for and against the occurrence of an event. The expert is given odds $X : Y$, which means that they lose Y if the event does not occur and win X if it does. The expert is asked if he/she would take those odds on the event occurring or if they would prefer those odds for the event not occurring. If they prefer to bet on the event occurring then $p(\text{Event}) > \frac{Y}{X+Y}$. If they prefer to bet on the event not occurring then $p(\text{Event}) < \frac{Y}{X+Y}$. The elicitor must start with two bets that bound the probability and then reduce the bounds to arrive at a final point of indifference. At indifference, $p(\text{Event}) = \frac{Y}{X+Y}$. This bounding approach avoids the anchoring heuristic if the initial high and low values are extreme

enough to bound the probability. The expert could still employ either the representativeness heuristic or the availability heuristic depending on the description of the event that is considered. However, the motivation of betting is used to engender additional effort in the thought process used to determine the expert's answer. This can avoid the use of simple thought processes.

Spetzler and Von Holstein [29] call the odds-based lottery a value method because the payoffs are varied and the probabilities are kept constant. The odds-based lottery technique works well with experts who are familiar with betting and for events that they are willing to place bets for. Weather forecasters familiar with betting might be willing to place bets on whether it will rain tomorrow, but oil tanker captains would probably not be willing to bet on a large oil spill occurring in the next year. The technique also assumes that the expert is risk neutral over the range of bets considered [39]. If the odds considered are too large, particularly on the loss side, then the expert's risk preferences can bias the results.

A reference lottery asks an expert whether he/she would like to bet on the event occurring or on another lottery with known probability (the reference lottery). A probability wheel is commonly used as the reference lottery to help the expert visualize the probability. However, red and blue balls drawn from a bag (or urn) and drawing cards from a deck are other examples that may work depending on the expert's familiarity with the specific visualization. The prizes for the event lottery and the reference lottery are fixed, so the same prize is awarded if the expert "wins" or "loses" either lottery. The expert must prefer the prize awarded on a "win." A high probability and a low probability reference lottery are offered first to create upper and lower bounds on the probability of the event, like with the betting approach, and then the bounds are narrowed until the expert reaches indifference between the event lottery and the reference lottery. At indifference, the probability of the event is equal to the probability of a "win" on the reference lottery. The prizes are fixed, so the expert's risk preferences do not bias the

results and framing effects are avoided. However, the technique is still based on betting, so the betting on the event must not be distasteful to the expert. Spetzler and Von Holstein [29] call the reference lottery a probability method because the probabilities of the reference lottery are varied and the payoffs are kept constant.

Fractile or quantile methods are also important when the distribution of an unknown quantity X is needed, not just a simple probability. However, we can elicit fractiles, $p(X \leq x)$, using either the betting or reference lottery method and specifying the event to be $X \leq x$. Fractiles can then be assessed for multiple values of x and the assessed cumulative distribution function plotted to provide feedback to the expert. Abbas *et al.* [40] empirically evaluate fixed probability methods to the fixed value methods for eliciting quantiles. They find a slight but consistent improvement in accuracy and precision with fixed-value methods. The subjects also responded faster and found the questions easier with fixed-value methods.

Verbal assessment uses descriptive words such as most likely, fairly, least likely, and so on. Wallsten *et al.* [41] observe that people prefer to communicate verbally and receive information numerically. In addition, Wallsten *et al.* determined that the overall quality of decisions made from verbal probability assessment is not different from those made with numerical probability assessment. The main difference between the two assessment methods is that, while overconfidence was prevalent in both models, the magnitude was greater in the verbal response method. This is partially due to the use of the 0.5 probability response with the numerical method as a hedge, whereas judges tended not to use the tossup (equally likely) response with the verbal method. Thus, the verbal method may lead to greater honesty. The authors suggest that the analyst provide a fixed list of approximately 11–15 chance-words when using the verbal method. The subject should then assign probability values for the words. The main advantage of verbal encoding is that the forecaster is more comfortable with their assessments. The verbal method relieves

the stress of defending the assessment because it allows vagueness. However, numerical encoding provides a higher level of resolution.

In summary, the literature shows that direct verbal assessment is as good as direct numerical assessment and is easier for the experts. However, indirect methods, particularly the reference lottery method, can motivate experts to think beyond the simple heuristic-based probability assessments.

CONCLUSIONS

In this article, we have discussed how problem decomposition and appropriate elicitation methods and processes can be used to improve probability judgments. We have discussed several approaches. First, we must choose the right events for which to request probability judgments. The analyst must ensure the clarity of what is being asked, decompose events to a reasonable level, and ensure that conditionalities are well understood. This helps to reduce biases caused by the representativeness heuristic. Second, the analyst must choose the right elicitation method to provide motivation to put forth sufficient effort and provide good judgments. It is also crucial to avoid framing effects and offering anchors. While effort must be maximized, it is also important to reduce the cognitive effort required to form a good probability judgment. This allows the mental effort to be spent on retrieving appropriate facts from memory and thoroughly considering their impact on the judgment. Third, incentives can increase motivation, such as scoring rules (see **Scoring Rules**). Last, training and feedback on performance (in terms of calibration, coherence, and resolution) have been shown to reduce biases that can be caused by the availability heuristic.

These methods can help reduce potential biases caused by simple, heuristic thought processes. However, we should point out that the title of this article uses the phrase “minimizing biases,” not removing them. Morgan and Henrion [42] and Benson and Onkal [31] conclude that biases can be reduced, but not removed. There are also numerous

methods for improving probability judgments during the aggregation process. For instance, we can find coherent probabilities “closest” to those provided by the experts (see Refs 43 and 44) and we can recalibrate judgments using performance on almanac questions (see *Eliciting Subjective Probability Distributions from Groups*, and Refs 27 and 45). Aggregation methods are discussed extensively in the article titled *Bayesian Aggregation of Experts’ Forecasts* in this encyclopedia.

REFERENCES

1. North DW, Stengel DN. Decision analysis of program choices in magnetic fusion energy development. *Manage Sci* 1982;37(4): 377–395.
2. Keeney RL, Sarin RK, Winkler RL. Analysis of alternative national ambient carbon monoxide standards. *Manage Sci* 1984;30(4): 518–528.
3. Winkler RL, Poses RM. Evaluating and combining physicians’ probabilities of survival in an intensive care unit. *Manage Sci* 1993; 39(12):1526–1543.
4. DeWispelare AR, Herren LT, Clemen RT. The use of probability elicitation in the high-level nuclear waste regulation program. *Int J Forecast* 1995;11:5–24.
5. Bier VM, Cox LA. Probabilistic risk analysis for engineered systems. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007. pp. 279–301.
6. Paté-Cornell ME. The engineering risk analysis method and some applications. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007. pp. 302–324.
7. Merrick JRW, van Dorp JR, Harrald J *et al.* A systems approach to managing oil transportation risk in Prince William Sound. *Sys Eng* 2000;3(3):128–142.
8. van Dorp JR, Merrick JRW, Harrald J *et al.* A risk management procedure for the Washington State ferries. *Risk Anal* 2001;21(1):127–142.
9. Merrick JRW, van Dorp JR, Mazzuchi T *et al.* The Prince William Sound risk assessment. *Interfaces* 2002;32(6):25–40.
10. Merrick JRW, van Dorp JR, Blackford JP *et al.* Traffic density analysis of proposed ferry service expansion in San Francisco Bay using a maritime simulation model. *Reliab Eng Sys Saf* 2003;81(2):119–132.
11. van Dorp JR, Merrick JRW. Towards the development of a comprehensive vessel traffic risk management procedure. *Ann Oper Res* 2010. In press.
12. Wallsten TS, Budescu DV. Encoding subjective probabilities: a psychological and psychometric review. *Manage Sci* 1983;29(2): 151–173.
13. Gigerenzer G, Hoffrage U, Kleonbolting H. Probabilistic mental models: a brunswikian theory of confidence. *Psychol Rev* 1991;98(4): 506–528.
14. Hora S. Eliciting probabilities from experts. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007.
15. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185:1124–1131.
16. Kahneman D, Tversky A. On the psychology of prediction. *Psychol Rev* 1973;80(4):237–251.
17. Tversky A, Kahneman D. Availability: a heuristic for judging frequency and probability. *Cognitive Psychol* 1973;5:207–232.
18. Combs B, Slovic P. Newspaper coverage of causes of death. *Journal Q* 1979;56(4): 837–843.
19. Kahneman D, Slovic P, Tversky A, editors. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press; 1982.
20. Clemen R, Reilly T. *Making hard decisions*. Duxbury Press; 2001.
21. Keren G. On the calibration of probability judgments: some critical comments on alternate perspectives. *J Behav Decis Making* 1997;10:269–278.
22. Liberman V, Tversky A. On the evaluation of probability judgments. *Psychol Bull* 1993; 114(1):162–173.
23. Erev L, Wallsten TS, Budescu DV. Simultaneous over- and underconfidence: the role of error in Judgment processes. *Psychol Rev* 1994;101:519–527.
24. Pfeifer EP. Are we overconfident in the belief that probability forecasters are overconfident? *Organ Behav Hum Decis Processes* 1994;58:203–213.

25. Juslin P. The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organ Behav Hum Decis Processes* 1994;57:226–246.
26. Murphy AH, Winkler RL. Reliability of subjective probability forecasts of precipitation and temperature. *J Roy Stat Soc Ser C* 1977; 26:41–47.
27. Lin S-W, Bier VM. A study of expert overconfidence. *Reliab Eng Syst Saf* 2007;93:711–721.
28. Tversky A, Kahneman D. Extension versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol Rev* 1983;90(4):293–315.
29. Spetzler CS, von Holstein CSS. Probability encoding in decision analysis. *Manage Sci* 1975;22(3):340–358.
30. Keeney RL, von Winterfeldt D. Eliciting probabilities from experts in complex technical problems. *IEEE Trans Eng Manage* 1991; 38(3):151–173.
31. Benson PG, Onkal D. The effects of feedback and training on the performance of probability forecasters. *Int J Forecast* 1992;8:559–573.
32. Gonzalez-Vallejo C, Bonham A. Aligning confidence with accuracy: revisiting the role of feedback. *Acta Psychopol* 2007;125:221–239.
33. Mellers B, Locke C. What have we learned from our mistakes? *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007. pp. 351–374.
34. Moskowitz H, Sarin RK. Improving the consistency of conditional probability assessments for forecasting and decision making. *Manage Sci* 1983;29(6):735–749.
35. Ravinder HV, Kleinmuntz DN, Dyer J. The reliability of subjective probabilities obtained through decomposition. *Manage Sci* 1988; 34(2):186–199.
36. Andradóttir S, Bier VM. Choosing the number of conditioning events in judgemental forecasting. *J Forecast* 1997;16:255–286.
37. Howard RA. Knowledge maps. *Manage Sci* 1989;35(8):903–922.
38. Benson PG, Curley SP, Smith GF. Belief assessment: an underdeveloped phase of probability elicitation. *Manage Sci* 1995;41(10): 1639–1653.
39. Kadane JB, Winkler RL. Separating probability elicitation from utilities. *J Am Stat Assoc* 1988;83:357–363.
40. Abbas AE, Budescu DV, Yu H-T *et al.* A comparison of two probability encoding methods: fixed probability vs. fixed variable values. *Decis Anal* 2008;5(4):190–202.
41. Wallsten TS, Budescu DV, Zwick R. Comparing the calibration and coherence of numerical and verbal probability judgments. *Manage Sci* 1993;39(2):176–190.
42. Morgan MG, Henrion M. *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press; 1990.
43. Lindley DV, Tversky A, Brown RV. On the reconciliation of probability assessments. *J R Stat Soc* 1979;142(2):146–180.
44. Predd JB, Osherson DN, Kulnarni SR, *et al.* Aggregating probability forecasts from incoherent and abstaining experts. *Decis Anal* 2008;5(4):177–189.
45. Clemen R, Lichtendahl KC. Debiasing expert overconfidence: a bayesian calibration model. In: *Presented at Probabilistic Safety Assessment and Management*; 2002 Jun 23–28; San Juan, Puerto Rico, Volume 6. 2002.

ELICITING SUBJECTIVE PROBABILITY DISTRIBUTIONS FROM GROUPS

ALIREZA DANESHKHAH

MATTHEW REVIE

TIM BEDFORD

LESLEY WALLS

JOHN QUIGLEY

Department of Management
Science, University of
Strathclyde, Glasgow, UK

INTRODUCTION

When eliciting uncertainty assessments from a group, the main problem is that of aggregation: how do we take potentially different assessments and obtain a single coherent assessment from them. There are many different approaches that could be adopted. Cooke [1] suggested that any methodology should conform to a set of five principles: reproducibility, accountability, empirical control, neutrality, and fairness. To ensure that the process is reproducible, it must be possible for a third party to reproduce all calculations. For accountability, the expert source of each input must be identified. Without some form of empirical control, it is difficult to assess the credibility of expert opinion. Neutrality means that the method adopted should encourage experts to state their true opinions. Finally, each expert should be treated equally *a priori*, that is, in the absence of evidence to the contrary. These principles apply largely to the general framework in which an aggregation technique is applied. However, they also have some implications for the structure of the mathematical form of the aggregation rule to be used.

Alongside Cooke's principles, one can consider mathematical requirements that an aggregation rule should satisfy [1,2]. We mention the two main possible desirable properties for an aggregation rule here.

- Marginalization is the process whereby we replace two disjoint events by the union of those two events, so that the probability coherently assessed for the union is the sum of the probabilities for the two events. An aggregation rule has the *marginalization property* if the same marginal aggregated probabilities are obtained when either marginalizing the individual expert probabilities first and then aggregating them, or aggregating the individual probabilities first and then marginalizing the aggregated probabilities. Intuitively, this means that the aggregated probabilities should not depend on the way in which the events have been grouped together.
- An aggregation rule has the *independence preservation property* if when all experts assess A and B as being independent the resulting aggregated assessment makes A independent of B .

It would be desirable if the function that aggregates the experts' assessments to a single distribution had both of the above properties. Unfortunately it can be shown that an aggregation rule cannot have these two properties [1,2]. Hence, we have to decide which is preferable.

Mathematical combination methods are broadly divided into two types: Bayesian methods and pooling methods. These are discussed below. In addition, we discuss issues arising from interaction of experts and the possibilities of behavioral aggregation. In the text below we distinguish between the roles of the expert and of the analyst. The latter is assumed to be conducting the study on behalf of a decision maker.

BAYESIAN APPROACHES

A Bayesian updating scheme is a natural way to take account of data from multiple experts [3–5]. This was first proposed by

Morris [6] (see also Clemen and Winkler [7] for details of this study). A key feature of Bayesian schemes is that the expert opinion is regarded as data. The analyst has to play the role of a metaexpert [8–10] by specifying his own prior distribution and also the likelihood of data from the set of experts (which potentially conflicts with Cooke’s principle of treating experts equally *a priori*, and also means that the analyst plays the role of an expert).

Suppose that n experts provide information separately to an analyst regarding some event or quantity of interest θ , so that this elicited information can be presented as a distribution $f_i(\theta)$ from each expert i . Then the analyst first uses Bayes’ theorem to update his own prior distribution $f(\theta)$ as follows:

$$f(\theta | D) \propto f(\theta)f(D | \theta),$$

where $D = \{f_1(\theta), \dots, f_n(\theta)\}$ is the set of the experts’ elicited distributions, and the likelihood function associated with the expert’s information is shown by $f(D | \theta)$.

This general rule can be used for the aggregation of any kind of information, such as the combination of point estimates and the aggregation of individual probabilities or probability distributions (see O’Hagan *et al.* [11] and references therein for details).

In practice, this approach is frustratingly complex and difficult to implement. The analyst needs to conduct a very sophisticated elicitation exercise on his prior distribution $f(\theta)$ and construct $f(D | \theta)$, which formulates the prior beliefs about what the experts are going to tell him, conditional on the true value of θ [11]. Clemen and Winkler [7] stressed that the likelihood function must account for the precision and bias of the individual information and it must also be able to model dependence among the experts’ elicited information.

Owing to this difficulty in assessing an appropriate likelihood function, considerable effort has gone into finding some models for aggregating single probabilities [4,12] and probability distributions [13]. Clemen and Winkler [7] and O’Hagan *et al.* [11] provide a review of a number of these models.

While Bayesian methods for aggregating expert judgment have been proposed, few have been applied in practice [1]. One exception to this is the model, proposed independently by Mosleh and Aspostolakis [14] and Winkler [15], for combining point estimates. Using the above notation, a method is proposed for assessing $f(D | \theta)$. Mosleh and Aspostolakis [14] suggest two models: one with additive errors and one with multiplicative errors.

For additive errors, it is assumed that the experts’ assessment, $f_i(\theta)$, can be modeled as $f_i(\theta) = \theta + \epsilon_i$. These errors are modeled as multivariate normal with mean μ_i and standard deviation σ_i . The analyst must provide an estimate for μ_i and σ_i , which reflects his assessment of the experts’ bias and accuracy. The likelihood function, $f(D | \theta)$, is simply the value of the normal density with mean $(\theta + \mu_i)$ and standard deviation σ_i . Alternatively, the multiplicative model assumes θ_i is log-normal and that $\theta_i = \epsilon_i\theta$. This is the same as the additive model after the assessments have been logged.

Where multiple experts are used, it seems reasonable to assume that dependency may exist between the different experts. This can be modeled using a normal or log-normal likelihood. However, this will substantially increase the burden on the analysts as they must assess a covariance matrix for the experts’ assessments.

As mentioned above, this approach assumes that the analyst is able to make assessments on the experts. This conflicts with the principle of equality *a priori*.

Other Bayesian approaches include that of Jouini and Clemen [16] which uses copulas (for definition of copula function refer to Dall’Agilo [17]) to construct a multivariate distribution (likelihood function) as a function of the experts’ elicited probability distributions. They also show how this copula-based joint distribution can be used in the Bayesian framework for aggregating expert information; see also Clemen and Winkler [7] for more details.

POOLING METHODS

Pooling methods are simple mathematical combination rules that are straightforward and widely used in practice. The two main approaches are the linear opinion pool and the logarithmic opinion pool. Letting $f_i(\theta)$, $i = 1, \dots, n$ represent the i th expert's probability distribution and w_i to be the weight attached to the i th expert's opinion ($w_i > 0$, $\forall w_i$ and $\sum_{i=1}^n w_i = 1$) then a *consensus distribution* $f(\theta)$ can be obtained as some function of the experts' distributions $D = \{f_1(\theta), \dots, f_n(\theta)\}$ in the following ways:

Linear Opinion Pool. This combination method is the weighted arithmetic mean of the densities given by $f(\theta) = \sum_{i=1}^n w_i f_i(\theta)$. The analyst might give equal weight to each expert by taking $w_i = 1/n$, or choose weights in terms of the experts' expertise. There are various ways of assigning weights [1,11,18], and we discuss two specific methods below. This is the only combination method that satisfies the marginalization property mentioned above.

Logarithmic Opinion Pool. This is a weighted geometric mean of the densities $f(\theta) = k \prod_{i=1}^n f_i(\theta)^{w_i}$, where k is the normalizing constant (a constant to ensure that the combined density integrates to 1). This method, unlike the linear opinion pool, does not satisfy the marginalization property but it does satisfy the independence preservation property mentioned above, and also a more general property called the *external Bayesian* principle. This principle tells us how an analyst updates the combined distribution when new information becomes available [3–5,19].

As discussed by O'Hagan *et al.* [11], the linear and logarithmic opinion pools lead to quite different aggregated distributions. One particular problem of the logarithmic pool is that when one expert believes an event to be impossible, then the aggregated opinion also says that it is impossible. Hence the logarithmic pool tends to concentrate probability very

highly on events for which there is consensus that they are not improbable. The linear opinion pool is widely used in practice, but logarithmic opinion pool is largely ignored because of unrealistically strong aggregated opinions.

Performance-Based Weighting

The main problem with the use of linear opinion pool is that of determining the weights. One way of doing this, which is consistent with the general principles discussed at the start, is to simply give equal weight to each expert. This method is widely used in Ortiza *et al.* [20], but is unsatisfactory because it does not allow us to use extra information that we might learn during elicitation about the experts. This extra information could be used to apply different weights without being inconsistent with the principle of treating the experts equally *a priori*.

There are two reasons why one should be interested in finding *informative weights* (that is, weights that are not just assigned uniformly). First, there may be differing levels of expertise between experts. Second, and more significantly, there may be considerable differences in the ability of subject matter experts to assess uncertainties.

Cooke's classical method gives a method for deriving weights based on the experts' performance in assessing distributions for *seed variables*, which are quantities whose true value is known to the analyst but not to the experts. The weight is determined from two measurements of expert performance on the seed variables. The first is a measure of calibration, and the second, of information. An expert is empirically calibrated if, upon examining those events for which the expert has assessed $p\%$ chance of occurrence, it turns out that $p\%$ actually occur. In the classical method this is assessed by judging how often the seed variables fall between the expert-specified 5%, 50%, and 95% quantiles. An expert's informativeness measures how narrow the uncertainty bands are, with higher informativeness associated with narrower bands. It is measured using the Kullback–Leibler distance between the expert distribution and

a uniform distribution covering the convex hull of all expert assessments. The expert's weight is determined by a combination of the calibration and information measures, judged relative to those of other experts. Experts who perform poorly with respect to the whole group may be assigned zero weight [1].

The weighting scheme proposed by Cooke [1] is asymptotically strictly proper, which means that an expert will achieve maximal weight by truthfully expressing his or her subject probability distribution (if the number of seed variables is large enough). Because of this, the classical method can be claimed to at least partially achieve the principle of neutrality mentioned above.

Cooke and Goossens [21] presented evidence that the Cooke's method produces better elicitation than simple equal weighting of the experts. O'Hagan *et al.* [11] reported that a relevant factor is neglected in the Cooke's method, which is the degree of correlation between experts. They concluded that groups of similar experts will receive too much weight, and that minority views will be underrepresented (see Genest and Zidek [5] and Clemen and Winkler [7,22] for more detailed reviews and discussion of opinion pooling).

One drawback of Cooke's method is that the scoring rule is only "asymptotically" proper. This means that the facilitator may have to ask a large number of questions in order to penalize potential game playing by experts. A recent improved method developed by Wisse *et al.* [23] gives a proper scoring rule that uses moments instead of quantiles. A loss function ϕ is defined as

$$\phi(a_1, \dots, a_n) = \sum_{j=1}^n c_j (x_j - a_j)^2,$$

where $a_1, \dots, a_n$ are the expert's assessed expectations of $X_1, \dots, X_n$ and x_j is the realization of variable X_j . A positive scaling variable, c_j is used to bring each quantity to a common monetary scale. As ϕ is a loss function, the most desirable expert is that which provides the smallest ϕ .

To develop a weighting scheme using the ϕ function, $\phi_1, \dots, \phi_n$ is calculated for

n experts. A cut-off value α is chosen such that any expert with $\phi \geq \alpha$ is given a weight of zero. Each of the remaining experts is assigned a score $\alpha - \phi_i$, which is normalized to give each expert an individual weight. α is chosen using optimization so that the loss of the combined experts is minimized. Wisse *et al.* [23] prove that the unnormalized weight of each expert is a proper scoring rule. While this method uses moments rather than quantiles, it is proposed to move between moments and quantiles using the simple Pearson–Tukey formula.

Performance of Mathematical Aggregation

Unfortunately, there is little literature on the performance of mathematical aggregation methods for probability distributions. We briefly report some important findings from these studies.

Clemen [24] argued that in the methods based on combining point estimates, the equal-weighted linear pool performs as well as more complex weighted averages. In addition, he concluded that the average generally performs almost as well as the best individual forecaster [11]. In a practical study by Winkler and Poses [25], the authors concluded that the best combination was an average of the two most experienced experts and also the least similar pairing.

The general message of the literature is that simple aggregation methods work well in comparison with more complex methods. It is clear that more sophisticated methods can beat a simple equal-weighted linear pool, but the conditions under which this improved performance can be achieved are not clear [11,25].

Cooke and Goossens [21] report very positive results with the classical method. They considered more than 20 applications including nuclear safety, volcanoes, dams, infectious diseases, pollution, and many others. Their findings show that the performance-weighted average generally calibrated as well as or better than either the best individual expert or the equal-weighted average on the seeding variables. These results support the idea that more complex methods might be justified in practical elicitation with real experts. In particular,

experts should be carefully chosen, to obtain good coverage and minimal redundancy, and should be given training in the elicitation techniques to be used; see also Cooke and Goossens [26] for more applications.

BEHAVIORAL AGGREGATION

The behavioral approach aims to generate agreement among the experts through interaction in order to share and exchange knowledge, information, and interpretations between themselves, which will cause them to reconsider their own assessments. This *may* lead to a consensus. If a single distribution is required, then the individual distributions may be combined using a mathematical approach or a consensus distribution may be sought. Such a consensus will naturally involve face-to-face group meetings via discussion and debate, interaction using computers, or sharing of information in some other way to achieve an agreement between the experts. In these situations the role of the facilitator becomes more important.

In general, these methods are not prescribed step-by-step methods but they can be flexible depending on the situation. There are some specific and controlled forms of interaction between the experts. These include the Delphi method, nominal group techniques, and Kaplan's approach.

The *Delphi method* can be applied to various kinds of group judgment and decision making. To elicit a group consensus probability distribution, the method proceeds by first eliciting the experts' assessments separately, then sharing each expert's opinions with all the other experts along with the explanation of that expert's reasoning. The experts then revise their probabilities by considering the new information they have received. This is an iterative process and it might stop when the experts' assessment have converged after a few rounds; otherwise one of the mathematical aggregation methods can be used to combine their assessments (for more details see Linstone and Turoff [27], Pill [28], and O'Hagan *et al.* [11] and references therein).

The *nominal group approach* is a variant of the Delphi method. In this method, each

expert presents her opinions to the group. The group then discusses these judgments with the help of the facilitator, and accordingly experts may revise their assessments [29]. DeGroot [30] gave another modification to this method by proposing that each expert revises her opinion by applying a linear opinion pool to all the experts' distributions, with weights that reflect the importance that each pool member puts on the opinions of each of the other participants.

The facilitator gets much more authoritative roles in the *Kaplan method*. Kaplan [31] believes that the emphasis of group elicitation should be on aggregating the experts' evidence rather than their views. The experts first present and discuss their information about the uncertain quantities in a group meeting. The facilitator then uses this information to deduce a probability distribution to present to the experts. The facilitator must obtain assurance from the experts that information has been interpreted correctly in the proposed distribution. This method is used in probability risk assessment of nuclear reactors, helicopter equipment, and space shuttle [31].

Performance of Behavioral Methods

O'Hagan *et al.* [11] reported that much of the evidence relating to behavioral aggregation has dealt with decision making or point estimation, rather than probability elicitation. In addition, the findings in the literature are generally again that the group often performs less well than a simple average of the individual judgments [32–35]. However, Reagan-Cirincione [36] finds that small group elicitation can outperform their best individual member if the elicitation combines the following three features: impartial facilitation that is responsive to the potential for biases in the group; a well-designed protocol that involves careful structuring and decomposition of the elicitation task; and continuous feedback via computer technology of the implications of the experts' judgments.

Clemen and Winkler [7] reported that there is no best method of combining expert opinion that would apply to all situations. They present several studies that attempt

to compare mathematical and behavioral methods, the results of which are mixed. Phillips [37], based on some experiments, concluded that the consensus distribution method is preferable to mathematical aggregation methods.

To choose the right method in behavioral aggregation, one should consider the type of interaction (face-to-face, via computer, anonymous); the nature of interaction (e.g., sharing information); the possibility of individual reassessment after interaction; and the role of the assessment team (e.g., facilitators). If these methods are not successful in deducing a single consensus distribution for all experts, then a mathematical aggregation may be applied.

DISCUSSION

A variety of methods have been discussed above. The choice between them depends on many “soft” factors. In public sector decision making there is a need for traceable methods that avoid biases and embrace diversity of views within the process. When the need for process validation is less, as might be the case for example in private corporations, then simpler methods may be more cost effective. In this article we have not addressed the issue of how to elicit information from individual analysts. Clearly, when eliciting information from individual experts, whether for aggregation or not, this should be done in ways that minimize known biases such as availability, anchoring, noninformativeness, and so on. In practice, combinations of behavioral approaches and mathematical aggregation may be used, for example by using small groups to help structure the modeling and hence to determine what the appropriate variables are on which further elicitation will take place.

REFERENCES

1. Cooke RM. Experts in uncertainty: opinion and subjective probability in science. New York: Oxford University Press; 1991.
2. Bedford T, Cooke RT. Probabilistic risk analysis: foundations and methods. Cambridge, UK: Cambridge University Press; 2001.
3. French S. Group consensus probability distributions: a critical survey (with discussion). In: Bernardo JM, *et al.*, editors. Bayesian statistics 2. Amsterdam: North-Holland Publishing Co.; 1985. pp. 183–197.
4. Lindley DV. Reconciliation of discrete probability distributions (with discussion). In: Bernardo JM, *et al.*, editors. Bayesian statistics 2. Amsterdam: North-Holland Publishing Co.; 1985. pp. 375–390.
5. Genest C, Zidek JV. Combining probability distributions: a critique and an annotated bibliography (with discussion). *Stat Sci* 1986;1:114–148.
6. Morris PA. Decision analysis expert use. *Manage Sci* 1974;20:1233–1241.
7. Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Anal* 1999;19:187–203.
8. French S. Updating of belief in the light of someone else’s opinion. *J R Stat Soc* 1980;A 143:43–48.
9. Gelfand AE, Mallick BK, Dey DK. Modeling expert opinion arising as a partial probabilistic specification. *J Am Stat Assoc* 1995;90:598–604.
10. Genest C, Schervish MJ. Modelling expert judgments for Bayesian updating. *Ann Stat* 1985;13:1198–1212.
11. O’Hagan A, Buck C, Daneshkhah A, *et al.* Uncertain judgements: eliciting expert probabilities. Chichester: Wiley; 2006.
12. Clemen RT, Winkler RL. Calibrating and combining precipitation probability forecasts. In: Vientl R, editor. Probability and Bayesian statistics. New York: Plenum; 1987. pp. 97–110.
13. Lindley DV. Reconciliation of probability distributions. *Oper Res* 1983;31:866–880.
14. Mosleh A, Aspostolakis G. The assessment of probability distributions from expert opinions with an application to seismic fragility curve. *Risk Anal* 1986;6(4):447–461.
15. Winkler RL. Combining probability distributions from dependent information sources. *Manage Sci* 1981;27:479–488.
16. Jouini MN, Clemen RT. Copula models for aggregating expert opinions. *Oper Res* 1996;44(3):444–457.
17. Dall’Agilo G, Kotz S, Salinetti G. Advances in probability distributions with given marginals: beyond the copulas. Dordrecht, Netherlands: Kluwer academic Publishers; 1991.

18. Genest C, McConway KJ. Allocating the weights in the linear opinion pool. *J Forecast* 1990;9:53–73.
19. McConway KJ. Marginalization and linear opinion pools. *J Am Stat Assoc* 1981;76:410–414.
20. Ortiza NR, Wheelera TA, Breedinga RJ, *et al.* Use of expert judgment in NUREG-1150. *Nucl Eng Des* 1991;126:313–331.
21. Cooke RM, Goossens LHJ. Procedures guide for structured expert judgement in accident consequence modelling; 2000.
22. Clemen RT, Winkler RL. Unanimity and compromise among probability forecasters. *Manage Sci* 1990;36:767–779.
23. Wisse B, Bedford T, Quigley J. Expert judgement combination using moment methods. *Reliab Eng Syst Saf* 2008;93(5): 675–686.
24. Clemen RT. Combining forecasts: a review and annotated bibliography. *Int J Forecas* 1989;5:559–583.
25. Winkler RL, Poses RM. Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Manage Sci* 1993;39:1526–1543.
26. Cooke RM, Goossens LHJ. TU Delft expert judgement data base. Technical Report. Delft, The Netherlands: Delft University of Technology; 2006.
27. Linstone, HA, Turoff, M, editors. The Delphi method: techniques and applications. Reading (MA): Addison-Wesley; 1975.
28. Pill J. The Delphi method: substance, context, a critique, and an annotated bibliography. *Socio Econ Plann Sciences* 1971;5: 57–71.
29. Delbecq AL, van de Ven AH, Gustafson DH. Group techniques for program planning. Glenview (IL): Scott, Foresman and Company; 1975.
30. DeGroot MH. Reaching a consensus. *J Am Stat Assoc* 1974;69:118–121.
31. Kaplan S. 'Expert information' vs 'expert opinions': another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliab Eng Syst Saf* 1992;35:61–72.
32. Gigone D, Hastie R. The common knowledge effect: Information sharing and group judgement. *J Pers Soc Psychol* 1993;65: 959–974.
33. Hill GW. Group versus individual performance: are $N + 1$ heads better than one? *Psychol Bull* 1982;91:517–539.
34. Kerr NL, MacCoun RJ, Kramer GP. Bias in judgement: comparing individuals and groups. *Psychol Rev* 1996;103:687–719.
35. Littlepage G, Robison W, Reddington K. Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organ Behav Hum Decis Processes* 1997;69:133–147.
36. Reagan-Cirincione P. Improving the accuracy of group judgement: a process intervention combining group facilitation, social judgement analysis, and information technology. *Organ Behav Hum Decis Processes* 1994;58:246–270.
37. Phillips LD. Group elicitation of probability distributions: Are many heads better than one? In: Shanteau J, *et al.*, editors. *Decision science and technology: reflections on the contributions of Ward Edwards*. Norwell (MA): Kluwer Academic Publishers; 1999. pp. 313–330.

ELLIPSOIDAL ALGORITHMS

RICARDO H. C. TAKAHASHI
Department of Mathematics,
Universidade Federal de
Minas Gerais, Belo Horizonte,
Minas Gerais, Brazil

INTRODUCTION

The first versions of *ellipsoidal algorithms* (EAs) have been proposed by Shor [1] and Yudin and Nemirovskii [2], based on the idea of generating outer ellipsoidal approximations for convex polytopes. In 1979, Khachiyan [3] employed the ellipsoid method for proving, for the first time, that linear programming problems could be solved in polynomial time. This important theoretical result attracted the attention of the optimization theory community to the EAs [4,5], and much research had been carried out on theoretical issues related to these algorithms in the years that followed [6–15]. However, although it has polynomial-time upper bound for convergence, the EAs have been found to present poor actual convergence rates. Owing to this issue, the EAs have never replaced the simplex algorithm for linear problems. After the introduction of EAs, another family of optimization algorithms, called *the interior point algorithms*, with polynomial-time upper bound for convergence has been developed, which can solve linear problems with competitive actual convergence rates, and are now replacing the simplex methods in several practical contexts [16–19]. Todd [20] presents a perspective of this history, which involves the simplex algorithm, EA, and interior point methods.

However, due to its convergence robustness (the guaranteed convergence to the global optimum of quasi-convex problems), capability to deal with nonsmooth functions, transparent handling of constraints, and

analytical simplicity, the EAs have kept a long-standing relevance in the optimization of nonlinear problems of medium-scale size (up to some dozens of variables and constraints). Several results related to optimization of nonlinear problems can be found, for instance, in the fields of automatic control theory [21–32], computational intelligence [33,34], and electromagnetic device design [35–37]. EAs continue to be employed for the purpose of deriving theoretical results in optimization theory: for instance, in the proof of convergence of algorithms for combinatorial problems [9,38,39], in deriving new algorithms [40], and in complexity analysis [41,42]. For linear programs and their relation with the ellipsoid method, the reader is referred to an extensive appendix given in Chvátal [43].

The goal of this article is to present the working principles behind the EAs. The basic EA is introduced in a constructive manner, and an enhanced version with faster convergence is motivated and presented. The article is organized as follows:

- The classes of problems for which the EAs are guaranteed to converge to the exact solutions (linear problems, convex problems, and quasi-convex problems) are stated.
- The general principle of *cutting-plane branch-and-cut* operation, that is valid for such classes of problems, is presented.
- The basic EA is presented, as an instance of such principles, and the convergence properties of this version of EA are discussed.
- The basic EA is reformulated, with the inclusion of the *deep cuts*, which accelerate the convergence. Specific forms of deep cuts that can be applied to linear- and convex problems are presented.
- Some further implementation issues related to EAs are briefly discussed, and some references are indicated.

MOTIVATION OF ELLIPSOIDAL ALGORITHMS

Consider the following inequality-constrained optimization problem:

$$\begin{aligned} x^* &= \arg \min_x f(x) \\ \text{subject to: } &\begin{cases} g(x) \leq 0 \\ x \in \mathcal{B}. \end{cases} \end{aligned} \quad (1)$$

In this expression, $x \in \mathbb{R}^n$ denotes the decision variable vector, $f(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ denotes the objective function, $g(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ denotes a function that means the “maximal violated constraint”, and $\mathcal{B} \subset \mathbb{R}^n$ denotes a “bounding box” that defines upper- and lower bounds for the admissible values of the decision variable values

$$\mathcal{B} \triangleq \left\{ x \mid \begin{bmatrix} -I \\ I \end{bmatrix} x + \begin{bmatrix} \underline{x} \\ -\bar{x} \end{bmatrix} \leq 0 \right\}, \quad (2)$$

where $\underline{x} \in \mathbb{R}^n$ stores the decision variable lower bounds and $\bar{x} \in \mathbb{R}^n$ stores the decision variable upper bounds. The function $g(\cdot)$ represents a short statement of a set of constraint functions

$$g(x) = \max \{g_1(x), g_2(x), \dots, g_m(x)\}, \quad (3)$$

which makes the following equivalence hold:

$$g(x) \leq 0 \Leftrightarrow \begin{cases} g_1(x) \leq 0 \\ g_2(x) \leq 0 \\ \vdots \\ g_m(x) \leq 0 \end{cases}. \quad (4)$$

The α -sublevel sets of functions $f(x)$ and $g(x)$, denoted respectively by $\mathcal{F}_\alpha$ and $\mathcal{G}_\alpha$, are defined as

$$\mathcal{F}_\alpha \triangleq \{x \mid f(x) \leq \alpha\} \quad \text{and} \quad \mathcal{G}_\alpha \triangleq \{x \mid g(x) \leq \alpha\}. \quad (5)$$

The feasible set $\mathcal{S}$ that contains the decision variable vectors which satisfy all constraints is given by

$$\mathcal{S} = \mathcal{G}_0 \cap \mathcal{B}. \quad (6)$$

Three particular classes of problems are of interest here, since the EAs are suitable to deal with problems within such classes:

- ($\mathcal{P}_1$) Linear problems: $f(x) = cx$ and $g(x) = \max \{Ax - b\}$, with $c \in \mathbb{R}^{1 \times n}$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^{m \times 1}$.
- ($\mathcal{P}_2$) Convex problems: $f(x)$ and all $g(x)$ are convex functions.
- ($\mathcal{P}_3$) Quasi-convex problems: $f(x)$ and all $g(x)$ are quasi-convex functions.

Convex functions are defined by the relation

$$\phi(x) \text{ is convex} \Leftrightarrow \phi(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda \phi(x_1) + (1 - \lambda)\phi(x_2) \quad (7)$$

for any $x_1, x_2 \in \mathbb{R}^n, 0 \leq \lambda \leq 1$. Quasi-convex functions are defined by

$$\begin{aligned} \phi(x) \text{ is quasi-convex} &\Leftrightarrow \\ \Phi_\alpha \text{ is a convex set } &\forall \alpha \in \mathbb{R} \end{aligned} \quad (8)$$

in which Φ_α is the α -sublevel set of $\phi(x)$. It should be noticed that linear problems are also convex problems, and convex problems are also quasi-convex problems:

$$\mathcal{P}_1 \subset \mathcal{P}_2 \subset \mathcal{P}_3. \quad (9)$$

All problems within class $\mathcal{P}_3$ (and consequently within classes $\mathcal{P}_1$ and $\mathcal{P}_2$ also) are endowed with the properties that any sublevel set $\mathcal{G}_\gamma$ is convex, and any sublevel set $\mathcal{F}_\alpha$ is convex. As the intersection of convex sets is a convex set, the sets $\mathcal{G}_\gamma \cap \mathcal{B}$ and $\mathcal{G}_0 \cap \mathcal{F}_\alpha \cap \mathcal{B}$ are also convex. The convexity of these sets is the key property that allows a cutting-plane branch-and-cut procedure that is employed within the EA. The following facts are relevant for allowing such procedures:

- Any point $x_k \in \mathcal{B}$ is on the boundary of the sets $\mathcal{F}_\alpha$ and $\mathcal{G}_\gamma$, for $\alpha = f(x_k)$ and $\gamma = g(x_k)$.

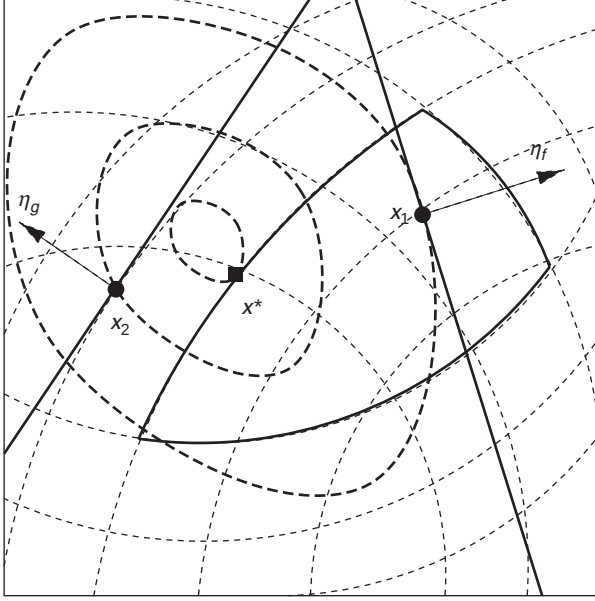


Figure 1. Cutting-plane branch-and-cut operations. The contour lines of the constraint functions g_1 , g_2 , and g_3 are depicted with thin dashed lines. The contour lines of the objective function f are depicted with thicker dashed closed lines. The feasible set is delineated by a thicker continuous closed line with three corners. A cutting plane on the feasible point x_1 would have a normal vector η_f , which is the normal vector of a contour curve of the objective function. A cutting plane on the infeasible point x_2 would have a normal vector η_g , which is the normal vector of the contour curve of the maximally violated constraint function. The point of optimum, x^* , is always kept after any cut that follows such rules.

- The convexity of the sets $\mathcal{F}_\alpha$ and $\mathcal{G}_\gamma$ implies that they have supporting hyperplanes on the point x_k , such that

$$\begin{aligned} \eta_f^T(x - x_k) &\leq 0 \quad \forall x \in \mathcal{F}_\alpha \\ \eta_g^T(x - x_k) &\leq 0 \quad \forall x \in \mathcal{G}_\gamma \end{aligned} \quad (10)$$

where η_f denotes a unitary vector normal to the supporting hyperplane of set $\mathcal{F}_\alpha$ and η_g denotes a unitary vector normal to the supporting hyperplane of set $\mathcal{G}_\gamma$.

The search for the point x^* starts with the only *a priori* information that $x^* \in \mathcal{B}$. A branch-and-cut algorithm can be employed in order to sequentially discard the portions of the set $\mathcal{B}$, such that the search continues on the remaining portions of that set, until a confidence region is small enough to guarantee that any point inside it is a reasonable approximation of x^* . The cutting-plane branch-and-cut algorithm is based on the idea that if one is able to compute a vector η_f and a vector η_g in any point x_k then, after this computation, it is possible to discard the points belonging to a semispace at one side of the cutting plane:

- In the case of infeasible x_k (in this case $\gamma > 0$), discarding the region where $\eta_g^T(x - x_k) > 0$ holds would take off a portion of the space composed of only infeasible points (since such points would be on the boundary of a set $\mathcal{G}_{\bar{\gamma}}$ with $\bar{\gamma} > \gamma$).
- In the case of feasible x_k (in this case $\gamma \leq 0$), discarding the region where $\eta_f^T(x - x_k) > 0$ holds would take off a portion of the space composed of points having objective function values higher than α (since such points would be on the boundary of a set $\mathcal{F}_{\bar{\alpha}}$ with $\bar{\alpha} > \alpha$).

The cutting-plane branch-and-cut operations are illustrated in Figure 1.

In the following section one kind of *cutting-plane branch-and-cut* procedure that can be used to solve problems of classes $\mathcal{P}_1$, $\mathcal{P}_2$, or $\mathcal{P}_3$ is discussed. The basic EA is presented, in the sequence, as a specific instance of such procedures.

Cutting-Plane Branch-and-Cut

Consider an optimization problem defined in expression (1) belonging to the class $\mathcal{P}_3$ (a quasi-convex problem). A *cutting-plane branch-and-cut procedure* that produces a

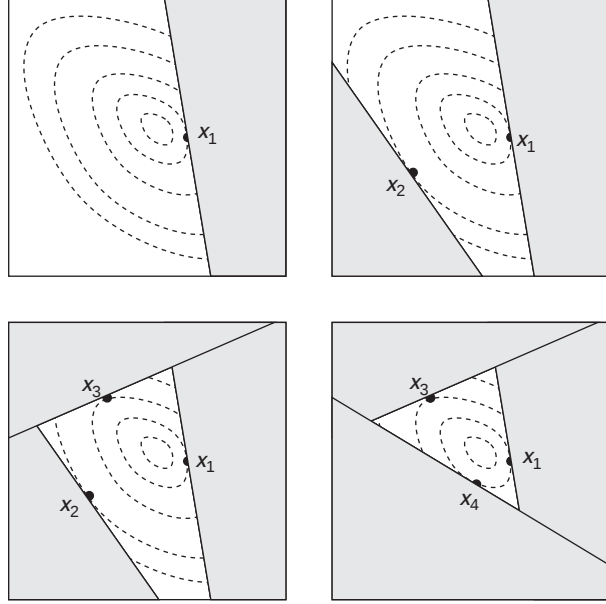


Figure 2. Four successive steps of the cutting-plane branch-and-cut procedure. The contour curves of a convex objective function are represented with dashed closed curves. In all part figures, a cutting plane is introduced on x_k , and the semispace that is excluded is shaded. The unshaded part is the remaining portion of Ω_k after the semispace exclusion.

sequence of points $[x_k]_{k=0}^{\infty}$, which, under certain conditions, converges to the solution x^* of problem (1) can be stated as in Algorithm 1. The idea of Algorithm 1 is to start the search with the information that x^* is located inside the set $\Omega_0 = \mathcal{B}$. In each step, the evaluation of a vector η_k brings new information about a region in which x^* could not be located—a nonoptimality region. The set Ω_{k+1} is calculated by discarding the intersection of Ω_k with such nonoptimality regions.

In each step, the set Ω_k has a “confidence region,” in which the optimum x^* is known to reside. The point x_k is chosen in each step as a “center” of the current confidence region Ω_k , as an attempt to get a maximal average volume-discarding rate along the algorithm iterations. A precise meaning must be attributed for “center,” in any realization of Algorithm 1. Few steps of the cutting-plane branch-and-cut procedure are illustrated in Figure 2.

Algorithm 1. Cutting-plane branch-and-cut procedure.

- 1: $k \leftarrow 0$
- 2: $\Omega_k \leftarrow \mathcal{B}$
- 3: **while not** stop criterion reached **do**
- 4: Choose a point $x_k \in \Omega_k$, such that x_k is located at “the center” of Ω_k
- 5: $\alpha \leftarrow f(x_k)$
- 6: $\gamma \leftarrow g(x_k)$
- 7: **if** $\gamma > 0$ **then**
- 8: $\eta_k \leftarrow \eta_g(x_k)$
- 9: **else**
- 10: $\eta_k \leftarrow \eta_f(x_k)$
- 11: **end if**
- 12: $\Omega_{k+1} \leftarrow \Omega_k / \{x \mid \eta_k^T \cdot (x - x_k) > 0\}$
- 13: $k \leftarrow k + 1$
- 14: **end while**

Define δ_k as the difference:

$$\delta_k = x^* - x_k. \quad (11)$$

It is expected that as the size of Ω_k becomes smaller, δ_k will also become smaller. Depending on the rule that is employed to choose x_k , it may be possible to guarantee that $\|\delta_k\| \rightarrow 0$, which would mean that the sequence $[x_k]_{k=1}^\infty$ converges to x^* .

The following sections discuss the realization issues of steps 4, 8, 10, and 12 of Algorithm 1.

Computing η_g and η_f . In the most general case of problems of class $\mathcal{P}_3$, the computation of η_g or η_f would require a local computation of the level set of function $f(\cdot)$ or $g(\cdot)$ (using for instance a level set method), followed by the computation of a normal vector to this level set. In particular cases, there are more straightforward ways for determining η_k :

- If the functions $f(\cdot)$ and $g(\cdot)$ are differentiable, then η_f and η_g are given by

$$\begin{aligned} \eta_f(x_k) &= \nabla f(x_k), \\ \eta_g(x_k) &= \nabla g(x_k). \end{aligned} \quad (12)$$

- If the functions $f(\cdot)$ and $g(\cdot)$ are convex (which means that the problem is of class $\mathcal{P}_2$, or convex), then η_f and η_g are given by

$$\begin{aligned} \eta_f(x_k) &= \nabla_s f(x_k), \\ \eta_g(x_k) &= \nabla_s g(x_k), \end{aligned} \quad (13)$$

where the notation $\nabla_s(\cdot)$ stands for a subgradient of the argument function.

- If the function $f(\cdot)$ is linear and $g(\cdot)$ is the maximum of affine functions (which means that the problem is of class $\mathcal{P}_1$, or linear), then η_f and η_g are given by

$$\begin{aligned} \eta_f(x_k) &= c^T, \\ \eta_g(x_k) &= A_i^T, \quad A_i x_k - b_i \geq A_j x_k - b_j \quad \forall j = 1, \dots, m, \end{aligned} \quad (14)$$

where A_i and b_i denote the i th row of matrices A and b .

Storing Ω_k and Choosing x_k . The problems of storing the set Ω_k and determining x_k as a “center” for such sets are interrelated. The initial set, $\Omega_0 = \mathcal{B}$, has the form of a box-constraint defined by Equation (2). Each step with a semispace exclusion action inserts a new constraint of the form $\eta_k^T x - \eta_k^T x_k \leq 0$. After k steps, the set Ω_k is given by

$$\Omega_k = \left\{ x \mid \begin{bmatrix} -I \\ I \\ \eta_1^T \\ \eta_2^T \\ \vdots \\ \eta_k^T \end{bmatrix} x + \begin{bmatrix} \frac{x}{2} \\ -\frac{x}{2} \\ \eta_1^T x_1 \\ \eta_2^T x_2 \\ \vdots \\ \eta_k^T x_k \end{bmatrix} \leq 0 \right\}. \quad (15)$$

While the center of the box Ω_0 can be determined by the simple relation $x_0 = \frac{x+\bar{x}}{2}$, finding x_k for $k \geq 1$ is a more involved problem. There is no longer an explicit representation of Ω_k as a convex combination of known vertices. The task of finding such a “center,” therefore, becomes increasingly difficult as k grows, since the constraint matrix receives one additional row per iteration.

An issue must be discussed at this point: comparing the set Ω_k , as stated in Equation (15), with the feasible solution set of a linear problem (a problem of class $\mathcal{P}_1$), which is given by $Ax - b \leq 0$, it becomes evident that both sets have a similar structure. The problem of finding a x_k associated to an Ω_k is similar to a problem of finding a feasible point for a linear problem. Therefore, the following points are suggested:

- In the case of linear problems, it makes no sense to use Algorithm 1 with such direct formulations, storing the set Ω_k via a representation like Equation (15), and calculating x_k from that representation. The reason is that the problem of finding x_k would become eventually more difficult than the original problem, and a solver for linear problems would be still necessary for solving it. Therefore, in the case of linear problems, Algorithm 1 only makes sense along with a kind of implicit representation of the set Ω_k , in order to store the information about each new cut without an exhaustive description as in Equation

(15). Such an implicit representation should also lead to a straightforward computation of a “center” x_k , in order to avoid reaching a subproblem that becomes as difficult as the original problem. The idea of the “centers” in the primal feasible-path versions of the *interior point methods* can be interpreted as an implicit encoding of progressive cuts that are added in each iteration (see **Basic CP Theory: Search**).

- In the case of nonlinear convex or quasi-convex problems (problems belonging to $\mathcal{P}_2/\mathcal{P}_1$ or $\mathcal{P}_3/\mathcal{P}_1$), finding x_k inside Ω_k can be interpreted as a sequence of linear feasibility problems, which replaces the original problem (1). This can be useful only if each linear feasibility subproblem is easier than the original problem. This is indeed the case of any problem that is suitable for being solved via the cutting-plane methods. The traditional cutting-plane methods promote an outer linear approximation of the nonlinear feasible set of increasing precision around the point of optimum. Algorithm 1, on the other hand, promotes an inner–outer approximation.

The basic EA can be interpreted as a realization of Algorithm 1, employing an ellipsoid as an implicit representation of the set Ω_k , and the ellipsoid center as the estimate of x_k .

THE BASIC ELLIPSOID ALGORITHM

It has been established that the raw cutting-plane branch-and-cut algorithm poses two main difficulties: (i) the confidence set Ω_k becomes increasingly complicated, as the number of algorithm iterations grows; and (ii) finding the center x_k of such an Ω_k becomes increasingly difficult. The basic EA can be interpreted as an attempt to solve such difficulties.

The main idea is to approximate the set Ω_k by an ellipsoid that contains it. The center of the ellipsoid becomes, in this case, an approximation for the center of Ω_k . Consider the problem (1). The basic EA begins with an

initial ellipsoid E_0 as follows:

$$E_0 \triangleq \{x | (x - x_0)^T Q_0 (x - x_0) \leq 1\}, \quad (16)$$

where $x_0 \in \mathbb{R}^n$ is the ellipsoid center and $Q_0 \in \mathbb{R}^{n \times n}$, $Q_0 = Q_0^T > 0$, is a symmetric positive-definite matrix. This ellipsoid is chosen such that $E_0 \supset \mathcal{B}$, which means that the solution x^* of problem (1) belongs to E_0 . At each iteration, the vector η_k is chosen as described in the section titled “Computing η_g and η_f .” This vector η_k defines the cut direction on the ellipsoid at each step. Using this η_k , the updates

$$\begin{aligned} x_{k+1} &= x_k - \beta_1 \frac{Q_k \eta_k}{(\eta_k^T Q_k \eta_k)^{\frac{1}{2}}}, \\ Q_{k+1} &= \beta_2 \left(Q_k - \beta_3 \frac{(Q_k \eta_k)(Q_k \eta_k)^T}{\eta_k^T Q_k \eta_k} \right), \end{aligned} \quad (17)$$

with

$$\beta_1 = \frac{1}{n+1} \quad \beta_2 = \frac{n^2}{n^2-1} \quad \beta_3 = \frac{2}{n+1} \quad (18)$$

generate a sequence of points x_k corresponding to the centers of new ellipsoids E_k :

$$E_k = \{x | (x - x_k)^T Q_k (x - x_k) \leq 1\}. \quad (19)$$

The derivation of these formulae is discussed in Refs 4 and 9.

Similar to the confidence set Ω_k in the cutting-plane branch-and-cut algorithm, all such ellipsoids E_k are guaranteed to contain the solution x^* . Geometrically, at each step, a hyperplane H_k with normal vector η_k is constructed passing through x_k . This hyperplane divides the ellipsoid E_k in half. The new ellipsoid is the smallest ellipsoid enclosing the half of E_k that contains the solution x^* . An iteration of the EA is illustrated in Figure 3. The evolution of the ellipsoids in a convex problem, from the first to the last iteration, is shown in Figure 4.

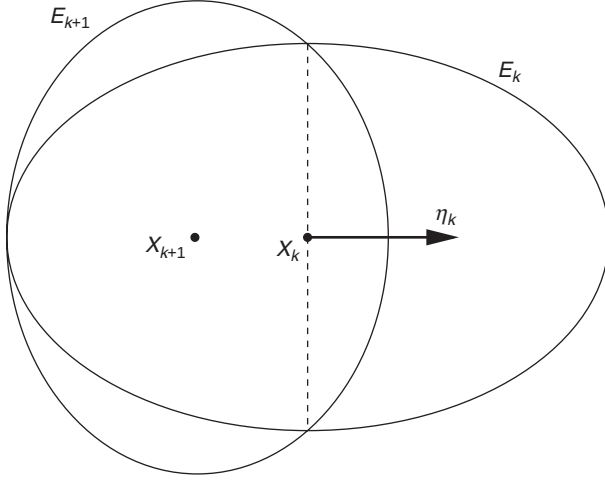


Figure 3. One step of the basic ellipsoid algorithm. The ellipsoid E_k receives a cut that passes its center, x_k , through the cutting plane normal to vector η_k . The new ellipsoid E_{k+1} has center x_{k+1} .

Convergence Analysis

The convergence rate of the basic EA is measured usually in terms of the reduction rate of the ellipsoid volume. This criterion is a direct consequence of the working principle behind the cutting-plane branch-and-cut procedure, which relies on the sequential exclusion of portions of the decision variable space, producing ultimately a sequence of smaller confidence regions, which are guaranteed to contain the point of optimum of quasi-convex problems.

In an idealized situation, the best possible case of that convergence rate for Algorithm 1 would correspond to a reduction rate of 0.5 in the confidence region volume per iteration. However, this situation cannot be reached in practice because it is not possible to guarantee that there is a “center” x_k of region Ω_k for which the region will be divided in two subregions of exactly the same volume for any cutting plane passing over it. As most of the cuts on arbitrary points x_k will discard the smaller portion of the confidence set (the reader can verify this statement by drawing the contour plot of a quasi-convex function inside a confidence region that contains the optimum, and check which regions are discarded when cutting planes are defined over random points), the following expression will hold in an average sense:

$$V[\Omega_k] \geq \frac{1}{2^k} V[\Omega_0] \quad (20)$$

where the function $V[\cdot]$ means the volume of the argument set.

In the case of the basic EA, the volume $V[\cdot]$ of the ellipsoid E_k decreases according to

$$V[E_k] = R_n^k \cdot V[E_0], \quad (21)$$

where

$$R_n = \frac{n}{n+1} \left(\frac{n^2}{n^2-1} \right)^{\frac{n-1}{2}} < 1. \quad (22)$$

The reduction ratio of the ellipsoid volume, R_n , depends on the dimension n of the space. For higher dimensions, R_n grows, and the algorithm presents slower ellipsoid volume reduction rate. The value of R_n as a function of dimension n is presented in Figure 5.

The explanation of such behavior comes from the excess of space volume circumscribed by an ellipsoid E_{k+1} in relation to the volume of the half of the former ellipsoid E_k . This issue is illustrated in Figure 6. Owing to this volume excess that is caused by the usage of ellipsoids as confidence regions, the rate of volume exclusion becomes slower when compared to an idealized instance of the cutting-plane branch-and-cut algorithm, which would present a volume compression rate of about 0.5 per iteration. As the excess becomes greater for higher dimensions, this effect causes the method to have a decreased efficiency as the problem

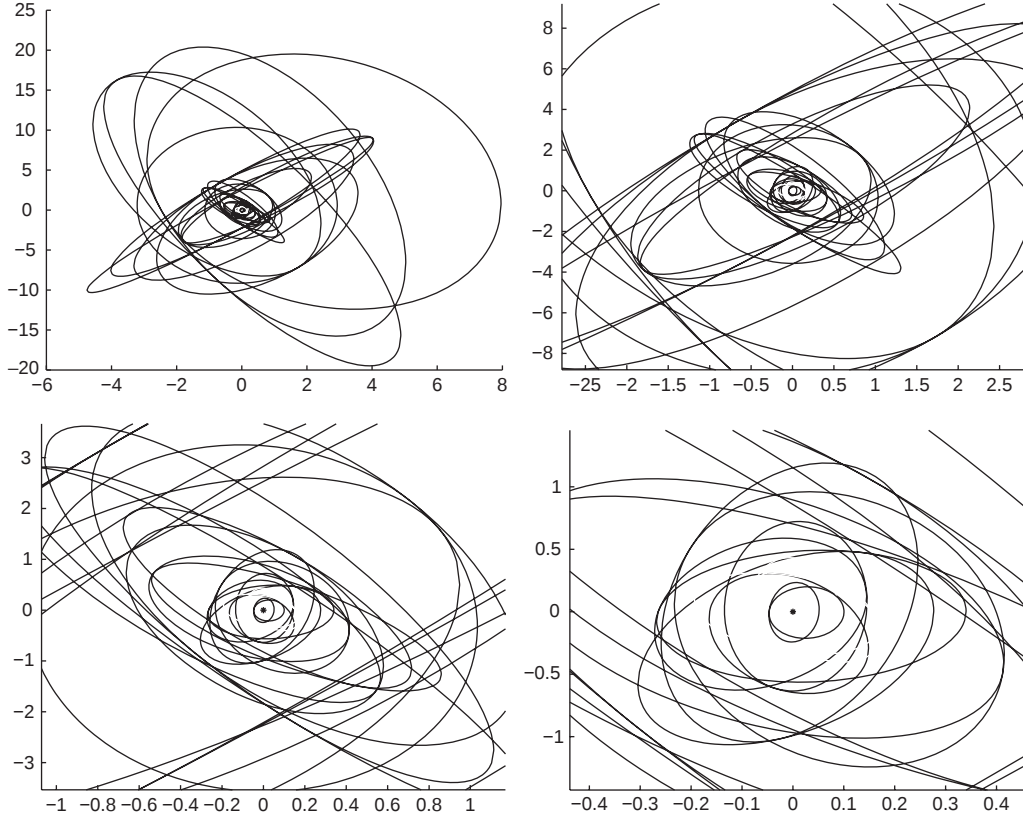


Figure 4. Evolution of ellipsoids generated by the basic ellipsoid algorithm in a convex problem, which has the point of optimum situated in coordinates $(0, 0)$. The part figure at the upper left shows all the ellipsoids. The sequence of part figures presents successive zooms of the region around the optimum, such that the last ellipsoid becomes visible in the lower right figure. It should be noticed that all the ellipsoids contain the point of optimum.

dimension increases. This behavior renders the basic EA ineffective even in problems of moderate dimensions. This in part explains why the EAs have received more attention, in the last years, in the context of optimization of nonlinear functions of relatively low dimension. In order to mitigate this difficulty, some variations in the method involving “deep cuts” have been developed. This is the subject of the next section of this article. Further discussions about the rate of convergence of EAs can be found in Refs 44 and 45.

Remark. A common mistake about the meaning of the convergence of the EA is the expectation that, as the volume of the ellipsoid E_k approaches zero, all points inside this

ellipsoid become arbitrary near the ellipsoid center x_k . This expectation is false because for an ellipsoid to reach a zero volume, it is enough that just one of its axes approaches zero size. Indeed, the ultimate ellipsoid in the EA very often has at least one axis that remains large (the ellipsoid in those cases tends to degenerate into a line, or into a plane, or into any other zero-measure flat). Nevertheless, even with a final ellipsoid with nonzero length axis, the center x_k usually terminates “near” the point of optimum x^* . This statement holds because not only the interior of the last ellipsoid of the sequence but also the “confidence region” is an intersection set of all the ellipsoids of the sequence $\{E_0, \dots, E_k\}$, and this intersection is usually small.

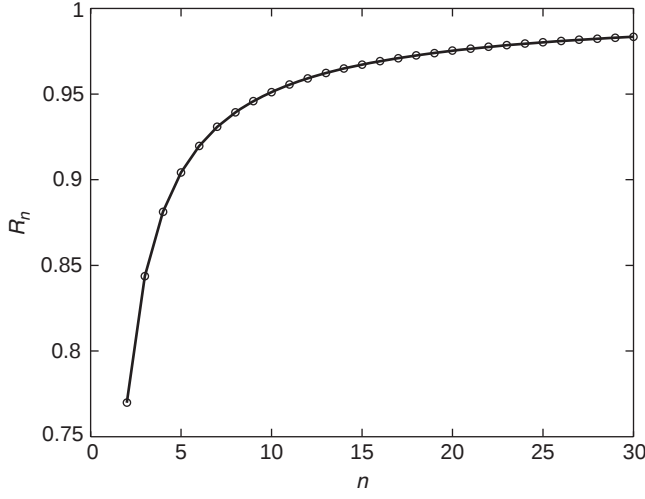


Figure 5. The volume contraction rate (R_n) of the basic ellipsoid algorithm, as a function of the space dimension n .

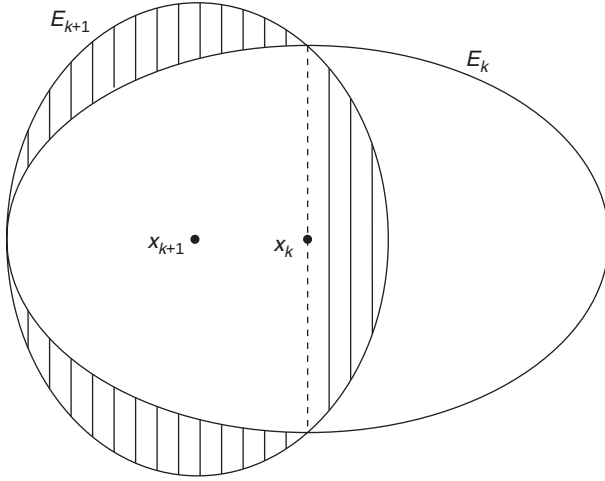


Figure 6. The same iteration of the basic ellipsoid algorithm presented in Figure 3, going from the ellipsoid E_k with center x_k , to the ellipsoid E_{k+1} with center x_{k+1} . The shaded areas within the ellipsoid E_{k+1} represent regions that were not contained in the half of the ellipsoid E_k that should be kept, which cause some loss of computational efficiency.

DEEP CUTS IN THE ELLIPSOID ALGORITHM

The *deep cut* procedure is obtained by using the update formulae (17), with the parameters β_1 , β_2 , and β_3 given by

$$\begin{aligned} \beta_1 &= \frac{1+n\psi}{n+1} & \beta_2 &= \frac{n^2(1-\psi^2)}{n^2-1} \\ \beta_3 &= \frac{2(1+n\psi)}{(n+1)(1+\psi)} \end{aligned} \quad (23)$$

instead of the ones stated in Equation (18). The update formulae using Equation (23)

generates a sequence of ellipsoids that contain the intersection of each former ellipsoid with a semispace H_k which does not pass through that former ellipsoid center:

$$H_k \triangleq \left\{ x \mid \eta_k^T(x - x_k) \leq -\psi \left[\eta_k^T Q_k \eta_k \right]^{\frac{1}{2}} \right\}. \quad (24)$$

Parameter ψ is called the *depth* of the cut, and represents the distance from x_k to the half-space H_k in the metric corresponding to matrix Q_k . Equation (23) is suitable for determining E_{k+1} for $-1/n \leq \psi \leq 1/n$. In the range $0 \leq \psi \leq 1/n$, the procedure leads

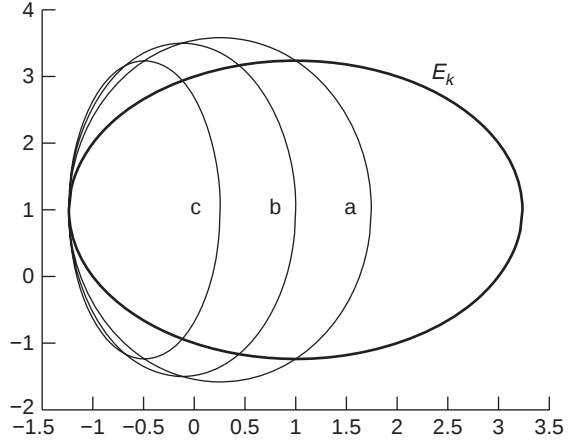


Figure 7. The effect of applying deep cuts. Starting from the ellipsoid E_k , (a) represents the ellipsoid E_{k+1} when a conventional cut (with no deep cut, or $\psi = 0$) is applied, (b) represents E_{k+1} for $\psi = 0.25$, and (c) represents E_{k+1} for $\psi = 0.5$.

to an accelerated convergence, since the intersection $H_k \cap E_k$ is smaller than half E_k for any $\psi > 0$ and the volume of the resulting ellipsoid E_{k+1} becomes smaller, in this range, than the volume that would result from the ordinary cut. This is illustrated in Figure 7.

The relevant questions are as follows:

- Given a problem, can a deep cut be applied within the EA?
- In a positive case, what cut depth ψ should be applied?

The answer to these questions will depend on the class of problems under consideration. In the case of quasi-convex problems (problems of class $\mathcal{P}_3$), which are the most general class of problems that can be dealt with by EAs, the application of deep cuts may lead to the loss of the solution—therefore, they should not be applied, unless the possibility of losing the optimum can be tolerated, as in the case of Refs 8 and 36. Fortunately, for problems of classes $\mathcal{P}_1$ (linear problems) and $\mathcal{P}_2$ (convex problems), there is a possibility of performing deep cuts while keeping the guaranteed convergence property [13].

Examine first the linear problems. In such problems, the cut normal vector in the case of infeasible points will have the form $\eta_k = A_i^T$, where i denotes the index of the most violated linear constraint. Notice that the feasibility of this constraint occurs for x such that

$$A_i x - b_i \leq 0. \quad (25)$$

The deep cut, instead of simply cutting the ellipsoid in its current center x_k , can directly use Equation (25), such that the maximal cut that still guarantees the constraint feasibility is applied. This can be performed by choosing ψ such that the deep cut expression

$$A_i(x - x_k) \leq -\psi (A_i Q_k A_i^T)^{\frac{1}{2}} \quad (26)$$

becomes equivalent to Equation (25). This leads to

$$\psi = \frac{A_i x_k - b_i}{(A_i Q_k A_i^T)^{\frac{1}{2}}}. \quad (27)$$

Noticing that ψ should not exceed $\frac{1}{n}$, the cut depth expression becomes

$$\psi = \min \left\{ \frac{A_i x_k - b_i}{(A_i Q_k A_i^T)^{\frac{1}{2}}}, \frac{1}{n} \right\}. \quad (28)$$

For determining the cut depth for feasible points, define $\hat{x}_k$ as the best feasible point that has been found up to iteration k (but not necessarily at iteration k). The following constraint applies to the problem optimum x^* :

$$c x^* \leq c \hat{x}_k. \quad (29)$$

This leads to the following expression for the cut depth when x_k is feasible:

$$\psi = \min \left\{ \frac{c x_k - c \hat{x}_k}{(c Q_k c^T)^{\frac{1}{2}}}, \frac{1}{n} \right\}. \quad (30)$$

In the case of problems of class $\mathcal{P}_2$ (convex problems), the following relations hold:

$$\begin{aligned} f(x) &\geq f(x_k) + \eta_f^T \cdot (x - x_k), \\ g(x) &\geq g(x_k) + \eta_g^T \cdot (x - x_k). \end{aligned} \quad (31)$$

These relations are direct consequences of convexity. In the case of x_k infeasible, the set defined by

$$g(x_k) + \eta_g^T \cdot (x - x_k) \leq 0 \quad (32)$$

strictly contains the feasible set $g(x) \leq 0$. Therefore, a deep cut that preserves the region in which Equation (32) holds can be applied. This leads to the expression

$$\psi = \min \left\{ \frac{g(x_k)}{\left(\eta_g^T Q_k \eta_g \right)^{\frac{1}{2}}}, \frac{1}{n} \right\}. \quad (33)$$

In the case of feasible x_k , taking $\hat{x}_k$ as the best feasible solution found up to iteration k , the following relation holds:

$$f(\hat{x}_k) \geq f(x^*). \quad (34)$$

The set defined by

$$\{x \mid f(x) \leq f(\hat{x}_k)\} \quad (35)$$

strictly contains the point of optimum x^* . Therefore, the deep cut given by

$$\psi = \min \left\{ \frac{f(x_k) - f(\hat{x}_k)}{\left(\eta_f^T Q_k \eta_f \right)^{\frac{1}{2}}}, \frac{1}{n} \right\} \quad (36)$$

can be applied.

FURTHER IMPLEMENTATION ISSUES

The above-mentioned points show the importance of using cuts that accelerate the ellipsoid contraction. This issue has been considered in several works:

- Dziuban *et al.* [8] Frenk *et al.* [13] study the issue of applying deep cuts, and Frenk and Gromicho [44] and Gromicho [45] study the convergence of such schemes.
- Ech-Cherif and Ecker [7], Ech-Cherif *et al.* [10], and Liao and Todd [11] discuss the application of two simultaneous cutting planes for cutting the ellipsoid.
- Kim *et al.* [12] discuss the formulation of a cut that incorporates the cut information of two successive iterations.

A very important issue that must be considered in the implementation of the EA is the numerical stability of the ellipsoid matrix Q_k . The successive application of the matrix update rule can lead to the loss of matrix positivity, which forces a premature termination of the algorithm, before reaching the user-defined stop condition. Some indications of matrix decompositions that can be used for avoiding such numerical problems are presented in Refs 4 and 6.

An extension of the EA for the case of equality constraints has been developed in Refs 14 and 15.

REFERENCES

1. Shor NZ. Cut-off method with space extension in convex programming problems. *Cybernetics* 1977;12:94–96.
2. Yudin DB, Nemirovskii AS. Informational complexity and effective methods for solving convex extremal problems. *Sov Math Matekon* 1977;13:24–45.
3. Khachiyan LG. A polynomial algorithm in linear programming. *Sov Math Matekon* 1979; 20:191–194.
4. Bland RG, Goldfarb D, Todd MJ. The ellipsoid method: a survey. *Oper Res* 1981;29(6): 1039–1091.
5. Akgül M. Volume 97, Topics in relaxation and ellipsoidal methods, research notes in mathematics. London: Pitman Publishing Inc.;1984.
6. Todd MJ. On minimum volume ellipsoids containing part of a given ellipsoid. *Math Oper Res* 1982;7(2):253–261.
7. Ech-Cherif A, Ecker JG. A class of rank-two ellipsoid algorithms for convex programming. *Math Program* 1984;29:187–202.

8. Dziuban ST, Ecker JG, Kupferschmid M. Using deep cuts in an ellipsoidal algorithm for nonlinear programming. *Math Program Study* 1985;25:93–107.
9. Grötschel M, Lovász L, Schrijver A. Geometric algorithms and combinatorial optimization. Berlin: Springer; 1988.
10. Ech-Cherif A, Ecker JG, Kupferschmid M. A numerical investigation of rank-two ellipsoid algorithms for nonlinear programming. *Math Program* 1989;43(1):87–95.
11. Liao A, Todd MJ. The ellipsoid algorithm using parallel cuts. *Comput Optim Appl* 1993; 2(4):299–316.
12. Kim S, Kim D, Chang KN. Using two successive subgradients in the ellipsoid method for nonlinear programming. *J Optim Theory Appl* 1994;82(3):543–554.
13. Frenk JBG, Gromicho J, Zhang S. A deep cut ellipsoid algorithm for convex programming: theory and applications. *Math Program* 1994;63(1–3):83–108.
14. Shah S, Mitchell JE, Kupferschmid M. An ellipsoid algorithm for equality-constrained nonlinear programs. *Comput Oper Res* 2001; 28(1):85–92.
15. Rugenstein EK, Kupferschmid M. Active set strategies in an ellipsoid algorithm for nonlinear programming. *Comput Oper Res* 2004;31(6):941–962.
16. Traub JF, Woniakowski H. Complexity of linear programming. *Oper Res Lett* 1982;1(2): 59–62.
17. Karmarkar N. A new polynomial-time algorithm for linear programming. *Proceedings of the 16th ACM Annual Symposium on Theory of Computing*. New York (NY): ACM; 1984. pp. 302–311.
18. Ye Y. Karmarkar's algorithm and the ellipsoid method. *Oper Res Lett* 1987;6(4):177–182.
19. Illes T, Terlaky T. Pivot versus interior point methods: pros and cons. *Eur J Oper Res* 2002;140:170–190.
20. Todd MJ. The many facets of linear programming. *Math Program* 2002;91(3):417–436.
21. Kupferschmid M, Mohrmann K, Ecker JG, *et al.* The ellipsoid algorithm: a new method for feedback gain optimization. *Annu Rev Automat Program* 1985;13(2):85–94.
22. Boyd S, Barratt CH. Linear controller design: limits of performance. Englewood Cliffs, New Jersey: Prentice-Hall; 1991.
23. Cheung MF, Yurkovich S, Passino KM. An optimal volume ellipsoid algorithm for parameter set estimation. *IEEE Trans Automat Control* 1993;38(8):1292–1296.
24. Boyd S, El-Ghaoui L, Feron E, *et al.* Linear matrix inequalities in system and control theory. Philadelphia, Pennsylvania: SIAM; 1994.
25. Rotstein H, Sideris A. $\mathcal{H}_\infty$ optimization with time-domain constraints. *IEEE Trans Automat Control* 1994;39(4):762–779.
26. Correa MV, Aguirre LA, Saldanha RR. Using steady-state prior knowledge to constrain parameter estimates in nonlinear system identification. *IEEE Trans Circuits Syst Fundam Theory Appl* 2002;49(9):1376–1381.
27. Kanev S, de Schutter B, Verhaegen M. An ellipsoid algorithm for probabilistic robust controller design. *Syst Control Lett* 2003; 49:365–375.
28. Cabral FB, Safonov MG. Unfalsified model reference adaptive control using the ellipsoid algorithm. *Int J Adapt Control Signal Process* 2004;18(8):683–696.
29. Gonçalves EN, Palhares RM, Takahashi RHC. Improved optimisation approach to the robust $\mathcal{H}_2/\mathcal{H}_\infty$ control problem for linear systems. *IEE Proc Control Theory Appl* 2005;152(2):171–176.
30. Gonçalves EN, Palhares RM, Takahashi RHC. “ $\mathcal{H}_2/\mathcal{H}_\infty$ filter design for systems with polytope-bounded uncertainty”. *IEEE Trans Signal Process* 2006;54(9):3620–3626.
31. Kanev S, Verhaegen M. Robustly asymptotically stable finite-horizon MPC. *Automatica* 2006;42(12):2189–2194.
32. Gonçalves EN, Palhares RM, Takahashi RHC. A novel approach for $\mathcal{H}_2/\mathcal{H}_\infty$ robust PID synthesis for uncertain systems. *J Process Control* 2008;18(1):19–26.
33. Teixeira RA, Braga AP, Takahashi RHC, *et al.* Improving generalization of MLPs with multi-objective optimization. *Neurocomputing* 2000;35(4):189–194.
34. Braga AP, Takahashi RHC, Costa MA, *et al.* Multi-objective algorithms for neural network learning. In: Jin Y, editor. *Multi-objective machine learning*. 1st ed. Berlin: Springer; 2006. pp. 151–171.
35. Saldanha RR, Coulomb JL, Sabonnadiere JC. An ellipsoid algorithm for the optimum design of magnetostatic problems. *IEEE Trans Magn* 1992;28(2):1573–1576.
36. Saldanha RR, Takahashi RHC, Vasconcelos JA, *et al.* Adaptive deep-cut method

- in ellipsoidal optimization for electromagnetic design. *IEEE Trans Magn* 1999;35(3): 1746–1749.
37. Takahashi RHC, Saldanha RR, Dias-Filho W, *et al.* A new constrained ellipsoidal algorithm for nonlinear optimization with equality constraints. *IEEE Trans Magn* 2003;39(5): 1289–1292.
 38. Rarp RM, Papadimitriou CH. On linear characterizations of combinatorial optimization problems. *IEEE Found Comput Sci* 1980;21: 1–9.
 39. Grotschel M, Lovasz L, Schrijver A. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* 1981;1(2):169–197.
 40. Liao A, Todd MJ. Solving LP problems via weighted centers. *SIAM J Optim* 1996;6(4): 933–960.
 41. Freund RM, Vera JR. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM J Optim* 1999;10(1):155–176.
 42. Freund RM, Vera JR. On the complexity of computing estimates of condition measures of a conic linear system. *Math Oper Res* 2003; 28(4):625–648.
 43. Chvátal V. *Linear programming*. New York: W. H. Freeman & Company; 1999.
 44. Frenk J, Gromicho J. An elementary rate of convergence result for the deep cut ellipsoid algorithm. *Recent advances in nonsmooth optimization*. Singapore: World Scientific; 1995. pp. 106–120.
 45. Gromicho J. *Quasiconvex optimization and location theory*. Dordrecht: Kluwer Academic Publishers; 1998.

EMERGENCY MEDICAL SERVICE SYSTEMS THAT IMPROVE PATIENT SURVIVABILITY

LAURA A. MCLAY
Department of Statistical
Sciences and Operations
Research, Virginia
Commonwealth University,
Richmond, Virginia

INTRODUCTION AND MOTIVATION

Responding to emergency cardiac arrest 911 calls is a major focus of emergency medical service (EMS) systems. EMS systems are often evaluated by how effectively they respond to and perform care for cardiac arrest calls because (i) effective treatment is available (i.e., CPR, early defibrillation); (ii) treatment is highly time dependent [Wik *et al.* [1] states that the likelihood of survival decreases by approximately 10% for every minute of delay in providing defibrillation after the onset of cardiac arrest]; (iii) each element in the community emergency response [early access, early CPR, early defibrillation, and early advanced life support (ALS)] must be strong for optimal survival rates; and (iv) if the EMS system can respond effectively to cardiac arrest and achieve a good survival rate, the same elements will usually yield similar outcomes for other life or death conditions such as major traumatic injury or stroke. This has been well-documented by medical research [2]. The EMS response time interval is highly correlated with cardiac arrest patient survival, whereas the link between the EMS response time interval and patient survival from other types of emergency medical conditions is unclear [3].

Despite the link between EMS response and patient outcomes being well-understood, the survival to hospital discharge rate after out-of-hospital cardiac arrest is low, with approximately 5–7% of cardiac arrest patients surviving [4]. Traditional EMS

systems depend on providing ALS to 911 calls within a fixed time frame (e.g., within 8 min 59 s or less of dispatch). However, medical research indicates that for cardiac arrest calls, it is often more effective if early patient care can be initiated by a first responder [using CPR and automated external defibrillators (AEDs)] rather than wait until ALS personnel respond several minutes later [2,4,5].

Erkut *et al.* [6] first propose how to optimally locate ambulances to improve patient survival rates by locating ambulances. They motivate many of the issues described here and highlight the need for further operations research applications in the EMS domain. By optimally utilizing EMS resources, operations research has the opportunity to play a critical role in designing next-generation EMS systems that save patient lives. Operations models for next-generation EMS system design must be driven by state-of-the-art emergency medical research to improve cardiac arrest patient survival rates. This article summarizes the key issues in defining EMS systems, using scarce emergency medical resources, and assessing EMS performance when applying operations research methodologies. For further information, the reader is referred to several excellent reviews of how operations research methodologies have been applied to EMS applications [7–11].

This article is organized as follows. First, background information about EMS systems and their resources are described. Second, ways to assess EMS performance are summarized and ways to measure patient survivability when addressing how to locate ambulances at rescue stations are discussed. Many of the performance measures have applications to other problems in EMS management and design. Next, important problems in next-generation EMS design that can benefit from systematic operations research modeling and analysis are discussed. The problems discussed extend beyond the scope of location models. Next, an illustrative example illustrates some of the key issues raised with data

extracted from Hanover County, Virginia. Finally, concluding remarks are given.

EMS BACKGROUND

Even small EMS systems may use several types of resources. Optimally using these scarce resources is challenging and often counterintuitive. This section summarizes the key resources used by EMS systems, including medical units, personnel, and computer-aided dispatching (CAD) systems.

EMS Resources

Historically, many EMS departments are under the same administration as fire departments and sometimes police departments. Sharing resources with other departments not only increases the number of resources that can be leveraged to respond to emergency medical calls, but also increases the complexity of the problems facing EMS agencies.

EMS departments contain a mixture of different types medical units as well as different types of staff. Ambulances are staffed by basic emergency medical technicians (EMTs) who provide basic life support (BLS) or paramedics (EMT's) who provide ALS. In some EMS systems, there may be several types of medical units that can respond to emergency 911 calls, including ambulances, nontransport quick response vehicles (QRVs), fire engines, and even police cars (police officers can provide CPR and basic first aid; some carry AEDs). Multiple medical units may be dispatched to a single 911 call. If a nontransport unit is dispatched first (such as a QRV or fire engine), then an ambulance must also be dispatched to the call to transport the patient to a hospital. In addition, medivacs (helicopter services operated by paramedics) may also be dispatched in the event that hospital transport is required.

For decades, EMS systems have been measured according to how they respond to and care for cardiac arrest patients [3]. As a result, since the 1960s, EMS systems increasingly provide more medical care at the scene and during transport. Most rural (and some suburban and smaller urban) EMS systems

in the United States have historically relied on volunteer EMTs. Many EMS systems have shifted from BLS to ALS ambulances, and hence, from EMTs to paramedics. At least one paramedic must operate a medical unit for it to be considered an ALS.

Ambulance staff may be volunteers or paid personnel. Owing to the decline of volunteerism, EMS systems increasingly rely on paid personnel. EMS systems are composed of a mixture of EMTs, paramedics, volunteers, and paid personnel. Many fire fighters are certified as either first responders or EMTs, and hence, they can provide BLS treatment.

Treating EMS Patients

Emergency medical 911 calls are typically classified as Priority 1, 2, 3, where Priority 1 calls are life-threatening emergencies, Priority 2 calls are emergencies that may be life threatening, and Priority 3 calls do not appear to be life-threatening emergencies. There is not a universal protocol for responding to Priority 1, 2, and 3 calls, since the response depends on the types of resources as well as the characteristics and demographics of the area that the EMS system serves. One such protocol for an EMS system with both ALS and BLS units is to deploy ALS units to Priority 1 calls, either type of ambulance for Priority 2 calls, and BLS units to Priority 3 calls. In addition, since many EMS agencies are under the same administration as the fire department, fire engines (with ALS or BLS equipment) can be dispatched to Priority 1 calls to stabilize patients and to Priority 2 and 3 calls if an ambulance is not readily available. Moreover, since not all Priority 1 calls turn out to be life threatening, they may be downgraded to Priority 2 after a unit has arrived on scene. Likewise, Priority 2 or 3 calls can be upgraded. Paramedics may be required to transport Priority 1 patients to the hospital, and hence, dispatching EMTs (BLS) to a Priority 1 call to stabilize a patient would necessitate dispatching a second (ALS) ambulance to the call if the patient requires hospital transport.

The following sequence of events occurs for each emergency medical call handed by EMS agencies [9].

1. An emergency medical 911 call is made and is handled by the dispatch center.
2. The dispatch center estimates the severity of the call.
3. The appropriate vehicles are dispatched to the scene of the call.
4. Service is provided.
5. The patient is transported to the hospital, if needed.
6. The vehicle returns to service and is directed to a predetermined location (e.g., station).

Dispatch Center

The dispatch center is a vital part of any EMS system. The dispatch center or primary public safety answering point receives emergency 911 calls, and determines the appropriate response. Calls arrive at a centralized center, and emergency medical, fire, and police calls are routed to separate dispatchers. In large cities, the EMS dispatch center may be physically located in a separate building from fire and police dispatch centers. When dispatchers answer calls, they usually interface with a CAD system to collect information from the caller, including the address and nature of the calls, dispatch units to the calls, and record information regarding the treatment of the patients.

The following sequence of events is typically recorded by CAD systems.

1. A record is entered into the CAD system.
2. A unit is dispatched to the call.
3. The unit leaves the station.
4. The unit arrives at the scene.
5. The unit leaves the scene for the hospital.
6. The unit arrives at the hospital.
7. The unit leaves the hospital.
8. The unit returns to service.
9. The unit arrives at the station.

Note that if a call does not result in a trip to the hospital, then 5–7 may not be recorded. The *response time* captures the difference between when the call was received (1) or

the unit is dispatched (2) and when it arrives at the scene (4) (or the curbside, when calls are in a skyscraper). The *turnaround time* captures the difference between when the call was received (1) or the unit is dispatched (2) and when the unit returns to service (8). Note that when a call arrives at the dispatch center, it is often not recorded, and hence, times are measured relative to when a unit is dispatched to a call, making it even more difficult to truly estimate how long a patient waits for the EMS response [3].

Large cities may have proprietary CAD systems that indicate how to dynamically dispatch and relocate medical units. In all but the largest US cities, CAD systems are generally not custom-designed, since they are often purchased from outside vendors. CAD systems record information to document EMS performance measures. However, the limitation is that it is often difficult to assess performance measures that are not captured explicitly by CAD systems. For example, it is often impossible to measure the response time interval from the time that the patient collapses. Rather, the “response time interval” is usually the time from dispatch to arrival of the unit on location (which is still not necessarily the time the team is at the patient’s side). It can be difficult to interface EMS CAD systems with fire and police CAD systems, which is necessary to understand the system response characteristics if fire and police respond to EMS calls and provide patient treatment. For example, the same patient may be assigned different incident numbers in the EMS and fire CAD systems. Matching calls by address and time may also be difficult since the difference in dispatch times between the types of units may be several minutes or longer. In addition, there are problems of time synchronization between different timepieces used in the EMS system, which results in documentation errors in response time intervals and other time measures [12].

CAD systems provide a large amount of data, but the data are often problematic [11]. For example, CAD systems capture events that are recorded by dispatchers (such as when an ambulance leaves the station) and do not always reflect when the events

actually occurred. Frequently, events are missing or are obviously incorrect (such as turnaround times that are 0 or negative, and response times that take hours). In addition, many calls are canceled before a medical unit reaches the destination. Hence, patients are “not treated” for a significant proportion of calls in the CAD database.

In EMS system design, the call locations are of interest to understand the nature of the response. Many CAD systems record the location of the call as well as its coordinates using a grid or district. Ideally, a geographic information system (GIS) is used to record latitude longitude coordinates. However, this has not been implemented in all CAD systems. Using a grid or district instead can capture approximate locations and is useful for aggregating call locations; but if this grid is coarse (e.g., 5 miles $\times$ 5 miles), it may limit the approaches used to design next-generation EMS systems.

PATIENT SURVIVABILITY PERFORMANCE MEASURES

This section summarizes traditional performance measures used by operations research models to evaluate EMS systems, and proposes new performance measures for next-generation EMS systems. The performance measures in this section focus on issues involving the response to emergency medical patients. Since the response time interval is closely tied with patient survival—the ultimate goal in EMS systems—all performance measures can be modeled as a *utility function associated with the response time*. The implications and challenges for using these performance measures are discussed as they relate to EMS operations and management of CAD systems. First, standards set by EMS systems are introduced, since these standards are the basis for operations research modeling efforts.

Performance measures used by operations research models often mimic the standards set by EMS systems, which almost always measure the response time. According to Fitch [3], the most common standards for an EMS system are as follows:

1. Respond to 90% of urban (Priority 1) calls in less than 9 min,
2. Respond to 90% of rural (Priority 1) calls within 15 min,
3. Respond to 90% of wilderness (Priority 1) calls within 30 min.

According to the first standard that requires that 90% of calls are responded to in no more than 8 min 59 s, a call is considered to be covered if the response time is 8 min and 59 s, whereas a response time of 1 s longer (9 min) is not considered to be covered. However, according to the medical literature, no difference in patient outcomes would be expected between these two scenarios. Formulated as a utility, the utility is 1 when the response time t_R (measured in minutes) is less than 9 min (i.e., $u(T_R) = 1, t_R < 9$) and 0 when the response time is 9 min or longer (i.e., $u(T_R) = 0, t_R \geq 9$). For this reason, these binary standards are not optimal for measuring patient survivability rates.

As noted elsewhere, little or no guidance is given as to how to interpret these standards based on time of day or location [11,6]. For example, it is not clear if it is acceptable to have, say, 80% of Priority 1 calls responded to in less than 9 min during the day and, say, 95% of Priority 1 calls responded to in less than 9 min overnight, so long as 90% of Priority 1 calls are responded to in less than 9 min across all times. Moreover, little guidance is given as to how to define the response time interval. The response time interval typically measures the time until an ALS ambulance responds. In such cases, the response time interval does not indicate if there were earlier responders (such as BLS ambulances or fire engines) or the time when patient care was initiated, which makes it difficult to assess the true EMS system response interval.

Although patient survivability (measured as patients surviving until hospital discharge) is the ultimate goal, performance measures set by EMS systems and used by operations research models quantify proxies for patient survivability. Until recently, this was a necessary measure, since having access to large amounts of CAD data is a relatively new phenomenon that has allowed for new and more detailed modeling

efforts. The proxies for patient survivability generally have one of the following forms [9]:

- maximize the proportion of the (geographic) area that can be responded to in a fixed time frame,
- maximize the proportion of the population that can be responded to in a fixed time frame,
- maximize the expected number of calls that can be responded to in a fixed time frame.

Medical research has shed some light on effective performance measures for EMS systems. Next-generation EMS systems and operations research models for such systems must be proactive in using this information to identify operational practices and strategies that translate to better patient outcomes.

Maximizing patient survival rates can be accomplished by using medical models that relate EMS response time intervals to patient survivability in cardiac arrest patients. Four of these models are explored here, which result in four patient survivability functions, each of which can be expressed as a utility. The excellent paper by Erkut *et al.* [6] explores these issues in greater detail, and some of the analysis is reproduced here. All survival models are interpreted to assume that EMS personnel do not witness a cardiac arrest. Let S denote the patient survivability, which is assumed to represent a probability (i.e., $0 \leq S \leq 1$) or utility. All times in the following models are measured relative to the time of collapse and are measured in minutes. Note that survival in each of the following equations can be thought of more generally as a utility associated with the response time.

The first survivability model is based on a study by Larsen *et al.* [13], which performed multiple linear regression for King County in Washington. Their model for patient survivability is given by

$$\begin{aligned} S(t_{\text{CPR}}, t_{\text{AED}}, t_{\text{ALS}}) \\ = \max\{0.67 - 0.023t_{\text{CPR}} \\ - 0.011t_{\text{AED}} - 0.021t_{\text{ALS}}, 0\}, \quad (1) \end{aligned}$$

where t_{CPR} denotes the time from collapse until CPR is performed on the patient, t_{AED}

denotes the time of defibrillation, and t_{ALS} denotes the time when ALS treatment is provided.

The second survivability model is based on a study by Valenzuela *et al.* [14]. It performs logistic regression to obtain an expression for patient survivability, and it is based on data obtained from Tucson, Arizona and King County, Washington. They estimate patient survivability as

$$\begin{aligned} S(t_{\text{CPR}}, t_{\text{AED}}) \\ = (1 + e^{-0.260 + 0.106t_{\text{CPR}} + 0.139t_{\text{AED}}})^{-1}. \quad (2) \end{aligned}$$

The third survivability model by Waelwijn *et al.* [15] performs logistic regression to obtain an expression for patient survivability from the perspective of the bystander that is based on data from Amsterdam, Netherlands. In this expression, it is assumed that the cardiac arrest is not witnessed by EMS providers. Patient survivability is expressed as

$$\begin{aligned} S(t_{\text{CPR}}, t_{\text{R}}) \\ = (1 + e^{0.04 + 0.3t_{\text{CPR}} + 0.14(t_{\text{R}} - t_{\text{CPR}})})^{-1}. \quad (3) \end{aligned}$$

The fourth survivability model by De Maio *et al.* [16] uses stepwise logistic regression to estimate patient survivability using data from Ontario, Canada, with

$$S(t_{\text{R}}) = (1 + e^{0.679 + 0.262t_{\text{R}}})^{-1}. \quad (4)$$

Figure 1 illustrates the utility for the four patient survival models considered as well as the standard requiring ALS care in less than 9 min under the following assumptions:

- It takes exactly 1 min for a call to EMS to be made and an ambulance to be dispatched after the patient collapses.
- CPR is performed by a paramedic or EMT immediately upon arrival, and CPR is not performed earlier by a bystander, resulting in $t_{\text{CPR}} = t_{\text{R}}$.
- An AED is used 1 min after arrival, resulting in $t_{\text{AED}} = t_{\text{R}} + 1$.
- ALS is provided 2 min after arrival ($t_{\text{ALS}} = t_{\text{R}} + 2$).

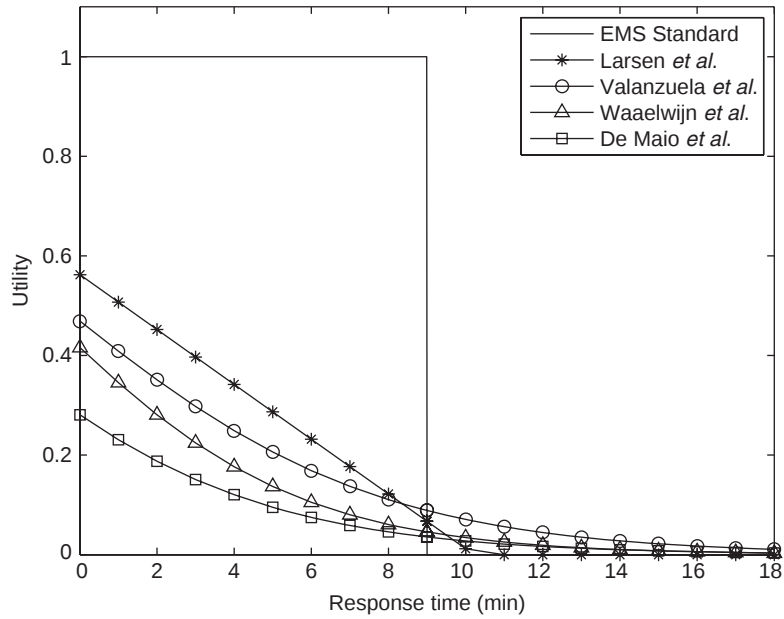


Figure 1. Patient survival utilities as a function of response time.

Figure 1 shows the utility curves for the four patient survivability functions and the 9-min EMS standard. Figure 1 indicates that patient survival is significantly less than 1, even for response times of 0. At 9 min, patient survival is less than 10% for all the models considered, and although patient survivability beyond 9 min is positive for all the models, cardiac arrest patients are extremely unlikely to survive.

The response time interval is an important predictor for short-term patient survival. Other factors that have been identified by the medical literature that affect short-term patient survival include the age of the patient, whether a bystander is present at the time of collapse, and whether bystanders attempted CPR and defibrillation before the first responder arrives.

Although Priority 1 patients are equally likely to survive whether an ALS or a BLS medical unit responds to the call, it is important for personnel to practice medical skills in order to maintain their efficacy in treating emergency medical patients [17]. One way to maximize patient survival rates, at least in the short term, is to employ as many medical units as possible to reduce response times.

This approach would, however, result in each medical unit responding to fewer calls, and hence, each paramedic and EMT would treat few patients, which may eventually erode their proficiency to treat patients. Therefore, there are drawbacks to simply increasing the number of medical units to meet performance standards. EMS systems should ideally consider the appropriateness of the response so that ALS providers have the most opportunities to provide patient care (i.e., dispatching ALS to Priority 1 calls and BLS to Priority 2 and 3 calls).

OPERATIONS RESEARCH OPPORTUNITIES

Designing next-generation EMS systems is truly a systems issue. This section overviews several broad areas in next-generation EMS design that could benefit from operations research methodologies. All potential application areas directly or indirectly affect patient survivability. In this article, patient survivability is motivated by the problem of locating medical units at stations. However, there are a number of other EMS problems that have the potential to improve patient survival rates.

Although numerous models have been analyzed for locating ambulances at a set of stations, nearly all models focus on meeting EMS standards and proxies for patient survivability, rather than explicitly modeling patient survivability [7–9]. Recently, McLay [18], Erkut *et al.* [6], and McLay and Mayorga [19] examined how to locate medical units when focusing on patient survivability.

McLay [18] evaluates the impact of the first responder on ambulance locations, a stepping stone to evaluating patient survival. In particular, McLay [18] investigates how to locate both ALS and BLS medical units such that response time is determined by the first responder, rather than the ALS medical unit, as traditionally done in EMS systems. McLay [18] examines the effect of dispatching policies for ALS and BLS medical units to Priority 1, 2, and 3 calls to illustrate how to use and locate medical units.

Erkut *et al.* [6] illustrate the importance of explicitly linking ambulance location to patient survivability, by providing the following example. Two demand locations A and B are located 18 min apart, with three potential stations located at A, B, and X, a station exactly 9 min away from both A and B. The demand at A is 10, and the demand at B is 1. Exactly one ambulance is available. Traditional operations research models that seek to maximize the number of calls responded to in less than 9 min locate the ambulance at X, which covers the entire demand of 11. If the probability of survival is given by $\exp(-t_R)$, where t_R denotes the response time (measured in minutes), then the expected number of survivors when an ambulance is located as X is $11 \times \exp(-9) = 0.001358$, whereas the expected number of survivors when the ambulance is located at A is 10, which differs by a factor of 700 [6]. This indicates that traditional covering location models may result in arbitrarily bad patient survival rates and motivates modeling efforts that explicitly focus on patient outcomes.

Erkut *et al.* [6] apply patient survivability models reported in the medical domain (see Fig. 1) to ambulance location models using discrete optimization models (linear and nonlinear) and Monte Carlo simulation. The probability of patient survival is a function of

response time, the time to initiate CPR, pre-travel delays, and the time to defibrillation. They demonstrate the benefit of using patient survival location models using data extracted from Edmonton, Canada. They explore how patient survival affects the location of ambulance by using survivability models as the objective function in set covering and maximal coverage models that consider one type of medical unit (ALS ambulances).

McLay and Mayorga [19] propose a methodology to evaluate a class of performance measures according to their resulting patient survival rates. They focus on response time thresholds (RTTs) and the number (or fraction) of calls that can be reached in a fixed time frame (the 9-min EMS standard is a 9-min RTT). They find that RTTs can act as proxies for patient survivability performance measures. For an example with real-world data, they find that 7- and 8-min RTTs always locate ambulances such that patient survival is also optimal. Similarly 9- and 10-min RTTs result in more equitable patient outcomes, with fewer disparities between rural and urban patient survival rates. These results are counterintuitive, since medical experts often recommend an EMS response time interval of less than 4 min. An example from this paper is summarized in section titled “Illustrative Example.”

Patient survivability largely depends on issues that EMS departments have traditionally not focused on. For example, medical literature has recognized that patient survivability depends on whether bystanders initiated CPR or defibrillation. Cities that have had large-scale CPR campaigns have benefited from a large proportion of cardiac arrest patients receiving bystander CPR prior to the EMS response and, as a result, have high patient survival rates [20]. We are not aware of any operations research modeling efforts that perform a systematic analysis of the effect of CPR campaigns on EMS and patient outcomes. Such a systems approach to maximize patient survivability may identify cost-effective ways to improve patient outcomes. A number of methodologies could be beneficial, including applied probability models and simulation.

In many cities and counties, emergency medical treatment of patients extends beyond ambulances and paramedics. Increasingly, campaigns that train people to apply CPR and the deployment of AEDs in public places results in care being applied by cardiac arrest bystanders. These campaigns have resulted in significant improvement in cardiac arrest patient survival rates [4,5]. Understanding where to locate AEDs as well as identifying how many should be purchased has the potential to save lives. For example, locating AEDs in Chicago airports resulted in a 75% survival rate for ventricular fibrillation cardiac arrests (9 of 12 patients survived) as compared to a 2% cardiac arrest survival rate in the city of Chicago [20]. A relative 40–45% increase in patient survival rate has been observed after large-scale AED deployments [20].

Another avenue of research using operations research is to design better EMS standards. For example, it has been noted that response time intervals should be less than 5 min and defibrillation should occur in less than 6 min from the time of collapse [3]. McLay and Mayorga [19] challenge this assumption by applying discrete optimization models, but more analysis is needed. Applying operations research models (using simulation, agent-based modeling, and decision analysis, for example) to analyze and create new standards has the potential to improve patient survivability. For example, performance measures used by operations research models almost always depend on a response time interval standard (such as 9 min), but a little guidance or explanation is used to explain why the selected response time interval is used [11]. Investigating how new standards correlate with patient outcomes can change how EMS systems are operated. Emergency medical research has shed light on some of these EMS features, such as illustrating that ALS patient care does not result in higher patient survival rates than BLS patient care [2]. Operations research can link state-of-the-art medical knowledge to practical guidelines for responding to and treating patients. Multiple-objective decision analysis and optimization can be used to design EMS

systems that perform well across multiple criteria (such as coverage and fairness). Chanta *et al.* [21] take an important first step in investigating the issue of equity using biobjective covering location models.

Response time is almost always used to measure EMS performance. Recall that the response time interval quantifies the time between when a medical unit is dispatched and when it arrives at the scene. It has been noted that the response time interval is often measured in different ways and does not reflect the true impact on patient survival. Budge *et al.* [22] indicate that the response time interval is usually defined to measure the sum of the travel time and the pretravel delay (the time between when the ambulance is dispatched and when the ambulance departs the station). The response time interval measures the response from the EMS point of view rather than the patient point of view since it is defined relative to dispatch rather than patient collapse. Exploring uncertainties in the components of response time from the patient point of view is critical for accurately predicting patient survival. Possible methodologies to be used here include data mining, stochastic processes, and simulation.

Nearly all EMS systems locate ambulances at rescue or fire stations when the ambulances are not busy treating patients. In order to improve patient survival rates, stations must be located in areas with high population density. Therefore, cities and counties zone the land for new stations before new development occurs. Although many location models have been developed, few consider long-term issues in building rescue stations based on changing demographics, anticipated zoning, and forecasted call volumes, which requires systematic planning and analysis. Not only are stations built under uncertainty with respect to forecasts in call volume, they are built under uncertainty with respect to EMS utilization and operations. For example, future EMS systems may rely even more heavily on paid personnel (rather than volunteers), may use BLS medical units rather than ALS medical units, and are not likely to locate ambulances at all stations at once. Exploring these issues

(using forecasting and optimization models, for example) can determine the best location for a future station across a wide spectrum of operating conditions.

Optimizing dispatching protocols—in addition to locating medical units—have potentially large effects on patient outcomes. To maximize patient survival rates, it is necessary to understand when to dispatch the closest medical unit to a call and when to dispatch a farther medical unit to a call. However, little attention has been given to identifying optimal dispatching policies. Carter *et al.* [23] use a queueing optimization model to determine the locations of two EMS response areas (i.e., beats) to balance the workload between two medical units. Each medical unit is assumed to always respond to calls within its response area, if the medical unit is available. If both medical units are busy, calls are served by medical units outside of the area. Carter *et al.* [23] show that it is not always best to dispatch the closest medical unit. Jarvis [24] introduces a Markov decision process model for the hypercube queueing model for determining optimal dispatching policies for a single type of medical unit, and it is illustrated with an example that minimizes the average distance traveled when responding to a call. Additional dispatching policies are needed to dispatch multiple types of medical units (including ambulances, nontransport units, fire engines, and police cars) to multiple patient priorities. Possible methodologies for this application include simulation optimization, decision support systems, and approximate dynamic programming. Henderson [25] proposes using system status management and approximate dynamic programming for relocating ambulances.

Providing guidelines for scheduling paid personnel given uncertainty in volunteer schedules has the potential to more efficiently use EMS resources. Many EMS stations supplement volunteer personnel with paid personnel. There are many challenges for scheduling in a mixed personnel system. For example, volunteers may have scheduling commitments that they cancel more frequently than paid personnel. When a volunteer cancels, usually very little notice

is given to the EMS management. The uncertainties in volunteer schedules can be addressed by planning for the worst-case scenario (i.e., not relying on volunteers at all), therefore, assuming that only paid personnel are available. This approach is suboptimal in two ways. First, it is more costly, since it is expensive to fully utilize paid personnel and overuse medical equipment. Second, with more personnel available, paramedics and EMTs respond to fewer emergency medical 911 calls and utilize their medical skills less frequently, which degrades their level of competence over time [17]. In addition, volunteers have lower levels of job satisfaction if they treat fewer emergency medical patients and transport more patients to the hospital, and volunteers are less likely to volunteer in the future if their skills are not utilized [26]. Sampson [27] outlines several key differences between volunteer and traditional (i.e., paid) scheduling problems and notes that cost is often not the bottom line when utilizing volunteers.

Suburban and rural EMS departments are not simply smaller versions of urban EMS departments, but rather they have unique characteristics that differentiate them from urban EMS departments. Suburban and rural EMS departments are characteristic of many EMS systems throughout the nation and illustrate issues that EMS systems frequently face on a daily basis. In particular, they typically operate in areas with low population densities, are not near hospitals, and rely on volunteer staff rather than career staff. Most notably, they typically leverage several types of resources to respond to emergency medical 911 calls. For example, personnel may include volunteer and paid staff, medical units may include ALS ambulances, BLS ambulances, nontransport QRVs, and fire engines. This can be contrasted to urban EMS systems, which may only have paid staff and one type of ambulance. Therefore, suburban and rural EMS management is extremely complex, even for small EMS systems, since determining the optimal portfolio of resources may be counterintuitive. Yet, these types of EMS systems typically lack the resources to determine how to optimally allocate and use their resources.

As operations research methodologies are applied to a broad set of EMS problems and operations, new EMS features and operations research models may become apparent, leading to new application areas within next-generation EMS system design.

ILLUSTRATIVE EXAMPLE

Models that improve patient survivability ultimately focus on the response time interval, and there are many factors that affect how the response time interval is measured and how medical units are used to respond to calls. This section explores several of these factors using real-world CAD data in order to illustrate several of the issues that influence the design of next-generation EMS systems and summarizes results from McLay and Mayorga [19] using this data to locate ambulances according to patient survival rates.

This section illustrates some of the challenges in designing effective and robust next-generation EMS systems using data extracted from Hanover County, a semisuburban, semirural county near Richmond, Virginia with a population of approximately 100,000 and an area of 471 square miles. It is described as approximately 70% rural with areas of suburbs. It uses a mixture of ALS and BLS ambulances, ALS QRVs, and fire engines for treating emergency medical patients. It employs paramedics (paid personnel) and EMTs (paid and volunteer personnel). AEDs are available on all types of medical units. Hanover County has 12 fire stations and 4 rescue stations at which they can locate ambulances. It is partitioned into 12 districts that correspond to the geographic areas surrounding each of the 12 fire stations. Figure 2 shows a map of Hanover County and its fire and rescue stations.

The CAD data set contains 9710 calls for service in a 1-year period, and it includes information regarding the location, response time, and service times for all calls. The dispatcher manually enters this information, and hence, there are occasionally problems with the data. For example, in 145 of the 8232 calls in which an ambulance responds (1.8%), the time when the ambulance leaves

its station is not recorded. In 60 of the 5764 calls in which an ambulance transports a patient to the hospital (1.0%), the time when the ambulance leaves for the hospital is not recorded. In addition, the response times vary from 0 to 604 min. Both extremes reflect scenarios that need to be taken into account when performing a data analysis. To understand these response times, note that the dispatcher enters the time into the CAD system by hitting a button that records the current time, not by physically recording the current time. Many of the scenarios with response times of 0 correspond to the cases when a second ambulance voluntarily responds to a call (before dispatched), and the dispatcher later enters this response into the CAD system, resulting in identical dispatch and responding times. The scenarios with extremely large response times correspond to the cases when the dispatcher forgets to enter the responding times into the CAD system. When this omission is noted, the dispatcher corrects the missing value by entering the current time, which may be hours after the ambulance responded. All of these issues indicate that there are many real issues in interpreting CAD data, although these issues are only present in a small proportion of the CAD data entries. Working with the EMS department is necessary to truly understand what phenomena the CAD data reflect and operational issues that the EMS department faces.

The CAD data indicate that 54, 22, and 24% of calls are classified as Priority 1, 2, and 3, respectively. These priorities are determined by call characteristics, medical protocols set by the medical director, and the CAD system. Therefore, the proportions of Priority 1, 2, and 3 calls can change abruptly when medical protocols or the CAD system change. For example, prior to changes in the Hanover County's medical protocols and the CAD system, 63, 26, and 11% of calls were classified as Priority 1, 2, and 3, respectively. The call priorities are also location-dependent, since they reflect the demographics and zoning in the vicinities of the stations. On the basis of the 12 fire districts, the proportion of calls originating in each district classified as Priority 1 varied from 49 to 56% of all calls.

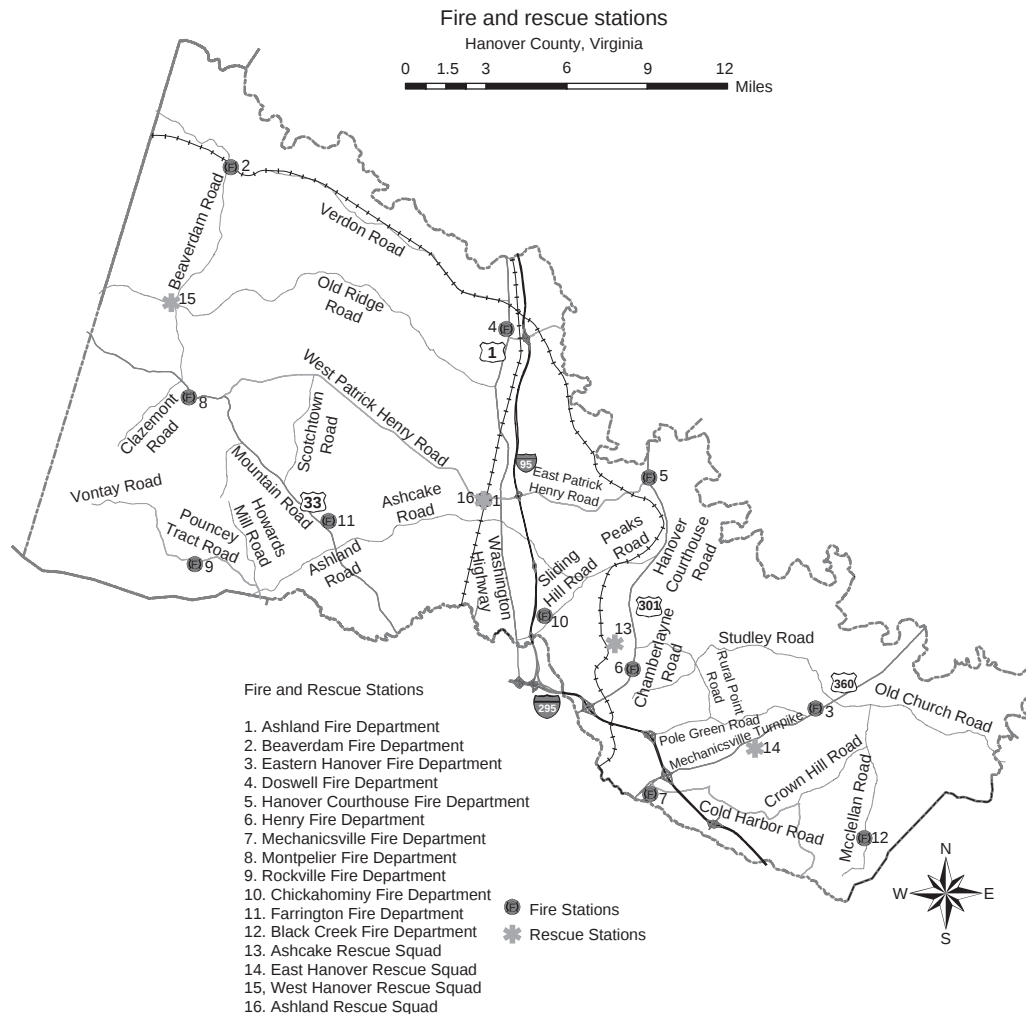


Figure 2. Fire and rescue stations in Hanover County, Virginia.

In a significant number of calls (11%), an ambulance is dispatched to a call, but the call is canceled before the ambulance arrives. This occurs for a number of reasons, and it occurs more frequently with high-priority calls, when multiple ambulances may be dispatched to a single call. A call is canceled for 14.2, 8.8, and 5.9% of Priority 1, 2, and 3 calls, respectively.

Improving response time to cardiac arrest patients can improve patient survival rates. It is difficult for a dispatcher to correctly diagnose a cardiac arrest over the phone, and cardiac arrest calls could be classified

as “Difficulty Breathing,” “Chest Pains,” or “Heart Problems” by the dispatcher. Not all calls with these classifications are truly cardiac arrest calls. Table 1 summarizes these call classifications by priority. Note that 24.5% of Priority 1 calls are one of these three call classifications, and hence, a large number of Priority 1 calls are potentially cardiac arrest calls. In addition, a small proportion of cardiac arrest calls could be misclassified as abdominal pain, for example, although these false negatives are difficult to identify from CAD and medical data.

Table 1. Number of EMS Calls by Classification and Priority

Classification	Priority 1	Priority 2	Priority 3
Difficulty breathing	697	160	9
Chest pains	360	159	14
Heart problems	100	14	8
All calls	4722	1925	2043

Table 2. Response Times According to District and Distance

District	All Responses			Responses ≤ 4 miles			Responses > 4 miles		
	$\bar{t}_R$ (min)	s_R	p_R	$\bar{t}_R$ (min)	s_R	p_R	$\bar{t}_R$ (min)	s_R	p_R
1	8.8	5	0.57	6.9	3.7	0.78	11.3	5.3	0.31
2	17.1	6.6	0.07	12.8	4.9	0.19	17.7	6.6	0.05
3	11.5	6.7	0.40	8.3	4.6	0.62	16.1	6.6	0.10
4	12.8	7.5	0.34	8.1	4.8	0.66	15.7	7.4	0.14
5	12.9	4.3	0.12	10.5	2.6	0.32	13.3	4.4	0.09
6	9.2	5	0.56	7.0	4.1	0.69	12.2	5.3	0.26
7	9.1	4.6	0.58	8.2	5.4	0.68	10.6	5.3	0.44
8	12.2	7.5	0.37	8.1	5.5	0.67	14.5	7.4	0.20
9	18.9	6.4	0.04	NA	NA	NA	19.5	6.5	0.03
10	8.7	5.1	0.62	8.0	4.9	0.70	10.6	5.1	0.40
11	14.2	5.5	0.11	10.6	5.3	0.38	14.2	5.4	0.10
12	15.3	4.9	0.07	11.4	2.7	0.19	16.6	4.8	0.03
Overall	10.2	5.8	0.49	7.8	4.1	0.69	12.7	6.2	0.28

The response time depends on the locations where the calls originate as well as the distances traveled. Table 2 shows the average response times ($\bar{t}_R$), the standard deviation of response times (s_R), and the proportion of calls responded to in less than 9 min (p_R) for all calls, calls when the ambulance travels 4 miles or less, and calls when the ambulance travels more than 4 miles for each of the 12 districts. The average response time is 10.2 min, which decreases to 7.8 min if an ambulance travels 4 miles or less and increases to 12.7 min if the ambulance travels more than 4 miles. Response times for all instances suggest that even when ambulances travel short distances, cardiac arrest patients are not likely to survive, which motivates using more comprehensive models and methodologies to address patient survivability.

The data set indicates that there is double response on 11% of the calls. In these calls, two ambulances are sent to the call. The first ambulance stabilizes the patient, and the second ambulance provides patient care and transports the patient to the hospital, if needed. The double response rate is 13,

7.4, and 5.7% of all Priority 1, 2, and 3 calls, respectively. The double response rate is higher on the weekends (12% of all calls) versus weekdays (10% of all calls). Since the weekends almost entirely rely on volunteer EMTs, there is likely a difference in how volunteer resources are used as compared to paid personnel. The double response rate also varies with the location where the call originates. Table 3 summarizes the number and proportion (p_D) of calls in each of the 12 fire districts for which there is double response. The most rural districts (2, 8, 9, 11) have high levels of double response as compared to suburban districts (1, 4, 7, 12).

The CAD data set indicates that 71% of calls result in a patient transported to the hospital. Understanding how an EMS system transports patients to hospitals is critical for understanding how the EMS system utilizes its resources, since patient transport time represents the majority of the ambulance service time. Table 3 summarizes the proportion (p_H) of calls in each of the 12 fire districts in which patients are transported to the hospital. In addition, Table 3

Table 3. Double Response and Hospital Call Characteristics According to District

District	Double Response	Hospital Calls		
	p_D	$\bar{t}_T$ (min)	s_T (min)	p_H
1	0.06	94	31	0.67
2	0.14	131	26	0.80
3	0.11	83	26	0.73
4	0.05	103	38	0.72
5	0.12	91	24	0.81
6	0.13	81	30	0.71
7	0.10	74	25	0.73
8	0.14	116	32	0.72
9	0.49	120	33	0.78
10	0.12	85	25	0.65
11	0.17	109	27	0.70
12	0.09	95	30	0.75
Overall	0.11	88	32	0.71

summarizes the average and standard deviation of the turnaround time (measured in minutes) when patients are transported to a hospital ($\bar{t}_T$ and s_T , respectively). Note that the average turnaround time for calls that transport patients to the hospital varies from 74 to 131 min, since the east side of Hanover County is closer to several hospitals than the west side. This motivates models for optimal dispatching policies, since an ambulance may be busy for more than 2 h from the time it is dispatched to a call.

CAD systems link the EMS system and its performance with operations research modeling efforts. As this section demonstrates, it is necessary to work closely with EMS personnel to determine how to interpret the data in the CAD system and to identify the most salient features in the EMS system. Moreover, evaluating EMS systems from the patient point of view is extremely challenging due to the lack of an accurate description of the patient experience.

McLay and Mayorga [19] use this data set to predict the expected additional number of lives lost if ambulances are located according to RTTs as opposed to patient survival rates. They evaluate the impact of T -minute RTTs, $T = 4, 5, \dots, 10$. Table 4 summarizes the expected number of lives lost over a 1-year period when seven ALS ambulances are located according to T -minute RTTs, $T = 4, 5, \dots, 10$, relative to the base case of 418 lives saved when ambulances are

Table 4. Expected Number of Lives Lost when Locating Seven Ambulances According to Response Time Thresholds

T -Minute RTT	Expected Number of Lives Lost	Ambulance Locations
4	1.49	1 , 1, 4, 6, 7, 7, 10
5	1.49	1 , 1, 4, 6, 7, 7, 10
6	0.81	1, 4, 6, 7, 7, 10, 14
7	0	1, 4, 6, 7, 7, 8, 10
8	0	1, 4, 6, 7, 7, 8, 10
9	2.10	1, 4, 6, 7, 8, 10, 14
10	2.10	1, 4, 6, 7, 8, 10, 14

optimally located according to patient survival. The resulting ambulance locations are listed as stations (refer to Fig. 2), with stations in bold in Table 4 denoting new stations used as compared to when ambulances are optimally located according to patient survival. Although the relative improvement in the number of lives saved is small (no more than 2.1 lives, on average), these lives could be saved at no additional cost, and it highlights the importance of appropriate performance measures in EMS system design and evaluation.

CONCLUSIONS

Operations research practitioners have the unique opportunity to make a difference in EMS system design. New directions in EMS systems need not merely be makeshift solutions for mending complex problems; they can be the result of modeling, analysis, and planning. This article summarizes how operations research can play a role in next-generation EMS system design that focuses on patient outcomes. Note that all EMS systems operate differently, and the issues raised in this article should be modified before being applied to other EMS systems.

Operations research has made an impact in EMS systems, particularly in the area of

locating medical units. There are many other areas within EMS system design and operation that could benefit from the application of operations research methodologies. The issues discussed represent but the tip of the iceberg. By using operations research methodologies to design and operate EMS systems, which save more patient lives, can drastically affect our nation's health and well-being.

Acknowledgments

It is a pleasure to acknowledge Dr. Joseph P. Ornato, M.D., Chair of the Department of Emergency Medicine at Virginia Commonwealth University and Medical Director for the Richmond Ambulance Authority, as well as Battalion Chief Henri Moore, Jr., Hanover Fire and EMS Chief Fred C. Crosby, II, Division Chief Eddie Buchanan, and Mr. Lawrence Roakes of the Hanover County Fire and EMS Department in Hanover County, Virginia, for the knowledge, experience, and data they provided to support this research effort as well as their feedback on a draft of this article. The data analysis was performed by Alisha Purohit for an Honors Summer Undergraduate Research Project program at Virginia Commonwealth University.

REFERENCES

- Wik L, Boye Hansen T, Fylling F, *et al.* Delaying defibrillation to give basic cardiopulmonary resuscitation to patients with out-of-hospital ventricular fibrillation. *J Am Med Assoc* 2003;289(11):1389–1395.
- Stiell IG, Wells GA, Field BJ. Improved out-of-hospital cardiac arrest survival through the inexpensive optimization of an existing defibrillation program: Opals study phase ii. Ontario prehospital advanced life support. *N Engl J Med* 1999;351(7):647–656.
- Fitch J. Response times: Myths, measurement and management. *J Emerg Med Serv* 2005;9:46–56.
- Ornato JP, McBurnie MA, Nichol G, *et al.* The public access defibrillation (PAD) trial: study design and rationale. *Resuscitation* 2003;56:135–147.
- Hallstrom AP, Ornato JP, Weisfeldt M, *et al.* Public-access defibrillation and survival after out-of-hospital cardiac arrest. *N Engl J Med* 2004;351(7):1–7.
- Erkut E, Ingolfsson A, Erdogan G. Ambulance location for maximum survival. *Nav Res Logist* 2008;55(1):42–55.
- Swersey AJ. Operations research and the public sector. Volume 6, Handbooks in operations research and management science. Chapter The Deployment of Police, Fire, and Emergency Medical Units. Amsterdam: Elsevier, North-Holland; 1994, pp. 151–200.
- Brotcorne L, Laporte G, Semet F. Ambulance location and relocation models. *Euro J Oper Res* 2003;147:451–463.
- Goldberg JB. Operations research models for the deployment of emergency service vehicles. *EMS Manag J* 2004;1(1):20–39.
- Green LV, Kolesar PJ. Improving emergency responsiveness with management science. *Manag Sci* 2004;50:1001–1014.
- Henderson SG, Mason AJ. Ambulance service planning: Simulation and data visualization. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. Operations research and health care: a handbook of methods and applications. Boston (MA): Kluwer Academic; 2004. pp. 77–102.
- Ornato JP, Doctor ML, Harbour LF, *et al.* Links synchronization of timepieces to the atomic clock in an urban emergency medical services system. *Ann Emerg Med* 1998;31(4):483–487.
- Larsen MP, Eisenberg MS, Cummins RO, *et al.* Predicting survival from out-of-hospital cardiac arrest—A graphic model. *Ann Emerg Med* 1993;22:1652–1658.
- Valenzuela TD, Roe DJ, Cretin S, *et al.* Estimating effectiveness of cardiac arrest intervention—a logistic regression survival model. *Circulation* 1997;96:3308–3313.
- Waelwijn RA, de Vos R, Tijssen JGP, *et al.* Survival models for out-of-hospital cardiopulmonary resuscitation from the perspectives of the bystander, the first responder, and the paramedic. *Resuscitation* 2001;51:113–122.
- De Maio VJ, Stiell IG, Wells GA, *et al.* Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Ann Emerg Med* 2003;42:242–250.
- Ornato JP. Personal interview, 20 March 2007.
- McLay LA. A maximum expected covering location model with two types of servers. *IIE Trans* 2009;41(8):730–741.

19. McLay LA, Mayorga ME. Evaluating emergency medical service performance measures. Technical report, Virginia Commonwealth University, Richmond, Virginia, 2009.
20. Caffrey S. Feasibility of public access to defibrillation. *Curr Opin Crit Care* 2002;8:195–198.
21. Chanta S, Mayorga M, McLay L, *et al.* A bi-objective covering location model for ems systems. In: Proceedings of the 2009 Industrial Engineering Research Conference, Miami, FL, 2009.
22. Budge S, Ingolfsson A, Erkut E. Optimal ambulance location with random delays and travel times. *Health Care Manag Sci* 2008;11:262–274.
23. Carter GM, Chaiken JM, Ignall E. Response areas for two emergency units. *Oper Res* 1972;20(3):571–594.
24. Jarvis JP. Optimization in stochastic systems with distinguishable servers. Technical report TR-19-75, MIT, Cambridge (MA), 1975.
25. Henderson SG. Operations research tools for addressing current challenges in emergency medical services. In: Cochran JJ, editor. *Wiley encyclopedia of operations research and management science*. Hoboken (NJ): Wiley; 2011. In press.
26. Studnek J, Fernandez AR. Non-urgent is no fun. *J Emerg Med Serv* 2007;32(10):38.
27. Sampson SE. Optimization of volunteer labor assignments. *J Oper Manag* 2006;24: 363–377.

ESTIMATING FAILURE RATES AND HAZARD FUNCTIONS

MASSIMILIANO GIORGIO
Department of Aerospace and
Mechanical Engineering,
Second University of Naples,
Aversa, Italy

GIANPAOLO PULCINI
Istituto Motori, National
Research Council (CNR),
Rome, Italy

The *failure rate* (or *hazard function*) $h(t)$ for a continuous lifetime distribution is defined as [1]:

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \frac{f(t)}{R(t)}, \quad t \geq 0, \end{aligned} \quad (1)$$

where $f(t)$ and $R(t)$ are the probability density function and the reliability function (also called *survival probability*), respectively, of the continuous nonnegative random variable T denoting the time-to-failure of a nonrepairable unit. The failure rate $h(t)$ specifies the instantaneous rate of failure at time t , given that the unit survives until t . Hence, the product $h(t) \cdot \Delta t$ is approximately the probability of failure in the time interval $[t, t + \Delta t)$, given survival till t . That is, in the short time Δt from t to $t + \Delta t$, a proportion $h(t) \cdot \Delta t$ of the population that reached age t is expected to fail [2]. Since $f(t) = -dR(t)/dt$, then $h(t) = -d \ln R(t)/dt$.

The *cumulative failure rate* (or *cumulative hazard function*) is defined as

$$H(t) = \int_0^t h(x) dx, \quad (2)$$

and is related to the reliability function $R(t)$ by the formula

$$H(t) = -\ln R(t). \quad (3)$$

It can be noted that, since $\lim_{t \rightarrow \infty} R(t) = 0$, the $\lim_{t \rightarrow \infty} H(t) = \infty$. Thus, the failure rate $h(t)$ for a continuous lifetime distribution possesses the properties

$$h(t) \geq 0 \quad \text{and} \quad \int_0^\infty h(x) dx = \infty.$$

In particular, note that being $\int_0^\infty h(x) dx = \infty$, $h(t)$ is not a probability function (i.e., only a pointwise probabilistic interpretation can be given for $h(t) \cdot \Delta t$) and no physical interpretation can be given to $H(t)$.

As the definition (Eq. 1) involves the concept of probability, the measurement of $h(t)$ and/or of $H(t)$ requires probabilistic and statistical methods. A number of classical nonparametric and parametric procedures for estimating the failure rate and the cumulative failure rate are illustrated here.

LIFETIME DATA COLLECTION

Any estimation procedure, or more generally, any method of lifetime data analysis, is strictly related to the way the data are collected. In particular, the exact value of the time-to-failure is only seldom known for all the units on test because to collect such a *complete* sample, it would be necessary to continuously monitor the units all along their life. On the contrary, very frequently, lifetime is known exactly only for a portion of the units on test, and incomplete information is available about the failure time of the other units (*censored* sample). Censoring arises in various ways and many different kinds of censored sampling are considered in literature [1,3,4]. The case of right censored observations is specifically addressed in the following section.

In general, a sample is called a *singly censored sample* if all the survivors have the same censoring time, whereas it is called a *multiply censored sample* if survivors have different censoring times. Random and staggered censorings usually give rise to multiply censored samples.

Only standard methods for censored data analysis are presented in this article. The use of these methods makes the assumption that censoring mechanisms are noninformative (or *independent*), that is, the probability of censoring at time t does not depend on the prognosis for failure at time t (i.e., the censoring time does not give information about the failure time distribution) [5].

NONPARAMETRIC PROCEDURES

Complete and Singly Right Censored Samples

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a complete sample of n data from a single distribution function, and let $\mathbf{x} = \{x_{(1)}, \dots, x_{(s)}\}$ indicate the ordered set of the s distinct values of the set $\mathbf{t}$ ($s \leq n$). We have that $s < n$ in presence of ties. From the relationship (Eq. 3) between $H(t)$ and $R(t)$, a natural nonparametric estimator of the cumulative hazard function $H(t)$ at the time t is given by

$$\tilde{H}(t) = \begin{cases} 0, & \text{for } 0 \leq t < x_{(1)} \\ -\ln \left(1 - \sum_{j=1}^i \frac{d_j}{n} \right), & \text{for } x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, 2, \dots, s-1, \end{cases} \quad (4)$$

where $x_{(0)} \equiv 0$, d_j is the number of units failed at $x_{(j)}$ and $n - \sum_{j=1}^i d_j$ is the number of units which survive time $x_{(j)}$. $\tilde{H}(t)$ is a staircase function, that is it is flat between successive failure times. Of course, the true cumulative hazard function of a large population of identical units operating under identical working conditions should be regarded as a smooth curve. Thus, for applicative purpose, it can be useful to imagine the $H(t)$

function as a smooth, nondecreasing curve which passes through the plotted points $(x_{(i)}, \tilde{H}(x_{(i)}))$.

In the case of singly censored sample data, where all failure times are smaller than or equal to the unique single censoring time, say $T_C \geq x_{(s)}$, the same approach can be adopted

$$\tilde{H}(t) = \begin{cases} 0, & \text{for } 0 \leq t < x_{(1)} \\ -\ln \left(1 - \sum_{j=1}^i \frac{d_j}{n} \right), & \text{for } x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, 2, \dots, s-1, \\ -\ln \left(1 - \sum_{j=1}^s \frac{d_j}{n} \right), & \text{for } x_{(s)} \leq t \leq T_C \end{cases} \quad (5)$$

where $\sum_{j=1}^s d_j$ is the total number of failed items in the sample. Estimator 5 can also be used in the case of multiply censored samples if all failure times are not greater than the smallest censoring time. In this case, T_C equals the largest censoring time.

In the case of complete or singly censored samples, the exact $(1 - \alpha)$ confidence intervals for $H(t)$, say $(\tilde{H}_L(t), \tilde{H}_U(t))$, can be easily obtained from Ref. 6 as follows:

$$\begin{aligned} \tilde{H}_L(t) : & \sum_{j=0}^{n \exp[-\tilde{H}(t)]} \binom{n}{j} (\exp[-\tilde{H}_L(t)])^j \\ & \times (1 - \exp[-\tilde{H}_L(t)])^{n-j} = \alpha/2 \\ \tilde{H}_U(t) : & \sum_{j=0}^{n \exp[-\tilde{H}(t)]-1} \binom{n}{j} (\exp[-\tilde{H}_U(t)])^j \\ & \times (1 - \exp[-\tilde{H}_U(t)])^{n-j} = 1 - \alpha/2. \end{aligned} \quad (6)$$

Note that, thanks to the relationship between the cumulative binomial function and the cumulative beta distribution, the limits (Eq. 6) can be also obtained by inverting the cumulative beta distribution,

so that

$$\begin{aligned}\tilde{H}_L(t) &= -\ln \left(B_{1-\alpha/2}(n \exp\{-\tilde{H}(t)\} \right. \\ &\quad \left. + 1, n[1 - \exp\{-\tilde{H}(t)\}]) \right) \\ \tilde{H}_U(t) &= -\ln \left(B_{\alpha/2}(n \exp\{-\tilde{H}(t)\}, \right. \\ &\quad \left. n[1 - \exp\{-\tilde{H}(t)\} + 1]) \right), \quad (7)\end{aligned}$$

where $B_\gamma(p, q)$ denotes the γ th quantile of the beta distribution with parameters p and q . Of course, in case of censoring, the time axis can be investigated up to the censoring time, say $x_{(s)}$ or T_C . An approximate $(1 - \alpha)$ confidence interval for $H(t)$ can be obtained by using the normal approximation for the distribution of $\tilde{H}(t)$,

$$\begin{aligned}\tilde{H}_L(t) &= \tilde{H}(t) + z_{\alpha/2} \cdot \tilde{\sigma}_{\tilde{H}} \quad \text{and} \\ \tilde{H}_U(t) &= \tilde{H}(t) - z_{\alpha/2} \cdot \tilde{\sigma}_{\tilde{H}}, \quad (8)\end{aligned}$$

where $z_{\alpha/2}$ is the $\alpha/2$ th quantile of the standard normal distribution, and

$$\tilde{\sigma}_{\tilde{H}} = \sqrt{\frac{\exp[\tilde{H}(t)] - 1}{n}} \quad (9)$$

is an estimate of the standard deviation of $\tilde{H}(t)$. The normal approximation works well for large samples but, for small and moderate samples, Equation (8) can provide a negative lower confidence limit, so that $\tilde{H}_L(t) = 0$ must be adopted. A way to overcome such a problem is to assume a lognormal approximation for $\tilde{H}(t)$, so that an approximate $(1 - \alpha)$ confidence interval for $H(t)$ results in

$$\begin{aligned}\tilde{H}_L(t) &= \tilde{H} \cdot \exp(z_{\alpha/2} \tilde{\sigma}_{\tilde{H}} / \tilde{H}) \quad \text{and} \\ \tilde{H}_U(t) &= \tilde{H} \cdot \exp(-z_{\alpha/2} \tilde{\sigma}_{\tilde{H}} / \tilde{H}). \quad (10)\end{aligned}$$

All the illustrated procedures for constructing confidence intervals for $H(t)$ can be used also in the case of multiple censored samples, provided that censoring times are equal to or greater than the largest failure time. In this case, the time axis can be investigated until the largest censoring time.

An estimate of the failure rate $h(t)$ at any time $t \leq x_{(s)}$ (or $t \leq T_C$) can be obtained by

smoothing the increments of the estimator (Eq. 4 or 5) of the cumulative failure rate from time $x_{(i-1)}$ up to time $x_{(i)}$ ($i = 1, 2, \dots, s$), say $\Delta \tilde{H}_i = \tilde{H}(x_{(i)}) - \tilde{H}(x_{(i-1)})$, as given by the following expression from Ref. 7:

$$\tilde{h}_k(t) = \frac{1}{b_n} \sum_{i=1}^s W\left(\frac{t - x_{(i)}}{b_n}\right) \cdot \Delta \tilde{H}_i, \quad (11)$$

where

$$\Delta \tilde{H}_i = \ln \frac{n - \sum_{j=1}^{i-1} d_j}{n - \sum_{j=1}^i d_j}, \quad i = 1, 2, \dots, s, \quad (12)$$

b_n is a sequence of smoothing parameters, known as the *bandwidth*, satisfying $b_n \rightarrow 0$ but $n b_n \rightarrow \infty$ as the sample size n tends to infinity, and the *kernel function* $W(z)$ is a bounded nonnegative weight function which satisfies $\int_{-\infty}^{\infty} W(z) dz = 1$. The kernel $W(z)$ is typically assumed to be a symmetric probability density function with zero mean and unit variance. For example, $W(z)$ can be the normal density function. The level of smoothness is set by the bandwidth b_n : the larger the b_n , the smoother the $\tilde{h}_k(t)$. Note that two relevant problems occur when applying kernel estimators: the boundary effects near the endpoints of the support of the failure rate and a substantial increase in the variance from left to right over the range of abscissae where the failure rate is estimated. Boundary-corrected kernel estimators and practical bandwidth choices can be found in Ref. 8. Further details about kernel estimators may be found in Ref. 9.

Multiply Censored Samples with Intermixed Failure and Censoring Times

Multiply censored samples are usually characterized by the fact that failures and censoring times are intermixed. In this case, the nonparametric estimator (Eq. 5) cannot be used and *ad hoc* methods must be used.

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a random sample of size n consisting of r failure times and $n - r$ censored observations, and let $\mathbf{x} = \{x_{(1)}, x_{(2)}, \dots, x_{(s)}\}$ denote the ordered set

of the s ($s \leq r$) distinct values of the set $\mathbf{t}$. Moreover, let d_i ($i = 1, 2, \dots, s$) denote the number of failures occurring at the same time $x_{(i)}$ ($\sum_{i=1}^s d_i = r$), let C_i denote the total number of censoring events that occurred before time $x_{(i)}$, and let k_i indicate the number of units at risk (neither failed nor withdrawn) immediately before $x_{(i)}$:

$$k_i = n - \sum_{j=1}^{i-1} d_j - C_i. \quad (13)$$

Due to Equation (3), a Kaplan–Meier (KM)-based estimate of $H(t)$ is as follows from Ref. 2:

$$\begin{aligned} \bar{H}(t) &= \begin{cases} 0, & 0 \leq t < x_{(1)} \\ -\sum_{j=1}^i \ln \left(1 - \frac{d_j}{k_j} \right), & x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, 2, \dots, s-1, \\ -\sum_{j=1}^s \ln \left(1 - \frac{d_j}{k_j} \right), & x_{(s)} \leq t \leq t_{\max} \end{cases} \end{aligned} \quad (14)$$

and $t_{\max}$ is the largest value in $\{t_1, t_2, \dots, t_n\}$. Note that, in case of a tie between a censoring time and the generic failure time $x_{(i)}$, the KM method assumes that the censoring time is infinitesimally larger than the failure time, so that the censored unit is still at risk at $x_{(i)}$.

The standard deviation $\sigma_{\bar{H}}$ of $\bar{H}(t)$, based on the Greenwood formula [1], may be obtained as follows:

$$\sigma_{\bar{H}} = \sqrt{\sum_{j: x_{(j)} \leq t} \frac{d_j}{k_j(k_j - d_j)}}. \quad (15)$$

An alternative to the KM-based estimator (Eq. 14) is the Nelson–Aalen (NA) estimator [1]:

$$\hat{H}(t) = \sum_{j: x_{(j)} \leq t} \frac{d_j}{k_j}. \quad (16)$$

For continuous models, $\bar{H}(t)$ and $\hat{H}(t)$ are asymptotically (i.e., for large n) equivalent

and do not differ greatly in most situations, except for large t values. In fact,

$$\begin{aligned} \bar{H}(t) &= - \sum_{j: x_{(j)} < t} \ln \left(1 - \frac{d_j}{k_j} \right) \\ &= \sum_{j: x_{(j)} < t} \left(\frac{d_j}{k_j} + \frac{d_j^2}{2k_j^2} + \dots \right) \end{aligned} \quad (17)$$

and thus, $\hat{H}(t)$ is a first-order approximation to $\bar{H}(t)$ [1]. This asymptotic equivalence makes it possible for the standard deviation $\sigma_{\hat{H}}$ of $\hat{H}(t)$ to be estimated from Equation (15).

Approximate confidence intervals for $H(t)$, at any fixed time $t \leq t_{\max}$, can be obtained through the normal or the lognormal approximation for the distribution of the estimator, using $\bar{H}(t)$ or $\hat{H}(t)$ in place of $\tilde{H}(t)$ in Equation (8) or (10), and estimating the standard deviation through Equation (15).

A kernel smoothing of the estimated increments $\Delta \bar{H}_i = \bar{H}(x_{(i)}) - \bar{H}(x_{(i-1)}) = \ln(1 - d_i/k_i)$ and $\Delta \hat{H}_i = \hat{H}(x_{(i)}) - \hat{H}(x_{(i-1)}) = d_i/k_i$ ($i = 1, 2, \dots, s$) of the cumulative failure rate from time $x_{(i-1)}$ up to time $x_{(i)}$, provides smooth estimates of the failure rate $h(t)$ at any time $t \leq t_{\max}$, as given by Ref. 7:

$$\begin{aligned} \bar{h}_k(t) &= \frac{1}{b_n} \sum_{i=1}^s W \left(\frac{t - x_{(i)}}{b_n} \right) \cdot \Delta \bar{H}_i \quad \text{and} \\ \hat{h}_k(t) &= \frac{1}{b_n} \sum_{i=1}^s W \left(\frac{t - x_{(i)}}{b_n} \right) \cdot \Delta \hat{H}_i. \end{aligned} \quad (18)$$

Example 1. Let us consider the data on field windings of $n = 16$ generators given by Nelson [2]. These data consist of $r = 7$ failure times (in months) of the failed windings and of $n - r = 9$ running times (in months) of the windings that did not fail (Table 1, where the asterisk denotes censoring times). The failure and censored times are intermixed because the units were put into service at different times. The largest time is a censoring time, equal to $t_{\max} = 130$ months, and there are no tied failure times, so that $s = r = 7$ and $d_i = 1$ for any failure time.

Table 1. The Winding Data in Example 1

i	t_i (mo)	i	t_i (mo)
1	31.7	9	87.5 ^a
2	39.2	10	88.3 ^a
3	57.5	11	94.2 ^a
4	65.0 ^a	12	101.7 ^a
5	65.8	13	105.8
6	70.0	14	109.2 ^a
7	75.0 ^a	15	110
8	75.0 ^a	16	130 ^a

^aDenotes a censoring time.

The distinct ordered times $x_{(i)}$ ($i = 1, 2, \dots, 7$) at which failures occurred, the total number C_i of windings censored before $x_{(i)}$, and the number k_i of windings at risk immediately before $x_{(i)}$, computed through Equation (13), are given in Table 2.

The cumulative hazard function $H(t)$ is estimated through the NA estimator (Eq. 16) and its standard deviation $\sigma_{\hat{H}}$ is estimated from Equation (15). Table 2 gives both $H(t)$ and $\hat{\sigma}_{\hat{H}}$ at each failure time $x_{(i)}$ ($i = 1, 2, \dots, 7$), together with the approximate 90% confidence intervals obtained from both, Equation (8) [i.e., based on the normal approximation for $\hat{H}(t)$] and from Equation (10) (i.e., based on the lognormal approximation). As expected, due to the small size of the sample, the lower confidence limits provided by Equation (8) are negative for some times t , and then for such times $\hat{H}_L(t) = 0$ should be adopted. For a comparative purpose, the last column of Table 2 gives the KM-based estimator $\bar{H}(t)$ too. As expected, the two estimates $H(t)$ and $\bar{H}(t)$ are quite close to each other, except for t values larger than 100 months when $\bar{H}(t)$ exceeds $\hat{H}(t)$ sensibly.

Figure 1 depicts the kernel estimates $\hat{h}_k(t)$ obtained for $b_n = 5$ and 10 months, and using a normal density function as kernel function $W(z)$. The kernel estimates give evidence of increasing failure rate with age t and, as expected, the larger value of the bandwidth b_n provides the smoother function.

For large samples, a suitable alternative to the estimates (Eqs 14 and 16) is available if the whole time interval $(0, t_{\max})$ is partitioned into a number m of subintervals, say $(\tau_{i-1}, \tau_i]$ ($i = 1, 2, \dots, m$), being $\tau_0 \equiv 0$ and

$\tau_m = t_{\max}$, and both failure and censoring data are grouped into the m subintervals. These subintervals may be of equal length, but this is not necessary; they only have to be contiguous. In this case, the actuarial-based estimator $\tilde{H}_A(\tau_i)$ of $H(\tau_i)$ [2] is

$$\tilde{H}_A(\tau_i) = - \sum_{j=1}^i \ln \left(1 - \frac{d_j}{k_j - c_j/2} \right),$$

where $i = 1, 2, \dots, m$, (19)

and k_i, c_i , and d_i , are the respective number of units at risk at the beginning of the i th subinterval $(\tau_{i-1}, \tau_i]$, the number of units withdrawn from the test during $(\tau_{i-1}, \tau_i]$, and the number of failures occurred in $(\tau_{i-1}, \tau_i]$. Of course, $k_1 = n$ and $k_{i+1} = k_i - d_i - c_i$ ($i = 1, 2, \dots, m-1$). The factor $-c_j/2$ adjusts for the censored units, which are assumed to run about half the j th subinterval. Other censoring adjustments have been proposed in the literature [10].

In contrast to the KM-based estimator (Eq. 14) and to the NA estimator (Eq. 16), the actuarial-based estimator (Eq. 19) gives estimates at the endpoints of the subintervals. If the i th subinterval contains one or more failures, then $\tilde{H}_A(t)$ is to be considered undefined within this subinterval. If zero failure occur in (τ_{i-1}, τ_i) , then $\tilde{H}_A(t) = \tilde{H}_A(\tau_{i-1})$ for any $\tau_{i-1} \leq t < \tau_i$.

Approximate confidence intervals for $H(t)$ can be obtained by assuming the lognormal distribution for $\tilde{H}_A(t)$ (Eq. 10), where the approximate standard deviation of $\tilde{H}_A(t)$ is estimated from the Greenwood formula (Eq. 16), taking into account that, approximately, $\tilde{\sigma}_{\tilde{H}} \cong \tilde{\sigma}_{\tilde{R}}/\tilde{R}(t)$ and using the adjusted $k_j - c_j/2$ instead of k_j :

$$\tilde{\sigma}_{\tilde{H}_A} = \sqrt{\sum_{j=1}^i \frac{d_j}{(k_j - c_j/2) \cdot (k_j - c_j/2 - d_j)}}. \quad (20)$$

When data are grouped, a nonparametric estimate of the failure rate $h(t)$ can be obtained from a combination of smoothing central rates of failures and a suitable transformation of the central failure rate, where

Table 2. Winding Failure Times, NA Estimate $\hat{H}(t)$, Estimated Standard Deviation $\hat{\sigma}_{\hat{H}}$, 95% Approximate Confidence Intervals Based on the Normal and Lognormal Approximation, and Kaplan–Meier-based (KM) Estimate $\bar{H}(t)$ in Example 1

i	$x_{(i)}$ (mo)	C_i	k_i	$\hat{H}(t)$ (NA)	$\hat{\sigma}_{\hat{H}}$	Approximate 90% Confidence Intervals		$\bar{H}(t)$ (KM)
						Normal approx.	Lognormal approx.	
1	31.7	0	16	0.063	0.065	(−0.044, 0.169)	(0.011, 0.342)	0.065
2	39.2	0	15	0.129	0.094	(−0.026, 0.285)	(0.039, 0.430)	0.134
3	57.5	0	14	0.201	0.120	(0.003, 0.398)	(0.075, 0.537)	0.208
4	65.8	1	12	0.284	0.148	(0.040, 0.528)	(0.120, 0.670)	0.295
5	70.0	1	11	0.375	0.176	(0.085, 0.665)	(0.173, 0.813)	0.390
6	105.8	7	4	0.625	0.338	(0.068, 1.181)	(0.256, 1.522)	0.678
7	110.0	8	2	1.125	0.784	(−0.164, 2.414)	(0.358, 3.539)	1.371

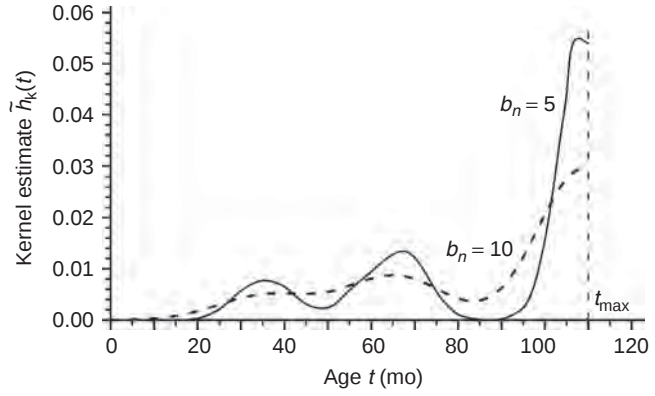


Figure 1. Plot of the kernel estimates of $h(t)$ for the data in Example 1.

the central failure rate for the i th subinterval is defined as follows from Ref. 11:

$$\tilde{h}_{c_i} = \frac{2d_i}{(k_i + k_{i+1})(\tau_i - \tau_{i-1})}, \quad i = 1, 2, \dots, m, \quad (21)$$

where $k_{m+1} = k_m - d_m - c_m$ is the number of units at risk at $t_{\max}$. Then, the smoothed central failure rate is obtained by smoothing the points $(\tau_i + \tau_{i-1})/2, \tilde{h}_{c_i}$ ($i = 1, 2, \dots, m$), for example with weighted least-squares smoothers or with spline smoothers, and a recommended transformation is discussed in Ref. 12.

PARAMETRIC PROCEDURES

Parametric estimation procedures require that a probability model is assumed for the failure time distribution, say

$$T \sim f(t; \boldsymbol{\theta}), \quad t \geq 0, \quad (22)$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_l\}$ is the vector of l parameters, some of which are unknown. Among all the parametric estimation procedures, only the maximum likelihood (ML) estimation method is discussed here, and the specific cases of the standard exponential and two-parameter Weibull distributions are considered in detail.

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a random sample of size n of exact failure times and right censored data from the failure time distribution $f(t; \boldsymbol{\theta})$. The idea behind ML estimation is that the value of model parameters which maximizes the probability (likelihood) of the sample data is their best estimate. In this context, the (joint) data probability function is called the *likelihood function*, say $L(\boldsymbol{\theta}; \mathbf{t})$, to highlight that it is regarded as a

function of the unknown parameters. Since multiplying the joint distribution function by a constant does not alter the ML estimates, the likelihood function is usually defined up to a *constant* multiplicative factor (where the term *constant* means that it does not depend on the unknown parameters, even if it may possibly depend on the data). Under such hypotheses, the likelihood function $L(\theta; \mathbf{t})$ can be formulated as follows:

$$L(\theta; \mathbf{t}) = \prod_{i=1}^n f(t_i; \theta)^{\delta_i} \cdot R(t_i; \theta)^{1-\delta_i}, \quad (23)$$

where $\delta_i = 1$ if t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation.

The ML estimate of θ , say $\hat{\theta} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l\}$, can be obtained (if it exists) as:

$$\hat{\theta} = \max_{\theta \in \Theta} L(\theta; \mathbf{t}), \quad (24)$$

with Θ being the space of the parameters. Hence, the ML estimate of the failure rate, say $\hat{h}(t)$, and of the cumulative failure rate, say $\hat{H}(t)$, for any time $t \geq 0$ can be easily obtained for any time t as

$$\hat{h}(t) = h(t; \hat{\theta}) \quad \text{and} \quad \hat{H}(t) = H(t; \hat{\theta}). \quad (25)$$

If an analytic closed-form expression exists for $h(t)$, the distribution function can be re-parameterized so that the value of the failure rate at whatever prefixed time x , say $h(x)$, explicitly appears as a parameter of the probability model. In this case, a direct estimation of $h(x)$ can be performed and the direct estimate, due to the invariance property of the ML method, will coincide with the indirect estimate (Eq. 25). Of course, the same result holds for $H(t)$.

Instead of maximizing the likelihood function (Eq. 23), it is generally much more convenient to maximize the log-likelihood function:

$$\begin{aligned} \ell(\theta; \mathbf{t}) &= \ln L(\theta; \mathbf{t}) \\ &= \sum_{i=1}^n \delta_i \cdot \ln f(t_i; \theta) + (1 - \delta_i) \cdot \ln R(t_i; \theta). \end{aligned} \quad (26)$$

In fact, the logarithm function is a monotone increasing function, and thus the

value $\hat{\theta}$ that maximizes $\ell(\theta; \mathbf{t})$, if it exists, is the same as the one that maximizes $L(\theta; \mathbf{t})$. The advantage in maximizing the log-likelihood function is that a logarithmic transformation usually yields simplified mathematics, since it transforms products in sums and often prevents such computational problems as overflows and/or underflows, often caused by Equation (23).

Approximate confidence intervals for $h(t)$ and $H(t)$, homolog to the confidence intervals for nonparametric estimates, can be obtained adopting the normal approximation for the distribution of $\hat{h}(t)$ and $\hat{H}(t)$, respectively, so that

$$\begin{aligned} (\hat{h}_L(t), \hat{h}_U(t)) &= \left(\hat{h}(t) + z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{h}(t))}, \right. \\ &\quad \left. \hat{h}(t) - z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{h}(t))} \right), \end{aligned} \quad (27)$$

$$\begin{aligned} (\hat{H}_L(t), \hat{H}_U(t)) &= \left(\hat{H}(t) + z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{H}(t))}, \right. \\ &\quad \left. \hat{H}(t) - z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{H}(t))} \right), \end{aligned} \quad (28)$$

where $z_{\alpha/2}$ is the $\alpha/2$ th quantile of the standard normal distribution, and $\hat{\text{Var}}(\hat{h}(t))$ and $\hat{\text{Var}}(\hat{H}(t))$ denote the estimates of the asymptotic variance of $\hat{h}(t)$ and $\hat{H}(t)$, respectively. Such variances can be obtained through the delta method [13]:

$$\begin{aligned} \hat{\text{Var}}(\hat{h}(t)) &= \sum_{i=1}^l \left[\frac{\partial h(t)}{\partial \theta_i} \Big|_{\hat{\theta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\theta}_i) \\ &\quad + 2 \cdot \sum_{i=1}^l \sum_{j=i+1}^l \frac{\partial h(t)}{\partial \theta_i} \Big|_{\hat{\theta}} \frac{\partial h(t)}{\partial \theta_j} \Big|_{\hat{\theta}} \\ &\quad \cdot \hat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j), \end{aligned} \quad (29)$$

$$\begin{aligned} \hat{\text{Var}}(\hat{H}(t)) &= \sum_{i=1}^l \left[\frac{\partial H(t)}{\partial \theta_i} \Big|_{\hat{\theta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\theta}_i) \\ &\quad + 2 \cdot \sum_{i=1}^l \sum_{j=i+1}^l \frac{\partial H(t)}{\partial \theta_i} \Big|_{\hat{\theta}} \frac{\partial H(t)}{\partial \theta_j} \Big|_{\hat{\theta}} \\ &\quad \cdot \hat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j), \end{aligned} \quad (30)$$

where $\hat{\text{Var}}(\hat{\theta}_i)$ and $\hat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j)$ ($i, j = 1, \dots, l$) are the elements (i, i) and (i, j) respectively, of the estimated $l \times l$ asymptotic variance–covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ of the parameter vector $\boldsymbol{\theta}$ (refer Appendix) and the partial derivatives of $h(t)$ and $H(t)$ with respect to the model parameters must be evaluated at the ML estimate $\hat{\boldsymbol{\theta}}$.

As an alternative, confidence limits can be based on the lognormal approximation for the distribution of $\hat{h}(t)$ and $\hat{H}(t)$, so that

$$\begin{aligned} &(\hat{h}_L(t), \hat{h}_U(t)) \\ &= \left(\hat{h}(t) \cdot \exp \left[z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{h}(t))/\hat{h}(t)} \right], \right. \\ &\quad \left. \hat{h}(t) \cdot \exp \left[-z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{h}(t))/\hat{h}(t)} \right] \right), \end{aligned} \quad (31)$$

$$\begin{aligned} &(\hat{H}_L(t), \hat{H}_U(t)) \\ &= \left(\hat{H}(t) \cdot \exp \left[z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{H}(t))/\hat{H}(t)} \right], \right. \\ &\quad \left. \hat{H}(t) \cdot \exp \left[-z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{H}(t))/\hat{H}(t)} \right] \right), \end{aligned} \quad (32)$$

which, unlike Equations (27) and (28), ensures that the lower limits are always positive for any $t > 0$, even for small samples.

If the probability density function can be re-parameterized in terms of the failure rate $h(x)$ evaluated at a specific time x of interest and in terms of $(l - 1)$ remaining parameters, say $\{\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_l\}$, then $h(x)$ becomes one of the model parameters, say $h(x) \equiv \theta_k$, and the estimated variance of $\hat{h}(x)$ is just the element (k, k) of the matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$. Clearly, the same result holds for $H(x)$.

Also the approximate confidence intervals (Eqs 31 and 32) generally work badly if the sample size is small. More accurate approximation can be obtained by using the parametric bootstrap procedure, which consists of choosing a suitable model for the observed data, estimating the model parameters (often by the ML method) and the quantity of interest, say $h(x)$, from the observed data, resampling from this (estimated) parametric model, and finally estimating $h(x)$ from the bootstrap sample [14]. If the resampling procedure

is repeated a large number of times, the empirical distribution of the bootstrap estimates, say $\{\hat{h}^{(1)}, \hat{h}^{(2)}, \hat{h}^{(3)}, \dots\}$, serves as an approximation to the sampling distribution of $\hat{h}(x)$, and an approximate confidence interval for $h(x)$ can be obtained.

Exact solutions for the interval estimation problem are sometimes available, depending on the lifetime distribution model and the censoring scheme. In the following, some procedures (both exact and approximate) for the one-parameter (or standard) exponential and two-parameter Weibull distributions are presented.

Data from the One-Parameter Exponential Distribution

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a random sample of size n consisting of r failure times and $n - r$ censored observations from a one-parameter exponential random variable with density function:

$$f(t; \lambda) = \lambda \exp(-\lambda \cdot t); \quad \lambda > 0, t \geq 0, \quad (33)$$

where the parameter λ is just the constant failure rate. The reliability function is

$$R(t; \lambda) = \exp(-\lambda \cdot t); \quad \lambda > 0, t \geq 0. \quad (34)$$

The likelihood function $L(\lambda; \mathbf{t})$ results in

$$\begin{aligned} L(\lambda; \mathbf{t}) &\propto \prod_{i=1}^n \lambda^{\delta_i} \exp(-\lambda \cdot \delta_i \cdot t_i) \\ &\quad \cdot \exp[-\lambda \cdot (1 - \delta_i) \cdot t_i] = \\ &= \prod_{i=1}^n \lambda^{\delta_i} \exp(-\lambda \cdot t_i) \\ &= \lambda^r \cdot \exp \left(-\lambda \cdot \sum_{i=1}^n t_i \right), \end{aligned} \quad (35)$$

where $\delta_i = 1$ if t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation. Thus, $\sum_{i=1}^n \delta_i = r$ is the number of exact failure times and the sum $\sum_{i=1}^n t_i$ is the total time on test. The log-likelihood function $\ell(\lambda; \mathbf{t})$ is

$$\ell(\lambda; \mathbf{t}) = \ln \left[\lambda^r \exp \left(-\lambda \cdot \sum_{i=1}^n t_i \right) \right]$$

$$= r \ln \lambda - \lambda \sum_{i=1}^n t_i, \quad (36)$$

and for $r > 0$, the ML estimate of the failure rate λ is equal to the number of observed failures over the total time on test [1]:

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^n t_i}. \quad (37)$$

The ML estimate of cumulative failure rate $H(t)$ can be straightforwardly obtained as

$$\hat{H}(t) = \int_0^t \hat{h}(x) dx = \hat{\lambda} t. \quad (38)$$

In the case of complete or Type II censored sample, the exact $(1 - \alpha)$ confidence interval for the failure rate λ is given by the following expression [15]:

$$\begin{aligned} (\hat{\lambda}_L, \hat{\lambda}_U) &= \left(\frac{\hat{\lambda} \chi_{\alpha/2}^2(2r)}{2r}, \frac{\hat{\lambda} \chi_{1-\alpha/2}^2(2r)}{2r} \right) \\ &= \left(\frac{\chi_{\alpha/2}^2(2r)}{2 \cdot \sum_{i=1}^n t_i}, \frac{\chi_{1-\alpha/2}^2(2r)}{2 \cdot \sum_{i=1}^n t_i} \right), \end{aligned} \quad (39)$$

where $\chi_{\gamma}^2(\nu)$ is the γ th quantile of the χ -square distribution with ν degrees of freedom. Since $H(t) = \lambda t$ for any fixed $t > 0$, the exact $(1 - \alpha)$ confidence interval for $H(t)$ is

$$\begin{aligned} (\hat{H}_L(t), \hat{H}_U(t)) &= (\hat{\lambda}_L t, \hat{\lambda}_U t) \\ &= \left(\frac{t \cdot \chi_{\alpha/2}^2(2r)}{2 \cdot \sum_{i=1}^n t_i}, \frac{t \cdot \chi_{1-\alpha/2}^2(2r)}{2 \cdot \sum_{i=1}^n t_i} \right). \end{aligned} \quad (40)$$

In the case of singly Type I censored sample or multiply censored sample, exact confidence intervals for λ , and hence for $H(t)$, are not available [16]. An approximate $(1 - \alpha)$ confidence interval for λ can

be obtained by employing the asymptotic normal approximation of $(\hat{\lambda} - \lambda)/\hat{\sigma}_{\hat{\lambda}}$, where $\hat{\sigma}_{\hat{\lambda}} = \hat{\lambda}/\sqrt{r}$ is the estimated asymptotic standard deviation of $\hat{\lambda}$ (obtained by inverting the estimated 1×1 information matrix $\hat{\mathbf{J}}(\hat{\theta})$, refer Appendix). Under such an approximation, we have from Ref. 1,

$$(\hat{\lambda}_L, \hat{\lambda}_U) = (\hat{\lambda} + z_{\alpha/2} \hat{\lambda}/\sqrt{r}, \hat{\lambda} - z_{\alpha/2} \hat{\lambda}/\sqrt{r}). \quad (41)$$

This approximation works well for moderate and large samples, whereas for small samples a better approximation [17] is given by

$$\begin{aligned} (\hat{\lambda}_L, \hat{\lambda}_U) &= \left(\left[\hat{\lambda}^{1/3} + z_{\alpha/2} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3, \right. \\ &\quad \left. \left[\hat{\lambda}^{1/3} - z_{\alpha/2} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3 \right), \end{aligned} \quad (42)$$

which employs the normal approximation of $\hat{\lambda}^{1/3}$ in small samples. Another approximate method (often attributed to Ref. 18), which works well in the case of singly Type I censored samples, treats $2r\lambda/\hat{\lambda}$ as being approximately distributed as a χ -square random variable with $(2r + 1)$ degrees of freedom, so that

$$(\hat{\lambda}_L, \hat{\lambda}_U) = \left(\frac{\hat{\lambda} \chi_{\alpha/2}^2(2r + 1)}{2r}, \frac{\hat{\lambda} \chi_{1-\alpha/2}^2(2r + 1)}{2r} \right). \quad (43)$$

Whichever asymptotic approximation is used to compute confidence limits for λ , an approximate $(1 - \alpha)$ confidence interval for $H(t)$ is given by

$$(\hat{H}_L(t), \hat{H}_U(t)) = (\hat{\lambda}_L t, \hat{\lambda}_U t). \quad (44)$$

Example 2. Let us consider the Type I censored data given in Ref. 19 which refers to 15 units of an electromechanical module tested at the temperature of 125°C. All the tests were stopped at the time $T_C = 1100$ h and 8 failures were observed at no tied times

$x_{(i)}$ ($i = 1, 2, \dots, 8$):

$$x_{(1)} = 88 \text{ h}, x_{(2)} = 105 \text{ h}, x_{(3)} = 141 \text{ h},$$

$$x_{(4)} = 344 \text{ h}, x_{(5)} = 430 \text{ h}, x_{(6)} = 516 \text{ h},$$

$$x_{(7)} = 937 \text{ h}, x_{(8)} = 1057 \text{ h}.$$

By assuming that the data follow the one-parameter exponential function, the ML estimate of the failure rate is

$$(\hat{\lambda}_L, \hat{\lambda}_U) = \begin{cases} (0.000217 \text{ h}^{-1}, 0.001197 \text{ h}^{-1}), & \text{from the normal approximation of } \hat{\lambda} \\ (0.000321 \text{ h}^{-1}, 0.001318 \text{ h}^{-1}), & \text{from the normal approximation of } \hat{\lambda}^{1/3} \\ (0.000334 \text{ h}^{-1}, 0.001334 \text{ h}^{-1}), & \text{from the } \chi^2 \text{-square approximation of } 2r\lambda/\hat{\lambda}, \end{cases} \quad (46)$$

being $\hat{\sigma}_{\hat{\lambda}} = \hat{\lambda}/\sqrt{r} = 0.000250 \text{ h}^{-1}$, $z_{0.025} = -1.960$, $\chi_{0.025}^2(17) = 7.564$, and $\chi_{0.975}^2(17) = 30.191$. From Equation (46), by using Equation (44), approximate 95% confidence

$$\hat{\lambda} = r / \sum_{i=1}^n t_i = 0.000707 \text{ h}^{-1}, \quad (45)$$

with r being the number of exact failure times equal to 8 and the total time on test $\sum_{i=1}^{15} t_i = 11,318 \text{ h}$. From Equations (41), (42) and (43), approximate 95% confidence intervals for λ are

intervals for $H(t)$ can be easily obtained for any t value of interest. For example, for $t = 500 \text{ h}$:

$$(\hat{H}_L(500), \hat{H}_U(500)) = \begin{cases} (0.109, 0.598), & \text{from the normal approximation of } \hat{\lambda} \\ (0.161, 0.659), & \text{from the normal approximation of } \hat{\lambda}^{1/3} \\ (0.167, 0.667), & \text{from the } \chi^2 \text{-square approximation of } 2r\lambda/\hat{\lambda}. \end{cases} \quad (47)$$

Since the size of the sample is not large, the confidence intervals based on the normal approximation of $\hat{\lambda}$ should be discarded. The remaining two approximations should both, work better and give confidence intervals close to each other.

Data from the Two-Parameter Weibull Distribution

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ be a random sample of size n consisting of r failure times and $n - r$ censored observations from a two-parameter Weibull random variable with density function

$$f(t; \theta, \beta) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1} \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right], \quad \theta, \beta > 0, t \geq 0 \quad (48)$$

and reliability function

$$R(t; \theta, \beta) = \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right], \quad \theta, \beta > 0, t \geq 0. \quad (49)$$

The likelihood function $L(\alpha, \beta; \mathbf{t})$ is

$$\begin{aligned} L(\theta, \beta; \mathbf{t}) &\propto \prod_{i=1}^n \left(\frac{\beta}{\theta} \right)^{\delta_i} \left(\frac{t_i}{\theta} \right)^{\delta_i(\beta-1)} \\ &\times \exp \left[-\delta_i \left(\frac{t_i}{\theta} \right)^{\beta} \right] \exp \left[-(1 - \delta_i) \cdot \left(\frac{t_i}{\theta} \right)^{\beta} \right] \\ &= \frac{\beta^r}{\theta^{r\beta}} \left(\prod_{i=1}^n t_i^{\delta_i} \right)^{\beta-1} \exp \left[- \left(\frac{\sum_{i=1}^n t_i^{\beta}}{\theta^{\beta}} \right) \right], \end{aligned} \quad (50)$$

where $\sum_{i=1}^n \delta_i = r$ is the number of exact failure times. The log-likelihood function $\ell(\theta, \beta; \mathbf{t})$ results in

$$\begin{aligned} \ell(\theta, \beta; \mathbf{t}) &= r \ln \beta - r \beta \ln \theta \\ &+ (\beta - 1) \cdot \sum_{i=1}^n (\delta_i \ln t_i) - \left(\frac{\sum_{i=1}^n t_i^{\beta}}{\theta^{\beta}} \right). \end{aligned} \quad (51)$$

By maximizing Equation (51), the ML estimate of the scale parameter θ is given by

$$\hat{\theta} = \left(\frac{\sum_{i=1}^n t_i^{\hat{\beta}}}{r} \right)^{1/\hat{\beta}}, \quad (52)$$

where the ML estimate $\hat{\beta}$ of the shape parameter β is the solution of

$$\frac{1}{\hat{\beta}} + \frac{\sum_{i=1}^n \delta_i \cdot \ln t_i}{r} - \frac{\sum_{i=1}^n t_i^{\hat{\beta}} \cdot \ln t_i}{\sum_{i=1}^n t_i^{\hat{\beta}}} = 0. \quad (53)$$

From definitions (Eqs 1 and 2), the failure rate and the cumulative failure rate for a two-parameter Weibull random variable are

$$h(t; \theta, \beta) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1} \text{ and} \\ H(t; \theta, \beta) = \left(\frac{t}{\theta} \right)^{\beta} \quad (54)$$

respectively, and their ML estimates at a given time t ($t \geq 0$) are

$$\hat{h}(t) = \frac{\hat{\beta}}{\hat{\theta}} \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}-1} \text{ and } \hat{H}(t) = \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}}. \quad (55)$$

The ML estimate of the cumulative failure rate at a given time x ($x > 0$) of interest, say $H_x \equiv H(x)$, can also be obtained directly re-parameterizing the log-likelihood function in terms of H_x and β . In this case, being $\theta = x/H_x^{1/\beta}$, the re-parameterized density and reliability functions result in

$$f(t; H_x, \beta) = \frac{\beta}{x} \left(\frac{t}{x} \right)^{\beta-1} \\ \times H_x \cdot \exp \left[-H_x \cdot \left(\frac{t}{x} \right)^{\beta} \right], \\ H_x > 0, \beta > 0, t \geq 0, \quad (56)$$

$$R(t; H_x, \beta) = \exp \left[-H_x \cdot \left(\frac{t}{x} \right)^{\beta} \right], \\ H_x > 0, \beta > 0, t \geq 0. \quad (57)$$

Expressions from Equations (56) and (57) can then be used to obtain the re-parameterized log-likelihood function $\ell(H_x, \beta; \mathbf{t})$. Maximizing $\ell(H_x, \beta; \mathbf{t})$ with respect to H_x and β gives the same ML estimates $\hat{H}_x = \hat{H}(x)$ and $\hat{\beta}$ as those obtained, via Equation (55), by maximizing the log-likelihood (Eq. 51). A similar procedure is to be used if we want to estimate directly the failure rate $h_x \equiv h(x)$ at a given time x , being $\theta = [\beta x^{\beta-1}/H_x]^{1/\beta}$.

In the case of complete samples, exploiting the fact that the distribution of $\hat{R}(t)$ depends only on $R(t)$ and n , the exact lower confidence limits for $R(t)$, say $\hat{R}_L(t)$, were obtained through a Monte Carlo simulation method and tabulated in Ref. 20. Then, from Equation (3), the exact upper confidence limit for $H(t)$ can be easily obtained:

$$\hat{H}_U(t) = -\ln [\hat{R}_L(t)]. \quad (58)$$

Alternatively, an iterative procedure given in Ref. 21 to compute $\hat{R}_L(t)$ allows $\hat{H}_U(t)$ to be obtained by using Equation (58). However, such a procedure is somewhat inconvenient, and the upper $(1 - \alpha)$ limit $\hat{H}_U(t)$ can be more easily computed by interpolating the $(1 - \alpha)$ lower confidence limits for $R(t)$ given in Table 7A of Ref. 21, for n ranging from 8 up to 100, $\hat{R}(t) = 0.50(0.02)0.98$, $\alpha = 0.25, 0.15, 0.10, 0.05$, and 0.02 .

For large samples, under the normal approximation for the distribution of $\hat{h}(t)$ and $\hat{H}(t)$, simple approximate $(1 - \alpha)$ confidence intervals for $h(t)$ and $H(t)$ can be obtained from Equations (27) and (28), respectively. The asymptotic variances $\hat{\text{Var}}(\hat{h}(t))$ and $\hat{\text{Var}}(\hat{H}(t))$ can be estimated through the delta method:

$$\hat{\text{Var}}(\hat{h}(t)) = \left[\frac{\partial h(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\theta}) \\ + \left[\frac{\partial h(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\beta})$$

$$\begin{aligned}
& + 2 \cdot \frac{\partial h(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \cdot \frac{\partial h(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \\
& \cdot \hat{\text{Cov}}(\hat{\theta}, \hat{\beta}), \quad (59) \\
\hat{\text{Var}}(\hat{H}(t)) = & \left[\frac{\partial H(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\theta}) \\
& + \left[\frac{\partial H(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \cdot \hat{\text{Var}}(\hat{\beta}) \\
& + 2 \cdot \frac{\partial H(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \cdot \frac{\partial H(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \\
& \cdot \hat{\text{Cov}}(\hat{\theta}, \hat{\beta}), \quad (60)
\end{aligned}$$

where $\hat{\text{Var}}(\hat{\theta})$, $\hat{\text{Var}}(\hat{\beta})$, and $\hat{\text{Cov}}(\hat{\theta}, \hat{\beta})$ are the elements (1,1), (2,2), and (1,2), respectively, of the estimated 2×2 asymptotic variance-covariance matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$, and the partial derivatives of $h(t)$ and $H(t)$ with respect to the model parameters, say

$$\begin{aligned}
\frac{\partial h(t)}{\partial \theta} &= - \left(\frac{\beta}{\theta} \right)^2 \left(\frac{t}{\theta} \right)^{\beta-1} \quad \text{and} \\
\frac{\partial h(t)}{\partial \beta} &= \frac{1}{t} \left(\frac{t}{\theta} \right)^{\beta} \left[1 + \beta \ln \left(\frac{t}{\theta} \right) \right], \quad (61) \\
\frac{\partial H(t)}{\partial \theta} &= - \left(\frac{\beta}{\theta} \right) \left(\frac{t}{\theta} \right)^{\beta} \quad \text{and} \\
\frac{\partial H(t)}{\partial \beta} &= \beta \left(\frac{t}{\theta} \right)^{\beta} \ln \left(\frac{t}{\theta} \right), \quad (62)
\end{aligned}$$

need to be evaluated at the ML estimates $\hat{\theta}$ and $\hat{\beta}$. Note that, for complete samples, the matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$ is approximately [1,20]

$$\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} \left(\frac{\hat{\theta}}{\hat{\beta}} \right)^2 \frac{1.1087}{n} & \hat{\theta} \cdot \frac{0.2570}{n} \\ \hat{\theta} \cdot \frac{0.2570}{n} & \hat{\beta}^2 \cdot \frac{0.6079}{n} \end{bmatrix}. \quad (63)$$

For small and moderate samples, the approximate $(1 - \alpha)$ confidence intervals (Eqs 31 and 32), based on the lognormal approximation for the distribution of $\hat{h}(t)$ and $\hat{H}(t)$, should work better than the normal approximations, since the lognormal approximation avoids obtaining negative values for the lower confidence limits.

Of course, any other method available for computing confidence limits for $R(t)$ can be

used for deriving confidence limits for $H(t)$. For example, the method suggested by [2] for $R(t)$ gives, as matter of fact, approximate $(1 - \alpha)$ confidence intervals for $U(t) = \ln H(t)$, in the form

$$\begin{aligned}
\hat{U}_L(t) &= \hat{U} + z_{\alpha/2} \\
&\cdot \left[\frac{1.1680 + 1.1000 \hat{U}^2 - 0.1913 \hat{U}}{n} \right]^{1/2}, \quad (64)
\end{aligned}$$

$$\begin{aligned}
\hat{U}_U(t) &= \hat{U} - z_{\alpha/2} \\
&\cdot \left[\frac{1.1680 + 1.1000 \hat{U}^2 - 0.1913 \hat{U}}{n} \right]^{1/2}, \quad (65)
\end{aligned}$$

where $\hat{U} \equiv \ln \hat{H}(t)$. Thus, we simply obtain: $\hat{H}_L(t) = \exp[\hat{U}_L(t)]$ and $\hat{H}_U(t) = \exp[\hat{U}_U(t)]$.

In the case of Type II censored samples, exact $(1 - \alpha)$ upper confidence limits for $H(t)$ can be obtained from the $(1 - \alpha)$ lower confidence limits for $R(t)$ given in Table 7B of Ref. 21, for $n = 40(20)120$, $\hat{R}(t)$ ranging from 0.70 up to 0.999, $\alpha = 0.10, 0.05, 0.02$, and 0.01, and censoring level $r/n = 0.75$ and 0.50. Such limits can be used as approximate limits for Type I censored samples, too. For large samples, approximate confidence intervals for $h(t)$ and $H(t)$ can be obtained from Equations (27) and (28), respectively, whereas for small or moderate samples the approximations (Eqs 31 and 32), based on the lognormal approximation, should work better. The asymptotic variances of $\hat{h}(t)$ and $\hat{H}(t)$ need to be estimated through Equations (29) and (30), respectively, and the asymptotic matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$ is obtained by inverting the estimated information matrix $\hat{\mathbf{J}}(\hat{\theta}, \hat{\beta})$ (Eq 75 in the Appendix). Note that, for singly censored samples, the asymptotic matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$ can be easily read from tables in Ref. 22 as a function of the standardized censoring time $\zeta = [\ln T_C - \ln \hat{\theta}] \hat{\beta}$, where T_C is the censoring time. The matrix is the same for singly censored Type I and II data.

Example 3. Let us consider the Type I censored data referring to $n = 12$ units of

Table 3. Life Test Data of a Certain Type of Electromechanical Device in Example 3

i	t_i (h)	δ_i	i	t_i (h)	δ_i
1	1138	1	7	5046	1
2	1944	1	8	5139	1
3	2764	1	9	5500	0
4	2846	1	10	5500	0
5	3246	1	11	5500	0
6	3803	1	12	5500	0

an electromechanical device given by Yang [19]. All the tests were stopped at the time $T_C = 5500$ h and $r = 8$ failures were observed at times $t_1, t_2, \dots, t_8$ (Table 3, where $\delta_i = 1$ if the time t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation). By solving Equation (53), the ML estimate of β is $\hat{\beta} = 2.055$, from which, by using Equation (52) the ML estimate of θ is

$$\hat{\theta} = \left(\frac{\sum_{i=1}^{12} t_i^{2.055}}{8} \right)^{1/2.055} = 5215 \text{ h.} \quad (66)$$

The failure rate and the cumulative failure rate at the time $t = 4000$ h are estimated from Equation (55): $\hat{h}(4000) = 0.000298 \text{ h}^{-1}$ and $\hat{H}(4000) = 0.580$. The approximate 90% confidence intervals for $h(t)$ and $H(t)$ at $t = 4000$ h, based on the lognormal approximations in Equations (31) and (32), result in

$$(\hat{h}_L(4000), \hat{h}_U(4000)) = (0.000153 \text{ h}^{-1}, 0.000580 \text{ h}^{-1}), \quad (67)$$

$$(\hat{H}_L(4000), \hat{H}_U(4000)) = (0.272, 1.235), \quad (68)$$

being

$$\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta}) = \begin{pmatrix} 825042. \text{ h}^2 & -90.815 \text{ h} \\ -90.815 \text{ h} & 0.4117 \end{pmatrix},$$

$$z_{0.050} = -1.645 \quad \hat{\text{Var}}(\hat{h}(4000)) = (0.000121 \text{ h}^{-1})^2, \text{ and}$$

$$\hat{\text{Var}}(\hat{H}(4000)) = (0.2666)^2.$$

GRAPHICAL PROCEDURE: CUMULATIVE HAZARD PLOTS

Estimate of the cumulative failure rate $H(t)$ can be often performed via *cumulative hazard plots* (often simply called *hazard plots* [2]). This procedure can be used in case the random variable T either pertains to the location-scale family or admits a transformation $g(t)$ which leads to a location-scale distribution. More formally, it is required that the cumulative density function $F(t) = 1 - R(t)$ of the random variable T can be expressed as

$$F(t; a, b) = \Phi \left(\frac{g(t) - a}{b} \right), \quad (69)$$

$\Phi(\cdot)$ being a function independent of any unknown parameter, so that, from Equation (3)

$$\begin{aligned} H(t; a, b) &= -\ln \left[1 - \Phi \left(\frac{g(t) - a}{b} \right) \right] \\ &= \Psi \left(\frac{g(t) - a}{b} \right). \end{aligned} \quad (70)$$

The parameters a and b are called *location* and *scale* parameters, respectively.

Then, the hazard plot is obtained by adopting a $g(t)$ scale on the data axis (the x axis) and a $\Psi^{-1}(H)$ scale on the probability axis (the y axis). On this *model-dependent* chart, the plot of $H(t)$ versus t appears as a straight line. For example, in the case of the two-parameter Weibull model, the cumulative failure rate $H(t) = (t/\theta)^\beta$ can be easily linearized as it follows

$$\ln H(t) = \beta \ln t - \beta \ln \theta, \quad (71)$$

so that, for a two-parameter Weibull random variable, $g(T) = \ln T$, $a = \ln \theta$, $b = 1/\beta$, and $\Psi^{-1}(H) = \ln H(t)$.

Commercial hazard plots are generally available (even if sometimes difficult to find) for certain common distributions, such as the standard exponential, the two-parameter Weibull, and the lognormal distributions. From Equation (71), we have that a two-parameter Weibull hazard plot is easy to construct since it has a log scale on both the axes. Even easier is to construct a standard

exponential hazard plot: it has a linear scale on both the axes because for such a model, $H(t) = \lambda t$ so that $g(T) = T, a = 0, b = \lambda$, and $\Psi^{-1}(H) = H(t)$.

In order to use hazard plots, each ordered failure time $t_{(i)}$ ($i = 1, 2, \dots$) against an estimate of $H(t)$ at that time, say H_i , must be plotted on the chart. These estimates of $H(t)$, named *plotting positions*, are obtained via a nonparametric approach. Only failure data are plotted on the hazard plot, and not censoring times. Thus, plotting positions must be evaluated only at failure times.

Once the data are plotted on the chosen hazard plot, it is necessary to fit a straight line through the points $(t_{(i)}, H_i)$, because the use of the plotting positions inevitably produces departure from linearity. Very often the fitting is performed by eyes, but a more accurate fit can be obtained by means of analytical methods, such as the least-squares method [1]. The graphical estimate of $H(x)$ at a given time x of interest is then given by the ordinate (i.e., the value on the cumulative hazard axis) of the point on the straight line having abscissa x .

Plotting Positions

In the case of complete or singly censored samples, the plotting position H_i for the i th ordered failure time $t_{(i)}$ ($i = 1, 2, \dots, r$) can be calculated using Ref. 2 as

$$H_i = \ln \left(\frac{n}{i - 0.5} \right), \quad i = 1, 2, \dots, r, \quad (72)$$

where n is the size of the sample and $r \leq n$ is the number of failure times in the sample. Other plotting positions, such as $H_i = \ln[(n + 1)/i]$, can be used. In the case of multiply censored samples without tied failure times, the plotting position H_i relative to the i th ordered failure time $t_{(i)}$ can be obtained via the following Herd–Johnson-based estimator [2]:

$$H_i = \sum_{j=1}^i \ln \left(\frac{k_j + 1}{k_j} \right), \quad i = 1, 2, \dots, r, \quad (73)$$

where k_i is the number of units at risk immediately before the ordered failure time $t_{(i)}$.

In the case of tied failure times, the plotting position H_i has to be estimated at any distinct (ordered) time $x_{(i)}$ ($i = 1, 2, \dots, s$) at which one or more failures occur, by using the NA estimator [Eq. (16); [1]]:

$$H_i = \sum_{j=1}^i \frac{d_j}{k_j}, \quad i = 1, 2, \dots, s, \quad (74)$$

where d_j denotes the number of failures that occurred at $x_{(j)}$.

Example 1 (continued). By using the winding data of Table 1, we can estimate via a graphical procedure, the cumulative failure rate of the electromechanical devices at the times $t = 50$ months and $t = 200$ months, under the assumption that failure times follow a two-parameter Weibull distribution. The plotting positions H_i are estimated via the NA estimator (Eq. 74) for each ordered failure time $t_{(i)}$ ($i = 1, 2, \dots, 8$), as given in the fifth column of Table 2 [note that, because there are no tied failure times, the Herd–Johnson-based estimator (Eq. 73) could also be used]. Once points $(t_{(i)}, H_i)$ are plotted on a Weibull hazard plot (i.e., a plot with both log scales) and the straight line is fitted by eye, the graphical estimates of the cumulative failure rate are $\hat{H}(50) = 0.19$ and $\hat{H}(200) = 2.5$, as depicted in Fig. 2. Using the least-squares method, we have $\hat{H}(50) = 0.173$ and $\hat{H}(200) = 2.941$, respectively. Note that, unlike the nonparametric procedure, the use of hazard plots allows $H(t)$ to be estimated also for t values greater than the largest failure or censoring time.

APPENDIX

The estimated $l \times l$ asymptotic variance–covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_l)$ is the inverse of the estimated information matrix $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$, whose entries are the negative second derivatives of the log-likelihood function $\ell(\boldsymbol{\theta}; t)$ with respect to the model parameters $\boldsymbol{\theta}$ evaluated at the ML estimates $\hat{\boldsymbol{\theta}} : \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}$.

For a right censored sample $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ from a two-parameter Weibull

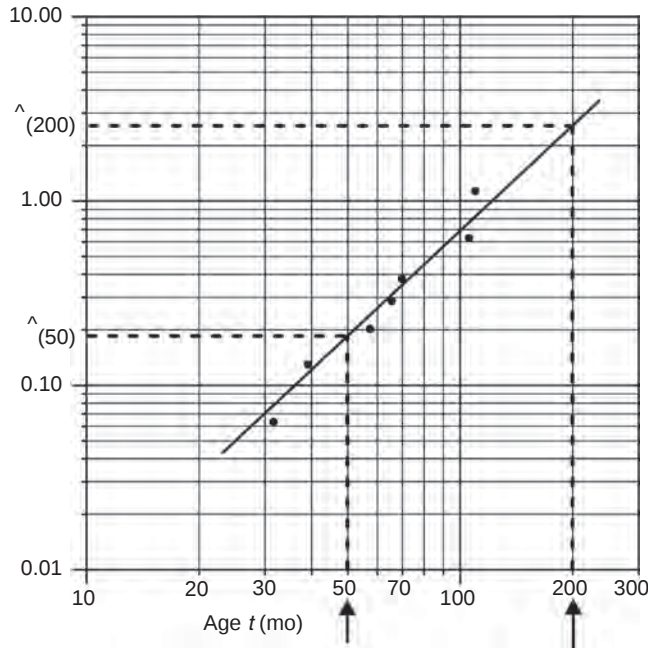


Figure 2. Graphical estimate of $H(50)$ and $H(200)$ in Example 1.

distribution with scale parameter θ and shape parameter β (refer section titled “Data from the Two-Parameter Weibull

Distribution”), the estimated 2×2 information matrix results in

$$\hat{\mathbf{J}}(\hat{\theta}, \hat{\beta}) = \begin{pmatrix} r \left(\frac{\hat{\beta}}{\hat{\theta}} \right)^2 & -\frac{\hat{\beta}}{\hat{\theta}} \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \ln \left(\frac{t_i}{\hat{\theta}} \right) \\ -\frac{\hat{\beta}}{\hat{\theta}} \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \ln \left(\frac{t_i}{\hat{\theta}} \right) & \frac{r}{\hat{\beta}^2} + \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \left[\ln \left(\frac{t_i}{\hat{\theta}} \right) \right]^2 \end{pmatrix}, \quad (75)$$

where r ($r \leq n$) is the number of exact failure times.

REFERENCES

1. Lawless JF. Statistical models and methods for lifetime data. New York: John Wiley & Sons, Inc.; 1982.
2. Nelson W. Applied life data analysis. New York: John Wiley & Sons, Inc.; 1982.
3. Kalbfleisch JD, Prentice RL. The statistical analysis of failure data. New York: John Wiley & Sons, Inc.; 1980.
4. Cox DR, Oakes D. Analysis of survival data. London: Chapman & Hall; 1984.
5. Kleinbaum GD, Klein M. Survival analysis: a self-learning text. 2nd ed. New York: Springer; 2005.
6. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. 3rd ed. New York: Mc Graw-Hill Higher Education; 1974.
7. Müller HG, Wang JL. Density and failure rate estimation. In: Ruggeri F, Kennet RS, Faltin FW, editors. Encyclopedia of statistics in quality and reliability. New York: John Wiley & Sons, Inc.; 2007.
8. Müller HG, Wang JL. Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics 1994;50(1):61–76. DOI: 10.2307/2533197.
9. Silverman BW. Density estimation for statistics and data analysis. London: Chapman and Hall; 1986.
10. Chiang CL. Introduction to stochastic processes in biostatistics. New York: John Wiley & Sons, Inc.; 1968.

11. Wang JL, Müller HG, Capra WB. Analysis of oldest-old mortality: Lifetables revisited. *Ann Stat* 1998;26(1):126–163. DOI: 10.1214/aos/1030563980.
12. Müller HG, Wang JL, Capra WB. From lifetables to hazard rates: the transformation approach. *Biometrika* 1997;84(4):881–892. DOI: 10.1093/biomet/84.4.881.
13. Oehlert GW A note on the delta method. *Am Stat* 1992;46(1):27–29. DOI: 10.2307/2684406.
14. Efron B. The jackknife, the bootstrap, and other resampling plans. Philadelphia (PA): Society of Industrial and Applied Mathematics; 1982.
15. Balakrishnan N, Basu AP. The exponential distribution: theory, methods, and applications. New York: Gordon and Breach; 1996.
16. Miller RG. Survival analysis. New York: John Wiley & Sons, Inc.; 1981.
17. Sprott DA. Normal likelihoods and their relation to large sample theory of estimation. *Biometrika* 1973;60(3):457–465. DOI:10.1093/biomet/60.3.457.
18. Cox DR. Some simple approximate tests for Poisson variates. *Biometrika* 1953;40(3–4): 354–360. DOI: 10.1093/biomet/40.3–4.354.
19. Yang G. Life cycle reliability engineering. New York: John Wiley & Sons, Inc.; 2007.
20. Thomas DR, Bain LJ, Antle CE. Maximum likelihood estimation, exact confidence intervals for reliability, and tolerance limits in the Weibull distribution. *Technometrics* 1970;12(2):363–371.
21. Bain LJ, Engelhardt M. Statistical analysis of reliability and life-testing models: theory and methods. 2nd ed. New York: Marcel Dekker; 1991.
22. Meeker WQ, Nelson W. Weibull variances and confidence limits by maximum likelihood for singly censored data. *Technometrics* 1977;19(4):473–476.

ESTIMATING INTENSITY AND MEAN VALUE FUNCTION

MASSIMILIANO GIORGIO
Department of Aerospace and
Mechanical Engineering,
Second University of
Naples, Aversa, Italy

GIANPAOLO PULCINI
Istituto Motori, National
Research Council (CNR),
Rome, Italy

The *complete* (or conditional) *intensity function* (CIF) of the stochastic point process $\{N(t) : t \geq 0\}$ is defined as

$$\lambda(t; H_t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr\{N(t, t + \Delta t) \geq 1 | H_t\}}{\Delta t}, \quad (1)$$

where $N(t, t + \Delta t)$ denotes the (integer-valued) random variable counting the number of failures that occur in the time interval $(t, t + \Delta t]$, and $H_t = \{N(0, s) : 0 \leq s \leq t\}$ represents the entire *history* of the process up to time t [1]. The way in which the history affects the CIF depends mainly on the kind of repair actions performed on the system. For example, if the system is renewed at each failure, then the history affects the CIF only through the occurrence time of the most recent failure prior to t , say $t_{N(t)}$, and the CIF at time t is a function of the time elapsed from $t_{N(t)}$ until t , say $\lambda(t; H_t) = \lambda(t - t_{N(t)})$. If the system is minimally repaired at failure, then the CIF does not depend on the history of the process but only, at the most, on time t . More complex forms for the CIF arise when the system is subject to imperfect repairs or to a maintenance policy consisting of a sequence of corrective repairs interspersed with major overhauls [2,3].

For small Δt values, the product $\lambda(t; H_t) \cdot \Delta t$ is approximately equal to the conditional probability that at least one failure (not necessarily the first) occurs in the time interval $(t, t + \Delta t]$, given the history up to t [4].

The expectation of the number $N(t) = N(0, t)$ of failures up to t , known as the *mean value function* $M(t) = E\{N(t)\}$, is a nondecreasing right continuous function. Its first derivative $v(t) = dM(t)/dt$, if it exists, is called the *rate of occurrence of failures* (ROCOF) and represents the time rate of change of $M(t)$ [5].

If simultaneous failures cannot occur, so that the arrival times of failures are distinct, then $M(t)$ is continuous everywhere and the point process is said to be *orderly* or *regular* [5,6]. Under orderliness conditions, the ROCOF function $v(t)$, if it exists, numerically equals the (*unconditional*) intensity function $\lambda(t)$ which is defined as

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr\{N(t, t + \Delta t) \geq 1\}}{\Delta t}, \quad (2)$$

and can be viewed as the expectation of the CIF (Eq. 1) over the entire space of possible histories. Of course, under orderliness conditions, the cumulative intensity function $\Lambda(t) = \int_0^t \lambda(z) dz$ numerically equals the mean value function $M(t)$.

If the CIF does not depend on H_t but only, at the most, on the operating time (i.e., the failure probability is unchanged by failure and successive repair), then the two intensity functions (Eqs 1 and 2) coincide. Also, for small Δt , the product $\lambda(t) \cdot \Delta t$ is approximately equal to the (unconditional) probability that at least one failure occurs in the time interval $(t, t + \Delta t]$ [6]. If the point process $\{N(t)\}$ is not orderly, then the ROCOF is greater than the intensity function $\lambda(t)$ [1] and does not specify the process uniquely [4].

The behavior of the intensity function $\lambda(t)$ and of the mean value function $M(t)$ is very representative of the failure/repair process. For example, a decreasing intensity function [or, equivalently, a downward concave $M(t)$] shows that failures become less and less frequent with the operating time t , whereas an increasing intensity function [or, equivalently, an upward concave $M(t)$] gives evidence of more and more frequent failures

with t . These situations are commonly known as *reliability growth* or *improvement* and *reliability decay* or *deterioration*.

Nonparametric and parametric estimation procedures of the intensity function and of the mean value function are available for data arising from both, a single repairable system and a multiple repairable system.

DATA FROM A SINGLE REPAIRABLE SYSTEM

Let $t_1 < t_2 < \dots < t_n \leq T$ denote the n failure times of a repairable system operating up to time T . Suppose that repair times are negligible, or are measured on a different timescale, so that the interarrival times $x_i = t_i - t_{i-1}$ (where $t_0 = 0$ and $i = 1, 2, \dots, n$) represent the operating times between the $(i-1)$ th and i th failures. A number of nonparametric and parametric estimation procedures for $\lambda(t)$ and $M(t)$ are illustrated. The main difference between the two procedures consists of the fact that, unlike the parametric one, the nonparametric procedure does not allow $\lambda(t)$ and $M(t)$ to be estimated for t values outside the observation interval $[0, T]$. In addition, nonparametric estimation generally provides only point estimates of $\lambda(t)$ and $M(t)$, and not interval estimation.

Nonparametric Estimators

A natural nonparametric estimate of $M(t)$ at the time t is given by the observed number of failures that occurred until t , say

$$\tilde{M}(t) = N(t), t \leq T. \quad (3)$$

It is a staircase function, that is, the function $\tilde{M}(t)$ is flat between successive failure times, equal to i ($i = 1, 2, \dots, n$) at $t = t_i$, but in a graphical representation of $\tilde{M}(t)$ only the points (t_i, i) need to be plotted and not the flat portions [7]. Of course, the true mean value function of a large population of identical systems operating under identical working conditions should be regarded as a smooth curve, so that it is useful to imagine a smooth curve which passes through the plotted points (t_i, i) .

From Equation (3), an estimate of the mean ROCOF function over each interval

$(t_{i-1}, t_i]$ ($i = 1, 2, \dots, n$) easily follows:

$$\tilde{v}(t) = \frac{1}{t_i - t_{i-1}}, \quad t_{i-1} < t \leq t_i, \quad i = 1, 2, \dots, n. \quad (4)$$

It is worth noting that, under orderliness conditions, any estimate of the ROCOF provides an estimate of $\lambda(t)$, too; similarly, any estimate of $M(t)$ provides an estimate of the cumulative intensity function $\Lambda(t)$. The graph of the estimator (Eq. 4) generally appears jagged, moving up when two successive failures are close each other, and down when the failures are sparse. A smoother estimator of $v(t)$ can be obtained by partitioning the observation interval $[0, T]$ into k subintervals: $[0, \tau_1], [\tau_1, \tau_2], \dots, [\tau_{k-1}, \tau_k]$, where $0 < \tau_1 < \dots < \tau_{k-1} < \tau_k = T$. These subintervals may be of equal length, but this is not necessary. If $N(\tau_{l-1}, \tau_l) = N(\tau_l) - N(\tau_{l-1})$ denotes the number of failures that occurred in the l th subinterval ($l = 1, 2, \dots, k$), then the estimate of the ROCOF function $v(t)$ is

$$\bar{v}(t) = \frac{N(\tau_{l-1}, \tau_l)}{\tau_l - \tau_{l-1}}, \quad \tau_{l-1} < t \leq \tau_l, l = 1, \dots, k, \quad (5)$$

where $\tau_0 = 0$ and $N(\tau_0) = 0$ [8]. The estimator (Eq. 5) is particularly suitable when the available data are aggregated, that is, they consist of the number of failures occurring in given nonoverlapping intervals, and not in the individual failure times. From Equation (5), an estimator of $M(t)$ suitable for aggregate data follows:

$$\bar{M}(t) = N(\tau_{l-1}) + \frac{t - \tau_{l-1}}{\tau_l - \tau_{l-1}} N(\tau_{l-1}, \tau_l), \quad \tau_{l-1} < t \leq \tau_l, l = 1, \dots, k; \quad (6)$$

this is a piecewise-linear function equal to $N(\tau_l)$ at $t = \tau_l$.

Another nonparametric estimate of $v(t)$, called a *natural estimate* in Ref. 5, is the following:

$$\bar{v}_{\text{nat}}(t) = \frac{N(t, t + \Delta t)}{\Delta t}, \quad (7)$$

which looks at the number $N(t, t + \Delta t)$ of failures that occurred in a moving window of

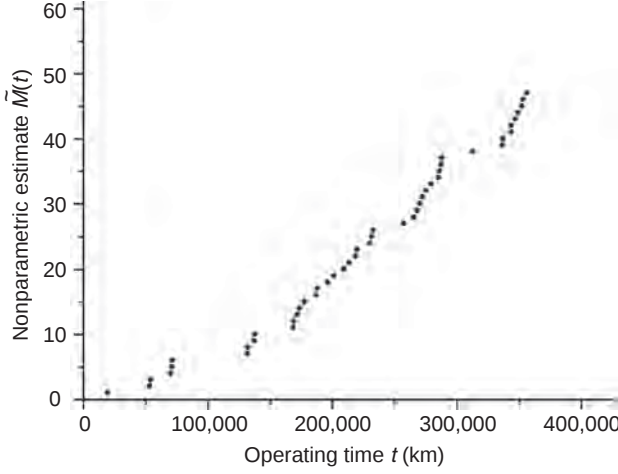


Figure 1. Nonparametric estimate of the mean value function $M(t)$ for the data in Example 1.

prefixed length Δt . When Δt is small, the plot of $\bar{v}_{\text{nat}}(t)$ will appear jagged, moving down when the failures are sparse, and up when the failures are frequent. Of course, the estimator in Equation (7) is seriously biased when t is in the interval $[T - \Delta t, T]$ because, when $t > T - \Delta t$, the interval $[t, t + \Delta t]$ will include times beyond the observation period. Thus, $\bar{v}_{\text{nat}}(t)$ should be calculated only for $0 \leq t \leq T - \Delta t$ [8].

Under orderliness conditions, a kernel smoothing of the estimated increments of $M(t)$ from time t_{i-1} up to t_i ($i = 1, 2, \dots, n$), say $\Delta \tilde{M}_i = \tilde{M}(t_i) - \tilde{M}(t_{i-1}) = 1$, provides a smooth estimate of $v(t)$ and $\lambda(t)$ as given by Lewis and Shedler [9]:

$$\tilde{v}_k(t) = \frac{1}{b_n} \sum_{i=1}^n W\left(\frac{t - t_i}{b_n}\right), \quad (8)$$

where b_n is a sequence of smoothing parameters, known as the *bandwidth*, satisfying $b_n \rightarrow 0$ but $nb_n \rightarrow \infty$ as the sample size $n \rightarrow \infty$, and the *kernel function* $W(z)$ is a bounded nonnegative weight function which satisfies $\int_{-\infty}^{\infty} W(z) dz = 1$. $W(z)$ is typically assumed to be a symmetric probability density function with zero mean and unit variance. For example, $W(z)$ can be the normal density function. The level of smoothness is set by the bandwidth b_n : the larger b_n , the smoother $\tilde{v}_k(t)$ is.

Example 1. Let us consider the failure data of the power-train system of a bus #524 built by Breda Menarinibus which consists of 47 failure times (in km) observed until $T = 373,716$ km [10]. Figure 1 depicts the nonparametric estimate $\tilde{M}(t)$ for such data. The imaginary smooth curve which passes through the plotted points (t_i, i) ($i = 1, 2, \dots, 47$) is upward concave, thus giving graphical evidence of reliability deterioration with t . Note that this analysis (as well as the following one) is for illustration purposes only. Figure 2 gives the plots of nonparametric estimates (Eqs 5 and 7) and of the kernel estimate, Equation (8), of the ROCOF $v(t)$: the plot of estimate 5, obtained by partitioning the interval $[0, T]$ into 8 subintervals of equal length of 46,714.5 km, is much less jagged than the plot of the natural estimate (7), obtained by setting $\Delta t = 20,000$ km. The kernel estimate $\tilde{v}_k(t)$ is obtained by setting $b_n = 20,000$ km, and using a normal density function as the kernel function. Accordingly, with the plot of $\tilde{M}(t)$ in Fig. 1, all the plots of Fig. 2 give evidence of increasing ROCOF (and intensity function) with operating time at least until $t = 200,000$ km, then becoming almost constant.

Parametric Estimators

Parametric estimation procedures imply the choice of a stochastic model able

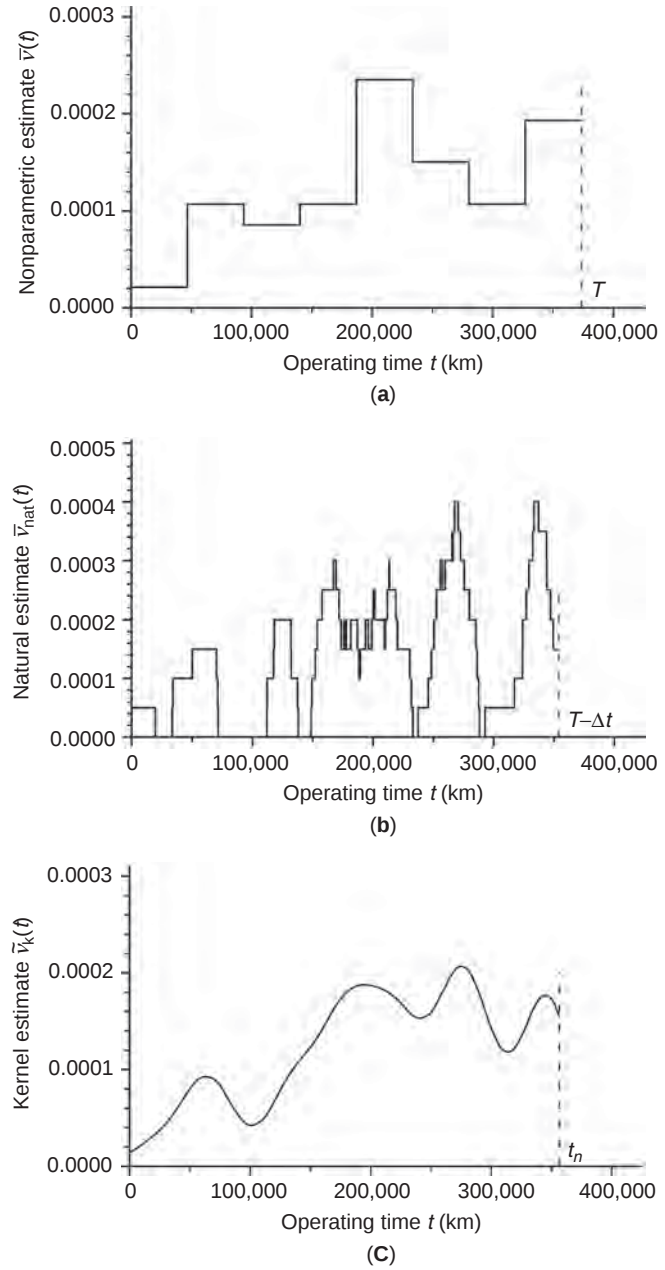


Figure 2. (a) Nonparametric, (b) natural, and (c) kernel estimates of $v(t)$ for the data in Example 1.

to adequately describe the observed failure/repair process. Such a choice should be based on technical information about the type of maintenance policy being employed and its effect on the operating conditions of the system. Of course, the

effect of maintenance actions depends not only on the action that is performed (repair or substitution of the failed part, major overhaul of the system, and so on) but also on the physical structure and complexity of the system.

Parametric estimation, based on the maximum likelihood method, is provided here for two special cases: (i) the system is subject to minimal repair (after repair, the system is in exactly the same condition it was in just before the failure) and its failure pattern does not show any trend with operating time t ; and (ii) the system is minimally repaired but can experience reliability improvement or deterioration with t . In the first case, the failure process is described through a homogeneous Poisson process (HPP) and in the second case a non-homogeneous Poisson process (NHPP) with power-law intensity (called the *Power-Law process*, PLP) is adopted here. Note that there are many choices for the intensity and the PLP is the simplest one for analytical procedures.

Homogeneous Poisson Process. The HPP is a Poisson process with constant intensity function $\lambda(t; H_t) = \lambda(t) = \lambda$, and times between successive failures, say $x_i = t_i - t_{i-1}$ ($i = 1, 2, \dots$), are independent and identically exponentially distributed random variables with mean $1/\lambda$. Since times between failures are independent and identically distributed random variables, then the HPP is also a renewal process. Due to the orderliness property of the Poisson processes, the ROCOF is equal to λ , too, and the mean value function is $M(t) = \lambda t$.

If the data arise from failure-truncated testing, that is, the process is observed up to the occurrence of a prefixed n th failure, the maximum likelihood estimate (MLE) of the intensity function is

$$\hat{\lambda} = n/t_n, \quad (9)$$

where t_n is the last failure time [8]. Since the quantity $2\lambda t_n$ is distributed as a chi-square random variable with $2n$ degrees of freedom, the equal-tail $(1 - \alpha)$ confidence interval for λ is

$$\left(\frac{\chi_{1-\alpha/2}^2(2n)}{2t_n}, \frac{\chi_{\alpha/2}^2(2n)}{2t_n} \right), \quad (10)$$

where $\chi^2(2n)$ denotes the critical value of the chi-square distribution with $2n$ degrees

of freedom. Due to the invariance property of the maximum likelihood estimators, the MLE of the mean value function is

$$\hat{M}(t) = t\hat{\lambda} = nt/t_n, \quad (11)$$

and the equal-tail $(1 - \alpha)$ confidence interval for $M(t)$ is

$$\left(\frac{t\chi_{1-\alpha/2}^2(2n)}{2t_n}, \frac{t\chi_{\alpha/2}^2(2n)}{2t_n} \right). \quad (12)$$

If the process is time truncated, that is, the testing is stopped at a prefixed time T , so that the number n of observed failures is random, the MLE of the intensity function is

$$\hat{\lambda} = n/T, \quad (13)$$

and the approximate equal-tail $(1 - \alpha)$ confidence interval for λ is [11]

$$\left(\frac{\chi_{1-\alpha/2}^2(2n)}{2T}, \frac{\chi_{\alpha/2}^2(2n+2)}{2T} \right). \quad (14)$$

The confidence interval in Equation (14) is conservative in the sense that the probability that it covers the true λ value in repeated sampling is equal to $(1 - \alpha)$ or greater. The MLE of the mean value function is

$$\hat{M}(t) = t\hat{\lambda} = nt/T, \quad (15)$$

and the conservative equal-tail $(1 - \alpha)$ confidence interval for $M(t)$ is

$$\left(\frac{t\chi_{1-\alpha/2}^2(2n)}{2T}, \frac{t\chi_{\alpha/2}^2(2n+2)}{2T} \right). \quad (16)$$

From Equations (11) and (15), we find that the MLE of the mean value function at the testing stopping time, say $\hat{M}(t_n)$ or $\hat{M}(T)$ for failure or time-truncated process, is equal to the observed number n of failures.

Power-Law Process. The Power-Law process is the best known and most frequently used NHPP. Its intensity function is given by Ref. 12 as

$$\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1}. \quad (17)$$

The intensity function (Eq. 17) can increase or decrease with t depending on whether the power-law parameter β is greater or smaller than 1. Thus, the PLP can describe both, reliability deterioration and improvement. If $\beta = 1$, the PLP reduces to the HPP. The expected number of failures $M(t) = (t/\theta)^\beta$ is linear with t on a log-log scale: $\ln[M(t)] = \beta \ln t - \beta \ln \theta$. Hence, if a failure data set follows the PLP, the plot of the cumulative number of failures $N(t_i) = i$ ($i = 1, 2, \dots$) versus the observed failure times t_i should be reasonably close to a straight line with slope β on log-log paper.

If the process is observed up to the occurrence of a prefixed n th failure (failure-truncated testing), the MLE of the PLP parameters is

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \ln(t_n/t_i)} \text{ and } \hat{\theta} = \frac{t_n}{n^{1/\hat{\beta}}}, \quad (18)$$

[12] and the MLE of the intensity function and of the mean value function, at a given time t , result in

$$\hat{\lambda}(t) = \frac{n\hat{\beta}}{t} \left(\frac{t}{t_n} \right)^{\hat{\beta}} \text{ and } \hat{M}(t) = n \left(\frac{t}{t_n} \right)^{\hat{\beta}}. \quad (19)$$

Simulation bounds on the cumulative number of events can be plotted [13] or approximate confidence intervals for $\lambda(t)$ and $M(t)$ can be obtained by applying the Delta method. The latter method allows approximated standard deviation of $\hat{\lambda}(t)$ and $\hat{M}(t)$ to be estimated once the first derivatives of $\lambda(t)$ and $M(t)$, with respect to the PLP parameters, have been evaluated at the MLE

$\hat{\theta}$ and $\hat{\beta}$:

$$\hat{\sigma}(\hat{\lambda}(t)) = \left[\hat{C}_{1,1} \left(\frac{\partial \lambda(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \right)^2 + 2\hat{C}_{1,2} \times \frac{\partial \lambda(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \frac{\partial \lambda(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} + \hat{C}_{2,2} \left(\frac{\partial \lambda(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \right)^2 \right]^{1/2}, \quad (20)$$

$$\hat{\sigma}(\hat{M}(t)) = \left[\hat{C}_{1,1} \left(\frac{\partial M(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \right)^2 + 2\hat{C}_{1,2} \times \frac{\partial M(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \frac{\partial M(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} + \hat{C}_{2,2} \left(\frac{\partial M(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \right)^2 \right]^{1/2}, \quad (21)$$

where $\hat{C}_{l,k}$ ($l, k = 1, 2$) denotes the (l, k) entry in the estimated covariance matrix $\mathbf{C}(\hat{\theta}, \hat{\beta})$ and

$$\begin{aligned} \frac{\partial \lambda(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} &= - \left(\frac{\hat{\beta}}{\hat{\theta}} \right)^2 \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}-1}, \\ \frac{\partial \lambda(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} &= \frac{1}{t} \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}} \left[1 + \hat{\beta} \ln \left(\frac{t}{\hat{\theta}} \right) \right], \\ \frac{\partial M(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} &= - \left(\frac{\hat{\beta}}{\hat{\theta}} \right) \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}}, \\ \frac{\partial M(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} &= \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}} \hat{\beta} \ln \left(\frac{t}{\hat{\theta}} \right). \end{aligned} \quad (23)$$

Hence, according to the approximation chosen for the asymptotic distribution of $\hat{\lambda}(t)$ and $\hat{M}(t)$, namely, the normal or the log-normal approximation, the approximate equal-tail $(1 - \alpha)$ confidence intervals for $\lambda(t)$ and $M(t)$ result in either

$$\hat{\lambda}(t) \pm z_{\alpha/2} \hat{\sigma}(\hat{\lambda}(t)) \text{ and } \hat{M}(t) \pm z_{\alpha/2} \hat{\sigma}(\hat{M}(t)) \quad (24)$$

or

$$\begin{aligned} \hat{\lambda}(t) \cdot \exp\{\pm z_{\alpha/2} \hat{\sigma}(\hat{\lambda}(t))/\hat{\lambda}(t)\} \text{ and} \\ \hat{M}(t) \cdot \exp\{\pm z_{\alpha/2} \hat{\sigma}(\hat{M}(t))/\hat{M}(t)\}, \end{aligned} \quad (25)$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value of the standard normal distribution. Since both $\lambda(t)$ and $M(t)$ are constrained to be nonnegative, the log-normal approximation should work better when the sample size is not large because it avoids negative lower confidence limits to be obtained, as often occurs under the normal approximation for small or moderate sample sizes.

For $t = t_n$, the MLE of the mean number of failures is $\hat{M}(t_n) = n$, that is, it equals the observed number of failures, and the MLE of $\lambda(t_n)$ is

$$\hat{\lambda}(t_n) = \frac{n\hat{\beta}}{t_n}. \quad (26)$$

The estimator in Equation (26) is highly biased for small n [14], since

$$E\{\hat{\lambda}(t_n)\} = \frac{n^2}{(n-1)(n-2)} \hat{\lambda}(t_n), \quad (27)$$

so that an unbiased estimators of $\lambda(t_n)$ results in the following equation [14]:

$$\hat{\lambda}_U(t_n) = \frac{(n-1)(n-2)\hat{\beta}}{nt_n}. \quad (28)$$

Note that the value of the intensity function when testing stops is a very important quantity for a system undergoing a reliability growth program. Indeed, this quantity represents the intensity for the system if it were put into service with no further improvements, and its estimate is generally used to decide when the reliability growth program can be stopped. On the other hand, for a system experiencing reliability deterioration, the estimate of that value of the intensity function can be used to decide when to remove the system from use or when to overhaul it.

By using the result that the quantity $\Omega = 4n^2\lambda(t_n)/\hat{\lambda}(t_n)$ is distributed as the product of two independent chi-square random variables with $2(n-1)$ and $2n$ degrees of freedom [15], respectively, $(1-\alpha)$ confidence intervals for $\lambda(t_n)$ are

$$(\omega_1 \hat{\lambda}(t_n), \omega_2 \hat{\lambda}(t_n)), \quad (29)$$

where ω_1 and ω_2 are, respectively, the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the distribution of $\Omega/(4n^2)$, and can be read from Ref. 8 (Table A.4) for selected values of α and $n = 2(1)100$ (values of $1/\omega_1$ and $1/\omega_2$ can be found in Table 1 of Ref. 16. For large n values, say $n \geq 100$, approximate $(1-\alpha)$ confidence intervals for $\lambda(t_n)$ are given by Crow [16]:

$$\left([1 + z_{\alpha/2}\sqrt{2/n}] \hat{\lambda}(t_n), [1 - z_{\alpha/2}\sqrt{2/n}] \hat{\lambda}(t_n) \right). \quad (30)$$

If testing is stopped at a prefixed time T the MLE of the PLP parameters [12] is

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \ln(T/t_i)} \quad \text{and} \quad \hat{\theta} = \frac{T}{n^{1/\hat{\beta}}}, \quad (31)$$

and the MLE of the intensity function and the mean value function, at a given time t , result in

$$\hat{\lambda}(t) = \frac{n\hat{\beta}}{t} \left(\frac{t}{T} \right)^{\hat{\beta}} \quad \text{and} \quad \hat{M}(t) = n \left(\frac{t}{T} \right)^{\hat{\beta}}. \quad (32)$$

As in the case of failure-truncated testing, approximate confidence intervals for $\lambda(t)$ and $M(t)$ can be obtained by applying the Delta method and using the estimated covariance matrix $\mathbf{C}(\hat{\theta}, \hat{\beta})$.

For $t = T$, the MLE of the mean number of failures is $\hat{M}(T) = n$, that is, it equals the observed number of failures, and the MLE of the intensity function is

$$\hat{\lambda}(T) = \frac{n\hat{\beta}}{T}. \quad (33)$$

Approximate equal-tail $(1-\alpha)$ confidence intervals for $\lambda(T)$ can be obtained from

$$(\rho_1 \hat{\lambda}(T), \rho_2 \hat{\lambda}(T)), \quad (34)$$

where ρ_1 and ρ_2 are listed in Ref. 8 (Table A.5) for selected values of α and $n = 2(1)100$ (values of $1/\rho_1$ and $1/\rho_2$ can be found in Table 2 of Ref. 16. For large n values, say $n \geq 100$, approximate $(1-\alpha)$ confidence

Table 1. Failure Times (in km) of the Power-Train System of Bus #524^a

19,635	53,464	54,125	70,207	71,161	71,687	131,964	132,123	137,412	137,988
168,600	169,111	171,903	173,648	177,711	187,956	188,132	196,231	201,534	209,389
213,735	218,836	220,115	230,287	231,821	233,084	257,455	265,657	268,300	270,535
272,544	275,580	279,850	285,329	286,304	287,481	288,087	312,890	336,773	337,135
343,871	343,945	347,031	349,276	352,414	353,403	356,534	—	—	

^aFrom Ref. 10, and to be read across the row).

intervals for $\lambda(T)$ are given by either of these equations [16]:

$$\left(\left[\frac{n + z_{\alpha/2}^2/4 - \sqrt{n z_{\alpha/2}^2/2 + z_{\alpha/2}^4/16}}{n} \right]^2 \hat{\lambda}(T), \left[\frac{n + z_{\alpha/2}^2/4 + \sqrt{n z_{\alpha/2}^2/2 + z_{\alpha/2}^4/16}}{n} \right]^2 \hat{\lambda}(T) \right) \quad (35)$$

or

$$\left((1 + z_{\alpha/2}/\sqrt{2n})^2 \hat{\lambda}(T), (1 - z_{\alpha/2}/\sqrt{2n})^2 \hat{\lambda}(T) \right). \quad (36)$$

Approximation from Equation (35) provides more accurate upper confidence limits, whereas lower limits are better approximated by Equation (36). A slightly different approximation can be found in Ref. 17.

Example 1 (continued). Let us continue considering the power-train data of Table 1. By assuming that the power-train system was very likely subject to deterioration phenomena with operating time and has undergone minimal repair at failure, the PLP is chosen to analyze the failure data. From Equation (31), the MLE of parameters are $\hat{\beta} = 1.555$ and $\hat{\theta} = 31422$ km. Figure 3 depicts the MLE of $M(t)$ together with the approximate 95% confidence interval, obtained by applying the Delta methods

and using the log-normal approximation for the distribution of $\hat{M}(t)$. The nonparametric estimate (Eq. 3), that is the points (t_i, i) ($i = 1, 2, \dots, 47$), is also depicted. Figure 3 shows a good fit of the PLP model to the observed failures. The approximate 95% confidence interval for $\lambda(T)$, based on the Delta method and the log-normal approximation for the distribution of $\hat{\lambda}(T)$, is compared in Table 2 to the confidence interval provided by Equation (34) and to the asymptotic approximations of Equations (35) and (36).

DATA FROM MULTIPLE REPAIRABLE SYSTEMS

A number of nonparametric and parametric estimation procedures for $\lambda(t)$ and $M(t)$ are available when failure data t_{ij} arise from m identical repairable systems, where t_{ij} denotes the i th failure time ($i = 1, 2, \dots, n_j$) from the j th system ($j = 1, 2, \dots, m$) observed over the interval $[0, T_j]$. Both common and uncommon observation periods for each system are considered.

Nonparametric Estimators

Let z_l ($l = 1, 2, \dots, N'$) denote the ordered *distinct* times at which failures of the observed systems occur, where N' is less than or equal to the total number of failures,

Table 2. Approximate 95% Confidence Intervals on $\lambda(T)$ (1/km) for the Data in Example 1

	Delta Method	Equation (34)	Equation (35)	Equation (36)
Lower limit	0.0001305	0.0001248	0.0001306	0.0001245
Upper limit	0.0002930	0.0002927	0.0002926	0.0002826

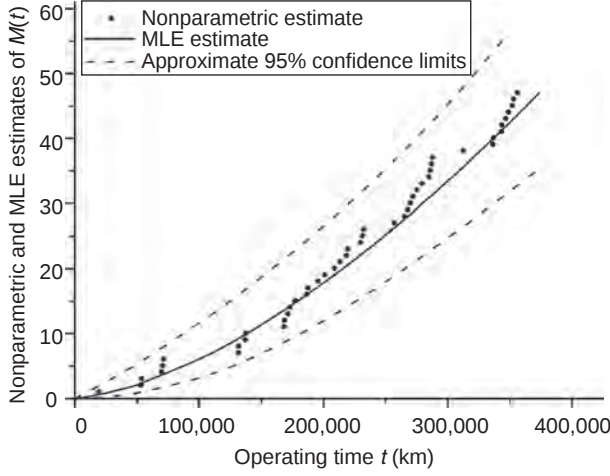


Figure 3. Nonparametric and MLE estimates of the mean value function $M(t)$ for the data in Example 1.

say $N = \sum_{j=1}^m n_j$, since data arise from multiple systems and some failure times can be tied. In addition, let δ_l ($l = 1, 2, \dots, N'$) denote the number of failures having occurred at z_l , so that $\sum_{l=1}^{N'} \delta_l = N$.

If censoring times T_j 's are common over all the systems, say $T_1 = T_2 = \dots = T_m = T$, then a nonparametric estimate of $M(t)$ at the time $t \leq T$ is given by the equation from Crowder *et al.* [18]:

$$\tilde{M}_C(t) = \frac{\sum_{j=1}^m N_j(t)}{m} = \frac{\sum_{l: z_l \leq t} \delta_l}{m}, \quad (37)$$

where $N_j(t)$ is the number of failures experienced by the j th system up to t .

If times T_j s are uncommon, than a nonparametric estimate of $M(t)$ is given by the recursive formula from Ref. 7:

$$\tilde{M}_U(t) = \tilde{M}_U(z_{l-1}) + \frac{\delta_l}{r(z_l)}, \quad z_{l-1} < t \leq z_l, l = 1, 2, \dots, \quad (38)$$

where $z_0 = 0$, $\tilde{M}_U(z_0) = 0$, and $r(z_l) \leq m$ is the number of systems at risk at time z_l which equals the total number of systems minus the number of systems censored before z_l . However, if the censoring time T_j of the j th system does not coincide with a failure time of that or another system, say $z_{l-1} < T_j < z_l$, the survival time interval (z_{l-1}, T_j) of the j th system is neglected by estimator (38),

and then it can overestimate $M(t)$. To take into account such possible survival times, the estimator from Ref. 10

$$\tilde{M}_T(t) = \tilde{M}_T(z_{l-1}) + \frac{\delta_l}{\text{TOT}(z_{l-1}, z_l)/(z_l - z_{l-1})}, \quad z_{l-1} < t \leq z_l, l = 1, 2, \dots, \quad (39)$$

can be used, where $\tilde{M}_T(z_0) = 0$ and $\text{TOT}(z_{l-1}, z_l)$ is the total operating time of all the observed systems during the time interval (z_{l-1}, z_l) . Each ratio $\text{TOT}(z_{l-1}, z_l)/(z_l - z_{l-1})$ represents an equivalent number of systems operating during the interval (z_{l-1}, z_l) ; it equals $r(z_l)$ if no system has been censored in that interval, and is included between $r(z_l)$ and $r(z_{l-1})$ if one or more systems have been censored in (z_{l-1}, z_l) . Then $\tilde{M}_T(t) \leq \tilde{M}_U(t)$, the equality holding when all censoring time prior to t coincide with a failure time.

Estimators (37), (38), and (39) are staircase functions which are flat between successive failure times z_l , but in a graphical representation of $\tilde{M}(t)$ only the points $(z_l, \tilde{M}_\bullet(z_l))$ ($l = 1, 2, \dots, N'$) need to be plotted, and not the flat portions. The true mean value function of a large population of identical systems operating under identical working conditions should be regarded as a smooth curve, so that it is useful to imagine a smooth curve which passes through the plotted points $(z_l, \tilde{M}_\bullet(z_l))$.

From Equations (37), (38), and (39), nonparametric estimates of the mean ROCOF

function over the interval $(z_{l-1}, z_l]$ easily follow, for common and uncommon times T_j s

$$\begin{aligned}\tilde{v}_C(t) &= \frac{\delta_l}{m \cdot (z_l - z_{l-1})}, \\ \tilde{v}_U(t) &= \frac{\delta_l}{r(z_l) \cdot (z_l - z_{l-1})}, \\ \tilde{v}_T(t) &= \frac{\delta_l}{\text{TOT}(z_{l-1}, z_l)}, z_{l-1} < t \leq z_l. \quad (40)\end{aligned}$$

Under orderliness conditions, any estimate of the ROCOF function provides an estimate of $\lambda(t)$ too. The graphs of the estimators (Eq. 40) generally appear jagged, especially when the number of observed systems is not large.

A kernel smoothing of the estimators $\tilde{M}_\bullet(z_l)$ ($l = 1, 2, \dots, N'$) provides smooth estimates of the ROCOF function

$$\begin{aligned}\tilde{v}_{\bullet,k}(t) &= \frac{1}{b_{N'}} \sum_{l=1}^{N'} W\left(\frac{t - z_l}{b_{N'}}\right) \\ &\quad \cdot [\tilde{M}_\bullet(z_l) - \tilde{M}_\bullet(z_{l-1})], \quad (41)\end{aligned}$$

where the *bandwidth* $b_{N'}$ satisfies $b_{N'} \rightarrow 0$ and $N'b_{N'} \rightarrow \infty$ as $N' \rightarrow \infty$, and the *kernel function* $W(z)$ is a bounded nonnegative weight function satisfying $\int_{-\infty}^{\infty} W(z)dz = 1$.

In case of common censoring times T_j 's, the observation interval $[0, T]$ can be partitioned into k subintervals: $[0, \tau_1]$, $(\tau_1, \tau_2]$, $\dots$, $(\tau_{k-1}, \tau_k]$, with $0 < \tau_1 < \dots < \tau_{k-1} < \tau_k = T$, which may be of equal or unequal length. If $N_j(\tau_l)$ denotes the number of failures experienced by the j th system until τ_l ($l = 1, 2, \dots, k$) and $N_j(\tau_{l-1}, \tau_l) = N_j(\tau_l) - N_j(\tau_{l-1})$, then the estimate of $M(t)$ is as follows [19]:

$$\begin{aligned}\bar{M}(t) &= \frac{1}{m} \sum_{j=1}^m N_j(\tau_{l-1}) + \frac{t - \tau_{l-1}}{\tau_l - \tau_{l-1}} \cdot \frac{1}{m} \\ &\quad \times \sum_{j=1}^m N_j(\tau_{l-1}, \tau_l), \tau_{l-1} < t \leq \tau_l, \\ &\quad l = 1, \dots, k. \quad (42)\end{aligned}$$

The estimator (42) is a piecewise-linear function and is particularly suitable when the available data are aggregated; that

is, they consist of the number of failures occurring in given nonoverlapping intervals, common over all the systems, and not in the individual failure times. From Equation (42), a smoother estimator of the mean ROCOF over the subinterval $(\tau_{l-1}, \tau_l]$ easily follows:

$$\begin{aligned}\bar{\mu}(t) &= \frac{1}{\tau_l - \tau_{l-1}} \cdot \frac{1}{m} \sum_{j=1}^m N_j(\tau_{l-1}, \tau_l), \\ \tau_{l-1} < t \leq \tau_l, l &= 1, \dots, k. \quad (43)\end{aligned}$$

Another nonparametric estimate of $\nu(t)$, called a *natural estimate* in Ref. 5, is

$$\bar{\nu}_{\text{nat}}(t) = \frac{1}{r(t)} \cdot \frac{\sum_{j=1}^m N_j(t, t + \Delta t)}{\Delta t}, \quad (44)$$

where $r(t) \leq m$ is the number of systems at risk at the operating time t , and $N_j(t, t + \Delta t)$ is the number of failures experienced by the j th system in a moving window of prefixed length Δt . Of course, $N_j(t, t + \Delta t) = 0$ for $t \geq T_j$. When Δt is small, the plot of $\bar{\nu}_{\text{nat}}(t)$ will appear jagged, moving down when the failures are sparse, and up when the failures are frequent. The estimator (44) is biased when one or more systems are censored in the interval $(t, t + \Delta t]$.

Example 2. Let us consider the failure data (in h) of 4 air-conditioning systems equipping Boeing 720 aircrafts, namely Plane #1, #2, #3, and #4, as given in Table 3 and copied out from a larger data set in Ref. 10. Suppose that the failure processes of Planes #1 and #2 were truncated at the occurrence of the sixth and ninth failure respectively, so that $T_1 \equiv t_{6,1} = 493$ h and $T_2 \equiv t_{9,2} = 1800$ h, whereas Planes #3 and #4 were observed up to time $T_3 = 800$ h and $T_4 = 648$ h respectively; they had experienced 6 and 2 failures, respectively. Then, a total of $N = 23$ no tied-failures were observed over a total operating time of 3741 h, so that $\delta_l = 1$ for $l = 1, 2, \dots, 23$ and $N' \equiv N$. Table 4 gives the nonparametric estimates $\tilde{M}_U(t)$ and $\tilde{M}_T(t)$ (evaluated at the failure times

Table 3. Failure Times (in h) of Four Air-Conditioning Systems of Boeing 720 Aircraft^a

Plane #1	Plane #2	Plane #3	Plane #4
194	359	50	130
209	368	304	623
250	380	309	—
279	650	592	—
312	1253	627	—
493	1256	639	—
—	1360	—	—
—	1362	—	—
—	1800	—	—

^aFrom Ref. 20.

z_l , $l = 1, 2, \dots, 23$). The estimate $\tilde{M}_T(z_l)$ departs from $\tilde{M}_U(z_l)$ as from $l = 18$, since the time interval $(z_{17}, z_{18}) = (639 \text{ h}, 650 \text{ h})$ includes the censoring time $T_4 = 648 \text{ h}$ of Plane #4 and the operating interval (z_{17}, T_4) without failures of Plane #4 is ignored by $\tilde{M}_U(t)$. A further departure occurs as from $l = 19$, since the time interval $(z_{18}, z_{19}) = (650 \text{ h}, 1253 \text{ h})$ includes the censoring time $T_3 = 800 \text{ h}$ of Plane #3. Note that all these analyses are for illustration purposes only.

Parametric Estimators

Maximum likelihood estimation procedures are provided here for two cases; in the first case, the failure patterns are modeled by identical homogeneous Poisson processes, in the second case an NHPP with power-law intensity (called the *PLP*) is adopted.

Homogeneous Poisson Processes. Suppose that m identical repairable systems are observed up to time T_j ($j = 1, 2, \dots, m$) and that n_j failures are experienced by the j th system. If each failure pattern is modeled by an HPP with common constant intensity λ , the MLE of λ is

$$\hat{\lambda} = \frac{\sum_{j=1}^m n_j}{\sum_{j=1}^m T_j}, \quad (45)$$

that is, it is the ratio of the total number of observed failures, say $N = \sum_{j=1}^m n_j$, over the total operating time of all the systems.

If all processes are failure truncated, so that $T_j \equiv t_{n_j, j}$ ($j = 1, 2, \dots, m$) is a random variable, the quantity $2\lambda / \sum_{j=1}^m T_j$ has a chi-square distribution with $2N$ degrees of freedom, and the equal-tail $(1 - \alpha)$ confidence interval for λ is

$$\left(\frac{\chi_{1-\alpha/2}^2(2N)}{2 \sum_{j=1}^m T_j}, \frac{\chi_{\alpha/2}^2(2N)}{2 \sum_{j=1}^m T_j} \right). \quad (46)$$

according to Ref. 8. The MLE of the mean value function is

$$\hat{M}(t) = t\hat{\lambda} = \frac{Nt}{\sum_{j=1}^m T_j}, \quad (47)$$

and the equal-tail $(1 - \alpha)$ confidence interval for $M(t)$ is

$$\left(\frac{t \cdot \chi_{1-\alpha/2}^2(2N)}{2 \sum_{j=1}^m T_j}, \frac{t \cdot \chi_{\alpha/2}^2(2N)}{2 \sum_{j=1}^m T_j} \right). \quad (48)$$

From Equation (47) we have $\sum_{j=1}^m \hat{M}(T_j) = \hat{M}(\sum_{j=1}^m T_j) = N$, that is, the MLE of the mean value function over the entire observation period equals the total number of observed failures.

If some of the systems are time truncated, so that, for some j 's, $T_j > t_{n_j, j}$, the MLE of λ and $M(t)$ are still given by Equations (45) and (47) respectively. In this case, an approximate $(1 - \alpha)$ confidence interval for λ can be obtained estimating the asymptotic standard deviation of $\hat{\lambda}$ by

$$\hat{\sigma}(\hat{\lambda}) = \sqrt{\left[-\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} \Big|_{\hat{\lambda}} \right]^{-1}} = \frac{\hat{\lambda}}{\sqrt{N}}, \quad (49)$$

where the second derivative of the log-likelihood function $\ell(\lambda) = N \ln \lambda + \lambda \sum_{j=1}^m T_j$

Table 4. Nonparametric Estimates of the Mean Value Function $M(t)$ for the Data in Example 2

l	Failure Time z_l (h)	No. $r(z_l)$ at Risk	$1/r(z_l)$	$\tilde{M}_U(z_l)$	$\frac{z_l - z_{l-1}}{\text{TOT}_{(z_{l-1}, z_l]}}$	$\tilde{M}_T(z_l)$
1	50	4	0.250	0.250	0.250	0.250
2	130	4	0.250	0.500	0.250	0.500
3	194	4	0.250	0.750	0.250	0.750
4	209	4	0.250	1.000	0.250	1.000
5	250	4	0.250	1.250	0.250	1.250
6	279	4	0.250	1.500	0.250	1.500
7	304	4	0.250	1.750	0.250	1.750
8	309	4	0.250	2.000	0.250	2.000
9	312	4	0.250	2.250	0.250	2.250
10	359	4	0.250	2.500	0.250	2.500
11	368	4	0.250	2.750	0.250	2.750
12	380	4	0.250	3.000	0.250	3.000
13	493	4	0.250	3.250	0.250	3.250
14	592	3	0.333	3.583	0.333	3.583
15	623	3	0.333	3.917	0.333	3.917
16	627	3	0.333	4.250	0.333	4.250
17	639	3	0.333	4.583	0.333	4.583
18	650	2	0.500	5.083	0.355	4.938
19	1253	1	1.000	6.083	0.801	5.739
20	1256	1	1.000	7.083	1.000	6.739
21	1360	1	1.000	8.083	1.000	7.739
22	1362	1	1.000	9.083	1.000	8.739
23	1800	1	1.000	10.083	1.000	9.739

with respect to λ is to be evaluated at the MLE $\hat{\lambda}$. Then, according to the approximation chosen for the asymptotic distribution of $\hat{\lambda}$, the normal or the log-normal one, the approximate equal-tail $(1 - \alpha)$ confidence intervals for $\lambda(t)$ results in either

$$\hat{\lambda} \pm z_{\alpha/2} \hat{\sigma}(\hat{\lambda}) \quad (50)$$

or

$$\hat{\lambda} \cdot \exp\{\pm z_{\alpha/2} \hat{\sigma}(\hat{\lambda}) / \hat{\lambda}\}. \quad (51)$$

An approximate $(1 - \alpha)$ confidence interval for λ can be also obtained by considering the likelihood ratio test of the null hypothesis $H_0 : \lambda = \lambda_0$ against the alternative hypothesis $H_1 : \lambda \neq \lambda_0$ [8]. Specifically, such a confidence interval for λ consists of all the values of λ_0 for which the null hypothesis is not rejected, that is, all the values of λ_0 for which

the log-likelihood statistic

$$\Lambda(\lambda_0) = -2 \left(N \ln \left[\lambda_0 \cdot \sum_{j=1}^m T_j \right] - N \ln N + N - \lambda_0 \cdot \sum_{j=1}^m T_j \right) \quad (52)$$

is less than the α critical value of the chi-square distribution with one degree of freedom, say $\chi_{\alpha}^2(1)$. In practice, once $\Lambda(\lambda_0)$ has been plotted against λ_0 , the approximate $(1 - \alpha)$ confidence interval for λ is given by all the values of λ_0 for which $\Lambda(\lambda_0) < \chi_{\alpha}^2(1)$ [8].

Once the confidence interval for λ , say $(\lambda_{\alpha/2}, \lambda_{1-\alpha/2})$, has been obtained, an approximate $(1 - \alpha)$ confidence interval for $M(t)$ is given by $(t\lambda_{\alpha/2}, t\lambda_{1-\alpha/2})$.

Example 2 (continued). Let us continue to consider the air-conditioning data of Table 3. By assuming that the

air-conditioning systems operated under identical working conditions, underwent minimal repair at failures, and experienced no trend with operating time, the HPP with intensity λ is used to model their failure process. From Equation (45), the MLE of λ is $\hat{\lambda} = 23/3741 = 0.006148 \text{ h}^{-1}$.

From Equation (49), the estimate of the asymptotic standard deviation of $\hat{\lambda}$ results in $\hat{\sigma}(\hat{\lambda}) = 0.006148/\sqrt{23} = 0.001282 \text{ h}^{-1}$. Choosing the log-normal approximation for the distribution of $\hat{\lambda}$, the approximate 95% confidence interval for λ is $(0.004086 \text{ h}^{-1}, 0.009252 \text{ h}^{-1})$. Alternatively, by using the method based on the log-likelihood statistic (Eq. 52), the approximate 95% confidence limits for λ results in $\lambda_{0.025} = 0.003965 \text{ h}^{-1}$ and $\lambda_{0.975} = 0.009014 \text{ h}^{-1}$.

Power-Law Process. Suppose that m identical repairable systems are observed up to time T_j ($j = 1, 2, \dots, m$) and that n_j failures are experienced by the j th system at times t_{ij} ($i = 1, 2, \dots, n_j$ and $j = 1, 2, \dots, m$). If each failure pattern is modeled by a PLP with common intensity function $\lambda(t) = (\beta/\theta)(t/\theta)^{\beta-1}$, the MLE of parameters are obtained by numerically solving

$$\hat{\beta} = \frac{N}{\sum_{j=1}^m (T_j/\hat{\theta})^{\hat{\beta}} \ln T_j - \sum_{j=1}^m \sum_{i=1}^{n_j} \ln t_{ij}} \text{ and}$$

$$\hat{\theta} = \left(\frac{\sum_{j=1}^m T_j^{\hat{\beta}}}{N} \right)^{1/\hat{\beta}}, \quad (53)$$

where $N = \sum_{j=1}^m n_j$ [12]. The MLE of the intensity function and of the mean value function, at a given time t , result in

$$\hat{\lambda}(t) = \frac{N \hat{\beta} t^{\hat{\beta}-1}}{\sum_{j=1}^m T_j^{\hat{\beta}}} \text{ and } \hat{M}(t) = \frac{N t^{\hat{\beta}}}{\sum_{j=1}^m T_j^{\hat{\beta}}}. \quad (54)$$

From Equation (54), we have $\sum_{j=1}^m \hat{M}(T_j) = N$, that is, the MLE of the mean value function over all the observation periods equals the total number of observed failures.

When $T_1 = T_2 = \dots = T_m = T$, the MLE of $\lambda(t)$ and $M(t)$ result in

$$\hat{\lambda}(t) = \frac{N \hat{\beta} t^{\hat{\beta}-1}}{m T^{\hat{\beta}}} \text{ and } \hat{M}(t) = \frac{N}{m} \left(\frac{t}{T} \right)^{\hat{\beta}}, \quad (55)$$

where $\hat{\beta} = N / \sum_{j=1}^m \sum_{i=1}^{n_j} \ln(T/t_{ij})$ and $\hat{\theta} = T/(N/m)^{1/\hat{\beta}}$.

For equal or unequal censoring times, approximate confidence intervals for $\lambda(t)$ and $M(t)$ can be obtained by applying the Delta method, as illustrated in the section titled "Data from a Single Repairable System" using the first derivatives of $\lambda(t)$ and $M(t)$ given by Equations (22) and (23) respectively, and the estimated covariance matrix $C(\hat{\theta}, \hat{\beta})$.

REFERENCES

1. Cox DR, Isham V. Point processes. London: Chapman and Hall; 1980.
2. Kijima M. Some results for repairable systems with general repair. J Appl Probab 1989;26(1):89–102. DOI: 20.2307/3214319.
3. Pulcini G. Mechanical reliability and maintenance models. In: Pham H, editor. Handbook of reliability engineering. London: Springer; 2003. pp. 317–348.
4. Lawless JF, Thiagarajah K. A point-process model incorporating renewals and time trends, with application to repairable systems. Technometrics 1996;38(2):131–138. DOI: 10.2307/1270406.
5. Ascher H, Feingold H. Repairable systems reliability: modeling, inference, misconceptions and their causes. New York: Marcel Dekker; 1984.
6. Thompson WA. Point process models with applications to safety and reliability. New York: Chapman and Hall; 1988.
7. Nelson W. Graphical analysis of system repair data. J Qual Tech 1988;20(1): 24–35.
8. Rigdon SE, Basu AP. Statistical methods for the reliability of repairable systems. New York: John Wiley & Sons, Inc.; 2000.

9. Lewis PAW, Shedler GS. Statistical analysis of non-stationary series of events in a data base system. *IBM J Res Dev* 1976;20(5): 465–482.
10. Pulcini G. Repairable system analysis for bounded intensity functions and various operating conditions. *J Qual Tech* 2008;40(1): 78–96.
11. Pearson ES, Hartley HO. Volume 1, *Biometrika tables for statisticians*. Cambridge: Cambridge University Press; 1966.
12. Crow LH. Reliability analysis for complex, repairable systems. In: Proschan F, Serfling RJ, editors. *Reliability and biometry*. Philadelphia (PA): SIAM; 1974. pp. 379–410.
13. Walls L, Bendell A. The structure and exploration of reliability field data: What to look for and how to analyse it. *Reliab Eng* 1986;15(2):115–143.
14. Tsokos CP, Rao ANV. Estimation of failure intensity for the Weibull process. *Reliab Eng Syst Saf* 1994;45(3):271–275. DOI: 10.1016/0951-8320(94)90143-0.
15. Lee L, Lee SK. Some results on inference for the Weibull process. *Technometrics* 1978;20(1):41–45. DOI: 10.2307/1268160.
16. Crow LH. Confidence interval procedures for the Weibull process with application to reliability growth. *Technometrics* 1982;24(1): 67–72. DOI: 10.2307/1267580.
17. Bain LJ, Engelhardt M. Inferences on the parameters and current system reliability for a time truncated Weibull process. *Technometrics* 1980;22(3):421–426. DOI: 10.2307/1268327.
18. Crowder MJ, Kimber AC, Smith RL, *et al.* *Statistical analysis of reliability data*. London: Chapman & Hall; 1991.
19. Henderson SG. Estimation for nonhomogeneous Poisson processes from aggregated data. *Oper Res Lett* 2003;31(5):375–382. DOI: 10.1016/S0167-6377(03)00027-0.
20. Cox DR, Lewis PAW. *The statistical analysis of series of events*. London: Chapman & Hall; 1966.

ESTIMATING SURVIVAL PROBABILITY

MASSIMILIANO GIORGIO
Department of Aerospace and
Mechanical Engineering,
Second University of Naples,
Aversa, Italy

GIANPAOLO PULCINI
Istituto Motori, National
Research Council (CNR),
Rome, Italy

The *survival probability* $R(t)$, often called the *reliability function*, is one of the widely used metrics for reliability of nonrepairable units. It is defined as the probability that a product performs its intended function without failure under specified operating conditions for a given period of time t [1]:

$$R(t) = \Pr(T > t), \quad t \geq 0, \quad (1)$$

where T is a continuous nonnegative random variable denoting the time to failure. Accordingly to definition (1), $R(t)$ is a right continuous nondecreasing function of t .

In addition, the conditional survivor function, that is, the probability that a (not-failed) unit of age t survives to w extra units of time, is given by

$$R(w | t) = \Pr(T - t > w | T > t) = \frac{R(t + w)}{R(t)}, \quad (2)$$

being $T - t$ its residual life.

As definition (1) involves the concept of probability, reliability measurement requires probabilistic and statistical methods. A number of classical nonparametric and parametric procedures for estimating the reliability function $R(t)$ of nonrepairable units from homogeneous samples are here illustrated. Methods for performing analysis in presence of heterogeneity are discussed

in Refs 2–5 and Bayesian approaches are treated in Refs 6 and 7.

RELIABILITY DATA DISTINGUISHING FEATURES

What distinguishes reliability data, or more in general lifetime data, from data used in other field of statistics is the frequent presence of *censoring*. To obtain the exact value of the time to failure of a unit it is necessary to continuously monitor the unit all through its life. Typically lifetime samples are incomplete or censored. In fact, very frequently, the lifetime is exactly known only for a portion of the units under study, whereas only incomplete information is available about the failure time of the others. Censoring arises in various ways and many different kinds of censored sampling are considered in literature [2,3,8]. The case of right censored observations is specifically addressed in the following. In general, a sample is called *singly censored* if all the survivors have the same censoring time, whereas it is called *multiply censored* if survivors have different censoring times.

Only standard methods for censored data analysis are presented in this article. The use of these methods requests the assumption that censoring mechanisms are noninformative (or *independent*), that is, the censoring time does not give information about the failure time distribution [9].

NONPARAMETRIC PROCEDURES

Complete and Singly Right Censored Samples

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a complete sample of n data from a single distribution function, and let $\mathbf{x} = \{x_{(1)}, \dots, x_{(h)}\}$ indicate the ordered set of the h distinct values of set $\mathbf{t}$ ($h \leq n$) at which one or more failures occur. From definition (1), a natural nonparametric estimator of the reliability function $R(t)$ at the time t is given by the fraction of units survived beyond t , say:

$$\tilde{R}(t) = \begin{cases} 1, & \text{for } 0 \leq t < x_{(1)} \\ 1 - \sum_{j=1}^i d_j/n, & \text{for } x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, 2, \dots, h-1, \\ 0, & \text{for } t \geq x_{(h)} \end{cases} \quad (3)$$

where $x_{(0)} \equiv 0$, d_j is the number of units failed at $x_{(j)}$ and $n - \sum_{j=1}^i d_j$ is the number of units that survive time $x_{(j)}$. $\tilde{R}(t)$ is a staircase function, that is, it is flat between successive failure times.

In the case of singly censored sample data, the same approach can be adopted. The main feature of singly right censored sample data is that all failure times are smaller than or equal to the unique single censoring time, say $T_C \geq x_{(h)}$. In this case,

$$\tilde{R}(t) = \begin{cases} 1, & \text{for } 0 \leq t < x_{(1)} \\ 1 - \sum_{j=1}^i d_j/n, & \text{for } x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, 2, \dots, h-1, \\ 1 - \sum_{j=1}^h d_j/n, & \text{for } x_{(h)} \leq t \leq T_C \end{cases} \quad (4)$$

where $\sum_{j=1}^h d_j$ is the total number of failed items in the sample. Estimator (4) can be used also in the case of multiply censored samples, provided that all failure times are not greater than the smallest censoring time. In this case, T_C equals the largest censoring time.

In the case of complete or singly censored samples, the exact $(1 - \alpha)$ confidence interval, say $(\tilde{R}_L(t), \tilde{R}_U(t))$, can be obtained by solving the following binomial-based equations [10]:

$$\begin{aligned} \tilde{R}_L(t) &: \sum_{j=0}^{n\tilde{R}(t)-1} \binom{n}{j} (\tilde{R}_L(t))^j (1 - \tilde{R}_L(t))^{n-j} \\ &= 1 - \alpha/2 \end{aligned}$$

$$\begin{aligned} \tilde{R}_U(t) &: \sum_{j=0}^{n\tilde{R}(t)} \binom{n}{j} (\tilde{R}_U(t))^j (1 - \tilde{R}_U(t))^{n-j} \\ &= \alpha/2. \end{aligned} \quad (5)$$

Of course, in case of censoring, the time axis can be investigated up to the censoring time, say $x_{(h)}$ or T_C . Exploiting the relationship between the cumulative binomial function and the cumulative beta distribution, the limits (Eq. 5) can be obtained by inverting the cumulative beta distribution, so that

$$\begin{aligned} \tilde{R}_L(t) &= B_{\alpha/2}(n\tilde{R}, n(1 - \tilde{R}) + 1) \quad \text{and} \\ \tilde{R}_U(t) &= B_{1-\alpha/2}(n\tilde{R} + 1, n(1 - \tilde{R})), \end{aligned} \quad (6)$$

where $B_\gamma(p, q)$ denotes the γ th quantile of the beta distribution with parameters p and q , and $\tilde{R}$ stands for $\tilde{R}(t)$. Alternatively, a conservative $(1 - \alpha)$ confidence interval for the reliability $R(t)$, at any specified time t not greater than the censoring time, is given by Blyth [11]:

$$\begin{aligned} \tilde{R}_L(t) &= \left[1 + \frac{[n(1 - \tilde{R}) + 1]F_{1-\alpha/2}(2n(1 - \tilde{R}) + 2, 2n\tilde{R})}{n\tilde{R}} \right]^{-1} \\ \tilde{R}_U(t) &= \left[1 + \frac{n(1 - \tilde{R})}{(n\tilde{R} + 1)F_{1-\alpha/2}(2n\tilde{R} + 2, 2n(1 - \tilde{R}))} \right]^{-1}, \end{aligned} \quad (7)$$

where $F_\gamma(v_1, v_2)$ is the γ th quantile of the F distribution with v_1 and v_2 degrees of freedom.

Approximate $(1 - \alpha)$ confidence intervals can be also constructed by using the method recommended by Ref. 12:

$$\begin{aligned} \tilde{R}_L(t) &= \frac{\tilde{R} + z_{\alpha/2}^2/(2n) + z_{\alpha/2}\sqrt{\tilde{R}(1 - \tilde{R})/n - z_{\alpha/2}^3/(4n^2)}}{1 + z_{\alpha/2}^2/n} \\ \tilde{R}_U(t) &= \frac{\tilde{R} + z_{\alpha/2}^2/(2n) - z_{\alpha/2}\sqrt{\tilde{R}(1 - \tilde{R})/n - z_{\alpha/2}^3/(4n^2)}}{1 + z_{\alpha/2}^2/n}, \end{aligned} \quad (8)$$

where $z_{\alpha/2}$ is the $(\alpha/2)$ th quantile of the standard normal distribution. This approach is substantiated on the grounds that it is the exact algebraic counterpart of the large-sample hypothesis test. It is also supported by investigations of Agresti and Coull [12]. One advantage of this method is that its worth does not strongly depend upon the value of n and/or $\tilde{R}$, and is recommended for virtually all combinations of n and/or $\tilde{R}$. Another advantage of this method is that the lower limit cannot be negative, although the upper limit can be sometimes larger than 1 for large $\tilde{R}$ and small n values.

As a further alternative, and perhaps the most frequently used, the normal approximation can be adopted:

$$\begin{aligned} \tilde{R}_L(t) &= \tilde{R}(t) + z_{\alpha/2} \tilde{\sigma}_{\tilde{R}} \\ \text{and } \tilde{R}_U(t) &= \tilde{R}(t) - z_{\alpha/2} \tilde{\sigma}_{\tilde{R}}, \end{aligned} \quad (9)$$

where

$$\tilde{\sigma}_{\tilde{R}} = \sqrt{\tilde{R}(1 - \tilde{R})/n} \quad (10)$$

is an estimate of the standard deviation of $\tilde{R}(t)$. For the normal approximation to not be poor, both the inequalities $n\tilde{R} \geq 5$ and $n(1 - \tilde{R}) \geq 5$ should be satisfied.

For small n values, better results are generally obtained if the normal approximation for $\text{logit}(\tilde{R}(t)) = \ln[(1 - \tilde{R}(t)) / \tilde{R}(t)]$ is used, which gives the following $(1 - \alpha)$ confidence interval [5]:

$$\tilde{R}_L(t) = \frac{\tilde{R}}{\tilde{R} + (1 - \tilde{R}) \times w_{\alpha/2}(\tilde{R})}$$

and

$$\tilde{R}_U(t) = \frac{\tilde{R}}{\tilde{R} + (1 - \tilde{R})/w_{\alpha/2}(\tilde{R})}, \quad (11)$$

where $w_{\alpha/2}(\tilde{R}) = \exp\{-z_{\alpha/2} \tilde{\sigma}_{\tilde{R}} / [\tilde{R}(1 - \tilde{R})]\}$. If the standard deviation $\sigma_{\tilde{R}}$ is estimated through Equation (10), we have that $w_{\alpha/2}(\tilde{R}) = \exp\{-z_{\alpha/2} / \sqrt{n\tilde{R}(1 - \tilde{R})}\}$. The interval in Equation (11), different from the interval in Equation (9), surely satisfies the constraints $0 \leq \tilde{R}_L(t), \tilde{R}_U(t) \leq 1$.

All the above-mentioned procedures for constructing confidence intervals for $R(t)$ can also be used in the case of multiply censored samples, provided that censoring times are equal to or greater than the largest failure time $x_{(h)}$. In this case, the time axis can be investigated until the largest censoring time.

Example 1. Let us consider the Type I censored data given in Ref. 13, which refers to 15 units of an electromechanical module tested at the temperature of 125°C. All the tests were stopped at the time $T_C = 1100$ h and eight failures were observed at no tied times, which, once ordered, are

$$\begin{aligned} x_{(1)} &= 88 \text{ h}, & x_{(2)} &= 105 \text{ h}, & x_{(3)} &= 141 \text{ h}, \\ x_{(4)} &= 344 \text{ h}, & x_{(5)} &= 430 \text{ h}, & x_{(6)} &= 516 \text{ h}, \\ x_{(7)} &= 937 \text{ h}, & x_{(8)} &= 1057 \text{ h}. \end{aligned}$$

The nonparametric estimate (4) of the reliability in each time interval $[x_{(i-1)}, x_{(i)}]$ is given in Table 1 and depicted in Fig. 1. Exact confidence intervals for the reliability at any fixed time $t \leq T_C$ are obtained from Equation (6). For example, for $t = 250$ h, the exact 95% confidence interval is

$$\tilde{R}_L(250) = 0.519 \quad \text{and} \quad \tilde{R}_U(250) = 0.957.$$

By using Equation (7), a conservative 95% confidence interval for the reliability at time $t = 250$ h is

Table 1. Nonparametric Estimate of the Reliability Function in Each Time Interval $[x_{(i-1)}, x_{(i)}]$, $i = 1, 2, \dots, 9$, for Example 1

i	$[x_{(i-1)}, x_{(i)}]$	$\tilde{R}(t)$
1	[0, 88)	1.000
2	[88, 105)	0.933
3	[105, 141)	0.867
4	[141, 344)	0.800
5	[344, 430)	0.733
6	[430, 516)	0.667
7	[516, 937)	0.600
8	[937, 1057)	0.533
9	[937, 1100]	0.467

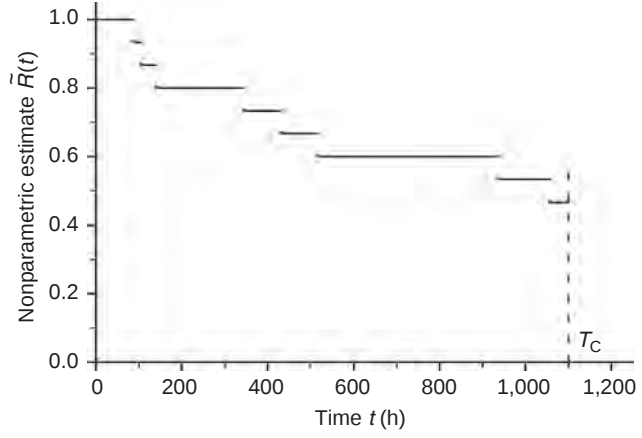


Figure 1. Plot of the nonparametric estimate $\tilde{R}(t)$ of the reliability function in Example 1.

$$\tilde{R}_L(250) = \left[1 + \frac{(15 \times 0.200 + 1) \times 2.779}{15 \times 0.800} \right]^{-1} = 0.519$$

and

$$\tilde{R}_U(250) = \left[1 + \frac{15 \times 0.200}{(15 \times 0.800 + 1) \times 5.097} \right]^{-1} = 0.957,$$

being $F_{1-\alpha/2}(2n(1-\tilde{R})+2, 2n\tilde{R}) = F_{0.975}(8, 24) = 2.779$ and $F_{1-\alpha/2}(2n\tilde{R}+2, 2n(1-\tilde{R})) = F_{0.975}(26, 6) = 5.097$. The normal approximate confidence interval (Eq. 9) provides:

$$\tilde{R}_L(250) = 0.800 + z_{0.025} \sqrt{\frac{0.800(1-0.800)}{15}} = 0.598$$

and

$$\tilde{R}_U(250) = 0.800 - z_{0.025} \sqrt{\frac{0.800(1-0.800)}{15}} = 1.002,$$

which, however, gives an upper limit $\tilde{R}_U(250)$ greater than 1, that is outside the possible range of values for $R(t)$. Thus, $\tilde{R}_U(250) = 1$ must be adopted. This undesirable result is due to the fact that the normal approximation here is poor for $t = 250$ h because, for $n = 15$ and $\tilde{R}(250) = 0.800$, the conditions $n\tilde{R}(250) \geq 5$ and $n[1-\tilde{R}(250)] \geq 5$ are not both satisfied, being $n[1-\tilde{R}(250)] = 3$. Finally, the approximate confidence interval

(Eq. 11) based on the logit transformations are

$$\begin{aligned} \tilde{R}_L(250) &= \frac{\tilde{R}}{\tilde{R} + (1-\tilde{R}) \times w_{0.975}(\tilde{R})} \\ &= \frac{0.800}{0.800 + (1-0.800) \times 3.544} \\ &= 0.530 \end{aligned}$$

and

$$\begin{aligned} \tilde{R}_U(250) &= \frac{0.800}{0.800 + (1-0.800)/3.544} \\ &= 0.934, \end{aligned}$$

where

$$\begin{aligned} w_{0.025}(\tilde{R}) &= \exp\{1.960/\sqrt{15 \cdot 0.800(1-0.800)}\} \\ &= 3.544. \end{aligned}$$

Multiply Censored Samples with Intermixed Failure and Censoring Times

Multiply censored samples are usually characterized by the fact that failure and censoring times are intermixed. In this case, the nonparametric estimation procedure (Eq. 4) cannot be applied and *ad hoc* methods must be used; among them, one of the most widely used is the Kaplan–Meier, or *product limit*, estimation method [1].

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a random sample of size n consisting of r failure times and $n-r$ censored observations, and let $\mathbf{x} = \{x_{(1)}, x_{(2)}, \dots, x_{(h)}\}$ denote the ordered

set of the h ($h \leq r$) distinct times at which one or more failures occur. Moreover, let d_i ($i = 1, 2, \dots, h$) denote the number of failures that occurred at the same time $x_{(i)}$ ($\sum_{i=1}^h d_i = r$), let C_i denote the total number of censorings that occurred before time $x_{(i)}$, and let k_i indicate the number of units at risk (neither failed nor withdrawn) immediately before $x_{(i)}$; $k_1 = n - C_1$ and

$$k_i = n - \sum_{j=1}^{i-1} d_j - C_i, \quad i = 2, \dots, h. \quad (12)$$

In the case of a tie between a censoring time and the generic failure time $x_{(i)}$, the Kaplan–Meier method assumes that the censoring time is infinitesimally larger than $x_{(i)}$, so that the censored unit is still at risk at $x_{(i)}$. Under this assumption, the Kaplan–Meier estimate of the reliability function is

$$\begin{aligned} \bar{R}(t) &= \begin{cases} 1, & 0 \leq t < x_{(1)} \\ \prod_{j=1}^i \frac{k_j - d_j}{k_j}, & x_{(i)} \leq t < x_{(i+1)}, \\ & i = 1, \dots, h-1, \\ \prod_{j=1}^h \frac{k_j - d_j}{k_j}, & x_{(h)} \leq t \leq t_{\max} \end{cases} \quad (13) \end{aligned}$$

where $t_{\max}$ is the largest value in $\{t_1, t_2, \dots, t_n\}$.

In presence of multiply censoring with intermixed failure and censoring times, an estimate of the standard deviation $\sigma_{\bar{R}}$ of $\bar{R}(t)$ may be obtained via the Greenwood's formula [2]:

$$\sigma_{\bar{R}} = \bar{R}(t) \sqrt{\sum_{j: x_{(j)} \leq t} \frac{d_j}{k_j(k_j - d_j)}}, \quad (14)$$

and approximate confidence intervals for $\bar{R}(t)$, at any fixed time $t \leq t_{\max}$, can be obtained through Equations (9) or (11), by using $\bar{R}$ and Equation (14) instead of $\tilde{R}$ and Equation (10), respectively. Note that, when there are no censored observations before a given time t , Equation (14) is numerically equal to Equation (10) up to that t value.

For large samples, suitable alternatives to the Kaplan–Meier estimator (13) are

the *actuarial* estimators, which are usually adopted to face situations in which the exact time to failure or the censoring time of some units is not available: it is only known that failure or censoring occurred after a some time point and before a later time point (i.e., interval-censored data). Nevertheless, especially in the case of large sample size, the actuarial estimators are often used also when the exact failure times are known in order to gain computational advantage (i.e., grouping of data is time saving). To perform the actuarial estimate, the whole interval $(0, t_{\max})$ has to be partitioned into a number m of subintervals, say $(\tau_{i-1}, \tau_i]$ ($i = 1, 2, \dots, m$), being $\tau_0 \equiv 0$ and $\tau_m = t_{\max}$. These subintervals may be of equal length, but this is not necessary. They only have to be contiguous.

Let c_i and d_i indicate, respectively, the number of units withdrawn from the test during the i th subinterval $(\tau_{i-1}, \tau_i]$ and the number of failures that occurred in $(\tau_{i-1}, \tau_i]$. Then, the number k_i ($i = 1, 2, \dots, m$) of units at risk at the beginning of the i th subinterval $(\tau_{i-1}, \tau_i]$ is $k_1 \equiv n$ and

$$k_i = n - \left(\sum_{j=1}^{i-1} d_j + \sum_{j=1}^{i-1} c_j \right), \quad i = 2, \dots, m.$$

Within such a notation, the *simple* actuarial estimate of the reliability function is

$$\tilde{R}_S(\tau_i) = \prod_{j=1}^i \left(1 - \frac{d_j}{k_j} \right), \quad i = 1, 2, \dots, m. \quad (15)$$

In order to take into account that the c_j units withdrawn from the test during the subinterval $(\tau_{j-1}, \tau_j]$ remain at risk only for a part of this subinterval, the *adjusted* actuarial estimator can be used [1]:

$$\tilde{R}_A(\tau_i) = \prod_{j=1}^i \left(1 - \frac{d_j}{k_j - c_j/2} \right), \quad i = 1, 2, \dots, m, \quad (16)$$

where the adjustment factor $-c_j/2$ arises from the assumption that c_j units run about half the subinterval $(\tau_{j-1}, \tau_j]$. Other

Table 2. Failure and Censoring Data of Diesel Generator Fans in Example 2

i	$x_{(i)}$	d_i	C_i	k_i
1	450	1	0	70
2	1150	2	1	68
3	1600	1	2	65
4	2070	2	11	55
5	2080	1	11	53
6	3100	1	16	47
7	3450	1	17	45
8	4600	1	27	34
9	6100	1	35	25
10	8750	1	51	8

censoring adjustments have been proposed in the literature [14].

In contrast to the Kaplan–Meier estimator, the actuarial estimators give reliability estimates at the endpoints of the subintervals. If the i th subinterval contains one or more failures, then $\tilde{R}_S(t)$ and $\tilde{R}_A(t)$ are to be considered undefined within this subinterval. If no failure occurs in (τ_{i-1}, τ_i) , then $\tilde{R}_S(t) = \tilde{R}_S(\tau_{i-1})$ and $\tilde{R}_A(t) = \tilde{R}_A(\tau_{i-1})$ for any $\tau_{i-1} \leq t < \tau_i$. As the lengths of the subintervals approach zero, the simple actuarial estimator (15) and the Kaplan–Meier estimator (13) coincide.

As in the case of the Kaplan–Meier estimator, confidence intervals for $R(t)$ can be computed through Equations (9) or (11). The Greenwood’s formula (14) is used to estimate the standard deviation of $\tilde{R}_S(t)$ or $\tilde{R}_A(t)$, where $k_j - c_j/2$ ($j = 1, 2, \dots, m$) should be used instead of k_j for estimating $\tilde{\sigma}_{\tilde{R}_A}$, so that

$$\tilde{\sigma}_{\tilde{R}_S} = \tilde{R}_S(\tau_i) \sqrt{\sum_{j=1}^i \frac{d_j}{k_j(k_j - d_j)}}$$

and

$$\tilde{\sigma}_{\tilde{R}_A} = \tilde{R}_A(\tau_i) \sqrt{\sum_{j=1}^i \frac{d_j}{(k_j - c_j/2)(k_j - c_j/2 - d_j)}}. \quad (17)$$

Example 2. Let us consider the data of $n = 70$ fans of diesel generators given by

Ref. 1. These data consist of the $r = 12$ failure times (in hours) of the failed fans and of the $n - r = 58$ running times (in hours) of the unfailed fans. The largest time is a censoring time, equal to $t_{\max} = 11,500$ h. Table 2 gives the distinct (ordered) times $x_{(i)}$ ($i = 1, 2, \dots, 10$) at which failures occurred (two pairs of failure times are recorded as equal), together with the number d_i of fans failed at $x_{(i)}$, the total number C_i of fans censored before $x_{(i)}$, and the number k_i of fans at risk immediately before $x_{(i)}$. Table 3 gives the Kaplan–Meier estimate (13) of the reliability function for each time interval between two successive failure times, together with the standard deviation $\bar{\sigma}_{\bar{R}}$ estimated from Equation (14). In Fig. 2 the reliability estimate $\bar{R}(t)$ is also depicted.

Approximate 90% confidence intervals for the reliability function at any time $t \leq t_{\max} = 11,500$ h are given in Table 3 by using the normal approximation of Equation (9) or the logit transformation of Equation (11). For example, at $t = 3800$ h, the normal approximation provides

$$\begin{aligned} \bar{R}_L(3800) &= \bar{R}(3800) + z_{0.05} \times \bar{\sigma}_{\bar{R}(3800)} \\ &= 0.852 - 1.645 \times 0.0460 \\ &= 0.777 \end{aligned}$$

and

$$\begin{aligned} \bar{R}_U(3800) &= \bar{R}(3800) - z_{0.05} \times \bar{\sigma}_{\bar{R}(3800)} \\ &= 0.852 + 1.645 \times 0.0460 = 0.928, \end{aligned}$$

being $\bar{R}(3800) = 0.852$ and $\bar{\sigma}_{\bar{R}(3800)} = 0.0460$, whereas the logit transformation gives:

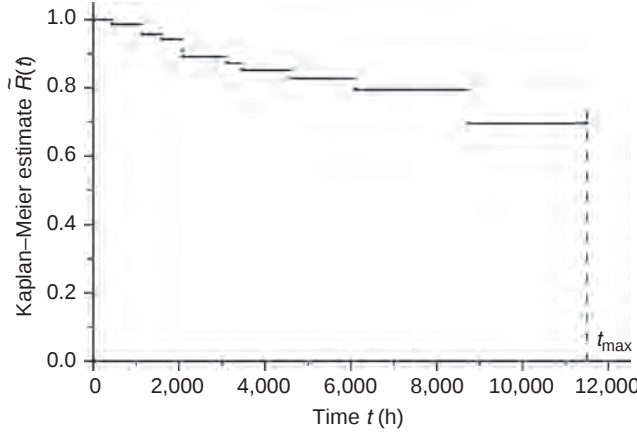
$$\begin{aligned} \bar{R}_L(3800) &= \frac{\bar{R}(3800)}{\bar{R}(3800) + (1 - \bar{R}(3800)) \times w_{0.05}(\bar{R})} \\ &= \frac{0.852}{0.852 + 0.148 \times 1.823} = 0.760 \end{aligned}$$

and

$$\begin{aligned} \bar{R}_U(3800) &= \frac{\bar{R}(3800)}{\bar{R}(3800) + (1 - \bar{R}(3800))/w_{0.05}(\bar{R})} \\ &= 0.913, \end{aligned}$$

Table 3. Kaplan–Meier Estimate $\bar{R}(t)$ of the Reliability Function, Estimate of its Standard Deviation $\sigma_{\bar{R}}$, and Approximate 90% Confidence Intervals for the Data in Example 2

i	Time Intervals (in hours)	$\bar{R}(t)$	$\sigma_{\bar{R}}$	90% Confidence Intervals	
				Normal Approximation	Logit Transformation
1	[0, 450)	1.000	0.0000	(1.000, 1.000)	(1.000, 1.000)
2	[450, 1150)	0.968	0.0142	(0.962, 1.009)	(0.929, 0.997)
3	[1150, 1600)	0.957	0.0244	(0.916, 0.997)	(0.893, 0.983)
4	[1600, 2070)	0.942	0.0282	(0.896, 0.988)	(0.874, 0.974)
5	[2070, 2080)	0.908	0.0361	(0.848, 0.967)	(0.829, 0.952)
6	[2080, 3100)	0.891	0.0392	(0.826, 0.955)	(0.808, 0.941)
7	[3100, 3450)	0.872	0.0427	(0.801, 0.942)	(0.784, 0.927)
8	[3450, 4600)	0.852	0.0460	(0.777, 0.928)	(0.760, 0.913)
9	[4600, 6100)	0.827	0.0510	(0.743, 0.911)	(0.727, 0.896)
10	[6100, 8750)	0.794	0.0587	(0.698, 0.891)	(0.681, 0.875)
11	[8750, 11500]	0.695	0.1061	(0.520, 0.869)	(0.500, 0.838)

**Figure 2.** Plot of the Kaplan–Meier estimate $\bar{R}(t)$ of the reliability function in Example 2.

where $w_{0.05}(\bar{R}) = \exp[1.645 \times 0.046 / (0.852 \times 0.148)] = 1.823$. Note that, for $450 \text{ h} \leq t < 1150 \text{ h}$, the normal approximation provides an upper limit for $R(t)$ greater than 1, and then the value $\bar{R}_L(t) = 1$ must be assumed here.

PARAMETRIC PROCEDURES

Parametric estimation procedures request that a probability model is assumed for the failure time distribution, say

$$T \sim f(t; \theta), \quad t \geq 0,$$

where $\theta = \{\theta_1, \theta_2, \dots, \theta_l\}$ is the vector of l model parameters, some of which are

unknown. Among all the parametric estimation procedures, only the maximum likelihood (ML) estimation method is here discussed, and the specific cases of the exponential and Weibull distributions are considered in detail.

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ denote a random sample of size n of exact failure times and right censored data from the failure time distribution $f(t; \theta)$. The idea behind ML estimation is that the value of model parameters, which maximizes the probability (likelihood) of the sample data, is their best estimate. In this context, the (joint) data probability function is called the *likelihood function*, say $L(\theta; \mathbf{t})$, to highlight that it is regarded as a function of the unknown parameters. Since

multiplying the joint distribution function by a constant does not alter the ML estimates, the likelihood function is usually defined up to a *constant* multiplicative factor (where the term *constant* is used to mean that it does not depend on the unknown parameters, even if it may possibly depend on the data). Under such hypotheses, the likelihood function $L(\boldsymbol{\theta}; \mathbf{t})$ is

$$L(\boldsymbol{\theta}; \mathbf{t}) = \prod_{i=1}^n f(t_i; \boldsymbol{\theta})^{\delta_i} R(t_i; \boldsymbol{\theta})^{1-\delta_i}, \quad (18)$$

where $\delta_i = 1$ if t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation.

The ML estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l\}$, can be obtained (if it exists) as

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{t}),$$

being Θ the space of the parameters. Then, the ML estimate of the reliability function at any time $t \geq 0$ can be easily obtained as

$$\hat{R}(t) = R(t; \hat{\boldsymbol{\theta}}). \quad (19)$$

If an analytic closed form expression exists for $R(t)$, it is generally possible to reparameterize the distribution function so that the reliability value $R_x = R(x)$ at whatever prefixed time x explicitly appears as a parameter of the probability model (see section titled “Data from the Two-Parameter Weibull Distribution”). In this case, the direct estimation of $R_x = R(x)$ can be performed. Owing to the invariance property of the ML method, such an estimate will coincide with the indirect estimate (19).

In performing the ML estimate, it is generally much more convenient to work with the log-likelihood function:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{t}) &= \ln L(\boldsymbol{\theta}; \mathbf{t}) = \sum_{i=1}^n [\delta_i \ln f(t_i; \boldsymbol{\theta}) \\ &\quad + (1 - \delta_i) \ln R(t_i; \boldsymbol{\theta})]. \end{aligned} \quad (20)$$

Indeed, since the logarithm function is a monotone increasing function, the value $\hat{\boldsymbol{\theta}}$ that maximizes $\ell(\boldsymbol{\theta}; \mathbf{t})$, if it exists, is the same that maximizes $L(\boldsymbol{\theta}; \mathbf{t})$. The advantage in maximizing the log-likelihood function is

that the logarithmic transformation usually yields simplified mathematics, since it transforms products in sums. Moreover, it often prevents such computational problems as overflows and/or underflows, often caused by Equation (18).

Approximate confidence intervals for $R(t)$, homolog to the confidence intervals (Eq. 9) for nonparametric estimates, can be obtained adopting the normal approximation for the distribution of $\hat{R}(t)$, so that

$$\hat{R}_L(t) = \hat{R}(t) + z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{R}(t))}$$

and

$$\hat{R}_U(t) = \hat{R}(t) - z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{R}(t))}, \quad (21)$$

where $z_{\alpha/2}$ is the $\alpha/2$ th quantile of the standard normal distribution and $\hat{\text{Var}}(\hat{R}(t))$ denotes the estimate of the asymptotic variance of $\hat{R}(t)$. Such a variance can be estimated through the delta method [15]:

$$\begin{aligned} \hat{\text{Var}}(\hat{R}(t)) &= \sum_{i=1}^l \left[\frac{\partial R(t)}{\partial \theta_i} \Big|_{\hat{\boldsymbol{\theta}}} \right]^2 \hat{\text{Var}}(\hat{\theta}_i) + 2 \sum_{i=1}^l \sum_{j=i+1}^l \\ &\quad \times \frac{\partial R(t)}{\partial \theta_i} \Big|_{\hat{\boldsymbol{\theta}}} \frac{\partial R(t)}{\partial \theta_j} \Big|_{\hat{\boldsymbol{\theta}}} \hat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j), \end{aligned} \quad (22)$$

where $\hat{\text{Var}}(\hat{\theta}_i)$ and $\hat{\text{Cov}}(\hat{\theta}_i, \hat{\theta}_j)$ ($i, j = 1, \dots, l$) are the elements (i, i) and (i, j) , respectively, of the estimated asymptotic variance-covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ of the parameter vector $\boldsymbol{\theta}$ (see section titled “Appendix”), and the partial derivatives of $R(t)$ with respect to the model parameters must be evaluated at the ML estimate $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l\}$.

Alternative to the normal approximation Equation (21), which does not work well for small and moderate samples, confidence limits based on the logit transformation can be evaluated:

$$\begin{aligned} \hat{R}_L(t) &= \frac{\hat{R}}{\hat{R} + (1 - \hat{R}) \times w_{\alpha/2}(\hat{R})} \quad \text{and} \\ \hat{R}_U(t) &= \frac{\hat{R}}{\hat{R} + (1 - \hat{R})/w_{\alpha/2}(\hat{R})}, \end{aligned} \quad (23)$$

where

$$w_{\alpha/2}(\hat{R}) = \exp \left[-z_{\alpha/2} \frac{\sqrt{\hat{\text{Var}}(\hat{R}(t))}}{\hat{R}(t)(1 - \hat{R}(t))} \right], \quad (24)$$

which assures, unlike Equation (21), that the constraints $0 \leq \hat{R}_L(t), \hat{R}_U(t) \leq 1$ are satisfied for any $t > 0$, even for small samples.

Note that, if the log-likelihood function can be reparameterized in terms of the reliability value $R_x = R(x)$ at the specific time x of interest and in terms of the $(l - 1)$ remaining parameters, say $\{\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_l\}$, then $R(x)$ becomes one of the model parameters, say $R(x) \equiv \theta_k$, and the estimated variance of $\hat{R}(x)$ is just the element (k, k) of the matrix $\hat{\mathbf{V}}(\hat{\theta})$.

Also the approximate confidence interval (Eq. 23) can work bad for small samples. More accurate approximate confidence intervals can be obtained adopting the likelihood ratio or a simulation based approach [5]. Moreover, exact solutions for the interval estimation problem are sometimes available, depending on the lifetime distribution model and the censoring scheme. In the following, some procedures for the one-parameter (or standard) exponential and two-parameter Weibull distributions are presented.

Data from the One-Parameter Exponential Distribution

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ be a random sample of size n consisting of exact failure and right censored data from a one-parameter exponential random variable with density function:

$$f(t; \lambda) = \lambda \exp(-\lambda t), \quad \lambda > 0, t \geq 0, \quad (25)$$

and reliability function:

$$R(t; \lambda) = \exp(-\lambda t), \quad \lambda > 0, t \geq 0. \quad (26)$$

The likelihood function $L(\lambda; \mathbf{t})$ relative to such a sample is

$$L(\lambda; \mathbf{t}) \propto \prod_{i=1}^n \lambda^{\delta_i} \exp(-\lambda \delta_i t_i) \times \exp[-\lambda(1 - \delta_i)t_i]$$

$$\begin{aligned} &= \prod_{i=1}^n \lambda^{\delta_i} \exp(-\lambda t_i) \\ &= \lambda^r \exp \left(-\lambda \sum_{i=1}^n t_i \right), \end{aligned}$$

where $\delta_i = 1$ if t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation. Thus, $\sum_{i=1}^n \delta_i = r$ is the number of exact failure times and the sum $\sum_{i=1}^n t_i$ is the total time on test. The log-likelihood function $\ell(\lambda; \mathbf{t})$ is

$$\begin{aligned} \ell(\lambda; \mathbf{t}) &= \ln \left[\lambda^r \exp \left(-\lambda \sum_{i=1}^n t_i \right) \right] \\ &= r \ln \lambda - \lambda \sum_{i=1}^n t_i, \end{aligned}$$

and, for $r > 0$, the ML estimate of the parameter λ (which represents the constant hazard rate) is equal to the number of observed failures over the total time on test [2]:

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^n t_i}. \quad (27)$$

The ML estimate of the reliability function $R(t; \lambda)$ can be straightforwardly obtained as

$$\begin{aligned} \hat{R}(t) &= R(t; \hat{\lambda}) = \exp(-\hat{\lambda} t) \\ &= \exp \left(-\frac{rt}{\sum_{i=1}^n t_i} \right), \quad t \geq 0. \end{aligned} \quad (28)$$

In the case of complete or Type II censored sample, the exact $(1 - \alpha)$ confidence interval for λ is given by Balakrishnan and Basu [16]:

$$(\hat{\lambda}_L, \hat{\lambda}_U) = \left(\frac{\hat{\lambda} \chi_{\alpha/2}^2(2r)}{2r}, \frac{\hat{\lambda} \chi_{1-\alpha/2}^2(2r)}{2r} \right), \quad (29)$$

where $\chi_{\gamma}^2(\nu)$ is the γ th quantile of the chi-square distribution with ν degrees of freedom. Since, for any fixed $t > 0$, $R(t)$ is a

decreasing function of λ , the exact $(1 - \alpha)$ confidence interval for $R(t)$ is

$$\begin{aligned} & (\hat{R}_L(t), \hat{R}_U(t)) \\ &= (\exp(-\hat{\lambda}_U t), \exp(-\hat{\lambda}_L t)) \\ &= \left(\exp \left[-\frac{t \chi_{1-\alpha/2}^2(2r)}{2 \sum_{i=1}^n t_i} \right], \exp \left[-\frac{t \chi_{\alpha/2}^2(2r)}{2 \sum_{i=1}^n t_i} \right] \right). \end{aligned} \quad (30)$$

In the case of singly Type I or multiply censored samples, exact confidence intervals for $R(t)$ are not available [17]. An approximate $(1 - \alpha)$ confidence interval for λ can be obtained by employing the asymptotic normal approximation of $(\hat{\lambda} - \lambda)/\hat{\sigma}_{\hat{\lambda}}$, where $\hat{\sigma}_{\hat{\lambda}} = \hat{\lambda}/\sqrt{r}$ is the estimated asymptotic standard deviation of $\hat{\lambda}$. Under such an approximation, we have [2]:

$$(\hat{\lambda}_L, \hat{\lambda}_U) = (\hat{\lambda} + z_{\alpha/2} \hat{\lambda}/\sqrt{r}, \hat{\lambda} - z_{\alpha/2} \hat{\lambda}/\sqrt{r}). \quad (31)$$

The approximation (31) works well for moderate and large samples, whereas for small samples a better approximation is given by Sprott [18]:

$$\begin{aligned} & (\hat{\lambda}_L, \hat{\lambda}_U) = \\ & \left(\left[\hat{\lambda}^{1/3} + z_{\alpha/2} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3, \left[\hat{\lambda}^{1/3} - z_{\alpha/2} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3 \right), \end{aligned} \quad (32)$$

which employs the normal approximation of $\hat{\lambda}^{1/3}$ in small samples. Another approximate method (often attributed to Ref. 19), which works well in the case of singly Type I censored, treats $2r\lambda/\hat{\lambda}$ as being approximately distributed as a chi-square random variable with $(2r + 1)$ degrees of freedom, so that

$$(\hat{\lambda}_L, \hat{\lambda}_U) = \left(\frac{\hat{\lambda} \chi_{\alpha/2}^2(2r + 1)}{2r}, \frac{\hat{\lambda} \chi_{1-\alpha/2}^2(2r + 1)}{2r} \right). \quad (33)$$

Once an approximate $(1 - \alpha)$ confidence interval for λ has been computed, then

irrespective of the chosen approximation, the approximate $(1 - \alpha)$ confidence interval for $R(t)$ is given by

$$(\hat{R}_L(t), \hat{R}_U(t)) = (\exp(-\hat{\lambda}_U t), \exp(-\hat{\lambda}_L t)). \quad (34)$$

Alternatively, approximate confidence limits based on the logit transformation can be evaluated from Equation (23), where the approximate estimated variance of $\hat{R}(t)$ is

$$\hat{\text{Var}}(\hat{R}(t)) = \left[\frac{dR(t)}{d\lambda} \Big|_{\hat{\lambda}} \right]^2 \hat{\sigma}_{\hat{\lambda}}^2 = \frac{(\hat{\lambda} t)^2 \exp(-2\hat{\lambda} t)}{r}. \quad (35)$$

Example 1 (continued). Let us assume that the electromechanical module data follow the one-parameter exponential function [13]. From Equation (27), the ML estimate of λ is $\hat{\lambda} = 0.000707\text{h}^{-1}$, since the number r of exact failure times is equal to 8 and the total time on test is equal to $\sum_{i=1}^{15} t_i = 11,318\text{h}$. The ML estimate of the reliability at time $t = 250\text{h}$ is

$$\begin{aligned} \hat{R}(250) &= \exp \left(-\frac{rt}{\sum_{i=1}^n t_i} \right) \\ &= \exp \left(-\frac{8 \cdot 250}{11318} \right) = 0.838. \end{aligned} \quad (36)$$

From Equations (31)–(33), approximate 95% confidence intervals for λ result in

$$\begin{aligned} & (\hat{\lambda} + z_{0.025} \hat{\lambda}/\sqrt{r}, \hat{\lambda} - z_{0.025} \hat{\lambda}/\sqrt{r}) \\ &= (0.000217\text{h}^{-1}, 0.001197\text{h}^{-1}), \end{aligned}$$

$$\begin{aligned} & \left(\left[\hat{\lambda}^{1/3} + z_{0.025} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3, \left[\hat{\lambda}^{1/3} - z_{0.025} \frac{\hat{\lambda}^{1/3}}{3\sqrt{r}} \right]^3 \right) \\ &= (0.000321\text{h}^{-1}, 0.001318\text{h}^{-1}), \end{aligned}$$

$$\left(\frac{\hat{\lambda} \cdot \chi_{0.025}^2(2r+1)}{2r}, \frac{\hat{\lambda} \cdot \chi_{0.975}^2(2r+1)}{2r} \right) \\ = (0.000334h^{-1}, 0.001334h^{-1}),$$

$$(\hat{R}_L(250), \hat{R}_U(250)) = \begin{cases} (0.741, 0.947), & \text{from the Normal approximation of } \hat{\lambda} \\ (0.719, 0.923), & \text{from the Normal approximation of } \hat{\lambda}^{1/3} \\ (0.716, 0.920), & \text{from the chi-square approximation of } 2r\lambda/\hat{\lambda}. \end{cases}$$

In addition, from Equation (23), the approximate 95% confidence interval for $R(250)$ based on the logit transformation results in (0.708, 0.917), since the estimated variance (35) is equal to 0.00274 and $w_{0.025}(0.838) = 2.130$.

Owing to the small size of the sample, the confidence interval (0.741, 0.947) based on the Normal approximation of $\hat{\lambda}$ should be discarded. The remaining three approximations should work all better, and give confidence intervals for $R(250)$ quite close each other.

Note that all the parametric confidence intervals for $R(250)$ are much narrower than the nonparametric intervals $(\hat{R}_L(250), \hat{R}_U(250))$ estimated in section titled ‘‘Complete and Singly Right Censored Samples.’’ This depends on the fact that the parametric approach uses not only the *sample* information (i.e., the information contained into the observed data), but also the *technical* information provided by the functional form chosen for the probability model. Of course, if the assumed model is not adequate, then both the estimate $\hat{R}(t)$ and the confidence interval $(\hat{R}_L(t), \hat{R}_U(t))$ are wrong, and the more the random variable deviates from exponentiality, the worse the parametric estimates are.

Data from the Two-Parameter Weibull Distribution

Let $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ indicate a random sample of size n of exact failure and right censored data from a two-parameter Weibull random variable with density function:

$$f(t; \theta, \beta) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1} \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right], \\ \theta, \beta > 0, t \geq 0, \quad (37)$$

where $z_{0.025} = -1.960$, $\chi_{0.975}^2(17) = 30.191$, and $\chi_{0.025}^2(17) = 7.564$. Then, by using Equation (34), approximate 95% confidence intervals for $R(250)$ are

and reliability function:

$$R(t; \theta, \beta) = \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right], \theta, \beta > 0. \quad (38)$$

The likelihood function $L(\theta, \beta; \mathbf{t})$ results in

$$L(\theta, \beta; \mathbf{t}) \propto \prod_{i=1}^n \left\{ \left(\frac{\beta}{\theta} \right)^{\delta_i} \left(\frac{t_i}{\theta} \right)^{\delta_i(\beta-1)} \right. \\ \left. \times \exp \left[-\delta_i \left(\frac{t_i}{\theta} \right)^{\beta} \right] \exp \left[-(1-\delta_i) \cdot \left(\frac{t_i}{\theta} \right)^{\beta} \right] \right\} \\ = \frac{\beta^r}{\theta^{r\beta}} \left(\prod_{i=1}^n t_i^{\delta_i} \right)^{\beta-1} \exp \left[- \left(\frac{\sum_{i=1}^n t_i^{\beta}}{\theta^{\beta}} \right) \right],$$

where $\sum_{i=1}^n \delta_i = r$ is the number of exact failure times. The log-likelihood function $\ell(\theta, \beta; \mathbf{t})$ is

$$\ell(\theta, \beta; \mathbf{t}) = r \ln \beta - r \beta \ln \theta + (\beta - 1) \\ \times \sum_{i=1}^n (\delta_i \ln t_i) - \left(\frac{\sum_{i=1}^n t_i^{\beta}}{\theta^{\beta}} \right), \quad (39)$$

and the ML estimate of the scale parameter θ is given by

$$\hat{\theta} = \left(\frac{\sum_{i=1}^n t_i^{\hat{\beta}}}{r} \right)^{1/\hat{\beta}}, \quad (40)$$

where the ML estimate $\hat{\beta}$ of the shape parameter β is the solution of

$$\frac{1}{\hat{\beta}} + \frac{\sum_{i=1}^n \delta_i \ln t_i}{r} - \frac{\sum_{i=1}^n t_i^{\hat{\beta}} \ln t_i}{\sum_{i=1}^n t_i^{\hat{\beta}}} = 0. \quad (41)$$

The ML estimate of the reliability $R(t; \theta, \beta)$ at any time $t \geq 0$ is

$$\hat{R}(t) = \exp \left[- \left(\frac{t}{\hat{\theta}} \right)^{\hat{\beta}} \right]. \quad (42)$$

The ML estimate of the reliability at a given time x , say $R_x \equiv R(x)$, can also be obtained directly reparameterizing the log-likelihood function in terms of R_x and β . In this case, being $\theta = x/[\ln(1/R_x)]^{1/\beta}$, the reparameterized density and reliability functions result in

$$f(t; R_x, \beta) = \frac{\beta}{x} \left(\frac{t}{x} \right)^{\beta-1} \ln(R_x) R_x^{(t/x)^\beta}, \quad 0 \leq R_x \leq 1, \beta > 0, t \geq 0, \quad (43)$$

$$R(t; R_x, \beta) = R_x^{(t/x)^\beta}, \quad 0 \leq R_x \leq 1, \beta > 0, t \geq 0. \quad (44)$$

Expressions (43) and (44) are then used to obtain the reparameterized log-likelihood function $l(R_x, \beta; \mathbf{t})$. Maximizing $l(R_x, \beta; \mathbf{t})$ with respect to R_x and β gives the same ML estimates $\hat{R}_x = \hat{R}(x)$ and $\hat{\beta}$ as those obtained by maximizing the log-likelihood Equation (39).

In the case of complete samples, the distribution of $\hat{R}(t)$ depends only on $R(t)$ and n . The ML estimator (42) is the nearly minimum variance unbiased estimator of $R(t)$: the bias $E\{\hat{R}(t) - R(t)\}$, given in Table 5 of Ref. 20, is quite small, and it does not seem worthwhile to attempt to eliminate it. For Type II censored samples, the distribution of $\hat{R}(t)$ depends on $R(t)$, n , and r , and the bias of $\hat{R}(t)$ is still small, at least in the large range of investigation of Ref. 21.

For complete data, exploiting the above-mentioned properties of the distribution

of $\hat{R}(t)$, exact lower confidence limits for $R(t)$ were obtained by using a Monte Carlo simulation method and are tabulated in Ref. 22. Alternatively, such limits could be obtained through an iterative procedure illustrated in Ref. 20. However, such a procedure is somewhat inconvenient, and then $(1 - \alpha)$ lower confidence limits can be read directly from the extensive Table 7A of Ref. 20, given for n ranging from 8 up to 100, $\hat{R}(t) = 0.50(0.02)0.98$, $\alpha = 0.25, 0.15, 0.10, 0.05$, and 0.02 .

Simple approximate $(1 - \alpha)$ confidence intervals for $R(t)$ can be obtained from Equation (21) (for large samples) or Equation (23) (for small or moderate samples), where an estimate of $\text{Var}(\hat{R}(t))$ can be obtained through the delta method:

$$\begin{aligned} \hat{\text{Var}}(\hat{R}(t)) &= \left[\frac{\partial R(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \hat{\text{Var}}(\hat{\theta}) \\ &+ \left[\frac{\partial R(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \right]^2 \hat{\text{Var}}(\hat{\beta}) \\ &+ 2 \frac{\partial R(t)}{\partial \theta} \Big|_{\hat{\theta}, \hat{\beta}} \frac{\partial R(t)}{\partial \beta} \Big|_{\hat{\theta}, \hat{\beta}} \\ &\times \hat{\text{Cov}}(\hat{\theta}, \hat{\beta}), \end{aligned} \quad (45)$$

where $\hat{\text{Var}}(\hat{\theta})$, $\hat{\text{Var}}(\hat{\beta})$, and $\hat{\text{Cov}}(\hat{\theta}, \hat{\beta})$ are the elements (1, 1), (2, 2), and (1, 2), respectively, of the estimated asymptotic variance-covariance matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$, and the partial derivatives of $R(t)$ with respect to the model parameters, say

$$\begin{aligned} \frac{\partial R(t)}{\partial \theta} &= \left(\frac{\beta}{\theta} \right) \left(\frac{t}{\theta} \right)^{\beta} \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right] \text{ and} \\ \frac{\partial R(t)}{\partial \beta} &= \left[\left(\frac{t}{\theta} \right)^{\beta} \beta \ln \left(\frac{t}{\theta} \right) \right] \exp \left[- \left(\frac{t}{\theta} \right)^{\beta} \right], \end{aligned}$$

are evaluated at the ML estimates $\hat{\theta}$ and $\hat{\beta}$. For complete samples, the asymptotic matrix $\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta})$ is [1,20]

$$\hat{\mathbf{V}}(\hat{\theta}, \hat{\beta}) = \begin{bmatrix} \left(\frac{\hat{\theta}}{\hat{\beta}} \right)^2 \frac{1.1087}{n} & \hat{\theta} \frac{0.2570}{n} \\ \hat{\theta} \frac{0.2570}{n} & \hat{\beta}^2 \frac{0.6079}{n} \end{bmatrix}. \quad (46)$$

As an alternative, the $(1 - \alpha)$ confidence interval proposed in Ref. 13 can be adopted:

$$(\hat{R}_L(t), \hat{R}_U(t)) = (\exp[-\exp(\hat{u}_U)], \exp[-\exp(\hat{u}_L)]), \quad (47)$$

where

$$(\hat{u}_L, \hat{u}_U) = \left(\hat{u} + z_{\alpha/2} \sqrt{\hat{Var}(\hat{u})}, \hat{u} - z_{\alpha/2} \sqrt{\hat{Var}(\hat{u})} \right),$$

$\hat{u} \equiv \ln[\hat{H}(t)] = \hat{\beta} \ln(t/\hat{\theta})$, and the asymptotic variance of $\hat{u}$ can be obtained through the delta method as follows:

$$\begin{aligned} \hat{Var}(\hat{u}) &= \left(\frac{\hat{\beta}}{\hat{\theta}} \right)^2 \hat{Var}(\hat{\theta}) + \left(\frac{\hat{u}}{\hat{\beta}} \right)^2 \\ &\quad \times \hat{Var}(\hat{\beta}) - 2 \frac{\hat{u}}{\hat{\theta}} \hat{Cov}(\hat{\theta}, \hat{\beta}). \end{aligned}$$

Using the asymptotic matrix (46), we have $\hat{Var}(\hat{u}) = (1.1087 + 0.6079\hat{u}^2 - 0.5140\hat{u})/n$. Confidence limits (47) are always within $(0, 1)$ for any $t > 0$ and any sample size. A very similar approximation is suggested by Ref. 1, which uses different coefficients to estimate the asymptotic variance of $\hat{u}$: $\hat{Var}(\hat{u}) = (1.1680 + 1.1000\hat{u}^2 - 0.1913\hat{u})/n$.

In the case of Type II censored sample, exact $(1 - \alpha)$ lower confidence limits can be read from Table 7B of Ref. 20, given for $n = 40(20)120$, $\hat{R}(t)$ ranging from 0.70 up to 0.999, $\alpha = 0.10, 0.05, 0.02$, and 0.01, and censoring level $r/n = 0.75$ and 0.50. Such limits can be used as approximate limits for Type I censored samples, too. As an alternative, the approximation (21) (suitable for large samples) or (23) (suitable for small or moderate samples) can be used, by estimating the asymptotic variance $\text{Var}(\hat{R}(t))$ through Equation (45) and computing the asymptotic matrix $\hat{V}(\hat{\theta}, \hat{\beta})$ as described in the section titled "Appendix." However, for singly censored samples, the matrix $\hat{V}(\hat{\theta}, \hat{\beta})$ can be easily read from tables in Ref. 23 too, as a function of the standardized censoring time $\zeta = [\ln T_C - \ln \hat{\theta}] \hat{\beta}$, where T_C is the censoring time. The matrix is the same for singly Type I and II censored data.

Table 4. Life Test Data of a Certain Type of Electromechanical Device in Example 4

i	t_i (h)	δ_i	i	t_i (h)	δ_i
1	1138	1	7	5046	1
2	1944	1	8	5139	1
3	2764	1	9	5500	0
4	2846	1	10	5500	0
5	3246	1	11	5500	0
6	3803	1	12	5500	0

Example 4. Let us consider the Type I censored data referring to $n = 12$ units of an electromechanical device given by Ref. [13]. All the tests were stopped at the time $T_C = 5500$ h and $r = 8$ failures were observed at times $t_1, t_2, \dots, t_8$ (see Table 4, where $\delta_i = 1$ if the time t_i is an exact failure time, and $\delta_i = 0$ if t_i is a right censored observation). By solving Equation (41), the ML estimate of β is $\hat{\beta} = 2.055$, from which, by using Equation (40), the ML estimate of θ is

$$\hat{\theta} = \left(\frac{\sum_{i=1}^{12} t_i^{2.055}}{8} \right)^{1/2.055} = 5215 \text{ h.}$$

The ML estimate of the reliability at the time $t = 4000$ h is computed from Equation (42): $\hat{R}(4000) = \exp[-(4000/5215)^{2.055}] = 0.560$. Unlike the nonparametric procedures, the ML procedure allows the reliability function $R(t)$ to be estimated also for times t greater than the largest observed time $t_{\max} = 5500$ h. For example, for $t = 8000$ h, still from Equation (42), we have $\hat{R}(8000) = \exp[-(8000/5215)^{2.055}] = 0.090$.

The approximate 90% confidence interval (23) for $\hat{R}(T)$ at $t = 4000$ h, based on the logit transformation, results in

$$(\hat{R}_L(4000), \hat{R}_U(4000)) = (0.320, 0.775),$$

being

$$\hat{V}(\hat{\theta}, \hat{\beta}) = \begin{pmatrix} 825042 \text{ h}^2 & -90.815 \text{ h} \\ -90.815 \text{ h} & 0.4117 \end{pmatrix},$$

$$\hat{Var}(\hat{R}(4000)) = 0.02230, \text{ and } w_{0.05}(\hat{R}) = 2.710.$$

GRAPHICAL PROCEDURE: PROBABILITY PLOTS

Estimate of the reliability function $R(t)$ can be easily performed via *probability charts*. This procedure can be used in the case the random variable T either pertains to the location-scale family or admits a transformation $\psi(T)$, which in turn leads to a location-scale distribution. More formally, it is required that the cumulative distribution function $F(t) = 1 - R(t)$ of the random variable T can be expressed as

$$F(t; a, b) = \Phi\left(\frac{\psi(t) - a}{b}\right), \quad (48)$$

being $\Phi(\bullet)$ a function independent of any unknown parameter. The parameters a and b are called *location* and *scale parameters*, respectively.

Then, the probability chart is obtained by adopting a $\psi(t)$ scale on the data axis (the x axis) and a $\Phi^{-1}(F)$ scale on the probability axis (the y axis). On this *model-dependent* chart, the plot of $F(t)$ versus t appears as a straight line. For example, in the case of the two-parameter Weibull model, the cumulative distribution function $F(t) = 1 - \exp[-(t/\theta)^\beta]$ can be linearized as it follows:

$$\ln \{\ln [1/(1 - F(t))]\} = \beta \ln t - \beta \ln \theta, \quad (49)$$

so that, for a two-parameter Weibull variable, $\psi(T) = \ln T$, $a = \ln \theta$, $b = 1/\beta$, and $\Phi^{-1}(F) = \ln \{\ln [1/(1 - F(t))]\}$, being $\Phi(\bullet) = 1 - \exp[-\exp(\bullet)]$ a function without unknown parameters.

Probability charts are available for many probability distributions widely used in reliability study, such as the standard exponential, the two-parameter Weibull, and the lognormal distributions.

In order to use probability charts, each ordered failure time $t_{(i)}$ ($i = 1, 2, \dots$) must be plotted on the chart against an estimate of $F(t)$ at that time, say F_i . These estimates of $F(t)$, named *plotting positions*, are obtained via a nonparametric approach. Only failure data are plotted on the probability chart, and not censoring times. Thus, plotting positions must be evaluated only at failure times.

Once the data are plotted on the chosen probability plot, it is necessary to fit a straight line through the points $(t_{(i)}, F_i)$, because the use of plotting positions inevitably produces departure from linearity. Very often the fitting is performed by sight, but a more accurate fit can be obtained by means of analytical methods, such as the least squares method [2]. The graphical estimate of $F(x)$ at a given time x of interest is then given by the ordinate (i.e., the value on the probability axis) of the point on the straight line having abscissa x .

Plotting Positions

In the case of complete or singly censored samples, the plotting position F_i for the i th ordered failure time $t_{(i)}$ ($i = 1, 2, \dots, r$) can be calculated as [2]:

$$F_i = \frac{i - 0.5}{n}, \quad i = 1, 2, \dots, r, \quad (50)$$

where n is the size of the sample and $r \leq n$ is the number of failure times in the sample. Other plotting positions, such as $F_i = i/(n + 1)$ and $F_i = (i - 0.3)/(n + 0.4)$, can be used. In the case of multiply censored samples without tied failure times, the plotting position F_i relative to the i th ordered failure time $t_{(i)}$ can be obtained via the Herd-Johnson estimator [1]:

$$F_i = 1 - \prod_{j=1}^i \left(\frac{k_j}{k_j + 1} \right), \quad i = 1, 2, \dots, r, \quad (51)$$

where k_i is the number of units at risk immediately before the ordered failure time $t_{(i)}$. In the case of tied failure times, the plotting position F_i has to be estimated at any distinct (ordered) time $x_{(i)}$ ($i = 1, 2, \dots, h$) at which one or more failures occur, by using

$$F_i = 1 - \prod_{j=1}^i \left(\frac{k_j - d_j + 1}{k_j + 1} \right), \quad i = 1, 2, \dots, h, \quad (52)$$

where d_j indicates the number of failures that occurred at $x_{(i)}$. Finally, for interval data, the plotting positions can be obtained from the actuarial estimators (15) or (16).

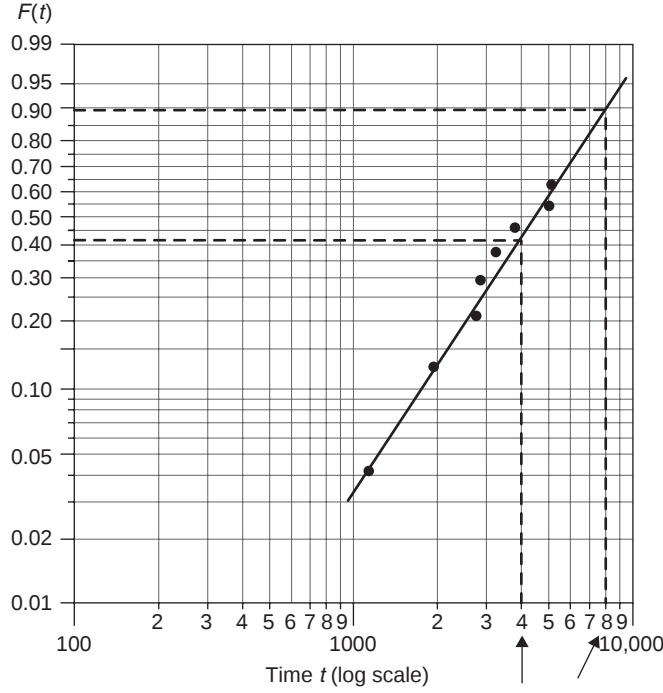


Figure 3. Graphical estimate of $F(4000)$ and $F(8000)$ in Example 4.

Table 5. Ordered Failure Times $t_{(i)}$ ($i = 1, 2, \dots, 8$) and the Corresponding Plotting Positions F_i in Example 4

i	$t_{(i)}$ (h)	F_i	i	$t_{(i)}$ (h)	F_i
1	1138	0.042	5	3246	0.375
2	1944	0.125	6	3803	0.458
3	2764	0.208	7	5046	0.542
4	2846	0.292	8	5139	0.625

Example 4 (continued). By using the life test data of the Type I censored sample of Table 4, we are interested in estimating, via a graphical procedure, the reliability of the electromechanical devices at the times $t = 4000$ h and $t = 8000$ h, under the assumption that failure times follow a two-parameter Weibull distribution. The plotting positions F_i are estimated via Equation (50) for each ordered failure time $t_{(i)}$ ($i = 1, 2, \dots, 8$), as given in Table 5. Once points $(t_{(i)}, F_i)$ are plotted on a Weibull chart and the straight line is fitted by sight, the

graphical estimate of the reliability is given by $\tilde{R}(4000) = 1 - \tilde{F}(4000) = 1 - 0.42 = 0.58$ and $\tilde{R}(8000) = 0.11$, as depicted in Fig. 3.

APPENDIX

The estimated $l \times l$ asymptotic variance-covariance matrix $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$ of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_l)$ is the inverse of the estimated information matrix $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$, whose entries are the negative second derivatives of the log-likelihood function $\ell(\boldsymbol{\theta}; t)$ with respect to the model parameters θ evaluated at the ML estimates $\hat{\boldsymbol{\theta}} : \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}$.

For a right censored sample $\mathbf{t} = \{t_1, t_2, \dots, t_n\}$ from a two-parameter Weibull distribution with scale parameters θ and shape parameter β (see section titled “Data from the Two-Parameter Weibull Distribution”), the estimated 2×2 information matrix results in

$$\hat{\mathbf{J}}(\hat{\theta}, \hat{\beta}) = \begin{pmatrix} r \left(\frac{\hat{\beta}}{\hat{\theta}} \right)^2 & -\frac{\hat{\beta}}{\hat{\theta}} \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \ln \left(\frac{t_i}{\hat{\theta}} \right) \\ -\frac{\hat{\beta}}{\hat{\theta}} \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \ln \left(\frac{t_i}{\hat{\theta}} \right) & \frac{r}{\hat{\beta}^2} + \sum_{i=1}^n \left(\frac{t_i}{\hat{\theta}} \right)^{\hat{\beta}} \left[\ln \left(\frac{t_i}{\hat{\theta}} \right) \right]^2 \end{pmatrix}, \quad (53)$$

where r ($r \leq n$) is the number of exact failure times.

REFERENCES

1. Nelson W. Applied life data analysis. New York: John Wiley & Sons; 1982.
2. Lawless JF. Statistical models and methods for lifetime data. New York: John Wiley & Sons; 1982.
3. Cox DR, Oakes D. Analysis of survival data. London: Chapman & Hall; 1984.
4. Nelson W. Accelerated testing. New York: John Wiley & Sons; 1990.
5. Meeker WQ, Escobar LA. Statistical methods for reliability data. New York: John Wiley & Sons; 1998.
6. Müller P, Quintana FA. Nonparametric Bayesian data analysis. *Stat Sci* 2004;19(1):95–110.
7. Martz HF, Waller RA. Bayesian reliability analysis. Malabar: Krieger Publishing Company; 1991.
8. Kalbfleisch JD, Prentice RL. The statistical analysis of failure data. New York: John Wiley & Sons; 1980.
9. Kleinbaum GD, Klein M. Survival analysis: a self-learning text. 2nd ed. New York: Springer; 2005.
10. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. 3rd ed. New York: Mc Graw-Hill Higher Education; 1974.
11. Blyth CR. Approximate binomial confidence limits. *J Am Stat Assoc* 1986;81(395):843–855. Sep;
12. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *Am Stat* 1998;52(2): 119–126. May; doi:10.2307/2685469).
13. Yang G. Life cycle reliability engineering. New York: John Wiley & Sons; 2007.
14. Chiang CL. Introduction to stochastic processes in biostatistics. New York: John Wiley & Sons; 1968.
15. Oehlert GW. A note on the delta method. *Am Stat* 1992;46(1):27–29. DOI: 10.2307/2684406.
16. Balakrishnan N, Basu AP. The exponential distribution: theory, methods, and applications. New York: Gordon and Breach; 1996.
17. Miller RG. Survival analysis. New York: John Wiley & Sons; 1981.
18. Sprott DA. Normal likelihoods and their relation to large sample theory of estimation. *Biometrika* 1973;60(3):457–465. DOI: 10.1093/biomet/60.3.457.
19. Cox DR. Some simple approximate tests for Poisson variates. *Biometrika* 1953;40 (3–4):354–360. DOI: 10.1093/biomet/40.3–4.354.
20. Bain LJ, Engelhardt M. Statistical analysis of reliability and life-testing models: theory and methods. 2nd ed. New York: Marcel Dekker; 1991.
21. Billmann BR, Antle CE, Bain LJ. Statistical inference from censored Weibull samples. *Technometrics* 1972;14(4):831–840.
22. Thomas DR, Bain LJ, Antle CE. Maximum likelihood estimation, exact confidence intervals for reliability, and tolerance limits in the Weibull distribution. *Technometrics* 1970;12(2):363–371.
23. Meeker WQ, Nelson W. Weibull variances and confidence limits by maximum likelihood for singly censored data. *Technometrics* 1977;19(4):473–476.

ESTONIAN OPERATIONAL RESEARCH SOCIETY

OTU VAARMANN
Institute of Cybernetics at
TUT, Department of
Mechanics and Applied
Mathematics, Tallinn,
Estonia

The *Estonian Operational Research Society* (EstORS) — *Eesti Operatsioonianalüüsi Selts* in Estonian — was founded by Dr. Toomas Meressoo [Estonian Scientific Foundation (ESF), Management Board] and Prof. Otu Vaarmann [Institute of Cybernetics at Tallinn University of Technology (TUT), Department of Mechanics and Applied Mathematics] on 18th March, 2010. At the extended meeting (General Assembly) of the society, including 22 participants from different Estonian universities and research institutes, on 6th April, 2010, the Management Board (Directive Board), composed of seven members, was elected: Otu Vaarmann was elected as president (Chairman of the Management Board); Assoc. Prof. Peep Miidla [University of Tartu (UT), Institute of Applied Mathematics]; vice president and IFORS representative; Mrs. Marje Tamm [Institute of Cybernetics at TUT, Library], secretary; Dr. Harald Kitzmann [Tallinn University of Technology (TUT), Tallinn School of Economics and Business Administration], Treasurer; and Prof. Jaan Janno [Tallinn University of Technology (TUT), Department of Mathematics], Prof. Jaan Lellep [University of Tartu (UT), Department of Mathematics], and Dr. Toomas Meressoo (ESF), members. Dr. Tiit Riismaa [Institute of Cybernetics at TUT, Department of Mechanics and Applied Mathematics] was elected as the auditor of the society. In accordance with the decision of the Management Board on 6th April, 2010, the application form for membership asking to admit the Estonian Operational Society to IFORS was sent to the IFORS secretary

on 22nd June, 2010, and in March 2011, the Estonian Operational Research Society has been accepted as the fifty-second national member of IFORS. Estonia, with its 1.34 million inhabitants, is one of the smallest members of the European Union.

EstORS is a sovereign voluntary nonprofit organization located in Estonian Republic, Tallinn (<http://www.ioc.ee/matem/estors/>). EstORS will be ruled and regulated by its statutes in accordance with the Constitution of the Estonian Republic as well as pertinent Estonian Laws and Regulations. The society unites the experts in the field of Operational Research/Industrial Mathematics and its applications, actively working for development of Estonian Republic as an information society. According to its statutory mission, EstORS aims at contributing to the progress of scientific and technical knowledge, integrating competence Engineering Science and Management Science/Operations Research (OR), creating and transferring scientific knowledge in Informatics and Decision Systems.

Every person or juridical person who is actively working for achieving the goals and objectives of the society and fulfilling the requirements of the statutes may be a member of the society. The members may belong to the following categories: founding members, active members, honorary members, and contributor members.

The founding members are those who are admitted within a 30-day period following this act.

The active members are those who have been included in the social register after 30 days passed from the society foundation event and who have regularly complied with the obligations imposed by the statutes.

The honorary members are those who attend to their merits or to relevant services given to the institution are judged by the General Assembly.

The contributor members are those who have not fulfilled all the requisites established to the active members.

In all cases, the members may be Estonian citizen or foreigners. The decision on admittance or withdrawal will be taken by the Management Board (Directive Board). The former member may be accepted once more to the society after one year of withdrawal under the common conditions.

The General Assembly consists of all the members having the right to participate, and for those having the right to vote, their vote will be counted equally. The General Assembly will meet either ordinarily or extraordinarily to treat exclusively the themes and items within a previously announced agenda. The General Assembly accepts the statutes, formulates the changes and basic directions of the society, and is convoked by the Management Board at least once a year.

The society will be administered and represented by the Management Board composed of at least one member and supremely of seven members. The members and the chairman of the Management Board (the president), as well as the members of Auditing Commission (or the Auditor), will be appointed by the General Assembly to hold their post for three years. Sessions of the Management Board will be held no rarer than once in a half year.

Main purposes of the society are

- to encourage theoretical and applied researchers in the field of OR to contribute to the creation of new knowledge in Informatics and Decision Making (DM) from the viewpoint of science, technology, and innovation;
- to participate with the other institutions in the elaboration of developmental policies for disciplines related to the scope of this society;
- to promote the establishment of international contacts and stimulate further advancement of academic links and free exchange of experience between its members and members of other similar scientific societies belonging to this country, this region, and the world;
- to organize national and international meetings (conferences, seminars, exhibitions, etc.) related to the subjects, scope, and objectives of the society;
- to contribute to the distribution of scientific and academic contributions obtained by the society and its members by advocating the achievements in the field of OR by means of lectures, publishing books and papers;
- to encourage the teaching of OR and to this end to create funds and courses for raising the level of the skills in the field of OR/DM.

The society plans to introduce a new Master Curriculum on Industrial Mathematics in accordance with ECMI (European Consortium For Mathematics in Industry) corresponding to Model Curriculum. Through the cooperative effort of mathematicians, engineers, and experts in science, the society also aims to develop educational materials that link mathematical topics with applications in engineering, management, and science that will be suitable for further education and self-study for students. Further, the society concerns itself with mathematical education and broader relationship between mathematics and society. The society intends to advance a variety of educational cooperation activities between partners in the area of lifelong learning from the countries in the Baltic and Nordic regions. Currently, specialization in OR can be found in Estonia mainly at TUT and UT.

Public decision making typically involves complex optimization problems with several conflicting goals from very different backgrounds (environmental planning and management of environmental problems, natural resource management, reallocations of land uses, conservation biodiversity, analysis of logistic problems, disaster management, etc.), and there is a great demand for the assessment of the ongoing sustainable management by use of criteria and indicators in questions. Therefore, in decision-making tasks where trade-offs between economic development and environmental quality are indispensable, a Pareto-optimal solution provided by parallel use of cost-benefit

analysis (CBA) and multicriteria decision analysis (MCDA) is desirable. The measurement of the level of sustainability for a certain economic activity is a difficult task involving uncertainties and subjective considerations, and there is not yet a general consensus on how to measure sustainability. Multicriteria analysis provides powerful and widely used techniques for the appraisal of complex decision problems and the analysis of sustainability. Many multicriteria decision problems can be posed as a multiobjective

optimization problem to find adequate solutions to many challenging problems in business, technology, and society as a whole.

Although our subsequent research is to a high degree focused on multicriteria analysis and large-scale and/or multiple-objective optimization problems, the society also plans more widely to embark on projects to improve the attitude toward mathematics, which, it hopes, will result in more widespread use of OR.

EULERIAN PATH AND TOUR PROBLEMS

SRIMATHY MOHAN
Department of Supply
Chain Management,
W.P. Carey School of
Business, Arizona
State University,
Tempe, Arizona

HISTORY OF THE EULERIAN TOUR PROBLEM

Routing problems play a very important role in service delivery. These problems arise in a large number of real-world applications such as postal delivery, electric and water meter reading, garbage collection, and snow clearance. Designing efficient routes for people and vehicles delivering these services has been a problem of interest for operations researchers and mathematicians. Such routing problems are of two types — node routing and arc routing problems — depending on whether the service request is at a node or on an arc/edge of the underlying graph. The underlying problem for most arc routing problems is determining a giant tour that starts and ends at a designated node and traverses all edges requiring service at least once. A tour (or circuit) is Eulerian if it starts at a designated node and traverses every edge in the graph exactly once before returning to the designated starting node. Given a graph and a starting node, the *Eulerian tour problem* (ETP) seeks to determine an Eulerian tour (or circuit), if one exists, or to show that the underlying graph does not contain an Eulerian tour.

The origins of the ETP can be traced back to the eighteenth century. The Pregel River flowed through the city of Königsberg in Eastern Prussia and enveloped the small island of Kneiphof before it split into two. The city had a total of seven bridges that were built across the river and connected the island and various parts of the mainland. Figure 1 shows

the orientation of the city and the bridges. Many people tried to determine if one could start at the end of a particular bridge and come back to the same point after traversing each of the seven bridges exactly once. It was believed that it was impossible, but no one could prove it. In 1735, Swiss mathematician, Leonard Euler showed that it is not possible to come back to the same point after traversing each bridge exactly once [1]. He showed that in order to come back to the same point after traversing each edge (or bridge) exactly once, every node in the underlying graph should be of even degree. In the Königsberg bridge problem, three of the nodes have a degree of three and one has a degree of five (Fig. 1).

Alexandersan [2] provides a detailed historical account of Leonard Euler's work on the Königsberg bridges and how and when his work was published. Facsimiles of Euler's original work have been published in Euler [3], and English translations of the original paper can be found in Fleischner [4]. Euler's work was essentially the beginning of many interesting graph theory studies, and hence, graphs that satisfy this condition have been termed *Eulerian graphs*. Eulerian graphs are also called *unicursal graphs* in graph theory.

In this paper, we present research related to the ETP. The ETP seeks to determine a tour that starts and ends at the same node and traverses all the edges in the graph exactly once or to show that the graph is not Eulerian. The next section introduces the basic notation and presents algorithms for determining an Eulerian tour in undirected, directed, and mixed Eulerian graphs. An undirected graph contains only undirected edges connecting the vertices of the graph, that is, the connecting links can be traversed in either direction. In a directed graph, every link connecting two vertices has a direction of traversal associated with it and the link (or arc) can be traversed only in that direction. A mixed graph contains a mix of both undirected and directed edges connecting the vertices. A connected undirected graph is

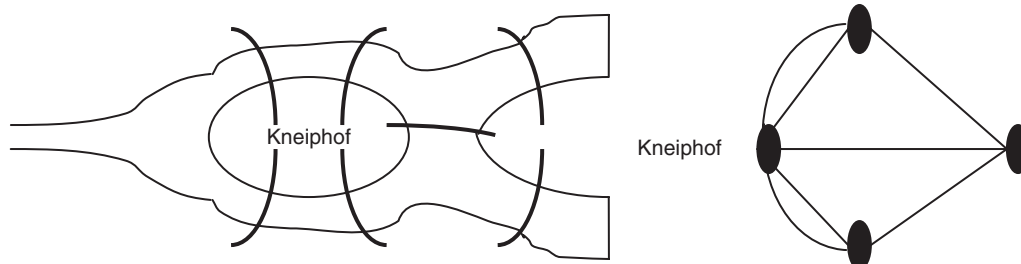


Figure 1. Seven bridges of Königsberg.

Eulerian if and only if every node has an even degree. A connected directed graph is Eulerian if and only if the in-degree of every vertex is equal to its out-degree. A mixed graph is Eulerian if all nodes have an even degree, and for each node, the number of arcs directed toward it is the same as the number of arcs directed away from it.

A graph is semi-Eulerian if and only if it is connected on all nodes except two that are of even degree. In a semi-Eulerian graph, we can determine an Eulerian path (or trail), that is, a path that traverses each arc or edge exactly once and starts and ends at the two nodes of odd degree. By adding an edge between the two nodes of odd degree, we get an Eulerian graph. We can then determine an Eulerian tour using the algorithms described below and then remove the edge that was added to obtain the Eulerian path (or trail). Hence, for the remainder of the paper, we have described the algorithms for Eulerian tours and concluded with areas of future research.

It is interesting to note that not all graphs are Eulerian, as in the case of the Königsberg bridge problem. In these cases, one might still be interested in determining an Eulerian tour on the underlying graph. This forms an interesting routing problem called the *Chinese postman problem* (CPP). The objective here is to determine the set of edges to be added to the graph at the lowest cost (distance) possible in order to make it Eulerian. Edmonds and Johnson [5] show that the CPP can be solved in polynomial time for undirected graphs using an adaptation of Edmond's blossom algorithm. This problem is called the *undirected CPP*. While the CPP

for directed graphs is also solvable in polynomial time, it is NP-hard when the underlying graph contains both arcs and edges. In this case, heuristics are used to make the graph Eulerian. For more detailed information on the least cost augmentation problem, see the articles *Matching and Assignment Algorithms* and *Chinese Postman Problem* in this encyclopedia.

ALGORITHMS FOR THE EULERIAN TOUR PROBLEM

In this paper, we assume that we are given an Eulerian graph $G = (V, E)$, and hence, the ETP seeks to determine an Eulerian tour (or circuit) on the underlying graph. Several algorithms have been proposed for determining Eulerian tours for directed, undirected, and mixed graphs. We next present a summary of these algorithms and direct the readers to appropriate references for detailed descriptions and proofs of the algorithms.

Undirected Eulerian Graphs

Euler [1] showed that an undirected graph is unicursal or Eulerian if and only if every node in the graph is of even degree. One of the earliest algorithms for the ETP on undirected graphs is described in Berge. Kaufmann later describes the same algorithm as Fleury's algorithm [6,7]. This is a very simple algorithm which depends on avoiding an *isthmus* or *bridge* in the graph. An *isthmus* or *bridge* is an edge in a connected graph whose removal disconnects the underlying graph.

Fleury's Algorithm.

- Step 1.* Mark all the edges in the graph as *unmarked*. Start at any node v_0 . Set $v_i = v_0$.
- Step 2.* From all the *unmarked* edges incident to v_i , pick an edge (v_i, v_j) that is not an *isthmus*. Remove edge (v_i, v_j) from the graph. If all edges have been removed, stop. If not, go to step 3.
- Step 3.* Set $v_i = v_j$ and go to step 2.

The algorithm is computationally expensive since at each step, one has to determine whether the chosen edge is an isthmus of the graph with unmarked edges. While the graph traversal in Fleury's algorithm is linear in the number of edges, that is, $O(|E|)$, we also need to factor in the complexity of detecting bridges, which is also linear in the number of edges, $O(|E|)$. Hence, Fleury's algorithm will have a time complexity of $O(|E|^2)$. Edmonds and Johnson [5] have proposed three alternate algorithms that are computationally simpler. These are the end-pairing algorithm, next-node algorithm, and the maze-search algorithm. All three algorithms essentially form one or more small tours that do not share any edges and then combine these tours to form a giant Eulerian tour. The three algorithms proposed by Edmonds and Johnson [5] are of $O(|E|)$ time complexity. We next present simple descriptions of the algorithms. See Edmonds and Johnson [5] for detailed algorithmic descriptions and proofs. The end-pairing algorithm represents the tour as $(v_0, e_1, v_1, e_2, \dots, v_0)$.

End-Pairing Algorithm.

- Step 1.* Start at node v_0 . Trace a simple tour starting and ending at node v_0 . If the tour includes all the edges in the graph, stop. If not, go to step 2.
- Step 2:* Consider any node v_i on the tour that has at least one edge incident to it and not on the tour. Trace another tour starting and ending at node v_i and using edges not on the first tour.
- Step 3:* If $v_i = v_0$, follow the first tour and then the second. If v_i is not equal to v_0 , then let e^1 and e^2 be two edges incident

to node v_i on the first tour and e^3 and e^4 be two edges incident to node v_i on the second tour. Interject the second tour on the first one as follows. Start at node v_0 . Follow the first tour to reach node v_i through edge e^1 , then follow the second tour through e^3 and return to node v_i through e^4 , and trace the remainder of the first tour through edge e^2 to return to v_0 . If the merged tour includes all edges in the graph, stop. If not, go to step 2.

Figure 2 shows a simple undirected Eulerian graph with five nodes and eight edges. Let us assume that the algorithm starts at node 1. The simple tour traced in step 1 is 1-2-5-1. If the next node chosen in step 2 is node 2 and the traced tour is 2-4-3-2, then $e^1 = (1, 2)$, $e^2 = (2, 5)$, $e^3 = (2, 4)$, and $e^4 = (3, 2)$. Step 3 would lead to the longer tour 1-2-4-3-2-5-1. The next simple traced tour will be 1-3-1. Since this tour originates at the origin of the appended tour, the final Eulerian tour will be 1-2-4-3-2-5-1-3-1.

In the next-node and maze-search algorithms, a tour is represented by associating a list of nodes for each node v_i . The list for each node v_i is obtained using the algorithm and is represented as $L_{v_i}(1), L_{v_i}(2), \dots, L_{v_i}(k)$. The list indicates the next node to go to when leaving node v_i .

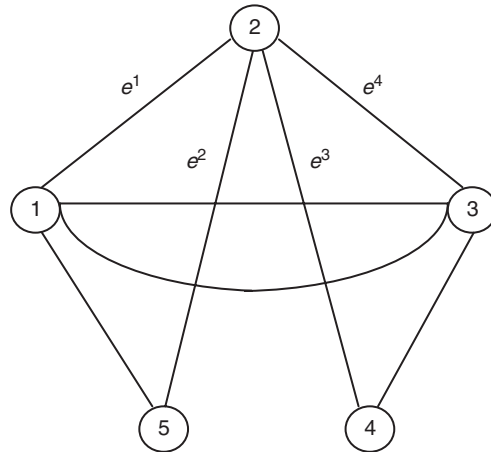


Figure 2. Example graph to demonstrate the end-pairing and next-node algorithms.

Next-Node Algorithm.

Step 1: Pick any node v_0 . Set $v = v_0$ and $k_i = 0$ for all nodes i . Let e be any edge incident to node v . All edges are *unmarked*.

Step 2: Let v_i be the node adjacent to node v on edge e . Let $k_{v_i} = k_{v_i} + 1$ and $L_{v_i}(k_{v_i}) = v$. Edge e is *marked* as used. If v_i has any other *unmarked* edges incident to it, go to step 3. If not, node v_i should now be v_0 . Go to step 4.

Step 3: Set $v = v_i$ and let e be any *unmarked* edge incident to v_i . Go to step 2.

Step 4: Let v_0 be any node with at least one *marked* and one *unmarked* edge incident to it. Set $v = v_0$, and go to step 2. If no such node exists, stop.

At the end of the algorithm, each node has its own list of nodes. The first time node v_i is reached, we leave it through node $L_{v_i}(k)$, the second time through node $L_{v_i}(k - 1)$, and the last time through node $L_{v_i}(1)$. For the graph in Figure 2, the algorithm starts at node 1, and scans the edges in the following order (1,2), (2,3), (3,4), (4,2), (2,5), (5,1), (1,3), and (3,1). Hence, the nodes have the following lists. $L_1(1) = 5$, $L_1(2) = 3$; $L_2(1) = 1$, $L_2(2) = 4$; $L_3(1) = 2$, $L_3(2) = 1$; $L_4(1) = 3$; and $L_5(1) = 2$. Now the Eulerian tour is represented as 1-3-1-5-2-4-3-2-1.

Maze-Search Algorithm.

The algorithm is similar to the next-node algorithm. While the end-pairing and next-node algorithms provided a representation of the Eulerian tour without traversing it, the maze-search algorithm actually traverses the tour. When the next-node algorithm has completed a tour and there are still some *unmarked* edges left, the algorithm moves to any node on the tour that is incident to at least one *unmarked* edge and forms another tour and interjects that tour on the original tour. On the other hand, the maze-search algorithm retraces the path from the start node until it finds the first node with at least one

unmarked edge and forms another tour. When there are no *unmarked* edges left, the algorithm retraces the tour from the last node to the start node. Thus, in the maze-search algorithm every edge is traversed twice. The order of the second traversal gives the actual Eulerian tour. Edmonds and Johnson [5, page 114] provide a detailed description of the maze-search algorithm.

Directed Eulerian Graphs

In a directed graph, every edge in the graph is either directed toward or away from a node. A directed graph is symmetric if the in-degree of every node is equal to its out-degree. A directed graph is Eulerian if and only if it is connected and symmetric. van Aardenne-Ehrenfest and de Bruijn [8] have described the spanning arborescence algorithm to determine an Eulerian tour in a directed Eulerian graph.

An arborescence is a directed graph in which, for a vertex v_0 called the *root* and any other vertex v , there is exactly one directed path from v_0 to v . In other words, an arborescence is a directed, rooted tree in which all edges point away from the root. If every node other than the root node of the arborescence has in-degree (out-degree) at most 1, it is referred to as an out-tree arborescence (in-tree arborescence). A spanning arborescence is an arborescence whose underlying undirected graph is a spanning tree.

Spanning Arborescence Algorithm.

Step 1. Construct a spanning arborescence with a root node v_0 .

Step 2. Label all the edges of the graph consecutively as follows. Start at the root node v_0 , and number the edges out of this node arbitrarily. The remaining nodes can be visited for numbering the outgoing edges in an arbitrary manner. For each node, number the edges out of the node consecutively while taking note of the following two requirements. If there is an edge from the node back

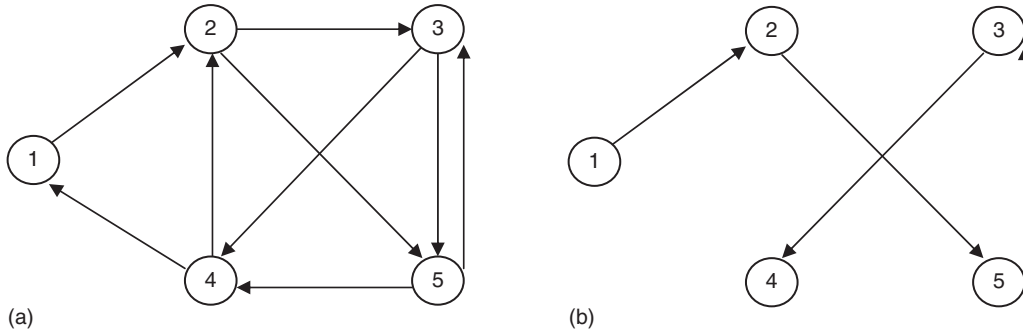


Figure 3. Directed graph and associated spanning arborescence.

to the root node, it will be numbered last. The edge used by the spanning arborescence outgoing from the node will receive the penultimate number in the ordering.

Step 3. The Eulerian tour can be traced by starting at the root node and traversing any edge incident to it. At all other nodes, traverse the lowest numbered edge. Stop when there are no edges remaining in the graph.

Figure 3a shows a directed Eulerian graph with five nodes and nine directed edges and a spanning arborescence for the graph. The graph is connected and symmetric. Hence, it is clear that the graph is Eulerian. The spanning arborescence is rooted at node 1 and has a directed path to the other four nodes from node 1. Figure 3b shows the numbering of the directed edges according to step 2 of the algorithm. For the given spanning arborescence and ordering, the Eulerian tour is 1-2-3-5-4-2-5-3-4-1. Note that one can use this algorithm to find more than one Eulerian tour. If we use a different ordering for visiting the nodes and for numbering the edges out of each node in step 2, we will trace a different Eulerian tour. A different spanning arborescence rooted at the same node or at a different node would again lead to another set of Eulerian tours. The important step in the algorithm is to find a spanning arborescence, which can be accomplished in $O(|E|)$ time complexity. The other steps require scanning each edge once, and hence, the time complexity of the algorithm is $O(|E|)$.

Mixed Eulerian Graphs

A mixed graph is one in which some edges are undirected and some are directed. A mixed graph is said to be balanced if for any subset S of the nodes in the graph, the differences between the number of edges directed from S to V/S and the number of edges directed from V/S to S is no greater than the number of undirected edges between S and V/S . A mixed graph is even if all nodes have an even degree, and the graph is symmetric if for every node, the number of edges directed toward the node is the same as the number of edges directed away from the node. A mixed graph is Eulerian if and only if it is connected, even, and balanced [9]. If a graph is even and symmetric, it is balanced and hence, Eulerian.

In order to prove the balanced condition of the graph, Ford and Fulkerson [[9], p. 60] show that we need to establish a circulation in G , and then direct some of the undirected edges according to the circulation. They force a unit flow on every directed edge of the graph and replace each undirected edge with a pair of opposite directed edges with a lower bound of 0 and an upper bound of 1. The balanced condition basically shows that there is sufficient escape capacity from S to V/S to take care of the forced flow on the directed edges. If this is true for every set S , then a feasible circulation exists. In order to prove this, Ford and Fulkerson suggest adding a source node (s), a sink node (t), and the arcs (s, t) and (t, s) with infinite capacity to the graph. If we solve the maximal flow problem on this network, we obtain a feasible circulation in

G. The details of the procedure are described below.

Converting a Mixed Graph to a Symmetric Graph.

- Step 1.* For each directed edge in the original graph, assign a lower bound of 1 and an upper bound of 1. Now replace each undirected edge with a pair of opposite directed edges, each of which has a lower bound of 0 and an upper bound of 1.
- Step 2.* Solve the network flow problem on this appended graph to obtain a feasible circulation.
- Step 3.* If (v_i, v_j) is an undirected edge in the original graph and the flow from v_i to v_j is 1 and the flow from v_j to v_i is 0, then orient the undirected edge (v_i, v_j) from v_i to v_j .

The maximum flow in the graph can be calculated on $O(|E|f)$ time using the Ford–Fulkerson algorithm, where f is the maximum flow. A variation of the algorithm, developed by Edmonds and Karp [10], can determine the maximum flow in $O(|V||E|^2)$ time. If the resulting graph contains only directed edges, the van Aardenne-Ehrenfest and de Bruijn algorithm described in the previous section can be used to determine the Eulerian tour. If the graph still contains undirected edges, we can easily assign directions to the remaining undirected edges using the following $O(|E|)$ procedure, and then use the spanning arborescence algorithm. The procedure to assign directions to the undirected edges is as described in Eiselt *et al.* [[11], p. XX].

- Step 1.* If there are no undirected edges, stop.
- Step 2.* Pick a node v that is incident to at least one undirected edge (v, w) . Set $v_i = v$ and $v_j = w$.
- Step 3.* Orient (v_i, v_j) from v_i to v_j . If $v_j = v$, go to step 1.
- Step 4.* Set $v_i = v_j$, and find an edge incident to v_i . Go to step 3.

At the end of this procedure, the mixed graph is transformed to a directed Eulerian graph, and one can use the spanning arborescence algorithm to determine an Eulerian tour.

CONCLUSION

In this paper, we have presented a historical overview of the ETP and summarized the main algorithms for finding an Eulerian tour in undirected, directed, and mixed unicursal graphs. There exist quite a few situations in practice, when not all the edges of the Eulerian graph need to be visited everyday. That is, the probability that an edge has to be visited on any given day could be less than one. This can be modeled as a stochastic Eulerian tour problem (SETP). While the deterministic version of the ETP presented in this paper has a long history and is well researched, the stochastic ETP is very new. Further research should lead to the investigation of the stochastic CPP and development of good heuristics for the SETP as well as the other stochastic arc routing problems.

REFERENCES

1. Euler L. Solution Problematis ad geometriam situs pertinentis. *Commentarii Academiae Petropolitanae* 1736;8:128–140.
2. Alexandersan GL. About the cover: Euler and Königsberg’s bridges: a historical view. *Bulletin of the Am. Math. Soc* 2006;43(4):567–573.
3. Euler L. (Newman JR, editor). Leonard Euler and the Königsberg bridges. *Sci. Am* 1953;189:66–70.
4. Fleischner H. Eulerian graphs and related topics (Part 1, Volume 1). *Annals of discrete mathematics*. Amsterdam: North-Holland; 1990. p.45.
5. Edmonds J, Johnson EL. Matching, Euler tours and the Chinese postman problem. *Math. Prog* 1973;5:88–124.
6. Berge C. The theory of graphs and its applications. New York: Wiley; 1962.
7. Kaufmann A. Graphs, dynamic programming and finite games. New York: Academic Press; 1961.

8. van Aardenne-Ehrenfest, de Bruijn NG. Circuits and trees in oriented graphs. *Simon Stevin* 1951;28:203–217.
9. Ford LR, Fulkerson DR. *Flows in networks*. Princeton, New Jersey: Princeton Univ. Press; 1962.
10. Edmonds J, Karp RM. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM* 1972;19(2):248–264.
11. Eiselt HA, Gendreau M, Laporte G. Arc routing problems. Part I: The Chinese postman problem. *Oper. Res* 1995;43:231–242.

EVACUATION PLANNING

EVA REGNIER

Associate Professor of Decision
Science, Defense Resources
Management Institute Naval
Postgraduate School, Monterey,
California,

Since the September 11 terrorist attacks, there has been an explosion of Operations Research (OR) work applied to the preparation for and response to disasters, primarily human-caused disasters. Hurricane Katrina in 2005 prompted a second wave of OR efforts to prevent and mitigate the impacts of natural disasters.

Yet, the existing OR literature on evacuation planning is quite sparse; Wright *et al.* [1] surveyed the use of OR methods in homeland security and identified 14 articles on evacuation models. Despite the recent surge of interest in response to disasters, including hurricanes, OR work on evacuation planning has not increased noticeably.

This article outlines the types of decisions involved in evacuation planning, and briefly describes the OR modeling and solution techniques that can and have been applied. For readers interested in a review, the Wright *et al.* review is fairly complete to the date of this writing, with a few recent contributions cited later in this article. This article also highlights challenges to deploying OR tools in evacuation planning and OR's limitations, and problems that OR cannot address without substantial contribution from or collaboration with other fields of expertise.

EVACUATIONS AND EVACUATION PLANNING

Evacuation is the rapid movement of people, usually away from a localized hazard, and with little or no lead time. Evacuations can occur from a large geographic region—for

example, to avoid a hurricane—or from a single building, aircraft, or ship. Evacuation may involve pedestrians, vehicles, aircraft, watercraft, rail networks or combinations. They may be centrally directed or (more frequently) resulting from the collective behavior of many individual agents, be they individuals, families, or vehicles. Evacuation planning therefore, provides many important problems to which OR can contribute.

Evacuations vary tremendously in scope and type, and may be very complex, involving many decisions that affect their progress and effectiveness. The following describes some of the dimensions across which evacuation scenarios differ, and that largely determine the relevant type of decisions, modeling, and analyses:

- *No Notice versus Short Notice.* In many cases, the hazard that necessitates an evacuation occurs without warning and therefore, evacuation is initiated after the hazard's impact has begun. Examples include evacuating an aircraft following a crash or failure, evacuating a city or region during a fire or following an industrial accident (e.g., nuclear or chemical), or a terrorist attack. Even when there is no notice of the precipitating event, the impacts may continue to worsen following the initial event; for example, a fire may be spreading through a building. In other cases—for example, an earthquake—the extent of the hazard may be reached before evacuation begins, but evacuation is necessary to be able to provide care for those in the region.

In contrast, some types of hazard provide advance warning (short notice) and therefore, the evacuation may be initiated before the hazard impact occurs. Hazards that may provide warning include hurricanes and some terrorist threats.

- *Lead Time.* A related factor is how much time is available to complete the evacuation. Almost by definition, evacuations are completed with urgency, and in many cases, OR formulations minimize evacuation time (or maximize the number of people evacuated within a given time). However, there are degrees of urgency. For example, evacuations to avoid a hurricane have a lead time of up to several days. Trade-offs between speed of evacuation and completeness or cost come into play, and the amount of lead time dictates how important speed will be relative to other objectives and other factors that may be modeled. The amount of lead time available for evacuation before the hazard impacts occur is usually uncertain, but there may be some basis for a deterministic or probabilistic estimate, as with a hurricane.
- *Scale—Number of People.* A key dimension differentiating evacuations is their scale. Evacuations from aircraft, vessels, or buildings may involve a handful to a few thousand people. Large-scale evacuations may involve tens or, potentially, even hundreds of thousands of people.
- *Scale—Spatial.* The spatial scale of the evacuation is highly correlated with the number of people involved; evacuations that involve tens of thousands of evacuees are regional and will involve moving evacuees tens to hundreds of miles. As will be discussed below, the types of decisions and the types of modeling involved in planning for small- versus large-scale decisions will be quite different.
- *Predictability.* A further set of issues relates to the degree to which important parameters—such as evacuation zones, number of evacuees, and capacity within the infrastructure—can be anticipated before the precipitating event. In most cases, a building evacuation is relatively predictable once it is initiated. Generally, the entire building would be evacuated at once; the number of people in each area can be known within relatively narrow bounds. In contrast, with wildfires, these parameters are much more variable: the evacuation zones are not known, the number of people in each area may be highly variable by time of day or season, the response rate is uncertain, and so on. Furthermore, some threats may evolve over time, requiring evacuation of more or different areas (e.g., through the spread of fire) or cutting off escape routes (e.g., flooding roadways during a hurricane). The evolution may be uncertain. These differences may require very different modeling techniques; for example, uncertainty must be modeled for highly unpredictable evacuations. The types of decisions that are most important will also be different as a function of predictability; for example, the importance of information-gathering and communication will be much greater for highly unpredictable evacuations.

Two important characteristics of evacuations have been omitted from the taxonomy: mode of transportation and type of physical infrastructure. The mode of transportation is tightly linked with scale, for instance

Table 1. Application of the Taxonomy to Three Evacuations: the World Trade Center (WTC) Evacuation on September 11, the 2008 Evacuation in Advance of Hurricane Ike, and the December 2008 Aircraft Evacuation at the Denver Airport

Dimension	WTC 2001	Hurricane Ike	Denver Jet
Lead time	None to minutes	Days to a week	None
Scale—people	Thousands	One million	115
Scale—spatial	Hundreds of vertical/horizontal feet	Tens to hundreds of miles	Hundreds of feet
Predictability	Medium	Low	High

evacuations of aircraft, vessels, and buildings are typically pedestrian evacuations. Smaller-scale evacuations, for example from within a few city blocks, may also be pedestrian-based. Larger-scale evacuations in developed regions will usually involve vehicles. The infrastructure usually provides structure and constraints on the evacuation and the modeling effort. Although transportation mode and infrastructure determine what an evacuation looks like, the same types of decisions and models may be used for pedestrians and for vehicles, for road networks, and within buildings.

Table 1 categorizes three recent evacuations according to these dimensions to illustrate some of the critical differences among evacuations. The evacuation of the WTC on September 11, 2001 was initiated after the precipitating event, that is, after the tower was struck by an attacking aircraft. Between the time of the first impact and the collapse of the second tower, less than 102 min passed [2]. Although evacuation recommendations were, in some cases, delayed or poorly communicated, 13,000–15,000 people exited, traveling down from their offices and leaving the immediate area. The evacuation of the WTC on September 11 was a pedestrian evacuation of a building or set of buildings. Some people in the second tower hit (the first to collapse) had minutes of warning that there was a hazard, so arguably the lead time was up to several minutes. The predictability is categorized as “medium” because although the specific nature of the hazard, the areas that were destroyed or became impassable were not anticipated, the need to evacuate the entire complex had been anticipated; there had even been drills and education of the employees working in the building prior to September 11 incident.

Similarly, the evacuation of the jet that ran off the runway in Denver in December 2008 [3] was not specifically anticipated; however, the need to evacuate the entire aircraft had been considered during the design of the aircraft. The passengers and crew escaped using inflatable slides that were provided specifically to be used in the event of an evacuation.

In contrast, the evacuation of many areas in the Houston and Galveston Texas area in advance of Hurricane Ike in 2008 was initiated for special needs residents over three days in advance of landfall. This was part of a much larger evacuation of the region, in which about a million people left their homes, and many moved tens to hundreds of miles. The evacuation was conducted primarily in personal vehicles, and via a transportation network. Unlike the other two evacuations, it was initiated several days before the hazardous conditions occurred. The storm itself had been tracked for several days prior to the actual evacuation, and preparations and decision making for the evacuation had begun. For major hurricane evacuations, the need for an evacuation may be considered up to a week before the actual impacts.

The remainder of this section briefly describes variables that planners can control that may influence the outcome of an evacuation, and are therefore the decisions whose effects may be analyzed in OR modeling efforts.

Ultimately, a complex network of interacting agents and organizations—ranging from individuals to community organizations to government agencies—collectively determines the course and effectiveness of an evacuation [4,5]. To their frustration, OR analysts cannot dictate individuals’ evacuation behavior, even if armed with an optimal and robust evacuation plan. However, both advance planning and interventions undertaken immediately preceding and during the course of an evacuation can determine constraints on and influence individuals’ evacuation behavior.

Decisions that may be taken after a hazard or threat of hazard necessitating an evacuation becomes known may be informed by details about the specific threat and in some cases, specific information about the population at risk and resources available at the time. The relevant decisions may be called *operational planning*. For large-scale evacuations, some of these actions may even be dynamically updated in real time in response to information about the progress of the evacuation. These actions include the following:

- *Communication.* In many cases, the most important short-term action that officials can take is to communicate with the affected population. Many types of decisions may be involved, such as the type of evacuation, which is mandatory, recommended, or voluntary [6]. In addition, the problem of targeting populations to receive evacuation warnings may be complex. “Shadow” evacuations of those who are not at risk in their homes may interfere with the evacuation of those who need to move, as the overevacuation during Hurricane Rita in 2005 illustrated. Moreover, messages may need to be coordinated across jurisdictional boundaries.
- *Short-Term Modifications to the Network.* Although the basic structure of the physical infrastructure will be determined far in advance of a specific threat, some adjustments in the structure of a transportation network may be made in response to a particular evacuation. This is most applicable in a large-scale evacuation occurring on a road network. With enough lead time, it is possible to change the flow patterns, by opening or closing routes, or by reversing traffic flow on key routes out of the area (contraflow). Within the road network, public transportation such as buses may also be provided. In addition, transportation network capacity may be increased; for example, rail resources may be pre-positioned by rescheduling passenger trains to transport people out of the evacuation zone.
- *Providing Resources to Facilitate/Support Evacuation.* Resources that either speed evacuation or encourage more people to evacuate by making the evacuation more comfortable, may also be deployed within evacuation time-scale. To make evacuation via a road network more efficient, officials may deploy personnel to direct traffic and provide roadside services, including fuel and equipment for removing disabled vehicles. In addition, providing supplies and other resources for evacuees en route or at designated destinations

may encourage evacuation, speed evacuation, or improve the outcomes for the evacuees, thus enhancing the likelihood that they will evacuate in the case of future emergencies. Resources that may be provided en route include food, water, fuel, and other necessities, medical supplies or care; these may be essential during an evacuation in which evacuees spend many hours on the road, as in the 1999 Hurricane Floyd and 2005 Hurricane Rita evacuations. Similarly, providing staffing and provisioning evacuation destinations such as shelters may encourage people to evacuate and then to proceed to these shelters. In the case of a large-scale evacuation, many evacuees travel hundreds of miles to stay with family or friends. This adds to the burden on the road network, and may slow the evacuation. In some cases, providing nearby shelters that people will consider acceptable options may reduce the total number of evacuee-miles traveled, and facilitate the evacuation overall.

Many of the decisions that affect the outcome of an evacuation are made years in advance of any specific hazard that necessitates an evacuation. The most obvious examples are the design of infrastructure, which largely determines evacuation routes and their capacity, and thus long-range planning decisions, described below, often determine the constraints on the short-term actions described above and on individual evacuation behavior.

The main long-range planning decisions that relate to the design of the transportation network or infrastructure. Evacuations are also commonly considered in building design, and may affect the number of exits; number, size, and location of stairwells; even nonphysical parameters such as dedication of “express” elevators; and occupation limits. Evacuation is not usually an important consideration in the planning and design of road networks [6], although this may be changing; OR tools and expertise can contribute, as discussed later.

It should be noted that actions that may be implemented with little or no lead time still require advance planning—the plan itself must be available before the hazard—for example, it is usually impossible to both develop and implement an evacuation plan within the evacuation time-scale. The US Marine Corps faced this problem when Hurricane Floyd threatened Camp Lejeune in coastal North Carolina, and there was no evacuation plan in place. Ultimately, Camp Lejeune did not evacuate its personnel [7]. Wildfires can create a similar situation because they are so unpredictable in timing, extent, and evolution. In general, evacuation plans will not be in place when a wildfire begins.

It may be important to identify and communicate evacuation zones, destinations, and routes in advance, so that the messages at the time of the actual evacuation may be more effective. Within a building or craft, signs or other installed indications of evacuation routes are also long-lead actions that affect evacuation outcomes.

Underlying decisions about how to implement an evacuation are the decisions about whether and when to take action. These decisions should be made on the basis of information about the hazards, possible courses of action, and their consequences. They may (and should) therefore be informed by the results of analyses of the implementation decisions described above. The whether-and-when decisions are rarely addressed in the OR literature, but may be addressed with some OR tools.

OR TOOLS FOR EVACUATION PLANNING

This section describes readily applicable modeling techniques and/or results that have been or may be used to address the decisions outlined above. There has been surprisingly little application of modeling and formal analysis to evacuation planning for transportation networks. Wolshon *et al.* [6] offer several explanations including the fact that mass evacuations are infrequent, relative to other events around which transportation systems are designed. Another explanation is

the expectation that evacuations will necessarily overload transportation networks, and analysis would merely reveal this failure, and therefore cannot add much value.

Most modeling is used either as decision support or in long-range planning. In decision support for operational decision making, model results inform decision makers' subjective decisions. Most of the OR contributions are in the long-range planning stage. Wright *et al.*'s [1] review of OR contributions in homeland security identifies 14 articles from the broader literature (including non-OR journals) that address the evacuation stage of the disaster life cycle. Only one of these is categorized as relating to the "response" phase of the hazard life cycle. The remainder address decisions taking place earlier, during planning prior to evacuation.

The vast majority of existing applications in OR use one of two modeling approaches: network flows, or agent-based simulation. This section outlines the most common application of these approaches to evacuation planning, and describes more briefly some other approaches.

Network Flows

Network flow models are the most common OR tool used in evacuation planning (see the section titled "Network Flows" in this encyclopedia). Network flows have been applied primarily to small-scale evacuation problems, such as building evacuation, although network flows have been used for large-scale road networks as well.

The infrastructure in which the evacuation occurs is modeled as a network, and the flows consist of evacuees, who may be pedestrian or vehicular. In general, nodes represent sources of evacuees and destinations. Nodes also represent transshipment points, which are usually physical flow-through points, such as stairways or highway intersections, interchanges, and merge points. Depending on the model, arcs (usually directed) may represent physical flow-through points, or may represent the possibility of travel between two nodes, and may be capacitated.

In a static network flow evacuation model, arc costs usually represent transit time or

distance, and the solution minimizes some evacuation-time-related objective, such as time-to-destination of last evacuee, or mean evacuation time. In some cases, for example, Yamada [8] and Han *et al.* [9] the problem reduces to an assignment problem.

More common in evacuation planning is the dynamic network flow model, reapplied at multiple time-steps, as the flow moves through the network. Generally, nodes and/or arcs are capacitated—by flow per unit time—and arcs are cost-less. Chalmet *et al.* [10] provide the first example in the OR literature, and make use of Jarvis and Ratliff's [11] result showing that in a dynamic network flow, minimizing the time to complete flow through the network simultaneously minimizes the mean (or other monotonic time-weighted sum of) evacuation time and maximizes the flow completed at every time-step. Chiu *et al.* [12] provide the most recent example, and bridge the gaps between the civil engineering and OR languages and approaches to evacuation modeling.

A dynamic model allows for changes over time in the underlying network, such as implementation of short-term network changes (e.g., contraflow), the closure of parts of the network due to disaster impacts (e.g., flooding cutting off hurricane evacuation routes) or changes in capacity due to overcrowding or other effects.

Either static or dynamic network flow models may be used to optimize flows within a network as given, or to test the effect of changes in the network. Within a given network, an optimal network flow solution may yield an evacuation plan, which may be used to assign evacuees to routes or destinations. The solution may be implemented by designating evacuation routes or destinations in advance of the need for an evacuation, either by communicating to the public in advance or during an evacuation either in a permanent (e.g., building exit sign) or temporary (e.g., mobile highway sign) fashion.

The optimal network flow may also be used to estimate evacuation completion times, and indicate the lead time required to complete an evacuation before disaster occurs or interferes with evacuation.

Network flow models may also be used iteratively in design of the network, for example, in identifying bottlenecks, sizing points of egress, or deciding whether to implement contraflow. A network flow analysis may also be used during design or redesign of infrastructure, the network solution may identify bottlenecks and therefore indicate areas that should be expanded or otherwise improved. The network analysis may be run varying the conditions of the evacuation. For example, the sources may vary with the nature or location of the hazard. In this way, the network model may be used as a tool in infrastructure design.

Usually, the optimization would be completed in advance, but in the case of a proactive evacuation, a network optimization could be completed using hazard-specific information about sources and capacities (accounting, for example, for construction or other short-term changes in the network) and its results implemented in real time. The results would be difficult to implement in a building, but in the case of evacuation via a road or rail network, it may be possible to either require or recommend evacuation routes or destinations with short-term planning.

Network flow modeling is the most common approach to evacuation planning appearing in the OR literature; this is consistent with the review of disaster operations management research in OR/MS generally [13], which found that mathematical programming was by far the most common modeling approach in the broad literature surveyed, with probability and statistics a distant second, and simulation third.

Agent-Based Models

In contrast, civil and transportation engineering researchers use simulation as their primary quantitative modeling approach for evacuation planning. The most common approach—especially in recent years—is agent-based simulation in which multiple decision making agents, with independent rules of behavior, interact within a modeled infrastructure. With increasing computing power and commercialization of simulation software, agent-based simulation is

applicable to planning for both large- and small-scale evacuations.

Agent-based modeling contrasts with network flow modeling in that it can (at additional computational cost) incorporate more and more detailed elements of the transportation and evacuation systems. For example, agent-based simulations can include diverse individual behavior patterns that have a substantial impact on evacuation time estimates [14] and highly detailed descriptions of the infrastructure [see Ref. 15, e.g., a model of the effect of traffic signal timing]. Some of these models combine mathematical programming to model individual decisions with agent-based simulation, to describe the cumulative impact on the larger system [16]. Effects such as capacity reductions caused by congestion emerge from agent-based simulations.

Another reason for simulation's dominance among civil and transportation engineering studies is that the purpose in evacuation modeling is often to evaluate designs of modifications to the physical infrastructure, rather than control and/or policy selection. This is changing with the growing importance of intelligent transportation systems, allowing for real-time control using optimized policies or direct real-time human intervention [17].

Agent-based models are already heavily in use in emergency planning. These models are already commercialized and operational; see Refs 18 and 19, which describe models in use as decision support systems (DSS) for emergency managers, in long-range transportation planning and in academic research. They use the term "micro" models to describe agent-based models, because until recently, the computational requirements of agent-based models restricted their use to modeling small-scale evacuations, or only a small portion of a larger evacuation.

Most, if not all, of the models they describe as "macro" or "meso" models use empirical relationships to describe the capacity of elements of the road network, as a function of the traffic geometric design (the physical structure of the road network) and control conditions (e.g., traffic lights).

A few of the DSS use the output of agent-based simulations as input (e.g., Hurrevac, www.hurrevac.com).

Logistics Models

As discussed above, OR tools that are used to plan for supply chain of relief supplies or services may be relevant to evacuation planning, because they may encourage evacuation and make it more efficient and effective. Logistics for evacuation is challenging because the network must be set up with a short lead time, and is intended to be temporary.

These challenges are just the sort that the OR community has begun to tackle under the rubric of disaster logistics or humanitarian logistics. Both *Interfaces* (volume 36, issue 6, November–December 2006) and *IIE Transactions* (volume 39, issue 1, January 2007) have offered special issues on applications to homeland security. The problems treated include several that could apply to logistics support for evacuations, such as setting up supply chains with short notice (e.g., for medical response to bioterrorism incidents). Moreover, there are many recent initiatives in humanitarian logistics that address supply chain, inventory [20], routing [21], and facility location problems [22]; see also the section titled "Introduction to facility location" in this encyclopedia that could be applicable to providing for evacuees.

Stochastic Modeling and Optimization

In the case of proactive evacuations, there may be flexibility in the timing of the initiation of an evacuation [23]. Hurricane evacuations are an obvious example, and evacuations following a disaster such as an earthquake may also allow officials flexibility in deciding when to initiate the evacuation. OR modeling may help in deciding when to evacuate, a decision that may be made in conjunction with or with cognizance of other evacuation planning decisions. OR approaches such as stochastic modeling and optimization may be employed to support this decision, as Regnier and Harr [24] illustrate in a stylized example.

Decision Support Systems

As discussed earlier, Wright *et al.* [1] cited only one OR article addressing the “response” phase of an evacuation. The response phase encompasses all operational planning, that is decisions that may be made in the evacuation time-scale, immediately before, during, and immediately after the precipitating event [5]. Reference 12 is an important recent contribution to the response phase: it offers a dynamic network flow model that can be implemented in real time to generate optimal decisions about destination and route assignment. They also offer a good overview of prior evacuation research and models underlying existing operational decision support.

The remainder of the OR contributions identified by Wright *et al.* [1], however, are in DSS; that is, anticipated disaster scenarios are determined with advance analysis, and called up in real time to provide human decision makers with assessments of the consequences of decisions so that they can consider the effects, within the DSS model, of various alternatives before subjectively choosing the best, based on their judgment, training, and experience with prior evacuations, and with the model. Many of the DSS integrate other types and sources of information, such as weather forecasts (in Hurrevac, www.hurrevac.com) and geographic information systems (GIS). De Silva and Eglese [25] and de Silva *et al.* [26] describe some of the issues and approaches in GIS-based decision support.

Although the existing DSS have been developed in the civil engineering community (and are reviewed by Hardy and Wunderlich [19]), it is an area that is developing rapidly (Wolshon, personal communication) with increasing use of agent-based simulation and better user interfaces, and integration among models and data sources. The OR community has much to offer in this area.

Multiple-Objective Models

Evacuation planning involves many competing objectives. Many of these can be thought of as simply balancing costs against effectiveness; for example, investing in

infrastructure or providing services and supplies to evacuees are costly but may speed evacuation. Most evacuations have complex socioeconomic consequences and therefore, naturally lend themselves to the benefits of structured processes for identifying important underlying objectives, especially in group contexts, and for balancing these in optimization [27]; see also the sections titled “Normative Structuring of Decision Problems” and “Decision Aids for Improving Group Decisions” in this encyclopedia. The many, often competing, objectives involved in evacuation planning may include speed, completeness, voluntariness, compliance rates for current and future evacuations, as well as security and equity (e.g., priority or support for subpopulations).

CHALLENGES AND LIMITATIONS

This section briefly highlights some of the challenges for researchers and practitioners and limitations to the application of OR in evacuation planning as posed by complications in real-world evacuations that make extending OR models to evacuation modeling nontrivial.

Perhaps the biggest challenge is human beings: ultimately, a collection of individuals implements an evacuation. Identifying an optimal policy is useless unless it can be communicated effectively to the individuals who implement the evacuation. Ensuring that there are conduits for messages to the individuals involved in implementation may itself be a key decision problem. Increasing the effectiveness of the communication, that is, the likelihood that the individuals will respond as intended, may also be challenging.

It is highly unlikely that all or even most of those involved in an evacuation will follow a globally optimized policy, even if it is well-communicated and logistically supported. Agent-based simulation can address some of these issues, by using research on human behavior as the basis for agent models. Like any model, these will be imperfect and subject to the limitations of the available data about human behavior, which is especially limited for behavior in unusual circumstances such as disasters.

In addition to the inability to dictate individuals' behavior, there are other complications of modeling the flow of human beings. For example, an optimal network flow might have flow from a single node (whether source or transshipment node) on multiple arcs. This would be difficult to implement or communicate when the flow is human, though it might emerge naturally by individual choice.

For large-scale evacuations, implementing OR solutions can be difficult because of multiple loci of authority. In the United States, even for a relatively small area to be evacuated, there may be multiple layers of government, each with one or more agency involved. If the evacuation zone crosses jurisdictional boundaries, additional authorities with different roles, rules, and systems come into play, and coordination becomes even more difficult. Both within and across jurisdictional boundaries, officials in different roles may also be separated by the professional language and culture of their area of specialization. Moreover, nonpublic officials, such as churches and other community organizations may be involved. The Washington D.C. metropolitan area, which includes two states, the district itself, and 17 city or county governments attempted to develop a single evacuation plan for the event of a terrorist strike. Ultimately, officials abandoned the attempt, and settled for documenting the individual entities' plans [28]. The most important future contributions, with the greatest practical impact, will likely require inputs and interaction from emergency management, civil and transportation engineers and researchers from other disciplines including social sciences.

There are also technical challenges to modeling and/or designing transportation networks, such as the following:

- Emergency responders and logistics may use the same routes (simultaneously or near-simultaneously) that evacuees are using. Moreover, the infrastructure and the solution must both allow multiple types of flow.
- In a related problem, different types of models may be appropriate for logistics supply chain and transportation

networks. The most effective global solution to evacuation planning might require integrating multiple models and optimizing for many complex, inter-related decisions [30]. Simulation may be used to integrate multiple model types, but this of course increases computational requirements and complicates the definition of a solution and the appropriate analysis.

- Disruptions to the network are more likely in the case of an evacuation than in other transportation planning problems, and network disruptions may be uncertain. For example, a bridge might go down, or an evacuation shelter might be flooded or destroyed. To the extent that these disruptions are predictable—either deterministically or stochastically—they may be handled with dynamic networks or agent-based simulation, but these considerations expand the scope of the analysis.
- For analyses completed before a specific evacuation, the evacuation scenario will be unknown; for example, sources, supply at each source, available routes, available destinations may not be known, but paths may need to be determined *a priori*. Therefore, for certain types of planning, robustness of decisions to potential evacuation scenarios is important [29].
- Another potentially critical problem is the number of people who, because of their health or lack of access to resources, do not have the means to self-evacuate.

Evacuation planning includes many sub-problems that fit into the framework of OR tools, including network modeling, stochastic optimization, and large-scale agent-based simulation. At the same time, evacuation planning poses many challenges to a straightforward application of these or any purely quantitative tools. Many of the assumptions required for the application of even complex OR models—such as the assumption of central planning and a single objective function—break down in this context.

There are many challenges for finding OR solutions that can be implemented in practice. Evacuation planning problems

- may be subject to substantial uncertainties about key parameters;
- involve many decision makers, often including multiple authorities and many legitimate stakeholders;
- are dependent on human behavior that may be difficult to predict and for which few data are available to support modeling;
- involve many interacting subsystems, which may require different modeling approaches;
- require extensive and ongoing collaboration with other disciplines and with multiple stakeholders.

However, these challenges play to the strengths of some within the broad community of OR researchers and practitioners. In particular, evacuation planning can benefit from the OR community's experience and expertise in structuring problems with multiple interacting subsystems, identifying robust solutions to complex problems, and playing a consulting role in public-sector and multistakeholder contexts.

REFERENCES

1. Wright DP, Liberatore MJ, Nydick RL. A survey of operations research models and applications in homeland security. *Interfaces* 2006;36(6):514–529. DOI: 10.1287/inte.1060.0253.
2. The 9/11 Commission (The National Commission on Terrorist Attacks Upon the United States) The 9/11 Commission Report. Available at: <http://www.9-11commission.gov/report/911Report.pdf>; <http://govinfo.library.unt.edu/911/report/911Report.pdf>. Accessed December 21, 2008.
3. Wald ML. 5 still in hospital after Denver jet crash. *The New York Times*. Available at: http://www.nytimes.com/2008/12/22/us/22crash.html?scp=5&sq=denver_plane_crash. December 21, 2008.
4. McEntire DA. Local emergency management organizations. Chapter 10. In: Rodríguez H, Quarantelli EL, Dynes RR, editors. *Handbook of disaster research*. New York, (NY): Springer; 2006. pp. 168–182.
5. Tierney KJ, Lindell M, Perry R. Facing the unexpected: Disaster preparedness and response in the United States. Washington, DC: Joseph Henry Press; 2001. p. 318.
6. Wolshon B, Urbina E, Wilmot C, Levitan M. Review of policies and practices for hurricane evacuation I: Transportation planning, preparedness and response. *Nat Hazards Rev* 2005;6(3):129–142. DOI: 10.1061/~ASCE!1527-6988.
7. Taylor BJ. Eastern North Carolina Marine Corps Forces and Installations high intensity hurricane evacuation decision support [MS Thesis]. Naval Postgraduate School is the university. Place is Monterey, California, USA, 2007.
8. Yamada T. A network flow approach to a city emergency evacuation planning. *Int J Syst Sci* 1996;27(10):931–936.
9. Han LD, Yuan F, Chin SM, *et al.* Global optimization of emergency evacuation assignments. *Interfaces* 2006;36(6):502–513. DOI: 10.1287/inte.1060.0251.
10. Chalmet LG, Francis RL, Saunders PB. Network models for building evacuation. *Manage Sci* 1982;28(1):86–105.
11. Jarvis JJ, Ratliff HD. Some equivalent objectives for dynamic network flow problems. *Manage Sci* 1982;28(1):106–109.
12. Chiu YC, Zheng H, Villalobos J, Gautam B. Modeling no-notice mass evacuation using a dynamic traffic flow optimization model. *IEE Trans* 2007;39(1):83–94. DOI: 10.1080/07408170600946473.
13. Altay N, Green III WG. OR/MS research in disaster operations management. *Eur J Oper Res* 2006;175(1):475–493.
14. Lindell MK, Prater CS. Critical behavior assumptions in evacuation time estimate analysis for private vehicles: Examples from hurricane research and planning. *J Urban Plan Dev* 2007;133(1):18–29.
15. Chen M, Chen L, Miller-Hooks E. Traffic signal timing for urban evacuation. *J Urban Plan Dev* 2007;133(1):30–42.
16. Murray-Tuite P. Perspectives for network management in response to unplanned disruptions. *J Urban Plan Dev* 2007;133(1):9–17.
17. Chiu YC, Mirchandani PB. Online behavior-robust feedback information routing strategy for mass evacuation. *IEEE Trans Intell Transp Syst* 2008;9(2):264–274.

18. Hamacher HW, Tjandra SA. Mathematical modeling of evacuation problems: A state of the art. In: Schreckenberg M, Sharma SD, editors. *Pedestrian and evacuation dynamics*. Berlin: Springer; 2001. pp. 227–266.
19. Hardy M, Wunderlich K. Evacuation management options (EMO) modeling assessment: Transportation modeling inventory. Research and Innovative Technology Administration, Intelligent Transportation Systems Program Office. Available at: http://www.its.dot.gov/its_publicsafety/emo/emo.pdf. Accessed 2010.
20. Beamon BM, Kotleba SA. Inventory modeling for complex emergencies in humanitarian relief operations. *Int J Logist: Res Appl* 2006;9(1):1–18. DOI: 10.1080/13675560500453667.
21. Campbell AM, Vandenbussche D, Hermann W. Routing for relief efforts. *Transp Sci* 2008;42(2):127–145. DOI: 10.1287/trsc.1070.0209.
22. Balcik B, Beamon BM. Facility location in humanitarian relief. *Int J Logist: Res Appl* 2008;11(2):101–121. DOI: 10.1080/13675560701561789.
23. Regnier E. Public evacuation decisions and hurricane track uncertainty. *Manage Sci* 2008;54(1):16–28.
24. Regnier E, Harr PA. A dynamic decision model applied to hurricane landfall. *Weather Forecast* 2006;21(5):764–780.
25. De Silva FN, Eglese RW. Integrating simulation modeling and GIS: Spatial decision support systems for evacuation planning. *J Oper Res Soc* 2000;51(4):423–430.
26. De Silva FN, Eglese RW, Pidd M. Evacuation planning and spatial decision making: Designing effective spatial decision support systems through integration of technologies. Chapter 21. In: Mora M, Forgionne GA, Gupta, JND, editors. *Decision making support systems: Achievements, trends and challenges*. London: Idea Group Publishing; 2002. pp. 358–373.
27. Keeney RL. Structuring objectives for problems of public interest. *Oper Res* 1988;36(3):396–405.
28. Sheridan MB. Guide to evacuate region reveals limitations. *The Washington Post*, February 4, 2008, B1.
29. Andreas AK, Smith JC. Decomposition algorithms for the design of a nonsimultaneous capacitated evacuation tree network. *Networks* 2008;53(2):91–103.
30. Hossam A, Bohar A. Emergency evacuation planning as a network design problem: A critical review. *Transp Lett Int J Transp Res* 2009;1(1):41–58.

EVALUATING AND COMPARING FORECASTING MODELS

SHOUYI WANG
WANPRACHA ART CHAOVALITWONGSE
Department of Industrial and Systems
Engineering, Rutgers University,
Piscataway, New Jersey

INTRODUCTION

Strategic planning for the future is extremely important in any business to make definitive action plans. Owing to its great importance, the knowledge of forecasting has been growing very rapidly in modern times. Many forecasting models have been developed to empower people in decision-making for various application areas. For example, accurate demand forecasts are essential for manufacturers to determine the optimal production rate by making a trade-off between stock-outs and high inventory levels. Successful climate forecasting models have been developed to provide early warnings of adverse climatic conditions, such as hurricanes, storms, or fogs [1]. In business activities, forecasting technologies have become indispensable tools in a wide range of managerial decision-making processes, such as finance, banking, investments, employment, mortgages, and loans [2].

Forecasts can be made based on either empirical qualitative analysis or mathematical quantitative analysis. Accordingly, forecasting models can be broadly classified as qualitative methods and quantitative methods. The flowchart of a typical forecasting modeling process is shown in Fig. 1. And the categorization of forecasting models is shown in Fig. 2. Fildes and Lusk [3] argued that no forecasting method could be the “best” method among the various competing forecasting methods. All forecasting methods have distinct advantages and disadvantages. Therefore, selecting the right forecasting

method is of critical importance to all decision makers. This article briefly reviews the most influential forecasting models, and discusses the evaluating and comparing criteria/algorithms for model selection.

QUALITATIVE METHODS

Qualitative forecasting techniques rely primarily on human judgment based on expertise, experience, or intuition. They can be used in a wide range of circumstances where historical data are not available, or circumstances that are changing so rapidly that a mathematical forecasting model based on past data may become irrelevant or questionable. A number of qualitative forecasting methods have been developed to make good use of qualitative judgments from experts. Five of the more popular qualitative methods are briefly presented in the following.

Delphi Method

The Delphi method is a group technique, which is based on structural surveys of a panel of experts about their perceptions of future events [4]. All of the experts answer a series of survey questions anonymously in several rounds. In the first round, the panel members are asked to write down their intuitive forecasts on the survey papers. Then, all the responses are summarized and fed back to each of the panel members. In the next round, the panel members may modify their original forecasts based on the responses of other panel members. This process usually generates a narrowing of opinions, and can be stopped by a predefined criterion, such as maximum number of rounds, degree of consensus, or stability of results. The final result of the process can be the mean or median results of the final round. The Delphi method provides more accurate forecasts than face-to-face group discussions, which are often highly influenced by those experts who have the best interaction and persuasion

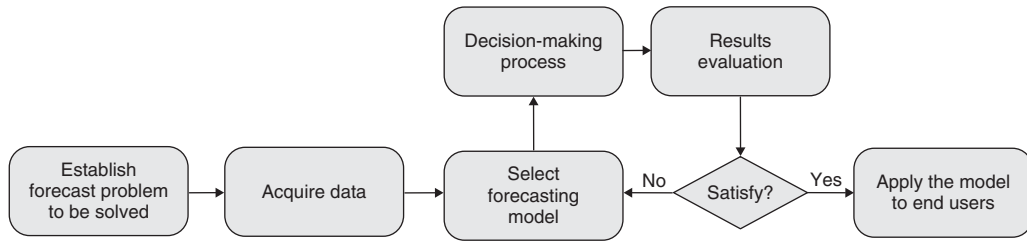


Figure 1. Flowchart of a typical forecasting modeling process.

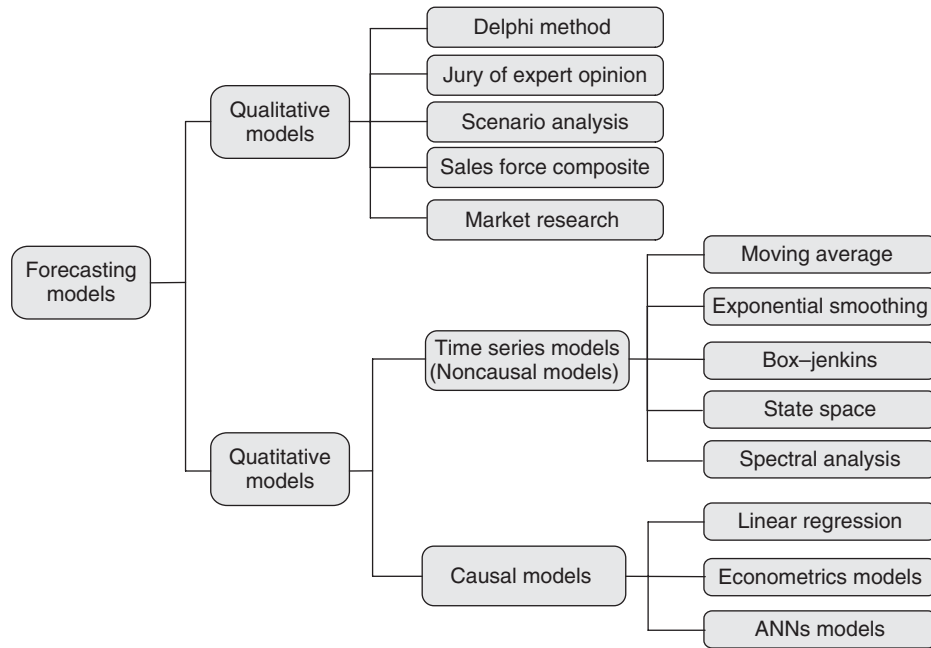


Figure 2. Categorization of forecasting models.

skills, or powerful background. With the advances in collaborative tools, such as e-mail, synchronous video conference, and web conference, the Delphi Method is very easy to implement in modern times. However, one drawback of the Delphi method is that it is usually very time consuming.

Jury of Expert Opinion

The jury of expert opinion is a method of developing forecasts in which the executives of a company are polled for their best estimate of the future trends, such as likely sales and demands [5]. It is also known as the *jury of*

executive opinion method. This method is one of the simplest and widely used forecasting methods in business. A typical outline of the “jury of expert opinion” method is as follows. First, each executive gets familiar with the background information as much as possible. Then in the meeting, all the executives write their estimates on survey papers. The final step is to combine all the options to produce an average and acceptable result among all the executives. The “jury of expert opinion” method can be considered as an informal variant of the Delphi method. The only difference between them is that the “jury of expert

opinion” method does not have a mechanism to prevent interactions amongst the meeting participants.

Scenario Analysis

Scenario analysis is a systematic thinking process to discover how various factors may function together to create the future [6]. Different scenarios can be generated by combining different policy plans with all known facts and anticipated changes about the future. Scenario analysis can often lead to plausible forecasts if the causal relationship between factors and their consequences have been appropriately simulated. Decision-makers can learn from the outcomes of different scenarios without risking failures and losses in real business. To approach scenario analysis, the major steps are briefly as follows:

1. Determine the overall assumptions and factors on which the scenarios will be built. All assumptions and factors should be carefully made so that they can be conveniently under the user’s control and inspection. A time scope of the scenarios are also given.
2. Combine all the selected factors and assumptions together to construct a baseline framework. Each setting of the factors corresponds to a scenario. Given different levels of the influential factors, the outcome evolution of each scenario can be quite different.
3. Compare outcomes of the scenarios and pick up the most sensibly fitted scenarios. This step usually requires a considerable amount of debate to determine the selected scenarios.

Sales Force Composite

Other than management and administration members, sales force can also be an important factor for qualitative forecasting analysis. The most popular method is known as *sales force composite*, which is based on forecasts of salespeople [7]. The assumption of the “sales force composite” method is that salespeople are most qualified to predict future

market development in their own territories. Since salespeople interact directly with customers, they are presumed to have good estimates of market changes and trends. One possible disadvantage of this method is that forecasting results of salespeople tend to be optimistic. This is mainly because salespeople may always choose an optimistic prediction if a low estimate could endanger their employment status. Moreover, since the sales force in one department may not be aware of impending changes in other areas, the opinions of salespeople in different territories should be combined into consideration in decision making.

Market Survey

To forecast future trends, another popular qualitative approach is to ask customers or potential users how they foresee their future consumption of a product or service. It is also called the *user’s expectation* method. Since there is no benefit conflict, individual users are supposed to provide forecasts without undue optimism. Market survey has gained its popularity and importance in forecasting the success of new products [8]. However, one problem is that user opinions may be heavily biased by the trend of the moment. For example, most people have pessimistic opinions during periods of recession and have optimistic opinions in periods of growth and prosperity. Another disadvantage of the “market survey” method is that it is often very expensive to obtain a comprehensive survey of customer expectations.

QUANTITATIVE METHODS

Quantitative methods make forecasts based on mathematical models rather than subjective judgment. These methods are the mainstream of forecasting techniques as a result of the great advances in mathematical modeling and computational power in modern times. Quantitative forecasting models have been utilized across a wide spectrum of business and industry. As shown in Fig. 2, quantitative methods can be classified as noncausal models and causal models [9]. Noncausal models are also known as *time-series models*, which

make forecasts by extracting systematic patterns (such as trends and seasonality) from historical time series data. Causal models are also known as *cause-and-effect models*, which investigate how the variable being forecasted is determined by its relevant influential factors. We will take a brief look at the leading quantitative methods in the following.

Time Series Models

Time series models predict values of the variable being forecasted based on historical patterns. Thus the basic underlying assumption of these methods is that future patterns are similar to historical patterns. To extract characteristics of time series patterns, four basic properties of data are often analyzed.

- *Trend*. Given a set of time series data, the term *trend* refers to a stable tendency of growth or decline exhibited in the data. The trend of a time series can be either linear or nonlinear. Accordingly, linear and nonlinear functions can be utilized to model the trend.
- *Seasonality*. If a pattern always repeats at a fixed interval, it is called a *seasonal pattern*. Seasonality is a very common characteristic of time series data. For example, air temperature exhibits a strong yearly seasonal pattern.
- *Cycles*. Cyclic patterns are similar to seasonal patterns, except that they repeat at varying intervals. For example, it is common to find nonstationary cycles in financial time series data.
- *Randomness*. Most time series data are assumed to contain both systematic patterns and random noises. The randomness usually makes the pattern difficult to identify. Most time series models include a noise term to take into account the effects of randomness.

Various time series methods have been developed to analyze these properties of time series data. Five of the most popular ones are moving average (MA), exponential smoothing, Box–Jenkins models, state–space models, and spectral methods.

Moving Average (MA). MA models are simple but popular forecasting methods in time series analysis. An MA model involves taking the arithmetic average of N most recent observations, where N is a specific number according to the nature of the data to be forecasted. For example, if you are forecasting monthly sales, you might use 12-month MA model, which takes the average sales over the past 12 months. A one-step-ahead MA model of N periods is given by

$$\begin{aligned} F_{t+1} &= \frac{1}{N} \sum_{i=t-N+1}^t x_i \\ &= \frac{x_t + x_{t-1} + \cdots + x_{t-N+1}}{N}, \end{aligned} \quad (1)$$

where F_{t+1} is the forecast for the $t+1$ th period, and x_i s, $i = t-N+1, \dots, t$ are the observations in the past N periods. The mean of N most recent observations is used as the forecast of the next period. The MAs method is probably the most commonly used technique to smooth out short-term fluctuations and capture characteristics of varying trends in time series data.

Exponential Smoothing. Exponential smoothing assigns exponentially decreasing weights as observations getting older. The most commonly used single exponential smoothing is given by

$$F_0 = x_0, \quad (2)$$

$$F_{t+1} = \alpha x_t + (1 - \alpha)F_t = F_t + \alpha(x_t - F_t), \quad (3)$$

where $\alpha \in [0, 1]$ is the smoothing factor, F_{t+1} is the new forecast for next period, x_t is the current observation at period t , and F_t is the last forecast made in period $t-1$. In the above formula, one can substitute $F_t = \alpha x_{t-1} + (1 - \alpha)F_{t-1}$, and continue so forth to obtain the infinite expansion of F_t as follows

$$F_t = \sum_{i=0}^{\infty} \alpha(1 - \alpha)^i x_{t-i-1} = \sum_{i=0}^{\infty} \alpha_i x_{t-i-1}, \quad (4)$$

where $\alpha_i = \alpha(1 - \alpha)^i$. From this expression, one can see clearly that the weight α_i

decreases exponentially with time. This illustrates why this method is called *exponential smoothing*. Single exponential smoothing works best only for stationary time series data without trends and seasonality. There are two variants of exponential smoothing, known as double and triple exponential smoothing, respectively. In particular, double exponential smoothing can be applied to handle time series data with linear trends, and triple exponential smoothing is capable of dealing with both linear trends and seasonality. A very detailed discussion of exponential smoothing techniques can be found in Gardner [10,11].

Box–Jenkins Methods. Box–Jenkins methods, named after the statisticians George Box and Gwilym Jenkins, who applied autoregressive moving average (ARMA) to make forecasts for time series data [12]. An ARMA model can be generally described by

$$\begin{aligned} x_t &= c + \epsilon_t + a_1x_{t-1} + a_2x_{t-2} + \cdots + a_px_{t-p}, \\ &\quad - b_1\epsilon_{t-1} - b_2\epsilon_{t-2} - \cdots - b_q\epsilon_{t-q}, \quad (5) \\ &= c\epsilon_t + \sum_{i=1}^p a_i x_i - \sum_{j=1}^q b_j \epsilon_j, \end{aligned}$$

where x_t is the current observation, $x_{t-1}, \dots, x_{t-p}$ are the observations in the past p periods, and the $a_1, \dots, a_p$ are the regression coefficients of the past p observations. The ϵ_t is the current prediction error, the $\epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-q}$ are the past q prediction errors, and the $b_0, b_1, \dots, b_q$ are the associated regression coefficients. ARMA models assume that the time series to be analyzed is stationary. To handle the non-stationarities such as trend and seasonality, Box and Jenkins proposed a differenced version of ARMA model, which is known as the *autoregressive integrated moving average* (ARIMA) model. The “I” stands for “integrated,” since the estimation process is performed on differenced data, and the time series needs to be integrated before making a forecast. For more mathematical details of the Box–Jenkins model, refer Box *et al.* [12].

State–Space Models. State–space models virtually build up a generalized representation of linear time series models in state space form. For example, one can represent an ARIMA model in state space form. Once a state space model is built, it can be conveniently analyzed by Kalman filter and the associated smoother. Kalman filter is a recursive procedure to compute the optimal estimator of the state vector at time t , based on the information available at time t . The parameters of a state space model are usually estimated by maximum likelihood functions. The state space method is a sophisticated form of forecasting models. A detailed discussion of various algorithms of state space models can be found in Harvey [13].

Spectral Analysis. Spectral analysis represents a group of methods, which decompose time series data into a few underlying sine and cosine functions of different frequencies. Compared to ARIMA or exponential smoothing techniques, for which the seasonal period is known as *a priori* in the analysis, spectrum analysis is suitable to deal with the seasonal series data for which lengths of cyclic patterns or fluctuations are changing rapidly or are difficult to estimate. In spectral analysis, some important recurring cycles of different frequencies in the time series can be discovered. Those patterns may be hidden in random noises and are extremely difficult to find out by other methods. The most common spectrum decomposition process is also referred as *Fourier analysis*, which can be considered as a linear multiple regression process. The dependent variable is the time series to be studied, and the independent variables are sine and cosine functions of all possible frequencies. In general, a spectral decomposition model is given by

$$\begin{aligned} x_t &= c + \sum_{j=1}^k (a_j \cos(\omega_j t) + b_j \sin(\omega_j t)), \\ 0 &< \omega_1 < \cdots < \omega_k < \pi, \quad (6) \end{aligned}$$

where $\omega_1 < \cdots < \omega_k$ are k possible wavelengths of the cyclic patterns in the time series, $a_1, a_2, \dots, a_k$, and $b_1, b_2, \dots, b_k$ are regression coefficients that represent the

degree of corresponding sinusoidal functions are correlated with the data. In other words, a large sine or cosine coefficient indicates a strong periodicity of the respective frequency in the data. There are various techniques available to perform spectral analysis for time series data. A comprehensive discussion of spectral analysis can be found in Koopmans [14].

Causal Models

Causal models are also known as *cause-and-effect models*, which establish a causal relationship between the variable being forecasted and all other related variables. The most well-known causal models are called *regression models*, which build up a rigorous mathematical model of causal relationship based on sound statistical techniques. In a regression model, the variable to be forecasted is called *dependent* or *response variable*, and the variables that represent the causal factors of the dependent variable are called *independent* or *explanatory variables*.

Regression Models. Regression models are in principle to investigate the relationship between one dependent variable and its relevant independent predictor variables. In general, a standard regression model takes the form as follows

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n, \quad (7)$$

where Y is the forecasted value of the dependent variable, b_0 is the intercept, and $b_1, b_2, \dots, b_n$ are the estimated regression coefficients representing the contribution of the independent predictor variables $X_1, X_2, \dots, X_n$, respectively. When linear regression models do not appear to adequately capture the relationships between dependent and predictor variables, nonlinear regression models (such as polynomial regression) can be used. An overview of the popular nonlinear regression models can be found in Seber and Wild [15].

Econometrics Models. In many real problems, the cause-and-effect relationship between dependent and independent variables are not straightforward. The estimated

model parameters by the standard regression analysis may become inappropriate due to the highly dynamic relationship between the dependent and independent variables. For example, we wish to forecast sales of a product, which is related to its price. However, the market price in turn is also affected by sales. Another typical example is the supply and demand model. The interaction of supply and demand jointly determines the equilibrium price and quantity of the product in the market. In such cases, it no longer makes sense to separate dependent and independent variables completely. To handle this problem, a set of simultaneous regression models is necessary to describe dynamics of these systems. The simultaneous regression models are called *econometrics models* in literature, since they are often applied to analyze the relationships between economic variables that should be jointly determined. One can consider that a single regression model is a special case of econometrics models. The rigorous mathematical formulations of econometrics can be found in Pindyck and Rubinfeld [16].

ANN Models. Artificial neural networks (ANNs) represent another important form of causal models, which have shown powerful capabilities of modeling complex relationships between inputs and outputs. An ANN model consists of a network of neurons connected by arcs with assigned weights. Neurons take some form of basic nonlinear functions. Therefore, an ANN model can be equivalently considered as a nonlinear regression model in mathematics. A typical ANN has three layers, an input layer, a hidden layer, and an output layer. As a causal model, the inputs to an ANN are independent or explanatory variables, and the outputs are dependent or response variables being forecasted. There are various algorithms available to train ANNs, such as Perceptron learning rule and backpropagation [17]. Once the structure and weights of an ANN are determined, it can be employed to perform forecasting. ANNs have been increasingly used in forecast modeling in the past decade. They are suitable for complicated problems, which are difficult to

be mathematically formulated by regression models or econometric models. In many real applications, ANN methods can often achieve good performance if given enough training data. An overview of the applications of ANNs in forecasting can be found in Zhang *et al.* [18].

COMPARING AND SELECTING FORECASTING MODELS

We have discussed the most popular forecasting models above. Table 1 below summarizes the appropriateness of the three major types of forecasting models. To evaluate forecast models, two terms that are often of concern are accuracy and bias. *Accuracy* refers to the distance between the forecasts and actual values. And a forecast is biased if the errors in one direction are significantly larger than those in other directions. In general, the basic objective of all forecast models is to maximize accuracy and minimize bias. After fitting several model candidates to a given data set, the next step is to compare and select the best forecasting model. A lot of criteria have been proposed to compare forecasting models, which will be discussed in the following.

Forecast Error Measures

To achieve high prediction accuracy is the primary objective in most forecasting tasks. To evaluate forecasting accuracy, four of the more popular direct error measures are mean squared error (MSE), and its variants root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage

error (MAPE). Minimizing these measures is usually the most essential criterion in comparing forecasting models. These measures are most frequently used due to their mathematical convenience. For each of these measures, a smaller value indicates higher prediction accuracy. Given a set of real data $y_i, i = 1, 2, \dots, n$, each of which has an associated forecast value $\hat{y}_i$, then these measures are defined as follows

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (8)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad (9)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (10)$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \quad (11)$$

The R-squared (R^2), also known as *coefficient of determination*, is another most commonly used criterion to evaluate a forecast model. The most general form of the R^2 is defined as follows

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}, \quad (12)$$

where $\bar{y} = \sum_{i=1}^n y_i / n$, it is the mean of the observed data. The second term compares the variance of the forecast errors with the

Table 1. The Applicable Situations for Different Types of Forecasting Models

Model Type	Applicable Situations	Horizon	Data Types
Qualitative models	No past data, no ideas about causality, data too costly to collect, short-run accuracy not required.	Either long-run or short-run	Background information, survey
Time series models	Historical time series data available, stable patterns exist, short-run accuracy needed	Short-run	Time series data
Causal models	Past data available, causal relationship is clear and stable, explanatory variables are controllable	Short-run	Response and explanatory data

total variance of the data. One minus the second term is the proportion of variability in a data set that can be explained by the forecast model. R^2 is used to measure how well a model approximates real data values. The magnitude of R^2 is usually restricted within 0 and 1. An R^2 close to 1.0 indicates that the model perfectly fits the data, while an R^2 close to 0 means that the model cannot explain the data at all.

Information Criterion

A number of model selection criteria have also been developed based on information theory. A well-known criterion is Akaike's information criterion (AIC) developed by Akaike in 1974 [19]. It makes a trade-off between accuracy and complexity in model construction. In general, this criterion can be defined as

$$\text{AIC}(K) = \log(\text{MSE}) + \frac{2K}{n}, \quad (13)$$

where K is the number of variables in the model, n is the number of observations in the data. The MSE here is the same as defined in (9) and it can be explained as the estimated residual variance in this criterion. In the formula of AIC, the first term indicates model accuracy, and the second term indicates model complexity in terms of the number of variables. Hence AIC not only rewards prediction accuracy, but also imposes more penalty to larger number of model variables. One major benefit of this penalty is to discourage overfitting. For a set of models, the one with the lowest AIC value is considered as the preferred model. This criterion is particular suitable for comparing a set of nested models. For example, compare an $\text{AR}(m)$ model with an $\text{AR}(m+1)$ for a given set of data. One drawback of AIC is that it might not be consistent, since as the number of observations grows, the probability of selecting the correct model does not approach one.

The Bayesian information criterion (BIC), also known as the *Schwarz criterion*, is another well-known criterion to select a set of parametric models with different choices of explanatory variables [20]. BIC is actually

a variant of AIC in a form of

$$\text{BIC}(K) = \log(\text{MSE}) + \frac{\log(n)K}{n}. \quad (14)$$

The BIC also makes a trade-off between accuracy and complexity of a model. For a set of models, the one with the lowest value of BIC is the one to be preferred. It differs from AIC in that the penalty coefficient of K becomes $\log(n)/n$ instead of $2/n$. The BIC generally penalizes free variables more strongly than does the AIC. In addition, Hannan and Quinn [21] also proposed an alternative to AIC and BIC called *Hannan–Quinn criterion* (HQC), which is given by

$$\text{HQC}(K) = \log(\text{MSE}) + \frac{2K \ln \ln(n)}{n}. \quad (15)$$

Similar to AIC and BIC, the model with the lower value of HQC is preferred. It has been shown that consistency can be obtained by the BIC [20] or HQC [21,22].

Cross-Validation

Cross-validation is also a commonly used technique to compare different predictive models in practice [23]. Given a set of data, the basic idea of cross-validation is to partition the data set into training and validating subsets, and estimate the predictive accuracy on the validating data by the model obtained from the training data set. The MSE, RMSE, MAE, and MAPE can be used to measure the expected level of fit of a predictive model. If a model fits the training data set very well but does not fit the validation data, it is called *overfitting*. A good predictive model is supposed to generate consistent results in both training and validating data sets. To reduce the variability of the model evaluation, multiple trials of cross-validation are usually performed with respect to different partitions to the original data set. The evaluation result of a predictive model is the averaged result over these trials. There are several different approaches to perform multiple steps of cross-validation.

- *Repeated Random Partition Validation.* This method simply divides the dataset into two subsets randomly each time

and repeats the same procedure a number of times. One problem of this method is that the validation subsets may overlap and some observations may never be selected in the validation subsets.

- *K-Fold Cross-Validation.* The dataset is partitioned randomly into K subsets. Then the cross-validation is repeated K times. Each time one of the subsets is reserved as the validation data, and the remaining $K - 1$ subsets are the training data sets. The validation result is the average over K results. This approach guarantees that each observation can be used for validation exactly once.
- *Leave-One-Out Cross-Validation.* This method is actually a special case of K -fold cross-validation, when the number of observations in each subset is one. In other words, only one observation is reserved for validation and the remaining observations are used for training. The procedure is repeated until each observation has been used once for validation.

Stepwise Model Selection

Stepwise model selection approaches are very useful for automatically selecting a set of nested predictive models, for which there are a large number of potential predictive variables [24]. The selection procedure is generally grounded in some statistical tests and usually takes the form of the partial F -test. Other measures can also be used, such as t -tests, R^2 , AIC, and BIC. Since the basic procedures are similar, only the case of the partial F -test is discussed here. To compare two nested models with different number of predictor variables, the partial F -test can be generally formulated as follows

$$F = \frac{\text{Extra sum of squares/Extra model df}}{\text{SSR of the larger model / Residual df of the larger model}}, \quad (16)$$

where SSR denotes the sum of squared residual. The key idea of this test is to check

if the “extra” predictors of the larger model explain significantly more of the variability compared to the variability that is explained by the predictors that are already in the smaller model. On the basis of the partial F -test, three approaches are commonly used for model selection:

- *Forward Selection.* This starts with the smallest number of possible predictors and adds predictors one by one until a stop criterion is satisfied or the largest model is reached. The current model satisfying the stop criterion is selected. In particular, suppose that the current model has P variables, and we want to test if one of the model with $P + 1$ variables is more preferred. If for all models of $P + 1$ variables, it satisfies

$$F = \frac{\text{SSR}(P + 1) - \text{SSR}(P)}{\text{SSR}(P + 1)/n - P - 2} < F_c, \quad (17)$$

where F_c is the critical value of the F -statistic for a chosen level of significance, n is the total number of observations to fit the models, and $n - P - 2$ is the residual df of a model with $P + 1$ variables. Then stop the process and the current model with P variables is preferred. Otherwise, select one preferred model of $P + 1$ variables, and repeat the test for all models with $P + 2$ variables, and so forth.

- *Backward Selection.* This starts with the largest number of possible predictors and removes predictors one by one. Suppose the current model has P variables. Then at each step, it compares all the possible models of $P - 1$ variables with the current model. The process is stopped and the current model is selected, if the F -test statistic of each smaller models is not significant. That is, for all smaller models, it has

$$F = \frac{\text{SSR}(P - 1) - \text{SSR}(P)}{\text{SSR}(P)/n - P - 1} < F_c. \quad (18)$$

- *Stepwise Selection.* This is a modified version of forward selection, which allows the elimination of predictors that become statistically insignificant in the model. At each step of the process, the p values of all predictors are computed. If the largest of these p values is greater than a critical value, then the corresponding predictor is eliminated. Other steps are all the same as with those of forward selection.

Residual Diagnostics

Residuals represent the portion of the validation data not explained by the model. The graphical residual analysis is commonly used complementarily with quantitative techniques. A typical residual diagnostics includes plots of residuals versus the predicted values, versus other predictors, and versus time; residual autocorrelation plots; residual histograms; and normal probability plots. In general, the residual analysis can be used to test the following:

- Whiteness test: a good predictive model should have uncorrelated residuals.
- Independence test: a good model should have residuals uncorrelated with past inputs.
- If there are some extreme influential observations. Identifying and deleting outliers from the training dataset may significantly improve the quality of a model.
- If the residuals exhibit systematic patterns and bias. The residuals of a good model should be approximately dispersed around zero evenly. If systematic patterns are found, the most probable reason is that one or several relevant predictors are missing. A forecast is biased if residuals in one direction are significantly larger than those of the other direction.

Qualitative Considerations

Any forecasting model will be inevitably used by people. Thus qualitative considerations can also be a very important factor in evaluating various forecasting models. The qualitative criteria include intuitive reasonableness of the model, simplicity of the model, ease of use in practice, ease of explanation and understanding, and quality of forecast plots and demonstrations. Moreover, if two models have similar predictive performance, the simpler model is usually preferred. One famous principle is called *Occam's razor*, which suggests that when competing theoretical models are equal in other respects, it is recommended to select the model that introduces the fewest assumptions while still sufficiently answering the question. For more details of Occam's razor, refer to Soklakov [25].

CONCLUSION

This article reviews the most commonly used forecasting methods including both qualitative and quantitative methods. Quantitative methods are based on rigorous mathematical formulations, while qualitative methods are based on subjective judgment. Different forecasting methods have distinct different characteristics and applicable areas. To select appropriate forecasting models, one generally needs to consider the following important factors before model construction:

- What is the objective of forecasting, short-run or long-run?
- What types of data or explanatory variables are available?
- How much accuracy is required?
- Whether to make a trade-off between the costs and gains of developing a forecasting model.

Once a number of model candidates are selected, one should use one or more evaluation criteria to select the best forecasting model. Although quantitative models are mainstream approaches in forecasting, they are often used in combination with

qualitative techniques involving human judgment. Broadly speaking, quantitative methods usually provide tools for decision support, while quantitative and qualitative techniques together are used for decision making in most of the real-world cases. Therefore, a comprehensive assessment of forecasting performance usually consists of both quantitative and qualitative analyses to enhance the forecasting rationality in practice. A very good review of combining qualitative methods with quantitative techniques can be found in Webby and O'Connor [26] and Armstrong and Collopy [27].

REFERENCES

- Shukla J. Predictability in the midst of chaos: a scientific basis for climate forecasting. *Science* 1998;282(5389):728–731.
- Evans MK. Practical business forecasting. Oxford (UK): Blackwell Publishers; 2002.
- Fildes R, Lusk EJ. The choice of a forecasting model. *Omega* 1984;12(5):427–435.
- Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. *Int J Forecast* 1999;15(4):353–375.
- Satorius LC, Mohn NC. Sales forecasting models: a diagnostic approach. Research monograph 69. Atlanta (GA): College of Business Administration, Georgia State University; 1976.
- Fahey L, Randall RM. Learning from the future: competitive foresight scenarios. New York: John Wiley & Sons, Inc.; 1997.
- Dalrymple DJ. Sales forecasting practices: results from a United States survey. *Int J Forecast* 1987;3(3–4):379–391.
- Kahn KB. An exploratory investigation of new product forecasting practices. *J Prod Innov Manage* 2002;19:379–391.
- Holden K, Peel DA, Thompson JL. Introduction to forecasting methods. Economic forecasting: an introduction. Cambridge (UK): Cambridge University Press; 1991.
- Gardner ES. Exponential smoothing: the state of the art. *J Forecast* 1985;4(1):1–28.
- Gardner ES. Exponential smoothing: the state of the art—part ii. *Int J Forecast* 2006;22(4):637–666.
- Box GEP, Jenkins GM, Reinsel GC. Time series analysis, forecasting and control. Englewood Cliffs (NJ): Prentice Hall; 1994.
- Harvey AC. Forecasting, structural time series models and the Kalman filter. Cambridge (MA): Cambridge University Press; 1991.
- Koopmans LH. The spectral analysis of time series. New York: Academic Press; 1995.
- Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley & Sons, Inc.; 1989.
- Pindyck R, Rubinfeld D. Econometric models and economic forecasts. Boston (MA): McGraw-Hill/Irwin; 1997.
- Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323(9):533–536.
- Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 1998;14(1):35–62.
- Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;19(6):716–723.
- Schwarz GE. Estimating the dimension of a model. *Ann Stat* 1978;6(2):461–464.
- Hannan EJ, Quinn BG. The determination of the order of an autoregression. *J R Stat Soc* 1979;B(41):190–195.
- Hannan EJ. The estimation of the order of an ARMA process. *Ann Stat* 1980;8(5):1071–1081.
- Geisser S. Predictive inference. New York: Chapman and Hall; 1993.
- Efroymson MA. Multiple regression analysis. Mathematical methods for digital computers. New York: Wiley; 1960. pp. 191–203.
- Soklakov AN. Occam's razor as a formal basis for a physical theory. *Found Phys Lett* 2002;15(2):107–135.
- Webby R, O'Connor M. Judgemental and statistical time series forecasting: a review of the literature. *Int J Forecast* 1996;12(1):91–118.
- Armstrong JS, Collopy F. Integration of statistical methods and judgment for time series forecasting: principles from empirical research. Forecasting with Judgment. Chichester (UK): John Wiley & Sons; 1998. pp. 269–293.

EVALUATIONS OF SINGLE- AND REPEATED-PLAY GAMBLER

DOUGLAS H. WEDELL
Department of Psychology,
University of South Carolina,
Columbia, South Carolina

We encounter single instantiations of risky prospects on a regular basis, such as when we consider which car to purchase, which investment to make, or which health treatment option to choose. In each case, there is a single, unique outcome that we will experience. Regardless of whether the probability that the car we purchase will be a lemon is 1 in 10 or 5 in 10, we will experience only one of these outcomes. So even if we pick the safe outcome (1 in 10) we may still get burned. At other times, we may be considering a set of repeated instantiations tied to our choice. For example, instead of purchasing a single car, one may purchase a fleet of 100 cars for a company. In such cases, it is likely that a distribution of outcomes will be observed so that 1 in 10 would translate into roughly 10 of the 100 cars in the fleet being a lemon. If we consider health decisions, the same type of distinction can be made. We may contrast the idea of choosing between risky procedures for a single person (often ourselves), or developing a choice policy applied to a population of individuals, as a health-care administrator might do.

What difference does it make to evaluate a single instantiation versus multiple instantiations of a risky prospect? Should it make a difference? This article reviews changes in decision-making behavior, which occur when evaluating single versus repeated-play gambles, beginning with a discussion of normative issues. That discussion is followed by a review of relevant behavioral studies and then a summary of theoretical explanations for these evaluations and choice patterns.

NORMATIVE ISSUES

When von Neuman and Morgenstern [1] provided a normative basis for selecting the alternative with the highest expected utility (EU), they essentially argued that a long-run perspective be applied to each unique choice set one encounters. Because the concept of expectation is based on an analysis of aggregated outcomes, EU theory naturally incorporates a long-run strategy. But, people do not always find it easy to adopt a long-run perspective when considering unique instantiations of a choice set. Instead, choices made in the short run for unique single event options and in the long run for repeated or multiple instantiations of risky prospects often differ considerably.

The Nobel prize-winning economist, Paul Samuelson, examined the distinction between short- and long-run perspectives in a brief article he wrote revolving around the choice behavior of one of his colleagues. He described how his colleague refused to risk \$100 for a 50/50 chance to win \$200 [2]. Although this gamble has quite a high expected value ($EV = 0.50[-\$100] + 0.50[\$200] = \$50$), such a refusal is not problematic for EU theory, because a concave utility function could be constructed in which its EU would be less than the status quo. However, this choice became problematic when the colleague insisted he would be willing to place the bet if he were guaranteed a hundred such opportunities. Samuelson described the colleague's unwillingness to play the bet once but insistence on playing it 100 times as reflecting two different decision rules, one focused on the short run and the other on the long run, and he considered this inconsistency as irrational and a violation of EU theory.

Is it a violation of normative principles to choose differently for single and repeated risky prospects? Samuelson [2] sketched out a proof that was later fleshed out more

fully by Tversky and Bar-Hillel [3] describing when such behavior was inconsistent with EU theory. Samuelson acknowledges that refusing the single play but endorsing multiple plays can be accommodated by EU theory under some specific utility functions. However, he states that if one would refuse the single bet under all wealth levels described by the possible range of outcomes (in this case, $100[-\$100] = -\$10,000$ to $100[\$200] = \$20,000$), then it is inconsistent to accept the repeated version. Tversky and Bar-Hillel presented a more general proof that does not depend on using EU theory as the normative standard but only depends on accepting the normative principles of transitivity and dominance. Both these proofs consider the choice of the aggregate as the same as a series of single-play choices within the range of wealth levels. However, Aloysius [4] pointed out that equating the multiple-play choice with a series of single-play choices is not a valid assumption, and he demonstrated how EU theory can be consistent with the choice pattern of Samuelson's colleague under the conditions outlined in the proofs when consideration of the aggregate is based on current one's wealth state rather than wealth states at each sequential choice point. Lopes [5,6] has also argued for the rationality of this behavior on the basis of considerations outside of the realm of EU theory. Thus, there is still a good deal of controversy concerning the normative status of switching choices between single plays of a risky prospect and multiple or repeated plays of that prospect.

BEHAVIORAL FINDINGS

Before describing behavioral results, it is useful to clarify the distributional representations of the single- and multiple-play situations, the latter being a binomial distribution of outcomes. Figure 1 presents the distributions that correspond to (i) a single play of Samuelson's bet, (ii) 10 plays of this bet, and (iii) 10 plays of the bet with payoffs reduced by a factor of 10, equating EVs for single play and repeated plays. There are several distinctions that are immediately apparent when comparing a single play to

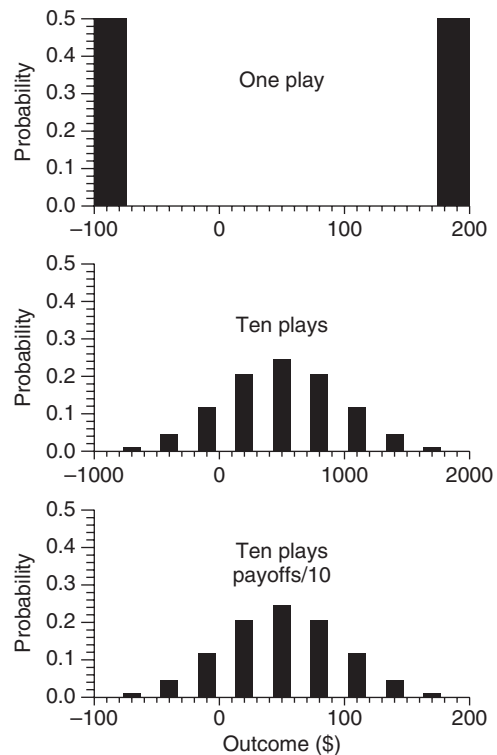


Figure 1. Outcome distributions for (a) one play of Samuelson's bet (50% chance to lose \$100 and 50% chance to win \$200), (b) 10 plays of the gamble, and (c) 10 plays of the gamble with outcomes reduced by a factor of 10.

multiple plays. As pointed out by Lopes [5,6], the EV of \$50 for Samuelson's bet does not correspond to any realized outcome in the single-play situation—one walks away either \$100 poorer or \$200 richer. In contrast, the EV is the modal value of the distribution for the 10-play situation (in this case, there is a probability of 0.2461 of walking away with the EV). Thus, with a single-play, EV is an abstract conception that does not match up with actual outcomes, whereas with repeated plays it is a tangible and typical result.

A second distinct feature of these distributions is that the probability of losing is quite different. With single plays, the probability of losing is simply $p = 0.50$, so that half of the time one will walk away a loser. With 10 plays, one needs to lose more than twice as often as win in order to end

up an overall loser, so that the probability of losing is only $p = 0.17$. If the bet were played 100 times, the probability of losing would reduce to less than 1 in 2000, radically different from the single-play situation. Of course, Samuelson [2] pointed out that the magnitude of the risk (i.e., variance of the distribution) increases with repeated plays. Thus, in the 10-play situation, one has a slight probability ($p = 0.00098$) to suffer a 10 times greater loss ($-\$1000$). Because of the differences in the range of losses and the EVs for single and multiple plays of a bet, researchers have often included comparison conditions in which EV and range of losses are equated for the two conditions, so that the influence of these factors can be isolated. Figure 1c shows the distribution of outcomes when payoffs are reduced by a factor equal to the number of plays. Note that Samuelson's objection does not apply to this situation where risk is clearly reduced, as evidenced by the much lower variance outcomes. This type of diversification strategy is a commonly used approach for reducing risk in a portfolio of assets.

Behavior for Samuelson's Bet

Several researchers have asked people to respond to bets that parallel the one Samuelson offered his colleague in order to consider the prevalence of this pattern of responding and evaluate rationales. In a study of college students, Wedell and Böckenholt [7] found that 79% of students refused a single play of the bet, but 44% of these were willing to play the bet 100 times. Benartzi and Thaler [8] surveyed undergraduates, graduate students, and coffee shop visitors with a problem that reduced the outcomes substantially from Samuelson's gamble and found only 34% refused the single-play version. Of these, roughly 44% were willing to play the bet 100 times, a similar conditional proportion as found by Wedell and Böckenholt. Keren [9] asked the 79% of participants, who refused to play Samuelson's bet one time, how many times it would have to be repeated before they were willing to play it. While 23% were willing to play it five times, 55% were willing to play it 100 times (a similar percentage as described above). Finally, Redelmeier and

Tversky [10] provided undergraduates with a bet similar to Samuelson's but with inflated amounts to win and lose. They found significantly fewer choices to play the bet once (43%) than five times (63%). In another sample, they asked participants to compare one play of the bet with five plays, five plays with six plays, and six plays with 12 plays. In each case, participants preferred the option of playing the bet more times. However, when asked whether after playing the gamble five times one would want to play it again, most participants refused, emphasizing the psychological difference between considering the set of gambles versus gamble one at a time. The results of these and other studies [11–18] indicate that the behavior of Samuelson's colleague to refuse a single play but endorse repeated plays of a gamble can be found in a substantial proportion of participants, although it may not be a majority response. This tendency may be greater for large outcome gambles, perhaps because of a higher refusal rate of the single gamble.

Reducing Decision Anomalies

It has long been known that human decision behavior often violates predictions of EU theory. To what extent is anomalous behavior found for single prospects reduced when participants evaluate repeated versions of those prospects? Keren and Wagenaar [19] investigated this question with respect to certainty and possibility effects, which Kahneman and Tversky [20] described as separable contributors to the Allais Paradox. In the certainty effect, a risky prospect A is compared to a certain (or near certain) prospect B, and then a control pair, A' and B', is created by reducing the probability of winning by a constant ratio. For example, A might be (0.50 to win \$250 else \$0) and B might be (1.00 to win \$100). Most people prefer the certain option, B, which means that the following inequality holds: $U(\$100)/U(\$250) < 0.50/1.00$. However, reducing each probability by a factor of 5 results in options A' (0.10 to win \$250 else \$0) and B' (0.20 to win \$100, else \$0). A majority of people prefer A' to B', which means that the following inequality holds: $U(\$100)/U(\$250) > 0.10/0.20$, a contradiction

of the choice of A over B. In the possibility effect, option probabilities are reduced by a constant ratio to create very low probability pairs that again create contradictory results. For example, one may prefer C (0.90 to win \$5000 else \$0) to D (0.45 to win \$12,000 else \$0), but prefer D' (0.01 to win \$12,000 else \$0) to C' (0.02 to win \$5000 else \$0).

Keren and Wagenaar [19] conducted two experiments in which participants chose between bets like those described above or the same types of bets represented as being played 10 or 100 times. For the unique, single-play gambles, they replicated both certainty and possibility effects. However, neither of these effects was significant for the repeated gambles. They concluded that different choice processes are invoked when considering the short-run unique gamble versus the long-run repeated gambles. In a follow-up, Keren [9] demonstrated a similar effect with gambles being played just once or five times. In these combined experiments, participants appeared to switch to a strategy of choosing the gamble with the higher EV when faced with multiple choices, consistent with results from Montgomery and Adelbratt [17] in which participants endorsed choosing on the basis of EV for multiple-play gambles but not for single-play gambles. The tendency for certainty and possibility effects to be reduced in the repeated-play situation implies that the choice process for repeated gambles is more consistent with EU theory than the process used for choosing single gambles.

Another violation of EU theory revolves around procedural invariance. Different preference elicitation methods can be used, such as choice, attractiveness ratings, minimum selling price, maximum buying price, and value matching. If the same utility and decision weighting processes underlie these different elicitation methods, then the preference ordering of risky prospects should be invariant across them. However, it has long been documented that preference order can be reversed by use of different methods for expressing preference [21]. In the classic response-based preference reversal, the same individual chooses between risky prospects and sets minimum selling prices for each. One prospect is referred to as a P-bet because

it has a relatively high probability to win a relatively small amount, such as a 0.90 chance to win \$2 else \$0. The other prospect is referred to as a \$-bet because one can win a relatively large amount but with a correspondingly smaller probability, such as a 0.10 chance to win \$20 else \$0. The typical finding is that the P-bet is preferred in choice, but the \$-bet is valued higher. Thus, people tend to choose the bet to which they assign a lower value.

Wedell and Böckenholt [22] examined whether the preference reversal phenomenon would obtain when gambles were represented as being played repeatedly. In both experiments, they found a significant reduction in preference reversals with an increase in the number of plays from one time to either 10 times or 100 times, although preference reversals were still significant for repeated-play conditions. Wedell and Böckenholt reported that the majority of subjects made fewer reversals in the multiple-play condition than in the single-play condition. When prompted to explain why they changed their pricing and choice behavior, the modal response was that the perceived chances of winning changed with increase in plays. Finally, note that Haisley *et al.* [23] demonstrated the tendency for choice to be more in line with EV for people purchasing actual lottery tickets (singly vs in groups of five), with the repeated-play version in this case leading to significantly fewer purchases.

A third violation of EU theory arises out of considerations of sources of uncertainty as described in the Ellsberg paradox [24]. Ellsberg found that when presented with a series of choices, the majority of people opt for the gamble with a known distribution of outcomes rather than an unknown distribution, the former representing uncertainty under risk and the latter uncertainty under ambiguity. This ambiguity aversion tendency was problematic, in that it led the majority to endorse choices that violated the independence axiom in EU theory. Liu and Colman [25] investigated the possibility that ambiguity aversion is reduced in the repeated-play case, and thus the potential violation of the independence axiom is likewise reduced. In two experiments in

which participants choose between risky and ambiguous options, they observed a strong reduction in ambiguity aversion for gambles represented as being played 100 times rather than once. These results then imply that decision makers tend to treat ambiguity and risk in a more equivalent manner when considering repeated-play gambles.

Medical Decisions

Redelmeier and Tversky [10,26] examined whether decisions of physicians would differ when considering a single instance versus aggregated instances. One study that compared aggregation across patients found that physicians were more willing to recommend a blood test when considering a single patient than a group of like patients (30% vs 17%). In another study that compared aggregation across time, physicians were more willing to recommend a risky medication when the term of the illness was expected to last many days than when it was to last a single day (42% vs 15%), even though the chronic nature of the disease made it difficult to justify treating the single instance differently from the repeated instances. However, DeKay and Kim [11] reviewed several studies that failed to replicate results of Redelmeier and Tversky [26]. They further suggested that choice differences for single and repeated situations depend on the perceived fungibility of the outcomes. Highly fungible monetary outcomes show large differences of the manipulation, but outcomes of low perceived fungibility, such as comparison across individuals, produce little effect of the single versus repeated perspective. In summary, the results in this area are mixed, with no simple pattern being observed across studies.

Investment Decisions

Investment decisions have a natural parallel to the structure of gambles, although it is less clear about the degree to which dependencies exist among investment options. Benartzi and Thaler [8] explored short- and long-run framing in a series of experiments that used hypothetical gambles and real retirement investments. One key finding from these studies was that the format in

which repeated gambles was presented was an important factor. While choices were more sensitive to EV for repeated than for single gambles, this difference was even more marked when tabular or graphic displays of the distribution of outcomes was shown (such as the displays in Fig. 1). This finding is not surprising when considering that limitations on cognitive resources would make it difficult for the average individual to accurately generate the probability distribution resulting from aggregation. (Consider an attempt at generating the outcome shown in Fig. 1b.) Benartzi and Thaler demonstrated in two additional studies that employees who were presented with the aggregate distribution over multiple years invested a much greater percentage of retirement funds in stock options than those given the one-year return rate (a 82% allocation to stocks vs a 41% allocation). Thus, when left to extrapolate consequences of repeated years of investment, people appear to be much more risk averse than when presented the resulting distribution of outcomes. Indeed, a study by Klos *et al.* [13] that examined investors' ability to judge various dimensions of the outcome distributions for investments and gambles found a decided advantage when the outcome distribution resulting from repeated instantiations was graphically represented.

While a number of studies have demonstrated greater acceptance of repeated or long-run investment strategies [8,12,18], Langer and Weber [14] demonstrated that the opposite effect could occur, depending on the structure of the investment. They replicated earlier work that utilized fairly large probabilities of moderate losses and then showed how gambles that included a low probability to lose a large amount were more likely to be endorsed for single plays than for repeated plays. They also demonstrated how both results were in line with prospect theory predictions.

THEORETICAL EXPLANATIONS

The clear differences in how many people respond to single and repeated instances of risky prospects can be explained by a

variety of theoretical mechanisms. The most prominent theoretical explanation has been derived from prospect theory and hinges around loss aversion. Kahneman and Lovallo [27] referred to this result as *myopic loss aversion* and showed how repeated prospects represented in their aggregated form largely eliminate loss aversion tendencies. This view of the phenomena implies that single and repeated gambles are essentially processed in the same way, although the latter requires editing into aggregate form. Others have referred to this as reflecting *narrow bracketing of outcomes* in the short run and *broad bracketing of outcomes* in the long run [18,23]. The take home message from this approach is not that there is a change in the valuation process, but rather that there is a change in the representation upon which valuation operates. For example, if we apply a simple loss aversion weighting scheme to Samuelson's bet in which losses are multiplied by a factor of 3, then one play of the bet is valued at $(0.5)(-100)(3) + (0.5)(200) = -50$ and so one would refuse to play it. However, applying this formulation to two plays of the bet results in a value of $0.25(-200)(3) + (0.50)(100) + (0.25)(400) = 0$, or indifference. According to this loss-averse weighting scheme, one would refuse a single play but choose three or more plays of the gamble. The key then is to aggregate across plays, or perhaps to see different investments as part of a broader portfolio across which one aggregates.

Other theoretical explanations focus on more distinct process differences between the situations, emphasizing shifting from qualitative to quantitative processing, a focus on aspiration levels, or on higher order features of distribution of outcomes related to variability, skewing, and rank [5–7,22]. Lopes [5,6] pointed out that EU theory's emphasis on one moment of the outcome distribution, the mean, may not lead to the wisest decisions. Other moments related to variability and skewing may be important, especially when some outcomes are catastrophic. It is possible that individuals use a qualitative comparison of outcomes to an aspiration level before proceeding to a more systematic valuation of the prospect [22]. If this is the case, then the

evaluation process would differ markedly for single- and multiple-play prospects because the multiple-play prospect would be more likely to exceed the aspiration level and proceed to a more integrative assessment. Consistent with this view are verbal reports justifying choice patterns [22] as well as evidence that risk assessment of single plays follows an additive integration pattern but that of multiple plays follows a multiplicative integration pattern [28]. Regardless of the mechanism, the extant literature is consistent with the view that people's evaluations of repeated risky prospects tend to conform more closely to normative standards, such as that described by EU theory, than their evaluations of risky prospects considered one at a time.

REFERENCES

1. Von Neuman J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1947.
2. Samuelson PA. Risk and uncertainty: a fallacy of large numbers. *Scientia* 1963;98:108–113.
3. Tversky A, Bar-Hillel M. Risk: the long and the short. *J Exp Psychol: Learn Mem Cogn* 1983;9:713–717. Aloysius JA. Decision making in the short and long run: repeated gambles and rationality. *Br J Math Stat Psychol* 2007;60:61–69.
4. Aloysius JA. Decision making in the short and long run: repeated gambles and rationality. *Br J Math Stat Psychol* 2007;60:61–69.
5. Lopes LL. Decision making in the short run. *J Exp Psychol: Hum Learn Mem* 1981;7:377–385.
6. Lopes LL. When time is of the essence: averaging, aspiration, and the short run. *Organ Behav Hum Decis Process* 1996;65:179–189.
7. Wedell DH, Böckenholt U. Contemplating single versus multiple encounters of a risky prospect. *Am J Psychol* 1994;107:499–518.
8. Benartzi S, Thaler RH. Risk aversion or myopia? Choices in repeated gambles and retirement investments. *Manage Sci* 1999;45:364–381.
9. Keren G. Additional tests of utility theory under unique and repeated condition. *J Behav Decis Mak* 1991;4:297–304.
10. Redelmeier DA, Tversky A. On the framing of multiple prospects. *Psychol Sci* 1992;3:191–193.

11. DeKay ML, Kim TG. When things don't add up: the role of perceived fungibility in repeated play decisions. *Psychol Sci* 2005;16:667–672.
12. Gneezy U, Potters J. An experiment on risk taking and evaluation periods. *Q J Econ* 1997; 112:631–645.
13. Klos A, Weber EU, Weber M. Investment decisions and time horizon: risk perception and risk behavior in repeated gambles. *Manage Sci* 2005;51:1777–1790.
14. Langer T, Weber M. Prospect theory, mental accounting and differences in aggregated and segregated evaluation of lottery portfolios. *Manage Sci* 2001;47:716–733.
15. Langer T, Weber M. Myopic prospect theory vs. myopic loss aversion: how general is the phenomenon?. *J Econ Behav Organ* 2005;56: 25–38.
16. Li S. The role of expected value illustrated in decision-making under risk: single-play vs. multiple-play. *J Risk Res* 2003;6:113–124.
17. Montgomery H, Adelbratt T. Gambling decisions and information about expected value. *Organ Behav Hum Perform* 1982;29:39–57.
18. Thaler RH, Tversky A, Kahneman D, *et al.* The effect of myopia and loss aversion on risk taking: an experimental test. *Q J Econ* 1997; 112:647–661.
19. Keren G, Wagenaar WA. Violation of utility theory in unique and repeated gambles. *J Exp Psychol: Learn Mem Cogn* 1987;13:387–391.
20. Kahneman DD, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47:263–291.
21. Lichtenstein S, Slovic P. Reversal of preferences between bids and choices in gambling decisions. *J Exp Psychol* 1971;89:46–55.
22. Wedell DH, Böckenholt U. Moderation of preference reversals in the long run. *J Exp Psychol Hum Percept Perform* 1990;16:429–438.
23. Haisley E, Mostafa R, Loewenstein G. Myopic risk-seeking: the impact of narrow decision bracketing on lottery play. *J Risk Uncertain* 2008;37:57–75.
24. Ellsberg D. Risk ambiguity and the Savage axioms. *Q J Econ* 1961;75:643–649.
25. Liu HH, Colman AM. Ambiguity aversion in the long run: repeated decisions under risk and uncertainty. *J Econ Psychol* 2009;30: 277–284.
26. Redelmeier DA, Tversky A. The discrepancy between medical decisions for individual patients and for groups. *New Engl J Med* 1990;322:1162–1164.
27. Kahneman DD, Lovallo D. Timid choices and bold forecasts: a cognitive perspective on risk taking. *Manage Sci* 1993;39:17–31.
28. Joag SG, Mowen JC, Gentry JW. Risk perception in a simulated industrial purchasing task: the effects of single versus multi-play decisions. *J Behav Decis Mak* 1990;3:91–108.

EVOLUTIONARY ALGORITHMS

ZBIGNIEW MICHAŁEWICZ
School of Computer Science,
University of Adelaide,
Adelaide, Australia

MARC SCHOENAUER
Project-Team TAO, INRIA
Saclay–Île-de-France,
Orsay, France

INTRODUCTION

Evolutionary computation (EC) techniques are stochastic algorithms whose search methods model some natural phenomena: genetic inheritance and Darwinian strife for survival. The idea behind evolutionary algorithms (EAs) is to do what nature does. Let us take rabbits as an example: at any given time there is a population of rabbits. Some of them are faster and smarter than other rabbits. These faster, smarter rabbits are less likely to be eaten by foxes, and therefore more of them survive to do what rabbits do best: make more rabbits. Of course, some of the slower, dumber rabbits will survive just because they are lucky. This surviving population of rabbits starts breeding. The breeding results in a good mixture of rabbit genetic material: some slow rabbits breed with fast rabbits, some fast with fast, some smart rabbits with dumb rabbits, and so on. And on the top of that, nature throws in a “wild hare” every once in a while by mutating some of the rabbit genetic material. The resulting baby rabbits will (on average) be faster and smarter than these in the original population because more faster, smarter parents survived the foxes. (It is a good thing that the foxes are undergoing similar process—otherwise the rabbits might become too fast and smart for the foxes to catch any of them). So the metaphor underlying EAs is that of natural evolution. In evolution, the problem each species faces

is one of searching for beneficial adaptations to a complicated and changing environment. The “knowledge” that each species has gained is embodied in the makeup of the chromosomes of its members. From the point of view of optimization, EC is a powerful stochastic zeroth order method (i.e., requiring only values of the function to optimize) that can find the global optimum of very rough functions. This allows EC to tackle optimization problems for which standard optimization methods (e.g., gradient-based algorithms requiring the existence and computation of derivatives) are not applicable. Moreover, most traditional methods are local in scope; thus, they identify only the local optimum closest to their starting point.

AN ALGORITHM

For the sake of clarity, we shall try to introduce a general framework, accounting as much as possible for most existing EAs.

Let the search space be a metric space E , and let F be a function $E \rightarrow \mathbb{R}$ called the *objective* function. The task of evolutionary optimization is to find the maximum of F on E (the case of minimization is easily handled by considering $-F$).

A *population* of size $P \in \mathbb{N}$ is a set of P *individuals* (points of E) not necessarily distinct. This population is generally initialized randomly (at time $t = 0$) and uniformly on E . Then the *fitness* of each individual is computed (on the basis of the values of the objective function); a fitness value is represented as a positive real number—the higher the number, the better the individual. The population then undergoes a succession of *generations*; the process is illustrated in Fig. 1.

Several aspects of the evolutionary procedure require additional comments:

- *Statistics and Stopping Criterion.* The simplest stopping criterion is based on the generation counter t (or on the number of function evaluations). However,


```

procedure evolutionary algorithm
begin
   $t \leftarrow 0$ 
  initialize population
  evaluate population
  while (not termination-condition) do
    begin
       $t \leftarrow t + 1$ 
      select individuals for reproduction
      apply variation operators
      evaluate newborn offspring
      select individuals for survival
    end
  end

```

Figure 1. The structure of an evolutionary algorithm.

it is possible to use more complex stopping criteria, which depend either on the evolution of the best fitness in the population along generations (i.e., measurements of the gradient of the gains over some number of generations), or on some measure of the diversity of the population.

- *Parental Selection.* Choice of some individuals that will generate offspring. Numerous selection processes can be used, either deterministic or stochastic. All are based on the fitness of the individuals. Depending on the selection scheme used, some individuals can be selected more than once. At that point, selected individuals give birth to copies of themselves (clones).
- *Application of Variation Operators.* To each one of these copies, some operator(s) are applied, giving birth to one or more offspring. The choice among possible operators is stochastic, according to user-supplied probabilities. These operators are always stochastic operators; it is common to distinguish between *crossover* (or *recombination*) and *mutation* operators:
 - crossover operators are operators from E^k into E , that is, some parents exchange genetic material to build up one offspring. In most cases, crossover involves two parents ($k = 2$), though more parents can be used.

- mutation operators are stochastic operators from E into E .

- *Evaluation.* Computation of the fitnesses of all newborn offspring. As mentioned earlier, the fitness measure of an individual is directly related to its objective function value.
- *Survival Selection.* Choice of which individuals will be part of next generation. The choice can be made either from the offspring only (in which case all parents “die”) or from both the offspring and the parents. In either case, this survival selection procedure can be deterministic or stochastic.

Sometimes the variation operators are defined on the same space as the objective function (called *phenotype space* or *behavioral space*); in other cases, an intermediate space is introduced (called *genotype space* or *representation space*). The mapping from the phenotype space in the genotype space is termed *coding*. The inverse mapping from the genotype space in the phenotype space is termed *decoding*. Genotypes undergo variation operators, and their fitness is evaluated on the corresponding phenotype. The properties of the coding mappings can greatly modify the global behavior of the EA. For more (implementational) details related to the structure outlined in Fig. 1 and further discussion on the above aspects of the evolutionary procedure, see Ref. 1.

AN EXAMPLE

In this section we provide an example (based on example given in Chapter 2 of Ref. 2) of action-steps of a standard genetic algorithm (GA)—the best-known paradigm within EAs—for a numerical optimization problem. These action-steps illustrate the general structure of EAs given in Fig. 1. Here, without any loss of generality, we assume the maximization problem where the objective function f takes positive values on its domain.

So let us suppose we wish to maximize a function of k variables, $f(x_1, \dots, x_k) : \mathbb{R}^k \rightarrow \mathbb{R}$. Suppose further that each variable x_i can

take values from a domain $D_i = [a_i, b_i] \subseteq \mathbb{R}$ and $f(x_1, \dots, x_k) > 0$ for all $x_i \in D_i$. We wish to optimize the function f with some required precision: suppose six decimal places for the variables' values are desirable.

It is clear that to achieve such precision each domain D_i should be cut into $(b_i - a_i) \cdot 10^6$ equal size ranges. Let us denote by m_i the smallest integer such that $(b_i - a_i) \cdot 10^6 \leq 2^{m_i} - 1$. Then, a representation having each variable x_i coded as a binary string of length m_i clearly satisfies the precision requirement. Additionally, the following formula interprets each such string:

$$x_i = a_i + \text{decimal}(\text{string}_2) \cdot \frac{b_i - a_i}{2^{m_i} - 1},$$

where $\text{decimal}(\text{string}_2)$ represents the decimal value of that binary string.

Now, each individual (in GA terminology, individuals are often called *chromosomes*) is represented by a binary string of length $m = \sum_{i=1}^k m_i$; the first m_1 bits map into a value from the range $[a_1, b_1]$, the next group of m_2 bits map into a value from the range $[a_2, b_2]$, and so on; the last group of m_k bits map into a value from the range $[a_k, b_k]$. A chromosome represents a potential solution to the problem.

To initialize the population, we can simply set some P number of chromosomes randomly in a bitwise fashion. However, if we do have some knowledge about the distribution of potential optima, we may use such information in arranging the set of initial (potential) solutions.

At each generation (while loop, see Fig. 1) we evaluate each chromosome (using the function f on the decoded sequences of variables), select new population with respect to the probability distribution based on fitness values, and alter the chromosomes in the new population by mutation and crossover operators. After some number of generations, when no further improvement is observed, the best chromosome represents an (possibly the global) optimal solution. Often we stop the algorithm after a fixed number of iterations depending on speed and resource criteria. For the selection process (selection of a new population with respect to the probability distribution based

on fitness values), a *roulette wheel* with slots sized according to fitness is used here; such a roulette wheel is constructed as follows:

- Calculate the fitness value $\text{eval}(v_i)$ for each chromosome v_i ($i = 1, \dots, P$).
- Find the total fitness of the population

$$F = \sum_{i=1}^P \text{eval}(v_i).$$

- Calculate the probability of a selection p_i for each chromosome v_i ($i = 1, \dots, P$):

$$p_i = \text{eval}(v_i) / F.$$

- Calculate a cumulative probability q_i for each chromosome v_i ($i = 1, \dots, P$):

$$q_i = \sum_{j=1}^i p_j.$$

The selection process is based on spinning the roulette wheel P times; each time we select a single chromosome for a new population in the following way:

- Generate a random (float) number r from the range $[0, 1]$.
- If $r < q_1$ then select the first chromosome (v_1); otherwise select the i -th chromosome v_i ($2 \leq i \leq P$) such that $q_{i-1} < r \leq q_i$.

Obviously, some chromosomes would be selected more than once. This is in accordance with the Schema Theorem (see the section titled "Comparison"): the best chromosomes get more copies, the average stay even, and the worst die off.

Now we are ready to apply the recombination operator, crossover, to the individuals in the new population. As mentioned earlier, one of the parameters of a genetic system is probability of crossover p_c . This probability gives us the expected number $p_c \cdot P$ of chromosomes which undergo the crossover operation. We proceed in the following way:

For each chromosome in the (new) population:

- Generate a random (float) number r from the range $[0,1]$;
- If $r < p_c$, select given chromosome for crossover.

Now we mate selected chromosomes randomly: for each pair of coupled chromosomes we generate a random integer number pos from the range $[1, m - 1]$ (m is the total length—number of bits—in a chromosome). The number pos indicates the position of the crossing point. Two chromosomes

$$(b_1 b_2 \dots b_{pos} b_{pos+1} \dots b_m) \text{ and} \\ (c_1 c_2 \dots c_{pos} c_{pos+1} \dots c_m)$$

are replaced by a pair of their offspring:

$$(b_1 b_2 \dots b_{pos} c_{pos+1} \dots c_m) \text{ and} \\ (c_1 c_2 \dots c_{pos} b_{pos+1} \dots b_m).$$

The next operator, mutation, is performed on a bit-by-bit basis. Another parameter of the genetic system, probability of mutation p_m , gives us the expected number of mutated bits $p_m \cdot m \cdot P$. Every bit (in all chromosomes in the whole population) has an equal chance to undergo mutation, that is, change from 0 to 1 or vice versa. So we proceed in the following way.

For each chromosome in the current (i.e., after crossover) population and for each bit within the chromosome:

- Generate a random (float) number r from the range $[0, 1]$;
- If $r < p_m$, mutate the bit.

Following selection, crossover, and mutation, the new population is ready for its next evaluation. This evaluation is used to build the probability distribution (for the next selection process), that is, for a construction of a roulette wheel with slots sized according to current fitness values. The rest of the evolution is just cyclic repetition of the above steps.

It is relatively easy to keep track of the best individual in the evolution process. It is customary (in GA implementations) to store

“the best ever” individual at a separate location; in that way, the algorithm would report the best value found during the whole process (as opposed to the best value in the final population)—this approach illustrates so-called *elitist* strategy.

HISTORICAL PARADIGMS

This section introduces the four historical paradigms that build what is today known as *evolutionary computation*.

Genetic Algorithms

In the canonical GA [3,4], the genotype space is $\{0, 1\}^n$. Note that the phenotype space can be any space, as long as it can be coded into bitstring genotypes. The parental selection is proportional selection (the best-known being the roulette wheel selection as discussed in the previous section): P random choices are made in the whole population, each individual having a probability proportional to its fitness of being selected. The crossover operators replace a segment of bits in the first parent string by the corresponding segment of bits from the second parent, and the mutation operator randomly flips the bits of the parent according to a fixed user-supplied probability. In the survival selection phase, all P offsprings replace all parents (also known as *generational survival*). The best fitness in the population can thus decrease: the original GA strategy is not elitist.

However, it rapidly became clear that the genotype space can be almost any space, as long as some crossover and mutation operators are provided [2,5]. Moreover, proportional selection has been gradually replaced by comparison-based selections, from *ranking* (the selection is performed on the rank of the individuals rather than on their actual fitness), or *tournament* selection (one selects the best individual among a uniform choice of T individuals, T ranging from two to some small proportion of the population size)—see Ref. 6 for a discussion on these selection methods. Finally, most users use the elitist variant of survival selection, in which the best individual of generation t is included in generation $t + 1$, whenever the best fitness value in the population decreases.

Evolution Strategies

Evolution strategies (ESs) [7,8] have been designed as *parametric optimization* algorithms, that is, optimizing functions of floating-point variables. The original ES handles a “population” made of a single individual given as a real-valued vector. This individual undergoes a Gaussian mutation: addition of zero-mean Gaussian variable of standard deviation σ to each of the real variables. The fittest from the parent and the offspring becomes the parent of next generation. The critical feature is the choice of parameter σ : Originally, the so-called 1/5 thumb rule [i.e., when more than 1/5 mutation are successful (respectively unsuccessful), increase (respectively decrease) σ] [7].

ES then evolved into population-based algorithms [9], termed (μ, λ) – ES or $(\mu + \lambda)$ – ES: μ parents generate λ offspring. (There is no parental selection, i.e., every parent produces λ/μ offspring on average). Moreover, the survival selection is deterministic, that is, the best μ individuals become the parents of the next generation, chosen among the $\mu + \lambda$ parents plus offspring in the elitist $(\mu + \lambda)$ – ES scheme, or among the λ offspring in the nonelitist (μ, λ) – ES scheme (with $\lambda \geq \mu$).

The main operator remains mutation, generalized into multivariate Gaussian mutation, that is, defined by a full covariance matrix C and a scaling factor σ , also called the *step-size* mutation. In the 1990s, the powerful paradigm of *self-adaptive mutation* was the rule: C and σ were added to the description of the individuals, and undergo mutation as well. The recent trend is to adapt C and σ based on the history of the evolution, leading to the state-of-the-art algorithm of *covariance matrix adaptation—evolution strategy* (CMA-ES) [10,11]. CMA-ES is almost parameter-free: it uses a $(\lambda/2, \lambda)$ – ES survival selection, where the parameter λ is chosen as $4 + \lfloor 3 \log(n) \rfloor$ [10], and increased in case of highly multimodal functions [11].

Evolutionary Programming

Evolutionary programming (EP) is one of the oldest EAs, originally proposed to evolve finite state machines [12]. As in ESs, there is no parental selection, and every individual in

the population generates one offspring. Moreover, the only evolution operator is mutation. Survival selection is what is today called a $(P + P)$ – ES, that is, the best P individuals among parents and offspring become the parents of the next generation.

As in the field of GAs, further works rapidly generalized the approach to handle any search space, still emphasizing the use of mutation as the only operator [13]. Several variants were then introduced, from stochastic survival selection to self-adaptive mutations (similar to the ones from ESs, though discovered completely independently [14]).

Genetic Programming

Genetic programming (GP) as a method for evolving computer programs first appeared as an application of GAs to tree-like structures [15]. Original GP evolves tree structures representing LISP-like S-expressions. The strength of this representation is that a closed crossover operator can easily be defined: by swapping subtrees between two valid S-expressions, one always gets a valid S-expression. Koza’s original work used a steady-state GA evolutionary mechanism [16]: a parent is selected by tournament (of size two to seven typically), and generates an offspring by crossover only. The offspring is then put back in the population using a reverse tournament: T individuals are uniformly chosen, and the one with the worst fitness gets replaced by the newborn offspring.

Since then, most published work considered mutation also. Further, several variants of tree-based GP have been proposed (e.g., linear GP [17], cartesian GP [18], push GP [19]). At present, GP is viewed more generally as program evolution [20].

MODERN TRENDS: HYBRID METHODS

Many researchers further modified EAs by “adding” some problem-specific knowledge to the algorithm. Several papers have discussed initialization techniques, different representations, decoding techniques (mapping from genetic representations to phenotypic representations), and the use of heuristics for

variation operators. Davis [21] wrote (in the context of classical, binary GAs):

“It has seemed true to me for some time that we cannot handle most real-world problems with binary representations and an operator set consisting only of binary crossover and binary mutation. One reason for this is that nearly every real-world domain has associated domain knowledge that is of use when one is considering a transformation of a solution in the domain [...]. I believe that genetic algorithms are the appropriate algorithms to use in a great many real-world applications. I also believe that one should incorporate real-world knowledge in one’s algorithm by adding it to one’s decoder or by expanding one’s operator set.”

Such hybrid/nonstandard systems enjoy a significant popularity in the EC community. Very often these systems, extended by the problem-specific knowledge, outperform other classical evolutionary methods as well as other standard techniques. For example, a system, Genetic-2N [2] constructed for the nonlinear transportation problem used a matrix representation for its chromosomes, a problem-specific mutation (main operator, used with probability 0.4), and arithmetical crossover (background operator, used with probability 0.05). It is hard to classify this system: it is not really a GA, since it can run with a mutation operator only without any significant decrease of the quality of results. Moreover, all matrix entries are floating-point numbers. It is not an ES, since it did not use Gaussian mutation, nor did it encode any control parameters in its chromosomal structures. Clearly, it has nothing to do with GP and very little (just matrix representation) with EP approaches. It is just an EC technique aimed at a particular class of problems.

COMPARISON

Many papers have been written on the similarities and differences between these approaches [9,13,22]. Clearly, these similarities and differences can be discussed from different perspectives.

- *The Representation Issue.* Original GAs, ESs, and EP address only bitstrings,

real numbers, and finite state machines, respectively. However, recent tendencies indicate that this is not a major difference. Moreover, it is still far from trivial to select appropriate variation operators for the chosen representation and the objective function, also known as the *fitness landscape* [2,5].

- *The Usefulness of Crossover.* According to the Schema Theorem [3,4], GAs’ main strength comes from the crossover operator: better and better solutions are built by exchanging *building blocks* from partially good solutions previously built, in a bottom-up approach. The mutation operator is then considered as a background operator. On the other hand, the philosophy behind EP and ESs is that such building blocks might not exist, at least for most real-world problems. Also some experimental results contradict the building block hypothesis; for example, uniform crossover [23] is very disruptive of short schemata whereas one- and two-point crossover are more likely to conserve short schemata and combine their defining bits in offspring. This top-down view considers that selective pressure plus genotypic variability brought by mutation are sufficient. This discussion on significance of the crossover has been going on for a long time. However, even when crossover is experimentally demonstrated beneficial to evolution, it could be because it acts like a large mutation [24]. Yet another example of the duality between crossover and mutation comes from GP’s history: the original GP algorithm [15] used only crossover, with no mutation at all, with very large population size: the rationale is that all building blocks that are necessary to represent at least one sufficiently good solution are already present in the initial population, and only need to be recombined.
- *Mutation Operators.* Whereas the usefulness of crossover has been heavily discussed, that of mutation is acknowledged by all trends (with the historical exception of Koza’s work in GP—see the

section titled “Genetic Programming”). Indeed, mutation is the only operator that can reintroduce *diversity* in the population, as crossover is *de facto* limited by the current population. From a theoretical point of view, only mutation can guarantee the ergodicity of the stochastic process (i.e., that all points of the search space can be reached whatever the current population). However, the way mutation operators are applied differ from one paradigm to another, and should be closely linked to the types of selection that are used. Traditional GAs use a low mutation rate, that is, the average number of bits that are flipped remains small (though it can take large values, with very small probabilities). Because all offspring replace all parents without competition, high mutation rates would totally prevent any convergence to any optimal solution, as mutation gets more and more destructive as the population contains better and better solutions. Within ESs, on the other hand, all individuals undergo mutation, and only the strength of the mutation is varied (e.g., the standard deviation of the Gaussian mutation, in one dimension). Together with the fact that several offspring are generated from each parent, this nevertheless still allows the algorithm to converge. Furthermore, it can converge to any arbitrary precision provided the mutation strength is decreased accordingly: this was first demonstrated by self-adaptive ES [8,25], and proved right for CMA-ES as well [10]. Indeed, adaptive and self-adaptive mutations can have a significant impact only when all individuals undergo mutation. Furthermore, it also requires some property of the mutation with respect to the fitness function called the *strong causality principle* emphasized in Ref. 7: mutation should be parameterized in such a way that small mutations have small effects on the fitness. This property is not specific to floating-point optimization, and should be kept in mind when designing mutation operators for a particular

representation. In particular, this property is not true when floating-point numbers are encoded into binary strings (as is the case in traditional style GAs).

- *The Selection Mechanisms.* They range from the totally stochastic fitness proportional parental selection of GAs with generational survival, to the deterministic (μ, λ) survival selection of ES, through the stochastic, but elitist (i.e., preserving the best), tournament-based survival step of EP and the steady-state scheme used in GP. Though some studies have been devoted to selection mechanisms [6], the choice of both selection steps for a given problem (fitness representation operators) is still an open question (and is very likely to be problem-dependent).

The current trend in the EC community is to mix up all these features on a very pragmatic basis, trying to best fit the application at hand: some ESs applications deal with discrete or mixed real-integer spaces [22], the early arguments [26] against the dogma “binary is the best” have diffused in the community, and the Schema Theorem has been extended to any representation [5]. Moreover, most ESs (including CMA-ES) use some crossover operator, mutation has been added to GP, and so on. And the different selection operators are more and more being used now by the whole community.

On the other hand, such hybrid algorithms, by getting away from the simple original algorithms, also escape the few available theoretical results, and EAs can today only rely on successful applications (see the section titled “Application Areas”) to demonstrate their usefulness as general-purpose optimization algorithms.

THEORETICAL RESULTS

Theoretical studies of EAs can be roughly categorized into three different types.

- An EA can be viewed as a Markov chain in the space of populations, as the population at time $t + 1$ only depends on

the population at time t . The full theory of Markov chains can then be applied. Some asymptotic results were obtained for general EAs [27], and for more focused algorithms [28], but for very specific and simple functions. Stronger results (convergence in finite time) were obtained using the powerful Friedlin—Wentzell theory [29] for very general functions, but a specific algorithm. However, those results have limited practical consequences as the real-world environment often requires short response times.

- The specific characteristics of ESs allowed precise theoretical studies on the *sphere function*, that is, the quadratic function. Early theoretical studies by the ES pioneers resulted in Rechenberg's 1/5th rule [7], and in Schwefel's first self-adaptive ES [8]. These works were later pursued in Schwefel's group in Dortmund, with the precise study of the so-called *progress rate*, that is, the improvement from one step of the algorithm to the next, by Beyer [25]. Finally, these results that had been obtained asymptotically for very high dimensions, have been rigorously justified in any dimension. In particular, all numerical values for optimal settings have been shown to actually be optimal, with the help of the theory of irreducible Harris recurrent Markov chains [30,31].
- Last, but not least, the classical Algorithm Complexity Theory has been used to study EAs on discrete spaces. The pioneer of this approach was Ingo Wegener, and most works in this area come from his group in Dortmund (and some spin-off groups led by his former PhD students). Here the algorithms under study are simple but actual algorithms (and they get progressively less "simple" every year), while the fitness functions are simple functions or classes of functions. Results range from effective complexity for hitting the optimum of linear Boolean functions with a $(1+1)$ -ES [32] to general

bounds for stochastic search algorithms in generic black box scenarios [33].

However, one should keep in mind that all the above theoretical analyses address some simple models of EAs. As stated earlier, the modern trends of EC gave birth to algorithms working on poorly structured search spaces (see next section), or hybrid algorithms, for which no theory is applicable at the moment.

APPLICATION AREAS

It is widely acknowledged that EC theory lags far behind practice. Indeed, lessons from successful applications are one of the main driving forces of EA research today. Several edited books are devoted to applications of EAs (e.g., see the recent edition [34]), and almost every event related to EAs has its own special session dedicated to applied works, and their proceedings provide a wide overview of actual applications [e.g., see several special sessions of the annual *IEEE Congress on Evolutionary Computation* (CEC), and the *Real-World Application* track run every year during the *ACM Genetic and Evolutionary Computation Conference* (GECCO)].

This section will quickly survey the preferred domains of application of EAs, and the different subdomains will be distinguished according to the type of search space they involve.

Combinatorial Optimization

Hard combinatorial optimization problems (NP hard, NP complete) involve huge discrete search spaces, and have been studied extensively by the operational research community. Two different situations should be considered: academic benchmark problems and large real-world problems.

As far as benchmark problems are concerned, it is now commonly acknowledged that pure EAs alone cannot compete with operations research (OR) methods (see all papers about combinatorial optimization in Ref. 35, for instance). However, in the last decade, hybrid algorithms termed *genetic local search*, or *memetic algorithms*, where

the EA searches the space of local optima with respect to some OR heuristic, have obtained the best-so-far results on a number of such benchmark problems [36,37].

The situation is slightly different for real-world problems: “pure” OR heuristics generally do not directly apply, and OR methods have to take into account problem specificities. This is true of course for EAs, and there are many success stories where EAs, carefully tuned to the problem at hand, have been very successful, as for instance in the broad area of scheduling [38,39]. A few commercial applications of EAs were also reported; some of them resulted in a significant return on investment [40]. However, these applications also required forecasting components, resulting in hybrid adaptive business intelligence systems [41].

Parametric Optimization

The optimization of functions with floating-point variables has been thoroughly studied by practitioners, and many very powerful methods exist. Though the most well-known methods address linear convex problems, there are many other cases that can be handled successfully [42]. However, the recently developed CMA-ES [10,11] can be viewed as the leading EA in the continuous domain, outperforming the best methods from both deterministic and stochastic domains for highly multimodal, irregular, ill-conditioned, and nonseparable functions. An impressive list of applications of CMA-ES is maintained on its inventor’s web page [43].

The situation is different when dealing with multiobjective problems: multiobjective evolutionary algorithms (MOEAs) are the only ones that can produce a set of best possible compromises (the *Pareto set*), and have recently received increased attention. MOEAs use the same variation operators as standard EAs, but the Darwinian components are modified to take into account the multivalued fitness [44,45]. Note that MOEAs are discussed in the section titled “Parametric Optimization” because prominent application results have been obtained in that area [46], but can be applied on any search space, as only the Darwinian part

of the algorithm is different from that of a single-objective EA.

Mixed Search Spaces

Mixed search spaces involve different types of variables, generally both continuous and discrete, and EAs are flexible enough to handle such search spaces easily. Once variation operators are known for continuous and discrete variables, constructing variation operators for mixed individuals is straightforward: crossover, for instance, can either exchange values of corresponding variables, or use the variable-level crossover operator. Many problems have been easily handled that way, like the optical filter optimization [47,48], where one is looking for a number of layers, the unknown being the layer thickness (continuous variable) and the material the layer is made of (discrete variable). Furthermore, some platforms now exist that help the non-EC experts to implement generic EAs for a given problem involving different types of mixed search spaces, requiring only a structured description of potential solutions, and of course a routine to compute the fitness of those candidate solutions [49].

Artificial Creativity

A very promising area of application of EAs, where EAs can be much more than yet another optimization method, is that of design, where the ability of EAs to handle almost any search space allows programmers (and artists!) to unveil their wildest ideas. The idea of component-based representations can boost innovation in structural design [50,51], architecture [52], as well as in many other areas including art [53]. But the most original idea in that direction is that of embryogenesis: the genotype is a program, and the phenotype is the result of applying that program to “grow an embryo”; the fitness of the genotype (the program) is obtained by testing that phenotype in a real situation. Such an approach has already led to astonishing results in analog circuit design, for instance, [54]—though exploring a huge search space (a space of programs) implies a heavy computational cost. Note that these results were achieved using GP (see the

section titled “Genetic Programming”), using both crossover and mutation, and a huge (distributed) population as well. Nevertheless, we firmly believe that great achievements can come from original ideas related to these developmental approaches [55].

CONCLUDING REMARKS

Natural evolution can be considered as a powerful problem solver, which evolved *Homo sapiens* from chaos, in only a couple of billion years. Computer-based evolutionary processes can also be used as efficient problem solvers for optimization, constraint handling, machine learning, and modeling tasks. Furthermore, many real-world phenomena from the study of life, economy, and society can be investigated by simulations based on evolving systems. Last but not least, evolutionary art and design form an emerging field of applications of the Darwinian ideas. We expect that computer applications based on evolutionary principles will gain popularity in the coming years in science, business, and entertainment.

An interesting question, which is being raised from time to time, asks for guidance on the types of problems for which evolutionary methods are more appropriate than, say, standard OR methods. From our perspective, the best answer to this question is given in a single phrase: *complexity*. Let us explain.

Real-world problems are usually difficult to solve for several reasons, and include [1] the following.

- The number of possible solutions is so large as to forbid an exhaustive search for the best answer.
- The evaluation function that describes the quality of any proposed solution is noisy or varies with time, thereby requiring not just a single solution but an entire series of solutions.
- The possible solutions are so heavily constrained that constructing even one feasible answer is difficult, let alone searching for an optimum solution.

Naturally, this list could be extended to include many other possible obstacles. For

example, we could include noise associated with our observations and measurements, uncertainty about given information, and the difficulties posed by problems that have multiple and possibly conflicting objectives (which may require a set of solutions rather than a single solution). All these reasons are just various aspects of the complexity of the problem.

Note that every time we solve a problem we must realize that we are in reality only finding the solution to a *model* of the problem. All models are a simplification of the real world—otherwise they would be as complex and unwieldy as the natural setting itself. Thus, the process of problem solving consists of two separate general steps: (i) creating a model of the problem, and (ii) using that model to generate a solution:

Problem $\Rightarrow$ Model $\Rightarrow$ Solution.

Note that the “solution” is only a solution in terms of the model. If our model has a high degree of fidelity, we can have more confidence that our solution will be meaningful. In contrast, if the model has too many unfulfilled assumptions and rough approximations, the solution may be meaningless, or worse.

So, in solving real-world problems there are at least two ways to proceed:

1. we can try to simplify the model so that traditional methods might return better answers or
2. we can keep the model with all its complexities, and use nontraditional approaches, to find a near-optimum solution.

In other words, the more complex the problem (e.g., size of the search space, evaluation function, noise, constraints), the more appropriate it is to use a nontraditional method, like EAs. Anyway, it is difficult to obtain a precise solution to a problem because we either have to approximate a model or approximate the solution. And a large volume of experimental evidence shows that this latter approach can often

be used to practical advantage (see, e.g., www.SolveITSoftware.com).

One can also look at EAs from a broader perspective of population-based methods, where a set of potential solutions is being processed in parallel. Apart from EAs, there are other population-based approaches that have been proposed over the last 20 years; these include differential evolution [56], artificial immune systems [57], particle swarm optimization [58], ant colony optimization [59], and cultural algorithms [60].

REFERENCES

1. Michalewicz Z, Fogel D. How to solve it: modern heuristics. 2nd ed. New York: Springer; 2004.
2. Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. 1st–3rd ed. New York: Springer; 1992–1996.
3. Holland JH. Adaptation in natural and artificial systems. Ann Arbor (MI): University of Michigan Press; 1975.
4. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading MA: Addison Wesley; 1989.
5. Radcliffe NJ. Equivalence class analysis of genetic algorithms. *Complex Syst* 1991;5:183–120.
6. Chakraborty U, Deb K, Chakraborty M. Analysis of selection algorithms: a Markov chain approach. *Evol Comput* 1996;4(2):133–168.
7. Rechenberg I. Evolutionstrategie: optimierung technischer systeme nach prinzipien des biologischen evolution. Stuttgart: Fromman-Holzboog; 1972.
8. Schwefel H-P. Numerical optimization of computer models. 2nd ed. New York: John Wiley & Sons, Inc.; 1995.
9. Bäck T, Schwefel H-P. An overview of evolutionary algorithms for parameter optimization. *Evol Comput* 1993;1(1):1–23.
10. Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. *Evol Comput* 2001;9(2):159–195.
11. Auger A, Hansen N. A restart CMA evolution strategy with increasing population size. *Proceedings of the CEC'05*. Piscataway (NJ): IEEE Press; 2005. pp. 1769–1776.
12. Fogel LJ, Owens AJ, Walsh MJ. Artificial intelligence through simulated evolution. New York: John Wiley & Sons, Inc.; 1966.
13. Fogel DB. Evolutionary computation: toward a new philosophy of machine intelligence. Piscataway (NJ): IEEE Press; 1995.
14. Saravanan N, Fogel DB, Nelson KM. A comparison of methods for self-adaptation in evolutionary algorithms. *Biosystems* 1995;36:157–166.
15. Koza JR. Genetic programming: on the programming of computers by means of natural evolution. Cambridge (MA): MIT Press; 1992.
16. Syswerda G. A study of reproduction in generational and steady state genetic algorithm. In: Rawlins GJE, editor. *Foundations of genetic algorithms*. San Francisco (CA): Morgan Kaufmann; 1991. pp. 94–101.
17. Nordin P. Evolutionary program induction of binary machine code and its applications. Muenster: Krehl; 1997.
18. Miller JF, Smith SL. Redundancy and computational efficiency in cartesian genetic programming. *IEEE Trans Evol Comput* 2006;10:167–174.
19. Spector L, Robinson A. Genetic programming and autoconstructive evolution with the push programming language. *Genet Program Evolvable Machine* 2002;3(1):7–40.
20. Banzhaf W, Nordin P, Keller R, Francone F. Genetic programming – an introduction on the automatic evolution of computer programs and its applications. San Francisco (CA): Morgan Kaufmann; 1998.
21. Davis L. Handbook of genetic algorithms. New York: Van Nostram Reinhold; 1991.
22. Bäck T. Evolutionary algorithms in theory and practice. New York: Oxford University Press; 1995.
23. Syswerda G. Uniform crossover in genetic algorithms. In: Schaffer JD, editor. *Proceedings of the 3rd International Conference on Genetic Algorithms*. San Mateo (CA): Morgan Kaufmann; 1989. pp. 2–9.
24. Jones T. Crossover, macromutation and population-based search. In: Eshelman LJ, editor. *Proceedings 6th ICGA*. San Mateo (CA): Morgan Kaufmann; 1995. pp. 73–80.
25. Beyer H-G. The theory of evolution strategies. Berlin: Springer; 2001.
26. Antonisse J. A new interpretation of schema notation that overturns the binary encoding constraint. In: Schaffer JD, editor. *Proceedings of the 3rd International Conference on Genetic Algorithms*. San Mateo (CA): Morgan Kaufmann; 1989. pp. 86–91.
27. Eiben A, Aarts E, Hee KV. Global convergence of genetic algorithms: a Markov chain

- analysis. In: Schwefel H-P, Männer R, editors. *Proceedings of the 1st Parallel Problem Solving from Nature*. Berlin, Germany: Springer; 1991. pp. 4–12.
28. Rudolph G. Convergence properties of evolutionary algorithms. Hamburg: Kovac; 1997.
29. Cerf R. An asymptotic theory of genetic algorithms. In: Alliot J-M, Lutton E, Ronald E, Schoenauer M, Snyers D, editors. *Artificial evolution*, volume 1063 of LNCS. Springer; 1996. pp. 37–53.
30. Auger A. Convergence results for $(1, \lambda)$ -SA-ES using the theory of φ -irreducible Markov chains. *Theor Comput Sci* 2005;334(1–3):35–69.
31. Auger A, Hansen N. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In: Cattolico M, editor. *Proceedings of the ACM GECCO'06*. New York: ACM Press. 2006; pp. 445–452.
32. Droste S, Jansen T, Wegener I. A rigorous complexity analysis of the $(1 + 1)$ evolutionary algorithm for separable functions with boolean inputs. *Evol Comput* 1998;6(2):185–196.
33. Droste S, Jansen T, Wegener I. Upper and lower bounds for randomized search heuristics in black-box optimization. *Theor Comput Syst* 2006;39(4):525–544.
34. Yu T, Davis L, Baydar C, Roy R, editors. *Evolutionary Computation in Practice*. Number 88 in *Studies in Computational Intelligence*. Berlin, Germany: Springer; 2008.
35. Bäck T, Fogel D, Michalewicz Z, editors. *Handbook of Evolutionary Computation*. New York: Oxford University Press; 1997.
36. Merz P, Freisleben B. Fitness landscapes and memetic algorithm design. In: Corne D, Dorigo M, Glover F, editors. *New Ideas in Optimization*. London: McGraw-Hill; 1999. pp. 245–260.
37. Merz P, Huhse J. An iterated local search approach for finding provably good solutions for very large tsp instances. In: Rudolph G, *et al.*, editors. *Proc. PPSN X*. Number 5199 in LNCS, Berlin, Germany: Springer; 2009. pp. 929–939.
38. Paechter B, Rankin R, Cumming A, Fogarty TC. Timetabling the classes of an entire university with an evolutionary algorithm. In: Bäck T, Eiben A, Schoenauer M, Schwefel H-P, editors. *Proceedings of the 5th Conference on Parallel Problems Solving from Nature*. Berlin, Germany: Springer; 1998.
39. Semet Y, Schoenauer M. On the benefits of inoculation, an example in train scheduling. In: *Proceedings of the GECCO'06*. New York: ACM Press; 2006. pp. 1761–1768.
40. Michalewicz Z, Schmidt M, Michalewicz M, Chiriac C. A Decision-Support System based on Computational Intelligence: A Case Study. *IEEE Intell Syst* 2005;20:44–49.
41. Michalewicz Z, Schmidt M, Michalewicz M, Chiriac C. *Adaptive Business Intelligence*. New York: Springer; 2006.
42. Bonnans F, Gilbert J, Lemarechal C, Sagastizábal C. *Optimisation numérique, aspects théoriques et pratiques*. Volume 23 of *Mathématiques & Applications*. Berlin, Germany: Springer; 1997.
43. Hansen N (2009+). References to CMA-ES applications.
44. Deb K. *Multi-objective optimization using evolutionary algorithms*. Chichester: John Wiley; 2001.
45. Coello CAC, Veldhuizen DAV, Lamont GB. *Evolutionary algorithms for solving multi-objective problems*. Boston (MA): Kluwer Academic Publishers; 2002.
46. Obayashi S. Pareto genetic algorithm for aerodynamic design using the Navier-Stokes equations. In: Quadraglia D, Périaux J, Poloni C, Winter G, editors. *Genetic algorithms and evolution strategies in engineering and computer sciences*. Chichester: John Wiley; 1997. pp. 245–266.
47. Bäck T, Schütz M. Evolution strategies for mixed-integer optimization of optical multi-layer systems. In: McDonnell JR, Reynolds RG, Fogel DB, editors. *Proceedings of the 4th Annual Conference on Evolutionary Programming*. Cambridge (MA): MIT Press; 1995.
48. Martin S, Rivory J, Schoenauer M. Synthesis of optical multi-layer systems using genetic algorithms. *Appl Optic* 1995;34:2267.
49. Da Costa L, Schoenauer M. Bringing evolutionary computation to industrial applications with GUIDE. In: Raidl G, *et al.*, editors. *Genetic and Evolutionary Computation Conference*. New York: ACM Press; 2009. pp. 1467–1474.
50. Gero J. Adaptive systems in designing: New analogies from genetics and developmental biology. In: Parmee I, editor. *Adaptive Computing in Design and Manufacture*. Berlin, Germany: Springer; 1998. pp. 3–12.
51. Hamda H, Schoenauer M. Topological optimum design with evolutionary algorithms. *J Convex Anal* 2002;9:503–517.

52. Rosenman M. Evolutionary case-based design. In: Artificial evolution'99. LNCS 1829. Berlin, Germany: Springer; 1999. pp. 53–72.
53. Bentley PJ, editor. Evolutionary design by computers. San Francisco (CA): Morgan Kaufman Publishers Inc; 1999.
54. Koza JR. Genetic programming III: automatic synthesis of analog circuits. Massachusetts: MIT Press; 1999.
55. Stanley KO. Compositional pattern producing networks: A novel abstraction of development. Genetic Programming and Evolvable Machines. Special Issue on Developmental Systems 2007;8(2):131–162.
56. Price KV, Storn RM, Lampien JA. Differential evolution. Berlin, Germany: Springer; 2005.
57. Dasgupta D, Niño LF. Immunological computation. Boca Raton (FL): CRC Press; 2009.
58. Clerc M. Particle swarm optimization. Chichester: Wiley; 2006.
59. Dorigo M, Stützle T. Ant colony optimization. Cambridge (MA): The MIT Press; 2004.
60. Cowan GS, Reynolds RG. Acquisition of software engineering knowledge: sweep, an automatic programming system based on genetic programming and cultural algorithms. Hackensack (NJ): World Scientific Publishing Company; 2004.

EVOLUTIONARY GAME THEORY AND EVOLUTIONARY STABILITY

ROSS CRESSMAN
Department of Mathematics,
Wilfrid Laurier University,
Waterloo, Ontario, Canada

Evolutionary game theory developed as a means to predict the expected distribution of individual behaviors in a biological system with a single species that evolves under the influence of natural selection [1,2]. The theory's predictions correspond to intuitive static solutions (e.g., the ESS of Definition 1 below) of the game formed through fitness (i.e., payoff) comparisons of different behaviors (i.e., strategies). The ESS then provides a definition of stability in the biological system that does not rely on an explicit dynamical model of behavioral evolution. When such a solution exhibits evolutionary stability (i.e., when it is a dynamically stable equilibrium of the evolving population), the detailed analysis of the underlying dynamical system can be ignored. Instead, it is the heuristic static conditions of evolutionary stability that become central to understanding behavioral evolution when complications such as genetic, spatial, and population size effects are added to evolutionary dynamics.

This biological perspective of evolutionary game theory [and especially its relationship to dynamic stability when the evolutionary system is modeled by the deterministic replicator equation (see Definition 1 below)] has been summarized in several survey monographs and books [3–7]. Evolutionary game theory has also become a standard method to analyze games that model human interactions between players that can be individuals or other economic entities such as firms or even nations [8,9]. Biological behavioral patterns then correspond to a player's strategy and the fitness generated when patterns interact become the strategy's

payoff. Evolutionary stability can then be interpreted as a strategic situation which has achieved an equilibrium in the sense of dynamical systems.

A fundamental result is that, at evolutionary stability of a behavioral distribution, no individual in the population can increase its fitness by unilaterally changing strategy. That is, evolutionary stability implies Nash equilibrium (NE) behavior: a result that has come to be known as (one aspect of) the folk theorem of evolutionary game theory [10,11]. However, as we will see, evolutionary stability requires more than NE. Thus, it can be viewed as an NE refinement or equilibrium selection technique. It is in this latter capacity that evolutionary game theory initially gained prominence in the economic literature when applied to rational decision making in classical noncooperative, symmetric games in either normal form or extensive form (See Refs 12, 13 and the references therein). These evolutionary games, that have linear payoffs and finitely many strategies, often consider other deterministic or stochastic evolutionary dynamics besides the replicator equation since these are thought to better represent decision making applied to economic or learning models [10,13–20]. The theory has been extended to nonsymmetric games that model two (or more) species and, more recently, to population games with nonlinear payoffs and/or a continuum of strategies [9,11,21].

Evolutionary stability is most clear when individuals in a single species can use only two possible strategies, denoted as p^* and p to match notation used later in this article, and payoffs are linear. Suppose that $\pi(p, \hat{p})$ is the payoff to a strategy p user against strategy $\hat{p}$ (or, in biological terms, the fitness of an individual using p in a random pairwise interaction in a large population exhibiting behavior $\hat{p}$). Then, an individual in a monomorphic population where everyone uses p^* cannot improve its fitness by switching to p if

$$\pi(p, p^*) \leq \pi(p^*, p^*) \quad \text{NE condition.} \quad (1)$$

If p does as well as p^* against p^* (i.e., if $\pi(p, p^*) = \pi(p^*, p^*)$), then evolutionary stability requires the extra condition that p^* must do better than p against p . That is,

$$\pi(p, p) < \pi(p^*, p) \text{ if } \pi(p, p^*) = \pi(p^*, p^*) \\ \text{stability condition.} \quad (2)$$

To see why both these conditions are necessary for dynamic stability, under the assumption of Maynard Smith [22] that “like begets like,” the per capita change in the number of individuals using strategy p is its expected payoff. This leads to the following continuous-time invasion dynamics of a resident monomorphic population p^* by a small proportion ε of mutants using p (it is the replicator equation for the game with two strategies, p and p^*).

$$\begin{aligned} \dot{\varepsilon} &= \varepsilon[\pi(p, \varepsilon p + (1 - \varepsilon)p^*) - \pi(\varepsilon p + (1 - \varepsilon)p^*, \\ &\quad \varepsilon p + (1 - \varepsilon)p^*)] \\ &= \varepsilon(1 - \varepsilon)[\pi(p, \varepsilon p + (1 - \varepsilon)p^*) \\ &\quad - \pi(p^*, \varepsilon p + (1 - \varepsilon)p^*)] \\ &= \varepsilon(1 - \varepsilon)[(1 - \varepsilon)(\pi(p, p^*) - \pi(p^*, p^*)) \\ &\quad + \varepsilon(\pi(p, p) - \pi(p^*, p))]. \end{aligned} \quad (3)$$

If p^* is a strict NE (i.e., the inequality in (Eq. 1) is strict), then $\dot{\varepsilon} < 0$ for all positive ε that are close to 0 since $(1 - \varepsilon)(\pi(p, p^*) - \pi(p^*, p^*))$ is the dominant term corresponding to the common interactions against p^* . Furthermore, if this term is 0 (i.e., if p^* is not strict), we still have $\dot{\varepsilon} < 0$ from the stability condition of Equation (2) corresponding to the less common interactions against p . Conversely, if $\dot{\varepsilon} < 0$ for all positive ε that are close to 0, then p^* satisfies conditions (1) and (2). Thus, the resident population p^* (i.e., $\varepsilon = 0$) is locally asymptotically stable¹

under Equation (3) (i.e., there is evolutionary stability at p^* under the replicator equation) if and only if p^* satisfies Equations (1) and (2). In fact, evolutionary stability occurs at such a p^* in these two-strategy games for any evolutionary dynamics whereby the proportion (or frequency) of users of strategy $\hat{p}$ increases if and only if its expected payoff is higher than that of the alternative strategy.

Moreover, if conditions (1) and (2) hold for all other alternative strategies in a symmetric game with finitely many pure strategies and linear payoffs, then p^* is an ESS (i.e., an *evolutionarily stable strategy*) according to John Maynard Smith in his influential book, *Evolution and the Theory of Games* [22, p. 14]. These concise mathematical conditions translate his informal definition (p. 10) as “a strategy such that, if all members of the population adopt it, then no mutant strategy could invade the population under the influence of natural selection.” This article begins with a thorough discussion of the connection between the static ESS conditions and evolutionary stability for these symmetric normal-form games. The complex relationship between these concepts here serves as the prototype of evolutionary game theory.

Additional complications in this relationship that arise in each of several types of more general games are briefly considered in the remainder of the article. Throughout, the biological perspective and terminology of evolutionary game theory is emphasized. The concepts and results in this interpretation can all be translated into a form more typical for games involving humans. In fact, it is this reinterpretation of the theory of biological evolutionary games as economic evolutionary games (and vice versa) that has historically fed the explosive growth of evolutionary game theory over the past thirty years.

¹Clearly, the unit interval $[0, 1]$ is (forward) invariant under the dynamics of Equation (3) (i.e., if $\varepsilon(t)$ is the unique solution of Equation (3) with initial value $\varepsilon(0) \in [0, 1]$, then $\varepsilon(t) \in [0, 1]$ for all $t \geq 0$). The rest point $\varepsilon = 0$ is (neutrally) stable if, for every

neighborhood U of 0 relative to $[0, 1]$, there exists a neighborhood V of 0 such that $\varepsilon(t) \in U$ for all $t \geq 0$ if $\varepsilon(0) \in V \cap [0, 1]$. It is *attracting* if, for some neighborhood U of 0 relative to $[0, 1]$, $\varepsilon(t)$ converges to 0 whenever $\varepsilon(0) \in U$. It is *(locally) asymptotically stable* if it is both stable and attracting.

EVOLUTIONARY GAME THEORY FOR SYMMETRIC NORMAL-FORM GAMES

In an evolutionary game with symmetric normal form, the population consists of individuals who must all use one of finitely many (say n) possible behaviors at any particular instant in time. These strategies are denoted as e_i for $i = 1, \dots, n$ and $S \equiv \{e_1, \dots, e_n\}$ is called the *pure strategy set*. An individual may also use a mixed strategy in $\Delta^n \equiv \{p = (p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1, p_i \geq 0\}$, where p_i is the proportion of the time this individual uses pure strategy e_i . If population size is large and the components of $\hat{p} \in \Delta^n$ are the current frequencies of strategy use in the population (i.e., $\hat{p}$ is the population state), then the payoff of an individual using p in a random pairwise interaction is given explicitly through the bilinear payoff function of the (two-player) symmetric normal-form game, $\pi(p, \hat{p}) \equiv \sum_{i,j=1}^n p_i \pi(e_i, e_j) \hat{p}_j$, where as before $\pi(e_i, e_j)$ is the payoff to e_i against e_j .²

The following definition generalizes the ESS concept and the replicator equation developed for the two-strategy game in the introductory section to games with finitely many strategies.

Definition 1. Suppose S is the pure-strategy set for a symmetric normal-form game.

- (a) $p^* \in \Delta^n$ is an *ESS* if it satisfies conditions (1) and (2) for all $p \in \Delta^n$ with $p \neq p^*$.
- (b) The (pure-strategy) *replicator equation*, which is invariant on Δ^n , is given by

$$\dot{p}_i = p_i (\pi(e_i, p) - \pi(p, p)) \quad \text{for } i = 1, \dots, n. \quad (4)$$

²If e_i is represented by the i^{th} unit column vector in R^n and $\pi(e_i, e_j)$ is entry A_{ij} in an $n \times n$ payoff matrix A , then $\pi(p, \hat{p})$ is the inner product $p \cdot A\hat{p}$ of the two column vectors p and $A\hat{p}$. For this reason, symmetric normal form games are often called *matrix games* with payoffs given in this latter form.

The replicator equation emerges from the biological interpretation [23] of fitness as reproductive success (i.e., the individual's expected number of offspring). It has the intuitive property that strategies with higher current payoff increase in relative frequency (i.e., become more common). For two-strategy games, the replicator equation then increases the frequency of the better strategy. One of the early successes of evolutionary game theory is the following result [4,23].

Theorem 1. *If $p^* \in \Delta^n$ is an ESS, then p^* is asymptotically stable with respect to the replicator equation (Eq. 4).*

That is, evolutionary stability of an ESS holds for the replicator equation for multiple pure strategies and not only for the invasion of p^* by a subpopulation using a single mutant strategy p . The converse of Theorem 1 is true when there are two pure strategies (i.e., $n = 2$) but not in general. In fact, there already exist non-ESS strategies p in three-strategy symmetric normal-form games (i.e., for $n = 3$) that are asymptotically stable under Equation (4) (such strategies p must be NE by the folk theorem). To illustrate this, consider the payoff matrix

$$\begin{array}{c} R \\ P \\ S \end{array} \begin{bmatrix} 0 & 2 - \alpha & 4 \\ 6 & 0 & -4 \\ -2 & 8 - \alpha & 0 \end{bmatrix}, \quad (5)$$

with parameter α . For $2 < \alpha < 8$, this is a generalized Rock–Paper–Scissors (RPS) game since it exhibits cyclic dominance (i.e., P strictly dominates in the two-strategy $\{R, P\}$ game, S strictly dominates in the two-strategy $\{P, S\}$ game, and R strictly dominates in the two-strategy $\{R, S\}$ game). For $-16 < \alpha < 14$, there is a unique symmetric NE; namely, $p^* = (28 - 2\alpha, 20, 16 + \alpha)/(64 - \alpha)$. This is an ESS if and only if $-16 < \alpha < 3.5$ but it is globally asymptotically stable under Equation (4) if and only if $-16 < \alpha < 6.5$. That is, p^* has evolutionary stability with respect to the replicator equation but is not an ESS for $3.5 \leq \alpha < 6.5$. On the other hand, p^* is asymptotically stable under all (mixed-strategy) replicator equations for which some

distribution of these mixed strategy types have average strategy p^* (i.e., p^* is *strongly stable* according to Refs 6 and 15) if and only if p^* is an ESS. Thus, an ESS is equivalent to evolutionary stability with respect to mixed-strategy replicator equations.

There are several important classes of symmetric games that always have an ESS. One such class consists of all matrix games with a symmetric payoff matrix (i.e., $\pi(\hat{p}, p) = \pi(p, \hat{p})$ for all $p, \hat{p} \in \Delta^n$) for which the quadratic function $\pi(p, p)$ on Δ^n has an isolated local maximum.³ In fact, p^* is an ESS if and only if p^* is an isolated local maximum. Moreover, only the ESSs have evolutionary stability with respect to the replicator equation. These results follow from the Fundamental Theorem of Natural Selection [4] that asserts population mean payoff $\pi(p, p)$ is strictly increasing except at equilibrium when fitness is modeled by a single diploid locus with finitely many alleles. It was recognized early on [25] that the replicator equation is then the selection equation of population genetics that models the evolution of allele frequencies at this locus when the viability of genotype $A_i A_j$ equals $\pi(e_i, e_j)$. In the game-theoretic context, these games are known as *partnership games* [4], *doubly symmetric games* [13], or *common interest games* [9].

Another important class that has a unique ESS is the set of *strictly stable games*, which are games with $\pi(p - \hat{p}, p - \hat{p}) < 0$ for all $p \neq \hat{p}$ in Δ^n [9]. Any strictly stable game has a unique NE and this strategy is an ESS that is globally asymptotically stable with respect to

the replicator equation. All games with a completely mixed ESS p^* (i.e., p^* is in the interior of Δ^n) are strictly stable [15]. So are war-of-attrition games with an increasing cost of waiting and a finite set of waiting times [10] as well as many congestion games that model traffic flow in transportation networks [9]. Single-species habitat selection games (with linear payoffs) are also strictly stable [26] with unique ESS more commonly known as the *ideal free distribution* (IFD) [27].

There are also games that have multiple ESSs. The two-strategy symmetric coordination game

$$\begin{matrix} e_1 & \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \\ e_2 & \end{matrix}$$

is the most elementary example. This has three symmetric NE; namely, both pure strategies as well as the mixed strategy $p^* = (\frac{2}{3}, \frac{1}{3}) \in \Delta^2$. Only the pure strategies are ESSs (they are also the only asymptotically stable strategies). Thus, equilibrium selection via evolutionary stability discards the unstable mixed NE, leaving both pure strategies as possible evolutionary outcomes. Stochastic stability [14, 28, 29] can then be used to select the ESS e_2 since it is the risk-dominant [30] strict NE. In terms of deterministic evolutionary dynamics, e_2 is selected since it has the larger basin of attraction (i.e., the set of initial conditions that evolve to e_2 is a larger interval than the interval that evolves to e_1). Stochastic effects, due to infrequent (but ongoing) individual experimentation of the other strategy, then make it more difficult for the population to shift from the larger basin of attraction to the smaller one than vice versa, implying the system will evolve near e_2 .

Unfortunately, there are many games that have no ESS or asymptotically stable NE.⁴ In particular, the standard normal form of an extensive form game often has no ESS since every such ESS must be pervasive (i.e., it must reach every information set when played against itself) and can have

³If the quadratic function $\pi(p, p)$ has a local maximum p on Δ^n that is not isolated, then the maximal connected set of local maxima E containing p is an ESSet (evolutionarily stable set; (i.e., every $p^* \in E$ is a NE and, if $\pi(p, p^*) = \pi(p^*, p^*)$ for $p \notin E$, then $\pi(p^*, p) > \pi(p, p)$)) that is setwise asymptotically stable with respect to the replicator equation [6, 24]. Every symmetric game with two pure strategies is payoff equivalent to a game with a symmetric payoff matrix (i.e., these two games have the same NE and the same ESS etc.) and so has at least one ESSet. In fact, each two-strategy game has an ESS unless it is selectively neutral (i.e., unless $\pi(\hat{p}, p) = \pi(p, p)$ for all $p, \hat{p} \in \Delta^n$).

⁴For example, take $3.5 \leq \alpha < 14$ or $6.5 \leq \alpha < 14$ respectively in the RPS game of payoff matrix (5).

no other realization-equivalent strategies [12]. To ease this latter problem, Selten [31] defined an ESS directly in terms of behavior strategies (i.e., strategies that specify the local behavior at each player information set of the game tree). His *direct ESS* is then a behavior strategy b^* that satisfies conditions, (1) and (2) for any other behavior strategy b . Each direct ESS corresponds to a set of normal-form strategies that is an ESSet but the converse is not true [32]. In particular, b^* must still be pervasive and so a pure-strategy nonpervasive NE outcome that yields the strictly highest possible payoff whenever either player uses this strategy (e.g., an outside option) cannot be a direct ESS. Furthermore, every direct ESS can be found by applying the backward induction technique with respect to the ES structure of the subgames and their corresponding truncations [31] and so is a subgame-perfect NE. However, this technique often fails to yield a direct ESS [12] as Selten originally hoped.⁵

As an elementary example, consider backward induction applied to the symmetric extensive form game of Fig. 1 taken from Ref. 12. Its second-stage subgame $\begin{matrix} \ell & \begin{bmatrix} -5 & 5 \\ -4 & 4 \end{bmatrix} \\ r & \end{matrix}$ has mixed ESS $b_2^* = (\frac{1}{2}, \frac{1}{2})$ and when the second decision point of player one u_2 is replaced by the payoff 0 from b_2^* , the truncated single-stage game $\begin{matrix} L & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ R & \end{matrix}$ also has a mixed ESS $b_1^* = (\frac{1}{2}, \frac{1}{2})$. Since both stage games have a mixed ESS (and so a unique NE since they are strictly stable), (b_1^*, b_2^*) is the only (behavior strategy) NE of Fig. 1 and it is pervasive. Surprisingly, this example has no direct ESS since (b_1^*, b_2^*) does not satisfy conditions (1) and (2) for the pure strategy b given by Rr [12].

Figure 1 shows that, although backward induction determines candidates for the ES structure, it is not useful for determining

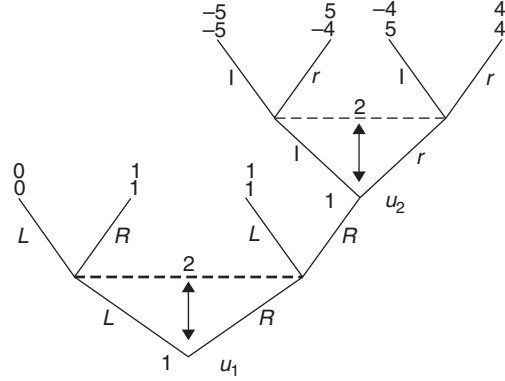


Figure 1. The extensive form of the van Damme example.

which candidates are actually direct ESSs.⁶ However, applying backward induction with respect to NE that exhibit evolutionary stability in the subgames and in their truncations shows more promise. For instance, each direct ESS b^* yields an ESSet in the game's normal form [10] and so is dynamically stable. More interestingly, for the class of (symmetric) simultaneity games where both players know all player actions at earlier stages and there are no moves by nature, Cressman [10] shows that, if b^* is a pervasive NE, then the set of realization-equivalent normal-form strategies exhibits evolutionary stability with respect to the replicator equation if and only if it comes from this backward induction process. In

⁵Similar problems arise for his related limit ESS concept [12,33] and for the ESS concept defined in terms of reduced normal forms of the extensive form game (e.g., the reduced-strategy normal form of Cressman [10]).

⁶The situation is more discouraging for nonpervasive NE. For example, the only NE outcome of the two-stage repeated Prisoner's Dilemma Game [34] with cumulative payoffs is mutual defection at each stage. This nonpervasive NE outcome cannot be an isolated behavior strategy (i.e., there is a corresponding NE component) and so there is no direct ESS. Worse, for typical single-stage payoffs such as $\begin{matrix} \text{defect} & \begin{bmatrix} -1 & 10 \\ -2 & 5 \end{bmatrix} \\ \text{cooperate} & \end{matrix}$, this component does not satisfy setwise extensions of the ESS (e.g., it is not an ESSet). It does, however, attract all trajectories of the replicator equation for the standard normal form (or reduced-strategy normal form) for which all pure strategies initially have positive frequency.

particular, there is evolutionary stability for the NE of Fig. 1 (that corresponds to the unique NE of its reduced-strategy normal form).

So far, we have summarized evolutionary stability with respect to the replicator equation that was originally developed through the biological translation of fitness into reproductive success. The replicator equation also emerges through imitation based on rational decision processes such as the proportional imitation rule [35] whereby each player samples a random individual's strategy and payoff, switching to this strategy (i.e., imitating) when it has a higher payoff at a rate proportional to the expected gain. Other imitation rules lead to different evolutionary game dynamics that satisfy various monotonicity properties [13,36,37] between higher payoff and an increase in (relative) strategy frequency. In general, there is a complex relationship between evolutionary stability under these dynamics and heuristic static stability conditions such as those defining an ESS [9,11].⁷

More is known for the best response dynamics, whose discrete-time version (called *fictitious play*) predates the replicator equation [38]. The replicator equation and best response dynamics, based on different assumptions of player information, are considered to be the two standard game dynamics. For the best response dynamics [39], each player knows the current payoff of all strategies and, in particular, the best reply to the current population state $p \in \Delta^n$ given by the set

$$BR(p) \equiv \{p' \in \Delta^n : \pi(p', p) \geq \pi(p'', p) \text{ for all } p'' \in \Delta^n\}.$$

If a small proportion of the population switches per unit time to the current best

reply, we have the best response dynamics

$$\dot{p} \in BR(p) - p, \quad (6)$$

a differential inclusion for which solutions exist [11].⁸ These solutions are piecewise linear trajectories that enjoy many of the same evolutionary stability properties as the replicator equation. In particular, each ESS exhibits evolutionary stability under this dynamics [40] and for strictly stable games, the unique ESS is globally asymptotically stable.

Nonlinear Payoff Functions

For symmetric normal-form games, the payoff $\pi_i(\hat{p})$ of pure strategy e_i is linear in the components of the population state $\hat{p}$ (i.e., $\pi_i(\hat{p}) = \sum_{j=1}^n \pi(e_i, e_j) \hat{p}_j$). Symmetric games with payoff $\pi_i(\hat{p})$ nonlinear in the population state $\hat{p}$ are quite common in biology and in economics [9,22] where they are called *playing-the-field models* or *population games* since payoffs are not assumed to result from two-player pairwise interactions. Following the approach for two-strategy games outlined in the introductory section using $\pi(p, \hat{p}) = \sum_i p_i \pi_i(\hat{p})$, evolutionary stability of p^* can be characterized by local versions of conditions (1) and (2) (i.e., p is assumed to be close to p^* in these inequalities). A more common equivalent characterization, called local or neighborhood superiority [13,41], is that

$$\pi(p^*, p) > \pi(p, p), \quad (7)$$

for all p sufficiently close (but not equal) to p^* . In fact, for multistrategy symmetric normal-form games, linearity implies condition (7) is equivalent to the ESS of Definition 1 [15].⁹

⁷Two-strategy symmetric games are the exception. For these, any reasonable evolutionary game dynamics (i.e., one that increases the frequency of the better strategy) will have evolutionary stability at p^* if and only if p^* is an ESS.

⁸Solutions are Lipschitz continuous functions of t whose derivative satisfies Equation (6) almost everywhere. There may be more than one solution for the same initial conditions since the vector field $BR(p) - p$ is discontinuous and can be multiple valued.

⁹In fact, a p^* satisfying inequality (7) is often called an ESS. The exact definition of "ESS" in games that are not of symmetric normal form is not universally

In biology, an example of evolutionary stability reasoning before the advent of (evolutionary) game theory is Fisher's [44] justification for the prevalence of the 50:50 sex ratio in diploid species (see also the "unbeatable strategy" concept of Hamilton [45] that predicts other sex ratios under different reproductive scenarios). Specifically, individuals who produce equal numbers of male and female offspring have higher fitness than average in a population with a biased sex ratio (see the sex-ratio game in Ref. 22 where nonlinear fitness is given by the expected number of grandchildren).

In general, nonlinearity complicates the application of evolutionary stability theory. In fact, Vickers and Cannings [46] give an example where a discrepancy between two ESS definitions already arises for a symmetric normal-form game with a countably infinite set of pure strategies S . Similar problems occur in the definition of an ESSet [47,48] for nonlinear payoff functions. However, these issues do not seem to emerge in practical applications and it is the local superiority condition that is typically used to prove its evolutionary stability under the replicator equation (and its mixed strategy counterpart).

A more pressing problem is that it is often quite difficult to determine the locally superior strategies in these games. Even two-strategy games can have several ESSs in the interior of Δ^2 . To illustrate these issues, consider the class of potential games [9]. These are games for which there is a continuously differentiable real-valued function V (called the *potential function*) defined on a neighborhood of Δ^n whose gradient satisfies $\nabla V(\hat{p}) = (\pi_1(\hat{p}), \dots, \pi_n(\hat{p}))$ for all $\hat{p} \in \Delta^n$. (The doubly symmetric normal-form games are of this type with potential function $V(\hat{p}) \equiv \frac{1}{2}\pi(\hat{p}, \hat{p})$.) If V is a strictly concave function, then the game is strictly stable and the absolute maximum point of V is the unique locally superior strategy. For general V , the isolated local maxima of V correspond to local

superiority. However, finding the local maxima of the potential function is a nontrivial task, especially when the potential function is nonquadratic or perhaps unknown. For non-potential games, local superiority becomes still more complicated.

ASYMMETRIC GAMES

Evolutionary theory also applies to nonsymmetric games. Following Ref. 49 (see also Refs 10 and 12), in a two-player asymmetric game, the players interact in pairwise contests after they are assigned a pair of roles from a finite set of N roles with probability that is independent of player designation. Furthermore, a player in a given role has a finite set of possible pure strategies and payoffs to mixed strategists are given by a bilinear payoff function. An asymmetric game with $N = 1$ is then a symmetric normal-form game. This section summarizes evolutionary stability when $N = 2$ with (pure) strategy sets $S = \{e_1, \dots, e_n\}$ and $T = \{f_1, \dots, f_m\}$ for roles one and two respectively.

The invasion dynamics of the resident monomorphic population $(p^*, q^*) \in \Delta^n \times \Delta^m$ by (p, q) generalizes Equation (3) to become

$$\begin{aligned} \dot{\varepsilon}_1 &= \varepsilon_1(1 - \varepsilon_1) \\ &\quad \times \begin{bmatrix} \pi_1(p; \varepsilon_1 p + (1 - \varepsilon_1)p^*, \varepsilon_2 q + (1 - \varepsilon_2)q^*) \\ -\pi_1(p^*; \varepsilon_1 p + (1 - \varepsilon_1)p^*, \varepsilon_2 q + (1 - \varepsilon_2)q^*) \end{bmatrix} \\ \dot{\varepsilon}_2 &= \varepsilon_2(1 - \varepsilon_2) \\ &\quad \times \begin{bmatrix} \pi_2(q; \varepsilon_1 p + (1 - \varepsilon_1)p^*, \varepsilon_2 q + (1 - \varepsilon_2)q^*) \\ -\pi_2(q^*; \varepsilon_1 p + (1 - \varepsilon_1)p^*, \varepsilon_2 q + (1 - \varepsilon_2)q^*) \end{bmatrix}, \end{aligned} \quad (8)$$

where $\pi_1(p; \varepsilon_1 p + (1 - \varepsilon_1)p^*, \varepsilon_2 q + (1 - \varepsilon_2)q^*)$ is the payoff to p when the current state of the population in roles one and two are $\varepsilon_1 p + (1 - \varepsilon_1)p^*$ and $\varepsilon_2 q + (1 - \varepsilon_2)q^*$, respectively. From the biological perspective, this is a two-species system where an individual's fitness depends on both interspecific and intraspecific interactions. By Cressman [6], (p^*, q^*) exhibits evolutionary stability under Equation (8) (i.e., $(0, 0)$ is always asymptotically stable under the above dynamics for all

agreed upon. There are many concepts that use different terminology but have ESS-like features in these games [42,43].

choices $p \neq p^*$ and $q \neq q^*$) if and only if

$$\begin{aligned} \text{either} \quad & \pi_1(p; p, q) < \pi_1(p^*; p, q) \\ \text{or} \quad & \pi_2(q; p, q) < \pi_2(q^*; p, q). \end{aligned} \quad (9)$$

Condition (9) is the two role analogue of neighborhood superiority (Eq. 7).

If condition (9) holds for all $(p, q) \in \Delta^n \times \Delta^m$ sufficiently close (but not equal) to (p^*, q^*) , then (p^*, q^*) is called a *two-species* ESS [10] or neighborhood superior [50]. This ESS enjoys similar evolutionary stability properties to the ESS of symmetric games. It is asymptotically for the replicator equation

$$\begin{aligned} \dot{p}_i &= p_i [\pi_1(e_i; p, q) - \pi_1(p; p, q)], \\ \dot{q}_j &= q_j [\pi_2(f_j; p, q) - \pi_2(q; p, q)], \end{aligned} \quad (10)$$

and for all its mixed strategy counterparts (i.e., (p^*, q^*) is strongly stable). Furthermore, if (p^*, q^*) is in the interior of $\Delta^n \times \Delta^m$, then it is globally asymptotically stable [10]. It is also globally asymptotically stable under the best response dynamics that generalizes Equation (6) to asymmetric games.

The dynamics (Eq. 10) is called the *two-species replicator equation* when players in role i are identified with species i . A biological example of two-species evolutionary stability analysis is the two-habitat selection model of Cressman *et al.* [51] when there is a Lotka–Volterra competitive two-species system in each patch and migration is always toward the patch with the highest species' payoff. An ESS always exists in this model and, depending on parameters, the ESS is mixed (i.e., both species coexist in each patch) in some cases while, in others, one of the species resides only in one patch at the ESS.

Bimatrix Games

A particularly important class of asymmetric games is the set of N -role truly asymmetric games where the two players are never assigned the same role (i.e., players in the same role do not interact or, in biological terms, there are no intraspecific interactions). Then, the N -tuple giving a strategy in each role is an ESS if and only if it is a

strict NE [49], and if and only if it has evolutionary stability with respect to the replicator equation (10) [10].

Thus, in contrast to symmetric games, we have an equivalence between the static ESS concept (Eq. 9) and evolutionary stability. However, this is an unsatisfactory result in the sense that a strict NE must be a pure-strategy pair and so is a very restrictive concept. In particular, a mixed ESS as in the habitat selection model referred to above is impossible in a truly asymmetric game. We again illustrate these issues when $N = 2$.

Since $\pi_1(e_i; p, q)$ does not depend on p , we can use the notation $\pi_1(e_i, q)$ instead (and also replace $\pi_2(f_j; p, q)$ with $\pi_2(p, f_j)$). Such a game is then represented by an $n \times m$ bimatrix (A, B) whose ij th entry is the pair of payoffs $(\pi_1(e_i, f_j), \pi_2(e_i, f_j))$ to the players in the two roles for the interspecific interaction between e_i and f_j . Standard examples, with two strategies for each player, include the Buyer–Seller Game [10] that has no ESS since its only NE is in the interior. Another is the Owner–Intruder Game (the bimatrix version of the Hawk–Dove Game) that has two strict NE which Maynard Smith [22] called the *bourgeois* ESS where the owners defend their territory (i.e., play Hawk if and only if the owner) and the paradoxical ESS where owners retreat (i.e., play Hawk if and only if the intruder).

One attempt to relax the ESS conditions in bimatrix games is through the concept of a strict equilibrium set (SESet), the analogue of the ESSet. This is a set F of NE (p^*, q^*) such that $(p, q^*) \in F$ if $\pi_1(p, q^*) = \pi_1(p^*, q^*)$ and $(p^*, q) \in F$ if $\pi_2(p^*, q) = \pi_2(p^*, q^*)$. Such sets correspond to the asymptotically stable sets of rest points of the replicator equation [10,11,52]. However, an SESet is still quite restrictive since F must contain all strategies p with support equal to or contained in any p^* with $(p^*, q^*) \in F$ and so cannot contain an interior NE (unless every strategy pair is Nash).

On the other hand, the Nash–Pareto pair of Hofbauer and Sigmund [4] does allow for isolated mixed strategy solutions. (p^*, q^*) is a Nash–Pareto pair if any pair of best replies (p, q) satisfies

$$\begin{aligned} &\text{either } \pi_1(p, q) \leq \pi_1(p^*, q) \\ &\quad \text{or } \pi_2(p, q) < \pi_2(p, q^*) \end{aligned}$$

and

$$\begin{aligned} &\text{either } \pi_2(p, q) \leq \pi_2(p, q^*) \\ &\quad \text{or } \pi_1(p, q) < \pi_1(p^*, q). \end{aligned}$$

Notice that both these inequality conditions are similar to condition (9). An example of a Nash–Pareto pair is the only NE of the Buyer–Seller Game. A noteworthy connection to evolutionary stability is that every Nash–Pareto pair is neutrally stable with respect to the replicator equation (10) [4].

CONCLUSION

The static ESS fitness comparisons introduced by Maynard Smith to predict the behavioral outcome of evolution in a single species have been extended in many directions during the intervening thirty years. These include biological extensions to multiple species and to population games as well as the equally important extensions to predict rational individual behavior in human conflict situations. As is apparent from this article, there is a complex relationship between these static conditions and evolutionary stability of the underlying dynamical system. Interestingly, the stability analysis has in many cases been determined before its connection with the ESS perspective was recognized.

This article has summarized evolutionary stability for deterministic evolutionary game dynamics based on random pairwise interactions in (symmetric or asymmetric) normal-form games. Evolutionary stability is also of much current interest for other game-theoretic models such as those that incorporate stochastic effects due to finite populations [20]; models of adaptive dynamics based on continuous (pure) strategy spaces [21,53,54]; models with assortative (i.e., nonrandom) interactions (e.g., games on graphs) [20]. There is also a great deal of interest, both theoretically and experimentally, on models for the evolution of human

cooperation whose underlying games are either the two-player Prisoner’s Dilemma Game or the multiplayer Public Goods Game [55]. As the evolutionary theory behind these models is a rapidly expanding area of current research, it is impossible to know in what guise evolutionary stability conditions will emerge in future applications. On the other hand, it is certain that Maynard Smith’s original idea will continue to play a central role.

REFERENCES

1. Maynard Smith J, Price G. The logic of animal conflicts. *Nature* 1973;246:15–18.
2. Maynard Smith J. The theory of games and the evolution of animal conflicts. *J Theor Biol* 1974;47:209–221.
3. Hines WGS. Evolutionarily stable strategies: A review of basic theory. *Theor Popul Biol* 1987;31:195–272.
4. Hofbauer J, Sigmund K. The theory of evolution and dynamical systems. Cambridge: Cambridge University Press; 1988.
5. Bomze IM, Pötscher BM. Game theoretical foundations of evolutionary stability, Lecture Notes in Economics and Mathematical Systems 324. Berlin: Springer; 1989.
6. Cressman R. The stability concept of evolutionary games (a dynamic approach), Lecture Notes in Biomathematics 94. Berlin: Springer; 1992.
7. Sigmund K. Games of life. Oxford: Oxford University Press; 1993.
8. Samuelson L. Evolutionary games and equilibrium selection. Cambridge (MA): MIT Press; 1997.
9. Sandholm WH. Population games and evolutionary dynamics. Cambridge (MA): MIT Press; 2010. Forthcoming.
10. Cressman R. Evolutionary dynamics and extensive form games. Cambridge (MA): MIT Press; 2003.
11. Hofbauer J, Sigmund K. Evolutionary game dynamics. *Bull Am Math Soc* 2003;40: 479–519.
12. van Damme E. Stability and perfection of Nash equilibria. 2nd ed. Berlin: Springer; 1991.
13. Weibull J. Evolutionary game theory. Cambridge (MA): MIT Press; 1995.
14. Vega-Redondo F. Evolution, games, and economic behavior. Oxford: Oxford University Press; 1996.

15. Hofbauer J, Sigmund K. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press; 1998.
16. Fudenberg D, Levine DK. *The theory of learning in games*. Cambridge (MA): MIT Press; 1998.
17. Young P. *Individual strategy and social structure*. Princeton (NJ): Princeton University Press; 1998.
18. Gintis H. *Game theory evolving*. Princeton (NJ): Princeton University Press; 2000.
19. Mesterton-Gibbons M. *An introduction to game-theoretic modelling*. 2nd ed. Providence (RI): American Mathematical Society; 2000.
20. Nowak MA. *Evolutionary dynamics*. Cambridge (MA): Harvard University Press; 2006.
21. Vincent TL, Brown J. *Evolutionary game theory, natural selection, and darwinian dynamics*. Cambridge: Cambridge University Press; 2005.
22. Maynard Smith J. *Evolution and the theory of games*. Cambridge: Cambridge University Press; 1982.
23. Taylor PD, Jonker L. Evolutionarily stable strategies and game dynamics. *Math Biosci* 1978;40:145–156.
24. Thomas B. On evolutionarily stable sets. *J Math Biol* 1985;22:105–115.
25. Akin E. *The geometry of population genetics*, Lecture Notes in Biomathematics 31. Berlin: Springer; 1979.
26. Krivan V, Cressman R, Schneider C. The ideal free distribution: A review and synthesis of the game theoretic perspective. *Theor Popul Biol* 2008;73:403–425.
27. Fretwell DS, Lucas HL. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheor* 1970;19:16–32.
28. Kandori M, Mailath GJ, Rob R. Learning, mutation, and long run equilibria in games. *Econometrica* 1993;61:29–56.
29. Foster D, Young HP. Stochastic evolutionary game dynamics. *Theor Popul Biol* 1990;38:219–232.
30. Harsanyi JC, Selten R. *A general theory of equilibrium selection in games*. Cambridge (MA): MIT Press; 1988.
31. Selten R. Evolutionary stability in extensive two-person games. *Math Soc Sci* 1983;5:269–363.
32. Cressman R. Evolutionarily stable sets in symmetric extensive two-person games. *Math Biosci* 1992;108:179–201.
33. Samuelson L. Limit evolutionarily stable strategies in two-player, normal form games. *Games Econ Behav* 1991;3:110–128.
34. Nachbar J. Evolution in the finitely repeated Prisoner's Dilemma. *J Econ Behav Organ* 1992;19:307–326.
35. Schlag KH. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *J Econ Theor* 1998;78:130–156.
36. Friedman D. Evolutionary games in economics. *Econometrica* 1991;59:637–666.
37. Samuelson L, Zhang J. Evolutionary stability in asymmetric games. *J Econ Theor* 1992;57:363–391.
38. Brown GW. Iterative solutions of games by fictitious play. In: Koopmans TC, editor. *Activity analysis of production and allocation*. New York: Wiley; 1951. pp. 374–376.
39. Matsui A. Best response dynamics and socially stable strategies. *J Econ Theor* 1992;57:343–362.
40. Sandholm WH. Local stability under evolutionary game dynamics. *Theor Econ* 2010;5:29–56.
41. Cressman R. Continuously stable strategies, neighborhood superiority and two-player games with continuous strategy spaces. *Int J Game Theor* 2009;38:221–247.
42. Apaloo J, Brown JS, Vincent TL. Evolutionary game theory: ESS, convergence stability, and NIS. *Evol Ecol Res* 2009;11:489–515.
43. Leimar O. Multidimensional convergence stability. *Evol Ecol Res* 2009;11:191–208.
44. Fisher RA. *The genetical theory of natural selection*. Oxford: Clarendon Press; 1930.
45. Hamilton WD. Extraordinary sex ratios. *Science* 1967;156:477–488.
46. Vickers G, Cannings C. On the definition of an evolutionarily stable strategy. *J Theor Biol* 1987;129:349–353.
47. Bomze IM. Uniform barriers and evolutionarily stable sets. In: Leinfellner W, Kohler E, editors. *Game theory, experience, rationality*. Dordrecht: Kluwer Academic Publishers; 1998. pp. 225–243.
48. Balkenborg D, Schlag KH. Evolutionarily stable sets. *Int J Game Theory* 2000;29:571–595.
49. Selten R. A note on evolutionarily stable strategies in asymmetrical animal contests. *J Theor Biol* 1980;84:93–101.
50. Cressman R. CSS, NIS and dynamic stability for two-species behavioral models with continuous trait spaces. *J Theor Biol* 2010;262:80–89.

51. Cressman R, Krivan V, Garay J. Ideal free distributions, evolutionary games, and population dynamics in multiple-species environments. *Am Nat* 2004;164:473–489.
52. Balkenborg D, Schlag KH. On the evolutionary selection of sets of Nash equilibria. *J Econ Theory* 2007;133:295–315.
53. Eshel I. Evolutionary and continuous stability. *J Theor Biol* 1983;103:99–111.
54. Dercole F, Rinaldi S. Analysis of evolutionary processes: the adaptive dynamics approach and its applications. Princeton (NJ): Princeton University Press; 2008.
55. Binmore K. Does game theory work: the bargaining challenge. Cambridge (MA): MIT Press; 2007.

EXACT SOLUTION OF THE CAPACITATED VEHICLE ROUTING PROBLEM

ROBERTO BALDACCI
DANIELE VIGO
DEIS, University of Bologna,
Cesena, Italy
PAOLO TOTH
DEIS, University of Bologna,
Bologna, Italy

INTRODUCTION

Vehicle routing problem (VRP) is a generic name given to a whole class of problems concerning the optimal design of routes to be used by a fleet of vehicles to serve a set of customers. Since it was first proposed by Dantzig and Ramser [1], hundreds of papers have been devoted to the exact and approximate solution of a member of the VRP class, known as the *capacitated vehicle routing problem* (CVRP), in which a homogeneous fleet of vehicles is available at a central depot, each vehicle can take a single route, and the only constraint considered is the vehicle capacity constraint.

An interesting survey covering the early exact methods for the CVRP is presented by Laporte and Nobert [2], and the book edited by Toth and Vigo [3] provides a comprehensive overview of exact methods for the CVRP and other variants proposed up to the end of the twentieth century. This work was updated from the survey of Cordeau *et al.* [4]. Recent surveys can be found in Baldacci *et al.* [5] and Bladacci *et al.* [6]. Many different heuristic algorithms are proposed in the literature for the CVRP and its variants. Among the various surveys on heuristic algorithms for the CVRP, we mention the surveys of Laporte and Semet [7] and of Gendreau *et al.* [8] in the book edited by Toth and Vigo [3].

In this article, we provide a review of the developments in the exact solution of the

CVRP on undirected graphs that had a major impact in the current state-of-the-art exact algorithms for this problem family. In particular, we present two mathematical formulations proposed in the literature that were used in the most successful exact approaches. We briefly survey the best-known exact algorithms that can be classified into the following categories: branch-and-bound, dynamic programming (DP), branch-and-cut, branch-and-cut-and-price, and set partitioning-based methods. As previously mentioned, we concentrate here on the undirected version of the problem since this is the most widely studied.

This article is organized as follows: In the following section we formally describe the CVRP and introduce the notation that will be used throughout the paper. The section titled “Mathematical Formulations of the CVRP” reviews two formulations and briefly describes the valid inequalities proposed for them. The section titled “Exact Methods for the CVRP” reviews the exact algorithms, while a comparison of the computational performances of the recent exact methods is reported in the final section.

PROBLEM DEFINITION AND BASIC NOTATION

In this section, we give a formal definition of the CVRP and introduce the relevant notation required to define the mathematical formulations.

In the CVRP all the customers correspond to deliveries, the demands are deterministic, known in advance, and may not be split. The vehicles are identical and based at a single central depot, each taking up exactly one route, and only the capacity restrictions for the vehicles are imposed. The objective is to minimize the total cost (i.e., their length or travel time) to serve all the customers.

More precisely, the CVRP may be described as the following graph theoretic problem.

Let $G = (V, E)$ be a complete and undirected graph where $V = \{0, \dots, n\}$ is the vertex set and E is the edge set. Vertex set $V_c = \{1, \dots, n\}$ corresponds to n customers, whereas vertex 0 corresponds to the depot. A nonnegative cost, d_{ij} , is associated with each edge $\{i, j\} \in E$ and represents the travel cost spent to go directly from vertex i to vertex j . In several practical cases, the cost matrix $D = [d_{ij}]$ satisfies the *triangle inequality*. Each customer $i \in V_c$ is associated with a known nonnegative demand, q_i , to be delivered. A set of m identical vehicles, each with capacity Q , is available at the depot. We assume that $q_i \leq Q$ for each $i \in V_c$. A *feasible route* (or simply *route*) is a simple cycle in G passing through the depot and a subset of customers of V_c such that the sum of the customer demands is less than or equal to Q .

In the variant of CVRP that is generally considered in the literature and that we study here, all available vehicles must be used, each performing exactly one route. We also assume that m is not smaller than $m_{\min}$, which is the minimum number of vehicles needed to serve all the customers. The value of $m_{\min}$ may be determined by solving the bin packing problem (BPP) associated with the CVRP, which calls for the determination of the minimum number of bins (i.e., vehicles), each with capacity Q , required to load all the n items (i.e., customers), each with nonnegative weight q_i , $i \in V_c$.

The CVRP consists of finding a collection of exactly m routes with minimum cost, defined as the sum of the costs of the edges belonging to the routes, and such that each customer vertex is visited by exactly one route.

The CVRP is known to be $\mathcal{NP}$ -hard (in the strong sense), and generalizes the well-known traveling salesman problem (TSP), which calls for the determination of a minimum cost simple cycle visiting all the vertices of G (Hamiltonian cycle), and arising when $Q \geq \sum_{i \in V_c} q_i$ and $m = 1$.

If G is a directed graph and the cost matrix D is asymmetric, the corresponding problem is called *asymmetric capacitated vehicle routing problem* (henceforth indicated as ACVRP).

An important variant of the basic CVRP, which received great attention in the scientific literature, arises when the vehicle fleet is characterized by different capacities and costs. This problem is generally known as *heterogeneous VRP*. A specific survey on heterogeneous VRPs can be found in Baldacci *et al.* [9].

MATHEMATICAL FORMULATIONS OF THE CVRP

In this section, we describe two mathematical formulations of the CVRP that were used in the most successful exact algorithms, namely, the *two-index vehicle flow* and *set partitioning* formulations.

In addition to the notation already introduced, for a subset $S \subseteq V_c$, let $\delta(S)$ be the *cutset* defined by S (i.e., $\delta(S) = \{\{i, j\} \in E : i \in S, j \notin S \text{ or } i \notin S, j \in S\}$). Moreover, we denote by $q(S)$ the total demand of the customers in S and by $\mathcal{S}$ the set $\{S : S \subseteq V_c, |S| \geq 2\}$.

Two-Index Vehicle Flow Formulation

The so-called vehicle flow formulations use binary decision variables to indicate if a vehicle travels between a pair of vertices in G . In *three-index* models there is a specific variable for each vehicle and vertex pair, whereas *two-index* models aggregate this information over the different vehicles.

The *two-index vehicle flow* formulation of the CVRP was originally proposed by Laporte *et al.* [10] and is as follows: Let x_{ij} be an integer variable which may take value in $\{0, 1\}$, $\forall \{i, j\} \in E \setminus \{\{0, j\} : j \in V_c\}$ and value in $\{0, 1, 2\}$, $\forall \{0, j\}$, $j \in V_c$, with $x_{ij} = 1$ when edge $\{i, j\}$ is traveled, and $x_{0j} = 2$ when a route including the single customer j is selected in the solution. Then, the CVRP can be formulated as the following integer linear program (LP):

$$\min \sum_{\{i, j\} \in E} d_{ij} x_{ij} \quad (1)$$

$$\text{s.t.} \quad \sum_{\{i, j\} \in \delta(\{h\})} x_{ij} = 2, \quad h \in V_c, \quad (2)$$

$$\sum_{\{i, j\} \in \delta(S)} x_{ij} \geq 2 \lceil q(S)/Q \rceil, \quad S \in \mathcal{S}, \quad (3)$$

$$\sum_{j \in V_c} x_{0j} = 2m, \quad (4)$$

$$x_{ij} \in \{0, 1\}, \quad \{i, j\} \in E \setminus \{\{0, j\} : j \in V_c\}, \quad (5)$$

$$x_{0j} \in \{0, 1, 2\}, \quad \{0, j\}, j \in V_c. \quad (6)$$

Constraints (2) are the degree constraints for each customer. Constraints (3) are the so-called *Rounded Capacity* (RC) inequality which, for any subset S of customers that does not include the depot, imposes the condition that at least $\lceil q(S)/Q \rceil$ vehicles enter and leave it. These constraints impose both, the capacity and the connectivity constraints. Constraint (4) states that m vehicles must leave and return to the depot, while constraints (5) and (6) are the integrality constraints.

Golden *et al.* [11] formulated the CVRP using three-index binary variables x_{ij}^k as vehicle flow variables to indicate whether vehicle k travels directly from customer i to customer j ($x_{ij}^k = 1$ if vehicle k travels directly from customer i to customer j , 0 otherwise). This formulation is generally used when additional route constraints, such as heterogeneous vehicle fleet and working time constraints, must be imposed.

The two- and three-index vehicle flow models have an exponential number of constraints to enforce connectivity. Several alternative models impose this requirement by using a set of continuous variables representing the flow of one or more “commodities” between the depot and the customers. These formulations were presented by Gavish and Graves [12], and subsequently explored by Gouveia [13] and Baldacci *et al.* [14].

Set Partitioning Formulation

The set partitioning formulation of the CVRP was originally proposed by Balinski and Quandt [15] and associates a binary variable with each *feasible* route.

Let $\mathcal{R}$ be the index set of all feasible routes and let $a_{i\ell}$ be a binary coefficient that is equal to 1 if vertex $i \in V$ belongs to route $\ell \in \mathcal{R}$ and takes the value 0 otherwise (note that $a_{0\ell} = 1, \forall \ell \in \mathcal{R}$). In the following we will use $R_\ell = \{0, i_1, i_2, \dots, i_{h_\ell}\}$ to indicate the subset of

vertices visited by route $\ell \in \mathcal{R}$. Such a route represents the trip of one vehicle leaving the depot, serving the demands of the customers in $R_\ell \setminus \{0\}$, and returning to the depot. Each route $\ell \in \mathcal{R}$ has an associated cost c_ℓ , that is equal to the optimal solution cost of the TSP instance defined by R_ℓ .

Let ξ_ℓ be a binary variable that is equal to 1 if and only if route $\ell \in \mathcal{R}$ belongs to the optimal solution. The set partitioning formulation for the CVRP is as follows:

$$\min \sum_{\ell \in \mathcal{R}} c_\ell \xi_\ell \quad (7)$$

$$\text{s.t.} \quad \sum_{\ell \in \mathcal{R}} a_{i\ell} \xi_\ell = 1, \quad i \in V_c, \quad (8)$$

$$\sum_{\ell \in \mathcal{R}} \xi_\ell = m, \quad (9)$$

$$\xi_\ell \in \{0, 1\}, \quad \ell \in \mathcal{R}. \quad (10)$$

Constraints (8) specify that each customer $i \in V_c$ must be covered by one route and constraint (9) requires that m routes are selected.

This model is valid for any type of cost matrix D . However, if the cost matrix satisfies the triangle inequality, then equality constraint (8) can be written as a ‘ $\geq$ ’ inequality, thus obtaining the so-called *set covering* formulation that preserves the optimal objective function value [16]. The set partitioning formulation is very general and can take into account several route constraints (e.g., time windows, precedences, route lengths) since route feasibility is implicitly considered in the definition of set $\mathcal{R}$.

The set partitioning formulation described above remains valid if the set of routes $\mathcal{R}$ is enlarged with the set $\hat{\mathcal{R}} \supset \mathcal{R}$ containing, in addition to elementary routes, *nonelementary* routes, which are routes in which the vehicle is permitted to visit customers more than once. In this case, coefficient $a_{i\ell}$ is a general integer coefficient that is equal to the number of times customer i is visited by route ℓ . Note that the overall integer programming formulation remains valid, since constraints (8) ensures that the variables representing nonelementary routes will be automatically eliminated when ξ is binary.

Relaxations and Valid Inequalities

Valid lower bounds on the CVRP can be derived from the LP-relaxations of the mathematical formulations described in the previous section. Some of the resulting LPs cannot be solved directly, even for moderate size instances, since either the number of variables or constraints is exponential in the problem size. Thus, the lower bounds are usually computed using cutting-plane and column generation techniques. In addition, to strengthen the relaxations, a variety of valid inequalities have been described in the literature for the different formulations.

Valid inequalities for the vehicle flow formulations have been described by Laporte and Nobert [17], Cornuéjols and Harche [18], Fischetti *et al.* [19], Augerat *et al.* [20], and Letchford *et al.* [21] and a description of some of them for the two-index formulation can be found in Naddef and Rinaldi [22]. Conditions under which some of these inequalities induce facets of the CVRP polytope are discussed in Cornuéjols and Harche [18].

Valid inequalities for the two-index formulation, similar to the RC inequalities, can be obtained according to the way in which the right-hand side of constraints (3) is computed. For example, the *fractional capacity* (FC) inequalities take the form

$$\sum_{\{i,j\} \in \delta(S)} x_{ij} \geq 2 \frac{q(S)}{Q}, \quad S \in \mathcal{S}. \quad (11)$$

Note that, as $q(S)$ can be smaller than Q , the following *subtour elimination* (SE) inequalities are not dominated by the FC inequalities:

$$\sum_{\{i,j\} \in \delta(S)} x_{ij} \geq 2, \quad S \in \mathcal{S}. \quad (12)$$

Obviously, the RC inequalities dominate the FC and SE inequalities. However, the separation problem for RC inequalities was shown to be strongly $\mathcal{NP}$ -hard [22], whereas FC and SE inequalities can be easily separated in polynomial time.

More involved inequalities for the CVRP, such as *generalized capacity*, *framed capacity*, and *hypotour* inequalities were proposed by Augerat *et al.* [20]. Araque *et al.* [23]

defined the *multistar* and *partial multistar* inequalities for the CVRP with unit demands that were generalized by Letchford *et al.* [21] to the CVRP, yielding the so-called *homogeneous* multistar and partial multistar inequalities.

Other classes of valid inequalities have been presented for the polytope associated with the two-index formulation. In particular, some of these valid inequalities are based on the successful results of polyhedral combinatorics developed for the TSP by Chvátal [24] and by Grötschel and Padberg [25,26]. These include, among others, *comb* inequalities and *path-bin* inequalities [22].

Concerning the set partitioning formulation, results on which inequalities are implied by its LP-relaxation have been reported by Baldacci *et al.* [14], Letchford and Salazar González [27], and Baldacci *et al.* [28]. The latter authors have considered other valid inequalities designed for the general set partitioning problem, such as *clique* inequalities.

Systematic studies on how the different formulations proposed for the CVRP are related with each other can be found in Baldacci *et al.* [14] and Letchford and Salazar González [27].

EXACT METHODS FOR THE CVRP

The best-known exact algorithms for the CVRP are briefly surveyed below. These methods can be classified into the following categories: *branch-and-bound*, *dynamic programming*, *branch-and-cut*, *branch-and-cut-and-price*, and *set partitioning-based methods*.

Branch-and-Bound

The effectiveness of branch-and-bound algorithms largely depends on the quality of the lower bounds used to limit the search tree.

Christofides *et al.* [29] described a Lagrangean bound for the CVRP based on the set partitioning formulation where the columns correspond to the set of q -routes. A q -route is not necessarily a simple cycle covering the depot and a subset of customers whose total demand is equal to q . Some customers may be visited more than once,

and thus the set of feasible CVRP routes is strictly contained in the set of q -routes. The resulting branch-and-bound algorithm solved benchmark instances with up to 25 customers. Christofides *et al.* [29] also described a lower bound based on spanning trees, where the depot has degree k , with $m \leq k \leq 2m$.

Fisher [30] described a Lagrangean bound for the CVRP based on m -trees, which are spanning trees having degree $2m$ in the depot, and an exact branch-and-bound algorithm that has optimally solved a well-known problem with 100 customers. Finally, Miller [31] dualized vehicle capacity constraints in a Lagrangean fashion and worked on the resulting b -matching relaxation. The resulting branch-and-bound algorithm solved several benchmark instances containing up to 50 customers.

As to the ACVRP, Laporte *et al.* [32] proposed a branch-and-bound algorithm based on the use of the assignment problem (AP) relaxation that was able to solve randomly generated instances involving some tens of customers and two to four vehicles. Fischetti *et al.* [33] described two new lower bounds based on disjunction on infeasible arc subsets and on min-cost flow that improved on the AP relaxation. The new lower bounds were combined into an effective overall additive bounding procedure, according to the additive approach proposed by Fischetti and Toth [34] that allows for the combination of different lower bounding procedures, each exploiting different substructures of the considered problem. The resulting branch-and-bound approach was able to solve randomly generated instances with up to 300 vertices and 4 vehicles, and real-world instances with up to 70 vertices and 3 vehicles.

Dynamic Programming

DP has been applied to solve several variants of the CVRP or to obtain tight lower bounds.

Christofides *et al.* [35] presented three formulations of the CVRP and introduced the state-space relaxation method for relaxing the DP recursions in order to obtain valid lower bounds on the value of the optimal solutions. The computational results show

that the ratio “lower bound/optimum value” varies between 93.1% and 99.6% when these state-space relaxations are used. Christofides [36] reported that a CVRP instance involving 50 customers has been solved exactly by this approach. Instances involving up to 125 customers were solved within 2% of the optimum in less than 15 min on a CYBER 855 computer.

Branch-and-Cut

The first sophisticated branch-and-cut algorithm based on the two-index formulation for the CVRP was proposed by Augerat *et al.* [20]. The algorithm used a variety of valid inequalities that led to significant improvements in the quality of the lower bound. A detailed description of this work can be found in Naddef and Rinaldi [22].

Based on the original work of Augerat *et al.* [20], other branch-and-cut methods based on the two-index formulation have been recently proposed by Ralphs *et al.* [37] and Lysgaard *et al.* [38]. Ralphs *et al.* [37] described a branch-and-cut algorithm based on the two-index formulation and on the addition of the RC inequalities in a cutting-plane fashion. Lysgaard *et al.* [38] used a variety of valid inequalities, including the RC, framed capacity, strengthened comb, multistar, partial multistar, extended hypotour inequalities, and classical Gomory mixed-integer cuts.

Baldacci *et al.* [14] described a branch-and-cut algorithm based on a two-commodity formulation of the CVRP, where the RC inequalities were used in a cutting-plane fashion to strengthen the lower bound obtained by the LP-relaxation of the two-commodity formulation.

The most important aspect of any branch-and-cut algorithm is designing exact or heuristic algorithms that effectively separate a given fractional point from the convex hull of the integer solutions. Given a (fractional) point x^* , the separation problem associated with a given family $\mathcal{F}$ of valid inequalities consists of finding a member $\alpha x \geq \beta$ of $\mathcal{F}$, such that $\alpha x^* < \beta$. Several heuristic and exact algorithms have been designed for the separation problems associated with the different valid inequalities derived for the

CVRP. A detailed description of some of these algorithms can be found in Augerat *et al.* [20], Naddef and Rinaldi [22], Lysgaard *et al.* [38], and Augerat *et al.* [39].

Valid inequalities which play a crucial role in the development of a cutting-plane algorithm for the CVRP are the RC inequalities. Thus, a number of heuristic algorithms have been designed for their separation [20,38,39] and one of the most simple and effective techniques is the so-called *greedy randomized algorithm* proposed by Augerat *et al.* [20]. Fukasawa *et al.* [40] implemented exact separation procedures for the RC inequalities, based on mixed-integer models. They observed that the heuristic procedures typically decrease the separation times by two orders of magnitude without any significant loss in terms of bound quality. Ralphs *et al.* [37] described a decomposition-based separation methodology for the RC inequalities that takes advantage of the ability to solve small instances of the TSP efficiently.

Letchford *et al.* [21] described separation heuristic algorithms for the multistar and partial multistar inequalities, while separation algorithms for the other classes of inequalities including the comb, framed capacity, and hypotour inequalities, have been described by Augerat *et al.* [20] and Lysgaard *et al.* [38]. The complete set of the separation routines used by Lysgaard *et al.* [38] can be found in the software package CVRPSEP [41], which is publicly available.

Two branching strategies have been used by the different branch-and-cut algorithms for the CVRP proposed in the literature: the *branching on sets* strategy and the *branching on variables* strategy. The branching on sets strategy has been introduced by Augerat *et al.* [20]. This strategy consists of choosing a set S , $S \subseteq V_c$, $|S| > 2$, for which $0 < p(S) < 2$, where $p(S) = \sum_{\{i,j\} \in \delta(S)} x_{ij} - 2\lceil q(S)/Q \rceil$, and creating two subproblems: one by adding the constraint $\sum_{\{i,j\} \in \delta(S)} x_{ij} = 2\lceil q(S)/Q \rceil$ and the other by adding the constraint $\sum_{\{i,j\} \in \delta(S)} x_{ij} \geq 2\lceil q(S)/Q \rceil + 2$. The branching on variables strategy involves the selection of an edge $\{i,j\}$ having a fractional value of x_{ij} , and the generation of two subproblems: one by fixing $x_{ij} = 1$ and the other by fixing $x_{ij} = 0$.

The edge $\{i,j\}$ is selected in such a way that is as close as possible to 0.5 and ties are broken by choosing the edge $\{i,j\}$ having maximum cost d_{ij} .

Branch-and-Cut-and-Price

Fukasawa *et al.* [40] suggested an approach to combine both two-index and set partitioning formulations. In the formulation considered by Fukasawa *et al.* [40], the variables correspond to the set of q -routes [29], while the set of constraints is defined by the set partitioning formulation constraints plus feasible CVRP inequalities designed for the two-index formulation. In addition to the RC inequalities, they also used other valid inequalities, all presented in Lysgaard *et al.* [38]. The family $\mathcal{F}$ of the different valid inequalities designed for the two-index formulation can be expressed in a general form as

$$\sum_{\{i,j\} \in E} \alpha_{ij}^t x_{ij} \geq \beta^t, \quad t \in \mathcal{F}. \quad (13)$$

Since the resulting formulation has an exponential number of both columns and rows, column-and-cut generation procedures for computing the lower bound and a branch-and-cut-and-price algorithm for solving the CVRP were proposed. More precisely, Fukasawa *et al.* [40] considered the following relaxation of the CVRP:

$$\min \sum_{\ell \in \hat{\mathcal{R}}} c_\ell \xi_\ell \quad (14)$$

$$\text{s.t. } \sum_{\ell \in \hat{\mathcal{R}}} a_{i\ell} \xi_\ell = 1, \quad i \in V_c, \quad (15)$$

$$\sum_{\ell \in \hat{\mathcal{R}}} \xi_\ell = m, \quad (16)$$

$$\sum_{\ell \in \hat{\mathcal{R}}} \alpha^t(R_\ell) \xi_\ell \geq \beta^t, \quad t \in \mathcal{F}, \quad (17)$$

$$\xi_\ell \geq 0, \quad \ell \in \hat{\mathcal{R}}, \quad (18)$$

where $\alpha^t(R_\ell) = \sum_{\{i,j\} \in E} \alpha_{ij}^t \eta_{ij}^\ell$ and the set of routes $\hat{\mathcal{R}}$ corresponds to the set of q -routes. The coefficient η_{ij}^ℓ is a general integer coefficient that is equal to the number of times edge $\{i,j\}$ is traversed by q -route ℓ .

Fukasawa *et al.* [40] used a pricing and cut generation technique to solve the LP above. At first, the LP (i.e., *master problem*) includes only the degree constraints (15) and (16) and a restricted set of q -routes. The resulting lower bound has been integrated by Fukasawa *et al.* [40] in an enumeration scheme to solve the CVRP to optimality. The enumeration scheme is based on the branching on sets strategy described in the previous section.

The cut generation is performed by converting a solution ξ of the LP into a corresponding solution x . If this solution is fractional, it is given as input to the CVRPSEP package in order to separate the different classes of valid inequalities. The violated inequalities found are then translated back into the ξ variables to be introduced in the LP.

The pricing subproblem consists of finding q -routes with minimum reduced cost. Although this problem is $\mathcal{NP}$ -hard, it can be solved in pseudopolynomial time. Additional features of the branch-and-cut-and-price algorithm of Fukasawa *et al.* [40] are *cycle elimination*, *heuristic acceleration*, and *dynamic selection of column generation* (see Ref. 40 for details).

Based on the branch-and-cut-and-price algorithm described above, Pessoa *et al.* [42] proposed an exact method for the ACVRP.

Set Partitioning-Based Method

Balinski and Quandt [15] introduced the set partitioning formulation of the CVRP, where each column corresponds to a route. Agarwal *et al.* [43] proposed a column-generation algorithm on a modified set partitioning problem where column costs are given by a linear function over the customers, yielding a lower bound on the actual route cost.

Hadjiconstantinou *et al.* [44] described a method for computing a feasible dual solution of the LP-relaxation of the set partitioning formulation that is based on the computation of k -shortest paths and q -paths. The resulting CVRP lower bound is superior to the lower bound obtained by Christofides *et al.* [29] and the corresponding branch-and-bound algorithm was able to optimally solve benchmark instances involving up to 50 customers.

Mingozi *et al.* [45] presented a new method for solving the set partitioning formulation of the CVRP. They described a procedure for computing a valid lower bound that combines in an additive way different dual heuristics to find a near-optimal dual solution of the LP-relaxation of the set partitioning model. These dual heuristics are based on a bounding method proposed by Christofides *et al.* [29,35]. The dual solution obtained is then used to limit the set of the feasible routes containing the optimal CVRP solutions. The resulting set partitioning problem is solved by using a branch-and-bound algorithm. Mingozi *et al.* [45] report optimal solutions of benchmark instances with up to 50 customers.

Baldacci *et al.* [28] improved the method of Mingozi *et al.* [45] and derived a new exact method for the CVRP based on the following set partitioning model with additional cuts.

Let $H = (\mathcal{R}, \mathcal{E})$ be the *conflict graph* where each vertex corresponds to a route and the edge set $\mathcal{E}$ contains every pair $\{\ell, \ell'\}$, $\forall \ell, \ell' \in \mathcal{R}, \ell < \ell'$, such that $R_\ell \cap R_{\ell'} \neq \emptyset$. Let $\mathcal{C}$ be the set of all cliques of H . The integer model proposed by Baldacci *et al.* [28] is as follows:

$$\min \sum_{\ell \in \mathcal{R}} c_\ell \xi_\ell \quad (19)$$

$$\text{s.t.} \quad \sum_{\ell \in \mathcal{R}} a_{i\ell} \xi_\ell = 1, \quad i \in V_c, \quad (20)$$

$$\sum_{\ell \in \mathcal{R}} \xi_\ell = m, \quad (21)$$

$$\sum_{\ell \in \mathcal{R}(S)} \xi_\ell \geq \lceil q(S)/Q \rceil, \quad S \in \mathcal{S}, \quad (22)$$

$$\sum_{\ell \in \mathcal{R}} \alpha^t(R_\ell) \xi_\ell \geq \beta^t, \quad t \in \mathcal{T}, \quad (23)$$

$$\sum_{\ell \in C} \xi_\ell \leq 1, \quad C \in \mathcal{C}, \quad (24)$$

$$\xi_\ell \in \{0, 1\}, \quad \ell \in \mathcal{R}, \quad (25)$$

where $\mathcal{R}(S) = \{\ell \in \mathcal{R} : R_\ell \cap S \neq \emptyset\}$, $\alpha^t(R_\ell) = \sum_{(i,j) \in E} \alpha_{ij}^t \eta_{ij}^\ell$ and $\eta_{ij}^\ell = 1$ for each edge $\{i, j\}$ covered by route ℓ , 0 otherwise. Constraints (22) and (24) correspond to the *strengthened capacity constraints* and *clique inequalities*, respectively.

Baldacci *et al.* [28] proposed an additive bounding procedure that combines two dual ascent heuristics, called H^1 and H^2 , to derive a near-optimal dual solution that is used by a column-and-cut generation algorithm, called H^3 , to initialize the master problem. H^3 tries to close the integrality gap by adding in a cutting-plane fashion both strengthened capacity and clique constraints. The final dual solution achieved is then used to generate a reduced set partitioning problem containing only the routes whose reduced cost is smaller than the gap between an upper bound and the lower bound obtained. The resulting reduced problem is then solved through an integer linear programming solver.

The key components of this method are (i) the dual ascent scheme used by the bounding procedures H^1 and H^2 , (ii) the use of a state-space relaxation [35] to extend the route set with a relaxation of feasible routes that is easier to compute, and (iii) the use of bounding functions to reduce the state-space graph computed by DP when solving the pricing problem and generating the final set partitioning model.

Recently, Baldacci *et al.* [46] further improved the method given in an earlier study [28]. In particular, they used a new route relaxation, called *ng-route*, that improves other relaxations of feasible routes proposed in the literature for the CVRP and increases the efficiency of the pricing algorithms. They improved procedure H^3 using subset-row inequalities [47] and a novel strategy for solving the pricing subproblem. Finally, a new fathoming criterion based on the dual solution achieved by H^2 is used to speed up the solution of the pricing subproblems and reduce the size of the final set partitioning model.

The method described above provides a general solution framework that can be specialized to solve a number of VRP variants [48].

COMPUTATIONAL EXPERIMENTS FOR THE CVRP

Both analytical and computational results on the CVRP can be found in Baldacci *et al.*

[14], Augerat *et al.* [20], Letchford *et al.* [21], Naddef and Rinaldi [22], Baldacci *et al.* [28], Ralph *et al.* [37], Lysgaard *et al.* [38], Fukasawa *et al.* [40], and Baldacci *et al.* [46]. However, extensive computational results over a common set of instances taken from the literature have been performed only by the last four of these exact methods. Moreover, according to the results reported, the last four methods represent the most effective exact methods currently available for the CVRP, both for the quality of the lower bounds produced and for the number of instances solved to optimality. Thus, in this section, we report a comparison of the results obtained by the exact algorithms given in Baldacci *et al.* [28], Lysgaard *et al.* [38], Fukasawa *et al.* [40], and Baldacci *et al.* [46] on five classes of CVRP instances from the literature, called A, B, E, M, and P. These instances are available at <http://branchandcut.org/VRP/data>. Classes A, B, and P were proposed by Augerat [49]. Instance class M was proposed by Christofides *et al.* [50], and class E was produced by Christofides and Eilon [51].

The algorithm of Lysgaard *et al.* [38] was run on an Intel Celeron 700 MHz, whereas the algorithms of Baldacci *et al.* [28] and Fukasawa *et al.* [40] used Pentium 4 processors with 2.6 GHz and 2.4 GHz, respectively. The algorithm of Baldacci *et al.* [46] was run on an IBM Intel Xeon X7350 Server (2.93 GHz-16 GB of RAM). According to SPEC benchmarks (<http://www.spec.org/benchmarks.html>), the machine used by Baldacci *et al.* [46] is about three times faster than the Pentium 4 2.6 GHz PC used earlier by Baldacci *et al.* [28] and the Pentium 4 2.4 GHz PC of Fukasawa *et al.* [40], and ten times faster than the Intel Celeron 700 MHz PC of Lysgaard *et al.* [38].

Table 1 reports a summary of the computational results obtained by the exact methods given in Baldacci *et al.* [28], Lysgaard *et al.* [38], Fukasawa *et al.* [40], and Baldacci *et al.* [46]. Column *np* of this table reports the total number of instances in the corresponding class. For each method and for each class, the table reports the following data: number of instances solved to optimality (*nopt*), average percentage ratio of the lower bound with respect to the optimal solution value (%LB),

Table 1. Summary Results for the CVRP

	Lysgaard <i>et al.</i> [38] ^a					Fukasawa <i>et al.</i> [40] ^b				Baldacci <i>et al.</i> [28] ^c				Baldacci <i>et al.</i> [46] ^d			
	<i>np</i>	<i>nopt</i>	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	<i>nopt</i>	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	<i>nopt</i>	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	<i>nopt</i>	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}
A	22	15	98.4	29	6,638	22	99.2	183	1,961	22	99.8	41	118	22	99.7	28	31
B	20	19	99.4	27	8,178	20	99.5	84	4,763	20	99.8	119	417	20	99.9	55	67
E-M	9	3	99.1	43	39,592	9	99.0	957	126,987	8	99.6	185	1,025	9	99.8	289	305
P	24	16	98.7	31	11,219	24	99.1	136	2,892	22	99.7	34	187	24	99.8	80	81
All	75	53				75				72				75			
Average			98.9	29	10,439		99.2	234	18,009		99.7	77	323		99.8	83	89

^acomputing time in seconds of an Intel Celeron 700 MHz.^bcomputing time in seconds of a Pentium 4 2.4 GHz.^ccomputing time in seconds of a Pentium 4 2.6 GHz.^dcomputing time in seconds of an IBM Intel Xeon X7350 Server 2.93 GHz.**Table 2. Results on Selected CVRP Instances**

Name	<i>z</i> *	Lysgaard <i>et al.</i> [38] ^a			Fukasawa <i>et al.</i> [40] ^b			Baldacci <i>et al.</i> [28] ^c			Baldacci <i>et al.</i> [46] ^d		
		%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}	%LB	<i>t</i> _{LB}	<i>t</i> _{TOT}
A-n54-k7	1,167	97.3	30	7,246	98.9	125	1,409	99.5	43	86	99.8	14	14
A-n64-k9	1,401	96.5	132	<i>tl</i>	98.9	265	11,254	99.5	32	120	99.6	21	24
A-n80-k10	1,763	97.0	201	<i>tl</i>	99.5	1,120	6,464	99.6	117	194	99.7	247	250
B-n50-k8	1,312	97.6	26	<i>tl</i>	98.7	97	2,845	99.7	640	662	99.8	118	147
B-n66-k9	1,316	98.7	80	<i>tl</i>	99.4	145	1,778	99.5	216	227	99.9	283	292
B-n68-k9	1,272	98.9	65	<i>tl</i>	99.3	260	87,436	99.5	254	6,168	99.7	336	526
E-n51-k5	521	99.6	24	59	99.5	51	65	100.0	13	13	100.0	2	2
E-n76-k7	682	97.7	72	118,683	98.2	264	46,520	99.0	146	3,371	99.8	290	291
E-n76-k8	735	97.7	136	<i>tl</i>	98.8	277	22,891	99.3	104	873	99.9	63	64
E-n76-k10	830	96.4	158	<i>tl</i>	98.5	354	80,722	99.5	60	174	99.5	19	21
E-n76-k14	1,021	95.0	181	<i>tl</i>	98.6	224	48,637	99.6	17	45	99.4	7	9
E-n101-k8	815	98.5	222	<i>tl</i>	98.8	1,068	801,963	99.0	250	—	100.0	638	638
E-n101-k14	1,067	96.2	555	<i>tl</i>	98.8	658	116,284	99.7	154	1,230	99.6	319	453
M-n101-k10	820	100.0	33	33	100.0	119	119	100.0	47	47	100.0	19	19
M-n121-k7	1,034	98.4	979	<i>tl</i>	99.7	5,594	25,678	99.8	944	2,448	99.9	1,247	1,249
P-n50-k8	631	95.4	28	<i>tl</i>	97.7	102	9,272	99.0	12	596	99.1	12	14
P-n55-k10	694	95.4	53	<i>tl</i>	98.2	107	9,076	99.3	16	66	99.3	6	8
P-n70-k10	827	96.2	90	<i>tl</i>	98.5	292	24,039	99.4	35	774	99.5	116	118
P-n76-k5	627	98.5	92	10,970	98.4	273	14,546	98.8	122	—	100.0	438	437
P-n101-k4	681	99.6	127	281	99.6	1,055	1,253	99.4	371	—	100.0	1,154	1,155

^acomputing time in seconds of an Intel Celeron 700 MHz.^bcomputing time in seconds of a Pentium 4 2.4 GHz.^ccomputing time in seconds of a Pentium 4 2.6 GHz.^dcomputing time in seconds of an IBM Intel Xeon X7350 Server 2.93 GHz.

average time in seconds for computing the lower bound (*t*_{LB}) and average total time in seconds (*t*_{TOT}) both computed over all the instances solved to optimality. The last two lines of Table 1 report, for each method, the total number of instances solved to optimality and the averages of lower bounds and computing times computed over all instances solved to optimality by the method.

Table 2 reports a detailed comparison of the methods in Baldacci *et al.* [28], Lysgaard *et al.* [38], Fukasawa *et al.* [40], and Baldacci *et al.* [46] on difficult CVRP instances. Columns %LB, *t*_{LB}, and *t*_{TOT} of this table have the same meaning as in Table 1. Column *z** reports the cost of the optimal solution of each instance. For the exact method of Lysgaard *et al.* [38], “*tl*” indicates that the time limit has been reached, and

for the exact method of Baldacci *et al.* [28], “—” denotes that the memory limit has been reached.

The lower bounds reported in the tables for the methods of Lysgaard *et al.* [38] and Fukasawa *et al.* [40] are relative to the lower bounds computed at the root node of the corresponding decision trees, while the lower bounds of the methods of Baldacci *et al.* [28] and Baldacci *et al.* [46] are relative to the lower bound computed by bounding procedure H^3 .

The tables show that the method of Baldacci *et al.* [28] was not able to solve to optimality three instances solved by Fukasawa *et al.* [40] and by Baldacci *et al.* [46]. Table 1 indicates that the lower bounds of Baldacci *et al.* 28 and 46 are on average superior to the lower bounds of Lysgaard *et al.* [38] and Fukasawa *et al.* [40] in all classes of instances considered. Taking the different computers used into account, Table 1 indicates that the method of Baldacci *et al.* [46] outperforms the other methods.

REFERENCES

1. Dantzig GB, Ramser JH. The truck dispatching problem. *Manage Sci* 1959;6(1):80–91.
2. Laporte G, Nobert Y. Exact algorithms for the vehicle routing problem. *Ann Discrete Math* 1987;31:147–184.
3. Toth P, Vigo D. Volume 9, The vehicle routing problem, SIAM monographs on discrete mathematics and applications. Philadelphia (PA); 2002.
4. Cordeau J-F, Laporte G, Savelsbergh MWP, *et al.* Vehicle routing. In: Barnhart C, Laporte G, editors. Volume 14, Transportation, Handbooks in operations research and management science. Amsterdam: North-Holland Publishing Co.; 2007. pp. 367–428.
5. Baldacci R, Toth P, Vigo D. Recent advances in vehicle routing exact algorithms. *4OR: A Q J Oper Res* 2007;5(4):269–298.
6. Baldacci R, Toth P, Vigo D. Exact algorithms for routing problems under vehicle capacity constraints. *Ann Oper Res* 2009. DOI: 10.1007/s10479-009-0650-0.
7. Laporte G, Semet F. Classical heuristics for the capacitated VRP. In: Toth P, Vigo D, editors. Volume 9, The vehicle routing problem, SIAM monographs on discrete mathematics and applications. Philadelphia (PA); 2002. pp. 109–128.
8. Gendreau M, Laporte G, Potvin J-Y. Metaheuristics for the capacitated VRP. In: Toth P, Vigo D, editors. Volume 9, The vehicle routing problem, SIAM monographs on discrete mathematics and applications. Philadelphia (PA); 2002. pp. 129–154.
9. Baldacci R, Battarra M, Vigo D. Routing a heterogeneous fleet of vehicles. In: Golden BL, Raghavan S, Wasil E, editors. Volume 43, The vehicle routing problem: latest advances and new challenges. Springer; Berlin 2008a. pp. 3–27.
10. Laporte G, Nobert Y, Desrochers M. Optimal routing under capacity and distance restrictions. *Opera Res* 1985;33:1058–1073.
11. Golden BL, Magnanti TL, Nguyen HQ. Implementing vehicle routing algorithms. *Networks* 1977;7:113–148.
12. Gavish B, Graves SC. Scheduling and routing in transportation and distribution systems: formulations and new relaxations. Technical report UBLCS-2004-16. Rochester (NY): Graduate School of Management, University of Rochester; 1982.
13. Gouveia L. A result on projection for the vehicle routing problem. *Eur J Oper Res* 1995;85:610–624.
14. Baldacci R, Hadjiconstantinou E, Mingozzi A. An exact algorithm for the capacitated vehicle routing problem based on a two-commodity network flow formulation. *Oper Res* 2004;52(5):723–738.
15. Balinski M, Quandt R. On an integer program for a delivery problem. *Oper Res* 1964;12:300–304.
16. Bramel J, Simchi-Levi D. Set-covering-based algorithms for the capacitated VRP. In: Toth P, Vigo D, editors. Volume 9, The vehicle routing problem, SIAM monographs on discrete mathematics and applications. Philadelphia (PA); 2002. pp. 85–108.
17. Laporte G, Nobert Y. Comb inequalities for the vehicle routing problem. *Methods Oper Res* 1984;51:271–276.
18. Cornuéjols G, Harche F. Polyhedral study of the capacitated vehicle routing. *Math Program* 1993;60:21–52.
19. Fischetti M, Salazar González JJ, Toth P. Experiments with a multi-commodity formulation for the symmetric capacitated vehicle routing problem. 3rd Meeting of the EURO

- Working Group on Transportation Barcelona, Barcelona, Spain. 1995. pp. 169–173.
20. Augerat P, Belenguer JM, Benavent E, *et al.* Computational results with a branch-and-cut code for the capacitated vehicle routing problem. Technical report 1 RR949-M. Grenoble France: ARTEMIS-IMAG; 1995.
 21. Letchford AN, Eglese RW, Lysgaard J. Multistars, partial multistars and the capacitated vehicle routing problem. *Math Program* 2002;94:21–40.
 22. Naddef D, Rinaldi G. Branch-and-cut algorithms for the capacitated VRP. In: Toth P, Vigo D, editors. Volume 9, The vehicle routing problem, SIAM monographs on discrete mathematics and applications, SIAM (Society for Industrial and Applied Mathematics), Philadelphia (PA); 2002. pp. 53–81.
 23. Araque JR, Hall LA, Magnanti TL. Capacitated trees, capacitated routing and associated polyhedra. Technical Report Discussion Paper. Center for Operations Research and Econometrics, Catholic University of Louvain, Belgium; 1990.
 24. Chvátal V. Edmonds polytopes and weakly Hamiltonian graphs. *Math Program* 1973; 5:29–40.
 25. Grötschel M, Padberg MW. On the symmetric traveling salesman problem: I and II. *Math Program* 1979;16:265–280.
 26. Grötschel M, Padberg MW. Polyhedral theory. In: Lawler EL, Lenstra JK, Rinnooy Kan AHG, *et al.*, editors. The traveling salesman problem: a guided tour of combinatorial optimization. Chichester: John Wiley & Sons, Ltd.; 1985. pp. 231–305.
 27. Letchford AN, Salazar González JJ. Projection results for vehicle routing. *Math Program* 2006;105(2–3):251–274.
 28. Baldacci R, Christofides N, Mingozzi A. An exact algorithm for the vehicle routing problem based on the set partitioning formulation with additional cuts. *Math Program Ser A* 2008b;115(2):351–385.
 29. Christofides N, Mingozzi A, Toth P. Exact algorithms for the vehicle routing problem based on spanning tree and shortest path relaxation. *Math Program* 1981a;10: 255–280.
 30. Fisher ML. Optimal solution of vehicle routing problems using minimum K -trees. *Oper Res* 1994;42:626–642.
 31. Miller DL. A matching-based exact algorithm for capacitated vehicle routing problems. *ORSA J Comput* 1995;7:1–9.
 32. Laporte G, Mercure H, Nobert Y. An exact algorithm for the asymmetrical capacitated vehicle routing problem. *Networks* 1986;16:33–46.
 33. Fischetti M, Toth P, Vigo D. A branch-and-bound algorithm for the capacitated vehicle routing problem on directed graphs. *Oper Res* 1994;42:846–859.
 34. Fischetti M, Toth P. An additive bounding procedure for combinatorial optimization problems. *Oper Res* 1989;37:319–328.
 35. Christofides N, Mingozzi A, Toth P. State space relaxation procedures for the computation of bounds to routing problems. *Networks* 1981b;11:145–164.
 36. Christofides N. Vehicle routing. In: Lawler EL, Lenstra JK, Rinnooy Kan AHG, *et al.*, editors. The traveling salesman problem: a guided tour of combinatorial optimization. Chichester: John Wiley & Sons, Ltd.; 1985. pp. 431–448.
 37. Ralphs TK, Kopman L, Pulleyblank WR, *et al.* On the capacitated vehicle routing problem. *Math Program Ser B* 2003;94:343–359.
 38. Lysgaard J, Letchford AN, Eglese RW. A new branch-and-cut algorithm for the capacitated vehicle routing problem. *Math Program* 2004;100(2):423–445.
 39. Augerat P, Belenguer JM, Benavent E, *et al.* Separating capacity constraints in the CVRP using Tabu search. *Eur J Oper Res* 1998;106:546–557.
 40. Fukasawa R, Longo H, Lysgaard J, *et al.* Robust branch-and-cut-and-price for the capacitated vehicle routing problem. *Math Program Ser A* 2006;106:491–511.
 41. Lysgaard J. CVRPSEP: a package of separation routines for the capacitated vehicle routing problem. Technical report. Department of Management Science and Logistics, Aarhus School of Business; Aarhus, Denmark 2003.
 42. Pessoa A, Poggi de Aragão M, Uchoa E. Robust branch-cut-and-price algorithms for vehicle routing problems. In: Golden BL, Raghavan S, Wasil E, editors. Volume 43, The vehicle routing problem: latest advances and new challenges. Springer; Berlin 2008. pp. 297–325.
 43. Agarwal Y, Mathur K, Salkin HM. A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks* 1989;19:731–749.
 44. Hadjiconstantinou E, Christofides N, Mingozzi A. A new exact algorithm for the vehicle

- routing problem based on q -paths and K -shortest paths relaxations. *Ann Oper Res* 1995;61:21–43.
45. Mingozzi A, Christofides N, Hadjiconstantinou E. An exact algorithm for the vehicle routing problem based on the set partitioning formulation. Technical report. University of Bologna; Bologna, Italy 1994.
 46. Baldacci R, Mingozzi A, Roberti R. New route relaxation and pricing strategies for the vehicle routing problem. 2010a. Submitted.
 47. Jepsen M, Petersen B, Spoorendonk S, *et al.* Subset-row inequalities applied to the vehicle-routing problem with time windows. *Oper Res* 2008;56(2):497–511.
 48. Baldacci R, Bartolini E, Mingozzi A, *et al.* An exact solution framework for a broad class of vehicle routing problems. *Comput Manage Sci* 2010b. DOI: 10.1007/s10287-009-0118-3.
 49. Augerat P. Approche polyédrale du problème de tournées de véhicules [PhD thesis]. Institut National Polytechnique de Grenoble; 1995.
 50. Christofides N, Mingozzi A, Toth P. The vehicle routing problem. In: Christofides N, Mingozzi A, Toth P, *et al.*, editors. *Combinatorial optimization*. John Wiley & Sons, Ltd.; Chichester 1979. Chapter 11. pp. 315–338.
 51. Christofides N, Eilon S. An algorithm for the vehicle dispatching problem. *Oper Res Q* 1969;20:309–318.

FAIRNESS AND EQUITY IN SOCIETAL DECISION ANALYSIS

L. ROBIN KELLER

YITONG WANG

The Paul Merage School of Business,
University of California Irvine,
Irvine, California

TIANJUN FENG

Fudan University, School of
Management, Shanghai, China

Imagine you divide three cookies between your two children, five-year-old Jack and two-year-old Jill. You might give Jack two cookies and Jill one cookie. But, then Jill will claim that this division is unfair, and will remain unconvinced even when you say that Jack is older and bigger and thus needs (and deserves) more food.

In such a parental role, you may decide to give equal cookie allocations in the future, ignoring arguably good reasons for giving an unequal allocation. Such an equal allocation is particularly appealing when the recipients all know what each person was given and have the opportunity to express their feelings about their allocation. But, sometimes, an unequal distribution will seem to be the most fair and equitable, given the special features of the situation. For example, whenever there is a single undividable item, such as a family heirloom, which needs to be somehow allocated, the resulting distribution will be unequal, but a fair process may be used to determine the allocation.

Just as it is important to consider fairness in making allocations within a family, fairness plays an important role in making societal decisions, such as where to locate hazardous facilities or how to apportion benefits and costs among different parties. In the operations research literature, fairness and equity are often used interchangeably, usually referring to perceptions of the appropriateness of allocations of risks, benefits, and costs (or of the allocation process). A variety of social science domains investigate

such allocations, and use a variety of terms to define specific types of fairness (such as *procedural* justice). The aim of this article is to focus primarily on how decision analysts and behavioral decision theorists (within the operations research community) have examined fairness, and provide some introduction and references to coverage in other disciplines. The section titled “Levels of Fairness Investigations” describes different levels of fairness investigations. The section titled “Fairness Factors and Issues to Consider in Societal Decision Models” presents fairness factors and issues to consider in constructing societal decision models. And the section titled “Decision Scenarios Involving Fairness and Experimental Results” presents some examples of decision scenarios involving fairness and experimental results for those scenarios.

LEVELS OF FAIRNESS INVESTIGATIONS

Fairness or equity can be examined from different perspectives. We present some of these below and provide some very brief examples of ways of looking at fairness at each of these levels.

To an Individual

Is it fair to pay a person less because he is single and has no family responsibilities? Is it fair to give priority to admitting a college student from a historically underrepresented racial or ethnic group? Is it fair to charge a customer more for an umbrella on the day it rains? Is it fair to pay a firefighter more than an office manager, since the firefighter faces more job safety risks?

Kahneman *et al.* [1,2] discuss different conditions under which an allocation of a cost or a benefit to an individual is seen as fair, and report results of a survey of the general (Canadian) public about scenarios such as fair pay and fair prices. In a laboratory experiment, Mellers [3] examines what fairness principles appear to be used by

experimental subjects in allocating salaries or taxes. When focusing on fairness to an individual, there still are usually implicit comparisons with what others in similar situations will receive or comparisons with the process that is used with others.

Within Families

How can heirs fairly decide which heir should get the sentimentally important stamp collection? During a divorce case, does the wife get a better deal than the husband, or is the settlement fair?

Pratt and Zeckhauser [4] describe how to fairly divide a number of silver heirlooms among heirs, where each heir's preferences for objects are represented by his or her utility function, then allocations to different heirs are made as they would be in a market for probability shares in the items, assuming each heir has an equal budget for making purchases and the item prices are the market-clearing equilibrium prices in a second-price auction.

In a married couple, each person brings some inputs to their marriage and receives benefits from the marriage, in some proportion to their inputs. The Walster *et al.* [5] book on equity theory addresses issues such as equity within a marriage (and in other settings).

Between Interested Entities/Parties

How fair is the \$2/h raise the university gave its lowest paid workers and the salary cut it gave other workers? How much of a movie's revenue should go to the actors versus the producers?

In cases involving disputes over revenue sharing between joint enterprise partners or negotiations in real estate transactions, involved parties (i.e., people, business, trade organizations, etc.) care about the fairness of the process and the eventual allocation. There is a large body of mathematical research dealing with this type of situation, which is called the *Cake-cutting problem* [6,7]. In this problem, "cake" could mean any desirable goods or benefits and "cutting" means the methods people use to divide them. The basic question—how to fairly cut

a cake—represents many issues involving multiple interested groups that want to share an entity fairly. Robertson and Webb [6] focus on three main issues in their mathematical survey of the cake-cutting problem: (i) the existence of a solution; (ii) the procedure to construct the solution; and (iii) what are the properties of the procedure to the solution. Brams and Taylor [7] address fair division from the cake-cutting problem to dispute resolution. Also see Raiffa's [8] book on negotiation methods and Camerer and Thaler [9] for related game theoretic experiments and results. Hoffman and Spitzer [10] examined participants' concepts of distributive justice in a laboratory experiment and Margolis' [11] book addresses how one person allocates outcomes to another in either a selfish or an altruistic way.

Within an Organization

Considering how much work I did on a project, did I get a fair amount of recognition for my contribution? How fair is the personnel process at my company? How fairly did my boss treat me during the last round of bonus allocations?

Organizational behavior researchers examine workplace fairness in terms of distributive justice, procedural justice, and interactional justice, and then examine the effects of perceived justice on behaviors, such as workplace revenge. Distributive justice refers to whether the outcome in a case is seen as fair, given the inputs [12,13]. Procedural justice focuses on the process by which outcomes are distributed [14]. Interactional justice is concerned primarily with how people are treated [15]. When observers within an organization see how others are treated they may want to believe that people get what they deserve and that there is a "just world" [16,17].

Comparing Within a Region or Nation

How fair is the distribution of land ownership or wealth in a country? How should we examine fairness of electric utility pricing within a region or nation? Do underprivileged groups face more environmental risks?

Rawls [18] discussed the theoretical notion of imagining oneself behind a veil of ignorance about which location in a society would ultimately be held by that person. Then, a person can consider what would be a fair distribution across the society, in ignorance of whether that person's ultimate fate would end up as being in a rich or a poor position. He then proposed that in the "original position" behind the veil a person would adopt a maximin strategy, which would maximize the position of the least well-off person (who has the minimum allocation). Economist John Harsanyi had a similar discussion in his work on the social welfare theory of utilitarianism. Suppes [19] also examined the distributive justice of income inequality. Zajac [20,21] examined fairness in public utility regulations. Zimmerman and colleagues [22–25] empirically examined the fairness of actual distributions of environmental risks, particularly to people in poorer areas, calling this field "environmental justice."

Comparing People Across Nations

How far apart are the wealthy and the poor in one country versus another? Do the same concepts of fairness apply in the United States and in China?

One can graph the distribution of family income within one country and compare it with the distribution in other countries. For a group of N families, the graph called a *Lorenz curve* plots the cumulative family income for the poorest x families on the y -axis against the number of families arranged on the x -axis from the poorest family (#1) to the richest family (# N). If all N families have the same income, which is presumably most fair, then the Lorenz curve is the straight 45° line. If many families have low incomes and only a few have high incomes, the Lorenz curve is convex and falls far below the straight 45° line. A popular way to make a comparison of the distribution in different countries is to use the Gini index, which is calculated from the Lorenz curve by computing the ratio of the area between a country's Lorenz curve and the 45° line divided by the entire triangular area under the 45° line. Therefore, a higher Gini index indicates more inequality. The scores

for different countries and an explanation of the calculation of the index are at <https://www.cia.gov/library/publications/the-world-factbook/fields/2172.html>.

There are differences in fairness perceptions between cultures. In some countries, the closest word to the English word fairness may be justice or balance. Interestingly, in the Canadian public survey [2], respondents answered on a bipolar scale, anchored by acceptable and unfair. For many in Western countries, these two concepts may indeed be two ends of a single scale. Certainly, when a child complains about unfairness in cookie allocations, it usually also means the allocation is unacceptable. However, from some other perspectives, such as those of Chinese, there may be two different dimensions: unfair to fair versus unacceptable to acceptable; see Bian and Keller [26, p. 317]. Bian and Keller found that Chinese [27] agreed with Americans [28] on fairness judgments involving societal risks, but disagreed on the best actions to take. In some cases, where there was a chance that a large group would die, the Americans would take the action that they saw as fair, but Chinese would not. In interviews with Chinese [27], it appeared that when the cultural value of treating everyone equally was in conflict with the value of preventing the extinction of the whole community, Chinese would take actions to avoid extinction of the community; see also Keller *et al.* [29] for additional details on these studies on "risk equity," which involved scenarios with probability distributions over health or safety outcomes.

FAIRNESS FACTORS AND ISSUES TO CONSIDER IN SOCIETAL DECISION MODELS

When I ask you if a particular allocation or transaction process is fair, you will likely respond, "it depends." Perceptions of fairness can depend on a variety of current and preexisting factors and issues. Fairness is also in the "eye of the beholder," so a person from one culture or one point of view may see unfairness when another person sees fairness. In the cookie example, Jack may see it as fair that he, the older child, gets two cookies, but Jill sees it as unfair.

While a parent may use an equal division of cookies among children, in many cases equal distribution is not feasible or prudent. A good introduction to different concepts of fair division is in Yaari and Bar-Hillel's [30] paper on dividing justly. They identified six categories of factors which "provide a possible justification for departure from equality," including the following:

1. differences in needs;
2. differences in tastes, or in the capacity to enjoy various goods;
3. differences in beliefs;
4. differences in endowments;
5. differences in effort, in productivity, or in contribution;
6. differences in rights or in legitimate claims.

When you are in charge of a societal decision analysis, which involves different outcomes (such as monetary benefits or costs, health and safety risks, etc.) to different people, you may want to ask yourself a number of questions to determine which factors should be considered in your decision process. Examples of such societal decisions are selecting a hazardous waste processing site in a region, choosing where to send emergency personnel in a major disaster, and budget allocation between elementary and middle schools. A few of the possible questions to ask before constructing your decision analysis model and process are listed below.

From the Perspective of an Outside Society Member, Does it Matter What the Different Parties Get?

Suppose an outsider does not receive any part of an allocation. Would that person care how the allocation is divided among those who receive an allocation? In such a case, the societal decision maker may wish to spread the allocation more fairly even if those affected do not know or do not care about what others get. See Lichtenstein *et al.* [31] and Wagenaar *et al.* [32] for discussions of the perspective of the societal decision maker and for some stylized societal decision scenarios. One thing that may matter to the

public or the decision maker is if the decision maker takes an action to give people unequal allocations or if the decision maker passively allows already occurring inequalities to remain.

Do the Different Affected Parties Care What Others Get?

Direct observation of others (such as watching children like Jack and Jill) or self-introspection will likely lead to the answer "yes." Yet, simple economic models often assume that a person merely cares about his own allocation. This makes it appear that economists tend not to be concerned about a person being selfish (since that is what such models would expect to occur) and tend to be baffled that a person would be altruistic. Indeed, they may work hard to explain altruistic behavior within the scope of such models by assuming that the altruism somehow helps the individual. For a different point of view, see the books by Margolis [11] on *Selfishness, Altruism, and Rationality* and Robert Frank [33] on *Choosing the Right Pond* for economic discussions in which people care about what others get.

Do the Different Parties Know What Others Get? Do They Know What Different Parties Started Out with or Did to "Deserve" Their Outcomes?

Often, fairness disputes may develop because affected parties do not know all the historical or preceding factors that may support an unequal allocation to one person in comparison with another. A raise may make up for not getting a raise in the last round of merit reviews, for example. See Mellers and Baron's edited book [34], *Psychological Perspectives on Justice*, for discussion of various components of fairness perception.

If the different parties care about what others get, and know what others get, then you may want your decision analysis model to keep track of allocation differences between neighbors or others who compare themselves, since such differences might be important. In some cases, you may choose to reveal the

allocation process and outcomes for all parties, even if the parties would not normally have been able to observe these.

Do the Different Parties Prefer that They Get the Same Outcome at the Same Time or, in Contrast, Do They Prefer That They Get the Outcome at Different Times?

The concept of common fate equity refers to situations where a set of people would like to experience the same “common” outcome simultaneously, even if it is a bad outcome. They prefer sharing a common fate over experiencing common fate inequity, where they would experience a bad (or good) outcome separately.

Parents of young children may choose to not fly together, so that they do not share the common fate of both dying together in the same airplane crash. Companies often have rules that top executives cannot fly together for a similar reason. In such cases, these rules avoid the parents or executives experiencing the common fate of dying from the same risky event. Sadly, such common fate events, though rare, do occur, such as the crash of the presidential plane carrying President Lech Kaczynski of Poland and 95 others including his wife, lawmakers, military commanders, priests, and others on April 10, 2010.

An axiom of preference for *common fate inequity* captures the preference to avoid common fates in decision analysis models. In other situations, a common fate will be preferred. See Fishburn’s [35] common fate equity axiom E4 and a newer treatment of a preference for “shared destinies” in Ref. 36. A decision scenario illustrating common fate preference among experiment participants [27,28] is in the section titled “Decision Scenarios Involving Fairness and Experimental Results” below.

Is It Possible for Parties to Trade Their Allocations to Achieve Better Outcomes?

Sometimes it is possible to adjust allocations among parties enough that they will see the final allocation as fair. One concept of fairness in this case is an *envy-free* allocation, one where no one envies what any other person has, and would not want to switch his own

allocation for another’s allocation. Keller and Sarin [28] present an envy-free risk-benefit model. However, there may be no envy-free allocation possible. In such cases, the societal decision maker must make difficult trade-offs, especially when health and safety risks are involved and there are shortages of life-saving resources (such as donated kidneys for kidney transplant surgeries). Calabresi and Bobbitt [37] identify four approaches in such “tragic choices”: pure market, accountable political, lottery (such as a military draft), and customary or evolutionary mechanisms.

Should Fairness to Future Generations be Considered?

When societal decisions may have effects on future generations, the fairness of the outcomes to those currently living compared to future generations may need to be considered. When called for, economists and decision analysts build considerations of intergenerational effects into their models. For example, Morton *et al.* [38] created a multiple objective decision analysis model to make recommendations to the UK government on what should be done with accumulating radioactive waste. One of their objectives was reducing “burden on future generations,” which was an influential objective when the decision of deep disposal of waste would be the preferred option over other storage options.

DECISION SCENARIOS INVOLVING FAIRNESS AND EXPERIMENTAL RESULTS

In this section, we give examples of decision scenarios involving allocations between people (of healthy foods or safety risks) and show what survey respondents chose.

Healthy Fruit Allocation Scenario

Yaari and Bar-Hillel [30] conducted an experimental survey to examine what choices participants would make in a number of scenarios where allocations across people would be made. They then examined the predominant choice pattern to see how it aligns with different concepts of fairness. Here is their first question, from Ref. 30:

A shipment containing 12 grapefruit and 12 avocados is to be distributed between Jones and Smith. The following information is given, and is also known to the two recipients:

- Doctors have determined that Jones's metabolism is such that his body derives 100 ml of vitamin F from each grapefruit consumed, while it derives no vitamin F whatsoever from avocado.
- Doctors have also determined that Smith's metabolism is such that his body derives 50 ml of vitamin F from each grapefruit consumed and also from each avocado consumed.
- Both persons, Jones and Smith, are interested in the consumption of grapefruit and/or avocados only insofar as such consumption provides vitamin F—and the more the better. All the other traits of the two fruits (such as taste, calorie content, etc.) are of no consequence to them.
- No trades can be made after the division takes place.

How should the fruits be divided between Jones and Smith if the division is to be just?

A large majority (82%) of Yaari and Bar-Hillel's 163 young adult respondents chose for Jones to have 8 grapefruits and 0 avocados (remembering that avocados did not benefit him at all) and for Smith to have 4 grapefruits and 12 avocados. Only 8% of the participants chose an equal division of six grapefruits and six avocados to each person. Therefore, in this case, the number of milligrammes of vitamin F [8×100 for Jones and $(4 \times 50) + (12 \times 50)$ for Smith] were equated on the basis of how much each person needed to eat to get the vitamin benefit, rather than the number of fruits.

Subsequent similar scenarios in Ref. 30 involve variations of needs, tastes, and beliefs. Proceeding through these examples, many subtle variations in fairness judgments are demonstrated. In societal decision analyses potentially involving similar subtle variations in fairness, it is important for decision models to keep track of such factors that would be important to the members of society for whom the decision is to be made.

Rescuer at Risk Scenario

Next, we consider the "Rescuer at Risk" scenario from an experimental survey conducted by Keller and Sarin [28] with American respondents and by Bian and Keller [27] with Chinese respondents:

On an island within your jurisdiction, 100 miners are trapped in one location in a mine. There is a way to rescue these miners by sending a rescue team through an unused tunnel. You have dispatched a rescue team of 10 rescuers to this tunnel. The team has come upon a portion of the tunnel that is dangerous. They need to station a rescuer at this point in the tunnel for the next 10 h to listen and watch for any signs that the trapped miners send to the team. However, there is a chance that, sometime in the next 10 h, a cave-in will occur, which will be fatal to the rescuer stationed there. The rest of the tunnel is safe, so the rescuers are not at a risk in other parts of the tunnel. The team is able to communicate with you at a command post via a portable radio. The team has contacted you for your orders about what to do next. They want to know if they should station one rescuer at the key point in the tunnel for 10 h or have each rescuer take a 1-h shift. There is a 10% chance that a cave-in will occur, and only one cave-in would occur, if any. The rescuers will definitely be able to save the 100 miners, no matter which option is taken.

Option 1.

One rescuer does the entire 10-h shift.
This rescuer has a 10% chance of death.
The other nine rescuers have a 0% chance of death.

Option 2.

Each of the 10 rescuers takes a 1-h shift.

Thus, each rescuer has a 1% chance of death because each would be in the tunnel one-tenth of the time, and one-tenth times 10% is 1%.

This scenario examined the *ex ante* fairness for the rescuers in doing their jobs, before the risk outcome (possible death of a rescuer) has the uncertainty resolved. Among respondents asked to indicate the fairer option, large majorities felt that Option 2, where each rescuer faced 1 h of risk, was fairer

(96% of 53 American adults, 96% of 51 young Chinese adults, and 82% of 49 middle-aged Chinese adults). For those who were asked which option would be chosen (not which is fairer), relatively large majorities chose Option 2 (83% of 53 Americans, 84% of 47 young Chinese adults, and 62% of 49 middle-aged Chinese adults).

For societal decision makers, it is often important to consider the *ex ante* risk to members of society, because society members may prefer a more equal distribution of *ex ante* risk, if feasible; see Keller *et al.* [29] for an example of a decision model that incorporates a preference for *ex ante* equity.

Miner Location Scenario

Now consider the “Miner Location” scenario from Keller and Sarin [28] with American respondents and by Bian and Keller [27] with Chinese respondents.

On an island within your jurisdiction, 100 miners are trapped, 50 in location A and 50 in location B. Two rescue options are possible.

Option 1.

Attempt to rescue all the miners in both locations.

The possible outcomes are

50% chance that none die (because the rescue operation is successful);

50% chance that all 100 die (because the rescue operation is not successful).

(Considered more fair by most Americans and Chinese, chosen as the action to take by most Americans and a weak majority of young Chinese adults)

Option 2.

Attempt to rescue only the miners in one location.

The possible outcomes are

50% chance that the 50 miners in location A live and the 50 miners in location B die because the rescue operation is sent to location A;

50% chance that the 50 miners in location B live and the 50 miners in location A die because the rescue operation is sent to location B.

(Chosen as the action to take by a weak majority of middle-aged Chinese)

In Option 1, the miners all experience the same common fate, since either all 100 would die or all would live, so this is an *ex post* equitable option, in the sense of Fishburn’s [35] common fate axiom. In the survey, a strong majority of American adults and a weak majority of young Chinese adults chose Option 1, where the miners all experience the same common fate. In contrast, a weak majority of middle-aged Chinese adults chose Option 2, which results in common fate inequity, but avoids the catastrophe of all 100 miners dying. A possible explanation for this difference in choices is that the traditional Chinese culture places more emphasis on protecting the group, as expressed in the cultural value of collectivism. (In contrast, the American culture places more emphasis on individualism.) Therefore, the middle-aged Chinese who chose Option 2 avoided the possible loss of the entire group.

For societal decision makers, it is sometimes important to consider the *ex post* risk to members of society; see Keller *et al.* [29] for an example of a decision model incorporating a preference for *ex post* equity. As illustrated in this common fate scenario, models may need to be flexible enough to be consistent with preference for equity or inequity, depending on specific contextual or cultural factors.

CONCLUSION

This article has provided a brief theoretical and empirical introduction to the examination of fairness and equity in societal decision analysis. For a more advanced discussion of “risk equity,” which considers risk in scenarios with probability distributions over health or safety outcomes, see Keller *et al.* [29].

In conclusion, societal decision makers need to consider the fairness of their allocations and their allocation process. In some situations, they may want to reason normatively about which fairness principles to follow. In others, they may wish to consult with stakeholders about perceptions of fairness or

gather experimental survey data to confirm what the affected population would prefer.

REFERENCES

1. Kahneman D, Knetsch J, Thaler R. Fairness and the assumptions of economics. *J Bus* 1986;59(4, part 2):S285–S300.
2. Kahneman D, Knetsch J, Thaler R. Fairness as a constraint on profit-seeking: entitlements in the market. *Am Econ Rev* 1986;76(4):728–741.
3. Mellers BA. “Fair” allocations of salaries and taxes. *J Exp Psychol Hum Percept Perform* 1986;12:80–91.
4. Pratt JW, Zeckhauser RJ. The fair and efficient division of the Winsor family silver. *Manage Sci* 1990;36(11):1293–1301.
5. Walster E, Walster GW, Berscheid E, *et al.* Equity: theory and research. Boston (MA): Allyn and Bacon; 1978.
6. Robertson J, Webb W. Cake-cutting algorithms: be fair if you can. Natick (MA): A K Peters; 1998.
7. Brams SJ, Taylor AD. Fair division: from cake-cutting to dispute resolution. Cambridge: Cambridge University Press; 1995.
8. Raiffa H. Art and science of negotiation. Cambridge (MA): The Belknap Press of Harvard University Press; 1982.
9. Camerer C, Thaler RH. Anomalies: ultimatums, dictators, and manners. *J Econ Perspect* 1995;9:209–219.
10. Hoffman E, Spitzer M. Entitlement, rights, and fairness: an experimental examination of subjects’ concepts of distributive justice. *J Leg Stud* 1985;14(2):259–297.
11. Margolis H. Selfishness, altruism, and rationality: theory of social choice. Chicago (IL): Cambridge University Press; 1982.
12. Adams JS. Inequity in social exchange. In: Berkowitz L, editors. *Advances in experimental social psychology*. New York: Academic Press; 1965. pp. 267–299.
13. Homans GC. Social exchange: its elementary forms. New York: Harcourt Brace Jovanovich; 1974.
14. Thibaut J, Walker L. Procedural justice: a psychological analysis. Hillsdale (NJ): Lawrence Erlbaum Associates; 1975.
15. Bies RJ, Moag JS. Interactional justice: communication criteria of fairness. In: Lewicki RJ, Sheppard BH, Bazerman MH, editors. *Research on negotiation in organizations*. Greenwich (CT): JAI Press; 1986. pp. 43–55.
16. Lerner MJ. Observer’s evaluation of a victim: justice, guilt, and veridical perception. *J Pers Soc Psychol* 1971;20:127–135.
17. Lerner MJ, Miller DT. Just world research and the attribution process: looking back and ahead. *Psychol Bull* 1978;85:1030–1051.
18. Rawls J. A theory of justice. Cambridge (MA): Belknap Press; 1971.
19. Suppes P. The distributive justice of income inequality. *Erkenntnis* 1977;11(2):233–250.
20. Zajac EE. Fairness or efficiency: an introduction to public utility pricing. Cambridge (MA): Ballinger Publishing Co., A Subsidiary of Harper & Row; 1978.
21. Zajac EE. Perceived economic justice: the example of public utility regulation. In: Peyton Young H, editor. *Cost allocation: methods, principles, applications*. North-Holland: Elsevier; 1985. Chapter 7.
22. Sexton K, Zimmerman R. The emerging role of environmental justice in decision making. In: Sexton K, Marcus AA, Easter W, *et al.*, editors. *Better environmental decisions: strategies for government, businesses and communities*. Washington (DC): Island Press; 1999. pp. 419–444.
23. Restrepo CE, Zimmerman R. In: Everitt B, Melnick E, editors. *Environmental Justice: Encyclopedia of Quantitative Risk Assessment*. New York: John Wiley Publishers; 2008.
24. Zimmerman R. Environmental justice. In: Molak V, editor. *Fundamentals of risk analysis and risk management*. Boca Raton (FL): CRC Press; 1997. pp. 281–291.
25. Naphtali ZS, Restrepo CE, Zimmerman R. Maps expand asthma hazards awareness: GIS helps policy makers see where childhood asthma, schools, and pollution sources collide. *Healthy GIS*. ESRI; 2008. pp. 4–5. <http://www.esri.com/library/newsletters/healthygis/-winter 2008.pdf>. Accessed winter 2008.
26. Bian WQ, Keller LR. Patterns of fairness judgments in north america and the people’s republic of China. *J Consum Psychol* 1999;8(3):301–320.
27. Bian WQ, Keller LR. Chinese and Americans agree on what is fair, but disagree on what is best in societal decisions affecting health and safety risks. *Risk Anal* 1999;19:439–452.
28. Keller LR, Sarin RK. Equity in social risk: some empirical observations. *Risk Anal* 1988;8(1):135–146.

29. Keller LR, Feng T, Wang Y. Measures of risk equity. *Encyclopedia of operations research and management science*. Hoboken (NJ): Wiley; 2010. In press.
30. Yaari ME, Bar-Hillel M. On dividing justly. *Soc Choice Welfare* 1984;1:1–24.
31. Lichtenstein S, Gregory R, Slovic P, *et al.* When lives are in your hands: dilemmas of the societal decision maker. In: Hogarth RM, editor. *Insights in decision making: a tribute to Hillel J. Einhorn*. Chicago (IL): University of Chicago Press; 1990. pp. 91–106.
32. Wagenaar WA, Keren G, Lichtenstein S. Islanders and hostages: deep and surface structures of decision problems. *Acta Psychol* 1988;67:175–189.
33. Frank RH. *Choosing the Right Pond: Human behavior and the quest for status*. New York: Oxford University Press; 1985.
34. Mellers B, Baron J, editors. *Psychological perspectives on justice: theory and applications*. New York: Cambridge University Press; 1993.
35. Fishburn PC. Equity axioms for public risks. *Oper Res* 1984;32(4):901–908.
36. Gajdos T, Weymark JA, Zoli C. Shared destinies and the measurement of social risk equity. *Ann Oper Res*. 2010;176(1):409–424.
37. Calabresi G, Bobbitt P. *Tragic choices*. New York: Norton; 1978.
38. Morton A, Airolidi M, Phillips L. Nuclear risk management on stage: the UK's Committee on Radioactive Waste Management. *Risk Anal* 2009;29(5):764–779.

FAST AND FRUGAL HEURISTICS

SVEN BERTEL

ALEX KIRLIK

Human Factors Division and Beckman
Institute for Advanced Science and
Technology, University of Illinois at
Urbana-Champaign, Urbana, Illinois

Cognitive heuristics are fundamental to human judgment and decision making. Often described as rules of thumb that are used in lieu of optimal procedures (e.g., when resources are limited, or knowledge is incomplete), cognitive heuristics may also be understood as essential psychological mechanisms that guide information search and produce decisions by effectively and efficiently exploiting information structures in the environment. This article describes a selection of simple but powerful, task-specific, fast and frugal heuristics (FFHs) as suggested by Gigerenzer, Todd, and the ABC Research Group [1]. Underlying assumptions related to concepts of ecological rationality are presented, as are basic mechanisms of, and research into the validity of FFHs.

Heuristics as a term has been used to designate a number of radically different concepts across the ages and disciplines. The modern English term is ultimately derived from the Ancient Greek *εὑρίσκειν* (“to find”), perhaps more widely known by the exclamation generally attributed to Archimedes of Syracuse after realizing, upon stepping in his bath and inspecting the amount of water displaced in the act, the basic principle of buoyancy: *Eureka!* (“I found it!”).

HEURISTICS AS PROCEDURES FOR PROBLEM SOLVING

In early modern times [2], *heuristics* came to describe the study of procedures and techniques that lead to discovering solutions to previously unsolved problems and to scientific discovery in general [3]. Descartes’ *Discours de la méthode* can be seen as an early

work in this tradition with its reduction of difficult problems into as many parts as is feasible and necessary to recognize their individual truths directly by human intuition [4]. Heuristic procedures often comprise an element of induction: Gigerenzer and Brighton [5, p. 108] provide a good example of a distinction made by George Polya between heuristics on the one hand and analytical methods on the other: while “heuristics are indispensable for finding a proof, [...] analysis is required to check a proof’s validity”.

The original reading of heuristics is still to be found in the psychological (e.g., Gestalt) lingua of the first half of the twentieth century. For example, with regard to human problem solving, Duncker defined heuristic methods of thinking as general procedures for which an “insistent” analyses of the situation, especially the endeavor to vary appropriate elements meaningfully sub specie of the goal, must belong to the essential nature of a solution through thinking” [6, p. 20f]. It is in Herbert Simon’s work [7,8] that we find classic functional descriptions of human reliance on heuristics, for instance, with regard to problem solving, along with explanations of what causes this reliance: According to his view of the matter, people usually do not maximize (i.e., they do not search for a problem’s optimal solution); rather, they *satisfice* (i.e., look for a solution that is deemed good enough). This satisficing behavior is caused by, internal factors such as cognitive limitations, as well as by external factors, such as in the environment and the problem structure, which may, for example, induce prohibitively high computational (or other) costs for actually finding and verifying an optimal solution.

It is also in this original reading of the term that heuristics came to be used in other fields, including computing. In recursive algorithm design, divide-and-conquer algorithmic paradigms, which include many standard data sorting and search algorithms, follow a similar pattern: a computational problem is broken down recursively until the individual parts can be solved easily enough; the

partial solutions are then combined into a global solution for the problem as a whole. The fields of computing in general and of artificial intelligence in particular are shaped by searches for good heuristics, which make computational problems that often are NP-hard in their general form (e.g., the traveling salesman problem, chess, etc.) computationally tractable for some relevant practical contexts; for example, by reformulating the problem representation (such as by making it more specific) or by reformulating the processes that operate on this representation (e.g., by defining optimality criteria locally rather than globally).

HEURISTICS AS PROCEDURES RESULTING FROM TRADE-OFFS BETWEEN EFFORT AND ACCURACY

Goldstein and Gigerenzer [9] point us toward the eventual changes in the use of heuristics that occurred in the late twentieth century. Heuristic procedures were then often contrasted with optimal procedures, with the former being seen simply as the poor man's version of the latter. Heuristics had a place (only) when the corresponding optimal procedures were unavailable; for instance, because mental operations or computation for the optimal procedures were too costly or because they would involve too much or unavailable knowledge. In other words, heuristics in this second reading of the term constitute suboptimal rules (i.e., rules of thumb), in particular, rules that insure effective cognitive functioning in the presence of strong limitations to cognitive processing: "The common procedure [...] is to start with a method that is considered optimal, eliminate some aspects, steps, or calculations, and propose that the mind carries out this naive version." [9, p. 75].

This view assumes, as a corollary, that results produced by heuristics, while perhaps on acceptable levels, will generally be of a poorer quality than results of ideal procedures. Quality of results may thus be improved by allocating more processing resources, by using more information, or by allowing more time for the processing

[5]. While heuristics in the first sense of the term do not emphasize optimality (or, rather, suboptimality) as a defining criterion, the second reading does. Heuristics are (computationally) cheap versions of optimal procedures that produce results inferior to those of optimal procedures, and you would only want to use heuristics when a running of corresponding optimal procedures is not an option; for instance because of cognitive limitations. Using less effort than required for the optimal procedures comes at the price of reduced accuracy.

HEURISTICS AS FUNDAMENTAL PSYCHOLOGICAL MECHANISMS

Fast and frugal heuristics (FFH)s as proposed by Gigerenzer *et al.* [1–12] add yet a third reading to the term heuristics that combines elements of the first two readings and also adds additional aspects. It shares with the first reading that heuristics enable us, as humans, to make judgments, decisions, or solve difficult problems; it differs from it in that heuristics in the third sense are not detailed, compositional procedures that one follows as one follows a recipe during cooking, but rather are a collection of (mostly, very) simple rules that are followed largely implicitly or subconsciously in human decision making.

This third reading of heuristics shares with the second reading that both types of heuristics are relatively simple and that their executing requires comparably little cognitive and computational resources. However, the third reading differs from the second in that there is no need for a comparison with alternative optimal procedures, not even a need for the existence of such procedures, but that heuristics instead constitute a group of fundamental psychological mechanisms that guide information search and produce decisions by effectively and efficiently exploiting information structures in the environment. As we will see in the following, the readings also differ in their assessment of how beneficial more processing and more information are for the accuracy of results of the decisions that are taken.

FAST AND FRUGAL HEURISTICS AT FIRST GLANCE

Bishop [13, p. 202] sums up the two basic tenets of the research program on FFHs carried out by Gigerenzer *et al.* as follows: (i) humans “are naturally disposed to use simple reasoning rules,” and (ii) “these rules are about as reliable in the long run as ideal rules.” In more detail, FFHs are frequently characterized by several or all the following properties: (i) FFH are ecologically, but not logically, rational; (ii) they have evolved as fundamental psychological mechanisms effective in human and animal adaptive problem solving; (iii) they operate in fast, frugal, and simple ways; (iv) they are precise and transparent enough to serve as basis for computational, model-based descriptions of cognition; (v) they are generally psychologically plausible. We will look at each of these properties in turn, beginning with a comparative discussion of ecological rationality of heuristic procedures and “normative” rationality in ideal procedures.

HEURISTICS AS ADAPTIVE PROCEDURES FOR SUCCESSFUL BEHAVIOR IN AN ENVIRONMENT

What does it mean for a set of heuristics to be ecologically rational? Let us first consider some more general points. Basically, the question comes down to one of relative performance with regard to the conditions that are prevalent in the structure of the environment. On a general level, any natural perceptual or cognitive system on this planet has evolved given specific environmental conditions; the systems have adapted to these conditions to a degree where effective functioning depends on, and only makes sense, given the conditions’ presence.

Exploiting environmental conditions for the functioning of biological systems has clear procedural benefit (e.g., as it leads to a reduction of functional complexity). For example, Lockhead and Pomerantz [14] write on the interrelationship between structures in the world and perceptual functioning: “The significance of structure

for understanding perception resides in three tenets: first, that structure abounds in the physical world; second, that a perceptual system that knew about and capitalized on such structure would enjoy advantages in the speed and accuracy of perception; and third, that biological perceptual systems do exploit structure.” [15, p. 10]. It is important to keep in mind that such adaptation can occur on different levels (i.e., it is not limited to perception) and that it can be induced by different kinds of processes, including evolutionary, social, or individual (i.e., learning) ones. Many environmental properties seem to be fairly constant even when we consider them under evolutionarily relevant time spans. As a matter of fact, many environmental properties are so basic that their presence has become (or, always has been) necessary for biological life itself: a specific distribution of certain physical elements in the immediate vicinity (such as of gases in the atmosphere, for creatures such as humans), a gravitational pull of certain strength, regular daylight of certain wavelength distribution, and so on. When it comes to human and animal judgment and decision making, properties of such kind are perhaps not too often involved, at least, they were not under evolutionary terms as there is/were not much variation. (Or, if there were, there would not be anyone left to make a decision.) Many other equally relevant environmental properties, however, do require adaptive judgments and decisions on actions on behalf of the individual; these include, among others, the availability and distribution of environmental features such as food, water, or suitable mates within a specific area. There is little doubt that many of such (and other) judgments and decisions in animals and humans, which require adaptive problem solving, are made based on the outcomes of heuristic procedures and that at least some of these procedures are surprisingly simple [16]. Consider, for example, strategies that a peahen may employ for choosing a suitable mate. She may try to base her decision on all the information that is potentially available to her and act as close to an “ideal” procedure as possible (e.g., she could compare all

available peacocks on a number of features, then compute expected overall utility values for each peacock based on a weighing of individual features, and finally choose the one peacock with the highest score), or she may employ a much simpler procedure (i.e., a heuristic), perhaps one with just a single cue: inspect only a few candidates and pick the one with the highest number of eyes on his train. There is an evidence that peahens do, in fact, act upon the second, simpler strategy [17]. How should one evaluate such choice of strategy? If we follow the notion that heuristics (in the sense of the second reading of the term) are suboptimal procedures that evolved from necessary trade-offs with limited processing resources (e.g., possessed by our peahen) then one should assume that the more inclusive first mate selection strategy should lead to better choices of mates. If one follows the notion that heuristics (in the sense of the third reading of the term) are basically mechanisms which have evolved to work well (perhaps, even be optimal?) given a certain type of environment, then such an assumption does not need to hold. Similar issues of strategy evaluation can be raised for pretty much any heuristics-based decision or judgment. It seems that only a look at some actual performance data can clarify this matter, so let us now turn to performance results of a selected simple heuristic, before again addressing questions of ecological rationality.

THE LESS-IS-MORE EFFECT

Let us inspect one example among the FFHs that have been proposed so far. Goldstein and Gigerenzer [9] present an effect which, at least on a surface level, should appear most counterintuitive to the reader trained to trust that possessing and considering more relevant information in a decision process will mean an increase in the quality of obtained results. Countering that belief, Goldstein and Gigerenzer show that, in some situations, a little ignorance (i.e., less relevant knowledge for a decision or judgment) on a topic can in fact lead to taking better decisions than having more

knowledge. They present a study in which they asked groups of students from Germany and the United States to decide whether San Antonio or San Diego is, in fact, the larger city. While about two-thirds of the American students correctly named San Diego, all of the German students in the study produced the correct answer despite possessing much less knowledge about US cities than their American counterparts. How is this possible? Surely, more knowledge about the cities in this study (and about any relevant domain in general) should have resulted in more accurate judgments on the cities' relative size. This means that the American students should have outperformed the German group. Goldstein and Gigerenzer argue that employment of a powerful FFH, the recognition heuristic, causes the high performance levels of the German students. This heuristic can be summed up as follows: *If only one of two alternatives is recognized, infer that the recognized alternative has the higher value on the decision criterion.* Most German students in the study had recognized the name of San Diego but not that of San Antonio, and they had subsequently decided on the recognized city. Most American students in the study, however, were unable to employ this particular heuristic as they knew too much (i.e., they recognized the names of both cities). Goldstein and Gigerenzer refer to the effect as the less-is-more effect.

A similarly strong result of the recognition heuristic has been demonstrated by Borges *et al.* in the area of investing in the stock market [18]. A test of the recognition heuristic (i.e., buy only those stocks whose names are recognized) against a number of different investment strategies showed that the recognition heuristic in fact, outperformed the strategies during the 6-month period of the test. On a related note, performance in inference-making may even benefit from the systematic forgetting of knowledge [19].

Let us now step back from the individual studies and return our attention to the more general picture: When is the recognition heuristic ecologically rational; what is its accuracy? Surely, this particular heuristic cannot be always rational, or we should immediately act and close all schools and

libraries, as well as hide all books, so that we can all know less and become more accurate at making judgments and decisions. While a reproduction of the actual deduction of when a less-is-more effect will occur is beyond what can be shown in this article [9], the claim is that “any strategy for solving multiple-choice problems that follows the recognition heuristic will yield a less-is-more effect if the probability α of getting a correct answer merely on the basis of recognition is greater than the probability β of getting a correct answer using more information” [9, p. 79f]. The discovery of the less-is-more effect has been described as “one of the major discoveries of the last decades” [20]: “Contrary to the belief in a general accuracy-effort trade-off, less information and computation can actually lead to higher accuracy, and in these situations the mind does not need to make trade-offs. Here, a less-is-more effect holds” [5, p. 109].

A SECOND TAKE ON ECOLOGICAL RATIONALITY

It should have become clear by now that the reasons for the existence of the less-is-more effect exhibited by the recognition heuristic have to lie in how the heuristic manages to exploit certain properties of the information structure in the environment. In other words, the effect is context-dependent. Let us now look more closely at which environmental properties it actually depends upon.

Imagine trying to fit a function to a set of observed data samples. It has been known for some time that models (i.e., functions) that achieve good data fits on observed data samples do not necessarily result in equally good predictions of additional data [21]. On the one hand, the higher the degree of the polynomial used as model, the lower the error between the model and the observed data. On the other hand, the higher the degree of the polynomial, the greater the model’s flexibility and the greater its ability to not only pick up the underlying function that one intends to model but also the accidental variation in the data sample that is due to noise or statistical variance. If the model is very sensitive to

noise or variance, it easily overfits the data sample and is likely to predict poorly.

In order to better make the point of why simple heuristics can lead to more accurate results than more information-rich procedures, Gigerenzer and Brighton [5] suggest an error function of the kinds common in machine learning, in which the total prediction error equals the sum of three terms: the squared bias, the variance, and the noise. Bias is then defined as the difference between the “true” function underlying the data and a mean function produced from the data sample through some assumed induction process. If both functions are identical, there is no bias. The variance term simply adds a measure of how susceptible the induction process of the mean function is to patterns specific to the individual data samples; Gigerenzer and Brighton define it as the sum squared difference between the functions induced from individual data samples and the mean function. Assuming that noise is an invariant property of the environment, or that it invariably comes with taking data samples from it, low prediction errors can only be ensured by keeping both bias and variance at bay. On the basis of their error function, Gigerenzer and Brighton make a two-part argument. First, they relate the goal of attaining low prediction errors to the bias-variance dilemma [22]: for a low bias on a broad predictor, a model must accommodate various patterns. This, however, likely comes with greater flexibility, which in turn leads to greater accommodation of accidental data patterns and ultimately, to greater variance and a higher total prediction error. The way to keep variance down, on the other hand, is to reduce the broadness of the model, which in turn leads to a likely increase in bias and again, a higher prediction error. With reference to Geman *et al.* [22], Gigerenzer and Brighton conclude that it is due to this dilemma, that “‘general purpose’ models tend to be poor predictors of the future when data are sparse” [5, p. 120]. Second, they bring in findings indicating that human performance in generalizing from only few observations is often remarkably accurate [23]. Both points can be consoled when one, first, considers heuristics in human cognition to be specialized

rather than broad in order to keep variance (and, thus, total prediction error) down and, second, when one considers a whole set of biased, specialized heuristics (an adaptive toolbox, in FFH terms) to account for accurate human performance in judgment and decision making across a broad spectrum of contexts.

Gigerenzer and Brighton argue that the relatively high predictive accuracy of heuristics is often simply due to their ability to control variance better than more flexible models can. They note that “the variance component of error is always an interaction between characteristics of the inference strategy, the structure of the environment, and the number of observations available. This is why the performance of heuristics in an environment is not reflected by a single number such as predictive accuracy, but by a learning curve revealing how bias and variance change as more observations become available.” [5, p. 120]. A use of heuristics can lead to more accurate performance than a use of strategies with more information, more computation, or more time “not because the world is similarly simple,” but instead “by betting on lower variance, not lower bias” [5, p. 124]. As a consequence, simple heuristics fare particularly well under those conditions, under which variance is the dominant source of prediction error. Naturally, these include those under which observations are sparse or noise is high.

AN ADAPTIVE TOOLBOX OF FAST AND FRUGAL HEURISTICS

The research program on FFHs has so far included a number of heuristics of various sorts. One of the program’s most recent publications [5] lists 10 well-studied examples, together with criteria of ecological validities and related surprising findings; for reasons of space, we will only name two more here, in addition to the recognition heuristics already discussed above. The fluency heuristic (if both alternatives are recognized, but at different speeds, choose the faster one) shows a less-is-more effect comparable to that of the recognition heuristics: systematic forgetting

can lead to more accurate predictions [19]. The imitate-the-majority heuristic is ecologically rational in stable environments or when search for information is very costly: consider the majority of people in your peer group and imitate their behavior.

OTHER PROPERTIES OF FAST AND FRUGAL HEURISTICS

Let us now return to the remaining properties of FFHs listed above under (ii) to (v). (ii) What does it mean for a heuristic to have evolved as a fundamental psychological mechanism? First, recall our discussion of the second and third reading of the term heuristics: the functioning of the psychological mechanism is directly represented as the heuristic procedure, not as an abstraction from a more complex ideal procedure that has occurred to cognitively function in spite of limited cognitive resources. One may similarly argue that such resource limitations are equally products of adaptation to environmental structures, as we have done above. Second, the individual judgment and decision making strategies may be direct consequences of natural selection; for instance, with regard to sexual selection or feeding behavior in animals [16]. (iii) Fast heuristics are those that allow for quick decisions and judgments; for example, they are easy to use. Frugal heuristics are those that (can) operate on only little of the available information and do not involve much and complex computation. This implies that not only are the processes that integrate information to make a decision frugal, but also that there exist rules that quickly stop search for and integration of information once a specific criterion has been reached [24]. (iv) What does it mean for heuristics to be precise enough to be amenable for computational modeling? Computational descriptions of behavior have been ascribed with a number of advantages over, for instance, conceptual descriptions in particular, because of the required levels of specificity [25]. FHH are precise and they can be sufficiently specified, for instance in terms of search rules, stop rules, and

decision rules that can (and have been) computationally modeled and tested. (v) Gigerenzer *et al.* [23] further suggest that general psychological plausibility is important for models of cognition in general and that the FFHs are, in fact, psychologically plausible. They break plausibility down to tractability in realistic scenarios rather than in artificial test boxes, to robustness (i.e., low model complexity, few or no free parameters), to frugality and speed in reaching decisions, and to evidence (i.e., the model should be consistent with other accepted findings and computational explanations of cognition).

EMPIRICAL EVALUATION AND MODELING

As laid down in Refs 5 and 1, the FFH research program is meant to include a systematic evaluation of the proposed heuristics. Such evaluation should be carried out with respect to three minimum criteria: (i) multiple models need to be tested against each other to determine relative predictiveness; (ii) models need to be tested for individuals as individuals differ in the strategies that they use; and (iii) a test of adaptive selection—test whether individuals employ heuristics where their use is ecologically rational. Also, with respect to the desired degree of psychological plausibility, models need to be tested in larger computational contexts, such as in the frameworks of the commonly used cognitive architectures and memory models (see Ref. 24, for a discussion on past and ongoing activities related to the ACT-R cognitive architecture). These approaches can again inform research on which heuristics humans and animals use, and on which heuristics they use given certain structures of information in the environment. Without the proper analysis of a heuristic's use in practice, it is little more than an interesting theoretical account of a procedural mechanism. With a proper analysis, it may in fact add much to our knowledge about cognitive procedures in judgment and decision making, as well as inform the design of environmental conditions and applications that improve human judgment and decision making.

TAKE-HOME MESSAGE FOR THE DECISION SCIENTIST

The FFH research effort offers an intriguing perspective on heuristics that constitute a set of fundamental psychological mechanisms through which humans make judgments, reach decisions, and act upon the environment. In this view, heuristics are no second-class procedures that are carried out only because one cannot computationally or cognitively “afford” to execute better procedures (e.g., due to cognitive limitations), but rather, heuristics constitute an effective and highly efficient group of procedures in their own right. Some FFHs exhibit so-called less-is-more effects by which accuracy of predictions is actually increased under less information. These effects are contrary to outcomes expected under traditional beliefs in general accuracy–effort trade-offs and they are further indications of the power of simple heuristics in judgment and decision making.

REFERENCES

1. Gigerenzer G, Todd PM, the ABC Research Group. Simple heuristics that make us smart. New York: Oxford University Press; 1999.
2. Jungius J. Heuretica 1622. Cited after [3].
3. Kleining G, Witt H. Discovery as basic methodology of qualitative and quantitative research. Forum Qual Sozialforschung/Forum: Qual Soc Res 2001;2.
4. Descartes R. Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences. 1637.
5. Gigerenzer G, Brighton H. Homo Heuristicus: why biased minds make better inferences. Top Cogn Sci 2009;1:107–143. (20f)
6. Duncker K. On problem solving. Psychol Monogr 1945;58:5 (Whole No. 270).
7. Simon HA. A behavioral model of rational choice. Q J Econ 1955;69:99–118.
8. Newell A, Simon HA. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall; 1972.
9. Goldstein DG, Gigerenzer G. Models of ecological rationality: the recognition heuristics. Psychol Rev 2002;109(1):75–90.
10. Todd PN, Gigerenzer G. Précis of simple heuristics that make us smart. Behav Brain Sci 2000;23:727–780.

11. Gigerenzer G, Selten R, editors. Bounded rationality: the adaptive toolbox. Cambridge, MA: MIT Press; 2001.
12. Gigerenzer G. Gut feelings: the intelligence of the unconscious. New York: Viking; 2007.
13. Bishop MA. Fast and frugal heuristics. *Philos Compass* 2006;1/2:201–223.
14. Lockhead G, Pomerantz J. The perception of structure. Washington, D.C.: American Psychological Association. 1991.
15. Neumann H, Stiehl H. Modelle der frühen visuellen Informationsverarbeitung. In: Görz G, editor. Einführung in die künstliche Intelligenz. Bonn: Addison-Wesley; 1995. pp. 583–657.
16. Hutchinson JMC, Gigerenzer G. Simple heuristics and rules of thumb: where psychologists and behavioural biologists might meet. *Behav Processes* 2005;69:97–124.
17. Petrie M, Halliday T. Experimental and natural changes in the peacock's (*Pavo cristatus*) train can affect mating success. *Behav Ecol Sociobiol* 1994;35:213–217.
18. Borges B, Goldstein D, Ortmann A, Gigerenzer G. Can ignorance beat the stock market. In: Gigerenzer G, Todd P, the ABC Research Group, editors. Simple heuristics that make us smart. Oxford: Oxford University Press; 1999.
19. Schooler LJ, Hertwig R. How forgetting aids heuristic inference. *Psychol Rev* 2005;112: 610–628.
20. Gigerenzer G. Rationality for mortals. New York: Oxford University Press; 2008.
21. Roberts S, Pashler H. How persuasive is a good fit? A comment on theory testing. *Psychol Rev* 2000;107:358–367.
22. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput* 1992;4:1–58.
23. Griffiths TL, Tenenbaum JB. Optimal predictions in everyday cognition. *Psychol Sci* 2006;17(9):767–773.
24. Gigerenzer G, Hoffrage U, Goldstein DG. Fast and frugal heuristics are plausible models of cognition: reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychol Rev* 2008;115(1):230–239.
25. Simon HA. What is an “explanation” of behavior? *Psychol Sci* 1992;3:150–161.

FEASIBLE DIRECTION METHOD

DING-ZHU DU

WEILI WU

Department of Computer Science,
University of Texas at Dallas,
Richardson, Texas

PANOS M. PARDALOS

Department of Industrial and Systems
Engineering and Biomedical
Engineering, University of Florida,
Center for Applied Optimization,
Gainesville, Florida

NASSIM SOHAEI

Department of Clinical Sciences,
University of Texas Southwestern
Medical Center, Dallas, Texas

ALGORITHM DESCRIPTION

The feasible direction method is an important class of algorithms for solving constrained nonlinear optimization problems. Each iteration of a feasible direction method consists of two steps:

- Step 1.* Compute a feasible direction (i.e., a direction along which there exists a piece of segment lying inside of feasible region.)
- Step 2.* Carry out a line-search along the feasible direction obtained in Step 1, to obtain a new feasible point.

There are many ways to compute a feasible direction and also many ways to do line-search. Computing a feasible direction and doing line-search are usually independent. Each method for computing a feasible direction can be combined with any line-search procedure to form a feasible direction method. However, the classification of feasible direction methods are mainly based on the way for computing a feasible direction. In the literature, there are three classical methods,

Zoutendijk's Method [1], Wolfe's method [2,3], and Rosen's method [4,5]. In this article, we survey various historical issues about them, including their global convergence and extension to nonlinear constrained optimization.

Zoutendijk's Method, Wolfe's method, and Rosen's method are initially proposed for solving linearly constrained nonlinear optimization problems.

Zoutendijk's Method

Consider the following problem:

$$\begin{aligned} & \text{maximize } f(\mathbf{x}) \\ & \text{subject to } h_j(\mathbf{x}) \geq 0, j = 1, 2, \dots, m, \end{aligned}$$

where f and h_j are continuously differentiable functions on R^n . Zoutendijk [1] gave the following algorithm.

Initially, choose a starting feasible point $\mathbf{x}_1$. At each iteration $k = 1, 2, \dots$, solve the following linear programming

$$\begin{aligned} & \text{maximize } z \\ & \text{subject to } \nabla f(\mathbf{x}_k)^T \mathbf{d} - z \geq 0, \\ & \quad \nabla h_j(\mathbf{x}_k)^T \mathbf{d} - z \geq 0 \quad \text{for } j \in J_k, \\ & \quad -1 \leq d_i \leq 1 \quad \text{for } i = 1, \dots, n, \end{aligned} \tag{1}$$

where d_i is the i th component of $\mathbf{d}$. Let $(z_k, \mathbf{d}_k)$ be an optimal solution of the linear programming. If $z_k = 0$, then stop; $\mathbf{x}_k$ is a Fritz John point. If $z_k > 0$, then choose an appropriate $\bar{\lambda}_k$ such that for $\lambda \in [0, \bar{\lambda}_k]$, $\mathbf{x}_k + \lambda \mathbf{d}_k$ is feasible, and find a new feasible point $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ ($0 < \lambda_k \leq \bar{\lambda}_k$) by a line-search procedure.

Wolfe's Reduced Gradient Method

Consider the following problem:

$$\begin{aligned} & \text{maximize } f(\mathbf{x}) \\ & \text{subject to } \mathbf{Ax} = \mathbf{b} \\ & \quad \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{A}$ is an $m \times n$ matrix of rank m , and $\mathbf{b}$ is an m -dimensional vector. Assume that every basic feasible solution is nondegenerate. Wolfe [3] suggested the following algorithm.

Choose an initial point $\mathbf{x}_1$. At iteration $k = 1, 2, \dots$ with point $\mathbf{x}_k$, the following steps are carried out.

1. Find a feasible basis I_k such that $x_{ki} > 0$ for all $i \in I_k$, where x_{ki} is the i component of $\mathbf{x}_k$. Divide n variables into two groups $\mathbf{x}_k, I_k = (x_{ki}, i \in I_k)$ and $\mathbf{x}_k, \bar{I}_k = (x_{ki}, i \in \bar{I}_k)$, where $\bar{I}_k = \{1, \dots, n\} \setminus I_k$. Computer

$$\mathbf{r} = \nabla f(\mathbf{x}_k) - (\mathbf{A}_{I_k}^{-1} \mathbf{A})^T \nabla_{I_k} f(\mathbf{x}_k),$$

where $\nabla_{I_k} f(\mathbf{x}_k)$ is a subvector of $\nabla f(\mathbf{x}_k)$ with indices in I_k and $\mathbf{A}_{I_k}$ is a submatrix of $\mathbf{A}$ with column indices in I_k . Choose the search direction $\mathbf{d}_k = (d_{k1}, \dots, d_{kn})^T$ by first setting that for $i \in \bar{I}_k$,

$$d_{ki} = \begin{cases} 0 & \text{if } r_{ki} \leq 0 \text{ and } x_{ki} = 0, \\ r_{ki} & \text{otherwise.} \end{cases}$$

and then setting

$$\mathbf{d}_k = \begin{pmatrix} -\mathbf{A}_I^{-1} \mathbf{A}_{\bar{I}_k} \mathbf{d}_{k, \bar{I}_k} \\ \mathbf{d}_{k, \bar{I}_k} \end{pmatrix}.$$

If $\mathbf{d}_k = \mathbf{0}$, then stop ($\mathbf{x}_k$ is a Kuhn-Tucker point). Otherwise go to the next step.

2. Define

$$\bar{\lambda}_k = \begin{cases} 1 & \text{if } d_{ki} \geq 0 \text{ for all } i, \\ \min \left\{ \frac{x_{ki}}{-d_{ki}} \mid d_{ki} < 0 \right\} & \text{otherwise.} \end{cases}$$

Find a new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ by a line-search procedure.

Rosen's Gradient Projection Method

Consider the following problem:

$$\begin{aligned} & \text{maximize} && f(\mathbf{x}) \\ & \text{subject to} && \mathbf{a}^T \mathbf{x} = b_j, \quad j = 1, 2, \dots, m', \\ & && \mathbf{a}^T \mathbf{x} \geq b_j, \quad j = m' + 1, \dots, m. \end{aligned} \quad (2)$$

Denote $M = \{1, 2, \dots, m\}$ and $M' = \{1, 2, \dots, m'\}$. For any $J \subseteq M$, denote $D_J = \{\mathbf{d} \mid \mathbf{a}^T \mathbf{d} = 0, j \in J\}$. P_J is the orthogonal projection operator for subspace D_J . Let $J(\mathbf{x})$ denote the active index set at $\mathbf{x}$. In particular, we denote $J_k = J(\mathbf{x}_k)$ and $J^* = J(\mathbf{x}^*)$. Throughout this article, we assume that for every feasible point $\mathbf{x}$, $\mathbf{a}_j, j \in J(\mathbf{x})$ are linearly independent.

Rosen's method is as follows:

Initially, choose a feasible point $\mathbf{x}_1$. At each iteration $k = 1, 2, \dots$, the algorithm carries out the following steps:

1. Choose a positive number c_k . Compute a search direction by the following formula.

$$\mathbf{d}_k = \begin{cases} P_{J_k} \mathbf{g}_k & \text{if } \|P_{J_k} \mathbf{g}_k\| > c_k u_{kh_k}, \\ P_{J_k \setminus h_k} \mathbf{g}_k & \text{otherwise,} \end{cases}$$

where

$$(u_{kj}, j \in J_k)^T = (\mathbf{A}_{J_k}^T \mathbf{A}_{J_k})^{-1} \mathbf{A}_{J_k}^T \mathbf{g}_k$$

and

$$u_{kh_k} = \max\{u_{kj} \mid j \in J \setminus M'\}.$$

2. If $\mathbf{d}_k = \mathbf{0}$, then stop; $\mathbf{x}_k$ is a Kuhn-Tucker point. If $\mathbf{d}_k \neq \mathbf{0}$, then compute

$$\bar{\lambda}_k = \begin{cases} 1 & \text{if } \mathbf{a}_j^T (\mathbf{x}_k + \mathbf{d}_k) \geq b_j \text{ for all } j \notin J_k, \\ \min \left\{ \frac{b_j - \mathbf{a}_j^T \mathbf{x}_k}{\mathbf{a}_j^T \mathbf{d}_k} \mid \mathbf{a}_j^T (\mathbf{x}_k + \mathbf{d}_k) < b_j \right\} & \text{otherwise,} \end{cases}$$

and find a new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ ($0 < \lambda_k \leq \bar{\lambda}_k$) by a line-search procedure.

There are many choices for the line-search procedure. The following are some examples.

Accurate Line-Search 1. The procedure computes a solution λ^+ for the problem of maximizing $f(\mathbf{x} + \lambda \mathbf{d})$ subject to $0 \leq \lambda \leq \bar{\lambda}$, and set $\mathbf{x}^+ = \mathbf{x} + \lambda^+ \mathbf{d}$.

Accurate Line-Search 2. The procedure finds the first local maximum λ^+ for the function $f(\mathbf{x} + \lambda \mathbf{d})$ with respect to λ over $[0, \bar{\lambda}]$, and set $\mathbf{x}^+ = \mathbf{x} + \lambda^+ \mathbf{d}$.

Goldstein Test. Select a constant α from $(0, 1/2)$. The procedure computes a new feasible point $\mathbf{x}^+ = \mathbf{x} + \lambda^+ \mathbf{d}$ ($0 < \lambda^+ \leq \bar{\lambda}$) such that

$$\alpha \lambda^+ \mathbf{g}(\mathbf{x})^T \mathbf{d} \leq f(\mathbf{x}^+) - f(\mathbf{x})$$

and either $\lambda^+ = \bar{\lambda}$ or

$$(1 - \alpha) \lambda^+ \mathbf{g}(\mathbf{x})^T \mathbf{d} \geq f(\mathbf{x}^+) - f(\mathbf{x}).$$

Wolfe Test. Select a constant α from $(0, 1/2)$. The procedure computes a feasible point $\mathbf{x}^+ = \mathbf{x} + \lambda^+ \mathbf{d}$ ($0 < \lambda^+ \leq \bar{\lambda}$) such that

$$\alpha \lambda^+ \mathbf{g}(\mathbf{x})^T \mathbf{d} \leq f(\mathbf{x}^+) - f(\mathbf{x})$$

and either $\lambda^+ = \bar{\lambda}$ or

$$(1 - \alpha) \mathbf{g}(\mathbf{x})^T \mathbf{d} \geq \mathbf{g}(\mathbf{x}^+)^T \mathbf{d}.$$

Armijo's Rule. The procedure finds a point $\mathbf{x}^+ = \mathbf{x} + \lambda^+ \mathbf{d}$ ($0 < \lambda^+ \leq \bar{\lambda}$) such that

$$\frac{1}{2} \lambda^+ \mathbf{g}(\mathbf{x})^T \mathbf{d} \leq f(\mathbf{x}^+) - f(\mathbf{x})$$

and either $\lambda^+ = \bar{\lambda}$ or

$$\lambda^+ \mathbf{g}(\mathbf{x})^T \mathbf{d} > f(\mathbf{x} + 2\lambda^+ \mathbf{d}) - f(\mathbf{x}).$$

Each line-search procedure can be considered as a point-to-set mapping from the point $(\mathbf{x}, \mathbf{d}, \bar{\lambda})$ to the set of all possible $\mathbf{x}^+$'s. Usually, a line-search procedure T have the following properties:

1. $\forall \mathbf{x}^+ \in T(\mathbf{x}, \mathbf{d}, \bar{\lambda}) : f(\mathbf{x}^+) > f(\mathbf{x})$.
2. If $\mathbf{g}(\mathbf{x})^T \mathbf{d} > 0$ and $0 < \bar{\lambda} < \infty$, then $T(\mathbf{x}, \mathbf{d}, \bar{\lambda}) \neq \emptyset$.

We call a line-search procedure T *normal* if in addition, it satisfies

$$\left. \begin{array}{l} \mathbf{x}_k \rightarrow \mathbf{x}^* \\ \mathbf{d}_k \rightarrow \mathbf{d}^* \\ \mathbf{x}_k^+ \in T(\mathbf{x}_k, \mathbf{d}_k, \bar{\lambda}_k) \\ f(\mathbf{x}_k^+) \rightarrow f(\mathbf{x}^*) \end{array} \right\} \Rightarrow \text{either } \lambda_k^+ \rightarrow 0 \text{ or } (\mathbf{g}^*)^T \mathbf{d}^* \leq 0 \quad (3)$$

and

$$\left. \begin{array}{l} \mathbf{x}_k \rightarrow \mathbf{x}^* \\ \mathbf{d}_k \rightarrow \mathbf{d}^* \\ \mathbf{x}_k^+ \in T(\mathbf{x}_k, \mathbf{d}_k, \bar{\lambda}_k) \\ f(\mathbf{x}_k^+) \rightarrow f(\mathbf{x}^*) \end{array} \right\} \Rightarrow \begin{array}{l} \text{either } (\mathbf{g}^*)^T \mathbf{d}^* \leq 0 \\ \text{or } \exists k_0 \forall k \geq k_0 : \lambda_k^+ = \bar{\lambda}_k \end{array}, \quad (4)$$

where $\mathbf{x}_k^+ = \mathbf{x}_k + \lambda_k^+ \mathbf{d}_k$.

In order for a feasible direction method to have a good performance, we usually choose to use a normal line-search procedure. All the above five line-search procedures (Accurate Line-Search 1 and 2, Armijo's Rule, Goldstein Test, and Wolfe Test) have been proved to be normal [6].

GLOBAL CONVERGENCE

Wolfe [3] gave a counterexample to show that Zoutendijk's method may generate a sequence of points converging to a point which is not a Fritz John point. There is a counterexample [2,7] to show that Wolfe's method may also generate a sequence convergent to a point which is not a Kuhn–Tucker point. The global convergence of Rosen's method was a well-known long-standing open problem, and its proof was found by Bazaraa and Shetty [7] and Du and Zhang [8,9].

Although Zoutendijk's method and Wolfe's method may generate a convergent sequence of points not convergent to a stationary point, such as a Fritz John point or a Kuhn–Tucker point, they still hold certain global convergence that if generated sequence is not convergent then every cluster point is a Fritz John point or a Kuhn–Tucker point. This is

due to a very powerful tool, slope lemma [6] given as follows:

Lemma 1 [Slope Lemma]. *Let $\{\mathbf{x}_k\}$ be a sequence of feasible points satisfying that $f(\mathbf{x}_k) < f(\mathbf{x}_{k+1})$ for $k = 1, 2, \dots$. Let $\mathbf{x}^*$ be a cluster point of the sequence satisfying that*

() for any subsequence $\{\mathbf{x}_k\}_{k \in K}$ convergent to $\mathbf{x}^*$, $\mathbf{x}_{k+1} - \mathbf{x}_k \rightarrow \mathbf{0}$ as $k \rightarrow \infty, k \in K$.*

If $\{\mathbf{x}_k\}$ is not convergent to $\mathbf{x}^$, then there exists a subsequence $\{\mathbf{x}_k\}_{k \in K}$ such that*

- (a) $\mathbf{x}_k \rightarrow \mathbf{x}^*$ as $k \rightarrow \infty, k \in K$, and
- (b) $\lim_{k \rightarrow \infty, k \in K} \frac{\nabla f(\mathbf{x}_k)^T (\mathbf{x}_{k+1} - \mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} = 0$.

Proof. Note that for any subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$, we have

$$\begin{aligned} \lim_{k \rightarrow \infty, k \in K} \frac{\nabla f(\mathbf{x}_k)^T (\mathbf{x}_{k+1} - \mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} \\ = \lim_{k \rightarrow \infty, k \in K} \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|}, \end{aligned}$$

since

$$\begin{aligned} & \left| \frac{\nabla f(\mathbf{x}_k)^T (\mathbf{x}_{k+1} - \mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} - \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} \right| \\ &= \left| \frac{(\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{y}_k))^T (\mathbf{x}_{k+1} - \mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} \right| \\ &\leq \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{y}_k)\|, \end{aligned}$$

where $\mathbf{y}_k$ is a point on the segment $[\mathbf{x}_k, \mathbf{x}_{k+1}]$. Suppose that the lemma is false. Then for any subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$,

$$\liminf_{k \rightarrow \infty, k \in K} \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} > 0.$$

By a diagonalization argument, we can see that there exists a positive number ε such that for every subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$,

$$\liminf_{k \rightarrow \infty, k \in K} \frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} \geq \varepsilon. \quad (5)$$

It follows that there exists a positive number δ such that for all k , $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \delta$ implies

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \geq \frac{\varepsilon}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \quad (6)$$

In fact, if such a positive number δ does not exist, then we can find a subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$ and satisfying that for $k \in K$,

$$\frac{f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} < \varepsilon/2,$$

contradicting Equation (6).

Since $\{\mathbf{x}_k\}$ does not converge to $\mathbf{x}^*$, there exists a positive number $\delta' < \delta$ such that $\|\mathbf{x}_k - \mathbf{x}^*\| > \delta'$ for infinitely many k . Define $B(\delta') = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^*\| \leq \delta'\}$. For each $\mathbf{x}_k \in B(\delta')$, denote by k' the smallest $k' \geq k$ such that $\mathbf{x}_{k'} \notin B(\delta')$. Choose an infinite sequence $\{k_i\}$ such that $\mathbf{x}_{k_i} \in B(\delta'/2)$ and $k_i < k'_i < k_{i+1} < k'_{i+1}$ for every i . Now, we have

$$\begin{aligned} f(\mathbf{x}^*) - f(\mathbf{x}_{k_1}) &\geq \sum_{i=1}^{\infty} \sum_{j=k_i}^{k'_i-1} (f(\mathbf{x}_{j+1}) - f(\mathbf{x}_j)) \\ &\geq \sum_{i=1}^{\infty} \sum_{j=k_i}^{k'_i-1} \frac{\varepsilon}{2} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \\ &\geq \sum_{i=1}^{\infty} \frac{\varepsilon}{2} \|\mathbf{x}_{k'_i} - \mathbf{x}_{k_i}\| \\ &\geq \sum_{i=1}^{\infty} \frac{\varepsilon}{2} \cdot \frac{\delta'}{2} = \infty, \end{aligned} \quad (7)$$

by Equation (7), a contradiction.

The following is a corollary of the above slope lemma.

Corollary 1. *Let $\{\mathbf{x}_k\}$ be a sequence of feasible points generated by an algorithm described as above. Suppose that $f(\mathbf{x}_{k+1}) > f(\mathbf{x}_k)$ for all k . Consider a cluster point $\mathbf{x}^*$ of sequence $\{\mathbf{x}_k\}$ satisfying that*

- (a) *for every subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$, $\{\mathbf{d}_k\}_{k \in K}$ is bounded, and*
- (b)

$$\left. \begin{aligned} \lim_{k \rightarrow \infty, k \in K} \mathbf{x}_k &= \mathbf{x}^* \\ \lim_{k \rightarrow \infty, k \in K} \mathbf{d}_k &= \mathbf{d}^* \\ \nabla f(\mathbf{x}^*)^T \mathbf{d}^* &> 0 \end{aligned} \right\}$$

$$\implies \mathbf{x}_{k+1} - \mathbf{x}_k \rightarrow \mathbf{0} \text{ as } k \rightarrow \infty, k \in K.$$

If $\{\mathbf{x}_k\}$ does not converge to $\mathbf{x}^$, then there exists a subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$ such that $\lim_{k \rightarrow \infty, k \in K} \mathbf{d}_k = \mathbf{d}^*$ and $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$.*

Proof. Consider two cases.

Case 1. There is a subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$ such that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| > \varepsilon$ for all $k \in K$ and a positive number ε . Since $\{\mathbf{d}_k\}_{k \in K}$ is bounded, there is a subsequence $\{\mathbf{x}_k\}_{k \in K'}$ ($K' \subset K$) such that $\{\mathbf{d}_k\}_{k \in K'}$ converges to a vector $\mathbf{d}^*$. Since $\mathbf{d}_k$ is an ascending direction at $\mathbf{x}_k$, we have $\nabla f(\mathbf{x}_k)^T \mathbf{d}_k > 0$ for every k . Thus, $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* \geq 0$. By (d), we can conclude that $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$.

Case 2. For every subsequence $\{\mathbf{x}_k\}_{k \in K}$ converging to $\mathbf{x}^*$, $\mathbf{x}_{k+1} - \mathbf{x}_k \rightarrow \mathbf{0}$ as $k \rightarrow \infty, k \in K$. By the first slope lemma, there exists a subsequence $\{\mathbf{x}_k\}_{k \in K}$ such that

$$\lim_{k \rightarrow \infty, k \in K} \frac{\nabla f(\mathbf{x}_k)^T (\mathbf{x}_{k+1} - \mathbf{x}_k)}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} = 0.$$

Let $\mathbf{d}^*$ be a cluster point of $\{\mathbf{d}_k\}_{k \in K}$. Then

$$\frac{\nabla f(\mathbf{x}^*)^T \mathbf{d}^*}{\|\mathbf{d}^*\|} = \lim_{k \rightarrow \infty, k \in K} \frac{\nabla f(\mathbf{x}_k)^T \mathbf{d}_k}{\|\mathbf{d}_k\|} = 0.$$

Therefore, $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$.

In the above corollary, the condition (d) is important to the contradiction argument. In fact, we see that for some algorithms, $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$ implies that $\mathbf{x}^*$ is a Kuhn–Tucker point or a Fritz John point. So, $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* > 0$ is an assumption for contradiction. A more important fact is that condition (d) holds for every normal line-search.

Next, let us apply this corollary to Zoutendijk's method, Wolfe's method, and Rosen's method.

Theorem 1. *If an infinite sequence of points generated by Zoutendijk's method does not converge, then every cluster point of the sequence is a Fritz John point.*

If an infinite sequence of points generated by Wolfe's method does not converge, then every cluster point of the sequence is a Kuhn–Tucker point.

If an infinite sequence of points generated by Rosen's method does not converge, then every cluster point of the sequence is a Kuhn–Tucker point.

Proof. Proofs for three statements are similar. Thus, we give the proof only for the first statement as an example.

First, we claim that (d) holds because we use a normal line-search procedure. Secondly, we claim that

$$\left. \begin{aligned} \lim_{k \rightarrow \infty, k \in K} \mathbf{x}_k &= \mathbf{x}^* \\ \lim_{k \rightarrow \infty, k \in K} \mathbf{d}_k &= \mathbf{d}^* \\ \nabla f(\mathbf{x}^*)^T \mathbf{d}^* &= 0 \end{aligned} \right\} \Rightarrow \mathbf{x}^* \text{ is a Fritz John point.}$$

To show our claim, let $\mathbf{x}_k \rightarrow \mathbf{x}^*$ and $\mathbf{d}_k \rightarrow \mathbf{d}^*$ such that $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$, where $\mathbf{d}_k$ is determined at $\mathbf{x}_k$ by Zoutendijk's method, that is, there is a number z_k such that $(z_k, \mathbf{d}_k)$ is an optimal solution of the linear programming (Eq. 1). Denote $z^* = \limsup_{k \rightarrow \infty, k \in K} z_k$. Since $\nabla f(\mathbf{x}_k)^T \mathbf{d}_k - z_k \geq 0$, we have $-z^* \geq 0$. However, since $(z, \mathbf{d}) = (0, \mathbf{0})$ is always a feasible solution of Equation (1), we have $z_k \geq 0$ for all k . Thus, $z^* = 0$. Without loss of generality, we assume that for every k , $J_k = J$ for a certain index set J . (Otherwise, we can study a subsequence with the property.) Next, we show that z^* is the optimal value of the following problem:

$$\begin{aligned} &\text{maximize} && z \\ &\text{subject to} && \nabla f(\mathbf{x}^*)^T \mathbf{d} - z \geq 0, \\ & && \nabla h_j(\mathbf{x}^*)^T \mathbf{d} - z \geq 0 \text{ for } j \in J, \\ & && -1 \leq d_i \leq 1 \text{ for } i = 1, \dots, n. \end{aligned} \quad (8)$$

In fact, if z^* is not the optimal value, then there exists a feasible solution of Equation (8), $(z', \mathbf{d}')$ with $z' > 0$. It follows that for sufficiently large k , $(z'/2, \mathbf{d}')$ is a feasible solution of Equation (1). So, $z_k \geq z'/2$ for such k 's. Letting $k \rightarrow \infty$, we have $0 = z^* \geq z'/2$, a contradiction. Therefore, $z^* = 0$ is the optimal value of Equation (8). It implies that the following system of inequalities has no solution.

$$\begin{aligned} &z > 0, \\ &\nabla f(\mathbf{x}^*)^T \mathbf{d} - z \geq 0, \\ &\nabla h_j(\mathbf{x}^*)^T \mathbf{d} - z \geq 0, \text{ for } j \in J. \end{aligned}$$

(If the system has a solution $(z, \mathbf{d})$, we can obtain a feasible solution with $z > 0$ for

Equation (8) by normalizing $\mathbf{d}$.) Hence, there exist $u > 0$, $v > 0$, and $w_j \geq 0$ for $j \in J$ such that

$$\begin{aligned} v \nabla f(\mathbf{x}^*) + \sum_{j \in J} w_j \nabla h_j(\mathbf{x}^*) &= \mathbf{0}, \\ u - v - \sum_{j \in J} w_j &= 0. \end{aligned}$$

$u > 0$ implies $v + \sum_{j \in J} w_j \neq 0$. Therefore, $\mathbf{x}^*$ is a Fritz John point.

Clearly, the search direction determined by Zoutendijk's method is bounded by $\sqrt{n}$. Let $\{\mathbf{x}_k\}$ be an infinite sequence generated by Zoutendijk's method and $\mathbf{x}^*$ a cluster point of the sequence. Suppose that $\{\mathbf{x}_k\}$ does not converge to $\mathbf{x}^*$. By Corollary 1, there is a subsequence $\{\mathbf{x}_{k_i}\}_{i \in K}$ converging to $\mathbf{x}^*$ such that $\{\mathbf{d}_{k_i}\}_{i \in K}$ converges to $\mathbf{d}^*$ with $\nabla f(\mathbf{x}^*)^T \mathbf{d}^* = 0$.

By our claim, $\mathbf{x}^*$ is a Fritz John point.

While convergent sequence generated by Zoutendijk's method and Wolfe's method may not converge to a Fritz John point or a Kuhn–Tucker point, convergent sequence generated by Rosen's method is still convergent to a Kuhn–Tucker point.

Theorem 2. *Let α and β be two positive numbers with $\alpha < \beta$. Suppose that $\alpha \leq c_k \leq \beta$ for all k . Then Rosen's method with a normal line-search procedure either stops at a Kuhn–Tucker point or generates an infinite sequence whose cluster points are all Kuhn–Tucker points.*

The proof of global convergence for Rosen's method for convergent generated sequence is based on the second slope lemma [6]. In general, several lemmas motivated from properties of slope are very interesting and helpful in establishing the global convergence of feasible direction methods.

SOME DEVELOPMENTS

Since Zoutendijk's method and Wolfe's method do not hold the global convergence when generated sequence is not convergent and the global convergence of Rosen's method

was an open problem for a long period, one discovered many variations [10–20,21] about them to improve their convergence. This research direction was very active 25 years ago. Meanwhile, one tried to improve their convergent rate by introducing various variable metric methods and conjugate gradient methods into feasible direction methods [15–17,22–25,28–30]. Those variations have superlinear convergent rate. While the feasible direction method is well-studied for linear constraints, their extensions to nonlinear constraints is also developed extensively [21–26].

REFERENCES

1. Zoutendijk G. Methods of feasible directions. New York: Elsevier; 1960.
2. Wolfe P. On the convergence of gradient methods under constraint. IBM J Res Dev 1972;16:407–411.
3. Wolfe P. Methods of nonlinear programming. In: Graves RL, Wolfe P, editors. Advances in mathematical programming. New York: McGraw-Hill; 1963.
4. Rosen JB. The gradient projection method for nonlinear programming, part I: linear constraints. SIAM J Appl Math 1960;8:181–217.
5. Rosen JB. The gradient projection method for nonlinear programming, part II: nonlinear constraints. SIAM J Appl Math 1961;9:514–553.
6. Du D-Z, Pardalos PM, Wu W. Mathematical theory of optimization. Boston: Kluwer Academic Publishers; 2001.
7. Bazaraa MS, Shetty CM. Nonlinear programming: theory and algorithms. New York: John Wiley & Sons, Inc.; 1979.
8. Du D-Z, Zhang X-S. A convergence theorem of Rosen's gradient projection method. Math Program 1986;36:135–144.
9. Du D-Z, Zhang X-S. Global convergence of Rosen's gradient projection methods. Math Program 1989;44:357–366.
10. Du D-Z, Zhang X-S. Notes on a new gradient projection method. Syst Sci Math Sci 1989;2:184–192.
11. Du D-Z. A modification of Rosen-Polak's algorithm. Kexue Tongbao 1983;28:301–305.
12. Du D-Z. A family of gradient projection algorithms. Acta Math Appl Sin Engl Ser 1985;2:1–13.

13. Polak E. On the convergence of optimization algorithms. *Rev Fr Infor Rech Oper* 1969;3:17–34.
14. Polak E. *Computational methods in optimization*. New York: Academic Press; 1971.
15. Ritter K. A superlinearly convergent method for minimization problems with linear inequality constraints. *Math Program* 1973;4:44–71.
16. Ritter K. A method of conjugate directions for linearly constrained nonlinear programming problems. *SIAM J Numer Anal* 1975;12:272–303.
17. Ritter K. Convergence and superlinear convergence of algorithms for linearly constrained minimization problems. In: Dixon LCW, Spedicato E, Szego GP, editors. *Nonlinear optimization: theory and algorithms*, part II. Boston (MA): Birkhauser; 1980. pp. 221–251.
18. Yue M, Han J. A new reduced gradient method. *Sci Sin* 1979;22:1099–1113.
19. Yue M, Han J. A unified approach to the feasible direction methods for nonlinear programming with linear constraints. *Acta Math Appl Sin Engl Ser* 1984;1:63–73.
20. Zhang X-S. Discussion on Polak's algorithm of nonlinear programming. *Acta Math Appl Sin* 1981;4:1–13.
21. Zhang X-S. An improved Rosen-Polak method. *Acta Math Appl Sin* 1979; 2:257–267.
22. Goldfarb D. Extension of Davidon's variable metric method to maximization under linear inequality and equality constraints. *SIAM J Appl Math* 1969;17:739–764.
23. Goldfarb D, Lapidus L. Conjugate gradient method for nonlinear programming problems with linear constraints. *Ind Eng Chem Fund* 1968;7:142–151.
24. Lai Y-L. An algorithm and its convergence for nonlinear constrained convex programming. *Acta Math Appl Sin* 1980;3:322–331.
25. Lai Y-L, Wu Fang, Gui X-Y. A reduced variable metric algorithm for linearly constrained convex programming. *Sci Sin Ser A* 1983;26:785–794.
26. Du D-Z. A gradient projection algorithm for convex programming with nonlinear constraints. *Acta Math Appl Sin* 1985;8:7–16.
27. Rosen JB, Kreuser J. A gradient projection algorithm for nonlinear constraints. In: Lootsma FA, editor. *Numerical methods for nonlinear optimization*. New York: Academic Press; 1972.
28. Rosen JB. Two-phase algorithm for nonlinear constraint problems. In: Mangasarian OL, Meyer RR, Robinson SM, editors. *Nonlinear programming*. New York: Academic Press; 1978. pp. 97–124.
29. Buckley A. An alternate implementation of Goldfarb's minimization algorithm. *Math Program* 1975;8:207–231.
30. Byrd RH. An example of irregular convergence in some constrained optimization methods that use the projected Hessian. *Math Program* 1985;32:232–237.

FEATURE EXTRACTION AND FEATURE SELECTION: A SURVEY OF METHODS IN INDUSTRIAL APPLICATIONS

MICHEL J. ANZANELLO

Department of Industrial Engineering,
Federal University of Rio Grande do
Sul, Porto Alegre, Brazil

The feature selection problem, also known as variable selection, is deemed one of the most challenging problems in statistical applications. Consider a set of J variables describing a hypothetical process or situation; many of these J variables do not contribute to any inference, such as model building or classification procedure. The challenge consists of defining a subset of p relevant variables, $p < J$, carrying most of the important information with minimal noise.

Problems related to high volume of data have become evident in many segments. In statistical and mathematical applications, where the goal is usually to find a model yielding good prediction on the response variable, the excessive number of predictors might only add noise to the model. These problems can be extended to chemical and industrial processes, where the increasing use of sensors for process monitoring and the widespread availability of computational resources have made the handling of large amounts of data a challenging task [1]. Hundreds or even thousands of noisy and correlated process variables are collected on a minute basis for monitoring and control purposes. Such volume of data may jeopardize the accuracy of process control activities.

Selecting process variables in industrial applications is an important effort for several reasons. A model composed of a large number of variables may fit well to historical data, but it does not assure high performance in terms of prediction or classification due to overfitting [2,3]. Overfitting occurs in excessively complex models (i.e., presenting too many degrees of freedom); an

overfitted model tends to describe mostly random error or noise instead of the underlying relationship. Second, identification of important variables based on engineers' empirical knowledge may or may not be correct. Third, variables that are noisy, due to sensor and other problems, may lead to a confusing or misleading model. Finally, it is preferable to have a parsimonious model for easier data collection and analysis.

It appears that the nature of data and field of application dictate the best statistical techniques to select variables. Variable selection approaches in industrial applications have ranged from very basic selection techniques, for example, stepwise, backward, and forward [4,5], to more complex multivariate techniques, such as partial least squares (PLS) [6–8], data mining tools [9], principal component analysis (PCA), and clustering [4], among others, and optimization tools, such as linear programming and genetic algorithms (GAs) [5,10].

Another widely applied technique for dimensionality reduction in practical applications is the feature extraction, or variable extraction. Apart from variable selection, which retains only a subset of significant variables, variable extraction actually transforms the variable. It projects the original dataset with higher dimensionality onto a smaller number of dimensions using linear or nonlinear combinations of the original variables [11]. This yields a much smaller and richer set of variables, which becomes particularly useful for data visualization such that a complex dataset can be visualized efficiently when reproduced in two or three dimensions. Models built on extracted variables usually present higher predictive quality than models built on original variables since the data is described by fewer, more meaningful new variables.

A few applications of variable extraction include analysis on data compression, decomposition and projection, pattern recognition [12], image processing [13], and biomedical applications [14]. Although applied on

industrial scenarios, this article is tailored to the variable selection problem.

For the objectives of this article, we assume there are two main purposes for selecting variables: (i) variable selection for prediction, where one aims at finding a limited subset of process variables, x , leading to the best prediction of response variable, y ; and (ii) variable selection for classification, where the goal is to identify the subset of process variables with the best discriminative ability to categorize observations in classes. Some approaches related to these two purposes are briefly described in the following sections.

VARIABLE SELECTION FOR PREDICTION

Consider the relation between a set of J process variables, $x_1, \dots, x_J$, and a response variable y represented by the generic multiple linear regression (MLR) model in Equation (1). The goal is to identify a set of p process variables, $p < J$, able to better explain and predict the response variable y

$$y = \sum_{i=1}^J b_i x_i. \quad (1)$$

Simplistic approaches for variable selection in MLR rely on backward, forward, and stepwise methods. In the backward multiple regression method (BMR), an initial model with J variables is constructed and a significance test, such as F or p value, is performed on the regression coefficients. The variable with the smallest F (or largest p value) is removed and a new regression based on $J - 1$ variables is generated. The process is repeated until the remaining variables have an F value larger than a defined threshold. The resulting variables are then used in the prediction model. Forward multiple regression (FMR) is similarly guided by a significance test, but variables are now introduced one-by-one in the model. The stepwise multiple regression (SMR) is a combination of the FMR and BMR, where variables can either enter or leave the model, based on specific thresholds [15–17]. These

methods have also been used as a reference to evaluate the performance of recent methods for variable selection, as in Xu and Zhang [5], Chong and Jun [6], and Anzanello *et al.* [9].

In addition, several criteria have been developed to evaluate the impact of adding/removing variables in MLR models. Some of these measures include Akaike's criterion (AIC) [18] and Schwarz's Bayesian criterion (SBC) [19], which penalize models when process variables are added. Similarly, the PRESS (prediction sum of squares) criterion assesses how prediction is affected when variable i is momentarily omitted from the model; this procedure is similar to the leave-one-out-at-a-time analysis. The variable yielding the worst PRESS is removed, such that the model has a better predictive ability when that variable is left out; the procedure is repeated for the $i - 1$ remaining variables and the PRESS is continuously monitored [17].

The prediction quality of basic regression models, such as the MLR, may be highly affected by industrial data, where variables, frequently, are correlated and noisy, and may present missing values and more variables than observations [1,20]. These conditions affect the estimation of regression coefficients and jeopardize the quality of variable selection generated by BMR, FMR, and SMR techniques, yielding imprecise predictions. An alternative regression technique capable of handling this sort of data is the PLS regression, described in the following section.

Variable Selection in PLS Regression

Owing to its robustness in handling ill-conditioned data, the PLS regression became the main tool for prediction in manufacturing and chemometrics applications. PLS is an iterative regression that generates linear combinations of process and response variables, aiming at maximizing the covariance between these combinations. For further information on PLS regression and its parameters, see Wold *et al.* [21,22]; a discussion on the properties of PLS and MLR is reported by Almoy [23] and Dingstad *et al.* [24].

Although more robust than the MLR, irrelevant and noisy variables also endanger the predictive ability of the PLS regression, which justifies the large number of methods aimed at identifying the most relevant variables for use in that regression. The general approach consists of applying statistical tools to the original dataset in order to prescreen and select relevant variables, and then using the resulting subset of variables in a PLS regression model for prediction [4]. In some applications, the subset selected by a statistical tool is refined using the PLS regression coefficients and a smaller subset is effectively used on the final model.

Some industrial segments are well known for developing and applying variable selection methods in PLS regression: Wang *et al.* [25] in refinery processes; Wold *et al.* [21] in paper recycling; Hoskuldsson [26] and Xu *et al.* [27] in near-infrared data; and Baumann *et al.* [28] and Zhai *et al.* [29] in quantitative structure–activity relationship/quantitative structure–property relationship (QSAR/QSPR). PLS has also been combined with a large variety of multivariate and optimization techniques for variable selection, as reported by Gauchi and Chagnon [4]; Xu and Zhang [5]; Chong and Jun [6]; and Huang and Wang [10]. We now briefly present some of these methods.

Gauchi and Chagnon [4] performed an extensive comparison of methods for variable selection using PLS regression in five datasets of chemical processes. The best method suggested by these authors is named backward- Q^2_{cum} (BQ). The index Q^2_{cum} is a measure between 0 and 1 of the predictive performance of the model (values close to 1 denote better predictions). The BQ method initially constructs a PLS model using the J process variables, evaluates Q^2_{cum} , and eliminates the variable with the smallest absolute PLS regression coefficient. A new PLS is created using $J - 1$ variables, Q^2_{cum} is again measured, and a new variable is eliminated. This procedure is repeated until there are no remaining variables. The final subset of variables is the one with the maximum Q^2_{cum} .

Another simple and efficient method presented in Ref. 4 is named simple regression

(SR). The response variable (y) is regressed against each of the J process variables one at a time, generating J regression coefficients. A test of significance on each of these coefficients defines the variables that will be used as predictors in a PLS model.

In a similar context, Chong and Jun [6] proposed a comparison between three methods for variable selection in a steel production plant: PLS-VIP (i.e., PLS regression associated to a variable selection importance index, the Variable Importance for the Projection), Lasso regression (LR) and SMR variable selection methods. The methods were tested in simulated datasets generated by the following simulation factors: (i) proportion of important variables, (ii) magnitudes of correlations, (iii) structure of regression coefficients, and (iv) noise. These authors found PLS-VIP as the best method for variable selection.

Some variable selection methods utilize the information on PLS regression parameters to iteratively identify the most relevant variables. Sarabia *et al.* [30] suggested a graphical method based on the reweighting of selected elements in the PLS weight vector, using cross-validation to evaluate the quality of predictions generated by the model. Forina *et al.* [31] proposed the sequential updating of relevant process variables by using an important vector obtained in a previous iteration. Variables with little information were given small weights, reducing or eliminating the influence of irrelevant variables. In Zarzo and Ferrer [32], a variable selection approach to evaluate variability on the product variable is proposed and compared to a blocked PCA approach. Lazraq *et al.* [33] suggested two inferential variable selection methods in PLS regression, namely PLS-Forward and PLS-Bootstrap, and compared those to the best methods in Ref. 4. These authors stated that there is no unanimous variable selection method, and a combination of methods should be used to corroborate a final subset.

PLS has also been utilized for selecting the most relevant variables in near-infrared (NIR) data, where datasets may present thousands of process variables. Hoskuldsson [26]

suggested the removal of process variables that present low correlation with y before applying PLS, yielding a more stable regression model. In addition, Xu *et al.* [27] utilized weighting vectors to calibrate the contribution of each process variable in NIR data during PLS iterations. Such weighting vectors are estimated by means of particle swarm optimization (PSO), an optimization technique that finds a solution for a problem in a search space. As for Fourier transform-infrared (FT-IR) spectral data, Leardi *et al.* [34] applied GAs to identify relevant variables, which are then used in a PLS model aimed at improving prediction of additive concentrations in polymer films. The method proposed in Ref. 34 also reduces the probability of overfitting the PLS model. Other approaches in infrared data can be found in Duranda *et al.* [35], Ye *et al.* [36], and Teofilo *et al.* [37].

Another field where variable selection has become an issue is the QSAR/QSPR analysis. QSAR/QSPR is defined as the process by which the chemical structure of a substance is correlated with a well-known process (e.g., chemical reactivity or biological activity). For instance, chemical reactivity of a certain substance can be expressed as the amount of that substance required to obtain a chemical response. Such quantitative representation enables precise characterization of substance properties and reduces costs of analyses. Variable selection approaches in QSAR/QSPR aim at selecting the most relevant variables for describing such chemical reactivity. For that matter, Zhai *et al.* [29] used a combination of PLS and projection pursuit (a technique that projects a high-dimensional dataset into low-dimensional subspace; see Ref. 38) to reduce the number of variables for explanation of chemical products. In addition, Baumann *et al.* [28] combined PLS, PCA, and tabu search to evaluate the prediction quality of models after variable selection. Finally, Roy and Roy [39] integrated PLS, k -means clustering technique and several coefficients of regression adherence to identify the best subset of variables in a predictive model. Other QSAR/QSPR related variable selection

procedures are reported in Liu *et al.* [40] and Gonzalez *et al.* [41].

Variable Selection and Other Tools for Prediction

Although PLS regression covers most of the predictive role in manufacturing and chemometrics applications, other statistical tools have been used to select variables with that purpose. Xu and Zhang [5] found that GAs and leaps-and-bounds performed better than the traditional SMR, FMR, and BMR methods in chemical data describing nitrobenzene toxicity. The selected variables could then be used in MLR or artificial neural networks to predict substances toxicity. With similar purposes, Gualdron *et al.* [42] suggested an efficient two-phase method to select variables in gas sensing applications. The first phase estimates a “figure of merit” index for each variable; variables whose indices are below certain threshold are removed. The second phase applies GAs and other stochastic approaches on the remaining variables to identify important variables, which are then used to build neural predictive models.

Another regression technique yielding higher performance after variable selection is the LR. Regression coefficients of the LR are estimated by weighting the PRESS toward a “shrinking parameter.” This parameter defines the variables included in the regression model by forcing coefficients of less important variables to assume zero value. The LR as variable selection method was applied to steel manufacturing data by Chong and Jun [6]. For details on LR fundamentals, see Refs 43 and 44.

Additional approaches for variable selection with prediction purposes include the use of GAs, as in Chatterjee *et al.* [45], and Gualdron *et al.* [46]; and Bayesian methods, as in Philips and Guttman [47].

Table 1 depicts a summary of the most cited approaches for variable selection with prediction purposes based on their mathematical background. Studies combining two or more techniques are listed under all used techniques.

Table 1. Summary of Approaches for Variable Selection with Prediction Purposes

PLS Regression	Variable Selection for Prediction		
	Genetic Algorithm	Stepwise	Clustering
Gauchi and Chagnon [4]	Leardi <i>et al.</i> [34]	Xu and Zhang [5]	Gauchi and Chagnon [4]
Wold <i>et al.</i> [21,22]	Xu and Zhang [5]	Gauchi and Chagnon [4]	Roy and Roy [39]
Sarabia <i>et al.</i> [30]	Gualdron <i>et al.</i> [42]	Chong and Jun [6]	
Forina <i>et al.</i> [31]	Chatterjee <i>et al.</i> [45]		<i>Lasso Regression</i>
Lazraq <i>et al.</i> [33]			Chong and Jun [6]
Hoskuldsson [26]			
Xu <i>et al.</i> [27]			<i>PCA</i>
Zhai <i>et al.</i> [29]			Zarzi and Ferrer [32]
Baumann <i>et al.</i> [28]			
Roy and Roy [39]			<i>Bayesian</i>
Chong and Jun [6]			Philips and Guttman [47]
Marengo <i>et al.</i> [7]			

VARIABLE SELECTION FOR CLASSIFICATION

In several manufacturing processes, the objective of variable selection is not prediction, but classification of observations into classes regarding some final specification, such as quality, level of profitability, or product reliability. Such observations usually refer to production batches or lots. A proper classification procedure, especially during the early stages of the process, allows adjustments on the process parameters to correct inconsistencies or, in many cases, enables the assignment of the product to a specific destination.

Most variable selection approaches with classification purposes rely on the combination of data mining tools and discriminant functions, and are usually focused on (i) improving the classification ability of existing algorithms, and (ii) reducing the processing time of the algorithm using a small subset of variables. Assessment of classification performance attained by such methods usually relies on measures such as accuracy, sensitivity or specificity, among others [12].

There are many examples of successful methods and approaches using data mining tools for variable selection in literature. A comprehensive theoretical survey of variable selection approaches is reported in Liu and Yu [48], while a comparison of algorithms for that purpose is presented by Kudo and Sklansky [49]: the best method combines leave-one-out rule and k -nearest neighbor

(KNN) classification technique. A method for variable selection and classification of observations into two classes relying on PLS regression and kernel was proposed by Tenenhaus *et al.* [50]. Although simple and easy to implement, the performance of the proposed method is equivalent to more complex data mining tools, such as support vector machine (SVM). Additionally, Krzanowski [51] and Glen [52] use discriminant analysis to select variables from a mixture of continuous and discrete variables, while variable selection approaches are combined with discriminant analysis by Pacheco *et al.* [53].

Anzanello *et al.* [9] proposed a method to select variables for classification of production batches in two quality levels in chemical processes. The framework integrated PLS regression and data mining tools, as KNN, SVM, and probabilistic neural network. The proposed method remarkably reduced the number of variables needed for classification, whereas yielded significant more accurate categorizations. The method was also innovated by developing new indices of variable importance for the classification purpose: these indices guide the variable removal process according to its impact on classification accuracy. Another variable selection framework was proposed by Anzanello [54] aimed at improving clustering results in a shoe manufacturing process. The original set of variables consisted of two groups of variables: (i) a set of process experts' variables

subjectively assessing the product complexity, and (ii) a set of learning related variables represented by the parameters of the learning curve (for details on learning curves, see Teplitz [55]). Variables were selected using a leave-one-variable-out-at-a-time, and the performance of the clustering procedure evaluated by means of the silhouette index. Clustering performance was improved significantly by using a reduced combination of the two sets of variables.

Some classical approaches for variable selection arise from the food manufacturing industry. Several approaches use PCA to identify the variables with most variance and then apply tools for classification (e.g., data mining, discriminant and clustering techniques) using the selected variables [56,57]. A method to reduce dimensionality in industrial wine manufacturing was reported by Urtubia *et al.* [58]: PCA was first used to select the variables carrying most of the metabolite interaction information, and classes of similar behavior were then generated by applying *K*-means clustering technique on the lower-dimensional dataset. In Camara *et al.* [59], PCA was associated to discriminant analysis to differentiate and classify wines; coefficients of discriminant functions were used to identify the relevant variables.

A similar study was conducted by Rebolo *et al.* [60] for testing the authenticity of wines produced in a specific region. Variables describing chemical substances were initially selected using cluster analysis and PCA techniques, followed by a stepwise Bayesian analysis. The study also aimed at categorizing the wines in classes according to their origin. With identical purposes, Marini *et al.* [61] studied two classification techniques, that is, soft independent modeling of class analogies (SIMCA) and UNEQ, to verify the authenticity of Italian wines. SIMCA describes the similarities among the samples of a category using a PCA [62], while the UNEQ is similar to a quadratic discriminant analysis [63]. The identification of the relevant variables was performed by means of stepwise linear discriminant analysis. In Capron *et al.* [64], a wine dataset consisting of 63 process variables was evaluated using

decision trees and modifications on PLS regression to identify the most important variables aimed at classifying wines into four different classes (i.e., origin countries). Also, in the food industry, integrated PCA, KNN, and Crombach's alpha are used to identify both the most relevant sensorial variables and the most consistent panelists in sensorial evaluations. The proposed method significantly reduced the number of sensorial variables needed for categorization of food samples into eight classes, besides reducing the number of panelists needed for assessment.

Other interesting applications of variable selection approaches might not be closely related to manufacturing or chemometrics, but could be easily adapted to such purposes: Chen *et al.* [65] integrated flexible neural trees to GAs aiming at identifying the important variables in breast cancer classification, while Chaovalitwongse *et al.* [66] used the KNN classification technique on time-series data of patients with abnormal brain activity. King and Jackson [67] tested four methods of variable selection based on PCA and Procrustes analysis in large environment-related datasets. Finally, a method for selecting gene-related variables was proposed by Diaz-Uriarte and Andres [68]; the method used random forest technique, and had comparable performance for classification of methods such as KNN and SVM.

The multicriteria variable selection using data mining tools has also received crescent attention in recent years. In these applications, sensitivity, specificity, and precision, among other performance measures, are simultaneously used to evaluate the prediction/classification performance. Li and Li [69] used sensitivity and specificity to select variables in biomedical applications. The same criteria were used by Su and Yang [70] in combination with SVMs to help in diagnosis of hypertension. Kirkos *et al.* [71] propose the combination of decision trees, neural networks, and SVM to select the most important variables in economic applications, while Gaganis *et al.* [72] compare KNN technique with discriminant analysis in financial evaluations. Other applications can be found

Table 2. Summary of Approaches for Variable Selection with Classification Purposes

Data Mining Tools	Variable Selection for Classification			
	PCA	Discriminant Analysis	PLS regression	Clustering
Kudo and Sklansky [49]	Mallet <i>et al.</i> [56]	Krzanowski [51]	Tenenhaus <i>et al.</i> [50]	Urtubia <i>et al.</i> [58]
Tenenhaus <i>et al.</i> [50]	Guo <i>et al.</i> [57]	Glen [52]	Anzanello <i>et al.</i> [9]	Rebolo <i>et al.</i> [60]
Anzanello <i>et al.</i> [9]	Urtubia <i>et al.</i> [58]	Pacheco <i>et al.</i> [53]	Capron <i>et al.</i> [64]	
Capron <i>et al.</i> [64]	Camara <i>et al.</i> [59]	Marini <i>et al.</i> [61]		
Anzanello <i>et al.</i> [76]	Rebolo <i>et al.</i> [60]			
Chen <i>et al.</i> [65]	Anzanello <i>et al.</i> [76]			
Chaovalitwongse <i>et al.</i> [66]	King and Jackson [67]			
Diaz-Uriarte <i>et al.</i> [68]				
Li and Li [69]				
Su and Yang [70]				
Kirkos <i>et al.</i> [71]				
Gaganis <i>et al.</i> [72]				

in text recognition, business decisions, and security systems [10,69,73–75]. As seen, the application of multicriteria approaches for variable selection in manufacturing and chemometrics applications is very incipient, and research on that topic is been currently developed by this author.

Table 2 depicts a summary of the most cited approaches for variable selection with classification purposes based on their mathematical background. Studies combining two or more techniques are listed under all used techniques.

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

The variable selection problem became a capital issue in many industrial processes due to the high volume of data collected by sensors and experiments, leading to a broad variety of efficient methods to identify the most relevant variables. This article briefly described methods to select the most relevant variables on manufacturing and chemometrics applications. For convenience, we deployed the presentation into two classical purposes of variable selection: prediction and classification. Variable selection for prediction has been mostly accomplished by means

of the PLS regression and other statistical tools, while methods focused on classification have integrated data mining and discriminant analysis to dimension reducing techniques, such as PCA, among others.

It seems that the search for efficient combinations of statistical techniques dictates the trend for future variable selection methods. One such promising approaches might be the insertion of parameters from regression models (e.g., PLS, MLR, LR) into data mining and optimization tools to calibrate variables during the selection process. Such parameters will capture relevant information on the variance of process and product variables, which might be used to penalize irrelevant variables before applying the selection method itself.

Techniques of set pruning might also be helpful in datasets consisting of several hundreds of variables. A possible approach would consist of two phases: a first phase would filter the original dataset to eliminate variables that do not contribute on model construction. A second phase would focus on the identification of the best subset among the remaining variables using either optimization techniques, for example, GAs, or approaches based on the leave-one-out-at-a-time analysis, among others. These approaches tend to be computationally more efficient than using the entire original set,

and could avoid the occurrence of local optimal solutions.

Efforts on multicriteria variable selection also pose a challenging proposition; methods will have to select variables assessing simultaneous criteria, for example, accuracy performance and costs of collecting/measuring variables. These methods should also incorporate experts' assessment on the selection process.

Finally, the availability of so many variable selection procedures and fields of application has offered some pitfalls for researchers and practitioners, as reported in Ref. 77. Many methods have yielded remarkable results when applied to specific fields, but failed on slightly different processes by misleading the variable selection and yielding conflicting results. This has caused some practitioners to use old-fashioned variable selection techniques, such as SMR and AIC, instead of applying the most recent methods. Studies on the robustness of recently proposed variable selection methods are mandatory to better understand their properties and ensure general application.

REFERENCES

1. Kettaneh N, Berglund A, Wold S. PCA and PLS in very large datasets. *Comput Stat Data Anal* 2005;48:69–85.
2. Reunanen J. Overfitting in making comparisons between variable selection methods. *J Mach Learn Res* 2003;3:1371–1382.
3. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;3:1157–1182.
4. Gauchi J, Chagnon P. Comparison of selection methods of exploratory variables in PLS regression with application to manufacturing process data. *Chemom Intell Lab Syst* 2001;58:171–193.
5. Xu L, Zhang W. Comparison of different methods for variable selection. *Anal Chim Acta* 2001;446:477–483.
6. Chong I, Jun C. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst* 2005;78:103–112.
7. Marengo E, Robotti E, Bobba M, *et al.* Application of partial least squares discriminant analysis and variable selection procedures: a 2D-PAGE proteomic study. *Anal Bioanal Chem* 2008;390(5):1327–1342.
8. Liu F, Jiang Y, He Y. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer. *Anal Chim Acta* 2009;635(1):45–52.
9. Anzanello M, Albin S, Chaovalitwongse W. Selecting the best variables for classifying production batches into two quality levels. *Chemom Intell Lab Syst* 2009;97(2):111–117.
10. Huang C, Wang C. A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst Appl* 2006;31(2):231–240.
11. Unser M, Eden M. Multiresolution feature extraction and selection for texture segmentation. *IEEE Trans Pattern Anal Mach Intell* 1998;11:717–728.
12. Duda R, Hart P, Stork D. *Pattern classification*. 2nd ed. New York: Wiley-Interscience; 2001.
13. Nixon M, Aguado A. *Feature extraction and image processing*. Hungary: Academic Press; 2008.
14. Morris J, Coombes K, Koomen J, *et al.* Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 2005;21(9):1764–1775.
15. Rencher A. *Methods of multivariate analysis*. New York: Wiley; 1995.
16. Montgomery D, Peck E, Vining G. *Introduction to linear regression analysis*. New York: Wiley; 2001.
17. Kutner M, Nachtsheim C, Neter J. *Applied linear regression models*. Boston (MA): McGraw-Hill; 2004.
18. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Control* 1974;19(6):716–723.
19. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6(2):461–464.
20. Nelson P, MacGregor J, Taylor P. The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemom Intell Lab Syst* 2006;80:1–12.
21. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 2001a;58:109–130.
22. Wold W, Trygg J, Berglund A, *et al.* Some recent developments in PLS modeling. *Chemom Intell Lab Syst* 2001b;58:131–150.

23. Almoy T. A simulation study on comparison of prediction methods when only a few components are relevant. *Comput Stat Data Anal* 1996;21:87–107.
24. Dingstad G, Westad F, Naes T. Three case studies illustrating the properties of ordinary and partial least squares regression in different mixture models. *Chemom Intell Lab Syst* 2004;71:33–45.
25. Wang D, Srinivasan R, Liu J, *et al.* Data-driven soft sensor approach for quality prediction in a refinery process. *Proc IEEE Int Conf Ind Inform* 2006:230–235.
26. Hoskuldsson A. Variable and subset selection in PLS regression. *Chemom Intell Lab Syst* 2001;55:23–38.
27. Xu L, Jiang J, Wu H, *et al.* Variable-weighted PLS. *Chemom Intell Lab Syst* 2007;85:140–143.
28. Baumann K, Alberto H, Korff M. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *J Chemom* 2002;16:339–350.
29. Zhai H, Chen X, Hu Z. A new approach for the identification of important variables. *Chemom Intell Lab Syst* 2006;80:130–135.
30. Sarabia L, Ortiz M, Sanchez A. Dimension wise selection in partial least squares regression a bootstrap estimated signal-noise relation to weight the loadings. *PLS and Related Methods, Proceedings of the PLS'01 International Symposium*. Paris: CISIA-CERESTA Editeur; 2001. pp. 327–339.
31. Forina M, Casolino C, Millan C. Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Chemom Intell Lab Syst* 1999;13:165–184.
32. Zarzo M, Ferrer A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. *Chemom Intell Lab Syst* 2004;73(1):15–27.
33. Lazraq A, Cleroux R, Gauchi J. Selecting both latent and exploratory variables in the PLS1 regression model. *Chemom Intell Lab Syst* 2003;66:117–126.
34. Leardi R, Seasholtz M, Pell R. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Anal Chim Acta* 2002;461(2):189–200.
35. Duranda A, Devosa O, Ruckebusch C, *et al.* Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles. *Anal Chim Acta* 2007;595(1–2):72–79.
36. Ye S, Wang D, Min S. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemom Intell Lab Syst* 2008;91(2):194–199.
37. Teofilo R, Martins J, Ferreira M. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J Chemom* 2009;23(1):32–48.
38. Daszykowski M, Walkzak B, Massart D. Projection methods in chemistry. *Chemom Intell Lab Syst* 2005;65:97–112.
39. Roy P, Roy K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci* 2008;27(3):302–313.
40. Liu S, Liu H, Yin C, *et al.* VSMP: a novel variable selection and modeling method based on the prediction. *J Chem Inf Model* 2003;43(3):964–969.
41. Gonzalez M, Teran C, Saiz-Urra L, *et al.* Variable selection methods in QSAR: an overview. *Curr Top Med Chem* 2008;8(18):1606–1627.
42. Gualdron O, Liobet E, Brezmes J, *et al.* Fast variable selection for gas sensing applications. *Proc IEEE Sensors* 2004: 892–895.
43. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser A Stat Soc* 1996;58(1):267–288.
44. Fu W. Penalized regressions: the bridge versus the lasso. *J Comput Graph Stat* 1998;7(3):397–416.
45. Chatterjee S, Laudato M, Lynch L. Genetic algorithms and their statistical applications: an introduction. *Comput Stat Data Anal* 1996;22:633–651.
46. Gualdron O, Torres I. Fast variable selection for gas sensing applications. *Proceedings of the Electronics, Robotics and Automotive Mechanics Conference*. Morelos, Mexico: CERMA; 2007. pp. 151–158.
47. Philips R, Guttman I. A new criterion for variable selection. *Stat Probab Lett* 1998;38:11–19.
48. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans Knowl Data Eng* 2005;17(4):491–502.

49. Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognit* 2000;33:25–41.
50. Tenenhaus A, Giron A, Viennet E, *et al.* Kernel logistic PLS: a tool for supervised non-linear dimensionality reduction and binary classification. *Comput Stat Data Anal* 2007; 51:4083–4100.
51. Krzanowski W. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Comput Stat Data Anal* 1995;19:419–431.
52. Glen J. An additive utility mixed integer programming model for nonlinear discriminant analysis. *J Oper Res Soc* 2008;59:1492–1505.
53. Pacheco J, Casado S, Nunez L, *et al.* Analysis of new variable selection methods for discriminant analysis. *Comput Stat Data Anal* 2006;51(3):1463–1478.
54. Anzanello M. Selecting relevant clustering variables in mass customization scenarios characterized by workers learning. In: Fogliatto FS, Da Silveira G, editors. *Mass customization: engineering and managing global operations*. New York: Springer; 2010. In press.
55. Teplitz C. *The learning curve deskbook: a reference guide to theory, calculations and applications*. New York: Quorum Books; 1991.
56. Mallet Y, De Vel O, Coomans D. Integrated feature extraction using adaptive wavelets. In: Liu H, Motoda H, editors. *Feature extraction, construction and selection: a data mining perspective*. Massachusetts: Kluwer Academic Publishers; 1998. pp. 175–189.
57. Guo Q, Wu W, Massart D, *et al.* Feature selection in principal component analysis of analytical data. *Chemom Intell Lab Syst* 2002;61:123–132.
58. Urtubia A, Perrez-Correa J, Soto A, *et al.* Using data mining techniques to predict industrial wine problem fermentation. *Food Control* 2007;18:1512–1517.
59. Camara J, Alves M, Marques J. Multivariate analysis for the classification and differentiation of Madeira wines according to the main grape varieties. *Talanta* 2006;68:1512–1521.
60. Rebolo S, Pena R, Latorre M, *et al.* Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. *Anal Chim Acta* 2000;417:211–220.
61. Marini F, Bucci R, Magri A, *et al.* Authentication of Italian CDO wines by class-modeling techniques. *Chemom Intell Lab Syst* 2006;84:164–171.
62. Wold S, Sjostrom M. In: Kowalski BR, editor. *Chemometrics, theory and application*, ACS symposium series 52. Washington, DC.: American Chemical Society; 1977. pp. 243–282.
63. Derde M, Massart D. Supervised pattern recognition: the ideal method? *Anal Chim Acta* 1986;184:33–51.
64. Capron X, Smeyers-Verbeke J, Massart D. Multivariate determination of the geographical origin of wines from four different countries. *Food Chem* 2007;101:1585–1597.
65. Chen Y, Abraham A, Yang B. Feature selection and classification using flexible neural tree. *Neurocomputing* 2006;70(1–3): 305–313.
66. Chaovalitwongse W, Fan Y, Sachdeo C. On the time series k-nearest neighbor classification of abnormal brain activity. *IEEE Trans Syst Man Cybern A* 2007;37:1005–1016.
67. King J, Jackson D. Variable selection in large environmental data sets using principal components analysis. *Environmetrics* 1999;10(1):67–77.
68. Diaz-Uriarte R, Andres S. Gene selection and classification of microarray data using random forest. *BMC Bioinform* 2006;7(3):1–13.
69. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinform* 2008;24(9): 1175–1182.
70. Su C, Yang C. Feature selection for the SVM: an application to hypertension diagnosis. *Expert Syst Appl* 2008;34(1):754–763.
71. Kirkos E, Spathis C, Manopoulos Y. Support vector machines, decision trees and neural networks for auditor selection. *J Comput Methods Sci Eng* 2008;8(3):213–224.
72. Gaganis C, Pasiouras F, Spathis C, *et al.* A comparison of nearest neighbors, discriminant and logit models for auditing decisions. *Intell Syst Account Finance Manag* 2007;15(1–2):23–40.
73. Rose-Pehrsson S, Shaffer R, Hart S, *et al.* Multi-criteria fire detection systems using a probabilistic neural network. *Sens Actuators B* 2000;69(3):325–335.
74. Doan S, Horiguchi S. An efficient feature selection using multi-criteria in text categorization. *Proceedings of the 4th International Conference on Hybrid Intelligent Systems; 2004, HIS '04*. 2004. pp. 86–91.
75. Pendaraki K, Zopounidis C, Doumpos M. On the construction of mutual fund portfolios: a

- multicriteria methodology and an application to the Greek market of equity mutual funds. *Eur J Oper Res* 2005;163(2):462–481.
76. Anzanello M, Fogliatto F, Rossini K. Data mining-based method for attribute selection in sensory profiling. *Food Qual Prefer* 2010. In press.
77. George E. The variable selection problem. *J Am Stat Assoc* 2000;95:1–12.

FICTITIOUS PLAY ALGORITHM

PABLO ZAFRA
Department of Mathematics,
Kean University, Union,
New Jersey

While a linear programming (LP) formulation may allow us to find an exact solution to matrix games, it is not always necessary (or practical) to find this exact solution particularly for a large payoff matrix (see **Linear Programming and Two-Person Zero-Sum Games**). There are game-theoretic methods that have been proposed: the numerical method of von Neuman [1], the differential equation approach of Brown and von Neuman [2], and the fictitious play method of Brown [3]. Brown's fictitious play method is particularly appealing because of its simplicity, which does not require heavy computation or calculation. It is an iterative procedure that provides an estimate of a solution to matrix (two-person zero-sum) games, albeit on a rather very slow rate.

When first proposed, the method was not known to converge, but it was applied to a few games with success. Convergence was proved soon after by Robinson [4], who showed that the strategies generated by fictitious play for both players yield game value approximations that converge to the true value of the game as the number of plays (iterations) becomes infinite. She did not, however, provide a rate of convergence. Brown suggested in his paper that the rate could be of order $O(1/k)$, but further computational experiments by Zafra and Peterson [5] suggests a slower rate of order $O(\sqrt{k})$, which follows Szép and Forgó's conjecture [6].

Gass *et al.* [7] later introduced an improvement to the fictitious play method. Their modified fictitious play (MFP) method exhibits a convergence rate that is much superior to the standard form. Washburn [8] provided further refinements of MFP that appear to converge even faster.

Although Brown's original method was developed for solving matrix games, various forms of fictitious play algorithms arose in the literature as researchers applied the concept to different classes of games such as nonzero-sum, n -person, and continuous-type games. A survey of the literature on these variants is presented in the section titled "Fictitious Play and Other Games."

BROWN'S FICTITIOUS PLAY METHOD

We assume an $(m \times n)$ payoff matrix $A = (a_{ij})$, where the a_{ij} represents the payment of P2 to P1. Denote the i th row of A by A_i and the j th column of A by A^j . Let $X = (x_1, \dots, x_i, \dots, x_m)$, $\sum x_i = 1$, $x_i \geq 0$ be a mixed strategy vector for P1, and let $Y = (y_1, \dots, y_j, \dots, y_n)$, $\sum y_j = 1$, $y_j \geq 0$ be a mixed strategy vector for P2. If some $x_i = 1$ ($y_j = 1$), then the corresponding strategy vector for P1 (P2) is termed a *pure strategy* and is denoted by X_i (Y_j).

Brown's iterative fictitious play method can be viewed as a statistician's way of solving matrix games, where a player will base future moves on his opponent's historical tendencies. As Brown pointed out, the method "... rests on the traditional statistician's philosophy of basing future decisions on the relevant past history." The procedure can be characterized as follows [9]:

Two players, perhaps unaware of the minimax theorem, engage in playing the same two-person zero-sum game repeatedly. They apply a greedy policy for choosing a particular pure strategy in each play which is based on the following intuitive consideration: each player chooses a strategy that is best against his opponent's past actions, relying on a statistician's belief that future actions will resemble the past. The relative frequencies of the pure strategies used in the plays provide estimates for the mixed strategies.

We now describe the method formally.

Let k be the index count of the number of iterations; we start with $k = 0$. For any

$(m \times n)$ matrix game A , we define two vectors $U(k) = [U_1(k), \dots, U_i(k), \dots, U_m(k)]$ and $V(k) = [V_1(k), \dots, V_j(k), \dots, V_n(k)]$, with $U(k)$ a column vector and $V(k)$ a row vector. U and V are the payoff vectors for P1 and P2, respectively. The process is initiated with $U(0) = (0, \dots, 0)$ and $V(0) = (0, \dots, 0)$. For $k = 1$, P1 arbitrarily plays the i th row and we form $V(1) = V(0) + A_i$. P2 would respond by playing column j , where $j = \operatorname{argmin}[V_1(1), \dots, V_n(1)]$. We then form $U(1) = U(0) + A^j$. This completes the first iteration.

In subsequent iterations k , P1 plays row i corresponding to $i(k) = \operatorname{argmax}[U_1(k-1), \dots, U_m(k-1)]$ and P2 plays column j corresponding to $j(k) = \operatorname{argmin}[V_1(k), \dots, V_n(k)]$. The vectors U and V are updated by the formulas

$$\begin{aligned} U(k) &= U(k-1) + A^{j(k)}, \\ V(k) &= V(k-1) + A_{i(k)}. \end{aligned}$$

For any $k \geq 1$, we see that the vectors $U(k)$ and $V(k)$ are the cumulative payoffs to P1 and P2, respectively, against the other player's set of iterative strategies. That is, if $a_i(k)$ is the number of times P1 has played strategy i and $b_j(k)$ is the number of times P2 has played strategy j , then

$$\begin{aligned} U(k) &= \sum_j b_j(k) A^j, \\ V(k) &= \sum_i a_i(k) A_i. \end{aligned}$$

For each iteration, the fictitious play method directs P1 to select row i that maximizes the payoff against P2's mixed strategy to date, while P2 would respond by selecting column j which minimizes his payout to P1. Ties are broken arbitrarily. At the end of the k th iteration, the mixed strategy vectors for P1 and P2 are $X(k) = (a_1(k)/k, \dots, a_m(k)/k)$ and $Y(k) = (b_1(k)/k, \dots, b_n(k)/k)$, respectively. Robinson's convergence theorem [4] showed that, for k large enough, $X(k)$ and $Y(k)$ approximate the optimal strategies. Moreover, note that for the vector $U(k)/k$, its maximum component, say $M(k)$, is an upper-bound estimate for v , since P2 can

select any strategy among the columns j and keep P1's expected earning to be at most $M(k)$. While for the vector $V(k)/k$, its minimum component, say $m(k)$, is a lower-bound estimate for v , since P1 can select any strategy among the rows i and make P2's expected payout to be at least $m(k)$. That is, $m(k) \leq v \leq M(k)$ [10].

Since convergence is not finite, the method terminates when (i) the specified iteration limit has been reached, or (ii) the desired precision for the game value error estimate (the difference between $m(k)$ and $M(k)$) has been achieved.

Example. To demonstrate Brown's fictitious play method, we reproduce in Table 1 the payoff matrix from his paper [3]:

Using a transformation to a linear program [11] (see **Linear Programming and Two-Person Zero-Sum Games**), we found that the game has value 1.441667 with the optimal mixed strategies $X = (0, 0.12037, 0.237037, 0.642593)$ and $Y = (0.083333, 0.666667, 0.25)$ for P1 and P2, respectively.

Summary of calculations for the cumulative payoffs for the first 25 steps of the algorithm are shown in Table 2. Note that $\Delta(k) = M(k) - m(k)$ in the table.

At $k = 0$, we initialize $U(0) = (0, \dots, 0)$ and $V(0) = (0, \dots, 0)$. To start constructing the table with the iteration $k = 1$, suppose P1 arbitrarily plays row 2 ($i = 2$). We form the payoff vector $V(1) = (1.3, 2, 0)$ from which P2 responds by playing column 3 ($j = 3$) to minimize his payout. We then form the payoff vector $U(1) = (1.2, 0, \underline{3.1}, 1.1)$. This completes the first iteration. Note that the underlined components in both vectors mentioned above will be used to calculate the lower and upper bound values of the game value. For

Table 1. Brown's Game Matrix

	$j = 1$	$j = 2$	$j = 3$
$i = 1$	3	1.1	1.2
$i = 2$	1.3	2	0
$i = 3$	0	1	3.1
$i = 4$	2	1.5	1.1

Table 2. Cumulative Payoffs

k	i	$j = 1$	$j = 2$	$j = 3$	$m(k)$	$M(k)$	$\Delta(k)$	j	$i = 1$	$i = 2$	$i = 3$	$i = 4$
1	2	1.3	2	0	0.00	3.10	3.10	3	1.2	0	3.1	1.1
2	3	1.3	3	3.1	0.65	2.10	1.45	1	4.2	1.3	3.1	3.1
3	1	4.3	4.1	4.3	1.37	1.77	0.40	2	5.3	3.3	4.1	4.6
4	1	7.3	5.2	5.5	1.30	1.60	0.30	2	6.4	5.3	5.1	6.1
5	1	10.3	6.3	6.7	1.26	1.52	0.26	2	7.5	7.3	6.1	7.6
6	4	12.3	7.8	7.8	1.30	1.55	0.25	2	8.6	9.3	7.1	9.1
7	2	13.6	9.8	7.8	1.11	1.46	0.35	3	9.8	9.3	10.2	10.2
8	3	13.6	10.8	10.9	1.35	1.46	0.11	2	10.9	11.3	11.2	11.7
9	4	15.6	12.3	12.0	1.33	1.59	0.26	3	12.1	11.3	14.3	12.8
10	3	15.6	13.3	15.1	1.33	1.53	0.20	2	13.2	13.3	15.3	14.3
11	3	15.6	14.3	18.2	1.30	1.48	0.18	2	14.3	15.3	16.3	15.8
12	3	15.6	15.3	21.3	1.27	1.44	0.17	2	15.4	17.3	17.3	17.3
13	2	16.9	17.3	21.3	1.30	1.48	0.18	1	18.4	18.6	17.3	19.3
14	4	18.9	18.8	22.4	1.34	1.49	0.15	2	19.5	20.6	18.3	20.8
15	4	20.9	20.3	23.5	1.35	1.51	0.16	2	20.6	22.6	19.3	22.3
16	2	22.2	22.3	23.5	1.39	1.52	0.13	1	23.6	23.9	19.3	24.3
17	4	24.2	23.8	24.6	1.40	1.52	0.12	2	24.7	25.9	20.3	25.8
18	2	25.5	25.8	24.6	1.37	1.49	0.12	3	25.9	25.9	23.4	26.9
19	4	27.5	27.3	25.7	1.35	1.47	0.12	3	27.1	25.9	26.5	28.0
20	4	29.5	28.8	26.8	1.34	1.48	0.14	3	28.3	25.9	29.6	29.1
21	3	29.5	29.8	29.9	1.40	1.49	0.09	1	31.3	27.2	29.6	31.1
22	1	32.5	30.9	31.1	1.40	1.48	0.08	2	32.4	29.2	30.6	32.6
23	4	34.5	32.4	32.2	1.40	1.47	0.07	3	33.6	29.2	33.7	33.7
24	3	34.5	33.4	35.3	1.39	1.47	0.08	2	34.7	31.2	34.7	35.2
25	4	36.5	34.9	36.4	1.40	1.47	0.07	2	35.8	33.2	35.7	36.7

this iteration $k = 1$, we have $m(1) = 0/1 = 0$, which is obtained by using the minimum component of $V(1)$, and $M(1) = 3.1/1 = 3.1$, which is obtained by using the maximum component of $U(1)$. Recall that $m(k)$ and $M(k)$ provide the lower bound and upper bound values for the game value, respectively.

In the next iteration ($k = 2$), P1 looks at $U(1)$ and plays row 3 ($i = 3$) to maximize his reward. The vector $V(2) = V(1) + A_3 = (1.3, 2, 0) + (0, 1, 3.1) = (1.3, 3, 3.1)$ is thus formed. P2 responds by playing column 1 ($j = 1$) to form the vector $U(2) = U(1) + A^1 = (1.2, 0, 3.1, 1.1) + (3, 1.3, 0, 2) = (4.2, 1.3, 3.1, 3.1)$. This completes the second iteration. Note that we now have $m(2) = 1.3/2 = 0.65$, the minimum component of $V(2)/2$, and $M(2) = 4.2/2 = 2.10$, the maximum component of $U(2)/2$. This completes the second iteration $k = 2$ as seen in the table.

The step is then repeated until the desired index is reached ($k = 25$ for this example) or

a desired accuracy is achieved in regards to $\Delta(k)$, the difference between $M(k)$ and $m(k)$. After 25 iterations, we see that $m(25) = 1.40$ and $M(25) = 1.47$. Observe that the values of $m(k)$ and $M(k)$ are not monotonic and they jog around the game value of 1.441667. This illustrates a typical behavior of the method.

After 25 iterations, the estimate for P1's mixed strategy can also be calculated by $X(25) = (4/25, 5/25, 7/25, 9/25)$, since row $i = 1$ has been played 4 times; row $i = 2$, 5 times; row $i = 3$, 7 times; and row $i = 4$, 9 times. For P2, the estimate for the mixed strategy is given by $Y(25) = (4/25, 14/25, 7/25)$, since column $j = 1$ has been played 4 times; $j = 2$, 14 times; and $j = 3$, 7 times. With only 25 iterations applied so far, observe that these are indeed very rough estimates of the optimal mixed strategies mentioned earlier. However, if large number of iterations are executed, say 10,000 iterations, the mixed strategies obtained are given by

$X = (0.0015, 0.1183, 0.2354, 0.6448)$ and $Y = (0.0843, 0.6652, 0.2505)$, which are fairly close to the optimal strategies.

Moreover, after 25 iterations, we find that $m(k)$ and $M(k)$ approach the game value somewhat rapidly with $m(25) = 1.40$ and $M(25) = 1.47$. However, at the end of 10,000 iterations, we found that $m(10,000) = 1.44071$ and $M(10,000) = 1.44192$ which gives error precision of only about 0.001, a slight improvement over the error 0.07 after only 25 iterations.

This illustrates another typical behavior of the algorithm, which gives seemingly fast convergence during early iterations but tends to drag later on. The convergence is painfully slow if extreme precision is desired. However, the method does provide a good approximation if high accuracy is not required, which is often true in practical applications.

CONVERGENCE AND MODIFICATIONS

Although the method has been proved to converge by Robinson, the rate of convergence is very slow. Shapiro [10] gives a rate of convergence estimate for fictitious play in terms of how fast $\Delta(k)$ goes to zero as k goes to infinity. His *a priori* estimate of the rate of convergence is of order $O(k^{-1/(m+n-2)})$ for an $(m \times n)$ general payoff matrix, which is rather loose. Shapiro suggested that this convergence rate could be improved. Brown actually suggested in his original paper that the rate could be of order $O(1/k)$. Szep and Forgo [6] later conjectured that the rate of convergence for a general game can be sharpened to $O(k^{-1/2})$. Further computational experiments by Zafra and Peterson [5] are consistent with Szep and Forgo's conjecture.

The first modification of the method was mentioned by Robinson in her proof [4], where she constructs a description of the fictitious method with the two players updating their beliefs (i.e., pure strategies) simultaneously rather than in an alternating manner. She then provided Brown's proposed method as an *alternate notion* of her vector system. However, she clearly pointed out that her proofs and results hold for both versions of the fictitious play method. It

turns out that this "simultaneous" version of the fictitious play is not as fast in empirical experiments compared to Brown's originally proposed sequence of alternate plays [5].

Gass *et al.* (GZQ) [7] later introduced a simple modification of Brown's method for symmetric games that greatly improved the convergence rate. If the original game is not symmetric, the basic idea is to transform the given general matrix game into an equivalent symmetric game (a game with a skew-symmetric matrix, that is, $a_{ij} = -a_{ji}$ for all i, j) and use the solution properties of symmetric games (the game value is zero and both players have the same optimal strategies). The symmetrization is done by using Gale, Kuhn, and Tucker's transformation matrix below [12]:

$$S = \begin{bmatrix} 0 & A & -1 \\ -A^T & 0 & 1 \\ 1^T & -1^T & 0 \end{bmatrix},$$

where A^T is the transpose of the given general $(m \times n)$ payoff matrix, 1 is a column vector of 1's with the appropriate number of components, and 0 is composed of zeros to fill out the matrix.

The fictitious play method is then applied to the enlarged skew-symmetric matrix with a modification that calls for the periodic restarting of the process. At restart, both players' strategies are made equal based on the following considerations: Select the maximizing or minimizing players' strategy that has a game value closest to zero. For both symmetric and general games, their modified fictitious play (MFP) procedure, where only the more successful player's strategy is retained, approximates the value of the game and optimal strategies in a greatly reduced number of iterations and in less computational time when compared to Brown's original method.

Limsakul [13] proposed a dynamic approach to the MFP method that involves a generalization of the matrix S above. It should first be noted that the solution to S requires that the matrix game A has a positive value. To meet this requirement, GZQ simply added a constant to all the entries of A . Although this *static* addition ensures

positive game value, they also suggested that a dynamic approach would produce better results. Following this suggestion, Limsakul used a dynamic procedure where the current best lower bound on the value of the game is subtracted from every entry. Note that this lower bound is necessarily negative because S is a symmetric game, and therefore subtracting this negative lower bound from the matrix entries would result in a modified matrix that has a small but nonnegative value. He reported a better computational experience compared to MFP.

Washburn [8] further extended and improved MFP by introducing not only a generalization of the matrix S but also a *minimal* symmetrization matrix B that omits the last row and column of S . The players of the game matrix B would then employ his improved MFP, in conjunction with a dynamically adjusted version of this minimal symmetric game transformation. This different symmetrization appears to be very efficient in solving for approximate solutions of general matrix games, particularly when a more accurate estimation is required.

FICTITIOUS PLAY AND OTHER GAMES

Brown originally proposed the fictitious play method to solve two-person zero-sum games. However, it can also be interpreted as a learning process whereby two (or more) players take turns playing the same finite game repeatedly, and where each player chooses a pure strategy in every turn that is a best response against the accumulated (historical) mixed strategy of the other player(s) from previous plays. This extended notion of fictitious play as a learning process has been applied to different types of games with mixed results in terms of convergence.

First, most of the research done by game theorists tried to identify classes of games where convergence holds. Berger [14] recently compiled a list of works in the literature where the fictitious play method converges. Works showing successful convergence of the method include: Miyasawa's (2×2) nonzero-sum game [15], geometric convergence proof of Miyasawa's example

[16], supermodular games with unique equilibria [17], supermodular games with diminishing returns [18], multiplayer games with identical player utilities [19], (3×3) games with strategic complementarities [20], (2×3) games [21], $(2 \times n)$ games [22], and continuous concave-convex zero-sum games [23]. More recent additions among convergent games include Berger's ordinal potential games and quasi-supermodular games with diminishing returns [24]; and his work on games with strategic complementarities [25]. Shamma and Arslan [26] also offered a "unified convergence proofs of continuous-time fictitious play" for certain special cases including zero-sum games, identical interest games, and (2×2) games.

Some of the authors above have extended Brown's discrete method to continuous version as he described in his original paper [27], and most of them have employed Robinson's modification of fictitious play [4] where players update their beliefs simultaneously. But Berger revived Brown's original fictitious play process in 2007 [14], where players update their beliefs in an alternating manner rather than simultaneously, and demonstrated that it converges to a large class of games called *nondegenerate ordinal potential games*. Convergence results are also available for other variants of fictitious play method including stochastic [28,29], (generalized) weakened form [30,31] and best response dynamics [32,33].

Although most research centered on identifying special classes of games where fictitious play converges, others have provided counterexamples demonstrating failure to converge. While Miyasawa's example showed convergence as applied to (2×2) nonzero-sum game, it has been quickly established that convergence need not apply to general nonzero-sum games due to Shapley [34]. Shapley showed a (3×3) example where the method follows a cycling process which exhibits runs of pure strategy selections that were repeated over and over again, and the lengths of runs also increase exponentially. Mixed strategies also cycled and stayed away from the equilibrium point. Sparrow *et al.* [35] recently generalized

Shapley's example by introducing a parameter that exhibits a chaotic behavior of the fictitious play. Another example was given by Foster and Young [36], who constructed an (8×8) coordination game where fictitious play did not converge to any equilibrium point. Other counterexamples showing that the method need not converge include three problems due to Jordan [37], three examples in Gaunersdorfer and Hofbauer [38], a degenerate (2×2) supermodular game with diminishing returns [39], and games with a unique and completely mixed strategy equilibrium where players use more than two pure strategies [40].

APPLICATIONS

While there are many research works available on various implementations, adaptations, modifications, and extensions of fictitious play as applied to different classes of games, very few applications regarding its use in solving problems from other disciplines or real-world problems are available from the literature. In this section, we mention some interesting and potential applications of fictitious play.

Gass and Zafra [41] described how their MFP can be used as a crash start to LP problems, namely, to combine MFP with a mathematical programming system (MPS), whether simplex-based or interior-point method. When MFP was applied to a game that was for an equivalent LP problem, it did not take many iterations to identify the variables that would be positive in the optimal LP solution. The information could then be used to form a crash basis for starting a simplex-based MPS. Similar ideas are used to generate a starting solution for interior-point methods, and in combining the MFP solution with an interior-point solution to find an optimal basic solution to a linear programming problem. Their experiments showed encouraging results in using MFP as an efficient tool for aiding in the solution of LP problems.

Garcia *et al.* [42] applied the fictitious play concept as a new procedure to find system optimal routings in dynamic traffic

networks. In formulating dynamic traffic networks as N -person games with identical interests, vehicles are treated as players and the average trip time experienced in the network as the common payoff function. Given the route choices of all vehicles, a routing mapping assigns a time-dependent shortest path from origin to destination for every vehicle. In game-theoretic viewpoint, this routing assignment is interpreted as a *best reply* of each vehicle in regard to the routing decisions of the other vehicles. Fictitious play is therefore implemented as an iterative routing-assignment algorithm such that, at each step, the time-dependent shortest paths are computed for each player based on the historical frequency of the other players' earlier routing decisions. This "decentralized" approach in producing system optimal routings yield encouraging results in their large-scale empirical experiments.

Motivated by Garcia's fictitious play approach to solve optimization problems, Lambert and Wang [43] further demonstrated its effectiveness as a heuristic optimization method. To test the fictitious play approach, they used the large-scale situational awareness simulation on "Low-Energy Electronic Design for Mobile Platforms." In this problem, mobile units keep track of each other's location over a specified period of time. The units use communication devices to broadcast their location information, and a fictitious play approach was used in designing the communication protocol. They found that the fictitious play solutions are similar to the simulated annealing approach, with comparable computational efforts and outperforming pure random search.

Lambert *et al.* [44] developed a sampled version of the fictitious play algorithm for games with identical payoffs. Calling it *sampled fictitious play*, it is a formalized approach to Garcia's fictitious play algorithm, which has also been applied earlier to their 2003 work on mobile unit situation awareness problem. They proposed this heuristic to solve any large-scale discrete optimization problems of the form $\max\{u(y) : y = (y^1, \dots, y^n) \in$

$Y^1 \times \dots \times Y^n$, where the feasible region is the Cartesian product of finite sets $Y^i, i = 1, \dots, n$. For games with identical payoff functions for players, fictitious play is viewed as a coordinate-search algorithm for optimization. They describe a way to avoid the prohibitively expensive calculations in each iteration, which are required of the best reply strategy, with an approximation based on sampling. Computational experiments were performed with their sampled fictitious play method using the dynamic traffic routing assignment and mobile units situation awareness problems as mentioned above.

On a different application, Hon-Snir *et al.* [45] analyzed a repeated first-price sealed-bid auction of a single item where the player types are known in advance. The paths generated by players are analyzed, where each player uses various classes of belief-based learning process including a generalized fictitious play scheme. In the generalized fictitious play, a player assigns a probability to the other players' bid vectors with the assumption that the opponents' next bid vector is determined from the historical (empirical) distribution of the past bid vectors. Under mild assumptions, they showed that, after a sufficiently long time, the players play an equilibrium of the one-shot auction where player's types are known. In effect, they showed that under their belief-based learning system, a repeated first-price auction behaves like a one-shot second-price auction in the long run.

REFERENCES

1. von Neumann J. A numerical method to determine optimum strategy. *Nav Res Log Q* 1954;1:109–115.
2. Brown GW, von Neumann J. Solutions of games by differential equations. In: Kuhn HW, Tucker AW, editors. Volume 1, Contributions to the theory of games, *Annals of Mathematics Study No. 24*. Princeton (NJ): Princeton University Press; 1950. pp. 73–79.
3. Brown GW. Iterative solutions of games by fictitious play. In: Koopmans TC, editor. *Activity analysis of production and allocation*, Cowles Commission Monograph 13. New York: Wiley; 1951. pp. 377–380.
4. Robinson J. An iterative method of solving a game. *Ann Math* 1951;54(2):296–301.
5. Peterson K. Methods of fictitious play [Master's Thesis] (Zafra P, advisor). Kean University Union, NJ; 1998.
6. Szép J, Forgó F. Introduction to the theory of games. Holland: D. Reidel Pub. Co.; 1985.
7. Gass SI, Zafra PMR, Qiu Z. Modified fictitious play. *Nav Res Log* 1996;43:955–970.
8. Washburn A. A new kind of fictitious play. *Nav Res Log* 2001;48:270–280.
9. Zafra P. Theoretical and computational advances in the fictitious play method and some applications to the solution of linear programming problems [PhD dissertation (Gass SI, advisor)]. College Park (MD): University of Maryland; 1993.
10. Shapiro HN. Note on a computation method in the theory of games. *Commun Pure Appl Math* 1958;XI:587–593.
11. Gass SI. Linear programming. 5th ed. New York: McGraw-Hill Book Company; 1985.
12. Gale D, Kuhn HW, Tucker AW. On symmetric games. In: Kuhn HW, Tucker AW, editors. Volume 1, Contributions to the theory of games, *Annals of Mathematics Study No. 24*. Princeton (NJ): Princeton University Press; 1950. pp. 73–79.
13. Limsakul P. Iterative algorithms for two-person zero-sum games [Master's Thesis (Washburn A, advisor)]. Naval Postgraduate School; 1999.
14. Berger U. Brown's original fictitious play. *J Econ Theory* 2007;135(1):572–578.
15. Miyazawa K. On the convergence of the learning process in a 2×2 non-zero-sum two-person game. *Econometric Research Program*, Princeton, University, Research Memorandum No. 33; 1961.
16. Metrick A, Polak B. Fictitious play in 2×2 games: a geometric proof of convergence. *Econ Theory* 1994;4:923–933.
17. Milgrom P, Roberts J. Adaptive and sophisticated learning in normal form games. *Games Econ Behav* 1991;3:82–100.
18. Krishna V. Learning in games with strategic complementarities. HBS Working Paper 92-073. Harvard University; 1992.
19. Monderer D, Shapley LS. Fictitious play property for games with identical interests. *J Econ Theory* 1996;68:258–265.
20. Hahn S. The convergence of fictitious play in 3×3 games with strategic complementarities. *Econ Lett* 1999;64(1):57–60.

21. Sela A. Fictitious play in 2×3 games. *Games Econ Behav* 2000;31(1):152–162.
22. Berger U. Fictitious play in $2 \times n$ games. *J Econ Theory* 2005;120(2):139–154.
23. Hofbauer J, Sorin S. Best response dynamics for continuous zero-sum games. *Discrete cont. dynamic sys. Series B*. 2006;6(1):215–224.
24. Berger U. Two more classes of games with the continuous-time fictitious play property. *Games Econ Behav* 2007;60(2):247–261.
25. Berger U. Learning in games with strategic complementarities revisited. *J Econ Theory* 2008;143:292–301.
26. Shamma JS, Arslan G. Unified convergence proofs of continuous-time fictitious play. *IEEE Trans Automat Control* 2004;49(7):1137–1142.
27. Brown GW. Some notes on computation of games solutions. RAND Report 78. Santa Monica (CA): The RAND Corporation; 1949.
28. Fudenberg D, Kreps DM. Learning mixed equilibria. *Games Econ Behav* 1993;5:320–367.
29. Hofbauer J, Sandholm W. On the global convergence of stochastic fictitious play. *Econometrica* 2002;70:2265–2294.
30. Leslie DS, Collins EJ. Generalized weakened fictitious play. *Games Econ Behav* 2006;56:285–298.
31. van der Genugten B. A weakened form of fictitious play in two-person zero-sum games. *Int Game Theory Rev* 2000;2(4):307–328.
32. Matsui A. Best response dynamics and socially stable strategies. *J Econ Theory* 1992;57:343–362.
33. Hofbauer J. Stability for the best response dynamics. Mimeo, University of Vienna; 1995.
34. Shapley LS. Some topics in two-person games. *Ann Math Stud* 1964;5:1–28.
35. Sparrow C, van Strien S, Harris C. Fictitious play in 3×3 games: the transition between periodic and chaotic behavior. *Games Econ Behav* 2008;63:259–291.
36. Foster DP, Young HP. On the nonconvergence of fictitious play in coordination games. *Games Econ Behav* 1998;25(1):79–96.
37. Jordan J. Three problems in learning mixed-strategy Nash equilibria. *Games Econ Behav* 1993;5:363–386.
38. Gaunersdorfer A, Hofbauer J. Fictitious play, shapley polygons and the replicator equation. *Games Econ Behav* 1995;11:279–303.
39. Monderer D, Sela A. A 2×2 game without the fictitious play property. *Games Econ Behav* 1996;14:144–148.
40. Krishna V, Sjostrom T. On the convergence of fictitious play. *Math Oper Res* 1998;23(2):479–511.
41. Gass SI, Zafra PMR. Modified fictitious play for solving matrix games and linear programming problems. *Comput Oper Res* 1995;22:893–903.
42. Garcia A, Reaume D, Smith RL. Fictitious play for finding system optimal routings in dynamic traffic networks. *Transp Res Part B* 2000;34:147–156.
43. Lambert TJ, Wang H. Fictitious play approach to a mobile unit situation awareness problem. Technical Report. Ann Arbor (MI): University of Michigan; 2003.
44. Lambert TJ, Epelman MA, Smith RL. A fictitious play approach to large-scale optimization. *Oper Res* 2005;53(3):477–489.
45. Hon-Snir S, Monderer D, Sela A. A learning approach to auctions. *J Econ Theory* 1998;82:65–88.

FINITE POPULATION MODELS—SINGLE STATION QUEUES

RAJA JAYARAMAN

Department of Industrial Engineering,
University of Arkansas, Fayetteville,
Arkansas

TIMOTHY I. MATIS

Department of Industrial Engineering,
Texas Tech University, Lubbock, Texas

INTRODUCTION

Queueing models consisting of a finite number of customers in the population arriving for service have several real-life applications including the classical machine-interference or machine–repairman model, where the number of machines that fail and are either waiting for or undergoing repair is finite. The fundamental characteristic of finite queues is that the arrival process is nonrenewal, as the calling population of new arrivals is finite. Machine failures, data packets, assembly jobs, number of calls to a trunk system, and so on, are finite and can be accommodated only if the service facility is idle, leading to a closed system. Many variations of the finite queueing model can be constructed depending on the assumptions made on the arriving population, service distributions, and queue discipline. In the following sections, we review several popular applications of finite queues.

An important application of finite queues arises in computer communication models. Consider N terminals requesting processing time by a central server. The total number of service requests from these N terminals and the amount of time the server takes to complete the tasks can be modeled using finite queues to obtain several performance measures of the system, such as server down time, idle time, and average time to process service requests. These measures play a pivotal role for decision makers to handle a variety of

issues including capacity planning for network optimization, client-server architecture models [1–3], web services scalability, and congestion control [4,5].

Another variation of finite queueing models is employed in telecommunication systems, where call arrivals requesting service are processed if the server is free. When the server is busy, all new calls are routed to a finite buffer (waiting pool) to wait a random amount of time until they reenter service or balk out of the system. Such a system is modeled as a retrial queueing system, which is discussed extensively in the article titled *Retrial Queues* and Falin and Templeton [6].

Finite queueing models are also extensively applied to the study of vacation models, most queueing models assume the continuous availability of servers, while in reality servers often take arbitrary vacations for random amounts of time and return to queue to serve waiting customers for a limited duration. A comprehensive treatment of such models can be found in Tian [7].

Finite queueing models also have strong applications in manufacturing and production systems [8,9]. They are used to study assembly systems that are closed with input buffers of a finite capacity k having dependent arrival distributions. Another application arises in fork-join queues, where several independent arrival buffers feed a service station. The assumption of the Poisson arrivals in these models is often unrealistic as input feed buffers are capacitated in buffer size. When the desired capacity of k is attained, all new arrivals are turned away. In addition, models that involve the probabilistic feedback of completed jobs into the input buffer for rework also invalidate the Poisson arrival assumption. Key performance measures of interest in these manufacturing systems include the queue length distributions of each buffer, average synchronization time, throughput of the system, idle time of the server, and so on.

Another interesting application of finite queues is in the study of reliability, where

component failures form the arrival process for servicing requests of repair and maintenance actions. The number of components in the system which can fail is finite, and the number of service requests either waiting or being repaired is also finite. Such models can be analyzed using finite population models. In many cases of reliability modeling, there is strong connection to finite population models as noted in Kavelenko *et al.* [10]. The performance measures of interest in reliability models include queue length distribution, mean waiting time, server idle time, and so on.

In the applications given above, arriving customers do not form a renewal process as they are dependent on the number of units currently in the system, which has a finite capacity. With reference to Kendall's notation, see the article titled **Queueing Notation**, an example of a finite-source model is given by $M/G/1/K/N$, which describes a single-server model with exponentially distributed arrival times and generally distributed service times. There is a maximum of K customers allowed both in service and waiting for service, and N available customers in the population to request service, which will be accommodated if the number of customers in the system is strictly less than K . The $M/G/1/N/N$ queue is a special case of the above system, where the number of allowed service requests in the system equals the number of available customers N . In the next section, we analyze $M/M/C/r/r$ queue, often referred to as *machine-repairman model*, and some key performance measures of the model.

$M/M/C/r/r$ QUEUE-MACHINE REPAIRMAN MODEL

The classical machine-repairman model (also known as *machine-interference model*, *machine-servicing models*) was developed in the late 1960s, and aimed to study the optimal assignment of repairman to service machines subject to random failures [11]. The importance of this class of model has found profound applications and extensions, many of which have been discussed above.

In simple terms, the machine-repairman model aims to obtain systems performance measures of both repairman and machines in the system, such as the average number of machine failures, throughput, number of machines in queue waiting for service, worker utilization and efficiency, and average number of idle operators among others.

Consider a system with r machines operating independently that are subject to random failures, and C repairman present to repair failed machines, with $C < r$. Assume the machine failures (arrivals) happen according to exponential distribution with rate λ , and machine repair times (service) are independent with rate μ . All machine failures are serviced according to first-come-first-served (FCFS) discipline, and serviced machines start operating upon completion of service. This model may be described as a $M/M/C/r/r$ queueing system, which is studied using birth-death process as given in Allen [12], Gross *et al.* [13], Kleinrock [14], and Trivedi [15].

Let $N(t)$ denote the number of machines that have failed at time t . Assuming that n machines have already failed in the system, the probability of a machine failure in $(t, t + \delta t)$ is given by

$$P(\text{a machine failure} | N(t) = n) = (r - n)\lambda.$$

Given that $N(t) = n$, the probability of one of the failed machines getting fixed is in $(t, t + \delta t)$ given by

$$P(\text{rectifying a failed machine} | N(t) = n) = \begin{cases} n\mu, & 0 \leq n \leq C \\ C\mu, & n > C \end{cases}.$$

The stochastic process $\{N(t), t \geq 0\}$ follows a birth-death process [13] with arrival and service rates given by

$$\lambda_n = \begin{cases} (r - n)\lambda, & 0 \leq n \leq r \\ 0, & n \geq C \end{cases}, \quad (1)$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n \leq C \\ C\mu, & n \geq C \end{cases}. \quad (2)$$

The Kolmogorov forward equations for the birth-death process are given by

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -r\lambda P_0(t) + \mu P_1(t) \\
\frac{dP_n(t)}{dt} &= -[(r-n)\lambda + n\mu]P_n(t) \\
&\quad + (n+1)\mu P_{n+1}(t) + (r-n+1) \\
&\quad \times \lambda P_{n-1}(t); \quad \text{for } 1 \leq n < C \\
\frac{dP_n(t)}{dt} &= -[(r-n)\lambda + C\mu]P_n(t) + C\mu P_{n+1}(t) \\
&\quad + (r-n+1)\lambda P_{n-1}(t); \\
&\quad \text{for } C \leq n < r \\
\frac{dP_r(t)}{dt} &= -C\mu P_r(t) + \lambda P_{r-1}(t). \quad (3)
\end{aligned}$$

In steady state,

$$\begin{aligned}
r\lambda P_0 &= \mu P_1 \\
[(r-n)\lambda + n\mu]P_n &= (n+1)\mu P_{n+1} \\
&\quad + (r-n+1)\lambda P_{n-1}; \quad \text{for } 1 \leq n < C \\
[(r-n)\lambda + C\mu]P_n &= C\mu P_{n+1} \\
&\quad + (r-n+1)\lambda P_{n-1}; \quad \text{for } C \leq n < r \\
C\mu P_r &= \lambda P_{r-1}. \quad (4)
\end{aligned}$$

Equation (4) can be solved recursively using the following relationship to obtain the state probabilities:

$$\begin{aligned}
(r-n)\lambda P_n &= (n+1)\mu P_{n+1}; \quad 0 \leq n < C \\
(r-n)\lambda P_n &= C\mu P_{n+1}; \quad n \geq C. \quad (5)
\end{aligned}$$

Solving Equation (4) recursively, we obtain the steady-state probability of having n machines failed in closed form

$$P_n = \begin{cases} \binom{r}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & 1 \leq n \leq C \\ \binom{r}{n} \frac{n!}{C!C^{n-C}} \left(\frac{\lambda}{\mu}\right)^n P_0, & C \leq n \leq r \end{cases}, \quad (6)$$

where P_0 is obtained using the normalizing condition given by $\sum P_n = 1$, or equivalently by

$$P_0 = \left[\sum_{n=0}^C \binom{r}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=C+1}^r \binom{r}{n} \frac{n!}{C!C^{n-C}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}. \quad (7)$$

Note that the steady-state probabilities in Equations (3) and (4) are applicable for any finite-source queue with exponential service, independent of the failure arrival distribution mean rate λ . The waiting time characteristics of this model can be obtained in closed-form expression based on the steady-state probabilities obtained above. In order to obtain the waiting time distribution of arrivals conditioned on n machine failures at the arrival epoch and unconditioning using the stationary probability distribution, we obtain waiting time distributions and other characteristics [13,36,37].

As noted, the arrival rates are conditioned on the state of the system with n machine failures, yet in steady state, we may express this as an unconditioned arrival rate that is commonly referred to as the *effective arrival rate* of machines is given by

$$\lambda_{\text{eff}} = \sum_{n=0}^{r-1} (r-n)\lambda P_n.$$

Performance Measures

Several key performance measures of the $M/M/C/r/r$ queue machine-repairman model are given in this section.

- Expected number of failed machines in steady state can be obtained from the definition of expectation as

$$E[L] = \sum_{n=0}^r nP_n.$$

- Average number of machines waiting in the service queue is given by

$$E[L_q] = \sum_{n=C+1}^r (n-C)P_n.$$

The associated mean waiting time from failure to recovery and the mean

waiting time in queue for service in closed form can be obtained from the above equations by employing Little formula [12,13]. It is intuitive that the number of working machines is the total number of machines less the expected number of failed machines, that is, $r - E[L]$.

- Machine utilization is the percentage of time that machines are productive/in use, and is given by the fraction of the number of working machines over the total number of machines r as

$$\rho_{\text{machine}} = \frac{r - E[L]}{r}.$$

- Repairman utilization or operator utilization is the percentage of time repairman is available to rectify machine failures. This is obtained as the sum of the expected number of failed machines being repaired by the C repairman and the sum of the probability of the number of failed machines that exceed the available number of repairman, and is given by

$$\rho_{\text{repairman}} = \sum_{n=0}^r \frac{nP_n}{C} + \sum_{n=C+1}^r P_n.$$

- The average waiting time of machines is the ratio of average number of machines waiting over the mean arrival rate times number of working machines as

$$\frac{E[L_q]}{\lambda[r - E[L]]}.$$

- The average down time of machines is the ratio of expected number of failed machines over the mean arrival rate times number of working machines as

$$\frac{E[L]}{\lambda[r - E[L]]}.$$

- The average number of idle repairman can be readily obtained using the repairman utilization as

$$\begin{aligned} C - C * \rho_{\text{repairman}} &= C(1 - \rho_{\text{repairman}}) \\ &= C \left[1 - \left(\sum_{n=0}^r \frac{nP_n}{C} + \sum_{n=C+1}^r P_n \right) \right] \\ &= \sum_{n=0}^C (C - n)P_n. \end{aligned} \quad (8)$$

- Percentage of time the machines are down (i.e., machine-interference time) awaiting service can be obtained by dividing $E[L_q]$ with the number of machines r as

$$\frac{\sum_{n=C+1}^r (n - C)P_n}{r}.$$

- Percentage of idle operators is obtained by dividing the average number of idle repairman by the total number of repairman C as

$$\frac{\sum_{n=0}^r (C - n)P_n}{C}.$$

In the next section, we discuss several extensions including parameter estimation methods and provide references to texts, survey articles, and other literature to the machine–repairman model.

EXTENSIONS AND FURTHER READING

A straightforward extension to the machine–repairman model is to incorporate general distributions for both arrival (machine failure) and service (repair) times, which have been studied by several authors [16–20]. Additional variations of the model include varying arrival and service rates [21], different types of failures, nonidentical machines [22], spare machine availability [13,23], spare machines with cold, warm, and hot standbys [24], priority queues [25–28], and server vacations [7,28]. Although closed-form solutions to the models are uncommon, many researchers have used numerical techniques to obtain the system performance measures.

In particular, a stationary distribution of number of working machines for an $M/G/1/N/N$ system was studied in Takacs [29], yet, resulted in a cumbersome evaluation of large factorials as the number of servers increased. A stable algorithm to compute these probabilities can be found in Gupta and Rao [18], and, as an alternative, many authors have analyzed the $M/G/1/N/N$ model using supplementary variables to recursively obtain the system probabilities. Likewise, embedded Markov chain techniques have also been used to obtain the probability distribution of machine down time at arbitrary time epochs.

In practice, the efficient estimation of parameters related to the machine failure (λ) and service (μ) time distributions are key to deriving meaningful performance measures. Since the machine operating and repair times are independent, maximum likelihood estimators are often employed to obtain reasonable estimates of the model parameters. Bayesian methods, however, also present an ideal alternative for inferences employing posterior distribution to approximate distributions of performance measure as given in Gaimon and Thompson [30].

In addition, we refer the readers to excellent survey articles on finite-source queueing systems and their applications [31,32], along with the books on modeling aspects and performance measures [28,33–35].

REFERENCES

1. Daigle JN. Queueing theory with applications to packet telecommunication. New York: Springer; 2004.
2. Sztrik J. Probability model for non-homogeneous multiprogramming computer system. *Acta Cybern* 1983;6:93–101.
3. Sztrik J. A queueing model for multiprogrammed computer systems with different I/O times. *Acta Cybern* 1985;7:127–135.
4. Menasce DA, Almeida VAF. Capacity planning for web services: metrics, models, and methods. New Jersey: Prentice Hall; 2001.
5. Nelson R. Probability, stochastic processes, and queueing theory - the mathematics of computer performance modeling. New York: Springer; 1995.
6. Falin G, Templeton J. Retrial queues. London: Chapman and Hall; 1997.
7. Tian N. Vacation queueing models: theory and applications - international series in operations research & management science. New York: Springer; 2006.
8. Altioik T. Performance analysis of manufacturing systems. New York: Springer; 1997.
9. Sztrik J. A finite-source queueing model for some manufacturing processes. *Probl Control Inf Theory* 1987;16(6):449–457.
10. Kavelenko IN, Kuznetsov NY, Pegg PA. Mathematical theory of reliability of time dependent systems with practical application. New York: John Wiley & Sons, Inc.; 1997.
11. Scherr AL. An analysis of time-shared computer systems. MIT research monograph. No. 36. Cambridge (MA): MIT Press; 1967.
12. Allen AO. Probability, statistics, and queueing theory with computer science applications. 2nd ed. Boston (MA): Academic Press; 1990.
13. Gross D, Shortle JF, Thompson JM, *et al.* Fundamentals of queueing theory. 4th ed. Hoboken (NJ): John Wiley & Sons, Inc.; 2008.
14. Kleinrock L. Queueing systems. Volume 1: theory and Volume 2: computer applications. New York: Wiley-Interscience; 1976.
15. Trivedi KS. Probability and statistics with reliability, queueing, and computer science applications. 2nd ed. New York: Wiley-Interscience; 2002.
16. Agnihothri SR. Interrelationships between performance measures for the machine-repairman problem. *Nav Res Log* 1989;36(3): 265–271.
17. Bunday BD, Scraton RE. The $G/M/r$ machine interference model. *Eur J Oper Res* 1980;4: 399–402.
18. Gupta UC, Rao TSS. A recursive method to compute the steady state probabilities of the machine interference model: $M/G/1/K$. *Comput Oper Res* 1994;21(6):597–605.
19. Sztrik J. On the finite-source $G/M/r$ queue. *Eur J Oper Res* 1985;20:261–268.
20. Sztrik J. The $G/M/1$ machine-interference model with state-dependent speeds. *J Oper Res Soc* 1988;39(2):201–207.
21. Kameda H. A finite-source queue with different customers. *J ACM* 1982;29(2):478–491.
22. Sztrik J. Some contributions to the machine interference problem with heterogeneous machines. *J Inf Process Cybern* 1988;24(3): 137–143.

23. Gross D. Queueing models for spares provisioning. *Nav Res Log* 1977;24(4):521–536.
24. Rao TSS, Gupta UC. Performance modelling of M/G/1 machine repairman model with cold, warm and hot standbys. *Comput Ind Eng* 2000;38(2):251–267.
25. Frostig E. Optimal policies for machine repairmen problems. *J Appl Probab* 1993;30(3):703–715.
26. Hsieh YC. Optimal assignment of priorities for the machine interference problems. *Microelectron Reliab* 1997;37(4):635–640.
27. Jaiswal N. Priority queues. New York: Academic Press; 1969.
28. Takagi H. Queueing analysis. A foundation of performance evaluation - Volume 1: Vacation and priority systems. Amsterdam: North-Holland Publishing Co.; 1991.
29. Takacs L. Introduction to the theory of queues. New York: Oxford University Press; 1962.
30. Gaimon C, Thompson GL. Optimal preventive and repair maintenance of a machine subject to failure. *Optim Control Appl Methods* 1984;5(1):57–67.
31. Sztrik J. Finite-source queueing systems and their applications - a bibliography. Debrecen, Hungary: Institute of Mathematics and Informatics - University of Debrecen; 2002. Retrived online: <http://irh.inf.unideb.hu/user/jsztrik/research/fsqreview.pdf>.
32. Sztrik J. Finite-source queueing systems and their applications. In: Ferenczi M, Pataricza A, Rnyai L, *et al.* Formal methods in computing. Budapest, Hungary: Akademia Kiado; 2005. Chapter 7.
33. Takagi H. Queueing analysis. A foundation of performance evaluation - Volume 2: Finite systems. Amsterdam: North-Holland Publishing Co.; 1993.
34. Takagi H. Queueing analysis. A foundation of performance evaluation - Volume 3: Discrete-time systems. Amsterdam: North-Holland Publishing Co.; 1993.
35. Buzacott JA, Shanthikumar JG. Stochastic models of manufacturing systems. Englewood Cliffs (NJ): Prentice Hall; 1993.
36. Sztrik J. On the machine interference. *Publ Math* 1983;30:165–175.
37. Sztrik J. Investigation of stationary characteristics of controlled systems $G/M/r$ with finite source. *Serdica* 1988;14(2):179–184.

FIRE DEPARTMENT DEPLOYMENT AND SERVICE ANALYSIS

JOHN R. HALL, JR.
National Fire Protection
Association, Quincy,
Maryland

One of the biggest success stories for operations research has been models for analysis of fire department service delivery. Pioneering work in the 1960s by researchers with the Rand Institute, New York City, won every major prize the operations research community has to offer. Applications of the work in New York and other cities have led to millions of dollars in documented savings while maintaining or nearly maintaining selected measures of service quality and quantity delivered [1–4].

Within the operations research community and literature, there have been a number of additional articles, but no dramatic shifts away from Rand's approaches. In the fire service community and its literature, however, developments have proceeded along different lines. While these alternative paths have produced potentially complementary work, there has been little effort to integrate them.

A GENERAL CONCEPTUAL MODEL OF FIRE DEPARTMENT SUPPRESSION SERVICE

A comprehensive timeline for fire development and fire department intervention might include the following elements [5]:

- *Detection Time.* Time from ignition to detection of fire.
- *Reporting Time.* Time from detection of fire to report to fire department.
- *Dispatch Time.* Time from receipt of report to dispatch of fire department resources.

- *Turnout Time.* Time from acknowledgment or awareness of call by responding forces until they begin travel to the site of the emergency.
- *Travel Time.* Time for travel to emergency site from station or other starting point to the place where the fire apparatus stops.
- *Access Time.* Time for firefighters to move from the apparatus to the point where firefighting will be staged.
- *Setup Time.* Time for firefighters to set up equipment prior to application of extinguishing agents.

There is a timeline running in parallel for the fire itself:

- smoldering, which is slow burning with no visible flame;
- open flaming period, when fire grows exponentially;
- flashover, when nearly everything in the first compartment is involved in fire;
- room-to-room spread;
- floor-to-floor spread;
- damage to structural stability;
- building-to-building spread.

Reducing any part of the first timeline by one minute will reduce fire damage by an amount that reflects the speed of fire growth at that point in the second timeline and the speed with which fire can be stopped by forces interceding at that point. If the fire department arrives while a fire is still smoldering, it is not worth arriving a minute sooner because the rate of damage during smoldering is so slow and control of the fire will occur so quickly. If the fire department arrives after flashover, it may not be able to control the fire at all until the first building is burned out. On the other hand, if the flashed-over building is a shed located some distance from any other building, there may not be much life, health, or property at risk, no matter when in the

fire timeline the fire department arrives. If the building is a high-rise with many people and considerable value, the consequences of arriving later in the fire can be enormous.

Dispatch and turnout times tend to be driven by a sequence of standard tasks and actions. There is limited opportunity to shorten those times. Setup time also involves a fairly standard set of tasks, although it can vary a great deal depending on the scale of the fire on arrival.

Detection and reporting times can be highly variable. They can be influenced by the use and performance of automatic detection, alarm, and reporting equipment, which in turn can be influenced by local codes. The use of fire sprinklers can keep fires smaller for a longer period of time.

Access time can vary significantly but is rarely addressed directly by deployment analysis models. Access time can depend on the height and area of the building and the complex or lot in which it is located, as well as the location of the fire within the building, security arrangements on site, and the extent to which fire department access complies with local requirements.

HISTORY OF TRAVEL TIME ANALYSIS AND CONTROL BEFORE THE NEW YORK CITY—RAND INSTITUTE

In the United States, for most of the twentieth century, controls and objectives on fire department travel time were greatly influenced by provisions of the *Grading Schedule for Municipal Fire Protection*, developed by the National Board of Fire Underwriters and later passed to the Insurance Services Office [1]. A community's insurance rates would be affected by the community's grade, which was calculated using an index number formula approach that included points for response times.

In the 1960s and more recently, the Grading Schedule formulas were stated as *maximum* acceptable response *distances*. All considerations that would affect the magnitude of loss associated with an extra minute of response time were channeled into index type formulas for defining categories of

estimated required fire flow (water delivered per unit time), that is, the rate of water delivery that would be needed to control a design fire at the location. Maximum acceptable response distances could be set at different values based on required fire flow.

During the 1960s and 1970s, there was growing technical criticism of the Grading Schedule [6]. The insurance industry's primary concern in creating the Grading Schedule was to avoid citywide conflagrations. A schedule designed for that purpose would not necessarily be valid for the more ordinary fires that drive total fire loss. The absence of better alternative formulas was not seen as a convincing defense, because actuarial decision making seemed capable of providing better results, at least for larger communities.

What Is the Objective Function?

If you frame the analysis to minimize travel time or to optimize on some other summary statistic based on travel time, then the basic analysis will not reflect potentially large variations in loss potential per unit time and will not reflect patterns in losses (like firefighter injuries and deaths) that may have little or no correlation to travel time at all. Those other considerations will then be either factored in through weights (e.g., hazard factors for types of buildings or for subareas) or constraints (e.g., maximums for response time or distance) or decomposed analysis (e.g., hazard groups of buildings or of subareas) or ignored.

The approach that is used will in turn make a difference, possibly a significant difference, in the relative emphasis placed on the "other considerations." If they are ignored, they will receive no emphasis. If they are handled through decomposed analysis, they may be regarded as second-order or sidebar concerns. If they are handled through weights or constraints, then the analyst may be back to something that looks very much like the Grading Schedule, at least in structure if not in specific parameters and thresholds.

Between weights and constraints, the latter approach has certain advantages. A constraints approach channels the limited

data and the interpretation of that data into a relatively small number of choices—the constraints—in a manner that resembles the standards-setting process already familiar to fire departments as the usual way of doing business. A weights approach typically requires the estimation of more parameters with more precision and more extensive use of the parameters in the analysis. If you have confidence in the weights, this is good, but if you do not—or your client does not—then the analysis may seem less transparent and so less acceptable. More specifically, with constraints it is easier to, for example, factor in the limited evidence that smaller crew sizes (fewer than four firefighters) are associated with higher firefighter injury rates per fire and a heightened risk of certain fatal scenarios. Also, incident rates by subarea or individual building may be less sensitive to small quantities of fire experience data.

THE NEW YORK CITY—RAND INSTITUTE'S PRINCIPAL TYPES OF OR MODELS FOR FIRE SERVICE DEPLOYMENT ANALYSIS

If the community's street network is laid out as a perfect grid, then all locations can be converted to graph coordinates, and travel distances will be given by the sum of distance along one dimension plus distance along the other dimension (called right-angle distance in the Rand studies). If the community's street network is more varied and dense and lacks major obstacles, like a river or railroad tracks, then it may be possible to achieve straight-line travel between many or most points in the community.

With the use of multipliers, one can adjust for the general influence of curves, hills, and one-way streets. A more complex approach uses different formulas as a function of distance (right angle for short distances, straight line for long distances, combination of both for intermediate distances).

Alternatively, one can set up the road system as a network and obtain actual link distances from appropriate data. Now you can adjust for actual locations of curves, hills, one-way streets, and even speed bumps, potholes and real-time road construction work.

The shortest-path algorithm can then be used to define the path from one location to another and calculate the travel distance associated with that path.

Travel time may not be simply proportional to travel speed if you take account of the need to accelerate and decelerate at the beginning and the end, respectively, of any trip and often at other points (e.g., when rounding corners, when street conditions are poor and will not permit peak speeds, when traffic is dense). Rand developed formulas for travel time as proportional to distance for longer distances and proportional to the square root of distance for shorter distances.

Using a queueing model framework, Rand also developed Poisson probability distributions for arrival of successive fire calls and exponential probability distributions for service time required at fire calls. Because a fire station, once located, will be expected to last for decades, they also developed bases for projecting future call volume. It would be possible to vary the parameters of the exponentially distributed service times based on hazard factors or hazard categories for the destination sites.

These simple models were then extended to address the case where a new fire call comes in while the assigned first-response company is still busy at another call.

Evaluating Alternative Locations for Fire Stations

Rand's version of a fire station location model was called the *Firehouse Site Evaluation Model* (FSEM). A very similar alternative model was developed independently at about the same time by Public Technology, Incorporated (PTI) [7,8].

The community is divided into subareas by a grid structure, and the resulting subareas are then associated with nodes for a network model of the community. Each candidate location for a station is associated with a node, and each possible site for a fire call is associated with a node. Using travel times based on either formula estimates (Rand FSEM) or collected data (PTI FSLP) for all links, the user can easily calculate travel times from

any station to any node. Using either a first-response protocol (which station goes with which node) or a closest-station approach, the user can calculate a travel time distribution associated with any choice of stations and response protocols. Nodes can be weighted by hazard factors such as the required fire flow values used by the Grading Schedule or the all-fire and structure-fire incidence rates used by Rand.

Many of the simplifications used in these models are a reflection of the times. Computational power required a coarse grid, with subareas no smaller than the coverage area of one alarm box and often several times that size. There was a simplification involved in treating travel time as constant for every structure in a single subarea, but even greater variation was ignored by treating the entire subarea as having a constant hazard factor and a single standard response complement.

Initial Dispatch Policies

In addition to decisions on how many companies to have and where to locate them, fire departments also face decisions on which and how many companies to send, based on where a fire call comes from.

The choices in framing the question for analysis are similar to those involved in station location decisions. The ideal analysis would estimate differences in loss—including direct damage, casualties, and possibly other types of loss—for each of several dispatch choices. The data required to conduct such an ideal analysis would be enormous in size, and the formulas required to make it manageable are in many cases based on a very narrow evidentiary base.

Rand elected to work with two levels of incident seriousness—called *serious fires* and *nonserious fires*—which were defined in terms of the number and type of companies they would tend to require in New York. This focused the analytical question into balancing the error of sending less than what is needed to a serious fire versus sending more than what is needed to a nonserious fire. This in turn involved analysis of patterns

of calls received from various locations and diagnostic error rates from various sources of calls (e.g., telephone report of fire vs. alarm box call).

When Is the Analysis Most Worthwhile?

The Institute published detailed documentation on their application of their models in several cities. Reading these applications, it seems clear that this kind of analysis has a hard time passing a cost-benefit test unless the community is planning to change the *number* of stations. Better siting of the same number of companies will produce marginal improvements in travel time profiles, and the scant data available on the value of time does not produce estimated benefits on the same scale as the costs of analysis and station relocations.

Applications where the number of companies is being changed come in two very different types. One is a community that plans to severely cut its budget by reducing the number of companies (e.g., New York). In such a situation, Rand found analysis could substantially reduce the impact of cuts on measures of overall average service, but strong objections from affected parts of the community were still likely and would often accuse modelers of promoting or causing the cuts.

The other type is a community facing rapid population growth and a need to add companies to respond to rising service demand (e.g., Denver's experience with Rand). In this situation, analysis can show that the same service objectives can be achieved with fewer additional companies if those companies are well located. Companies that do not exist do not have political support, so those kinds of changes in plans are less likely to produce political resistance.

Is OR-based Deployment Analysis Useful Only for Large Communities?

Rand cites six cities as the pilot users of their models—New York City, Yonkers (NY), Jersey City (NJ), Trenton (NJ), Denver (CO), and Tacoma (WA). All six had over 100,000 population at the time of the studies, and all but Trenton had over 150,000 population.

Communities of 100,000 or more population account for just over one-third of the US population but for just over 1% of US fire departments, or just over 300 fire departments. These primarily big-city departments average nearly 20 fire stations per department [9]. (Populous communities include large cities and some large urbanized counties.)

Communities of 50,000–99,999 population average five stations per department, and smaller communities average even fewer. The average for all US fire departments is less than two fire stations per department. It is not clear how large a community must be or how rapidly its population must be growing in order for the likely benefits of OR analysis to exceed the costs, but in terms of the Rand models as they were when they were created, it appears that most communities would not have found them to be worth the money.

How Is Deployment Analysis Conducted Today?

The rapid increases in computer power and calculation affordability, coupled with widespread availability of global positioning geographic databases, have created a world in which the cost of this kind of deployment analysis might be within the reach of far more communities. However, the websites for the Rand Corporation and Public Technology Institute (the modern name for Public Technology, Inc.) do not offer tools or consulting services for fire department or emergency response deployment analysis. Rand keeps its many reports and publications in print, while PTI has shifted its emphasis to IT services that do not use OR type models.

If you search the Internet for fire department deployment analysis tools and consulting, you will find vendors, but the models they are emphasizing are not the ones developed by Rand and still being refined in the OR literature.

At about the same time that the Institute was working in New York, the fire service community was pursuing “master planning.” More recently, the dominant phrase has been “community risk assessment.” In both cases,

the models, methods, and formats favor unquantified flow charts and checklists, as is common with planning approaches, rather than the quantified methods of OR.

Rating Deployment as Part of Fire Department Accreditation

In the last decade, the Commission on Fire Accreditation International (CFAI) developed an accreditation protocol that included fire service deployment elements [10].

The assessment of deployment capability is channeled through “standards of response coverage,” a term borrowed from the United Kingdom. Response time objectives are set for sites based on their risk category. The protocol defines four risk categories for structures based on high or low probability and high or low consequence. The protocol also anticipates that their simple two-part separation of consequence may be elaborated into many categories.

The names given to the four risk categories make clear that the emphasis is on consequence rather than probability. High probability/low consequence buildings are called *moderate risk* while low probability/high consequence buildings are called *high/special risk*. Inasmuch as the categories are meant to generate quantities of resources and speed of response required to minimize loss, it makes sense that probability is not particularly relevant.

Recent NFPA standards provide candidate values for response time objectives [11]. NFPA 1710 includes minimum standards for resources, including personnel, required at fire calls, based on hazard. NFPA 1710 also includes response time requirements which are stated in terms of the maximum response time achieved for 90% of fire incidents. Also in 2001, NFPA published the first edition of NFPA 1720, a companion document for volunteer fire departments [12].

The CFAI protocol identifies a process for local authorities to use in setting response time and other objectives. Accreditation depends partly on the completion of the objective-setting process and partly on success in meeting objectives once set. CFAI references NFPA standards but does not require they be used as the basis for

local objectives. CFAI also includes specific examples of individual communities, with the suite of objectives they chose. While not presented as standards or default choices, these examples could take on the aura of best practices or default choices.

The consulting firms that offer deployment analysis services are set up to work to the CFAI model. In this manner, the CFAI protocol has acquired the status that the Institute models had two decades ago—a widely used complete protocol with no competitor having comparable usage, resulting in a status as de facto standard or at least starting point for any other approach that might be offered.

In 2001, the US Fire Administration (USFA) announced the availability of software for Risk, Hazard and Value Evaluation (RHAVE), a tool to be used with the CFAI accreditation protocol [13]. The core of RHAVE is an Occupancy Vulnerability Assessment Profile (OVAP), which is a menu-based, qualitatively structured procedure for assigning hazard categories. The procedure includes information on building construction, height and area, distance to other buildings, number and characteristics of occupants, on-site fire safety features and systems relative to applicable codes, relative accessibility from the outside, types of on-site activities, types (degree of hazard) and quantities of on-site combustibles, and even degree of enforcement of codes by local fire marshals and inspectors. Water demand (fire flow) is entered by the user, not calculated. The inputs are reduced to four factors—building factor, life safety factor, risk factor, and water demand factor. The sum of these four factors is the factor score, and ranges of the score map into four RHAVE categories.

Shortly after announcing the availability of RHAVE, the USFA withdrew as its supplier and tech support. It is not clear how much RHAVE is being used by CFAI, the other original sponsor.

How Could Operations Research be Applied to Fire Department Deployment Analysis Today, Given Common Practices?

An OR deployment analysis for the twenty-first century could refine the consequence/

hazard methods and concepts developed for CFAI into weights. In partnership with fire protection engineering modelers, OR modelers could develop a more quantitative and less subjective representation of hazard (loss potential) as a function of site-specific characteristics (including access time, which is still generally ignored), scenario, and travel time.

It could use the response cover objectives as alternative measures for optimization (e.g., maximize the percentage of hazard-weighted incidents with response times that meet the objectives) or as constraints (e.g., minimize average response time to hazard-weighted incidents subject to a constraint of achieving the response time objectives).

Recent Operations Research Work on Fire Department Deployment

Most OR research on the subject in the past couple of decades has not followed the path described above but has pursued one or more of the following discussed below.

Extension to Ambulances. The last several decades have seen tremendous growth in the involvement of fire departments in the delivery of emergency medical service, either staffed ambulances or emergency medical technicians on fire apparatus who will link up with the ambulances, arriving separately and possibly under a separate authority. The CFAI protocol includes objectives for emergency medical service (EMS) that parallel those for fire suppression. As of 2005, two-thirds (67%) of US municipal fire departments provided EMS service, including nearly all departments serving communities of 100,000 or more population [9]. In 1980, NFPA statistics show municipal fire departments responded to 2,988,000 fire calls and 5,045,000 medical emergency calls, for a ratio of 1.7 EMS calls to every fire call. By 2006, municipal fire departments responded to only 1,642,500 fire calls and 15,062,500 medical emergency calls, for a ratio of 9.2 EMS calls to every fire call.

Meanwhile, a robust body of OR work on ambulance location has not grown out of the

fire department deployment literature but has its roots in health-care OR applications and facility location methods generally. This means the OR modeling may not fully address calls that involve both medical and nonmedical aspects, such as a tanker truck collision which may result in a fire, a hazardous material spill endangering the environment, and a medical emergency involving trapped or injured people. However, the NFPA standards and CFAI protocols are expressed in terms of response time and percent coverage standards; this formulation is quite compatible with the approaches used in OR ambulance location models.

It is beyond the scope of this article to characterize the state of OR work on ambulance deployment. It is also unnecessary. Three review articles provide an excellent overview—one by Henderson in this encyclopedia [14], one by Daskin and Dean [15], and one by Erkut *et al.* [16]. The latter two are recent works, both of which refer to Marianov and ReVelle [17], a 1995 piece that also remains useful today.

In emergency service delivery, it is common to treat distance as a proxy for time and time as a proxy for loss. At any particular fire, property damage may grow exponentially as a function of time. At any particular fire or medical aid call, the likelihood of survival may decline nonlinearly as a function of time. Much of the evolution in fire department and ambulance deployment models may be seen as attempts to make distance a better proxy for loss. For example, a coverage model to define successful coverage not by an acceptably short distance to the site of the emergency but by an acceptably short time for units with resources sufficient to the emergency. Other aspects of state of the art models for both fire and medical aid emergency service deployment include analysis of the solution under alternative projections of patterns of future demand and incorporation of the likelihood that the nearest unit will be busy with an earlier call.

Extension to Natural Disasters and Homeland Security. Firefighters are all-purpose first responders to every type of emergency.

In recent years, there has been greater awareness of the need for deployment and response analysis for the special challenges associated with natural disasters and homeland security. Here are three examples:

- *Post-Earthquake Fire.* Unique fire challenges can arise in the aftermath of a significant earthquake, including many simultaneous fire starts, initiated by breaks in gas lines and electrical distribution lines, and significant disruptions in the road and water supply distribution networks. Normal response protocols will need to be significantly modified to reflect these conditions, and triage protocols for multiple fires will need to be implemented. The OR community has been underrepresented in work on this subject. A major flood or snowstorm can also create some of these unique challenges, reducing availability of road networks and water supply, while also creating conditions for some simultaneous fire starts.
- *Resources Diverted by a Major Incident.* Interviewing New York City fire department officials for their retrospective on the New York Rand work, Green and Kolesar learned that the Rand relocation algorithms had allowed the city to maintain “adequate levels” of fire protection throughout the city when 200 fire companies were committed to the response to the World Trade Center fires on September 11, 2001 [4]. Many other communities have been faced with threatening flood conditions from nearby swollen rivers, forcing most if not all of their emergency response forces to commit to mitigation of the natural disaster. A community may compensate by being extra careful during such a challenge, but such behavior cannot be indefinitely sustained. Therefore, the question remains of how to achieve reasonable fire protection coverage under such conditions. OR analysis could help create better pre-plans to deal with such situations.

- *Terrorist and Ordinary Non-Fire Emergency Incidents.* The 1995 Sarin gas terrorist attack in the Tokyo subway system and the events of September 11, 2001, heightened concerns about all types of possible terrorist attacks, including chemical, biological, radiological, or nuclear attacks.

With such incidents, it is not only possible but also likely that emergency response will involve agencies beyond the local area where the incident occurs. Written agreements to guide such responses and to coordinate local, state, and national resources are now commonplace for wildland firefighting but not yet for antiterrorism or for natural disasters such as floods and hurricanes. This also appears to be a problem area ripe for new OR work.

Refining the Math. There have been numerous published literature surveys on the topic of fire service deployment analysis. The impression one gets from any of these surveys is that most recent work on fire service or ambulance deployment analysis has been limited to mathematical refinements. A new paper describes a model of the same type as one of the older models with some changes in specifications, equations, data sources, or solution methods. There may be one real community that has been used to test the method.

Other issues have kept the market for deployment analysis small and should be priority focuses for any new work: (a) ways to frame deployment analysis so that it is practical, affordable, and useful for small departments and (b) modifications to deployment analysis models to incorporate volunteers and/or mutual aid from neighboring departments [18].

The Future of Deployment Analysis Using Operations Research

Green and Kolesar said that further progress in deployment analysis would depend on better success in finding client organizations that “can act as full-fledged partners,” in establishing projects that encompass multiple agencies and even multiple regions, and

in adapting models and showing validity for extreme events, such as the 9/11 attacks [4].

Given the success of CFAI in establishing a detailed protocol for fire department accreditation—and the interest of fire departments in obtaining the status and benefits associated with accreditation—it would seem prudent for OR professionals to seek to align their models and methods with the CFAI protocol. This seems a promising way to obtain better motivated clients and to reduce resistance from potential clients based on a likely greater allegiance to and comfort with a different approach to such analysis.

REFERENCES

1. Walker WE, Chaiken JM, Ignall EJ, editors. Fire department deployment analysis. New York: North Holland; 1979.
2. Kolesar P, Swersey A. The deployment of urban emergency units: a survey. *Delivery of urban services*. New York: North Holland; 1986. pp. 87–120.
3. Swersey AJ. The deployment of police, fire and emergency medical units. In: Pollock SM, Rothkopf M, Barnett A, editors. Volume 6, *Handbooks in operations research and management science*. New York: North Holland; 1994. pp. 151–190.
4. Green LV, Kolesar PJ. Improving emergency responsiveness with management science. *Manage Sci* 2004;50(8):1001–1014.
5. Johnson R, Price M. GIS for fire station locations and response protocols. In: Cote AE, editor. *Fire protection handbook*. Quincy (MA): National Fire Protection Association; 2008. pp. 12-215–12-233.
6. Homer P, Lawton J, Toregas C. Challenging the ISO fire rating system. *Pub Manage* 1977;59(7):2–6.
7. Fire station location package. Washington (DC): Public Technology, Inc.; 1974.
8. Toregas C, Swain R, ReVelle C, *et al.* The location of emergency service facilities. *Oper Res* 1971;19(6):1363–1373.
9. Four years later—a second needs assessment of the U.S. fire service. Quincy (MA): U.S. Fire Administration and National Fire Protection Association; 2006.
10. CFAI fire & emergency service self-assessment manual. 7th ed. Chantilly (VA): Center

- for Public Safety Excellence/Commission on Fire Accreditation International; 2006.
11. NFPA 1710. Standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by career fire departments. Quincy (MA): National Fire Protection Association; 2004.
 12. NFPA 1720. Standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by volunteer fire departments. Quincy (MA): National Fire Protection Association; 2004.
 13. R.H.A.V.E., risk, hazard and value evaluation, installation and users manual. Chantilly (VA): U.S. Fire Administration and Commission on Fire Accreditation International; 2001.
 14. Henderson S. Operations research tools for addressing current challenges in emergency medical services. *Encycl Oper Res Manage Sci* 2009. In preparation.
 15. Daskin M, Dean LK. Location of health care facilities. In: Brandeau M, Sainfort F, Pier-skalla WP, editors. *Operations research and health care—a handbook of methods and applications*. New York: Kluwer; 2004. pp. 43–76.
 16. Erkut E, Ingolfsson A, Sim T, *et al.* Computational comparison of five maximal covering models for locating ambulances. Paper published directly on http://www.business.ualberta.ca/aingolfsson/documents/PDF/Comparison_of_five_covering_models_May_31_07.pdf. 2007 May 31.
 17. Marianov V, ReVelle C. Siting emergency services. In: Drezner Z, editor. *Facility location: a survey of applications and methods*. Series in Operations Research. New York: Springer; 1995. pp. 199–222.
 18. Jennings CR. The promise and pitfalls of fire service deployment analysis methods. *Fire Technol* 2001;37(3):195–197.

FLUID MODELS OF QUEUEING NETWORKS

DAVID GAMARNIK
Operations Research Center and
Sloan School of Management,
MIT, Cambridge, Massachusetts

INTRODUCTION

Fluid models are a powerful approximation technique for studying important dynamic properties of queueing networks. The model is typically derived from an underlying queueing system by replacing a stochastic discrete model with a continuous deterministic model, using a certain law of large numbers type of rescaling. The obtained process reveals a lot fundamental dynamic structures in the underlying system. Historically, the fluid models have been derived in order to analyze the stability (positive recurrence) properties of an underlying queueing system. But since their inception, the technique proved to be useful for other applications such as designing optimal control policies, as well as for the analysis of queueing systems in the heavy traffic regime.

This article introduces basics of fluid models and illustrates them with a lot of examples. We begin with an informal construction to be followed by rigorous definitions and basic properties. A particular focus is given to the notions of stability of fluid models and methods of stability analysis such as the Lyapunov function technique. Examples of stable fluid models are given as well as some counterexamples, illustrating subtle issues in stability analysis. The formality of connections between fluid models and underlying stochastic queueing systems is discussed at a very high level, since the underlying techniques are highly involved and fall outside of the scope of the article. Instead, we point the reader to a rich literature on this subject.

The remainder of the article is organized as follows. A stochastic multiclass queueing

network (MQNET) is defined in the section titled “Multiclass Queueing Network: Model Formulation.” This model serves as the main motivation for introducing and studying the main object of this article—fluid model of an MQNET. Fluid MQNET is introduced informally in the section titled “Informal Derivation of Fluid Equations” using a couple of simple queueing models. Formally, the model is described in the section titled “A Fluid Model of a Multiclass Queueing Network.” Stability of fluid models is discussed in the section titled “Stability of a Fluid Model.” Some basic stability results are established in this section. Lyapunov method of stability analysis is introduced in the section titled “Lyapunov Function Method.” The connections between stability of fluid models and stability of an MQNET are discussed in the section titled “Connections with Stability of Stochastic Queueing Networks.” Apart from this section, all the other sections contain a sufficient amount of technical details and some (simple) proofs. Discussing results in the section titled “Connections with Stability of Stochastic Queueing Networks” in detail would require introduction of sophisticated measure-theoretic notions such as positive Harris recurrence, which falls outside of the scope of this article. Thus instead we give a somewhat informal summary of the main results in this area. The Final section provides notes and sources for material discussed in this article.

Let us introduce some notational conventions. $\mathbb{R}(\mathbb{R}_+)$ denotes the set of (nonnegative) real values. $\mathbb{R}^N$ is the N -dimensional Euclidian space and $\mathbb{R}_+^N$ is the nonnegative quadrant in this space—space of vectors with nonnegative components. Given a matrix $\mathbf{A}$, $\mathbf{A}'$ denotes its transpose. All vectors are assumed to be column vectors. $\mathbf{e}'$ denotes the (row) vector of ones. $\mathbf{I}$ denotes the indicator function. In other words, $\mathbf{I}\{A\}$ equals one if the event A holds true, and equals zero otherwise.

We use without defining some basic objects in queueing theory such as M/M/1 and G/G/1

queueing systems (refer to Kleinrock [1], or any other introductory text on queueing models). The latter is sometimes denoted by GI/GI/1 in order to stress the stochastic independence of interarrival and service times. We simply write it as G/G/1, since in the framework of fluid models only the first order parameters such as arrival rates and service rates play a role. In all of our queueing models no reneging is assumed and all the buffers are assumed to have infinite capacity.

MULTICLASS QUEUEING NETWORK: MODEL FORMULATION

We now introduce the MQNET model. One of earliest and important motivation for this model is wafer fabrication lines [2] and [3]. We introduce the model at a sufficient level of detail in order to facilitate the discussion of fluid models, but some details are omitted in the interest of space. The network consists of J single-server stations $s(1), \dots, s(J)$ and N job classes $1, \dots, N$. Each class $i = 1, \dots, N$ is associated with a unique buffer where the jobs associated with this class are queued. For convenience this buffer is also denoted by i and each buffer is associated with a unique server. The association is described by a $J \times N$ matrix $\mathbf{C} = (C_{ji}, 1 \leq j \leq J, 1 \leq i \leq N)$ with 0–1 entries, where $C_{ji} = 1$ if class i is associated with server $s(j)$ and $C_{ji} = 0$ otherwise.

Classes $i = 1, \dots, N$ are associated with external arrival processes $(A_i(t), t \geq 0)$ and service processes $(S_i(t), t \geq 0)$. For every real value $t \geq 0$, $A_i(t)$ denotes the total number of external arrivals into this class until time t , and $S_i(t)$ denotes the total number of service completions of the class i jobs that the associated server is capable of achieving during the time interval $[0, t]$. In practice, the actual number is typically lower if the server did not have jobs to work on or was working on jobs in a different class. A common assumption is to assume that $A_i(t), S_i(t)$ are renewal processes arising from random i.i.d. interarrival times. Let α_i represent the external arrival rate of the jobs in class i . We understand it in the sense that $\mathbb{E}[A_i(t)] = \alpha_i t + o(t)$ as $t \rightarrow \infty$. Similarly μ_i denotes the service

rate associated with class i . Let $\mathbf{M}$ denote the diagonal $N \times N$ matrix with μ_i in the i th place.

Upon service completion some of the jobs are rerouted to the same or different buffer and the jobs in the same buffer represent the mixture of external and internal arrivals. The routing is given by an $N \times N$ substochastic matrix $\mathbf{P} = (P_{ik} \geq 0, 1 \leq i, k \leq N)$ with the spectral radius strictly less than unity. In other words, $\sum_k P_{ik} \leq 1$ for each i and $\mathbf{P}^n \rightarrow 0$ coordinate-wise as $n \rightarrow \infty$. The second condition is equivalent to saying that every job *eventually* leaves the network (exercise). Here P_{ik} represents the probability that a job which has finished a service in class i is routed to class k . Upon arrival into buffer k the job occupies the last position in the queue. It is assumed that all the routing decisions are made independently for all the jobs. An important special case is when $P_{ik} = 0, 1$, which corresponds to the deterministic routing mechanism. In this case, each row contains at most one entry 1 with column indicating the destination class of the job from class i . When all the entries in a row are zero, the job leaves the network upon the service completion. The parameters $\alpha, \mu, \mathbf{P}, \mathbf{C}$ constitute the basic first order parameters of an MQNET and we write $(\alpha, \mu, \mathbf{P}, \mathbf{C})$ to denote an MQNET.

An important role is played by the load factors which we introduce now. Consider a system of traffic equations in variables $\lambda_i, 1 \leq i \leq N$ given by $\lambda_i = \alpha_i + \sum_{k=1}^N \lambda_k P_{ki}$. In vector form, we have for the (column) vector $\lambda = (\lambda_i, 1 \leq i \leq N)$

$$\lambda = \alpha + \mathbf{P}'\lambda, \quad (1)$$

which admits a unique solution explicitly given as

$$\lambda = [\mathbf{I} - \mathbf{P}']^{-1}\alpha. \quad (2)$$

Intuitively, λ_i represents the total (external plus internal) arrival rate into the buffer i . For every class i and server $s(j)$, define $\rho_i = \lambda_i / \mu_i$ to be the traffic intensity associated with the class i and define $\rho_{s(j)} = \sum_{i \in s(j)} \rho_i$ to be the traffic intensity of the server $s(j)$. Let

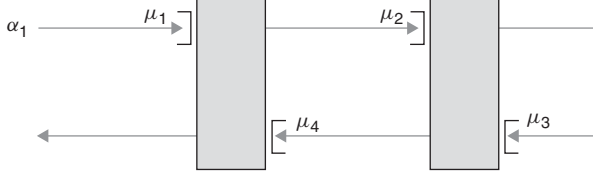


Figure 1. An example of an unstable queueing network satisfying Equation (18).

also $\rho = (\rho_i, 1 \leq i \leq N)$. Then

$$\rho = M^{-1}\lambda = M^{-1}[I - P']^{-1}\alpha.$$

Let us illustrate our discussion with an example. Consider the networks depicted in Fig. 1. This is an MQNET with

$$C = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

No external arrivals are associated with classes 2, 3, and 4. Suppose the arrival rate for the class 1 is $\alpha_1 = 1$. Suppose the service rates are $\mu_1 = \mu_3 = 10, \mu_2 = \mu_4 = 5/3$. The traffic equations (1) and (2) are solved easily to give $\lambda_i = 1, i = 1, 2, 3, 4$. This is consistent with the intuition: class 1 admits only external arrival streams with rate 1. Classes 2, 3, and 4 admit only internal arrival streams, all with rate 1. We also find that $\rho = (1/10, 3/5, 1/10, 3/5)'$, and $\rho_{s(1)} = \rho_{s(2)} = 7/10$.

Another example is described in Fig. 2. The parameters of the network are as follows:

$$N = 3, J = 2, \alpha = (\alpha_1, 0, \alpha_3), \mu = (\mu_1, \mu_2, \mu_3)$$

$$P = \begin{pmatrix} 0 & p & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \quad (3)$$

We obtain $\rho_1 = \alpha_1/\mu_1, \rho_2 = \alpha_1 p/\mu_2, \rho_3 = \alpha_3/\mu_3$.

An important aspect of an MQNET is scheduling policy specification. Given a choice of several jobs how does the server decide which job to work on or when to idle? In this article, we consider a broad class of policy for which we require only two assumptions: (i) the jobs within each class are selected using the first-in-first-out (FIFO) rule (this property is sometimes called *head-of-the-line* type policy), and (ii) scheduling policy is work conserving, namely, servers never idle when they have at least one job waiting to be processed on a server. Most of the scheduling policies considered in the queueing literature, such as FIFO, processor sharing, and static buffer priority (which we simply call a priority policy), satisfy these two properties. We refer to Bramson [4], Dai [5], and Dai and Weiss [6] for definitions and properties of these policies in the stochastic MQNET setting.

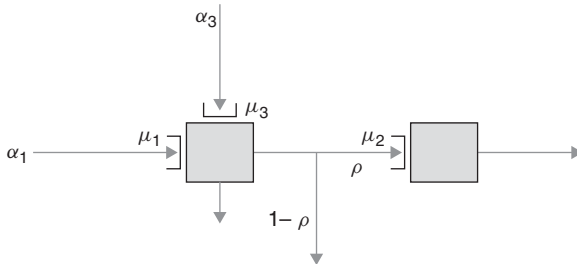


Figure 2. A two-server network with three classes.

INFORMAL DERIVATION OF FLUID EQUATIONS

Let us begin by constructing a fluid model of a $G/G/1$ type queueing system, arguing very informally. We begin with a “birds-eye” view of this queueing system in order to intuit the dynamics of this system at large timescales when viewed from a “large distance.”

Fluid Model of a $G/G/1$ Queueing System: “Birds-Eye” View of Stochasticity

Consider a $G/G/1$ queueing system which has a very large number of jobs $Q(0) = n$ at time $t = 0$. Let the arrival rate and service rate of our queueing system be α and μ , respectively. Imagine that you observe this system from a very large distance such that, for example, the difference between queue length 100 and 102 is blurred. Imagine also that you also observe this system for a long time period comparable with n . What do you expect to see?

Consider the initial time interval $[0, nt]$ where $t > 0$ is fixed to be a constant and n grows to infinity. From our assumption $\mathbb{E}[A(t)] = \alpha t + o(t)$, you expect approximately αnt new jobs arriving into the system during this time period. By the same token, the server is capable of serving roughly μnt jobs. When t is sufficiently small, specifically $t < 1/\mu$, this number is smaller than n , and we do not expect a shortage of jobs during this time period, so the server is working all the time during the time interval $[0, t]$. Thus, roughly speaking, we have $Q(nt) \approx n + \alpha nt - \mu nt$ jobs in the system at the end of this period. Another way to write this approximation is

$$\frac{Q(nt)}{n} \approx 1 + \alpha t - \mu t,$$

which is n free in the right-hand side. In order to make the left-hand side also n free, introduce $\bar{Q}(t) = Q(nt)/n$ and rewrite the equation as our first fluid equation: $\bar{Q}(t) \approx 1 + \alpha t - \mu t$. Anticipating that in fact all the approximations $\approx$ can be justified rigorously, we replace it by equality from now on.

What happens for larger t ? The answer depends on the relative magnitude of α and

μ , namely, whether $\rho = \alpha/\mu$ satisfies $\rho > 1$, $\rho = 1$, or $\rho < 1$.

- $\rho > 1$. In this case the arrival rate is faster than the service rate. We expect that the server never runs out of jobs and the fluid equation becomes $Q(nt) \approx n + \alpha nt - \mu nt$, that is,

$$\bar{Q}(t) = 1 + \alpha t - \mu t, \forall t \geq 0.$$

Specifically $\bar{Q}(t)$ diverges to infinity at a linear rate $\alpha - \mu > 0$.

- $\rho = 1$. In this case, our reasoning suggests again that the server never runs out of jobs, but since $\alpha = \mu$, we simply obtain $Q(nt) \approx n + \alpha nt - \mu nt = n$, that is,

$$\bar{Q}(t) = 1, \forall t \geq 0. \quad (4)$$

A way to reason and justify this is to observe that $Q(t)$ is a symmetric random walk constrained to stay on nonnegative integers. Starting from state $Q(0) = n$ during time interval $[0, nt]$ it fluctuates at the level $O(\sqrt{nt})$ (in fact the fluctuation becomes a driftless Brownian motion). This correction $O(\sqrt{nt})$ is washed away on the fluid n -scale, and we obtain the limiting equation (4).

- $\rho < 1$. This is the most interesting case. Since $\alpha < \mu$, the total number of jobs steadily decreases until the server empties for the first time. The fluid equation is again $Q(nt) \approx n + \alpha nt - \mu nt$, but it is valid only until the first emptying time which we expect to happen around time $n/(\mu - \alpha)$. After this, we expect the system to have usual steady-state type fluctuations that do not depend on n —the initial queue length. In other words, $Q(nt)/n \approx 0$. We conclude

$$\bar{Q}(t) = \begin{cases} 1 + \alpha t - \mu t, & t \leq \frac{1}{\mu - \alpha}; \\ 0, & t > \frac{1}{\mu - \alpha}. \end{cases} \quad (5)$$

We can summarize all three cases by writing $\bar{Q}(t) = \max(1 + \alpha t - \mu t, 0)$. If we begin with nx , many jobs at time zero, instead of n , but still keep n as our scaling parameter,

we obtain $\bar{Q}(t) = \max(x + \alpha t - \mu t, 0)$. Yet, another way to write the fluid equation (5), for the interesting case $\rho < 1$ is $\bar{Q}(t) = x + \alpha t - \mu T(t)$, where $T(t) = t$ for $t \leq x/(\mu - \alpha)$ and $T(t) = x/(\mu - \alpha) + \rho t$ otherwise (check the validity of this). There is a good physical interpretation of the function $T(t)$ —it is the scaled cumulative amount of time that the server was working. Up until time $x/(\mu - \alpha)$ the queue was positive and the server was working full time, thus giving $T(t) = t$. After this time, the system is more or less in steady state, where we know from Little's Law, the average time that the server is busy is ρ . Hence the expression for T after the critical time $x/(\mu - \alpha)$.

More General Constructs: Many Classes, Scheduling Policies, Routing

How can we generalize our informal derivations of a fluid model to more general and interesting cases, for example, networks of queues with many classes such as MQNET described in the previous section?

Consider a multiclass single-server ($J = 1$) queueing system with classes $i = 1, 2, \dots, N$. For simplicity, let us assume that there is no feedback in the system. In other words, all the jobs leave the system after their first service completion. Using the terminology of the previous section, we consider $(\alpha, \mu, \mathbf{P})$ with $\mathbf{P} = 0$. Suppose we begin with $nx_i, x_i > 0$ jobs in class $i = 1, 2, \dots, N$, where n is again very large. We focus first on the workload at time t , defined for a moment as $W(t) = \sum_{1 \leq i \leq N} Q_i(t)/\mu_i$. The workload changes at the rate approximately $\sum_i \alpha_i/\mu_i - 1 = \rho - 1$ per unit of time. Specifically, if $\rho < 1$, we expect it to hit zero for the first time around time t_0 satisfying $\sum_i nx_i + (\rho - 1)t_0 = 0$, namely, $t_0 = \sum_i nx_i/(1 - \rho)$. After this moment, just as in the $G/G/1$ case, the system experiences the steady-state type stochastic fluctuations which are n independent. Thus, we expect the following fluid equation for the fluid model of a workload process $\bar{W}(t) \triangleq W(nt)/n$

$$\bar{W}(t) = \begin{cases} \sum_i \frac{x_i}{\mu_i} + (\rho - 1)t, & t \leq t_0; \\ 0, & t > t_0. \end{cases} \quad (6)$$

What about the individual queues $Q_i(t)$? Naturally, their dynamics depend on the choice of the scheduling policy. Yet, without specifying the policy, we can still say something in the fluid equations. Let $T_i(t)$ be the total time spent by the server working on class i jobs during the interval $[0, t]$ and let $\bar{T}_i(t) = T_i(nt)/n$. Trivially, $\sum_i T_i(nt) \leq nt$, that is, $\sum_i \bar{T}_i(t) \leq t$. Since during the time interval $[0, nt_0]$, we expect the system to be fully busy, the following relation seems reasonable: $\sum_i \bar{T}_i(t) = t, \forall t \leq t_0$.

We expect roughly $\mu_i T_i(nt)$ type i jobs to be processed during the time interval $[0, nt]$. Also roughly $\alpha_i nt$ jobs arrive during this time interval. We obtain $Q_i(t) = nx_i + \alpha_i nt - \mu_i T_i(nt)$, $t \leq t_0$, which leads to a fluid equation $\bar{Q}_i(t) = x_i + \alpha_i t - \bar{T}_i(t)$, $t \leq t_0$. The following set of fluid equations is then obtained:

$$\bar{Q}_i(t) = \begin{cases} x_i + \alpha_i t - \mu_i \bar{T}_i(t), & \forall t \leq t_0, \\ 0, & \forall t > t_0, \end{cases} \quad 1 \leq i \leq N.$$

$$\sum_{i=1}^N \bar{T}_i(t) = t, \forall t \leq t_0,$$

$$\sum_{i=1}^N \bar{T}_i(t) \leq t, \forall t \geq 0.$$

How do we incorporate scheduling policy into the fluid equations? In fact, we already did it via quantities $\bar{T}_i(t)$ as they encode the relative effort dedicated by the server to each individual class during a given time interval. Sometimes additional conditions on $\bar{T}_i(t)$ are added resulting from details of a policy.

It remains to derive the fluid equation in the presence of feedback and routing, specifically probabilistic routing. Consider the following network in Fig. 2 described in the earlier section. Rather than deriving the full set of fluid equations, let us focus on the second queue $Q_2(t)$. Suppose initially we have $Q_i(0) = nx_i, i = 1, 2$. During the timer interval $[0, nt]$, the first server processes roughly $\mu_1 T_1(nt)$ jobs out of which roughly $\mu_1 T_1(nt)p$ will be routed to the second queue. Also during the same time period roughly $\mu_2 T_2(t)$ jobs will be served in the second server. Thus

we can write the following fluid equation for the second queue: $\bar{Q}_2(t) = x_2 + \mu_1 \bar{T}_1(t)p - \mu_2 \bar{T}_2(t)$, in addition to the already derived equation $\bar{Q}_1(t) = x_1 + \alpha_1 t - \mu_1 \bar{T}_1(t)$.

With this in mind we now turn to a formal definition of fluid models.

A FLUID MODEL OF A MULTICLASS QUEUEING NETWORK

Consider an MQNET $(\alpha, \mu, \mathbf{P}, \mathbf{C})$. The network consists of N classes and J service stations $s(1), \dots, s(J)$.

Definition 1. A fluid model associated with a network $(\alpha, \mu, \mathbf{P}, \mathbf{C})$ is defined to be the set of solutions $\bar{\mathbf{Q}}(t) = (\bar{Q}_1(t), \dots, \bar{Q}_N(t))$, $\bar{\mathbf{T}}(t) = (\bar{T}_1(t), \dots, \bar{T}_N(t))$ to the following system of functional equations and inequalities.

$$\bar{Q}_i(t) = \bar{Q}_i(0) + \alpha_i t + \sum_{j=1}^N \mu_j p_{ji} \bar{T}_j(t) - \mu_i \bar{T}_i(t), \quad (7)$$

$$\sum_{i \in s(j)} \bar{T}_i(t) - \bar{T}_i(s) \leq t - s, \quad (8)$$

$$\bar{T}_i(t) \text{ is nondecreasing, } \bar{T}_i(0) = 0, \text{ and } \bar{Q}_i(t) \geq 0 \quad (9)$$

for all $i = 1, 2, \dots, N, t, s \in \mathbb{R}_+, t \geq s$.

Any solution $(\bar{\mathbf{Q}}(t), \bar{\mathbf{T}}(t))$ to the system of equations 7–9 is called a *fluid solution* to the fluid model. We think of $\bar{\mathbf{Q}}(0)$ as the initial state of a fluid solution. We can rewrite Equation (7) in the matrix form:

$$\bar{\mathbf{Q}}(t) = \bar{\mathbf{Q}}(0) + \alpha t + (\mathbf{P}' - \mathbf{I})\mathbf{M}\bar{\mathbf{T}}(t). \quad (10)$$

From this we immediately obtain that for every $s < t$

$$\bar{\mathbf{Q}}(t) = \bar{\mathbf{Q}}(s) + \alpha(t - s) + (\mathbf{P}' - \mathbf{I})\mathbf{M}(\bar{\mathbf{T}}(t) - \bar{\mathbf{T}}(s)). \quad (11)$$

Thus, for each $s \geq 0$ we obtain a new fluid solution $\bar{\mathbf{Q}}'(t) = \bar{\mathbf{Q}}(t + s)$, $\bar{\mathbf{T}}'(t) = \bar{\mathbf{T}}(t + s) - \bar{\mathbf{T}}(s)$, $t \geq 0$ (check that this is indeed a fluid solution). Multiplying both

sides of Equation (10) by $M^{-1}[\mathbf{I} - \mathbf{P}']^{-1}$, we obtain

$$M^{-1}[\mathbf{I} - \mathbf{P}']^{-1}\bar{\mathbf{Q}}(t) = M^{-1}[\mathbf{I} - \mathbf{P}']^{-1}\bar{\mathbf{Q}}(0) + \rho t - \bar{\mathbf{T}}(t).$$

Define $\bar{\mathbf{Q}}_{s(j)} = \sum_{i \in s(j)} \bar{Q}_i(t)$, $\bar{\mathbf{T}}_{s(j)} = \sum_{i \in s(j)} \bar{T}_i(t)$. For the reasons which will become clear in future, define also $\bar{\mathbf{W}}(t) = (\bar{W}_i(t), 1 \leq i \leq N) \triangleq M^{-1}[\mathbf{I} - \mathbf{P}']^{-1}\bar{\mathbf{Q}}(t) \geq 0$ to be the *fluid workload* of the fluid solution at time t . Observe that $\bar{\mathbf{W}}(t) \neq 0$ if and only if $\bar{\mathbf{Q}}(t) \neq 0$ (show this). Also define $\bar{W}_{s(j)}(t) \triangleq \sum_{i \in s(j)} \bar{W}_i(t)$ to be the *fluid workload* in server $s(j)$ at time t . In terms of these definitions, we obtain

$$\bar{\mathbf{W}}(t) = \bar{\mathbf{W}}(0) + \rho t - \bar{\mathbf{T}}(t), \quad (12)$$

$$\bar{W}_{s(j)}(t) = \bar{W}_{s(j)}(0) + \rho_{s(j)}t - \sum_{i \in s(j)} \bar{T}_i(t), \quad \forall j = 1, \dots, J. \quad (13)$$

Basic Properties of a Fluid Model

What properties should a fluid solution $(\bar{\mathbf{Q}}(t), \bar{\mathbf{T}}(t))$ satisfy? The proof of the following observation is an easy exercise.

Proposition 1. Every fluid solution $(\bar{\mathbf{Q}}(t), \bar{\mathbf{T}}(t))$ is Lipschits continuous.

Looking at the fluid solution $\bar{\mathbf{Q}}(t) = \max(\bar{\mathbf{Q}}(0) + \alpha t - \mu t, 0)$ we see that the solution while continuous is not necessarily differentiable. This turns out not to be a major problem. The following fact is known from real analysis: for every Lipschits continuous function f there exists a set $A \subset \mathbb{R}$ such that for every $t \in A$, $\dot{f}(t)$ exists and $\mathbb{R} \setminus A$ has Lebesgue measure zero [7]. In other words, while the derivative does not exist everywhere, it exists *almost* everywhere, in the measure-theoretic sense. Throughout this article, we consider derivatives at various time instances implicitly assuming that we consider only times when derivatives exist. In light of Proposition 1, the derivatives of $\bar{\mathbf{Q}}(t)$ and $\bar{\mathbf{T}}(t)$ exist almost everywhere. Using a simplifying notation $\mathbf{u}(t) = d\bar{\mathbf{T}}(t)/dt$,

we obtain from Equations (10), (8) and (9), respectively,

$$\frac{d\bar{Q}(t)}{dt} = \alpha + [\mathbf{P}' - \mathbf{I}]\mathbf{M}\mathbf{u}(t), \quad (14)$$

$$\sum_{i \in s(j)} u_i(t) \leq 1, \forall j = 1, \dots, J, \quad (15)$$

$$\mathbf{u}(t) \geq 0. \quad (16)$$

Let $u_{s(j)}(t) \triangleq \sum_{i \in s(j)} u_i(t)$. Informally, $u_{s(j)}(t)$ represents the *rate* at which server $s(j)$ works at time t . The amount of time that the server $s(j)$ was idling during some time period $[t_1, t_2]$ can be measured in terms of fluid solution as $t_2 - t_1 - \sum_{i \in s(j)} (\bar{T}_i(t_2) - \bar{T}_i(t_1))$. It thus makes sense to define a fluid solution to be work conserving if the amount of idling is zero whenever the fluid level in the server is positive.

Definition 2. A fluid solution $(\bar{Q}(t), \bar{T}(t))$ is defined to be *work conserving* or *non-idling* if for every server $s(j)$ and every $t_1 < t_2$

$$\inf_{t_1 \leq s \leq t_2} \sum_{i \in s(j)} \bar{Q}_i(s) > 0,$$

we have

$$\sum_{i \in s(j)} (\bar{T}_i(t_2) - \bar{T}_i(t_1)) = t_2 - t_1.$$

Examples of Fluid Models and Fluid Solutions

Given a fluid model, does it have at least one fluid solution? The answer is trivially yes: set $\bar{T}(t) = 0$ for all t and $\bar{Q}(t) = \bar{Q}(0) + \alpha t$. Check that this is indeed a fluid solution. But it is not a very interesting one. It essentially corresponds to the case when all servers are idling all the time and, as a result, the queues are building up at a linear rate. A more interesting case is existence of at least one work-conserving fluid solution. The answer is again positive but this is harder to prove and we leave it as an exercise. Let us now turn to specific examples of fluid model.

G/G/1 Queueing System. We assume $\rho < 1$. For this and the following examples let us focus exclusively on work-conserving fluid solutions. We seek solutions to the system of equations and inequalities:

$$\bar{Q}(t) = \bar{Q}(0) + \alpha t - \mu \bar{T}(t),$$

$$\bar{T}(t) - \bar{T}(s) \leq t - s, \forall s < t,$$

Satisfying, in addition, the work-conservation requirement and $\bar{T}(0) = 0$. Observe that $\bar{T}(t) \leq t$ inequality implies that $\bar{Q}(t) \geq \bar{Q}(0) + \alpha t - \mu t > 0$ for all $t < \tau \triangleq \bar{Q}(0)/(\mu - \alpha)$. The work-conservation condition then implies $\bar{T}(t) = t$ for all $t < \tau$. As a result, we must have $\bar{Q}(t) = \bar{Q}(0) + \alpha t - \mu t$ for all such $t < \tau$. The continuity of the fluid solution implies that $\bar{Q}(\tau) = \bar{Q}(0) + \alpha \tau - \mu \tau = 0$. We have established the existence and uniqueness of the fluid solution at least on the interval $[0, \tau]$. What happens next? It is easy to check that $\bar{Q}(t) = 0, \bar{T}(t) = \tau + \rho(t - \tau)$ for $t \geq \tau$ is a fluid work-conserving solution (recall that we already informally derived this). Let us now show that it is in fact unique. Suppose there exists a solution $(\bar{Q}(t), \bar{T}(t))$ such that $\bar{Q}(t') \neq 0$ for some $t' > \tau$. By continuity of $\bar{Q}(t)$ we have that the set $\{t : \bar{Q}(t) \neq 0\}$ is an open subset of $[0, \infty)$. A theorem from real analysis [7] states that every open subset of a real line can be represented as a finite or countable union of non-overlapping open intervals $\cup_i (a_i, b_i)$. Thus t' must fall into one of those intervals, say (a_k, b_k) . Notice that $\tau \leq a_k$ (why?). By work conservation we have $\bar{T}(s) - \bar{T}(a_k) = s - a_k$ for all $s \in (a_k, b_k)$. We have $\bar{Q}(a_k) = 0$ and then from Equation (11) we obtain

$$\begin{aligned} \bar{Q}(t') &= \bar{Q}(a_k) + \alpha(t' - a_k) - \mu(t' - a_k) \\ &= \alpha(t' - a_k) - \mu(t' - a_k) \\ &< 0, \end{aligned}$$

since $\rho < 1$, but this is a contradiction. We conclude that the unique work-conserving fluid solution is

$$\bar{Q}(t) = \max(0, \bar{Q}(0) + \alpha t - \mu t),$$

$$\bar{T}(t) = \mathbf{1}\{t \leq \tau\}t + \mathbf{1}\{t > \tau\}(\tau + \rho(t - \tau)),$$

where $\tau = \bar{Q}(0)/(\mu - \alpha)$.

Characterizing fluid solutions explicitly for more complicated networks would be a daunting task. Instead, we are content with getting an insight into the *qualitative* behavior of fluid models. This becomes particularly important when we turn to the issues of stability of fluid models.

STABILITY OF A FLUID MODEL

We now turn to the important property—stability of a fluid model and its relationship to the stability of underlying queueing systems. From basic queueing theory we know that $G/G/1$ queue is stable if and only if $\rho < 1$. On the other hand, as we have seen, $\rho < 1$ is precisely the condition under which the unique work-conserving fluid solution of a $G/G/1$ queueing system converges to zero and stays at zero after some initial time. This is not a coincidence. We can think of $G/G/1$ being stable under the $\rho < 1$ condition as the property that *no matter how large* was the initial number of jobs in the queue, this large backlog will be cleared at some point and the system will enter the steady-state regime.

Definition 3. A fluid solution $(\bar{\mathbf{Q}}(t), \bar{\mathbf{T}}(t))$ is defined to be stable if there exists $\tau > 0$ such that $\bar{\mathbf{Q}}(t) = 0, \forall t \geq \tau$. A fluid model $(\alpha, \mu, \mathbf{P}, \mathbf{C})$ is defined to be globally stable if every work-conserving fluid solution is stable.

The word global is used in the definition to indicate the property of the network itself rather than the property of a particular fluid solution. The emptying time τ per our definition is solution specific, even though we see that in many interesting cases there exists one τ that works for *all* fluid solutions.

Observe that if the fluid solution $\bar{\mathbf{Q}}(t)$ becomes zero after time τ , then from Equation (10) we obtain

$$0 = \frac{d\bar{\mathbf{Q}}(t)}{dt} = \alpha + (\mathbf{P}' - \mathbf{I})\mathbf{M}\mathbf{u}(t), \quad (17)$$

namely, $\mathbf{u}(t) = \mathbf{M}^{-1}[\mathbf{I} - \mathbf{P}']^{-1}\alpha = \rho$. In other words, $\bar{\mathbf{T}}(t) = \bar{\mathbf{T}}(\tau) + \rho t$ for all $t \geq \tau$. We can interpret this observation as servers working just enough time to prevent fluid from building up from level zero.

Let us establish some preliminary results on stability. Our first result connects stability with the traffic intensity parameter ρ .

Theorem 1. *A necessary condition for a global stability is*

$$\rho_{s(j)} < 1 \quad (18)$$

for every $j = 1, \dots, J$.

Proof. Suppose the traffic intensity of some server $s(j)$ satisfies $\rho_{s(j)} \geq 1$. We have from Equations (13) and (8)

$$\begin{aligned} \bar{W}_{s(j)}(t) &= \bar{W}_{s(j)}(0) + \rho_{s(j)}t - \sum_{i \in s(j)} \bar{T}_i(t) \\ &\geq \bar{W}_{s(j)}(0) + \rho_{s(j)}t - t \\ &\geq \bar{W}_{s(j)}(0). \end{aligned}$$

Thus it suffices to exhibit a starting point $\bar{\mathbf{Q}}(0)$ such that $\bar{W}_{s(j)}(0) > 0$. Letting $\mathbf{C}_j$ denote the j th row of $\mathbf{C}$, we have

$$\begin{aligned} \bar{W}_{s(j)}(0) &= \mathbf{C}'_j \bar{\mathbf{W}}(0) \\ &= \mathbf{C}'_j \mathbf{M}^{-1}[\mathbf{I} - \mathbf{P}']^{-1} \bar{\mathbf{Q}}(0) \\ &= \mathbf{C}'_j \mathbf{M}^{-1} \left(\mathbf{I} + \sum_{m \geq 1} (\mathbf{P}')^m \right) \bar{\mathbf{Q}}(0) \\ &\geq \mathbf{C}'_j \mathbf{M}^{-1} \bar{\mathbf{Q}}(0) \\ &= \sum_{i \in s(j)} \mu_i^{-1} \bar{Q}_i(0). \end{aligned}$$

Thus every fluid solution with a starting point $\bar{\mathbf{Q}}(0)$ satisfying $\bar{Q}_{s(j)} > 0$ never reaches zero.

Thus, from now on we focus on the case when Equation (18) holds. Next we establish that this condition is also sufficient in the case of a single-server queueing system.

Theorem 2. *A single-server ($J = 1$) fluid model $(\alpha, \mu, \mathbf{P})$ is globally stable if and only if $\rho < 1$.*

Proof. Using Theorem 1, it suffices to assume $\rho < 1$. From Equation (13), we have that

the workload $\bar{W}(t)$ of our unique server satisfies

$$\bar{W}(t) = \bar{W}(0) + \rho t - \sum_i \bar{T}_i(t). \quad (19)$$

Owing to constraint (8), we have $\bar{W}(t) > 0$ for all $t \leq \tau \triangleq \bar{W}(0)/(1 - \rho)$, which implies $\bar{Q}(t) \neq 0, \forall t < \tau$, which further implies that $\sum_i \bar{T}_i(t) = t, \forall t < \tau$, by work conservation. It then follows that $\bar{W}(\tau) = \bar{W}(0) + \rho\tau - \tau = 0$, implying $\bar{Q}(\tau) = 0$, and we established that every work-conserving fluid solution reaches zero exactly at time τ . We claim further that $\bar{Q}(t) = 0, \forall t > \tau$, and the proof of this is a generalization of an argument already used in the special case of a fluid model corresponding to a $G/G/1$ queueing system. We leave it to the reader to complete the argument.

Unstable Fluid Model

The stability property would be simple if the traffic intensity condition (18) was also sufficient for stability. However, this is, unfortunately, not the case. We now present an example of a network such that Equation (18) is satisfied, yet an unstable fluid solution exists. Consider the two-server four classes network depicted in Fig. 1. We assume $\alpha = (\alpha, 0, 0, 0), \mathbf{P}_{ii+1} = 1, i = 1, 2, 3, P_{ij} = 0$ for all other j , otherwise; $\mathbf{C}_{11} = \mathbf{C}_{14} = \mathbf{C}_{22} = \mathbf{C}_{23} = 1, \mathbf{C}_{ij} = 0$ otherwise. This network is called *Lu-Kumar network* due to authors [8] who introduced this network. A similar phenomenon was introduced for the first time in a related work by Kumar and Seidman [9].

Theorem 3. *The Lu-Kumar network depicted in Fig. 1 is unstable if $\rho_2 + \rho_4 \geq 1$.*

The proof of this result is involved, but conceptually not complicated. The idea is to build a concrete solution satisfying the required properties. It leads to an alternating cycle behavior where a large amount of fluid is accumulated at either first or second server at different times. This behavior induces a starvation where a server nominally capable of processing jobs waits for these jobs for too long (“starves”).

Feedforward Fluid Models are Stable

As we said before, the reason that an unstable fluid solution exists in the network depicted in Fig. 1 has to do with the property that the two servers alternate in “starvation,” thus creating an increasing backlog of work. This was possible due to a mutual feedback existing between the two servers. It turns out that this feedback is necessary for the nontrivial mode of instability (i.e., instability in underloaded networks) to take place. We now show that if the feedback in the network is one sided and the network is underloaded, then it is also stable.

A fluid network $(\alpha, \mu, \mathbf{P}, \mathbf{C})$ is called *feedforward* or *acyclic* if for every $1 \leq j_1 < j_2 \leq J$, and every $i \in s(j_1), l \in s(j_2)$, we must have $p_{li} = 0$. In other words, jobs go through the servers in a nondecreasing order. A simple example of such a network is a tandem queueing network consisting of J servers. Another example is the network in Fig. 2.

Theorem 4. *A feedforward fluid network is stable if condition (18) is met.*

The proof is based on a certain induction argument. One can show that for every j , the server $s(j)$ and all the servers preceding it become empty at a certain time t_j , and the outflow from these servers becomes equal to the inflows. In other words, these servers practically “disappear” from the network. Each successive server can be analyzed as a single-server fluid MQNET.

LYAPUNOV FUNCTION METHOD

Now that we realize that stability is indeed a nontrivial property and is not simply captured by the load condition (18), it would be useful to have general methods that can help us to determine at least sufficient conditions for stability. Unfortunately, as we discuss in the last section, one cannot hope for establishing stability property exactly for every given network. However, for practical applications perhaps we can identify some methods that provide *sufficient* if not always *necessary* conditions for stability. One of such

most powerful methods is the Lyapunov function method, which draws from a general theory of differential equations and dynamical systems.

Definition 4. A function $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}_+$ is called a Lyapunov function if $\Phi(x) = 0$ if and only if $x = 0$ and there exists $\gamma > 0$ such that for every time interval (t_1, t_2) and every fluid solution satisfying $\bar{Q}(t) \neq 0, \forall t \in (t_1, t_2)$, we have

$$\Phi(\bar{Q}(t_2)) - \Phi(\bar{Q}(t_1)) \leq -\gamma(t_2 - t_1).$$

If Lyapunov function Φ is differentiable, then, using the chain differentiation rule, the condition can be restated equivalently as

$$\nabla \Phi \cdot \frac{d\bar{Q}(t)}{dt} \leq -\gamma,$$

for all t where $\bar{Q}(t) \neq 0$ and the derivative of $\bar{Q}(t)$ is defined. Here, $\nabla \Phi$ denotes the gradient of Φ .

The idea behind the Lyapunov function construction is to find a certain function (sometimes also called *potential*) of the state space which decreases with time as long as the state $\bar{Q}(t)$ is nonzero. The trick is to turn a multidimensional process $\bar{Q}(t)$ into a one-dimensional process $\Phi(\bar{Q}(t))$.

Let us now show that Lyapunov functions are useful because they provide sufficient conditions for stability.

Theorem 5. *Given a fluid network (α, μ, P, C) suppose a Lyapunov function Φ exists. Then the network is globally stable.*

Proof. Fix any $\tau > \Phi(\bar{Q}(0))/\gamma$. If $\bar{Q}(t) \neq 0$ for all $0 \leq t \leq \tau$, then $\Phi(\bar{Q}(\tau)) \leq \Phi(\bar{Q}(0)) - \gamma\tau < 0$ —contradiction. Thus $\bar{Q}(t_0) = 0$ for some $t_0 \leq \Phi(\bar{Q}(0))/\gamma$. Suppose $\bar{Q}(t) \neq 0$ for some $t > t_0$. In this case, $t_1 = \inf\{t \geq t_0 : \bar{Q}(t) \neq 0\}$ is finite. Then, by continuity, we can find that $t_2 > t_1 \geq t_0$ such that $\bar{Q}(t_1) = 0$ and $\bar{Q}(t) \neq 0$ for all $t \in (t_1, t_2]$. But then $\Phi(\bar{Q}(t_2)) \leq \Phi(\bar{Q}(t_1)) - \gamma(t_2 - t_1) = -\gamma(t_2 - t_1) < 0$ —contradiction again. We conclude, $\bar{Q}(t) = 0, \forall t \geq t_0$.

Let us now demonstrate Lyapunov function method with some examples.

Multiclass Single-Server Queueing System.

Recall (Theorem 2) that such a system is stable provided $\rho < 1$. In fact our proof is essentially based on constructing a Lyapunov function. Let $\Phi(x) = e'M^{-1}[I - P]^{-1}x$. Check that Φ is zero only at $x = 0$. Then $\Phi(\bar{Q}(t))$ is simply the workload $\bar{W}(t)$ of the server for which we have a fluid equation (19). Whenever $\bar{Q}(t) \neq 0$ over some interval, (t_1, t_2) we have $\sum_i (\bar{T}_i(t_2) - \bar{T}_i(t_1)) = t_2 - t_1$, implying

$$\bar{W}(t_2) - \bar{W}(t_1) \leq (\rho - 1)(t_2 - t_1).$$

Thus, the linear function Φ is indeed a Lyapunov function with $\gamma = 1 - \rho$.

Let us consider more sophisticated applications of the Lyapunov function method to the question of stability of fluid models. Our first example is generalized Jackson network (GJN) considered in the section titled “Multiclass Queueing Network: Model Formulation”. For simplicity, we call this model fluid Jackson network and denote it by (α, μ, P) . An interesting and deep result is that fluid Jackson network is always stable under the usual load conditions. In other words, the instability can only be caused by multiplicity of classes in the network.

Theorem 6. *A fluid Jackson network (α, μ, P) is globally stable provided the load condition (18) holds.*

The proof is based on showing that $\Phi(x) = e'M^{-1}[I - P]^{-1}x$ is a linear Lyapunov function [5] and [10]. It can be shown that a fluid Jackson network has, in fact, a unique non-idling solution for every starting state $Q(0)$. Thus, global stability in this case is equivalent to just stability.

For another demonstration of the power of the Lyapunov function method, let us return to the Lu–Kumar network depicted in Fig. 1. Recall Theorem 3 stating that this network can be unstable under the assumption $\rho_2 + \rho_4 \geq 1$. It turns out that this is the necessary condition for the existence of an unstable trajectory, in addition, of course, to Equation (18). The result below then gives us a complete global stability picture for this network.

Theorem 7. *The Lu-Kumar network is globally stable if and only if the load condition (18) holds and $\rho_2 + \rho_4 < 1$.*

Proof. In light of Theorem 3, it suffices to consider the case $\rho_2 + \rho_4 < 1$. We construct an appropriate Lyapunov function. It turns out that the following choice works (though it is not obvious yet why). Let

$$\theta_1 = 1 - \rho_4 - \min \left((1 - \rho_1 - \rho_4)/2, \right. \\ \left. \times (1 - \rho_2 - \rho_4)/3 \right) \quad (20)$$

$$\theta_2 = \rho_2 + \min \left((1 - \rho_2 - \rho_3)/2, \right. \\ \left. \times (1 - \rho_2 - \rho_4)/3 \right). \quad (21)$$

For every $x = (x_1, \dots, x_4) \in \mathbb{R}_+^4$, let

$$\Phi_1(x) = \theta_1 x_1 + (1 - \theta_1) \sum_{1 \leq i \leq 4} x_i, \\ \Phi_2(x) = \theta_2(x_1 + x_2) + (1 - \theta_2)(x_1 + x_2 + x_3).$$

Now define the Lyapunov function Φ as $\Phi(x) = \max(\Phi_1(x), \Phi_2(x))$. Observe that $\Phi(x)$ is a Lipschitz continuous function. Thus $\Phi(\bar{Q}(t))$ has derivatives for almost all t , along with $\bar{Q}_i(t), \bar{T}_i(t)$. Given such t , let $\mathcal{I} = \mathcal{I}(t) \subset \{1, 2\}$ such that $\Phi_i(\bar{Q}(t)) = \Phi(\bar{Q}(t))$ for every $i \in \mathcal{I}$ (the only choices are $\mathcal{I} = \{1\}, \{2\}, \{1, 2\}$). The following is left as an exercise:

$$\frac{d\Phi(\bar{Q}(t))}{dt} = \frac{d\Phi_i(\bar{Q}(t))}{dt} \quad (22)$$

for every $i \in \mathcal{I}$.

Now let us show that for each server $s(i), i = 1, 2$, if the total fluid level in this server is nonzero at time t , then $\frac{d\Phi_i(\bar{Q}(t))}{dt} \leq -\gamma$, for some positive γ . Consider server $s(1)$. We have

$$\frac{d\Phi_1(\bar{Q}(t))}{dt} = \theta_1(\alpha - \mu_1 u_1(t)) + (1 - \theta_1) \\ \times (\alpha - \mu_4 u_4(t)) \\ = \alpha - \theta_1 \mu_1 u_1(t) - (1 - \theta_1) \mu_4 u_4(t).$$

Now if $\sum_{i \in s(1)} \bar{Q}_i(t) = \bar{Q}_1(t) + \bar{Q}_4(t) > 0$, then $u_1(t) + u_4(t) = 1$. In light of this the

derivative is indeed negative, provided that $\alpha < \theta_1 \mu_1$ and $\alpha < (1 - \theta_1) \mu_4$. Check that this is indeed the case.

Now consider the server $s(2)$. We have

$$\frac{d\Phi_2(\bar{Q}(t))}{dt} = \theta_2(\alpha - \mu_2 u_1(t)) + (1 - \theta_2) \\ \times (\alpha - \mu_3 u_3(t)) \\ = \alpha - \theta_2 \mu_2 u_2(t) - (1 - \theta_2) \mu_3 u_3(t).$$

Again if $\sum_{i \in s(2)} \bar{Q}_i(t) > 0$, then $u_2(t) + u_3(t) = 1$, and the derivative is negative provided that $\alpha < \theta_2 \mu_2$ and $\alpha < (1 - \theta_2) \mu_3$. Check again that this is indeed the case.

Now we claim that if $\bar{Q}(t) \neq 0$ and $\sum_{i \in s(1)} \bar{Q}_i(t) = 0$, then $\Phi_1(\bar{Q}(t)) \leq \Phi_2(\bar{Q}(t))$, and similarly, if $\sum_{i \in s(2)} \bar{Q}_i(t) = 0$, then $\Phi_2(\bar{Q}(t)) \leq \Phi_1(\bar{Q}(t))$. Let us first see what this claim gives us. In light of Equation (22), this claim provides us with the conclusion that the derivative of $\Phi(\bar{Q}(t))$ equals the derivative of $\Phi_i(\bar{Q}(t)), i = 1, 2$, such that the corresponding server $s(i)$ has a positive amount of fluid. But we just showed that in this case the derivative of $\Phi_i(\bar{Q}(t))$ is negative and the proof of the theorem is complete. Thus, it remains to check the claim. First, suppose $\sum_{i \in s(2)} \bar{Q}_i(t) = 0$. Then,

$$\Phi_2(\bar{Q}(t)) = \bar{Q}_1(t) \leq \bar{Q}_1(t) \\ + (1 - \theta_1) \bar{Q}_4(t) = \Phi_1(\bar{Q}(t)),$$

and the assertion is established.

Now suppose $\sum_{i \in s(1)} \bar{Q}_i(t) = 0$. Then,

$$\Phi_1(\bar{Q}(t)) = (1 - \theta_1)(\bar{Q}_2(t) + \bar{Q}_3(t)), \\ \Phi_2(\bar{Q}(t)) = \bar{Q}_2(t) + (1 - \theta_2) \bar{Q}_3(t).$$

The claim follows since $1 - \theta_1 < 1 - \theta_2$, which is easily checked.

CONNECTIONS WITH STABILITY OF STOCHASTIC QUEUEING NETWORKS

Up until now we have not discussed two important questions. How does the stability property of a fluid model connect with the stability property of a stochastic MQNET, and do we have tools to determine stability of fluid

models exactly, (as opposed to just obtaining sufficient conditions provided, e.g., by constructing a Lyapunov functions)? These are perhaps the two most important questions. Because of the complexity of the underlying techniques involved, we only give a brief and somewhat informal summary of results. The appropriate references can be found in the next section.

The first one has to begin with a formal definition of stability of a queueing network. Intuitively, we say that a queueing network is stable if the queue length does not grow without a bound in some sense, say in expectation, as time goes to infinity. In classical queueing models, such as M/M/N queueing system, the formal definition of stability is positive recurrence of the underlying Markovian process. This notion can be generalized to multiclass queueing networks using the notion of so-called positive Harris recurrence. The Harris recurrence is an appropriate framework to discuss recurrence properties of Markov chains in general (not necessarily countable) state space. It turns out that there indeed exists a very important, albeit one directional, connection between the stability of a fluid model and its stochastic counterpart.

Theorem 8. *Suppose a fluid model is globally stable. Then, the underlying MQNET is positive Harris recurrent for every work-conserving scheduling policy.*

The theorem as stated is not completely precise. In fact, it was stated and proven for a broad class of so-called head-of-the-line type scheduling policies—policies that within each class use the FIFO rule. The notion of scheduling policies in MQNETs requires a special technical discussion on its own which we skip. Unfortunately, the converse result does not necessarily hold.

Theorem 9. *There exists an MQNET and a (priority) scheduling policy such that the corresponding fluid solution is unstable, yet the network is positive Harris recurrent.*

A similar result is known also for the FIFO policy. Again care is needed in order to properly define the “corresponding” fluid solution.

Sometimes this is defined as a weak limit of stochastic MQNET processes under the law of large numbers type of rescaling. In general, this weak limit is sometimes called *fluid limit* as opposed to a fluid solution (which, in general, might not be a weak limit corresponding to any policy). Other times it is defined as fluid solution satisfying some *implications* of implementing a particular scheduling policy. This choice can be problematic as there is no guarantee that all the relevant implications of a given scheduling rule are properly captured in the fluid model.

In some special cases, more concrete results can be established.

Theorem 10. *Suppose an MQNET with two servers ($J = 2$) admits a fluid solution which diverges to infinity. Then, there exists a work-conserving scheduling policy in the underlying MQNET such that the network is not positive Harris recurrent.*

In fact, a stronger result holds—under this scheduling policy the queue lengths diverge to infinity with probability 1 as time diverges to infinity. This result still leaves open the intermediate case when the fluid model is not globally stable, and no fluid solution exists which diverges to infinity. Several more general converse results are known (see next section for references).

As far as obtaining tight stability characterization of fluid models or of stochastic MQNETs for that matter, the positive results are limited. The only known result concerning general networks is the following characterization of globally stable fluid networks with two servers.

Theorem 11. *There exists a homogeneous linear programming problem associated with an arbitrary fluid MQNET. When the MQNET has two servers ($J = 2$), the network is globally stable if and only if this linear programming problem has a nonzero solution. In the special case when the routing process is deterministic, namely all entries of $\mathbf{P}$ are zero or 1, one can compute explicitly a certain parameter ρ^* called maximum*

virtual traffic intensity. The fluid network is stable if and only if $\rho^* < 1$.

There are additional tight stability results which, however, concern networks with very specific assumptions, such as the four class two-server network Of Lu–Kumar, which is depicted in Fig. 1. Short of these results, there does not exist a general stability criterion for stochastic or fluid MQNET. The explanation for the lack of such results comes from negative results established in Gamarnik [11] and Gamarnik and Katz-Rogozhnikov [12], where the question is placed on an algorithmic theoretic framework. Specifically, it has been shown that stability of a multiclass queueing networks is an undecidable property, namely, it cannot be resolved using any effective computational procedure. Because of space constraint the notion of undecidability is not defined in this article; the interested reader is referred to a very nice book by Sipser [13] instead, or to any textbook on the theory of algorithms.

On the positive side, there are a variety of scheduling policies and their fluid counterparts which stabilize MQNET. For example, when the network has a special structure known as *re-entrant line* MQNET [6,14], it is known that two particular priority policies known as first-buffer-first-serve and last-buffer-first-serve stabilize the system both in fluid and stochastic setting. In more general setting, it is known that the so-called global FIFO (policy that gives priority to jobs which arrive earlier to the *entire* network) is stable [15]. Finally, fluid models play a critical role in the analysis of diffusion approximation of queueing networks, especially in the multiclass setting, and are connected to a very important concept of *multiplicative* state space collapse [16,17].

NOTES AND SOURCES

Instability phenomena in underloaded multiclass queueing networks were first observed by Kumar and Seidman [9], who demonstrated the existence of an unstable work-conserving scheduling policy. A similar phenomenon was described in a related work

by Lu and Kumar [8], who introduced the network, which is depicted in Fig. 1. Rybko and Stolyar [18] obtained a similar result in a similar network and used fluid model technique for their analysis. Bramson [19] and Seidman [20] have shown that these instability phenomena can take place even for the FIFO service discipline. A comprehensive reference for stability and positive recurrence properties of general state space Markov chains is Meyn and Tweedie [21]. Fluid models of queueing systems have been considered for a long time [1], with transport processes being one of the earliest applications. The applications to MQNET have been developed in a series of papers [5,18,22,23]. A good book reference for this topic is [10]. The results showing that stability of fluid models implies stability of underlying stochastic networks (Theorem 8) were established by Dai [5] for general multiclass queueing networks and by Stolyar [23] for networks with exponentially distributed interarrival and service times. Partial converse results were established by Dai [24], Meyn [25], and Puhalskii and Rybko [26]. Some counterexamples showing that stability of fluid models does not necessarily coincide with stability of stochastic networks, including Theorem 9, were obtained by Dai *et al.* [27] and Bramson [28]. Theorem 10 was obtained by Gamarnik and Hasenbein [29]. Tight characterization of stable fluid models with two servers (Theorem 11) was obtained by Bertsimas *et al.* [30] using a trajectory decomposition method, and in the special case of 0-1 matrices $\mathbf{P}$ by Dai and Vande-Vate [31] using the Lyapunov function technique. Stability of feedforward networks (Theorem 4) was established by Dai [5] for fluid networks, and by Down and Meyn [32] directly for stochastic feedforward networks. Stability of stochastic Jackson network with generally distributed interarrival and service times was shown by Sigman [33] and Meyn and Down [34], and for fluid models (Theorem 2) by Dai [5]. The undecidability results mentioned in the previous section were obtained by Gamarnik and Katz [11], building on an earlier work by Gamarnik [12] for stability of random walks and their fluid models in a nonnegative orthant [35]. For a good reference on Turing

machines and decidability (computability), refer to Sipser [13]. A comprehensive study of random walks in a nonnegative orthant can be found in Fayolle *et al.* [36].

REFERENCES

1. Kleinrock L. Queueing systems. Canada: John Wiley and Sons, Inc.; 1975.
2. Kumar PR. Scheduling semiconductor manufacturing plants. *IEEE Control Syst Mag* 1994;14(6):30–40.
3. Kumar S, Kumar PR. Queueing network models in the design and analysis of semiconductor wafer fabs. *IEEE Trans Rob Autom* 2001;17(5):548–561.
4. Bramson M. Convergence to equilibria for fluid models of head-of-the-line proportional processor sharing queueing networks. *Queueing Syst Theory Appl* 1996;23(1-4):1–26.
5. Dai JG. On the positive Harris recurrence for multiclass queueing networks: a unified approach via fluid models. *Ann Appl Probab* 1995;5:49–77.
6. Dai JG, Weiss G. Stability and instability of fluid models for certain re-entrant lines. *Math Oper Res* 1996;21:115–134.
7. Royden H. Real analysis. 3d ed. New Jersey: Prentice Hall; 1988.
8. Lu SH, Kumar PR. Distributed scheduling based on due dates and buffer priorities. *IEEE Trans Autom Control* 1991;36:1406–1416.
9. Kumar PR, Seidman TI. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans Autom Control* 1990;AC-35:289–298.
10. Chen H, Yao D. Fundamentals of queueing networks: performance, asymptotics and optimization. New York: Springer; 2001.
11. Gamarnik D, Katz-Rogozhnikov D. On deciding stability of queueing networks under priority scheduling policy. *Ann Appl Probab* 2009;19:2008–2037.
12. Gamarnik D. On deciding stability of constrained homogeneous random walks and queueing systems. *Math Oper Res* 2002;27(2):272–293.
13. Sipser M. Introduction to the theory of computability. Boston (MA): PWS Publishing Company; 1997.
14. Kumar PR. Re-entrant lines. *Queueing Syst Theory Appl* 1993;13(1-3):87–110. (Special Issue on Queueing Networks.)
15. Bramson M. Stability of earliest-due-date first-served queueing networks. *Queueing Syst* 2001;39(1):79–102.
16. Bramson M. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst Theory Appl* 1998;30:89–148.
17. Williams RJ. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst Theory Appl* 1998;30:27–88.
18. Rybko A, Stolyar A. On the ergodicity of stochastic processes describing open queueing networks. *Probl Peredachi Inform* 1992;28:3–26.
19. Bramson M. Instability of FIFO queueing networks. *Ann Appl Probab* 1994;2:414–431.
20. Seidman TI. First come first serve can be unstable. *IEEE Trans Autom Control* 1994;39:2166–2170.
21. Meyn SP, Tweedie RL. Markov chains and stochastic stability. London: Springer; 1993.
22. Chen H, Mandelbaum A. Discrete flow networks: bottleneck analysis and fluid approximations. *Math Oper Res* 1991;16:408–446.
23. Stolyar A. On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes. *Markov Processes Relat Fields* 1995;1(4):491–512.
24. Dai JG. A fluid-limit model criterion for instability of multiclass queueing networks. *Ann Appl Probab* 1996;6:751–757.
25. Meyn SP. Transience of multiclass queueing networks and their fluid models. *Ann Appl Probab* 1995;5:946–957.
26. Puhalskii AA, Rybko AN. Nonergodicity of a queueing network under nonstability of its fluid model. *Probl Inf Trans* 2000;36:26–41.
27. Dai JG, Hasenbein JJ, Vande Vate JH. Stability and instability of a two-station queueing network. *Ann Appl Probab* 2004;14:326–377.
28. Bramson M. A stable queueing network with unstable fluid model. *Ann Appl Probab* 1999;9(3):818–853.
29. Gamarnik D, Hasenbein J. Instability in stochastic and fluid queueing networks. *Ann Appl Probab* 2005;15(3):1652–1690.
30. Bertsimas D, Gamarnik D, Tsitsiklis J. Stability conditions for multiclass fluid queueing networks. *IEEE Trans Autom Control* 1996;41:1618–1631.

31. Dai JG, Vande Vate JH. The stability of two-station multi-type fluid networks. *Oper Res* 2000;48:721–744.
32. Down DD, Meyn SP. Stability of acyclic multi-class queueing networks. *IEEE Trans Autom Control* 1995;40(5):916–920.
33. Sigman K. The stability of open queueing networks, stochastic processes and their applications. *Stoch Processes Appl* 1990;35: 11–25.
34. Meyn SP, Down D. Stability of generalized Jackson networks. *Ann Appl Probab* 1994;4: 124–148.
35. Gamarnik D. Computing stationary probability distribution and large deviations rates for constrained homogeneous random walks. The undecidability results. *Math Oper Res* 2007;27(2):272–293.
36. Fayolle G, Malyshev VA, Menshikov MV. Topics in the constructive theory of countable Markov chains. Cambridge: Cambridge University Press; 1995.

FORECASTING APPROACHES FOR THE HIGH-TECH INDUSTRY

MICHAEL GILLILAND
SAS Institute, SAS Campus Drive,
Cary, North Carolina

Forecasting in high-tech industries is beset by most of the same challenges faced in other industries. However, two factors that make high-tech forecasting more difficult are frequent new product introductions (with short product life cycles) and long supply chain lead times (due to off-shore production). New products—since they have no sales history to go by—are inherently more difficult to forecast than existing products. Long supply lead times require forecasts to be made farther into the future—and longer range forecasts tend to be less accurate than forecasts for the near term. These challenges make it highly unrealistic to expect accurate forecasts in the high-tech industry, yet there are methods for high-tech organizations to deal with this extra uncertainty. This article will focus on two such methods: The “structured analogy” approach to new product forecasting and forecast value added (FVA) analysis. Neither approach can guarantee highly accurate forecasts (nothing can do that). However, these methods permit the high-tech organization to generate forecasts efficiently (in less time, at less cost), and to better understand and deal with the uncertainties and risks.

NEW PRODUCT FORECASTING

There are many types of new product forecasting (NPF) situations that high-tech organizations can encounter:

- new-to-the-world products (entirely new types of products),
- new markets for existing products (geographical expansion or new sales channels),
- refinements to existing products (such as “new and improved” versions, or packaging changes).

Kenneth Kahn’s book, *New Product Forecasting: An Applied Approach* [1], provides a practical “toolbox” approach to illustrate how new product forecasts can be generated. The approach can vary significantly depending on the nature of the new product (whether new-to-the-world or the simple refinement of an existing product), and the nature of the market into which it will be sold (new or existing). This article will focus on the forecasting of new products in existing markets, where the most commonly used method is forecasting by analogy. Forecasting by analogy assumes that demand for a new product will be similar to demand for *like-items* from the past. Like-items are identified by common attributes or other criteria, and the forecast for the new item is based on a composite of the sales history of the like-items.

The real estate market provides a good example of the use of analogies. To determine the listing price for a property that is new on the market, the sales agent will prepare a list of “comps” (comparable houses in the area) that are currently on the market or have been recently sold. The prices of these comps (which serve as like-items) are used to suggest an appropriate listing price for the new property (which is, in effect, a forecast of what the property should sell for).

All NPF forecasting methods utilize judgment to some extent. While the use of judgment is necessary, it is frequently biased with over-optimism, or can be overly influenced by recent events. Judgment can also be clouded by personal or political agendas, where the forecast is used to represent what the organization wants to happen, rather than what anyone honestly believes will happen.

In practice, the forecasting by analogy approach is subject to similar misuse. Like-items that failed in the marketplace

may be forgotten (or purposely ignored), leading to overly optimistic expectations for the new product. There is also the danger of overconfidence in the accuracy of forecasts for the new product, and failure to appreciate the true uncertainty and risks involved.

New Product Forecasting by Structured Analogy

An emerging method for new product forecasting helps address some of these concerns. The “structured analogy” approach utilizes data visualization and analytical software to augment the use of analogies with structured judgment. This approach seeks to remove judgmental bias by providing a historical context for each decision [2].

The structured analogy approach requires two types of data: (i) attributes (of prior products and new products) and (ii) historical sales (of prior products). Product attributes might include:

- Product type (PC, mainframe, router, cpu, graphics chip, sound card, etc.)
- Target Market (consumer, industrial)
- Time of introduction (summer, winter, fiscal quarter, etc.)
- Financial characteristics (price point, competitor prices, etc.)
- Physical characteristics (memory, storage, processor speed, case size, etc.)

Historical information on past new product introductions is also needed. Sales of all

prior new product introductions (e.g., the first 20 weeks) are extracted and manipulated to align them all on a common time scale (weeks since product release). This permits the sales profiles of like-items to be overlaid to allow for statistical and visual assessment of the range of past new product outcomes.

Structured Analogy Process

The structured analogy process for new product forecasting has six main steps:

Step 1: Query. Identify a set of *candidate* products that have similar attributes to the new product.

Step 2: Filter. Option to manually remove inappropriate or outlier products from the set of candidate products.

Step 3: Cluster. Cluster the candidate products according to their sales pattern and select the most appropriate cluster to serve as the *surrogate* products.

Step 4: Model. Select the most appropriate statistical model for the cluster of surrogate products.

Step 5: Forecast. Use the statistical model to generate a forecast for the new product.

Step 6: Override. Option to make manual adjustments to the statistical model’s forecast.

Figure 1 illustrates the result of the query step for DVD movie sales, where like-items (based on user-specified product attributes for Genre and MPAA Rating) have been extracted and their profiles aligned. These

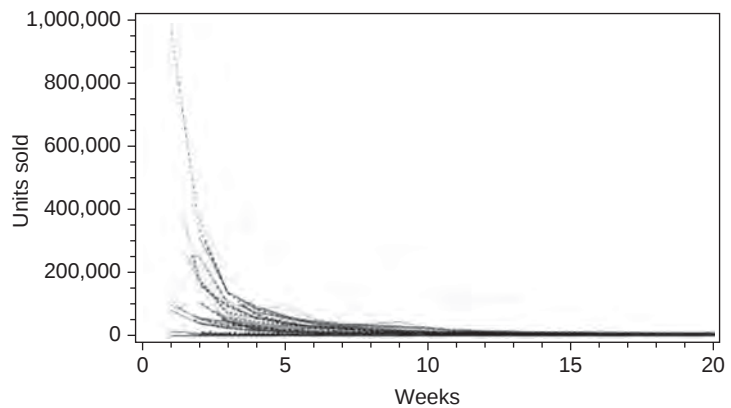


Figure 1. Profiles of like-item sales over first 20 weeks.

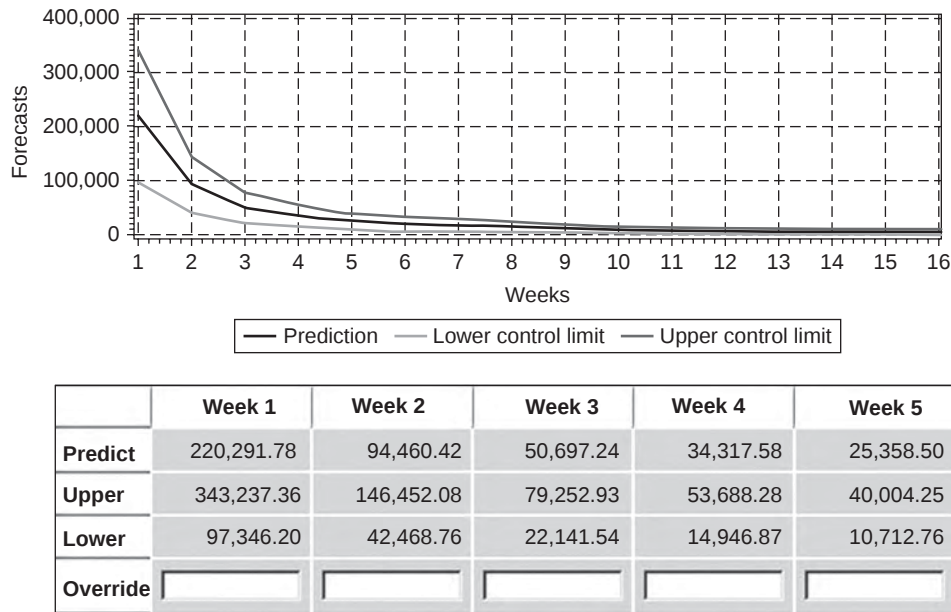


Figure 2. Statistical forecast and option for manual override.

particular like-items display a common type of profile (sales rapidly fall after the first week of release), yet the volume of sales varies considerably (from under a thousand to nearly one million units in the first week).

Note that judgment was used to determine which attributes were most relevant (here, Genre and MPAA rating). Judgment is again used in the filter step, where the forecaster can remove items from the candidate pool. After the filtering is applied, the cluster step groups the remaining candidates according to similarity of their sales pattern. Again, judgment is applied to select which cluster is most appropriate for forecasting the new product. This forms the group of *surrogate* products which are used to create the forecast model.

The model step generates a statistical model fitting the surrogate cluster. This can be a simple composite average of all the surrogate patterns, or more elaborate modeling techniques may be used. Once the model is specified, the forecast step generates a forecast for the new product. The final override step allows the forecaster to manually adjust the system generated forecast, and pass the final forecasts to downstream planning systems (Fig. 2).

Applying the Structured Analogy Approach

The structured analogy approach can be useful in many (but not all) new product forecasting situations [3]. It augments human judgment by automating the historical data handling and extraction, by incorporating statistical analysis, and by providing visualization of the range of historical outcomes. Software makes it possible to quickly extract candidate products based on the user-specified attribute criteria. It aligns, scales, and clusters the historical patterns automatically, making it easier to visualize the behavior of past new products. This visualization helps the forecaster realize the risks, uncertainties, and variability in new product behavior.

Expectations for the accuracy of new product forecasts should be modest, and acknowledgment of this uncertainty should be at the forefront. The structured analogy approach allows the organization to both statistically and visually assess the likely range of new product demand, so that it can manage accordingly. Rather than lock in elaborate sales and supply plans based on a point forecast that is likely to be wrong (and

possibly very wrong), the organization can use the structured analogy process to assess alternative demand scenarios and mitigate risk.

Judgment is always going to be a big part of new product forecasting, but judgment needs assistance to minimize biases and keep it as objective as possible. While the structured analogy approach can be used to generate new product forecasts, it is also of great value in assessing the *reasonableness* of forecasts that are provided from elsewhere in the organization. The role of structured analogy software is to do the heavy computational work and provide guidance—making the NPF process as automated, efficient, and objective as possible.

FORECAST VALUE ADDED ANALYSIS

Forecast accuracy is largely determined by the nature of the demand pattern being forecast—its *forecastability*. Smooth, stable, and repeating patterns can be forecast accurately with simple methods, but the high-tech industry rarely enjoys this luxury. Instead, elaborate models and processes are developed to forecast the volatile demand patterns that high-tech organizations face. But do these costly and time-consuming efforts work—do they make the forecast any better than a simple model, such as a moving average? This is the question addressed by forecast value added (FVA) analysis [3].

Traditional forecasting performance metrics, such as mean absolute percent error (MAPE) indicate the magnitude of the forecast error. But traditional metrics tell the organization nothing about what accuracy is reasonable to expect, or whether their efforts are “adding value” by making the forecast better. By definition

FVA $\equiv$ The change in a forecasting performance metric that can be attributed to a particular step or participant in the forecasting process.

FVA analysis compares the forecast at each step in a process to the forecast in the prior step. For example, in the simple process shown in Fig. 3, the analyst override is compared to the statistical forecast (generated by the forecasting software), which in turn is compared to a “naïve” forecast. A naïve forecast is something easy to compute, requiring the minimum amount of cost and effort. Typical examples are a random walk (using the last known actual as the future forecast), a seasonal random walk (such as using the actual from the corresponding period last year as the forecast for each period this year), or a moving average. If a process step (such as an elaborate statistical model or a manual override) is not making the forecast better, it can be eliminated from the forecasting process.

FVA analysis is consistent with a “lean” approach to supply chain management. The method has been utilized at major organizations inside and outside the high-tech industry, as a way to identify the waste and inefficiency in a forecasting process. The naïve model provides what is, in effect, a “placebo” forecast against which all other efforts are compared. Many organizations have found that their elaborate forecasting processes are just making the forecast worse.

Figure 4 illustrates an FVA report for the process in Fig. 3. This “stairstep” report shows what each process step was able to achieve, and computes its “forecast value added” compared to prior steps. In this example, we see that the statistical forecast reduced forecast error (MAPE) by 10% points compared to the naïve model. The analyst override did perform 5% points better than the naïve model, but 5% points worse than the statistical model. In such a situation, the organization may decide to eliminate (or sharply restrict) analyst overrides, and rely on the statistical forecast.

FVA analysis has several uses in high-tech organizations. Given the volatile demand patterns common in high-tech organizations,

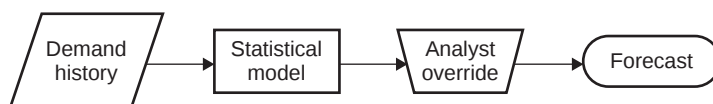


Figure 3. Simple forecasting process.

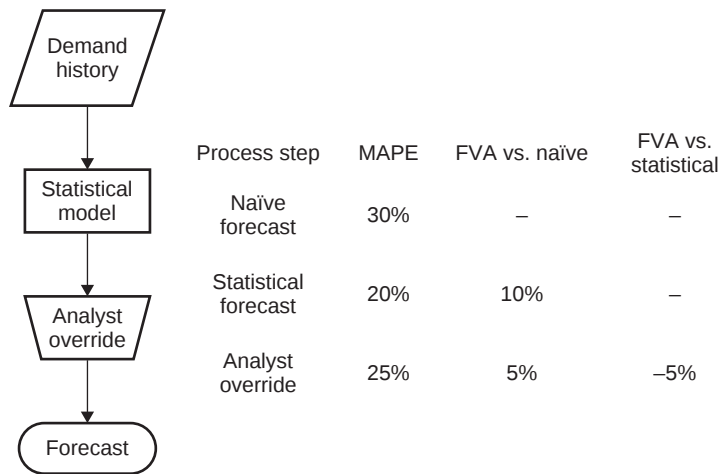


Figure 4. FVA report for simple forecasting process.

they tend to have unrealistic expectations for forecast accuracy. Achieving the desired degree of accuracy may be impossible, leading the organization to over-invest in the forecasting function rather than look for other (nonforecasting) solutions to their business challenges.

The naïve model used in FVA analysis provides a realistic baseline for forecast accuracy expectations. It is reasonable to expect the organization to forecast *no worse than* the naïve model, and any process steps making the forecast worse can be safely eliminated. Even after bad practices have been eliminated, it can be surprisingly difficult to forecast significantly better than the naïve model.

FVA Shortcut: In-Sample Analysis of Forecastability

FVA analysis can be viewed as an “after-the-fact” approach that compares the out-of-sample performance of a complicated statistical technique (or elaborate forecasting process) with the out-of-sample performance of a naïve method. As such, it is appropriate for evaluating an organization’s existing forecasting approach, and determining (based on historical performance) whether the organization’s forecasting activities were making the forecast any better—or just making it worse. Such information would be of great value if it were known *before* elaborate and expensive methods were enacted—to avoid

the cost of enacting those methods that will fail to provide improvement. As a shortcut to full FVA analysis, in-sample analysis of *forecastability* may provide some answers.

The Spring 2009 issue of *Foresight: The International Journal of Applied Forecasting* contains several articles on the topic of forecastability. As Editor Len Tashman states in his preview to this special section,

Assessing the forecastability of a time series can give us a basis for judging how successful we’ve been in modeling the historical data and how much improvement we can still hope to attain [4].

Analysis of forecastability can drive better decisions on the configuration of complex forecasting systems. For example, if a time series is found to be overly volatile and “nonforecastable,” then overly elaborate methods may tend to “overfit” the historical behavior and not generate better forecasts. In this situation a simple statistical technique (e.g., random walk, moving average, or single exponential smoothing) could be used to generate forecasts as accurate as can reasonably be expected given the nature of the behavior, with no further effort applied. Effort could instead be applied to the more forecastable patterns—those exhibiting systematic behavior—that could benefit from the application of more sophisticated modeling or process approaches.

In-sample analysis cannot provide the more thorough and definitive performance answers that FVA analysis can provide. However, it can be done “before-the-fact” of major resource investments (such as a new forecasting system, or a more elaborate process), and help drive more informed decisions on the likely value of these new investments.

Handling Unforecastable Demand

For items that are determined to be unforecastable, or at least not forecastable to the degree of accuracy desired, the organization can seek nonforecasting related approaches. Two such approaches are supply chain reengineering and demand smoothing.

When an organization’s supply chain is extremely fast and flexible, and it can respond to demand as it occurs, then highly accurate forecasting (at least at the product level) is no longer a requirement. Forecasting efforts can be pushed to higher levels of aggregation, focusing on raw materials, or staffing, or capacity requirements, rather than on individual products. The supply chain reengineering approach seeks improvements in supply chain performance, to render forecast accuracy less critical.

The demand smoothing approach endeavors to “make the demand forecastable” by minimizing organizational practices that increase demand volatility. Common sales, marketing, and financial practices (such as sales contests, pricing and promotional activity, and quarter end revenue targets), tend to increase the volatility of demand, and thus make it more difficult to forecast accurately. Policies and incentives (both internal, and for customers) can be redesigned to encourage more predictable behavior. This can result in better forecasts without any new investment in forecasting systems or processes.

SUMMARY

This article has addressed two frequently occurring problems in high-tech industry forecasting. Regarding new product forecasting, the structured analogy approach helps management better understand the

uncertainty in new product forecasts, and the likely range of outcomes. This avoids the common mistake of putting too much confidence in the forecast, and taking on unwarranted risks. The structured analogy approach does not assure highly accurate forecasts—no method can do that. Its purpose is to augment management judgment—delivering objective and data-based guidance on the reasonableness of new product forecasts, whether these forecasts were generated by the process itself or provided from elsewhere in the organization.

The objective of the forecasting process is to generate forecasts as accurate as can reasonably be expected, and to do this as efficiently as possible. Forecasts are often made worse by organization politics and personal agendas that contaminate the process with needless manual overrides and executive approval steps. Traditional forecasting performance metrics, by themselves, do not identify these wasted efforts. FVA analysis is a method now used, in high-tech and other industries, to identify wasted efforts that cost time and money, yet fail to improve the forecast.

REFERENCES

1. Kahn K. New product forecasting: An applied approach. Armonk, NY: M.E. Sharp; 2006.
2. Gilliland M, Guseman S. Forecasting new products by structured analogy. *J Bus Forecast* 2009–2010;28(4):12–15.
3. Gilliland M. The business forecasting deal: exposing myths, eliminating bad practices, providing practical solutions. Hoboken, NJ: John Wiley & Sons; 2010.
4. Tashman L. Preview: Special feature on assessing forecastability. *Foresight: Int J Appl Forecast* 2009;13(Spring):23.

FURTHER READING

- Makridakis Spiros, Wheelwright S, Hyndman Rob. Forecasting methods and applications. 3rd ed. New York: John Wiley & Sons; 1998.
- Wallace T, Stahl R. Sales forecasting: A new approach. Cincinnati, OH: T.F. Wallace & Co.; 2002.

FORECASTING FOR INVENTORY PLANNING UNDER CORRELATED DEMAND

MELIKE BAYKAL-GÜRSOY
Department of Industrial and
Systems Engineering, RUTCOR,
CAIT, Rutgers University, New
Brunswick, New Jersey

NESIM K. ERKIP
Department of Industrial
Engineering, Bilkent University,
Bilkent, Ankara, Turkey

INTRODUCTION

When selling similar products, mainly due to product substitution by customers in case of stock outs and price/brand concerns, demand for a particular product may depend on the inventory positions of other products. Thus, the effective demand for a particular product is crosscorrelated to the demand of other products. Also, demand for each product might be autocorrelated. Demand created by advertisement or price discounts today might reduce demand in the future. From time to time, the demand might also experience other disturbances that are due to the current economic or political conditions. These dependencies make it challenging for a firm to generate accurate demand forecast for each product and to determine the “right” order quantities so as to maximize its own profit. Survey results reported at the Harvard/Wharton Merchandising Effectiveness Project [1] have found that these subjective forecasts tend to have an average forecasting error of 50% or more. As a result, some firms buy too little of some products resulting in lost sales and profit margin, and some firms buy too much of some products resulting in excess supply that must be marked down after a while, frequently to the point where the product is sold at a loss. A survey conducted by

a major retailer and reported in the New York Times on June 2, 1994, concluded that 50% of customers did not purchase products when they visited the store and of these 40% stated they did so due to their inability to find a given product. Similar under- and oversupply costs have been reported for other products such as automobiles [2] and computers [3]. Demand uncertainty is highest when a new product is introduced against competition. In some industries, for example, in the pharmaceutical industry, firms prefer to be responsive to the customer demand and thus incur high inventory costs to reduce lost sale. This is where accuracy in forecasting becomes crucial.

Firms spend a great deal of time and resources trying to predict, as accurately as possible, the future demand for their products and services. Clearly, it is beneficial to know about the future, and, in most cases, the near future where accuracy is most needed. Knowing about the future enables making better decisions of ordering when inventories are reviewed—both in pull- and push-type systems. Also, production plans and resource scheduling are done more effectively in push-type systems if we have a good idea about the future.

Two issues that are of concern here are forecasting and inventory control for multi-item systems where demand for these items can be correlated, as well as temporally correlated for each item. Classical forecasting procedures are usually carried out for single items. Standard coverage of well-known text books includes the following topics: times series analysis, exponential smoothing methods (see *Holt–Winters Exponential Smoothing*), moving average (MA) methods, regression with various possible extensions [4–8]. These references also include the study of nonstationary autoregressive integrated moving average (ARIMA) models (see *Forecasting Nonstationary Processes*). Nevertheless, a number of procedures can be utilized for the forecasting activities under correlated demand. For example,

regression models can be utilized to express several dependencies. However, more specialized tools, for example, multivariate time series modeling, state-space methods (see ***Forecasting: State-Space Models and Kalman Filter Estimation***), are needed for the purpose of forecasting for multiple items with correlation across items and time.

Most of the initial work done in the area of inventory theory has ignored the correlation between demand for different items as well as time periods. Instead, the demand is assumed to follow an independent stochastic process in order to ensure tractability of analytical results. However, more realistic models incorporate the past history of the market on the present demand behavior. Intuitively, it can be said that autocorrelated demand would worsen performance measures such as expected number in the warehouse and expected number of stock outs. It is important to investigate how profound the effect of demand autocorrelation is on system performance. Blinder [9] investigates the change in the optimum price and optimum order size to the demand fluctuations and shows that they will be larger when demand is highly correlated. Kahn [10], assuming a first-order autoregressive, AR(1), demand model, demonstrates that the order size is more variable than sales if demand is positively correlated. Altioek and Melamed [11] study the impact of autocorrelation on several production/inventory systems. They consider three different environments; an M/M/1 (see the section titled “Single-Station Queues: CTMC Models” in this encyclopedia) type workstation receiving autocorrelated arrival process; an M/G/1 (see the section titled “Single-Station Queues: Non-CTMC Models” in this encyclopedia) workstation with deterministic processing times, and autocorrelated operation dependent time to failures; and a single-stage manufacturing process with a raw material buffer and finished product inventory. They incorporate autocorrelation in random elements in the system, such as demand, lead time, time to failure, and repair times and investigate the effect on some system performance measures, such as average product waiting/flow times, throughput, and customer service

levels. The authors conclude that finished product level and customer service level are highly affected by the autocorrelation in the demand process [11].

We think that the following questions are valid for inventory analysis:

- How can one represent the demand function(s) that consider(s) the existing correlation structure?
- How does the level of demand correlation affect order quantities and expected profits over time?
- What is the optimal inventory control policy under correlated demand?

The first question represents the forecasting aspect, and the second and third ones represent the inventory control aspect. We aim to overview the inventory approaches in order to ascertain the demand-forecasting requirements. We then review forecasting procedures that will be helpful in realizing the aforementioned inventory models.

The next section briefly discusses the previous work on single and multiple item inventory control under correlated demand. In the section titled “Evolving Forecasts Through Time”, we examine the notion of evolving forecasts, including the Martingale model of forecast evolution (MMFE). We then summarize our conclusions.

INVENTORY PLANNING UNDER CORRELATED DEMAND

The bulk of inventory theory is based on single-item problems, with known independent and identically distributed demand. For the newsvendor problem (single period), the optimal policy is given by the *critical fractile* solution for the optimal stock level S^* [12] (see ***Newsvendor Models***). In the case of a normally distributed demand with mean μ and variance σ^2 , the solution simplifies to

$$S^* = \mu + z\sigma,$$

where z is obtained by evaluating the cumulative distribution function of the standard normal at the critical fractile. Scarf [13] has

shown the optimality of (s_t, S_t) policy for the dynamic inventory control problem with setup cost, K , under the condition that the optimal value function at each period t is K -convex. In such a policy, it is optimal to order up to the optimal stock level S whenever the inventory level drops to the reorder level, s , or below. If the setup cost, K , is zero, the optimal policy reduces to the base-stock policy that prescribes ordering up to S whenever the inventory level is less than S . It was shown by Veinott [14] and Sobel [15] that under certain conditions the optimal base-stock policy is myopic, that is, it can be decided only by considering the current period. Myopic policies are easier to implement and thus are preferable to the more complicated time-dependent policies. Veinott [14] has extended his results on the optimality of myopic policies to the demand distributions that are independent but not identical as long as they are stochastically increasing over time.

In this section, we review inventory models with correlated demand. The subsections consider single/multi item–single location and single/multi-item–multiple locations, with correlated demand.

Single Item–Single Location with Correlated Demand

We first review a broader group of studies referred to here as *inventory problems under nonstationary demand*. In these models, the demand distribution can vary during each period and is considered mostly in the context of single-product single-firm inventory control. Karlin and Fabens [16] introduce a Markov-modulated demand model that depends on the changing environmental factors. In this model, environmental factors are represented as the states of a Markov chain (see the section titled “Discrete Time Markov Chains” in this encyclopedia) and the demand in any given period is a random variable with a distribution function that depends on the current state of the environment. Thus, in this model, demands are independent and identically distributed as long as the environment’s state remains the same. Aware of the complexity of the analysis, the authors concentrate on finding a single optimal ordering policy irrespective of the state of the

environment. Song and Zipkin [17] study the continuous-time problem where the demand process is a Markov-modulated Poisson process (MMPP). In such a process, demands arrive as a Poisson process with demand rate changing according to a Markov chain. They show that the optimal ordering policy is of state-dependent (s, S) -type policy when there is a fixed ordering cost in addition to the linear cost (see also [18]). Treharne and Sox [19] analyze the same model when the current state of the environment is not observable. They give heuristics and present computational results. Without assuming a Markovian structure on the environment, Morton and Pentico [20] study the finite horizon, zero lead time, nonstationary problem, and derive upper and lower near-myopic bounds. They assume that demands in successive periods are independent but not necessarily identically distributed. They develop a number of heuristics and show that the best near-myopic heuristic is robust. Anupindi *et al.* [21] studying the stationary lead-time problem, obtain near-myopic bounds and give heuristics. They show that the average error increases with increase in variance of the lead-time distribution. Bollapragada and Rao [22] examine a single-product, nonstationary inventory problem with both supply and demand uncertainty, capacity limits on ordering quantities, and service-level requirements. A scenario-based stochastic program for the static, finite-horizon problem, is presented to determine replenishment orders over the horizon. A heuristic based on the first two moments of the random variables and a normal approximation is proposed.

Other than on Markov-modulated structures, there are a few studies in the area of correlated demands. In these studies, it is assumed that the demand is correlated between products as well as time periods. Although it is very difficult to obtain analytical solutions under correlated demands, it is worth studying correlation since it tells something about the new information which is observed every period. Ray [23] studies AR(1) and MA(1) demand models under stochastic lead times. By writing the moment-generating function of lead-time demand

in terms of demand variance–covariance matrix, and the lead-time distribution, he obtains the first two moments of the lead-time demand. These moments are then utilized in computing the reorder level that will provide a specified service level [24]. Johnson and Thompson [25] consider a single product, periodic-review system under stationary and nonstationary demand processes with no fixed cost. There are linear holding and shortage costs and the deliveries are instantaneous. They show that under some additional assumptions on the demand process, the optimal policies are myopic. Miller [26] introduces a multiplicative AR(1)-type uncertainty on the demand process and shows the optimality of myopic base-stock policies. Lovejoy [27], assuming a more general dependent demand structure, gives conditions under which the optimal replenishment policy is myopic. In production/inventory systems, a certain amount of information becomes available as time advances from one period to the next. It is important to incorporate this available information into the planning decisions of the system. This task can be achieved by means of forecast revisions that include correlation and evolve through time. Charnes *et al.* [28] consider a periodic-review inventory replenishment model with an order-up-to policy for the case of deterministic lead times and a covariance-stationary stochastic demand process. A method is derived for setting the inventory safety stock to achieve an exact desired stock-out probability when the autocovariance function for Gaussian demand is known. Graves [29] considers a special nonstationary demand structure, where the demand follows an integrated moving average (ARIMA (0, 1, 1)) process. For this demand process, the optimal order-up-to level is characterized and implications are discussed. Urban [30] analyzes the effect of autocorrelated demand to determine reorder levels. Specifically, first-order autoregressive and moving average (ARMA) demand processes are examined. Finally, considering nonstationary, correlated and evolving stochastic demands, Levi *et al.* [31] provide an approximation algorithm to obtain computationally efficient policies with constant worst-case performance guarantees.

Another group of studies that can be placed in this category are models with inventory and/or backorder-dependent demand structures. Inventory dependent demand is a special case of the general nonstationary demand model; within this category, the backorder-dependent demand structure has attracted some attention, as it is frequently observed in practice. Argon *et al.* [32] propose a single item, periodic-review model to investigate the effects of changes in the demand process that occur after shortage realizations. They investigate a system where the demands in successive periods are deterministic, but the level is affected by the backorder realizations. Urban [33] develops a periodic-review model with products that have autocorrelated demand, as well as demands dependent on the amount of inventory displayed to the customer. Urban [34] gives a comprehensive review of inventory models with inventory-level-dependent demand.

Single/Multi-Item Multiple Locations with Correlated Demand

In multilocation problems, demand interactions among the locations can be modeled using correlation structures. The work by Federgruen and Zipkin [35] is one of the initial studies that consider demand correlation across locations. They approach the periodic-review problem via dynamic programming to provide results in line with Clark and Scarf [36]. Erkip *et al.* [37] investigate the single-warehouse, n -retailer, multiechelon supply-chain model by Eppen and Schrage [38], and extend the results to the case where demand is correlated across locations as well as through time. The extension of the model to multiple items is possible, as long as the correlation structure is known. Closed-form expressions are obtained for safety stock values. Gilbert [39] presents a multistage supply-chain model that is based on the ARIMA time series models. Given an ARIMA model of consumer demand and the lead times at each stage, it is shown that the orders and inventories at each stage also follow ARIMA models, and closed-form expressions for these models are obtained. Graves and Willems [40] consider a single-product problem with a short life cycle. Given

a short product life cycle, product demand is increasingly difficult to forecast and is never really stationary because the demand rate evolves over the life of the product. The problem of placing strategic safety stocks with minimum cost is studied. They extend previous results for stationary demand to the case of nonstationary demand.

Another line of research within the multiechelon inventory systems that consider demand correlations across the items is the research on postponement. Several researchers study the benefits of product and process design that calls for delaying the differentiation of products; in other words, defer the stage after which the products assume their unique identities. Note that the end products have correlated demands, and hence aggregating them in manufacturing will help reduce the expected costs. Lee and Tang [41] develop a simple model that captures the costs and benefits associated with the use of product differentiation issues and focus on products having only one point of differentiation. Garg and Tang [42] extend these results to product families which have several points of differentiation. The analysis indicates that demand correlations in addition to some other factors play an important role in determining which point of differentiation should be delayed. Aviv and Federgruen [43] characterize the benefits of delayed differentiation and quick response programs when sales forecasts are updated over time. They consider more general settings, where parameters of the demand distributions fail to be known with accuracy or where consecutive demands are correlated.

Another stream of models that are of interest is models with advance demand information (ADI). Among many recent articles, we summarize a few. DeCroix and Mookerjee [44] study a periodic-review problem in which there is an option of purchasing advance demand information at the beginning of each period. Gallego and Özer [45] model ADI through a vector of future demands and show the optimality of a state-dependent order-up-to policy. Contrary to assuming that advance demand information is always known, Tan *et al.* [46] investigate a periodic-review inventory policy with

stochastic demand and stochastic advance demand information. In this model, each signal of advance demand can be treated as a potential demand. The evolution of the ADI is modeled as a Markov chain. Finally, Tan [47] considers the demand-forecasting problem in a business-to-business environment, where some customers provide information on their future orders, that are subject to change. The inaccurate information is used to specify a metric that influences forecasting.

EVOLVING FORECASTS THROUGH TIME

Introducing updating procedures for forecasting brings the Bayesian approaches into picture. There are a number of inventory problems that incorporate the Bayesian approach in decision making. There are two types of approaches: one assumes fully observable demand and the other, partially observable demand. Scarf [48], Azoury [49], Eppen and Iyer [50], Hill [51,52], among others, have assumed fully observable demand. For example, Lariviere and Porteus [53], as an example, considered Bayesian updating with partially observed sales. There are a number of studies that consider the notion of quick response, and hence apply Bayesian updates of forecasts to planning problems. As an example, Milner and Kouvelis [54] propose a single-period inventory modeling framework, with two ordering opportunities. The second order reacts to updated demand information and potentially capitalizes on supply-chain flexibility. The authors analyze the total inventory cost of a firm for alternate demand types: the standard assumption of independent demand over the period, fashion-driven innovative products through a Bayesian model, and innovative products with evolving demand through a Martingale process (see the section titled “Martingales” in this encyclopedia). The three demand processes exhibit very different behaviors with respect to the value of the alternate forms of flexibility.

Modeling the evolution of forecasting has been studied in the literature by several researchers. Graves *et al.* [55] and Heath and Jackson [56] suggest a general descriptive

model that accommodates simultaneous evolution of demand for many time periods and captures correlations between products and time periods, which is named as the *Martingale model of forecast evolution (MMFE)*. They argue that the MMFE approach is preferred to a direct time-series approach since it can capture the potential impact of expert judgment other than past data information. Below, we give a brief description of the model.

Consider a multi-item environment. Let D_n^j denote a forecast vector for item j in period n . We assume that there are J items and the forecasts are available for M periods. Hence,

$$D_n^j = (d_{n,n}^j, d_{n,n+1}^j, \dots, d_{n,n+M}^j, \mu^j, \mu^j, \dots),$$

where $d_{n,n}^j$ is the demand realization for item j in period n , $d_{n,n+1}^j$ is the demand forecast for period $(n+1)$ made at period n . It is assumed that the periods beyond the forecast horizon have a demand forecast of μ^j (stationary).

The evolution of demand forecasts from one period to the next can be modeled according to either an additive evolution model or a multiplicative one. Here we present the MMFE method through the additive model. Accordingly, the new forecast for the $(n+1)$ th period demand can be written in terms of the previous forecast and an error term as

$$\begin{aligned} d_{n+1,n+1}^j &= d_{n,n+1}^j + \epsilon_{n+1,1}^j \\ d_{n+1,n+2}^j &= d_{n,n+2}^j + \epsilon_{n+1,2}^j \\ &\vdots \\ d_{n+1,n+M+1}^j &= \mu^j + \epsilon_{n+1,M+1}^j, \end{aligned}$$

where for $n \geq 1$, $\bar{\epsilon}_n^j = (\epsilon_{n,1}^j, \epsilon_{n,2}^j, \dots, \epsilon_{n,M+1}^j)$ is the vector that represents the evolution of demands and forecasts from period $n-1$ to n . Note that the first component $\epsilon_{n,1}^j$ is the one-step ahead forecast error. The $\bar{\epsilon}_n^j$ vectors are assumed to be independent, identically distributed, multivariate normal random vectors with mean 0. The demand forecast update equation states that as we get closer to a demand period, say period

$(n+1)$, by updating the forecasts for this period successively, we reduce the variability of forecast errors. These assumptions also imply that the i -th step-ahead demand forecast for any j , $d_{n,n+i}^j$, is a martingale, and it is the conditional expectation of the $(n+i)$ -th period demand, given the history up to time n [55,56]. The definition of $\bar{\epsilon}_n^j$ enables us to allow for correlation among future forecasts of an item, as well as cross-correlations among forecasts of different items for a given period, n .

The above structure holds under the assumptions stated by Heath and Jackson [56]. The methodology suggested by MMFE is useful, as it also incorporates consistent knowledge on the forecasts. Many extensions or analyses regarding MMFE are available. In the remaining part of this section we summarize some of these studies, which also have a relation to or application in the supply chain area.

Güllü [57] considers a single-item production/inventory system that incorporates forecasts into planning decisions. He uses an additive forecast evolution model, which is a special case of MMFE and is able to capture the structure of dependent and nonstationary demand. He assumes that the production capacity is limited and unsatisfied demand is backlogged. As a benchmark, he considers a standard inventory model, which assumes the same distributional properties of demand. The suggested model yields lower expected costs and inventory levels when compared to a standard inventory model. In a similar study, Güllü [58] investigates a two-echelon allocation model consisting of a depot and several retailers, again integrating forecast revisions. The depot does not hold any inventory and unsatisfied demand is backlogged. He compares this system to a standard allocation model. The standard model results in higher order-up-to levels and higher system costs. Toktay and Wein [59] study a capacitated single-item make-to-stock environment facing a stationary, correlated demand process. They utilize the MMFE approach to generate forecast revisions. The forecast updates are then incorporated into the system to obtain approximate results for the base-stock level

that yields minimum expected inventory holding and backorder costs. Iida and Zipkin [60] propose a dynamic inventory model with the MMFE model. As a special case, it includes the ARMA demand model, represented by MMFE. They combine the MMFE with a linear programming model of production/distribution system in order to find an economical safety factor level and quantify the effects of forecast error on the total system cost. Lu *et al.* [61], also considering the MMFE model of demand forecast evolution in a periodic-review inventory model, develop bounds on the optimal base-stock levels as was done in Iida and Zipkin [60]. In addition, they provide bounds on the cost error of any heuristic with respect to the optimal policy, which is computationally intractable to evaluate. They give the necessary and sufficient conditions for the optimality of myopic policies.

Dong and Lee [62] revisit the serial multi-echelon inventory system of Clark and show that the structure of the optimal stocking policy of Clark and Scarf [36] holds under time-correlated demand processes using an MMFE. Then, the authors extend the approximation to an autoregressive demand model.

Aviv [63] considers a supply chain in which the underlying demand process, possibly evolves according to a vector autoregressive time series. The author proposes an adaptive inventory replenishment policy that utilizes the Kalman filter technique (see **Forecasting: State-Space Models and Kalman Filter Estimation**).

FORECASTING PROCEDURES FOR CORRELATED DEMAND AND APPLICATIONS

A typical forecasting process involves analyzing historical data. Historical data may exhibit trends, seasonal variations, as well as correlations. Other factors may come into the picture such as impact of other products or services in the market on the demand of the product being forecasted. Typically, the performance of the forecasting process is evaluated on the basis of the relative error, that is the percentage difference of the actual from its forecasted value, or the mean-squared forecast error.

Time series analysis techniques are utilized to carry out the forecasting process. Demand for each product is forecasted over a planning horizon in such a way that the mean-squared forecast error is minimized. Standard time series analysis techniques consider correlation structures over time, but are not effective when there are multiple items with correlated demand.

One way of avoiding forecasting demands of multiple items with cross-correlation is to aggregate the items. On the other hand, there is a need to forecast at the item level, as well. The following is the question here: Is it better to forecast the aggregate (and hence utilize a top-down approach) and then disaggregate, or to forecast the individual items directly (and hence utilize a bottom-up approach), and then the total? In any specific application, it will be hard to say which one would be better, unless ample data is available to support both approaches. Of course, the general understanding and approach are to utilize, as much as possible, the possibilities brought by the available data. Note that there are examples of both; consider Zhou *et al.* [64], where a top-down approach is utilized, and Chen *et al.* [65], which presents different (but not independent) representations for a two-item forecasting problem—an example of the bottom-up strategy.

In case we only need to consider autocorrelation, demand data can be modeled either as a stationary ARMA process or a nonstationary ARIMA process with other controllable and uncontrollable inputs [4–8]. Stationary models keep their distributional behavior in equilibrium around a constant mean. On the other hand, the nonstationary, ARIMA, models have no constant mean level over time. Nonstationary data quite often arise in industrial forecasting. Demand processes could also be modified to incorporate the effects of external factors such as economic, political, and geographic conditions. For multi-item systems with cross-correlation, multivariate time series models [66] such as a vector ARMA(p,q) could be utilized.

Aviv [63], considering a two-stage supply chain, employs state-space methods for modeling demand process as vector AR(1)(VAR(1)) time series. The minimum mean-squared

error (MMSE) estimates of the demand can then be obtained recursively via the Kalman filter (see *Forecasting: State-Space Models and Kalman Filter Estimation*) as new observations are included over time. These observations may not include full demand information. The adaptive replenishments follow a base-stock policy that depends on the aggregate demand estimate during lead time.

In order to forecast intermittent demand, Croston [67] presents two separate exponential smoothing forecasts, one on the demand size and the other on the demand interarrival times. Several studies demonstrate that Croston's method performs better than some competing approaches [68,69]. Shenshstone and Hyndman [70] attempt to identify stochastic models that underlie Croston's method in order to obtain confidence intervals for the forecasts. One such model assumes a nonstationary ARIMA(0,1,1) time series model for both the demand size and the demand interarrival times.

There are a few studies that combine forecasting performance with inventory performance. The difficulty lies in creating benchmarks, and we believe that more work in this area is needed.

One example is by Sani and Kingsman [69] who compare various inventory replenishment heuristics of (s,S) type in addition to the comparison of forecasting methods for a low demand item.

Chen *et al.* [65] report another study that combines forecasting and inventory control. The authors use a bivariate VAR(1) time-series model to investigate the effects of aggregating two interrelated demands. The paper further explores the properties of the aggregated time series and provides guidelines for practitioners to determine proper aggregation and forecasting approaches. A method is proposed to estimate the parameters of the demand model.

Other examples are from the production planning environments where forecast procedures are integrated with the planning procedures. Heath and Jackson [56] combine the MMFE with a linear programming model of production/distribution system in order to

find an economic safety stock level and quantify the effects of forecast error on the total system cost. A similar study is carried out by Kayhan *et al.* [71], where the application of MMFE to an industrial case is considered. Procedures to set safety stocks for an LP (linear programming)-based production planning model are proposed utilizing forecast error correlations across items and time modeled by MMFE.

CONCLUSIONS

In this article, we consider single-/multi-item supply chains under stationary or nonstationary correlated demand. We review inventory approaches in order to establish the demand-forecasting requirements. We also discuss the notion of evolving forecasts, including MMFE. We summarize various approaches for forecasting the demand of such inventory systems.

Forecasting is a very important tool for supply chains. Customer behavior and retail performance are factors affecting the demand for products. More comprehensive approaches for forecasting are needed that will consider all aspects of demand. On the other hand, we observe that there are still challenges ahead in achieving well-performing plans in a dynamically changing, uncertain environment, for more classical forecasting.

REFERENCES

1. Fisher M, Raman A. Survey results reported at the Harvard/Wharton Merchandising Effectiveness Project; 1998.
2. White JB. Hughes charge give GM net loss of 357 million. Wall Street Journal. 1992 August 7.
3. Stewart TA. Brace for Japan's hot new strategy. Fortune. 1992 Sep 21.
4. Box GEP, Jenkins GM, C G. Reinsel. Time series analysis: forecasting and control. 3rd ed. Englewood Cliffs (NJ): Prentice Hall; 1994.
5. Brockwell PJ, Davis RA. Time series: theory and methods. 2nd ed. New Jersey: Springer; 1991.

6. Brockwell PJ, Davis RA. Introduction to time series and forecasting. 2nd ed. New Jersey: Springer; 2002.
7. Montgomery DC, Jennings CL, Kulahci M. Introduction to time series analysis and forecasting. New York: John Wiley & Sons, Inc.; 2008.
8. Makridakis SG, Wheelwright SC, Hyndman RJ. Forecasting: methods and applications. 3rd ed. New York: John Wiley & Sons, Inc.; 1998.
9. Blinder A. Inventories and sticky prices: more on the microfoundations of macroeconomics. *Am Econ Rev* 1982;72(3):334–348.
10. Kahn JA. Inventories and the volatility of production. *Am Econ Rev* 1987;77(4):667–679.
11. Altiok T, Melamed B. The case for modeling correlation in manufacturing systems. *IIE Trans* 2001;33:779–791.
12. Arrow K, Harris T, Marschak J. Optimal inventory policy. *Econometrica* 1951;19:250–272.
13. Scarf H. The optimality of (s, S) policies in the dynamic inventory problem. In: Arrow J, Karlin S, Suppes P, editors. *Mathematical methods in the social sciences*. Stanford (CA): Stanford University Press; 1960: pp. 196–202.
14. Veinott AF. Optimality policy for multi-product, dynamic, nonstationary inventory problem. *Manage Sci* 1965;12:206–222.
15. Sobel MJ. Myopic solutions of Markov decision processes and stochastic games. *Oper Res* 1981;29:995–1009.
16. Karlin S, Fabens A. The (s, S) inventory model under Markovian demand process. In: Arrow J, Karlin S, Suppes P, editors. *Mathematical methods in the social sciences*. Stanford (CA): Stanford University Press; 1960.
17. Song J-S, Zipkin P. Inventory control in a fluctuating demand environment. *Oper Res* 1993;41:351–370.
18. Sethi SP, Cheng F. Optimality of (s, S) policies in inventory models with Markovian demand. *Manage Sci* 1997;45:931–939.
19. Treharne JT, Sox CR. Adaptive inventory control for nonstationary demand and partial information. *Manage Sci* 2002;48:607–624.
20. Morton TE, Pentico D. The finite horizon nonstationary stochastic inventory problem: near-myopic bounds, heuristics, testing. *Manage Sci* 1995;41:334–343.
21. Anupindi R, Morton TE, Pentico D. The nonstationary stochastic lead-time inventory problem: Near-myopic bounds, heuristics, and testing. *Manage Sci* 1996;42:124–129.
22. Bollapragada R, Rao US. Replenishment planning in discrete-time, capacitated, nonstationary, stochastic inventory systems. *IIE Trans* 2006;38(7):583–595.
23. Ray WD. The significance of correlated demands and variable lead times for stock control policies. *J Oper Res Soc* 1980;31(2): 187–190.
24. Ray WD. Computation of reorder levels when the demands are correlated and the lead time random. *J Oper Res Soc* 1981;32(1): 27–34.
25. Johnson GD, Thompson HE. Optimality of myopic inventory policies for certain dependent demand processes. *Manage Sci* 1975;21:1303–1307.
26. Miller BL. Scarf's state reduction method, flexibility, and a dependent demand inventory model. *Oper Res* 1986;34(1):83–90.
27. Lovejoy WS. Myopic policies for some inventory models with uncertain demand distributions. *Manage Sci* 1990;36(6):724–738.
28. Charnes JM, Marmorstein H, Zinn W. Safety stock determination with serially correlated demand in a periodic-review inventory system. *J Oper Res Soc* 1995;46(8):1006–1013.
29. Graves SC. A single-item inventory model for a non-stationary demand process. *Manuf Serv Oper Manage* 1999;1(1):50–61.
30. Urban TL. Reorder level determination with serially-correlated demand. *J Oper Res Soc* 2000;51(6):762–768.
31. Levi R, Pál M, Roundy RO, *et al.* Approximation algorithms for stochastic inventory control models. *Math Oper Res* 2007;32(2): 284–302.
32. Argon NT, Güllü R, Erkip N. Analysis of an inventory system under backorder correlated deterministic demand and geometric supply process. *Int J Prod Econ* 2001;71(1-3): 247–254.
33. Urban TL. A periodic-review model with serially-correlated, inventory-level-dependent demand. *Int J Prod Econ* 2005;95(3):287–295.
34. Urban TL. Inventory models with inventory-level-dependent demand: a comprehensive review and unifying theory. *Eur J Oper Res* 2005;162(3):792–804.
35. Federgruen A, Zipkin P. Computational issues in an infinite-horizon, multi-echelon inventory model. *Oper Res* 1984;32(4):818–836.

36. Clark A, Scarf H. Optimal policies for a multi-echelon inventory problem. *Manage Sci* 1960;6:475–490.
37. Erkip NK, Hausman WH, Nahmias S. Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands. *Manage Sci* 1990;36(3):381–392.
38. Eppen G, Schrage L. Centralized ordering policies in a multi-warehouse system with leadtimes and random demand. In: Schwartz L, editor. *Multi-level Production/Inventory control systems: theory and practice*. Amsterdam: North-Holland; 1981. pp. 51–69.
39. Gilbert K. An ARIMA supply chain model. *Manage Sci* 2005;51(2):305–310.
40. Graves SC, Willems SP. Strategic inventory placement in supply chains: Non-stationary demand. *Manuf Serv Oper Manage* 2008;10(2):278–287.
41. Lee HL, Tang CS. Modeling the costs and benefits of delayed product differentiation. *Manage Sci* 1997;53(1):40–53.
42. Garg A, Tang CS. On postponement strategies for product families with multiple points of differentiation. *IIE Trans* 1997;29(8):641–650.
43. Aviv Y, Federgruen A. Design for postponement: a comprehensive characterization of its benefits under unknown demand distributions. *Oper Res* 2001;49(4):578–598.
44. DeCroix GA, Mookerjee VS. Purchasing demand information in a stochastic-demand inventory. *Eur J Oper Res* 1997;102:36–57.
45. Gallego G, Özer O. Integrating replenishment decisions with advance demand information. *Manage Sci* 2001;47:1344–1360.
46. Tan TR, Güllü R, Erkip N. Modelling imperfect advance demand information and analysis of optimal inventory policies. *Eur J Oper Res* 2007;177:897–923.
47. Tan TR. Using imperfect advance demand information in forecasting. *IMA J Manage Math* 2008;19:163–173.
48. Scarf H. Some remarks on Bayes solutions to the inventory problem. *Nav Res Log Q* 1960;7:591–596.
49. Azoury K. Bayes solution to dynamic inventory models under unknown demand distribution. *Manage Sci* 1985;31:1150–1160.
50. Eppen GD, Iyer AV. Improved fashion buying with bayesian updates. *Oper Res* 1997;45:805–819.
51. Hill RM. Applying bayesian methodology with a uniform prior to the single period inventory model. *Eur J Oper Res* 1997;98:555–562.
52. Hill RM. Bayesian decision-making in inventory modeling. *IMA J Math Appl Bus Ind* 1999;10:147–163.
53. Lariviere MA, Porteus LE. Stalking information: Bayesian inventory management with unobserved lost sales. *Manage Sci* 1999;45:346–363.
54. Milner JM, Kouvelis P. Order quantity and timing flexibility in supply chains: The role of demand characteristics. *Manage Sci* 2005;51(6):970–985.
55. Graves SC, Meal HC, Dasu S, *et al.* Two-stage production planning in a dynamic environment. In: Schneeweiss C, Axsater S, Silver E, editors. *Volume 266, Multi-stage production planning and inventory control, Lecture notes in economics and mathematical systems*. Berlin: Springer; 1986. pp. 9–43.
56. Heath DC, Jackson PL. Modeling the evolution of demand forecasts with application to the to safety stock analysis in production/distribution systems. *IEE Trans* 1994;26(3):17–30.
57. Güllü R. On the value of information in dynamic production/inventory problems under forecast evolution. *Nav Res Log* 1996;43:289–303.
58. Güllü R. A two-echelon allocation model and the value of information under correlated forecasts and demands. *Eur J Oper Res* 1997;99:386–400.
59. Toktay LB, Wein LM. Analysis of forecasting-production-inventory system with stationary demand. *Manage Sci* 2001;47(9):1268–1281.
60. Iida T, Zipkin P. Approximate solutions of a dynamic forecast-inventory model. *Manuf Serv Oper Manage* 2006;8(4):407–425.
61. Lu X, Song J-S, Regan A. Inventory planning with forecast updates: Approximate solution and cost error bounds. *Oper Res* 2006;54(6):1079–1097.
62. Dong LX, Lee HL. Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Oper Res* 2003;51(6):969–980.
63. Aviv Y. A time-series framework for supply-chain inventory management. *Oper Res* 2003;51(2):210–227.
64. Zhou S, Jackson PL, Roundy RO, *et al.* The evolution of family level sales forecasts into product level forecasts: Modeling and estimation. *IIE Trans* 2007;39(9):831–843.
65. Chen A, Hsu CH, Blue J. Demand planning approaches to aggregating and forecasting interrelated demands for safety stock and

- backup capacity planning. *Int J Prod Res* 2007;45(10):2269–2294.
66. Reinsel GC. *Elements of multivariate time series analysis*. New York: Springer; 1993.
67. Croston JD. Forecasting and stock control for intermittent demands. *Oper Res Q* 1972; 23(3):289–303.
68. Willemain TR, Smart CN, Shocker JH, *et al.* Forecasting intermittent demand in manufacturing: A comparative evaluation of Croston's method. *Int J Forecast* 1994;10:529–538.
69. Sani B, Kingsman BG. Selecting the best periodic inventory control and demand forecasting methods for low demand items. *J Oper Res Soc* 1997;48(7):700–713.
70. Shenstone L, Hyndman RJ. Stochastic models underlying Croston's method for intermittent demand forecasting. *J Forecast* 2005; 24:389–402.
71. Kayhan M, Güllü R, Erkip N. Integrating forecast evolution with production/inventory planning: an implementation framework for a fast moving consumer goods manufacturer. Working Paper, Ankara: Middle East Technical University; 2005.

FORECASTING NONSTATIONARY PROCESSES

JUI-HONG CHIEN

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

PETROS XANTHOPOULOS

PANOS M. PARDALOS

VERA TOMAINO

Department of Industrial and Systems
Engineering and Biomedical
Engineering, University of Florida,
Center for Applied Optimization,
Gainesville, Florida

INTRODUCTION

Most of our daily life phenomena can be recorded as time series, such as wind speed, heart rate, and stock exchange price. Sometimes, it is convenient to assume that there are exact mathematical formulas describing the dynamics of the observable stochastic process. Understanding the underlying mechanism of these processes is very crucial in engineering and science. Although it is impossible to fully understand all the details of processes relating to the outcomes of a process, it is intuitively reasonable to assume that there are rules behind observable measurements and, once we understand those, we can forecast how the time series may develop to some extent. However, real-life phenomena are rarely simple and usually involve many parameters and variables that can influence the results with different time lags with different degrees. Still, with careful control of the recording environment, it is possible to reduce unpredictable factors and forecast a process within a margin of confidence. To understand all those principles and rules controlling these black-boxed processes simultaneously is a really tough task. Therefore, we would like to study simple nonstationary paradigms before moving

to the more complicated time series. As mentioned, a straightforward assumption is that a process possesses certain fixed properties in itself such that it evolves with some principle. One of the most convenient properties in a process is stationarity. Stationarity in the strict sense means that the statistics of a stochastic process does not change with time, meaning

$$f_{x_1, x_2, \dots, x_n}(t_1, t_2, \dots, t_n) = f_{x_1, x_2, \dots, x_n} \\ \times (t_1 + \tau, t_2 + \tau, \dots, t_n + \tau), \quad (1)$$

where f denotes probability mass function and τ denotes a time lag and of any size. When a process fails to fulfill Equation (1), we claim that the process is not stationary in the strict sense. Stationary processes possess relatively stable statistical properties such as fixed mean and variance compared to nonstationary ones. Nevertheless, fixed mean and variance by themselves do not imply stationarity in the strict sense. Besides, with those fixed statistics we can begin predicting the foregoing process with more confidence and less complexity. The first section of this article focuses on some necessary background for methods that will be introduced later. In the second section of this article, we will present two well-known practical methods to forecast nonstationary processes.

BACKGROUND KNOWLEDGE

Stationarity

Stationarity in the strict sense is almost impossible in real life. Even a parametric simulated process can only theoretically have stationarity, not to mention a time series recorded from real-world signals which have much more complex dynamics. One main reason why we bring up stationarity is that no model can include all explanatory variables affecting on the process, not to mention some factors that are beyond one's control. Although we can hardly have a stationary

process in the strict sense, we want to take advantages of those convenient properties of stationarity when studying nonstationary time series. Fortunately, one can still claim that a process has stationarity in a wide sense if it has not only a fixed mean and variance but also time-invariant autocovariance. Still, even when a process has stationarity, it does not mean that one can fully describe the process with only deterministic terms. To better describe effectors that are not controllable in a process, we should model a process with one or several deterministic and random terms. Actually, it is the random terms that define whether the process is stochastic or not. To make a further step toward real-life processes, we do not deal with stochastic processes but nonstationary stochastic processes. One way to dissect a nonstationary process is to use the concept of piecewise stationarity. Even though the whole process is nonstationary, we can still treat each segment of a reasonable length as if they were stationary, after splitting it into segments. Based on piecewise stationary concept in time domain, one can make an analogue to time and frequency domain and that is the idea of local stationary wavelet analysis that will be introduced in nonparametric method section of this article.

Local Stationarity. One of the crucial parameters for using the piecewise stationary concept is how to choose the length of a segment that is neither too short for calculating statistical estimates nor too long so that the process is no longer stationary [1]. It should be noted that the criterion for selecting the segment length should always be dependent on the method that we are using in order to model the nonstationary process. For example, for a quasi-stationary process, the underlying assumption is that the parametric properties might change their states between different time segments. As a result, an ideal segment should not be too long to overlap segments having different parametric values. Compared to quasi-stationary processes, parameters of a time-varying autoregressive model do not suddenly change but gradually evolve with time. Proper selection of a method to model

a process depends on some basic properties of the raw data. For instance, a seismic wave recording or an electroencephalograph recording containing a seizure shows sudden state changes. As a result, it is reasonable to use a quasi-stationary method for modeling. For some other processes such as the population of some species or the price of a stock, one might only be interested in their smooth evolution property instead of their sudden change due to special or dramatic events, so it would be reasonable to use the time-varying autoregressive model, which will be introduced later on in this article. For more details of segmentation implementation, readers are referred to the work of Appel and Brandt [1], and Adak [2]. Those methods of modeling a nonstationary process are called *parametric methods*. In addition, one can use nonparametric methods that do not assume the existence of a model or distribution family in a process. Compared to parametric methods, nonparametric methods still involve choosing the parameter to some extent but not as much as in parametric methods. If one intends to observe a nonstationary process as a piecewise, quasi, or evolutionary stationary process, a very well-known and widely used nonparametric tool called *time-dependent spectrum* in the time–frequency analysis field can be used. For a more detailed introduction of the classes of nonstationary process, readers are referred to the work of Adak [2].

Parametric Method for Analyzing Nonstationary Process

The benefits of using a parametric method are its simplicity and lucid structure. The main feature of parametric methods is the use of parameters as structural elements to introduce interpretation and interaction so that one can reduce the ambiguity within the data itself. This, however, might be achieved at the risk of misinterpretation [3]. Some other parametric approaches are based on specific assumptions that process outputs belong to the family or families of the distributions used in the model or can be described as the output of a stochastic process [4], whereas nonparametric approaches do not impose specific constraints with regard to

the distribution family. One has to first carefully scrutinize the statistical properties of the observed data and the necessary assumption for the parametric method being applied for the estimation [5]. In sum, parametric methods are beneficial only if there is a strong reason to look at the process in a particular structural way or as a member of a certain distribution family. As one can imagine, a parametric approach might not perform well on a process having a huge future random term which does not constantly fall into any distribution family and a weak link between current and future states of the process. The time-varying autoregressive (AR) model [6], just like a stationary process, can use a parametric or nonparametric method to start the analysis on a nonstationary process. One of the most commonly used model for simulating or fitting nonstationary time series is the time-varying AR model. A p th order linear time-varying AR model can be written in a general form as

$$x_k = \sum_{i=1}^p \phi_{i,k} x_{k-i} + \varepsilon_k, \quad (2)$$

where $\phi_{i,k}$ are time-varying parameters that differentiate a time-varying AR model from an AR model that has fixed coefficient ϕ_i instead of $\phi_{i,k}$. The main difficulty of using a parametric model for analysis lies in choosing proper values for parameters. As Equation (2) shows, p and $\phi_{i,k}$ are parameters that need to be chosen on the basis of the training data, while ε_k is a zero-mean white-noise process. One advantage of a parametric method is that it renders a lucid dynamic relationship between data points and gives us some sense of how data evolve over time. However, the model may not be unique because of the uncertainty of parameters. We should bear in mind that we can fit models of different orders into the same data sets. In this case, the time-varying AR model parameter values ($\phi_{i,k}$) might be significantly different especially for a complex data set. Several ways have been proposed to find the optimal order of an AR model, such as the Akaike information criterion (AIC) [7], Bayesian information criterion (BIC) [8], and the likelihood ratio test [9,10]. Once the order of

the AR model is decided, ϕ_i are calculated from Yule–Walker (YW) equations. There exist other methods for estimating regression coefficients such as Burg’s method (Burg), the least-squares (LS) approach, modified covariance method, covariance method, parametric spectral estimation method, Prony’s method, and so on. However, choosing the order of a time-varying AR model is more complex than choosing an AR model itself. A recently proposed method for the time-varying AR model order uncertainty is the generalized likelihood ratio test [11].

Autoregressive Integrated Moving Average (ARIMA) Model. In this case, we deal with nonstationary processes whose derivative, of proper order, is a stationary process. Compared to an autoregressive moving average (ARMA) model, which can be represented by the bilinear equation

$$x_n = - \sum_{k=1}^p a_k x_{n-k} + \sum_{k=0}^q b_k u_{n-k}, \quad p, q \in \mathbb{Z}, \quad (3)$$

where $\{u_n\}$ is a Gaussian white noise process, an autoregressive integrated moving average model (ARIMA) can exhibit not only ARMA but also nonstationarity having neither a fixed mean nor a fixed variance. A very important property of ARIMA is that, if one takes the d th derivative of the process, it will result in a stationary process that can be represented by another ARMA model $\{x_n\}$ [12]. This relationship can be represented by the following equation:

$$\Delta^d z_n \triangleq x_n = - \sum_{k=1}^p a_k x_{n-k} + \sum_{k=0}^q b_k u_{n-k}, \quad p, q, d \in \mathbb{Z}. \quad (4)$$

This also gives us a powerful linkage for changing a nonstationary process into a stationary one which can be more easily analyzed. However, one should notice that the high-frequency components of the process would be amplified after differentiating [13]. By considering an ARMA (p, q), with $q \leq p$, it is possible that the AR part has more terms than the moving average (MA) part. The

realization of an ARMA state space model is given as:

$$\tilde{x}_{k+1} = \begin{bmatrix} -a_{1,k} & 1 & 0 & 0 & \cdots & 0 \\ -a_{2,k} & 0 & 1 & 0 & \cdots & 0 \\ -a_{3,k} & 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 \\ -a_{p-1,k} & 0 & 0 & \cdots & 0 & 1 \\ -a_{p,k} & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tilde{x}_k + \begin{bmatrix} b_{1,k} - a_{1,k} \\ b_{2,k} - a_{2,k} \\ \vdots \end{bmatrix} \tilde{u}_k, \quad (5)$$

$$z_k = [1 \ 0 \ \cdots \ 0] \tilde{x}_k + u_k, \quad (6)$$

where $\tilde{x}_k = \{x_k, x_{k-1}, \dots, x_{k-p+1}\}^T$ and $\{\tilde{u}_k\}, \{u_k\}$ are both Gaussian white-noise processes. The tilde indicates a vector. If we denote the transition matrix as F_k , the noise coupling matrix as G_k , and the observation matrix as $\tilde{h} = [1, 0, \dots, 0]$, Equations (5) and (6) and Equations (3) and (4) become

$$\tilde{x}_{k+1} = F_k \tilde{x}_k + G_k \tilde{u}_k, \quad (7)$$

$$z_k = \tilde{h} \tilde{x}_k + u_k. \quad (8)$$

We can make a comparison between ARIMA and time-varying AR which has the following dynamic matrix form:

$$\tilde{x}_{k+1} = F_k \tilde{x}_k + \varepsilon_k, \quad (9)$$

$$z_k = \tilde{h} \tilde{x}_k, \quad (10)$$

where ε_k is a zero-mean innovation with time-varying variance and

$$F_k = \begin{bmatrix} \phi_{k,1} & \phi_{k,2} & \cdots & \phi_{k,p-1} & \phi_{k,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \quad (11)$$

Bayesian Autoregressive Model. Another widely used parametric model for time series is the Bayesian autoregressive model. There are a number of Bayesian forecasting

techniques that are available in the literature. A Bayesian multivariate regression was used to analyze the vector autoregressive model. The model is provided by the equation

$$x_t = \tilde{v} + A_1 y_{t-1} + \cdots + A_p y_{t-p} + \tilde{u}_t, \quad (12)$$

where $\tilde{x}_t$ is a column vector and A_i are the autoregressive coefficients, $\tilde{v}$ is the vector of intercepts, and u_t is the Gaussian white noise. Now, the above can be rewritten as the Bayesian multivariate regression model. A detailed analysis of the model is provided in the refs 14 and 15. A natural extension of the multivariate regression model would be the multivariate dynamic model in the previous section. This provides us with possibility of discounting, which is a process where more weights can be added to the most recent information and external information can also be considered. Barnett *et al.* [14] provide a Bayesian modeling of autoregressive time series through parameter estimation, outlier identification, and interpolation of missing information. They provide a Markov-chain Monte Carlo sampling technique to simultaneously handle this and demonstrate it through a real-life example. A Bayesian approach to nonstationary process has also been studied [15].

In this procedure, the data is divided into blocks and an autoregressive model is fitted to each block. The smoothness parameter that is controlled by the autoregressive coefficients of successive blocks is determined by AIC or BIC. McCullough and Tsay extend the Bayesian analysis for autoregressive time series in order to predict the time point at which the variance and mean shifts occur [16]. They use a Gibbs sampler to make random draws from conditional distributions of the variables involved, which are related to the probability of shift at a given time point. A number of surveys and expository articles that deal with Bayesian forecasting techniques are available in the literature for an interested reader [17]. Besides parametric and nonparametric methods, there are semiparametric models [18] combining both parametric distributions and nonparametric

models. There have been works using semi-parametric models to analyze nonstationary processes. Interested readers can refer to Velasco [19] and Damian *et al.* [20] for details. One should notice that, since semi-parametric methods contain parametric components, they make the assumptions that are made in the parametric methods and might suffer from the same disadvantages as the parametric methods.

Time–Frequency Analysis: Wavelet Transforms

A useful tool for time–frequency analysis is called *wavelets*. Most of the stationary processes are generated in the time domain and also recorded in time domain. However, a weakly stationary process can be characterized by its frequency components. That is why Fourier transform is so widely used in almost every field of science and engineering. The Fourier transform represents a process formula as a superposition of pure frequencies, as follows:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt, \quad (13a)$$

where i denotes an imaginary unit.

Although the representation through Fourier transform can clearly show those weighted components of a process in terms of frequencies, it does not include any time component. One can hardly perceive when those frequency components participate in a process. Especially, for real data such as seismic waves or biomedical signals, a functional transform showing results in both the time and frequency domains can be much more informative and intuitive. A very widely used time–frequency analysis is the wavelet transform, which renders a perspective with both time and frequency resolutions by imposing a mask function before the original signal as

$$X(t, f) = \int_{-\infty}^{\infty} w(t - \tau)x(\tau)e^{-i2\pi f\tau} d\tau, \quad (13b)$$

where $w(t - \tau)$ is the mother wavelet function. When $w(t - \tau)$ is a rectangular window, the above equation is also called *short-time Fourier transform*, *windowed Fourier transform*, or *time-dependent Fourier transform*.

One may see the potentiality of wavelet from Equation (13b) that it has a t parameter that can characterize the signal at a certain range in time domain. As we can imagine, a real-life signal such as seismic waves that shows the characteristics of a short impulsive feature can hardly be described merely through Fourier transform decomposing the seismic wave into a combination of sinusoidal functions covering from negative infinity to infinity [21]. Because of the dual property of a wavelet both in the time and frequency domains, one can not only use a wavelet for time–frequency analysis but also for denoising a process that has noise with a certain unique frequency band and that happen sporadically with time [22,23].

Local Stationary Wavelets. Assuming that a process $\{X_t\}, t \in \mathbb{Z}$ possesses stationarity, we can estimate second-order statistics such as variance or auto-covariance through only a segment of observation and have confidence in the estimated statistics because its stationarity has rendered characteristic quantitative homogeneity. As we have longer term observation, we can estimate statistics better through prolonging our observational period. In contrast, for a nonstationary process, if we assume that the variance of a process changes slowly over time, saying that variance is a continuous and differentiable function of time, the variance can still be estimated with certain precision by gathering information around the time point of interested variance. This demand exactly fits the property of wavelet transform, which preserves local characteristics of a process to some extent on both time and frequency domains. A statistically slowly changing process can be called a *local stationary process* [24]. For a covariance-stationary process X_t , there is a Cramér representation

$$X_t = \int_{(-\pi, \pi]} A(\omega) \exp(i\omega t) dZ(\omega), \quad t \in \mathbb{Z}, \quad (14)$$

where $Z(\omega)$ is a stochastic process with orthonormal increments [25]. Nonstationary processes can be written in almost the same way as Equation (14) but $A(\omega)$ is replaced by a function of time. Along with

the same rationale, a nonstationary process can be represented using local stationary wavelet and nondecimated discrete wavelets $\psi_{j,k}(t)$, $j = -1, -2, \dots, k \in \mathbb{Z}$, where j indicates the scale and k indicates time location. To express a local stationary process, one can write it as

$$X_t = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} w_{j,k} \psi_{j,k}(t) \xi_{j,k}, \quad (15)$$

where $\xi_{j,k}$ is mutually orthogonal zero-mean random innovation. From Equations (14) and (15), one should notice that $w_{j,k}$ in Equation (14) corresponds to $A(\omega)$ in Equation (14) and both of them indicate the amplitude of each analysis component. One should also notice that $\psi_{j,k}(t)$ is a replacement of the Fourier harmonics $\exp(i\omega t)$, in Equation (14). The subscript indices j and k in $w_{j,k}$ represent scale and time location, respectively, just like those in $\psi_{j,k}(t)$. The reason for using nondecimated discrete wavelet is that it can be shifted to any time point and not only confined to shifts of 2^{-j} compared to discrete wavelet. Note that nondecimated wavelets do not have orthogonalities and are an over-complete collection of shifted vectors [24]. An example of discrete nondecimated wavelet is the Haar wavelet. Nason *et al.* [24] also proposed the evolutionary wavelet spectrum (EWS) to quantify how the size of $w_{j,k}$ changes over time and defined it as

$$S_j(z) = S_j(k/T) \approx w_{j,k}^2. \quad (16)$$

Note that in EWS, $S_j(z)$ still has resolution in both frequency and time domains but with a different timescale which rescales the whole observation time into $z \in (0, 1)$: saying that the rescaled time z is calculated by dividing k by the whole observation time length T , and $k = 0, 1, \dots, T-1$. Our goal is to estimate the variance of a local stationary process. To accomplish that, one way is to put together the localized information around the time point of interest, with the concept of local stationary wavelet and EWS. Let $c(z, \tau)$ represent localized autocovariance and be defined as:

$$\begin{aligned} \lim_{T \rightarrow \infty} \text{cov}(X_{[zT]-\tau}, X_{[zT]+\tau}) &= c(z, \tau) \\ &= \sum_{j=-\infty}^{-1} S_j(z) \Psi_j(\tau), \end{aligned} \quad (17)$$

where $\Psi_j(\tau)$ is the autocorrelation function of $\psi_{j,k}(t)$ and $[\cdot]$ denotes the integer part of the real number. That is,

$$\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{j,k}(0) \psi_{j,k}(\tau). \quad (18)$$

Once the localized autocovariance is defined, localized variance at time z can be defined by

$$v(z) = c(z, 0) = \sum_{j=-\infty}^{-1} S_j(z). \quad (19)$$

Another evidence for justifying that a time–frequency domain autocovariance in terms of wavelets preserves the intuitive sense of an autocovariance in time domain can be seen through the relationship between Allan variances [26], which is defined by

$$\bar{X}_t(\tau) = \frac{1}{\tau} \sum_{n=0}^{\tau-1} X_{t-n}, \quad (20)$$

and Haar wavelet variances [24]. It is worth noting that Allan variance involves merely localized time-domain representation. However, a Haar wavelet involves both localized time and frequency domain representations and, then, an Allan variance is

$$\sigma_X^2(\tau) = \frac{1}{2} E(|\bar{X}_t(\tau) - \bar{X}_{t-\tau}(\tau)|^2), \quad (21)$$

while a Haar coefficient on finest scale or smallest dilation, $j = -1$, on any translation or location k is

$$\begin{aligned} \bar{d}_{-1,k} &= 1/[\sqrt{2}(X_{2k+1} - X_{2k})], \\ k &= 0, \dots, T/2 - 1, \end{aligned} \quad (22)$$

then we can see that

$$\text{var}\{\hat{d}_{j,k}\} = E\hat{d}_{j,k}^2 = \tau_j \sigma_X^2(\tau_j). \quad (23)$$

Forecasting

Based on the background mentioned above, we are ready to present prediction algorithms for nonstationary processes. A Kalman filter can be used for prediction on a more parametric basis. Compared to a Kalman filter, local stationary wavelets do not necessarily need to fit data with a parametric model and may concern only measurement *per se*. In contrast to parametric analysis, a nonparametric analysis forsakes the idea that there is a model governing evolution of data, which in any case lets the observation characterize the data by itself.

Kalman Filter. The Kalman filter is a recursive filter which can estimate linear dynamic parameters given some noisy measurements. Here we should not take the word “filter” as only a passive data-processing algorithm but think of it as a computer program that can gather input information in order to optimally estimate the transition of a system. Another problem arises when one needs to determine whether a Kalman filter is able to properly estimate a nonstationary process. As we mentioned in the previous section on ARIMA, a nonstationary process can be treated as a stationary ARMA after taking appropriate orders of derivative on the ARIMA [27,28]. As a result, we can focus on only ARMA now without loss of generality, and the following is a predictive method using the concept of Kalman filter. If an ARIMA fails to be differentiated properly into an ARMA, one can still implement a Kalman filter onto an ARIMA process even if the ARIMA has missing observations [29]. Since the Gaussianity and linearity possessed in a Kalman filter are independent of the nonstationarity of a process, we can assume linearity and Gaussianity in a Kalman filter and use it to analyze a nonstationarity process without loss of generality. Readers are referred to Gordon [30], and Kitagawa [31] for more detailed methods for modeling non-Gaussian state space. According to McGonigal and Ionescu [28], a state space model must be found first before applying a Kalman filter for prediction. They proposed a simplified model based on Aoki’s method

which could model an ARMA state space [32]. In Huang *et al.*, [13], the authors gave an implementation of using Kalman filter to predict a time-varying AR-2 process, which is a nonstationary and Gaussian–Markovian process and can be described in a general form as

$$x_k = \sum_{i=1}^p \phi_{i,k} x_{k-i} + \varepsilon_k + \sigma_k. \quad (24)$$

If one compares Equation (24) with Equation (2), one should notice that the time-varying model mentioned in this section has an extra term σ_k . Huang added σ_k , a time-varying term, to better track the moving trend of wind speed, which is one of the applications discussed in his work. To describe this time-varying AR-2 model in a matrix form, we first define a vector space $\Psi = (\phi_1 \ \phi_2 \ \sigma)^T$ containing parameters that we need to decide later and then represent the differences of parameters as they evolve with time using their dynamic relationship

$$\begin{pmatrix} \Psi_k \\ \Gamma_k \end{pmatrix} = \begin{pmatrix} I & I \\ 0 & H \end{pmatrix} \begin{pmatrix} \Psi_{k-1} \\ \Gamma_{k-1} \end{pmatrix} + \begin{pmatrix} 0 \\ I \end{pmatrix} \Omega_k, \quad (25)$$

where I is the identity matrix and Ω is a zero-mean white-noise vector. If we define H as an identity matrix, Γ_k is a random walk process and Ψ_k ends up being an integrated random walk process. If H is a zero matrix, then Γ_k should be a white Gaussian noise process and Ψ_k ends up being a random walk process. Note that the role of Γ_k is the first-order difference of Ψ_k and it describes the changes between Ψ_k and Ψ_{k-1} as time moves on. We can predict our nonstationary process, which is a time-varying AR-2 model, once the crucial step of choosing parameters in our model is carried out, and this can be accomplished by the Kalman algorithm. A general way of implementation of a Kalman filter starts with setting up the parameter model in the state space form as

$$\tilde{z}_k = A \tilde{z}_{k-1} + B \tilde{\zeta}_k, \quad (26)$$

$$\tilde{y}_k = C \tilde{z}_k + \tilde{\eta}_k, \quad (27)$$

where $z_k = \begin{pmatrix} \Psi_k \\ \Gamma_k \end{pmatrix}$ is the state vector, A is state matrix, B is input matrix, C is output matrix, and y_k is the observation vector of dimension p , meaning $\tilde{y}_k = (x_k \ \cdots \ x_{k-p+1})^T$.

Furthermore, $\tilde{\zeta}$ and $\tilde{\eta}$ are zero-mean white-noise vector processes. Given the estimation of z_k as $\hat{z}_k$, one can write the estimation dynamics as

$$\hat{z}_{k|k-1} = A\hat{z}_{k-1}. \quad (28)$$

And then we define the covariance matrix of estimation errors $P_k = \langle (\hat{z}_k - \tilde{z}_k)^T (\hat{z}_k - \tilde{z}_k) \rangle$, state disturbance matrix $R = \langle \tilde{\zeta}_k^T \tilde{\zeta}_k \rangle$, and the noise variance matrix of measurement $Q = \langle \tilde{\eta}_k^T \tilde{\eta}_k \rangle$ where $\langle \cdot \rangle$ denotes expectation. $\hat{z}_k$ can be calculated through a recursive algorithm [33].

Local Stationary Wavelet (LSW). Given a series of t measurements, in which we want to use those t measurements to predict the process step h after t , we can write our interested process as $X_{0,T}, \dots, X_{t-1,T}$ where $T = t + h$ and then define a linear predictor as

$$\hat{X}_{t+h-1,T} = \sum_{s=0}^{t-1} b_{t-1-s,T}^{(t,h)} X_{s,T}. \quad (29)$$

The next task is to optimize $b_{t-1-s,T}^{(t,h)}$ so that mean square prediction error (MSPE) $E(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2$ is minimized. To utilize the properties of local stationary wavelet (LSW), we represent the approximate MSPE in a wavelet matrix form

$$\begin{aligned} E(\hat{X}_{t+h-1,T} - X_{t+h-1,T})^2 \\ = \tilde{b}_{t+h-1}' B_{t+h-1,T} \tilde{b}_{t+h-1}, \end{aligned} \quad (30)$$

where $\tilde{b}_t = (\tilde{b}_{t-1,T}^{(1)}, \dots, \tilde{b}_{0,T}^{(1)})$, and $B_{t,T}$ is the covariance matrix consisting of localized autocovariance whose size is $(t+1) \times (t+1)$ and (m,n) th element is $c(n+m/2T, n-m)$, which is the localized autocovariance involving the wavelet form as mentioned in the background LSW section. The whole procedure of forecasting a nonstationary process with long enough observations involves two training sets. First, divide the long-term

observations into two training sets and then use the first one to generate the $B_{t,T}$ matrix. Second, use the other training set to optimize $\tilde{b}_t$ vector such that MSPE is minimized. In practice, the first part involves selection of two more parameters, and readers are referred to Fryzlewicz *et al.* [34] for specific details. Once those parameters are selected, the model would be ready for prediction.

APPLICATION

In Gomez *et al.* [29], Kalman filter method is applied to five datasets (totally 621 observations) of monthly observations on international airline passengers. An ARIMA model is used to fit observations, and all parameters were estimated before the application of Kalman filter. Root mean square errors (RMSE) of predictions of 1 month ahead up to 12 months ahead were reported. For the one-month-ahead prediction, the RMSE is around 0.045 for all five datasets. For 12-month-ahead prediction, the RMSE increases to more than 0.080 for all five datasets. In Fryzlewicz *et al.* [34], LSW is applied to forecast how the wind speed anomaly index evolves up to nine steps (nine months) ahead. Modeling the wind speed anomaly index with LSW is intuitively legitimate because the variation of wind speed is expected to be smooth. After parameter training, their one-step-ahead prediction can have more than 90% of actual observations fall in the 95% prediction interval which is based on the assumption of Gaussianity and defined as

$$I_t = [-1.96\hat{\sigma}_t + \hat{X}_t, 1.96\hat{\sigma}_t + \hat{X}_t]. \quad (31)$$

The results of nonstationary forecasting in other references using other methods also are promising for only one-step-ahead prediction but not accurate enough for a longer prediction horizon.

MULTIVARIATE PROCESS MODELING AND FORECASTING

In this section we discuss in brief some extensions for multivariate nonstationary time series forecasting for the multivariate

processes. A direct extension of multivariate models can be obtained from the univariate models by recording the observational errors at each time point as vectors instead of scalars. Thus, the observational errors obtained from the multivariate models follow a multivariate normal distribution. However, the theory for univariate models cannot be directly extended to the multivariate models because of the uncertainty in the variance matrix and hence we do not have the luxury of performing the same conjugate analyses as what we did for univariate models. In other words, the univariate theory could be extended only under the assumption that the observational error variance matrix is known ahead of time, and this is usually not the case. For $t = 1, \dots$, we have $\tilde{x}_t$ as a column vector with r observations. The multivariate model is provided by the four-tuple system: $\{F, G, V, W\}_y = \{F_t, G_t, V_t, W_t\}$, where F_t is $n \times r$ dynamic regression matrix, G_t is $n \times n$ state evolution matrix, V_t is $r \times r$ observational variance matrix, and W_t is a $n \times n$ evolution variance matrix, and we assume that all of them are known. Now, the defining equations of the model are given by

$$\tilde{x}_t = F_t \tilde{\theta}_t + \tilde{v}_t, \quad \tilde{v}_t \sim N[0, V_t], \quad (32)$$

$$\tilde{\theta}_t = G_t \tilde{\theta}_{t-1} + \tilde{\omega}_t, \quad \tilde{\omega}_t \sim N[0, W_t], \quad (33)$$

where $\tilde{v}_t$ and $\tilde{\omega}_t$ are the error sequences that are internally mutually independent and $\tilde{\theta}_t$ is the n -dimensional state vector. Equation (32) provides the sampling distribution of $\tilde{X}_t$ conditional on $\tilde{\theta}_t$, and Equation (33) is the state equation that provides the state of the system over time. Now that the variance matrices are assumed to be known, we can directly extend the methodologies developed for univariate models to the above multivariate model. A detailed analysis of this extension is provided in Ref. 35. A number of models have been developed for the case where the covariance matrix for the multivariate time series is not known ahead of time. A framework was developed for this case [36,37]. The model assumes that the analyses are performed on different time series that are alike and each of these time series follow the case of univariate dynamic model. The components of the four-tuple systems are assumed

to be common for each of the time series and the defining equations for each of the time series is provided in a similar manner. Once again, the error sequences in the equations are independent. This framework permits the development of various models for the multivariate time series analysis with unknown variance matrix. The observational error vector is assumed to follow a multivariate normal distribution and the evolution error matrix follows a matrix-variate normal distribution. Several models were developed from this framework and extended [38–40].

CONCLUSION

In this article, the concept of nonstationarity in stochastic processes was presented, explaining how to preprocess nonstationary processes and forecast their trend parametrically and nonparametrically. The applications of forecasting nonstationary processes span widely from air pollution to electroencephalograph analysis. Those applications and their forecasting results in detail can be found in the respective references. Besides these two methods we introduced, there are other methods for forecasting a nonstationary process. Here, we only present a general concept of this topic. In different applications, there may be a favorable method with more specific parameter selection criteria to better model and forecast the process.

REFERENCES

1. Appel U, Brandt AV. Adaptive sequential segmentation of piecewise stationary time series. *Inf Sci* 1983;29(1):27–56.
2. Adak S. Time-dependent spectral analysis of nonstationary time series. *J Am Stat Assoc* 1998;93(444):1488–1489.
3. Harrison PJ, Stevens CF. Bayesian forecasting. *J R Stat Soc Ser B (Methodological)* 1976;38:205–247.
4. Pereda E, Quiroga RQ, Bhattacharya J. Nonlinear multivariate analysis of neurophysiological signals. *Prog Neurobiol* 2005;77(1–2):1–37.
5. Hlaváčková-Schindler K, Paluš M, Vejmelka M, *et al.* Causality detection based on information-theoretic approaches in time series analysis. *Phys Rep* 2007;441(1):1–46.

6. West M, Prado R, Krystal AD. Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *J Am Stat Assoc* 1999;94(446):375–376.
7. Akaike H. Information theory and an extension of the maximum likelihood principle. Budapest, Hungary; 1973. pp. 267–281.
8. Hannan EJ, Quinn BG. The determination of the order of an autoregression. *J R Stat Soc* 1979;41(2):190–195.
9. Cox DR, Hinkley DV. Theoretical statistics. London: Chapman and Hall; 1974.
10. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989;57(2):307–333.
11. Abramovich YI, Spencer NK, Turley MDE. Time-varying autoregressive (TVAR) adaptive order and spectrum estimation. 1928. pp. 89–93.
12. Box GEP, Jenkins G. Time series analysis, forecasting and control. San Francisco (CA): Holden-Day, Incorporated; 1990.
13. Huang Z, Chalabi ZS. Use of time-series analysis to model and forecast wind speed. *J Wind Eng Ind Aerodyn* 1995;56(2–3):311–322.
14. Barnett G, Kohn R, Sheather S. Bayesian estimation of an autoregressive model using markov chain monte carlo. *J Econ* 1996;74(2):237–254.
15. Tamura YH. An approach to the nonstationary process analysis. *Ann Inst Stat Math* 1987;39(1):227–241.
16. McCulloch RE, Tsay RS. Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J Am Stat Assoc* 1993;88:968–978.
17. Geweke J, Whiteman C. Bayesian forecasting. In: Elliott G, Granger CWJ, Timmermann A, editors. *Handbook of economic forecasting*; Amsterdam, Netherlands; Elsevier B.V; 2006.
18. Robinson PM. Root-N-consistent semiparametric regression. *Econometrica (J Economet Soc)* 1988;56:931–954.
19. Velasco C. Gaussian semiparametric estimation of non-stationary time series. *J Time Ser Anal* 1999;20(1):87–127.
20. Damian D, Sampson PD, Guttorp P. Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics* 2001;12(2):161–178.
21. Qian S, Chen D. Joint time-frequency analysis. *Signal Process Mag IEEE* 1999;16(2):52–67.
22. Taswell C, Toolsmiths C, Stanford CA. The what, how, and why of wavelet shrinkage denoising. *Comput Sci Eng [See also IEEE Computational Science and Engineering* 2000;2(3):12–19.
23. Coifman RR, Donoho DL. Translation-invariant de-noising, *Lecture Notes in Statistics*. New York: Springer; 1995. pp. 125–125.
24. Nason GP. Wavelets in time-series analysis. *Philos Trans R Soc Math Phys Eng Sci* 1999;357(1760):2511–2526.
25. Brillinger D. Time series: data analysis and theory. Philadelphia (PA); Society for industrial Mathematics; 1975.
26. Allan DW. Statistics of atomic frequency standards. *Proc IEEE* 1966;54(2):221–230.
27. Kalman RE. A new approach to linear filtering and prediction problems. *J Basic Eng* 1960;82(1):35–45.
28. McGonigal D, Ionescu D. An outline for a Kalman filter and recursive parameter estimation approach applied to stock market forecasting. In: *Electrical and Computer Engineering, 1995. Canadian Conference on Volume 2*; Montreal, Canada; 1995 Sep 5–8. pp. 1148–1151.
29. Gomez V, Maravall A. Estimation, prediction, and interpolation for nonstationary series with the Kalman filter. *J Am Stat Assoc* 1994;89(426):611–624.
30. Gordon NJ, Salmond DJ, Smith AFM. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc F, Radar and Signal Processing* 1993;140(2):107–113.
31. Kitagawa G, Kohn R, Ansley CF, *et al.* Non-gaussian state-space modeling of non-stationary time series. *J Am Stat Assoc* 1987;82(400):1032–1041.
32. Aoki M, Havenner A. State space modeling of multiple time series. *Economet Rev* 1991;10(1):1–59.
33. Young P. Recursive approaches to time series analysis. *Bull Inst Math Appl* 1974;10:209–224.
34. Fryzlewicz P, Van Belleghem S, von Sachs R. Forecasting non-stationary time series by wavelet process modelling. *Ann Inst Stat Math* 2003;55(4):737–764.
35. West M, Harrison J. Bayesian forecasting and dynamic models. Springer; 1997.
36. Quintana JM, West M. Multivariate time series analysis: new techniques applied to

- international exchange rate data. *Statistician* 1987;36:275–281.
37. Quintana JM, West M. Time series analysis of compositional data. *Bayesian Stat* 1988;3:747–756.
38. Mardia KV, Kent JT, Bibby JM. *Multivariate analysis*. New York: Academic press; 1979.
39. Harvey AC. Analysis and generalisation of a multivariate exponential smoothing model. *Manage Sci* 1986;32:374–380.
40. Highfield RA. *Forecasting with Bayesian state space models*. Chicago (IL): Graduate School of Business, University of Chicago; 1987.

FORECASTING: STATE-SPACE MODELS AND KALMAN FILTER ESTIMATION

KEMAL GÜRSOY

Department of Mathematics,
Boğaziçi University, Bebek,
Istanbul, Turkey

MELIKE BAYKAL-GÜRSOY

Department of Industrial and
Systems Engineering, RUTCOR,
CAIT, Rutgers University, New
Brunswick, New Jersey

INTRODUCTION

Humans observe their environment, learn its properties, and by forecasting the future events, they make plans to influence future outcomes to their benefit. There are risks that those plans, for which actions are based, are not good at all. Hence, it is desirable to set up expectations, based on the detailed analysis of empirical evidence and sound knowledge in order to reduce the risk of future regrets.

Forecasting methods to set up reasonable expectations for the future events are mainly based on the properties of relevant historical data (time series) assuming that the environment will not change significantly. Forecasting models express relationships between what the past evidence shows and what could possibly take place in the future.

System models are useful concepts to represent relations, interactions, and to build up tools for predicting the collective behavior of entities. An input–output model expresses the behavior of an open system that interacts with its environment, in a causal manner (the past and the present of the system shape its future), which takes from its environment (input, or stimulus) and gives to its environment (output, or response). If one is interested in the behavior of an observable (measurable) open system, then one should observe the system for some time for

stimulus–response relations and construct an input–output model that suitably fits to these observations.

A similar causal model (local in time) is the state-space representation of an open system, where the underlying system's behavior is represented by its state and its response to an input at that state. The collection of all possible (admissible) states of a system is known as the *state-space* [1–3]. In order to predict the future states of a dynamic system, with minimum mean square error, we can design a recursive predictor based on the current estimate and the current filtered estimation error. The best estimate of the state of a linear stochastic system from partial observations is known as *Kalman filter* [4,5] (also known as *Kalman–Bucy filter*).

Previous work on this subject has been done by Kolmogorov [6], Wiener [7], Kalman [4], Bucy [8], and many others in the last century. We can also mention closely related works by Wald [9], and Robbins and Monroe [10], where the *stochastic approximation* algorithm developed in Ref. 10 may be utilized for parameter estimation.

Stationary and nonstationary time series models such as autoregressive (AR), moving average (MA), mixed autoregressive and moving average (ARMA), autoregressive integrated moving average (ARIMA), autoregressive and moving average with external input (ARMAX), can be translated into state-space models [2,3,11–13], thus benefit from the recursive prediction algorithm. Multivariate time series are especially suitable to state-space representation. For example, Aviv [14] uses state-space models and Kalman filter in demand estimation and inventory control. Bayesian estimation can also exploit the Kalman filtering methodology, as discussed in Refs 15 and 16.

We introduce the general state-space model in the next section, and then focus on linear Gaussian model in the section titled “Linear Gaussian Model.” We describe Kalman filter iterations in the section titled “Kalman Filter for State Estimation.” The

adaptive and nonlinear generalizations follow in the sections titled “Adaptive State Estimation” and “Nonlinear State Estimation,” respectively. The parameter estimation can also be approached by utilizing a Kalman filter, as presented in the section titled “Parameter Estimation.”

STATE-SPACE STRUCTURE

A mathematical model of a dynamic system with multiple inputs and outputs can be constructed as follows. Let x_t be the multidimensional (vector) state of the system, $\mathcal{X}$ be the state-space, u_t be an observable multidimensional input to the system, and y_t be an observable multidimensional output of the system, at time t . Also, let v_t and w_t denote multidimensional random noise processes, at time t . Assume $x_{t+1} = f_t(x_t, u_t, v_t)$, $y_t = g_t(x_t, w_t)$, where f, g are measurable functions at all time. By using this simple model, we try to capture and express the relationships between the state of the system, the input that drives it, and the output of the system. The k -step ahead state, x_{t+k} , depends upon the present state, x_t , as well as the future inputs, such that, $x_{t+k} = f_{t+k-1}(f_{t+k-2}(\dots(f_t(x_t, u_t, v_t))\dots), u_{t+k-1}, v_{t+k-1})$. This recursion is the basis for state transitions from period t to k -step ahead period $t+k$.

LINEAR GAUSSIAN MODEL

In this section, we will consider linear state-space models operating in continuous or discrete time. Let the initial (starting) state be denoted as $x(t_0)$ for the continuous and $x(k_0)$ for the discrete-time process.

Continuous-Time Model. Assume that the state dynamics, or evolution is as follows,

$$\frac{d}{dt}x(t) = A_t x(t) + B_t u(t) + G_t v(t).$$

Also the observed output, or response is given by

$$y(t) = C_t x(t) + w(t),$$

where $x(t), u(t), y(t)$ are finite real vectors, A_t, B_t, C_t, G_t are real matrices with proper dimensions and $t \geq t_0$ is a real number.

The noise processes $\{v(t)\}, \{w(t)\}$ are independent Gaussian random vectors, with zero means and finite variances, conditionally independent of $x(t)$ and $u(t)$. Let $\{v(t)\}$ be a white Gaussian stochastic process with covariance

$$E[v(t)v^T(t-\tau)] = Q(t)\delta(t-\tau),$$

and $\{w(t)\}$ be a white Gaussian stochastic process with covariance

$$E[w(t)w^T(t-\tau)] = R(t)\delta(t-\tau),$$

where $\delta(\cdot)$ is the Kronecker delta function, that is, its value is one when its argument is zero, its value is zero otherwise. Since these stochastic processes are uncorrelated, we have

$$\begin{aligned} E[v(t)w^T(t)] &= E[x(0)v^T(t)] \\ &= E[x(0)w^T(t)] \\ &= E[u(t)v^T(t)] \\ &= E[u(t)w^T(t)] = 0. \end{aligned}$$

Discrete-Time Model. Assume that the state dynamics is given by

$$x(k+1) = A_k x(k) + B_k u(k) + G_k v(k),$$

and the observation equation is

$$y(k) = C_k x(k) + w(k),$$

where $x(k), u(k), y(k)$ are finite real vectors, A_k, B_k, C_k, G_k are real matrices with proper dimensions and $k \geq k_0$ is an integer.

The noise processes $\{v(k)\}, \{w(k)\}$ are independent Gaussian random vectors, with zero means and finite variances, conditionally independent of $x(k)$ and $u(k)$, and temporally uncorrelated, that is, $E[v(k)v^T(j)] = Q(k)\delta_{kj}$, $E[w(k)w^T(j)] = R(k)\delta_{kj}$, where δ_{kj} is the Kronecker delta, that is, its value is one when $k=j$, and zero otherwise.

In these models, the system's evolution can be represented by state transition,

$$x(t|t_0) = \Phi(t, t_0)x_{t_0} + \int_{t_0}^t \Phi(t, s)B(s)u(s) ds,$$

where $\Phi(., .)$ is the state-transition function (a matrix, or a dynamic semigroup operator on the state-space), $x(t_0)$ is the initial state of the system and integral is a summation operator in the Lebesgue sense [17] that operates both in the continuous time and in the discrete time, componentwise.

Note that for linear systems, the initial state is a critical point for the future trajectory of the system. On the other hand, if the system is not linear, then the initial state of the system is not sufficient for the system's behavior, system's trajectories would depend upon other factors as well. Many trajectories (perhaps infinitely many) from the same initial state are possible. There are two essential questions:

- Q1. Is the system *observable*, that is, is it possible to determine the states of the system $\{x(s)|t_0 \leq s < t\}$, from the observations of $y(s)$ and $u(s)$, for $s \in [t_0, t)$?
- Q2. Is the system *controllable*, that is, is it possible to drive the system into a desired state $x(t_f)$, from an initial state $x(t_0)$, by selecting a stream of feasible inputs (stimuli) $\{u(s)|t_0 \leq s \leq t_f\}$, in a finite interval of time $[t_0, t_f]$?
- For the discrete-time system dynamics, we replace t, t_0, t_f by k, k_0, k_f , respectively.

The duality of these questions, as well as the state-space formulation of linear systems for predicting and controlling the system's behavior were explained by many researchers, for one particular approach, see Ref. 5.

The first question is relevant to forecasting, or estimating the states of a linear dynamical system, and can be approached by the Wiener filter [7]. Wiener filter provides the best estimates (in the sense of minimum mean square error [18]) of the Markovian signals $x(t)$, by using a set of observations $\{y(s), s \in [t_0, t]\}$. State estimates, $\hat{x}(s|t)$ may

be obtained for prediction or $\hat{x}(t|t) := \hat{x}(t)$, for filtering when $s > t > t_0$. The Wiener filter is based on a function, $h(t)$, such that its convolution with the past observations, $h \star y$, would generate the state estimator,

$$\hat{x}(t) = \int_0^\infty h(s)y(t-s) ds.$$

The function, $h(t)$ is selected to minimize the expected squared estimation error,

$$E[(x(t) - \hat{x}(t))^2] = E[x^2(t)] - 2E[\hat{x}(t)x(t)] + E[\hat{x}^2(t)].$$

Since $E[\hat{x}^2(t)] = \int_0^\infty h(s) ds \int_0^\infty h(\tau) R_{yy}(t, \tau) d\tau$ and $E[\hat{x}(t)x(t)] = \int_0^\infty h(s) R_{xy}(t, s) ds$, where R_{yy} is the autocorrelation function of the observations $y(t)$ and R_{xy} is the correlation function of $x(t)$ and $y(t)$, then the optimum filter, say h^* satisfies,

$$\int_0^\infty h^*(\tau) R_{yy}(t, \tau) d\tau = R_{xy}(t),$$

which is known as the *Wiener-Hopf equation*, and its solution for h^* yields the best linear estimator for state $x(t)$ [7].

KALMAN FILTER FOR STATE ESTIMATION

The Kalman filter, also known as *Kalman-Bucy filter*, is a sequential estimator for the states of a stationary linear dynamical system and it is based on the Wiener filter, subject to a white Gaussian noise. Without loss of generality, we can set the initial time t_0 to be 0. The usual Kalman filter formulation, such as given in Ref. 19, for a known distribution of the initial state, $x(0)$, is carried on the following simplified state-space equations. Note that, the control component is not included in this model, thus one is concerned only with the state estimation problem.

Continuous Time:

$$\begin{aligned} \frac{d}{dt}x(t) &= A_t x(t) + G_t v(t) \text{ and} \\ y(t) &= C_t x(t) + w(t), \text{ for } t \geq 0. \end{aligned}$$

Discrete Time:

$$\begin{aligned} x(k+1) &= A_k x(k) + G_k v(k) \text{ and} \\ y(k) &= C_k x(k) + w(k), \text{ for } k \geq 0. \end{aligned}$$

The Kalman filter first estimates the state of the system, then updates the estimation process, according to the estimation error that manifests itself in the difference of actual observation and the estimated observation which is based on the state estimation. Hence, these two stages of estimation–correction process repeats itself for the entire forecasting horizon. In our model, noises and states are independent Gaussian random processes with finite second moments. Also let $E[x(0)] = \bar{x}(0)$ and $\text{Var}(x(0)) = P_0$.

Continuous-Time Case

Let $\hat{x}(s|t)$ be the estimate of $x(s)$, given all $y(\tau)$, for $0 \leq \tau \leq t$ and $s > t \geq 0$. Also, let $e(s|t) = x(s) - \hat{x}(s|t)$ be the estimation error, at time s .

Therefore, one must find the linear filter $K(t, \tau)$ such that $\hat{x}(s|t) = \int_0^t K(s, \tau) y(\tau) d\tau$, where $\hat{x}(0|0) = \bar{x}(0)$. Also, the mean squared error, $E[\|e(s|t)\|^2]$ is minimized by this filter, where $\|e(s|t)\|$ denotes the Euclidean norm of the error vector.

If this filter, $K(t, \tau)$, satisfies the Wiener–Hopf equation, then $\hat{x}(s|t)$ will be an unbiased minimum variance estimator of $x(s)$ (for the proof see Ref. 8). Hence, this filter will satisfy,

$$\frac{d}{dt} \hat{x}(t|t) = \int_0^t \frac{\partial}{\partial t} K(t, \tau) y(\tau) d\tau + K(t, t) y(t),$$

and

$$\frac{d}{dt} \hat{x}(t|t) = A_t \hat{x}(t|t) + K(t, t) [y(t) - C_t \hat{x}(t|t)],$$

where $\hat{x}(0|0) = \bar{x}(0)$. The estimation error, $e(t|t) = x(t) - \hat{x}(t|t)$, will be subject to the following evolution,

$$\begin{aligned} \frac{d}{dt} e(t|t) &= [A_t - K(t, t) C_t] e(t|t) + G_t v(t) \\ &\quad - K(t, t) w(t). \end{aligned}$$

Let the covariance of the estimation error be $E[e(t|t)e^T(t|t)] = P(t)$, and $P(0) = P_0$.

In order to obtain the least square error estimator, the estimation error, $e(t|t)$, must be orthogonal to the state estimate, $\hat{x}(t|t)$, that is, $E[e(t|t)\hat{x}(t|\tau)] = 0$, for $t \geq \tau \geq 0$.

Thus, by solving for the above estimation error differential equation, we obtain the evaluation of the error function as follows:

$$\begin{aligned} e(t|t) &= \Psi(t, 0) e(0|0) \\ &\quad - \int_0^t \Psi(t, \tau) [K(\tau, \tau) w(\tau) - G_\tau v(\tau)] d\tau, \end{aligned}$$

where $\Psi(\cdot, \cdot)$ denotes the transition function for e . Therefore,

$$\begin{aligned} P(t) &= \Psi(t, 0) E[e(0|0)e^T(0|0)] \Psi^T(t, 0) \\ &\quad - \int_0^t \int_0^t \Psi(t, \tau) [K(\tau, \tau) \\ &\quad \times E[w(\tau)w^T(s)] K^T(s, s)] \Psi^T(t, s) ds d\tau \\ &\quad + \int_0^t \int_0^t \Psi(t, \tau) G_\tau E[v(\tau)v^T(s)] G_s^T \\ &\quad \times \Psi^T(t, s) ds d\tau. \end{aligned}$$

By taking the time derivative of the above function, we have the following Riccati equation for the covariance of the estimation error,

$$\begin{aligned} \frac{d}{dt} P(t) &= G_t Q(t) G_t^T + A_t P(t) + P(t) A_t^T \\ &\quad - P(t) C_t^T R^{-1}(t) C_t P(t), \end{aligned}$$

where, $P(t) C_t^T = E[e(t|t)e^T(t|t)] C_t^T = K(t, t) R(t)$. Hence, this yields $K(t, t) = P(t) C_t^T R^{-1}(t)$.

In summary, the Kalman filter is defined by the following equations, for $t \geq 0$:

$$\frac{d}{dt} \hat{x}(t|t) = A_t \hat{x}(t|t) + K(t, t) [y(t) - C_t \hat{x}(t|t)],$$

where $\hat{x}(0|0) = \bar{x}(0)$,

$$\begin{aligned} K(t, t) &= P(t) C_t^T R^{-1}(t), \\ \frac{d}{dt} P(t) &= A_t P(t) + P(t) A_t^T + G_t Q(t) G_t^T \\ &\quad - P(t) C_t^T R^{-1}(t) C_t P(t). \end{aligned}$$

The forecasting of future states can be done by $\hat{x}(s|t) = \Phi(s, t) \hat{x}(t|t)$, for $s > t > 0$ and by using the state-transition function $\Phi(\cdot, \cdot)$.

Discrete-Time Case

Let $\hat{x}(k|m) = E[x(k)|y(0), y(1), \dots, y(m)]$ be the state estimator and $e(k|m) = x(k) - \hat{x}(k|m)$ be the estimation error, for $k \geq m$. The best estimator of states has the minimum mean square estimation error property, which is obtained by the orthogonal projection of the state estimator on the observations, with the following Kalman recursions.

$$\begin{aligned}\hat{x}(k|k) &= \hat{x}(k|k-1) + P(k-1)C_k^T \\ &\quad \times [C_k P(k-1)C_k^T + R(k)]^{-1}(y(k) \\ &\quad - C_k \hat{x}(k|k-1)), \\ P(k|k) &= P(k-1) - P(k-1)C_k^T [C_k P(k-1)C_k^T \\ &\quad + R(k)]^{-1} C_k P(k-1),\end{aligned}$$

where $P(k|k) = E[e(k|k)e^T(k|k)]$. Also, $\hat{x}(k+1|k) = A_k \hat{x}(k|k)$, that is,

$$\begin{aligned}\hat{x}(k+1|k) &= A_k \hat{x}(k|k-1) + A_k P(k-1)C_k^T \\ &\quad \times [C_k P(k-1)C_k^T + R(k)]^{-1}(y(k) \\ &\quad - C_k \hat{x}(k|k-1)), \quad \text{and} \\ e(k+1|k) &= A_k e(k|k-1) + A_k P(k-1)C_k^T \\ &\quad \times [C_k P(k-1)C_k^T + R(k)]^{-1}(y(k) \\ &\quad - C_k \hat{x}(k|k-1) - G_k v(k)).\end{aligned}$$

The estimation error covariance, $P(k) = E[e(k+1|k)e^T(k+1|k)]$, that is,

$$\begin{aligned}P(k) &= A_k P(k-1)A_k^T - A_k P(k-1)C_k^T \\ &\quad \times [C_k P(k-1)C_k^T + R(k)]^{-1} C_k P(k-1) \\ &\quad \times A_k^T + G_k Q(k)G_k^T,\end{aligned}$$

where $\hat{x}(0|0) = E[x(0)]$ and $P(0) = \text{Var}(x(0))$.

This discrete-time Kalman filter is a minimum variance and minimum mean square error sequential estimator of the state.

We can construct a predictor, based on this discrete Kalman filter, to forecast a distant future, $\hat{x}(n|k) = \prod_{j=k, \dots, n} A_j \hat{x}(k|k)$, for $n > k > 0$.

In this approach, predictor-corrector structure of the Kalman filter is generating an *innovation* sequence, $y(k) - \hat{y}(k)$, which is a zero-mean and independent stochastic

process (based on the maximum likelihood estimation, or, orthogonal projection method). There are other approaches for optimal state estimators, such as recursive Bayesian estimators, but when the system is linear and the noise structure is independent Gaussian, these methods converge to the Kalman filter [8].

ADAPTIVE STATE ESTIMATION

When there are uncertainties in the system parameters, Kalman filter cannot provide the state estimation. We can approach to those cases with a Bayesian construct [15,16,20]. Let the parameter vector θ is subject to uncertainty, where the linear system dynamics are as before.

For continuous time: $\frac{d}{dt}x(t) = A(t, \theta)x(t) + G(t, \theta)v(t)$ and $y(t) = C(t, \theta)x(t) + w(t)$, where $v(t)$ and $w(t)$ are independent zero-mean white Gaussian random vectors with $Q(t)\delta(t)$ and $R(t)\delta(t)$ covariances, respectively. The initial state is an independent Gaussian random vector with mean $\hat{x}(0|0, \theta)$ and variance $P(0|0, \theta)$.

The system evolution is subject to its unknown parameter vector, $\theta \in \Omega$, with its prior probability density $f(\theta)$. On the basis of the observation data, $Y_t = \{y(s)|0 \leq s \leq t\}$, the minimum mean square estimate of the state, $\hat{x}(t|t) = \int_{\Omega} \hat{x}(t|t, \theta) f(\theta|Y_t) d\theta$, where $\hat{x}(t|t) = E[x(t)|Y_t]$, $\hat{x}(t|t, \theta) = E[x(t)|Y_t, \theta]$, over the sample space of the parameter, Ω .

The posterior probability density of the parameter, based on the observations, would be as follows:

$$\begin{aligned}f(\theta|Y_t) &= f(\theta) \exp \left(- \int_0^t \hat{x}(\tau|\tau, \theta) C^T(\tau, \theta) \right. \\ &\quad \times R^{-1}(\tau) y(\tau) d\tau \\ &\quad \left. - \frac{1}{2} \int_0^t \|C(\tau, \theta) \hat{x}(\tau|\tau, \theta)\|^2 R^{-1}(\tau) d\tau \right) \\ &\quad \times \exp \left(- \int_0^t \hat{y}^T(\tau|\tau) R^{-1}(\tau) y(\tau) d\tau \right. \\ &\quad \left. - \frac{1}{2} \int_0^t \|\hat{y}(\tau|\tau)\|^2 R^{-1}(\tau) d\tau \right),\end{aligned}$$

where $\hat{y}(\tau|\tau) = \int_{\Omega} C(\tau, \theta) \hat{x}(\tau|\tau) f(\theta|Y_t) d\theta$.

The conditional covariance of the state estimation error would be,

$$P(t) = \int_{\Omega} P(t|t, \theta) [\hat{x}(t|t, \theta) - \hat{x}(t|t)] [\hat{x}(t|t, \theta) - \hat{x}(t|t)]^T f(\theta|Y_t) d\theta,$$

where $P(t|t, \theta) = E[(x(t) - \hat{x}(t|t, \theta))(x(t) - \hat{x}(t|t, \theta))^T | Y_t, \theta]$, for all $\theta \in \Omega$.

The same procedure is also suitable for the discrete-time estimations, as follows.

For discrete time: $x(k+1) = A(k, \theta)x(k) + v(k)$ and $y(k) = C(k, \theta)x(k) + w(k)$, where $v(k)$, $w(k)$ are independent zero-mean Gaussian

vectors with covariances $Q(k)\delta_{kj}$ and $R(k)\delta_{kj}$, respectively.

The initial state is an independent random vector with mean $\hat{x}(0|0, \theta)$ and covariance $P(0|0, \theta)$. Let the unknown parameter has a prior probability density $f(\theta)$. Hence, based on a realization of past observations, $Y_k = \{y(j) | j = 1, 2, \dots, k\}$, the minimum mean square error state estimate would be $\hat{x}(k|k) = \int_{\Omega} \hat{x}(k|k, \theta) f(\theta|Y_k) d\theta$, where $\hat{x}(k|k) = E[x(k)|Y_k]$ and $\hat{x}(k|k, \theta) = E[x(k)|Y_k, \theta]$.

The posterior probability of the parameter, based on the observation data,

$$f(\theta|Y_k) = \frac{\left[|P_z(k|k-1, \theta)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|y(k) - C(k, \theta)\hat{x}(k|k-1, \theta)\|^2 P_z^{-1}(k|k-1, \theta)\right) f(\theta|Y_{k-1}) \right]}{\left[\int_{\Omega} f(\theta|Y_k) |P_z(k|k-1, \theta)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \|y(k) - C(k, \theta)\hat{x}(k|k-1, \theta)\|^2 P_z^{-1}(k|k-1, \theta)\right) d\theta \right]},$$

where $y(k) - C(k, \theta)\hat{x}(k|k-1, \theta)$ is the conditional white noise process with covariance $P_z(k|k-1, \theta) = C(k, \theta)P(k|k-1, \theta)C^T(k, \theta) + R(k)$.

Moreover, the state estimation error covariance would be $P(k|Y_k) = \int_{\Omega} (P(k|k, \theta) + [\hat{x}(k|k, \theta) - \hat{x}(k|k)][\hat{x}(k|k, \theta) - \hat{x}(k|k)]^T) f(\theta|Y_k) d\theta$.

NONLINEAR STATE ESTIMATION

The state estimation for nonlinear systems is very difficult, to say the least, unlike the linear systems. Under suitable conditions, a linearization could be an approximation of a nonlinear system's representation, then we may employ the Kalman filter approach, but always be mindful of the divergence of the approximation from the actual system.

Let us modify our state-space model for the nonlinear structure, as follows:

1. $\frac{d}{dt}x(t) = g(x, t) + v(t)$ and $y(t) = h(x, t) + w(t)$, for continuous time,
2. $x(k+1) = g(x, k) + v(k)$ and $y(k) = h(x, k) + w(k)$, for discrete time,

where x, y are state and output (observation) vectors, also let g, h are differentiable vector functions, representing the system dynamics (evolution), and v, w are zero-mean independent Gaussian random vectors with time-varying covariances, Q and R , respectively.

1. *Continuous Time.* At time t , the first-order Taylor series expansion for g and h , around the estimated state, $\hat{x}$, would be as follows:

$$g(x, t) = g(\hat{x}, t) + A(\hat{x}, t)(x - \hat{x}) + \epsilon$$

$$\text{and } h(x, t) = h(\hat{x}, t) + C(\hat{x}, t)(x - \hat{x}) + \varepsilon,$$

where ϵ, ε are the approximation errors, such that

$$A(\hat{x}, t) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \dots & \frac{\partial g_n}{\partial x_n} \end{pmatrix},$$

$$C(\hat{x}, t) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \frac{\partial h_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial h_m}{\partial x_1} & \dots & \frac{\partial h_m}{\partial x_n} \end{pmatrix},$$

$$\text{and } H_i(\hat{x}, t) = \begin{pmatrix} \frac{\partial^2 h_i}{\partial x_1^2} & \dots & \frac{\partial^2 h_i}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 h_i}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 h_i}{\partial x_n^2} \end{pmatrix},$$

for $i = 1, \dots, m$.

Here, the functions g and h are evaluated at $\hat{x}(t)$.

Then the extended Kalman filter for the state estimation would be as follows [21]:

$$\begin{aligned} \frac{d}{dt}\hat{x}(t) &= g(\hat{x}, t) + \hat{P}(t)C(\hat{x}, t)R^{-1}(t) \\ &\quad \times [y(t) - h(\hat{x}, t)], \\ \frac{d}{dt}\hat{P}(t) &= \hat{P}(t)A^T(\hat{x}, t)\hat{P}(t) + Q(t) - \hat{P}(t) \\ &\quad \times C^T(\hat{x}, t)R^{-1}(t)C(\hat{x}, t)\hat{P}(t) \\ &\quad + \sum_{i=1, \dots, m} \hat{P}(t)H_i(\hat{x}, t)I[i]R^{-1}(t) \\ &\quad \times [y(t) - h(\hat{x}, t)], \end{aligned}$$

where $I[i]$ is the i th row of the identity matrix I .

2. *Discrete Time.* The extended Kalman filter construct can be carried out as follows,

$$\begin{aligned} \hat{x}(k+1|k+1) &= g(\hat{x}(k|k)) + \hat{P}(k+1|k+1) \\ &\quad \times C^T(\hat{x}, k+1)R^{-1}(k+1) \\ &\quad \times [y(k+1) - C(\hat{x}, k+1) \\ &\quad \times g(\hat{x}(k|k))], \\ \hat{P}(k+1|k+1) &= [R(k) + C^T(\hat{x}, k+1) \\ &\quad \times P(k+1|k)C(\hat{x}, k+1)]^{-1} \\ &\quad \times \hat{P}(k+1|k), \\ \hat{P}(k+1|k) &= A(\hat{x}, k)\hat{P}(k|k)A^T(\hat{x}, k) \\ &\quad + Q(k). \end{aligned}$$

There are methods to improve the convergence of extended Kalman filter [19,22] and apply these improved extended Kalman filters to adaptive controller design [23,24].

The linearization could also be made around a reference (nominal) trajectory, $\bar{x}(t)$, with a given $\bar{x}(0)$, such that $\frac{d}{dt}\bar{x}(t) = g(\bar{x}(t), t)$, for $t \geq 0$.

Also, let $\Delta x(t) = x(t) - \bar{x}(t)$ be a Gaussian perturbation from the reference trajectory. Hence, $\frac{d\Delta x(t)}{dt} = g(x(t), t) - g(\bar{x}(t), t) + v(t)$. Therefore, the first-order Taylor series approximation would be, $g(x(t), t) - g(\bar{x}(t), t) = A(\bar{x}(0), t)\Delta x(t)$, where $A(\bar{x}(0), t)$

is evaluated along the reference trajectory, $\frac{d\Delta x(t)}{dt} = A(\bar{x}(0), t)\Delta x(t) + v(t)$. Hence, the Kalman filter approach could be used with $\hat{x}(0) = \bar{x}(0)$.

PARAMETER ESTIMATION

We can estimate the parameter, together with the state of a system, by using the above constructed extended Kalman filter structure. Let the state vector x be augmented with the parameter vector θ , such that,

$$z(t) = \begin{pmatrix} x(t) \\ \theta(t) \end{pmatrix} \text{ and } V(t) = \begin{pmatrix} v(t) \\ 0 \end{pmatrix}.$$

1. $\frac{d}{dt}z(t) = g(z, t) + V(t)$ and $y(t) = h(z, t) + w(t)$, for continuous time,
2. $z(t+1) = g(z, k) + V(k)$ and $y(k) = h(z, k) + w(k)$, for discrete time.

A linearization is obtained by the first-order Taylor series expansion, around the augmented state estimate $\hat{z}$. Without loss of generality, in discrete time a linearized Kalman filter that generates state and parameter estimation would be as follows,

$$\begin{aligned} \hat{z}(k+1) &= \hat{z}(k+1|k) + K(k+1) \\ &\quad \times [y(k+1) - h(\hat{z}(k+1|k), k)], \end{aligned}$$

with $\hat{z}(0|0) = z(0)$,

$$h(\hat{z}(k+1|k), k) = C(k, \hat{\theta}_k)\hat{x}(k+1|k),$$

$$\hat{z}(k+1|k) = F(k)\hat{z}(k),$$

$$K(k+1) = P(k+1|k+1)G(k)R^{-1}(k),$$

$$\begin{aligned} P(k+1|k+1) &= P(k+1|k) - P(k+1|k)G(k) \\ &\quad \times [G^T(k)P(k+1|k)G(k) \\ &\quad + R(k)]^{-1}G^T(k)P(k+1|k), \end{aligned}$$

$$\begin{aligned} P(k+1|k) &= F(k)P(k|k)F^T(k) \\ &\quad + Q(k)G(k)G^T(k), \end{aligned}$$

where F, G, R, Q are appropriate matrices for the augmented state-parameter vector z , and θ_0 is a priori information about the parameter.

These methods provide sequential estimation of the state and the parameter for stationary and linear, or linearized systems yet they suffer immensely in real-time computations. Divergency is a common problem for linearized models [24], that is the success of estimation depends upon a very good initial point for the nominal trajectory and persistent corrections for very well-behaved systems.

CONCLUSION

The Kalman Filter is a recursive state estimator for a linear system, subject to a random noise. It has many technological applications in diverse engineering fields, which are implemented by digital systems. Thus the discrete version of the Kalman filter is much more common.

The Kalman filter has a sequence of two phases; it predicts the state of a linear system then compares it with the observations and adjusts the filter gain. Henceforth, this recursive instrument for estimating the states and refining the estimation process has been utilized in adaptive controllers, particularly in the linear-quadratic-Gaussian controllers, where a linear system is subject to an independent Gaussian random noise, and the controller minimizes a quadratic error function (performance measure). When the system structure has nonlinearities, the filter is not appropriate for state estimation. For those cases, an extended Kalman filter (EKF) is constructed, where a Taylor series expansion around the presently estimated state, or around a nominal state value, linearizes the system dynamics.

There are divergence problems in the Kalman filter implementations, and also actual computations suffer from the curse of the dimensionality. Thus, persistent adjustments are needed and the Kalman filter begs for clever tricks to expedite the filter gain computations.

REFERENCES

1. Astrom KJ. Introduction to stochastic control. New York: Academic Press; 1970.
2. Kumar PR, Varaiya P. Stochastic systems: estimation, identification and adaptive control. New Jersey: Prentice Hall; 1986.
3. Goodwin GC, Sin KS. In: Kailath T, editor. Adaptive filtering prediction and control: information and system science series. Englewood Cliffs (NJ): Prentice Hall; 1984.
4. Kalman R. A new approach to linear filtering and prediction problems. *J Basic Eng* 1960;82:35–45.
5. Kalman R. Canonical structure of linear dynamical systems. *Proc Natl Acad Sci USA* 1962;3:596–600.
6. Kolmogoroff A. Interpolation und extrapolation von stationaren zufalligen folgen. *Bull Acad Sci USSR Ser Math* 1941;5:3–14.
7. Wiener N. Extrapolation, interpolation, and smoothing of stationary time series with engineering applications. New York: MIT and Wiley; 1949.
8. Kalman R, Bucy R. New results in linear filtering and prediction theory. *Trans ASME J Basic Eng* 1961;83:95–108.
9. Wald A. Sequential analysis. New York: Wiley; 1947.
10. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22:400–425.
11. Box GEP, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. New Jersey: Prentice Hall; 1994.
12. Brockwell PJ, Davis RA. Introduction to time series and forecasting. New York: Springer; 2002.
13. Hamilton JD. Time series analysis. Princeton (NJ): Princeton University Press; 1994.
14. Aviv Y. A time-series framework for supply-chain inventory management. *Oper Res* 2003;51(2):210–227.
15. Meinhold RJ, Singpurwalla ND. Understanding the Kalman filter. *Am Stat* 1983;37(2):123–127.
16. West M, Harrison J. Bayesian forecasting and dynamic models. 2nd ed. New York: Springer; 1999.
17. Lebesgue H. Measure and the integral. San Fransisco (CA): Holden-Day; 1966.
18. Gauss CF. Theoria combinationis observationum erroribus minimis obnoxiae: translated by G.W. Steward as theory of the combination of observations least subject to error. Philadelphia (PA): SIAM; 1995.
19. Saridis G. Self-organizing control of stochastic systems. New York: Dekker; 1977.

20. Lainiotis D. Optimal adaptive estimation: Structure and parameter adaptation. IEEE Trans Automat Contr 1971;AC-16:160–170.
21. Gelb A. Applied optimal estimation. Massachusetts (MA): MIT press; 1974.
22. Ljung L. Asymptotic behaviour of the EKF as a parameter estimator. IEEE Trans Automat Control 1979;AC-24:36–50.
23. Gulcur HO, Akman G, Gursoy K. Extended Kalman filter design for time-varying system identification: an adaptive controller. IFAC Trans 1982;1:452–473.
24. Gursoy K. Time varying parameter estimation with extended Kalman filters for adaptive control applications [Master's thesis]. Turkey: METU; 1981.

FORMULATING GOOD MILP MODELS

JOHN N. HOOKER
Carnegie Mellon University,
Pittsburgh, Pennsylvania

Mixed integer/linear programming (MILP) provides a highly developed technology for solving a wide variety of optimization problems with continuous and discrete variables. To harness this technology, however, one must know how to formulate a given problem as an MILP model. In many cases, it is not enough to simply write a correct model. One must also write a “good” model, meaning that the formulation is amenable to fast solution. A poor formulation can require, by orders of magnitude, more solution time than a good one.

MILP problem formulation is generally regarded as an art rather than a science, but it need not be unprincipled. An underlying theory tells us precisely which problems can be given in an MILP formulation, and this in turn leads to useful guidelines for writing good models.

The guidelines are based on the fact that MILP formulations involve two modeling ideas: knapsack constraints and choices between discrete alternatives (disjunctions). By identifying these two elements in a given problem, one can obtain practical MILP formulations in a reasonably systematic way. The resulting models incorporate many of the “tricks” that one might otherwise learn only through familiarity with the folklore of modeling.

The characteristics of a good model have never been precisely defined, but in many cases, it is useful to have a “tight” continuous relaxation. The relaxation is important because its solution provides a bound on the optimal value of the original problem. A bound from a tight relaxation is closer to the optimal value and can accelerate solution,

because all general-purpose MILP solvers make essential use of bounds to prune the search tree.

BASIC DEFINITIONS

An MILP problem consists of linear inequality constraints and a linear objective function. At least some of the variables must take integer values, while the remaining variables are continuous. Such a problem can be written in the form

$$\begin{aligned} \min \text{ (or max) } & cx \\ Ax & \geq b \\ x & \in \mathbb{R}^n, \quad x_j \in \mathbb{Z} \quad \text{for } j \in J. \end{aligned} \quad \begin{matrix} (a) \\ (b) \end{matrix} \quad (1)$$

where $x = (x_1, \dots, x_n)$ and A is an $m \times n$ matrix. Variables x_j for $j \in J$ must take integer values. A value of x is *feasible* for Equation (1) if it satisfies the constraints (a) and (b), and the *feasible set* is the set of feasible values. The objective is to find a feasible x that minimizes (or maximizes) cx .

The *continuous relaxation* of Equation (1) is obtained by dropping the integrality condition on x_j for $j \in J$. The continuous relaxation is generally much easier to solve than Equation (1), because it is a linear programming problem. Its optimal value is a lower bound on the optimal value of Equation (1) if the objective is to minimize, and an upper bound if the objective is to maximize.

KNAPSACK MODELING

Mixed integer formulations frequently involve counting ideas that can be expressed as *knapsack inequalities*. For present purposes, we can define a knapsack inequality to be one of the form $\sum_j a_j x_j \leq \beta$, or $ax \leq \beta$ for short, where some (or all) of the variables x_j may be restricted to integer values. We also regard $ax \geq \beta$ as a knapsack inequality.

Knapsack Problems

The term *knapsack inequality* derives from the fact that the *integer knapsack problem* can be formulated with such an inequality. The problem is to pack a knapsack with items that have the greatest possible value, while not exceeding a maximum weight β . There are n types of items. Each item of type j has weight a_j and adds value c_j . If x_j is the number of items of type j put into the knapsack, the problem can be written as

$$\begin{aligned} \max \quad & cx \\ ax \leq & \beta \\ x_j \in \mathbb{Z}, \quad & \text{all } j. \end{aligned} \quad (2)$$

This is an MILP model because it has the form of Equation (1).

A wide variety of modeling situations involve this same basic idea. A classic example is the capital budgeting problem, in which the objective is to allocate a limited amount of capital to projects so as to maximize revenue. Here, β is the amount of capital available. There are n types of projects, and each project j has initial cost a_j and earns revenue c_j . Variable x_j represents the number of projects of type j that are funded.

Typically, an MILP model contains many knapsack constraints. There may also be pure linear constraints in which all the variables are continuous. This article focuses on modeling ideas that require integer variables, while the article on **An Introduction to Linear Programming** discusses purely linear modeling.

Example: Freight Transport

A product can be manufactured at several plants, from which it can be shipped to n customers (Fig. 1). Each customer j requires a quantity D_j of the product, and each plant i can manufacture at most C_i . The product must be transported in vehicles, which have capacity K . The cost of sending a vehicle along the route from factory i and to customer j is c_{ij} , regardless of how much cargo it carries. The objective is to assign vehicles to routes to minimize cost while meeting customer demand.

The movement of product from the plants to customers is described by a standard network flow model, which is purely linear. If x_{ij} is the quantity shipped from plant i to customer j , the constraints are

$$\begin{aligned} \sum_j x_{ij} &\leq C_i, \quad \text{all } i, \quad (a) \\ \sum_i x_{ij} &= D_j, \quad \text{all } j. \quad (b) \end{aligned} \quad (3)$$

Constraint (a) ensure that plant capacity is not exceeded, and constraint (b) meet customer demand. The vehicles assigned to each route must have enough capacity to carry the flow x_{ij} along that route. If y_{ij} is the number of vehicles assigned to the route from i to j , this condition is captured by the knapsack inequalities

$$Ky_{ij} \geq x_{ij}, \quad \text{all } i, j, \quad (4)$$

where y_{ij} takes integer values. These are knapsack inequalities because they can be written as $x_{ij} - Ky_{ij} \leq 0$. The objective is to minimize cost

$$\min \sum_{ij} c_{ij} y_{ij}. \quad (5)$$

The MILP model consists of Equations (3)–(5), plus the condition that $x_{ij} \geq 0$ and $y_{ij} \in \mathbb{Z}$ for all i, j .

A variation of this problem allows several types of vehicles to be used. A vehicle of type t has capacity K_t and costs c_{ijt} to operate on the route from i to j . Then if y_{ijt} is the number of vehicles of type t traveling from i to j , the knapsack constraint (4) becomes

$$\sum_t K_t y_{ijt} \geq x_{ij}, \quad \text{all } i, j$$

and the objective Equation (5) becomes

$$\min \sum_{ijt} c_{ijt} y_{ijt}.$$

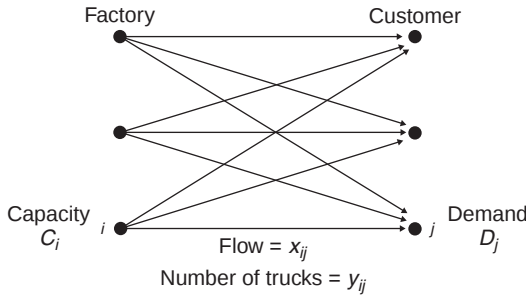


Figure 1. A freight transport problem.

Set Packing, Set Covering, and Set Partitioning

Special cases of knapsack inequalities occur in set packing, set covering, and set partitioning problems. The *set packing* problem begins with a collection of finite sets S_j for $j = 1, \dots, n$ that may partially overlap. It seeks a largest subcollection of sets that are pairwise disjoint.

Suppose, for example, that there are n surgeries to be performed, and the objective is to perform as many as possible this morning. Surgery j requires a specific set S_j of surgeons and other personnel. Because the surgeries must proceed in parallel, no two surgeries with overlapping personnel can be performed. This is a set packing problem.

The set packing problem can be formulated with 0–1 knapsack inequalities. Let $A_{ij} = 1$ when item i belongs to set S_j , and $A_{ij} = 0$ otherwise. Let variable $x_j = 1$ when set j is selected. The knapsack inequality $\sum_{j=1}^n A_{ij}x_j \leq 1$ prevents the selection of more than one set containing item i . Thus the system $Ax \leq e$ of knapsack inequalities, where e is a vector of ones, prevents the selection of any two overlapping sets. The objective is to maximize $\sum_{j=1}^n x_j$ subject to $Ax \leq e$ and $x \in \{0, 1\}^n$, which is an MILP problem.

The *set covering* problem likewise begins with a collection of sets S_j but seeks the minimum subcollection that contains all the elements in the union of the sets. For example, one may wish to buy a minimum collection of songbooks that contains all the songs that appear in at least one book. Here, S_j is the set of songs in book j .

If A_{ij} and x_j are as before, the knapsack inequality $\sum_{j=1}^n A_{ij}x_j \geq 1$ ensures that item i is covered. The set covering problem is to

minimize $\sum_{j=1}^n x_j$ subject to $Ax \geq e$ and $x \in \{0, 1\}^n$. The objective function in this or the set packing problem can be generalized to cx by attaching a weight c_j to each set S_j , to represent the cost or benefit of selecting S_j .

The *set partitioning* problem seeks a subcollection of sets such that each element is contained in exactly one of the sets selected. The constraints are therefore $Ax = e$, which are a combination of the knapsack constraints $Ax \leq e$ and $Ax \geq e$. The problem is to minimize or maximize cx subject to these constraints.

Example: Airline Crew Rostering

An important practical example of set partitioning is the airline crew rostering problem. Crews must be assigned to sequences of flight legs, while observing complicated work rules. For example, there are restrictions on the number of flight legs a crew may staff in one assignment, the total duration of the assignment, the layover time between flight legs, and the locations of the origin and destination.

Let S_j be a set of flight legs that can be assigned to a single crew, where j indexes all possible such sets. There may be millions of sets S_j and therefore millions of variables x_j . A set S_j is selected ($x_j = 1$) when it is assigned to a crew, incurring cost c_j . This is a partitioning problem because each flight leg must be staffed by exactly one crew and must therefore appear in exactly one selected S_j .

There are ways to formulate the problem with a much smaller model that explicitly represents the work rules with constraints. Yet, the set partitioning model has the advantage that it can be solved by *branch-and-price* methods. These methods solve

the continuous relaxation by *column generation*; that is, by adding variables (and the associated columns of A) to the problem only when they can improve the solution. Typically, only a tiny fraction of the columns are generated. This is an instance in which a large model is “better” than a small one in the sense that it permits faster solution.

Clique Inequalities

A collection of set packing constraints in a model can sometimes be replaced or supplemented by a *clique inequality*, which substantially tightens the relaxation. For example, the 0–1 inequalities

$$\begin{aligned} x_1 + x_2 &\leq 1, \\ x_1 + x_3 &\leq 1, \\ x_2 + x_3 &\leq 1. \end{aligned} \quad (6)$$

are equivalent to the clique inequality $x_1 + x_2 + x_3 \leq 1$. One can see that the clique inequality provides a tighter relaxation from the fact that $x_1 = x_2 = x_3 = 1/2$ violates it but satisfies Equation (6). It should therefore replace Equation (6) in the model.

To generalize this idea, define a graph whose vertices j correspond to 0–1 variables x_j . The graph contains an edge (i, j) whenever $x_i + x_j \leq 1$ is implied by a constraint in the model. If the induced subgraph on some subset C of vertices is a clique, then $\sum_{j \in C} x_j \leq 1$ is a valid inequality and can be added to the model.

Logical Conditions

Logical conditions on 0–1 variables can be formulated as knapsack inequalities that are similar to set covering constraints. Suppose, for example, that either plants 2 and 3 must be built, or else plant 1 must not be built. For the moment, regard x_j as a Boolean variable that is true when plant j is built, and false otherwise. The condition can be written as

$$\neg x_1 \vee x_2 \vee x_3, \quad (7)$$

where $\vee$ means “or” and $\neg$ means “not.” Such a condition is a *logical clause*, meaning that it is a disjunction of *literals* (Boolean variables or their negations). Because the clause states

that at least one of the literals must be true, it can be written as an inequality $(1 - x_1) + x_2 + x_3 \geq 1$ by viewing x_j as true when $x_j = 1$ and false when $x_j = 0$. This is equivalent to the knapsack inequality $-x_1 + x_2 + x_3 \geq 0$.

Any logical condition built from “and,” “or,” “not,” and “if” can be converted to a set of clauses and given an MILP model on that basis. One need to observe only three equivalences:

- $A \Rightarrow B$ (i.e., “ A implies B ,” or “ B if A ”) is equivalent to $\neg A \vee B$;
- $\neg(A \wedge B)$ is equivalent to $\neg A \vee \neg B$ (De Morgan’s Law), where $\wedge$ means “and;”
- $A \vee (B \wedge C)$ is equivalent to $(A \vee B) \wedge (A \vee C)$, a form of distribution.

Consider, for example, the condition, “If plants 1 and 2 are built, then plants 3 and 4 must be built.” It can be written as

$$(x_1 \wedge x_2) \Rightarrow (x_3 \wedge x_4).$$

The implication is first eliminated to obtain $\neg(x_1 \wedge x_2) \vee (x_3 \wedge x_4)$. The negation is then brought inside using De Morgan’s Law, resulting in the expression $\neg x_1 \vee \neg x_2 \vee (x_3 \wedge x_4)$. The conjunction is now distributed as

$$(\neg x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_4).$$

This conjunction of two clauses can be written as two knapsack constraints

$$\begin{aligned} -x_1 - x_2 + x_3 &\geq -1, \\ -x_1 - x_2 + x_4 &\geq -1. \end{aligned}$$

A model can sometimes be further tightened by generating *resolvents* of the clauses. Suppose, for example, that a model contains clauses $x_1 \vee x_2$, $\neg x_1 \vee \neg x_3$. Because exactly one variable (x_1) occurs in the two clauses with different signs, one can infer the resolvent $x_2 \vee \neg x_3$, which contains all the literals in either clause except x_1 and $\neg x_1$. The resolvent is converted to knapsack form and added to the MILP model. The process can be repeated as desired, and clauses implied by others (if any) can be dropped.

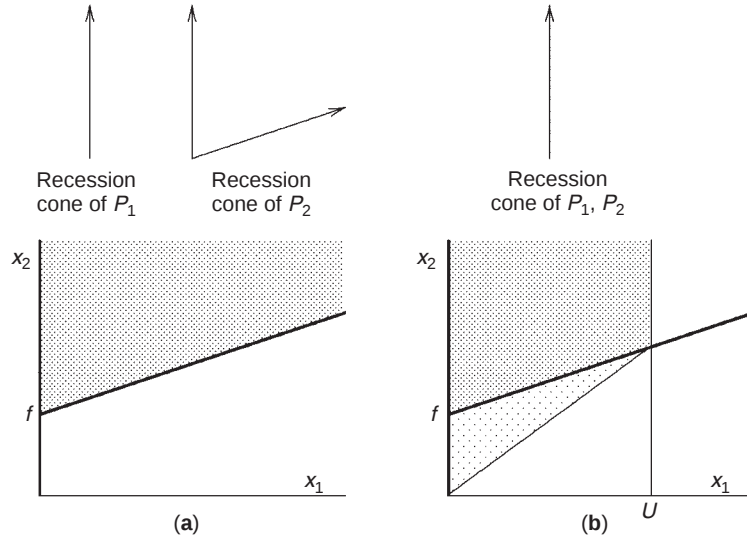


Figure 2. (a) Feasible set of a fixed-charge problem, consisting of the union of polyhedra P_1 (heavy vertical line) and P_2 (shaded area). (b) Feasible set $P_1 \cup P'_2$ of the same problem with the bound $x_1 \leq U$, where P'_2 is the darker shaded area. The convex hull of the feasible set is the entire shaded area.

MILP REPRESENTABILITY

Knapsack-based modeling is useful but harnesses only a fraction of MILP modeling capability. The full resources of MILP modeling appear when one introduces auxiliary 0–1 variables to represent discrete alternatives. A choice among alternatives can be represented by a disjunction in which each term describes one of the alternatives. If each term is a system of knapsack inequalities, the problem has an MILP model, provided the systems satisfy a fairly innocuous technical condition. In fact, the problem has an MILP model *if and only if* it can be written as such a disjunction of knapsack systems.

Example: Fixed-Charge Function

A simple fixed-charge problem illustrates disjunctive modeling. Suppose the cost x_2 of manufacturing quantity x_1 of some product is to be minimized. The cost is zero when $x_1 = 0$ and is $f + cx_1$ otherwise, where f is the fixed cost and c the unit variable cost.

The problem can be viewed as minimizing x_2 subject to $(x_1, x_2) \in S$, where S is the set

depicted in Fig. 2a. There are two discrete alternatives: the cost is zero or the cost is a fixed cost plus a variable cost. They are captured by the disjunctive formulation

$$\begin{aligned} \min x_2 \\ \left(\begin{array}{l} x_2 \geq 0 \\ x_1 = 0 \end{array} \right) \vee \left(\begin{array}{l} x_2 \geq cx_1 + f \\ x_1 \geq 0 \end{array} \right). \end{aligned} \quad (8)$$

The disjuncts define two polyhedra P_1 and P_2 , respectively, illustrated in the figure. In general, there would be additional constraints in the problem, but the focus here is on the fixed-charge formulation.

It is impossible to write an MILP model for the disjunction in Equation (8) because it fails the technical condition mentioned earlier. The condition is that the polyhedra defined by the disjunctions (ignoring the integrality condition) must have the same *recession cone*. The recession cone is the set of directions in which one can go forever without leaving the polyhedron. That is, the recession cone of a polyhedron P is the set of vectors r such that for any $x \in P$, $x + \alpha r \in P$ for all $\alpha \geq 0$. If the polyhedron is bounded, its recession cone is the origin.

The recession cone of P_1 in the example is P_1 itself (Fig. 2a), and the recession cone of P_2 is the set of all vectors (x_1, x_2) with $x_2 \geq cx_1 \geq 0$. However, if an upper bound U is placed on x_1 , the problem becomes

$$\begin{aligned} \min x_2 \\ \left(\begin{array}{l} x_1 = 0 \\ x_2 \geq 0 \end{array} \right) \vee \left(\begin{array}{l} x_2 \geq cx_1 + f \\ 0 \leq x_1 \leq U \end{array} \right). \end{aligned} \quad (9)$$

The recession cone of each of the resulting polyhedra P_1, P'_2 (Fig. 2b) is the same (namely, P_1), and the problem is therefore MILP representable. In fact, it will be shown that the problem has the following MILP model:

$$\begin{aligned} \min x_2 \\ 0 \leq x_1 \leq U\delta, \quad x_2 \geq f\delta + cx_1, \quad \delta \in \{0, 1\}. \end{aligned} \quad (10)$$

which can be simplified to

$$\begin{aligned} \min f\delta + cx_1 \\ 0 \leq x_1 \leq U\delta, \quad \delta \in \{0, 1\} \end{aligned} \quad (11)$$

The 0–1 variable δ encodes whether the quantity produced is zero or positive, in the former case ($\delta = 0$) forcing $x_1 = 0$, and in the latter case, incurring the fixed charge f .

The Representability Theorem

A problem is *MILP representable* if there is some MILP model for it. To make this more precise, suppose one wishes to model an optimization problem $\mathcal{O}$, which can in general be conceived as the problem of minimizing some variable x_j subject to $x \in S$, where $x = (x_1, \dots, x_n)$. Any objective function $f(x)$ can be accommodated by adding variable x_j and constraint $x_j \geq f(x)$ to S , and minimizing or maximizing x_j .

Consider an MILP model $\mathcal{M}$ that minimizes x_j subject to constraints that contain the variables $x_1, \dots, x_n$ (some of which may be restricted to integers), possibly along with auxiliary continuous variables $y_1, \dots, y_m$ and 0–1 variables $\delta_1, \dots, \delta_k$. Model $\mathcal{M}$ represents problem $\mathcal{O}$ if the projection of $\mathcal{M}$'s feasible set onto x is S ; that is, $x \in S$ if and only if (x, y, δ) is feasible in $\mathcal{M}$ for some y, δ . This

means that one can solve $\mathcal{O}$ by solving $\mathcal{M}$, because if (x, y, δ) is an optimal solution of $\mathcal{M}$, then x is an optimal solution of $\mathcal{O}$.

MILP representability is completely characterized by the following theorem, which is proved in Refs 1 and 2. (Technically, the proof assumes that A, c consist of rational data.)

Theorem 1. *An optimization problem that minimizes x_j subject to $x \in S$ is MILP representable if and only if S is the feasible set of some constraint of the form*

$$\begin{aligned} \bigvee_{k \in K} (A^k x \geq b^k), \\ x \in \mathbb{R}^n, \quad x_j \in \mathbb{Z} \quad \text{for } j \in J, \end{aligned} \quad (12)$$

where K is finite and the polyhedra $\{x \mid A^k x \geq b^k\}$ for $k \in K$ have the same recession cone.

Thus, an optimization problem has an MILP model if and only if its feasible set can be described by a disjunction of knapsack systems that define polyhedra with the same recession cone.

If it is difficult to identify the recession cones, one can always ensure that they are the same by placing fixed lower and upper bounds on all the variables—bounds wide enough so as not to be violated by any reasonable solution. This results in bounded polyhedra for every disjunct, so that every recession cone is the origin.

Conversion to an MILP Model

There remains the question as to how to convert the disjunctive problem Equation (12) into an MILP model. There are two standard ways: a big- M formulation and a convex hull formulation. Both require 0–1 auxiliary variables. The convex hull formulation is generally tighter, but at the cost of introducing continuous auxiliary variables as well. First, the big- M formulation.

Theorem 2. *If an optimization problem satisfies the conditions of Theorem 1, it is represented by the big- M formulation*

$$\begin{aligned}
 & \min cx \\
 & A^k x \geq b^k - M^k(1 - \delta_k), \quad k \in K, \\
 & x \in \mathbb{R}^n, \quad x_j \in \mathbb{Z} \quad \text{for } j \in J, \quad \delta_k \in \{0, 1\} \\
 & \text{for } k \in K,
 \end{aligned} \tag{13}$$

where

$$\begin{aligned}
 M^k = b^k - \min_{\ell \neq k} \left\{ \min_x \left\{ A^\ell x \mid A^\ell x \geq b^\ell, x \in \mathbb{R}^n, \right. \right. \\
 \left. \left. x_j \in \mathbb{Z} \quad \text{for } j \in J \right\} \right\}.
 \end{aligned} \tag{14}$$

When auxiliary variable $\delta_k = 1$, the k th linear system $A^k x \geq b^k$ is enforced. When $\delta_k = 0$, the k th system is deactivated by subtracting a vector M^k of large numbers from the right-hand side. The components of M^k are chosen in Equation (14) to be as small as possible, so as to result in a tighter relaxation, but larger values yield a correct model.

The *convex hull MILP formulation* introduces continuous auxiliary variables by disaggregating x into a sum $\sum_{k \in K} x^k$, where each x^k corresponds to the k th disjunct.

Theorem 3. *If an optimization problem $\mathcal{O}$ satisfies the conditions of Theorem 1, it is represented by the formulation*

$$\begin{aligned}
 & \min cx \\
 & x = \sum_{k \in K} x^k A^k x^k \geq b^k \delta_k, \quad k \in K, \\
 & \sum_{k \in K} \delta_k = 1 \quad x \in \mathbb{R}^n, \quad x_j \in \mathbb{Z} \quad \text{for } j \in J, \\
 & \delta_k \in \{0, 1\} \quad \text{for } k \in K.
 \end{aligned} \tag{15}$$

The convex hull formulation (Eq. 15) has the advantage that it provides the tightest possible linear relaxation of $\mathcal{O}$, namely a *convex hull relaxation*. That is, its continuous relaxation describes a set that, when projected onto x , is the closure of the convex hull of $\mathcal{O}$'s feasible set. This assumes that the individual systems $A^k x \geq b^k$ provide convex hull relaxations—as they do, for example, when all the x_j s are continuous.

The model represented by Equation (15) continues to provide a valid convex hull

relaxation of $\mathcal{O}$ when the recession cone condition is not satisfied, even though Equation (15) does not correctly represent $\mathcal{O}$ in this case. In addition, the model (15) is *locally ideal*, meaning that the 0–1 variables δ_k take integer values at every extreme point of the polyhedron described by the model's linear relaxation.

The big- M formulation (Eq. 13) likewise provides a valid relaxation of $\mathcal{O}$ even when the polyhedra do not have the same recession cone, although not in general a convex hull relaxation. The big- M formulation need not be locally ideal.

It is hard to say, in general, when a big- M formulation is preferable to a convex hull formulation. It tends to be more attractive when there are a large number of disjuncts, because the convex hull model introduces a vector x^k of auxiliary continuous variables for every disjunct. In some cases, the big- M model provides a convex hull relaxation. On the other hand, the convex hull model may simplify, and some variables may drop out.

In practice, a problem typically has several disjunctive constraints. (A system of knapsack constraints can be regarded as a disjunction with a single disjunct.) The MILP formulations of the disjunctions are pooled to form an MILP model. Note that while convex hull formulations provide convex hull relaxations of the individual disjunctions, when taken together they do not necessarily provide a convex hull relaxation of the problem as a whole.

Modeling the Fixed-Charge Function

The fixed-charge problem can be modeled by writing a big- M or a convex hull formulation of the disjunctive problem (Eq. 8). In this case, the two formulations turn out to be identical.

The big- M formulation (Eq. 13) of Equation (8) is

$$\begin{aligned}
 0 &\leq x \leq M_1^1 \delta & 0 &\leq x \leq U + M_1^2(1 - \delta), \\
 z &\geq -M_2^1 \delta & z &\geq cx + f - M_2^2(1 - \delta), \\
 \delta &\in \{0, 1\},
 \end{aligned} \tag{16}$$

where δ_1, δ_2 are written as $\delta, 1 - \delta$ because they sum to one. From Equation (14), $(M_1^1, M_2^1, M_1^2, M_2^2) = (U, -f, 0, f)$. So Equation (16) simplifies to Equation (11).

The convex hull formulation (Eq. 15) of Equation (8) is

$$\begin{aligned} x &= x^1 + x^2 & z &= z^1 + z^2, \\ x^1 &= 0 & 0 &\leq x^2 \leq U\delta, \\ z^1 &\geq 0 & z^2 &\geq cx^2 + f\delta, \\ \delta &\in \{0, 1\}. \end{aligned}$$

Because δ_1 does not appear in the model for the first disjunct, δ_2 is written as δ . The fact that $x^1 = 0$ means that x can replace x^1 . Also z^1 can be removed by writing $z \geq z^2$. Because z_2 is minimized, z can now replace z_2 , and the model again simplifies to Equation (11). The feasible set of Equation (11)'s continuous relaxation projects onto the convex hull of the feasible set in x_1, x_2 -space, as illustrated in Fig. 2b.

EXAMPLES OF MILP MODELING

The disjunctive modeling approach may seem a roundabout way to obtain a fairly obvious model for the fixed-charge function. However, it provides a general tool for constructing good formulations that are much less obvious for more complex problems. This is illustrated by the following examples.

Capacitated Facility Location

The transportation problem described earlier assumes that factories already exist at all of the factory sites. The problem can be modified to address a location as well as a routing issue, at which sites should factories be built?

The problem poses two discrete alternatives for each site i . Either a factory is built or it is not. If it is built, the total shipments out of the location must be at most C_i , and a fixed cost f_i is incurred. Otherwise nothing is shipped out of the location. Thus if x_{ij} is again the quantity transported from i to j , the situation at site i is represented by a disjunction of two knapsack systems

$$\left(\begin{aligned} \sum_{j=1}^n x_{ij} &\leq C_i \\ 0 \leq x_{ij} &\leq K_{ij}y_{ij}, \text{ all } j \\ z_i &= f_i \\ y_{ij} &\in \mathbb{Z}, \text{ all } j \end{aligned} \right) \vee \left(\begin{aligned} x_{ij} &= 0, \text{ all } j \\ z_i &= 0 \end{aligned} \right), \quad (17)$$

where variable z_i represents the fixed cost incurred. Again, each customer j must receive adequate supply

$$\sum_{i=1}^m x_{ij} = D_j, \text{ all } j. \quad (18)$$

This can be viewed as a disjunction with one disjunct. The problem is to minimize

$$\sum_{i=1}^m \left(z_i + \sum_{j=1}^n c_{ij}x_{ij} \right) \quad (19)$$

subject to Equation (18), and subject to Equation (17) for all i .

The polyhedra defined by the two disjuncts in Equation (17) have different recession cones. The cone for the first polyhedron is $\{(x_i, y_i) \mid x_i = 0, y_i \geq 0\}$, where $x_i = (x_{i1}, \dots, x_{in})$ and $y_i = (y_{i1}, \dots, y_{in})$, while the cone for the second is $\{(x_i, y_i) \mid x_i = 0\}$. However, if we add the innocuous constraint $y_i \geq 0$ to the second disjunct, the two disjuncts have the same recession cone and can be given a convex hull formulation

$$\begin{aligned} \sum_{j=1}^n x_{ij} &\leq C_i\delta_i, \\ 0 \leq x_{ij} &\leq K_{ij}y_{ij}, \text{ all } j \\ z_i &= f_i\delta_i, \quad \delta_i \in \{0, 1\}, \quad y_{ij} \in \mathbb{Z}, \text{ all } j \end{aligned} \quad (20)$$

The resulting MILP model for the entire problem is

$$\min \sum_{i=1}^m \left(f_i\delta_i + \sum_{j=1}^n c_{ij}y_{ij} \right),$$

$$\begin{aligned}
\sum_{j=1}^n x_{ij} &\leq C_i \delta_i, \quad \text{all } i, \\
0 \leq x_{ij} &\leq K_{ij} y_{ij}, \quad \text{all } i, j, \\
\sum_{i=1}^m x_{ij} &= D_j, \quad \text{all } j, \\
\delta_i &\in \{0, 1\}, \quad y_{ij} \in \mathbb{Z}, \quad \text{all } i, j.
\end{aligned} \tag{21}$$

If the big- M formulation (Eq. 13) of the disjunctions is used, rather than the convex hull formulation, the same MILP model results.

Uncapacitated Facility Location

An uncapacitated location problem shows how a disjunctive approach can lead to a tighter relaxation than one might obtain otherwise. The problem is the same as the previous one, except that there is no limit on the capacity of each facility.

A beginner's mistake is to model this as a special case of the capacitated problem. Although there is no capacity limit, one can observe that each factory i will ship at most $\sum_j D_j$ units and therefore let $C_i = \sum_j D_j$ in the formulation (21) for the capacitated problem. This is a valid formulation of the uncapacitated problem, but there is a much tighter one.

Let us begin again with a choice of alternative. If factory i is installed, it supplies at most D_j to each customer j and incurs cost f_i . If it is not installed, then it supplies nothing

$$\left(\begin{array}{l} 0 \leq x_{ij} \leq D_j, \quad \text{all } j \\ 0 \leq x_{ij} \leq K_{ij} y_{ij}, \quad \text{all } j \\ z_i = f_i \\ y_{ij} \in \mathbb{Z}, \quad \text{all } j \end{array} \right) \vee \left(\begin{array}{l} x_{ij} = 0, \quad \text{all } j \\ z_i = 0 \end{array} \right).$$

The convex hull formulation of this disjunction is

$$\begin{aligned}
x_{ij} &\leq D_j \delta_i, \quad \text{all } j, \\
0 \leq x_{ij} &\leq K_{ij} y_{ij}, \quad \text{all } j, \\
z_i &= f_i \delta_i, \quad \delta_i \in \{0, 1\}, \quad y_{ij} \in \mathbb{Z}, \quad \text{all } j.
\end{aligned} \tag{22}$$

This is a tighter MILP formulation than Equation (20) with $C_i = \sum_j D_j$. To see this, note that setting $\delta_i = 1/2$ can satisfy Equation (20) when $x_{ij} > D_j/2$ for some j ,

provided $x_{ik} < D_k/2$ for some $k \neq j$, whereas this solution cannot satisfy Equation (22).

A rule of thumb for MILP modeling is that a more disaggregated model like Equation (22) is tighter. Yet, Equation (22) results simply from using all available information when formulating each disjunct.

Lot Sizing with Setup Costs

A lot sizing problem with setup costs illustrates how logical relations among linear systems can be captured with logical constraints.

There is a demand D_t for a product in each period t . No more than K_t units of the product can be manufactured in period t , and any excess over demand is stocked to satisfy future demand. If there is no production in the previous period, then a setup cost of f_t is incurred. The unit production cost is p_t , and the unit holding cost per period is h_t . A starting stock level s_0 is given. The objective is to choose production levels in each period so as to minimize total cost over all periods.

Let x_t be the production level in period t and s_t the stock level at the end of the period. In each period t , there are three options to choose from: (i) start producing (with a setup cost), (ii) continue producing (with no setup cost), and (iii) produce nothing. If v_t is the setup cost incurred in period t , these correspond respectively to the three disjuncts

$$\left(\begin{array}{l} v_t \geq f_t \\ 0 \leq x_t \leq K_t \end{array} \right) \vee \left(\begin{array}{l} v_t \geq 0 \\ 0 \leq x_t \leq K_t \end{array} \right) \vee \left(\begin{array}{l} v_t \geq 0 \\ x_t = 0 \end{array} \right). \tag{23}$$

There are logical connections between the choices in consecutive periods. If we schematically represent the disjunction Equation (23) as $A_t \vee B_t \vee C_t$, the logical connections can be written as

$$\begin{aligned}
B_t &\Rightarrow (A_{t-1} \vee B_{t-1}), \\
A_t &\Rightarrow (\neg A_{t-1} \wedge \neg B_{t-1}).
\end{aligned} \tag{24}$$

The inventory balance constraints are

$$\begin{aligned}
s_{t-1} + x_t &= D_t + s_t, \quad s_t \geq 0, \\
t &= 1, \dots, n,
\end{aligned} \tag{25}$$

where s_t is the stock level in period t and s_0 is given. The problem is to minimize

$$\sum_{t=1}^n (p_t x_t + h_t s_t + v_t) \quad (26)$$

subject to Equations (23) and (24) for all $t \geq 1$ and Equation (25).

A convex hull formulation for Equation (23) is

$$\begin{aligned} v_t^1 &\geq f_t \alpha_t, & v_t^2 &\geq 0, & v_t^3 &\geq 0, \\ 0 &\leq x_t^1 \leq K_t \alpha_t, & 0 &\leq x_t^2 \leq K_t \beta_t, & x_t^3 &= 0, \\ v_t &= v_t^1 + v_t^2 + v_t^3, & x_t &= x_t^1 + x_t^2 + x_t^3, \\ \alpha_t + \beta_t + \gamma_t &= 1, & \alpha_t, \beta_t, \gamma_t &\in \{0, 1\}. \end{aligned} \quad (27)$$

Thus, $\alpha_t = 1$ indicates a start-up, $\beta_t = 1$ continued production, and $\gamma_t = 1$ no production in period t . To simplify Equation (27), first eliminate γ_t , so that $\alpha_t + \beta_t \leq 1$. Because $x_t^3 = 0$, one can set $x_t = x_t^1 + x_t^2$, which allows one to replace the two capacity constraints in Equation (27) by $0 \leq x_t \leq K_t(\alpha_t + \beta_t)$. Finally, v_t can replace v_t^1 because v_t is minimized. The convex hull formulation (27) becomes

$$\begin{aligned} v_t &\geq f_t \alpha_t, & 0 &\leq x_t \leq K_t(\alpha_t + \beta_t), & (a) \\ \alpha_t + \beta_t &\leq 1, & & & (b) \end{aligned} \quad (28)$$

where $\alpha_t, \beta_t \in \{0, 1\}$. The logical constraints (Eq. 24) can be formulated as

$$\begin{aligned} \alpha_{t-1} + \beta_{t-1} - \beta_t &\geq 0, & (a) \\ -\alpha_{t-1} - \alpha_t &\geq -1, & -\beta_{t-1} - \alpha_t \geq -1. & (b) \end{aligned} \quad (29)$$

The set packing inequalities in Equation (29b) and $\alpha_{t-1} + \beta_{t-1} \leq 1$ from Equation (28b) can be replaced by the clique inequality

$$\alpha_{t-1} + \beta_{t-1} + \alpha_t \leq 1. \quad (30)$$

The entire problem can now be formulated as minimizing Equation (26) subject to Equation (25) along with Equations (28a), (29a), (30), and $\alpha_t, \beta_t \in \{0, 1\}$ for all $t \geq 1$.

Package Delivery

A final example, adapted from Refs 3 and 4, illustrates how a principled modeling approach incorporates two modeling tricks that one might otherwise miss. A collection of packages are to be delivered by several trucks, and each package j has size a_j . Each available truck i has capacity Q_i and costs c_i to operate. The problem is to decide which trucks to use, and which packages to load on each truck, to deliver all the items at minimum cost.

The decision problem consists of two levels: the choice of which trucks to use, followed by the choice of which packages to load on each truck. The trucks selected must provide sufficient capacity, which leads naturally to a 0–1 knapsack constraint

$$\sum_i Q_i \delta_i \geq \sum_j a_j, \quad (31)$$

where each $\delta_i \in \{0, 1\}$ and $\delta_i = 1$ when truck i is selected.

The secondary choice of which packages to load on truck i depends on whether that truck is selected. If the truck i is selected, then a cost c_i is incurred, and the items loaded must fit into the truck (a 0–1 knapsack constraint). If truck i is not selected, then no items can be loaded. The disjunction is

$$\left(\begin{array}{l} z_i \geq c_i \\ \sum_j a_j x_{ij} \leq Q_i \\ 0 \leq x_{ij} \leq 1, \text{ all } j \\ x_{ij} \in \mathbb{Z}, \text{ all } j \end{array} \right) \vee (x_{ij} = 0, \text{ all } j), \quad (32)$$

where z_i is the fixed cost incurred by truck i , and $x_{ij} = 1$ when package j is loaded into truck i . The associated polyhedra have the same recession cone if one adds $z_i \geq 0$ to the second disjunct. Because $\delta_i = 1$ when truck i is selected, the convex hull formulation of Equation (32) is

$$\begin{aligned} z_i &\geq c_i \delta_i, & (a) \\ \sum_j a_j x_{ij} &\leq Q_i \delta_i, & (b) \\ x_{ij} &\leq \delta_i, \text{ all } j, & (c) \\ \delta_i, x_{ij} &\in \{0, 1\}, \text{ all } j. \end{aligned} \quad (33)$$

Finally, set covering inequalities ensure that each package is shipped

$$\sum_i x_{ij} \geq 1, \quad x_{ij} \in \{0, 1\}, \quad \text{all } j. \quad (34)$$

One can now minimize total fixed cost $\sum_i z_i$ subject to Equations (31) and (34), as well as Equation (33) for all i .

This formulation differs in two ways from a formulation that one might initially write for this problem. One might omit the constraint (33c) because the model is correct without it. Yet this constraint makes the relaxation tighter. Also, one might not include constraint (31), because due to Equation (34), it is the sum of constraints (33b) over all i . Yet the presence of Equation (31) allows the solver to deduce lifted knapsack cuts, which create a tighter continuous relaxation. This results in much faster solution.

PIECEWISE LINEAR FUNCTIONS

Piecewise linear functions are a very useful modeling tool because they can be used to approximate separable nonlinear functions in an MILP model. Typically, one wishes to model a function of the form $g(x) = \sum_j g_j(x_j)$ by approximating each nonlinear term $g_j(x_j)$ with a piecewise linear function $f_j(x_j)$. The function $f_j(x_j)$ is set to a value equal to (or close to) $g_j(x_j)$ at a finite number of values of x_j and is defined between these points by a linear interpolation.

It is convenient to drop the subscripts and refer to $f_j(x_j)$ as $f(x)$. For the sake of generality, suppose that $f(x)$ is piecewise linear in the sense that it is linear on possibly disjoint intervals $[a_i, b_i]$ and undefined outside these intervals, as illustrated in Fig. 3. More precisely, $x \in \bigcup_{i=1}^k [a_i, b_i]$ and

$$f(x) = \begin{cases} f(a_i) + \frac{x - a_i}{b_i - a_i} [f(b_i) - f(a_i)] & \text{if } x \in [a_i, b_i] \text{ and } a_i < b_i, \\ f(a_i) & \text{if } x = a_i = b_i. \end{cases} \quad (35)$$

In many applications, each $b_i = a_{i+1}$, which means $f(x)$ is continuous.

A disjunctive formulation is natural and convenient for a function of this form

$$\bigvee_i \left(\begin{array}{l} x = \lambda a_i + \mu b_i \\ z = \lambda f(a_i) + \mu f(b_i) \\ \lambda + \mu = 1, \quad \lambda, \mu \geq 0 \end{array} \right), \quad (36)$$

where disjunct i corresponds to $x \in [a_i, b_i]$, and z is a variable that represents $f(x)$. Because the disjuncts define polyhedra with the same recession cone (all the polyhedra are bounded), Equation (36) has a locally ideal, convex hull formulation

$$\begin{aligned} x &= \sum_i \lambda_i a_i + \mu_i b_i \\ z &= \sum_i \lambda_i f(a_i) + \mu_i f(b_i) \\ \lambda_i + \mu_i &= \delta_i, \quad \text{all } i \\ \sum_i \delta_i &= 1 \\ \lambda_i, \mu_i &\geq 0, \quad \delta_i \in \{0, 1\}, \quad \text{all } i \end{aligned} \quad (37)$$

This is similar to a well-known textbook model that dispenses with the multipliers μ_i but applies only when $f(x)$ is continuous

$$\begin{aligned} x &= \sum_{i=1}^{k+1} \lambda_i a_i, \quad z = \sum_{i=1}^{k+1} \lambda_i f(a_i), \quad \sum_{i=1}^{k+1} \lambda_i = 1 \quad (a) \\ \lambda_i &\leq \delta_{i-1} + \delta_i, \quad i = 2, \dots, k \quad (b) \\ \lambda_1 &\leq \delta_1, \quad \lambda_{k+1} \leq \delta_k, \quad \sum_{i=1}^k \delta_i = 1 \quad (c) \\ \lambda_i &\geq 0, \quad i = 1, \dots, k+1; \quad \delta_i \in \{0, 1\}, i = 1, \dots, k. \end{aligned} \quad (38)$$

where $a_{k+1} = b_k$. The model is designed to require that at most two λ_i s can be positive, and if two are positive they must be adjacent. However, this model is not as tight as Equation (37) and is not locally ideal [5]. For continuous functions $f(x)$, one can use the *incremental cost model*, which contains no more variables than Equation (38) but is equivalent to the tight model Equation (37) and locally ideal [6]

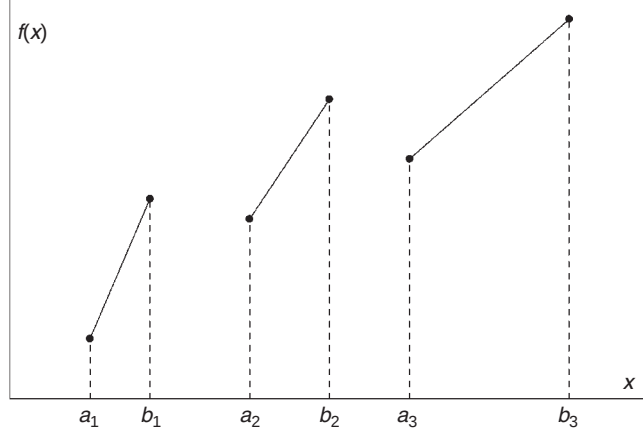


Figure 3. A piecewise linear function.

$$\begin{aligned}
 x &= a_1 + \sum_{i=2}^{k+1} x_i, \quad z = f(a_1) + \sum_{i=2}^{k+1} \frac{f'_i}{a'_i} x_i \\
 0 &\leq x_i \leq a'_i, \quad i = 2, \dots, k+1 \\
 a'_i \delta_i &\leq x_i \leq a'_i \delta_{i-1}, \quad i = 3, \dots, k \\
 a'_2 \delta_2 &\leq x_2 \leq a'_2, \quad 0 \leq x_{k+1} \leq a'_{k+1} \delta_k \\
 \delta_i &\in \{0, 1\}, \quad i = 2, \dots, k.
 \end{aligned} \tag{39}$$

Here, $a'_i = a_i - a_{i-1}$ and $f'_i = f(a_i) - f(a_{i-1})$.

There are recent proposals for modeling piecewise linear functions with a logarithmic number of 0–1 variables [7–9].

SYMMETRY

Symmetry occurs in an MILP model when values or variables can be exchanged without affecting the problem in any significant way. Symmetries can greatly retard solution of the problem, because the solver may waste time enumerating a large number of equivalent solutions. Symmetries can sometimes be broken by adding constraints to the model, or by reformulating it completely. There are also techniques that deal with symmetries during the solution process.

Value symmetry occurs when the values can be interchanged in a solution without affecting the feasibility or optimality of that solution. For example, any two colors in a graph coloring can be swapped. *Variable symmetry* occurs when variables can be

interchanged in a model without affecting the set of feasible or optimal solutions.

Symmetry-Breaking Constraints

Variable symmetry can often be avoided by adding constraints that prioritize symmetric variables. This is illustrated by the package delivery problem described earlier. Suppose that there are four trucks, with capacities 80, 80, 80, and 100, respectively. There are seven packages, having sizes 30, 30, 30, 40, 40, 50, and 50.

Because trucks 1, 2, and 3 are identical, the variables δ_1, δ_2 , and δ_3 can be interchanged with no effect on the problem. This symmetry can be broken by requiring that truck 1 be used first, followed by truck 2, and then truck 3. This is accomplished with the symmetry-breaking constraint $\delta_1 \geq \delta_2 \geq \delta_3$.

Symmetry-Breaking Reformulation

Because some packages are the same size in the example, they give rise to another variable symmetry. Packages 1, 2, and 3 have the same size, which means that variables x_{i1}, x_{i2} , and x_{i3} can be interchanged for any i , and similarly for the other package sizes. These symmetries can be removed by redefining variables. Let y_{ik} be the number of packages of size k loaded on truck i . Then, the problem is to minimize $\sum_i c_i \delta_i$ subject to Equation (31) and

$$\sum_k \bar{a}_k y_{ik} \leq Q_i \delta_i, \quad \delta_i \in \{0, 1\}, \text{ all } i;$$

$$y_{ik} \geq 0, y_{ij} \in \mathbb{Z}, \text{ all } i, k$$

plus the symmetry-breaking constraints for trucks, where $\bar{a}_k$ is the space consumed by packages of size k .

An alternate approach is to remove both truck and package symmetries by formulating the problem in a manner similar to the crew rostering problem described earlier. A truck of size 80 can be loaded in four ways: 30 + 30, 30 + 40, 30 + 50, and 40 + 40, which can be indexed $\ell = 1, \dots, 4$. A truck of size 100 can be loaded in six ways: 30 + 30 + 30, 30 + 30 + 40, 30 + 50, 40 + 40, 40 + 50, and 50 + 50. Incomplete loadings need not be enumerated. Let $d_{t\ell}$ be a vector whose k th component is the number of packages of size k used in loading pattern ℓ on a truck of type t ($t = 1, 2$). Then the four loading patterns for truck type 1 (capacity 80) are encoded by $d_{11}, \dots, d_{14}$, which are the columns on the left below, while the loading patterns for truck type 2 (capacity 100) appear on the right

2	1	1	0	3	2	1	0	0	0
0	1	0	2	0	1	0	1	2	0
0	0	1	0	0	0	1	1	0	2

The rows correspond to packages of sizes 30, 40, and 50. If variable $z_{t\ell}$ is the number of trucks of type t loaded with pattern ℓ , the MILP model is

$$\min \sum_t \bar{c}_t \sum_\ell z_{t\ell},$$

$$\sum_{t\ell} d_{t\ell} z_{t\ell} \geq b; \quad z_{t\ell} \geq 0, z_{t\ell} \in \mathbb{Z}, \text{ all } t, \ell,$$

where, $\bar{c}_t$ is the capacity of truck type t . The requirements vector b contains the number of packages of each type. In the example, $b = (3, 2, 2)$.

There may be a very large number of loading patterns, but as in the case of crew rostering, the problem can be solved by branch-and-price method.

EXTENSIONS

MILP solvers typically supply features that go beyond MILP modeling and allow more efficient solution. The most common among them include special ordered sets, indicator variables, and semicontinuous variables.

Special Ordered Sets

A *special ordered set of type 1* (SOS1) is a variable set $S = \{x_1, \dots, x_k\}$ for which exactly one variable is allowed to be positive, and the rest must be zero. SOS1 is most commonly used along with a set partitioning constraint $\sum_{j=1}^k x_j = 1$. If the model declares S to be SOS1, there is no need to restrict the variables in the set partitioning constraint to be integer. The solver enforces integrality through a efficient branching scheme that creates a branch for each $x_j \in S$ by fixing all the other variables in S to zero.

Variable set S is a *special ordered set of type 2* (SOS2) when exactly two of the variables are allowed to be positive, the rest must be zero, and any two positive variables must be adjacent (i.e., x_j and x_{j+1} for some j). SOS2 is designed to simplify the model (Eq. 38) of a piecewise linear function. By declaring $\{\lambda_1, \dots, \lambda_{k+1}\}$ to be SOS2, only constraints (a) of Equation (38) need be included in the model. SOS2 is also implemented by an efficient branching scheme, but it sacrifices the convex hull relaxation provided by the tighter MILP models (37) and (39). In addition, it is unsuitable for a discontinuous piecewise linear function like that in Fig. 3. Further options are discussed in Ref. 10.

Indicator Constraints and Semicontinuous Variables

Indicator constraints are constraints that are enforced only when a 0–1 variable is equal to 1. Thus if one wishes to enforce $ax \geq \beta$ only when $\delta = 1$, the pair $(\delta, ax \geq \beta)$ is declared an indicator constraint. This device can obviously be encoded with MILP constraints. The indicator constraint obviates the encoding but sacrifices the continuous relaxation it provides.

A *semicontinuous* variable x is a related idea in which x is forced to zero when $\delta = 0$

and to be within bounds $L \leq x \leq U$, when $\delta = 1$. The MILP encoding is similar to that of a discontinuous piecewise linear function. No encoding is necessary if x is declared a semicontinuous variable.

FURTHER READING

General guidance on MILP modeling can be found in Refs 1, and 11–13. Modeling of logical conditions is further explained in Refs 14 and 15. The traveling salesman problem presents an interesting challenge for MILP modeling, and several formulations are discussed in Ref. 16. A recent trend is to integrate MILP modeling with constraint programming models, because this provides a richer modeling framework that can allow the solver to exploit special structure in the problem [14,17].

REFERENCES

1. Hooker JN. A principled approach to mixed integer/linear problem formulation. In: Chinneck JW, Kristjansson B, Saltzman M, editors. *Operations research and cyber-infrastructure (ICS 2009 Proceedings)*. Berlin: Springer; 2009. pp. 79–100.
2. Jeroslow RG. Representability in mixed integer programming, I: characterization results. *Discrete Appl Math* 1987;17:223–243.
3. Aardal K. Reformulation of capacitated facility location problems: how redundant information can help. *Ann Oper Res* 1998;82:289–309.
4. Trick M. Formulations and reformulations in integer programming. In: Barták R, Milano M, editors. *Volume 3524, Integration of AI and OR techniques in constraint programming for combinatorial optimization problems (CPAIOR 2005)*. Lecture Notes in Computer Science. Berlin: Springer; 2005. pp. 366–379.
5. Padberg M. Approximating separable nonlinear functions via mixed zero-one programs. *Oper Res Lett* 2000;27:1–5.
6. Sherali HD. On mixed-integer zero-one representations for separable lower-semicontinuous piecewise-linear functions. *Oper Res Lett* 2001;28:155–160.
7. Li H-L, Lu H-C, Huang C-H, *et al.* A superior representation method for piecewise linear functions. *INFORMS J Comput* 2009;21:314–321.
8. Vielma JP, Nemhauser GL. Modeling disjunctive constraints with a logarithmic number of binary variables and constraints. In: *Integer Programming and Combinatorial Optimization Proceedings (IPCO 2008)*, volume 5035. Lecture Notes in Computer Science. Berlin: Springer; 2008. pp. 199–213.
9. Vielma JP, Ahmed S, Nemhauser G. A note on: A superior representation method for piecewise linear functions by Li, Lu, Huang, and Hu. *INFORMS J Comput* 2010;22:493–497.
10. Keha AB, de Farias IR, Nemhauser GL. Models for representing piecewise linear cost functions. *Oper Res Lett* 2004;32:44–48.
11. Guéret C, Heipcke S, Prins C, *et al.* *Applications of optimization with Xpress-MP*. White paper, Dash Optimization, 2000.
12. Williams HP. *Model building in mathematical programming*. 4th ed. New York: Wiley; 1999.
13. Williams HP. The formulation and solution of discrete optimization models. In: Appa G, Pitsoulis L, Williams HP, editors. *Handbook on modelling for discrete optimization*. Berlin: Springer; 2006. pp. 3–38.
14. Hooker JN. *Integrated methods for optimization*. Berlin: Springer; 2007.
15. Williams HP. *Logic and integer programming*. Berlin: Springer; 2009.
16. Orman AJ, Williams HP. A survey of different integer programming formulations of the travelling salesman problem. In: Kon-toghiorghes EJ, Gatu C, editors. *Optimization, economics and financial analysis. Advances in computational management science*. Berlin: Springer; 2006. pp. 933–106.
17. Hooker JN. Hybrid modeling. In: Van Hentenryck P, Milano M, editors. *Hybrid optimization: the 10 years of CPAIOR*. Berlin: Springer. In press.

FOUNDATIONS OF CONSTRAINED OPTIMIZATION

SVEN LEYFFER
ASHUTOSH MAHAJAN
Mathematics and Computer
Science Division, Argonne
National Laboratory, Argonne,
Illinois

BACKGROUND AND INTRODUCTION

Nonlinearly constrained optimization problems (NCOs) are an important class of problems with a broad range of engineering, scientific, and operational applications. For ease of presentation, we consider NCOs of the form

$$\underset{x}{\text{minimize}} \ f(x) \quad \text{subject to} \ c(x) = 0 \text{ and } x \geq 0, \quad (1)$$

where the objective function, $f: \mathbb{R}^n \rightarrow \mathbb{R}$, and the constraint functions, $c: \mathbb{R}^n \rightarrow \mathbb{R}^m$, are twice continuously differentiable. We denote the multipliers corresponding to the equality constraints, $c(x) = 0$, by y and the multipliers of the inequality constraints, $x \geq 0$, by $z \geq 0$. An NCO may also have unbounded variables, upper bounds, or general range constraints of the form $l_i \leq c_i(x) \leq u_i$, which we omit for the sake of simplicity.

In general, one cannot solve Equation (1) directly or explicitly. Instead, an iterative method is used that solves a sequence of simpler, approximate subproblems to generate a sequence of approximate solutions, $\{x_k\}$, starting from an initial guess, x_0 . Every subproblem may in turn be solved by an iterative process. These inner iterations are also referred to as *minor iterations*. A simplified algorithmic framework for solving Equation (1) is as follows:

Algorithm 1. Framework for nonlinear optimization methods

```

Given initial estimate  $(x_0, y_0, z_0) \in \mathbb{R}^{n+m+n}$ ,
set  $k = 0$ ;
while  $x_k$  is not optimal do
    repeat
        Approximately solve and refine an approximate subproblem of Equation (1) around  $x_k$ .
    until improved solution estimate  $x_{k+1}$  found;
    Check whether  $x_{k+1}$  is optimal; set  $k = k + 1$ .
end

```

In this article, we review the basic components of methods for solving NCOs. In particular, we review the four fundamental components of Algorithm 1: the *convergence test* that checks for optimal solutions or detects failure; the *approximate subproblem* that computes an improved new iterate; the *globalization strategy* that ensures convergence from remote starting points, by indicating whether a new solution estimate is better than the current estimate; and the *globalization mechanism* that truncates the step computed by the local model to enforce the globalization strategy, effectively refining the local model.

Algorithms for NCOs are categorized by the choice they implement for each of these fundamental components. In the next section, we review the fundamental building blocks of methods for nonlinearly constrained optimization. Our presentation is implementation-oriented and emphasizes the common components of different classes of methods.

Notation. Throughout this paper, we denote iterates by $x_k, k = 1, 2, \dots$, and we use subscripts to indicate functions evaluated at an iterate, for example, $f_k = f(x_k)$ and $c_k = c(x_k)$. We also denote the gradients by $g_k = \nabla f(x_k)$ and the Jacobian by $A_k = \nabla c(x_k)$. The Hessian of the Lagrangian is denoted by H_k .

CONVERGENCE TEST AND TERMINATION CONDITIONS

We start by describing the convergence test, a common component among all NCO

algorithms. The convergence test also provides the motivation for the approximate subproblems that are described next. The convergence analysis of NCO algorithms typically provides convergence only to Karush-Kuhn-Tucker (KKT) points. A suitable approximate convergence test is thus given by

$$\begin{aligned} \|c_k\| \leq \epsilon \quad \text{and} \quad \|g_k - A_k y_k - z_k\| \leq \epsilon \\ \text{and} \quad \|\min(x_k, z_k)\| \leq \epsilon, \end{aligned} \quad (2)$$

where $\epsilon > 0$ is the tolerance and the min in the last expression corresponding to complementary slackness is taken componentwise.

In practice, it may not be possible to ensure convergence to an approximate KKT point, for example, if the constraints fail to satisfy a constraint qualification [1, Chapter 7]. In that case, we replace the second condition by

$$\|A_k y_k + z_k\| \leq \epsilon,$$

which corresponds to a Fritz–John point.

Infeasible Stationary Points

Unless the NCO is convex or some restrictive assumptions are made, methods cannot guarantee convergence even to a feasible point. Moreover, an NCO may not even have a feasible point, and we are interested in a (local) certificate of infeasibility. In this case, neither the approximate subproblem nor the convergence test is adequate to achieve and detect convergence. A more appropriate convergence test and feasibility subproblem can be based on the following feasibility problem:

$$\underset{x}{\text{minimize}} \quad \|c(x)\| \quad \text{subject to} \quad x \geq 0, \quad (3)$$

which can be formulated as a smooth optimization problem by introducing slack variables. Algorithms for solving Equation (3) are analogous to algorithms for NCOs, because the feasibility problem can be reformulated as a smooth NCO by introducing additional variables. In general, we can replace this objective by any weighted norm. A suitable convergence test is then

$$\|A_k y_k - z_k\| \leq \epsilon \quad \text{and} \quad \|\min(x_k, z_k)\| \leq \epsilon,$$

where y_k are the multipliers or weights corresponding to the norm used in the objective of Equation (3). For example, if we use the ℓ_1 norm, then if $[c_k]_i < 0$, $[y_k]_i = -1$, if $[c_k]_i > 0$, $[y_k]_i = 1$, and $-1 \leq [y_k]_i \leq 1$ otherwise. The multipliers are readily computed as a by-product of solving the subproblem.

APPROXIMATE SUBPROBLEM: IMPROVING A SOLUTION ESTIMATE

One key difference among nonlinear optimization methods is how the approximate subproblem is constructed. The goal of the approximate subproblem is to provide a *computable* step that improves on the current iterate. We distinguish three broad classes of approximate subproblems: sequential linear models, sequential quadratic models, and interior-point models. Methods that are based on the augmented Lagrangian method are more suitably described in the context of globalization strategies in the section titled “Globalization Strategy: Convergence from Remote Starting Points.”

Sequential Linear and Quadratic Programming

Sequential linear and quadratic programming methods construct a linear or quadratic approximation of Equation (1) and solve a sequence of such approximations, converging to a stationary point.

Sequential Quadratic Programming (SQP) Methods. Sequential quadratic programming (SQP) methods successively minimize a quadratic model, $m_k(d)$, subject to a linearization of the constraints about x_k [2–4] to obtain a displacement $d := x - x_k$.

$$\begin{aligned} \underset{d}{\text{Minimize}} \quad m_k(d) &:= g_k^T d + \frac{1}{2} d^T H_k d \\ \text{subject to} \quad c_k + A_k^T d &= 0, \quad x_k + d \geq 0, \end{aligned} \quad (4)$$

where $H_k \simeq \nabla^2 \mathcal{L}(x_k, y_k)$ approximates the Hessian of the Lagrangian and y_k is the multiplier estimate at iteration k . The new iterate is $x_{k+1} = x_k + d$, together with the multipliers y_{k+1} of the linearized constraints of Equation (4). If H_k is not positive-definite on the null-space of the active constraint

normals, then the quadratic programming (QP) is nonconvex, and SQP methods seek a local minimum of Equation (4). The solution of the QP subproblem can become computationally expensive for large-scale problems because the null-space method for solving QPs requires the factorization of a dense reduced-Hessian matrix. This bottleneck has led to the development of other methods that use linear programming (LP) solutions in the approximate subproblem, and these approaches are described next.

Sequential Linear Programming (SLP) Methods. Sequential linear programming (SLP) methods construct a linear approximation to Equation (1). In general, this LP will be unbounded, and SLP methods require the addition of a trust region (discussed in more detail in the next section):

$$\begin{aligned} & \underset{d}{\text{minimize}} \quad m_k(d) = g_k^T d \\ & \text{subject to} \quad c_k + A_k^T d = 0, \quad x_k + d \geq 0, \\ & \text{and} \quad \|d\|_\infty \leq \Delta_k, \end{aligned} \quad (5)$$

where $\Delta_k > 0$ is the trust-region radius. Griffith and Stewart [5] used this method without a trust region but with the assumption that the variables are bounded. In general, $\Delta_k \rightarrow 0$ must converge to zero to ensure convergence. SLP methods can be viewed as steepest descent methods and typically converge only linearly. If, however there are exactly n active and linearly independent constraint normals at the solution, then SLP reduces to Newton's method for solving a square system of nonlinear equations and converges superlinearly.

Sequential Linear/Quadratic Programming (SLQP) Methods. Sequential linear/quadratic programming (SLQP) methods combine the advantages of the SLP method (fast solution of the LP) and SQP methods (fast local convergence) by adding equality-constrained quadratic programming (EQP) to the SLP method [6–8]. SLQP methods thus solve two subproblems: first, an LP is solved to obtain a step for the next iteration and also an estimate of the active set $\mathcal{A}_k := \{i : [x_k]_i + \hat{d}_i = 0\}$ from a solution $\hat{d}$ of Equation (5). This

estimate of the active set is then used to construct an EQP, on the active constraints,

$$\begin{aligned} & \underset{d}{\text{minimize}} \quad q_k(d) = g_k^T d + \frac{1}{2} d^T H_k d \\ & \text{subject to} \quad c_k + A_k^T d = 0, \\ & \quad [x_k]_i + d_i = 0, \quad \forall i \in \mathcal{A}_k. \end{aligned} \quad (6)$$

If H_k is second-order sufficient (i.e., positive-definite on the null-space of the constraints), then the solution of Equation (6) is equivalent to the following linear system obtained by applying the KKT conditions to the EQP:

$$\begin{bmatrix} H_k & -A_k & -I_k \\ A_k^T & 0 & 0 \\ I_k^T & 0 & 0 \end{bmatrix} \begin{pmatrix} x \\ y_A \end{pmatrix} = \begin{pmatrix} -g_k + H_k x_k \\ -c_k \\ 0 \end{pmatrix},$$

where $I_k = [e_i]_{i \in \mathcal{A}_k}$ are the normals of the active inequality constraints. By taking a suitable basis from the LP simplex solve, SLQP methods can ensure that $[A_k : I_k]$ has full rank. Linear solvers such as MA57 can also detect the inertia; and if H_k is not second-order sufficient, a multiple of the identity can be added to H_k to ensure descent of the EQP step.

Sequential Quadratic/Quadratic Programming (SQQP) Methods. Sequential quadratic/quadratic programming (SQQP) methods have recently been proposed as SQP types of methods [9,10]. First, a convex QP model constructed by using a positive-definite Hessian approximation is solved. The solution of this convex QP is followed by a reduced inequality constrained model or an EQP with the exact second derivative of the Lagrangian.

Theory of Sequential Linear/Quadratic Programming Methods. If H_k is the exact Hessian of the Lagrangian and if the Jacobian of the active constraints has full rank, then SQP methods converge quadratically near a minimizer that satisfies a constraint qualification and a second-order sufficient condition [4].

It can also be shown that, under the additional assumption of strict complementarity, all four methods identify the optimal active set in a finite number of iterations.

The methods described in this section are also often referred to as *active-set methods*, because the solution of each LP or QP provides not only a suitable new iterate but also an estimate of the active set at the solution.

Interior-Point Methods

Interior-point methods (IPMs) are an alternative approach to active-set methods. Interior-point methods are a class of perturbed Newton methods that postpone the decision of which constraints are active until the end of the iterative process. The most successful IPMs are primal–dual IPMs, which can be viewed as Newton’s method applied to the perturbed first-order conditions of Equation (1):

$$0 = F_\mu(x, y, z) = \begin{pmatrix} \nabla f(x) - \nabla c(x)^T y - z \\ c(x) \\ Xz - \mu e \end{pmatrix}, \quad (7)$$

where $\mu > 0$ is the barrier parameter, $X = \text{diag}(x)$ is a diagonal matrix with x along its diagonal, and $e = (1, \dots, 1)$ is the vector of all ones. Note that, for $\mu = 0$, these conditions are equivalent to the first-order conditions except for the absence of the nonnegativity constraints $x, z \geq 0$.

IPMs start at an “interior” iterate $x_0, z_0 > 0$ and generate a sequence of interior iterates $x_k, z_k > 0$ by approximately solving the first-order conditions Equation (7) for a decreasing sequence of barrier parameters. IPMs can be shown to be polynomial-time algorithms for convex nonlinear programming (NLP) [11].

Newton’s method applied to the primal–dual system of Equation (7) around x_k gives rise to the approximate subproblem,

$$\begin{bmatrix} H_k & -A_k & -I \\ A_k^T & 0 & 0 \\ Z_k & 0 & X_k \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = -F_\mu(x_k, y_k, z_k), \quad (8)$$

where H_k approximates the Hessian of the Lagrangian, $\nabla^2 \mathcal{L}_k$, and the step $(x_{k+1},$

$y_{k+1}, z_{k+1}) = (x_k, y_k, z_k) + (\alpha_x \Delta x, \alpha_y \Delta y, \alpha_z \Delta z)$ is safeguarded to ensure that $x_{k+1}, z_{k+1} > 0$ remain strictly positive. A sketch of IPM is given in Algorithm 2.

Algorithm 2. Primal–dual interior-point method

Given an initial solution estimate (x_0, y_0, z_0) ,
such that $(x_0, z_0) > 0$.
Choose a barrier parameter μ_0 , $0 < \sigma < 1$, and a
sequence $\epsilon_k \searrow 0$.
repeat
 Set $(x_{k,0}, y_{k,0}, z_{k,0}) = (x_k, y_k, z_k)$, $l = 0$.
 repeat
 Approximately solve the Newton system (8)
 for a new iterate $(x_{k,l+1}, y_{k,l+1}, z_{k,l+1})$.
 Set $l = l + 1$.
 until $\|F_{\mu_k}(x_{k,l}, y_{k,l}, z_{k,l})\| \leq \epsilon_k$;
 Reduce the barrier parameter $\mu_{k+1} = \sigma \mu_k$,
 and set $k = k + 1$.
until x_k, y_k, z_k optimal;

Relationship to Barrier Methods. Primal–dual IPMs are related to earlier barrier methods [12]. These methods were given much attention in the 1960s but soon lost favor because of the ill-conditioning of the Hessian. They regained attention in the 1980s after it was shown that these methods can provide polynomial-time algorithms for LP problems. See the surveys [13–15] for further material. Barrier methods approximately solve a sequence of barrier problems,

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad f(x) - \mu \sum_{i=1}^n \log(x_i) \\ & \text{subject to} \quad c(x) = 0, \end{aligned} \quad (9)$$

for a decreasing sequence of barrier parameters $\mu > 0$. The first-order conditions of Equation (9) are given by

$$\nabla f(x) - \mu X^{-1}e - A(x)y = 0 \quad \text{and} \quad c(x) = 0. \quad (10)$$

Applying Newton’s method to this system of equations results in the following linear system:

$$\begin{bmatrix} H_k + \mu X_k^{-2} & -A_k \\ A_k^T & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = - \begin{pmatrix} g_k - \mu X_k^{-1}e - A_k y_k \\ c_k \end{pmatrix}.$$

Introducing first-order multiplier estimates $Z(x_k) := \mu X_k^{-1}$, which can be written as $Z(x_k)X_k = \mu e$, we obtain the system

$$\begin{bmatrix} H_k + Z(x_k)X_k^{-1} & -A_k \\ A_k & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = - \begin{pmatrix} g_k - \mu X_k^{-1}e - A_k y_k \\ c_k \end{pmatrix},$$

which is equivalent to the primal–dual Newton system (8), where we have eliminated

$$\Delta z = -X^{-1}Z\Delta x - Ze - \mu X^{-1}e.$$

Thus, the main difference between classical barrier methods and the primal–dual IPMs is that Z_k is not free for barrier methods but is chosen as the primal multiplier $Z(x_k) = \mu X_k^{-1}$. This freedom in the primal–dual method avoids some difficulties with ill-conditioning of the barrier Hessian.

Convergence of Barrier Methods. If there exists a compact set of isolated local minimizers of Equation (1) with at least one point in the closure of the strictly feasible set, then it follows that barrier methods converge to a local minimum [13].

GLOBALIZATION STRATEGY: CONVERGENCE FROM REMOTE STARTING POINTS

The approximate subproblems of the preceding section guarantee convergence only in a small neighborhood of a regular solution. Globalization strategies are concerned with ensuring convergence from remote starting points to stationary points (and should not be confused with global optimization). To ensure convergence from remote starting points, we must monitor the progress of the iterates generated by the approximate subproblem. Monitoring is easily done in unconstrained optimization, where we can measure progress by comparing objective values. In constrained optimization, however, we must take the constraint violation into account. Three broad classes of strategies exist: augmented Lagrangian methods, penalty and

merit-function methods, and filter and funnel methods.

Augmented Lagrangian Methods

The augmented Lagrangian of Equation (1) is given by

$$\mathcal{L}(x, y, \rho) = f(x) - y^T c(x) + \frac{\rho}{2} \|c(x)\|_2^2, \quad (11)$$

where $\rho > 0$ is the penalty parameter. The augmented Lagrangian is used in two modes to develop algorithms for solving Equation (1): by defining a linearly constrained problem or by defining a bound-constrained problem.

Linearly Constrained Lagrangian Methods.

These methods successively minimize a shifted augmented Lagrangian subject to a linearization of the constraints. The shifted augmented Lagrangian is defined as

$$\bar{\mathcal{L}}(x, y, \rho) = f(x) - y^T p_k(x) + \frac{\rho}{2} \|p_k(x)\|_2^2, \quad (12)$$

where $p_k(x)$ are the higher-order nonlinear terms at the current iterate x_k , that is,

$$p_k(x) = c(x) - c_k - A_k^T(x - x_k). \quad (13)$$

This approach results in the following approximate subproblem:

$$\begin{aligned} & \underset{x}{\text{minimize}} \quad \bar{\mathcal{L}}(x, y_k, \rho_k) \\ & \text{subject to} \quad c_k + A_k^T(x - x_k) = 0, \quad x \geq 0. \end{aligned} \quad (14)$$

We note that if $c_k + A_k^T(x - x_k) = 0$, then minimizing the shifted augmented Lagrangian is equivalent to minimizing the Lagrangian over these constraints. Linearly constrained, augmented Lagrangian methods solve a sequence of problems (14) for a fixed penalty parameter. Multipliers are updated by using a first-order multiplier update rule,

$$y_{k+1} = y_k - \rho_k c(x_{k+1}), \quad (15)$$

where x_{k+1} solves Equation (14). We observe, that augmented Lagrangian methods iterate on the dual variables, unlike the active-set, or IPMs of the previous sections.

Bound-Constrained Lagrangian Methods.

These methods approximately minimize the augmented Lagrangian,

$$\underset{x}{\text{minimize}} \mathcal{L}(x, y_k, \rho_k) \quad \text{subject to } tx \geq 0. \quad (16)$$

The advantage of this approach is that efficient methods for bound-constrained optimization can readily be applied, such as the gradient-projection conjugate-gradient approach [16], which can be interpreted as an approximate Newton method on the active inequality constraints.

Global convergence is promoted by defining two forcing sequences, $\omega_k \searrow 0$, controlling the accuracy with which every bound-constrained problem is solved, and $\eta_k \searrow 0$, controlling progress toward feasibility of the nonlinear constraints. A typical bound-constrained Lagrangian method can then be stated as follows:

Algorithm 3. Bound-constrained augmented Lagrangian method

Given an initial solution estimate (x_0, y_0) , and an initial penalty parameter ρ_0 .

repeat

Set $x_{k,0} = x_k$, $\rho_{k,0} = \rho_k$, $l = 0$, $\text{success} = \text{false}$.

repeat

Find an ω_k -optimal solution $x_{k,l+1}$ of
minimize $\mathcal{L}(x, y_k, \rho_{k,l})$ s.t. $x \geq 0$

if $\|c(x_{k,l+1})\| \leq \eta_k$ **then**

Perform a first-order multiplier update:

$y_{k+1} = y_k - \rho_{k,l} c(x_{k,l+1})$.

Set $\rho_{k+1} = \rho_{k,l}$, and $\text{success} = \text{true}$.

else

Increase penalty: $\rho_{k,l+1} = 10\rho_{k,l}$;
set $l = l + 1$.

end

until $\text{success} = \text{true}$;

Set $x_{k+1} = x_{k,l+1}$, and $k = k + 1$.

until x_k, y_k is optimal;

We note that the inner loop in Algorithm 3 updates the penalty parameter until it is sufficiently large to force progress toward feasibility, at which point the multipliers are updated. Each minimization can be started from the previous iterate, $x_{k,l}$.

Theory of Augmented Lagrangian Methods.

Conn *et al.* [17] show that a bound-constrained Lagrangian method converges globally if the sequence $\{x_k\}$ of iterates is bounded and if the Jacobian of the constraints at all limit points of $\{x_k\}$ has column rank no smaller than m . Conn *et al.* [17] also show that if some additional conditions are met, then their algorithm is R-linearly convergent. Bertsekas [18] shows that the method converges Q-linearly if $\{\rho_k\}$ is bounded, and superlinearly otherwise. Linearly constrained augmented Lagrangian methods can be made globally convergent by adding slack variables to handle infeasible subproblems [19].

Penalty and Merit-Function Methods

Penalty and merit functions combine the objective function and a measure of the constraint violation into a single function whose local minimizers correspond to local minimizers of the original problem (1). Convergence from remote starting points can then be ensured by forcing descent of the penalty or merit function, using one of the mechanisms of the next section.

Exact penalty functions are an attractive alternative to augmented Lagrangians and are defined as

$$p_\rho(x) = f(x) + \rho \|c(x)\|,$$

where $\rho > 0$ is the penalty parameter. Most approaches use the ℓ_1 norm to define the penalty function. It can be shown that a local minimizer, x^* , of $p_\rho(x)$ is a local minimizer of problem (1) if $\rho > \|y^*\|_D$, where y^* are the corresponding Lagrange multipliers and $\|\cdot\|_D$ is the dual norm of $\|\cdot\|$ (i.e., the ℓ_∞ -norm in the case of the ℓ_1 exact-penalty function) [20, Chapter 14.3]. Classical approaches using $p_\rho(x)$ have solved a sequence of penalty problems for an increasing sequence of penalty parameters. Modern approaches attempt to steer the penalty parameter by comparing the predicted decrease in the constraint violation to the actual decrease over a step [21].

A number of other merit functions also exist. The oldest, the quadratic penalty function, $f(x) + \rho \|c(x)\|_2^2$, converges only if

the penalty parameter diverges to infinity. Augmented Lagrangian functions and Lagrangian penalty functions such as $f(x) + y^T c(x) + \rho \|c(x)\|$ have also been used to promote global convergence. A key ingredient in any convergence analysis is to connect the approximate subproblem to the merit function that is being used in a way that ensures a descent property of the merit function; see the section titled “Line-Search Methods.”

Filter and Funnel Methods

Filter and funnel methods provide an alternative to penalty methods that does not rely on the use of a penalty parameter. Both methods use step acceptance strategies that are closer to the original problem, by separating the constraints and the objective function.

Filter Methods. Filter methods keep a record of the constraint violation, $h_l := \|c(x_l)\|$, and objective function value, $f_l := f(x_l)$, for some previous iterates, $x_l, l \in \mathcal{F}_k$ [22]. A new point is acceptable if it improves either the objective function or the constraint violation compared to all previous iterates. That is, $\hat{x}$ is acceptable if

$$f(\hat{x}) \leq f_l - \gamma h_l \quad \text{or} \quad h(\hat{x}) \leq \beta h_l, \quad \forall l \in \mathcal{F}_k,$$

where $\gamma > 0, 0 < \beta < 1$, are constants that ensure that iterates cannot accumulate at infeasible limit points. A typical filter is shown in Fig. 1(a), where the straight lines correspond to the region in the (h, f) -plane that is dominated by previous iterations and the dashed lines correspond to the envelope defined by γ, β .

The filter provides convergence only to a feasible limit because any infinite sequence of iterates must converge to a point, where $h(x) = 0$, provided that $f(x)$ is bounded below. To ensure convergence to a local minimum, filter methods use a standard sufficient reduction condition from unconstrained optimization,

$$f(x_k) - f(x_k + d) \geq -\sigma m_k(d), \quad (17)$$

where $\sigma > 0$ is the fraction of predicted decrease and $m_k(d)$ is the model reduction from the approximate subproblem. It makes sense to enforce this condition only if the model predicts a decrease in the objective function. Thus, filter methods use the switching condition $m_k(d) \geq \gamma h_k^2$ to decide when Equation (17) should be enforced. A new iterate that satisfies both conditions is called an *f-type iterate*, and an iterate for which the switching condition fails is called an *h-type iterate* to indicate that it mostly reduces the constraint violation. If a new point is accepted, then it is added to the current iterate to the filter, $\mathcal{F}_k$, if $h_k > 0$ or if it corresponds to h-type iterations (which automatically satisfy $h_k > 0$).

Funnel Methods. The method of Gould and Toint [23] can be viewed as a filter method with just a single filter entry, corresponding to an upper bound on the constraint violation. Thus, the filter contains only a single entry, $(U_k, -\infty)$. The upper bound is reduced during h-type iterations, to force the iterates toward feasibility; it is left unchanged during f-type iterations. Thus, it is possible to converge without reducing U_k to zero (consistent with the observation that SQP methods converge locally). A schematic interpretation of the funnel is given in Fig. 1(b).

Maratos Effect and Loss of Fast Convergence

One can construct simple examples showing that arbitrarily close to an isolated strict local minimizer, the Newton step will be rejected by the exact penalty function [24], resulting in slow convergence. This phenomenon is known as the *Maratos effect*. It can be mitigated by computing a second-order correction step, which is a Newton step that uses the same linear system with an updated right-hand side [20,25]. An alternative method to avoid the Maratos effect is the use of nonmonotone techniques that require descent over only the last M iterates, where $M > 1$ is a constant.

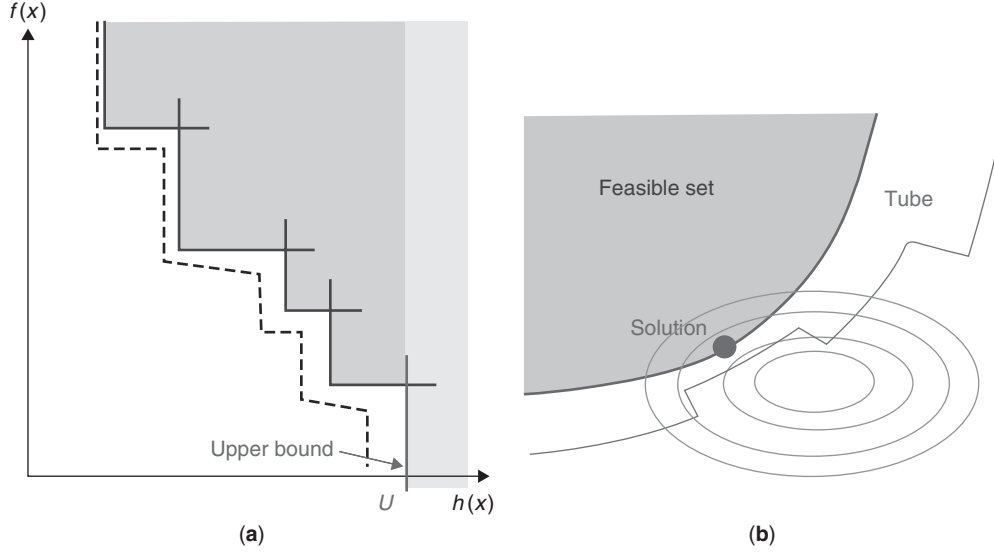


Figure 1. (a) shows a filter where the grey area corresponds to the points that are rejected by the filter; (b) shows a funnel around the feasible set.

GLOBALIZATION MECHANISMS

In this section, we review two mechanisms to reduce the step that is computed by the approximate subproblem: line-search methods and trust-region methods. Both mechanisms can be used in conjunction with any of the approximate subproblems and any of the global convergence strategies, giving rise to a broad family of algorithms. In the article *Software for Nonlinearly Constrained Optimization*, we describe how these components are used in software for NCOs.

Line-Search Methods

Line-search methods enforce convergence with a backtracking line search along the direction s . For IPMs, the search direction, $s = (\Delta_x, \Delta_y, \Delta_z)$, is obtained by solving the primal-dual system (8). For SQP methods, the search direction is the solution of the QP in Equation (4), $s = d$. It is important to ensure that the model produces a descent direction, e.g., $\nabla \Phi(x_k)^T s < 0$ for a merit or penalty function

$\Phi(x)$; otherwise, the line search may not terminate. A popular line search is the Armijo search [25], described in Algorithm 4 for a merit function $\Phi(x)$. The algorithm can be shown to converge to a stationary point, detect unboundedness, or converge to a point where there are no directions of descent.

Algorithm 4. (Armijo) Line-Search Method for Nonlinear Optimization

Given initial estimate $x_0 \in \mathbb{R}^n$, let $0 < \sigma < 1$, and set $k = 0$;

while x_k is not optimal **do**

 Approximately solve an approximate subproblem of Equation (1) around x_k to find a search direction s .

 Make sure that s is a descent direction, e.g. $\nabla \Phi(x_k)^T s < 0$.

 Set $\alpha^0 = 1$ and $l = 0$.

repeat

 Set $\alpha^{l+1} = \alpha^l/2$ and evaluate $\Phi(x_k + \alpha^{l+1}s)$.

 Set $l = l + 1$.

until $\Phi(x_k + \alpha^l s) \leq f_k + \alpha^l \sigma s^T \nabla \Phi_k$;

 set $k = k + 1$.

end

Line-search methods for filters can be defined in a similar way. Instead of checking descent in the merit function, a filter method is used to check acceptance to a filter. Unlike merit functions, filter methods do not have a simple definition of descent; hence, the line search is terminated unsuccessfully once the step size α^l becomes smaller than a constant. In this case, filter methods switch to a restoration step, obtained by solving a local approximation of Equation (3).

Trust-Region Methods

Trust-region methods explicitly restrict the step that is computed by the approximate subproblem, by adding a trust-region constraint of the form $\|d\| \leq \Delta_k$ to the approximate subproblem. Most methods use an ℓ_∞ -norm trust region, which can be represented by bounds on the variables. The trust-region radius, $\Delta_k > 0$, is adjusted at every iteration depending on how well the approximate subproblem agrees with the NCO, Equation (1).

Algorithm 5. Trust-Region Methods for Nonlinear Optimization

Given initial estimate $x_0 \in \mathbb{R}^n$, choose $\Delta_0 \geq \underline{\Delta} > 0$, and set $k = 0$;

repeat

Reset $\Delta_{k,l} := \Delta_k \geq \underline{\Delta} > 0$; set `success` = false, and $l = 0$.

repeat

Solve an approximate subproblem in a *trust-region*, e.g. Equation (4) with $\|d\| \leq \Delta_{k,l}$.

if $x_k + d$ sufficiently better than x_k **then**

Accept the step: $x_{k+1} = x_k + d$; possibly increase $\Delta_{k,l+1}$; set `success` = true.

else

Reject the step and decrease the trust-region $\Delta_{k,l+1} = \Delta_{k,l}/2$.

end

until `success` = true;

Set $k = k + 1$.

until x_k is optimal;

Step acceptance in this algorithm can be based either on filter methods, or on

sufficient decrease in a merit or penalty function, see Algorithm 4. Trust-region methods are related to regularization techniques, which add a multiple of the identity matrix, $\sigma_k I$, to the Hessian, H_k . Locally, the solution of the regularized problem is equivalent to the solution of a trust-region problem with an ℓ_2 trust-region. One disadvantage of trust-region methods is the fact that the subproblem may become inconsistent as $\Delta_{k,l} \rightarrow 0$. This situation can be dealt with in three different ways: (i) a penalty function approach; (ii) a restoration phase in which the algorithm minimizes the constraint violation [26]; or (iii) a composite step approach [27].

Acknowledgments

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, US Department of Energy, under Contract DE-AC02-06CH11357. This work was also supported by the US Department of Energy through grant DE-FG02-05ER25694.

REFERENCES

1. Mangasarian O. Nonlinear programming. New York: McGraw-Hill Book Company; 1969.
2. Han S. A globally convergent method for nonlinear programming. J Optim Theory Appl 1977;22(3):297–309.
3. Powell M. A fast algorithm for nonlinearly constrained optimization calculations. In: Watson G, editor. Numerical analysis. Berlin: Springer; 1978. pp. 144–157.
4. Boggs P, Tolle J. Sequential quadratic programming. Acta Numer 1995;4:1–51.
5. Griffith R, Stewart R. A nonlinear programming technique for the optimization of continuous processing systems. Manage Sci 1961;7(4):379–392.
6. Fletcher R, de la Maza ES. Nonlinear programming and nonsmooth optimization by successive linear programming. Math Program 1989;43:235–256.
7. Chin C, Fletcher R. On the global convergence of an SLP-filter algorithm that takes EQP steps. Math Program 2003;96(1):161–177.

8. Byrd RH, Gould NIM, Nocedal J, *et al.* An algorithm for nonlinear optimization using linear programming and equality-constrained subproblems. *Math Program Ser B* 2004;100(1):27–48.
9. Gould NIM, Robinson DP. A second derivative SQP method: global convergence. *SIAM J Optim* 2010;20(4):2023–2048.
10. Gould NIM, Robinson DP. A second derivative SQP method: local convergence. *Numerical Analysis Report 08/21*. Oxford University Computing Laboratory. 2008. (To appear in *SIAM Journal on Optimization*).
11. Nesterov Y, Nemirovskii A. Interior point polynomial algorithms in convex programming. Volume 13, *Studies in applied mathematics*. Philadelphia (PA): SIAM; 1994.
12. Fiacco AV, McCormick GP. Nonlinear programming: sequential unconstrained minimization techniques. Volume 4, *Classics in applied mathematics*. Philadelphia (PA): SIAM; 1990. (Reprint of the original book published in 1968 by Wiley, New York).
13. Wright M. Interior methods for constrained optimization. *Acta Numer* 1992;1:341–407.
14. Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. *SIAM Rev* 2002;4(4):525–597.
15. Nemirovski A, Todd M. Interior-point methods for optimization. *Acta Numer* 2008; 17:181–234.
16. Moré JJ, Toraldo G. On the solution of quadratic programming problems with bound constraints. *SIAM J Optim* 1991;1(1): 93–113.
17. Conn AR, Gould NIM, Toint PL. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J Numer Anal* 1991;28(2):545–572.
18. Bertsekas D. *Constrained optimization and Lagrange multiplier methods*. Belmont (MA): Athena Scientific; 1996.
19. Friedlander MP, Saunders MA. A globally convergent linearly constrained Lagrangian method for nonlinear optimization. *SIAM J Optim* 2005;15(3):863–897.
20. Fletcher R. *Practical methods of optimization*. Chichester: John Wiley & Sons, Ltd.; 1987.
21. Byrd RH, Nocedal J, Waltz RA. Steering exact penalty methods for nonlinear programming. *Optim Method Softw* 2008;23(2):197–213.
22. Fletcher R, Leyffer S. Nonlinear programming without a penalty function. *Math Program* 2002;91:239–270.
23. Gould NIM, Toint PL. Nonlinear programming without a penalty function or a filter. *Math Program* 2010;122(1):155–196.
24. Maratos N. Exact penalty function algorithms for finite-dimensional and control optimization problems [PhD thesis]. University of London. 1978.
25. Nocedal J, Wright S. *Numerical optimization*. New York: Springer; 1999.
26. Fletcher R, Leyffer S. Filter-type algorithms for solving systems of algebraic equations and inequalities. In: di Pillo G, Murli A, editors. *High performance algorithms and software for nonlinear optimization*. Dordrecht, The Netherlands: Kluwer academic publishers; 2003. pp. 259–278.
27. Omojokun EO. Trust-region algorithms for optimization with nonlinear equality and inequality constraints [PhD thesis]. Boulder (CO): University of Colorado; 1989.

FOUNDATIONS OF DECISION THEORY

LOUIS ANTHONY COX, JR.
Cox Associates and Department of
Mathematical and Statistical
Sciences, University of Colorado,
Denver, Colorado

A FRAMEWORK FOR DECISION ANALYSIS: ACTS, STATES, AND CONSEQUENCES

Modern decision analysis is based on the following fundamental framework. A decision maker (DM) must choose one among many *alternatives* (often referred to as *acts*) from a set of possible alternatives, A . The DM's choice will be followed by a *consequence*, belonging to a set of possible consequences, C . Which consequence is produced may depend not only on the DM's choice of act, but also on what else happens; this is modeled by calling everything that affects the consequence, other than the DM's choice, the *state of nature* (or, more briefly, the *state*), and by describing the decision problem in terms of a *consequence function* that associates a specific consequence in C with each act–state pair, say (a, s) , from $A \times S$, where A is the “choice set” of possible alternatives (or choices, decisions, and acts); and S is the set of possible states. The consequence function represents the causal relation between acts and their consequences, for each state. (If this relation is uncertain, then it can be included in the definitions of states.) Let $c(a, s)$ denote the consequence that results when the DM takes act a and the state is s . Decision analysis addresses the fundamental problem of deciding *which act in A to choose*, given a consequence function, and informed by preferences for consequences in C , and by any available beliefs or information about states in S .

Example [Choosing the lesser of two evils]. Table 1 represents a simple decision problem, with two possible states (state 1 and state 2) in S and two possible acts (act A and

act B) in the choice set A . This “payoff matrix” representation of the decision problem shows consequence, $c(a, s)$, for each act–state pair.

For example, $c(A, 1) = 2$ deaths, and $c(B, 2) = 5$ deaths. The DM must choose between acts A and B, based on the information summarized in this table, preferences for consequences (which we assume to be such that fewer deaths are preferred to more), and any available information or beliefs about the states. Methods for making such a decision are discussed next.

ALTERNATIVE FORMULATIONS AND NORMATIVE MODELS FOR DECISION MAKING

If the state is modeled as a random variable, with known probabilities for each possible state (or with known probability measures or probability densities, if the state space S is continuous instead of discrete), then another common version of the fundamental decision problem is to decide which act in A to choose, given the *probabilities of consequences* for each act. This version can be derived from the previous one using applied probability. It can be applied even if the consequence function is probabilistic, rather than deterministic. For example, if C, A , and S are all discrete (finite or countable) sets, with known probability $\Pr(s)$ for state s , and if the probability of consequence c occurring when act a is taken and the state is s is denoted by $\Pr(c | a, s)$, then the probability of each consequence c , if act a is selected, can be calculated by “marginalizing out” the state variable, via the following formula (expressing the law of total probability):

$$\begin{aligned}\Pr(c | a) &= \Pr(\text{consequence is } c \text{ if act } a \\ &\quad \text{is selected}) \\ &= \sum_s \Pr(c | a, s) \Pr(s).\end{aligned}\quad (1)$$

(If the state space is continuous, or mixed discrete and continuous, then the above sum

Table 1. A Simple Decision Problem, Represented by a Payoff Matrix

	State 1	State 2
Act A	2 people die	4 people die
Act B	3 people die	5 people die

is replaced by a Lebesgue–Stieltjes integral.) If there are too many states for such explicit enumeration and summation to be convenient, then *Monte Carlo sampling* (in which states are randomly sampled, and probabilities of consequences are estimated by recording how frequently they occur, for each choice of act) may provide a viable alternative way to estimate the probabilities of consequences for each act. $\Pr(c \mid a)$ is called a *consequence model*. The field of *risk assessment* provides constructive methods for estimating consequence models from data.

Most decision analysis formulations of decision problems assume that the DM has *preferences* for consequences and *beliefs* or *information* about states and/or about probabilities of consequences if different acts are chosen. Preferences for consequences are represented by an ordinal *value function*, that is, a function that assigns numbers to consequences so that higher numbers are assigned to preferred consequences. Consequences that are indifferent to each other are assigned the same value number. Examples of value functions include measures of net return from financial investments; lives saved (or life-years saved) by public health interventions; and the objective functions used in optimization problems. Beliefs about states are often represented by (perhaps subjective) probabilities calculated after conditioning on all relevant knowledge and information. Decision analysis then seeks to identify a *most-preferred act*, typically by constructing an ordinal value function for probability distributions over consequences (representing the acts), and then optimizing it, that is, finding an act that maximizes it.

This basic framework of acts, states, and consequences, underpins many specific decision models, developed in other articles. Table 2 lists some important examples. See also the articles titled *Average Reward of a Given MDP Policy* and *Total Expected*

Discounted Reward MDPs: Value Iteration Algorithm

in this encyclopedia. Creatively interpreting A , S , and C allows many powerful decision support and optimization methods to be brought to bear to assist “rational” (consequence-driven) decision making. All of these methods use predicted consequences of acts to optimize decisions.

SOME PRINCIPLES FOR SOLVING DECISION PROBLEMS

Principle 1: Dominance and Stochastic Dominance Criteria

The task of selecting a “best” or most-preferred act in A , given a consequence model $c(a, s)$ can sometimes be simplified by using the principle of *strict dominance*: if one act yields better (preferred) consequences than another for *every* state of nature, that is, no matter how uncertainties are resolved, then it should be preferred. The former act is said to *strictly dominate* the latter. Dominated acts can be eliminated from further consideration, thus simplifying the decision problem. For example, in Table 1, act A strictly dominates act B, as it causes fewer deaths in each state of nature; thus, act B can be eliminated. Such simplification via elimination can be applied even if the probabilities of different states are unknown or have not yet been assessed, and even if the value functions for consequences and acts have not yet been quantified. However, strictly dominated (or strictly dominant acts) do not exist, in many decision problems.

Example [A decision with no strictly dominant acts, solved by stochastic dominance]. Table 3 is a slight rearrangement of the example in Table 1. Now, neither act strictly dominates the other. The best choice depends on the probabilities of the two possible states, which we denote by p for the probability of state 1, and by $1 - p$ for the probability of state 2. If the DM is certain that state 1 will occur ($p = 1$), then the best decision is act A (causing 2 deaths instead of 5). If state 2 is certain to occur ($p = 0$), then the preferred act is B (causing 3 deaths instead of 4).

Table 2. Some Normative Decision-Modeling Frameworks

Framework/ Representation	Act	State	Consequence	Optimization Algorithms
<i>Normal form</i> • Consequence model $c(a, s)$ • State probability model $Pr(s)$ • Utility model $u(c)$	a = act (input to the consequence model)	s = state (input to consequence model)	$c(a, s)$ = consequence model (a matrix or bivariate function)	Mathematical programming: $\text{Max}_{a \in A} \sum_c u(c) \Pr(c a)$ s.t. $\Pr(c a) = \sum_s c(a, s) \Pr(s a)$
<i>Utility table</i> (also called a payoff matrix)	Row of matrix (rows represent the possible acts, in A .)	Column of matrix representing states in S	Value (or payoff) of consequence in cell (usually an NM utility or expected utility)	Eliminate dominated rows, then choose act to maximize expected utility (EU)
<i>Decision tree</i> [1]	Choice of an act at each decision node	Outcome at each chance node	Utilities at tips of tree	Use backward dynamic programming to make choices that maximize EU
<i>Influence diagrams and Bayesian networks</i> [2]	Choice of an act at each decision node	Outcome at each chance node	Value at value node	• Gibbs sampling • Bucket elimination • Graph algorithms (e.g., arc reversal)
<i>Markov decision process (MDP); semi-Markov decision process</i>	Choice of act in each state	Transition time and next state	Reward per unit time in states and/or at transitions; transition rates among states	• Value iteration • Policy iteration • Neurodynamic programming • Linear programming

(continued overleaf)

Table 2. (Continued)

Framework/ Representation	Act	State	Consequence	Optimization Algorithms
<i>Partially observable MDP (POMDP)</i>	Decision rule mapping observations to acts	Underlying state of process	Function of uncertain state and choice of acts, and/or transitions	Similar to MDP. (A POMDP is an MDP with probabilities of states for its state.) Grid and sampling algorithms for large problems
<i>Stochastic optimal control and robust control of uncertain systems</i>	Decision rule (maps observed history to choice of controlled input values)	Probability of next state, given history to date (and inputs)	Function of state and/or control trajectory	Stochastic dynamic programming; adaptive optimization, robust control, reinforcement learning algorithms
<i>Portfolio optimization</i>	Allocation of limited resources among risky investments	Yield from each investment	Return (or loss) from selected resource allocation	Large-scale mathematical programming, robust optimization, and stochastic optimization
<i>Simulation model</i>	Controllable inputs	States of model components	Function of component states and input sequence	Simulation-optimization heuristics and algorithms
<i>Unknown model</i>	Sequence of acts		Sequence of rewards	Adaptive learning, on-line, and minimal-regret algorithms

Table 3. A Payoff Matrix without Strict Dominance

	State 1	State 2
Act A	2 people die	4 people die
Act B	5 people die	3 people die
	$\Pr(\text{state 1}) = p$	$\Pr(\text{state 2}) = 1 - p$

If a DM has no objective basis for assessing p very precisely, but only knows (or judges) that state 1 is at least as probable as state 2, then what should he/she do? Strict dominance does not provide an answer. However, a different criterion called (first-order) *stochastic dominance* can be applied to conclude that any DM who prefers fewer deaths to more in this example should choose act A. The rationale is as follows. Suppose first that $p = 0.5$. (This is the case that is most favorable to selecting act B, given the constraint $p \geq 1 - p$, which is equivalent to $p \geq 0.5$.) Then, choosing act A gives equal chances of 2 or 4 people dying, while choosing act B gives equal chances of 3 or 5 people dying. Thus, the cumulative probability distribution function (CDF) for the number of deaths if B is chosen lies strictly to the right of the CDF for the number of deaths if A is chosen. Therefore, choosing A minimizes not only the expected number of deaths, but the probability that the number of deaths will not exceed any specified number. By this criterion, A is clearly superior to B when $p = 0.5$. If $p > 0.5$, then A is even more preferred to B, since increasing p shifts the CDF for B further to the right of the CDF of A. Thus, the criterion of stochastic dominance allows a clear choice between acts A and B, given partial knowledge of state probabilities (that $p \geq 1 - p$) and partial information about preferences for consequences (fewer deaths are preferred to more). Other dominance criteria (e.g., convex stochastic dominance, which eliminates acts that are dominated by a probability mixture of other acts [3]) have likewise been developed to eliminate some acts, given partial information about preferences and beliefs.

Principle 2: Expected Utility (EU) Criterion

After dominated alternatives have been identified and removed, a “best” act can be

selected from the remaining set of undominated alternatives using optimization methods, provided that preferences and beliefs can be modeled with enough consistency and precision so that a well-defined objective function (an ordinal value function for acts) can be constructed. *Expected utility (EU) theory* provides the best-known normative procedure for extending preferences for consequences to preferences (represented by a value function) for acts. It does so by calculating the EU of an act as follows

$$EU(a) = \sum_c \Pr(c | a)u(c). \quad (2)$$

Here, $u(c)$ is a number called the *utility* [or von Neumann–Morgenstern (NM) utility] of consequence c . In principle, $u(c)$ is defined so that the DM is indifferent between receiving consequence c with certainty and receiving the most-preferred consequence in C with probability $u(c)$, and otherwise receiving the least-preferred consequence in C . This is, admittedly, a rather theoretical and conceptual definition of utility. However, it can be proved that making choices that maximize EU is the *unique* procedure that is consistent with certain normative axioms [1] (e.g., the axioms of *reduction*, stating that choices should be based on probabilities for final consequences, not on interim resolutions of uncertainties; *weak ordering*, stating that alternatives can be consistently rank-ordered by preference, although with ties allowed; *independence*, stating that preferences among alternatives are not changed by uncertainty about whether an opportunity to choose among them will arise; and *continuity*, stating that a unique utility, $u(c)$, can be identified, by the above procedure, for any consequence, c .) Any procedure other than maximizing EU violates one or more of these axioms, giving EU theory a unique normative status as the “gold standard” for a theory of rational decision making.

Real people may have trouble assessing such indifference probabilities in hypothetical lotteries with useful coherence and consistency. In practice, therefore, decision analysts use a variety of other methods to quantify utility functions, $u(c)$, as discussed

in the related articles on *single-attribute utility theory* (SAUT) and *multiattribute utility theory* (MAUT). For example, if C is a set of possible monetary consequences, represented as an interval of the real line, then simple conditions (such as that a risk-averse DM be willing to either buy or sell an uncertain prospect for the same amount, interpreted as the “value” of the prospect) imply that the utility function must have a simple analytic form [such as $u(c) = 1 - \exp(-kc)$, where k expresses the DM’s coefficient of relative risk aversion]. Estimation of utility functions may then be simplified to estimating parameters (such as k), usually from forced-choice data in which the DM expresses preferences between pairs of simple prospects. In any case, the EU functional $EU(a)$ provides an ordinal value function for acts, constructed from the consequence model, $\Pr(c | a)$, and the utility function, $u(c)$. The corresponding normative theory of decision making (justified by axioms that imply it) prescribes choosing an act that maximizes $EU(a)$.

Example [EU Decision Making].

Problem: In Table 3, suppose that p is at most 0.2. Does this constraint give a risk-neutral DM (defined as one who optimizes the expected value of the consequence) enough information to select an act that maximizes EU (here assumed to be equivalent to minimizing expected deaths)?

Solution: The expected number of deaths if A is selected is $2p + 4(1 - p) = 4 - 2p$. The expected number of deaths if B is selected is $5p + 3(1 - p) = 3 + 2p$. Hence, the DM should prefer A to B only if $4 - 2p < 3 + 2p$, or, equivalently, if $p > 0.25$. Hence, knowing that p is at most 0.2 does indeed provide enough information to conclude that B is the EU-maximizing decision for a risk-neutral DM.

Principle 3: Satisficing and Heuristic Decision Making

The set of possible acts may be complex (e.g., all decision rules mapping observations to controlled inputs to a complex system) and

large (e.g., the set of all possible allocations of limited resources among competing investment opportunities). If searching for acts that have relatively high value is itself a costly or time consuming process, then heuristics may be used to abbreviate or simplify the search. This can lead to decisions that are suboptimal in principle, in that they do not necessarily truly optimize the DM’s objective function (or value function for acts), and yet that are pragmatic, in that they outperform other choices that can readily be identified and evaluated, and perhaps reduce the sum of search costs and costs from suboptimal decision making. *Satisficing* refers to decision making, in which the DM stops searching for a better alternative as soon as he/she finds one that is good enough (according to one or more criteria). *Fast, frugal heuristics* are simple decision rules (such as “Do what most others do,” or “Choose the first alternative that is clearly better than others on at least one attribute, and not obviously significantly worse than others on any attribute”) that often work well in familiar task environments, yet require relatively little deliberative thought or analysis [4]. In decision analysis, which seeks to clarify principles and techniques for improving and optimizing decisions, such pragmatic heuristics can usually be criticized, correctly, as leading to needlessly poor (dominated) choices, compared to those that would be produced by more thorough and rigorous analyses. Moving away from the prescriptions of EU theory typically exacts real losses in terms of obtaining less preferred outcomes more often. However, work on decision heuristics does draw attention to the importance, even for normative theories, of considering the costs (including cognitive burden) of decision making itself, and of balancing these against the consequences of suboptimal decisions.

EXTENSIONS OF THE BASIC ACT-STATE CONSEQUENCE FRAMEWORK

The fundamental decision problem, of choosing an act a from A to maximize some value function $V(a)$ [such as EU, if $V(a) = EU(a)$] can be generalized in the following ways.

- *Probabilities of states may depend on the DMs choice of act.* For example, a large company's decision to invest in a certain technology might change the probabilities that others make similar investments, which could affect the probabilities of different consequences for the first investor. If probabilities of states depend on the decision, a , then the probability of consequence c , given act a , is $\Pr(c | a) = \sum_s \Pr(c | a, s) \Pr(s | a)$. EU calculations and stochastic dominance comparisons can then be made using these consequence probabilities.
- *Preferences for consequences may depend on which state occurs.* In this case, the utility function $u(c)$ can be generalized to a state-dependent utility function, $u(c, s)$ [5,6]. State-dependent utilities can be used to model uncertain future preferences, or preferences for a defined set of consequences (such as returns from a financial investment portfolio) that depend on other random events or conditions, not included in the consequence set (such as the life or health of the recipient when the consequence is received).
- *Preferences for consequences may change, depending on previous acts and consequences.* Examples include acquired tastes, adaptation, habit formation, and addiction. Such changes in preferences over time can be modeled in various ways, including via state-dependent utility functions, or explicit dynamic utility functions.
- *Unknown and ambiguous probabilities.* As illustrated in the preceding examples, it may sometimes be necessary or convenient to make decisions with probabilities that are not precisely known, and for which there is no adequate factual basis for selecting a unique prior. [If a unique prior can be specified, then Bayesian inference can be used to condition it on data, producing an updated posterior distribution. Advances in computational methods, such as Markov chain Monte Carlo (MCMC) algorithms and importance

sampling, have made Bayesian inference computationally practical in a wide range of cases, even if convenient conjugate priors are not available, provided that a prior can be specified.] *Robust optimization* of decisions does not require a unique prior, but allows the state probabilities (or probability measures) to belong to a set of possibilities, called an *uncertainty set*. See also the article titled **Introduction to Robust Optimization** in this encyclopedia. Uncertainty sets are usually convex, and may be generated from constraints (such as upper and lower bounds on probabilities of particular states, or of particular events, corresponding to subsets of states.) Normative decision analysis has been extended to allow such sets of probabilities. The *multiple priors model* [7], originally used to model ambiguity-aversion in decision analysis with uncertain probabilities, but subsequently applied to robust control problems [8], provides normative axioms that imply that a DM should select an act from A to maximize the minimum possible value of EU, where the minimum possible value is found by minimizing EU over all probability measures in the uncertainty set.

- *Some acts may have intrinsic value, apart from their consequences.* Intervening to help a person in need, fulfill a duty, prevent an injustice, or right a wrong may carry satisfaction, and strongly affect preferences among acts, no matter what consequences follow. Less benignly, the act of gambling may carry a thrill, quite apart from its probable consequences. In such cases, the distinction between act and consequence is less clear than the idealized normative models suggest, and the separation implicit in the $c(a, s)$, $u(c, s)$, and $\Pr(c | a, s)$ notations may be more artificial than useful. The limited precision and detail used in describing acts and consequences for purposes of decision analysis may also obscure information that is crucial for making decisions. For example, in Table 1, it

seems obvious that act A is preferable to act B, since, no matter which state occurs, act A always results in 1 fewer people dying than would act B. But suppose that we add more detail to this description, by revealing that act B is “do nothing” and act A is “kill Mr. Jones and harvest his blood and organs, then use them to protect the others.” If act A is to sacrifice an innocent, unwilling, and healthy Mr. Jones for the benefit of others, then act A, far from being the obvious dominant choice that it seemed, may now be understood to be an unacceptable violation of Mr. Jones’s rights, with act B being vastly preferable.

- *Unknown consequence functions, and decision making without beliefs.* In new and unfamiliar environments, a rational agent may not at first have any beliefs—or at least no defensible and useful ones—about the probable consequences of alternative acts. Nothing in the A, S, C formalism dictates what should go into these sets. Then, empirical experimentation and learning (from one’s own actions and observed results, or from those of other agents, or both), rather than model-based prediction and optimization, may be crucial to an agent’s survival and well-being. In *evolutionary* and *learning* models of decision making, beliefs and preferences are not needed. Instead, the DM’s behavior is represented by a *decision rule*, mapping observations to probabilities of selecting acts. A DM may then use *reinforcement learning* (see also **Reinforcement Learning Algorithms for MDPs**), in which acts are made more probable when they are associated with larger average experienced rewards; and a population of DMs may undergo *evolution*, to develop decision rules that have high average reward and/or survival probabilities, compared to other possible decision rules. Learning and evolution enable emergence of highly effective (e.g., asymptotically optimal or near-optimal, using criteria such as zero average

regret) decision rules in many environments, even without models of beliefs about states [9,10]. However, the trial and error needed for learning and evolution may be slow, expensive, or potentially fatal. Although the modeling and calculations used in decision analysis typically require more information collection, cognition, and calculation than simple adaptive learning and evolution rules, they can reduce the costs and time to arrive at effective decision rules, at least if the costs of developing usefully accurate consequence models, $c(a, s)$ or $\Pr(c \mid a, s)$, are relatively small compared to the costs of learning or evolving decision rules. Moreover, in many important decision contexts (e.g., one-time major investment decisions), learning and evolution are not practical options, satisficing approaches are too costly, and decision analysis may be the only viable approach.

- *Uncertain acts, consequences, and future preferences.* Some important policy decisions, such as those with very diffuse and/or long-delayed consequences (e.g., regulation of environmental emissions, disposal of high-level radioactive wastes, or interventions to modify possible future climate) have consequences that current decision makers may never observe or know with any confidence, and yet that matter greatly. Moreover, how present decisions affect eventual consequences may depend in part on future acts and adaptations that are difficult or impossible to foresee—for example, because changes in tastes, technology, or understanding will reveal new options, currently unanticipated consequences, or new preferences that are not currently envisioned. Even, the opportunity set A_t of acts available t years in the future may depend on the (currently uncertain) state of the world. A growing literature in economics, management science, and marketing science attempts to extend the foundations and boundaries of decision analysis to include decisions with uncertain future opportunity sets,

consequences, and preferences using concepts such as the economic value of flexibility and reversibility of current decisions when future information becomes available [11–13].

CONCLUSIONS

The basic framework of acts, states, and consequences discussed in this article has provided a very useful foundation for many ingenious practical decision analysis models and algorithms (Table 2) that are now widely used in areas as diverse as economics, manufacturing and logistics, marketing, R&D and risky investments, control engineering, and environmental and public health risk management policy analysis. When the sets A , S , and C are clearly defined, well understood, and capture everything of interest to the decision maker, then decision analysis methods based on these foundations can help to overcome the many limitations and psychological traps of holistic intuitive judgment and decision making, and help practitioners identify better (e.g., dominant or stochastically dominant) decisions [14]. However, our limited ability to know and describe all relevant details of acts, states, and consequences, as well as of beliefs and preferences, in practice—especially future ones—provides a strong incentive to continue refining and extending the normative foundations of decision analysis. Practitioners currently emphasize the need to iteratively refine and check decision models to make sure that they truly capture all (and only) aspects of the situation that should significantly affect a decision, for example, as revealed by sensitivity analysis of simple approximate models [15]. Representing alternative acts by controlled stochastic processes that produce consequences (and perhaps future opportunities, preferences, and decisions) over time, rather than only by probability distributions for final consequences, appears to be a promising direction for generalizing and extending the foundations of decision analysis [13]. Some of the most exciting current directions of research for making decision

analysis better applicable to hard practical problems have been described in this article. After over half-a-century of axiomatic development and in-depth philosophical and mathematical discussion, the foundations of modern decision analysis remain an area of vibrant research. Although the act–state–consequence formalism has proved its value in an enormous range of business, engineering, and policy applications, it seems likely that the most useful and general normative decision theories are yet to come. These will probably build on the foundational work now being done on decision making without clearly known and understood acts, states, consequences, preferences, and beliefs.

SUPPLEMENTARY REFERENCES

Selected on-line references for decision optimization frameworks and algorithms in Table 2.

- Decision trees:
 - Game Trees for Decision Analysis—Shenoy (1996), <http://citeseer.ist.psu.edu/shenoy96game.html>
- Influence diagrams and Bayesian networks:
 - Sampling Methods for Action Selection in Influence Diagrams—Ortiz, Kaelbling (2000), <http://citeseer.ist.psu.edu/ortiz00sampling.html>
 - A Forward Monte Carlo Method for Solving Influence Diagrams—Charnes, Shenoy (2000), <http://citeseer.ist.psu.edu/charnes00forward.html>
 - A Simple Method to Evaluate Influence Diagrams—Xiang And Ye (2001), <http://citeseer.ist.psu.edu/ye01simple.html>
 - Learning Bayesian Networks with R—<http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/BottcherDethlefsen.pdf>; see also—<http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html>
- Markov decision processes (MDPs) and Partially observable MDPs (POMDPs):

- Reinforcement Learning for Factored Markov Decision Processes—Sallans (2002), <http://citeseer.ist.psu.edu/sallans02reinforcement.html>
- Symbolic Dynamic Programming for First-Order MDPs—Boutilier *et al.* (2001), <http://citeseer.ist.psu.edu/boutilier01symbolic.html>
- Speeding Up the Convergence of Value Iteration in POMDPs—Zhang, Zhang (2001), <http://citeseer.ist.psu.edu/zhang01speeding.html>
- Solving POMDP by On-Policy Linear Approximate Learning Algorithm—He, Shayman (1999), <http://citeseer.ist.psu.edu/335710.html>
- Optimal and robust control and reinforcement learning for uncertain and nonlinear systems:
 - Feedback Control Methodologies for Nonlinear Systems—Beeler, Tran, Banks (2000), <http://citeseer.ist.psu.edu/beeler00feedback.html> (for deterministic nonlinear systems)
 - An Overview of Industrial Model Predictive Control Technology—Qin, Badgwell (1997), <http://citeseer.ist.psu.edu/qin97overview.html>
 - <http://citeseer.ist.psu.edu/kaelbling96reinforcement.html>
 - <http://citeseer.ist.psu.edu/sutton98reinforcement.html>
 - <http://www.princeton.edu/~noahw/palgrave2.pdf> (introduces robust control)
- Simulation-optimization:
 - A Survey of Simulation Optimization Techniques and Procedures—Swisher, Jacobson *et al.* (2000), <http://citeseer.ist.psu.edu/517471.html>
 - Simulation Optimization of Stochastic Systems with Integer Variables by Sequential Linearization—Abspoel *et al.* (2000), <http://citeseer.ist.psu.edu/516176.html>
 - Simulation Optimization: Methods And Applications—Carson, Maria (1997), <http://citeseer.ist.psu.edu/carson97simulation.html>
 - <http://opttek.com/simulation.html> (overview and link to commercial software)
- Minimal-regret, on-line and adaptive learning algorithms:
 - Minimizing Regret: The General Case—Rustichini (1999), <http://citeseer.ist.psu.edu/rustichini98minimizing.html>
 - Adaptive Strategies and Regret Minimization in Arbitrarily Varying Markov Environments—Mannor, Shimkin, <http://citeseer.ist.psu.edu/467490.html>
 - Nearly Optimal Exploration-Exploitation Decision Thresholds—Dimitrakakis (2006), <http://citeseer.ist.psu.edu/dimitrakakis06nearly.html>
 - Combinatorial Online Optimization in Real Time—Grötschel, Krumke, Rambau, (2001), <http://citeseer.ist.psu.edu/448491.html>; see also <http://citeseer.ist.psu.edu/foster97regret.html>

REFERENCES

1. Raiffa H. Decision analysis: introductory lectures on choices under uncertainty. Reading (MA): Addison-Wesley; 1968.
2. Clemen RT. Making hard decisions: an introduction to decision analysis. 2nd ed. Boston (MA): PWS-Kent; 1996.
3. Lodwick LA. A generalized convex stochastic dominance algorithm. IMA J Manage Math 1989;2(3):225–246. <http://imaman.oxfordjournals.org/cgi/content/abstract/2/3/225>
4. Gigerenzer G. Adaptive thinking: rationality in the real world. Oxford: Oxford University Press; 2000.
5. Karni E. State-dependent preferences. 2005. Available at <http://www.econ.jhu.edu/people/Karni/sdp1.pdf>. Accessed in 2010.
6. Dreze JH, Rustichini A. State-dependent utility and decision theory. No 2000007. CORE Discussion Papers from Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). 2000. <http://www.core.ucl.ac.be/services/psfiles/dp00/dp2000-7.pdf>.

7. Gilboa I, Schmeidler D. Maxmin expected utility with non-unique prior. *J Math Econ* 1989;18:141–153.
8. Williams N. Robust control: an entry for The New Palgrave. 2009. Available at <http://www.ssc.wisc.edu/~nwilliam/palgrave2.pdf>. Accessed in 2010.
9. Young PH. Strategic learning and its limits. New York: Oxford University Press. 2004.
10. Sandholm WH. Population games and evolutionary dynamics. Cambridge (MA): MIT Press; 2010. <http://www.ssc.wisc.edu/~whs/book/index.html>.
11. Benjaafar S, Morin TL, Talavage JJ. The strategic value of flexibility in sequential decision making. *Eur J Oper Res* 1995; 82(3):438–457.
12. Walsh JW. Flexibility in consumer purchasing for uncertain future tastes. *Mark Sci* 1995;14(2):148–165.
13. Nehring K. Preference for flexibility in a Savage framework. *Econometrica* 1999;67(1): 101–119. <http://www.econ.ucdavis.edu/faculty/nehring/papers/flex1296.pdf>
14. Hammond JS, Keeney RL, Raiffa H. The hidden traps in decision making. *Clin Lab Manage Rev* 1999;13(1):39–47.
15. Howard R. The foundations of decision analysis revisited. In: Edwards W, Miles R, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. Chapter 3. <http://www.usc.edu/dept/create/assets/001/50843.pdf>

FOUNDATIONS OF SIMULATION MODELING

WAI KIN

VICTOR CHAN

Department of Industrial and Systems
Engineering, Rensselaer Polytechnic
Institute, Troy, New York

Simulation modeling is an important step of a simulation study. It concerns with the creation of an efficient simulation model as accurate as needed to mimic the behavior of the underlying system and to reproduce useful observations for subsequent analysis. This introductory article discusses three main simulation modeling methods: discrete-event simulation, continuous simulation, and agent-based simulation. In particular, three discrete-event simulation modeling paradigms are introduced: event scheduling, process interaction, and activity scanning. Several modeling issues are exemplified by a simple single-server queueing system and a slightly more complex multiple-server finite capacity queueing system with impatient customers. Continuous simulation is briefly discussed to draw the differences between discrete-event and continuous simulations. Simulating emergent behavior as a common purpose of agent-based simulation is demonstrated by a classic game of life model and a variation model. Monte-Carlo simulation and other simulation modeling formalisms are briefly highlighted.

INTRODUCTION

Basically, a simulation study involves three steps: modeling, analysis, and presentation. These three steps can be repeated until satisfaction. Modeling is about the creation of an efficient simulation model. Analysis here in a broad sense includes studies of inputs to and outputs from the simulation model, design of experiments, and optimization. The importance of the presentation step is sometimes

underestimated. Presentation is about how to effectively collect and present simulation results and how to communicate with relevant people to improve, verify, and validate (or accredit) the model and results. None of the three steps should be omitted in a professional simulation study.

This introductory article focuses on the first step—modeling. In particular, it deals with computer-based simulation modeling. A computer-based simulation model has the advantages of being low cost, fast, and repeatable. Beginners sometimes find modeling a time consuming task. This is perhaps due to unfamiliarity with either the modeling approach or the software package (or both). This article addresses the former by introducing different simulation modeling approaches. References for using software packages are provided in subsequent discussion. One mistake that beginners often make is the inclination of including as many details as possible into their simulation models. It should be noted that a simulation model is created to satisfy the need of a simulation study, rather than to duplicate the real system exactly. Therefore, the objective of the study matters in simulation modeling.

The modeling power of a simulation modeling approach and the functionalities of the underlying simulation package should also be evaluated separately. It is possible that a less convenient modeling approach implemented by a powerful software platform receives more attention than a more flexible and scalable modeling approach that was executed in a lower-level, less user-friendly programming language. This article does not review the usage, user interface, and functionalities of simulation packages.

Depending on the state of a system and the need of a simulation study, the corresponding simulation model can be discrete, continuous, or hybrid. Strictly speaking, a simulation model is discrete (continuous) if *all* its state variables are discrete (continuous). A model with both discrete and continuous state variables is a hybrid simulation model. There are

also several other ways of classifying simulation models including, static versus dynamic, deterministic versus stochastic, analytic versus numeric, and physical versus computer-based. For example, a simulation model is dynamic if time (simulation clock) is present and static otherwise. Monte-Carlo simulation is one type of static simulation models. Discrete-event and continuous simulations are usually dynamic simulation models [1].

Besides discrete, continuous, and Monte-Carlo simulations, a relatively new simulation paradigm—agent-based simulation—has received increasing attention in various fields over the past two decades. Basically, agent-based simulation, being a bottom-up approach, is capable of modeling systems with intelligent, heterogeneous, and autonomous entities. An agent-based simulation model can be discrete, continuous, or both depending on the type of its state variables. One of the main purposes of agent-based simulation is to simulate known or unknown macroscopic behavior emerging from microscopic interactions.

The rest of the article is organized as follows. The section titled “Discrete-Event Simulation Models” discusses three discrete-event simulation modeling approaches. The section titled “Continuous Simulation Models” outlines continuous simulation models. The section titled “Agent-Based Models” highlights several key concepts in agent-based simulation. The section titled “Monte-Carlo Models” briefly introduces Monte-Carlo simulation. The final section offers a conclusion.

DISCRETE-EVENT SIMULATION MODELS

One special property of a *discrete-event system* is that its state changes only at discrete instants of time. These instants of time represent the occurrences of certain events that change the state of the system and are referred to *event occurrence times* (or simply, *event times*). The system state remains constant in between event times because during which no events occur. This special property facilitates the modeling and analysis of discrete-event simulation because it

allows modelers to focus on the events, their relationships, and durations. To put it in a simple way, creating a discrete-event simulation model is to identify these elements and put them together in a correct manner. This will be demonstrated by examples in the following.

Several concepts in discrete-event simulation modeling are of importance: randomness, dynamics (time evolutions), future-event list (or pending events list or event calendar), and resident and transient entities.

Randomness is the uncertainty in the occurrences of events and durations. Time evolutions are the footprints of the system state with respect to time (both occurred events and their times). Randomness and time evolutions together differentiate a discrete-event simulation model from a standard computer program that is deterministic and static (i.e., no time involved). Almost all simulation packages provide simple mechanisms to implement randomness (i.e., by specifying distributions for event durations and relationships) and handle time evolutions (i.e., by providing access to the current and future simulated times).

A future-event list keeps track of the events scheduled to occur in the simulated future. This list is updated whenever a new event is scheduled to occur in the simulated future. This list is automatically generated and updated in most simulation packages. Users sometimes can also access and modify this list.

Resident entities are objects that reside in the model for a long period of time compared to the simulation run length, for example, a facility, a server, or a queue. Transient entities on the other hand appear in the model for a short period of time, for example, a job, a customer, or a message. The population of resident entities is usually much smaller than that of transient entities. In a typical queueing model, the number of servers is usually less than a hundred while the number of jobs/customers can be tens of thousands or millions. The large population of transient entities poses a computational burden if each of them is stored in memory and requires a certain amount of computation. A more

efficient modeling approach is to keep track of the aggregated information of the transient entities, such as the total number of jobs in queue or the total waiting time of jobs in the system, rather than the individual information, such as the individual waiting time of each job. Because this aggregated information is usually associated with some resident entities, e.g., the queue size, this approach is called the resident-entity approach. The drawback of this approach is that the individual characteristics of the transient entities are lost. Nevertheless, this approach has been shown to significantly improve simulation speed and memory usage [2, 3]. Indeed, it is sometimes possible to avoid memorizing individual characteristics and still be able to obtain the same accuracy

in the simulation results. This is demonstrated in the following example.

A discrete-event simulation model can be created using different world views, which were first discussed in Kiviat [4] and Lackner [5]. Three popular world views: event scheduling, process interaction, and activity scanning, are introduced in the following. Before that, let us first outline a typical algorithm for executing a discrete-event simulation model. This algorithm is essentially an event-scheduling algorithm. Most software packages employ a similar execution algorithm. If one prefers not to use existing simulation packages, he/she can also implement the algorithm by using a general-purpose programming language, such as C/C++.

Discrete-Event Simulation Execution Algorithm

Initialize

Do until stop condition is satisfied

 Advance clock to the time of next event

 Execute event

 Perform actions: change state variables, schedule events, collect statistics, etc.

Event Scheduling

We will use a single-stage multiple-server queueing system with finite capacity and impatient customers (i.e., $G/G/c/K$) to illustrate several basic modeling issues of discrete-event simulation. In this system, there is only one type of transient entities (called *customers*) and they arrive at the system in accordance with an arbitrary stochastic process. Each arriving customer is admitted if the current queue size is less than its capacity (denoted by K). Upon joining the queue, each customer starts a random impatient time period; if he/she does not receive service by the end of the impatient time, he/she will depart the system without being served. The c servers are identical with no vacation and breakdown. This simple model can represent many real-world systems, for example, a bank waiting line with multiple clerks, or a computer communication system (or a call center) in which the entities are the messages or requests stored in buffer and the

servers are the parallel CPUs (or operators) handling these messages/requests.

A good starting point of building a discrete-event simulation model is to have a textual description of a simplified model and try to identify and combine the key events, event relationships, and conditions into a desired logic. After the simplified model is finished, other features can be added to augment the simplified model to create the final model. We will employ this practice to first develop a single-server queueing system with infinite capacity and no impatient customers, and then modify this simplified model to model the original system. A textual description of the single-server queueing simulation is given in the following.

“A simulation **begins** by having the first job **enter** the system. New jobs **enter** the system at random intervals of arrivals. When a job **enters**, *if the resource is available* the job can **start** service. Once a job **starts**, it will **leave** after a service time delay. Whenever

a job **leaves**, *if there is more work waiting*, then service can **start** on the next job” [6].

Almost all information necessary to build this single-server model is embedded in above paragraph. To see this, we highlight the key words in three different formats: bold face, underline, and italics. The words in bold face represent the events: “enter,” “start,” “leave,” and the initial event “begin,” which only executes once. The underlined words denote the random time durations. Here, “random intervals of arrivals” are the interarrival times of the arriving entities, and “service time delay” is the service time duration of each service. The italicized words signify event-scheduling conditions. To build the actual model, circle the events and add arcs starting from the corresponding preceding event and ending at the corresponding succeeding event in each sentence (see Fig. 1). These arcs represent the event relationships. The delay time on each arc is specified by the underlined words if present, or zero otherwise. The condition on each arc is given by the italicized words on that sentence if present, or unconditional otherwise. The resulting simulation model (the right panel of Fig. 1) is an event-scheduling-based simulation model, which is represented by a simulation modeling language called *event relationship graphs*, or simply *event graphs* [7]. Two state variables (Q and R) and two random variables (a and s) are needed: Q returns or sets the current number of customers in the queue, R returns or sets the current number of available servers, a

returns an interarrival time generated from a given distribution (i.e., $a = F_a^{-1}(u)$, where u is a uniform random number between 0 and 1), s returns a service time duration generated from a given distribution (i.e., $s = F_s^{-1}(u)$).

We now ready to extend this model to our final multiple-server system with queue capacity and impatient customers. Two important discrete-event simulation modeling techniques are used: Boolean expressions and look-ahead capability. To model the queue capacity, we note that when a customer arrives, the queue size will increase by one if the current queue size is less than K (i.e., the customer is admitted), and remains unchanged otherwise (i.e., the customer is rejected). This logic can be implemented by using the following Boolean expression:

$$Q = Q + (Q < K), \quad (1)$$

where $(Q < K) = 1$ if the condition is true and zero otherwise.

To model impatient customers without storing the waiting times of all customers (i.e., without resorting to the transient entity approach), we use the look-ahead capability to obtain information in the simulated future. For ease of exposition, we will use a two-server queueing system (i.e., $c = 2$) to illustrate the idea (it can be extended to more than two servers). The idea is as follows. When a customer arrives (say customer i), if he knows the maximum duration he needs

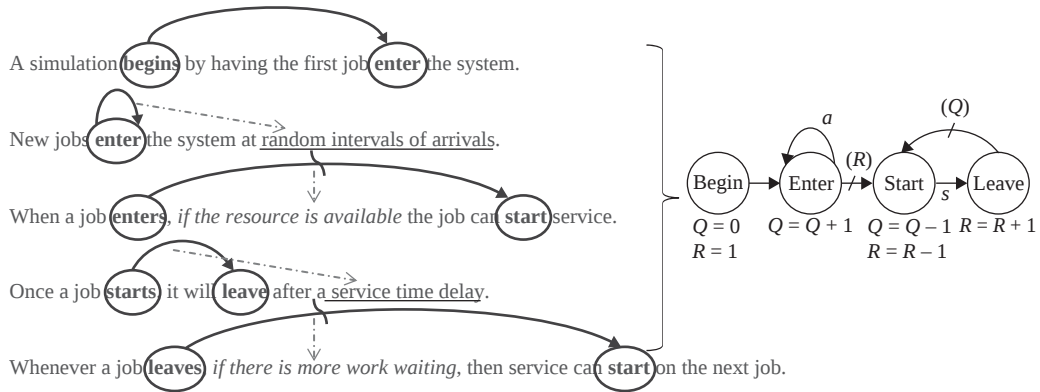


Figure 1. From textual description to simulation model.

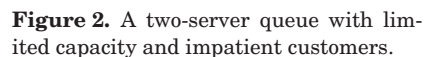
In particular, let D be the decision variable of whether to stay or not (stay if $D == 1$ and leave if $D == 0$). Let T_c be current simulation time, which can be obtained by using some system function in most simulation packages (e.g., CLK in Sigma [6] and TNOW in Arena [8]). Let p be the impatient time period of a customer generated from a given distribution, that is, $p \sim F_p(p)$. The state changes that take place when an Enter event occurs are as follows:

The first equation says that a customer will stay (i.e., $D == 1$) if the queue size is less than K (i.e., $(Q < K) == 1$) and the time

The state changes under the Stay event are

The first equation simply generates and stores the service time duration of a customer in s . The second equation computes the actual waiting time of the customer. It is equal to 0 if the arriving customer finds an available server (i.e., $T_1 - T_c < 0$). Otherwise, it is equal to the amount of time before one of the servers is available (i.e., $T_1 - T_c$). The notation “ $(.)^+$ ” represents the positive part of the waiting time (i.e., $\max\{T_1 - T_c, 0\}$). The third and the last equations update the two earliest available times of the servers due to this new arriving customer.

One remark observed from this model is regarding the trade-off between speed and memory. Sometimes, a small amount of auxiliary memory (variables) can greatly reduce



the speed or complexity of the simulation model, such as this one.

Process Interaction

Another popular simulation modeling paradigm is the process-interaction approach. Loosely speaking, for simple systems the process interaction approach is WYSIWYG (*What You See Is What You Get*) in the sense that what you create in the model is what you observe in the real simple system (see Fig. 3 and explanation in next paragraph). (However, this intuitive view diminishes when the complexity of the system increases). This intuitive view is attractive to beginners because it deals directly with the actual objects of a real system. Together with their animation capability, process-interaction-based simulation packages are by far the most common in industry.

To illustrate the intuitive view of process-interaction models, we model the single-server queueing system by using three process-interaction modules to mimic the three processes of the real system: (i) the process of customer arrivals, (ii) the process of queueing customers for service, and (iii) the process of customer departures. We then connect the three modules sequentially so that the entities (customers) can flow through the three modules as they do in the real system. Figure 3 shows the implementation of this model using the simulation package Arena 11 [8]. The mapping between the three modules and the three real processes are also shown in the figure.

To create the model in Arena, just drag the modules from a built-in template of Arena, paste them into the model workspace, connect them based on the job flow, and specify

the data for each module (such as interarrival time and service time distributions, and defining the server) by double clicking the modules. See Kelton *et al.* [8] for details of using Arena.

Two issues of the process-interaction approach should be mentioned: deadlock and virtual entities. Deadlock could occur when two entities are trying to seize the same resources in a wrong sequence. For example, consider two entities (A1 and A2) and two machines (M1 and M2). Let us suppose that A1 is holding M1 and is trying to seize M2 to start a process. At the same time, E2 is holding M2 and needs M1 to start another process. In this situation, unless some mechanism is introduced to break the conflict, both entities and machines will be waiting forever. Therefore, acquisition of resources should be carefully planned when building a process-interaction model. One good rule in practice is to seize *all* required resources at *one* time. In the case where several resources must be seized sequentially (e.g., Process 1 requires only Machine 1 and then immediately Process 2 requires both Machines 1 and 2), then release the previously seized resources first and then seize all needed resources at one time (e.g., release Machine 1 first and then seize both Machines 1 and 2).

Virtual entities are needed when one needs to model some abstract mechanisms or logics that have no physical counterparts in the real system. This is because almost everything in a process-interaction model (including activities, logics, controls, and assigning values to variables) must be driven by flowing entities. For instance, to simulate a traffic intersection, the right-of-way can be modeled as a “virtual resource” and

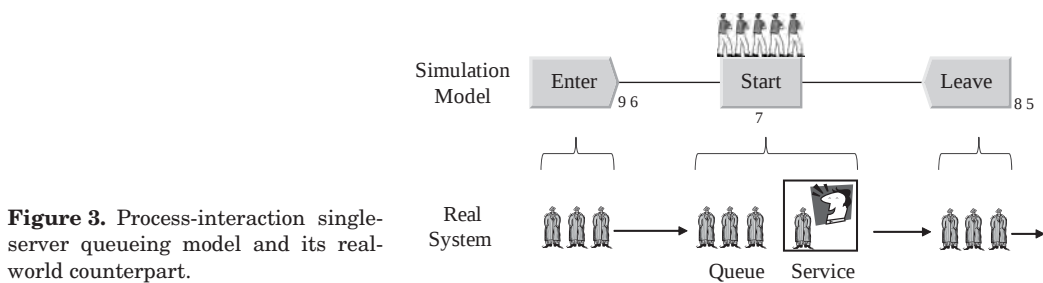


Figure 3. Process-interaction single-server queueing model and its real-world counterpart.

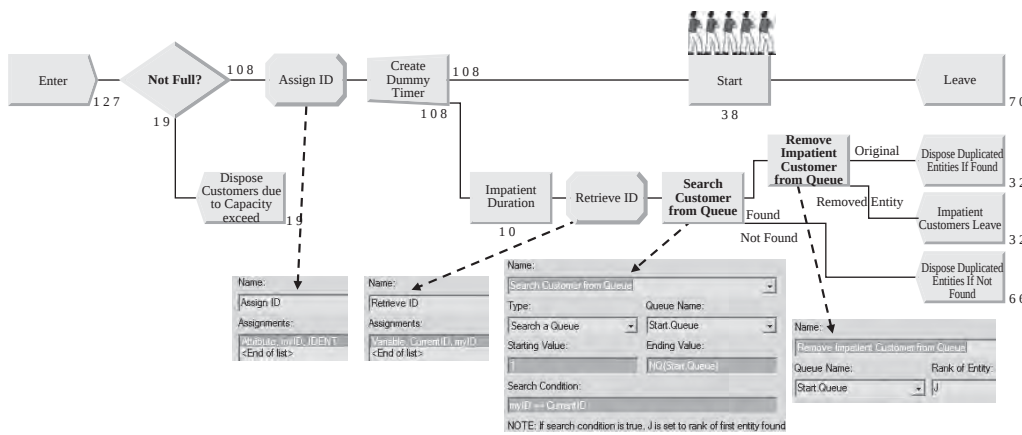


Figure 4. Process-interaction model of a capacitated queue with multiple servers and impatient customers.

traffics from any directions will need to seize this virtual resource in order to pass the intersection.

Another example of “virtual entities” appears in the multiple-server queueing system with impatient customers. In a process-interaction model, an entity in queue is a passive object in the sense that it cannot proactively trigger activities by itself when certain condition is met. Some mechanism must be created to trigger the corresponding activity. For example, to trigger the departure of an impatient customer whose impatient time has expired, a virtual entity is needed. We use the Arena model shown in Fig. 4 to explain this issue.

An arriving customer, if accepted by the diamond “Not Full?” module based on the current queue size, is assigned an ID in the “Assign ID” module. This customer is then duplicated before joining the queue (the “Start” module). This duplicate is a virtual entity and will stay in the “Impatient Duration” module for a random impatient duration. When the impatient period expires, this virtual entity first identifies itself in the “Retrieve ID” module (because there could be other virtual entities representing other customers). It then enters the “Search Customer from Queue” module to see if the customer that it represents is still waiting in the queue. If not found, the customer either is being served or has left, and this

virtual entity can be disposed. If found, the virtual entity will enter the “Remove Impatient Customer from Queue” module and trigger the departure of the customer from the queue. Both the virtual entity and impatient customer (if found) are then disposed.

From a computational point of view, the process of searching a customer from a queue could be inefficient when the queue size is large. Moreover, the search is done for every customer regardless of whether the customer has left or not. A more efficient approach such as Fig. 2 is therefore recommended for large-scale congested systems.

Other situations where virtual entities are needed include the modeling of resource preventive maintenance and failures, or random events that could change the state of the system. Indeed, virtual entities, when appropriately and professionally handled, can be a powerful modeling technique.

Activity Scanning

An activity-scanning simulation model works by scanning activities of the model and *fires* (i.e., activates) the ones that meet certain given conditions. One of the main benefits of the activity-scanning approach is its analytical property, which facilitates the construction of mathematical models and theoretical analysis. The main representative approach of activity scanning is Petri-net

models [9]. A Petri net is a bipartite graph (i.e., a graph with two types of elements and no two elements of the same type connect to each other). The two types of elements are circles (called *places*) and bars (called *transitions*). Besides these two types of elements, users can also add tokens to the graph to make the Petri net run. Tokens, which are solid dots, are used to represent objects (e.g., jobs or machines) and model logics (e.g., failure events). Each transition represents a certain activity of the model. All the input places to a transition denote the conditions required for the transition to fire. All the output places from a transition are the results of the firing of the transition. In a standard Petri net, a transition is enabled when there is at least one token in each input place. After a time delay associated with the transition (which can be random, deterministic, or zero), the transition fires and one token is removed from each input place. Immediately, one token is added to each output place. The number of tokens entering and leaving from a transition may or may not equal. If the number of tokens in any place is always below a certain number at any time, the Petri net is called *bounded*, otherwise, *unbounded*. At any time of the simulation, a snapshot of the token distribution (their locations) gives the current state of the simulation. This token distribution is called a *marking*.

The aforementioned firing rule is called *remove-on-fire*, in that the tokens are removed when the transition fires. Another firing rule is *remove-on-enable*, which removes the tokens when the transition is enabled. It can be shown that these two rules are equivalent.

To build a Petri-net simulation model, the following four-step procedure can be used:

1. Identify all activities, including zero time activities (one transition for each activity) of the system,
2. Identify the conditions for each activity (one place for each condition),
3. Connect the places and transitions to make the net alive according to the model logic, and
4. Add tokens to represent the initial state of the system.

Fig. 5 presents the four-step implementation of the Petri-net simulation model of the single-server queueing system without capacity and no impatient customers. There are two activities in this simple system: arrivals of customers and services for customers, which are modeled by the two transitions, E and S (which respectively stand for “Enter” and “Start”). The single token in Place A denotes the arrival stream of customers (i.e., the process of arrivals). Every a random time units, Transition E fires; it then places one token to Place Q, denoting an arrival of a customer to the queue, and also puts one token back to Place A, initiating the arrival of the next customer. The token in Place R represents the single machine which is available right now (adding more tokens here extends the model to a multiple-server system). Transition S has two input places, signifying the two conditions for a service to start (i.e., an available server and a waiting customer). When these two conditions are satisfied, Transition S is enabled; and after s random time units, it fires, denoting the completion of a service. One token is then placed in Place R, which allows the server to work on the next customer if there is any. One token is also deposited to Place L, which represents a processed customer.

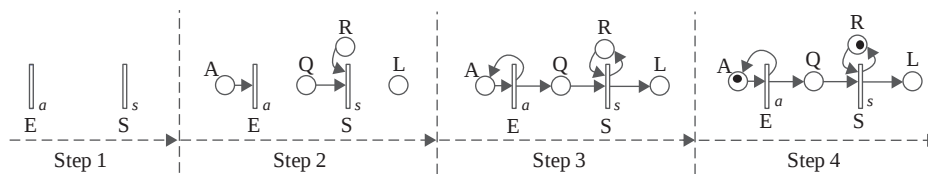


Figure 5. Activity scanning model of a single-server queueing system.

Enrichments of Petri nets include using colored tokens to represent different types of entities, allowing multiple tokens to be removed from or deposited into a place, and using inhibitor arcs to resolve conflictions by assigning priorities to different activities. The inclusion of inhibitor arcs increases the modeling power of Petri nets to a Turing machine.

Other Discrete-Event Modeling World Views

A discrete-event system can also be characterized by a state machine represented by a state-transition diagram. A state machine includes all the system states, transitions between states, and the corresponding actions. For example, in the single-server queueing system, the state could be the number of customers in the system, the transitions are the two events “arrival of a new customer” and “departure of finished customer,” and the actions are either to increase the number of customers in the system by one (when the arrival transition occurs) or decrease the number of customers in the system by one (when a departure transition occurs). There is no need to include the transition for the start of a service because the state is the number of customers in the system, which does not change when a service begins. There is a mapping between the states, transitions, and actions of a state machine and the state variables, events, and conditions and state changes of an event relationship graph. Therefore, the modeling powerful of an event relationship graph is equivalent to a state machine (or a Turing machine) (see a formal proof in Savage *et al.* [10]).

Besides graphical representations, a discrete-event can also be represented formally as a mathematical model. One example is the formalism called *Discrete event system specification* (DEVS). The DEVS formalism describes a discrete-event system by defining all the events, how the states change, how the times of different activities advance, how the events change the states and vice versa both externally and internally [11]. Another well-known discrete-event modeling formalism is the generalized semi-Markov process (GSMP). GSMP is a formal way of

describing the dynamics of a discrete-event system. It facilitates mathematical analysis of the underlying system [12]. Another representation of discrete-event systems is based on optimization models. In this representation, events are represented as decision variables, event relationships are modeled by constraints, event times are captured by the right-hand-sides of the constraints, and event scheduling (i.e., execute an event as soon as feasible) is realized by the objective function of the optimization model. The benefits of the optimization representation include using dual variables for sensitivity analysis, identifying deadlocks by examining feasibility of the optimization model, obtaining more efficient simulation models by eliminations of redundant constraints [13].

CONTINUOUS SIMULATION MODELS

Continuous simulation models simulate systems with state variables that change continuously in time. One popular type of continuous simulation is *system dynamics*. This method is suitable for systems with feedback structures and for modeling high-level of abstraction of a real system. In a system dynamics simulation model, the state of the system is represented by stocks and the rate of change in the state as flows. The stock variables accumulate all past state values and therefore, represent the current system state (e.g., $x(t)$ and $y(t)$ in the example below). The flow variables (can also be parameters) are measured per unit of time such as the changes in a state variable per second (e.g., the growth and death rates in the example below). Another type of continuous simulation models frequently used in modeling and simulating queueing systems is stochastic fluid (flow) models. Continuous simulation has been widely used in many fields, including electrical engineering, chemistry, biology, mechanical engineering, business, and so on.

One simple continuous simulation model is the Lotka-Volterra’s predator-prey model, which models the competition between two types of populations: prey and predators [1]. Let $x(t)$ and $y(t)$ denote, respectively, the number of individuals in the prey and

predator populations at time t . Each population is determined by two quantities, the *pure change rate* of the population itself and the *interaction rate* due to the interaction with the other population. The pure change rate of the prey is their growth rate in the absence of the predators (denoted by a), while the pure change rate of the predators is their death rate (or starvation rate) due to the absence of the prey (denoted by $-b$, which is negative because it is decreasing). The interaction rate of the prey is the decrease in their numbers due to the presence of the predators (denoted by $-cy(t)$, which is proportional to the number of predators) and the interaction rate of the predators is the increase in their population due to the presence of the prey (denoted by $dx(t)$, which is proportional to the number of prey).

This model assumes exponential growth (or death) in both populations in the absence of the other, that is, the prey's growth is $x'(t) = ax(t)$ and the predators' death is $y'(t) = -by(t)$. The interaction between the two populations are captured by the product of their population sizes, that is, the decrease in the prey due to the predators is $-cx(t)y(t)$ and the increase in the predators due to the prey is $dx(t)y(t)$. Combining the pure change and interaction terms for each population gives two differential equations

$$x'(t) = ax(t) - cx(t)y(t), \quad (2)$$

$$y'(t) = -by(t) + dx(t)y(t). \quad (3)$$

This simple differential-equation model can be solved exactly to obtain analytical solutions for $x(t)$ and $y(t)$. With boundary conditions and given values for all parameters, one can determine (i.e., simulate) the system state at any time t . There is no randomness in this simple model. To model randomness, one can include (e.g., add) random factors, which can also be functions of t , to above equations, for example, Gaussian noises. This will increase the complexity of the model. Indeed, simulation models of many real-world systems are so complex that no closed-form solutions can be obtained. Numerical methods, such as the Runge-Kutta method, can be used to find approximate solutions [14].

AGENT-BASED MODELS

In this section, we first discuss the connection between agent-based simulation and the two simulation approaches: discrete-event and continuous simulations. We then introduce some concepts of agent-based simulation (such as agents and interactions) and outline a general algorithm for executing an agent-based simulation model. We exemplify these concepts by using a simple agent-based simulation model.

Loosely speaking, an agent-based simulation model is a hybrid discrete-continuous simulation model with proactive, autonomous, and intelligent entities. Recall that entities in a discrete-event simulation model are rather simple with limited capabilities, and can only react to state changes. In the process-interaction queueing model with impatient customers, the customer entities cannot actively leave the system when their impatient times expire. A separated mechanism by using virtual entities is needed to trigger the departure of the impatient customers. The servers in this model are also reactive. For instance, if we model vacations or breakdown (e.g., sick), then separated mechanisms (such as vacation timers or failure timers) will be needed.

Agent-based simulation extends the capability of entities to allow autonomy and intelligence. The resulting autonomous objects (called *agents*) are able to proactively interact with other agents and the environment in which they situate. If an agent-based simulation model includes both discrete and continuous state variables, it is a hybrid discrete-continuous simulation model with proactive autonomous entities.

The popularity of agent-based simulation is driven by the increasing complexity of real-world systems and its capability of handling these complex systems. A real-world complex system often contains a large number of interacting units that are autonomous, goal-driven, and adaptive (i.e., "*agents*"). For example, a large-scale health-care network may consist of a number of units from large facilities such as hospitals, medical supply warehouses, blood banks to small service units such as ambulances,

operating rooms, and from personnel such as doctors and nurses to patients and other service providers. All these units can be considered as autonomous agents. For example, warehouses react to demands by providing inventory based on a certain rule, ambulances operate in accordance with some dispatch policies and physical mechanical rules, and patients try to intelligently select doctors. In many cases, agents, while operating based on their own strategies, cooperate or compete with each other in accordance with certain mutual rules or agreements.

It is natural to model agents as objects in a computer program and simulate their interactions to see what interaction rules lead to what consequences (i.e., emergent behavior). Mechanisms or policies can then

be enforced to discourage bad behavior and promote good ones. Emergent behavior here means arising coherent patterns of macroscopic behavior resulting from microscopic actions and interactions. This leads to the agent-based simulation approach, which has been developed rapidly over the past two decades. It has been widely used in various fields including military, biology, social science, economics, business, and so on. The theoretical basis of agent-based simulation lies mainly in *cellular automata* [15–17], *complex system modeling*, *artificial life*, and *swarm intelligence* [18,19].

The algorithm for executing an agent-based simulation model varies greatly across different agent-based platforms. A general scheme is outlined below.

Agent-Based Simulation Execution Algorithm

Initialize

Do until stop condition is satisfied

 Advance clock by a given amount of time (e.g., a tick count)

 Update all continuous state variables

 For each agent (select an agent randomly or based on a specific order)

 Perform actions: change state, communicate with other agents or environment, etc.

We now use the *Conway's Game of Life* model to illustrate the modeling of agents, their interactions, and the resulting emergent behavior. Imagine that there is a two-dimensional grid of size $n \times n$, where n is the number of rows (or columns) in the grid. Each entry of the grid lives a cell. The cells cannot move. Each cell has only two states: alive and dead. Each cell changes its state based on the following three simple rules that describe its interactions with its eight neighbors (up, down, left, right, and four diagonal cells). See the left panel of Fig. 6 for a visual description of the rules.

1. A live cell with exactly two or three live neighbors will remain alive in the next time step.
2. A dead cell with exactly three live neighbors will come to life in the next time step.

3. Otherwise, the cell will die either of loneliness or crowdedness in the next time step.

Implementing this model in a standard agent-based simulation package (such as Repast [20] or NetLogo [21]) and running it for about one hundred iterations yields a distribution of relatively stable patterns of live cells as shown in the right panel of Fig. 6. Note that the middle panel of Fig. 6 is the initial distribution of live cells.

In fact, this example is so simple that it can be considered as a cellular automaton. A cellular automaton model considers a set of cells residing in a two-dimensional lattice. Moreover, this model only includes discrete state variables and therefore can also be modeled by using the discrete-event simulation approach.

To demonstrate the flexibility of this model in arriving different emergent behaviors, we generalize the original Game of Life model

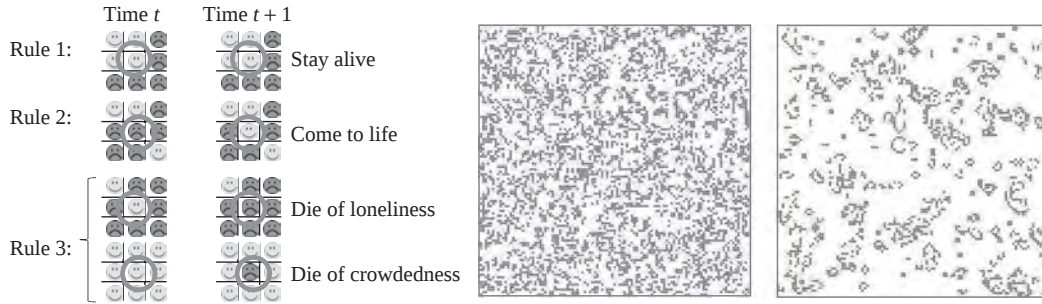


Figure 6. Agent-based simulation model of Conway's game of life.

by simply changing the number of live neighbors in Rules 1 and 2 to X , Y , and Z as defined below. To better visualize the emergent patterns, we also color the cells based on the number of live neighbors (see the legend of Fig. 7).

The variation of Conway's game of life model is as follows:

1. A live cell with the number of live neighbors between X and Y , inclusively, will remain alive in the next time step,
2. A dead cell with exactly Z live neighbors will come to life in the next time step,
3. Otherwise, the cell will die either of loneliness or crowdedness in the next time step,

where $0 \leq X, Y, Z \leq 8$ and $X \leq Y$.

This variation has a total of 288 possible combinations of rules, (X, Y, Z) . Some of them are trivial but some produce clear emergent patterns as shown in Fig. 7. These patterns also evolve in time. One can also use a sequence of multiple rules to obtain different combinations of patterns. The patterns shown in Fig. 7 are obtained by using this sequence of rules: $(0, 7, 8) \rightarrow (0, 7, 3) \rightarrow (0, 7, 6) \rightarrow (0, 7, 5) \rightarrow (0, 7, 4) \rightarrow (0, 5, 4) \rightarrow (3, 5, 4)$, where the first rule is to initialize the states of all cells to dead.

Although cellular automata are already Turing complete, agent-based simulation allows agents to have a greater degree of flexibilities and functionalities (such as learning and adaptation) and can accommodate more complicated interaction rules both between agents and agents, and between

agents and their environment. Agents can also have internal memories and perceptions to respond to outside and adjust their inside. Agent-based simulation also allows different kinds of spatial environments, such as ring, lattice, random, or maps (such as GIS maps). Agents can move freely in its environment. This makes it applicable for modeling and visualizing complex behavior in physical systems, such as simulations of traffic, evacuations, biological systems, mechanical systems, and so on.

One issue in agent-based simulation is the order in which the agents' states are updated in each time step. If one agent is updated before its neighbors are updated, then its state at the current time step can affect the states of its neighbors at this time step, otherwise, its state at the current time step will affect its neighbors in the next time step. One common practice is to randomize the updating order.

The computational requirement of the agent-based simulation execution algorithm is high because it has to loop through all agents in each time step. Methods have been proposed to avoid looping all agents, for example, only updating those agents whose state changes matter. Another approach is to employ the discrete-event simulation algorithm to advance the clock based on events and only update the continuous variables in every time step. Different hardware (such as parallel CPUs or graphics processing units—GPU) have also been used to improve the efficiency of agent-based simulation models. It has been shown that execution speed can be improved by thousands of times when using GPU [22].

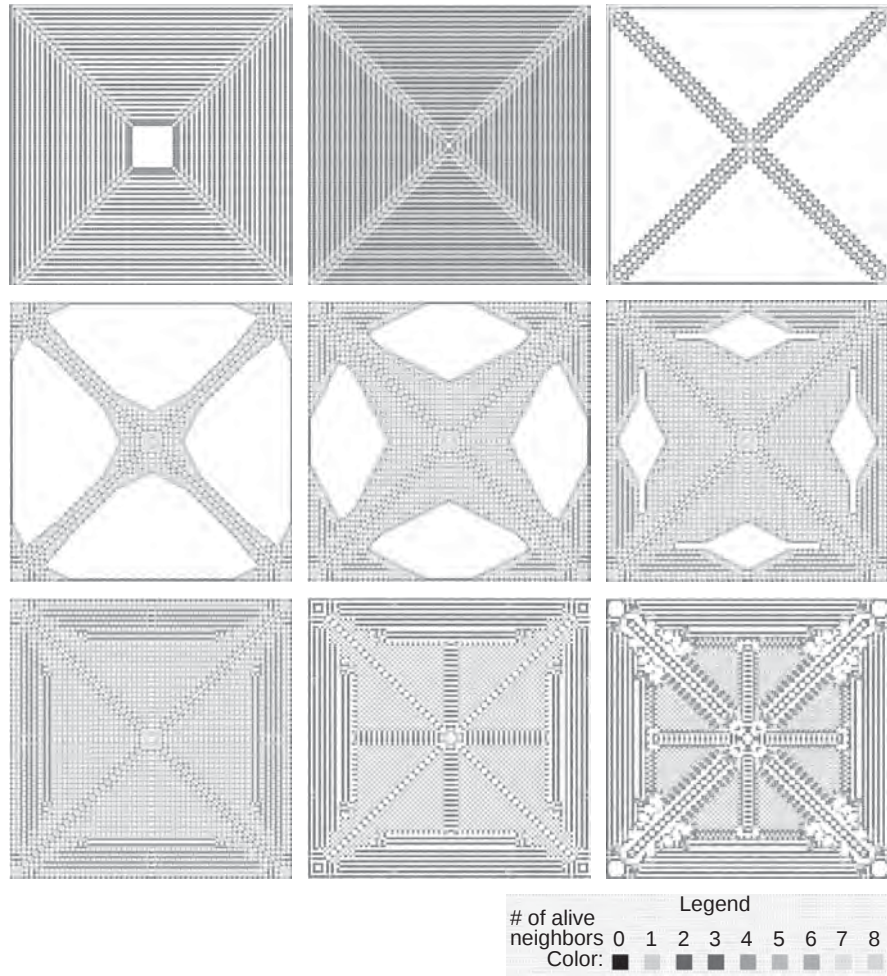


Figure 7. Emergent behavior of variation of Conway's game of life model.

Monte-CARLO MODELS

Monte-Carlo simulation approach is based on repeated sampling from an algorithm, a set of equations, or a given computer model. Sampling here means two steps: (i) generate random numbers and (ii) evaluate the target values using the generated random numbers. In general, the sampling targets can be static or dynamics, continuous or discrete, or deterministic or stochastic. Sampling from a deterministic target is needed when the response space of the target is large and only a certain amount of computing budget is allocated to sample a fraction of the whole space.

The relationship between Monte-Carlo simulation and other simulations such as discrete-event or continuous simulations is subtle and depends on how one defines Monte-Carlo simulation. If one considers Monte-Carlo simulation as a static model that does not involve time or represents a system at a particular time (as discussed in the section titled "Introduction"), then Monte-Carlo simulation is a special case of other simulations that involve time. The relationship turns around if one allows the sampling target to evolve in time.

While different definitions of Monte-Carlo simulations can be found in the

literature, one popular definition is by Law (2006) in which Monte-Carlo simulation is defined as “a scheme employing random numbers...used for solving certain stochastic or deterministic problems where the passage of time plays no substantive role.” This definition is well accepted in the discrete-event simulation area as it clearly differentiates discrete-event dynamic simulation and static pure-sampling-based simulation.

Monte-Carlo simulation has various applications in physics, mathematics, and finance, such as simulating molecular dynamics, global warming impacts, earthquakes, solving diverse mathematical problems (from numerical integration to Markov chains, and from heuristics optimization algorithms to games), and evaluation and prediction (e.g., simulation of Econometrics models such as geometric Brownian motion, diffusion models, and GARCH models) [23].

CONCLUSION

This introductory article introduces basic modeling concepts in discrete-event simulation, continuous simulation, and agent-based simulation. It also highlights the relationship among these modeling approaches as well as different world views of discrete-event simulation models.

Besides these modeling approaches, new simulation modeling methods are also being proposed and developed. Classical approaches tend to focus on aspects of efficiency, easy-to-learn, and animation-friendly. New approaches are moving toward the need of an integrated simulation framework, in which modeling, analysis, and presentation are integrated. How a simulation model is created directly influences experimentations (analysis), and data collections and interpretation (presentation), which in turn affect the validity (or credibility) of the simulation model and results. Therefore, the construction of a simulation model should be tied to the corresponding analysis and presentation based on the need of the underlying simulation study.

REFERENCES

1. Law A. Simulation modeling and analysis. 4th ed. Boston: McGraw-Hill; 2006.
2. Law A, Kelton WD. Simulation modeling and analysis. 3rd ed. Boston: McGraw-Hill; 2000.
3. Schruben LW, Roeder TM. Fast simulations of large-scale highly congested systems. *Simulation* 2003;79(3):115–125.
4. Kiviat PJ. Digital computer simulation: computer programming languages. Santa Monica (CA): The Rand Corporation; 1969. Memorandum RM-5883-PR.
5. Lackner MR. Digital simulation and system theory. Santa Monica (CA): System Development Corporation, SDC SP- 12; 1964.
6. Schruben D, Schruben LW. Graphical simulation modeling using SIGMA, Ithaca (NY): Custom Simulation; 2001.
7. Schruben L. Simulation modeling with event graphs. *Comm ACM* 1983;26(11):957–963.
8. Kelton WD, Sadowski RP, Sturrock DT. Simulation with Arena. 4th ed. New York (NY): McGraw-Hill; 2007.xxiv,668.
9. Peterson JL. Petri net theory and the modeling of systems. New Jersey: Prentice-Hall; 1981.290.
10. Savage EL, Schruben LW, Yucesan E. On the generality of event graph models. *INFORMS J Comput* 2005;17(1):3–9.
11. Zeigler BPA. Theory of modelling and simulation. Melbourne: Krieger Publishing Company; 1984. p. 460. Reprinted.
12. Glynn P. A GSMP Formalism for Discrete Event Systems, Proceedings of the IEEE; 1989.
13. Chan WKV, Schruben LW. Optimization models of discrete-event system dynamics. *Oper Res* 2008;56:1218–1237.
14. Cellier F, Kofman E. Continuous system simulation. New York: Springer; 2006.
15. Von Neumann J. Theory of self-reproducing automata. Champaign (IL): University of Illinois Press; 1966. p. 388.
16. Beyer WA, Sellers PH, Waterman MS. Stanislaw M. Ulam’s contributions to theoretical theory. *Lett Math Phys* 1985;10:231–242.
17. Wolfram S. Universality and complexity in cellular automata. *Phys Nonlinear Phenom* 1984;10D:1–35.
18. Bonabeau EA, Dorigo MA, Theraulaz GA. Swarm intelligence: from natural to artificial systems. New York: Oxford University Press; 1999.

19. Macal CM, North MJ. Tutorial on agent-based modeling and simulation part 2: how to model with agents. Proceedings of the 2006 Winter Simulation Conference. Monterey (CA), Piscataway (NJ): Institute of Electrical and Electronics Engineers; 2006.
20. North MJ, Collier NT, Vos JR. Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans Model Comput Simulat* 2006;16(1):1–25.
21. Wilensky U, NetLogo. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston (IL). 1999.
22. Lysenko M, D'Souza RM. A framework for megascale agent based model simulations on graphics processing units. *J Artif Soc Soc Simulat* 2008;11(4):10.
23. Hammersley JM, Handscomb DC. Monte Carlo methods. London: Methuen; 1964.

FRITZ–JOHN AND KKT OPTIMALITY CONDITIONS FOR CONSTRAINED OPTIMIZATION

APRIL ANDREAS
Engineering and Mathematics,
McLennan Community
College, Waco, Texas

In this article, we introduce Fritz–John and Karush–Kuhn–Tucker (KKT) optimality conditions for constrained optimization, following in large part the discussion of this material in Bazaraa *et al.*[1]. We first review some critical definitions that may be already familiar and further define particular sets related to feasible and improving regions for an optimization problem, establishing some basic relationships among the same. Next, we show a simple and direct derivation of the Fritz–John optimality conditions, which when present ensure optimality of a given point $\bar{x}$. We then show how to derive the KKT conditions for optimality. Next, we provide a generalization of Fritz–John and KKT theory, with necessary and sufficient conditions for optimality, and conclude with some simple numerical examples.

DEFINITIONS

Consider an optimization problem of the form

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \quad \forall i \in 1, \dots, m \end{aligned}$$

where f and g_i are differentiable functions, and where x is a finite-dimensional real vector.

We begin first with several definitions. Recall the definitions for convexity and concavity of $f(x)$, summarized in the table below, where each condition must hold for all feasible vectors x_1 and x_2 , $0 < \tau < 1$, to fit the related definition.

Define F at some point $\bar{x}$ as the set of all descent directions d , that is, $f(\bar{x} + \lambda d <$

$f(\bar{x}) \quad \forall \lambda \in (0, \varepsilon)$, for some ε . That is, $d \in F$ if the objective initially decreases by moving from $\bar{x}$ along direction d .

Definition	Condition
$f(x)$ is convex	$f(\tau x_1 + (1 - \tau)x_2) \leq \tau f(x_1) + (1 - \tau)f(x_2)$
$f(x)$ is strictly convex	$f(\tau x_1 + (1 - \tau)x_2) < \tau f(x_1) + (1 - \tau)f(x_2)$
$f(x)$ is concave	$f(\tau x_1 + (1 - \tau)x_2) \geq \tau f(x_1) + (1 - \tau)f(x_2)$
$f(x)$ is strictly concave	$f(\tau x_1 + (1 - \tau)x_2) > \tau f(x_1) + (1 - \tau)f(x_2)$

If we also define $F_0 = \{d : \nabla f(\bar{x})^T d < 0\}$ and define $F'_0 = \{d \neq 0 : \nabla f(\bar{x})^T d \leq 0\}$, it is clear that $F_0 \subseteq F \subseteq F'_0$. Note that $F_0 = F$ if and only if f is convex, and that $F = F'_0$ if and only if f is strictly concave.

Define G at some point $\bar{x}$ as the set of all feasible directions d , that is $\bar{x} + \lambda d$ is feasible $\forall \lambda \in (0, \varepsilon)$, for some ε .

Let all the binding constraints from $i = 1, \dots, m$ be enumerated in the set I . Now, if we also define $G_0 = \{d : \nabla g_i(\bar{x})^T d < 0 \quad \forall i \in I\}$ and $G'_0 = \{d \neq 0 : \nabla g_i(\bar{x})^T d \leq 0 \quad \forall i \in I\}$, it is also clear that $G_0 \subseteq G \subseteq G'_0$. Note that $G_0 = G$ if and only if g_i is strictly convex $\forall i \in I$, and that $G = G'_0$ if and only if g_i is concave $\forall i \in I$.

The following derivations will rely also on Gordon's lemma and Farkas' lemma, which are two theorems of the alternative. Gordon's lemma states that there exists a solution to exactly one of the following two systems:

$$\text{System I : } Ax < 0$$

$$\text{System II : } A^T y = 0, y \geq 0, y \neq 0,$$

and Farkas' lemma states that there exists a solution to exactly one of the following two systems:

$$\text{System I : } Ax \leq 0, c^T x > 0$$

$$\text{System II : } A^T y = c, y \geq 0.$$

DERIVATION OF FRITZ-JOHN OPTIMALITY CONDITIONS

With respect to the optimization problem introduced in the previous section, it is clear that if $\bar{x}$ is a local minimum, then $F \cap G = \emptyset$, since no directions are both feasible (G) and improving (F). Since $F_0 \subseteq F$ and $G_0 \subseteq G$, we can also state that $F_0 \cap G_0 = \emptyset$. If $F_0 \cap G_0 = \emptyset$ then there is no direction d such that

$$\begin{pmatrix} \nabla f(\bar{x})^T \\ \nabla g_{i_1}(\bar{x})^T \\ \nabla g_{i_2}(\bar{x})^T \\ \vdots \\ \nabla g_{i_p}(\bar{x})^T \end{pmatrix} d < \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (1)$$

where $\nabla f(\bar{x})$ and $\nabla g_{i_1}(\bar{x}), \dots, \nabla g_{i_p}(\bar{x})$ are column vectors, and where $I = \{i_1, \dots, i_p\}$. If we let

$$A = \begin{pmatrix} \nabla f(\bar{x})^T \\ \nabla g_{i_1}(\bar{x})^T \\ \nabla g_{i_2}(\bar{x})^T \\ \vdots \\ \nabla g_{i_p}(\bar{x})^T \end{pmatrix},$$

we can then rewrite Equation (1) as $Ad < 0$. Since there is no solution to $Ad < 0$, Gordon's lemma implies that there must be a solution to $A^T u = 0$, that is,

$$(\nabla f(\bar{x}) \quad \nabla g_{i_1}(\bar{x}) \quad \dots \quad \nabla g_{i_p}(\bar{x})) \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_p \end{pmatrix} = 0, \quad (2)$$

for some $u \geq 0, u \neq 0$. Equation (2) can be rewritten as $u_0 \nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) = 0$, where $u_0 \geq 0, u_i \geq 0 \forall i \in I, u \neq 0$.

A summary of this logic follows:
 $\bar{x}$ is a local minimum $\rightarrow$

$$F \cap G = \emptyset \rightarrow$$

$$F_0 \cap G_0 = \emptyset \leftrightarrow$$

there exists no d such that $Ad < 0 \leftrightarrow$; there is a solution to $u_0 \nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) = 0$, where $u_0 \geq 0, u_i \geq 0 \forall i \in I, u \neq 0$.

We may now formally define a Fritz-John point $\bar{x}$ as one that satisfies the following conditions for some vector u .

Primal feasibility

- $\bar{x}$ is feasible

Dual feasibility

- $u_0 \nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) = 0, u \geq 0, u \neq 0$

Complementary slackness

- $g_i(\bar{x})u_i = 0 \forall i = 1, \dots, m$

The implication is that if $\bar{x}$ is locally optimal, then it is a Fritz-John point, and hence the three conditions above are *necessary* conditions. There are two single-direction implication arrows in the first two lines of the logical progression. Our next task will be to work this logical progression backwards, and by this means we will derive *sufficient* Fritz-John conditions for a point to be optimal.

First, examine the statement $F \cap G = \emptyset \rightarrow F_0 \cap G_0 = \emptyset$. The reverse statement is true when $F_0 = F$ and $G_0 = G$. We noted in the definitions section that $F_0 = F$ if and only if f is convex and $G_0 = G$ if and only if g_i is strictly convex $\forall i \in I$. If we make these assumptions, we can conclude that $F_0 \cap G_0 = \emptyset \rightarrow F \cap G = \emptyset$.

Next, examine the statement that $\bar{x}$ is a local minimum $\rightarrow F \cap G = \emptyset$. We cannot assume that $F \cap G = \emptyset$ implies that $\bar{x}$ is a local minimum, because we may be ignoring nonlinear feasible directions. However, if we have assumed that g_i is strictly convex $\forall i \in I$, then the feasible region is locally convex and we can restrict our attention to straight-line directions. By assuming a locally convex feasible region,

we can conclude that $F \cap G = \emptyset \rightarrow \bar{x}$ is a local minimum. Therefore, if $f(x)$ is convex and all binding constraints are strictly convex, then a Fritz–John point is a local minimum.

DERIVATION OF KKT OPTIMALITY CONDITIONS

We begin our discussion of KKT conditions by following some of the same logic used for deriving the Fritz–John conditions. Again, note that if $\bar{x}$ is a local minimum, then $F \cap G = \emptyset$, and since $F_0 \subseteq F$, we can also state that $F_0 \cap G = \emptyset$. If we assume that g_i is concave $\forall i \in I$, then $G = G'_0$, and we can conclude that $F_0 \cap G'_0 = \emptyset$. Therefore, there is no direction d such that

$$\nabla f(\bar{x})^T d < 0 \text{ and } \nabla g_i(\bar{x})^T d \leq 0 \forall i \in I.$$

Writing the first inequality as $-\nabla f(\bar{x})^T d > 0$, we can employ Farkas' lemma to guarantee that there exists a vector $u \geq 0$ such that

$$\nabla g_i(\bar{x})u_i = -\nabla f(\bar{x})^T \forall i \in I.$$

We can rewrite the above conditions as the following:

$$\begin{aligned} \sum_{i=1}^m u_i \nabla g_i(\bar{x}) &= -\nabla f(\bar{x}) \\ u_i &\geq 0, i \in 1, \dots, m, \end{aligned}$$

which is equivalent to the Fritz–John points described above when $u_0 = 1$.

We now summarize this logic as follows:
 $\bar{x}$ is a local minimum $\rightarrow$

$$\begin{aligned} F \cap G &= \emptyset \rightarrow \\ F_0 \cap G &= \emptyset \leftarrow \\ F_0 \cap G'_0 &= \emptyset \leftrightarrow \end{aligned}$$

there exists no d such that $\nabla g_i(\bar{x})^T d \leq 0 \forall i \in I$ and $-\nabla f(\bar{x})^T d > 0 \leftrightarrow$ there is a vector $u \geq 0$ such that $\nabla g_i(\bar{x})u_i = -\nabla f(\bar{x})^T \forall i \in I$.

The third arrow is bidirectional if, for instance, g_i is concave $\forall i \in I$.

Formally, we say that $\bar{x}$ is a KKT point if there exists a vector u such that the following conditions hold.

Primal feasibility

- $\bar{x}$ is feasible

Dual feasibility

- $\nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) = 0, u \geq 0$

Complementary slackness

- $g_i(\bar{x})u_i = 0 \forall i = 1, \dots, m$

We have shown that a local minimum is a KKT point if constraints are concave. Constraint concavity is just one possible constraint qualification under which KKT conditions are necessary for optimality; see the article titled **Constraint Qualifications** for an expanded discussion. We now pursue this logic backwards. If KKT conditions are met, then we can state that $F_0 \cap G = \emptyset$. If we also assume that f is convex, then $F_0 = F$ and we can state that $F \cap G = \emptyset$. Finally, if we assume that the feasible region is locally convex, then $F \cap G = \emptyset$ implies that $\bar{x}$ is a local minimum. Therefore, a KKT point is a local minimum if f is convex and all binding constraints are convex. Note that if all constraints are convex, then a KKT point is a global optimum.

COMPARISON OF FRITZ–JOHN AND KKT POINTS

In the previous discussion, we identified a KKT point $\bar{x}$ as a Fritz–John point where $u_0 = 1$. That is, all KKT points are Fritz–John points, but not all Fritz–John points are KKT points.

Recall that if $\bar{x}$ is a Fritz–John point, the following dual feasibility condition must hold true:

$$\sum_{i=1}^m u_i \nabla g_i(\bar{x}) = -u_0 \nabla f(\bar{x}), u \geq 0, u \neq 0.$$

There are two options for u_0 : either $u_0 = 0$ or $u_0 \neq 0$. In the former case, we can scale the dual solution by dividing each element of u by u_0 to obtain a vector in which $u_0 = 1$,

which in turn proves that $\bar{x}$ is in fact a KKT point.

If $u_0 = 0$, then the dual feasibility condition becomes

$$\sum_{i=1}^m u_i \nabla g_i(\bar{x}) = 0, u \geq 0, u \neq 0. \quad (3)$$

Equation (3) implies that any one vector $\nabla g_i(\bar{x})$ can be written as a linear combination of the remaining vectors; that is, vectors $\nabla g_i(\bar{x})$ are linearly dependent.

Suppose however that the binding constraint gradients are known to be linearly independent. This removes the possibility that u_0 could equal zero, leaving that u_0 must be greater than 0, and hence that the Fritz-John conditions are equivalent to the KKT conditions. In this case, the KKT conditions are necessary for local optimality.

FURTHER GENERALIZATION

Now consider the following optimization problem, which includes a set of equality constraints:

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \quad \forall i = 1, \dots, m \\ & h_i(x) = 0 \quad \forall i = 1, \dots, l \end{aligned}$$

where f, g_i , and h_i are differentiable functions. We can expand the definition of Fritz-John points for these problems to include any point $\bar{x}$ such that the following conditions hold.

Primal feasibility

- $\bar{x}$ is feasible

Dual feasibility

- $u_0 \nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) + \sum_{i=1}^l v_i \nabla h_i(\bar{x}) = 0, u \geq 0, u \neq 0, v$ unrestricted

Complementary slackness

- $g_i(\bar{x})u_i = 0 \quad \forall i = 1, \dots, m$

Note that if we simply replace each equality constraint $h_i(x) = 0$ with two inequalities $h_i(x) \leq 0$ and $h_i(x) \geq 0$, then every feasible $\bar{x}$ solution is a Fritz-John point, thus making the Fritz-John conditions extremely weak.

We summarize the relationship with Fritz-John points and local optima below:

- If $\bar{x}$ is a local minimum, then it is a Fritz-John point.
- If $\bar{x}$ is a Fritz-John point, it is a local minimum if all of the following conditions are met:
 - f is convex;
 - binding $g_i(\bar{x})$ are strictly convex;
 - h_i are affine $\forall i = 1, \dots, l$;
 - $\nabla h_i(\bar{x})$ are linearly independent.

Similarly, we can expand the definition of KKT points to include any point $\bar{x}$ such that the following conditions hold.

Primal feasibility

- $\bar{x}$ is feasible

Dual feasibility

- $\nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) + \sum_{i=1}^l v_i \nabla h_i(\bar{x}) = 0, u \geq 0, u \neq 0, v$ unrestricted

Complementary slackness

- $g_i(\bar{x})u_i = 0 \quad \forall i = 1, \dots, m$

We summarize the relationship with optimality below:

- *Necessary.* If $\bar{x}$ is a local minimum, then it is a KKT point if
 - binding $\nabla g_i(\bar{x})$ are linearly independent;
 - all $\nabla h_i(\bar{x})$ are linearly independent.
- *Sufficient.* If $\bar{x}$ is a KKT point, then it is a local minimum if
 - f is convex;
 - binding $g_i(\bar{x})$ are convex;
 - $h_i(\bar{x})$ are convex if $v_i > 0$, concave if $v_i < 0$, and unrestricted if $v_i = 0$.

Also, as an alternative to the necessary conditions above, if all f_i are concave and all h_i are affine, then a local minimum must be a KKT point.

EXAMPLE USING FRITZ–JOHN THEORY

Consider the following optimization problem:

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \geq 2 \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

Since Fritz–John conditions are necessary for local optimality, and since this problem is clearly bounded (and an optimal solution exists), we can determine an optimal solution by finding all the Fritz–John points, and choosing one having the smallest objective as an optimal solution. First, we rewrite the problem as follows:

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & -(x_1 - 1)^2 - (x_2 - 1)^2 \leq -2 \\ & -x_1 \leq 0 \\ & -x_2 \leq 0 \end{aligned}$$

we then establish the following conditions.

Primal feasibility

$$\begin{aligned} \bullet \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \geq 2 \\ & x_1 \geq 0, x_2 \geq 0 \end{aligned}$$

Dual feasibility

$$\begin{aligned} \bullet \quad & \begin{pmatrix} -2(\bar{x}_1 - 1) \\ -2(\bar{x}_2 - 1) \end{pmatrix} u_1 + \begin{pmatrix} -1 \\ 0 \end{pmatrix} u_2 \\ & + \begin{pmatrix} 0 \\ -1 \end{pmatrix} u_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} u_0 \end{aligned}$$

Complementary slackness

$$\begin{aligned} \bullet \quad & u_1(-(x_1 - 1)^2 - (x_2 - 1)^2 + 2) = 0 \\ & u_2(-x_1) = 0 \\ & u_3(-x_2) = 0 \end{aligned}$$

Again, there are two options for u_0 : either $u_0 = 0$ or $u_0 > 0$ (with $u_0 = 1$ without loss of generality in the second case). If $u_0 = 0$, we must have $u_1 > 0$ since $\nabla g_2(\bar{x})$ and $\nabla g_3(\bar{x})$ are linearly independent. Since $u_1 > 0$, constraint 1 must be binding:

$$(\bar{x}_1 - 1)^2 + (\bar{x}_2 - 1)^2 = 2.$$

Since $x_1 = x_2 = 1$ is infeasible, we know that $\nabla g_1(\bar{x}) \neq 0$ for any feasible $\bar{x}$, and so at least

one of u_2 and u_3 must be positive. First, suppose that $u_2 > 0$, and so constraint 2 must be also binding, which implies that $\bar{x}_1 = 0$. Therefore

$$(0 - 1)^2 + (\bar{x}_2 - 1)^2 = 2 \rightarrow \bar{x}_2 = 0 \text{ or } \bar{x}_2 = 2.$$

Now, we determine whether $(\bar{x}_1, \bar{x}_2) = (0, 0)$ or $(\bar{x}_1, \bar{x}_2) = (0, 2)$ is a Fritz–John point:

$$\begin{aligned} (\bar{x}_1, \bar{x}_2) = (0, 0) : \\ \begin{pmatrix} 2 \\ 2 \end{pmatrix} u_1 + \begin{pmatrix} -1 \\ 0 \end{pmatrix} u_2 + \begin{pmatrix} 0 \\ -1 \end{pmatrix} u_3 &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \rightarrow u_2 = 2u_1, u_3 = 2u_1 \end{aligned}$$

Choosing u_1 to be any positive value, we have that $(0, 0)$ is a Fritz–John point.

$$(\bar{x}_1, \bar{x}_2) = (0, 2) :$$

$$\begin{aligned} \begin{pmatrix} 2 \\ -2 \end{pmatrix} u_1 + \begin{pmatrix} -1 \\ 0 \end{pmatrix} u_2 + \begin{pmatrix} 0 \\ -1 \end{pmatrix} u_3 \\ = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rightarrow \text{has no solution} \end{aligned}$$

Therefore, $(0, 2)$ is not a Fritz–John point when $u_0 = 0$.

The case of $u_3 > 0$ is symmetric, and does not reveal any further Fritz–John points. Next, if $u_0 = 1$, we have the following:

$$\begin{aligned} \begin{pmatrix} -2\bar{x}_1 + 2 \\ -2\bar{x}_1 + 2 \end{pmatrix} u_1 + \begin{pmatrix} -1 \\ 0 \end{pmatrix} u_2 + \begin{pmatrix} 0 \\ -1 \end{pmatrix} u_3 \\ = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \end{aligned}$$

If $u_1 = 0$, then $u_2 = u_3 = 1$ and g_2 and g_3 must be binding, so we again have the point $(\bar{x}_1, \bar{x}_2) = (0, 0)$. Continuing this process, we compute the following four points:

- $\bar{x} = (0, 0)$ —all three constraints are binding: $f(\bar{x}) = 0$;
- $\bar{x} = (0, 2)$ —constraints 1 and 2 are binding: $f(\bar{x}) = 2$;
- $\bar{x} = (2, 0)$ —constraints 1 and 3 are binding: $f(\bar{x}) = 2$;
- $\bar{x} = (2, 2)$ —constraint 1 is binding: $f(\bar{x}) = 4$.

Since our objective is of the minimization sense, $\bar{x} = (0, 0)$ is optimal.

EXAMPLE USING KKT THEORY

Consider the following optimization problem:

$$\begin{aligned} \min \quad & -x_1 \\ \text{subject to} \quad & x_1^2 + x_2 \leq 4 \\ & x_1 + x_2 = 2 \end{aligned}$$

We first show that $\bar{x} = (2, 0)$ is a KKT point. Observe that $g(x) = x_1^2 + x_2 - 4$ and $h(x) = x_1 + x_2 - 2$. To show that $(2, 0)$ is a KKT point, we must show that it satisfies all primal feasibility, dual feasibility, and complementary slackness conditions.

Primal feasibility

- $x_1^2 + x_2 \leq 4$
 $x_1 + x_2 = 2$

Dual feasibility

- $\begin{pmatrix} -1 \\ 0 \end{pmatrix} + \begin{pmatrix} 2x_1 \\ 1 \end{pmatrix} u + \begin{pmatrix} 1 \\ 1 \end{pmatrix} v = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

Complementary slackness

- $(x_1^2 + x_2 - 4)u = 0$

We can satisfy these conditions by setting $\bar{x} = (2, 0)$, $u = 1/3$, and $v = -1/3$; therefore, $\bar{x} = (2, 0)$ is a KKT point.

Since f is convex, the binding $g_i(\bar{x})$ are convex, and $h_i(\bar{x})$ are affine and linearly independent, we can conclude that $\bar{x}$ is a KKT point and is locally optimal. Furthermore, since the feasible region is convex (in addition to having a convex objective function), a local optimal solution is a global optimal solution, and so $\bar{x} = (2, 0)$ is globally optimal.

REFERENCES

1. Bazaraa MS, Sherali HD, Shetty CM. Nonlinear programming: Theory and algorithms. Hoboken (NJ): John Wiley & Sons; 1993.

FURTHER READING

- Bazaraa MS, Jarvis JJ, Sherali HD. Linear programming and network flows. Hoboken (NJ): John Wiley & Sons; 1990.
- Karush W. Minima of functions of several variables with inequalities as side constraints. [M.Sc. Dissertation]. Chicago (IL): Department of Mathematics, University of Chicago; 1939.
- Kuhn HW, Tucker AW. Nonlinear programming. Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press; 1951. pp. 481–492.
- Mangasarian OL. Nonlinear programming (Classics in applied mathematics). Madison (WI): Society for Industrial Mathematics; 1987.
- Mordecai A. Nonlinear programming: Analysis and methods. New York (NY): Dover Publishing; 2003.
- Nocedal J, Wright SJ. Numerical optimization. New York (NY): Springer; 2000.

FUZZY MEASURES AND INTEGRALS IN MULTICRITERIA DECISION ANALYSIS

MICHEL GRABISCH
University of Paris I, Paris School
of Economics, Paris, France

INTRODUCTION

Most of the methods in multicriteria decision making (MCDM) use the weighted arithmetic mean (WAM) (weighted sum) when criteria have to be aggregated. This simple aggregation function, used from the ancient Greeks, is universally used and has the advantage of simplicity and readability, in the sense that any decision maker without special knowledge can understand the meaning of its parameters, namely the weights on the criteria. However, the drawbacks of the weighted sum have been pointed out many times: for instance, its sensitivity to extreme values and, more importantly, its inability to explore the concave parts of the Pareto frontier in multiobjective optimization problems. On a more practical side, the weighted sum is unable to translate behaviors in aggregation such as favor high scores; focus on weak points; and eliminate alternative if criterion 3 is not well satisfied; if criteria 1 and 2 are satisfied, do not overestimate their importance. The first two examples in the previous list are still relatively simple, as they can be handled by a weighted sum on the ordered list of scores in increasing order: this is called the ordered weighted average (OWA), and it contains, as particular cases, the median, the minimum, and the maximum. However, the last two examples are more complex and cannot be treated, neither by the weighted sum nor by an OWA. They are typical examples of *interaction* between criteria. Interaction means that the way the score on a criterion is handled depends on the value of the scores on the other criteria. A typical example

given in Ref. 1 is the following: suppose that some students are evaluated on three topics, which are mathematics, physics, and literature, with an emphasis on scientific topics. There is some interaction between mathematics and physics in the sense that, as usually students good at maths are also good at physics and vice versa, a good mark in both math and physics should not be overestimated with respect to literature. On the contrary, as scientific topics are more important, bad marks in both mathematics and physics become very important. The existence of interaction phenomena, as well as the inability of the weighted sum to represent them, has been recognized from a long time ago, and the absence of an adequate tool to consider them has provoked the appearance of the following general rule for the practitioner: *choose/define criteria so that they are independent (no interaction)*. Obviously, this rule cannot always be applied, or leads to excessive simplification.

The appearance of the Choquet integral in MCDM has permitted to solve this problem, and has brought a precise definition of what interaction is. The Choquet integral not only generalizes both the weighted sum and the OWA but also can model the examples of interaction described earlier. So far, many theoretical studies and real applications have been done (see the overviews [1–3] and [4, Chapter 5] for further references and details).

FUZZY MEASURES AND INTEGRALS

We consider a set of criteria $N = \{1, \dots, n\}$, and denote the set of its subsets by 2^N . An alternative is an element of $X = X_1 \times \dots \times X_n$, where X_i is the set of values for attribute i . We write $x = (x_1, \dots, x_n)$ for $x \in X$. We make the assumption that the utility functions $u_i : X_i \rightarrow \mathbb{R}$ have been determined, up to a positive affine transformation.

A *fuzzy measure* [5] or *capacity* [6] on N is a mapping $\mu : 2^N \rightarrow [0, 1]$ satisfying the following two conditions:

1. $\mu(\emptyset) = 0$, $\mu(N) = 1$.
2. $A \subseteq B$ implies $\mu(A) \leq \mu(B)$ (monotonicity).

In applications in MCDM, $\mu(A)$ is often interpreted as the *weight of importance* of the group or coalition $A \subseteq N$. This is, however, an imprecise statement, and it should be interpreted in a more rigorous way as follows. Suppose that for each criterion $i \in N$ there exist two levels with an absolute meaning, for example, a *neutral level*, denoted by $\mathbf{0}_i$, and a *satisfactory level*, denoted by $\mathbf{1}_i$ (like in the MACBETH methodology [7]). These levels are elements of X_i , and represent, respectively, a level that is neither good nor bad and a level that would make the decision maker totally satisfied if he or she could obtain it. Alternatively, the neutral level can be replaced by the *unsatisfactory level*, which is a level that is not acceptable by the decision maker. We set for convenience $u_i(\mathbf{1}_i) = 1$ and $u_i(\mathbf{0}_i) = 0$ for all $i \in N$. We consider now for any subset $A \subseteq N$ the *binary alternative* $(\mathbf{1}_A, \mathbf{0}_{-A})$, whose coordinates are $\mathbf{1}_i$ for all $i \in A$, and $\mathbf{0}_i$ otherwise. This alternative is satisfactory for all criteria $i \in A$ and neutral (or unsatisfactory) otherwise. Then, $\mu(A)$ is interpreted as the overall evaluation of $(\mathbf{1}_A, \mathbf{0}_{-A})$.

Under this interpretation, the axioms defining a fuzzy measure are natural: the overall evaluation of an alternative being neutral (or unsatisfactory) for every criterion should be 0, and its evaluation should be 1 if all criteria are satisfactory. Moreover, turning one criterion on the neutral level to a satisfactory level cannot decrease the overall evaluation (monotonicity condition).

In summary, a fuzzy measure determines the evaluation of every binary alternative. It remains to determine the evaluation of any other alternative. This is done using the *Choquet integral* [6], defined as follows. Consider an alternative $x = (x_1, \dots, x_n)$ and the vector $u(x) = [u_1(x_1), \dots, u_n(x_n)]$ of its values for the utility functions. We call $u_i(x_i)$ the *score* of x with respect to criterion i , and we denote it by a_i if there is no ambiguity. Rearrange the criteria so that $a_1 \leq a_2 \leq \dots \leq a_n$. Then, the Choquet integral of $u(x) = (a_1, \dots, a_n)$, provided all components a_i are nonnegative, is

computed as follows:

$$\begin{aligned} \mathcal{C}_\mu(a_1, \dots, a_n) \\ = a_1 \underbrace{\mu(N)}_{=1} + \sum_{i=2}^n (a_i - a_{i-1}) \mu(\{i, \dots, n\}), \\ (a_1, \dots, a_n \geq 0) \end{aligned} \quad (1)$$

or equivalently

$$\begin{aligned} \mathcal{C}_\mu(a_1, \dots, a_n) \\ = \sum_{i=1}^{n-1} a_i (\mu(\{i, \dots, n\}) \\ - \mu(\{i+1, \dots, n\})) + a_n \mu(\{n\}). \end{aligned} \quad (2)$$

Example 1 Consider $N = \{1, 2, 3\}$ and the fuzzy measure μ defined in the following table

A	$\emptyset$	{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	N
$\mu(A)$	0	0.2	0.2	0.4	0.5	0.6	0.7	1

Consider now the vector $u(x) = (3, 1, 6)$. Criteria should be first rearranged in the following order: 2, 1, 3, as $u_2(x_2) < u_1(x_1) < u_3(x_3)$. Then, using Equation (1), the Choquet integral is computed as follows:

$$\begin{aligned} \mathcal{C}_\mu(u(x)) &= 1 \times 1 + (3 - 1) \times \mu(\{1, 3\}) \\ &\quad + (6 - 3) \times \mu(\{3\}) \\ &= 1 + 2 \times 0.6 + 3 \times 0.4 = 3.4 \end{aligned} \quad (3)$$

Thus, the overall evaluation of x is 3.4. Observe that this value lies in $[1, 6]$, the interval of minimal and maximal utility values on the criteria.

Formula (2) clearly shows that the Choquet integral is a kind of weighted sum, the weights being $\mu(\{i, \dots, n\}) - \mu(\{i+1, \dots, n\})$. However, be careful that these weights depend on x , as criteria must be first reordered to get an increasing sequence of scores. Hence, precisely, as there are n possible orderings, the Choquet integral is an assembling of $n!$ weighted sums, each for a specific order of criteria (and this explains why the Choquet integral can model interaction). It can be proved that the

Choquet integral is nevertheless continuous. Other important properties are:

1. $C_\mu(u(\mathbf{1}_A, \mathbf{0}_{-A})) = \mu(A)$, for all $A \subseteq N$;
2. $C_\mu(\alpha u(x)) = \alpha C_\mu(u(x))$ for all $\alpha > 0$;
3. $u_i(x_i) \leq u_i(x'_i)$ for $i = 1, \dots, n$ implies $C_\mu(u(x)) \leq C_\mu(u(x'))$;
4. $\min_{i=1}^n u_i(x_i) \leq C_\mu(u(x)) \leq \max_{i=1}^n u_i(x_i)$.

The first property says that the Choquet integral extends the overall evaluation provided by the fuzzy measure. The second one says that the Choquet integral is compatible with a ratio scale. The third one says that an increase of the scores cannot lead to a decrease of the overall evaluation. The last one says that the overall evaluation is comprised between the minimum and the maximum of all scores. The two last properties imply *idempotency*, that is, $C_\mu(a, a, \dots, a) = a$ for any $a \in \mathbb{R}_+$.

So far, the Choquet integral is defined for nonnegative arguments, preventing it to be used with interval scales. In order to extend it to any real-valued argument, there are two possibilities. The first one is called the *symmetric integral*, denoted by $\check{C}_\mu$. Its name is due to the fact that $\check{C}_\mu(-u(x)) = -\check{C}_\mu(u(x))$, and it is compatible with ratio scales. However, it is not compatible with interval scales, as it is not invariant to positive affine transformations: $\check{C}_\mu(\alpha u(x) + \beta) \neq \alpha \check{C}_\mu(u(x)) + \beta$ in general, for $\alpha > 0$ and $\beta \in \mathbb{R}$. Its expression is, using the notation $u(x) = (a_1, \dots, a_n)$:

$$\begin{aligned} \check{C}_\mu(a_1, \dots, a_n) &= \sum_{i=1}^{p-1} (a_i - a_{i+1}) \mu(\{1, \dots, i\}) \\ &\quad + a_p \mu(\{1, \dots, p\}) + a_{p+1} \mu(\{p+1, \dots, n\}) \\ &\quad + \sum_{i=p+2}^n (a_i - a_{i-1}) \mu(\{i, \dots, n\}), \end{aligned}$$

having rearranged the criteria, so that $a_1 \leq \dots \leq a_p < 0 \leq a_{p+1} \leq \dots \leq a_n$.

The second possibility is called the *asymmetric integral*, and its definition merely amounts to keep the same formula (1) for $(a_1, \dots, a_n) \in \mathbb{R}^n$. The asymmetric integral is invariant to positive affine transformations and is, therefore, compatible with interval scales.

RELATION WITH OTHER AGGREGATION FUNCTIONS

The Choquet integral contains, as particular case, two important families of aggregation functions: the weighted sum (WAM) and the OWA [8]. We recall their definitions:

$$\begin{aligned} \text{WAM}(a_1, \dots, a_n) &= \sum_{i=1}^n w_i a_i \\ \text{OWA}(a_1, \dots, a_n) &= \sum_{i=1}^n w_i a_{(i)}, \end{aligned}$$

where $(a_{(1)}, \dots, a_{(n)})$ is the vector of scores rearranged in nondecreasing order, that is, $a_{(1)} \leq \dots \leq a_{(n)}$, and $w_1, \dots, w_n$ are weights in $[0, 1]$. We recall that OWA contains, as particular case, the arithmetic mean (with $w_1 = w_2 = \dots = w_n = \frac{1}{n}$), the minimum ($w_1 = 1, w_2 = \dots = w_n = 0$), the maximum ($w_1 = \dots = w_{n-1} = 0, w_n = 1$), the median, and every order statistic. We stress the fact that the weights in OWA are not weights on criteria, but weights on the rank of the scores. Hence, high values for weights with low (respectively, high) indices indicate that the OWA focuses on the smallest (respectively, greatest) scores obtained, that is, it has a conjunctive (respectively, disjunctive) behavior.

The WAM is recovered with an additive fuzzy measure, whereas the OWA is recovered with a symmetric fuzzy measure [9]. We define these terms below.

An *additive fuzzy measure* μ satisfies $\mu(A) = \sum_{i \in A} \mu(\{i\})$ for every $A \subseteq N$. It is then easy to see that the Choquet integral reduces to the WAM with $w_i = \mu(\{i\})$, for $i = 1, \dots, n$. A *symmetric fuzzy measure* μ satisfies $\mu(A) = \mu(B)$ whenever $|A| = |B|$, that is, μ depends only on the cardinality of the subsets. Letting $m_i = \mu(A)$ with $|A| = i$, we obtain that the Choquet integral reduces to an OWA with $w_i = m_{n-i+1} - m_{n-i}$, $i = 2, \dots, n$, and $w_1 = 1 - \sum_{i=2}^n w_i$.

This shows the versatility of the Choquet integral, which is able to consider both importance of criteria and importance of ranks. In particular, it can be more or less disjunctive or conjunctive.

SHAPLEY VALUE AND INTERACTION INDICES

The versatility of the Choquet integral is hindered by its exponential complexity. Indeed, the size of the model is in $O(2^n)$, which makes its use impossible for high values of n , and render its interpretation nearly impossible for the decision maker. Therefore, one needs some tool to interpret fuzzy measures, and also submodels of polynomial size. This section focuses on the first point, whereas the second one will be addressed in the next section.

The most popular tool in MCDM being the weighted sum, decision makers are used to deal with weights of importance for criteria. The Choquet integral being much more complex than a weighted sum, is there some similar notion to importance weights? An adequate answer to this question comes from cooperative game theory, through the notion of Shapley value [10, 11]. As μ quantifies the overall score of alternatives being satisfactory on some set of criteria, a criterion i should be said to be important; if whenever i enters a group of criteria A , the difference $\mu(A \cup i) - \mu(A)$ is high. The *Shapley importance index* $\phi_i(\mu)$ for criterion i is a weighted average of this difference taken over all possible $A \subseteq N \setminus i$:

$$\phi_i(\mu) = \sum_{A \subseteq N \setminus i} \frac{|A|!(n - |A| - 1)!}{n!} \times (\mu(A \cup i) - \mu(A)),$$

and the vector $(\phi_1(\mu), \dots, \phi_n(\mu))$ is called the *Shapley value* of μ . The weights in the average ensure that $\sum_{i=1}^n \phi_i(\mu) = 1$ and invariance to permutations on N . Also, $\phi_i(\mu) \geq 0$; therefore, the coefficients $\phi_1(\mu), \dots, \phi_n(\mu)$ act like weights of importance in a weighted sum.

The Shapley value offers a simple summary of the fuzzy measure μ , but is by no means a substitute to it; that is, infinitely many fuzzy measures have the same Shapley value. We need, therefore, some other quantity to enrich the summary. This is brought by the *interaction index*, a notion that is intuitively appealing in MCDM. Let us first consider two criteria i, j . By a *positive interaction* or *synergy* between i and j , we mean

that the satisfaction of both criteria is much more valuable than the satisfaction of them separately. We may say that they are *complementary*. This is translated quantitatively as follows, taking any set $A \subseteq N \setminus \{i, j\}$:

$$\begin{aligned} \mu(A \cup \{i, j\}) - \mu(A) &\geq (\mu(A \cup i) - \mu(A)) \\ &\quad + (\mu(A \cup j) - \mu(A)), \end{aligned}$$

which can be rewritten as

$$\mu(A \cup \{i, j\}) - \mu(A \cup i) - \mu(A \cup j) + \mu(A) \geq 0.$$

Similarly, a *negative interaction* or *synergy* between i and j means that the satisfaction of both is not that better than the satisfaction of one of them, leading to the reverse inequality; and in this case, we may say that they are *redundant* or *substitutable*. The case of equality means that the added value by both criteria is exactly the sum of the individual added values, hence there is no interaction between i and j . The *interaction index* $I_{ij}(\mu)$ [12] is obtained by taking a weighted average of the above expression for all possible groups $A \subseteq N \setminus \{i, j\}$:

$$\begin{aligned} I_{ij}(\mu) &= \sum_{A \subseteq N \setminus \{i, j\}} \frac{|A|!(n - |A| - 2)!}{(n - 1)!} \\ &\quad \times (\mu(A \cup \{i, j\}) - \mu(A \cup i) - \mu(A \cup j) + \mu(A)). \end{aligned}$$

Note that the weights are similar to those of the Shapley value.

The Shapley value together with the interaction indices for all pairs $i, j \in N$ gives a good summary of the fuzzy measure μ , although still not complete. It is possible to define interaction indices for more than two criteria. The general definition for a group of criteria $B \subseteq N$ is given as follows [13]:

$$\begin{aligned} I_B(\mu) &= \sum_{A \subseteq N \setminus B} \left(\frac{|A|!(n - |A| - |B|)!}{(n - |B| + 1)!} \right. \\ &\quad \left. \times \sum_{K \subseteq B} (-1)^{|B \setminus K|} \mu(A \cup K) \right). \end{aligned}$$

Note that the Shapley value is a particular case of it when $B = \{i\}$, $i \in N$. It can be

shown that the set of 2^n interaction indices $I_B(\mu)$ for $B \subseteq N$ is an exact representation of μ , in the sense that there is a one-to-one correspondence between them.

Example 2 Taking again the fuzzy measure given in Example 1, we find for the interaction indices and Shapley value:

$$\begin{aligned}
 \phi_1(\mu) &= \frac{1}{3}(0.2 - 0) + \frac{1}{6}(0.5 - 0.2) \\
 &\quad + \frac{1}{6}(0.6 - 0.4) + \frac{1}{3}(1 - 0.7) = 0.25 \\
 \phi_2(\mu) &= \frac{1}{3}(0.2) + \frac{1}{6}(0.3) + \frac{1}{6}(0.3) \\
 &\quad + \frac{1}{3}(0.4) = 0.3 \\
 \phi_3(\mu) &= \frac{1}{3}(0.4) + \frac{1}{6}(0.4) + \frac{1}{6}(0.5) \\
 &\quad + \frac{1}{3}(0.5) = 0.45 \\
 I_{12}(\mu) &= \frac{1}{2}(0.5 - 0.2 - 0.2 + 0) \\
 &\quad + \frac{1}{2}(1 - 0.6 - 0.7 + 0.4) = 0.1 \\
 I_{13}(\mu) &= \frac{1}{2}(0.6 - 0.2 - 0.4 + 0) \\
 &\quad + \frac{1}{2}(1 - 0.5 - 0.7 + 0.2) = 0 \\
 I_{23}(\mu) &= \frac{1}{2}(0.7 - 0.2 - 0.4 + 0) \\
 &\quad + \frac{1}{2}(1 - 0.5 - 0.6 + 0.2) = 0.1 \\
 I_{123}(\mu) &= 1 - 0.5 - 0.6 - 0.7 + 0.2 + 0.2 \\
 &\quad + 0.4 - 0 = 0.
 \end{aligned}$$

One sees that criterion 3 is the most important, criteria 1 and 3 are independent, and there is a slight complementarity between criteria 1 and 2, and between 2 and 3.

K-ADDITIVE FUZZY MEASURES

The main advantage of the representation of a fuzzy measure μ with interaction indices is that it gives a clear meaning to μ : the Shapley value expresses the importance of

criteria, the interaction index for pairs of criteria expresses when they are complementary or redundant, and so on. However, the interpretation of $I_B(\mu)$ is less and less clear as the size of B grows: it seems hardly possible to have a clear view of what expresses an interaction index for four criteria. This suggests that one should truncate the representation up to some level. We call *k-additive fuzzy measure* [13] a fuzzy measure for which the interaction indices are zero for any group of criteria of size greater than k . Hence, a 1-additive measure reduces to a vector of importance weights, and is, therefore, equivalent to an additive fuzzy measure. In this case, we know that the Choquet integral is merely a weighted sum. More interesting is a 2-additive fuzzy measure, as complementarity and redundancy within pairs of criteria can be expressed. It can be shown that the Choquet integral takes the following particular form [14]:

$$\begin{aligned}
 C_\mu(a_1, \dots, a_n) &= \sum_{i,j: I_{ij} > 0} (a_i \wedge a_j) I_{ij} + \sum_{i,j: I_{ij} < 0} (a_i \vee a_j) |I_{ij}| \\
 &\quad + \sum_{i \in N} a_i \left(\phi_i - \frac{1}{2} \sum_{j \neq i} |I_{ij}| \right), \quad (5)
 \end{aligned}$$

where $\wedge$ and $\vee$ stand for minimum and maximum, respectively, and we have omitted μ in $I_{ij}(\mu)$ and $\phi_i(\mu)$. This expression offers two important advantages:

1. It gives a clear meaning to the interaction index in terms of aggregation behavior. A positive interaction between i and j implies a conjunctive aggregation (*both* criteria must be satisfied), whereas a negative interaction implies a disjunctive behavior (it is enough to satisfy one of the criteria). In addition, one can see that the Shapley value indeed acts like a weight of importance in a weighted sum; however, its effect on criterion i is diminished by the sum of interaction indices pertaining to i . It follows that if criterion i has no interaction (is independent) with the other criteria,

it appears only in the linear part of the Choquet integral with weight ϕ_i .

2. The Choquet integral is a sum of terms being a minimum, a maximum, or a monomial of degree 1. It can be shown that all coefficients of these terms are nonnegative; moreover, they sum up to 1: in short, Equation (5) is a convex combination of terms representing elementary behaviors in decision making, disjunction and conjunction of two criteria, and dictators. All these local behaviors can be put in a pie chart, which offers a very clear view for the decision maker.

Finally, observe that the 2-additive model is in $O(n^2)$, as the number of free coefficients defining a 2-additive fuzzy measure is $n - 1 + \binom{n}{2} = \frac{n^2+n-2}{2}$.

Example 3 We take again the fuzzy measure μ of Example 1, with $u(x) = (3, 1, 6)$. Indeed, we see by Example 2 that μ is 2-additive. Therefore, Equation (5) can be applied, and it gives (see Example 2 for the values of interaction):

$$\begin{aligned} C_\mu(u(x)) &= 0.1(3 \wedge 1) + 0 + 0.1(1 \wedge 6) \\ &\quad + 3 \left(0.25 - \frac{0.1}{2} \right) + 1 \left(0.3 - \frac{0.2}{2} \right) \\ &\quad + 6 \left(0.45 - \frac{0.1}{2} \right) = 3.4 \end{aligned}$$

We observe that, as expected, the result is the same as our computation in Example 1.

ORDINAL MODELS

So far, it was assumed that utility functions were real-valued functions. In many cases, however, scores are given on a finite ordinal scale, such as $\{bad, medium, good\}$. As the fuzzy measure is interpreted as an overall score on binary alternatives, it follows that the fuzzy measure too is defined on the same ordinal scale. However, the arithmetic operations used in the Choquet integral cannot be used on an ordinal scale, and another kind of

integral has to be used, involving in its definition only minimum and maximum, which are the only operations being meaningful on an ordinal scale.

Let us denote the finite ordinal scale by L , with 0 and 1 the lower and upper bounds of L , respectively. Imposing that the overall score be 0 (respectively, 1), if all scores on criteria are 0 (respectively, 1), it can be shown [15] that the only possible aggregation function using only minimum and maximum is the *Sugeno integral* [5], defined by

$$S_\mu(a_1, \dots, a_n) = \bigvee_{i=1}^n (a_i \wedge \mu(\{i, \dots, n\})),$$

where $(a_1, \dots, a_n)$ is the vector of scores, and criteria have been reordered so that $a_1 \leq \dots \leq a_n$, as for the Choquet integral. To be meaningful, the fuzzy measure and all utility functions must take values in L , with $\mu(N) = 1$.

The Sugeno integral keeps many of the properties of the Choquet integral:

1. $S_\mu(u(\mathbf{1}_A, \mathbf{0}_{-A})) = \mu(A)$, for all $A \subseteq N$;
2. $S_\mu(\alpha \wedge u(x)) = \alpha \wedge S_\mu(u(x))$, $S_\mu(\alpha \vee u(x)) = \alpha \vee S_\mu(u(x))$;
3. $u_i(x_i) \leq u_i(x'_i)$ for $i = 1, \dots, n$ implies $S_\mu(u(x)) \leq S_\mu(u(x'))$;
4. $\min_{i=1}^n u_i(x_i) \leq S_\mu(u(x)) \leq \max_{i=1}^n u_i(x_i)$.

However, an important drawback of the Sugeno integral is that it can happen that two alternatives x, x' have the same overall evaluation although $u_i(x_i) > u_i(x'_i)$ for all $i \in N$ (Example 4).

Example 4 Let us consider the scale $L = \{0, 1, \dots, 10\}$ and use the fuzzy measure of Example 1 multiplied by 10, together with $u(x) = (3, 1, 6)$. We obtain:

$$S_\mu(u(x)) = (1 \wedge 10) \vee (3 \wedge 6) \vee (6 \wedge 4) = 4.$$

Note that this is relatively close to 3.4 found in Example 1 with the Choquet integral. Now consider $u(x') = (2, 0, 5)$. We find

$$S_\mu(u(x')) = (0 \wedge 10) \vee (2 \wedge 6) \vee (5 \wedge 4) = 4,$$

that is, $S_\mu(u(x)) = S_\mu(u(x'))$, although x has better scores than x' on each criterion.

IDENTIFICATION OF THE MODEL

Once the utility functions have been determined, the identification of the model amounts to the identification of the fuzzy measure, which, most of the time, will be the solution of an optimization problem under constraints. We give here two main types of optimization problem (we refer the reader to [16] and [4, Chapter 11] for a survey and more detail):

1. *Minimizing the total squared error:* Denoting by $\mathcal{O}$ the set of available alternatives, the data set consists of the vectors $u(x)$ of scores on criteria together with the overall score $y(x)$, for all $x \in \mathcal{O}$. The aim is to minimize the distance to the desired overall scores:

$$\min_{\mu} \sum_{x \in \mathcal{O}} (\mathcal{C}_\mu(u(x)) - y(x))^2.$$

The constraints of the problem are the monotonicity constraints of the fuzzy measure:

$$\mu(A \cup i) - \mu(A) \geq 0, \quad \forall i \in N, \forall A \subseteq N \setminus i,$$

and possibly some information on the Shapley value and interaction indices, such as a partial order or the sign of some interaction index.

2. *Maximum separation of alternatives:* The data consists of a preference relation $\succeq$ on a set $\mathcal{O}$ of alternatives. The objective is to maximize the distance between the overall scores obtained for these alternatives, provided they are compatible with the preference relation (constraints):

$$\begin{aligned} \max \quad & \epsilon \\ \text{subject to} \quad & \mathcal{C}_\mu(u(x)) - \mathcal{C}_\mu(u(x')) \geq \delta + \epsilon, \\ & \forall x, x' \in \mathcal{O}, \text{ such that } x \succeq x' \\ & \mu(A \cup i) - \mu(A) \geq 0, \\ & \forall i \in N, \forall A \subseteq N \setminus i, \end{aligned}$$

where δ is a fixed threshold. As above, additional constraints can be given, pertaining to the Shapley value and interaction indices. It should be noted that contrarily to the first case, this optimization problem may have no solution.

The above optimization problems concern solely the Choquet integral, and are either quadratic (in the first case) or linear (in the second case) programs. For the first case, standard quadratic programming methods can be avoided, and a faster (although suboptimal) heuristic algorithm (HLMS, heuristic least mean squares) can be used instead [17, 18]. For the Sugeno integral, only heuristic methods such as genetic algorithms can be used [19].

REFERENCES

1. Grabisch M. The application of fuzzy integrals in multicriteria decision making. *Eur J Oper Res* 1996;89:445–456.
2. Grabisch M, Labreuche Ch. Fuzzy measures and integrals in MCDA. In: Figueira J, Greco S, Ehrgott M, editors. *Multiple criteria decision analysis*. Kluwer Academic Publishers; 2005. p 563–608.
3. Grabisch M, Labreuche Ch. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Ann Oper Res* 2010;175:247–286. doi: 10.1007/s10479-009-0655-8.
4. Grabisch M, Marichal J-L, Mesiar R, *et al.* Aggregation functions. Volume 127, *Encyclopedia of Mathematics and its Applications*. Cambridge: Cambridge University Press; 2009.
5. Sugeno M. *Theory of fuzzy integrals and its applications* [PhD thesis]. Tokyo Institute of Technology; 1974.
6. Choquet G. Theory of capacities. *Ann Inst Fourier* 1953;5:131–295.
7. Bana e Costa CA, Vansnick J-C. The MACBETH approach: basic ideas, software and an application. In: Meskens N, Roubens M, editors. *Advances in decision analysis*. Kluwer Academic Publishers; 1999. p 131–157.
8. Yager RR. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Trans Syst Man Cybern* 1988;18:183–190.

9. Murofushi T, Sugeno M. Some quantities represented by the Choquet integral. *Fuzzy Sets Syst* 1993;56:229–235.
10. Shapley LS. A value for n -person games. In: Kuhn HW, Tucker AW, editors. *Contributions to the theory of games. Volume II (28)*, *Annals of Mathematics Studies*. Princeton University Press; 1953. p 307–317.
11. Roth AE, editor. *The Shapley value. Essays in Honor of Lloyd S. Shapley*. Cambridge University Press; 1988.
12. Murofushi T, Soneda S. Techniques for reading fuzzy measures (III): interaction index. 9th *Fuzzy System Symposium*, Sapporo, Japan, May 1993. p 693–696. In Japanese.
13. Grabisch M. k -order additive discrete fuzzy measures and their representation. *Fuzzy Sets Syst* 1997;92:167–189.
14. Grabisch M. Alternative representations of discrete fuzzy measures for decision making. *Int J Uncertain Fuzz Knowl Based Syst* 1997;5:587–607.
15. Marichal J-L. Weighted lattice polynomials. *Discrete Math* 2009;309:814–820.
16. Grabisch M, Kojadinovic I, Meyer P. A review of capacity identification methods for Choquet integral based multi-attribute utility theory—applications of the Kappalab R package. *Eur J Oper Res* 2008;186:766–785.
17. Grabisch M. A new algorithm for identifying fuzzy measures and its application to pattern recognition. *International Joint Conference of the 4th IEEE International Conference on Fuzzy Systems and the 2nd International Fuzzy Engineering Symposium*, Yokohama, Japan, March 1995. p 145–150.
18. Grabisch M, Raufaste E. An empirical study of statistical properties of Choquet and Sugeno integrals. *IEEE Trans Fuzzy Syst* 2008;16(4):839–850.
19. Wang Z, Leung KS, Wang J. A genetic algorithm for determining nonadditive set functions in information fusion. *Fuzzy Sets Syst* 1999;102:462–469.

GAME-THEORETIC METHODS IN COUNTERTERRORISM AND SECURITY

VICKI M. BIER

Department of Industrial and Systems
Engineering, University of
Wisconsin-Madison, Madison,
Wisconsin

After the September 11, 2001, terrorist attacks on the World Trade Center and the Pentagon, and the subsequent anthrax attacks in the United States, there has been an increased interest in security. However, security problems pose some significant challenges to conventional methods for risk analysis [1,2]. One critical challenge is the intentional nature of terrorist attacks [3].

Protecting against intentional attacks is fundamentally different from protecting against accidents or acts of nature. In particular, an adversary may adopt a different offensive strategy in response to observed security measures, in order to circumvent or disable those measures. For example, Ravid [4] argues that since the adversary can change targets in response to defensive investments, “investment in defensive measures, unlike investment in safety measures, saves a lower number of lives (or other sort of damages) than the apparent direct contribution of those measures.” Game theory [5] provides a way of taking this into account. In particular, game theory identifies the optimal defense against and optimal attack. While it may not be possible to actually predict terrorist attacks, protecting against an optimal attack can arguably be described as conservative. This article discusses approaches for applying game theory to the problem of defending against intentional attacks.

APPLICATIONS OF GAME THEORY TO SECURITY

Game theory has a long history of being applied to security, beginning with military

applications [6–10]. It has also been extensively used in political science [11–13]; for example, in the context of arms control. Recently, game theory has also been applied to computer security [14,15]. Many applications of game theory to security have been by economists [16–27]. Much of this work has been designed to provide *policy insights* [25]—for example, on the relative merits of public versus private funding of defensive investments [23], or deterrence versus defense [19,20,22,27]. Since the events of September 11, however, there has been increased interest in using game theory not only to explore the effects of different policies, but to generate detailed guidance in support of operational-level decisions as well—for example, which assets to protect, or how much to charge for terrorism insurance.

For example, Enders and Sandler [18] and Hausken [21] study substitution effects in security. Enders and Sandler observe that “installation of screening devices in US airports in January 1973 made skyjackings more difficult, thus encouraging terrorists to substitute into other kinds of hostage missions.” Similarly, they note: “If ... the government were to secure its embassies or military bases, then attacks against such facilities would become more costly on a per-unit basis. If, moreover, the government were not at the same time to increase the security for embassy and military personnel when outside their facilities, then attacks directed at these individuals (e.g., assassinations) would become relatively cheaper” (and hence presumably more frequent).

Clearly, security improvements that appear to be justified without taking into account the fact that attacks may be deflected to other targets may turn out to be wasteful (at least from a public perspective) if they merely deflect attacks to other targets of comparable value. Therefore, the Brookings Institution [28] has recommended that “policy-makers should focus primarily on those targets at which an attack would involve a large number of casualties, would

entail significant economic costs, or would critically damage sites of high national significance.”

The Brookings recommendation constitutes a reasonable *zero-order* suggestion about how to prioritize targets for investment. However, it is important to go beyond the zero-order heuristic of protecting only the most valuable or potentially damaging targets, to account for the success probabilities of different attacks, since terrorists appear to take the probability of success into account in their choice of targets. For example, Woo [29] has observed that “Al-Qaeda is . . . sensitive to target hardening,” and that “Osama bin Laden has expected very high levels of reliability for martyrdom operations.” Thus, even if a successful attack against a given target would be highly damaging, that target may not merit as much defensive investment as a target that is less valuable but more vulnerable (and hence more attractive to attackers). Early models that take the success probabilities of potential attacks into account include Woo [29], Bier *et al.* [30], and Major [31].

Both the Brookings recommendations [28] and the results in Bier *et al.* [30] still represent “weakest-link” models, in which defensive investment is allocated only to the target(s) that would cause the most damage if attacked. However, such weakest-link models can be unrealistic in practice. For example, Arce and Sandler [16] note that weakest-link solutions “are not commonly observed among the global and transnational collective action problems confronting humankind.” In particular, real-world decision makers typically *hedge* by defending additional targets, to cover contingencies such as misestimating which targets are most attractive to attackers.

The models in Woo [29] and Major [31] achieve the more realistic result of defensive hedging at optimality, but do so with a restrictive (and possibly questionable) assumption. In particular, these models assume that the attacker can observe “the marginal effectiveness of defense . . . at [each] target” [31], but not which defenses have been implemented. More recent work

[32–34] attempts to achieve the result of defensive hedging at equilibrium in a different manner—assuming (perhaps conservatively) that attackers can observe defensive investments perfectly, but that defenders are uncertain about the attractiveness of possible targets to the attackers. Interestingly, hedging does not always occur in this model. In particular, the results suggest that it is often optimal for the defender to put no investment at all into some targets—especially when the defender is highly resource constrained, and the various potential targets differ greatly in their values (both of which are likely to be the case in practice).

Recently, researchers have begun to focus on allocation of defensive resources in the face of threats from both terrorism and natural disasters [35–38]. In addition, some researchers have begun formally taking into account the effect of deterrence or retaliation on terrorist recruitment [24,39].

Security as a Game between Defenders

The work discussed above views security primarily as a game between attacker and defender. However, defensive strategies adopted by one agent can also affect the incentives faced by other defenders. Some types of defensive actions (such as installation of visible burglar alarms or car alarms) may increase risk to other potential victims (a negative externality), leading to overinvestment in security. Conversely, other types of defensive actions—such as vaccination, fire protection, or use of antivirus protection software—decrease the risk to other potential victims (a positive externality). This type of situation can result in underinvestment in security, and *free riding* by defenders [21,25].

In order to better account for security investments with positive externalities, Kunreuther and Heal [40] propose a model of interdependent security (IDS) where agents are vulnerable to *infection* from other agents; Heal and Kunreuther [41] apply this model to the case of airline security. More specifically, this body of work assumes that the consequences of even a single successful

attack would be catastrophic—leading, for example, to business failure. Under this assumption, Kunreuther and Heal show that failure of one agent to invest in security can make it unprofitable for other agents to invest, even when they would otherwise do so. Moreover, this game can have multiple equilibrium solutions (e.g., one in which all defenders invest in security, and another in which none invest), so coordinating mechanisms can help to ensure that the socially optimal level of investment is reached.

Recent work [42] has extended these results to attacks occurring over time (rather than a *snapshot* model). In this case, differences in discount rates among agents can lead some agents with low discount rates not to invest in security, if other agents (e.g., those with high discount rates) choose not to invest. Thus, heterogeneous time preferences can complicate the task of achieving security.

Unlike the above work, Keohane and Zeckhauser [22] consider both positive and negative externalities among defenders. In particular, they note that reduction of population in a major city (e.g., due to relocation decisions by members of the public) may make all targets in that city less attractive to terrorists, while government investment to make a location safer will tend to attract increased population, making that location more attractive to potential attackers.

COMBINING RISK ANALYSIS AND GAME THEORY

Conventional risk analysis does not explicitly model the adaptive response of potential attackers, and hence may overstate the effectiveness and cost-effectiveness of defensive investments. However, much game-theoretic security work focuses on nonprobabilistic games. Even those models that explicitly consider the success probabilities of potential attacks [29–34,38,40–42] generally treat individual assets in isolation, and fail to consider their possible role as components of a larger system. Therefore,

combining risk analysis with game theory could be fruitful in studying intentional threats to complex systems such as critical infrastructure.

Hausken [43] has integrated probabilistic risk analysis and game theory (although not in the security context), by interpreting system reliability as a public good, and considering the incentives of players responsible for maintaining particular components of a larger system. In particular, by viewing reliability as a game between defenders responsible for different portions of the system, he identifies relationships between the series or parallel structure of the system being analyzed and classic games such as the coordination game, battle of the sexes, chicken, and prisoner's dilemma.

Rowe [44] argues that the implications of “the human variable” in terrorism risk have yet to be adequately appreciated, and presents a framework for evaluating possible defenses in light of terrorists’ ability to “learn from experience and alter their tactics.” This approach has been used in practice to help prioritize defensive investments among multiple potential targets and threat types. Similarly, Paté-Cornell and Guikema [45] discuss the need for “periodic updating of the model and its input” to account for the dynamic nature of terrorism, but do not formally use game theory.

Banks and Anderson [46] apply similar ideas to the threat of intentionally introduced smallpox. Their approach embeds risk analysis in a game-theoretic formulation of the defender’s decision, accounting for both the adaptive nature of terrorism, and also uncertainty about the costs and benefits of particular defenses. They conclude that this approach “captures facets of the problem that are not amenable to either game theory or risk analysis on their own.”

Bier *et al.* [30] use game theory to study security of simple series and parallel systems, as a building block to more complex systems. Results suggest that defending series systems against intentional attack is extremely difficult, if the attacker knows about the system’s defenses. In particular, the attacker’s

ability to respond to the defender's investments deprives the defender of the ability to allocate defensive investments according to their cost-effectiveness; instead, defensive investments in series systems must equalize the strength of all defended components, consistent with results in Dresher [8]. Recent work [47] begins to extend these types of models to more complicated system structures (including both parallel and series subsystems).

Recently, controversy has developed over the use of game theory in homeland security, with some arguing [48] that the assumptions of game theory are unrealistic, and that probabilistic risk analysis provides an adequate means of assessing security threats, and others arguing [49] that game-theoretic insights or other new approaches that explicitly account for adversary objectives are crucial to protection against intelligent adversaries. While both methods have strengths and weaknesses, of course, it seems likely that any good method for assessing security threats will involve some method for explicitly incorporating adversary objectives, even if not in a formally game-theoretic manner.

CONCLUSIONS

As noted above, protecting complex systems against intentional attacks is likely to require a combination of game theory and risk analysis. Risk analysis by itself does not account for the attacker's responses to security measures, while game theory often deals with individual assets in isolation, rather than with complex networked systems (such as computer systems, electrical transmission systems, or transportation systems). Approaches that combine risk models and game methods can take advantage of the strengths of both approaches.

Acknowledgments

This material is based upon work supported in part by the US Army Research Laboratory and the US Army Research Office under grant number DAAD19-01-1-0502, by the US National Science Foundation under

grant number DMI-0228204, by the Midwest Regional University Transportation Center under project number 04-05, and by the United States Department of Homeland Security through the Center for Risk and Economic Analysis of Terrorism Events (CREATE) under grant numbers EMW-2004-GR-0112 and 2007-ST-061-000001. Any opinions, findings, and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect the views of the sponsors. I would also like to acknowledge the numerous colleagues and students who have contributed to the body of work discussed here.

Revised and adapted from chapters in *Modern Statistical and Mathematical Methods in Reliability*, World Scientific Publishing Company, ISBN: 981-256-356-3; *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Springer, ISBN: 038-732-904-8; and *Encyclopedia of Quantitative Risk Analysis and Assessment*, Wiley, ISBN: 047-003-549-8.

REFERENCES

1. Bedford T, Cooke R. Probabilistic risk analysis: foundations and methods. Cambridge: Cambridge University Press; 2001.
2. Bier VM. An overview of probabilistic risk analysis for complex engineered systems. In: Molak V, editor. *Fundamentals of risk analysis and risk management*. Boca Raton (FL): Lewis Publishers; 1997. pp. 67–85.
3. Brannigan V, Smidts C. Performance based fire safety regulation under intentional uncertainty. *Fire Mater* 1998;23(6):341–347.
4. Ravid I. Theater ballistic missiles and asymmetric war. The Military Conflict Institute; 2002.
5. Fudenberg D, Tirole J. *Game theory*. Cambridge (MA): MIT Press; 1991.
6. Berkovitz LD, Dresher M. A game-theory analysis of tactical air war. *Oper Res* 1959;7(5):599–620.
7. Berkovitz LD, Dresher M. Allocation of two types of aircraft in tactical air war: a game-theoretic analysis. *Oper Res* 1960;8(5):694–706.

8. Dresher M. Games of strategy: theory and applications. Englewood Cliffs (NJ): Prentice-Hall; 1961.
9. Haywood OG. Military decision and game theory. *J Oper Res Soc Am* 1954;2(4): 365–385.
10. Leibowitz ML, Lieberman GJ. Optimal composition and deployment of a heterogeneous local air-defense system. *Oper Res* 1960;8(3):324–337.
11. Brams SJ. Game theory and politics. New York: Free Press; 1975.
12. Brams SJ. Superpower games: applying game theory to superpower conflict. New Haven (CT): Yale University Press; 1985.
13. Brams SJ, Kilgour DM. Game theory and national security. Oxford: Basil Blackwell; 1988.
14. Anderson R. Why information security is hard: an economic perspective. 18th Symposium on Operating Systems Principles: Banff, Canada. 2001. <http://www.cl.cam.ac.uk/~rja14/Papers/econ.pdf>.
15. Schneier B. Secrets and lies: digital security in a networked world. New York: Wiley; 2000.
16. Arce M, Daniel G, Sandler T. Transnational public goods: strategies and institutions. *Eur J Polit Econ* 2001;17(3):493–516.
17. Brauer J, Roux A. Peace as an international public good: an application to southern Africa. *Def Peace Econ* 2000;11(4):643–659.
18. Enders W, Sandler T. What do we know about the substitution effect in transnational terrorism? In: Silke A, editor. Researching terrorism: trends, achievements, failures. London: Frank Cass; 2004. pp. 119–137.
19. Frey BS. Dealing with terrorism - stick or carrot? Cheltenham: Edward Elgar; 2004.
20. Frey BS, Luechinger S. How to fight terrorism: alternatives to deterrence. *Def Peace Econ* 2003;14(4):237–249.
21. Hausken K. Income, interdependence, and substitution effects affecting incentives for security investment. *J Account Public Policy* 2006;25(6):629–665.
22. Keohane NO, Zeckhauser RJ. The ecology of terror defense. *J Risk Uncertain* 2003;26(2–3):201–229.
23. Lakdawalla D, Zanjani G. Insurance, self-protection, and the economics of terrorism. *J Public Econ* 2005;89(9–10):1891–1905.
24. Rosendorff B, Sandler T. Too much of a good thing? The proactive response dilemma. *J Conflict Resolut* 2004;48(5):657–671.
25. Sandler T, Daniel G, Arce M. Terrorism and game theory. *Simul Gaming* 2003;34(3):319–337.
26. Sandler T, Enders W. An economic perspective on transnational terrorism. *Eur J Polit Econ* 2004;20(2):301–316.
27. Siqueira K, Sandler T. Terrorists versus the government: strategic interaction, support, and sponsorship. *J Conflict Resolut* 2006;50(6):1–21.
28. O'Hanlon M, Orszag P, Daalder I, *et al.* Protecting the American Homeland. Washington (DC): Brookings Institution; 2002.
29. Woo G. Insuring against Al-Qaeda. Insurance Project Workshop. 2003. Available at <http://www.nber.org/~confer/2003/insurance03/woo.pdf>. Accessed in 2010.
30. Bier VM, Nagaraj A, Abhichandani V. Protection of simple series and parallel systems with components of different values. *Reliab Eng Syst Saf* 2005;87(3):313–323.
31. Major J. Advanced techniques for modeling terrorism risk. *J Risk Financ* 2002;4(1):15–24.
32. Bier VM, Oliveros S, Samuelson L. Choosing what to protect: strategic defensive allocation against an unknown attacker. *J Public Econ Theory* 2007;9(4):563–387.
33. Bier VM. Choosing what to protect. *Risk Anal* 2007;27(3):607–620.
34. Bier VM, Hapuriwat N, Menoyo J, *et al.* Optimal resource allocation for defense of targets based on differing measures of attractiveness. *Risk Anal* 2008;28(3):763–770.
35. Powell R. Defending against terrorist attacks with limited resources. *Am Polit Sci Rev* 2007;101(3):527–541.
36. Woo G. Joint mitigation of earthquake and terrorism risk. Proceedings of 8th U.S. National Conference on Earthquake Engineering: San Francisco (CA). 2006.
37. Zhuang J, Bier VM. Balancing terrorism and natural disasters: defensive strategy with endogenous attacker effort. *Oper Res* 2007;55(5):976–991.
38. Golany B, Kaplan EH, Marmur A, *et al.* Nature plays with dice—terrorists do not: allocating resources to counter strategic versus probabilistic risks. *Eur J Oper Res* 2008;192(1):198–208.
39. Das SP, Lahiri S. A strategic analysis of terrorist activity and counter-terrorism policies. *Top Theor Econ* 2006;6(2): Available at <http://www.bepress.com/cgi/viewcontent.cgi?article=1295&context=bejte>.

40. Kunreuther H, Heal G. Interdependent security. *J Risk Uncertain* 2003;26(2–3):231–249.
41. Heal G, Kunreuther H. IDS models of airline security. *J Conflict Resolut* 2005; 49(2):201–217.
42. Zhuang J, Bier VM, Gupta A. Subsidies in interdependent security with heterogeneous discount rates. *Eng Econ* 2007;52(1): 1–19.
43. Hausken K. Probabilistic risk analysis and game theory. *Risk Anal* 2002;22(1): 17–27.
44. Rowe WD. Vulnerability to terrorism: addressing the human variables. In: Haines YY, Moser DA, Stakhiv EZ, editors. *Risk-based decision making in water resources X*. Reston (VA): American Society of Civil Engineers; 2002. pp. 155–159.
45. Pate-Cornell E, Guikema S. Probabilistic modeling of terrorist threats: a systems analysis approach to setting priorities among countermeasures. *Mil Oper Res* 2002;7(4):5–20.
46. Banks DL, Anderson S. Combining game theory and risk analysis in counterterrorism: a smallpox example. In: Wilson AG, Wilson GD, Olwell DH, editors. *Statistical methods in counterterrorism: game theory, modeling, syndromic surveillance, and biometric authentication*. New York: Springer; 2006. pp. 9–22.
47. Azaiez N, Bier VM. Optimal resource allocation for security in reliability systems. *Eur J Oper Res* 2007;181(2):773–786.
48. Ezell BC, Bennet SP, von Winterfeldt D, *et al.* Probabilistic risk analysis and terrorism risk. *Risk Anal* 2010;30(4):575–589.
49. Parnell GS, Smith CM, Moxley FI. Intelligent adversary risk analysis: a bioterrorism risk management model. *Risk Anal* 2010;30(1):32–48.

GENERATING HOMOGENEOUS POISSON PROCESSES

RAGHU PASUPATHY
Department of Industrial and Systems
Engineering, Virginia Tech,
Blacksburg, Virginia

Recall that a *counting process* $\{N_t, t \geq 0\}$ is a stochastic process defined on a sample space Ω such that for each $\omega \in \Omega$, the function $N_t(\omega)$ is a “realization” of the number of “events” happening in the interval $(0, t]$, with $N_0(\omega) = 0$. By this definition, $N_t(\omega)$ is automatically integer valued, nondecreasing, and right-continuous for each ω . A nonhomogeneous Poisson process is a type of counting process that is characterized as follows.

Definition 1. A counting process $\{N_t, t \geq 0\}$ is called a nonhomogeneous Poisson process if:

- (i) $\forall t, s \geq 0$, and $0 \leq u \leq t$, $N_{t+s} - N_t$ is independent of N_u ;
- (ii) $\forall t, s \geq 0$, $\Pr\{N_{t+s} - N_t \geq 2\} = o(s)$; and
- (iii) $\forall t, s \geq 0$, $\Pr\{N_{t+s} - N_t = 1\} = \lambda(t)s + o(s)$, where $\lambda(t)$ is some positive-valued function.

Definition 1 has been adopted from Billingsley [1, p. 297], with the notation $o(s)$ used in the usual sense—to denote a function $v(s)$ that satisfies $\lim_{s \rightarrow 0} v(s)/s = 0$ [2, p. 1]. The function $\lambda(t)$ appearing in Definition 1, called the *rate function*, completely characterizes the Poisson process. Various other definitions of a Poisson process are available and quite prevalent. See, for instance, Cox and Lewis [3], Ross [4], Gnedenko and Kovalenko [5], and Çinlar [6] (Section 23 in Ref. 1 discusses the equivalence of various definitions of the Poisson process.) For example, Çinlar [6] provides the following definition based on the sample-paths of N_t that is particularly relevant to the current context.

Definition 2. A counting process $\{N_t, t \geq 0\}$ is called a nonhomogeneous Poisson process if:

- (i) $\forall t, s \geq 0$, and $0 \leq u \leq t$, $N_{t+s} - N_t$ is independent of N_u ;
- (ii) $\forall t \geq 0$ and for almost all ω , the mapping $t \rightarrow N_t(\omega)$ has jumps of unit magnitude only, that is, $N_t(\omega) - \lim_{s \uparrow t} N_s(\omega) = 0$ or $1 \forall t \geq 0$ and for almost all ω .

It is worth noting that Definitions 1 and 2 are not equivalent—the latter definition allows (unit) jumps of positive probability in the sample-paths, in which case, the rate function $\lambda(t)$ appearing in Definition 1 simply does not exist.

In what follows, we discuss existing methods to generate pseudorandom numbers from a nonhomogeneous Poisson process. By this, we mean generating, on a digital computer, a realization of event times $\{t_i, i = 1, 2, \dots\}$ from a Poisson process. The Poisson process is specified either through its rate function $\lambda(t)$ (when it exists), or more generally through its expectation function $\Lambda(t) \equiv E[N_t]$. When the rate function $\lambda(t)$ exists, $\Lambda(t) = \int_0^t \lambda(y) dy$.

GENERATING NONHOMOGENEOUS POISSON PROCESSES

As discussed in (see **Generating Nonhomogeneous Poisson Processes**) homogeneous Poisson processes can be generated very efficiently and in a fairly straightforward fashion. This is in some contrast with nonhomogeneous Poisson processes, where generation methods tend to be much less straightforward. We categorize the available methods for generating nonhomogeneous Poisson processes into three broad groups: (i) inversion methods, (ii) order statistics methods, and (iii) acceptance–rejection methods. We discuss each of these in what follows.

Inversion

The earliest known inversion method seems to be the technique devised by Çinlar [6, p. 96]. It is based on an interesting property of nonhomogeneous Poisson processes with a continuous expectation function $\Lambda(t)$.

Theorem 1 [Çinlar, 1975]. *Let $\Lambda(t), t \geq 0$ be a positive-valued, continuous, nondecreasing function. Then the random variables $T_1, T_2, \dots$ are event times corresponding to a nonhomogeneous Poisson process with expectation function $\Lambda(t)$ if and only if $\Lambda(T_1), \Lambda(T_2), \dots$ are the event times corresponding to a homogeneous Poisson process with rate one.*

Theorem 1 provides a method to generate event times from a nonhomogeneous Poisson process that is straightforward in principle. First, generate event times from a homogeneous Poisson process with rate one, and then invert $\Lambda(\cdot)$ to obtain the event times of the required process.

Algorithm 1 (Çinlar's Method).

- (0) Initialize $s = 0$.
- (1) Generate $u \sim U(0, 1)$.
- (2) Set $s \leftarrow s - \log(u)$.
- (3) Set $t \leftarrow \inf\{v : \Lambda(v) \geq s\}$.
- (4) Deliver t .
- (5) Go to Step (1).

When the expectation function $\Lambda(t)$ falls in a particular family of functions that facilitates efficient analytic inversion in Step (3), Algorithm 1 tends to be very fast. This happens only infrequently, however, and implementation often involves employing expensive numerical quadrature for inversion [7] in Step (3). Adoption of Çinlar's method should thus be considered accordingly. Code for a rudimentary version of Çinlar's method can be found through the website <https://filebox.vt.edu/users/pasupath/pasupath.htm>.

Theorem 1 assumes the continuity of the expectation function $\Lambda(t)$. When $\Lambda(t)$ has discontinuities, Algorithm 1 can be adapted in

a somewhat straightforward fashion. See, Çinlar [6, p. 100] and Resnick [8] for different treatments.

Another inversion approach to generating nonhomogeneous Poisson processes stems from the distribution of inter-event times. Specifically, consider the i th inter-event time $X_i = T_{i+1} - T_i$ conditional on the first i event times $T_1 = t_1, T_2 = t_2, \dots, T_i = t_i$. We can derive the cdf of X_i (conditional on $T_1, T_2, \dots, T_i$) as follows.

$$\begin{aligned}
 F_{t_i}(x) &= \Pr\{X_i \leq x | T_j = t_j, j = 1, 2, \dots, i\} \\
 &= \Pr\{N_{t_i+x} - N_{t_i} \geq 1 | T_j = t_j, \\
 &\quad j = 1, 2, \dots, i\} \\
 &= \Pr\{N_{t_i+x} - N_{t_i} \geq 1\} \\
 &= 1 - \Pr\{N_{t_i+x} - N_{t_i} = 0\} \\
 &= 1 - \exp(-\Lambda(t_i + x) + \Lambda(t_i)), \quad (1)
 \end{aligned}$$

where the third equality is from the independent increments property of the Poisson process. (We again note here that Equation (1) is not true in general if $\Lambda(t)$ has jumps. It can, however, be shown that Equation (1) will still hold if the location of the jumps has a Poisson number of events with mean equal to the value of the jump [8].)

The cdf appearing in Equation (1) provides a natural method of generating a nonhomogeneous Poisson process—given the previous i event times, generate the $(i+1)$ th event time as the sum of the i th event time and the i th inter-event time distributed according to F_{t_i} .

Algorithm 2.

- (0) Initialize $t = 0$.
- (1) Generate $x \sim F_t$ given by Equation (1).
- (2) Set $t \leftarrow t + x$.
- (3) Deliver t .
- (4) Go to Step (1).

Like Çinlar's method, Algorithm 2 is direct, exact, and elegant. However, and again like Çinlar's method, Algorithm 2 is efficient only when the form of the expectation function $\Lambda(t)$ renders easy generation from

F_t . Specifically, suppose the cdf-inverse method is used to generate from F_t in Step (1). If the nonhomogeneous Poisson process is specified through the rate function $\lambda(t)$, this amounts to finding x satisfying

$$\int_t^{t+x} \lambda(y) dy = -\ln(1-u), \text{ where } u \sim U(0,1). \quad (2)$$

In general, the problem appearing in Equation (2) can be solved only using numerical quadrature [7], and can hence be computationally expensive. (Public domain software is available for this purpose [9].)

By contrast, if enough structure is present and known, efficient tailor-made techniques can be devised to solve Equation (2). One such example is the piecewise-linear rate function considered by Klein and Roberts [10], and by Lee *et al.* [11], with success. Specifically, Klein and Roberts [10] assume that $\lambda(t)$ is a continuous, piecewise-linear function connecting the points $(0, \lambda_0), (t_1, \lambda_1), (t_2, \lambda_2), \dots$. Therefore, generating from the distribution F_t in Step (1) of Algorithm 2 amounts to identifying x such that $-\log(1-u) = \Lambda(t_i, x)$ where $\Lambda(t_i, x) \equiv \Lambda(t_i + x) - \Lambda(t_i)$. Through straightforward integration of the piecewise-linear function, Klein and Roberts [10] come up with an expression for the function $\Lambda(t_i, x)$ and the corresponding root x that satisfies $-\log(1-u) = \Lambda(t_i, x)$.

One important point needs to be noted in implementing the inversion methods that we have discussed. Depending on the nature of the expectation function $\Lambda(t)$, the distribution of the time until the next event may not be well defined. In particular, there may be a positive probability of observing no events past a particular point in time, that is, there exists $\delta > 0$ such that $\Pr\{X_i > x\} > \delta$ for all $x > 0$. In such a case, Step (3) in Çinlar's method may have no solution, and F_t in Equation (1) may not be a proper cdf.

Order Statistics

Let us now state a general result on the distribution of the event times of a nonhomogeneous Poisson process with expectation function $\Lambda(t)$.

Theorem 2 [Cox and Lewis, 1962]. *Let $T_1, T_2, \dots$ be random variables representing the event times of a nonhomogeneous Poisson process with continuous expectation function $\Lambda(t)$, and let N_t represent the total number of events occurring before time t in the process. Then, conditional on the number of events $N_{t_0} = n$, the event times $T_1, T_2, \dots, T_n$ are distributed as order statistics from a sample with distribution function $F(t) = \Lambda(t)/\Lambda(t_0)$ for $t \in [0, t_0]$.*

Theorem 2 is a generalization of the result for homogeneous Poisson processes. It naturally gives rise to Algorithm 3 for generating random variates from a nonhomogeneous Poisson process with expectation function $\Lambda(t)$ in a fixed interval $[0, t_0]$.

Algorithm 3.

- (1) Generate $n \sim \text{Poisson}(\Lambda(t_0))$.
- (2) Independently generate n random variates $t'_1, t'_2, \dots, t'_n$ from the cdf $F(t) = \Lambda(t)/\Lambda(t_0)$.
- (3) Order $t'_1, t'_2, \dots, t'_n$ to obtain $t_1 = t'_{(1)}, t_2 = t'_{(2)}, \dots, t_n = t'_{(n)}$.
- (4) Deliver $t_1, t_2, \dots, t_n$.

The efficiency of Algorithm 3 depends critically on Step (2), where n random variates are to be generated from the cdf $F(t)$. If $\lambda(t)$ is assumed to be log-linear ($\lambda(t) = \lambda e^{\alpha_1 t}$) as in Ref. 12, Algorithm 3 becomes very efficient since $F(t) = (e^{\alpha_1 t} - 1)/(e^{\alpha_1 t_0} - 1), t \in [0, t_0], \alpha_1 \neq 0$ is easily invertible. Lewis and Shedler, in the same article [12], provide an even faster variant based on gap statistics. This is extended to the case of second degree exponential polynomial rate functions (i.e., $\lambda(t) = \exp(\alpha_0 + \alpha_1 t + \alpha_2 t^2)$) in Ref. 13.

Acceptance-Rejection

Currently the most popular method for generating nonhomogeneous Poisson processes is the “process analogue” of acceptance-rejection called *thinning* [14]. The intuitive idea behind thinning is to first find a constant rate function $\lambda_u(t) = \lambda_u$, which dominates the desired rate function $\lambda(t)$, next generate from

the implied homogeneous Poisson process with rate $\lambda_u(t) = \lambda_u$, and then reject an appropriate fraction of the generated events so that the desired rate $\lambda(t)$ is achieved. The following theorem is the basis for this procedure.

Theorem 3 [Lewis and Shedler, 1979]. *Consider a nonhomogeneous Poisson process with rate function $\lambda_u(t), t \geq 0$. Suppose that $T_1^*, T_2^*, \dots, T_n^*$ are random variables representing event times from the nonhomogeneous Poisson process with rate function $\lambda_u(t)$, and lying in the fixed interval $(0, t_0]$. Let $\lambda(t)$ be a rate function such that $0 \leq \lambda(t) \leq \lambda_u(t)$ for all $t \in [0, t_0]$. If the i th event time T_i^* is independently deleted with probability $1 - \lambda(t)/\lambda_u(t)$ for $i = 1, 2, \dots, n$, then the remaining event times form a nonhomogeneous Poisson process with rate function $\lambda(t)$ in the interval $(0, t_0]$.*

A thinning algorithm for generating random variates from a nonhomogeneous Poisson process follows immediately.

Algorithm 4.

- (0) Initialize $t = 0$.
- (1) Generate $u_1 \sim U(0, 1)$.
- (2) Set $t \leftarrow t - \frac{1}{\lambda_u} \log u_1$.
- (3) Generate $u_2 \sim U(0, 1)$ independent of u_1 .
- (4) If $u_2 \leq \lambda(t)/\lambda_u$ then deliver t .
- (5) Go to Step (1).

In the above thinning algorithm, at any time t , a variate generated from the majorizing rate function $\lambda_u(t) = \lambda_u$ is accepted with probability $\lambda(t)/\lambda_u$. The efficiency thus depends critically on how “snugly” λ_u approximates $\lambda(t)$ —rate functions $\lambda(t)$ that exhibit heavy fluctuations in time will render the thinning algorithm inefficient. A proof that the thinning algorithm in fact produces the desired stochastic process can be found in Ref. 14.

Ross [15] provides a straightforward modification to the thinning method with the objective of mitigating excessive rejection (Also see Ref. 16, p. 179.) The intuitive idea behind this extension is *piecewise*

thinning—majorize the desired rate function using an appropriately chosen piecewise constant function having k pieces, and then perform regular thinning within each piece. Specifically, first divide the horizon of interest $[0, t_0]$ into k intervals $[s_{j-1}, s_j], j = 1, 2, \dots, k$, and choose constants λ_j satisfying $\lambda_j \geq \sup_{t \in [s_{j-1}, s_j]} \{\lambda(t)\}$. Random variates from a homogeneous Poisson process are then generated in the interval $[s_{j-1}, s_j], j = 1, 2, \dots, k$, by generating exponential random variates with mean $1/\lambda_j, j = 1, 2, \dots, k$. The resulting event times are each accepted independently with probability $\lambda(t)/\lambda_j, j = 1, 2, \dots, k$. Certain simple edge corrections are required when a generated exponential inter-event time straddles two pieces of the majorizing function. A complete algorithm listing can be found in Ref. 15 (p. 85).

While piecewise thinning alleviates excessive rejection, it can still be inefficient depending on how well the majorizing step function approximates the rate function. Modern methods, thus, go one step further than piecewise thinning and construct a majorizing function $\lambda_u(t)$ that is both a good approximation of the rate function and has a form that lends itself to easy generation. This results in a method that is a hybrid of two methods—usually an inversion method such as Algorithm 2 used to generate random variates from the constructed majorizing function $\lambda_u(t)$, and a thinning method to reject a correct fraction of the generated random variates. We now state this in the form of an algorithm.

Algorithm 5.

- (0) Initialize $t = 0$.
- (1) Generate $x \sim F_t$ where $F_t(x)$ is given in Equation (1) with $\Lambda(t) = \int_0^t \lambda_u(t) dt$.
- (2) Set $t \leftarrow t + x$.
- (3) Generate $u \sim U(0, 1)$ independent of x .
- (4) If $u \leq \lambda(t+x)/\lambda_u(t+x)$ then deliver t .
- (5) Go to Step (1).

Such a hybrid algorithm has been used with much success to generate a nonhomogeneous Poisson process on a fixed interval $[0, t_0]$ in

Ref. 11. The majorizing function $\lambda_u(t)$ used in Ref. 11 is a continuous piecewise-linear function that is optimal in a certain precise sense.

GENERATING TWO-DIMENSIONAL POISSON PROCESSES

A counting process $\{N(t), t \geq 0\}$ is said to constitute a two-dimensional nonhomogeneous Poisson process on $C \subseteq \mathbb{R}^2$ with rate function $\lambda(x, y) > 0$ if

- (i) The number of events in a region $R \subseteq C$ is Poisson distributed with parameter $\Lambda(R) = \iint_R \lambda(x, y) dx dy$.
- (ii) The number of events occurring in any finite set of nonoverlapping regions are mutually independent.

(We do not go into details about the complete characterization of the nature of C or the subsets R , for which (1) and (2) should be satisfied. See Ref. 8 (p. 303) for a more rigorous treatment.) The Poisson process is said to be a homogeneous Poisson process if $\lambda(x, y)$ in the above definition is constant on C , that is, $\lambda(x, y) = \lambda$ for $(x, y) \in C$ [17]. For generation from a two-dimensional homogeneous Poisson process, consult the section titled “Generation of Homogeneous Poisson Processes.”

For generating nonhomogeneous Poisson processes in a bounded region C , a simple first algorithm is based on the idea that the location (X, Y) of an event in C , conditional on the number of events $N = n$ that occur in C , is distributed over the region C according to the probability density function

$$f(x, y) = \frac{\lambda(x, y)}{\iint_C \lambda(x, y) dx dy}, \quad (x, y) \in C. \quad (3)$$

(See Chapter 4 of Ref. 8 for more on this. Particularly, the above idea extends seamlessly even into higher dimensions.) This, combined with the fact that the number of events N in C is Poisson distributed with mean $\iint_C \lambda(x, y) dx dy$, gives rise to a conceptually simple algorithm for generating events.

Algorithm 6.

- (1) Generate $n \sim \text{Poisson}(\iint_C \lambda(x, y) dx dy)$.
- (2) If $n = 0$ then exit. Otherwise independently generate n events $(x_i, y_i), i = 1, 2, \dots, n$ that are distributed in C according to the density $f(x, y)$ given in Equation (3).
- (3) Deliver $(x_i, y_i), i = 1, 2, \dots, n$.

Of course, the ease with which the above algorithm is implemented will depend on the nature of the rate function $\lambda(x, y)$. Particularly, it will depend on how easily $\lambda(x, y)$ is integrated over C , and whether $f(x, y)$ lends itself to easy generation. A lot is known about generation from continuous densities with arbitrary form—see Ref. 18 for more on this.

Another popular method for generating nonhomogeneous Poisson processes is the multidimensional analogue of thinning, and stems from a theorem by Lewis and Shedler [14].

Theorem 4 [Lewis and Shedler, 1979]. *Let $(X_i, Y_i), i = 1, 2, \dots$ be the Cartesian coordinates of a two-dimensional Poisson process on $C \subseteq \mathbb{R}^2$ with rate function $\lambda^*(x, y) \geq 0$. If the i th event is independently deleted with probability $1 - \lambda(x, y)/\lambda^*(x, y)$ where $0 \leq \lambda(x, y) \leq \lambda^*(x, y)$ for all $(x, y) \in C$, the resulting process is a Poisson process on C with rate function $\lambda(x, y)$.*

A basic thinning algorithm would thus involve identifying λ_u such that $\lambda_u \geq \lambda(x, y)$ for all $(x, y) \in C$, generating random variates from a homogeneous Poisson process with rate λ_u (see the section titled “Generation of Homogeneous Poisson Processes”), and then deleting events with probability $1 - \lambda(x, y)/\lambda_u$. As in one-dimensional thinning, such an algorithm would be very inefficient in contexts where the rate function $\lambda(x, y)$ exhibits major fluctuations. Methods that are analogous to piecewise thinning and the hybrid method for one-dimensional thinning, if devised, will potentially prove much more efficient.

A more recent idea for generating nonhomogeneous Poisson processes is provided

by Saltzman *et al.* [19], and is based on the idea of conditioning. (See also Section 4.10 of Ref. 8 for a deeper treatment of this idea.) Specifically, they note the following two facts about a two-dimensional nonhomogeneous Poisson process with rate function $\lambda(x, y) > 0$ on $C \subset \mathbb{R}^2$.

- (i) If $\{(X_i, Y_i)\}$ are events corresponding to the two-dimensional process, then the abscissae $\{X_i\}$ correspond to a one-dimensional nonhomogeneous Poisson process with rate function $\lambda_X(x) = \int_{C(x)} \lambda(x, y) dy$, where $C(x) = \{y : (x, y) \in C\}$.
- (ii) If (X, Y) denotes the location of an event from the two-dimensional process, the conditional random variable $Y|X = x$ has the probability density function $\lambda(x, y)/\lambda_X(x)$.

The above two facts give rise to a generation algorithm that first generates the abscissa of an event and then generates its ordinate through the conditional density function provided in (2). Like Algorithm 6, the efficiency of the resulting algorithm depends on the tractability of the rate functions $\lambda(x, y)$ and $\lambda_X(x)$. We summarize this as Algorithm 7 below.

Algorithm 7.

- (1) Initialize $i = 0$.
- (2) Generate x_i according to the one-dimensional nonhomogeneous Poisson process with rate function $\lambda_X(x)$.
- (3) Generate y_i according to the probability density function $\lambda(x_i, y)/\lambda_X(x_i)$.
- (4) Deliver (x_i, y_i) .
- (5) Set $i = i + 1$ and go to Step (2).

REFERENCES

1. Billingsley P. Probability and measure. New York: Wiley; 1995.
2. Serfling RJ. Approximation theorems of mathematical statistics. New York: John Wiley & Sons, Inc.; 1980.
3. Cox DR, Lewis PAW. The statistical analysis of series of events. London: Methuen; 1966.
4. Ross S. Stochastic processes. New York: Wiley; 1995.
5. Gnedenko BV, Kovalenko I. Introduction to queueing theory. Jerusalem: Israel Program for Scientific Translations, Jerusalem; 1968.
6. Çinlar E. Introduction to stochastic processes. New Jersey: Prentice-Hall; 1975.
7. Ortega JM, Rheinboldt WC. Iterative solution of nonlinear equations in several variables. New York: Academic Press; 1970.
8. Resnick S. Adventures in stochastic processes. Boston (MA): Birkhäuser; 1992.
9. Kuhl ME, Wilson JR. User's manual for mp3mle and mp3sim: software for estimating and simulating nonhomogeneous Poisson processes. Technical report. North Carolina State University, Department of Industrial Engineering; Raleigh, North Carolina; 1996.
10. Klein RW, Roberts SD. A time-varying Poisson arrival process generator. Simulation 1984;43(4):193–195.
11. Lee SH, Crawford MM, Wilson JR. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. Commun Stat Simul 1991;20(2 & 3):777–809.
12. Lewis PAW, Shedler GS. Simulation of nonhomogeneous Poisson processes with log-linear rate function. Biometrika 1976;63:501–505.
13. Lewis PAW, Shedler GS. Simulation of nonhomogeneous Poisson processes with degree-two exponential polynomial rate function. Oper Res 1979;27(5):1026–1039.
14. Lewis PAW, Shedler GS. Simulation of nonhomogeneous Poisson processes by thinning. Nav Res Log Q 1979;26(3):403–413.
15. Ross S. Simulation. Burlington (MA): Elsevier; 2006.
16. Bratley P, Fox BL, Schrage LE. A guide to simulation. New York: Springer; 1987.
17. Karlin S, Taylor HM. A first course in stochastic processes. New York: Academic Press; 1975.
18. Devroye L. Non-uniform random variate generation. New York: Springer; 1986.
19. Saltzman E, Drew J, Leemis L. Simulation of multivariate nonhomogeneous poisson processes by inversion and composition. ACM Trans Model Comput Simulation. In press.

GENERATING NONHOMOGENEOUS POISSON PROCESSES

RAGHU PASUPATHY

Department of Industrial and
Systems Engineering,
Virginia Tech, Blacksburg,
Virginia

Recall that a *counting process* $\{N_t, t \geq 0\}$ is a stochastic process defined on a sample space Ω such that for each $\omega \in \Omega$, the function $N_t(\omega)$ is a “realization” of the number of “events” happening in the interval $(0, t]$, with $N_0(\omega) = 0$. By this definition, $N_t(\omega)$ is automatically integer valued, nondecreasing, and right-continuous for each ω . A homogeneous Poisson process is a type of counting process that is characterized as follows:

Definition 1. A counting process $\{N_t, t \geq 0\}$ is called a homogeneous Poisson process if:

- (i) $\forall t, s \geq 0$, and $0 \leq u \leq t$, $N_{t+s} - N_t$ is independent of N_u ;
- (ii) $\forall t, s \geq 0$, $\Pr\{N_{t+s} - N_t \geq 2\} = o(s)$; and
- (iii) $\forall t, s \geq 0$, $\Pr\{N_{t+s} - N_t = 1\} = \lambda s + o(s)$, where $\lambda > 0$.

Definition 1 has been adopted from Ross [1], Chapter 2, with the notation $o(h)$ used in the usual sense—to denote a function $v(h)$ that satisfies $\lim_{h \rightarrow 0} v(h)/h = 0$ [2, p. 1]. The constant $\lambda > 0$ appearing in Definition 1 is called the *rate of the Poisson process*. Stipulation (i) ensures that the resulting process has *independent increments*, that is, the number of events happening in disjoint time intervals is independent. Likewise, the stipulations (ii) and (iii) ensure that the process has *stationary increments*, that is, the distribution of the number of events happening in an interval depends only on the length of the interval. Definition 1 implies that the time between events are independent and identically distributed

(i.i.d.) exponential random variables with mean $1/\lambda$ [1, p. 64]. Moreover, the number of events in any fixed interval of time with length t has a Poisson distribution with mean λt . The latter fact motivates an alternative equivalent definition of a Poisson process.

Definition 2. A counting process $\{N_t, t \geq 0\}$ is called a homogeneous Poisson process if:

- (i) $\forall t, s \geq 0$, and $0 \leq u \leq t$, $N_{t+s} - N_t$ is independent of N_u ; and
- (ii) $\forall t, s \geq 0$, $\Pr\{N_{t+s} - N_t = j\} = \exp\{-\lambda s\} \frac{(\lambda s)^j}{j!}$, $j = 0, 1, \dots$.

That Definitions 1 and 2 are equivalent follows in a rather straightforward manner (see Ref. 1, p. 61 for a proof).

A remarkable, completely qualitative definition of a homogeneous Poisson process is provided by Çinlar [3, p. 71].

Definition 3. A counting process $\{N_t, t \geq 0\}$ is called a homogeneous Poisson process if:

- (i) $\forall t \geq 0$ and almost all ω , the mapping $t \rightarrow N_t(\omega)$ has jumps of unit magnitude only; that is, $N_t(\omega) - \lim_{s \uparrow t} N_s(\omega) = 0$ or $1 \forall t \geq 0$ and for almost all ω ;
- (ii) $\forall t, s \geq 0$, and $0 \leq u \leq t$, $N_{t+s} - N_t$ is independent of N_u ; and
- (iii) $\forall t, s \geq 0$, the distribution of $N_{t+s} - N_t$ does not depend on t .

(Çinlar [3, p. 71] demonstrates that the above assumptions are sufficient to ensure that assumptions (ii) and (iii) of Definition 1 hold.)

In what follows, we discuss existing methods to generate pseudorandom numbers from a homogeneous Poisson process. By this we mean generating, on a digital computer, a realization of event times $\{t_i, i = 1, 2, \dots\}$ from a homogeneous Poisson process that is specified through its rate λ . (For a more general definition of a Poisson process and its generation on a digital computer, see **Generating Homogeneous Poisson Processes**.)

GENERATION ON THE NONNEGATIVE REAL LINE

Generating random variates from a homogeneous Poisson process is straightforward upon noting that the interevent times in such a process are i.i.d. exponential random variables with mean $1/\lambda$. This is owing to the fact that exponential random variates can be generated with ease through the cdf-inverse method [4,5]. Algorithm 1 thus obtains the required Poisson process realization by independently generating exponential random variates and then summing them.

Algorithm 1.

- (0) Initialize $t = 0$.
- (1) Generate $u \sim U(0, 1)$.
- (2) Set $t \leftarrow t - \frac{1}{\lambda} \log(u)$.
- (3) Deliver t .
- (4) Go to Step (1).

(Standard packages use more elaborate techniques that avoid the use of the “log” function in Step (2).) Another straightforward method to generate such realizations arises from a well-known property of a homogeneous Poisson process. Conditioned on the number of events $N(t_0) = n$ that occur in the fixed interval $[0, t_0]$, the event times $T_1, T_2, \dots, T_n$ of a homogeneous Poisson process are distributed as order statistics from a uniform distribution supported on $[0, t_0]$ [6]. (See *Generating Homogeneous Poisson Processes* for a more general result and a proof.) This property gives rise to an algorithm to generate events from a homogeneous Poisson process confined to the fixed interval $[0, t_0]$. We list this as Algorithm 2.

Algorithm 2.

- (1) Generate $n \sim \text{Poisson}(\lambda t)$.
- (2) If $n = 0$ then exit. Otherwise independently generate $u_1 \sim U(0, 1)$, $u_2 \sim U(0, 1)$, $\dots$, $u_n \sim U(0, 1)$, and sort (in increasing order) to obtain $u_{(1)}, u_{(2)}, \dots, u_{(n)}$.
- (3) Set $t_i \leftarrow t_0 u_{(i)}$, $i = 1, 2, \dots, n$.
- (4) Deliver t_i , $i = 1, 2, \dots, n$.

In Step (1), the Poisson random variate n can be generated using any number of available

methods, most notably table-based methods [7–10] and cdf-inverse adaptations [11]. See Ref. 5 for a complete treatment. The sorting operation in Step (2) is $O(n \log n)$ and so the efficiency of Algorithm 2 depends on the magnitudes of t_0 and λ .

GENERATION ON THE PLANE

A counting process $\{N(t), t \geq 0\}$ is said to constitute a two-dimensional homogeneous Poisson process on $C \subseteq \mathbf{R}^2$ with rate $\lambda \geq 0$ if:

- (i) the number of events in a region $R \subseteq C$ is Poisson distributed with parameter $\Lambda(R) = \int_R \lambda \, dV = \lambda A(R)$, where $A(R)$ is the volume of the region R ;
- (ii) the number of events occurring in any finite set of nonoverlapping regions are mutually independent.

(For a more general definition of a Poisson process on a plane, and corresponding generation methods on a digital computer, see *Generating Homogeneous Poisson Processes*.)

Generation from a two-dimensional homogeneous Poisson process with rate λ on a circle of radius $r > 0$ turns out to be easy based on the following result by Lewis and Shedler [12]. (See also Rubinstein and Kroese [13, p. 70].)

Theorem 1. [Lewis and Shedler, 1979]. Suppose $(R_1, \theta_1), (R_2, \theta_2), \dots, (R_N, \theta_N)$ are the polar coordinates of $N > 0$ events from a homogeneous Poisson process on the circle $C \equiv \{(x, y) : x^2 + y^2 \leq r^2\}$. Then, conditional on $N = n$, the ordered event radii $R_{(1)}, R_{(2)}, \dots, R_{(n)}$ are order statistics from the density $f(z) = 2z/r^2$, $z \in [0, r]$. Furthermore, $\theta_1, \theta_2, \dots, \theta_n$ are i.i.d. uniform on $[0, 2\pi]$ and independent of $R_1, R_2, \dots, R_n$.

The resulting algorithm for generating events from a homogeneous Poisson process on a circle of radius r thus involves first generating the number of events $N = n$ from a Poisson distribution with parameter $\pi r^2 \lambda$, then generating the radii $R_{(1)}, R_{(2)}, \dots, R_{(n)}$ as order statistics from the density $f(z)$,

and then finally generating the angles $\theta_1, \theta_2, \dots, \theta_n$ uniformly (and independently) from $[0, 2\pi]$.

Algorithm 3.

- (0) Generate $n \sim \text{Poisson}(\pi r^2 \lambda)$. If $n = 0$ then exit. Otherwise generate $u_1 \sim U(0, 1)$, $u_2 \sim U(0, 1), \dots, u_n \sim U(0, 1)$ independently.
- (1) Set $R_1 \leftarrow r\sqrt{u_1}, R_2 \leftarrow r\sqrt{u_2}, \dots, R_n \leftarrow r\sqrt{u_n}$.
- (2) Sort $R_1, R_2, \dots, R_n$ in ascending order to obtain $R_{(1)}, R_{(2)}, \dots, R_{(n)}$.
- (3) Generate $u_{n+1} \sim U(0, 1)$, $u_{n+2} \sim U(0, 1), \dots, u_{2n} \sim U(0, 1)$ independently.
- (4) Set $\theta_1 \leftarrow 2\pi u_{n+1}, \theta_2 \leftarrow 2\pi u_{n+2}, \dots, \theta_n \leftarrow 2\pi u_{2n}$.
- (5) Deliver $(R_{(1)}, \theta_1), (R_{(2)}, \theta_2), \dots, (R_{(n)}, \theta_n)$.

(In the above algorithm, Step (2) is not required if there is no notion of ordering that is implicit in the abscissa and ordinate.)

Ross [14] gives a method to extend Algorithm 3 to the context of more irregular regions. For instance, suppose we want to generate events from a homogeneous Poisson process in the region bounded by the first quadrant, the positive valued function $f(x)$, and the line $x = T$, that is, the region of interest $C \equiv \{(x, y) : x \geq 0, y \geq 0, y \leq f(x), x \leq T\}$. Ross [14] shows that the order statistics $X_{(1)}, X_{(2)}, \dots, X_{(N)}$ of the x -coordinates of the events are such that $\int_{X_{(i-1)}}^{X_{(i)}} f(x) dx, X_{(0)} = 0, i = 1, 2, \dots$, are i.i.d. exponential with mean $1/\lambda$. The corresponding i th y -coordinate Y_i is uniformly distributed between 0 and $f(X_{(i)})$. This gives rise to an algorithm that is straightforward, at least in principle.

Algorithm 4.

- (0) Initialize $j = 1, n = 0, w = 0, x_0 = 0$.
- (1) Generate $u_j \sim U(0, 1)$ independently.
- (2) Set $w_j \leftarrow -(1/\lambda) \ln(1 - u_j)$.
- (3) Set $w \leftarrow w + w_j$.
- (4) If $w \leq \int_0^T f(x) dx$ then set $n \leftarrow n + 1$ and go to Step (1).
- (5) If $n = 0$ then exit. Otherwise solve for $x_{(i)}, i = 1, 2, \dots, n$ satisfying $\int_{x_{(i-1)}}^{x_{(i)}} f(x) dx = w_i$.
- (6) Generate $u_{n+1} \sim U(0, 1), u_{n+2} \sim U(0, 1), \dots, u_{2n} \sim U(0, 1)$ independently.
- (7) Set $y_i \leftarrow u_{n+if}(x_i)$.
- (8) Deliver $(x_{(i)}, y_i), i = 1, 2, \dots, n$.

The efficiency of Algorithm 4 depends critically on the inversion in Step (5). Theorem 1 can be generalized even further to give a result that is the multidimensional analogue of the order statistics property discussed in the section titled “Generation on the Nonnegative Real Line.” Specifically, consider a two-dimensional homogeneous Poisson process defined on an arbitrary region C . Then it is seen somewhat easily that the location (X, Y) of an event on the plane, conditional on the number of events $N = n$, is uniformly distributed over the region C [15, p. 359]. This then gives rise to a very simple algorithm that is analogous to Algorithm 2, and generalizes Algorithm 3.

Algorithm 5.

- (1) Generate $n \sim \text{Poisson}(\lambda A(C))$.
- (2) If $n = 0$ then exit. Otherwise independently generate n points $(x_i, y_i), i = 1, 2, \dots, n$ that are uniformly distributed in C .
- (3) Deliver $(x_i, y_i), i = 1, 2, \dots, n$.

If the region C is irregular, Step 2 may not be straightforward. In such a case, a two-dimensional homogeneous process can be generated on any convenient region $\bar{C}$ (e.g., a rectangle) that covers C , and the events falling within C can be delivered as a realization of the required process.

REFERENCES

1. Ross S. Stochastic processes. New York: Wiley; 1995.
2. Serfling RJ. Approximation theorems of mathematical statistics. New York: John Wiley & Sons, Inc.; 1980.
3. Çinlar E. Introduction to stochastic processes. New Jersey: Prentice-Hall; 1975.
4. Asmussen S, Glynn PW. Stochastic simulation: algorithms and analysis. New York: Springer; 2007.
5. Devroye L. Non-uniform random variate generation. New York: Springer; 1986.
6. Lewis PAW, Shedler GS. Simulation of nonhomogenous Poisson processes with log-linear rate function. Biometrika 1976;63:501–505.
7. Chen HC, Asau Y. On generating random variates from an empirical distribution. AIIE Trans 1974;6:163–166.

8. Fishman GS, Moore LR. Sampling from a discrete distribution while preserving monotonicity. *Am Stat* 1984;38:219–223.
9. Peterson AV Jr, Kronmal RA. Analytic comparison of three general-purpose methods for the computer generation of discrete random variables. *Appl Stat* 1983;32(3):276–286.
10. Walker AJ. An efficient method for generating discrete random variables with general distributions. *ACM Trans Math Softw* 1977;3:253–256.
11. Kemp CD, Kemp AW. Poisson random variate generation. *Appl Stat* 1991;40(1):143–158.
12. Lewis PAW, Shedler GS. Simulation of nonhomogenous Poisson processes by thinning. *Nav Res Log Q* 1979;26(3):403–413.
13. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo method*. New York: John Wiley & Sons, Inc.; 2007.
14. Ross S. *Simulation*. Burlington (VT): Elsevier; 2006.
15. Resnick S. *Adventures in stochastic processes*. Boston (MA): Birkhäuser; 1992.

GENERIC STOCHASTIC GRADIENT METHODS

BERNARD BERCU
Université Bordeaux 1, Institut de
Mathématiques de Bordeaux,
Talence Cedex, France

JEAN-CLAUDE FORT
Université Paris Descartes, Paris
Cedex, France

STOCHASTIC GRADIENT METHODS

Gradient descent methods (Borkar [1], Guler [2], Spall [3]) are based on the elementary observation that if f is a continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$, then f decreases faster in the opposite direction of its gradient ∇f . More precisely, any solution $\theta(t)$ of the differential equation

$$\frac{\partial \theta}{\partial t} = -\nabla f(\theta) \quad (1)$$

satisfies

$$\frac{\partial f(\theta(t))}{\partial t} = \left\langle \nabla f(\theta(t)), \frac{\partial \theta}{\partial t} \right\rangle = -\|\nabla f(\theta(t))\|^2,$$

which indicates that f is decreasing on the path $\theta(t)$. Suppose that now f is a convex function, then f reaches a minimum at a critical point θ^* such that $\nabla f(\theta^*) = 0$, and $\theta(t)$ converges to θ^* as t tends to infinity. Gradient descent methods have been introduced in order to estimate the critical points θ^* via a discretization of the previous differential equation. They are based on the recursive construction of a sequence (θ_n) satisfying

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n)$$

where the step size γ_n has to be carefully chosen. In general, the assumptions needed for the convergence of θ_n to θ^* are that the sequence (γ_n) decreases to zero and that

$$\sum_{n=1}^{\infty} \gamma_n = \infty.$$

We can observe that if the above sum converges, then θ_n may not converge to θ^* . From a practical point of view, one can choose

$$\gamma_n = \frac{c}{d+n}$$

with a suitable choice of the positive constants c and d . More details on the practical choice of the sequence (γ_n) may be found in Spall ([3], p. 106). Let us consider the illustrative example of quadratic forms on $\mathbb{R}^d$. Suppose that

$$f(\theta) = \theta' A \theta + \langle b, \theta \rangle$$

where A is a symmetric, positive-definite square matrix of order d and $b \in \mathbb{R}^d$. Then, we clearly have $\nabla f(\theta) = 2A\theta + b$ so that f reaches a minimum at $\theta^* = -A^{-1}b/2$. Then, the gradient descent algorithm can be rewritten as

$$\theta_{n+1} = \theta_n - \gamma_{n+1}(2A\theta_n + b).$$

Then, $\theta_{n+1} - \theta^* = (I - 2A\gamma_{n+1})(\theta_n - \theta^*)$, which implies that θ_n converges to θ^* as soon as the step size $\gamma_n \leq 1/(2\lambda_d)$, where λ_d stands for the largest eigenvalue of A .

We shall now focus our attention on stochastic gradient methods. Let (X_n) be a sequence of random vectors with values in $\mathbb{R}^p$. For a continuously differentiable function f and a given value $\theta_n \in \mathbb{R}^d$, we suppose that we observe or we are able to evaluate a random vector $H(\theta_n, X_{n+1})$ such that

$$\mathbb{E}[H(\theta_n, X_{n+1}) | \mathcal{F}_n] = \nabla f(\theta_n) \quad a.s.$$

where $\mathcal{F}_n$ stands for the σ -algebra of the events occurring up to time n .

Then, in order to find the critical points, in fact the local minima of f , we make use of the stochastic gradient algorithm

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, X_{n+1}) \quad (2)$$

where the initial value θ_0 can be arbitrarily chosen and (γ_n) is a deterministic positive sequence decreasing toward zero, satisfying

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

We clearly have

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \varepsilon_{n+1},$$

where $\varepsilon_{n+1} = \nabla f(\theta_n) - H(\theta_n, X_{n+1})$. Consequently, the stochastic gradient algorithm appears as a perturbed discretization of the ordinary differential equation given by (1). We can observe that the stochastic gradient method belongs to the wide family of stochastic algorithms. Moreover, the Robbins–Monro algorithm (Robbins and Monro [4], Robbins and Siegmund [5]) is a very efficient procedure to estimate the zeros of a nonlinear continuous function F . The stochastic gradient algorithm can be seen as a special case of Robbins–Monro algorithm with $F = \nabla f$.

The Kiefer and Wolfowitz algorithm [6,7] was proposed to search the maxima of a strictly concave function f .

It only uses finite difference of perturbed evaluations of the function f . It is given by the recursive relation

$$\theta_{n+1} = \theta_n - \frac{\gamma_{n+1}}{2c_{n+1}} (Y_{n+1} - Z_{n+1}) \quad (3)$$

with

$$\mathbb{E}[Y_{n+1} | \mathcal{F}_n] = f(\theta_n - c_n) \quad \text{and}$$

$$\mathbb{E}[Z_{n+1} | \mathcal{F}_n] = f(\theta_n + c_n) \quad a.s.,$$

where (γ_n) and (c_n) are deterministic positive sequences decreasing to zero such that

$$\sum_{n=1}^{\infty} \gamma_n = \infty, \quad \sum_{n=1}^{\infty} \gamma_n c_n < \infty, \quad \sum_{n=1}^{\infty} \left(\frac{\gamma_n}{c_n} \right)^2 < \infty.$$

A persistently excited version of the Kiefer–Wolfowitz algorithm is the simultaneous perturbation stochastic approximation introduced by Spall [3,8]. A recent survey concerning Spall algorithm as well as some

developments of other finite-difference algorithms along the same lines may be found in Bhatnagar [9]. We also refer the reader to the related contributions of Gerencser [10] and Granichin [11–13].

One can realize that these stochastic algorithms are widely used in many fields such as competitive learning, signal processing, wireless protocol of communication, financial engineering, and stock management.

MAIN RESULTS

Almost Sure Convergence

We start with the simplest situation in which the Robbins–Monro procedure can be applied [4,5].

Theorem 1. *Assume that f is a continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$ and denote by θ^* an isolated critical point of f . In addition, suppose that for all $\theta \neq \theta^*$,*

$$\langle \theta - \theta^*, \nabla f(\theta) \rangle > 0. \quad (4)$$

Consider the stochastic gradient algorithm given by (2), where (γ_n) is a deterministic positive sequence decreasing to zero such that

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty. \quad (5)$$

If it exists $a > 0$ such that for all $\theta \in \mathbb{R}^d$, $\mathbb{E}[\|H(\theta, X_{n+1})\|^2 | \mathcal{F}_n] \leq a(1 + \|\theta\|^2)$ a.s., then

$$\lim_{n \rightarrow \infty} \theta_n = \theta^* \quad a.s. \quad (6)$$

Assumption (4) indicates that there exists a unique minimum for f , which is also a global minimum. In this situation, the Robbins–Monro algorithm converges almost surely to this global minimum. Similar result is also true for the Kiefer–Wolfowitz algorithm under suitable assumptions [6,7]. One can obtain the optimal rates of convergence by averaging [14]. In a more general setting, Polyak and Tsybakov [15] have also investigated the optimal rates of convergence when considering not only a function f but also a class of smooth functions.

Hereafter, we consider a more general situation Brandiere [16], Brandiere and Duflo [17], Lazarev [18], Pemantle [19] leading to less precise results.

Theorem 2. *Assume that f is a twice continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$ and denoted by Θ^* the set of all critical points of f . Assume that Θ^* is made of isolated points and that*

$$\lim_{\|\theta\| \rightarrow \infty} |f(\theta)| = \infty.$$

In addition, suppose that for any $\theta^ \in \Theta^*$, either the Hessian matrix $\nabla^2 f(\theta^*)$ is positive definite, which indicates that θ^* is a nondegenerate local minimum of f , or that it exists at least a direction $v^* \in \mathbb{R}^d$ depending on θ^* , such that $\langle v^*, \nabla^2 f(\theta^*) v^* \rangle < 0$, which indicates that θ^* is a saddle point or a local maximum of f . Consider the stochastic gradient algorithm given by (2), where the sequence (γ_n) satisfies (5). Assume that it exists $a > 0$ such that for all $\theta \in \mathbb{R}^d$, $\mathbb{E}[\|H(\theta, X_{n+1})\|^2 | \mathcal{F}_n] \leq a(1 + \|\theta\|^2)$ a.s. Furthermore, assume that it exists $c > 0$ such that for all unit $v \in \mathbb{R}^d$*

$$\mathbb{E}[\langle \varepsilon_{n+1}, v \rangle | \mathcal{F}_n] \geq c > 0, \quad (7)$$

where $\varepsilon_{n+1} = H(\theta_n, X_{n+1}) - \nabla f(\theta_n)$, which indicates that the random noise ε_{n+1} weights all possible directions of $\mathbb{R}^d$. Then,

$$\lim_{n \rightarrow \infty} \theta_n = \theta^* \quad \text{a.s.}, \quad (8)$$

where the random vector θ^ belongs to the set of the local minima of f in Θ^* .*

Theorem 2 gives a reasonable condition to ensure that the algorithm converges to a local minimum avoiding saddle points or local maxima. Assumption (7) indicates that there is always noise in an unstable direction of any saddle point or maximum. One way to realize this condition is to add an artificial Gaussian noise in all directions.

Central Limit Theorem

In all the sequel, denoted by

$$\mathcal{C}(\theta^*) = \{\theta_n \rightarrow \theta^*\}$$

the event corresponding to the convergence of the stochastic gradient algorithm to θ^* . The asymptotic normality of the gradient algorithm requires more conditions on the function f and on the step size γ_n . It was established by Fabian [20] for the Robbins–Monro algorithm and Bouton [21] in a more general framework, see also Kushner and Huang [22] and Major and Revesz [23].

We shall restrict ourselves to the particular case where $\gamma_n = \gamma/n$ with $\gamma > 0$.

Theorem 3. *Assume that f is a twice continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$ and denoted by θ^* an isolated critical point of f . In addition, suppose that for all $\theta \neq \theta^*$,*

$$\nabla f(\theta) = \nabla^2 f(\theta^*)(\theta - \theta^*) + O(\|\theta - \theta^*\|^2), \quad (9)$$

where the Hessian matrix $\nabla^2 f(\theta^)$ is positive definite. More precisely, assume that its minimum eigenvalue λ satisfies $\lambda > \zeta$, where $\zeta = 1/(2\gamma)$. If it exists some constant $b > 2$ such that*

$$\sup_{n \geq 1} \mathbb{E}[\|\varepsilon_{n+1}\|^b | \mathcal{F}_n] < \infty \quad \text{a.s.} \quad (10)$$

and

$$\lim_{n \rightarrow \infty} \mathbb{E}[\varepsilon_{n+1} \varepsilon'_{n+1} | \mathcal{F}_n] = \Gamma \quad \text{a.s.}, \quad (11)$$

where $\varepsilon_{n+1} = H(\theta_n, X_{n+1}) - \nabla f(\theta_n)$, and Γ is a positive-definite covariance matrix, then we have on $\mathcal{C}(\theta^)$, the asymptotic normality*

$$\sqrt{n}(\theta_n - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \gamma \Sigma), \quad (12)$$

where the limiting covariance matrix Σ is the solution of the Lyapunov equation

$$(\nabla^2 f(\theta^*) - \zeta I_d) \Sigma + \Sigma (\nabla^2 f(\theta^*) - \zeta I_d) = \Gamma.$$

Remark 1. The solution of the Lyapunov equation is given by

$$\Sigma = \int_0^\infty \exp(-t(\nabla^2 f(\theta^*) - \zeta I_d)) \Gamma \exp(-t(\nabla^2 f(\theta^*) - \zeta I_d)) dt.$$

Theorem 3 also holds in the particular case where $\gamma_n = \gamma/n^\alpha$ with $\gamma > 0$ and $1/2 < \alpha < 1$.

Remark 2. It could be possible to obtain the asymptotic efficiency for the stochastic gradient algorithm by choosing the matrix step size

$$\gamma_n = \frac{(\nabla^2 f(\theta^*))^{-1}}{n}.$$

In this particular situation that is not far from Newton's stochastic algorithm, we can deduce from Kushner and Yin ([24] p. 302) that θ_n is an asymptotically efficient estimator of θ^* with

$$\sqrt{n}(\theta_n - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (\nabla^2 f(\theta^*))^{-1/2} \Sigma (\nabla^2 f(\theta^*))^{-1/2}).$$

However, in most cases, it is not possible to explicitly calculate $\nabla^2 f(\theta^*)$.

The asymptotic normality given by (12) allows to build confidence intervals around the value of θ^* . It only requires the estimation of the limiting covariance matrix Σ . We will see in the following text that the quadratic strong law enables us to compute such an estimate.

Almost Sure Central Limit Theorem

We shall now focus our attention on the law of iterated logarithm given by Gaposhkin and Krasulina [25] and the quadratic strong law established by Pelletier [26] associated with the stochastic gradient algorithm.

Theorem 4. *Under the same assumptions as in Theorem 3, we have almost surely on $\mathcal{C}(\theta^*)$ and for all eigenvectors $v \in \mathbb{R}^d$ of the Hessian matrix $\nabla^2 f(\theta^*)$.*

$$\begin{aligned} \limsup_{n \rightarrow \infty} \left(\frac{n}{2 \log \log n} \right)^{1/2} \langle v, \theta_n - \theta^* \rangle \\ = - \liminf_{n \rightarrow \infty} \left(\frac{n}{2 \log \log n} \right)^{1/2} \langle v, \theta_n - \theta^* \rangle \\ = \sqrt{\gamma v' \Sigma v} \quad a.s. \end{aligned} \quad (13)$$

In particular,

$$\limsup_{n \rightarrow \infty} \left(\frac{n}{2 \log \log n} \right) \|\theta_n - \theta^*\|^2 \leq \gamma \text{tr}(P' \Sigma P) \quad a.s., \quad (14)$$

where P is the orthogonal matrix whose columns are the eigenvectors of $\nabla^2 f(\theta^*)$. Finally, almost surely on $\mathcal{C}(\theta^*)$, we have the quadratic strong law

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n (\theta_k - \theta^*)(\theta_k - \theta^*)' = \gamma \Sigma \quad a.s. \quad (15)$$

An easy way to estimate the covariance matrix in (12) is to make use of the quadratic strong law (15) replacing θ^* by θ_n .

Remark 3. Theorem 4 is also true in the particular case where $\gamma_n = \gamma/n^\alpha$ with $\gamma > 0$ and $1/2 < \alpha < 1$, as soon as $\alpha b > 2$. This is of course immediately satisfied if $b \geq 4$. According to Pelletier [26], it is much more appropriate to make use of $\gamma_n = \gamma/n$. As a matter of fact, if $\gamma_n = \gamma/n^\alpha$, the quadratic strong law (15) becomes

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n^{1-\alpha}} \sum_{k=1}^n (\theta_k - \theta^*)(\theta_k - \theta^*)' \\ = \frac{\gamma}{1-\alpha} \Sigma \quad a.s. \end{aligned} \quad (16)$$

As $\log n$ is always lesser than $n^{1-\alpha}$, the choice $\gamma_n = \gamma/n$ outperforms the second one.

The almost sure central limit theorem was established by Pelletier [27], see Cenac and Cenac [28] for a recent contribution on the convergence of moments.

Theorem 5. *Under the same assumptions as in Theorem 4, we have almost surely on $\mathcal{C}(\theta^*)$, the almost sure central limit theorem*

$$\frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} \delta_{\sqrt{k}(\theta_k - \theta^*)} \Rightarrow \mathcal{N}(0, \gamma \Sigma) \quad a.s. \quad (17)$$

Remark 4. We can observe that the quadratic strong law given by (15) is simply the convergence of the second moment in (17). In addition, for any continuous function h from $\mathbb{R}^d$ to $\mathbb{R}$ such that for some constant $c > 0$, $|h(x)| \leq c(1 + \|x\|^2)$, we have almost surely on $\mathcal{C}(\theta^*)$,

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} h(\sqrt{k}(\theta_k - \theta^*)) = \mathbb{E}[h(Z)] \quad a.s.,$$

where Z stands for a random vector with $\mathcal{N}(0, \gamma \Sigma)$ distribution.

EXAMPLES

Least-Square Approximation

Assume that we observe an input–output relationship

$$Y = F(X) + \xi,$$

where $X \in \mathbb{R}^p$ is the input, $Y \in \mathbb{R}$ is the output, and ξ is the driven noise. The link function F is unknown and we look for an approximation of F from a basis of functions $(\varphi_1, \dots, \varphi_d)$ from $\mathbb{R}^p$ to $\mathbb{R}$. We can modelize the input as random vectors and thus we want to minimize, for instance,

$$f(\theta) = \frac{1}{2} \mathbb{E}[(Y - \langle \theta, \Phi(X) \rangle)^2],$$

where

$$\Phi(X) = \begin{pmatrix} \varphi_1(X) \\ \vdots \\ \varphi_d(X) \end{pmatrix}.$$

We suppose that Y and $\Phi(X)$ are square integrable. We clearly have

$$\begin{aligned} \nabla f(\theta) &= -\mathbb{E}[(Y - \langle \theta, \Phi(X) \rangle) \Phi(X)] \quad \text{and} \\ \nabla^2 f(\theta) &= \mathbb{E}[\Phi(X) \Phi(X)']. \end{aligned}$$

Assume that the matrix $\mathbb{E}[\Phi(X) \Phi(X)']$ is positive definite and denoted by θ^* the minimizer of the function f , which is the solution of the equation

$$\mathbb{E}[\Phi(X) \Phi(X)'] \theta = \mathbb{E}[Y \Phi(X)].$$

We can estimate θ^* by the stochastic gradient algorithm

$$\theta_{n+1} = \theta_n - \gamma_{n+1} (Y_{n+1} - \langle \theta_n, \Phi(X_{n+1}) \rangle) \Phi(X_{n+1}),$$

where (X_n, Y_n) is a sequence of independent random vectors sharing the same distribution as (X, Y) . All the results of the previous section are satisfied under additional moment conditions.

Empirical Quantiles

Let X be a random variable with continuous and strictly increasing distribution function F . We are interested in the estimation of the α -quantile θ_α of F , which satisfies

$$F(\theta_\alpha) = \mathbb{P}(X \leq \theta_\alpha) = \alpha.$$

Consider the function

$$f(\theta) = \int_{-\infty}^{\theta} F(x) dx - \alpha \theta.$$

We clearly have $f'(\theta) = F(\theta) - \alpha$ and θ_α is the unique minimum of f . In addition, if $H(\theta, X) = I_{\{X \leq \theta\}} - \alpha$, we can immediately see that $\mathbb{E}[H(\theta, X)] = f'(\theta)$. Consequently, for any $0 < \alpha < 1$, we can estimate θ_α by the gradient descent algorithm

$$\theta_{n+1} = \theta_n - \gamma_{n+1} (I_{\{X_{n+1} \leq \theta_n\}} - \alpha),$$

where (X_n) is a sequence of independent random variables sharing the same distribution as X . As before, all the results of the previous section are satisfied.

Maximum Likelihood Estimation

Let X be a random vector with values in $\mathbb{R}^p$. Assume that the distribution function F_θ of X depends on an unknown parameter $\theta \in \mathbb{R}^d$ and let $\varphi(\theta, \cdot)$ be the probability density function associated with X . In what follows, suppose that the function $\varphi(\theta, \cdot)$ is known. Our goal is to find θ^* that maximizes $\mathbb{E}[\log \varphi(\theta, X)]$. This maximization problem is of course the same as the minimization of $-\mathbb{E}[\log \varphi(\theta, X)]$. Denote

$$H(\theta, X) = \nabla \log \varphi(\theta, X) = \frac{\nabla \varphi(\theta, X)}{\varphi(\theta, X)},$$

where the gradient is taken with respect to θ and suppose that

$$\mathbb{E}[H(\theta^*, X)] = \nabla f(\theta^*) = 0.$$

Moreover, denote by $I(\theta)$ the Fisher information matrix

$$I(\theta) = \mathbb{E}[\nabla \log \varphi(\theta, X) (\nabla \log \varphi(\theta, X))']$$

and assume that $I(\theta^*)$ is positive definite. If the maximum likelihood estimator of θ is very complicated to calculate, we can estimate θ^* by the stochastic gradient algorithm

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(\theta_n, X_{n+1}),$$

where (X_n) is a sequence of independent random vectors sharing the same distribution as X . Under some regularity assumptions, all the results of the previous section are satisfied. We can also make use of a Newton-like stochastic algorithm

$$\theta_{n+1} = \theta_n + \gamma_{n+1} I(\theta_n)^{-1} H(\theta_n, X_{n+1}),$$

which performs pretty well as in the deterministic case.

Quantization

Quantization is a field where the stochastic gradient is the most used method. Let us explain what is quantization as in Graf and Luschgy [29]. We assume that a phenomenon is well described by a continuous random vector X with values in $\mathbb{R}^p$.

In order to drastically simplify the information contained in X , for example, in view of transmission, we look for a code book with finitely many values that represents, as faithfully as possible, the random vector X . Hence, our goal is to find a discrete random vector Y with q values, which minimizes the quadratic distance

$$\mathbb{E}[\|X - Y\|^2] = \min_{\theta, P} \mathbb{E} \left[\sum_{i=1}^q \|X - \theta_i\|^2 I_{P_i} \right],$$

where $\theta = (\theta_1, \dots, \theta_q) \in (\mathbb{R}^p)^q$ is associated with the discrete values of Y and $P \in \mathcal{P}$ is a partition of the probability space. We can easily observe that, for a given $\theta \in \mathbb{R}^{pq}$, the best choice for P is the Voronoï partition associated with θ , that is $V = (V_1, \dots, V_q)$, where

$$V_i = \left\{ x \in \mathbb{R}^p / \min_{1 \leq k \leq q} \|x - \theta_k\|^2 = \|x - \theta_i\|^2 \right\}.$$

Consequently, our goal is to minimize

$$f(\theta) = \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^q \|X - \theta_i\|^2 I_{V_i}(X) \right],$$

where $V = (V_1, \dots, V_q)$ is the Voronoï partition associated with θ , so that f is well defined as a function of θ . Such minimizers θ^* are called *optimal quantizers* and generally are not unique.

In the particular case $p = 1$, if we know the probability distribution of X , for example, exponential, normal, or Gamma distribution, we can use a deterministic gradient method, namely, the Newton algorithm, in order to find optimal quantizers very accurately.

As soon as $p \geq 2$, we cannot use such an algorithm, because of the high complexity of the Voronoï partition. Therefore, an alternative is to make use of the stochastic gradient algorithm. It is given, for all $1 \leq i \leq q$, by

$$\theta_{i,n+1} = \theta_{i,n} - \gamma_{n+1} (\theta_{i,n} - X_{n+1}) I_{V_{i,n}}(X_{n+1}),$$

where (X_n) is a sequence of independent random vectors sharing the same distribution as X and $V_n = (V_{1,n}, \dots, V_{q,n})$ is the Voronoï partition associated with θ_n .

We can observe that, from step n to step $n + 1$, only one quantizer is modified. It is precisely the one for which X_{n+1} falls in the Voronoï cell.

The quantization can be extended to stochastic processes. Once again, stochastic gradient algorithm in Hilbert spaces is the most popular way to find a numerical approximation of optimal quantizers.

SUMMARY

The method of gradient descent was probably introduced by Cauchy in early nineteenth century with Cauchy's intermediate value theorem. In the middle of nineteenth century, pushed by economy, the optimization became a very important field of mathematics. During all the first part of the twentieth century, scientific advances arose. One of the pioneers in this area was Hotelling [30] who proposed an experimental way to find the value θ^* for which an unknown polynomial function f reaches a maximum or minimum.

At the half of twentieth century, optimization in a random framework appears in various domains: physics, chimics, transmission of information, etc., leading to a

new development of optimization techniques: stochastic optimization such as stochastic gradient methods.

Since the seminal work of Robbins and Monro [4], a wide range of literature is available on stochastic gradient methods. It is very difficult to provide an exhaustive list of references on this domain of applied mathematics.

We refer the reader to the excellent books of Benveniste et al. [31], Duflo [32], Hall and Heyde [33], Kushner and Yin [24], Ljung et al. [34], and Spall [3], and the references therein. A whole history remains to be written. We strongly believe that it is possible to provide much more applications of stochastic gradient methods in applied mathematics.

REFERENCES

1. Borkar VS. Stochastic approximation: a dynamical systems viewpoint. Cambridge: Cambridge University Press; 2008.
2. Guler O. Foundations of optimization. Volume 258: Graduate texts in mathematics. New York: Springer; 2010.
3. Spall JC. Introduction to stochastic search and optimization: estimation, simulation, and control, Wiley-Interscience Series in Discrete Mathematics and Optimization. Hoboken (NJ): Wiley-Interscience; 2003.
4. Robbins H., Monro S. A stochastic approximation method. *Ann Math Stat* 1951; 22: 400–407.
5. Robbins H., Siegmund D. A convergence theorem for non negative almost supermartingales and some applications. In: *Optimizing methods in statistics*. Ohio State Univ, New York: Academic Press; 1971. p 233–257.
6. Kiefer J., Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Ann Math Stat* 1952;23:462–466.
7. Wolfowitz J. On the stochastic approximation method of Robbins and Monro. *Ann Math Stat* 1952;23:457–461.
8. Spall JC. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans Automat Control* 1992;37:332–341.
9. Bhatnagar S. Simultaneous perturbation and finite difference methods. *Wiley Encycl Oper Res Manage Sci* 2011;7:4969–4991.
10. Gerencsér L. Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans Automat Control* 1999;44:894–905.
11. Granichin ON. A stochastic approximation procedure with disturbance at the input. *Avtomat Telemekh* 1992;2: 97–104.
12. Granichin ON. Randomized algorithms for stochastic approximation under arbitrary noise. *Avtomat Telemekh* 2002;2:44–55.
13. Granichin ON. Optimal rate of convergence of randomized algorithms for stochastic approximation under arbitrary noise. *Avtomat Telemekh* 2003;2:88–99.
14. Polyak BT., Juditsky A. Acceleration of stochastic approximation by averaging. *SIAM J Control Optim* 1992;30(4):838–855.
15. Polyak BT., Tsybakov AB. Optimal orders of accuracy for search algorithms of stochastic optimization. *Problemy Peredachi Informatsii* 1990;26:45–53.
16. Brandière O. Some pathological traps for stochastic approximation. *SIAM J Control Optim* 1998;36(4):1293–1314.
17. Brandière O., Duflo M. Les algorithmes stochastiques contournent-ils les pièges? *Ann Inst H Poincaré Prob Stat* 1996;32(3): 395–427.
18. Lazarev VA. Convergence of stochastic approximation procedures in the case of several roots of a regression equation. *Problemy Peredachi Informatsii* 1992;28(1):75–88.
19. Pemantle R. Nonconvergence to unstable points in urn models and stochastic approximations. *Ann Prob* 1990;18(2):698–712.
20. Fabian V. On asymptotic normality in stochastic approximation. *Ann Math Stat* 1968;39: 1327–1332.
21. Bouton C. Approximation gaussienne d'algorithmes stochastiques à dynamique markovienne. *Ann Inst H Poincaré Prob Stat* 1988;24(1):131–155.
22. Kushner HJ., Huang H. Rates of convergence for stochastic approximation type algorithms. *SIAM J Control Optim* 1979;17(5):607–617.
23. Major P., Révész P. A limit theorem for the Robbins-Monro approximation. *Z Wahrscheinlichkeitstheorie und Verw Gebiete* 1973;27:79–86.
24. Kushner HJ., Yin GG. Stochastic approximation and recursive algorithms and applications. Volume 35: Stochastic modelling and applied probability, Applications of Mathematics (New York) 2nd ed. New York: Springer-Verlag, 2003.

25. Gaposhkin VF., Krasulina TP. The law of the iterated logarithm in stochastic approximation processes. *Teor Veroyatnost i Primenen* 1974;19:879–886.
26. Pelletier M. On the almost sure asymptotic behaviour of stochastic algorithms. *Stoch Process Appl* 1998;78(2):217–244.
27. Pelletier M. An almost sure central limit theorem for stochastic approximation algorithms. *J Multivariate Anal* 1999; 71(1):76–93.
28. Cenac P. On the convergence of moments in the almost sure central limit theorem for stochastic approximation algorithms. *ESAIM Prob Stat* 2013;17:179–194.
29. Graf S., Luschgy H. Foundations of quantization for probability distributions. Volume 1730: *Lecture Notes in Mathematics*. Berlin: Springer-Verlag; 2000.
30. Hotelling H. Experimental determination of the maximum of a function. *Ann Math Stat* 1941;12:20–45.
31. Benveniste A., Métivier M., Priouret P. Adaptive algorithms and stochastic approximations. Volume 22: *Applications of mathematics*. Berlin: Springer-Verlag; 1990.
32. Duflo M. Algorithmes stochastiques. Volume 23: *Mathématiques & Applications*. Berlin: Springer-Verlag; 1996.
33. Hall P., Heyde CC. Martingale limit theory and its application, *Probability and mathematical statistics*. New York: Academic Press Inc.; 1980.
34. Ljung L., Pflug G., Walk H. Stochastic approximation and optimization of random systems. Volume 17: *DMV Seminar*. Basel: Birkhäuser Verlag; 1992.

GENETIC ALGORITHMS

MARTIN PELIKAN

Department of Mathematics and
Computer Science, University of
Missouri in St. Louis, St. Louis,
Missouri

INTRODUCTION

Genetic algorithms (GAs) [1,2] are stochastic optimization methods inspired by natural evolution and genetics. GAs work with a population or multiset of candidate solutions. Maintaining a *population of candidate solutions*—as opposed to a *single solution*—has several advantages. For example, using a population allows simultaneous exploration of multiple basins of attraction, it allows statistical decision making based on the entire sample of promising solutions even when the evaluation procedure is affected by external noise, and it enables the use of learning techniques for identifying problem regularities. The initial population is typically generated at random using the uniform distribution over all admissible solutions. The population is then updated for a number of iterations using the operators of selection, variation, and replacement. The selection operator ensures that the search for the optimum focuses on candidate solutions of high quality. Variation operators exploit information from the best solutions found so far to generate new high quality solutions. Finally, the replacement operator provides a method for updating the original population of solutions using the new solutions created by selection and variation.

Over the last few decades, GAs have been successfully applied to many problems of business, engineering, and science. Because of their operational simplicity and wide applicability, genetic algorithms are now playing an increasingly important role in computational optimization and operations research.

The purpose of this article is to provide an introduction to GAs. The article is organized as follows. The section titled “Genetic Algorithm Procedure” outlines the basic GA procedure. The section titled “Genetic Algorithm Operators” details popular GA operators and discusses how GAs can be applied to problems defined over various representations. The section titled “Dealing with Multiple Objectives, Constraints, and Noise” focuses on the use of GAs for solving problems with multiple objectives, noise, and constraints. The section titled “Efficiency Enhancement” outlines several approaches to enhancing efficiency of GAs. The section titled “Guidelines for Applying Genetic Algorithms to a New Problem” provides basic guidelines for applying GAs to a new problem. Finally, the section titled “Starting Points for Obtaining Additional Information” outlines some of the main starting points for the reader interested in learning more about GAs.

GENETIC ALGORITHM PROCEDURE

An optimization problem may be defined by specifying (i) a representation of potential solutions to the problem and (ii) a measure for evaluating quality of each candidate solution. The goal is to find a solution or a set of solutions that perform best with respect to the specified measure. For example, in the traveling salesman problem, potential solutions are represented by permutations of cities to visit and the measure of solution quality gives preference to shorter tours.

GAs pose no hard restrictions on the representation of candidate solutions or the measure for evaluating solution quality, although it is important that the underlying representation is consistent with the chosen operators. Representations can vary from binary strings to vectors of real numbers, to permutations, to production rules, to schedules, and to program codes. Performance measures can be based on a computer procedure, a simulation, an interaction with the human, or a

combination of the above. Solution quality can also be measured using a partial ordering operator on the space of candidate solutions, which only compares relative quality of two solutions without expressing solution quality numerically. In addition, evaluation can be probabilistic and it can be affected by external noise. Nonetheless, for the sake of simplicity, in this section it is assumed that candidate solutions are represented by binary strings of fixed length and that the performance of each candidate solution is represented by a real number called *fitness*. The task is to find a binary string or a set of binary strings with the highest fitness.

Components of a Genetic Algorithm

The basic GA procedure consists of the following key ingredients:

Initialization. GAs usually generate the initial population of candidate solutions randomly according to a uniform distribution over all admissible solutions. However, the initial population can sometimes be biased using prior problem-specific knowledge or other optimization procedures [3–6].

Selection. Each GA iteration starts by selecting a set of promising solutions from the current population based on the quality of each solution. A number of different selection techniques can be used [7,8], but the basic idea of all these methods is the same—make more copies of solutions that perform better at the expense of solutions that perform worse. For example, *tournament selection* of order s selects one solution at a time by first choosing a random subset of s candidate solutions from the current population and then

selecting the best solution out of this subset [9]. Random tournaments are repeated until there are sufficiently many solutions in the selected population. The size of the tournaments determines selection pressure—the larger the tournaments, the higher the pressure on the quality of each solution. The tournament size of $s = 2$ is a common setting. Typically, the size of the selected set of solutions is the same as the size of the original population.

Variation. Once the set of promising solutions has been selected, new candidate solutions are created by applying recombination (crossover) and mutation to the promising solutions. Crossover combines subsets of promising solutions by exchanging some of their parts. Mutation perturbs the recombined solutions slightly to explore their immediate neighborhood. Most of the commonly used crossover operators combine pairs of promising solutions. For example, *one-point crossover* randomly selects a single position in the two strings (crossing site) and exchanges the bits on all the subsequent positions (see Fig. 1). Typically, the selected population is first randomly divided into pairs of candidate solutions, and crossover is then applied to each of these pairs with a specified probability p_c . The probability p_c of crossover is usually in the range $[0.6, 0.95]$. One of the most common mutation operators for binary strings is *bit-flip mutation*, in which each bit is modified (flipped) with the same probability (see Fig. 1); the probability of changing a bit is typically small, changing only one or a few bits at a time.

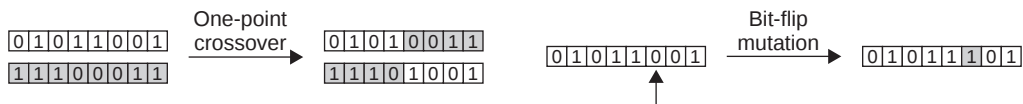


Figure 1. In one-point crossover, a site is randomly chosen and the strings exchange all bits after this site. In bit-flip mutation, each bit is modified with the same probability. The probability is typically small and only one or a few bits are changed.

Replacement. After applying crossover and mutation to the set of promising solutions, the population of new candidate solutions replaces the original one or its part, and the next iteration is executed (starting with selection) unless termination criteria are met. For example, the run can be terminated when the population converges to a singleton, the population contains a good enough solution, or an upper bound on the number of iterations has been reached.

The pseudocode of the basic genetic algorithm is as follows:

```
Genetic algorithm {
  t = 0;
  generate initial population P(0);
  while (not done) {
    select population of promising solutions
      S(t) from P(t);
    apply variation operators on S(t) to
      create O(t);
    create P(t+1) by combining O(t) and P(t);
    t = t+1;
  }
}
```

Since genetic algorithms are inspired by biology, common GA terminology is strongly influenced by that in biology; see Table 1 for an overview of GA terminology.

Table 1. Common GA Terminology

Term	Alternative Name(s)
Candidate solution	Individual, chromosome, string
Decision variable	Variable, locus, string position
Value of decision variable	Bit, allele
Measure of solution quality	Fitness function, objective function
Numeric representation of solution quality	Fitness, fitness value
Population of promising solutions, $S(t)$	Parent population, parents, selected solutions
Population of new solutions, $O(t)$	Offspring population, offspring, children
Iteration	Generation

Simulation of a Genetic Algorithm by Hand

To illustrate the basic GA procedure, consider solving the onemax problem, which assumes that candidate solutions are represented by binary strings of n bits, and the fitness is computed as the sum of bits in the input binary string:

$$\text{onemax}(X_1, X_2, \dots, X_n) = \sum_{i=1}^n X_i,$$

where $(X_1, X_2, \dots, X_n)$ denotes the input binary string. onemax is a simple unimodal function with the optimum in the string $(1, 1, \dots, 1)$.

The simulation will use binary tournament selection (i.e., $s = 2$), one-point crossover, bit-flip mutation and full replacement (the new population of solutions fully replaces the original population). The remaining parameters of the simulation by hand will be set as follows:

Parameter	Description	Value
n	String length	8
N	Population size	10
p_c	Crossover probability	0.6
p_m	Mutation probability	1/8

The first step consists of generating $N = 10$ binary strings of length $n = 8$. The resulting initial population is shown below:

	Candidate Solution	Fitness
1	10111001	5
2	00010010	2
3	01101001	4
4	00001001	2
5	00100110	3
6	00001010	2
7	11011000	4
8	11111000	5
9	11111000	5
10	00011001	3

4 GENETIC ALGORITHMS

Selection of the parent population starts by random selection of 10 pairs of candidate solutions from the original population. The winner of each of these tournaments is selected based on the fitness function (for ties the winner is chosen arbitrarily):

Tournament	Winner	Fitness
1 11111000 & 00011001	11111000	5
2 00010010 & 00011001	00011001	3
3 11011000 & 10111001	10111001	5
4 01101001 & 00001010	01101001	4
5 00100110 & 11111000	11111000	5
6 00010010 & 11111000	11111000	5
7 01101001 & 10111001	10111001	5
8 00001001 & 00010010	00010010	2
9 00100110 & 11011000	11011000	4
10 10111001 & 00011001	10111001	5

Since each parent solution was chosen independently, it is not necessary to pair solutions for crossover randomly, but consequent pairs of selected strings can be considered for crossover. There are five pairs of parents and the probability of crossover is 0.6; therefore, about three pairs of parents are expected to be recombined using crossover. Mutation is then applied to all the resulting strings. Mutation is expected to change 1 bit on average since $p_m = 1/8$ and $n = 8$. The results after applying crossover follow:

Parents	Crossing Site	After Crossover
1 11111000	3	11011001
2 00011001		00111000
3 10111001	—	10111001
4 01101001		01101001
5 11111000	2	11111000
6 11111000		11111000
7 10111001	—	10111001
8 00010010		00010010
9 11011000	6	11011001
10 10111001		10111000

The results of applying mutation to the resulting population follow:

	After Crossover	After Mutation	Fitness
1	11011001	11111001	6
2	00111000	00111000	3
3	10111001	11111001	6
4	01101001	00111101	5
5	11111000	11110000	4
6	11111000	11111000	5
7	10111001	10011011	5
8	00010010	10010010	3
9	11011001	11010001	4
10	10111000	10111000	4

The population after mutation will then replace the original population of solutions, which serves as the starting point for the next GA iteration.

One of the ways to analyze the results of the simulation is to look at the change in the average fitness of the population. The average fitness of the original population is 3.5, whereas the average fitness of the new population is 4.5. Therefore, one round of selection, crossover, and mutation increased the average quality of solutions in the population. At the same time, most of the new solutions are different from those in the original population, although they share many similarities. We were thus able to create a population of new solutions, which are of better quality than those that we started with. Simulating the following few iterations of the GA should further increase solution quality, until the global optimum would be found.

Simulation of a Genetic Algorithm on a Larger Problem

In this section, the simulation of the GA by hand is extended to deal with a larger Onemax problem of $n = 100$ bits. Due to the increased problem size, in order to ensure that the global optimum is found with a high probability, we also increase the population size. Specifically, we set the population size as $N = 100$, the probability of crossover as

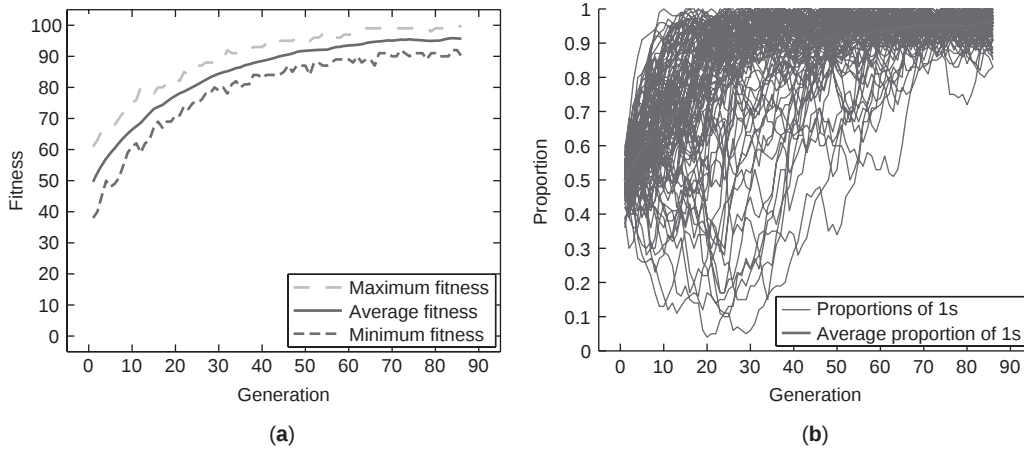


Figure 2. GA simulation on onemax of $n = 100$ bits: (a) fitness values and (b) proportions of 1s in different positions.

$p_c = 0.6$, and the probability of mutation as $p_m = 1/100$. The results of the simulation are presented in Fig. 2.

The results presented in Fig. 2a show that the average population fitness increases over time and so does the maximum and minimum fitness in the population, although the minimum and maximum fitness often decrease temporarily. Furthermore, Fig. 2b shows that the proportions of 1s in most string positions increase over time, although some of these proportions decrease temporarily and fluctuate substantially. Nonetheless, the average proportion of 1s in different positions steadily increases and eventually the probability of a 1 in any position becomes 80% or more. Continuing the simulation would lead to a further increase in the proportions of 1s until almost every solution in the population would be globally optimal.

Of course, onemax is a relatively simple problem and it is not a surprise that GAs can solve this problem without any difficulty. Onemax is used to only illustrate the basic GA procedure on a problem that is easy to understand. For onemax, theoretical models [10–12] exist which estimate that, assuming that the expected number of bits that do not converge to the correct values is upper bounded by an arbitrary constant, the number of evaluations until convergence is expected to grow at most as $O(n \log n)$ where n is the number of bits in the problem.

GENETIC ALGORITHM OPERATORS

Previous section presented the basic GA procedure and some of the simplest GA operators. This section presents several other popular selection, replacement, and variation operators. In addition, the section discusses how GAs can be used to solve problems where candidate solutions are not represented by fixed-length binary strings and how linkage learning can be incorporated into GAs to improve their performance and scalability.

Selection

The main task of the selection operator is to select a population of promising candidate solutions from the current population of candidates, ensuring that solutions of higher quality have a higher probability of being selected than others. There are two main types of selection methods: (i) fitness proportionate selection, and (ii) ordinal selection.

Fitness Proportionate Selection. In the *fitness proportionate selection* methods, each selected solution is drawn from the same probability distribution and the probability of selecting the solution is in some way proportional to its fitness value. Most commonly, the fitness is assumed to be positive and the probability of selecting an individual X with

fitness $f(X)$ is linearly proportional to its fitness. Therefore, the probability $p_{\text{select}}(X)$ of selecting X from population $P(t)$ is

$$p_{\text{select}}(X) = \frac{f(X)}{\sum_{Y \in P(t)} f(Y)}.$$

Ordinal Selection. In *ordinal selection* methods, the probability of selecting a particular member of the population does not directly depend on the actual value of its fitness but it depends on the relative quality of this solution compared to other members of the current population. *Tournament selection* discussed in the previous section is one of the most popular ordinal selection methods [9]. Another popular ordinal selection method is *truncation selection*, which selects a given percentage τ of the best solutions from the population [11]. For example, for $\tau = 0.5$, the best half of the population is selected. If the selected population should have the same size as the original, multiple copies of each candidate solution in the selected portion of the population can be created.

Ordinal selection methods are invariant to linear transformations of fitness and they pose fewer restrictions on the fitness function than fitness proportionate selection methods do. In addition, ordinal selection methods exert a sustained pressure toward solutions of higher quality and the strength of this pressure is often easier to control. These are the main reasons why ordinal selection methods are generally more popular than fitness proportionate selection methods.

The selection operator can also be used to promote population diversity by adjusting the probability of selecting each candidate solution based on how this solution differs from the solutions that have already been selected. To promote diversity, preference should be given to solutions that differ from the solutions selected so far. The methods that aim at maintaining diversity of GA populations are often referred to as *nicing methods* [13–16].

One of the most popular approaches to niching via selection is *fitness sharing*, in

which fitness values are updated so that the individuals similar to the previously selected individuals are penalized [13,15–18].

Replacement

The main task of the replacement operator is to incorporate the population of new candidate solutions into the original population of candidates. There are two main approaches to replacement: (i) delete-all and (ii) steady state.

Delete-All or Full Replacement. With delete-all or full replacement, the entire original population is replaced with the new solutions as was done in the simulations presented in the previous section.

Steady-State Replacement. It may often be beneficial to allow candidate solutions in the original population to become a part of the new population, providing overlap between the populations in subsequent iterations of the GA. One way to accomplish this is to combine the original population of candidate solutions with the newly generated ones, and then select the best solutions out of the combined population to form the population for the next iteration. Another approach is to use selection and variation to create fewer individuals than is the size of the original population, and then replace only some of the original solutions with the new ones. The solutions to replace may be selected at random or they may be chosen to be the worst solutions in the original population. GAs where the best solutions found so far are guaranteed to survive to the next iteration are often referred to as *elitist* GAs. The proportion of the population that is going to be replaced by new individuals is typically called *generation gap*. GAs with small generation gap in which a substantial portion of the original population survives to the next generation are often called *steady-state* GAs. In practice, steady-state replacement schemes are more popular than delete-all replacement.

Replacement can also be used to promote diversity in the population and ensure that the populations of candidate solutions represent a diverse sample of candidate solutions. For example, in a crowding factor model (or, more simply, crowding) [19], for each new individual, a subset of CF candidate solutions is first selected from the original population. The new solution then replaces the most similar solution from this subset. The parameter CF is called *crowding factor*. The crowding factor provides a way to control crowding; the higher the value of CF , the stronger the pressure on diversity maintenance. Another replacement strategy that promotes diversity is called *restricted tournament selection* [20]. Restricted tournament selection proceeds similar to crowding; however, the most similar solution from the selected subset is replaced only if the new solution is better (with respect to the fitness function).

Variation

One-point crossover and bit-flip mutation are among the simplest and most popular variation operators. However, many other variation operators are common in practice. This subsection reviews a few such operators applicable to candidate solutions represented by binary strings with the focus on crossover operators applicable to two parents. The topic of variation is revisited in the sections titled “Dealing with Other Representations” and “Adaptive Operators and Linkage Learning,” which discuss linkage learning and some of the more advanced variation operators.

One-Point, Two-Point and k -Point Crossover. In *one-point crossover*, a crossing site is chosen at random and the two parents exchange all bits after the selected position. In *two-point crossover*, two crossing sites are randomly selected and the two parents exchange all bits between these two sites. The two-point crossover can be further generalized to k -point crossover, in which k crossing sites are selected. Two-point crossover can be approximated by two applications of one-point crossover, whereas k -point crossover can be approximated by k -

applications of one-point crossover. Most frequently, one-point or two-point crossover operators are used.

Uniform Crossover. *Uniform crossover* exchanges the bits in each position of the two parents with probability 50%. About half of the bits are expected to be exchanged on average.

The main difference between k -point crossover and uniform crossover lies in the way these operators select the positions in which the bits are going to be exchanged. In one-point crossover or, more generally, in k -point crossover there is an implicit linkage between positions located close to each other in solution strings. This is because positions that are close to each other are likely to be treated together (either exchanged or not). On the other hand, uniform crossover treats every string position independently, and it is therefore independent of the ordering of decision variables in the solution strings. The topic of linkage in crossover will be revisited in section titled “Adaptive Operators and Linkage Learning.”

It is of note that in GAs crossover is often considered the primary variation operator [2,21], whereas in evolution strategies [22,23] mutation is considered the primary variation operator. Consequently, the probability of crossover in GAs is typically quite high (60% or more) and the strength of mutation is set so that mutation causes only small changes in the solution. For an introduction to evolution strategies, see Bäck *et al.* [24] and Beyer and Schwefel [25].

Dealing with Other Representations

In practice, candidate solutions are often not represented as binary strings of fixed length; for example, candidate solutions may be represented by permutations, vectors of real values, or variable length strings. There are two main approaches to applying GAs to such problems:

Map Binary Strings to the Original Representation. Introduce a mapping from binary strings of fixed length to candidate solutions in the original representation. Then, a standard GA for

binary strings can be applied and the mapping function can be used to map each obtained candidate solution to the original representation so that the solution can be evaluated. In some cases, it may be difficult to design a one-to-one mapping and, as a result, not all admissible solutions will be considered. Another potential complication is that it may sometimes be far from straightforward to design an appropriate mapping between the set of binary strings and the actual problem representation. On the other hand, the mapping may introduce additional opportunities for discovery and exploitation of problem regularities, which may not be apparent in the original representation.

Modify Variation Operators. Modify variation operators to make them work with the other representation. Extensions of standard variation operators to some representations are straightforward. For example, if candidate solutions are represented by strings over a finite (nonbinary) alphabet, all aforementioned crossover operators can be applied without any modification and the bit-flip mutation can be adjusted to change the value being modified to any other admissible value. Nonetheless, if the used representation is qualitatively different, variation operators may have to be modified.

The remainder of this section discusses the two approaches to dealing with other representations using GAs through examples for two of the most popular representations: (i) real-valued vectors and (ii) permutations.

Mapping between Other Representations and Binary Strings. Solutions to many real-world problems can be encoded as vectors of real-valued variables. There are numerous ways for mapping real-valued vectors into binary strings and the other way around; however, since each real-valued variable can obtain infinitely many values, the mapping cannot be one-to-one and some accuracy will have to be lost when representing real-valued vectors using binary strings of fixed

length. For simplicity, let us assume that each real-valued variable can obtain values from $[a, b]$ where $a < b$, and the task is thus to represent vectors from $[a, b]^n$ where n is the number of real-valued variables.

One way to approach this problem is to use a fixed number k of bits to represent each real-valued variable in the vector [2]. Binary strings that represent n real-valued variables would therefore consist of $n \times k$ bits. The k bits for each real-valued variable X_i are interpreted as binary representation of an integer k_i from 0 to $2^k - 1$. The value of X_i for a particular combination of the k bits representing X_i is given by

$$X_i = a + (b - a) \frac{k_i}{2^k - 1}.$$

Before evaluating each binary string, the values of the real-valued variables are obtained, and the resulting real-valued vector is evaluated using the fitness function. One may also consider mapping real-valued vectors to binary strings, where the value of each variable X_i would determine the closest k_i , and, consequently, the most accurate k -bit representation of X_i .

Clearly, the more bits are used to represent each real value, the more accurate the representation becomes. An appropriate value of k depends on the desired accuracy of the solution and the ruggedness of the fitness landscape, but in many cases, one can use between 5 and 30 bits per real variable. The final solutions can be further tuned with a local search method, using for example, a variant of the steepest descent/ascent method.

Intuitively, after several iterations of a GA, most variables will have their values concentrated in one or several smaller areas of the interval $[a, b]$. This is the reason why using a mapping function that assumes that the binary strings can represent only equidistant values becomes inefficient, as most of these values will lead to solutions of low quality and will therefore not be used. Hence one may choose to use an adaptive discretization scheme, in which more points are allocated to subintervals that appear in the population most frequently [26–29].

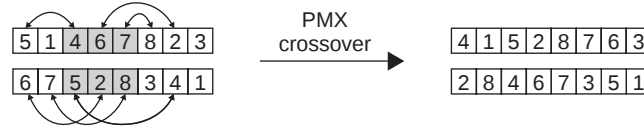


Figure 3. Partially matched crossover starts by selecting two positions at random. Then, the elements between the two positions are matched against the other parent solution and the corresponding pairs of elements define pairs of elements to exchange in each parent solution. In this example, the elements between the two selected positions are (4, 6, 7) and (5, 2, 8); therefore, in both solutions, 4 will exchange with 5, 6 will exchange with 2, and 7 will exchange with 8.

Binary strings can also be used to encode candidate solutions from other representations, such as permutations, tree graphs, or combinations of several different representations. Although there are certain benefits of mapping a particular representation to binary strings, such as the potential for finding and exploiting various problem regularities, except for the domain of real-valued vectors, practitioners often prefer to use the original representation of the problem to represent candidate solutions and use modified variation operators capable of directly manipulating solutions in the representation used. Some examples for the domain of permutations and real-valued vectors are presented next.

Variation Operators for Permutations and Real-Valued Vectors. Many important optimization problems deal with candidate solutions represented by permutations; examples of such problems include the quadratic assignment problem and the traveling salesman problem. Although the variation operators presented thus far can be applied to permutations, doing this may often result in candidate solutions that are not valid permutations. One way to deal with this problem is to define a *repair operator*, which corrects candidate solutions after variation so that they become valid. Another approach is to modify the variation operators themselves so that only valid permutations are obtained. Two crossover operators and several mutation operators will be presented next, all of which directly manipulate permutations without making them invalid.

Partially Matched Crossover (PMX). PMX [30] combines two parent permutations by first selecting two permutation positions at random. Then, for each permutation element between the two positions, the element is exchanged with the corresponding element in the other parent solution. All exchanges are done in both parent solutions. An illustrative example of the partially matched crossover is shown in Fig. 3.

Uniform Order-Based Crossover. Uniform order-based crossover [31] starts by generating a mask to determine which elements will be copied from the original parent and which will be taken from the other parent permutation. The mask may be defined, for example, by generating a 1 in each position with probability 50% and generating a 0 otherwise. The two parents retain the elements in the positions where the mask contains a 1. The remaining elements are filled in from the other parent in the order they appear in the other parent. An illustrative example is shown in Fig. 4.

Swaps, Inversion, and Scramble Mutation.

Swap mutation starts by selecting two permutation elements at random and it then exchanges these two elements. *Inversion mutation* starts by selecting two locations in the permutation and it then inverts the permutation elements between these two locations. Finally, *scramble mutation* selects two locations at random and it then randomly

Figure 4. Uniform order-based crossover starts by selecting random positions for which the permutation elements do not change. The remaining positions are filled with the missing elements copied in the order they appear in the other parent.

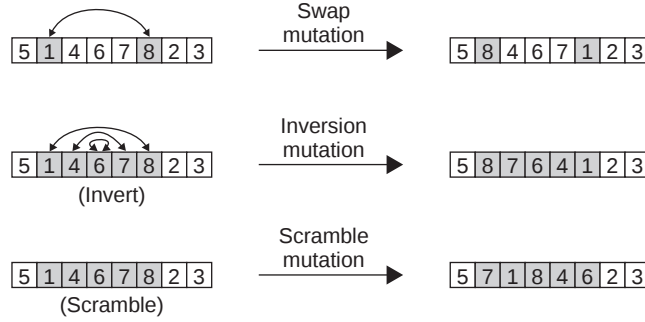
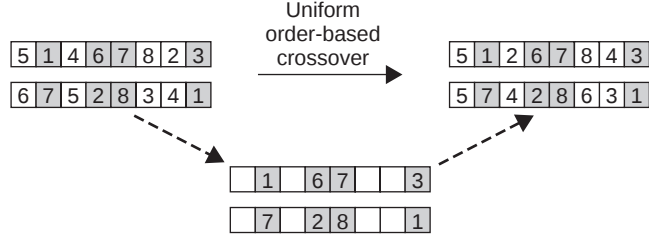


Figure 5. Mutation operators for permutations. All three mutation operators shown here start by randomly selecting two locations in the permutation (in this example, the second and sixth positions are selected). Swap mutation exchanges the elements in the selected locations. Inversion mutation inverts the order of the elements between the two locations. Scramble mutation randomly reorders all the elements between the two locations.

reorders all elements between the two selected positions. All three mutation operators are illustrated in Fig. 5.

Numerous variation operators have been designed for real-valued vectors. We mention a few simple ones here.

Arithmetic Crossover. Arithmetic crossover starts by selecting a location i in the two parents at random and generating a random parameter $\alpha \in [0, 1]$ according to the uniform distribution. Denoting the variable at location i in the first parent by X_i and the corresponding variable in the second parent by Y_i , the values of X_i and Y_i are mapped to X'_i and Y'_i , respectively, as follows:

$$\begin{aligned} X'_i &= (1 - \alpha)X_i + \alpha Y_i, \\ Y'_i &= (1 - \alpha)Y_i + \alpha X_i. \end{aligned}$$

Arithmetic crossover can also be applied to several or even all variables.

Simulated Binary Crossover (SBX). In simulated binary crossover [32], a location for the crossover site is first chosen, similarly as in arithmetic crossover. Let us denote the variables of the two parents for the chosen crossover site by X_i and Y_i , and assume (without loss of generality) that $X_i < Y_i$. The new values of X_i and Y_i are computed as

$$\begin{aligned} X'_i &= \frac{X_i + Y_i}{2} - \frac{\beta}{2} (Y_i - X_i), \\ Y'_i &= \frac{X_i + Y_i}{2} + \frac{\beta}{2} (Y_i - X_i), \end{aligned}$$

where $\beta > 0$ is a randomly generated parameter (spread factor). One way to generate β is to use the following probability density function [32]:

$$c(\beta) = \begin{cases} 0.5(n+1)\beta^n & \text{for } \beta \leq 1 \\ 0.5(n+1)\beta^{-n-2} & \text{for } \beta > 1. \end{cases}$$

This procedure can be applied, similarly as in arithmetic crossover, to several or even all variables in the two parents.

Gaussian Mutation. In Gaussian mutation, each variable is modified by adding a random number according to a Gaussian distribution with zero mean. The variance of the Gaussian distribution defines the strength of mutation, and is typically significantly smaller than the range of values of a variable.

Adaptive Operators and Linkage Learning

The combination of selection and variation is the main driving force in the search for the optimum in a GA. Selection selects the most promising candidate solutions, whereas variation combines parts of the selected candidate solutions and explores their immediate neighborhood. To ensure that the combination of selection and variation enables robust, efficient, and scalable search for the global optimum, it is necessary that variation exploits regularities in the problem landscape and that it combines and modifies the best candidate solutions in a meaningful way. Nonetheless, all variation operators discussed in the previous section process candidate solutions in the same way regardless of the specifics of the problem being solved. Consequently, for some classes of difficult problems, practitioners may often have to either modify these variation operators to better suit the problem or modify the representation of candidate solutions to make the existing variation operators more effective. One way to alleviate this problem is to use *adaptive variation operators* that are capable of adapting to the problem being solved. In this section, we discuss the use of adaptive operators with the main focus on recombination operators capable of *linkage learning* [21,33–35].

The previous section pointed out that in one-point crossover implicit linkage

is introduced on variables encoded in consequent locations of solution strings. Consequently, combinations of values in consequent positions of solution strings are often preserved by variation, whereas combinations of values in positions located far from each other may be often modified. This has a positive effect on GA performance if the decision variables located close to each other are strongly correlated in the problem being solved. Nonetheless, for many problems variable correlations may appear both between variables that are located close to each other and those that are located far away from each other. Furthermore, a practitioner may not be aware of correlations that are important for solving the problem in a scalable manner. This is why practitioners are often forced to define new, problem-specific variation operators or to modify the representation of candidate solutions so that the correlations between problem variables correspond more closely to the implicit linkage introduced by the variation operators used, such as the one-point crossover.

To alleviate this problem, adaptive variation operators can be designed that are capable of discovering the most important interactions between problem variables and modifying the recombination operator accordingly. The process of discovering important interactions between problem variables in order to improve effectiveness of variation is often referred to as *linkage learning* [21,33–37]. The term *linkage* is also common in biological systems, where linkage refers to the level of association in the inheritance of two or more nonallelic genes that is higher than that expected if these genes were independent [38].

There are two main approaches to linkage learning [35]. In the *unimetric approach*, linkage learning is based solely on the fitness values obtained during the GA run. In the *multimetric approach*, linkage learning is based on an additional metric besides the fitness function itself. While unimetric approaches appear more biologically plausible, multimetric approaches have been somewhat more successful in practice [35].

GAs capable of linkage learning can be split into three main categories: (i)

perturbation techniques, (ii) linkage adaptation techniques, and (iii) estimation of distribution algorithms. Linkage adaptation techniques are the only category based on the unimetric approach.

Perturbation Techniques. In perturbation-based linkage learning GAs, linkages are identified via perturbing candidate solutions and analyzing the effects of these perturbations on the fitness values. The main representatives of perturbation-based linkage learning GAs are the messy GA [36,39], the fast messy GA [37,40], the gene expression messy GA [41], the linkage identification by nonmonotonicity detection GA [42,43], and the dependency-structure-matrix-based GA [44,45].

Linkage Adaptation Techniques. In linkage adaptation techniques, linkages are evolved along with candidate solutions. Selection thus remains the main driving force of linkage learning. Linkage adaptation is used, for example, in the linkage learning GA [46].

Estimation of Distribution Algorithms (EDAs). In EDAs [47–50]—also called *probabilistic model-building genetic algorithms* and *iterated density estimation algorithms*—standard variation operators of genetic algorithms are replaced by building and sampling probabilistic models of selected candidate solutions. Since many classes of probabilistic models enable the discovery and use of dependencies between problem variables, the use of probability distributions provides a powerful tool for multimetric linkage learning. Some of the linkage learning GAs based on the EDA framework are the Bayesian optimization algorithm (BOA) [51,52], the estimation of Bayesian networks algorithm (EBNA) [53], the learning factorized distribution algorithm (LFDA) [54], and the extended compact genetic algorithm (ecGA) [55]. For more information on EDAs, please see other literature [47–50].

DEALING WITH MULTIPLE OBJECTIVES, CONSTRAINTS, AND NOISE

Many challenging real-world problems contain multiple objectives or constraints, or the evaluation of solution quality is affected by noise. GAs provide a fairly powerful solution even in these scenarios [56,57]. In fact, robustness and the ability to deal with multimodal, noisy, constrained, and multiobjective optimization problems are among the key advantages of GAs over many other optimization techniques. This section reviews some of the most important concepts for dealing with problems with multiple objectives, noise, and constraints. The main focus is on providing the interested reader with pointers.

Multiobjective Genetic Algorithms

In multiobjective optimization, the task is to maximize or minimize several objectives. For example, in engine design, one may want to maximize performance and minimize fuel consumption. The main difficulty in dealing with multiobjective problems is that the objectives are typically conflicting because increasing solution quality with respect to one objective leads to decreased quality with respect to another objective. There are two basic approaches to solving multiobjective optimization problems using GAs and other evolutionary algorithms [56,57]:

Transform Multiple Objectives into a Single Objective. One way to deal with a multiobjective problem is to transform all objectives into a single objective using a weighted sum, utility theory, or another approach. The main difficulty lies in the appropriate selection of weights or utility functions because there is typically only little problem-specific knowledge to guide these design choices.

Find the Entire Pareto-Optimal Set (Trade-Off between Objectives). Another approach is to find the entire Pareto-optimal set (front) of solutions or a representative subset of the entire Pareto-optimal set [58,59]. A set of candidate solutions is called

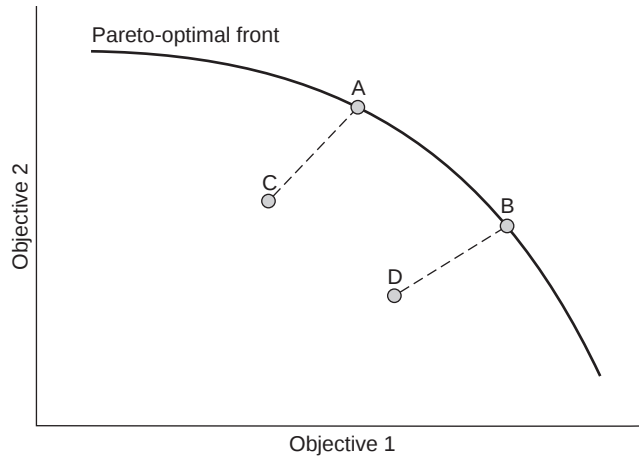


Figure 6. In this example, the task is to maximize two objectives and each axis in the graph corresponds to one of these objectives. The two solutions *A* and *B* are members of the Pareto-optimal front (set), and they thus cannot be dominated by any other admissible solution in both objectives. Two solutions *C* and *D* are not members of the Pareto-optimal set, because they are dominated by other solutions. For example, *C* is dominated by *A* and *D* is dominated by *B*.

Pareto optimal if the set contains only solutions that are not dominated (outperformed) by any other admissible solution in all objectives. See Fig. 6 for an illustration of the Pareto-optimal set. The set of Pareto-optimal solutions defines the trade-off between the objectives and can serve as the main input for finding the solution of desired properties. The main difficulty in finding the entire Pareto-optimal set is that the size of the set typically grows very fast with the number of objectives and so does the difficulty of finding enough representatives of this set. Two most popular variants of a GA for finding the Pareto-optimal set of solutions are (i) the nondominated sorting GA-II (NSGA-II) [60], and (ii) the improved strength Pareto evolutionary algorithm (SPEA2) [61]. In both algorithms, the modifications take place in selection and replacement operators; other operators remain the same as in standard, single-objective GAs.

For an excellent introduction to solving multiobjective optimization problems using GAs, see Deb [56] and Coello *et al.* [57].

Noisy Evaluation

In many real-world problems, evaluation of candidate solutions is affected by noise. The sources of noise include physical measurements, approximate models, stochastic simulation models, incomplete information, and human-computer interaction [62].

Genetic and other evolutionary algorithms are known to be capable of dealing with problems that include noise particularly well [62,63]. This is due to several reasons. Most importantly, one of the consequences of using a population of candidate solutions as opposed to a single solution is that the use of a population alleviates the effects of noise by averaging them out. In addition, GAs do not require derivatives or gradients, which are difficult to approximate in the presence of noise.

One way to alleviate the effects of noise in fitness evaluation is to increase the population size [10,21,63]. In addition to the increased population-sizing requirements, in the presence of noise, the overall number of iterations until convergence can be expected to increase [21,63].

Besides increasing the population size and the number of iterations, one may deal with noise by using *fitness averaging*. The main idea of fitness averaging is to evaluate each candidate solution several times, and use the average fitness of these evaluations as the fitness of this solution. Given the available time for the GA and the specifics of the problem, one can compute the optimal number of evaluations for each candidate solution [64].

Constraints

In many real-world optimization problems, the set of admissible solutions is defined

using a set of constraints, which may be defined, for example, by providing lower bounds for linear combinations of variables. The task is to maximize or minimize a given objective function subject to the given constraints.

There are three main types of approaches to using GAs for solving optimization problems with linear or nonlinear constraints:

Preserve the Feasibility of Candidate Solutions. Specialized variation operators may be defined that are always guaranteed to preserve feasibility of candidate solutions. If the initial population is generated so that it contains only feasible solutions and if the specialized variation operators are used, all solutions obtained during the GA run will be guaranteed to be feasible with respect to the given constraints. Assuming that the search space is convex and that the feasible solutions are represented by real-valued vectors, one may use, for example, the genetic algorithm for numerical optimization of constrained problems (GENOCOP) [65]. Another way to ensure feasibility of candidate solutions is to reject infeasible solutions [66] or to control the proportion of infeasible solutions in some way [67]. Finally, one may introduce *repair* operators, which modify each infeasible solution to make it feasible [68].

Penalize the Infeasible Solutions. If the population is allowed to contain solutions that violate some of the constraints, infeasible solutions may be penalized so that the search is biased toward feasible regions of the search space. This can be done, for example, by introducing a static or dynamic penalty function that reduces fitness of infeasible solutions [69–71] or by modifying the selection operator to consider constraint violation in addition to the fitness value [72].

Hybrid Approaches. The above two classes of approaches to solving constrained optimization problems can be combined in various ways, creating hybrid

approaches to dealing with constrained problems.

For an overview of the work on solving optimization problems with constraints using genetic and other evolutionary algorithms, the interested reader should consult Michalewicz and Janikow [65], Coello [73], and Michalewicz [74].

EFFICIENCY ENHANCEMENT

Practical applications of GAs often necessitate the use of additional efficiency enhancement techniques, such as parallelization and hybridization. There are four main categories of efficiency enhancement techniques: (i) parallelization, (ii) hybridization, (iii) evaluation relaxation, and (iv) time continuation. This section reviews these main types of efficiency enhancement techniques.

Parallelization

One way to speed up GAs is to distribute the computation to multiple processors [75–79]. There are three main types of parallel GAs [78]: (i) single-population master-slave GAs, (ii) single-population fine-grained GAs, and (iii) multiple-population coarse-grained GAs (see Fig. 7 for an illustration of these types of parallel GAs).

Single-Population Master-Slave GAs. In the master-slave architecture, the computation is controlled by a master processor, which stores the population and executes GA operators. Slave processors are typically used only to compute the fitness values of candidate solutions. Other GA operators may also be distributed, but due to the relatively high cost of communication between different processors, it is important to ensure that the savings in computational time are more significant than the time required for communication between the master and slave processors.

Single-Population Fine-Grained GAs. Fine-grained GAs consist of one spatially structured population, which is

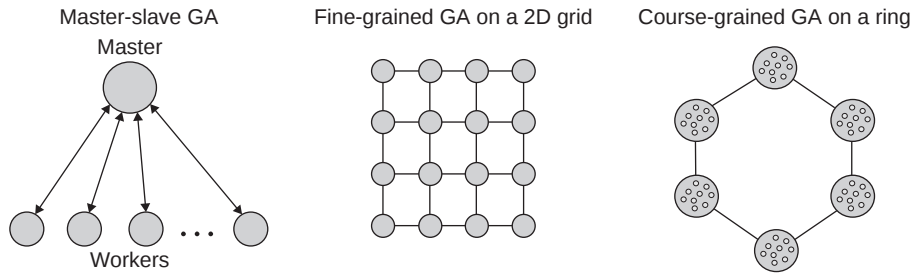


Figure 7. Illustration of different types of parallel GAs.

distributed among a large number of processing units. Fine-grained GAs are best suited for massively parallel computers. Ideally, each processing unit considers only one candidate solution. Selection and variation are restricted based on the spatial distribution and communication structure of the parallel computer. For example, the individuals may be distributed on a 2D grid and communication may be restricted to consider only four neighbors of each candidate solution.

Multiple-Population Coarse-Grained GAs.

The multiple-population coarse-grained GAs split the population into several subpopulations, which develop in isolation. Infrequent migration of individuals from one subpopulation to another is allowed. There are numerous parameters that define the migration patterns, such as when and how many individuals are exchanged between subpopulations, and which subpopulations are allowed to exchange individuals.

Ideally, if the GA is executed in parallel using p processors, the speedup compared to the standard GA implementation on a single processor should be approximately p or even more; otherwise, the computational resources are not used efficiently. For introduction and theoretical guidelines for designing efficient parallel GAs, see Goldberg [76] and Cantú-Paz [79,80].

Hybridization

The central idea in hybridization is to combine two or more optimization methods

into a single hybrid technique. Typically, global–local hybrids that combine two optimization procedures are used; one of these methods is capable of effective global exploration of the search space, whereas the other is capable of quickly reaching a local optimum. In hybrid GAs, a GA would be typically used as a global optimizer and some form of hill climbing would be used as a local searcher. The local searcher may be run on every solution in the population or only on a portion of the population (e.g., on the best solution obtained). Hybrid GAs are sometimes referred to as *memetic GAs*. A survey to hybrid GAs can be found in Sinha and Goldberg [81].

One of the key issues in the design of efficient hybrid procedures is the appropriate coordination of the global and local search. The two main possibilities are captured in the literature by the terms *Baldwinian* and *Lamarckian* hybrid [82]:

Baldwinian Hybrid. In a Baldwinian hybrid, the local search procedure uses the global procedure's value as a starting point, searches until reaching an optimum, and then passes back the function value found *without* backsubstituting the solution value found.

Lamarckian Hybrid. In a Lamarckian hybrid, the local search procedure again uses the global procedure's value as a starting point, searches until reaching an optimum, and then passes back the function value found *with* backsubstitution of the solution value found.

At first sight, Lamarckian hybrids may seem more intuitive, but in some cases

backsubstitution of solution values found by local search has been found to reduce solution variance to a point such that subsequent exploration is diminished. Various rules of thumb have suggested Baldwinian moves with a small percentage of Lamarckian moves as desirable in practice [68].

Finding an adequate way to split resources between the global and local search requires a systems-level understanding of the roles of the two procedures. Goldberg and Voessner [83] proposed such a theory and Sinha and Goldberg [84], Sinha [85], and Sinha *et al.* [86] explored the theory in the context of simple GAs. Simply stated, the theory suggests that the key role of the global searcher is to find one or a number of good regions and the key role of the local searcher is to find local optima within these regions.

Evaluation Relaxation

For many optimization problems, evaluation of the objective function is computationally expensive. The number of evaluations until a solution is found can be substantially decreased using hybridization and prior knowledge, and a number of studies exist that support this fact [21,34,54,87–89]. However, to deal with computationally expensive objective functions, one may also have to use past evaluations to estimate fitness values for some of the new solutions. There are two basic approaches to using past evaluations to estimate fitness values of new solutions.

Exogenous Models. Past evaluations can be used to build a fast, approximate model of the objective function, which can be subsequently used to evaluate some of the new candidate solutions instead of the original, computationally expensive objective function [89–92]. For example, a simple linear model can be adjusted to fit the fitness values in the population of parents and then sampled to estimate fitness values of some of the new solutions [90]. More advanced models can be developed based on probabilistic models discovered by EDAs [89,91] or neural networks [92].

Endogenous Models. The fitness of new solutions can also be estimated directly from parental fitness values [93] because new solutions have often fitness values similar to the parent solutions used to create these new solutions. For example, the fitness of a new solution can be estimated as an average fitness of the parents that were used to create this new solution [93].

Using an approximate model to evaluate some candidate solutions has typically a similar effect as if the fitness function was affected by additive external noise [90]. Consequently, the overall number of candidate solutions that must be considered until an accurate solution is obtained can be expected to increase. However, since many of the solutions are evaluated using a fast approximation of the computationally expensive fitness function, the overall computational time can be reduced substantially. In practice, even speedups of over 50 can be obtained if advanced exogenous models are used in combination with model building using EDAs [91].

Time Continuation

Different variation operators have different function and effects. Recombination combines bits and pieces between pairs of promising solutions whereas mutation modifies candidate solutions to explore their immediate neighborhood. The optimal division of labor between the two operators depends on the specifics of the problem and the time available to complete the task. Furthermore, depending on the time available and problem specifics, in some situations it may be preferable to search for solutions using a large population, whereas in other cases it may be preferable to search for solutions using a small population with multiple convergence epochs. In time continuation [21,94,95], the focus is on the optimal division of labor between different search operators and other parameters of a GA with the focus on either (i) maximizing solution quality within a limited computational budget or (ii) minimizing computational time for obtaining a solution of given quality.

Theoretical results for time continuation indicate that for additively separable problems with subproblems of equal salience, multiple convergence epochs with a rather small population are expected to be more efficient. However, in the presence of noise and overlap between subproblems, a single epoch with a large population is expected to be more efficient. For an overview of work on time continuation and theoretical results, please see Srivastava and Goldberg [95] and Srivastava [96].

GUIDELINES FOR APPLYING GENETIC ALGORITHMS TO A NEW PROBLEM

One of the main advantages of GAs over other optimization methods is that it is relatively straightforward to apply GAs to new optimization problems. This section provides guidelines that can serve as a starting point for novices who want to apply a GA to a new problem.

Parameters. There are numerous parameters that must be set in order to apply a GA to a new problem. The simple GA implementation presented in the section titled “Genetic Algorithm Procedure” contains three main parameters (besides the problem-dependent ones). The best values of these three parameters depend on the problem being solved, but there are some general rules of thumb that provide a fairly good starting point.

- *Population size, N .* The population size is probably the most important parameter to tune and using populations that are too small is a frequent mistake that GA researchers and practitioners make. Generally, the population size should increase with problem size (the number of bits or variables), and the larger the population size, the better the obtained solution should be. Unless the evaluation of the objective function is prohibitively slow, the population size can range from tens to thousands or even more. As a starting point

for initial runs, the population size may be set to the problem size, $N = n$. If using $N = n$ is infeasible, one may start with smaller values of N , such as $N = 50$. Changing the population size and rerunning the GA with larger and smaller populations (e.g., $N/2$ and $2N$) provides an indication of whether the current population size is adequate. If increasing the population size leads to better solutions, the population size should be increased. If increasing the population size leads to solutions of about the same quality, the current population size is likely to suffice. Of course, if the evaluation of the objective function is computationally intensive, large populations may become intractable and, to run a sufficient number of generations, one is often forced to use a relatively small population size.

- *Crossover probability, p_c .* The probability of crossover in GAs is typically quite large. For example, DeJong and Spears [97] suggest that $p_c = 0.6$ regardless of the number of bits n . Many researchers use an even larger probability of crossover, for example, $p_c = 0.9$ or $p_c = 1$.
- *Mutation probability, p_m .* The probability of mutation is typically set so that the expected number of modified bits is fixed regardless of n and that only a few bits are expected to change on average. For example, one may use $p_m = 1/n$. Some researchers use larger values of p_m , but in many cases the reasons for increasing the probability of mutation are that the remaining parameter settings are inadequate.

Providing the GA with adequate values of parameters is a long-standing topic in GA research, and for more information, there are numerous publications that the reader may consult Goldberg [21] and Lobo *et al.* [98] provide a good starting point for these efforts.

Termination. Typically, each GA run is terminated when the number of generations reaches a predefined threshold, when a solution of sufficient quality has been achieved, or when the GA run starts to stagnate without improving the quality of the population for a given number of iterations. One of the frequent mistakes of GA practitioners is to underestimate the population size and then overestimate the number of generations. This can lead to an unnecessary increase in computational requirements of the GA by executing a large number of GA iterations with no or only little improvement. The number of generations is typically only several times larger than the number of bits in the problem. One way to avoid setting a strict upper bound on the number of generations is to terminate the run when the average fitness of the current population does not improve much for a given number of iterations (e.g., 5 or 10 generations). Of course, the best way to terminate a GA run depends on the problem being solved and several trial runs should provide enough input to make an informed decision.

Representation. Representation of candidate solutions is a critical component of a GA. One may either use the original representation of candidate solutions or one may map the original representation into binary strings, real-valued vectors, or another representation. A good starting point is to use the original representation for the particular problem with no significant modifications or transformations. In the initial stage, modifications should be done only if the chosen implementation necessitates them. If the results obtained are not satisfactory, alternative representations should be examined.

Operators. Numerous operators were described in this article and more can be found in the references provided. What operators should one choose? The best way to start, similarly as for the representations, is to try simple

operators provided in the literature for the representation used. If one starts with a computer implementation of a GA downloaded from the Internet, the implementation most likely includes numerous operators to consider; sticking with the default choices should be a great way to start. If the results are not satisfactory, different operators may be examined and specialized operators may be designed.

Fitness Function. The fitness function resides right at the heart of a GA—the fitness function defines what solutions are good and what solutions are bad. Since the fitness function depends completely on the problem being solved, it is difficult to give a set of general guidelines for designing adequate fitness functions. It is important to keep in mind that the GA does not know what your original problem is; instead, the GA tries to maximize the fitness function over the space of admissible solutions defined by the representation used. The best way to start in the design of a fitness function is to look at the examples from the specific area.

STARTING POINTS FOR OBTAINING ADDITIONAL INFORMATION

Introductory Books and Tutorials

Numerous books and other publications exist that provide introduction to GAs and additional starting points. The following list of references provides some of them: [1,2,24,31,99–105].

Software

The following list provides some of the popular GA implementations available online free of charge. These implementations should provide a good starting point for the interested reader.

- *A C++ Library of Genetic Algorithm Components*
<http://lancet.mit.edu/ga/>

- *Algorithm::Evolutionary, Perl Library for Evolutionary Computation*
<http://opeak.sourceforge.net/>
- *Genetic ALgorithm Optimized for Portability and Parallelism System (GALLOPS)*
<http://garage.cse.msu.edu/software/gallops/index.html>
- *Open BEAGLE, a Versatile Evolutionary Computation Framework*
<http://beagle.gel.ulaval.ca/>
- *Simple GA Implementation in C (based on Goldberg [2])*
<http://www.illigal.uiuc.edu/pub/src/simpleGA/C/>
- *Other GA Software Packages*
<http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/genetic/ga/systems/0.html>

Journals

The following journals are key venues for papers on GAs and evolutionary computation, although papers on GAs can be found in many other journals focusing on optimization, artificial intelligence, machine learning, and applications.

- *Evolutionary Computation* (MIT Press)
<http://www.mitpressjournals.org/loi/>
- *Evolutionary Intelligence* (Springer)
<http://www.springer.com/engineering/journal/12065>
- *Genetic Programming and Evolvable Machines* (Springer)
<http://www.springer.com/computer/ai/journal/10710>
- *IEEE Transactions on Evolutionary Computation* (IEEE Press)
<http://ieeexplore.ieee.org/servlet/opac?punumber=4235>
- *Natural Computing* (Springer)
<http://www.springer.com/computer/theoretical+computer+science/journal/11047>

Conferences

The following conferences provide the most important venues for publishing papers on

GAs and evolutionary computation, although similarly as for journals, papers on GAs are often published in other venues.

- *ACM SIGEVO Genetic and Evolutionary Computation Conference (GECCO)*
- *European Workshops on Applications of Evolutionary Computation (EvoWorkshops)*
- *IEEE Congress on Evolutionary Computation (CEC)*
- *Main European Events on Evolutionary Computation (EvoStar)*
- *Parallel Problem Solving in Nature (PPSN)*
- *Simulated Evolution and Learning (SEAL)*.

Acknowledgments

This project was sponsored by the National Science Foundation (NSF) under CAREER grant ECS-0547013, and by the University of Missouri in St. Louis through the High Performance Computing Collaboratory sponsored by ITS and the Research Award and Research Board programs. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, or the U.S. Government.

REFERENCES

1. Holland JH. Adaptation in natural and artificial systems. Ann Arbor (MI): University of Michigan Press; 1975.
2. Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Reading (MA): Addison-Wesley; 1989.
3. Niesse J, Mayne H. Global geometry optimization of atomic clusters using a modified genetic algorithm in space-fixed coordinates. J Chem Phys 1996;105(11):4700–4706.
4. Sastry K. Efficient atomic cluster optimization using a hybrid extended compact genetic algorithm with seeded population. IlliGAL

- Report No. 2001018. Urbana-Champaign (IL): Genetic Algorithms Laboratory, University of Illinois; 2001.
5. Louis SJ. Learning from experience: case injected genetic algorithm design of combinational logic circuits. In: Parmee IC, editor. *Proceedings of the 5th International Conference on Adaptive Computing in Design and Manufacturing*. Devon, UK: Springer; 2002. pp. 295–306.
 6. Krasnogor N, Smith J. A tutorial for competent memetic algorithms: model, taxonomy, and design issues. *IEEE Trans Evol Comput* 2005;9(5):474–488.
 7. Goldberg DE, Deb K. A comparative analysis of selection schemes used in genetic algorithms. *Found Genet Algorithm* 1991;1:69–93.
 8. Cantú-Paz E. Comparing selection methods of evolutionary algorithms using the distribution of fitness. Technical Report UCRL-JC-138582. Lawrence Livermore National Laboratory, University of California; 2000.
 9. Brindle A. Genetic algorithms for function optimization. Doctoral dissertation and technical report tr81-2. Edmonton: Department of Computer Science, University of Alberta; 1981.
 10. Harik G, Cantú-Paz E, Goldberg DE, *et al.* The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Evol Comput* 1999;7(3):231–253.
 11. Mühlenbein H, Schlierkamp-Voosen D. Predictive models for the breeder genetic algorithm: I. continuous parameter optimization. *Evol Comput* 1993;1(1):25–49.
 12. Thierens D, Goldberg D. Convergence models of genetic algorithm selection schemes. *Parall Probl Solving Nat* 1994;866:116–121.
 13. Goldberg DE, Wang L. Adaptive niching via coevolutionary sharing. *Genetic algorithms and evolution strategy in engineering and computer science*. West Sussex: John Wiley & Sons, Ltd.; 1997. Chapter 2. pp. 21–38.
 14. Oei CK, Goldberg DE, Chang S-J. Tournament selection, niching, and the preservation of diversity. *IlliGAL Report No. 91011*. Urbana (IL): Genetic Algorithms Laboratory, University of Illinois; 1991.
 15. Horn J. Finite Markov chain analysis of genetic algorithms with niching. *Proceedings of the International Conference on Genetic Algorithms (ICGA-93)*. Urbana-Champaign (IL); 1993. pp. 110–117.
 16. Mahfoud SW. Niching methods for genetic algorithms [doctoral dissertation]. Urbana (IL): University of Illinois; 1995. Unpublished.
 17. Goldberg DE, Richardson J. Genetic algorithms with sharing for multimodal function optimization. *Proceedings of the International Conference on Genetic Algorithms (ICGA-87)*. Cambridge (MA); 1987. pp. 41–49.
 18. Horton B. The restricted breed GA for multimodal genetic algorithms: comparison with sequential niching, crowding and sharing. *Evol Comput* 2010. In Press.
 19. De Jong KA. An analysis of the behavior of a class of genetic adaptive systems [PhD thesis]. Ann Arbor (MI): University of Michigan; 1975. (University Microfilms No. 76-9381.)
 20. Harik GR. Finding multimodal solutions using restricted tournament selection. *Proceedings of the International Conference on Genetic Algorithms (ICGA-95)*. Pittsburgh (PA); 1995. pp. 24–31.
 21. Goldberg DE. The design of innovation: lessons from and for competent genetic algorithms. Volume 7, *Genetic algorithms and evolutionary computation*. Boston, Dordrecht, London: Kluwer Academic Publishers; 2002.
 22. Rechenberg I. *Evolutionsstrategie: optimierung technischer systeme nach prinzipien der biologischen evolution*. Stuttgart: Frommann-Holzboog; 1973.
 23. Rechenberg I. *Evolutionsstrategie '94*. Stuttgart: Frommann-Holzboog Verlag; 1994.
 24. Bäck T, Fogel DB, Michalewicz Z. *Handbook of evolutionary computation*. Oxford: Oxford University Press; 1997.
 25. Beyer H-G, Schwefel H-P. Evolution strategies—a comprehensive introduction. *Nat Comput* 2002;1(1):3–52.
 26. Cantú-Paz E. Supervised and unsupervised discretization methods for evolutionary algorithms. *Workshop Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001)*. San Francisco (CA); 2001. pp. 213–216.
 27. Tsutsui S, Pelikan M, Goldberg DE. Evolutionary algorithm using marginal histogram models in continuous domain. *Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES-2001)*. Osaka, Japan; 2001. pp. 112–121.

28. Pelikan M, Goldberg DE, Tsutsui S. Combining the strengths of the Bayesian optimization algorithm and adaptive evolution strategies. IlliGAL Report No. 2001023. Urbana (IL): Illinois Genetic Algorithms Laboratory, University of Illinois; 2001.
29. ping Chen Y, Chen C-H. Enabling the extended compact genetic algorithm for real-parameter optimization by using adaptive discretization. *Evol Comput* 2010; 18(2):199–228.
30. Goldberg DE, Lingle R Jr. Alleles, loci, and the traveling salesman problem. *Proceedings of the International Conference on Genetic Algorithms and their Applications*. Pittsburgh (PA); 1985. pp. 154–159.
31. Davis L. *Handbook of genetic algorithms*. New York: Van Nostrand Reinhold; 1991.
32. Deb K, Agrawal RB. Simulated binary crossover for continuous search space. *Complex Syst* 1995;9:115–148.
33. Harik GR, Goldberg DE. Learning linkage. *Found Genet Algorithm* 1996;4:247–262.
34. Pelikan M. *Hierarchical Bayesian optimization algorithm: Toward a new generation of evolutionary algorithms*. Berlin, Heidelberg: Springer; 2005.
35. ping Chen Y, Yu T-L, Sastry K, *et al.* A survey of genetic linkage learning techniques. IlliGAL Report No. 2007014. Urbana (IL): Illinois Genetic Algorithms Laboratory, University of Illinois; 2007.
36. Goldberg DE, Korb B, Deb K. Messy genetic algorithms: motivation, analysis, and first results. *Complex Syst* 1989;3(5):493–530.
37. Kargupta H. *SEARCH, polynomial complexity, and the fast messy genetic algorithm* [PhD thesis]. Urbana (IL): University of Illinois; 1995.
38. Hartl DL, Jones EW. *Genetics: principles and analysis*. 4th ed. Boston (MA): Jones and Bartlett Publishers; 1998.
39. Deb K. *Binary and floating-point function optimization using messy genetic algorithms* [PhD thesis]. Urbana (IL): University of Illinois; 1991.
40. Goldberg DE, Deb K, Kargupta H. Rapid, accurate optimization of difficult problems using fast messy genetic algorithms. *Proceedings of the International Conference on Genetic Algorithms*. Urbana-Champaign (IL); 1993. pp. 56–64.
41. Bandyopadhyay S, Kargupta H, Wang G. Revisiting the GEMGA: Scalable evolutionary optimization through linkage learning. *Proceedings of the International Conference on Evolutionary Computation (ICEC-98)*. Seoul, Korea; 1998. pp. 603–608.
42. Munetomo M, Goldberg DE. Linkage identification by non-monotonicity detection for overlapping functions. *Evol Comput* 1999;7(4):377–398.
43. Heckendorn RB, Wright AH. Efficient linkage discovery by limited probing. *Evol Comput* 2004;12(4):517–545.
44. Yu T-L, Goldberg DE, Chen Y-P. A genetic algorithm design inspired by organizational theory: a pilot study of a dependency structure matrix driven genetic algorithm. IlliGAL Report No. 2003007. Urbana (IL): Illinois Genetic Algorithms Laboratory, University of Illinois; 2003.
45. Yu T-L, Goldberg DE, Sastry K, Lima CF, Pelikan M. Dependency structure matrix, genetic algorithms, and effective recombination. *Evol Comput* 2009;17(4):595–626.
46. Harik GR. *Learning gene linkage to efficiently solve problems of bounded difficulty using genetic algorithms* [PhD thesis]. Ann Arbor (MI): University of Michigan; 1997.
47. Pelikan M, Goldberg DE, Lobo F. A survey of optimization by building and using probabilistic models. *Comput Optim Appl* 2002;21(1):5–20.
48. Larrañaga P, Lozano JA, editors. *Estimation of distribution algorithms: a new tool for evolutionary computation*. Boston (MA): Kluwer Academic Publishers; 2002.
49. Pelikan M, Sastry K, Cantú-Paz E, editors. *Scalable optimization via probabilistic modeling: From algorithms to applications*. Berlin, Heidelberg: Springer; 2006.
50. Lozano JA, Larrañaga P, Inza I, *et al.* editors. *Towards a new evolutionary computation: Advances on estimation of distribution algorithms*. Berlin, Heidelberg: Springer; 2006.
51. Pelikan M, Goldberg DE, Cantú-Paz E. BOA: the Bayesian optimization algorithm. Volume I, Genetic and Evolutionary Computation Conference (GECCO-99). Orlando (FL); 1999. pp. 525–532.
52. Pelikan M, Goldberg DE, Cantú-Paz E. Bayesian optimization algorithm, population sizing, and time to convergence. In: Whitley D, Goldberg D, Cantu-Paz E, *et al.* editors. *Genetic and Evolutionary Computation Conference (GECCO-2000)*. San Francisco (CA): Morgan Kaufmann; 2000. pp. 275–282.

53. Larrañaga P, Etxeberria R, Lozano J, *et al.* Combinatorial optimization by learning and simulation of Bayesian networks. Proceedings of the Uncertainty in Artificial Intelligence (UAI-2000). Stanford (CA): 2000. pp. 343–352.
54. Mühlenbein H, Mahnig T. FDA—A scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evol Comput* 1999;7(4):353–376.
55. Harik G. Linkage learning via probabilistic modeling in the ECGA. IlliGAL Report No. 99010. Urbana (IL): Illinois Genetic Algorithms Laboratory, University of Illinois; 1999.
56. Deb K. Multi-objective optimization using evolutionary algorithms. Chichester: John Wiley & Sons, Ltd.; 2001.
57. Coello CAC, Lamont GB, Veldhuizen DA. Evolutionary algorithms for solving multi-objective problems. Genetic and evolutionary computation. 2nd ed. New York: Springer; 2007.
58. Fonseca CM, Fleming PJ. Genetic algorithms for multiobjective optimization: formulation, discussion and generalization. In: Forrest S, editor. Proceedings of the International Conference on Genetic Algorithms (ICGA-93). San Mateo (CA): Morgan Kaufmann; 1993. pp. 416–423.
59. Horn J, Nafpliotis N. Multiobjective optimization using the niched pareto genetic algorithm. IlliGAL Report No. 93005. Urbana (IL): University of Illinois; 1993 Jul.
60. Deb K, Agrawal S, Pratap A, *et al.* A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Trans Evol Comput* 2002;6(2):182–197.
61. Zitzler E, Laumanns M, Thiele L. SPEA2: improving the strength pareto evolutionary algorithm. TIK-Report 103. Zurich: Department of Electrical Engineering, Swiss Federal Institute of Technology (ETH) Zurich; 2001.
62. Arnold D. Noisy optimization with evolution strategies. Genetic algorithms and evolutionary computation. Norwell (MA): Kluwer Academic Publishers; 2002.
63. Miller BL, Goldberg DE. Genetic algorithms, selection schemes, and the varying effects of noise. *Evol Comput* 1996;4(2):113–131.
64. Miller BL, Goldberg DE. Optimal sampling for genetic algorithms. *Intell Eng Syst Artif Neural Netw* 1996;6:291–297.
65. Michalewicz Z, Janikow CZ. Handling constraints in genetic algorithms. Proceedings of the International Conference on Genetic Algorithms (ICGA-91). San Diego (CA); 1991. pp. 151–157.
66. Baeck T, Hoffmeister F, Schwefel H-P. A survey of evolution strategies. Proceedings of the International Conference on Genetic Algorithms (ICGA-91). San Diego (CA); 1991. pp. 2–9.
67. Schoenauer M, Xanthakis S. Constrained GA optimization. Proceedings of the International Conference on Genetic Algorithms (ICGA-93). Urbana-Champaign (IL); 1993. pp. 573–580.
68. Orvosh D, Davis L. Shall we repair? Genetic algorithms, combinatorial optimization, and feasibility constraints. Proceedings of the International Conference on Genetic Algorithms (ICGA-93). Urbana-Champaign (IL); 1993. pp. 650.
69. Powell D. Using genetic algorithms in engineering design optimization with non-linear constraints. Proceedings of the International Conference on Genetic Algorithms (ICGA-93). Urbana-Champaign (IL); 1993. pp. 424–430.
70. Homaifar A, Qi CX, Lai SH. Constrained optimization via genetic algorithms. *Simulation* 1994;62(4):242–254.
71. Joines JA, Houck CR. On the use of non-stationary penalty functions to solve nonlinear constrained optimization problems with gas. Proceedings of the International Conference on Evolutionary Computation (ICEC-94). Orlando (FL); 1994. pp. 579–584.
72. Deb K. An efficient constraint handling method for genetic algorithms. *Comput Method Appl Mech Eng* 2000;186(2–4):311–338.
73. Coello CAC. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Comput Method Appl Mech Eng* 2002;191(11–12):1245–1287.
74. Michalewicz Z. A survey of constraint handling techniques in evolutionary computation methods. Proceedings of the 4th Annual Conference on Evolutionary Programming. San Diego (CA); 1995. pp. 135–155.
75. Cohoon JP, Hegde SU, Martin WN, *et al.* Punctuated equilibria: a parallel genetic algorithm. Proceedings of the International Conference on Genetic Algorithms (ICGA-87). Cambridge (MA); 1987. pp. 148–154.

76. Goldberg DE. Sizing populations for serial and parallel genetic algorithms. *Proceedings of the International Conference on Genetic Algorithms (ICGA-89)*. Fairfax (VA); 1989. pp. 70–79.
77. Mühlenbein H. Evolution in time and space—the parallel genetic algorithm. *Found Genet Algorithm* 1991;1:316–337.
78. Cantú-Paz E. A survey of parallel genetic algorithms. *Calculateurs Parall Reseaux Systeme Repartis* 1998;10(2):141–171.
79. Cantú-Paz E. *Efficient and accurate parallel genetic algorithms*. Boston (MA): Kluwer Academic Publishers; 2000.
80. Cantú-Paz E. Efficient parallel genetic algorithms: theory and practice. *Comput Method Appl Mech Eng* 2000;186(4):221–238.
81. Sinha A, Goldberg DE. A survey of hybrid genetic and evolutionary algorithms. *IlliGAL Report No. 2002XXX*. Urbana (IL): Illinois Genetic Algorithms Laboratory, University of Illinois; 2002.
82. Hinton GE, Nowlan SJ. How learning can guide evolution. *Complex Syst* 1987; 1:495–502.
83. Goldberg DE, Voessner S. Optimizing global-local search hybrids. *Genetic and Evolutionary Computation Conference (GECCO-99)*. Orlando (FL); 1999. pp. 220–228.
84. Sinha A, Goldberg DE. Verification and extension of the theory of global-local hybrids. *Genetic and Evolutionary Computation Conference (GECCO-2001)*. San Francisco (CA); 2001. pp. 591–597.
85. Sinha A. *Designing efficient genetic and evolutionary hybrids* [Master's thesis]. *IlliGAL Report No. 2003020*. Urbana (IL): University of Illinois; 2003.
86. Sinha A, Chen Y-P, Goldberg DE. Designing efficient genetic and evolutionary algorithm hybrids. In: Hart WE, Krasnogor N, Smith J, editors. *Recent advances in memetic algorithms*. Berlin, Heidelberg: Springer; 2004. pp. 259–288.
87. Pelikan M, Hartmann AK. Searching for ground states of Ising spin glasses with hierarchical BOA and cluster exact approximation. In: Cantú-Paz E, Pelikan M, Sastry K, editors. *Scalable optimization via probabilistic modeling: From algorithms to applications*. Berlin, Heidelberg: Springer; 2006. pp. 333–349.
88. Radetic E, Pelikan M, Goldberg DE. Effects of a deterministic hill climber on hBOA. *Genetic and Evolutionary Computation Conference (GECCO-2009)*. Montreal, Canada; 2009. pp. 437–444.
89. Sastry K, Pelikan M, Goldberg DE. Efficiency enhancement of estimation of distribution algorithms. In: Pelikan M, Sastry K, Cantú-Paz E, editors. *Scalable optimization via probabilistic modeling: From algorithms to applications*. Berlin, Heidelberg: Springer; 2006. pp. 161–185.
90. Sastry K, Goldberg DE, Pelikan M. Don't evaluate, inherit. *Genetic and Evolutionary Computation Conference (GECCO-2001)*. San Francisco (CA); 2001. pp. 551–558.
91. Pelikan M, Sastry K. Fitness inheritance in the Bayesian optimization algorithm. Volume 2, *Genetic and Evolutionary Computation Conference (GECCO-2004)*. Seattle (WA); 2004. pp. 48–59.
92. Jin Y. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Comput* 2005;9(1):3–12.
93. Smith RE, Dike BA, Stegmann SA. Fitness inheritance in genetic algorithms. *Proceedings of the ACM Symposium on Applied Computing*. Nashville (TN); 1995. pp. 345–350.
94. Goldberg DE. Using time efficiently: genetic-evolutionary algorithms and the continuation problem. *Genetic and Evolutionary Computation Conference (GECCO-99)*. Orlando (FL); 1999. pp. 212–219.
95. Srivastava RP, Goldberg DE. Verification of the theory of genetic and evolutionary continuation. *Genetic and Evolutionary Computation Conference (GECCO-2001)*. San Francisco (CA); 2001. pp. 551–558.
96. Srivastava RP. *Time continuation in genetic algorithms* [Master's thesis]. Urbana (IL): University of Illinois; 2002.
97. DeJong KA, Spears WM. An analysis of the interacting roles of population size and crossover in genetic algorithms. *International Workshop on Parallel Problem Solving from Nature*; Dortmund, Germany: University of Dortmund; 1990. 38–47.
98. Lobo FG, Lima CF, Michalewicz Z, editors. *Parameter setting in evolutionary algorithms*. Studies in computational intelligence series. Berlin, Heidelberg: Springer; 2007.
99. Davis LD, editor. *Genetic algorithms and simulated annealing*. London: Pitman; 1987.
100. Whitley D. A genetic algorithm tutorial. *Stat Comput* 1994;4:65–85.

101. Reeves CR. Genetic algorithms. In: Reeves CR, editor. Modern heuristic techniques for combinatorial problems. New York: McGraw-Hill; 1995.
102. Michalewicz Z. Genetic algorithms + data structures = evolution programs. 3rd ed. Berlin: Springer; 1996.
103. Mitchell M. An introduction to genetic algorithms. Cambridge (MA): MIT Press; 1996.
104. Eiben AE, Smith JE. Introduction to evolutionary computing. Berlin, Heidelberg: Springer; 2003.
105. Sastry K, Goldberg DE, Kendall G. Genetic algorithms: a tutorial. In: Introductory tutorials in optimization, search and decision support methodologies. Berlin: Springer; 2005. Chapter 4. pp. 97–125.

GEOMETRIC PROGRAMMING

JAYANT RAJGOPAL

Department of Industrial Engineering,
University of Pittsburgh, Pittsburgh,
Pennsylvania

Geometric programming (GP) is a methodology for solving algebraic nonlinear optimization problems that possess a specific structure. In this article, we present a very brief historical context for GP, followed by principal results for the original GP problem, various subsequent generalizations, and an overview of computational issues and solution algorithms.

HISTORICAL CONTEXT

The original GP formulation and much of the supporting theory were developed between 1961 and 1967. The initial motivation was provided by the physicist Clarence Zener, who was the director of Westinghouse Electric Corporation Research Labs in Pittsburgh. While looking for more efficient ways to solve engineering design problems that were of interest to Westinghouse, Zener published two papers in 1961–1962 [1,2] where he showed that problems with a certain structure could be solved very simply but elegantly, and that the solution also had invariance properties that were interesting from an engineering perspective. Coincidentally, during the same time period Richard Duffin, a mathematician at Carnegie Institute of Technology (now Carnegie Mellon University) in Pittsburgh, was working on a duality theory with applications to nonlinear programming (NLP). On learning of each other's work, the two began a collaborative effort to ground Zener's technique with the proper mathematical foundation, resulting in the formal birth of GP. Subsequently, Elmor Peterson, a doctoral student of Duffin at the time, worked on formulating and developing the theory for constrained problems.

Possibly the most important feature of GP is that a problem with highly nonlinear constraints can be stated equivalently in the dual form with only linear constraints; this is obviously very desirable when developing solution procedures. A second important feature of GP is that it yields invariance properties with very useful practical implications. For example, in some cases for a given set of design constraints the different cost components of an optimal design must make the same *relative* contributions to total cost, regardless of the individual cost coefficients for the components. These invariance properties are especially significant in engineering design.

The joint efforts of Duffin *et al.* culminated in the publication of the seminal text in this area [3] in 1967. Since then, GP's attractive structural properties and its elegant theoretical basis have led to the development of a number of algorithms for solving GP problems, several extensions of the original formulation to more general forms, and a wide and eclectic range of applications. It is now a mature area and many texts on NLP cover the subject in varying degrees of detail. Books and collected volumes specifically focused on GP include Duffin *et al.* [3], Zener [4], Beightler and Phillips [5], Avriel [6], Cao [7], Chiang [8].

POSYNOMIAL GEOMETRIC PROGRAMMING

In this section we present the original version of GP, along with its duality theory.

The Primal-Dual Pair in GP

The primal geometric program may be stated as follows:

Program P

$$\text{Minimize } g_0(\mathbf{t}) \quad (1)$$

$$\text{s.t. } g_k(\mathbf{t}) \leq 1, \quad k = 1, 2, \dots, p, \quad (2)$$

$$\mathbf{t} \in \mathbf{R}_{++}^m,$$

$$\begin{aligned} \text{where } g_k(\mathbf{t}) &= \sum_{i \in [k]} c_i \prod_{j=1}^m t_j^{a_{ij}}, \\ k &= 0, 1, \dots, p; \quad c_i \in \mathbf{R}_{++} \forall i, \\ a_{ij} &\in \mathbf{R} \forall i, j. \end{aligned} \quad (3)$$

Each function $g_k (k = 0, 1, \dots, p)$ is a *generalized* polynomial consisting of a sum of terms, where each term within the polynomial has a strictly positive coefficient followed by a product of the design variables, with each raised to an exponent. Unlike a standard polynomial, these exponents are not restricted to the set of nonnegative integers but may take on any real value (including negative values). These generalized polynomials with positive coefficients are referred to as *posynomials*. The domain of a posynomial is thus necessarily restricted to the set of strictly positive real numbers. There are a total of n terms included in the $p + 1$ posynomials, and the index set $I = [1, 2, \dots, n]$ is used to number these sequentially so that the first term of posynomial 0 has the index 1 and the last term of posynomial p has the index n . The index subset $[k]$ numbers the terms in the k th posynomial, with $I = [0] \cup [1] \cup \dots \cup [p]$ and $[k] \cap [l] = \emptyset$ for $k \neq l$. The inputs to the problem are (i) the (positive) coefficient vectors $\mathbf{c}^k, k = 0, 1, \dots, p$ and (ii) the $n \times m$ exponent matrix $\mathbf{A}$, where the entry in row i and column $j (= a_{ij})$, is the exponent for variable t_j in term i . As an illustrative example consider the following formulation [5]:

$$\begin{aligned} \text{Minimize } & t_1^{-1} t_2^{-1} t_3 + 25 t_1 t_2^{-1} \\ \text{s.t. } & t_1 t_2 \leq 1 \\ & 5 t_1 t_3^{-1} + t_2 t_3^{-1} \leq 1; \quad t_1, t_2, t_3 > 0 \end{aligned}$$

Here $m = 3; n = 5; p = 2; [0] = \{1, 2\}, [1] = \{3\}, [2] = \{3, 4\}$, with $I = [0] \cup [1] \cup [2] = \{1, 2, 3, 4, 5\}$

$$\mathbf{c} = \begin{bmatrix} 1 \\ 25 \\ 1 \\ 5 \\ 1 \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} -1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

The dual corresponding to Program P may be stated as

Program D

$$\text{Maximize } v(\boldsymbol{\lambda}, \boldsymbol{\delta}) = \left(\prod_{i=1}^n \left(\frac{c_i}{\delta_i} \right)^{\delta_i} \right) \left(\prod_{k=0}^p \lambda_k^{\lambda_k} \right) \quad (4)$$

$$\text{s.t. } \sum_{i \in [k]} \delta_i = \lambda_k, \quad k = 1, 2, \dots, p, \quad (5)$$

$$\sum_{i \in [0]} \delta_i = \lambda_0 = 1, \quad (6)$$

$$\sum_{i=1}^n \delta_i a_{ij} = 0, \quad j = 0, 1, \dots, m, \quad (7)$$

$$\begin{aligned} \delta_i, \lambda_k &\geq 0 \text{ for } i \in I \text{ and} \\ k &= 1, 2, \dots, p. \end{aligned}$$

The constraint specified by Equation (6) is referred to as the *normality* condition and those specified in Equation (7) are referred to as the *orthogonality* conditions; the dual vector $\boldsymbol{\delta}$ must be orthogonal to each column of $\mathbf{A}$. It is also shown in Duffin *et al.* [3] that λ_k for $k \geq 1$ is equivalent to the Lagrange multiplier for constraint k . Note that effectively there are n variables (one δ_i corresponding to each term in the primal) and $(m + 1)$ constraints, since the vector $\boldsymbol{\lambda}$ could be eliminated by substituting $\lambda_k = \sum_{i \in [k]} \delta_i$ for all k . Also note that while the term $(c_i/\delta_i)^{\delta_i}$ is undefined if $\delta_i = 0$, we define this to be 1, since this quantity goes to 1 in the limit as $\delta_i \rightarrow 0^+$. The earlier example yields

$$\begin{aligned} \text{Maximize } v(\boldsymbol{\lambda}, \boldsymbol{\delta}) &= \left(\frac{1}{\delta_1} \right)^{\delta_1} \left(\frac{25}{\delta_2} \right)^{\delta_2} \left(\frac{1}{\delta_3} \right)^{\delta_3} \\ &\quad \left(\frac{5}{\delta_4} \right)^{\delta_4} \left(\frac{1}{\delta_5} \right)^{\delta_5} (\lambda_0)^{\lambda_0} (\lambda_1)^{\lambda_1} (\lambda_2)^{\lambda_2} \\ \text{s.t. } &\delta_1 + \delta_2 = \lambda_0; \delta_3 = \lambda_1; \delta_4 + \delta_5 = \lambda_2; \\ &\lambda_0 = 1 \\ &-\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0; \\ &-\delta_1 - \delta_2 + \delta_3 + \delta_5 = 0; \\ &-\delta_4 - \delta_5 = 0; \\ &\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\delta} \geq \mathbf{0}. \end{aligned}$$

Next, we present a transformation that yields an attractive result for the primal. Suppose we define $z_j = \ln t_j$, so that $t_j = e^{z_j}$. We obtain the *transformed* primal, which is stated as follows:

Program P_z

$$\text{Minimize } g_0(\mathbf{z}) \quad (8)$$

$$\text{st } g_k(\mathbf{z}) \leq 1, k = 1, 2, \dots, p, \quad (9)$$

$$\mathbf{z} \in \mathbb{R}^m,$$

$$\begin{aligned} \text{where } g_k(\mathbf{z}) &= \sum_{i \in [k]} c_i e^{\sum_{j=1}^m a_{ij} z_j} \\ &= \sum_{i \in [k]} e^{\chi_i + \sum_{j=1}^m a_{ij} z_j}, \\ k &= 0, 1, \dots, p, c_i > 0. \end{aligned} \quad (10)$$

Note that we define $\chi_i = \ln c_i$ and that the transformed functions $g_k(\mathbf{z})$ are all positive exponential functions. To examine some of the results that follow from P_z , let us define

$$x_i = \sum_{j=1}^m a_{ij} z_j, \quad i \in I. \quad (11)$$

Thus the vector $\mathbf{x} \in \mathbb{R}^n$ is restricted to lie in the column space of the exponent matrix $\mathbf{A}$. Now consider the equivalent problem of minimizing $\sum_{i \in [0]} c_i e^{x_i}$, st $\sum_{i \in [k]} c_i e^{x_i} \leq 1, k = 1, 2, \dots, p$. If the rank of $\mathbf{A}$ is less than m , then there is some partition of its column vectors into a basis with linearly independent vectors and another set with the remaining columns, each of which is linearly dependent on the columns in the basis. The variables z_j corresponding to the columns in the latter set can then be assigned arbitrary values without having any effect on the minimum value of P_z ; thus they can essentially be eliminated from the problem along with their columns in $\mathbf{A}$. Thus, the rank of $\mathbf{A}$ is at least m . Now suppose $m = n$ (so that $\mathbf{A}$ is square with full rank). Consider a vector $\mathbf{M} \in \mathbb{R}^n$ with every element equal to $-M$, where M is some arbitrarily large value. We can always find $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{A}\mathbf{z} = \mathbf{M}$, that is, such that $\sum_{j=1}^m a_{ij} z_j = -M$ for every i . It follows from Equation (10) that each $g_k(\mathbf{z})$ can then be made arbitrarily close to zero, and thus P_z

reduces to the trivial case where it is feasible but has an infimum of 0 that is never attained. These observations lead to the following proposition.

Proposition 1. *Every nontrivial GP problem can be formulated so that (i) the exponent matrix has rank m , and (ii) $m < n$.*

The nonnegative quantity $D = n - (m + 1)$ is referred to as the *degree of difficulty* of the GP problem; the term is related to how underdetermined the dual constraints system ((6) and (7)) is. For our earlier example, $D = 1$. Another important result that follows from the above transformation is the following theorem, the proof for which may be found in Duffin *et al.* [3].

Theorem 1. *Program P_z is a convex program.*

Note that not all posynomials are convex, so that not all primal problems in the format of Program P are convex programs; however, they can all be restated as convex programs using the above transformation.

Duality Theory for GP

We now present (without detailed proofs) the main duality theorems for posynomial GP. The proofs hinge on a generalization of the classical inequality relating the arithmetic and geometric means of a set of numbers (the former is at least as large as the latter), and make use of the standard Lagrange multiplier approach. The interested reader is referred to Duffin *et al.* [3] for details. For the sake of notational convenience, we assume that the λ_k variables in the dual formulation have been eliminated by substituting $\lambda_k = \sum_{i \in [k]} \delta_i$ for all k , and thus, we use $v(\lambda, \delta)$ and $v(\delta)$ interchangeably.

Theorem 2. *Suppose $\mathbf{t} \in \mathbb{R}^m$ is feasible in the primal (Program P), and the vector $\delta \in \mathbb{R}^n$ is feasible in the dual (Program D). Then $g_0(\mathbf{t}) \geq v(\delta)$.*

Thus, any solution to the linear system represented by the dual constraints readily generates a lower bound on the primal

optimum using the above “weak” duality theorem. Next, we state conditions that guarantee the existence of feasible vectors and a stronger duality result. Before doing so we make the following definition.

Definition 1. Program P is said to be superconsistent if there is at least one vector $\mathbf{t} \in \mathbf{R}_{++}^m$ such that $g_k(\mathbf{t}) < 1, k = 1, 2, \dots, p$.

Theorem 3. Suppose (i) Program P is superconsistent and (ii) the primal objective $g_0(\mathbf{t})$ attains its constrained minimum at a point $\mathbf{t}^*$ feasible in P . Then Program D is consistent and the dual objective $v(\delta)$ attains its constrained maximum at a point (δ^*) feasible in D . Moreover,

$$1. \quad g_0(\mathbf{t}^*) = v(\delta^*) \quad (12)$$

$$2. \quad \delta_i^* = \begin{cases} \frac{c_i \prod_{j=1}^m (t_j^*)^{a_{ij}}}{g_0(\mathbf{t}^*)} = \frac{c_i \prod_{j=1}^m (t_j^*)^{a_{ij}}}{v(\delta^*)}, i \in [0] \\ \lambda_k^* \left(c_i \prod_{j=1}^m (t_j^*)^{a_{ij}} \right), i \in [k], \\ k = 1, 2, \dots, p \end{cases} \quad (13)$$

The above “strong” duality theorem also provides via Equation (13) a relationship between the optimal values of the primal and dual variables. Note that superconsistency is like a regularity condition in general NLP and the other assumption eliminates pathological cases where the GP might have an infimum that is not attained. The following theorem specifies conditions under which this latter assumption holds.

Theorem 4. If Program P is consistent and there exists $\delta \in \mathbf{R}_{++}^n$ that is feasible in Program D , then the objective for Program P attains its constrained minimum at a feasible point.

While the duality results just presented are adequate for most practical situations, they are not comprehensive because of requirements such as superconsistency and the attainment of minima/maxima. We

conclude this section with a discussion of the so-called *refined* duality theory of GP, which dispenses with these requirements. Since the results from this are primarily of theoretical interest, our discussion is very brief.

We now generalize the problem so that the primal objective is to find the constrained *infimum* of $g_0(\mathbf{t})$, while the dual objective is to find the constrained *supremum* of $v(\lambda, \delta)$. Much of the refined duality theory is based on determining if for each value of $i = 1, 2, \dots, n$ there is at least one vector $\delta \in \mathbf{R}_+^m$ that satisfies the orthogonality constraints (7) and also has $\delta_i > 0$. The set of all i for which this is true determines the so-called *irreducible integer set* (Γ). If Γ is identical to the original index set I , the corresponding primal–dual pair is said to be *canonical*; otherwise it is said to be *degenerate*, and if $\Gamma \cap [0] = \emptyset$ then it is said to be *totally degenerate*. Essentially, canonical programs may be viewed as well-behaved with strong results similar to those presented earlier in this section, while totally degenerate programs represent the other end of the spectrum with inconsistent duals and a primal infimum of zero even when the primal is consistent. Degenerate programs that are not totally degenerate lie in between; they can be reduced to canonical form by essentially deleting those terms that do not belong to Γ . For more details on these concepts as well as other results, the interested reader is referred to Duffin *et al.* [3] or Rajgopal [9].

EXTENSIONS OF GEOMETRIC PROGRAMMING

Typically, the term *generalized geometric programming* (GGP) is used to describe extensions of the original (posynomial) GP formulation to more general forms. However, there appears to be a fair amount of ambiguity in terms of what exactly this term encompasses. For example, Beightler and Phillips [5] describe a GGP as one where the coefficient c_i in term i does not have to be strictly positive, and where the directions of the inequalities in the constraints could be reversed; this latter situation, in particular, is referred to as *reversed geometric programming*, and the class of problems just described

is more generally referred to by the term *signomial geometric programming (SGP)*. The initial work in this area was done by Passy and Wilde [10]. On the other hand, Peterson [11,12] uses the term GGP to describe a more general class of separable problems with an underlying algebraic structure that is linear or for which linearity can be induced. He then derives general duality and optimality results for this class of problems based mainly on conjugate transforms [13]. More recently, Boyd *et al.* [14] and Chiang [8] describe a GGP as a program in which the objective and constraint functions are *generalized posynomials*. The latter term is used to describe any function that can be formed from posynomials using the operations of addition, multiplication, (positive) exponentiation, and maximum. Such functions have the advantage of maintaining convexity.

We take a neutral view and simply use this section to consider various extensions to the original (posynomial) GP formulation that have been discussed in the literature. While these extensions encompass a much broader class of algebraic nonlinear optimization problems, convexity is unfortunately not guaranteed any longer in many cases, and thus the duality results for these are not as elegant. Nevertheless, several of these extensions have been well studied (especially from a computational perspective) since they capture many practical problems.

Signomial GP (SGP)

The first extension uses signomial functions. These are posynomials where each term is multiplied by a *signum* factor of ± 1 . Moreover, to account for reversed constraints the RHS of each constraint could now be 1 or -1. The primal may be stated as follows:

Program SP

Minimize $g_0(\mathbf{t})$

$$\text{st } g_k(\mathbf{t}) \leq \sigma_{k0}, \quad k = 1, 2, \dots, p, \quad (14)$$

$$\mathbf{t} \in \mathbf{R}_{++}^m,$$

$$\text{where } g_k(\mathbf{t}) = \sum_{i \in [k]} \sigma_i c_i \prod_{j=1}^m t_j^{a_{ij}},$$

$$k = 0, 1, \dots, p, c_i > 0, \quad (15)$$

$$\text{and } \sigma_{k0} = \pm 1, k = 1, 2, \dots, p;$$

$$\sigma_i = \pm 1, i = 1, 2, \dots, n.$$

The formulation is almost identical to Program P , with the exception of the signum factors σ_i and σ_{k0} that capture the sign of the coefficient for term i and the RHS of constraint k respectively; note that these are *given* as part of the problem formulation. This also allows for a much more generalized constraint structure including “ $\geq$ ” and equality constraints. To construct the dual of this problem one needs to introduce a new signum factor σ_{00} , which represents the sign of the primal objective at the optimum. This is not known *a priori* but for many practical problem formulations it can be guessed intelligently, and in engineering applications it is positive in the vast majority of cases. In the worst case, a wrong guess can be discovered during the course of the dual computations.

Program SD

$$\begin{aligned} \text{Maximize } v(\lambda, \delta) = & \sigma_{00} \left(\prod_{i=1}^n \left(\frac{c_i}{\delta_i} \right)^{\delta_i \sigma_i} \right. \\ & \left. \times \prod_{k=1}^p \lambda_k^{\lambda_k \sigma_{k0}} \right)^{\sigma_{00}} \end{aligned} \quad (16)$$

$$\text{s.t. } \sigma_{k0} \sum_{i \in [k]} \sigma_i \delta_i = \lambda_k, \quad k = 1, 2, \dots, p, \quad (17)$$

$$\sigma_{00} \sum_{i \in [0]} \sigma_i \delta_i = 1, \quad (18)$$

$$\sum_{i=1}^n \sigma_i \delta_i a_{ij} = 0, \quad j = 0, 1, \dots, m, \quad (19)$$

$$\delta_i, \lambda_k \geq 0 \text{ for } i \in I \text{ and}$$

$$k = 1, 2, \dots, p.$$

The constraints given by Equations (18) and (19) are referred to as *generalized normality* and *generalized orthogonality* conditions, respectively. Unlike the dual program in the section titled “Duality Theory for GP,” this

Program SD does not in general yield any strong duality results and since Program SP cannot be reformulated as a convex program, there will in general, be a duality gap. Thus this dual is not particularly useful since duality results for this pair are not as strong as the ones in the section titled “Duality Theory for GP,” and as a result, they also have limited use from an algorithmic/computational perspective. For more discussion on special cases of the SGP problem the reader is referred to Beightler and Phillips [5].

Transcendental Geometric Programs (TGP)

Introduced by Lidor and Wilde [15], this extension to GP allows some of the variables (referred to as *transcendental variables*) to also appear as exponents or logarithms in the primal geometric program, resulting in the so-called posynomial functions. Suppose that in addition to the usual m “algebraic” variables $t_1, t_2, \dots, t_m$, we have q such transcendental variables; as a matter of notational convenience we denote these via $s_1, s_2, \dots, s_q$. Then the primal transcendental geometric programming (TGP) is defined as

Program TGP

$$\text{Minimize } g_0(\mathbf{t}, \mathbf{s}) \quad (20)$$

$$\text{st } g_k(\mathbf{t}, \mathbf{s}) \leq 1, \quad k = 1, 2, \dots, p, \quad (21)$$

$$\mathbf{t} \in \mathbf{R}_{++}^m, \mathbf{s} \in \mathbf{R}_{++}^q,$$

$$\begin{aligned} \text{where } g_k(\mathbf{t}, \mathbf{s}) = & \sum_{i \in [k]} c_i \left(\prod_{j=1}^m t_j^{a_{ij}} \right) \\ & \times \left(\prod_{l=1}^q s_l^{b_{il}} \exp(d_{il} s_l) \right), \\ & k = 0, 1, \dots, p; \end{aligned} \quad (22)$$

$$\text{and } c_i \in \mathbf{R}_{++} \forall i;$$

$$a_{ij}, b_{il}, d_{il} \in \mathbf{R} \forall i, j, l, c_i > 0.$$

Note that if either (i) $d_{il} = 0$ for all i and l , or (ii) $b_{il} = 0$ for all i and l , we have a regular GP in $m + l$ variables (in case (ii) we can define $s'_l = \exp(s_l)$ to obtain the necessary form). An example of a function that leads to a transcendental program is where each term in

the summation in Equation (22) is a logarithmic function of the form $c_i \prod_{j=1}^m x_j^{a_{ij}} (\log x_j)^{b_{ij}}$; here we get the necessary form by defining $s_j = \log x_j$. Unfortunately, because of the lack of convexity Program TGP could have several local minima and even though a dual can be derived for it, the latter is not very useful in that it does not provide a lower bound on the optimum value of the primal. However, if we *fix* the transcendental variables, conditions can be specified whereby the primal–dual pair reduces to the traditional GP primal–dual pair. For more details and other interesting properties, the interested reader is referred to Lidor and Wilde [15].

Quadratic Posylognomials (QPL)

These functions were introduced and analyzed as extensions to GP by Hough and Goforth [16]. To understand the motivation, consider a term in a posynomial: $u = c \prod_j t_j^{a_j}$. This may be rewritten as $\ln u = \ln c + \sum_j a_j \ln t_j$, that is, as a linear function of the logarithm of the design variables. This type of log-linear relationship has long been known in many engineering and scientific contexts; this was even pointed out by Zener during the original development of GP. However, it has been postulated that in some cases (e.g., the machining economics problem with Taylor’s tool life equation) a quadratic—rather than a linear—logarithmic relationship of the following form is more appropriate:

$$\ln u = \ln c + \sum_{j=1}^m a_j \ln t_j + \frac{1}{2} \sum_{j=1}^m \sum_{l=1}^m b_{jl} \ln t_j \ln t_l. \quad (23)$$

Upon exponentiation this leads to the form

$$\begin{aligned} u &= c \prod_{j=1}^m t_j^{a_j} \cdot \prod_{j=1}^m t_j^{\frac{1}{2} \sum_{l=1}^m b_{jl} \ln t_l} \\ &= c \prod_{j=1}^m t_j^{a_j + \frac{1}{2} \sum_{l=1}^m b_{jl} \ln t_l}. \end{aligned} \quad (24)$$

Hough and Goforth define a *quadratic posylognomial (QPL)* as any finite sum of terms such as those in Equation (24) and

regular posynomial terms, and attempted to develop extensions of traditional GP theory to programs involving such functions. However, there is no way to reduce these programs to the GP form so that, once again, the results are not as clean or as powerful as in traditional GP.

Quadratic Geometric Programming (QGP)

Consider the symmetric $m \times m$ matrix $\mathbf{B} = [b_{jl}]$ of quadratic term multipliers in Equation (23). It can be shown that if $\mathbf{B}$ is positive semidefinite, then each term as given by Equation (24) is convex, and hence the entire quadratic posylognomial is convex in $\mathbf{z} = \ln \mathbf{t}$. Under such conditions the QPL formulation of the previous section reduces to a convex program in the logarithm of the original design variables, similar to Program P_z in the section titled “The Primal–Dual Pair in GP.” Now consider the functions $g_k(\mathbf{z})$ given by Equation (10) that were used to define Program P_z . Jefferson and Scott [17] redefine this by appending to the linear function, a quadratic term that is defined by a symmetric positive definite matrix $\mathbf{B}$:

$$g_k(\mathbf{z}) = \sum_{i \in [k]} \exp(\chi_i + \mathbf{a}^i \mathbf{z} + \mathbf{z}^T \mathbf{B} \mathbf{z}); \quad k = 0, 1, \dots, p. \quad (25)$$

As before, we define $\chi_i = \ln c_i$, and $\mathbf{a}^i$ is a row vector of size m representing the i th row of the original exponent matrix $\mathbf{A}$. They then define a quadratic geometric programming (QGP) as follows:

Program QGP

$$\text{Minimize } g_0(\mathbf{z}) \quad (26)$$

$$\text{s.t. } g_k(\mathbf{z}) \leq 1, k = 1, 2, \dots, p, \quad (27)$$

$$\mathbf{z} \in \mathbf{R}^m,$$

where $g_k(\mathbf{z}), k = 0, 1, \dots, p$ is given by Equation (25), and can be shown to be convex in $\mathbf{z}$.

Composite Geometric Programming (CGP)

Finally, we describe a further generalization of Program QGP described in the previous

section; this was formalized by Jefferson *et al.* [18]. Consider Equation (25) once again, but with the quadratic form in the exponentiation of each term replaced with some *general* convex function:

$$g_k(\mathbf{z}) = \sum_{i \in [k]} \exp(\chi_i + \mathbf{a}^i \mathbf{z} + h_i(\mathbf{z})); \quad k = 0, 1, \dots, p. \quad (28)$$

The optimization problem in composite geometric programming (CGP) is the same as the one described in Equations (26) and (27), but with $g_k(\mathbf{z})$ now described via Equation (28). Since each function in Equation (28) is still convex, it is clear that CGP is also a convex program but is more general than QGP.

COMPUTATIONAL ISSUES AND ALGORITHMS

While posynomial GPs are convex programs and traditional NLP methods can be applied to solve these, the special structure of the primal and (especially) the dual program, and the relationships between the optimum solutions to the two have led to the development of a number of computational algorithms designed specifically for geometric programs. It is not particularly surprising that such algorithms are more robust and also tend to perform better; indeed, a computational study done by Dembo [19] confirmed this fact. These structures are examined further in the section titled “Posynomial Structure and Primal–Dual Relationships at the Optimum.”

Broadly speaking, algorithms for GP may be classified into two categories: posynomial GP methods and signomial GP methods. Solution techniques for posynomial programs may be categorized as either primal-based algorithms that directly solve the original, completely nonlinear problem, or dual-based algorithms that solve the equivalent linearly constrained dual. There are two schools of thought on which approach is superior, and in order to understand the pros and cons of each it is worth examining the structure of the dual problem given by expressions (4–7) in the section titled “The Primal–Dual Pair in GP.”

Posynomial Structure and Primal–Dual Relationships at the Optimum

First, note that the dual has a nonlinear objective but is linearly constrained. Second, since the log function is monotone increasing in its argument one could choose to replace the dual objective with its logarithm and maximize the latter instead. There are some advantages in doing so. If the λ_k are eliminated by substituting from Equations (5) and (6) into Equation (4), the resulting log-dual objective is a concave function to be maximized over a polyhedral set, which is a well-studied convex program for the solution of which numerous efficient algorithms have been developed. Alternatively, if one were to retain the λ_k and work with the log-dual objective, the resulting function—although no longer concave over its domain—is separable in the δ_i and λ_k variables. Again, this is a structural feature that is attractive from a computational perspective.

Unfortunately, there are also certain disadvantages associated with dual-based methods. These problems arise when one or more primal constraints are inactive at the optimum. First, if primal constraint k is inactive at the optimum, then λ_k is equal to zero at the dual optimum. As noted in the section titled “The Primal–Dual Pair in GP” λ_k is equivalent to the Lagrange multiplier for constraint k , so that this result follows from the complementary slackness property. Then Equation (5) and nonnegativity imply that each $\delta_i, i \in [k]$ must also be equal to zero. This causes several difficulties from a computational standpoint. First, terms such as $(c_i/\delta_i)^{\delta_i}$ and $\lambda_k^{\lambda_k}$ (or $\delta_i \ln \delta_i$ and $\lambda_k \ln \lambda_k$ if the logarithm of the dual objective is being used) are now undefined. However, in the limit these quantities approach 1 (or 0 with the log-dual objective) and may therefore be explicitly defined as being equal to 1 (or 0) when δ_i and λ_k are zero; in a computer code special provisions need to be made for this. Typically, this is done by forcing the entire “block” of δ_i values to zero when several of them fall below some small tolerance. Yet another problem is that the objective function is also nondifferentiable now; this causes difficulties for any algorithm that requires the computation of gradients. To overcome these

problems, several procedures have been proposed in the literature. Unfortunately, all of them call for some relatively arbitrary rule to decide on how to address points where the dual variables approach zero. There is always some degree of ambiguity in their implementation (e.g., what is meant by “close” to zero?), and very small tolerances can often lead to numerical problems.

A second (albeit, less common) problem with slack primal constraints lies in the recovery of the primal solution vector ($\mathbf{t}$) from the optimal dual solution vector (λ, δ). With slack constraints at the optimum, there may occasionally be insufficient information to recover the primal solution vector. In order to see this, recall the system of equations representing the relationship between the primal and dual solutions at their respective optima as given by Equation (13) as part of Theorem 3. Typically, we take logarithms on both sides and obtain a system that is linear in the logarithm of t_i and can thus be easily solved to recover the optimal primal vector from the optimal dual one. However, if constraint k is inactive, then λ_k and $\delta_i, i \in [k]$ are all equal to zero, so that the equations in Equation (13) deriving from terms in k are of no use in that they provide no information! It might thus happen that when we consider all the useful equations, that is, corresponding to each term in the objective, and to each term from a constraint for which the corresponding value of λ_k is strictly *positive* at the dual optimum, we might have a system that is underdetermined and therefore impossible to solve for the optimal value of $\mathbf{t}$.

In such situations, one must resort to solving what are referred to as *subsidiary problems* in order to recover the primal vector $\mathbf{t}^*$. In fact, it might actually be necessary to solve a *sequence* of several subsidiary problems in order to obtain a primal solution. Each subsidiary problem is a separate GP problem and has the role of activating one constraint, so that the values of the dual variables corresponding to terms in the constraint are no longer zero and thus additional information becomes available in the form of Equation (13) for recovering the values of the primal variables. It was shown by

Duffin *et al.* [3] that as constraints are activated one by one and additional information becomes available, eventually one will have sufficient information to recover a complete primal solution. A further discussion of subsidiary problems is provided by Rajgopal and Bricker [20]. General issues on dual to primal conversion in posynomial GP are also discussed in detail by Dembo [21].

In summary, while the dual is more attractive on account of its structure, it also presents numerous computational problems in the presence of slack primal constraints at the optimum. This has caused some researchers to abandon the linear structure and address the primal directly, where of course, one has to contend with a highly non-linear problem. Nevertheless, the primal has several characteristics that are conducive to good algorithms [19]. Derivatives of any order for the objective and constraints are always available and relatively inexpensive to compute once the value of the posynomial is known. Also, if the problem is restated as Program P_z (Eqs (8)–(10)), we have a convex program for which very good algorithms exist. One important issue to account for is the fact that strict positivity of the decision variables could lead to a noncompact feasible set and special provisions must be made to ensure that no variable goes to zero.

Algorithms

During the first 10 or so years after its inception, a number of new algorithms were developed specifically for GP. Since then the development has been somewhat slower, although new techniques continue to appear in the literature every now and then. While general-purpose algorithms for NLP can of course be applied to GP, algorithms developed specifically around the GP structure perform better [19]. Other articles that summarize and compare various computational strategies and algorithms include Dinkel *et al.* [22], Dinkel *et al.* [23], Rijckaert and Martens [24], and Sarma *et al.* [25].

We now provide a brief overview of some of the approaches used. Many good algorithms employed a procedure called *condensation*

[26,27]. In this approach, a posynomial $g(\mathbf{t})$ is approximated by a monomial (a posynomial with one term) that is denoted by $g(\mathbf{t}, \tilde{\mathbf{t}})$ and is formed at some fixed $\tilde{\mathbf{t}} > \mathbf{0}$. Let us denote the i th term of the posynomial by $u_i = c_i \prod_{j=1}^m t_j^{a_{ij}}$ and define $\varepsilon_i(\tilde{\mathbf{t}}) = u_i/g(\tilde{\mathbf{t}})$; clearly ε_i are non-negative and sum to 1. Then the condensation at the point $\tilde{\mathbf{t}}$ is given by

$$g(\mathbf{t}, \tilde{\mathbf{t}}) = c(\tilde{\mathbf{t}}) \prod_{j=1}^m (t_j)^{a_j(\tilde{\mathbf{t}})}, \quad (29)$$

where

$$a_j(\tilde{\mathbf{t}}) = \sum_i \varepsilon_i(\tilde{\mathbf{t}}) a_{ij} \quad \text{and} \\ c(\tilde{\mathbf{t}}) = \prod_i \left(\frac{c_i}{\varepsilon_i(\tilde{\mathbf{t}})} \right)^{\varepsilon_i(\tilde{\mathbf{t}})}. \quad (30)$$

It is easily shown that the condensation is equal to the original posynomial at the point of condensation, and as a result of the arithmetic-geometric mean inequality, any point that is feasible in the original constraint using this posynomial is also feasible for the condensation. Thus condensation provides us with an outer approximation of the feasible region. This leads to the natural idea of a cutting plane because it also allows us to linearize the problem: we can take the logarithm of $g(\mathbf{t}, \tilde{\mathbf{t}})$ and obtain a linear function in the log of the design variables.

Looking at the primal problem, Rijckaert and Martens [28] took the route of iteratively solving the Kuhn–Tucker optimality conditions using condensation and a Newton–Raphson type approach. Another algorithm that deals directly with the primal and also uses condensation with a cutting plane approach was developed by Dembo [29] and is still considered one of the better algorithms for GP. Dembo also showed how this could be adapted to solve signomial programs by solving a sequence of GP approximations. Dinkel *et al.* [30] also used condensation and linearization and their procedure is actually quite similar to the method in Avriel *et al.* [29]. Gochet and Smeers [31] used a cutting plane approach, but solve Program P_z (the convex version of the primal). More recently, Maranas and Floudas [32] described another

algorithm that is used to find the global optimum for a signomial problem. Their method works by expressing the objective and constraints as differences of two convex functions and then using a branch and bound type approach. There have been several new interior point methods, for example, Bricker and Chang [33] and Kortanek *et al.* [34]. The latter in particular is probably one of the best algorithms currently available and the authors report being able to efficiently solve very large problems.

Coming to the dual problem an early approach by Beck and Ecker [35] was to use a variation of the convex simplex method that allowed for blocks of variables to go to zero simultaneously in order to counter problems with differentiability. Blau and Wilde [36] developed another early method that solves for the Kuhn–Tucker optimality conditions, but for the dual problem; it can also handle signomial problems. Kochenberger *et al.* [37] describe a method where they use separable programming for a version with the log-dual objective. Kortanek and No [38] used an interior point approach with affine scaling on the dual. More recently, Rajgopal and Bricker [39] describe a method that is based on a reformulation of the primal as a semi-infinite linear program [40] where the problem is restated with a finite number of variables but infinitely many constraints. Essentially, these constraints may be viewed as linear supports for the original nonlinear constraints. Rather than solve this problem, they use its dual (which is a so-called *generalized linear program*), and implement a column generation approach to solve it. The authors report that their method avoids all of the computational problems typically associated with dual-based algorithms.

Finally, it is worth mentioning that in addition to the open availability of codes for several GP algorithms [34,39] there are also implementations on MATLAB that are available to solve GP problems [41,42].

SUMMARY

GP is a mature area within the broader area of NLP. Most of the important theoretical results relevant to GP as well as

major generalizations and extensions were developed during roughly a 15-year span of time starting in the mid-1960s. Two comprehensive sources that capture the important work during that period in terms of theory, algorithms, and applications are the bibliographic note by Rijckaert and Martens [43] and the review article by Ecker [44].

While the pace of development since then has been slower, extensions and new insights, as well as new applications and algorithms, continue to be reported regularly in the literature. As examples of later extensions, Choi and Bricker [45] have an interesting article that extends GP to the case where some of the variables are discrete. Uncertainty in GP has also attracted the attention of some researchers; for example, Jagannathan [46] examines a stochastic GP with recourse, while Liu [47] considers the case where coefficients and exponents are assumed to lie within their own intervals as opposed to taking on fixed values. Cao [7,48] combines GP with fuzzy mathematics to develop the area of fuzzy GP. Recently Hsiung *et al.* [49] incorporate parameter uncertainty and develop a method for addressing robust GP where the optimization is over the worst-case scenario. Another interesting theoretical article by Peterson [50] attempts to relate GP duality to regular Lagrangian duality and duality for parametric programming (in the form of parameter perturbations).

Development of new applications has possibly been the biggest area of emphasis within GP over the last 10–15 years. There is in some circles a misconception that GP is a narrow subset of NLP with a rigid format requirement that precludes most optimization problems from being addressed by this technique. As the literature indicates, this is not true; new applications for GP continue to be regularly reported beyond the initial ones in engineering design and the multitude of areas cited in Rijckaert and Martens [43]. These applications span many fields including mathematics and statistics [51–53], manufacturing [54–57], production and inventory control [58–61], and most recently, electronics and computer engineering [62–65]. With the availability of good software as well, this trend will likely continue in the near future.

REFERENCES

1. Zener C. A mathematical aid in optimizing engineering design problems. *Proc Natl Acad Sci U S A* 1961;47:537–539.
2. Zener C. A further mathematical aid in optimizing engineering designs. *Proc Natl Acad Sci U S A* 1961;48:518–522.
3. Duffin RJ, Peterson EL, Zener C. *Geometric programming—theory and applications*. New York: John Wiley & Sons, Inc.; 1967.
4. Zener C. *Engineering design by geometric programming*. New York: John Wiley & Sons, Inc.; 1971.
5. Beightler C, Phillips DT. *Applied geometric programming*. New York: John Wiley & Sons, Inc.; 1976.
6. Avriel M. *Advances in geometric programming*. New York: Plenum Press; 1980.
7. Cao BY. *Fuzzy geometric programming*. Norwell (MA): Kluwer Academic Publishers; 2002.
8. Chiang M. *Geometric programming for communication systems*. Hanover (MA): Now Publishers Inc.; 2005.
9. Rajgopal J. An alternative approach to the refined duality theory of geometric programming. *J Math Anal Appl* 1992;167:266–288.
10. Passy U, Wilde DJ. Generalized polynomial optimization. *SIAM J Appl Math* 1967;15:1344–1356.
11. Peterson EL. Optimality conditions in generalized geometric programming. *J Optim Theory Appl* 1978;26:3–13.
12. Peterson EL. Saddle points and duality in generalized geometric programming. *J Optim Theory Appl* 1978;26:15–41.
13. Rockafellar RT. *Convex analysis*. Princeton (NJ): Princeton University Press; 1996.
14. Boyd S, *et al.* A tutorial on geometric programming. *Optim Eng* 2007;8:67–127.
15. Lidor G, Wilde DJ. Transcendental geometric programs. *J Optim Theory Appl* 1978;26:77–96.
16. Hough CL Jr, Goforth RE. Quadratic posynomials: an extension of posynomial geometric programming. *AIIE Trans* 1981;13:47–54.
17. Jefferson TR, Scott CH. Quadratic geometric programming with application to machining economics. *Math Program* 1985;31:137–152.
18. Jefferson TR, Wang YP, Scott CH. Composite geometric programming. *J Optim Theory Appl* 1990;64(1):101–118.
19. Dembo RS. Current state of the art of algorithms and computer software for geometric programming. *J Optim Theory Appl* 1978;26:149–183.
20. Rajgopal J, Bricker DL. On subsidiary problems in geometric programming. *Eur J Oper Res* 1992;63:102–113.
21. Dembo RS. Dual to primal conversion in geometric programming. *J Optim Theory Appl* 1978;26:243–252.
22. Dinkel JJ, Kochenberger GA, McCarl B. A computational study of methods for solving polynomial geometric programs. *J Optim Theory Appl* 1976;1:233–259.
23. Dinkel JJ, Elliott WH, Kochenberger GA. Computational aspects of cutting plane methods in geometric programming. *Math Program* 1977;13:200–220.
24. Rijckaert MJ, Martens XM. Comparison of generalized geometric programming algorithms. *J Optim Theory Appl* 1978;26:205–242.
25. Sarma PVLN, *et al.* A comparison of computational strategies for geometric programs. *J Optim Theory Appl* 1978;26:185–203.
26. Duffin RJ. Linearizing geometric programs. *SIAM Rev* 1970;12:211–227.
27. Dinkel JJ, Kochenberger GA. Some remarks on condensation methods for geometric programs. *Math Program* 1979;17:109–113.
28. Rijckaert MJ, Martens XM. A condensation method for generalized geometric programming algorithms. *Math Program* 1976;11:89–93.
29. Avriel M, Dembo R, Passy U. Solution of generalized geometric programs. *Int J Numer Methods Eng* 1975;9:149–169.
30. Dinkel JJ, Elliott WH, Kochenberger GA. A linear programming approach to geometric programs. *Nav Res Log Q* 1978;25:39–53.
31. Gochet W, Smeers Y. On the use of linear programs to solve prototype geometric programs. *Cah Centre Etud Rech Opér* 1974;16(1). 16.
32. Maranas CD, Floudas CA. Global optimization in generalized geometric programming. *Comput Chem Eng* 1997;21:351–369.
33. Bricker DL, Chang HY. A path-following algorithm for posynomial geometric programming 25th Anniversary Geometric Programming Conference; 1992 Aug 3–5; Colorado School of Mines Golden, CO. 1992.
34. Kortanek KO, Xu X, Ye Y. An infeasible interior-point algorithm for solving primal

- and dual geometric programs. *Math Program* 1997;76:155–181.
35. Beck PA, Ecker JG. Some computational experience with a modified convex simplex algorithm for geometric programming. *J Optim Theory Appl* 1975;15:189–202.
 36. Blau GE, Wilde DJ. A Lagrangean algorithm for equality constrained generalized polynomial optimization. *AIChE J* 1971;17:235–240.
 37. Kochenberger GA, Woolsey RED, McCarl BA. On the solution of geometric programs via separable programming. *Oper Res Q* 1973;24:285–296.
 38. Kortanek KO, No H. A second order affine scaling algorithm for the geometric programming dual with logarithmic barrier. *Optimization* 1992;23:303–322.
 39. Rajgopal J, Bricker DL. Solving posynomial geometric programming problems via generalized linear programming. *Comput Optim Appl* 2002;21:95–109.
 40. Rajgopal J, Bricker DL. Posynomial geometric programming as a special case of semi-infinite linear programming. *J Optim Theory Appl* 1990;66:455–475.
 41. Andersen ED. 1998; The MOSEK optimization toolbox for MATLAB. Available at http://www.mosek.com/fileadmin/products/6_0/tools/doc/pdf/toolbox.pdf. Accessed 2010 Sept 20.
 42. Mutapcic A, *et al.* 2006. GGPLAB version 1.00; A Matlab toolbox for geometric programming. Available at <http://www.stanford.edu/~boyd/ggplab>. Accessed 2010 Sept 20.
 43. Rijckaert MJ, Martens XM. Bibliographical note on geometric programming. *J Optim Theory Appl* 1978;26:325–337.
 44. Ecker JG. Geometric programming: methods, computations and applications. *SIAM Rev* 1980;22:338–362.
 45. Choi JC, Bricker DL. Geometric programming with several discrete variables: algorithms employing generalized Benders' decomposition. *Eng Optim* 1995;25:201–212.
 46. Jagannathan R. A stochastic geometric programming problem with multiplicative recourse. *Oper Res Lett* 1990;9:99–104.
 47. Liu S-T. Posynomial geometric programming with interval exponents and coefficients. *Eur J Oper Res* 2008;186:17–27.
 48. Cao B-Y. Fuzzy geometric programming. *J Fuzzy Sets Syst* 1993;53:135–153.
 49. Hsiung K-L, Kim S-J, Boyd S. Tractable approximate robust geometric programming. *Optim Eng* 2008;9:95–118.
 50. Peterson EL. The fundamental relations between geometric programming duality, parametric programming duality, and ordinary Lagrangian duality. *Ann Oper Res* 2001;105:109–153.
 51. Mazumdar M, Jefferson TR. Maximum likelihood estimates for multinomial probabilities via geometric programming. *Biometrika* 1983;70:257–261.
 52. Bricker DL, Kortanek KO, Xu L. Maximum likelihood estimates with order restrictions on probabilities and odds ratios. *J Appl Math Decis Sci* 1997;1:53–65.
 53. Cheng H, Fang S-C, Lavery JE. Univariate cubic L_1 splines—a geometric programming approach. *Math Methods Oper Res* 2002;56:197–229.
 54. Gupta R, Batra JL, Lal GK. Profit rate maximization in multipass turning with constraints: a geometric programming approach. *Int J Prod Res* 1994;32:1557–1569.
 55. Choi JC, Bricker DL. Effectiveness of a geometric programming algorithm for optimization of machining economics models. *Comput Oper Res* 1996;23:957–961.
 56. Sönmez A, *et al.* Dynamic optimization of multipass milling operations via geometric programming. *Int J Mach Tools Manuf* 1999;39:297–320.
 57. Liu S-T. Fuzzy geometric programming approach to a fuzzy machining economics model. *Int J Prod Res* 2004;42:3253–3269.
 58. Jung H, Klein CM. Optimal inventory policies under decreasing cost functions via geometric programming. *Eur J Oper Res* 2001;132:628–642.
 59. Lee WJ. Determining order quantity and selling price by geometric programming: optimal solution, bounds, and sensitivity. *Decis Sci* 2007;24:76–87.
 60. Mandal NK, Roy TK, Maiti M. Inventory model of deteriorated items with a constraint: a geometric programming approach. *Eur J Oper Res* 2006;173:199–210.
 61. Leung K-NF. A generalized geometric programming solution to “An economic production quantity model with flexibility and reliability considerations.” *Eur J Oper Res* 2007;176:240–251; Mandal P, Visvanathan V. CMOS op-amp sizing using a geometric programming formulation. *IEEE Trans Comput Aided Des Integr Circuits Syst* 2001;20:22–38.
 62. Li X, *et al.* Robust analog/RF circuit design with projection-based posynomial modeling.

- International Conference on Computer-Aided Design (ICCAD'04) 2004 Nov 7–11; San Jose (CA). 2004. pp. 855–862.
63. Singh J, *et al.* Robust gate sizing by geometric programming. Proceedings of the 42nd Annual Conference on Design Automation; 2005 Jun 13–17; Anaheim, CA. 2005. pp. 315–320.
64. Boyd SP, *et al.* Digital circuit optimization via geometric programming. *Oper Res* 2005;53:899–932.
65. Chiang M, *et al.* Power control by geometric programming. *IEEE Trans Wirel Commun* 2007;6:2640–2651.

GERMAN OPERATIONS RESEARCH SOCIETY (GOR) (GESELLSCHAFT FÜR OPERATIONS RESEARCH)

HORST W. HAMACHER
Department of Mathematics,
University of Kaiserslautern,
Kaiserslautern, Germany

In Germany, operations research has a long tradition, manifested since the 1950 s in various organization forms, and—until the German reunification in 1990—in two different countries. Excellent surveys on the history of OR in the two Germanies can be found in the articles of Müller-Merbach and Lassmann (OR News 36 (2009), pp. 6–12). Part of this history is the existence of several predecessor OR societies which, in 1998, led to the foundation of the GOR as the single OR society. Since then, six presidents have been serving: Peter Kleinschmidt (1998), Ulrich Derigs (1999–2000), Uwe Zimmermann (2001–2002), Gerhard Wäscher (2003–2006), Thomas Spengler (2007–2008), and Horst W. Hamacher (2009–2010). The GOR with its 1100 individual and corporate members is the second largest OR society in Europe. It is a member of IFORS and EURO and has recently, in July 2009, been organizer of the EURO conference in Bonn with some 2400 participants.

One of the main features and strengths of the GOR is its—never ending—effort to bring together various streams, in particular OR practitioners and basic researchers.

The first important tool in this effort is our publication instruments. The *OR News* appearing thrice per year and edited since its beginning by Joachim Minnemann is—similar to its US counterpart ORMS Today—an excellent source of information

on OR events in the world and on OR articles which are accessible to laymen as well as specialists. The professional journals, *Operations Research Spectrum* and *Mathematical Methods of Operations Research* (currently edited by Stefan Minner and Alexander Martin, respectively) are well established in the international OR community. The former journal is devoted to the practice of OR while the latter focuses on the development of new mathematical methods within OR. An excellent opportunity to meet and exchange ideas is the *annual GOR conference*, which is always taking place in the first week of September. Every 4 years this conference is organized as a trinational event with our sister organizations in Austria and Switzerland. Refereed extended abstracts of selected papers presented at these meeting are published in the Springer Series *Operations Research Proceedings*.

The thirteen working groups compose another invaluable instrument to unify practice and theory within the GOR. In general, the groups meet twice a year in workshops to exchange ideas, discuss problems posed by industry and administrations, or to present new results. These meetings are in general open to nonmembers of the GOR. Each issue of the OR News contains an update of the current activities of the GOR working groups.

In order to recognize excellence, five types of GOR awards have been established: Bachelor's, Diploma/Master's, Ph.D., Scientific, and Enterprise Awards. With the exception of the first one, they are awarded during the annual GOR conference and are generously supported by sponsors who at present include INFORM GmbH, GAMS, BASF, and Springer Publishers.

GOMORY CUTS

RICARDO FUKASAWA
Department of Combinatorics and
Optimization, University of
Waterloo, Waterloo, Ontario,
Canada

Gomory fractional (GF) cuts were introduced in 1958 with an announcement by Ralph E. Gomory [1] of an “algorithm for integer solutions to linear programs (LPs)” (a full version of the algorithm appears in Ref. 2), which gave a finite cutting-plane algorithm for pure integer programming. In a later paper [3], he also introduced the so called *Gomory mixed-integer (GMI)* cuts. Since then, the field of (mixed-) integer programming has developed immensely, but throughout the years and up to this day, Gomory’s fractional cuts, the closely related Chvátal–Gomory (CG) cuts, and the mixed-integer cuts of Gomory play a significant role in much of the research that is done in the field. The purpose of this paper is to do a brief survey of the research that has been carried out around Gomory cuts throughout the years. The interested reader is also referred to some surveys on general cutting planes [4–6], as well as related articles about cutting planes (in the section titled “Cutting Planes”) in this encyclopedia.

The organization of this paper is as follows. In what remains of this section, we define what are the GF cuts, the closely related CG cuts, and the GMI cuts. We then introduce mixed-integer rounding (MIR) cuts and explore their relationship with GMI cuts. The section titled “Closures and Rank” is devoted to defining and reviewing some of the results related to rank and closure for each of the classes of cuts. The section titled “Special Cases, Variations, Strengthening, and Others” surveys some variations, special cases, and relations that Gomory cuts have with other types of cuts. The section titled “Implementation and Computational Aspects” is devoted to computational issues, while the section titled “Multirow Gomory

Cuts” talks about the recent interest in the so-called *multirow Gomory cuts*.

Throughout this paper, we refer to the following general mixed-integer sets:

$$P_I = \{(x, z) \in \mathbb{Z}_+^n \times \mathbb{R}_+^p : Ax + Dz \leq b\} \quad (1)$$

and

$$Q_I = \{(x, z) \in \mathbb{Z}^n \times \mathbb{R}^p : Ax + Dz \leq b\}. \quad (2)$$

GOMORY’S FRACTIONAL CUTS

Gomory’s fractional cuts are defined for the pure integer case and when all variables are assumed to be nonnegative, that is, they are cuts that are valid for the following set (P_I with $p = 0$):

$$P_I = \{x \in \mathbb{Z}_+^n : Ax \leq b\}. \quad (3)$$

In addition, Gomory’s fractional cuts were derived in Ref. 1 when the constraints were in equality form. Since, in the context of mixed-integer programming, it is assumed that all data is rational, we may assume that A and b are integral and introduce slack variables to obtain $P'_I = \{(x, s) \in \mathbb{Z}_+^{n+m} : Ax + s = b\} := \{z \in \mathbb{Z}_+^{n+m} : Gz = b\}$, where $z = (x, s)$ and $G = (A, I)$.

We can relax P'_I by multiplying all equations by a vector $u \in \mathbb{R}^m$ to obtain the set $\{z \in \mathbb{Z}_+^{n+m} : uGz = ub\}$. We can further relax this set to obtain the set $P_G = \{z \in \mathbb{Z}_+^{n+m} : uGz \equiv ub \pmod{1}\}$, which means that uGz and ub differ by an integer. Moreover, let $f(a) := a - \lfloor a \rfloor$ denote the fractional part of a real number a . Now since P_G is expressed in terms of a modular equation, it can be rewritten as $P_G = \{z \in \mathbb{Z}_+^{n+m} : f(uG)z \equiv f(ub) \pmod{1}\}$, where $f(uG)$ is the vector obtained by taking the fractional part of each component of uG . Notice that $f(uG) \geq 0$ and $z \geq 0$, hence $f(uG)z \geq 0$. Finally, since $f(uG)z \equiv f(ub) \pmod{1}$ and $f(uG)z \geq 0$, then $f(uG)z$ must be equal to

$f(ub) + k$ for some integer $k \geq 0$. This argument shows that the following inequality is valid for P_G :

$$f(uG)z \geq f(ub). \quad (4)$$

If we substitute in Equation (4) the identities $f(uG) = uG - \lfloor uG \rfloor$, $f(ub) = ub - \lfloor ub \rfloor$, and $s = b - Ax$, we obtain the *GF cut* on the original x variables:

$$\lfloor uA \rfloor x - \lfloor u \rfloor Ax \leq \lfloor ub \rfloor - \lfloor u \rfloor b. \quad (5)$$

We would like to point out that GF cuts are defined here by using any multiplier $u \in \mathbb{R}^m$, but Gomory's cutting-plane algorithm uses a specific set of multipliers u that are associated with basic feasible solutions of $P' = \{(x, s) \in \mathbb{R}_+^{n+m} : Ax + s = b\}$. For more details on this distinction and its implications, see Ref. 7.

CHVÁTAL-GOMORY CUTS

CG cuts were defined by Chvátal [8] and are also defined for the pure integer case (that is, when $p = 0$). However, unlike in the case of the GF cuts, CG cuts are valid for the following set

$$Q_I = \{x \in \mathbb{Z}^n : Ax \leq b\}, \quad (6)$$

that is, even when there is no nonnegativity assumption on the variables. Similar to Gomory's fractional cuts, we can aggregate the constraints in Q_I by using a multiplier $u \in \mathbb{R}_+^m$ to obtain the relaxation $\{x \in \mathbb{Z}^n : uAx \leq ub\}$. Now, if $uA \in \mathbb{Z}^n$, then $uAx \in \mathbb{Z}$ and hence the following CG cut is valid for Q_I .

$$uAx \leq \lfloor ub \rfloor. \quad (7)$$

If our original feasible set is described as

$$P_I = \{x \in \mathbb{Z}_+^n : Ax \leq b\}, \quad (8)$$

that is, with a nonnegativity assumption on the variables, a slightly different argument can be used to derive the CG cut.

Note that if $Q_I = \{x \in \mathbb{Z}^n : Ax \leq b ; -x \leq 0\}$, that is, if the nonnegativity constraints

are included in the constraint matrix, rather than explicitly imposed, we have that any cut of the form (7) is

$$(uA - \lambda)x \leq \lfloor ub \rfloor,$$

where $(u, \lambda) \in \mathbb{R}_+^{m+n}$ are the CG multipliers.

In this case, we do not need to require that $uA \in \mathbb{Z}^n$, because, since $x \geq 0$, $\lfloor uA \rfloor x \leq uAx \leq ub$ and, since $\lfloor uA \rfloor x \in \mathbb{Z}$, the CG cut

$$\lfloor uA \rfloor x \leq \lfloor ub \rfloor \quad (9)$$

is valid for P_I . Equivalently, we require $(uA - \lambda) \in \mathbb{Z}^n$ and any nondominated CG cut will have $\lambda = uA - \lfloor uA \rfloor$, which leads to Equation (9).

It is easy to see that any CG cut of the form (9) can be obtained as a CG cut of the form (7) when the nonnegativity constraints are not explicitly imposed, but are present in the constraint matrix. Conversely, under the nonnegativity assumption, any cut of the form (7) can either be obtained as a cut of the form (9) or is dominated by one.

The name CG cuts comes from the fact that, under the nonnegativity assumption on the variables, any cut of the form (5) can be obtained as a cut of the form (9) and vice versa (see Refs 7 and 9 for more details). This is the reason why these two types of cuts are often said to be equivalent. However, note that the cut (7) cannot always be obtained as a cut of the form (5), since the latter requires the nonnegativity assumption.

GOMORY MIXED-INTEGER CUTS

The arguments for the derivation of CG cuts and GF cuts rely on the fact that all variables are required to be integral and therefore the same arguments cannot be directly applied when we have a mixed-integer set

$$P_I = \{(x, z) \in \mathbb{Z}_+^n \times \mathbb{R}_+^p : Ax + Dz \leq b\}. \quad (10)$$

By adding slack variables, we obtain the following set in equality form

$$P'_I = \{(x, y) \in \mathbb{Z}_+^n \times \mathbb{R}_+^{p+m} : Ax + Gy = b\}, \quad (11)$$

where $G = (D, I)$ and $y = (z, s)$. We again multiply the equations by a vector $u \in \mathbb{R}^m$ to obtain the set

$$P'_I(u) = \{(x, y) \in \mathbb{Z}_+^n \times \mathbb{R}_+^{p+m} : \bar{a}x + \bar{g}y = \bar{b}\}, \quad (12)$$

where $\bar{a} = uA$, $\bar{g} = uG$, and $\bar{b} = ub$ and we assume that $f(\bar{b}) > 0$. Let $f_i = f(\bar{a}_i)$ and $f_0 = f(\bar{b})$. We are now ready to derive the GMI cut. Notice that $P'_I(u)$ can be rewritten as

$$\begin{aligned} & \sum_{i: f_i \leq f_0} f_i x_i + \sum_{i: f_i > f_0} (f_i - 1)x_i + \bar{g}y \\ &= \left(\lfloor \bar{b} \rfloor - \sum_{i=1}^n \lfloor \bar{a}_i \rfloor x_i - \sum_{i: f_i > f_0} x_i \right) + f_0. \end{aligned}$$

Since $k := \left(\lfloor \bar{b} \rfloor - \sum_{i=1}^n \lfloor \bar{a}_i \rfloor x_i - \sum_{i: f_i > f_0} x_i \right) \in \mathbb{Z}$, then either $k \leq -1$, in which case

$$- \sum_{i: f_i \leq f_0} \frac{f_i}{1-f_0} x_i + \sum_{i: f_i > f_0} \frac{1-f_i}{1-f_0} x_i - \frac{\bar{g}}{1-f_0} y \geq 1 \quad (13)$$

or $k \geq 0$, in which case

$$\sum_{i: f_i \leq f_0} \frac{f_i}{f_0} x_i + \sum_{i: f_i > f_0} \frac{f_i - 1}{f_0} x_i + \frac{\bar{g}}{f_0} y \geq 1. \quad (14)$$

Since both these cuts are of the form $a^1 w \geq 1$ and $a^2 w \geq 1$, and $w \geq 0$, then the cut $\sum_j \max\{a_j^1, a_j^2\} w_j \geq 1$ is valid for $P'_I(u)$ and hence valid for P'_I .

Therefore, the following inequality, called the *GMI cut*, is valid for P'_I

$$\begin{aligned} & \sum_{i: f_i \leq f_0} \frac{f_i}{f_0} x_i + \sum_{i: f_i > f_0} \frac{1-f_i}{1-f_0} x_i + \sum_{j: \bar{g}_j \geq 0} \frac{\bar{g}_j}{f_0} y_j \\ & - \sum_{j: \bar{g}_j < 0} \frac{\bar{g}_j}{1-f_0} y_j \geq 1 \end{aligned} \quad (15)$$

(for this and other derivations of this cut see Refs 2, 4, 9).

It is worth noting that, in the pure integer case, the GMI cut is still valid and is of the form

$$\sum_{i: f_i \leq f_0} \frac{f_i}{f_0} x_i + \sum_{i: f_i > f_0} \frac{1-f_i}{1-f_0} x_i \geq 1. \quad (16)$$

We can compare that cut with the GF cut

$$\sum_{i=1}^n \frac{f_i}{f_0} x_i \geq 1$$

and notice that when $f_i > f_0$, we have that $\frac{1-f_i}{1-f_0} < \frac{f_i}{f_0}$. Therefore, the GMI cut dominates the GF cut in the pure integer case.

MIXED-INTEGER ROUNDING CUTS

In this section, we introduce MIR cuts. Notice that this class of cuts is the topic of another completely independent article of this encyclopedia. However, MIR cuts are so closely related to GMI cuts that it is worth devoting some attention to this relationship in more detail. First, we need to point out that there are multiple cuts that are called *MIR* cuts. We use here the definition from Ref. 10. See also Refs 9, 11–13 for other references on MIR cuts. Consider the general mixed-integer set

$$Q_I = \{(x, z) \in \mathbb{Z}^n \times \mathbb{R}^p : A'x + D'z \leq b'\}. \quad (17)$$

Suppose we multiply the inequalities in Equation (17) by two different vectors $\mu^1, \mu^2 \in \mathbb{R}_+^{n'}$, where $\mu^1 D' = \mu^2 D'$ and $(\mu^2 - \mu^1)A' \in \mathbb{Z}^n$ then the inequality

$$\begin{aligned} & (\mu^2 - \mu^1)A'x \\ & + \frac{\mu^1 A'x + \mu^1 D'z - \mu^1 b'}{1 - ((\mu^2 b' - \mu^1 b') - \lfloor \mu^2 b' - \mu^1 b' \rfloor)} \\ & \leq \lfloor \mu^2 b' - \mu^1 b' \rfloor \end{aligned} \quad (18)$$

is called the *MIR inequality* and is valid for Q_I . Since MIR cuts are the subject of another article titled of this encyclopedia, we will refrain from proving that the MIR inequality is valid for Q_I . We will instead point out that Q_I does not explicitly require the variables to be nonnegative. However, if $A' = [A, -I, \mathbf{0}]^T$, $D' = [D, \mathbf{0}, -I]^T$, and $b' = [b, \mathbf{0}, \mathbf{0}]^T$ (that is, the nonnegativity constraints are included in $A'x + D'z \leq b'$), then $Q_I = P_I$ and the MIR inequalities (Eq. 18) can be derived as *GMI inequalities* (Eq. 15) and vice versa.

To see how, let us assume that the nonnegativity constraints are contained in the inequalities $A'x + D'z \leq b'$. Therefore, the multipliers μ^1 and μ^2 can be written as $\mu^1 = [\mu_o^1, \mu_x^1, \mu_z^1]^T$ and $\mu^2 = [\mu_o^2, \mu_x^2, \mu_z^2]^T$, where μ_o^i is the multiplier for the constraints $Ax + Dz \leq b$, μ_x^i for the $-x \leq 0$ constraints and μ_z^i for the $-z \leq 0$ constraints, for $i = 1, 2$. Notice that even if we vary μ^1 and μ^2 , as long as $\mu^2 - \mu^1$ does not change, the only term that varies in Equation (18) is the numerator of the fraction. Therefore, since $Ax + Dz - b' \leq 0$, and $\mu^1 \geq 0$, if we increase μ^1 and μ^2 by the same amount, we get a weaker inequality. So, we may assume that for every i , either $\mu_o^1 = 0$ or $\mu_o^2 = 0$. So we may define $\mu = \mu^1 - \mu^2 \in \mathbb{R}^{m'}$ with no sign restriction and $\mu^1 = \mu^+$ and $\mu^2 = (-\mu)^+$, with Equation (18) becoming

$$-\mu A'x + (\mu)^+ \frac{Ax + Dz - b'}{1 - (-\mu b' - \lfloor -\mu b' \rfloor)} \leq \lfloor -\mu b' \rfloor. \quad (19)$$

Now, notice that $\lfloor -\mu b' \rfloor = -\lceil \mu b' \rceil$, so we can rewrite Equation (19) as

$$\begin{aligned} &(\mu b' - \lfloor \mu b' \rfloor) \mu A'x + (\mu)^+(b' - A'x - D'z) \\ &\geq (\mu b' - \lfloor \mu b' \rfloor) \lceil \mu b' \rceil. \end{aligned} \quad (20)$$

Moreover, if we expand μ , A' , D' , and b' , we get

$$\begin{aligned} &(\mu_o b - \lfloor \mu_o b \rfloor)(\mu_o A - \mu_x)x + (\mu_o)^+(b - Ax - Dz) \\ &+ (\mu_x)^+x + (\mu_z)^+z \geq (\mu_o b - \lfloor \mu_o b \rfloor) \lceil \mu_o b \rceil. \end{aligned} \quad (21)$$

But recall that $\mu A' \in \mathbb{Z}^n$ and $\mu D' = 0$. This means that $\mu_o A - \mu_x \in \mathbb{Z}^n$ and that $\mu_o D - \mu_z = 0$. Now $\mu_o A - \mu_x \in \mathbb{Z}^n$ implies that $\mu_x = \hat{a} + t$ for some integral vector $t \in \mathbb{Z}^n$, where $\hat{a} = \mu_o A - \lfloor \mu_o A \rfloor$. If $t_i \geq 1$, then $(\mu_x)_i \geq 1$ and hence $(\mu_x)_i^+ = (\mu_x)_i$. So decreasing t_i to 0 decreases the coefficient of x_i . If $t_i < -1$, then $(\mu_x)_i \leq -1$ and hence $(\mu_x)_i^+ = 0$. So the coefficient of x_i can be decreased by increasing t_i . Hence $t_i \in \{0, -1\}$ for any nondominated MIR inequality.

Defining $\mu_o b - \lfloor \mu_o b \rfloor = f_0$, we can rewrite Equation (21) as

$$\begin{aligned} &(\mu_x)^+x - f_0 \mu_x x + f_0 \hat{a} x + f_0 \lfloor \mu_o A \rfloor x + (\mu_o)^+ \\ &\times (b - Ax - Dz) + (\mu_o D)^+ z \geq f_0 \lceil \mu_o b \rceil. \end{aligned} \quad (22)$$

When $t_i = 0$, we have that $(\mu_x)_i = \hat{a}_i$ so the coefficient of x_i in the first three terms of Equation (22) is $\hat{a}_i$. When $t_i = -1$, we have that $(\mu_x)_i = \hat{a}_i - 1$, so the coefficient of x_i in the first three terms of Equation (22) is f_0 . Therefore, it is easy to see that any nondominated MIR inequality will have the form

$$\begin{aligned} &\min\{\mu_o A - \lfloor \mu_o A \rfloor, f_0 \mathbf{1}\}x + f_0 \lfloor \mu_o A \rfloor x + (\mu_o)^+ \\ &\times (b - Ax - Dz) + (\mu_o D)^+ z \geq f_0 \lceil \mu_o b \rceil. \end{aligned} \quad (23)$$

It is not hard to see, after some algebraic manipulations, that the MIR inequality (Eq. 23) is exactly the GMI inequality (Eq. 15) that we obtain if we do as follows:

1. add slack variables s to $Ax + Dz \leq b$;
2. use μ_o as a multiplier for $Ax + Dz + s = b$ and obtain Equation (15);
3. multiply Equation (15) by $f_0(1 - f_0)$, add it to f_0 times $\mu_o Ax + \mu_o Dz + \mu_o s = \mu_o b$ and substitute out the slack variables.

Therefore, under the nonnegativity assumption, MIR inequalities and GMI inequalities are said to be equivalent. However, note that Equation (15) was derived using the nonnegativity of the variables, while Equation (18) was not. In that sense, the relationship between MIR inequalities and GMI inequalities is analogous to the relationship between CG cuts and GF cuts. For more details, see Refs 7, 14, 15.

CLOSURES AND RANK

The concepts of the *elementary closure* and *rank* were introduced by Chvátal [16] for CG cuts and are important topics, which have received much attention since they are theoretical concepts that give an indication of the strength of a family of cuts or alternatively a measure of its complexity. We now define these concepts for Gomory cuts and survey a few of the results that are known for ranks and closures of polyhedra.

For each of the cuts presented in the previous section, let P denote the polyhedron obtained by dropping the integrality requirements on the sets from which the cuts were derived (P_I with $p = 0$ for GF,

Q_I with $p = 0$ for CG , P_I for GMI). Notice that all the cuts presented in the previous section were derived by aggregating the set of linear constraints in P with some multipliers $u \in \mathbb{R}^m$ ($u \in \mathbb{R}_+^m$ in the case of CG cuts). This observation means that, once the multipliers are fixed, the derivation of the cuts is unique and it makes sense to define $GF(u, P)$, $CG(u, P)$, and $GMI(u, P)$ as the Gomory fractional/ CG / GMI cut obtained by using u as multipliers to aggregate the linear constraints in P . This immediately gives rise to families of these inequalities $F_{GF}(P) := \{GF(u, P) : u \in \mathbb{R}^m\}$, $F_{CG}(P) := \{CG(u, P) : u \in \mathbb{R}_+^m\}$, and $F_{GMI}(P) := \{GMI(u, P) : u \in \mathbb{R}^m\}$ obtained by considering all possible values of u .

We can now define

$$P_{CG}^1 = \{x \in P : x \text{ satisfies all inequalities in } F_{CG}(P)\} \quad (24)$$

as the *rank-1 CG -closure* of P , and one can similarly define P_{GF}^1 and P_{GMI}^1 , which are the *rank-1 GF -closure* and *rank-1 GMI closure* of P . The rank-1 closure is also called the *elementary closure* of a given family of cuts derived from P and valid for P_I .

It is also useful to apply the same concept recursively and define for $k \geq 1$:

$$P_{CG}^k = \{x \in P_{CG}^{k-1} : x \text{ satisfies all inequalities in } F_{CG}(P_{CG}^{k-1})\} \quad (25)$$

as the *rank- k CG -closure* where $P_{CG}^0 = P$. Similarly define the *rank- k GF closure* P_{GF}^k and the *rank- k GMI closure* P_{GMI}^k . Note that (25) is not well-defined unless $P_{CG}^1/P_{GF}^1/P_{GMI}^1$ is a polyhedron. This is in fact true and will be dealt with in the subsection “Seperation and Polyhedrality of Closures”. With that, we can also define the CG rank of an inequality being k if it is valid for P_{CG}^k but not for P_{CG}^{k-1} . Finally, we say that the CG rank of a class of inequalities is the maximum CG rank among all inequalities in that class and we say that the rank of a polyhedron P is the highest rank among all inequalities defining $\text{conv}(P \cap \mathbb{Z}^n)$ (we can derive similar definitions for GF / GMI ranks).

Relationships Between Closures

As mentioned earlier, GF cuts are equivalent to CG cuts in the presence of nonnegativity constraints and, therefore, the immediate result is that $P_{CG}^1 = P_{GF}^1$. Similarly, in the mixed-integer case, the rank-1 GMI closure is the same as the rank-1 MIR closure, under the nonnegativity assumption. In addition, if the variables are nonnegative, the rank-1 GMI closure is the same as the rank-1 split closure of P , obtained from another important family of cuts: split cuts [17,18]. Because of this fact, GMI , MIR , and split cuts are considered “equivalent” under the nonnegativity assumption. For more details on MIR and split cuts, see the article title and ***Split Cuts*** in this encyclopedia. As shown before, in the pure integer case, the GMI cut dominates the GF cut, which means that $P_{GMI}^1 \subseteq P_{CG}^1$. In addition, the inclusion is strict, so not only does the individual GMI cut derived using a specific set of multipliers dominate the CG cut using the same multiplier but the class of rank- k GMI cuts is also “stronger” than the class of rank- k CG cuts. In fact, when compared to several different classes of cuts, the rank-1 GMI closure is stronger than most of the other rank-1 closures. For more details on these and other related results see the detailed study on the relationship between several elementary closures by Cornuéjols and Li [7].

Separation and Polyhedrality of Closures

Note that the way rank-1 CG / GMI closures are defined, they consist of infinitely many inequalities. Therefore, one natural question is whether there exists a finite set of those inequalities that suffice to describe P_{CG}^1 , P_{GF}^1 , and P_{GMI}^1 .

The answer is positive in both cases. In the pure integer case, P_{CG}^1 (and hence P_{GF}^1) was shown to be polyhedral by Chvátal [16] and Schrijver [19,20]. In addition, Bockmayr and Eisenbrand [21] show that, in fixed dimension, the number of inequalities needed to describe P_{CG}^1 is polynomial in the size of the input. In the pure and mixed-integer case, P_{GMI}^1 was shown to be polyhedral by Cook *et al.* [18], in terms of the split-closure (equivalent to P_{GMI}^1 under nonnegativity).

Alternative proofs of this fact were later given by Andersen *et al.* [22,23], Vielma [24], and Dash *et al.* [15].

Now the number of inequalities required to describe P_{CG}^1 , P_{GF}^1 , and P_{GMI}^1 is usually very large and, therefore, one natural question that arises is to solve the separation problem, that is, given a point $x^* \in P$, either show that $x^* \in P_{CG}^1/P_{GF}^1/P_{GMI}^1$ or provide a hyperplane separating x^* from $P_{CG}^1/P_{GF}^1/P_{GMI}^1$. Unfortunately, the separation problem cannot be solved in polynomial time unless $P = NP$: Separation over P_{CG}^1 is shown to be strongly NP-complete by Eisenbrand [25]. Even if P is contained in the nonnegative orthant, separation over P_{CG}^1 (equivalently over P_{GF}^1), and over P_{GMI}^1 , was shown to be strongly NP-complete by Caprara and Letchford [26]. In spite of that, in the case of set covering, Bienstock and Zückerberg [27] have recently shown that one can approximate the rank- k CG closure to an arbitrary fixed precision in polynomial time (for a fixed k).

Bounds on the Rank

Another interesting question that has received significant research attention relates to obtaining bounds on the rank of polyhedra. These bounds give an idea on how much does each family of cutting planes approximate the convex hull of integer solutions.

For pure integer programs (IPs), it is well known [8,19] that all valid inequalities have finite CG rank. However, the rank of certain inequalities can be arbitrarily large, even in dimension 2 (see Chvátal [8]). This implies that, even if one could generate and add all possible rank-1 cuts, this procedure needs to be repeated an arbitrarily large number of times to obtain the convex hull, which gives an indication of how hard it is to solve integer programs by using only CG cuts.

Cook *et al.* [28] and Gerards [29] show that there is an upper bound on the CG-rank that only depends on the coefficients of the constraint matrix (not on the right-hand side); see Refs 9 and 20. There are, however, special cases that have some more reasonable bounds on the CG rank of polyhedra. For

problems where P is contained in the 0/1 cube, Bockmayr and Eisenbrand [30] give a bound of $O(n^3 \log n)$ on the CG rank of a polyhedron, where n is the number of integer variables. This bound was later improved to $O(n^2 \log n^{n/2})$ by Bockmayr *et al.* [31] and then to $O(n^2 \log n)$ by Eisenbrand and Schulz [32]. In that same paper, they give a different upper bound of $n + \|c\|_1$ on the rank of any inequality $cx \leq \delta$ (with $c \in \mathbb{Z}^n$) valid for the integer hull of P . A very similar bound was obtained by Chvátal *et al.* [33] for monotone polyhedra.

In that same paper, Chvátal *et al.* [33] also give an example of a polyhedron P with a CG rank of n (where n is the number of variables), thereby showing that the lower bound on the CG-rank of polyhedra is n . Several other results are known about rank of polyhedra. For instance, Hartmann [34] gives conditions under which an inequality has rank greater than 1 and Eisenbrand and Schulz [32] give a lower bound on the rank of value $(1 + \epsilon)n$ for some $\epsilon > 0$.

These results on the rank are for CG cuts in the pure integer case and do not immediately translate to the mixed-integer case. For instance, Cook *et al.* [18] give an example of a simple mixed-integer program (MIP) that has infinite GMI-rank, which means that even if we add all possible GMI cuts at every round of cut generation, we still do not get even a finite algorithm for solving a very simple MIP.

However, like in the case of the CG rank, the GMI-rank becomes much smaller when the polyhedra are contained in the 0/1 cube. In particular, Cornuéjols and Li [35] show that for a mixed-integer program where all the n integer variables are restricted to be binary the GMI rank of the polyhedron P is bounded above by n . This result actually follows from the result of Balas [17] on facial disjunctive programs and is best possible as there exists a lower bound of n given by Cornuéjols and Li [35] themselves. The example they use to provide the lower bound is exactly the same example proposed for the CG-rank by Chvátal *et al.* [33]. Notice that, in particular, these results also imply that, for pure 0–1 integer programs, the GMI-rank of polyhedra has an upper bound of n , in

contrast with the CG rank for which only an upper bound of $O(n^2 \log n)$ is currently known.

SPECIAL CASES, VARIATIONS, STRENGTHENING, AND OTHERS

The importance of CG and GMI cuts led several researchers to consider different special cases, variations, and ways to strengthen these cuts. This section highlights a few of these endeavors.

$\{0, \frac{1}{2}\}$ -Cuts and mod- k Cuts

The term $\{0, \frac{1}{2}\}$ -CG cuts ($\{0, \frac{1}{2}\}$ -cuts for short) was introduced by Caprara and Fischetti [36]. The idea of $\{0, \frac{1}{2}\}$ -cuts is to restrict the multipliers $u \in \mathbb{R}_+^m$ used to aggregate the linear constraints of the LP relaxation to be in the set $\{0, \frac{1}{2}\}$. The importance of this special case of Gomory cuts is that many classes of inequalities for combinatorial optimization problems are $\{0, \frac{1}{2}\}$ -cuts, for example, comb inequalities for the traveling salesman problem (TSP). In Ref. 36, it was shown that, even though the CG multipliers are very restricted, separation of $\{0, \frac{1}{2}\}$ -cuts remains strongly NP-hard. In Ref. 26, the authors show that, even under the nonnegativity assumption, separation of $\{0, \frac{1}{2}\}$ -cuts remains strongly NP-hard. Recently, Letchford *et al.* [37] have shown that, even if we restrict ourselves to $\{0, 1\}$ problems and allow some additional structure like set-packing, the separation of $\{0, \frac{1}{2}\}$ -cuts remains strongly NP-hard.

Nevertheless, computationally, Andreello *et al.* [38] show that $\{0, \frac{1}{2}\}$ -cuts can significantly decrease the solution time of branch-and-cut codes for some specific combinatorial optimization problems. Moreover, there are some positive theoretical results involving $\{0, \frac{1}{2}\}$ -cuts.

In order to introduce these results, let us first introduce the mod- k cuts, which generalize $\{0, \frac{1}{2}\}$ -cuts. The mod- k cuts are CG cuts where the multipliers $u \in \mathbb{R}_+^m$ used to aggregate the linear constraints of the LP relaxation are restricted to be in the set $\{0, \frac{1}{k}, \dots, \frac{k-1}{k}\}$ for some integer $k > 1$. It

is not hard to see that all nondominated CG cuts can be written as a mod- k cuts for some k . Clearly, the NP-hardness results for $\{0, \frac{1}{2}\}$ -cuts immediately imply that the separation of mod- k cuts is also NP-hard. This particular class of cuts was studied by Caprara *et al.* [39] where they showed that if we only restrict ourselves to maximally violated mod- k cuts, then we can solve the separation problem in polynomial time. This positive result obviously extends to $\{0, \frac{1}{2}\}$ -cuts and generalizes a result of Applegate *et al.* [40] and Fleischer and Tardos [41] for the separation of maximally violated comb inequalities for the TSP. Later on, Letchford [42] extended their results to show that *totally tight* mod- k -cuts (of which maximally violated mod- k cuts are a subclass) can also be separated in polynomial time. Though Caprara *et al.* [39] had already derived this result for totally tight mod- k cuts as well, in Caprara *et al.*'s result one must specify in advance which class of mod- k cuts one wishes to separate (i.e, we must know what k is in advance). Letchford's results show that this is actually not necessary.

Another generalization of $\{0, \frac{1}{2}\}$ -cuts are binary clutter inequalities [43] for which separation is again strongly NP-hard, but polynomially solvable under certain conditions (this, in particular, gives other conditions under which $\{0, \frac{1}{2}\}$ -cuts can be separated in polynomial time).

Projected CG Cuts

Notice that CG cuts are only defined for pure integer programs. However, if one considers the general mixed integer set

$$P_I = \{(x, z) \in \mathbb{Z}_+^n \times \mathbb{R}_+^p : Ax + Dz \leq b\}, \quad (26)$$

one can still use CG cuts to obtain important cuts for P_I . The idea is as follows. Let P be the polyhedron obtained by dropping the integrality constraints from P_I , that is,

$$P = \{(x, z) \in \mathbb{R}_+^n \times \mathbb{R}_+^p : Ax + Dz \leq b\}. \quad (27)$$

Now, define $P(x)$ as the projection of P in the x -space. Finally, define $P_I(x) := P(x) \cap \mathbb{Z}^n$.

Then, since $P_I(x)$ is a pure integer set, one can generate CG cuts for it. Moreover, due to the structure of how $P(x)$ is obtained, it is easy to see that any CG cut generated for $P_I(x)$ is a cut that is valid for P_I . For more details on this, see Ref. 44. In particular, the authors of that paper also show that such cuts correspond to a specific class of split cuts (and hence to a specific class of GMI cuts). Another interesting result related to projected CG cuts is that, if all the continuous variables have a zero coefficient in the objective function, then one can optimize over the MIP by using a finite number of rounds of projected CG cuts.

Other Variations

There have been several other variations, strengthenings, and studies of Gomory cuts. We mention a few of them as follows:

- k -cuts [45] are GMI cuts obtained from integer multiples of tableau rows. The above paper analyzes when k -cuts are better than the simple GMI cuts obtained by a tableau row and give a study when k -cuts seem to help in practice.
- In reduce-and-split cuts [46], the authors observe that the coefficients of the continuous variables in the GMI cut are an important factor to determine the distance from the cut to the point being cut off. Since this distance is one of the measures of “cut quality,” the authors give an algorithm to try to reduce the coefficients of the continuous variables and show that by doing so there is a significant improvement over only using GMI cuts from tableau rows.
- Balas and Perregard [47] give a precise correspondence between lift-and-project cuts [48] for mixed 0–1 programs, split cuts [18], and GMI cuts. Such correspondence gives rise to the upper bound on the GMI rank of n for pure 0/1 IPs and to an algorithm that tries to obtain better cuts than the GMI cut from a tableau row by looking at GMI cuts obtained from a (possibly infeasible) basis of the tableau.

- Letchford and Lodi [42] show a simple way to strengthen CG cuts, obtaining a cut with CG rank 2.
- Köppe and Weismantel [49] use basis reduction to generate strong GMI cuts.
- Ceria *et al.* [50] consider GMI cuts obtained by aggregating the rows of the optimal tableau from the LP relaxation with integer multipliers.
- Glover and Sherali [51] give a closed-form expression for a cut obtained by repeating the CG procedure several times, thus obtaining a closed form for some CG cuts whose rank is higher than 1.

These are just a few examples and the literature of papers that fall into this same category is enormous and would not fit into the available space we have. Therefore, we refrain from making an exhaustive list of the same.

Gomory Cuts as Group Cuts

In this section, we introduce Gomory’s corner and group relaxations, which are important relaxations for MIPs. The *corner relaxation* is a common relaxation of the feasible region of an MIP, obtained from the system of inequalities defining the tableau rows of the LP relaxation and relaxing the nonnegativity constraints on the basic variables.

Specifically, consider the feasible region for an MIP:

$$\begin{aligned} Ax &= b \\ x &\geq 0 \\ x_i &\in \mathbb{Z}, \forall i \in I \subseteq \{1, \dots, n\}. \end{aligned} \quad (28)$$

The tableau rows from the LP relaxation yield the system:

$$\begin{aligned} x_B &= f + Rx_N \\ x_B, x_N &\geq 0 \\ x_i &\in \mathbb{Z}, \forall i \in I \subseteq \{1, \dots, n\}, \end{aligned} \quad (29)$$

where $f = A_B^{-1}b$ and $R = -A_B^{-1}A_N$. For simplicity, we assume that $I = \{1, \dots, n\}$ and that

$|B| = m$. If we drop the nonnegativity constraints corresponding to variables x_B , we get the *corner relaxation*:

$$\left\{ \begin{array}{l} x_B = f + \sum_{j \in N} r^j x_j \\ x : \quad x_N \geq 0 \\ \quad x_B \in \mathbb{Z}^m \\ \quad x_i \in \mathbb{Z}, \forall i \in N \end{array} \right\}, \quad (30)$$

where r^j are the columns of R .

If we add one variable for every possible element of the group generated by the r^j vectors, we get the master group relaxation. If $m = 1$, we call it the master cyclic group relaxation. For a more thorough discussion on the group problem, corner polyhedron, and master group relaxations, the reader is referred to Refs 52–54 and the article titled ***Inequalities from Group Relaxations*** in this encyclopedia.

Now we call a function ψ valid for the group relaxation if, for every feasible solution x for the group relaxation, we have that $\sum \psi(r)x_r \geq 1$. Valid functions give rise to valid inequalities for Equation (30) and as a consequence valid inequalities to the original MIP. Also, GMI cuts can be obtained by extreme valid functions (the analogous to a facet-defining inequality), which shows that GMI cuts are important in the context of group cuts. We further discuss the importance of GMI cuts in the context of group cuts in the next section.

IMPLEMENTATION AND COMPUTATIONAL ASPECTS

We now turn our attentions to some of the computational aspects and results that are related to Gomory cuts.

Even though there was interest in Gomory cuts after the 1958 paper [1], they were thought to be computationally ineffective for quite some time [12,55]. Among the many criticisms, Gomory cuts were considered to converge very poorly and their use inside a branch-and-cut framework was questioned since it would require some complex bookkeeping because cuts generated in the branch-and-cut tree are not valid for the

whole problem. Much to the surprise of the community, Balas *et al.* [56] showed in 1996 that GMICs can be used effectively within a branch-and-cut framework for 0–1 MIPs. A more detailed and accurate description of the history of the above paper can be found in Ref. 57. In particular, in that paper, they show a way to lift Gomory cuts for 0–1 MIPs so that they become globally valid regardless of where in the branch-and-cut tree are they generated. They also highlight some computational strategies, some of which are still used today, like adding multiple cutting planes at a time to the LP relaxation.

The importance of that paper in reigniting the interest in Gomory cuts as an important tool to solve general MIPs is remarkable. Indeed, currently, the ability to generate GMICs is widely considered an essential part of most MIP solvers [58]. This fact can be evidenced by a series of papers by Bixby *et al.* [59,60] where, among other things, they test which of the features in CPLEX [61] are more important for the solution of MIPs. They tested the average performance degradation on the solution times for a series of benchmark instances. Their conclusions are that “The clear winner in these tests was cutting planes. Disabling this feature resulted in a deterioration of a *factor* of almost 54 in overall performance, a remarkable difference.”, and among the cutting planes “Gomory cuts are the clear winner by this measure.”.

Though it is hard to say what is the exact influence of such a paper in the research efforts of the following years, we believe it is safe to say the Balas *et al.* paper [56] generated a large interest in Gomory cuts decades after their introduction.

In addition to the already mentioned variations and improvements over Gomory cuts, for which several computational tests have been performed, recently there have been some other papers focusing on computational aspects of Gomory cuts, which we highlight below.

Strength of Gomory Cuts

The practical success of Gomory cuts described above comes mostly by using heuristics for generating such cuts inside commercial codes. While their practical success shows the quality of these heuristics, it remains to be answered if we could do a better job in using these (or other) cutting planes. In other words, the natural questions are as follows: How well are we using Gomory cuts? Could we do much better by trying to generate different Gomory cuts or other classes of cuts? In this section, we review some of the efforts to answer such questions.

One answer to those questions has recently been studied by Fischetti and Lodi [62]. In the above paper, they study the effect of rank-1 CG cuts by modeling the separation problem as an MIP. They obtain good results, in particular, being able to solve an open instance of MIPLIB 2003 [63] for the first time. Their approach was later extended [64,65] to pose the separation of GMI cuts (equivalently MIR/split cuts) as an MIP and use them to evaluate the rank-1 GMI closure. Their results also show that there is still much to gain by using rank-1 Gomory cuts if one is able to separate them well or identify the useful ones. Indeed, recently, Dash and Goycoolea [66] proposed a heuristic for generating rank-1 GMI cuts that seems to perform very well in practice.

The papers mentioned above show that it may be worth spending time to generate “better” Gomory cuts. However, one may switch focus and try to generate cuts that are different than Gomory cuts in hopes of improving the practical performance of an MIP solver. However, even though several other classes of cutting planes for integer programs have been proposed throughout the years, the practical success of several of those inequalities when compared to GMI cuts is limited. This has led researchers to investigate if there is an intrinsic reason for this to happen. Indeed, when compared to group cuts, Dash and Günlük [67] show that interpolated subadditive cuts are dominated by GMI cuts (in a probabilistic sense). In addition, they show in Ref. 68 that after adding

GMI cuts for the tableau rows of the LP relaxation, it is often the case that there are no other violated cyclic group cuts from those same tableau rows. In particular, in those instances, one would not gain anything by attempting to add other cyclic group cuts from the same tableau rows in addition to the GMI cuts.

These results are in line with the computational results of Fischetti and Saturni [69], who show that the performance of GMI is seldom improved when using other cyclic group cuts obtained by interpolation (something that Cornuéjols *et al.* [45] also observe for k -cuts instead of cyclic group cuts). Another indication that GMI are the most important cuts among all the cyclic group cuts comes from Gomory *et al.*’s shooting experiment [70], where they computationally show that a random sample of facets of the cyclic group have most of its elements as GMI inequalities.

In contrast, Fischetti and Monaci [71] study the corner relaxation and show that often one can have a significant bound improvement over GMIs if the whole corner/group relaxation is considered (that is considering all the rows of those relaxations simultaneously—not just a single row at a time as in the case of the cyclic group).

Another empirical study of the strength of GMI cuts is performed in Ref. 72, where they show that adding several rounds of GMI cuts based on the initial tableau rows of the LP relaxation often approximates the intersection of the convex hull of integer points of each of those tableau rows well. In other words, even if one could use any cut valid for the convex hull of integer points satisfying one of the tableau rows, there would often be a negligible gain in terms of bounds, thereby showing the strength of GMI cuts even outside the context of group cuts.

All these computational results seem to indicate that, if one is to significantly improve the performance of Gomory cuts, one either needs to consider cuts that are valid for relaxations with several constraints (as opposed to one constraint as is the case of Gomory cuts) or find a way to identify important single-constraint relaxations and

generate Gomory cuts for it. In fact, in the section “Multirow Gomory Cuts,” we discuss recent efforts toward the first alternative, namely, the study of multirow Gomory cuts.

Numerical Issues and Pure Cutting-Plane Algorithm

When designing a computer code to solve a mixed-integer program, several issues start becoming relevant to the accuracy and effectiveness of such code. For a more complete survey on these issues and the progress of mixed-integer programming computations, we refer the reader to Ref. 58.

We now focus on two of these computational issues that are faced by Gomory cuts. The first one is the issue of cut validity. Note that the derivation of Gomory cuts is purely algebraic and hence relies on the accuracy of numerical computations. However, algebraic calculations in computers are usually performed in floating-point arithmetic, which has intrinsic errors. Usually these errors are very small, but even a very small error can cause a big difference in the coefficient obtained in a Gomory cut. For example, the CG cut for $x_1 + 2x_2 \leq 2.005$ is (assuming $x_1, x_2 \in \mathbb{Z}_+$) is $x_1 + 2x_2 \leq 2$. However, if there is an error of 0.01 in the calculations, we may obtain $x_1 + 2x_2 \leq 1.995$ and hence the cut $x_1 + 2x_2 \leq 1$, which may be an invalid cut. Since the generation of an invalid cut may lead to an MIP solver cutting off the optimal solution and returning the incorrect solution, one needs to be careful when implementing Gomory cuts using floating-point arithmetic.

The usual approach taken by MIP solvers is to implement Gomory cuts with care using heuristic rules to try to avoid such numerical problems. However, as reported by Margot [73], modern cutting-plane generators still generate invalid cuts once in a while. Neumaier and Shcherbina [74], Cook *et al.* [75], and Zanette *et al.* [76] propose methods to generate Gomory cuts that will not suffer from these numerical problems by having a structured way to control the error in the floating-point calculations.

Another issue that is often faced when using Gomory cuts (and that is somewhat

related to the issue of accuracy) is that, after adding a few rounds of cuts, the cut coefficients tend to start getting large and the determinant of the basis matrix in the optimal tableau also tends to grow. This can lead to numerical problems in the LP solver and is one of the reasons why commercial solvers tend not to add too many rounds of Gomory cuts and also why there is a common knowledge that pure cutting-plane algorithms using Gomory cuts are ineffective.

Recently, however, the above has been discussed again in Ref. 76, where they show that pure cutting-plane algorithms can actually perform well, provided certain care is taken. They show that, if one is to perform Gomory’s cutting-plane algorithm with the lexicographic dual simplex, a significant improvement can be obtained in the final bound. Moreover, the authors also show that the problem of obtaining large cut coefficients and basis determinants becomes much less pronounced.

In Ref. 77, the authors extend their study to understand what is the relationship between cutting-plane methods and some enumerative schemes, proposing different variants of Gomory’s cutting-plane method. This line of research is extremely recent and shows that there are still several issues (both computational and theoretical) that we do not fully understand with respect to Gomory cuts and the cutting-plane method.

MULTIROW GOMORY CUTS

We now focus on an extension of Gomory cuts that consider multiple-row relaxations of an MIP. Recently, there has been a significant amount of work devoted to these so-called *multirow Gomory cuts*. The idea is to extend the fact that GMI cuts are, among the cyclic group cuts, the ones that have the best coefficients for the continuous variables. Indeed, Dash and Günlük [68] observe that this is one of the reasons why GMI cuts are (computationally) the most important cuts among all cyclic group cuts. The extension of such an idea to multiple rows is to consider a tableau of the linear programming relaxation of an

IP, derive the cuts with the best coefficients for the continuous variables, and lift back the integer variables.

To achieve this goal, we follow the same steps that were developed in the section titled “Gomory Cuts as Group Cuts” to obtain the corner and group relaxations. If we now consider Equation (30) and further drop all integrality constraints from x_N and the non-negativity constraints for x_B , we get the following set:

$$\left\{ x : \begin{array}{l} f + \sum_{j \in N} r^j x_j \in \mathbb{Z}^m \\ x_N \geq 0. \end{array} \right\}. \quad (31)$$

So if we obtain a facet-defining inequality for Equation (31) and then lift the integer variables back to obtain an inequality for Equation (30), then we get an inequality with good coefficients for any continuous variable, thereby emulating the same property that GMI cuts have in the case $m = 1$.

Andersen *et al.* [78] show that facet-defining inequalities for Equation (31) when $m = 2$ are intersection cuts and are related to lattice-free convex sets. Borozan and Cornuéjols [79] show that minimal inequalities for a semiinfinite extension of Equation (31) obtained by adding one variable for all possible real coefficients $r^j \in \mathbb{R}^m$ correspond to maximal lattice-free convex sets. Cornuéjols and Margot [80] study which of these inequalities define facets of Equation (31), while the papers [81–83] study the lifting of the integer variables to obtain cuts valid for Equation (30). Other recent papers related to multirow Gomory cuts are Refs 84, 85.

Extensions of multirow Gomory cuts are studied in Refs 86–88 by considering tighter relaxations than Equation (31) and its semiinfinite extension. In addition, the papers [89,90] show computational experiments with multirow Gomory cuts and there are several other groups performing computational experiments as well. Finally, the papers [91,92] study the GMI rank of the multirow Gomory cuts. One particularly interesting result is that some of these cuts have an infinite GMI rank, thereby showing

that a lot can be gained by using these cuts in addition to GMI cuts.

Summing up, this is one particular area of research that is very recent and very active. Indeed the earliest of the 15 papers cited above is dated 2007 and we know of several different research groups actively researching questions related to multirow Gomory cuts.

CONCLUSION

In this article, we tried to do a very brief review of Gomory cuts, trying to be at the same time thorough and concise to fit into the allowed space. While this topic is as old as the field of integer programming, this remains a very active research topic, with much research still being carried out today. We hope that we have been able to provide a glimpse into the topic, stating the main results and with enough references so that the interested reader can explore further.

REFERENCES

1. Gomory RE. Outline of an algorithm for integer solutions to linear programs. Bull Am Math Soc 1958;64:275–278.
2. Gomory RE. An algorithm for integer solutions to linear programs. In: Graves P, Wolfe P, editors. Recent advances in mathematical programming. New York: McGraw-Hill; 1963. pp. 269–302.
3. Gomory RE. An algorithm for the mixed integer problem. Technical Report RM-2597. Santa Monica (CA): RAND Corporation; 1960.
4. Cornuéjols G. Valid inequalities for mixed integer linear programs. Math Program Ser B 2008;112:3–44.
5. Marchand H, Martin A, Weismantel R, *et al.* Cutting planes in integer and mixed integer programming. Discrete Appl Math 2002; 123(1-3):397–446.
6. Padberg M. Classical cuts for mixed-integer programming and branch-and-cut. Math Methods Oper Res 2001;53:173–203.
7. Cornuéjols G, Li Y. Elementary closures for integer programs. Oper Res Lett 2001;28:1–8.
8. Chvátal V. Edmonds polytopes and a hierarchy of combinatorial optimization. Discrete Math 1973;4:305–337.

9. Nemhauser GL, Wolsey LA. Integer and combinatorial optimization. Discrete mathematics and optimization. New York: Wiley-Interscience; 1999.
10. Nemhauser GL, Wolsey LA. A recursive procedure for generating all cuts for 0-1 mixed integer programs. *Math Program* 1990;46:379–390.
11. Marchand H, Wolsey LA. Aggregation and mixed integer rounding to solve MIPs. *Oper Res* 2001;49:363–371.
12. Nemhauser GL, Wolsey LA. Integer programming. In: Nemhauser GL, Rinnooy Kan AHG, Todd MJ, editors. *Handbook in operations research and management science 1: optimization*. Amsterdam: North-Holland; 1989. pp. 447–527.
13. Wolsey LA. Integer programming. New York: Wiley-Interscience; 1998.
14. Bonami P, Cornuéjols G. A note on the MIR closure. *Oper Res Lett* 2008;36:4–6.
15. Dash S, Günlük O, Lodi A. On the MIR closure of polyhedra. *Math Program* 2010;121(1):33–60.
16. Chvátal V. Edmonds polytopes and weakly hamiltonian graphs. *Math Program* 1973;5:29–40.
17. Balas E. Disjunctive programming. *Ann Discrete Math* 1979;5:3–51.
18. Cook W, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program* 1990;47:155–174.
19. Schrijver A. On cutting planes. *Ann Discrete Math* 1980;9:291–296.
20. Schrijver A. *Theory of linear and integer programming*. Chichester: Wiley-Interscience; 1986.
21. Bockmayr A, Eisenbrand F. Cutting planes and the elementary closure in fixed dimension. *Math Oper Res* 2001;26:304–312.
22. Andersen K, Cornuéjols G, Li Y. Split closure and intersection cuts. In: Cook WJ, Schulz AS, editors. *Volume 2337, IPCO, Lecture notes in computer science*. Berlin: Springer; 2002. pp. 127–144.
23. Andersen K, Cornuéjols G, Li Y. Split closure and intersection cuts. *Math Program* 2005;102:457–493.
24. Vielma JP. A constructive characterization of the split closure of a mixed integer linear program. *Oper Res Lett* 2007;35:29–35.
25. Eisenbrand F. On the membership problem for the elementary closure of a polyhedron. *Combinatorica* 1999;19:297–300.
26. Caprara A, Letchford A. On the separation of split cuts and related inequalities. *Math Program* 2003;94(2-3):279–294.
27. Bienstock D, Zuckerberg M. Approximate fixed-rank closures of covering problems. *Math Program* 2006;105:9–27.
28. Cook W, Gerards AMH, Schrijver A, *et al.* Sensitivity theorems in integer linear programming. *Math Program* 1986;34:251–264.
29. Gerards AMH. On cutting planes and matrices. In: Cook W, Seymour PD, editors. *Polyhedral combinatorics*. Providence (RI): DIMACS; 1990. pp. 29–32.
30. Bockmayr A, Eisenbrand F. On the Chvátal rank of polytopes in the 0/1 cube. Technical Report MPI-I-97-2-009. Saarbrücken: Max-Planck-Institut für Informatik; 1997.
31. Bockmayr A, Eisenbrand F, Hartmann M, *et al.* On the Chvátal rank of polytopes in the 0/1 cube. *Discrete Appl Math* 1999;98:21–27.
32. Eisenbrand F, Schulz AS. Bounds on the Chvátal rank of polytopes in the 0/1 cube. *Combinatorica* 2003;23(2):245–261.
33. Chvátal V, Cook W, Hartmann M. On cutting-plane proofs in combinatorial optimization. *Linear Algebra Appl* 1989;114/115:455–499.
34. Hartmann M, Queyranne M, Wang Y. On the Chvátal rank of certain inequalities. In: Cornuéjols G, Burkhard RE, Woeginger GJ, editors. *Volume 1610, IPCO, Lecture notes in computer science*. Berlin: Springer; 1999. pp. 218–233.
35. Cornuéjols G, Li Y. On the rank of mixed 0,1 polyhedra. *Math Program A* 2002;91:391–397.
36. Caprara A, Fischetti M. {0-1/2}-Chvátal-Gomory cuts. *Math Program* 1996;74:221–236.
37. Letchford AN, Pokutta S, Schulz AS. On the membership problem for the {0,1/2}-closure. In press. Available at http://www.lancs.ac.uk/staff/letchfoa/articles/zero_half_hard.pdf.
38. Andreello G, Caprara A, Fischetti M. Embedding cuts in a branch-and-cut framework: a computational study with {0,1/2}-cuts. *INFORMS J Comput* 2007;19(2):229–238.
39. Caprara A, Fischetti M, Letchford AN. On the separation of maximally violated mod-k cuts. *Math Program* 2000;87:37–56.
40. Applegate D, Bixby RE, Chvátal V, *et al.* Finding cuts in the TSP. Technical Report 95-05. New Brunswick (NJ): DIMACS; 1995.
41. Fleischer L, Tardos E. Separating maximally violated comb inequalities in planar graphs.

- Math Oper Res 1999;24:130–148.
42. Letchford AN, Lodi A. Strengthening Chvátal-Gomory cuts and Gomory fractional cuts. *Oper Res Lett* 2002;30(2):74–82.
 43. Letchford AN. Binary clutter inequalities for integer programs. *Math Program Ser B* 2003;98:201–221.
 44. Bonami P, Cornuéjols G, Dash S, *et al.* Projected Chvátal-Gomory cuts for mixed integer linear programs. *Math Program* 2008;113:241–257.
 45. Cornuéjols G, Li Y, Vandenbussche D. K-cuts: A variation of Gomory mixed integer cuts from the LP tableau. *INFORMS J Comput* 2003;15(4):385–396.
 46. Andersen K, Cornuéjols G, Li Y. Reduce-and-split cuts: Improving the performance of mixed-integer Gomory cuts. *Manage Sci* 2005;51:1720–1732.
 47. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0-1 programming. *Math Program* 2003;94:221–245.
 48. Balas E, Ceria S, Cornuéjols G. A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math Program* 1993;58:295–324.
 49. Köppe M, Weismantel R. A mixed-integer Farkas lemma and some consequences. *Oper Res Lett* 2004;32:207–211.
 50. Ceria S, Cornuéjols G, Dawande M. Combining and strengthening Gomory cuts. In: Balas E, Clausen J, editors. Volume 920, IPCO, Lecture notes in computer science. Berlin: Springer; 1995. pp. 438–451.
 51. Glover F, Sherali H. Chvátal-Gomory-tier cuts for general integer programs. *Discrete Optimization* 2005;2:51–69.
 52. Dey SS, Richard JP. The Group-theoretic approach in mixed integer programming. 50 years of integer programming 1958–2008. Heidelberg: Springer; 2009. pp. 727–801.
 53. Gomory RE, Johnson EL. Some continuous functions related to corner polyhedra I. *Math Program* 1972;3:23–85.
 54. Gomory RE, Johnson EL. Some continuous functions related to corner polyhedra II. *Math Program* 1972;3:359–389.
 55. Williams HP. Model building in mathematical programming. New York: Wiley; 1985.
 56. Balas E, Ceria S, Cornuéjols G, *et al.* Gomory cuts revisited. *Oper Res Lett* 1996;19:1–9.
 57. Cornuéjols G. Revival of the Gomory cuts in the 1990's. *Ann Oper Res* 2007;149:63–66.
 58. Lodi A. Mixed integer programming computation. 50 years of integer programming 1958–2008. Heidelberg: Springer; 2009. pp. 619–645.
 59. Bixby RE, Fenelon M, Gu Z, *et al.* MIP: Theory and practice - closing the gap. System Modelling and Optimization: Methods, Theory, and Applications. Dordrecht: Kluwer Academic Publishers; 1999. pp. 19–50.
 60. Bixby RE, Gu Z, Rothberg E, *et al.* Mixed integer programming: a progress report. The sharpest cut: the impact of Manfred Padberg and his work, MPS/SIAM Series on Optimization. Philadelphia: SIAM; 2004. pp. 309–326.
 61. CPLEX. Available at <http://www.ilog.com/products/cplex>.
 62. Fischetti M, Lodi A. Optimizing over the Chvátal closure. In: Juenger M, Kaibel V, editors. Volume 3509, IPCO, lecture notes in computer science. Berlin: Springer; 2005. pp. 12–22.
 63. Achterberg T, Koch T, Martin A. MIPLIB 2003. *Oper Res Lett* 2006;34(4):1–12. See Available at <http://miplib.zib.de>.
 64. Balas E, Saxena A. Optimizing over the split closure. *Math Program* 2008;113(2): 219–240.
 65. Dash S, Günlük O, Lodi A. On the MIR closure of polyhedra. In: Fischetti M, Williamson D, editors. Volume 4513, IPCO, lecture notes in computer science. Berlin: Springer; 2007. pp. 337–351.
 66. Dash S, Goycoolea M. A heuristic to generate rank-1 GMI cuts. Technical report available at Optimization Online. 2009 Oct.
 67. Dash S, Günlük O. Valid inequalities based on the interpolation procedure. *Math Program* 2006;106(1):111–136.
 68. Dash S, Günlük O. On the strength of Gomory mixed-integer cuts as group cuts. *Math Program* 2008;115:387–407.
 69. Fischetti M, Saturni C. Mixed-integer cuts from cyclic groups. *Lect Notes Comput Sci* 2005;3509:1–11.
 70. Gomory RE, Johnson EL, Evans L. Corner polyhedra and their connection with cutting planes. *Math Program* 2003;96(2):321–339.
 71. Fischetti M, Monaci M. How tight is the corner relaxation? *Discrete Optim* 2008;5:262–269.
 72. Fukasawa R, Goycoolea M. On the exact separation of mixed integer knapsack cuts. *Math Program*. In press.
 73. Margot F. Testing cut generators for mixed-integer linear programming. *Math Program*

- Comput 2009;1:69–95.
74. Neumaier A, Shcherbina O. Safe bounds in linear and mixed-integer linear programming. *Math Program* 2004;99:283–296.
 75. Cook WJ, Dash S, Fukasawa R, *et al.* Numerically safe Gomory mixed-integer cuts. *INFORMS J Comput* 2009;21(4):641–649.
 76. Zanette A, Fischetti M, Balas E. Lexicography and degeneracy: Can a pure cutting plane algorithm work? *Math Program* 2010. In press.
 77. Balas E, Fischetti M, Zanette A. On the enumerative nature of Gomory’s dual cutting plane method. *Math Program B* 2010. In press.
 78. Andersen K, Louveaux Q, Weismantel R, *et al.* Inequalities from two rows of a simplex tableau. In: Fischetti M, Williamson D, editors. Volume 4513, IPCO, lecture notes in computer science. Berlin: Springer; 2007. pp. 1–15.
 79. Borozan V, Cornuéjols G. Minimal valid inequalities for integer constraints. *Math Oper Res* 2009;34:538–546.
 80. Cornuéjols G, Margot F. On the facets of mixed integer programs with two integer variables and two constraints. *Math Program* 2009;120:429–456.
 81. Basu A, Campelo M, Conforti M, *et al.* On lifting integer variables in minimal inequalities. 2009. Available at <http://integer.tepper.cmu.edu>.
 82. Conforti M, Cornuéjols G, Zambelli G. A geometric perspective on lifting. 2009. Available at <http://integer.tepper.cmu.edu>.
 83. Dey SS, Wolsey LA. Lifting integer variables in minimal inequalities corresponding to lattice-free triangles. In: Lodi A, Panconesi A, Rinaldi G, editors. Volume 5035, IPCO, lecture notes in computer science. Berlin: Springer; 2008. pp. 463–476.
 84. Basu A, Bonami P, Cornuéjols G, *et al.* On the relative strength of split, triangle and quadrilateral cuts. *Math Program A* 2008. In press.
 85. Zambelli G. On degenerate multi-row Gomory cuts. *Oper Res Lett* 2009;37:21–22.
 86. Basu A, Conforti M, Cornuéjols G, *et al.* Minimal inequalities for an infinite relaxation of integer programs. 2009. Available at <http://integer.tepper.cmu.edu>.
 87. Dey SS, Wolsey LA. Constrained infinite group relaxations of MIPs. 2009.
 88. Fukasawa R, Günlük O. Strengthening lattice-free cuts using nonnegativity. 2009. Manuscript available at optimization online.
 89. Espinoza DG. Computing with multi-row Gomory cuts. In: Lodi A, Panconesi A, Rinaldi G, editors. Volume 5035, IPCO, lecture notes in computer science. Berlin: Springer; 2008. pp. 214–224.
 90. Basu A, Bonami P, Cornuéjols G, *et al.* Experiments with two-row cuts from degenerate tableaux. 2009. Available at <http://integer.tepper.cmu.edu>.
 91. Dey SS. A note on split rank of intersection cuts. 2009. Manuscript available at optimization online.
 92. Dey SS, Louveaux Q. Split rank of triangle and quadrilateral inequalities. 2009. Manuscript available at optimization online.

GRADIENT-TYPE METHODS

FOAD MAHDAVI PAJOUH
BALABHASKAR BALASUNDARAM
School of Industrial Engineering and
Management, Oklahoma State
University, Stillwater,
Oklahoma

INTRODUCTION

This article reviews gradient-type descent methods for the unconstrained minimization problem:

$$\min \{f(x) : x \in \mathbb{R}^n\}, \quad (1)$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$. We first focus on the *gradient method* for differentiable functions and then briefly discuss the analogous *subgradient method* for nondifferentiable convex functions. The gradient method, which is also called the *method of steepest descent*, was proposed by Cauchy in 1847 [1] for solving systems of equations. This algorithm is one of the most fundamental and simple derivative-based procedures for unconstrained minimization of a differentiable function. The performance of the method in terms of speed of convergence is lacking, and it tends to suffer from very slow convergence especially as a stationary point is approached. However, it does guarantee global convergence under reasonable conditions, and admits a thorough mathematical analysis of its behavior. For this reason, the gradient method has been used as a starting point in the development of more sophisticated globally convergent algorithms with better convergence properties for unconstrained minimization. Numerous excellent books [2–11] already cover this topic in varying amounts of detail. Our goal here is to present a cogent overview of this fundamental method and consolidate the treatment

of this topic in various books and the general body of nonlinear programming literature.

THE GRADIENT METHOD

We review some basic definitions and properties before discussing the gradient method in detail.

Definition 1. A nonzero direction d is called a *descent direction* at $\bar{x}$ if $f(\bar{x} + \lambda d) < f(\bar{x})$ for all $\lambda \in (0, \delta)$ for some $\delta > 0$.

Definition 2. Consider a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, the directional derivative of f at $\bar{x}$ along a nonzero direction d denoted by $f'(\bar{x}, d)$, is given by the following limit if it exists, which reduces to the inner product whenever the gradient exists

$$f'(\bar{x}, d) = \lim_{\lambda \downarrow 0} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \nabla f(\bar{x})^t d. \quad (2)$$

If $f'(\bar{x}, d) < 0$ then d is a descent direction at $\bar{x}$. Given a differentiable function f , the gradient method in each step, moves along the direction that minimizes $f'(\bar{x}, d)$, that is, the direction of steepest descent. The method stops when such a move is not possible.

Proposition 1. If a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at $\bar{x}$ and $\nabla f(\bar{x}) \neq \mathbf{0}$ then $\bar{d} = -\nabla f(\bar{x}) / \|\nabla f(\bar{x})\|$ minimizes $f'(\bar{x}, d)$ subject to $\|d\| \leq 1$.

Proposition 2. Suppose $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at $\bar{x}$. If $\bar{x}$ is a local minimum, $\nabla f(\bar{x}) = \mathbf{0}$.

Hence, whenever the gradient is nonzero, $-\nabla f(\bar{x})$ gives us the direction of steepest descent at $\bar{x}$. When it is zero, the method has reached a *stationary point*, which is a necessary condition for optimality. Further note that when $\nabla f(\bar{x}) = \mathbf{0}$, $f'(\bar{x}, d) = 0$ for all d . However, it is possible for descent directions of f to exist at $\bar{x}$.

Gradient Method with Exact Line Searches

Given a point $\bar{x}$, the gradient method conducts a line search at $\bar{x}$ along the direction $-\nabla f(\bar{x})/\|\nabla f(\bar{x})\|$ or equivalently along $-\nabla f(\bar{x})$ to find the next point. This step is repeated until some appropriate termination condition is satisfied. Algorithm 1 describes the gradient method. Theorem 1 addresses the convergence of the gradient method.

Algorithm 1. The Gradient Method.

- 1: **Initialization.** Select a starting point x^0 , the termination tolerance $\epsilon > 0$ and let $k = 0$
- 2: **while** $\|\nabla f(x^k)\| > \epsilon$ **do**
- 3: $\lambda_k \leftarrow \operatorname{argmin} \{f(x^k - \lambda \nabla f(x^k)) : \lambda \geq 0\}$
- 4: $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$
- 5: $k \leftarrow k + 1$
- 6: **end while**

Exact line search guarantees descent in the function value and it can also be verified that two consecutive search directions are orthogonal to each other. Since λ_k minimizes $\phi_k(\lambda) = f(x^k - \lambda \nabla f(x^k))$ over $\lambda \geq 0$ and $-\nabla f(x^k) \neq \mathbf{0}$ is a descent direction (i.e., $\lambda_k > 0$), we must have $\phi'_k(\lambda_k) = 0$, or

$$\begin{aligned} \frac{df(x^k - \lambda \nabla f(x^k))}{d\lambda} \Big|_{\lambda=\lambda_k} \\ = -\nabla f(x^{k+1})^t \nabla f(x^k) = 0. \end{aligned} \quad (3)$$

Theorem 1 [3]. Consider the unconstrained minimization of a continuously differentiable function $f: \mathbb{R}^n \rightarrow \mathbb{R}$. Assume that the sequence $\{x^k : k \geq 0\}$ generated by the gradient method described in Algorithm 1 is contained in a compact set. Then every limit point of the sequence $\{x^k : k \geq 0\}$ is a stationary point.

The gradient method with exact line search is globally convergent under the stated assumptions of Theorem 1. The next relevant question deals with the speed of convergence of the gradient method, the analysis of which focuses on convex quadratic and nonquadratic functions. Consider first the convex quadratic case, $f(x) = \frac{1}{2}x^t Qx$, where

Q is symmetric positive definite. Clearly, the global minimum is $\mathbf{0}$ and $f(\mathbf{0}) = 0$. Note that the analysis of the gradient method for convex quadratic functions is very important for understanding the behavior of the method near the local minimum of any twice differentiable function.

Theorem 2 [3]. Suppose the gradient method described in Algorithm 1 is applied to $f(x) = \frac{1}{2}x^t Qx$, then the following inequality holds.

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{\rho - 1}{\rho + 1} \right)^2, \quad k = 0, 1, 2, \dots, \quad (4)$$

where ρ is the condition number of Q .

We can see from Theorem 2, the sequence $\{f(x^k) : k \geq 0\}$ converges to 0 at least linearly. When the condition number is 1, we have convergence in a single step as observed when $f(x) = \frac{1}{2}(x^t x)$. When the condition number is much larger than 1, we say the problem is *ill-conditioned* and the speed of convergence is rather slow. The latter case gives rise to the phenomenon of “zigzagging” that is illustrated in Fig. 1 for a two-dimensional problem with long, narrow level sets. The poor convergence of the gradient method in such cases can be intuitively explained as follows.

The gradient method only utilizes first-order information about the function in determining the direction of movement. It uses the linear approximation of f ,

$$\begin{aligned} f(x^k + \lambda d) &= f(x^k) + \lambda \nabla f(x^k)^t d \\ &\quad + \lambda \|d\| \alpha(x^k; \lambda d), \end{aligned} \quad (5)$$

in which the term $\lambda \|d\| \alpha(x^k; \lambda d)$ is ignored. Close to the stationary point, when $\|\nabla f(x^k)\|$ is almost zero, the term $\lambda \nabla f(x^k)^t d$ is small. Now even for small values of λ , the term $\lambda \|d\| \alpha(x^k; \lambda d)$ may contribute significantly to the description of f . So moving along the directions generated as we approach a stationary point does not effectively decrease the objective function value [2].

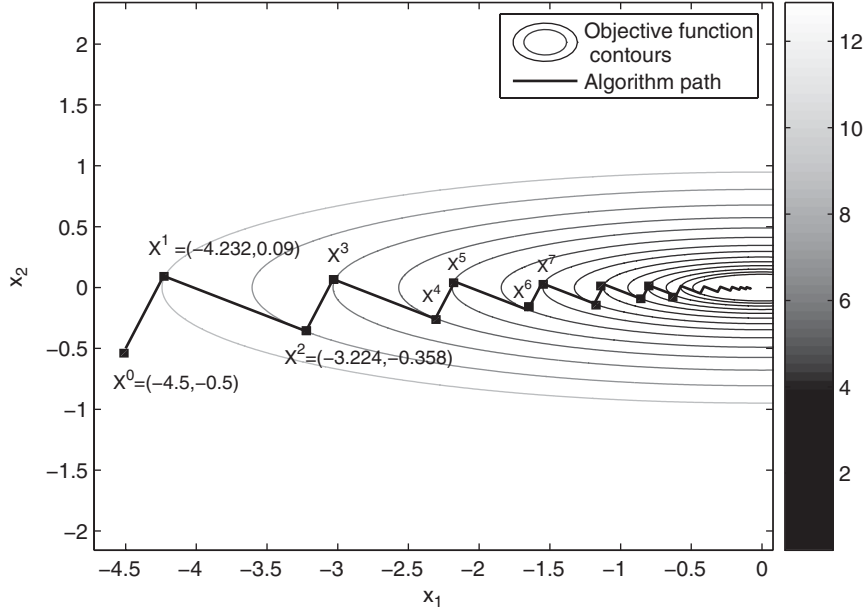


Figure 1. The progress of the gradient method for minimizing $f(x_1, x_2) = (1/2)(x_1^2 + 20x_2^2)$. The minimizing point is $(0, 0)$ and the starting point is $(-4.5, -0.5)$.

Before concluding this section, we provide an extension of Theorem 2 to certain strongly convex functions. Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be *strongly convex* if there exists $m > 0$ such that $\nabla^2 f(x) \geq mI$ for all $x \in \mathbb{R}^n$, where $A \geq B$ means that $A - B$ is positive semidefinite. Note that by Taylor's theorem, for any $x, y \in \mathbb{R}^n$ there exists z on the line segment joining x, y such that

$$f(y) = f(x) + \nabla f(x)^t(y - x) + \frac{1}{2}(y - x)^t \nabla^2 f(z)(y - x). \quad (6)$$

In addition, if f is strongly convex, we obtain

$$f(y) \geq f(x) + \nabla f(x)^t(y - x) + \frac{m}{2}(\|y - x\|)^2, \quad (7)$$

which is stronger than the basic supporting hyperplane inequality that characterizes differentiable convex functions (obtained by setting $m = 0$). In particular, this implies that the sublevel sets of f are bounded and hence, the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is

bounded. Further, $\min\{f(x) : x \in \mathbb{R}^n\}$ has a unique minimum x^* since f is strictly convex.

Theorem 3 [3]. Suppose f is twice continuously differentiable and its Hessian satisfies the condition,

$$mI \leq \nabla^2 f(x) \leq MI, \quad \forall x \in \mathbb{R}^n, \quad (8)$$

where $0 < m \leq M$. Then, for every x^0 the sequence $\{x^k : k \geq 0\}$ generated by the gradient method described in Algorithm 1 satisfies

$$\limsup_{k \rightarrow \infty} \frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \left(\frac{\rho - 1}{\rho + 1} \right)^2, \quad (9)$$

where ρ is the condition number of $\nabla^2 f(x^*)$ and x^* is the unique global minimum.

Similar to Theorem 2 for convex quadratics, Theorem 3 shows that the sequence $\{f(x^k) : k \geq 0\}$ converges to $f(x^*)$ at least linearly. The ill-conditioning effects are also explained by this result.

Gradient Method with Inexact Line Searches

Performing exact line searches is often computationally impractical due to the excessive function evaluations required. The descent property guaranteed by exact line search, however, is important for the convergence of the algorithm. Alternatives to exact line search that still preserve descent and global convergence are important issues in this context. The gradient method can be implemented with a constant step size $\tau > 0$ as

$$x^{k+1} = x^k - \tau \nabla f(x^k), \quad k = 0, 1, 2, \dots \quad (10)$$

Under certain conditions, this approach can also be shown to be convergent.

Theorem 4 [3]. *Assume that the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and its gradient is Lipschitz continuous with the Lipschitz constant $L > 0$. Further assume that the function f is bounded from below*

and the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is bounded. If the step size τ satisfies $0 < \tau < \frac{1}{L}$, then the sequence $\{x^k : k \geq 0\}$ generated using Equation (10) is bounded and every limit point of this sequence is a stationary point.

While the convergence of the gradient method with constant step size is interesting, it requires knowledge of the Lipschitz constant L which is a considerable drawback. A more attractive alternative is the use of Armijo's rule to conduct an inexact line search. We conclude this section with the gradient method that uses Armijo's inexact line search described in Algorithm 2. Theorem 5 establishes global convergence of the algorithm with Armijo's inexact line search. Note that this algorithm is better suited to computer implementation for any differentiable function even though convergence guarantee requires Lipschitz continuity among other conditions.

Algorithm 2. The gradient method with Armijo's line search.

```

1: Initialization. Select a starting point  $x^0$ , the termination tolerance  $\epsilon > 0$ , fixed step
   length parameter  $\bar{\lambda} \in (0, \infty)$ , line search parameters  $\alpha \in (1, \infty)$  and  $\beta \in (0, 1)$  and let  $k = 0$ 
2: while  $\|\nabla f(x^k)\| > \epsilon$  do
3:   Define line search function  $\phi(\lambda) = f(x^k - \lambda \nabla f(x^k))$  for  $\lambda \geq 0$ 
4:   Define Armijo's function  $\hat{\phi}(\lambda) = \phi(0) + \lambda \beta \phi'(0)$  for  $\lambda \geq 0$ 
5:   if  $\phi(\bar{\lambda}) \leq \hat{\phi}(\bar{\lambda})$  then
6:     Find the largest integer  $t \geq 0$  such that  $\phi(\alpha^t \bar{\lambda}) \leq \hat{\phi}(\alpha^t \bar{\lambda})$ 
7:      $\lambda_k = \alpha^t \bar{\lambda}$ 
8:   else
9:     Find the smallest integer  $t \geq 1$  such that  $\phi(\alpha^{-t} \bar{\lambda}) \leq \hat{\phi}(\alpha^{-t} \bar{\lambda})$ 
10:     $\lambda_k = \alpha^{-t} \bar{\lambda}$ 
11:   end if
12:    $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$ 
13:    $k \leftarrow k + 1$ 
14: end while

```

Theorem 5 [2]. *Assume that the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant $L > 0$ on the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$. Suppose Algorithm 2 generates the sequence $\{x^k : k \geq 0\}$, then $\lim_{k \rightarrow \infty} \nabla f(x^k) = 0$. In addition, if the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is bounded, then every limit point of the sequence $\{x^k : k \geq 0\}$ is a stationary point.*

SUBGRADIENT METHOD FOR CONVEX FUNCTIONS

Consider the unconstrained minimization problem (1), in which function f is convex but not necessarily differentiable. A natural analog of the gradient method employs subgradients of the function in place of gradients. However, not all properties of the

gradient method such as descent in each iteration can be guaranteed. Nonetheless, globally convergent subgradient methods can be developed based on some interesting observations.

Definition 3. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. Then $\eta \in \mathbb{R}^n$ is called a *subgradient* of f at $\bar{x}$ if the following condition holds:

$$f(x) \geq f(\bar{x}) + \eta^t(x - \bar{x}) \quad \forall x \in \mathbb{R}^n. \quad (11)$$

The collection of all subgradients at $\bar{x}$ is called the *subdifferential* of f at $\bar{x}$ denoted by $\partial f(\bar{x})$. If f is differentiable at $\bar{x}$, then $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$, the singleton set.

Proposition 3. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. Consider the problem $\min\{f(x) : x \in \mathbb{R}^n\}$. The point $\bar{x} \in \mathbb{R}^n$ is an optimal solution to this problem if and only if $\mathbf{0} \in \partial f(\bar{x})$.

Descent in the function value in each iteration is a key requirement for showing convergence of the gradient method under various settings. It is important to note that $\eta \in \partial f(\bar{x})$, $\eta \neq \mathbf{0}$ need not be a descent direction of f at $\bar{x}$. The negative of the subgradient of minimum norm when it is different from $\mathbf{0}$ is the direction of steepest descent. Furthermore,

Algorithm 3. The subgradient method.

```

1: Initialization. Select a starting point  $x^0$ , the termination tolerance  $\epsilon > 0$ , and let  $k = 0$ 
2:  $UB_k \leftarrow f(x^k)$ , incumbent solution  $x^* \leftarrow x^0$ 
3: Find a subgradient  $g^k \in \partial f(x^k)$  of  $f$  at  $x^k$ 
4: if  $g^k = \mathbf{0}$  then
5:   Stop and return  $x^k$  as the solution
6: else
7:   Let  $d^k = -g^k / \|g^k\|$ , select a step size  $\lambda_k > 0$ 
8:    $x^{k+1} \leftarrow x^k + \lambda_k d^k$ 
9:   if  $f(x^{k+1}) < UB_k$  then
10:     $UB_{k+1} \leftarrow f(x^{k+1})$  and  $x^* \leftarrow x^{k+1}$ 
11:   else
12:     $UB_{k+1} \leftarrow UB_k$ 
13:   end if
14:    $k \leftarrow k + 1$ 
15:   Go to Step 3
16: end if
```

Since Algorithm 3 selects $g^k \in \partial f(x^k)$ arbitrarily, even for some k , $\mathbf{0} \in \partial f(x^k)$, the termination condition $g^k = \mathbf{0}$ may not be achieved. Other practical termination criteria that

even if $\bar{x}$ is optimal, we may fail to detect that $\mathbf{0} \in \partial f(\bar{x})$. Hence, employing a line search along the negative of the subgradient in each iteration is not well motivated. However, from the definition of a subgradient we have

$$\{x : f(x) \leq f(\bar{x})\} \subseteq \{x \in \mathbb{R}^n : \eta^t(x - \bar{x}) \leq 0\}. \quad (12)$$

Hence, taking a step along the negative of the subgradient intuitively takes us “closer” to the optimal point. In fact, if a step of prescribed length is taken along the negative of the subgradient in each iteration, the method can be shown to converge.

Subgradient Method and Convergence

Given an iterate x^k , positive step size λ_k and a subgradient of f at x^k , $g^k \in \partial f(x^k)$, the next iterate x^{k+1} is constructed using the following relationship:

$$x^{k+1} = x^k + \lambda_k d^k, \quad (13)$$

where $d^k = -g^k / \|g^k\|$ is the direction of movement. The scaling coefficient $\|g^k\|$ in d^k can be replaced by $\max(c, \|g^k\|)$ or $\max(c, \|g^k\|^2)$ where $c > 0$ is a fixed constant [3].

limit the maximum number of iterations, or terminate when $\|x^{k+1} - x^k\| < \epsilon$ or $\|x^{k+1} - x^k\| / \|x^k\| < \epsilon$ for a sufficiently small $\epsilon > 0$, can be used. Theorem 6 addresses the

convergence of the subgradient method which can be guaranteed if the step sizes satisfy certain conditions.

Theorem 6 [2,3]. *Suppose f is convex and assume an optimal solution to problem (1) exists. Suppose the subgradient algorithm 3 is used to solve this problem and assume the nonnegative step size λ_k satisfies the following conditions:*

$$\{\lambda_k\} \rightarrow 0^+, \quad (14)$$

$$\sum_{k=0}^{\infty} \lambda_k = \infty. \quad (15)$$

Then, either the algorithm terminates finitely with an optimal solution, or the sequence $\{UB_k : k \geq 0\}$ generated by the subgradient method satisfies

$$\{UB_k\} \rightarrow f^* = \min \{f(x) : x \in \mathbb{R}^n\}. \quad (16)$$

Observe that sequence $\{f(x^k) : k \geq 0\}$ is generally nonmonotone, and we have effectively shown in Theorem 6 that $\liminf_{k \rightarrow \infty} f(x^k) = f^*$. Under stronger conditions, even the convergence of $\{x^k : k \geq 0\}$ can be established.

Theorem 7 [2,3]. *Suppose f is convex and assume an optimal solution to problem (1) exists. Suppose the subgradient algorithm 3 is used to solve this problem and assume the nonnegative step size λ_k satisfies the following conditions:*

$$\sum_{k=0}^{\infty} \lambda_k = \infty, \quad (17)$$

$$\sum_{k=0}^{\infty} \lambda_k^2 < \infty. \quad (18)$$

Then, either the algorithm terminates finitely with an optimal solution, or the sequence $\{x^k : k \geq 0\}$ generated by the subgradient method converges to an optimal solution of Equation (1).

Conditions of Theorem 6 on selecting step sizes guarantees convergence of the subgradient method. But in practice, these

conditions necessarily do not result in the algorithm's convergence. For example, selecting $\lambda = 1/k$ satisfying conditions of Theorem 6 as well as Theorem 7 has been observed to be unsatisfactory in practice [2,3]. Hence, the choice of step length is important and nontrivial. Better insight into the feature of the subgradient method that drives convergence can be gained by carefully exploring the intuition presented by Equation (12). Let x^* be an optimal solution to the problem (1) and x^k be a nonoptimal iterate obtained from the algorithm. Since f is convex, $f(x^*) \geq f(x^k) + (x^* - x^k)^t g^k$ which results in $(x^* - x^k)^t (-g^k) \geq f(x^k) - f(x^*) > 0$. Hence, direction $d^k = -g^k / \|g^k\|$, even when not a descent direction, leads to a point which is closer to x^* in Euclidean norm than x^k . This fundamental fact facilitates convergence of the subgradient method. From this point of view, an ideal step size is one which takes us closest to x^* . In the other words, this ideal step size is the one which makes $x^{k+1} - x^*$ orthogonal to d^k . So we require λ_k^* to satisfy

$$\lambda_k^* = (x^k - x^*)^t g^k / \|g^k\|, \quad (19)$$

which needs to be approximated as x^* is unknown. Since f is convex, $f(x^*) \geq f(x^k) + (x^* - x^k)^t g^k$. Using Equation (19), we have $\lambda_k^* \geq (f(x^k) - f(x^*)) / \|g^k\|$. The step size can then be chosen using the following equation:

$$\lambda_k = \beta_k (f(x^k) - \underline{f}) / \|g^k\|, \quad (20)$$

where $\underline{f}$ is a lower bound on $f(x^*)$ and $\beta_k > 0$. By selecting $\epsilon_1 < \beta_k \leq 2 - \epsilon_2, \forall k$ for some $\epsilon_1, \epsilon_2 > 0$ and using $f(x^*)$ instead of $\underline{f}$, the sequence $\{x^k : k \geq 0\}$ generated by the subgradient algorithm can be shown to converge to x^* [2]. Practical implementations of this method based on Equation (20) employ what is called the *block halving scheme*. For more details on this scheme, the reader is referred to Bazaraa *et al.* [2]. Interested readers can find fundamental results on nondifferentiable optimization in Shor [12]. A recent survey of nondifferentiable optimization algorithms such as the subgradient method, the subgradient method with space

dilation, and the r -algorithm can be found in Shor [13].

CONCLUSION

As it can be observed from Theorems 1, 4, and 5, the global convergence of the gradient method is quite impressive. However, it is offset by its typically slow convergence as observed from Theorems 2 and 3 even for convex minimization. Despite its drawbacks, applications of the gradient method can be found in training neural networks [14], for solving a system of nonlinear equations by solving the sum of squares expression [15], for solving a system of fuzzy linear equations [16], and in multicriteria optimization [17,18]. The gradient method is used to solve 0,1-nonlinear programs by transforming them to an unconstrained optimization problem in Anjidani and Effati [19]. The gradient method with affine scaling is developed for linear programming in Gonzaga [20]. The notion of steepest descent among other classical approaches is used to develop the gravitational method for linear programming in Chang and Murty [21].

Some of the most competitive globally convergent algorithms for unconstrained minimization of differentiable functions can be viewed as extensions of the gradient method that modify the direction of search in a manner that accelerates convergence. Pure Newton method (see **Newton-Type Methods**) converges rapidly when started close to the solution, but does not provide global convergence. A common approach to overcome the poor convergence of gradient method is to deflect the gradient akin to gradient deflection by the inverse of the Hessian matrix used in Newton's method. In this approach, rather than using the direction $d = -\nabla f(x)$, we use direction $d = -D\nabla f(x)$ giving rise to quasi Newton methods (see **Quasi-Newton Methods**), or $d = -\nabla f(x) + g$ giving rise to conjugate gradient methods (see **Nonlinear Conjugate Gradient Methods**). By virtue of its simplicity and amenability to mathematical analysis, the gradient method will remain central to the study of algorithms for nonlinear optimization.

REFERENCES

1. Cauchy A. Méthode générale pour la résolution des systèmes d'équations simultanées. CR Acad Sci I (Paris) 1847;25:536–538.
2. Bazaraa MS, Sherali HD, Shetty CM. Nonlinear programming: theory and algorithms. 3rd ed. New York: Wiley-Interscience; 2006.
3. Ruszczyński A. Nonlinear optimization. Princeton (NJ): Princeton University Press; 2006.
4. Boyd S, Vandenberghe L. Convex optimization. Cambridge (MA): Cambridge University Press; 2004.
5. Bertsekas DP. Nonlinear programming. 2nd ed. Belmont (MA): Athena Scientific; 1999.
6. Luenberger DG, Ye Y. Linear and nonlinear programming. New York: Springer; 2008.
7. Griva I, Nash SG, Sofer A. Linear and nonlinear optimization. 2nd ed. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2008.
8. Nocedal J, Wright SJ. Numerical optimization. New York: Springer; 2000.
9. Bonnans FJ, Gilbert JC, Lemaréchal C, *et al.* Numerical optimization. 2nd ed. Berlin, Heidelberg, New York: Springer-Verlag; 2006.
10. Fletcher R. Practical methods of optimization. 2nd ed. Chichester, New York, Brisbane, Toronto, Singapore: Wiley; 2000.
11. Gill PE, Murray W, Wright MH. Practical optimization. New York: Academic Press; 1981.
12. Shor NZ. Minimization methods for nondifferentiable functions. Berlin: Springer; 1985. (English translators Naukova Dumka, Kiev, 1979.)
13. Shor NZ, Zhurbenko NG, Likhovid AP, *et al.* Algorithms of nondifferentiable optimization: development and application. Cybern Syst Anal 2003;39(4):537–548.
14. Parisi R, Di Claudio ED, Orlandi G, *et al.* A generalized learning paradigm exploiting the structure of feedforward neural networks. IEEE Trans Neural Netw 1996;7:1450–1460.
15. Matveev VA. Method for the approximate solution of a system of nonlinear equations. USSR Comput Math Math Phys 1964;4:1–16.
16. Abbasbandy S, Jafarian A. Steepest descent method for system of fuzzy linear equations. Appl Math Comput 2006;175:823–833.
17. Fliege J, Svaiter BF. Steepest descent methods for multicriteria optimization. Math Methods Oper Res 2000;51(3):479–494.

18. Drummond LMG, Svaiter BF. A steepest descent method for vector optimization. *J Comput Appl Math* 2005;175:395–414.
19. Anjidani M, Effati S. Steepest descent method for solving zero-one nonlinear programming problems. *Appl Math Comput* 2007;193(1):197–202.
20. Gonzaga CC. Polynomial affine algorithms for linear programming. *Math Program* 1990;49(1):7–21.
21. Chang SY, Murty KG. The steepest descent gravitational method for linear programming. *Discrete Appl Math* 1989;25:211–239.

GRAPH SEARCH TECHNIQUES

BOTING YANG
Department of Computer Science,
University of Regina, Regina,
Canada

INTRODUCTION

Graph searching problems serve as mathematical models for many real-world problems, such as capturing a robber in a building, clearing a complex system of interconnected pipes which is contaminated by some noxious gas, or killing a computer virus in a network system. The meaning of a cleared or contaminated edge varies with problems. For example, in the problem of capturing robbers, a cleared edge means that there is no robber hiding along this edge, while a contaminated edge means that a robber may be hiding along this edge. Knowing that our goal is to clear a graph (capture all robbers), a basic question is what is the fewest number of searchers required? We call this the *search number*. There are many graph searching problems motivated by applied problems or inspired by some theoretical issues in computer science and discrete mathematics. These problems are defined by the class of graphs (e.g., undirected graphs, digraphs, or hypergraphs), the actions of searchers and robbers, conditions of captures, speed, visibility, and so on.

A large number of graph searching models have been introduced. We will focus on some basic models. The reader is referred to survey papers [1–3] for a more thorough exposition of graph searching problems. Graph searching has a variety of names in the literature, such as cops-and-robber games and pursuit-evasion problems. In some problems, searchers are also called cops, sweepers, or pursuers, and robbers are called intruders, fugitives, or evaders.

All graphs and digraphs in this article are finite. We use $G = (V, E)$ ($D = (V, E)$) to denote a graph (digraph) with vertex set V and edge set E . We use uv to denote an edge in a graph with two end vertices u and v and (u, v) to denote a directed edge in a digraph with tail u and head v .

This article surveys graph searching problems on undirected graphs and digraphs. There are three major parts of the survey. The first part deals with the undirected graph searching problems in which the robber can move at very large speed. The second part deals with the digraph searching problems in which searchers and the robber may or may not obey the edge directions. The third part deals with the cops-and-robber games in which cops and the robber take turns to move.

SEARCHING UNDIRECTED GRAPHS

The first graph searching model was introduced by Parsons [4], after being approached by Breisch [5] (a paper dealing with finding a spelunker lost in a system of caves). Let G be a connected, undirected, finite graph embedded in $\mathbb{E}^3$ (i.e., $\mathbb{R}^3$ with the Euclidean distance metric) such that no pair of edges intersect at a point that is not a common endpoint. A point in G is either a vertex or an interior point of an edge in G . Imagine there is a robber in G who can be located at any point of G , that is, anywhere along an edge or at a vertex. We want to capture the robber using the minimum number of searchers. For each positive integer k , let $C_k(G)$ be the set of all families $F = \{f_1, f_2, \dots, f_k\}$ of continuous functions $f_i : [0, \infty) \rightarrow G$. A continuous search strategy for G is a family $F \in C_k(G)$ such that for every continuous function $h : [0, \infty) \rightarrow G$ (corresponding to the robber), there is a $t_h \in [0, \infty)$ and $f_i \in F$ satisfying $h(t_h) = f_i(t_h)$. We say that f_i captures the robber at time t_h . The continuous search number of a graph G is the smallest k such that there exists a search strategy for G in $C_k(G)$.

There are some contexts for which Parsons searching is the model required. For example, in the case of searching for someone lost or a robber hiding in a cave system, the robber and the searchers move continuously. But the notion of Parsons searching presents certain difficulties because the model involves the action of continuous functions defined on the nonnegative real numbers.

There are three basic discrete versions of Parsons searching: edge searching, node searching, and mixed searching, which involves placing some restriction on searchers, but placing no restrictions on a robber. In these search models, the discrete time intervals are introduced. The basic idea is that for each time interval, only one searcher can have one of the three actions: placing, removing, and sliding. Initially, G contains a robber who is located at a point in G , and G does not contain any searchers. Each searcher has no information on the whereabouts of the robber (invisible), but the robber has complete knowledge of the location of all searchers. The goal of the searchers is to capture the robber, and the goal of the robber is to avoid being captured. The robber always chooses the best strategy so that he evades capture. Suppose we start at time t_0 and the robber is captured at time t_N , and the search time is divided into N intervals $(t_0, t_1]$, $(t_1, t_2]$, ..., $(t_{N-1}, t_N]$ such that in each interval $(t_i, t_{i+1}]$ (also called *step*) exactly one searcher performs one action. The robber can move from a point x to a point y in G at any time in the interval (t_0, t_N) if there exists a path between x and y which contains no searcher. In the *edge search model* introduced by Megiddo *et al.* [6], there are three actions for searchers: placing a searcher on a vertex, removing a searcher from a vertex, and sliding a searcher along an edge from one end to the other. The robber is captured if a searcher and the robber occupy the same point on G . In the *node search model* introduced by Kirousis and Papadimitriou [7], there are two actions for searchers: placing a searcher on a vertex and removing a searcher from a vertex. The robber is captured if a searcher and the robber occupy the same vertex of G or the robber is on an edge whose endpoints

are both occupied by searchers. In the *mixed search model* introduced by Bienstock and Seymour [8], searchers have the same actions as those in the edge search model. The robber is captured if a searcher and the robber occupy the same point on G or the robber is on an edge whose endpoints are both occupied by searchers.

The edge search number of G , denoted by $es(G)$, is the smallest positive integer k such that k searchers can capture the robber. Similarly, we can define the node search number of G ($ns(G)$), and mixed search number of G ($ms(G)$). The following relationships between search numbers are given in Refs 7 and 8: $ns(G) - 1 \leq es(G) \leq ns(G) + 1$; $ms(G) \leq es(G) \leq ms(G) + 1$; and $ms(G) \leq ns(G) \leq ms(G) + 1$.

Megiddo *et al.* [6] showed that the edge search problem is NP-hard. This problem also belongs to the NP class. This follows from the monotonicity result of LaPaugh [9], in which she showed that recontamination of edges cannot reduce the number of searchers needed to clear a graph. A search strategy is monotonic if the set of cleared edges before any step is always a subset of the set of cleared edges after the step. Monotonicity is a very important issue in graph search problems. Bienstock and Seymour [8] proposed a method that gives a succinct proof for the monotonicity of the mixed search problem, which implies the monotonicity of the edge search problem and the node search problem. Fomin and Thilikos [10] provided a general framework that can unify monotonicity results in a unique minmax theorem.

Search numbers have close relationships with pathwidth and treewidth [11,12]. Given a graph $G = (V, E)$, a tree decomposition of G is a pair (T, W) with a tree $T = (I, F)$, $I = \{1, 2, \dots, m\}$, and a family of nonempty subsets $W = \{W_i \subseteq V : i = 1, 2, \dots, m\}$, such that

1. $\bigcup_{i=1}^m W_i = V$;
2. for each edge $uv \in E$, there is an $i \in I$ with $\{u, v\} \subseteq W_i$; and
3. for all $i, j, k \in I$, if j is on the path from i to k in T , then $W_i \cap W_k \subseteq W_j$.

The width of a tree decomposition (T, W) is $\max\{|W_i| - 1 : 1 \leq i \leq m\}$. The *treewidth* of G , denoted by $\text{tw}(G)$, is the minimum width over all tree decompositions of G . A tree decomposition (T, W) is a path decomposition if T is a path; the *pathwidth* of a graph G , denoted by $\text{pw}(G)$, is the minimum width over all path decompositions of G .

Another related graph parameter is the vertex separation introduced by Ellis *et al.* [13]. The following relationships are given in Kinnersley [14], Kirousis and Papadimitriou [7], and Yang [15] : $\text{sms}(G) = \text{pw}(G) = \text{vs}(G) = \text{ns}(G) - 1$, where $\text{sms}(G)$ is the strong-mixed search number of G introduced in Yang [15] and $\text{vs}(G)$ is the vertex separation of G .

Dendris *et al.* [16] introduced the *inert-robber game*, which has the same setting as the node searching model except that the robber stays only on vertices and moves just before a searcher is going to be placed on the vertex v currently occupied by the robber (given that there is a vertex u that can be reached via a searcher-free path from v to u ; otherwise the robber is captured). They showed that the inert-robber game is monotonic and the search number of a graph in the inert-robber game is equal to the treewidth of the graph plus one.

Seymour and Thomas [17] introduced another variant of graph searching, *visible-robber game*. Its setting differs from node search in that the robber stands only on vertices and is visible to searchers. They showed that the visible-robber game is monotonic and the search number of a graph in the visible-robber game is equal to the treewidth of the graph plus one. They also showed that if there is a nonlosing strategy for the robber, then there is a nice nonlosing strategy with a particularly simple form.

The problem of determining whether k searchers can capture the robber in the visible-robber game (or inert-robber game) is NP-complete because it is monotonic and determining treewidth is NP-complete [18].

There are many variants of graph searching models, for example, connected search [19,20], domination search [21], and robber-and-marshals games [22].

SEARCHING DIGRAPHS

Johnson *et al.* [23] generalized the concept of treewidth to digraphs. For a digraph $D = (V, E)$, let Z and S be two disjoint subsets of V . S is Z -normal if for every directed path, $v_1 \dots v_k$, in D such that $v_1, v_k \in S$, either $v_i \in S$ for all $1 \leq i \leq k$, or there exists $j \leq k$ such that $v_j \in Z$. Given a directed tree T with edges oriented away from a unique vertex $r \in V(T)$ (called the *root*), we write $t > e$ for $t \in V(T)$ and $e \in E(T)$ if e occurs on the unique directed path from r to t , and $e \sim t$ if e is incident with t . An arboreal decomposition of a digraph D is a triple (T, X, W) where T is a directed tree with a unique root, and $X = \{X_t \subseteq V : t \in V(T)\}$ and $W = \{W_e \subseteq V : e \in E(T)\}$ satisfy the following:

- X is a partition of V into nonempty sets.
- If $e \in E(T)$, then $\bigcup\{X_t : t \in V(T) \text{ and } t > e\}$ is W_e -normal.

The width of an arboreal decomposition (T, X, W) is the minimum k such that for all $t \in V(T)$, $|X_t \cup \bigcup_{e \sim t} W_e| \leq k + 1$. The *directed treewidth* of D is the least integer k such that D has an arboreal decomposition of width k . Johnson *et al.* [23] introduced a generalization of the visible-robber game, called the *strong directed visible-robber game*. In the strong directed visible-robber game, the searchers and robber occupy only vertices and must obey the edge directions when they move along edges. The searchers have two actions—placing and removing. The robber is visible, and the robber can move from vertex u to v if there is a searcher-free directed cycle containing u and v . Let $k \geq 1$ be an integer. A *haven* of order k in a digraph D is a function β assigning to every $Z \subseteq V(D)$ with $|Z| < k$ the vertex set of a strong component of $D \setminus Z$ in such a way that if $Z' \subseteq Z \subseteq V(D)$ with $|Z'| < k$, then $\beta(Z) \subseteq \beta(Z')$. If $k - 1$ searchers have a winning strategy on the digraph D , then D has no haven of order k . If β is a haven of order k in D , then the robber wins against $k - 1$ searchers by staying in $\beta(Z)$, where Z is the set of vertices occupied by the searchers. Johnson *et al.* [23] proved that if D has a haven of order k , then its directed treewidth

is at most $3k - 2$. Adler [24] showed that the strong directed visible-robber game is not robber-monotone and when the directed treewidth of D is at least $k - 1$, it is not implied that D has a haven of order k .

Berwanger *et al.* [25] and Obdržálek [26] introduced another complexity measure for digraphs, DAG-width. Given a digraph $D = (V, E)$, a set $X \subseteq V$ guards a set $Y \subseteq V$ if $X \cap Y = \emptyset$ and whenever there is an edge $(u, v) \in E$ such that $u \in Y$ and $v \notin Y$, then $v \in X$. A DAG-decomposition of D is a pair (T, W) with T , an acyclic digraph, and $W = \{W_d \subseteq V : d \in V(T)\}$, a family of nonempty subsets, such that

1. $\bigcup_{d \in V(T)} W_d = V$;
2. for all $d, d', d'' \in V$, if d' is on a directed path from d to d'' in D , then $W_d \cap W_{d''} \subseteq W_{d'}$;
3. for all edges $(d, d') \in E(T)$, $W_d \cap W_{d'}$ guards $W_{\geq d'} \setminus W_d$, where $W_{\geq d'} = \bigcup \{W_{d''} : d'' \in V(T) \text{ and there is a directed path from } d' \text{ to } d'' \text{ in } T\}$; for any root d , $W_{\geq d}$ is guarded by $\emptyset$.

The width of a DAG-decomposition (T, W) is $\max\{|W_d| : d \in V(T)\}$. The DAG-width of D is the minimum width over all DAG-decompositions of D . Corresponding to the DAG-width, Berwanger *et al.* [25] and Obdržálek [26] introduced the *directed visible-robber game*, which has the same setting as the strong directed visible-robber game except that the robber can move from vertex u to v if there is a searcher-free directed path from u to v . They proved that for any digraph D , k searchers have a monotone winning strategy if and only if D has a DAG-decomposition of width k . They also proved that if a graph has DAG-width k , its directed treewidth is at most $3k + 1$. However, there are graphs with directed treewidth 1 and arbitrarily large DAG-width.

Hunter and Kreutzer [27] extended the inert-robber game to the *directed inert-robber game* on digraphs, which has the same setting as the inert-robber game on graphs except that the searchers and robber occupy only vertices and must obey the edge directions when they move along edges. They introduced another complexity measure for

digraphs (Kelly-width), and showed that k searchers have a monotone winning strategy to capture an inert robber in a digraph D if and only if D has Kelly-width k . For the relation between the directed inert-robber game and the directed visible-robber game, if k searchers can capture an inert robber in a digraph D with a robber-monotone strategy, then $2k - 1$ searchers can capture a visible robber in D . For every integer k , there is a digraph such that $3k$ searchers have a robber-monotone winning strategy in the directed inert-robber game but no fewer than $4k$ searchers can capture a visible robber in the directed visible-robber game.

Kreutzer and Ordyniak [28] proved that both directed visible-robber games and directed inert-robber games are nonmonotonic. In particular, within the directed visible-robber game setting, for every $p \geq 2$ there exists a digraph D_p such that the monotonic search number of D_p is $4p - 2$ and the search number of D_p is $3p - 1$. Within the directed inert-robber game setting, for every $p \geq 2$ there exists a digraph D_p such that the monotonic searcher number of D_p is $7p$ and the search number of D_p is $6p$. The problem of determining whether k searchers can capture the robber in the directed visible-robber game (respectively, directed inert-robber game or strong directed visible-robber game) is NP-hard.

Nowakowski [29] and Yang and Cao [30–32] generalized the edge searching problem to digraphs by introducing three digraph search models: directed search, strong search, and weak search. In all three models, the robber is invisible and searchers have three types of actions: placing, removing, and sliding. These models differ in the abilities of the searchers and robber depending on whether or not they must obey the edge directions. In the *directed search* model, both searchers and robber must move in the edge directions; in the *strong search* model, the robber must move in the edge directions but searchers need not; and in the *weak search* model, searchers must move in the edge directions but the robber need not. In particular, in the directed and strong search models, the robber can move from vertex u to vertex v along a searcher-free directed

path from u to v at a great speed at any time; in the weak search model, the robber can move from vertex u to vertex v along a searcher-free undirected path between u and v at a great speed at any time.

For a digraph D , let $ds(D)$, $ss(D)$, and $ws(D)$ denote the directed, strong, and weak search numbers, respectively. Let $es(D)$ be the edge search number of the underlying undirected graph of D . Yang and Cao [30–32] proved that the directed, strong, and weak search problems are all monotonic and NP-complete. They gave the following relationships between search numbers: $ss(D) \leq ds(D) \leq ws(D)$, $ss(D) \leq es(D) \leq ws(D)$, $ds(D) - 1 \leq ss(D) \leq ds(D)$, $ds(D) \leq es(D) + 1$, and $ws(D) \leq es(D) + 2$.

Reed, Seymour, and Thomas introduced the directed pathwidth. Given a digraph $D = (V, E)$, a directed path decomposition of D is a sequence of subsets of vertices $W_1, W_2, \dots, W_m$ such that

1. $\bigcup_{i=1}^m W_i = V$;
2. for $1 \leq i < j < k \leq m$, $W_i \cap W_k \subseteq W_j$;
3. for each edge in D , it either has both endpoints in the same W_i or has its tail in W_i and head in W_j , where $i < j$.

The width of a directed path decomposition of D is $\max_{1 \leq i \leq m} (|W_i| - 1)$. The *directed pathwidth* of D , denoted by $dpw(D)$, is the minimum width over all directed path decompositions of D .

Yang and Cao [33] extended the vertex separation to digraphs. Given a digraph $D = (V, E)$ and a linear layout $L : V \rightarrow \{1, 2, \dots, |V|\}$, let $DV_L(i) = \{x \in V : \text{there exists } y \in V \text{ such that edge } (y, x) \in E \text{ and } L(x) \leq i \text{ and } L(y) > i\}$. The directed vertex separation of D with respect to L , denoted by $dvs_L(D)$, is defined as $dvs_L(D) = \max\{|DV_L(i)| : 1 \leq i \leq |V|\}$. The *directed vertex separation* of D is defined as $dvs(D) = \min\{dvs_L(D) : L \text{ is a linear layout of } D\}$. The following relationships are given in Yang and Cao [33]: $dvs(D) = dpw(D)$, $dvs(G) + 1 \leq ds(G) \leq dvs(G) + 2$, and $dvs(G) \leq ss(G) \leq dvs(G) + 2$.

There are several other searching models on digraphs, such as arc searching [34] and

directed node searching [35]. See also the thesis of Hunter for more discussion [36].

COPS-AND-ROBBER GAMES

A different branch of research on graph searching is the cops-and-robber game, in which cops first locate themselves on the vertices of a graph, and then a robber chooses a vertex; all participants have complete information on the location of all other participants. At even ticks of a clock (starting at 0), some subset of the cops move to adjacent vertices, and, on odd ticks of the clock, the robber moves to an adjacent vertex or stays at his current vertex. Capture takes place when a cop and the robber occupy the same vertex at the same time. This game was introduced by Nowakowski and Winkler [37], and independently by Quilliot [38]. They presented a nice characterization of 1-cop-win graphs. The characterization involves an order relation defined on the vertex set of a graph. We use $N(u)$ to denote the set of all neighbors of u and $N[u] = N(u) \cup \{u\}$. Let $V(G) = \{v_1, v_2, \dots, v_n\}$ and $v \in V(G)$. Define $N_i[v] = N[v] \cap \{v_j : i \leq j \leq n\}$. Nowakowski and Winkler showed that a connected graph is 1-cop-win if and only if its vertices can be linearly ordered $v_1, v_2, \dots, v_n$ such that for each $1 \leq i < n$, there is a j , $i < j \leq n$, satisfying $N_i[v_i] = N_i[v_j]$. A graph is bridged if each of its cycles of length four or more has a shortcut, that is, a pair of vertices whose distance in the graph is shorter than their distance on the cycle. Anstee and Farber [39] gave another characterization of 1-cop-win graphs. They proved that a connected graph is bridged if and only if it is 1-cop-win and contains no induced cycles of length four or five.

The cop number of a graph G , denoted $cn(G)$, is the smallest k such that k cops can capture any robber in the cops-and-robber game. Frankl [40] proved that $cn(G) > d^t$ for a connected graph G with girth at least $8t - 3$ and minimum degree at least $d + 1$. Schroeder [41] showed that for a graph G of orientable genus k , $cn(G) \leq \lfloor \frac{3k}{2} \rfloor + 3$.

Excluding some minors can guarantee a small cop number. Let u be a vertex of a graph

H such that $H - u$ has no isolated vertices. Andreae [42] proved that if G is a graph with no H -minor, then $\text{cn}(G) \leq |E(H - u)|$.

Given a graph containing k cops and one robber, Goldstein and Reingold [43] showed that the problem of determining if k cops can capture a robber from the given initial positions is EXPTIME-complete. For a constant k , they showed that it is polynomial to decide if a graph is k -cop win. Hahn and MacGillivray [44] extended this result by giving a polynomial time algorithm that determines whether k cops can capture l robbers on a graph or digraph, where both k and l are constants.

There are many bounds on cop numbers on different classes of graphs. See Hahn [45] and references therein for details. Several variants of the cops-and-robber game can be found in Clarke and Nowakowski [46,47] and Fomin *et al.* [48].

Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments and suggestions, which improved the presentation of this article. Research was supported in part by NSERC.

REFERENCES

1. Fomin F, Thilikos D. An annotated bibliography on guaranteed graph searching. Theoretical Computer Science 2008;399:236–245.
2. Alspach B. Searching and sweeping graphs: a brief survey. Le Matematiche 2006;59:5–37.
3. Bienstock D. Graph searching, path-width, tree-width and related problems (a survey). DIMACS Ser Discrete Math Theor Comput Sci 1991;5:33–49.
4. Parsons TD. Pursuit-evasion in graphs. In: Alavi Y, Lick DR, editors. Theory and applications of graphs. Berlin: Springer; 1976. pp. 426–441.
5. Breisch R. An intuitive approach to speleotopology. Southwest Cavers 1967;6:72–78.
6. Megiddo N, Hakimi SL, Garey M, *et al.* The complexity of searching a graph. J ACM 1998; 35:18–44.
7. Kirousis LM, Papadimitriou CH. Searching and pebbling. Theor Comput Sci 1986; 47:205–216.
8. Bienstock D, Seymour P. Monotonicity in graph searching. Journal of Algorithm 1991; 12:239–245.
9. LaPaugh AS. Recontamination does not help to search a graph. J ACM 1993;40:224–245.
10. Fomin F, Thilikos D. On the monotonicity of games generated by symmetric submodular functions. Discrete Appl Math 2003;131: 323–335.
11. Robertson N, Seymour P. Graph minors I: excluding a forest. J Comb Theory Ser B 1983;35:39–61.
12. Robertson N, Seymour P. Graph minors II: algorithmic aspects of tree-width. J Algorithm 1986;7:309–322.
13. Ellis J, Sudborough I, Turner J. The vertex separation and search number of a graph. Inf Comput 1994;113:50–79.
14. Kinnersley NG. The vertex separation number of a graph equals its path-width. Inf Process Lett 1992;42:345–350.
15. Yang B. Strong-mixed searching and path-width. J Comb Optim 2007;13:47–59.
16. Dendris N, Kirousis L, Thilikos D. Fugitive-search games on graphs and related parameters. Theor Comput Sci 1997;172:233–254.
17. Seymour P, Thomas R. Graph searching and a min-max theorem for tree-width. J Comb Theory Ser B 1993;58:22–33.
18. Arnborg S, Corneil D, Proskurowski A. Complexity of finding embeddings in a k -tree. SIAM J Algebra Discrete Methods 1987;8: 277–284.
19. Barrière L, Fraigniaud P, Santoro N *et al.* Searching is not jumping. Proceedings of the 29th Workshop on Graph Theoretic Concepts in Computer Science (WG'03). Lecture Notes in Computer Science 2880. Berlin: Springer; 2003. pp. 34–45.
20. Yang B, Dyer D, Alspach B. Sweeping graphs with large clique number. Discrete Math 2009;309:5770–5780.
21. Fomin F, Kratsch D, Müller H. On the domination search number. Discrete Appl Math 2003;127:565–580.
22. Gottlob G, Leone N, Scarcello F, *et al.* game theoretic and logical characterizations of hypertree width. J Comput Syst Sci 2003; 66:775–808.
23. Johnson T, Robertson N, Seymour P, *et al.* Directed tree-width. J Comb Theory Ser B 2001;82:138–154.
24. Adler I. Directed tree-width examples. J Comb Theory Ser B 2007;97:718–725.

25. Berwanger D, Dawar A, Hunter P, *et al.* DAG-width and parity games. Proceedings of the 23rd Symposium of Theoretical Aspects of Computer Science (STACS'06). Lecture Notes in Computer Science 3884. Berlin: Springer; 2006. pp. 524–436.
26. Obdržálek J. DAG-width: connectivity measure for directed graphs. Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'06). Boston: ACM; 2006. pp. 814–821.
27. Hunter P, Kreutzer S. Digraph measures: kelly decompositions, games, and orderings. *Theor Comput Sci* 2008;399:206–219.
28. Kreutzer S, Ordyniak S. Digraph decompositions and monotonicity in digraph searching. Proceedings of the 34th International Workshop on Graph-Theoretic Concepts in Computer Science (WG'08). Lecture Notes in Computer Science 5344. Berlin: Springer; 2008. pp. 336–347.
29. Nowakowski R. Search and sweep numbers of finite directed acyclic graphs. *Discrete Appl Math* 1993;41:1–11.
30. Yang B, Cao Y. Monotonicity in digraph search problems. *Theor Comput Sci* 2008;407:532–544.
31. Yang B, Cao Y. Monotonicity of strong searching on digraphs. *J Comb Optim* 2007;14:411–425.
32. Yang B, Cao Y. On the monotonicity of weak searching. Proceedings of the 14th International Computing and Combinatorics Conference (COCOON'08). Lecture Notes in Computer Science 5092. Berlin: Springer; 2008. pp. 52–61.
33. Yang B, Cao Y. Digraph searching, directed vertex separation and directed pathwidth. *Discrete Appl Math* 2008;156:1822–1837.
34. Alspach B, Dyer D, Hanson D, *et al.* Arc searching digraphs without jumping. Proceedings of the 1st International Conference on Combinatorial Optimization and Applications (COCOA'07). Lecture Notes in Computer Science 4616. Berlin: Springer; 2007. pp. 354–365.
35. Barat J. Directed path-width and monotonicity in digraph searching. *Graphs Comb* 2005;22:161–172.
36. Hunter P. Complexity and infinite games on finite graphs [PhD thesis]. University of Cambridge, Computer Laboratory; 2007.
37. Nowakowski R, Winkler P. Vertex-to-vertex pursuit in a graph. *Discrete Math* 1983;43:235–239.
38. Quilliot A. Problemes de jeux, depoint fixe, de connectivité et de représentation sur des graphes, des ensembles ordonnés et des hypergraphes. 1983.
39. Anstee R, Farber M. On bridged graphs and cop-win graphs. *J Comb Theory Ser B* 1988;44:22–28.
40. Frankl P. Cops and robbers in graphs with large girth and Cayley graphs. *Discrete Appl Math* 1987;17:301–305.
41. Schroeder B. The copnumber of a graph is bounded by $\lfloor \frac{3}{2} \text{genus}(G) \rfloor + 3$. In: Koslowski J, Melton A, editors. *Categorical perspectives*. Trends in Mathematics Boston (MA): Birkhäuser Boston; 2001. pp. 243–263.
42. Andrae T. On a pursuit game played on graphs for which a minor is excluded. *J Comb Theory Ser B* 1986;41:37–47.
43. Goldstein A, Reingold E. The complexity of pursuit on a graph. *Theor Comput Sci* 1995;143:93–12.
44. Hahn G, MacGillivray G. A note on k -cop, l -robber games on graphs. *Discrete Math* 2006;306:2492–2497.
45. Hahn G. Cops, robbers and graphs. *Tatra Mt Math Publ* 2007;36:1–14.
46. Clarke N, Nowakowski R. Cops, robber, and photo radar. *Ars Comb* 2000;56:97–103.
47. Clarke N, Nowakowski R. Tandem-win graphs. *Discrete Math* 2005;299:56–64.
48. Fomin F, Golovach P, Hall A, *et al.* How to guard a graph? Proceedings of 15th International Symposium on Algorithms and Computation (ISAAC'08). Lecture Notes in Computer Science 5369. Berlin: Springer; 2008. pp. 318–329.

GRAPHICAL METHODS FOR RELIABILITY DATA

BARIS SURUCU

Department of Statistics,
Middle East Technical University,
Ankara/Turkey

HAKAN S. SAZAK

Department of Statistics,
Ege University, Bornova,
İzmir/Turkey

Graphical methods for reliability analysis have received considerable attention in the literature. In statistical data analysis, graphical techniques are of great importance, as they are very effective and practical tools in summarizing, describing, and presenting data [1–3]. Nowadays, statistical software provide an extensive range of parametric and nonparametric graphical approaches; see, for example, Meeker [4]. Parametric techniques enable model verification, estimation of parameters, and outlier-inlier detection under many statistical distributions such as Weibull, exponential, extreme value, normal, lognormal, and gamma. Although parametric approaches result in high efficiency when underlying distributions are true, assuming a wrong distribution may yield inefficient and misleading results. Nonparametric techniques, which do not require any distributional assumption, generally use empirical functions based on sample observations. Nonparametric techniques are advantageous in the sense that they are simple to carry out and they lead to fast and straightforward summarization of data. On the other hand, they do not make use of distributional structures, which may cause information loss and lack of efficiency, especially for small sample sizes.

In this article, we review some existing nonparametric and parametric graphical methods for reliability data. We first give general definitions and concepts in the area. Then, we discuss some well-known

basic graphical techniques including empirical cumulative distribution function (cdf) plot, hazard plot, probability-probability (P-P) plot, quantile-quantile (Q-Q) plot, the Weibull plot, the Duane plot, mean cumulative function (MCF) plot, and Total Time on Test (TTT) plot. We also describe some problem-specific graphical techniques that explore different applications in the reliability area.

GENERAL OVERVIEW OF GRAPHICAL TECHNIQUES

With the development of technology and industry, more complicated systems have emerged increasing the need for reliability assessment of components and systems. Owing to their easiness of use and practical interpretations, graphical methods have appeared to become one of the most effective assessment tools for analyzing reliability data, starting from the second half of the past century. More emphasis was placed on estimating life distributions and their parameters [1,2,5–10] as well as reliability monitoring through survival curve [11], hazard plotting [12,13–14], the Duane plot [15], and MCF plot [16,17]. For system reliability, see the work by Proctor and Singh [18]. Some other related graphical methods assessing distributional assumptions, such as P-P plotting, were discussed in various works; see, for example, Lawless [19] and Nelson [20]. Besides these, some generalizations of TTT were developed following the initial works by Epstein and Sobel [21] and Barlow and Campo [22]; see Bergman and Klefsjo [23] for details. Toward the end of the past century, the concept of reliability growth models was of much interest to researchers and a number of reliability growth models were proposed to predict and continuously improve the reliability of a component or a system [24–26].

Recent literature mainly revolves around these basic concepts. Many useful variants

of these graphical techniques pertaining to reliability data have been proposed, and new graphical methods have been developed for other practical applications in the field. The succeeding sections summarize these methods and the related literature.

GRAPHICAL TECHNIQUES

Determining Probability Distributions

Many studies in the literature consider parametric graphical approaches to analyze and display reliability data visually. A number of graphical methods have been developed to determine the underlying distributional structure. These easy-to-use methods are summarized in the following sections.

Empirical cdf. Empirical cdf plot is one of the basic and useful tools in examining the behavior of the distribution. One can describe the shape of the sample data by simply plotting the empirical cdf against failure times or lifetimes of components. The theoretical cdf of the hypothesized distribution for the

observations can also be depicted on the same plot to see how well the assumed distribution fits the data. Figure 1 shows an example of an empirical cdf plot (together with the theoretical cdf) of a data set simulated from the Weibull distribution with the cdf,

$$F(x) = P(X \leq x) = 1 - e^{-(x/\sigma)^p} \quad (1)$$

$$x > 0, \sigma > 0, p > 0$$

where σ and p are the scale and the shape parameters, respectively. For more references, see King [27], Lawless [19], Nelson [20], and Abernethy [28].

Probability-Probability (P-P) Plot. The P-P plot is used to check the validity of the assumed distribution. It is obtained by plotting the empirical cdf of data against the specified theoretical cdf. The construction of a P-P plot necessitates that the location and scale parameters of the specified distribution be known. The linearity of the line gives a strong evidence for the validity of the specified distribution; see Wilk and Gnanadesikan [29] and Gnanadesikan [30]. It is also possible

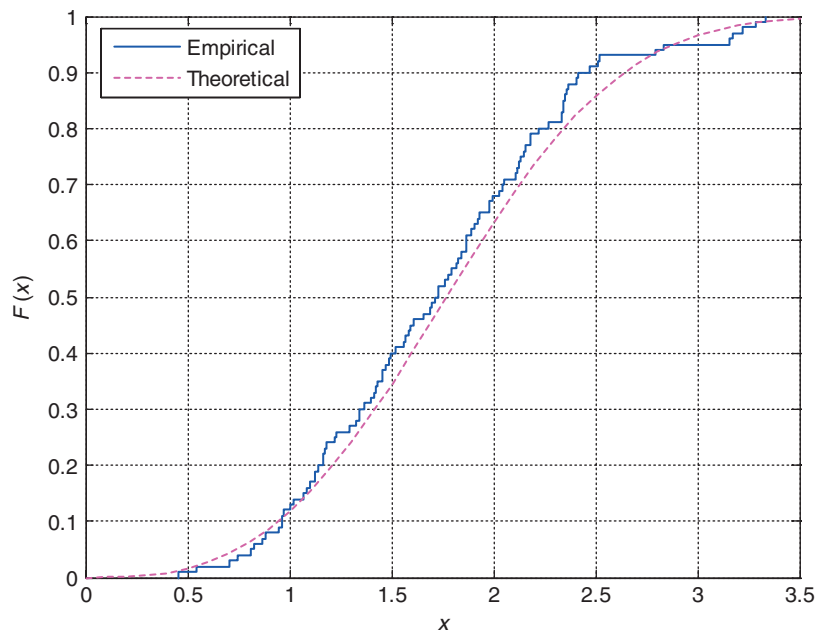


Figure 1. Empirical cdf plot of the simulated data set from Weibull ($\sigma = 2, p = 3$); $n = 100$.

to obtain a P-P plot for censored data as one only needs to plot uncensored observations against their empirical cdf values.

There are many applications and variants of P-P plots in the literature. For example, Chang and Tang [31] studied a simple graphical technique that is analogous to probability plotting for the Birnbaum–Saunders distribution for failure time modeling of fatigue process. They used P-P plot for goodness-of-fit for the Birnbaum–Saunders distribution as well as for the estimation of parameters. See also Wang *et al.* [32] for the use of P-P plot in managing improvements in the strengths of Oriented Strand Board.

Quantile-Quantile (Q-Q) Plot. The Q-Q plot is another type of plot for the verification of the assumed distribution. The horizontal axis consists of the expected values of the standardized order statistics of the assumed distribution and the vertical axis consists of the ordered observations. Figure 2 shows the Q-Q plot of a data set under the assumption of the Weibull distribution with shape parameter 3. The linearity is an evidence of

the true underlying distribution. It should be noted that the shape parameter of the Weibull distribution should be specified to plot the expected values of the standardized distribution ($\sigma = 1$). On the other hand, one does not need to specify the parameters of a location-scale distribution to obtain a Q-Q plot, as they do not affect its linearity.

The Weibull Plot. In the field of reliability, it is quite common to assume Weibull, exponential, extreme value, normal, lognormal, and gamma as the distributions of lifetimes or failure times of components. Among these, especially the Weibull distribution gained a lot of attention in evaluating reliability. Numerous data sets have been analyzed assuming two-parameter Weibull distribution. To assess the validity of a two-parameter Weibull distribution, Weibull plot is often utilized. It is a combination of P-P and Q-Q plots because the graph is obtained by plotting the empirical cdf against the order statistics with scaled axis. For linearity purposes, $\ln[-\ln(1 - F(x))]$ and $\ln x$ can also be used in the axes; see Nelson

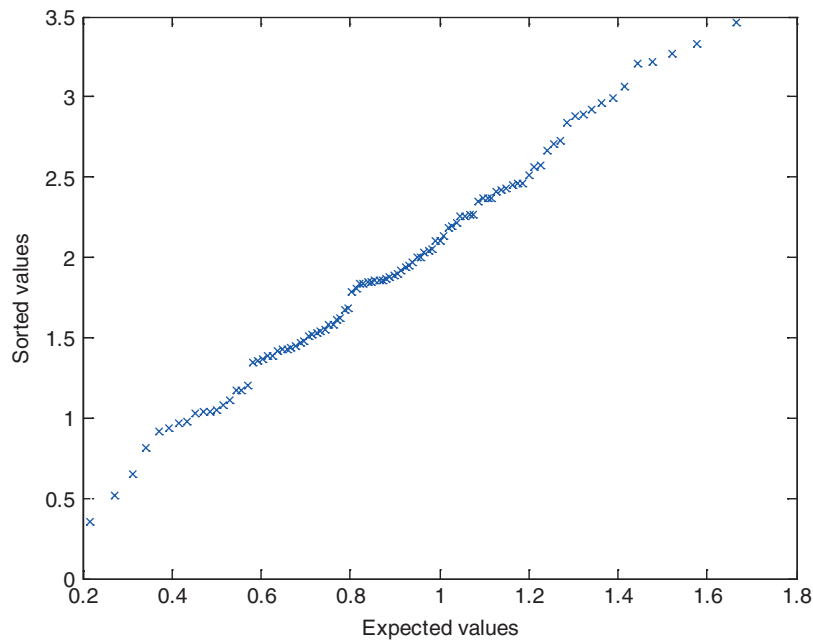


Figure 2. Q-Q plot of the simulated data set from Weibull ($\sigma = 2$, $p = 3$); $n = 100$.

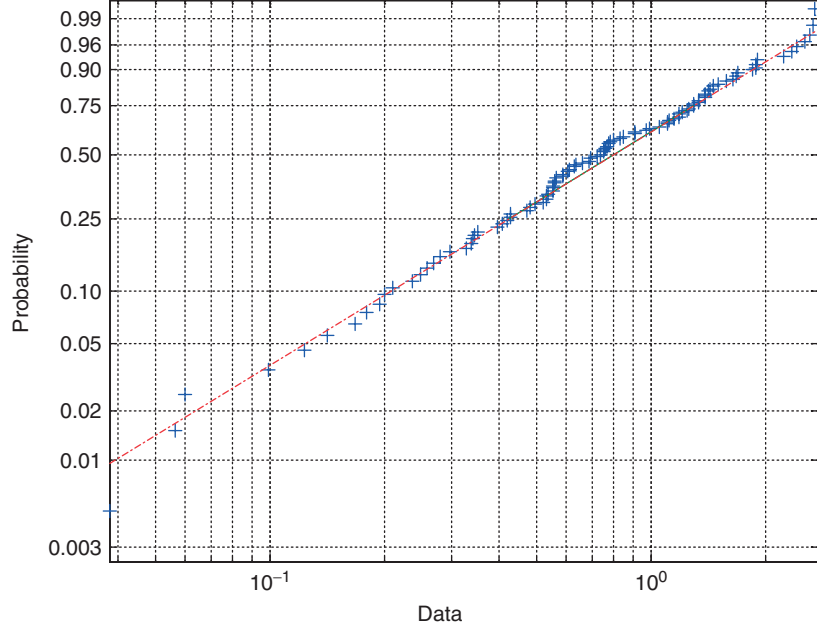


Figure 3. Weibull plot of the simulated data set from Weibull ($\sigma = 1, p = 1.5$); $n = 100$.

[20] for more details. Figure 3 shows an example of the Weibull plot obtained from a simulated data with $\sigma = 1$ and $p = 1.5$.

Apart from the classical Weibull plot, Hartler [33] suggested the use of Weibull process plot that works for repairable systems failing according to a Weibull process. Hartler [33] stated that the new graphical method is applicable to preliminary studies of system failure pattern, parameter estimation, early failure detection, failure pattern comparison, and determination of the failure process. In another work by Jiang and Kececioğlu [34], the performances of the two estimation methods based on the Weibull plot for different mixed-Weibull distributions were explored. Using an additive model, Xie and Lai [26] combined two Weibull distributions with decreasing and increasing failure rates to obtain a combined distribution having a bathtub shape. They utilized a simple graphical technique similar to the Weibull plot. The technique is also useful for assessing the goodness-of-fit of the distribution and estimating the unknown

distribution parameters (see also Zhao *et al.* [35] and Zhang and Xie [36]).

ESTIMATING AND MONITORING RELIABILITY

Estimating and monitoring the reliability of components or a system have become quite crucial since it reflects almost all the information about the ongoing process. Many effective graphical tools have been suggested in the literature for various different purposes. These tools are discussed in the following sections.

Hazard Plots

The hazard rate of a component or a system at time t is defined as

$$h(t) = \frac{f(t)}{R(t)}, \quad t > 0, \quad (2)$$

where

$$R(t) = P(X > t) = 1 - F(t), \quad t > 0; \quad (3)$$

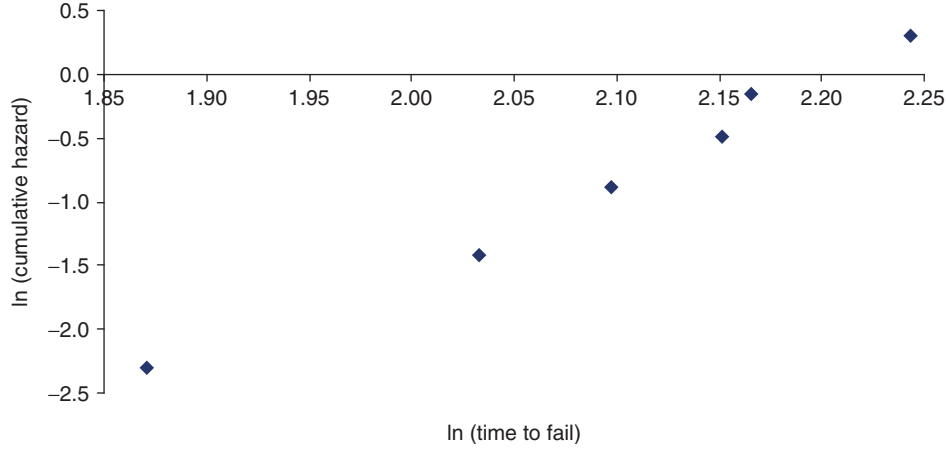


Figure 4. Hazard plot of a censored Weibull data set.

$F(t)$ being the cdf. It shows the instantaneous failure rate of a component or a system. The cumulative hazard rate is also often used in reliability analyses and is given by

$$H(t) = \int_0^t h(u)du. \quad (4)$$

Similar to P-P and Q-Q plotting, hazard plotting is a graphical goodness-of-fit method to assess the distributional structure of the data and is obtained by plotting times of failure against the empirical cumulative hazard rates on a log-log scale. Figure 4 shows an example of a hazard plot of a simulated censored Weibull data set having a sample size of 10, 4 of which are censored. The linearity is an evidence for the Weibull distribution.

Numerous papers discussed the use of hazard plotting in the literature. Nelson [14], for example, presented hazard plots for left-truncated lifetime data with illustrations from a real-life data set on compressors in heat pumps. More recently, Persson *et al.* [37] proposed a new method resembling the Kolmogorov–Smirnov method for the comparison of two groups with different hazard functions and conducted a simulation study to compare the effectiveness of six graphical techniques in identifying nonproportionality of hazards. They reported that the Arjas plot

is the best performing one among the six graphical techniques.

The Kaplan–Meier Plot

Graphical techniques are utile for estimating reliability function when the sample is of complete or censored type. The most well-known method of estimating the reliability or survival function is the Kaplan–Meier estimator; see Kaplan and Meier [11] for details. As a graphical tool, the Kaplan–Meier plot is one of the practical and effective graphical techniques in estimating reliability. It is especially advantageous for censored reliability data, as it can handle censored structures. It is obtained by plotting time against the estimated survival function. In Figure 5, an example of the Kaplan–Meier plot of a right censored data set generated from a Weibull distribution is displayed.

There are some other works that considered the problem of estimating reliability. Zhao *et al.* [35] studied storage reliability estimation. They initially considered nonparametric estimation for current reliability and then carried out parametric estimation based on the Weibull plot. Smooth estimation of the reliability function was achieved by Kulasekera *et al.* [38]. On the basis of many graphical results, they examined the performance of the Kaplan–Meier estimator smoothed with a suitable kernel estimator. Martínez *et al.* [39] proposed a

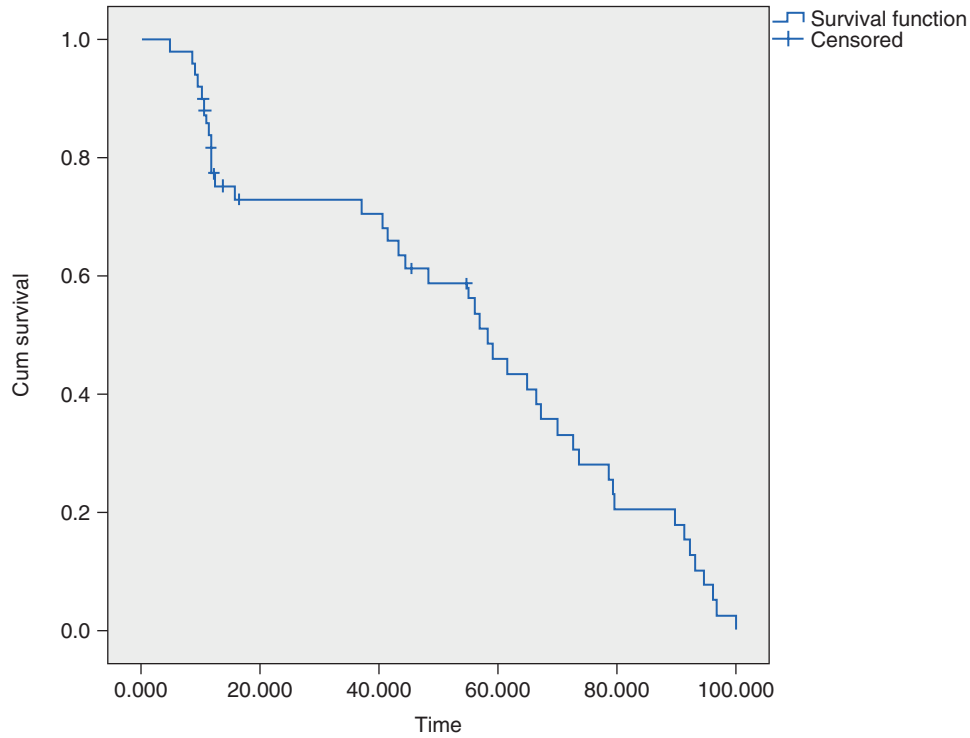


Figure 5. The Kaplan–Meier plot of data from Weibull.

support tool for maintenance management decision making using a nonparametric estimator for the reliability function. A number of graphs and diagrams were provided for visualization.

Reliability Growth Models

Many reliability data can be analyzed under the general framework of the nonhomogeneous Poisson process (NHPP) with a mean value function, $m(t)$, defined as the expected number of failures in $[0, t]$ [24]. The NHPP models are often utilized for analyzing reliability growth. Reliability growth analyses are quite useful to see the improvement in a product's reliability over time. It is also helpful in predicting future reliability of components or systems.

For analyzing reliability growth, the simplest and commonly used NHPP model is the Duane model [15]. Duane proposed a graphical approach that is based on plotting the

times between failures and the number of failures on a log-log scale. The plot tends to give a straight line if the assumed model is valid. One can also take the logarithm of the values and use normal scales. The plotted line also enables the estimation of parameters. Figure 6 shows the Duane plot of the data set by Currit *et al.* [40] based on a power-law model ($m(t) = at^b$); see Xie and Zhao [24]. The straight line provides a strong evidence that the power-law model is quite convenient, although some other models for this data set were also suggested by Xie and Zhao [24].

Xie and Zhao [24,25] proposed the use of the plotting the cumulative number of failures against the running time depending on the Duane model and other various models. They suggested using the plots for model verification and then for the estimation of parameters by giving illustrations for many real-life data sets. Wang [41] suggested a nonparametric graphical method to interpret

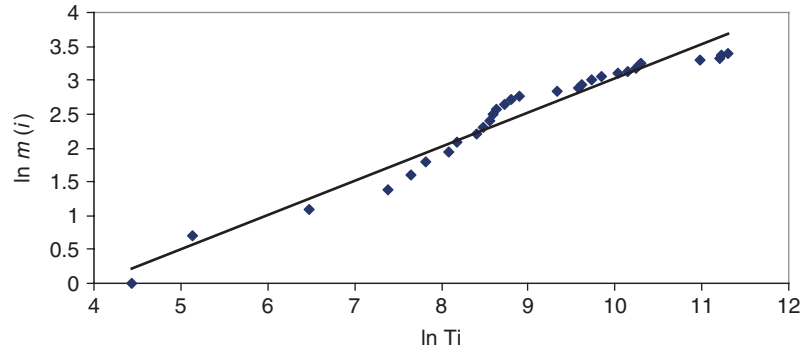


Figure 6. The Duane plot of data set by Currit *et al.* [40].

ill-collected data for repairable systems in vehicles. He presented the Mean Cumulative Repair Function plot, reliability growth plot based on power-law model, and reliability function plots based on the suggested method. Donovan and Murpy [42] proposed a new reliability growth model and developed the instantaneous mean time between failure (MTBF) equation for the new model. In another work by the same authors, some advantages of the new model over the Duane model were reported [43]. For more references about reliability growth models, see Codier [44].

Mean Cumulative Function (MCF) Plot

Some reliability data have repetition times of failures of repairable components. Since classical tools are not enough to describe, model, and analyze such data, methods that take into account the repetitive feature of the system are required. $M(t)$ (MCF) is the mean number (or cost) of recurrences up to time t , and the MCF plot is a useful and effective tool in expressing, modeling, and analyzing recurrence data. Nelson [16] suggested the use of nonparametric estimate of MCF in constructing its plot against time. It can be used for censored data and does not need any assumption. The plot is very simple to draw and can be used to estimate failure rates and other characteristics such as burn in time. See also Nelson [45,46] for a number of examples of the application of graphical analysis of repair data.

As for the reliability monitoring, control charting has been a popular graphical tool in the past decade. There have been quite a few works that studied the problem of monitoring failure data. Zhang *et al.* [47] considered exponential control charts to monitor the failure process of a reliability component or system. Batson *et al.* [48] utilized control charts based on normal, lognormal, exponential, and Weibull distributions to monitor field failure data. Xie *et al.* [49] implemented a control chart to monitor failure processes. Sürücü and Sazak [50] proposed a method to monitor the cumulative time elapsed between r failures using the three-parameter Weibull distribution. They also provided control charts and average run length curves to show the usefulness of the method in the reliability monitoring step.

Total Time on Test (TTT) Plot

TTT statistics were first introduced by Epstein and Sobel [21] who worked with exponential order statistics in life testing problems. Later on, the approach was applied in analyzing failure data for repairable and nonrepairable systems, aging analysis, maintenance planning, and dependent failure data analysis, starting with the works by Barlow *et al.* [51], Barlow and Campo [22], Barlow and Davis [52], and Barlow [53]. Bergman and Klefsjo [23] analyzed maintenance periods of repairable equipments by TTT plotting. Høyland and Rausand [54] utilized the TTT plot for data from nonrepairable systems. Kumar *et al.* [55] aimed

to estimate the operational reliability of machines in a Swedish mine and obtain the duration of optimal preventive maintenance by using TTT and double-TTT plots. Rao and Prasad [56] discussed the use of TTT plots for maintenance interval estimation for failure data having increasing failure rate. They also considered the TTT plots including covariate and showed its use with an example. Further theory and applications for the TTT plotting can be seen in a number of papers by Klefsjo and Westberg [57], Kumar and Westberg [58], Kvaloy and Lindqvist [59], Sun and Kececioglu [60], and Lai and Xie [61]. A generalization of the TTT plot was proposed by Kunitz and Pamme [62,63] who introduced aging plot for the identification of life distributions having bathtub shapes.

To explain the basic concept of TTT plotting, consider the following failure times of a certain component or a system: $0 \leq t_0 \leq t_1 \leq t_2 \leq \dots \leq t_n$, where t_i is the i th ordered failure time in a sample of size n . To see if there is a trend in failure structure, define the total time on the test at time t_i as

$$T_i = n \int_0^{t_i} R(t) dt, \quad i = 1, 2, \dots, n. \quad (5)$$

The scaled TTT plot is obtained by plotting S_i versus $F(t_i)$ where $S_i = \frac{T_i}{T_n}$.

The total time on the test at t_i is defined by

$$T_i = \sum_{j=1}^i (n - j + 1) (t_j - t_{j-1}), \quad i = 1, 2, \dots, n. \quad (6)$$

Then, the TTT plot is obtained by plotting S_i against the empirical cdf, $F_e(t_i) = i/n$. For other TTT plots based on different estimation methods of the survival function $R(t)$, readers can refer to Sun and Kececioglu [60] and Rao and Prasad [56].

OTHER GRAPHICAL METHODS

Many other studies have been published for analyzing reliability data through graphical methods. They are summarized below.

Balagurusamy and Misra [64] proposed a graphical method for the determination of the number of active redundant units and the computation of reliability parameters in an m -order system. Newby [65] utilized some graphical methods based on the aging properties of a number of failure distributions. Launer [66] presented graphical data analysis techniques based on the α -percentile residual life function [67,68] to study the empirical behavior of the hazard rate. Nicol *et al.* [69] described and discussed a graphical-model-based reliability estimation tool in their work. To study dependent failures, Walls [70] developed a graphical technique to estimate the reliability of repairable systems. Liu and Chiou [71] made use of Petri nets for failure analysis. Burrell [72] used a reliability plot technique to model citation data. The theory of limited expected value function was considered and extended by Quigley and Walls [73] for model identification of censored data. They utilized many graphical methods for their analysis. Trindade and Nathan [74] worked on field data and proposed some simple graphical methods for the reliability analysis of repairable systems. Halim and Tang [75] studied age heterogeneity for the MCF in repairable systems and presented some graphical representations. Different from other works, Donat *et al.* [76–78] studied probabilistic graphical models known as *graphical duration models* for analyzing reliability. Halim and Tang [79] suggested another graphical approach to obtain the confidence interval for the optimal preventive maintenance. Al-Garni *et al.* [80] used cumulative and MCF plots for field failures of aircraft systems. In a recent work by Saleh and Castet [81], many graphical methods were utilized to study the spacecraft reliability. For other related literature, the reader can refer to Hovis [82], Bergman and Klefsjo [83], Martin [84], Meeker and Escobar [85], Reineke *et al.* [86], and Nelson [87,88].

CONCLUDING REMARKS

Graphical methods are very useful tools in statistical analysis. They have widespread

use especially in the reliability area. Many analyses such as model verification and estimation of parameters for failure distributions are effectively conducted through these methods. In this article, we have presented the literature on graphical methods for reliability data, including the recent developments in the area.

REFERENCES

1. Cox DR. Some remarks on the role in statistics of graphical methods. *Appl Stat* 1978;27:4–9.
2. Cox DR. A note on the graphical analysis of survival data. *Biometrika* 1979;66:188–190.
3. Gentleman R, Crowley J. Graphical methods for censored data. *J Am Stat Assoc* 1991;86:678–683.
4. Meeker WQ. Trends in the Statistical Assessment of Reliability. Unpublished manuscript; 2010.
5. Kao JHK. A graphical estimation of mixed Weibull parameters in life-testing of electron tubes. *Technometrics* 1959;1:389–407.
6. Herd GR. Estimation of reliability from incomplete data. *Proceedings of the 6th National Symposium on Reliability and Quality Control*. New York: IEEE; 1960. p 202–217.
7. Johnson LG. *The Statistical Treatment of Fatigue Experiments*. New York: Elsevier; 1964.
8. Nelson W. Graphical analysis of accelerated life test data with a mix of failure models. *IEEE Trans Reliab* 1975;24(10): 230–237.
9. Cran GW. Graphical estimation methods for Weibull distributions. *Microelectron Reliab* 1976;15:47–52.
10. Jensen F, Petersen NE. *Burn-in: An Engineering Approach to the Design and Analysis of Burn-in Procedures*. Chichester: Wiley; 1982.
11. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–481.
12. Nelson W. Graphical analysis of accelerated life test data with the inverse power law model. *IEEE Trans Reliab* 1972a;R-21:2–11.
13. Nelson W. Theory and application of hazard plotting for censored failure data. *Technometrics* 1972b;14:945–966.
14. Nelson W. Hazard plotting of left truncated life data. *J Qual Technol* 1990;22:230–238.
15. Duane JT. Learning curve approach to reliability monitoring. *IEEE Trans Aerosp* 1964;2:563–566.
16. Nelson W. Graphical analysis of system repair data. *J Qual Technol* 1988;20:2–35.
17. Nelson W. Confidence limits for recurrence data - applied to cost or number of product repairs. *Technometrics* 1995;37:147–157.
18. Proctor CL, Singh B. The graphical reliability evaluation of three-state device networks. *Microelectron Reliab* 1975;14:203–214.
19. Lawless JF. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons; 1982.
20. Nelson W. *Applied Life Data Analysis*. New York: John Wiley and Sons; 1982.
21. Epstein L, Sobel M. Life testing. *J Am Stat Assoc* 1953;48:486–502.
22. Barlow RE, Campo R. Total time on test processes and applications to failure data analysis. In: Barlow RE, Fussell J, Singpurwalla ND, editors. *Reliability and Fault Tree Analysis*. Philadelphia (PA): SIAM; 1975. p 451–481.
23. Bergman B, Klefsjo B. The total time on test concept and its use in reliability theory. *Oper Res* 1984;32:596–606.
24. Xie M, Zhao M. On some reliability growth models with graphical interpretations. *Microelectron Reliab* 1993;33(2): 149–167.
25. Xie M, Zhao M. Reliability growth plot - an underutilized tool in reliability analysis. *Microelectron Reliab* 1996;36:797–805.
26. Xie M, Lai CD. Reliability analysis using an additive Weibull model with bathtub shaped failure rate function. *Reliab Eng Syst Saf* 1995;52:87–93.
27. King JR. *Probability Charts for Decision Making*. New York: Industrial Press; 1971.
28. Abernethy RB. *The New Weibull Handbook*. 2nd ed. North Palm Beach (FL): R.B. Abernethy; 1996.
29. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. *Biometrika* 1968;49:525–545.
30. Gnanadesikan R. *Statistical Data Analysis of Multivariate Observations*. New York: John Wiley and Sons; 1997.
31. Chang DS, Tang LC. Graphical analysis for Birnbaum-Saunders distribution. *Microelectron Reliab* 1994;34(1): 17–22.
32. Wang Y, Young TM, Guess FM, León RV. Exploring reliability of oriented strand board's

- tensile and stiffness strengths. *Int J Reliab Appl* 2007;8(1): 111–124.
33. Hartler G. Graphical Weibull analysis of repairable systems. *Qual Reliab Eng* 1985;1:23–26.
 34. Jiang S, Kececioglu D. Graphical representation of two mixed-Weibull distributions. *IEEE Trans Reliab* 1992;41(2): 241–247.
 35. Zhao M, Xie M, Zhang Y. A study of a storage reliability estimation problems. *Qual Reliab Int Eng* 1995;11:123–127.
 36. Zhang T, Xie M. On the upper truncated Weibull distribution and its reliability implications. *Reliab Eng Syst Saf* 2011;96:194–200.
 37. Persson I, Astra ZR, Södertälje D, Khamis H. A comparison of graphical methods for assessing the proportional hazards assumption in the Cox model. *Festschrift in Honor of Distinguished Professor Mir Masoom Ali, On the Occasion of His Retirement, 2007 May 18–19.* p 23–43.
 38. Kulasekera KB, Williams CL, Coffin M, Manatunga A. Smooth estimation of the reliability function. *Lifetime Data Anal* 2001;7:415–433.
 39. Martínez LB, Márquez AC, Gunckel PV, Andreani AA. The Graphical analysis for maintenance management method: a quantitative graphical analysis to support maintenance management decision making. *Qual Reliab Eng Int* 2012. DOI: 10.1002/qre.1296.
 40. Currit PA, Dyer M, Mills HD. Certifying the reliability of software. *IEEE Trans Softw Eng* 1986;12(1): 3–11.
 41. Wang, CJ. Graphical analysis of ill-collected interval data for a repairable system in vehicles. *Proceedings: Annual Reliability and Maintainability Symposium (RAMS)*. New York: IEEE; 1991. p 93–97.
 42. Donovan J, Murphy E. Reliability growth-a new graphical model. *Qual Reliab Eng Int* 1999;15:167–174.
 43. Donovan J, Murphy E. Simulation and comparison of reliability growth models. *Int J Qual Reliab Manage* 2002;19(3): 259–271.
 44. Codier EO. Reliability growth in real life. *Proceedings Annual Symposium on Reliability*; 1968. p 458–469.
 45. Nelson W. An application of graphical analysis of repair data. *Qual Reliab Eng Int* 1998;14:49–52.
 46. Nelson W. Graphical comparison of sets of repair data. *Qual Reliab Eng Int* 2000;16:235–241.
 47. Zhang CW, Xie M, Goh TN. Design of exponential control charts using a sequential sampling scheme. *IIE Trans* 2006;38(12): 1105–1116.
 48. Batson RG, Jeong Y, Fonseca DJ, Ray PS. Control charts for monitoring field failure data. *Qual Reliab Eng Int* 2006;22(7): 733–755.
 49. Xie M, Goh TN, Ranjan P. Some effective control chart procedures for reliability monitoring. *Reliab Eng Syst Saf* 2002;77:143–150.
 50. Sürücü B, Sazak HS. Monitoring reliability for a three-parameter Weibull distribution. *Reliab Eng Syst Saf* 2009;94:503–508.
 51. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. *Statistical Inference Under Order Restrictions*. New York: John Wiley and Sons; 1972.
 52. Barlow RE, Davis B. Analysis of time between failures for repairable components. In: Fussell JB, Burdick GR, editors. *Nuclear Systems Reliability Engineering and Risk Assessment*. Philadelphia (PA): SIAM; 1977. p 543–561.
 53. Barlow RE. Geometry of the total time on test transform. *Nav Res Logist Q* 1979;26:393–402.
 54. Høyland A, Rausand M. *System Reliability Theory: Models and Statistics Methods*. New York: Wiley; 1994.
 55. Kumar U, Klefsjo B, Granholm S. Reliability investigation for a fleet of load haul dump machines in a Swedish mine. *Reliab Eng Syst Saf* 1989;26:341–361.
 56. Rao KRM, Prasad PVN. Graphical methods for reliability of repairable equipment and maintenance planning. *Proceedings of the Annual Reliability and Maintainability Symposium*. Philadelphia, USA: IEEE; 2001. p 123–128.
 57. Klefsjo B, Westberg U. TTT plotting and maintenance policies. *Qual Eng* 1996;9(2): 229–235.
 58. Kumar D, Westberg U. Maintenance scheduling under age replacement policy using proportional hazards model and TTT plotting. *Euro J Oper Res* 1997;99(3): 507–515.
 59. Kvaloy JT, Lindqvist BH. TTT-based tests for trend in repairable systems data. *Reliab Eng Syst Saf* 1998;60:13–28.
 60. Sun FB, Kececioglu DB. A new method for obtaining TTT plot for a censored sample. *Proceedings of the Annual Reliability and Maintainability Symposium*. Washington DC, USA: IEEE; 1999.
 61. Lai CD, Xie M. *Stochastic Ageing and Dependence for Reliability*. New York: Springer; 2006.

62. Kunitz H, Pamme H. Graphical tools for life time data analysis. *Stat Pap* 1991;32:85–113.
63. Kunitz H, Pamme H. The mixed gamma ageing model in life data analysis. *Stat Pap* 1993;34(1): 303–318.
64. Balagurusamy E, Misra KB. A graphical method for computing reliability parameters of m-order systems. *Int J Math Edu Sci Technol* 1976;7(2): 193–199.
65. Newby M. Failure models and data analysis based on ageing properties. *Int J Qual Reliab Manage* 1987;4(3): 36–46.
66. Launer RL. Graphical techniques for analyzing failure data with the percentile residual-life function. *IEEE Trans Reliab* 1993;42:71–80.
67. Haines AL, Singpurwalla ND. In: Proschan F, Serfling RJ, editors. Some contributions to the stochastic characterization of wear. *Reliability and Biometry, Statistical Analysis of Life Length*. Philadelphia (PA): SIAM; 1974. p 47–80.
68. Barlow E, Proschan F. Importance of system components and fault tree events. *Stoch Process Appl* 1975;3:153–173.
69. Nicol DM, Palumbo DL, Ulrey ML. A graphical model-based reliability estimation tool and failure mode and effects simulator. *Proceedings of the Annual Reliability and Maintainability Symposium*. Washington DC, USA: IEEE; 1995. p 74–81.
70. Walls L. A graphical approach to identification of dependent failures. *J Royal Stat Soc* 1996;45(2): 185–196.
71. Liu TS, Chiou SB. The application of Petri nets to failure analysis. *Reliab Eng Syst Saf* 1997;57:129–142.
72. Burrell QL. Modelling citation age data: simple graphical methods from reliability theory. *Scientometrics* 2002;55(2): 273–285.
73. Quigley JL, Walls LA. Conditional lifetime data analysis using the limited expected value function. *Qual Reliab Eng Int* 2004;20(3): 185–192.
74. Trindade D, Nathan S. Simple plots for monitoring the field reliability of repairable systems. *Proceedings of the Annual Reliability and Maintainability Symposium*; 2005. p 539–544.
75. Halim T, Tang L-C. An age-adjusted comparison of field failure data for repairable systems. *Proceedings of the 2008 Annual Reliability and Maintainability Symposium*. Las Vegas, USA: IEEE; 2008. p 54–58.
76. Donat R, Bouillaut L, Aknin P, Leray P, Levy D. A generic approach to model complex system reliability using graphical duration models. *Proceedings of the 5th International Mathematical Methods in Reliability Conference*; 2007 July 1–4.
77. Donat R, Bouillaut L, Aknin P, Leray P. Reliability analysis using graphical duration models. *Proceedings of the 3rd International Conference on Availability, Reliability and Security*. Washington DC, USA: IEEE; 2008a. p 795–800.
78. Donat R, Bouillaut L, Aknin P, Leray P, Bondeux S. Specific graphical models for analyzing the reliability. *Proceedings of 16th IEEE Mediterranean Conference on Control and Automation*; 2008b. p 621–626.
79. Halim T, Tang L-C. A graphical approach for confidence limits of optimal preventive maintenance cycles. *Qual Reliab Eng Int* 2009;25:199–213.
80. Al-Garni AZ, Tuzan M, Abdelrahman WG. Graphical techniques for managing field failures of aircraft systems and components. *J Aircraft* 2009;46(2): 608–616.
81. Saleh JH, Castet J-F. *Spacecraft reliability and multi-state failures: A Statistical Approach*. Chichester: John Wiley and Sons, Ltd; 2011.
82. Hovis JB. Effectiveness of reliability system testing on quality and reliability. *Proceedings of the Annual Reliability and Maintainability Symposium*; 1977.
83. Bergman B, Klefsjo B. A graphical method applicable to age replacement problems. *IEEE Trans Reliab* 1982;R-31:478–481.
84. Martin N. Failure models and data analysis based on ageing properties. *Int J Qual Reliab Manage* 1987;4(3): 36–46.
85. Meeker WQ, Escobar LA. *Statistical Methods for Reliability Data*. New York: Wiley Interscience; 1998.
86. Reineke DM., Paul EA, Murdock WP. Survival analysis and maintenance policies for a series system with highly censored data. *Proceedings of the Annual Reliability and Maintainability Symposium*. Anaheim, USA: IEEE; 1998.
87. Nelson W. Recurrent events data analysis for product repairs, disease recurrences, and other applications. Philadelphia, USA: ASA-SIAM; 2003.
88. Nelson W. *Applied Life Data Analysis*. New York: John Wiley and Sons; 2004.

GRASP: GREEDY RANDOMIZED ADAPTIVE SEARCH PROCEDURES

MAURICIO G. C. RESENDE

Algorithms and Optimization Research
Department, AT&T Labs Research,
Florham Park, New Jersey

RICARDO M. A. SILVA

Computational Intelligence and
Optimization Group, Department of
Computer Science, Federal
University of Lavras, Lavras, Brazil

Combinatorial optimization can be defined by a finite ground set $E = \{1, \dots, n\}$, a set of feasible solutions $F \subseteq 2^E$, and an objective function $f: 2^E \rightarrow \mathbb{R}$, all three defined for each specific problem. In this article, we consider the minimization version of the problem, where we seek an optimal solution $S^* \in F$ such that $f(S^*) \leq f(S)$, $\forall S \in F$. Combinatorial optimization finds applications in many settings, including routing, scheduling, inventory and production planning, and facility location.

While much progress has been made in finding provably optimal solutions to combinatorial optimization problems employing techniques such as branch and bound, cutting planes, and dynamic programming, as well as provably near-optimal solutions using approximation algorithms, many combinatorial optimization problems arising in practice benefit from heuristic methods that quickly produce good-quality solutions. Many modern heuristics for combinatorial optimization are based on guidelines provided by metaheuristics. Among these, we find genetic algorithms, simulated annealing, tabu search, variable neighborhood search, scatter search, path-relinking, iterated local search, ant colony optimization, swarm optimization, and greedy randomized adaptive search procedures (GRASPs).

In this article, we review the basic building blocks of GRASP, including solution construction schemes, local search methods, and

use of path-relinking as a memory mechanism in GRASP. The article is organized as follows. In the first section, we introduce a basic local search scheme. In the second section, we examine the relationship between the metaheuristic GRASP and local search. The third section covers construction schemes while in the fourth section, we introduce a memory mechanism in GRASP through path-relinking. Concluding remarks are made in the fifth section.

LOCAL SEARCH

Local search is a fundamental operator in GRASP heuristics. It is fully specified by a feasible starting solution $s_0 \in F$, an objective function $f(\cdot)$, for which we seek a local optimum, and a local neighborhood structure $N(\cdot)$ that restricts feasible moves in the search space. Local search in GRASP is applied from many distinct feasible starting solutions. Each application of local search in a GRASP heuristic results in a locally optimal solution, the best of which is output as the GRASP solution.

Consider the *neighborhood solution space* graph $\mathcal{X} = (S, M)$, where the node set S represents all feasible solutions of a combinatorial optimization problem and the edge set M corresponds to *moves* connecting neighboring solutions. A solution $s \in S$ is in the neighborhood $N(t)$ of a solution $t \in S$ if s can be obtained from t by making a small predefined change in t . If a solution $s \in N(t)$, then $t \in N(s)$, $(s, t) \in M$, and $(t, s) \in M$. Different neighborhood structures can be defined for a given combinatorial optimization problem.

The basic local search schemes take an initial solution, make it the current solution, and search the current solution's neighborhood for an improving solution. If one is found, it is made the current solution and local search is recursively applied to the current solution. The procedures terminate when no better solution exists in the current solution's neighborhood. Such local

search schemes require that the size of the neighborhood be such that its exploration can be done efficiently.

Algorithm 1. Basic local search algorithm.

```

 $t \leftarrow s_0$ ;
while there exists  $s \in N(t)$  such that  $f(s) < f(t)$  do
  |  $t \leftarrow s$ ;
end
return  $t$ ;

```

Algorithm 1 shows pseudocode for a basic local search algorithm. Local search starts at some node $s_0 \in S$ and makes it the current solution t , that is, $t = s_0$. At each iteration, it seeks a neighboring solution having a better objective function value than the current solution, that is, a solution $s \in N(t)$ such that $f(s) < f(t)$. If such a solution is found, it becomes the current solution, that is, $t = s$. These iterations are repeated until there is no better solution in the neighborhood $N(t)$ of the current solution t . In this case, solution t is called a *local optimum*.

Local search can be thought of as starting at a node $s \in S$ and examining adjacent nodes in graph $\mathcal{X}$ for an improving solution. In the *first-improving* variant, local search moves to the first-improving solution that it finds, whereas in the *best-improving* all neighboring solutions are evaluated and the local search moves to one of the best found. A compromise solution is used in *sampled local search* [1] where the neighborhood is sampled and the local search moves to the best of the sampled neighbors. One can think of this part of the search as *intensification*, since we concentrate on a small portion of the solution space.

Several approaches have been proposed to extend the basic local search scheme described above. These include methods such as variable neighborhood descent [2–7], variable neighborhood search [3, 8–12], short-term memory tabu search [13–24], simulated annealing [25–27], iterated local search [28–34], and very large-scale neighborhood search [35–37], which explore beyond the current solution’s neighborhood by allowing cost-increasing moves, by exploring multiple neighborhoods, or by exploring very large neighborhoods.

GREEDY RANDOMIZED ADAPTIVE SEARCH PROCEDURES

An effective search method needs to enable *diversification*. In carrying out local search, a good strategy should not focus the search on one particular region of the solution space, for example, around a solution built with a greedy algorithm. One alternative is to start local search from many randomly generated solutions with the expectation that there is a cost-improving path in $\mathcal{X}$ from one of those solutions to an optimal or near-optimal solution.

A GRASP repeatedly applies local search, starting from solutions that are constructed using a randomized greedy algorithm. The best local optimum found over all local searches is returned as the solution of the heuristic. Algorithm 2 shows pseudocode for a generic GRASP heuristic. At the completion of the randomized greedy phase the solution on hand is, for many problem instances, feasible. However, for some problem instances, it may be infeasible and requires that a repair procedure is applied to restore feasibility. Examples of GRASP implementations where repair procedures are needed to restore feasibility of the constructed solution can be found in Duarte *et al.* [38, 39], Nascimento *et al.* [40], and Mateus *et al.* [1].

An especially appealing characteristic of GRASP is the ease with which it can be implemented. Few parameters need to be set and tuned, and therefore, development can focus on implementing algorithms and data structures to assure efficiency. As we will see later, basic implementations of GRASP rely exclusively on two parameters. The first controls the number of construction/local search iterations that will be applied and the second controls the blend of randomness and greediness in the solution construction procedure. In spite of its simplicity and ease of implementation, GRASP is a very effective metaheuristic and produces the best-known solutions for many problems.

GRASP was first introduced by Feo and Resende [41]. See Feo and Resende [42], Pitsoulis and Resende [43], Resende and Ribeiro [44], and Resende [45] for surveys of GRASP

and Festa and Resende [46–48] for annotated bibliographies.

Algorithm 2. A basic GRASP in pseudocode.

```

 $x^* \leftarrow \infty;$ 
while stopping criterion not satisfied do
   $x \leftarrow \text{RandomizedGreedy}(\cdot);$ 
  if  $x$  is not feasible then
     $x \leftarrow \text{repair}(x);$ 
  end
   $x \leftarrow \text{LocalSearch}(x);$ 
  if  $f(x) < f(x^*)$  then
     $x^* \leftarrow x;$ 
  end
end
return  $x^*;$ 

```

GREEDY RANDOMIZED CONSTRUCTION

The greedy randomized construction methods of GRASP seek to produce a diverse set of good-quality starting solutions from which to start local search. This is achieved by adding randomization to the greedy algorithm. We illustrate here eight ways to do this. Solutions are built by adding one ground set element at a time to a partially constructed solution.

Semigreedy Construction

In the first construction scheme, called *semigreedy algorithm* by Hart and Shogan [49], at each step let the candidate set C denote all of the remaining ground set elements that can be added to the partial solution, and let $RCL(C)$ be a *restricted candidate list* made up of high-quality candidate elements. The quality of a candidate element is determined by its contribution, at that point, to the cost of the solution being constructed. A *greedy function* $g(c)$ measures this contribution for each candidate $c \in C$. Membership can be determined by rank or by quality relative to other candidates. Membership by rank, also called *cardinality based*, is achieved if the candidate is one of the q candidates with smallest greedy function value, where q is an input parameter that determines how greedy or random the construction will be. Membership by *quality* relative to other candidates determines a greedy

function cutoff value and only considers candidates with a greedy value no greater than the cutoff. To implement this, one usually makes use of a real-valued RCL parameter $\alpha \in [0, 1]$. Let $\bar{g} = \max\{g(c) \mid c \in C\}$ and $\bar{g} = \min\{g(c) \mid c \in C\}$. A candidate $c \in C$ is placed in the RCL only if $\bar{g} \leq g(c) \leq \bar{g} + \alpha \cdot (\bar{g} - \bar{g})$. Input parameter α determines how greedy or random the construction will be. Of these restricted candidates, one is selected at random and added to the partial solution. The construction is repeated until there are no further candidates. Algorithm 3 shows pseudocode for this construction procedure.

Algorithm 3. Greedy randomized construction: randomizing the greedy algorithm—semigreedy algorithm.

```

 $S \leftarrow \emptyset; C \leftarrow E;$ 
while  $|C| > 0$  do
  For all  $c \in C$  compute greedy function value  $g(c)$ ;
  Define  $RCL(C) \leftarrow \{c \in C \mid g(c) \text{ has a low value}\};$ 
  Select at random  $c^* \in RCL(C)$ ;
  Add  $c^*$  to partial solution:  $S \leftarrow S \cup \{c^*\};$ 
  Let  $C$  be the set of ground set elements that can
    be added to  $S$ ;
end
return  $S$ ;

```

Sample Greedy Construction

In the second construction scheme, called *sample greedy* by Resende and Werneck [50], instead of randomizing the greedy algorithm, a greedy algorithm is applied to each solution in a random sample of candidates. At each step a fixed-size subset of the candidates in C is sampled and the incremental contribution to the cost of the partial solution is computed for each sampled element. An element with the best incremental contribution is selected and added to the partial solution. This process is repeated until, as before, the construction terminates when no further candidate exists. Algorithm 4 shows pseudocode for this construction procedure. Sample greedy uses a parameter p to control the balance between greediness and randomness in the construction. Small values of p lead to more random solutions, while large values lead to more greedy solutions.

Algorithm 4. GRASP greedy randomized construction: sample greedy.

```

 $S \leftarrow \emptyset; C \leftarrow E;$ 
while  $|C| > 0$  do
    Randomly sample  $\min\{p, |C|\}$  elements from  $C$ 
    and put them in  $RCL(C)$ ;
    Select  $c^* = \operatorname{argmin}\{g(c) \mid c \in RCL(C)\}$ ;
    Add  $c^*$  to partial solution:  $S \leftarrow S \cup \{c^*\}$ ;
    Let  $C$  be the set of ground set elements that can
    be added to  $S$ ;
end
return  $S$ ;

```

Random Plus Greedy Construction

A possible shortcoming of the greedy randomized construction based on restricted candidate list is its complexity. At each step of the construction, each yet unselected candidate element has to be evaluated by the greedy function. In cases where the difference between the number of elements in the ground set and the number of elements that appear in a solution is large, this may not be very efficient. The third construction scheme, *random plus greedy*, was introduced in Resende and Werneck [50]. As is the case with the *sample greedy* construction scheme, the random plus greedy scheme also addresses this shortcoming.

In random plus greedy, a portion of the solution is constructed by randomly choosing p candidate elements and the remaining solution is completed in a greedy fashion. The resulting solution is randomized greedy. The value of p determines how greedy or random the construction will be. Small values of p lead to more greedy solutions where large values lead to ones that are more random.

Proportional Greedy Construction

The fourth construction scheme is the *proportional greedy* construction scheme. It was also introduced in Resende and Werneck [50]. In each iteration of proportional greedy, we compute the greedy function $g(c)$ for every candidate element $c \in C$ and then pick a candidate at random, but in a biased way: the probability of a given candidate $c' \in C$ being selected is inversely proportional to $g(c') - \min\{g(c) \mid c \in C\}$.

Reactive GRASP Construction

The fifth construction scheme is *reactive GRASP*. In randomized greedy construction procedures, the algorithm designer must decide how to balance greediness and randomness. A simple approach is to balance at random. For example, in the semigreedy construction with membership in the RCL by quality relative to the other candidates, the parameter α can be selected uniformly at random from the interval $[0, 1]$, so that each GRASP iteration uses a different α value and therefore has a different balance between greediness and randomness.

Prais and Ribeiro [51] showed that using a single fixed value for the value of RCL parameter α in the membership by quality relative to other candidates of the semigreedy algorithm often hinders finding a high-quality solution, which eventually could be found if another value was used. They proposed an extension of the basic GRASP procedure, which they call *reactive GRASP*, in which the parameter α is not fixed, but instead is selected at each iteration from a discrete set of possible values. The solution values found along the previous iterations serve as a guide for the selection process.

Prais and Ribeiro [51] define $\Psi = \{\alpha_1, \dots, \alpha_m\}$ to be the set of possible values for α . The probabilities associated with the choice of each value are all initially made equal to $p_i = 1/m$, $i = 1, \dots, m$. Furthermore, let z^* be the incumbent solution and let A_i be the average value of all solutions found using $\alpha = \alpha_i$, $i = 1, \dots, m$. The selection probabilities are periodically reevaluated by taking $p_i = q_i / \sum_{j=1}^m q_j$, with $q_i = z^* / A_i$ for $i = 1, \dots, m$. The value of q_i will be larger for values of $\alpha = \alpha_i$ leading to the best solutions on average. Larger values of q_i correspond to more suitable values for the parameter α . The probabilities associated with these more appropriate values will then increase when they are reevaluated. This reactive strategy is not limited to semigreedy procedures where membership in the RCL depends on relative quality. It can be extended to the other greedy randomized construction schemes, all of which need to balance greediness with randomization.

Long-Term Memory in Construction

The sixth construction scheme introduces long-term memory structures. Though reactive GRASP uses long-term memory (information gathered in previous iterations) to adjust the balance of greediness and randomness, Fleurent and Glover [52] observed that the basic GRASP does not and proposed a long-term memory scheme to address this issue in multistart heuristics. Long-term memory is one of the fundamentals on which tabu search [16,17] relies. Their scheme maintains a pool of elite solutions to be used in the construction phase. Later in this article, we discuss pools of elite solutions in detail. For now, all we need to know is that to become an elite solution, a solution must be either better than the best member of the pool, or better than its worst member and sufficiently different from the other solutions in the pool.

Fleurent and Glover [52] define a *strongly determined variable* to be one that cannot be changed without eroding the objective or changing, significantly, other variables and a *consistent variable* to be one that receives a particular value in a large portion of the elite solution set. Let $I(e)$ be a measure of the strongly determined and consistent features of solution element e from the ground set E . $I(e)$ becomes larger as e appears more often in the pool of elite solutions. It is used in the construction phase as follows. Recall that $g(e)$ is the greedy function value for candidate $e \in C$, that is, the incremental cost associated with the incorporation of element $e \in C$ into the solution under construction. Let $K(e) = F(g(e), I(e))$ be a function of the greedy and the intensification functions. For example, $K(e) = \lambda g(e) + I(e)$. The intensification scheme biases selection from the set C of candidate solutions to those elements $e \in C$ with a high value of $K(e)$ by setting their selection probability to be $p(e) = K(e) / \sum_{s \in \text{RCL}} K(s)$. The function $K(e)$ can vary with time by changing the value of λ , for example, initially λ may be set to a large value that is decreased when diversification is called for.

Biased Sampling Construction

The seventh construction scheme was introduced by Bresina [53] as *heuristic-biased stochastic sampling*. It is another way to depart from the uniform selection of candidate elements in the construction of a greedy randomized solution. Instead of choosing the next candidate element to add to the partial solution uniformly at random, heuristic-biased stochastic sampling suggests that any probability distribution can be used to bias the selection toward some particular candidates.

In the construction mechanism proposed by Bresina [53], a family of such probability distributions is introduced. They are based on the rank $r[\sigma]$ assigned to each candidate element σ , according to its greedy function value. The element with the smallest greedy function value has rank 1, the second smallest has rank 2, and so on. Several bias functions $b(\cdot)$ are introduced by Bresina, such as random bias where $b(r) = 1$, linear bias where $b(r) = 1/r$, log bias where $b(r) = \log^{-1}(r + 1)$, exponential bias where $b(r) = e^{-r}$, and polynomial bias of order n where $b(r) = r^{-n}$. Once all elements of the RCL have been ranked, the probability $\pi(\sigma)$ of selecting element $\sigma \in \text{RCL}$ can be computed as $\pi(\sigma) = b(r[\sigma]) / (\sum_{\sigma' \in \text{RCL}} b(r[\sigma']))$. In this scheme one can restrict candidates to be selected from the RCL or from the entire set of candidates, that is, make $\text{RCL} = C$.

Construction with Cost Perturbation

In the eighth construction scheme, called *construction with cost perturbation*, random perturbations are introduced in the original problem as a way of achieving randomness. The idea of introducing noise into the original costs to randomize a solution construction procedure is similar to that in the so-called *noising method* of Charon and Hudry [54,55].

Construction with cost perturbation can be more effective than the greedy randomized construction of the basic GRASP procedure in circumstances where the construction algorithms are insensitive to standard randomization strategies, such as selecting an element at random from a restricted candidate list. Ribeiro *et al.* [6] showed that

this is the case for the shortest-path heuristic of Takahashi and Matsuyama [56], which is used as one of the main building blocks of the construction phase of a hybrid GRASP they proposed for the Steiner problem in graphs.

Another situation where cost perturbations can be effective arises when no greedy algorithm is available for straightforward randomization as was the case of the hybrid GRASP developed by Canuto *et al.* [9] for the prize-collecting Steiner tree problem, which makes use of the primal–dual approximation algorithm of Goemans and Williamson [57] to build initial solutions using perturbed costs.

ADDING MEMORY TO GRASP THROUGH PATH-RELINKING

Path-relinking was originally proposed by Glover [58] as an intensification strategy exploring trajectories connecting elite solutions obtained by tabu search or scatter search [59–61]. Starting from one or more elite solutions, paths in the solution space graph leading toward other elite solutions are generated and explored in the search for better solutions. To generate paths, moves are selected to introduce attributes in the current solution that are present in

Algorithm 6. A basic GRASP with path-relinking heuristic.

```

P ← ∅;
f* ← ∞;
for i = 1, ..., imax do
    x ← GreedyRandomizedConstruction();
    if x is not feasible then
        x ← repair(x);
    end
    x ← LocalSearch(x);
    if i ≥ 2 then
        Randomly choose pool solutions  $\mathcal{Y} \subseteq P$  to relink with x;
        for y ∈  $\mathcal{Y}$  do
            xp ← PathRelinking(x, y);
            Update the elite set P with xp;
            if f(xp) < f* then
                f* ← f(xp);
                x* ← xp;
            end
        end
    end
end
P ← PostOptimize(P);
x* ← argmin{f(x), x ∈ P};
return x*;

```

the elite guiding solution. Path-relinking may be viewed as a strategy that seeks to incorporate attributes of high-quality solutions, by favoring these attributes in the selected moves.

The pseudocode in Algorithm 5 illustrates the mixed path-relinking procedure applied to a pair of solutions x_s and x_t . Mixed path-relinking [62] interchanges the roles of starting and guiding solutions after each move.

Algorithm 5. Mixed path-relinking procedure between solutions x_s and x_t .

```

Compute symmetric differences  $\Delta(x_s, x_t)$  and  $\Delta(x_t, x_s)$ ;
f* ← min{f(xs), f(xt)};
x* ← argmin{f(xs), f(xt)};
x ← xs; y ← xt;
while |Δ(x, y)| > 1 do
    m* ← argmin{f(x ⊕ m) : m ∈ Δ(x, y)};
    x ← x ⊕ m*;
    Update Δ(x, y) and Δ(y, x);
    if f(x) < f* then
        f* ← f(x);
        x* ← x;
    end
    t ← y; y ← x; x ← t;
end
x* ← LocalSearch(x*);
return x*;

```

The procedure starts by computing the symmetric differences $\Delta(x_s, x_t)$ and $\Delta(x_t, x_s)$ between the two solutions, that is, the set of moves needed to reach x_t from x_s and vice versa. Two paths of solutions are generated, one starting at x_s and the other at x_t . These paths grow out and meet to form a single path between x_t from x_s . Local search is applied to the best solution x^* in this path and the local minimum is returned by the algorithm. Initially, x and y are set to x_s and x_t , respectively. At each step, the procedure examines all moves $m \in \Delta(x, y)$ from the current solution x to solutions that contain an additional attribute of y and selects the one which results in the least-cost solution, that is, the one which minimizes $f(x \oplus m)$, where $x \oplus m$ is the solution resulting from applying move m to solution x . A best move m^* is made, producing solution $x \oplus m^*$. If necessary, the best solution x^* is updated. The sets of available moves are updated and the roles of x and y are interchanged. The procedure terminates when x and y are in each other's local neighborhood, that is, when $|\Delta(x, y)| = 1$.

Path-relinking is a major enhancement to the basic GRASP procedure, leading to significant improvements in solution time and quality. The use of path-relinking in GRASP was first proposed by Laguna and Martí [63]. It was followed by several extensions, improvements, and successful applications [6,9,44,50,64–68]. A survey of GRASP with path-relinking is presented in Resende and Ribeiro [69]. Two basic strategies are used. In one, path-relinking is applied to all pairs of elite solutions, either periodically during the GRASP iterations or after all GRASP iterations have been performed as a postoptimization step. In the other, path-relinking is applied as an intensification strategy to each local optimum obtained after the local search phase.

Applying path-relinking as an intensification strategy to each local optimum seems to be more effective than simply using it only as a postoptimization step. In general, combining intensification with postoptimization results in the best strategy. In the context of intensification, path-relinking is applied to pairs (x, y) of solutions, where x is a locally optimal solution produced by each GRASP

iteration after local search and y is one of a few elite solutions randomly chosen from a pool with a limited number `Max Elite` of elite solutions found along the search. Uniform random selection is a simple strategy to implement. Since the symmetric difference is a measure of the length of the path explored during relinking, a strategy biased toward pool elements y with large symmetric differences with respect to x is usually better than one using uniform random selection [50].

The pool is originally empty. Since we wish to maintain a pool of good but diverse solutions, each locally optimal solution obtained by local search is considered as a candidate to be inserted into the pool if it is sufficiently different from every other solution currently in the pool. If the pool already has `Max Elite` solutions and the candidate is better than the worst of them, then a simple strategy is to have the former replaces the latter. Another strategy, which tends to increase the diversity of the pool, is to replace the pool element most similar to the candidate among all pool elements with cost worse than the candidate's. If the pool is not full, the candidate is simply inserted.

Postoptimization is done on a series of pools. The initial pool P_0 is the pool P obtained at the end of the GRASP iterations. The value of the best solution of P_0 is assigned to f_0^* and the pool counter is initialized $k = 0$. At the k th iteration, all pairs of elements in pool P_k are combined using path-relinking. Each result of path-relinking is tested for membership in pool P_{k+1} following the same criteria used during the GRASP iterations. If a new best solution is produced, that is, $f_{k+1}^* < f_k^*$, then $k \leftarrow k + 1$ and a new iteration of postoptimization is done. Otherwise, postoptimization halts with $x^* \in \operatorname{argmin}\{f(x) \mid x \in P_{k+1}\}$ as the result.

The pseudocode in Algorithm 6 illustrates such a procedure. Each GRASP iteration now has three main steps. In the *construction phase* a greedy randomized construction procedure is used to build a feasible solution. In the *local search phase* the solution built in the first phase is progressively improved by a neighborhood search strategy, until a local minimum is found. In the *path-relinking phase* the path-relinking algorithm is applied

to the solution obtained by local search and to a randomly selected solution from the pool. The best solution found along this trajectory is also considered as a candidate for insertion in the pool and the incumbent is updated. At the end of the GRASP iterations, a *postoptimization phase* combines the elite solutions in the pool in the search for better solutions.

For some combinatorial optimization problems, a move m^* guided by the target solution is not guaranteed to result in a feasible solution $x \oplus m^*$ and the above-described scheme fails. Mateus *et al.* [1] proposed a new variant of GRASP with path-relinking suitable for such problems.

CONCLUDING REMARKS

This article considered basic building blocks needed to engineer heuristics based on the guiding principles of GRASP. These include randomized solution construction schemes, local search procedures, and the introduction of memory structures by way of path-relinking. One important topic that was left out of this article is that of parallel GRASP. The interested reader is directed to the survey of Resende and Ribeiro [70].

Acknowledgments

R. M. A. Silva was partially funded by the Brazilian Council for the Development of Science and Technology, CNPq.

REFERENCES

1. Mateus GR, Resende MGC, Silva RMA. GRASP with path-relinking for the generalized quadratic assignment problem. Technical report. Florham Park (NJ): AT&T Labs Research, Shannon Laboratory; To appear in J. Heuristics. 2009.
2. Andrade DV, Resende MGC. A GRASP for PBX telephone migration scheduling. Dallas, Texas: Proceedings of the 8th INFORMS Telecommunications Conference. 2006.
3. Hansen P, Mladenović N. An introduction to variable neighbourhood search. In: Voss S, Martello S, Osman IH, *et al.* editors. Metaheuristics: advances and trends in local search procedures for optimization. Norwell (MA): Kluwer; 1999. pp. 433–458.
4. Martins SL, Pardalos PM, Resende MGC, *et al.* Greedy randomized adaptive search procedures for the Steiner problem in graphs. In: Pardalos PM, Rajasegaran S, Rolim J, editors. Volume 43, Randomization methods in algorithmic design, Providence (R): DIMACS Series on Discrete Mathematics and Theoretical Computer Science. American Mathematical Society; 1999. pp. 133–145.
5. Ribeiro CC, Souza MC. Variable neighborhood search for the degree constrained minimum spanning tree problem. Discrete Appl Math 2002;118:43–54.
6. Ribeiro CC, Uchoa E, Werneck RF. A hybrid GRASP with perturbations for the Steiner problem in graphs. INFORMS Journal on Computing 2002;14:228–246.
7. Ribeiro CC, Vianna DS. A GRASP/VND heuristic for the phylogeny problem using a new neighborhood structure. International Transactions in Operational Research 2005; 12:325–338.
8. Beltrán JD, Calderón JE, Cabrera RJ, *et al.* GRASP/VNS hybrid for the strip packing problem. Valencia, Spain: Proceedings of Hybrid Metaheuristics (HM2004). 2004. pp. 79–90.
9. Canuto SA, Resende MGC, Ribeiro CC. Local search with perturbations for the prize-collecting Steiner tree problem in graphs. Networks 2001;38:50–58.
10. Drummond L, Vianna LS, Silva MB, *et al.* Distributed parallel metaheuristics based on GRASP and VNS for Solving the traveling purchaser problem. Taiwan, ROC: Proceedings of the 9th International Conference on Parallel and Distributed Systems (ICPADS'02). 2002. pp. 257–266.
11. Festa P, Pardalos PM, Resende MGC, *et al.* Randomized heuristics for the MAX-CUT problem. Optim Methods Softw 2002;7: 1033–1058.
12. Ochi LS, Silva MB, Drummond L. GRASP and VNS for solving traveling purchaser problem. Porto, Portugal: Proceedings of the 4th Metaheuristics International Conference (Mic2001). 2001. pp. 489–494.
13. Abdinnour-Helm S, Hadley SW. Tabu search based heuristics for multi-floor facility layout. Int J Prod Res 2000;38:365–383.
14. de Souza MC, Duhamel C, Ribeiro CC. A GRASP heuristic for the capacitated minimum spanning tree problem using a memory-based local search strategy. In Resende MGC, de Sousa JP, editors. Metaheuristics: computer decision-making. Norwell (MA): Kluwer

- Academic Publishers; 2003. pp. 627–658.
15. Delmaire H, Díaz JA, Fernández E, *et al.* Reactive GRASP and Tabu Search based heuristics for the single source capacitated plant location problem. *INFOR* 1999;37:194–225.
 16. Glover F. Tabu search –Part I. *ORSA J Comput* 1989;1:190–206.
 17. Glover F. Tabu search –Part II. *ORSA J Comput* 1990;2:4–32.
 18. Laguna M, González-Velarde JL. A search heuristic for just-in-time scheduling in parallel machines. *J Intell Manuf* 1991;2:253–260.
 19. Li Z, Guo S, Wang F, *et al.* Improved GRASP with tabu search for vehicle routing with both time window and limited number of vehicles. In: Orchard B, Yang C, Ali M, editors. Volume 3029, *Innovations in Applied Artificial Intelligence – Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2004)*, Berlin, Germany: Lecture Notes in Computer Science. Springer; 2004. pp. 552–561.
 20. Lim A, Wang F. A smoothed dynamic tabu search embedded GRASP for m-VRPTW. Boca Raton, Florida: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004). 2004. pp. 704–708.
 21. Moura A, Oliveira JF. A GRASP approach to the container-loading problem. *IEEE Intell Syst* 2005;20:50–57.
 22. Pu GG, Chong Z, Qiu ZY, *et al.* A hybrid heuristic algorithm for HW-SW partitioning within timed automata. Volume 4251, *Proceedings of Knowledge-based Intelligent Information and Engineering Systems*, Berlin, Germany: Lecture Notes in Artificial Intelligence. Springer; 2006. pp. 459–466.
 23. Serra D, Colomé R. Consumer choice and optimal location models: Formulations and heuristics. *Pap Reg Sci* 2001;80:439–464.
 24. Souza MJF, Maculan N, Ochi LS. A GRASP-tabu search algorithm to solve a school timetabling problem. Porto, Portugal: Proceedings of the 4th Metaheuristics International Conference (Mic2001). 2001. pp. 53–58.
 25. de la Peña MGB. Heuristics and metaheuristics approaches used to solve the rural postman problem: a comparative case study. Island of Madeira, Portugal: Proceedings of the 4th International ICSC Symposium on ENGINEERING OF INTELLIGENT SYSTEMS (EIS 2004). 2004. Available at <http://www.x-cd.com/eis04/22.pdf>.
 26. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. *Science* 1983;220(4598):671–680.
 27. Liu X, Pardalos PM, Rajasekaran S, *et al.* A GRASP for frequency assignment in mobile radio networks. In: Badrinath BR, Hsu F, Pardalos PM, *et al.*, editors. Volume 52, *Mobile networks and computing*, Providence (R): DIMACS Series on Discrete Mathematics and Theoretical Computer Science. American Mathematical Society; 2000. pp. 195–201.
 28. Baum EB. Iterated descent: a better algorithm for local search in combinatorial optimization problems. Technical report. California Institute of Technology; 1986.
 29. Baum EB. Towards practical ‘neural’ computation for combinatorial optimization problems. *AIP Conference Proceedings* 151 on Neural Networks for Computing. Woodbury (NY): American Institute of Physics Inc.; 1987. pp. 53–58.
 30. Baxter J. Local optima avoidance in depot location. *J Oper Res Soc* 1981;32:815–819.
 31. Johnson DS. Local optimization and the traveling salesman problem. Volume 443, *Proceedings of the 17th Colloquium on Automata*, LNCS. Berlin, Germany: Springer; 1990. pp. 446–461.
 32. Martin O, Otto SW. Combining simulated annealing with local search heuristics. *Ann Oper Res* 1996;63:57–75.
 33. Martin O, Otto SW, Felten EW. Large-step Markov chains for the traveling salesman problem. *Complex Syst* 1991;5:299–326.
 34. Ribeiro CC, Urrutia S. Heuristics for the mirrored traveling tournament problem. *Eur J Oper Res* 2007;127:775–787.
 35. Ahuja RK, Orlin JB, Sharma D. Multi-exchange neighborhood structures for the capacitated minimum spanning tree problem. *Math Program Ser A* 2001;91:71–97.
 36. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123:75–102.
 37. Geng Y, Li Y, Lim A. A very large-scale neighborhood search approach to capacitated warehouse routing problem. Hong Kong: 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 05). 2005. p. 8.
 38. Duarte A, Ribeiro CC, Urrutia S. A hybrid ILS heuristic to the referee assignment problem with an embedded MIP strategy. *Lect Notes Comput Sci* 2007;4771:82–95.

39. Duarte AR, Ribeiro CC, Urrutia S, *et al.* Referee assignment in sports leagues. *Lect Notes Comput Sci* 2007;3867:158–173.
40. Nascimento MCV, Resende MGC, Toledo FMB. GRASP with path-relinking for the multi-plant capacitated plot sizing problem. *Eur J Oper Res* 2010; 200:747–754.
41. Feo TA, Resende MGC. A probabilistic heuristic for a computationally difficult set covering problem. *Oper Res Lett* 1989;8:67–71.
42. Feo TA, Resende MGC. Greedy randomized adaptive search procedures. *J Glob Optim* 1995;6:109–133.
43. Pitsoulis LS, Resende MGC. Greedy randomized adaptive search procedures. In: Pardalos PM, Resende MGC, editors. *Handbook of applied optimization*. New York (NY): Oxford University Press; 2002. pp. 168–183.
44. Resende MGC, Ribeiro CC. Greedy randomized adaptive search procedures. In: Glover F, Kochenberger G, editors. *Handbook of metaheuristics*. Norwell (MA): Kluwer Academic Publishers; 2003. pp. 219–249.
45. Resende MGC. Metaheuristic hybridization with greedy randomized adaptive search procedures. In: Chen Z-L, Raghavan S, editors. *Tutorials. Hanover (MD): Institute for Operational Research and the Management Sciences (INFORMS) 2008*. pp. 295–319.
46. Festa P, Resende MGC. GRASP: an annotated bibliography. In: Ribeiro CC, Hansen P, editors. *Essays and surveys in metaheuristics*. Norwell (MA): Kluwer Academic Publishers; 2002. pp. 325–367.
47. Festa P, Resende MGC. An annotated bibliography of GRASP, Part I: algorithms. *Int Trans Oper Res* 2009;16:1–24.
48. Festa P, Resende MGC. An annotated bibliography of GRASP, Part II: applications. *Int Trans Oper Res* 2009;16:131–172.
49. Hart JP, Shogan AW. Semi-greedy heuristics: an empirical study. *Oper Res Lett* 1987;6: 107–114.
50. Resende MGC, Werneck RF. A hybrid heuristic for the p -median problem. *J Heuristics* 2004;10:59–88.
51. Prais M, Ribeiro CC. Reactive GRASP: an application to a matrix decomposition problem in TDMA traffic assignment. *INFORMS J Comput* 2000;12:164–176.
52. Fleurent C, Glover F. Improved constructive multistart strategies for the quadratic assignment problem using adaptive memory. *INFORMS J Comput* 1999;11:198–204.
53. Bresina JL. Heuristic-biased stochastic sampling. *Proceedings of the 13th National Conference on Artificial Intelligence*. Portland; 1996. pp. 271–278.
54. Charon I, Hudry O. The noising method: a new method for combinatorial optimization. *Oper Res Lett* 1993;14:133–137.
55. Charon I, Hudry O. The noising methods: a survey. In: Ribeiro CC, Ribeiro CC, editors. *Essays and surveys in metaheuristics*. Kluwer Academic Publishers; Norwell (MA): 2002. pp. 245–261.
56. Takahashi H, Matsuyama A. An approximate solution for the Steiner problem in graphs. *Math Jap* 1980;24:573–577.
57. Goemans MX, Williamson DP. The primal dual method for approximation algorithms and its application to network design problems. In: Hochbaum D, editor. *Approximation algorithms for NP-hard problems*. Boston (MA): PWS Publishing Co.; 1996. pp. 144–191.
58. Glover F. Tabu search and adaptive memory programming –Advances, applications and challenges. In: Barr RS, Helgason RV, Kennington JL, editors. *Interfaces in computer science and operations research*. Norwell (MA): Kluwer; 1996. pp. 1–75.
59. Glover F. Multi-start and strategic oscillation methods –Principles to exploit adaptive memory. In: Laguna M, González-Velarde JL, editors. *Computing tools for modeling, optimization and simulation: interfaces in computer science and operations research*. Norwell (MA): Kluwer; 2000. pp. 1–24.
60. Glover F, Laguna M. *Tabu search*. Norwell (MA): Kluwer; 1997.
61. Glover F, Laguna M, Martí R. Fundamentals of scatter search and path relinking. *Control Cybern* 2000;39:653–684.
62. Ribeiro CC, Rosseti I. A parallel GRASP for the 2-path network design problem. *Lect Notes Comput Sci* 2002;2004:922–926.
63. Laguna M, Martí R. GRASP and path relinking for 2-layer straight line crossing minimization. *INFORMS J Comput* 1999;11:44–52.
64. Aiex RM, Pardalos PM, Resende MGC, *et al.* GRASP with path-relinking for three-index assignment. *INFORMS J Comput* 2005; 17:224–247.
65. Festa P, Pardalos PM, Pitsoulis LS, *et al.* GRASP with path-relinking for the weighted MAXSAT problem. *ACM J Exp Algorithmics* 2006;11:Article 2.4: 1-16.

66. Oliveira CA, Pardalos PM, Resende MGC. GRASP with path-relinking for the quadratic assignment problem. In: Ribeiro CC, Martins SL, editors. Volume 3059, Proceedings of III Workshop on Efficient and Experimental Algorithms (WEA2004). Berlin, Germany: Springer; 2004. pp. 356–368.
67. Resende MGC, Martí R, Gallego M, *et al.* GRASP and path relinking for the max-min diversity problem. *Comput Oper Res* 2010; 37:498–508.
68. Resende MGC, Werneck RF. A hybrid multistart heuristic for the uncapacitated facility location problem. *Eur J Oper Res* 2006; 174:54–68.
69. Resende MGC, Ribeiro CC. GRASP with path-relinking: recent advances and applications. In: Ibaraki T, Nonobe K, Yagiura M, editors. *Metaheuristics: progress as real problem solvers*. New York (NY): Springer; 2005. pp. 29–63.
70. Resende MGC, Ribeiro CC. Parallel greedy randomized adaptive search procedures. In: Alba E, editor. *Parallel metaheuristics: a new class of algorithms*. Hoboken (NJ): John Wiley & Sons, Inc.; 2005. pp. 315–346.

GROUP DYNAMICS PROCESSES FOR IMPROVED DECISION MAKING

LAWRENCE D. PHILLIPS
Department of Management,
London School of Economics
and Political Science,
London, UK

Many operations research (OR) practitioners work with groups. They do this for a variety of reasons: to ensure a full exchange of viewpoints on the issues being considered, to obtain data and information shared by several sources, to engage participants in constructive dialogue, to develop shared understanding of the issues, to create a sense of common purpose, and to gain commitment to the way forward. Models can help. They provide an output that integrates input data in a way that extends human capability.

But can they extend the capabilities of groups? Regan-Cirincione [1] showed that groups can outperform even their best member, given three conditions: (i) impartial facilitation of the group process, (ii) modeling of group judgments, and (iii) use of information technology to display the output of the computer model. These conditions enabled her groups to engage in discussions that challenged shared assumptions and developed new perspectives. This led to conclusions that lay outside the original range of individual participants' answers to the given task.

Encouraged by the improved communication and performance of facilitated work groups that use interactive modeling, facilitators of these groups recognized that some groups worked more effectively than others, and the differences did not seem to be attributable to the facilitation, the modeling, or the technology. It was apparent that group dynamics were at work, that covert processes were engaged beneath the surface, sometimes interfering with the task at hand. Knowledge of those processes could help the facilitator to help the group accomplish its task more effectively.

Because the literature on group dynamics is vast, it is necessary here to draw selectively from that literature, focusing only on findings that could make a difference to an OR practitioner engaged in problem solving with work groups. Four areas provide relevant findings: experiences with groups, experimental research with groups, studies of on-going groups, and systems for improved decision making.

EXPERIENCES WITH GROUPS

An experiential approach to the study of groups that began in the United Kingdom as group therapy for returning WWII military people showed that groups exhibited their own dynamic. This discovery initiated the study of group processes for ordinary working groups, that is, groups formed to achieve an objective. The studies, which have been ongoing since 1957 [2], are conducted in the "here-and-now," with participants taking part in observing and experiencing the work of the group, which is to learn about group dynamics. Participants are engaged in plenary sessions, large groups, small study groups, and intergroup exercises among the small groups, over a period of one to two weeks [3]. These temporary learning communities, first developed in the United Kingdom by the Tavistock Institute of Human Relations, and subsequently in the United States by the A. K. Rice Institute [4], as Working Conferences on Group Relations, are directed and staffed by qualified professionals drawn from the helping professions, at least some of whom are psychoanalysts or psychotherapists. They serve as consultants to the body of paying participants, who come on the conferences to learn about group dynamics and to explore how their learning might be utilized back in their home institutions.

The "Primary Task" and Group Competence

Achieving the group's objective is considered

to be the “primary task.” Without a clear understanding of the primary task, the group can become fragmented and unproductive. But even if the task is clear, the requisite competence must exist in the group to accomplish the task, and the group must be given adequate resources to carry out the activities needed to achieve their goal. For example, a large organization formed a new group whose lead manager asked his direct reports to devise a strategy for the group. The group’s discussion seemed to the facilitator to be unfocused, and a general air of dreariness pervaded the room, so the facilitator asked what was going on, and within a few minutes established that none of the participants knew what strategy was [5]. After a short tutorial on strategy by the facilitator, the group came alive and got on with the task.

The Group Life and Anxiety

Not surprisingly, the dynamics uncovered in these working conferences show that groups have many of the features of individuals. The group itself has a life, which is different from that of other groups, and often is quite distinctive. Just as individuals pursue goals, so the group pursues its goals. In doing so, the group life can shift and take on emotional tones that might facilitate or hinder achievement of the goals. A major source of anxiety experienced by participants comes into being when individual and group goals conflict, when what is best for the group is not experienced by one or more of its members as personally fulfilling. The resulting anxiety is dealt with in a mature group when individuals hold the anxiety and work it through, either by constructive debate within the group, by reconsidering their own goals, or just by waiting to see if a later resolution becomes possible.

“Basic Assumption” Behavior

Less mature groups might demonstrate one of five behavior patterns that can interfere with work on the primary task. Three of these were first identified by Wilfred Bion, who called them “basic assumption behavior” [2]: *fight/flight*, *dependency*, and *pairing*.

As anxiety builds, a flight from the primary task may be felt as the only way to preserve the security of the group, and one way to do this is to instigate a fight with the leadership or facilitator. A participant or group of participants might walk out or threaten to walk out if the present task continues unaltered, a *fight/flight* response to anxiety. Another way to respond to the anxiety is to depend on some real or imagined authority figure to act as the group’s savior. Often, it is the facilitator who is asked, “What do you think is best for us at this point?” Or, some outside authority is quoted, or a rule, or an institutional process. Whatever is chosen, this may be a sign that the group is abandoning responsibility for achieving the objective, thereby creating a *dependency* relationship. Group members act as if they have no skills, and depend on a leader or the facilitator to accomplish the primary task. The third pattern is *pairing*, an abandonment of the primary task in favor of developing the relationship between a pair of its members who are thought to be imbued with special capabilities to carry out the primary task. The group life is one of hope for the pair, who are expected to produce the goods—though to actually do so would destroy the hope, which is the objective the group wishes to maintain to deal with their anxiety.

Two more patterns identified since the publication of Bion’s book are *oneness* and *me-ness*. Pierre Turquet observed in some groups a culture of wanting to be at one with some higher power, to surrender themselves, and become one with a powerful force [6]. Achieving and maintaining this powerful sense of wholeness becomes the focus of group activity, replacing the feeling of anxiety in working on the primary task. This is, of course, seen in some religious groups, but it can also happen with no religious connotations in ordinary working groups when the good feelings of togetherness, or *oneness*, becomes the primary aim. The opposite of this behavior pattern, according to Gordon Lawrence [7], is a denial of the group’s social reality; the only reality is the core identity of each participant—*me-ness*. One or more participants resist becoming contributing members of the group. They maintain

an aloof stance, seldom contributing, never exposing their feelings. When many of the members take this stance, the group feels calm, unproductive, even dead. The facilitator feels increasingly that he or she must energize the group, but every attempt fails. This behavior pattern is more common in newly formed groups, particularly if members feel threatened by external social events, such as financial crises, threatened job losses, and volatility in the business or political environment.

The group relations conferences have built a substantial literature on group dynamics [8], which provides rich and in-depth perspectives on the functioning of groups, roles, authority, leadership, and organizations as systems. The experiential learning approach provides opportunities to develop skilled knowledge that cannot be obtained by direct teaching. The conferences broaden participants' repertoires of choices and enable them to be more effective as group members. Information about the UK's Tavistock/Leicester Conferences is available at <http://www.tavistock.org>, and about the US's Group Relations Conferences at <http://www.akriceinstitute.org>.

EXPERIMENTAL RESEARCH WITH GROUPS

Social psychology has a long tradition of research on group processes, extending back to the 1920s when Floyd Allport, a prominent American psychologist, argued that the proper task of sociology was to describe collective behavior, while the task of social psychology was to explain that behavior in terms of the individual. In other words, he believed that the psychology of groups was essentially the psychology of individuals. That pretty well killed experimental work on groups for a couple of decades, and it only revived during the 1950s, the subject becoming respectable once more as researchers became clearer about the individual/group distinction, and with the publication of proper experimental approaches that manipulated independent variables to see their effects on dependent variables [9,10].

It is now a vast literature. The fifth edition of Donelson Forsyth's comprehensive

textbook, *Group Dynamics* [11] includes about 3000 references. And it does not include any references to publications about group relations conferences, nor does "problem solving" appear in the index even though some of the research he reports is relevant to that topic. On the other hand, the whole of Maier's now out-of-print book [10] was devoted to applied research methods to determine how problem solving could be enhanced in groups. In between these extremes, whether or not indexed, the relevance of group dynamics to problem solving can be found in several recent books and journals [11–16]. Drawing from that literature, this section highlights seven of the many factors that influence group dynamics in problem-solving groups: group size, stage of development, norms, roles, status, cohesion, and leadership style.

Group Size

The size of the group greatly affects communication among participants. A very small group, of less than six participants, often works like a jazz group: frequent individual, brief solos, punctuated by the group working together. A small group, of 6–12 people, requires a degree of self-discipline on the part of each member in deciding whether or not to say something, which can lead to rising anxiety from unspoken disagreements. In mature groups, this can be a rich stimulus to problem solving, for the group is small enough for disagreements to be aired, but too large to allow unfettered displays of individual aggrandizement, thus creating the potential for constructive problem solving; the group works together like a chamber orchestra. Beyond about 12–15 people, the scope for individual contributions decreases, as in an orchestra, where a conductor is required to bring the orchestra to its full capability. For OR practitioners, groups composed of up to 12 people work best, though facilitators or leaders with strong facilitation skills [17,18] can effectively manage larger groups. Beyond about 25 people, problem solving becomes very difficult, though Schein claims that dialoging processes can be accomplished even in large groups [19].

Stage of Development

Newly formed groups often behave initially rather differently from established groups in which people are well acquainted. Facilitators of new groups often find that, after introductions and an initial polite exchange of views, the groups begin to become more aggressive. Sometimes, this arises as individuals jockey for leadership positions, like stags battling on the field for supremacy, and at other times to relieve pent-up frustrations about inadequate leadership, organizational dysfunction, or personal animosities. Eventually, members adapt to each other, develop agreed ways of behaving, and the group life becomes more stable; effective problem solving becomes possible. Described initially by Tuckman [20] and now described as “forming, storming, norming and performing” [21], research shows that this sequence can happen, but it by no means describes all newly formed groups.

Norms

Groups that continue to work together rather quickly develop ways of working, accepted practices, approved behaviors, and a culture that collectively define the “norms” against which all behavior is deemed either acceptable or unacceptable. In some cases, these norms are derived from individual and organizational core values, often embodying principles of fairness, equity, and respect. For example, the first Committee on Radioactive Waste Disposal, charged in 2004 with recommending a management strategy for the United Kingdom’s radioactive waste, began by developing a set of five key principles, the first of which was “To be open and transparent.” They then conducted themselves in a manner consistent with these principles, to the extent that a retrospective analysis of their successful work revealed an important implicit norm that had guided their work: “To respect alternative points of view” [22].

Norms usually contribute to accomplishing the group’s primary task, but norms can also have an inhibiting effect. As an example, a small group of three people believed they

had developed a new optical process for correcting dyslexia. As they discussed in a facilitated workshop how their system worked, it became less and less clear to the facilitator what the physiological basis for their system was. The group rejected all suggestions to validate their approach through plausible theory and proper scientific research—they constantly referred to successful cases, decrying the ability of research to “prove” their case. An anti-scientific norm had been established, and the facilitator was unable to help them.

It behooves outsiders brought in to help the group in its work to be sensitive to these norms, facilitating in concert with the norms where they support the group, but also challenging the norms when they seem to inhibit effective work.

Roles

At the beginning of a meeting with a group, a facilitator will be better able to understand the group dynamics in subsequent interchanges between members if he or she notes each person’s job role as they introduce themselves. Discussions of the issues relevant to the primary task are often role-related. In considering a new e-commerce enterprise, the managing director of a financial services company emphasized the opportunity presented by all three options being considered, the director of the organization’s bank favored the option that was most like a bank, the operations manager was most worried about the risk that all three of these untried ventures presented, and so forth. They spoke from their own experience, bringing to the discussion what they knew best. Some groups then engage in criticizing opposing points of view, hoping that will allow just one option to emerge as best. Instead, the facilitator encouraged everyone to speak freely, and used the diversity to populate a decision model that represented everyone’s concerns in the form of evaluation criteria. Role differences can be used to identify criteria against which the options are then appraised.

Roles can also be conferred on members by the group, usually explicitly as the work is partitioned and assigned to different participants, but sometimes unknowingly.

For example, a person who is an optimist and is known to have a good sense of humor can be cast in the role of the clown, expected to make light of difficult patches in the group's work. That defuses building frustration, and can help a group to deal with its mounting anxiety. A normally pessimistic individual may become the group's skeptic, always calling attention to the downside, alerting the group to the risks and uncertainties. Other roles, observed in group relations conferences, include that of savior, father, mother, victim, observer, spectator, commentator, expert, warrior, or leader. This person, implicitly "elected" by the group, becomes the spokesperson for feelings bubbling up in the group. The skeptic expresses everyone's skepticism, the savior provides a vision for the group, the expert provides the knowledge to accomplish the primary task, the observer maintains a quiet watchful eye on everything, and so forth. While each of those functions may help a group, an excessive implicit demand by the group may show that the group is unwilling to take on those responsibilities, and is relieved of anxiety as the nominated person takes over expressing what others are feeling. Usually, that person feels increasingly uncomfortable in the role thrust on him or her without quite knowing why. The facilitator can ask the group if that person is perhaps expressing feelings that the rest of the group are experiencing. This may defuse the growing burden and shift responsibilities back to the whole group.

Status

The status of individuals in the group can affect group productivity either favorably or unfavorably. Junior people are often reluctant to speak when high status individuals are included as members. That is particularly true in Eastern countries where saving face is important. It is necessary for the facilitator to address questions first to lower status participants because if a high status person speaks first, no lower status person will disagree. To a lesser degree, this also happens in the West.

However, if at the start of the group, the most senior person, usually the sponsor of the event, establishes the primary task and explains that anyone who holds information

that would be relevant to the task should feel free to speak, status differences will be less threatening. Information is established as a neutral commodity, to be shared by everyone. That enhances group productivity, unless people have been included in the group who do not have the capability to deal with the issues. This can happen if participants are drawn from more than two or three levels in the organization when strategy is being discussed, for their comments will often be seen as "down in the weeds" or paying too much attention to the trees with no view of the forest. The information they are sharing is at the wrong level of granularity.

Cohesion

The degree to which members of a group like each other can powerfully affect the group's performance; so common sense tells us. But research shows that groups can be effective even if people in it do not all like each other. Two forces are at work here. One is indeed interpersonal attraction, but the more important feature is commitment to the group's goals, norms and primary task [23]. The relationship between cohesiveness and group performance is reflexive: successful performance strengthens cohesiveness and cohesiveness enhances performance. Of course, if developing group cohesion replaces the primary task, even if everyone is committed to it, then performance toward achieving the goal will suffer. But if work on the primary task is maintained, and participants believe in the value of achieving their objectives, then positive interpersonal relationships help to strengthen the norms and the willingness of individuals to work hard in achieving the goals.

One threat to cohesiveness is worth noting: anonymity. Anonymous voting is sometimes cited as a good way to ensure that one person does not dominate a group, or to protect minority viewpoints. Most facilitators would agree that these two concerns can be handled with good facilitation skills, like asking everyone first to write down their view on something, then revealing them in turn, or simply asking participants who have not spoken for their views. One unidentified black ball in an anonymous vote can lower

trust among members as participants discuss who cast the dissenting vote. Research on this topic shows no consistent effect, so facilitators need to take care in using anonymous votes to ensure that the risk is justified by any possible benefits.

Leadership Style

Much of the research on leadership shows that different tasks require different leadership styles. What is appropriate in one context may not work in another. A group facing an immediate and serious crisis will benefit from a leader with a clear vision, the competence to take the group forward, and the energy to persist in spite of difficulties. Charisma is not necessary; it may help in some circumstances, but is not a necessary condition for group effectiveness [24].

In the setting of a work group in which the OR professional is acting as a facilitator, it is helpful to distinguish between process and content [25]. The processes adopted by a group in accomplishing its primary task can be guided by a facilitator, who then relieves the leader of attending to process. That leaves the leader of the group free to contribute to content. But here it is best for the leader first to listen to what others have to say and only later at an appropriate time to present his or her view, if that seems appropriate. A leader who is too directive can stifle the group. In general, group problem solving requires a consultative and participative style of leadership.

STUDIES OF ON-GOING GROUPS

Another approach to group dynamics is to study existing groups. Researchers in the three sets of studies reported here used various combinations of observation, interviewing, and examination of available records. All were concerned with decisions taken by individuals and teams in a variety of real-life settings. All used systematic methods to analyze their data and draw reasonable inferences and conclusions, but none engaged in laboratory-based research. For some, decision making was the focus, not group dynamics, but inevitably all commented on team functioning.

Small Work Groups

A team of researchers led by Richard Hackman [12] studied 27 teams at different levels in diverse organizations that covered top management groups, task forces, professional support teams, performing groups, human service teams, customer service teams, and production teams. Their observations of these teams at work showed the importance of structure: high performing teams are clear about their goals and objectives, are motivated to achieve their objectives and highly engaged in the primary task, are operating within well-defined roles, have been given the requisite authority and resources to fulfill the primary task, and are clear about the limits of their authority. All these findings are consistent with those of the group relations conferences.

Naturalistic Decision-Making Groups

Gary Klein [26] is a main contributor to the field of naturalistic decision making (see *The Naturalistic Decision Making Perspective*). He studied the decisions taken by highly proficient individuals in settings of “time pressure, high stakes, personal responsibility and shifting conditions.” The main contributions of his work to an understanding of group dynamics are twofold: (i) experience is crucial to accomplishing the primary task, and (ii) effective decision making under pressure is not analytical; there is always a time–accuracy trade-off (see *Decision Making Under Pressure and Constraints: Bounded Rationality*). He acknowledges that both non-analytical and analytical ways of problem solving are appropriate in many situations, which implies that effective teams are able to integrate intuition with thinking.

Political Decision Making

Janis and Mann [27,28] used the term *group-think* to describe a common feature of foreign policy decisions from 1940 to 1970 in the United States. They discovered “a deterioration of mental efficiency, reality testing, and moral judgment that results from in-group pressures.” They identified the conditions that stimulate group thinking: a cohesive

group which is under stress, dominated by a directive leader and insulated from information from outside the group, with participants focusing on one course of action rather than searching for alternatives whose merits can be appraised. They noted several features of the group dynamic:

- tightly knit group sharing an illusion of unanimity and correctness
- pressure exerted on dissenters to conform;
- unquestioned belief in the group's morality;
- scapegoating of "out-groups" as inferior;
- self-appointed "mindguards" shielding the group from contradictory information.

Leaders can protect against group-think by encouraging the airing of differences in viewpoints and listening to others before stating their own preferences. Also, inviting outside experts to join the group and establishing other groups to work on the same or different aspects of the problem legitimizes alternative viewpoints. For example, the European Medicines Agency (EMA) establishes two separate expert groups from different European Union countries to prepare reports on the benefits and risks of new medicinal products submitted to the EMA for approval. Each country mobilizes its own experts in preparing their reports, which are subsequently compared and discussed by the EMA's regulatory committee.

SYSTEMS FOR IMPROVING DECISION MAKING

Many of the above findings are integrated in group-based systems that can mobilize the capabilities of individuals to create high-performing teams. Two are mentioned briefly here: the Margerison and McCann Team Management System, and Decision Conferences. Each of these brings together many of the above research findings to create high-performing teams.

Team Management System

The research of Charles Margerison and Dick McCann identified nine types of work that characterize the activities carried out by effective teams. They found that people differ in their preferences for these types of work, so they developed a method for measuring people's work preferences [29] based on the psychoanalyst Carl Jung's theory of psychological types [30]. By focusing on work preferences rather than personality types, they developed a way of mapping people's work preferences onto the eight types of work. Their research shows that a balance of work preferences in a team helps to establish high levels of performance. They also developed an instrument for evaluating a person's skills in linking tasks and people. It is the combination of effective linking skills with group members contributing according to their work preferences that enables a group to function smoothly.

Decision Conferences

Since 1979 several thousand 2–3-day work groups, facilitated by impartial specialists in group processes and decision analysis, using on-the-spot modeling of information and participants' judgments, have enabled groups of key players to arrive quickly at an agreed way forward. Developments since 1982 at the London School of Economics brought together the technology of decision analysis (see *Foundations of Decision Theory*) with an understanding of group processes. The resulting intellectual technology, Decision Conferences [31], has demonstrated the advantages of this approach over ordinary working-group meetings [32]. *Decision Conferencing* uses a mixture of Decision Conferences, workshops, and other activities for sustained work with a client on large and difficult problems of strategy, evaluation, choice, prioritization, and resource allocation.

REFERENCES

1. Regan-Cirincione P. Improving the accuracy of group judgment: a process intervention combining group facilitation, social judgment analysis, and information technology.

- Organ Behav Hum Decis Processes 1994;58: 246–270.
2. Bion WR. Experiences in groups. London: Tavistock Publications; 1961.
 3. Higgin G, Bridger H. The psychodynamics of an inter-group experience. London: Tavistock Publications; 1965.
 4. Rice AK. Learning for leadership. London: Tavistock Publications; 1965.
 5. Phillips LD, Phillips MC. Facilitated work groups: theory and practice. *J Oper Res Soc* 1993;44(6):533–549.
 6. Turquet PM. Leadership: the individual and the group. In: Gibbard GS, Hartman JJ, Mann RD, editors. *Analysis of groups: contribution to theory, research, and practice*. San Francisco (CA): Jossey-Bass; 1974.
 7. Lawrence WG, Bain A, Gould L. The fifth basic assumption. *Free Assoc* 1966;6(37):28–56.
 8. Trist E, Murray H, editors. *The social engagement of social science: a Tavistock anthology*. Volume I, *The Socio-psychological perspective*. London: Free Association Books; 1990.
 9. Thibaut JW, Kelley HH. *The social psychology of groups*. New Brunswick (NJ): Transaction Publishers; 1986 (Original edition 1959, John Wiley & Sons).
 10. Maier NRF. *Problem-solving discussions and conferences: leadership methods and skills*. New York: McGraw-Hill Book Company; 1963.
 11. Forsyth DR. *Group dynamics*. 5th ed. Belmont (CA): Wadsworth, Cengage Learning; 2009.
 12. Hackman JR, editor. *Groups that work (and those that don't): creating conditions for effective teamwork*. San Francisco (CA): Jossey-Bass; 1990.
 13. Baron RS, Kerr NL. Group process, group decision, group action. In: Manstead T, editor. *Mapping social psychology*. 2nd ed. Buckingham: Open University Press; 2003.
 14. Levi D. *Group dynamics for teams*. 2nd ed. Thousand Oaks (CA): Sage Publications; 2007.
 15. *Group dynamics: theory, research and practice*. Quarterly, American Psychological Association, Washington, DC.
 16. Brown R. *Group processes: dynamics within and between groups*. 2nd ed. Blackwell: Oxford UK & Cambridge USA; 2000.
 17. Schein EH. *Process consultation revisited: building the helping relationship*. Reading (MA): Addison-Wesley; 1999.
 18. Schwarz R. *The skilled facilitator: a comprehensive resource for consultants, facilitators, managers, trainers and coaches*. San Francisco (CA): Jossey-Bass; 2002.
 19. Schein EH. On dialogue, culture, and organizational learning. *Organ Dyn* 1993;22(2): 40–51.
 20. Tuckman BW. Developmental sequence in small groups. *Psychol Bull* 1965;63(6): 384–399.
 21. Tuckman BW, Jensen MAC. Stages of small group development revisited. *Group Organ Stud* 1977;2:419–427.
 22. Morton A, Airolidi M, Phillips L. Nuclear risk management on stage: the UK's Committee on Radioactive Waste Management. *Risk Anal* 2009;29(5):764–779.
 23. Mullen B, Copper C. The relation between group cohesiveness and performance: an integration. *Psychol Bull* 1994;115(2):210–217.
 24. Jaques E, Clement SC. *Executive leadership: a practical guide to managing complexity*. Arlington (VA): Cason Hall & Co. Publishers Ltd.; 1991.
 25. Eden C. The unfolding nature of group decision support—two dimensions of skill. In: Eden C, Radford J, editors. *Tackling strategic problems: the role of group decision support*. London: Sage; 1990. pp. 48–52.
 26. Klein G. *Sources of power: how people make decisions*. Cambridge (MA): MIT Press; 1998.
 27. Janis IL, Mann L. *Decision making: a psychological analysis of conflict, choice, and commitment*. New York: Free Press; 1977.
 28. Janis IL. *Groupthink: psychological studies of policy decisions and fiascos*. 2nd ed. Boston (MA): Houghton Mifflin; 1982.
 29. Margerison CJ. *Team leadership: a guide to success with team management systems*. London: Thomson; 2002.
 30. Jung CJ. *Psychological types*. London: Keegan-Paul; 1923.
 31. Phillips LD. Decision Conferencing. In: Edwards W, Miles RF, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007.
 32. Schilling MS, Oeser N, Schaub C. How effective are decision analyses? Assessing decision process and group alignment effects. *Decis Anal* 2007;4(4):227–242.

GUIDED LOCAL SEARCH

CHRISTOS VOUDOURIS
EDWARD P. K. TSANG
ABDULLAH ALSHEDDY
School of Computer Science and
Electronic Engineering,
University of Essex,
Colchester, UK

Many optimization problems found in the real world are computationally intractable (NP-Hard) due to the exponential increase in the number of possible solutions when increasing the problem size. Classical methods such as complete search often encounter great difficulty in solving such problems within reasonable computational times. This motivates the development of heuristic methods that sacrifice completeness. Some of the best known heuristics are local search (LS) methods. LS works by starting with an initial solution (randomly or heuristically generated), and then iteratively moving to “neighbor” solutions that improve the objective function. A “neighbor” solution is usually obtained by making a local (i.e., small) change to the current solution. LS terminates when it reaches a local minimum¹ (i.e., a state where the current solution is superior to all its neighbors), or computational resources run out. LS can obtain solutions of good quality very quickly. However, a major issue that negatively affects its efficiency is that it can be trapped in local minima, which halt the search, preventing the algorithm from reaching the global minimum. To overcome this problem, a class of techniques, which is referred to as, metaheuristics, has been introduced and grown in popularity over the years. Simulated annealing (SA) and tabu search (TS) are well known and representative techniques in this class.

¹We assume a minimization problem here but all the findings equally apply to a maximization problem.

Guided local search (GLS) [1] is another and relatively newer general metaheuristic algorithm. It is a higher level procedure that can sit on top of other heuristic methods to guide them to escape locally optimal solutions. GLS can be seen as a generalization of techniques such as GENET [2–4] and the min-conflicts heuristic repair method by Minton *et al.* [5] developed for constraint satisfaction problems. GLS also relates to ideas from the area of search theory as applied to naval search and rescue operations or deep-space surveillance on how to distribute the searching effort [6,7].

One key concept in GLS is to replace the objective function used by LS with an augmented objective function, which includes penalties associated to features of the candidate solutions. When LS is trapped in a local minimum, GLS dynamically modifies the augmented objective function by selectively increasing penalties for features present in the local minimum. This causes LS to escape by forcing it to move toward neighbor solutions that exclude these features. For the scheme to work, GLS requires the definition of solution features (i.e., attributes that can be used to discriminate between different solutions). It also requires a mechanism to select which features are penalized in a local minimum. The selection is normally done using a utility function, which for each feature present in a local minimum defines the utility of penalizing it. This utility function should be defined so that the search focuses where promise is shown while not neglecting the less-promising areas. From another viewpoint, the GLS penalization scheme should balance search intensification versus diversification.

Different metaheuristics apply different strategies to improve search and overcome the problem of local minima as shown in Table 1. SA and genetic algorithm (GA) employ a randomness component in a probability-based and population-based schema respectively to modify themovement

strategy. TS [8] exploits historical information in a memory-based approach to modify the set of neighbor solutions (i.e., neighborhood). GLS employs information in a penalty-based approach to dynamically modify the objective function.

GUIDED LOCAL SEARCH

GLS applies a penalty-based approach that can be superimposed on an LS algorithm with the aim of guiding it to escape local minima. The objective function is augmented with penalties which are increased when the LS settles in local minimum. Penalties are associated to solution features. A key component in GLS is thus the definition of the solution features. Solution features are defined to distinguish between solutions with different characteristics. GLS aims to penalize poor and costly characteristics that hopefully will be removed by the LS algorithm. Features are problem dependant and can be derived directly from the objective function which usually comprises one or more features. Examples of features are *whether the candidate tour travels immediately from city X to city Y* in the traveling salesman problem (TSP), and *whether the item I is assigned to item J* in the assignment problem. A simple indicator function can be used to determine

whether a solution exhibits a specific feature and therefore formally defines it.

GLS associates a cost and a penalty to each feature. Feature costs can often be defined by looking at the terms of the objective function and their coefficients. For instance, the costs of the example features mentioned above could be that of the traveling distance from city X to city Y in the TSP, and the carried cost of assigning item I to item J in the assignment problem. Feature penalties are initialized to 0 and they are incremented every time the LS settles in a local optimum and the feature is selected to be penalized.

Once features and their costs have been determined, GLS defines an augmented function h that will be used by LS (replacing the problem's objective function g):

$$h(s) = g(s) + \lambda \times \sum (p_i \times I_i(s)), \quad (1)$$

where s is a candidate solution, g an objective function that maps every candidate solution s to a numerical value, λ is a parameter to the GLS algorithm, i ranges over the features, p_i is the penalty for feature i (all p_i 's are initialized to 0), and I_i is an indicator function of whether s exhibits feature i :

$$I_i(s) = 1 \text{ if } s \text{ exhibits feature } i; 0 \text{ otherwise.} \quad (2)$$

Table 1. Overview of Strategies and Components of GLS and Other Well-Known Metaheuristics

Metaheuristic	The High-Level Strategy	Basic Component	Tuning Parameters
GLS	Using a penalty-based approach to dynamically modify the cost function	Features cost	λ
TS	Using a memory-based approach to dynamically change the neighborhood method	Solution attributes Short-term memory (STM) Long-term memory (LTM)	Tabu activation rules Tabu tenure [Aspiration criterion] [LTM measures and parameters]
SA	Using a probabilistic-based approach to allow nonimproving moves	Annealing schedule	Initial temperature Cooling rate
GA	Using a population-based approach that includes a mutation operator to involve (partially) random solutions	Genetic operators: reproduction, crossover, and mutation, Replacement strategy	Population size Crossover rate Mutation rate

Starting from an initial solution and using the augmented objective function h , an LS algorithm is applied until it reaches a local minimum. Provided with the local minimum solution, GLS selects which features to penalize to increase the value of h and force LS to move to a neighboring solution. The process is repeated until a termination criterion is met. The basic procedure of GLS is shown in Fig. 1. The basic idea of using the augmented objective function to force search out of local minima is illustrated in Fig. 2.

With regards to the feature selection mechanism, the intention is to gradually penalize “unfavorable” features when LS settles in a local minimum. The feature that has high cost affects the overall cost more and guiding the search toward solutions that avoid it when escaping out of the local minimum provides a strong intensification

mechanism, which favors high-quality solutions.

However, this intensification of search needs to be balanced with a diversification effect, which allows penalizing features of low cost when the high-cost features have been penalized several times. The current penalty value of a feature readily provides this information (i.e., the frequency of penalizing a feature). Combining the two factors above, the utility of penalizing feature i , given by $util_i$, under a local optimum s_* , is defined as follows:

$$util_i(s_*) = I_i(s_*) \times \frac{c_i}{1 + p_i}, \quad (3)$$

where c_i is the cost and p_i is the current penalty value of feature i . Note here that if a feature is not exhibited in the local optimum (as indicated by I_i), then the utility of

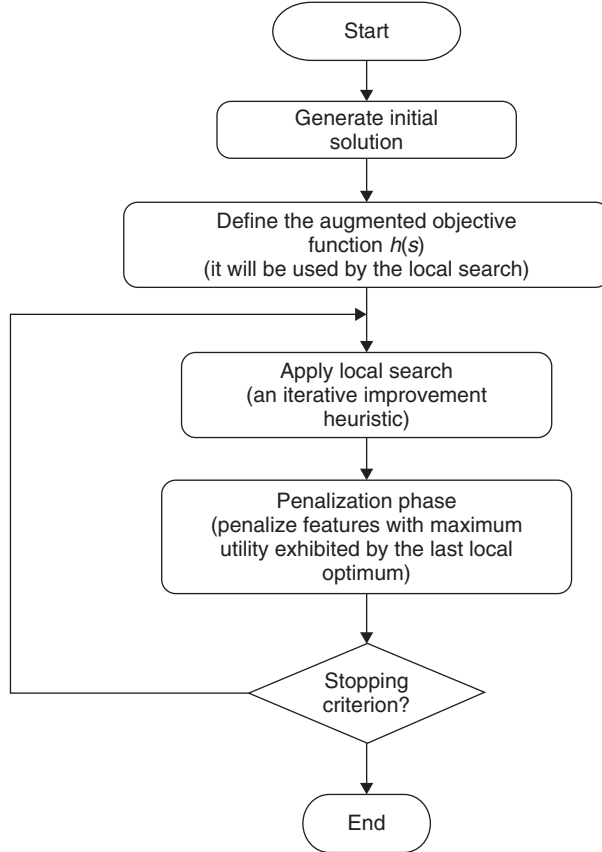


Figure 1. The basic procedure of guided local search.

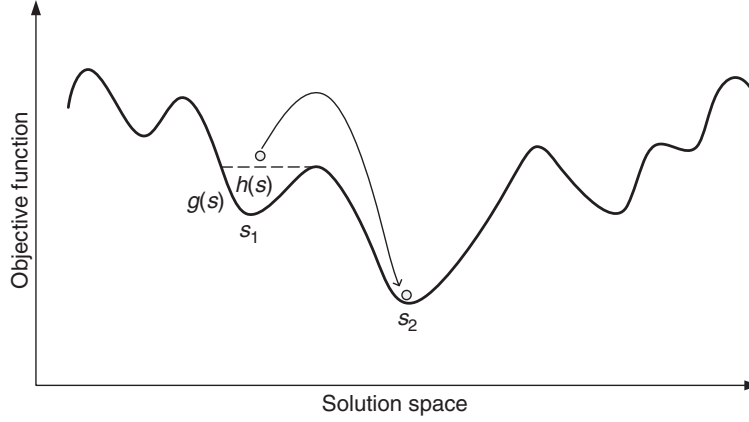


Figure 2. The basic idea of GLS. An augmented objective function h is created and dynamically changed to force local search out of local minima of the original objective function g .

penalizing it is 0. The higher the cost of a feature (the greater c_i), the greater the utility of penalizing it. Besides, the more times that it has been penalized (the greater p_i), the lower the utility of penalizing it again. In a local optimum, the feature(s) with the greatest utility value will be penalized. When a feature is penalized, its penalty value is always increased by 1. The scaling of the penalty is adjusted by λ .

Considering the feature's cost and penalty as factors in selecting the feature to penalize helps GLS to focus its search effort on more promising areas of the search space: areas that contain candidate solutions that exhibit "good features," that is, features involving lower cost. On the other hand, penalties help to prevent the search from directing all effort to any particular region of the search space. Naturally the choice of the features, their costs, and the setting of λ may affect the efficiency of search. Experience shows that the features and their costs if derived directly from the objective function tend to work well with providing guidance to the LS. In many problems, the performance of GLS is also not too sensitive to the value of λ . This means one does not require too much effort to apply GLS to a new problem.

Measures to reduce the sensitivity of the λ parameter can also be adopted and they have been proven to improve algorithm performance in certain cases [9].

FAST LOCAL SEARCH

Another factor that affects the efficiency of LS algorithms is the size of the neighborhood that is checked at each iteration. LS algorithms, given a move operator, which introduces small changes to the current solutions, consider some or all the neighbors that can be derived from a candidate solution using that move operator. This usually results in a very large number of neighbors and, in cases such as greedy LS, we need to evaluate all solutions in the neighborhood of the current solution to decide the best one to move to. This is obviously time-consuming, and also affects metaheuristics such as GLS, which utilize LS in an iterative manner to examine numerous local minima. New algorithms that focus on redefining the neighborhood function have been proposed such as approximate 2-opt [10] for TSP.

Fast local search (FLS) was developed to restrict the neighborhood size, which leads to an improvement in the efficiency of LS especially in the context of the GLS metaheuristic. The aim of FLS is to focus LS on the promising neighbor solutions. The basic idea of FLS is to partition the neighborhood into a number of subneighborhoods. Each neighbor solution belongs to at least one subneighborhood.

To illustrate this partition, consider the single swap move operator used to obtain

the neighbors of a solution represented by a permutation of entities, for example, cities or jobs. In this case, the neighborhood of a solution is the set of all solutions obtained by swapping a variable's value with the value of another variable. A subneighborhood SN_i can be defined as the set of neighbors obtained by swapping variable i with other variables.

In FLS, each subneighborhood is associated with an activation bit. There are two states of any subneighborhood: active or inactive (i.e., the activation bit is either 1 or 0, respectively). Only active subneighborhoods are examined during the search. At the beginning of the search, all subneighborhoods are active. During the search, a subneighborhood becomes inactive when all candidate solutions it includes have been evaluated, but none of which provide a better move. However, if an improvement is found in a subneighborhood, then it remains active and the move is executed. This movement may include the process of activating some inactive subneighborhoods if they are expected to involve promising solutions. FLS stops when all subneighborhoods become inactive, that is, all activation bits equal 0, and there are no subneighborhoods to be examined.

The overall procedure could be many times faster than conventional LS. The bit setting scheme encourages chains of moves that improve specific parts of the overall solution. As the solution becomes locally better, the process is settling down, examining fewer moves, and saving enormous amounts of time which would otherwise be spent on examining predominantly bad moves.

A COMBINATION OF GLS AND FLS

Combining GLS with FLS is straightforward. The idea is to associate features to subneighborhoods. The associations to be made are such that for each feature we know which subneighborhoods contain moves that have an immediate effect upon the state of the feature (i.e., moves that remove the feature from the solution). After the penalties of features are increased, only the subneighborhoods corresponding to these features are activated.

The combined technique is called *guided fast local search* (GFLS).

GFLS is more sophisticated than the basic GLS algorithm using a number of FLS subneighborhoods corresponding to solution features, which are enabled/disabled during the search process. By reducing the size of the neighborhood, the computation time of an LS iteration is significantly reduced. Moreover, GFLS can provide better search focus after penalty increase by GLS by selectively activating only related subneighborhoods. This can dramatically shorten the time required by an improvement method to reoptimize the solution each time the augmented objective function is modified.

Overall, GFLS enables more LS iterations in a fixed amount of time and therefore offers more opportunities to GLS to guide the underlying LS process. It is much faster than basic GLS especially in large problem instances where repeatedly and exhaustively searching the whole neighborhood is computationally expensive. There is a chance some improving moves may be missed. The hope and premise is that the gain in "search speed" outweighs the loss in "search quality."

EXTENSIONS AND HYBRIDS

GLS has proved itself an efficient approach to tackle hard optimization problems. This has resulted in further research on GLS extensions and hybrids with other heuristics and metaheuristics. The outputs from this research, as will be discussed in this section, demonstrate the flexibility and generality of GLS.

Extensions to Guided Local Search

Use of tabu lists in GLS was found promising in the radio link frequency assignment problem (RLFAP) [11]. The motivation there was to avoid the possibility of misguiding the LS if too many penalties were built up during the search. Resembling the tabu lists idea, a limited number of penalties are used. When the list is full, old penalties are overwritten.

Aspiration criterion is another idea that comes from TS. Mills *et al.* [9] have described an extended guided local search

(EGLS), which utilizes the *improved-best* aspiration criterion modified specifically for penalty-based schemes. The basic idea is to use aspiration to favor promising moves. This can be done by simply examining the original objective function to check whether a new, better than previously found solution exists in the current neighborhood, before following the standard GLS scheme. In addition, the idea of random moves has been incorporated into EGLS to enhance GLS. LS is allowed to make a move at random according to some probability. The resulting algorithms improved the robustness of GLS over a range of parameter settings in the case of the quadratic assignment problem (QAP).

A distributed version of GLS was introduced by Basharu *et al.* [12] for solving distributed constraint satisfaction problems. The distributed guided local search (Dist-GLS) is following an agent-based paradigm introducing a messaging protocol communicating penalty decisions between problem solving agents working on different parts of the problem. It also incorporates additional heuristics to enhance GLS's efficiency in distributed scenarios.

GLS with a multiphase penalization procedure was proposed by Hifi *et al.* [13]. The modified procedure works either as a guidance for LS to escape local optima or as a restarting strategy. Tamura *et al.* [14] suggested a GLS procedure with a modified augmented objective function.

Guiding other Heuristics

Instead of simple LS, GLS can be generalized to guide other sophisticated heuristics in order to improve their efficiency. GLS has been hybridized with more complex LS methods such as variable neighborhood search (VNS) and large neighborhood search (LNS). The addition to VNS and LNS is the use of GLS to escape local minima encountered. Guided variable neighborhood search (GVNS), which is proposed by Kytöjoki *et al.* [15], combines GLS with VNS, which was successfully applied to the vehicle routing problem (VRP). The hybrid of GLS and LNS was proposed by Xiaohu and Haubrich [16]

in solving a network planning optimization problem.

GLS can sit on top of other metaheuristics such as GAs [17,18]. This has been demonstrated in guided genetic algorithm (GGA) [19–22]. GGA is a hybrid of GA and GLS. Besides the objective of extending the domain of both GA and GLS, a major objective of GGA is to further improve GLS by the means of population-based search. The basic idea is to employ GLS to bring GA out of plateau, a state when no progress has been made after n iterations (n is a parameter to GGA). GLS modifies the fitness function (which is the objective function) by means of penalties, using the criteria defined in Equation (3). GA will then use the modified fitness function in future generations. The penalties are also used to bias crossover and mutation in GA—genes that contribute more to the penalties are made more susceptible to changes by these two GA operators. This allows GGA to be more focused in its search. On the other hand, GGA can roughly be seen as a number of GLSs from different starting points running in parallel, exchanging material in a GA manner. The difference is that only one set of penalties is used in GGA, whereas parallel GLSs could have used one independent set of penalties per run. Besides, learning in GGA is more selective than parallel GLS: the updating of penalties is only based on the best chromosome found at the point of penalization.

GLS Hybrids

GLS has been hybridized with single point-based metaheuristics [e.g., TS [8]] and population-based metaheuristics (e.g., evolutionary computation), creating new frameworks that have been applied to a wide range of applications. Guided tabu search (GTS), as proposed by Tarantilis *et al.* [23,24], is a combination of GLS with TS. The basic idea is to control the exploration of TS by a guiding mechanism, based on GLS, which continuously modifies the objective function of the problem. Another hybrid of GLS and TS is proposed in the work of De Backer *et al.* [25].

GLS can sit on top of the basic LS in ant colony optimization (ACO). This has been

Table 2. GLS Extensions and Hybrids

Extensions/Hybrids	Description
Extended guided local search	Incorporates aspiration criterion (inspired from TS) and/or randomness components to the GLS.
Guided genetic algorithm	GLS operating on top of GA; it uses features and their penalties to bias GA operators.
Guided variable neighborhood search (GVNS), guided large neighborhood search (GLNS)	GLS combined with sophisticated local search algorithms: VNS and LNS.
Guided tabu search	Incorporates into a tabu search framework a guiding mechanism (inspired from GLS).
EDA/GLS	EDA is hybridized with GLS to help the latter from being trapped in a small local area.
Guided evolution strategy	ES is hybridized with GLS to help the latter from being trapped in a small local area.
Guided constraint search	A fast constraint search by using the GLS approach to ignore unfavored values in a variable's domain.
Dist-GLS	A distributed version of GLS.

demonstrated by the work of Hani *et al.* [26], which proposes ACO_GLS, a hybrid ACO approach coupled with a GLS, and applies the algorithm to an industrial layout problem, which is modeled as a QAP.

Zhang *et al.* [27] proposed a framework that incorporates GLS within an estimation of distribution algorithm (EDA) framework, in which each solution in the population of EDA performs a GLS. The resulting framework outperforms the stand-alone EDA and GLS in the QAP. The hybrid of GLS, FLS, and evolution strategy (ES) was introduced by Mester and Braysy [28]. An iterative two-stage procedure was proposed, where GLS plays a role in both phases to improve the LS in the first stage and to regulate the objective function and the neighborhood of the modified ES in the second one. Penalty variable neighborhood, which is incorporated in the algorithm, resembles the FLS principle, by which only a small set of the neighbors of the penalized feature are considered by the LS.

GLS's ideas of features, penalties, and utilities have been incorporated into new algorithms. Guided constraint search, which is proposed by Gomes *et al.* [29], borrows ideas from GLS to enhance the efficiency of pure constraint programming (CP) methods. The basic idea is to define a utility function, penalty parameter, and cost for each feature

(pair of variable/value). According to the utility function, at each iteration, only the best n most promising values of each variable's domain are examined, defining a subspace for the CP method. Those features which were considered but do not belong to a new best solution returned by the CP method are deemed bad features to be penalized.

A summary of the main extensions and hybrids of GLS as presented in this section is given in Table 2.

APPLICATIONS

GLS has been successfully applied to a wide range of problems. This section attempts to provide a comprehensive survey of these applications which are grouped here into five categories of applications: routing and scheduling problems, assignment problems, resource allocation problems, constraint optimization problems, and continuous optimization.

Routing and Scheduling Problems

Significant results for GLS and FLS have been obtained in their application to problems of this category, particularly the TSP and the VRP.

Given the positions of n cities, the task in TSP is to find the shortest closed tour

in which all cities are visited only once. The Lin–Kernighan (LK) algorithm is a specialized algorithm for TSP that has long been perceived as the champion of this problem [30] with some of its newer iterated implementations [31] still improving further the original method. GFLS using the 2-opt move operator was tested against LK [32] in a set of benchmark problems from the public TSP library TSBLib [33]. Given the same amount of time, GFLS found better results than LK in average. GFLS also outperformed other metaheuristics, namely, implementations of SA [34], TS [35], and GA [36] reported on the TSP. One must be cautious when interpreting such empirical results as they could be affected by many factors, including implementation details. But given that the TSP is an extensively studied problem, it takes something special for an algorithm to outperform sophisticated algorithms such as LK under any reasonable measure (“find me the best results within a given amount of time” must be a realistic requirement). It must be emphasized that LK is specialized for TSP but GLS and FLS are much simpler general purpose algorithms. Other applications of GLS combined with memetic algorithms [37] and with dynamic programming-based search technique [38], to the TSP were also reported. The application of GLS has also been extended to other variants of TSP. Padron and Balaguer [39] have applied GLS to the related rural postman problem (RPP), Vansteenwegen *et al.* [40] applied GLS to the related team orienteering problem (TOP), and Mester *et al.* [41] applied the guided ES hybrid metaheuristic to the genetic ordering problems [a unidimensional wandering salesperson problem (UWSP)]. In addition, the approach of applying GLS to TSP was deployed to solve other optimization problems, namely, network planning problems [16,42] and query reformulation [43]. Another application of GENET includes rail traffic control [44].

The VRP is another major application of GLS in this category. In a VRP, one is given a set of vehicles, each with its specific capacity and availability, and a set of customers to serve, each with specific weight and/or time

demand on the vehicles. The vehicles are grouped in one or more depots. Both the depots and the customers are geographically distributed. The task is to serve the customers using the vehicles, satisfying time and capacity constraints. This is a practical problem. Like many practical problems, it is NP-hard. Kilby *et al.* applied GLS to VRPs and achieved outstanding results [45,46]. As a result, their work was incorporated in Dispatcher, a commercial package developed by ILOG [25]. Recently, GLS and its hybrids have been applied to several variants of the VRP. GLS has been applied to the VRP with backhauls and time windows [47], and the capacitated arc routing problem [48]. GTS has been applied to VRP with time window [23,24], and also extended to other variants of VRP, including VRP with two-dimensional loading constraints [49], VRP with simultaneous pickup and delivery [50], and VRP with replenishment facility [24]. GLS hybrids with VNS [15] and with ES [28] have been applied to large-scale VRPs.

Processors configuration problem is a scheduling problem, in which one is given a set of processors, each of which with a fixed number of connections. In connecting the processors, one objective is to minimize the maximum distances between processors. Another possible objective is to minimize the average distance between pairs of processors [51]. In applying GGA to the processors configuration problem, representation was a key issue. To avoid generating illegal configurations, only mutation is used. GGA found configurations with shorter average communication distance than those found by other algorithms reported previously [20,21].

Assignment Problems

A generic problem under this category is the generalized assignment problem (GAP). In this problem, the task is to assign agents to jobs. There are two main constraints that should be satisfied: each job can be handled by only one agent, and each agent has a finite resource capacity that limits the number of jobs that it can be assigned. Assigning different agents to different jobs bears different utilities. Different agents will consume different amounts of resources when doing

the same job. In a set of benchmark problems, GGA found results as good as those produced by a state-of-the-art algorithm (which was also a GA algorithm) by Chu and Beasley [52], with improved robustness [22].

GLS hybrids have been proposed to the related QAP. Zhang *et al.* [27] proposed the GLS/EDA hybrid metaheuristic. In addition, the hybrid of GLS with ACO (ACO GLS) has been applied to a variation of the QAP [26].

Resource Allocation Problems

Problems in this category include the path assignment problem [53], maximum channel assignment problem [54], workforce scheduling problem (WSP) [55], and others.

The WSP [55] is a resource allocation problem, in which the task is to assign technicians from various bases to serve the jobs, which may include customer requests and repairs, at various locations. Customer requirements and working hours restrict the times that certain jobs can be served by certain technicians. The objective is to minimize a function that takes into account the traveling cost, overtime cost, and unserved jobs. In the WSP, GFLS holds the best-published results in the benchmark problem available to the authors [56].

The knapsack problem is another optimization problem, which can be considered as a resource allocation problem. Hifi *et al.* [13] applied a variation of GLS to the multiple choice multidimensional knapsack problem, a variation of the knapsack problem.

Constraint Satisfaction/Optimization Problems

Several combinatorial optimization problems concern the search of a feasible solution with respect to a set of constraints. In some cases, finding such a solution is impossible, and then the task becomes finding a solution that minimizes the number of unsatisfied constraints (relaxations).

More specifically, given a set of propositions in conjunctive normal form, the satisfiability (SAT) problem is to determine whether the propositions can all be satisfied while the maximum-satisfiability problem (MAX-SAT) asks for the maximum number of clauses which can be satisfied. The weighted

MAX-SAT problem is a variation of the MAX-SAT problem in which each clause is given a weight. The task is to minimize the total weight of the violated clauses (or maximize the total weight of satisfied clauses).

SAT, MAX-SAT, and weighted MAX-SAT problems have received significant attention over the years [57]. GLSSAT, an extension of GLS, was applied to both the SAT and weighted MAX-SAT problem [58]. On a set SAT problems from DIMACS, GLSSAT produced results comparable to those produced by WalkSAT [59], a variation of GSAT [60], which was specifically designed for the SAT problem.

On a popular set of benchmark weighted MAX-SAT problems, GLSSAT produced better or comparable solutions, more frequently than state-of-the-art algorithms, such as DLM [61], WalkSAT [59], and GRASP [62].

The RLFAP is another constraint optimization problem. In RLFAP, the task is to assign available frequencies to communication channels satisfying constraints that prevent interference [63]. In some RLFAPs, the goal is to minimize the number of frequencies used. Bouju *et al.* [63] is an early work that applied GENET, a method related to GLS, to radio length frequency assignment. For the CALMA set of benchmark problems, which has been widely used, GFLS reported the best results compared to all work published previously [64]. In the NATO Symposium on RLFAP in Denmark, 1998, GGA was demonstrated to improve the robustness of GLS [21]. In the same symposium, new and significantly improved results by GLS were reported [11]. At the time, GLS and GGA held some of the best known results in the CALMA set of benchmark problems.

GLS and FLS have been successfully applied to the related 3D bin packing problem and its variants [65–67], VLSI design problem [68], the natural language parsing problem [69], and graph set T-colorings problem [70]. A distributed version of GLS has been applied to graph coloring [12].

In another application, Lee and Tam [71] and Stuckey and Tam [72] embedded GENET in logic programming languages in order to enhance programming efficiency. In these

logic programming implementations, unification is replaced by constraint satisfaction [73]. This enhances efficiency and extends applicability of logic programming. Hoos and Tsang [74] provide an overview of LS in CP.

Continuous Optimization Problems

GLS has been applied to general function optimization problems [75]. Results show that GLS spreads its search effort across solution candidates depending on their quality (as measured by the objective function). Besides, GLS consistently found solutions in a difficult landscape with many local optima.

SUMMARY AND CONCLUSIONS

For many years, general heuristics for combinatorial optimization problems with most prominent examples, the methods of SA and GAs, heavily relied on randomness to generate good approximate solutions to difficult NP-hard problems. The introduction and acceptance of TS [8] by the operations research community initiated an important new era for heuristic methods, where deterministic algorithms exploiting historical information started appearing and being used in real-world applications. GLS described in this article follows this trend. While TS is a class of algorithms (where a lot of freedom is given to the management of the tabu list), GLS is more prescriptive (the procedures are more concretely defined). GLS heavily exploits information (not only the search history) to distribute the search effort in the various regions of the search space. Important structural properties of solutions are captured by solution features.

Solutions features are assigned costs and LS is biased to spend its efforts according to these costs. Penalties on features are utilized for that purpose. When LS settles in a local minimum, the penalties are increased for selected features present in the local minimum. By penalizing features appearing in local minima, GLS not only avoids the local minima visited (exploiting historical

information), but also diversifies choices for the various structural properties of solutions captured by the solution features. Features of high costs are penalized more times than features of low cost: the diversification process is directed and deterministic rather than undirected and random.

The GLS and GFLS methods are still in their early stages and future research is required to develop them further. The use of incentives implemented as negative penalties, which encourage the use of specific solution features, is one promising direction to be explored. Other interesting directions include fuzzy features with indicator functions returning real values in the $[0, 1]$ interval, automated tuning of the λ or α parameters, definition of effective termination criteria, alternative utility functions for selecting the features penalized and also studies about the convergence properties of GLS.

REFERENCES

1. Voudouris C. Guided local search for combinatorial optimisation problems [PhD thesis]. Colchester: Department of Computer Science, University of Essex; 1997.
2. Davenport A, Tsang EPK, Wang CJ, *et al.* GENET: a connectionist architecture for solving constraint satisfaction problems by iterative improvement. Proceedings of the 12th National Conference for Artificial Intelligence; 1994 July 31–Aug 4; Seattle, NA. Menlo Park (CA): AAAI Press; 1994. pp. 325–330.
3. Tsang EPK, Wang CJ. A generic neural network approach for constraint satisfaction problems. In: Taylor JG, editor. Neural network applications. Springer; 1992. pp. 12–22.
4. Tsang EPK, Wang CJ, Davenport A, *et al.* A family of stochastic methods for constraint satisfaction and optimisation. Proceedings of the First International Conference on the Practical Application of Constraint Technologies and Logic Programming (PACLP), London. 1999. pp. 359–383.
5. Minton S, Johnston MD, Philips AB, *et al.* Minimizing conflicts: a heuristic repair method for constraint satisfaction and scheduling problems. *Artif Intell Med* 1992;58(1–3):161–205 (Special volume on constraint based reasoning).

6. Koopman BO. The theory of search, part III, the optimum distribution of search effort. *Oper Res* 1957;5:613–626.
7. Stone LD. The process of search planning: current approaches and continuing problems. *Oper Res* 1983;31:207–233.
8. Glover F, Laguna M. *Tabu search*. Boston (MA): Kluwer Academic Publishers; 1997.
9. Mills P, Tsang E, Ford J. Applying an extended guided local search to the quadratic assignment problem. *Ann Oper Res* 2003;118(1–4):121–135.
10. Bentley JJ. Fast algorithms for geometric traveling salesman problems. *ORSA J Comput* 1992;4:387–411.
11. Voudouris C, Tsang E. Solving the radio link frequency assignment problems using guided local search. *Proceedings of NATO Symposium on Frequency Assignment, Sharing and Conservation in Systems (AEROSPACE), AGARD, RTO-MP-13, Paper No. 14a; Aalborg, Denmark. 1998.*
12. Basharu M, Arana I, Ahriz H. Distributed guided local search for solving binary DisCSPs. *Proceedings of FLAIRS 2005. Menlo Park (CA): AAAI Press; 2005. pp. 660–665.*
13. Hifi M, Michrafy M, Sbihi A. Heuristic algorithms for the multiple-choice multidimensional knapsack problem. *J Oper Res Soc* 2004;55:1323–1332.
14. Tamura H, Zhang Z, Tang Z, *et al.* Objective function adjustment algorithm for combinatorial optimization problems. *IEICE Trans Fundam Electron Commun Comput Sci* 2006;E89-A(9):2441–2444.
15. Kytöjoki J, Nuortio T, Braysy O, *et al.* An efficient variable neighborhood search heuristic for very large scale vehicle routing problems. *Comput Oper Res* 2007;34(9):2743–2757.
16. Xiaohu T, Haubrich H-J. A hybrid meta-heuristic method for the planning of medium-voltage distribution networks. *Proceedings of 15th Power Systems Computation Conference (PSCC 2005), Liege, Belgium; 2005.*
17. Goldberg DE. *Genetic algorithms in search, optimization, and machine learning*. Reading (MA): Addison-Wesley Publishing Company, Inc.; 1989.
18. Holland JH. *Adaptation in natural and artificial systems*. Ann Arbor (MI): University of Michigan Press; 1975.
19. Lau TL. *Guided genetic algorithm [PhD thesis]*. Colchester: Department of Computer Science, University of Essex; 1999.
20. Lau TL, Tsang EPK. Solving the processor configuration problem with a mutation-based genetic algorithm. *Int J Artif Intell Tools* 1997;6(4):567–585.
21. Lau TL, Tsang EPK. Guided genetic algorithm and its application to the radio link frequency allocation problem. *Proceedings of NATO Symposium on Frequency Assignment, Sharing and Conservation in Systems (AEROSPACE), AGARD, RTO-MP-13, Paper No. 14b; Aalborg, Denmark. 1998.*
22. Lau TL, Tsang EPK. The guided genetic algorithm and its application to the general assignment problem. *IEEE 10th International Conference on Tools with Artificial Intelligence (ICTAI'98); Taiwan; 1998. pp. 336–343.*
23. Tarantilis CD, Zachariadis EE, Kiranoudis CT. A guided tabu search for the heterogeneous vehicle routing problem. *J Oper Res Soc* 2008;59(12):1659–1673.
24. Tarantilis CD, Zachariadis EE, Kiranoudis CT. A hybrid guided local search for the vehicle-routing problem with intermediate replenishment facilities. *INFORMS J Comput* 2008;20(1):154–168.
25. Backer BD, Furnon V, Shaw P, *et al.* Solving vehicle routing problems using constraint programming and metaheuristics. *J Heuristics* 2000;6(4):501–523.
26. Hani Y, Amodeo L, Yalaoui F, *et al.* Ant colony optimization for solving an industrial layout problem. *Eur J Oper Res* 2007;183(2):633–642.
27. Zhang Q, Sun J, Tsang EPK, *et al.* Combination of guided local search and estimation of distribution algorithm for solving quadratic assignment problem. *Bird of a Feather Workshops, Genetic and Evolutionary Computation Conference; Chicago (IL). 2003.*
28. Mester D, Braysy O. Active guided evolution strategies for large-scale vehicle routing problems with time windows. *Comput Oper Res* 2005;32(6):1593–1614.
29. Gomes N, Vale Z, Ramos C. Hybrid constraint algorithm for the maintenance scheduling of electric power units. *Proceedings of the International Conference on Intelligent Systems Application to Power Systems (ISAP 2003); Lemnos, Greece. 2003.*
30. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling salesman problem. *Oper Res* 1973;21:498–516.
31. Helsgaun K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. *Eur J Oper Res* 2000;126:106–130.

32. Voudouris C, Tsang EPK. Guided local search and its application to the traveling salesman problem. *Eur J Oper Res* 1999;113(2):469–499.
33. Reinelt G. A traveling salesman problem library. *ORSA J Comput* 1991;3:376–384.
34. Johnson D. Local optimization and the traveling salesman problem. *Proceedings of the 17th Colloquium on Automata Languages and Programming. Lecture Notes in Computer Science* 443. Berlin: Springer; 1990. pp. 446–461.
35. Knox J. Tabu search performance on the symmetric travelling salesman problem. *Comput Oper Res* 1994;21(8):867–876.
36. Freisleben B, Merz P. A genetic local search algorithm for solving symmetric and asymmetric travelling salesman problems. *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*. Nagoya: IEEE Press; 1996. pp. 616–621.
37. Holstein D, Moscato P. Memetic algorithms using guided local search: a case study. In: Corne D, Glover F, Dorigo M, editors. *New ideas in optimisation*. London: McGraw-Hill; 1999. pp. 235–244.
38. Congram RK, Potts CN. Dynasearch algorithms for the traveling salesman problem. Presentation at the Travelling Salesman Workshop, CORMSIS. University of Southampton. 1999.
39. Padron V, Balaguer C. New methodology to solve the RPP by means of isolated edge. In: Tuson A, editor. *2000 Cambridge Conference Tutorial Papers Young OR 11*. Cambridge: UK Operational Research Society; 2000.
40. Vansteenwegen P, Souffriau W, Berghe G, *et al.* A guided local search metaheuristic for the team orienteering problem. *Eur J Oper Res* 2009;196(1):118–127.
41. Mester DI, Ronin YI, Nevo E, *et al.* Fast and high precision algorithms for optimization in large-scale genomic problems. *Comput Biol Chem* 2004;28(4):281–290.
42. Flores Lucio G, Reed M, Henning I. Guided local search as a network planning algorithm that incorporates uncertain traffic demands. *Comput Netw* 2007;51(11):3172–3196.
43. Moghrabi I. Guided local search for query reformulation using weight propagation. *Int J Appl Math Comput Sci* 2006;16(4):537–549.
44. Jose R, Boyce J. Application of connectionist local search to line management rail traffic control. *Proceedings of International Conference on Practical Applications of Constraint Technology (PACT'97)*; London. 1997.
45. Kilby P, Prosser P, Shaw P. Guided local search for the vehicle routing problem with time windows. In: Voss S, Martello S, Osman IH, *et al.* editors. *Meta-heuristics: advances and trends in local search paradigms for optimization*. Boston: Kluwer Academic Publishers; 1999. pp. 473–486.
46. Kilby P, Prosser P, Shaw P. A comparison of traditional and constraint based heuristic methods on vehicle routing problems with side constraints. *Constraints* 2000;5(4):389–414.
47. Zhong Y, Cole MH. A vehicle routing problem with backhauls and time windows: a guided local search solution. *Transp Res Part E: Log Trans Rev* 2005;41(2):131–144.
48. Beullens P, Muyldermans L, Cattrysse D, *et al.* A guided local search heuristic for the capacitated arc routing problem. *Eur J Oper Res* 2003;147(3):629–643.
49. Zachariadis E, Tarantilis C, Kiranoudis C. A guided Tabu search for the vehicle routing problem with two-dimensional loading constraints. *Eur J Oper Res* 2008;195(3):729–743.
50. Zachariadis E, Tarantilis C, Kiranoudis C. A hybrid metaheuristic algorithm for the vehicle routing problem with simultaneous delivery and pick-up service. *Expert Syst Appl* 2008;36(2):1070–1081.
51. Chalmers AG. A minimum path parallel processing environment. *Research Monographs in Computer Science*. Bristol: Alpha Books; 1994.
52. Chu P, Beasley JE. A genetic algorithm for the generalized assignment problem. *Comput Oper Res* 1997;24:17–23.
53. Anderson CA, Fraughnaugh K, Parker M, *et al.* Path assignment for call routing: an application of tabu search. *Ann Oper Res* 1993;41:301–312.
54. Simon HU. Approximation algorithms for channel assignment in cellular radio networks. *Proceedings 7th International Symposium on Fundamentals of Computation Theory. Lecture Notes in Computer Science* 380. Berlin: Springer, 1989. pp. 405–416.
55. Azarmi N, Abdul-Hameed W. Workforce scheduling with constraint logic programming. *BT Technol J* 1995;13(1):81–94.
56. Tsang EPK, Voudouris C. Fast local search and guided local search and their application to British Telecom's workforce scheduling problem. *Oper Res Lett* 1997;20(3):119–127.
57. Gent IP, van Maaren H, Walsh T. SAT2000, Highlights of satisfiability research in the year

2000. Frontiers in artificial intelligence and applications. Netherlands: IOS Press; 2000.
58. Mills P, Tsang EPK. Guided local search for solving SAT and weighted MAX-SAT problems. *J Autom Reason* 2000;24:205–223.
 59. Selman B, Kautz H. Domain-independent extensions to GSAT: solving large structured satisfiability problems. Proceedings of 13th International Joint Conference on AI; Chambéry, France. 1993. pp. 290–295.
 60. Selman B, Levesque HJ, Mitchell DG. A new method for solving hard satisfiability problems. Proceedings of AAAI-92; San Jose (CA). 1992. pp. 440–446.
 61. Shang Y, Wah BW. A discrete lagrangian-based global-search method for solving satisfiability problems. *J Glob Optimization* 1998;12(1):61–99.
 62. Resende MGC, Feo TA. A GRASP for satisfiability. Cliques, coloring, and satisfiability: second DIMACS implementation challenge. In: Johnson DS, Trick MA, editors. Volume 26, DIMACS series on discrete mathematics and theoretical computer science. American Mathematical Society; 1996. pp. 499–520.
 63. Bouju A, Boyce JF, Dimitropoulos CHD, *et al.* Intelligent search for the radio link frequency assignment problem. Proceedings of the International Conference on Digital Signal Processing; Cyprus. 1995.
 64. Voudouris C, Tsang EPK. Partial constraint satisfaction problems and guided local search. Proceedings of PACT'96; London. 1996. pp. 337–356.
 65. Egeblad J, Nielsen B, Odgaard A. Fast neighborhood search for two- and three-dimensional nesting problems. *Eur J Oper Res* 2007;183(3):1249–1266.
 66. Faroe O, Pisinger D, Zachariasen M. Guided local search for the three-dimensional bin packing problem. Technical Report nr 99-13, Department of Computer Science, University of Copenhagen; 1999.
 67. Langer Y, Bay M, Crama Y, *et al.* Optimization of surface utilization using heuristic approaches. Proceedings of the International Conference COMPIT'05; Hamburg, Germany. 2005. pp. 419–425.
 68. Faroe O, Pisinger D, Zachariasen M. Guided local search for final placement in VLSI design. *J Heuristics* 2003;9:269–295.
 69. Daum M, Menzel W. Parsing natural language using guided local search. Proceedings of 15th European Conference on Artificial Intelligence (ECAI-2002); Lyon, France. 2002. pp. 435–439.
 70. Chiarandini M, Stutzle T. Stochastic local search algorithms for graph set T-colouring and frequency assignment. *Constraints* 2007;12(3):371–403.
 71. Lee JHM, Tam VWL. A framework for integrating artificial neural networks and logic programming. *Int J Artif Intell Tools* 1995;4:3–32.
 72. Stuckey P, Tam V. Semantics for using stochastic constraint solvers in constraint logic programming. *J Funct Logic Program* 1998;2.
 73. Tsang EPK. Foundations of constraint satisfaction. London: Academic Press; 1993.
 74. Hoos H, Tsang EPK. Local search for constraint satisfaction. In: Rossi F, van Beek P, Walsh T, editors. Handbook of constraint programming. Netherlands: Elsevier; 2006. pp. 245–277.
 75. Voudouris C. Guided local search an illustrative example in function optimisation. *BT Technol J* 1998;16(3):46–50.

HAZARD RATE FUNCTION

SATYANSHU KUMAR UPADHYAY
Department of Statistics, DST Centre
for Interdisciplinary Mathematical
Sciences, Banaras Hindu
University, Varanasi, India

INTRODUCTION

The statistical analysis of lifetime data has become a topic of immense importance to researchers and practitioners engaged in many diverse areas such as engineering, medicine, biology, and social sciences. The lifetimes, sometimes also referred to as the *failure times*, *survival times*, and so on, are the results of life-testing experiments, where prototype of items or organisms are subjected to stresses and the environmental conditions that characterize the intended operating circumstances. These experiments may be laboratory-conducted experiments or may also be considered as natural processes giving rise to lifetimes. In addition, the lifetime should not always be treated as the complete duration of life from birth to death of an item or an individual rather it can simply be taken to mean an observation of a nonnegative random variable. A few typical examples may include the study of the incubation period of a disease such as the AIDS, the remission time of a specific type of cancer, employment duration of an individual in a particular company, the length of marriages of divorce cases in a country, and the time to failure of an item or equipment in engineering considerations, number of revolutions before the failure of, say, ball-bearings, and so on.

Given the results of a life test experiment, the accounts of the theory of statistics start from the premises that the observations correspond to random variables and there

exists a family of probability distributions for these random variables. The true distribution, giving rise to the process under consideration, is an unknown member of the family and the objective of the analysis is connected with some aspects of the unknown true distribution. Therefore, in a life-testing experiment, the first and the foremost task is to formulate a mathematical model or a probability distribution that adequately represents some features of the process. If we carefully see the mechanism of the process involved in the occurrence of an event, we find that there are many causes acting either in isolation or in a group, which may be held responsible for the happening of an event at a particular instant. It may not often be possible to isolate these causes for their separate mathematical accountability and, as a result, the choice of actual lifetime distribution becomes an art. The attempt to distinguish among various available models merely based on the actual observed data may lead to outright wrong conclusions since these distributions often exhibit absolutely different behavior at the tails and the observed data are sparse mainly on the right tail. Therefore, we need to appeal to a concept that is capable of differentiating among various possible distributions. Such a concept is known in the literature of lifetime data analysis as the *hazard rate*. The hazard rate is also referred by several other names in different disciplines. For example, it is named as the *force of mortality* in actuarial science to compute mortality tables [1] and as *intensity function* in the extreme value theory [2]. In statistics, its reciprocal for the normal distribution is known as *Mill's ratio* [3]. Tukey [4] obtains qualitative results concerning order statistics from distributions with monotonic hazard rates; such distributions are referred to as the *subexponential*. The hazard rate considered as a function of time is known as the *hazard rate function*.

Definition and Interpretation of the Hazard Rate

The notion of the hazard rate is particularly useful in lifetime distributions as it describes the way in which the instantaneous rate of failure for an individual changes with time. To clarify the ideas, we shall consider a few basic concepts and discuss separately the cases for continuous and discrete variables.

Case of Continuous Variate. Let T be a nonnegative random variable representing the lifetimes or the time until the occurrence of an event although the time units might be taken as “cycles,” “stress,” and so on, especially with reference to engineering statistics. Let $f(t)$ denote the density function and $F(t)$ denote the cumulative distribution function of T . Obviously, $F(t) = P(T \leq t)$ gives the probability that an event has occurred by the time t . Another quantity of interest may be the survival function, which is defined as

$$S(t) = 1 - F(t) = P(T \geq t) = \int_t^\infty f(x) dx, \quad (1)$$

which gives the probability of being alive at duration t , or more generally, the probability that the event of interest has not occurred by the time t . $S(t)$ is also called the *reliability function* at time t , when one is concerned with the lifetimes of manufactured items. It is to be noted that in general $S(t)$ is a monotone nonincreasing left continuous function with $S(0) = 1$, and $\lim_{t \rightarrow \infty} S(t) = 0$.

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event. This can be defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2)$$

The numerator in Equation (2) represents the conditional probability of the occurrence of the event in the interval $(t, t + \Delta t)$ given that it has not occurred before time t , and the denominator is nothing but the width of the interval. Dividing the numerator and the denominator, the expression represents the rate of occurrence of the event per unit of time and taking the limit as the width of the

interval Δt goes to zero, we get the instantaneous rate of occurrence of the event at t given that it has not occurred to time t . In particular, $h(t)\Delta t$ is the approximate probability of occurrence of the event in the small interval of time t to $t + \Delta t$ given survival up to time t . If we put the usual definition of conditional probability for the numerator of Equation (2), the expression for $h(t)$ can be written as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)}. \quad (3)$$

Note that the joint probability T is in the interval $(t, t + \Delta t)$ and $T \geq t$ is, of course, the probability that T lies in the interval $(t, t + \Delta t)$. Thus $h(t)$ can be equivalently written as

$$h(t) = \frac{f(t)}{S(t)}. \quad (4)$$

It is easy to see that the hazard function $h(t)$ provides an alternative formulation for representing the distribution of T . In fact, since $f(t) = -\frac{dS(t)}{dt}$, Equation (4) can be rewritten as

$$h(t) = -\frac{d}{dt} \log S(t),$$

which on integrating and using $S(0) = 1$, yields

$$S(t) = \exp\left(-\int_0^t h(y) dy\right). \quad (5)$$

That is, for any positive real t , the value of the survival function $S(t)$, of T at t , may be regarded as $\exp(-\text{area under } h(\cdot) \text{ to the left of } t)$ although the interpretation in terms of area may not always be justified. In any case, an obvious question arises that what happens at infinity. Since $S(\infty) = 0$, it implies that $S(\infty) = \exp(-\text{total area under } h(\cdot))$ must be zero. Therefore, all hazard functions must satisfy the following conditions:

$$\begin{aligned} \int_0^t h(u) du &< \infty, \\ \int_0^\infty h(u) du &= \infty, \end{aligned}$$

and $h(t) \geq 0$ for all $t \geq 0$. Thus, the hazard rate should not be regarded in any way as a probability density although it has a probabilistic interpretation.

The hazard function $h(t)$ can also be used to obtain the probability density function $f(t)$. Thus, from Equation (4) using Equation (5), we can easily write

$$f(t) = h(t) \exp\left(-\int_0^t h(y) dy\right). \quad (6)$$

The argument of the exponential term in Equation (6) is defined as the cumulative hazard function or the integrated hazard function and it can be considered as the sum of the risks that one faces in going from the time zero to t . Thus,

$$S(t) = \exp(-H(t)), \quad (7)$$

where

$$H(t) = \int_0^t h(y) dy. \quad (8)$$

Like the hazard function, the cumulative hazard function is not a probability. However, it is also a measure of risk: the greater the value of $H(t)$, the greater is the risk of failure by the time t .

Case of Discrete Variate. Sometimes the random variable T may be treated as a discrete variable, especially when the lifetimes are given in the form of a group data or when they represent integral number of cycles of some form. Suppose T can take on values $t_1, t_2, \dots$, and so on, with $0 \leq t_1 < t_2 < \dots$ and let $\varphi(t_i)$ denote the probability that T equals $t_i, i = 1, 2, \dots$. Obviously, $\varphi(\cdot)$ can be referred to as the *probability mass function* for the discrete random variable T . The survival function $S(t)$ can then be written as

$$S(t) = P(T \geq t) = \sum_{i: t_i \geq t} \varphi(t_i). \quad (9)$$

Note that $S(t)$, as for the continuous case, is a monotone decreasing but left continuous function with $S(0) = 1$ and $S(\infty) = 0$. The

hazard function at t_i can now be defined as

$$h(t_i) = P(T = t_i | T \geq t_i) \\ = \frac{\varphi(t_i)}{S(t_i)}, \quad i = 1, 2, \dots \quad (10)$$

Moreover, if we use the fact that $\varphi(t_i) = S(t_i) - S(t_{i+1})$, Equation (10) can be written as

$$h(t_i) = 1 - \frac{S(t_{i+1})}{S(t_i)}, \quad i = 1, 2, \dots \quad (11)$$

Thus, corresponding to Equations (5) and (6), $S(t)$ and $\varphi(t_i)$ can be written as

$$S(t) = \prod_{i: t_i < t} [1 - h(t_i)], \quad (12)$$

$$\varphi(t_i) = h(t_i) \prod_{j=1}^{i-1} (1 - h(t_j)), \quad (13)$$

which suggests that the probability, survival and hazard functions provide equivalent specifications of the distribution of a discrete random variable T as it was noted in continuous case.

An analog of $H(t)$ can be defined for discrete case as well but the cumulative hazard function is not an appropriate name for it [5]. It is so because $H(t)$ in this case, does not equal to the cumulative sum of $h(t_i)$ ($i: t_i < t$). One can therefore, simply consider $H(t) = -\log(S(t))$, where $S(t)$ is given in Equation (12).

Occasionally, situations arise when the distribution of T may have both discrete and continuous components. In such cases, the density function is sum of both discrete and continuous parts. The hazard function in such cases can be defined as above, whereas the survival function involves product of terms similar to Equations (5) and (12), which as usual is a monotone nonincreasing left continuous function on $[0, \infty)$ [5].

Further Discussion. The probability density (mass) function and the survival function (or, more generally, the cumulative distribution function $P(T \leq t)$) are the common representations of the distribution of a random variable T . The readers might not be

comfortable with the hazard rate, although it is a more specialized characterization of the distribution of a random variable especially when dealing with the lifetime data. As an immediate use of the hazard rate function one can say that greater hazard rate implies shorter life on an average. Let us consider, for instance, that lifetimes T_1, T_2 of two different objects have hazard rates $h_1(t)$ and $h_2(t)$, respectively. Further suppose that $h_1(t) > h_2(t)$ for all $t > 0$. Then,

$$\int_0^t h_1(y) dy > \int_0^t h_2(y) dy,$$

which implies from Equations (1) and (4) that

$$1 - F_1(t) < 1 - F_2(t)$$

and, therefore,

$$E(T_1) < E(T_2)$$

since $E(T_1) = \int_0^\infty [1 - F_1(y)] dy$.

Often, in many applications, the qualitative information may be available on how the hazard rate changes with the amount of time when an individual or item is subjected to testing. This information can be used to define $h(t)$, which can then be translated to define the functional forms of probability density (mass) or cumulative distribution (or survival) functions. For example, there may be models with nondecreasing (nonincreasing) hazard rate shapes or there may be models with hazard functions having some other well-defined characteristics.

An Example Based on Exponential Distribution. The simplest lifetime distribution may be obtained by assuming a constant hazard rate over time so that $h(t) = \lambda$ for all t . The corresponding survival function using Equation (5) may be written as

$$S(t) = \exp(-\lambda t). \quad (14)$$

This distribution is called the *exponential distribution* with parameter λ . The density may be obtained using Equation (4) as

$$f(t) = \lambda \exp(-\lambda t). \quad (15)$$

This distribution plays a central role in lifetime data analysis, although it is probably too simple to be useful in many applications [5,6].

The semiconductor devices can be considered to be important examples which do seem to exhibit constant hazard rate behavior and hence their lifetimes can be expressed by exponential distribution. It is to be noted that constant hazard behavior of devices does not necessarily mean that such devices never fail rather it should be interpreted that the devices having this characteristic do not appear to age. Thus, it can be said that a 20-year-old semiconductor device is just as likely to fail in the next moment as a brand new such device.

Some Typical Shapes of the Hazard Rate Function

Different hazard rate shapes can provide different kinds of distributions. Distributions with monotone hazard rates are of considerable importance and such distributions constitute a very large class. If $h(t)$ is increasing (decreasing) in $t \geq 0$ then $f(t)$ is said to be increasing hazard rate IHR (decreasing hazard rate DHR) distributions. For a mathematical convenience and added generality, concavity (convexity) properties have also been used for defining IHR (DHR) distributions whether or not a density exists [7]. Hence, we say that a distribution F has a DHR if its support is of the form $[a, \infty)$, and $h(t)$ is decreasing in $t \geq a$. For convenience, we may take $a = 0$. If the distribution F has a density, it can be verified by differentiating $\log [1 - F(t)]$ that $h(t)$ is increasing if and only if the support of F is an interval, and if on that interval, $\log [1 - F(t)]$ is concave. Similarly, F is DHR if and only if the support of F is $[0, \infty)$ and $\log [1 - F(t)]$ is convex in $t \geq 0$ [3].

A well-known example can be the Weibull distribution with density

$$f(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{(\beta-1)} \exp \left[- \left(\frac{t}{\theta} \right)^\beta \right]; \theta, \beta, x > 0, \quad (16)$$

where θ defines the scale parameter and β determines the shape of the distribution. The

hazard rate of this distribution can be written as

$$h(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{(\beta-1)}. \quad (17)$$

Obviously, the Weibull density has IHR for $\beta > 1$, DHR for $\beta < 1$, and it coincides with the well-known exponential distribution for $\beta = 1$. It is to be noted that the form of the exponential distribution given in Equation (15) is a reparameterized form of the model that can be obtained from Equation (16) taking $\lambda = \theta^{-1}$ and $\beta = 1$.

Figure 1 provides the increasing, decreasing, and constant shapes of the hazard rate function with the corresponding densities shown in Fig. 2. These figures are drawn using Weibull distribution when the scale parameter θ is set to unity. A similar pattern can be observed when considering gamma distribution [6] with density

$$\frac{1}{\Gamma \kappa} \left(\frac{t}{\theta} \right)^{\kappa-1} \exp \left[- \left(\frac{t}{\theta} \right) \right]; \quad \theta, \kappa, t > 0.$$

That is, the gamma model is IHR for $\kappa > 1$, DHR for $\kappa < 1$, and it has a constant hazard rate for $\kappa = 1$. In the discrete case, a well-known example is provided by the negative binomial distribution with probability mass

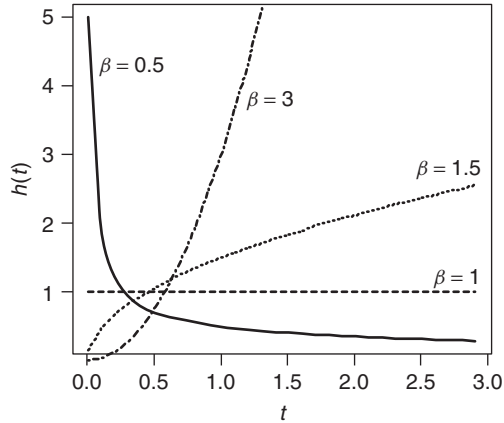


Figure 1. Hazard rate for the Weibull distribution showing increasing ($\beta > 1$), decreasing ($\beta < 1$) and constant trends ($\beta = 1$).

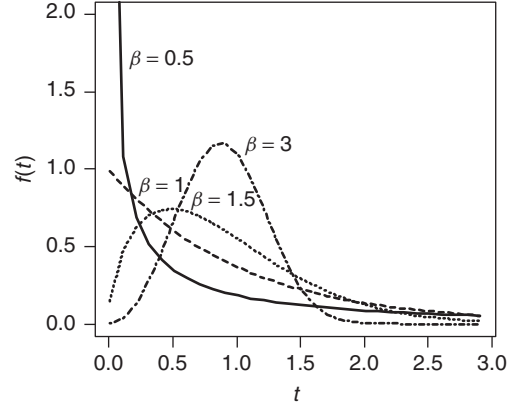


Figure 2. Probability density function of the Weibull distribution for different values of shape parameter.

function given by

$$\binom{-\alpha}{j} p^\alpha (p-1)^j; \quad \alpha > 0, 0 < p < 1, \\ \text{and } j = 0, 1, \dots$$

This model is also IHR for $\alpha > 1$ and DHR for $\alpha < 1$ and coincides with the geometric distribution for $\alpha = 1$. The case $\alpha = 1$ is characterized by the constant hazard rate [3].

A typical hazard rate may generally exhibit a so-called bathtub or “U”-shaped appearance. Such a hazard rate shape is shown in Fig. 3. If, for example, individuals

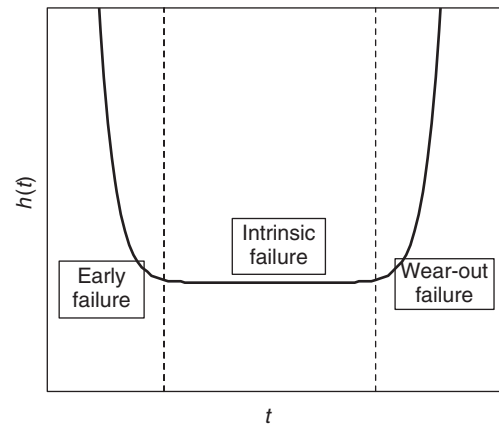


Figure 3. A typical bathtub-shaped hazard rate function.

in a population are followed right from their actual birth to death, such shapes are usually found appropriate. Bathtub or “U”-shaped hazard rate clearly indicates three distinct regions with two change points. The first, called the *initial failure region* or a region due to birth defect, is characterized as a DHR. In engineering context, this region represents early failures due to material or manufacturing defects. Good quality control and burn-in product testing usually eliminate many substandard devices, and thus avoid this high initial failure rate region. Similarly, we may follow the pattern in human population, where we often have early deaths resulting from birth defects or various other causes leading to high infant mortality. The second region (Fig. 3), called the *chance* or *random failure region*, is characterized by a near constant hazard rate. It represents chance failures caused by sudden stresses, unusually severe and unpredictable operating conditions, and so on. To eliminate these, one would require a device that is over designed for the vast majority of situations. The situation of a near constant hazard rate can be noticed in human population as well where it is usually observed that the death rate drops after attaining an initial age of, say, 5 years and then becomes relatively constant until the age of, say, 30 years or so. The third and the final segment in the bathtub shape is called the *wear-out failure region* and it is typified by an increasing failure rate (Fig. 3). In engineering context, such a situation is usually observed because of the equipment deterioration, accumulated shocks, fatigue, and so on. In human population, this deteriorating pattern in the form of IHR occurs primarily because of aging.

Modeling for the data sets that exhibit bathtub shapes has always been a tough task. There have been several proposals given in the literature to model such behavior [8–10]. One such proposal is to combine at most three different distributions, often the Weibull or some similar model, separately for decreasing, constant, and increasing shapes of the hazard rate. Sometimes combination of two different models may also result in the required bathtub shape. Generalization

or modification of Weibull density can be another important approach to obtain the bathtub shape of the hazard rate function. A number of models have been proposed in the literature using this approach. The key idea is simple. Consider the functional form of the hazard rate of Weibull distribution and modify this form so that it is capable of representing the bathtub shape as well. This modification can be done by either adding a new parameter in the hazard rate of the Weibull model or by manipulating the mathematical form of the hazard rate function so that it does provide bathtub shape as well. Exponentiated Weibull, Weibull extension, and modified Weibull models are a few important families obtained by generalizing or modifying the two-parameter Weibull family [11].

Model Detection Using Empirical Hazard Function

Given a model, one can always find the hazard rate function with the help of Equation (4), (10), or (11) using the fact that the concerned variate is continuous or discrete. A question, however, arises that can we get an approximate shape of the hazard rate function from the data in order to facilitate the tentative selection of a model. The answer to the question is “yes,” we can. This can be accomplished by obtaining the nonparametric estimate of the hazard rate function from the given set of observations and then drawing the estimated hazard function so obtained with respect to time [12]. This curve will help us to tentatively guess for a model for the data in hand. For example, if the curve so drawn is increasing, one can try to rely on the models such as Weibull and gamma, with shape parameter greater than unity. A word of remark: this is only a tentative guessing; one should always consider one of the various available tools for testing the goodness of fit of the observed data with the considered model. In addition, if more than one model is appearing proper for the data in hand based on the nonparametric estimated hazard plot, one can even try to compare the various proposed models in order to obtain the most appropriate candidate [13].

REFERENCES

1. Steffensen JF. Some recent researches in the theory of statistics and actuarial science. Cambridge: Cambridge University Press; 1930.
2. Gumbel EJ. Statistics of extremes. New York: Columbia University Press; 1958.
3. Barlow RE, Marshall AW, Proschan F. Properties of probability distributions with monotone hazard rate. *Ann Math Stat* 1963;34:375–389.
4. Tukey JW. A problem of Berkson, and minimum variance orderly estimators. *Ann Math Stat* 1958;29:588–592.
5. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley; 1980.
6. Lawless JF. Statistical models and methods for lifetime data. New York: Wiley; 1982.
7. Schoenberg IJ. On Polya frequency functions. *J Anal Math* 1951;1:331–374.
8. Krohn CA. Hazard versus renewal rate of electronic items. *IEEE Trans Reliab* 1969;18:64–73.
9. Mann NR, Schafer RE, Singpurwalla ND. Methods for statistical analysis of reliability and life data. New York: Wiley; 1974.
10. Wondmagegnehu ET, Navarro J, Hernandez PJ. Bathtub shaped failure rates from mixtures: a practical point of view. *IEEE Trans Reliab* 2005;54:270–275.
11. Murthy DNP, Xie M, Jiang R. Weibull models. New Jersey: Wiley-IEEE; 2004.
12. Martz HF, Waller RA. Bayesian reliability analysis. New York: Wiley; 1982.
13. Gupta A. MCMC based solutions of some problems related to bathtub shaped hazard rate function [Unpublished Ph.D. Thesis]. Banaras Hindu University, 2008.
- Bain LJ. Statistical analysis of reliability and life-testing models. New York: Marcel Dekker; 1991.
- Barlow RE, Proschan F. Mathematical theory of reliability. New York: Wiley; 1965.
- Cohen AC, Whitten BJ. Parameter estimation in reliability and life span models. New York: Marcel Dekker; 1988.
- Crowder MJ, Kimber AC, Smith RL, *et al.* Statistical analysis of reliability data. London: Chapman & Hall; 1991.
- Greenwich M. A unimodal hazard rate function and its failure distribution. *Stat Pap* 1992;33:187–202.
- Gross AJ, Clark VA. Survival distributions: reliability applications in the biomedical sciences. New York: Wiley; 1975.
- Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. New York: Wiley; 2002.
- Klutke GA, Kiessler PC, Wortman MA. A critical look at the bathtub curve. *IEEE Trans Reliab* 2003;52(1):125–129.
- Lawless JF. Statistical models and methods for lifetime data. 2nd ed. New York: Wiley; 2002.
- Nelson W. Applied life data analysis. New York: Wiley; 1982.
- Singpurwalla ND. Reliability and risk: a Bayesian perspective. New York: Wiley; 2006.
- Sinha SK. Reliability and life testing. New Delhi: Wiley; 1986.
- Wajid RA, Khan MS. Comparative distributions of hazard modeling analysis. *Pak J Stat Oper Res* 2006;2(2):127–134.

FURTHER READING

- Aalen OO, Gjessing HK. Understanding the shape of the hazard rate: a process point of view. *Stat Sci* 2001;16(1):1–14.
- Asher HE, Feingold H. Repairable system reliability. New York: Marcel Dekker; 1984.

HAZARDOUS MATERIALS TRANSPORTATION

VEDAT VERTER
Desautels Faculty of Management,
McGill University, Montreal,
Quebec, Canada

Hazardous materials (hazmats), such as gases, flammables, explosives, and radioactive materials are an integral part of our industrial lifestyle. In most cases, the point of production and consumption for these materials are different, and hence they need to be transported in significant volumes. For example, oil extracted from oil fields is sent to refineries, which ship their products (such as heating oil and gasoline) to storage tanks at different locations within a country. As another example, polychlorinated biphenyls (PCBs) are collected at many industrial installations, such as old power generation and transfer stations, and shipped to special waste management facilities for incineration. Accidental release of hazmats during transportation can be harmful to people, property, and the environment. Despite deregulation of transportation industry around the globe, governments at federal, provincial, and municipal levels remain involved in assessing and mitigating the public and environmental risks of dangerous goods shipments. Note that there are many stakeholders including the shippers, the carriers, the nongovernmental organizations, the public at large as well as the governmental agencies. To complicate hazmat logistics even further, different government levels often have different jurisdiction. For example in Canada, the railroads are under federal authority, the highways and hazmat storage sites are under provincial jurisdiction, whereas the city roads are governed by municipalities. A wide range of hazmat types with different impact zones are shipped between many origin–destination pairs in each jurisdiction. The economic viability of the transport sector is also a crucial

factor for regulators in designing policies to mitigate the public and environmental risks of dangerous goods.

In its 2005–2006 biennial report on hazardous materials transportation, the US Department of Transportation (US DOT) estimated that every day there are over 800,000 hazmat shipments in the United States, which amounts to about 9 million tons. The US Chemical Manufacturers Association forecast that the total volume of hazmat transportation will surpass 5 billion tons by 2020. The most recent Trucking Commodity Origin and Destination (TCOD) Survey by Statistics Canada, however, estimated that 106 million tons of dangerous goods were shipped by trucks in 2004, which amounts to 17.4% of all freight shipments across the Canadian roads (Provencher, 2008). Note that the survey focuses on for-hire trucks only, excluding the shipments carried by the companies' own trucks (e.g., Petro-Canada) and by foreign companies. In addition, 36 million tons of dangerous goods were moved by Canadian National (CN) and Canadian Pacific (CP) in 2004, and this represents 12.5% of all rail freight. The US DOT estimated that around 140 million tons of hazmats were carried via rail across the country in 2007. Overall, a significant majority of hazmat shipments are by road and rail, whereas marine shipments in bulk and in containers represent the third major mode of transportation.

There are about 40 major hazmat accidents in Canada and about 75 serious incidents in the US annually. In August 2008, for example, an explosion at Sunrise Propane's distribution facility in Toronto affected 580 homes and caused 1.5 million dollars in cleanup costs. There were also two fatalities and a number of nearby residents injured by the blast and flying debris. Note that this was only one of the 150 large propane distribution centers in Ontario. Within a few days of this accident, a CSX train derailment in Tennessee became a hazmat incident because one of the 17

derailed railcars—all carrying automotive products—hit a chemical storage facility owned by White House Utility district. Although such accidents are often not catastrophic, increasing shipment volumes raise the possibility of a gruesome event, such as the 1979 train derailment in Ontario that resulted in a chlorine leak and consequently required the evacuation of 200,000 people. A more recent example is the November 2005 collision in Sinaloa, Mexico that involved an ammonia truck and caused 39 fatalities.

The people living and/or working around the roads heavily used for dangerous goods shipments incur most of the transport risk. The policies that are available to government agencies for mitigating hazmat transport risk can be categorized into two main groups with respect to their underlying philosophy: proactive and reactive. The latter group of policies aims at confining the undesirable consequences of a hazmat incident after the occurrence of the accident. The development of emergency response plans—that involves the establishment of a network of responder teams specializing on different hazmat types—is the most popular example of reactive policies. The consequence of an accident can be reduced by better coordination of responder teams and faster response times to the incident site. In contrast, the proactive risk mitigation policies aim at reducing the likelihood and consequences of hazmat incidents *a priori*. The restrictions on the use of certain road segments by trucks carrying dangerous goods, the establishment of inspection centers for such trucks, and the improvements in the durability of railcar containers used for hazmat transportation are common examples in this category.

TRANSPORT MODES FOR HAZMAT

The majority of the literature on hazmat transportation focuses on road shipments. The development of reliable techniques for transport risk assessment [1–3] and the optimization of the routing decisions for potentially hazardous shipments [4,5] are the two main goals of these papers. Typically, a carrier's perspective is adopted by focusing

on transportation of a single type of hazmat between a single origin–destination (O–D) pair. Although train shipments can reach comparable levels to truck shipments from a total tonnage perspective, particularly in Europe and Canada, the literature on rail transport of hazmats is sparse. There are a number of differences between rail and highway routing of hazmat transportation. Rail infrastructure is typically owned and maintained by private rail companies. Consequently, railroad networks are sparse and do not contain as many potential alternative routes as highway networks. More importantly, railroads do not have tracks circumventing major population centers that are comparable to interstate beltways around metropolitan areas. A given shipment is likely to be handled by more than one railroad carrier, whereas truck shipments are usually limited to a single company. The rail carriers are motivated to maximize their portion of the movement. In a recent paper, Verma and Verter [6] highlighted additional differences between the two transport modes. A train usually carries nonhazardous and hazardous cargo together, whereas these two types of cargo are almost never mixed in a truck shipment. Furthermore, a rail tank car has roughly three times the capacity of a truck-tanker (80 tons and 25–30 tons, respectively), and the number of hazmat railcars varies significantly among different trains. Another important characteristic of trains is the possibility of incidents that involve multiple railcars.

The recent studies mostly focus on improving tank car safety through better container design, in an effort to reduce the likelihood of a release in the event of a derailment. By studying the risks associated with non-pressurized materials, Raj and Pritchard [7] report that the DOT-105 tank car design constitutes a safer option than DOT-111. Barkan *et al.* [8] show that tank cars equipped with surge pressure reduction devices experienced lower release rates than those without this technology. Barkan *et al.* [9] conclude that the speed of derailment and the number of derailed cars are highly correlated with hazmat release accidents. Recently, Barkan *et al.*

[10] study the trade-off between increased damage resistance and greater exposure to accidents, while minimizing the probability of release. They point out that the optimal container thickness varies with tank car configuration.

The comparison of rail and road from the viewpoint of hazmat transport risk attracted the attention of some researchers. In particular, Glickman [11] observed that the accident rate for significant spills (when release quantities exceed 5 gallons or 40 pounds) is higher for truck tankers than for rail tank cars, and that rail tank cars are more prone to small spills. Saccomanno *et al.* [12] showed that the safer mode varies with the hazmat being shipped, and differing volumes complicate comparison between the two transport modes. Leeming and Saccomanno [13] reported that although hazmat railway shipments pose more risk to residents in the vicinity of railroad tracks, the total risk of these two transport modes does not differ significantly. Their conclusion is based on a single case study in England. There is no consensus among researchers with regards to the best transport mode in terms of public and environmental safety. The transport mode choice varies with the type of hazmat to be carried.

Intermodal transportation is increasingly becoming popular since it enables the carriers to take advantage of the synergy between the scale economies in rail and the flexibility of road shipments. The US DOT estimates that multimodal transportation of dangerous goods will increase drastically in the near future. The transfer of freight from one mode to another takes place at intermodal terminals that require significant capital outlays at the construction phase. Consequently, the intermodal transport network design literature has focused on cost minimization, and we are aware of only two exceptions: Rondinelli and Berry [14] make a brief reference to the possible environmental impact of such facilities, and Taniguchi *et al.* [15] incorporate the resulting traffic congestion when deciding the location of similar facilities in Japan.

TRANSPORT RISK ASSESSMENT

The main concern associated with hazardous materials shipments is the undesirable consequences of an incident that involves the release of the cargo from its container. Typically, such incidents are caused by the involvement of the vehicle in an accident, such as a truck collision or a train derailment. The resulting leakage, fire, and/or explosion may cause fatalities, major and minor injuries as well as environmental damage. In an effort to minimize such undesirable consequences, the emergency response guidelines mandate the initial evacuation of a certain area around the incident site. The size of the evacuation area varies with the type of hazmat involved in the accident. Some hazmats are not only extremely dangerous, but they can travel long distances upon being airborne. Chlorine, for example, can travel up to 10 km depending on the wind conditions. The residents who live near the roads and rail tracks that are often used for moving hazmats are exposed to the risk of suffering the undesirable consequences of an incident. The public's resistance to the use of nearby routes for hazmat transportation usually depends on their perception of the associated risks rather than the expert assessments.

Various definitions of transport risk have been proposed in the earlier literature: a popular measure is the number of people living within a threshold distance from the routes used by hazmat trucks. This model is originally suggested by Batta and Chiu [16] and it emphasizes population exposure to hazmats rather than the occurrence of an incident. Revelle *et al.* [17] used a weighted combination of population exposure and transportation cost in finding routes for radioactively contaminated fuel rods. Alternatively, incident probability is suggested as a risk measure by Saccomanno and Chan [18] and Abkowitz *et al.* [19]. This model focuses on the likelihood of having a hazmat incident during transportation and ignores the possible undesirable consequences. Consequently, it is more suitable for the hazmats with relatively small danger zones. The expected risk model, which is also endorsed by the US

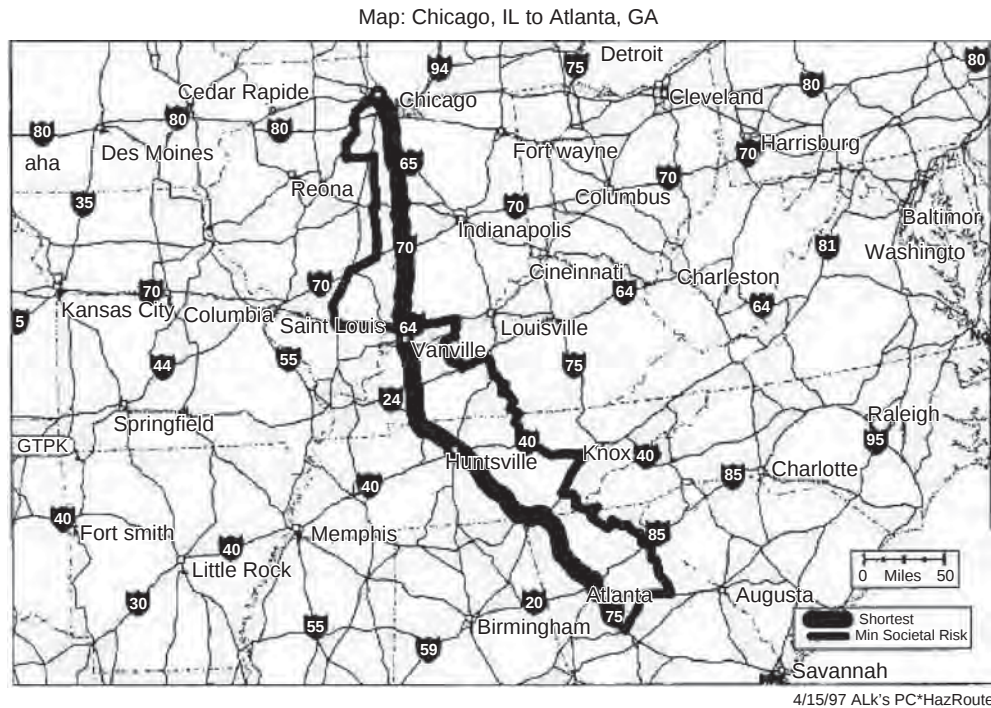


Figure 1. Alternative routes between Chicago and Atlanta. [Source: Adapted from Erkut and Verter [21].]

DOT provides a means to incorporate both the probability and the consequence of a hazmat incident [20]. The single commodity, single origin–destination problems studied in these papers are typically faced by hazmat carriers and they can be reduced to a shortest path model using the risk measure as arc impedance.

Through a detailed analysis of the United States road network, Erkut and Verter [21] demonstrated that the optimal routes for potentially hazardous trucks vary considerably with the risk representation in a mathematical model, and hence the choice of risk measure may have important repercussions. For example, Fig. 1 depicts the path that minimizes the expected risk of a hazmat shipment from Chicago to Atlanta as well as the shortest path for this (O–D) pair. Note that these two paths are geographically very different. Erkut and Verter [21] also showed that the best path for a transport risk

measure could perform poorly under some of the other risk measures.

Verter and Kara [22] incorporated risk assessment techniques in a GIS framework and presented a critical evaluation of the hazmat transportation activity in Quebec and Ontario. To date, this paper remains as the largest-scale risk assessment study in the scholarly literature. There are more than 100,000 gasoline truck shipments in these two Provinces annually. Figure 2 depicts the possible outcome of routing all these shipments according to four different criteria. The thickness of each road segment is proportional to the number of gasoline trucks using that link. Note that the routes that minimize the total distance and the total incident probability are similar, because the highest quality roads with the least likelihood of an accident are those that connect the urban centers. The apparent similarity between the minimum population exposure and the minimum expected risk routes is due to the fact

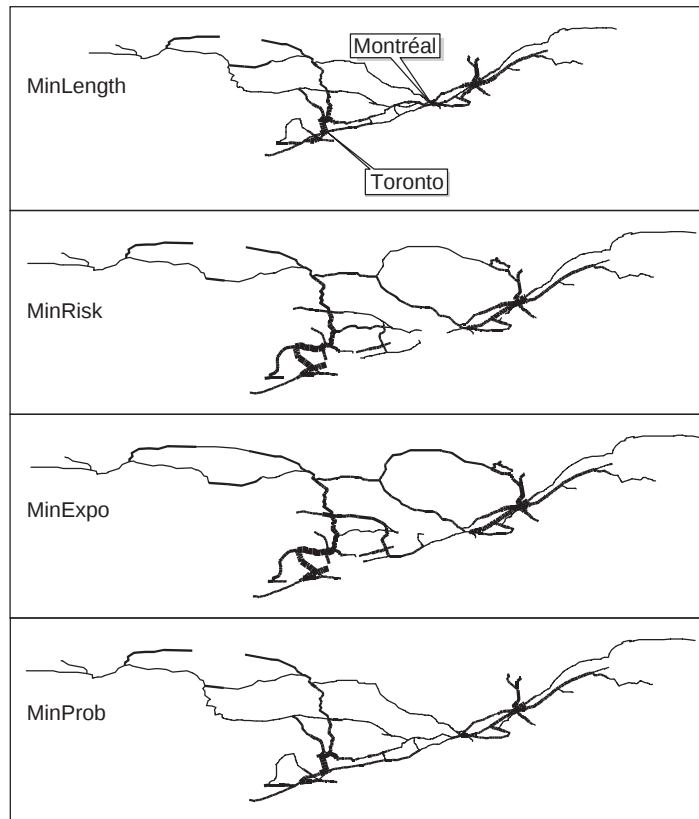


Figure 2. Alternative routing of gasoline shipments in Ontario and Quebec. [Source: Adapted from Verter and Kara [22].]

that both of these risks are driven by the spatial distribution of the population.

Verter and Kara [22] highlighted the significance of the topology of the road network by showing that a risk mitigation measure that would be effective in Toronto (i.e., the use of minimum population exposure routes) would not be as effective in Montreal, since the hub-shaped road network does not allow the trucks to bypass the heavily populated island.

For more information on health and environmental risk assessment.

THE REGULATORS' PERSPECTIVE

There are only a few exceptions to the carrier-oriented perspective of the prevailing literature. List and Mirchandani [23] presented the first multicommodity formulation for routing hazmats and locating hazardous

waste treatment facilities. The application of their model in Albany, New York, is based on a number of simplifications, including the representation of each population center as a point and the shipment of a single type of hazmat. Iakovou *et al.* [24] provide a multicommodity network flow model for the problem of routing hazardous vessels. Their model is applied to the transportation of crude oil and refined petroleum products in the Gulf of Mexico. The aim was to avoid overloading certain links of the transport network, which usually happens when the optimal route for each shipment is identified independent of the other shipments.

Kara and Verter [25] presented the first mathematical formulation for the problem of identifying the road segments that need to be closed for hazmat trucks. In the context of the road network of Western Ontario, they compare alternative policies for regulating the use of road segments. Recognizing the extent

of the cost-risk trade-off between the governments' and carriers' ideal solutions, Verter and Kara [26] developed a formulation that can generate compromise solutions for the hazmat road network design problem. Marcotte *et al.* [27] demonstrated that the use of tolls for hazmat trucks is a more efficient policy for mitigating transport risk than prohibiting the use of certain road segments. Following up on Kara and Verter [25], Erkut and Alp [28] propose an alternative formulation as a Steiner tree selection problem and proposed a greedy heuristic. This topology takes away the carriers' freedom in route selection and reduces the bi-level problem to a single level. Erkut and Gzara [29] consider a bi-objective cost and risk minimization version of the network design problem, and present a heuristic algorithm that exploits the network flow structure at both levels.

Focusing on other policy tools available to policy makers, Berman *et al.* [30] develop a model to determine the best locations for specialized hazmat emergency teams, whereas Verter and Erkut [31] highlight insurance requirements as an effective means for influencing major hazmat carriers' route choices. The former study shows that it is possible to drastically improve the responsiveness of specialized emergency response teams in the Quebec–Windsor corridor by relocating their stations. The latter study, however, points out that the insurance requirements can be effective in incentivizing the use of routes with less transport risk only for the carriers with large fleets. Gendreau *et al.* [32] develop heuristics for determining the location of inspection stations to be established on a network so as to ensure compliance with the regulator's requirements pertaining to route choices, maximum load and speed, driver's fatigue level, insurance requirements, container specifications, and so on.

AVENUES FOR FUTURE RESEARCH

In closing this introductory article on hazardous materials transportation, the reader is referred to Erkut *et al.* [33] for a comprehensive review of the recent research on the topic. For an overview of the earlier

literature, the reader is referred to Erkut and Verter [34] and List *et al.* [35]. This section highlights two fruitful avenues for future research: (i) transportation of hazmats in urban centers, and (ii) the use of intermodal transportation for dangerous goods shipments.

The prevailing research on truck shipments of dangerous goods is focused on highway transportation rather than the shipments through congested city streets. A significant majority of the trucks carrying potentially dangerous cargo, however, enters urban areas to make their delivery, for example, to gas stations. There is a need for the development of analytical approaches that incorporate the differentiating characteristics of urban transportation. In particular, the traffic congestion on the streets and the dynamic patterns of population distribution within urban centers need to be incorporated in transport risk assessment and truck routing models. The development of new risk assessment methods for shipments through urban streets is needed. In this domain, the potential harm to the other people on the road is as significant as the risk imposed on the people living within a threshold distance from the road. Note that the latter has been the sole focus of the prevailing risk assessment methods for highway and shipments. From the perspective of the regulator, risk mitigation policies need to be adapted to the urban environment. For example, the arrival of specialized emergency response teams to the incident site might be delayed due to heavy traffic congestion on the city roads. Also, the implementation of toll pricing policies for the use of certain city road segments by dangerous goods trucks may become a necessity [27].

The other potentially significant area for future research is the use of intermodal transportation networks for hazmat shipments. As a means of competing with the trucking industry, major railway companies are increasingly investing in intermodal transportation (see *Decision Problems and Applications of Operations Research at Marine Container Terminals*). This gives them the ability to benefit from the flexibility of short haul truck shipments (i.e., inbound

and outbound drayage), while enjoying the scale economies in long haul rail shipments. Note that the transfer of the cargo from trucks to railcars is done at intermodal terminals, and many of these yards are being built near (or within) urban centers. The alternative strategies regarding the construction of multimodal transport yards, which are becoming increasingly common for dangerous goods shipment, need to be studied. Given their significant impact on the spatial distribution of transport risk, these fixed facilities need to be a significant element of the risk assessment and policy development frameworks. Early work on intermodal transportation of hazmats identified the major risk and cost trade-offs between the drayage and rail operations [36]. Although intermodal transportation constitutes a major opportunity for improving hazmat logistics, the arising problems are often quite complex from a methodological perspective.

REFERENCES

- Glickman TS. An expeditious risk assessment of the highway transportation of flammable liquids in bulk. *Trans Sci* 1991;25(2):115–123.
- Abkowitz M, Cheng PDM. Hazardous materials transport risk estimation under conditions of limited data availability. *Trans Res Rec* 1989;1245:14–22.
- Zhang JJ, Hodgson J, Erkut E. Using GIS to assess the risks of hazardous materials transport in networks. *Eur J Oper Res* 2000;121(2):316–329.
- Akgün V, Parekh A, Batta R, *et al.* Routing of a hazmat truck in the presence of weather systems. *Comput Oper Res* 2007;34(5):1351–1373.
- Erkut E, Ingolfsson A. Catastrophe avoidance models for hazardous materials route planning. *Trans Sci* 2000;34(2):165–179.
- Verma M, Verter V. Railroad transportation of dangerous goods: population exposure to airborne toxins. *Comput Oper Res* 2007;34:1287–1303.
- Raj PK, Pritchard EW. Hazardous materials transportation on U.S. railroads. *Trans Res Rec* 2000;1707:22–26.
- Barkan CPL, Treichel TT, Widell GW. Reducing hazardous materials releases from railroad tank car safety vents. *Trans Res Rec* 2000;1707:27–34.
- Barkan CPL, Dick CT, Anderson R. Railroad derailment factors affecting hazardous materials transportation risk. *Trans Res Rec* 2003;1825:64–74.
- Barkan CPL, Ukkusuri S, Waller S. Optimizing railroad tank cars for safety: the tradeoff between damage resistance and probability of accident involvement. *Comput Oper Res* 2007;34:1266–1286.
- Glickman TS. Benchmark estimates of release accident rates in hazardous materials transportation of rail and truck. *Trans Res Rec* 1988;1193:22–28.
- Saccomanno FF, Shortreed JH, Van Aerde M, *et al.* Comparison of risk measures for the transport of dangerous commodities by truck and rail. *Trans Res Rec* 1989;1245:1–13.
- Leeming DG, Saccomanno FF. Use of quantified risk assessment in evaluating the risks of transporting chlorine by road and rail. *Trans Res Rec* 1994;1430:27–35.
- Rondinelli D, Berry M. Multimodal transportation, logistics, and the environment: managing interactions in a global economy. *Eur Manage J* 2000;18(4):398–410.
- Taniguchi E, Noritake M, Yamada T, *et al.* Optimal size and location planning of public logistics terminals. *Trans Res Part E* 1999;35:207–222.
- Batta R, Chiu SS. Optimal obnoxious paths on a network: transportation of hazardous materials. *Oper Res* 1988;36:84–92.
- Revelle C, Cohon J, Shobrys D. Simultaneous siting and routing in the disposal of hazardous wastes. *Trans Sci* 1991;25(2):138–145.
- Saccomanno FF, Chan A. Economic evaluation of routing strategies for hazardous road shipments. *Trans Res Rec* 1985;1020:12–18.
- Abkowitz M, Lepofsky M, Cheng P. Selecting criteria for designating hazardous materials highway routes. *Trans Res Rec* 1992;1333:30–35.
- Alp E. Risk-based transportation planning practice: overall methodology and a case example. *INFOR* 1995;33:4–19.
- Erkut E, Verter V. Modeling of transport risk for hazardous materials. *Oper Res* 1998;46(5):625–642.
- Verter V, Kara BY. A GIS-based framework for hazardous materials transport risk assessment. *Risk Anal* 2001;21(6):1109–1120.

23. List GF, Mirchandani PB. An integrated network/planar multiobjective model for routing and siting for hazardous materials and wastes. *Trans Sci* 1991;25(2):146–156.
24. Iakovou E, Douligieris C, Li H, *et al.* A maritime global route planning model for hazardous materials transportation. *Trans Sci* 1999;33(1):34–48.
25. Kara BY, Verter V. Designing a road network for hazardous materials transportation. *Trans Sci* 2004;38(2):188–196.
26. Verter V, Kara B. A path-based approach for the hazmat transport network design problem. *Manage Sci* 2008;54(1):29–40.
27. Marcotte P, Mercier A, Savard G, *et al.* Toll policies for mitigating hazardous materials transport risk. *Trans Sci* 2009;43(2):228–243.
28. Erkut E, Alp O. Designing a road network for dangerous goods shipments. *Comput Oper Res* 2007;34(5):1389–1405.
29. Erkut E, Gzara F. Solving the hazmat transport network design problem. *Comput Oper Res* 2008;35(7):2234–2247.
30. Berman O, Verter V, Kara BY. Designing emergency response networks for hazardous materials transportation. *Comput Oper Res* 2007;34(5):1374–1388.
31. Verter V, Erkut E. Incorporating insurance costs in hazardous materials routing models. *Trans Sci* 1997;31(3):227–236.
32. Gendreau M, Laporte G, Parent I. Heuristics for the location of inspection stations on a network. *Nav Res Log* 2000;47(4):287–303.
33. Erkut E, Tjandra S, Verter V. Hazardous materials transportation. In: Barnhart C, Laporte G, editors. Volume 14, *Handbooks in operations research & management Science*. The Netherlands: Elsevier; 2007. pp. 539–621.
34. Erkut E, Verter V. Hazardous materials logistics. In: Drezner Z, editor. *Facility location: a survey of applications and methods*. New York: Springer; 1995. Chapter 20.
35. List GF, Mirchandani PB, Turnquist MA, *et al.* Modeling and analysis for hazardous materials transportation - risk analysis, routing scheduling and facility location. *Trans Sci* 1991;25(2):100–114.
36. Verma M, Verter V. A lead-time approach to intermodal transportation of hazardous materials. *Eur J Oper Res* 2010;202:696–706.

HELLENIC OPERATIONAL RESEARCH SOCIETY

NIKOLAOS MATSATSINIS
Department of Production Engineering and
Management, Technical University of
Crete, Crete, Greece

HISTORICAL ASPECTS

The Hellenic Operational Research Society (HELORS—www.eeee.org.gr) was founded in 1963 by pioneering Greek scientists, with the objective of promoting the study and applications of operational research (OR) methodology, for the benefit of the Hellenic economy and society. HELORS' corporate headquarters is located in Athens, where the administrative council sits in privately-owned offices.

HELORS is a member of the "Association of European Operational Research Societies—EURO" within the "International Federation of Operational Research Societies," (IFORS).

The society promotes and disseminates OR/scientific management by every possible means, as a tool for making rational and documented decisions. In the 46 years since its inception, HELORS has evolved into a scientific entity with an important presence in the scientific and economic life of the country, with hundreds of members who are outstanding for their theoretical backgrounds, entrepreneurial endeavors, and professionalism.

HELORS initially went through a period (1963–1980) where most of its efforts lay in the dissemination and advancement of the methods and techniques of OR to organizations and staff of the private and public sectors. To meet these objectives, seminars, lectures, and conferences were organized under the auspices of HELORS that extended the use of OR in making rational

decisions in topics such as strategy, political organization, and managerial control.

Current results of a company's activities, *inter alia*, are the employment of operational researchers in businesses and organizations, national defense general staff, and elsewhere. The emergence of the value and importance of OR from HELORS led to the introduction of relative courses in universities. As a consequence, numerous students became acquainted with the principles and the philosophy of OR, and in turn, commonly apply them in their employment environments. Enterprise and organization managers, by experiencing the constructive results of operational research work, recognized OR as a valuable tool of decision making and problem solving.

Subsequently, a second creative and fruitful period follows (1980–today), in parallel with socioeconomic, technological, and political changes in Europe and other parts of the world. These changes influence significantly the Hellenic economy, which should in turn, become more productive and competitive [1–5].

In 1984, the Macedonia–Thrace annex was founded, aiming primarily at the growth of OR in the greater area of Balkans and at an improved organization and communication of the members of northern Greece. This annex enumerates more than 120 members and is located in Thessaloniki. Among other activities, the Macedonia–Thrace annex organizes four different Balkan conferences, various seminars in industrial management and regularly participates in educational programs of neighboring Balkan countries when funded by the European Union.

GOALS

Adaptable to current trends, HELORS' main objective is to participate and contribute substantially to business and organization development of the private and as the public sectors, by upgrading their human

resources. This can only be achieved with appropriate education and training. Thus, HELORS organizes educational seminars, lectures of broad social thinking and business, as well as conferences dedicated to functional demands of businesses, and so on, and hence contributing to the establishment of crucial relationships between expert researchers of the field of management and the wider social community.

This new reality extends HELORS field of action which approaches better the enterprise and organization; presents and analyses a broad spectrum of recent management, marketing, decision making, financial, and new technology practices. HELORS' prestige and successful choices of people ensures the quality and success of its events. With all these activities, HELORS has gained an outstanding position in the field of scientific management. Today, HELORS is recognized for

- educating business managers;
- organizing conferences and workshops in topics related to crucial issues in management science;
- presenting economic and social reality.

The conditions under which HELORS will operate in the next few years are expected to create

- a Greece that is more "European," with the upgrading of institutions, management, the way of thinking and acting;
- enterprises and organizations that are more rationally organized, that will have to a greater extent assimilated modern management and information technologies;
- a very competitive market for management staff where greater importance will be given to the timely update of information and new technologies and the opportunities for exchange of ideas and experiences, especially at an international level.

These predicted conditions define HELORS' strategic choices as these are

shaped by the Executive Council and the members' General Council and are summarized by the following triptych:

- extroversion, participation in common issues, development of initiatives for the researching of problems, cooperation with other associated bodies, further strengthening of international contacts;
- close cooperation with dynamic enterprises in a mutual relationship, support for managerial staff development programs and the implementation of new technologies and ensuring material and moral support since the objective is for enterprises to determine that HELORS is "their issue";
- ensure that the members of HELORS receive all the benefits that provide for professional achievement and social recognition.

By all indications, the aspirations of earlier generations of management scientists for the emergence of a new way of thinking (i.e., systematic approach to the problems of businesses and justified business decision making) are gradually fulfilled in our country.

The large family of operational researchers should continue to play a catalytic role, which has faithfully served over the years, to account for the human potential of the field of science management and to enhance business practices by adopting modern technologies and methods.

HELORS will continue its effort for the consolidation of all management scientists who have the knowledge and the willingness to develop initiatives for further development and exploitation of these resources. The service package for their members will include:

- information on international and local activities;
- information on technological advances in their field through seminars, workshops, lectures, and so on;
- exchange of experiences on the subject of their work through workshops, work groups, and seminars;

- member participation in continuous training, either as lecturers, or as participants;
- provision of specialized software from the HELORS offices, special library, and so on.

With these services to its members, HELORS will try to help in the promotion of the modern, systematic way of thinking in addressing problems, professional competence in complex roles, and social sensitivity in order to upgrade our economy.

MEMBERS

Today the number of members of HELORS stands at 750. This human potential is the most important element of the assets and comprises its invaluable capital. The number of members, ranks HELORS as the fifth largest among the countries of IFORS (41 countries, in which HELORS has been a member since 1966).

Members are mainly engineers, economists, mathematicians, and so on. Over 60% of the members hold a Master's degree in operational research/business administration, and 15% hold a doctoral degree.

The important input that operational researchers have provided is recognized and demonstrated by the fact that many members of this organization are now senior executives in private companies and organizations, and several other members have repeatedly held top management positions in public organizations.

Also many members of the company have become professors in higher education institutions and by their educational and research work, they contribute efficiently to the establishment of OR as a management problem-solving tool. Members have published theoretical and applied articles on OR, in prestigious Greek and foreign scientific journals. The new members easily find employment in the private or public sector.

ACTIVITIES

The following are indicative activities of the HELORS.

Presidents of HELORS

The following are the Presidents of HELORS:

1963–1974:	General Radamathis Spanogiannakis, Founder of HELORS
1974–1976:	Professor Dimitris Ksirokostas, National Technical University
1976–1978:	Professor Ioannis Pappas, National Technical University
1979–1980:	Professor Dimitris Xirokostas, National Technical University
1981–1984:	Professor Dimitris Blesios, University of Piraeus
1985–1988:	Professor Dimitris Papoulias, Kapodistrian University of Athens
1989–1990:	Professor Konstantinos Papis, University of Piraeus
1991–1994:	Konstantinos Vasiliadis, Executive Manager
1995–1996:	Spiros Pashentis, Executive Manager
1997–1998:	Professor Konstantinos Papis, University of Piraeus
1999–2004:	Professor Ioannis Siskos, University of Piraeus
2005–2008:	Professor Dionisios Giannakopoulos, Technical Institute of Piraeus
2009...	Professor Nikolaos Matsatsinis, Technical University of Crete

Golden Medals

HELORS has decided to establish the award “Gold Medal of Operational Research” for distinguished Greek scientists, who have made important advances in operations research. “Gold Medals of Operational Research” have been awarded to the following:

1999	General Radamathis Spanogiannakis, Hellas
2000	Professor Dimitris Bertsekas, Massachusetts Institute of Technology (MIT), United States
2001	Professor Panagiotis Pardalos, University of Florida, United States
2002	Professor Spiros Makridakis, INSEAD, France
2003	Professor Dimitris Ksirokokostas, National Technical University, Hellas
2004	Professor Dimitris Bertsimas, MIT, United States
2005	Professor Ioannis Siskos, University of Pireaus, Greece
2006	Professor Biron Papathanasiou, Aristotle University of Thessaloniki, Greece
2007	Professor Evangelos Pashos, University Paris-Dauphine, France
2008	Professor Aristidis Sissouras, University of Patras, Greece
2009	Professor Konstantinos Zopounidis, Technical University of Crete, Greece

Work Groups

For better organization and efficiency, the HELORS operates in working groups. Currently, the following working groups have been approved and operate specialized in different topics:

- Multicriteria Decision Analysis (Group coordinator: Yannis Siskos), and

- OR in health (Group coordinator: Prof. Aris Sissouras).

The Multicriteria Decision Analysis group has already organized seven meetings in multicriteria analysis and one seminar on multiple criteria decision aid.

In the near future, more working groups are expected to be approved and become operational.

Publications

Scientific Journal. HELORS issues the scientific journal “Operational Research: An International Journal–ORIJ.” ORIJ publishes high quality scientific papers that contribute significantly to the fields of OR/MS. ORIJ covers all aspects of OR including optimization methods, decision theory, stochastic models, simulation, game theory, queueing systems, inventory and reliability, among others. Articles published in ORIJ present new theoretical insights and developments, as well as real-world case studies illustrating the practical implementation of OR. Contents include papers exploring the interactions of OR/MS with other relevant disciplines such as information technology, computer science, artificial intelligence, soft computing, and electronic services. This is a unique feature of ORIJ as compared to other existing OR journals and provides a means to explore new directions in OR/MS research in an interdisciplinary context. ORIJ also publishes overview papers from eminent scientists in significant fields of OR/MS; these review the state-of-the-art in these fields. As of 2008, ORIJ is published by Springer (<http://www.springerlink.com/content/1109-2858>).

Conference Proceedings. HELORS issues proceedings of the conference that it organizes.

Newsletter. Every 3 months the HELORS prepares an electronic press releases (newsletter) which includes:

- activities of HELORS,
- HELORS conferences,

- OR—an international journal,
- upcoming surveys,
- call for papers in European and international conferences,
- call for papers in special issues magazine in EU,
- fellowships on EU issues outside Greece,
- links to interesting sites,
- EU journals.

Conferences

Twenty-one national conferences have been organized since 1977, covering the current problems of the Hellenic economy and society. The choice of topics, exchange ideas, and intense questioning during the conference contributed significantly to the promotion of OR in our country and to the promotion of the Hellenic operational researcher in Hellenic businesses. HELORS has also organized six special conferences in more specialized topics.

More than 4500 people have participated in HELORS conferences. More than 1000 announcements have been made during these conferences and can be divided into applied research works and those that promote basic scientific research. Business executives, professors, and executives of public sector have made announcements in topics related to strategy, organization, management and control, technology, and so on.

The participants have benefited in multiple ways from the conferences, by listening to the opinions and attitudes of other members and thus, gaining important experience. HELORS also organized six international conferences. One of the top conferences was the 12th International Conference on OR of IFORS which was held in Athens, Greece, 25–29 June 1990, and was recognized as the best in terms of quality and content of IFORS conferences. It brought together about 700 scientists from around the world. The company has also organized the European Conference EURO XX OR and the Management of Electronic Services in Rhodes from 4–7 July 2004.

Special mention should be made of the organization under the initiative and

responsibility of Macedonia–Thrace annex: four Balkan conferences in OR (1988, 1993, 1995, and 2002) have been held in Thessaloniki with the participation of researchers from the Balkan countries (Albania, Bulgaria, Serbia, Former Yugoslav Republic of Macedonia, Romania, Turkey, and so on).

Training and Seminars

So far, more than 100 lectures have been given by prestigious members of the HELORS, on current issues of OR and the methodology for making rational decisions.

From its inception until today, HELORS has held more than 700 seminars, gaining thus exceptional, extensive, and valuable experience. This is one of the most remarkable and important efforts of HELORS in education and vocational training. The training topics, since early years, fully cover all methods and techniques of operations research and many areas, directives, and questions developed in the management science.

The seminars offered by HELORS are divided into seven groups:

- marketing,
- finance,
- computers and new technologies,
- media-planning resources,
- quality management,
- management of materials-distribution,
- decision systems administration.

Approximately 13,500 executives from the private and public sectors have participated in the company's seminars. Participants, through an evaluation procedure, recognized the practical value and usefulness of the topics covered by the seminars. The majority of the trainees' implemented techniques and methods of OR in their employment companies with positive results.

The workshops of the organization are characterized by topics of current interest and high quality services. The instructors are professors of theoretical training and practical experience, with great knowledge and experience of the field.

Each seminar is structured with appropriate balance between theory and practice. This also comprises the main difference between HELORS and other similar companies. HELORS tries to teach the participants a new way of thinking, analyzing, and solving business problems, so that they instead of randomly applying methods and tools in their businesses, they will be able to choose the most appropriate methodology.

Projects

The most recent indicative studies undertaken by the HELORS are given below:

1. completion of part of the project "SPEKS—Creating value through change: anthropocentric approach bringing together social partners, enterprises, and knowledge providers" (contracting body: Labor Institute of GSEE);
2. research implementation on the Hellenic police officer satisfaction (contracting body: Institute of Police Studies, Training and Documentation);
3. implementation of two studies within the research project entitled WFP CRETE "continuing adaptation of human resources to new conditions through an integrated approach to management of change through intelligent use of business information systems" (contracting body: TUC);
4. external evaluation implementation of the Action 20 of the Activity 3 project ARTEMIS: EQUALITY IN ARMED FORCES "European Project EQUAL" (contracting body: Research Center, University of Piraeus).

CONCLUSIONS

As obvious from its activities, HELORS is an active organization and based on its members' quality, tries to achieve its objectives and increase leverage.

In upcoming years, high priority will be given to collaborations with several scientific unions, such as the Hellenic Mathematical Society, Hellenic Society of Artificial Intelligence, and Greek Computer Society, with an aim to jointly organizing conferences of similar interest.

HELORS intends to organize a student conference in OR in which undergraduate, postgraduate, and doctoral students can present their works and attend lectures from distinguished researchers of this field.

In parallel, it will establish the OR award medals for the best work (undergraduate, graduate, and doctoral); this is aimed at providing motivation for young scientists and researchers to participate in the activities of the HELORS to ensure continuity and proliferation of OR concepts, and will create new incentives for the country's scientific potential.

REFERENCES

1. Pappis CP. A survey of OR utilization in Greece. *Eur J Oper Res* 1981;6:248–251.
2. Pappis CP, Nikitas GD, Patrikalakis GS. Perspectives for practice OR in practice: results of a survey in Greece. *Eur J Oper Res* 1988;34:405–410.
3. Pappis CP, Dimopoulou M. OR practice in Greece: status and challenges. *Eur J Oper Res* 1995;87:445–451.
4. Papoulias DB, Darzentas J. OR in Greece: myth and reality. *Eur J Oper Res* 1990;49:289–294.
5. Bornstein CT, Rosenhead J. The role of operational research in less developed countries: A critical approach. *Eur J Oper Res* 1990;49:156–178.

HEURISTICS AND THEIR USE IN MILITARY MODELING

RAYMOND R. HILL
Department of Operational Sciences,
Air Force Institute of Technology,
Wright-Patterson Air Force Base,
Ohio

EDWARD A. POHL
Industrial Engineering Department,
University of Arkansas, Bell
Engineering Center, Fayetteville,
Arkansas

INTRODUCTION

The US Department of Defense (DoD) employs many models, and these models vary tremendously by type and purpose. The more common types of models used by the DoD include those listed by Gueret *et al.* [1]:

- simulation models;
- optimization or mathematical programming models;
- financial models;
- logistics and inventory models;
- statistical models.

These models span a range of applications such as

- analysis of combat operations;
- analyses of weaponry effectiveness;
- development and analysis of deployment plans;
- programming for and analysis of weapons procurement;
- personnel scheduling;
- logistics forecasting.

An optimization model develops a best allocation of resources, subject to various resource limitations. This optimization process employs some computerized algorithm

that conducts a search among potential solutions to locate some best feasible allocation. This computerized search requires a computational representation of the optimization problem of interest.

A general form of a constrained optimization problem is

$$\begin{array}{ll}\text{Minimize} & f(x) \\ \text{subject to} & x \in \Omega.\end{array}$$

The function $f(x)$ represents some objective or cost function we wish to minimize by finding values for the vector of decision variables, x . These variables x are elements of the space of feasible solutions denoted by Ω . In linear programming, Ω is defined as the intersection of the linear constraints defined for the problem. In nonlinear programming, the definition of Ω may involve nonlinear constraint functions. In integer programming, the continuity of Ω is lost due to integrality constraints placed on some or all of the decision variables. A special case of integrality constraints placed on Ω involve binary restrictions on x ; particular x variables take on values of 0 or 1. Binary programming problems are among the more difficult problems to solve and are thus often solved using heuristic search methods.

Classic optimization algorithms use objective function derivative information to search among feasible solutions (i.e., searches within the feasible region of the problem defined by the problem constraints). These approaches provide guarantees of convergence to global optimal solutions for linear models or local optimal solutions for nonlinear models.

Heuristic optimization algorithms are search procedures based on nonderivative information. These are sometimes called *local search procedures*. With increased computing power, such approaches have become increasingly popular when solving large, complex problems. Although they provide no guarantee of optimality, heuristic solutions

are generally quite good but more importantly, they can accommodate more complex problem formulations than can be accommodated by classical optimization methods.

Our current focus is on the use of heuristic search methods applied to military optimization and modeling problems. We can distinguish heuristic search techniques from classical techniques by the following characteristics:

- They do not guarantee convergence to a global optimal (in general).
- They do not employ gradient-based techniques.
- They involve some form of local neighborhood evaluation.
- They have some analogy to a heuristic problem-solving approach or an analogy to some natural process (that has some optimization aspect).

Despite our focus on heuristic search methods, we must note that there is increased interest in the computational research area in combining classical optimization methods with heuristic search methods. Such hybridized approaches exploit the search properties of heuristic approaches with the focused local search capabilities of classical methods. A thorough discussion of these approaches are a natural extension to this article, but outside its current focus.

The remainder of this article is organized as follows. In the section titled “Heuristic Search Concepts,” we provide a fundamental understanding of important concepts pertaining to heuristic search methods for optimization. These concepts provide a foundation for subsequent sections that overview various heuristic search techniques. Subsequent sections provide overviews of simulated annealing (SA), genetic algorithms (GAs), tabu search, scatter search, ant colony optimization (ACO), and particle swarm optimization (PSO). In each of these sections, we introduce the motivations for the technique, the fundamental concepts pertaining to how the technique functions and, where appropriate, discuss some of the advanced concepts associated with the technique. In the section

titled “Military Applications,” we provide an overview of how some of these heuristic search methods have been applied to various military-focused modeling applications. We end with concluding remarks pertaining to the use of heuristic search methods in military optimization applications.

HEURISTIC SEARCH CONCEPTS

According to Silver [2], “the term heuristic means a method which, on the basis of experience or judgment, seems likely to yield a reasonable solution to a problem, but which cannot be guaranteed to produce the mathematically optimal solution.” The use of heuristic methods is not a new trend. Early research examined a variety of heuristic solution methods for solving hard discrete optimization problems. For example, heuristics such as the Hungarian method, the nearest-neighbor route construction approach, the minimum spanning tree algorithm, or the cost–benefit ratio calculations for knapsack problems are often covered in a basic optimization course. These heuristics select decision variable values that produce an immediate gain based on the current information available to the algorithm. A common label of “greedy” is applied to these approaches. These approaches are greedy in the sense that current decisions are based just on currently available data versus basing those decisions on globally available information.

Unfortunately greedy heuristic approaches sometimes yield poor quality solutions. With a greedy approach, a “good” variable selection earlier in the solution construction process is usually offset with “poor” variable choice later in the process due to the inescapable limitations on variable selections; the greedy approaches suffer from a loss of freedom of choice later in the process. This is referred to as getting “trapped in regions of local optimality” and the regions where the heuristic search is trapped too often contain poor solutions.

Modern heuristics first construct a solution and then improve upon that solution by including mechanisms to escape from

the locally optimal regions. These modern heuristics, many of which were discovered when examining algorithms based on some natural phenomena, are collectively referred to as metaheuristics. According to Osman and Kelly [3], a “metaheuristic is an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search spaces using learning strategies to structure information in order to find efficiently near-optimal solutions.” The nature of how the subordinate search is guided, how the solutions are structured, and how the process learns varies by heuristic and specifics of each distinguish the particular metaheuristic. The remainder of this section provides some of the common terminology and concepts underlying metaheuristic search strategies.

Any metaheuristic seeks good *solutions*, x , to the optimization problem. These solutions provide decision variable values that provide the best available solution for some specified evaluation criteria while meeting all restrictions on the solution. All solutions have *attributes*, or characteristics, that define the solution. For instance, in routing problems, typical attributes are the number of routes, number of vehicles, the customers assigned to each route, and the order in which the customers are visited within each route. In assignment problems, attributes might be how jobs are assigned to workstations, the order of the jobs within the workstations, when jobs need to be completed, how much time jobs require for completion, and when jobs should be completed. The solution *evaluation* is a function that maps the combination of decision variable values to some numerical indicator of quality; for instance, the objective function for the problem. The solution *restrictions* are the constraints on the problem. A metaheuristic systematically searches among problem solutions, moving among those solutions via some *transition function*, $t(x)$. A transition function is defined for the problem and provides the basis for changing a current solution to a new solution, thus constituting a *move* among the solutions in the problem solution space. Metaheuristics exploit this *neighborhood* concept to control

the search. The neighborhood of a solution consists of those solutions obtained using the transition function. An improving search continues to find better solutions through the search process. As the rate of solution improvement drops, the search is said to have slowed. When search improvement stops, the search is said to have stagnated, stabilized or in some cases, to have converged.

The general approach for a heuristic search algorithm is the following:

1. Initialize any search parameters; generate an initial solution or set of solutions.
2. Generate the neighborhood of the solution, denote this as $N(x)$.
3. Evaluate members of the neighborhood.
4. Select some best element of the neighborhood.
5. Set the new solution to the best element by transitioning from the current solution to the new solution.
6. If not done searching, go to Step 2, else return current best-found solution as final solution.

The general approach described above is more specific in actual practice and varies by algorithm. Some approaches may vary how the initial solutions are generated or may periodically reinitialize the search. Some algorithms change the definition of the neighborhood especially during the execution of the algorithm. Muller-Merbach [4] describes various search neighborhoods. Evaluation of the members of a neighborhood can be a computationally expensive process, therefore some algorithms employ strategies to reduce the neighborhood evaluation effort. Examples of neighborhood strategies include the following:

- evaluate all members and choose the best;
- evaluate members until the first feasible solution is found;
- evaluate members until a first improving solution is found;

- evaluate a subset of members choosing the best in the subset;
- evaluate and use some randomly chosen member.

Some heuristic implementations employ data structures to store a set of best solutions found during the search. For instance, this set of best solutions found can be used to restart the search process, provide a range of potentially good solutions, or provide the best solution found by the heuristic search algorithm.

A defining aspect of modern heuristic search algorithms is their use of an explicit strategy to escape regions of local optimality. Classic methods, such as linear programming, gradient descent (or ascent) for nonlinear problems, even greedy heuristic search algorithms stop when there are no improving moves. In the case of linear programming, the solution returned is optimal. In the case of nonlinear problems, the solution returned is a local optimal. For greedy heuristic search algorithms, lack of an improving move stops the process and the process is deemed to be at a local optimal (albeit sometimes not a very good solution in terms of global optimality). The power of modern heuristic search algorithms is their ability to escape these regions of local optima and continue transversing the search space. Some of the more common escape mechanisms include

- acceptance of nonimproving moves as a means to transition the search out of the local optima regions;
- restarting of the search from a predefined set of solutions or from a newly generated solution;
- generation of and transition to some new random solution;
- modification of the transition function such as in variable neighborhood search strategies found in routing strategies [5].

We close this background section with definitions of other terms commonly found in the heuristic search literature. First, *diversification* within a search algorithm means moving

into new, potentially unexplored regions of the search space. Diversification is used to expand the search and ensure more thorough coverage of the solution space. Strategies employed for diversification vary by the algorithm considered. Conversely, *intensification* within a search algorithm refers to a reinforcement of the local improvement strategies to move toward locally optimal solutions. Intensification is a focusing of the search effort to hone in on some local optimal point. The more strategic search algorithms incorporate both mechanisms transitioning between periods of intensification and periods of diversification.

The concept of *learning* is often used in the discussion of a heuristic search algorithm. In general, learning entails using search history to tune the search based on the results of the search. In some algorithms, for instance SA and GAs, these learning mechanisms might involve changes to the search parameters. In other algorithms, like tabu search or scatter search, these mechanisms might involve the strategic use of data structures to store information regarding search progress and then using that data, the *memory* of the search, to modify specific aspects of the search. Finally, a *dynamic* search algorithm can change or adapt search parameters during the search to make it more productive and efficient.

SIMULATED ANNEALING

Background and Motivation

SA is one of the earliest search heuristics for optimization inspired by a natural analogy. The analogy is “to the process of physical annealing with solids, in which a material is heated and then allowed to cool very slowly until it achieves its most regular possible crystalline state, with corresponding minimum energy” [6].

As discussed in Kirkpatrick *et al.* [7], Metropolis *et al.* [8] first introduced a computational form of SA as a means to simulate “a collection of atoms at equilibrium at a given temperature.” The Metropolis approach involved the probabilistic acceptance of nonimproving steps of a process. In their case, when an atom is displaced causing a decrease in

system energy, the displacement is accepted and the simulation proceeds. If the energy is increased, then the displacement can be rejected with some probability. That probability function involves a parameter called the *temperature*. This temperature parameter decreases during the execution of the algorithm, causing changes to the probability of accepting the nonimproving moves.

Kirkpatrick *et al.* [7] tied the Metropolis algorithm to optimization and demonstrated the algorithm on various optimization problems.

The Algorithm

SA is a local search algorithm with a probabilistic element that provides the SA with an escape mechanism to depart from the local optima region. Early in the search, random moves are generally accepted. As the search progresses, nonimproving random moves are accepted less frequently. Early search processes promote diversification, later search processes promote intensification.

The SA search process, assuming we want to find a maximum value, is as follows:

1. Initialize the search process:
 - set initial solution x_0 via some selected method;
 - set temperature reduction factor, $0 < \alpha < 1$, initial temperature, f_0 , and $f_{\min}$ as minimum temperature;
 - set $nreps$ as maximum iterations per temperature and $reps = 0$ as current iteration counter;
 - set $x = x_0$ and $f = f_0$ as initial solution and initial temperature, respectively;
 - determine neighborhood of any solution, $N(x, t)$.
2. Randomly select $x^0 \in N(x, t)$. Increment $reps$.
3. Calculate $\delta = f(x) - f(x^0)$.
4. If $\delta < 0$ then set $x = x^0$, else generate $u \in U(0, 1)$ and then, if $u < \exp(-\delta/f)$, $x = x^0$.
5. If $reps \leq nreps$, go to Step 2.
6. Set $f = \alpha f$, $reps = 0$. If $f \geq f_{\min}$, go to Step 2, else Return x as final solution.

In Step 2, the transition function randomly selects a neighbor of the current solution x . The change in solution is calculated in Step 3. In Step 4, improving solutions are always accepted while nonimproving solutions are selected with probability that is a function of the current temperature, f . When the temperature is high, such as early in the search process, the probability of selection is higher giving diversifying activity to the search. As the temperature decreases, the probability of selection decreases giving intensifying activity to the search. At high temperatures the SA algorithm resembles a random search while at low temperatures the SA algorithm resembles an ascent function (again, assuming maximization is the objective). The $nreps$ parameter provides the number of iterations at each temperature level (Step 5). In Step 6, the temperature is lowered by the factor α . When α is high, the process mimics slow cooling while when α is low, it mimics rapid cooling. When the temperature reaches some predefined minimum, $f_{\min}$, the SA process stops and returns the current solution, x .

The SA algorithm requires problem-specific decisions prior to implementation and execution. These decisions are

- a formulation of the problem,
- a neighborhood definition and transition function,
- an objective function,
- a strategy for handling any problem constraints,
- some procedure for obtaining an initial solution.

The SA can actually be thought of as a set of algorithms, each distinguished by the parameter values assigned. These more generic algorithm-specifying decisions include setting the values for α , f_0 , $f_{\min}$, and $nreps$. Unfortunately, few guidelines exist for setting these parameters; generally, the parameter values are empirically tuned on sample problems and then set to the derived best values for solving actual problems.

Algorithm Enhancements

The fundamental trade-off in an SA is the cooling rate, α , and the number of iterations

at each temperature, $nreps$. These together form the *cooling schedule* for the algorithm. Theory suggests using large values of α and large values of $nreps$ but this combination however, can lead to computationally expensive procedures. Practical implementations choose smaller values, in some cases even setting $nreps = 1$, offset by large α values.

Practical SA implementations use data structures to save the best solutions encountered during the search. Upon algorithm completion, the best solution in this list is returned. One SA enhancement is called *reheating*. This involves the resetting of f to a higher temperature to promote additional diversification in the search process.

Additional enhancements are more problem-specific. For instance, transition functions and neighborhoods can be adjusted as a function of temperature to oscillate between periods of diversification or intensification. Restricted neighborhoods can be used to accommodate problem constraints or to improve processing time. Data structures can save the search history and avoid revisiting solutions. Finally, the SA can be hybridized with other search methods such as steepest descent (or ascent) to further improve solutions prior to returning the final solution from the algorithm.

Final Comments

SA is one of the earliest metaheuristics. As such, it has been used in many applications and has achieved success in complex optimization settings. While often viewed as a random search, SA is really a neighborhood search that interweaves varying degrees of random escape mechanisms with local improvement to achieve reasonably good solutions to hard (discrete) optimization problems.

GENETIC ALGORITHMS

Background and Motivation

GAs are search procedures inspired by biology and the workings of natural selection. The GA name “originates from the analogy between the representation of a complex

structure by means of a vector of components, and the idea, familiar to biologists, of the genetic structure of a chromosome” [9]. In biology, natural selection reinforces characteristics most amenable to a species’ survival. Genes within the chromosomes of the stronger members, corresponding to the more desirable characteristics, pass to subsequent generations through reproduction.

In optimization problems, solutions are considered members of a population. Each member of this population is assigned a value indicative of its strength or quality; for instance, the objective function value or penalized objective function value. A population reproduction process is then used to match current solutions, biased by each solution’s strength, to produce new solutions, or offspring, by pairing up current solutions in the population. Better solutions, or stronger population members, are paired more often and thus provide greater influence on future solutions, or population members, for the next generation. Problem solutions are encoded so that problem characteristics can be passed onto subsequent population members. Starting with a population of solutions and allowing a “survival of the fittest” evolution process to occur over some number of generations produces a final population of strong solutions that ideally contain the optimal solution from within that population. The best such solution is returned as the problem solution.

The Algorithm

A GA requires an encoding of problem solutions. A typical encoding scheme is to represent the solution as a binary stream. This binary stream is referred to as the *chromosome* while sections representing logical components of the solution are called *genes*; for instance, each decision variable is represented by some binary substream (the gene) which when collectively concatenated yields the chromosome. The actual representation of the problem solution is referred to as the *phenotype* while the encoded representation is referred to as the *genotype*. The mapping from phenotype to genotype is problem-specific and user-defined.

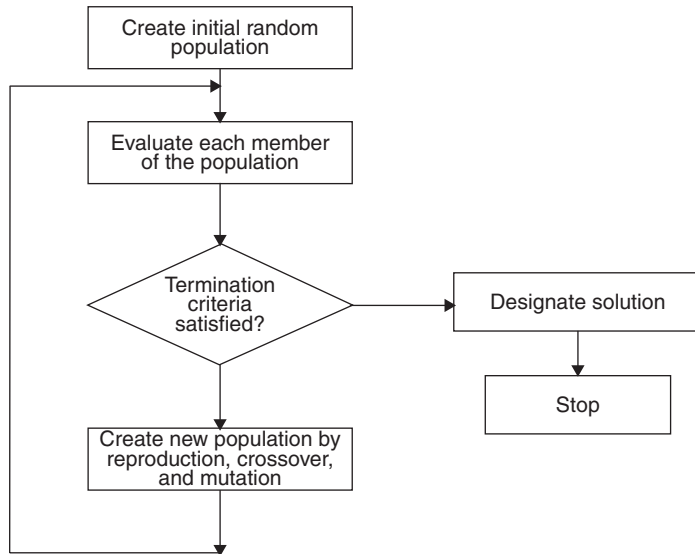


Figure 1. Genetic algorithm flowchart.

Each solution has a strength value or *fitness function*. This fitness function, evaluated in the phenotype and associated with the genotype, is exploited during the evolutionary processes within the GA. This fitness function must account for objective function values and solution feasibility; quite often, penalty functions are used to accommodate problem constraints.

The GA is a population-based heuristic. The GA starts with some initial population and evolves that population through subsequent *generations*. Evolutionary processes produce offspring that potentially replace (weaker) individuals in the current population. This is called the GA *birth and death process*. The resulting population dynamics change the makeup of the population with each generation; ideally, each subsequent generation should be stronger than the previous generation. In optimization terms, each new generation should have a higher average value associated with the population members and the “best” member of the population should never be worse than the best member of the previous generations. These are key attributes for a GA. When the average solution value of the generations does not change or the best solution is not replaced for some period, the GA is said to have *converged*. Intuitively, it means

the population is unable to evolve newer, stronger members, and thus the population is unable to evolve a better solution than that which has already been found.

Figure 1 [10] is a general flowchart of a GA.

Three main genetic operators are used in population dynamics: reproduction, crossover, and mutation. Varying probabilities of applying these operators control the speed of convergence of the GA. The crossover and mutation operators must be carefully designed.

Reproduction is a process in which a new population is created by combining individual strings from the present population, according to their fitness function values [11]. Copying strings according to their fitness values means that strings with higher fitness values have a higher probability of contributing one or more offspring for the next generation. This improves the possibility of survival for very strong solutions, an element of immortality to the population member. There are a variety of ways to select population members for reproduction:

- *Random*. Pick two elements of the population randomly and allow these two to serve as parents.

- *Rank*. Rank order population members and pair parents by their ranking.
- *Elitist*. Only choose among the strongest population members to serve as parents.
- *Tournament*. Pick a random subset from the population, put the best of the subset into the pool of parents allowing multiple copies of any population member.
- *Proportional*. Normalize population member fitness functions and randomly pick parents using the normalized fitness as their probability of selection;

allow multiple copies of any population member.

New individuals are created as offspring of two parents via the *crossover* operation. One or more crossover points are selected at random within the chromosome of each parent. The parts that are delimited by the crossover points are then interchanged between the parents. The individuals formed as a result of such an interchange are the offspring [12]. Common types of crossover are depicted in Fig. 2 [10,12]. Burst crossover is another type of crossover where the

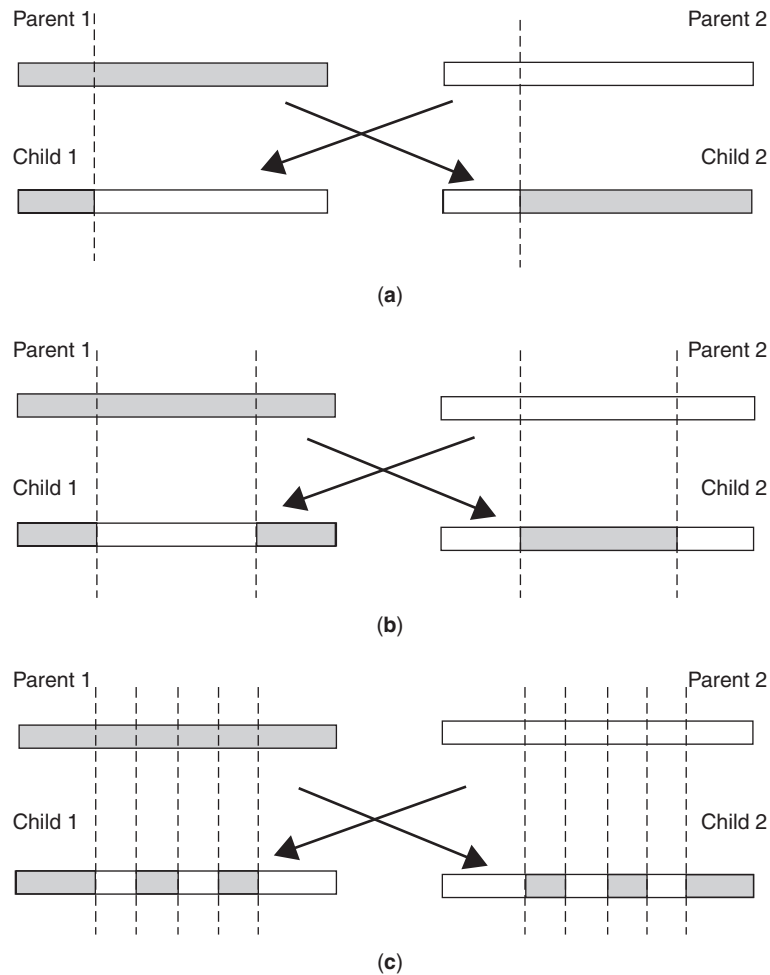


Figure 2. Different types of crossover operators: (a) one-point crossover, (b) two-point crossover, and (c) multipoint crossover.

crossover is made at every bit position [13].

The final reproductive operator is the *mutation* operator. Mutations are random changes made to the offspring chromosome. For instance, for a binary encoded chromosome, each bit in the string is examined and with some small probability, its value is complemented. Mutation provides an element of diversification in the GA plus serves the important task of restoring potentially good solution characteristics that might otherwise get lost during reproduction.

There are a variety of choices in terms of how generations evolve. The production of offspring expands the size of the population, or the current generation. A GA normally retains a constant population size, thus some of this expanded population must be removed, or killed off, to restore the population to its proper size. Once restored, the resulting population represents the next generation of the GA. Three replacement strategies are

- generational, where the offspring replace parents yielding a new population;
- incremental, where proportionally select members for the next generation from the current, expanded population;
- best, where only the strongest members of the expanded population is kept.

The GA continues for some defined number of generations, or until convergence of the population is recognized. The best solution in the population is returned as the estimated “optimal” solution to the problem.

Algorithm Enhancements

There have been a variety of enhancements to the basic GA described above. Only some of those enhancements are discussed here. More complete information can be found in the various texts available on GAs. The interested reader is also directed to Illinois Genetic Algorithms Laboratory, <http://www.illgal.uiuc.edu/web/>, for some of the latest developments in GAs.

GA improvements involve relaxing the underlying natural analogy. The first such

improvement ensures the best member, or members, of a population survive into the next generation. Increasing the mutation rate helps diversify a stagnated search. Reinitializing the population provides a means to restart the research. Expanding the population size for some period of time is another way to increase diversity. Changes to population dynamics operators include using more intricate crossover patterns, fixing certain sections of chromosomes to not allow their destruction during crossover, and ensuring that offspring are feasible solutions.

The reproduction process is sometimes improved. For instance, one adjustment to a GA is to ensure all population members are unique. A *Lamarchian* process takes a solution from the population, applies a local search heuristic to that solution, and if the solution is improved, replaces that solution in the population with the improved solution. This can be applied to multiple members of the population.

Finally, there are also ways to improve the initial population. Elements of a randomly selected population can be improved via local improvement heuristics. A randomly selected population can be improved by intelligently modifying the probabilities associated with the solution selection process [14]. The initial population can also be seeded with strong solutions found via other optimization methods or via specific domain knowledge about the problem. Typically, the stronger the initial population of a GA, the quicker that GA can converge and return its recommended solution.

Final Comments

The GA is a popular metaheuristic solution approach. Because the GA works on an encoded version of the optimization problem, the algorithm is applicable to a wide variety of problems; once encoded with a suitable fitness function, the GA is quite problem-independent. As a result, the GA is arguably the most general-purpose metaheuristic and also arguably one of the two most popular metaheuristics in terms of use and implementation.

The GA does have distracting elements. Some, such as Ross [15] view the GA as an

inferior approach. Since the GA works on a population of solutions, each of which require evaluation, the computational overhead can be troublesome. In some applications, non-binary encoding schemes are required and crossover operations require a good deal of problem-specific knowledge. In such cases, the GA label may be inappropriate. The GA can also experience convergence problems when later generations become “too similar” to effectively produce strong, new solutions. As with most metaheuristics, the GA employs algorithm parameters that may need tuning to work effectively on the problem at hand. This requires additional investigation before one can actually solve the problem of interest. Finally, understanding the schema theory, Reeves [16] proposed that GA performance is most likely not suitable for casual reading.

Additional details can be found in the work of Beasley *et al.* [17,18], the article by Reeves [19], and the overview chapter by Reeves [9].

TABU SEARCH

Background and Motivation

Tabu search’s analogy is to mimic the human problem-solving process. These approaches involve the use of memory to help avoid repeating self-defeating sequences and the search approach when the current search strategy stagnates.

The Algorithm

Fundamentally, tabu search is a local search procedure but it is augmented by additional rules to provide additional search capabilities. The fundamental approach within local search is to examine the neighborhood of the current solution (which includes evaluation of those neighbors), select some neighboring solution, transition to that neighboring solution, and repeat the process until some termination criteria is met. This simplistic approach however has issues associated with cycling among solutions and getting stuck in regions of the solution space.

The tabu search framework introduced the concept of the *tabu list*, a data structure of solution attributes deemed “off limits” for some number of search moves (called their

tabu tenure). Ignoring tabu attributes causes good features to remain in solutions, or out of solutions, and provides a mechanism to allow the search to move out of regions of local optimality. Use of multiple tabu lists is sometimes used to further enrich the search process.

The *aspiration criteria* allows the search process to override tabu restrictions and visit strong solutions examined within the current neighborhood that might otherwise be ignored. As with tabu lists, multiple aspiration criteria can be employed during a tabu search process. This criteria ensures that good solutions are not ignored during the search.

In a 1977 paper, Glover [20], inspired by Senju and Toyoda [21], introduced the concept of *strategic oscillation*. Strategic oscillation allows a search process to systematically cross barriers other search approaches strictly apply. The most common barrier for Integer Programming (IP) problems is feasibility, either feasibility due to the constraints or feasibility due to the integrality restrictions. A systematic path through the infeasible region coupled with a search path back into the feasible region can compel the tabu search to reach new, potentially favorable regions of the search space.

The basic tabu search framework can be summarized as follows:

1. Initialize.
 - Determine the search neighborhood(s) and solution-evaluation criteria.
 - Define tabu list(s) structure(s) and processes.
 - Define aspiration criteria and memory structures to employ.
 - Create initial solution to problem.
2. Local search. Examine and evaluate neighboring solutions. Select best non-tabu neighbor or best tabu neighbor that meets aspiration criteria.
3. Implement move to the selected neighbor. Update tabu list structures (and other memory structures used). Return to Step 2 if stopping conditions are not met.
4. Return best solution found.

The basic tabu structure outlined above can be expanded to include predefined strategic oscillation phases. This structure can also involve dynamic aspects. Battiti and Techionli [22] dynamically change the tabu tenure employed, based on recorded search progress. In their reactive tabu search, if memory structures indicate the search progress has slowed or is cycling among sequences of solutions, the tabu tenure is (temporarily) adjusted to diversify the search.

Algorithm Enhancements

The basic tabu search outlined above provides a flexible, robust, and generally effective search process. However, on more difficult problems, additional structure can be exploited to try and improve the solution returned from the basic tabu search. Within the tabu search framework, these additional structures are referred to as *intermediate* and *long-term memory approaches*.

Intermediate-Term Memory Approaches. The key component of intermediate-term memory is the *elite list*. This list contains good solutions encountered during the search process. For example, the list may contain good neighbors evaluated but not selected for a move during the basic tabu search process. Kinney *et al.* [23] build an elite list via a variety of solution construction methods for routing problems.

When invoked during the tabu search process, intermediate-term memory search efforts will extract a solution from the elite list and restart the tabu search from that solution. This restart process may or may not alter the current status of any memory structures to tabu lists; this is a strategic choice made by the tabu search designer. The extraction and restart process continues using elements from the elite list for a predetermined period of time or until the elite list is exhausted.

Placing solutions on the elite list is tabu search specific; the search design process determines the membership criteria. Examples of criteria include: solution is some percentage of overall best solution, some percentage from best solution in the current neighborhood, or is among some set of top

solutions in the neighborhood. Use of the elite list diversifies the search into new regions and intensifies the search by using good solutions from the elite list to restart the search process.

Long-Term Memory Approaches. The long-term memory aspects of tabu search are intended to radically diversify the search and involves processes that can be fairly complex. The long-term memory aspects of the tabu search framework are unique in the realm of heuristic search methods.

The *vocabulary building* process involves the structured use of memory to build new solutions from which the tabu search process can restart. During the search process, as moves are implemented, a catalog of solution attributes can be maintained. This catalog is then accessed at some point during the search process to ascertain attributes often used and attributes infrequently used in solutions. Such attributes are then combined to create new solutions. For instance, using infrequently found attributes can yield a new solution. The solution might be of poor quality but may be so diverse a solution that it places the tabu search process into a new region of the search space, a region that may have really good solutions within it. Combining frequently used attributes may yield a very strong solution. Sometimes the solution may be infeasible but may yield a search trajectory back to the feasible region and into a new region of the search space. Building such tabu search implementations requires consideration of attribute data structures, updates to these data structures, and mechanisms to use the data structures.

Path relinking involves the use of good solutions and a constructed search along a path connecting the two solutions. This involves a problem-specific routine that transforms one target solution to another target solution via a sequence of attribute changes. One rationale for the path relinking approach is that better solutions may lie between good solutions and local search mechanisms may be unable to follow the path between the good solutions using just the defined neighborhood search functions; the path must be forced by the search process.

Controlled randomization can come into play in the tabu search framework and this involves modifying the search neighborhood such that some set of good solutions, from the current neighborhood, are selected and then one of the elements from this set are picked probabilistically. For instance, the approach may be to take the top five neighbors of a current solution. The probability of selection among any of the five neighbors is then a function of their solution quality (i.e., objective function value). The stronger neighbors thus have a better chance of selection. The random component adds an extra element of diversity to the search process.

Each of the above long-term memory strategies are evoked based on triggering rules embedded within the tabu search process. Such rules generally emanate from a *target analysis* effort performed on the problem type. Target analysis is a key learning process of tabu search. It involves a systematic off-line study of solution search performance, a conjecture regarding new search rules, and experimentation to support or reject the conjecture regarding the candidate rule. Effective search rules are subsequently embedded into the tabu search framework for use on general instances of the problem.

Final Comments

Tabu search features three key themes [24]:

1. flexible memory structures;
2. a mechanism to control the freeing and constraining of the search process;
3. memory structures of differing time spans to promote intensification and diversification.

The resulting framework of tabu search provides a flexible, powerful mechanism for solving IP problems of varying complexity. The flexibility of the framework has yielded a variety of innovative optimization search processes. Each of these search processes feature the interplay between intensification of the search in promising areas of the search space and then diversification of the search into new regions of the search space when the search process is deemed to have slowed or even to have stalled.

Details of the tabu search framework are plentiful. However, the interested reader can find the key insights in Glover's work [24–26], the text by Glover and Laguna [27], and the chapter by Reeves [9].

SCATTER SEARCH

Background and Motivation

Scatter search was first proposed by Glover in a 1977 paper [20] and hit its stride in 1990 [28]. The motivation for the concept came from formulations that obtained solutions by combining decision rules and constraints (see work by Laguna and Marti [28] or Glover *et al.* [29] for details).

Another way to consider scatter search is to assume that given some set of good solutions to a problem, potentially better solutions may exist in the subspaces spanned by subsets of these good solutions. In a simple example, given two good solutions, x^1 and x^2 , consider potential solutions on the line containing both x^1 and x^2 . Scatter search is sometimes related to the concepts found in simplex search.

The Algorithm

The basic scatter search template has five methods [28].

The *diversification method* is defined as a way to generate a fairly large set of trial solutions. Typically, some seed solution is generated and the diversification method is used to create the trial pool of solutions with sufficient diversity to adequately cover the solution space.

An *improvement method* is used to transform some trial solution into an enhanced (i.e., better) solution. This method varies by problem type.

A *reference set update method* maintains some “best” set of solutions of high quality and sufficient diversity. This best set is called the *reference set*.

A *subset generation method* extracts members of the reference set to create new combined solutions for the problem of interest.

A *solution combination method* uses the results from the reference set to create the combined solutions.

These template methods are used to create the specific scatter search implementation. The generic scatter search process is defined as the following four step process:

- Step 1. Generate starting set of solutions; improve solutions; extract reference set.
- Step 2. Create new pool of solutions via solutions combination method.
- Step 3. Attempt to improve the solutions in the new pool using the improvement method.
- Step 4. Update reference set with best solutions from the improved pool. Return to Step 2 until stopping condition is met.

Upon completion, the final reference contains the best solution found by the algorithm. Additionally, a dedicated memory structure can be employed to specifically retain the best or set of best solutions encountered during the search process.

Algorithm Enhancements

While the basic scatter search is quite flexible, additional features have been added to improve the power of the algorithm, particularly when used on difficult problems.

The basic scatter search employs a static method for reference set updating. This static approach examines all combinations from the reference set to create the trial pool from which the reference set is updated. A dynamic update method puts quality trial solutions into the reference set and makes those new reference set members available for the combination method. This quickly removes poor solutions from the reference set and gets the better solutions involved in the search process sooner.

When scatter search progress slows, the reference set can be rebuilt. One way to rebuild the reference set is to create a new reference set with maximally diverse solutions [30].

The reference set can be tiered to avoid search stagnation. While the reference set is generally updated based on solution quality, a two-tier reference set would add updates

based on solution diversity as well. Thus, the reference set will use both, high quality solutions and diverse solutions in the combination generation method. This helps keep the search aggressive (with the high quality solutions) and productive (with the diverse solutions).

Combination methods typically consider reference set solutions in pairs. Advanced combination methods employ three or more reference set members to create a new combination. This use of multiple members, when coupled with improved combination methods, provides tremendous diversification capability for the scatter search process.

Finally, memory structures can enhance the search. Hashing functions can be used to store all solutions encountered during the search, thereby avoiding the reexamination of solutions. Attributes of solutions can be tracked using data structures and the information exploited to generate diverse solutions or to intensify the search. Such strategic uses of memory are influenced by the success of these approaches in tabu search.

Final Comments

Scatter search is a relatively new meta-heuristic search technique. However, scatter search is easily hybridized with other methods, making the approach quite flexible. The scatter search technique is also quite apt for simulation optimization applications and is found in the OptQuest optimization package [28].

ANT COLONY OPTIMIZATION

Background and Motivation

Ant colony algorithms are inspired by the foraging behavior of a colony of ants. Despite their lack of communication skills (as we know them), foraging ants demonstrate the ability to collectively find a shortest path route from their nest to a food source. They even demonstrate an ability to adapt their routes when those routes get disrupted.

Ants employ an indirect form of communication wherein ants deposit pheromones along their selected trail [31]. Shorter trails

are transited quicker, and thus more often and so are reinforced with more pheromone deposits influencing other ants to follow those paths. This *stigmergistic* behavior is mimicked in optimization algorithms where the path length is used to reinforce segments selected for a path. The formulation challenge is to characterize the optimization problem as segments over which the “ants” travel.

The Algorithm

The ACO algorithm computationally mimics the environmental reinforcement behavior observed in real ants. The ACO algorithm requires the following elements:

- an optimization problem mapped into a set of nodes, connected via segments that an ant travels over to construct a solution to the problem;
- a means to evaluate the quality of the solution obtained by the ant;
- a data structure used to track the artificial pheromone deposited (and remaining) on each of the problem segments;
- a means for updating the pheromone data structure based on solutions obtained by the ants;
- a transition policy employed by the ants to choose among segments balancing random segment selection, segment selection based on quality of the segment, and segment selection based on pheromone levels.

The ACO is a population-based heuristic; a set of artificial ants are allowed to individually arrive at solutions to the problem. The artificial ants are given memory (to avoid duplicate segment selection) and a look-ahead capability (to select among available segments during solution construction).

Initially, ants move randomly through the network of segments. From a given segment, the ant views the remaining segments, picks one at random, and adds that segment to its partial solution. Once all ants have constructed their solutions, the set contains good and poor solutions.

The next step is for each ant to deposit pheromone on the segments used in their

solutions but deposit an amount related to the quality of that ant’s attained solution. One such pheromone deposit function is

$$\Delta\tau_{ij}^k = \begin{cases} \frac{Q}{L_k} & \text{if edge (i,j) used,} \\ 0 & \text{otherwise,} \end{cases}$$

where Q is some user-defined value, L_k is the solution constructed by ant k , and τ_{ij}^k is the pheromone deposited by ant k on segment (i, j) . This pheromone change is coupled with a pheromone decay function so that a typical pheromone update function resembles the following expression:

$$\tau_{ij} = (1 - \rho)\tau_{ij} + \Delta\tau_{ij},$$

$$\Delta\tau_{ij} = \sum_{k=1}^m \tau_{ij}^k.$$

The decay factor, ρ , is used to ensure the segment pheromone level, τ_{ij} , is not reinforced too much and that it loses reinforcement over time. The value $\Delta\tau_{ij}$ is the sum of pheromone deposited by the set of m ants.

Each ACO uses a function to guide the ant’s transition behavior. This transition function provides probability values associated with any available segments biased by both, the heuristic “desirability” of the segment and the segment’s current pheromone level. For instance, Velayudhan *et al.* [32] use

$$p_{ij}^k = \begin{cases} \tau_{ij}^\alpha v_{ij}^\beta / \sum_{\ell \in S} [\tau_{i\ell}^\alpha v_{i\ell}^\beta], & j \in S, \\ 0, & j \notin S, \end{cases}$$

in the context of a traveling salesman problem application. The probability ant k , currently at node i , selects segment (i, j) among their set of possible segments, $j \in S$, is p_{ij}^k . The τ_{ij} is the pheromone level on segment (i, j) , the v_{ij} is a desirability measure of the segment (such as the distance, weight or cost), and the parameters α and β are user-defined and control the influence of pheromone and desirability, respectively, on the transition tendencies of the ant.

The set of ants iterate through some predetermined number of tours, after which the best solution found is returned.

Algorithm Enhancements

The ACO has been applied to many classes of optimization problems. Velayudhan *et al.* [32] have compiled a list of these applications.

A common enhancement to the ACO is to change the pheromone update strategy. For instance, rather than allowing all ants to update, an elitist strategy allows only those ants with good solutions to update the pheromone trail. Another strategy is to always let the best overall solution update its used segments. Both strategies keep selection pressure on those segments associated with the better solutions. Another change is to update pheromones during their tour construction instead of at the completion of all tours.

Influences on the transition function can also be used. Changing the transition function form, that is changing the v_{ij} during algorithm execution, further evaporating τ_{ij} , evaporating the τ_{ij} for the segments (i,j) not used, and modifying the α and β parameters are methods that have been explored. When used, computational study is required to fine-tune the changes to the problem parameters investigated.

The ACO can quickly converge to solutions. The pressure caused by the pheromone update scheme can cause all ants to effectively construct the same solution. Too often, this convergence occurs too early in the search process which can lead to solutions of inferior quality. A common enhancement to promote diversification in the search is to restart the search process, add periods of random solution construction, or drastically change the α and β parameters.

Final Comments

The ACO is a reinforcement learning system. The ACO is sometimes implemented as an agent-based approach and called an *agent-based optimization method*. Despite being relatively new, the ACO algorithm has quickly established credibility in solving hard combinatorial optimization problems. Among its detracting features are its sometimes early convergence characteristics, the computational overhead of a population-based approach, the need for tuning the

parameters of the algorithm, and its sometimes tenuous application to general IP or continuous-variable optimization problems.

For further reading on the ACO, see Dorigo *et al.* [31] and the complete section on the ACO in Corne *et al.* [33].

PARTICLE SWARM OPTIMIZATION

Background and Motivation

PSO is a specific algorithm that is a subset of the class of swarm intelligence algorithms that are based on the collective social behavior of species that compete for food. Ants, bees, fish, termites, and birds are examples of species that exhibit cooperative social behavior by indirect communication, which enables the members of the group to select appropriate movements within a defined search space to look for food [34]. PSO is a stochastic population-based metaheuristic which mimics the behavior of a school of fish or flock of birds in their search for food. The “school or flock” work together by coordinating their movements without any central control. This algorithm is relatively simple to implement and has been shown to be effective on a variety of problem structures. The algorithm is biologically inspired with strong links to evolutionary computation. “This algorithm belongs ideologically to that philosophical school that allows wisdom to emerge rather than trying to impose it, that emulates nature rather than trying to control it, and that seeks to make things simpler rather than more complex” [35].

The Algorithm

In the basic model, we assume we have a flock of birds or school of fish which we represent as N particles. Each particle (bird or fish) is composed of three D -dimensional vectors; the current position $\vec{x}_i$, the previous best position $\vec{p}_i$, and the velocity $\vec{v}_i$, where D is the dimensionality of the search space. Each particle i is a candidate solution to the problem being studied and therefore the objective function is calculated for each particle at its current location. Each particle then determines its movement through

the search space by successively adjusting its position toward the global optimum by combining aspects of its own history and current position with those of other members of the swarm. The next iteration takes place after all particles move to their next position. The selection of the next position is based on two factors: the best position visited by the individual particle $p_i = (p_{i1}, p_{i2}, \dots, p_{iD})$ which we designate $pbest_i$, and the best position visited by the entire swarm, $gbest$ denoted as $p_g = (p_{g1}, p_{g2}, \dots, p_{gD})$. These particles serve as leaders, and assist the search algorithm in finding the better regions of the decision space. New points are selected by adding $\vec{v}_i$, coordinates to the current position, $\vec{x}_i$, where $\vec{v}_i$ serves as the equivalent of a step size for the next move. The velocity, which defines the amount of change that is applied to the current particle position, is updated using the following relation:

$$v_i(t) = v_i(t-1) + \rho_1 C_1 \times (p_i - x_i(t-1)) \\ + \rho_2 C_2 \times (p_g - x_i(t-1)),$$

where ρ_1 and ρ_2 are two uniform random numbers. The constants C_1 and C_2 represent

the cognitive and social learning factors, respectively. The cognitive learning factor represents the attraction a particle has toward its own success, while the social learning factor represents the attraction the particle has toward the success of its neighbors. The elements of $\vec{v}_i$ are ordinarily limited to some defined range, $[-V_{\max}, V_{\max}]$, which has the effect of influencing the balance between exploration and exploitation of the search space. Each particle will then update its coordinate position vector as follows:

$$x_i(t) = x_i(t-1) + v_i(t).$$

Each particle then updates the best local and global solutions, if appropriate, as follows:

$$\text{if } f(x_i) < pbest_i, \text{ then } p_i = x_i. \\ \text{if } f(x_i) < gbest, \text{ then } g_i = x_i.$$

The pseudocode for the PSO algorithm is given below:

Random initialization of the entire swarm;

Repeat

Evaluate $f(x_i)$

For all particles, i

Update velocities:

$$v_i(t) = v_i(t-1) + \rho_1 C_1 \times (p_i - x_i(t-1)) + \rho_2 C_2 \times (p_g - x_i(t-1))$$

Move to the new position: $x_i(t) = x_i(t-1) + v_i(t)$

IF $f(x_i) < pbest_i$, **then** $p_i = x_i$;

IF $f(x_i) < gbest$, **then** $g_i = x_i$;

Update (x_i, v_i) .

EndFor

Until Stopping Criteria

The final issue to resolve is the selection of values for the various parameters in the PSO algorithm. First, one has to decide how many particles (or members of the flock) to have in

the algorithm. The value of N , the number of particles, is inversely related to the number of required iterations for the algorithm. Like any population-based heuristic, increasing

the population size tends to improve results at the expense of more computation time. Talbi [34] recommends using a value between 20 and 60 for N . The acceleration coefficients are generally set at values ≤ 2 .

Algorithm Enhancements

Inertia Weight. An inertia weighting factor, w can be added to the velocity update equation in order to control the impact of the previous velocity on the current one as shown below:

$$v_i(t) = w \times v_i(t-1) + \rho_1 C_1 \times (p_i - x_i(t-1)) \\ + \rho_2 C_2 \times (p_g - x_i(t-1)).$$

The motivation for the inertia weight is to gain better control of the search and reduce the impact the selection of $V_{\max}$ has on the convergence of the algorithm. A large inertia weight encourages global exploration while a smaller inertia weight encourages local exploitation. Researchers have found that that the best performance is obtained by initially setting w to a value of 0.9 and gradually reducing it to a much lower value of about 0.4 as you move toward homing in on the local optimal value.

Constriction Coefficients. The use of a constriction coefficient helps dampen the dynamics of the particles beyond what was capable with the use of $V_{\max}$. Clerc and Kennedy [36] describe many ways to implement constriction coefficients. The simplest method is given as follows:

$$v_i(t) = \chi (v_i(t-1) + \rho_1 C_1 \times (p_i - x_i(t-1)) \\ + \rho_2 C_2 \times (p_g - x_i(t-1)),)$$

where $C = C_1 + C_2 > 4$ and

$$\chi = \frac{2}{C - 2 + \sqrt{C^2 - 4C}}.$$

Using this method, the particles have been found to converge without using any $V_{\max}$. Another approach by Shi and Eberhard [37] recommended using the constriction coefficient approach and limit $V_{\max}$ to $V_{\max}$, which is the dynamic range of each variable

in each dimension. This results in a PSO implementation that has no problem-specific parameters.

Fully Informed Particle Swarm. In the implementations discussed above, the PSO algorithm is constantly making the trade between the amounts of influence the local position has, versus the amount of influence the best particle has on the next move. Kennedy and Mendes [38] developed the fully informed particle swarm (FIPS) approach that allows the velocity update to be influenced by all its neighbors and possibly have negligible influence from its previous success. The velocity update formula for the FIPS implementation is as follows:

$$v_i(t) = \chi \left(v_i(t-1) + \frac{1}{K_i} \sum_{n=1}^{K_i} \rho_1 C \right. \\ \left. \times (p_{nbr_n} - x_i(t-1)) \right),$$

where K_i is the number of neighbors for particle i , and p_{nbr_n} is the best position of i 's n th neighbor. For $K_i = 2$, the velocity is updated using only the particle's current position and neighborhood best.

Final Comments

PSO has been used in a diverse range of application settings. Poli [39] provides a detailed review article that categorizes the large number of recent publications that deal with PSO applications in the IEEE Xplore database. The main PSO application areas include image analysis (7.6%), design and restructuring of electrical loads and networks (7.1%), general control applications (7.0%), antenna design (5.8%), power generation and power systems (5.8%), and scheduling (5.6%) [40]. Approximately 1.3% of the more than 1100 papers published using PSO were in the security and military applications domain. These papers included network security, intrusion detection, cryptography, border security, weapon target assignment, and missile effectiveness optimization. Wang *et al.* [41] have used a modified PSO to

perform unmanned combat aerial vehicle (UCAV) path planning. Foo *et al.* [42] utilized a PSO algorithm in their three-dimensional path planning for an unmanned aerial vehicle (UAV). Cui and Potok [43] have used a PSO social model in their multiagent insurgency warfare simulation.

MILITARY APPLICATIONS

The range of military model applications is quite vast. Battilega and Grange [44] required an entire text to review the diverse range of models then available. However, their review did not even include heuristics as a major technique for military optimization applications [44]. This section thus focuses on heuristic methods applied to military optimization applications. Further, the focus is narrowed to the use of the metaheuristic search techniques previously reviewed.

Assignment Problems

The classic assignment problem uniquely pairs objects from one set with objects from another set, while seeking that assignment that either minimizes some cost function or maximizes some profit function. Extensions add additional conditions on the assignment problem realized as constraints within the model. Military applications of the assignment problem include

- maintenance tasks to maintainers;
- aircrews to aircraft;
- personnel to organizations;
- radio frequencies to user channels;
- weapons to targets;
- supplies to delivery modes.

The weapons assignment model application is crucial to military analyses, from developing requisitions to purchasing the weapons to determining which mode of delivery will be used to employ the weapons. Wacholder [45] used a neural network to solve a nonlinear weapon assignment problem for ballistic missile defense. Green *et al.* [46] use a rounding heuristic to find solutions to the arsenal exchange model (AEM), a

model used to assign nuclear weapons to targets. Their rounding heuristic was shown to outperform the existing AEM assignment heuristic. Cullenbine *et al.* [47] apply reactive tabu search to solve the weapon assignment model (WAM). A WAM solution also involves the assignment of nuclear warheads to targets but those solutions are evaluated with respect to multiple goals the plan should achieve. Their solution approach included tabu search mechanisms for search diversification and was found to yield reasonable solutions faster than the currently existing heuristics. Lee and Su [48] use an ant colony algorithm but augment their algorithm with immunity concepts inspired by Jiao and Wang [49]. Immunity concepts include vaccination and immunity processes. A vaccination process uses problem-specific knowledge to modify specific aspects of the solution to avoid degradations in solution quality. An immunity process ensures weaker solutions are not used to reinforce subsequent solutions. Lee and Su [48] compare their approach to SA and GA approaches on four randomly generated problems.

Blodgett *et al.* [50] provide an interesting application of tabu search to the assignment of naval defense systems against anti-ship missiles. Rather than a simple assignment, which may require replanning once operations commence, the authors develop a decision tree-based solution to account for the probabilistic outcomes that arise in actual operations. The intent is to derive solutions more robust to the uncertainty of combat. Erdem and Ozdemirel [51] apply a GA to a force-versus-force allocation problem. Their problem allocates blue (friendly) force components against red (enemy) force components to achieve defined attrition goals at minimal overall effort. The force components on each side are modeled as heterogeneous and divisible components. The GA employs problem-specific chromosome repair operators to ensure feasibility of population members and it employs a local improvement phase to further improve population members (such as found in memetic algorithm approaches).

Military organizations, particularly those deployed in combat zones, require effective communications for operational success. The

frequency assignment or channel assignment problem assigns frequencies to wireless connections to ensure communication can occur and interference from other frequencies is minimized. Aardal *et al.* [52] provide a thorough survey of solution techniques for these frequency assignment problems. Their review includes classic optimization and heuristic search methods. Heuristic methods found in the literature, and covered in their survey, include greedy algorithms, local search methods, tabu search, SA, GAs, and ant colony algorithms. Lau and Tsang [53] was not included in the survey but present a guided genetic algorithm (GGA). The GGA is devised to retain specific knowledge of solution characteristics of the population members. This knowledge structure is used to modify the reproduction process (crossover and mutation) to reinforce or limit strong or weak solution features, respectively. The results indicate that the GGA provides good solutions, but at higher computational costs as compared to local search procedures.

Routing Problems

A routing problem assigns a sequence of destinations to a vehicle so that some cost function is minimized as the vehicle transits the segments connecting the destinations in the vehicle's route. Military applications of routing include

- routing a UAV or set of UAVs over some set of targets;
- routing a surveillance asset over some search region;
- determining the sequence of stops for a mobility aircraft mission;
- determining how to route a missile or aircraft to a target to avoid threats and obstacles;
- determining how to deliver required supplies from supply locations to destination locations.

The routing problem associated with UAVs has received a great deal of attention. Additional constraints added to these problems include considerations of no-fly

zones, arrival time restrictions, time-over-target restrictions, flight duration considerations, and threats within the route space. Ryan *et al.* [54] examine the UAV routing problem with time windows and time precedence considerations. The authors couple a reactive tabu search route finder with a simulation route evaluator to determine "robust routes." The simulation is used to assess target coverage and travel time considerations using stochastic processes for the threats and the weather. Multiple iterations yield route segment information used to determine those better segments to use within the route. O'Rourke *et al.* [55] build upon the optimization portion of this work, yielding a reactive tabu search that solves a more comprehensive routing problem. Harder *et al.* [56] develop a general-purpose tabu search-based routing solver and apply the tool specifically to UAV routing. The tool consists of an interface by which the user can enter a UAV routing problem and a Java-based solver is used to return potential solutions. A subsequent empirical study involved academic benchmark problems and operationally motivated test instances. An open software version of the solver engine is available as OpenTS. Kinney *et al.* [23] consider a set of local searches, to include tabu search, for quickly solving routing problems, specifically UAV routing problems. This study included analyses regarding when to discontinue local search efforts and return the suggested solution. This quick-running heuristic approach was implemented in the UAV routing framework from Harder *et al.* [56]. Shetty *et al.* [57] recently considered the routing of UCAVs. Their problem considers target priorities, minimum and maximum levels of target damage, payload capacities, and range restrictions. They use tabu search to guide subordinate heuristic-solving aspects of the decomposed problem; their tabu search guides the iterative application of an assignment and a multiple traveling salesman heuristic.

A closely related problem is that of choosing flight paths. This is particularly applicable in mission planning to avoid terrain obstacles or known threats and in search planning. John *et al.* [58] use a GA

to determine a flight route for a regional surveillance problem. The GA incorporates constraints for no-fly zone considerations and was able to obtain reasonable solutions in a fraction of the time required by an IP-based solver. Kierstead and DelBalzo [59] employ a GA to design efficient search paths. While applicable to a variety of military applications, the specific application in this case is the detection of submarines within some enclosed search region.

Packing Problems

A packing problem seeks to place some set of objects within some specified, confined space. Attributes of these objects typically include dimensions and weight, but can expand to include myriad other attributes such as interactions with other objects requiring placement. The confined spaces can be single-dimensional, like a knapsack; two-dimensional, such as floor space on a truck; or three-dimensional, such as a shipping container. Typical packing applications can be found in shipping operations, military deployment load planning or storage facilities.

Packing heuristics have been widely used in mobility planning. Early approaches employed greedy algorithms. Cochard and Yost [60] describe the deployable mobility execution system, an early PC-based aircraft load planning system. The army automated load planning system (AALPS) was another early system and was deployed in 1978 [61]. Each tool provided initial load plans, using greedy construction approaches to accommodate various aircraft loading constraints, and allowed the user to interactively tailor the initial load plans. Heidelberg *et al.* [61] provide an improved two-dimensional packing algorithm for the gross load planning of aircraft.

Gueret *et al.* [1] consider heuristics for loading aircraft. They employ two construction heuristics modified by either of two local search algorithms. The first local search approach systematically changes the order items are loaded onto the aircraft to try and improve the solution. The second local search tries to reduce the number of aircraft

required by attempting to empty a selected aircraft's contents into other aircraft having available capacity. Harwig *et al.* [62] apply tabu search to the two-dimensional packing problem applied to the load planning for mobility operations. This problem arises when planning shipments of material or determining airlift requirements for a planned deployment. This approach could apply to the loading of pallets or the loading of aircraft. Baltacioglu *et al.* [63] develop a heuristic for three-dimensional packing of cargo pallets. Their heuristic mimics how humans create packings for a confined three-dimensional space. Related work includes a tabu search approach in Lodi *et al.* [64], a GA approach in Bortfeldt and Gehring [65], and an SA approach in Faina [66]. Morgan *et al.* [67] consider the problem of determining a mix of space launch vehicles for assigning a set of satellites to those vehicles. Modeled as a bin-packing problem, the authors generated over 3600 test problems, solved each to optimality or an upper bound, and examined how well a local improvement heuristic performed. The local heuristic found best solutions in 43% of the problems and on average came within 3–5% of the best-known solutions for the test problems.

Scheduling and Planning

Scheduling and planning involves the sequencing of tasks and the assignment of resources to the tasks. As tasks are sequenced, start times, finish times, and resource expenditures are compiled and examined to ensure solution feasibility. In some applications, such as aircrew scheduling, the plethora of requirements that must be satisfied makes mathematical programming approaches impractical due to the complexity of the formulation and the time required to obtain the proven optimal solution.

Sklar *et al.* [68] consider the problem of assigning crews to airlift routes to minimize the number of crews required to complete the assigned missions. They employ two constructive heuristics iteratively on decompositions of the problem. In their first heuristic, crews are assigned. These crews are used in the second heuristic to determine flight departure times. The departures are

used to reassign the crews when iterating back to the first heuristic. The process is complete when the entire solution stabilizes or an iteration limit is reached. Combs and Moore [69] apply tabu search to the crew scheduling problem. Their tabu search simultaneously determined crew rotations and crew schedules. The exact optimization method, a set-partitioning problem, was used to perform vocabulary building. Their approach then uses all of the partial crew rotations/schedules as columns of the set-partitioning problem (with tabu search as the column generator). Once the set-partitioning problem is solved, it would improve the best current solution by bring rotations/schedules previously not used into the solution. The improved solutions then become new points from which to restart the tabu search. The algorithm accounts for all crew rest and flight restrictions associated with the mobility air crew schedule. Lambert *et al.* [70] apply tabu search to the strategic airlift problem (SAP). The SAP solutions are

- pair aircraft and deployment requirements;
- determine aircraft missions required;
- time phase the aircraft along the mobility distribution network.

McGinnis *et al.* [71] develop a mathematical model for scheduling training resources associated with army initial entry training. A multiphase heuristic is used to develop a resource schedule, based on the projected trainee arrivals. The heuristic solutions are compared to exact solutions obtained via dynamic programming and are found to fall within 10% of the optimal solution, on average, but at a fraction of the computational time required. Butler *et al.* [72] present an integrated system to support the scheduling of combat engineers. They employ a GA to the task of assigning the combat engineering teams to a variety of combat engineering tasks. The scheduling challenge includes assigning and sequencing combat engineering units with differing capabilities to arrive at an efficient and effective schedule to accomplish the set of tasks. Their discussion includes specifics of

how the GA paradigm was tailored to the combat engineering scheduling problem.

Schlabach *et al.* [73] describe FOX-GA, a GA used to support military maneuver planning. The FOX-GA implementation involves three notable contributions. First, the authors use a fast-running aggregate model to evaluate potential solutions by using the time-consuming war games that are generally used. Such solution-evaluation efficiency is crucial for successful use of a GA. Second, the GA offers the decision maker a set of solutions, with a single solution via math programming or the full population of solutions. Finally, the authors implement niching, a GA strategy to ensure population diversity is maintained during the execution of the algorithm. Maintaining the diversity of the population ensures more efficient sampling of the solution search space.

Imsand *et al.* [74] use a GA to conduct a vulnerability analysis on the commander's analysis and planning simulation (CAPS), which is used for the operational analysis of missile defense systems. The GA was used to determine those factors most influencing the simulation output and thus, affecting the planning effectiveness.

Satellite range scheduling involves the assignment of tracking stations to supported satellites to satisfy requested satellite support. Gooley *et al.* [75] develop a heuristic for the Air Force Satellite Control Network scheduling problem. Their approach is to build an initial schedule and apply a local search involving interchange operators to improve the schedule. Results on an actual test problem were fair but obtained in minutes versus the longer time required by the manual process.

Final Comments

Our application review is by no means comprehensive; it is hopefully representative of how these heuristics are employed. We necessarily excluded from our review the vast material found in conference papers, working papers, technical reports, and graduate research documents. A scan of these materials will find the use of metaheuristics in nearly all types of optimization applications. We also did not address the use of heuristics

within a simulation-based optimization setting. Simulation-based optimization methods often use procedures based on SA, GAs, tabu search or scatter search. These methods are “wrapped” around military models and used to guide those models toward some best set of specified input parameters. For instance, Hill and McIntyre [76] use a GA, and later a scatter search-based technique, to find robust solutions to a multiscenario force planning problem involving a linear mathematical program.

CONCLUDING REMARKS

A variety of trends have encouraged the development and employment of search heuristics for optimization applications. Some of these were mentioned in the preceding sections. Others are discussed in the following paragraphs.

Modern computers are tremendously powerful computational devices. Some of us still remember punch-card computer programming or the initial “personal” computers that ran just on floppy diskettes. Modern computers are fast and have lots of memory. This means we can do more calculations and search among more solutions in less time than ever before. Modern computers make heuristic search methods feasible and practical.

Modern problems are very complex. This is not to say that past problems were not complex. In fact past problems were also complex but analytical methods of the time required we simplify these complex problems to accommodate the limited analytical methods. Our modern ability to solve large problems means we sometimes no longer want to simplify reality for the sake of an algorithm. We no longer always want exact solutions to approximations of the true system of interest. We now are often more satisfied with approximate solutions to problems that are more accurate reflections of the system of interest. These more accurate problems often create intractable formulations for classical solution methods. Current decision support requirements for problem solutions make heuristic search methods preferable.

The modern decision-making environment is too fast-paced to wait for a perfect answer (i.e., a guaranteed optimal answer). Getting that guaranteed optimal solution can sometimes take an inordinate amount of time. Getting an answer pretty close to an optimal can now be provided rapidly. These good, feasible solutions can be worth more to the decision maker in time-sensitive, decision-making situations. The pace of modern decision making has thus made heuristic search methods an attractive problem-solving methodology.

To close, the reader should not believe we advocate exclusive use of heuristic search methods for optimization applications; we do not. When practical and appropriate, mathematical programming methods should be preferred and exploited. However, we have hopefully provided the reader with the rationale, background, and motivation to appreciate the valuable role heuristic search methods can serve in an optimization model-based decision-making setting. When practical and appropriate, heuristic search methods should be explored, examined, and employed.

REFERENCES

1. Gueret C, Jussien N, Lhomme O, *et al.* Loading aircraft for military operations. *J Oper Res Soc* 2003;54:458–465.
2. Silver EA. An overview of heuristic solution methods. *J Oper Res Soc* 2004;55:936–956.
3. Osman IH, Kelly JP. Meta-heuristics: an overview. In: Osman IH, Kelly JP, editors. *Meta-heuristics: theory and applications*. Norwell (MA): Kluwer Academic Publishers; 1996.
4. Muller-Merbach H. Heuristics and their design: a survey. *Eur J Oper Res* 1981;8(1):1–23.
5. Hansen P, Mladenovic N. Variable neighborhood search. In: Glover F, Kochenberger G, editors. *Handbook of metaheuristics*. Norwell (MA): Kluwer Academic Publishers; 2003.
6. Henderson D, Jacobson SH, Johnson AW. The theory and practice of simulated annealing. In: Glover F, Kochenberger G, editors. *Handbook on metaheuristics*. Norwell (MA): Kluwer Academic Publishers; 2003. pp. 287–320.
7. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science* 1983;220(4598):671–680.

8. Velayudhan SP. Empirical analysis of randomness in ant colony optimization algorithms applied to the traveling salesman problem [Master's thesis]. Dayton (OH): Wright State University; 2004.
9. Reeves CR. Modern heuristics techniques for combinatorial problems. New York (NY): John Wiley & Sons, Inc.; 1993.
10. Renner G, Ekart A. Genetic algorithms in computer-aided design. *Comput Aided Des* 2003;35:709–726.
11. Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Norwell (MA): Addison-Wesley; 1989.
12. Zalazala, MS, Fleming PJ, editors. Genetic algorithms in engineering systems. London: Institution of Electrical Engineers; 1997.
13. Hoff A, Lokketangen A, Mittet I. Genetic algorithms for 0/1 multidimensional knapsack problems. *Proceedings Norsk Informatikk Konferanse, NIK '96*; Molde College, Norway. 1996. pp. 291–301.
14. Hill RR, Hiremath C. Improving genetic algorithm convergence using problem structure and domain knowledge in multidimensional knapsack problems. *Int J Oper Res* 2005;1(1/2):145–159.
15. Ross P. What are genetic algorithms good at? *INFORMS J Comput* 1997;9(3):260–262.
16. Reeves CR. Genetic algorithms. Modern heuristics techniques for combinatorial problems. New York: John Wiley & Sons, Inc.; 1993.
17. Beasley D, Bull DR, Martin RR. An overview of genetic algorithms, part 1: fundamentals. *Univ Comput* 1993;15(2):58–69.
18. Beasley D, Bull DR, Martin RR. An overview of genetic algorithms, part 2: research topics. *Univ Comput* 1993;15(4):170–181.
19. Reeves CR. Genetic algorithms for the operations researcher. *INFORMS J Comput* 1997;9(3):231–250.
20. Glover F. Heuristics for integer programming using surrogate constraints. *Decis Sci* 1977;8(1):156–166.
21. Senju S, Toyoda Y. An approach to linear programming with 0-1 variables. *Manage Sci* 1968;15(4):B196–B207.
22. Battiti R, Tecchiolli G. The reactive tabu search. *ORSA J Comput* 1994;6(2):126–140.
23. Kinney GW, Hill RR, Moore JT. Devising a quick-running heuristic for an unmanned aerial vehicle (UAV) routing systems. *J Oper Res Soc* 2005;56(7):776–786.
24. Glover F. Tabu search: a tutorial. *Interfaces* 1990;20(4):74–94.
25. Glover F. Tabu search - part I. *ORSA J Comput* 1989;1:190–206.
26. Glover F. Tabu search - part ii. *ORSA J Comput* 1990;2(1):4–32.
27. Glover F, Laguna M. Tabu search. Kluwer Academic Publishers; 1997.
28. Laguna M, Marti R. Scatter search: methodology and implementations in C. Norwell (MA): Kluwer Academic Publishers; 2003.
29. Glover F, Laguna M, Marti R. Fundamentals of scatter search and path relinking. *Control Cybern* 2000;29(3):653–684.
30. Marti R, Laguna M, Glover F. Principles of scatter search. *Eur J Oper Res* 2006;169(4):359–372.
31. Dorigo M, Maniezzo V, Coloni A. The ant system: optimization by a colony of cooperating ants. *IEEE Trans Syst Man Cybern Part B* 1996;26(1):1–13.
32. Velayudhan SP, Hill RR, Hiremath C, *et al.* Empirical analysis of randomness in ant colony optimization algorithms applied to the traveling salesman problem. *Int J Inf Syst Logist Manage* 2007;2(2):69–76.
33. Corne D, Dorigo M, Glover F, editors. New ideas in optimization. New York: McGraw Hill; 1999.
34. Talbi E. Metaheuristics: from design to implementation. Hoboken (NJ): John Wiley & Sons, Inc.; 2009.
35. Kennedy J, Eberhard R. Particle swarm optimization. *IEEE International Conference on Neural Networks*; Perth, Australia. 1995. pp. 1942–1948.
36. Clerc M, Kennedy J. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans Evol Comput* 2002;6:58–73.
37. Shi Y, Eberhard R. Empirical study of particle swarm optimization. Volume 3, *Proceedings of the 1999 Congress on Evolutionary Computation*. 1999. pp. 1945–1950.
38. Kennedy J, Mendes R. Population structure and particle swarm performance. *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. Honolulu (HI). 2002. pp. 1671–1676.
39. Poli R. Analysis of the publication on the applications of particle swarm optimisation. *J Artif Evol Appl* 2008;ID 685175:10.
40. Poli R, Kennedy J, Blackwell T. Particle swarm optimization: an overview. *J Swarm Intell* 2007;1:33–57.

41. Wang G, Chen X, Fan Y. UCAV path planning based on modified PSO algorithm. *Aeronaut Comput Tech* 2007;37(4).
42. Foo J, Knutzon J, Oliver J, *et al.* Three-dimensional path planning of unmanned aerial vehicles using particle swarm optimization. 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, AIAA 2006-6995; 2006 Sep 6–8; Portsmouth, Virginia. 2006.
43. Cui X, Potok T. A particle swarm model for multi-agent based insurgency warfare simulation. Fifth International Conference on Software Engineering Research, Management and Applications. Mt. pleasant (MI). 2007. pp. 177–183.
44. Battilega JA, Grange JK. The military applications of modeling. Ohio: Wright-Patterson Air Force Base: Air Force Institute of Technology; 1984.
45. Wacholder E. A neural network-based optimization algorithm for the static weapon-target assignment problem. *ORSA J Comput* 1989;1(4):232–246.
46. Green DJ, Moore JT, Borsi JJ. An integer solution heuristic for the arsenal exchange model (AEM). *Mil Oper Res* 1997;3(2):5–15.
47. Cullenbine CA, Gallagher MA, Moore JT. Assigning nuclear weapons with reactive tabu search. *Mil Oper Res* 2003;8(1):57–69.
48. Lee ZJ, Lee CY, Su SF. An immunity-based any colony optimization algorithm for solving weapon-target assignment problem. *Appl Soft Comput* 2002;2(1):39–47.
49. Jiao L, Wang L. A novel genetic algorithm based on immunity. *IEEE Trans Syst Man Cybern A* 2000;30(5):552–561.
50. Blodgett DE, Gendreau M, Guertin F, *et al.* A tabu search heuristic for resource management in naval warfare. *J Heuristics* 2003;9(2): 145–169.
51. Erdem E, Ozdemirel NE. An evolutionary approach for the target allocation problem. *J Oper Res Soc* 2003;54(10):958–969.
52. Aardal KI, van Hoesel SPM, Koster AMCA, *et al.* Models and solution techniques for frequency assignment problems. *Ann Oper Res* 2007;153(1):79–129.
53. Lau TL, Tsang EPK. Guided genetic algorithm and its application to radio link frequency assignment problems. *Constraints* 2001; 6(4):373–398.
54. Ryan JL, Bailey TG, Moore JT. Unmanned aerial vehicle (UAV) route selection using reactive tabu search. *Mil Oper Res* 1999;4(3): 5–24.
55. O'Rourke KP, Bailey TG, Hill RR, *et al.* Dynamic routing of unmanned aerial vehicles using reactive tabu search. *Mil Oper Res* 2001;6(1):5–30.
56. Harder RW, Hill RR, Moore JT. A java universal vehicle router for routing unmanned aerial vehicles. *Int Trans Oper Res* 2004; 11(3):259–276.
57. Shetty VK, Sudit M, Nagi R. Priority-based assignment and routing of a fleet of unmanned combat aerial vehicles. *Comput Oper Res* 2008;35(6):1813–1828.
58. John M, Panton D, White K. Mission planning for regional surveillance. *Ann Oper Res* 2001;108(1/4):157–173.
59. Kierstead DP, DelBalzo DR. A genetic algorithm applied to planning search paths in complicated environments. *Mil Oper Res* 2003; 8(2):45–60.
60. Cochard DD, Yost KA. Improving utilization of air force cargo aircraft. *Interfaces* 1985;15(1):53–68.
61. Heidelberg KR, Parnell GS, Ames JE IV. Automated air load planning. *Naval Res Logist* 1998;45(8):751–768.
62. Harwig JM, Barnes JW, Moore JT. An adaptive tabu search approach for 2-dimensional orthogonal packing problems. *Mil Oper Res* 2006;11(2):5–26.
63. Baltacioglu E, Moore JT, Hill RR. The distributor's three-dimensional pallet-packing problem: a human intelligence-based heuristic approach. *Int J Oper Res* 2006;1(3):249–266.
64. Lodi A, Martello S, Vigo D. Heuristic algorithms for the three-dimensional bin packing problem. *Eur J Oper Res* 2002;141(2): 410–420.
65. Bortfeldt A, Gehring H. A hybrid genetic algorithm for the container loading problem. *Eur J Oper Res* 2001;131(1):143–161.
66. Faina L. A global optimization algorithm for the three-dimensional packing problem. *Eur J Oper Res* 2000;126(2):340–354.
67. Morgan LO, Morton AR, Daniels RL. Simultaneously determining the mix of space launch vehicles and the assignment of satellites to rockets. *Eur J Oper Res* 2006;17(3):747–760.
68. Sklar MG, Armstrong RD, Samn S. Heuristics for scheduling aircraft and crew during airlift operations. *Transp Sci* 1990;24(1):63–76.
69. Combs TE, Moore JT. A hybrid tabu search/set partitioning approach to tanker crew scheduling. *Mil Oper Res* 2004;9(1):43–56.

70. Lambert GR, Barnes JW, Van Veldhuizen DA. A tabu search approach to the strategic airlift problem. *Mil Oper Res* 2007;12(3):59–79.
71. McGinnis ML, Gernandez-Gaucherand E, Mirchandani PB. Military training resource scheduling: System model, optimal and heuristic decision processes. *Mil Oper Res* 1996;2(1):53–69.
72. Butler CD, Boggess G, Bridges S. Knowledge based planning of military engineering tasks. *Expert Syst Appl* 1999;16(2):221–232.
73. Schlabach JL, Hayes CC, Goldberg DE. Foxga: a genetic algorithm for generating and analyzing battlefield courses of action. *Evol Comput* 1999;7(1):45–68.
74. Imsand ES, Evans G, Dozier G, *et al.* Using genetic algorithms to aid in a vulnerability analysis of national missile defense simulation software. *J Defense Model Simul* 2004;1(4):215–223.
75. Gooley TD, Borsi JJ, Moore JT. Automating air force satellite control network (AFSCN) scheduling. *Math Comput Model* 1996;24(2):91–101.
76. Hill RR, McIntyre GA. A methodology for robust, multi-scenario optimization. *Phalanx* 2000;33(3):27–31.

HEURISTICS FOR THE TRAVELING SALESMAN PROBLEM

MARY E. KURZ

Department of Industrial Engineering,
Clemson University, Clemson, South
Carolina

INTRODUCTION

In the context of this article, we consider a *heuristic* to be a method to construct a feasible solution to a TSP instance. TSP heuristics can be considered to be one of three types: constructive, improvement, or composite. We specifically do not include *metaheuristics* in our discussion, though one might consider them to be improvement heuristics. Composite TSP heuristics use a constructive heuristic followed by an improvement phase, so this article does not explicitly consider composite heuristics.

We define a TSP instance by n , the number of cities, and a distance matrix $\mathbf{D}$, whose entries d_{ij} represent the distance associated with traveling directly from city i to city j . Every nondiagonal entry in $\mathbf{D}$ is assumed to exist. Unless otherwise noted, the distance matrix $\mathbf{D}$ is assumed to be asymmetric, non-Euclidean, and not to satisfy the triangle inequality. The interested reader is encouraged to consult one or more of the comprehensive surveys of TSP heuristics, including Refs 1–5. Laporte [6] very recently provided an aptly named concise guide to the TSP, including key results about heuristics for both symmetric and asymmetric TSPs and an extensive bibliography. Emphasis is placed in this article on the most referenced or enduring techniques, rather than providing a historical view of the development of these heuristics. Furthermore, heuristics for the plethora of TSP variants, such as the generalized TSP, are not considered here. References [1–6] also contain extensive computational results, while Bentley [7] details fast implementations of several methods.

CONSTRUCTION HEURISTICS

Construction heuristics iteratively build one feasible solution through the application of a deterministic set of steps. A random permutation of cities can create an initial solution as well. Following the detailed exposition by Reinelt [8], the construction heuristics are divided into nearest neighbor heuristic, insertion heuristics, spanning tree-based heuristics, and savings heuristics. We then consider heuristics explicitly developed for the asymmetric TSP.

Nearest Neighbor Heuristic

Nearest-neighbor heuristics build a solution directly by adding cities to the end of a partial solution. Nearest-neighbor heuristics on a set N of cities have the following basic steps:

1. select a starting city, j . $N = N \setminus \{j\}$;
2. as long as N is not empty,
 - a. select the next city $i : i = \arg \min_N (d_{ji})$;
 - b. append i to the end of the partial solution;
 - c. $N = N \setminus \{i\}$.

The nearest-neighbor heuristic is an $O(n^2)$ algorithm with an arbitrarily poor performance [1,8].

Insertion Heuristics

Insertion heuristics build a solution directly by inserting cities one at a time *into* a partial solution, as opposed to the end of a partial solution. Insertion heuristics vary in their criteria for selecting the next city and determining the position for this city. These two steps are usually considered in tandem, meaning that the determination of *which* city is to be inserted depends on *where* in the partial solution the city is to be inserted. We assume that the city is to be inserted at the location that minimally increases the partial solution total distance, unless otherwise noted. Let the set of cities in any partial solution be called *solution cities*, represented by P .

Basic Insertion Heuristics. The basic insertion heuristics on a set N of cities have the following basic steps, assuming starting with an empty partial solution:

1. select a starting city, j . $N = N \setminus \{j\}$;
2. as long as N is not empty,
 - a. select the next city i ;
 - b. insert i into the partial solution;
 - c. $N = N \setminus \{i\}$.

Selection and Insertion Criteria. Several criteria are summarized [1,8,9] for application in the basic insertion heuristics; no doubt, several more could be conceived. A subset is included as follows:

- *Nearest insertion*—inserts the city that has the shortest distance to a solution city: $i = \arg \min_N (d_{ji} | j \in P)$.
- *Farthest insertion*—inserts the city that has the longest distance to a solution city: $i = \arg \max_N (d_{ji} | j \in P)$.
- *Cheapest insertion*—inserts the city that increases the total length of the partial solution with its inclusion the least.
- *Random insertion*—inserts a randomly selected city.

In general, these criteria, when used with basic insertion heuristics, result in $O(n^2)$ methods [except for cheapest insertion, which can be $O(n^2 \log n)$] [8].

Minimum Spanning Tree Heuristics

If the distance matrix $\mathbf{D}$ satisfies the triangle inequality and is symmetric, then two methods based on the concepts of minimum spanning trees can be used to find a feasible solution to the TSP. We let a graph G represent the distance matrix, so that each node is a city and each edge in G has the corresponding distance from $\mathbf{D}$. The interested reader should also reference material on Basic notation and graph theory concepts.

The Doubling method can be summarized as follows:

1. compute a minimum spanning tree of the graph G and call it M ;

2. double every edge of M , yielding an Eulerian graph called E ;
3. find an Eulerian trail of E ;
4. find a Hamiltonian cycle of E .

The Christofides method differs from the Doubling method in step 2. Christofides suggested finding a minimum weight perfect matching, W , of the odd-degree nodes in M . Then $W + M$, formed by adding the edges in W to M , is an Eulerian graph. The Doubling and Christofides methods are the primary construction heuristics for the TSP with a distance matrix that is symmetric and satisfies the triangle inequality presented in Papadimitriou and Steiglitz [10].

The Doubling method has a running time of n^2 but a worse case bound of two times the optimal tour length. The Christofides method has a running time of n^4 or n^3 (depending on the implementation of the minimum weight perfect matching) and a worst case bound of 1.5 times the optimal tour length [10].

Savings Heuristic

Two comparative studies of heuristics [1,8] found that the Savings Heuristic, also known as the *Clarke and Wright savings heuristic*, performed quite well. The savings heuristic was developed for the vehicle routing problem and adapted to the TSP by Golden [1,8]. The savings heuristic requires the following steps:

1. select a starting city, b . $N = N \setminus \{b\}$;
2. compute the savings $s_{ij} = d_{bi} + d_{bj} - d_{ij}$, for $i, j \in N$;
3. sort the list of savings in descending order;
4. in the order of the sorted savings list, link the edges i and j until a single feasible solution is formed.

On the basis of the definition of the savings, it is presumed that a symmetric distance matrix is required.

A Construction Heuristic for Asymmetric TSP

Laporte [6] identifies the Karp Patch heuristic as being best in its class. In this method, an assignment problem is created using the

distance matrix **D**. If the optimal solution to the assignment problem is not a feasible TSP solution, corrections are made in a process called *patching*. The steps as described in Ref. 11 are as follows:

1. create and solve the related assignment problem. If a feasible TSP solution is found, stop;
2. otherwise, while a feasible TSP solution does not exist,
 - a. select the two cycles with the most cities;
 - b. intelligently combine them by deleting one arc in each cycle and connecting the cycles with arcs between them.

IMPROVEMENT HEURISTICS

Improvement heuristics assume the existence of a feasible solution to the TSP, created in some manner. Improvement heuristics may be good additions to metaheuristic techniques for the TSP (see the articles titled **Guided Local Search**, **Evolutionary Algorithms**, **Ant Colony Optimization**, **Memetic Algorithms**, **Genetic Algorithms**, **Simulated Annealing** in this encyclopedia), providing neighborhoods for local search techniques.

Basic Improvement Methods

Though not often reported in research, simple improvement methods may be desirable for use in metaheuristic approaches due to their speed. These two methods are described in Ref. 12.

The node insertion heuristic considers moving a city from its current location to a different location. The detailed steps can be described as follows:

1. Let T be a current feasible solution.
2. Until $done = \text{TRUE}$,
 - a. For each city $i \in N$: Consider moving city i to all possible other locations in the tour. If the best new location reduces the current tour length, move city i to that location.

- b. If no potential moves for any city provide improvement, set $done = \text{TRUE}$.

The edge insertion heuristic considers moving an edge from its current location to a different location. We could interpret this as moving two adjacent cities from their current location to a new location. The detailed steps can be described as follows:

1. Let T be a current feasible solution;
2. Until $done = \text{TRUE}$,
 - a. For each city $i \in N$: Let $j =$ the city following i in T . Consider moving cities i and j to all possible other locations in the tour. If the best new location reduces the current tour length, move cities i and j to that location by resetting T .
 - b. If no potential moves for any pairs of adjacent cities provide improvement, set $done = \text{TRUE}$.

Reinelt [12] provides computational analysis of the effectiveness of these methods in improving solutions developed with random, nearest neighbor, and savings heuristics, reporting that they are very slow and perform poorly on randomly generated initial tours. However, these form the basis of understanding the Edge Exchange methods below.

Edge Exchange Methods: 2-opt, 3-opt, Lin-Kernighan

Extending the idea of moving one edge from one location to another, edge exchange methods are built on the idea of deleting some number of edges and reattaching the remaining tour sections in a feasible manner.

The most basic of these exchange methods, 2-opt, involves deleting two edges, say (i,j) and (k,l) . Assuming that cities i,j,k , and l appear in this order in the current tour T , then the only way to reconnect them is with the edges (i,k) and (j,l) , with the cities between j and k being reversed in the new tour T' . In fact, Croes called this concept *inversion* [13]. Lin [14] generalized these ideas with the concept of λ -optimality, which is the dominant feature in good improvement methods for the TSP. A tour is λ -optimal if

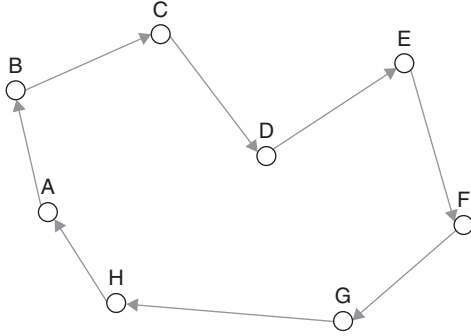


Figure 1. A feasible eight-city tour.

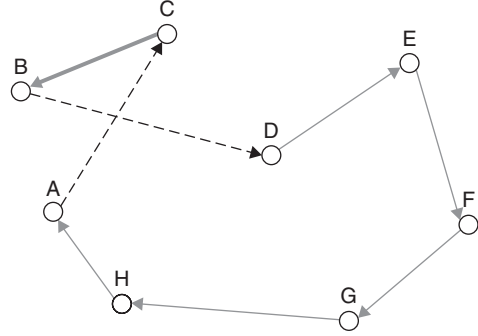


Figure 2. After a node insertion.

any sets of λ edges do not exist that can be replaced by another set of λ edges, reducing the total tour length. Unfortunately, the concept of λ -optimality and the methods deriving from the idea of λ -optimality are both called by some variant of “ λ -opt”. These statements illustrate the two usages of the phrase: (i) “The tour is 3-opt”; (ii) “I applied 2-opt to the problem.” The reader should use care when interpreting these terms. Golden [9] observes that a λ -opt heuristic is an $O(n^\lambda)$ method.

Figure 1 is a feasible tour of eight cities labeled A to H that are used to illustrate some concepts of the edge exchange methods. We can refer to tour as a list of the cities, arbitrarily beginning with A: ABCDEFGH. The figure provides directional information for the tour through the explicitly included arrows on the arcs.

2-opt Heuristic. 2-opt, as a heuristic, can be implemented as follows [12]:

1. Let T be a current feasible solution;
2. Until $done = \text{TRUE}$,
 - a. For each city $i \in N$: Let j = the city following i in T . Consider exchanging cities i and j with all other pairs of cities k and l in the tour. If the best exchange (say with k^* and l^*) reduces the current tour length, reset T as follows: (i) delete the edges (i,j) and (k^*,l^*) ; (ii) insert the edges (i,k^*) and (j,l^*) ; (iii) reverse the cities between j and k ;
 - b. If no potential exchanges provide improvement, set $done = \text{TRUE}$.

The resulting tour is said to be 2-optimal.

Note that the node insertion local search is a special case of the 2-opt, in which k is the city following j in T . Figure 2 shows how the tour ABCDEFGH can be transformed through a node insertion; specifically, node B is inserted between nodes C and D, resulting in the tour ACBDEFGH. Note that arcs AB and CD have been removed, arcs AC and BD have been added, and the arc BC is now inverted to be arc CB. In general, 2-opt need not invert only a single arc, as illustrated in Fig. 3. Figure 3 shows how a general 2-opt move may be performed on the tour in Fig. 1. Edges CD and GH are removed and edges CG and DH are added. Moreover, edges DE, EF, and FG are now inverted.

3-opt Heuristic. The definition of k -optimality implies the creation of $n - 1$ possible heuristics. On the basis of the computational complexity, each increase in

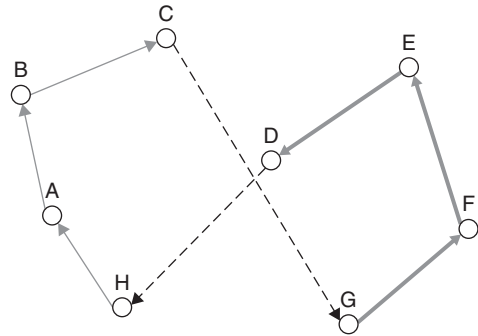


Figure 3. After a 2-opt move.

k significantly increases the computational complexity. 3-opt is the largest version that is commonly seen. The intent is to find triples of edges that should be deleted and replaced with three new edges, while maintaining a feasible solution. Lin [14] notes that this is conceptually equivalent to identifying a set of cities to be removed from their current location (breaking two edges), possibly inverted and inserted between two cities (breaking the third edge).

On the basis of this observation, 3-opt, as a heuristic, can be implemented as follows:

1. Let T be a current feasible solution;
2. Until $done = \text{TRUE}$,
 - a. For each city $i \in N$ and $j \in N \setminus \{i\}$: Consider inserting the inverted and uninverted set of cities between and including i and j with all other adjacent pairs of cities k and l in the tour. If the best exchange reduces the current tour length, reset T by making the insertion (and inversion, if required) in T ;
 - b. if no potential exchanges provide improvement, set $done = \text{TRUE}$.

The resulting tour is said to be 3-optimal.

Note that the edge insertion local search is a special case of the 3-opt, in which the edge to be moved in edge insertion is not inverted and therefore uniquely identifies the edges to be removed and reformed. Figure 4 shows how the tour ABCDEFGH can be transformed through an edge insertion or 3-opt move. First, consider how the

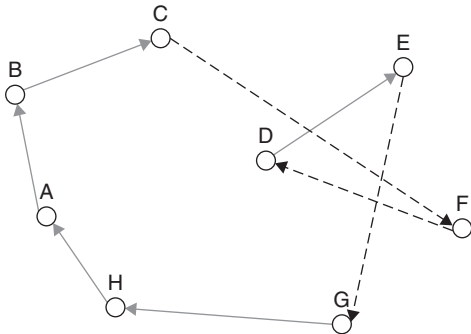


Figure 4. After an edge insertion OR 3-opt move.

result is achieved from the edge insertion point of view. The edge DE is removed from its original location between C and F and inserted between the nodes F and G, resulting in the tour ABCFDEGH. From the 3-opt point of view, we can see that this breaks three edges: CD, EF, and FG. The new edges are CF, FD, and EG.

Lin–Kernighan Heuristic. Laporte [6] reports that the dominant heuristics for symmetric TSPs are based on the Lin–Kernighan heuristic [15]. References [5] and [6] identify Helsgaun’s variant as performing very well, even for the asymmetric case [11]. The Lin–Kernighan Heuristic is much more complicated than any other method presented in this article. The basic idea is that the value of λ that should be used is unknown; the decision to use 2-opt or 3-opt, or a different value, is arbitrary. Therefore, begin by considering 2-opt and then incrementally consider increasing values of λ .

The Golden *et al.* [9] presentation of the Lin–Kernighan heuristic lies between the length in Refs 4 and 12. It is presented, slightly adapted, here for the convenience of the reader. G_p^* is the improvement that can be gained if the p edges currently under consideration are deleted and replaced with p different edges.

1. Let T be a current feasible solution.
2. Set $G_0^* = 0$. Select a city l and either edge including it from T , (l, k) for possible removal. Set $p = 1$.
3. Find the edge (k, i) not in T such that $i = \arg \max(g_1)$, where $g_1 = d_{lk} - d_{ki} > 0$. The value g_1 is positive when the selected edge is smaller than (l, k) ; we select the smallest of all such edges.
4. At this point, we anticipate that edge (l, k) will be removed and (k, i) will be added. Some edge currently incident to i must leave, or i will have degree 3. Select one of these edges, (i, j) , to leave. We could complete a 2-opt-type move at this time if we select the edge (j, l) to be added to T . If we did this (remove (l, k) and (i, j) , and add (k, i)

and (j, l) , $G_1^* = g_1 + d_{ij} - d_{jl}$. For now, assume that making this move will reduce the total tour length, meaning $G_1^* > 0$. If not, go to 2 and select a different city l .

5. However, it may be that edge (j, l) is not the one to be added. It may be that there is a different edge, adjacent to j , which would contribute more to reducing the tour length. To find the best of the edges adjacent with j , find the edge (j, m) not in T such that $m = \arg \max(g_p)$, where $g_1 = d_{ij} - d_{jm}$ (the edge (j, l) is one of the potential edges to be considered). Compute $G_p = \sum_{s=1}^p g_s$ and $G_p^* = G_p + c_{mn} - c_{nl}$. Let $G^* = \max_{s=1 \dots p} \{G_s^*\}$. Increment p and go to step 5 unless
 - No further feasible swaps exist.
 - the current configuration is a tour.
 - $G_p \leq 0$ (the attempt does not reduce the tour length).
 - $G_p \leq G^*$ (the attempt is worse than the best option seen so far).

Stop and select the best of the tours associated with the $G_s^*, s = 1 \dots p$ values.

The increased coding and computational complexity introduced by this algorithm, and particularly by Helsgaun's variant [16], appear to be justified by the quality of solutions found by it. Helsgaun finds several new upper bounds to symmetric problem instances in TSPLIB, as well as performing well for asymmetric instances.

CONCLUSION

This article has provided an overview of five construction heuristic classes and two improvement heuristic classes. While the literature contains endless variants of the TSP, these heuristic examples contain those that form the basis of understanding the TSP literature. The interested reader is directed to Refs 1–6 for more detailed comprehensive surveys of TSP heuristics, with extensive computational results, and to Ref. 7 for fast implementations of several methods.

REFERENCES

1. Golden B, Bodin L, Doyle T, Stewart W Jr. Approximate traveling salesman algorithms. *Oper Res* 1980;28(3):694–711. DOI: 10.1287/opre.28.3.694.
2. Lawler EL, Lenstra JK, Rinooy Kan AHG, Shmoys DB, editors. The traveling salesman problem: A guided tour of combinatorial optimization. Chichester: John Wiley; 1985.
3. Reinelt G. The traveling salesman: computational solutions for TSP applications. *Lecture Notes in Computer Science* 840. Berlin: Springer; 1994. DOI: 10.1007/3-540-48661-5.
4. Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Boston: Kluwer; 2002.
5. Applegate DL, Bixby RE, Chvatal V, Cook WJ. The traveling salesman problem: A computational study. Princeton: Princeton Press; 2006. pp. 425–488.
6. Laporte G. A concise guide to the traveling salesman problem. *J Oper Res Soc* 2010;61:35–40. DOI: 10.1057/jors.2009.76.
7. Bentley JJ. Fast algorithms for geometric traveling salesman problems. *ORSA J Comput* 1992;4(4):387–411.
8. Reinelt G. The traveling salesman: Computational solutions for TSP applications. *Lecture Notes in Computer Science* 840. Berlin: Springer; 1994. pp. 73–99. DOI: 10.1007/3-540-48661-5_6.
9. Golden BL, Stewart WR. Empirical analysis of heuristics. In: Lawler EL, Lenstra JK, Rinooy Kan AHG, Shmoys DB, editors. The traveling salesman problem: A guided tour of combinatorial optimization. Chichester: John Wiley; 1985: 207–250. 1985.
10. Papadimitriou CH, Steiglitz K. Combinatorial optimization: Algorithms and complexity. Mineola: Dover; 1998. pp. 410–418.
11. Johnson DS, Gutin G, McGeoch LA, Yeo A, Zhang W, Zverovitch A. Experimental analysis of heuristics for the ATSP. In: Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Boston: Kluwer; 2002. 445–487.
12. Reinelt G. The traveling salesman: Computational solutions for TSP applications. *Lecture Notes in Computer Science* 840. Berlin: Springer; 1994. 100–132. DOI: 10.1007/3-540-48661-5_7.
13. Croes GA. A method for solving traveling-salesman problems. *Oper Res* 1958;6(5): 791–812.

14. Lin S. Computer solutions of traveling salesman problem. *Bell Syst Tech J* 1965;44(10): 2245–2269.
15. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling-salesman problem. *Oper Res* 1973;21:498–516.
16. Helgsaun K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. *Eur J Oper Res* 2000;126: 106–130.

HEURISTICS IN MIXED INTEGER PROGRAMMING

MATTEO FISCHETTI
DEI, Università di Padova, Padova,
Italy

ANDREA LODI
DEIS, Università di Bologna,
Bologna, Italy

INTRODUCTION

We consider a generic mixed integer linear program (MILP) in the form

$$\min c^T x \quad (1)$$

$$Ax \geq b \quad (2)$$

$$x_j \in \{0, 1\} \quad \forall j \in \mathcal{B}, \quad (3)$$

$$x_j \text{ integer} \quad \forall j \in \mathcal{G}, \quad (4)$$

$$x_j \text{ continuous}, \quad \forall j \in \mathcal{C}, \quad (5)$$

where A is an $m \times n$ input matrix, and b and c are input vectors of dimension m and n , respectively. Here, the variable index set $\mathcal{N} := \{1, \dots, n\}$ is partitioned into $(\mathcal{B}, \mathcal{G}, \mathcal{C})$, where $\mathcal{B}$ is the index set of the 0–1 variables (if any), while the sets $\mathcal{G}$ and $\mathcal{C}$ index the general integer and the continuous variables, respectively. Bound on the variables, including the 0–1 bounds on the binary ones, are assumed to be part of system (2). Removing the integrality requirement on variables indexed by $\mathcal{I} := \mathcal{B} \cup \mathcal{G}$ leads to the *LP relaxation* $\min\{c^T x : x \in P\}$ where $P := \{x \in \mathcal{R}^n : Ax \geq b\}$.

MILP heuristics aim at finding a feasible (and hopefully good) solution of the problem above, which is an NP-hard problem by itself. We next present the main ideas underlying some of the heuristics proposed in the literature. In this article, in particular, we focus on those algorithms developed with the aim of being tightly integrated within MILP solvers. We group and discuss the algorithms

depending on the main ingredient (building block) they use, namely LP-based heuristics in the section titled “LP-Based Heuristics” and MILP-based approaches in the section titled “MILP-Based Heuristics.”

With a little abuse of terminology, in what follows, we will say that a point x is *integer* if x_j is integer for all $j \in \mathcal{I}$ (no matter what the value of the other components), whereas the *rounding* $\tilde{x}$ of a given x will be defined as $\tilde{x}_j := \lfloor x_j \rfloor$ if $j \in \mathcal{I}$ and $\tilde{x}_j := x_j$ otherwise ($\lfloor \cdot \rfloor$ representing scalar rounding to the nearest integer).

We end the introduction by noting that, although Achterberg [1] has shown that the impact of heuristics is not dramatic in terms of ability of an MILP solver to prove optimality in a (much) quicker way, the psychological impact for the user who sees a high quality feasible solution provided very early is huge. For this reason, the availability of very effective general-purpose heuristics for MILP is among the most crucial improvements over the last 10 years.

LP-BASED HEURISTICS

In this section we address some basic heuristics that only assume the availability of a “black-box” LP solver.

Folklore

The so-called *rounding* as well as *diving* heuristics belong to folklore.

Rounding methods solve the LP relaxation to get an optimal point x^* , whose fractional components x_j^* with $j \in \mathcal{I}$ are rounded, for example, to their nearest integer value. Continuous variables, if any, can optimally be recomputed once the integer variables have been fixed by solving an LP.

Diving (also known as *Relax-and-Fix*) methods, instead, mimic an enumerative scheme by sequentially fixing some integer-valued variables that assume a fractional value in the solution of the current LP-relaxation. The sequence can be viewed

as “diving” on a branch-and-bound tree until a feasible (and hopefully good) MILP solution is found. A limited backtracking is often allowed to investigate alternative fixing patterns. Various variable-fixing (i.e., branching) policies have been used by different authors, all meant to hopefully reach deep branching nodes while maintaining a good LP relaxation lower bound value.

Pivoting Methods

These methods originate from the seminal work of Balas and Martin [2] in the late 1970s. The basic heuristic, called *Pivot and Complement* (P&C), applies to pure 0–1 integer linear programs (ILPs) with $\mathcal{G} = \mathcal{C} = \emptyset$. It is based on the observation that a feasible solution is just a basic solution of the LP relaxation where all 0–1 variables are non-basic (either at their lower or upper bound), that is, only the slack (continuous) variables on Equation (2) are allowed to be basic. Given an optimal basis of the LP relaxation, P&C then performs a systematic series of “non-standard” pivots in the attempt of driving as many 0–1 variables as possible out of the basis, while preserving primal feasibility and worsening the objective function as little as possible.

More specifically, the method prescribes the use of three types of pivots: (i) Type 1 pivots that maintain primal feasibility and increase the number of nonbasic 0–1 variables; (ii) Type 2 pivots that also maintain primal feasibility but leave the number of nonbasic 0–1 variables unchanged, while reducing the *infeasibility degree* of the current LP solution x^* , computed as $\sum_{j \in B} \min\{x_j^*, 1 - x_j^*\}$; (iii) Type 3 pivots that increase the number of nonbasic 0–1 variables at the price of producing a primal-infeasible point—a situation to be repaired at a later step of the procedure.

Nonbasic 0–1 variables can also be complemented with the aim of improving the solution quality or reducing its primal infeasibility, and reduced costs are exploited to fix variables when mathematically correct. The P&C heuristic was later improved by Løkketangen *et al.* [3] by adding tabu search to the basic mechanisms.

In the mid-1980s Balas and Martin [4] introduced a generalization of P&C to MILPs, called *pivot and shift*, where general integer variables are treated as 0–1 variables “centered” on the current LP solution, and some new types of “nonstandard” pivots are introduced. An elaboration with the same name of this idea was later proposed by Balas *et al.* [5]. Different extensions called *pivot*, *cut*, and *dive* and *pivot and Gomory cut* were proposed by Nediak and Eckstein [6] and by Gosh and Hayward [7], respectively.

OCTANE and Line Search Methods

Line search methods are based on the following construction. One starts from an LP relaxation optimal vertex x^* and draws a line toward a second point, for example, y , to be chosen according to a certain criterion. Thus, moving in a discretized way from x^* to y traces a discrete sequence of (say) K points $x^k := x^* + \alpha_k(y - x^*)$ for $0 = \alpha_1 < \dots < \alpha_K = 1$. Each x^k can therefore be associated (e.g., by rounding) with an integer point $\tilde{x}^k$ which is then checked for feasibility and possibly used to update a current-best MILP solution. Local shifts of the rounded variables are typically allowed in the attempt of improving the feasibility and/or quality of the rounded points.

Hillier [8] introduced this scheme by defining y so as to go inside P along a so-called *interior path*, the rational being supported by the geometrical intuition that (mainly for general integer MILPs) going inside P increases the chances the rounded point be feasible. A similar scheme was later adopted by Ibaraki *et al.* [9] and by Faaland and Hiller [10], where the sampling line segment $[x^*, y]$ is replaced by a sequence of linear segments $[x^*, y^1]$, $[y^1, y^2]$, $\dots$ approximating a curved search trajectory. Very recently, Naoum-Sawaya and Elhedhli [11] investigated the effect of replacing the LP solver by an analytic-center method.

OCTANE (for Octahedral Neighborhood Enumeration) is a line search method proposed by Balas *et al.* [12] for 0–1 ILPs. It also starts with an LP optimal vertex x^* and moves it along a certain search direction $y - x^*$. However, the integer (possibly infeasible) points discovered along this direction

are not defined through rounding, but exploit the following construction. Consider the center $x^0 = (1/2, \dots, 1/2)$ of the unit hypercube. For any vertex $\tilde{x}$ of the unit hypercube (not necessarily in P), define the hyperplane $H(\tilde{x})$ passing through $\tilde{x}$ and orthogonal to $\tilde{x} - x^0$.¹ Starting from x^* and going toward y one crosses, in sequence, certain hyperplanes $H(\tilde{x}^1), H(\tilde{x}^2), \dots$ whose generating 0–1 points $\tilde{x}^1, \tilde{x}^2, \dots$ are checked for feasibility and used to possibly update the current-best feasible solution. *OCTANE* prescribes to visit only a limited number of hyperplanes, and exploits an effective method for their enumeration.

Feasibility Pump

The *Feasibility Pump* (FP) is a method originally proposed by Fischetti *et al.* [13] for 0–1 MILPs ($\mathcal{G} = \emptyset$), and then extended by Bertacco *et al.* [14] to the general case. It is based on the observation that a feasible MILP solution is a point x of P that coincides with its rounding. Replacing “coincides” with “is as close as possible” leads to the following iterative scheme. FP works with a *pair* of points $(x^*, \tilde{x})$ with $x^* \in P$ and $\tilde{x}$ integer, that are iteratively updated with the aim of reducing as much as possible their L_1 -distance $\Delta(x^*, \tilde{x}) := \sum_{j \in \mathcal{I}} |x_j^* - \tilde{x}_j|$. In a sense, the method is aimed at “pumping” the integrality of $\tilde{x}$ into x^* , and the LP-feasibility of x^* into $\tilde{x}$ —hence its name.² To be more specific, one starts with any $x^* \in P$, for example, an optimal LP vertex, and initializes an integer $\tilde{x}$ as the rounding of x^* . At each FP iteration, called a *pumping cycle*, $\tilde{x}$ is fixed and one finds a point $x^* \in P$ which is as close as possible to $\tilde{x}$ by solving the auxiliary LP $\min\{\Delta(x, \tilde{x}) : x \in P\}$. If $\Delta(x^*, \tilde{x}) = 0$, then x^* is integer and the method is completed. Otherwise, $\tilde{x}$ is replaced by the rounding of x^* so as to hopefully reduce $\Delta(x^*, \tilde{x})$ even further, and the process is iterated. The basic

FP scheme above may stall in case $\Delta(x^*, \tilde{x})$ is not reduced at some iteration, hence a number of diversification mechanisms are introduced.

As stated, FP is meant to only produce a feasible MILP solution, as the objective function is only taken into account implicitly in the definition of the very first x^* . In order to produce good quality solutions, the original FP approach [13] uses a “sliding window” method that introduces the cut $c^T x \leq UB$ into the LP model and updates UB , on the fly, by taking an intermediate value between the optimal LP-relaxation and the UB values. (Note that the value UB can also be a “guess” of a feasible solution value and in such a case the cut can be possibly invalid.) A different approach, called *objective FP*, was developed by Achterberg and Berthold [15], who modified the objective function of the auxiliary LP to $\Delta(x^*, \tilde{x}) + \alpha c^T x$, where α is a dynamically updated parameter. Very recently, Fischetti and Salvagnin [16] proposed a more elaborated FP version, called FP 2.0, where the original rounding operation used to construct $\tilde{x}$ from x^* is replaced by a more clever rounding heuristic based on constraint propagation—a basic tool in constraint programming (see **Constraint Programming Links with Math Programming**).

MILP-BASED HEURISTICS

We next address a new generation of MILP heuristics that emerged in the late 1990s. Their hallmark is the use of a “black-box” external MILP solver to explore a solution neighborhood defined by invalid linear constraints. The use of an exact MILP solver inside an MILP heuristic may appear naive at first glance, but it turns out to be effective in the cases where the added invalid constraints lead to a structural simplification of the MILP at hand and allow, for example, for a more powerful instance preprocessing and/or for extensive node pruning.

Local Branching

The *local branching* (LB) scheme of Fischetti and Lodi [17] appears to be the first method

¹The half-spaces induced by these hyperplanes that contain x^0 define an octahedron, hence the name of the method.

²The name was actually inspired by the *Electron Pump* of the Asimov’s science fiction novel “The Gods Themselves”.

embedding an MILP solver within a general MILP heuristic framework. Suppose a feasible *reference solution* $\bar{x}$ of an MILP with $B \neq \emptyset$ is given, and one aims at finding an improved solution that is “not too far” from $\bar{x}$. To this end, one can define the *k-OPT neighborhood* $\mathcal{N}(\bar{x}, k)$ of $\bar{x}$ as the set of the MILP solutions satisfying the invalid *local branching constraint*

$$\Delta(x, \bar{x}) := \sum_{j \in B: \bar{x}_j = 0} x_j + \sum_{j \in B: \bar{x}_j = 1} (1 - x_j) \leq k, \quad (6)$$

for a small parameter k (typically, $k = 10$ or $k = 20$), and explore it by means of an external MILP solver, often heuristically, that is, within a prefixed number of branch-and-bound nodes. The method is in the spirit of local search metaheuristics and in particular of *large neighborhood search* [18], with the novelty that neighborhoods are obtained through a so-called *soft fixing*, that is, through invalid cuts to be added to the original MILP model. Diversification cuts can be defined in a similar way, thus leading to a flexible toolkit for the definition of metaheuristics for general MILPs. For example, Hansen *et al.* [19] embedded LB within variable neighborhood search, whereas Fischetti *et al.* [20] specialized LB to MILPs with a two-level variable structure. Integration between FP and LB was instead proposed by Fischetti and Lodi [21].

As its name suggests, LB can also be used within an exact enumerative scheme by branching on the “local” disjunction $\Delta(x, \bar{x}) \leq k$ (to be explored first) or $\Delta(x, \bar{x}) \geq k + 1$ with respect to the incumbent MILP solution $\bar{x}$ (no matter the LP solution at a given node). This branching strategy is intended to favor the detection of good solutions very early in the enumeration.

Relaxation Induced Neighborhood Search and Variants

The *relaxation induced neighborhood search* (RINS) framework of Danna *et al.* [22] also uses the (heuristic) solution of a simplified MILP through an external MILP solver as a main ingredient, but extends the idea by taking the solution of the LP relaxations into

account. At specified nodes of the branch-and-bound tree, the current LP relaxation solution x^* and the incumbent $\bar{x}$ are compared and all integer-constrained variables that agree in value are fixed and projected out of the MILP. In this way, such an MILP is not only simplified but generally speaking (provided the number of components fixed is sufficiently large) it is also reduced in size. This is probably the main difference between LB and RINS, namely the fact that the former performs a *soft* fixing by means of the cardinality constraint (6), while in the latter the fixing is *hard* with clear benefits associated with dealing with smaller MILPs.

The RINS framework is also particularly suitable for an integration within a classical branch-and-bound (and then within classical MILP solvers) because it can be virtually invoked at any node of the tree where solutions of the LP relaxations are always different. Clearly, invoking the algorithm too often would result in slowing down the overall running time for proving optimality, thus RINS is only executed every L (say) nodes in the tree. On the other hand, the size of the resulting *reduced* MILP cannot be predicted in advance and the MILP is solved only if the reduction is relevant enough.

The *distance* induced neighborhood search (DINS) approach by Ghosh [23] is a RINS variant which takes an intermediate step between LB and RINS by performing both soft and hard fixing within the same algorithm. The idea behind DINS is that the most promising solutions are those “close” to the solution of the continuous relaxation, thus the neighborhood is defined by the distance inequality

$$\sum_{j \in B \cup G} |x_j - x_j^*| \leq \sum_{j \in B \cup G} |\bar{x}_j - x_j^*|, \quad (7)$$

where again x^* and $\bar{x}$ denote the current LP solution of the LP relaxation and the incumbent, respectively. For a variable x_j such that $|\bar{x}_j - x_j^*| < 0.5$ the only way for the new solution to decrease its distance with respect to $\bar{x}_j$ is to hard fix it, that is, $x_j = \bar{x}_j$. For the other variables, instead, a *rebounding* phase is performed before the neighborhood is explored (as usual by means of a call to an

external MILP solver), in which the bounds on the variables are set so as the absolute value of the distance between the new solution and $\bar{x}$ cannot increase. A bit more of flexibility can be granted by allowing some of the variables to increase their distance but reducing the overall distance (as guaranteed by Equation 7). This is again obtained by a sophisticated mixture of hard and soft fixing.

The relaxation *enforced* neighborhood search (RENS) scheme by Berthold [24] is another RINS approach whose main characteristic is to define a reduced MILP (to be explored by a call to an external solver) as the set of all integer solutions which can be obtained by rounding a solution $x^* \in P$. More precisely, for a given x^* the reduced MILP is constructed by

1. *hard fixing* those integer-constrained variables whose current value is already integer, that is, $x_j = x_j^*, \forall j \in F$ where $F := \{j : x_j^* = \lfloor x_j^* \rfloor, j \in \mathcal{B} \cup \mathcal{G}\}$;
2. *rebounding* the remaining variables by using the two rounded values as lower ℓ_j and upper u_j bounds, that is, $\ell_j = \lfloor x_j^* \rfloor$ and $u_j = \lceil x_j^* \rceil$ for all $j \in (\mathcal{B} \cup \mathcal{G}) \setminus F$.

Interestingly, RENS does not require an incumbent solution to work with and it can be used offline as a tool for the analysis of the effectiveness of MILP rounding heuristics. Note that pivot-and-shift [5] uses a similar mechanism to restrict the range of the integer variables depending on their LP value and such a rebounding is similar to that of DINS as well.

An Evolutionary Algorithm within MILP

The farthest (to date) step in the direction of using metaheuristic ideas within an MILP has been taken by Rothberg [25]. The resulting approach, often called *polishing* algorithm, implements an evolutionary heuristic which is invoked at selected nodes of a branch-and-bound tree and includes all classical ingredients of genetic computation. More precisely,

- *Population*. A fixed-size population of feasible solutions is maintained. Those

solutions are either obtained within the tree (by other heuristics, including diving) or computed by the polishing algorithm itself.

- *Combination*. Two or more solutions (the parents) are combined with the aim of creating a new member of the population (the child) with “improved” characteristics. The RINS scheme is adopted, that is, all variables whose value coincides in the parents are fixed and the reduced MILP is heuristically solved by an external MILP solver within a limited number of branch-and-bound nodes. This scheme, clearly, is much more time-consuming than a classical combination step in evolutionary algorithms, but guarantees that the child solution is feasible (for the original problem).
- *Mutation*. A sufficient degree of diversification is guaranteed by performing a classical mutation action which consists in (i) selecting at random a *seed* solution in the population, (ii) fixing at random some of its variables, and (iii) heuristically solve the resulting reduced MILP.
- *Selection*. The parents’ selection is performed in a simple way by randomly picking a solution in the population and then choosing, again at random, the second one but only among those solutions with a better objective value. Although it is easy to extend this criterion to select more than two parents, in Rothberg [25] either two or all solutions are used as parents, that is, in the latter case no selection is performed.

The polishing algorithm is then a fully general metaheuristic implemented within an MILP framework and, in turn, making use of an external MILP solver as main “engine.”

REFERENCES

1. Achterberg T. Constraint integer programming. PhD thesis, Berlin: ZIB; 2007.
2. Balas E, Martin CH. Pivot-and-complement: a Heuristic for 0-1 programming. *Manage Sci* 1980;26:86–96.

3. Løkketangen A, Jornsten K, Storoy S. Tabu search within a Pivot and complement framework. *Int Trans Oper Res* 1994;1:305–317.
4. Balas E, Martin CH. Pivot and shift: a heuristic for mixed integer programming. Technical report, GSIA, Carnegie Mellon University; 1986.
5. Balas E, Schmieta S, Wallace C. Pivot and shift - a mixed integer programming Heuristic. *Disc Optim* 2004;1:3–12.
6. Nediak M, Eckstein J. Pivot, cut, and dive: a Heuristic for 0-1 mixed integer programming. *J Heuristics* 2007;13:471–503.
7. Ghosh S, Hayward RB. Pivot and Gomory cut: a MILP feasibility Heuristic. Technical report, University of Alberta; 2009.
8. Hillier FS. Efficient Heuristic procedures for integer linear programming with an interior. *Oper Res* 1969;17:600–637.
9. Ibaraki T, Ohashi T, Mine H. A Heuristic algorithm for mixed-integer programming problems. *Math Program Study* 1974;2:115–136.
10. Faaland BH, Hillier FS. Interior path methods for heuristic integer programming procedures. *Oper Res* 1979;27:1069–1087.
11. Naoum-Sawaya J, Elhedhli S. An interior point cutting plane Heuristic for mixed integer programming. Technical report, University of Waterloo; 2010.
12. Balas E, Ceria S, Dawande M, *et al.* OCTANE: a new Heuristic for pure 0-1 programs. *Oper Res* 2001;49:207–225.
13. Fischetti M, Glover F, Lodi A. The feasibility pump. *Math Program* 2005;104:91–104.
14. Bertacco L, Fischetti M, Lodi A. A feasibility pump Heuristic for general mixed-integer problems. *Disc Optim* 2007;4:63–76.
15. Achterberg T, Berthold T. Improving the feasibility pump. *Disc Optim* 2007;4:77–86.
16. Fischetti M, Salvagnin D. Feasibility pump 2.0. *Math Program Comput* 2009;1:201–222.
17. Fischetti M, Lodi A. Local branching. *Math Program* 2003;98:23–47.
18. Shaw P. Using constraint programming and local search methods to solve vehicle routing problems. In: Maher M, Puget JF, editors. *Principles and practice of constraint programming CP98*. Volume 1520, Lecture notes in computer science. Heidelberg: Springer; 1998. pp. 417–431.
19. Hansen P, Mladenović N, Urošević D. Variable neighborhood search and local branching. *Comput Oper Res* 2006;33:3034–3045.
20. Fischetti M, Polo C, Scantamburlo M. A local branching Heuristic for mixed-integer programs with 2-level variables. *Networks* 2004;44:61–72.
21. Fischetti M, Lodi A. Repairing MILP infeasibility through local branching. *Comput Oper Res* 2008;35:1436–1445.
22. Danna E, Rothberg E, Le Pape C. Exploring relaxation induced neighborhoods to improve MILP solutions. *Math Program* 2005;102:71–90.
23. Ghosh S. DINS, a MILP improvement Heuristic. In: Fischetti M, Williamson DP, editors. *Integer programming and combinatorial optimization, IPCO 2007*. Volume 4513, Lecture notes in computer science. Springer; 2007. pp. 310–323.
24. Berthold T. RENS - relaxation enforced neighborhood search. Technical Report 07-28. Berlin: ZIB; 2007.
25. Rothberg E. An evolutionary algorithm for polishing mixed integer programming solutions. *Inform J Comput* 2007;19:534–541.

HISTORY OF CONSTRAINT PROGRAMMING

ROMAN BARTÁK
Department of Theoretical
Computer Science and
Mathematical Logic, Faculty of
Mathematics and Physics,
Charles University, Prague,
Czech Republic

Constraint programming (CP) is a technology for solving combinatorial optimization problems in areas such as scheduling, configuration, vehicle routing, networking, and bioinformatics. It can be seen as a field on the edge between “mathematical programming” and “computer programming.” In “mathematical programming,” the users declaratively specify the problem to be solved, and the underlying solution mechanism generates a solution. In “computer programming,” the users—the programmers—have to explicitly instruct the computer how to produce the solution. In CP also, the users declaratively specify the problems—this is called *constraint modeling* and the problems are formulated as constraint satisfaction problems (CSP), but the users can also help the solution mechanism, for example, by programming a specific strategy to search for a solution. Constraint modeling is more flexible and closer to formal problem specification than, for example, the problem formulation in mathematical programming. Hence, CP is sometimes seen as *one of the closest approaches computer science has yet made to the Holy Grail of programming: the user states the problem; the computer solves it* [1].

The formulation of a CSP is extremely simple—the problem consists of a finite set X of variables, each variable $x_i \in X$ has assigned a set D_i of possible values that can be used for instantiation of x_i —its domain—and finally there is a finite set of constraints that restrict possible combinations of values assigned to the variables. The constraint on a set of variables can

be expressed in extension as a set of value tuples that can be consistently assigned to these variables or used as an arithmetical or a logical formula. More formally, the constraint is any relation over the set of variables—it is defined as a subset of the Cartesian product of domains of the constrained variables. A solution to a CSP is a complete instantiation of variables satisfying all constraints. A nice example of a CSP is the well-known Sudoku puzzle, where the variables correspond to the cells, each domain consists of numbers 1 to 9, and there is a constraint forcing inequality between any pair of variables corresponding to the cells in the same row, in the same column, and in the same subgrid. Another well-known example is the graph-coloring problem, that can be directly encoded as a CSP (the nodes specify the variables; the colors are possible values for the variables, and the arcs define binary constraints forcing inequality).

There are two important special classes of CSPs: Boolean CSPs, where the domains contain only 0 and 1 and the constraints describe logical relations, and binary CSPs, where all constraints are binary relations. While Boolean CSPs naturally model satisfiability problems and hence show that CSPs, in general, are NP-complete problems (though there are tractable classes as described further), binary CSPs were the topic of initial studies in the area of constraint satisfaction, and many constraint satisfaction algorithms were initially formulated and studied for binary constraints. There also exist extensions of classical CSPs, for example, soft CSPs and constraint optimization problems (COP). In soft CSPs, some constraints can be relaxed and the task is to satisfy as many constraints as possible (other formulations also exist [2]), while in a COP a given objective function evaluates the solutions of the underlying CSP and the task is to minimize or maximize the value of such a function.

The solving technology behind CP originates mostly in artificial intelligence, but nowadays it also absorbs techniques from

fields such as databases, discrete mathematics, numerical mathematics, and operations research. This article gives a brief overview of how constraint satisfaction technology evolved. We focus mainly on constraint satisfaction over discrete finite domains, which forms the vast majority of research in the area of constraint satisfaction. There are also studies of usually more complex problems over real numbers and other domains.

THE ANCIENT TIMES

Though frequently (and wrongly) CP is used as a synonym for enumeration or search, the truth is that a specific form of inference called *constraint propagation* is much more typical to CP. Nevertheless, taking search as one of the core technologies behind CP, we can trace the roots of CP as back as to ancient Greeks, namely to the myth on Theseus who used backtrack-like search to escape from the Crete's Labyrinth inhabited by a Minotaur [3]. Backtrack search was used in the nineteenth century to solve problems in recreational mathematics [4], and it was an early subject of study in fields such as operations research. The backtracking algorithm was informally described by Golomb and Baumert [5], but a nonrecursive formulation of Bitner and Reingold [6] is used more frequently.

As mentioned above, inference techniques are more typical to CP. Again, going to ancient Greeks, Diophantine equations named after the Greek mathematician Diophantus can be seen as a special form of a CSP [7]. These equational constraints with integer domains were solved by the Indian mathematician Brahmagupta in the seventh century [8]. Probably more known works in this area are results by Gauss, who studied systematic methods for solving linear equations by variable elimination [9], and the works by Fourier who solved linear inequality constraints [10]. The ideas of Gauss and Fourier are reflected in some modern approaches of constraint solving [11].

THE BEGINNING

Artificial intelligence has its birth year in 1956, when this name for a new research

field was proposed by McCarthy at the Dartmouth workshop [12]. The birth year of CP can be set in 1963, when Ivan Sutherland in his MIT PhD Thesis described the system Sketchpad—an interactive graphics application that allowed the user to draw and manipulate constrained geometric figures on computer display [13]. This ancestor of modern drawing programs and CAD tools had a follower in the system ThingLab by Alan Borning [14], where preferential constraints can be precompiled to resatisfy them quickly after changing values of some variables representing positions of a computer mouse. Yes, the computer mouse existed before Apple Macintosh appeared in 1984. The common idea of these systems is that rather than writing ad hoc procedures for the manipulation of graphical objects, the user declares the constraints about these objects and the system automatically maintains these constraints while the user manipulates any object. Since Sketchpad in 1963, two parallel streams of CP developed, that Freuder and Mackworth [15] call the “language” stream and the “algorithm” stream.

The language stream represents the effort on developing novel programming languages and systems motivated by applications in various domains. For example, the research in the field of electric circuit analysis and design at MIT led to the language CONSTRAINTS [16]. Fikes [17] described REF-ARF, which was a part of a general problem-solving system that employed constraint satisfaction as one of its parts. Another notable language was Alice by Lauriere [18]. This purely declarative language used high level objects such as functions and relations to model combinatorial problems. Constraint solving was handled by consistency methods as well as by reasoning about abstract functions. Other systems and languages that include some form of constraint reasoning and that were developed in this period are Absys—a declarative language based on manipulation of equational constraints [19], MOLGEN—an expert system for problems in the area of molecular genetics [20], and PLANNER—a language for proving theorems in robots [21]. From the current perspective, the real landmark in the

area of languages for constraint satisfaction was the development of the programming language Prolog by Alain Colmerauer and Philippe Roussel [22]. Prolog (programmation en logique—French for programming in logic) is a logic programming language that was originally developed for the area of computation linguistics, and is based on Robert Kowalski's procedural interpretation of Horn clauses [23]. Prolog can be seen as an early CP language, where the equality constraints over the terms are solved by the unification algorithm. Colmerauer further developed this idea in Prolog II that supported equations and inequalities over the rational trees [24]. Prolog II can be seen as the first logic programming language that explicitly used constraints. The successor—Prolog III—is a real constraint logic programming language providing the constraints over the Booleans, linear arithmetic over the rational numbers, and the constraints over the lists [25]. It was used in several application areas such as chemical reasoning, diagnosis problems, and scheduling problems.

The algorithm stream was another constraint-related stream in artificial intelligence in the 1970s. It was motivated by the research in the area of machine vision, and the first landmark was the development of filtering algorithms for scene labeling by David Waltz [26]. Scene labeling was one of the first problems formulated as a CSP—the lines in the scene define the variables, each line is to be annotated either as convex, concave or boundary (these are the values in the domain), and the junctions of the lines define the constraints describing physically feasible combinations of edge types. Waltz suggested an algorithm for pruning the infeasible values, which is called *domain filtering*. This constraint propagation (arc consistency) algorithm is today known as AC-2, and it started the research in the area of constraint consistency techniques. Ugo Montanari proposed the notion of path consistency and established the general framework for reasoning with constraints [27]. The notions of arc and path consistency are derived from the graph representation of a binary CSP, where the nodes describe the variables and the arcs describe the

binary constraints. In arc consistency, we require that each value in the domain of a variable has a compatible value (support) in the domain of a connected variable, where compatibility means that the pair of values satisfies all constraints between these two variables. In path consistency, we require that two values from the domains of two variables have a common support in the domain of any third variable. Alan Mackworth described the basic arc and path consistency algorithms AC-1, AC-2, AC-3, PC-1, and PC-2 [28] and together with Eugene Freuder studied their complexity [29]. The idea of AC-3 is still in the core of many current constraint solvers. Mackworth also extended the definitions and the consistency algorithms to nonbinary constraints [30], while Freuder generalized arc and path consistency to k -consistency [31] and later to (i, j) consistency [32]. In k -consistency, we require that any consistent instantiation of $k - 1$ different variables can be consistently extended to any additional variable, while (i, j) -consistency generalizes this idea by requiring any consistent instantiation of i different variables to be consistently extendable to any additional j variables.

THE REAL START

The research in the 1970s prepared a base for expansion of constraint satisfaction techniques and for joining the language and the algorithm streams. As already mentioned, the development of Prolog was one of the landmarks in the area of constraint satisfaction. Researchers realized that unification in Prolog is just a special form of constraint reasoning, and this observation started the systematic use of constraints in programming. Constraint logic programming (CLP) was born [33]. There were three independent research teams working on concepts of CLP. We already mentioned Alain Colmerauer and his Prolog II and Prolog III systems. The actual notion of CLP was proposed by Jaffar and Lassez in 1986 [33]. They gave a general schema and semantics for the class of CLP languages, and they developed the language CLP(R) which was an extension of Prolog with the arithmetic constraints over the real

numbers [34]. They used an incremental simplex algorithm for solving the linear constraints, while the evaluation of the nonlinear constraints was delayed until they became linear or sufficiently ground. This system was used in financial modeling and in several engineering applications. The real join of the algorithm and the language streams can, however, be seen in the third system. The team led by Mehmet Dincbas with notable contributions from Pascal Van Hentenryck, Helmut Simonis, and others in the European Computer-Industry Research Center (ECRC) developed the system CHIP (constraint handling in Prolog) [35]. This system solved the combinatorial problems, and it combined Prolog's backtracking search with the consistency techniques. It was the first CLP language focusing on constraint satisfaction over the finite domains. The main application areas were scheduling, circuit diagnosis, and cutting stock problems. The architecture of CHIP, the combination of its backtracking search and consistency techniques, can be seen as the mainstream constraint satisfaction approach. Many later systems employed the same schema. The first book on CP [35], which described the principles of CHIP, was published at that time.

DEVELOPING THE CORE CONCEPTS

When the base solving approach for CP, the combination of its search techniques with inference techniques based on constraint propagation, was established, the research focused on both these areas independently as well as on their efficient integration.

The research on search techniques followed the work on backtracking search by providing several improvements. One branch focused on improving the inefficiencies of chronological backtracking when going back during search. *Backjumping* was introduced by Gaschnig [36] to recover from a known problem of backtracking called *thrashing* (when going back, the search algorithm does not exploit information about the reason of the failure). On the other hand, backjumping can jump to the variable causing the conflict during the search rather than trying different instantiations of the variables that are

not related to the conflict. A graph-based variant of backjumping that exploits the structure of the constraint graph and, hence, can jump back several times was proposed by Dechter [37]. Prosser then combined both ideas and suggested conflict-directed backjumping [38]. One of the disadvantages of backjumping is that it forgets the assignment of variables that are jumped across. *Dynamic backtracking* proposed by Ginsberg [39] solves this problem by dynamic variable reordering during search. Another approach for improving chronological backtracking was based on remembering the result of constraint checks, so the next time the same check is done, the result can be retrieved and used. This technique for removing redundant constraint checks called *backmarking* was proposed by Haralick and Elliot [40]. Together with other techniques that, for example, record no-goods discovered during search, these improvements of chronological backtracking are known under the umbrella term *intelligent backtracking* or *look-back* search techniques.

Other efforts focused on exploiting heuristics during search. When instantiating a variable, there are basically two types of heuristics necessary to guide the search: the variable ordering heuristics to select the variable for instantiation, and the value ordering heuristics to select the value to be assigned to the variable. Already Golomb and Baumert proposed variable ordering heuristic, suggesting to instantiate first the variable with the smallest domain [5]. This is called the *dom heuristic* today. Haralick and Elliot [40] later generalized this principle and give it a name—a *fail-first principle* (select the variable whose instantiation will lead to a failure with the highest chance). Motivated by the graph-coloring problems, Brélaz [41] suggested combining the dom heuristic with information about the number of constraints in which the variable participates (the variables participating in more constraints are tried first). This is called the *dom+deg heuristic* (select the variable with the smallest domain and break ties by preferring the variable participating in more constraints). Another variation of this heuristic called *dom/deg* was suggested by

Bessiere and Regin [42]. All these heuristics try to exploit the graphical structure of a CSP. There is also an “independent” research branch studying how the structure of CSPs can be exploited during problem solving. Already Freuder [43] showed that ordering the variables based on the width of the constraint graph can lead to solving the problem without any backtrack. The variable ordering giving the smallest width of the constraint graph is used, where the width is the maximal number of backward edges in the ordered graph. In particular, Freuder proved that if the problem is k -consistent for any k between 1 and $w + 1$, where w is the width of the graph, then the above ordering leads to backtrack-free search. Dechter and Pearl [44] suggested another structure-based heuristic that instantiates first the variables that cut cycles in the constraint graph. The motivation is hidden in the observation that tree-structured CSPs can be solved in polynomial time [43], and this heuristic tries to reduce the problem into a tree-structured one. These structure-based heuristics are not frequently used today because they can break down in the presence of global constraints.

While the generally applicable variable ordering heuristics are widely used, the value ordering heuristics are more frequently problem dependent. The values are selected based on their meaning in the problem. Nevertheless, the problem-independent value ordering heuristics were also proposed and studied. Dechter and Pearl [44] suggested approximating the number of solutions by solving a tree relaxation of the problem and preferring the values that lead to more solutions in such tree relaxations. An easier way is to assign the value to the variable, propagate this instantiation to other variables through the constraints (e.g., by making the problem arc consistent), and then to select the variable based on the size of the resulting domains. Ginsberg *et al.* [45] proposed to select first the value that maximizes the product of the remaining domain sizes—a so-called *promise heuristic*. Frost and Dechter [46] suggested selecting first the value that maximizes the sum of the remaining domain sizes—a so-called *min-conflict heuristic*.

While the heuristics drive the search algorithm toward the solution, they cannot guarantee to be always correct, that is, to lead to the solution without any backtrack. Hence a research branch studying how to recover from the failure of the heuristics appeared. It was pioneered by Harvey and Ginsberg [47], who developed a so-called *limited discrepancy search* motivated by two features of good value ordering heuristics. First, the number of times the heuristic is wrong—a so-called discrepancy—is usually small. Second, if there are any discrepancies, then they appear earlier in the search tree. Other discrepancy-based search techniques were proposed later; for example, an improved limited discrepancy search by Korf [48], the discrepancy-bounded depth-first search by Beck and Perron [49], the interleaved depth-first search by Meseguer [50], and the depth-bounded discrepancy search by Walsh [51].

Another research branch focused on consistency techniques. The original arc and path consistency algorithms proposed in Mackworth [28] were not optimal regarding the time efficiency. Mohr and Henderson [52] proposed an optimal worst-case time complexity algorithm AC-4 to achieve arc consistency. This algorithm was based on the idea of keeping all supports for each value and removing that value from the domain when it lost all supports. The main disadvantage of this algorithm was large memory consumption (to remember all supports) and also the feature that the average time complexity was close to the worst case, so the nonoptimal AC-3 was better in average. Hence, Bessiere [53] proposed an improvement called AC-6 based on the idea that it is enough to keep a single support for each value and when the support is lost, the algorithm tries to find another one (if not possible, then the value is removed from the domain). Further improvement led to AC-7 [54] that exploits bidirectionality of constraints; that is, each value is a support for its own support. While AC-6 and AC-7 are based on the original idea of AC-4, the optimal versions of AC-3 algorithm were also proposed at an IJCAI 2001 conference. Zhang and Yap proposed

the algorithm AC-3.1 [55], and Bessiere and Regin proposed the algorithm AC-2001 [56]. Both algorithms combined the idea of AC-6 with the structure of AC-3. In the above listing of AC algorithms, we intentionally skipped algorithm AC-5 [57], which was going in a different direction. AC-5 is a general schema for AC algorithms, and depending on which subroutines are used, it can coincide with AC-3, AC-4, or other concrete algorithm for arc consistency. Moreover, this schema covers special filtering techniques for functional ($=$), antifunctional ($\neq$), and monotonic ($<$) constraints showing that the semantics of the constraint can be exploited during domain filtering. The concept of global constraints that heavily exploit the semantics of the constraint during domain filtering is discussed later.

Today, arc consistency is the most prominent inference technique in CP, but other consistency techniques were also studied. As mentioned, the notion of path consistency was introduced by Montanari [27], and Mackworth proposed the first path consistency algorithms PC-1 and PC-2 [28]. Mohr and Henderson [52] tried to develop a more efficient version of the path consistency algorithm using the same idea as in AC-4. This algorithm, called PC-3, is not sound as it may filter out a value that is consistent. Two years later, Han and Lee corrected this algorithm and called the new algorithm PC-4 [58]. The research of path consistency algorithms closely followed the novel ideas developed for the arc consistency algorithms, and Singh proposed the algorithm PC-5 [59] exploiting the idea of AC-6. At the same time, Berlandier also proposed a novel consistency level between arc and path consistency [60]. He called it *restricted path consistency* and the algorithm integrated some ideas of path consistency into the AC-4 algorithm. For a small overhead, this algorithm can remove more inconsistent values than arc consistency without the deficiencies of path consistency such as large memory consumption and changing the structure of the constraint graph. Other consistency levels were also studied, namely inverse consistency [61], which is basically (i, j) -consistency, where $i = 1$; neighbourhood

inverse consistency [62], where a value of a variable is required to be consistently extendable to neighbourhood variables; and singleton consistency [63] that can strengthen other consistency levels. For example, singleton arc consistency requires that the problem can be made arc consistent after instantiating any variable.

The early research on constraint satisfaction focused on binary CSPs and many of the above described results were formulated for binary constraints. Nevertheless, the constraints in practical applications frequently have an arity greater than two. Fortunately, n -ary constraints can be converted to the equivalent set of binary constraints. It was Montanari, who suggested the first conversion of an n -ary constraint to a set of binary constraints by projection [27]. However, the obtained set of binary constraints is not equivalent to the original n -ary constraint (some value tuples forbidden by the n -ary constraint may be allowed by the set of binary constraints). The equivalent encodings of n -ary constraints as binary CSPs were proposed and studied by Bacchus and van Beek [64] and Rossi *et al.* [65]. From the other perspective, Mackworth extended the definitions and the consistency algorithms to nonbinary constraints [30], which opened the area for constraint satisfaction algorithms for n -ary constraints. It is possible to directly extend the concept of arc consistency to n -ary constraints, which is called *generalized arc consistency*, or sometimes hyper arc consistency or domain consistency. The critical point for the efficiency of these algorithms is how to realize domain filtering for n -ary constraints. One can generalize the techniques from the binary constraints, but it is usually more efficient to exploit the semantics of the constraint as first suggested in AC-5 [57]. This is when the so-called global constraints were born. A *global constraint* is an n -ary constraint, where the arity is not fixed (the filtering algorithm can work with any arity). Probably the most well-known example of global constraint is *all-different* that encapsulates a set of binary inequality constraints. An efficient filtering algorithm for this constraint was proposed by Régin [66]. Since that time, many other

global constraints, frequently motivated by some application, have been proposed. Régim [67] generalized the all-different constraint into the *global cardinality constraint* (each value appears a given number of times in the set of variables). Pesant proposed the *regular constraint* (the sequence of values assigned to variables forms a word that is accepted by a finite automaton) [68], but the idea of a filtering algorithm was already presented by Barták [69]. The *grammar constraint* uses the same motivation, but the sequence of values is given by a grammar rather than by an automaton [70]. It was the CP system CHIP (a follower of the original CHIP from ECRC) marketed by Cosytec that popularized the idea of global constraints. Beldiceanu *et al.* attempted to classify the global constraints using the formalism of graph invariants [71].

So far we have discussed two separate streams of research that we can call the search stream and the consistency (or inference) stream. Both of them are integrated in a typical CP system. We already mentioned the look-back schemes, where the constraints are used mainly to check whether or not the partial instantiation of the variables satisfies the constraints. The inference techniques can instead be used over the not-yet instantiated variables to detect a possible future failure of the search algorithm. This is called a *look ahead scheme*. The simplest way is propagating information about the currently instantiated variable to its direct neighbors. This is called *forward checking* and it was proposed by McGregor for binary constraints [72], and extended to the n -ary constraints by van Hentenryck [35]. One can propagate the instantiation further to even more distant variables but still in the direction from the currently instantiated variable to the not-yet instantiated variables. This is called *partial look ahead*, and it was proposed by Haralick and Elliot [40]. Even before, Gaschnig suggested maintaining arc consistency during search. That is, after instantiating the variable, the problem is made arc consistent and, if this is not possible, then backtracking is forced. He called this technique *full look ahead* [73]. The very same idea was explored under the name *maintaining arc consistency* (MAC) in Sabin and Freuder [74]. For a

long time, researchers were convinced that partial look ahead was more efficient than full look ahead, which pruned more of the search space but also added more overhead. It was Sabin and Freuder [74] who showed that for the difficult problems MAC solves the problems much faster than partial look ahead. The MAC algorithm forms the most frequently used solution approach in modern CP systems. The recent trend in integration of search and inference goes toward randomized versions of backtracking search with random restarts, where the information from one search iteration can be exploited during the inference in subsequent iterations [75].

COMMUNITY ASPECTS

So far we focused on the research aspects in the history of CP, but it is also interesting to see how the community evolved. At the beginning, there were only a few people developing this new solution technology, mainly under the umbrella of artificial intelligence. We gave names of some of them in the above paragraphs. It was in 1993 when the first workshop on Principles and Practice of Constraint Programming was organized at Newport, Rhode Island, United States. After the second workshop in 1994, the workshop changed to a conference with formal proceedings published in the Springer LNCS series. The first international conference on Principles and Practice of Constraint Programming was organized in 1995, in Cassis, France. This year can be seen as the year when the CP community was formally born. This started the series of annual conferences known under the acronym CP that is a prominent forum, where researchers working in the area of CP meet and discuss novel developments. At the end of the twentieth century, there was a series of sister conferences on practical applications of CP called PACT (The Practical Application of Constraint Technology), but it disappeared after a few editions. As the scope of CP was broadening and the originally AI-based technology started to absorb results from other research areas, namely operations research, another CP-related conference appeared. After a series of successful

workshops, in 2004, the first international conference on Integration of AI and OR Techniques in Constraint Programming (CP-AI-OR) was held in Nice, France. Since that time, the CP-AI-OR conference is organized annually and complements the CP conference by focusing on integration of techniques from different fields. In addition to these two major CP-related conferences, there is a special publication forum for researchers in the area of CP. In 1996, the Constraints journal (Kluwer) was started with Eugene Freuder as the first Editor-In-Chief. This journal “provides a common forum for the many disciplines involved in CP, constraint satisfaction, and optimization, and the many application domains in which constraint technology is employed” [76].

The CP community is now governed by the Association for Constraint Programming. This nonprofit international organization was formally established in 2005, and it undertook the role of the Constraint Programming Organizing Committee that started with the first CP conference in 1995 and coordinated the activities around the CP conference. “The Association for Constraint Programming aims at promoting CP in every aspect of the scientific world, by encouraging its theoretical and practical developments, its teaching in the academic institutions, its adoption in the industrial world, and its use in the application fields” [77]. In addition to taking care of the CP conferences, the association supports other CP-related activities such as summer schools.

The evolution of CP can also be nicely observed in published books. We already mentioned the pioneering book *Constraint Satisfaction in Logic Programming* by Pascal Van Hentenryck published in 1989 [35]. The book shows how the CP started, and the ideas from this book are still used in modern CP systems. The first book that can be assumed as a textbook on CP is *Foundations of Constraint Satisfaction* by Edward Tsang published in 1993 [78]. Prior to this book, the material on constraint satisfaction was scattered and the lack of organization made it hard to study. *Foundations of Constraint Satisfaction* consolidated

the material in the way that is appropriate for teaching and study of this area. In 1998, a new book following the style of *Constraint Satisfaction of Logic Programming* was published—*Programming with Constraints: An Introduction* by Marriott and Stuckey [79]. This book is targeted at practitioners of constraint satisfaction technology. The reader is guided using examples through problem modeling, using appropriate data structures, and controlling search. The book represents the language stream of research. In 2003, three new books on CP were published. The book *Essentials of Constraint Programming* by Abdennadher and Frühwirth [80] presents constraint handling rules as the universal language for writing constraint solvers. The book *Principles of Constraint Programming* by Apt [81] brings a novel unifying view of CP, where the constraint solvers and local consistency are presented by means of rules and the constraint propagation algorithms by means of the generic iteration algorithms. Finally, the book *Constraint Processing* by Dechter [82] is a good textbook covering basics of constraint processing as well as advanced constraint satisfaction methods. Among the other books on CP published later, *The Handbook of Constraint Programming* [83] must be mentioned as it provides a comprehensive survey of the area of CP written in the encyclopedic style. It was also a valuable source of information for this article.

REFERENCES

1. Freuder EC. In pursuit of the holy grail. *Constraints* 1997;2(1):57–61.
2. Barták R. Modelling soft constraints: a survey. *Neural Netw World* 2002;12(5):421–431.
3. Wikipedia. Theseus. Available at <http://en.wikipedia.org/wiki/Theseus>.
4. Lucas E. *Récréations mathématiques*. Paris: Gauthier-Villars; 1891.
5. Golomb S, Baumert L. Backtrack programming. *J ACM* 1965;12:516–524.
6. Bitner JR, Reingold EM. Backtracking programming techniques. *Commun ACM* 1975; 18(11):651–656.

7. Wikipedia. Diophantus. <http://en.wikipedia.org/wiki/Diophantus>.
8. Wikipedia. Brahmagupta. <http://en.wikipedia.org/wiki/Brahmagupta>.
9. Gauss CF. *Disquisitiones arithmeticae*. 1801.
10. Schrijver A. *Theory of linear and integer programming*. Chichester: John Wiley & Sons; 1998.
11. Harvey W, Stuckey PJ, Borning A. Compiling constraint solving using projection. *Proceedings of Principles and Practice of Constraint Programming—CP97*; 1997 Oct 29–Nov 1; Linz, Austria. 1997. pp. 491–505.
12. McCarthy J, Minsky M, Rochester N, *et al.* A proposal for the Dartmouth summer research project on artificial intelligence. *The Dartmouth Summer Research Conference on Artificial Intelligence*. Dartmouth; 1955.
13. Sutherland IE. SKETCHPAD: A man-machine graphical communications system. *Proceedings of the Spring Joint Computer Conference*; 1963 May 21–23; Detroit, MI. 1963. pp. 329–346.
14. Borning A. The programming language aspects of ThingLab, a constraint-oriented simulation laboratory. *ACM Trans Program Lang Syst* 1981;3(4):252–387.
15. Freuder EC, Mackworth AK. Constraint satisfaction: an emerging paradigm. In: Rossi F, Van Beek P, Walsh T, editors. *Handbook of constraint programming*. Amsterdam: Elsevier; 2006. pp. 13–27.
16. Sussman GJ, Steele GL. CONSTRAINTS: a language for expressing almost hierarchical descriptions. *Artif Intell* 1980;14(1):1–39.
17. Fikes RE. REF-ARF: a system for solving problems stated as procedures. *Art Intell J* 1970;1(1):27–120.
18. Lauriere JL. A language and a program for stating and solving combinatorial problems. *Art Intell* 1978;10:29–127.
19. Foster A, Elcock EW. Absys 1: an incremental compiler for assertions: an introduction. *Proceedings of the Fourth Annual Machine Intelligence Workshop*. Volume 4, Machine Intelligence. Edinburgh: Edinburgh University Press; 1969.
20. Stefik M. Planning with constraints (MOLGEN: Part 1). *Artif Intell J* 1981;16:111–140.
21. Hewitt C. PLANNER: a language for proving theorems in robots. *Proceedings of the First International Joint Conference on Artificial Intelligence*; 1969 May 7–9; Washington, DC. 1969. pp. 295–301.
22. Colmerauer A, Roussel P. The birth of Prolog. *Proceedings of the Conference on History of Programming Languages*; 1993 Apr 20–23; Cambridge, MA. 1993. pp. 37–52.
23. Kowalski RA. Predicate logic as programming language. *Proceedings of IFIP Congress*; 1974 Aug 5–10; Stockholm, Sweden. 1974. pp. 569–574.
24. Colmerauer A. Equations and inequations on finite and infinite trees. *Proceedings of the International Conference on Fifth Generation Computer Systems*; 1984 Nov 6–9; Tokyo, Japan. 1984. pp. 85–99.
25. Colmerauer A. Opening the Prolog III universe. *BYTE Mag* 1987; 12(9):177–182.
26. Waltz DL. Understanding line drawings of scenes with shadows. In: Winston PH, editor. *Psychology of computer vision*. New York: McGraw-Hill; 1975.
27. Montanari U. Networks of constraints fundamental properties and applications to picture processing. *Inf Sci* 1974;7:95–132.
28. Mackworth AK. Consistency in networks of relations. *Artif Intell* 1977;8:99–118.
29. Mackworth AK, Freuder EC. The complexity of some polynomial network consistency algorithms for constraint satisfaction problems. *Artif Intell* 1985;25:65–74.
30. Mackworth AK. On reading sketch maps. *Proceedings IJCAI* 1977; 1977 Aug; Cambridge, MA. 1977. pp. 598–606.
31. Freuder EC. Synthesising constraint expressions. *Commun ACM* 1978;21(11):958–966.
32. Freuder EC. A sufficient condition for backtrack-bounded search. *J ACM* 1985;32(4):755–761.
33. Jaffar J, Lassez JL. Constraint logic programming. *Proceedings of the 14th ACM Symposium on Principles of Programming Languages*; 1987 Jan 21–23; Munich, Germany. 1987. pp. 111–119.
34. Jaffar J, Michaylov S, Stuckey PJ, *et al.* The CLP(R) language and system. *TOPLAS* 1992;14(3):339–395.
35. Van Hentenryck P. *Constraint satisfaction in logic programming*. Cambridge (MA): The MIT Press; 1989.
36. Gaschnig J. Performance measurement and analysis of certain search algorithms. *Technical Report CMU-CS-79-124*. Pittsburgh (PA): Carnegie-Mellon University; 1979.
37. Dechter R. Enhancement schemes for constraint processing: backjumping, learning,

- and cutset decomposition. *Artif Intell* 1990; 41:273–312.
38. Prosser P. Hybrid algorithms for constraint satisfaction problems. *Comput Intell* 1993;9(3):268–299.
 39. Ginsberg ML. Dynamic backtracking. *J Artif Intell Res* 1993;1:25–46.
 40. Haralick RM, Elliot GL. Increasing tree search efficiency for constraint satisfaction problems. *Artif Intell* 1980;14:263–314.
 41. Brélaz D. New methods to color the vertices of a graph. *Commun ACM* 1979;22:251–256.
 42. Bessiere C, Régin J-C. MAC and combined heuristics: two reasons for forsake FC (and CBJ?) on hard problems. *Proceedings of the Second International Conference on Principles and Practice of Constraint programming (CP)*; 1996 Aug 19–22; Cambridge, MA. 1996. pp. 61–75.
 43. Freuder EC. A sufficient condition for backtrack-free search. *J ACM* 1982;29:24–32.
 44. Dechter R, Pearl J. Network-based heuristics for constraint satisfaction problems. *Artif Intell* 1988;34:1–38.
 45. Ginsberg ML, Frank M, Halpin MP, *et al.* Search lessons learned from crossword puzzles. *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI)*; 1990 Jul 29–Aug 3; Boston, MA. 1990. pp. 210–215.
 46. Frost D, Dechter R. Look-ahead value ordering for constraint satisfaction problems. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*; 1995 Aug 20–25; Montréal, Québec, Canada. 1995. pp. 572–578.
 47. Harvey WD, Ginsberg ML. Limited discrepancy search. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*; 1995 Aug 20–25; Montréal, Québec, Canada. 1995. pp. 607–613.
 48. Korf RE. Improved limited discrepancy search. *Proceedings of National Conference on Artificial Intelligence (AAAI-96)*; 1996 Aug 4–8; Portland, OR. 1996. pp. 286–291.
 49. Beck JC, Perron L. Discrepancy-bounded depth first search. *Proceedings of CP-AI-OR*; 2000 Mar 8–10; Paderborn, Germany. 2000. pp. 7–17.
 50. Meseguer P. Interleaved depth-first search. *Proceedings of 15th International Joint Conference on Artificial Intelligence*; 1997 Aug 23–29; Nagoya, Japan. 1997. pp. 1382–1387.
 51. Walsh T. Depth-bounded discrepancy search. *Proceedings of 15th International Joint Conference on Artificial Intelligence*; 1997 Aug 23–29; Nagoya, Japan. 1997. pp. 1388–1393.
 52. Mohr R, Henderson TC. Arc and path consistency revised. *Artif Intell* 1986;28:225–233.
 53. Bessiere C. Arc-consistency and arc-consistency again. *Artif Intell* 1994;65: 179–190.
 54. Bessiere C, Freuder EC, Régin J-R. Using constraint metaknowledge to reduce arc consistency computation. *Artif Intell* 1999;107: 125–148.
 55. Zhang Y, Yap R. Making AC-3 an optimal algorithm. *Proceedings of IJCAI-01*; 2001 Aug 4–10; Seattle, WA. 2001. pp. 316–321.
 56. Bessiere C, Régin J-C. Refining the basic constraint propagation algorithm. *Proceedings of IJCAI-01*; 2001 Aug 4–10; Seattle, WA. 2001. pp. 309–315.
 57. Van Hentenryck P, Deville Y, Teng CM. A generic arc-consistency algorithm and its specializations. *Artif Intell* 1992;57:291–321.
 58. Han C, Lee C. Comments on Mohr and Henderson's path consistency algorithm. *Artif Intell* 1988;36:125–130.
 59. Singh M. 1995. Path consistency revised. *Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence* 1995 Nov 5–8; Washington, DC. 1995. pp. 318–325.
 60. Berlandier P. Improving domain filtering using restricted path consistency. *Proceedings of the IEEE CAIA-95*; 1995 Feb 20–23; Los Angeles, CA. 1995.
 61. Verfaillie G, Martinez D, Bessiere C. A generic customizable framework for inverse local consistency. *Proceedings of the AAAI National Conference*; 1999 Jul 18–22; Orlando, FL. 1999. pp. 169–174.
 62. Freuder EC, Elfe CD. Neighbourhood inverse consistency preprocessing. *Proceedings of the AAAI National Conference*; 1996 Aug 4–8; Portland, OR. 1996. pp. 202–208.
 63. Debruyne R, Bessiere C. Some practicable filtering techniques for the constraint satisfaction problem. *Proceedings of the 15th IJCAI*; 1997 Aug 23–29; Nagoya, Japan. 1997. pp. 412–417.
 64. Bacchus F, van Beek P. On the conversion between non-binary and binary constraint satisfaction problems. *Proceedings of the National Conference on Artificial Intelligence (AAAI-98)*; 1998 Jul 26–30; Madison, WI. 1998.

65. Rossi F, Dahr V, Petrie C. On the equivalence of constraint satisfaction problems. Proceedings of the European Conference on Artificial Intelligence (ECAI-90); 1990 Aug 6; Stockholm, Sweden. 1990.
66. Régin JC. A filtering algorithm for constraints of difference in CSPs. Proceedings of the National Conference on Artificial Intelligence (AAAI-94); 1994 Jul 31–Aug 4; Seattle, WA. 1994. pp. 362–367.
67. Régin JC. Generalized arc consistency for global cardinality constraint. Proceedings of National Conference on Artificial Intelligence (AAAI-96); 1996 Aug 4–8; Portland, OR. 1996. pp. 209–215.
68. Pesant G. A regular language membership constraint for finite sequences of variables. Proceedings of the Tenth International Conference on Principles and Practice of Constraint Programming (CP); 2004 Sep 27–Oct 1; Toronto, Canada. 2004. pp. 183–195.
69. Barták R. Modelling resource transitions in constraint-based scheduling. Proceedings of SOFSEM 2002: Theory and Practice of Informatics 2002 Nov 22–29; Milovy, Czech Republic. 2002. pp. 186–194.
70. Quimper CG, Walsh T. Global grammar constraint. Proceedings of Principles and Practice of Constraint Programming—CP 2006; 2006 Sep 25–29; Nantes, France. 2006. pp. 751–755.
71. Beldiceanu N, Carlsson M, Rampon JX. Global constraint catalogue. Technical Report T2005-06. SICS; 2005.
72. McGregor JJ. Relational consistency algorithms and their application in finding subgraph and graph isomorphisms. *Inf Sci* 1979;19(3):229–250.
73. Gaschnig J. A constraint satisfaction method for inference making. Proceedings of the 12th Annual Allerton Conference on Circuit and System Theory; 1974 Oct 2–4; Monticello, IL. 1974. pp. 866–874.
74. Sabin D, Freuder EC. Contradicting conventional wisdom in constraint satisfaction. Proceedings of ECAI; 1994 Aug 8–12; Amsterdam, The Netherlands. 1994. pp. 125–129.
75. Gomes C, Selman B, Kautz H. Boosting combinatorial search through randomization. Proceedings of National Conference on Artificial Intelligence (AAAI); 1998 Jul 26–30; Madison, WI. 1998. pp. 432–437.
76. Constraints Journal. <http://www.springer.com/computer+science/ai/journal/10601>.
77. Association for Constraint Programming. <http://www.4c.ucc.ie/a4cp/>. Accessed on August 12th 2010.
78. Tsang EPK. Foundations of constraint satisfaction. London: Academic Press; 1993.
79. Marriott K, Stuckey PJ. Programming with constraints: an introduction. Cambridge (MA): The MIT Press; 1998.
80. Abdennadher S, Frühwirth T. Essentials of constraint programming. Berlin: Springer; 2003.
81. Apt K. Principles of constraint programming. Cambridge: Cambridge University Press; 2003.
82. Dechter R. Constraint processing. San Francisco: Morgan Kaufmann Publishers; 2003.
83. Rossi F, Van Beek P, Walsh T, editors. Handbook of constraint programming. Amsterdam: Elsevier; 2006.

HISTORY OF LP DEVELOPMENT

KATTA G. MURTY

Department of Industrial and Operations
Engineering, University of Michigan,
Ann Arbor

Systems Engineering Department, King
Fahd University of Petroleum and
Minerals, Dhahran, Saudi Arabia

MATHEMATICAL MODELING, ALGEBRA, SYSTEMS OF LINEAR EQUATIONS, AND LINEAR ALGEBRA

One of the most fundamental ideas of the human mind, discovered more than 5000 years ago by the Chinese, Indians, Iranians, and Babylonians, is to represent quantities that we want to determine by symbols—usually letters of the alphabet like x, y, z —then express the relationships between the quantities represented by these symbols in the form of equations, and finally, use these equations as tools to find out the true values represented by the symbols. The symbols representing the unknown quantities to be determined are now called *unknowns*, *variables*, or *decision variables*.

The process of representing the relationships between the variables through equations or other functional relationships is called *modeling* or *mathematical modeling*. The earliest mathematical models constructed were *systems of linear equations*, and soon after, the famous *elimination method* for solving them was discovered in China and India.

The Chinese text *Chiu-Chang Suanshu* (*Nine Chapters on the Mathematical Art*) compiled over 2000 years ago describes the method using a problem of determining the yield (measured in units called *tou*) from three types of grain—inferior, medium, superior—given the yield data from three experiments, each using a separate combination of the three types of grain. See

Kangshen *et al.* [1] for information on this ancient work, also a summary of this ancient Chinese text can be seen at the following website: http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/Nine_chapters.html.

Ancient Indian texts *Sulva Suutrah* (*Procedures Based on Ropes*), and the *Bakhshali Manuscript*, which date back to the same period, describe the method in terms of solving systems of two (three) linear equations in two (three) variables; see Refs 2 and 3 for information on these texts. For a summary and review of this book see <http://www.tlca.com/adults/origin-math.html>.

This effort culminated around 825 AD in the writing of two books by the Persian mathematician Muhammad ibn-Musa Alkhawarizmi in Arabic; this attracted international attention. The first was *Al-Maqala fi Hisab al-jabr w'almuqabilah* (*An essay on algebra and equations*). The term *al-jabr* in Arabic means “restoring” in the sense of solving an equation. In Latin translation, the title of this book became *Ludus Algebrae*, the second word in this title surviving as the modern word *algebra* for the subject, and Alkhawarizmi is regarded as the father of algebra. *Linear algebra* is the name subsequently given to the branch of algebra dealing with systems of linear equations. The word *linear* in “linear algebra” refers to the *linear combinations* in the spaces studied, and the *linearity* of “linear functions” and “linear equations” studied in the subject.

The second book, *Kitab al-Jam'a wal-Tafreeq bil Hisab al-Hindi*, appeared in a Latin translation under the title *Algoritmi de Numero Indorum*, meaning *Al-Khwarizmi Concerning the Hindu Art of Reckoning*; it was based on earlier Indian and Arabic treatises. This book survives only in its Latin translation, because all copies of the original Arabic version have been lost or destroyed. The word *algorithm* (meaning procedures for solving algebraic systems) originated from the title of this Latin translation. Algo-

rithms seem to have originated in the work of ancient Indian mathematicians on rules for solving linear and quadratic equations.

An Example Application

Here is an example application from Murty [4] that leads to a model involving simultaneous linear equations. A steel company has four different types of scrap metal (called *SM-1* to *SM-4*) with compositions as given in Table 1. They need to blend these four scrap metals into a mixture for which the composition by weight is Al, 4.43%; Si, 3.22%; C, 3.89%; Fe, 88.46%. How should they prepare this mixture?

To answer this question, we first define the decision variables, denoted by x_1, x_2, x_3, x_4 , where for $j = 1$ to 4 , x_j = proportion of *SM-j* by weight in the mixture to be prepared. Then the percentage by weight of the element Al in the mixture will be $5x_1 + 7x_2 + 2x_3 + x_4$, which is required to be 4.43. Arguing the same way for the percentage by weight in the mixture of the elements Si, C, and Fe, we find that the decision variables x_1 to x_4 must satisfy each equation in the following system of linear equations in order to lead to the desired mixture:

$$\begin{aligned} 5x_1 + 7x_2 + 2x_3 + x_4 &= 4.43, \\ 3x_1 + 6x_2 + x_3 + 2x_4 &= 3.22, \\ 4x_1 + 5x_2 + 3x_3 + x_4 &= 3.89, \\ 88x_1 + 82x_2 + 94x_3 + 96x_4 &= 88.46, \\ x_1 + x_2 + x_3 + x_4 &= 1. \end{aligned}$$

The last equation in the system stems from the fact that the sum of the proportions of various ingredients in a blend must always be equal to 1. From the definition of

Table 1. Compositions of Available Scrap Metals

Type	% in type, by weight, of element			
	Al	Si	C	Fe
SM-1	5	3	4	88
SM-2	7	6	5	82
SM-3	2	1	3	94
SM-4	1	2	1	96

the variables given above, it is clear that a solution to this system of equations makes sense for the blending application under consideration, only if all the variables in the system have nonnegative values in it. The nonnegativity restrictions on the variables are *linear inequality constraints*. They cannot be expressed in the form of linear equations, and since nobody knew how to handle linear inequalities at that time, they ignored them, and considered this system of equations as the mathematical model for the problem.

Necessary and Sufficient Conditions for the Existence of a Solution

Consider the general system of m equations in the column vector $x = (x_1, \dots, x_n)^T$, of unknowns

$$Ax = b, \quad (1)$$

where $A, b = (b_i)$ are $m \times n, m \times 1$ matrices. The elimination method for solving this system of equations, directly leads to the following classical theorem for the existence of a feasible solution [4].

Theorem 1. *Theorem of alternatives for systems of linear equations: The system of linear equations (1) has no feasible solution x iff there is a linear combination of equations in it which becomes the fundamental inconsistent equation " $0 = 1$;" that is, it has no solution iff there exists a row vector $\pi = (\pi_1, \dots, \pi_m)$ such that*

$$\begin{aligned} \pi A &= 0, \\ \pi b &= \alpha \neq 0, \end{aligned} \quad (2)$$

where α is some nonzero number (e.g., $\alpha = 1$). In linear algebra, the system (2) is known as the alternate system corresponding to Equation (1); it shares the data in the original system (Eq. 1).

That is why when Equation (2) has a feasible solution, any solution of it, π , is known as a *certificate of infeasibility* for the original system (Eq. 1).

The Gaussian and the Gauss–Jordan Elimination Methods

The elimination method for solving systems of linear equations remained unknown in Europe until Gauss rediscovered it at the beginning of the nineteenth century while calculating the orbit of the asteroid Ceres based on recorded observations in tracking it earlier. It was lost from view when the astronomer tracking it, Piazzi, fell ill. Gauss got the data from Piazzi, and tried to approximate the orbit of Ceres by a quadratic formula using the data. He designed the *method of least squares* for estimating the best values for the parameters in the formula, to give the closest fit to the observed data; this gives rise to a system of linear equations to be solved. He rediscovered the elimination method to solve that system. Even though the system was quite large for hand computation, Gauss's accurate computations helped in relocating the asteroid in the skies in a few months' time, and his reputation as a mathematician soared.

Gaussian elimination method and *Gauss–Jordan (GJ) elimination method* are two commonly used variants of the elimination method for solving systems of linear equations such as Equation (1); they are still the leading methods in use today. When applied on a system of linear equations (Eq. 1), both of them will terminate by either finding a feasible solution x of the original system or a certificate π (feasible solution for the alternate system (Eq. 2)) establishing that the original system of Equation (1) is infeasible.

LACK OF A METHOD TO SOLVE LINEAR INEQUALITIES UNTIL MODERN TIMES

Even though linear equations had been conquered thousands of years ago, systems of linear inequalities remained inaccessible until modern times. The set of feasible solutions to a system of linear inequalities is called a *polyhedron* or *convex polyhedron*, and geometric properties of polyhedra were studied by the Egyptians earlier than 4000 BC while building the pyramids, and later by the Greeks, Chinese, Indians, and others.

The following theorem [4] relates systems of linear inequalities to systems of linear equations.

Theorem 2. *Consider the system of linear inequalities*

$$Ax \geq b, \quad (3)$$

where $A = (a_{ij})$ is an $m \times n$ matrix and $b = (b_i) \in R^m$. Let A_i denote the i th row vector of A . So, the constraints in the system are $A_i x \geq b_i$, $i \in \{1, \dots, m\}$. If this system has a feasible solution, then there exists a subset $P = \{p_1, \dots, p_s\} \subset \{1, \dots, m\}$ such that every solution of the system of equations

$$A_i x = b_i, i \in P,$$

is also a feasible solution of the original system of linear inequalities (Eq. 3).

This theorem states that every nonempty polyhedron has a nonempty face that is an affine space. It can be used to generate a finite enumerative algorithm to find a feasible solution to a system of linear constraints containing inequalities. It involves enumeration over subsets of the inequalities in the system. For each subset do the following: eliminate all the inequality constraints in the subset from the system. If there are any inequalities in the remaining system change them into equations. Find any solution of the resulting system of linear equations. If that solution satisfies all the constraints in the original system, terminate. Otherwise, repeat the same procedure with the next subset of inequalities. At the end of the enumeration, if no feasible solution of the original system has turned up, it must be infeasible.

However, if the original system has m inequality constraints, in the worst case this enumerative algorithm may have to solve 2^m systems of linear equations before it either finds a feasible solution of the original system, or concludes that it is infeasible. The effort required grows exponentially with the number of inequalities in the system in the worst case, so it is not practical. But this theorem presents an interesting *paradox* about whether systems of linear equations,

or systems of linear inequalities, are more fundamental.

As you know, linear equations can be transformed into linear inequalities by replacing each equation with the opposing pair of inequalities; but there is no way a linear inequality can be transformed into linear equations. This indicates that linear inequalities are more fundamental than linear equations.

But this theorem shows that linear equations are the key to solving linear inequalities, and hence are more fundamental. This is the paradox.

THE IMPORTANCE OF LINEAR INEQUALITY CONSTRAINTS, AND THEIR RELATION TO LINEAR PROGRAMS

The first interest in inequalities arose from studies in mechanics, beginning in the eighteenth century. Crude examples of applications involving linear inequality models started appearing in published literature around the 1700s.

Linear programming (LP) involves optimization of a linear objective function subject to linear inequality constraints. Examples of LP models started appearing in published literature from about the middle of the eighteenth century. An example of an application of LP is the fertilizer maker's product mix problem discussed in Ref. 4. It leads to the following LP model

$$\begin{array}{ll} \text{Maximize } z(x) = 15x_1 + 10x_2 & \text{Item} \\ \text{Subject to} & \\ & 2x_1 + x_2 \leq 1500 \text{ RM 1} \\ & x_1 + x_2 \leq 1200 \text{ RM 2} \\ & x_1 \leq 500 \text{ RM 3} \\ & x_1 \geq 0, x_2 \geq 0 \end{array}$$

in which the decision variables x_1, x_2 are the tons of high-pH, low-pH fertilizers manufactured per day using three raw materials RM 1, 2, 3 for which the available supply is at most 1500, 1200, 500 tons/day, respectively. The limit on the supply of each of these raw materials leads to a constraint in the model; that is why these raw materials are called the *items* corresponding to those constraints in the model. The objective function to be

maximized is the daily net profit from fertilizer manufacturing activities.

In this example, all the constraints on the variables are inequality constraints. In the same way, inequality constraints appear much more frequently and prominently than the equality constraints in most real-world applications. Before the advent of LP, linear equations were used to model problems, mostly because an efficient method to solve them is known.

Fourier was one of the first to recognize the importance of inequalities as opposed to equations for applying mathematics. Also, he was a pioneer who, in the early nineteenth century, observed the link between linear inequalities and linear programs.

Necessary and Sufficient Conditions for the Existence of Feasible Solutions to General Systems of Linear Constraints

The problem of finding a feasible solution to a general system of linear constraints: $Ax = b$, $Dx \geq d$; can itself be posed as a linear program through a Phase I formulation (see any of the Refs 4 to 8) and by solving this Phase I linear program, we can obtain a feasible solution for the system if the system is feasible, or obtain a certificate of infeasibility for it (similar to the one discussed for systems of linear equations in the section titled "Mathematical Modeling, Algebra, Systems of Linear Equation, and Linear Algebra").

Associated with every LP problem, there is another linear program called its *dual*, involving a different set of variables but sharing the same data. When referring to the dual problem of an LP, the original LP is called *the primal* or *the primal problem*. Together, the two problems are referred to as a *primal, dual pair* of linear programs. The names *primal, dual* for the two problems were coined by Tobias Dantzig, father of George Dantzig, around 1955 in conversations with his son.

A duality-type result for systems of linear equations only (no inequalities) is Theorem 1 discussed above, has been known for a long time (by the eighteenth century or even earlier) but similar results for systems of linear constraints including linear inequalities were unknown until recently. These

important duality-type results for systems of linear constraints including inequalities, known as *either/or theorems* or *theorems of alternatives* started appearing in published literature from the mid-nineteenth century.

These theorems show that a given system of linear constraints has a feasible solution iff another system in a different set of variables but constructed using the same data as the original system, has no feasible solution. The first such necessary and sufficient condition for feasibility of a system involving linear inequalities seems to be that of Gordon, given in an 1873 paper:

Gordon's Theorem. *The system: $Ax < 0$ has a solution iff the alternate system: $yA = 0, y \geq 0$ has no nonzero solution.*

At the end of the nineteenth and the first half of twentieth centuries, many such results were published by Farkas, Minkowski, Stienke, Motzkin, and several others. The most famous among them, that appeared in 1894, is

Farkas' Lemma. *The system $Ax = b, x \geq 0$ has a feasible solution iff the alternate system $yA \leq 0, yb > 0$ has no feasible solution.*

The fundamental notion of *duality* and the term itself, were introduced by von Neumann in conversations with Dantzig in 1947, and stated in a working paper written by him in the same year. Gale, Kuhn, and Tucker independently formulated the duality theorem of LP and proved it rigorously using Farkas' lemma in 1951; and they received the John von Neumann Theory prize of INFORMS in 1980 for this work.

Duality plays a big role in LP theory. One of the corollaries of the duality theorems of LP is that the general linear program

$$\begin{aligned} &\text{Minimize } cx & (4) \\ &\text{subject to } Ax = b \\ &\text{and } Dx \geq d \end{aligned}$$

and the system of linear constraints

$$\begin{aligned} Ax &= b, \\ Dx &\geq d, \\ \pi A + \mu D &= c, \\ cx - \pi b - \mu d &\leq 0, \\ \mu &\geq 0, \end{aligned} \tag{5}$$

are mathematically equivalent problems, in the sense that $\bar{x}$ is an optimum solution of Equation (4) iff there exists a $(\bar{\pi}, \bar{\mu})$ such that $(\bar{x}, \bar{\pi}, \bar{\mu})$ is feasible with respect to Equation (5) and conversely if $(\bar{x}, \bar{\pi}, \bar{\mu})$ is any feasible solution to Equation (5), then $\bar{x}$ is an optimum solution of the original LP Eq. (4), and $(\bar{\pi}, \bar{\mu})$ is an optimum solution of the dual of the LP Eq. (4) [4].

This corollary of the duality theorems shows that any linear program can be posed as a problem of solving a system of linear inequalities without any optimization. Thus, solving linear inequalities and LPs are mathematically equivalent problems. Both problems of comparable sizes can be solved with comparable efficiencies by available algorithms. So, the additional aspect of "optimization" in linear programs does not make LPs any harder either theoretically or computationally.

FOURIER ELIMINATION METHOD FOR LINEAR INEQUALITIES

By 1827, Fourier generalized the elimination method to solve a system of linear inequalities. The method now known as the *Fourier* or *Fourier–Motzkin elimination method*, is one of the earliest methods proposed for solving systems of linear inequalities. It consists of successive elimination of variables from the system. However, starting with a system of m inequalities, the number of inequalities can jump to $O(m^2)$ after eliminating only one variable from the system, so this method is not practically viable except for very small problems.

HISTORY OF THE SIMPLEX METHOD FOR LP

In 1827, Fourier also published a geometric version of the principle behind the simplex algorithm for an LP (vertex-to-vertex descent along the edges to an optimum, a rudimentary version of the simplex method) in the context of a specific LP in three variables (an LP model for a Chebyshev approximation problem) but did not discuss how this descent can be accomplished computationally on systems stated algebraically. In 1910, De la Vallée Poussin designed a method for the Chebyshev approximation problem that is an algebraic and computational analog of this Fourier's geometric version; this procedure is essentially the primal simplex method applied to that problem.

In a parallel effort, Gordon (in 1873), Farkas (in 1896), and Minkowski (in 1896) studied linear inequalities and laid the foundations for the algebraic theory of polyhedra and derived necessary and sufficient conditions for a system of linear constraints including linear inequalities to have a feasible solution.

Studying LP models for organizing and planning production, Kantorovich (in 1939) developed the idea of dual variables (*resolving multipliers*) and derived a dual-simplex type method for solving a general LP.

Full citations for references before 1939 mentioned so far can be seen from the list of references in Refs 5 and 6.

This work culminated in the mid-twentieth century with the development of the primal simplex method by Dantzig. This was the first complete, practically and computationally viable method for solving systems of linear inequalities. So, LP can be considered as the branch of mathematics which is an extension of linear algebra to solve systems of linear constraints including linear inequalities. The development of LP as a decision making tool is a landmark

event in the history of mathematics, and its application brought our ability to solve general systems of linear constraints (including linear equations, inequalities) to a state of completion.

Now the simplex method is one of the popular methods for solving either systems of linear constraints including linear inequality constraints, or linear programs. The primary computational tool used by the simplex method for solving an LP is the same one used in the GJ elimination method for linear equations (the GJ Pivot step on the detached coefficient table of the problem), hence the simplex method can be viewed as the extension of the GJ method for solving systems of linear equations to systems of linear constraints including inequalities or LPs.

LP has now become a dominant subject in the development of efficient computational algorithms, the study of convex polyhedra, and in algorithms for decision making. But for a short time in the beginning, its potential was not well-recognized.

Dantzig tells the story of how when he gave his first talk on LP and his simplex method for solving it at a professional conference, Hotelling dismissed it as unimportant "since everything in the world is nonlinear." But von Neumann came to the defense of Dantzig saying that the subject will become very important (see p. xxvii of Ref. 7). The preface in this book contains an excellent account of the early history of LP.

von Neumann's early assessment of the importance of LP turned out to be astonishingly correct. Today, the applications of LP in almost all areas of science are so numerous, so well known, and recognized that they need no enumeration. Also, LP seems to be the basis for most of the efficient algorithms for many problems in other areas of mathematical programming. Many of the successful approaches in nonlinear programming, discrete optimization, and other

Linear Algebra

Study of linear equations. Originated over 2000 years ago.

→

Linear Programming

Study of linear constraints including inequalities. Twentieth century extension of linear algebra.

branches of optimization are based on LP in their iterations. Also, with the development of duality theory and game theory, LP has assumed a central position in economics.

Using the revised simplex algorithm, by the 1960s, Dantzig and Wolfe had developed *column and row generation techniques* for solving LP models without having all the constraints explicitly on hand, but generating each column (or row) vector in the model as needed; see Ref. 5 for details.

WORST-CASE COMPUTATIONAL COMPLEXITY OF THE SIMPLEX METHOD AND OF LP

By 1970, software systems for LP based on the simplex method were well developed and were able to handle very satisfactorily even the largest LP models appearing in applications in those days. So, at that time, everyone considered LP as a well-solved problem, and the simplex method was the ultimate most efficient algorithm for solving it. No one expected any further significant algorithmic developments in LP.

But small groups of theoretical researchers in several centers were busy trying to determine the worst-case computational complexity of the simplex method (i.e., whether it is a polynomial-time algorithm or not), and of LP itself (i.e., whether it belongs to the class P of polynomially solvable problems or not). Then, around 1970, Klee and Minty constructed a class of LPs with $2n$ constraints in n variables $(x_1, \dots, x_n)^T$ for each $n \geq 2$, with the feasible region for the n th problem being a slightly perturbed unit cube of dimension n having 2^n extreme points, and showed that the problem of maximizing x_n in it by a version of the simplex method takes $2^n - 1$ pivot steps to solve. So, to solve the n th problem in this class, this version of the simplex method passes through every extreme point of the feasible region before terminating with the unique optimum solution of the problem. This has conclusively established that even though the simplex method solves LP models very efficiently in practice, at least this particular version of it is not a polynomial-time algorithm in the worst case [8].

By now (2009) all popular versions of the simplex method for LP have been shown to require exponential time effort in the worst case. The question “is there a version of the simplex method for LP that is a polynomial-time algorithm?” remains an unresolved theoretical research problem in LP.

Ever since the work of Klee and Minty, the question “is there a polynomial-time algorithm for LPs?” became a challenge for algorithmic researchers. This challenge was settled with an affirmative by a young Russian mathematician, Leonid Khachiyan, in 1979. He developed the ellipsoid method for LP and showed that the computational effort to solve an LP of size L in n variables using it is bounded above by $O(n^4 L)$; thus establishing that the ellipsoid method is a polynomial-time algorithm for LP. Even though in practice the computational performance of the ellipsoid method is very poor in comparison with the simplex method, Khachiyan’s result established that LP belongs to the class P and ushered in the era of polynomial-time methods for LP [8].

For finding a feasible solution to a system of linear inequalities, like the simplex method, the ellipsoid method also has the remarkable property that it does not need all the constraints in the model explicitly but only a scheme for checking at each stage whether the current iterate is feasible, or a constraint in the model that it violates.

DEVELOPMENT OF INTERIOR POINT METHODS FOR LP

In Ref. 9, Dikin published the first interior point method (IPM) for LP, this method has subsequently been named as the *affine scaling method*. Dikin proved convergence results on this method in a subsequent publication in 1974. But his papers went unnoticed until the event mentioned in the next paras.

It is still not known whether the affine scaling method for LP is a polynomial-time algorithm.

In the early 1980s, Narendra Karmarkar pioneered a new polynomial-time method for LP, an IPM, which claims to be many times faster than the simplex method for solving

large-scale sparse LPs. These claims helped to focus researchers' attention on it. His work attracted worldwide attention, and his talks on it at that time stimulated a lot of work on these methods that stay "away" from the boundary of the feasible region.

In the flush of Karmarkar's international fame, his company at that time, AT&T saw a lucrative business opportunity. They started a small division with several high-level programmers to develop software for Karmarkar's algorithm for LP. Soon they put on the market a bundled system including a computer and software under the trade name *Korbex system* for 8 or 9 million dollars. I believe that they have been able to sell only a couple of systems, because soon after Karmarkar's result became known, many researchers developed IPMs for LP much more efficient than Karmarkar's algorithm, and released software based on these algorithms as free downloads. So, within a very short time the Korbex division of AT&T was closed and the remaining units of the Korbex system were mothballed.

In the tidal wave of research that ensued many different classes of IPMs, barrier algorithms, and central path approaches have been developed for LP and extended to wider classes of problems including convex quadratic programming, monotone linear complementarity problem, and semidefinite programming problems. Among the many IPMs, the *primal-dual IPM*, has become the most popular IPM for solving LPs and LP software implementations.

This primal-dual IPM has been observed to take much smaller number of iterations to solve LPs than the simplex method, and this number grows very slowly with the size of the LP model being solved, with the result that IPMs have gained the reputation of *taking almost a constant number of iterations* even as the size of the LP grows. Now, most large-scale LPs are typically solved using software based on IPMs, commercial solvers of LP like CPLEX are able to solve models arising in applications very fast (a few minutes of CPU time) and thus, meet the needs of practitioners very satisfactorily.

But the simplex method has some subtle advantages too. It is an ideal method for solving small LPs by hand. The IPMs on the other hand, are not suitable for hand computation even on small LPs. Another is that if the LP being solved has an optimum solution, then at termination the simplex method outputs an optimum basic feasible solution (BFS) even when the LP has alternate optimum solutions. On LPs with alternate optimum solutions, the IPMs will output a relative interior point of the optimum face and not a BFS; to move from that point to an optimum BFS requires additional computation. Practitioners typically seem to prefer an optimum BFS over a relative interior point of the optimum face for implementation in real-world applications.

NEWER METHODS TO SOLVE LPS USING MATRIX INVERSION OPERATIONS SPARINGLY

All the methods discussed so far, which are based solely on matrix inversion operations, work very well if the models are very sparse, and their performance depends critically on being able to exploit the sparsity in the model to advantage. For solving large-scale models which do not fit this mold, these methods have difficulty.

So, recent algorithmic research on LP has focused on the methods that can solve LPs fast without using matrix inversion operations, or using them only sparingly. Also, in the methods discussed above, every constraint in the model appears in the matrix inversion operations carried out in every step. If the new methods operate with only a subset of constraints selected intelligently in matrix inversion operations in each step, they may have an advantage. The future of algorithmic research in LP is in this area sphere methods discussed in [4,10,11] are in this category.

REFERENCES

1. Kangshen S, Crossley JN, Lun AWC. The nine chapters on the mathematical art: companion and commentary. China: Oxford University Press (ISBN 0-19-853936-3); 1999.

- pp. xiv + 596. Beijing: Oxford, England and Science Press.
2. Joseph GG. The crest of the peacock: non-European roots of mathematics. Harmondsworth, Middlesex: Penguin Books; 1992.
 3. Lakshmikantham V, Leela S. The origin of mathematics. Lanham (MD): University Press of America, Inc.; 2000. p. 104. (visit www.univpress.com for information, also a summary of this book can be seen at: <http://www.tlca.com/adults/origin-math.html>).
 4. Murty KG. Optimization for decision making; linear and quadratic models. New York: Springer; 2009.
 5. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
 6. Schrijver A. Theory of linear and integer programming. New York: Wiley-Interscience; 1986.
 7. Dantzig GB, Thapa MN. Linear programming: Volume 1 introduction, Volume 2 theory and extensions. New York: Springer; 1997.
 8. Murty KG. Linear programming. New York: Wiley; 1983.
 9. Dikin II. Iterative solution of problems in linear and quadratic programming. Sov Math Dokl 1967;8:674–675.
 10. Murty KG, Oskoorouchi MR. Note on implementing the new sphere method for LP using matrix inversions sparingly. Optimization letters 2008;3:137–160.
 11. Murty KG, Oskoorouchi MR. Sphere methods for LP. Algorithmic OR 2010;5:21–33.

HOLT–WINTERS EXPONENTIAL SMOOTHING

RONG PAN

School of Computing, Informatics, Decision
Systems Engineering, Arizona State
University, Tempe, Arizona

The Holt–Winters method is a class of forecasting procedures for time series with trend and seasonality. It was initially developed based on empirical evidence and intuitive reasoning. Over decades its statistical properties and practical power have been discovered and tested by numerous studies [1–3]. Besides being simple and practical, the Holt–Winters method and its variants are remarkably robust for a wide range of applications. Today, it is one of the most popular forecasting systems employed in demand forecasting, production planning, inventory control, and so on.

During the 1950s, Holt developed an exponential smoothing method for additive trends and seasonal data. His work was originally reported to the Office of Naval Research in 1957 and was recently reprinted in *International Journal of Forecasting* [4]. Winters [5] tested Holt’s method with empirical data, and this method and its variants became known as the *Holt–Winters forecasting system*.

Throughout this article, x_t will be used to denote both a time series and its parent stochastic process, $\hat{x}_t(k)$ to denote a k -step-ahead forecast from time t . The remaining paper is organized as follows: first, an example will be given for illustrating the trend and seasonal components commonly appeared in time series, followed by the description of the additive and multiplicative versions of the Holt–Winters method. Then, the estimation method of smoothing parameters and the diagnosis of forecasting method are discussed. The software implementation of the Holt–Winters method is described. Finally, some further developments of the Holt–Winters method are presented.

EXAMPLE: AIRLINE PASSENGER DATA

The following time series (Fig. 1) shows the totals of international airline passengers from January 1949 to December 1960 (12 years, 144 observations). Data are collected monthly and are listed as Series G in Brown [6]. The series clearly shows seasonal pattern—more traffic in July and August but less in November—and an increasing trend. It also shows that the amplitude of the cycle enlarges over years.

After a natural logarithm transformation of the data, the series is identified as a seasonal ARIMA (0,1,1)X(0,1,1)₁₂ in Box *et al.* [7]. The forecasts of the passenger amount in the last two years (Years 1959 and 1960) as well as one year in future according to this model are plotted in Fig. 2a. One can see that the cyclical feature has been captured by the model, but the modeled increasing trend and increasing amplitude of cycle are lower than expected. Figure 2b shows the forecasts using the Holt–Winters method using the original data set with three smoothing parameters as 0.323, 0.045, and 0.980. It appears that the result of the Holt–Winters method is better than the previous one. The mean square error of the ARIMA method calculated for the last 2 years (January 1959–December 1960) is 1706. The mean square error of the Holt–Winters method is 1434.

HOLT–WINTERS METHOD

The Holt–Winters forecasting method is a heuristic decomposition procedure for seasonal data. It decomposes the forecast of a time series to three components—level, trend, and seasonal. The estimations of these components are updated by applying exponential smoothing technique (See, *The Exponentially Weighted Moving Average*). Therefore, there are three smoothing parameters and they are determined either *a priori* or by fitting historical data. Depending on the types of trend and seasonal components, the Holt–Winters

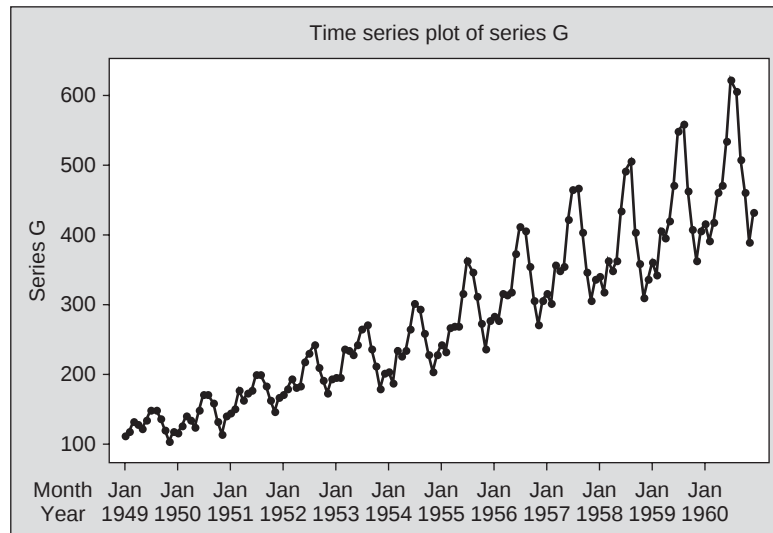


Figure 1. Monthly international airline passenger (in thousands).

forecasting method can be specified as for a time series of linear trend, exponential trend, or damped trend, and additive seasonal or multiplicative seasonal, and any combination of trend and seasonal. In this section we will describe two basic Holt–Winters forecasting procedures for linear trend with additive seasonality and linear trend with multiplicative seasonality. For other variants of the Holt–Winters method, one may refer to Ref. 1. In general, if the amplitude of the

seasonal pattern is independent of the level, then an additive model is appropriate; however, if the amplitude of the seasonal pattern is proportional to the level, a multiplicative seasonal effect needs to be considered.

The basic Holt–Winter model assumes a linear trend. Adopting the notations used in most textbooks and literature, we write the forecasting procedure as the following for the additive seasonal version of the

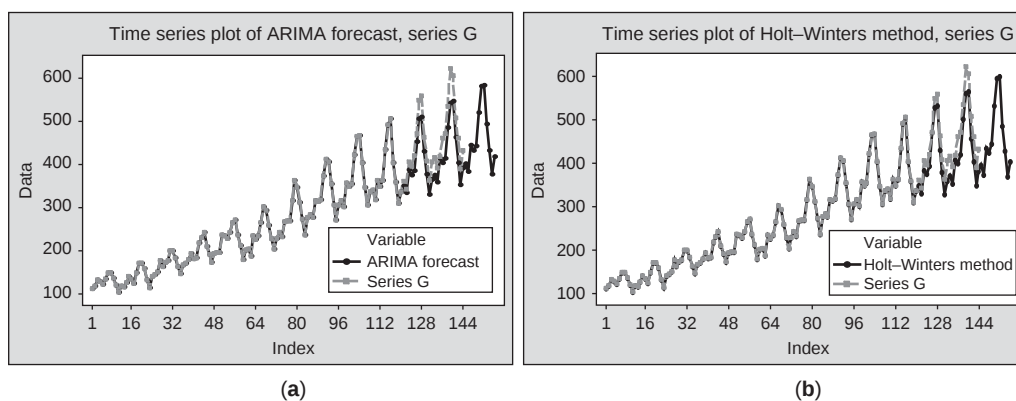


Figure 2. Forecasting using (a) ARIMA (0,1,1)X(0,1,1)₁₂ model and (b) the Holt–Winters method.

Holt–Winters method:

$$L_t = \alpha(X_t - S_{t-p}) + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad (1)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}, \quad (2)$$

$$S_t = \gamma(X_t - L_t) + (1 - \gamma)S_{t-p}, \quad (3)$$

where L_t is the level component, T_t the trend component, and S_t the seasonal component; α , β , and γ are exponential smoothing parameters. Let p be the length of cyclical period and k be the forecasting lead time. The k -step-ahead forecast of a time series and the one-step-ahead forecasting error are given by

$$\hat{x}_t(k) = L_t + kT_t + S_{t+k-p} \quad (4)$$

and

$$e_t = x_t - (L_{t-1} + T_{t-1} + S_{t-p}), \quad (5)$$

respectively.

For the multiplicative seasonal version,

$$L_t = \alpha(X_t/S_{t-p}) + (1 - \alpha)(L_{t-1} + T_{t-1}), \quad (6)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}, \quad (7)$$

$$S_t = \gamma(X_t/L_t) + (1 - \gamma)S_{t-p}, \quad (8)$$

$$\hat{x}_t(k) = (L_t + kT_t)S_{t+k-p}, \quad (9)$$

and

$$e_t = x_t - (L_{t-1} + T_{t-1})S_{t-p}. \quad (10)$$

It is well known that the underlying optimal model of the additive version of the Holt–Winters method is an ARIMA time series model. By optimal, we mean that the additive Holt–Winters method provides the minimum one-step-ahead mean square error forecasting for the underlying ARIMA model. The discussion of the equivalency of exponential smoothing and the models to certain types of ARIMA models can be found in Roberts [8]. However, the multiplicative Holt–Winters method does not have an ARIMA equivalent, because it is a nonlinear forecasting method in nature. Even for the additive Holt–Winters method, Chatfield [9] argued that its ARIMA equivalent is so complex that the model identification through

the time series’ autocorrelation structure becomes impossible. Recent theoretical development has shown that exponential smoothing methods are optimal for a very general class of state–space models, which is broader than the ARIMA class.

It has long been criticized that the Holt–Winters method, as a heuristic, does not have a solid theoretical foundation, which, in turn, hampers the identification of a time series that could be best forecasted by the Holt–Winters method. In fact, using the Box–Jenkins methodology [7], most time series generated from the additive Holt–Winters models cannot be correctly identified. Gardner [1] summarized a practical identification procedure. First, we plot the series and look of trend, seasonal variation, outliers, and changes in structure that may be slow or sudden and may indicate that exponential smoothing is not appropriate in the first place. We should examine by outliers, consider making adjustments, and then decide on the form of the trend and seasonal variation. We should also consider the possibility of transforming the data, either to stabilize the variance or to make the seasonal effect additive. Next, we fit an appropriate method, produce forecasts, and check the adequacy of the method by examining the one-step-ahead forecast errors, particularly their autocorrelation function. The finding may lead to a different method or a modification of the selected method.

Empirical studies have shown that the Holt–Winters forecasting system is robust at short horizons, but tends to overshoot the data at longer horizons (more than three or four time periods ahead); so it may be advisable to damp a linear trend at long horizons. Parzen [10] tackled the problem by creating “short-memory” forecast and “long-memory” forecast, where the “short-memory” utilizes the autocovariance of stationary time series and the “long-memory” contains a nonstationary trend component. Lewandowski’s FORSYS program [11] added the damped trend as the lead time increases and the rate of damping increases with the level of noise. Gardner and McKenzie [12] discussed a generalization of the Holt model with

damped trends to avoid the overshoot of forecast over long lead time. Their study showed that this method improves long-term forecast accuracy compared to the traditional Holts method, while it does not lose much power in short-term forecast.

FITTING TIME SERIES AND FORECASTING

The smoothing parameters used in the Holt–Winters method can be determined by data analysts based on their experience. It is often suggested that the smoothing parameter value should be in the range of 0.1–0.3. However, many studies [13] found that the values of α and γ often exceed 0.3, but the value of β is frequently less than 0.1. One of the common practices of fitting a time series is to find the parameter values that minimize the sum of the squares of the one-step-ahead forecast errors (SSE), that is,

$$\text{Min}_{\alpha, \beta, \gamma} \sum_t e_t^2 \quad (11)$$

$$\text{S.T. } 0 \leq \alpha, \beta, \gamma < 1.$$

Some other optimization criteria, such as minimizing mean absolute percentage error (MAPE) or mean absolute deviation (MAD), have also been used, but they are not as popular as SSE. Since the additive Holt–Winters model has the ARIMA equivalence, the prediction can be written as a linear combination of past observations, that is, $\hat{x}_t(1) = \phi_t x_t + \phi_{t-1} x_{t-1} + \dots + \phi_1 x_1$. The coefficients in the autoregressive function are functions of the three smoothing parameters. Therefore, the optimal parameter values can be found using the standard method such as the conditional likelihood method [7]. In the multiplicative model, the forecast depends on the past values of the series in a nonlinear manner. Typically numerical search algorithms are required for finding the optimal parameter values.

Finding the optimal smoothing parameters can be a difficult task. According to Ref. 14, this is due to (i) the fitting criteria, such as likelihood functions, being highly nonlinear functions of the parameters and (ii) the likelihood functions not required to

be concave so that the numerical optimization routines cannot guarantee global optima without a full grid search. Snyder and Shami [14] proposed a parsimonious model, which contains only two smoothing parameters and has certain advantages in simplifying parameter estimation. Smoothing parameters are typically restricted to the (0,1) interval. However, McClain [15] and Sweet [16] showed that some parameter values in this interval produce noninvertible models, that is, the impact on forecasts of values in the distant past is larger than for recent values. This is counterintuitive and may produce poor forecasts. Archibald [17] used a parameter space, which is a subset of the additive invertible region, to avoid this problem. With proper approximation, the multiplicative model can be handled in this way too.

The Holt–Winters method has been implemented in many statistical software packages. In Minitab v14 (Time Series $\rightarrow$ Winters' Method), one has to specify the smoothing parameter values *a priori*. Minitab only implements the Holt–Winters method with known smoothing parameters. In SAS, the FORECAST procedure can be used while specifying “method = WINTERS” (for the multiplicative method) or “method = ADDWINTERS” (for the additive method). The Time Series Forecasting System (a graphical data analysis application) in SAS allows one to find the optimal smoothing parameters.

In the remainder of this section, several other issues related to forecasting using the Holt–Winters method are discussed. They include how to find the starting values for recursively estimating the level, trend, and seasonal components (Eqs 1–3 for the additive method and Eqs 6–8 for the multiplicative method); how to normalize seasonal components; how to produce prediction intervals; and the diagnosis of forecasts.

Starting Values

The starting values of level, trend, and seasonal components ($L_0, T_0, S_{1-p}, \dots, S_0$) must be specified at the beginning of the series in order to initiate the updating procedure. The simplest way is to use the average of the observations in the first cycle period, that is,

let $L_0 = \sum_{t=1}^p x_t/p$, $T_0 = 0$ and $S_{t-p} = x_t - L_0$ for the additive version, and $S_{t-p} = x_t/L_0$ for the multiplicative version.

Some researchers suggested using the first two or three cycle periods, instead of one. Some even suggested using the entire observation in calculating the initial values [5,18]. Other less simple methods can be found in Chatfield [13]. Makridakis *et al.* [19] suggested a backcasting method, where the time order of observations is reversed and the most recent data are used to start up the procedure in the reverse direction. The values at the end of the series are then used as the starting values.

Normalization of Seasonal Factor

Seasonal component estimates can be contaminated by some of the other factors being modeled. Because it was observed in practice that seasonal terms are ceasing over time to be simply seasonal effects, many authors have suggested renormalizing the seasonal terms either once a year or at every observation [20].

The seasonal terms are originally set up to reflect the seasonal pattern only; so, in an additive version of the method they should sum to zero and in a multiplicative version they should multiply to one. However, they are not longer so after the calculation over several cycles. Examples have been used to demonstrate that without seasonal factor renormalization, although the forecasts will perform well, the estimates of the individual components (level and seasonal) will be unreliable. The common practice of renormalization is to subtract from each seasonal term a specific amount so that they sum to zero (for the additive version). However, when this renormalization is carried out, forecasts will be biased for the next few periods before the method adjusts level estimates. Therefore, Lawton [21] recommended adjusting the level as well by adding the same amount. Roberts [8] and McKenzie [22] suggested a variant of the Holt–Winters method, which maintains the balance between good forecasts and component estimates and does not require the adjustment of level. The modified seasonal estimate is (for the additive

version)

$$S_t = \gamma \frac{p-1}{p} (x_t - L_t) + \left(1 - \gamma \frac{p-1}{p}\right) S_t(p), \quad (12)$$

$$S_t(j) = S_{t-1}(j-1) - \frac{\gamma}{p} (x_t - L_t - S_{t-1}(p-1)),$$

$$\text{for } 1 \leq j \leq p-1, \quad (13)$$

where $S_t(j)$ is the seasonal estimate at time t for the season which occurred j periods earlier.

Archibald and Koehler [23] put forward a practical method that makes the renormalization equations in a common form. Using state–space equations, they adapted the renormalization equations for the multiplicative version. They recommended that renormalization of seasonal factors should be done in every time period and in the multiplicative version the level and trend should be adjusted along with the seasonal component.

Prediction Intervals

Because the additive Holt–Winters model can be converted to an equivalent optimal ARIMA model, the theoretical result of the prediction interval can be obtained. Yar and Chatfield [24] proposed a method of constructing prediction intervals for the additive Holt–Winters forecasting system. It was extended to the multiplicative version in Chatfield and Yar [25]. It was found that the width of multiplicative prediction intervals will depend on both the lead time and the forecasting time, and it may either decrease or increase with lead time. The exact prediction intervals are complicated and difficult for calculation and comprehension. Oftentimes, simple normal approximations of prediction intervals are employed [26] and the k -step-ahead 95% prediction intervals are given below.

For the additive Holt–Winters method,

$$\left[\hat{x}_t(k) \pm z_{0.025} \sqrt{c_k s^2} \right], \quad (14)$$

where $z_{0.025}$ is the percentile of standard normal distribution; s^2 is the sample variance of prediction errors, $s^2 = \sum_{t=0}^{n-1} (x_{t+1} -$

$\hat{x}_t(1))^{2/(n-3)}$; c_k is a correction factor, $c_1 = 1$, $c_k = 1 + \sum_{j=1}^{k-1} \alpha^2(1+j\beta)^2$ for $2 \leq k \leq p$, $c_k = 1 + \sum_{j=1}^{k-1} [\alpha(1+j\beta) + d_{j,p}(1-\alpha)\gamma]^2$ for $p < k$, and $d_{j,p} = 1$ if j is an integer multiple of p , and 0 otherwise.

For the multiplicative Holt-Winters method,

$$\left[\hat{x}_t(k) \pm z_{0.025} S_{t+k-p} \sqrt{c_k s_r^2} \right], \quad (15)$$

where S_{t+k-p} is the seasonal component of one cycle before; s_r^2 is the sample variance of relative prediction errors, $s_r^2 = \sum_{t=0}^{n-1} [(x_{t+1} - \hat{x}_t(1))/\hat{x}_t(1)]^2/(n-3)$; for the correction factor, $c_1 = (L_t + T_t)^2$, $c_2 = \alpha^2(1+\beta)^2(L_t + T_t)^2 + (L_t + 2T_t)^2$, and $c_k = \alpha^2(1+(k-1)\beta)^2(L_t + T_t)^2 + \dots + \alpha^2(1+\beta)^2(L_t + (k-1)T_t)^2 + (L_t + kT_t)^2$ for $2 \leq k \leq p$.

Monitoring Forecasts

Tracking signal is a tool for monitoring the goodness of a forecasting system. Brown [6] suggested a simple cumulative sum tracking signal. It is a ratio of the cumulative sum of one-step-ahead forecast errors to the smoothed MAD. Trigg [27] suggested a smoothed-error tracking signal. Trigg's signal is simply the ratio of the smoothed error to the smoothed MAD. The value of the signal will fluctuate between 0 and 1. A large signal value indicates that the forecasting system is producing errors that are consistently larger or less than 0; therefore, it is biased.

Directly monitoring one-step-ahead forecasting errors is another common way for

detecting any deficiency of the forecasting system. It is recommended to plot the time series of the errors and their sample autocorrelation and sample partial autocorrelation functions. A good forecasting system will produce random errors, so no pattern is expected from these plots; otherwise, an investigation should be performed to modify the forecast procedure.

SOFTWARE IMPLEMENTATION

We revisit the airline passenger data set, which was discussed in the previous section. The data is taken from Box *et al.* (7, p. 547). By examining the time series plot such as Fig. 1, it is found that a strong increasing trend and the amplitude of seasonal cycle increase along with the trend. Thus, a multiplicative Holt-Winters model is appropriate.

Figure 3a is the command window of the Winter's Method in Minitab. One can select the multiplicative or additive model and specify the seasonal period and parameter values. To generate forecasts, one needs to specify the number of forecasts and the origin time of forecasts. In the "Storage" option, one may select "Forecasts", "Residuals", "Upper 95% prediction limit", "Lower 95% prediction limit", and so on. Note that residuals are the one-period-ahead forecast errors. These residuals are used to calculate MAPE, MAD, and MSD (mean square deviation). These residuals can also be used to generate diagnostic plots using Autocorrelation. After running this model, results are presented in the

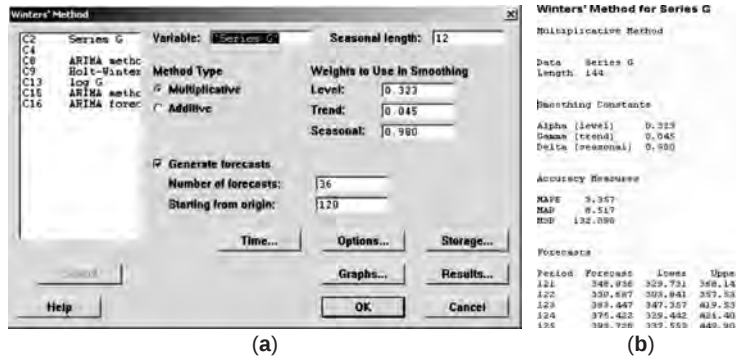


Figure 3. The Winter's Method command window in Minitab (a) and outputs (b).

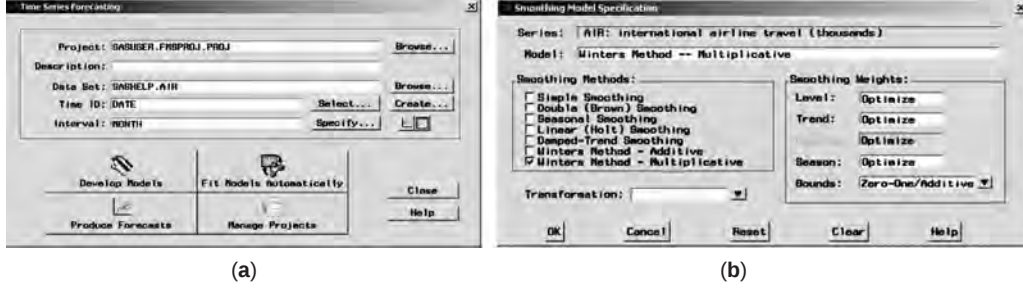


Figure 4. (a) SAS Time Series Forecasting window and (b) Smoothing Model Specification window.

Minitab session window. A snapshot of these results is given in Fig. 3b.

SAS v9.1 also has a graphical interface for time series analysis, which can be found through its top menu, Solutions → Analysis → Time Series Forecasting Systems. The airline passenger data set has already been stored in the SASHELP directory and the data file is named as “air.” In the Time Series Forecasting window, one may select “Develop Models,” then click on the blank and select “Fit smoothing model...” A smoothing model specification window will pop up, where one can choose the Winters’ multiplicative method, as well as weights on level, trend, and season. If “Optimize” is selected, then the optimal weights will be found by using an optimization criterion, such as sum of square error, which is selected in the previous “Develop Models” window. Figure 4 depicts these graphical interfaces.

FURTHER DEVELOPMENT

Irregular Observation Intervals

An *ad hoc* modification of Holt–Winters method with missing observations (irregular observation intervals) was discussed in Cipra *et al.* [28]. The basic idea is to adjust the smoothing parameter values according to the length of intermittent time. The three smoothing parameters used in the Holt–Winters method are modified as

$$V_{t_n} = \frac{V_{t_{n-1}}}{(1 - \alpha)^{t_n - t_{n-1}} + V_{t_{n-1}}}, \quad (16)$$

$$U_{t_n} = \frac{U_{t_{n-1}}}{(1 - \beta)^{t_n - t_{n-1}} + U_{t_{n-1}}}, \quad (17)$$

$$W_{t_n} = \frac{W_{t_{n-1}}}{(1 - \gamma)^{(t_n - t_n^*)/p} + W_{t_{n-1}}}, \quad (18)$$

where α, β, γ are the smoothing parameters used in the Holt–Winters method for regular time series; U, V, W are new smoothing parameters; t_n is the time of observation; and t_n^* is the corresponding time of the same seasonal period as t_n in the last cycle. Note that the time index of an irregular time series, t_n , is the actual time or normalized time, instead of a simple sequence index as used in regular time series.

State–Space Formulation

Ord *et al.* [29] draw the connection of the exponential smoothing method to a general class of state–space models as its underlying optimal model, which is broader than the traditional ARIMA class in that it allows the error variance to depend on the level and seasonal component. The optimal refers to that conditioning on the assumed model for which the exponential smoothing method gives the minimum mean square error forecast. This, in return, explains the robust nature of the Holt–Winters method, which was shown in Chen [30] using Monte Carlo simulation and empirically tested in Makridakis and Hibon [31]. The single source of error (SSOE) state–space models given below are special cases of Refs 29 and 32–34.

For the additive Holt–Winters model, observation equation

$$x_t = L_{t-1} + T_{t-1} + S_{t-p} + \varepsilon_t, \quad (19)$$

state equations

$$L_t = L_{t-1} + T_{t-1} + \alpha \varepsilon_t, \quad (20)$$

$$T_t = T_{t-1} + \alpha \beta \varepsilon_t, \quad (21)$$

$$S_t = S_{t-p} + (1 - \alpha) \gamma \varepsilon_t. \quad (22)$$

The error term ε_t has a value from the normal distribution that has the zero mean and the standard deviation σ .

For the multiplicative Holt–Winters model,

observation equation

$$x_t = (L_{t-1} + T_{t-1}) S_{t-p} (1 + \varepsilon_t), \quad (23)$$

state equations

$$L_t = L_{t-1} + T_{t-1} + \alpha (L_{t-1} + T_{t-1}) \varepsilon_t, \quad (24)$$

$$T_t = T_{t-1} + \alpha \beta (L_{t-1} + T_{t-1}) \varepsilon_t, \quad (25)$$

$$S_t = S_{t-p} + (1 - \alpha) \gamma S_{t-p} \varepsilon_t. \quad (26)$$

Note that, unlike the additive model, in the multiplicative model the variances in three state equations are not constants—they depend on the most recent level and trend components in the level and trend state equations, and the last corresponding seasonal component in the seasonal state equation.

Gardner [2] praised the development of Holt–Winters state–space models as the most significant advancement of this method in recent years. These models can also be easily modified to accommodate damped trends. More recently, the state–space models are utilized in modeling multiple seasonal patterns, such as the daily and hourly patterns of traffic flows [35].

REFERENCES

- Gardner ES. Exponential smoothing: the state of the art. *J Forecast* 1985;4:1–28.
- Gardner ES. Exponential smoothing: the state of the art—part II. *Int J Forecast* 2006;22:637–666.
- Makridakis S. Empirical evidence versus personal experience. *J Forecast* 1983;2:295–306.
- Holt CC. Forecasting seasonals and trends by exponentially weighted moving averages. *Int J Forecast* 2004;20:5–10.
- Winters PR. Forecasting sales by exponentially weighted moving averages. *Manage Sci* 1960;6:324–342.
- Brown RG. Smoothing, forecasting and prediction of discrete time series. Englewood Cliffs (NJ): Prentice Hall; 1962.
- Box GEP, Jenkins GM, Reinsel GC. Time series analysis: forecasting and control. 3rd ed. Upper Saddle River (NJ): Prentice Hall; 1994.
- Roberts SA. A general class of Holt–Winters type forecasting models. *Manage Sci* 1982;28(7):808–820.
- Chatfield C. Some recent developments in time series analysis. *J R Stat Soc [Ser A]* 1977;140:492–510.
- Parzen E. ARARMA models for time series analysis and forecasting. *J Forecast* 1982;1:67–82.
- Lewandowski R. Sales forecasting by FORSYS. *J Forecast* 1982;1:205–214.
- Gardner ES, McKenzie E. Forecasting trends in time series. *Manage Sci* 1985;31(10):1237–1246.
- Chatfield C. The Holt–Winters forecasting procedure. *Appl Stat* 1978;27:264–279.
- Snyder RD, Shami RG. Exponential smoothing of seasonal data: a comparison. *J Forecast* 2001;20:197–202.
- McClain JL. Dynamics of exponential smoothing with trend and seasonal terms. *Manage Sci* 1974;20:1300–1304.
- Sweet AL. Computing the variance of the forecast error for the Holt–Winters seasonal models. *J Forecast* 1985;4:235–243.
- Archibald BC. Parameter space of the Holt–Winters’ model. *Int J Forecast* 1990;6:199–209.
- Montgomery DC, Johnson LA. Forecasting and time series analysis. New York: McGraw-Hill; 1976.
- Makridakis S, Andersen A, Carbone R, *et al.* The accuracy of extrapolation (time series) methods: results of a forecasting competition. *J Forecast* 1982;1:111–153.
- Chatfield C, Yar M. Holt–Winters forecasting: some practical issues. *Statistician* 1988;37:129–140.
- Lawton R. How should additive Holt–Winters estimates be corrected? *Int J Forecast* 1998;14:393–403.

22. McKenzie E. Renormalization of seasonals in Winters' forecasting systems: is it necessary? *Oper Res* 1986;34:174–176.
23. Archibald BC, Koehler AB. Normalization of seasonal factor in Winters' methods. *Int J Forecast* 2003;19:143–148.
24. Yar M, Chatfield C. Prediction intervals for the Holt-Winters forecasting procedure. *Int J Forecast* 1990;6:1–11.
25. Chatfield C, Yar M. Prediction intervals for multiplicative Holt-Winters. *Int J Forecast* 1991;7:31–37.
26. Bowerman BL, O'Connell RT, Koehler AB. *Forecasting, time series, and regression: an applied approach*. 4th ed. Belmont (CA): Thomson Brooks/Cole; 2005.
27. Trigg DW. Monitoring a forecasting system. *Oper Res Q* 1964;15:271–274.
28. Cipra T, Trujillo J, Rubio A. Holt-Winters method with missing observations. *Manage Sci* 1995;41(1):174–178.
29. Ord JK, Koehler AB, Snyder RD. Estimation and prediction for a class of dynamic nonlinear statistical models. *J Am Stat Assoc* 1997;92:1621–1629.
30. Chen C. Robustness properties of some forecasting methods for seasonal time series: a Monte Carlo study. *Int J Forecast* 1997;13:269–280.
31. Makridakis S, Hibon M. The M3-competition: results, conclusions and implications. *Int J Forecast* 2000;16:451–476.
32. Chatfield C, Koehler AB, Ord JK, *et al.* A new look at models for exponential smoothing. *J R Stat Soc [Ser D]* 2001;50:147–159.
33. Koehler AB, Snyder RD, Ord JK. Forecasting modes and prediction intervals for the multiplicative Holt-Winters method. *Int J Forecast* 2001;17:269–286.
34. Hyndman RJ, Koehler AB, Snyder RD, *et al.* A state space framework for automatic forecasting using exponential smoothing methods. *Int J Forecast* 2002;18:439–454.
35. Gould PG, Koehler AB, Ord JK, *et al.* Forecasting time series with multiple seasonal patterns. *Eur J Oper Res* 2008;191:207–222.

HORSE RACING

JOHN BACON-SHONE
Social Sciences Research Centre,
The University of Hong Kong,
Hong Kong, China

Horse racing is an ancient sport, dating back to Roman times. It has a strong hold in Great Britain and many former British colonies, including Canada, United States, Australia, New Zealand, Hong Kong, India, and Singapore. However, it is also popular in many other countries, including France, Italy, Japan, South Korea, United Arab Emirates, and more. While the most popular form is racing of thoroughbred horses on a flat course, racing over a track with obstacles that require the horses to jump (steeplechase racing) is also popular in some countries and there is also racing where the horses pull people in carts around tracks (known as *trotting*). In all forms of horse racing, the outcome of the race is the rank order of the horses, so only relative time to complete the course matters. Table 1 shows the top 20 countries that had the largest number of races in 2006, according to the worldwide horseracing body, the International Federation of Horseracing Associations [1]. This shows that the United States has by far the largest number of races on the flat (51,491), trots (53,610) and overall (105,278), while Great Britain has the largest number of jump races (3,380). The total number of races in 2006 worldwide was around 320,000, which is over 800 races per day.

In any given horse race, there are usually 6–15 horses competing (although it can be as low as 4 or as high as 40), with prize money usually offered to at least the first three horses. The performance of a horse in a race depends on many variables, including the trainer, the jockey, the distance, the track type and condition, the pedigree, age and condition of the horse, and so on.

Of all the sports, horse racing has attracted the widest range of academic interest from economists, psychologists, statisticians and others studying risk, operational research, and finance, arguably because of the richness of data available from race-tracks across the world for up to a century, allowing detailed analysis of both performance and gambling data from multiple perspectives. The large betting pools mean that successful models can be financially very rewarding. Compared to many financial markets, it has the simplicity of regular known end points. Although often only the aggregate gambling data at the end point is available, for some gambling pools the odds offered on each horse or combination of horses may be publicly available and regularly updated until the start of the race.

GAMBLING

In practice, the major interest in horse racing is because of the gambling, which usually dominates the prize money, although the prize money can be substantial for major races, like the Dubai World Cup, which offers a prize of US\$6 million (about €4 million). Table 2 shows the top 20 countries ordered by the average prize money per race in euros. Hong Kong, UAE, and Singapore rank 1, 2, and 4 on this metric although none of them appear in the top 20 countries ranked on races as shown in Table 1. What is most noticeable is that the prize money per race is usually less than 10% of the betting turnover per race. Although the United States is a clear number one in terms of total races, it is only ranked 3rd on total betting turnover behind Japan and Great Britain and 18th on betting turnover per race. Hong Kong dominates in turnover per race, followed by South Korea, Great Britain, and Ireland. In terms of percentage return to the bettor, Great Britain, Ireland, and Australia are clearly the most generous (88–89%).

Table 1. Total Number of Horse Races in 2006: Top 20 Countries

Country	Flat Races	Jump Races	Trot Races	Total Races
USA	51,491	177	53,610	105,278
Australia	19,821	142	15,666	35,629
Canada	5,281	0	27,183	32,464
Italy	5,550	272	16,751	22,573
Japan	17,806	133	0	17,939
France	4,582	2,179	10,731	17,492
Sweden	629	15	8,872	9,516
Great Britain	5,554	3,380	0	8,934
Germany	1,720	60	5,734	7,514
Argentina	6,737	0	418	7,155
Chile	6,136	0	0	6,136
New Zealand	2,855	132	2,435	5,422
Brazil	4,742	0	0	4,742
Norway	253	7	4,424	4,684
South Africa	3,883	0	0	3,883
Turkey	3,237	0	0	3,237
India	2,854	0	0	2,854
Ireland	874	1,394	0	2,268
Peru	1,918	0	0	1,918
South Korea	1,670	0	0	1,670
Total	161,297	8,163	149,482	318,942

Source: IFHA.

Table 2. Prize Money per Race in 2006 in Euros: Top 20 Countries

Country	Prize Money	Betting Turnover	Turnover/Race	Prize/Race	Return (%)
Hong Kong	66,470,899	6,207,983,843	8,550,942	91,558	83
UAE	26,925,364	0		88,280	
South Korea	101,091,877	4,104,169,139	2,457,586	60,534	72
Singapore	24,561,393	924,524,975	1,313,246	34,888	81
Japan	545,832,655	20,335,405,413	1,133,586	30,427	74
Macau	20,865,185	276,613,099	352,373	26,580	84
Ireland	55,061,000	3,611,439,828	1,592,346	24,277	88
France	349,501,759	8,301,610,124	474,595	19,981	73
Qatar	4,567,680	0		18,568	
Turkey	57,913,530	783,456,251	242,032	17,891	45
Great Britain	153,509,828	15,480,510,598	1,732,764	17,183	89
Greece	15,952,041	337,517,392	280,563	13,260	80
Saudi Arabia	4,512,276	0		10,693	
USA	980,779,104	11,221,983,915	106,594	9,316	79
Spain	4,468,810	25,575,000	52,408	9,157	29
Italy	195,235,678	2,900,713,355	128,504	8,649	N/A
Sweden	76,527,000	1,240,569,000	130,367	8,042	70
Mauritius	1,856,923	79,719,000	332,163	7,737	75
Australia	274,739,351	6,992,917,680	196,270	7,711	88
Cyprus	7,490,028	89,634,818	89,456	7,475	76

Source: IFHA.

In addition to betting on the winning horse, there are many other types of bet on a single race usually available, including identifying any one of the first two or three horses (place or show), the first two horses in finishing order (exacta or perfecta) or unordered (quinella), the first three horses in finishing order (tierce or trifecta) or unordered (trio). There are also often bets that cover two (double), three (treble or pick three), or even up to six races, which usually require identifying all the winners and may have a jackpot that rolls over if there is no winner.

There are two major types of bets, namely, fixed odds bets, usually placed with a bookmaker and variable odds bets, usually placed through a pari-mutuel system that pays out the winning bets pro rata using all of the pool after removing tax and overheads, so that you are in effect gambling against the rest of the bettors rather than a bookmaker. While betting originally required visiting the race-track in order to place the bet, the development of telecommunications and the internet have facilitated the growth of remote betting, so the bettor does not even need to be in the same country and may be in a country where such betting is illegal. Fixed odds bets are now often available not only from bookmakers, but through betting exchanges [2], where other bettors act as the bookmaker and provide the other side of the bet, with the exchanges taking a percentage of the bet. Because offshore bookmakers and betting exchanges can operate while only paying limited tax, they can often offer more attractive odds than onshore operators, although some countries such as the United Kingdom have reduced tax in order to remain competitive. Betting exchanges also usually offer better rates of return than bookmakers as their risk and overheads are usually lower.

PREDICTING THE FINISHING ORDER OF A RACE

Winning a bet requires predicting the outcome of a race. While some gamblers may select bets in a very subjective way, the large money at stake has facilitated the development of sophisticated proprietary models

that incorporate all the information available about the horses, jockeys, trainers, tracks, and so on. The simplest scenario requires predicting the winning horse.

Bolton and Chapman [3] used multivariate logistic regression incorporating 10 horse- or jockey-specific variables, utilizing the discrete choice model of McFadden [4]. What is sometimes missed is that the multinomial logit discrete choice model assumes the “independence from irrelevant alternatives” (IIA). This means that the Bolton and Chapman model, like the model proposed by Harville [5] (which assumes that the performances of the horses expressed in terms of running times follow independent exponential distributions), leads to the conclusion that the chance of horse i beating horse j does not depend on the performance of any other horse. This means that the chance that a horse comes second has a simple relationship to the chance of winning as follows:

$$\begin{aligned} &\Pr(\text{horse } j \text{ comes second} | \text{horse } i \text{ comes first}) \\ &= p_j / (1 - p_i), \end{aligned}$$

where p_i is the chance that horse i comes first. A simpler formulation is in terms of the log odds that horse i beats horse j for k th position, if neither horse comes in the top $k - 1$ positions,

$$LO(i, j|k) = \log(p_i/p_j),$$

which does not depend on k .

This allows modeling the performance not only of the winning horse, but also of all other positions in the race, as it treats each position as a competition amongst all horses that did not finish in an earlier position. However, this would imply that horses compete as strongly for second last place as they do for first place, which is implausible. One way to avoid this problem is to only model the winning position, but that ignores the information about second and third position, where there are clear incentives for the horse and jockey, given the prizes and gambling.

More complex bets also require predicting the second and third placed horses, again raising the question of how the chance of a horse coming second relates to its chance

of winning. Lo and Bacon-Shone [6] summarized different models for the underlying abilities of the horses, which all enable estimation of the probability distribution for the finishing order, given the winning probabilities. Other possible underlying models include the multivariate independent normal distribution proposed by Henery [7] and the generalization of the exponential to the gamma distribution proposed by Stern [8], both of which require numerical solution of a system of integral equations in order to fit the models. Lo *et al.* [9] proposed an extension to the Harville model that provides a good, simple approximation to the Stern and hence the Henery models.

This model assumes that

$$\text{LO}(i, j|k) = \lambda_{k,n} \log(p_i/p_j),$$

where $\lambda_{k,n}$ is decreasing in k and depends weakly on n , the number of horses in the race and the racetrack through the shape parameter in Stern's gamma model, which has Harville's model as the limiting distribution. This model also includes Henery's model as the simplification of $\lambda_{k,n} = 1$. Lo *et al.* [9] showed, using this approximation, that the Henery model fits well for US and Hong Kong data sets, while the Stern model is needed for Japan and that the Harville model does not fit any of the data sets adequately. Lo *et al.* [9] also identified an even simpler approximation, which is that

$$\lambda_{k,n} \approx \mu^k,$$

which fits moderately well in the data sets they examine.

This formulation can be seen as providing guidance how to discount information from lower finishing positions, which is rational given the reducing incentives for lower finishing positions in terms of both prize and gambling returns, without needing to exclude this information completely.

TESTING THE EFFICIENCY OF THE BETTING MARKETS

Betting on horse races has attracted wide interest from economists interested in testing

under what conditions this market is efficient. Ziemba [10] provided a recent review of the well-developed field. Roberts [11] defined efficiency as weak, semistrong, and strong when prices reflect all price, all public, and all information respectively. For weak form efficiency, the initial paper by Griffith [12] found evidence of greater expected returns on horses at shorter odds than those at longer odds for pari-mutuel betting at US tracks (commonly known as the *favorite/long-shot bias*) although he did not do any formal statistical test. Snyder [13] showed statistical evidence that the betting public and newspaper experts both overbet long shots, while Asch *et al.* [14] not only showed statistical evidence of inefficiency in US win pools, with overbet long shots, but also that late bettors have better information, although in all cases the expected return was negative, given the track take. This was replicated in Australia by Bird *et al.* [15] and Dowie [16] in the United Kingdom using bookmaker data, unlike the papers above which were all restricted to pari-mutuel betting. Bird and McCrae [17] also examined the public information in newspaper tipsters odds and the change in bookmaker odds, but did not find any evidence of inefficiency, unless there is prior knowledge of the change in odds.

However, Busche and Hall [18] showed that this inefficiency does not exist in Hong Kong. One possibility is that the large pools in Hong Kong (as seen in Table 2, they are on average more than 50 times the US pools per race and 5 times the Japanese pools per race) have attracted sufficient serious gamblers, like Benter, to remove the bias. Lo and Bacon-Shone [19] used a more sensitive log-odds test instead of grouping by odds or position as used by Busche and Hall. They conclude that Hong Kong and Shanghai have no bias toward long shots, while at least some tracks in the United States and Japan have a positive bias, although Busche [20] showed that other tracks in Japan do not show the bias and Walls and Busche [21] suggest this mixed result in Japan may reflect that tracks showing the bias have much lower turnover than the tracks that show the bias. Walls and Busche [21] also showed that at least

some of the bias in favor of long shots is in fact bias caused by breakage (the rounding down of payouts to the nearest 10 cents).

Interestingly, Smith *et al.* [22] found significant evidence of long-shot bias even in betting exchange data for Great Britain, although the amount of bias was significantly less than for bets with bookmakers.

In recent years attention has focused on trying to identify the underlying source of this bias. In simple terms, the bias can be caused either by a supply issue (i.e., bookmakers protecting themselves against inside information) or demand (bettors who prefer more risky bets). Aggregate data from win bets alone is not sufficient to distinguish between these alternatives.

One key step forward was Shin's papers that developed a model [23] to show how inside knowledge could explain the long-shot preference and then applied [24] that model to Dowie's data, concluding that the incidence of insider trading was around 2%. Shin's model assumes that the bookmakers know that insiders always back the winner, while the proportion of money on all other horses is consistent with the chances of those horses winning and they adjust the odds to derive zero maximal profit. Williams and Paton [25] showed that the evidence of insider trading appears to be much less in higher handicap races, where there is more public information. However, Schnytzer and Shilony [26] showed that Shin's formulation is not sufficient to fully explain the long-shot preference as there is still bias after allowing for the inside information.

One important point is that analyzing pari-mutuel betting alone eliminates supply side explanations and Bruce *et al.* [27] examined a data set that allows segmentation of the pari-mutuel market into three parts: home race track, other racetracks, and off-track betting. They argue that given the travel and admission costs of attending a track and the preference of social gamblers for on-track atmosphere at the home track, there will be differences in the long-shot bias in the three markets. Using the Lo and Bacon-Shone log-odds test mentioned above, they show a positive bias for the home track and negative bias for other tracks and

off-track betting, which they explain as a rational exploitation of the mispriced betting of home track bettors.

This still does not provide conclusive evidence to distinguish between inside information explanations and risk preference explanations. Snowberg and Wolfers [28] investigated this distinction for a large US data set of about 650,000 races. They formulate two extreme positions, unbiased expectations and risk-loving (preferences model) and biased expectations and risk-neutral (perceptions model). The key innovation is that they investigate exacta and other combinatoric bets. Their initial analysis is flawed as it uses the Harville model to predict the conditional probabilities for the horse coming second given the knowledge of the winning horse. However, as noted above, Lo and Bacon-Shone [6] have shown conclusively that for United States, Hong Kong, and Japan (i.e., both in situations that do and do not show the long-shot preference), the Harville model fits second and third place conditional probabilities very poorly and that Stern's independent gamma model fits measurably better. However, Snowberg and Wolfer go on to estimate the conditional probabilities for a grid of combinations of win probabilities for the two horses. Even this approach is arguably flawed as the analysis of Lo and Bacon-Shone showed that for the gamma model the relationship between the win and exacta probabilities depends (although only weakly) on the number of horses in the race. Despite these possible limitations, this paper provides strong evidence that the perceptions model provides predictions of exacta, quinella, and trifecta bets given the win bets that are much more consistent than the preference model. The key remaining question is to explain the Hong Kong anomaly of an anomaly as presented by Busche and Hall [18].

Sung and Johnson [29] reviewed the literature on semistrong efficiency in horse race betting and conclude that while bettors are remarkably successful in discounting relatively simple information, the semistrong form inefficiency observed appears to arise from complex, subtle, and possible changing relationships.

BETTING SYSTEMS FOR HORSE RACING

Hausch *et al.* [30] showed how to use win bet information to take advantage of inefficiencies in the place and show markets. This system combined estimates of place and show probabilities using Harville's model and the optimal capital growth model of Kelly [31] to select the bet sizes to maximize asymptotic growth of capital. This criterion yields a relative bet size of

$$b = (t_p - 1)/(t - 1),$$

where p is the estimate for the bet to win, t is the gross return and the bet is placed if t_p is greater than one. Hausch and Ziemba [32] extended the system to cover multiple horses, multiple bets, and account for breakage and take. Gramm and Ziemba [33] summarized subsequent improvements made to their system, including breeding and performance data on the horses. Hausch and Ziemba [34] focused specifically on risk-free hedges (referred to as a *lock* in horse racing) and identified that for pools that require a minimum payout on show bets, a heavily backed favorite can provide a lock in the show market.

Bolton and Chapman [3] not only developed estimation procedures for predicting horse performance, they also combined it with various betting strategies to see if they would yield a profit, given the semistrong form inefficiencies known to exist. Although their strategies were better than a random strategy, they were unable to achieve a profit. Chapman and Winchester [35] tried the model on a larger sample of 2000 Hong Kong race data and concluded that a profit was possible.

Details of a successful gambling system built on the Bolton and Chapman model can be found in the two papers by Benter [36,37]. This system also used the ideas of Hausch [30] *et al.* and Lo *et al.* [9], although it also contained an important innovation in that Benter used a two-stage process. Firstly, he used multivariate linear regression of the finishing position (normalized to the $[-0.5, +0.5]$ interval) on all the performance data he collected on horses, trainers, and jockeys.

He then applied these regression coefficients to construct a performance estimate for each horse in a race and fit the Bolton and Chapman type model to the linear combination of this performance estimate and the public log odds for the horse to win that race in order to estimate the weights in the linear combination.

Edelman [38] has recently examined replacing the first regression stage of Benter's approach with the Support Vector Machine (SVM) of Hearst *et al.* [39] and showed that he can generate profits on a 100-race holdout sample taken from a small sample of only 200 races in Australia, outperforming the linear regression approach of Benter and a feedforward Neural Network model. Lessman *et al.* [40] revisited Edelman's approach, but they only use a binary win-lose outcome in the first SVM stage instead of the normalized finishing position. They then apply their approach to 400 races from Great Britain with a holdout sample of 156 races for evaluation. They illustrate that their model generates higher profits than Edelman's approach, which in turn generates higher profits than the original approach of Bolton and Chapman. However, they exclude information on positions other than winning, on the grounds that the distribution of second place is different from first place. This argument is flawed given the results of Lo *et al.* [9] mentioned above which show that although positions other than winning provide less information, they do contain valuable information, even if Edelman's model did not take full advantage of that information.

BETTING TURNOVER FOR HORSE RACES

While the total money bet on a race is not of direct interest to many gamblers, it is clearly of great interest to bookmakers, operators of pari-mutuel pools, and governments receiving tax revenue as it has direct impact on their revenues. It is also of interest to large-scale gamblers as large bets can change the payouts substantially in smaller pools. It is also of interest in understanding the impact of changes in the regulatory regimes and

of the competitiveness of horse racing with other forms of leisure activity.

Thalheimer and Ali [41] reviewed evaluations of the competition with pari-mutuel betting on horse races from within and outside the industry. Total turnover of pari-mutuel wagering on horse racing declined 44% from 1977 to 2002. They show that competition from casino gaming has had a significant negative impact, while slot machines inside pari-mutuel betting shops also reduces the turnover. They also show negative impacts from state lotteries and a professional sporting event. While simulcasts reduce the local turnover, there is evidence of increased total pari-mutuel turnover. However, their analysis looks primarily at gross turnover at the daily or monthly level.

In terms of betting at the race level, Ray [42] examined the predictors of pari-mutuel betting turnover for just 200 individual races across a one-month period at a single track, which greatly limits generalizability. Her model has an R^2 of about 64%. Gramm *et al.* [43] examined 2,957 races across 12 US racetracks from October to December 2002. Their data covers from five to nine different pools on each race. Their regression model for log of total turnover has an unadjusted R^2 of about 82% and shows increases with an increase in betting pools, more horses, more competitive races, later races, and larger prizes. While they also examine separate models for win, place, show, exacta, and trifecta pools, they do not directly examine the distribution of money across pools. Bacon-Shone and Woods [44] provided the most detailed study within location (5271 races at two tracks over seven seasons from 2000 to 2007). They include economic (employment), meteorological (humidity, temperature, and rainfall), and race data. Their model yields an adjusted R^2 of about 84% for the log of total turnover for nine pari-mutuel pools, with increases for lower class, more horses, middle distances, later races, less rain, races earlier in the week. They also show the impact of a rebate on losing win bets of 10% on losing bets of at least HK\$10,000 in the win, place, quinella, and quinella-place pools introduced at the start of the 2006 season. They also examine

the distribution of money between pools eligible for the rebate and other pools. This shows that the increased bet options move money away from the rebate pools, as does a strong first or second favorite, while the introduction of the rebate moved money back to the rebate pools.

CONCLUSION

In summary, horse racing data continues to yield fascinating research that provides insight into the efficiency of markets and what determines human behavior and horse performance as well as the development of sophisticated consumer choice models that will undoubtedly be utilized in many other fields that are also rewarding, even if not as financially rewarding as gambling.

Interesting future research directions include identifying better models for the finishing order of horses that incorporates explicitly the discounting for lower positions; confirmation that the perceptions model is more consistent generally with gambling data than the preference model, including in Hong Kong where the long-shot bias is absent; further examination of the returns when using SVM modeling; and examining whether efficiency of betting on horse races is increasing over time, given the wide availability of sophisticated models to assist gamblers and the increasing role of gambling exchanges.

REFERENCES

1. International Federation of Horseracing Associations. Facts and figures. 2006. Available at <http://www.ifhaonline.org/wageringDisplay.asp?section=4>. Accessed 2009 May 17.
2. Smith MA, Williams LV. Betting exchanges: a technological revolution in sports betting. In: Hausch DB, Ziemba WT, editors. Handbook of sports and lottery markets. San Diego (CA): North-Holland; 2008. pp. 403–418.
3. Bolton R, Chapman R. Searching for positive returns at the track: a multinomial logit model for handicapping horse races. *Manage Sci* 1986;32:1040–1060.
4. McFadden D. Conditional logit analysis of qualitative choice behavior. *Front Econ* 1974;8:105–142.

5. Harville D. Assigning probabilities to the outcomes of multi-entry competitions. *J Am Stat Assoc* 1973;68:312–316.
6. Lo VSY, Bacon-Shone J. Approximating the ordering probabilities of multi-entry competitions by a simple method. In: Hausch DB, Ziemba WT, editors. *Handbook of sports and lottery markets*. San Diego (CA): Elsevier; 2008. pp. 51–65.
7. Henery R. Permutation probabilities as models for horse races. *J Roy Stat Soc Ser B (Methodological)* 1981;43:86–91.
8. Stern H. Models for distributions on permutations. *J Am Stat Assoc* 1990;85:558–564.
9. Lo V, Bacon-Shone J, Busche K. The application of ranking probability models to racetrack betting. *Manage Sci* 1995;41:1048–1059.
10. Ziemba W. Efficiency of racing, sports and lottery betting markets. In: Hausch DB, Ziemba WT, editors. *Handbook of sports and lottery markets*. San Diego (CA): Elsevier; 2008. pp. 183–222.
11. Roberts H. Statistical versus clinical prediction of the stock market. 1967, Unpublished.
12. Griffith R. Odds adjustments by American horse-race bettors. *Am J Psychol* 1949;62:290–294.
13. Snyder W. Horse racing: testing the efficient markets model. *J Finance* 1978;33:1109–1118.
14. Asch P, Malkiel BG, Quandt RE. Racetrack betting and informed behavior. *J Financ Econ* 1982;10(2):187–194.
15. Bird R, McCrae M, Beggs J. Are gamblers really risk takers? *Aust Econ Pap* 1987; 26(49):237–253.
16. Dowie J. On the efficiency and equity of betting markets. *Economica* 1976;43:139–150.
17. Bird R, McCrae M. Tests of the efficiency of racetrack betting using bookmaker odds. *Manage Sci* 1987;33:1552–1562.
18. Busche K, Hall C. An exception to the risk preference anomaly. *J Bus* 1988;61:337–346.
19. Lo V, Bacon-Shone J. Probability and statistical models for racing. *J Quant Anal Sports* 2008;4(2):11.
20. Busche K. Efficient market results in an Asian setting. In: Hausch DB, Ziemba WT, editors. *Efficiency of racetrack betting markets*. Singapore: World Scientific Publishing Company; 2008 ed. 1994. pp. 615–616.
21. Walls W, Busche K. Breakage, turnover, and betting market efficiency. In: Williams LV, editor. *The economics of gambling*. London: Routledge; 2003;43.
22. Smith M, Paton D, Williams L, *et al.* Market efficiency in person-to-person betting. *Economica* 2006;73(292):673–689.
23. Shin H. Prices of state contingent claims with insider traders, and the favourite-longshot bias. *Econ J* 1992;102:426–435.
24. Shin H. Measuring the incidence of insider trading in a market for state-contingent claims. *Econ J* 1993;103:1141–1153.
25. Williams L, Paton D. Why is there a favourite-longshot bias in British racetrack betting markets. *Econ J* 1997;440:150–158.
26. Schnytzer A, Shilony Y. Is the presence of inside traders necessary to give rise to a favorite-longshot bias? In: Williams LV, editor. *The economics of gambling*. London: Routledge; 2003. pp. 14.
27. Bruce A, Johnson J, Peirson J, *et al.* An examination of the determinants of biased behaviour in a market for state contingent claims. *Economica* 2009;76(302):282–303.
28. Snowberg E, Wolfers J. The favorite-longshot bias: understanding a market anomaly. In: Hausch DB, Ziemba WT, editors. *Examining explanations of a market anomaly: preferences or perceptions?* San Diego (CA): Elsevier; 2008.
29. Sung M, Johnson JEV. Semi-strong form efficiency in horse race betting. In: Hausch DB, Ziemba WT, editors. *Handbook of sports and lotteries markets*. San Diego (CA): Elsevier; 2008. pp. 275–306.
30. Hausch D, Ziemba W, Rubinstein M. Efficiency of the market for racetrack betting. *Manage Sci* 1981;27:1435–1452.
31. Kelly J Jr. A new interpretation of information rate. *Inform Theor IRE Trans* 1956;2(3): 185–189.
32. Hausch D, Ziemba W. Transactions costs, extent of inefficiencies, entries and multiple wagers in a racetrack betting model. *Manage Sci* 1985;31:381–394.
33. Gramm M, Ziemba WT. The dosage breeding theory for horse racing predictions. In: Hausch DB, Ziemba WT, editors. *Handbook of sports and lottery markets*. San Diego (CA): Elsevier; 2008. pp. 307–340.
34. Hausch D, Ziemba W. Locks at the racetrack. *Interfaces* 1990;20:41–48.
35. Chapman R, Winchester M. Still searching for positive returns at the track: empirical results from 2,000 Hong Kong races. In: Hausch DB, Ziemba WT, editors. *Efficiency of racetrack*

- betting markets Singapore: World Scientific Publishing Company; 2008 ed. 1994. p. 173.
36. Benter W. Advances in the mathematical modeling of horse race outcomes. 12th International conference on gambling and risk-taking Vancouver, Canada. 2003.
 37. Benter W. Computer based horse race handicapping and wagering systems: a report. In: Hausch DB, Ziemba WT, editors. Efficiency of racetrack betting markets Singapore: World Scientific Publishing Company; 2008 ed. 1994. p. 183.
 38. Edelman D. Adapting support vector machine methods for horserace odds prediction. *Ann Oper Res* 2007;151(1):325–336.
 39. Hearst M, Dumais S, Osman E, *et al.* Support vector machines. *IEEE Intell Syst* 1998;13(4):18–28.
 40. Lessmann S, Sung M, Johnson J, Identifying winners of competitive events: a SVM-based classification model for horserace prediction. *Eur J Oper Res* 2009;196(2):569–577.
 41. Thalheimer R, Ali MM. Pari-Mutuel horse race wagering - competition from within and outside the industry. In: Hausch DB, Ziemba WT, editors. Handbook of sports and lottery markets. San Diego (CA): Elsevier; 2008. pp. 3–15.
 42. Ray M. Determinants of simulcast wagering: the demand for harness and thoroughbred horse races. Unpublished ms. 2002.
 43. Gramm M, McKinney CN, Owens DH, *et al.* What do bettors want? Determinants of pari-mutuel betting preference. *Am J Econ Sociol* 2007;66(3):465–491.
 44. Bacon-Shone J, Woods A. Modeling money bet on horse races in Hong Kong. In: Donald BH, William TZ, editors. Handbook of sports and lottery markets. San Diego (CA): Elsevier; 2008. pp. 17–24.

HOUSING AND COMMUNITY DEVELOPMENT

MICHAEL P. JOHNSON
Department of Public Policy and
Public Affairs, University of
Massachusetts Boston,
Boston, Massachusetts

Housing is a key component of the US economy: in 2001, housing comprised more than a third of the nation's tangible assets, and housing consumption and related spending represented more than 21% of the US gross domestic product [1]. Housing that is of a decent quality and *affordable* is associated with increased household wealth, family stability, mental and physical health, labor market participation, educational achievement, and neighborhood quality [2]. Decent and affordable housing also contributes to the improved physical, economic, environmental, and social health—the *sustainability*—of communities. These impacts are especially important for lower-income and disadvantaged households.

However, the benefits of housing, and of sustainable communities, are unequally distributed. Examples include increases in renters in market-rate housing paying excessive fractions of incomes in rent or living in severely substandard housing [3], shortages in affordable and “workforce” housing [4], flat or declining funding levels for public housing authorities [1], persistent gaps in homeownership rates by race and ethnicity [5], excessive housing and transportation burdens on working families [6–8], challenges in economic revitalization facing older urban centers [9], and displacement of the urban poor as a consequence of public housing renovation and economic renewal [10]. These concerns are summarized as inequalities across class, race, and ethnicity in the “geography of opportunity” linking housing, education, employment, and other services [11].

Most recently, the US economic recession of 2007–2009 [12] has had adverse effects across the economy, particularly residential housing. Since 2005, there have been substantial decreases in median housing values, home equity, existing home sales, mortgage refinances, total housing starts and new home sales, and substantial increases in home mortgage delinquencies and foreclosures [13]. It is estimated that American homeowners have lost more than \$4 trillion in wealth associated with housing equity between July 2006 and July 2008 [14]. Foreclosures resulting from the housing market decline are associated with increased crime, decreased property values, and losses in tax revenues [15]. These negative impacts have been borne disproportionately by communities in the west, southeast, upper midwest, and portions of the northeast, by low- and middle-income households, and by black and Hispanic households [13]. The foreclosed housing crisis, combined along with longer-term adverse trends in rental housing, has resulted in increased pressure on the rental housing market and a mismatch between affordable rental housing supply and demand [16].

Public policy analysis and design can increase the efficiency of market-rate housing production. It can also reduce the magnitude of inequities within housing markets, as well as inequalities related to housing-related social outcomes. Management science and operations research, and related disciplines, can help determine when, where, what type and by what means affordable housing and sustainable communities might be built, redeveloped, and maintained. This article argues that decision modeling research that is interdisciplinary, that combines multiple analytical perspectives and that addresses the needs of multiple stakeholders and multiple policy objectives may provide useful guidance for policy analysis in housing and community development. Most of the policy motivation and definitions, and much of the

research described in this article, reflects a US perspective.

We review a number of key definitions used throughout this article. “Subsidized” and “assisted” housing refers to housing intended for low- and moderate-income families (in the United States, typically, those with incomes at or below 80% of the area median income) is produced with funds or incentives provided by federal, state or local governments. Means by which this is done for rental housing include government ownership and operation of newly built housing; government contracts or tax credits with private developers to produce housing, or direct subsidies to families to secure private-market housing on their own [17]. Owner-occupied housing can be produced using mortgage interest and real estate tax deductions, below-market interest rate mortgages, and mortgage insurance and support for government-sponsored enterprises that increase market demand for affordable housing financing instruments [4].

“Affordable” housing is targeted at middle- and lower-income families (typically, those with incomes at or below 120% of the area median income) and which does not consume more than 30% of a household’s income. Such housing, whether renter- or owner-occupied, may receive direct support from government sources, as for subsidized/assisted housing, or may be provided in exchange for economic incentives provided by political jurisdictions or entities responsible for them, or to conform with zoning rules [4]. “Workforce housing” is affordable housing intended specifically for households whose members provide essential municipal services [18].

Housing and communities can be built or modified to be economically, socially or environmentally “sustainable” [19]. For example, public policies or design standards may reduce the negative environmental impacts, such as energy consumption or pollution, of residential and community development [20]. Also, public policies or design standards may enable diverse families to live in housing and neighborhoods that allow broad access to economic and social mobility. Such communities may have the potential to smoothly

adapt to changes over time in employment, infrastructure, and demographics.

Decision models for housing and community development build on research and best practices in supply and demand analysis, land acquisition, construction, and management. However, these models, though numerous, diverse, and well-represented in high-quality scholarly research outlets, have not achieved their potential in policy and practice. Why, then, are they worth studying?

First, these are policy-relevant decision models. They address problems that affect choices or resources of individual families (“person-based” strategies) as well as problems that affect physical infrastructure or human resources of entire communities (“place-based” strategies) [21]. They address important policy and practice concerns such as minimizing costs, maximizing benefits, maximizing geographic deconcentration, and maximizing fair access to affordable housing and sustainable communities. Secondly, decision models may provide direct assistance to urban and regional planners who may rely on more traditional methods such as land-use zoning, subdivision regulation, growth management, smart growth, equitable development, and inclusionary zoning [22]. Finally, decision models may generate recommendations that balance national-level policy priorities, such as creating new housing choices, protecting current housing choices, and changing attitudes and preferences regarding housing choice [11].

Analysis of the research literature in decision models for housing and community development yields a number of insights. First, while long-lived, rich, and diverse, this literature is relatively thin and disconnected within the research community and disconnected from practice: innovative models and applications are published in diverse journals, often with nonoverlapping disciplinary audiences. Secondly, the lack of validation of decision models makes the research less attractive to policy-makers (who demand evidence that an innovative program is likely to improve outcomes) or to field managers (who seek guidance regarding routine scheduling and resource allocation decisions). Thirdly, the primary importance of these decision

models for practitioners is not their specific prescriptions or methodological innovations, but instead the potential for improved systems and process knowledge that makes better use of training and expertise. Finally, the current crisis in housing markets provides opportunities for novel applications for acquiring and redeveloping foreclosed housing units, designing more socially and economically sustainable housing construction and community development strategies, and enabling nonprofit housing providers to collaborate more effectively.

The remainder of this article is organized as follows: the first section is a review of the literature of decision modeling applications in housing and community development, with emphasis on those that address issues of affordability and sustainability. The next section synthesizes previous research and identifies key unanswered questions. This is followed by a section containing a research agenda for decision models in housing and community development; the last section concludes the article.

SURVEY OF DECISION MODELING APPLICATIONS

Our survey of decision modeling applications in affordable housing and sustainable community development is divided into three areas. *Descriptive* research seeks to explain what we observe; specifically, evidence regarding policy initiatives or strategies. *Prescriptive* research supports the identification of a most-preferred policy alternative or set of alternatives. *Decision support systems* (DSS) are information technology applications that automate the policy analysis and recommendation processes.

Descriptive Research

Descriptive research provides increased understanding of social phenomena that motivate or are influenced by policy initiatives. This research can identify the causes of specific events or observed outcomes, provide evidence regarding the efficacy of policy initiatives or social interventions, and can motivate and justify policy-relevant decision

models. Descriptive research for housing and community development encompasses disciplines within the social sciences (economics, policy analysis, urban and regional planning), engineering (civil engineering, industrial and systems engineering), as well as operations research and management science. The literature in these areas is vast, and we make no effort to provide a comprehensive review. In particular, we acknowledge that entire research journals and books are devoted to the study of housing, community development, urban economics, and urban affairs. The purpose of this brief review of a portion of this domain is to identify that subset of research which intends, or can be readily adapted, to support the design of prescriptive models for systems design/redesign. Nearly all of the research discussed in this section describes social phenomena in developed countries, especially the United States.

We distinguish between descriptive research that is *retrospective* (analysis of particular data sets, or literature reviews) and that which is *prospective* (systems models, sensitivity analyses, and simulations). Certain descriptive OR/MS models such as stochastic processes are, for the purposes of this article, classified as “prescriptive” because they are situated within a research tradition of analysis for system design/redesign.

Retrospective policy-analytic research, relevant to housing and community development, includes a historical survey of US low-income housing policy culminating in a proposal to fundamentally redesign housing policy to better meet the needs of low- and moderate-income families and communities [23], and a survey of current trends in transit-friendly, mixed-use development and redevelopment of distressed inner-city neighborhoods into mixed-income communities via public–private partnerships [24]. In addition, it is demonstrated in Ref. 25 that US consumers still overwhelmingly prefer the traditional suburban model of detached, single-family owner-occupied housing. A research review asserts that the presence of low- and moderate-income families in the low-density suburbs of Australia that

lack certain amenities, such as transit access, is not in itself evidence of “locational disadvantage” justifying social interventions; the role of individual social status and locational decisions balancing multiple criteria must be considered as well [26]. An examination of the Canadian rental housing sector since World War II [27] distinguishes markets for housing stock and for rental housing accommodations, describes trends in market outcomes, and identifies social policies that may have produced these outcomes.

There are many applications of housing construction engineering to increase affordability, energy efficiency, and structural integrity, as well as to decrease negative environmental impacts. Best practices are presented in Ref. 28 for increased use of alternative energy sources and more efficient heating and air conditioning systems. Advanced computer simulation methods and architectural methods to maximize passive solar exposure and minimize building footprints have resulted in significant potential energy savings [29]; similar technologies can be applied to rehabilitation of existing housing in low-income areas [30]. Process engineering methods such as concurrent engineering [31] and knowledge management [32] can increase the speed and quality of housing construction.

Descriptive models that are more prospective in nature can be classified according to the unit of observation and the level of aggregation. We consider first, those models that take a systems view, without explicit consideration of individual housing units or of individual households. A systems dynamics approach may be used to identify variables to measure environmental sustainability of different development strategies [33]. A critical review of conceptual models and planning frameworks for sustainable affordable housing results in a proposal of a new approach for affordable housing planning that addresses different development phases, is cross-disciplinary, and involves multiple experts and stakeholders [19]. Systems dynamics is used to model the social, economic, and environmental sustainability of housing and community development [34]. Alternative national-level

policies for renovation of the public housing stock are evaluated on the basis of empirically estimated survivor functions for housing estimated at the city level [35]. A proposed research agenda for social development—including housing—lies at the intersection of systems analysis, sustainable development, and MS/OR [36].

Another descriptive modeling approach addresses local, regional or national markets using housing units or households as the basis of analysis. Variations in economic relationships between private developers and public housing managers, and physical configurations of subsidized and market-rate housing are shown to have impacts on regional housing markets [37]. Analysis of production levels of social (subsidized) housing in Ontario is used to determine whether individual planning areas are receiving their “fair share” of social housing [38]. Housing mobility programs, which enable low-income families to choose housing in socially and economically diverse neighborhoods, may decrease regional-level well-being even if they achieve their programmatic goals [39]. Agent-based models may be used to simulate the spatial impacts of residential location on urban sprawl [40].

Other descriptive models use individual housing units, or families, as the unit of observation, and aggregate these actors to derive characteristics of larger systems. The relationship of an individual housing unit to its environment is used to define measures of environmental sustainability, and thus to measure sustainability impacts of housing-level performance targets [20]. Observations of individual housing units in a city are inputs to estimated survivor functions based on actuarial methods [41]. A life cycle analysis of housing using a stocks-and-flows-based model supports the evaluation of the benefits and costs of different housing construction and maintenance practices [42]. Observations of homes advertised for sale can help identify clusters of housing with shared attributes, as well as key physical determinants of housing choice [43]. Surveys and focus groups of residents in affordable housing are used to

build and validate a neural network model of residential satisfaction [44].

Much descriptive research that is relevant to decision models derived from observations of individual units and households, has a policy focus on affordable and assisted housing. A method for evaluating the benefits and costs of housing mobility programs is applied to early evaluations of a housing mobility experiment [45]. The American Housing Survey is used to estimate the likelihood of homeownership as a function of a variety of affordable lending policies [46] and to demonstrate that regulations that restrict the supply of newly constructed, market-rate housing can reduce the size of affordable housing stock [47].

Prescriptive Models

Prescriptive models for housing and community development can be classified as to their temporal and geographical scope, as well as programmatic and spatial specificity. Certain systems models abstract away the details of specific markets or population regions to concentrate on mathematical properties. A systems model of population that flows between living arrangements in the wake of a natural disaster is used to formulate multiple partial differential equations whose transient solutions allow evaluation of different public policies [48]. An optimal control model of a generic housing mobility initiative is used to identify stable and unstable long-term equilibria associated with different housing mobility policies [49,50].

Other models introduce limited spatial as well as programmatic specificity. Optimization of the social surplus associated with stylized land-use alternatives, subject to allocation and capacity constraints, yields insights regarding the influence of a parameter governing maximization of individual utility or of community utility [51].

Some prescriptive housing models are motivated directly by policy initiatives in defined geographic regions, but place less emphasis on specific spatial characteristics. Formula-based allocations and a gravity model are used to derive proposed allocations of affordable housing to portions of a metropolitan region [52]. A linear

programming model to allocate low-income households in defined zones in order to minimize total commuting and housing costs can contribute to the design and implementation of a fair housing policy [53]. Multiobjective math optimization is used to generate alternative potential allocations of households using rental vouchers to census tracts across a county that balances net social benefit and equity (see *Nonlinear Multiobjective Programming*) [54,55]; a variation of this model [56] incorporates uncertainty as to the actual locational outcomes of voucher recipients. A multiobjective model for location of project-based subsidized rental housing optimizes social efficiency and equity measures [57]; this model is extended to address affordable housing produced by governmental and nongovernmental organizations [58]. Production levels for affordable housing can be chosen to minimize total costs while accounting for environmental impacts and construction technology requirements [59].

Another stream of research develops regional planning strategies using detailed representations of programs and/or planning units. Urban renewal programs are designed based on the assignment of specific building types, levels, and prices to land parcels to optimize net social benefit [60]. A multiobjective land acquisition problem is solved on a grid to optimize compactness, acquisition, and development cost as well as land area [61], and proximity to high- and low-amenity communities and compactness [62]. A multicriteria planning problem for “smart growth” using actual, nonuniform land parcels and reflecting the perspectives of a government planner, an environmentalist, a conservationist, and a land developer yields a nonlinear integer multiobjective math program whose nondominated solutions provide specific guidance to diverse decision makers and stakeholders [63].

Tactical planning models for housing construction or redevelopment in specific, localized areas include Ref. 60, as well as other models where spatial concerns are less important. The Analytic Hierarchy Process is used to incorporate customer requirements for industrialized housing [31]. Math programs related to production scheduling

problems are used to design policies for relocating families in public housing communities undergoing renovations, to minimize total development time while ensuring that as few families as possible are displaced [64–66] (see the section titled “Single Machine Scheduling” in this encyclopedia).

Operational models for real estate practice are relatively rare. Multiattribute utility theory is used to rank applicants to a large commercial development; such a model could easily be applied to residential housing [67]. An application of queueing theory (see the section titled “The $M/M/1$ Queue” in this encyclopedia) evaluates the impacts of race-based versus non-race-based tenant assignment policies in public housing on levels of racial segregation and waiting times for available units [68].

Decision Support Systems

Research applications of DSS for housing and regional planning are numerous (see Ref. 69), but those focusing specifically on affordable housing and/or sustainable communities are less common. Computerized applications assist the US Army in forecasting demand and allocating resources for military housing [70,71]. Using an application for housing mobility planning [54] as the model engine, a spatial decision support system (SDSS) can provide strategic guidance to planners and counselors regarding housing mobility policy design [72]. A “proof-of-concept” DSS for individual housing mobility counselling is developed using the Analytic Hierarchy Process [73]; this work is extended in Ref. 74 to develop a Web-accessible prototype SDSS that reflects the needs of housing clients, counselors, and landlords. A spatial DSS supports property management and sales, as well as identification of potential properties for purchase [75].

SYNTHESIS OF PREVIOUS RESEARCH

There is a strong base of evidence from descriptive models upon which to develop prescriptive models and DSS for housing and community development planning, construction, and management. While there

is no single “canonical” prescriptive model that captures all elements of housing and community development, a generic multiobjective math programming model for this purpose might be presented as follows. Suppose $\mathbf{X}$ is a vector of decision variables that represent the location, type, and number of housing units to be built. Let $B(\mathbf{X})$ represent the net social benefit associated with housing provision strategy $\mathbf{X}$. $B(\cdot)$ can be nonlinear and capture impacts to residents of the housing under consideration, other community residents, and taxpayers. Let $E(\mathbf{X})$ represent the environmental impacts of housing provision strategy $\mathbf{X}$. Let $F(\mathbf{X})$ represent the perceived fairness, or equity, of the housing provision strategy. Then, a social planner would wish to solve

$$\text{Optimize } \{B(\mathbf{X}), E(\mathbf{X}), F(\mathbf{X})\}, \quad (1)$$

$$\text{s.t. } C(\mathbf{X}) \leq B, \quad (2)$$

$$D_{\text{Min}} \leq N(\mathbf{X}) \leq D_{\text{Max}}, \quad (3)$$

$$\mathbf{X} \in S, \quad (4)$$

where objective (1) balances social, environmental, and political considerations, constraint (2) ensures that development strategies obey resource limitations, and constraint (3) ensures that the level of housing provided lies within given estimates of minimum and maximum demand. Constraint (4) ensures that the decision variables obey all relevant spatial and programmatic constraints. This model could address policy design over multiple periods and incorporate uncertainty in housing markets. A large research literature addresses the solutions of models (1–4): well-known “classical” approaches include the weighting method and the constraint method [76]; more recent approaches include interactive methods, fuzzy optimization, and evolutionary algorithms [77] (see also *Nonlinear Multiobjective Programming*). Models such as 1–4 could be incorporated into SDSS.

As discussed above, other modeling approaches such as agent-based simulation, optimal control, and queueing theory could equally well generate alternative housing provision strategies. The practical utility of these planning models is a function of the

quality of evidence provided by descriptive research which generates insight into the social, environmental, and political impacts of housing and community development policies.

Research in housing and community development described in the previous section and summarized in model 1–4 above faces significant limitations. Prescriptive planning models which are partial-equilibrium, deterministic, and single-period in nature may be insufficiently flexible to address the diversity of problems typical of housing and community development. There does not appear to be a “theory” of affordable housing or sustainable community development that can be adapted to diverse regions, economies or housing types. There are few examples of explicit modeling of stakeholder values to generate decision models that achieve social goals shared across stakeholders (see the section titled “Basic Assumptions, Principles and Axioms of Normative Decision Analysis” in this encyclopedia). Few models attempt to jointly and explicitly address social welfare, equity (see *Fairness and Equity in Societal Decision Analysis*), and environmental impacts. There is little research on models linking strategic, tactical, and operational concerns that correspond to the process of policy planning, design, implementation, and evaluation, though Refs 47, 53, 70, and 72 address separately strategic, tactical, and operational aspects of housing mobility programs. With some exceptions [66,71], most models have neither been applied in the field nor rigorously evaluated according to best practices and specific real-world data.

These limitations motivate many questions that future research might address. Some of these include: How can descriptive models for policy analysis address multiple policy alternatives simultaneously? What types of prescriptive planning models are most appropriate for housing and community development? When are spatial versus non-spatial models most appropriate? What barriers prevent consistent use of prescriptive models? How can prescriptive models explicitly address the “geography of opportunity”?

How can speculative, forward-looking models be validated? How can modelers trade off model realism and detail against tractability?

RESEARCH AGENDA

We discuss a number of promising extensions to the decision modeling research described in this article.

Descriptive Models

It was argued above that a key element of prescriptive models for affordable housing and community development is the formulation and measurement of social welfare. Despite innovative models for measuring regional impacts of housing mobility programs that incorporate scale economies and variations across neighborhood types [39] and estimates of individual-level impacts of housing mobility programs [45], no similar work has addressed similar impacts associated with housing redevelopment programs. In addition, there is no research known to this author that evaluates the full range of social benefits and costs of housing and community development initiatives.

It would be of interest to formulate multistate dynamic models of affordable housing and community development programs that incorporate multiple transitions over space, class, and time associated with policy interventions such as particular assisted housing programs as well as normal class and household mobility; there has been initial limited effort in this area [78]. Another approach could extend current research on agent-based models for residential location, such as given in Ref. 40, to affordable housing and community development policy design. Here, agents could be individual families who participate in a housing mobility or housing redevelopment initiative.

The field of construction engineering and architecture can also benefit from research that applies current technologies to the design of sustainable housing units and communities. Agent-based modeling and cellular automata are applied to the design of virtual cities [79]. These designs are based on simulation models for energy-efficient

residential housing based on data from daily activities of residents [80].

Prescriptive Models

Housing providers and community development experts can choose from many affordable housing planning and management tools [81]. A multicriteria decision analysis could assist organizations in choosing the methods that are best-suited to their technical capabilities, funding streams, and service area characteristics. Another application might adapt methods from human-computer interaction research to evaluate the benefits perceived by housing and community development practitioners of decision models.

US Hurricanes Katrina and Rita in 2005, as well as the Indian Ocean earthquake/tsunami of 2004 are examples of natural disasters that can result in immense human and physical losses. While the urban planning profession has performed admirably in developing redevelopment strategies for New Orleans in the wake of Hurricane Katrina [82], there are opportunities for OR/MS to provide prescriptive planning models for urban redevelopment in the wake of natural disasters (see *Operations Research to Improve Disaster Supply Chain Management* and *Evacuation Planning*). Work by Jung *et al.* [83] is an initial effort in this direction, but much more research is needed.

Affordable housing providers face intense competition for increasingly limited development resources. A regional decision model could provide guidance regarding alternative strategies for collaboration across jurisdictional boundaries, service categories, and client populations.

Of great interest in the current foreclosure crisis is the development of strategies to select foreclosed housing units for purchase, redevelopment and resale, or ongoing management. A multiobjective discrete math programming model for strategy design in foreclosed housing acquisition is developed and applied to a small suburban municipality [84]. Current research extensions motivated by real-world practice include: selecting specific foreclosed housing units while addressing social benefits and costs,

alignment with strategic policy concerns; and extending the multiobjective acquisition model to incorporate uncertainty and risk.

Decision Support Systems

Information technology-enabled decision support is ubiquitous. However, practitioners in affordable housing and community development, especially those serving distressed communities, have limited access to IT applications that might help them make better decisions about the services to use, or provide a means for them to participate in the policy design. SDSS for affordable housing and community development could be developed so that they are easy for inexperienced users to master, use detailed housing market and community-level data, allow users to identify and rank decision alternatives with a variety of methods, and facilitate collaboration between multiple stakeholders. Preliminary analysis [85] has described the prospects of a professional-quality SDSS that might fulfill the promise of the prototype [74]. The application NEO CANDO ties together data from many different sources to support CDC investment decisions [86]. DSS might also be used to facilitate the design of energy-efficient residential units using engineering and architectural best practices. Finally, a recent graduate student service learning project has developed a DSS for management of services for the homeless [87, p. 3].

CONCLUSION

In this article, we have reviewed models for housing and community development, especially those that seek to address issues of affordability and sustainability. The research literature in this area is large, diverse, and long-lived but with substantial opportunities to perform research across disciplines, develop a general theory of prescriptive modeling and decision support, and generate real-world applications. A number of promising extensions may both, shed light on the dynamics of real-world behaviors of actors in the housing and community development process and provide

specific guidance to allow them to more fully enjoy the benefits of affordable housing and sustainable communities.

Acknowledgments

This research was funded in part by the National Science Foundation Faculty Early Career Development (CAREER) Program, "CAREER: Public Sector Decision Modeling for Facility Location and Service Delivery." Thanks to Felicia Sullivan (University of Massachusetts Boston) and Philip Akol, Changmi Jung, Jeannie Kim, and Vincent Chiou (Carnegie Mellon University) for their research assistance.

REFERENCES

1. Joint Center for Housing Studies of Harvard University. 2006. The state of the nation's housing. Available at <http://www.jchs.harvard.edu/publications/markets/son2006/son2006.pdf>. Accessed 2009 Jun 25.
2. Rohe WM, McCarthy G, Van Zandt S. The social benefits and costs of homeownership: a critical assessment of the research. 2001. Joint Center for Housing Studies, Harvard University, Cambridge, MA. Working Paper LIHO 01.12.
3. US Department of Housing and Urban Development. Affordable housing needs 2005: a report to Congress. Office of Policy Development and Research. 2007. Available at <http://www.huduser.org/intercept.asp?loc=/Publications/pdf/AffHsgNeeds.pdf>. Accessed 2009 Jun 5.
4. Millennial Housing Commission. Meeting our nation's housing challenges: a report of the Bipartisan Millennial Housing Commission appointed by the Congress of the United States. 2002. Available at <http://govinfo.library.unt.edu/mhc/MHCReport.pdf>. Accessed 2009 Jun 5.
5. NAACP and National Association of Home Builders. Building on a dream. 2006. Available at http://www.nahb.org/fileUpload_details.aspx?contentTypeID=7&contentID=2858. Accessed 2009 Jun 5.
6. Lippman BJ. A heavy load: the combined housing and transportation burdens of working families. National Housing Conference. Washington, DC: Center for Housing Policy; 2006. Available at http://www.nhc.org/pdf/pub_heavy_load_10_06.pdf. Accessed 2009 Jun 5.
7. Haas PM, Makarewicz C, Benedict A. Center for Neighborhood Technology, Sanchez TW, Dawkins CJ, Virginia Tech). Housing and transportation cost trade-offs and burdens of working households in 28 metros. National Housing Conference. Washington, D.C.: Center for Housing Policy; 2006. Available at <http://www.nhc.org/pdf/chp-pub-hl06-cnt-report.pdf>. Accessed 2009 Jun 5.
8. Cervero R, Chapple K, Landis J Wachs M (Institute for Transportation Studies, University of California, Berkeley). Making do: how working families in seven U.S. metropolitan areas trade off housing costs and commuting times. National Housing Conference. Washington, DC: Center for housing policy; 2006. Available at <http://www.nhc.org/pdf/chp-pub-hl06-berkeley-report.pdf>. Accessed 2009 Jun 5.
9. Fox RK, Treuhaft S, Douglass R. Shared prosperity, stronger regions: an agenda for rebuilding America's older core cities. PolicyLink and community development partnerships network. 2005. Available at <http://www.policylink.org/atf/cf/%7B97c6d565-bb43-406d-a6d5-eca3bbf35af0%7D/SHAREDPROSPERITY-CORECITES-FINAL.PDF>. Accessed 2009 Jun 5.
10. Goetz EG. Clearing the way: deconcentrating the poor in urban America. Washington, DC: Urban Institute Press; 2003.
11. de Souza Briggs X. Politics and policy: changing the geography of opportunity. In: de Souza Briggs X, editor. The geography of opportunity: race and housing choice in metropolitan America. Washington, DC: Brookings Institution Press; 2005. pp. 310–341.
12. National Bureau of Economic Research. Determination of the December 2007 peak in economic activity. Business Cycle Dating Committee, December 11, 2008. Available at <http://www.nber.org/cycles/dec2008.html>. Accessed 2009 Jun 6.
13. Joint Center for Housing Studies of Harvard University. The state of the nation's housing. 2008. Available at <http://www.jchs.harvard.edu/publications/markets/son2008/son2008.pdf>. Accessed 2009 Jun 5.
14. Baker D, Rosnick D. The impact of the housing crash on family wealth. Washington (DC): Center for Economic and Policy Research; 2008. Available at http://www.cepr.net/documents/publications/wealth_2008_07.pdf. Accessed 2009 Jun 6.

15. Atlas J, Dreier P. Stemming the red tide. *Shelterforce* 2008;Spring:8–15.
16. Joint Center for Housing Studies of Harvard University. America's rental housing: the key to a balanced national policy. 2008. Available at http://www.jchs.harvard.edu/publications/rental/rh08_americas_rental_housing/rh08_americas_rental_housing.pdf. Accessed 2009 Jun 5.
17. Olsen EO. Housing programs for low-income households. In: Moffitt RA, editor. *Means-tested transfer programs in the United States*. Chicago (IL): The University of Chicago Press; 2003. pp. 365–441.
18. McIlwain J. Housing for moderate-income households in the European Union and the United States. Washington, DC: Urban Land Institute; 2003.
19. Salama AM, Alshuwaikhat HM. A trans-disciplinary approach for a comprehensive understanding of sustainable affordable housing. *Glob Built Environ Rev* 2006;5(3):35–50. Available at <https://www.edgehill.ac.uk/gber/pdf/vol5/issue3/Article%204.pdf>. Accessed 2009 Jun 5.
20. Priemus H. How to make housing sustainable? The Dutch experience. *Environ Plann B Plann Des* 2005;32(1):5–19.
21. Glaeser EL. Places, people, policies: an agenda for America's urban poor. *Harvard Magazine* 2000;103(2). Available at <http://harvardmagazine.com/2000/11/places-people-policies.html>. Accessed 2009 Jun 5.
22. Pendall R, Nelson AC, Dawkins CJ, *et al.* Connecting smart growth, housing affordability, and racial equality. In: de Souza Briggs X, editor. *The geography of opportunity: race and housing choice in metropolitan America*. Washington, DC: Brookings Institution Press; 2005. pp. 219–246.
23. von Hoffman A. High ambitions: the past and future of American low-income housing policy. *Housing Policy Debate* 1996;7(3):423–446.
24. Bohl CC. New urbanism and the city: potential applications and implications for distressed inner-city neighborhoods. *Housing Policy Debate* 2000;11(4):761–801.
25. Myers D, Gearin E. Current preferences and future demand for denser residential environments. *Housing Policy Debate* 2001;12(4): 633–659.
26. Maher C. Residential mobility, locational disadvantage and spatial inequality in Australian cities. *Urban Policy Res* 1994;12(3): 185–191.
27. Miron JR. Private rental housing: the Canadian experience. *Urban Stud* 1995;32(3): 579–604.
28. Steven Winter Associates, Inc. US Department of Energy. Building America field project: results for the consortium for advanced residential buildings (CARB). Golden, Colorado: National Renewable Energy Laboratory; 2001. Available at <http://www.nrel.gov/docs/fy03osti/31380.pdf>. Accessed 2009 Jun 5.
29. Balcomb JD, Hancock CE, Barker G. US Department of Energy. Design, construction, and performance of the grand canyon house. Golden, Colorado: National Renewable Energy Laboratory. 1999. Available at <http://www.nrel.gov/docs/fy99osti/24767.pdf>. Accessed 2009 Jun 5.
30. Katrakakis JT, Knight PA, Cavallo JD. Energy-efficient rehabilitation of multifamily buildings in the Midwest. Argonne National Laboratory, Decision and Information Sciences Division. 1994. Available at <http://www.osti.gov/bridge/servlets/purl/10186292-Xtn9vf/webviewable/10186292.pdf>. Accessed 2009 Jun 5.
31. Armacost RL, Paul J, Mullens MA, *et al.* An AHP framework for prioritizing customer requirements in QFD: an industrialized housing application. *IIE Trans* 1994;26(4):72–80.
32. Ibrahim R, Nissen M. Emerging technology to model dynamic knowledge creation and flow among construction industry stakeholders during the critical feasibility-entitlements phase. In: Flood I, editor. *Information technology 2003: towards a vision for information technology in civil engineering*. Reston (VA): American Society of Civil Engineers; 2004. pp. 1–14. Available at <http://cedb.asce.org/cgi/WWWdisplay.cgi?0305909>. Accessed 2009 Jun 5.
33. Kelly KL. A systems approach to identifying decisive information for sustainable development. *Eur J Oper Res* 1998;109:452–464.
34. Hjorth P, Bagheri A. Navigating towards sustainable development: a systems dynamics approach. *Futures* 2006;38(1):4–92.
35. Gleeson ME. Renovation of public housing: suggestions from a simple model. *Manage Sci* 1992;38(5):655–666.
36. Midgley G, Reynolds M. Systems/operational research and sustainable development: towards a new agenda. *Sus Dev* 2004;12(1): 56–64.

37. Tiesdell S. Integrating affordable housing within market-rate developments: the design dimension. *Environ Plann B Plann Des* 2004; 31:195–212.
38. Skelton I. Planning under welfare pluralism: social housing targets and allocations in Ontario, Canada. *Geoforum* 1997;28(1): 79–89.
39. Galster GC. Consequences from the redistribution of urban poverty during the 1990s: a cautionary tale. *Econ Dev Q* 2005;19(2): 119–125.
40. Brown DG, Robinson DT. Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl. *Ecol Soc* 2006;11(1):22. Available at <http://www.ecologyandsociety.org/vol11/iss1/art46/ES-2006-1749.pdf>. Accessed 2009 Jun 8.
41. Gleeson ME. Estimating housing mortality from loss records. *Environ Plann A* 1985;17:647–659.
42. Johnstone IM. Development of a model to estimate the benefit–cost ratio performance of housing. *Constr Manage Econ* 2004;22:607–617.
43. Day L. Choosing a house: the effect of cost constraints on single-family house design and construction. *Environ Plann B Plann Des* 1995;22:603–622.
44. Paris DE, Kangari R. Multifamily affordable housing: residential satisfaction. *J Perform Construct Facility* 2005;19(2):138–145.
45. Johnson MP, Ladd HF, Ludwig J. The benefits and costs of residential-mobility programs for the poor. *Housing Stud* 2002;17(1):125–138.
46. Quercia RG, McCarthy GW, Wachter SM. The impacts of affordable lending efforts on homeownership rates. *J Housing Econ* 2003;12(1):29–59.
47. Tsuruel Somerville C, Mayer CJ. Government regulation and changes in the affordable housing stock. *FBRNY Econ Policy Rev* 2003;9(2):45–62.
48. Nikolopoulos CV, Tzanetis DE. A model for housing allocation of a homeless population due to a natural disaster. *Nonlinear Anal Real World Appl* 2003;4(4):561–579.
49. Caulkins JP, Feichtinger G, Grass D, *et al.* Placing the poor while keeping the rich in their place: separating strategies for optimally managing residential mobility and assimilation. *Demogr Res* 2005;13(1):1–34.
50. Caulkins JP, Feichtinger G, Johnson MP, *et al.* Skiba thresholds in a model of controlled migration. *J Econ Behav Organ* 2005;57(4):490–508.
51. Brotchie JF. A new approach to urban modelling. *Manage Sci* 1978;24(16):1753–1758.
52. Frech GM, Thyagarajan S. Housing allocation model for low-income and moderate-income households in Albany Schenectady Troy SMSA. *J Environ Syst* 1975;5(1):59–77.
53. Kim TJ. Alternative model for fair share allocation of lower-income housing. *Socioecon Plann Sci* 1979;13(2):113–116.
54. Johnson MP, Hurter AP. Decision support for a housing relocation program using a multi-objective optimization model. *Manage Sci* 2000;46(12):1569–1584.
55. Johnson MP. Single-period location models for subsidized housing: tenant-based subsidies. *Ann Oper Res* 2003;123:105–124.
56. Johnson MP. Tenant-based subsidized housing location planning under uncertainty. *Socioecon Plann Sci* 2001;35(3):149–173.
57. Johnson MP. Single-period location models for subsidized housing: project-based subsidies. *Socioecon Plann Sci* 2006;40(4):249–274.
58. Johnson MP. Planning models for affordable housing development. *Environ Plann B Plann Des* 2007;34(3):501–523.
59. Tiwari P, Parikh J, Sharma V. Performance evaluation of cost-effective buildings—a cost, emissions and employment pint of view. *Build Environ* 1996;31(1):75–90.
60. Atkins RS, Krokosky EM. Optimum housing allocation model for urban areas. *J Urban Plan Dev Division* 1971;97(1):41–53.
61. Wright J, ReVelle C, Cohon J. A multiobjective integer programming model for the land acquisition problem. *Reg Sci Urban Econ* 1983;13:31–53.
62. Gilbert KC, Holmes DD, Rosenthal RE. A multiobjective discrete optimization model for land allocation. *Manage Sci* 1985;31(12):1509–1522.
63. Gabriel SA, Faria JA, Moglen GE. A multi-objective optimization approach to smart growth in land development. *Socioecon Plann Sci* 2006;40:212–248.
64. Kaplan EH. Relocation models for public housing redevelopment programs. *Environ Plann B Plann Des* 1986;13:5–19.
65. Kaplan EH, Amir A. A fast feasibility test for relocation problems. *Eur J Oper Res* 1987;35(2):201–206.
66. Kaplan EH, Berman O. OR hits the heights: relocation planning at the Orient Heights

- housing project. *Interfaces* 1988;18(6): 14–22.
67. Yau C, Davis T. Using multi-criteria analysis for tenant selection. *Decis Support Syst* 1994;12:233–244.
 68. Kaplan EH. Analyzing tenant assignment policies. *Manage Sci* 1987;33(3):395–408.
 69. Timmermans H, editor. *Decision support systems in urban planning*. London: E & FN SPON; 1997.
 70. Forgionne GA. HANS: a decision support system for military housing managers. *Interfaces* 1991;21(6):37–51.
 71. Forgionne GA, Frager YS. The US Army relies on decision support systems in allocating housing resources. *Interfaces* 1998;28(2):72–79.
 72. Johnson MP. A spatial decision support system prototype for housing mobility program planning. *J Geogr Syst* 2001;3(1):49–67.
 73. Johnson MP. Decision support for family relocation decisions under the section 8 housing assistance program using GIS and the analytic hierarchy process. *J Housing Res* 2002;12(2):277–306.
 74. Johnson MP. Spatial decision support for assisted housing mobility counseling. *Decis Support Syst* 2005;41(1):296–312.
 75. Zeng TQ, Zhou Q, Int J. Optimal decision making using GIS: a prototype of a real estate graphical information system (REGIS). *Geogr Inf Sci* 2001;15(4):207–321.
 76. Cohon JL. *Multiobjective programming and planning*. New York: Academic Press; 1978.
 77. Ehrgott M, Gandibleux X, editors. *Multiple criteria optimization: state-of-the-art annotated bibliographic surveys*. Norwell (MA): Kluwer Academic Publishers; 2002.
 78. Johnson MP, Caulkins JP. Can housing mobility programs make a long-term impact on the lives of poor families and the health of middle-class communities: a policy simulation. *Heinz School Working Paper Series* 2005-33. 2006.
 79. Chiou YS, Davison C. Virtual metropolis for smart growth policy analysis: an agent-based/cellular automata dual-layered complex system model. *EcoCity World Summit*, 2008, San Francisco.
 80. Chiou YS, Carley KM. Characterizing US single-family detached house through social network analysis. Unpublished working paper.
 81. Washington Area Housing Partnership. Toolkit for affordable housing development. Washington, D.C.; 2005. Available at http://www.wahpdc.org/Toolkit_for_Affordable_Housing_Dev_2-17-06.pdf. Accessed 2009 Jun 8.
 82. Hartman C, Squires GD. *There is no such thing as a natural disaster: race, class, and Hurricane Katrina*. New York: Routledge; 2006.
 83. Jung C, Johnson MP, Williams JC. Mathematical models for reconstruction planning in urban areas. *Heinz School Working Paper Series* 2007-10. 2007.
 84. Johnson MP, Turcotte D, Sullivan FM. What foreclosed homes should a municipality purchase to stabilize vulnerable neighborhoods? *Netw Spat Econ* 2009. In press.
 85. Johnson MP. Can a spatial decision support system improve low-income service delivery? Analysis of tools and requirements for a computer-assisted mobility counseling system. *Heinz School Working Paper Series* 2005–32. 2005.
 86. Case Western Reserve University, NEO CANDO. Mandel School of Applied Social Sciences. Available at <http://neocando.case.edu/cando/index.jsp>. Accessed 2010 May 13.
 87. Nesmith R. Working out for the best: panel discussion highlights students presenting computing for good projects. *The Whistle* 34(32). p. 3. Available at www.whistle.gatech.edu/archives/09/dec/14/dec14.pdf. Accessed 2010 Jan 3.

FURTHER READING

- Boardman AE, Greenberg DH, Vining AR, *et al*. *Cost-benefit analysis: concepts and practice*. 3rd ed. Upper Saddle River (NJ): Pearson Education, Inc.; 2006.
- Eiselt HA, Sandblom C-L. *Decision analysis, location models, and scheduling problems*. Berlin: Springer; 2004.
- Fotheringham AS, O'Kelly ME. *Spatial interaction models: formulations and applications*. Dordrecht: Kluwer Academic Publishers; 1989.
- Green RK, Malpezzi S. *A primer on US housing markets and housing policy*. Washington, DC: The Urban Institute Press; 2003.
- Hannon B, Ruth M. *Dynamic modeling*. 2nd ed. New York: Springer; 2001.
- Johnson MP, Smilowitz K. Community-based operations research. In: Klastorin T, editor. *Tutorials in operations research* 2007. Hanover (MD):

- Institute for Operations Research and the Management Sciences; 2007. pp. 102–123.
- Keeney RL. Value-focused thinking: a path to creative decisionmaking. Cambridge (MA): Harvard University Press; 1992.
- Larson RC, Odoni AR. Urban operations research. New Jersey: Prentice-Hall; 1981. Available at http://web.mit.edu/urban_or_book/www/book/. Accessed 2009 Jun 26.
- Malczewski J. GIS and multicriteria decision analysis. New York: John Wiley & Sons, Inc.; 1999.
- Marianov V, Serra D. Location problems in the public sector. In: Drezner Z, Hamacher HW, editors. Facility location: applications and theory. Berlin: Springer; 2004. pp. 119–150.
- Pollock SM, Rothkopf MH, Barnett A, editors. Handbooks in operations research and management science. Volume 6, Operations Research and the Public Sector. Amsterdam: North-Holland; 1994.

HYPER-HEURISTICS

GRAHAM KENDALL
EDMUND K. BURKE
University of Nottingham

INTRODUCTION

The purpose of this article is to present a brief overview of hyper-heuristics, which can be viewed as *heuristics to choose heuristics* [1] or as *heuristics to generate heuristics* [2]. This article provides an introduction to these two approaches, in addition to highlighting some of the relevant literature. For a comparison between meta-heuristics and hyper-heuristics, the reader is referred to Burke and Kendall [3], which provides tutorials on many common meta-heuristics as well as describing hyper-heuristics [4].

One of the key motivations of hyper-heuristic research is to raise the level of generality at which search algorithms operate, with a long-term research objective of being able to build search methodologies that are able to adapt themselves to different problem instances/domains with minimum human input. Figure 1 shows the current state of search methodologies, where the research effort with respect to hyper-heuristics is concentrated in the central position.

In addition, the study of hyper-heuristics raises a number of other research challenges. For example,

- To what extent can hyper-heuristic algorithms be self-tuning? One of the criticisms that could be levelled at some meta-heuristic approaches is that they require significant tuning for specific problem instance(s).
- Significant recent hyper-heuristic research has concentrated on *heuristics to choose heuristics*. While there are still many research directions to be investigated in this area, there is growing

interest in automating the process of creating or generating heuristics. This has received relatively little attention to date (but there has been a recent focus on using genetic programming to generate heuristics). This area is likely to become more important as it will enable users to utilise hyper-heuristics without having to implement a set of low-level heuristics.

- Hyper-heuristic research opens up the possibility of having an off the shelf system, which is able to address many different problem types, with minimal user intervention. A key research challenge that needs to be addressed to enable this type of system is how the problem can be represented.

As mentioned above, an important motivating goal of hyper-heuristic research is to raise the level of generality of search systems. However, it would be naïve to propose one general purpose search algorithm [5]. There is, nonetheless, significant scope to investigate more general search algorithms that are able to operate across a wider range of problems than is currently possible using bespoke algorithms, which are typically designed with a single-problem domain in mind, or even just a single-problem instance. Whilst there is much scientific merit in doing this, the algorithm is often only ever suitable for this domain and, possibly, just the particular benchmark instance(s) under consideration. Rather than attempting to get better and better solutions on a specific problem, one of the goals of hyper-heuristic research is to be able to get high-quality solutions across a range of problems, even if they are not as good as a bespoke algorithm on a single problem. Of course, we are still interested in finding solutions that are of as high a quality level as possible, recognising that if the main objective is to find the highest quality solution possible, for a given problem, then a bespoke algorithm is probably more appropriate. Therefore,

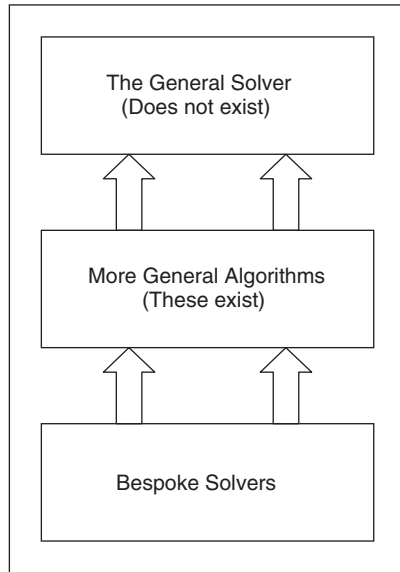


Figure 1. The drive toward more general search algorithms.

the goal is to build a methodology that can operate across different problem types without human recoding, redesign, or parameter setting. The solutions that are returned by a hyper-heuristic should, of course, be acceptable for an end user. However, the cost of producing the solution also needs to be acceptable as problem-specific approaches can be resource intensive (in terms of developmental time for other problems), especially if the software cannot be reused. A motivation of this research area is to recognise that this cost-quality trade-off underpins a series of key scientific challenges.

After briefly discussing the historical roots of hyper-heuristic research, we present the two major research directions that are currently being undertaken. The first, selects a heuristic from a supplied set to be applied to the current problem state. The second direction has the goal of generating heuristics. This is a less mature area, and we discuss the main research breakthroughs that have been made.

As hyper-heuristic algorithms are often implementations of meta-heuristics on heuristic search spaces we also provide

references to the relevant literature from this area.

There are several introductory chapters/papers which would be useful starting points for those new to the area [4,6–8].

HISTORICAL ROOTS

Although the term hyper-heuristics is relatively new, being first used in a published paper in 1997 [9], its roots can be traced back to the 1960s [10–12]. Work carried out more recently (but still not using the term hyper-heuristics) includes Dorndorf and Pesch [13], Hart [14,15], Schaffer and Eshelman [16], and Fang [17]. In Hart [14,15], for example, a two stage genetic algorithm [18] was employed. The genetic algorithm representation utilised both a direct and an indirect encoding, with the indirect encoding representing a sequence of heuristics in order to derive the schedule. This approach was shown to produce good quality solutions on a real-world scheduling problem, and was able to do so in a few minutes. Ross [19] studied a hyper-heuristic approach to the bin packing problem. Their hyper-heuristic was a Learning Classifier System [20]. Benchmark bin packing problems were initially solved using a set of heuristics taken from the literature. The learning classifier system was then used to find rules to associate each heuristic with the problem state, after which the problem state is consulted to decide which heuristic to apply next. This alters the problem state, so calling for another heuristic. When comparing the hyper-heuristic against using each heuristic in isolation, it was able to solve 78% of the benchmark problems to optimality and the best performing heuristic was only able to solve 73% to optimality. Bin packing was also the subject of Ross [1], with the hyper-heuristic now based on a genetic algorithm. The hyper-heuristic solved 80% of the benchmark datasets to optimality; slightly outperforming the learning classifier of Ross [19] and, again, producing superior results when using single heuristics in isolation. In Ross [21], a genetic algorithm based hyper-heuristic is utilised to construct

timetables for class and examination problems. The construction methodology does not involve any search or backtracking, and so the run time is predictable. The genetic algorithm searches over a set of heuristics, drawn from the literature, and decides which heuristic to invoke at each decision point. The results show that this hyper-heuristic performs well across a range of problem instances. Potvin [22] utilise a slightly different approach, to those described above, for vehicle routing problems. Their system, Micro-ALTO, supports algorithm designers via a graphical, interactive environment in order for them to specify problem solving strategies. We also note that adaptive algorithms have been in existence for some time. One well-known example is the adaptive strategy of Rechenberg [23], where control parameters were adjusted at run time. Although not a hyper-heuristic in the way that we have defined them, it does represent an algorithm that can adapt to its environment.

The first journal paper to use the term hyper-heuristic [24] investigated a tabu search based methodology across two different problems (university course timetabling and personnel rostering). The motivation for applying the methodology across two different domains was to test the hypothesis that the *same* algorithm is able to produce good results on two distinct problem types. The hyper-heuristic was based on tabu Search [25,26], together with a reinforcement learning [27] mechanism. The method performed well when compared against bespoke algorithms for these two problems.

HEURISTICS TO CHOOSE HEURISTICS

Many of the earlier hyper-heuristic approaches have a set of heuristics available that the hyper-heuristic is able to call upon, and the problem reduces to which heuristic to call at each decision point. In order to address the motivation of attempting to be as general as possible, a potential research direction is to build a hyper-heuristic approach which has no knowledge of the problem domain. In this situation, the hyper-heuristic only has a number of heuristics available and these can be called at any time. It only has to decide which heuristic to call. Figure 2 shows a suggested hyper-heuristic framework. At the top level, there is the hyper-heuristic domain. This is separated from the problem domain by a *domain barrier*. This restricts the hyper-heuristic from *knowing* anything about the domain on which it is operating, although there is no reason why we could not (or should not) *leak* some problem-specific information through this barrier with the goal of sacrificing *some* generality for *more* effectiveness. The implication is that we are able to switch the problem domain but still utilise the same hyper-heuristic algorithm while minimising other changes. That is, the *same* search algorithm can be used at the top level, with problem domain information being relegated to the lower level, so that it can be changed in order to address a different problem in the expectation that the hyper-heuristic will adapt itself to the new problem that has been presented.

Information that might be allowed to pass through the domain barrier could include

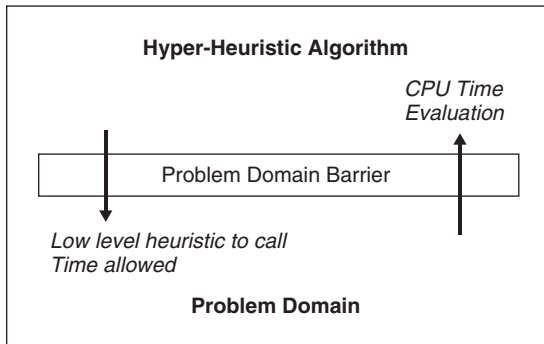


Figure 2. A hyper-heuristic framework.
Source: Adapted from Cowling *et al.* [28].

changes in the evaluation function, the execution time of a given heuristic, the past performance of each heuristic, and so on. The problem domain (lower level) of the framework consists of three distinct elements.

Set of Low-Level Heuristics

The heuristics that are provided as a part of the framework are typically drawn from existing heuristics in the literature. Typically, the framework is provided with as many heuristics as the researcher thinks are relevant, with a bias often being to provide too many, rather than too few. This bias is recommended because if a heuristic is not to be useful to the search, then an effective hyper-heuristic should adapt and not select that heuristic anyway.

By employing a standard template for each heuristic, the problem domain can be altered with no changes being required at the higher level of the framework.

One suggested, C++ style interface, is presented here

```
bool H1 (double & evaluation, int CPU time,
int iterations);
```

where

- The return value (Boolean) is true if the heuristic has modified the current solution in such a way that it satisfies all the hard constraints of the problem. That is, the current solution state represents a feasible solution. If the current solution is infeasible, then false is returned.
- The evaluation parameter is a reference parameter, so that the heuristic is able to return the evaluation of the current solution.
- The CPU time parameter specifies the amount of time the higher level strategy is allocating to H_n . Once that time limit has been exceeded, the heuristic must terminate and return a solution.
- The iterations parameter determines how many times the heuristic should be executed before returning.

Of course, this template is just one possibility. For example, the interface could

be modified so that it can search a search space which includes infeasible solutions which might be beneficial, if not essential, for some problems.

Representation of the Problem

Although the higher level hyper-heuristic often has no information about the problem on which it is operating, the framework must still represent the problem at the lower level. The most suitable representation can be derived from the scientific literature. The goal of developing representations which have some level of generality represents a major scientific challenge.

Evaluation Function

Utilising the framework shown in Fig. 2, various hyper-heuristics have been investigated. The following presents some of these research directions. It is not a comprehensive list, but a representative sample from recent work.

In Cowling [28] and Kendall [29], a *choice function* was used. This rates heuristic selection in three ways.

1. The performance of individual heuristics, rewarding heuristics which perform well and punishing poorly performing heuristics.
2. The performance of pairs of heuristics, on the basis that a heuristic may perform poorly but may be required in order to improve the performance of another heuristic.
3. How long is it since a heuristic has been called? This measure is a diversification strategy to ensure that a heuristic that has not been called for some time has the opportunity to be called again, even though it has not performed well in the past.

The choice function method was shown to be effective when applied to three different real-world problems.

A case-based heuristic selection method was investigated in Burke [30]. The paper is motivated by the idea of storing previously solved problems, together with the heuristics that were used on that problem. When

presented with a new problem, a similar problem is found in the case-base and the same heuristics are utilised.

The focus of the work in Bai [31] is shelf-space allocation within supermarkets. It utilises a variety of methodologies, including greedy heuristics, GRASP (greedy randomised adaptive search procedure) [32], and simulated annealing [33], to produce results against which the hyper-heuristic is compared against. The major contribution of Bai [31] is the presentation of three different hyper-heuristics, a tabu search based hyper-heuristic based on Burke [24], a simulated annealing hyper-heuristic that was proposed in Bai and Kendall [34], and a hybridisation of these two hyper-heuristics. All the three hyper-heuristics use the same set of low-level heuristics, which comprised four types of n -opt exchange operators. The results demonstrate that the hyper-heuristic approaches produced the same results, or were superior to the other methods presented in the paper.

There have been a number of studies which have utilised tabu search as a hyper-heuristic. One tabu search approach [1] has been discussed above (see the section titled “Introduction”), but other papers [35,36] also draw on the idea that certain heuristics should be made *tabu* at a given time in order to stop them being called.

Other hyper-heuristic approaches include those based on genetic algorithms [37], graphs [38,39], and simulated annealing [40].

HEURISTICS TO GENERATE HEURISTICS

One of the criticisms that can be levelled at approaches which are based around *heuristics to choose heuristics* is that for each problem domain, the researcher (or user) has to define and implement a new set of low-level heuristics. This is not only time consuming but it also often requires knowledge of the area and the ability to access and interpret the scientific literature.

In order to address this issue, there has been some work recently on automatically generating heuristics. There are at least two motivations for doing this.

1. If we are presented with a problem that does not have a suitable heuristic available (or we do not have the time/money to develop a suitable heuristic ourselves), we may be able to automatically generate a heuristic that is able to produce a good quality solution to the problem at hand.
2. We may be able to automatically build new, perhaps unintuitive, heuristics that we can use as our set of low-level heuristics.

The work we discuss in this section starts to address point 1. The research for point 2 is still in its very early stages.

The approach described in Burke and Newall [41] generated heuristic orderings for examination timetabling, which can be viewed as finding new heuristics. The approach drew upon the squeaky wheel optimisation method [42] and was applied to examination timetabling. The work is motivated by the following observations. When scheduling examinations, a common method is to sort the examinations into some order (representing a view of what is most difficult), and then schedule the examinations in the order they appear in the list. The approach in Burke and Newall [41] considers a heuristic as being such an ordering. It uses human designed heuristics as a starting point, and looks at those exams which cause the highest penalty and promotes them up the order. At the end of the algorithm, a new heuristic (ordering) has been generated. In Burke and Newall [41], this was referred to as an *adaptive* heuristic. The method was shown to be effective on a range of benchmarks instances.

Another research initiative for generating heuristics is based on Genetic Programming [43], applied to n -dimensional packing problems [44]. In this approach [2], genetic programming is utilised to evolve a function that determines into which bin to place a given piece. This is done by placing an item into a bin and using this problem state as input to the genetic programming function. The function is applied to all item/bin pairs, and the item/pair which returns the maximal value is the placement that is used. The approach is competitive with human designed heuristics

but has also been developed to operate across different dimensions of the problem (one-, two-, and three-dimensional), where only the problem domain changes and the hyper-heuristic remains constant.

CONCLUSIONS

We have presented a very brief introduction and overview of hyper-heuristics, including references to some of the historical work which did not use the term, but which can be recognised as representing the scientific roots of the area.

The article has focussed on two types of hyper-heuristics [6]. *Heuristics to choose heuristics* select the next heuristic to apply from a given set of (usually) human designed heuristics. As one of the key aims of this methodology is to raise the level of generality, the hyper-heuristic algorithm is often separated from the low-level heuristics, and the evaluation function, so that we can switch the low-level heuristics and so address a different problem domain.

An emerging research direction is represented by *heuristics to generate heuristics*. To date there has been limited work in this area, mainly focusing on evolving heuristics using genetic programming. The aim of this approach is to generate heuristics which a human could not (or would not have the time/resources) design.

Although hyper-heuristics have their roots in the 1960s, they have only recently started to receive significant research interest. We hope that this article has provided the reader with a starting point on which to increase their knowledge and understanding of hyper-heuristics and that it enables other researchers to make a contribution to this still emerging area.

REFERENCES

1. Ross P, Marin-Blázquez JG, Schulenburg S, Hart E. Learning a procedure that can solve hard bin-packing problems: a new GA-based approach to hyper-heuristics. Proceedings of the genetic and evolutionary computation conference. Chicago, Illinois, USA: Springer; 2003. pp. 1295–1306.
2. Burke EK, Hyde M, Kendall G, Woodward J. A genetic programming hyper-heuristic approach for evolving two dimensional strip packing heuristics. IEEE Trans Evol Comput. In press (DOI: 10.1109/TEVC.2010.2041061).
3. Burke EK, Kendall G, editors. Search methodologies: introductory tutorials in optimization and decision support techniques. New York, USA: Springer; 2005.
4. Ross P Hyper-heuristics. In: Burke EK, Kendall G, editors. Search methodologies: introductory tutorials in optimization and decision support techniques. New York, USA: Springer; 2005. Chapter 17. pp. 529–556.
5. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Trans Evol Comput 1997;1(1):67–82.
6. Burke E, Hyde M, Kendall G, Ochoa G, Özcan E, Woodward J. A classification of hyper-heuristics approaches. Technical Report No. NOTCS-TR-SUB-0907061259-5808. Nottingham: School of Computer Science and Information Technology, University of Nottingham Computer Science; 2009. (to appear in Handbook of Meta-heuristics).
7. Burke EK, Hyde M, Kendall G, Ochoa G, Özcan E, Qu R. A survey of hyper-heuristics, school of computer science and information technology. Computer Science Technical Report No. NOTCS-TR-SUB-0906241418-2747. Nottingham, UK: University of Nottingham; 2009.
8. Burke EK, Hart E, Kendall G, Newall J., Ross P, Schulenburg S. In: Glover F, Kochenberger G, editors. Handbook of metaheuristics. 2003. pp. 457–474.
9. Denzinger J, Fuchs M. High performance ATP systems by combining several AI methods. Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI 97). 1997. Nagoya, Aichi, Japan, pp. 102–107.
10. Fisher H, Thompson GL. Probabilistic learning combinations of local job-shop scheduling rules. Factory Scheduling Conference; 1961 May 10–12. Carnegie Institute of Technology, USA, pp. 10–12.
11. Fisher H, Thompson GL. Probabilistic learning combinations of local job-shop scheduling rules. In: Muth JF, Thompson GL, editors. Industrial scheduling. New Jersey: Prentice-Hall, Inc.; 1963. pp. 225–251.
12. Crowston WB, Glover F, Thompson GL, Trawick JD. Probabilistic and parametric learning combinations of local job shop scheduling rules. Pittsburgh (PA): GSIA, Carnegie Mellon

- University; 1963. (ONR Research memorandum No. 117).
13. Dorndorf U, Pesch E. Evolution based learning in a job shop scheduling environment. *Comput Oper Res* 1995;22(1):2540.
 14. Hart E, Ross P, Nelson JAD. Scheduling chicken catching - an investigation into the success of a genetic algorithm on a real world scheduling problem. *Ann Oper Res* 1999;92(0):363–380.
 15. Hart E, Ross P, Nelson JAD. Solving a real-world problem using an evolving heuristically driven schedule builder. *Evol Comput* 1998;6(1):61–80.
 16. Schaffer JD, Eshelman LJ. Combinatorial optimisation by genetic algorithms: the value of the genotype/phenotype distinction. 1st International Conference on Evolutionary Computation and its Applications EvCA'96. Springer; 1996. pp. 110–120. (Presidium of the Russian Academy of Sciences).
 17. Fang H-L, Ross PM, Corne D. A promising hybrid GA/Heuristic approach for open-shop scheduling problems. In: Cohn A, editor. *Proceedings of ECAI 94: 11th European Conference on Artificial Intelligence*. Amsterdam, The Netherlands: John Wiley & Sons, Ltd.; 1994. pp. 590–594.
 18. Sastry K, Goldberg D, Kendall G. Genetic algorithms. In: Burke EK, Kendall G, editors. *Search methodologies: introductory tutorials in optimization and decision support techniques*. New York, USA: Springer; 2005. Chapter 4. pp. 97–125.
 19. Ross P, Schulenburg S, Marin-Blázquez JG, Hart E. Hyper-heuristics: learning to combine simple heuristics in bin-packing problem. *Proceedings of the Genetic and Evolutionary Computation Conference*. New York, USA, 2002. pp. 215–226.
 20. Sigaud O, Wilson SW. Learning classifier systems: a survey. *Soft Comput Fusion Found Methodol Appl* 2007;11(11):1065–1078.
 21. Ross P, Marin-Blázquez JG, Hart E. Hyper-heuristics applied to class and exam timetabling problems. *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*. Portland, Oregon, USA: IEEE Press; 2004. pp. 1691–1698.
 22. Potvin J-Y, Lapalme G, Rousseau J-M. A microcomputer assistant for the development of vehicle routing and scheduling heuristics. *Decis Support Syst* 1994;12(1):41–56.
 23. Rechenberg I. *Evolutionsstrategie. Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Technical University of Berlin, Germany: Frommann Holzboog; 1973. ISBN 3-7728-0373-3. (reprint of dissertation from 1970).
 24. Burke EK, Kendall G, Soubeiga E. A Tabu-search hyperheuristic for timetabling and rostering. *J Heuristics* 2003;9(6):451–470.
 25. Glover F, Laguna M. *Tabu search*. Massachusetts, USA: Kluwer Academic Publishers; 1998.
 26. Gendreau M, Potvin J-Y. Tabu Search. In: Burke EK, Kendall G, editors. *Search methodologies: introductory tutorials in optimization and decision support techniques*. New York, USA: Springer; 2005. Chapter 6. pp. 165–186.
 27. Gosavi A. Reinforcement learning: a tutorial survey and recent advances. *INFORMS J Comput* 2009;21(2):178–192.
 28. Cowling P, Kendall G, Soubeiga E. A hyper-heuristic approach for scheduling a sales summit. *Selected Papers of the 3rd International Conference on the Practice and Theory of Automated Timetabling, PATAT 2000, LNCS 2079*. Konstanz Germany: Springer; 2000. pp. 176–190.
 29. Kendall G, Soubeiga E, Cowling P. Choice function and random hyperheuristics. *Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning*. pp. 667–671.
 30. Burke EK, Petrovic S, Qu R. Cased-based heuristic selection for timetabling problems. *J Schedul* 2006;9(2):115–132.
 31. Bai R, Burke EK, Kendall G. Heuristic, meta-heuristic and hyper-heuristic approaches for fresh produce inventory control and shelf space allocation. *J Oper Res Soc* 2008;59(10):1387–1397.
 32. Feo TA, Resende MGC. Greedy randomized adaptive search procedures. *J Glob Optim* 1995;6(2):109–133.
 33. Aarts E, Korst J, Michiels W. Simulated annealing. In: Burke EK, Kendall G, editors. *Search methodologies: introductory tutorials in optimization and decision support techniques*. New York, USA: Springer; 2005. Chapter 7. pp. 187–210.
 34. Bai R, Kendall G. An investigation of automated planograms using a simulated annealing based hyper-heuristics. In: Ibaraki T, Nonobe K, Yagiura M, editors. *Volume 32, Metaheuristics: progress as real problem solvers. Operations research/computer science interfaces series*. New York, USA: Springer; 2005. pp. 87–108.

35. Kendall G, Hussin MA. Tabu search hyper-heuristic approach to the examination timetabling problem at the MARA University of Technology. In: Burke E, Trick T, editors. Practice and theory of automated timetabling V (PATAT 2004). Lecture notes in computer science 3616. Pittsburgh: Springer; 2005. pp. 270–293. (Revised Selected Papers of the 5th International Conference).
36. Han L, Kendall G. Investigation of a tabu assisted hyper-heuristic genetic algorithm. Proceedings of Congress on evolutionary computation. Berlin, Germany: IEEE Press; pp. 2230–2237.
37. Cowling P, Kendall G, Han L. An investigation of a hyper-heuristic genetic algorithm. Proceedings of Congress on evolutionary computation. Hawaii, USA: IEEE Press; 2002. pp. 2230–2237.
38. Qu R, Burke EK. Hybridisations within a graph based hyper-heuristic framework for university timetabling problems. *J Oper Res Soc* 2009;60:1273–1285.
39. Burke EK, McCollum B, Meisels A, Petrovic S, Qu R. A graph-based hyper-heuristic for timetabling problems. *Eur J Oper Res* 2007;176(1):177–192.
40. Dowsland K, Soubeiga E, Burke EK. A simulated annealing hyper-heuristic for determining shipper sizes. *Eur J Oper Res* 2007;179(3):759–774.
41. Burke EK, Newall JP. Solving examination timetabling problems through adaptation of heuristic orderings. *Ann Oper Res* 2004;129:107–134.
42. Joslin DE, Clements DP. Squeaky wheel optimization. *J Artif Intell Res* 1999;10:353–373.
43. Koza JR, Poli R. Genetic programming. In: Burke EK, Kendall G, editors. Search methodologies: introductory tutorials in optimization and decision support techniques. New York, USA: Springer; 2005. Chapter 5. pp. 127–164.
44. Martello S, Toth P. Knapsack problems: algorithms and computer implementations. Chichester, UK: John Wiley & Sons; 1990.

ICE HOCKEY

ARMANN INGOLFSSON
School of Business,
University of Alberta,
Edmonton, Alberta,
Canada

Ice hockey is a popular team sport, particularly in cold climates such as in Canada, the northern part of the United States, Russia, and northern Europe. The primary professional ice hockey league is the National Hockey League (NHL), which operates in the United States and Canada and draws players from throughout the world. As it is the case with *baseball*, *football* (see ***American Football: Rules and Research***), *soccer* (see ***Soccer/World Football***), and *basketball*, various operations research (OR) methods and statistical analyses have been used to address questions of interest to ice hockey fans, coaches, and managers. This article outlines published work on such questions and elaborates on two issues: (i) determining the optimal time to “pull the goalie” near the end of an ice hockey game and (ii) estimating the probability that a team will qualify for the NHL playoffs. Rules vary among ice hockey leagues, but all the descriptions and empirical results described in this article refer to the NHL, unless otherwise noted.

An ice hockey team normally has six players on the ice during a game: a goaltender (“goalie”), two defensemen, and three forwards (center, right wing, and left wing). Players rotate onto the ice in shifts of about 1 min. The players use sticks to propel the “puck” (a black rubber disc), with the aim of scoring goals by shooting the puck into the opposing team’s goal. Penalties to players for rule infractions are common (an average 2007–2008 NHL game had 27.3 penalty minutes), resulting in the penalized team playing “shorthanded”; the opposing team is said to be on the “power play.” A game lasts for 60 min, comprising three 20-min periods. During the NHL’s regular season

of 82 games for each team, if the score is tied after three periods (so-called “regulation time”), the teams play one overtime period that lasts for 5 min or until one of the teams scores. If the score remains tied after the 5-min overtime period, the game outcome is determined by a “shootout,” during which a chosen set of players from the two teams alternate in taking penalty shots (where a single skater attempts to score on an undefended goalie). The overtime rules change during the NHL playoffs in that overtime continues, in 20-min periods, until one team scores a goal. During the regular season, a win is worth two points and an overtime loss is worth one point in the standings. A loss during regulation time is worth no points.

The NHL has 30 teams, divided into two 15-team Conferences, each of which is subdivided into three five-team Divisions. The top eight teams from each Conference advance to the playoffs, in which pairs of teams play a series of a maximum of seven games, with the first team to win four games advancing to the next round of the playoffs. The top team from each Conference plays in the final series, and the winning team is awarded the Stanley Cup.

Researchers from various fields, including OR, statistics, economics, and management strategy, have developed mathematical models of ice hockey. Although I focus on OR models, I touch on contributions from these other fields. The interplay between statistical analysis and decision modeling is important. For example, the utility of models to choose an optimal player order for a shootout [e.g., Ref. 1] depends on how precisely the shootout scoring probabilities of different players are known.

POISSON GOAL SCORING PROCESSES

Perhaps the most important and useful modeling device for understanding the dynamics of hockey games is the assumption that scoring follows a *Poisson process* (see ***Poisson Process and its Generalizations***), an

assumption that can be justified by taking a first-principles approach. Call the teams in a hockey game A and B and imagine dividing the 60 min of regulation play into intervals of, say, $\Delta t = 30$ s. Assume that during each interval, Team A scores with probability $\alpha \Delta t$ and does not score with probability $1 - \alpha \Delta t$, independent of scoring in other intervals and independent of whether or not Team B scores in the interval. The parameter α is Team A's goal-scoring rate. Given these assumptions, N_A , the number of goals that Team A scores in the game, will follow a binomial distribution with mean αT , where T is the length of the game. Letting Δt approach zero, the binomial distribution approaches a Poisson distribution with mean αT [e.g., (2, Section 5.2)]. A symmetric set of assumptions for Team B implies that its scoring will follow a Poisson process with rate β , which implies that, N_B , the number of goals that Team B scores in the game, has a Poisson distribution with rate βT . Thus, we have reduced the dynamics of the game to two independent and competing Poisson processes.

Before scrutinizing the validity of the assumptions that led to the Poisson processes, I illustrate their usefulness by demonstrating how one can compute the probability that Team A wins the game during "regular time," that is, before overtime. (Ref. 3 describes an alternative approach, based on the same assumptions.) Starting from the law of total probability and the property that the superposition of two independent Poisson processes is a Poisson process, we see that

$$\begin{aligned} \Pr\{A \text{ wins}\} &= \Pr\{N_A > N_B\} \\ &= \sum_{n=1}^{\infty} \Pr\{N_A > N_B \mid N_A + N_B = n\} \\ &\quad \times \frac{((\alpha + \beta)T)^n e^{-(\alpha + \beta)T}}{n!}. \end{aligned} \quad (1)$$

Now, let us view N_A as being obtained from $N_A + N_B$ by thinning; that is, whenever a goal is scored, Team A scored the goal with probability $p = \alpha/(\alpha + \beta)$, independent of all other goals. This implies that if the total number of goals in the game

is n , then the number of goals scored by Team A follows a binomial distribution with parameters n and p . Therefore, the term $\Pr\{N_A > N_B \mid N_A + N_B = n\}$ equals the probability that a binomial random variable with parameters n and p is larger than $n/2$, which is easily calculated. We approximate the infinite series in Equation (1) with the sum of the first m terms. The resulting truncation error is bounded by $\sum_{n=m+1}^{\infty} ((\alpha + \beta)T)^n \exp(-(\alpha + \beta)T)/n!$ —that is, by the complementary distribution function for a Poisson random variable with mean $(\alpha + \beta)T$. To estimate how many terms are needed for an accurate approximation, suppose that $(\alpha + \beta)T$, which is the average number of goals per game, equals 6 (which is roughly consistent with the last few NHL seasons). With this goals-per-game average, truncating the infinite series after 20 terms results in a truncation error of about 10^{-6} .

Near the end of the NHL's 2008–2009 regular season, the Detroit Red Wings were the top scoring team, with a goal scoring average of 3.63 per game, followed by the Boston Bruins, with an average of 3.28 per game. The lowest scoring team, the Phoenix Coyotes, averaged 2.43 goals per game. The calculation I have outlined results in the Red Wings having a 0.48 probability of winning a game against the Bruins in regular time (with a 0.37 probability of a loss and 0.15 probability of a tie) and a 0.61 probability of winning a game against the Coyotes in regular time (with a 0.24 probability of a loss and 0.15 probability of a tie). Table 1 illustrates the calculation for the Red Wings versus the Bruins. To refine this analysis, instead of using average goal-scoring rates against all teams, one could attempt to classify teams based on ability and estimate average goal-scoring rates for each team as a function of the opposing team's ability.

Let us now revisit the assumptions that led us to model scoring by Teams A and B as independent Poisson processes. References 3 and 4 demonstrate empirically that the Poisson approximation is fairly accurate. Two assumptions are particularly suspect, if taken literally. First, there are various reasons to expect that the goal-scoring rate of

Table 1. Calculation of the Probability that the Detroit Red Wings Win in Regular Time, if Playing the Boston Bruins

Number of Goals (n)	$\Pr\{n \text{ Goals}\}$	Number of Goals Needed to Win	$\Pr\{\text{Red Wings Win}\}$
1	0.015	1	0.525
2	0.045	2	0.276
3	0.089	2	0.538
4	0.134	3	0.351
5	0.161	3	0.547
—	—	—	—
20	3.7×10^{-6}	11	0.503
$\Pr\{\text{Red Wings win}\} = 0.48$		Error bound: 1.2×10^{-5}	

one team may depend on the goal-scoring rate of the other team. For example, one could expect that if Team A scores the first goal, then Team B will respond by “trying harder” and its goal-scoring rate will increase. Such hypotheses say nothing, however, about whether or not the dependence between the two processes is large enough to cause concern. This approximation appears to be defensible, as the only study that, to my knowledge, has investigated this issue [3] found no statistical evidence against the independence assumption. Secondly, one would expect that a certain minimum time must elapse between one goal and the next, thus violating the assumption that the probability of a goal in the next infinitesimal interval does not depend on when the last goal was scored. Reference 5 investigated this issue in detail, and found that a generalization of the Poisson process, whereby the goal-scoring rate increases gradually from zero to a constant rate as time from the last goal increases and provides a good fit to empirical data on the times at which goals were scored in the NHL. Based on the analysis in Ref. 5, the “warm-up” time required before the goal-scoring rate reaches the constant value appears to be approximately 20 s. Although the model in Ref. 5 is undoubtedly more realistic, the Poisson model provides a reasonable and relatively tractable first approximation to the dynamics of a hockey game.

Goal scoring by individual players has also been modeled as a Poisson process. For example, Refs 6 and 7 demonstrate that the distribution of the number of points (goals plus assists) per game for former

NHL player Wayne Gretzky—the “Great One”—is remarkably well approximated by a Poisson distribution, with a mean of 2.39 points per game.

The frequency of “hat tricks” (a player scoring three or more goals in a game) is another measure that can be understood in terms of a Poisson process model. In the fall of 2008, Teemu Selanne scored his twenty-first career hat trick. This was described as a “statistical oddity” [8] by one sports writer, because Selanne has scored more hat tricks in 16 seasons than Gordie Howe, the NHL’s most durable player, had scored in 26 seasons, while scoring fewer total goals than Howe did. Is Selanne’s achievement unusual, from a statistical point of view? Let us use the Poisson model and try to answer that question. At the time, Selanne had played 1080 regular season NHL games and had scored 560 goals, for a 0.52 goals-per-game average. If Selanne’s goal production follows a Poisson process with a constant rate of 0.52 goals per game, then we can compute the probability that he scores zero goals ($\exp(-0.52) = 0.60$), one goal ($0.52 \exp(-0.52) = 0.31$), two goals ($0.52^2 \exp(-0.52)/2 = 0.08$), or three or more goals ($1 - 0.60 - 0.31 - 0.08 = 0.016$) in a game. Using the 0.016 estimate of Selanne’s hat trick probability, we can estimate an expected number of hat tricks for him to be $1080 \times 0.016 = 17.10$. Because a Poisson distribution’s variance equals the mean, the standard deviation equals $\sqrt{17.10} = 4.13$; if the Poisson model is accurate, therefore, we would expect Selanne’s actual number of hat tricks (21) to have a 68% probability (based

on a normal distribution approximation) of falling within 17.10 ± 4.13 , as indeed it does, albeit at the high end of the range. From this perspective, Selanne's achievement is close to what one would expect, given his total number of games played and total number of goals scored. If one does the same analysis for Howe, one obtains an expected value of 19.60 total hat tricks with a standard deviation of 4.43. Howe's actual number of hat tricks was 19—very close to the expected value. Thus, the reasoning that players with a higher total number of goals and higher total number of seasons are likely to have a higher total number of hat tricks is false. In order to be likely to score hat tricks, a player must have a high goals-per-game average, and this matters more than the total number of goals.

If one repeats the hat trick analysis for several other players who were mentioned in Ref. 8, two stand out as having had a significantly greater number of hat tricks than expected: Wayne Gretzky (50 actual, 34.55 expected) and Glenn Anderson (21 actual, 11.64 expected). The players mentioned in Ref. 8 are primarily ones that are at the top of the list of hat trick scorers, so these players do not constitute a random sample of NHL players. For that reason alone, one would expect the number of hat tricks for this group of players to be biased upwards, a point on which Ref. 9 elaborates. However, there are other factors involved. Let us take a closer look at Gretzky. His expected number of hat tricks is 34.55 but he actually had 50 hat tricks, which is well outside the standard deviation of 5.88. One of many unusual things about Gretzky is how much his goals-per-game average changed during his career, from 1.15 in the 1981–1982 season to 0.13 in the 1999–2000 season. Computing the expected number of hat tricks separately for every season (which corresponds to assuming that Gretzky's goal-scoring process was a Poisson process with a time-varying rate) should result in a better estimate. Adding the expected number of hat tricks for each season results in an expected value of 47.6 career hat tricks, with a standard deviation of 6.91. With this approach, Gretzky's 50 actual

career hat tricks are well within one standard deviation of the expected value.

EMPIRICAL EVIDENCE ABOUT ICE HOCKEY

A number of other questions about ice hockey have been investigated empirically, including the following:

- Can a team's win/loss percentage be predicted from the number of goals, for (GF) and against (GA)? Baseball statistician Bill James [10] found that for major league baseball, teams' win/loss percentages were well approximated by $(\text{runs scored})^2 / [(\text{runs scored})^2 + (\text{runs allowed})^2]$ and he referred to this as the Pythagorean method. In an attempt to develop a similar formula for ice hockey, Cochran and Blackstock [11] found that NHL teams' regulation time win/loss percentages over a season (adjusted for ties) were well approximated by $\text{GF}^{1.927} / (\text{GF}^{1.927} + \text{GA}^{1.927})$.
- How likely is a player to score on a penalty shot during a shootout? According to Ref. 12, the overall average scoring probability is 0.33. The probability is only slightly higher (0.34) for the first six shooters (who were presumably chosen for their scoring ability) than the later shooters (0.31). According to the same reference, scoring probabilities differ only marginally, depending on the importance of the shot ("must make" vs. "could win" vs. "no pressure").
- The longest shootout in NHL history involved 30 shooters, in a 2005 game between the New York Rangers and the Washington Capitals. How long will we have to wait for a 40-shooter shootout? According to estimates in Ref. 12, 40-shooter shootouts can be expected to occur once every 156 seasons, on average.
- More than half the regular season games that reach overtime remain tied at the end of overtime and are therefore determined by a shootout. Is the overtime period too short? According

to estimates in Ref. 13, if the overtime period were extended from 5 to 10 min, then the fraction of overtime games that are decided by a shootout would drop from 0.53 to 0.28.

- If a team is leading after two periods, how likely is it that it will win the game? According to Ref. 4, the answer is 0.72 if the team is playing at home and is leading by one goal, 0.92 if it is leading by two goals, and 0.98 if leading by three goals. The probabilities are slightly lower for teams playing away from home.
- A related question is “when is a goal most valuable?” Or, in other words, under what circumstances is goal scoring most likely to increase the probability of winning the game? According to Ref. 5, not surprisingly, a goal is most valuable if the score is close to being tied and little time is left in the game. As an example, Ref. 5 estimates that a game-tying goal with 5 min remaining in regulation time increases the probability of winning the game for the scoring team from 0.10 to 0.40.
- Are referees biased, or as Ref. 14 puts the opposite possibility, are referees Markov (i.e., “memoryless”)? These authors find that the home team receives fewer penalties on average, a finding that is consistent with similar analyses for soccer. (Instead of referee bias, a possible alternative explanation that has not been investigated to my knowledge, is that the home team commits fewer penalties, on average.) A second type of bias that Ref. 14 investigated is “follow-up bias”, in which a referee that penalized Team A last is more likely to penalize Team B next. According to estimates in Ref. 14, if the home team gets the first penalty, the probability that the away team gets the second penalty is 0.62. If the home team receives the first two penalties, the probability that the away team gets the third penalty is 0.65.
- If Wayne Gretzky and Mario Lemieux had reached their peak in the same year, and played on equally strong teams, who would have scored more points? Reference 7 develops a “statistical time machine” to answer such questions and it gives the edge to Lemieux.
- Is the “hot goalie” a myth, as Ref. 15 argues is the case for the “hot hand” in basketball? Reference 16 examines the performance of the Detroit Red Wings in their four-game sweep of the Stanley Cup final series in 1997, with Mike Vernon in goal, and concludes that the “hot goalie hypothesis is alive and well.”
- When a team performs poorly, a typical reaction is to fire the coach. Does firing the coach work? Management strategists have investigated team leadership in the context of NHL hockey, particularly for coach succession. According to Refs 17 and 18, when a coaching change occurs between seasons, the team’s performance remains about the same, on average. When a coach is replaced during a season, the team performance worsens, on average, during the remainder of the season. If a coach taking over during a season continues to coach in the following season, the team’s performance improves, on average during the second season. The explanation offered by Ref. 17 is that a coaching change during the season gives the new coach an opportunity to evaluate players, make roster changes, and influence player training during the offseason, all of which result in improved performance during the next season.
- Does awarding one point for an overtime loss cause teams to play more defensively if they are tied near the end of regulation time? A series of articles by economists on this topic [19–22] uses models of goal scoring (similar to our Poisson process assumptions) and empirical data to suggest that the answer is yes, and that this phenomenon has increased the frequency of overtime games.
- What determines player salaries? There is a sizable literature in sports

economics (see Ref. 23 for a review) that attempts to relate player salaries to such factors as market structure, nationality, and player type. For example, Ref. 24 found that NHL players from Sweden and Finland were higher-paid, on average, than players from Canada, the United States, Russia, and the former Czechoslovakia, perhaps because Swedish and Finnish players have relatively more attractive options for playing in their home countries rather than the NHL. As another example, after the short-lived World Hockey Association (WHA; an NHL competitor) folded, Jones and Walsh [25] estimated NHL players' contribution to team revenue and compared them to players' salaries. They found that, on average, player salaries exceeded their contribution to team revenue—a result that is consistent with the NHL having increased player salaries beyond the point of profitability, in order to drive the WHA out of business.

WHEN TO PULL THE GOALIE

Moving on to decision models, operations researchers have studied one decision particularly intensely: When should a coach “pull the goalie?” If a hockey team is down by one goal near the end of a game, the coach will often replace the goalie with another player, so that the team has six skating players against the opponent's five. Pulling the goalie increases the team's goal-scoring rate, thus increasing the probability that the team will tie the game. The opponent's goal-scoring rate is likely to increase even more, thus increasing the probability that the team loses by more than one goal, but in terms of points in the standings, a loss is worth zero points, regardless of how many goals separate the teams. (Although goal differential could matter at the end of the season, as it is used as a tiebreaker in determining which teams enter the playoffs.) Therefore, when the probability of a win is close to zero, a strategy that increases the probability of

a tie and thus entering overtime is worthwhile, even if the strategy also increases the probability of losing by a larger margin.

How do NHL coaches decide when to pull the goalie? Former Los Angeles Kings head coach, Andy Murray, was quoted in 1994 [26] as saying “if you're down by one goal, you're looking at the 1 min mark.” The “one minute when trailing by one” rule is also mentioned in Refs 27–31, all of whom analyze the decision mathematically and conclude that it would be better to pull the goalie more than 1 min before the end of the game. Recall our assumption that scoring by the two teams can be modeled as independent Poisson processes with rates α and β . If Team A pulls its goalie, suppose that its goal-scoring rate increases from α to α' and Team B's goal-scoring rate increases from β to β' , and further assume that $\beta' > \alpha'$. With T min remaining in the game and Team A trailing by one goal, suppose that Team A decides to pull its goalie when $t < T$ min are left, if it is still trailing by one goal at that point. This strategy will result in a tie if Team A scores the next goal. Following Refs 28–30, I computed the probability that Team A scores the next goal separately for two cases: (i) the next goal is scored in the interval from T to t min and (ii) the next goal is scored in the last t min. Under case (i), Team A scores the next goal if the following two events occur:

- The teams score at least one goal in the interval from T to t (with probability $1 - \exp(-(\alpha + \beta)(T - t))$).
- Conditional on the preceding event, Team A scores the first goal (with probability $\alpha/(\alpha + \beta)$).

Under case (ii), Team A scores the next goal if the following three events occur:

- Neither team scores in the interval from T to t min (with probability $\exp(-(\alpha + \beta)(T - t))$).
- The teams score at least one goal in the last t min (with probability $1 - \exp(-(\alpha' + \beta')t)$).
- Conditional on the preceding event, Team A scores the first goal (with probability $\alpha'/(\alpha' + \beta')$).

By combining terms, we obtain the probability that Team A obtains a tie sometime before the game ends, as a function of T and t :

$$f(T, t) = (1 - e^{-(\alpha+\beta)(T-t)}) \frac{\alpha}{\alpha + \beta} + e^{-(\alpha+\beta)(T-t)} (1 - e^{-(\alpha'+\beta')t}) \frac{\alpha'}{\alpha' + \beta'} \quad (2)$$

One obtains the optimal time to pull the goalie by maximizing f with respect to t . One can find the optimal time t^* analytically [28] and t^* has the desirable property that it does not depend on T , but it is perhaps more insightful to plot how f depends on t for a given T . As an example, suppose that in a game between the Red Wings and the Coyotes, the Red Wings are trailing by one goal with $T = 5$ min remaining. Figure 1 shows how the probability that the Red Wings will tie the game before it ends (i.e., $f(5, t)$) depends on the time t at which the Red Wings coach pulls the goalie, if the Red Wings have not tied it already. The probability of a tie is computed using Equation (2) and the goal-scoring rate estimates discussed earlier. Pulling the goalie with 1 min remaining results in a tie probability of 0.279. The optimal time to pull the goalie is considerably earlier than 1 min before the end of the game: approximately 2.5 min. By pulling the goalie earlier, the probability of a tie increases from 0.279 to 0.295.

The models and the goal-scoring rate estimates in Refs 27–31 differ, but all of these

papers reach the same conclusion: it is optimal to pull the goalie with more than one min remaining (the estimates of t^* in these papers range from 1.3 to 4 min). As one would expect, strong teams (with high α) should pull their goalie sooner than weak teams do, and teams playing against a strong opponent (with high β) should pull their goalie later than should teams playing a weak opponent [29].

WILL MY TEAM QUALIFY FOR THE PLAYOFFS?

As of 20 March 2009, the Edmonton Oilers had played 70 of their 82 regular season games and had earned 77 points, putting them in seventh place in the NHL's Western Conference. As is usually the case at this time of the year, for me and other Oilers fans there was considerable uncertainty about whether or not the Oilers would be one of the eight Western Conference teams to earn a playoff spot. *Simulation* and *optimization* can provide different and complementary answers to this question. Optimization [e.g., [32,33]] can sometimes provide a definite “yes” or “no” answer, whereas simulation [34] can provide an estimate of the probability that the answer will be “yes.” Simulation can also provide additional information—about how many points will most likely be needed to qualify for the playoffs, for example, or the probability that a team will qualify, given that it earns X points in its last Y games.

I now describe the Monte Carlo simulation model in Ref. 34 in greater detail. The web

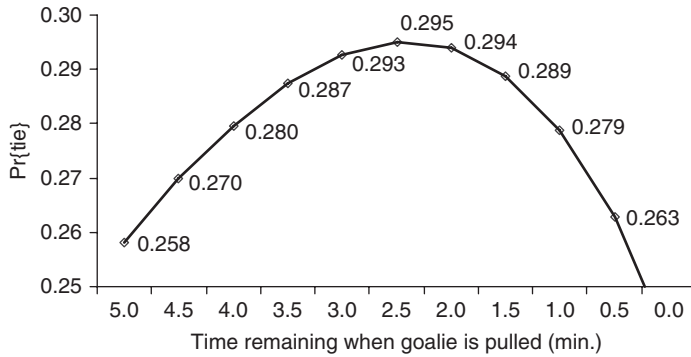


Figure 1. Probability of a tie as a function of the time the goalie is pulled.

site <http://www.sportsclubstats.com/NHL.html> shows the results of a similar simulation model. The building block of the model is the simulation of the outcome of every remaining regular season game. To simulate the outcome of a game between Teams A and B, one requires as input the probability that Team A will win the game. A simple approach, adopted in Ref. 34, is to use the average of Team A's winning percentage and Team B's losing percentage at that point in the season (Ref. 34 provides additional details, including how to simulate overtime games). Alternatively, one could use the approach described at the beginning of this article, an approach that involves computing win probabilities based on the team's goal-scoring rates and Poisson assumptions. Another option is a statistical model that estimates win probabilities by incorporating such factors as home-ice advantage, offensive capability, defensive capability, and player injuries.

In each replication of the simulation, one simulates every remaining regular season game. Then, one computes the simulated standings at the end of the season, according to the NHL's ranking criteria, which are the following, in decreasing order of priority:

- Top ranking within a Division guarantees a top three standing in the Conference;
- Total points;
- Total wins;
- Results of head-to-head games for teams that are tied based on the first three criteria;
- Total goals for (GF) minus total goals against (GA).

Interestingly, before 1970, the NHL used GF instead of GF–GA as the final criterion; as discussed in Ref. 35, this led to perverse incentives near the end of the 1969–1970 regular season. In the Montreal Canadiens' final game that season, trailing 5–2 against the Chicago Blackhawks, the Canadiens could have qualified for the playoffs if they had scored three more goals, regardless of how many more goals the Blackhawks

scored. The Canadiens followed the rational strategy, given the ranking criteria, of pulling their goalie at the beginning of the third period—but without success. Reference 35 uses regression analysis to argue that GF–GA is a better tie breaker, and the NHL has since adopted this approach.

Simulation has an advantage over optimization in handling arbitrarily complex ranking criteria. The approach used in Ref. 34 is to rank teams based on “pseudo points,” computed as follows:

$$\text{Pseudo points} = M(1) \times (\text{Criterion 1}) + \dots + M(n) \times (\text{Criterion } n), \quad (3)$$

where $M(1) \gg \dots \gg M(n) > 0$, in order to enforce the strict prioritization of the factors. Put differently, the teams are ordered lexicographically, based on the ranking criteria. If some of the criteria are difficult to implement in the simulation model (e.g., the results of head-to-head games) and are unlikely to make a difference, then one can either ignore them (and eliminate replications where those criteria made a difference) or replace those criteria with a random number in order to break ties that remain after applying the criteria that are explicitly included in the simulation.

Table 2 shows an example of simulation results for the NHL's Western Conference, as of 19 March 2009 and based on 5000 replications, focusing on the Edmonton Oilers. Each row in the table corresponds to one replication. On the basis of these replications, one can estimate the probability that the Oilers qualify for the playoffs (0.624) and the average number of points needed to qualify (88.9), by computing simple averages.

I show two examples to illustrate how one can dig deeper into the simulation results to answer interesting questions. First, how does the probability that the Oilers qualify for the playoffs depend on the number of points they earn from their remaining 12 games? Table 3 shows a cross-tabulation obtained from the simulation results and it answers this question (there were no replications in which the Oilers earned 0, 23, or 24 points in their final 12 games). If Edmonton were to earn 12 of the 24 available remaining points,

Table 2. NHL Western Conference Simulation Results, as of 19 March 2009

Replication	Edm. Rank	Edm. Points	Edm. Qualifies?	Points Needed to Qualify
1	6	94	Yes	92
2	11	86	No	*
3	10	86	No	87
—	—	—	—	—
5 000	5	95	Yes	89

*Simulated criteria insufficient to determine top 8 teams

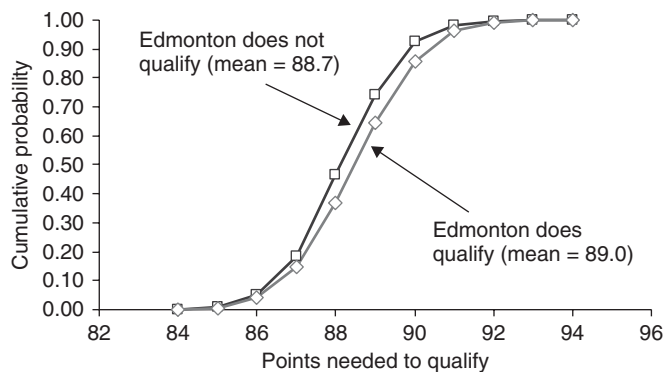
their probability of qualifying would be 0.73; to bring the probability to 0.99, they would need to earn 14 of the 24 available points.

Secondly, sports writers sometimes use a two-step process to estimate how many points a team will need to achieve to qualify for the playoffs: (i) estimate the number of points X that will be needed to qualify and (ii) subtract the team's current number of points from X . Implicit in this approach is an assumption that the random variable X is independent of whether or not the team of interest qualifies. Put in more concrete terms, the question is whether the distribution of the minimum number of points needed to qualify for the playoffs depends on whether the Edmonton Oilers qualify. By cross-tabulating the simulated values of X against an indicator variable for whether or not Edmonton qualifies, one can produce a graph like the one shown in Fig. 2. As that figure shows, more points are needed to qualify, on average, if Edmonton qualifies than if it does not. To understand this outcome, keep in mind that

Table 3. Cross-tabulation of Edmonton Oilers Points and Whether Edmonton Qualifies for the Playoffs

Pts. in Last 12 Games	Edm. does not Qualify	Edm. Qualifies	Pr (Edm. Qualifies)
1	1		
2	4		
3	6		
4	22		
5	37		
6	86		
7	152		
8	278	2	0.01
9	315	17	0.05
10	404	115	0.22
11	327	207	0.39
12	173	469	0.73
13	66	537	0.89
14	8	583	0.99
15	2	416	1.00
16		316	1.00
17		206	1.00
18		143	1.00
19		55	1.00
20		37	1.00
21		11	1.00
22		5	1.00
Totals	1881	3119	

in the current standings, Edmonton is in the top 8. If Edmonton ends up not qualifying, it means that the Oilers performed relatively poorly during the remainder of the season, which would tend to reduce the qualification threshold for teams that are below the Oilers in the current standings.

**Figure 2.** Distribution of number of points needed to qualify, depending on whether Edmonton qualifies.

In conclusion, operations research and statistical analyzes can provide insight and predictions about game strategies, whether a team will make the playoffs, the consequences of rule changes, and a variety of other issues of interest to ice hockey sportswriters, analysts, and fans. I hope that this article will stimulate readers to learn more about the many issues touched upon, and that it will motivate other researchers to undertake further study of ice hockey.

REFERENCES

1. Hurley WJ. How should you order the players in a shootout? *Chance* 1998;8(1):25–26.
2. Bertsekas DP, Tsitsiklis JN. Introduction to probability. Belmont (MA): Athena Scientific; 2002.
3. Mullet GM. Simeon Poisson and the National Hockey League. *Am Stat* 1977;31(1):8–12.
4. Gill PS. Late-game reversals in professional basketball, football, and hockey. *Am Stat* 2000;54(2):94–99.
5. Thomas AC. Inter-arrival times of goals in ice hockey. *J Quant Anal Sports* 2007;3(3). Available at <http://www.bepress.com/jqas/vol3/iss3/5>.
6. Schmuland B. Shark attacks and the Poisson approximation. *Pi Sky* 2001 December: 12–14.
7. Berry SM, Reese CS, Larkey PD. Bridging different eras in sports. *J Am Stat Assoc* 1999;94(447):661–676.
8. Matheson J. Late bloomer Selanne still defies odds; Ducks veteran has 21 career hat tricks despite starting play in NHL at age 22. *Edmonton J* 2008 Nov 4:C.3.
9. Hurley WJ. Jussi Jokinen, Regression to the mean, and the assessment of exceptional performance. *Chance* 2009;22(1):28–33.
10. James B. The Bill James Abstract. Self-published, 1980.
11. Cochran JJ, Blackstock R. Pythagoras and the National Hockey League. *J Quant Anal Sports* 2009;5(2). Available at <http://www.bepress.com/jqas/vol5/iss2/11>. DOI: 10.2202/1559-0410.1181.
12. Hurley WJ. Using operational research to answer some interesting questions about NHL shootouts. *Int J Oper Res* 2009;4(1):117–123. DOI: 10.1504/IJOR.2009.021621.
13. Brimberg J, Hurley WJ. Is the overtime period in an NHL game long enough? An example for teaching estimation and hypothesis testing in the presence of censored data. *Am Stat* 2008;62(2):151–154.
14. Brimberg J, Hurley WJ. Are National Hockey League Referees Markov? Working paper.
15. Tversky A, Gilovich T. The cold facts about the “hot hand” in basketball. *Chance* 1989;2(1):16–21.
16. Morrison DG, Schmittlein DC. It takes a hot goalie to raise the Stanley Cup. *Chance* 1998;11(1):3–7.
17. Rowe WG, Cannella AA Jr, Rankin D, *et al.* Leader succession and organizational performance: integrating the common-sense, ritual scapegoating, and vicious-circle succession theories. *Leader Q* 2005;16:197–219.
18. Audas R, Goddard J, Rowe WQ. Modelling employment durations of NHL head coaches: turnover and post-succession performance. *Manag Decis Econ* 2006;27:293–306. DOI: 10.1002/mde.1259.
19. Abrevaya J. Fit to be tied: the incentive effects of overtime rules in professional hockey. *J Sports Econ* 2004;5(3):292–306. DOI: 10.1177/1527002503260560.
20. Barnejee AN, Swinne JFM, Weersink A. Skating on thin ice: rule changes and team strategies in the NHL. *Can J Econ* 2007;40(2):493–514.
21. Longley N, Sankaran S. The incentive effects of overtime rules in professional hockey: a comment and extension. *J Sports Econ* 2007;8(5):546–554. DOI: 10.1177/1527002506294938.
22. Shmanske S, Lowenthal F. Overtime incentives in the National Hockey League (NHL): more evidence. *J Sports Econ* 2007;8(4):435–442. DOI: 10.1177/1527002506292581.
23. Leadley JC, Zygmunt ZX. National Hockey League. In: Fize J, editor. Handbook of sports economics research. Armonk (NY): M. E. Sharpe; 2006. pp. 49–98.
24. Williams B, Williams D. Are salaries in the National Hockey League related to nationality? *Chance* 1997;10(3):20–24.
25. Jones JCH, Walsh WD. The World Hockey Association and player exploitation in the National Hockey League. *Q Rev Econ Bus* 1987;27(2):87–101.
26. Laskaris S. Behind the bench: pulling your goalie. *Hockey Player Mag* 1994. Available at http://www.hockeyplayer.com/artman/publish/article_462.shtml.

27. Morrison DG. On the optimal time to pull the goalie: a Poisson model applied to a common strategy used in ice hockey. In: Machol RE, Ladany SP, Morrison DG, editors. TIMS Studies in Management Science. volume 4. Management science in sports. Amsterdam: North-Holland; 1976. pp. 137–144.
28. Morrison DG, Wheat RD. Misapplications reviews: pulling the goalie revisited. *Interfaces* 1986;16(6):28–34.
29. Erkut E. More on Morrison and Wheat's pulling the goalie revisited. *Interfaces* 1987;17(5):121–123.
30. Nydick RL Jr, Weiss HJ. More on Erkut's more on Morrison and Wheat's 'pulling the goalie revisited'. *Interfaces* 1989;19(5):45–48.
31. Washburn A. Still more on pulling the goalie. *Interfaces* 1991;21(2):59–64.
32. Adler I, Erera AL, Hochbaum DS, *et al.* Baseball, optimization, and the world wide web. *Interfaces* 2002;32(2):12–22.
33. Russell T, van Beek P. Determining the number of games needed to guarantee an NHL playoff spot. Proceedings of the Sixth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CPAIOR 2009), Pittsburgh, 2009, 233–247.
34. Ingolfsson A. Simulating NHL games to motivate student interest in OR/MS. *INFORMS Trans Educ* 2004;5(1):37–46.
35. Morris C. Breaking deadlocks in hockey: prediction and correlation. In: Mosteller F, editor. *Statistics by example: detecting patterns*. Reading (MA): Addison-Wesley; 1973. pp. 119–140.

ICELANDIC OPERATIONS RESEARCH SOCIETY

HLYNUR STEFÁNSSON
EYJÓLFUR INGI ÁSGEIRSSON
School of Science and Engineering,
Reykjavik University, Reykjavik,
Iceland

SNJÓLFUR ÓLAFSSON
School of Business, University of
Iceland, Reykjavik, Iceland

HISTORY AND ACTIVITIES

ICORS (Icelandic Operations Research Society) was founded in January 1985. The membership quickly increased to just more than 100 members in the first five years, and has stayed relatively stable at around 100 ever since. The population of Iceland is only around 300,000 so one might not expect to find an OR society in such a small country. A part of the explanation is that ICORS is somewhat different from most of the other OR societies in that only a small number of the members are academics and that very few have a degree in OR. Many of the members have had one or two courses in OR, including a large number of the engineers, and they want to explore possible uses of OR methods in their firms or institutions. The concept of OR is relatively well known in Iceland for several reasons: one is that the concept of OR (“aðgerðarannsóknir” in Icelandic) is discussed in newspapers and magazines every now and then and another is that one of the early OR people was for a while Minister of Fisheries and a member of the parliament for many years.

ICORS arranges meetings approximately once every month during winter and publishes a newsletter in Icelandic.

ICORS collaborates with the Nordic OR societies, which have arranged biennial conferences, now under the name Nordic operations research society (NOAS), but formerly

NOAK. In 1980, an extra NOAK conference was held in Iceland between two regular ones.

ICORS is a member of the EURO (Association of European Operational Research Societies) and INFORMS (Institute for Operations Research and Management Sciences).

HONORARY MEMBERS

The ICORS has six honorary members:

1. Helgi Sigvaldason, is a pioneer in teaching OR in Iceland. He wrote the first text book on OR in Icelandic. He also applied OR methods to various practical projects in Iceland [1].
2. Frederick S. Hillier is a Professor of OR at Stanford University: Well known for his widely used text books on OR, his roots are Icelandic and he has contributed to the development of OR in Iceland.
3. Kjartan Jóhannson was one of the pioneers in OR in Iceland, and first worked as a consultant. Later he became a member of parliament and a minister, and he supported OR in different ways.
4. Þorkell Helgason was a pioneer in applying OR methods within the governmental sector in Iceland.
5. Páll Jensson is a Professor of OR at University of Iceland. He has been involved in various research projects and practical OR projects in Iceland [2,3].
6. Snjólfur Ólafsson, Professor of OR at University of Iceland, was the first president of ICORS. He has been involved in various research projects and practical OR projects in Iceland [4,5].

ICORS PRESIDENTS

ICORS has had three presidents from its establishment in 1985. The first president was Prof. Snjólfur Ólafsson. He was the

president for 17 years until Prof. Birna Kristinsdóttir took over in 2002. She was the president until 2007 when Prof. Hlynur Stefánsson took over.

EURO XXI CONFERENCE IN ICELAND

The EURO XXI Conference was held by ICORS in Reykjavik (July 2–5, 2006). The theme of the conference was “OR for better management of sustainable development.” The conference was a great success with 1541 delegates from 63 countries. In the following section we summarize some key information about the conference.

Organization

In 2003, at the EURO conference in Istanbul, ICORS was formally appointed to host the EURO XXI Conference in 2006 in Reykjavik, Iceland. The following members were appointed to the program committee: Tuula Kinnunen (chair), Gerhard-Wilhelm Weber (plenary chair), Raymond Bisdorff, Jean-Pierre Brans, Ulrike Leopold-Wildburger, Snjólfr Ólafsson, Maurice Shutler, Marino Widmer, and Martin Zachariassen. The following members were appointed to the organizing committee: Snjólfr Ólafsson (chair), Gülay Barbarosoglu, Magnús M. Halldórsson, Ingjaldur Hannibalsson, Thorkell Helgason, Páll Jensson, Tuula Kinnunen, Jakob Krarup, Birna P. Kristinsdóttir, Bjarni Kristjánsson, and Hlynur Stefánsson.

The program and organizing committees had four formal meetings to prepare the conference, one every year from 2003 to 2006.

Program and Scope of the Conference

Following the regular format for EURO conferences, the EURO XXI 2006 conference featured plenary, semiplenary, tutorial, keynote sessions and panels along with a large number of invited and contributed sessions on topic areas. The plenary sessions were given by the EURO 2006 Gold Medalist Luk van Wassenhove and the IFORS Distinguished Lecturer Saul Gass who gave a very informative talk titled “An annotated timeline

of operations research.” In total, there were 14 semiplenary sessions, 8 tutorial sessions, 5 keynote sessions, and 4 panel sessions. Other presentations were organized in 84 topic-specific streams containing 3–17 sessions each and in total 1477 papers were presented. A total of nine exhibitors attended the conference to display books and software. The conference was held at the University of Iceland in 47 meeting rooms spread across seven buildings. The sessions were organized in specific buildings, based on streams and subject themes.

The social program was very well attended. The welcome reception took place at the Blue Lagoon. However, due to the high number of participants it was impossible to take them all to the Blue Lagoon on the same day. Therefore half the participants went to the Blue Lagoon on Sunday and to a reception in Perlan (a popular restaurant with panorama view over entire Reykjavik) on Monday, and the other half in opposite order. The Soul Gass 7K Run took place with 27 participants, many of whom enjoyed using the outdoor swimming pools of Reykjavik afterwards. Finally the social program ended with the gala banquet, which was held in a place that could seat all the attending participants.

OR AT THE ACADEMIC LEVEL IN ICELAND

OR in Icelandic Universities

At the University of Iceland there are three faculty positions more or less devoted to OR: one in the Faculty of Mathematics, one in the Faculty of Engineering and one in the Faculty of Business and Economics. At Reykjavik University there are another three faculty positions devoted to OR. Introductory courses to OR are taught to approximately 250 students annually and several advanced courses are offered at undergraduate and graduate levels.

OR Students and Faculty Abroad

There are at least four members of ICORS who have OR/management science faculty positions in foreign universities, including

Georgia Institute of Technology, London School of Business, Stanford University, and University of Atlanta.

Quite a few students have completed MSc or PhD degrees in OR from various foreign universities, mostly in North America and Europe.

OR PROJECTS—SHORT HISTORY OF THE USE OF OR IN ICELAND

During the early years of the use of OR methods in Iceland (1965–1990) the main emphasis was on fishing and the fishing industry. This emphasis is not surprising since the fishing industry is one of the backbones of the Icelandic economy. The first major Icelandic OR model was created in 1966. The OR model was a simulation of the fishing and landing of herring, written in Fortran. The project was a collaboration between Icelandic and Danish scientists.

In 1969 a small group made a model for demersal fishing, mainly to estimate whether small or large trawlers would be more economical.

In the late seventies, Þorkell Helgason and others began to build different fisheries models for demersal fishing. These models have been used, both for academic studies and also directly by government authorities. In the beginning, OR people presented optimization models [6,7] but these were never accepted by people in the industry. The model that was used was a detailed simulation model [4] which was based on the data for the last few decades of fishing in Icelandic waters. At that time, this data was probably more extensive than that available in most other countries.

The stock size of demersal species is one of the basic elements of the model. The estimate of stock size is based on data on the catch in earlier years, divided by age. Extensive statistical analysis was performed in order to improve the estimation methods [8,9]. Also, some research work was done concerning stock estimation for whales.

Iceland is rich in energy, both in the form of waterfalls and geothermal power. Many projects, in the field of OR and related fields,

have been carried out in connection with energy.

In terms of Icelandic software related to ICORS, Maximal Software, Inc. is one of the leading providers of software for developing and formulating models in the field of optimization. Maximal Software Inc. is founded and owned by Bjarni Kristjánsson, who is an active participant in ICORS.

Over the last 10–15 years a number of companies specializing in OR or OR-related methods have been founded. The earliest companies include Mímisbrunnur ehf. and Bestun og ráðgjöf ehf. which later merged, forming a company called AGR ehf. AGR specializes in forecasting and inventory management. AGR has developed software that uses forecasting to create automatic purchase proposals for small- to medium-sized companies, with the goal of improving stock turnover and decreasing the stock in inventory. Other companies include Vaktaskipan ehf. which focuses on staff scheduling. There are also OR departments in a couple of the larger companies in Iceland. Marel Food System, a global provider of equipment and systems for the food processing industry, and Icelandair, the largest airline company in Iceland, both have OR departments and use OR methods to improve their products and services and to increase their revenue.

REFERENCES

1. Helgason T. Optimal fishing patterns for Icelandic cod. In: Chapman DC, Gallucci VG, editors. Quantitative population dynamics. Fairland (MA): International Co-operative Publishing House; 1981. pp. 243–265.
2. Helgason T, Ólafsson S. An Icelandic fisheries model. *Eur J Oper Res* 1988;33:191–199.
3. Guðmundsson G. Statistical considerations in the analysis of catch-at-age observations. *J Cons Int Explor Mer* 1986;43:83–90.
4. Guðmundsson G. Time series models of fishing mortality rates. ICES C.M. 1987/d: 6. Copenhagen: International Council for the Exploration of the Sea; 1987.
5. Hannibaldsson I. Optimal allocation of boats to factories during the capelin fishing season in Iceland [PhD thesis]. Ohio State University. 1978.

6. Jensson P. A simulation model of the capelin fishing in Iceland. In: Haley KB, editor. Applied Operations Research in Fishing, Proceedings from OR in Fisheries. NATO Symposium in Trondheim; 1981. pp. 187–198.
7. Jensson P. Daily production planning in fish processing firms. Eur J Oper Res 1988;36: 410–415.
8. Ólafsson S. Weighted matching in chess tournaments. J Oper Res Soc 1990;41:17–24.
9. Sigvaldason H. Beslutningsproblemer ved et Hydro-termisk elforsyningssystem (Decision Problems Concerning a System of Hydro and Powerplant) [Lic Tech Thesis]. Denmark: IMSOR; 1964.

IFORS: BRINGING THE WORLD OF OR TOGETHER FOR 50 YEARS

ELISE DEL ROSARIO
One Small Step Forward Foundation,
Inc., Bagumbayan, Quezon City,
Philippines

GRAHAM RAND
Department of Management Science,
Lancaster University, Lancaster,
UK

Operational Research (OR) was born at the time of ideological confrontation at the end of the 1930s, but the International Federation of Operational Research Societies (IFORS) was born some 20 years later out of professional cooperation. IFORS officially came into existence in January 1959, as a result of discussions inaugurated at the first international conference, held in Oxford in 1957 [1].

COMPOSITION

Member Societies

Founded by the United States (ORSA Operations Research Society of America), United Kingdom (ORS Operational Research Society), and France (SOFRO Société Française de Recherche Opérationnelle) 50 years ago, IFORS is now composed of 50 active member societies. Table 1 shows the development of the federation as national societies joined IFORS. The national member societies represent some 25,000 individual members, ranging from 50 to 10,000 per society.

Regional Groups

In 1976, the Association of European Operational Research Societies (EURO) within IFORS was established. The Association of Latin–Ibero American Operations Research Societies (ALIO) was founded in 1982, followed immediately in 1985 by the Association of Asia–Pacific Operations Research Societies (APORS) within IFORS. The Association of North American Operations

Research Societies (NORAM) within IFORS (composed of the OR societies of Canada and the USA) was created so that a Vice President representing the region may be appointed to the IFORS Administrative Committee (AC).

Kindred Societies and International Affiliation

Bodies other than the national societies, such as the Airlines Group of the International Federation of Operational Research Societies (AGIFORS), the Mathematical Programming Society (MPS), and the Fellowship for Operational Research (FOR) were welcomed as kindred societies. IFORS joined four other international federations—the International Federation of Automatic Control (IFAC), the International Federation for Information Processing (IFIP), the International Federation for Mathematics and Computers in Simulation (IMACS), and the International Measurement Confederation (IMEKO) to form the Five International Associations Coordinating Committee (FIACC).

LEADERSHIP AND ORGANIZATIONAL STRUCTURE

The initial structure of IFORS did not require a President. The affairs of the federation were to be administered by a “foster” society for three-year periods. The statutes provided for a secretariat consisting of an executive officer (the secretary) and a treasurer; both nominated by the designated foster society. The initial foster society was the ORS in the UK, which appointed Sir Charles Goodeve Secretary. ORSA took over as foster society for 1962–1964, with Philip Morse as the secretary. A change in structure of the administration of IFORS came into effect in 1968, after France had taken its turn as the third foster society, with the creation of an AC. A further change was introduced, when the vice presidential component of the AC became four regional VPs, plus one VP-at-large, elected by the IFORS membership as a

Table 1. Growth of IFORS: Date of Joining of Each National Society

Year	OR Society	Year	OR Society
1959	France, United Kingdom, United States of America	1978	Singapore
1960	Australia, Belgium, Canada, India, The Netherlands, Norway, Sweden	1979	Austria
1961	Japan	1982	China, Portugal
1962	Argentina, Germany, Italy	1983	Hong Kong, Yugoslavia
1963	Denmark, Spain, Switzerland	1986	Iceland
1966	Greece, Ireland, Mexico	1988	Malaysia
1969	Brazil, Israel	1990	Philippines, Poland
1970	New Zealand	1992	Hungary
1972	Korea	1993	Bulgaria
1973	South Africa	1994	Croatia, Czech Republic, Slovakia
1975	Chile, Finland	1998	Belarus
1976	Egypt	2002	Bangladesh, ^a Colombia, Lithuania
1977	Turkey	2007	Slovenia
		2009	Uruguay, Iran

^a Bangladesh withdrew from IFORS in 2009.

Table 2. Presidents of IFORS

Period	President	Period	President
1959–61	Sir Charles Goodeve (United Kingdom)	1983–85	Heiner Muller Merbach (Germany)
1962–64	Philip Morse (United States of America)	1986–88	Jacques Lesourne (France)
1965–66	Marcel Boiteux (France)	1989–91	Bill Pierskalla (United States of America)
1967	Charles Salzmann (France)	1992–94	Brian Haley (United Kingdom)
1968–70	Alec Lee (Canada)	1995–97	Peter Bell (Canada)
1971–73	Arne Jensen (Denmark)	1998–2000	Andres Weintraub (Chile)
1974–76	Takehiko (Bill) Matsuda (Japan)	2001–03	Paolo Toth (Italy)
		2004–06	Thomas Magnanti (United States of America)
1977–79	David Hertz (United States of America)	2007–09	Elise del Rosario (Philippines)
1980–82	Roger Colcutt (United Kingdom)	2010–12	Dominique de Werra (Switzerland)

whole. The presidents of IFORS are as shown in Table 2 [2].

IFORS is run by a board composed of the representatives of each member country who vote on major issues confronting IFORS. The AC is responsible for the execution of activities and drafting of proposals for approval by the board. The AC is elected for a period of 3 years, except for the treasurer whose term is renewable twice at the most. The secretary's location corresponds to the headquarters of IFORS.

TRIENNIAL CONFERENCES

The 1957 Oxford Meeting was followed by preparation of the statutes, which set the purpose as “the development of operational research as a unified science and its advancement in all nations of the world.” The statutes contain a provision that in all formal votes taken by the board, the voting strength of each member society is in proportion to the square root of the qualified membership. This is meant to give greater weight to larger societies but not to overwhelm the

smaller societies. The first conference was followed by a series of triennial conferences of which eight have been held in Europe, five in the North American region, two in the Asia–Pacific region, and one each in Africa and Latin America.

PUBLICATIONS

International Abstracts in Operations Research (IAOR)

The history of *International Abstracts in Operations Research* (IAOR) goes back to 1958 when Hugh Miser, the then secretary of ORSA, proposed to Sir Charles Goodeve that one of the early items of IFORS business should be the possibility of funding and operating a journal that would bring all “unclassified” OR work to the attention of the worldwide OR community; IAOR was first published in 1961 [3]. The method of collecting, indexing, and publishing material did not change until 2007 when the plan to respond to the challenges and opportunities of the internet were put in motion. The IAOR system development sought to modernize IAOR by making it an “online-based” rather than a “print-based” product. Development activities to improve speed and comprehensiveness in acquiring abstracts and allowing access to online subscribers were carried out. With effect from the third issue of 2009, all issues are being produced using the “Editor’s Workbench” which facilitates the electronic capture of materials.

International Transactions in Operational Research (ITOR)

From the first conference in 1957 up to the Athens Conference in 1990, a volume of proceedings was published. These volumes include a selection of papers presented, as well as records of discussions at workshops. After the 1993 Lisbon Conference and under the editorship of Peter Bell, the first volume of the *International Transactions in Operational Research* (ITOR) was published. A 2006 review committee to address the problems being encountered by the journal proposed a strategic plan that began to

be implemented in 2007. Various innovations on the composition of the editorial board and launching of special issues have resulted in increased paper submissions. An electronic editorial office was launched in 2008, enabling a decreased refereeing time and in general, greater effectiveness of the refereeing procedure. In 2003–2006, ITO published citations of 23 inductees into the IFORS OR Hall of Fame.

IFORS DISTINGUISHED LECTURE (IDL)

In 1999 IFORS established a International Federation of Operational Research Societies Distinguished Lectures (IDL) program to recognize distinguished OR scholars and analysts and support member societies and regional groupings. Through this program, IFORS sponsors lectures by distinguished OR scholars and analysts at conferences of its regional groupings. So far, the nomination, screening, and selection process have yielded 29 IDL.

DEVELOPING COUNTRY ACTIVITIES

The IFORS 1975 Japan Conference with the theme “OR in the Service of Developing Economies” put OR for development on the IFORS agenda. The 1984 Washington Conference started several initiatives in the area of OR for development. OR practice in developing countries continues to be one of the focal points of IFORS.

ICORD

In December 1992, participants of the first International Conference on Operations Research for Development (ICORD) held in Ahmedabad, India, agreed on recommendations on how OR could best be advanced in developing countries. The Ahmedabad Declaration, as it has become known, called for a range of actions from IFORS to support and strengthen OR in developing countries. Successive ICORDs were held in Rio de Janeiro (1996), Manila (1997), Berg-en-dal (2001), Jamshedpur (2005), and Fortaleza (2008). A call for proposals that would lead to

the organization of smaller workshops leading up to the next ICORD was subsequently issued. The new structure aims to achieve a greater momentum for the ORD (Operations Research for Development) program through greater frequency and visibility of actions and improved focus for ORD activities on selected problem-oriented themes [4], and is financially supported by IFORS.

IFORS Prize for OR in Development

The IFORS Prize, awarded at the triennial conferences, aims to bring to the fore outstanding works that show how OR has been used to help specific organizations in their decision-making process within the context of a developing country and in emphasizing developmental issues.

Africa Initiatives

With the help of EURO, IFORS has initiated several activities, including sponsoring conferences and scholarships in the African continent in an effort to address the lack of organized OR activity in the area. IFORS has provided funding for the Operations Research Practice for Africa (ORPA), conferences for 2007 (ORPA2), and 2008 (ORPA4). While ORPA2 focused on applications of OR outside South Africa, ORPA 4 focused on the use of operations research to address urban transportation and water resource management issues in Africa. During the 2008 IFORS Conference held in South Africa, EURO and IFORS jointly sponsored 25 participants coming from Africa and other developing countries. The Local Organizing Committee screened and selected from among the applicants the recipients of the registration, travel, and accommodation funding as required.

OR EDUCATION

Teachers' Workshops

One of the initiatives that came out of the 1984 IFORS triennial conference in Washington was the launching of teachers' workshops which aimed to increase the use of OR as a practical tool for decision makers in Asia by

influencing the way OR is taught. Regional workshops for teachers of OR in developing countries were subsequently held in Ahmedabad (1986), Mumbai (1987), Kuala Lumpur (1988), Singapore (1989), Jakarta (1990), and Manila (2004).

More recently, IFORS cosponsored the teachers' workshops initiative of the US national society, INFORMS (Institute for Operations Research and the Management Sciences), during the 2008 Latin-Ibero American Congress on Operations Research (CLAIO) held in Colombia and the 2009 APORS conference held in India. This program aims to hone the skills of OR teachers and improve their teaching effectiveness. Under this program, speaker expenses were paid by the sponsoring organizations.

Student Programs

Since 2000, IFORS has continually provided scholarship funds for participants to join summer institutes and workshops as part of the regional activities. So far, several scholars have been sent to the EURO, ALIO, and APORS workshops. The sponsorship is aimed at encouraging the establishment of groups of promising young OR scientists who will continue to work and network in the future. Under this program, IFORS pays for the air fare of the IFORS scholars while the local organizers provide the accommodation, program fees, and living expenses. IFORS also contributes to the expenses of the local organizers.

Most recently, IFORS scholars were sent, and a contribution to the local accommodation expenses was made, to the ELAVIO (Escuela Latinoamericana de Verano de Investigacion Operativa) 2008 in Peru and the ELAVIO 2009 in Mexico. Two IFORS scholars were sponsored to the EURO Summer–Winter Institutes (ESWI) 2009 in Spain. IFORS has also sponsored the EURO Working Group workshop on *OR for Developing Countries* by helping to defray expenses of the participants needing support and seconding a speaker to the workshop. Held regularly, the workshop provides a forum for doctoral students and others involved in OR and development to share and discuss their

research activities; encourages students to establish and maintain a research network; and fosters a research environment to support career development in a practical OR discipline.

Web-Based Educational Resources

As part of fostering excellence in OR education, the *Educational Resources* portion of the IFORS website is being developed to gather and make available on the web, high quality educational material such as case studies, methodological readings, and algorithms. This work is in progress. The other component is the *tutORial*, which has been completed and features a collection of interactive web-based tutorial modules on generic OR topics.

IFORS WEBSITE

Apart from serving as a valuable resource for the OR practitioner and academic, the IFORS website <http://www.ifors.org>, aims to be a key tool in promoting OR internationally. In line with this, the site includes information about local and international conferences, activities of member societies, updates on IFORS committee activities, links to national societies and regional groupings, and all IFORS newsletter issues. The website is kept current through the constant update of information and continuous upgrade of layout, navigation, news management system, RSS feeds, search and translation facilities. An area for members is available for online voting and discussions [5].

IFORS NEWS

The IFORS News takes over from the IFORS Bulletin last published in 2000. The IFORS News is published quarterly, featuring events from member societies, reports from the IFORS scholars, perspectives from IFORS distinguished lecturers, past and present IFORS officers, write-ups of IFORS prize-winning papers, and scholarly takes on various OR topics. It has also incorporated the newsletter of the Developing Countries Committee, which aimed to keep people in contact between the ICORD conferences. All the IFORS News issues are downloadable from the IFORS website and alerts are sent to the national societies when the issues become available. Only the annual report issues are printed for distribution at conferences and sent to the members by post.

REFERENCES

1. Rand GK. IFORS: the formative years. *Int Trans Oper Res* 2000;7:101–107.
2. Rand GK. Forty years of IFORS. *Int Trans Oper Res* 2001;8:611–633.
3. Rand GK. IAOR comes of age. In: Brans JP, editor. *Operational research '81*. Amsterdam: North-Holland; 1981. pp. 25–38.
4. Del Rosario EA., Rand GK. IFORS: 50 at 50. *BEIO* 2010;26(1):83–86.
5. Del Rosario EA. Three years into the IFORS 50th. *IFORS News* 2009;3(4):2–4.

IMPLEMENTING THE SIMPLEX METHOD

MATTHEW J. SALTZMAN
Department of Mathematical
Sciences, Clemson University,
Clemson, South Carolina

Linear programs (LPs) are the most widely employed optimization models. Throughout the history of operations research, the simplex method has been the workhorse algorithm for solving LPs.

Linear programming was first formalized by Dantzig [1]. The simplex method was originally presented at a 1949 conference and published in the proceedings [2]. One of the first published reports of a computer implementation of the simplex method for linear programming appeared in 1953 [3]. It compared the simplex method to two other algorithms, solving problems of up to 10 variables and 10 inequality constraints on the SEAC computer at the US National Bureau of Standards (now the National Institute of Standards and Technology). The basic revised simplex method—the foundation of all efficient simplex implementations—is described by Dantzig and Orchard-Hays [4]. Development of more efficient implementation techniques proceeded apace through the 1970s, but performance reached something of a plateau after that.

In 1979, Khachian [5] announced a polynomial-time algorithm for linear programming. (Most variations of the simplex method are known to require worst case running time that grows exponentially with problem size—none are known to be polynomial.) However, Khachian’s “ellipsoid method” cannot be implemented efficiently enough to compete with simplex codes on practical problems. In 1984, Karmarkar announced (with great fanfare) his interior-point algorithm [6], for which he claimed a highly efficient implementation. His announcement was received amid much

controversy, and was taken as a challenge by simplex implementers.

Dramatic improvements in both interior-point and simplex algorithms followed during the late 1980s and the 1990s. Bixby [7] details the history of computational LP from its origins to the turn of the millennium. Bixby reports that, from the first release of the CPLEX solver in 1988 through 2002, the speed of CPLEX’s simplex implementation (controlling for hardware advances) has improved by a factor of 50. Since then, progress seems to have stabilized again, though incremental improvements continue.

The basic simplex method is a simple and straightforward algorithm. But modern efficient implementations represent a tour de force of algorithm engineering and numerical methods designed to exploit opportunities for efficiency wherever they arise.

This article surveys the key techniques necessary to implement a robust, efficient simplex code. These techniques support the following strategic design elements:

- compute only what you need;
- minimize the number of iterations;
- update rather than recompute;
- exploit sparsity ruthlessly;
- control the growth of roundoff errors.

Complete descriptions of the detailed data structures and algorithms required for implementing the techniques described here are much too involved to include in this short article at a level of detail suitable for implementation. Resources for programmers desiring to create a simplex code are listed at the end of the article.

THE BOUNDED-VARIABLE LINEAR PROGRAM

The standard form of linear program for implementation handles implicitly both

upper and lower bounds as well as free variables. For *decision variables* $\mathbf{x} \in \mathbb{R}^n$ and constants $\mathbf{c}, \mathbf{l}, \mathbf{u} \in \mathbb{R}^n$, $\mathbf{b}_l, \mathbf{b}_u \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times n}$, the bounded-variable linear program is:

$$\begin{aligned} & \text{minimize} && \mathbf{c}^T \mathbf{x} \\ & \text{subject to} && \mathbf{b}_l \leq A\mathbf{x} \leq \mathbf{b}_u \\ & && \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}, \end{aligned} \quad (1)$$

where $l_j, u_j, (\mathbf{b}_l)_i$, and $(\mathbf{b}_u)_i$ need not be finite, but $l_j = u_j$ and $(\mathbf{b}_l)_i = (\mathbf{b}_u)_i$ are permitted. Most LP codes transform the input problem into one of two internal forms:

- A nonhomogeneous form (used by CPLEX, among other solvers):

$$\begin{aligned} & \text{minimize} && \mathbf{c}^T \mathbf{x} \\ & \text{subject to} && A\mathbf{x} + I\mathbf{s} = \mathbf{b} \\ & && \mathbf{l}_x \leq \mathbf{x} \leq \mathbf{u}_x, \mathbf{l}_s \leq \mathbf{s} \leq \mathbf{u}_s, \end{aligned}$$

where b_i is $(\mathbf{b}_u)_i$ or $(\mathbf{b}_l)_i$ and the *slack* (or *logical*) variable s_i is nonnegative for less-or-equal constraints, nonpositive for greater-or-equal constraints, nonnegative and bounded for range constraints, or fixed to zero for equality constraints. (Obvious variations involving choice of b_i and signs on the slacks are also possible).

- A homogeneous form (used by the COIN-OR LP (CLP) solver):

$$\begin{aligned} & \text{minimize} && \mathbf{c}^T \mathbf{x} \\ & \text{subject to} && A\mathbf{x} - I\mathbf{s} = \mathbf{0} \\ & && \mathbf{l}_x \leq \mathbf{x} \leq \mathbf{u}_x, \mathbf{b}_l \leq \mathbf{s} \leq \mathbf{b}_u. \end{aligned}$$

Both of these forms can be represented in standard bounded-variable form by merging the original slack and decision variables into a single vector $\mathbf{x}$:

$$\begin{aligned} & \text{minimize} && \mathbf{c}^T \mathbf{x} \\ & \text{subject to} && A\mathbf{x} = \mathbf{b} \\ & && \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \end{aligned} \quad (2)$$

We denote the set of row indices by $\mathcal{I} = \{1, 2, \dots, m\}$ and the set of column indices by $\mathcal{J} = \{1, 2, \dots, n\}$. A *basis* for Equation (2) is defined by a partition of $\mathcal{J}$ into ordered subsets $\mathcal{B}, \mathcal{L}, \mathcal{U}, \mathcal{F}$, called the *basis header*. Up to

column ordering, Equation (2) can be rewritten as

$$\begin{aligned} & \text{minimize} && \zeta = \mathbf{c}_B^T \mathbf{x}_B + \mathbf{c}_L^T \mathbf{x}_L + \mathbf{c}_U^T \mathbf{x}_U \\ & && + \mathbf{c}_F^T \mathbf{x}_F \\ & \text{subject to} && B\mathbf{x}_B + L\mathbf{x}_L + U\mathbf{x}_U + F\mathbf{x}_F = \mathbf{b} \quad (3) \\ & && \mathbf{l}_B \leq \mathbf{x}_B \leq \mathbf{u}_B, \\ & && \mathbf{l}_L \leq \mathbf{x}_L \leq \mathbf{u}_L, \\ & && \mathbf{l}_U \leq \mathbf{x}_U \leq \mathbf{u}_U, \\ & && \mathbf{l}_F \leq \mathbf{x}_F \leq \mathbf{u}_F, \end{aligned}$$

where the columns of the submatrix B of A are indexed by the members of $\mathcal{B}$, etc. The partition must satisfy the following criteria:

- $|\mathcal{B}| = m$ and B has rank m ;
- all members of $\mathcal{L}$ have finite lower bounds;
- all members of $\mathcal{U}$ have finite upper bounds and;
- all members of $\mathcal{F}$ have infinite upper and lower bounds.

The order of indices in each set is arbitrary but static.

We use the following notation to describe these sets: For $\beta = 1, 2, \dots, m$, let $j(\beta)$ be the original index value appearing in position β in $\mathcal{B}$, and $\beta(j)$ be the position in $\mathcal{B}$ containing index j . The members of $\mathcal{L}$ are indexed similarly by $\lambda = 1, 2, \dots, n_L = |\mathcal{L}|$; $\mu = 1, 2, \dots, n_U = |\mathcal{U}|$, and $\phi = 1, 2, \dots, n_F = |\mathcal{F}|$. It is occasionally useful to refer to the nonbasic indices as a group. Thus, we define $\mathcal{N} = \mathcal{L} \cup \mathcal{U} \cup \mathcal{F}$, $A = [BLUF] = [BN]$ (up to column order), and index the entries of $\mathcal{N}$ by $v = 1, 2, \dots, n_N = n_L + n_U + n_F = n - m$. The variables $\mathbf{x}$ and rim vectors $\mathbf{l}, \mathbf{u}$, and $\mathbf{c}$ are partitioned similarly.

Associated with each possible partition is a *basic solution* defined as follows:

$$\mathbf{x}_L = \mathbf{l}_L, \quad (4)$$

$$\mathbf{x}_U = \mathbf{u}_U, \quad (5)$$

$$\mathbf{x}_F = \mathbf{0}, \quad (6)$$

$$\mathbf{x}_B = B^{-1}\mathbf{b} - B^{-1}L\mathbf{x}_L - B^{-1}U\mathbf{x}_U - B^{-1}F\mathbf{x}_F \quad (7)$$

$$= B^{-1}\mathbf{b} - B^{-1}L\mathbf{l}_L - B^{-1}U\mathbf{u}_U \quad (8)$$

$$= B^{-1}\hat{\mathbf{b}} \quad (9)$$

$$\begin{aligned} \zeta &= \mathbf{c}_B^T B^{-1}\mathbf{b} + (\mathbf{c}_L^T - \mathbf{c}_B B^{-1}L)\mathbf{x}_L \\ &\quad + (\mathbf{c}_U^T - \mathbf{c}_B^T B^{-1}U)\mathbf{x}_U \\ &\quad + (\mathbf{c}_F^T - \mathbf{c}_B^T B^{-1}F)\mathbf{x}_F \end{aligned} \quad (10)$$

$$\begin{aligned} &= \mathbf{c}_B^T B^{-1}\mathbf{b} + (\mathbf{c}_L^T - \mathbf{c}_B B^{-1}L)\mathbf{l}_L \\ &\quad + (\mathbf{c}_U^T - \mathbf{c}_B^T B^{-1}U)\mathbf{u}_U. \end{aligned} \quad (11)$$

If $\mathbf{l}_B \leq \mathbf{x}_B \leq \mathbf{u}_B$, then this is a *basic feasible solution* (BFS). The variables in $\mathcal{B}$ are referred to as *basic*, those in $\mathcal{L}$ and $\mathcal{U}$ are *nonbasic* at their lower and upper bounds, respectively, and those in $\mathcal{F}$ are *superbasic*. More generally, a *superbasic variable* is a nonbasic variable that is not equal to either bound; a *superbasic solution* is one that contains superbasic variables.

In fact, the values of $\mathbf{x}_F$ are arbitrary. The basis headers $\mathcal{B}, \mathcal{L}, \mathcal{U}, \mathcal{F}$ together with Equations (4), (5), (7), and (10) define a family of superbasic solutions parameterized by $\mathbf{x}_F$. The convention is to set $\mathbf{x}_F = \mathbf{0}$, as this simplifies the computation of $\mathbf{x}_B$ and ζ to Equations (8) and (11), respectively.

THE BOUNDED-VARIABLE SIMPLEX METHOD

The bounded-variable simplex method works as follows:

1. *Initialization*. Identify a BFS (see the section titled “Phase 1”). Factor B , solve $B\mathbf{x}_B = \hat{\mathbf{b}} = \mathbf{b} - L\mathbf{l}_L - U\mathbf{u}_U$, and solve $B^T\mathbf{y} = \mathbf{c}_B$ (see the section titled “Basis Factorization”).
2. *Optimality Test/Entering Variable Selection (Pricing)*. If $\mathbf{c}_L - L^T\mathbf{y} \geq \mathbf{0}$, $\mathbf{c}_U - U^T\mathbf{y} \leq \mathbf{0}$, and $\mathbf{c}_F - F^T\mathbf{y} = \mathbf{0}$, then *stop* (the current BFS is optimal). Otherwise, select an index $j^* \in \mathcal{N}$ that violates these conditions. Let $\bar{c}_{j^*} = c_{j^*} - \mathbf{y}^T A_{j^*}$.

3. Direction/Ratio Test/Leaving Variable Selection.

- If $j^* \in \mathcal{L}$, solve $B\mathbf{d}_B = -L_{\cdot j^*}$.
- If $j^* \in \mathcal{U}$, solve $B\mathbf{d}_B = U_{\cdot j^*}$.
- If $j^* \in \mathcal{F}$ and $(\mathbf{c}_F - F^T\mathbf{y})_{\phi(j^*)} < 0$, solve $B\mathbf{d}_B = -F_{\cdot \phi(j^*)}$.
- If $j^* \in \mathcal{F}$ and $(\mathbf{c}_F - F^T\mathbf{y})_{\phi(j^*)} > 0$, solve $B\mathbf{d}_B = F_{\cdot \phi(j^*)}$.

Compute

$$\begin{aligned} \beta^* &= \arg \min_{\beta \in \mathcal{B}} \left\{ \min_{\beta: (\mathbf{d}_B)_\beta < 0} \frac{(\mathbf{x}_B)_\beta - (\mathbf{l}_B)_\beta}{-(\mathbf{d}_B)_\beta}, \right. \\ &\quad \left. \min_{\beta: (\mathbf{d}_B)_\beta > 0} \frac{(\mathbf{u}_B)_\beta - (\mathbf{x}_B)_\beta}{(\mathbf{d}_B)_\beta} \right\} \\ \alpha^* &= \begin{cases} \frac{(\mathbf{x}_B)_{\beta^*} - (\mathbf{l}_B)_{\beta^*}}{-(\mathbf{d}_B)_{\beta^*}} & \text{if } (\mathbf{d}_B)_{\beta^*} < 0 \\ \frac{(\mathbf{u}_B)_{\beta^*} - (\mathbf{x}_B)_{\beta^*}}{(\mathbf{d}_B)_{\beta^*}} & \text{if } (\mathbf{d}_B)_{\beta^*} > 0 \end{cases} \end{aligned}$$

$$\gamma^* = \min\{u_{j^*} - l_{j^*}, \alpha^*\}$$

If $\gamma^* = \infty$, then *stop* (the LP is unbounded).

4. Pivot and Update.

- (a) Set $\mathbf{x}_B := \mathbf{x}_B + \gamma^* \mathbf{d}_B$.
- (b) If $\gamma^* = u_{j^*} - l_{j^*}$ then move j^* from $\mathcal{L}$ to $\mathcal{U}$ or vice versa. Otherwise ($\gamma^* = \alpha^* < u_{j^*} - l_{j^*}$):
 - Add $j(\beta^*)$ to $\mathcal{L}$ (if $(\mathbf{d}_B)_{\beta^*} < 0$) or $\mathcal{U}$ (if $(\mathbf{d}_B)_{\beta^*} > 0$), and set the β^* entry of $\mathcal{B}$ to j^* .
 - Set $(\mathbf{x}_B)_{\beta^*} := \alpha^*$.
- (c) Set $\mathbf{y} := \mathbf{y} + \tau^* B_{\cdot \beta^*}^{-T}$, where $\tau^* = \bar{c}_{j^*}/(\mathbf{d}_B)_{\beta^*}$.
- (d) Set $\zeta := \zeta + \gamma^* \bar{c}_{j^*}$.
- (e) Replace $B_{\cdot \beta^*}$ with A_{j^*} and update the factorization of B .
- (f) Go to Step 2.

Pricing. The values $\mathbf{y} = B^{-T}\mathbf{c}_B$ computed in Step 2 are called the *shadow prices* of the constraints. As long as the current basis remains feasible for small changes in b_i , y_i is the rate of change of ζ per unit increase in b_i . The values $\bar{\mathbf{c}} = \mathbf{c} - A^T\mathbf{y}$ are referred to as

reduced costs of the variables. For $j \in \mathcal{N}$, $\bar{c}_j$ is the rate of change of ζ per unit increase in x_j , while maintaining feasibility with respect to the constraints $A\mathbf{x} = \mathbf{b}$, that is, per unit step in the direction $\mathbf{d} = (\mathbf{d}_B, \mathbf{d}_N)$, where $\mathbf{d}_N$ is a unit vector with ± 1 in position $\nu(j^*)$ (depending on whether x_{j^*} is increasing or decreasing).

There are usually several nonbasic variables that violate the optimality conditions in Step 2. The section titled “Pricing” describes several rules for choosing among those candidates.

Ratio Test. Note that a variable in the basis with infinite lower and upper bounds never leaves.

Ties for leaving variable may occur if more than one variable reaches a bound simultaneously. If any basic variable is equal to one of its bounds, the basis is *degenerate*. In this case, it is possible that $\gamma^* = 0$, in which case the pivot is called *degenerate* as well. In the absence of degenerate pivots, the algorithm terminates in Step 2 or Step 3 after a finite number of iterations. Modifications to handle degeneracy are described in the section titled “Degeneracy”.

Pivot. The update of B in Step 4 involves replacing the column of B corresponding to the variable leaving the basis with the column of A corresponding to the entering variable.

Note that $B\mathbf{d}_B = \pm A_{j^*}$ where the sign depends on whether the entering variable is increasing or decreasing. For simplicity, in the remainder of this discussion and in the sections titled “Steepest Edge Pricing” and “Basis Update,” we take $B\mathbf{d}_B = -A_{j^*}$. Then the update can be written as $B := BE$, where E is an *eta matrix* (η -matrix), that is, an identity matrix with one column replaced:

$$E = I - (\mathbf{d}_B + \mathbf{e}^{\beta^*}) (\mathbf{e}^{\beta^*})^T. \quad (12)$$

The effect of post-multiplying B by E is to replace the β^* column of B by A_{j^*} .

Refactoring the updated basis at every iteration would be impractical. Techniques for updating the factorization after changing

B are described in the section titled “Basis Update.”

Reducing the number of iterations is accomplished through the initialization step, the pricing step, and mitigation of degeneracy in the selection of the leaving variable. Also, the dual simplex method (see the section titled “The Bounded-Variable Dual Simplex Method”) is more efficient for many problems.

In general, the most costly operations of each iteration are the pivot/update and the direction computation (see the section titled “Sparsity and Linear Algebra”).

Roundoff control is accomplished by regularly recomputing from scratch values that are computed using updates and by careful selection of tolerances for making finite-precision comparisons (see the section titled “Controlling Roundoff Errors”).

Phase I. The problem of finding a starting BFS is usually solved using a two-phase approach. Phase II is the algorithm described above. Phase I implements the same approach, applied to a modified problem with a known starting BFS and an objective for which optimality is achieved when a BFS for the original problem is found or the original problem is proved to be infeasible.

For the bounded-variable LP (2), a known-feasible phase I problem and starting BFS can be constructed by selecting an arbitrary basis B , then fixing the nonbasic variables as follows:

- Let $j \in \mathcal{L}$ if $l_j > -\infty$ and $c_j \geq 0$ or if $u_j = \infty$.
- Let $j \in \mathcal{U}$ if $u_j < \infty$ and $c_j < 0$ or if $l_j = -\infty$.
- Let $j \in \mathcal{F}$ if $l_j = -\infty$ and $u_j = \infty$.

Next, temporarily replace bounds on $x_j, j \in \mathcal{B}$ so that basic variables that violate their original bounds are feasible and set the objective coefficients to push those variables toward their original bounds.

- If $x_j > u_j$, $l'_j := u_j$, $u'_j := \infty$ and choose $c'_j > 0$.
- If $x_j < l_j$, $u'_j := l_j$, $l'_j := -\infty$ and choose $c'_j > 0$.

- Otherwise, $l'_j := l_j$ and $u'_j := u_j$, and set $c'_j = 0$.

Numerous variations on this basic idea are possible.

A simple choice for initial $\mathcal{B}$ is the basis consisting of the slack variables added to LP (1) to produce LP (2). Numerous other methods for selecting $\mathcal{B}$ from among the original problem variables have been proposed as well. These generally go by the name *basis crashing*. Several crashes are described in Ref. 8, including those from Refs 9 and 10.

Pricing

The entering-variable selection rule that is presented in most expositions of the simplex method (including Dantzig's seminal book [11]) is to choose the entering variable that reduces the objective function the most per unit change in the entering variable. This rule (known as *Dantzig pricing*) selects the largest magnitude reduced cost among the negative reduced costs for variables at their lower bounds, positive reduced costs for variables at their upper bounds, and nonzero reduced costs for superbasic variables.

In early implementations, pricing all nonbasic variables was considered too expensive, so techniques were developed to avoid full pricing. These ideas reduced the cost per iteration, but did not address the issue of reducing the iteration count. The main techniques were combinations of *partial pricing* and *multiple pricing* (see the section titled "Partial Pricing and Multiple Pricing").

More recently, pricing techniques have been developed on the basis of choosing the entering variable that reduces the objective function the most per unit step in the corresponding direction. These *steepest edge pricing* methods require complete pricing at every iteration, but in general, substantially reduce the number of iterations required (see the section titled "Steepest Edge Pricing").

Partial Pricing and Multiple Pricing. In *partial pricing*, a contiguous subset of nonbasic variables is considered and the one with the best reduced cost is selected. In the next iteration, the next block is considered.

In *multiple pricing*, a subset of candidate nonbasic variables is considered, and the subproblem consisting of those variables plus the current basic variables is solved to optimality. Then a new subset is selected.

Numerous variations and combinations of these themes have been implemented, with varying degrees of success. For example, starting with a block of 100 variables, one might select the 10 most promising candidates, iterate considering only those candidates and the current basic variables until that subproblem is optimized, then move on to the next block of 100 variables.

One difficulty with using reduced cost as a selection criterion is that the reduced cost does not always correlate well with the actual decrease in objective value resulting from the chosen pivot. *Largest decrease* pricing looks at the actual decrease in the objective function resulting from pivoting in the candidate variable. This technique is far too expensive to apply in practice, because it requires a ratio test for every variable considered (although it can work reasonably effectively with multiple pricing).

Another difficulty with reduced costs is that they are not scale invariant. That is, scaling columns changes the reduced costs proportionally.

Steepest Edge Pricing. In *steepest edge* pricing, the direction vector $\mathbf{d} = (\mathbf{d}_B, \mathbf{d}_L, \mathbf{d}_U, \mathbf{d}_F)$ for each candidate variable is normalized. The nonbasic part $(\mathbf{d}_L, \mathbf{d}_U, \mathbf{d}_F)$ is a (positive or negative) unit vector, so $\|\mathbf{d}\|^2 = \|\mathbf{d}_B\|^2 + 1$. The entering variable is the one with the best value of $\mathbf{c}^T \mathbf{d} / \|\mathbf{d}\|$. It is easy to see that these weights are scale invariant.

Computing the norms and descent rates from scratch at each iteration would be impractically expensive, but there is an efficient update procedure [12]. Following the development in Ref. 13, suppose we know

$$\delta_j = \|\mathbf{d}_B^j\|^2 = \|B^{-1}A_{\cdot j}\|^2, \quad j \in \mathcal{N},$$

the squared norms of the basic portions of direction vectors used in the previous iteration, in which x_{j^*} entered and x_{i^*} left the basis. For convenience, denote $\beta^* = \beta(i^*)$.

Recall that the current basis matrix is formed from the previous one by $B := BE$ where the eta matrix E is formed from the identity by replacing a column, as in Equation (12). Then E^{-1} is also an eta matrix with the same column replaced:

$$E^{-1} = I + \frac{1}{(\mathbf{d}_B)_{\beta^*}} (\mathbf{d}_B - \mathbf{e}^{\beta^*}) (\mathbf{e}^{\beta^*})^T. \quad (13)$$

Thus the updated basic squared norm is

$$\begin{aligned} \delta_j &:= (\mathbf{A}_j)^T B^{-T} E^{-T} E^{-1} B^{-1} \mathbf{A}_j \\ &= (\mathbf{A}_j)^T B^{-T} \left(I + \frac{1}{(\mathbf{d}_B)_{\beta^*}} (\mathbf{e}^{\beta^*}) \right. \\ &\quad \times (\mathbf{d}_B - \mathbf{e}^{\beta^*})^T \Big) \\ &\quad \times \left(I + \frac{1}{(\mathbf{d}_B)_{\beta^*}} (\mathbf{d}_B - \mathbf{e}^{\beta^*}) (\mathbf{e}^{\beta^*})^T \right) B^{-1} \mathbf{A}_j. \end{aligned}$$

Let $\mathbf{v} = B^{-T} \mathbf{e}^{\beta^*}$ and $\mathbf{w} = B^{-T} \mathbf{d}_B$. Expanding the product above and substituting $\mathbf{v}$ and $\mathbf{w}$, the updated basic squared norms are computed from the previous values by

$$\begin{aligned} \delta_j &:= \delta_j - 2 \frac{(\mathbf{A}_j)^T \mathbf{v} (\mathbf{w} - \mathbf{v})^T \mathbf{A}_j}{(\mathbf{d}_B)_{\beta^*}} \\ &\quad + \left((\mathbf{A}_j)^T \mathbf{v} \right)^2 \frac{\|\mathbf{d}_B - \mathbf{e}^{\beta^*}\|^2}{(\mathbf{d}_B)_{\beta^*}^2}. \end{aligned}$$

Utilizing the update procedure requires maintaining complete prices at every iteration, and the additional cost of computing $\mathbf{w}$, plus the inner products involving $\mathbf{w}$ for columns where $(\mathbf{A}_j)^T \mathbf{v}$ is nonzero. (For sparse problems, $(\mathbf{A}_j)^T \mathbf{v}$ is frequently zero.) Also, memory is required for δ and the initial value of δ must be computed. The reduction in the average number of iterations versus the reduced-cost pricing methods, however, is worth the increase in cost of each iteration when the update procedure is used.

Forrest and Goldfarb [14] describe techniques to maintain the exact norms in a subspace of the variables, which they call “projected steepest edge” techniques. A dynamic projected steepest edge method maintains exact norms on a subset of

variables that can grow or shrink over the course of the algorithm.

Devex and Other Pricing Techniques. Although it was developed first, *Devex pricing* [15] can be viewed as an approximation of the steepest edge prices with a more efficient procedure. Benichou *et al.* [16] developed a modification of Devex that can be incorporated into a partial/multiple scheme. Swietanowski [17] proposes a different steepest edge approximation. Maros [8] generalizes all of the aforementioned techniques into a general framework for implementing pricing strategies.

Degeneracy

If there are ties for the leaving variable (multiple choices for β^*), then the simplex method can make a sequence of *degenerate pivots* with step size $\gamma^* = 0$, visiting a collection of *degenerate bases* (multiple bases describing the same BFS, where some basic variables take on values equal to their lower or upper bounds), and returning to a basis that has been visited previously. Thus, it is possible for the simplex method to cycle forever without terminating.

Cycling is extremely rare in practice. For highly degenerate problems, however, *stalling* (long sequences of degenerate pivots) is a frequent occurrence. Several anti-cycling techniques are known to guarantee termination of the simplex method in theory, including *random perturbation* (the “guarantee” here is probabilistic), *lexicographic perturbation*, and *Bland’s rule*.

As a practical matter, anti-degeneracy techniques add complexity to each iteration and interfere with other techniques designed to reduce the iteration count. For example, Bland’s rule dictates both entering and leaving variables, thus destroying any benefit from advanced pricing techniques. Most such techniques can be enabled when stalling is encountered and disabled when a positive reduction is made in the objective value.

The theoretical anti-cycling measures are generally not that effective in practice and they can be subject to numerical difficulties.

Several practical anti-degeneracy techniques are described in Ref. 8.

- When degeneracy is detected, tests on the entering column can be performed to identify and avoid columns that are guaranteed to result in degenerate pivots, because the pivot element corresponding to a degenerate variable is nonzero with the appropriate sign. Identifying columns that are guaranteed not to result in degenerate pivots is more difficult. However, Raymond *et al.* [18] describe techniques for selecting subproblems (rows and columns) that are likely to exhibit good behavior, and they describe computational experiments indicating that the effort invested yields improvements in number of pivots and overall solution time.
- The “ad hoc” method of Wolfe [19] systematically introduces “virtual perturbations” of the degenerate current basic values so that the basic direction is nondegenerate for those variables. Ryan and Osborne [20] provide a nice interpretation of Wolfe’s method as seeking a direction of unboundedness with respect to the degenerate constraints. They describe an implementation that effectively adds no overhead to the simplex method. Ryan and Osborne’s implementation marks the degenerate variables, then pivots against only the degenerate constraints until optimality is proved, a nondegenerate direction is found, or a degenerate pivot is encountered in the restricted constraint set. If the latter occurs, the problem is restricted recursively to the subset of degenerate constraints with respect to the current direction.
- The EXPAND method of Gill *et al.* [21] uses an expanding feasibility tolerance (see the section titled “Tolerances”) to avoid altogether the need to consider degeneracy.
- Direct perturbation of the degenerate basic variable values or their bounds during the ratio test is also effective. If the algorithm terminates with

perturbations activated, the problem is re-optimized with the perturbations removed.

THE BOUNDED-VARIABLE DUAL SIMPLEX METHOD

For many problems, a variation on the simplex method performs better than the version described above. The dual simplex method starts with a possibly infeasible basic solution that satisfies the optimality conditions, and maintains optimality through each iteration but reduces infeasibility.

The dual algorithm is critical in post-optimality sensitivity and ranging analysis and in applications such as integer programming, where an optimal LP solution is used to create one or more related problems for which that solution is infeasible (by branching or by adding cutting planes). In these situations, the related problem is generally optimized starting from that now infeasible, but still optimal, solution.

In addition, the dual simplex method often takes fewer iterations and may be less susceptible to stalling due to degeneracy than the primal algorithm. Even for solving general LPs from scratch, the dual steepest edge algorithm is generally the most efficient simplex method (though that depends on problem characteristics) [7].

The Dual LP

The dual of LP (2) is

$$\begin{aligned} &\text{maximize} && \mathbf{b}^T \mathbf{y} - \mathbf{l}^T \mathbf{w} + \mathbf{u}^T \mathbf{z} \\ &\text{subject to} && \mathbf{A}^T \mathbf{y} - \mathbf{w} + \mathbf{z} = \mathbf{c}, \\ &&& \mathbf{w} \geq \mathbf{0}, \mathbf{z} \geq \mathbf{0}. \end{aligned} \quad (14)$$

In addition, if $l_j = -\infty$ then $w_j = 0$ and if $u_j = \infty$ then $z_j = 0$. Applying complementary slackness, one can derive a one-to-one correspondence between the bases for Equations (2) and (14). Consider the basis described in

Equation (3). In dual form:

$$\begin{aligned}
& \text{maximize} \\
& \mathbf{b}^T \mathbf{y} - \mathbf{l}_B^T \mathbf{w}_B - \mathbf{l}_L^T \mathbf{w}_L - \mathbf{l}_U^T \mathbf{w}_U - \mathbf{l}_F^T \mathbf{w}_F \\
& + \mathbf{u}_B^T \mathbf{z}_B + \mathbf{u}_L^T \mathbf{z}_L + \mathbf{u}_U^T \mathbf{z}_U + \mathbf{u}_F^T \mathbf{z}_F, \\
& B^T \mathbf{y} - \mathbf{w}_B + \mathbf{z}_B = \mathbf{c}_B, \\
& L^T \mathbf{y} - \mathbf{w}_L + \mathbf{z}_L = \mathbf{c}_L, \\
& U^T \mathbf{y} - \mathbf{w}_U + \mathbf{z}_U = \mathbf{c}_U, \\
& F^T \mathbf{y} - \mathbf{w}_F + \mathbf{z}_F = \mathbf{c}_F, \\
& \mathbf{w}_B \geq \mathbf{0}, \mathbf{w}_L \geq \mathbf{0}, \mathbf{w}_U \geq \mathbf{0}, \mathbf{w}_F \geq \mathbf{0}, \\
& \mathbf{z}_B \geq \mathbf{0}, \mathbf{z}_L \geq \mathbf{0}, \mathbf{z}_U \geq \mathbf{0}, \mathbf{z}_F \geq \mathbf{0}.
\end{aligned} \tag{15}$$

Then the complementary dual basic solution to Equations (4)–(11) meets the following criteria:

- $\mathbf{y}$ is basic, $\mathbf{w}_B$ and $\mathbf{z}_B$ are not. So $\mathbf{y} = B^{-T} \mathbf{c}_B$.
- $\mathbf{w}_L, \mathbf{z}_U$ are always basic, $\mathbf{w}_U$ and $\mathbf{z}_L$ are not. So $\mathbf{w}_L = L^T \mathbf{y} - \mathbf{c}_L$ and $\mathbf{z}_U = \mathbf{c}_U - U^T \mathbf{y}$.
- If $(F^T \mathbf{y})_\phi \geq (\mathbf{c}_F)_\phi$ then $(\mathbf{w}_F)_\phi$ is basic and $(\mathbf{w}_F)_\phi = (F^T \mathbf{y} - \mathbf{c}_F)_\phi$, otherwise $(\mathbf{z}_F)_\phi$ is basic and $(\mathbf{z}_F)_\phi = (\mathbf{c}_F - F^T \mathbf{y})_\phi$.

This solution is dual feasible if and only if the optimality conditions from the primal bounded-variable simplex method are satisfied. A primal-dual complementary basic pair where both solutions are feasible is optimal for both problems.

The w_j and z_j associated with infinite lower and upper bounds, respectively, can be considered as artificial variables and given dual objective coefficients of zero. This is irrelevant for variables with one infinite bound, as the dual slack corresponding to the infinite bound is never basic. For free variables, any feasible dual solution will have both dual slacks equal to zero. For dual solutions that violate this condition, setting the dual objective coefficient to zero gives the same objective value for both complementary solutions.

The Bounded-Variable Dual Simplex Method

For simplicity, assume that $\mathbf{l}$ and $\mathbf{u}$ are finite (so $\mathcal{F} = \emptyset$) and $\mathbf{l} < \mathbf{u}$.

The dual simplex method works with the primal basic form of Equation (3), but begins with a dual-feasible basis. If the corresponding primal solution is infeasible, then an infeasible leaving variable is selected. A ratio test is performed to find an entering variable that preserves dual feasibility. Then the standard pivot step is performed. The algorithm terminates when optimality or dual unboundedness (primal infeasibility) is detected.

1. *Initialization.* Identify a dual BFS (see the section titled “Dual Phase I”). Factor B , solve $B\mathbf{x}_B = \mathbf{b} - L\mathbf{l}_L - U\mathbf{u}_U$, and solve $B^T \mathbf{y} = \mathbf{c}_B$ (see the section titled “Basis Factorization”). Compute $\bar{\mathbf{c}}_N = \mathbf{c}_N - N^T \mathbf{y}$.
2. *Dual optimality test/leaving variable.* If $\mathbf{l}_B \leq \mathbf{x}_B \leq \mathbf{u}_B$, then *stop* (the current dual BFS is primal feasible/dual optimal). Otherwise, select an index $i^* \in \mathcal{B}$ such that $x_{i^*} < l_{i^*}$ or $x_{i^*} > u_{i^*}$ (see the section titled “Dual Steepest Edge Pricing”).
3. *Direction/ratio test/entering variable.* Solve $B^T \mathbf{v} = \mathbf{e}^{\beta(i^*)}$ (where $\mathbf{e}^{\beta(i^*)}$ is the unit vector with 1 in the position in $\mathcal{B}$ of the index i^*). Compute $\mathbf{q}_L = L^T \mathbf{v}$, $\mathbf{q}_U = U^T \mathbf{v}$.
 - If $x_{i^*} < l_{i^*}$, let $\bar{\mathcal{L}} = \{j \in \mathcal{L} : q_j < 0\}$, $\bar{\mathcal{U}} = \{j \in \mathcal{U} : q_j > 0\}$.
 - If $x_{i^*} > u_{i^*}$, let $\bar{\mathcal{L}} = \{j \in \mathcal{L} : q_j > 0\}$, $\bar{\mathcal{U}} = \{j \in \mathcal{U} : q_j < 0\}$.
 If $\bar{\mathcal{L}} = \bar{\mathcal{U}} = \emptyset$ then *stop* (the dual is unbounded and the primal is infeasible). Otherwise, let $j^* = \arg \min_{j \in \bar{\mathcal{L}} \cup \bar{\mathcal{U}}} |\bar{c}_j / q_j|$. Solve $B\mathbf{d} = \mathbf{A}_{j^*}$. Let

$$\gamma^* = \begin{cases} (x_{i^*} - l_{i^*}) / q_{j^*} & \text{if } x_{i^*} < l_{i^*}, \\ (x_{i^*} - u_{i^*}) / q_{j^*} & \text{if } x_{i^*} > u_{i^*}. \end{cases}$$

4. *Pivot and update.*
 - (a) Set $\mathbf{x}_B := \mathbf{x}_B - \gamma^* \mathbf{d}$.
 - (b) • Add i^* to $\mathcal{L}$ if $x_{i^*} = l_{i^*}$ or to $\mathcal{U}$ if $x_{i^*} = u_{i^*}$, and set the $\beta(i^*)$ entry of $\mathcal{B}$ to j^* .
 - $x_{j^*} := x_{j^*} + \gamma^*$.
 - (c) Set $\bar{c}_{i^*} := -\bar{c}_{j^*} / q_{j^*}$, $\bar{c}_j := \bar{c}_j + \bar{c}_{i^*} q_j$, for $j \in \mathcal{L} \cup \mathcal{U}$.
 - (d) Set $\mathbf{y} := \mathbf{y} + \tau^* B_{\beta^*}^{-T}$, where $\tau^* = \bar{c}_{j^*} / (\mathbf{d}_B)_{\beta^*}$.

- (e) Set $\zeta := \zeta + \gamma^* \bar{c}_j^*$.
- (f) Replace B_{β^*} with A_{j^*} and update the factorization of B .
- (g) Go to Step 2.

In the case where some bounds are infinite, the corresponding dual slacks do not appear in the problem. This situation leads to a rather straightforward extension to the algorithm, but it adds significant complexity to the bookkeeping.

Dual Phase I

In many contexts (such as post-optimality analysis and integer programming), a problem with a known optimal basis is modified, and a primal-optimal (hence dual-feasible) basis for the modified problem can be easily constructed from the optimal basis for the unmodified problem. But for solving problems from scratch using dual simplex, it is necessary to construct a dual-feasible basis directly.

If all l_j and u_j are finite, there is an easy dual-feasible basis:

- Let B be the all-slack basis, so $y = c_B = 0$.
- For the remaining variables, let $j \in \mathcal{L}$ if $c_j \geq 0$, $j \in \mathcal{U}$ if $c_j < 0$.

If there is no obvious dual-feasible basis, then a dual phase I procedure must be implemented.

As with primal phase I, there are many possible variations involving changing objective function coefficients so that the starting basis is dual feasible and changing bounds to make dual feasibility for the original objective correspond to optimality for the phase I problem.

Dual Steepest Edge Pricing

The reduced costs for the dual problem are given by the values of the primal variables; basic variables that violate their bounds are candidates to leave the primal basis. Analogs of exact and approximate steepest edge pricing rules, in which the true dual ascent rates are computed instead of the ascent rates per unit change in the pivot variable, can

be derived for the dual simplex method as well [14]. To make these methods efficient, analogs of the update methods for steepest edge pricing are also necessary.

SPARSITY AND LINEAR ALGEBRA

The main expense of a simplex iteration is the requirement to solve systems of equations involving the basis matrix B . Because LPs in practice are generally extremely sparse (only a small fraction of entries in A are nonzero), implementation of solution techniques specialized for sparse, asymmetric systems is critical for good performance. It is also critical to update solutions and factorizations from iteration to iteration, rather than to re-solve from scratch.

Data Structures

Rather than store all entries in a sparse vector or matrix, it is usually advantageous with respect to both memory and running time to store only the nonzero entries and their locations. Consider the following example matrix (blanks represent zero entries):

$$A = \begin{bmatrix} 5 & 3 & & 4 & & \\ & -1 & 2 & & 1 & \\ & 3 & 3 & -1 & & \\ 1 & & & 1 & & \end{bmatrix}$$

If rows are indexed from 0 to 3 and columns from 0 to 5 (the common array indexing rule for most modern programming languages), A can be represented in column-major form using the following data structure:

cstart	0	2	5	6	8	10	11				
rindex	0	3	0	1	2	2	1	3	0	2	1
value	5	1	3	-1	3	3	2	1	4	-1	1

Here, $cstart[j]$ is the index in the other two arrays of the first entry in column j , and $cstart[n]$ is the index of the entry after the last one in the other two arrays, that is, the number of nonzeros in the

matrix. Thus, the nonzeros in column j of A are described in `rindex[cstart[j]]` through `rindex[cstart[j+1]]-1`, which contain the row indices of nonzeros in column j , and `value[cstart[j]]` through `value[cstart[j+1]]-1`, which contain the values of those entries. A similar data structure can hold the matrix entries in row-major order. Each can be constructed from the other in time linear in the number of nonzeros.

Matrix-vector multiplication ($\mathbf{y} = A\mathbf{x}$) can be carried out efficiently if A is stored in column-major order by computing a dense linear combination of columns ($\mathbf{y} = \sum_{j \in \mathcal{J}} x_j A_j$) and collecting the nonzeros in $\mathbf{y}$. If A is stored in row-major order, then inner products of $\mathbf{x}$ and rows of A can be computed efficiently by forming a dense copy of $\mathbf{x}$ and summing the pairwise products over the nonzeros in each row A_i .

A modification to this structure supports the inclusion of some buffer entries in each column so that rows can be added efficiently. The modification introduces an additional `cend` array that points to the first unused entry in each block of column entries. Obviously, an analogous structure can be used to store the matrix in row-major order.

Basis Factorization

Rather than computing B^{-1} explicitly to solve $B\mathbf{x} = \mathbf{b}$, it is preferable for both accuracy and efficiency to factor B into the product of triangular matrices, $PB = LU$, where P is a permutation matrix (permutation of the rows of the identity matrix), L is unit lower triangular, and U is upper triangular. The solution is computed by solving $L\mathbf{v} = P\mathbf{b}$ by forward substitution and then solving $U\mathbf{x} = \mathbf{v}$ by back substitution.

For “right-hand” systems, $\mathbf{y}^T B = \mathbf{c}^T$, the procedure is to solve $U^T \mathbf{w} = \mathbf{c}$ by forward substitution, then solve $L^T \mathbf{v} = \mathbf{w}$ by back substitution, and finally compute $\mathbf{y} = P^T \mathbf{v}$.

The factorization $PB = LU$ is computed using a variation of Gaussian elimination. The primary operation in Gaussian elimination is the addition of a multiple of one row to another, with the goal of zeroing a particular entry in the target row. It is possible that a row update operation introduces

a nonzero value into an entry that was previously zero. Such an entry is called *fill-in*. The two conflicting goals in selecting the sequence of operations for initial factorization are to minimize fill-in and to minimize the error in the computation of the solution.

Rudimentary LU Factorization. In Gaussian elimination, the subdiagonal elements of each column in turn are zeroed out by first swapping a pair of rows on and below the diagonal (if necessary) to place a nonzero element on the diagonal (the pivot), then adding multiples of the pivot row to the rows below to zero the entries in the pivot column. The same updates are applied to the right-hand side. The resulting system $U\mathbf{x} = \bar{\mathbf{b}}$ is solved by back substitution. It is apparent that $U = L^{-1}PB$ and $\bar{\mathbf{b}} = L^{-1}P\mathbf{b}$ for an appropriate permutation matrix P (representing the collected row interchanges) and lower-triangular matrix L^{-1} (representing the elimination steps).

In LU factorization, the row interchanges and the inverses of the row updates are recorded in P and L . The right-hand side is not updated—instead, the solution is computed with two triangular solves, as described above.

Numerical stability of Gaussian elimination is achieved by avoiding small pivot elements. The most common implementation for dense matrices utilizes *partial pivoting*: selecting the largest magnitude element on or below the diagonal in the pivot column to use as the pivot element.

The problem of choosing a pivot sequence that minimizes fill-in is NP -complete. The best-known heuristic technique for avoiding fill-in is from Markowitz [22]. Suppose we are pivoting in row k and column k . Let p_i^k be the number of nonzeros in row i in column k or greater, and q_j^k be the number of nonzeros in column j in row k or greater. Then Markowitz proposes selecting a pivot element in row i' and j' that minimizes $(p_{i'}^k - 1)(q_{j'}^k - 1)$.

The Factorization of Suhl and Suhl. There are any number of strategies for making the trade-off between numerical stability and sparsity. A particularly effective strategy for the simplex method is that of Suhl and

Suhl [23]. This method first selects pivots from singleton columns, then singleton rows. It then searches rows and columns with small numbers of nonzeros for a pivot that exceeds a dynamically adjusted threshold. If none is found, denser rows and columns are searched. If the search fails to find a suitable pivot, the factorization fails. As the factors are computed, entries below a cutoff value are set to zero.

Basis Update

Refactoring the basis matrix B from scratch at each simplex iteration would be prohibitively expensive. Fortunately, the change in B at each iteration is restricted to a single column, so that the factorization can be updated efficiently. There are two main strategies for carrying out the updates: product-form updates and rank-one updates.

The Basic Product-Form Update. The standard simplex pivot is based on an update that constructs a matrix E_k such that $B_{k+1}^{-1} = E_k^{-1}B_k^{-1}$. Pre-multiplication of $B_k^{-1}[A \ b]$ by E_k^{-1} achieves the effect of pivoting in the tableau simplex method (except for the objective row update). Recall from Equation (12) that E_k is an eta matrix, which differs from the identity matrix in exactly one column—in our case, in column β^* , corresponding to the leaving variable. That column is $\pm \mathbf{d}_B$, as computed in Step 2 of the algorithm in the section titled “Bounded-Variable Simplex Method.” The inverse of E_k is the eta matrix given in Equation (13).

Instead of actually multiplying out $E_k^{-1}B_k^{-1}$, the *product form of the inverse* simply keeps a list of the eta matrices from iteration to iteration: $B_k^{-1} = E_{k-1}^{-1}E_{k-2}^{-1}\dots E_0^{-1}B_0^{-1}$. Then $\mathbf{x} = B_k^{-1}\mathbf{b}$ is computed by pre-multiplying $\mathbf{b}$ by B_0^{-1} and then iteratively by $E_0^{-1}, E_1^{-1}, \dots, E_{k-1}^{-1}$. This procedure is often referred to as *ftran* or *forward transformation*. To compute $\mathbf{y} = B^{-T}\mathbf{c}$, pre-multiply $\mathbf{c}$ iteratively by $E_{k-1}^{-T}, E_{k-2}^{-T}, \dots, E_0^{-T}$ and then by B_0^{-T} . This procedure is known as *btran*, or *backward transformation*.

As noted in the section titled “Basis Factorization,” it is not desirable to maintain explicit inverses. Instead, the *product form of*

the basis factors keeps the *LU*-factorization of B_0 and a list of uninverted eta matrices: $L_0U_0E_0E_1\dots E_{k-1}\mathbf{x} = P_0\mathbf{b}$. This system can be solved using the following *ftran*: solve $L_0\mathbf{w} = P_0\mathbf{b}$ and $U\mathbf{v}_0 = \mathbf{w}$, then iteratively solve $E_0\mathbf{v}_1 = \mathbf{v}_0$, $E_1\mathbf{v}_2 = \mathbf{v}_1, \dots, E_k\mathbf{v}_{k+1} = \mathbf{v}_k$. Then $\mathbf{x} = \mathbf{v}_{k+1}$. Right-hand systems $E_{k-1}^TE_{k-2}^T\dots E_1^TU_0^TL_0^TP_0\mathbf{y} = \mathbf{c}$ can be solved by the following *btran*: let $\mathbf{v}_0 = \mathbf{c}$, iteratively solve $E_1^T\mathbf{v}_1 = \mathbf{v}_0$, $E_2^T\mathbf{v}_2 = \mathbf{v}_1, \dots, E_k^T\mathbf{v}_k = \mathbf{v}_{k-1}$, then solve $U_0^T\mathbf{w}_1 = \mathbf{v}_k$, $L_0^T\mathbf{w}_2 = \mathbf{w}_1$, and compute $\mathbf{y} = P^T\mathbf{w}_2$.

A difficulty with the product-form update is that the pivot element is dictated by the choice of entering and leaving variables. If the pivot is small, then numerical difficulties can result. Also, there is limited opportunity to control sparsity. A more effective update would give the implementation some control over stability and sparsity.

Rank-One Updates. Any matrix of rank one can be written as $\mathbf{p}\mathbf{q}^T$ where $\mathbf{p}$ and $\mathbf{q}$ are vectors. A rank-one update of B_k to produce B_{k+1} is accomplished by computing the product of B_k and E_k defined in Equation (12), first distributing B_k over the terms in the sum:

$$\begin{aligned} B_k E_k &= B_k \left(I - (\mathbf{d}_B^k + \mathbf{e}^{\beta^*}) (\mathbf{e}^{\beta^*})^T \right) \\ &= B_k - (B_k \mathbf{d}_B^k + B_k \mathbf{e}^{\beta^*}) (\mathbf{e}^{\beta^*})^T \\ &= B_k + (A_{j^*} - (B_k)_{\cdot\beta^*}) (\mathbf{e}^{\beta^*})^T. \end{aligned}$$

Ignoring row permutations for now, if $B_k = LU$, then $L^{-1}B_{k+1} = U + (L^{-1}A_{j^*} - U_{\cdot\beta^*}) (\mathbf{e}^{\beta^*})^T$. This matrix is not upper triangular, however—the new column represents a *spike* that extends below the diagonal:

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & + & - & - & - & - \\ & & | & * & * & * & * \\ & & | & & * & * & * \\ & & | & & & * & * \\ & & & & & & * \end{bmatrix}.$$

There are several techniques for restoring the upper-triangular structure of U . The Bartels-Golub [24], Forrest-Tomlin [25], and Saunders [26] methods all proceed by permuting

the spike to the last column (shifting intervening columns to the left), and then permuting the pivot row to the bottom of the matrix (shifting lower rows up):

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & * & * & * & * & | \\ & & & * & * & * & | \\ & & & & * & * & | \\ & & & & & * & | \\ & - & - & - & - & + & \end{bmatrix}.$$

Reid [27] shifts the spike to the left only as far as necessary:

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & * & * & * & | & * \\ & & & * & * & | & * \\ & & & & * & | & * \\ & - & - & - & + & - & * \end{bmatrix}.$$

The methods then differ in how they perform the elimination of the spike row. Bartels-Golub performs partial pivoting. The others take steps to limit fill-in. In these methods, fill-in is restricted to the L matrix.

Suhl and Suhl [28] shift the spike column to the left as far as necessary to make the resulting matrix upper Hessenberg; that is, upper triangular except for possible nonzeros in the first subdiagonal:

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & - & - & - & + & - \\ & & * & * & * & | & * \\ & & & * & * & | & * \\ & & & & * & | & * \\ & & & & & & * \end{bmatrix}.$$

The rows below the pivot row are used to eliminate the pivot row entries to the left of the relocated spike:

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & & & & + & - \\ & & * & * & * & | & * \\ & & & * & * & | & * \\ & & & & * & | & * \\ & & & & & & * \end{bmatrix}.$$

Then the pivot row is shifted down to restore the upper triangular structure:

$$\begin{bmatrix} * & * & * & * & * & * & * \\ & * & * & * & * & * & * \\ & & * & * & * & | & * \\ & & & * & * & | & * \\ & & & & * & | & * \\ & & & & & + & - \\ & & & & & & * \end{bmatrix}.$$

Pivots are stored in L , which is kept as a collection of column and row eta matrices.

Again, the details of implementing this procedure efficiently are too involved to present here, but they can be found in Ref. 8.

Other Sparse Operations

Various steps in the algorithm require operations with vectors and matrices. Depending on the nature of the operations and on the problem characteristics, it may be possible to exploit vector sparsity in the various solves and updates. The detailed algorithms need to be analyzed to see where sparsity can be exploited.

CONTROLLING ROUND-OFF ERRORS

Tolerances

Most present-day floating-point implementations are based on the IEEE-754 standard. An excellent introduction to floating-point arithmetic is Ref. 29.

The representation of floating-point numbers in computers is of limited precision. This means that the arithmetic on those quantities is subject to rounding errors. Over the course of a sequence of several such operations, these errors can accumulate enough to render the final result significantly inaccurate or even meaningless. In addition, comparisons of the results of such computations to other results or to fixed values must allow for small errors. The selection of tolerances for such comparisons and for accuracy tests is critical to the reliability of a simplex method implementation. It is also something of an art.

The need for tolerances in the simplex method arises at several points in the algorithm:

- *Optimality test/entering variable.* Is $\bar{c}_j$ for $j \in \mathcal{N}$ significantly different from zero?
- *Direction/ratio test/leaving variable.* Is a basic variable close to or outside one of its bounds? Is a candidate pivot element too close to zero? Is there more than one candidate for leaving variable?
- *Pivot and update.* Has the factorization or basis update resulted in a fortuitous cancellation? Are the current factorizations and solutions accurate enough; that is, are $Bx_B = \hat{b} = b - Nx_N$ and $B^T y = c_B$?

For example, for standard double-precision arithmetic with 15-digit precision, Murtagh [30] suggests values of 10^{-5} for the optimality tolerance, 10^{-8} for pivot rejection, and 10^{-6} for the tolerance on $\|\hat{b} - Bx_B\|_\infty$. Murtagh also suggests 10^{-10} for annihilating a matrix coefficient, but the Suhl and Suhl factorization and update described in the sections titled “Basis Factorization” and “Basis Update” use a dynamic annihilation tolerance.

The use of tolerances in determining the leaving variable is somewhat more complex. For feasibility, Maros [8] suggests a tolerance in the range $10^{-8} \leq \epsilon_f \leq 10^{-6}$. This value is used in the computation of β^* to find the distances for each basic variable to slightly relaxed lower or upper bounds.

Harris [15] proposes a technique for applying this relaxed ratio to improve numerical stability. Her idea is to use the minimum relaxed ratio as a bound on acceptable exact ratios, and to choose the numerically most attractive pivot from among candidate variables with an exact ratio smaller than this bound. Note that this method allows the solution to take on small infeasibilities (on the order of ϵ_f), which must eventually be eliminated. Some of the anti-degeneracy techniques mentioned in the section titled “Degeneracy” (notably the EXPAND procedure [21]) are designed to overcome this difficulty.

Refactoring

As updates accumulate over several iterations, the accuracy of `ftran` and `btran`—as well as x_B , y , and any other updated values—deteriorates and the cost of carrying them out increases. When the accuracy gets too low or the cost gets too high, the updates must be discarded. The basis matrix is then refactored from scratch, and the primal and dual solutions are recomputed using the fresh factorization. Product-form updates require refactorization after 30–50 pivots. The triangular updates reduce the frequency of refactorization to every 100–150 iterations.

Scaling

One source of numerical difficulties is a large range of coefficient values in the instance input. Model builders prefer to work in the natural units associated with the quantities they are using, but the choice of units can result in large differences in the magnitudes of coefficients.

Often, scaling the rows and columns before solving the LP can alleviate much of the resulting numerical instability. Of course, scaling columns requires unscaling the solution before reporting it. If the unscaled solution does not satisfy termination criteria for the unscaled problem, the unscaled problem is resolved starting from the scaled problem’s final basis.

Note that for standard floating-point representations on current computers, a scale factor that is a power of 2 is desirable, so that precision of the mantissa is preserved. In addition, scaling is not always advantageous in the simplex method—its usefulness depends on problem characteristics. For example, in problems with many zeros and ones, fortuitous cancellation can be exploited, but scaling may destroy that structure.

Common scaling methods include:

- *Equilibration.* In equilibration scaling, each row is scaled so that the largest element has magnitude 1; then each column is also scaled the same way. Thus, each row and column has the

maximum magnitude of any entry equal to 1.

- *Mean-Based Scaling.* Here, the geometric mean of the smallest and largest nonzero magnitudes or the arithmetic mean of nonzeros (or the nearest power of 2) is used to scale rows, then columns.
- *Curtis and Reid's Method.* Curtis and Reid [31] propose a conjugate gradient method to compute a scale factor that approximately minimizes $\sum_{i,j:a_{ij} \neq 0} \log^2 |a_{ij}|$. The method converges rapidly and generally requires negligible time compared to that required to solve the LP.

Often, the methods are used in combination and applied repeatedly some small number of times. However, a study by Elble and Sahinidis [32] indicates that equilibration scaling alone generally provides as much benefit as combined techniques.

OTHER ISSUES

Preprocessing

Often, problem size and difficulty can be reduced by preprocessing ("presolving") the problem. Brearley *et al.* [33] proposed the basic techniques for detecting redundant constraints and bounds and removing or tightening them, fixing variables, converting constraints to bounds, and converting columns to bounds on shadow prices. The procedures are applied repeatedly until no further changes are detected or for a fixed number of passes, and each transformation is recorded so that the solution to the presolved problem can be transformed to a solution to the original problem. At the end of the solution process, the changes are backed out and the solution to the original problem is verified or repaired.

Parallelism

The simplex method, with all sparsity-exploiting techniques in play, does not offer many opportunities for large-grain parallelism of the sort that is well suited for cluster computing.

There are some opportunities for exploiting more fine-grained parallelism for highly rectangular problems. The most obvious place is in pricing for problems with many variables, which can exploit multicore CPUs or symmetric multiprocessing machines with multiple CPUs accessing shared memory via a high-speed bus. Here, the relatively expensive complete pricing operation in steepest edge techniques can be farmed out across CPUs.

The ratio test is another trivially parallelizable step, but it is also cheap, so the savings from parallelizing that procedure must be traded off carefully against the cost of starting up and managing the parallel tasks.

RESOURCES

There are several books on linear programming that focus on [8,30,34] or devote significant space to [35,36] implementation issues. Of the former, Maros's book [8] is the most current and complete, describing in detail the algorithms and data structures required to implement nearly all of the techniques described here.

Maros's book does not include code, but there are several open-source or source-available codes accessible on the World Wide Web. (Proprietary codes are not described here, as they do not provide source code that could be studied to learn about implementations.)

- The COIN-OR LP solver (CLP) (<http://projects.coin-or.org/Clp/>) is an open-source C++ code written by John Forrest of IBM and others. It is the preferred LP solver for the COIN-OR Branch-and-Cut (CBC) solver (<http://projects.coin-or.org/Cbc/>), also written by Forrest.
- The Dynamic LP solver (DyLP) (<http://projects.coin-or.org/DyLP/>) is an open-source C code written by Lou Hafer of Simon Fraser University. It is the LP solver in Hafer's bonsaiG mixed integer programming (MIP) code.
- LPSolve (<http://lpsolve.sourceforge.net/>) is an open-source linear and mixed

integer programming solver written in C originally by Michel Berkelaar and currently maintained by Kjell Eikland and Peter Notebaert.

- The GNU Linear Programming Kit (GLPK) (<http://www.gnu.org/software/glpk/>) is an open-source C code for linear and MIP developed by Andrew Makhurin.
- SoPlex (<http://soplex.zib.de/>) is free for academic and nonprofit use. It is an implementation of the simplex method in C++, developed at Konrad-Zuse-Zentrum für Informationstechnik Berlin.
- Robert Vanderbei makes educational implementations of simplex and interior-point methods in C available (<http://www.orfe.princeton.edu/~rvdb/LPbook/src/>) in conjunction with his textbook [13].

Related articles in this encyclopedia include the following:

- General LP topics
 - *An Introduction to Linear Programming*
 - *History of LP Development*
 - *LP Duality and KKT Conditions for LP*
- Simplex-related topics
 - *Dual Simplex*
 - *Randomized Simplex Algorithms*
 - *Simplex-Based LP Solvers*
 - *The Simplex Method and Its Complexity*
- Other algorithmic LP topics
 - *Parametric LP Analysis*
 - *Sensitivity Analysis in Linear Programming*
 - *Central Path and Barrier Algorithms for Linear Optimization*
 - *Self-Dual Embedding Technique for Linear Optimization*
 - *Interior-Point Linear Programming Solvers*
 - *Introduction to Polynomial Time Algorithms for LP*

- *Linear Programming Projection Algorithms*
- *Lagrangian Optimization for LP*

REFERENCES

1. Dantzig GB. Programming in a linear structure. Technical report. Washington (DC): Comptroller, USAF; 1948.
2. Dantzig GB. Maximization of a linear function of variables subject to linear inequalities. In: Koopmans TC, editor. Activity analysis of production and allocation. New York: Wiley; 1951. pp. 330–335.
3. Hoffman A, Mannos M, Solokowsky D, *et al.* Computational experience in solving linear programs. SIAM J 1953;1:1–33.
4. Dantzig GB, Orchard-Hays W. Notes on linear programming: Part V—alternate algorithm for the revised simplex method using product form for the inverse. Technical Report Research Memorandum RM-1268. Santa Monica (CA): The RAND Corporation; 1953;
5. Khachian LG. A polynomial algorithm in linear programming. Sov Math Dokl 1979;20:191–194.
6. Karmarkar N. A new polynomial time algorithm for linear programming. Combinatorica 1984;4(4):373–395.
7. Bixby R. Solving real-world linear programs: a decade and more of progress. Oper Res 2002;50(1):3–15.
8. Maros I. Computational Techniques of the Simplex Method. Boston (MA): Kluwer; 2003.
9. Bixby R. Implementing the simplex method: the initial basis. ORSA J Comput 1992; 4(3):267–284.
10. Gould N, Reid J. New crash procedures for large systems of linear constraints. Math Program 1989;45:475–501.
11. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
12. Goldfarb D, Reid JK. A practical steepest edge simplex algorithm. Math Program 1977;12:361–371.
13. Vanderbei R. Linear Programming: Foundations and Extensions. 3rd ed. New York: Springer; 2008.
14. Forrest J, Goldfarb D. Steepest edge simplex algorithms for linear programming. Math Program 1992;57(3):342–374.

15. Harris P. Pivot selection methods of the Devex LP code. *Math Program* 1973;5:1–28.
16. Benichou M, Gautier J, Hentges G, Ribiere G. The efficient solution of large-scale linear programming problems. *Math Program* 1977;13:280–322.
17. Swietanowski A. A new steepest edge approximation for the simplex method for linear programming. *Comput Optim Appl* 1998; 10(3):271–281.
18. Raymond V, Soumis F, Orban D. A new version of the improved primal simplex for degenerate linear programs. *Comput Oper Res* 2010;37:91–98.
19. Wolfe P. A technique for resolving degeneracy in linear programming. *SIAM J Appl Math* 1963;11:205–211.
20. Ryan DM, Osborne MR. On the solution of highly degenerate linear programmes. *Math Program* 1988;41:385–392.
21. Gill P, Murray W, Saunders M, *et al.* A practical anti-cycling procedure for linearly constrained optimization. *Math Program* 1989;45:437–474.
22. Markowitz H. The elimination form of the inverse and its application to linear programming. *Manage Sci* 1957;3(3):255–269.
23. Suhl U, Suhl L. Computing sparse LU factorizations for large-scale linear programming problems. *ORSA J Comput* 1990;2(4): 325–335.
24. Bartels R, Golub G. The simplex method of linear programming using LU decomposition. *Commun ACM* 1969;12:266–268.
25. Forrest J, Tomlin J. Updating triangular factors of the basis to maintain sparsity in the product-form simplex method. *Math Program* 1972;2:263–278.
26. Saunders M. A fast and stable implementation of the simplex method using Bartels-Golub updating. In: Bunch JR, Rose DJ, editors. *Sparse matrix computation*. New York: Academic Press; 1976. pp. 213–226.
27. Reid J. A sparsity-exploiting variant of the Bartels-Golub decomposition for linear programming bases. *Math Program* 1982;24(1):55–69.
28. Suhl U, Suhl L. A fast LU update for linear programming. *Ann Oper Res* 1993;43: 33–47.
29. Overton ML. *Numerical computing with IEEE floating point arithmetic*. Philadelphia: SIAM; 2001.
30. Murtagh B. *Advanced linear programming: computation and practice*. New York: McGraw-Hill; 1981.
31. Curtis R, Reid J. On the automatic scaling of matrices for Gaussian elimination. *J Inst Math Appl* 1972;10:118–124.
32. Elble JM, Sahinidis NV. *Scaling linear programs prior to the application of the simplex method*. Working Paper, Pittsburgh (PA): Carnegie Mellon University; 2010.
33. Brearly A, Mitra G, Williams H. Analysis of mathematical programs prior to applying the simplex method. *Math Program* 1975;8:54–83.
34. Nazareth J. *Comput Solut Linear Programs*. New York: Oxford University Press; 1987.
35. Chvátal V. *Linear programming*. New York: Freeman Press; 1983.
36. Gill P, Murray W, Wright M. *Numerical linear algebra and optimization*. New York: Addison Wesley; 1991.

IMPOSSIBILITY THEOREMS AND VOTING PARADOXES IN COLLECTIVE CHOICE THEORY

ELIZABETH MAGGIE PENN
Department of Political Science,
Washington University,
St. Louis, Missouri

In the field of positive political theory, the term *impossibility theorem* refers to a collection of results that take as given a set of normatively appealing criteria and deductively prove that such a set of criteria are inconsistent: that is, it is impossible for a voting system—any voting system—to simultaneously satisfy the given set of criteria. This approach to the study of voting systems is called *axiomatic*, because the normative criteria that we might like our voting systems to satisfy are used as starting points in these results, or taken as a given. In other words, these normative criteria are *axioms*, which in mathematics are defined as statements that are assumed solely for the sake of studying the consequences that follow from them.

The mathematical study of voting systems is motivated by the fact that any group seeking to make a collective decision must choose some method of translating the preferences of the group into a societal outcome. Moreover, there are an infinite number of such methods: from simple majority rule, to weighted voting, to choosing a “dictator” to unilaterally make societal decisions. A brief consideration of the types of voting rules that have been used throughout the course of human history reveals an assortment that varies not only with respect to procedure and formula but also with respect to the goals that each system was intended to further. In ancient Athens, for example, the election of delegates was done by lottery, and it is estimated that half of all citizens over the age of 30 served as public officers at some point in their lives. This process formed the basis of Aristotle’s concept of democracy, which espoused “ruling and being ruled by turns” [1, pp. 20–22].

Conversely, many modern elections utilize runoffs to ensure that electoral winners have received the support of a majority of voters. An advisory board may utilize weighted voting to ensure that more knowledgeable experts have more say in a final decision.

The concepts of “ruling by turns,” “winners should receive the support of a majority,” and “greater voting weight to more knowledgeable individuals” can all be thought of as decision-making desiderata of the particular groups described above. Axiomatic approaches to the study of voting systems are informative and useful because they take these types of *goals* of a voting system as a given, and focus on whether any voting system is capable of satisfying the goals and, if so, what such a voting system should look like. For example, suppose a group wishes to utilize a voting system in which no individual has an incentive to misrepresent his or her true preferences, in other words, a voting system in which all individuals are always best served by voting truthfully. The Gibbard–Satterthwaite theorem, an important impossibility theorem, would then tell us that this goal is impossible to attain with any reasonably democratic voting system.

The remainder of this article presents a brief survey of several of the most well-known impossibility theorems, including Arrow’s theorem and the Gibbard–Satterthwaite theorem mentioned above. This article also discusses various widely known *voting paradoxes*, or particular examples of how seemingly sensible voting procedures can produce undesirable outcomes. These paradoxes can be thought of as impossibility results in their own right, as they demonstrate an incompatibility between sensibility in procedure and sensibility in outcome.

NOTATION AND DEFINITIONS

The most well-known impossibility theorems in positive political theory involve voting systems, or systems of *preference aggregation*. Thus, the theorems concern

procedures that take a collection of individuals with heterogeneous preferences as an input, and produce some collective preference ordering or collective choice as an output. It should be noted, however, that there are many other impossibility theorems in positive political theory that do not explicitly involve preference aggregation, *per se*.

To present these results, some simple notation is needed. First, we assume that there exists a collection of individuals seeking to make a collective choice. This group is termed N , and it is of size n . We assume that N contains at least two individuals, or that $n \geq 2$. Second, we assume that there is a finite set of alternatives, or policies, under consideration by the group. We call this set X , and assume (to begin with) that X contains at least three different alternatives. And third, we assume that each individual has his or her own preference ordering of the alternatives under consideration. This *preference relation* is denoted $\succeq_i$ for person i . Thus, if person i likes alternative x more than alternative y , it is denoted by $x \succeq_i y$.

We assume that each individual's preference relation $\succeq_i$ is a *weak order*, which means that it is a way of comparing pairs of alternatives that is *reflexive*, *transitive*, and *complete*. The $\succeq_i$ relation is intended to be similar to $\geq$, the "greater than or equal to" relation, which we use to compare numbers on the real line. Reflexivity means that for any alternative x , it is the case that $x \succeq_i x$, or that i likes x at least as much as x . Completeness means that for any two alternatives, x and y , it is the case that either $x \succeq_i y$, or that $y \succeq_i x$, or both. In other words, i either likes x at least as much as y , or y at least as much as x , or is indifferent between the two. Last, transitivity means that if $x \succeq_i y$ and $y \succeq_i z$, then $x \succeq_i z$. This final condition is a rationality condition; voters cannot have preferences that "cycle" in the sense of ordering the alternatives $x \succ y$, $y \succ z$, and $z \succ x$.

At times we decompose $\succeq_i$ into two parts: $\succ_i$ and $\sim_i$. While $\succeq_i$ represents Person i 's preference relation, $\succ_i$ represents his *strict* preferences, and $\sim_i$ represents his *indifference*. Thus, $x \succ_i y$ implies that $x \succeq_i y$, but not $y \succeq_i x$, or that i strictly prefers x to y . Similarly, $x \sim_i y$ implies that both $x \succeq_i y$ and

that $y \succeq_i x$, or that i is indifferent between x and y . Last, we use the term ρ to describe the entire collection of individual preferences. Thus, $\rho = (\succeq_1, \succeq_2, \dots, \succeq_n)$.

More concretely, suppose we have two people in our group, $N = \{1, 2\}$, and that we have three alternatives under consideration, $X = \{x, y, z\}$. Suppose Person 1 prefers $x \succ_1 y \succ_1 z$ and Person 2 has preferences $y \succ_2 x \succ_2 z$. In other words, Person 1 strictly prefers x to y and z , and strictly prefers y to z , and Person 2 strictly prefers y to z and x , and z to x . Then

$$\rho = \begin{matrix} x \succ_1 y \succ_1 z \\ y \succ_2 z \succ_2 x \end{matrix} \quad (1)$$

Thus, ρ characterizes our group and the preferences of its members.

Now that we have described the collection of alternatives, people, and their preferences over the alternatives, we can begin to consider various ways of conceptualizing collective choice. We consider two different types of mechanisms for generating a group choice. The first is termed a *preference aggregation rule*, and is denoted f . This mechanism takes a preference profile ρ as an input, and generates a group preference relation, $\succeq$, over all alternatives. This group preference relation will be reflexive and complete, but not necessarily transitive.

As an example, suppose f is the method of Borda count (let us call it f_B), and ρ is the profile described in Equation (1). As there are three alternatives under consideration, Borda count works as follows: an alternative that a voter ranks first receives two points, an alternative he ranks second receives one point, and an alternative he ranks third receives zero points. The social ranking is then the sum of these scores across individuals. Thus, given the ρ in Equation (1), x receives two total points (two from Voter 1 and zero from Voter 2), y receives three total points (one from Voter 1 and two from Voter 2), and z receives one total point (from Voter 2). It follows that Borda count ranks the alternatives $y \succ x \succ z$, or more formally,

$$f_B(\rho) = y \succ x \succ z.$$

While a preference aggregation rule determines a social preference ordering of the

entire collection of alternatives under consideration, a *choice function*, F , takes a preference profile as an input and simply generates a single winner. We can also think of Borda count as choice function, F_B . In this case, our function would return the alternative with the highest Borda score, breaking a tie if need be. Using our profile ρ in Equation (1),

$$F_B(\rho) = y.$$

A choice function or preference aggregation rule satisfies a property termed *unrestricted domain* if its domain is the set of all possible preference profiles. In other words, unrestricted domain requires that these collective choice mechanisms be capable of yielding a social choice or a social preference relation when provided with *any* possible collection of preference orderings of the individuals; we cannot restrict the preferences that individuals may have. For the purposes of this article, we always assume that our preference aggregation rules and choice functions satisfy this property.

A SURVEY OF IMPOSSIBILITY THEOREMS IN VOTING THEORY

This section first briefly presents three impossibility theorems concerning preference aggregation rules: Arrow's theorem, May's theorem, and Sen's theorem. Arrow's theorem is the most well known of the impossibility theorems, and partially motivates the theorems that follow. Arrow lays out four simple axioms that any reasonable aggregation rule should (arguably) satisfy, and then proves that these axioms are incompatible with each other; no rule can simultaneously satisfy all four. May looks specifically at majority rule over two alternatives, and shows that it is the only voting system that can simultaneously satisfy several desirable axioms including *anonymity* and *neutrality*, the properties of treating all voters and all alternatives candidates equally. Sen demonstrates that a very minimal requirement that a voting rule be responsive to the preferences of more than one individual is incompatible with that rule being capable of generating a social ranking

of the alternatives that is both efficient and capable of yielding a best alternative.

This section concludes with a discussion of the Gibbard–Satterthwaite theorem, an impossibility theorem concerning choice functions. Loosely speaking, this theorem tells us that it is impossible to design a system for generating a collective choice that both takes the preferences of more than one individual into account and provides no incentive for individuals to misrepresent their preferences.

Arrow's Theorem

We begin by laying out Arrow's four axioms, or conditions that any reasonable aggregation rule should satisfy. Perhaps one of the least demanding requirements a group would seek to impose on its voting system is that a group decision is minimally responsive to the preferences of the members of that group. To this end, Arrow first defines the condition of *Pareto efficiency*.

An aggregation rule f is *Pareto efficient* if, for any profile ρ , if $x \succ_i y$ for all individuals i in N , then $x \succ y$. In words, this condition says that if *every* individual i strictly prefers x to y , then our aggregation rule f should generate a collective ranking of the alternatives that ranks x strictly higher than y . This condition rules out aggregation rules that, for example, always rank $x \succ y$, regardless of the group's x, y preferences.

Second, we might want our voting system to not consider “irrelevant” alternatives when generating a social ranking between two alternatives. In other words, the group members' preferences between alternatives c and d should not affect the social ranking of two different alternatives, a and b . This leads to Arrow's second condition of *independence of irrelevant alternatives*.

An aggregation rule f is *independent of irrelevant alternatives* (IIA) if for any two profiles, ρ and ρ' , if every individual's x, y ranking under ρ agrees with their x, y ranking under ρ' , then the social x, y ranking generated by $f(\rho)$ should agree with the social x, y ranking generated by $f(\rho')$. The example below is intended to make this concept more concrete, and to illustrate why the Borda

count procedure violates IIA. Consider the following two profiles, ρ and ρ' :

$$\rho = \begin{matrix} x \succ_1 y \succ_1 z \\ y \succ_2 z \succ_2 x \end{matrix}, \quad \rho' = \begin{matrix} x \succ_1 z \succ_1 y \\ y \succ_2 x \succ_2 z \end{matrix}. \quad (2)$$

As discussed earlier, Borda count produces the social ranking $y \succ x \succ z$ when applied to profile ρ . However, at profile ρ' , $f_B(\rho')$ generates the social ranking $x \succ y \succ z$, because at this profile x receives three combined points, y receives two, and z receives one. If we look solely at the two individuals' x, y rankings, these two profiles look identical: Voter 1 prefers $x \succ_1 y$ under both ρ and ρ' , and Voter 2 prefers $y \succ_2 x$ under both ρ and ρ' . However, $f_B(\rho)$ generates $y \succ x$, while $f_B(\rho')$ generates $x \succ y$. This is a violation of IIA.

Arrow's third axiom concerns the ability of a preference aggregation rule to generate an unambiguous winner, or a collection of winners, if there is a tie. A procedure that, for example, generates the social ranking $x \succ y$, $y \succ z$, and $z \succ x$ might not be particularly useful to a group seeking to make a collective choice, because it does not generate an unambiguously "best" collection of alternatives. Any given alternative is strictly worse than another. This leads to Arrow's third condition of *transitivity*.

An aggregation rule f is *transitive* if it always produces a transitive ordering of the alternatives. Thus, if f produces an ordering in which $x \succeq y$ and $y \succeq z$, then it must also be the case that $x \succeq z$. This condition guarantees that the social ordering generated by f cannot cycle, in the sense that it cannot produce a ranking in which $x \succ y$, $y \succ z$, and $z \succ x$; it guarantees that the social preference ordering satisfies the same rationality condition as the individual preference orderings that it was constructed from. Moreover, it ensures the existence of an alternative or collection of alternatives that are not strictly worse than anything else.

Last, Arrow's axiom of *no dictator* concerns the responsiveness of the preference aggregation rule to the preferences of more than one individual. An aggregation rule is *dictatorial* if there is one particular voter

whose individual preferences always determine the social preference ordering, irrespective of the preferences of the other voters. Formally, this condition says that there exists one voter i , so that every time $x \succ_i y$, the aggregation rule f produces the ranking $x \succ y$. An aggregation rule f satisfies *no dictator* if it is not dictatorial. We are now ready to state Arrow's theorem.

Theorem 1 [Arrow, 1950 and 1963]. *With three or more alternatives, any aggregation rule satisfying unrestricted domain, Pareto efficiency, IIA, and transitivity is dictatorial.*

Arrow's theorem then tells us that if a group wishes to design a preference aggregation rule that is Pareto efficient, transitive, and independent of irrelevant alternatives, and if we place no restrictions on the preferences that individuals may have, then the rule must grant all decision-making authority to a single individual [2,3]. Thus, any aggregation rule that satisfies no dictator *must* violate either transitivity, or Pareto efficiency, or IIA. And practically speaking, it will violate either transitivity or IIA, as the only nondictatorial aggregation rules that are ruled out by the addition of Pareto efficiency are those rules that are either null (generate a tie over all alternatives) or inverse dictatorships. This extension of Arrow's theorem to non-Pareto efficient rules was proved by Wilson [4], and is known as *Wilson's impossibility theorem*.

May's Theorem

May's Theorem comes as a response to the axiom of independence of irrelevant alternatives utilized by Arrow, which requires that group choices between pairs of alternatives depend only on individual preferences over those pairs. May argues that this condition implies that knowing a group's choice between every pair of alternatives is akin to knowing the group's entire preference relation; in this sense, Arrow's result can be thought of as a problem concerning choice over pairs. Thus, while Arrow's Theorem requires at least three alternatives in the group's choice set, May focuses solely on the

case in which there are only two alternatives under consideration.

May lays out three axioms that any preference aggregation rule should satisfy when there are two alternatives under consideration. His first condition is *anonymity*, which states that the preference aggregation rule f should treat every voter equally. This means that in $f(\succeq_1, \succeq_2, \dots, \succeq_n)$ any two arguments (e.g., $\succeq_5$ and $\succeq_3$) could be interchanged without changing the group preference relation produced by f . In other words, changing the voters' names (as defined by the subscripts on $\succeq_i$) can have no effect on how f ranks the alternatives.

May's second condition is *neutrality*, which says that the aggregation rule f does not favor either alternative. In other words, the labeling of the alternatives cannot matter to f . May's final condition is *positive responsiveness*, which says that the group preference relation should respond to individual preferences in a positive way. Thus, if we have a preference profile ρ in which $f(\rho)$ generates $x \succeq y$, then changing a single individual's preferences such that he is *more* favorable to x (i.e., changing his preferences from $y \succ_i x$ to $x \succeq_i y$, or from $x \sim_i y$ to $x \succ_i y$) should make the group strictly favor x over y (i.e., change the group preference from $x \sim y$ to $x \succ y$, or keep the group at $x \succ y$ if that was the group's ordering before i 's preferences changed).

Last, define the *method of majority rule* to be a preference aggregation rule, f_M over a pair of alternatives, x and y , in which $x \succ y$ if more people strictly prefer x to y than y to x , and $y \succ x$ if more people strictly prefer y to x than x to y , and $x \sim y$ if the same number of people prefer x to y as prefer y to x . We are now ready to state May's theorem.

Theorem 2 [May, 1952]. *The only preference aggregation rule that is anonymous, neutral, and positively responsive is the method of majority rule.*

A particularly succinct explanation of May's theorem is provided by May himself: "In Arrow's terms our theorem may be expressed by saying that any [preference aggregation rule] that is not based on simple majority decision, i.e., does not decide

between any pair of alternatives by majority vote, will either ... favor one individual over another, favor one alternative over the other, or fail to respond positively to individual preferences." [5, p. 683].

Sen's Theorem (the Liberal Paradox)

Like May's theorem, Sen's theorem can also be viewed as building upon Arrow's results. Sen argues that certain kinds of choices are personal, and that individuals should have the freedom to make these kinds of choices for themselves. In this vein, Sen defines an axiom termed *minimal liberalism*, which states that there are at least two individuals and two pairs of distinct alternatives (x, y) and (w, z) such that Person 1 is free to decide between (x, y) and Person 2 is free to decide between (w, z) . Minimal liberalism is a particular version of Arrow's condition of no dictator, as it characterizes a situation in which decision-making power is minimally decentralized. Sen shows that even a weak condition such as minimal liberalism is incompatible with a procedure that is capable of both generating transitive outcomes and being Pareto efficient.

Theorem 3 [Sen, 1970]. *No aggregation rule that produces transitive outcomes can simultaneously satisfy unrestricted domain, Pareto efficiency, and minimal liberalism.*

Sen provides the following illustration of his theorem. Consider a social choice over three alternatives: either Mr A can read a copy of *Lady Chatterly's Lover* (alternative a), or Mr B can read it (alternative b), or neither of them can read it (alternative c). Mr A, the prude, prefers neither of them reading it to himself reading it (in order to protect the impressionable Mr B), to Mr B reading it. Thus he prefers $c \succ_A a \succ_A b$. Mr B, the lascivious, prefers Mr A reads it (in order for Mr A to be exposed to Lawrence's prose), to himself reading it, to neither of them reading it, and thus prefers $a \succ_B b \succ_B c$.

An argument can be made that Mr. A should decide between the (a, c) states of the world, the states in which only he reads the book or neither reads the book. Similarly,

Mr. B should decide between the (b, c) states of the world. Thus, the condition of minimal liberalism is satisfied, and yields the social ordering $b \succ c$ (by Mr. B's preferences), and $c \succ a$ (by Mr. A's preferences). However, both individuals prefer a to b , and so Pareto efficiency along with minimal liberalism yields the cycle $a \succ b, b \succ c$, and $c \succ a$. Thus, the procedure is incapable of generating an optimal choice [6, pp. 80–81].

The Gibbard–Satterthwaite Theorem

Our final impossibility theorem, proved independently by Gibbard [7] and Satterthwaite [8], differs from the previous theorems in several important ways. First, it concerns choice functions rather than preference aggregation rules (i.e., voting systems that produce a single winner, as opposed to a social ordering of the alternatives), and second, it does not assume that a voting system takes a “true” collection of preferences as an input. Rather, it considers voting systems that take individuals' reported preferences, or ballots, as an input.

Gibbard and Satterthwaite are concerned with the *strategy-proofness* of a choice function, or the ability of a choice function to discourage insincere behavior by a voter. Suppose that $\rho = (\succeq_1, \dots, \succeq_n)$ is a “true,” or “sincere,” preference profile, whereas $\rho' = (\succeq_1, \dots, \succeq'_i, \dots, \succeq_n)$ is a profile that only differs from ρ in that Voter i reports the “insincere” preferences $\succeq'_i$, as opposed to his true preferences $\succeq_i$. Then a choice function F is *strategy-proof* if $F(\rho')$ will never generate an outcome that Voter i strictly prefers to the outcome generated by $F(\rho)$, the outcome he would have received by sincerely reporting his preferences. Thus, F being strategy-proof implies that no voter can ever strictly benefit by claiming to have preferences that are different than what they actually are. Unfortunately, Gibbard and Satterthwaite demonstrate that strategy-proofness and Pareto efficiency are not attainable with a nondictatorial procedure. Note that the definitions of “dictator” and “Pareto efficiency” used in this theorem are modified slightly from our previous definitions to accommodate the fact that we are considering choice functions.

Theorem 4 [Gibbard, 1973; Satterthwaite 1975]. *With three or more alternatives and unrestricted domain, any choice function that is strategy-proof and Pareto efficient is dictatorial.*

The Gibbard–Satterthwaite theorem proves that the possibility of strategic voting, or voting against one's true preferences, is endemic to the vast majority of voting systems. Despite their differences, the Gibbard–Satterthwaite theorem is mathematically similar to Arrow's theorem, and it has been demonstrated that the two results can be derived from what is essentially a single proof [9]. It should also be noted that Pareto efficiency can be replaced with the far weaker condition of *nonimposition*, which states that any alternative can be achieved as a social choice, given some collection of ballots.

VOTING PARADOXES

Impossibility theorems are paradoxical, in the sense that they prove that seemingly sensible conditions are incompatible with each other. While these results are quite general, in that they typically apply to a wide range of voting systems, it is also well known that specific voting systems can suffer from specific anomalies. A *voting paradox* occurs when a particular voting system yields an end result that is highly counterintuitive. The paradoxes discussed here are well-known problems stemming from widely used voting rules, and represent only a handful of the documented paradoxes.

Condorcet's Paradox

The most famous voting paradox is known as *Condorcet's paradox*, and was introduced by Marquis de Condorcet, a French mathematician and philosopher, in his 1785 treatise *Essay on the Application of Analysis to the Probability of Majority Decision*. This paradox demonstrates that majority rule with three or more alternatives may yield intransitive outcomes. It can be summed up by the following preference profile:

$$\begin{aligned} x &>_1 y >_1 z \\ \rho = y &>_2 z >_2 x . \\ z &>_3 x >_3 y \end{aligned}$$

Given profile ρ , it is clear that a majority vote between x and y would yield the social ranking $x > y$ (by the votes of persons 1 and 3); a majority vote between y and z would yield $y > z$ (by the votes of persons 1 and 2); and that a majority vote between x and z would yield the ranking $z > x$ (by the votes of persons 2 and 3). Thus, majority voting over pairs of alternatives is not guaranteed to yield transitive outcomes when there are three or more alternatives, even when we assume that individual preferences are transitive. In this case, majority voting yields the cycle $x > y$, $y > z$, and $z > x$.

Ostrogorski's Paradox

Ostrogorski's paradox involves the role of political parties as intermediaries in the process of voting. His paradox stems from the fact that the following two procedures can yield different outcomes: (i) Each voter votes for the party whose stand is closest to his own on the most issues. The winner is the party with the most votes. (ii) On a series of issues, each voter votes for the party whose stance is closest to his own. The winner is the party that wins the most issues. The following table demonstrates this paradox [10, p. 72].

<i>Voter</i>	<i>Favorite Party on Issue 1</i>	<i>Favorite Party on Issue 2</i>	<i>Favorite Party on Issue 3</i>	<i>Party Supported</i>
1	X	Y	Y	Y
2	Y	X	Y	Y
3	Y	Y	X	Y
4	X	X	X	X
5	X	X	X	X

The above table shows that a majority of voters prefer Party Y to Party X, overall. At the same time, if we examine the table issue by issue, we can see that Party X defeats Party Y on every issue. Thus, a party winning an election (Y, in this case), could fail to represent the views of a majority on every single issue.

The Additional Support (Monotonicity) Paradox

One straightforward criterion that we might like a voting system to satisfy is that increased support for a winning candidate does not turn that candidate into a loser. It is well known, however, that many two-round voting systems do not satisfy this property, and thus suffer from the additional support, or monotonicity, paradox. The following two tables demonstrate this phenomenon in a majority runoff election, or an election in which, if no candidate receives a majority, the top two vote-getters proceed to a runoff.

<i>Number of Voters</i>	<i>Preferences</i>
6	$a > b > c$
7	$b > a > c$
8	$c > a > b$

In the above election, 11 votes are required out of 21 votes for a candidate to win. Since no candidate receives a majority, a runoff is held between b and c (the two highest vote-getters), and as the a voters prefer b to c , b wins the runoff. Now suppose that three voters switch their preferences from $c > a > b$ to $b > c > a$. This is represented in the following table:

<i>Number of Voters</i>	<i>Preferences</i>
6	$a > b > c$
10	$b > a > c$
5	$c > a > b$

In this case, a and b proceed to the runoff, where a ends up defeating b by 11 to 10. Thus, by receiving strictly greater support in the second election b changes from being a winning candidate to being a losing candidate.

The Alabama Paradox

In the United States, congressional seats are awarded to the 50 states following every decennial census. The Constitution specifies that these seats should be awarded in proportion to state populations, but does not

specify a particular formula by which the seats should be distributed. Prior to 1911, the total number of congressional seats was readjusted every 10 years to account for population growth. It was because of such a readjustment that in 1882 the *Alabama paradox* was discovered: as the total number of seats to be distributed increased, the state of Alabama ended up losing a seat. This paradox is a consequence of the particular apportionment formula being used at the time, called *Hamilton's method*, which worked as follows: First, a quota is determined to represent the ideal number of people to be represented by a single congressional seat. If 435 seats are to be distributed we would get

$$\text{Quota} = \frac{\text{Population of the United States}}{435}.$$

Second, a state's ideal number of seats is determined by dividing the state's population by this quota:

$$\begin{aligned} &\text{States ideal number of seats} \\ &= \frac{\text{Population of the state}}{\text{Quota}}. \end{aligned}$$

In general, however, a state's ideal number of seats (its entitlement) will involve a fractional part. Hamilton's method specifies that each state should be awarded the whole number portion of its entitlement, and that any remaining seats to be distributed go to the states with the largest fractional portions of their entitlements. The following tables demonstrate how Hamilton's method works, and its vulnerability to the Alabama paradox.

<i>Seat Allocations with 10 Seats to Distribute</i>			
<i>State</i>	<i>Population</i>	<i>State's Ideal # of Seats</i>	<i>Seats Awarded</i>
A	1200	4.29	4
B	1200	4.29	4
C	400	1.42	2

In the above table there was one remaining seat to distribute after the states were awarded their integer amounts. This seat

went to State C, because it had the largest fractional entitlement (0.42). The next table shows what happens when the total number of seats to be distributed increases from 10 to 11.

<i>Seat Allocations with 11 Seats to Distribute</i>			
<i>State</i>	<i>Population</i>	<i>State's Ideal # of Seats</i>	<i>Seats Awarded</i>
A	1200	4.71	5
B	1200	4.71	5
C	400	1.57	1

The above table shows that when the number of seats to be distributed is increased State C's fractional entitlement dips below those of States A and B. Thus, even though there are more seats to be awarded, State C ends up losing a seat.

FURTHER READINGS

The field of voting theory is vast and growing, with many topics for the interested reader to discover. Austen-Smith and Banks [11,12] provide an extremely thorough and technical presentation of this material. A less technical introduction to this material with many real-world applications can be found in Saari [13]. Nurmi [10] provides an excellent overview of the various voting paradoxes, and a classification of their types. Balinski and Young [14] provide a history of representation paradoxes that have occurred in the context of United States congressional apportionment. Taylor [15] focuses exclusively on the manipulability of voting systems, and in particular the Gibbard–Satterthwaite theorem and its many extensions. In addition, the interested reader will find the classic texts of Arrow, May, Sen, Gibbard, and Satterthwaite to be readable and thought provoking.

REFERENCES

1. Colomer J. The strategy and history of electoral system choice. In: Colomer J, editor. *Handbook of electoral system choice*. New York: Palgrave Macmillan; 2004. pp. 3–73.

2. Arrow KJ Social choice and individual values. New Haven: Yale University Press; 1963.
3. Arrow KJ. *J Polit Econ* 1950;58:(4):328–346.
4. Wilson R. *J Econ Theory* 1972;5:(3):478–486.
5. May KO. *Econometrica* 1952;20:(4):680–684.
6. Sen AK. *Collective choice and social welfare*. San Francisco: Holden-Day; 1970.
7. Gibbard A. *Econometrica* 1973;41:(4):587–601.
8. Satterthwaite MA. *J Econ Theory* 1975;10:(2):187–217.
9. Reny PJ. *Econ Lett* 2001;70:(1):99–105.
10. Nurmi H *Voting paradoxes and how to deal with them*. Berlin: Springer; 1999.
11. Austen-Smith D, Banks JS. *Positive political theory I: collective preference*. Ann Arbor: University of Michigan Press; 1999.
12. Austen-Smith D, Banks JS. *Positive political theory II: strategy and structure*. Ann Arbor: University of Michigan Press; 2005.
13. Saari D *Decisions and elections: explaining the unexpected*. Cambridge: Cambridge University Press; 2001.
14. Balinski ML, Young HP *Fair representation: meeting the ideal of one man, one vote*. Washington, DC: Brookings Institute Press; 2001.
15. Taylor AD *Social choice and the mathematics of manipulation*. New York: Cambridge University Press; 2005.

IMPROVING PACKAGING OPERATIONS IN THE PLASTICS INDUSTRY

RAJESH TYAGI
SRINIVAS BOLLAPRAGADA
GE Global Research, Niskayuna,
New York

INTRODUCTION

The plastics company in question is a multibillion dollar global materials business that manufactures a variety of plastics and specialty materials such as thermoplastics, silicones, fused quartz, and ceramics. Its customers are in a wide range of industries such as automotive, appliances, optical media, construction, aerospace, transportation, medical, computers, and sports equipment. It includes several divisions whose products differ primarily in their heat resisting and sealing properties. As such, these divisions have their own niche markets. For example, one division makes sealants for consumer and industrial use, while another division supplies plastics that are widely used in the automotive industry. This project pertains to yet another division that makes materials used in high heat applications like electronics, power tools, cellular phones, and appliances. The division manufactures over 160 million lbs of material worth \$800 million annually.

Starting in late 2001, the division began to experience a manufacturing bottleneck at the manufacturing facility. In particular, the packaging operations had become a major bottleneck, impacting the production throughput of the plant. The primary reason was that the product offerings had become very diverse over time, providing little opportunity to aggregate customer orders when creating production lots. Smaller lot sizes imply more frequent cleanups, and packaging stations were especially impacted as they service many extrusion lines (or

extruders) that convert raw plastics called *resins* into finished products in the form of plastic pellets and thus, encountered frequent lot changes. As packaging failed to keep up with production, extrusion lines had to be stopped (or blocked) to prevent overflowing storage bins. Blocked extrusion lines led to the manufacturing schedules, and therefore customer orders, being delayed past their due dates. The blocking also effectively reduced the throughput capacity of the extruder lines. Management found a temporary fix by outsourcing some of the packaging operations—material was loaded onto trucks and railcars, and transported to an outside vendor for packaging at an annual cost of \$250,000. In 2003, as the demand outlook for the products brightened with the improving economy, it was clear that the problem would only get worse unless packaging operations were improved. The management looked at various possible causes and even though they felt that packaging scheduling could be improved, they were not convinced that the improvement alone would eliminate the bottleneck. Faced with an impending increase in production, management decided that increasing packaging capacity along with reconfiguration of packaging operations—assignment of packaging stations to extrusion lines—was the safer solution. To that end, they identified four alternative configurations of linking extruders to packaging stations. As will be explained later, the manufacturing under consideration is a very complex process involving many different types of materials flowing through a number of common bins and where different materials must be kept separate and thus cannot be placed in the same bin at the same time. It soon became apparent the management needed operations research expertise and the authors were brought into the project team to provide that expertise in evaluating the alternatives.

The team's initial charter was to identify the best alternative configuration—one that would minimize blockages of the extrud-

ers while requiring the minimum increase in packaging capacity (packaging stations as well as packagers). At first, team members thought of performing some spreadsheet analysis to estimate the capacities and utilizations of the packaging stations. However, the network of capacitated containers that shared the packaging stations was too complex to estimate the number of cleanups and throughput statistics using spreadsheet capability alone. Hence, the team decided to use simulation technology for the analysis.

Using simulation models, the team determined that the existing capacity was more than sufficient for packaging the output of extruders. The team's analysis showed that the bottleneck was caused not due to insufficient capacity but due to poor scheduling. The team suggested that a scheduling system that more efficiently schedules the packaging operations would significantly reduce the packaging bottleneck. In addition, it would free up some of the packagers' time to focus on other activities. The division then funded a project to develop this packaging scheduling system. The team implemented the system in November 2003, resulting in significant benefits that included increased capacity, improved productivity, and elimination of outsourced packaging cost. These benefits are described in greater detail toward the end.

The rest of the article is organized into the following sections: (i) packaging operations (in which packaging operations are described in detail), (ii) capacity analysis (analysis of the packaging capacity using discrete event simulation) that includes a literature review, (iii) the packaging problem statement, (iv) the packaging scheduling algorithm (a novel heuristic algorithm for scheduling packaging operations), (v) system implementation and use, (vi) benefits and summary, and (vii) references.

PACKAGING OPERATIONS

The packaging system that was simulated, developed, and eventually implemented is part of a plastics manufacturing process. The manufacturing of plastics consists of two separate operations: (i) feedstock conversion and

(ii) compounding. First, feedstocks are converted into a number of core resins, which are stored in silos. Then, as customer orders are released for production, these resins are combined using an appropriate bill of materials and compounded to produce the final product. This end product is then packaged and shipped to the customer. Because products are offered in a wide variety of colors and grades, most customer orders end up being produced as separate lots.

The feedstock conversion operation is performed by resin plants that convert the primary raw materials of plastics, crude oil, and natural gas, into a number of different resins. Each resin has its own unique characteristics such as hardness, heat-resistance, and corrosion. The resins are stored in silos prior to their conversion to finished products by the compounding operation.

The compounding operation, which is the focus of this study, is ordinarily performed by 13 extruders, which extrude the finished product as pellets. After extrusion, the product is placed in containers called *v-bins* where quality tests are conducted on samples (Fig. 1). If samples are satisfactory, the finished product is moved into another type of container called a *hopper*, where it waits to be packaged. An extrusion line generally has two or three *v-bins* and one or two hoppers. Once in a hopper, the product must be packaged in either 55-lb bags or 1100-lb boxes, depending on customer order requirement. The division utilizes two bag-packaging stations and two box-packaging stations. Thus, the 13 extruders with a total of 22 hoppers must be packaged using these four packaging stations (Fig. 2). The existing configuration has six lines assigned to one pair of packaging stations (Bag 1 and Box 1), while the remaining seven lines utilize the other pair of packaging stations (Bag 2 and Box 2). A packager must be in attendance at each station during the packaging operation to help with sealing bags or boxes. At the time of the study, one packager attended to both Bag 1 and Box 1, while Bag 2 and Box 2 had dedicated packagers assigned to them.

As stated earlier, the packaging operations were unable to keep pace with the output from extruders. Consequently, as

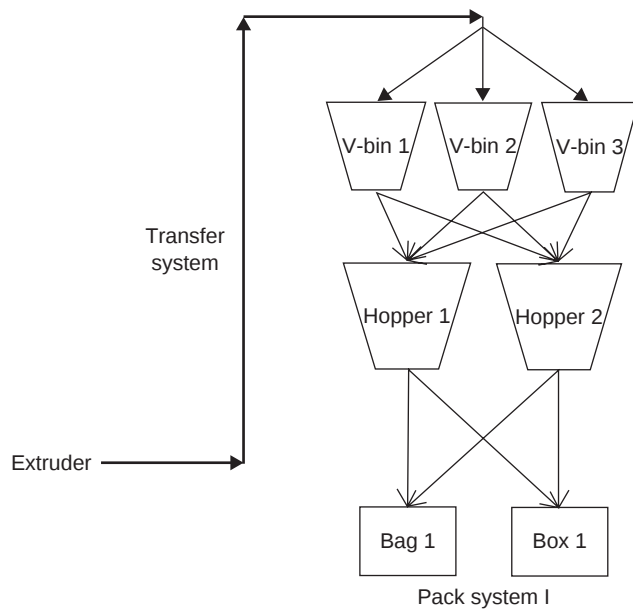


Figure 1. After extrusion, the material goes to v-bins for quality tests before being placed in hoppers where they await packaging on a bag- or box-packaging station depending on customer requirement. Each extruder has 1 to 3 v-bins and 1 or 2 hoppers.

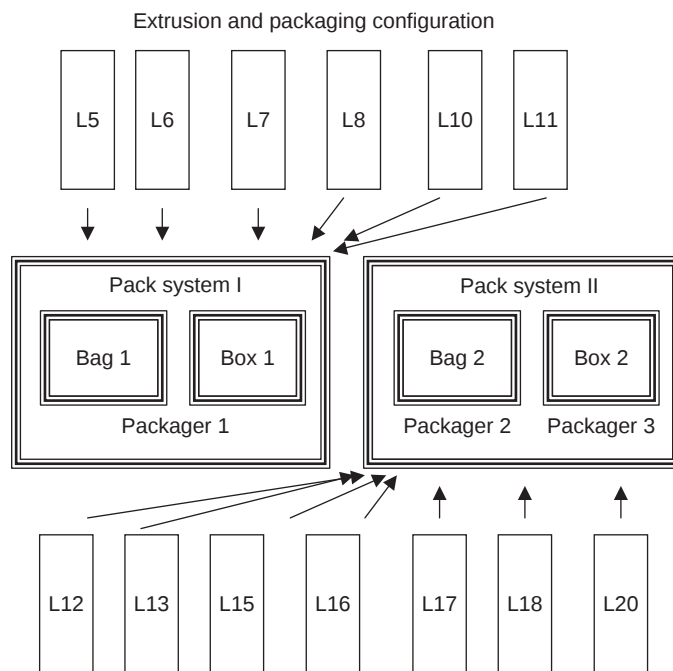


Figure 2. The 13 extruding lines are served by two pack systems, consisting of a bag- and a box-packaging station each. Six of the extruders connect to Pack system I, while the remaining seven connect to Pack system II. For clarity, the v-bins and hoppers between each extruder and the packaging system are not displayed.

material in hoppers and v-bins awaited packaging, the extruders had to be frequently stopped to prevent overflowing of these containers or mixing of different materials in the same containers. As a general

rule, if an extruder is blocked for more than one hour, it must be completely purged; that is, its contents must be discarded to avoid contamination. When extrusion lines are blocked, subsequent lots on the extruders

cannot be processed as scheduled, resulting in delays in fulfilling customer orders. In effect, they severely curtail the throughput of the plant. To alleviate these problems, the business had to outsource some of the packaging operations; the material was transported to a vendor, who packaged them in bags or boxes as necessary at a total cost of over \$250,000 per year. Management felt that the packaging bottleneck was due to insufficient packaging capacity. They created teams which identified four alternative packaging configurations that they believed would alleviate the packaging bottleneck. "Alternative 1" envisioned a separate pack system for lines L12 and L13 (Fig. 2); "Alternative 2" required a separate bagging system for lines L5 and L6 with L12 and L13 rerouted to Pack 1; "Alternative 3" would utilize connector lines such that lines utilizing Pack 1 could be routed to Pack 2 if the former was in use and the latter was available, and vice versa; and "Alternative 4" was to replace the two existing pack systems with a single new high speed high capacity bag pack for all lines. The management then requested the team to analyze the system to recommend the best alternative to implement. The team used a simulation model to analyze the packaging operations, as described in the next section.

CAPACITY ANALYSIS

To determine whether capacity was a cause of the problem of packaging operations' inability to keep up with production, the adequacy of the existing packaging capacity, and production scheduling practices were evaluated. Below, the authors discuss the manufacturing environment, production scheduling, and explain their discoveries.

The manufacturing environment in the plastics industry is characterized by multiple production (or extrusion) lines that are collocated in the same physical plant. While some of these lines may be dedicated to producing a single product, others can produce multiple products. The production schedules are determined considering the product demand, setup costs, inventory costs, and shortage cost.

Many papers have been published dealing with production scheduling in this manufacturing environment. In the case of a single production line, various aspects of scheduling have been investigated by Hsu [1], Gallego and Shaw [2], Jones and Inman [3], Gallego *et al.* [4], Dobson [5], Zipkin [6], Anderson [7], and Gallego and Moon [8]. Multiple production lines have been studied by Maxwell and Singh [9], Carreno [10], and Bollapragada and Rao [11]. Much of the literature deals with developing cyclical schedules under relatively constant product demands.

The company in which the packaging system that was simulated, developed, and eventually implemented has used a multistep proprietary system that includes the following steps. First, the corporate office looks at the book of business and determines which production orders must be fulfilled in the upcoming week. This is done while taking into consideration the aggregate production capacity and order delivery dates. The released orders are then forwarded to the manufacturing, which aggregates them into lots and develops production orders. It then assigns the production orders to individual extruders based on extruders' production capacities and the products that they can manufacture. Thus, at any point in time, the manufacturing facility has a rolling seven-day production schedule for the extruders.

Neither the literature nor the company's own legacy scheduling system, however, had addressed packaging operations. Plants generally manage packaging operations two ways. When production is stable with few cleanups between lots, sufficient packaging capacity capable of coping with the production can usually be installed. On the other hand, plants with unstable schedules often have to outsource part of the packaging operations to keep the extrusion lines operating. The latter case characterized the present operations as well. The division's book of business had become highly customized to the extent that production was characterized by large number of small lot sizes, requiring frequent cleanups of packaging containers and stations.

To effectively simulate the packaging system and evaluate its capacity, the team

started by base lining the existing system. Base lining refers to creating benchmark performance measures for the current system against which alternatives can be evaluated. This was done by developing a discrete-event simulation model of the existing plant configuration using ProModel simulation software. The model captured the material production and flows through extruders, v-bins, hoppers, and packaging stations as discussed earlier (Figs 1 and 2). Whenever a v-bin, hopper, or a packaging station processed a different lot, we added cleanup times requiring 15 min of packager's time to wash the station followed by 30 min of unattended drying. Other model features were v-bin and hopper capacities, extrusion rates, packager shifts, packager breaks and safety meetings, and packaging station breakdowns. We conducted time studies to determine packaging times to pack a bag or a box. When a packaging facility finished processing a lot, the next hopper was selected for processing by following the rule in force at that time: identify all hoppers that are at least half full, and then select the one that contains the lot with the earliest due date.

From historical data, a production week that had a normal production schedule was selected. All extrusion schedules were entered for that week into the simulation model to develop the baseline performance metrics. The team then successively modified the baseline model to implement each of the four alternative configurations and recorded the resulting performance metrics. The entire process was repeated with a light weekly schedule and a heavy weekly schedule.

Thus, the performance of the current system in each of the four alternatives was simulated under similar demand conditions. All configurations were evaluated on packaging cycle times, packaging stations utilization, packager utilization, average lateness, and number of line blockings. Analysis revealed that none of the alternatives fared significantly better than the existing system. But it was surprising to discover that the baseline system should be causing any bottlenecks

to begin with since the utilization of packagers and packaging stations was less than 30 percent.

The team concluded that the existing packaging capacity was more than sufficient to handle the output from extruders. Management was at first surprised by the team's conclusions, but eventually accepted our findings once we explained the actual simulation runs. It was obvious that actual packaging operations were not running as effectively as could have been expected. Further investigation as to why the packagers were unable to follow the prescribed rules revealed that the team found it very difficult to select the right hopper to package from, at the right time, from the more than 40 hoppers that are in the manufacturing facility. Consequently, packagers made up their own rules and would often process only the box stations, which were less physically demanding than the bag stations. The hoppers with material waiting to be bagged were being left unattended, leading to line blocking. Therefore, when presenting our analysis to the managers, the team proposed that a scheduling system be developed to generate efficient packaging schedules that could be easily followed by the packagers. The scheduling system would create this schedule so as to minimize the blocking of extruders. A formal statement of the packaging system formulation is presented next.

THE PACKAGING PROBLEM STATEMENT

The packaging problem that was used for the simulation of the packaging system can be stated formally as follows

Minimize Total number of extruder blockages

Subject to constraints on:

- lot processing times.
- lot sequencing.
- packager to facility assignments.
- setup times.
- V-bins and hopper contents and connectivity.

The objective is to minimize the total number of times that extruders are blocked over the time horizon for which schedules are generated. There are five sets of constraints: (i) lot-processing constraints incorporate scheduled start times and extrusion times once lot processing begins; (ii) lot-sequencing constraints should ensure that all bags or boxes of a lot on an extruder are completely packaged before packaging of the next lot on that extruder begins; (iii) packager-facility assignment constraints are necessary because Bag 1 and Box 1 share the same packager and thus, can never be used simultaneously; (iv) setup times constraints require packaging stations be cleaned after a lot change—this entails the packager washing the facility for 15 min and then letting the facility dry for 30 min. And finally, there are constraints that model the connectivity between each extruder's v-bins and hoppers and their contents. These constraints should make sure material flow from v-bins to hopper is restricted by the physical connectivity limitations and that a container never contains material from two different lots at the same time.

The team formulated the above problem as a mathematical model. This resulted in a large mixed-integer programming model that could not be solved in a reasonable amount of time using existing commercial math programming software. We therefore developed a heuristic scheduling algorithm, which is described in the next section.

THE PACKAGING SCHEDULING ALGORITHM

We use a discrete-time simulation-based heuristic scheduling algorithm to develop packaging schedules for a rolling horizon of one week. We divide the time horizon of one week into consecutive time periods of 15-min durations. The state of the system is maintained by each time period for the entire time horizon. The state of the system in a given time period include the current contents of the hoppers and v-bins, the jobs being processed by the extruders and the packaging stations, and the availability of the packagers. Our algorithm schedules

the packaging stations for the entire week, one period at a time starting from the first period. The underlying philosophy behind the algorithm is to package the output of the extruder that will be blocked at the earliest if no packaging stations were available. Then as some of the material is packaged from the extruder, the v-bins and hoppers start to empty out because the packaging rate is much higher than the extrusion rate. Even before the lot is completely packaged, the hoppers may be sufficiently empty so that this extruder will no longer be the earliest to be blocked. At this time, the extruder that now has the earliest blocking time is used. To avoid incurring frequent set-ups on the packaging equipment, it is ensured that a facility packages for a minimum amount of time before it switches to a new lot that requires a set up. This philosophy is implemented as explained below.

At the start of each time period, the blockage time of each extruder is computed. Blockage time is defined as the time before the extruder is blocked under the assumption that none of the materials in its v-bins and hoppers is packaged, and that the line continues to extrude its scheduled lots. The blockage time is computed using the jobs scheduled on the extruder, the production rate of the extruder, and the total available capacity in the v-bins and hoppers for storing the extruded material. The available storage capacity does not include unfilled storage space in a hopper or v-bin if it contains a different lot because different grades of plastics cannot be mixed. This situation may occur because it is possible for multiple lots to be awaiting packaging at an extruder. For example, the extruder may extrude a complete lot and place it in one hopper if the hopper is sufficiently large, and then proceed to place another extruded lot in the second hopper. Even the v-bins can be used for storage to prevent line blocking.

After computing the blockage times of all extruders, the extruders are sorted in the increasing order of blockage time. Starting from the first extruder in the sorted list, we attempt to schedule packaging for each extruder in the sorted list. To schedule packaging for a given extruder, we determine if

both the packaging facility needed to service that extruder and the packager assigned to that facility are available. If so, packaging for that extruder is scheduled. If this is a different lot for the packaging facility, a 45-min cleanup on the facility is scheduled. The next extruder in the list sorted by blockage times is then processed similarly. After all extruders have been processed for this time period, the state of the system for the rest

of the time horizon is updated. The updating included the quantities in the v-bins and hoppers based on the material that would be extruded and packaged during that time period, and the availability of the stations and packagers in the future time periods. The process for the next time period is then repeated, and continued until the end of the planning horizon. Our packaging algorithm can be summarized as follows.

Procedure Packaging_Algorithm

```

For period  $t = 1$  To NUM_PERIODS
    Compute extruder blockage times
    Sort extruders in the increasing order of blockage time
    For each extruder in the sorted list (starting from the top of the list)
        Check if the packaging facility needed by the extruder and the packager who
        operates that facility are available.
        If so, schedule packaging on the extruder for one time period. If this is a
        different lot for the packaging facility, schedule cleanup on that facility
        lasting 3 periods (45 min)
    Loop (next extruder)
    Update the system state for the rest of the time horizon.
Loop (next period  $t$ )
End procedure

```

SYSTEM IMPLEMENTATION AND USE

We implemented the scheduling algorithm in a web-based system, which we designed using Microsoft technologies: ASP, Visual Basic, VBScript, and MS-Access. The scheduling algorithm was coded in Visual Basic. The packaging schedules generated by the algorithm are saved in the packaging database that is implemented in Microsoft Access. A web interface using Visual Interdev was designed to extract the packaging schedules from the database and present to the user in a HTML format. The web server and the database reside on a Windows NT server. We provided a rich user interface with several options to view the schedules by date, shift, packaging facility, and packager. The packaging system interfaces with the smart tags database to retrieve information on current contents of v-bins and hoppers. These smart tags were installed to support the packaging system. We retrieved information on lots scheduled for production on each extruder line from the database supporting the master scheduling

system. Figure 3 shows the architecture of the system.

Shortly before the start of each shift, the shift leader ensures that the data in the system is accurate. The shift leader updates the database if any packaging facility is unavailable due to scheduled maintenance or breakdown, or if a packager is not available for any reason. He or she then issues the command to generate the packaging schedule for the next one week. The system now obtains real-time data on the current lot quantity in each v-bin and hopper, quantities that have already been packaged in bags and boxes, and the status of each packaging facility and packager. The extruder schedules are then obtained from the extruder schedules database.

The system then executes the scheduling algorithm and stores the results in the packaging database. The two most important outputs of the algorithm are packaging schedules for each packaging station, and the packager schedules for each packager. The packaging schedule (Fig. 4) shows the different extrusion lines and lots the facility would be packaging. Cleanup operations between

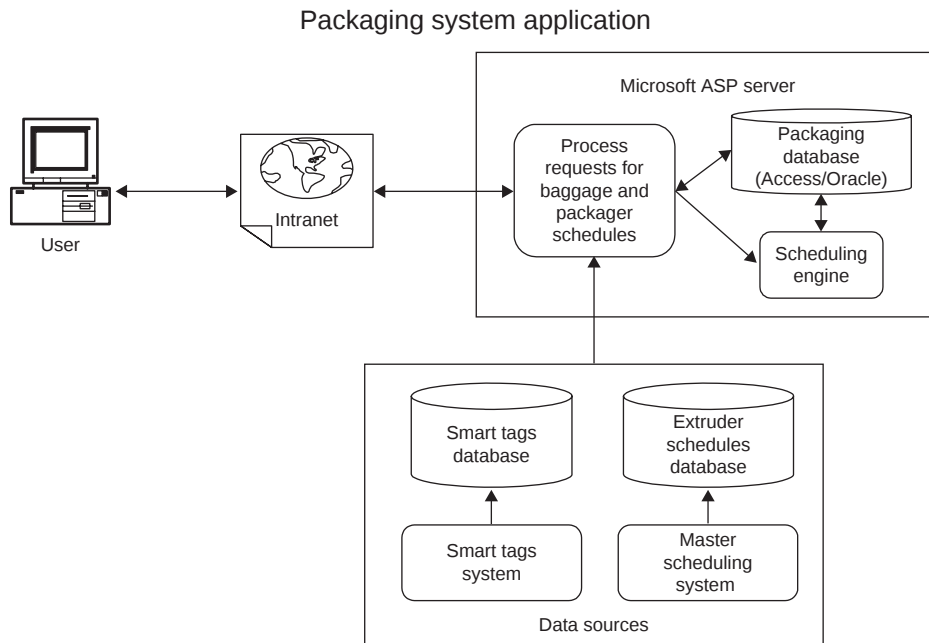


Figure 3. The packaging system resides on a Microsoft ASP server. It retrieves information on current contents of v-bins and hoppers from the smart tag application database and the extruder line schedules from the extruder schedules database. The scheduling engine written in Visual Basic generates schedules for packaging lines and packagers and saves them into the packaging database.

lot changes are also included. An important benefit of this report is that it shows times when the facility is idle so that maintenance operations can be performed if necessary. The packager schedules (Fig. 5) are created for each packager and show the packaging stations and corresponding lots the packager must attend to during the shift. Packagers' scheduled lunch and coffee breaks, and mandatory meetings are also built into the report. Idle times are also displayed which are used by packagers to plan nonpackaging related activities.

The packaging schedule is regenerated with fresh data at the beginning of each shift for a one-week rolling time horizon. Clearly, only the schedule for the shift is of interest to the packagers; however, the week-long schedule gives management visibility to upcoming load at the packaging stations. There are occasions when a new schedule needs to be developed in-between a shift if

some events cause the actual packaging operations to deviate from the proposed schedules. For example, a packaging facility may break down causing some of the extruders feeding into it to be blocked. Or, the business may rush through an important last minute order, which would also throw the actual operations out of kilter with the schedule. When these events occur, the personnel on the plant floor deal with the immediate situation as best as they can and then the shift lead regenerates the schedule from that time onwards.

BENEFITS AND SUMMARY

After an initial two-month phase-in period, during which packaging schedules were thoroughly reviewed for completeness and efficiency, and system bugs identified and fixed, the packaging system was placed into production in January, 2004, just as the demand for

The screenshot displays the 'Bagging Schedule' window. At the top, a menu bar includes Home, Line Schedule, Line Status, Bagging Facility Status, Bagging Facility Schedule, Packager Schedule, Bagging Facility Downtime, Packager Breaks, Packager Unavailability, Packaging Status Sheet, Admin, and Help. Below the menu, a date selector is set to '5/24/04' and a shift selector is set to 'Days'. The main content area is titled 'Day: 5/24/04 Shift: DAYS' and contains three tables for different facilities and packagers.

StartTime	EndTime	Line	Lot	Product
8:00:00 AM	8:15:00 AM	2	Under Cleaning	
8:15:00 AM	9:15:00 AM	0	Idle	
9:15:00 AM	9:45:00 AM	5	CV2WXD	780-BK1066
9:45:00 AM	10:00:00 AM	0	Idle	
10:00:00 AM	11:30:00 AM	5	CV2WXD	780-BK1066
11:30:00 AM	12:00:00 PM	0	Idle	
12:00:00 PM	12:45:00 PM	2	Under Cleaning	
12:45:00 PM	1:45:00 PM	6	CV2XBR	771-1001
1:45:00 PM	4:00:00 PM	0	Idle	

StartTime	EndTime	Line	Lot	Product
8:00:00 AM	8:30:00 AM	0	Idle	
8:30:00 AM	9:15:00 AM	7	CV2XGN	4521-8244
9:15:00 AM	12:15:00 PM	0	Idle	
12:15:00 PM	12:45:00 PM	7	CV2XGN	4521-8244
12:45:00 PM	1:45:00 PM	0	Idle	
1:45:00 PM	2:15:00 PM	7	CV2XGN	4521-8244
2:15:00 PM	2:30:00 PM	0	Idle	
2:30:00 PM	3:00:00 PM	7	CV2XGN	4521-8244
3:00:00 PM	3:15:00 PM	0	Idle	
3:15:00 PM	4:00:00 PM	7	CV2XGN	4521-8244
4:00:00 PM	4:00:00 PM	0	Idle	

StartTime	EndTime	Line	Lot	Product
8:00:00 AM	8:30:00 AM	0	Idle	
8:30:00 AM	9:45:00 AM	16	CV2X93	310-BK1066
9:45:00 AM	10:00:00 AM	0	Idle	
10:00:00 AM	11:30:00 AM	16	CV2X93	310-BK1066

StartTime	EndTime	Line	Lot	Product
8:00:00 AM	8:15:00 AM	2	Under Cleaning	
8:15:00 AM	8:30:00 AM	0	Idle	
8:30:00 AM	9:45:00 AM	18	CV2XGG	508-BK1066
9:45:00 AM	10:00:00 AM	0	Idle	

Figure 4. The packaging facility schedule, created for each facility, shows which extrusion line will be packaged by facility. It also shows the times when the facility is under cleaning and idle.

The screenshot displays the 'Packager Schedule' window. At the top, a menu bar includes Home, Line Schedule, Line Status, Bagging Facility Status, Bagging Facility Schedule, Packager Schedule, Bagging Facility Downtime, Packager Breaks, Packager Unavailability, Packaging Status Sheet, Admin, and Help. Below the menu, a date selector is set to '5/24/04' and a shift selector is set to 'Days'. The main content area is titled 'Day: 5/24/04 Shift: DAYS' and contains three tables for different packagers.

StartTime	EndTime	Facility	Line	Lot
8:00:00 AM	8:30:00 AM	Busy/Break		
8:30:00 AM	9:00:00 AM	Box_1	7	CV2XGN
9:15:00 AM	9:30:00 AM	Bag_1	5	CV2WXD
9:45:00 AM	10:00:00 AM	Busy/Break		
10:00:00 AM	11:15:00 AM	Bag_1	5	CV2WXD
11:30:00 AM	12:00:00 PM	Busy/Break		
12:00:00 PM	12:15:00 PM	Bag_1	2	Under Cleaning
12:15:00 PM	12:30:00 PM	Box_1	7	CV2XGN
12:45:00 PM	1:30:00 PM	Bag_1	6	CV2XBR
1:45:00 PM	2:00:00 PM	Box_1	7	CV2XGN
2:15:00 PM	2:30:00 PM	Busy/Break		
2:30:00 PM	2:45:00 PM	Box_1	7	CV2XGN
3:00:00 PM	3:15:00 PM	Busy/Break		
3:15:00 PM	3:45:00 PM	Box_1	7	CV2XGN

StartTime	EndTime	Facility	Line	Lot
8:00:00 AM	8:30:00 AM	Busy/Break		
8:30:00 AM	9:30:00 AM	Bag_2	16	CV2X93
9:45:00 AM	10:00:00 AM	Busy/Break		
10:00:00 AM	11:15:00 AM	Bag_2	16	CV2X93
11:30:00 AM	12:00:00 PM	Busy/Break		
12:00:00 PM	2:00:00 PM	Bag_2	16	CV2X93
2:15:00 PM	2:30:00 PM	Busy/Break		
2:30:00 PM	2:45:00 PM	Bag_2	16	CV2X93
3:00:00 PM	3:15:00 PM	Busy/Break		
3:15:00 PM	3:45:00 PM	Bag_2	16	CV2X93

StartTime	EndTime	Facility	Line	Lot
8:00:00 AM	8:30:00 AM	Busy/Break		
8:30:00 AM	9:30:00 AM	Box_2	18	CV2XGG

Figure 5. For each packager, a schedule is generated showing which stations the packager should be attending during the shift. It also displays the times when the packager is on a break or busy in mandatory meetings.

the division's products began to pick up with improving economy. The project resulted in the following benefits:

- *Reduced Capital Investment Costs.* The simulation analysis clearly showed that existing packaging capacity and the number of packagers utilized were sufficient for a capacity of up to 3.5 million lbs per week. This prevented an unnecessary capacity expansion project that would have required over a million dollars in capital expenditure.
- *Increased Capacity.* After the system had been in use for a while, often with increased loads, the management concluded that by removing packaging as bottleneck, the throughput capacity of the plant has been effectively increased by 20%, or about 30 million lbs per year.
- *Increased Efficiencies.* Packagers now get their schedule at the beginning of each shift showing where they need to be at what time for the duration of the shift. Since the packagers have visibility to their schedules, they can better utilize their "idle time" for other tasks.
- *Personnel Savings.* The manufacturing operations are performed 24 hours a day, 7 days a week, in three daily shifts. To accommodate a five-day workweek, four teams of packagers rotate through the shifts during a week. Improved scheduling resulted in one full time equivalent (FTE) worth of personnel idle time per shift. Packagers now use this idle time to load the material into trailers themselves, a task previously done by the warehouse personnel. This has allowed for a total reduction of four FTEs from the warehouse staff of 16 personnel that formed the four shift teams, or 25% of the personnel cost.
- *Outsourcing Costs.* Outsourcing of packaging was reduced by 50% (some orders still require specialized packaging), saving another \$125,000 per year.
- *Maintenance Planning.* The system generates a rolling seven-day schedule at the beginning of each shift. This gives sufficient advance visibility to idle times

at packaging stations so that plant personnel can perform maintenance operations without affecting plant throughput.

In this article, we have described a scheduling problem encountered in the plastics industry. The bottleneck in the packaging operations was affecting the efficiency of the extruding operations. After an initial simulation-based capacity analysis that showed that packaging stations had sufficient capacity, we turned our attention to improving scheduling at those stations. We developed and implemented a packaging scheduling system that creates a schedule for each packaging station and for each packager. By removing the packaging operations as a bottleneck, the system effectively increased the throughput capacity of the manufacturing operations and also eliminated the need for outsourcing of packaging operations. The main lesson we learnt was that when dealing with complex manufacturing systems, the obvious solution might not be the correct solution. In this project, the obvious solution of increasing the packaging capacity would not have achieved the desired results, as we learnt from the capacity analysis that was performed via simulation. This project clearly illustrated the need to involve operations research experts in analyzing complex manufacturing systems.

Acknowledgments

This work was undertaken for SABIC Innovative Plastics, a business unit of Saudi Basic Industries Corporation. The authors also wish to thank Diderico van Eyl of SABIC Innovative Plastics for reviewing this manuscript and making suggestions to improve its quality and readability.

REFERENCES

1. Hsu W. On the general feasibility test of scheduling lot sizes for several products on one machine. *Manage Sci* 1983;29:93–105.
2. Gallego G, Shaw DX. Complexity of the ELSP with general cyclic schedules. *IIE Trans* 1997;29:109–113.

3. Jones PC, Inman RR. When is the economic lot scheduling problem easy? *IIE Trans* 1989;21:11–20.
4. Gallego G, Queyranne M, Simchi-Levi D. Single resource multiitem inventory systems. *Oper Res* 1996;44:580–595.
5. Dobson G. The economic lot-scheduling problem: achieving feasibility using time-varying lot sizes. *Oper Res* 1987;35:764–771.
6. Zipkin P. Computing optimal lot sizes in the economic lot scheduling problem. *Oper Res* 1991;39:56–63.
7. Anderson E. Testing feasibility in the lot scheduling problem. *Oper Res* 1990;38:1079–1088.
8. Gallego G, Moon I. The effect of externalizing setups in the economic lot scheduling problem. *Oper Res* 1992;40:614–619.
9. Maxwell WL, Singh H. Scheduling cyclic production on several identical machines. *Oper Res* 1986;34:460–463.
10. Carreno JJ. Economic lot scheduling for multiple products on parallel identical processors. *Manage Sci* 1990;36:348–358.
11. Bollapragada R, Rao U. Single-stage resource allocation and economic lot scheduling on multiple, nonidentical production lines. *Manage Sci* 1999;45:889–904.

IMPROVING PUBLIC HEALTH IN DEVELOPING COUNTRIES THROUGH OPERATIONS RESEARCH

PRASHANT YADAV
MIT-Zaragoza International
Logistics Program, Zaragoza
Logistics Center, Zaragoza,
Spain

Public health involves improving the health of a community (as opposed to an individual) by preventive measures, control and treatment of communicable diseases, health education, regular monitoring, and surveillance of diseases. Activities in ensuring public health can be generally classified as “public goods.” The benefits from public health activities are of a community nature and its mode of delivery does not always allow targeting and exclusion. It typically costs very little for an additional individual to enjoy the benefits accruing from a public health intervention and it is generally difficult or impossible to exclude individuals from consuming the benefits. A fundamental problem with public health (and more generally all public goods) is the inability of all segments of the population to pay for health and in some cases, the difficulty of motivating individuals to pay because a portion of the resulting benefits from public health expenditures are benefit externalities for them. Poverty and more competing needs for expenditure severely limit the ability of large segments of the population in the developing world to pay for better health. Public health is therefore mostly financed by national governments with the objective of controlling the spread of disease and ensuring better health of their citizens or by supranational agencies to ensure better global health. The resources committed to public health are generated from tax and other governmental revenues and thus lend themselves as ideal candidates for using operations research (OR) techniques in resource allocation and deployment. In the last century, we have seen remarkable progress in

the field of public health in large parts of the world, a part of which can be attributed to the use of scientific methods to allocate and deploy public health interventions.

Unfortunately, the spread of infectious diseases such as HIV/AIDS and the resurgence of malaria combined with severe resource constraints have resulted in little progress on this front in low income countries, especially in sub-Saharan Africa. As an example, the average life expectancy is merely 39.6 years in Swaziland and 42.1 years in Zambia as compared to 82.6 years in Japan or 78.2 years in the United States (the world average is 67 years) [1]. The public health situation in low income countries today corresponds in many respects to the situation observed in the rich OECD countries a few decades ago. However, the technical and scientific knowledge for overcoming some of the diseases such as diphtheria, diarrhea, typhoid fever, and malaria already exists now. In addition, reasonable (although not sufficient) financial resources have been committed by institutions such as the Global Fund to Fight AIDS, TB, and Malaria and the US President’s Emergency Program for AIDS Relief to counter these diseases. Unlike in the OECD countries where health care is financed either through government or individual private expenditure, financial resources for improvement of health in developing countries come from a variety of sources including bilateral government aid, multilateral aid institutions, private philanthropic organizations, other non-governmental organizations (NGOs), and government loan readjustments with the World Bank (Fig. 1).

This variety in the sources of financing further exacerbates the complexities in resource allocation because not-for-profit, government, and for-profit stakeholders very often have different objectives. Given the overall scarcity of resources and a complex multistakeholder environment, the foremost need in developing countries is now to apply OR techniques to ensure the optimal

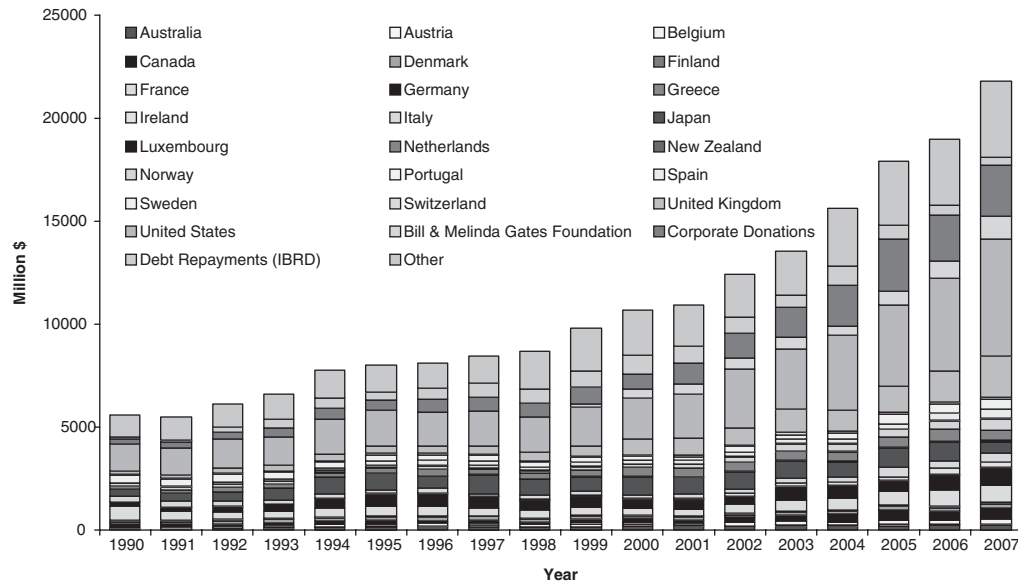


Figure 1. Sources of financing for global health. [Source: Source data obtained from the Institute for Health Metrics and Evaluation. <http://www.healthmetricsandevaluation.org/resources/datasets/2009/dah.html>]

allocation, utilization, and deployment of the available economic and material resources. OR can be applied to optimize efficiency of different public health programs in countries with high disease burdens and scarce resources. As an illustration, the budget allocated to OR in grants allocated by the Global Fund increased from a total of US \$42 million for all grants between 2001 and 2005 to US \$27.3 million for grants in 2006 alone [2].

OR techniques have traditionally been used in the field of health care for a range of problems such as hospital capacity planning, health-care facility location, outcomes and pharmaco-economic analysis [3]. This article will provide an introductory overview for an OR practitioner/researcher considering applications of OR which have specific relevance to public health in the developing world. The purpose is not to provide a comprehensive survey nor a taxonomy of OR problems in public health in developing countries but instead to provide the reader an overview of the nature of problems in global public health that are suitable for the use of OR techniques.

OPERATIONS RESEARCH IN PUBLIC HEALTH: DEFINITIONAL ISSUES

A review of literature and technical reports in public health reveals that the definition of public health and even more so the definition of OR varies significantly amongst public health practitioners and researchers. A quick review of the existing definitions is presented before we delve further into the topic.

Winslow [4] first defined public health as “the science and art of preventing disease, prolonging life, and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals.” In a more recent definition, the Institute of Medicine [5] defines public health as “fulfilling society’s interest in assuring conditions in which people can be healthy by applying scientific and technical knowledge to prevent disease and promote health.” For a more detailed discussion on activities that are considered as public health, see Ref. 6.

Similarly, a review of the OR activity description of the Global Fund, the WHO

and various other agencies involved in public health revealed that a commonly accepted definition is “the use of systematic research techniques to provide policy makers and program managers with evidence that they can use to improve program performance.” The definition in public health practice is in tune with the early definition by Morse and Kimball [7] and Waddington [8], which includes activities such as impact assessment, outcome measurement, and metric design as a key part of OR. In this article, we work with a more restrictive definition, that is, “the use of mathematical models, statistics and algorithms to aid in decision making with the goal of optimizing performance on specific metrics.”

KEY METRICS IN PUBLIC HEALTH

Although the core techniques of OR do not change much upon their application in public health, the nature of the objective function and output metrics are considerably different in many cases. As public health deals with the overall state of health of the population, almost all analysis hinges on being able to measure the state of health. Public health operational research also focuses on the efficiency of an intervention through the use of cost-effective measures. The costs and public health effects of an intervention are assessed to determine whether the public health intervention is worthwhile from an economic perspective. When resources are inadequate to meet all possible health needs—as they almost always are in the case of developing countries—being able to quantify the outcomes using a common health metric aids the efficiency of resource allocation. Several different measures have thus been proposed and we present the most commonly used ones. Some of these metrics are also widely used in health care OR.

Mortality

Mortality rate is a key measure of the health of a population and captures the number of individuals who die each year by age, sex, and the cause of death classified according to standard medical criterion. Most countries

collect such data but the quality, coverage, and completeness varies significantly. A 2003 WHO study [9] found that there are 28 countries where less than 70% of the mortality data are complete or an incorrect cause of death was assigned to more than 20% of deaths.

Morbidity

The number of cases of a particular disease occurring in a year is usually reported as cases per 1000. When looking at morbidity it is important to differentiate between *prevalence* and *incidence*.

Incidence describes the occurrence of new disease in the population in a given time period, whereas *prevalence* is a static measure of the proportion of a population that has the given disease, whether the disease cases occurred recently or at some previous point in time. For infectious diseases of short duration such as malaria or diarrhea, the differences are not significant between the two, but for more chronic communicable diseases such as HIV/AIDS and sexually transmitted infections, there are significant differences between prevalence and incidence.

Life Expectancy

Life expectancy measures the average remaining life span of individuals in a given population group. A commonly used measure is life expectancy at birth.

The above metrics only measure the impact of disease or ill-health as either death or prevalence/incidence. The longer term impacts of disease such as disability and loss of quality of life have a significant cost to the society and an overall public health impact but they are not considered in the three metrics presented earlier. Quality-adjusted life year (QALY) and Disability-adjusted life year (DALY) are two economic measures of health that combine the duration and quality of life.

Quality-Adjusted Life Year (QALY)

QALY is a metric which measures the duration and quality of life [10]. The number of QALYs lived by an individual in one year

is equal to the health-related quality of life weight attached to that year of the individual's life. An individual's quality-adjusted life expectancy at age a can be written as $\sum_{t=a}^{a+L} Q_t$ where L is the remaining life expectancy at a . However, the value of life in a given year is not the same as life in future years and hence a discount factor δ is used. Typically a 3% discount factor is used [11].

Thus, QALYs gained as a result of a specific intervention can be written as $\sum_{t=a}^{a+L^i} \frac{Q_t^i}{(1+\delta)^t} - \sum_{t=a}^{a+L} \frac{Q_t}{(1+\delta)^t}$, where Q_t^i and Q_t represent the quality of life weight with and without the intervention respectively for each time period t and L^i is the remaining life expectancy after the intervention. It is important to note that QALYs do not have an age-weighting function, that is, apart from discounting one QALY has the same value regardless of the age at which it is lived.

Disability-Adjusted Life Year (DALY)

A DALY is a health outcome measure which like the QALY measures both the quality of life reduced due to a disability after the disease and the lifetime lost due to premature mortality. One DALY can be viewed as one lost year of "healthy" life. The concept of DALY was first presented in World Bank [12]. $DALY = YLL + YLD$, where YLL are the years of life lost due to premature mortality (average life expectancy at age of death due to disease) and YLD are the years lost due to disability. YLD are calculated by multiplying the number of cases of the disease by the average duration of a case and a specifically chosen disability weight for each health condition. The disability weight is determined by expert valuations on a scale from 0 (perfect health) to 1 (death).

A key difference between the DALY and the QALY is that in DALY calculations the value attached to life is weighted so that years of life in childhood and old age are counted less [13]. Also, DALY involves continuous discounting, whereas QALY is based on discrete discounting. Ignoring continuous discounting, age weighting and the way the disability/quality scores are computed, one DALY saved compares to one QALY gained, that is, $QALY = 1 - DALY$.

There has been considerable critique of the DALY (and QALY) as a metric for resource allocation decisions in public health [13,14]. One of the main critiques is its use of different age weights in order to capture lesser economic value creation as a child or as an older person. However, such weighting may work against resource allocations that alleviate diseases prevalent only in children in favor of those diseases that are more prevalent in working adults. Additionally, public health interventions often result in nonhealth related economic benefits and using DALY minimization as the objective in resource allocation may ignore some of these interrelationships [14]. Also, when used for resource allocation, DALY discriminates against those who are already disabled as their DALYs are lower.

Public health OR modelers require metrics so that health outcomes can be expressed in common units to analyze the trade-offs between health and economic benefits. Measuring health outcomes in common units is such a complex process that some of these shortcomings are likely to be present in any economic metric created for this purpose. Despite the shortcomings, QALY and DALY continue to be used as a metric for measuring the impact of health interventions. While QALYs are more commonly used in developed countries, DALYs are used for measuring burden of disease and cost-effectiveness of health interventions in low income countries. Threshold values on cost per DALY saved with a public health intervention in developing countries have emerged as thumb-rules for practitioners in the field of public health to accept or reject an intervention. A common threshold for interventions in low income countries is \$100 per DALY saved [15].

Table 1 shows some common interventions in developing countries and their cost-effectiveness ratios measured in US\$/DALY averted. The ranges of the cost-effectiveness ratios for many of these interventions are wide and vary according to the local setting due to varying ability to target specific populations. It is also evident that although the \$100 per DALY averted thumb rule is mostly used, in instances such as HIV/AIDS where the disease can have a much wider

Table 1. Cost-effectiveness (US\$/DALY) for Some Common Public Health Interventions in Developing Countries

Disease	Intervention	Cost Effectiveness (US\$/DALY)	Cost Effectiveness Range (US\$/DALY)
HIV/AIDS	Antiretroviral therapy	922	350–1494
HIV/AIDS	Condom promotion and distribution	82	52–112
HIV/AIDS	Mother-to-child transmission prevention	192	7–377
HIV/AIDS	Peer and education programs for high-risk groups	37	6–68
HIV/AIDS	Treatment of opportunistic infections	156	3–310
HIV/AIDS	Voluntary counseling and testing	47	10–85
Malaria	Insecticide treated bed-nets	11	5–17
Malaria	Intermittent preventive treatment in pregnancy with sulfadoxinepyrimethamine	19	13–24
Malaria	Residual household spraying	17	9–24
Tuberculosis	Directly observed short-course chemotherapy	301	84–551
Traffic accidents	Increased speeding penalties, enforcement, media campaigns, and speed bumps	21	3–38

[Source: Data obtained from disease control priorities project. <http://www.dcp2.org>.]

social impact than captured in DALYs, resources are allocated to recommended interventions such antiretro viral treatment despite their high cost per DALY averted.

DISTRIBUTIONAL EQUITY CONSIDERATIONS IN PUBLIC HEALTH

OR models for public health are distinct from other OR models because apart from commonly used objectives such as profit, output, and costs, they often also consider *equity of outcomes*. Effective public health requires that there are no large disparities between the availability of medicines, preventive health services, and other inputs across different socioeconomic segments of the population. OR models commonly incorporate equity constructs by setting constraints on minimum and/or maximum levels of an output provided to a given socioeconomic segment and in some cases by minimizing the sum of absolute deviations from the mean level of product or service provision.

A commonly used measure is the concentration curve, which is a plot of the cumulative proportion of the specific health-related variable being measured (income, health, medicine availability, distance to nearest facility etc.) against the cumulative

proportion of the population ranked by income, beginning with the poorest, and ending with the richest. Equality on the curve is represented by a diagonal line of slope 1, and the greater the deviation of the curve from this line, the greater the inequality. So for instance, if everyone, irrespective of his or her income, has exactly the same value of the variable which measures poor health status, the concentration curve will be a 45-degree line. On the other hand, if the health variable takes higher values (reflecting poorer health status) among poorer people, the concentration curve will lie above the line of equality. The farther the curve is above the line of equality, the more concentrated is poor health among the poor.

Figure 2 shows the infant mortality measured as deaths of children under five years of age against the cumulative births ranked by income. The curve for India lies above the line of equality, indicating that under-five child deaths in India are concentrated among the poor. In comparison, the Mali curve lies everywhere below that of India implying there is less inequality in under-five child deaths in Mali than in India.

Such concentration curves provide useful visual representation of inequality but provide limited ability to rigorously compare the

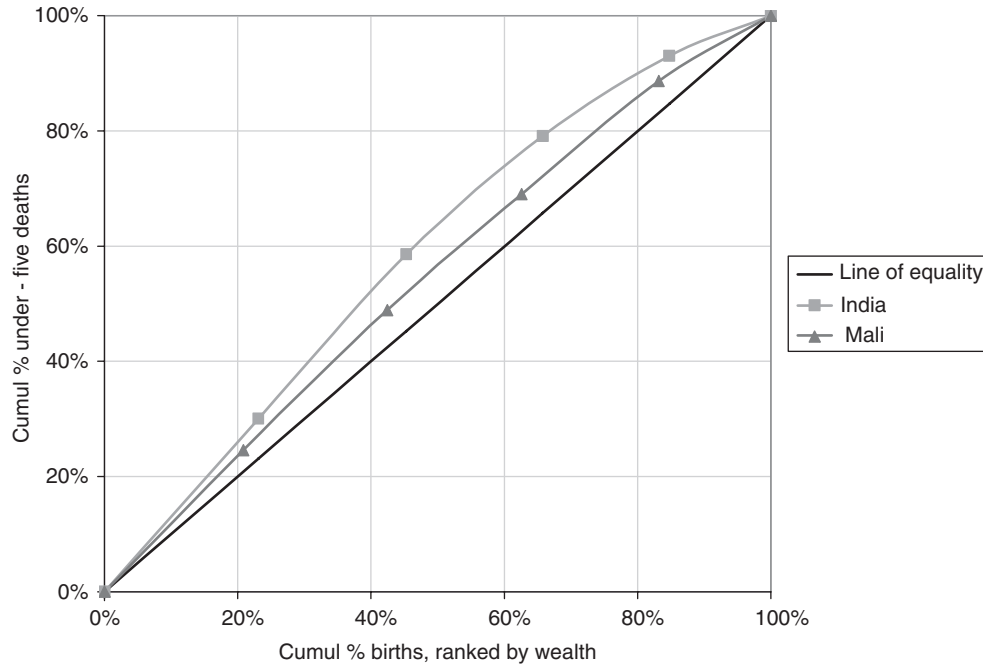


Figure 2. Lorenz curves for under-five deaths in India and Mali. *Source:* Doorslae. *et al.* 2008 [16].

Table 2. Gini Index of Health Facilities in the Lagos State of Nigeria

Facility Type	Gini Index
Primary health center	0.12107
Public secondary health facility	0.32567
Private hospital	0.23163
Public and private secondary health facilities	0.14015

[Source: Data obtained from Oguntade and N.A. Yusuph [18].]

extent of inequality across units of geographical analysis and over time. Different curves can be superimposed on each other as in Fig. 2 to see the difference but the number of such comparisons is limited.

An alternative method is thus to compute the Gini coefficient [17] of the public health input or output metric. The Gini coefficient G is defined as the mean of absolute differences between all pairs of individuals for a chosen measure.

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \bar{x}},$$

where x_i = observed value of the measure, $\bar{x}$ = mean of the observations, and n = number of observations. The minimum value of G is 0 implying perfect distributional equality and the maximum value is 1 which is perfect inequality (implying one individual possesses all of that measure).

A very common use of the Gini coefficient is to measure the distribution of access to health facilities by population in facility location problems or the distribution of availability of drugs by socioeconomic quintiles. Table 2 shows the Gini coefficient for the state of Lagos in Nigeria for different types of health facilities. If the entire population has perfectly equal access to a health facility, the Gini coefficients would be 0.

The success of vaccination programs in low income countries typically is measured not just on the total number or proportion of children vaccinated but the Gini coefficient is calculated to understand any distributional inequities. Design of targeted programs for vaccination, the location of new vaccination clinics and mass distribution campaigns are all driven by such analysis.

BASICS OF DISEASE MODELING

Understanding and quantifying the impact of public health interventions requires some preliminary understanding of disease transmission. Although a detailed discussion of disease transmission model is beyond the scope of this article, a basic model of disease transmission is discussed to give the OR practitioner a general idea. Ross [19] first developed a mathematical model of infectious disease transmission, which was used extensively by Macdonald [20] for studying malaria transmission. Commonly known as the Ross and McDonald model, it is now the basic building block in our understanding of disease transmission.

Most disease models are compartmental models and in the simplest sense they split the human population into a proportion of those who are susceptible to an infection, those who are infected, and those who have recovered and acquired temporary or permanent immunity (Fig. 3). Recovered and immune individuals can return back to the susceptible state after a certain period and birth and natural or disease induced death occur in the different compartments. The changes in the proportions of these three categories are described by differential equations where the parameters are estimated from field studies and vary depending upon the nature of the disease and infection dynamics. Models with more number of

states (or compartments) are closer at depicting reality but the resulting system of differential equations becomes complex to solve. For a detailed discussion of dynamic models of disease progression, see Anderson and May [21].

A key parameter of the disease model is the basic reproduction rate R_0 , which is the number of secondary infections that can arise when a single infected individual is introduced into a population where everyone is susceptible. Very simply stated, if $R_0 = 2$, there will be two new individuals who will be infected in the first round of transmission, 4 in the second, 8 in the third, and so on. If $R_0 = 100$, there will be 100 secondary infections in the first round, 10,000 in the second, and so on. In reality, the dynamics are more complicated as the susceptible population itself changes. Theoretically, if $R_0 > 1$, the number of infected people will grow exponentially until the entire population is infected. In reality, however, the development of immunity and other factors may arrest the development of disease. If $R_0 < 1$, the number of cases declines exponentially until the infectious disease can be eliminated or eradicated. Using malaria as an example we present some more details of the Ross and McDonald model

$$R_0 = \frac{ma^2 p^n l}{-\log_e p},$$

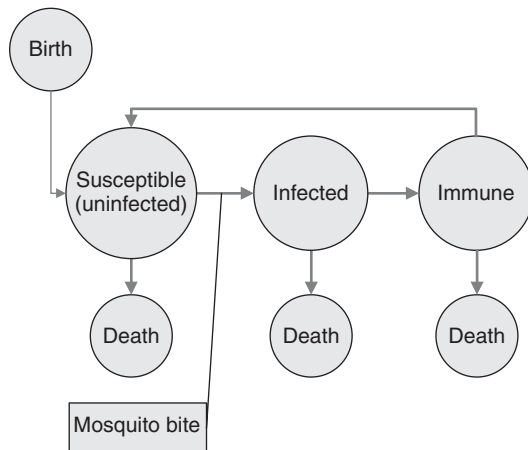


Figure 3. Basic compartmental model of malaria transmission.

where m is the number of female mosquitoes per person in the region, a is the rate at which each mosquito bites, p is the proportion of mosquitoes that survive from one day to the next, n is the maturation period of the parasite in the mosquito, and l is the duration of a single infection.

The rate a at which mosquitoes bite appears as a squared term because a cycle of transmission requires one bite to infect the mosquito and one to pass the infection back to a human. More involved versions of the model also include a probability that a bite from an infected mosquito results in human infection.

The interventions to control malaria can be broadly categorized into three main types: indoor residual spraying, distributing insecticide treated bed-nets, and treatment or prophylaxis with drugs. Transmission of malaria is influenced by many factors and an operations researcher needs to understand which interventions will have the strongest impact on reducing transmission in order to be able to develop allocation models. The above model makes it clear that measures which can reduce p , the proportion of (infected) mosquitoes that survive, is far more beneficial in reducing disease transmission than, for example, insecticidal spraying on mosquito breeding sites, which only impacts m . We can also see that parasite elimination through prompt treatment of infected individuals is also very beneficial as it decreases the duration of an infection l but may also indirectly decrease a . This demonstrates why a good understanding of the underlying disease transmission model is essential for building usable OR models for public health.

Admittedly, the above model was a simplified model only to illustrate how understanding and modeling of disease models remain key for optimal resource allocation. In reality, however, most infection transmission models have nonlinearities that can either lead to stable or unstable equilibrium points. Analysis of resource allocation decisions where disease progression follows a nonlinear transmission presents a complex problem which requires sophisticated OR techniques to solve.

OPERATIONS RESEARCH AND RESOURCE ALLOCATION FOR PUBLIC HEALTH INTERVENTIONS

Public health policy makers at different levels (global, national regional) face a range of resource allocation problems. First, they have to decide the allocation of resources between different diseases (e.g., malaria, HIV/AIDS, diarrhea, tuberculosis, maternal health); secondly, the allocation of resource between multiple interventions within each disease area (e.g., indoor insecticide spraying, distribution of bed-nets, distribution of medicines for treatment, and chemoprophylaxis); and thirdly, the allocation of resources to specific operational and tactical models for delivery of each intervention. As discussed in the section titled “Key Metrics in Public Health,” a commonly used method for resource allocation is cost-effectiveness analysis based on cost per DALYs saved or cost per QALY gained. Notwithstanding the shortcomings of QALYs and DALYs, other issues further exacerbate the complexity of public health resource allocation decisions. Some of these include nonconstant returns to scale; differences in measures of cost-effectiveness; high dependence of cost-effectiveness on the spatial distribution of incidence; trade-off between increased efficiency and effectiveness through intervention targeting versus ensuring distributional equity; and interconnectedness of returns from each intervention. We start with a simple example to illustrate resource allocation in public health.

Imagine there are two diseases: Disease A kills 5000 people per year, and Disease B kills only 50 people per year. It costs \$100 per person to prevent people from dying of Disease A and \$50 per person to prevent death from Disease B. The public health resource allocator has \$10,000 for disease control. How should he spend the money? Note that a naïve public resource allocator would set priorities based on the high mortality from Disease A and allocate all the resources available (\$10,000) to Disease A and save 100 lives. If the objective is to maximize the total number of lives saved, a simple linear programming formulation would give us the optimal resource

allocation.

x_A = Budget allocated to Disease A,

x_B = Budget allocated to Disease B

Maximize number of lives saved = $\frac{x_A}{100} + \frac{x_B}{50}$

Subject to

$$x_A + x_B = 10,000$$

$$\frac{x_A}{100} \leq 5,000; \frac{x_B}{50} \leq 50.$$

Clearly, the optimal allocation is $x_A = \$7500$ and $x_B = \$2500$ with a total of 125 lives saved.

Typically, the cost of interventions are not linear. For example, it is easier and cheaper to conduct insecticide spraying (or drug distribution) in urban areas as populations are easier to reach and population density is higher. The cost of spraying (or drug distribution) is much higher for rural populations, which are more difficult to reach [22]. Thus, the cost per-unit of output for an intervention depends upon the level of that intervention. A piecewise linear model with decreasing returns is commonly used to capture this in public health resource allocation models. Often, it is also difficult to separate the benefits from each intervention.

The cost per DALY saved in Table 3 is an average measure for each intervention and they vary considerably across and within countries depending on the degree of malaria

incidence [24]. The wide ranges in the cost per DALY saved is attributed to challenges in measurement due to varying scales of the intervention and nonlinearity in the costs. This shows that the design of optimal resource allocation models has to be region- and context-specific as they will depend on the level of disease incidence, current scale of existing interventions, geo-spatial distribution of populations at risk, and other health infrastructure related issues. Resource allocation managers and public health OR analysts must acknowledge that it is not going to be easy to obtain specific recent estimates of the cost-effectiveness of every intervention for each specific context. Therefore, OR models need to be built based on results from other contexts using carefully selected extrapolation and statistical matching methods.

One approach to enhance the allocative efficiency between different interventions is to stratify regions in terms of their epidemiology, sociology, and other characteristics and select targeted region-specific interventions. However, this approach may also lead to lower distributional equity of the interventions or their outcomes. Carefully managing the dual objectives of efficiency and equity is a challenge that resource allocation managers in public health constantly face. In addition to the distributional equity considerations, epidemiological or sociodemographic data is hard to obtain in low income countries, which often makes the costs of identifying and/or stratifying the target populations extremely high.

In the long term, resource allocation decisions are a vector of allocations to each intervention for each time period and the optimal allocation will vary considerably based on whether the objective is to eradicate a disease or control a disease. The dynamics of disease progressions imply that there is high degree of complementarity between the interventions and the presence of one intervention positively impacts the benefits from the other. Such interrelationships become complex to model in traditional OR models and sometimes require systems dynamic models [25]. Similarly, there are positive

Table 3. Cost-Effectiveness of Selected Malaria Interventions

Intervention	Cost per DALY Saved (Range)
Insecticide treated bed-nets (net + insecticide treatment)	\$11–\$17
Indoor residual spraying (1 round)	\$9–\$12
Indoor residual spraying (2 rounds)	\$17–\$24
Intermittent preventive therapy during pregnancy	\$13–\$24

[Source: Data obtained from Jamieson *et al.* [23], *Disease Control Priorities in Developing Countries*, 2nd edition, 2006.]

complementarities across disease-specific interventions implying that a disease-specific public health intervention has benefits for other diseases.

EXAMPLES OF OR PROBLEMS IN PUBLIC HEALTH

With a background on the metrics, objectives, and other notable considerations in OR for public health, we now turn to a few application areas. In Table 4, we present the most commonly observed challenges in developing country public health programs and the nature of OR models that could potentially be used to address that problem.

We describe a few important classes of problems which have particular relevance to public health in the developing world.

Facility Location Problems

Facility location problems have been a cornerstone of OR since the work of Weber [26], who considered the problem of locating a warehouse to minimize the total travel distance between the warehouse and a set of customers. Much later, Hakimi [27] considered the more general problem of locating one or more facilities to minimize either the sum of distances or the maximum distance between facilities and customers. Determining the optimal number of facilities was also considered in many location models. For a detailed review on facility location problems, see Brandeau and Chiu [28].

In traditional facility location models, the objective is to minimize either the weighted sum of distance or the maximum distance traveled. However, a public health planner may want to minimize the maximum distance (or a nonlinear function of maximum distance) traveled to capture the equity considerations presented earlier. The model of Maimon [29], which looks at facility location for public sector applications by incorporating equity and variance of distance traveled in the formulation is particularly applicable to public health settings. Similarly, the model of Schilling [30] considers multiple objectives in locating facilities with public sector objectives.

Traditional location models work under the assumption that individuals choose the facility that is closest to them. Given the social stigma associated with diseases such as HIV/AIDS and sexually transmitted infections, individuals in a public health setting may not always choose facilities closest to them. Literature in public health also shows that attendance at public health clinics falls with increasing distance to the facility [31,32]. For diseases such as TB and HIV/AIDS that require long-term treatment nonadherence to treatment can lead to higher costs as nonadhering patients migrate from first-line to more costly second-line treatment. For TB, Shargie and Lindtjörn [32] show that treatment adherence decreases as distance traveled by patients to the clinic increases. Using that premise [33] investigated the impact of patients' travel

Table 4. Public Health Challenges in Developing Countries and Potential OR Models

Public Health Challenge in Developing Countries	Potential OR Models
Poor physical accessibility to health facilities	• Facility location
High waiting time at public health facilities	• Capacity management
Acute shortage of health care workers	• Capacity management • Staff deployment
Nonavailability/shortages of drugs, vaccines, and other health products at public health facilities	• Network design • Optimal multiechelon inventory control • Inventory rationing

distance to HIV/AIDS clinics on treatment adherence and developed a model for locating new HIV/AIDS clinics in Zambia to maximize adhering population. Their simulation model shows that taking into account patient attendance and adherence in optimal ARV facility location placement could lead to a potential 3% increase in patients starting and adhering to treatment.

Another area of concern to public health is the average or maximum response time of a facility location model to contain disease outbreaks. In such circumstances, opening new temporary facilities to quickly distribute (vaccines, drugs, and therapeutics) to the affected population [34,35] is an effective outbreak containment strategy. The economic trade-off is between the number of facilities to open and the facility capacity versus the distance traveled and its impact on outbreak containment. OR models [35] determine *how many* new facilities should be opened and where and also rules on the assignment of individuals to the facilities. Heier *et al.* [36] consider the more general problem of decentralized facility assignment in such a scenario when the users pick the facility themselves and establish the cost of “anarchy” in such a system. They show that in some cases when patients pick their facilities based on full information about facility status and travel distances, it results in greater total travel time than a centralized solution where patients are allocated to facilities or when patients do not have full information about facility status.

Another key concern for public health in developing countries is the performance of a facility model when hospital infections such as MRSA (methicillin-resistant *Staphylococcus aureus*) or machine and equipment failures that are common in developing countries could temporarily shut down a health clinic. Berman *et al.* [37] look at the classical p -median facility location problem and introduce the possibility that a facility might suffer a disruption. Using the example of Toronto hospitals, they show that considering the probability of failure in the facility location model results in a higher degree of centralization. Their results corroborate the colocation and centralization

observed in Toronto and many other hospital systems.

The choice of the location of public health facilities is a key strategic decision variable in the design of public health systems. OR techniques with carefully selected objectives functions can help find facility locations which can maximize public health impact through better access to those requiring treatment or preventive interventions. Future research could focus on incorporating patient's health facility choice models and other benefit externalities into the optimal facility location problem.

Capacity Management Models

The emergence of new infectious diseases or the resurgence of old ones like malaria suddenly presented additional demand on the public health system, but the system could not develop as quickly due to financial and human resource constraints. The lack of human resource capacity is illustrated by the fact that sub-Saharan Africa which has over 25% of the global burden of disease has only 2–3% of the world's health workers. Unfortunately, the money to hire, train, and sustain new health care workers is not available and is not likely to be available in the medium term [38]. This acute scarcity of health care workers in developing countries, especially in sub-Saharan Africa, offers limited ability in the short- to medium-term to apply models that help determine optimal staffing or capacity levels for clinics. The proper deployment of health staff and careful management of the available capacity in clinics can be a key approach to maximizing public health outcome in such an environment. This is currently lacking and long queues continue to form at many points within the public health system.

Many different queueing models for capacity management in health-care setting have been developed [39,40]. An important point of difference in the case of public health clinics in developing countries is that in many cases average capacity to serve patients is less than the arrival capacity. Thus, in such systems, patients renegeing or service rationing is the only way that a system can attain the state of equilibrium;

otherwise, queue lengths keep increasing with time (the analysis of transient queue behavior is beyond the scope of this article). When server utilization is very high, average waiting time can be minimized by giving priority to patients who require shorter service times. Thus, the shortest processing time rule instead of the first-in first-out queue discipline could minimize total waiting times. However, such a rule is difficult to implement in practice due to its perceived unfairness amongst the patients especially in single server systems. If there were two servers, a system like the one observed commonly in supermarkets where there is dedicated express service counter for shorter processing time customers could be envisaged. Using a lesser trained health worker to treat shorter processing time patients can be an effective means of capacity management in public clinics. Thus, a triage step to know patient needs when they enter the queue is important whether to enforce the shortest processing time discipline or direct them to the lesser trained health worker. In reality, health systems are organized as tiered networks with referrals from one stage to the other. Thus, an overall analysis of public health systems requires analyzing queueing networks. Discrete event simulation could thus be a very important OR technique to utilize in the public health system. An overview of applications of discrete event simulation in health care can be found in Ref. 41. Also Hopp and Spearman [42] provide easy to use queueing approximations which can be used for modeling multistation public health systems and to analyze the impact of different parameters and flow arrangements in a public health clinic on waiting time, throughput, and variability propagation.

Since public health is usually free and does not involve any price mechanisms to regulate demand and supply, revenue management based methods of service differentiation or capacity management are not applicable in this setting. Pure rationing to restrict demand is not ethical and creates many sociocultural problems. The renege process works in a manner that individuals who can afford to obtain care at other locations quickly drop out of the system. The

long waiting time in public health clinics in some developing countries acts as a rationing system which favors those who have a low opportunity cost of time and penalizes those who have short or long term formal employment [43]. Thus, on the surface it appears that long queueing automatically leads to socially efficient rationing on the basis of socioeconomic factors. In reality, however, informal arrangements lead to queue jumping which benefits those with more resources making the queue renege phenomenon into an inequitable and unjust rationing approach. Queueing based analysis, nevertheless, can be an important tool to devise capacity management strategies for public health clinics in developing countries. Future research could focus on modeling the impact of different task shifting interventions (point-of-care testing, nurse triage, telemedicine, decoupling drug dispensing from care provision) and varying queue disciplines in developing country health programs with severe capacity constraints.

Stock Rationing Models

Availability of drugs, vaccines, and other health supplies is very low at the clinic levels in many developing countries. Average availability at public health facilities in certain regions within a group of 36 surveyed countries was as low as 29.4% [44]. The typical structure for distribution of medicines in the public sector in most developing countries consists of a central/primary warehousing and distribution point which supplies to one or two downstream stock holding points depending upon the distribution of population and administrative structure of a country. The level of availability is low at each of the stocking points in the supply chain and thus each stage in the system has to engage in stock rationing. In systems where stock rationing is very common, if the rationing is done based on a fixed proportion of the order size, each downstream unit (clinics or sub-warehouses) inflates its orders to get a bigger share of the available supply [45,46]. If past consumption is used to allocate stock [46], changes in seasonality

and consumption patterns as a result of epidemiological patterns or differential coverage of preventive interventions are not adequately captured. In addition, the degree of lost sales at each clinic continues to get repeated in the future.

The traditional models of inventory rationing in OR analyze the problem of how to allocate inventory to different customer classes [47,48] by setting a threshold of inventory for each class of customers. Most of these models assume a known revenue or cost for each customer class and define the optimal threshold levels for rationing based on overall revenue maximization or cost minimization. Pibernik and Yadav [49] consider a model of capacity reservation, based on service-level objectives when long-term cost or revenue measures of each customer class are not available. Zenios *et al.* [50], in their model for allocating cadaveric kidneys among various patient segments, consider an objective function incorporating QALYs. Also, in most of these models, the customer segments are independent with no overlap. Optimal rationing of drugs at the clinics to achieve better public health outcomes will be based on customer segments that are defined by demographic and health characteristics (medical threshold e.g., CD4 count or severity of disease or income level, sex, age). This implies that their membership in a customer segment is itself dependent upon the rationing decision. Deo [51] considers such a model where a public health planner who wants to maximize the total expected QALYs over a long time horizon, creates the optimal rationing strategy for HIV/AIDS patients. The determination of future QALYs from the rationing decision is captured through a simple disease transmission model. Through numerical analysis he shows that a rationing policy that follows open enrollment with enforced prioritization of current patients out-performs rationing heuristics that are commonly used in practice.

Stock rationing and allocation problems in public health require a model which considers the long-term objective of maximizing public health outcomes (DALYs, QALYs) with disease models to understand the

impact of the rationing strategy on long-term disease progression. Better understanding of behavioral issues involved in rationing health commodities is also needed for the rationing models to be put into practice.

CONCLUSION

This article presents an overview of the use of OR in public health in developing countries by first introducing the objective functions and output metrics commonly used by public health decision makers. A discussion of DALY and QALY metrics reveals that although they are commonly used in OR models for public health, their design is not flawless. The simple Ross and Macdonald model of disease transmission applied to malaria illustrates that resource allocation problems in OR need basic understanding of disease transmission. The concept of distribution equity is key to many public health OR problems and besides simple max–min formulations, commonly used measures of inequity such as the Gini coefficient and concentration curve are also required in many public health OR models. The number and scope of OR models for public health in developing countries continues to grow as work in this field continues apace. Understanding the main needs for OR in developing country public health programs and working closely with implementation partners are crucial for the OR/MS practitioner. Application areas of OR/MS in developing country health programs are highlighted along with opportunities for future research in this area.

Acknowledgments

The author gratefully acknowledges the comments of two referees and the associate editor in improving the content and presentation of this article.

REFERENCES

1. World Population Prospects. 2006. United Nations. Available at http://www.un.org/esa/population/publications/wpp2006/WPP2006_Highlights_rev.pdf. Accessed 2009 June 25.

2. Korenromp E, Komatsu R, Katz I, *et al.* Operational Research on HIV/AIDS, tuberculosis and malaria control in Global Fund-supported programs: rounds 1–6 grants. In: Proceedings of the 5th European Conference on Tropical Medicine and International Health; Amsterdam: 2007 Aug.
3. Brandeau ML, Pierskalla WP, Sainfort F, editors. Operations research and health care: a handbook of methods and applications, International Series in Operations Research and Management Science, 70. Boston (MA): Kluwer Academic Publishers; 2004.
4. Winslow CEA. The untilled field of public health. *Mod Med* 1920;2:183–191.
5. The Future of Public Health. Committee for the study of the future of public health; division of health care services 1998. Washington (DC): National Academy of Science Press.
6. Turnock BJ. Public health: what it is and how it works. Sudbury (MA): Jones & Bartlett Publishers; 2004.
7. Morse PM, Kimball GE. Methods of operations research. New York: Wiley; 1951.
8. Waddington CH. Operational Research. *Nature* 1948;161:404.
9. Mathers CD, Fat DM, Inoue M, *et al.* Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ* 2005;83: 171–177.
10. Zeckhauser R, Shepard DS. Where now for saving lives? *Law Contemp Probl* 1976;40:5–45.
11. World Health Organization (WHO). The global burden of disease: 2004 update. Geneva: World Health Organization.
12. World Bank. World Bank Development Report 1993: investing in health. New York: Oxford University Press; 1993.
13. Sassi F. Calculating QALYs, comparing QALY and DALY calculations. *Health Policy Plann* 2006;21(5):402–408. DOI: 10.1093/heapol/czl018 2006.
14. Anand S, Hanson K. Disability-adjusted life years: a critical review. *J Health Econ* 1997;16:685–702.
15. Banerjee AV, Benabou R, Mookherjee D, editors. Understanding poverty. New York: Oxford University Press; 2006.
16. van Doorslaer E, Wagstaff A, Lindelow M. The World Bank. Washington (DC); 2008. Available at www.worldbank.org/analyzing_healthequity. Accessed 2009 July 01.
17. Gini C. Variabilità e mutabilità. In: Pizetti E, Salvemini T, editors. 1912 reprinted in *Memorie di metodologica statistica*. Rome: Libreria Eredi Virgilio Veschi; 1955.
18. Oguntade AE, Yusuph NA. Health infrastructure inequality—a case study of lagos state, Nigeria. *Soc Sci* 2007;2(1):51–55.
19. Ross R. The prevention of malaria. New York: E.P. Dutton & Co.; 1910.
20. Macdonald G. The analysis of equilibrium in malaria. *Trop Dis Bull* 1952;49:813–829.
21. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. USA: Oxford University Press; 1992.
22. Barat LM, Palmer N, Basu S, *et al.* Do malaria control interventions reach the poor? A view through the equity lens. *Am J Trop Med Hyg* 2004;71(2 Suppl):174–178.
23. Jamison DT, *et al.* editors. Disease control priorities in developing countries. USA: Oxford University Press; 2006.
24. Goodman CA, Coleman PG, Mills AJ. Cost-effectiveness of malaria control in sub-Saharan Africa. *Lancet* 1999;354:378–385.
25. Homer J, Hirsch G. Opportunities and demands in public health systems. *Am J Public Health* 2006;96(3):452.
26. Weber A. *Über den Standort der Industrien*, 1909; translated as Alfred Weber's *Theory of the Location of Industries*. Chicago (IL): University of Chicago; 1929.
27. Hakimi SL. Optimum locations of switching centres and the absolute centres and medians of a graph. *Oper Res* 1964;12:450–459.
28. Brandeau ML, Chiu SS. An overview of representative problems in location research. *Manage Sci* 1989;35:645–674.
29. Maimon O. The variance equity measure in locational decision theory. *Ann Oper Res* 1986;6(5):147–160.
30. Schilling DA. Dynamic location modeling for public sector facilities: a multicriteria approach. *Decis Sci* 1980;11:714–724.
31. Rahaman MM, Aziz KM, Munsh MH, *et al.* A diarrhea clinic in rural Bangladesh: influence of distance, age, and sex on attendance and diarrheal mortality. *Am J Public Health* 1982;72(10):1124–1128.
32. Shargie EB, Lindtjörn B. Determinants of treatment adherence among smear-positive pulmonary tuberculosis patients in southern ethiopia. *PLoS Med* 2007;4(2):37.
33. Carr CC, Jallah JD. Improving spatial accessibility to antiretroviral treatments for

- HIV/AIDS [MS thesis]. Zaragoza, Spain; MIT-Zaragoza International Logistics Program; 2008.
34. Kaplan EH, Craft DL, Wein LM. Emergency response to a smallpox attack: the case for mass vaccination. *Proc Natl Acad Sci U S A* 2002;99(16):10935–10940.
 35. Lee EK, Smalley HK, Zhang Y, *et al.* Facility location and multi-modality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. *Int J Risk Assess Manage* 2008;12:311–351.
 36. Heier J, Ergun O, Keskinocak P, *et al.* Decentralized facility location models. In: *INFORMS Annual Conference 2008*. Washington (DC); 2008.
 37. Berman O, Krass D, Menezes M. Facility reliability issues in network p-median problems: strategic centralization and co-location effects. *Oper Res* 2007;55(2):2332–350. DOI: 10.1287/opre.1060.0348.
 38. Conway MD, Gupta S, Khajavi K. Addressing Africa's health workforce crisis. *McKinsey Q* 2007.
 39. Bailey NTJ. Queuing for medical care. *Appl Stat* 1954;3:137–145.
 40. Green L. Queueing analysis in healthcare. In: Hall RW, editor. *Patient flow: reducing delay in healthcare delivery*. New York: Springer; 2006. pp. 281–308.
 41. Jacobson S, Hall S, Swisher J. Discrete-event simulation of health care systems. In: Hall RW, editor. *Patient flow: reducing delay in healthcare delivery*. New York: Springer Science; 2006. pp. 211–252, 281–308.
 42. Hopp W, Spearman L. *Factory physics*. New York: McGraw-Hill; 2000.
 43. Rosen S, Sanne I, Collier A, *et al.* Rationing antiretroviral therapy for HIV/AIDS in Africa: choices and consequences. *PLoS Med* 2005;2:11.
 44. Cameron A, Ewen M, Ross-Degnan D, *et al.* Medicine prices, availability, and affordability in 36 developing and middle-income countries: a secondary analysis. *Lancet* 2008;373(9659):240–249.
 45. Gonçalves P. Why do shortages inflate to huge bubbles? Working Paper. Cambridge (MA): Sloan School of Management, Massachusetts Institute of Technology; 2008.
 46. Lariviere M. Capacity allocation using past sales: when to turn-and-earn. *Manage Sci* 1999;45(5):685–703.
 47. Topkis DM. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Manage Sci* 1968;15:160–176.
 48. Nahmias S, Demmy WS. Operating characteristics of an inventory system with rationing. *Manage Sci* 1981;27:1236–1245.
 49. Pibernik R, Yadav P. Dynamic capacity reservation and due date quoting in a make-to-order system. *Nav Res Log* 2009;55(7):593–611.
 50. Zenios SA, Wein LM, Chertow GM. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Oper Res* 2000;49:549–569.
 51. Deo S. Rationing of HIV treatment in resource-constrained settings under supply uncertainty. In: *Proceedings of the Manufacturing & Service Operations Management Society Meeting*; 2008; Baltimore (MD).

INEQUALITIES FROM GROUP RELAXATIONS

JEAN-PHILIPPE P. RICHARD
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

INTRODUCTION

We consider mixed integer programs (MIPs) of the form

$$\min \left\{ \sum_{j \in N} c_j x_j \mid \sum_{j \in N} A_j x_j = d, x_j \in \mathbb{Z}_+ \right. \\ \left. \forall j \in I, x_j \in \mathbb{R}_+ \forall j \in C \right\} \quad (1)$$

where $N = \{1, \dots, n\}$, (I, C) forms a partition of N , $d \in \mathbb{Z}^m$, $A_j \in \mathbb{Z}^m$, and $c_j \in \mathbb{R} \forall j \in N$. We refer to the optimization problem obtained by relaxing the condition $x_j \in \mathbb{Z}_+$ to $x_j \in \mathbb{R}_+$ as the linear programming (LP) relaxation of problem (1). A fundamental problem in the design of cutting-planes and LP-based branch-and-cut algorithms is the derivation of valid inequalities for problem (1). This problem is particularly difficult for MIPs without special structure. The group approach is a method to generate cutting-planes for such problems. This approach requires that, any time a cutting-plane is generated, an optimal solution of the LP relaxation and a corresponding simplex tableau are known. Cuts are then generated using the information contained in one or several rows of this tableau.

As an illustration, consider

$$\min \left\{ -x_1 \mid -x_1 + 2x_2 \leq 4, 2x_1 - x_2 \leq 6, \right. \\ \left. 6x_1 + 5x_2 \leq 38, (x_1, x_2) \in \mathbb{Z}_+^2 \right\} \quad (2)$$

which can be put in the form of problem (1) by adding (integer) slack variables x_3, x_4 , and x_5 . The feasible region of problem (2) and its LP relaxation are represented in Fig. 1a. The optimal solution $x^{\text{LP}} = (4.25, 2.5)$ of the LP relaxation is highlighted with a star. The corresponding simplex tableau is

$$\begin{aligned} \min & -\frac{68}{16} + \frac{5}{16}x_4 + \frac{1}{16}x_5 \\ \text{s.t. } & x_1 + \frac{5}{16}x_4 + \frac{1}{16}x_5 = \frac{68}{16} \\ & x_2 - \frac{6}{16}x_4 + \frac{2}{16}x_5 = \frac{40}{16} \\ & x_3 + \frac{17}{16}x_4 - \frac{3}{16}x_5 = \frac{52}{16} \end{aligned} \quad (3)$$

where $B = \{1, 2, 3\}$ and $N = \{4, 5\}$ are the sets of indices of basic and nonbasic variables respectively.

We now describe how the group approach derives a cut for problem (2) from simplex tableau (Eq. 3). To this end, we consider the relaxation of problem (2) obtained by ignoring the restriction that basic variables x_1, x_2 , and x_3 are nonnegative but maintaining the restriction that they are integer. This relaxation K is described in the space of nonbasic variables (x_4, x_5) as

$$\begin{aligned} \min & -\frac{68}{16} + \frac{5}{16}x_4 + \frac{1}{16}x_5 \\ \text{s.t. } & \frac{5}{16}x_4 + \frac{1}{16}x_5 \equiv \frac{4}{16} \pmod{1} \\ & \frac{10}{16}x_4 + \frac{2}{16}x_5 \equiv \frac{8}{16} \pmod{1} \\ & \frac{1}{16}x_4 + \frac{13}{16}x_5 \equiv \frac{4}{16} \pmod{1} \\ & x_4 \in \mathbb{Z}_+, x_5 \in \mathbb{Z}_+ \end{aligned} \quad (4)$$

where the effect of integer variables x_1, x_2 , and x_3 is emulated by the “modulo 1” expression. The relaxation (Eq. 4) of problem (2) is itself an MIP for which facet-defining inequalities can be derived. We represent the set K of solutions (x_4, x_5) that satisfy the constraints of problem (4)

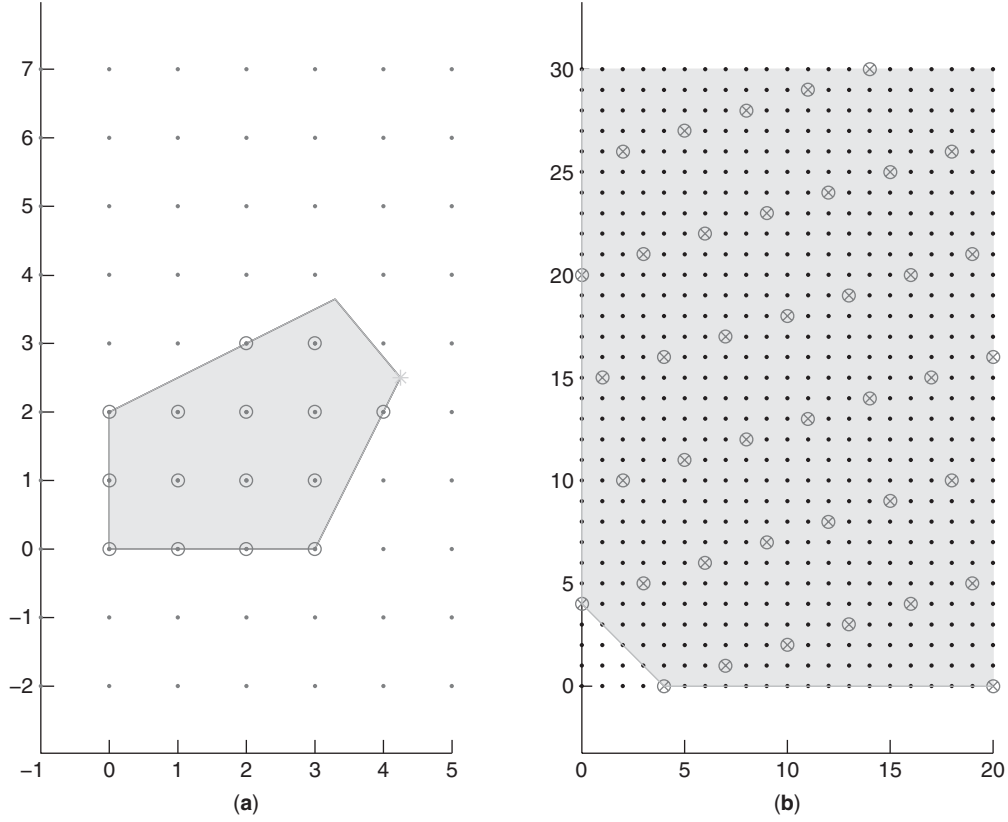


Figure 1. Feasible region of Equation (2) and corner relaxation.

with thicker bullets in the Fig. 1b. From this illustration, it can be verified that $(K) = \{(x_4, x_5) \in \mathbb{R}^2 \mid x_4 + x_5 \geq 4, x_4 \geq 0, x_5 \geq 0\}$. Substituting slack definitions for x_4, x_5 , we obtain the following description of the corner relaxation in the space of original variables $\{(x_1, x_2) \in \mathbb{R}^2 \mid 2x_1 - x_2 \leq 6, 6x_1 + 5x_2 \leq 38, 2x_1 + x_2 \leq 10\}$, which is represented in Fig. 2a. Because the basic solution associated with problem (3) is nondegenerate, relaxation (Eq. 4) can be graphically visualized as the set of integer solutions in the cone formed by the active constraints at x^{LP} . For this reason, it is often referred to as *corner relaxation* of problem (2) associated with basis B . It is clear that valid inequalities for problem (4) are also valid for problem (2) since problem (4) is a relaxation of problem (2).

As far as cutting-planes algorithms are concerned, we see in Fig. 2a that $2x_1 + x_2 \leq 10$ is a valid inequality for problem (2) that cuts off x^{LP} . More generally, nontrivial facets of the corner relaxation always yield violated cutting-planes when used in a simplex tableaux whose basic solutions are fractional. After adding $2x_1 + x_2 \leq 10$ to problem (2), we observe in Fig. 2b that the improved formulation still does not give the convex hull of feasible solutions to problem (2). However, its LP relaxation produces an optimal solution to problem (2).

In the previous example, it was simple to obtain a cut since the corner relaxation only had the two variables (x_4, x_5) . For more general cases, the above approach will only be effective if the inequalities describing the convex hull of K can be easily characterized. To obtain such a simple characterization, we

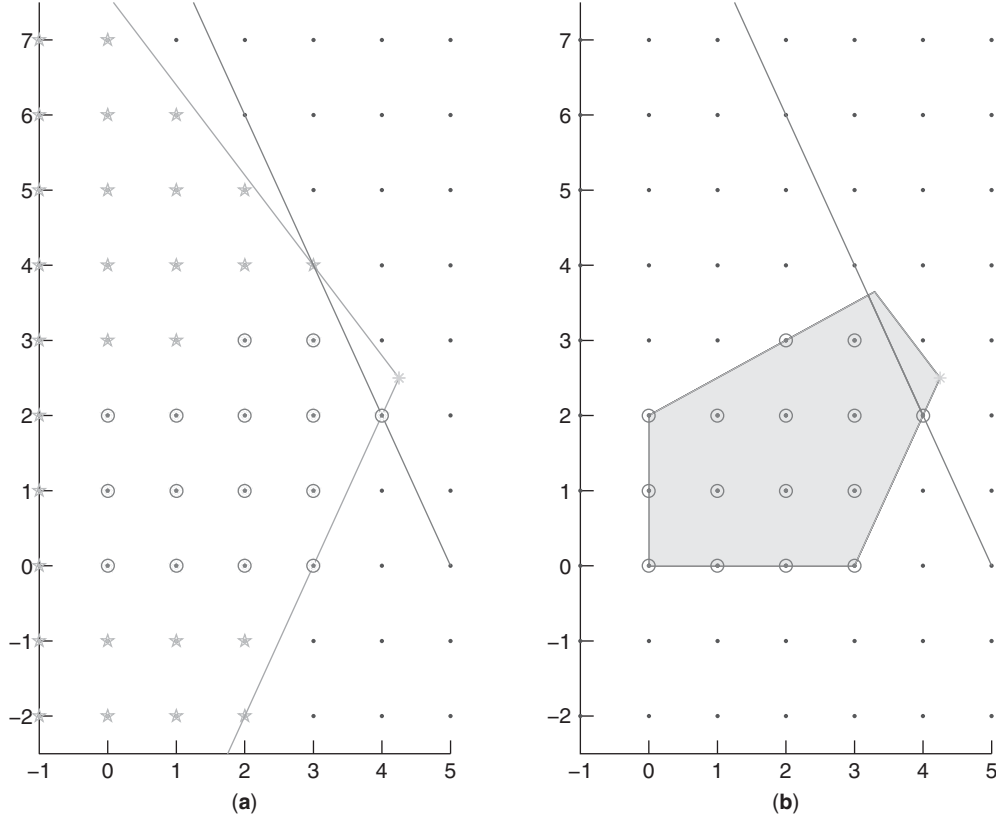


Figure 2. Corner relaxation and cutting-plane.

introduce a suitable set of new variables in the problem. In the case of problem (4), we study

$$\left\{ z \in \mathbb{Z}_+^{16 \times 16 \times 16} \mid \sum_{i=0}^{15} \sum_{j=0}^{15} \sum_{k=0}^{15} \frac{1}{16} \binom{i}{j}{k} z_{i,j,k} \equiv \frac{1}{16} \binom{4}{8}{4} \pmod{1} \right\} \quad (5)$$

where a variable z_g is introduced for each element g of the group $(G, +)$ where $G = C^{16} \times C^{16} \times C^{16}$ with $C^{16} = \{0, \frac{1}{16}, \frac{2}{16}, \dots, \frac{15}{16}\}$ and where the addition is taken modulo 1. It is clear that problem (5) is a relaxation of problem (4) since problem (4) can be obtained from problem (5) by fixing all variables z_g to 0 except for $z_{5,10,1}$ (which corresponds to x_4) and

$z_{1,2,13}$ (which corresponds to x_5). The advantage of problem (5) over problem (4) is that it applies to all simplex tableaux, whose entries are integer multiples of $\frac{1}{16}$ and whose right-hand sides are equal to $(\frac{4}{16}, \frac{8}{16}, \frac{4}{16})' \pmod{1}$ instead of applying only to problem (4). For this reason, this set is typically referred to as *master corner relaxation* of problem (3) and its convex hull is referred to as *master corner polyhedron* associated with problem (3). It is remarkable that, although the facetial structure of MIPs is typically complex, the facets of master corner polyhedra have a simple structure. Further, it can be shown that all facet-defining inequalities of the corner relaxation can be obtained from facet-defining inequalities of master corner relaxations; see Theorem 4.

The group-theoretic approach to MIP is the study of how corner (and more generally master corner) relaxations can be used to solve MIPs. Since corner relaxations are indeed relaxations of MIPs, they have been used both as a source of cutting-planes for MIPs and as a substitute for LP relaxations in branch-and-bound procedures; see Ref. 1 for a more detailed presentation. This article focuses solely on cutting-planes.

HISTORY

In 1958, Gomory [2] described the outline of the first finitely convergent cutting-planes algorithm for bounded MIP; see also Ref. 3. This algorithm successively generates cutting-planes to expose an optimal solution of the problem. The approach was later generalized to the mixed integer case [4]. Although the cuts central to this algorithm are group-theoretic cuts, they were not derived as such. The common principles underlying their derivation was discovered later. In fact, in a series of papers [5–7], Gilmore and Gomory studied how to solve cutting stock problems by column generation. While studying the solution of the integer knapsack problems that were required to solve their pricing problem, the authors discovered a periodic behavior to optimal solutions. The analysis of this behavior led Gomory to develop an asymptotic theory of MIP and to introduce corner and master corner polyhedra [8–10]. Following its formal introduction in Ref. 10, the group-theoretic approach for the generation of cutting-planes in MIP experienced a flurry of activities in the early 1970s. Gomory and Johnson [11,12] extended the theory to different types of groups. In particular, they studied infinite groups that are easier to use in practice, but are more difficult to study theoretically. The approach was further generalized by Johnson [13], who also considered in Ref. 14 models where constraints are imposed on basic variables. Johnson also investigated relationships to duality; see Ref. 15 for an exposition.

Early experiments with the Gomory cutting-planes algorithm concluded that,

although theoretically convergent, the algorithm did not often yield optimal solutions because it was prone to numerical problems. As a result, group cuts remain largely unused in computer implementations. A dramatic revival of interest occurred in the mid-1990s following the study [16] where Gomory mixed integer cuts (GMICs) were first shown to be useful in computation. GMICs were subsequently incorporated in all major commercial software. In 2003, new results and insights were presented in Refs 17 and 18. These papers provided a new impetus in the study of group problems. In particular, new extreme inequalities were developed for both finite and infinite group problems with a single constraint [19–22]; extreme inequalities were developed for group problems with multiple constraints [23,24]; the computational potential and quality of group cutting-planes was evaluated [25,26] and the strength of the corner relaxation was analyzed [27]. Most of this research focused on the “traditional” group framework in which integer variables are prevalent. In the last few years, however, a new stream of research has emerged that considers group problems in which all nonbasic variables are continuous [28]. This model and its variations have been vigorously investigated [29–33]. In particular, their study reveals interesting connections between group cutting-planes and intersection cuts from lattice-free convex sets.

VALID INEQUALITIES FROM GROUP RELAXATIONS

Corner and Master Corner Relaxations

We now introduce the notion of corner and master corner relaxation (or *group relaxation*) more precisely. We assume that the LP relaxation of problem (1) is feasible and bounded so that there is a basic feasible solution that is optimal. Let B and N represent the set of indices of basic and nonbasic variables respectively. Let A_B (respectively A_N) be the matrix whose columns correspond to the variables x_i for $i \in B$ (respectively $i \in N$). The optimal simplex tableau of the LP relaxation associated with B can be written as

$$x_B + A_B^{-1}A_N x_N = A_B^{-1}d. \quad (6)$$

The corner relaxation is obtained by relaxing the nonnegativity on basic variables x_B . For a tableau row in which the basic variable is continuous, relaxing the lower bound on the basic variable is equivalent to dropping the constraint. Therefore, we may assume without loss of generality that $B \cap C = \emptyset$. For the remaining tableaux rows in which basic variables are integer, relaxing lower bound on the basic variable amounts to requiring that the nonbasic variables satisfy the modular relation $A_B^{-1}A_N x_N \equiv A_B^{-1}d \pmod{1}$. The corner relaxation associated with B can then be defined as the set of solutions $(x_{N \cap I}, x_{N \cap C}) \in \mathbb{Z}_+^{|N \cap I|} \times \mathbb{R}_+^{|N \cap C|}$ that satisfy

$$A_B^{-1}A_N x_N \equiv A_B^{-1}d \pmod{1}. \quad (7)$$

In Equation (7), computation is carried in the group $(I^m, +)$ of real vectors whose components are in $[0, 1)$ and where the addition is taken modulo 1.

In order to obtain master corner relaxations (or group relaxations), we perform two additional steps. In the first step, we aggregate “duplicate” variables. In particular, we aggregate all integer variables whose columns in Equation (7) are identical modulo 1. We denote the set of all remaining columns by G^* . Similarly, we aggregate all continuous variables whose columns in Equation (7) are a positive multiple of each other. We also perform a normalization of the remaining columns so that they belong to the boundary S^m of the hypercube $\mathcal{H} = [-1, 1]^m$. We denote the set of all these columns by S^* . In the second step, we insert Equation (7) into a higher-dimensional space by adding variables. To this end, we identify a group $(G, +)$, where the addition is taken modulo 1 that is such that $G^* \subseteq G$. Similarly, we select a subset S of S^m that is such that $S^* \subseteq S$. Adding a variable for each element of G and each element of S yields the following definition where $\mathcal{F}$ is the function that associates to each real vector, the vector of its fractional parts and where $r = A_B^{-1}d \pmod{1}$.

Definition 1. Given $r \in I^m \setminus \{0\}$, a subgroup G of I^m , and a subset S of S^m , we define the mixed integer group problem, $\Gamma(G, S, r)$ to be the set of all functions $z : G \rightarrow \mathbb{R}$ and $y : S \rightarrow \mathbb{R}$ that satisfy the following properties:

1. $\sum_{g \in G} gz(g) + \mathcal{F}(\sum_{s \in S} sy(s)) = r$,
2. $z(g) \in \mathbb{Z}_+ \quad \forall g \in G$,
3. $y(s) \in \mathbb{R}_+ \quad \forall s \in S$,
4. z and y have finite supports.

In Definition 1, $z(g)$ and $y(s)$ are, respectively, the integer and continuous variables of the group relaxation. Condition 1 represents the defining equality of the master corner relaxation constructed from the simplex tableau. We refer to a problem for which there are m constraints in Condition 1 as being an m -dimensional mixed-integer group problem. Condition 4 is assumed to guarantee that the summation in Condition 1 is well defined. In Definition 1, we also implicitly assume that S is chosen in such a way that $\Gamma(G, S, r) \neq \emptyset$.

There are many possible choices for G in Definition 1. Assuming that the determinant of the basis matrix A_B is D , one could select the group $(G, +)$ such that $G = C^D \times \dots \times C^D$ and $C^D = \{0, \frac{1}{D}, \dots, \frac{D-1}{D}\}$. In this case, $\Gamma(G, S, r)$ is referred to as a *finite group problem*. Further, $(G, +)$ can always be chosen to be $(I^m, +)$. In this case, $\Gamma(G, S, r)$ is referred to as an *infinite group problem*. Similarly, there are many possible choices for S . It is typically chosen to be the empty set if cuts for pure integer programs are desired or chosen to be S^m otherwise. In the remainder of this article, we often use G and I^m to represent the groups $(G, +)$ and $(I^m, +)$, respectively.

Cuts from Group Relaxations

Although unusual in appearance, $\Gamma(G, S, r)$ is an MIP that is a relaxation of problem (1). As our ultimate goal is to generate valid inequalities for problem (1) from this relaxation, we now focus on characterizing strong valid inequalities for $\Gamma(G, S, r)$.

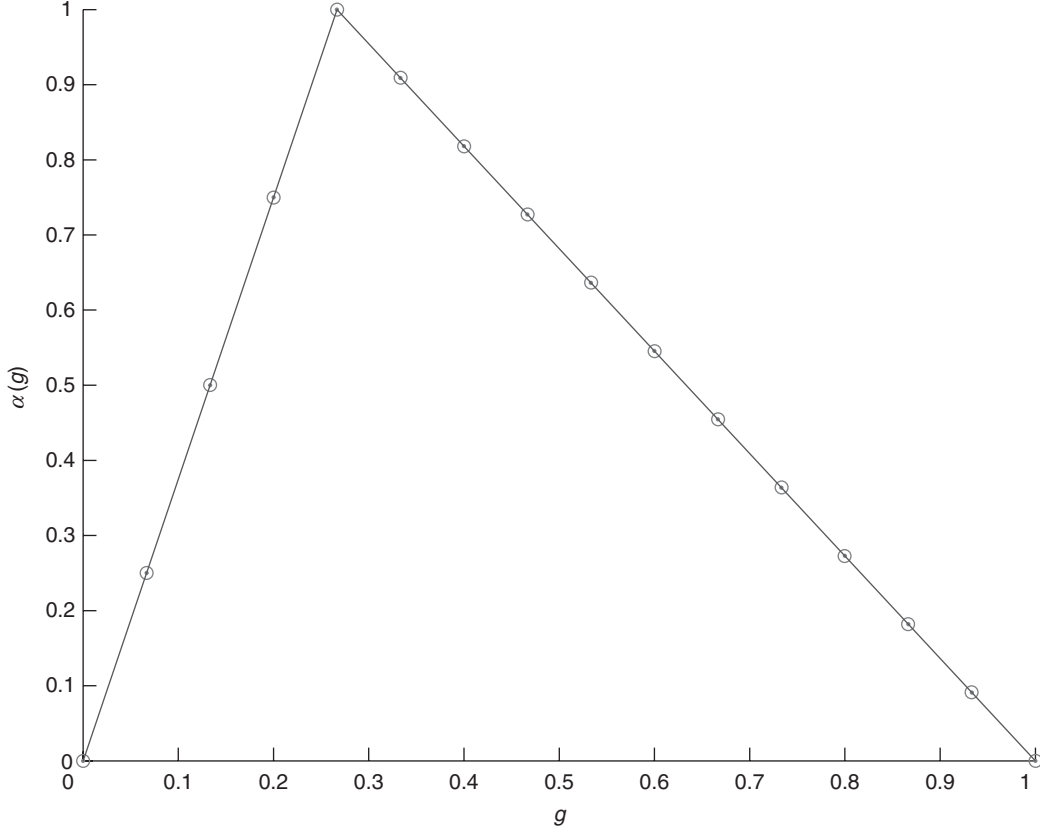


Figure 3. Gomory mixed integer cut.

Valid Inequalities. In order to describe a valid inequality, we must associate a coefficient $\alpha(g)$ to each integer variable $z(g)$ and associate a coefficient $\beta(s)$ to each continuous variable $y(s)$. For this reason, valid inequalities are often defined and represented as functions over G and S , respectively (Fig. 3). We then define the notion of valid inequality as one that is satisfied by all solutions of the problem.

Definition 2. Given $r \in I^m \setminus \{0\}$, a subgroup G of I^m and a subset S of S^m , we say that the functions $\alpha : G \rightarrow \mathbb{R}_+$ and $\beta : S \rightarrow \mathbb{R}_+$ define a valid inequality for $\Gamma(G, S, r)$ if $\sum_{g \in G} \alpha(g)z(g) + \sum_{s \in S} \beta(s)y(s) \geq 1$, $\forall (z, y) \in \Gamma(G, S, r)$, where $\alpha(o) = 0$.

Definition 2 does not include the trivial inequalities $z(g) \geq 0$ and $y(s) \geq 0$ as they do

not yield violated cutting-planes for MIPs. It focuses on nontrivial inequalities that can be rescaled to have right-hand sides equal to 1. Definition 2 requires that the coefficients of the variables are nonnegative in any valid inequality, which is without loss of generality when G is finite. Although Definition 2 does not require that $\alpha(r) = 1$, we assume implicitly that $\alpha(r) = 1$ in valid inequalities whenever $r \in G$.

Subadditive Inequalities. The concept of valid inequality does not intrinsically carry a notion of strength. The following concept of subadditive inequality identifies, among valid inequalities, those that support $\Gamma(G, S, r)$. The definition is taken from Ref. 13, although variants appeared earlier in Refs 10–12.

Definition 3. A valid inequality for $\Gamma(G, S, r)$ is said to be a subadditive valid inequality for $\Gamma(G, S, r)$ if it satisfies the following requirements:

1. $\alpha(g_1) + \alpha(g_2) \geq \alpha(g_1 + g_2), \forall g_1, g_2 \in G$
2. $\alpha(g_1) + \sum_{s \in S} \beta(s)y(s) \geq \alpha(g_2), \forall g_1, g_2 \in G$ and $\forall y$ with finite support s.t. $g_1 + \mathcal{F}(\sum_{s \in S} sy(s)) = g_2$;
3. $\sum_{s_1 \neq s_2} \beta(s_1)y(s_1) \geq \beta(s_2), \forall y$ with finite support s.t. $s_2 = \sum_{s_1 \neq s_2} s_1 y(s_1)$ and $\sum_{s_1 \neq s_2} s_1 y(s_2) \in S$.

It is proven in Ref. 11 that valid inequalities for master group problems are dominated by subadditive inequalities. Therefore, when deriving strong inequalities for $\Gamma(G, S, r)$, it is sufficient to study subadditive inequalities. Definition 3 is often cumbersome to verify. However, when dealing with group relaxations with only integer variables, that is, $S = \emptyset$, it reduces to verifying Condition 1. For piecewise linear functions, necessary and sufficient conditions have been developed to help show that Condition 1 holds [18, 22, 24].

Minimal Inequalities. A natural strengthening of the concept of subadditive inequality occurs when requiring that the inequality is not dominated coefficient-wise by another valid inequality. In the context of the group problem, such inequalities are said to be *minimal*. A general characterization of minimal inequalities of $\Gamma(G, S, r)$ is given in Ref. 13 that extends the more specific results given in Refs [10–12]. Since this characterization is technical, we restrict its presentation to two particular cases that are commonly encountered and are simpler to describe. The first is for finite group problems that have only integer variables.

Theorem 1. Given $r \in G \setminus \{o\}$, a finite subgroup G of I^m and a valid inequality α for $\Gamma(G, \emptyset, r)$, α is a minimal valid inequality if and only if

1. $\alpha(g_1) + \alpha(g_2) \geq \alpha(g_1 + g_2), \forall g_1, g_2 \in G$
2. $\alpha(g) + \alpha(r - g) = \alpha(r), \forall g \in G$.

When continuous variables are involved, the coefficients β of continuous variables must be related to the slope of the function α at the origin in the direction of the constraints coefficients of these variables.

Theorem 2. Given $r \in I^m \setminus \{o\}$, a subset S of S^m and an inequality (α, β) for $\Gamma(I^m, S, r)$, the inequality (α, β) is minimal for $\Gamma(I^m, S, r)$ if and only if the following conditions hold:

1. $\alpha(g_1) + \alpha(g_2) \geq \alpha(g_1 + g_2), \forall g_1, g_2 \in I^m$
2. $\beta(s) = \lim_{h \rightarrow 0} \frac{\alpha(\mathcal{F}(hs))}{h}, \forall s \in S$
3. $\alpha(g) + \alpha(r - g) = 1, \forall g \in I^m$.

When an inequality is known to be subadditive, verifying that it is minimal reduces to establishing that it satisfies Conditions 2 and 3. This verification is typically direct.

Extreme Inequalities. Although a minimal inequality is not dominated coefficient-wise by any other single valid inequality, it might be dominated coefficient-wise by a convex combination of other valid inequalities. This motivates the following definition.

Definition 4. Given a valid inequality (α, β) for $\Gamma(G, S, r)$, we say that it is extreme for $\Gamma(G, S, r)$ if whenever $\alpha = \frac{1}{2}\alpha^1 + \frac{1}{2}\alpha^2$ and $\beta = \frac{1}{2}\beta^1 + \frac{1}{2}\beta^2$ for some valid inequalities (α^1, β^1) and (α^2, β^2) of $\Gamma(G, S, r)$ then $\alpha = \alpha^1 = \alpha^2$ and $\beta = \beta^1 = \beta^2$.

It is shown in Ref. 13 that extreme valid inequalities for $\Gamma(G, S, r)$ are minimal for $\Gamma(G, S, r)$. Therefore, it is sufficient to search for extreme inequalities among minimal inequalities. We next describe important results regarding extreme inequalities. We differentiate the finite and infinite cases.

When the group G used in $\Gamma(G, S, r)$ is finite, there is an elegant characterization of the extreme inequalities of $\Gamma(G, S, r)$. In this case, the function $\alpha(\cdot)$ can be represented as a vector with $|G|$ entries that we denote as α_g . Only vectors satisfying the conditions of Theorem 1 correspond to minimal inequalities. Among them, only those that are not convex combinations of one another can be extreme. This leads to the following theorem of Gomory [10].

Theorem 3. Given $r \in G \setminus \{o\}$ where G be a finite subgroup of I^m , the extreme inequalities of $\Gamma(G, \emptyset, r)$ are $z_g \geq 0$ for $g \in G$ and $\sum_{g \in G} \alpha_g z_g \geq 1$ where $\alpha \in \mathbb{R}_+^{|G|}$ are the extreme points of $\Gamma^*(G, \emptyset, r)$, where

$$\Gamma^*(G, \emptyset, r) = \left\{ \alpha \in \mathbb{R}_+^{|G|} \left| \begin{array}{ll} \alpha_{g_1} + \alpha_{g_2} \geq \alpha_{g_1+g_2}, & \forall g_1, g_2 \in G \\ \alpha_g + \alpha_{r-g} = \alpha_r, & \forall g \in G \\ \alpha_r = 1, \\ \alpha_g \geq 0, & \forall g \in G \end{array} \right. \right\}.$$

Proving that an inequality α is extreme using Theorem 3 is conceptually simple as it reduces to proving that the vector of its coefficients yields an extreme point of $P^*(G, \emptyset, r)$. Given an extreme inequality of a master corner relaxation, one can create a valid inequality for its initial corner relaxation by setting all the coefficients of introduced variables to zero. Gomory [10] shows that all extreme inequalities for a corner relaxation can be obtained in this way.

Theorem 4. All extreme inequalities of a corner relaxation can be obtained from the extreme inequalities of its finite master group relaxations.

When the group G used in $\Gamma(G, S, r)$ is infinite, the situation is more complex. The first proof that a given inequality is extreme was described in Ref. 12 for piecewise linear subadditive inequalities with two slopes. Streamlined arguments were proposed in a slightly different framework in Ref. 17. Variants can also be found in Refs 24 and 34. These proofs are typically constructed using contradiction as follows. A subadditive minimal inequality α is considered and is assumed to not be extreme. In this case, $\alpha = \frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2$ where $\alpha_1 \neq \alpha_2$ and α_1, α_2 are minimal valid inequalities for $\Gamma(I^m, \emptyset, r)$. The “equality” set $E(\alpha) = \{(g_1, g_2) \in I^m \times I^m \mid \alpha(g_1) + \alpha(g_2) = \alpha(g_1 + g_2)\}$ is then constructed. Since α_1 and α_2 are minimal functions, they are subadditive. Thus $E(\alpha_1) \supseteq E(\alpha)$ and $E(\alpha_2) \supseteq E(\alpha)$. The final contradiction is then obtained by showing that if $\alpha_1(g_1) + \alpha_1(g_2) = \alpha_1(g_1 + g_2)$ for all $g_1, g_2 \in E(\alpha)$, then $\alpha_1 = \alpha$. The proof that this equality holds is typically done by considering various intervals over which

a type of result called *Interval Lemma* is invoked [17,24,34]. We refer the reader to Ref. 1 for a more detailed description of the approach. This approach is well suited for proving that piecewise linear subadditive inequalities are extreme. We mention, however, that not all extreme inequalities of infinite group problems are piecewise linear [35].

Extreme inequalities of group relaxations can be applied to simplex tableaux in the manner illustrated in the section titled “Introduction.” Because extreme inequalities for one-dimensional group problems are easiest to describe, all but one row of the simplex tableau are typically relaxed before a group cut is generated. Among extreme inequalities for one-dimensional group problems, the GMIC is particularly significant as it is probably the simplest function that satisfies the conditions of Theorems 1 and 2 and is also the most important computationally as seen in the section titled Computational Potential of Group Cuts.

Theorem 5. Let G be a subgroup of I^1 . The GMIC $\alpha(g) = \min\{\frac{g}{r}, \frac{1-g}{1-r}\}$ is extreme for $\Gamma(G, \emptyset, r)$.

In Fig. 3, we represent the GMIC for C^{15} and I^1 when $r = \frac{4}{15}$. In this figure, we represent the elements g of the groups C^{15} and I^1 on the x axis. The values of $\alpha(g)$ are represented on the y axis with dots for C^{15} and with a solid line for I^1 .

OBTAINING STRONG/EXTREME INEQUALITIES FOR GROUP PROBLEMS

We now address the problem of deriving strong cuts for group problems. In the section titled “Relations between Extreme Inequalities,” we describe how to use the extreme inequalities of a group problem to generate extreme inequalities for other group problems. In the section titled “Constituent Functions,” we present methods to generate extreme inequalities *ab nihilo*. These inequalities are the building blocks used by other procedures.

Relations between Extreme Inequalities

We now discuss relations between the extreme inequalities of various group problems.

From Infinite Group Problems. We assume in this section that we know an extreme inequality α for $\Gamma(I^m, \emptyset, r)$. When $m = 1$, inequality α can typically be used to create extreme inequalities for finite group problems as follows [11,12].

Theorem 6. *Given an extreme function $\alpha : I^1 \rightarrow \mathbb{R}_+$ for $\Gamma(I^1, \emptyset, r)$ that is continuous and piecewise linear, and a finite cyclic subgroup G of I^m where G contains all the points at which α is nondifferentiable, the function obtained by restricting α to G is extreme for $\Gamma(G, \emptyset, r)$.*

Strong inequalities for $\Gamma(I^m, \emptyset, r)$ can also be used to generate strong inequalities for other infinite group problems as the following results suggest [13,24].

Theorem 7. *Given a surjective map $\varsigma : G_1 \rightarrow G_2$ that is a homomorphism, the function α is subadditive and minimal for $\Gamma(G_1, \emptyset, r)$ if and only if the function $\alpha(\varsigma(x))$ is subadditive and minimal for $\Gamma(G_2, \emptyset, v)$, for any v such that $\varsigma(v) = r$.*

For particular types of homomorphisms, Theorem 7 can be strengthened. Given $n_i \neq 0$ for $i = 1, \dots, m$, we define the multiplicative homomorphism $\lambda : I^m \rightarrow I^m$ as $\lambda(x) = (n_1x_1, n_2x_2, \dots, n_mx_m)$.

Theorem 8. *Given a multiplicative homomorphism λ , the function α is extreme for $\Gamma(I^m, \emptyset, r)$ if and only if $\alpha \circ \lambda$ is extreme for $\Gamma(I^m, \emptyset, v)$, where $\lambda(v) = r$.*

Theorem 7 can also be strengthened if the homomorphism ς is an automorphism [13].

Theorem 9. *Given an automorphism ω of I^m and an extreme inequality α of $\Gamma(I^m, \emptyset, r)$, the function $\alpha \circ \omega$ is extreme for $\Gamma(I^m, \emptyset, \omega^{-1}(r))$.*

Combination procedures can also be used to generate extreme inequalities. In particular, one can “aggregate” the rows of a group problem into a single constraint before applying a one-dimensional group cut. More precisely, let $\alpha : I^1 \rightarrow \mathbb{R}_+$ be a valid inequality for $\Gamma(I^1, \emptyset, r)$ and let $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{Z}^m$ be weights chosen so that $\lambda \neq 0$. The aggregation function $\alpha^\lambda : I^m \rightarrow \mathbb{R}_+$ is defined as $\alpha^\lambda(x) = \alpha(\sum_{i=1}^m \lambda_i x_i)$. When the inequality α is known to be extreme for $\Gamma(I^1, \emptyset, r)$, $\alpha^\lambda(x)$ can be shown to be extreme in a variety of contexts [24,34,36].

Another such combination procedure is *sequential-merge*. This procedure generates an inequality α^1 for $\Gamma(I^{m+1}, \emptyset, r)$ by combining an inequality α^2 for $\Gamma(I^m, \emptyset, r)$ with an inequality α^3 for $\Gamma(I^1, \emptyset, r)$. Intuitively, the procedure applies α^2 to the first m rows of $\Gamma(I^{m+1}, \emptyset, r)$, aggregates the resulting inequality with the last row of $\Gamma(I^{m+1}, \emptyset, r)$, and applies α^3 to the aggregate constraint. Under some mild conditions [23,37], α^1 can be shown to be extreme for $\Gamma(I^{m+1}, \emptyset, r)$ if α^2 and α^3 are extreme for $\Gamma(I^m, \emptyset, r)$ and $\Gamma(I^1, \emptyset, r)$, respectively.

From Finite Group Problems. We assume in this section that we know an extreme inequality α for $\Gamma(G, S, r)$, where G is a finite group. Theorems 10 and 11 are due to Gomory [10] and relate extreme inequalities of finite group problems.

Theorem 10. *Given two finite subgroups G^1 and G^2 of I^m , $r \in G^1 \setminus \{o\}$ and a surjective homomorphism $\varsigma : G^1 \rightarrow G^2$, the inequality $\alpha \circ \varsigma$ is extreme for $\Gamma(G^1, \emptyset, r)$ whenever $\alpha : G^2 \rightarrow \mathbb{R}_+$ is an extreme inequality for $\Gamma(G^2, \emptyset, \varsigma(r))$ and $r \notin \text{Kern}(\varsigma)$.*

Theorem 11. *Given an automorphism $\omega : G \rightarrow G$, the inequality ϕ is extreme for $\Gamma(G, \emptyset, r)$ if and only if $\phi \circ \omega$ is extreme for $\Gamma(G, \emptyset, \omega^{-1}(r))$.*

Interpolation is a natural idea to create inequalities for infinite group problems from finite group problems. Interpolation, however, does not necessarily create extreme inequalities for $\Gamma(I^m, \emptyset, r)$ even in the

case where $m = 1$, a situation where the interpolation is unique. We refer to Ref. 34 for a discussion. A stronger procedure to generate inequalities for infinite group problems from finite group problems is *fill-in*. Fill-in has its origin in Ref. 12 and is described in its generality in Ref. 13. This procedure is analogous to lifting. Let G be a subgroup of I^m and $S \subseteq S^m$ and assume that S is sufficiently diverse that every vector $s \in \mathbb{R}^m$ can be written as a nonnegative combination of elements of S . Let (α, β) be a valid subadditive function for $\Gamma(G, S, r)$. To obtain a valid subadditive inequality $(\hat{\alpha}, \hat{\beta})$ for $\Gamma(I^m, S^m, r)$, we use the following procedure. First, we construct $\tilde{\beta}(s) = \min \{ \sum_{s' \in L} \beta(s')y(s') \mid s = \sum_{s' \in L} s'y(s'), y(s') \geq 0 \}$ where L is a linearly independent subset of S . Then we construct the function $\hat{\beta} : \mathbb{R}^m \rightarrow \mathbb{R}_+$ as $\hat{\beta}(s) = \lambda \tilde{\beta}(s')$ where $s' \in S^m$, $\lambda \geq 0$, and $s = \lambda s'$. In other words, $\hat{\beta}$ is the homogenous extension of $\tilde{\beta}$ to $\mathbb{R}^m$. Finally, we compute $\hat{\alpha}(g) = \min \{ \alpha(g') + \hat{\beta}(s) \mid g' \in G, s \in S, g' + \mathcal{F}(s) = g \}$.

For group problems with a single constraint, the fill-in procedure produces inequalities that have two slopes and are therefore extreme [12]. For group problems with multiple constraints, however, the inequality can typically only be proven to be subadditive even if an extreme inequality is used as a starting point of the procedure. Sufficient conditions for fill-in inequalities to be extreme are given in Refs 31 and 38.

Constituent Functions

Finite Group Problems. For discrete problems, deriving a violated group-theoretic inequality reduces to solving an LP over Gomory's characterization; see Theorem 3. The solution of this LP might be computationally demanding if the size of the group is large. As a result, families of extreme points of $\Gamma^*(G, \emptyset, r)$, which result in extreme inequalities for finite group relaxations, are typically derived analytically. The inspiration for the derivation of these families inequalities come either from an analytical study of its structure or from shooting experiments [18] where a random objective is optimized over $\Gamma^*(G, \emptyset, r)$ and the pattern of the resulting inequality is

studied and generalized. Such analyses have produced many families of inequalities including two- and three-slope functions [10,39]. A form of lifting called *tilting* has also been used to derive families of extreme inequalities from lower-dimensional faces of finite group problems [39]. Obtaining the extreme inequalities of simple subfamilies of subadditive functions has also been proposed as a way to obtain strong inequalities for group problems in Refs 21 and 22, where the procedure is coined *approximate lifting*.

Infinite Group Problems. Generating families of extreme inequalities for infinite group problems is harder as it is not possible to use shooting experiments. Many known extreme inequalities for infinite group problems are interpolations of extreme inequalities for finite group problems [17,21]. The approximate lifting procedure of Miller *et al.* and Richard *et al.* [21,22] also yields candidate extreme inequalities.

Extreme inequalities for infinite group problems can also be derived from simple sets. In this approach, a simple mixed integer set F with free integer variables and at least a continuous variable is considered. An aggregation scheme is then derived that aggregates the group problem variables into the variables of F . Any strong inequality for F can then be converted into a group-theoretic cut using the previously obtained aggregation scheme. This procedure is successfully used in Refs 19 and 40 and also provides an alternate derivation of GMIC in Ref. 41.

A new direction of research that has attracted much attention over the past few years is the derivation of strong inequalities for *continuous group problems*, that is, problems of the form $\Gamma(\{o\}, S, r)$ for which it is typically not assumed that $S \subseteq S^m$. A likely reason for this interest is that valid inequalities for $\Gamma(\{o\}, S, r)$ have a strong geometrical grounding. To describe it, we define a full-dimensional convex set P in $\mathbb{R}^m$ to be a lattice-free convex set if $\text{int}(P) \cap Z^m = \emptyset$. Further, we assume that P is a maximal lattice-free set if there does not exist another lattice-free convex set P' such that $P' \neq P$ and $P' \supseteq P$. It is shown in Ref. 29 that nontrivial minimal valid

inequalities for $\Gamma(\{o\}, \mathbb{R}^m, r)$ are equivalent either to $\beta(s) \geq 0$ or to $\sum_{s \in \mathbb{R}^m} \beta(s)y(s) \geq 1$, where the set $P(\beta) = \{u \in \mathbb{R}^m \mid \beta(u+r) \leq 1\}$ is a maximal lattice-free set with $-r$ in its interior. The study of minimal inequalities for these sets therefore reduces to the study of maximal lattice-free convex sets. We refer to Refs 28,30,31,42, and 43 for discussions and mention that valid inequalities for $\Gamma(\{o\}, S, r)$ can be extended to $\Gamma(G, S, r)$ using the fill-in procedure described above [38].

COMPUTATIONAL POTENTIAL OF GROUP CUTS

Various recent studies have tried to gauge the strength of group cuts for MIPs. In particular, Fischetti and Monaci [27] evaluates the strength of corner relaxation on a family of test MIP problems. The authors find that the corner relaxation cuts a substantial fraction of the integrality gap of MIPs, especially for problems with integer variables that are not binary. This study therefore suggests that there is value in studying group-theoretic cuts for MIPs. An important question is therefore to determine which group cuts are most useful in solving MIPs.

Often, group cuts are generated from single rows of simplex tableaux. Among one-dimensional group cuts, GMIC is the most extensively used in commercial computer implementations; see Bixby and Rothberg [44]. Many other cuts, however, have been tested. They include two-step MIR (mixed integer rounding) inequalities [25,45], k -cuts (homomorphisms of GMIC) [46] and interpolated cuts [26]. The overwhelming conclusion of these studies is that GMICs are most effective among one-dimensional group-theoretic cuts. Several families of two-slope functions can also be helpful [25,45]. All other families of cuts seem to have a marginal impact when GMICs are also added [26].

While there might not be many practical reasons to derive further families of cuts for one-dimensional group problems, the investigation of cuts for group problems with multiple constraints is promising, as suggested in Ref. 27. In fact, very few strong

inequalities are known in this case, and multidimensional equivalents of GMIC are yet to be found. Experiments with the continuous group relaxation approach in Ref. 47 give empirical evidence that some multirow group cuts might be helpful in computation. Determining whether the computational potential of these cuts can be harnessed is an active area of research.

REFERENCES

1. Richard J-PP, Dey SS. The group-theoretic approach in mixed integer programming. In: Jünger M, Liebling TM, Naddef D, *et al.* editors. 50 Years of Integer Programming 1958–2008. Berlin: Springer; 2009. pp. 727–801.
2. Gomory RE. Outline of an algorithm for integer solutions to linear programs. Bull Am Math Soc 1958;64:275–278.
3. Gomory RE. An algorithm for integer solutions to linear programs. In: Graves RL, Wolfe P, editors. Recent advances in mathematical programming. New York: McGraw-Hill Book Company Inc.; 1963. pp. 269–302.
4. Gomory RE. An algorithm for the mixed integer problem. Technical Report RM-2597. Santa Monica (CA): RAND Corporation; 1960.
5. Gilmore PC, Gomory RE. A linear programming approach to the cutting-stock problem. Oper Res 1961;9:849–859.
6. Gilmore PC, Gomory RE. A linear programming approach to the cutting-stock problem: Part ii. Oper Res 1963;11:863–888.
7. Gilmore PC, Gomory RE. Multistage cutting stock problems of two and more dimensions. Oper Res 1965;13:94–120.
8. Gomory RE. Some relation between integer and non-integer solutions of linear program. Proc Natl Acad Sci U S A 1965;53:250–265.
9. Gomory RE. Faces of an integer polyhedron. Proc Natl Acad Sci U S A 1967;57:16–18.
10. Gomory RE. Some polyhedra related to combinatorial problems. Linear Algebra Appl 1969;2:341–375.
11. Gomory RE, Johnson EL. Some continuous functions related to corner polyhedra, part I. Math Program 1972;3:23–85.
12. Gomory RE, Johnson EL. Some continuous functions related to corner polyhedra, part II. Mathe Program 1972;3:359–389.

13. Johnson EL. On the group problem for mixed integer programming. *Math Program Study* 1974;2:137–179.
14. Johnson EL. Characterization of facets for multiple right-hand side choice linear programs. *Math Program Study* 1981;14: 112–142.
15. Johnson EL. *Integer programming—facets, subadditivity and duality for group and semi-group problems*. Philadelphia (PA): SIAM Publications; 1980.
16. Balas E, Ceria S, Cornuéjols G, *et al.* Gomory cuts revisited. *Oper Res Lett* 1996;19:1–9.
17. Gomory RE, Johnson EL. T-space and cutting planes. *Math Program* 2003;96:341–375.
18. Gomory RE, Johnson EL, Evans L. Corner polyhedra and their connection with cutting planes. *Math Program* 2003;96:321–339.
19. Dash S, Günlük O. Valid inequalities based on simple mixed-integer sets. *Math Program* 2006;106:29–53.
20. Dash S, Günlük O. Valid inequalities based on the interpolation procedure. *Math Program* 2006;106:111–136.
21. Miller LA, Li Y, Richard J-PP. New facets for finite and infinite group problems from approximate lifting. *Nav Res Log* 2008;55: 172–191.
22. Richard J-PP, Li Y, Miller LA. Valid inequalities for MIPs and group polyhedra from approximate liftings. *Math Program* 2009;118:253–277.
23. Dey SS, Richard J-PP. Relations between facets of low- and high-dimensional group problems. *Math Program* 2010;123:285–313.
24. Dey SS, Richard J-PP. Facets of the two-dimensional infinite group problems. *Math Oper Res* 2008;33:140–166.
25. Dash S, Günlük O. On the strength of Gomory mixed integer cuts as group cuts. *Math Program* 2008;115:387–407.
26. Fischetti M, Saturni C. Mixed integer cuts from cyclic groups. *Math Program* 2007;109:27–53.
27. Fischetti M, Monaci M. How tight is the corner relaxation? *Discrete Optim* 2008;5:262–269.
28. Andersen K, Louveaux Q, Weismantel R, *et al.* Inequalities from two rows of a simplex tableau. In: Fischetti M, Williamson DP, editors. *Proceedings 12th Conference on Integer and Combinatorial Optimization*. Berlin: Springer; 2007. pp. 30–42.
29. Basu A, Conforti M, Cornuéjols G, *et al.* Minimal inequalities for an infinite relaxation of integer programs. *SIAM J Discrete Math* 2010;24:158–168.
30. Cornuéjols G, Margot F. On the facets of mixed integer programs with two integer variables and two constraints. *Math Program* 2009;120:429–456.
31. Dey SS, Wolsey LA. Two row mixed integer cuts via lifting. *Math Program* 2010;124: 143–174.
32. Dey SS, Wolsey LA. *Constrained infinite group relaxations of MIPs*. Technical Report CORE DP 33. Belgium: Université catholique de Louvain, Louvain-la-Neuve; 2009.
33. Fukasawa R, Günlük O. Strengthening lattice-free cuts using non-negativity; 2009. Available at http://www.optimization-online.org/DB_HTML/2009/05/2296.html. Accessed 2009.
34. Dey SS, Richard J-PP, Li Y, *et al.* Extreme inequalities for infinite group problems. *Math Program* 2010;121:145–170.
35. Basu A, Conforti M, Cornuéjols G, *et al.* A counterexample to a conjecture of Gomory and Johnson. *Math Program*. In press.
36. Dey SS. *Strong cutting planes for unstructured mixed integer programs using multiple constraints* [PhD thesis]. West Lafayette (IN): Purdue University; 2007.
37. Dey SS, Richard J-PP. Sequential-merge facets for two-dimensional group problems. In: Fischetti M, Williamson DP, editors. *Proceedings 12th Conference on Integer and Combinatorial Optimization*. Berlin: Springer; 2007. pp. 30–42.
38. Dey SS, Wolsey LA. Lifting integer variables in minimal inequalities corresponding to lattice-free triangles. In: Lodi A, Panconesi A, Rinaldi G, editors. *Proceedings 13th Conference on Integer and Combinatorial Optimization*. Berlin: Springer; 2008. pp. 463–475.
39. Aráoz J, Evans L, Gomory RE, *et al.* Cyclic groups and knapsack facets. *Math Program* 2003;96:377–408.
40. Kianfar K, Fathi Y. Generalized mixed integer rounding valid inequalities: facets for infinite group polyhedra. *Math Program* 2009;120:313–346.
41. Marchand H, Wolsey LA. Aggregation and mixed integer rounding to solve MIPs. *Oper Res* 2001;49:363–371.
42. Borozan V, Cornuéjols G. Minimal inequalities for integer constraints. *Math Oper Res* 2009;34:538–546.
43. Zambelli G. On degenerate multi-row Gomory cuts. *Oper Res Lett* 2009;37:21–22.

- 44. Bixby RE, Rothberg EE. Progress in computational mixed integer programming - a look back from the other side of the tipping point. *Ann Oper Res* 2007;149:37–41.
- 45. Dash S, Goycoolea M, Günlük O. Two step MIR inequalities for mixed integer programs. *INFORMS J Comput* 2010;22:236–249.
- 46. Cornuéjols G, Li Y, Vandenbussche D. K-cuts: a variation of gomory mixed integer cuts from the LP tableau. *INFORMS J Comput* 2003;15:385–396.
- 47. Espinoza D. Computing with multiple-row Gomory cuts. In: Lodi A, Panconesi A, Rinaldi G, editors. *Proceedings 13th Conference on Integer Programming and Combinatorial Optimization*. Berlin: Springer; 2008. pp. 214–224.

INFINITE HORIZON PROBLEMS

ARCHIS GHATE
University of Washington,
Seattle

INTRODUCTION

A typical discrete-time sequential decision problem involves a system whose state is assumed to evolve either deterministically or probabilistically over time-periods that are often called *stages*. This evolution is affected by the decisions a planner makes at the beginning of each stage after observing the system state. The decision maker's goal then is to optimize some measure of system performance over a certain time-horizon. For example, in a typical production and inventory management problem, the system state is given by the inventory on hand at the beginning of a stage, and the decision corresponds to the production level in that stage. The inventory beginning the next stage equals the old inventory plus the production quantity minus the stochastic demand filled. Unsatisfied demand may be lost. The planner's goal may be to maximize the expected total discounted profit, where revenue is generated by selling the product, and costs are incurred for production, inventory holding, and shortage.

The dynamic systems in such optimization problems often do not have a predetermined time of extinction. Thus, using a finite planning horizon typically introduces end-of-study effects on early decisions. Indeed, a finite horizon formulation of a production planning problem essentially amounts to assuming that the demand in all subsequent time-periods after this horizon is zero. Then the decision maker is likely to plan initial production such that the inventory ending this finite horizon is zero, forcing him/her to produce additional units later when the actual demand beyond the initial study horizon is

revealed. In fact, if the production costs in these later periods turn out to be high, the decision maker may realize rather late that it would have been optimal to produce more earlier on and carry inventory forward to later periods. More generally, finite horizon formulations of dynamic optimization problems can lead to short-sighted decisions, and therefore, infinite horizon versions of such problems are studied extensively.

The literature on infinite horizon optimization is vast and encompasses diverse fields including electrical engineering, especially control systems; economics/finance; operations research/management science; statistics; and mathematics. It spans both discrete as well as continuous-time problems, and includes research where the states and decisions may belong to abstract infinite-dimensional spaces. Here, we focus only on discrete-time models under the assumption that both the state and decision sets are finite because these appear to be more common in the operations research and management science literature (see Refs 1 and 2 for continuous-time and/or metric space extensions). Within this context, there are two somewhat distinct bodies of research—stationary models and nonstationary models.

A majority of the work assumes that the data do not change over time, leading to *stationary* infinite horizon problems. More specifically, stationary infinite horizon Markov decision problems (MDPs) with the expected total discounted reward criterion are the most studied and best understood among such models. Extensive theoretical and algorithmic treatments of such models along with their diverse applications as well as discussions of MDPs with other optimality criteria, and their continuous-time and semi-Markov extensions can be found in Refs 3–7; these will be reviewed in detail elsewhere.

A considerable amount of research on the other hand focuses on *nonstationary* infinite horizon problems. Here, we are

interested in truly nonstationary problems in contrast to problems in which data trends over time can be completely specified with a finite set of parameters. These problems present notoriously difficult challenges that are circumvented by (the time-homogeneity) assumption in stationary problems. While computing an optimal policy in a stationary infinite horizon MDP can be reduced to a finite-dimensional problem through Bellman's equations, a nonstationary problem is necessarily infinite-dimensional. Even in a deterministic nonstationary problem, computing only a first period decision that is infinite horizon optimal may require knowledge of an *infinite* number of data parameters, for example, a reward function for every time-period, rendering the task impossible. The question arises as to whether it is possible to forecast data parameters for only a finite number of periods N , and compute an N horizon optimal (with respect to these specific N period data parameters) first period decision that remains optimal for all infinite horizon problems no matter what future data parameters for periods $N + 1$ and beyond are revealed. A finite horizon with this property is called a *forecast horizon*. In short, when data parameters beyond period N are irrelevant to computing an infinite horizon optimal first period decision, N is a forecast horizon.

In the fortunate case where such a forecast horizon exists and can be discovered using a finite procedure, we could implement the corresponding first period decision and then roll forward to construct a new infinite horizon problem with a new initial state. The process could then be repeated recursively ad infinitum to deliver an infinite horizon optimal strategy. This rolling (or receding) horizon procedure renders the nonstationary infinite horizon problem "solvable." When such horizons fail to exist, it may be possible to obtain horizon-dependent bounds on the error induced by solving finite horizon truncations of the infinite horizon problem [8–10]. These ideas go at least as far back as early work on production planning problems in the 1950s [11–14]. Since then, a range of theoretical, algorithmic, and application-specific papers

in areas such as production planning and inventory management [15–28], equipment replacement [29–31], capacity expansion [32–35], research and development planning [36], and cash management [37] have appeared in the infinite horizon optimization literature. An extensive review of these works with a detailed classified bibliography of over two hundred articles is available in Ref. 38. Below we outline some of the main ideas in this field. We focus on discounted deterministic problems and only briefly mention extensions to their stochastic counterparts.

CHALLENGES IN NONSTATIONARY INFINITE HORIZON OPTIMIZATION

Owing to infinite data requirements, it is not in general possible even to completely specify an instance of a truly nonstationary infinite horizon problem. We must instead speak of a *class* Φ of nonstationary infinite horizon problems characterized by certain structural properties of their data parameters. For example, Φ could be the class of all discounted deterministic cost minimization production planning problems with stationary discount factors strictly less than one and the following cost and demand data properties: (i) linear production and inventory holding costs, (ii) strictly positive integer-valued unit production and unit inventory costs bounded above by given positive integers c and h respectively, and (iii) strictly positive integers-valued demands bounded above by a given positive integer d . For $N \geq 1$, let ϕ_N denote the data parameters for the first N periods from a forecast ϕ in Φ . For instance, in the above class of production planning problems, ϕ_N corresponds to specific values of discount factor $0 < \alpha < 1$, integer-valued unit production costs $0 < c_1, c_2, \dots, c_N \leq c$, integer-valued unit inventory holding costs $0 < h_1, h_2, \dots, h_N \leq h$, and integer demands $0 < d_1, d_2, \dots, d_N \leq d$. Let $\phi(N)$ be the set of all forecasts θ in Φ whose data parameters for the first N periods match that of ϕ , that is, $\theta_N = \phi_N$.

The idea of a forecast horizon can then be made more precise as follows: N is a *forecast*

horizon for ϕ if there exists a first period decision optimal to ϕ_N that remains optimal to all infinite horizon forecasts θ in $\Phi(N)$. According to this definition, the discovery that N is a forecast horizon can be made based on the knowledge of specific values of data parameters in the first N periods as well as on properties of class Φ ; it must however be made *without* any reference to ϕ . Such a forecast horizon could be viewed as a forecast horizon for all forecasts in $\Phi(N)$.

The weaker notion of a solution horizon arises when the class Φ is a singleton forecast ϕ : N is a *solution horizon* if a first period decision optimal to ϕ_N is also optimal to ϕ . Note here that whether N is a solution horizon for ϕ or not may depend on data parameters in ϕ in periods $N + 1$ and beyond, whereas this is not allowed in the definition of a forecast horizon.

A stronger notion is that of a forecast horizon for class Φ : N^* is a *forecast horizon for class Φ* if it is a forecast horizon for *every* ϕ in Φ . Hence, the conclusion that N^* is a forecast horizon for class Φ must be drawn based solely on the defining properties of Φ without reference to any specific data parameter values. Finally, if N^* is a forecast horizon for class Φ , then so are all horizons longer than N^* . Therefore, it is of some interest to find the *minimum* forecast horizon owing to the computational difficulties and costs involved in forecasting data parameters for the distant future. The above ideas also specialize to finite horizon problems [38]. See Refs 38–41 for complete formal details on forecast and solution horizons.

Five main sources of these horizons were identified in the classified bibliography in Ref. 38. These are (i) cost structure, (ii) constraints, (iii) discounting, (iv) turnpike properties, and (v) ergodicity. For instance, the relative value of fixed costs in relation to holding costs is an important factor in existence of forecast horizons in the dynamic lot-size model of Wagner and Whitin [14]. Nonnegativity and warehouse capacity constraints on inventory values lead to forecast horizons in the wheat-trading model of [42]. The turnpike property, well studied in the economics literature for a long time [43], can also lead to solution horizons [28,44,45]. This arises when

the problem horizon is sufficiently long with given initial and terminal states and an optimal trajectory quickly passes from the initial state to a steady-state, remains there for a long time, ultimately returning to the final state as late as possible. This is typical in maximization problems with concave objective functions and state dynamics functions. Chand *et al.* [38] comment that unfortunately researchers often do not explicitly identify the precise source of forecast and solution horizons as they result from complex interactions between several problem characteristics.

A portion of the literature over the years has attempted to establish general conditions for existence of solution and forecast horizons and on developing algorithms for their discovery. It is often easy to establish that for any forecast ϕ in class Φ , the finite horizon optimal values $v^*(\phi_N)$ converge to the infinite horizon optimal value $v^*(\phi)$ as $N \rightarrow \infty$. Uniqueness of an infinite horizon optimal solution $x^*(\phi)$ to a forecast ϕ is then sufficient for solution and forecast horizons for forecast ϕ to exist. See Refs 36, 41, 46–49 for detailed formal discussions of sufficiency of unique infinite horizon optima. Unfortunately, uniqueness is often impossible to verify and is especially rare in discrete optimization problems [1,49,50]. Researchers have thus attempted to instead establish convergence of sets of finite horizon optima to the set of infinite horizon optima and then have applied tie-breaking procedures to force solution convergence [1,49]. Here, we discuss an alternative approach that circumvents this issue for a class of problems where optimal first period decisions are monotone in horizon length—a property succinctly described as policy monotonicity.

EXPLOITING POLICY MONOTONICITY

We illustrate the main ideas in this type of work through a production planning example taken from Ref. 27. Let Φ be the class of all infinite horizon production planning problems with the following defining characteristics:

1. Non-negative integer demands.

2. *Convex* non-negative production cost functions where
 - the cost of not producing in a period is zero,
 - the marginal cost of production is at least $c > 0$,
 - the marginal cost of production is at most γ .
3. *Convex* nonnegative inventory holding cost functions where
 - the cost of not holding inventory is zero,
 - the marginal cost of holding is at least $h > 0$.
4. Discount factors fixed at $0 < \alpha < 1$.
5. Nonnegative initial inventory.
6. There exists at least one production-inventory schedule with finite total discounted infinite horizon cost.

Any specific forecast ϕ from this class results in the following infinite horizon production planning problem over time-periods $n = 1, 2, \dots$

$$\begin{aligned}
 (P^\phi) \quad & \min \sum_{n=1}^{\infty} \alpha^{n-1} [c_n^\phi(x_n) + h_n^\phi(i_n)] \\
 & i_0 \text{ given} \\
 & i_{n-1} + x_n - d_n^\phi = i_n, \quad n = 1, 2, \dots \\
 & x_n \geq 0, \quad n = 1, 2, \dots \\
 & i_n \geq 0, \quad n = 1, 2, \dots \\
 & x_n \text{ integer}, \quad n = 1, 2, \dots \\
 & i_n \text{ integer}, \quad n = 1, 2, \dots
 \end{aligned}$$

Here, $c_n^\phi(\cdot)$ and $h_n^\phi(\cdot)$ are production and inventory cost functions, $d_1^\phi, d_2^\phi, \dots$ is the demand stream, and i_0^ϕ is the initial inventory on hand, all specific to forecast ϕ in Φ . The production schedule is denoted $x_1, x_2, \dots$ and given the initial inventory and the demand stream, it induces the inventory schedule $i_1, i_2, \dots$. Note however that for a truly nonstationary problem, we can never entirely specify ϕ and therefore problem (P^ϕ) is only an abstract entity. We instead focus on its finite-horizon truncations

$(P^{\phi N})$:

$$\begin{aligned}
 (P^{\phi N}) \quad & \min \sum_{n=1}^N \alpha^{n-1} [c_n^\phi(x_n) + h_n^\phi(i_n)] \\
 & i_0 \text{ given} \\
 & i_{n-1} + x_n - d_n^\phi = i_n, \quad n = 1, 2, \dots, N \\
 & x_n \geq 0, \quad n = 1, 2, \dots, N \\
 & i_n \geq 0, \quad n = 1, 2, \dots, N \\
 & x_n \text{ integer}, \quad n = 1, 2, \dots, N \\
 & i_n \text{ integer}, \quad n = 1, 2, \dots, N.
 \end{aligned}$$

A well-known monotonicity result on convex production planning problems [51] provides that an optimal response to a unit increase in one of the demands is to increase optimal production by one unit. More formally,

Lemma 1 [51]. *Suppose $x_1^*, x_2^*, \dots, x_N^*$ is an optimal production schedule for demands $d_1, d_2, \dots, d_N$. If one of these demands is increased by one unit, it is optimal to increase one of these production levels by one unit.*

A special case of this monotonicity result that is presented in Ref. 4 was employed in Ref. 27 to first prove monotonicity of optimal production levels in horizon length and then establish the existence of forecast horizons for class Φ . Specifically, they embedded an N horizon problem $(P^{\phi N})$ in an $N+1$ horizon problem by setting demand in the $N+1$ st period to zero. Because it is optimal to end period N with zero inventory in the N horizon problem, an optimal production schedule for $(P^{\phi N})$, when appended with zero production in period $N+1$, results in an optimal production schedule for the embedded $N+1$ horizon problem. Then, if we increase the demand in the $N+1$ st period to its actual value d_{N+1}^ϕ , one unit at a time, it is optimal to increase production levels unit by unit. This leads to the conclusion that *increasing the problem horizon does not decrease optimal production levels*.

The idea of vanishing tail costs in discounted problems and the fact that optimal production levels remain bounded (owing to

strictly positive marginal production costs) were then used in Ref. 27 to prove value convergence of finite horizon truncations: optimal finite horizon costs denoted $C^*(\phi_N)$ converge to the optimal infinite horizon cost denoted $C^*(\phi)$ as $N \rightarrow \infty$. Finite horizon optimal production levels $x_n^*(\phi_N)$ incrementally constructed using the above procedure then converge monotonically upward to an infinite horizon optimal production level $x_n^*(\phi)$ for every period n as $N \rightarrow \infty$ *without requiring uniqueness of infinite horizon optimal production schedules*. In fact, because production levels are integers, convergence implies that for some horizon $N^*(\phi)$ that is long enough,

$$x_1^*(\phi_N) = x_1^*(\phi) \text{ for all } N \geq N^*(\phi). \quad (1)$$

Thus $N^*(\phi)$ is a solution horizon for ϕ (see Ref. 27 for proofs and other details).

It is even more interesting that monotonicity leads to *closed form formulas* for forecast horizons. For instance, note in the above production planning situation that when the horizon increases from N to $N+1$, optimal first period production either increases or remains the same. If the first period production *must* increase, then the horizon cannot be longer than a certain length. To see this, observe that if the first period production increases as a result of an increase in horizon length, this extra production in the first period must be carried all the way to period $N+1$ (because optimal production levels in intermediate periods are nondecreasing in horizon). This implies that carrying inventory forward from the first period to the $N+1$ st period must be cheaper than producing in the $N+1$ st period. In particular, the least expensive way of producing in the first period and carrying inventory forward to period $N+1$ must be cheaper than the most expensive way of producing in period $N+1$. This leads to the following inequality:

$$c + h \frac{1 - \alpha^N}{1 - \alpha} \leq \gamma \alpha^N. \quad (2)$$

In other words, if optimal first period production must increase as a result of the increase

in horizon length from N to $N+1$, then N cannot be more than N^* given by [see formula (29) in Ref. 27]

$$N^* = \left\lceil \frac{(1 - \alpha)c + h}{(1 - \alpha)\gamma + h} \right\rceil, \quad (3)$$

where $\lceil y \rceil$ denotes the smallest integer strictly greater than y . Intuitively, N^* is the longest number of periods over which it may be optimal to carry inventory forward; the first period decisions are thus decoupled from periods beyond N^* . Most importantly, this closed form formula does not depend on any specific data parameter values but rather only on properties of class Φ . That is, N^* in Equation (3) is a forecast horizon for class Φ . More strongly, it was proven in Ref. 21 that this N^* is the minimum forecast horizon for class Φ , that is, it is not possible to provide a shorter forecast horizon without explicitly knowing specific data parameter values. Numerical illustrations in Ref. 27 demonstrate that N^* is often very short, that is, a few days or a few weeks.

The above device of embedding an N horizon problem in an $N+1$ horizon problem and appealing to optimal production level monotonicity in demand to establish optimal production level monotonicity in horizon length may not always work. For example, it fails when backlogging is allowed. This is because even though it may be optimal to end the N horizon problem with zero inventory, it may no longer be optimal to end the N th period of the embedded $N+1$ horizon problem with zero inventory. In particular, it may be optimal to carry negative inventory, that is, to backorder in the N th period of the embedded problem. Analysis similar to the one presented above however was successfully carried out in Ref. 21 for the backlogging case by embedding the N horizon problem in an infinite horizon problem with zero demand in periods $N+1$ and beyond. This was possible because the embedded infinite horizon problem reduces to an $N+M$ horizon problem with zero terminal inventory, where M is an upper bound on the number of successive periods in which inventory may be optimally backordered. Lemma

0 then extends to this $N + M$ horizon problem with backlogging leading to a closed-form forecast horizon formula.

SOME NEW RESULTS AND FUTURE DIRECTIONS

The idea of exploiting policy monotonicity has also been extended to stochastic problems. For example, a closed form formula for a forecast horizon in production planning problems with stochastic demands and linear costs was provided in Ref. 20. A general theory for employing policy monotonicity to algorithmically discover forecast horizons in a class of nonhomogeneous infinite horizon MDPs is presented in Ref. 40. The results in Ref. 40 use classic work on establishing monotone optimal policies in MDPs through supermodular functions [52]. There does not seem to be too much work, however, on identifying diverse applications where the first period decision is monotone in horizon length, and this may be a fruitful direction for research in the near future. The close connection between Bellman's equations for *stationary* MDPs and finite-dimensional linear programs has recently proven quite helpful in approximate solution of large-scale stationary MDPs earlier considered intractable [53]. Unfortunately, the linear programs that arise from *nonstationary* MDPs have a countably infinite number of variables and a countably infinite number of constraints, and hence present several mathematical difficulties [54–56]. However, there has been some recent progress in better understanding algebraic and geometric properties of such infinite-dimensional linear programs [57–63]. These results may provide an interesting foundation for further exploring nonstationary infinite horizon optimization problems.

REFERENCES

1. Schochetman IE, Smith RL. Infinite horizon optimization. *Math Oper Res* 1989;14: 559–574.
2. Sethi SP, Thompson GL. *Optimal control theory: applications to Manag Sci and economics*. 2nd ed. New York: Springer; 2005.
3. Bertsekas DP. Volume 1 and 2, *Dynamic programming and optimal control*. 3rd ed. Nashua (NH): Athena Scientific; 2007.
4. Denardo EV. *Dynamic programming: models and applications*. Mineola (NY): Dover; 2003.
5. Heyman DP, Sobel MJ. *Stochastic models in operations research*. Volume II, *Stochastic optimization*. Mineola, New York: Dover; 2003.
6. Puterman M. *Markov decision processes: discrete stochastic dynamic programming*. New Jersey: John Wiley & Sons, Inc.; 1994.
7. Ross SM. *Introduction to stochastic dynamic programming*. New York: Academic Press; 1983.
8. Alden JM, Smith RL. Rolling horizon procedures in nonhomogeneous markov decision processes. *Oper Res* 1984;40:S183–S194.
9. Bean JC, Birge JR, Smith RL. Aggregation in dynamic programming. *Oper Res* 1987;35:215–220.
10. Lee C, Denardo EV. Rolling planning horizons: error bounds for the dynamic lot size model. *Math Oper Res* 1986;11:423–432.
11. Charnes A, Cooper WW, Mellon B. A model for optimizing production by reference to cost surrogates. *Econometrica* 1955;23:307–323.
12. Johnson SM. Sequential production planning over time at minimum cost. *Manag Sci* 1957;3:435–437.
13. Modigliani F, Hohn FE. Production planning over time and the nature of the expectation and planning horizon. *Econometrica* 1955;23:46–66.
14. Wagner HM, Whitin TM. Dynamic version of the economic lot size model. *Manag Sci* 1958;5:89–96.
15. Bean JC, Smith RL, Yano C. Forecast horizons for the discounted dynamic lot size model allowing speculative motive. *Nav Res Logist* 1987;34:761–774.
16. Blackburn J, Kunreuther HC. Planning horizons for the dynamic lot size model with backlogging. *Manag Sci* 1974;21:251–255.
17. Chand S, Morton TE. Minimal forecast horizon procedures for dynamic lot size models. *Nav Res Logist* 1986;33:111–122.
18. Chand S, Sethi SP, Proth JM. Existence of forecast horizons in undiscounted discrete-time lot size models. *Oper Res* 1990;38:884–892.
19. Chand S, Sethi SP, Sorger G. Forecast horizons in the discounted dynamic lot size model. *Manag Sci* 1992;38:1034–1048.

20. Cheevaprawatdomrong T, Smith RL. Infinite horizon production scheduling in time varying systems under stochastic demand. *Oper Res* 2004;52(1):105–115.
21. Ghate A, Smith RL. Optimal backlogging over an infinite horizon under time-varying convex production and inventory costs. *Manuf Serv Oper Manag* 2009;11(2):362–368.
22. Kunreuther HC, Morton TE. Planning horizons for production smoothing with deterministic demands. *Manag Sci* 1973;20:110–125.
23. Kunreuther HC, Morton TE. General planning horizons for production smoothing with deterministic demands. *Manag Sci* 1974;20:1037–1046.
24. Lee DR, Orr D. Further results on planning horizons in the production smoothing problem. *Manag Sci* 1977;23:490–498.
25. Morton TE. The non-stationary infinite horizon inventory problem. *Manag Sci* 1978;24:1474–1482.
26. Morton TE. Universal planning horizons for generalized convex production scheduling. *Oper Res* 1978;26:1046–1058.
27. Smith RL, Zhang R. Infinite horizon production planning in time varying systems with convex production and inventory costs. *Manag Sci* 1998;44(9):1313–1320.
28. Thompson GL, Sethi SP. Turnpike horizons for production planning. *Manag Sci* 1980;26:229–241.
29. Bean JC, Lohmann J, Smith RL. Equipment replacement under technological change. *Nav Res Logist* 1994;41:117–128.
30. Schochetman IE, Smith RL. Infinite horizon optimality criteria for equipment replacement under technological change. *Oper Res Lett* 2007;35:485–492.
31. Sethi SP, Chand S. Planning horizon procedures in machine replacement models. *Manag Sci* 1979;25:140–151.
32. Bean JC, Smith RL. Optimal capacity expansion over an infinite horizon. *Manag Sci* 1985;31:1523–1532.
33. Ryan SM. Forecast frequency in rolling horizon hedging heuristics for capacity expansion. *Eur J Oper Res* 1998;109:550–558.
34. Smith RL. Planning horizons for the deterministic capacity problem. *Comput Oper Res* 1981;8:209–220.
35. Udaybhanu V, Morton TE. Planning horizons for capacity expansion. *Eur J Oper Res* 1988;34:297–307.
36. Hopp WJ. A sequential model of R and D investment over an unbounded time horizon. *Manag Sci* 1987;33:500–508.
37. Sethi SP. A note on a planning horizon model of cash management. *J Financ Quant Anal* 1971;6:659–665.
38. Chand S, Hsu VN, Sethi SP. Forecast, solution, and rolling horizons in operations management problems: a classified bibliography. *Manuf Serv Oper Manag* 2002;4(1):25–43.
39. Bes C, Sethi SP. Concepts of forecast and decision horizons: applications to dynamic stochastic optimization problems. *Math Oper Res* 1988;13:295–310.
40. Cheevaprawatdomrong T, Schochetman IE, Smith RL, *et al.* Solution and forecast horizons for infinite horizon nonhomogeneous markov decision processes. *Math Oper Res* 2007;32(1):51–72.
41. Hopp WJ. Identifying forecast horizons in nonhomogeneous markov decision processes. *Oper Res* 1989;37(2):339–343.
42. Sethi SP, Thompson GL. Planning and forecast horizons in a simple wheat trading model. In: Fleichtinger G, Kall P, editors. *Operations research in progress*. Boston (MA): Reidel Publishing Company; 1982. pp. 203–214.
43. McKenzie L. Turnpike theory. *Econometrica* 1976;44:841–864.
44. Bylka S. Strong turnpike policies in the single-item capacitated lot-sizing problem with periodical dynamic parameter. *Nav Res Logist* 1997;44:775–790.
45. Shapiro JF. Turnpike planning horizons for a markovian decision model. *Manag Sci* 1968;14:292–300.
46. Bean JC, Smith RL. Conditions for the existence of planning horizons. *Math Oper Res* 1984;9:391–401.
47. Bean JC, Smith RL. Conditions for the existence of solution horizons. *Math Program* 1993;59:215–229.
48. Bhaskaran S, Sethi SP. Conditions of the existence of decision horizons for discounted problems in stochastic environment: a note. *Oper Res Lett* 1985;4:61–65.
49. Ryan SM, Bean JC, Smith RL. A tie-breaking rule for discrete infinite horizon optimization. *Oper Res* 1992;40:S117–S126.
50. Ryan SM, Bean JC. Degeneracy in infinite horizon optimization. *Math Program* 1989;43:305–316.

51. Veinott AF Jr. Production planning with convex costs: a parametric study. *Manag Sci* 1964;10:441–460.
52. Topkis DM. Supermodularity and complementarity. New Jersey: Princeton University Press; 1998.
53. De Farias DP, Roy BV. The linear programming approach to approximate dynamic programming. *Oper Res* 2003;51:850–865.
54. Anderson EJ, Nash P. Linear programming in infinite-dimensional spaces: theory and applications. Chichester, Great Britain: John Wiley & Sons, Ltd.; 1987.
55. Grinold RC. Infinite horizon programs. *Manag Sci* 1971;18:157–170.
56. Grinold RC. Finite horizon approximations of infinite horizon linear programs. *Math Program* 1977;12:1–17.
57. Cross WP, Romeijn HE, Smith RL. Approximating extreme points of infinite dimensional convex sets. *Math Oper Res* 1998;23:433–442.
58. Ghaté A, Sharma D, Smith RL. A shadow simplex method for infinite linear programs. *Oper Res* 2009. DOI: 10.1287/opre.1090.0755.
59. Ghaté A, Smith RL. Characterizing extreme points as basic feasible solutions in infinite linear programs. *Oper Res Lett* 2009;37: 7–10.
60. Hernandez-Lerma O, Lasserre JB. Approximation schemes for infinite linear programs. *SIAM J Optim* 1998;24:1246–1260.
61. Romeijn HE, Smith RL. Shadow prices in infinite-dimensional linear programming. *Math Oper Res* 1998;23:239–256.
62. Romeijn HE, Sharma D, Smith RL. Extreme point solutions for infinite network flow problems. *Networks* 2006;48:209–222.
63. Sharkey TC, Romeijn HE. A simplex algorithm for minimum cost network flow problems in infinite networks. *Networks* 2008;52:14–31.

INFINITE LINEAR PROGRAMS

THOMAS C. SHARKEY
Department of Decision Sciences
and Engineering Systems,
Rensselaer Polytechnic Institute,
Troy, New York

Many important optimization problems that can be formulated as linear programs involve decision making or planning over time. If the horizon of the problem is finite and we are concerned with making decisions at discrete points in time, we obtain a classic finite-dimensional linear program. However, in many cases, there is no natural horizon or we are interested in making decisions for every point in time in a continuous setting. We then obtain a linear program in an infinite-dimensional space or, equivalently, an infinite linear program. Infinite linear programs have many applications in solving classic optimization problems over an infinite horizon including deterministic or stochastic dynamic programs with countable states [1], production planning problems [2–4], equipment replacement problems [5,6], capacity expansion problems [5,7], problems involving exploiting resources [8], and network flow problems [4,9].

The analysis of infinite linear programs and the development of solution methods for them is quite difficult due to the fact that many of the fundamental properties of finite-dimensional linear programming fall apart in an infinite-dimensional space. It is not straightforward to characterize an extreme point of the feasible region of an infinite linear program through a decomposition of basic and nonbasic variables [4,10], which is a building block of the simplex method for linear programs. In fact, an infinite linear program may not even have an extreme point even when all variables are bounded [11]. Weak duality does not hold using the natural dual of an infinite linear program [2], which is analogous to the familiar finite-dimensional linear programming dual problem. Weak

duality may not hold even when the infinite linear program has a single feasible solution [9]. This obstacle can be overcome in a relatively straightforward manner by posing the infinite linear program and its dual in the appropriate infinite-dimensional spaces. However, these spaces are often too restrictive to obtain strong duality results [11,12]. Another complication in extending the simplex method to infinite linear programs is that there can exist an infinite sequence of adjacent extreme points with strictly decreasing objective functions (in a minimization problem) that do not converge to the optimal objective function value [1].

The focus of this article is on infinite linear programs where both the number of variables and the number of constraints are infinite. For an introduction to infinite linear programs where either the number of variables or the number of constraints is finite, that is, semi-infinite linear programs, we refer the interested reader to Ref. 13 or Ref. 14. The remainder of this article is split into two sections on different classes of infinite linear programs. In the section titled “Countable Infinite Linear Programs”, we discuss infinite linear programs where the number of variables and the number of constraints are countable, and, in the section titled “Continuous Linear Programs”, we discuss infinite linear programs where the number of variables and the number of constraints are uncountable, that is, the class of so-called continuous linear programs (CLPs).

COUNTABLE INFINITE LINEAR PROGRAMS

A countable infinite linear program (CILP) is a linear program that has a countably infinite number of variables and constraints. CILPs arise in many of the motivating applications of infinite linear programs such as infinite-horizon deterministic and stochastic dynamic programs with countable states and infinite-horizon production planning problems. More generally, sequential

decision-making problems over time in which the decision points are discrete and there does not exist a natural finite planning horizon can often be formulated as CILPs. The mathematical formulation of a CILP in standard form (1,10) is given by

$$\text{minimize } \sum_{j=1}^{\infty} c_j x_j$$

subject to

$$\begin{aligned} \sum_{j=1}^{\infty} a_{ij} x_j &= b_i \quad \text{for } i = 1, 2, \dots \\ x_j &\geq 0 \quad \text{for } j = 1, 2, \dots \end{aligned} \quad (1)$$

The standard form of a CILP is analogous to the standard form of a traditional finite-dimensional linear program; however, it begins to highlight the difficulties in analyzing infinite linear programs. For example, the infinite sums involved in the CILP may not be well defined (if $x_j = 1$ for all $j = 1, 2, \dots$ is feasible and $c_j = (-1)^j$ for $j = 1, 2, \dots$ then the objective function will not be well defined for this solution). There are many common assumptions placed on CILPs including (i) assuming directly that the infinite sums are well defined, (ii) each constraint contains a finite number of variables (which implies that the constraints are well defined), (iii) there exists nonnegative numbers u_j for $j = 1, \dots$ such that for every feasible solution to the CILP that $x_j \leq u_j$ for $j = 1, 2, \dots$ and $\sum_{j=1}^{\infty} |c_j| u_j < \infty$ (implying that the feasible region is compact and the objective function is well defined and continuous). These assumptions are often naturally satisfied by the motivating applications of CILPs since, typically, we are making a finite set of decisions from a finite set of alternatives at each decision epoch in which the costs of the decisions will be discounted over an infinite horizon. There are many obstacles in the analysis of CILPs and the development of solution methods (e.g., a simplex method) for them. In this section, we will review these obstacles and the methods that have been used to overcome them. We will first discuss an example of a class of CILPs, network flow

problems with a countably infinite number of nodes and arcs [4,9].

In particular, we consider an infinite directed network, $G = (N, A)$, where $N = 1, 2, \dots$ is the set of nodes and $A \subset N \times N$ is the set of arcs. We will let $x \in \mathbb{R}^{|A|}$ denote a vector of flows in this infinite network. The regularity assumption that each node has a finite in- and out degree ensures that the set A is countable. Our infinite-dimensional minimum cost network flow problem can be formulated as

$$\text{minimize } \sum_{(i,j) \in A} c_{ij} x_{ij}$$

subject to

$$\begin{aligned} \sum_{j:(j,i) \in A} x_{ji} - \sum_{j:(i,j) \in A} x_{ij} &= b_i \quad \text{for } i = 1, 2, \dots \\ x_{ij} &\leq u_{ij} \quad \text{for } (i,j) \in A \\ x_{ij} &\geq 0 \quad \text{for } (i,j) \in A, \end{aligned} \quad (2)$$

where $c \in \mathbb{R}^{|A|}$ is a vector of costs, $u \in \mathbb{R}_+^{|A|}$ is a vector of nonnegative upper bounds, and $b \in \mathbb{R}^{|N|}$ is a vector of supplies. This formulation is analogous to the finite-dimensional minimum cost network flow problem [15] and ensures that a sufficient amount of flow leaves supply nodes (those nodes i such that $b_i < 0$) and is delivered to demand nodes (those nodes i such that $b_i > 0$) while respecting the flow bounds on the arcs. This problem can be easily converted to the standard form of a CILP by adding slack variables to the flow-bound constraints. Many of the obstacles that arise in the analysis of CILPs also arise in the analysis of these infinite-dimensional network flow problems.

Finite-Dimensional Truncations

One of the main goals in the study of CILPs is to develop solution methods for them. In order for a solution method to be viable in a real-world application, it needs to provide solutions in a finite amount of time. However, it is clear that we need an infinite amount of information (e.g., the cost parameters of the variables, the coefficients associated with each variable in each constraint, and the constants associated with each constraint)

to even characterize a CILP. Therefore, a fundamental issue in the study of CILPs is to develop methods that perform a finite amount of work using a finite amount of information and provide (partial) solutions to the CILP. Many methods that examine finite-dimensional truncations of the problem for various special classes of CILPs have been developed. These finite-dimensional truncations typically consider the first n variables and the constraints that contain only the first n variables, although we may not consider a truncation for every possible value of n . For many of the motivating applications of CILPs, the truncations are readily available by considering each decision epoch as a truncation. Moreover, any CILP in which each constraint contains only a finite number of variables with nonzero coefficients has a natural set of truncations [1,2]. There has been much attention devoted to examining various classes of infinite-horizon optimization problems through a series of truncations to establish *planning* or *forecast* horizons [6,16–26]. These planning or forecast horizons are finite-dimensional truncations that have the desirable property that the first decision (for sequential decision-making problems) is optimal or near-optimal to the infinite-horizon problem. From a theoretical standpoint, the analysis of finite-dimensional truncations of CILPs is used in developing duality theory for CILPs [2,3] and guaranteeing infinite-dimensional simplex methods converge to the optimal solution value of the CILP [1,9].

Extreme Points of the Feasible Region

The study of the extreme points of CILPs is important since we know that the optimum of a CILP occurs at an extreme point [27] and extensions of the simplex method to infinite-dimensional spaces would traverse the extreme points of the feasible region. It is shown in Ref. 28 that the extreme points of finite-dimensional projections of a CILP converge in the product topology to the set of extreme points of the CILP. The finite-dimensional projection of a CILP onto the space $\mathbb{R}^n$ contains a particular n -dimensional vector if and only if the vector corresponds to the values of the first n variables for some feasible solution of the

CILP. More precisely, the projection of an infinite-dimensional set $S \in \mathbb{R}^\infty$ (e.g., the constraint set of a CILP) onto the space $\mathbb{R}^n$ is defined as $S_n = \{p_n(x) : x \in S\}$ where $p_n(x)$ is the first n entries of the infinite-dimensional vector x . The relationship between the projections of a CILP and appropriately defined truncations of CILPs is important in the development of a Shadow Simplex method for CILPs in Ref. 1.

The characterization of an extreme point of a CILP cannot typically be done through a decomposition of basic and nonbasic variables. As an example, recall that for a finite network flow problem, a solution x is an extreme point if and only if the graph containing all “free” arcs, that is, (i, j) such that x_{ij} is strictly between its bounds, does not contain a cycle [15]. For infinite network flow problems with a countably infinite number of nodes and arcs, if x is an extreme point, then the free arc graph does not contain a cycle but the converse is not necessarily true [4]. Example 1 discusses a problem that illustrates this issue.

Example 1. Consider the infinite network flow problem in Fig. 1 (which is modified from an example in Ref. 4) where node 1 is a supply node that must ship one unit of demand (i.e., $b_1 = -1$) from it and all other nodes $i = 2, 3, \dots$ are transshipment nodes (i.e., $b_i = 0$). The flow bound on each arc is 1. The only two extreme points of this problem are the solution that sends the unit of flow along the “upper” path (nodes $1 - 2 - 4 - \dots$) and the solution that sends the unit of flow along the “lower” path (nodes $1 - 3 - 5 - \dots$). Any solution where the unit of the flow is split between these two paths is not an

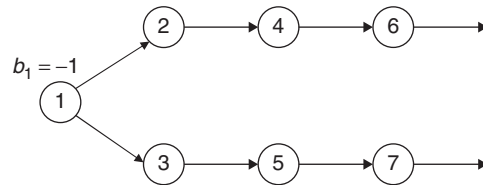


Figure 1. An example problem where an acyclic free arc graph does not imply an extreme point.

extreme point; however, the free arc graph of the solution will be acyclic.

The work in Ref. 4 provides necessary and sufficient conditions for a solution of an infinite network flow problem to be an extreme point of the feasible region. The main issue that limits Ref. 4 from characterizing an extreme point through a decomposition of basic and nonbasic variables is that the support of the free arcs may go to zero, that is, there can exist an edge (i, j) such that x_{ij} is arbitrarily close to its lower or upper bound. The *nonvanishing support* condition developed in Ref. 4 guarantees that an extreme point can be characterized through a decomposition of basic and nonbasic variables. This condition is met by all extreme points of problems with integral capacities and demand. Therefore, all extreme points of problems with integral capacities and demand can be characterized through a decomposition of basic and nonbasic variables. The motivating applications of infinite network flow problems typically have integral capacities and demand, so this condition is not necessarily restrictive. It is shown in Ref. 10 that the nonvanishing support condition guarantees that an extreme point can be decomposed into basic and nonbasic variables for general CILPs.

Duality Theory for Countable Infinite Linear Programs

Duality theory is of critical importance in the study of linear programs. The dual problem provides lower bounds (or upper bounds) on the objective function value of the primal problem and the dual variables have important economic interpretations as the shadow prices of the constraints. There are many examples of duality failing to hold for CILPs [9,11,29] when using the “natural” dual [2] of the CILP, which is analogous to the dual linear program in finite-dimensional linear programming. In fact, weak duality fails to hold even for the simple infinite network flow problem discussed in Example 1.

Example 1 Revisited. (This example is modified from an example in Ref. 9.) The

mathematical formulation of this problem is given by

$$\text{minimize } c_{12}x_{12} + \sum_{i=3}^{\infty} c_{i-2,i} x_{i-2,i}$$

subject to

$$\begin{aligned} -x_{12} - x_{13} &= -1 \\ x_{12} - x_{24} &= 0 \\ x_{i-2,i} - x_{i,i+2} &= 0 \quad \text{for } i = 3, 4, \dots \\ 0 &\leq x_{12} \leq 1 \\ 0 &\leq x_{i-2,i} \leq 1 \quad \text{for } i = 3, 4, \dots \end{aligned} \quad (3)$$

The natural dual of this problem, which is analogous to the dual of a finite-dimensional network flow problem [15], is given as

$$\text{maximize } -\pi_1 - \theta_{12} - \sum_{i=3}^{\infty} \theta_{i-2,i}$$

subject to

$$\begin{aligned} \pi_2 - \pi_1 - \theta_{12} &\leq c_{12} \\ \pi_i - \pi_{i-2} - \theta_{i-2,i} &\leq c_{i-2,i} \quad \text{for } i = 3, 4, \dots \\ \theta_{12} &\geq 0 \\ \theta_{i-2,i} &\geq 0 \quad \text{for } i = 3, 4, \dots \end{aligned} \quad (4)$$

where π_i is the dual multiplier associated with the flow-balance constraint of the i th node and θ_{ij} is the dual multiplier associated with the flow-bound constraint of arc (i, j) . For Problem 3) when all arc costs are zero, the optimal solution to it will have cost 0. Any solution of the form $\pi_i = \Pi$ for $i = 1, 2, \dots$, $\theta_{12} = 0$, and $\theta_{i-2,i} = 0$ for $i = 3, 4, \dots$ will be feasible to the dual problem (Eq. 4). Therefore, for solutions of this form where $\Pi < 0$, the objective function value of the dual solution exceeds the optimal objective function solution of the primal problem and thus weak duality fails to hold.

Weak duality can be obtained somewhat easily by posing the CILP and its dual in appropriate infinite-dimensional spaces, but these spaces are often too restrictive to guarantee strong duality. Therefore, the

establishment of a dual problem for CILPs that obtains both weak and strong duality remains a challenging problem. The ideal approaches to obtain this dual problem would employ, to the greatest extent possible, the deep understanding that we have of the dual problem in finite-dimensional linear programming.

The vast majority of the work in duality theory for CILPs focuses on the case where each constraint has a finite number of variables. Strong duality results are obtained for a general class of CILPs in Ref. 30 under a set of assumptions relating to degeneracy; however, these assumptions are quite restrictive in the infinite-dimensional setting. The work in Refs 16, 17, 31, and 32 provides duality results for special classes of CILPs. These works along with Refs 2 and 3 develop the duality results indirectly through the inheritance of properties from finite-dimensional truncations of the CILP. Weak and strong duality results are obtained for CILPs with inequality constraints in Refs 2 and 3 under a particular restriction on the constraint matrix. This restriction requires that “past” decisions in “future” constraints eventually have only nonnegative coefficients. For example, in production planning problems, this restriction is naturally satisfied since the inventory coming into a time period is the “past” decision and the production/inventory-balance constraint of the time period is the “future” constraint. However, this restriction does not allow the results of Refs 2 and 3 to be applied to equality-constrained CILPs. The *transversality* condition is introduced in Ref. 2 to obtain strong duality results. This condition requires that the dual variable (or multiplier) associated with the i th constraint goes to zero as i goes to infinity. The transversality condition is used in Ref. 9 to develop weak and strong duality results for infinite network flow problems with equality constraints. Duality theory for a general class of CILPs with equality constraints is developed in Ref. 33.

Infinite-Dimensional Simplex Methods

There has been success in extending the simplex method for linear programming to CILPs. The network simplex method

was extended to solve problems in infinite networks in Ref. 9. This was a true infinite-dimensional network simplex method, that is, it moved between adjacent extreme points of the feasible region of the problem using the characterization of extreme points from Ref. 4. However, in general, each pivot operation could require an infinite amount of work. For the class of infinite network flow problems whose flow-balance constraints are inequalities (i.e., we do not need to ship all supply from a node and we can leave more at a node than its demand), which include production planning problems, it is shown that each pivot operation can be performed in a finite amount of time.

A shadow simplex method for CILPs is developed in Ref. 1. The term *shadow* comes from the fact that the method performs computations (in particular, pivot operations from the simplex method for finite-dimensional linear programming) on the projections, that is, the shadows, of the feasible region of the CILP into finite-dimensional spaces. It is assumed that appropriately defined finite-dimensional truncations of the CILP are equivalent to the projections of the infinite-dimensional feasible region. This assumption guarantees that an extreme point of the finite-dimensional truncation can be extended to an extreme point for all truncations that are in a larger-dimensional space (including the infinite-dimensional space) using the results of Ref. 28. Although this assumption appears restrictive, it is shown in Ref. 1 that we can make this assumption without loss of generality for infinite-horizon deterministic or stochastic dynamic programs using a “Big-M” method and for production planning problems with a feasible solution using a “valid inequality” approach. This Shadow Simplex method solves projections of increasing dimension using the traditional simplex method. It does not, however, correspond to a true infinite-dimensional simplex method in all situations since the move from the optimal solution to a particular finite-dimensional projection to the initial solution of the next projection in higher dimensional space does not need to be between two adjacent extreme points. However, for the important cases of

infinite-horizon deterministic and stochastic dynamic programs, it is shown that the method does indeed correspond to a true infinite-dimensional simplex method.

CONTINUOUS LINEAR PROGRAMS

CLPs are infinite linear programs where the infinite dimension arises from the fact that we need to make decisions at every point in time in a continuous setting. The formulation of a CLP is given by

$$\begin{aligned} & \text{maximize } \int_0^T c(t)^\top x(t) dt \\ & \text{subject to} \\ & B(t)x(t) + \int_0^t K(s,t)x(s) ds \leq b(t) \\ & \quad \text{for all } t \in [0, T] \\ & x(t) \geq 0 \quad \text{for all } t \in [0, T]. \end{aligned} \quad (5)$$

This problem was discussed in Refs 34 and 35 to model some economic processes. It is clear that CLPs have an uncountable number of decision variables and an uncountable number of constraints since we are considering all points in time over a continuous and finite horizon. The obstacles of linear programming in infinite-dimensional spaces have been overcome to some degree for CLPs and various special cases of them. Typically, these special cases of CLPs come in the form of restricting the functional spaces associated with the parameters in the problem and the functional space of the solutions. Extreme points of CLPs have been characterized for various forms of functional spaces of the parameters in Ref. 36. There have been duality results for CLPs [37] and interior point methods have been analyzed to solve them [38]. For a complete overview of the theory developed for CLPs, we refer the reader to Ref. 11.

As an example of a CLP, we will consider the problem of maximizing the total flow sent from a source node to a sink node over a finite horizon in a continuous-time setting [39,40]. In particular, we consider a directed network,

$G = (N, A)$, where we are interested in sending the flow from the source node u to the sink node v over the horizon $[0, T]$. Each intermediate node $i \in N \setminus \{u, v\}$ has an associated function a_i that represents its storage capacity as a function of time and each arc $(i, j) \in A$ has an associated function u_{ij} that represents its flow bound as a function of time. We then seek a collection of nonnegative functions x_{ij} over $[0, T]$ where $x_{ij}(t)$ represents the amount of flow on arc (i, j) at time $t \in [0, T]$ that maximizes the flow from the source to the sink. The mathematical formulation of this problem is as follows:

$$\begin{aligned} & \text{maximize } \int_0^T \left(\sum_{i:(i,v) \in A} x_{iv}(t) - \sum_{i:(v,i) \in A} x_{vi}(t) \right) dt \\ & \text{subject to} \\ & \int_0^t \sum_{j:(j,i) \in A} x_{ji}(t) dt - \int_0^t \sum_{j:(i,j) \in A} x_{ij}(t) dt \leq a_i(t) \\ & \quad \text{for } i \in N \setminus \{u, v\}, t \in [0, T] \\ & x_{ij}(t) \leq u_{ij}(t) \text{ for } (i, j) \in A, t \in [0, T] \\ & x_{ij}(t) \geq 0 \text{ for } (i, j) \in A, t \in [0, T]. \end{aligned} \quad (6)$$

The objective function of this problem maximizes a function that represents the total amount of flow that is absorbed by the sink node v over the horizon of the problem. The first set of constraints in this formulation is a generalization of the flow-balance constraints. The first integral in these constraints represents the total amount of flow that has arrived at the node i in the interval $[0, t]$ while the second integral represents the total amount of flow that has left the node i in the interval $[0, t]$. Therefore, these constraints guarantee that the amount of flow left the node i at any point in time t does not exceed the storage capacity of the node at that time.

Separated Continuous Linear Programs

An important subclass of CLPs is the so-called separated continuous linear programs (SCLPs) where the integral constraints and the instantaneous constraints are separated.

In other words, the formulation of a SCLP is given by

$$\text{maximize } \int_0^T c(t)^\top x(t) dt$$

subject to

$$\begin{aligned} \int_0^t Gx(s) ds + y(t) &= a(t) \text{ for all } t \in [0, T] \\ Hx(t) + z(t) &= b(t) \text{ for all } t \in [0, T] \\ x(t) &\geq 0 \text{ for all } t \in [0, T]. \end{aligned} \quad (7)$$

SCLPs have applications in job-shop scheduling [41] and continuous-time network flow problems [39,40,42]. The extreme points of SCLPs were characterized in Ref. 43 in a manner that is analogous to the basic solutions of finite-dimensional linear programming problems. The work in Ref. 44 provides characterizations on the structure of the optimal solutions to SCLPs (such as piecewise constant or piecewise analytic) based on various assumptions on the problem data.

Clearly, we can discretize both a CLP and an SCLP in order to end up with a finite-dimensional linear program. This is similar to approximating the CLPs through a finite-dimensional truncation. It turns out that the discretization of SCLPs proves quite powerful in analyzing them. The work in Ref. 45 developed an algorithm for classes of SCLPs under the assumptions that $a(t)$ is piecewise linear and continuous, $b(t)$ is piecewise constant, and the functions $c(t)$ are piecewise linear. This work develops an improvement step for any feasible, nonoptimal solution to the SCLP, which relies on examining an appropriate discretization of the problem. This improvement step is used to develop an algorithm to solve the SCLP and to prove a strong duality result. The convergence of this algorithm was not proven until the work in Ref. 46. However, the ideas behind the algorithm (especially the discretization) were used in developing duality theory for SCLPs in Ref. 47. This algorithm has been extended to the case where the cost functions $c(t)$ are piecewise analytic in Ref. 48.

An algorithm based on the simplex method was developed in Ref. 49 for SCLPs under the

assumptions that $a(t)$ is linear, the functions $c(t)$ are linear, and $b(t)$ is constant. The algorithm in Ref. 49 solves these SCLPs directly, that is, without discretizing the problem. A limitation of the algorithm is that its pivot operations are well defined only under a nondegeneracy assumption (although any problem can be perturbed to a nondegenerate problem). However, simple conditions for a problem to be nondegenerate are provided. In Ref. 49, an appropriate dual problem is defined, a pivotlike operation is developed to move from extreme point to extreme point, and a proof that the algorithm solves the problem by pivoting to a *finite number* of extreme points is provided.

Continuous-Time Network Flow Problems

There has been much work done for the special class of CLPs with an underlying network structure. In other words, we are given a network and we need to determine the amount of flow on each arc as a function of time (i.e., the value $x_{ij}(t)$ corresponds to the amount of flow on edge (i, j) at time t) so that each node satisfies a generalization of the flow-balance constraints at each point in time [39]. These continuous-time network flow problems have various objectives including maximizing the amount of flow from a source to a sink over a finite horizon [39], minimizing the amount of time needed to empty the network of excess supply [50], and minimizing the total cost incurred by meeting specified demands over the horizon (which is the extension of the minimum cost flow problem to the continuous setting, see Ref. 42). The work in Ref. 39 extends the classic max-flow min-cut result to continuous maximum flow problems where the edges have zero traversal time. This result is generalized to the case where edges have traversal times in Ref. 40. The network simplex algorithm was extended to separated continuous network flow problems in Ref. 42; however, this method is known not to converge to the optimal solution of the problem. A solution method that specializes the algorithm of Ref. 45 to separated continuous network flow problems was developed in Ref. 51 and shown to converge to the optimal solution.

Another stream of research related to continuous-time network flow problems is to develop efficient (i.e., polynomial-time) algorithms or approximation algorithms for their special classes. The work in Ref. 52 extends various polynomial-time algorithms for dynamic network flow problems in a discrete time setting to a continuous time setting under the assumptions that edge-capacities and traversal times are integral and constant over time. Approximation schemes for various continuous-time network flow problems are given in Refs 53 and 54. The assumptions placed on the data in Ref. 53 guarantee that the optimal solution to the problem is piecewise constant using the work of Ref. 43. However, it is not known if the number of pieces of the optimal solution is polynomial in the input size of the problem. This problem is overcome (to a certain extent) in Ref. 53 by developing an algorithm that returns a solution whose value is arbitrarily close to the optimal objective function value and whose size (i.e., the number of pieces characterizing it) is polynomial in terms of the input into the algorithm. This solution is obtained by solving a discretized version of the continuous problem.

SUMMARY

We have discussed the many challenges that arise in the study of infinite linear programs. Many of these challenges come from the fact that the fundamental properties of finite-dimensional linear programming (e.g., the characterization of extreme points or duality theory) do not apply or extend in a straightforward manner to infinite linear programs. We discussed the two main classes of infinite linear programs, CILPs and CLPs. For CILPs, we saw that the analysis of finite-dimensional truncations of the infinite linear program were used to overcome many of the challenges faced by this class of infinite linear programs. Similarly, finite-dimensional discretizations of CLPs are important in their analysis. These methods use our vast knowledge of finite-dimensional linear programming to aid in the analysis of infinite linear programs.

There has been success in developing simplex (or simplex-like) methods to solve infinite linear programs that can perform operations in finite time with finite information. The development of simplex (or simplex-like) methods for new classes of infinite linear programs is an important area of future research, especially those methods that reduce to “true” simplex methods, that is, ones that move between adjacent extreme points.

REFERENCES

1. Ghate A, Sharma D, Smith RL. A shadow simplex method for infinite linear programs. *Oper Res* 2009. In press.
2. Romeijn HE, Smith RL, Bean JC. Duality in infinite dimensional linear programming. *Math Program* 1992;53:79–97.
3. Romeijn HE, Smith RL. Shadow prices in infinite dimensional linear programming. *Math Oper Res* 1998;23(1):239–256.
4. Romeijn HE, Sharma D, Smith RL. Extreme point characterizations for infinite network flow problems. *Networks* 2006;48(4):209–222.
5. Hopkins DSP. Infinite-horizon optimality in an equipment replacement and capacity expansion model. *Manage Sci* 1971;18(3):145–156.
6. Bean JC, Lohmann JR, Smith RL. A dynamic infinite horizon replacement economy decision model. *Eng Econ* 1985;30:99–120.
7. Bean JC, Smith RL. Optimal capacity expansion over an infinite horizon. *Manage Sci* 1985;31:1523–1532.
8. Garcia A, Smith RL. Solving nonstationary infinite horizon dynamic optimization problems. *J Math Anal Appl* 2000;244:304–317.
9. Sharkey TC, Romeijn HE. A simplex algorithm for minimum-cost network-flow problems in infinite networks. *Networks* 2008;52(1):14–31.
10. Ghate A, Smith RL. Characterizing extreme points as basic feasible solutions in infinite linear programs. *Oper Res Lett* 2009;37:7–10.
11. Anderson EJ, Nash P. *Linear programming in infinite dimensional spaces*. New York: Wiley; 1987.
12. Luenberger DG. *Optimization by vector space methods*. New York: Wiley; 1969.
13. Goberna MA, Lopez MA. *Linear semi-infinite optimization*. New York: Wiley; 1998.

14. Goberna MA, Lopez MA. Linear semi-infinite programming theory: An updated survey. *Eur J Oper Res* 2002;143:390–405.
15. Ahuja RK, Magnanti TL, Orlin JB. *Network flows: theory, algorithms, and applications*. Englewood Cliffs (NJ): Prentice-Hall; 1993.
16. Grinold RC. Infinite horizon programs. *Manage Sci* 1971;18:157–170.
17. Grinold RC. Finite horizon approximations of infinite horizon linear programs. *Math Program* 1977;12:1–17.
18. Grinold RC. Convex infinite horizon programs. *Math Program* 1983;25:64–82.
19. Bean JC, Smith RL. Conditions for the existence of planning horizons. *Math Oper Res* 1984;9:391–401.
20. Bès C, Sethi SP. Concepts of forecast and decision horizons: Application to dynamic stochastic optimization problems. *Math Oper Res* 1988;13:295–310.
21. Schochetman IE, Smith RL. Infinite horizon optimization. *Math Oper Res* 1989;14:559–574.
22. Ryan S, Bean JC. Degeneracy in infinite horizon optimization. *Math Program* 1989;43:305–316.
23. Schochetman IE, Smith RL. Convergence of best approximations from unbounded sets. *J Math Anal Appl* 1992;166:112–128.
24. Federgruen A, Tzur M. The dynamic lot-sizing model with backlogging: a simple $O(n \log n)$ algorithm and minimal forecast horizon procedure. *Nav Res Logist* 1993;40:459–478.
25. Federgruen A, Tzur M. Minimal forecast horizons and a new planning procedure for the general dynamic lot-sizing model: nervousness revisited. *Oper Res* 1994;42:456–468.
26. Federgruen A, Tzur M. Fast solution and detection of minimal forecast horizons in dynamic programs with a single indicator of the future: applications to dynamic lot-sizing models. *Manage Sci* 1995;41:874–893.
27. Roy NM. Extreme points of convex sets in infinite dimensional spaces. *Am Math Mon* 1987;94:409–422.
28. Cross WP, Romeijn HE, Smith RL. Approximating extreme points of infinite dimensional convex sets. *Math Oper Res* 1998;23(2):433–442.
29. Grinold RC, Hopkins DSP. Duality overlap in infinite linear programs. *J Math Anal Appl* 1973;41:333–335.
30. Clark SA. An infinite-dimensional LP duality theorem. *Math Oper Res* 2003;28(2):233–245.
31. Evers J. A duality theory for infinite horizon optimization of concave input/output processes. *Math Oper Res* 1983;8:479–497.
32. Jones P, Zydiak J, Hopp W. Stationary dual prices and depreciation. *Math Program* 1988;41:357–366.
33. Ghaté A, Smith RL. Duality theory for countably infinite linear programs. Technical report, Ann Arbor (MI): Department of Industrial and Operations Engineering, The University of Michigan; 2006.
34. Bellman R. Bottleneck problems and dynamic programming. *Proc Natl Acad Sci U S A* 1953;39:947–951.
35. Bellman R. *Dynamic programming*. Princeton (NJ): Princeton University Press; 1957.
36. Perold AF. Extreme points and basic feasible solutions in continuous time linear programming. *SIAM J Control Optim* 1981;19(1):52–63.
37. Grinold RC. Symmetric duality for continuous linear programs. *SIAM J Appl Math* 1970;18:32–51.
38. Ito C, Kelley T, Sachs EW. Inexact primal dual interior point iteration for linear programs in function spaces. *Comput Optim Appl* 1995;4:189–202.
39. Anderson EJ, Nash P, Philpott AB. A class of continuous network flow problems. *Math Oper Res* 1982;7:501–514.
40. Philpott AB. Continuous-time flows in networks. *Math Oper Res* 1990;15(4):640–661.
41. Anderson EJ. A new continuous model for job-shop scheduling. *Int J Syst Sci* 1981;12:1469–1475.
42. Anderson EJ, Philpott AB. A continuous-time network simplex algorithm. *Networks* 1989;19:394–425.
43. Anderson EJ, Nash P, Perold AF. Some properties of a class of continuous linear programs. *SIAM J Control Optim* 1983;21(5):758–765.
44. Pullan MC. Forms of optimal solutions for separated continuous linear programs. *SIAM J Control Optim* 1995;33(6):1952–1977.
45. Pullan MC. An algorithm for a class of continuous linear programs. *SIAM J Control Optim* 1993;31(6):1558–1577.
46. Pullan MC. Convergence of a general class of algorithms for separated continuous linear programs. *SIAM J Optim* 2000;10(3):722–731.
47. Pullan MC. A duality theory for separated continuous linear programs. *SIAM J Control Optim* 1996;34(3):931–965.

48. Pullan MC. An extended algorithm for separated continuous linear programs. *Math Program* 2002;93:415–451.
49. Weiss G. A simplex based algorithm to solve separated continuous linear programs. *Math Program* 2008;115:151–198.
50. Hajek B, Ogier RG. Optimal dynamic routing in communication networks with continuous traffic. *Networks* 1984;14(3):457–487.
51. Philpott AB, Craddock M. An adaptive discretization algorithm for a class of continuous network programs. *Networks* 1995;26(1):1–11.
52. Fleischer L, Tardos E. Efficient continuous-time dynamic network flow algorithms. *Oper Res Lett* 1998;23:71–80.
53. Fleischer L, Sethuraman J. Efficient algorithms for separated continuous linear programs: The multicommodity flow problem with holding costs and extensions. *Math Oper Res* 2005;30:916–938.
54. Fleischer L, Skutella M. Quickest flows over time. *SIAM J Comput* 2007;36:1600–1630.

INFORMATION SHARING IN SUPPLY CHAINS

ZHONG JUSTIN REN
Boston University School of
Management, Boston and MIT
Sloan School of Management,
Cambridge, Massachusetts

INTRODUCTION

Information sharing is crucial for supply chain coordination, as no supply chain can run smoothly without information. In this article, we first provide an anatomical view of the different types of information that exist in a supply chain. Then, we delve into each by reviewing relevant literature. The goal of this article is to provide a comprehensive but *nontechnical* introduction to the literature. For a technical treatment of this topic, readers are referred to Chen [1].

The literature on information sharing in supply chains is vast. While an exhaustive review is not possible here, we try to highlight those that make recent contribution to the literature, especially those appearing in the last 15 years (1994–2009). Whenever possible we juxtapose empirical research along with analytical papers on a given topic so as to highlight its relevance to practice. This article also intersects with a number of other articles in this book, such as those on supply chain coordination, (see *Supply Chain Coordination*), inventory record accuracy (see *Inventory Record Inaccuracy in Retail Supply Chains*), vendor managed inventory (see *Vendor-Managed Inventory*) and RFID applications (See *Inventory Inaccuracies in Supply Chains: How Can RFID Improve the Performance?*). Readers are referred to those articles for discussion on specific topics.

RESEARCH OVERVIEW

In this section, we first provide an overview of different types of information, then we enumerate the major research themes associated with information in supply chains.

Different Types of Information in a Supply Chain

Let us start by looking at a simple three-stage serial supply chain for a product depicted below (Fig. 1).

In this typical supply chain, demand arrives at the retailer, who orders from its distribution channel, who in turn orders from the manufacturer. As one of the supply chain parties, what information is needed in order for it to optimize its operations? Let us enumerate the types of information that are explicit or implicit in the supply chain:

- *Demand / Orders.* Supply chains exist in order to match demand with supply, therefore demand or order information is perhaps the most commonly shared type of information in a supply chain.
- *Forecasts.* Sometimes what are received from the downstream party or sent to its upstream party are not orders per se, but just forecasts; that is, intention to buy with no commitment, either in the form of a point estimate or a distribution.
- *Inventory / Order Status Information.* This includes order status information such as the location and quantity of inventory at each stage in the system. It may also include what inventory policies are being used at each stage of the system.
- *Lead Time information.* It includes order lead time and demand lead time (whose exact definition will be dealt with later).
- *Costs.* These may include the cost of manufacturing, distributing, or selling

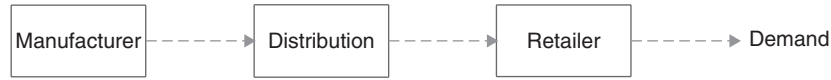


Figure 1. A Simple three-stage serial supply chain

products. Such information is not always available to all supply chain parties.

There are several other less-researched categories of information. For example, product development information of manufacturers. (Does it offer any benefit if a manufacturer shares its product development information with its downstream parties?) Other examples include information on supply chain structures, capacity, and types of contracts. We will discuss them in the latter sections of this article.

Most of the research in information sharing assumes that information existing in the supply chain is accurate (except for those dealing with strategic information-sharing behavior). However, that is frequently not true in practice. Supply chain parties have to deal with inaccurate information. What if we know some information is subject to errors, and may be inaccurate? Is it beneficial to have the knowledge that some information is bad? We will cover literature dealing with such questions as well. Next, we briefly summarize major research themes in the literature that are related to information sharing in supply chains.

Research Themes

Research related to information and information sharing in the OR/OM (operations management) field has been evolving around, but is not limited to, the following research questions:

1. What is the value of having a certain type of information in a supply chain, such as demand forecast or order information? Or, put in another way, what is the impact of *not* having a certain type of information?
2. How can information be used in improving system efficiency? What are the

mechanisms through which information would create value in improving the efficiency of supply chain parties?

3. What are the incentive issues in information sharing?
4. How is the value of information affected by contracts, organizational factors of supply chain parties, and market structure?

With these questions in mind, we review relevant literature by each category of information. Within each category, we organize articles by their research topics and contribution to the literature.

LITERATURE

Demand-Related Information

Demand or its related information may be on the top of the list if one talks about information sharing in a supply chain. There has been intense interest during the past decade on studying the value of sharing such demand information in various types of inventory systems. As a result, a significant body of knowledge has been created in this area.

Value of Demand Information. Chen [2] has contributed to one of the earliest works that study the value of demand information in an N -stage serial inventory system. In its model, each stage uses a stage-specific recorder point/order quantity, namely (R, nQ) policy. That is, whenever the inventory position at a stage falls below R , order nQ units to bring its inventory position back to be above R . Here n is the minimum integer required to achieve that. The paper quantifies the value of demand information by comparing two types of (R, nQ) policy. The first is based on echelon-stock calculation, which is the inventory position of the subsystem, which includes the stage itself and all

the downstream stages. This policy, therefore uses centralized demand information of all stages. The other policy is based on installation stock, which is its local inventory position. The difference of cost between the two policies is the value of having centralized information. Results show that such a value ranges from 0 to 9%.

Gavirneni *et al.* [3] studied a two-stage capacitated inventory system with a supplier and a retailer, where the retailer uses an (s, S) policy. Value of information is studied in a progressive fashion with three scenarios. The base scenario is where the only information the supplier observes is the retailer's orders, which is assumed to be i.i.d. In the second scenario, the supplier knows the following information: demand distribution of the retailer, that the retailer uses an (s, S) policy and its specifics, and when the last order from the retailer occurred. In the third scenario, the supplier has full information. It is shown that from the base scenario to the second, the supplier's cost decreases from 10 to 90%, whereas the cost decrease from the second to the third ranges from 1 to 5%.

Lee *et al.* [4] study a two-stage serial supply chain with nonstationary demand. In particular, end-product demand is assumed to follow an AR(1)(Autoregressive) process. The key question here is how large the benefit is for the supplier to observe the end-product demand, instead of just the retailer's orders. They show that the benefit of sharing demand information can be quite large, especially when demand correlation is high. Gaur *et al.* [5] study a model with autoregressive moving average (ARMA) demand, and find that the value of sharing demand information (or inferring demand information from orders) can be significant.

Aviv and Federgruen [6] study the value of demand information in the context of the Vendor Managed Inventory (VMI) program. In their model setup, there is one supplier who replenishes for multiple independent retailers. The value of information is obtained by comparing the supply chain costs in a decentralized model with no information sharing (Model 1) against that with a decentralized one with information sharing (Model 2), and finally a supplier-controlled

system with full information, that is VMI. The authors find that the average savings in cost from Model 1 to Model 2 is about 2%, and that from Model 2 to Model 3 is about 4.7%. Article 4.5.3.4 discusses in further detail the benefits (and challenges) in VMI.

It should be noted that not all inventory models lead to the conclusion that there is significant benefit from sharing demand information. Raghunathan [7] studies the supply chain setting presented in Lee *et al.* [4], and shows that if the manufacturer can improve its forecasting method by using the retailer's entire order history, the benefit of sharing demand information may not be necessarily large. Another example is given in the study by Graves [8], where a two-stage serial system with autoregressive integrated moving average (ARIMA) $(0, 1, 1)$ demand was studied. He first shows that the demand process that the supplier faces is also of ARIMA $(0, 1, 1)$. In addition, there is no benefit for the supplier in observing any extra information such as the end-market demand. Chen and Lee [9] provide a unified framework that is able to resolve the differences in literature. Moreover, they show that the value of information sharing is maximized when the retailer shares its order forecasts truthfully with the supplier.

Sharing Demand Information with Competition. Recent papers have further deepened our understanding about information sharing in different supply chain settings such as among competing retailers, or among competing supply chains. Ha and Tong [10] investigate competition between two serial supply chains, each with one manufacturer and one retailer. In each supply chain, the manufacturer decides whether to invest in sharing demand size information with its respective retailer. Retailers then engage in Cournot competition based on selling quantities. The authors show that interestingly, the value of information sharing now critically depends on the contracts each supply chain adopts. With the simple linear contracts, the value of information sharing is actually negative. This is because more information allows the manufacturer to increase its profit at the

expense of the whole supply chain performance. For quantity-based contract menus, however, information sharing does bring positive value to the supply chain.

In practice, when competing customers source from the same manufacturer, information leakage has been a major concern for customers who do not intend to share their order information with their competitors [11]. Li and Zhang [12] study the implication of multiple competing retailers sharing demand information with each other, and propose a “confidentiality” scheme in information sharing. In their model, the retailers engage in Bertrand (i.e., price) competition, but they all source from one common supplier. When each of the retailers shares its information confidentially with the manufacturer, they can only infer demand information from the wholesale price that the manufacturer sets. Alternatively, retailers could choose to disclose their private information to everyone, or a selected set of all supply chain parties. Interestingly, they show that information confidentiality actually improves supply chain efficiency. This is because confidentiality alleviates the “double marginalization” problem existing when the manufacturer sets the wholesale price. Confidential information sharing vertically centralizes each retailer’s information at the manufacturer without horizontal direct disclosure to other retailers. Such strategic signaling lowers the equilibrium wholesale price and therefore improves the total supply chain profits.

Value of Advance Orders Information. Another class of demand information is advanced orders, or “pre-orders.” Sometimes, customers are willing to wait for a certain length of time before receiving and consuming a product. This advance demand information should permit suppliers to better forecast demand and replenish inventory. Tang *et al.* [13] study such benefits of having advance demand information. They consider a seasonal product with uncertain demand and long lead time (think of the production of “mooncake,” a type of pastry that is consumed for the annual “mid-autumn festival” in many Asian cultures). The

retailer can induce customers to pre-order for a discounted price before the selling season, which the authors call “Advance Book Discount” (ABD) program. The time between placement and fulfillment of these precommitted orders provides an opportunity for the retailer to update demand forecasts. The authors solve for the optimal discount rate and demonstrate the benefit of such advance demand information.

It turns out that in inventory systems there exists an elegant relationship between demand lead time, that is the time from customer’s pre-orders to their delivery dates, and supply lead time. Hariharan and Zipkin [14] formally establish the following insight: demand lead time is the opposite of supply lead times. This is intuitive in the sense that demand lead time essentially allows planners to “peek into the future” and see perfectly what orders need to be fulfilled within the demand lead time window. Therefore, while long supply lead time increases supply chain cost, obtaining such advance demand information decreases it. See also the work of Chen [15] that further generalizes such a result when customer orders are heterogeneous in their lead time. Wang and Toktay [16] study the case where demand lead time is flexible; that is, early delivery is permissible. They find that with delivery flexibility increasing, the demand lead time is more beneficial than decreasing the supply lead time.

Sharing Forecast Information

Forecast information has very different characteristics, compared to “real” information. First of all, forecasts are, by definition, imprecise. Their accuracy is dependent upon which forecasting methodologies are being used, and the forecaster’s knowledge about the underlying state of the world to be forecasted. Secondly, forecasts very often represent forecasters’ intentions, and are not contractible. Therefore, incentives in sharing forecast information become important in driving supply chain relationships. Below, we discuss each theme in detail.

How to Utilize Forecast Information in Supply Chains. Incorporating demand forecasts into production or inventory system is one of

the most fruitful areas in operations literature, and we therefore devote more space to this section.

With the increasing application of material requirement planning (MRP) system in industry, the role of forecast updating becomes even more relevant to practice. Motivated by an industry study, Heath and Jackson [17] propose a martingale model of forecast evolution (MMFE) to characterize the process of forecast evolution, and apply this model in a simulation study to analyze safety stock levels for a production and distribution system. Graves *et al.* [18] independently developed an MMFE model to study requirement planning in multistage production-inventory system. Later works that use MMFE include Güllü [19], Chen *et al.* [20], Toktay and Wein [21], and Chen and Lee [9]. Kaminsky and Swaminathan [22] consider a capacitated production planning model where forecasts are updated each period. Forecasts are modeled to become more accurate over time, as the confidence interval of forecasts in which true demand lies, decreases.

Another approach in modeling forecast updating is to assume a joint statistical distribution between two random variables (e.g., demands in two periods). Then conditional upon the realization of the first quantity, forecasts updating of the second quantity can be modeled as a conditional distribution. This approach is advantageous when those two random variables are correlated. Some of the recent examples that adopt this approach include the works of Fisher and Raman [23], Eppen and Iyer [24], Donohue [25], Gurnani and Tang [26], and Agrawal *et al.* [27].

However, Cattani and Hausman [28] question the presumption that forecasts should become more accurate over time. They show both, theoretically and empirically, that demand forecasts do not necessarily become more accurate as they are updated. They argue that such forecast churning can cause inefficiencies if the firm reacts to the wrong forecast update. Cohen *et al.* [29] and Terwiesch *et al.* [30] empirically confirm this observation, based on their study of the semiconductor equipment industry. Miyaoka and Hausman [31] show with an

inventory model using base-stock policy that “stale” forecasts do not necessarily hurt, and actually improve, supply chain benefit. See also Refs 32 and 33 on two different models on how better forecasts from downstream firms may not uniformly provide benefits for everyone in the supply chain. While some of the findings regarding the effectiveness of forecast sharing are statistical, others uncover that incentives also play an important role in determining the efficiency and effectiveness of forecast sharing. Next, we discuss related incentive issues in detail.

Incentives in Sharing Forecast Information.

Most of the work in this category is analytical. Their models set themselves apart from those inventory optimization ones in that they assume supply chain parties are strategic. That is, they seek to maximize their own profit or utility, and do not always fully cooperate in a supply chain. In the context of forecast sharing, this means that if they are the ones generating forecasts, they may not send their true forecasts; if they are at the receiving end, they may not treat received information as true, or react to it as in a centralized environment.

Lee *et al.* [34] provide one of the earliest academic discussions of problems related to order forecasts and their implication on supply chain coordination. They refer to order forecasts that are eventually cancelled as “phantom orders,” and see them as a key contributor to the bullwhip effect in supply chains. Given that for some demand outcomes the supplier may be capacity constrained, the buyer has an incentive to forecast extra orders (phantom orders), especially if scarce supplier capacity is rationed based on placed orders. Armony and Plambeck [35] investigate how such false orders can lead a manufacturer to overestimate the demand and make faulty decisions about capacity investment.

Another cause of inefficiency in the supply chain may be the misuse of statistical forecasting method. Chen *et al.* [20] show that the bullwhip effect can result if a customer uses moving average forecasting method to forecast a serially correlated demand. Watson

and Zheng [36] show that when customers overestimate the demand correlation, or use the improper forecasting tools, it can increase order volatility and in turn the bullwhip effect, and the resulting outcome can be worse than not forecasting at all.

When a supply chain party possesses private information that the other parties do not know, how to induce truth-telling in sharing forecast is an important topic. Cachon and Lariviere [37] formulate a capacity procurement game, where they distinguish between two types of buyer–supplier relations concerning forecast sharing. Under forced compliance, the supplier’s behavior is observable to the buyer. In contrast, under voluntary compliance, behavior is not verifiable and could be attributed to stochastic influences from the environment. Therefore, in order to induce the supplier to provide enough capacity for high demand, the high demand type customer has to exert extra incentive to serve as signaling device, such as fixed payment or a large firm commitment, to separate himself from a low-demand type customer. Özer and Wei [38] show that in the setting of Cachon and Lariviere [37], channel coordination is possible when combining capacity commitment contract with payback contract. Ren *et al.* [39] propose a long-term perspective of this problem. When the supply chain parties expect to transact repeatedly over a long horizon, they show that under some conditions a simple wholesale price can achieve truthful information sharing and system coordination asymptotically. See the article titled **Supply Chain Coordination** in this encyclopedia for a related discussion on supply chain coordination.

In the area of collaborative planning, forecasting, and replenishment (CPFR), Aviv [40,41] study a supply chain system where its members jointly maintain and update a single forecasting process in the system. Three types of two-stage supply chain configurations are considered. The first is where retailer and supplier coordinate inventory policy but do not share demand forecast information. In the second setting, the supplier is in charge of supply chain inventory, but retailer does not share

demand signal, while the last model is a centrally managed system with full information sharing. He shows that the benefit of sharing forecast information increases as supply chain members are able to explain a large portion of demand uncertainty through early demand information, especially in the case of correlated demand. However in this setting, firms’ incentives in forecast sharing are not considered. Deshpande *et al.* [42] propose a secure collaborative planning, forecasting, and replenishment (SCPFR) protocol that facilitates forecast sharing and inventory planning while at the same time preventing information leakage.

Inventory-Related Information

Research in this category is generally centered on studying the value of having inventory information or its related status information. The advance of information technology such as RFID has made such information, and research on this topic, increasing relevant to supply chain management practice.

Value of Inventory or Inventory Status Information. One basic question in inventory system optimization is “Is it helpful to know inventory information of downstream supply chain parties?” Researchers have studied this question with a variety of approaches: analytical modeling, controlled experiments, and empirical studies.

Cachon and Fisher [43] study this question with a one-warehouse multiretailer system. In the base case with no information sharing, both the retailers and the warehouse use an (R, nQ) policy, but the warehouse only observes the retailers’ orders. With information sharing, however, the warehouse can see retailers’ inventory status at any time. This gives the warehouse opportunities to allocate inventory optimally; this is based on the inventory need of each retailer. Using a numeric study, they show that the benefit of sharing such inventory information is 2.2% of the total supply chain cost on the average. They also study two other benefits of information technology: reducing order lead time, and reducing batch size; both improve the

physical flow of goods (as opposed to improving the information flow). It is found that the two benefits associated with improving physical flow of goods are more significant than those in sharing inventory information.

Zhang *et al.* [44] analyze a repeated game with a retailer and a supplier under asymmetric inventory information. In each period of the game, the supplier offers a contract, but he/she does not observe the inventory position nor the sales of the retailer. This class of problems are called *dynamic adverse-selection models with hidden Markovian states*, and are typically hard to solve. However, the authors show that under exponential demand and when the cost parameters are in a certain critical region, the optimal infinite-horizon contract for the supplier is stationary: the supplier gives a take-it-or-leave-it offer to the retailer in which a fixed quantity can be purchased for a fixed amount.

In a controlled experimental environment, Croson and Donohue [45] study the benefit of sharing inventory information in the context of “the beer game.” First, they find that decision makers consistently underweigh the pipeline stock even when the normal operational causes (e.g., batching, price fluctuations, and demand estimation) are removed, and as a result the bullwhip effect persists. Will sharing inventory information help remove some of the variability? It turns out sharing inventory information has little effect on the orders of downstream chain members. On the positive side, inventory information seems to substantially reduce the variance of orders for upstream members.

The empirical evidence on the effect of sharing inventory information is relatively few, but the results are generally positive. Clark and Hammond [46] find a correlation of VMI practices with performance improvements. Kulp *et al.* [47] conduct an extensive study with the food and consumer packaged goods industry, and find that sharing retail store inventory levels provide benefit to the manufacturer’s profitability. So does such collaborative planning on replenishment initiatives such as VMI.

Value of Sharing Order Status Information.

What about knowing the status of orders in transit or in process? Will that bring any value at all? Before the advent of sophisticated information technology, this question is not that relevant since such information is plainly unobtainable. However, as technologies such as RFID and GPS systems are becoming widely available and are increasingly being deployed and utilized in recent years, it becomes interesting to ask what the value of having order status information is. Gaukler *et al.* [48] analyze such potential of knowing order progress information. In a traditional supply system, a retailer placing an order only knows the time of its order release and the distribution of its stochastic lead time before it arrives. Now, with new tracking technology such as RFID, the retailer is able to track the status and location of its orders. This not only helps the retailer in deciding when to order and how much to order, but also enables the retailer to place an emergency order in the event of a likely delay. They solve the optimal inventory control policy [within the class of (Q, R) policies] with such added visibility, and show that the cost savings could be significant, especially in supply chains where lead times are long and variable. See also Ref. 49 and the article titled 4.5.3.4 in this encyclopedia for related studies and recent developments on RFID technologies.

In another stream of research, Jain and Moinszadeh [50] consider the value of manufacturers sharing inventory availability information with retailers. In their model of a typical two-tier serial supply chain, the manufacturer informs the retailer whether the product will be backordered before the retailer places its order. By doing so, the manufacturer shares information on inventory availability in the manufacturer’s finished-goods inventory. The retailer can then take advantage of such information by changing to a new ordering policy. But does the manufacturer have an incentive to share inventory availability information? The authors reveal that indeed, sharing such information will induce the retailer to increase its order rate.

Lead Time Information

Mainstream inventory optimization models have long embraced the uncertain nature of lead times. As a result, we know quite a lot about what to do when precise lead time information is not available in a system. For example, Song and Zipkin [51] studied a single-location inventory system of, say, a retailer, whose order lead time was random and followed a Markov process. Given that the retailer is able to have the lead time information for each period before it places an order, the authors show that a state-dependent base-stock policy is optimal.

But what if the retailer does not have such lead time information? What is the value of having such information? Chen and Yu [52] study this question. The authors compare two scenarios: one without information sharing, and the other with information sharing as in Ref. 51. Without information sharing on lead time information, the retailer can only infer current lead time from past order data in making replenishment decisions, whereas if the supplier cares to share order lead time information, the retailer can then implement the optimal state-dependent base-stock replenishment policy. They find that the value from sharing lead time information can be quite high, especially for high volume items.

Cost Information

In the past, it has generally been assumed in supply chain research that the cost structure of supply chain parties (e.g., cost of manufacturing, distribution, and selling for each stage of a supply chain) is public knowledge. In practice, however, supply chain parties often do not have perfect information on the other parties' costs. Recent research has challenged this assumption, and has since made substantial progress by incorporating asymmetry cost information.

Supply Chain Coordination under Information Asymmetry. Ha [53] studies a supply chain with a buyer and a supplier, but the marginal cost information of the buyer is unknown to the supplier. In order to induce truthful information, the supplier offers a

menu of contracts to distinguish the different buyer types. It is shown that because of information asymmetry, supply chain profit-maximizing solution is not attainable with a linear price contract. Corbett and de Groote [54] provide another example of screening contract in the context of a two-stage economic lot-sizing problem, where the buyer's cost information is private. Chen [15] studies a buyer who sources from multiple suppliers whose costs are only privately known. The author solves for an optimal procurement method, which is a quantity-payment schedule. This contract is auctioned off among the suppliers, and the winner can choose any quantity to deliver and gets paid according to the pre-announced payment schedule.

Recently, researchers began to study the impact of asymmetric cost information in a repeated relationship setting. Taylor and Plambeck [55] study a repeated capacity procurement game with one customer and one supplier. In their model setting, production cost is uncertain and therefore *ex ante* binding contracting is infeasible. They identify an optimal "relationship contract" that maximizes the customer's expected profits. Taylor and Plambeck [56] further study two simple forms of relational contracts, where the buyer can either promise to pay a specific price, or can promise to purchase a specific quantity. These contractual arrangements are both easy to implement, and the authors show that which option is optimal to the buyer depends on the production and capacity cost, as well as the discount factor.

Lee *et al.* [57] study a finitely repeated game. The setting is similar to the "selling to the newsvendor" model described in Ref. 58, except that the buyer's cost information is private, but they have opportunities to transact repeatedly. They find that the buyer's "information advantage" persists in the initial phase of the repeated game. However, as the game goes on, such an information advantage gradually dissipates, and the buyer starts to randomize between acceptance and rejection of the contract. Therefore, his/her private information is revealed in a probabilistic fashion. The supplier's long-run equilibrium strategy is of a threshold type: he/she starts with offering a low selling

price and keeps doing so until his/her belief about the buyer's type is over a threshold. These results indicate that repeated supply chain relationships have significantly different dynamics than those observed in a one-shot supply chain relationship.

Sharing Information about Product Development

Product development sits at the top of a supply chain. Therefore, it is relevant to study information sharing that is related to product development. There are two dimensions to this. The first is information sharing within the process of product development. There, the research question is how to best share information in order to come up with successful products with faster lead times. The second dimension is about the value of sharing information about product development, with other echelons of the supply chain. Below we review both areas of research.

The difficulty in sharing product information within the product development functional area is similar to what is observed in sharing forecast information. In both environments, the sharing of preliminary information is a common practice. Development teams frequently begin their work on a new product prior to receiving detailed design specifications from the customer and/or from the market research department. Therefore, how to utilize preliminary information is a central question. In this line of research, Krishnan *et al.* [59] and Loch and Terwiesch [60] model the situation faced by a concurrent engineering team where an information-receiving development activity must decide on how to rely on the preliminary information provided by the information sender. The information receiver always wants to start early, in an attempt to gain from parallel task execution. Starting early, however, uses a lower quality of information and thus has a higher likelihood of costly rework. Consequently, the information receiver faces the dilemma "Rush and be Wrong or Wait and be Late" [61]. This fundamental trade-off is similar to the supplier's problem as described above. These studies suggest that the time a project should commit resources, based on preliminary information, should be delayed

if the information received is still unstable or if experience from similar projects suggests that the information is likely to change in the near future.

Somewhat surprisingly, there has been little OM research identifying the value of sharing product development-related information with other stages of supply chains. This can be an interesting topic as in practice, firms take very different approaches. Take product preannouncement, for example. In the high-tech industry, it is well known that there are firms who never give product announcements (e.g., the famous Apple secrecy), and there are others who tend to preannounce their pipeline products well ahead of time (e.g., Microsoft). What is the impact on their supply chains? How to coordinate production and distribution under these different strategies? These are the questions that analytical modeling can help answer. On the other hand, empirical evidence seems to argue for the more open approach. For example, Kulp *et al.* [47] study the food and consumer packaged goods industry, and find that when manufacturers collaborate closely with their retailers and share product development-related information, they experience enhanced profit margins.

Accuracy of Information Shared

So far we have assumed that information shared in supply chains is accurate. Is that a universally valid assumption? Empirical evidence suggests not. We refer readers to article (See *Inventory Record Inaccuracy in Retail Supply Chains*) in this encyclopedia for an in-depth discussion, but here we briefly mention a few recent studies. Raman *et al.* [62] find that data inaccuracy persists in the retailing inventory, which results in substantial operational inefficiency and profit loss. DeHoratius and Raman [63] study inventory record accuracy with one nation-wide retailer, and find that 65% of the records have some forms of inaccuracy.

Recent analytical research has taken notice, and has begun to address this issue. Kok and Shang [64] study an inventory system with inaccurate records. A manager can conduct costly inventory inspection at the beginning of each period to correct

record inaccuracy. They show that an inspection-adjusted base-stock (IABS) policy is optimal for the single-period problem, and is nearly optimal for the finite-horizon problem. Lee and Özer [49] study the value of RFID technology in improving information accuracy. They distinguish inventory records from actual inventory, and attribute inventory inaccuracy to two causes: inventory misplacement, and transaction errors. They also show that RFID technologies allow for better control and can reduce inventory-related costs. Using a classical multiperiod periodic review inventory model, DeHoratius *et al.* [65] model inventory inaccuracy as an invisible demand process. As a result, the decision makers have imperfect information of the true inventory status. They propose a Bayesian information updation method and derive corresponding replenishment policies that are shown to effectively recoup much of the cost incurred by inventory record inaccuracy.

SUMMARY AND FUTURE RESEARCH

We have made great strides in understanding and managing information sharing in supply chains. However, supply chain innovation constantly brings about new questions and challenges. Here, we conclude with a few thoughts on some emerging issues as well as some less-explored research areas:

1. *Innovation in Supply Chain Structure.* OM literature typically deals with various supply issues under a given supply chain structure (i.e., a serial supply chain, or a distribution network). With the help of information technology, business models are undergoing tremendous innovations. This includes innovations in supply chain structures. For example, there are situations where customers may not know exactly who their suppliers are. In this setting, it is not clear if sharing information will necessarily lead to better outcome. With more innovations on supply chain structure, more research in this area will be needed.
2. *Information Technology and New Information.* With information technology, it is possible to have access to information at a more granular level. For example, it may be possible to track product-related information at every step in the supply chain. This may mean significant cost saving for perishable products. For example, Ferguson and Ketzenberg [66] analyze the value of sharing product age information. In their model, a retailer sells a perishable product to consumers and receives replenishment from a supplier. The product lifetime is fixed once received by the retailer, but the age of replenished items varies stochastically over time. Because the product is perishable, any unsold inventory remaining after the lifetime elapses must be discarded. With information sharing, the retailer then is informed of the product age, prior to placing an order, and is able to make better ordering decisions. They find that the retailer benefits the most from information sharing when the variability of demand is high, product lifetimes are short, and the cost of the product is high.
3. *Information in Service Supply Chains.* Service is dominant in today's economy. However, we do not have as much knowledge in understanding service supply chain as we do with inventory supply chains. More research is needed to study the connection between literature in the two areas, and what is different in information sharing in service supply chains. For example, what information should be shared, and what is the value of sharing such information in health-care delivery systems? We do not have good answers to such questions yet. An important distinction of service supply chains is their interaction with customers. Hence, information sharing with customers in addition to supply chain parties adds another interesting dimension of complexity. Much more research can be done in

better managing information to better customer experience and improve system efficiency.

4. *Empirical Research.* Finally, we need more empirical work analyzing and understanding the benefits of information sharing in supply chains, the drivers of such benefits, and what are the significant facilitators and mediators in realizing those benefits. Among the various areas of literature on information sharing we reviewed above, empirical research is distinctively lacking in many of them. With more empirical evidence in helping us answering those questions thoroughly, we will be able to identify new research questions and spur further analytical research.

REFERENCES

1. Chen F. In: De Kok AG, Graves SC, editors. Information sharing and supply chain coordination. Handbooks in operations research and management science. Amsterdam: Elsevier; 2003. p. 11.
2. Chen F. Echelon reorder points, installation reorder points, and the value of centralized demand information. *Manag Sci* 1998;44(12): S221–S234.
3. Gavirneni S, Kapuscinski R, Tayur S. Value of information in capacitated supply chains. *Manag Sci* 1999;45(1):16–24.
4. Lee HL, So KC, Tang CS. The value of information sharing in a two-level supply chain. *Manag Sci* 2000;46(5):626–643.
5. Gaur V, Giloni A, Seshadri S. Information sharing in a supply chain under ARMA demand. *Manag Sci* 2005;51(6):961–969.
6. Aviv Y, Federgruen A. The operational benefits of information sharing and vendor managed inventory (VMI) programs. Working Paper, The John M. Olin School of Business, Washington University, 1998.
7. Raghunathan S. Information Sharing in a Supply Chain. A Note on its Value when Demand Is Nonstationary. *Manage Sci*, 2001;47(4):605–610.
8. Graves SC. A single-item inventory model for a nonstationary demand process. *Manuf Serv Oper Manag* 1999;1(1):50–61.
9. Chen L, Lee HL. Information sharing and order variability control under a generalized demand model. *Manag Sci* 2009;55(5): 781–797.
10. Ha AY, Tong SL. Contracting and information sharing under supply chain competition. *Manag Sci* 2008;54(4):701–715.
11. Anand KS, Goyal M. Strategic information management under leakage in a supply chain. *Manag Sci* 2009;55(3):438–452.
12. Li LD, Zhang HT. Confidentiality and information sharing in supply chain coordination. *Manag Sci* 2008;54(8):1467–1481.
13. Tang CS, Rajaram K, Alptekinoglu A, *et al.* The benefits of advance booking discount programs: model and analysis. *Manag Sci* 2004;50(4):465–478.
14. Hariharan R, Zipkin. P Customer-Order Information, Leadtimes and Inventories. *Manage Sci*, 1995;41(10): 1599–1607.
15. Chen F. Market segmentation, advanced demand information, and supply chain performance. *Manuf Serv Oper Manag* 2001;3(1): 53–67.
16. Wang T, Toktay BL. Inventory management with advance demand information and flexible delivery. *Manag Sci* 2008;54(4):716–732.
17. Heath DC, Jackson PL. Modeling the evolution of demand forecasts with application to safety stock analysis in production distribution-systems. *IIE Trans* 1994;26(3): 17–30.
18. Graves S, Kletter DB, Hetzel WB. A dynamic model for requirements planning with application to supply chain optimization *Oper Res* 1998;46(3):S35–S49.
19. Güllü R. On the value of information in dynamic production inventory problems under forecast evolution. *Nav Res Logist* 1996;43(2): 289–303.
20. Chen F, Drezner Z, Ryan JK, *et al.* Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead times, and information. *Manag Sci* 2000;46(3): 436–443.
21. Toktay LB, Wein LM. Analysis of a forecasting-production-inventory system with stationary demand. *Manag Sci* 2001;47(9): 1268–1281.
22. Kaminsky P, Swaminathan J. Utilizing forecast band refinement for capacitated production planning. *Manuf & Serv Oper Manage* 2001;3(1):68–81.

23. Fisher M, Raman A. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper Res* 1996;44(1): 87–99.
24. Eppen GD, Iyer AV. Improved fashion buying with Bayesian updates. *Oper Res* 1997;45(6):805–819.
25. Donohue KL. Efficient supply contracts for fashion goods with forecast updating and two production modes. *Manag Sci* 2000;46(11):1397–1411.
26. Gurnani H, Tang CS. Note: optimal ordering decisions with uncertain cost and demand forecast updating. *Manag Sci* 1999;45(10): 1456–1462.
27. Agrawal N, Smith SS, Tsay AA. Multi-vendor sourcing in a retail supply chain. *Prod Oper Manag* 2002;11(2):157.
28. Cattani K, Hausman W. Why are forecast updates often disappointing? *Manuf Serv Oper Manag* 2000;1(2):119–127.
29. Cohen MA, Ho TH, Ren ZJ, *et al.* Measuring imputed cost in the semiconductor equipment supply chain. *Manag Sci* 2003;49(12): 1653–1670.
30. Terwiesch C, Ren ZJ, Ho TH, *et al.* An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Manag Sci* 2005;51(2):208–220.
31. Miyaoka J, Hausman W. How a Base Stock Policy Using “Stale” Forecasts Provides Supply Chain Benefits. *Manuf & Serv Oper Manage* 2004; 6(2):149–162.
32. Miyaoka J, Hausman W. How Improved Forecasts Can Degrade Decentralized Supply Chains. *Manuf & Serv Oper Manage* 2008. 10(3):547–562.
33. Taylor T, Xiao W. Does a manufacturer benefit from selling to a better-forecasting retailer? *Manage Sci*, In press. 2008.
34. Lee HL, Padmanabhan V, Whang SJ. Information distortion in a supply chain: the bullwhip effect. *Manag Sci* 1997;43(4):546–558.
35. Armony M, Plambeck EL. The impact of duplicate orders on demand estimation and capacity investment. *Manage Sci*, 2005;51(10):1505–1518.
36. Watson N, Zheng Y-S. Adverse effects of overreaction to demand changes and improper forecasting. Working paper, The Wharton School, 2002.
37. Cachon GP, Lariviere MA. Contracting to assure supply: how to share demand forecasts in a supply chain. *Manag Sci* 2001;47(5): 629–646.
38. Özer O, Wei W. Strategic commitment for optimal capacity decision under asymmetric forecast information *Manag Sci* 2006;52: 1238–1257.
39. Ren ZJ, Cohen MA, Ho TH, *et al.* Information Sharing in a Long-Term Supply Chain Relationship The Role of Customer Review Strategy. *Oper Res*. 2010;58(1):81–93.
40. Aviv Y. The effect of collaborative forecasting on supply chain performance. *Manag Sci* 2001;47(10):1326–1343.
41. Aviv Y. Gaining benefits from joint forecasting and replenishment processes: the case of auto-correlated demand. *Manuf Serv Oper Manag* 2002;4(1):55–74.
42. Deshpande V, Schwarz L, Atallah M, *et al.* Secure collaborative planning, forecasting, and replenishment. Working Paper, 2008.
43. Cachon GP, Fisher M. Supply chain inventory management and the value of shared information. *Manag Sci* 2000;46(8):1032–1048.
44. Zhang H, Nagarajan M, Sosic G. Dynamic supplier contracts under asymmetric inventory information. Working Paper, 2009.
45. Croson R, Donohue K. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Manag Sci* 2006;52(3):323–336.
46. Clark T.H, Hammond J. Reengineering Channel Reordering Process to Improve Total Supply Chain Performance. *Prod and Oper Manage*. 1997;6(3):248–265.
47. Kulp SC, Lee HL, Ofek E, Manufacturer Benefits from Information Integration with Retail Customers. *Manage Sci*;50(4):431–444. 2004.
48. Gaukler GM, Ozer O, Hausman WH. Order progress information: improved dynamic emergency ordering policies. *Prod Oper Manag* 2008;17(6):599–613.
49. Lee H, Özer O. Unlocking the value of RFID. *Prod Oper Manag* 2007;16(1):40–64.
50. Jain A, Moinszadeh K. A supply chain model with reverse information exchange. *Manuf Serv Oper Manag* 2005;7(4):360–378.
51. Song J-S, Zipkin PH. Inventory control with information about supply conditions. *Manag Sci* 1996;42(10):1409–1419.
52. Chen F, Yu B. Quantifying the value of leadtime information in a single-location inventory system. *Manuf Serv Oper Manag* 2005;7(2):144–151.
53. Ha AY. Supplier-buyer contracting: Asymmetric cost information and cutoff level

- policy for buyer participation. *Nav Res Logist* 2001;48(1):41–64.
54. Corbett CJ, de Groote X. A supplier's optimal quantity discount policy under asymmetric information. *Manag Sci* 2000;46(3):444–450.
55. Taylor TA, Plambeck EL. Supply Chain Relationships and Contracts. The Impact of Repeated Interaction on Capacity Investment and Procurement. *Manage Sci* 2006;53(10):1577–1593.
56. Taylor TA, Plambeck EL. Simple Relational Contracts to Motivate Capacity Investment. Price Only vs. Price and Quantity. *Manuf Serv Oper Manage*. 2007;9(1):94–113.
57. Lee CC, Ren ZJ, Zhou D. The role of asymmetric cost information in a finitely repeated supply chain relationship. Working Paper, 2008.
58. Lariviere MA, Porteus EL. Selling to the newsvendor: an analysis of price-only contracts. *Manuf Serv Oper Manag* 2001;3(4): 293–305.
59. Krishnan V, Eppinger SD, Whitney DE. A model-based framework to overlap product development activities. *Manag Sci* 1997;43(4): 437–451.
60. Loch CH, Terwiesch C. Communication and uncertainty in concurrent engineering. *Manag Sci* 1998;44(8):1032–1048.
61. Terwiesch C, Loch CH, De Meyer A. Exchanging preliminary information in concurrent engineering: alternative coordination strategies. *Org Sci* 2002;13(4):402–419.
62. Raman A, DeHoratius N, Ton Z. Execution: the missing link in retail operations. *Calif Manag Rev* 2001;43(3):136–152.
63. DeHoratius N, Raman A. Inventory record inaccuracy: an empirical analysis. *Manag Sci* 2008;54(4):627–641.
64. Kok AG, Shang KH. Inspection and replenishment policies for systems with inventory record inaccuracy. *Manuf Serv Oper Manag* 2007;9(2):185–205.
65. DeHoratius N, Mersereau AJ, Schrage L. Retail inventory management when records are inaccurate. *Manuf Serv Oper Manag* 2008;10(2):257–277.
66. Ferguson M, Ketzenberg M. Information Sharing to Improve Retail Product Freshness of Perishables. *Prod and Oper Manage* 2006;15(1):57.

INITIAL TRANSIENT PERIOD IN STEADY-STATE SYSTEMS

K. PRESTON WHITE JR
University of Virginia
Charlottesville, Virginia

STEWART ROBINSON
Warwick Business School,
University of Warwick,
Coventry, UK

THE PROBLEM

The problem of the *initial transient*, also known as the *start-up* or *warm-up* or *initialization-bias problem*, arises in estimating the limiting distribution of a nonterminating stochastic process from output generated by one or more replications of a discrete-event simulation. Because the steady-state operating regime does not offer natural boundary conditions for starting or stopping simulation runs, initial conditions typically are chosen for convenience and terminating conditions are sought which provide satisfactory point and interval estimates. It is well known that arbitrary initialization can introduce bias in estimators of the limiting statistics.

In this article, we discuss the problem of the initial transient and outline methods for dealing with the problem. We focus on two specific methods: one graphical (Welch's method) and the other heuristic (MSER-5).

A PRACTICAL EXAMPLE

Figure 1 provides an example of the initial transient in a model of a real system. The simulation model represents a user support help desk [1]. Requests for support arrive via telephone, email, and face-to-face contact. Some requests can be dealt with immediately, but others queue for service or require a call-out. This leads to delays of a week or more in some cases. The graph shows the total waiting time

for the first 500 calls (entities) to leave the simulation model and the cumulative average waiting time. Because the model starts (for convenience) in the infrequent state of no calls in the system, a clear bias exists in the early observations.

A SIMPLE EXAMPLE ILLUSTRATING ISSUES AND NOTATION

The fundamental issues can be elucidated using the straightforward example of an M/M/1 queue for which we are interested in estimating interval statistics for the mean number in system from a simulation experiment. The output of a single simulation run is a realization of the discrete-state, continuous-parameter stochastic process $\{Y_i; i = 0, 1, 2, \dots, n\}$, where $Y_i = Y(t_i)$ is the i th sequential observation from a simulation with run length n and initial condition $Y_0 = y_0$. The sample mean and large-sample variance are

$$\bar{Y}(n | y_0) = \sum_{i=0}^k \omega_i Y_i,$$
$$S^2(n | y_0) = \sum_{i=0}^k \omega_i Y_i^2 - \bar{Y}(n | y_0)^2,$$

where the weight $\omega_i = 1/(n + 1)$ and $k = n$ for variables which are tallied (such as customer waiting times for the help-desk example), or $\omega_i = (t_{i+1} - t_i)/(t_n - t_0)$ and $k = n - 1$ for variables which are time-averaged (such as the number in system for the M/M/1 queue). Since the process generating these observations is ergodic, these are consistent estimators for the steady-state mean θ and variance σ^2 , independent of the choice of initial condition, that is, $\lim_{n \rightarrow \infty} \bar{Y}(n | y_0) = \theta$ and $\lim_{n \rightarrow \infty} S^2(n | y_0) = \sigma^2$, $\forall y_0 \geq 0$. However, because the observations are sequentially correlated, for finite n and arbitrary $y_0 \neq \theta$, both estimators are biased. Thus, there is bias in both the location (accuracy) and

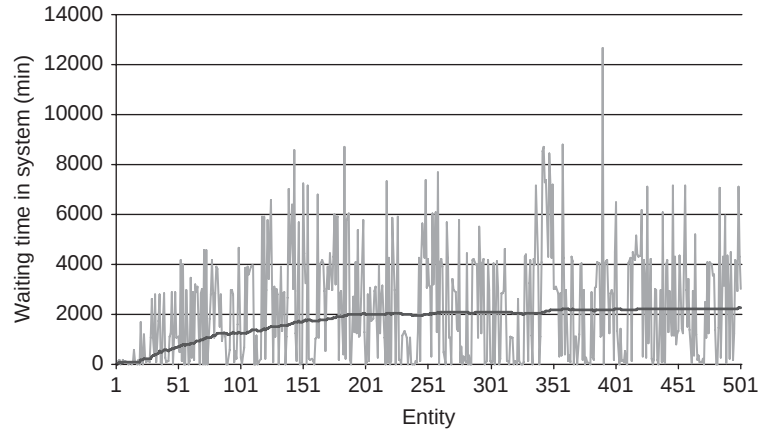


Figure 1. The user support model, waiting time in the system by entity as it departs the system and the cumulative average waiting time.

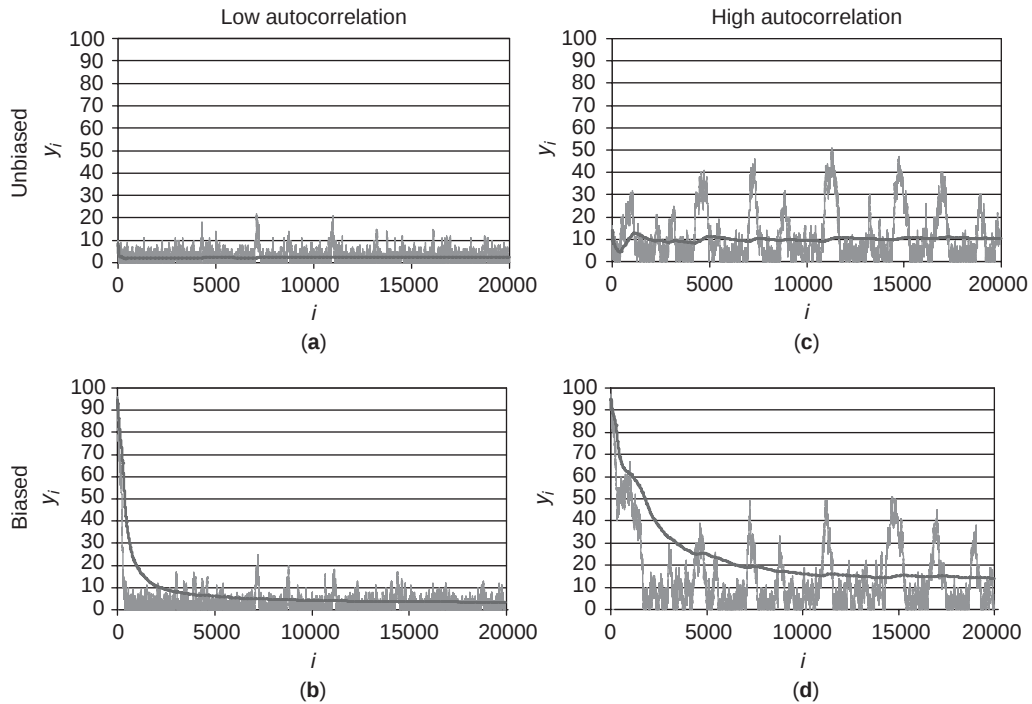


Figure 2. The time-domain effects of initial conditions and autocorrelation on the running estimate of mean number in system for two M/M/1 queues differing in traffic intensity.

width (precision) of the estimated confidence interval on the mean.

This is illustrated in the time-domain in Fig. 2. Each panel shows the sequence $\{y_i; i = 0, 1, 2, \dots, n\}$, the number in system

for an M/M/1 queue as a function of the observation number for a single simulation run, where observations are made at each change of state. The mean arrival rate is $\lambda = 1$ and the terminating event

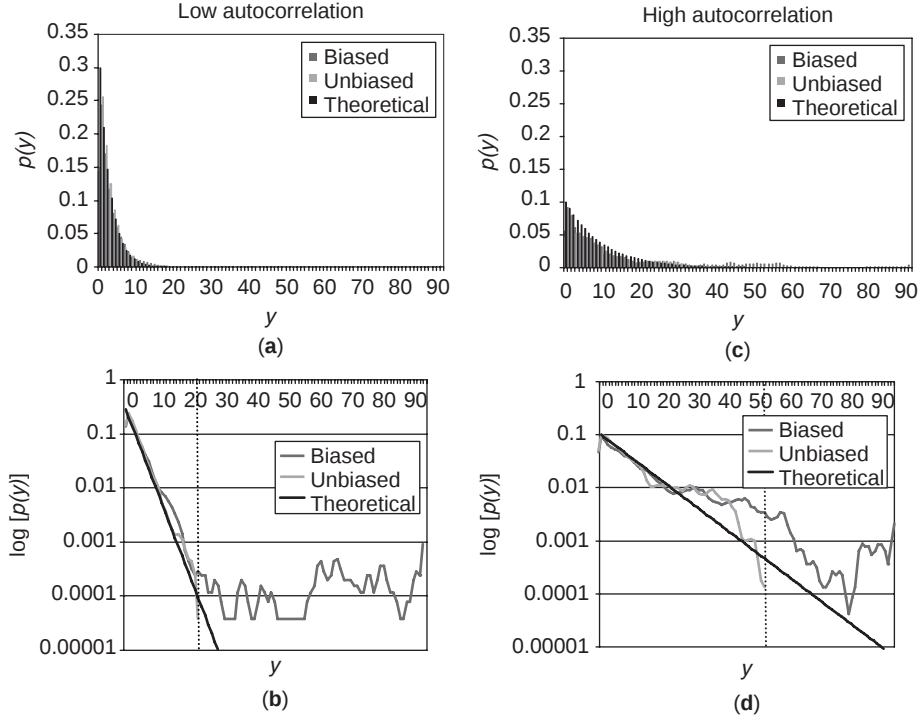


Figure 3. The frequency-domain effects of initial conditions and autocorrelation on the running estimate of mean number in system for two M/M/1 queues differing in traffic intensity.

is 10,000 departures, resulting in a run length $t_n \cong 10,000$ with $n \cong 20,000$ arrival and departure events. Also shown is the corresponding cumulative sample mean, the sequence $\{\bar{y}(i | y_0), i = 0, 1, \dots, n\}$. The queues in panels (a) and (b) have traffic intensities of $\rho = 0.7$ and these queues are initialized, respectively, at $y_0 = 90$ (a rare event, with $\Pr(Y_i \geq 90) = 6.86 \times 10^{-15}$) and $y_0 = 4$ (the approximate steady-state mean). The queues in panels (c) and (d) have traffic intensities of $\rho = 0.9$, initialized respectively at $y_0 = 90$ ($\Pr(Y_i \geq 90) = 6.86 \times 10^{-5}$) and $y_0 = 9$ (the steady-state mean). The bias introduced by the rare initial conditions is clear for both queues; the severity of this bias is clearly greater for the queue with higher traffic intensity (greater autocorrelation).

These effects are perhaps even more apparent in the (relative) frequency domain, as shown in Fig. 3. The stationary state for the queue is distributed geometrically, $Y_{ss} \sim \text{Geom}(1 - \rho)$, with mean $\theta = \rho/(1 - \rho)$ and

variance $\sigma^2 = \rho/(1 - \rho)^2$. Panels (a) and (c) compare the histograms of the estimated distributions for both the unbiased and biased initial conditions (for the entire run) to the probability mass function (pmf) of this theoretical distribution. These comparisons are repeated in panels (b) and (d), where the steady-state probabilities are now plotted on a logarithmic scale. This transforms the geometric pmf into a straight line and exaggerates the differences between the estimated and theoretical distributions at lower probabilities.

This example illustrates the fundamental issue. Every state is *positive recurrent*, $\Pr(Y_i = y) > 0$ for all $y \geq 0$ at steady state, and therefore the states themselves are *not* transient in the usual sense. For both the unbiased and biased cases, however, there is a transient period associated with the *sample mean*, $\bar{Y}(i, y_0)$, during which a sufficiently large number of observations must be collected for the sampling distribution

to approximate the true stationary distribution. For the biased case, there is also a transient period associated with the system state, Y_i , during which the state decreases from an improbable artificial initial value to subsequent values more representative of the steady-state operating regime. *Both* transients are prolonged by increasing autocorrelation in the values of the state. The duration of the mean transient is considerably longer than that of the state transient, because of the additional autocorrelation induced by calculating the cumulative mean.

With these observations, it is apparent that there is some minimum run length required to reduce the sampling error to an acceptable level for estimation of any statistic. This minimum run length increases with the degree of autocorrelation and is an inescapable artifact of the system under study. This minimum run length also increases with the increasing rarity of the initial value of the state. The problem of the initial transient, therefore, is to reduce the run length required to yield accuracy and precision in the sample statistics comparable to that in the unbiased case; or, alternatively, for a given computing budget, to improve the accuracy and precision in the sample statistics by removing the contribution of the state transient.

Dealing with the Problem

The problem of the initial transient has been studied extensively and two principle approaches exist for its mitigation. *Intelligent initialization* attempts to select initial conditions that are representative of the steady-state operating regime for each simulation run. In the simple case of the M/M/1 queue, for example, this implies initializing the number in system with a value close to the steady-state mean or mode [2,3]. If practical, this approach is always wise.

Banks *et al.* [4] offer two strategies for intelligent initialization—(i) observation and collection of data from the system under study, or a similar system, if such systems exist and (ii) solution of simplified, analytically tractable models approximating the

system or subsystems. Information regarding the choice of more representative initial conditions also may be available from pilot simulation runs undertaken during model development and verification. If the frame for output analysis employs multiple replications, Kelton [5] suggests that initial conditions be chosen at random from simple (uniform or binomial) probability distributions, as crude approximations to the stationary distributions. The parameters for these distributions can be based on expert opinion.

Intelligent initialization clearly has its limitations. Systems or comparable systems often simply do not exist. When these do, the time and expense of collecting the data required for intelligent initialization may be prohibitive. Analytically tractable models that adequately capture steady-state response also may be unavailable, or at least difficult to determine with any degree of confidence or fidelity.

Deletion or truncation uses the simulation itself to select initial conditions. This is accomplished by allowing the simulation to *warm-up* by running the simulation for some period of time, until a more representative state is achieved as the system finds its way into a more typical steady-state operating regime. Observations collected during this period are *deleted* in the computation of statistics for the remainder of the simulation run. This second approach can be employed as an alternative to intelligent simulation, or in conjunction with a judicious choice of initial conditions.

Using deletion, the steady-state mean estimator is

$$\bar{Y}(n, d | y_0) = \sum_{i=d+1}^k \omega_{i,d} Y_i,$$

where d is the length of the warm-up period and $\omega_i = 1/(n + 1 - d)$ and $k = n$ for variables which are tallied, or $\omega_i = (t_{i+1} - t_i)/(t_n - t_d)$ and $k = n - 1$ for variables which are time-averaged. The central issue in the deletion approach is to determine an appropriate value for d . If the period is too short, the effects of initialization may be reduced, but not eliminated. If the period is too long,

useable observations are discarded and the precision of the estimate suffers.

A very large number of methods have been proposed for determining the length of the warm-up period. These can be classified into five types [6]:

- *graphical methods* that plot the sample path and determine the truncation point by visual inspection;
- *initialization-bias tests* that directly identify bias in the sample path;
- *heuristics* that apply simple rules to the sample path;
- *statistical methods* that apply basic statistical principles (such as hypothesis testing) to the sample path;
- *hybrid methods* that typically combine graphical methods with heuristics.

Various methods have been proposed in each of these categories and are listed in Table 1. In addition to the original papers cited, surveys listed among the references provide further details and comparisons.

Graphical methods are almost certainly the most common in practice and one prevailing view is that visualization of the evolution of the output and statistics of interest is the easiest and most accurate strategy. Certainly, the human faculty for recognizing visual patterns should not be underestimated. To the extent practicable, individual and cumulative observations should be plotted and inspected visually as a part of any warm-up analysis.

Initialization-bias tests are not strictly the methods for identifying the warm-up period, but can be readily adapted to this end through repeated application to alternative truncated series. Heuristics and statistical methods are highly desirable, since these can be applied to suggest, refine, validate, and in some instances automate the choice of a truncation point. Rules and procedures that can be automated are being incorporated in some simulation software and are particularly useful when a simulation study requires a large number of runs. Given the widespread use of graphical methods and the growing importance of good heuristics, examples of these

two approaches are provided in the following sections.

GRAPHICAL METHODS AND WELCH'S PROCEDURE

Welch's graphical procedure [6,51–53] has been advanced as a general approach to aid in visualization. The following development parallels Law [54], who also provides recommendations for choosing parameters for the procedure. The procedure relies on first averaging observations across m independent replications, where each replication has length n observations and the same initial value y_0 . The result is the averaged process

$$\left\{ \bar{Y}_i(y_0) = \frac{1}{m} \sum_{j=1}^m Y_{ij}(y_0), i = 0, 1, 2, \dots, n \right\},$$

where $Y_{ij}(y_0)$ is the i th observation from the j th replication. Averaging preserves the transient mean, $E[\bar{Y}_i(y_0)] = E_j[Y_{ij}]$, while reducing the transient variance by a factor of m , $\text{Var}[\bar{Y}_i(y_0)] = \text{Var}_j[Y_{ij}(y_0)]/m$, for all $j = 1, \dots, m$. Five to ten replications are recommended initially.

Secondly, the averaged process is smoothed by a moving average with a window of length w , where w is a positive integer such that $w \leq \lfloor n/4 \rfloor$, as follows:

$$\bar{Y}_i(w, y_0) = \begin{cases} \frac{\sum_{s=-w}^w \bar{Y}_{i+s}(y_0)}{2w+1}; & \text{if } i = w, \dots, n-w, \\ \frac{\sum_{s=-(i-1)}^{i-1} \bar{Y}_{i+s}(y_0)}{2i-1}; & \text{if } i = 0, \dots, w-1. \end{cases}$$

The intent is to filter high-frequency oscillations, leaving a smooth moving-averaged process

$$\{\bar{Y}_i(w, y_0), i = 0, 1, 2, \dots, n-w\},$$

which closely approximates the transient cumulative mean for m sufficiently large.

Finally, the moving-averaged process is plotted and the length of the warm-up period

Table 1. Methods for Determining the Warm-Up Period [After Hoad *et al.* [7]]

Type	Method	References
Graphical	Simple time series inspection	8
	Ensemble (batch) average plots	4
	Cumulative-mean rule	2,4,8–15
	Deleting-the-cumulative-mean rule	11,12
	CUSUM plots	10
	Welch's procedure	4,14–20, Law and Kelton (2000)
	Variance plots (or Gordon rule)	2,8,9,17
	Exponentially weighted moving average control charts	21
	Statistical process control (SPC)	Law and Kelton (2000), 20,22
Heuristic	Ensemble (batch) average plots with Schriber's rule	2,3,17
	Conway rule or forward data-interval rule	2,3,9,17,20,23–27
	Modified Conway rule or backward data-interval rule	2,9,27,28
	Crossing-of-the-mean rule	2,3,9,17,20,27,28
	Autocorrelation estimator rule	2,17,29
	Marginal confidence rule or marginal standard error rules (MSER)	19,27,30
	Marginal standard error rule m , (e.g., $m = 5$, MSER-5)	15,20,30,31
	Telephone network rule	32
	Relaxation heuristics	11,12,17,19,33
	Beck's approach for cyclic output	34
	Tocher's cycle rule	17
	Kimble's double exponential smoothing method	33
	Euclidean distance (ED) method	28
	Neural networks (NN) method	28
Statistical	Goodness-of-fit test	17
	Algorithm for a static data set (ASD)	14
	Algorithm for a dynamic data set (ADD)	14
	Kelton and law regression method	11,12,16,17,19,33,35,36, Law and Kelton (2000)
	Glynn and Iglehart bias deletion rule	37
	Wavelet-based spectral method (WASSP)	38–40
	Queueing approximations method (MSEASVT)	41
	Chaos theory methods (methods M1 and M2)	42
	Kalman filter method	36, Law and Kelton (2000)
	Randomization tests for initialization bias	20,26
Initialization-bias tests	Schruben's maximum test (STS)	16,26,43–45, Law and Kelton (2000)
	Schruben's modified test	10,16,30,43, Law and Kelton (2000)
	Optimal test (Brownian bridge process)	17,33,44,46, Law and Kelton (2000)
	Rank test	46,47, Law and Kelton (2000)
	Batch means based tests–max test	30,42,48,49, Law and Kelton (2000)
	Batch means based tests–batch means test	30,45,48,49, Law and Kelton (2000)
	Batch means based tests–area test	45,48,49, Law and Kelton (2000)
	Ockerman and Goldsman students t -tests method	45

Table 1. (Continued)

Type	Method	References
Hybrid	Ockerman and Goldsman (t -test)	45
	compound tests	
	Pawlikowski's sequential method	17
	Scale invariant truncation (SIT) point method	50

is determined by visual inspection as the point at which the process “flattens out.” If this process is not sufficiently smooth to make a determination, w is increased and the moving-averaged process is recalculated. If w is at its limit and the process still is not sufficiently smooth, additional replications are required and the moving-averaged process again is recalculated. If the process does not appear to flatten, then n must be increased, or it is possible that the model does not reach a steady state.

Welch's procedure can be effective, but expensive to implement. If the process is highly variable, the run length n and the number of replications m required to capture the transient and achieve a smooth moving-averaged process can be large. If the computational intensity of the simulation is also large relative to the computing budget, then the $n \times m$ observations required to develop a clear understanding of the artificial initialization transient might be applied more effectively to observing the steady state. Additionally, visual determination of d remains subjective, as in all purely graphical methods, and cannot be directly automated.

Figure 4 illustrates these issues. Welch's procedure was applied to the M/M/1 queue with traffic intensity of $\rho = 0.9$ discussed above, with $n \cong 10,000$ events, $m = 10$ replications, and a window of $w = 1500$. The number in system is plotted for each replication as a function of the observation index i , as are the average and moving-averaged processes. Inspection of the moving-averaged process does not clearly indicate the length of the warm-up period. Indeed, what we might infer from this application is the truth that the cumulative mean approaches its steady-state value asymptotically from above. This is doubtless and very useful to know, but comes at considerable computational expense ($n \times m = 100,000$ observations). Moreover, the central issue of how to select a truncation point for subsequent output analysis remains unclear.

HEURISTICS AND MSER

Heuristics are the oldest nongraphical approaches to determining the length of the warm-up period. In many instances, these are based on the calculation of what

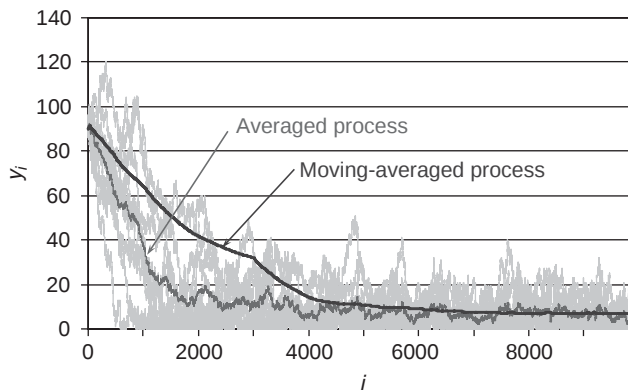


Figure 4. Results of the application of Welch's procedure (10 replications and a window of 1500) for mean number in system for an M/M/1 queue with traffic intensity $\rho = 0.9$.

amounts to a visual cue (such as the state trajectory crossing the mean trajectory), which is easy to understand and equally easy to compute. In a recent study, Hoad *et al.* [7] investigated the methods listed in Table 1. On the basis of criteria that included accuracy and robustness, simplicity, potential for automation, number of estimated parameters, and computation speed, they selected the MSER-5 heuristic as the most suitable for automation.

The original MSER [27,55] determines the length of the warm-up period by solving the following optimization problem:

$$\begin{aligned} d^* &= \arg \min_{n \gg d \geq 0} [MSER(n, d | y_0)] \\ &= \arg \min_{n \gg d \geq 0} \left[\frac{1}{(n-d)^2} \sum_{i=d+1}^n \right. \\ &\quad \left. \times (Y_i - \bar{Y}(n, d | y_0))^2 \right]. \end{aligned}$$

The intuition is that this choice minimizes the width of marginal confidence interval on the mean. While this interval is calculated using the sample variance of the reserved sequence (the sequence remaining after truncation) as an estimator for the variance, Franklin and White [31] have shown that substituting an unbiased estimator is less efficient and does not improve the performance of the heuristic. Because the state transient is much shorter than the mean transient, truncation at d^* in effect eliminates initialization bias and (approximately) minimizes the mean-squared error in the estimate of the mean.

MSER- m [30,56] is an improvement in MSER, which replaces the output series $\{Y_i; i = 0, 1, \dots, n\}$ with the series of batch means $\{Z_j, j = 0, 1, \dots, \lfloor n/m \rfloor\}$, where

$$Z_j = (1/m) \sum_{p=1}^m Y_{m(j-1)+p}. \quad (1)$$

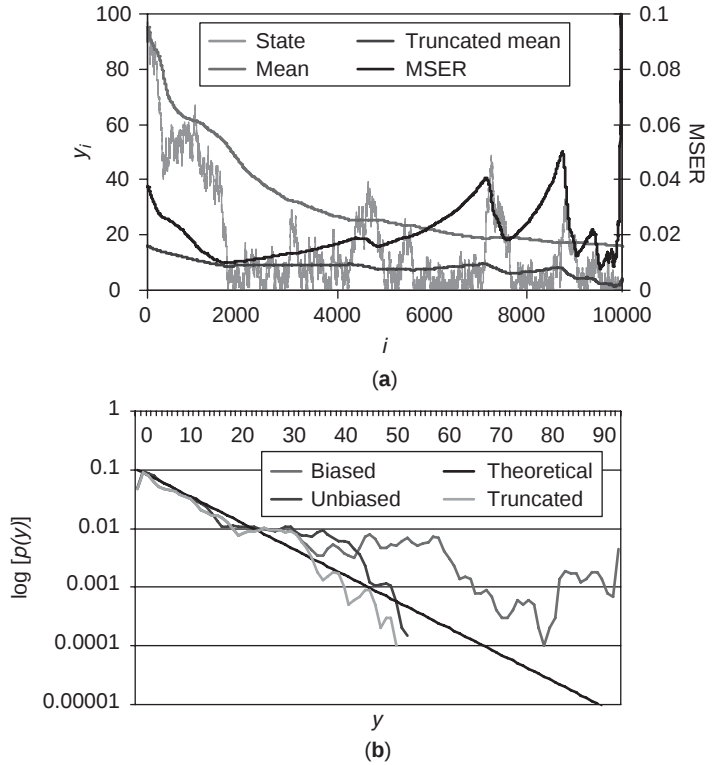


Figure 5. Results of the application of MSER-5 for mean number in system for an M/M/1 queue with traffic intensity $\rho = 0.9$.

The length of the batch is somewhat arbitrary, but the value $m = 5$ has shown to yield superior results over a wide range of test problems.

Figure 5 illustrates the application of MSER-5 to the M/M/1 queue with traffic intensity of $\rho = 0.9$ with $n \cong 10,000$ events. In addition to the output sequence and cumulative mean, panel (a) also shows the truncated mean $\bar{Y}(n, d | y_0)$ and $\text{MSER}(n, d | y_0)$ statistic as functions of the truncation point $d = i$. The minimum value of the statistic over the initial component of the run obtains at $d^* = 1607$ ($t_{d^*} = 781.8$), with $Y_{d^*+1} = 18$ as the initial value of the reserved sequence and $Y(n, d^* + 1 | y_0) = 8.871$ as the truncated point estimate of the mean.

In addition to the biased, unbiased, and theoretical log-densities, panel (b) shows the log-density of the reserved sequence for the MSER-5 truncation point. Clearly the state transient has been removed. Equally clearly, there remains a substantial mean transient, which overrepresents observations in the range of 21–47 and causes a high estimate of the mean for this run. This is true for the unbiased and biased sequences as well, and is not an artifact of initialization. Given the high autocorrelation, a significantly longer run length is required to provide reasonable precision in the estimate.

HYBRID WELCH/MSER

A principle shortcoming of Welch's procedure—that it yields a subjective determination of the truncation point by visual inspection—can be overcome by applying a heuristic to the moving-averaged process. For example, Fig. 6 plots the moving-averaged process from Fig. 4, together with the truncated mean and MSER statistic for this process. MSER recommends a truncation point of $d^* = 6653$, corresponding to a value of $\bar{Y}_i(w, y_0) = 7.9212$. Adopting an initial condition of $y_0 = 8$ for subsequent output analysis is a very reasonable choice.

How many Replications?

There are two standard frameworks for steady-state output analysis. *Replication/deletion* uses m independent replications, deleting the mean transient from each. Its advantage is that the mean estimates from each replication are uncorrelated and easily computed. Its disadvantage is inefficiency. The mean transient is repeated in each replication at a cost of $d \times m$ deleted observations, if the same warm-up period is used for each replication (which is typically the case). This is compounded by the need for a conservative value for d , which is sufficiently large to capture the transient for each replication.

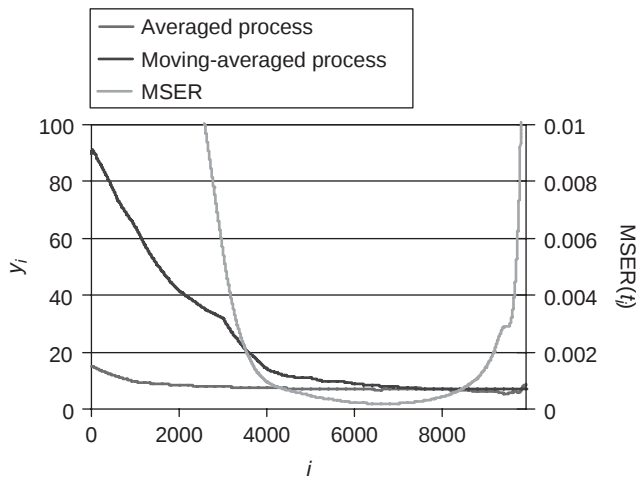


Figure 6. Results of the application of MSER to the moving-averaged process developed using Welch's procedure for an M/M/1 queue with traffic intensity $\rho = 0.9$ initialized at 90 in system.

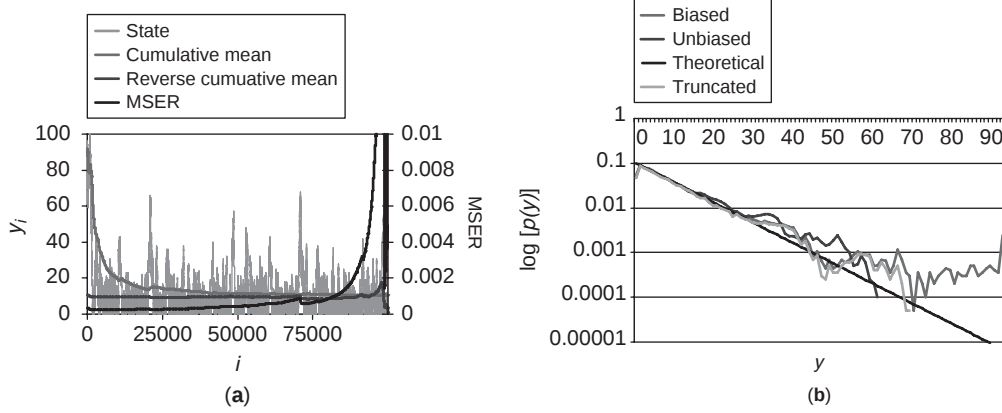


Figure 7. Results of the application of MSER-5 to a single long run for an M/M/1 queue with traffic intensity $\rho = 0.9$ initialized at 90 in system.

The alternative to replication/deletion is a *single, long run*, with the advantage that the transient is identified and deleted only once. The disadvantage is the need to construct an unbiased estimator for the variance, which corrects for autocorrelation of the observations. This is most often accomplished by the method of *batch means*. The reserved time series is divided into some number (typically 20–40) sequential batches, the mean is computed for each batch, and the variance of the batch means is used to estimate the confidence interval on the mean. This approach is effective if the batches are sufficiently large to result in uncorrelated batch means. While many authors express a strong preference for replication/deletion approach, the method of batch means can be vastly more efficient [57].

We illustrate this observation by applying MSER directly to one long run of the example M/M/1 simulation, given the same computing budget of $n \cong 100,000$ observations employed in the determination of a warm-up period applying Welch's procedure. The results are shown in Fig. 7. MSER-5 recommends a truncation point of $d^* = 3350$, approximately half that determined previously for the moving-averaged sequence. Applying batch means to the reserved sequence with 20 batches, the estimated 95% confidence interval on the mean is

$\bar{Y}(n, d | y_0) = 9.5061 \pm 1.774$. Note that in applying Welch's procedure, the entire budget was expended in developing a clear sense of the initial transient, in order to determine a suitable initial condition for further output analysis. In contrast, using this same budget to generate a single, long replication, and applying MSER-5 to the result, yield point and interval estimates of the mean without additional computation. In our research, this economy has been consistent across all instances of a wide range of test problems.

CONCLUSIONS

This article describes the problem of the initial transient that arises when estimating the limiting distribution of a nonterminating stochastic process. Concepts are explored with reference to simple M/M/1 queueing model with an artificially induced initial bias. A range of solution methods have been proposed and here we explore two such approaches, graphical and heuristic. We also demonstrate how these methods might be combined. Results from test problems consistently show that truncation of output data from a single long run is more economical than the replication/deletion approach for nonterminating simulations.

REFERENCES

1. Robinson S. Soft with a hard centre: discrete-event simulation in facilitation. *J Oper Res Soc* 2001;52(8):905–915.
2. Wilson JR, Pritsker AAB. A survey of research on the simulation startup problem. *Simulation* 1978a;31(2):55–58.
3. Wilson JR, Pritsker AAB. Evaluation of startup policies in simulation experiments. *Simulation* 1978b;31(3):79–89.
4. Banks J, Carson JS, Nelson BL, *et al.* Discrete-event system simulation. 4th ed. New Jersey: Prentice-Hall; 2001.
5. Kelton W.D., Random initialization methods in simulation. *IIE Transactions*; 1989;21:355–367.
6. Robinson S. Simulation: the practice of model development and use. England: John Wiley & Sons, Inc.; 2004.
7. Hoad K, Robinson S, Davies R. Automating warm-up length estimation. Proceedings of the 2008 Winter Simulation Conference; Miami (FL): 2008. pp. 532–540.
8. Gordon G. System simulation. New Jersey: Prentice-Hall; 1969.
9. Gafarian AV, Ancker CJ Jr, Morisaku T. Evaluation of commonly used rules for detecting 'steady state' in computer simulation. *Nav Res Log Q* 1978;25:511–529.
10. Nelson BL. Statistical analysis of simulation results. Handbook of industrial engineering. 2nd ed. New York: John Wiley & Sons, Inc.; 1992.
11. Roth E, Josephy N. A relaxation time heuristic for exponential-Erlang queuing systems. *Comput Oper Res* 1993;20(3):293–301.
12. Roth E. The relaxation time heuristic for the initial transient problem in M/M/k queuing systems. *Eur J Oper Res* 1994;72:376–386.
13. Fishman GS. Discrete-event simulation: modeling, programming, and analysis. New York: Springer; 2001.
14. Bause F, Eickhoff M. Truncation point estimation using multiple replications in parallel. Proceedings of the 2003 Winter Simulation Conference; New Orleans (LA): 2003. pp. 414–421.
15. Sandikci B, Sabuncuoglu I. Analysis of the behavior of the transient period in non-terminating simulations. *Eur J Oper Res* 2006;173:252–267.
16. Law AM. Statistical analysis of simulation output data. *Oper Res* 1983;31:983–1029.
17. Pawlikowski K. Steady-state simulation of queuing processes: a survey of problems and solutions. *Comput Surv* 1990;122(2):123–170.
18. Alexopoulos C, Seila AF. Output data analysis. Handbook of simulation. New York: Wiley; 1998. pp. 225–272.
19. Linton JR, Harmonosky CM. A comparison of selective initialization bias elimination methods. Proceedings of the 2002 Winter Simulation Conference; San Diego (CA): 2002. pp. 1951–1957.
20. Mahajan PS, Ingalls RG. Evaluation of methods used to detect warm-up period in steady state simulation. Proceedings of the 2004 Winter Simulation Conference; Washington (DC): 2004. pp. 663–671.
21. Rossetti MD, Li Z, Qu P. Exploring exponentially weighted moving average control charts to determine the warm-up period. Proceedings of the Winter Simulation Conference; Orlando (FL): Volume 539; 2005. pp. 771–780.
22. Robinson S. A statistical process control approach to selecting a warm-up period for a discrete-event simulation. *Eur J Oper Res* 2005;176:332–346.
23. Conway RW. Some tactical problems in digital simulation. *Manag Sci* 1963;10(1):47–61.
24. Fishman GS. Concepts and methods in discrete event digital simulation. New York: Wiley; 1973.
25. Bratley P, Fox B, Schrage L. A guide to simulation. 2nd ed. New York: Springer; 1987.
26. Yucesan E. Randomization tests for initialization bias in simulation output. *Nav Res Log* 1993;40:643–663.
27. White KP Jr. An effective truncation heuristic for bias reduction in simulation output. *Simulation* 1997;69(6):323–334.
28. Lee Y-H, Kyung K-H, Jung C-S. On-line determination of steady state in simulation outputs. *Comput Ind Eng* 1997;33(3):805–808.
29. Fishman GS. Estimating sample size in computing simulation experiments. *Manag Sci* 1971;18:21–38.
30. White KP Jr, Cobb MJ, Spratt SC. A comparison of five steady-state truncation heuristics for simulation. Proceedings of the 2000 Winter Simulation Conference. Orlando (FL): 2000. pp. 755–760.
31. Franklin WW, White KP Jr. Stationarity tests and MSER-5: exploring the intuition behind mean-squared-error reduction in detecting and correcting initialization bias. Proceedings

- of the 2008 Winter Simulation Conference; Miami (FL): 2008. pp. 541–546.
32. Zobel CW, White KP Jr. Determining a warm-up period for a telephone network routing simulation. Proceedings of the 1999 Winter Simulation Conference; Phoenix (AR): 1999. pp. 662–665.
 33. Kimbler DL, Knight BD. A survey of current methods for the elimination of initialization bias in digital simulation. *Ann Simul Symp* 1987;20:133–142.
 34. Beck AD. Consistency of warm up periods for a simulation model that is cyclic in nature. Proceedings of the Simulation Study Group, OR Society; Volume 538; 2004. pp. 105–108.
 35. Kelton WD, Law AM. A new approach for dealing with the startup problem in discrete event simulation. *Nav Res Log Q* 1983;30:641–658.
 36. Gallagher MA, Bauer KW Jr, Maybeck PS. Initial data truncation for univariate output of discrete-event simulations using the Kalman Filter. *Manag Sci* 1996;42(4):559–575.
 37. Glynn PW, Iglehart DL. A new initial bias deletion rule. Proceedings of the 1987 Winter Simulation Conference; Atlanta (GA): 1987. pp. 318–319.
 38. Lada EK, Wilson JR, Steiger NM. A wavelet-based spectral method for steady-state simulation analysis. Proceedings of the 2003 Winter Simulation Conference; New Orleans: 2003. pp. 422–430.
 39. Lada EK, Wilson JR, Steiger NM, *et al.* Performance evaluation of a wavelet-based spectral method for steady-state simulation analysis. Proceedings of the 2004 Winter Simulation Conference; Orlando (FL): 2004. pp. 694–702.
 40. Lada EK, Wilson JR. A wavelet-based spectral procedure for steady-state simulation analysis. *Eur J Oper Res* 2006;174:1769–1801.
 41. Rossetti MD, Delaney PJ. Control of initialization bias in queuing simulations using queuing approximations. Proceedings of the 1995 Winter Simulation Conference; Arlington (VA): 1995. pp. 322–329.
 42. Lee Y-H, Oh H-S. Detecting truncation point in steady-state simulation using chaos theory. Proceedings of the 1994 Winter Simulation Conference; Orlando (FL): 1994. pp. 353–360.
 43. Schruben LW. Detecting initialization bias in simulation output. *Oper Res* 1982;30(3):569–590.
 44. Schruben L, Singh H, Tierney L. Optimal tests for initialization bias in simulation output. *Oper Res* 1983;31(6):1167–1178.
 45. Ockerman DH, Goldsman D. Student t-tests and compound tests to detect transients in simulated time series. *Eur J Oper Res* 1999;116:681–691.
 46. Ma X, Kochhar AK. A comparison study of two tests for detecting initialization bias in simulation output. *Simulation* 1993;61(2):94–101.
 47. Vassilacopoulos G. Testing for initialization bias in simulation output. *Simulation* 1989;52(4):151–153.
 48. Cash CR, Dippold DG, Long JM, *et al.* Evaluation of tests for initial-condition bias. Proceedings of the 1992 Winter Simulation Conference; Arlington (VA): 1992. pp. 577–585.
 49. Goldsman D, Schruben LW, Swain JJ. Tests for transient means in simulated time series. *Nav Res Log* 1994;41:171–187.
 50. Jackway PT, deSilva BM. A methodology for initialization bias reduction in computer simulation output. *Asia Pac J Oper Res* 1992;9:87–100.
 51. Welch PD. The statistical analysis of simulation results. In: Lavenberg SS, editor. *The computer performance modeling handbook*. New York: Academic Press; 1983.
 52. Banks J. *Getting started with automod*. Bountiful (UT): AutoSimulations, Inc.; 2000.
 53. Rossetti MD. *Simulation modeling and arena*. New York: Wiley; 2010.
 54. Law AM. *Simulation modeling and analysis*. 4th ed. New York: McGraw-Hill; 2007.
 55. McClarnon MA. Detection of steady-state in discrete-event dynamic systems: an analysis of heuristics [MS thesis]. Department of Systems Engineering, University of Virginia; 1990.
 56. Spratt SC. An evaluation of contemporary heuristics for the startup problem [MS thesis]. Faculty of the School of Engineering and Applied Science, University of Virginia; 1998.
 57. Alexopoulos C, Goldsman D. To batch or not to batch? *ACM Trans Model Comput Simul* 2004;14(1):76–114.

INSPECTION GAMES

RYUSUKE HOHZAKI
Department of Computer Science
National Defense Academy,
Yokosuka, Japan

INTRODUCTION

Dresher [1] began the inspection game to analyze the treaty of arms reduction in 1962. In the field of compliance of treaties or laws, the expansion of the inspection game has three phases in accordance with the situation of international relations in the world. Dresher's work belongs to Phase I in the 1960s under the commission of the US Arms Control and Disarmament Agency (ACDA). Dresher proposed a recursive game to analyze the effectiveness of on-site inspections for the treaty. Maschler [2] generalized Dresher's modeling. They considered the game where a player called *a violator* wished to violate the treaty in secret for his national interest, and the other player called *an inspector* wanted to deter illegal behaviors of the violator by effective inspection. In their models, the violator must pay one of the penalties if his violation is exposed by an inspection but he can escape the exposure by side payment of penalty q . Dresher discussed special cases of $q = 1/2$ and $q = 1$, and Maschler did a general case of $0 \leq q \leq 1$. As well as Dresher and Maschler, Kuhn [3] contributed to the game-theoretical analysis on Nuclear test ban treaty in Phase I.

Phase II has a span from the late 1960s to the 1980s. In this phase, they discussed the compliance of the Non-Proliferation Treaty for Nuclear Weapons. The International Atomic Energy Agency (IAEA) is an international organization founded in 1957 to promote safe, secure, and peaceful nuclear sciences and technology all over the world. The treaty on the Non-Proliferation of Nuclear Weapons (NTP) makes use of the

inspection the IAEA supplies to prevent the spread of nuclear weapons and technology. Bierlein [4], Jaech [5], Avenhaus [6], and Bowen and Bennett [7] promoted the application of game theory to such treaties as the NTP.

In Phase III after the Cold War, the inspection game was developed to deal with new disarmament treaties, for example, the treaty on intermediate nuclear forces and the treaty on conventional forces in Europe, in the mid-1980s. Brams and Davis [8], Kilgour [9], Ruckle [10], and Avenhaus and Cauty [11] belong to this phase.

The level on frequency of inspections by the IAEA is precisely prescribed in agreement such as the Information Circulars 153. On the basis of its scientific rules, most of the inspections were carried out in Japan, Germany, and Canada until 1993 because these countries have many nuclear-powered electric plants. These countries were believed to have no intention or no motivation to develop nuclear weapons. On the other hand, the inspection cost accounts for a large part of the annual budget of the IAEA, and the inspections were thought to be a waste of the money. After that, there occurred serious problems such as Dr. Khan's black market and we had the suspicion that some countries were developing nuclear weapons using nuclear technologies in black market. Through these problems, the IAEA changed its policy and tried to strengthen the safeguard system by additional protocols and NTP Comprehensive Safeguard Agreement. Some researchers analyzed the intention of inspectees to make illegal actions from the game-theoretical point of view, for example, Avenhaus [12], Avenhaus and Kilgour [13], Avenhaus and Cauty [14], and Hohzaki [15].

Dresher's research was branched out to the so-called smuggling game. Thomas and Nisgav [16] extended Maschler's model to a game with Customs and a smuggler, in which Customs patrols for illegal actions of the smuggler by using one or two patrol

boats. They formulated the problem into a multistage recursive game and adopted a simple numerical method to repeat solving a one-stage matrix game step-by-step. The smuggling game reflects real serious problems that the United States faced during the 1980s through the 1990s. The US Customs Service, the Drug Enforcement Agency, and the Department of Defense were eager to interdict the smuggling of drug into the United States by inspection or surveillance around the border between the United States and other countries. At that time, they focused on network interdiction games; for example, see Mitchell and Bell [17], Caulkins *et al.* [18], and Washburn and Wood [19]. Washburn and Wood constructed their model on a network and applied graph theory as well as game theory to solve the problem. The network interdiction problem could be applied to many problems in the real world, such as tactical air strikes, interdiction of drug invasion, controlling infections in a hospital, chemically treating raw sewage, or flood control. Many researchers have studied on this topic, such as McMasters and Mustin [20], Ghare *et al.* [21], Fulkerson and Harding [22], Golden [23], Gal [24], Wood [25], and Whiteman [26]. The ambush game is a model that is similar to the interdiction game [27].

Not related to any network structure, Baston and Bostock [28] first gave us a closed form of equilibrium point for the typical type of inspection game as an extension of Dresher's model. Furthermore, they modified the perfect-capture assumption, that Customs certainly captures the smuggler when both players meet, to the nonperfect-capture model. They succeeded in solving the game by introducing capture probability depending on the number of patrol boats. However, the opportunity for the smuggler to ship its cargo of contraband is still assumed to be once at most, as assumed in the preceding papers. Garnaev [29] extended their work to a model with three patrol boats.

Compared with these studies, Sakaguchi [30] introduced the assumption that the smuggler possibly takes several actions in the perfect-capture model. His model is a

kind of repeated game with several opportunities of smuggling and then the value of the game is given by multiplying the value of the once-opportunity game by the number of opportunities. Ferguson and Melolidakis [31] is an extension of Sakaguchi's model. They assumed that the smuggler can get rid of the capture by paying some cost $q (\leq 1)$. He must pay unit cost 1 on being captured but he loses nothing if not captured. Hohzaki *et al.* [32] also extended Sakaguchi's model such that the capture of smuggler terminates the game, and the encounter of the smuggler and Customs stochastically results in one of capture, success of smuggling, or no-event. Almost all researches adopted the assumption that the smuggler has two options (smuggling or not-smuggling) as his strategy. Hohzaki [33] first analyzed the smuggling game with the smuggler's decision on the amount of contraband. In almost all past models, they assumed that players know the behavior their opponents took in the past. If information about their opponents is perfectly hidden to competitors, players have no chance to change their strategies, and therefore the problem would become a one-shot game. Hohzaki and Maebara [34] considered a one-shot smuggling game during a finite number of days without getting any information.

In all models surveyed so far, the information acquisition is symmetric between players. Harsanyi [35] made us notice that information is crucial to the results of the game, and proposed the concept of complete information and incomplete information. The concept has been applied to a variety of game models. In the smuggling game, it would be natural that players acquire information about their opponents in an asymmetric manner because Customs is generally thought to be a public organization and the smuggler would be a secret society. Under the situation, the information acquisition must be asymmetrical between players. Hohzaki [36] first dealt with a smuggling game with asymmetric information, where Customs decides to patrol or not to patrol without any smuggler's information, but the smuggler chooses his strategy, knowing the past decision of Customs. He proposed a

methodology to derive Bayesian equilibrium and evaluated the value of information.

We have traced the history of the inspection game stemming from Dresher's work. From this point, we first focus on some models of the smuggling game and explain its basics and methodology to derive Nash equilibriums while paying attention to the value of information. And we look at the models of the interdiction game and the ambush game.

A BASIC MODEL OF INSPECTION GAME WITH MULTIPLE STAGES

We consider the following two-person zero-sum (TPZS) game with Customs and a smuggler, taken in Reference [32], because it is more general and more applicable than Dresher's original modeling.

- A1. Customs and a smuggler compete with each other during N days, on each of which two players take an action once at most. We denote each stage by the number of residual days until the last day like $N = \{N, N-1, \dots, 1\}$.
- A2. Customs can afford to patrol K times at most and the smuggler can smuggle L times at most. In the case of $K > N$ or $L > N$, players lose excess opportunities of patrolling for Customs and smuggling for the smuggler.
- A3. At each stage, Customs has two options: patrol (P) or not-patrol (NP). The smuggler also has a two-option strategy: smuggling (S) or not-smuggling (NS).
- A4. If the smuggler smuggles and Customs dispatch a patrol boat on the same day, Customs can capture the smuggler with probability p and, at the same time, the smuggler succeeds in smuggling with probability q . With the residual probability of $1 - p - q$, there happens nothing. In the case of no patrol, the smuggler successfully fulfills his smuggling.
- A5. Customs gains reward $\alpha > 0$ by the capture of the smuggler but loses reward 1 on success of smuggling by the smuggler. We assume $\alpha p - q > 0$

because the smuggler dare not smuggle on the patrol day of Customs.

The payoff of the game is zero-sum, which means that the reward of Customs is the same amount of loss for the smuggler and vice versa. We define the payoff of the game by the Customs' reward.

- A6. If Customs does not capture the smuggler at a stage, the game transfers to the next stage, and the number of stage decreases by one. The game ends on capture of smuggler or at the last stage with no residual day.
- A7. Assumptions A1 through A6 are common knowledge for both players. Each player knows the strategies taken by his opponent at the past stages.

Our TPZS game model is slightly complicated than Baston and Bostock's model or Sakaguchi's model. However, the methods or formulations for equilibrium points are very similar, although the formulations have individual difficulties for their solutions. The multistage inspection game is usually solved through the construction of a recursive game and a system of recursive equations or difference equations. Now let us proceed to the next step to construct the recursive game for our model.

By $\Gamma(n, k, l)$, we denote the game starting from Stage n , at which Customs has k opportunities to patrol and the smuggler can smuggle l times at most, and by $v(n, k, l)$, we denote the value of the game $\Gamma(n, k, l)$. Considering each combination of two players' strategies at Stage n and its resultant situation, we can represent the value $v(n, k, l)$ by the value of the following recursive 2×2 matrix game.

$$v(n, k, l) = \text{val} \begin{pmatrix} \alpha p - q + (1-p)v(n-1, k-1, l) & v(n-1, k-1, l-1) \\ -1 + v(n-1, k, l-1) & v(n-1, k, l) \end{pmatrix}, \quad (1)$$

where val represents the value of its sequent matrix game. Two columns of the left and the right represent the smuggler's strategies: S and NS, respectively. The upper row and the

lower row represent the Customs' choices: P and NP, respectively. Taking the combination of P and S, for example, Customs can expect reward $\alpha p - q$ at the present stage. The game ends at this stage with probability p if the smuggler is captured; otherwise it steps forward to the next stage with probability $1 - p$ to transfer to a new stage game $\Gamma(n - 1, k - 1, l - 1)$. Other elements of the payoff matrix are also explainable.

Starting from initial conditions of $v(0, k, l) = 0$, $v(n, k, 0) = 0$, and $v(n, 0, l) = -l$ and using Equation 1, we can numerically calculate $v(n, k, l)$ for any n , k , and l . Hohzaki *et al.* [32] proved that four elements in Equation 1, substituted by four alphabets a , b , c , and d , have the following relation:

$$v(n, k, l) = \text{val} \left(\begin{array}{ccc} a & > & c \\ \vee & & \wedge \\ b & \leq & d \end{array} \right),$$

and that any optimal strategy always becomes a mixed one. The value of the game is given by $v(n, k, l) = (ad - bc)/(a - b - c + d)$, and the expression is regarded as a difference equation held between $v(n, \cdot)$ and $v(n - 1, \cdot)$. In some cases, we can solve the difference equation with respect to three indices n , k , and l , and have an analytical form of solution. In the case of $l = 1$, $v(n, k, 1)$ is given by

$$v(n, k, 1) = - \binom{n-1}{k} / \sum_{r=0}^k (\alpha p - q)^{k-r} \binom{n}{r} \quad (2)$$

In this case, the probability for Customs to patrol is kept the same through all stages by his optimal strategy so that the smuggler cannot anticipate the most risky day for him when Customs is most likely to patrol. The solution 2 is a generalization of the results that Baston and Bostock [28] and Sakaguchi [30] obtained. In other cases of $l > 1$, we can derive their results as special cases of our model by setting $\alpha p - q = 1$.

THE VALUE OF INFORMATION IN THE INSPECTION GAME

We have overviewed the general basic model of the inspection game in Section 2, where

both players choose their actions after knowing their opponents' past strategies. The information the players get makes a lot of effects on the results of the game, as Harsanyi [35] pointed out. Here, we take two models for analysis: a no-information model in which both players do not get any information about their opponents, and an asymmetric information model in which there occurs an asymmetry between the information of both players.

A Model with No Information

Let us first discuss the inspection game with no information. We precisely define the change of the assumption from the basic model in Section 2. We replace Assumption A7 with the assumption of no information, as follows.

- S1. Each player makes his decision at each stage without knowing any strategy of his opponent in the past. However, both players can recognize the occurrence of capture of the smuggler.

Even under Assumption S7, it is natural that both players distinguish between capture and no-capture, and the game steps forward to the next stage only if the smuggler is not captured. The small change of the assumption makes the problem a one-shot game because all players have no chance to revise their strategies in process of the game without any information to use and would make their decisions once in advance of the game. We derived a closed-form expression for the value of the multistage game in a special case in Section 2. For the one-shot game, we can pursue more elaborate methodology to derive equilibriums.

For all stages N , we denote an pure strategy of Customs by $\mathbf{x} = (x_N, x_{N-1}, \dots, x_1)$, where x_j is 1 for P and 0 for NP at the Stage j and satisfies $\sum_j x_j \leq K$, and a smuggler's pure strategy by $\mathbf{y} = (y_N, y_{N-1}, \dots, y_1)$, where y_j is 1 for S and 0 for NS at the stage j and has $\sum_j y_j \leq L$. For a combination of a Customs' pure strategy $\mathbf{x}$ and a smuggler's pure strategy $\mathbf{y}$, we can evaluate the payoff by

$$R(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^N (1-p)^{T_n(\mathbf{x}, \mathbf{y})} y_n \{x_n(\alpha p - q + 1) - 1\}, \quad (3)$$

where $T_n(\mathbf{x}, \mathbf{y}) \equiv \sum_{j=n+1}^N x_j y_j$. $(1-p)^{T_n(\mathbf{x}, \mathbf{y})}$ indicates the survival probability of the smuggler until Stage n .

Because there are a finite number of feasible pure strategies for both of $\mathbf{x}$ and $\mathbf{y}$, we can numerically solve the matrix game with the payoff $R(\mathbf{x}, \mathbf{y})$ if we do not mind the size of the problem. However, this model looks interesting as an example of application of dynamic programming (DP) and probability theory such that each player estimates his opponent's mixed strategy $\mathbf{x}$ or $\mathbf{y}$ stage-by-stage by using the information of capture or no-capture. The smuggler who took S strategy at the present stage n might revise the estimated probability of Customs' taking strategy $\mathbf{x}$, denoted by $\rho(\mathbf{x})$, conditioned on no-capture by

$$\Gamma_n \rho(\mathbf{x}) = \frac{\rho(\mathbf{x})(1 - p x_n)}{1 - p \Pr(\text{Customs takes any strategy with } x_n = 1)}$$

at the next stage. On the other hand, the smuggler who took NS would not revise the probability $\rho(\mathbf{x})$ because the capture never occurs. By DP formulation, we can find the rational behavior of the smuggler based on the revised probability on the Customs' mixed strategy stage-by-stage and finally derive an optimal Customs' mixed strategy as an equilibrium point. The procedure is equivalent to solving the maximin optimization problem with respect to the payoff $R(\mathbf{x}, \mathbf{y})$. In a similar manner, we can find the rational behavior of Customs while revising the estimation on the mixed strategy of the smuggler stage-by-stage and then finally derive an optimal smuggler's strategy, which is the same as an optimal solution of the minimax optimization problem. Anyway, the readers could refer to Reference [34] for more details.

A Model with Asymmetric Information

Let us discuss the second model of the inspection game, where information acquisition is asymmetric between players; Player B can

obtain some information but Player A cannot. In such a situation, Player A has to make a decision based on his belief on the characteristics of his opponent, the present state of the game, and others. In the model with incomplete information, the payoff of the game is expressed in different ways, relying on individual belief of each player. Player B can definitely recognize his reward or the payoff based on his correct information, but Player A would have some uncertainty in his payoff even though the payoff has the same meaning for both players. In many countries, Customs would be a public organization but the smuggler could be a secret society. Therefore, the behavior of Customs is comparatively open to outside but the smuggler's information is kept in secret. For the game with incomplete information, game theory presents us the concept of Bayesian–Nash equilibrium, by which we can understand the rationality on the player's behavior, taking account of the belief. Here, we are going to discuss the way to handle the inspection game with incomplete information or asymmetric information, in accordance with Reference [36]. The Bayesian–Nash equilibrium is the extended concept of Nash equilibrium, in which the entire combination of players' strategies and their beliefs has to be consistent with each other.

To deal with the incomplete information, we change Assumption A7 in Section 2 to the following.

- I1. At the beginning of a stage, the smuggler knows the strategy taken by Customs at the previous stage but the smuggler's past strategy is in secret to Customs. Both players recognize the occurrence of capture of the smuggler. The initial state of the game, (N, K, L) , is common knowledge for both players.

We can denote an exact state of the game by (n, k, l) , where Customs holds k chances to patrol and the smuggler has l chances to smuggle at Stage n . At the beginning of the stage n , the smuggler correctly recognizes the state (n, k, l) . But Customs knows only (n, k) and anticipates the number l by his belief on it. Let us denote his belief

by a probability distribution $q_n = \{q_n(l), l = \underline{l}_n, \underline{l}_n + 1, \dots, \bar{l}_n\}$, where $q_n(l)$ is the probability that the smuggler has l chances of smuggling left and satisfies conditions $\sum_l q_n(l) = 1$ and $q_n(l) \geq 0$. The lower bound $\underline{l}_n$ of l is estimated by everyday smuggling, and the upper bound $\bar{l}_n$ is calculated by no smuggling, as follows: $\underline{l}_n = \max\{L - N + n, 0\}$, $\bar{l}_n = \min\{L, n\}$. We refer to the smuggler with l chances of smuggling as the l -type smuggler. Customs has the perfect belief of

$$q_N(L) = 1, \quad q_N(l) = 0 (l \neq L) \quad (4)$$

only at the first stage N .

Before discussing the difference between payoffs of two players, we are going to define strategies of players. For Customs knowing the state (n, k) , we denote the probability of taking P or NP by x_n or $1 - x_n$, respectively. For the l -type smuggler recognizing the state (n, k, l) , we use $y_n(l)$ or $1 - y_n(l)$ as the probability of taking smuggling (S) or not-smuggling (NS). We illustrate a stage game in State (n, k, l) by Table 1, similarly to a matrix game.

Table 1. A Table of Payoffs in State (n, k, l)

Belief	$q_n(\underline{l}_n)$	$q_n(l)$	$q_n(\bar{l}_n)$
		...	$y_n(l), 1 - y_n(l)$
	C/S	S, NS	...
		S, NS	S, NS
x_n	P	...	(1), (2)
$1 - x_n$	NP	...	(3), (4)

where (1) through (4) have to be replaced with

1. $R_l(P, S) \equiv \alpha p - q + (1 - p)v(n - 1, k - 1, l - 1; \Gamma_P(q_n))$,
2. $R_l(P, NS) \equiv v(n - 1, k - 1, l; \Gamma_P(q_n))$,
3. $R_l(NP, S) \equiv -1 + v(n - 1, k, l - 1; \Gamma_N(q_n))$,
4. $R_l(NP, NS) \equiv v(n - 1, k, l; \Gamma_N(q_n))$.

$\Gamma_P(q_n)$ and $\Gamma_N(q_n)$ are the posterior beliefs revised from q_n at the beginning of the next stage $n - 1$, depending on Customs' present strategy P and NP , respectively,

and $v(n, k, l; q_n)$ is the expected payoff of the smuggler from the current state (n, k, l) to the end of the game. Please note that x_n or $y_n(l)$ ought to be zero in the case of $k = 0$ or $l = 0$, respectively, because Customs or the smuggler has only one choice of NP or NS in such cases.

Now let us derive the expected payoffs of two players and see a difference between them. Customs with belief q_n would expect the following payoff $P(x_n, y_n)$ and want to maximize the payoff by changing his strategy x_n .

$$\begin{aligned} P(x_n, y_n) = & \sum_{l=\underline{l}_n, l \neq 0}^{\bar{l}_n} q_n(l) [x_n \{y_n(l) R_l(P, S) \\ & + (1 - y_n(l)) R_l(P, NS)\} \\ & + (1 - x_n) \{y_n(l) R_l(NP, S) \\ & + (1 - y_n(l)) R_l(NP, NS)\}]. \end{aligned} \quad (5)$$

On the other hand, the type- l ($l \in [\underline{l}_n, \bar{l}_n]$, $l \neq 0$) smuggler correctly knows the present state (n, k, l) and would desire to minimize the following expected payoff by changing his strategy $y_n(l)$.

$$\begin{aligned} V_l(x_n, y_n(l)) = & x_n \{y_n(l) R_l(P, S) \\ & + (1 - y_n(l)) R_l(P, NS)\} \\ & + (1 - x_n) \{y_n(l) R_l(NP, S) \\ & + (1 - y_n(l)) R_l(NP, NS)\}. \end{aligned} \quad (6)$$

We notice that the asymmetry of information acquisition between two players bears some difference between expressions (5) and (6), but we also notice that $V_l(x_n, y_n(l))$ is contained in $P(x_n, y_n)$ as a part. Therefore, we can regard the stage game in State (n, k, l) as a TPZS game with the expected payoff $P(x_n, y_n)$ between Customs and all types of smugglers.

We derive a Bayesian-Nash equilibrium at each stage n by finding the optimal strategy x_n^* of maximizing $P(x_n, y_n)$, the optimal strategy $y_n^*(l)$ of minimizing $V_l(x_n, y_n(l))$ for all l s, and the rational belief $\{q_n^*(l)\}$ of being revised from $\{q_{n+1}(l)\}$ by $y_n^*(l)$ and having a consistency with its initial condition 4. A combination of optimal strategies $x_n^*, n \in N$, $\{y_n^*(l), l \in [\underline{l}_n, \bar{l}_n], n \in N\}$ and rational belief

$\{q_n^*(l), l \in [l_n, \bar{l}_n], n \in \mathbf{N}\}$ through all stages $\mathbf{N}$ is the *Bayesian-Nash equilibrium*. The derivation of the Bayesian-Nash equilibrium is usually tough in reality, and its numerical derivation requires a smart algorithm (refer to Hohzaki [36,37] for more details).

In Section 2, we described the multistage inspection game, in which both players can acquire their opponents' information. In this section, we have looked through the models with no-information and asymmetric information. By comparing with the values of the games, we can evaluate the value of information.

INTERDICTION GAMES AND AMBUSH GAMES

As mentioned in "Introduction", the study on the interdiction game was accelerated since the United States was confronted with the smuggling problem of drugs from Latin America in the 1980s. In a network theoretical context, Ford and Fulkerson [38] found the minimum cost interdiction against all paths from a start node s to a destination node t by the well-known max-flow min-cut theorem, although their problem is not a game. If we associate an arc (i, j) with a cost r_{ij} to break it, the minimum total cost to disrupt all paths from s to t is given by a minimum $s-t$ cut after corresponding r_{ij} to the capacity of arc (i, j) . For the interdiction game, Wollmer [39] discussed a game between an interdictor and an evader in a military context.

Let us see the basic model of Washburn and Wood [19]. An evader attempts to traverse a path from node s to node e through a network, which consists of a set of nodes $\mathbf{N}$ and a set of arcs $\mathbf{A}$, without being detected by an interdictor. The interdictor selects an arc k and sets up an inspection site. If the evader traverses the arc k , he is detected with probability p_k . Otherwise, he goes undetected. The interdiction game is the TPZS game with the payoff of the interdiction probability, in which an evader's strategy is the selection of a path and an interdictor's strategy is the selection of an arc for inspection. Let

us denote an evader's mixed strategy by $\hat{\mathbf{y}} = \{\hat{y}_l, l \in L\}$, where L is a set of all paths from s to e and $\hat{y}_l$ is the probability of taking path l , and an interdictor's mixed strategy by $\mathbf{x} = \{x_k, k \in \mathbf{A}\}$, where x_k is the probability to inspect the arc k . For an interdictor's mixed strategy $\mathbf{x}$ and an evader's mixed strategy $\hat{\mathbf{y}}$, the expected payoff is given by $V(\mathbf{x}, \hat{\mathbf{y}}) = \sum_{k \in \mathbf{A}} \sum_{l \in L} x_k p_k d_{kl} \hat{y}_l$, where d_{kl} is 1 if path l includes arc k and 0 otherwise, referred to as the *arc-path incident matrix*. The minimax optimization of this basic problem is easily formulated as follows: $\min_v$ s.t. $\sum_{l \in L} p_k d_{kl} \hat{y}_l \leq v (k \in \mathbf{A})$, $\sum_{l \in L} \hat{y}_l = 1$, $\hat{y}_l \geq 0 (l \in L)$. The authors showed that an optimal strategy x_k^* of the interdictor is inversely proportional to p_k if k belongs to a minimum cut on the network with capacity p_k and is zero otherwise.

The authors also extended the basic problem to some advanced versions with multiple sources and sinks, multiple evaders, or multiple interdictors from the point of view of mathematical programming. There are also some researches to deal with interdiction games for analysis on communication networks. Kodialam and Lakshman [40] developed a sampling strategy of network packets to effectively detect network intrusions within a given sampling budget. The interdiction game is sometimes called the *infiltration game* from the opposite coloration [41,42].

The ambush game has almost the same situation as the interdiction game. If we regard the research on the passage control of a submarine by anti-submarine warfare (ASW) airplanes in straits as an ambush game, we can say that Morse and Kimball discussed the first example of the ambush game in *Methods of Operations Research* [43]. While the interdiction game is focused on networks in many models, the ambush game is usually considered in general regions such as a rectangular lattice or a continuous rectangle.

Ruckle [27,44] introduced an original ambush game, in which Blue wishes to move across a rectangular lattice to avoid being ambushed by Red, who wants to place obstacles in the Blue's path. Let us consider the lattice having m rows and n

columns. Blue sets up a function f from domain $\{1, 2, \dots, n\}$ into range $\{1, 2, \dots, m\}$ as his path, and Red chooses a set of some points A from mn points in the lattice area under some constraints. For f and A , the payoff of the game $R(f, A)$ is defined to be 1 if $(i, f(i)) \notin A$ and 0 if $(i, f(i)) \in A$. Ruckle found optimal solutions in three cases. Let us see his results in two cases of them. In Case 1, Red cannot choose more than $k (< m)$ points from the lattice. An optimal mixed strategy of Blue is to pick up every integer j from $\{1, 2, \dots, m\}$ as $f(i) = j$ with probability $1/m$. An optimal mixed strategy of Red is to choose one from all combinations of choosing k from m points in a column with equal probability $1/\binom{m}{k}$. These strategies bring the value of the game $1 - k/m$. In Case 2, Red has the constraints that the number of points to pick up in a column is $c (< m)$ at most as well as the total number of points up to k . In this case, an optimal strategy of Red consists of first choosing $p + 1$ columns at random and then randomly picking up c points in p columns out of the $p + 1$ columns and q points in the rest one column, where p and q are the integers satisfying $k = pc + q (0 \leq q < c)$. In such a manner, each combination is taken with equal probability $1/\binom{m}{c}^p \binom{m}{q}$. An optimal strategy of Blue is to choose a row j and keep $f(i) = j$ for every column i with probability $1/m$ or to pick a point in every column at completely random, that is, to take one out of all combinations of $f(i) = j$ for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ with probability $1/m^n$. The value of the game becomes $(1 - c/m)^p (1 - q/m)$. The similar game is modeled on a two-dimensional continuous space.

Some researchers considered other situations of ambushing by Red. Red has two barriers of lengths a and b , which can be laid anywhere in an interval, and can detect the Blue's crossing at a point that intersects these barriers. The problem is referred to as *Ruckle problem*. Garnaev [45] considered a discrete version of the problem on a discrete interval $[1, n]$ of integers. The ambush game has been now gotten up-to-date such that a variety of barriers are available to Red, for example.

REFERENCES

1. Drescher M. A sampling inspection problem in arms control agreements: a game-theoretic analysis, Memorandum RM-2972-ARPA, Santa Monica (CA): The RAND Corporation; 1962.
2. Maschler M. A price leadership method for solving the inspection's non-constant-sum game. *Nav Res Log Q* 1966;13:11–33.
3. Kuhn HW. Recursive inspection games. In: Anscombe FJ., et al., editors. Applications of statistical methodology to arms control and disarmament, Final report to the U.S. Arms Control and Disarmament Agency under contract No. ACDA/ST-3 by Mathematica, Part III Princeton (NJ); 1963, p 169–181.
4. Bierlein D. Direkte inspektionssysteme. *Oper Res Verfah* 1968;6:57–68.
5. Jaech JL. Statistical methods in nuclear material control, Technical Information Center, United States Atomic Energy Commission TID-26298, Washington (DC); 1973.
6. Avenhaus R. Safeguards systems analysis. New York: Plenum; 1986.
7. Bowen WM., Bennett CA., editors. Statistical methodology for nuclear management. Report NUREG/CR-4604 PNL-5849. prepared for the U.S. Nuclear Regulatory Commission, Washington (DC): U.S. Nuclear Regulatory Commission; 1988.
8. Brams S., Davis MD. The verification problem in arms control: a game theoretic analysis. In: Ciotti-Revilla C., Merritt RL., Zinnes DA., editors. Interaction and communication in global politics. London: Sage; 1987. p 141–161.
9. Kilgour DM. Site selection for on-site inspection in arms control. *Arms Control* 1992;13:439–462.
10. Ruckle WH. The upper risk of an inspection agreement. *Oper Res* 1992;40:877–884.
11. Avenhaus R., Cnaty MJ. Compliance quantified. Cambridge: Cambridge University Press; 1996.
12. Avenhaus R. Inspection games. In: Aumann RJ., Hart S., editors. Handbook of game theory. Volume 3: New York: Elsevier Science; 2002. p 1947–1986.
13. Avenhaus R., Kilgour D. Efficient distributions of arm-control inspection effort. *Nav Res Log Q* 2004;51:1–27.
14. Avenhaus R., Cnaty MJ. Playing for time: a sequential inspection game. *Eur J Oper Res* 2005;167:475–492.

15. Hohzaki R. An inspection game with multiple inspectees. *Eur J Oper Res* 2007;178: 894–906.
16. Thomas M., Nisgav Y. An infiltration game with time dependent payoff. *Nav Res Log Q* 1976;23:297–302.
17. Mitchell T., Bell R. Drug interdiction operations by the coast guard. Alexandria (VA): Center for Naval Analysis; 1980.
18. Caulkins JP., Crawford G., Reuer P. Simulation of adaptive response: a model of drug interdiction. *Math Comput Mod* 1993;17: 27–52.
19. Washburn A., Wood K. Two-person zero-sum games for network interdiction. *Oper Res* 1995;43:243–251.
20. MacMasters AW., Mustin TM. Optimal interdiction of a supply network. *Nav Res Log Q* 1970;17:261–268.
21. Ghare PM., Montgomery DC., Turner WC. Optimal interdiction policy for a flow network. *Nav Res Log Q* 1971;18:37–45.
22. Fulkerson DR., Harding GC. Maximizing the minimum source-sink path subject to a budget constraint. *Math Prog* 1977;13: 116–118.
23. Golden B. A problem in network interdiction. *Nav Res Log Q* 1978;25:711–713.
24. Gal S. Search games. New York: Academic Press; 1980.
25. Wood K. Deterministic network interdiction. *Math Comput Mod* 1993;17:1–18.
26. Whiteman PS. Improving single strike effectiveness for network interdiction. *Milit Oper Res* 1999;4:15–30.
27. Ruckle WH. Geometric games and their applications. London: Pitman; 1983.
28. Baston V., Bostock F. A generalized inspection game. *Nav Res Log* 1991;38:171–182.
29. Garnaev A. A remark on the customs and smuggler game. *Nav Res Log* 1994;41: 287–293.
30. Sakaguchi M. A sequential game of multi-opportunity infiltration. *Math Japon* 1994;39: 157–166.
31. Ferguson T, Melolidakis C. On the inspection game. *Nav Res Log* 1998;45:327–334.
32. Hohzaki R., Kudoh D., Komiya T. An inspection game: taking account of fulfillment probabilities of players' aims. *Nav Res Log* 2006;53:761–771.
33. Hohzaki R. An inspection game with smuggler's decision on the amount of contraband. *J Oper Res Soc Jpn* 2011;54:25–45.
34. Hohzaki R., Maehara H. A single-shot game of multi-period inspection. *Eur J Oper Res* 2010;207:1410–1418.
35. Harsanyi JC. Games with incomplete information played by Bayesian players: Part I. *Manage Sci* 1967;8:159–182.
36. Hohzaki R. A smuggling game with the secrecy of smuggler's information. *J Oper Res Soc Jpn* 2012;55:23–47.
37. Hohzaki R., Masuda R. A smuggling game with asymmetrical information of players. *J Oper Res Soc* 2012;63:1434–1446.
38. Ford LR., Fulkerson DR. Flows in networks. Princeton (NJ): Princeton University; 1962.
39. Wollmer RD. Removing ARCS from a network. *J Oper Res Soc Am* 1964;12:934–940.
40. Kodialam M., Lakshman TV. Detecting network intrusions via sampling: a game theoretical approach. Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications (IEEE INFOCOM), Volume 3; 2003; p 1880–1889.
41. Auger JM. An infiltration game on (k) . *Nav Res Log* 1991;38:511–529.
42. Gal S. Search games. Wiley Encyclopedia of Operations Research and Management Science. New York: John Wiley and Sons; Volume 7: 2011 p 4743–4750.
43. Morse PM., Kimball GE. Methods of operations research. Cambridge: MIT Press; 1951.
44. Ruckle WH., Fennell R., Holmes PT., Fennemore C. Ambushing random walks I: finite models. *Oper Res* 1976;24:314–324.
45. Garnaev A. On a Ruckle problem in discrete games of ambush. *Nav Res Log* 1997;44: 353–364.

INSTANCE FORMATS FOR MATHEMATICAL OPTIMIZATION MODELS

HORAND GASSMANN
School of Business Administration,
Dalhousie University, Halifax,
Canada

JUN MA
Department of Industrial Engineering
and Management Sciences,
Northwestern University,
Evanston, Illinois

KIPP MARTIN
Booth School of Business, The
University of Chicago, Chicago,
Illinois

In order to use an optimization solver, it is necessary to communicate a model instance to the solver. Early users of optimization solvers often created a specific model instance from input data by writing a *matrix generator*. Mixed-integer linear programming (MILP) solvers such as IBM's MPSX/370 package and LINDO provided an application program interface that allowed a user to write a matrix generator in a procedural programming language such as PL/I or Fortran. The matrix generator was then linked to the solver library to produce an executable program that solved the specific instance. The term *matrix generator* arose because, initially, the most widely used optimization solvers were for MILP, and specifying the constraint matrix and objective function vector was sufficient to describe a model instance [along with a characterization of variable types and bounds and constraint right-hand sides (RHS)].

Matrix generators were tightly coupled with the solver because the model instance was communicated directly in-memory to the solver. However, modelers still needed to store the model instance in a persistent state for archival, communication, and debugging purposes. The classic MPS format was the first widely adopted instance format. It was

developed by William Orchard-Hays [1,2], it is based on an earlier format developed for the IBM 704 [3]. Even today, most LP (linear programming) and MILP solvers read and write files in MPS format.

The development of *algebraic modeling languages* represented a major breakthrough for the field of operations research because they greatly reduced the effort required to produce an optimization model instance. Modeling languages such as GAMS [4] and AMPL [5] that communicate with solvers by writing an instance to a file, have resulted in additional instance formats. Indeed, both GAMS and AMPL have their own instance representation formats. There now exists a plethora of proposed instance formats, depending on the optimization type. In Table 1 we list some key optimization types and corresponding formats. This table is illustrative, but by no means complete.

It is crucial to understand the difference between an algebraic model and a specific instance of that model. The description of a specific optimization instance is very different from the description of a generic algebraic model. An algebraic model description is *symbolic, general, concise, and understandable* [6]. An algebraic model contains declarations of high-level algebraic components such as sets, parameters, variables, objectives, and constraints. It is designed to be easily written and read by humans and often does not contain specific parameter values. By contrast, a model instance is designed to be read directly by a solver and all parameter values must be specified. An algebraic modeling language takes a model and compiles it into an instance using a specific data set. An LP model instance is illustrated in Fig. 1. (The details of this representation are discussed in the section titled "Optimization Instance Format.") An appropriate analogy from object-oriented programming is that the algebraic model corresponds to a class and the model instance corresponds to an *instantiated* object in that class.

```

1  NAME  ParInc. From Anderson, Sweeney, Williams, Camm, and Martin
2  ROWS
3    N      OBJ
4    L  R0000000
5    L  R0000001
6    L  R0000002
7    L  R0000003
8  COLUMNS
9    x0      OBJ      -10
10   x0      R0000000      0.7  R0000001      0.5
11   x0      R0000002      1    R0000003      0.1
12   x1      OBJ      -9
13   x1      R0000000      1    R0000001      0.8333
14   x1      R0000002      0.6667  R0000003      0.25
15  RHS
16   RHS1    R0000000      630
17   RHS1    R0000001      600
18   RHS1    R0000002      708
19   RHS1    R0000003      135
20  BOUNDS
21  ENDATA

```

Figure 1. A sample MPS file.

Table 1. Optimization Classes and Proposed Formats

Optimization Type	Instance Format
Mixed-integer linear program	MPS, lpsolve-lp, CPLEX-lp, AMPL nl, OSiL, OptML, GAMS dat
Nonlinear	xMPS, MOI, SIF, OSiL, AMPL nl, GAMS dat, LINDO MPI
Stochastic linear programming	SMPS, OSiL
Quadratic	MPS, OSiL
Semidefinite and second-order cone	sparse SDPA, SDPLR, OSiL
Linear network optimization	NETGEN, NETFLO, DIMACS, RELAX4
Traveling salesman problem	tsp

In this article we review existing instance formats for optimization problems, solver options, and solver results. We describe the research issues involved in designing instance formats. We do not address algebraic models.

In the next section we discuss key issues that must be considered when designing

an instance format. In the section titled “Optimization Instance Format” we describe formats for representing optimization problem instances. In addition to optimization instances, it is necessary to have formats for solver options (section titled “Solver Option Instance Format”) and results (section titled “Solver Result Instance Format”). There are obviously numerous types of optimization models. Extensions to some of the more recently studied optimization areas are the topic of the final section.

KEY DESIGN ISSUES

There are several key design issues when designing an instance representation. First is the decision of which instance to represent. Is the proposed format only for the problem instance, or does it include instances of solver options, solver results, or instance modifications? This is discussed more thoroughly in the following section. Then there is the key issue of how to format or represent problem instance information. This is discussed in the section on key issue 2. The next issue is how comprehensive and extensible the representation should be. This is discussed in the

section on design of the format. Finally, there is the issue of how the instance representations map to in-memory objects on the solver side. This is discussed in the section on key issue 4.

Key Issue 1: Separation of Functionality

In a modeling language-solver infrastructure, an instance passed to the solver must contain information on the problem to be optimized and may specify solver options. Likewise, a result is returned from the solver. There are actually four distinct instance formats that can be specified.

1. Optimization *problem* instance format—this is discussed briefly in the section on key issue 2 and in greater detail in the section titled “Optimization Instance Format”.
2. Solver *option* instance format—it is often necessary to pass options to a solver. This is a challenge given the wide variety of solvers and a complete lack of a standard for option formats. See the corresponding section for a proposed standard.
3. Solver *result* instance format—after a problem is solved (or terminated) results must be passed back from the solver to the client. See the corresponding section for more discussion on result instance formats.
4. Instance *modification* format—there are many solution procedures such as column generation and cutting plane algorithms that require modifying the original problem instance. Aside from the possibility of specifying multiple objectives and RHS in MPS, and switching between them, there is no standard instance modification format that the authors are aware of. This is a good topic for further research.

Key Issue 2: Problem Instance Format

A typical optimization problem instance might include millions of nonzero coefficients in linear terms and hundreds, if not thousands, of nonlinear terms. A

central problem in representing instances is managing these entries and ensuring that they are handled correctly and efficiently. There are three basic styles or formats that have been used to represent a model instance.

Natural Algebraic Format. Consider the simple constraint $.5x_0 + .8333x_1 \leq 600$. One way to represent this constraint instance is in its natural algebraic format. For example, in the LINDO solver [7], this constraint is expressed as

```
R0000001) +0.5 x0 + 0.8333 x1 < +600.
```

Other solvers have similar algebraic formats, for example CPLEX-lp file format [8], lpsolve lp-format, and Xpress lp-format [9,10]. Three important features of this format are that (i) it is easily read by a human and convenient for debugging purposes; (ii) it naturally extends to the nonlinear case; and (iii) it can exploit supersparsity [11]. With regard to the third point, we could rewrite

```
3.141592654*x1 +3.141592654*x2
+3.141592654*x3 + ...
```

as

```
3.141592654*( x1 + x2 + x3 + ... )
```

Modeling Language Specific Format. Modeling languages take an algebraic model and compile it with the data to produce a model instance. In the case of modeling languages such as AMPL [5] and GAMS [4], the model instance is temporarily written to a file which is then read by the solver. In order to make this process efficient, modeling languages use a parsimonious representation and do not store variable names. For example, the body of the constraint $.5x_0 + .8333x_1 \leq 600$ has the following representation in the AMPL nl format [12]:

```
J1 2
0 0.5
1 0.8333.
```

For more on how to interpret these lines and on representing the RHS, see the section titled “Optimization Instance Format”. Parsimony is important since problem instances are often exchanged over a network.

Markup Language Format. Yet another option for model instance representation is to use a markup language such as extensible markup language (XML). Two proposed XML-based instance formats are OSiL [13] and OptML [14]. An instance representation in OSiL is provided in the section titled “Optimization Instance Format.” Although apparently verbose, there are good reasons to use XML for an instance format. First, an instance represented in XML has both markup and data which allow for easy parsing and compression. Second, an instance in XML format can be validated against a schema which is important for error checking. The W3C XML schema standard permits developers of XML schemas to define their own data types called `complexType`s. The `complexType` is used to specify the elements and attributes that are allowed to appear in a valid XML instance file such as the one shown later in Fig. 3. The use of XML also facilitates the design of the corresponding in-memory instance (see the section on key issue 4). Third, XML is becoming the *lingua franca* of data storage in industry and using this format makes optimization modeling conform with modern information technology infrastructures (a more thorough discussion can be found in Ref. 15).

Key Issue 3: Design of Format: Level of Detail and Flexibility

An optimization problem instance is a well-defined entity. A certain amount of detail is required and it is not possible to be flexible in deciding whether or not certain problem components should be present. For example, we cannot have a linear program without constraints or variables. However, there is more flexibility in reporting the solution of an optimization problem. For example, in a LP solution, we may or may not have reduced costs or RHS sensitivity information. In general, different optimization solvers may present their results in different formats,

and some may include more detail than others. The level of solution detail is up to the solver developer and would be difficult to standardize. Similar logic applies to solver options. Thus, when representing a problem instance, it is necessary to express the instance in a very rigorous fashion. However, when designing option or result instances, flexibility may be more desirable. We elaborate on this issue in the sections titled “Solver Option Instance Format” and “Solver Result Instance Format.”

Key Issue 4: Design of an In-Memory Instance Object

In various situations, for example when using a modeling language such as GAMS or AMPL, the problem instance may be serialized (e.g., written to a file) and then deserialized (e.g., read back into memory) by the solver. The way in which a problem instance is represented in memory may look very different from the format that persists in a string or file. There is no “standard way” to represent, for example, an instance in the MPS format as an in-memory object. The optimization services (OS) framework is the first to specify a set of mapping or binding rules on how to transform the persistent instance format into in-memory objects. The OS format (OSiL) for representing an instance is based on a W3C XML schema. The use of a schema to specify the XML instance format is key to providing a binding between the format and the `OSInstance` class which is the in-memory representation of the optimization problem instance. For example, each XML schema `complexType` corresponds to a class in `OSInstance`. For more details on this binding see Ref. 13.

OPTIMIZATION INSTANCE FORMAT

Instance formats are used to store problem instances in a file or string before sending them to the solver. These can be created by hand, or be output from a matrix generator or an algebraic modeling language. Formats can be either text or binary.

An instance format must be able to represent all components of a problem instance.

Special purpose formats for narrow problem classes, such as traveling salesman problems, might have more restricted requirements and be more compact than formats for general problems but some problem components such as decision variables, objectives, and constraints are universal to all problem classes.

Information about the number of variables and constraints must be contained in the instance format, but there are at least two ways to accomplish this goal. For example, the AMPL nl format [12] has a header section that contains explicit information about the number of continuous, integer, and binary variables and breaks this down further into variables that appear in nonlinear expressions in the objective, constraints, both, or neither. The MPS format [16], on the other hand, contains the information implicitly and requires the reader to keep count of the variables as the instance is being processed. This is similar for constraints. Many instance formats (one notable exception being the MPS format) assume that the problem has a unique objective, so that there is no need to record the number of objectives. Obviously, formats designed to handle multiobjective problems must be able to express the number of objectives found in the problem instance.

Along with the number of decision variables, it is necessary to store information about them, such as upper and lower bounds. Finally, for report generation purposes and communication of the optimization result to the user, it is desirable that each variable have a unique identifier, which may or may not be communicated to the solver. For that reason, some instance formats refer to variables as well as constraints by number, while others, including the MPS format, refer to them by name. The latter allows more flexibility in the ordering of certain problem components but is more costly to parse.

The biggest challenge for a problem instance format is handling nonzero elements in the constraints. In problems of realistic size, it is essential to exploit sparsity in the linear, quadratic, and nonlinear components of the problem. Sparse representations of linear constraint matrices

can be achieved in several different ways. MPS records the matrix coefficients column by column as triples of (*column*, *row*, *value*), where *column* and *row* are names instead of numbers. This storage scheme results in the variable name being repeated for each row in which the variable has a nonzero value. AMPL nl format stores the constraint matrix by rows with markers separating one row from another, and with column number and coefficient value for every nonzero value in the row. OSiL uses the sparse matrix format [17], which is a combination of three arrays giving the column starts, row indices, and coefficient values when storing the constraint matrix by columns. OSiL can also store the constraint matrix by row using row starts, column indices, and coefficient values. Row and column representations work equally well for representing a constraint matrix. However, as we see at the end of this section, row-wise storage is more general and natural for nonlinear programs.

As an illustration of the above ideas, we offer the following small LP problem [18]:

$$\begin{aligned} \max \quad & 10x_0 + 9x_1 \\ \text{s.t.} \quad & 0.7x_0 + 1.0x_1 \leq 630, \\ & 0.5x_0 + 0.8333x_1 \leq 600, \\ & 1.0x_0 + 0.6667x_1 \leq 708, \\ & 0.1x_0 + 0.25x_1 \leq 135. \end{aligned} \tag{1}$$

Figure 1 is a representation of instance (1) in MPS format. Lines 3–7 set up the objective and constraint rows, the number being determined implicitly. Lines 9–11 give the objective and constraint coefficients associated with the first column, x_0 ; lines 12–14 give the coefficients for the second column, x_1 . Lines 16–19 specify the RHS values associated with the four constraints.

The AMPL nl format can be stored in either binary or text format; we give the text form for instance (1) in Fig. 2. The annotated section of the problem (lines 1–10) is used to set up the problem dimension; the next ten lines are used to show any nonlinear expressions (of which there are none, as indicated by the special code `n0` in lines 12,

```

1  g3 0 1 0 # problem parinc
2  2 4 1 0 0 # vars, constraints, objectives, ranges, eqns
3  0 0 # nonlinear constraints, objectives
4  0 0 # network constraints: nonlinear, linear
5  0 0 0 # nonlinear vars in constraints, objectives, both
6  0 0 0 1 # linear network variables; functions; arith, flags
7  0 0 0 0 0 # discrete variables: binary, integer, nonlinear (b,c,o)
8  8 2 # nonzeros in Jacobian, gradients
9  0 0 # max name lengths: constraints, variables
10 0 0 0 0 0 # common exprs: b,c,o,c1,o1
11 C0
12 n0
13 C1
14 n0
15 C2
16 n0
17 C3
18 n0
19 O0 1
20 n0
21 r
22 1 630
23 1 600
24 1 708
25 1 135
26 b
27 2 0
28 2 0
29 k1
30 4
31 J0 2
32 0 0.7
33 1 1
34 J1 2
35 0 0.5
36 1 0.8333
37 J2 2
38 0 1
39 1 0.6667
40 J3 2
41 0 0.1
42 1 0.25
43 G0 2
44 0 10
45 1 9

```

Figure 2. A sample nl file.

14, 16, 18, and 20) in the constraints and the objective. The RHS are set up in lines 21–25. (The number 1 preceding the RHS values marks them as upper bounds.) Lines 26–28 give lower bounds of 0 on the decision variables. The constraint matrix is given in row format starting in line 31. Line 31 specifies that the first row (numbered 0) contains two

entries, and the following two lines contain index-value pairs. This pattern repeats for the remaining three constraints. A similar format is used in lines 43–45 to specify the objective row.

Finally, we give the OSiL format for the same problem (Fig. 3). OSiL is an extremely general problem instance format

```

1  <?xml version="1.0" encoding="UTF-8"?>
2  <osil xmlns="os.optimizationservices.org">
3    <instanceHeader>
4      <name>ParInc. </name>
5    </instanceHeader>
6    <instanceData>
7      <variables numberOfVariables="2">
8        <var name="x0" lb="0" type="C" />
9        <var name="x1" lb="0" type="C" />
10     </variables>
11     <objectives numberOfObjectives="1">
12       <obj name = "Par, Inc. Objective Function"
13         maxOrMin="max" numberOfObjCoef="2">
14         <coef idx="0">10</coef>
15         <coef idx="1">9</coef>
16       </obj>
17     </objectives>
18     <constraints numberOfConstraints="4">
19       <con name="cutanddye" ub="630"/>
20       <con name="sewing" ub="600"/>
21       <con name="finishing" ub="708"/>
22       <con ub="135"/>
23     </constraints>
24     <linearConstraintCoefficients numberOfValues="8">
25       <start>
26         <el>0</el> <el>4</el> <el>8</el>
27       </start>
28       <rowIdx>
29         <el>0</el> <el>1</el>
30         <el>2</el> <el>3</el>
31         <el>0</el> <el>1</el>
32         <el>2</el> <el>3</el>
33       </rowIdx>
34       <value>
35         <el>0.7</el><el>.5</el>
36         <el>1.</el> <el>.1</el>
37         <el>1.0</el><el>0.8333</el>
38         <el>0.6667</el><el>0.25</el>
39       </value>
40     </linearConstraintCoefficients>
41   </instanceData>
42 </osil>

```

Figure 3. A sample OSiL file.

for representing linear, integer, quadratic, and nonlinear problem instances as well as a large number of special features, such as multiobjectives, stochastic programs, cone programs, special ordered sets (SOS), and others. The OSiL format is largely self-documenting, so we only highlight a few main points. Header information is given in lines 3 to 5. The actual optimization data start in line 6. Separate sections give

information about variables (lines 7–10: bounds, variable type, and an optional name), objective function (lines 12–16: direction of optimization and objective coefficients) and constraints (lines 19–22: RHS values).

In OSiL XML markup the constraint matrix is represented in sparse matrix format [17], either by row or by column. In this example we are storing by column (often called the column major). In column-major

storage, the k^{th} `<start>` element in line 26 points to the start of the column k row indexes (contained in lines 29–32) and the column k coefficient values (contained in lines 35–38). The last `<start>` element (in this case 8) value is the number of nonzero elements.

Due to the availability of fast LP codes, linear models were initially of primary interest. However, as progress is made with nonlinear solution algorithms, there is increased interest in formats for nonlinear instances. For example, there are a number of solvers that now support objectives and constraints with quadratic terms and second-order cone constraints. There are even extensions of the MPS format to allow for quadratic terms (QMATRIX) and second-order cone constraints (CSECTION). See for example, Ref. 19, for an illustration of both the QMATRIX and CSECTION MPS formats. Rather than detail these specific extensions we address representation of general nonlinear optimization problems.

There are several ways in which general nonlinear expressions can be captured in an instance format. We illustrate with a modification of the Rosenbrock function [20]. The model is as follows:

$$\begin{aligned}
 \min \quad & (1 - x_0)^2 + 100 * (x_1 - x_0^2)^2 + 9 * x_1 \\
 \text{s.t.} \quad & x_0 + 10.5 * x_0^2 + 11.7 * x_1^2 \\
 & + 3 * x_0 * x_1 \leq 25, \\
 & \ln(2 * x_0 * x_1) + 7.5 * x_0 \\
 & + 5.25 * x_1 \geq 10.
 \end{aligned} \tag{2}$$

In instance (2), the terms $100 * (x_1 - x_0^2)^2$ and $\ln(2 * x_0 * x_1)$ are not quadratic and must use a more general representation. We illustrate approaches to modeling general nonlinear terms using the $\ln(2 * x_0 * x_1)$ term. The standard algebraic format $\ln(2 * x_0 * x_1)$ is *infix notation* [21] and is ambiguous with regard to the order of operations. The ambiguity is resolved by using parentheses (or by convention with regard to operator precedence and left or right associativity), where, for example, $\ln((2 * x_0) * x_1)$ specifies an order of operations. Due to the

infix ambiguity, most instance formats store each nonlinear term as a *prefix* or *postfix* sequence of operands and operators. The list of operators and operands is also called an *instruction list*. The prefix and postfix sequences for $\ln((2 * x_0) * x_1)$ are

```
ln * * 2 x0 x1
```

and

```
x1 2 x0 * * ln,
```

respectively. A prefix or postfix instruction list is trivial to evaluate using a stack data structure and is also convenient for algorithmic or automatic differentiation. The postfix or prefix instruction list is a key part of several instance formats. For example, the AMPL nl format uses prefix combined with numerical codes for the operators `ln` and `*` and writes

```
o43
o2
o2
n2
v0
v1,
```

where `o43` is the code for the logarithm operator, `o2` is the code for the multiplication operator, and `n` and `v` denote references to numbers and variables, respectively.

The OSiL instance format also corresponds to a prefix instruction list. The term $\ln((2 * x_0) * x_1)$ in OSiL is

```
<nl idx="1">
  <ln>
    <times>
      <times>
        <number type="real" value="2.0"/>
        <variable coef="1.0" idx="0"/>
      </times>
      <variable coef="1.0" idx="1"/>
    </times>
  </ln>
</nl>.
```

The sequence of open tags is

```
<ln><times><times><number><variable><variable>
```

and corresponds directly to the prefix instruction list.

A nonlinear extension of MPS based on postfix is described by Halldórsson *et al.* [22]. It puts the postfix instruction list into the MPS record structure. For example, $\ln((2 * x_0) * x_1)$ is written as

```
NONLINEAR
CON1 V1 mult 2 X0
CON1 V2 mult V1 X1
CON1 RES ln V2.
```

Lines 2 and 3 compute the product $2x_0x_1$ and store it into the intermediate object V2; the last line computes the natural logarithm of V2 and stores into the reserved object RES, which is used to designate the final result of the computation, which gets added to the linear and quadratic terms for constraint CON1.

A completely different format, SIF (standard input format), was developed by Conn *et al.* [23,24]. A SIF file consists of two pieces, a declarative portion in which a number of declarations of auxiliary quantities are interspersed with declarations of variables, objectives, and constraints, and a procedural portion where the nonlinear functions are defined in terms of the original variables and coefficients as well as the auxiliary quantities previously defined. SIF is a very complex and far-reaching proposal that did not achieve universal acceptance.

Formats for specialized problems can occasionally streamline the information needs and reduce, often greatly, the size of the resulting instance file. For instance, a stochastic programming problem may need to record values for problem coefficients that take different values in a number of different scenarios, along with a number of deterministic coefficients that do not change from one scenario to the next. If the random variables are time-staged, this can lead to a geometric explosion of the event tree. The SMPS format for stochastic programs reduces redundancy by storing only the stochastic components explicitly. When the random variables are independent from one stage to the next, the SMPS format can store the event tree in linear space. See Ref. 25 for more details. Similar features are available in OSiL [26]. Compact specialized instance formats exist for other problem classes,

such as semidefinite cone programming, network problems, and traveling salesman problems.

SOLVER OPTION INSTANCE FORMAT

Solver options are an integral part of the solution process and are usually tied to the problem instance. Traditionally, there have not been any standards for the communication of options since each solver tended to have its own application programming interface (API) with which the user communicates either through files, command line options or interactive keyboard inputs.

Standardizing the input format for solver options is hard since the options tend to be very specific to the solver and algorithm. In designing a common format it is useful to break down the process into syntax (how to represent options) and semantics (how to interpret their meanings). While the former is relatively easy to standardize, the latter is nearly impossible, even though a number of common options may be identified that are implemented in the majority of solvers. Standard option formats therefore, need to be both rigid in format and flexible in content, and must be extensible to leave room for future solver development.

The development of option formats is driven by other considerations as well. Since solver options form only part of the input sent to a solver, it is desirable to have some way to ensure that the options are linked to the correct problem instance. This is necessary particularly where the options provide values for the elements of an instance component, such as initial variable values. Some solvers such as LINDO allow solver options to have different scopes, applying either to one particular problem only, or to the entire solver environment. This is useful because it allows the user to set an output level once for a series of problems, instead of having to repeat the information for each separate problem.

Solvers are not the only component of a mathematical programming system; options could be sent to an optimization analyzer or simulation tool. If a large environment has

such diverse tools, it seems advantageous to design option formats that can be shared among all components. In order to establish the correct match of option file and analysis tool, the intended target must be recorded in the option file. Additionally, especially in a commercial environment, basic security features such as license information, username, and password may be needed in the options file. Finally, especially in a distributed environment, additional facilities are often needed to monitor the progress of the solution process on the remote computer, to verify capabilities of the computer system on which the solver is to be run, and perhaps to perform ancillary file operations before and after the solution.

In order to describe the design of an options format, we distinguish several classes of solver options. First, options may apply to the problem instance. For example, the MPS input format does not allow specifying the direction of optimization (min or max). Solver options (so called *agenda cards*) [16] can be used to communicate this information.

A second category of options controls the flow of the solution algorithm. In this category, there are options to select a particular algorithm class (e.g., primal simplex, dual simplex or interior point method in a linear program) or algorithm variant (for instance a particular pricing scheme). Options such as whether to scale a coefficient matrix or not, which crashing and preprocessing procedures to use, as well as limits on the number of iterations or the solution time, would also fall into this category. For nonlinear problems, one can also specify an initial point.

Other solver options include various tolerances, the amount and type of output generated (including the level of diagnostic messages when parsing the input), the location and name of input and output files, whether the solution needs to be saved for subsequent runs, perhaps with modified data, and so on. It does not stop there. Especially in distributed computing, it may be necessary to control the environment in which the solution is obtained. For example, it may be known that the solution of an

instance places space requirements on a server, which may limit the choice of servers that could potentially be used.

The optimization services option language (OSoL) schema is to our knowledge the first attempt at standardizing the specification of solver and process options [27,28]. We end this section with a small file in the OSoL format (Fig. 4). This file should be self-explanatory. It first specifies that the job is to be run only on computers with at least 1 Gb of memory, provides initial values for two decision variables, and gives four solver options. Two of these options are intended for the Ipopt solver [29]; the other two are for the LINDO solver [30]. This demonstrates that an option file can be shared between different solvers. The two Ipopt options control the amount of output to be generated and where the output is to be directed; the LINDO options illustrate that the same option can be used with different scope, applying to a single model in the first case and to the entire session in the second case.

SOLVER RESULT INSTANCE FORMAT

It is necessary for an optimization solver to communicate the solution result back to the user. Most solvers have their own result format or API for communicating the solution result. Although not nearly as widespread as the MPS instance format, there is also an MPS format for reporting solver results [16]. The MPS result format is a table-based one and separated into three sections: header, rows, and columns. The header section lists such information as problem instance name, objective value, status, and number of iterations. The rows section provides results on the constraints such as constraint values, slacks, and duals. The columns section provides results on the variables such as solution values and reduced costs. The two major problems with the MPS result instance format are that it is optimization-centric and is not easily extensible.

An alternative result instance format is OSrL (optimization services result language) [27,28]. OSrL supports multiple solutions with multiobjectives. Each solution contains

```

<?xmlversion="1.0"encoding="UTF-8"?>
<osol>
  <system>
    <minMemorySizeunit="gigabyte">1.0</minMemorySize>
  </system>
  <optimization>
    <variablesnumberOfOtherVariableOptions="0">
      <initialVariableValuesnumberOfVar="2">
        <varidx="0"value="1"/>
        <varvalue="4.742999643577776e-2"idx="1"/>
      </initialVariableValues>
    </variables>
    <solverOptionsnumberOfSolverOptions="4">
      <solverOptionname="print_level"solver="ipopt"
        type="integer"value="5"/>
      <solverOptionname="output_file"solver="ipopt"
        type="string"value="ipopt.out"/>
      <solverOptionname="LS_IPARAM_LP_PRINTLEVEL"solver="lindo"
        category="model"type="integer"value="0"/>
      <solverOptionname="LS_IPARAM_LP_PRINTLEVEL"solver="lindo"
        category="environment"type="integer"value="1"/>
    </solverOptions>
  </optimization>
</osol>

```

Figure 4. A sample OSoL file.

a separate set of variable, objective, constraint, and other results. For results that are uncommon, or solver specific, or subject to different interpretations, OSrL provides the mechanism of `otherSolutionResult` and `otherSolverOutput`. We show below a small segment of OSrL from an actual application that illustrates “nonstandard” solver output that is important to communicate (Fig. 5).

In addition to the optimization category, which communicates output directly related to the optimization solution, the OSrL format contains four further categories: general, system, service, and job. The general result section contains such information as general status and message, service name and uniform resource identifier (URI); instance name, job ID,

solver invoked, and time stamp. The system section contains information about available disk space, memory, and CPU. The service section contains information about the host optimization service, such as its current state (busy or idle) and service utilization. The job section contains information about the particular optimization job process such as its submit/scheduled/start/end time, as well as CPU, disk, and memory usage. Each of these sections provides the “other” mechanism for custom results.

EXTENSIONS

If a problem instance format is well designed, it should be flexible enough to extend the

```

<otherSolutionResultname="ReqDocPairings"numberOfItems='55'>
  <item>778,Liberty,Surgical,Pod1,1,M,AM,1</item>
  <item>778,Liberty,Surgical,Pod2,1,T,AM,1</item>
  <item>551,Liberty,Surgical,Pod3,1,M,AM,1</item>
  ...
</otherSolutionResult>

```

Figure 5. A section of an OSrL file.

standard “core” optimization types such as continuous and mixed integer linear, quadratic, and general nonlinear programs. In this section we list potential extensions to the core instance format. It is critical that the core instance format is not affected by the extensions and that whenever a new optimization type is added, the core format remains unchanged, keeping the standards backward compatible.

1. **Logic and Constraint Programs.** If the instance format has a built-in nonlinear expression tree, it should be quite natural for it to be extended to this type of programs, as logic and relational operators are expressed in the same way as all regular nonlinear operators. For example `if` is simply an operator that takes three operands. Constraint programs use other special operators such as `and`, `or`, `xor`, `not`, `if`, `implies`, `eq`, `neq`, `geq`, `gt`, `leq`, `lt`, and `alldiff`.
2. **Complementarity Problems.** Similar to extending the logic and constraint programs, we can add a `complements` operator to express such problems. The `complements` operator takes two operands. Each operand must be either an inequality or an equality operand (in which case the condition is redundant). The `complements` operator evaluates to true if both operands are true and at least one inequality is tight.
3. **Special ordered sets (SOS).** There are several types of SOS [31–34], but each type is essentially expressed as a set of indices of variables already defined in the core programs, plus optional information such as a convexity row.
4. **Semicontinuous Constraints (SC).** These are typically constraints on variables that require the value of a variable to either be equal to 0 or lie between a lower and upper bound where the lower bound is greater than zero.
5. **Cone Programs.** There are many widely used cone types including nonnegative cone, quadratic cone, rotated quadratic cone, and semidefinite cone [35].
6. **Disjunctive Programs.** Among all the possible extensions, the disjunctive program extension is probably the one that benefits the most from building on the core, as almost all the data have already been represented in the core and each disjunction usually just alters a tiny piece of information from the core; for example, change of a matrix coefficient or constraint bound. Each disjunction is essentially trying to construct a separate problem instance with the least amount of extra data.
7. **Robust Optimization and Stochastic Optimization.** These are two ways for dealing with uncertainty in the problem parameters. Robust optimization can be thought of as a special case of stochastic programming, and it is a design choice whether to define it as a separate extension or just subsume it under the more general heading of stochastic programming.
8. **Instance Modification.** Popular algorithms such as branch-cut-and-price often require addition or deletion of constraints and variables. Currently there is no instance format for specifying model changes.
9. **Real-Time Data.** In a tightly coupled modeling environment, a modeler creates a problem instance using a modeling language. The instance file is self-contained and has all of the parameters necessary for optimization by a solver. However, as data are updated in real time, it is desirable from an efficiency standpoint to update only the data that have changed and not regenerate the entire instance.
10. **User-Defined Functions.** No matter how many nonlinear operators and functions an instance format supports, it cannot exhaust all of them and should support custom-defined functions, to be used in the nonlinear expression tree along with other standard functions.
11. **Optimization via Simulation.** Not all functions (constraints or objectives) have explicit forms and if the function

is a black box simulation that cannot be easily modeled by any user function, one needs the support of optimization via simulation.

12. Network and Graph Problems. Network and graph problems traditionally are built around the concepts of *vertices* and *arcs*, instead of *constraints* and *variables*. Special formats may be needed to deal with them effectively.
13. Instance Communication over a Network. Given the trend toward cloud computing and computing on demand, there will almost certainly be a need to communicate instance formats over a network. The optimization problem, solver options, and solver results, along with header information, must all be packed in the body of a larger “instance” that is communicated over a network. For a discussion of higher-level network instance formats for optimization see Refs 27 and 28.
14. Optimization over Vector Spaces. Discrete-time optimal control problems are often merely an approximation of an underlying continuous-time problem. The discretization step is arbitrary and should occasionally be allowed to be refined by the solution algorithm. It would therefore be useful to have a way to express the underlying continuous-time problem directly.

REFERENCES

1. Orchard-Hays W. History of mathematical programming systems. In: Greenberg H, editor. Design and implementation of optimization software. The Netherlands: Sijthoff & Nordhoff; 1978.
2. Orchard-Hays W. History of mathematical programming systems. IEEE Ann Hist Comput 1984;6:296–312.
3. Orchard-Hays W, Cutler L, Judd H. Manual for the RAND-IBM code for linear programming on the 704. RAND Corporation. 1956. Document P-842.
4. Gams Development Corporation. Available at <http://www.gams.com>. Accessed 2010 July.
5. AMPL Optimization LLC. <http://www.ampl.com>. Accessed 2010 July.
6. Fourer R. Modeling languages versus matrix generators for linear programming. ACM Trans Math Softw 1983;9(2):143–183.
7. Geeknet, Inc. LINDO lp files. Available at <http://lpsolve.sourceforge.net/5.5/LINDO-format.htm>. Accessed 2010 Aug.
8. Mittelmann HD. CPLEX LP format. http://plato.asu.edu/cplex_lp.pdf. Accessed 2010 Aug.
9. Geeknet, Inc. LP file format. Available at <http://lpsolve.sourceforge.net/5.5/lp-format.htm>. Accessed 2010 Aug.
10. Geeknet, Inc. XPRESS lp files. Available at <http://lpsolve.sourceforge.net/5.5/Xpress-format.htm>. Accessed 2010 Aug.
11. Kalan JE. Aspects of large-scale in-core linear programming. Proceedings of the 1971 26th Annual Conference. New York: ACM, Association for Computing Machinery; 1971. pp. 304–313.
12. Gay DM. Hooking your solver to AMPL. Technical Report 97-4-06. Murray Hill (NJ): Computing Sciences Research Center, Bell Laboratories; 1997.
13. Fourer R, Ma J, Martin K. OSiL: an instance language for optimization. Comput Optim Appl 2010;45(1):181–203.
14. Kristjánsson B. Optimization modeling in distributed applications, Presented at INFORMS Round Table; 2003 Oct; Atlanta (GA). Available at <http://www.maximal-usa.com/slides/Svna01Max/sld036.htm>. Accessed 2010 July.
15. Fourer R, Lopes L, Martin RK. LPFML: a W3C XML schema for linear and integer programming. INFORMS J Comput 2005; 17(2):139–158.
16. Murtagh BA. Advanced linear programming: computation and practice. New York: McGraw-Hill Publishers; 1981.
17. Duff IS, Erisman AM, Reid JK. Direct methods for sparse matrices. Oxford: Oxford University Press; 1986.
18. Anderson DR, Sweeney DJ, Williams TA, et al. An introduction to management science: quantitative approaches to decision making. 13th ed. Mason (OH): South-Western Cengage Publishing. In press.
19. The Mosek C API manual version 5.0 (revision 137). Available at http://www.mosek.com/fileadmin/products/5_0/tools/doc/html/capi/node022.html. Accessed 2010 July.

20. Rosenbrock HH. An automatic method for finding the greatest or least value of a function. *Comput J* 1960;3:175–184.
21. Jinks PJ. Infix, postfix and prefix. <http://www.cs.man.ac.uk/~pjj/cs212/fix.html>. Accessed 2010 July.
22. Halldórsson BV, Thorsteinsson ES, Kristjánsson B. A modelling interface to non-linear programming solvers: An instance—xMPS, the extended MPS format 2000. Available at <http://www.maximalsoftware.com/resources/xmps/xmps.pdf>. Accessed 2010 July.
23. Conn AR, Gould NIM, Toint PL. Volume 17, LANCELOT: a fortran package for large-scale nonlinear optimization (release A), Springer series in computational mathematics. Berlin: Springer; 1992.
24. Conn AR, Gould NIM, Toint PL. The SIF reference document. Available at <http://www.numerical.rl.ac.uk/lancelot/sif/sifhtml.html>. Accessed 2010 July.
25. Gassmann HI. The SMPS format for stochastic linear programs. Available at <http://myweb.dal.ca/gassmann/smeps2.htm>. Accessed 2010 July.
26. Fourer R, Gassmann HI, Ma J, *et al.* An XML-based schema for stochastic programs. *Ann Oper Res* 2009;166:313–337.
27. Fourer R, Gassmann HI, Ma J, *et al.* COIN-OR optimization services project. Available at <http://www.optimizationservices.org>. Accessed 2010 July.
28. Fourer R, Ma J, Martin K. Optimization services: a framework for distributed optimization. *Oper Res*. In press.
29. Waechter A. Introduction to ipopt. Available at <http://www.coin-or.org/Ipopt/documentation/>. Accessed 2010 July.
30. LINDO Systems, Inc. Available at <http://www.lindo.com>. Accessed 2010 July.
31. Beale EML, Forrest JJH. Global optimization using special ordered sets. *Math Program* 1976;10:52–69.
32. Tomlin JA. A suggested extension of special ordered sets to non-separable non-convex programming problems. In: Hansen P, editor. *Studies on graphs and discrete programming*. Amsterdam: North-Holland Publishing Co; 1981. pp. 359–370.
33. Escudero LF. S3 sets, an extension of the Beale-Tomlin special ordered sets. *Math Prog Ser A and B* 1988;42(1):113–123.
34. Geeknet, Inc. lp_solve 5.0 online documentation <http://lpsolve.sourceforge.net/5.0/SOS.htm>. Accessed 2010 July.
35. Pólik I, Gassman HI, Ma J, *et al.* Modeling cone optimization problems with coin OS. Paper presented at INFORMS 2009; 2009 Oct 12; San Diego (CA).

INTEGER PROGRAMMING DUALITY

MENAL GUZELSOY
School of Industrial and Systems
Engineering, Georgia
Institute of Technology,
Atlanta, Georgia

TED K. RALPHS
Department of Industrial and
Systems Engineering, Lehigh
University, Bethlehem,
Pennsylvania

INTRODUCTION

This article describes what is known about duality for integer programs. It is perhaps surprising that many of the results familiar from linear programming (LP) duality do extend to integer programming. However, this generalization requires adoption of a more general point of view on duality than is apparent from studying the linear programming case. Duality theory, as we shall define it here, is generally concerned with the development of methods for determining the effect on the optimal solution value of perturbations in the right-hand side of a mathematical program, which can be thought of as specifying the resources available to operate a system. The study of duality is hence the study of a parameterized family of mathematical programming instances and is closely related to parametric programming and sensitivity analysis. More formally, duality theory can be thought of as the study of the so-called *value function*, which is a function that takes a right-hand side vector as input and returns the optimal solution value of one of a family of integer programming instances parameterized by that right-hand side vector. In the linear programming case, the value function is piecewise linear and convex (assuming minimization), which is the reason for the tractability of the linear programming dual problem. In the integer programming case,

the value function has a more complex structure, as we shall see.

In this article, we consider a mixed integer linear program (MILP) (referred to as the *primal problem*) of the form

$$z_{\text{IP}} = \min_{x \in S} cx, \quad (1)$$

where $c \in \mathbb{R}^n$ is the objective function vector and the feasible region

$$S = \{x \in \mathbb{Z}_+^r \times \mathbb{R}_+^{n-r} \mid Ax = b\} \quad (2)$$

is defined by a constraint matrix $A \in \mathbb{Q}^{m \times n}$ and a right-hand side vector $b \in \mathbb{R}^m$. Note that for simplicity of notation, we omit the transpose symbol for the inner product of vectors and for matrix–vector products. By convention, $z_{\text{IP}} = \infty$ if $S = \emptyset$ and $z_{\text{IP}} = -\infty$ if for every $x \in S$, there exists $y \in S$ such that $cy < cx$. The goal is to determine z_{IP} and an *optimal solution* x^* such that $cx^* = z_{\text{IP}}$ (if z_{IP} is finite), where cx is the *solution value* of $x \in S$. The variables indexed by the set $I = \{1, \dots, r\} \subseteq N = \{1, \dots, n\}$ (if $r > 0$, otherwise $I = \emptyset$) are the *integer variables* and the remaining variables, indexed by the set $C = N \setminus I$, are the *continuous variables*. The LP obtained from a given MILP by removing the integrality requirements on the variables, that is, setting $r = 0$, is referred to as the associated *LP relaxation*. Similarly, the associated *pure integer linear program* (PILP) is obtained by requiring *all* variables to be integer, that is, setting $r = n$.

By parameterizing (1), we obtain the *value function* $z : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\pm\infty\}$ defined as

$$z(d) = \min_{x \in S(d)} cx, \quad (3)$$

where, for a given $d \in \mathbb{R}^m$, $S(d) = \{x \in \mathbb{Z}_+^r \times \mathbb{R}_+^{n-r} \mid Ax = d\}$. A *dual function* $F : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is a function that bounds the value function over the set $\mathbb{R}^m$, that is,

$$F(d) \leq z(d) \quad \forall d \in \mathbb{R}^m. \quad (4)$$

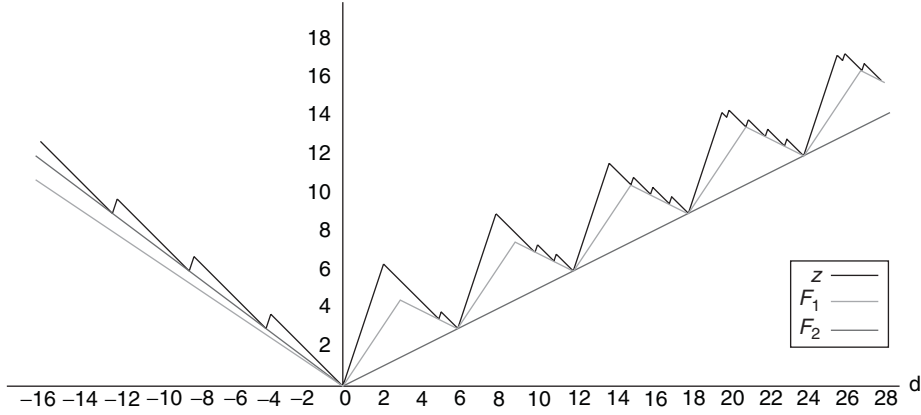


Figure 1. The value function and sample dual functions for (6).

Dual functions and methods for constructing them are the central focus of duality theory. The aim in constructing a dual function F is for it to closely approximate the value function z . In particular, we usually want a dual function whose value at b is as close as possible to $z(b)$. The most general form of the *dual problem* is then

$$z_D = \max \{F(b) : F(d) \leq z(d), d \in \mathbb{R}^m, \\ F : \mathbb{R}^m \rightarrow \mathbb{R}\}, \quad (5)$$

which is an optimization problem over the space of dual functions. We call F a *weak dual function* if it is feasible for (5), that is, it satisfies (4), and a *strong dual function* with respect to b if it is optimal for (5), that is, satisfies $F(b) = z(b)$. Note that we allow only real-valued functions, so that (5) is infeasible if $z_{IP} = -\infty$ and $z_D \leq z_{IP}$ otherwise. If z_{IP} is finite, the value function itself (or a real-valued extension of it) is a strong dual function, that is, $z_D = z_{IP}$.

Example 1. Consider the following MILP instance with a fixed right-hand side b :

$$\begin{aligned} \min \quad & 3x_1 + \frac{7}{2}x_2 + 3x_3 + 6x_4 + 7x_5 \\ \text{s.t.} \quad & 6x_1 + 5x_2 - 4x_3 + 2x_4 - 7x_5 = b \text{ and} \\ & x_1, x_2, x_3 \in \mathbb{Z}_+, x_4, x_5 \in \mathbb{R}_+. \end{aligned} \quad (6)$$

The value function z of this sample problem together with dual functions F_1 and F_2 are given in Fig. 1. Note that the value function is better approximated by F_1 for right-hand side values $b > 0$ and by F_2 otherwise. In addition, observe that the dual functions are strong for a subset of right-hand side values and weak otherwise.

PROPERTIES OF THE VALUE FUNCTION

The study of duality begins with the study of the value function itself. In particular, we want to know what the structure of the value function is and to what class of functions it may belong. Our first observation is that the value function is *subadditive*. A function F is *subadditive* over a domain Θ if $F(\lambda_1) + F(\lambda_2) \geq F(\lambda_1 + \lambda_2)$ for all $\lambda_1, \lambda_2, \lambda_1 + \lambda_2 \in \Theta$. The value function itself is subadditive over the domain $\Omega = \{d \in \mathbb{R}^m \mid S(d) \neq \emptyset\}$. To study this, let $d_1, d_2 \in \Omega$ and suppose $z(d_i) = cx_i$ for some $x_i \in S(d_i)$, $i = 1, 2$. Then, $x_1 + x_2 \in S(d_1 + d_2)$ and hence $z(d_1) + z(d_2) = c(x_1 + x_2) \geq z(d_1 + d_2)$. Subadditivity reflects the intuitively natural property of increasing returns to scale that we expect the value function to have.

For PILPs, Blair and Jeroslow [1] showed that a procedure similar to Gomory's cutting plane procedure can be used to construct

the value function in a finite number of steps, although the representation from this procedure may have exponential size. They were thus able to characterize the class of functions to which value functions belong, namely, *Gomory functions*, a subset of a more general class called *Chvátal functions*.

Definition 1. Let $\mathcal{L}^m = \{f : \mathbb{R}^m \rightarrow \mathbb{R} \mid f \text{ is linear}\}$. Chvátal functions are the smallest set of functions $\mathcal{C}$ such that

1. If $h \in \mathcal{L}^m$, then $h \in \mathcal{C}$.
2. If $h_1, h_2 \in \mathcal{C}$ and $\alpha, \beta \in \mathbb{Q}_+$, then $\alpha h_1 + \beta h_2 \in \mathcal{C}$.
3. If $h \in \mathcal{C}$, then $\lceil h \rceil \in \mathcal{C}$.

Gomory functions are the smallest set of functions $\mathcal{G} \subseteq \mathcal{C}$ with the additional property that

4. If $h_1, h_2 \in \mathcal{G}$, then $\max\{h_1, h_2\} \in \mathcal{G}$.

Note that both Chvátal and Gomory functions are subadditive and every Gomory function can be written as the maximum of finitely many Chvátal functions. Blair and Jeroslow [1] show for a PILP in the form (1) with $z(0) = 0$ that (i) the value function over the domain Ω is a Gomory function and (ii) there always exists an optimal solution to the dual problem (5) that is a Chvátal function.

For MILPs, Blair and Jeroslow [2] argue that the value function can still be represented by a Gomory function if $c_j = 0 \forall j \in C$ or can be written as a minimum of finitely many Gomory functions. A deeper result is contained in the subsequent work of Blair [3], who showed that the value function of an MILP can be written as a *Jeroslow formula*, consisting of a Gomory function and a correction term. For a given $d \in \Omega$, let the set $\mathcal{E}$ consist of the index sets of dual feasible bases of the linear program

$$\min \{c_C x_C : A_C x_C = d, x \geq 0\}. \quad (7)$$

By the rationality of A , we can choose $M \in \mathbb{Z}_+$ such that $MA_E^{-1}\alpha^j \in \mathbb{Z}^m$ for all $E \in \mathcal{E}$, $j \in I$, where α^j is the j th column of A . For $E \in \mathcal{E}$,

let v_E be the corresponding basic feasible solution to the dual of

$$\min \left\{ \frac{1}{M} c_C x_C : \frac{1}{M} A_C x_C = d, x \geq 0 \right\}, \quad (8)$$

which is a scaled version of (7). Finally, for a right-hand side d and $E \in \mathcal{E}$, let $\lfloor d \rfloor_E = A_E \lfloor A_E^{-1} d \rfloor$.

Theorem 1 [3]. For the MILP (1), there is a $g \in \mathcal{G}$ such that

$$z(d) = \min_{E \in \mathcal{E}} g(\lfloor d \rfloor_E) + v_E(d - \lfloor d \rfloor_E) \quad (9)$$

for all $d \in \Omega$.

THE SUBADDITIVE DUAL

The dual problem (5) is rather general, so a natural question is whether it is possible to restrict the space of functions in order to obtain a tractable method for constructing a strong dual function. Note that

$$\begin{aligned} F(d) &\leq z(d), d \in \mathbb{R}^m \\ \iff F(d) &\leq cx, x \in S(d), d \in \mathbb{R}^m \\ \iff F(Ax) &\leq cx, x \in \mathbb{Z}_+^r \times \mathbb{R}_+^{n-r}. \end{aligned} \quad (10)$$

Restricting F to be linear, (5) reduces to $\max\{vb \mid vA \leq c, v \in \mathbb{R}^m\}$ by (10), which is the dual of the LP relaxation of the original MILP. Jeroslow [4] showed that restricting the search space to only convex functions also results in the same optimal solution value as obtained in the linear case. Therefore, these restrictions result in a dual that is no longer guaranteed to produce a strong dual function.

Johnson [5–7] and later Jeroslow [8] developed the idea of restricting the search space to a certain subset of subadditive functions. It is clear that this restriction retains the desirable property of strong duality, since the value function is itself subadditive over Ω .

From (10), if we set $\Gamma^m \equiv \{F : \mathbb{R}^m \rightarrow \mathbb{R} \mid F(0) = 0, F \text{ is subadditive}\}$, then we can

rewrite (5) as the *subadditive dual*

$$\begin{aligned} z_D &= \max F(b) \\ F(\alpha^j) &\leq c_j \quad \forall j \in I, \\ \bar{F}(\alpha^j) &\leq c_j \quad \forall j \in C, \text{ and } \\ F &\in \Gamma^m, \end{aligned} \quad (11)$$

where the function $\bar{F}$ is the *upper d -directional derivative* of F at zero and is defined by

$$\bar{F}(d) = \limsup_{\delta \rightarrow 0^+} \frac{F(\delta d)}{\delta} \quad \forall d \in \mathbb{R}^m. \quad (12)$$

Intuitively, the role of $\bar{F}$ in (11) is to ensure that solutions to (11) have gradients that do not exceed those of the value function near zero, since the subadditivity of F alone is not enough to ensure this in the case of MILP.

The main results concerning the relations between the primal and dual problems in LP duality can be carried to the MILP case.

Theorem 2. *For the primal problem (1) and its subadditive dual (11), the following statements hold:*

1. (Weak duality by Jeroslow [4,8]). Let x be a feasible solution to the primal problem and let F be a feasible solution to the subadditive dual. Then, $F(b) \leq cx$.
2. If the primal problem (respectively, the dual) is unbounded then the dual problem (respectively, the primal) is infeasible.
3. If the primal problem (respectively, the dual) is infeasible, then the dual problem (respectively, the primal) is infeasible or unbounded.
4. (Strong duality by Jeroslow [4,8] and Wolsey [9]) If the primal problem (respectively, the dual) has a finite optimum, then so does the dual problem (respectively, the primal) and they are equal.

The proof of weak duality and the consequent results follow directly from the construction

of dual problem. However, in order to show that strong duality holds for the case where $\Omega \subset \mathbb{R}^m$, we need an additional result of Blair and Jeroslow [10], which states that any given MILP can be “extended” to one that is feasible for all right-hand sides and whose value function agrees with that of the original MILP for all right-hand sides $d \in \Omega$.

Example 2. For the MILP (6), the subadditive dual problem is

$$\begin{aligned} \max \quad & F(b) \\ & F(6) \leq 3 \\ & F(5) \leq \frac{7}{2} \\ & F(-4) \leq 3 \\ & \bar{F}(2) \leq 6 \\ & \bar{F}(-7) \leq 7 \\ & F \in \Gamma^1. \end{aligned} \quad (13)$$

Note from Fig. 1 that $F_2(d) = \max\{\frac{d}{2}, -\frac{3d}{4}\}$ $\forall d \in \mathbb{R}$ is feasible for the dual problem and furthermore, optimal for the right-hand side values $b \in \{\dots, -8, -4, 0, 6, 12, \dots\}$.

The subadditive dual (11) can also be used to extend familiar concepts such as *reduced costs* and the *complementary slackness* conditions to MILPs. For a given optimal solution F^* to (11), the reduced costs can be defined as $c_j - F^*(\alpha^j)$ for $j \in I$ and $c_j - \bar{F}^*(\alpha^j)$ for $j \in C$. These reduced costs have an interpretation similar to that in the LP case, except that we do not have the same concept of “sensitivity ranges” within which the computed bounds are exact. On the other hand, complementary slackness conditions, if satisfied, yield a certificate of optimality for a given primal–dual pair of feasible solutions and can be stated as follows.

Theorem 3 [7–9,11]. *For a given right-hand side b , let x^* and F^* be feasible solutions to the primal problem (1) and the subadditive dual problem (11). Then, x^* and F^* are*

optimal if and only if

$$\begin{aligned} x_j^*(c_j - F^*(\alpha^j)) &= 0, \forall j \in I, \\ x_j^*(c_j - \bar{F}^*(\alpha^j)) &= 0, \forall j \in C, \text{ and} \\ F^*(b) &= \sum_{j \in I} F^*(\alpha^j)x_j^* + \sum_{j \in C} \bar{F}^*(\alpha^j)x_j^*. \end{aligned} \quad (14)$$

The subadditive duality framework also allows the use of subadditive functions to obtain inequalities valid for the convex hull of the feasible region. It is easy to see that for any $d \in \Omega$ and $F \in \Gamma^m$ with $\bar{F}(\alpha^j) < \infty \forall j \in C$, the inequality

$$\sum_{j \in I} F(\alpha^j)x_j + \sum_{j \in C} \bar{F}(\alpha^j)x_j \geq F(d) \quad (15)$$

is satisfied for all $x \in S(d)$. Using this property, Johnson [5] and later Jeroslow [8] showed that any valid inequality for $S(d)$ is either equivalent to or dominated by an inequality in the form of (15).

As an extension to this result, Bachem and Schrader [11] showed that the convex hull of $S(d)$ can be represented using only subadditive functions and that rationality of A is enough to ensure the existence of such a representation, even if the convex hull is unbounded. For a fixed right-hand side, it is clear that only finitely many subadditive functions are needed to obtain a complete description, since every rational polyhedron has finitely many facets. In addition, Wolsey [12] showed that for PILPs, there exists a finite representation that is valid for all right-hand sides.

Theorem 4 [12]. *For a PILP in the form (1), there exist finitely many subadditive functions F_i , $i = 1, \dots, k$, such that*

$$\text{conv}(S(d)) = \left\{ x : \sum_{j=1}^n F_i(\alpha^j)x_j \geq F_i(d), \right. \\ \left. i = 1, \dots, k, x \geq 0 \right\} \quad (16)$$

for any $d \in \Omega$.

CONSTRUCTING DUAL FUNCTIONS

Dual functions can be grouped into three categories: (i) those obtained from known families of relaxations, (ii) those obtained as a by-product of a primal solution algorithm, such as branch-and-bound, and (iii) those constructed explicitly in closed form using a finite procedure.

Cutting Plane Method

Cutting plane algorithms are a broad class of methods for obtaining lower bounds on the optimal solution value of a given MILP by iteratively generating inequalities (called *cutting planes* or *cuts*) valid for the convex hull of S . In iteration k , the algorithm solves the following linear program:

$$\begin{aligned} \min \quad & cx \\ \text{s.t.} \quad & Ax = b \\ & \Pi x \geq \Pi_0 \\ & x \geq 0, \end{aligned} \quad (17)$$

where $\Pi \in \mathbb{R}^{k \times n}$ and $\Pi_0 \in \mathbb{R}^k$ represents the cutting planes generated so far. The LP dual of (17), that is,

$$\begin{aligned} \max \quad & vb + w\Pi_0 \\ & vA + w\Pi \leq c \\ & v \in \mathbb{R}^m, w \in \mathbb{R}_+^k, \end{aligned} \quad (18)$$

is also a dual problem for the original primal MILP. Assuming that a subadditive representation (15) of each cut is known, the i th cut can be expressed parametrically as a function of the right-hand side $d \in \mathbb{R}^m$ in the form

$$\begin{aligned} \sum_{j \in I} F_i(\sigma_i(\alpha^j))x_j + \sum_{j \in C} \bar{F}_i(\bar{\sigma}_i(\alpha^j))x_j \\ \geq F_i(\sigma_i(d)), \end{aligned} \quad (19)$$

where F_i is the subadditive function representing the cut, and the functions $\sigma_i, \bar{\sigma}_i : \mathbb{R}^m \rightarrow \mathbb{R}^{m+i-1}$ are defined by

- $\sigma_1(d) = \bar{\sigma}_1(d) = d$,
- $\sigma_i(d) = [dF_1(\sigma_1(d)) \cdots F_{i-1}(\sigma_{i-1}(d))] \text{ for } i \geq 2$, and

- $\bar{\sigma}_i(d) = [d\bar{F}_1(\bar{\sigma}_1(d)) \cdots \bar{F}_{i-1}(\bar{\sigma}_{i-1}(d))]$ for $i \geq 2$.

Furthermore, if (v^k, w^k) is a feasible solution to (18) in the k th iteration, then the function

$$F_{CP}(d) = v^k d + \sum_{i=1}^k w_i^k F_i(\sigma_i(d)) \quad (20)$$

is a feasible solution to the subadditive dual problem (11).

Wolsey [9] showed how to construct a dual function optimal to the subadditive dual for a given PILP using the Gomory fractional cutting plane algorithm under the assumption that cuts are generated using a method guaranteed to yield a sequence of LPs with lexicographically increasing solution vectors (this method is needed to guarantee termination of the algorithm in a finite number of steps with either an optimal solution or a proof that the original problem is infeasible). In Gomory's procedure, the subadditive function F_i , generated for iteration i , has the following form:

$$F_i(d) = \left[\sum_{k=1}^m \lambda_k^{i-1} d_k + \sum_{k=1}^{i-1} \lambda_{m+k}^{i-1} F_k(d) \right],$$

where $\lambda^{i-1} = (\lambda_1^{i-1}, \dots, \lambda_{m+i-1}^{i-1}) \geq 0$. (21)

Assuming that $b \in \Omega$, $z(b) > -\infty$, and that the algorithm terminates after k iterations, the function F_G defined by

$$F_G(d) = v^k d + \sum_{i=1}^k w_i^k F_i(d) \quad (22)$$

is optimal to the subadditive dual problem (11).

In practice, it is generally not computationally feasible to determine a subadditive representation for each cut added to the LP relaxation. An alternative approach that is feasible for some classes of valid inequalities is simply to track the dependency of each cut on the original right-hand side in some other way. As an example of this, Schrage and Wolsey [13] showed how to construct a function tracking dependency on the right-hand side for cover inequalities by expressing the right-hand side of a cut of this type as an

explicit function of the right-hand side of the original knapsack constraint.

Corrected Linear Dual Functions

A natural way in which to account for the fact that linear functions are not sufficient to yield strong dual functions in the case of MILPs is to consider dual functions that consist of a linear term (as in the LP case) and a correction term accounting for the duality gap. One way to construct such a function is to consider the well-known *group relaxation*. Let B be the index set of the columns of a dual feasible basis for the LP relaxation of a PILP and denote the index set of the remaining columns by $N \setminus B$. Consider the function F_B defined as

$$\begin{aligned} F_B(d) = \min \quad & c_B x_B + c_{N \setminus B} x_{N \setminus B} \\ \text{s.t.} \quad & A_B x_B + A_{N \setminus B} x_{N \setminus B} = d \\ & x_B \in \mathbb{Z}^m, x_{N \setminus B} \in \mathbb{Z}_+^{n-m}. \end{aligned} \quad (23)$$

Substituting $x_B = A_B^{-1}d - A_B^{-1}A_{N \setminus B}x_{N \setminus B}$ in the objective function, we obtain the group relaxation (Gomory [14]).

$$\begin{aligned} F_B(d) = c_B A_B^{-1}d - \max \quad & \bar{c}_{N \setminus B} x_{N \setminus B} \\ & A_B x_B + A_{N \setminus B} x_{N \setminus B} = d \\ & x_B \in \mathbb{Z}^m, x_{N \setminus B} \in \mathbb{Z}_+^{n-m}, \end{aligned} \quad (24)$$

where $\bar{c}_{N \setminus B} = (c_B A_B^{-1} A_{N \setminus B} - c_{N \setminus B})$. Here, dual feasibility of the basis A_B is required to ensure that $\bar{c}_{N \setminus B} \leq 0$.

Note that F_B is subadditive and hence feasible for the subadditive dual (11). Observe that $F_B(b) = z(b)$ when there exists an optimal solution to (24) with $x_B \geq 0$.

Another way to construct an optimal solution to the subadditive dual using a linear function with a correction term is given by Klabjan [15].

Theorem 5 [15]. *For a PILP in the form (1), and a given vector $v \in \mathbb{R}^m$, define the function F_v as*

$$\begin{aligned} F_v(d) = vd - \max \{ & (vA_{\mathcal{D}_v} - c_{\mathcal{D}_v})x \\ & | A_{\mathcal{D}_v}x \leq d, x \in \mathbb{Z}_+^{|\mathcal{D}_v|} \}, \end{aligned} \quad (25)$$

where $\mathcal{D}_v = \{j \in I : v\alpha^j > c_j\}$. Then, F_v is a feasible solution to the subadditive dual problem (11) and furthermore, if $b \in \Omega$ and $z(b) > -\infty$, there exists a $v \in \mathbb{R}^m$ such that $F_v(b) = z(b)$.

Klabjan [15] also introduced an algorithm that finds the optimal dual function utilizing a subadditive approach from Burdet and Johnson [16] together with a row generation approach that requires the enumeration of feasible solutions.

Lagrangian Relaxation

A mathematical program obtained by relaxing and subsequently penalizing the violation of a subset of the original constraints, called the *complicating constraints*, is a *Lagrangian relaxation* (Fisher [17]). Suppose that for a given $d \in \mathbb{R}^m$ the inequalities defined by matrix A and right-hand side d are partitioned into two subsets defined by matrices A^1 and A^2 and right-hand sides d^1 and d^2 . Furthermore, let $\mathcal{S}_{LD}(d^1) = \{x \in \mathbb{Z}_+^r \times \mathbb{R}_+^{n-r} : A^1x = d^1\}$. Then, for a given penalty multiplier $v \in \mathbb{R}^{m-l}$, the corresponding Lagrangian relaxation can be formulated as

$$L(d, v) = \min_{x \in \mathcal{S}_{LD}(d^1)} cx + v(d^2 - A^2x). \quad (26)$$

Assuming $z(0) = 0$ and that $x^*(d)$ is an optimal solution to the original MILP with right-hand side d , we have $L(d, v) \leq cx^*(d) + v(d^2 - A^2x^*(d)) = cx^*(d) = z(d) \forall v \in \mathbb{R}^{m-l}$. Thus, the Lagrangian function defined by

$$L_D(d) = \max \{L(d, v) : v \in V\}, \quad (27)$$

with $V \equiv \mathbb{R}^{m-l}$, is a feasible dual function in the sense that $L_D(d) \leq z(d) \forall d \in \Omega$. Observe that for a given $d \in \Omega$, $L(d, v)$ is a concave, piecewise-polyhedral function.

L_D above is a weak dual function in general, but Blair and Jeroslow [18] showed that it can be made strong for PILP problems by introducing a quadratic term.

Theorem 6 [18]. *For a PILP in the form (1), denote the quadratic Lagrangian dual function as*

$$L_D(d, v) = \max_{\rho \in \mathbb{R}_+} L(d, v, \rho), \quad (28)$$

with

$$L(d, v, \rho) = \min_{x \in \mathbb{Z}_+^n} \left\{ (c - vA)x + \rho \sum_{i=1}^m (A_i x - d_i)^2 + vd \right\}, \quad (29)$$

where $v \in \mathbb{R}^m$, $\rho \in \mathbb{R}_+$ and A_i is the i^{th} row of A . Then for a given $v \in \mathbb{R}^m$, $L_D(d, v) \leq z(d) \forall d \in \Omega$. Furthermore, if $b \in \Omega$ and $z(b) > -\infty$, then for any $v \in \mathbb{R}^m$, $L_D(b, v) = z(b)$.

Branch and Bound

Let us assume now that the MILP (1) has a finite optimum and has been solved to optimality by a branch and bound algorithm. Let T be the set of leaf nodes of the resulting search tree. To obtain a bound for the node $t \in T$, we solve the LP relaxation of the following problem:

$$\begin{aligned} z^t(b) = \min \quad & cx \\ \text{s.t.} \quad & Ax = b \\ & x \geq l^t \\ & -x \geq -u^t \\ & x \in \mathbb{Z}^r \times \mathbb{R}^{n-r}, \end{aligned} \quad (30)$$

where $u^t, l^t \in \mathbb{Z}_+^n$ are the branching bounds applied only to the integer variables.

For each feasibly pruned node $t \in T$, let $(v^t, \underline{v}^t, \bar{v}^t)$ be the corresponding solution to the dual of the LP relaxation that is used to obtain the bound that allowed the pruning of node t . Note that such a solution is always available if the LP relaxations are solved using a dual simplex algorithm. For each infeasibly pruned node $t \in T$, let $(v^t, \underline{v}^t, \bar{v}^t)$ be a corresponding dual feasible solution with $v^t b + \underline{v}^t l^t - \bar{v}^t u^t \geq z(b)$ that can be obtained from the dual solution of the parent of node t and a dual ray that makes the dual problem unbounded.

Theorem 7 [9]. *If $b \in \Omega$ and $z(b) > -\infty$, then the function*

$$F_{BC}(d) = \min_{t \in T} \{v^t d + \underline{v}^t l^t - \bar{v}^t u^t\} \quad (31)$$

is an optimal solution to the dual (5).

Note that it is also possible to consider the dual information revealed by the intermediate tree nodes. Schrage and Wolsey [13] give a recursive algorithm to evaluate a dual function that approximates the value function better by considering the relations between intermediate nodes and their offsprings.

Generating Functions

Lasserre [19,20] introduced a different method for constructing the value function of PILPs that utilizes generating functions. This methodology does not fit well into a traditional duality framework, but nevertheless gives some perspective about the role of basic feasible solutions of the LP relaxation in determining the optimal solution of a PILP.

Theorem 8 [21]. *For a PILP in the form (1) with $A \in \mathbb{Z}^{m \times n}$, define*

$$z(d, c) = \min_{x \in S(d)} cx, \quad (32)$$

and let the corresponding summation function be

$$\hat{z}(d, c) = \sum_{x \in S(d)} e^{cx} \forall d \in \mathbb{Z}^m. \quad (33)$$

Then the relationship between z and $\hat{z}$ is

$$\begin{aligned} e^{z(d, c)} &= \lim_{q \rightarrow -\infty} \hat{z}(d, qc)^{1/q} \quad \text{or, equivalently,} \\ z(d, c) &= \lim_{q \rightarrow -\infty} \frac{1}{q} \ln \hat{z}(d, qc). \end{aligned} \quad (34)$$

A closed form representation of $\hat{z}$ can be obtained by solving the two sided $\mathbb{Z}$ -transform $\hat{F} : \mathbb{C}^m \rightarrow \mathbb{C}$ defined by

$$\begin{aligned} \hat{F}(s, c) &= \sum_{d \in \mathbb{Z}^m} s^{-d} \hat{z}(d, c) \\ &= \prod_{j=1}^n \frac{1}{1 - e^{c_j} s^{-a^j}}, \end{aligned} \quad (35)$$

with $s^d = s_1^{d_1} \dots s_m^{d_m}$ for $d \in \mathbb{Z}^m$ and where the last equality is obtained by applying Barvinok's [22] short form equation for summation problems over the domain of

integral points. This equality is well defined if $|s^{a^j}| > e^{c_j}$, $j = 1, \dots, n$ and the function $\hat{z}$ is then obtained by solving the inverse problem

$$\begin{aligned} \hat{z}(d, c) &= \frac{1}{(2i\pi)^m} \int_{|s_1|=\gamma_1} \dots \\ &\times \int_{|s_m|=\gamma_m} \hat{F}(s, c) s^{d-1^m} ds, \end{aligned} \quad (36)$$

where γ is a vector satisfying $\gamma^{a^j} > e^{c_j}$, $j = 1, \dots, n$ and $1^m = (1, \dots, 1) \in \mathbb{R}^m$.

Although it is possible to solve (36) directly by Cauchy residue techniques, the complex poles make this difficult. One alternative is to apply Brion and Vergne's (see Lasserre [21] and Brion and Vergne [23] for details) lattice points counting formula in a polyhedron to get the reduced form, which, for each $d \in \mathbb{R}^m$, is composed of the optimal solution value of the LP relaxation and a correction term. The correction term is the minimum of the sum of the reduced costs of certain nonbasic variables over all basic feasible solutions, obtained by the degree sum of certain real-valued univariate polynomials. Another approach using generating functions is to apply Barvinok's [24] algorithm for counting lattice points in a polyhedron of fixed dimension to a specially constructed polyhedron that includes for any right-hand side the corresponding minimal test set (see De Loera *et al.* [25,26] for details).

CONCLUSION

Considering the fact that computing the value function for just a single right-hand side is an NP-hard optimization problem in general, computing the value function must necessarily be extremely difficult. The theory of duality, in turn, enables us to derive tractable methods for generating dual functions that approximate the value function and are at the core of a variety of MILP optimization algorithms including sensitivity analysis, warm starting, stochastic and multilevel programming algorithms.

It has so far been difficult to replicate for integer programming the functional and efficient implementations of dual methods we have become so accustomed to in the linear programming case. However, with recent advances in computing, an increased attention is being devoted to integer programming duality in the hope that new computational tools can contribute to more successful implementations.

REFERENCES

1. Blair CE, Jeroslow RG. The value function of an integer program. *Math Program* 1982;23:237–273.
2. Blair CE, Jeroslow RG. Constructive characterization of the value function of a mixed-integer program: I. *Discrete Appl Math* 1984;9:217–233.
3. Blair CE. A closed-form representation of mixed-integer program value functions. *Math Program* 1995;71:127–136.
4. Jeroslow RG. Minimal inequalities. *Math Program* 1979;17:1–15.
5. Johnson EL. Cyclic groups, cutting planes and shortest paths. In: Hu TC, Robinson SM, editors. *Mathematical programming*. New York: Academic Press; 1973. pp. 185–211.
6. Johnson EL. On the group problem for mixed integer programming. *Math Program Stud* 1974;2:137–179.
7. Johnson EL. On the group problem and a subadditive approach to integer programming. *Ann Discrete Math* 1979;5:97–112.
8. Jeroslow RG. Cutting plane theory: algebraic methods. *Discrete Math* 1978;23:121–150.
9. Wolsey LA. Integer programming duality: price functions and sensitivity analysis. *Math Program* 1981;20:173–195.
10. Blair CE, Jeroslow RG. The value function of a mixed integer program: I. *Discrete Math* 1977;19:121–138.
11. Bachem A, Schrader R. Minimal equalities and subadditive duality. *SIAM J Control Optim* 1980;18(4):437–443.
12. Wolsey LA. *The b-hull of an integer program*. London: London School of Economics and CORE; 1979.
13. Schrage L, Wolsey LA. Sensitivity analysis for branch and bound integer programming. *Oper Res* 1985;33(5):1008–1023.
14. Gomory RE. Some polyhedra related to combinatorial problems. *Linear Algebra Appl* 1969;2:451–558.
15. Klabjan D. A new subadditive approach to integer programming: theory and algorithms. *Proceedings of the Ninth Conference on Integer Programming and Combinatorial Optimization*; Cambridge, MA. 2002. pp. 384–400.
16. Burdet C, Johnson EL. A subadditive approach to solve integer programs. *Ann Discrete Math* 1977;1:117–144.
17. Fisher ML. The Lagrangian relaxation method for solving integer programming problems. *Manage Sci* 1981;27:1–17.
18. Blair CE, Jeroslow RG. The value function of a mixed integer program: II. *Discrete Math* 1979;25:7–19.
19. Lasserre JB. Generating functions and duality for integer programs. *Discrete Optim* 2005;2(1):167–187.
20. Lasserre JB. Integer programming, duality and superadditive functions. *Contemp Math* 2005;374:139–150.
21. Lasserre JB. Duality and farkas lemma for integer programs. In: Hunt E, Pearce CEM, editors. *Optimization: structure and applications*. Applied optimization series. Berlin, Heidelberg: Kluwer Academic Publishers; 2009.
22. Barvinok AI. Computing the volume, counting integral points, and exponential sums. *Discrete Comput Geom* 1993;10:1–13.
23. Brion M, Vergne M. Residue formulae, vector partition functions and lattice points in rational polytopes. *J Am Math Soc* 1997;10:797–833.
24. Barvinok AI. Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed. *Math Oper Res* 1994;19:769–779.
25. De Loera J, Haws D, Hemmecke R, *et al.* Short rational functions for toric algebra and applications. *J Symb Comput* 2004;38:959–973.
26. De Loera J, Haws D, Hemmecke R, *et al.* Three kinds of integer programming algorithms based on barvinok’s rational functions. 10th International IPCO Conference. *Lecture Notes in Computer Science* 3064. Berlin, Heidelberg: Springer; 2004. pp. 244–255.

INTEGRATED SUPPLY CHAIN DESIGN MODELS

HO-YIN MAK

Department of Industrial
Engineering and Logistics
Management, The Hong Kong
University of Science and
Technology, Clear Water Bay,
Kowloon, Hong Kong

ZUO-JUN MAX SHEN

Department of Industrial
Engineering and Operations
Research, University of
California, Berkeley, California

INTRODUCTION

The past few decades have seen tremendous improvements in logistics processes. Logistics costs in the United States dropped from 16% of GDP in 1981 to below 9% in 2002 [1], signifying the efficiency gains. A large part of this trend can be attributed to the long line of academic and industrial research on logistics management that has emerged since World War II. From the 1990s onward, much of the research focus has been shifted to the broader area of supply chain management (SCM). In addition to the efficient flow and storage of materials studied in logistics, researchers are now dealing with a much wider range of business problems including the management of information, incentives to align conflicting interests, mitigation of risks, and the interfaces with different business functions such as marketing and finance. We have seen much research effort on identifying the best practices to manage and operate a network of facilities (e.g., factories, warehouses, retail stores), parts of which may be owned by different parties, to maximize long-term profitability. However, a necessary condition to ensure the successful implementation of these policies is the existence of a

well-designed facilities network as a backbone for all operations.

Traditionally, researchers and practitioners tackle the *logistics network design* problem using facility location models. Throughout the history of development of the location research field, the most important factor of consideration has been *distance*. All the classical location models, such as the median, center, and covering models, focus on the tradeoff between measures of distance and measures of budget. These correspond to the line-haul shipping and facility investment components of the logistics costs. For an excellent coverage of the field, the reader may refer to the texts by Daskin [2], Drezner [3] and Drezner and Hamacher [4].

In practice, however, the *supply chain* activities at the *tactical* and *operational* levels after a network of facilities is constructed often go far beyond the scope of shipping and investment costs. Classical location models, when applied to supply chain settings, tend to generate myopic solutions because they ignore the subtle effects of these SCM activities. For example, a distance-focused location model would suggest identical network designs for firms selling fashion items, machine spare parts, or grocery, as long as the transportation and facility investment costs are the same. This is clearly inconsistent with the wide range of SCM practices that have been developed to cater to different product natures. For example, spare parts supply chains typically operate under tight delivery time constraints, because of the destructive circumstances of late deliveries, such as shutting down of factories for assembly machines and delays of flights for aircrafts. To model the response and delivery lead times, one must pay attention to the inventory dynamics of the supply chain [5,6]. Because of these special considerations, supply chain design (SCD) models formulated for spare parts applications are usually not directly applicable to other settings, and vice versa.

The emerging stream of research on integrated supply chain design (ISCD) attempts to acknowledge this gap in the classical theory. Since network design decisions have a long-lasting impact on future costs and revenues, it is important to consider activities in the tactical and operational phases when making strategic decisions. This is the main idea behind the family of ISCD models. If tactical and operational costs incurred by future decisions such as holding inventory and distribution management are carefully estimated when making location decisions, it is more likely that the resulting supply chain will generate higher long-term profit. The reason is that some fundamental tradeoffs in supply chain design cannot be adequately reflected by classical location models that focus on distance–budget considerations. For example, the consideration of inventory holding costs alters the fundamental tradeoff between fixed and shipping costs in the classical uncapacitated facility location (UFL) model. Computational studies conducted by Shen *et al.* [7] suggest that, owing to economies of scale in inventory replenishment and risk pooling, it is optimal for a firm that holds inventory at a warehouse to locate fewer warehouses than suggested by the UFL model.

In this article, we illustrate the key ideas of ISCD modeling by reviewing several recent advances. We do not attempt to review the literature exhaustively. The interested reader may refer to Shen [8] for a more comprehensive survey. Owing to the wide coverage of the research stream, some important study areas are inevitably omitted. For example, the design of spare parts inventory networks [5,6,9] is one of the active research areas not being covered in this article. The remainder of this article is organized as follows. In the section titled “Basic Model”, we discuss the basic model upon which various ISCD models can be built. Then in sections titled “Model with Inventory Considerations,” “Multiple Sourcing,” and “Reliability and Robustness Issues”, we review several important models that have influenced the development of the research stream. Finally, in the section titled

“Future Research Directions”, we discuss some promising future research directions.

BASIC MODEL

Among the classical location models, the UFL model (also known as the UFL or the fixed charge model) is frequently used for SCD. It is even introduced in undergraduate textbooks [10] as a basic model for SCD. To formulate this model, we first introduce the following notations:

SETS

I = set of retailers

J = set of candidate warehouse locations

COST PARAMETERS

f_j = fixed (annualized) cost of opening a warehouse at location j

μ_i = demand volume at retailer i

d_{ij} = shipping cost per unit demand from location j to retailer i

DECISION VARIABLES

X_j = 1 if a warehouse is located at j , 0 otherwise;

Y_{ij} = 1 if demand at retailer i is assigned to warehouse at j , 0 otherwise.

Then the UFL model can be formulated as the following integer program:

$$(\text{UFL}) : \min \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} \mu_i d_{ij} Y_{ij} \quad (1)$$

subject to

$$\sum_{j \in J} Y_{ij} = 1, \text{ for each } i \in I \quad (2)$$

$$X_j \geq Y_{ij}, \text{ for each } i \in I, j \in J \quad (3)$$

$$X_j \in \{0, 1\}, \text{ for each } j \in J \quad (4)$$

$$Y_{ij} \in \{0, 1\}, \text{ for each } i \in I, j \in J. \quad (5)$$

In the above, the objective (1) is to minimize the sum of fixed location and variable shipping costs. The constraints (2) require that each retailer be assigned to one warehouse, and constraints (3) require that a warehouse be opened if it is used to serve any demand. As discussed in the section titled “Introduction”, the model focuses on the strategic tradeoff between fixed costs of opening warehouses and the distance, or shipping costs, from warehouses to retailers (i.e., customers). Note that our framework can also be used to model different type of facilities. For example, instead of warehouses and retailers, we may modify our model to study the location of factories that serve distribution centers.

To account for impacts of future SCM activities, it is natural to extend the UFL by adding terms that reflect the costs incurred in the tactical and operational phases. For example, we may use $G_j(\mathbf{Y})$ to denote the tactical and operational (except shipping) costs for a warehouse at site j , given demand assignments according to $\mathbf{Y}$.¹ The formulation now becomes

$$\begin{aligned} \min \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} \mu_i d_{ij} Y_{ij} \\ + \sum_{j \in J} G_j \left(\sum_{i \in I} \mu_i Y_{ij} \right) \end{aligned} \quad (6)$$

subject to Equations (2–5).

In the following sections, we discuss several representative models that fit the above general framework.

MODEL WITH INVENTORY CONSIDERATIONS

The UFL model captures the basic tradeoff between facility investment costs and shipping costs. However, this simple tradeoff fails to cover certain crucial issues in a

supply chain setting. In particular, tactical inventory management costs at warehouses exhibit economies of scale with respect to demand volume. In classical inventory theory [11], such economies of scale are achieved through economic ordering for cycle stock replenishments and risk pooling of safety stocks. In this section, we discuss a variant of model (6) that captures such impacts of inventory management on facility location decisions. The model was originally proposed by Shen *et al.* [7].

We begin by modifying and defining the following parameters:

DEMAND PARAMETERS

- μ_i = mean (daily) demand at retailer i
- σ_i^2 = variance of (daily) demand at retailer i
- α = desired probability of having no stock-outs during the lead time of a replenishment cycle
- z_α = standard normal deviate, where $P(Z \leq z_\alpha) = \alpha$
- L = order lead time at warehouse

COST PARAMETERS

- h_j = inventory holding cost at site j per unit per year
- F_j = fixed administrative and handling cost of placing an order at warehouse j
- g_j = fixed shipment cost per shipment from supplier to warehouse j
- $\bar{a}_j$ = variable cost per unit shipment from supplier to warehouse j
- χ = number of days in a year

In this model, we assume that retailers face independent, normally distributed demand. The goal is to decide the number and locations of warehouses, considering the resulting investment, shipment, and inventory costs. Inventory is controlled at each warehouse using a continuous review (r, Q) policy, subject to Type 1 service level requirements (α) . To simplify the ordering decisions, the cycle order quantities are determined by the economic order quantity (EOQ)

¹In this article, a letter in boldface represents the vector or matrix of variables denoted by the same letter. For example, $\mathbf{Y}$ represents the matrix with elements being the Y_{ij} 's.

based on the mean demand served by the warehouse.

Cycle Stock Costs

We first note that the daily expected demand going through warehouse j is given by $D_j = \sum_{i \in I} \mu_i Y_{ij}$. Suppose that the number of replenishments in a year at warehouse j is n_j . Then the average shipment size is $\chi D_j / n_j$ and the average inventory on-hand is approximately $\chi D_j / (2n_j)$. Each shipment from the supplier to warehouse j incurs a cost of $g_j + \bar{a}_j \chi D_j / n_j$, which consists of a fixed and a variable component. Then the annual working inventory cost at warehouse j is given by

$$F_j n_j + (g_j + \bar{a}_j \chi D_j / n_j) n_j + h_j \chi D_j / (2n_j).$$

We may observe that the expression above is similar in structure to the cost expression in the EOQ problem. Therefore, the optimal number of replenishments per year is given by

$$n_j^* = \sqrt{\frac{h_j \chi D_j}{2(F_j + g_j)}}.$$

Correspondingly, the annual working inventory cost associated with warehouse j is given by:

$$\sqrt{2h_j \chi D_j (F_j + g_j)} + \bar{a}_j \chi D_j.$$

Safety Stock Cost

We now formulate the safety stock cost for the model. The classical newsvendor model [12] suggests that the amount of safety stock needed to meet a Type-1 service level of α is given by z_α times the standard deviation of demand during lead time. Using our notation, the holding cost for the safety stock at warehouse j is then given by

$$h_j z_\alpha \sqrt{L \sum_{i \in I} \sigma_i^2 Y_{ij}}. \quad (7)$$

Complete Model Formulation

Adding the above cost expressions to a UFL-type formulation, we obtain the following

ISCD model:

$$\begin{aligned} \min \sum_{j \in J} & \left[f_j X_j + \sum_{i \in I} (d_{ij} + a_j) \chi \mu_i Y_{ij} + \sqrt{2h_j (F_j + g_j)} \right. \\ & \times \left. \sqrt{\sum_{i \in I} \chi \mu_i Y_{ij} + h_j z_\alpha \sqrt{\sum_{i \in I} L \sigma_i^2 Y_{ij}}} \right] \\ = \sum_{j \in J} & \left[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + K_j \sqrt{\sum_{i \in I} \mu_i Y_{ij}} \right. \\ & \left. + \eta_j \sqrt{\sum_{i \in I} \sigma_i^2 Y_{ij}} \right] \end{aligned} \quad (8)$$

subject to Equations (2–5)
where

$$\begin{aligned} \hat{d}_{ij} &= (d_{ij} + a_j) \chi \mu_i \\ K_j &= \sqrt{2h_j (F_j + g_j)} \\ \eta_j &= h_j z_\alpha \sqrt{L}. \end{aligned}$$

Solution Approaches

The above model has been studied by Shen *et al.* [7] and Daskin *et al.* [13]. One interesting property uncovered was that unlike the traditional UFL model, in which it is always optimal to assign a retailer to the open warehouse that incurs the lowest shipping cost, it is sometimes optimal to assign demand to more remote warehouses in the new model in order to achieve economies of scale by inventory sharing. In the case where the sets of retailer and candidate warehouse locations are the same ($I = J$), it is even possible that a warehouse does not even serve a retailer at the same location. These pathological situations, however, do not occur when the mean–variance ratio of demand is the same for all retailers (i.e., $\sigma_i / \mu_i = \gamma, \forall i \in I$). This condition holds in many practical settings. For example, if demands follow Poisson distributions, the mean–variance ratio will be equal to 1 for all retailers. Both papers develop algorithms based on this assumption, which allows the objective function to

be rewritten as

$$\min \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + \hat{K}_{ij} \sqrt{\sum_{i \in I} \mu_i Y_{ij}} \right],$$

where $\hat{K}_j = K_j + q_j \sqrt{\gamma}$.

The algorithms developed in the two papers are closely related. Shen *et al.* [7] utilize a column generation approach, which involves reformulating the problem as a set-covering problem, solving restrictions of the problem with only a subset of variables, and successively adding variables that reduce the cost to improve the formulation. Daskin *et al.* [13], on the other hand, improve the computational results by using a Lagrangian relaxation procedure by relaxing constraints (2). For a tutorial on the Lagrangian relaxation approach, the interested reader may refer to Fisher [14]. In both algorithms, a subproblem of the following form (or pricing problem) has to be solved in each iteration, for each candidate site $j \in J$:

$$\min_{Z \in \{0,1\}^{|I|}} \sum_{i \in I} b_i Y_i + \sqrt{\sum_{i \in I} c_i Y_i}, \quad (9)$$

where $c_i \geq 0$, but $b_i < 0$. A lower order polynomial time procedure, developed by Shen *et al.*, was utilized in both algorithms to solve the above subproblem. The procedure is based on the special structure of the above subproblem. The first term in the objective function is linearly decreasing in Y_i , while the second term is concave increasing in Y_i . Intuitively, this structure of this subproblem is somewhat related to that of the knapsack problem, in the sense that one wants to activate variables (set Y_i to 1) to improve the cost linearly, while limited by a penalty that works in a similar fashion as the capacity constraint in the knapsack problem. On the basis of this observation, Shen *et al.* prove that after sorting the i 's by the b_i/c_i ratio, it is possible to identify an optimal solution by checking at most $|I|$ candidate solutions. Therefore, the complexity of solving the subproblem for each $j \in J$ is $O(|I| \log |I|)$ and that of solving all $|J|$ subproblems in one iteration of column generation or Lagrangian relaxation is $O(|I||J| \log |I|)$.

To solve the problem without the restriction that all retailers have equal mean-variance ratio for demand, Shu *et al.* [15] build on the Shen *et al.* column generation algorithm. The resulting pricing problem is similar to Equation (9), but contains one more square root term in the objective function. Exploiting the fact that the subproblem is equivalent to minimizing a submodular set function together with other structural properties, they develop an efficient algorithm that solves the subproblem in $O(|I|^2 \log |I|)$ time. Solving all $|J|$ subproblems in one iteration thus has complexity of $O(|I|^2 |J| \log |I|)$.

Recently, Berenguer *et al.* [16] studied an ISCD model with unreliable supply and transportation. Their solution approach involves reformulating ISCD models as mixed-integer second-order cone programs (MISOCPs), the continuous relaxations of which can be efficiently solved with conic programming solvers in polynomial time. In addition, several classes of cutting planes [17–19] can be used to exploit the special structure of MISOCPs and speed up the branch-and-bound procedure. It is interesting to note that problem (6) can also be written as the following MISOCP:

$$\min \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + K_j V_j + q W_j \right],$$

subject to

$$V_j \geq \sqrt{\sum_{i \in I} \mu_i Y_{ij}^2}, \quad (10)$$

$$W_j \geq \sqrt{\sum_{i \in I} \sigma_i^2 Y_{ij}^2}, \quad (11)$$

plus Equations (2–5).

Inequalities (10) and (11) are second-order conic constraints. Note that the two formulations are equivalent because $Y_{ij} = Y_{ij}^2$ for $Y_{ij} \in \{0, 1\}$.

Key Insights

One might question how much this more complex model improves over the relatively simpler UFL. Computational comparisons

between these models show that the solution obtained by solving the UFL always overestimates the number of facilities needed. This is due to the fact that the UFL does not account for the economies of scale in inventory costs. From the computational experiments conducted by Shen *et al.* [7], supply chain networks based on UFL solutions, on average, incur about 7% extra cost compared to the optimal solution obtained by solving the model discussed above. This suggests the potential benefits of building more sophisticated ISCD models to capture the underlying dynamics that are not accounted for in the classical models.

Extensions

Impact of Capacity Constraints. In practical settings, warehouses typically have limited capacity. The traditional way to model warehouse capacity is to constrain the maximum demand volume that can be assigned to a facility, as is done in the classical capacitated facility location (CFL) problem. However, in a supply chain setting, capacity restrictions at a warehouse depend more on the maximum inventory level than on (average) demand volume. On the basis of this observation, Ozsen *et al.* [20] formulate a capacitated version of model (6). Their capacity constraint is defined on the basis of the maximum level of inventory kept at the warehouse. This constraint allows an additional layer of freedom in demand assignment, because instead of strictly prohibiting the assignment of high demand volumes to warehouses, this model allows warehouses handling high demand volumes to lower peak inventory levels by ordering frequently in small lots. As a result, the model captures the tradeoff between locating more warehouses to ensure efficient capacity utilization versus ordering smaller lots to operate under capacity.

Recall that inventory is managed at each warehouse using an (r, Q) policy. Note that the absolute maximum inventory level at a warehouse is reached in the following (unlikely) situation: after a reorder request, the demand during the lead time is zero. Therefore, the maximum inventory level is equal to the reorder point plus the order quantity. Denoting the capacity and cycle

order quantity at warehouse j by C_j and Q_j , respectively, the capacity constraint can be written as

$$Q_j + z_\alpha \sqrt{L \sum_{i \in I} \sigma_i^2 Y_{ij}} + \sum_{i \in I} L \mu_i Y_{ij} \leq C_j. \quad (12)$$

Recall that for (r, Q) inventory systems, the reorder point is equal to the safety stock plus expected lead time demand. These are equal to the second and third terms on the left hand side of Equation (12), respectively.

To solve this problem, Ozsen *et al.* extend the Lagrangian relaxation algorithm of Daskin *et al.* [13]. Their subproblem has a similar form to Equation (9), with the square root function replaced by a function that is neither convex nor concave. They prove that each of the new subproblems can be solved with the sorting algorithm [7] with an additional subroutine to search for roots in at most $|I|$ subintervals within a line segment. They then modify the Lagrangian relaxation algorithm of Daskin *et al.* accordingly to solve the capacitated SCD problem.

Impact of Operational Shipment Consolidation. In daily logistics operations where demand points have relatively small demand volumes, it is possible to lower delivery cost by consolidating shipments using milk runs, that is, delivering to multiple demand points with the same vehicle on the same route. Therefore, when retailers have low-demand volumes and/or place frequent replenishment requests, there is a possibility to achieve greater economies of scale using vehicle routing. Recall that in all the previous models, it is assumed that shipping costs from warehouses to retailers are linear in demand volumes. To capture the impact of such operational economies of scale on the optimal network configuration, it is necessary to further extend the objective function to include operational vehicle routing costs.

Like most facility location problems, the vehicle routing problem (VRP) is NP-hard. Owing to the high complexities of both problems, it is not practical to directly combine a vehicle routing problem and an ISCD problem. Fortunately, we are mainly concerned about the strategic design aspects rather

than the operational details at this stage. Therefore, it suffices to approximate the operational delivery costs by some tractable functions of demand volumes. The concept of continuous approximation in logistics analysis (see Daganzo [21] for a thorough treatment of the subject) is then very helpful. By using a concise summary of problem characteristics (e.g., the total demand volume and the number of retailers), it is possible to obtain simple tractable approximations for analysis or optimization. One such approximation for the optimal VRP length (V_j), that is, total distance traveled by all vehicles needed to visit all retailers once, is given by Haimovich and Rinnooy Kan [22]:

$$V_j \approx \frac{2}{\kappa} \sum_{i=1}^m d_{ij} \mu_i + \left(1 - \frac{1}{\kappa}\right) \psi_j. \quad (13)$$

In the above equation, m is the number of retailers visited, κ is the capacity of a vehicle, and ψ_j can be interpreted as a local detour distance. Intuitively, the first term gives the total distance traveled by all vehicles from the depot to the respective local neighborhoods and back. The second term (the detour lengths) represents the distance traveled within the respective local neighborhoods to visit all local retailers. Furthermore, based on the results by Daganzo [23], Shen and Qi [24] show that if N retailers are uniformly distributed in a region of area A , m of which are visited and m is large (> 100), then the expected detour length ψ_j can be approximated by

$$\psi_j \approx \rho m \sqrt{\frac{A}{N}}. \quad (14)$$

In the above, ρ is a constant that equals 0.75 for the Euclidean metric. Combining Equations (13) and (14), the VRP cost can then be approximated by

$$V_j \approx \frac{2}{\kappa} \sum_{i=1}^m d_{ij} \mu_i + \left(1 - \frac{1}{\kappa}\right) \rho m \sqrt{\frac{A}{N}}. \quad (15)$$

Shen and Qi assume that there is a dedicated truck in each warehouse that delivers to the retailers every period according to a certain route, and that the transportation

cost related to this route is an increasing and concave function $R_j(T_j)$ in the distance traveled, under reasonable conditions (e.g., driver does not have to work overtime). Then, the new ISCD problem incorporating strategic location, tactical inventory control, and operational vehicle routing elements can be formulated as

$$\min \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + \hat{K}_{ij} \sqrt{\sum_{i \in I} \mu_i Y_{ij}} + R_j \left(\sum_{i \in I} \hat{b}_{ij} Y_{ij} \right) \right],$$

where $\hat{b}_{ij} = 2\mu_i d_{ij}/\kappa + \chi(1 - 1/\kappa)\rho\sqrt{\frac{A}{N}} \geq 0$.

Shen and Qi further develop a Lagrangian relaxation algorithm by exploiting the structure of a subproblem that is similar to Equation (9), but involves three concave terms. They generalize the procedure proposed by Shu *et al.* [15] to solve the subproblem in $O(|I|^2 \log |I|)$ time.

MULTIPLE SOURCING

One key assumption made in the models discussed in the section titled “Model with Inventory Considerations” is that all retailers use single sourcing, that is, receive shipments from only one warehouse. In practical settings, such an arrangement may reduce operational complexity, because demand allocation rules are relatively simple. However, this simplicity does not come without a cost. First, the single sourcing requirement restricts efficient utilization of capacity. Second, a single sourcing arrangement is not responsive to dynamic changes of demand. In this section, we discuss models that address these two issues, respectively. Both models allow multiple sourcing arrangements, in which each retailer can receive shipments from multiple warehouses in such a way that minimizes cost.

Enhancing Capacity Management by Static Multiple Sourcing

In classical UFL problems such as the UFL and P-median problems, there exist single

sourcing solutions that are optimal even if multiple sourcing is allowed. However, when the facilities have capacity limits, multiple sourcing allows splitting of a retailer's demand to fit within the assigned warehouses' capacity limits and therefore achieves lower costs. In the ISCD setting, we have discussed in the section titled "Key Insights" how warehouse capacity can be defined on the basis of storage space for inventory. We now discuss the multiple sourcing version of the same model, as developed by Ozsen *et al.* [25].

The multiple sourcing arrangements allowed in this model are static, meaning that a retailer's demand has to be split among multiple warehouses in fixed proportions, which are to be determined endogenously together with the location decisions. For example, at the SCD stage, it is determined that a retailer will place half of its orders at warehouse 1 and the other half at warehouse 2, and will not adjust this ratio over time. The reason for maintaining such consistent sourcing ratios is to efficiently allocate demand to warehouses such that capacities can be utilized at a low cost manner. As done in the classical CFL, static multiple sourcing can be modeled by allowing the demand assignment variables (Y_{ij}) to be fractional instead of binary. Therefore, the model is identical to Equation (6), except that constraints (5) are replaced by

$$0 \leq Y_{ij} \leq 1, \text{ for each } i \in I, j \in J. \quad (16)$$

The solution algorithm is similar to the one designed for the single-sourcing version. The authors show that, despite the freedom to split demands, it is optimal to do only a limited degree of splitting. For example, there exists an optimal solution if the number of multisourced retailers is smaller than the number of warehouses open.

Responding to Demand Fluctuations by Dynamic Sourcing

The static multiple sourcing arrangement discussed in the previous section makes demand assignment a strategic decision that cannot respond to fluctuations of demand and other problem parameters. As SCD decisions usually govern long-term supply

chain performance and accurate forecasts for future market conditions are seldom available, it is important to insert certain degrees of flexibility in the network configuration. One possible example of flexibility is to use multiple sourcing, and allow sourcing proportions to be determined subject to realization of the initially unknown parameters, such as demand volumes. Mak and Shen [26] formulate a two-stage ISCD model to study the cost-benefit tradeoff for such flexibility.

In the first (*design*) stage, the network structure, that is, location of DCs and the allocation of retailers to DCs, is to be determined. Then in the second (*management*) stage, the inventory and shipment decisions are made for a number of periods given the network structure determined in the design stage. We first introduce the following additional notations:

Cost and Other Parameters

- c_{ij} = fixed (annualized) cost of creating an arc connecting DC j and retailer i
- T = number of periods in management stage
- D^t = demand vector in period t , following a known probability distribution

Decision Variables

- $Z_{ij} = 1$ if we build an arc between DC j and retailer i , 0 otherwise.

The two-stage stochastic programming problem can be formulated as follows:

$$\min \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} c_{ij} Z_{ij} + v(\mathbf{X}, \mathbf{Z}, D^1, \dots, D^T) \quad (17)$$

subject to

$$\begin{aligned} \sum_{j \in J} Z_{ij} &\geq 1, \text{ for each } i \in I \\ X_j &\geq Z_{ij}, \text{ for each } i \in I, j \in J \\ X_j &\in \{0, 1\}, \text{ for each } j \in J \\ Z_{ij} &\in \{0, 1\}, \text{ for each } i \in I, j \in J. \end{aligned}$$

In the above formulation, $v(\cdot)$ is defined to be the expected recourse function, which includes the inventory, shipping, and shortage costs in the management stage. Mak and Shen [27] provide the detailed formulation and prove various properties, including the optimality of using stationary base stock policies at the warehouses. They provide an efficient algorithm to obtain high-quality solutions to the problem.

From computational experiments, Mak and Shen obtain the following insights. First, adding sourcing flexibility can dramatically enhance the network's ability to mitigate demand risks. Second, by adding a small degree of flexibility, it is possible to achieve most of the benefits of complete flexibility (i.e., any retailer can receive shipments from any warehouse). This finding echoes the classical results in process flexibility [28,29]. Third, the optimal degree of flexibility increases with demand variability and decreases with demand correlation across retailers.

RELIABILITY AND ROBUSTNESS ISSUES

SCD involves long-term decisions that are difficult to reverse. Therefore, the network that we design today will be subject to changes in the business environment in the future years and even decades. Unfortunately, it is well known as one of the fundamental principles of forecasting that the longer the forecast horizon, the less accurate the forecast will be [30, Chapter 2]. Unlike operational decisions such as machine scheduling, strategic network design decisions have to be made on the basis of imprecise long-term forecasts of input parameters such as fuel cost and demand rate. Therefore, it is in the best interest of the organization to design a robust network that performs reasonably well under any possible realization of the uncertain parameters for which only rough estimates are available.

There has been a significant amount of work on facility location under uncertainty. Snyder [31] provides an excellent review of the subject. However, these models, just like the classical location models, do not

account for the implications of tactical and operational activities of the system. On the other hand, there has been relatively little work in the ISCD stream that involves uncertainty issues. In the section titled "Parameter Uncertainty and Scenario Analysis", we first discuss an ISCD model that takes the uncertainty of input parameters into account. Then, we move on to models that handle two important risk factors, namely, the randomness of supply yield and potential disruptions of facilities, in sections titled "Unreliable Supply" and "Dynamic Sourcing under Disruptions," respectively. It is worth noting that the three categories of uncertainty models are closely related, yet distinct. Parameter uncertainty models (section titled "Parameter Uncertainty and Scenario Analysis") assume that certain input parameters follow certain probability distributions (or lie within certain support intervals), and optimize either the expected or the worst-case performance accordingly. The other two categories focus on the unreliability of supply. Random yield models (section titled "Unreliable Supply") assume that the amount of products received by a retailer or a customer is a random proportion of the ordered amount. Disruption models (section titled "Dynamic Sourcing under Disruptions") assume that, with a certain probability, a warehouse cannot function (i.e., is disrupted) and supplies zero amount to retailers that it is supposed to serve. It is possible to model unreliable supply using parameter uncertainty, by assuming that the supply level is one of the uncertain parameters. It is also possible to model disruptions with random yield, by assuming that the yield follows a Bernoulli (all-or-nothing) distribution. However, these approaches are not desirable, because they do not take into account specific characteristics of the respective problems. For example, strategies such as keeping safety stocks are effective measures to adapt to yield uncertainty, while parameter uncertainty models do not take this into account. On the other hand, as has been pointed out by Chopra *et al.* [32], recurrent supply risk (random yield) and disruption risks are different fundamentally

in terms of mitigation strategies and ought not be confused.

Parameter Uncertainty and Scenario Analysis

Addressing the issue that it is impossible to accurately predict the future business environment at the design stage, Snyder *et al.* [33] formulate a scenario-based model. Scenario-based optimization is a key idea that is widely used in stochastic programming [34], and is consistent with scenario-based planning frequently adopted by managers. With some parameters (e.g., demand volume) being uncertain at the time of design, we may define a finite set of possible realizations or scenarios (denoted by S), and assign a probability or weight ($q_s, s \in S$) to each of them. The objective is to minimize the expected cost over this distribution of scenarios. In this model, the facility location decisions are strategic decisions that must be made before the scenario realization is observed, based on the probability distribution discussed above. Then, when the supply chain network is in operation and the scenario realization is observed, the manager may adjust the retailer-warehouse assignments to minimize cost.

There are two possible interpretations for this framework. First, we may assume that one of the set of parameter scenarios will be realized, and the parameter values remain as they are from then on. Second, we may assume that the parameters are dynamically changing over time, and the long-term proportion of time that each scenario s is realized is equal to the probability q_s . To formulate this model, we redefine the following uncertain parameters and recourse decision variables to make them dependent on the realized scenario:

Parameters

- μ_{is} = mean daily demand at retailer i under scenario s
- σ_{is}^2 = variance of daily demand at retailer i under scenario s
- d_{ijs} = shipping cost per unit from warehouse j to retailer i under scenario s

Decision Variables

$Y_{ijs} = 1$ if demand at retailer i is assigned to warehouse at j when scenario s is realized, 0 otherwise.

Then the scenario-based problem can be formulated as follows:

$$\sum_{s \in S} q_s \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} \hat{d}_{ijs} Y_{ijs} + K_j \sqrt{\sum_{i \in I} \mu_{is} Y_{ijs}} + \eta_j \sqrt{\sum_{i \in I} \sigma_{is}^2 Y_{ijs}} \right]$$

subject to

$$\begin{aligned} \sum_{j \in J} Y_{ijs} &= 1, \text{ for each } i \in I, s \in S; \\ X_j &\geq Y_{ijs}, \text{ for each } i \in I, j \in J, s \in S; \\ X_j &\in \{0, 1\}, \text{ for each } j \in J; \\ Y_{ijs} &\in \{0, 1\}, \text{ for each } i \in I, j \in J, s \in S; \end{aligned}$$

where $\hat{d}_{ijs} = (d_{ijs} + a_j) \chi \mu_{is}$.

Snyder *et al.* [33] devise a Lagrangian relaxation algorithm to solve the problem. The subproblem at each iteration turns out to be identical to the one encountered in solving the model in the section titled “Model with Inventory Considerations”. Their computational results show that the solution obtained by solving the scenario-based model is often very different from the ones obtained by solving the problem assuming no uncertainty. The latter solutions typically perform poorly in the long run in the presence of parameter uncertainty.

Independently, Shen [35] develops a similar model with a different interpretation. Instead of being scenarios of realizations of uncertainty, the set S represents the collection of different commodities sold by the firm in Shen’s model. These commodities have separate demand volumes and inventories, but are handled at the same warehouses. Comparing the two models, Shen’s is slightly more general because it uses a general increasing and concave inventory cost term, which includes Snyder *et al.*’s EOQ-based inventory cost function as a special case.

Unreliable Supply

We have seen that firms that were not prepared to respond to unreliable supplies, such as Ericsson in the 2000 Albuquerque fire, suffered huge losses and lost their market leadership. As the backbone of all supply chain operations, the supply chain network must be designed by taking supply risk factors into account to ensure ample supply.

Qi and Shen [36] study a problem in which supply is not perfectly reliable. They consider a multiple-period problem in which the amount that a retailer receives from its assigned warehouse is a random fraction (the reliability coefficient) of the original order quantity. This assumption is also known as *random yield* [37] in the production control literature. Besides facility location costs, working inventory costs, safety stock costs, and shipping costs that have been considered in the section titled “Model with Inventory Considerations”, the Qi and Shen model also takes into account the penalty costs of not meeting the retailers’ demand in full. We begin by introducing the following additional notations:

Parameters

- c = purchasing price per unit of product from supplier
- τ = number of periods per year
- R_j = reliability coefficient associated with facility j
- p_i = retailer price per unit of product at retailer i
- π_i = penalty (goodwill) cost for missing each unit of demand at retailer i
- u_i = salvage value of product at retailer i

Decision Variables

- ω_{ij} = quantity that retailer i orders from facility j in each period (assumed to be constant over time).

Utilizing the result on inventory control under unreliable supply provided by Dada *et al.* [38], they formulate a model of the following structure:

$$\begin{aligned} \max - \sum_{j \in J} \left[f_j X_j + c \tau \sum_{i \in I} \omega_{ij} + a_j \tau \sum_{i \in I} \omega_{ij} \right. \\ \left. + K_j \sqrt{\sum_{i \in I} \omega_{ij}} \right] + \tau \sum_{i \in I} \Upsilon_i(\mathbf{Q}) \end{aligned}$$

subject to

$$\begin{aligned} 1 - e^{-\beta \omega_{ij}} &\leq X_j, \text{ for each } i \in I, j \in J \\ \omega_{ij} &\geq 0, \text{ for each } i \in I, j \in J \\ X_j &\in \{0, 1\}. \end{aligned} \quad (18)$$

The objective is to maximize the expected annual profit of the entire system. The first two terms represent the fixed location costs and the purchase costs of products. The third and fourth terms represent the sum of inventory and delivery costs. Finally, the last term represents the difference between sales revenues of the retailers and inventory costs. The exact form is derived by Dada *et al.* [38] as follows:

$$\begin{aligned} \Upsilon_i(\mathbf{Q}) = E \left\{ p_i D_i + u_i \left[\sum_{j \in J} R_j \omega_{ij} - D_i \right]^+ \right. \\ \left. - (p_i + \pi_i) \left[D_i - \sum_{j \in J} R_j \omega_{ij} \right]^+ \right. \\ \left. - \sum_{j \in J} d_{ij} R_j \omega_{ij} \right\}. \end{aligned}$$

The constraints (18) ensure that retailers cannot place orders at warehouses that are not opened. In devising an efficient Lagrangian relaxation algorithm to solve the problem, Qi and Shen found that a constraint of this form is more effective than one of the form $MX_j \geq \omega_{ij}$.

Dynamic Sourcing under Disruptions

Following a series of catastrophic events over the past decade, including the September 11 terrorist attack and the strike of Hurricane Katrina, researchers and practitioners have put increasingly more emphasis on ensuring the reliability of supply chain networks.

Owing to the practical significance, strategies to deal with reliable supply have become a key issue in supply chain research. Most of the existing works focus on inventory control [39] and risk mitigation and contingency planning [40,41]. Mitigation strategies are those in which the firm takes action in advance and incurs a cost to prepare for uncertainty. Holding safety stock is an example of a mitigation strategy to adapt to uncertain demand. In contrast, contingency strategies are those in which the firm takes action after some event occurs. For example, emergency inventory transshipment during stock-outs can be regarded as a contingency strategy to adapt to demand uncertainty.

There is also a significant stream of research on facility location under disruption risks. Snyder and Daskin [42] study the optimal network configuration that minimizes a weighted average of regular shipping costs and the expected failure costs in case of disruptions, assuming that probabilities of disruptions at all candidate facility locations are equal. The equal-probability assumption is relaxed in related works by Berman *et al.* [43], who focus on asymptotic properties of the problem; Zhan *et al.* [44], who propose and solve a mixed-integer nonlinear programming version of the problem; Shen *et al.* [45], who formulate and solve a two-stage stochastic program and then a nonlinear integer program and provide an approximation algorithm with constant worst case bound for the special case where the probability that a facility fails is a constant; and Cui *et al.* [46], who provide a linear formulation and study qualitative insights by using a continuous approximation model. Lim *et al.* [47] study a closely related problem in which there is an option to locate disruption-proof, but more expensive, facilities. Despite the contributions of this stream of work, these works do not consider operational and tactical level decisions.

In one of the very few works on ISCD models under disruption risks, Mak and Shen [26] extend the model outlined in the section titled “Responding to Demand Fluctuations by Dynamic Sourcing” to study the impact of

warehouse disruptions in networks with and without the flexibility of dynamic sourcing. It is interesting to see that the dynamic multiple sourcing arrangement discussed in the section titled “Responding to Demand Fluctuations by Dynamic Sourcing” fits the definition of contingency under demand fluctuations. In fact, such an arrangement can also help limiting disruption risks, by dynamically adjusting the demand allocation proportions among the working facilities. The key insight behind their model is the following. Economies of scale in inventory replenishment and risk pooling of demand encourage the location of fewer warehouses, each handling a larger demand volume, as suggested in the section titled “Key Insights”. On the other hand, the risk of disruptions makes risk diversification, by locating more facilities, each handling a smaller demand volume—a desirable strategy. These two effects go against each other and it is difficult to find a balance. However, dynamic sourcing enhances inventory sharing across multiple locations (as suggested by the study covered in the section titled “Responding to Demand Fluctuations by Dynamic Sourcing”). Therefore, it is possible to achieve risk diversification by locating a large number of warehouses, while still achieving risk-pooling benefits using dynamic sourcing. From computational experiments, Mak and Shen show that dynamic sourcing is very effective in limiting losses incurred by facility disruptions. Furthermore, only small degrees of flexibility are needed to hedge against high disruption risks.

FUTURE RESEARCH DIRECTIONS

Interface with Marketing Strategies

Traditionally, logistics and marketing are treated as two separate business functions. Past research has treated the two areas independently. However, one can easily observe the close link between the two: logistics treats demand as given, and optimizes the supply process; marketing treats the supply process as given, and manages demand to maximize revenue. It is clear that decisions in

one function can greatly impact the performance in the other. In fact, some companies that manage to jointly formulate strategies considering both facets fare extremely well. For example, Zara allows frequent stock-outs of its products on purpose to induce consumers to purchase early and manages a supply chain network with redundant capacity to respond quickly to demand changes [48]. In this case, managing the two functions separately would have prohibited such success.

Not until recently has research on the interface between marketing and operations emerged. Some examples of active research areas include rationing and pricing strategies in the presence of strategic consumers [49–51], and the coordination of pricing and inventory decisions [52]. There are thus many promising research questions that fall within the context of retail network design that remain unanswered. For example, how does network design enable or prohibit effective market segmentations based on factors such as geographical location, pricing, product availability, and quality of service? Many of these factors, together with logistics operations such as inventory control and distribution planning, carry implications on the optimal network configuration. It will be a challenge to model and analyze the subtle relationships between these activities and the resulting implications on the optimal network configuration.

It is also important to study the impact of competition. Research on this topic would help us better understand some interesting phenomena such as the cluster entry strategy of 7-Eleven Japan of opening 50–60 stores at a time when entering new markets where competition exists [53]. Much of the existing research on competitive facility location (see Drezner [3] for a comprehensive review) focuses on capturing consumers from competitors by geographical proximity, capacity, and exogenous factors such as brand name. However, in supply chain settings, firms compete on many other factors, such as product offering and availability, logistics services, and even loyalty programs. Some of these factors have subtle implications on the strategic design

of the retail network, while some others may not. To synchronize the firm's marketing and logistics strategies with the network design decisions, one ought to investigate the potential impacts of these marketing and operation activities and carefully select the most important ones to be endogenized in the design model.

Designing Effective Implementation Mechanisms

The research directions previously outlined make the assumption that all design and operation decisions are made by a single manager with the goal of maximizing total supply chain profit. However, horizontal (e.g., an alliance of retailers) and vertical (e.g., a supply chain consisting of a retailer, distributor, and manufacturer) cooperation is common in modern supply chains. In such settings, there are multiple distinct parties with possibly conflicting interests. In an attempt to maximize the total supply chain profit, it is necessary to provide incentives such that all parties commit to a centralized solution. The centralized optimal solution is often impossible to implement, but it is usually possible to coordinate a solution in which every party can benefit compared to the case where decisions are completely centralized.

There exists a stream of research on incentives to induce inventory sharing [54]. However, there has been very limited research on cooperative SCD. The existing research on cooperative facility games [55] mostly deals with simple location problems such as the UFL and set-covering problems. Moreover, the focus of such research is mainly on algorithmic aspects such as computing equilibrium points, rather than obtaining managerial insights on how to implement such mechanisms in practice.

REFERENCES

1. MacroSys Research and Technology. Logistics costs and U.S. gross domestic product. 2005. Available at http://ops.fhwa.dot.gov/freight/freight_analysis/econ_methods/lcdp_rep/index.htm. Accessed 2007 April 5.
2. Daskin MS. Network and discrete location: models, algorithms, and applications. New York: John Wiley & Sons, Inc.; 1995.

3. Drezner Z, editor. Facility location: a survey of applications and methods. New York: Springer; 1995.
4. Drezner Z, Hamacher HW, editors. Facility location: applications and theory. New York: Springer; 2002.
5. Candas MF, Kutanoglu E. Benefits of considering inventory in service parts logistics network design problems with time-based service constraints. *IIE Trans* 2007;39(2):159–176.
6. Mak HY, Shen ZJ. Dynamic sourcing in supply chain design. Working Paper. Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology; 2009a.
7. Shen ZJ, Coullard C, Daskin MS. A joint location-inventory model. *Transp Sci* 2003; 37: 40–55.
8. Shen ZJ. Integrated supply chain design models: a survey and future research directions. *J Ind Manage Optim* 2006;3(1):1–27.
9. Jeet V, Kutanoglu E, Partani A. Logistics network design with inventory stocking for low-demand parts: modeling and optimization. *IIE Trans* 2009;41(5):389–407.
10. Chopra S, Meindl P. Supply chain management. 4th ed. Upper Saddle River (NJ): Prentice Hall; 2009.
11. Zipkin PH. Foundations of inventory management. Boston (MA): McGraw-Hill; 2000.
12. Axsäter S. Inventory control. 2nd ed. New York: Springer; 2006.
13. Daskin MS, Coullard C, Shen ZJ. An inventory-location model: formulation, solution algorithm and computational results. *Ann Oper Res* 2002;110:83–106.
14. Fisher M. An applications oriented guide to lagrangian relaxation. *Interfaces* 1985; 15(2):2–21.
15. Shu J, Teo CP, Shen ZJ. Stochastic transportation-inventory network design problem. *Oper Res* 2005;53(1):48–60.
16. Berenguer G, Atamtürk A, Shen ZJ. Conic programming for a capacitated location-inventory model with unreliable supply and transportation. Working Paper. Berkeley: Department of Industrial Engineering and Operations Research, University of California; 2009.
17. Atamtürk A, Narayanan V. Polymatroids and mean-risk minimization in discrete optimization. *Oper Res Lett* 2008;36(5):618–622.
18. Atamtürk A, Narayanan V. The submodular knapsack polytope. *Discrete Optim* 2009;6(4):333–344.
19. Atamürk A, Narayanan V. Conic mixed-integer rounding cuts. *Math Program Ser A* 2010;122:1–20.
20. Ozsen L, Daskin M, Coullard C. Capacitated facility location model with risk pooling. *Nav Res Log* 2008a;55(4):295–312.
21. Daganzo CF. Logistics systems analysis. 4th ed. Berlin: Springer; 2005.
22. Haimovich M, Rinnooy Kan AHG. Bounds and heuristics for capacitated routing problems. *Math Oper Res* 1985;10(4):527–542.
23. Daganzo CF. The distance traveled to visit N points with a maximum of C stops per vehicle: an analytic model and an application. *Transp Sci* 1984;18(4):331–350.
24. Shen ZJ, Qi L. Incorporating inventory and routing costs in strategic location models. *Eur J Oper Res* 2006;179:372–389.
25. Ozsen L, Daskin M, Coullard C. Facility location modeling and inventory management with multi-sourcing. *Transp Sci* 2009; 43(4):455–472.
26. Mak HY, Shen ZJ. Risk pooling and risk diversification in supply chain design. Working Paper. Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology. 2009b.
27. Mak HY, Shen ZJ. Two-echelon inventory-location problem with service considerations. *Nav Res Log* 2009c;56(8):730–744.
28. Jordan WC, Graves SC. Principles and benefits of manufacturing process flexibility. *Manage Sci* 1995;41:577–594.
29. Graves SC, Tomlin BT. Process flexibility in supply chain. *Manage Sci* 2003;49(7):907–919.
30. Nahmias S. Production and operations management. Chicago: Irwin; 2005.
31. Snyder LV. Facility location under uncertainty: a review. *IIE Trans* 2006;38: 537–554.
32. Chopra S, Reinhardt G, Mohan U. The importance of decoupling recurrent and disruption risks in a supply chain. *Nav Res Log* 2007;54(5):544–555.
33. Snyder LV, Daskin MS, Teo CP. The stochastic location model with risk pooling. *Eur J Oper Res* 2007;179(3):1221–1238.
34. Birge J, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
35. Shen ZJ. A multi-commodity supply chain design problem. *IIE Trans* 2005;37: 753–762.

36. Qi L, Shen ZJ. A supply chain design model with unreliable supply. *Nav Res Log* 2007;54(8):829–844.
37. Yano CA, Lee HL. Lot sizing with random yields: a review. *Oper Res* 1995;43(2):311–334.
38. Dada M, Petrucci NC, Schwarz LB. A newsvendor's procurement problem when suppliers are unreliable. *Manuf Serv Oper Manage* 2007;9(1):9–32.
39. Qi L, Shen ZJ, Snyder LV. A continuous-review inventory model with disruptions at both supplier and retailer. *Prod Oper Manage* 2009;18(5):516–532.
40. Tomlin B. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Manage Sci* 2006;52:39–657.
41. Tomlin B. Disruption-management strategies for short life-cycle products. *Nav Res Log* 2009;56(4):318–347.
42. Snyder LV, Daskin MS. Reliability models for facility location: the expected failure cost case. *Transp Sci* 2005;39(3):400–416.
43. Berman O, Krass D, Menezes MBC. Facility reliability issues in network p-median problems: strategic centralization and co-location effects. *Oper Res* 2007;55:332–350.
44. Zhan RL, Shen ZJ, Daskin MS. System reliability with location-specific failure probabilities. Working paper. Berkeley: University of California; 2007.
45. Shen Z-J, Zhan R, Zhang J. The reliable facility location problem: formulations, heuristics, and approximation algorithms. *INFORMS J Comput* 2010. In press.
46. Cui T, Ouyang Y, Shen ZJ. Reliable facility location design under the risk of disruptions. *Oper Res* 2010. In press. doi:10.1287/opre.1090.0801.
47. Lim M, Daskin MS, Bassamboo A, *et al.* A facility reliability problem: formulation, properties and algorithm. *Nav Res Log* 2009;57(1):58–70.
48. Ferdows K, Lewis MA, Machuca JAD. Rapid-fire fulfillment. *Harv Bus Rev* 2004;82(11):104–110.
49. Liu Q, van Ryzin GJ. Strategic capacity rationing to induce early purchases. *Manage Sci* 2008;54(6):1115–1131.
50. Su X., Zhang F. 2009. On the value of commitment and availability guarantees when selling to strategic consumers. *Manage Sci* 2009;55(5):713–726.
51. Su X. Intertemporal pricing with strategic customer behavior. *Manage Sci* 2007;53(5):726–741.
52. Chen X, Simchi-Levi D. Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: the finite horizon case. *Oper Res* 2004;52(6):887–896.
53. Chopra S. Seven eleven Japan case. Evanston (IL): Kellogg School of Management, Northwestern University; 1995.
54. Chen X, Zhang J. A stochastic programming duality approach to inventory centralization games. *Oper Res* 2009;57(4):840–851.
55. Goemans MX, Skutella M. Cooperative facility location games. *J Algorithm* 2000;50(2):194–214.

INTERIOR POINT METHODS FOR NONLINEAR PROGRAMS

JIMING PENG

Department of Industrial and
Enterprise System Engineering,
University of Illinois at
Urbana–Champaign, Illinois

MAZIAR SALAHI

Department of Mathematics,
Faculty of Sciences, University
of Guilan, Rasht, Iran

INTRODUCTION

In general, interior point methods (IPMs) refer to a special class of optimization algorithms for solving classes of linear and nonlinear programming problems. As indicated by their names, IPMs usually move in the interior of the feasible region of the underlying optimization problem and try to find an optimal solution by following a particular path that converges to the solution set of the corresponding optimization problem.

The study of IPMs can be dated back to the mid-1950s when Frisch [1,2] proposed the logarithmic barrier method for linear programming (LP). Similar ideas were also used by Fiacco and McCormick in Fiacco [3] and Fiacco and McCormick [4] to transfer constrained optimization problems into parameterized unconstrained problems by using penalty or barrier terms. The name of interior point methods first appeared in Fiacco and McCormick [4] where the relationships between minimizers of the barrier and penalty function and solutions of the original constrained problems were discussed in detail. However, despite the good features of barrier methods, Lootsma [5] and Murray [6] observed independently in the late 1960s that the Hessian of a barrier function becomes increasingly ill-conditioned as the solution is approached and is singular in the limit. Discouraged by

this serious drawback of barrier methods, many optimization experts around that time turned their interest to other methods such as augmented Lagrangian and sequential quadratic programming (SQP) methods [7–10] that seemed to be more efficient in practice without having ill-conditioning.

The modern era of IPMs dates back to 1984 when Karmarkar [11] announced his seminal work on a polynomial-time algorithm for linear programming that is efficient both in theory and practice. Since then, many variants of IPMs for linear and nonlinear optimization problems subject to certain constraints, including the primal, the dual, and the primal–dual IPMs, have been proposed and analyzed; for more details we refer to Alizadeh and Goldfarb [12], Nesterov and Nemirovski [13], Terlaky [14], Vandenberghe and Boyd [15], Wolkowicz *et al.* [16], Wright [17], and Ye [18] and the references therein.

A GENERIC FRAMEWORK OF PRIMAL–DUAL IPMS

In this survey, we focus only on the primal–dual framework, which is the most widely used one both in theoretical analysis [19–21] and practical implementation [22–24]. To start with, we consider NLPs in the following form:

$$\begin{aligned} \min_{x \in R^n} f(x) \\ c_i(x) = 0 \quad i \in \mathcal{E}, \\ c_i(x) \geq 0 \quad i \in \mathcal{I}, \end{aligned} \quad (1)$$

where f and $c_i(x)$'s are twice-continuously differentiable functions from R^n to R . By introducing the slack vector s we write Equation (1) as

$$\begin{aligned} \min_{x \in R^n} f(x) \\ c_i(x) = 0 \quad i \in \mathcal{E}, \\ c_i(x) - s_i = 0 \quad i \in \mathcal{I}, \\ s \geq 0. \end{aligned} \quad (2)$$

The Lagrangian function associated with this problem is

$$L(x, s, \lambda) = f(x) + \lambda_{\mathcal{E}}^T c_{\mathcal{E}}(x) - \lambda_{\mathcal{I}}^T (c_{\mathcal{I}}(x) - s), \quad (3)$$

where $\lambda_{\mathcal{E}}, \lambda_{\mathcal{I}}$ are Lagrange multipliers. The Karush-Kuhn-Tucker (KKT) conditions associated with Equation (2) can be written as

$$\begin{aligned} \nabla_x L(x, s, \lambda) &= 0, \\ c_i(x) &= 0 \quad i \in \mathcal{E}, \\ c_i(x) - s_i &= 0 \quad i \in \mathcal{I}, \\ S \Lambda_{\mathcal{I}} e &= 0, \\ s, \lambda_{\mathcal{I}} &\geq 0, \end{aligned} \quad (4)$$

where $S = \text{diag}(s)$ and $\Lambda_{\mathcal{I}} = \text{diag}(\lambda_{\mathcal{I}})$. Let us consider the following barrier problem for Equation (2):

$$\begin{aligned} \min_{x \in R^n} \Phi_{\mu}(x, s) &:= f(x) - \mu \sum_{i \in \mathcal{I}} \log s_i, \\ c_i(x) &= 0 \quad i \in \mathcal{E}, \\ c_i(x) - s_i &= 0 \quad i \in \mathcal{I}, \end{aligned} \quad (5)$$

where $\mu > 0$ is the barrier parameter. The Lagrangian function associated with Equation (5) is

$$\mathcal{L}(x, s, \lambda, \mu) = \Phi_{\mu}(x, s) + \lambda_{\mathcal{E}}^T c_{\mathcal{E}}(x) - \lambda_{\mathcal{I}}^T (c_{\mathcal{I}}(x) - s). \quad (6)$$

For convenience, let $A_{\mathcal{E}}$ and $A_{\mathcal{I}}$ denote the Jacobian matrices of the mappings $c_{\mathcal{E}}(x)$ and $c_{\mathcal{I}}(x)$, respectively. Then, we can write the first-order optimality conditions for Equation (5) as follows:

$$\begin{aligned} \nabla f(x) + A_{\mathcal{E}}^T \lambda_{\mathcal{E}} - A_{\mathcal{I}}^T \lambda_{\mathcal{I}} &= 0, \\ S \Lambda_{\mathcal{I}} e - \mu e &= 0, s \geq 0, \lambda_{\mathcal{I}} \geq 0 \quad (7) \\ c_i(x) &= 0 \quad i \in \mathcal{E}, \\ c_i(x) - s_i &= 0 \quad i \in \mathcal{I}. \end{aligned}$$

At each iteration, a primal-dual IPM obtains a search direction by computing the

linearized Newton direction for the nonlinear system of Equation (7) as follows [25–27]:

$$\begin{pmatrix} W(z, \lambda, \mu) & A(x)^T \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta z \\ \Delta \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_z \mathcal{L}(z, \lambda, \mu) \\ c(z) \end{pmatrix}, \quad (8)$$

where

$$\begin{aligned} z &= \begin{pmatrix} x \\ s \end{pmatrix}, \Delta z = \begin{pmatrix} \Delta x \\ \Delta s \end{pmatrix}, \\ \Delta \lambda &= \begin{pmatrix} \Delta \lambda_{\mathcal{E}} \\ \Delta \lambda_{\mathcal{I}} \end{pmatrix}, c(z) = \begin{pmatrix} c_{\mathcal{E}}(x) \\ c_{\mathcal{I}}(x) - s \end{pmatrix}. \end{aligned} \quad (9)$$

and

$$\begin{aligned} A(x) &= \begin{pmatrix} A_{\mathcal{E}} & 0 \\ A_{\mathcal{I}} & -I \end{pmatrix}, \\ W(z, \lambda, \mu) &= \begin{pmatrix} \nabla_{xx}^2 \mathcal{L}(z, \lambda, \mu) & 0 \\ 0 & -S^{-1} \Lambda_{\mathcal{I}} \end{pmatrix}. \end{aligned} \quad (10)$$

After solving Equation (8), we update the new iterate by

$$z^+ = z + \alpha_z \Delta z, \quad \lambda^+ = \lambda + \alpha_{\lambda} \Delta \lambda, \quad (11)$$

where $\alpha_z, \alpha_{\lambda}$ are step lengths that are chosen to keep s and $\lambda_{\mathcal{I}}$ strictly positive. Then a backtracking line search is used to compute the final step length, which provides sufficient decrease of a merit function or ensures acceptability by a filter. This simple approach is the basis for line search IPMs; however, it should be modified for the nonconvex problem to prevent convergence to nonstationary point. The optimality test is based on a merit function that measures the violation of the optimality conditions

$$m(x, s, \lambda) = \max \left\{ \|\nabla f(x) + A_{\mathcal{E}}^T \lambda_{\mathcal{E}} - A_{\mathcal{I}}^T \lambda_{\mathcal{I}}\|, \|S \Lambda_{\mathcal{I}} e\|, \|c_{\mathcal{E}}(x)\|, \|c_{\mathcal{I}}(x) - s\| \right\}$$

A generic framework for primal-dual IPMs is summarized as follows:

Primal-Dual interior point algorithm.

1. Choose (x_0, s_0, λ_0) such that $s_0 > 0, \lambda_{\mathcal{I}}^0 > 0$ and select an initial barrier parameter $\mu_0 > 0$.

2. For $k = 0, 1, \dots$, do the following:

- If $m(x, s, \lambda) \leq \text{tol}$, then stop (The optimality test).
- Compute the search direction $d = (\Delta z; \Delta \lambda)$ by solving Equation (8);
- Compute the step length $\hat{\alpha}_s$ and $\hat{\alpha}_{\lambda_{\mathcal{I}}}$ by

$$\hat{\alpha}_s = \frac{-1}{\min(S_k^{-1} \Delta s_k, -1)},$$

$$\hat{\alpha}_{\lambda_{\mathcal{I}}} = \frac{-1}{\min(\Lambda_k^{-1} \Delta \lambda_{\lambda_{\mathcal{I}}}, -1)}.$$

- Find $\tau_k \in (0, 1]$ and $\alpha \in (0, 1]$ such that

$$m(x + \alpha \Delta x, s + \alpha_s \Delta s, \lambda_{\mathcal{E}} + \alpha \Delta \lambda_{\mathcal{E}}, \lambda_{\mathcal{I}} + \alpha_{\lambda_{\mathcal{I}}} \Delta \lambda_{\mathcal{I}}) \leq m(x, s, \lambda) + \beta \alpha \nabla m(x, s, \lambda)^T d$$

for some $\beta \in (0, 1)$ where $\alpha_s = \min(1, \tau_k \hat{\alpha}_s)$, $\alpha_{\lambda_{\mathcal{I}}} = \min(1, \tau_k \hat{\alpha}_{\lambda_{\mathcal{I}}})$.

- Update the iterate by

$$x_{k+1} = x_k + \alpha \Delta x_k, \quad s_{k+1} = s_k + \alpha_s \Delta s_k,$$

$$(\lambda_{\mathcal{E}})_{k+1} = (\lambda_{\mathcal{E}})_k + \alpha (\Delta \lambda_{\mathcal{E}})_k,$$

$$(\lambda_{\mathcal{I}})_{k+1} = (\lambda_{\mathcal{I}})_k + \alpha_{\lambda_{\mathcal{I}}} (\Delta \lambda_{\mathcal{I}})_k.$$

Update the barrier parameter μ_{k+1} ¹.

GLOBAL AND LOCAL CONVERGENCE OF THE ALGORITHM

As it is known, one of the main features of IPMs that distinguishes them from many other optimization methods is their polynomial iteration complexity. This feature holds for some NLPs as well. (see Monteiro and Adler, Wright, and Ye [17,18,28] and reference therein). Nesterov and Nemirovski [13], showed that IPMs enjoy polynomial

complexity for linear programming problems over the intersection of an affine set and the so-called self-dual cone, which includes the second-order cone and the cone of semidefinite matrices. These classes of problems allow us to find a global optimal solution of some nonconvex quadratically constrained quadratic and fractional quadratic problems [29,30].

For generic convex and nonconvex programming problems, however, the theoretical results on IPMs are weaker. In both the theoretical analysis and practical implementation of primal–dual IPMs for nonconvex NLPs, a key is how to ensure convergence from a general starting point. One of the most popular choices is to require a sufficient decrease in a merit function that encourages the early iterates to move toward the trajectory. However, it should be noted that the choice of an appropriate merit function for nonconvex problems is nontrivial. For example, $\|F_{\mu}(v)\|$ is the one that has been used within many algorithms, where $F_{\mu}(v)$ is the function on the left-hand side of Equation (7). But it only ensures convergence to points satisfying the first-order conditions. However, since many primal–dual algorithms try to find points satisfying second-order necessary conditions, this merit function is not appropriate for such a purpose. For example, in Conn *et al.* [31] the authors have used the barrier function as the merit function.

For illustration, in the sequel we present a specific variant of primal–dual IPMs that is globally convergent under certain conditions [19]. Let us denote the function on the left-hand side of Equation (4) by

$$F(x, s, \lambda) = \begin{bmatrix} G(x, s, \lambda) \\ S \Lambda_{\mathcal{I}} e \end{bmatrix},$$

and

$$G(x, s, \lambda) = \begin{bmatrix} \nabla_x L(x, s, \lambda) \\ c_{\mathcal{I}} \\ c_{\mathcal{E}} - s \end{bmatrix}.$$

Further, for simplicity, let $v = (x, s, \lambda)$ and $v(\alpha) = v + \alpha \Delta v$, where

$$F'(v) \Delta v = -F(v) + \mu \hat{e}, \quad (12)$$

$\hat{e} = (0, \dots, 0, 1, \dots, 1)$ with p ones and $p = |\mathcal{E}|$. For a given starting point v_0 with $(s_0, \lambda_{\mathcal{I}}^0) > 0$,

¹See the section titled “Implementation Strategies and Software Development” for more details on the barrier parameter updating strategy.

let

$$\tau_1 = \frac{\min(S_0 \Lambda_{\mathcal{I}}^0 e)}{[s_0^T \lambda_{\mathcal{I}}^0 / p]}, \quad \tau_2 = \frac{s_0^T \lambda_{\mathcal{I}}^0}{\|G(v_0)\|}.$$

Define

$$f^I(\alpha) = \min(S(\alpha) \Lambda_{\mathcal{I}}(\alpha)) - \gamma \tau_1 \frac{s(\alpha)^T \lambda_{\mathcal{I}}(\alpha)}{p},$$

$$f^{II}(\alpha) = s(\alpha)^T \lambda_{\mathcal{I}}(\alpha) - \gamma \tau_2 \|G(v(\alpha))\|,$$

where $\gamma \in (0, 1)$ is a constant. It can be easily shown that if α_k are chosen such that $f^I(\alpha) \geq 0$, for all $\alpha \in [0, \alpha_k]$, then at every iteration we have $(s_k, \lambda_{\mathcal{I}}^k) > 0$ and

$$\frac{\min(S_k \Lambda_{\mathcal{I}}^k e)}{[(s_k)^T \lambda_{\mathcal{I}}^k / p]} \geq \gamma_k \tau_1, \quad \gamma_k \in (0, 1).$$

On the basis of the above observation, to choose a suitable size α_k , we require α_k to satisfy the following relations:

$$f^i(\alpha_k) \geq 0, i = I, II, \quad f^I(\alpha_k) \geq 0, \forall \alpha \in [0, \alpha_k].$$

For $i = I, II$, let

$$\alpha_i = \max_{\alpha \in [0, 1]} \{\alpha : f^i(\alpha') \geq 0, \forall \alpha' \leq \alpha\}. \quad (13)$$

In other words, α' is either one or the smallest positive root for the functions $f^i(\alpha)$ in $(0, 1]$. One can show that $\alpha^i > 0$. Since $f^I(a)$ is piecewise quadratic, α^I can be found easily.

The globalized algorithm is a perturbed and damped Newton method combined with a backtracking line search. The merit function used for the line search is the squared l_2 -norm of the residual, that is,

$$\phi(v) = \|F(v)\|^2.$$

Let ϕ_k denote the value of the function $\phi(v_k)$ and $\phi_k(\alpha)$ denote $\phi(v_k + \alpha \Delta v_k)$. Under mild conditions, one can show that the perturbed Newton step,

$$\Delta v_k = -F'_{\mu_k}(v_k)^{-1} F_{\mu}(v_k)$$

is a descent direction for the merit function $\phi(v)$.

The globalized algorithm in *El-Bakry et al.* [19] is described as follows.

A globalized algorithm.

Step 0. Choose $v_0 = (x_0, s_0, \lambda^0)$ such that $(s_0, \lambda_{\mathcal{I}}^0) > 0$, $\rho \in (0, 1)$, and $\beta \in (0, \frac{1}{2}]$. Set $k = 0$, $\gamma_{k-1} = 1$, and compute $\phi_0 = \phi(v_0)$. For $k = 0, 1, \dots$, do the following steps.

Step 1. If $\phi_k \leq \epsilon_{\text{exit}}$, stop.

Step 2. Choose $\sigma_k \in (0, 1)$; for $v = v_k$, compute the perturbed Newton direction Δv_k from Equation (12) with $\mu_k = \sigma_k \frac{s_k^T \lambda_{\mathcal{I}}^k}{p}$.

Step 3. Step length selection:

- Choose $\frac{1}{2} \leq \gamma_k \leq \gamma_{k-1}$: compute $\alpha^i, i = I, II$, from (13) and let

$$\bar{\alpha}_k = \min(\alpha^I, \alpha^{II}).$$

- Let $\alpha_k = \rho^t \bar{\alpha}_k$, where t is the smallest nonnegative integer satisfying

$$\phi_k(\alpha_k) \leq \phi_k(0) + \alpha_k \beta \phi'_k(0).$$

Step 4. Let $v_{k+1} = v_k + \alpha_k \Delta v_k$ and $k = k + 1$ and go to Step 1.

For a given $\epsilon \geq 0$, let us define

$$\Omega(\epsilon) = \left\{ v : \epsilon \leq \phi(v) \leq \phi_0, \min(S \Lambda_{\mathcal{I}} e) / (s^T \lambda_{\mathcal{I}} / p) \geq \frac{\tau_1}{2}, s^T \lambda_{\mathcal{I}} / \|G(v)\| \geq \frac{\tau_2}{2} \right\}.$$

We now state some assumptions that are needed to prove the global convergence of the algorithm.

(C1) In the set $\Omega(0)$, the functions $f(x), c_{\mathcal{I}}(x), c_{\mathcal{E}}(x)$ are twice-continuously differentiable and the derivative of $G(v)$, is Lipschitz continuous with constant L . Moreover, the columns of $\nabla c_{\mathcal{I}}(x)$ are linearly independent.

(C2) The iteration sequence $\{x_k\}$ is bounded.²

²This can be ensured by imposing box constraints $-M \leq x \leq M$ for sufficiently large $M > 0$.

- (C3) The matrix $\nabla^2(x, s, \lambda) + \nabla c_{\mathcal{E}}(x)S^{-1}\Lambda_T\nabla c_{\mathcal{E}}^T(x)$ is invertible for v in any compact subset of $\Omega(0)$, where s is bounded away from zero.
- (C4) Let I_s^0 be the index set $\{i : 1 \leq i \leq p = |\mathcal{E}|, \liminf [s_k]_i = 0\}$. Then, the set of gradients $\{\nabla c_1(x_k), \dots, \nabla c_{|\mathcal{I}|}(x_k), \nabla c_i(x_k), i \in I_s^0\}$ is linearly independent for sufficiently large k .

The following theorem is from El-Bakry *et al.* [19].

Theorem 1. *Let $\{v_k\}$ be generated by the globalized algorithm, with $\epsilon_{\text{exit}} = 0$ and $\{\sigma_k\} \subset (0, 1)$ bounded away from zero and one. Under assumptions (C1)–(C4), $\{F(v_k)\}$ converges to zero and, for any limit point $v^* = (x^*, s^*, \lambda^*)$ of $\{v_k\}$, x^* is a KKT point of problem (1).*

We mention that if (C1)–(C4) are violated, then this particular algorithm may not converge. For example in Wachter and Biegler [32], a class of one-dimensional examples are identified for which certain line-search barrier-SQP methods fail to find a feasible point and do not converge to a meaningful point. This is caused by the failure of regularity condition (C4). There are many variants of IPMs that use regularization and relatively weaker assumptions other than (C1)–(C4) to ensure global convergence. For more details, we refer to the articles by Ulbrich *et al.* [20], and Yamashita [21], Akrotirianakis and Rustem [33], Griva *et al.* [34] and the references therein.

Local convergence results for primal–dual IPMs have been given by Wright [35], Conn *et al.* [31], El-Bakry *et al.* [19], Yamashita *et al.* [21,36], and Benson *et al.* [25]. The local convergence results of these methods are inherited from the use of Newton’s method to generate search directions. As with SQP methods, the drawbacks of Newton’s method are also inherited. Global convergence results have been discussed by El-Bakry *et al.* [19], Yamashita [21], and Akrotirianakis and Rustem [33]. Here, we present only the local convergence of the

algorithm from El-Bakry *et al.* [19]. Let $\mathcal{A}(x^*)$ be the set of active constraints. The following assumptions are crucial:

1. There is a solution (x^*, s^*, λ^*) satisfying the KKT conditions for Equation (1).
2. The Hessian matrices $\nabla^2 f(x), \nabla^2 c_{\mathcal{E}}(x), \nabla^2 c_{\mathcal{I}}(x)$ exist and are locally Lipschitz continuous at x^* .
3. The set $\{\nabla c_1(x^*), \dots, \nabla c_{|\mathcal{E}|}(x^*)\} \cup \{\nabla c_i(x^*), i \in \mathcal{A}(x^*)\}$ is linearly independent.
4. For all $\eta \neq 0$, $\nabla c_i(x^*)^T \eta = 0$, $i \in \mathcal{E} \cup \mathcal{A}(x^*)$ we have $\eta^T \nabla_{xx}^2 \mathcal{L}(x^*) \eta > 0$.
5. For all $i \in \mathcal{I}$, $s_i + \lambda_i > 0$.

Theorem 2. [19] *Suppose for Equation (1) at $v^* = (z^*, \lambda^*, \mu^*)$ assumptions 1 to 5 hold. Given $\hat{\tau} \in (0, 1)$, there exists a neighborhood D of v^* and a constant $\hat{\sigma}$ such that for any v^0 in D and $\tau_k \in [\hat{\tau}, 1)$ and $\sigma_k \in (0, \hat{\sigma})$, primal–dual interior point algorithm is well defined and iteration sequence generated by it converges Q -linearly to v^* .*

More results on local convergence analysis can be found in the articles by Akrotirianakis and Rustem [33] Griva *et al.* [34] Ulbrich *et al.* [20].

IMPLEMENTATION STRATEGIES AND SOFTWARE DEVELOPMENT

As shown in the previous section, we need to update the iterate carefully to ensure the convergence of the algorithm for NLPs. In this section, we review some other updating strategies that are crucial for the success of primal–dual IPMs in both theory and practice.

We start with the procedure for updating the barrier parameter μ . Two types of update strategies have been considered in the literature: adaptive and monotone. Adaptive strategies [19,20,37] allow changes in the barrier parameter at every iteration, and are often efficient in practice, but they generally do not guarantee the global convergence of the algorithm. For example, the algorithms in El-Bakry *et al.* [19] and Ulbrich *et al.*

[20] only converge to stationary points. This is because these algorithms enforce only reduction of a measure of stationarity but do not necessarily reduce the objective function value. The most important monotone strategy is the so-called Fiacco–McCormick approach that fixes the barrier parameter until an approximate solution of the barrier problem is found. It has been employed in various nonlinear interior algorithms [21,24,38] and has been implemented, for example, in the IPOPT and KNITRO software packages. The Fiacco–McCormick strategy provides a framework for establishing global convergence [39,40], but sometimes suffers from several limitations. For example, it could be very sensitive to the choice of the initial point, the initial value of the barrier parameter, and the scaling of the problem, and it is often unable to recover quickly when the iterates approach the boundary of the feasible region prematurely. The numerical experiments of Nocedal *et al.* [41] with IPOPT and KNITRO suggests that more dynamic update strategies are needed to improve the efficiency of nonlinear IPMs. The two widely used adaptive approaches are Mehrotra’s technique and the one used in LOQO [23,24]. Computational experiments show that the adaptive strategies outperform the monotone approach but they are not very robust. To resolve this problem, Nocedal *et al.*, [41] introduced an efficient adaptive strategy using a linear quality function within Mehrotra’s predictor corrector algorithm with a globalization framework. The role of the quality function is to determine the best value of σ in $\mu = \sigma \frac{s^T w}{n}$ by penalizing large errors in the KKT conditions.

Another key step in primal–dual IPMs is that of solving the Newton system for a search direction. An effective implementation must exploit the sparsity or other structures in the matrix. The Newton system in primal–dual IPMs is well defined when the matrix $W(z, \lambda, \mu)$ is nonsingular. Several conditions such as the second-order optimality, strict complementarity, and regularity conditions can guarantee the nonsingularity of $W(z, \lambda, \mu)$ in a small neighborhood of the optimal solution. When the iterate is far from the solution, we may regularize the Hessian

$\nabla_{xx}^2 \mathcal{L}(z, \lambda, \mu)$, if necessary, to get a nonsingular $W(z, \lambda, \mu)$ [42,43]. An alternative approach is to use a safe-guarded trust region step to stabilize the iteration, which can be found in Waltz *et al* [44].

As it is known, merit functions are usually used to measure the progress of the algorithm. Another approach to ensure the convergence of primal–dual IPMs for NLPs is the so-called filter method. Differing from the methods using merit functions, the filter methods separately deal with the two primary goals: reduce the objective function value and the infeasibility. Such a strategy makes the filter methods insensitive to the choice of merit functions and allows more freedom in selecting the next trial point. More details on filter methods and their numerical performance can be found in Vanderbei [23], Wachter and Biegler [24], Benson *et al.* [45], Fletcher *et al.* [46].

The above discussed strategies have been implemented in several software packages and successfully applied to solve large-scale nonlinear and nonconvex programming problems. These include IPOPT [24], KNITRO [22], and LOQO [23]. One may find information regarding these packages on NEOS server (<http://www-neos.mcs.anl.gov/neos/>), which provides online solution for optimization problems.

CONCLUSIONS

The research on IPMs has reshaped many areas in mathematical programming and still continues to improve the optimization theory and methods. In this article we have given a brief overview of primal–dual IPMs for NLPs. We have presented some global and local convergence results. Key implementation strategies have been discussed. More details on the theory, algorithms, convergence results, practical performance of algorithms can be found in the list of references and related papers.

REFERENCES

1. Frisch R. The logarithmic potential method for solving linear programming problems.

- Memorandum, Oslo: University Institute of Economics; 1955.
2. Frisch R. The logarithmic potential method for convex programming problems. Memorandum, Oslo: University Institute of Economics; 1955.
3. Fiacco AV. Barrier methods for nonlinear programming. In: Holzman A, editor. Operations research support methodology. New York: Marcel Dekker; 1979. pp. 377–440.
4. Fiacco AV, McCormick GP. Nonlinear programming, Classics in Applied Mathematics 4. Philadelphia (PA): SIAM; 1990. Reprint of the 1968 original.
5. Lootsma FA. Hessian matrices of penalty functions for solving constrained-optimization problems. Philips Res Rep 1969;24:322–330.
6. Murray W. Analytical expressions for the eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions. J Optim Theory Appl 1971;7:189–196.
7. Bazaraa MS, Sherali HD, Shetty CM. Nonlinear programming: theory and algorithms. 2nd ed. New York: John Wiley & Sons, Inc.; 1993.
8. Fletcher R. Practical methods of optimization. 2nd ed. Chichester: John Wiley & Sons, Ltd.; 1987.
9. Gill PE, Murray W, Wright MH. Practical optimization. London, New York: Academic Press; 1981.
10. Nocedal J, Wright SJ. Numerical optimization. New York: Springer; 1999.
11. Karmarkar N. A new polynomial-time algorithm for linear programming. Combinatorica 1984;4:373–395.
12. Alizadeh F, Goldfarb D. Second-order cone programming. Math Program 2003;95:3–51.
13. Nesterov Y, Nemirovskii A. Interior point polynomial algorithms in convex programming. Philadelphia: SIAM; 1994.
14. Terlaky T, editor. Interior point methods of mathematical programming. Boston (MA): Kluwer Academic Publishers; 1996.
15. Vandenberghe L, Boyd S. Semidefinite programming. SIAM Rev 1996;38:49–95.
16. Wolkowicz H, Saigal R, Vandenberghe L, editors. Handbook on semidefinite programming. Boston: Kluwer; 2000.
17. Wright SJ. Primal-dual interior-point methods. Philadelphia (PA): SIAM; 1997.
18. Ye Y. Interior point algorithms, theory and analysis. Chichester: John Wiley & Sons, Ltd.; 1997.
19. El-Bakry AS, Tapia RA, Tsuchiya T, et al. On the formulation and theory of the Newton interior-point method for nonlinear programming. J Optim Theory Appl 1996;89:507–541.
20. Ulbrich M, Ulbrich S, Vicente LN. A globally convergent primal-dual interior-point filter method for nonlinear programming. Math Program 2004;100(2):379–410.
21. Yamashita H. A globally convergent primal-dual interior point method for constrained optimization. Optim Methods Softw 1998;10:443–469.
22. Byrd RH, Nocedal J, Waltz RA. KNITRO: an integrated package for nonlinear optimization. In: di Pillo G, Roma M, editors. Large-scale nonlinear optimization. New York: Springer Verlag; 2006. pp. 35–59.
23. Vanderbei RJ. LOQO: an interior point code for quadratic programming. Optim Methods Softw 1999;12:451–484.
24. Wachter A, Biegler LT. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math Program Ser A and B 2006;106(1):25–57.
25. Benson H, Sen A, Shanno DF. Interior-point method for nonconvex nonlinear programming: convergence analysis and computational experience. working paper. Drexel University; 2009.
26. Byrd RH, Hribar ME, Nocedal J. An interior point algorithm for large-scale nonlinear programming. SIAM J Optim 1999;9:877–900.
27. Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. SIAM Rev 2002;44(4):525–597.
28. Monteiro RDC, Adler I. Interior path following primal-dual algorithms. Part II: Convex quadratic programming. Math Program 1998;44:43–66.
29. Beck A, Ben-Tal A. On the solution of the Tikhonov regularization of the total least squares problem. SIAM J Optim 2006;17(1):98–118.
30. Ye Y, Zhang S. New results on quadratic minimization. SIAM J Optim 2003;14(1):245–267.
31. Conn AR, Gould NIM, Orban D, et al. A primal-dual trust-region algorithm for non-convex nonlinear programming. Math Program 2000;87:215–249.
32. Wachter A, Biegler LT. Failure of global convergence for a class of interior point methods for nonlinear programming. Math Program 2000;88:565–587.

33. Akrotirianakis I, Rustem B. A globally convergent interior point algorithm for nonlinear problems. *J Optim Theory Appl* 2005;125(3):497–521.
34. Griva I, Shanno DF, Vanderbei RJ, *et al.* Global convergence of a primal-dual interior-point method for nonlinear programming. *Algorithmic Oper Res* 2008;3(1):12–19.
35. Wright MH. Interior methods for constrained optimization. *Acta Numer* 1992;1:341–407.
36. Yamashita H, Yabe H. Superlinear and quadratic convergence of some primal-dual interior point methods for constrained optimization. *Math Program* 1996;75:377–397.
37. Vanderbei RJ, Shanno DF. An interior-point algorithm for nonconvex nonlinear programming. *Comput Optim Appl* 1999;13:231–252.
38. Forsgren A, Gill PE. Primal-dual interior methods for nonconvex nonlinear programming. *SIAM J Optim* 1998;8(4):1132–1152.
39. Byrd RH, Gilbert JC, Nocedal J. A trust region method based on interior point techniques for nonlinear programming. *Math Program* 2000;89:149–185.
40. Wachter A, Biegler LT. Line search filter methods for nonlinear programming: motivation and global convergence. *SIAM J Optim* 2005;16(1):1–31.
41. Nocedal J, Wachter A, Waltz RA. Adaptive barrier update strategies for nonlinear interior methods. Report RC 23563. Yorktown (NY): IBM T. J. Watson Research Center; 2005.
42. Forsgren A, Gill PE, Shinnerl JR. Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization. *SIAM J Matrix Anal Appl* 1996;17:187–211.
43. Wright SJ. Stability of linear equation solvers in interior-point methods. *SIAM J Matrix Anal Appl* 1995;16:1287–1307.
44. Waltz RA, Morales JL, Nocedal J, *et al.* KNITRO-direct: a hybrid interior algorithm for nonlinear optimization. Technical report 2003-6. Evanston (IL): Optimization Technology Center, Northwestern University; 2003.
45. Benson HY, Shanno DF, Vanderbei RJ. Interior-point methods for nonconvex nonlinear programming: filter methods and merit functions. Report ORFE-00-06. Princeton (NJ): Department of Operations Research and Financial Engineering, Princeton University; 2000.
46. Fletcher R, Leyffer S, Toint P. A brief history of filter methods. *SIAG/Optimization Views-and-News* 2007;18(1):2–12.

INTERIOR-POINT LINEAR PROGRAMMING SOLVERS

HANDE Y. BENSON
Drexel University,
Philadelphia,
Pennsylvania

We present an overview of available software for solving linear programming (LP) problems using interior-point methods. We consider both open-source and proprietary commercial codes, including BPMPD [1,2], COIN Linear Program code (CLP) [3], FortMP [4], GIPALS32 [5], GNU Linear Programming Kit (GLPK) [6], HOPDM [7,8], IBM ILOG CPLEX [9], LINDO [10], LIPSOL [11], LOQO [12], Microsoft Solver Foundation [13], MOSEK [14], PCx [15], SAS/OR [16], and Xpress Optimizer [17].

Some of the codes discussed include primal and dual simplex solvers as well, but we focus the discussion on the implementation of the interior-point solver. For each solver, we present types of problems solved, available distribution modes, input formats and modeling languages, as well as algorithmic details, including problem formulation, use of higher corrections, presolve techniques, ordering heuristics for symbolic Cholesky factorization, and the specifics of numerical factorization. Many of the solvers allow user-control over some algorithmic details such as the type of presolve techniques applied to the problem and the choice of ordering heuristic. As the following discussion will show, the solvers tend to all have similar characteristics, and performance differences are generally due to subtle changes in the presolve phase, the implementation of the linear algebra routines, and other special considerations, such as scaling, to promote numerical stability.

We start with a short discussion of input formats and modeling languages available for LPs. Then, we will present details of a wide-range of interior-point codes.

INPUT FORMATS AND MODELING LANGUAGES FOR LINEAR PROGRAMMING

For a standard LP of the form

$$\begin{array}{ll} \text{maximize} & c^T x \\ \text{subject to} & Ax = b \\ & x \geq 0, \end{array} \quad (1)$$

it suffices to pass the matrix A and the vectors b and c to the solver for a full-description of the problem. Of course, the general form of an LP can incorporate inequality constraints as well as bounds on variables, but such deviations from Equation (1) require either the user to include slack variables to convert the problem to Equation (1) or the modeling mechanism to accommodate the specification of additional data vectors, such as upper and lower bounds on the variables and indicator arrays for constraint types.

We will discuss two types of modeling mechanisms here: input formats, which allow for a specifically formatted means to enter problem data which can then be interpreted by the solver, and modeling languages, which allow for a flexible algebraic expression of the model as well as the data and are accompanied by full-fledged modeling systems. Many of the solvers we will discuss can also be used as callable libraries, so the model can be expressed in a variety of programming languages, such as C, Fortran, and Java, and various routines implemented in the solver library can be called to load and solve the LP.

MPS Format

Mathematical Programming System (MPS), is the oldest input format for LPs and is the most widely accepted input format among both academic and commercial solvers. We refer the reader to Ref. 18 for a detailed description and provide a brief overview here. MPS uses ASCII text files to store the problem and is modeled after punch cards. Header cards specify which component of Equation (1) is to be introduced next. The ROWS card

specifies the names and types of objective functions and constraints, the COLUMNS card specifies variable names and the entries of c and A , the RHS card specifies the elements of the right-hand side vector b , and the BOUNDS card specifies any upper and lower bounds on x . The format is column-oriented: for example, rather than using the ROWS card to enter each of the variable coefficients appearing in a constraint, the coefficients of a single variable as it appears in the objective function and any constraints are listed together in the COLUMNS card. All model components are assigned names of eight characters or less, and the information is presented in fields that begin in columns 2, 5, 15, 25, 40, and 50. Each data element has to be individually listed in the file, and at most two elements can be specified per line.

Because of its punch card format, MPS files may be hard for a user to read, but it is very easy for a solver or a modeling environment to read and write in this format. With additional card types, the format has been extended to quadratic problems (QPs) and mixed-integer programming (MIPs), and free format variations exist, as well.

LP Format

The LP input format is an alternative to the MPS format that allows for entering algebraic expressions to describe the problem. While it may be more natural to read a model in LP format, it is nonetheless not a standardized format and different solvers/parsers may interpret the problem components differently. Two common variants are CPLEX LP format and Xpress LP format.

As in the MPS format, the model is expressed in an ASCII text file, and there are specified sections to express each type of model component. Each section starts with a header, and there are Objectives, Constraints, Bounds, and Variable Type sections available for LPs. Within the Constraints section, for example, each constraint is given a name, followed by a colon, and then the constraint expression, and each constraint is written on a separate line. Unlike the MPS format, there are no fixed field widths for separating problem components. For further

details of the LP format, we refer the reader to Ref. 19.

AMPL

AMPL [20] is a modeling language for algebraically expressing a wide-range of optimization problems. Originally developed at Bell Laboratories, it has become the standard for formulating highly complex problems. AMPL is more than just a compiler for the optimization model, as it also includes pre and postprocessing routines, as well as first- and second-order differentiation capabilities for nonlinear programming NLPs. The indexing and looping constructs in AMPL allow for a concise expression of large-scale LPs, and the processor allows for the separation of the model and data into multiple files. Many LP solvers have AMPL interfaces, and a tool for converting MPS format to AMPL, *m2a*, is also available.

GAMS

GAMS [21], or the General Algebraic Modeling System, is another widely used modeling system for optimization. Originally funded by The World Bank, it is the oldest modeling language for mathematical programming, and it is currently developed and distributed by the GAMS Development Corporation. It is similar to AMPL in terms of capabilities, and the GAMS-AMPL service available through the NEOS Server [22] allows for the solution of models formulated in GAMS by AMPL solvers.

Other Input Formats and Modeling Languages

Almost every LP solver to be discussed accepts models expressed in one or more of the four input formats and modeling languages mentioned above. There are several other means of communicating LPs to the solvers and we will briefly mention them here:

- LP-Plain (LPP), is a variant of the LP input format;
- Optimization Problem Format (OPF), is a variant of the CPLEX LP input format;

- Mathematical Programming Language (MPL), is an algebraic language and modeling system;
- Optimization Modeling Language (OML), is the input format used by Microsoft Solver Foundation.

INTERIOR-POINT LP SOLVERS

In this section, we will provide an overview of some of the most widely used LP solvers. The solvers are presented in alphabetical order.

BPMPD

BPMPD [1,2] is a solver developed by Cs. Mészáros for solving LPs and convex QPs. It is written in C and is available both as an executable and as a callable library. It accepts input in LP and MPS formats for LP and in QPs format for convex QP. The solver can also be called from AMPL and is available for use on the NEOS Server [22]. There is a MEX interface [23] to use the callable library from within Matlab.

BPMPD implements an infeasible interior-point method as described in Refs 24 and 25, and uses Mehrotra's [26] predictor-corrector framework and barrier parameter updates, along with higher order corrections [27]. The strength of the solver is in the extensive linear algebra routines, particularly for presolve and Cholesky factorization. During presolve, BPMPD removes empty rows and columns, singleton rows, and redundant variable bounds and constraints, eliminates free or implied free variables, substitutes duplicated variables, and sparsifies the constraint matrix. A heuristic is used to decide between the normal equation and augmented system approaches for solving for the step directions, and minimum degree and minimum local-fill options are available for symbolic ordering. For the numerical factorization, a left-looking supernodal factorization algorithm is implemented, and a supernodal backsolve is used. During both factorization and backsolve, loop unrolling is used.

CLP

CLP [3] is an open-source solver distributed by the COIN-OR project [28]. It is implemented in C++ and is available as an executable and a callable library. It can also be used over the NEOS server [22] to solve LPs written in MPS format. While mainly a simplex solver, CLP also implements a barrier method for handling LPs and convex QPs. We focus here on the barrier code.

CLP implements Mehrotra's [26] predictor-corrector algorithm and uses higher order corrections [27]. The distribution contains native code for sparse Cholesky factorization, but it is recommended that a third party factorization code be used. There are interfaces for WSSMP [29], which provide parallel enabled ordering and factorization, TAUCS [30], which uses third party ordering and implements parallel enabled factorization, and AMD [31], which provides an approximate minimum degree ordering and can be used with CLP's native code for factorization.

FortMP

FortMP [4] is a commercial solver distributed by Optirisk systems. It can handle LPs and convex QPs, as well as mixed-integer linear and QPs. The code is written in Fortran 77 and accepts models written in AMPL as well as input in MPS format. It is available for use on the NEOS Server [22]. FortMP is primarily a simplex solver and an interior-point method implementation, referred to as *FortMP IPM*, is included for large-scale LPs and convex QPs.

FortMP IPM is an infeasible primal-dual interior-point method. There are three alternatives implemented for step direction calculations: an affine scaling algorithm, a barrier algorithm, and a predictor-corrector algorithm. According to the documentation, the latter is recommended for most LPs, whereas the affine scaling algorithm is to be preferred for extremely sparse problems, and the barrier algorithm is backup in case of failure. For barrier parameter updates, the user can choose between a multiple of the current duality gap and the updating scheme of [26]. The code includes native routines for

Cholesky factorization and allows for the use of supernodes.

GIPALS32

GIPALS32 [5] is a commercial solver distributed by Optimalon software. It is a Windows-based callable library and includes an application programming interface (API). GIPALS32 is the solver library for the LP environment GIPALS, which uses a graphical user interface that includes tables and forms for easily entering problem data and can read and write MPS files.

GIPALS32 implements Mehrotra's [26] predictor-corrector algorithm and uses higher order corrections [27].

GLPK

GLPK [6] is a toolkit for LP and MIP, distributed by the Free Software Foundation, Inc. as part of the GNU Project under the GNU General Public License. It is written in ANSI C and is available as a callable library. The solver routine within the toolkit is *glpsol*. GLPK includes both simplex and interior-point codes, and accepts input in the GNU MathProg modeling language, a subset of the AMPL language, and in GAMS as CoinGLPK. It is available for use on the NEOS Server [22].

GIPALS32 implements an infeasible primal-dual interior-point method with Mehrotra's [26] predictor-corrector algorithm. For symbolic Cholesky factorization, it has multiple ordering options, including quotient minimum degree and approximate minimum degree, and includes interfaces for third party ordering codes AMD [31] and COLAMD [32]. One should note, however, that the simplex methods implemented within GLPK are more mature than the interior-point method, which does not include some important features such as dense column processing and scaling to provide numerical stability.

HOPDM

HOPDM [7,8] is a package for solving large-scale linear, convex quadratic, and convex non-LP problems, written by Gondzio and

distributed by the University of Edinburgh. It is written in C and is available as an executable and a callable library. It accepts models written in MPS format and can be called from AMPL.

HOPDM implements an infeasible primal-dual interior-point method. It uses multiple centrality correctors [27], and the number of correctors is chosen automatically to reduce the overall solution time. HOPDM includes an extensive presolve routine [33,34], and chooses between the normal equations and the augmented system approach in order to improve the efficiency of the factorization method [35]. Symbolic Cholesky factorization approaches include multiple minimum degree ordering [36] and the quotient tree minimum degree ordering [37]. New versions also include iterative solvers for the Newton system. The algorithm is also capable of warm-starting [38] after a data perturbation in the objective, right-hand side, or the constraint matrix.

IBM ILOG CPLEX

CPLEX [9] is a package for solving LPs, convex QPs, problems with convex quadratic constraints, integer, and MIPs. It is currently distributed as an executable and a callable library by ILOG, an IBM company. It accepts models written in MPS and CPLEX LP input formats and can be called from ILOG OPL Development Studio, AMPL, GAMS, MPL, Matlab, Microsoft Excel, and Python. A parallel implementation is also available.

CPLEX implements three algorithms: a primal simplex method, a dual simplex method, and a primal-dual logarithmic barrier method. We will focus on the barrier method here, which is a primal-dual interior-point method using the predictor-corrector framework [26]. It uses multiple centrality correctors [27], and the maximum number of correctors can be chosen by the user or automatically determined by the algorithm. For symbolic Cholesky factorization, there are three options: approximate minimum degree, approximate minimum local fill, and nested dissection. CPLEX can automatically choose the best ordering heuristic to use, or this can be user-specified. The numerical factorization can be performed out-of-core

to improve memory usage and efficiency, and special handling of dense columns is available.

LINDO Systems

LINDO Systems, Inc. distributes three software packages [10] for mathematical programming: LINDO API, the callable library of solvers; LINGO, the modeling system for building and solving problems; and What's Best, the Microsoft Excel add-in. These packages can model and solve a variety of different types of optimization problems, including LPs, QPs, and NLPs. For LPs, the LINDO API can handle problems written in the MPS input format and in LINDO, and What's Best accepts spreadsheet models developed using Microsoft Excel. Besides the interior-point method, the packages also contain implementations of the primal and dual simplex methods. A parallel implementation is available.

LIPSOL

LIPSOL [11], or LP Interior-Point SOLvers, is a Matlab-based package for solving LPs, written and distributed by Zhang. It takes advantage of the sparse matrix operations available in Matlab and also uses Fortran and C routines linked via MEX. Besides problem data stored in a MAT file, LIPSOL can also accept LPs written in MPS and LPP input formats.

LIPSOL implements a primal-dual infeasible interior-point method and uses Mehrotra's predictor–corrector framework [26]. The barrier parameter updates are a modified version of that in Ref. 26. Instead of using Matlab's built-in sparse Cholesky factorization code, which would need to be modified for numerical stability, LIPSOL uses two Fortran codes: the multiple minimum degree ordering package [39] and the sparse Cholesky factorization package [40]. For efficiency, dense columns are split, so a preconditioned conjugate gradient approach is used to ensure numerical stability.

LOQO

LOQO [12] is a solver for LPs, QPs, and NLPs, written by Vanderbei and distributed

by Princeton University. It is written in C and is available as an executable or a callable library. It accepts LPs written in the MPS input format and can be called from AMPL and Matlab, via a MEX interface. It is available for use over the NEOS Server [22].

Despite its current focus as an NLP code, LOQO was originally developed as an infeasible interior-point method for LPs and convex QPs. It implements an infeasible interior-point method and uses Mehrotra's predictor–corrector framework [26]. Free variables and equality constraints are split and allow for the use of primal or dual orderings in the symbolic Cholesky factorization. Furthermore, priorities, or tiers, are assigned to the columns of the matrix [41] by constraint and variable types, and within each tier, minimum local fill and multiple minimum degree orderings [42] are available. For the numerical factorization, supernodes are used with loop unrolling to improve efficiency.

Microsoft Solver Foundation

Microsoft Solver Foundation [13] is an integrated modeling and solver environment for LPs, convex QPs, second-order cone programming problems, and integer and MIPs, as well as constraint programming and unconstrained non-LP problems. The models can be expressed in OML format. For LP, the environment contains both a simplex and an interior-point solver, and we focus here on the latter.

Microsoft Solver Foundation implements a primal-dual interior-point method that solves a homogeneous self-dual model. Each equality constraint is split into a range constraint with equal bounds, and free variables are handled as inequality constraints, that is, a free variable x_j is replaced by the requirement that $x_j + s_j \geq 0$ with $s_j \geq 0$. Any numerical issues arising from such a treatment of equality constraints and free variables are further alleviated by the regularization scheme of [43]. The normal equations approach is used when solving the Newton system. The default ordering method for Cholesky factorization is the approximate minimum degree ordering, but an option to switch to the approximate minimum fill is

also available. For the numerical factorization, a left-looking supernodal factorization algorithm is implemented, and it is multi-core-enabled.

MOSEK

MOSEK [14] is a package for solving LPs, conic QPs, general convex NLPs, and mixed-integer problems, distributed by MOSEK ApS. It accepts LPs written in the MPS, LP, and OPF input formats and can be called from AMPL and Matlab. APIs for use with programs written in C/C++, C#, .NET, Java, Delphi, and Python are provided. It is also available for use over the NEOS Server [22]. Additionally, there is a parallel implementation of MOSEK [44].

MOSEK implements the homogeneous self-dual algorithm described in Ref. 45. Prior to solving the LP, a presolve phase is used to check for and remove redundancies and linear dependence, eliminate fixed variables, substitute out free variables, and reduce problem size. A dualizer routine decides whether to solve the primal or the dual LP using heuristics. Scaling is performed by multiplying the necessary constraints and variables by constants to improve numerical stability. For the step direction calculations, Mehrotra's predictor-corrector framework [26] is used along with higher-order corrections [27]. The solution of the Newton system is done via the normal equations approach. The default ordering method for Cholesky factorization is the approximate minimum degree heuristic [31], but implementations of multiple minimum degree and minimum local-fill are also provided. For the numerical factorization, supernodes are split into blocks to improve cache usage, and up to 16-way loop unrolling is available to further increase efficiency. A push Cholesky framework, where an iteration consists of a particular column performing all its updates on all other remaining columns to the right, is used.

PCx

PCx [15] is an LP solver distributed by the Optimization Technology Center at Argonne

National Laboratory and Northwestern University. The code is written in C and incorporates a modified version of Ng and Peyton's sparse Cholesky factorization package, written in Fortran 77. It accepts LPs written in the MPS input format and can be called from AMPL and Matlab. Additionally, a graphical user interface (PCxGUI) written in Java is available. PCx is available for use over the NEOS Server [22]. A parallel implementation, pPCx, [46] also exists.

PCx implements Mehrotra's predictor-corrector algorithm [26] and also uses higher-order corrections [27]. The scaling technique of Curtis and Reid [47], which minimizes the deviations of the nonzero elements of the coefficient matrix A from 1.0, is applied prior to the solution to improve the numerical stability of the code. A presolver removes redundancies, empty rows and columns, singleton rows, and fixed variables, eliminates singleton columns, and substitutes duplicated variables. For solving the Newton system, PCx defaults to the normal equations approach, and uses Ng and Peyton's sparse Cholesky factorization package [40], modified to handle small pivot elements. The ordering scheme is multiple minimum degree as described in Refs 36 and 48. For the numerical factorization, supernodes are split into blocks to make better use of hierarchical memory, and loop unrolling further increases efficiency.

SAS/OR

SAS/OR software [16] is a commercial package which includes codes for mathematical programming, project and resource scheduling, and discrete event simulation. It is distributed by SAS Institute, Inc. Models are expressed in the OptModel language, also distributed by SAS, and the input format MPS can also be used for LPs. The optimization routines include solvers for LPs, QPs, NLPs, and mixed-integer problems. There are three options for solving LPs: primal simplex, dual simplex, and interior-point methods. We will focus on the interior-point method here.

SAS/OR incorporates an infeasible primal-dual interior-point method, as described in Ref. 24 and uses Mehrotra's [26]

predictor–corrector framework. A presolve routine is used to check for and remove redundancies, handle singleton rows and columns, eliminate fixed variables, and reduce problem size. The minimum degree heuristic is used for symbolic Cholesky factorization. For numerical factorization, the left-looking, or pull, Cholesky [49], where an iteration consists of a particular column receiving all its updates from all applicable columns to the left, is used.

Xpress Optimizer

Xpress Optimizer [17] is a commercial solver for LP and QPs, distributed as a part of the Xpress-MP suite by FICO/Dash Optimization. It can be used via Console Xpress, which is the graphical user interface of the suite, or as a callable library with C, C++, Java, Fortran, VB6, and .NET programming interfaces. It accepts models written in MPS and LP input formats and can also be called from AMPL or GAMS. Another tool from the Xpress-MP suite, Xpress-Mosel, can be used to take advantage of the object-oriented construction for building advanced models. Xpress Optimizer, as well as the other tools of Xpress-MP, are available for use over the NEOS Server [22]. There is also a parallel implementation. There are three solvers implemented in the Xpress Optimizer: primal simplex, dual simplex, and a barrier interior-point method. We will focus on the interior-point method here.

Xpress Optimizer implements the homogeneous self-dual algorithm described in Ref. 45 and uses Mehrotra’s predictor–corrector framework [26] with Gondzio’s higher-order corrections [27]. A presolve stage determines redundant constraints and may tighten variable bounds as necessary. By default, the code determines whether to solve the primal or the dual LP using the barrier method. For sparse Cholesky factorization, the user can choose among three options for finding an ordering: minimum degree, minimum local fill, or nested dissection. For the numerical factorization, they are given a choice between pull Cholesky, where an iteration consists of a particular column receiving all its updates from all other applicable columns to the left,

and push Cholesky, where an iteration consists of a particular column performing all its updates on all other applicable columns to the right. The default is to use the pull Cholesky.

CONCLUSION

We have presented an overview of available interior-point codes for LP. This list is not exhaustive, as there are numerous codes for conic optimization, such as SDPT3 and SeDuMi, and for nonlinear optimization, such as IPOPT and KNITRO, for which LP problems constitute a special case. We have chosen to focus here on solvers that are either primarily for LP (and by natural extension, QP) or have special features for handling LPs.

Many of the solvers share common features. For example, the use of the predictor–corrector framework and even higher order corrections is quite common, and most solvers implement a minimum degree ordering or a close variation for symbolic Cholesky factorization. The presolve phase also plays an important role, and many of the solvers balance the loss of computing time during presolve with possible gains in efficiency and numerical stability.

REFERENCES

1. Mészáros Cs. Fast cholesky factorization for interior point methods of linear programming. *Comput Math Appl* 1996;31(4/5):49–51.
2. Mészáros Cs. The efficient implementation of interior point methods for linear programming and their applications [PhD thesis]. Eötvös Loránd University of Sciences; 1996.
3. Forrest J, de la Nuez D, Lougee-Heimer R. CLP user’s guide. Technical report. Yorktown Heights, (NY): IBM Corporation; 2004.
4. Ellison EFD, Hajian M, Jones H, *et al.* FortMP manual. Technical report. OptiRisk Systems; April 2008.
5. Optimalon Software. GIPALS32. Available at http://www.optimalon.com/product_gipals32.htm. Accessed 2010.
6. Makhorin A. GLPK (GNU linear programming kit). Available at <http://www.gnu.org/software/glpk/glpk.html>. Accessed 2008.

7. Gondzio J. HOPDM (version 2.12)-a fast LP solver based on a primal-dual interior point method. *Eur J Oper Res* 1995;85(1):221–225.
8. Gondzio J, Makowski M. HOPDM - modular solver for LP problems, user's guide to version 2.12. Working Paper WP-95-50. Laxenburg: International Institute for Applied Systems Analysis; 1995.
9. IBM/LOG. CPLEX 11.0 reference manual. Available at <http://www.ilog.com/products/cplex/>. Accessed 2010.
10. Inc. LINDO Systems. LINDO API 6.0, user manual. Available at <http://www.lindo.com>. Accessed 2010.
11. Zhang Y. Solving large-scale linear programs by interior-point methods under the Matlab environment. *Optim Meth Software* 1998;10(1):1–31.
12. Vanderbei RJ. LOQO user's manual—version 3.10. *Optim Methods Softw* 1999;12:485–514.
13. Microsoft Corporation. Microsoft Solver Foundation. Available at <http://www.solverfoundation.com>. Accessed 2010.
14. Andersen E, Andersen K. The MOSEK optimization software. Denmark: EKA Consulting ApS. Available at <http://www.mosek.com>.
15. Czyzyk J, Mehrotra S, Wagner M, *et al.* PCx: an interior-point code for linear programming. *Optim Meth Software* 1999;11(1):397–430.
16. SAS Institute, Inc. SAS/OR software. <http://www.sas.com/technologies/analytics/optimization/or/>.
17. Dash Optimization. Xpress optimizer reference manual. Release 2001;13.
18. Nazareth JL. Computer solutions of linear programs. Oxford: Oxford University Press; 1987.
19. Cook W. LP format: BNF definition. Available at http://www2.isye.gatech.edu/~wcook/qsop/hlp/ff_lp_bnf.htm. Accessed 2003.
20. Fourer R, Gay DM, Kernighan BW. AMPL: a modeling language for mathematical programming. San Francisco, (CA): Scientific Press; 1993.
21. Brooke A, Kendrick D, Meeraus A. GAMS: A User's Guide. San Francisco, (CA): Scientific Press; 1988.
22. Czyzyk J, Mesnier M, Moré J. The neos server. *IEEE J Comput Sci Eng* 1998;5:68–75.
23. Murillo-Sánchez CE. A Matlab MEX interface for the BPMPD interior point solver. Available at <http://www.pserc.cornell.edu/bpmpd/>. Accessed 2009.
24. Carpenter TJ, Lustig IJ, Mulvey JM, *et al.* Higher order predictor-corrector interior point methods with application to quadratic objectives. *SIAM J Optim* 1993;3:696–725.
25. Monteiro RDC, Adler I. Interior path following primal-dual algorithms: Part II: convex quadratic programming. *Math Program* 1989;44:43–66.
26. Mehrotra S. On the implementation of a (primal-dual) interior point method. *SIAM J Optim* 1992;2:575–601.
27. Gondzio J. Multiple centrality corrections in a primal dual method for linear programming. *Comput Opt Appl* 1996;6:137–156.
28. Lougee-Heimer R. The Common Optimization INterface for operations research. *IBM J Res Dev* 2003;47(1):57–66.
29. Gupta A, Joshi M, Kumar V. Volume 1541/1998, WSSMP: a high-performance serial and parallel symmetric sparse linear solver. Umea, Sweden: Springer; 2006.
30. Toledo S, Chen D, Rotkin V. A library of sparse linear solvers. 2003. Available at <http://www.tau.ac.il/~stoledo/taucs>.
31. Amestoy P, Davis TA, Duff IS. Algorithm 837: AMD, an approximate minimum degree ordering algorithm. *ACM Trans Math Software* 2004;30(3):381–388.
32. Davis TA, Gilbert JR, Larimore S, *et al.* Algorithm 836: Colamd, a column approximate minimum degree ordering algorithm. *ACM Trans Math Software* 2004;30(3):377–380.
33. Gondzio J. Splitting dense columns of constraint matrix in interior point methods for large scale linear programming. *Optimization* 1992;24:285–297.
34. Gondzio J. Presolve analysis of linear programs prior to applying an interior point method. *ORSA J Comput* 2001;9(1):73–91.
35. Gondzio J. Implementing cholesky factorization for interior point methods of linear programming. *Optimization* 1993;27:121–140.
36. George A, Liu JWH. The evolution of the minimum degree ordering algorithm. *SIAM Rev* 1989;31:1–19.
37. George A, Liu J. Computer solution of large sparse positive definite systems. Englewood Cliffs, (NJ): Prentice-Hall; 1981.
38. Gondzio J, Grothey A. Reoptimization with the primal-dual interior point method. *SIAM J Optim* 2003;13(3):842–864.
39. Liu JW, Ng EG, Peyton BW. On finding supernodes for sparse matrix computations. *SIAM J Matrix Anal Appl* 1993;1:242–252.

40. Ng EG, Peyton BW. Block sparse cholesky algorithms on advanced uniprocessor computers. *SIAM J Sci Comput* 1993;14: 1034–1056.
41. Berger AJ, Mulvey JM, Rothberg E, *et al.* Solving multistage stochastic programs using tree dissection. Technical Report, Statistics and Operations Research, SOR-95-07. 1995.
42. Vanderbei RJ. A comparison between the minimum-local-fill and minimum-degree algorithms. Technical report. Murray Hill, (NJ): AT&T Bell Laboratories; 1990.
43. Anjos MF, Burer S. On handling free variables in interior-point methods for conic linear optimization. *SIAM J Optim* 2007;18(4): 1310–1325.
44. Andersen ED, Andersen KD. A parallel interior-point based linear programming solver for shared-memory multiprocessor computers: A case study based on the XPRESS LP solver. Technical report. Belgium: COTE, UCL; 1997.
45. Andersen ED, Andersen KD. The MOSEK interior-point optimizer for linear programming: an implementation of the homogeneous algorithm. In: Frenk H, *et al.*, editors. High performance optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000. pp. 197–232.
46. Coleman TF, Czyzyk J, Sun C, *et al.* pPCx: parallel software for linear programming. In: Proceedings of the Eighth SIAM Conference on Parallel Processing in Scientific Computing, Minneapolis, Minnesota; Mar 1997.
47. Curtis AR, Reid JK. On the automatic scaling of matrices for gaussian elimination. *J Inst Math Appl* 1972;10:118–124.
48. George JA, Liu JWH, Ng E. User guide for SPARSPAK, Waterloo sparse linear equations package. Technical Report 78-30 (revised 1980). Ontario, Canada: Department of Computer Science, University of Waterloo; 1980.
49. George JA, Liu JWH, Peyton BW. Computer solution of positive definite systems. Unpublished book available from the authors.

INTRODUCTION TO BRANCHING PROCESSES

PETER JAGERS
Department of Mathematical
Sciences, Chalmers
University of Technology
and University of
Gothenburg, Gothenburg,
Sweden

BACKGROUND

Branching processes describe sets of individuals (populations) evolving over time, their size and composition changing by members generating new individuals (giving birth), and possibly disappearing (dying). Thus, they form the basic stochastic model for biological population dynamics, including demography. They can, however, also be used to describe other phenomena where the same pattern occurs, like in nuclear physics, molecular replication, or the performance of a certain computer algorithm.

There are branching processes of varying complexity. In the simple Galton–Watson process, real time is replaced by generation counting: individuals of one generation individually and independently give birth to a number of new individuals according to the same offspring number distribution. These then form the next generation. Another simple form is known as the birth-and-death process (see *Birth-and-Death Processes*). Here time is real, but there is no aging of individuals and reproduction is reduced to binary splitting, or death without children.

On the other end of the spectrum, are general branching processes, where individuals can be of various types, life spans follow arbitrary, type-dependent distributions and births occur as general point processes (see *Introduction to Point Processes*) during life, that is, in any random pattern. These can also be dependent upon the

mother's type, and children can have new types. In classical theory, individuals are still supposed to reproduce asexually and otherwise independently of one another. Recent work has dealt with population size dependence, so that small populations for instance, could reproduce more than big populations do.

Historically, branching processes were born out of the so-called *extinction paradox*; that is, the fact that a rapid growth of a population as a whole does not exclude frequent extinction of separate family lines. This seems first to have been noticed in a human context, as the family (name) extinction problem, by Malthus and a circle of French demographers. (In the fight for survival in biological evolution, extinction is even more prevalent, but we are not considering extinction as a consequence of competition.) The first mathematical treatment seems to be due to I. J. Bienaymé (1845), and later by F. Galton and H. Watson (1873, 1875), who however were led to conclude that all families must die out, even if reproduction is supercritical, that is, the mean offspring number per individual exceeds one. In spite of the implausibility of this result, and a certain controversy about it at the time, it was corrected only half a century later, by Haldane (1927) and definitely, by Steffensen (1930).

The correct extinction theorem reads: Consider a single-type branching process, that is a population where individuals reproduce independently and according to one distribution. If the mean offspring number per individual, m , is less than or equal to one, then the population must die out, while if $m > 1$ there is a positive chance of survival. The extinction probability q for a population with one ancestor is the smallest nonnegative root of the equation $f(q) = q$, where f is the probability generating function of the offspring number distribution. Thus formulated, the theorem is valid not only for Galton–Watson processes but for quite general branching processes. Properly rephrased, it even extends to multitype processes.

The source of confusion was probably that for natural populations, the reproduction law often has both a big mean m and a comparatively large extinction probability q , due to high probability of begetting no children combined with a strong convexity of the reproduction-generating function f .

Another source of inspiration for branching process theory is the growth of populations and the ensuing stabilization of composition. Here, theory goes back to Euler, and many contributions were formulated in the late nineteenth and early twentieth centuries by demographers and actuaries, mostly in a deterministic framework. The exponential rate of free population growth, ascribed to Malthus, is a fundamental result, another is the stable age distribution, in its first version due to Euler himself.

The extinction problem can be formulated in terms of simple branching processes, whereas the growth rates and composition matters can only be meaningfully resolved within models with more realistic time structure. In the case of large populations, the need for a stochastic formulation of individual behavior is also not so clear and as a consequence, it is only rather recently that branching process theory has come to embrace demographic theory and deterministic, so-called structured population dynamics. When it comes to population size dependence and other interplay between environment and individual reproduction, only simple special cases have been treated.

GALTON–WATSON PROCESSES

A (*Bienaymé*–)Galton–Watson or simple branching process $\{Z_n, n = 0, 1, \dots\}$, where Z_n stands for the size of the n th generation, is usually defined from an array of independent, identically distributed non-negative integer-valued random variables $\{X_{ni}; n, i = 1, 2, \dots\}$, with the interpretation of X_{ni} being the number of children of the i th individual of the n th generation, and a starting value Z_0 , conventionally taken as 1, unless something else is mentioned. The technical definition is then recursive

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{ni}, n = 0, 1, 2, \dots$$

The idea is either that time is disregarded and only generations counted, or else that the scheme should depict the evolution of seasonal populations, say with a life span of one year, like for many plants or insects. Processes are referred to as *supercritical* if the reproduction mean $m = E[X_{ni}] > 1$, critical if $m = 1$, and subcritical otherwise. The *reproduction law* is usually denoted $p_k = P(X_{ni} = k), k = 0, 1, 2, \dots$, so that $m = \sum k p_k$, and the reproduction-generating function, referred to in the previous section is $f(s) = \sum s^k p_k, 0 \leq s \leq 1$.

The most important problem naturally solved within this context is the extinction problem, already mentioned, where a finer time structure is irrelevant. Simple branching processes, however, also serve as a touchstone of results for more refined models. Beyond the extinction theorem, the following basic facts have been proved and in principle, extend to more general forms of branching processes:

1. Exponential growth: If a process does not die out, then it grows like m^n , if and only if there is a finite logarithmic moment of reproduction: $E[X_{ni} \log X_{ni}] < \infty$, the so-called $x \log x$ condition.
2. In a subcritical process, the survival probability decreases like m^n , again under the $x \log x$ condition. Given that the population has not died out, its size stabilizes.
3. In a critical process with finite reproduction variance, $\sigma^2 = \text{Var}[X_{ni}] < \infty$, the survival probability decreases like $2/n\sigma^2$. Conditionally upon nonextinction, critical populations grow linearly.

SPECIAL GALTON–WATSON PROCESSES, GENERALIZATIONS, RAMIFICATIONS

The case of *binary splitting* is of special biological interest, $p = p_2 = 1 - p_0$, occurring in

cells and bacteria. Then obviously $m = 2p$. A variation takes place in molecular replication, where the molecule remains, if it does not double: $p = p_2 = 1 - p_1$ [1].

Various forms of *linear fractional* reproduction can be easily studied and thus command more of purely mathematical interest. The reason is that here the chief analytical tool, reproduction-generating functions, can be explicitly iterated and thus, the behavior of $\{Z_n\}$ often easily found. The reproduction probabilities have the geometric form $p_0 = 1 - a, p_k = ab^{k-1}(1 - b), 0 \leq a \leq 1, 0 \leq b < 1, k = 1, 2, \dots$, yielding generating functions of precisely a linear fractional form. Studies of these processes can give hints about more general or scientifically more relevant branching processes.

Generalizations include *branching processes with immigration* and branching processes in *random environments*. In the former, an immigration component is added to the children of the preceding generation so that the recursion runs like

$$Y_{n+1} = \sum_{i=1}^{Y_n} X_{ni} + I_n,$$

the last variable being the immigration, usually assumed independently and identically distributed (i.i.d.).

Random environments were initially thought to cover cases like weather fluctuations, the mathematical formalization then being that first there is an independent choice of reproduction law, then all reproductions occur in an i.i.d. fashion according to that. Gradually, they have developed into other forms, one of special interest being the case where the environment in which individuals multiply is influenced by the population size itself. In this way, processes can be constructed which are supercritical under and subcritical above a critical population size, in ecology often referred to as the *carrying capacity*.

Interestingly, such processes retain the basic branching process property of either dying out or growing beyond limits. Indeed, the dichotomy between extinction and explosion, that is the impossibility of population size ultimately stabilizing, is an inherent

property of population models that allow leeway for individual variation in reproduction [2, p. 108]. However, population size dependence can give rise to interesting phenomena like *quasi-stationarity*, the time to extinction including a long-time (exponential in the carrying capacity) plateau of oscillation around it. Binary models for molecular replication that include Michaelis–Menten type population size dependence of the replication success probability lead to linear growth of the type observed in experiments, rather than classical exponential growth [2, p. 231].

In the Galton–Watson context of nonoverlapping generations, processes with two sexes and mating have also been investigated, often under the heading of *bisexual* branching processes.

MULTITYPE BRANCHING PROCESSES IN DISCRETE TIME

If mating were not required, but females could reproduce independently, getting both daughters and sons, the latter being sterile, this would provide an example of two-type branching. More generally, *multitype* Galton–Watson processes can be formulated by replacing the random variables X_{ni} by vectors, giving the numbers of children of various types. The X_{ni} are still thought of as independent, but their distributions may be influenced by the mother's type. For further reading refer article titled this encyclopedia or Ref. 3. There is also a theory of general type spaces, the case of discrete time with nonoverlapping generations to be found in Ref. 4, general theory being touched upon in Ref. 2, where further references are also provided.

If types communicate in the sense that an individual of any type can have progeny of any other type, much of single-type theory carries over. The mean matrix $M = \{m_{ij}\}$, where the elements give the expected number of j -children to an i -individual plays a crucial role, and its largest eigenvalue determines the criticality properties of the process.

In genetics, processes where types do not communicate, but fresh types can arise out of

mutation, are of great interest. An extreme version of this is the infinite-alleles model, where all mutations introduce new alleles. Another is provided by branching processes with immigration, where the immigration can be viewed as progeny from a type that besides this progeny, always gets one child of its own type.

AGE-DEPENDENT AND GENERAL BRANCHING PROCESSES

The simplest type of branching processes in continuous time, are known as *birth-and-death* processes. Here, the idea is that individuals give birth at a constant rate λ , at the same time being subject to a constant death risk, μ . This means that the intensity of something, whatever it is, happening in an individual's life is $\lambda + \mu$, and when something happens it is death with probability $\mu/(\lambda + \mu)$ and birth with the complementary probability. An alternative interpretation is therefore, that the individual has an exponentially distributed life span, with parameter $\lambda + \mu$, ended by binary splitting. Since exponential life spans mean that there is no aging, we need not discern between a mother and a newborn.

Birth-and-death processes lend themselves to an immediate generalization *Markov branching processes*: individuals having exponentially distributed life spans end with independent division into a random number of offspring. This is the continuous-time counterpart to Galton–Watson processes, which are discrete-time Markov chains, and results about the latter are readily transferred to continuous time. Sometimes matters are even more easy here [5,6].

Markov branching processes in their turn, lead on to *age-dependent* or *Bellman–Harris* processes. Here, the life span distribution is arbitrary, and the cell or particle is supposed to split into an offspring of random size, independently of the length of life. Thus, the process is age-dependent in the sense that for a Markovian description, ages of all individuals are indispensable. Otherwise, the future development is not well determined, as opposed to the Markov branching case, where

the exponential distribution means that age does not influence future development. On the other hand, reproduction remains independent of the mother's age at death. This was generalized by Sevastyanov into truly age-dependent processes [7].

The next step in the development of branching process theory was the introduction of *general* branching processes, also known under the acronym of CMJ (Crumpp-Mode-Jagers)-processes [2,8]. Here complete generality is achieved, within the frames of independent, asexual reproduction. Individuals can have arbitrary life span distributions, and give birth according to an arbitrary random pattern, very erratically like human females, repeatedly in litters or broods of random size, or splitting like cells or elementary particles. The type space can also be quite general, the reproduction process then taking the form of a marked point process, marks being the types of the newborns.

The extinction or explosion dichotomy persists, and under a generalized $x \log x$ condition, the population exhibits exponential growth, at a so-called *Malthusian rate* if it is supercritical, and the survival chance decreases at an exponential rate in the subcritical case. Also in the critical case, the basic pattern from simple Galton–Watson processes carries over.

Technically, the analysis of general branching processes is much more involved. Martingales and renewal theory (in multi-type formulations Markov renewal theory, (see **Markov Renewal Function and Markov Renewal-Type Equations**), replace the subtle, elementary analysis of reproduction-generating functions that is crucial in the analysis of Bienaymé–Galton–Watson processes or the differential equations of Markov branching.

The concept of *random characteristics* [2] permits the measuring or counting of populations in a multitude of ways. We can count the population of all those alive, or (mathematically fundamental) all those born, or those below a certain age, or those having a certain ancestry, or progeny, in a certain phase of life, or measure the total body mass or DNA content of the population. Limit

theorems (see *Limit Theorems for Branching Processes*) will then exhibit stable compositions in populations not dying out. A very specific example is the stable age distribution of demography, others are the mitotic index of cell kinetics, and even the phylogeny of a random, typical individual.

This all concerns populations not dying out. Interestingly enough, a branching process known to die out, remains a branching process; that is, individuals remain independent in spite of the global condition of extinction. If the original process was supercritical, the new process will be subcritical, and subcritical general branching processes emerge as the natural model for time and path to extinction of threatened populations.

As the starting number grows, a clear pattern emerges: the time T to extinction behaves proportionally to the logarithm of the starting number minus a constant plus a Gumbel random variable. There is also convergence of the suitably normed population size u -ways to extinction, that is at times $uT, 0 \leq u < 1$ [9].

REFERENCES

1. Jagers P, Klebaner FC. Random variation and concentration effects in PCR. *J Theor Biol* 2003;224:299–304.
2. Jagers P. Branching processes with biological applications. Chichester: Wiley; 1975.
3. Kimmel M, Axelrod DE. Branching processes in biology. New York: Springer; 2002.
4. Harris TE. The theory of branching processes. Berlin: Springer; 1963.
5. Athreya KB, Ney P. Branching processes. Berlin: Springer; 1972.
6. Mode C. Multitype branching processes. Oxford: Elsevier; 1971.
7. Sevastyanov BA. Vetvyashchiesya protsessy (Branching processes). Moscow: Nauka; 1971. Also available in German: Verzweigungsprozesse. Berlin: Akademie Verlag; 1974.
8. Haccou P, Jagers P, Vatutin VA. Branching processes: variation, growth, and extinction of populations. Cambridge: Cambridge University Press; 2005.
9. Jagers P, Klebaner FC, Sagitov S. On the path to extinction. *Proc Natl Acad Sci U S A* 2007;104:6107–6111.

INTRODUCTION TO DIFFUSION PROCESSES

ARKA P. GHOSH
Department of Statistics,
Iowa State University,
Ames, Iowa

INTRODUCTION

The history of diffusion processes begins with the botanist Brown, who during 1826–1827 observed that grains of pollen suspended in a water display a certain type of erratic motion, which did not fit into any of the contemporary mathematical models [1]. This motion came to be known as the Brownian motion. In 1905, Einstein used physical principles to conduct a mathematical analysis of this motion and Wiener provided a rigorous mathematical foundation for the Brownian motion (hence Brownian motion is also called the *Wiener process*). See ***Brownian Motion and Queueing Applications*** for more details on the Brownian motion. Diffusion processes, in some sense, are generalized versions of the Brownian motion. To fully understand the diffusion process, the reader should be familiar with the concept of stochastic calculus. We begin by introducing stochastic differential equations (SDEs) using heuristic arguments, followed by a brief introduction to stochastic calculus which is used to make the SDE formulation more rigorous. General diffusion process is defined next as a solution to SDEs. Finally, we discuss some major properties of diffusion processes and some common examples of diffusion processes.

PRELIMINARIES

It is common to characterize the evolution of a system through differential equations, which typically describes the rate of change of the system, and in some ways, as functions of other variables (e.g., current value of the

state, other variables etc.). Here, the state of the system can be, for example, the length of a queue in a (deterministic) model: Denoting the length of a queue at time $t \geq 0$ as $q(t)$, one may have the following characterization of the evolution of the queue over time,

$$q(0) = q_0, \frac{d(q(t))}{dt} = \theta(t) q(t), \quad t \geq 0. \quad (1)$$

Here, $\theta(t)$ represents the rate of growth (or “decay”) of the queue-length at time $t \geq 0$, and q_0 is the initial queue-length at time $t = 0$. This “population-growth”-type model is quite common in many applications, and makes perfect sense if the growth parameter is a completely known function. However, in reality it is often not known completely; we may know that the rate function is “approximately” $b(t)$ but it is reasonable to assume that it is subject to some random environmental effects, so that we have

$$\theta(t) = b(t) + \text{“noise”},$$

where we do not know the exact behavior of the noise term, but we know its probability distribution. The following text describes how Equation (1) is dealt with, in the presence of random noise.

Heuristic formulation of SDEs

One usually formulates Equation (1) in this more general version:

$$\frac{dQ(t)}{dt} = b(t, Q(t)) + \sigma(t, Q(t)) \cdot \text{“noise”}, \quad (2)$$

where b and σ are some given functions. For a good model for the “noise” part of the above equation, it is common to use a white-noise type model using the “increments” of the Brownian motion $\{W(t)\}$ (see Section 2.1.6.3). Writing Equation (2) in the increment form, we get that for any $0 = t_0 < t_1 < \dots < t_m = t$,

$$(Q_{k+1} - Q_k) = b(t_k, Q_k)(\Delta t_k) + \sigma(t_k, Q_k)(\Delta W_k), \quad (3)$$

where $Q_j = Q(t_j)$, $\Delta W_j = W(t_{j+1}) - W(t_j)$, $\Delta t_k = t_{j+1} - t_j$. Thus, using Equation (2), we get the following formulation describing the dynamics of the process Q , assuming $Q(0) = q_0$:

$$Q(t) = q_0 + \sum_{k=1}^{m-1} b(t_k, Q_k)(\Delta t_k) + \sum_{k=1}^{m-1} \sigma(t_k, Q_k)(\Delta W_k). \quad (4)$$

Now this formulation depends on the time-discretization, and it would be more acceptable in the limiting form as $\Delta t_k \rightarrow 0$, if the limit on the right hand side exists. This led to the formulation of stochastic integrals (which we describe below) and led to the following equation:

$$Q(t) = q_0 + \int_0^t b(s, Q(s)) ds + \int_0^t \sigma(s, Q(s)) dW(s), \quad (5)$$

Here the last integral is defined as the stochastic integral. The above equation is called the Stochastic Differential Equation (SDE) describing the dynamics of the process Q . In addition to the above “integral form”, one often uses an equivalent “differential form” to describe the dynamics of Equation (5):

$$Q(0) = q_0, \quad dQ(t) = b(t, Q(t)) dt + \sigma(t, Q(t)) dW(t), \quad t \geq 0. \quad (6)$$

A diffusion is the solution to the Equation (5) or (6) described above, when it exists. Below, we give a brief introduction to the stochastic calculus that is used to define the stochastic integral in these equations.

Stochastic Integration

Note that there are two types of integral expressions (involving some stochastic process, say $\{X(s)\}$ and some function $f(\cdot, \cdot)$) in (5),

$$\int_0^t f(s, X(s)) ds, \quad \int_0^t f(s, X(s)) dW(s),$$

where W is a Brownian Motion. In most situations, the first integral can be interpreted simply as the Riemann integral, in the classical way. There are several ways of interpreting the second integral; most of them involve viewing this integral as a suitable limit of the partial sum

$$\sum_{k=1}^m f(t_k, X(t_k)) \Delta W_k,$$

as $\Delta t_j \rightarrow 0$ (here, $\{t_j\}$, Δt_j , W_k is as defined in Equation (3)). This leads to the most widely used version of stochastic integrals, called the *Itô integral* (Refs 1 or 2 for more detail). This is the most common approach to define stochastic integrals, where the function-value at left-end-point of the intervals $[t_j, t_{j+1})$ is used to approximate the integrand. One popular alternative (where the middle point of the interval is used) to this is the *Stratonovich Integral* (2, Chapter 3).

Itô Integrals

Let $f(t, x)$ be a continuous function in t and x , such that

$$\int_0^t E(f^2(u, X(u))) du < \infty.$$

Then the Ito integral of $f(t, X(t))$ with respect to $W(t)$ is defined as

$$\int_0^t f(s, X(s)) dW(s) = \lim_{\Delta_k \rightarrow 0} \sum_k (f(t_{k+1}, X_{k+1}) - f(t_k, X_k)) \Delta W_k, \quad (7)$$

where $\Delta_k = \max_k \Delta t_k$, and the limit is defined in the mean-squared sense (a random variable V is called the “mean square limit” of a sequence $\{V_n\}$ if $\lim_{n \rightarrow \infty} E[(V_n - V)^2] = 0$). For the above definition to make sense, we also need to have that the process X is “adapted” to the filtration generated by the Brownian motion W . In simple words, if $\mathcal{F}_t$ denotes the sigma-algebra generated by $\{W(s) : s \leq t\}$ (it contains all the information of W -process till the time t), then $X(t)$ has to be measurable with respect to $\mathcal{F}_t$, for all $t \geq 0$, that is, $X(t)$ can be completely determined using the knowledge of $W(s)$ for $s \leq t$.

DIFFUSION PROCESS

As defined earlier (Eqs 5–6), a diffusion process $\{X(t)\}$ is a solution to the following equation:

$$X(t) = x_0 + \int_0^t b(s, X(s)) ds + \int_0^t \sigma(s, X(s)) dW(s), \quad (8)$$

or, equivalently, to

$$X(0) = x_0, \quad dX(t) = b(t, X(t)) dt + \sigma(t, X(t)) dW(t), \quad (9)$$

where $b(t, x)$ and $\sigma(t, x)$ are continuous functions of t and x , such that

$$\int_0^t E(\sigma^2(s, X(s))) ds < \infty.$$

Here, $b(\cdot, \cdot)$ is called its drift function, and $\sigma(\cdot, \cdot)$ the diffusion function. Here b and σ can be thought of as the infinitesimal mean and variance of the increments of the process X (see **Brownian Motion and Queueing Applications** for more details).

First note that a diffusion process is defined as a solution to a (stochastic differential) equation, and we have to discuss the conditions under which the solution exists. But at this point, it is easy to consider some very basic examples:

Example 1. Setting $x_0 = 0$, $b(t, x) = 0$ and $\sigma(t, x) = 1$ yields that $X(t) = W(t)$, the standard Brownian motion starting at the origin. Similarly, for constant coefficients $b(x, t) = b$, $\sigma(x, t) = \sigma$ yields a Brownian motion with drift b , diffusion parameter σ . So, diffusion processes are more general than the Brownian motion processes described in section 2.1.6.3.

The following result gives a general set of conditions under which diffusion exists (i.e., there is a solution for the SDE in Eq. 8):

Existence Conditions for Diffusion Processes

Let $T > 0$ and $b(\cdot, \cdot), \sigma(\cdot, \cdot)$ be measurable functions satisfying for all real x , and for all $t \in [0, T]$

$$|b(t, x)| + |\sigma(t, x)| < C(1 + |x|), \quad \text{and} \\ |b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| < D|x - y|,$$

for some constants $C > 0, D > 0$, then there exists solution of the SDE in Equation (8). In addition, this solution is also unique. Under much weaker conditions (e.g., continuity of b and σ are strictly nonnegative), one gets uniqueness in the sense that any two solutions will have the same probability distribution (“weak” uniqueness). See Chapter 5.2 of Ref. 2 for more discussion on this.

Example 1 describes one of the simplest types of such processes and now we know conditions under which other diffusion processes exist. Now, suppose $X(t)$ is a diffusion process and $Z(t) = g(X(t))$ for a “nice” function $g(\cdot)$. Can we characterize the dynamics of such a process $\{Z(t)\}$? The answer is provided by the celebrated *Itô’s formula* described below:

Itô’s Formula

Let $\{X(t)\}$ be a diffusion process (i.e., a solution of Eq. 8). Let $g(t, x)$ be a function that is continuously differentiable in t and twice continuously differentiable in x . The stochastic process defined as $Z(t) = g(t, X(t))$ satisfies the following stochastic integral equation:

$$Z(t) = Z(0) + \int_0^t (g_t(s, X(s)) + b(s, X(s))g_x(s, X(s)) + \frac{1}{2}\sigma^2(s, X(s))g_{xx}(s, X(s))) ds + \int_0^t \sigma(s, X(s))g_x(s, X(s)) dW(s), \quad (10)$$

or, equivalently, the stochastic differential equation

$$\begin{aligned}
Z(0) &= g(0, x), \\
dZ(t) &= \left(g_t(t, X(t)) + b(t, X(t))g_x(t, X(t)) \right. \\
&\quad \left. + \frac{1}{2}\sigma^2(t, X(t))g_{xx}(t, X(t)) \right) dt \\
&\quad + \sigma(t, X(t))g_x(t, X(t)) dW(t). \quad (11)
\end{aligned}$$

Here g_t, g_x denote the partial derivatives of g with respect to t and x , respectively and g_{xx} is the second order partial derivative with respect to x .

The above formula is extremely important for dealing with diffusion processes and getting explicit solutions for them (by solving the defining SDEs as in Eq. 8). The following give some applications of this formula, as well as some standard examples of diffusion processes.

EXAMPLES

In Example 1, we already saw that Brownian motion is a simple example of a diffusion process. Here we discuss a few more standard examples.

Example 2 [Geometric Brownian Motion Process]. Recall the motivating example at the beginning of this section (Eq. 2). In this situation, a queue-length process is modeled as the following diffusion process:

$$Q(0) = q_0, \quad dQ(t) = bQ(t)dt + \sigma Q(t)dW(t), \quad (12)$$

where b and $\sigma > 0$ are constants. To solve the above SDE, first we write it as follows:

$$\frac{dQ(t)}{Q(t)} = b dt + \sigma dW(t). \quad (13)$$

Next, applying Itô's formula with $g(t, x) = \ln(x)$ to the diffusion Q described in Equation (12), we get (after simplification) that

$$d\ln(Q(t)) = \frac{dQ(t)}{Q(t)} - \frac{1}{2}\sigma^2 dt. \quad (14)$$

Using Equation (13) and (14), we get that

$$b dt + \sigma dW(t) = d\ln(Q(t)) - \frac{1}{2}\sigma^2 dt,$$

and hence, together with the fact that $Q(0) = q_0$, we have that

$$Q(t) = q_0 \exp\left(\left(b - \sigma^2/2\right)t + \sigma W(t)\right).$$

This is the explicit expression of the diffusion process (implicitly) defined in Equation (12). This diffusion process is known as the *Geometric Brownian Motion process* with parameters b and σ .

Example 3 [Ornstein-Uhlenbeck (OU) Process]. This is another common diffusion process defined by

$$X(0) = x, \quad dX(t) = -bX(t)dt + \sigma dW(t), \quad (15)$$

where $b > 0$ and $\sigma > 0$ are constants (here, b and σ are called the parameters of the OU process). One can show using calculations similar to the ones above that the explicit expression of this diffusion process is as follows:

$$X(t) = \exp(-bt) \left(x + \sigma \int_0^t \exp(bs) dW(s) \right). \quad (16)$$

As we mentioned earlier, diffusion equations are extremely important in operations research, especially in the context of modeling queues, storage processes and so on. In continuous-time models, it is often natural to use discrete-state processes (such as DTMC models in the section titled “Definition and Examples of CTMCs” in this encyclopedia), because of discrete nature of the state-space (e.g., number of people in the queue). However, under suitable scalings (“diffusion”-scaling), the discrete-state system can be approximated by a suitable diffusion. Often the physical models have natural constraints (such as queue-lengths being nonnegative) which lead to constrained diffusions. Since there is a large literature on analysis of the diffusion processes, understanding the approximate system (driven by diffusions) is often possible and usually provides useful insights for the discrete-state physical systems. We provide a list of references below for further reading on these topics.

FURTHER READING

More detail about diffusion processes can be found in classical textbooks on probability, such as Ref. 3. For more recent developments, we encourage the reader to consult Refs 4–7 (incomplete list). For students and researchers learning from this material for the first time, some more accessible texts would be Refs 1 and 2. In many situations, diffusions are Markov processes and hence mathematical tools for Markov processes can be used for diffusions. A reference for general Markov processes would be Ref. 8.

Diffusion processes are often used in physics [9], financial applications [10,11] and many other disciplines. For operations research applications, a general reference would be Ref. 12. For more detailed reference for diffusion processes in queueing theory models, we recommend Refs 13–15.

REFERENCES

1. Karatzas I, Shreve S. Brownian motion and stochastic calculus. 2nd ed. New York: Springer; 1991.
2. Oksendal BK. Stochastic differential equations: an introduction with applications. 6th ed. Berlin, Heidelberg: Springer; 2003.
3. Feller W. An introduction to probability theory and its applications. Volume 2. 2nd ed. New York: John Wiley & Sons; 1991.
4. Ikeda N. Ito's stochastic calculus and probability theory. New York: Springer; 1996.
5. Itô K, McKean HP. Diffusion processes and their sample paths. New York: Springer; 1996.
6. Protter P. Stochastic integration and differential equations. 2nd ed. New York: Springer; 2005.
7. Stroock DW, Varadhan SRS. Multidimensional diffusion processes. New York: Springer; 2005.
8. Ethier SN, Kurtz TG. Markov processes. Characterization and convergence. Wiley series in probability and mathematical statistics: probability and mathematical statistics. New York: John Wiley & Sons, Inc.; 1986.
9. Kadanoff LP. Statistical physics: statics, dynamics and renormalization. World Scientific; 2000.
10. Karatzas I, Shreve S. Methods of mathematical finance. New York: Springer; 2001.
11. Steele JM. Stochastic calculus and financial applications. New York: Springer; 2000.
12. Kulkarni VG. Modeling and analysis of stochastic systems. 1st ed. Chapman & Hall; 1996.
13. Whitt W. Stochastic-process limits. New York: Springer. 2002.
14. Kushner HJ. Heavy traffic analysis of controlled queueing and communications networks. 1st ed. New York: Springer; 2001.
15. Chen H, Yao D. Fundamentals of queueing networks. 1st ed. New York: Springer; 2001.

INTRODUCTION TO DISCRETE-EVENT SIMULATION

RICKI G. INGALLS
School of Industrial Engineering
and Management, Oklahoma
State University, Stillwater,
Oklahoma

DEFINITION OF SIMULATION

Simulation, according to Shannon [1], is “the process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behavior of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of the system.” According to this definition, a simulation can be a discrete-event simulation, as we will discuss in this article. Many people are familiar with the term “MRP simulation.” This is a model (actually a copy) of the real system (the MRP system of record) on which experiments (or scenarios) can be run to evaluate various strategies (such as how to respond to a drastic change in the forecast). Although we do not teach courses in our curriculum on “MRP simulation,” an MRP simulation is no less a simulation than the type of simulation we will discuss in this article.

The difference, and the power, of discrete-event simulation is the ability to mimic the *dynamics* of a real system. Many models, including high-powered optimization models, cannot take into account the dynamics of a real system. It is the ability to mimic the dynamics of the real system that gives discrete-event simulation its structure, its function, and its unique way to analyze results. So, to take liberties with one of my mentors in this field, we will say that simulation is the process of designing a *dynamic*

model of an actual *dynamic* system for the purpose either of understanding the behavior of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of the system.

Throughout this article, we reference ideas and thoughts from Shannon [1], Law and Kelton [2], Banks *et al.* [3], and Kelton *et al.* [4].

A DRIVE-THROUGH EXAMPLE

In order to give us a reference model for discussion purposes, let us consider the example of a drive-through window at a fast-food restaurant whose logic is shown in Fig. 1. As a car enters from the street, the driver, who we call Fred, decides whether or not to get in line. If Fred decides to leave the restaurant, he leaves as a dissatisfied customer. One of the great things about a simulation is that it is easy to track these type of customers. In most real-world systems, it is usually difficult to track customers who leave dissatisfied.

If Fred decides to get in line, then he waits until the menu board (with the speaker no normal human being can understand) is available. At that time, Fred gives the order to the order taker.

After the order is taken, two things occur simultaneously: (i) Fred moves forward if there is room. If there is no room, then he has to wait at the menu board until there is room to move forward. As soon as there is room, he moves so the next customer can order. (ii) The order is sent electronically back to the kitchen where it is prepared as soon as the cook is available.

As soon as Fred reaches the pickup window, he pays and picks up his food, if it is ready. If the food is not ready, then Fred has to wait until his order is prepared. In this drive-through, Fred never hears, “Please pull forward and we will bring your order out

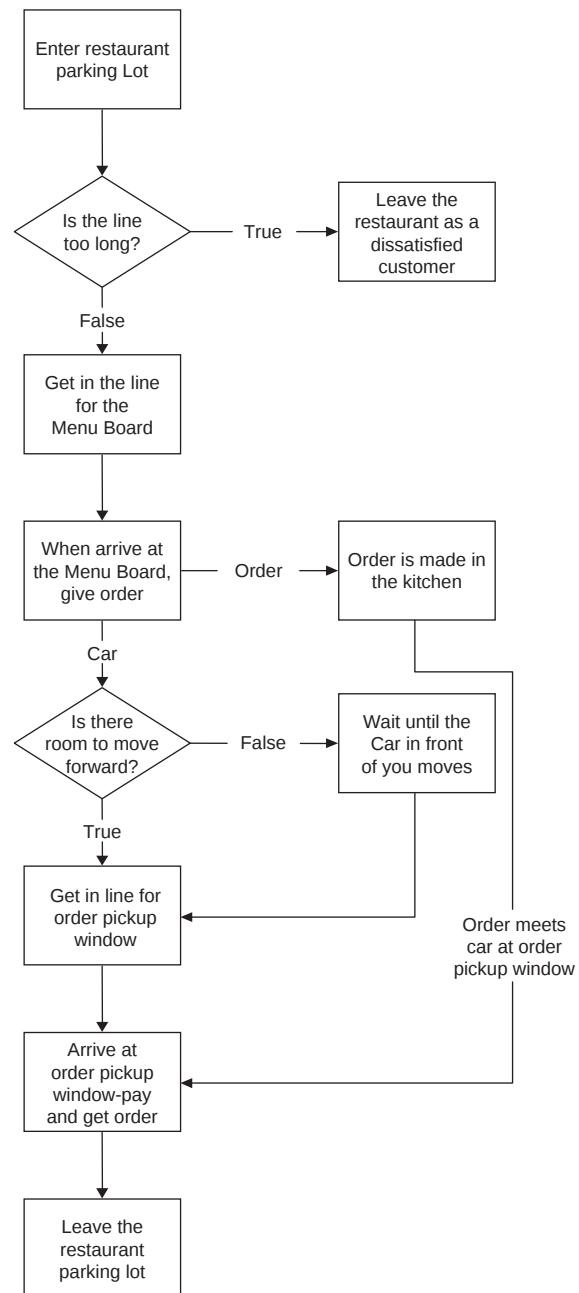


Figure 1. Drive-through example logic flow.

to you.” That is one of the reasons that Fred and I both like this place.

As soon as Fred parts with his money and gets his food, he leaves the pickup window as a satisfied customer.

DISCRETE-EVENT SIMULATION STRUCTURE

Although there are various flavors and paradigms in discrete-event simulation, there has evolved a basic structure that is

used by most simulation packages. Regardless of how complex a discrete-event simulation package may be, it is likely to contain the basic components that we describe in this section.

The structural components of a discrete-event simulation include *entities*, *activities and events*, *resources*, *global variables*, a *random number generator*, a *calendar*, *system state variables*, and *statistics collectors*.

Entities

The best way to understand the function of an entity is to understand that *entities cause changes in the state of the simulation*. Without entities, nothing would happen in a simulation. As a matter of fact, one stopping condition for a simulation model is the condition where there are no active entities in the system.

Entities have *attributes*. Attributes are characteristics of a given entity that are unique to that entity. Attributes are critical to the understanding of the performance and function of entities in the simulation.

In our example, the primary entity type is the cars coming to the restaurant. Also, there is a second entity type created when the order is taken, and that is the order itself. It has a relatively short life in the simulation, lasting only from the time the order is taken until it meets back up with the car at the pickup window.

We also have two attributes in our simulation. The first one is the time of day that the car enters the restaurant parking lot. We call this attribute *StartTime*. The second is the value of the order. We call this attribute *OrderValue*. Both these are unique to the customers who are represented in the simulation.

Another example of an entity is a part that is flowing through a factory. Entities that represent parts in a factory could be created randomly or according to a schedule. A common attribute would be the time that the part started in the factory. Each part would have a unique time that it started in the factory. It may also have other attributes such as priority, the type of part, and the cost incurred to produce the part. In a normal

factory simulation that is tracking each individual part, it would not be unusual to have thousands of entities active in the simulation simultaneously.

Entities, however, do not need to be parts in a manufacturing facility or cars in a drive-through. It is also a common practice to use entities to represent the flow of *information*. This information could be a customer order, an e-mail alert, a packet in a computer network, and so on. It can be anything nonphysical that causes a change in the status of the system. These entities also have attributes. For example, an entity representing a packet in a computer network would have attributes such as packet size, packet destination, the time that the packet would time-out, and so on.

Activities and Events

Activities are processes and logic in the simulation. *Events* are conditions that occur at a point in time which cause a change in the state of the system. An entity *interacts* with activities. Entities *interacting with* activities *create* events.

There are three major types of activities in a simulation: *delays*, *queues*, and *logic*. The *delay* activity is when the entity is delayed for a definite period of time. In our example, there are three delays. The first is when Fred is ordering at the menu board, the second is when the order is being cooked in the kitchen, and the third is when Fred picks up his order at the pickup window. In general, the length of time of a delay is either constant or is randomly generated. At the point that the entity starts the delay, an event occurs. This event schedules the entity on the *calendar* (which we will get to later). If the delay is for d time units, then the entity is scheduled to complete the delay d time units after the current time of the simulation. At that time, the delay expires and another event is generated.

Queues are places in the simulation where entities wait for an unspecified period of time. Entities can be waiting on *resources* (which we will get to later) to be available or for a given system condition to occur. *Queues* are most commonly used for waiting in line for a resource or storing material that will be taken out of the queue when the right

conditions exist. In our example, there are three queues: the first is the part of the line that waits on the menu board to be available, the second is the order waiting on the kitchen becoming available, and the third is the part of the car line waiting on the order pickup window to become available. All three of these queues are waiting for *resources* to be available.

Logic activities simply allow the entity to effect the state of the system through the manipulation of state variables (which we get to later) or decision logic. The first of several logic activities in our example is the decision whether to get in the order line in the first place. This decision is effected by the length of the line in front of the menu board.

Resources

In a simulation, resources represent anything that has a restricted (or constrained) capacity. Common examples of resources include workers, machines, nodes in a communication network, traffic intersections, and so on. In our example, we have three different resources. The first resource is the menu board, which only one car can use at a time. The menu board is utilized from the time a car moves in front of it until the car moves away from it. In our model, the car does not automatically move away from the menu board after the order has been placed. There must also be room to move forward. So, this resource is occupied for both productive time (when the order is being placed) and unproductive time (when there is not enough room to pull the car forward). The second resource is the kitchen. It is utilized from the time an order arrives until the order is finished cooking. The third resource is the order pickup window. This resource can also be unproductive. That occurs when the car has arrived at the window, but the order is not ready from the kitchen yet. In the course of the simulation, we can track the key statistics of each of these resources, including utilization and costs.

It should also be noted that very complex resources can be utilized in a simulation. In a manufacturing simulation, conveyors are a very complex resource that many simulation packages offer. Also, transportation options

such as trucks are offered as resources. A third complex resource is a vat or container that has a (continuous) flow of material both in and out of the resource. Depending on the target market of the simulation package, many complex resources are available for use.

Global Variables

If you are a programmer, then the idea of having *global variables* is nothing new. A *global variable* is a variable that is available to the entire model at all times. A global variable can track just about anything that is of interest to the entire simulation. In our model, we have four global variables, two of which help us configure the problem and two of which collect revenue information. The two variables that help us configure the problem are the length of the line allowed at the restaurant. The first is *MenuBoardLineSize*, which gives the maximum line size for the menu board, including the car that is at the menu board. The second is *OrderPickupLineSize*, which gives the maximum line size after the menu board, including the car at the pickup window. By changing these two variables, the simulator can analyze the effectiveness of different line configurations.

The two variables that help use collect revenue information are *Revenue* and *LostRevenue*. *Revenue* is the total amount of revenue collected in the simulation, and *LostRevenue* is the total amount of revenue lost because the customer thought that the line was too long.

Random Number Generator

Every simulation package has a random number generator. The random number generator (technically called a *pseudo-random number* generator) is a software routine that generates a random number between 0 and 1 that is used in sampling random distributions. For example, let us assume that one has determined that a given process delay is uniformly distributed between 10 min and 20 min. Then every time an entity went through that process, the random number generator would generate a number between 0 and 1 and evaluate the uniform distribution formula that has a

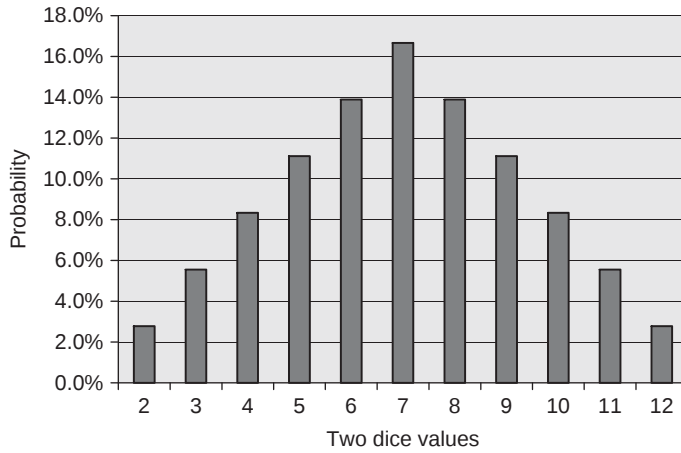


Figure 2. Distribution of rolling two dice.

minimum of 10 and a maximum of 20. As an example, let us assume that the generated random number is 0.7312, then the delay would be $10 + (0.7312) \times (20 - 10) = 17.312$. So, the entity would delay for 17.312 time units in the simulation. Everything that is random in the simulation uses the random number generator as an input to determine values.

In our example model, we will use a physical random number generator, two dice. The probability distribution for the roll of two dice is seen in Fig. 2. In our model, the roll of two dice becomes the basis for every random delay and randomly assigned value in the model. We have five randomly assigned delays and values in the model: (i) the time between arrivals of cars to the restaurant, (ii) the value of the order for the car, (iii) the delay at the menu board, (iv) the delay at the kitchen, and (v) the delay at the order pickup window. The formulas for these random values are as follows:

1. Time between arrivals of cars to the restaurant = $(\text{Dice} \times 10)$ (s).
2. The value of the order for the car = $(\text{Dice} \times 2) - 2$ (\$).
3. The delay at the menu board = $(\text{Dice} \times 10)$ (s).
4. The delay at the kitchen = $(\text{Dice} \times 8)$ (s).
5. The delay at the order pickup window = $(\text{Dice} \times 10)$ (s).

The Calendar

The *calendar* for the simulation is a list of events that are scheduled to occur in the future. In every simulation, there is only one calendar of future events and it is ordered by the earliest scheduled time first. In a later example, it will become clearer how the calendar works and why it is important in the simulation. At this point, just remember that, at any given point in time, every event that has already been scheduled to occur in the future is held on the calendar.

System State Variables

Depending on the simulation package, there can be several system state variables, but the system state variable that every simulation package has is the current time of the simulation. In order to keep from offending any simulation vendors, we choose a different name for our simulation time variable. Our name of the current time of the simulation is *CurrentTime*. This variable is updated every time an entity is taken from the calendar.

Statistics Collectors

Statistics collectors are a part of the simulation that collect statistics on certain states (such as the state of a resource), or the value of global variables, or certain performance statistics based on attributes of the entity. There are three different types of statistics

that are collected: *counts*, *time-persistent*, and *tallies*. Counts are very straightforward, they *count*. In our model, we count the number of *Lost Customers* because the line was too long. *Time-persistent* statistical collectors give the time-weighted values of different variables in the simulation. A common variable to track is the utilization of a resource. In our model, we collect six different time-persistent statistics, which are the number of busy resources of the three resources that we have in the model and the number of entities in each of the three queues that serve the three resources. *Tally statistics* are collected one observation at a time without regard to the amount of time between observations. In our model, we collect a very common statistic, which is the amount of time that an entity stays in the system. Since we assign the value of *StartTime* to the attribute of the entity when it enters the restaurant parking lot, then the value $CurrentTime - StartTime$ is the total amount of time in the system. If we are to improve this system, we would want to minimize this statistic without hurting the revenue too much.

A WALK THROUGH THESE CONCEPTS

What we are going to do now is walk through a couple of steps in our simulation model. As one can see later, we track the state of the system, the entities on the calendar, the

values of attributes and state variables, and the statistics we are collecting.

The State of the System at Noon

Figure 3 shows the state of the system at noon during a lunch rush at our fast-food restaurant. To help us design the line at the restaurant, we have set two of our global variables, *MenuBoardLineSize* and *Order-PickupLineSize*, both equal to 3. As one can see, it is the total capacity of the line at the restaurant. The state of our system at noon is as follows:

- Car 1 (the Corvette) is at the Order Pickup Window receiving its order.
- Cars 2 (the convertible VW Bug) and 3 (the minivan) are in line waiting for the Order Pickup Window.
- Car 4 (the 2-seat convertible) is ordering at the Menu Board.
- Cars 5 (the Police Car) and 6 (the Porsche) are waiting in line for the Menu Board.
- The Order for Car 2 is being cooked in the Kitchen.
- The Order for Car 3 is waiting in line for the Kitchen.
- Car 7 (of undetermined type) is scheduled to arrive at the restaurant in the future.

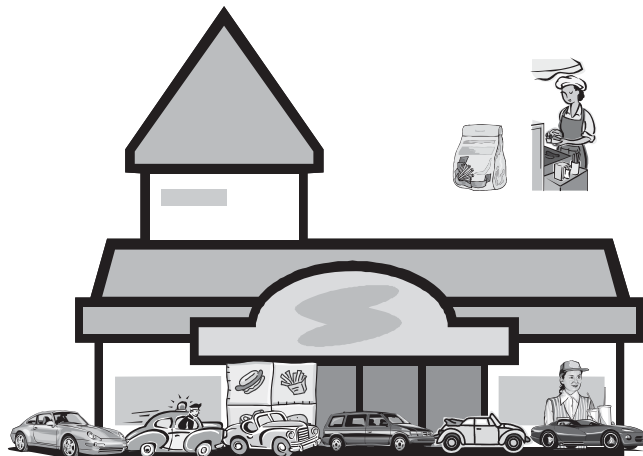


Figure 3. State of the system at noon.

Table 1. The Calendar at Noon

Entity	Event	Event Time
Car(7)	Arrive at restaurant	12:00:20 p.m.
Car(1)	Order pickup window complete	12:00:40 p.m.
Order(2)	Kitchen complete	12:00:56 p.m.
Car(4)	Menu board complete	12:01:10 p.m.

The calendar of this system is made up of the entities that are scheduled to complete an activity with a specific time duration. These entities are Car 1, Car 4, the Order for Car 2, and Car 7. The calendar for these four entities is shown in Table 1.

We also know the values of the attributes of the entities in the system. As you recall, we have two attributes, *StartTime* and *OrderValue*. The values for those attributes for each of the entities in the system is outlined in Table 2.

Other important information concerns our system state variable, *CurrentTime*, which is set to 12:00:00 p.m.

The statistics that we are tracking in the simulation have the values listed in Table 3 as of noon. The “Time/Obs” column gives the amount of time that we have been collecting the statistic (which is since 11:00 a.m.) for time-persistent statistics or the number of observations for tally statistics.

The State of the System at 12:00:20 p.m.

Here is where we get one of the key ideas about a *discrete-event simulation*. A discrete-event simulation moves the current time of the simulation through events and their timing instead of distinct time intervals. If we had distinct time intervals of 1 s, we would go through 20 time intervals before anything would happen. So instead

Table 2. Entity Attributes at Noon

Entity	<i>StartTime</i>	<i>OrderValue</i> (\$)
Car(1)	11:54:20 a.m.	10
Car(2)	11:55:50 a.m.	6
Car(3)	11:57:10 a.m.	4
Car(4)	11:58:20 a.m.	14
Car(5)	11:59:30 a.m.	14
Car(6)	12:00:00 p.m.	10
Car(7)	—	—

Table 3. Statistics at Noon

Statistic	Value	Time/Obs
Revenue	\$504	—
Lost revenue	\$74	—
MenuBoard utilization	0.9984	1:00:00
Kitchen utilization	0.7006	1:00:00
OrderWindow utilization	0.9678	1:00:00
MenuBoard waiting in line	1.2306	1:00:00
Kitchen waiting in line	0.0822	1:00:00
OrderWindow waiting in line	1.0311	1:00:00
Time in system	5.5969	44

of going through 20 time intervals with nothing happening, we go straight to the next scheduled event, which is the first event on the calendar, which is scheduled to occur at 12:00:20 p.m. That event is the arrival of Car 7 to the restaurant.

When Car 7 arrives at the restaurant, the first activity in the simulation is to set its attributes. Obviously, *StartTime* is set to 12:00:20 p.m., and the *OrderValue* is set using the formula $(\text{Dice} \times 2) - 2$ after we roll the dice. The roll of the dice gives us a 9, so the value of the order is \$16. After the attributes are assigned, Car 7 makes a decision whether to get in line to order. Since we have set the *MenuBoardLineSize* = 3 and there are 3 cars in the line, Car 7 leaves as a dissatisfied customer. When Car 7 leaves, we want to capture some important information, namely how much revenue we have lost because of dissatisfied customers, which is tracked in our global variable *RevenueLost*. *RevenueLost* is incremented from \$74 to \$90.

Now that Car 7 has arrived at (and subsequently, left) the restaurant, we also need to schedule the arrival of the next car (Car 8) to the restaurant. The time between arrivals to the restaurant is set using the formula $(\text{Dice} \times 10)$ in seconds. So, we roll the dice and get a 7. Car 8 will arrive 70 s in the future at time 12:01:30 p.m. Car 8 is now put on the calendar with the event “Arrive at Restaurant” that will occur at the event time of 12:01:30 p.m.

So what has changed? Car 7 has arrived and left and Car 8 is scheduled to arrive in the future. We have also updated *RevenueLost*. We have a new calendar, which is shown in Table 4.

Table 4. The Calendar at 12:00:20 p.m

Entity	Event	Event Time
Car(1)	Order pickup window complete	12:00:40 p.m.
Order(2)	Kitchen complete	12:00:56 p.m.
Car(4)	Menu board complete	12:01:10 p.m.
Car(8)	Arrive at restaurant	12:01:30 p.m.

Table 5. Statistics at 12:00:20 PM

Statistic	Value	Time/Obs
Revenue (\$)	504	—
Lost revenue (\$)	90	—
MenuBoard utilization	0.998409	1:00:20
Kitchen utilization	0.702254	1:00:20
OrderWindow utilization	0.967978	1:00:20
MenuBoard waiting in line	1.234851	1:00:20
Kitchen waiting in line	0.087271	1:00:20
OrderWindow waiting in line	1.036453	1:00:20
Time in system	5.5969	44

The statistics can be updated at this point as well. (However, most simulation packages only update statistics when the value that is being tracked changes.) We have two types of statistics, the counting of *Revenue* and *RevenueLost*, the resource utilization statistics and the queue length statistics. The counting statistics is simply the values of the variables we are tracking, and *RevenueLost* has gone to \$90. The other statistics is time-dependent statistics that are time-weighted averages of a given value. As an example, let us calculate the new value of the average number of cars waiting in line for the menu board. At noon, the simulation had been running for 1 h and the average value was 1.2306. From noon to 12:00:20 p.m., the number of cars waiting in line for the menu board has been 2. So the new time-weighted average is $[(1.2306 \times 1:00:00) + (2 \times 0:00:20)]/1:00:20 = 1.234851$. If we were to convert this formula to seconds, it would be $[(1.2306 \times 3600) + (2 \times 20)]/3620 = 1.234851$. All time-dependent statistics are calculated in this way. The value for all of the statistics at time 12:00:20 p.m. is in Table 5.

The State of the System at 12:00:40 p.m.

We want to take one more step in the simulation because we actually see the line move!

Looking at Table 4, we know that the next scheduled event is that Car 1 finishes its time at the Order Pickup Window at time 12:00:40 p.m. The first thing to do is set the *CurrentTime* to 12:00:40 p.m. When Car 1 finishes its time at the Order Pickup Window, we know that Car 1 has paid for its order, and so *Revenue* must be increased by the *OrderValue* attribute of Car 1. So the value *Revenue* is increased from \$504 to \$514. We also need to calculate a new average *Time in System* statistic. The time in the system for Car 1 is the *CurrentTime StartTime*, which would be 12:00:40 p.m. – 11:54:20 a.m., which is 6 min and 20 s. Since we are collecting *Time in System* in minutes, the new average *Time in System* is $[(5.5969 \times 44) + 6.3333]/45 = 5.613265$. The other thing that occurs is that the *Order Pickup Window* resource is no longer utilized. Since the *Order Pickup Window* resources is available for other cars, the simulation will allocate the *Order Pickup Window* resource to the car that is first in the line that is waiting to use the *Order Pickup Window*. Our Car 2 (the convertible VW Bug) is first in line for the *Order Pickup Window*, so Car 2 is allocated the *Order Pickup Window* resource. Can Car 2 start picking up its order? No, it cannot. The reason is that the order for Car 2 is still in the Kitchen. So even though the *Order Pickup Window* resource is occupied, no productive work is going on. At the next step in the simulation, the order for Car 2 will complete in the Kitchen and Car 2 will be able to start the process of picking up its order. But for now, it just has to wait!

Since Car 2 has moved forward, Car 3 moves to first place in the line waiting for the *Order Pickup Window* resource. Car 4 is still occupies the *Menu Board* resource and is still in the process of giving its order. This process is scheduled to end at time 12:01:10 p.m. (See Table 4). Cars 5 and 6 do no change their positions in the line and Car 8 is still scheduled to arrive at the restaurant at time 12:01:30 p.m. The new state of the system is shown in Fig. 4.

We also update our statistics at this time. As we described in at time 12:00:20 p.m., we look at what has happened since the last time we updated the statistics in order to determine the new values for the statistics.

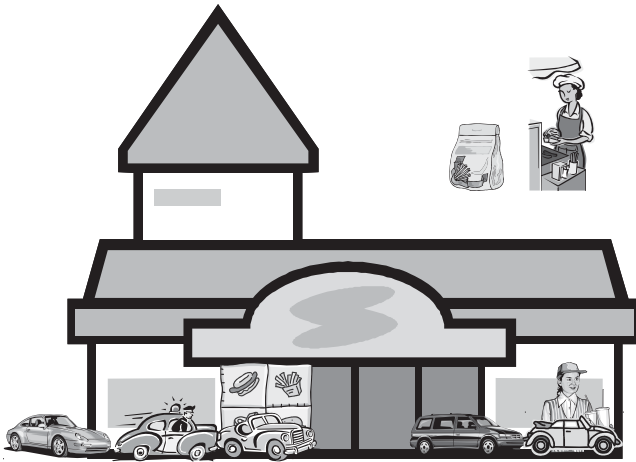


Figure 4. The state of the system at 12:00:40 p.m.

Table 6. Statistics at 12:00:40 p.m

Statistic	Value	Time/Obs
Revenue (\$)	514	—
Lost revenue (\$)	90	—
MenuBoard utilization	0.998418	1:00:40
Kitchen utilization	0.703890	1:00:40
OrderWindow utilization	0.968154	1:00:40
MenuBoard waiting in line	1.239055	1:00:40
Kitchen waiting in line	0.092286	1:00:40
OrderWindow waiting in line	1.041747	1:00:40
Time in system	5.613265	45

So, what number will we use to calculate the new average number of cars waiting for the *Order Window*, 2 (the number in line at 12:00:20) or 1 (the number in line now)? The answer is 2 because from time 12:00:20 p.m. to 12:00:40 p.m. there were 2 cars waiting in line. The new values for the statistics are shown in Table 6.

INTERPRETING OUTPUT STATISTICS

Let us assume that our model has run from 11:00 a.m. to 2:00 p.m. At 2:00 p.m., we stop the simulation and we have all of our statistics calculated. The “answer” for the simulation is in Table 7.

But what does this data really tell us? Is it saying that every day, from 11:00 a.m. to 2:00 p.m., the revenue for the restaurant will be \$1638? Is it saying that every day, from 11:00 a.m. to 2:00 p.m., the average amount

Table 7. Statistics at End of Simulation

Statistic	Value	Time/Obs
Revenue (\$)	638	—
Lost revenue (\$)	170	—
MenuBoard utilization	0.9724	3:00:00
Kitchen utilization	0.7250	3:00:00
OrderWindow utilization	0.9846	3:00:00
MenuBoard waiting in Line	0.8785	3:00:00
Kitchen waiting in line	0.133	3:00:00
OrderWindow waiting in line	1.1567	3:00:00
Time in system	5.0487	140

of time that a car will be in line is 5.0487 min? What is this really saying?

The answer to that question is, “These numbers give us a random answer to the performance of the system.” It is a random answer because there are random inputs to the system that give us this answer. So how can we be sure that the answer is any good?

The answer generated by only one run is not really an answer at all. To make the point, let us take the following 100 rolls of the dice. You would think that 100 rolls of the dice would tell us the average. But if you take the 100 rolls in Table 8, the average is only 6.72. Is that close enough? Would you be willing to bet that the next 100 rolls would come up with an average of 6.72. Probably not. (If you are, please contact me as soon as possible on any other wagers you may be considering.)

So what is the answer to this problem? The answer is that, if we really want to estimate the average value of the roll of the dice if we

Table 8. Rolls of the Dice

6	5	8	6	5	4	10	6	7	8
5	9	8	7	8	6	3	8	7	6
9	9	8	8	6	8	9	7	10	5
2	10	11	8	6	8	7	3	8	8
4	6	8	11	2	4	8	9	8	5
9	3	8	7	2	3	9	10	7	7
3	9	5	7	7	7	9	4	8	7
4	10	7	4	10	8	4	8	9	7
3	6	6	3	6	3	10	9	7	4
6	8	5	9	12	6	8	6	4	2

Average: 6.72.

Table 9. 30 Iterations of Throwing Dice 100 Times

6.72	6.95	6.78	7.14	6.62	6.81
6.75	7.17	6.62	7.3	6.92	7.04
6.79	7.13	7.17	7.12	6.82	7.29
7.13	7.26	7.19	6.52	6.8	7.3
6.95	6.96	6.78	7.17	7.13	6.68

Average: 6.967; standard deviation: 0.22992; 95% confidence interval: [6.8847, 7.0493].

roll the dice 100 times, we need to run the simulation more times. Each time that we run a simulation is called an *iteration*. So, as an example, let us say that we have run our dice-throwing simulation for 30 iterations, and Table 9 has the average values for each of those iterations.

If we are simply trying to estimate the average of 100 rolls of the dice, we do not need to worry about the standard deviation for each iteration. However, we do need to worry about the standard deviation of the averages from each iteration. As is shown in Table 9, the average of the averages (so to speak) is 6.967. The standard deviation of those 30 averages is 0.22992. With that information, we can calculate a *confidence interval*.

A confidence interval is a statistical measure where we want to bound some statistic. The level of confidence (95% in our example) is the statistical probability that the statistic that we are considering lies in the interval. So, the interpretation of the 95% confidence interval of [6.8847, 7.0493] would be “There is a 95% chance that the true average of rolling the dice 100 times lies between 6.8847

Table 10. Confidence Intervals (CIs) after 30 Iterations

Statistic	Value	CI Lower	CI Upper
Revenue (\$)	1649	1627	1670
Lost revenue (\$)	155	135	176
Lost customers	12.83	10.99	14.67
MenuBoard utilization	0.97	0.96	0.98
Kitchen utilization	0.73	0.72	0.74
OrderWindow utilization	0.99	0.99	0.99
MenuBoard waiting in line	0.95	0.89	1.01
Kitchen waiting in line	0.12	0.11	0.13
OrderWindow waiting in line	1.23	1.18	1.28
Time in system	5.31	5.18	5.44

and 7.0493.” Most statistical and simulation packages automatically calculate confidence intervals for you. Even Microsoft Excel has a function to calculate confidence intervals.

So, in our example, we want to run 30 iterations so that we can have good confidence intervals for each of our statistics. (Although we will not get into the topic in this article, one should run 30 iterations (or more if you can) in order to get good confidence intervals.) Table 10 shows the confidence intervals for each of our statistics.

These statistics are very interesting on several fronts. First, we could improve our revenue every day at lunch 9.4% if we could capture the lost revenue. We need to find a way to lower the number of lost customers from the 12.83 that we have seen in the current system. We also know that both the *Menu Board* and *Order Window* are highly utilized, which means that lines are forming in front of those two resources. We would also want to look at reducing the car’s time in the system. The minimum average theoretical time for a car would be 70 s at the *Menu Board* and 70 s at the *Order Window*, which is 140 s, or 2.33 min. So, about 3 min of the car’s time is simply waiting. A key customer satisfaction metric is to minimize the amount of time they have to wait in line.

FINDING WAYS TO IMPROVE A SYSTEM

What we have accomplished to this point is the analysis of a system that exists. We have

learned that we can improve revenue and that we have very busy resources. So let us run some alternative scenarios to determine what (if any) improvements we should make.

Scenario 1: No Lines

Scenario 1 is based on the “eliminating inventory” thought. If we want to get the customer out of the system quicker, we would simply need to eliminate some (or all) of the line. We will take drastic action and eliminate all waiting in the system. If a car cannot pull right up to the *Menu Board*, then it will leave. If the *Order Window* is not available after the order is placed, then the car waits at the *Menu Board* until the *Order Window* is free. This should greatly lower the *Time in System* statistic. Some would say that this would be the best way to run the drive-through line since the system is *balanced*. The arrival rate is 1 every 70 s, the *Menu Board* rate is 1 every 70 s, and the *Order Window* rate is 1 every 70 s. Some would say that this is a perfect production line.

In order to accomplish this in the model, we simply change two variables, *MenuBoardLineSize* and *OrderPickupLineSize*, and set them both to a value of 1. Then we run the model for 30 iterations and get the statistics in Table 11.

Well, the strategy worked the way we expected it to work. We reduced the *Time in System* statistic by 37.7%. Now the cars are whipping through the drive-through. But there is a catch. We are losing 5.5 times

the number of customers that we were losing before! We have lowered our revenues by 40%! This would hardly be considered a viable alternative!

What was it about eliminating the lines that caused such a problem? Without the lines, we caused excessive *blocking* in the system. Blocking occurs when something in the system will not allow a resource to become available even though its service is finished. Throughout this scenario, the car at the *Menu Board* was finished ordering, but the car could not move forward because the *Order Window* was still occupied. In this scenario (or in any balanced system with random variation and no queueing space), each car at the *Menu Board* has a 50% chance of being blocked by the *Order Window*. This can be reduced by adding queues between the resources.

In industrial problems, it is often difficult to determine where and why blocking occurs. That is why there is so much emphasis on eliminating and/or controlling the *bottleneck* of the system. It is likely that the bottleneck is responsible for blocking in other parts of the system.

Scenario 2: Longer Lines at the Menu Board

Well, since Scenario 1 did not work very well, let us try another. Let us rearrange the parking lot so that we can get room for 6 cars waiting in line for the *Menu Board* (including the car whose order is being taken). We will keep the line for three cars for the *Order Pickup Window*. The results for Scenario 2 is shown in Table 12.

Well, we did add revenue (\$51 per day) and decrease the number of lost customers, that is good; but now the average customer waits an additional minute to get the order, which is bad. You can see why this occurs. With the *Menu Board* line being longer, more cars can wait for the menu board and they do not become lost customers in this scenario, where they would have become lost customers in the original model. The price they pay for having room to get in line is that the line is longer, and so the time through the system increases. This is another typical trade-off. Queues are good because they allow more throughput in

Table 11. Scenario 1 Statistics

Statistic	Value	CI Lower	CI Upper
Revenue (\$)	989	968	010
Lost revenue (\$)	851	827	875
Lost customers	70.43	68.87	71.99
MenuBoard utilization	0.68	0.67	0.69
Kitchen utilization	0.43	0.42	0.44
OrderWindow utilization	0.84	0.83	0.85
MenuBoard waiting in line	0.00	0.00	0.00
Kitchen waiting in line	0.00	0.00	0.00
OrderWindow waiting in line	0.00	0.00	0.00
Time in system	3.31	3.28	3.34

Table 12. Scenario 2 Statistics

Statistic	Value	CI Lower	CI Upper
Revenue (\$)	1690	1665	1715
Lost revenue (\$)	99	77	122
Lost customers	8.50	6.73	10.27
MenuBoard utilization	0.99	0.99	0.99
Kitchen utilization	0.73	0.72	0.74
OrderWindow utilization	0.99	0.99	0.99
MenuBoard waiting in line	3.07	2.90	3.24
Kitchen waiting in line	0.13	0.12	0.14
OrderWindow waiting in line	1.28	1.23	1.33
Time in system	6.33	6.14	6.52

a system, but they add cycle time for the system.

Scenario 3: Improved Service Times

Let us try one more scenario. This scenario has found new technology that will cut the average service time at the Menu Board and the Order Window by 20% from 70 s to 56 s. To implement this technology, it would take \$30,000 at each store and must be paid for by increased revenue at the store. This is the only thing that changes from the original model. The results are shown in Table 13.

Well, now we see some improvement! The *Time in System* has dropped from 5.31 min to 3.42 min. We have virtually eliminated any lost customers, and have increased revenue by \$174 per lunch shift. We would pay back

Table 13. Scenario 3 Statistics

Statistic	Value	CI Lower	CI Upper
Revenue (\$)	1822	1794	1850
Lost revenue (\$)	10	5	15
Lost customers	0.70	0.33	1.07
MenuBoard utilization	0.81	0.80	0.82
Kitchen utilization	0.79	0.78	0.80
OrderWindow utilization	0.97	0.97	0.97
MenuBoard waiting in line	0.26	0.23	0.29
Kitchen waiting in line	0.23	0.21	0.25
OrderWindow waiting in line	0.85	0.82	0.88
Time in system	3.42	3.36	3.48

the \$30,000 for the implementation of the new technology in 172 days or less than 6 months! It is well worth the investment and you are a hero for figuring it out!

Why did this scenario show such improvement? It was because our resources were not so highly loaded. The *Menu Board* utilization dropped from 97 to 81%, and so the line for the Menu Board would be much smaller (it dropped by 72%). We would have seen a similar drop in the *Order Window* utilization, but now the Order Window is being often being blocked by the Kitchen.

Generating Other Scenarios

This simulation and any other simulation can be used to evaluate many different scenarios if the person creating the model allows some flexibility in the model structure. One also has to consider the amount of time it takes to run a new scenario. In a large-scale simulation, to run and evaluate a new scenario could take several days.

WHAT HAVE WE LEARNED?

This exercise points out several things about simulation in general. First, simulation can mimic the *dynamic* behavior of a system. That is what it is built to do. Regardless of how complex a system may be, it is likely that a simulation *expert* will be able to create a model that will evaluate it. However, the more complex a system is, the longer it takes to model, run, and evaluate. But do not be discouraged, there are very good simulation people available to model large systems.

Second, you (or the person analyzing the system) must have a good understanding of simulation statistics. It is important during the creation of the model so that input distributions are used properly. It is important during the analysis of the output statistics so that the output is not misinterpreted. Mistakes with either the inputs or the outputs will cause the simulation analysis to be invalid.

Third, to analyze a system, simulation is used to evaluate different scenarios. It does not choose the best scenario for you. This may seem to be a problem, but most managers

have no shortage of scenarios to evaluate. The trade-off for this is that you can analyze the *dynamics* of the system and not just the average behavior.

Fourth, the scenarios that you do choose are generated by you and not the system. This is where familiarity with the system under study and a familiarity with system dynamics concepts are very valuable.

Discrete-event simulation is a complex and powerful tool that you can be used effectively to analyze large systems.

REFERENCES

1. Shannon RE. Systems simulation—The art and science. New Jersey: Prentice-Hall; 1975.
2. Law AM, Kelton WD. Simulation modeling and analysis. 3rd ed. New York: McGraw-Hill; 2000.
3. Banks J, Carson JS II, Nelson BL, *et al.* Discrete event system simulation. 3rd ed. New Jersey: Prentice-Hall; 2000.
4. Kelton WD, Sadowski R, Sadowski D. Simulation with Arena. 2nd ed. New York: McGraw-Hill; 2001.

INTRODUCTION TO FACILITY LOCATION

LAWRENCE V. SNYDER
Department of Industrial and
Systems Engineering,
Lehigh University,
Bethlehem, Pennsylvania

Facility location problems involve choosing locations for facilities in order to serve customers while achieving some balance between cost and service. If we open too few facilities, customers will tend to be far from their nearest facility, while if we open too many, infrastructure costs will be large. The facilities may be factories, warehouses, distribution centers, and so on, while customers may represent actual consumers or downstream facilities such as retailers. Many location problems involve two sets of decisions: where to locate, and which customers should be assigned to which facilities. They are therefore sometimes known as *location-allocation problems*.

Facility location is one of the oldest classes of operations research (OR) problems, and the field reflects OR's roots as a multidisciplinary endeavor. Facility location problems are studied by academics and practitioners from economics, mathematics, computer science, geography, and marketing, as well as from OR and its siblings such as industrial engineering and management science. Articles on the subject range in tone from highly theoretical mathematical and algorithmic investigations to highly applied, practical case studies.

Facility location models have been widely applied in both public and private sectors to choose locations for such facilities as emergency medical services (EMS) [1], fire stations [2], airline hubs [3], blood banks [4], fast-food restaurants [5], public swimming pools [6], schools [7], vehicle inspection stations [8], and bus stops [9]. (See Current *et al.* [10] for additional examples). They have been used to locate entities that one would not

normally consider to be facilities, such as wildlife reserves [11], satellite orbits [12], and bank accounts [13]. They have also been applied to problems in which no physical entities are to be located, but the facilities and customers are a modeling construct in order to make other types of decisions, such as tool selection in flexible manufacturing systems [14], platform positioning by political parties [15], and screening tests for cervical cancer [16]. They also sometimes arise as subproblems for other OR problems such as vehicle routing [17].

Since most classical facility location problems are simple to describe and formulate, yet difficult (NP-hard) to solve, they have been a "proving ground" for many optimization algorithms. Nearly every algorithmic method that has been developed for integer programming (IP) problems has, at one time or another, been applied to facility location. These include exact approaches such as branch and bound, branch and cut, and relaxation and decomposition methods, as well as metaheuristics such as genetic algorithms and tabu search.

This article provides a broad overview of the field of facility location (also known as *location theory*, or *location analysis*). The article is not meant to be a complete review of the literature; indeed, such a review would be virtually impossible given the huge number of articles, books, and other references that have been written about the topic. (The online bibliography by Hale [18] cites over 3400 scholarly works.) Many other reviews and surveys of facility location have been written; see Table 1 for a partial list. Textbooks on location theory include those by Handler and Mirchandani [19], Hurter and Martinich [20], Mirchandani and Francis [21], Daskin [22], Drezner [23], Drezner and Hamacher [24], Eiselt and Sandblom [25], and Eiselt and Marianov [26]. Professional society subdivisions devoted to facility location include the INFORMS Section on Location Analysis (SOLA; [27]) and the EURO Working Group on Locational Analysis (EWGLA; [28]).

Table 1. Partial List of Reviews and Surveys of the Facility Location Literature

Review Type	References
Broad-based reviews	[29–35]
p -Median and p -center problems	[36,37]
Algorithms for p -median problem	[38,39]
Weber problem	[40]
Covering problems	[21,41]
Obnoxious location and dispersion problems	[42,43]
Hierarchical (multiechelon) problems	[44,45]
Competitive location problems	[46,47]
Hub location problems	[48,49]
Problems with congested facilities	[50,51]
Dynamic and stochastic location problems	[52]
Stochastic location problems	[53,54]
Problems with disrupted facilities	[55–57]
Integrated problems	[58–61]
Public sector problems	[62]
Application to supply chain management	[60,63]

The section titled “Taxonomy” discusses several dimensions along which location models are often classified. Six classical models, including their mathematical programming formulations and solution methods, are described in the section titled “Classical Models.” The section titled “Extensions” discusses some of the subfields that extend and expand the classical models to richer settings. In general, the focus is on discrete and network location models (see the section titled “Topology”) throughout the article since they are the most commonly studied models.

TAXONOMY

This section discusses some of the most common dimensions along which facility location models are classified.

Topology

Perhaps the most important question in describing a facility location problem is, where may the facilities be located? The answer has important implications on how the problem is formulated, analyzed, and solved. Most problems fall into one of

the following three categories: continuous, network, or discrete problems.

In *continuous* location models, facilities may be located anywhere on the plane (or on some other continuous region). The earliest location problems were of this type. For example, the Weber problem [64] involves choosing the location of a single facility to minimize the total Euclidean distance between the location and a set of fixed customer locations (see the section titled “Classical Models”). A special case of continuous models, known as *analytical* models, assumes that the customers are distributed continuously over some region. (In contrast, in the Weber problem and most other continuous models, the decision space for the facility locations is continuous, but the customers are located at fixed, discrete points.) The goal is generally to derive the optimal number of facilities in closed form. Analytical models are meant to provide insights rather than implementable solutions [34,65].

In *network* problems, we are given a graph or network with customers located on the nodes. Facilities may be located at nodes or anywhere along the arcs, and travel between customers and facilities occurs along the arcs. Specialized problems involving simple network topologies (such as trees) or a restricted number of facilities (1 or 2) may often be solved in polynomial time [66]. In general, network problems are formulated and solved using IP techniques.

A key property of network problems is the *Hakimi property* [67], which states that an optimal solution exists in which the facilities are all located at the nodes (rather than along the arcs of the network). This property reduces an infinite solution space to a finite (though combinatorially large) one. Unfortunately, the Hakimi property does not hold for all network problems; for example, it holds for the p -median problem on a network but not for the p -center problem (see the section titled “Classical Models”). When it does not hold, it is sometimes possible to guarantee that the optimal locations will be part of an augmented node set in which dummy nodes have been added along the arcs [68].

In *discrete* location models, we are given a discrete set of points at which the facilities

may be located. These problems are usually formulated and solved using IP techniques. Network models can be considered as a special case of discrete models if the Hakimi property applies (to the original node set or an augmented one); the distance between a given facility and customer is simply set to be the network distance between those two nodes.

In this article, the primary focus is on network and discrete models since they are the most extensively studied and the most commonly extended models.

Distance Measures

Most location models can accommodate a variety of distance measures, although many algorithms depend critically on the use of a specific measure, or on properties such as symmetry or the triangle inequality.

In many cases, travel times or costs are required, rather than distances. These can sometimes be obtained by multiplying the distance by a constant, or one might assume that the times or costs are specified explicitly in a matrix. In general, the terms *distance*, *travel time*, and *transportation cost* are used interchangeably here.

Several common distance measures are discussed below.

Euclidean. If points are located in the plane, the familiar Euclidean distance is a natural choice for the distance measure: the distance between points (x_1, y_1) and (x_2, y_2) is given by

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

Rectilinear. This measure assumes that travel is possible only along lines that are parallel to one of the two axes. It is sometimes known as the *Manhattan metric* since it can represent travel in a city with a grid of perpendicular streets. The distance between (x_1, y_1) and (x_2, y_2) is given by

$$|x_1 - x_2| + |y_1 - y_2|. \quad (2)$$

Although this measure accurately reflects the distance in some situations, the reason for

using it is usually that it is mathematically more convenient than the Euclidean distance since it is piecewise linear rather than non-linear.

ℓ_p Norm. For a fixed constant $p \geq 1$, the ℓ_p distance between (x_1, y_1) and (x_2, y_2) is given by

$$(|x_1 - x_2|^p + |y_1 - y_2|^p)^{1/p}. \quad (3)$$

By setting $p = 1$ and $p = 2$, we get the rectilinear and Euclidean measures, respectively, as special cases. For general p , this norm is of more mathematical than practical interest.

Great Circle. All the measures discussed above assume that the points are located on the plane and do not represent geographical locations. For example, if locations are specified using latitude and longitude, the Euclidean distance formula does not apply. Instead, one can use great-circle distances, whose formula correctly interprets coordinates given as latitude and longitude and also accounts for the curvature of the Earth, assuming that travel occurs over a great circle (the shortest route over the surface of a sphere). Let (α_1, β_1) and (α_2, β_2) be the latitude and longitude of two points and let $\Delta\alpha$ and $\Delta\beta$ be the differences in the latitude and longitude (respectively). Then the great-circle distance between them is given by

$$2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\Delta\alpha}{2} \right) + \cos \alpha_1 \cos \alpha_2 \sin^2 \left(\frac{\Delta\beta}{2} \right)} \right), \quad (4)$$

where r is the radius of the Earth, approximately 3958.76 miles or 6371.01 km (on average).

Highway. Highway distances reflect the true travel distance between two geographical points after solving a shortest-path problem on the road network that connects them. These distances can be obtained from web sites such as Mapquest or Google Maps, or from many geographic information systems (GIS). For most practical instances, this is the distance measure of choice.

Matrix. In some cases, the distance matrix, which lists the distance between each pair of points, is simply given explicitly. This is the most general measure: all of the measures listed above are special cases. Distance matrices are useful when the distance does not follow a predictable pattern—for example, if it represents freight rates or airline ticket prices. In the remainder of this article, it is assumed that this matrix is given and the details about how the distances were calculated are ignored.

Aggregate Distance Function

Another important element of a location model is the way in which it calculates an aggregate measure to summarize the distances between customers and their assigned facilities. The three most common aggregate distance functions are total distance, maximum distance, and coverage. Any of these can occur in either the objective function or the constraints, although the first two are most commonly found in the objective function.

In *total distance* models, the aggregate distance is calculated by simply summing the distances over all the customers; that is,

$$\sum_{\text{customers } i} \text{distance from } i \text{ to assigned facility.} \quad (5)$$

Often, the distances are weighted based on the demand of the customer or some other weighting factor:

$$\sum_{\text{customers } i} \text{weight}_i \cdot \text{distance from } i \text{ to assigned facility.} \quad (6)$$

Problems featuring this aggregate distance function are sometimes known as *minisum problems*. Examples include the *p*-median problem and the uncapacitated fixed-charge location problem (UFLP) (see the section titled “Classical Models”).

In *maximum distance* models, the aggregate distance is set equal to the maximum distance from a customer to its assigned facility:

$$\max_{\text{customers } i} \text{distance from } i \text{ to assigned facility.} \quad (7)$$

(This measure, too, has a weighted version.) This function is used to ensure some degree of service equity among the customers. Problems featuring this aggregate distance function are sometimes known as *minimax problems*. Examples include the *p*-center problem.

Finally, *coverage* models assume that a customer is “covered” if it is within a pre-specified radius of its assigned facility. Such models may require that all customers be covered by an open facility (as in the set covering location problem [SCLP]) or may cover as many (weighted or unweighted) customers as possible using a fixed number of facilities (as in the maximal covering location problem [MCLP]).

Limiting the Number of Facilities

Most aggregate distance functions encourage facilities to be close to customers, and this leads to a tendency to open many (or all) of the available facilities. To counteract this tendency, most facility location models include some mechanism for preventing too many facilities from being open. The two most common approaches are to use a fixed cost or to place an explicit restriction on the number of open facilities.

Fixed costs apply to each open facility and are independent of the volume of demand assigned to the facility. They represent overhead costs associated with the facility, such as real estate-related costs (construction, mortgage, or rent costs), heating and cooling, and some personnel costs. This is the approach taken by the UFLP (see the section titled “Classical Models”). Importantly, these costs must be expressed in the same time units as the demands. For example, if a facility costs \$24 million to build and is expected to be in use for 10 years, and if monthly demands are used as an input to the model, then the fixed cost should be given as \$200,000 ($=24,000,000/(10 \times 12)$) per month.

The other common way to limit the number of facilities is simply to include a *cardinality constraint* requiring at most *p* facilities to

be opened, for a constant integer $p \geq 1$. This is the approach taken by the p -median and p -center problems.

Some models use a hybrid approach consisting of a budget constraint restricting the total fixed cost of the opened facilities, rather than their cardinality.

CLASSICAL MODELS

Combining the various options in the dimensions listed above, a wide range of “fundamental” models can be obtained. From these, an even greater array of extensions and variants can be formulated. This section discusses six classical models that have served as the starting point for the majority of models that have been developed subsequently. Mathematical formulations and a brief discussion of the algorithmic approaches used are provided. The first model, the Weber problem, is a continuous model, while the other models are discrete or network models. The first three models are total distance models, the fourth is a maximum distance model, and the last two are coverage models.

The following notation is used throughout this section and the next. (Not all notation is used by every model.)

SETS

- I set of customers
- J set of potential facility sites

PARAMETERS

- h_i demand (items per unit time) at customer i , $\forall i \in I$
- c_{ij} cost to ship one item from facility j to customer i , $\forall i \in I, j \in J$
- f_j fixed cost per unit time to open facility j , $\forall j \in J$
- p maximum number of facilities to open

DECISION VARIABLES

- x_j 1 if facility j is opened, 0 otherwise, $\forall j \in J$
- y_{ij} 1 if facility j serves customer i , 0 otherwise, $\forall i \in I, j \in J$

The Weber Problem

Background and Formulation. In the Weber problem, we have to locate a single facility anywhere in a region Ω in order to minimize the total demand-weighted Euclidean distance from the facility to a set of fixed customer locations, also located in Ω . Since this is a continuous problem, we replace the notation c_{ij} with a function $c_i(x, y)$ that represents the distance from customer i to the point (x, y) , as given by Equation (1). The Weber problem is then formulated as follows:

$$(\text{WEBER}) \quad \text{minimize} \quad \left\{ \sum_{i \in I} h_i c_i(x, y) \right\} \quad (8a)$$

$$\text{subject to} \quad (x, y) \in \Omega \quad (8b)$$

This is a generalization of a 3-“customer” problem that is usually attributed to the seventeenth-century mathematician Pierre de Fermat.

Solution Methods. The Weber problem can be solved using a variety of iterative methods, most notably the so-called Weiszfeld procedure [69], or using standard nonlinear programming techniques. Special cases of the problem can be solved using geometric constructions such as Torricelli points or Simpson lines, and even by physical devices such as the Varignon frame [40].

Extensions include alternate distance metrics [70], multiple facilities [71], forbidden regions [72], travel barriers [73], and the location of dimensional (nonpoint) facilities such as lines [74]. For a more comprehensive discussion of the Weber problem and its variants, refer to Drezner *et al.* [40].

The p -Median Problem

Background and Formulation. The objective of the p -median problem (sometimes known as the k -median, m -median, etc.) is to locate p facilities so as to minimize the total demand-weighted transportation cost between customers and their assigned facilities. The most common formulation, due to Revelle and Swain [75], is as follows:

$$(p\text{MP}) \text{ minimize } \sum_{i \in I} \sum_{j \in J} h_i c_{ij} y_{ij} \quad (9a)$$

$$\text{subject to } \sum_{j \in J} y_{ij} = 1 \quad \forall i \in I \quad (9b)$$

$$y_{ij} \leq x_j \quad \forall i \in I, \forall j \in J \quad (9c)$$

$$\sum_{j \in J} x_j \leq p \quad (9d)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (9e)$$

$$y_{ij} \geq 0 \quad \forall i \in I, \forall j \in J. \quad (9f)$$

The objective function (9a) calculates the total demand-weighted transportation cost between customers and their assigned facilities. Constraints (9b) require each customer to be assigned to exactly one facility, and that facility must be open by constraints (9c). Constraint (9d) requires at most p facilities to be open. (At optimality, exactly p facilities will be open in most instances.) Constraints (9e) require the x variables to be binary, while (9f) require the y variables to be nonnegative. Although the y variables were defined to be binary, if we relax them to continuous variables, there always exists an optimal solution in which they are binary.

Solution Methods. The p -median problem is polynomially solvable for fixed p but is NP-hard for variable p [76]. Among the class of NP-hard problems, p -median is considered a relatively “easy” problem since it can be solved to optimality relatively quickly. One reason for this is that the LP relaxation is often extremely tight and in fact often produces integer solutions “for free” [75]. Nevertheless, a huge range of solution methods—both heuristic and exact—have been proposed for the p -median problem. Some of these are summarized next.

Heuristics. Perhaps the earliest and simplest heuristic for the p -median problem is the *greedy-add* heuristic [77], in which we start with all facilities closed and open the single facility that results in the smallest cost; then force that facility open and find a second facility that will reduce the objective function the most; then repeat until p facilities have been opened.

The *neighborhood search* heuristic [78] begins with any feasible solution and iteratively repeats two main steps. In the first, we find the 1-median (a trivial problem) in the “neighborhood” of each facility, defined as the set of customers the facility serves. If any facility is not equal to the 1-median in its neighborhood, we replace it with the 1-median, guaranteeing a decrease in the objective function. If any facilities have been replaced in this manner, we proceed to the second step, in which we reassign each customer to its nearest open facility, creating new neighborhoods. We repeat this process until a given iteration makes no modifications to the solution.

Another well-known improvement heuristic is the *interchange* (or *exchange*) heuristic [79], which searches for pairs of facilities—one open, one closed—such that the objective function would improve if we “swap” the two facilities—that is, open the closed facility and close the open one. We repeat this until no improving swaps are found. Variants of this heuristic have been applied to a wide range of facility location problems, not just to the p -median. If applied directly, this heuristic is more computationally burdensome than the greedy-add heuristic since it relies on evaluating $O(|J|^2)$ swaps at each iteration, while greedy-add evaluates $O(|J|)$ moves. However, the *fast interchange* heuristic of Whitaker [80] (later refined by Hansen and Mladenović [81]) implements the same idea in a much more efficient manner so that the moves have similar complexity to those in greedy-add.

Since the mid-1990s, there has been an explosion in the development of *metaheuristics*—heuristics that are designed to be general enough to apply to many combinatorial optimization problems with little customization required. A great many of these have been applied to the p -median problem; these include tabu search [82], genetic algorithms [83], variable neighborhood search [81], simulated annealing [84], heuristic concentration [85], ant colony optimization [86], and neural networks [87]. For a review of metaheuristics applied to the p -median problem, see Mladenović *et al.* [39].

Exact Algorithms. The p -median problem can be solved using straightforward branch and bound methods because of its tight LP relaxation [75,88]. A host of other exact algorithmic approaches have been introduced for the problem, many of which were preferable to branch and bound when those methods were first developed since LP codes at that time were significantly less efficient than they are now. Moreover, some of these algorithms are more easily extended to problem variants (e.g., nonlinear objectives) compared to LP-based methods. Classical optimization methods that have been applied to the p -median problem include branch and bound [75], branch and cut [89], dual-based approaches [90], LP decomposition [91], and graph-theoretic bounds [92]. For further information on the range of techniques that have been proposed, see the annotated bibliography by Reese [38].

The most popular exact algorithm for solving the p -median problem (and many of its variants) seems to be *Lagrangian relaxation*, in which a set of constraints is removed and a penalty for violation is inserted into the objective function. The method was first applied to a p -median-like problem by Cornuejols *et al.* [13], although their model also includes fixed costs in the objective function and is therefore a hybrid between the p -median problem and the UFLP (see the section titled “The Uncapacitated Fixed-Charge Location Problem”). Fisher [93] discusses the method for a problem identical to that of Problem (9). The idea is to relax constraints (9b) to obtain the following problem for fixed Lagrange multipliers λ :

$$\begin{aligned} (pMP-LR_\lambda) \quad & \text{minimize} \\ & \sum_{i \in I} \sum_{j \in J} h_i c_{ij} y_{ij} + \sum_{i \in I} \lambda_i \left(1 - \sum_{j \in J} y_{ij} \right) \\ & = \sum_{i \in I} \sum_{j \in J} (h_i c_{ij} - \lambda_i) y_{ij} + \sum_{i \in I} \lambda_i \end{aligned} \quad (10a)$$

subject to

$$y_{ij} \leq x_j \quad \forall i \in I, \forall j \in J \quad (10b)$$

$$\sum_{j \in J} x_j \leq p \quad (10c)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (10d)$$

$$y_{ij} \geq 0 \quad \forall i \in I, \forall j \in J. \quad (10e)$$

This problem is trivial to solve. If we open facility j , we would set $y_{ij} = 1$ if $h_i c_{ij} - \lambda_i < 0$ and to 0 otherwise. Therefore, the “benefit” from opening facility j is given by

$$\beta_j = \sum_{i \in I} \min\{0, h_i c_{ij} - \lambda_i\}. \quad (11)$$

To solve $(pMP-LR_\lambda)$, we open the p facilities with the smallest values of β_j and assign customers to facilities based on the sign of $h_i c_{ij} - \lambda_i$. Of course, this solution will almost certainly not be feasible for (pMP) since some customers will be assigned to multiple facilities and some not at all. However, for any λ , the optimal objective value of $(pMP-LR_\lambda)$ provides a lower bound on that of (pMP) , which is useful for evaluating feasible solutions. In fact, the solution to $(pMP-LR_\lambda)$ itself can be used to generate a feasible solution: we simply keep the same p facilities open and assign each customer to its nearest open facility.

These steps—generating lower and upper bounds—are performed iteratively, and at the end of each iteration, the Lagrange multipliers λ are updated using subgradient optimization or some other method. The algorithm terminates when the lower and upper bounds are sufficiently close or when another stopping criteria (such as a limit on the number of iterations) is triggered. In the latter case, branch and bound can be used to narrow the gap between the bounds; at each node of the branch and bound tree, lower bounds are obtained using Lagrangian relaxation, rather than using LP relaxation as in the standard branch and bound method. This method is capable of solving problems with hundreds or thousands of facilities in seconds on a desktop computer [94]. For additional details about this method and other possible relaxations, see Daskin [22] and Fischer [93].

The Uncapacitated Fixed-Charge Location Problem

Background and Formulation. The UFLP is similar to the p -median problem except that it has no restriction on the number of open

facilities and it includes a fixed cost for opening each facility in the objective function. The UFLP (also known as the *simple plant location problem* [SPLP] as well as a variety of other similar names) can be formulated as follows:

(UFLP) minimize

$$\sum_{j \in J} f_j x_j + \sum_{i \in I} \sum_{j \in J} h_i c_{ij} y_{ij} \quad (12a)$$

$$\text{subject to } \sum_{j \in J} y_{ij} = 1 \quad \forall i \in I \quad (12b)$$

$$y_{ij} \leq x_j \quad \forall i \in I, \forall j \in J \quad (12c)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (12d)$$

$$y_{ij} \geq 0 \quad \forall i \in I, \forall j \in J. \quad (12e)$$

The constraints are identical to those for (pMP) except for the exclusion of (9d). Hence we do not discuss them further.

Solution Methods. The UFLP is NP-hard [95] but, similar to the p -median problem, it is considered relatively “easy” nevertheless. Variants of many of the heuristics described in the section titled “Solution Methods” on p.6 have also been applied to the UFLP, for example, the greedy heuristic [77]. So have the exact solution methods, such as branch and bound [88]. The Lagrangian relaxation approach described in the section titled “Solution Methods” on p.6 is also readily applicable; the only change is that, rather than opening the p facilities with the smallest β_j , we instead open *any* facility for which $\beta_j + f_j < 0$ (that is, for which the benefit outweighs the fixed cost).

Another important exact algorithm for the UFLP is Erlenkotter’s DUALOC algorithm [96], which operates on the dual of the LP relaxation. The first step is a *dual ascent procedure* that starts with a given dual solution, attempts to improve it, and then derives a primal feasible solution from it. If these solutions are optimal, the procedure terminates; otherwise, a *dual adjustment procedure* attempts to reduce complementary slackness violations in order to improve the primal/dual pair. Branch and bound is used in the event that the resulting primal solution is not an integer.

The p -Center Problem

Background and Formulation. The p -center problem locates p facilities in order to minimize the maximum distance between a customer and its assigned facility. The p -center problem was first proposed by Hakimi [67]. On a network, the Hakimi property does not apply—that is, the optimal locations may not lie at the nodes of the network. (Imagine locating a single facility on a network consisting of two nodes and an arc connecting them; the optimal solution is to locate halfway along the arc itself.) The literature, therefore, often distinguishes between the “vertex” p -center problem, in which facilities must be located at nodes, and the “absolute” p -center problem, in which facilities may be located on arcs. In either case, the demand nodes may be weighted or unweighted. Using the notation already introduced, the weighted vertex p -center problem can be formulated as follows:

$$(pCP) \text{ minimize } w \quad (13a)$$

$$\text{subject to } \sum_{j \in J} y_{ij} = 1 \quad \forall i \in I \quad (13b)$$

$$y_{ij} \leq x_j \quad \forall i \in I, \forall j \in J \quad (13c)$$

$$\sum_{j \in J} x_j \leq p \quad (13d)$$

$$\sum_{j \in J} h_i c_{ij} y_{ij} \leq w \quad \forall i \in I \quad (13e)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (13f)$$

$$y_{ij} \in \{0, 1\} \quad \forall i \in I, \forall j \in J. \quad (13g)$$

The objective function (13a) consists simply of a new decision variable, w , which represents the maximum weighted distance of a customer to its assigned facility. This interpretation of w is enforced through constraints (13e), which require w to be no smaller than the weighted assigned distance for each customer. Constraints (13b)–(13d) are identical to constraints (9b)–(9d) in the p -median problem. Note that the assignment variables (y) must be enforced as binary in constraint (13g); it is not sufficient to relax these constraints, as it is in the p -median problem.

Solution Methods. Similar to the p -median problem, the p -center problem can be solved in polynomial time for fixed p but is NP-hard for variable p [97]. Largely due to its minimax structure, the p -center problem does not lend itself nearly as well as the p -median problem to LP-based approaches or to Lagrangian relaxation. One common approach for solving the unweighted version is to iteratively solve the SCLP (see next section) by applying a line search on the coverage radius. The optimal objective value of the p -center problem is the smallest r such that the optimal objective value of the set covering problem with coverage radius r equals p [98]. The method can be extended fairly easily to the weighted case [19]. An important algorithm for the continuous version of the problem, due to Drezner and Wesolowsky [99], uses numerical integration of ordinary differential equations.

The Set Covering Location Problem

Background and Formulation. The SCLP, and the MCLP discussed in the next section, both make use of the notion of *coverage*: A customer is considered “covered” if there is an open facility within a prespecified radius r (measured by distance, travel time, or transportation cost). The objective of the SCLP is to minimize the total number of facilities located while covering every customer. For more details on the SCLP, see Mirchandani and Francis [21] and Snyder [41].

The SCLP was first introduced in the context of facility location by Hakimi [100]. (The graph-theoretic vertex cover problem, which had been studied well before Hakimi’s article, can be considered a special case of the SCLP in which $I = J$; $d_{ij} = 1$ for all i, j ; and $r = 1$. In fact, it is this special case that Hakimi’s article considers, although he is usually credited with introducing the general SCLP.)

Let $N_i = \{j \in J : c_{ij} \leq r\}$ be the set of facilities that cover customer i . The following mathematical programming formulation is due to Toregas *et al.* [101]:

$$(\text{SCLP}) \text{ minimize } \sum_{j \in J} x_j \quad (14a)$$

$$\text{subject to } \sum_{j \in N_i} x_j \geq 1 \quad \forall i \in I \quad (14b)$$

$$x_j \in \{0, 1\} \quad \forall j \in J. \quad (14c)$$

The objective function (14a) simply calculates the number of open facilities. (Alternately, the x_j can be weighted by f_j to minimize the total fixed cost of the opened facilities.) Constraints (14b) ensure that, for each customer i , at least one facility that covers i is open.

The Hakimi property does not apply to the SCLP; a counterexample consists of a single edge of length 1 connecting two nodes, and a coverage radius of $r = 0.5$.

Solution Methods. The SCLP is NP-hard [95]. Hakimi [100] proposed a solution method using Boolean functions that requires enumerating all coverings and is therefore impractical for realistic instances. Instead, the SCLP is generally solved using IP techniques, the first of which was proposed by Toregas *et al.* [101], who formulated the problem as in problem (14) and used LP branch and bound to solve it. This formulation typically has a small (often zero) integrality gap [102], and Toregas *et al.* [101] suggest a cut that strengthens the LP relaxation further. The size of the IP itself can often be reduced by eliminating dominated columns (i.e., facilities whose covered customers are a subset of another facility’s) and dominated rows (i.e., customers whose covering facilities are a superset of another customer’s) [103]. A large body of literature also exists for solving general (i.e., not facility-location-specific) set covering problems [104], and these algorithms also apply to the SCLP.

The Maximal Covering Location Problem

Background and Formulation. The MCLP [105] can be thought of as the inverse of the SCLP: the SCLP minimizes the number of facilities required to cover all customers, while the MCLP maximizes the total demand of covered customers subject to a limit on the number of facilities opened. One important advantage of the MCLP over the SCLP is that it accounts for differences in customer

demands (and therefore importance) rather than treating all customers as identical, as the SCLP does.

Let z_i be a decision variable that is equal to 1 if customer i is covered and 0 otherwise, and define N_i as the set of facilities that cover customer i , as above. Then the MCLP can be formulated as follows:

$$\text{(MCLP) maximize } \sum_{i \in I} h_i z_i \quad (15a)$$

$$\text{subject to } z_i \leq \sum_{j \in N_i} x_j \quad \forall i \in I \quad (15b)$$

$$\sum_{j \in J} x_j \leq p \quad (15c)$$

$$x_j \in \{0, 1\} \quad \forall j \in J \quad (15d)$$

$$z_i \in \{0, 1\} \quad \forall i \in I. \quad (15e)$$

The objective function (15a) calculates the total covered demands. Constraints (15b) prevent a customer from being considered covered unless some covering facility has been opened. Constraints (15c) require at most p facilities to be opened.

Solution Methods. The MCLP is NP-hard [106] but, like several of the problems discussed so far, it generally has a fairly tight LP bound. Church and ReVelle [105] report that the LP relaxations of the majority of their test instances have integer solutions, and they use branch and bound to solve the remainder of the problems. They also propose a method for solving a problem for consecutive values of p ; it attempts to create an integer solution for the p -facility problem using a fractional solution for that problem and an integer solution for the $(p - 1)$ -facility problem. Lagrangian relaxation, relaxing constraints (15b), is also effective [107,108].

Heuristic methods have been proposed, including greedy-add [109]; greedy adding with substitution (GAS) [105], which combines greedy-add with an exchange heuristic; and metaheuristics [110,111].

EXTENSIONS

A broad literature has arisen that builds on the classical models discussed in the section

titled “Classical Models” to develop richer models that address more realistic and more nuanced issues. In this section several topics that have emerged over the past few decades are discussed briefly.

Hazardous Facility Location Problems

The models discussed in the section titled “Classical Models” all aim to locate facilities close to customers. However, for some facilities—for example, prisons and nuclear waste sites—the opposite is desired. So-called *hazardous facility location problems* account for this by maximizing the distance between facilities and customers or among the facilities themselves.

Obnoxious facility location problems attempt to maximize the distance between customers and facilities. The *maxisum problem* is the opposite of the p -median problem in that it locates p facilities in order to *maximize* the demand-weighted distance between customers and their nearest facilities. The formulation is a bit more complicated than simply changing the sense of constraint (9a) from minimize to maximize, since doing so would cause the model to assign customers to their farthest facilities. Instead so-called “closest assignment” constraints ensure that customers are assigned to their nearest opened facilities. Goldman and Dearing [112] and Church and Garfinkel [113] introduce the 1-facility version; Drezner and Wesolowsky [114] study the p -facility version. Similarly, the *maximin problem* is a maximization version of the p -center problem.

The *p -dispersion problem* [115,116] seeks to maximize the minimum distance between facilities—customers are entirely omitted from the problem. For example, the military may wish to disperse arms depots geographically to reduce the risk of a simultaneous attack, or a retailer may wish to keep branch outlets at some distance from one another to avoid cannibalization.

See Erkut and Neuman [42] and Cappanera [43] for reviews of these and other hazardous location problems.

Dynamic Location Problems

Dynamic location problems assume that facilities may be opened, closed, or moved

over time, either due to limited budgets that prevent all facilities from being opened at once, or in response to changes in the problem parameters (demands, costs, etc.). (A closely related body of literature assumes that the parameters change stochastically over time, but these models usually assume that facility locations are static; see the section titled “Stochastic and Robust Models”). Examples include [117–123]; for a review, see Owen and Daskin [52].

Hierarchical Location Problems

The classical models make facility location decisions for a single echelon (or layer)—for example, plants, warehouses, or retailers. In practice, it is often important to make decisions about multiple echelons simultaneously, including both location and flow decisions. Such problems are known as *hierarchical location problems* [124–126]; see the reviews by Narula [44] and Sahin and Süral [45]. Many hierarchical problems also include multiple commodities, building on the seminal (single-echelon) multicommodity location model by Geoffrion and Graves [127]. This topic is also closely related to network design problems, in which the objective is to locate arcs, rather than nodes, of the network [128,129].

Competitive Location Problems

In *competitive location problems*, two or more players locate facilities in order to maximize their own profit or market share. The typical example concerns two competing retail or restaurant chains (Home Depot and Lowe’s, McDonald’s and Burger King) that are near substitutes so that consumers choose one over the other, largely based on proximity. The earliest such model is by the economist Hotelling [130], who considered a continuous version of the problem. Most subsequent works consider discrete or network versions [131,132]; for reviews, see Plastria [46] and Eiselt *et al.* [47].

Hub Location Problems

Whereas classical facility location problems involve flows between facilities and customers, *hub location problems* also involve

flows among the facilities. The motivation comes from airlines, third party logistics companies, and other firms that operate hub-and-spoke systems. In such systems, interfacility flows are as important as, or more important than, facility–customer flows, in part because they often use a more cost-effective mode of transportation (e.g., larger-capacity trucks or airplanes). Demands in hub location problems are usually expressed as customer-to-customer flows; that is, they are origin-and-destination-specific, whereas classical problems use destination-specific demands. The standard formulation for the basic problem of locating p hubs to minimize the total transportation cost [3] has the same constraints as the p -median problem, but its objective is quadratic in the assignment variables to capture the intrafacility flows, making the problem significantly more difficult to solve. For surveys of the literature on hub location, see O’Kelly and Miller [48] and Campbell *et al.* [49].

Facility Congestion

Facility congestion refers to the fact that a customer wishing to use a facility may find it already in use. A classic example is the location of ambulances or fire trucks. Facility location problems with congestion account for this possibility when deciding on where to locate facilities (and, sometimes, the number of servers to locate at each facility). These are essentially queueing problems in which the facility or server is the service mechanism; customers finding no servers available must wait in queue. The hypercube model of Larson [133,134] models this queueing process explicitly, but the resulting expressions are generally too complex to embed into optimization models. Instead, most of the research has focused on approximating the probability that a customer finds an available server, and building enough redundancy to ensure that this probability is sufficiently high. Daskin’s [135,136] maximum expected covering location model (MEXCLM) is an extension of the MCLP that maximizes the total expected coverage assuming that all facilities have the same availability probability; it has served as the basis for a host of other models that

relax this assumption in various ways; for reviews, see Daskin *et al.* [50] and Berman and Krass [51].

Stochastic and Robust Models

Stochastic and robust location models account explicitly for the fact that demands, costs, and other parameters may be stochastic or unknown at the time when facility location decisions are made. Uncertain parameters are generally represented as scenarios (for which probabilities may or may not be known) or as intervals in which the parameters are assumed to lie. The problems are modeled and solved using stochastic optimization (which optimizes the expected value of the objective function) or robust optimization (which optimizes some other measure of uncertainty, typically reflecting some measure of risk aversion). Stochastic problems tend to be easier to solve than robust problems, which are usually solved exactly for special cases or heuristically for general problems. The earliest articles on this topic extend classical models such as the p -median or UFLP, minimizing the expected cost [137,138], a mean–variance objective [139], or the worst-case cost or regret [140]. For reviews of stochastic and robust problems, see Owen and Daskin [52], Louveaux [53], and Snyder [54].

A related stream of literature considers uncertainty from the supply side in the form of facility or supplier disruptions, during which customers must be assigned to backup facilities. One set of models assumes that facilities are stochastically disrupted [141–143], while another assumes that facilities are interdicted by a malicious entity. The latter models either determine the worst-case damage that can be caused for a given set of facilities [144] or determine which facilities to “fortify” to protect against this worst-case damage [145]. For reviews of facility location problems with disruptions, see Snyder and Daskin [55] and Snyder *et al.* [56,57].

Integrated Location Models

The term *integrated location models* generally refers to models that integrate facility location (a strategic decision) with tactical or

operational decisions, most commonly vehicle routing or inventory. Location–routing models recognize that deliveries are often made using less-than-truckload (LTL) shipments, in which the transportation cost from a facility to a customer depends on which other customers are also served by that facility. In contrast, classical location models assume full truckload (TL) deliveries, in which the transportation cost is independent of other deliveries. Location–routing problems are generally quite difficult to solve, since they combine two NP-hard problems, facility location and vehicle routing. They typically either model the outbound flows from facilities to customers [146,147], or model both the outbound flows and the inbound (TL) flows from suppliers to facilities [148]. For reviews of location–routing models, see Laporte [58], Min *et al.* [59], and Daskin *et al.* [60].

Location–inventory models incorporate the cost of managing inventory into the objective function of a facility location problem. Nozick and Turnquist [149] formulate a model in which the cost of a base-stock policy at each facility is approximated by a linear function that is embedded into a model similar to the UFLP. Shen *et al.* [150] and Daskin *et al.* [151] formulate a similar model but include the nonlinear inventory costs of both cycle stock and safety stock, accounting for economies of scale in ordering as well as the risk-pooling effect [152]. As a result, the solutions of their model tend to open fewer facilities than those of the UFLP. The model can be solved nearly as efficiently as the UFLP using either column generation [150] or Lagrangian relaxation [151]. For reviews of location–inventory models, see Shen [61] or the article titled **Introduction to Facility Location** in this encyclopedia.

Public Sector Models

Although most facility location models can be applied either in the private or in the public sector, models that focus primarily on cost or profit objectives (such as the p -median and UFLP) tend to be most applicable in the private sector, while those that use coverage or equity objectives (such as the p -center, SCLP, and MCLP) are more readily applied to the public sector. Marianov and Serra

[62] review these latter models, in addition to models with congestion (see the section titled “Facility Congestion”), and public sector modifications of the p -median problem. These models and their extensions have been applied to the location of subsidized housing [153], humanitarian relief agencies [154], and wildlife reserve sites [11].

CONCLUSION

This article has provided a broad overview of the field of facility location. For the sake of brevity, many related topics are omitted, which can be found described in the articles listed below, or in the references cited therein. It is hoped that this article will serve as a useful reference for those actively working in the field of facility location, as well as a helpful starting point for those just beginning to explore it.

Most of the extensions discussed in the section titled “Extensions” are active areas of research. Future research areas will likely include modeling and solution techniques for problems with nonlinearities (such as economies of scale), large-scale problems with a great many nodes or scenarios, problems that exhibit multiple temporal or geographical scales, models to hedge against supply chain risk, and problems with humanitarian and environmental benefits.

REFERENCES

1. ReVelle C, Bigman D, Schilling D, *et al.* Facility location: a review of context-free and EMS models. *Health Serv Res* 1977;12(2):129–146.
2. Schilling D, ReVelle C, Cohon J, *et al.* Some models for fire protection locational decisions. *Eur J Oper Res* 1980;5:1–7.
3. O’Kelly ME. A quadratic integer program for the location of interacting hub facilities. *Eur J Oper Res* 1987;32:393–404.
4. Price WL, Turcotte M. Locating a blood bank. *Interfaces* 1986;16:17–26.
5. Min H. A multiobjective retail service location model for fast food restaurants. *Omega* 1987;15:429–441.
6. Goodchild MF, Booth PJ. Location and allocation of recreational facilities: Public swimming pools in London, Ontario. *Ont Geogr* 1980;15:35–51.
7. Tewari VK, Sidheswar J. High school location decision making in rural India and location-allocation models. In: Ghosh A, Rushton G, editors. *Spatial analysis and location-allocation models*. New York: Van Nostrand Reinhold Co.; 1987. pp. 137–162.
8. Hodgson JM, Rosing KE, Zhang J. Locating vehicle inspection stations to protect a transportation network. *Geogr Anal* 1996;28:299–314.
9. Gleason J. A set covering approach to bus stop location. *Omega* 1975;3:605–608.
10. Current J, Daskin MS, Schilling D. Discrete network location models. In: Drezner Zvi, Hamacher HW, editors. *Facility location: applications and theory*. New York: Springer; 2002. Chapter 3.
11. ReVelle C, Williams JC. Reserve design and facility siting. In: Drezner Z, Hamacher HW, editors. *Facility location: applications and theory*. New York: Springer; 2002. Chapter 10. pp. 309–330.
12. Drezner Z. Location strategies for satellites’ orbits. *Nav Res Log* 1988;35:503–512.
13. Cornuejols G, Fisher ML, Nemhauser GL. Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms. *Manage Sci* 1977;23(8):789–810.
14. Daskin MS, Jones PC, Lowe TJ. Rationalizing tool selection in a flexible manufacturing system for sheet metal products. *Oper Res* 1990;38:1104–1115.
15. Ginsberg V, Pestieau P, Thisse JF. A spatial model of party competition with electoral and ideological objectives. In: Ghosh A, Rushton G, editors. *Spatial analysis and location-allocation models*. New York: Van Nostrand Reinhold Co.; 1987. pp. 101–117.
16. Brotcorne L, Laporte G, Semet F. Fast heuristics for large scale covering-location problems. *Comput Oper Res* 2002;29(6):651–665. DOI: 10.1016/S0305-0548(99)00088-X.
17. Bramel J, Simchi-Levi D. A location-based heuristic for general routing problems. *Oper Res* 1995;43:649–660.
18. Hale T. Trevor hale’s location science references. 2010. Available at <http://gator.dt.uh.edu/~halet/>. Accessed 2010 June 17.
19. Handler GY, Mirchandani PB. *Location on networks*. Cambridge (MA): MIT Press; 1979.

20. Hurter AP, Martinich JS. Facility location and the theory of production. Boston (MA): Kluwer Academic Publishers; 1989.
21. Mirchandani PB, Francis RL, editors. Discrete location theory. New York: Wiley-Interscience; 1990.
22. Daskin MS. Network and discrete location: models, algorithms, and applications. New York: Wiley; 1995.
23. Drezner Z, editor. Facility location: a survey of applications and methods. New York: Springer; 1995.
24. Drezner Z, Hamacher HW, editors. Facility location: applications and theory. New York: Springer; 2002.
25. Eiselt HA, Sandblom C-L. Decision analysis, location models, and scheduling problems. New York: Springer; 2004.
26. Eiselt HA, Marianov V, editors. Foundations of location analysis. New York: Springer; 2010. In press.
27. INFORMS Section on Location Analysis. Available at location.section.informs.org.
28. EURO Working Group on Locational Analysis. Available at www.ewgla.eu.
29. Brandeau ML, Chiu SS. An overview of representative problems in location research. *Manage Sci* 1989;35(6):645–674.
30. Hale TS, Moberg CR. Location science research: a review. *Ann Oper Res* 2003;123:21–35.
31. Klose A, Drexel A. Facility location models for distribution system design. *Eur J Oper Res* 2005;162:4–29.
32. ReVelle CS, Eiselt HA. Location analysis: a synthesis and survey. *Eur J Oper Res* 2005;165(1):1–19.
33. ReVelle CS, Eiselt HA, Daskin MS. A bibliography for some fundamental problem categories in discrete location science. *Eur J Oper Res* 2008;184(3):817–848.
34. Daskin MS. What you should know about location modeling. *Nav Res Log* 2008;55:283–294. DOI: 10.1002/nav.20284.
35. Smith HK, Laporte G, Harper PR. Locational analysis: highlights of growth to maturity. *J Oper Res Soc* 2009;60(S1):S140–S148.
36. Tansel BC, Francis RL, Lowe TJ. Location on networks: a survey. Part I: the p -center and p -median problems. *Manage Sci* 1983a;29(3):482–497.
37. Tansel BC, Francis RL, Lowe TJ. Location on networks: a survey. Part II: exploiting tree network structure. *Manage Sci* 1983b;29(4):498–511.
38. Reese J. Methods for solving the p -median problem: an annotated bibliography. Technical report. Department of Mathematics, Trinity University; 2005.
39. Mladenovic N, Brimberg J, Hansen P, *et al.* The p -median problem: a survey of metaheuristic approaches. *Eur J Oper Res* 2007;179(3):927–939. DOI: 10.1016/j.ejor.2005.05.034.
40. Drezner Z, Klamroth K, Schöbel A, *et al.* The Weber problem. In: Drezner Z, Hamacher HW, editors. Facility location: applications and theory. New York: Springer; 2002. Chapter 1.
41. Snyder LV. Covering problems. In: Eiselt HA, Marianov V, editors. Foundations of location analysis. Springer; 2010. In press.
42. Erkut E, Neuman S. Analytical models for locating undesirable facilities. *Eur J Oper Res* 1989;40(3):275–291. DOI: 10.1016/0377-2217(89)90420-7.
43. Cappanera P. A survey on obnoxious facility location problems. Technical report. University of Pisa; 1999.
44. Narula SC. Minisum hierarchical location-allocation problem on a network: a survey. *Ann Oper Res* 1986;6:257–272.
45. Sahin G, Süral H. A review of hierarchical facility location models. *Comput Oper Res* 2007;34(8):2310–2331. DOI: 10.1016/j.cor.2005.09.005.
46. Plastria F. Static competitive facility location: an overview of optimisation approaches. *Eur J Oper Res* 2001;129(3):461–470. DOI: 10.1016/S0377-2217(00)00169-7.
47. Eiselt HA, Laporte G, Thisse J-F. Competitive location models: a framework and bibliography. *Transp Sci* 1993;27(1):44–54.
48. O’Kelly ME, Miller HM. The hub network design problems: a review and synthesis. *J Transp Geogr* 1994;2:31–40.
49. Campbell JF, Ernst AT, Krishnamoorthy M. Hub location problems. In: Drezner Z, Hamacher HW, editors. Facility location: applications and theory. New York: Springer; 2002. Chapter 12. pp. 375–410.
50. Daskin MS, Hogan K, ReVelle C. Integration of multiple, excess, backup, and expected covering models. *Environ Plann B Plann Des* 1988;15(1):15–35.
51. Berman O, Krass D. Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher HW, editors. Facility location: applications and theory. New York: Springer; 2002. pp. 331–373.

52. Owen SH, Daskin MS. Strategic facility location: a review. *Eur J Oper Res* 1998;111(3):423–447.
53. Louveaux FV. Stochastic location analysis. *Locat Sci* 1993;1(2):127–154.
54. Snyder LV. Facility location under uncertainty: a review. *IIE Trans* 2006;38(7):537–554.
55. Snyder LV, Daskin MS. Models for reliable supply chain network design. In: Murray AT, Grubbs TH, editors. *Reliability and vulnerability in critical infrastructure: a quantitative geographic perspective*. New York: Springer; 2007. Chapter 13. pp. 257–289.
56. Snyder LV, Scaparra MP, Daskin ML, *et al.* Planning for disruptions in supply chain networks. In: Johnson MP, Norman B, Secomandi N, editors. *Tutorials in operations research*. Baltimore: INFORMS; 2006. Chapter 9. pp. 234–257.
57. Snyder LV, Bulut Z, Peng P, *et al.* Supply chain disruptions: a review. Working paper. Lehigh University; 2010.
58. Laporte G. Location routing problems. In: Golden BL, Assad AA, editors. *Vehicle routing: methods and studies*. Amsterdam: North-Holland Publishing Co.; 1988. pp. 163–197.
59. Min H, Jayaraman V, Srivastava R. Combined location routing problems: a synthesis and future research directions. *Eur J Oper Res* 1998;108:1–15.
60. Daskin MS, Snyder LV, Berger RT. Facility location in supply chain design. In: Langevin A, Riopel D, editors. *Logistics systems: design and operation*. New York: Springer; 2005. Chapter 2. pp. 39–66.
61. Shen Z-JM. Integrated supply chain design models: a survey and future research directions. *J Ind Manage Optim* 2007;3(1):1–27.
62. Marianov V, Serra D. Location problems in the public sector. In: Drezner Z, Hamacher HW, editors. *Facility location: applications and theory*. New York: Springer; 2002. Chapter 4. pp. 121–152.
63. Melo MT, Nickel S, da Gama FS. Facility location and supply chain management - a review. *Eur J Oper Res* 2009;196(2):401–412. DOI: 10.1016/j.ejor.2008.05.007.
64. Weber A. *Über den Standort der Industrien, Erster Teil: Reine Theorie des Standortes*. Tübingen: Mohr; 1909.
65. Daganzo CF. *Logistics systems analysis*. Berlin: Springer; 1991.
66. Goldman AJ. Optimal center location in simple networks. *Transp Sci* 1971;5(2):212–221.
67. Hakimi SL. Optimum locations of switching centers and the absolute centers and medians of a graph. *Oper Res* 1964;12(3):450–459.
68. Church RL, Meadows B. Location modelling using maximum service distance criteria. *Geogr Anal* 1979;11:358–373.
69. Weiszfeld E. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Math J* 1937;43:355–386.
70. Brimberg J, Love RF. Properties of ordinary and weighted sums of order p used for distance estimation. *RAIRO-Rech Opér* 1995;29(1):59–72.
71. Miehele W. Link-length minimization in networks. *Oper Res* 1958;6(2):232–243.
72. Hansen P, Peeters D, Thisse J-F. An algorithm for a constrained Weber problem. *Manage Sci* 1982;28(11):1285–1295.
73. Katz IN, Cooper L. Facility location in the presence of forbidden regions, I: formulation and the case of Euclidean distance with one forbidden circle. *Eur J Oper Res* 1981;6(2):166–173.
74. Wesolowsky GO. Location of the median line for weighted points. *Environ Plann A* 1975;7(2):163–170.
75. Revelle CS, Swain RW. Central facilities location. *Geogr Anal* 1970;2:30–42.
76. Kariv O, Hakimi SL. An algorithmic approach to network location problems. II: The p -medians. *SIAM J Appl Math* 1979b;37(3):539–560.
77. Kuehn AA, Hamburger MJ. A heuristic program for locating warehouses. *Manage Sci* 1963;9(4):643–666.
78. Maranzana FE. On the location of supply points to minimize transport costs. *Oper Res Q* 1964;15(3):261–270.
79. Teitz MB, Bart P. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Oper Res* 1968;16(5):955–961.
80. Whitaker RA. A fast algorithm for the greedy interchange of large-scale clustering and median location problems. *INFOR* 1983;21(2):95–108.
81. Hansen P, Mladenović N. Variable neighborhood search for the p -median. *Locat Sci* 1997;5(4):207–226.
82. Rolland E, Schilling DA, Current JR. An efficient tabu search procedure for

- the p -median problem. *Eur J Oper Res* 1997;96(2):329–342.
83. Hosage CM, Goodchild MF. Discrete space location-allocation solutions from genetic algorithms. *Ann Oper Res* 1986;6(2):35–46.
 84. Murray AT, Church RL. Applying simulated annealing to location-planning models. *J Heuristics* 1996;2(1):31–53.
 85. Rosing KE, ReVelle CS. Heuristic concentration: two stage solution construction. *Eur J Oper Res* 1997;97(1):75–86. DOI: 10.1016/S0377-2217(96)00100-2.
 86. Levanova TV, Loresh MA. Algorithms of ant system and simulated annealing for the p -median problem. *Automat Remote Control* 2004;65(3):431–438.
 87. Merino ED, Pérez JM, Aragonés JJ. Neural network algorithms for the p -median problem. *ESANN 2003 Proceedings: European Symposium on Artificial Neural Networks. Bruges (Belgium); 2003. pp. 385–391.*
 88. Efraymson MA, Ray TL. A branch-bound algorithm for plant location. *Oper Res* 1966;14(3):361–368.
 89. de Farias IR A family of facets for the uncapacitated p -median polytope. *Oper Res Lett* 2001;28(4):161–167.
 90. Galvão RD. A dual-bounded algorithm for the p -median problem. *Oper Res* 1980;28(5):1112–1121.
 91. Garfinkel RS, Neebe AW, Rao MR. An algorithm for the m -median plant location problem. *Transp Sci* 1974;8(3):217–236.
 92. Galvão RD. A graph theoretical bound for the p -median problem. *Eur J Oper Res* 1981;6(2):162–165.
 93. Fisher ML. The Lagrangian relaxation method for solving integer programming problems. *Manage Sci* 1981;27(1):1–18.
 94. Daskin MS. {SITATION} software. 2010. Available at sitemaker.umich.edu/msdaskin/software. Accessed 2010.
 95. Garey MR, Johnson DS. *Computers and intractability: a guide to the theory of NP-completeness*. New York: W. H. Freeman and Company; 1979.
 96. Erlenkotter D. A dual-based procedure for uncapacitated facility location. *Oper Res* 1978;26(6):992–1009.
 97. Kariv O, Hakimi SL. An algorithmic approach to network location problems. I: the p -centers. *SIAM J Appl Math* 1979a;37(3):513–538.
 98. Minieka E. Short notes: the m -center problem. *SIAM Rev* 1970;12(1):138–139.
 99. Drezner Z, Wesolowsky GO. A new method for the multifacility minimax location problem. *J Oper Res Soc* 1978;29(11):1095–1101.
 100. Hakimi SL. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Oper Res* 1965;13(3):462–475.
 101. Bramel J, Simchi-Levi D. On the effectiveness of set covering formulations for the vehicle routing problem with time windows. *Oper Res* 1997;45(2):295–301.
 102. Toregas C, Swain R, ReVelle C, *et al.* The location of emergency service facilities. *Oper Res* 1971;19(6):1363–1373.
 103. Toregas C, ReVelle C. Optimal location under time or distance constraints. *Papers Reg Sci Assoc* 1972;28:133–143.
 104. Christofides N, Paixão J. Algorithms for large scale set covering problems. *Ann Oper Res* 1993; 43(5):259–277.
 105. Church R, ReVelle C. The maximal covering location problem. *Papers Reg Sci Assoc* 1974;32:101–118.
 106. Megiddo N, Zemel E, Hakimi SL. The maximum coverage location problem. *SIAM J Algebr Discrete Methods* 1983;4(2):253–261.
 107. Daskin MS, Haghani AE, Khanal M, *et al.* Aggregation effects in maximum covering models. *Ann Oper Res* 1989;18:115–140.
 108. Galvão RD, ReVelle C. A Lagrangean heuristic for the maximal covering location problem. *Eur J Oper Res* 1996; 88:114–123.
 109. Church R. *Synthesis of a class of public facilities location models* [PhD thesis]. Baltimore (MD): The Johns Hopkins University; 1974.
 110. Resende MGC. Computing approximate solutions of the maximum covering problem with GRASP. *J Heuristics* 1998;4(2):161–177. DOI: 10.1023/A:1009677613792.
 111. Jaramillo JH, Bhadury J, Batta R. On the use of genetic algorithms to solve location problems. *Comput Oper Res* 2002;29(6):761–779. DOI: 10.1016/S0305-0548(01)00021-1.
 112. Goldman AJ, Dearing PM. Concepts of optimal location for partially noxious facilities. *ORSA Bull* 1975;23(1):B–31.
 113. Church RL, Garfinkel RS. Locating an obnoxious facility on a network. *Transp Sci* 1978;12(2):107–118.
 114. Drezner Z, Wesolowsky GO. Location of multiple obnoxious facilities. *Transp Sci* 1985;19:193–202.

115. Chandrasekaran R, Daughety A. Location on tree networks: p -centre and n -dispersion problems. *Math Oper Res* 1981;6(1):50–57.
116. Kubj MJ. Programming models for facility dispersion: the p -dispersion and maximum dispersion problems. *Geogr Anal* 1987;19:315–329.
117. Ballou RH. Dynamic warehouse location analysis. *J Mark Res* 1968;5:271–276.
118. Scott AJ. Dynamic location-allocation systems: some basic planning strategies. *Environ Plann* 1971;3:73–82.
119. Sheppard ES. A conceptual framework for dynamic location-allocation analysis. *Environ Plann A* 1974;6:547–564.
120. Wesolowsky GO, Truscott WG. The multiperiod location-allocation problem with relocation of facilities. *Manage Sci* 1976;22(1):57–65.
121. Van Roy TJ, Erlenkotter D. A dual-based procedure for dynamic facility location. *Manage Sci* 1982;28(10):1091–1105.
122. Campbell JF. Locating transportation terminals to serve an expanding demand. *Transp Res Part B* 1990;24B(3):173–192.
123. Drezner Z. Dynamic facility location: the progressive p -median problem. *Locat Sci* 1995;3(1):1–7.
124. Narula SC, Ogbu UI. An hierarchical location–allocation problem. *Omega* 1979;7(2):137–143.
125. Weaver JR, Church RL. The nested hierarchical median facility location problem. *INFOR* 1991;29-2:100–115.
126. Serra D, ReVelle C. The pq -median problem: location and districting of hierarchical facilities. *Locat Sci* 1993;1(4):299–312.
127. Geoffrion AM, Graves GW. Multi-commodity distribution system design by Benders decomposition. *Manage Sci* 1974;20(5):822–844.
128. Magnanti TL, Mireault P, Wong RT. Tailoring Benders decomposition for uncapacitated network design. *Math Program Study* 1986;26:112–154.
129. Balakrishnan A, Magnanti TL, Wong RT. A dual-ascent procedure for large-scale uncapacitated network design. *Oper Res* 1989;37(5):716–740.
130. Hotelling H. Stability in competition. *Econ J* 1929;39(153):41–57.
131. Hakimi SL. On locating new facilities in a competitive environment. *Eur J Oper Res* 1983;12:29–55.
132. Serra D, ReVelle C. Competitive location and pricing on networks. *Geogr Anal* 1999;31:109–129.
133. Larson RC. A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput Oper Res* 1974;1:67–95.
134. Larson RC. Approximating the performance of urban emergency service systems. *Oper Res* 1975;23(5):845–868.
135. Daskin MS. Application of an expected covering model to emergency medical service system design. *Decis Sci* 1982;13:416–439.
136. Daskin MS. A maximum expected covering location model: formulation, properties and heuristic solution. *Transp Sci* 1983;17(1):48–70.
137. Weaver JR, Church RL. Computational procedures for location problems on stochastic networks. *Transp Sci* 1983;17(2):168–180.
138. Mirchandani PB, Oudjit A, Wong RT. ‘Multidimensional’ extensions and a nested dual approach for the m -median problem. *Eur J Oper Res* 1985;21(1):121–137.
139. Jucker JV, Carlson RC. The simple plant-location problem under uncertainty. *Oper Res* 1976;24(6):1045–1055.
140. Serra D, Marianov V. The p -median problem in a changing network: the case of Barcelona. *Locat Sci* 1998;6:383–394.
141. Drezner Z. Heuristic solution methods for two location problems with unreliable facilities. *J Oper Res Soc* 1987;38(6):509–514.
142. Snyder LV, Daskin MS. Reliability models for facility location: the expected failure cost case. *Transp Sci* 2005;39(3):400–416.
143. Berman O, Krass D, Menezes MBC. Facility reliability issues in network p -median problems: strategic centralization and co-location effects. *Oper Res* 2007;55(2):332–350.
144. Church RL, Scaparra MP, Middleton RS. Identifying critical infrastructure: the median and covering facility interdiction problems. *Ann Assoc Am Geogr* 2004;94(3):491–502.
145. Church RL, Scaparra MP. Protecting critical assets: the r -interdiction median problem with fortification. *Geogr Anal* 2006;39(2):129–146.
146. Laporte G, Nobert Y, Pelletier J. Hamiltonian location problems. *Eur J Oper Res* 1983;12:82–89.
147. Berger RT. Location-routing models for distribution system design [PhD thesis].

- Department of Industrial Engineering and Management Sciences, Northwestern University; 1997.
148. Perl J, Daskin MS. A warehouse location-routing problem. *Transp Res Part B* 1985;19B(5):381–396.
 149. Nozick LK, Turnquist MA. Inventory, transportation, service quality and the location of distribution centers. *Eur J Oper Res* 2001;129(2):362–371.
 150. Shen Z-JM, Coullard CR, Daskin MS. A joint location-inventory model. *Transp Sci* 2003;37(1):40–55.
 151. Daskin MS, Coullard CR, Shen Z-JM. An inventory-location model: Formulation, solution algorithm and computational results. *Ann Oper Res* 2002;110:83–106.
 152. Eppen GD. Effects of centralization on expected costs in a multi-location newsboy problem. *Manage Sci* 1979;25(5):498–501.
 153. Johnson MP, Hurter AP. An optimization model for location of subsidized housing in metropolitan areas. *Locat Sci* 1998;6(1-4):257–279. DOI: 10.1016/S0966-8349(98)00044-8.
 154. Balcik B, Beamon BM. Facility location in humanitarian relief. *Int J Log Res Appl* 2008;11(2):101–121.

INTRODUCTION TO LARGE-SCALE LINEAR PROGRAMMING AND APPLICATIONS

STEPHEN J. STOYAN
MAGED M. DESSOUKY
XIAOQING WANG

Daniel J. Epstein Department of Industrial
and Systems Engineering, University of
Southern California, Los Angeles,
California

INTRODUCTION

Linear programming (LP) has been used as an effective tool to solve a number of applications in manufacturing, transportation, military, health care, and finance. In order to accurately model practical systems, however, there may be many variables and constraints necessary to capture all of the attributes in the problem. In many applications, the size of the formulation in terms of the number of variables and constraints can be quite large. For example, Gondzio and Grothey [1] solve problems with variable sizes up to 10^9 and Leachman [2] formulates an LP model of over 10^5 constraints to solve a production planning problem for semiconductors at Harris Corporation. With such large-size problems, specialized procedures such as column generation and cutting-plane methods may need to be employed to solve them. In some cases, the problem may have a special structure, where decomposition methods can also be used. The purpose of this article is to provide an overview of these specialized procedures, and we conclude with an example of a large-scale formulation for solving a route balancing problem. For general discussion on the progress made with LP, problem size, and algorithmic developments one may refer to Bixby [3]. In this article, we describe some of the fundamental ideas used in large-scale LP with an emphasis on the simplex method. However, the reader may also refer to the literature on interior point methods [4,5,11].

LINEAR PROGRAMMING PROBLEM

Consider the following standard LP problem:

$$\min \mathbf{c}^\top x \quad (1)$$

$$\text{s.t. } \mathbf{A}x = \mathbf{b} \quad (2)$$

$$x \geq 0, \quad (3)$$

where $x \in \mathbb{R}^n$, $\mathbf{A}$ is an $m \times n$ matrix, and $\mathbf{c}$ and $\mathbf{b}$ are vectors of length n and m , respectively. Now suppose the problem is so large, for example, the $\mathbf{A}$ matrix cannot be stored in memory, that the LP cannot be solved. Using the simplex method, large-scale problems can be solved by adding a few extra steps to the algorithm. Depending on the structure of the $\mathbf{A}$ matrix it may be possible to decompose the problem such that solving smaller subproblems provides sufficient optimality conditions. Recall from the simplex method, an LP is partitioned into Basic (B) and Nonbasic (N) elements. Such a partitioning on Equations (1)–(3) gives

$$\min \mathbf{c}_B^\top x_B + \mathbf{c}_N^\top x_N \quad (4)$$

$$\text{s.t. } \mathbf{B}x_B + \mathbf{N}x_N = \mathbf{b} \quad (5)$$

$$x_B, x_N \geq 0, \quad (6)$$

where $\mathbf{B}$ is an $m \times m$ matrix composed of the set of linearly independent columns $\mathbf{A}_i \forall i \in B$ ($\mathbf{A}_i$ refers to column i in the $\mathbf{A}$ matrix), $\mathbf{N}$ is a matrix composed of the remaining set of columns $\mathbf{A}_j \forall j \in N$, $\mathbf{c}_B$ refers to the vector in the objective function containing the elements in B , and $\mathbf{c}_N$ refers to the elements in N . Also, x_B and x_N refer to the basic and nonbasic decision variables, respectively. Finally, $\mathbf{B}^{-1}$ exists since the columns of $\mathbf{B}$ are linearly independent. Then, solving for x_B in Equation (5) and making the substitution in Equation (1) provides

$$\min \mathbf{c}_B^\top \mathbf{B}^{-1} \mathbf{b} + (\mathbf{c}_N^\top - \mathbf{c}_B^\top \mathbf{B}^{-1} \mathbf{N}) x_N. \quad (7)$$

A basis matrix $\mathbf{B}$ is then said to be optimal

if $\mathbf{c}_N^\top - \mathbf{c}_B^\top \mathbf{B}^{-1} \mathbf{N} \geq 0$ and $\mathbf{B}^{-1} \mathbf{b} \geq 0$. The first argument is provided in Equation (7); if the value in parenthesis is nonnegative then introducing a nonbasic element to B offers no further reduction to the current objective value, which proves the optimality of the solution. Although we will not go through all the details of the algorithm, the simplex method tries to find a set of basic variables that satisfy the optimality criteria. To do so, given an initial basis, reduced costs r_j are defined $\forall j \in N$, where

$$r_j = \mathbf{c}_j - \mathbf{c}_B^\top \mathbf{B}^{-1} \mathbf{A}_j. \quad (8)$$

The primal simplex method then allows variables to enter and leave the basis while upholding $\mathbf{B}^{-1} \mathbf{b} \geq 0$, and, thus, for a given iteration if $r_j \geq 0 \forall j \in N$ the current basis is optimal. This observation is a critical step in many large-scale applications. Given the current basis, one method is to search for a new variable $x_j \forall j \in N$ to enter the basis if its reduced cost is *not* nonnegative; otherwise, the current solution is optimal. In the following sections, we investigate such applications.

COLUMN GENERATION

Many practical problems involve large-scale models where there are many more variables than constraints; hence, $n \gg m$. As mentioned in the last section, these instances may be quite difficult to solve. For such problems, column generation is an effective solution method. Column generation can also be used for problems where columns may be dynamically generated to form possible solutions. The general idea behind the algorithm is to find the subset of m basic variables that needs to be considered in the solution to the problem since nonbasic variables are equal to zero. To do so, the method forms a subset of the columns in Equations (1)–(3) called the *master problem*, which contains the columns associated with a basic feasible solution to the original problem and possibly a few additional columns. The method then searches for variables that have the potential to reduce the current objective value, which

is computed in a *subproblem* that identifies an entering variable. This implies finding a variable $x_j \forall j \in N$ with a reduced cost $r_j < 0$. The column generation algorithm proceeds as follows: first a master problem is generated involving a subset of the original variables, which contains an initial basis. Then a subproblem is solved to find an entering variable and the corresponding column is added to the master problem. The master problem is then re-solved and this process continues until the subproblem cannot identify an entering variable (i.e., $r_j \geq 0 \forall j \in N$), in which case the current solution is optimal. The master problem is defined as

$$\min \sum_{\omega \in \Psi} \mathbf{c}_\omega x_\omega \quad (9)$$

$$\text{s.t.} \quad \sum_{\omega \in \Psi} \mathbf{A}_\omega x_\omega = \mathbf{b} \quad (10)$$

$$x \geq 0, \quad (11)$$

where, for example, $\Psi = \{i \in B, j\}$, which contains all of the current basic columns and a nonbasic column $\mathbf{A}_j$. Then a subproblem is introduced that is used to seek a new column with $r_j < 0$ to add to the master problem. Re-solving the master problem will generate a new solution and the process is continued until $r_j \geq 0$ for all variables in the subproblem. We provide an example of a possible subproblem using Equation (16). One may also note that there are different techniques used to store the number of elements in set Ψ of the master problem. In the example described above, the size of Ψ is kept to a minimum of $m + 1$ variables. Other methods never remove any variables that have entered the basis, and some allow Ψ to grow to a certain size before removing variables. Using any of these sizes of Ψ and solving Equations (9)–(11) by adding columns in an iterative manner as described above is guaranteed to terminate in the absence of degeneracy since it is simply a special implementation of the revised simplex method. If degeneracy is present in the problem, one can use a method to prevent cycling such as the lexicographic rule.

Column generation is also an effective tool at solving problems where instead of

defining a large set of decisions, the problem may be more efficiently approached if it is structured to generate candidate solutions (columns) until a better solution cannot be found. Such problems exist in inventory, transportation, and scheduling. In the section titled “Route Balancing Example,” we provide an example where road crews are assigned to clean various areas of Los Angeles based on distance. As is presented, the problem can be expressed as

$$\min \mathbf{e}^\top x \quad (12)$$

$$\text{s.t. } Ax = \mathbf{b} \quad (13)$$

$$x \geq 0, \quad (14)$$

where $\mathbf{e}$ is a vector of all ones. As mentioned above, matrix A may not involve the set of all possible combinations to solve the initial problem and candidate solutions will be generated to do so. In this case, candidate solutions are equivalent to adding columns A_j to the problem. In the route balancing example, finding a candidate solution (column A_j) simply requires finding an area that a crew may be allocated to. Thus, providing an initial basis to Equations (12)–(14) and possibly a few additional columns forms the master problem. Suppose that we have an initial basis, then the subproblem will generate an additional candidate solution j with a negative reduced cost, that is,

$$r_j = 1 - \mathbf{e}^\top B^{-1}A_j < 0. \quad (15)$$

Now say that the subproblem generated columns associated with the most negative reduced cost. This is equivalent to taking the column associated with

$$\min\{r_j\} = \min \left\{ 1 - \mathbf{e}^\top B^{-1}A_j \right\}. \quad (16)$$

If the solution to the reduced cost found in Equation (16) is nonnegative, then the current solution is optimal. Notice that solving Equation (16) is equivalent to the following inequality:

$$\max \left\{ \mathbf{e}^\top B^{-1}A_j \right\} \leq 1. \quad (17)$$

Thus, the master problem is optimal if the solution to the subproblem satisfies Equation (17). In addition to Equation (17), the subproblem also involves constraints that satisfy the original problem such that only feasible solutions are considered. In some cases, the subproblem may simply require solving the knapsack problem. For more information and examples of such issues the reader may refer to Refs 6–8.

CUTTING-PLANE METHODS

In the section titled “Column Generation,” we discussed large-scale problems where $n \gg m$. We now examine the case in which the problem involves a large number of constraints. This is equivalent to investigating the dual of the model presented in the section titled “Column Generation.” In this case, we define problems where *cutting-plane* methods provide a valid solution approach. Considering the dual of Equations (1)–(3)

$$\max \mathbf{b}^\top y \quad (18)$$

$$\text{s.t. } A^\top y \leq \mathbf{c}, \quad (19)$$

where the resulting problem above has many rows or constraints. In order to solve the problem in Equations (18) and (19), the cutting-plane method defines a less-constrained problem

$$\max \mathbf{b}^\top y \quad (20)$$

$$\text{s.t. } A_\omega^\top y \leq \mathbf{c}_\omega, \quad \forall \omega \in \Omega, \quad (21)$$

where Ω is a subset of the constraints in the initial problem (i.e., $A_\omega^\top$ corresponds to a row in Equation (19)). The less-constrained problem in Equations (20)–(21) is referred to as the *relaxed* problem. The cutting-plane algorithm begins by first solving the relaxed problem, then uses a subproblem to check the feasibility of the solution with that of the initial problem in Equations (18)–(19). If a constraint is violated in the initial problem, then the violated constraint is added to subset Ω in the relaxed problem and it is re-solved. This continues until the optimal solution of the relaxed problem produces a

feasible solution with respect to the initial problem, at which point we have found an optimal solution. Thus, the algorithm continuously solves the relaxed problem with at least one additional constraint being added to Ω after each iteration, until an optimal solution is found that does not violate the initial problem. Given that y_Ω^* is an optimal solution to Equations (20)–(21), using the cutting-plane method there are two issues to consider with respect to the optimality of Equations (18)–(19), namely:

- (i) If y_Ω^* is a feasible solution to Equations (18)–(19), then $y_D^* = y_\Omega^*$; where y_D^* refers to an optimal solution of Equations (18)–(19).
- (ii) If y_Ω^* is infeasible with respect to Equations (18)–(19), then we find the violated constraint(s) and add it to the relaxed problem (Eqs 20 and 21). Thus, given that row $A_j^\top$ provides the violation, set Ω is updated to include row j and the problem is re-solved.

Figure 1 is a graphical illustration of cases (i) and (ii) above. Given that the feasible region of the initial problem is A_1 , if constraint L_4 is not included in the relaxed problem (Eqs 20 and 21), then the feasible region expands to contain regions $A_1 \cup A_2$. In this case, solving the relaxed problem produces an optimal solution to Equations (18)–(19) and, thus, $y_\Omega^* = y_D^*$. However, if constraint L_2 is not included in the relaxed problem (Eqs 20 and 21), then the feasible region contains $A_1 \cup A_3$. Depending on the slope of the objective function and the rest of the region, the optimal solution may be at vertex y_Ω^* , as depicted in the figure. In this case, $y_\Omega^* \neq y_D^*$, and violated inequality L_2 would be generated as a cutting plane.

The last issue to consider with respect to the cutting-plane algorithm is how to check that the solution is feasible with respect to the initial problem. Also, if the solution is not feasible, an efficient method to identify the violated constraint(s) is necessary. Given that y^k is the current solution to the relaxed problem, one method to identify a violated

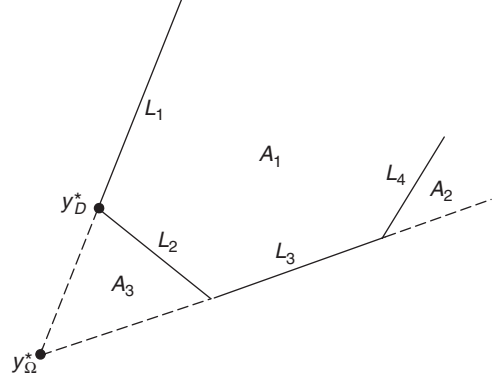


Figure 1. Illustrative example of issues related to Cutting-plane methods.

constraint(s) involves solving

$$\min_{\ell \notin \Omega} \mathbf{c}_\ell - \mathbf{A}_\ell^\top y^k \quad (22)$$

from the variables in the initial problem (Eqs 18 and 19). This forms the subproblem in the cutting-plane method, and if Equation (22) is nonnegative for all elements ℓ then we have a feasible and, therefore, an optimal solution. Otherwise, we add a constraint that satisfies the inequality $\mathbf{A}_\ell^\top y^k > \mathbf{c}_\ell$ to the relaxed problem (Eqs 20 and 21). Depending on the problem, solving Equation (22) may be challenging. There are techniques that may be employed for such instances; the reader may refer to Refs 6 and 7 for more information on this topic. One may also note that the cutting-plane method we describe in this section is different than the method used in integer programming, where constraints (cuts) are added to the problem to form the convex hull of the integer feasible set; please refer to Ref. 9 for more on this topic.

SPECIAL DECOMPOSITIONS

In addition to the methods mentioned in the sections titled “Column generation” and “Cutting-Plane Methods,” many large-scale problems may have special structures that can be further exploited. For example, some problems may contain a *sparse* $\mathbf{A}$ matrix and/or have a unique *block-matrix* structure.

In some cases it is possible to decompose the large-scale system into smaller, more tractable problems. In a simple example, consider the following problem with a block diagonal matrix structure:

$$\min \bar{\mathbf{c}}_1^\top \bar{\mathbf{x}}_1 + \bar{\mathbf{c}}_2^\top \bar{\mathbf{x}}_2 + \cdots + \bar{\mathbf{c}}_n^\top \bar{\mathbf{x}}_n \quad (23)$$

$$\text{s.t. } \hat{\mathbf{D}}_1 \bar{\mathbf{x}}_1 = \bar{\mathbf{b}}_1 \quad (24)$$

$$\hat{\mathbf{D}}_2 \bar{\mathbf{x}}_2 = \bar{\mathbf{b}}_2 \quad (25)$$

$\vdots$

$$\hat{\mathbf{D}}_n \bar{\mathbf{x}}_n = \bar{\mathbf{b}}_n \quad (26)$$

$$\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_n \geq 0, \quad (27)$$

where $\hat{\mathbf{D}}_1, \dots, \hat{\mathbf{D}}_n$ represent block matrices, and $\bar{\mathbf{b}}_1, \dots, \bar{\mathbf{b}}_n$, $\bar{\mathbf{c}}_1, \dots, \bar{\mathbf{c}}_n$, and $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_n$ represent vectors of appropriate dimension. Equations (23)–(27) can be easily decomposed into n subproblems involving $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_n$, and solved independently. The sum of the subproblem solutions will give the optimal value to Equation (23). Now consider problems that have the following structure:

$$\min \bar{\mathbf{c}}_1^\top \bar{\mathbf{x}}_1 + \bar{\mathbf{c}}_2^\top \bar{\mathbf{x}}_2 + \cdots + \bar{\mathbf{c}}_n^\top \bar{\mathbf{x}}_n \quad (28)$$

$$\text{s.t. } \hat{\mathbf{A}}_1 \bar{\mathbf{x}}_1 + \hat{\mathbf{A}}_2 \bar{\mathbf{x}}_2 + \cdots + \hat{\mathbf{A}}_n \bar{\mathbf{x}}_n = \bar{\mathbf{b}}_0 \quad (29)$$

$$\hat{\mathbf{D}}_1 \bar{\mathbf{x}}_1 = \bar{\mathbf{b}}_1 \quad (30)$$

$$\hat{\mathbf{D}}_2 \bar{\mathbf{x}}_2 = \bar{\mathbf{b}}_2 \quad (31)$$

$\vdots$

$$\hat{\mathbf{D}}_n \bar{\mathbf{x}}_n = \bar{\mathbf{b}}_n \quad (32)$$

$$\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_n \geq 0, \quad (33)$$

where $\hat{\mathbf{A}}_1, \dots, \hat{\mathbf{A}}_n$ also represent block matrices. The constraint structure presented in Equations (28)–(33) is quite common for many practical problems. For example, problems in production, finance, and transportation often involve models where some subset of constraints only pertain to particular elements/areas of the problem (e.g., time periods) and others span the whole problem domain (e.g., all time periods). For such instances, *Dantzig–Wolfe decomposition* is an efficient solution approach [6,7,10]. Similar to the methods described earlier, Dantzig–Wolfe decomposition forms a master problem of the system in Equations

(28)–(33), which contains a basis. Then, the method uses a subproblem to generate admissible columns to add to the master problem, which is re-solved until no further improvement can be made with respect to the objective function. First, the master problem is formed by defining

$$\chi_i = \left\{ \hat{\mathbf{D}}_i \bar{\mathbf{x}}_i = \bar{\mathbf{b}}_i, \bar{\mathbf{x}}_i \geq 0 \right\} \quad \forall i = 1, \dots, n, \quad (34)$$

and assuming $\chi_i \neq \emptyset \forall i = 1, \dots, n$, Equations (28)–(33) become

$$\min \sum_{i=1}^n \bar{\mathbf{c}}_i^\top \bar{\mathbf{x}}_i \quad (35)$$

$$\text{s.t. } \sum_{i=1}^n \hat{\mathbf{A}}_i \bar{\mathbf{x}}_i = \bar{\mathbf{b}}_0 \quad (36)$$

$$\bar{\mathbf{x}}_i \in \chi_i, \quad \forall i = 1, \dots, n. \quad (37)$$

Then, using the *representation theorem* [6,7,10], we have that any element $\bar{\mathbf{x}}_i \in \chi_i$ can be expressed as

$$\bar{\mathbf{x}}_i = \sum_{q \in Q_i} \lambda_i^q v_i^q + \sum_{k \in K_i} \mu_i^k d_i^k, \quad (38)$$

where $v_i^q \forall q \in Q_i$ consist of the set of extreme points in χ_i and $d_i^k \forall k \in K_i$ is the set of extreme rays in χ_i ; also note that $\lambda_i^q \geq 0$, $\mu_i^k \geq 0$, and $\sum_{q \in Q_i} \lambda_i^q = 1 \forall i = 1, \dots, n$. Therefore, from the representation theorem Equations (35)–(37) are equivalent to

$$\min \sum_{i=1}^n \left(\sum_{q \in Q_i} \bar{\mathbf{c}}_i^\top (\lambda_i^q v_i^q) + \sum_{k \in K_i} \bar{\mathbf{c}}_i^\top (\mu_i^k d_i^k) \right) \quad (39)$$

$$\text{s.t. } \sum_{i=1}^n \left(\sum_{q \in Q_i} \hat{\mathbf{A}}_i (\lambda_i^q v_i^q) + \sum_{k \in K_i} \hat{\mathbf{A}}_i (\mu_i^k d_i^k) \right) = \bar{\mathbf{b}}_0 \quad (40)$$

$$\sum_{q \in Q_i} \lambda_i^q = 1, \quad \forall i = 1, \dots, n, \quad (41)$$

$$\lambda_i^q \geq 0, \quad \forall i = 1, \dots, n, q \in Q_i, \quad (42)$$

$$\mu_i^k \geq 0, \quad \forall i = 1, \dots, n, k \in K_i, \quad (43)$$

which is the master problem of Dantzig–Wolfe decomposition. One of the key issues with the master problem is that Equations (39)–(43) typically contain more variables than the original problem in Equations (28)–(33), since a variable is created for every extreme point and extreme ray of each polyhedron $\chi_i \forall i = 1, \dots, n$. This issue can be re-solved by using column generation methods (described in the section titled “Column Generation”) to facilitate the increased number of variables. For further discussion on the Dantzig–Wolfe algorithm the reader may refer to Refs 6–8 and 10.

Taking the dual of Equations (28)–(33) forms a problem that can be solved using *Benders decomposition*. Such problems describe the structure of a two-stage stochastic linear program and can be found in many practical models where uncertainty is present. Benders decomposition is a Dantzig–Wolfe decomposition algorithm applied to the dual, and, hence, the master problem contains many constraints. The key idea in Benders decomposition is to solve a relaxed master problem similar to what was done in cutting-plane methods, described in the section titled “cutting-Plane Methods”. For more information on Benders decomposition one may refer to Refs 6–8 and 10.

ROUTE BALANCING EXAMPLE

In this section, we present an example of an LP formulation to balance the routes for road crews used by street sweepers. Several counties across California have begun to switch from diesel-powered street sweepers to compressed natural gas (CNG) street sweepers in order to comply with Federal and State air quality regulations. However, due to a number of reasons, which include slow refueling time at CNG fueling stations and some design characteristics of the CNG sweepers, the productivity of the sweeping operations have decreased with these new sweepers.

The goal of the study is to identify new operating policies to improve the productivity of the CNG street sweepers. From a recent analysis of the assigned road miles, it was evident that the workload varied significantly

from crew to crew. Hence, a better balance of the assigned miles-to-crews could significantly improve the productivity of the crews. In this section, we present a mathematical model that optimizes the assignment of road miles to crews.

Input Parameters:

Let,

- $1, 2, \dots, n$: be the possible road crews with a total of n ;
- r_{ij} : be the maximum possible road miles that crew i can shift to crew j , $\forall i = 1, 2, \dots, n, j = 1, 2, \dots, n$, where $r_{ii} = 0$;
- a_i : be the original assigned miles of crew $i = 1, 2, \dots, n$;
- u : be the average miles of the region.

Decision Variables:

Let,

- x_{ij} : be the road miles that crew i shifts to crew j , $\forall i = 1, 2, \dots, n, j = 1, 2, \dots, n$, where $x_{ii} = 0$;
- z_i : be the deviation between the workload of crew $i = 1, 2, \dots, n$ and u .

Using the variable and parameter definitions above, the road crew linear program is as follows:

$$\min \sum_{i=1}^n z_i \quad (44)$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{j=1}^n x_{ji} + a_i - \sum_{j=1}^n x_{ij} - u \leq z_i \\ & \forall i = 1, 2, \dots, n, \\ & \sum_{j=1}^n x_{ij} - a_i - \sum_{j=1}^n x_{ji} + u \leq z_i \end{aligned} \quad (45)$$

$$\forall i = 1, 2, \dots, n, \quad (46)$$

$$x_{ij} \leq r_{ij}$$

$$\forall i = 1, 2, \dots, n, j = 1, 2, \dots, n, \quad (47)$$

$$x_{ij} \geq 0$$

$$\forall i = 1, 2, \dots, n, j = 1, 2, \dots, n. \quad (48)$$

The objective in problem (44)–(48) is to minimize the deviation between the assigned miles of each crew and the average miles of the region, which is equivalent to

$$\sum_{i=1}^n \left| \sum_{j=1}^n x_{ji} + a_i - \sum_{j=1}^n x_{ij} - u \right|. \quad (49)$$

Here, variable z_i is used to represent the deviation and linear transformations represented by constraints (45) and (46). The objective function is simply the sum of the deviation of all crews. The smaller the objective value, the more balanced the routes are, and an objective value of zero means that each route in the region has exactly the same number of miles. The third constraint (47) limits the miles that crew i can shift to another crew. The data establishment of r_{ij} is based on the physical connection between two adjacent crews, since we only allow the reassignment of miles if the two crews are next to each other on the free-way. The last constraint (48) is simply a sign restriction on the nonnegativity of variable x_{ij} . The nonnegativity of variable z_i has already been ensured by the first and second constraints.

On the basis of the LP in Equations (44)–(48), we constructed an example that consists of 37 crews ($n = 37$) and has 31.25 miles average for the region ($u = 31.25$ miles). The resulting LP model contains 1406 variables and 2812 constraints, which can be easily increased in size. Thus, Equations (44)–(48) may be solved using cutting-plane methods or taking the dual and applying column generation. In addition, the problem used preassigned miles for

Table 1. Crew Road Miles

Crew	Current Solution (miles)	Optimized Solution (miles)
1	20.90	31.25
2	30.6	29.01
3	39.76	31.25
4	10.30	10.13
5	28.57	31.25
6	30.90	47.05
7	44.29	77.95
8	79.02	31.25
9	38.29	31.25
10	20.84	23.13
11	41.03	34.03
12	12.64	19.64
13	9.2	15.11
14	22.96	18.13
15	18.98	31.25
16	28.15	25.59
17	20.21	23.64
18	29.45	31.25
19	19.63	12.82
20	19.36	21.40
21	39.48	33.05
22	30.16	31.25
23	40.32	31.25
24	21.77	11.71
25	28.79	31.25
26	20.49	31.25
27	20.23	28.13
28	25.84	31.25
29	35.99	31.25
30	47.22	38.65
31	27.36	33.55
32	31.67	31.25
33	44.11	38.34
34	54.2	77.04
35	57.39	31.25
36	38.04	38.04
37	27.95	31.25

road crews; however, using column generation, alternative road-mile combinations may be generated—as presented in the section titled “Column Generation”. For the given example, Table 1 lists the assigned miles to each crew for the current and optimized solutions, which was solved using CPLEX 9.0. The optimal solution is 273.00 miles, which is a 29.39% improvement over the current actual practice of 386.64 miles.

REFERENCES

1. Gondzio J, Grothey A. Direct solution of linear systems of size 10^9 arising in optimization with interior point methods. *Lecture Notes in Computer Science*. Berlin: Springer-Verlag; 2006. pp. 513–525.
2. Leachman RC. Modeling techniques for automated production planning in the semiconductor industry. In: Ciriani TA, Leachman RC, editors. *Optimization in industry*. New York: John Wiley & Sons, Inc.; 1993.
3. Bixby RE. Solving real-world linear programs: a decade and more of progress. *Oper Res* 2002;50(1):3–15.
4. Roos C, Terlaky T, Vial J-P. *Interior point methods for linear optimization*. New York: Springer; 2006.
5. Ye Y. *Interior point algorithms: theory and analysis*. Toronto: John Wiley & Sons; 1997.
6. Bazaraa MS, Jarvis JJ, Sherali HD. *Linear programming and network flows*. Hoboken (NJ): John Wiley & Sons, Inc.; 2005.
7. Bertsimas D, Tsitsiklis JN. *Introduction to linear optimization*. Belmont (MA): Athena Scientific; 1997.
8. Winston WL. *Operations research: applications and algorithms*. Toronto: Thomson Brooks/Cole; 2004.
9. Wolsey LA. *Integer programming*. Toronto: John Wiley & Sons; 1998.
10. Dantzig GB, Thapa MN. *Linear programming 2: theory and extensions*. New York: Springer; 2003.
11. Vanderbei RJ. *Linear programming: foundations and extensions*. New York: Springer; 2001.

INTRODUCTION TO LÉVY PROCESSES

YUNAN LIU
CHIEN-CHIA L. HUANG
North Carolina State University,
Department of Industrial and
Systems Engineering, Raleigh,
North Carolina

INTRODUCTION

Lévy processes, now widely used to construct and analyze financial models, were named in honor of the French mathematician *Paul Lévy* (1886–1971). As one of the founders of modern probability theory, Lévy made major contributions to the study of Gaussian processes, stable laws, infinitely divisible distributions, and processes with independent and stationary increments, which is now known as *Lévy processes*.

Lévy processes provide ingredients for building a rich class of continuous-time stochastic processes, such as *Poisson processes* and *Brownian motions*, which are two fundamental examples. For more details, see [1] and Chapter 5 of [2]. The applications of Lévy processes have immensely been emerging in many areas, for instance, queuing theory and financial engineering.

In this review article, we intend to give an introductory lecture and provide fundamental results on Lévy processes. In the section titled “Preliminary,” we state preliminaries on probability theory and stochastic processes; in the section titled “Lévy Processes,” we formally introduce Lévy processes and their distributional properties; in the section titled “Examples of Lévy Processes,” we illustrate five important examples of Lévy processes; in the section titled “Poisson Random Measures,” we present *Poisson random measures* (PRMs), the building blocks of the pure-jump part of Lévy processes; in the sections titled “Lévy–Itô Decomposition” and

“Lévy–Khintchine Formula,” we introduce the two most important theorems: *Lévy–Itô decomposition* and *Lévy–Khintchine formula*; and finally in the section titled “Path Properties,” we introduce four kinds of Lévy processes distinguished by different path properties.

PRELIMINARY

Probability Space and Random Variables

Let the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ denote a *probability space*, where $\mathcal{F}$ is the σ -field (σ -algebra) of the underlying *sample space* Ω , and $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is the *probability measure* associated. A *random variable* is a real-valued $\mathcal{F}$ -measurable function $X : \Omega \rightarrow \mathbb{R}$, where $\mathcal{F}$ -measurability means that for any set B in the Borel σ -field $\mathcal{B}$, we have $X^{-1}(B) \in \mathcal{F}$. Note that the random variable X induces a probability measure μ_X on $\mathbb{R}$, defined by $\mu_X(B) \equiv \mathbb{P}(X^{-1}(B)) \equiv \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$ for any Borel set B in $\mathcal{B}$.

Let $F_X(\cdot) \equiv \mathbb{P}(X \leq \cdot)$ and $F_X^c \equiv 1 - F_X$ be the *cumulative distribution function (CDF)* and *complement cumulative distribution function (CCDF)* of a random variable X . Two random variables X and Y are said to be *equal in distribution*, denoted by $X \stackrel{D}{=} Y$, if $F_X = F_Y$. Two random variables X and Y are said to be *independent*, denoted by $X \perp Y$, if $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$ for all $x, y \in \mathbb{R}$. Given the independence, the distribution (induced measure) of $X + Y$ is given by convolution of measures, that is, $\mu_{X+Y} = \mu_X * \mu_Y$. $\mu_X * \mu_Y(A) \equiv \int_{\mathbb{R}} \mu_X(A - y)\mu_Y(dy)$, $A \in \mathcal{B}$ and $\mu_Y(dy) \equiv \mathbb{P}(\{\omega \in \Omega : y < Y(\omega) < y + dy\})$.

Let $\mathbb{E}[X] \equiv \int_{\Omega} X d\mathbb{P} = \int_{\mathbb{R}} x \mu_X(dx) = \int_{\mathbb{R}} x dF(x)$ be the *expectation* of X , provided that $\int_{\mathbb{R}} |x| dF(x) < \infty$. Note that, if two random variables X and Y are independent, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Characteristic Functions

The *characteristic function (CF)* of a random variable X is defined by $\Psi_X(\theta) \equiv \mathbb{E}[e^{i\theta X}]$, with

$\theta \in \mathbb{R}$ and $i \equiv \sqrt{-1}$. It has the following properties: (i) $|\Psi_X(\theta)| \leq 1$; (ii) $\Psi_X(\theta)$ is real-valued if and only if X is symmetric; (iii) $\Psi_X(\theta)$ is Hermitian, that is, $\Psi_X(-\theta) = \overline{\Psi_X(\theta)}$; and (iv) $\mathbb{E}[X^n] = i^{-n} \frac{d^n}{d\theta^n} \Psi_X(\theta) \Big|_{\theta=0}$. Note that the CF uniquely determines the distribution of a random variable.

The independence among the random variables can be understood through their CFs. Namely, the random variables $X_1, \dots, X_n$ are independent if and only if

$$\mathbb{E} \left[e^{i \sum_{k=1}^n \theta_k X_k} \right] = \prod_{k=1}^n \Psi_{X_k}(\theta_k). \quad (1)$$

Counting Processes and Counting Measures

A *stochastic process* $\{X(t), t \geq 0\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a parameterized collection of random variables $\{X(t)\}_{t \geq 0}$, assuming values in $\mathbb{R}$, that is, $X : [0, \infty) \times \Omega \rightarrow \mathbb{R}$. For each fixed $\omega \in \Omega$, $\{X(t, \omega), t \geq 0\}$ a sample path of the process, is a deterministic function and, for each fixed $t \geq 0$, $X(t, \cdot)$ is a random variable. A *counting process* $\{N(t), t \geq 0\}$ is a stochastic process that counts the total number of events occurred, and thus must be (i) nonnegative; (ii) integer-valued; and (iii) nondecreasing. Note that to understand (iii), we interpret $N(t) - N(s)$ as the total number of events occurred within time interval $(s, t]$.

A *counting measure* can be viewed as an induced measure by the associated counting process. Let $0 \leq T_1 \leq T_2 \leq \dots$ be the sequence of the occurrence times of events. For a measurable set $A \subset [0, \infty)$, define the measure $M(\omega, A) \equiv \#\{i \geq 1, T_i(\omega) \in A\}$. Note that the measure $M(\omega, \cdot)$ depends on $\omega \in \Omega$; it is thus a *random measure*. More details on random measures will be discussed in the section titled “Poisson Random Measures.”

LÉVY PROCESSES

A *random walk*, sum of independent and identically distributed (IID) random variables, provides the simplest example of discrete-time stochastic processes. We hereby introduce the definition of *Lévy processes*, the continuous-time analogs of random walks.

Definition 1. [Lévy Processes]. A continuous-time process $\{X(t), t \geq 0\}$ is called a *Lévy process* if

- (i) $X(0) = 0$;
- (ii) it has *independent* increments, that is, $X(t_4) - X(t_3) \perp X(t_2) - X(t_1)$ for all $0 < t_1 < t_2 < t_3 < t_4$;
- (iii) it has *stationary* increments, that is, $X(t_2) - X(t_1) \stackrel{D}{=} X(t_2 - t_1)$ for all $0 < t_1 < t_2$;
- (iv) its path is *stochastically continuous*, that is, $\lim_{s \rightarrow t} \mathbb{P}(|X(t) - X(s)| > \epsilon) = 0$, for $\epsilon > 0$.

Note that the *stochastic continuity* condition in property (iv) does not imply that sample paths of Lévy processes are continuous. For instance, *Poisson processes*, special cases of Lévy processes, are pure-jump processes. Property (iv) serves to rule out the processes with discontinuities at deterministic times (known as the “calender effect”).

Infinite Divisibility

A distribution F of a random variable X is called *infinitely divisible* if there exists n IID random variables $Y_{n,1}, \dots, Y_{n,n}$ such that the sum $\sum_{i=1}^n Y_{n,i}$ follows distribution F for all $n \geq 2$.

Famous examples of infinitely divisible distributions are *Poisson*, *Gaussian*, and *Gamma*. A random variable X following the above distributions can be expressed by a sum of n IID random variables following the same distribution but with modified parameter (dependent on n). For instance, a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is equal in distribution to $\sum_{i=1}^n Y_i$, where $Y_1, \dots, Y_n$ are IID following $\mathcal{N}(\mu/n, \sigma^2/n)$. Other examples are *Pareto*, *lognormal*, *Cauchy*, *stable*, and *student* distributions.

Distributional Properties of Lévy Processes

Lévy processes possess the *infinitely divisible* property. Consider a fixed time $t > 0$, we have

$$X(t) = \sum_{k=1}^n Y_k^{(t)}, \text{ where } Y_k^{(t)} \equiv X\left(\frac{kt}{n}\right) - X\left(\frac{(k-1)t}{n}\right),$$

denotes the increment of the process in the interval $((k-1)t/n, kt/n]$, $1 \leq k \leq n$. The infinite divisibility of $X(t)$ easily follows by properties (ii) and (iii) of the definition of Lévy processes.

Following Equation (1) and the definition of infinite divisibility, the CF of $X(t)$ with $t = n$ can be written as

$$\Psi_{X(n)}(\theta) = \prod_{k=1}^n \Psi_{Y_k^{(n)}}(\theta) = (\Psi_{X(1)}(\theta))^n,$$

where the last equality holds as $Y_k^{(n)} \stackrel{D}{=} Y_1^{(n)} = X(n/n) = X(1)$. We can then easily extend the above argument from integer-valued t to all $t > 0$. Therefore, for all $t > 0$, we can write the CF

$$\Psi_{X(t)}(\theta) = (\Psi_{X(1)}(\theta))^t \equiv e^{-t\psi(\theta)},$$

where the function ψ is called the *characteristic exponent* (CE) of the Lévy process $\{X(t), t \geq 0\}$. As $\Psi_{X(1)}(\theta) = e^{\psi(\theta)}$, the distribution (law) of the whole process is determined solely by the CE ψ , or equivalently by the distribution of the random variable $X(1)$. It is not hard to prove that for a distribution F that is infinitely divisible, there exists a Lévy process $\{X(t), t \geq 0\}$ such that the $X(1)$ follows F .

EXAMPLES OF LÉVY PROCESSES

We next introduce some concrete examples of Lévy processes. Some of these examples are fundamental building blocks to construct general Lévy processes.

Poisson Processes

A Lévy process $\{N(t), t \geq 0\}$ is a *Poisson process* with rate $\lambda > 0$ if, for a fixed t , the random variable $N(t)$ follows Poisson distribution with mean λt , that is, the *probability mass function* (PMF) $\mathbb{P}(N(t) = k) = e^{-\lambda t}(\lambda t)^k / k!$, for $k = 0, 1, 2, \dots$

Poisson processes are pure-jump processes with positive unit jumps and are thus counting processes. The intertransition times between these jumps $T_1, T_2, \dots$ form a sequence of IID *exponential* distribution with rate λ , having a *probability density*

function (PDF) $f_{T_1}(x) = \lambda e^{-\lambda x} \mathbf{1}_{\{x \geq 0\}}$. The CE of a Poisson process with rate λ is given by

$$\psi(\theta) = \lambda(1 - e^{i\theta}). \quad (2)$$

Besides Lévy processes, Poisson processes are special cases of several other classical processes, such as *renewal processes* and *continuous-time Markov chains*.

Compound Poisson Processes

Given a Poisson process $\{N(t), t \geq 0\}$ and a sequence of IID random variables $\{\xi_i : i \geq 1\}$ that are independent with N , then the process $\{Y(t), t \geq 0\}$ with $Y(t) \equiv \sum_{i=1}^{N(t)} \xi_i$ is called a *compound Poisson process*. Compound Poisson processes generalize Poisson processes from unit jump sizes to general and random jump sizes. In general, compound Poisson processes are no longer counting processes because there may be negative jumps. Let G be the CDF of ξ_1 , which describes the sizes of the jumps.

The fact that compound Poisson processes have stationary and independent increments simply follows from those properties of Poisson processes. The CE is given by

$$\psi(\theta) = \lambda \int_{\mathbb{R}} (1 - e^{i\theta y}) dG(y). \quad (3)$$

Note that Equation (3) reduces to Equation (2) when the measure induced by ξ_1 degenerates to $\delta_1(\cdot)$, the dirac measure at point 1.

Brownian Motions

A Lévy process with continuous sample path (with probability 1) $\{W(t), t \geq 0\}$ is called a standard *Brownian motion* when, for a fixed t , the random variable $W(t) \sim \mathcal{N}(0, t)$, that is, $W(t)$ follows a Gaussian distribution with mean 0 and variance t . Let

$$X(t) \equiv a W(t) + b t. \quad (4)$$

Then $\{X(t), t \geq 0\}$ is called a *Brownian motion with drift*. It is obvious that $X(t) \sim \mathcal{N}(bt, a^2 t)$ with PDF

$$f_{X(t)}(x) = \frac{1}{\sqrt{2\pi a^2 t}} e^{-\frac{(x-bt)^2}{2a^2 t}}.$$

The CE of X is given by

$$\psi(\theta) = a^2\theta^2/2 - i\theta b. \quad (5)$$

Inverse Gaussian Processes

Regarding a standard Brownian motion with drift $\{X(t), t \geq 0\}$ defined in Equation (4) with $a = 1$ and $b > 0$, define the first passage time

$$\tau(t) \equiv \inf\{s > 0 : X(s) > t\}, \quad (6)$$

that is, the first time a Brownian motion crosses above the level $t \geq 0$. Then the process $\{\tau(t), t \geq 0\}$ is called an *inverse Gaussian process*.

We use the so-called *strong Markov* property of Brownian motions to show that $\{\tau(t), t \geq 0\}$ indeed is a Lévy process. Define $X^*(t) \equiv X(\tau(s) + t) - s$, then the new process $\{X^*(t), t \geq 0\}$, describing the evolution of the Brown motion after hitting level s , is equal in distribution to $\{X(t), t \geq 0\}$. The dynamics of X^* is also independent of $X(u)$ for $0 \leq u \leq \tau(s)$. For $0 \leq s < t$ let $\tau^*(t-s)$ be the first passage time (in the form of Equation (6)) of X^* at level $t-s$. We thus have that the remaining time to level t (given hitting s)

$$\begin{aligned} \tau(t) - \tau(s) &= \tau^*(t-s) \stackrel{\mathcal{D}}{=} \tau(t-s) \\ \text{and } \tau(t) - \tau(s) &\perp \tau(s), \end{aligned}$$

which prove that the inverse Gaussian process has stationary and independent increments.

The PDF of $\tau(t)$ is given by

$$f_{\tau(t)}(x) = \frac{t}{\sqrt{2\pi x^3}} e^{tb} e^{-\frac{1}{2}(t^2 x^{-1} + b^2 x)},$$

and the CE takes the form

$$\psi(\theta) = \sqrt{-2i\theta + b^2} - b.$$

Furthermore, the inverse Gaussian process is an example of a *subordinator*, a subclass of Lévy processes having nondecreasing paths (see section titled “Path Properties”).

Stable Processes

A random variable Y is said to follow a *stable* distribution if, for all $n \geq 1$,

$$n^{1/\alpha} Y + b_n \stackrel{\mathcal{D}}{=} \sum_{i=1}^n Y_i, \quad (7)$$

where $Y_1, \dots, Y_n$ are IID copies of Y , $\alpha \in (0, 2]$ and $b_n \in \mathbb{R}$. We say the distribution is *strictly stable* if $b_n = 0$. By subtracting b_n/n from each term of the right-hand side of Equation (7) and divide both sides by $n^{1/\alpha}$, one can easily verify that stable distributions are infinitely divisible. *Stable processes* thus form one class of Lévy processes whose CEs correspond to those of stable distributions.

We point out that the case $\alpha = 2$ corresponds to zero mean Gaussian random variables. For $\alpha \in (0, 2)$, CE has the form

$$\begin{aligned} \psi(\theta) &\equiv c|\theta|^\alpha \Gamma(\theta, \alpha, \beta) + i\theta\eta, \quad \text{where} \\ \Gamma(\theta, \alpha, \beta) &\equiv \begin{cases} 1 - i\beta \tan(\frac{\pi\alpha}{2}) \text{sgn}(\theta), & \text{if } \alpha \neq 1, \\ 1 + i\beta \frac{2}{\pi} \text{sgn}(\theta) \log|\theta|, & \text{if } \alpha = 1, \end{cases} \end{aligned}$$

where $-1 \leq \beta \leq 1$, $c > 0$, $\eta \in \mathbb{R}$ and the sign function $\text{sgn}(x) \equiv \mathbf{1}_{\{x > 0\}} - \mathbf{1}_{\{x < 0\}}$.

POISSON RANDOM MEASURES

Because Poisson and compound Poisson processes play important roles in the construction of Lévy processes, we next introduce the concept PRM that is exploited to study the jump structure of Lévy processes.

Definition 2. [Poisson Random Measure]. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $E \subset \mathbb{R}^d$. Given a Radon measure μ (i.e., $\mu(A) < \infty$ for any compact set A). A PRM on E with intensity measure μ is an integer-valued random measure such that

- (i) for $\omega \in \Omega$, $M(\omega, \cdot)$ is an integer-valued Radon measure on E ;
- (ii) for each measurable set $A \subset E$, $M(A) \equiv M(\cdot, A)$ is a Poisson random variable with mean $\mu(A)$;

- (iii) for disjoint measurable sets $A_1, \dots, A_n$, $M(A_1), \dots, M(A_n)$ are independent Poisson random variables.

We can construct a PRM with any given μ . Assume $\mu(E) < \infty$ without loss of generality, we (i) first generate IID random variables $X_1, X_2, \dots$ with $\mathbb{P}(X_i \in A) = \frac{\mu(A)}{\mu(E)}$; (ii) second generate an independent Poisson random variable $M(E)$ on $(\Omega, \mathcal{F}, \mathbb{P})$ with mean $\mu(E)$; (iii) and finally let $M(A) \equiv \sum_{i=1}^{M(E)} \mathbf{1}_{\{X_i \in A\}}$.

Constructing Jump Processes Using PRM

PRMs can be employed to construct jump processes. Consider $E = [0, T] \times \mathbb{R} \setminus \{0\}$, we can write

$$M = \sum_{n \geq 1} \delta_{(T_n, \xi_n)}, \quad \text{or} \quad M(E) = \sum_{n \geq 1} \mathbf{1}_{\{(T_n, \xi_n) \in E\}},$$

where $(T_n)_{n \geq 1}$ is an increasing sequence. Each point $[T_n(\omega), \xi_n(\omega)] \in [0, T] \times \mathbb{R} \setminus \{0\}$ corresponds to jump with size $\xi_n(\omega)$ at time $T_n(\omega)$. The second expression counts the total number of jumps in $[0, T]$. We hereby exclude 0 so that every jump has a nonzero size.

By integrating a measurable function f with respect to the PRM M , we can construct a jump process

$$\begin{aligned} X(t) &= \int_0^t \int_{\mathbb{R} \setminus \{0\}} f(s, y) M(ds \times dy) \\ &= \sum_{T_n \in [0, t]} f(T_n, \xi_n), \end{aligned}$$

The above jump process becomes the compound Poisson process introduced in the section titled “Examples of Lévy Processes” when $f(s, y) = y$ and $\mu(ds \times dy) = \lambda \mu_{\xi_1}(dy)ds$, where μ_{ξ_1} is the measure induced by ξ_1 . It becomes a Poisson process when $\mu_{\xi_1} = \delta_1$, see section titled “Examples of Lévy Processes.”

LÉVY–ITÔ DECOMPOSITION

We are now ready to present the most fundamental result of Lévy processes, the *Lévy–Itô*

decomposition. The Lévy–Itô decomposition enables us to understand the path structures of a general Lévy process by decomposing the Lévy process into three independent auxiliary Lévy processes.

Consider a Lévy process $\{X(t), t \geq 0\}$. We let $\Delta X(t) \equiv X(t+) - X(t-)$ be the size of a jump, if any, at time t and let A be a measurable set in $\mathbb{R}$. A *jump measure* is defined by

$$M_X(B) \equiv \#\{[t, \Delta X(t)] \in B\}, \quad (8)$$

for a measurable set $B \subset [0, \infty) \times \mathbb{R}$. Note that $M_X([t_1, t_2] \times A)$ counts the number of jumps of the Lévy process X in $[t_1, t_2]$ with jump sizes in A . The *Lévy measure* associated with X is defined as

$$\nu_X(A) \equiv \mathbb{E} \left[\sum_{0 \leq t \leq 1} \mathbf{1}_{\{\Delta X(t) \neq 0, \Delta X(t) \in A\}} \right], \quad (9)$$

which describes the expected number of jumps with sizes in A , per unit time. Obviously, we have

$$\nu_X(A) = \mathbb{E}[M_X([0, 1] \times A)]. \quad (10)$$

Lévy–Itô Decomposition

A Lévy process $\{X(t), t \geq 0\}$ can be written as

$$X(t) = X^c(t) + X^l(t) + X^s(t), \quad (11)$$

where $X^c(t)$, $X^l(t)$, and $X^s(t)$ are three independent processes; $X^c(t) \equiv bt + aW(t)$ is a Brownian motion with drift as defined in Equation (4); the two terms $X^l(t)$ and $X^s(t)$ are defined by

$$X^l(t) \equiv \int_0^t \int_{|x| \geq 1} x M_X(du \times dx) \quad (12)$$

and

$$X^s(t) \equiv \lim_{\epsilon \rightarrow 0} \int_0^t \int_{\epsilon \leq |x| < 1} x \tilde{M}_X(du \times dx), \quad (13)$$

with $\tilde{M}_X(du \times dx) \equiv M_X(du \times dx) - \nu_X(dx)du$ being the centered (compensated) version of the jump measure M_X , as defined in

Equation (8); the Lévy measure ν_X , as defined in Equation (9), is a Radon measure on $\mathbb{R} \setminus \{0\}$ such that

$$\int (|y|^2 \wedge 1) \nu_X(dy) < \infty; \quad (14)$$

and $a, b \in \mathbb{R}$ are constants.

To better understand this important result, we remark on the meanings of the three terms of the right-hand side in Equation (11). First, the Brownian motion with drift $X^c(t)$, described by the drift b and volatility term a , characterizes the continuous part of the path of a Lévy process.

The second term in Equation (11), as elaborated in Equation (12), is a compound Poisson process with a finite number of jumps with size > 1 . Note that the finiteness here is guaranteed by condition Eq. (14). Here the threshold separating big jumps and small jumps is set to be 1 as a convention. Changing the threshold from 1 to an arbitrary constant $c > 0$ results in a refinement of the drift term b . See [3] for alternative representations with a general c .

Finally, as the measure ν_X may have a singularity at 0, there can be infinitely many small jumps whose sum does not necessarily converge. This prevents us from letting ϵ go to 0 directly for a compound Poisson process with amplitude between ϵ and 1. The third term in Equation (11), as elaborated in Equation (13), thus has to be centered by the measure $\nu_X(dx)du$ to obtain convergence.

In summary, the Lévy–Itô decomposition implies that an arbitrary Lévy process can be decomposed into the sum of three independent components: (i) a Brownian motion with drift, (ii) a compound Poisson process with finite and big jumps, and (iii) a discontinuous process with an infinite number of small jumps. The distribution of a Lévy process is thus characterized by three parameters a , b , and ν_X .

LÉVY–KHINTCHINE FORMULA

Using the Lévy–Itô decomposition, we can quickly obtain the second fundamental result: the *Lévy–Khinchine formula*.

Lévy–Khinchine Formula

Consider a Lévy process $\{X(t), t \geq 0\}$ with parameter (a, b, ν_X) , its CE has the form

$$\begin{aligned} \psi(\theta) = & \left(\frac{1}{2} a^2 \theta^2 - i b \theta \right) + \int_{|x| \geq 1} (1 - e^{i \theta x}) \nu_X(dx) \\ & + \int_{|x| < 1} (1 - e^{i \theta x} + i \theta x) \nu_X(dx). \end{aligned} \quad (15)$$

This result can be quickly proved using the Lévy–Itô decomposition and Equation (1). It can be easily seen from Equations (5) and (3) that the first two terms of the right-hand side of Equation (15) are the CE's of a Brownian motion with drift and a compound Poisson process with jump amplitudes greater than 1, corresponding to the first two terms of the right-hand side in Equation (11). In addition, the third term in Equation (15) is the CE of Equation (13).

PATH PROPERTIES

In the section titled “Lévy–Itô Decomposition,” the distribution of a Lévy process $\{X(t), t \geq 0\}$ is characterized by its parameters a , b , and ν_X . We now introduce four typical kinds of Lévy processes distinguished by different path properties, reflected by different parameters.

Continuous Paths

As both X^l and X^s in Equation (11) represent jumps, a Lévy process having a continuous sample path must be a Brownian motion with drift. As a result, we must have only the first term in Equation (11) with $\nu_X = 0$ and Equation (15) thus degenerates to Equation (5).

Piecewise Constant Paths

On the contrary, as X^c in Equation (11) represents the continuous part, a pure-jump Lévy process (thus having piecewise constant paths) must be a compound Poisson process. Hence, the Lévy process involves only the second two terms in Equation (11) with $a = b = 0$ and $\nu_X(\mathbb{R}) < \infty$, see Ref. 4 for more details. Consequently, Equation (15) thus

degenerates to Equation (3) with $\nu_X(dx) = \lambda dG(x)$.

Finite Variation

Consider an interval $[0, T]$, a function f is said to be of *finite variation* if, for each partition $0 = t_0 < t_1 < \dots < t_{n-1} < t_n = T$, the total variation thereof is finite, that is,

$$TV(f, [0, T]) \equiv \sup_{\{t_k\}_{k=1}^n} \sum_{i=1}^n |f(t_i) - f(t_{i-1})| < \infty.$$

As Brownian motions have infinite variation over any finite interval [5], a Lévy process with finite variation must not involve the Brownian term W , that is, we have $a = 0$ in Equation (15). In addition, as compound Poisson processes [denoted by X^I in Equation (12)] always have finite variation, we impose extra conditions such that X^s has finite variation. We require $\int_{|x| \leq 1} |x| \nu_X(dx) < \infty$. In this case, Equation (11) simplifies to

$$X(t) = \tilde{b}t + \int_0^t \int_{x \in \mathbb{R}} x M_X(du \times dx),$$

$$\text{where } \tilde{b} \equiv b - \int_{|x| < 1} x \nu_X(dx).$$

Subordinators

A Lévy process having nondecreasing paths over time is called a *subordinator*. In this case, the Brownian term is again not involved because the path of a Brownian motion is not monotone. Hence, $a = 0$. In addition, we require the Lévy process to have merely positive jumps of finite variation and positive drift, that is, $\nu_X((-\infty, 0]) = 0$, $\int (x \wedge 1) \nu_X(dx) < \infty$ and $b \geq 0$. See Ref. 3 for more discussion on subordinators.

FURTHER READING

More detailed contents on Lévy processes can be referred to the following books [3, 4, 6–8]. Readers who are interested in infinite divisibility properties are referred to [9]. Regarding applications of Lévy processes in queuing theory, we refer to the book [10] and a recent survey on Lévy-driven queues [11].

REFERENCES

1. Yao DD. Brownian Motion and Queueing Applications, Encyclopedia of Operations Research and Management Science. Wiley-Interscience; 2010.
2. Insua DR, Ruggeri F, Wiper MP. Bayesian analysis of stochastic process models. Wiley-Interscience; 2012.
3. Cont R, Tankov P. Financial modelling with jump processes. New York: Chapman & Hall/CRC; 2004.
4. Applebaum D. Lévy processes and stochastic calculus. New York: Cambridge University Press; 2004.
5. Øksendal B. Stochastic differential equations: an introduction with applications. 6th edn. Heidelberg: Springer-Verlag; 2004.
6. Barndorff-Nielsen OE, Mikosch T, Resnick SI. Lévy processes: theory and applications. Boston (MA): Birkhäuser; 2001.
7. Bertoin J. Lévy processes. Cambridge: Cambridge University Press; 1998.
8. Kyprianou AE. Introductory lectures on fluctuations of Lévy processes with applications. Berlin: Springer-Verlag; 2006.
9. Meerschaert MM, Scheffler H-P. Limit distributions for sums of independent random vectors. New York: Wiley-Interscience; 2001.
10. Whitt W. Stochastic-process limits. New York: Springer-Verlag; 2002.
11. Debicki K, Mandjes M. Lévy-driven queues. Surv Oper Res Manage Sci 2012;17(1): 15–37.

INTRODUCTION TO MULTIATTRIBUTE UTILITY THEORY

JOHN C. BUTLER
and JAMES S. DYER
McCombs School of Business,
University of Texas at Austin,
Austin, TX

INTRODUCTION

In this article, we provide a review of multiattribute utility (MAU) theory. We begin with a brief review of single-attribute utility theory, and explore preference representations that measure a decision maker's preferences for risky alternatives. Following convention (Keeney and Raiffa, 1993),¹ we adopt the term utility function for preferences in risky situations and use the term value function for preferences measured based on ordinal comparisons and strength of preference. Next, we characterize the multiattribute decision problem, and discuss the conditions that allow a MAU function to be decomposed into three common functional forms. Finally, the relationships between multiattribute preference functions under conditions of certainty and risk are discussed, and we survey issues related to their assessment and applications.

MAU theory provides an axiomatic foundation for choices involving multiple criteria. As a result, one can examine these axioms and determine whether or not they are reasonable guides to rational behavior. Often arguments are made that decision makers do not always make decisions that are consistent with these axioms, and, therefore, the axioms should not be viewed as necessary for a theory of choice. While this may be

true as a descriptive statement for individual decision-making, it is much more difficult to identify situations involving significant implications for other parties where a cavalier disregard for normative theories of choice can be defended. Further, theories that do not provide consistent recommendations can be compared on first principles by evaluating the underlying axioms.

UTILITY THEORY: PREFERENCE REPRESENTATIONS UNDER RISK

Utility theory studies the fundamental aspects of individual choice behavior, such as how to identify and quantify an individual's preferences over a set of alternatives, and how to construct appropriate preference representation functions for decision-making under risk. An important feature of utility theory is that it is based on rigorous axioms, which characterize an individual's choice behavior. These preference axioms are essential for establishing preference representation functions, and provide the rationale for the quantitative analysis of preference.

Utility theory is primarily concerned with properties of a binary preference relation $\succ$ on a choice set P , where P is a convex set of probability distributions or lotteries $\{p, q, r, \dots\}$ on a nonempty set X of outcomes. As an example, we might present an individual with a pair of alternatives, say p and q (e.g., two gambles with one nonzero outcome from the set X of monetary values) where $p, q \in P$ (e.g., the set of all monetary gambles with one nonzero outcome), and ask how they compare (e.g., do you prefer p or q ?). For example, p could be a fair coin flip that pays \$10 for heads and \$0 for tails and q could be a fair die roll that pays \$12 for a number larger than 4 and \$0 otherwise. If the individual says that p is preferred to q , then we write $p \succ q$, where $\succ$ means strict preference. If the individual states that he or she is indifferent between p and q , then we represent this preference as $p \sim q$. Alternatively, we can define $\sim$ as the absence of strict preference; that is, not $p \succ$

¹*Decisions with Multiple Objectives* by R. L. Keeney and H. Raiffa was originally published by Wiley in 1976. The Cambridge University Press version was published in 1993.

q and not $q > p$. If it is not the case that $q > p$, then we write $q \succsim p$, where $\succsim$ represents a weak preference (or preference-indifference) relation. We can also define $\succsim$ as the union of strict preference $>$ and indifference $\sim$; that is, both $q > p$ and $p \sim q$.

Preference studies begin with some basic assumptions (or axioms) of individual choice behavior. First, it seems reasonable to assume that an individual can state preference over a pair of alternatives without contradiction; that is, the individual does not strictly prefer p to q and q to p simultaneously. This leads to the following definition for *preference asymmetry*: preference is asymmetric if there is no pair p and q in P such that $p > q$ and $q > p$. In other words, asymmetry ensures preference consistency.

Furthermore, if an individual makes the judgment that p is preferred to q , then he or she should be able to place any other alternative r somewhere on the ordinal scale determined by the following: either better than q or worse than p , or both. Formally, we define *negative transitivity* by saying that preferences are negatively transitive if given $p > q$ in P and any third element r in P , it follows that either $p > r$ or $r > q$, or both.

If the preference relation $>$ is asymmetric and negatively transitive, then it is called a *weak order*. The weak order assumption implies some desirable properties of a preference ordering and is a basic assumption in many preference studies. If the preference relation $>$ is a weak order, then the associated indifference and weak preference relationships are well behaved. More specifically, if strict preference $>$ is a weak order, then

1. strict preference $>$ is *transitive* (if $p > r$ and $r > q$, then $p > q$);
2. indifference $\sim$ is *transitive*, *reflexive* ($p \sim p$ for all p); and *symmetric* ($p \sim q$ implies $q \sim p$);
3. exactly one of $p > q$, $q > p$, or $p \sim q$ holds for each pair p and q ; and
4. weak preference $\succsim$ is *transitive* and *complete* (for a pair p and q , either $p \succsim q$ or $q \succsim p$).

Thus, an individual whose strict preference can be represented by a weak order can specify a unique ranking of all alternatives under consideration. Further discussions of the properties of binary preference relations are presented in Refs 1 and 2.

Perhaps the most significant contribution to the theory of risky choice was the formalization of *expected utility theory* by von Neumann and Morgenstern [3]. This development has been refined by a number of scholars and is most commonly presented in terms of three basic axioms [1].

For lotteries p, q, r in P and all $\lambda, 0 < \lambda < 1$, the expected utility axioms are:

1. *Ordering*: $>$ is a weak order;
2. *Independence*: If $p > q$, then $(\lambda p + (1-\lambda)r) > (\lambda q + (1-\lambda)r)$ for all $r \in P$;
3. *Continuity*: If $p > r > q$, then there exist some $0 < \alpha < 1$ and $0 < \beta < 1$ such that $(\alpha p + (1-\alpha)q) > r > (\beta p + (1-\beta)q)$.

The von Neumann-Morgenstern expected utility theory asserts that the above axioms hold if and only if there exists a real-valued function u defined on the set X of outcomes such that for all p, q in P , $p > q$ if and only if

$$\sum_{x \in X} p(x)u(x) \geq \sum_{x \in X} q(x)u(x)$$

where $p(x)$ and $q(x)$ are the probabilities of outcome x for lotteries p and q , respectively. Moreover, such a u is unique up to a positive linear transformation.

The expected utility model can also be used to characterize an individual's risk attitude (Chapter 4, [4, 5]). If an individual is indifferent between the expected value of a lottery and the lottery itself, then we say that the individual is risk neutral. Intuitively, the "risk" associated with the gamble does not have an impact on this individual's preferences, and, therefore, he or she perceives the value of the risky alternative to be equal to its actuarial expected value. However, if he or she feels that his or her subjective assessment of the value of the lottery, known as his or her certainty equivalent for the lottery, is less than its expected value, then we say that he or she is risk averse. Conversely,

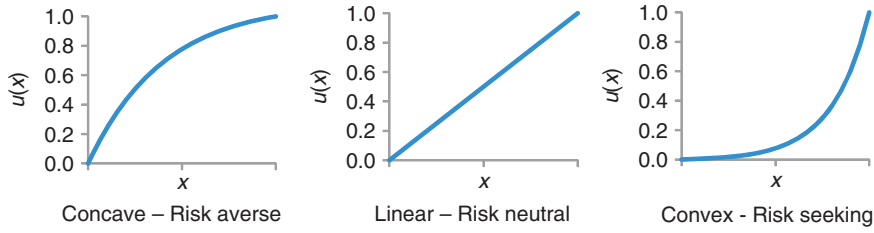


Figure 1. Example single-attribute utility function shapes.

if he or she feels that his or her certainty equivalent for the lottery is more than its expected value, we say that he or she is risk seeking.

A utility function defined on the outcome set X may be plotted on any arbitrary scale, but typically the range from 0 to 1 is used. If an individual is risk averse, risk neutral, or risk seeking over the range of X , then the individual's utility function is concave, linear, or convex, respectively. Examples of these utility function shapes are shown in Fig. 1.

The von Neumann-Morgenstern theory of risky choice presumes that the probabilities of the outcomes of lotteries are provided to the decision maker. Savage [6] extended the theory of risky choice to allow for the simultaneous determination of subjective probabilities for outcomes and a utility function u defined over those outcomes. The probabilities in Savage's model are personal or subjective probabilities, and the model itself is a subjective expected utility representation.

The assessment of von Neumann-Morgenstern utility functions will almost always involve the introduction of risk in the form of simple lotteries. Two basic risky stimuli are shown in Fig. 2. In the certainty equivalent assessment method (left panel of Fig. 2), the decision maker is asked to provide a sure amount $\$X$ such that he or she is indifferent to receiving $\$X$ for sure and receiving the lottery that has an expected monetary value of $\$140$. A risk-neutral decision maker would specify $\$X = 140$; risk aversion would be consistent with $\$X < 140$ and a risk seeker would specify $\$X > 140$. The right panel of Fig. 2 depicts the probability equivalent

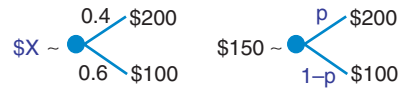


Figure 2. Utility function assessment examples.

method for utility assessment where the decision maker specifies a probability p that leads to equivalence between $\$150$ and the stimulus lottery. A risk-neutral decision maker would specify $p = 0.5$, whereas risk-averse (risk-seeking) utility functions would be consistent with $p > (<) 0.5$. To assess a utility function, these questions would be repeated for different stimuli that cover the range of dollar outcomes. For a more general discussion of these assessment approaches, see [5] or [7].

As a normative theory, the expected-utility model has played a major role in the prescriptive analysis of decision problems. However, for descriptive purposes, the assumptions of this theory have been challenged by empirical studies (e.g., [8]). Some of these empirical studies demonstrate that subjects may choose alternatives that imply a violation of the independence axiom. One implication of the independence axiom is that the expected utility model is "linear in probabilities." For a discussion, see Fishburn and Wakker [9]. A number of contributions have been made by relaxing the independence axiom and developing some *nonlinear utility models* to accommodate actual decision behavior [10–12].

MULTIATTRIBUTE UTILITY FUNCTIONS

Suppose that the outcomes defined for the preference functions are now considered to be vectors. That is, suppose that $X = X_1 \times X_2 \times \dots \times X_n$ where X_i represents the performance of an alternative on attribute i . In this section, we use the letters x , y , and z to indicate distinct elements of X . For example, $w \in X$ is represented by $(w_1, w_2, \dots, w_n)$, where w_i is a level in the nonempty attribute set X_i for $i = 1, \dots, n$. We let $I \subset \{1, 2, \dots, n\}$ be a subset of the attribute indices, define X_I as the subset of the attributes designated by the subscripts in I , and let $\bar{X}_I$ represent the complementary subset of the n attributes. We may write $w = (w_I, \bar{w}_I)$ or use the notations $(z_i, \bar{w}_i)$ and $(x_i, \bar{w}_i)$ to denote two elements of X that differ only in the level of the i th attribute. Finally, we also assume that the preference relation $\succsim$ on X is consistent with the expected-utility axioms 1–3, which means that there exists a von Neumann-Morgenstern utility function $u(x_1, x_2, \dots, x_n)$ defined on X .

If the decision maker's preferences are consistent with some additional independence conditions, then $u(x_1, x_2, \dots, x_n)$ can be decomposed into additive, multiplicative, and other well-structured forms that simplify assessment.

Preference Independence

The concept of mutual preference independence can be defined as follows:

1. X_I is *preference independent* of $\bar{X}_I$ if $(w_I, \bar{w}_I) \succ (x_I, \bar{w}_I)$ for any $w_I, x_I \in X_I$ and $\bar{w}_I \in \bar{X}_I$ implies $(w_I, \bar{x}_I) \succ (x_I, \bar{x}_I)$ for all $\bar{x}_I \in \bar{X}_I$.
2. The attributes $X_1, \dots, X_n$ are *mutually preference independent* if for every subset $I \subseteq \{1, \dots, n\}$ the set X_I of these attributes is preference independent of $\bar{X}_I$.

Attributes X_i and X_j are preference independent if the tradeoffs (substitution rates) between X_i and X_j are independent of all other attributes. Mutual preference independence requires that preference independence

holds for all pairs X_i and X_j . Essentially, mutual preference independence implies that the indifference curve for any pair of attributes is unaffected by the fixed levels of the remaining attributes. The indifference curve is the set of all combinations of X_i and X_j that have the same utility value when the other attributes are held constant and is typically plotted as a line or curve.

In practice, the condition of preference independence among a set of attributes is often a reasonable assumption. For example, suppose that a decision maker is choosing from a set of automobiles based on the attributes of cost, horsepower, and attractiveness. Further, assume that the decision maker prefers more horsepower to less, lower costs, and more attractive automobiles. If he or she prefers an automobile with the attribute values (\$25K, 170 hp, ugly) to one with the attribute values (\$28K, 200 hp, ugly), then if cost and horsepower are preference independent of appearance, he or she will still prefer the first set of attribute values to the second set if the appearance for both alternatives is changed to beautiful. This would also be true for the preference order of any two alternatives that differ in the first two attributes but have a common value for the third attribute.

Preference independence does imply the existence of an additive MAU function that exists in the case of certainty for the comparison of alternatives where there is no risk regarding any of the outcomes. However, the approaches that can be used to construct this additive function require tradeoffs by the decision maker among pairs of attributes in a manner that would be difficult to implement in practice (see [5], Sections 3.4.6 and 3.4.7). Nevertheless, mutual preference independence is an important concept because it can be combined with other independence conditions to allow a MAU function to be decomposed into a form that is easy to assess using much simpler expressions of preference by a decision maker.

Utility Independence

An attribute X_i is said to be *utility independent* of its complementary attributes if

preferences over lotteries with different levels of X_i do not depend on the fixed levels of the remaining attributes. Attributes $X_1, X_2, \dots, X_n$ are mutually utility independent if all proper subsets of these attributes are utility independent of their complementary subsets. Further, it can be shown that if these same attributes are mutually preference independent, then they will also be mutually utility independent if any pair of the attributes is utility independent of its complementary attributes [5].

Once again, suppose that a decision maker is choosing from a set of automobiles based on the attributes of cost, horsepower, and attractiveness, but now there is some uncertainty regarding new environmental laws that may impact the cost and the horsepower of a particular automobile. The current performance levels of one of the alternatives may be (\$25.0K, 170 hp, ugly), but if the legislation is passed a new device will have to be added that will increase cost and decrease horsepower leading to a new performance vector (\$25.7K, 150 hp, ugly). An alternative automobile might have possible outcomes of (\$28.0K, 200 hp, ugly) if the legislation does not pass and (\$29.0K, 175 hp, ugly) if it does. We will assume that a decision maker estimates the legislation will pass with probability 0.5.

Therefore, the decision maker is evaluating choices between lotteries such as the one shown in Fig. 3. For example, the decision maker may prefer Auto 1 to Auto 2, because the risks associated with the cost and the horsepower for Auto 1 are more acceptable to his or her than the risks associated with the cost and horsepower of Auto 2. If the decision maker's choices for these lotteries do not depend on common values of the third

attribute, then cost and horsepower are utility independent of appearance.

If the decision maker's preferences are such that each attribute is utility independent of its complimentary set of attributes, then these preferences may be represented by a MAU function $u(x_1, x_2, \dots, x_n)$ with the multilinear form

$$\begin{aligned} u(x) = & \sum_{i=1}^n k_i u_i(x_i) + \sum_{i=1}^n \sum_{j>i}^n k_{ij} u_i(x_i) u_j(x_j) \\ & + \sum_{i=1}^n \sum_{j>i}^n \sum_{m>j>i}^n k_{ijm} u_i(x_i) u_j(x_j) u_m(x_m) + \dots \\ & + k_{123\dots n} u_1(x_1) u_2(x_2) \dots u_n(x_n) \end{aligned} \quad (1)$$

where u_i is a single-attribute function over X_i scaled from 0 to 1, $0 < k_i < 1$ and k_{ijm} is a scaling constant that represents the impact of the interactions among attributes i, j , and m on preferences, for example [5]. These scaling constants are sometimes called *weights* on the criteria in the literature, but we will use the former term to avoid the implication that they reflect a "weight of importance" of a criterion, which would be misleading as we shall discuss.

If the stronger condition of mutual utility independence is consistent with the preferences of the decision maker, then the multiplicative model is appropriate

$$1 + k u(x_1, x_2, \dots, x_n) = \prod_{i=1}^n [1 + k k_i u_i(x_i)] \quad (2)$$

where k is the lone additional scaling constant, suggesting that the strength of all interactions is the same [5]. If the scaling constant k is determined to be 0 through the appropriate assessment procedure, then (2)

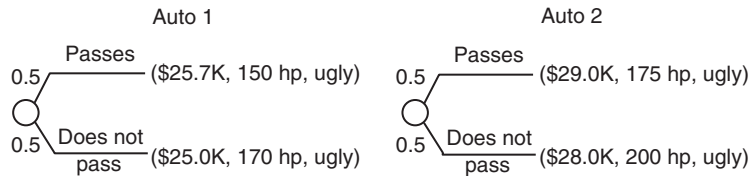


Figure 3. Choice between two lotteries.

reduces to the additive form

$$u(x_1, x_2, \dots, x_n) = \sum_{i=1}^n k_i u_i(x_i) \quad (3)$$

where $\sum_{i=1}^n k_i = 1$. To see this, expand the right hand side of (2), subtract 1 from both sides of the equation, and divide by λ .

Additive Independence

A majority of the applied work in multiattribute utility theory deals with the case when the von Neumann-Morgenstern utility function is decomposed into the additive form 3. As noted earlier, the additive form can be a special case of the multilinear utility function, and this can be determined through the assessment procedure for the scaling constants. However, it is useful to be able to determine the existence of this important form of the multiattribute utility function by directly testing the appropriate preference conditions.

Fishburn [13] has derived necessary and sufficient conditions for a utility function to be additive. The key condition for additivity is that the preferences for any lotteries $p, q \in P$ should depend only on the marginal probabilities of the attribute values and not on their joint probability distributions.

Returning to the automobile example once again, an additive utility model (3) will be appropriate if the decision maker is indifferent between the two lotteries shown in Fig. 4, and for all other permutations of the attribute values that maintain the same marginal probabilities for each.

Notice that in either lottery in Fig. 4, the marginal probability of receiving the most preferred outcome or the least preferred outcome on each attribute is identical (0.5). However, a decision maker may prefer Auto 2 over Auto 1 if he or she wishes to avoid a 0.5 chance of the poor outcome (\$29K, 150 hp, ugly) on all three attributes, or he or she may have the reverse preference if he or she is willing to accept some risk in order to have a chance at the best outcome on all three attributes. In either of the latter cases, utility independence may still be satisfied, and the

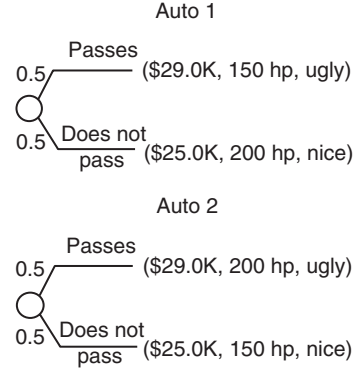


Figure 4. Additive independence criterion for risk.

multilinear 1 or the multiplicative 2 decomposition of the multiattribute utility function may be appropriate.

The assessment of the multiplicative (2) or additive form (3) implied by the condition of mutual utility independence is simplified by the fact that each of the single-attribute utility functions may be assessed independently (more accurately, while all of the other attributes are held constant at arbitrarily selected values), using the well-known utility function assessment techniques suitable for single-attribute utility functions. Other independence conditions have been identified that lead to more complex nonadditive decompositions of a multiattribute utility function. These general conditions are reviewed in Farquhar [14] and more recently in Abbas and Bell [15].

We present a hypothetical example of the assessment of an additive two-attribute utility model for a decision maker considering a car purchase based on the attributes of cost (x_c) and horsepower (x_h) using the attribute ranges of [\$25K, \$29K] for cost and (150 hp, 200 hp) for horsepower in the previous example. We have assumed a linear utility function for cost and a concave exponential utility function for horsepower, but these could be assessed using one of the processes described in Fig. 2 to determine utility function values at discrete points and then by fitting continuous functions to these points. Next, we detail how to assess the scaling constants for an additive utility function to

arrive at $u(x_c, x_h) = 0.625 u_c(x_c) + 0.375 u_h(x_h)$ as shown in Fig. 5.

Intuitively, the scaling constants of a multicriteria decision model are measures of the importance of the *increase* from the worst to the best level of performance on one objective compared to the *increase* from the worst to the best level of performance on another objective. For example, if one were planning to purchase a car from the set of cars with costs between \$10,000 and \$1,000,000, then cost would have relatively more influence on the decision than if we limit the choice set to the Honda Accord and the Mazda 626 so that the range of cost would be between \$25,000 and \$29,000. Other features of the cars would become more relevant when making the latter decision.

One of the common techniques for assessing the scaling constants is the tradeoff approach that explicitly considers the ranges over which the attributes vary. The tradeoff

approach begins by having the decision maker consider a pair of hypothetical alternatives. For example, he or she might be asked to compare the alternatives A and B shown below.

	Alternative A	Alternative B
Cost	\$25K	\$29K
Horsepower	150 hp	200 hp

Note that alternative A has the best level of cost and the worst level of horsepower, whereas the opposite is true for B. Often there are more than two criteria that determine the performance of an alternative. When this occurs, the decision maker is asked to imagine that A and B have identical outcomes on all other criteria.

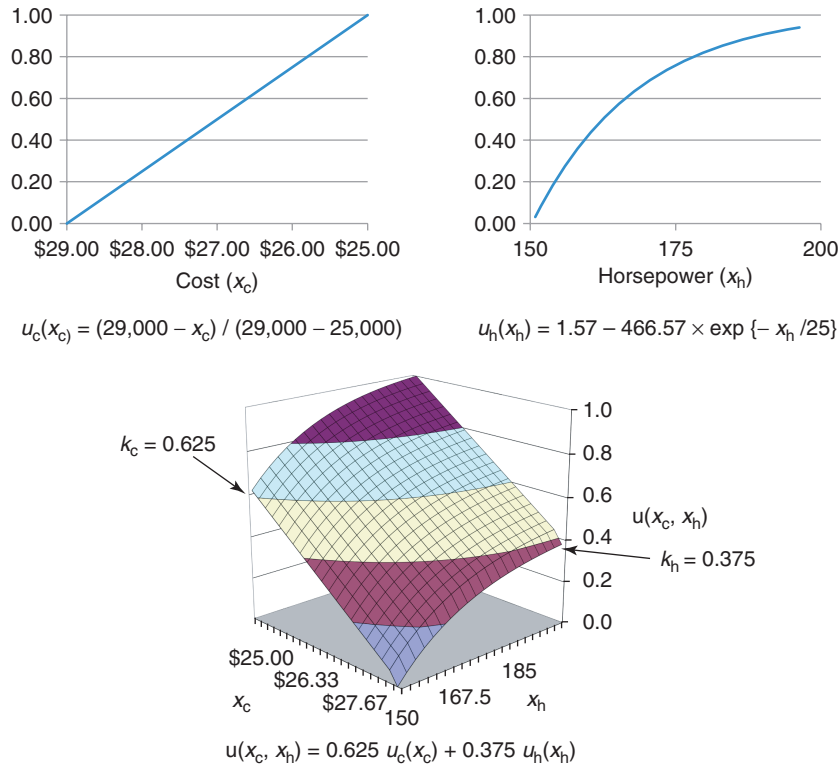


Figure 5. Additive two-attribute model.

The decision maker is asked if he or she prefers one alternative to the other. If he or she is indifferent between the two alternatives, the scaling constant assessment for this particular pair of measures is complete and the attribute scaling constants are equal. Typically, he or she will prefer one alternative to another. For example, let us assume that alternative A is preferred. This implies that it is more important to shift cost from \$29K to \$25K than to shift horsepower from 150 to 200 hp. In other words, the scaling constant on cost is greater than the scaling constant on horsepower.

In this case, the decision maker is asked to provide x'_c such that alternative A is indifferent to alternative B:

	Alternative A	Alternative B
Cost	\$25K	x'_c
Horsepower	150 hp	200 hp

When indifference has been achieved, we know that

$$k_c u_c(\$25K) + k_h u_h(150) = k_c u_c(x'_c) + k_h u_h(200)$$

or

$$k_c [u_c(\$25K) - u_c(x'_c)] = k_h [u_h(200) - u_h(150)]. \quad (4)$$

For scaling purposes, we set the utility of the most preferred level of each alternative to 1 and the utility of the least preferred alternative to 0 and (4) reduces to

$$k_c [1 - u_c(x'_c)] = k_h [1 - 0] \quad (5)$$

Thus, the ratio of the scaling constant for cost to the scaling constant for horsepower is

$$k_c/k_h = 1/[1 - u_c(x'_c)] \quad (6)$$

Given our linear utility function for cost, $u_c(x_c) = (29 - x_c)/(29 - 25)$ as $u_c(x_c)$ is decreasing in x_c . If the decision maker indicates that

he or she would pay \$2.4K to shift horsepower from 150 to 200, then he or she has specified $x'_c = 25K + 2.4K = \$27.4K$. Hence, $u_c(27.4) = 0.4$, $k_c/k_h = 5/3$ and $k_c = 0.625$.

To see the impact of the range of performance on the relative importance of an attribute, assume that we decide to consider cars that cost up to \$50K rather than our initial assumption of a maximum price of \$29K. There is no reason that this change in the range should impact x'_c — it is still worth \$2.4K to increase the horsepower from worst to best—but it will affect $u_c(x'_c)$. Now, $u_c(x_c) = (50 - x_c)/(50 - 25)$, $u_c(27.4) \cong 0.904$, $k_c/k_h \cong 10.417$, and $k_c \cong 0.912$. As we suggested at the start, when the range of the cost of cars considered increases while the ranges of the other attributes are held constant, the scaling constant for cost increases. However, the additive utility function would still give the same ranking of the alternatives because the increase in the scaling constant for cost would be just offset by a decrease in the corresponding utility function score on cost, as x'_c would be a smaller proportion of the range of the utility function.

For a multiattribute problem with n criteria, $n-1$ tradeoffs will be required in order to determine $n-1$ equations in n unknowns. Using these ratios and the restriction that the sum of the scaling constants is unity, we can solve for all the scaling constants. See Keeney and Raiffa [5] or [16] for additional details and examples of scaling constant assessment.

VALUE FUNCTIONS: PREFERENCE FUNCTIONS FOR CERTAINTY

Keeney and Raiffa [5] emphasized the use of multiattribute preference models based on the theories of von Neumann and Morgenstern [3], which rely on axioms involving risk. As a result, this approach has become synonymous in the view of many scholars with multiattribute preference theory. However, this theory is not the appropriate one for decisions involving multiple objectives when the outcomes of the alternatives are known with certainty.

Single-Attribute Value Functions

In order to model the preferences of a decision maker, we may wish to consider a “strength of preference” notion that involves comparisons of preference differences between pairs of alternatives. To do so, we need more restrictive preference assumptions, including that of a weak order over preferences between exchanges of pairs of alternatives (Chapter 4, [17]). We use the term *measurable value function* for a value function that may be used to order the differences in the strength of preference between pairs of alternatives or, more simply, the “preference differences” between the alternatives.

Once again, we begin with a discussion of a preference model for a single attribute. Let X denote the set of all possible consequences in a decision situation, with $w, x, y, z, w', x', y' \in X$; define X^* as a nonempty subset of $X \times X$, and $\succsim^*$ as a binary relation on X^* . That is, X^* is a set of paired values for the original set of outcomes or consequences X , so if X is a set of automobiles, then X^* would be a pair of automobiles. We will be interested in how much difference we perceive in our strength of preference for one of these automobiles versus our strength of preference for the other one. We shall interpret $wx \succsim^* yz$ to mean that the strength of preference for w over x is greater than or equal to the strength of preference for y over z . The notation $wx \sim^* yz$ means both $wx \succsim^* yz$ and $yz \succsim^* wx$, and $wx \succ^* yz$ means not $yz \succsim^* wx$.

We emphasize again that we are comparing the strength of preference for one alternative over another from the same set of alternatives. One way to think about this comparison is to focus on the notion of having the opportunity to exchange x for w , or to exchange y for z . If the decision maker would prefer the former exchange to the latter one, then $wx \succsim^* yz$.

We use the axiom system for positive difference structures presented by Krantz *et al.* [17]. This set of axioms includes several technical assumptions that have no significant implications for behavior. However, a key axiom that does have an intuitive interpretation in terms of preferences is the following one: If $wx, xy, w'x', x'y' \in X^*$, $wx \succsim^* w'x'$, and $xy \succsim^* x'y'$,

then $wy \succsim^* w'y'$. That is, if the difference in the strength of preference between w and x exceeds the difference between w' and x' , and the difference in the strength of preference between x and y exceeds the difference between x' and y' , then the difference in the strength of preference between w and y must exceed the difference between w' and y' .

To provide some intuition for this axiom, suppose that a shopper is trying to choose a car based on its color; the cars are identical on all other attributes. He or she might feel, for example, that he or she would prefer to exchange a blue car for a black one versus the exchange of a white car for a silver one. And, he or she might prefer to exchange a red car for a blue one versus the exchange of a yellow car for a white one. Then, it seems logical that he or she would prefer to exchange a red car for a black one versus the exchange of a yellow car for a silver one.

The axioms of Krantz *et al.* [17] imply that there exists a real-valued function v on X such that, for all $w, x, y, z \in X$, if w is preferred to x and y to z , then $wx \succsim^* yz$ if and only if

$$v(w) - v(x) \geq v(y) - v(z) \quad (7)$$

Further, v is unique up to a positive linear transformation, so it is a *cardinal function* (i.e., v provides an *interval scale of measurement*). That is, if v' also satisfies (7), then there are real numbers $a > 0$ and b such that $v'(x) = av(x) + b$ for all $x \in X$ ([17], Theorem 4.1).

We define the binary preference relation $\succsim$ on X from the binary relation $\succsim^*$ on X^* in the natural way by requiring $wx \succsim^* yx$ if and only if $w \succsim y$ for all $w, x, y \in X$. Then, from (7), it is clear that $w \succsim y$ if and only if $v(w) \geq v(y)$. Thus, v is a value function on X and, by virtue of (7), it is a *measurable value function*.

The ideas of strength of preference and of measurable value functions are important concepts that are often used implicitly in the implementation of preference theories in practice. Further, the measurable value function can be assessed using questions for subjects that do not require choices among lotteries, which may be artificial distractions in cases where subjects are trying to choose among alternatives that do not require

the consideration of risk. Examples of methods for assessing measurable value functions would include direct rating of alternatives on a cardinal scale, or direct comparisons of preference differences. For a detailed discussion of these approaches, see Keller [17], Farquhar and Keller [18], von Winterfeldt and Edwards [7], and Kirkwood [16].

It is important to note that the measurable value function and the utility function of an individual assessed over the same attribute range may be different. This would mean that an individual's preferences among risky alternatives captured by $u(x)$ could be different from his or her preferences for certain outcomes captured by $v(x)$. Therefore, it would not be logically consistent to take the expected value of a preference function $v(x)$ that is assessed using questions that do not involve risk.

As an example of a situation where utility and value functions can be distinct for the same individual, consider asking an individual to specify the price he or she would be willing to pay for a coin flip that pays \$20 for heads and \$0 for tails. In our classroom experiences with students, there are always a few who specify a price of \$10—the expected value of the lottery—but the modal response is typically around \$5 or \$6, indicating that these students tend to be very risk averse and, therefore, that their utility functions would be concave. However, suppose that you ask these same students, “How much money would you have to receive to feel twice as satisfied as if you received \$10?” Most students will say “\$20” indicating that the value function for money in this range from \$0 to \$10 is linear, which implies that the marginal value of \$1 is \$1. In this case, the value and utility functions for the typical student would have different shapes.

We acknowledge that the students are giving quick reactions that may highlight their risk aversion in the question about the coin flip, and that some of them might agree that they should be willing to pay closer to the expected value for a coin flip with such small outcomes if this is pointed out to them. Nevertheless, it is also possible that some students would feel that their utility functions representing their risky preferences are

quite different from their value functions representing the marginal value of additional dollars.

As a practical matter, however, the difference between $v(x)$ and $u(x)$ may not be significant in a decision context [7]. The relationship between these two measures of preferences is also discussed in Dyer and Sarin [19].

Measurable Multiattribute Preference Functions for the Case of Certainty

Let $X^* = \{wx | w, x \in X\}$ be a nonempty subset of $X \times X$, and $\succsim^*$ denote a weak order on X^* . Once again, we may interpret $wx \succsim^* yz$ to mean that the preference difference between w and x is greater than the preference difference between y and z . In this case, however, we allow the outcomes defined for the preference functions to be vectors. That is, suppose that $X = X_1 \times X_2 \times \dots \times X_n$ where X_i represents the performance of an alternative on attribute i .

It seems reasonable to assume a relationship between $\succsim$ on X and $\succsim^*$ on X^* as follows. Suppose that the attributes $X_1, \dots, X_n$ are mutually preference independent. These two orders are *difference consistent* if, for all $w_i, x_i \in X_i$, $(w_i, \bar{w}_i) \succsim (x_i, \bar{w}_i)$ if and only if $(w_i, \bar{w}_i)(x_i^\circ, \bar{w}_i) \succsim^* (x_i, \bar{w}_i)(x_i^\circ, \bar{w}_i)$ for some $x_i^\circ \in X_i$ and some $\bar{w}_i \in \bar{X}_i$, and for any $i \in \{1, \dots, n\}$, and if $w \sim x$, then $wy \sim^* xy$ or $yw \sim^* yx$ or both for any $y \in X$. Loosely speaking, this means that if one multiattributed alternative is preferred to another, then the preference difference between that alternative and some common reference alternative $(x_i^\circ, \bar{w}_i)$ will be larger than the preference difference between the alternative that is not preferred and this reference alternative.

Weak Difference Independence

In the case of certainty, weak difference independence plays a role similar to the utility independence condition in multiattribute utility theory. Specifically, the subset of attributes X_I is weak difference independent of $\bar{X}_I$ if, given any $w_I, x_I, y_I, z_I \in X_I$ and some $\bar{w}_I \in \bar{X}_I$ such that the subject's judgments regarding strength of preferences between pairs of multiattributed alternatives

is as follows: $(w_I, \bar{w}_I)(x_I, \bar{w}_I) \succ^*(y_I, \bar{w}_I)(z_I, \bar{w}_I)$, then the decision maker will also consider $(w_I, \bar{x}_I)(x_I, \bar{x}_I) \succ^*(y_I, \bar{x}_I)(z_I, \bar{x}_I)$ for any $\bar{x}_I \in \bar{X}_I$. That is, the ordering of preference differences depends only on the values of the attributes X_I and not on the fixed values of $\bar{X}_I$. The attributes are mutually weak difference independent if all proper subsets of these attributes are weak difference independent of their complementary subsets.

A measurable multiattribute value function $v(x_1, x_2, \dots, x_n)$ on X can be of similar form as a utility function. For example, X may have the multiplicative form

$$1 + \lambda v(x_1, x_2, \dots, x_n) = \prod_{i=1}^n [1 + \lambda \lambda_i v(x_i)] \quad (8)$$

if and only if $X_1, \dots, X_n$ are mutually weak difference independent, where v_i is a single-attribute measurable value function over X_i scaled from 0 to 1, the λ_i are positive scaling constants, and λ is an additional scaling constant. If the scaling constant λ is determined to be 0 through the appropriate assessment procedure, then 8 reduces to the additive form

$$v(x) = \sum_{i=1}^n \lambda_i v_i(x_i)$$

where $\sum_{i=1}^n \lambda_i = 1$.

Difference Independence

Once again, we see that the additive form for measurable value functions is a special case of the multiplicative form, and this can be discovered through the assessment procedures. However, we may wish to identify the stronger preference conditions that do imply an additive measurable multiattribute preference function for the case of certainty. Mutual preference independence guarantees the existence of an additive preference function for the case of certainty that will provide an ordinal ranking of alternatives, but it may not capture the underlying strength of preference of the decision maker.

The required condition for additivity that also provides a measurable preference

function is called *difference independence*. The attribute X_i is difference independent of $\bar{X}_i$ if, for all $w_i, x_i \in X_i$ such that $(w_i, \bar{w}_i) \succ (x_i, \bar{w}_i)$ for some $\bar{w}_i \in \bar{X}_i$, $(w_i, \bar{w}_i)(x_i, \bar{w}_i) \sim^*(w_i, \bar{x}_i)(x_i, \bar{x}_i)$ for any $\bar{x}_i \in \bar{X}_i$. Intuitively, the preference difference between two multiattributed alternatives differing only on one attribute does not depend on the common values of the other attributes.

The attributes are mutually difference independent if all proper subsets of these attributes are difference independent of their complementary subsets. For the case of $n \geq 3$, mutual difference independence along with some additional structural and technical conditions² ensure that if $wx, yz \in X^*$, then $wx \succ^* yz$ if and only if

$$\begin{aligned} \sum_{i=1}^n \lambda_i v_i(w_i) - \sum_{i=1}^n \lambda_i v_i(x_i) &\geq \sum_{i=1}^n \lambda_i v_i(y_i) \\ &- \sum_{i=1}^n \lambda_i v_i(z_i) \end{aligned}$$

and $x \succ y$ if and only if

$$\sum_{i=1}^n \lambda_i v_i(x_i) \geq \sum_{i=1}^n \lambda_i v_i(y_i)$$

where v_i is a single-attribute measurable value function over X_i scaled from 0 to 1, and $\sum_{i=1}^n \lambda_i = 1$. Further, if $v_i', i = 1, \dots, n$ are n other functions with the same properties, then there exist constants $\alpha > 0, \beta_1, \dots, \beta_n$ such that $v_i' = \alpha v_i + \beta_i, i = 1, \dots, n$.

Verification of the independence conditions necessary to apply multiattribute value theory is discussed in detail by Keeney and Raiffa [5], von Winterfeldt and Edwards [7], and Kirkwood [6]. Assessing individual measureable value functions is discussed

²Specifically, we assume restricted solvability from below, an Archimedean property, at least three attributes are essential, and that the attributes are bounded from below. If $n = 2$, we assume that the two attributes are preferentially independent of one another and that the Thomsen condition is satisfied (see Krantz *et al.* [17]).

in Dyer and Sarin [20] and Salo and Hamalainen [21].

The necessary conditions for the additive and multiplicative measurable value functions and risky utility functions, notably mutual preference independence, are very similar and so it is natural to investigate the relationships among them. To illustrate an important result regarding the decision maker's preferences, consider the following statements:

1. $X_1, \dots, X_n$ are difference consistent and X_1 is difference independent of $\bar{X}_1$.
2. There exists a utility function u on X and preferences over lotteries on $X_1, \dots, X_n$ depend only on their marginal probability distributions and not on their joint probability distributions.

If statements 1 and 2 are both true, then $v(x_1, x_2, \dots, x_n) = u(x_1, x_2, \dots, x_n)$ and both are additive. Loosely speaking, the additive independence concepts for utility functions and measurable value functions imply that the values of the other attributes have no impact on either the shape of the preference function on each attribute or on the relative importance of each attribute. Therefore, the additive preference function can be assessed using either questions involving risk, or questions involving strength of preference and the results should be identical. Otherwise, it may be possible to transform a multiattribute value function into a multiattribute utility function and vice versa [20].

Finally, there are many examples of multiattribute preference models that are presented as alternatives to multiattribute utility or value theory. One of the most commonly utilized multiattribute preference approaches is the analytic hierarchy process (AHP). The AHP estimates ratios of importance and preference and integrates them into an additive model of preference under certainty. However, as typically implemented, the AHP does not require decision makers to consider the range of attribute performance when making these ratio judgments which may lead to arbitrary rankings of alternatives or to rank reversals when new alternatives are introduced [22].

However, the subjects can be asked to make ratio comparisons of preference *differences* rather than to make ratio judgments where there is no natural zero-point on a preference scale, and the results may provide a valid alternative for assessing a measurable multiattribute value function (e.g., see [21–24]).

DISCUSSION

In the theoretical development presented here, it is assumed that the vector of attributes $X = X_1, X_2, \dots, X_n$ is given. Whether the decision maker is operating in a setting of certainty or uncertainty, the identification of the correct set of attributes is difficult and a key precursor to a successful multiattribute decision analysis. Prescriptive decision analysis suggests identifying the fundamental objectives—what the decision maker really cares about—and then constructing a value hierarchy by decomposing these objectives until quantifiable attributes can be identified. Keeney's (1992) concept of value-focused thinking is one of the earliest attempts to formalize the process of identifying fundamental objectives that form the basis for the value hierarchy. [25]

In this article, we have presented an informal discussion of “multiattribute utility theory.” In fact, this discussion has emphasized that there is no *single* version of multiattribute utility theory that is relevant to *all* multicriteria decision analyses. Instead, the appropriate method should be driven by the decision context. Specifically, we focused on two distinct theories of multiattribute preference functions that may be used to represent a decision maker's preferences under conditions of risk or certainty, respectively.

Multiattribute utility theory is an elegant and useful model of preference suitable for applications involving risky choice. The seminal work of Keeney and Raiffa [5] has made this theory synonymous with multicriteria decision-making, and the measurable theories are often overlooked or ignored as a result. In fact, the latter approaches may provide more attractive and appropriate theories for many applications of multicriteria decision analysis.

REFERENCES

1. Fishburn PC. Utility theory for decision making. New York: Wiley; 1970.
2. Kreps DM. A course in microeconomic theory. Princeton University Press; 1990.
3. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1947.
4. Pratt JW. Risk aversion in the small and in the large. *Econometrica* 1964;32:122–136.
5. Keeney RL, Raiffa H. Decision with multiple objectives: preference and value trade-offs. New York: Cambridge University Press; 1993.
6. Savage LJ. The foundations of statistics. New York: Wiley; 1954.
7. von Winterfeldt D, Edwards W. Decision analysis and behavioral research. Cambridge University Press; 1986.
8. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica* 1979;47:276–287.
9. Fishburn PC, Wakker P. The invention of the independence condition for preferences. *Manag Sci* 1995;41(7):1130–1144.
10. Wakker PP. Prospect Theory: For Risk and Ambiguity. Cambridge, UK: Cambridge University Press; 2010.
11. Fishburn PC. Nonlinear preference and utility theory. Johns Hopkins University Press; 1988.
12. Camerer C. Individual decision making. In: Kagel JH, Roth AE, editors. Handbook of experimental economics. Princeton, NJ: Princeton University Press; 1995.
13. Fishburn PC. Independence in utility with whole product sets. *Oper Res* 1965;13:28–45.
14. Farquhar PH. A survey of multiattribute utility theory and applications. In: Starr MK, Zeleny M, editors. Multiple criteria decision making, Vol. 6 of TIMS studies in the management sciences. North-Holland, Amsterdam; 1977. pp. 59–90.
15. Abbas AE, Bell DE. One-switch independence for multiattribute utility functions. *Oper Res* 2011;59(3):764–771.
16. Kirkwood CW. Strategic decision making. Belmont, Calif: Duxbury Press; 1996.
17. Krantz DH, Luce RD, Suppes P, *et al.* Foundations of measurement, volume I, additive and polynomial representations. Academic Press; 1971.
18. Keller LR. An empirical test of relative risk aversion. *IEEE Trans Man Syst Cybern* 1985;SMC-15:476–482.
19. Farquhar PH, Keller LR. Preference intensity measurement. *Ann Oper Res* 1989;19:205–217.
20. Dyer JS, Sarin RK. Relative risk aversion. *Manag Sci* 1982;28:875–886.
21. Dyer JS, Sarin RA. Measurable multiattribute value functions. *Oper Res* 1979;22:810–822.
22. Salo AA, Hamalainen RP. On the measurement of preferences in the analytic hierarchy process. *J Multi-Crit Decis Anal* 1997;6:309–319.
23. Dyer JS. Remarks on the analytic hierarchy process. *Manag Sci* 1990;36:249–258.
24. Kamenetzky RD. The relationship between the analytic hierarchy process and the additive value function. *Decis Sci* 1982;13:702–713.
25. Keeney RL. Value-Focused Thinking: A Path to Creative Decisionmaking. Cambridge, MA: Harvard University Press; 1992.

INTRODUCTION TO POINT PROCESSES

FREDERIC PAIK SCHOENBERG
UCLA Department of Statistics,
Los Angeles, California

A point process is a random collection of points falling in some space. In most applications, each point represents the time and/or location of an event, such as a lightning strike or earthquake. When modeling purely temporal data, the space in which the points fall is simply a portion of the real line.

Figure 1 shows an example of origin times of Southern California earthquakes of magnitude at least 3.0 occurring in 2008, obtained from the Southern California Earthquake Data Center. Point process data arise in a wide variety of scientific applications, including epidemiology, ecology, forestry, mining, hydrology, astronomy, ecology, and meteorology.

FORMAL DEFINITIONS

There are several ways of characterizing a point process. The usual approach is to define a point process N as a random measure on a complete separable metric space S taking values in the nonnegative integers $\mathbf{Z}^+$ (or infinity). In this framework the measure $N(A)$ represents the number of points falling in the subset A of S . Attention is typically restricted to the case where N may only contain finitely many points on any bounded subset of S .

The random measure definition above is perhaps not the most intuitive means of characterizing a point process, and several alternatives exist. Consider the case of a temporal point process, for example, the times of events occurring between time 0 and time T . One may characterize N as an ordered list $\{\tau_1, \dots, \tau_n\}$ of event times. One may alternatively convey the equivalent information about N via the interevent times $\{u_1, \dots, u_n\}$,

where $u_i = \tau_i - \tau_{i-1}$, with the convention that $\tau_0 = 0$.

A temporal point process N may alternatively be described by a counting process $N(t)$, where for any time t between 0 and T , $N(t)$ is the number of points occurring at or before time t . The resulting process $N(t)$ must be nondecreasing and right-continuous, and take only nonnegative integer values. One could thus define a temporal point process as any nondecreasing, right-continuous $\mathbf{Z}^+$ -valued process. This sort of definition has been used by Jacod [1], Brémaud [2], Andersen *et al.* [3], and others.

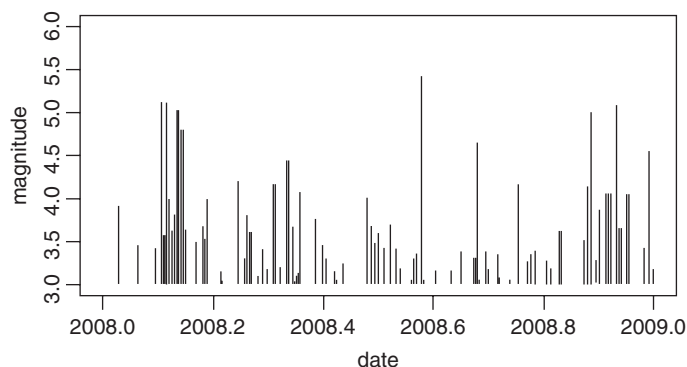
The counting process definition of a point process lends itself naturally to discussions of martingales and distributional theory. On the other hand, the random measure formulation has the advantage of generalizing immediately to point processes in higher dimensions and in abstract spaces. For simplicity, the remainder of this article will deal only with purely temporal point processes, where the domain S is simply a portion of the real (time-)line.

A point process is called *simple* if with probability one, all its points τ_i occur at distinct times. A closely related concept is orderliness: a point process N is *orderly* if for any time t , $P\{N(t, t + \Delta t) > 1\}/\Delta t \rightarrow 0$ as $\Delta t \rightarrow 0$.

The points of a point process are typically indistinguishable other than by their times and/or locations. When there is other important information to be stored along with each point, such as a list of times of wildfire origin times along with their corresponding burn areas, or the earthquake times and magnitudes shown in Fig. 1, the result may be viewed as a *marked* point process.

The relationship between time series and point processes is worth noting. Many data sets that are traditionally viewed as realizations of (marked) point processes could, in principle, also be regarded as time series, and vice versa. For instance, a sequence of earthquake origin times is typically viewed as a temporal point process; however, one

Figure 1. Times and magnitudes of Southern California earthquakes occurring in 2008, with magnitude at least 3.0, compiled by the Southern California Earthquake Data Center.



could also store such a sequence as a time series consisting of zeros and ones, with the ones corresponding to earthquakes. The main difference is that, for a point process, the points can occur at any time in a continuum, whereas the time intervals are discretized in the time series case.

POINT PROCESS MODELS

The most important point process model is the *Poisson process*, which is a simple point process N such that the number of points in any set follows a Poisson distribution and the numbers of points in disjoint sets are independent. That is, N is a Poisson process if $N(A_1), \dots, N(A_k)$ are independent Poisson random variables, for any disjoint, measurable subsets $A_1, \dots, A_k$ of S .

The behavior of a simple temporal point process N is typically modeled by specifying its *conditional intensity*, $\lambda(t)$, which represents the infinitesimal rate at which events are expected to occur around a particular time t , conditional on the prior history of the point process prior to time t . Formally, the conditional intensity associated with a temporal point process N may be defined via the limiting conditional expectation

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} E\{N[t, t + \Delta t] | H_t\} / \Delta t,$$

provided this limit exists, where H_t is the history of the point process N over all times strictly prior to time t . Some authors instead define λ as the conditional probability

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} P\{N[t, t + \Delta t] > 0 | H_t\} / \Delta t,$$

since the two definitions are equivalent for orderly point processes. As all finite-dimensional distributions of N are uniquely determined by the conditional intensity [4,5], in modeling N , it suffices to prescribe a model for λ .

For a temporal point process originating at time 0, one may define the *compensator* $A(t)$ as the integral of the conditional intensity from time 0 to time t . The compensator may equivalently be defined as the unique nonnegative, nondecreasing predictable process $A(t)$ such that $N[0, t] - A(t)$ is a martingale [1,2].

λ may be estimated nonparametrically [6–9], or via a parametric model. In general, $\lambda(t)$ depends not only on t but also on the times τ_i of preceding events. When N is a Poisson process, however, λ is deterministic; that is, $\lambda(t)$ depends only on t .

A stationary Poisson process has constant conditional rate: $\lambda(t) = \alpha$, for all t . This model posits that the risk of an event is the same at all times, regardless of how frequently such events have occurred previously. For a nonstationary Poisson process, $\lambda(t)$ is some function of t . A generalization is the Cox process, or doubly-stochastic Poisson process, which is a Poisson process whose intensity function is randomly generated.

Another important elementary type of temporal point process is the *renewal process*. A renewal process is a simple point process such that the interevent times $\{u_1, \dots, u_n\}$ are independent (typically i.i.d.) random variables. In the i.i.d. case, the density function governing each interevent time is called the *renewal density*. Note that the renewal density f is ordinarily taken to be a density

on the half-line, that is, $f(t) = 0$ for $t < 0$. Renewal models embody the notion that the hazard of an event occurring at a particular time depends only on the time since the most recent event. In wildfire hazard analysis, for example, such a model is consistent with the theory of fuel loading followed by complete fuel depletion in the event of a fire.

A point process N may be called *self-exciting* if $\text{cov}\{N(s, t), N(t, u)\} > 0$ for $s < t < u$. N is *self-correcting* if instead this covariance is negative. Thus the occurrence of points in a self-exciting point process causes other points to be more likely to occur, whereas in a self-correcting process, the points have an inhibitory effect. By definition, a Poisson process is neither self-exciting nor self-correcting.

Self-exciting point process models are often used to model events that are temporally clustered. A commonly used example is the *Hawkes process*, where the conditional intensity is given by

$$\lambda(t) = \mu(t) + \sum_{i: \tau_i < t} \nu(t - \tau_i),$$

where $\mu(t)$ represents the deterministic background rate and the function ν governs the clustering density. An example where $\mu(t)$ is constant was offered by Hawkes and Adamopoulos [10]. Hawkes models are commonly used in seismology, where they are sometimes called *epidemic-type aftershock sequence* (ETAS) models, encompassing the notion that earthquakes can have aftershocks, and those aftershocks can have aftershocks, and so on. A form of the clustering density ν that is commonly used in modeling earthquake aftershocks is the Omori–Utsu law [11]

$$\nu(t) = \kappa(t + c)^{-p},$$

which corresponds to power-law decay in the clustering behavior over time.

Alternative versions of clustered point processes are formed by generating a sequence of parents and then placing clusters of points (offspring) around each parent. For the Neyman–Scott process, for instance, the offspring points are independently and identically distributed around the parents, and

for the Bartlett–Lewis process the offspring points are each generated via a renewal process originating at the corresponding parent.

Self-correcting models are used in ecology, forestry, and other fields to model occurrences that are well-dispersed. Such models may be useful in describing births of species, for example, or in seismology for modeling earthquake catalogs after aftershocks have been removed [12]. An important example is the Markovian point process model (so called because the conditional rate obeys the Markov property), where

$$\lambda(t) = g\{t, N[0, t)\}.$$

For example, the function g may be selected so that λ takes the form

$$\lambda(t) = \exp\{\alpha + \beta(t - \rho N[0, t))\},$$

where α , β , and ρ are constants [13–15]. The parameters α and β govern the background rate and trend in the occurrences, while the product $\beta\rho$ represents the decrease in conditional rate of future events caused by each event, which may be due to diminished fuel load in the case of wildfire modeling, for instance, or the release of strain energy in the seismological case.

TRANSFORMATIONS

Some important operations on point processes include superposition, thinning, and rescaling. Graphically, these operations correspond, respectively, to overlaying points of one process onto those of another point process, deleting certain points, and stretching out the time axis.

The Poisson process frequently arises as a limiting process resulting from transformations of point processes. For example, under quite general conditions, when k independent copies of a temporal point process are superposed and the time axis is rescaled by a factor of k , the resulting process converges to a Poisson process (for details see Refs 16 and 17). Similar results can be obtained for randomly thinned point processes [18] or rescaled point processes [19].

ESTIMATION AND SIMULATION

The parameter vector θ for a point process model with conditional intensity $\lambda(\theta)$ is typically estimated by maximizing the log-likelihood function

$$L(\theta) = \sum_i \log \lambda(\tau_i; \theta) - \int_S \lambda(t) dt.$$

The consistency, asymptotic normality, and efficiency of maximum likelihood estimators have been established under standard conditions [20,21], and standard errors may be obtained via the Hessian of the log-likelihood [22–24].

Alternatively, simulations may be useful for obtaining approximate standard errors and for other types of inference. An effective simulation method for point processes based on random thinning was devised in Refs 25 and 26. The procedure, which works for point processes whose conditional rate λ is bounded (or locally bounded) above by some constant b , involves simulating a (locally) stationary Poisson process with intensity b and thinning it, keeping each point τ_i with probability $\lambda(\tau_i)/b$.

MODEL EVALUATION

The fit of a parametric point process model is often assessed using a likelihood score such as the Akaike Information Criterion (AIC), which is defined as $2p - 2L(\theta)$, where p is the number of fitted parameters. The AIC rewards a model for higher likelihood and penalizes a model for overfitting; lower AIC indicates better fit.

Residual analysis methods for point processes often involve comparing the number of points observed with the integrated conditional intensity within specified portions of S . Methods for normalizing and inspecting such residuals are described in Ref. 27.

Another useful technique for evaluating point process models is via rescaling. The method essentially involves rescaling the time axis of the observed point process N at time t by a factor of $\lambda(t)$. More specifically, if the points are observed from time 0 to time t , then each point τ_i is moved to the new time

$A(\tau_i)$, where A is the compensator. The resulting process M is a stationary Poisson process with unit rate, provided the original point process is simple [19]. Similarly, one may inspect residuals obtained by randomly thinning the process, that is, keeping each point τ_i independently with probability inversely proportional to $\lambda(\tau_i)$. As with rescaled residuals, the resulting thinned residuals will be distributed according to a stationary Poisson process [28]. In practice, one may use the estimated intensity or compensator in place of the true intensity or compensator and inspect the rescaled or thinned process for uniformity. Several tests exist for this purpose, with different uses depending on the alternative hypotheses [3,29–33]. Some of the most useful are tests based on second- and higher-order properties [34–37]. There are also more general second-order tests for point processes that do not rely on a stationary Poisson null hypothesis [38–40].

For further reading on point processes, see Refs 5 and 17.

REFERENCES

1. Jacod J. Multivariate point processes: predictable projections, Radon-Nikodym derivatives, representation of martingales. *Z Wahrsch Verw Geb* 1975;31:235–253.
2. Brémaud P. Point processes and queues: Martingale dynamics. New York: Springer; 1981.
3. Andersen P, Børgan O, Gill R, *et al.* Statistical models based on counting processes. New York: Springer; 1993.
4. Fishman PM, Snyder D. The statistical analysis of space-time point processes. *IEEE Trans Inf Theory* 1976;22:257–274.
5. Daley DJ, Vere-Jones D. An introduction to the theory of point processes, volume i: elementary theory and methods. 2nd ed. New York: Springer; 2003.
6. Diggle P. A kernel method for smoothing point process data. *Appl Stat* 1985;34:138–147.
7. Guttorp P, Thompson M. Nonparametric estimation of intensities for sampled counting processes. *J Roy Stat Soc B* 1990;52:157–173.
8. Vere-Jones D. Statistical methods for the description and display of earthquake catalogs. In: Statistics in the environmental and earth sciences. Walden A, Guttorp P, editors. London: Edward Arnold; 1992; pp. 220–246.

9. Brillinger D. Some wavelet analyses of point process data. 31st Asilomar Conference on signals, systems and computers. Pacific Grove, California: IEEE Computer Society; 1998. pp. 1087–1091.
10. Hawkes A, Adamopoulos L. Cluster models for earthquakes—regional comparisons. *Bull Int Stat Inst* 1973;45(3):454–461.
11. Ogata Y. Statistical models for earthquake occurrences and residual analysis for point processes. *J Amer Stat Assoc* 1988;83:9–27.
12. Vere-Jones D. Earthquake prediction—a statistician's view. *J Phys Earth* 1978;26: 129–146.
13. Isham V, Westcott M. A self-correcting point process. *Stoch Proc Appl* 1979;8:335–347.
14. Ogata Y, Vere-Jones D. Inference for earthquake models: a self-correcting model. *Stoch Proc Appl* 1984;17:337–347.
15. Vere-Jones D, Ogata Y. On the moments of a self-correcting process. *J Appl Probab* 1984;21:335–342.
16. Kallenberg O. Random measures. 3rd ed. Berlin: Akademie-Verlag; 1983.
17. Daley DJ, Vere-Jones D. An introduction to the theory of point processes Volume II: General theory and structure. 2nd ed. New York: Springer; 2008.
18. Westcott M. Simple proof of a result on thinned point processes. *Ann Probab* 1976;4: 89–90.
19. Meyer P. Démonstration simplifiée d'un théorème de Knight. *Lect Notes Math* 1971;191:191–195.
20. Brillinger D. Statistical inference for stationary point processes. *Stoch Process Relat Top* 1975;1:55–100.
21. Ogata Y. The asymptotic behavior of maximum likelihood estimators for stationary point processes. *Ann Inst Stat Math* 1978;30:243–262.
22. Kutoyants Y. Parameter estimation for stochastic processes. Berlin: Helderman; 1984.
23. Dzharaparidze K. On asymptotic inference about intensity parameters of a counting process. *Proc 45th Sess Int Stat Inst* 1985;4(23.2):1–15.
24. Rathbun S, Cressie N. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Adv Appl Probab* 1994;26:122–154.
25. Lewis P, Shedler G. Simulation of nonhomogeneous Poisson processes by thinning. *Nav Res Logist Q* 1979;26:403–413.
26. Ogata Y. On Lewis' simulation method for point processes. *IEEE Trans Inf Theory* 1981;IT-27:23–31.
27. Baddeley A, Turner R, Møller J, *et al.* Residual analysis for spatial point processes. *J Roy Stat Soc B* 2005;67(5):617–666.
28. Schoenberg FP. Multi-dimensional residual analysis of point process models for earthquake occurrences. *J Amer Stat Assoc* 2003;98(464):789–795.
29. Diggle P. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* 1979;35:87–101.
30. Dijkstra J, Rietjens T, Steutel F. A simple test for uniformity. *Stat Neerl* 1984;38:33–44.
31. Lisek B, Lisek M. A new method for testing whether a point process is Poisson. *Statistics* 1985;16:445–450.
32. Lawson A. On tests for spatial trend in a non-homogeneous Poisson process. *J Appl Stat* 1988;15:225–234.
33. Arsham H. A modified one-sided K-S confidence region narrower in the tail. *Commun Stat A* 1987;16:17–28.
34. Bartlett M. The spectral analysis of two-dimensional point processes. *Biometrika* 1964;51:299–311.
35. Davies R. Testing the hypothesis that a point process is Poisson. *Adv Appl Probab* 1977;9:724–746.
36. Ripley B. Tests of 'randomness' for spatial point patterns. *J Roy Stat Soc B* 1979; 41:368–374.
37. Heinrich L. Goodness-of-fit tests for the second moment function of a stationary multidimensional Poisson process. *Statistics* 1991;22: 245–278.
38. Baddeley A, Møller J, Waagepetersen R. Non and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat Neerl* 2000;54(3):329–350.
39. Veen A, Schoenberg FP. Assessing spatial point process models for California earthquakes using weighted K-functions: analysis of California earthquakes. In: Case studies in spatial point process models. Baddeley A, Gregori P, Mateu J, *et al.*, editors, New York: Springer; 2005. pp 293–306.
40. Adelfio G, Schoenberg FP. Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *Ann Inst Stat Math* 2009;61(2):75–96.

INTRODUCTION TO POLYNOMIAL TIME ALGORITHMS FOR LP

TAMÁS TERLAKY

Department of Industrial and Systems
Engineering, Lehigh University,
Bethlehem, Pennsylvania

INTRODUCTION

Since the path-breaking paper of Khachiyan [1], a tremendous amount of research was devoted to the study of polynomial time algorithms for linear programming (LP)¹ Khachiyan [1] has developed the foundations of the theory underlying all polynomial time algorithms for LP. He presented the first *condition number*, the input bit-length of an LP problem given by integer (or rational) data. In the second section, we present some frequently used condition numbers, and then in the third section, a simple variant of the ellipsoid method is presented.

Over the last quarter century, interior point methods (IPMs) dominate the literature of polynomial time algorithms for LP. In 1984, Karmarkar [3] presented the first polynomial time IPM. By now, path-following IPMs are among the most efficient methods for solving linear, and also wide classes of general convex optimization problems. Hundreds of polynomial time IPM variants were developed for LP. Interior point software implementations have challenged simplex method implementations, and frequently surpassed their performance [4–6]. All aspects of linear optimization (LO) had to be revisited. Not only the research on algorithms exploded, but so did the research on duality theory [2,7]. The book of Roos *et al.* [2] presents a thorough treatment of the IPM-based theory—duality, initialization, complexity, sensitivity analysis,

implementation strategies—and large classes of IPMs for LP.

This article is structured as follows. The section LP Duality and Condition Numbers summarizes key duality results of LP, and presents condition numbers that indicate the difficulty of the problem. The ellipsoid method is presented in the section The Ellipsoid Method, while in the section IPMs for LP, we give a brief introduction to IPMs for LP. In the fifth section, we review some extensions of polynomial time algorithms (mainly IPMs) to convex conic and nonlinear optimization.

Notation

In this article, $\mathbb{R}_+^n$ denotes the set of nonnegative vectors in $\mathbb{R}^n$. Throughout, we use $\|\cdot\|_p$ ($p \in \{1, 2, \infty\}$) to denote the p -norm on $\mathbb{R}^n$, with $\|\cdot\|$ denoting the Euclidean norm $\|\cdot\|_2$. The identity matrix is denoted by I , e denotes the vector that has all of its components equal to one. Given an n -dimensional vector x , the matrix X denotes the $n \times n$ diagonal matrix whose diagonal entries are the coordinates x_j of x . If $x, s \in \mathbb{R}^n$ then $x^T s$ denotes the dot product of the two vectors, while xs denotes the coordinatewise product of the vectors x and s .

LP DUALITY AND CONDITION NUMBERS

The Linear Optimization Problem, Duality

We consider the general LP problem (P) and its dual (D) in canonical form:

$$\min \{c^T x : Ax \geq b, x \geq 0\}, \quad (P)$$

$$\max \{b^T y : A^T y \leq c, y \geq 0\}, \quad (D)$$

where A is an $m \times \ell$ matrix, $b, y \in \mathbb{R}^m$ and $c, x \in \mathbb{R}^\ell$. It is well known that by using only elementary transformations, any given LP problem can easily be transformed into a “minimal” canonical form. These transformations [2] can be summarized as follows:

¹Linear programming is frequently referred to as *linear optimization (LO)* [2].

- introduce slack variables in order to get equations (if a variable has both lower and upper bounds, then one of these bounds is transformed to an inequality constraint);
- shift the variables with lower or upper bound so that the respective bound becomes 0 and, if needed replace the variable by its negative;
- eliminate free variables;
- use Gaussian elimination to transform the problem into a form, where all equations have a singleton column (i.e., choose a basis and multiply the equations by the basis inverse), while dependent constraints are eliminated.

The weak duality theorem sets the stage for establishing the fundamental duality relationship between the primal problem (P) and its dual (D).

Theorem 1 [Weak duality for linear optimization]. *Let us assume that $x \in \mathbb{R}^\ell$ and $y \in \mathbb{R}^m$ are feasible solutions for the primal problem (P) and dual problem (D), respectively. Then one has*

$$c^T x \geq b^T y,$$

where equality holds if and only if²

- (i) $x_i(c - A^T y)_i = 0$ for all $i = 1, \dots, \ell$ and
- (ii) $y_j(Ax - b)_j = 0$ for all $j = 1, \dots, m$.

The Weak Duality Theorem implies the following sufficient optimality condition.

Corollary 1. *Let a primal and dual feasible solution $x \in \mathbb{R}^\ell$ and $y \in \mathbb{R}^m$ with $c^T x = b^T y$ be given. Then x is an optimal solution of*

²These conditions are in general referred to as the *complementarity conditions*. Using the coordinatewise notation we may write $x(c - A^T y) = 0$ and $y(Ax - b) = 0$, or by introducing the slack variables $z = Ax - b$ and $s = c - A^T y$ we have $xs = 0$ and $yz = 0$. By the weak duality theorem complementarity and feasibility imply optimality.

the primal problem (P) and y is an optimal solution of the dual problem (D).

Although the Weak Duality Theorem provides a sufficient condition to check optimality of a feasible solution pair, it does not prove that for every feasible primal-dual LP problem pair, an optimal pair of solutions with zero duality gap exists. This is the so-called Strong Duality Theorem that can be proved for example, by using only simple calculus and fundamental concepts of IPMs [2].

As, optimistically, we are looking for optimal solutions with zero duality gap, we utilize the results of the Weak Duality Theorem (Theorem 1); thus, we need to find a solution of the inequality system formed by the primal and the dual feasibility constraints and by requiring that the dual objective is at least as large as the primal one (reversed weak duality inequality). For this system, we already know that any solution of this system gives both a primal and a dual feasible solution with equal objective values; thus, by Corollary 1, these solutions are optimal.

$$\begin{aligned} -Ax &\leq -b, & x &\geq 0, \\ A^T y &\leq c, & y &\geq 0, \\ -b^T y + c^T x &\leq 0. \end{aligned}$$

The Strong Duality Theorem of LP [2,8] says that solving this inequality system is equivalent to solving the primal and dual LP problems (P) and (D). By introducing the notation

$$\tilde{A}^T = \begin{pmatrix} 0 & -A \\ A^T & 0 \\ -b^T + c^T \end{pmatrix}, \quad \tilde{c} = \begin{pmatrix} -b \\ c \\ 0 \end{pmatrix},$$

$$\tilde{y} = \begin{pmatrix} y \\ x \end{pmatrix} \geq 0,$$

solving the LP problem is equivalent to solving the feasibility problem

$$\tilde{A}^T \tilde{y} \leq \tilde{c}, \quad \tilde{y} \geq 0. \quad (FP)$$

Optimal Partition

Roos *et al.* [2] and Goldman and Tucker [9] proved that whenever optimal solutions exist, the primal-dual LP problem pair (P) and (D)

always has strictly complementary optimal solutions (x, z) and (s, y) , where slack variables are defined as $z = Ax - b$, and $s = c - A^T y$. Strictly complementary optimal solutions are feasible for the primal and dual problems, respectively and satisfy

$$xs = 0, \quad x + s > 0, \quad zy = 0, \quad y + z > 0.$$

The existence of strictly complementary optimal solutions allow us to introduce the *optimal partition* of the index sets:

$$\begin{aligned} \mathbf{B}_x &= \{j : x_j > 0\}, & \mathbf{B}_z &= \{i : z_i > 0\} \\ \mathbf{N}_x &= \{j : s_j > 0\}, & \mathbf{N}_z &= \{i : y_i > 0\}. \end{aligned}$$

One easily see that

$$\begin{aligned} \mathbf{B}_x \cap \mathbf{N}_x &= \emptyset, \quad \mathbf{B}_x \cup \mathbf{N}_x = \{1, \dots, \ell\}, \quad \text{and} \\ \mathbf{B}_z \cap \mathbf{N}_z &= \emptyset, \quad \mathbf{B}_z \cup \mathbf{N}_z = \{1, \dots, m\}. \end{aligned}$$

A Condition Number

We first introduce the following two numbers:

$$\begin{aligned} \sigma^P &:= \min \left\{ \min_{j \in \mathbf{B}_x} \max_{x \in \mathcal{P}^*} \{x_j\}, \min_{i \in \mathbf{B}_z} \max_{z \in \mathcal{P}^*} \{z_i\} \right\}, \\ \sigma^D &:= \min \left\{ \min_{j \in \mathbf{N}_x} \max_{s \in \mathcal{D}^*} \{s_j\}, \min_{i \in \mathbf{N}_z} \max_{y \in \mathcal{D}^*} \{y_i(x)\} \right\}, \end{aligned}$$

where $\mathcal{P}^*$ and $\mathcal{D}^*$ denote the set of primal and dual optimal solutions, respectively. By convention we take $\sigma^P = \infty$ if both $\mathbf{B}_x$ and $\mathbf{B}_z$ are empty, and analogously $\sigma^D = \infty$ if both $\mathbf{N}_x$ and $\mathbf{N}_z$ are empty. Owing to the definition of strictly complementary solutions, both σ^P and σ^D are positive. As a consequence, the number

$$\sigma := \min \left\{ \sigma^P, \sigma^D \right\}$$

is positive as well. This number plays a crucial role in the further analysis and is called the *condition number* of LP problem (P) .³

³This condition number seems to be a natural quantity for measuring the degree of hardness of (P) and (D) . The smaller the number the more difficult it is to find a strictly complementary solution,

For the calculation of σ we need to know the maximal value of each coordinate of the vector (x, z, s, y) when this vector runs through all possible optimal solutions of (P) and (D) , and σ is the smallest of all of those maximal values. The following lemma enables us to give an estimate for σ_{SP} .

Lemma 1. *Let A be an integer $m \times n$ matrix with columns A_j and $b \in \mathbb{Z}^m$ so that the set*

$$\mathcal{F} := \{x : Ax = b, x \geq 0\}$$

is bounded and contains a positive vector. Then for each j , with $1 \leq j \leq n$,

$$\max_x \{x_j : x \in \mathcal{F}\} \geq \frac{1}{\prod_{j=1}^n \|A_j\|}.$$

Proof. Observe that each column A_j of A must be nonzero, due to the boundedness of $\mathcal{F}$. Fixing the index j , let $x \in \mathcal{F}$ be such that x_j is maximal. Note that such an x exists since $\mathcal{F}$ is bounded. Moreover, since $\mathcal{F}$ contains a positive vector, we must have $x_j > 0$. Let J be the support of x :

$$J = \{j : x_j > 0\}.$$

We may assume that the columns of the submatrix A_J are linearly independent. Let A_{KJ} be any nonsingular submatrix of A_J . Here K denotes a suitable subset of the row indices. Then we have

$$A_{KJ} x_J = b_K,$$

since the coordinates x_j of x with $j \notin J$ are zero. By using Cramer's rule⁴ we have

$$x_j = \frac{\det A'_{KJ}}{\det A_{KJ}}, \quad (1)$$

and the more difficult it is to differentiate a nonzero coordinate of a solution from a zero coordinate. In a more general context, this condition number was introduced by Ye [10]; see also Ye and Pardalos [11]. For a discussion of other condition numbers and their relation to the size of a problem we refer the reader to Vavasis and Ye [12].

⁴The idea of using Cramer's rule in this way was applied first by Khachiyan [1].

where A'_{KJ} denotes the matrix arising from A_{KJ} by replacing the j th column by b_K . We know that $x_j > 0$, consequently $|\det A'_{KJ}| > 0$. Since all entries of A'_{KJ} are integral, the absolute value of its determinant is at least 1, implying

$$x_j \geq \frac{1}{|\det A_{KJ}|}.$$

Owing to the Hadamard [2,13] inequality, we have that $|\det A_{KJ}|$ is bounded above by the product of the norms of its columns.⁵ We have

$$x_i \geq \frac{1}{\prod_{j \in J} \|A_{Kj}\|} \geq \frac{1}{\prod_{j \in J} \|A_j\|} \geq \frac{1}{\prod_{j=1}^n \|A_j\|}.$$

The second inequality is obvious and the last inequality follows since A has no zero columns and hence the norm of each column of A is at least 1. This proves the lemma.

Theorem 2. *Let assume that all problem data A, b, c are integers. Let $\tilde{A}_j$ denote the j th column of $\tilde{A}$. Then the condition number σ of (P) satisfies*

$$\sigma \geq \frac{1}{\prod_{j=1}^{m+\ell} \|\tilde{A}_j\|}.$$

One can give an upper bound for the coordinates of the basic solutions of problem (P) . Analogous to Lemma 1, using Cramer's rule and Hadamard's inequality one can derive the following bound.

Lemma 2. *Let A^B be an $m \times m$ nonsingular matrix with columns A_j^B and b a vector of dimension m such that the unique solution x^B of $A^B x^B = b$ is nonnegative. Let all the entries in A^B and b be integral. Then for each j one has*

$$x_j \leq \|b\| \prod_{j=1}^m \|A_j^B\|.$$

⁵The idea of using Hadamard's inequality for deriving bounds on the coordinates of x_j from Equation (1) was applied earlier by Kladfazy and Terlaky [13] in a similar context.

This provides a uniform upper bound for all the coordinates of the primal and dual basis solutions.

Theorem 3. *Let us assume that all problem data A, b, c are integers. Then, all the coordinates of the basis solutions of problems (P) and (D) are bounded by*

$$\bar{\sigma} := \|\tilde{c}\| \prod_{j=1}^{m+\ell} \|\tilde{A}_j\|.$$

Input Length L

Another characteristic number, analogous to the condition number σ , that also allows to give bounds on the size of the coordinates of basis solutions. Assume that all problem data A, b , and c are integers. Then the binary input length of problem (P) and (D) are given by

$$\begin{aligned} L = & \sum_{i=1}^m \sum_{j=1}^{\ell} \lceil \log_2 (|A_{ij}| + 1) \rceil \\ & + \sum_{i=1}^m \lceil \log_2 (|b_i| + 1) \rceil \\ & + \sum_{j=1}^{\ell} \lceil \log_2 (|c_j| + 1) \rceil \\ & + m\ell + m + \ell + 1. \end{aligned}$$

It is known [2,14] that $\sigma \geq 2^{-L}$, and that 2^L is an upper bound for all the coordinates of the basic solutions; thus, 2^{-L} can be considered as a condition number for LP problems (P) and (D) .

Polynomial Complexity

An algorithm for solving LP problems (P) and (D) is called a *polynomial time* algorithm, if the following hold:

- The number of iterations is bounded by a polynomial of the problem dimension $n \geq \max\{m, \ell\}$ and the logarithm of the condition number. *Note:* The reader may find in the literature polynomial iteration bounds expressed for example, as $\mathcal{O}(n^\gamma \log(\frac{1}{\sigma}))$ or $\mathcal{O}(n^\gamma L)$, where

usual values of γ are 0.5, 1, or 2. For algorithms that produce ϵ precise solutions, the iteration bound needs to depend only logarithmically on the precision, that is, the iteration bound may be $\mathcal{O}(n^\gamma \log \frac{n}{\epsilon})$.

- The number of arithmetic operations for each iterations is bounded by a polynomial of the dimension n . *Note:* Usually the number of arithmetic operations per iterations is in the range of $\mathcal{O}(n^2) - \mathcal{O}(n^3)$.
- If the algorithm does not produce an exact solution, then from a solution sufficiently close to optimality, a polynomial time rounding procedure allows to identify an exact optimal solution.

In the sequel we shortly discuss a variant of Khachiyan's polynomial time *ellipsoid method*, and a simple variant of *central path-following IPMs*.

THE ELLIPSOID METHOD

The presented variant is not the sharpest version of the ellipsoid method. Our aim is to present the main ideas, and present a simple, complete version. Proofs and details of this variant can be found in Klafszky and Terlaky [13]; see also Schrijver [14].

The next lemma shows, that if system (FP) has a solution, that is, LP problems (P) and (D) have optimal solutions, then by Theorem 3 (FP) has a solution in the “ $\bar{\sigma}$ cube,” that is, the infinity norm of the solution vector is less than or equal to $\bar{\sigma}$.

Lemma 3.

1. If $\tilde{A}^T \tilde{y} \leq \tilde{c}$ is consistent, then $\tilde{A}^T \tilde{y} \leq \tilde{c}$; $|\tilde{y}_i| \leq \bar{\sigma}$ for all index i is consistent. (Any basic solution has this property.)
2. If $\tilde{A}\tilde{x} = 0$, $\tilde{c}^T \tilde{x} = -1$, $\tilde{x} \geq 0$ is consistent, then $\tilde{A}\tilde{x} = 0$, $\tilde{c}^T \tilde{x} = -1$, $\tilde{x} \geq 0$, $\tilde{x}_j \leq \bar{\sigma}$ for all j is consistent. (Any basic solution has this property.)

This result and the Farkas Lemma [2,8] allows us to prove the following theorem.

Theorem 4. *Inequality system $\tilde{A}^T \tilde{y} \leq \tilde{c}$ is solvable if and only if the inequality system*

$$\begin{aligned} \tilde{A}^T \tilde{y} &\leq \tilde{c} + \frac{1}{m\ell\bar{\sigma}}e \\ |\tilde{y}_i| &\leq \bar{\sigma} + \frac{1}{m\ell\bar{\sigma}}, \quad \text{for all } i \end{aligned}$$

is solvable.

Corollary 2. *If the inequality system*

$$\begin{aligned} \tilde{A}^T \tilde{y} &\leq \tilde{c} + \frac{1}{m\ell\bar{\sigma}}e \\ |\tilde{y}_i| &\leq \bar{\sigma} + \frac{1}{m^2\ell\bar{\sigma}^2}, \quad \text{for all } i \end{aligned}$$

is solvable, then the set of solutions contains a cube with edge length $\frac{1}{m^2\ell\bar{\sigma}^2}$.

These results provide a polyhedron that is empty if and only if the original polyhedron is empty, and if not empty the new polyhedron contains a “small” cube, which guarantees that its volume is strictly positive.

The ellipsoid algorithm of the section titled “The Ellipsoid Algorithm and its Complexity”, and the procedure of the section titled “Find an Exact Solution” applied to the optimality conditions, represented by the matrix $\tilde{A}$ and vector $\tilde{c}$, yield a polynomial time algorithm that identifies an optimal solution of the LP problem, if it exists. For simplicity of notation, instead of $\tilde{A}^T \tilde{y} \leq \tilde{c}$ the notation $A^T y \leq c$ will be used in the next sections.

Ellipsoids

Ellipsoids and an ellipsoid transformation are examined in this section. It will be shown that the transformed ellipsoid contains a suitable half ellipsoid of the original ellipsoid and its volume is less than the volume of the original ellipsoid.

Definition 1. Let $P : m \times m$ be a positive definite matrix, and $z \in R^m$. Then the ellipsoid centered at z is given by

$$E = \text{ell}(z, P) = \left\{ x \in R^m \mid (x - z)^T P^{-1} (x - z) \right\}.$$

It is well known, that the set $\{x|a^T(x-z) \leq 0\}$ is the half-space, whose boundary contains point z and the normal vector of this half-space is vector a . A half-space through the center of an ellipsoid intersected by the ellipsoid provides a half ellipsoid.

Let an ellipsoid $E = \text{ell}(z, P)$ and a vector $a \in R^m$ be given. Let $E' = \text{ell}(z', P')$ be given as follows:

$$\begin{aligned} z' &= z - \frac{1}{m+1} \frac{Pa}{\sqrt{a^T P a}} \\ P' &= \frac{m^2}{m^2-1} \left(P - \frac{2}{m+1} \frac{Paa^T P}{a^T P a} \right). \end{aligned} \quad (2)$$

Now we will verify that E' is really an ellipsoid and it contains the half ellipsoid $E \cap$

$\{x|a^T(x-z) \leq 0\}$, and its volume is smaller than that of E .

Lemma 4. *For matrix P' and ellipsoid E' the following hold:*

- Matrix P' is positive definite.
- Ellipsoid E' contains the half ellipsoid $E \cap \{x|a^T(x-z) \leq 0\}$.
- $\frac{\text{vol}(E')}{\text{vol}(E)} < \exp\left(\frac{-1}{2(m+1)}\right)$.

The Ellipsoid Algorithm and its Complexity

The following ellipsoid method decides if the linear inequality system $A^T y \leq c$ is consistent or not.

Algorithm

Initialization: Let a matrix $A : m \times \ell$ and a vector $c \in R^\ell$ be given. Let us compute $\bar{\sigma}$ and let $E_0 = \text{ell}(z^0, P_0) = \text{ell}(0, 2\bar{\sigma}E)$ be the sphere with radius $2\bar{\sigma}$, be the initial ellipsoid.

Termination:

- If z^k satisfies $A^T z^k \leq c + \frac{1}{m\ell\bar{\sigma}}e$, then STOP, inequality system $A^T y \leq c$ is consistent.
- If $k = K > 2m(m+1)\log(4m^2\ell\bar{\sigma}^3)$, then STOP, the inequality system is inconsistent.

General step:

- If z^k does not satisfies $A^T z^k \leq c + \frac{1}{m\ell\bar{\sigma}}e$, then there is an index j for which

$$a_j^T z^k \leq c_j + \frac{1}{m\ell\bar{\sigma}}.$$

Let $a := a_j$, and according to formulas (2), let $z^{k+1} = z'$, $P_{k+1} = P'$, and $k = k+1$.

By Corollary 2 we know that inequality system $A^T y \leq c$ is consistent if and only if $A^T y \leq c + \frac{1}{m\ell\bar{\sigma}}e$ is consistent. So, we have to show that in case of $k=K$ the inequality system is indeed inconsistent.

Theorem 5. *If the above algorithm takes more than $2m(m+1)\log(4\bar{\sigma}^3 m^2 \ell)$ steps, then inequality system $A^T y \leq c$ is inconsistent.*

Proof. By Corollary 2 it is enough to show that inequality system $A^T y \leq c + \frac{1}{m\ell\bar{\sigma}}e$, $|y_i| \leq \bar{\sigma} + \frac{1}{m^2\ell\bar{\sigma}^2}$, for all i is inconsistent in this case. Corollary 2 says that its solution set, if it is

not empty, contains a cube with edge length $\frac{1}{m^2\ell\bar{\sigma}^2}$. The volume of this cube is $\left(\frac{1}{m^2\ell\bar{\sigma}^2}\right)^m$.

Lemma 4 implies that this cube is contained in all of the ellipsoids generated by the algorithms, and the volume of the actual ellipsoid after K steps can be estimated by Lemma 4 as follows:

$$\begin{aligned} \text{vol}(E_K) &< \exp\left(\frac{-K}{2(m+1)}\right) \text{vol}(E_0) \\ &< \exp\left(\frac{-K}{2(m+1)}\right) (4\bar{\sigma})^m. \end{aligned}$$

If this value is less than $\left(\frac{1}{m^2\ell\bar{\sigma}^2}\right)^m$, that is the volume of the ellipsoid is less than the

volume of the contained cube, then we have a contradiction, that is the inequality system is inconsistent,

$$\exp\left(\frac{-K}{2(m+1)}\right)(4\bar{\sigma})^m < \left(\frac{1}{m^2\ell\bar{\sigma}^2}\right)^m,$$

which holds if

$$K > 2m(m+1)\log(4m^2\ell\bar{\sigma}^3).$$

The proof is complete.

Since the number of steps of the algorithm can be bounded by a polynomial, and the computational cost per iteration is quadratic in the matrix dimension, the ellipsoid algorithm is a *polynomial time algorithm*. This way a polynomial time algorithm is constructed, which decides whether inequality system $A^T y \leq c$ is consistent or not. Unfortunately our algorithm gives a solution only for inequality system $A^T y \leq c + \frac{1}{m\ell\bar{\sigma}}e$, and the solution of this system is not necessarily a solution for system $A^T y \leq c$. Thus, we need

Algorithm

Step 0. Decide if the inequality system

$$a_1^T y \leq c_1, \dots, a_\ell^T y \leq c_\ell$$

is consistent or not. If it is inconsistent STOP.

Step 1. If it is consistent then decide if the inequality system

$$a_1^T y = c_1, \quad a_2^T y \leq c_2, \dots, a_\ell^T y \leq c_\ell$$

is consistent. If it is inconsistent, then it is enough to solve inequality system $a_2^T y \leq c_2, \dots, a_\ell^T y \leq c_\ell$, since in this case for all solutions of this system $a_1^T y < c_1$ holds. Let $\ell := \ell - 1$ and return to the previous step.

Step 2. If it is consistent then decide if the inequality system

$$a_1^T y = c_1, \quad a_2^T y = c_2, \quad a_3^T y \leq c_3, \dots, a_\ell^T y \leq c_\ell$$

is consistent.

$\vdots$

Step k. If it is consistent then decide if the inequality system

$$a_1^T y = c_1, \dots, a_k^T y = c_k, \quad a_{k+1}^T y \leq c_{k+1}, \dots, a_\ell^T y \leq c_\ell$$

to construct a polynomial time procedure for generating a solution of the original inequality system. This procedure subsequently calls the algorithm of this section.

Find an Exact Solution

Using Gaussian elimination and a polynomial time algorithm, which decides whether an inequality system is consistent or not, a solution of inequality system $A^T y \leq c$ can be found in polynomial time.

Theorem 6. *If we have an algorithm (e.g., ellipsoid method), which decides the consistency of a linear inequality system, then a solution can be found in polynomial time.*

Proof. The proof is constructive. By calling the algorithm at most ℓ times, and solving an equation system by Gaussian elimination at most ℓ times, a solution can be obtained.

is consistent. If it is inconsistent, then it is enough to solve inequality system $a_1^T y = c_1, \dots, a_{k-1}^T y = c_{k-1}$, $a_{k+1}^T y \leq c_{k+1}, \dots, a_\ell^T y \leq c_\ell$, since for all solutions of this system $a_k^T y < c_k$ holds. Let $\ell := \ell - 1$ and return to the previous step.

Step $k + 1$. If it is consistent then decide if the inequality system

$$a_1^T y = c_1, \dots, a_k^T y = c_k, a_{k+1}^T y = c_{k+1}, a_{k+2}^T y \leq c_{k+2}, \dots, a_\ell^T y \leq c_\ell$$

is consistent. ...

This algorithm finds a solution in polynomial time for inequality system $A^T y \leq c$, since in the last step an equation system is to be solved, which can be done by Gaussian elimination.

IPMs FOR LP

After decades of intensive research, deep understanding of IPMs has been developed. Easy to understand, simple variants of polynomial IPMs, as well as computationally highly efficient variants are published. The self-dual embedding model [2,7,15] provides an elegant, and powerful solution for the initialization problem of IPMs. As in case of the ellipsoid algorithm, we also demonstrate that IPMs not only converge to an optimal solution (if it exists), but after a finite number of iterations a strongly polynomial rounding procedure [2,16] generates an exact solution. This section is based on the paper of Terlaky [7], and Part I and the Appendix of Roos *et al.* [2].

Duality and the Self-Dual Embedding Model

We consider the general LP problem (P) and its dual (D) in the standard canonical form as defined in the section “LP Duality and Condition Numbers.” Theorem 1, the weak duality theorem and its corollary leads to the optimality conditions that are represented as the linear inequality system (FP). By introducing slack variables, and by homogenizing the inequality system, the Goldman–Tucker model [9,17] is obtained.

$$\begin{array}{rclcl} Ax - \tau b - z & = & 0, & x \geq 0, & z \geq 0 \\ -A^T y & + \tau c & -s & = & 0, & y \geq 0, & s \geq 0 \\ b^T y - c^T x & & -\rho & = & 0, & \tau \geq 0, & \rho \geq 0. \end{array} \quad (3)$$

It is easy to verify that for any solution (y, x, τ, z, s, ρ) of the Goldman–Tucker system (3) $\tau \rho > 0$ cannot hold. Further, any optimal pair (x, y) of problems (P) and (D) with zero duality gap is a solution of the Goldman–Tucker system with $\tau = 1$ and $\rho = 0$. We are looking for nontrivial solutions of this system with $\tau > 0$, because those give primal and dual optimal solutions $(\frac{x}{\tau}, \frac{y}{\tau})$ with zero duality gap, as $\tau > 0$ implies $\rho = 0$.

On the other hand, if the Goldman–Tucker system admits a feasible solution $(\hat{y}, \hat{x}, \hat{\tau}, \hat{z}, \hat{s}, \hat{\rho})$ with $\hat{\tau} = 0$ and $\hat{\rho} > 0$, then we may conclude that either of (P) or (D), or both of (P) and (D), are infeasible. Summarizing these results, we have the following theorem.

Theorem 7. *Let a primal dual pair (P) and (D) of LP problems be given. The following statements hold for the solutions of the Goldman–Tucker system (3).*

1. *Any optimal pair (x, y) of (P) and (D) with zero duality gap is a solution of the corresponding Goldman–Tucker system with $\tau = 1$.*
2. *If (y, x, τ, z, s, ρ) is a solution of the Goldman–Tucker system then either $\tau = 0$, or $\rho = 0$, or both, that is, $\tau \rho > 0$ cannot happen.*
3. *Any solution (y, x, τ, z, s, ρ) of the Goldman–Tucker system, where $\tau > 0$ and $\rho = 0$, gives a primal and dual optimal pair $(\frac{x}{\tau}, \frac{y}{\tau})$ with zero duality gap.*
4. *If the Goldman–Tucker system admits a feasible solution $(\hat{y}, \hat{x}, \hat{\tau}, \hat{z}, \hat{s}, \hat{\rho})$ with $\hat{\tau} = 0$ and $\hat{\rho} > 0$, then we may conclude that either (P), or (D), or both of them are infeasible.*

IPMs lead us to a solution of the Goldman–Tucker system where either $\tau > 0$ or $\rho > 0$. Thus, the IPM approach ensures $\tau + \rho > 0$. Now we give a simple form of the Goldman–Tucker model. Let

$$Mu \geq 0, \quad u \geq 0, \quad w(u) = Mu, \quad (4)$$

where

$$u = \begin{pmatrix} y \\ x \\ \tau \end{pmatrix}, \quad w(u) = \begin{pmatrix} z \\ s \\ \rho \end{pmatrix}, \quad \text{and} \\ M = \begin{pmatrix} 0 & A & -b \\ -A^T & 0 & c \\ b^T & -c^T & 0 \end{pmatrix}$$

is a skew-symmetric $n \times n$ matrix, that is, $M^T = -M$ and $n = m + \ell + 1$. The Goldman–Tucker Theorem [2,9,17] states that system (4) admits a *strictly complementary* solution.

Theorem 8 [Goldman–Tucker]. *System (4) has a strictly complementary feasible solution, i.e., a solution for which $uw(u) = 0$ and $u + w(u) > 0$.*

This theorem ensures that either case 3 or case 4 of Theorem 7 must occur when one solves the Goldman–Tucker system of LP. This is in fact the strong duality theorem of LP.

Theorem 9 [Strong duality for LP]. *Let a primal and dual LP problem be given. Exactly one of the following statements hold:*

- Both (P) and (D) are feasible, and there are optimal solutions x^* and y^* such that $c^T x^* = b^T y^*$.
- Either problem (P), or (D), or both of them are infeasible.

The Skew-Symmetric Self-Dual Embedding Model. Problem (4) is a homogeneous feasibility problem, and we transform it into an equivalent problem, which satisfies the interior point condition (IPC), that is, that admits a solution where $u > 0$ and $w(u) > 0$. This happens by embedding the problem and defining

an appropriate nonnegative vector q that will be both the right-hand side and the objective vector in the larger LP problem.

Let us take $u = w(u) = e$. These vectors are positive, but they do not satisfy the equality conditions in Equation (4). Let us further define the error vector r obtained this way by

$$r := e - Me \quad \text{and let} \quad \lambda := n + 1.$$

Then we have

$$\begin{pmatrix} M & r \\ -r^T & 0 \end{pmatrix} \begin{pmatrix} e \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ \lambda \end{pmatrix} = \begin{pmatrix} Me + r \\ -r^T e + \lambda \end{pmatrix} = \begin{pmatrix} e \\ 1 \end{pmatrix}.$$

Hence, the following problem

$$\begin{aligned} \min \left\{ \lambda \vartheta : \begin{pmatrix} M & r \\ -r^T & 0 \end{pmatrix} \begin{pmatrix} u \\ \vartheta \end{pmatrix} + \begin{pmatrix} w \\ v \end{pmatrix} \right. \\ \left. = \begin{pmatrix} 0 \\ \lambda \end{pmatrix}; \begin{pmatrix} u \\ \vartheta \end{pmatrix}, \begin{pmatrix} w \\ v \end{pmatrix} \geq 0 \right\} \quad (\overline{\text{SP}}) \end{aligned}$$

satisfies the IPC because the all-one vector is feasible for this problem. This problem is a self-dual skew-symmetric LP problem, because by introducing the notation

$$\overline{M} = \begin{pmatrix} M & r \\ -r^T & 0 \end{pmatrix}, \quad \overline{u} = \begin{pmatrix} u \\ \vartheta \end{pmatrix}, \quad \text{and} \quad \overline{q} = \begin{pmatrix} 0 \\ \lambda \end{pmatrix},$$

one can write problem $(\overline{\text{SP}})$ as

$$\min \left\{ \overline{q}^T \overline{u} : \overline{M} \overline{u} \geq -\overline{q}, \overline{x} \geq 0 \right\}. \quad (\text{SPe})$$

Finding a strictly complementary solution to (4) is equivalent to finding a strictly complementary optimal solution to problem (SPe). This claim is valid, because Equation (SPe) satisfies the IPC and it admits a strictly complementary optimal solution. As the objective function is just a constant multiple of ϑ , and the optimal value is zero, ϑ must be zero in any optimal solution. Having a strictly complementary solution of Equation (SPe), we either have an optimal solution of the embedded LP problem, or we can conclude that the LP problem does not have an optimal solution.

Basic Properties of the Skew-Symmetric Self-Dual Model. We continue to study our skew-symmetric self-dual embedding problem (SPe) while utilizing that the IPC holds for this LP problem. For ease of discussion let us simplify our notation, so the embedding problem (SPe) can be given as

$$\min \left\{ q^T u : Mu \geq -q, u \geq 0 \right\}, \quad (\text{SP})$$

where the matrix $M \in \mathbb{R}^{n \times n}$ is skew-symmetric, $0 \leq q \in \mathbb{Z}^n$, and problem (SP) satisfies the IPC, i.e., there exists an $u > 0$ with $Mu > 0$. The set of *feasible* and the set of *optimal* solutions of (SP) are denoted by

$$\begin{aligned} \text{SP} &:= \{u : u \geq 0, Mu \geq -q\} \quad \text{and} \\ \text{SP}^* &:= \{u : u \geq 0, Mu + q = w(u) \geq 0, \\ &\quad uw(u) = 0\}. \end{aligned}$$

The only nontrivial result that one needs to prove is the existence of a strictly complementary solution. The IPC implies this result by implying the existence of the *central path* [18–21] of problem (SP), which plays a crucial role in path-following IPMs.

Definition 2. Let the IPC be satisfied. The *central path* of problem (SP) is defined as the set of solutions

$$\begin{aligned} \{(u(\mu), w(u(\mu))) : Mu + q = w, \quad uw = \mu e, \\ u > 0, \quad \text{for some } \mu > 0\}. \end{aligned}$$

If no confusion is possible, instead of $w(u(\mu))$, the notation $w(\mu)$ will be used. The following key theorem establishes the existence of the central path. For an elementary proof—using only calculus—the reader is referred to Terlaky [7].

Theorem 10. *The next statements are equivalent.*

- (i) (SP) satisfies the interior point condition;
- (ii) For each $0 < \mu \in \mathbb{R}$ there exists a unique solution $(u(\mu), w(\mu)) > 0$ such that

$$Mu + q = w$$

$$uw = \mu e.$$

Analogous to the definitions in the section titled “Optimal Partition”, the *optimal partition* $(\mathbf{B}, \mathbf{N})$ is given by

$$\begin{aligned} \mathbf{B} &:= \{i : u_i > 0, \text{ for some } u \in \text{SP}^*\}, \\ \mathbf{N} &:= \{i : w(u)_i > 0, \text{ for some } u \in \text{SP}^*\}. \end{aligned}$$

The central path converges to a strictly complementary optimal solution, that is, if $u^* = \lim_{\mu \rightarrow 0} u(\mu)$ then $u^* w(u^*) = 0$ and $u^* + w(u^*) > 0$, and thus $\mathbf{B} \cup \mathbf{N} = \{1, \dots, n\}$. By using the embedding model, this result implies the Goldman–Tucker theorem (Theorem 8) for the general LP problem.

Theorem 11. *If the IPC holds then there exists an optimal solution u^* and $w^* = w(u^*)$ of problem (SP) such that $u^* w^* = 0$, $u_{\mathbf{B}}^* > 0$, $w_{\mathbf{N}}^* > 0$ and $u^* + w^* > 0$.*

A key result in this respect is that the central path converges to a unique point, the so-called *analytic center* [21] of the optimal set SP^* . Let $\bar{u} \in \text{SP}^*$, $\bar{w} = w(\bar{u})$ maximize the product $\prod_{i \in \mathbf{B}} u_i \prod_{i \in \mathbf{N}} w_i$ over $u \in \text{SP}^*$. Then $\bar{u}$ is called the *analytic center* of SP^* .

Theorem 12. *The limit point u^* of the central path is the analytic center of SP^* .*

Identifying the Optimal Partition. As in the section titled “Optimal Partition” we define a condition number of problem (SP), which characterizes the magnitude of the nonzero variables on the optimal set SP^* . Let

$$\sigma^u := \min_{i \in \mathbf{B}} \max_{u \in \text{SP}^*} \{u_i\} \quad \sigma^w := \min_{i \in \mathbf{N}} \max_{u \in \text{SP}^*} \{w(u)_i\}.$$

Then the *condition number* of problem (SP) is defined as

$$\sigma = \min\{\sigma^u, \sigma^w\} = \min_i \max_{u \in \text{SP}^*} \{u_i + w(u)_i\}.$$

As before, if M and q are integral and all the columns of M are nonzero. By using the

notation $\pi(M) = \prod_{i=1}^n \|M_i\|$, one can give an easily computable lower bound for σ as

$$\sigma \geq \frac{1}{\pi(M)}. \quad (5)$$

The Size of the Variables along the Central Path. By using the condition number σ , one can derive lower and upper bounds for the variables along the central path.

Lemma 5. *Let $(\mathbf{B}, \mathbf{N})$ be the optimal partition of problem (SP). For each positive μ one has*

$$u_i(\mu) \geq \frac{\sigma}{n} \quad i \in \mathbf{B}, \quad u_i(\mu) \leq \frac{n\mu}{\sigma} \quad i \in \mathbf{N},$$

$$w_i(\mu) \leq \frac{n\mu}{\sigma} \quad i \in \mathbf{B}, \quad w_i(\mu) \geq \frac{\sigma}{n} \quad i \in \mathbf{N}.$$

Identifying the Optimal Partition. For sufficiently small μ , the bounds presented in Lemma 5 enable us to identify the optimal partition $(\mathbf{B}, \mathbf{N})$. One just has to calculate the μ value that ensures that the coordinates going to zero are smaller than the coordinates that stay larger than $\frac{\sigma}{n}$. One can easily verify that having $\mu < \frac{\sigma^2}{n^2}$ and a central solution $u(\mu) \in SP$, the optimal partition $(\mathbf{B}, \mathbf{N})$ can be identified. Note that these results can be generalized to the case when a vector (u, w) is not on, but just in a neighborhood of the central path. For proofs the interested reader is referred to Roos *et al.* [2].

Rounding to an Exact Solution. A strictly complementary solution can be identified by moving along the central path as $\mu \rightarrow 0$. In fact, we do not have to do that; we can stop at a sufficiently small $\mu > 0$, and round-off the current “almost optimal” solution to a strictly complementary optimal one. We need some new notation. Let $\omega := \|M\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |M_{ij}|$ and $\pi := \pi(M) = \prod_{i=1}^n \|M_i\|$.

Lemma 6. *Let M and q be integral and all the columns of M be nonzero. If $(u, w) := (u(\mu), w(u(\mu)))$ is a central solution with*

$$u^T s = n\mu \leq \frac{\sigma^2}{n^{\frac{3}{2}}(1+\omega)^2\pi},$$

which certainly holds if

$$n\mu \leq \frac{1}{n^{\frac{3}{2}}(1+\omega)^2\pi^3},$$

then by a simple rounding procedure a strictly complementary optimal solution can be found in $\mathcal{O}(n^3)$ arithmetic operations.

Note that this rounding result can also be generalized to the situation when a vector (u, w) is not on, but just in a certain neighborhood of the central path. For details the reader is referred again to Roos *et al.* [2]. This result makes it clear that when one solves an LP problem by using an IPM, the iterative process can be stopped in a proper neighborhood of the central path at a sufficiently small value of μ , and at that point one can round-off to a strictly complementary optimal solution.

Optimal Basis Identification. It is well known that simplex methods provide an optimal basis. A reassuring fact is that having a complementary optimal solution pair, an optimal basis can be identified in strongly polynomial time, in at most n pivots; thus, an interior optimal solution pair allows efficient identification of an optimal basis. For a proof the reader may consult Roos *et al.* [2], Megiddo [22], and Ye [23].

Summary of the Theoretical Results

Let us summarize the results for the general LP problem in canonical form

$$\min \left\{ c^T x : Ax - z = b, x \geq 0, z \geq 0 \right\}, \quad (P)$$

$$\max \left\{ b^T y : A^T y + s = c, y \geq 0, s \geq 0 \right\}, \quad (D)$$

where for convenience the slack variables are already included in the problem formulation.

- In the section titled “Duality and the Self-Dual Embedding Model” we have seen that to solve the LP problem it is

sufficient to find a strictly complementary solution to the Goldman–Tucker model (4).

- This *homogeneous* system always admits the zero solution; however, to get a solution for the original problems we need a strictly complementary solution pair for which $\tau + \rho > 0$ holds.
- The Goldman–Tucker model can be transformed into a skew-symmetric self-dual problem (SP) satisfying the IPC.
- If problem (SP) satisfies the IPC then
 - the central path exists (Theorem 10);
 - the central path converges to a strictly complementary solution (Theorem 11);
 - the central path converges to the analytic center of the optimal set (Theorem 12);
 - if the problem data is integral and a solution on the central path with a sufficiently small μ is given, then the optimal partition and an exact strictly complementary optimal solution (Lemma 6) can be found.
- If (x^*, z^*) is optimal for (P) and (s^*, y^*) for (D) , then $(y^*, x^*, 1, z^*, s^*, 0)$ is a solution for the Goldman–Tucker model with the requested property $\tau + \rho > 0$ (Theorem 7).
- Any solution of the Goldman–Tucker model (y, x, τ, z, s, ρ) with $\tau > 0$ yields an optimal solution pair (scale the variables (x, z) and (y, s) by $\frac{1}{\tau}$) for LP (Theorem 7).
- Any solution of the Goldman–Tucker model (y, x, τ, z, s, ρ) with $\rho > 0$ provides a certificate of primal or dual infeasibility (Theorem 7).
- If $\tau = 0$ in every solution (y, x, τ, z, s, ρ) then (P) and (D) have no optimal solutions.
- From any complementary optimal pair of primal-dual solutions, an optimal basis can be identified in strongly polynomial time

In the next section a generic IPM is presented and the complexity of some variants is discussed.

A General Scheme of IP Algorithms for Linear Optimization

In this section we briefly review the main elements of IPMs. For easy reference to the majority of the literature, the algorithmic concepts are presented for the standard form of LP problems:

$$\min \{c^T x : Ax = b, x \geq 0\} \quad (P_s)$$

$$\max \{b^T y : A^T y + s = c, y \geq 0\}. \quad (D_s)$$

For this form, for any interior feasible $x, s > 0$ solution, the Newton system for the central path equations

$$\begin{array}{rcl} Ax & = & b \\ A^T y & +s & = c \\ xs & = & \mu e \end{array} \quad (6)$$

is given as

$$\begin{array}{rcl} A\Delta x & = & 0 \\ A^T \Delta y & +\Delta s & = 0 \\ s\Delta x & +x\Delta s & = \mu e - xs. \end{array} \quad (7)$$

Assuming that $\text{rank}(A) = m$, where A is an $m \times n$ matrix, and $x, s > 0$, the above Newton system (7) has a unique solution. Having determined this unique solution, a possibly damped Newton step can be made as

$$x := x + \alpha \Delta x, \quad s := s + \alpha \Delta s.$$

IPMs make full or damped Newton steps at each iterations, and the process is controlled by careful selection of the step length and the targeted central path parameter value μ or, in broader sense, the right-hand side of the third Newton equation. A key tool in this process is the choice of a proximity measure that allows to quantify the distance from the central path and ensure that the iterates remain positive.

Proximity Measures. We have seen that the central path leads to a specific strictly complementary solution, to the analytic center of the set of optimal solutions. However, due to the nonlinearity of the equation system determining the central path, the iterates cannot stay on the central path, even if one uses

the embedding model and the initial interior point was on the central path. For this reason we need some proximity measures that allow us to control the Newton process and keep the iterates in a proper neighborhood of the central path. These proximity measures depend on the current primal-dual iterate x and s , and the central path parameter value μ , that is either chosen independent of x and s or as $\mu = x^T s / n$. The goal is to quantify how far the iterate is from the point corresponding to μ on the central path. In general the proximity measure is denoted by $\delta(x, s, \mu)$.

Various centrality measures were developed in the literature of IPMs, see the books [2,15,23–26]. Here we present just two of them. Both of the proximity measures allow us to design polynomial IPMs. Observe that on the central path all the coordinates of the vector xs are equal. This observation indicates that the proximity measure

$$\delta_c(xs) := \frac{\max(xs)}{\min(xs)},$$

where $\max(xs)$ and $\min(xs)$ denote the largest and smallest coordinate of the vector xs , is an appropriate measure of centrality. One has $\delta_c(xs) \geq 1$, and the larger the value of $\delta_c(xs)$ is, the further the iterate is from the central path. If $\delta_c(xs) = 1$, then the current solution $(x, s) = (x(\mu), s(\mu))$, with $\mu = x^T s / n$, is on the central path. Another proximity measure, extensively used in Roos *et al.* [2] is

$$\delta_0(xs, \mu) := \frac{1}{2} \left\| \left(\frac{xs}{\mu} \right)^{\frac{1}{2}} - \left(\frac{\mu}{xs} \right)^{\frac{1}{2}} \right\|.$$

One has $\delta_0(xs, \mu) \geq 0$, and the larger the value of $\delta_0(xs)$ is, the further the iterate is from the target point on the central path. If $\delta_0(xs, \mu) = 0$, then the current solution is on the central path.

A Generic Interior Point Algorithm. Algorithm 1 gives a general framework for a large family polynomial time IPMs.

Algorithm 1. Generic Interior Point Newton Algorithm

Input:

A proximity parameter γ ; an accuracy parameter $\varepsilon > 0$;
a variable step length α ; update parameter $0 < \theta < 1$;
a feasible (x^0, s^0) and $\mu^0 > 0$ s.t. $\delta(x^0, s^0, \mu^0) \leq \gamma$.

begin

$x := x^0$; $s := s^0$; $\mu := \mu^0$;

while $n\mu \geq \varepsilon$ **do**

begin

$\mu := (1 - \theta)\mu$;

while $\delta(x, s, \mu) \geq \gamma$ **do**

begin

solve the Newton system (7);

choose the damping factor α ;

$x := x + \alpha \Delta x$;

$s := s + \alpha \Delta s$;

end

end

end

Proper choices of the parameters allow to prove polynomial iteration complexity of the resulting algorithm. Here we present three sets of choices that give polynomial IPMs. For complexity proofs, see for example, Roos *et al.*

[2] and the papers cited therein. To specify a concrete algorithm the following choices need to be made:

- choose the proximity measure $\delta(x, s, \mu)$, and the proximity parameter γ ,

- choose the update scheme for μ , and specify the step length for the Newton step.

The first IPM, that also enjoys the best possible iteration complexity bound, is a *primal-dual algorithm with full Newton steps* [2]. Let us make the following choices:

$$\begin{aligned} \delta(x, s, \mu) &:= \delta_0(xs, \mu); \quad \mu^0 := 1; \\ \theta &:= \frac{1}{2\sqrt{n}}; \quad \gamma = \frac{1}{\sqrt{2}}; \alpha = 1. \end{aligned}$$

Theorem 13 [2, Theorem II.52]. *With the given parameter set, the full Newton step algorithm requires at most*

$$\left\lceil 2\sqrt{n} \log \frac{n}{\varepsilon} \right\rceil$$

iterations to produce a feasible solution (x, s) such that $\delta_0(xs, \mu) \leq \gamma$ and $n\mu \leq \varepsilon$.

The second IPM is a *primal-dual large-update algorithm* [2]. The choices are

$$\begin{aligned} \delta(x, s, \mu) &:= \delta_0(xs, \mu); \\ \mu^0 &:= 1; \quad 0 < \theta < \frac{n}{n + \sqrt{n}}; \\ \gamma &= \frac{\sqrt{R}}{2\sqrt{1 + \sqrt{R}}}, \text{ where } R := \frac{\theta\sqrt{n}}{1 - \theta}; \end{aligned}$$

and α is the result of a line search, when along the search direction the primal-dual logarithmic barrier function $\frac{x^T s}{\mu} - \sum_{i=1}^n \log \frac{x_i s_i}{\mu}$ is minimized.

Theorem 14 [2, Theorem II.74]. *With the given parameter set the primal-dual large-update algorithm requires at most*

$$\left\lceil \frac{1}{\theta} \left[2 \left(1 + \sqrt{\frac{\theta\sqrt{n}}{1 - \theta}} \right)^4 \right] \log \frac{n}{\varepsilon} \right\rceil$$

iterations to produce a feasible solution (x, s) , such that $\delta_0(xs, \mu) \leq \tau$ and $n\mu \leq \varepsilon$.

When we choose $0 < \theta < 1$ as a constant independent of n , for example, $\theta = 0.9$, then the complexity becomes $\mathcal{O}(n \log \frac{n}{\varepsilon})$, while $\theta = \frac{\nu}{\sqrt{n}}$,

with any fixed positive value ν gives a complexity of $\mathcal{O}(\sqrt{n} \log \frac{n}{\varepsilon})$.

The third IPM is the *Dikin step algorithm* studied for example, in Roos *et al.* [2]. This is one of the simplest IPMs, with an extremely elementary complexity analysis. The price for simplicity is that the polynomial complexity result is not the best possible. For this algorithm we choose $\delta(x, s, \mu) := \delta_c(xs)$. Recall that this measure is always larger than or equal to 1. Further let $\gamma = 2$, $\mu^0 := 0$ imply that μ stays equal to zero, thus θ is irrelevant. The Newton step $(\Delta x, \Delta s)$ is the solution of Equation (7), when the right-hand side of the last equation is replaced by $-\frac{x^2 s^2}{\|xs\|}$, and finally let $\alpha = \frac{1}{2\sqrt{n}}$.

Theorem 15 [2, Theorem I.27]. *The Dikin step algorithm requires not more than*

$$\left\lceil 2n \log \frac{n}{\varepsilon} \right\rceil$$

iterations to produce a feasible solution (x, s) for $(\overline{\text{SP}})$ such that $\delta_c(xs) \leq 2$ and $n\mu \leq \varepsilon$.

EXTENSIONS

IPMs are developed much beyond LP. The books [23–25, 27, 28] discuss extensions of IPMs for classes of nonlinear problems. IPMs also provide polynomial time algorithms for second-order cone and semidefinite optimization problems that both have numerous interesting applications, not only in such traditional areas as combinatorial optimization [29] and control [30, 31], but also in various areas of engineering, including structural [30, 32] and electrical engineering [33, 34]. For surveys on algorithmic and complexity issues the reader may consult [23, 27, 35–40].

REFERENCES

1. Khachiyan L. A polynomial algorithm in linear programming. Sov Math Dokl 1979;20: 191–194.
2. Roos C, Terlaky T, Vial J.P. Interior point methods for linear optimization. 2nd ed. New York: Springer; 2006.
3. Karmarkar N. A new polynomial-time algorithm for linear programming. Combinatorica 1984;4:373–395.

4. Bixby R. Solving real-world linear programs: a decade and more of progress. *Oper Res* 2002;50:3–15.
5. Gondzio J, Terlaky T. A computational view of interior point methods for linear programming. In: Beasley J, editor. *Advances in linear and integer programming*. Oxford: Oxford University Press; 1996. pp. 103–185.
6. Andersen E, Gondzio J, Mészáros S, Xu X. Implementation of interior point methods for large scale linear programming. In: Terlaky T, editor. *Interior point methods of mathematical programming*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1996. pp. 189–252.
7. Terlaky T. An easy way to teach interior-point methods. *Eur J Oper Res* 2001;130:1–19.
8. Dantzig G. *Linear programming and extensions*. Princeton (NJ): Princeton University Press; 1963.
9. Goldman A, Tucker A. Theory of linear programming. In: Kuhn H, Tucker A, editors. *Volume 38, Linear inequalities and related systems. Annals of mathematical studies*. Princeton (NJ): Princeton University Press; 1956. pp. 53–97.
10. Ye Y. Toward probabilistic analysis of interior-point algorithms for linear programming. *Math Oper Res* 1994;19:38–52.
11. Ye Y, Pardalos P. A class of linear complementarity problems solvable in polynomial time. *Linear Algebra Appl* 1991;152:3–17.
12. Vavasis S, Ye Y. Condition numbers for polyhedra with real number data. *Oper Res Lett* 1995;17:209–214.
13. Klafszky E, Terlaky T. On the ellipsoid method. *Sigma* 1992;8:269–280. (in Hungarian).
14. Schrijver A. *Theory of linear and integer programming*. Chichester: John Wiley and Sons, Ltd.; 1986.
15. Ye Y, Todd M, Mizuno S. An $O(\sqrt{n}L)$ -iteration complexity homogeneous and self-dual linear programming algorithm. *Math Oper Res* 1994;19:53–67.
16. Mehrotra S, Ye Y. On finding the optimal facet of linear programs. *Math Program* 1993;62:497–515.
17. Tucker A. Dual systems of homogeneous linear relations. In: Kuhn H, Tucker A, editors. *Volume 38, Linear inequalities and related systems. Annals of mathematical studies*. Princeton (NJ): Princeton University Press; 1956. pp. 3–18.
18. Fiacco A, McCormick G. *Nonlinear programming: sequential unconstrained minimization techniques*. New York: John Wiley and Sons, Inc.; 1968.
19. Frisch R. *The logarithmic potential method for solving linear programming problems*. Oslo: University Institute of Economics; 1955.
20. Megiddo N. Pathways to the optimal set in linear programming. In: Megiddo N, editor. *Progress in mathematical programming: interior point and related methods*. New York: Springer; 1989. pp. 131–158.
21. Sonnevend G. An “analytic center” for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming. In: Prékopa A, Szelecsán J, Strazicky B, editors. *Volume 84, System Modeling and Optimization: Proceedings of the 12th IFIP-Conference, Lecture notes in control and information sciences*. Berlin: Springer; 1986. pp. 866–876.
22. Megiddo N. On finding primal and dual optimal bases. *ORSA J Comput* 1991;3:63–65.
23. Ye Y. *Interior-point algorithms: theory and analysis*. Wiley-interscience series in discrete mathematics and optimization. New York: John Wiley and Sons, Inc.; 1997.
24. Hertog D. *Interior point approach to linear, quadratic and convex programming*. Volume 277, *Mathematics and its applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1994.
25. Jansen B. *Interior point techniques in optimization. complexity, sensitivity and algorithms*. Volume 6, *Applied optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1996.
26. Wright S. *Primal-dual interior-point methods*. Philadelphia (PA): SIAM; 1997.
27. Nesterov Y, Nemirovski A. *Interior-point polynomial algorithms in convex programming*. Volume 13, *SIAM studies in applied mathematics*. Philadelphia (PA): SIAM Publications; 1994.
28. Terlaky T, editor. *Interior point methods of mathematical programming*. Volume 5, *Applied optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1996.
29. Alizadeh F. *Combinatorial optimization with interior point methods and semidefinite matrices [PhD thesis]*. Minneapolis (MN): University of Minnesota; 1991.
30. Ben-Tal A, Nemirovski A. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. MPS-SIAM

- series on optimization. Philadelphia (PA): SIAM; 2001.
31. Pólik I, Terlaky T. A survey of the S-Lemma. *SIAM Rev* 2007;49:371–418.
 32. de Klerk E, Roos C, Terlaky T. Semidefinite problems in truss topology optimization. Technical Report 95-128. The Netherlands: Faculty of Technical Mathematics and Informatics, TU Delft. 1995.
 33. Vandenberghe L, Boyd S. Semidefinite programming. *SIAM Rev* 1996;38:49–95.
 34. Boyd S, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004.
 35. de Klerk E, Roos C, Terlaky T. Initialization in semidefinite programming via a self-dual, skew-symmetric embedding. *Oper Res Lett* 1997;20:213–221.
 36. de Klerk E, Roos C, Terlaky T. Infeasible-start semidefinite programming algorithms via self-dual embeddings. In: Pardalos P, Wolkowicz H, editors. Topics in semidefinite and interior point methods. Volume 18, Fields institute communications. Providence (RI): AMS; 1998. pp. 215–236.
 37. de Klerk E, Roos C, Terlaky T. On primal-dual path-following algorithms for semidefinite programming. In: Gianessi F, Komlósi S, Rapcsák T, editors. New trends in mathematical programming. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1998. pp. 137–157.
 38. Nesterov Y, Todd M. Self-scaled barriers and interior-point methods for convex programming. *Math Oper Res* 1997;22:1–42.
 39. Sturm J. Primal-dual interior point approach to semidefinite programming. In: Frenk J, Roos C, Terlaky T, *et al.* High performance optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999.
 40. Wolkowicz H, Saigal R, Vandenberghe L, editors. Handbook of semidefinite programming: theory, algorithms, and applications. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000.

INTRODUCTION TO RARE-EVENT SIMULATION

JOHN F. SHORTLE

Department of Systems Engineering &
Operations Research, George Mason
University, Fairfax, Virginia

PIERRE L'ECUYER

Département d'Informatique et de
Recherche Opérationnelle, Université
de Montréal, Montréal (Québec),
Canada

INTRODUCTION

The evaluation of rare-event probabilities can pose significant challenges for simulation. To illustrate, let γ be the probability of a rare event. The standard way to estimate γ is to conduct n i.i.d. replications. Assuming that the outcome of each replication is either a 1 (denoting the occurrence of the rare event) or a 0 (denoting the nonoccurrence of the event), then the standard estimator $\hat{\gamma}$ is the number of observed occurrences divided by the number of replications. This is a binomial random variable with mean γ and variance

$$\text{Var}[\hat{\gamma}] = \frac{\gamma(1-\gamma)}{n} \approx \frac{\gamma}{n}.$$

Although $\text{Var}[\hat{\gamma}]$ gets smaller as the rare-event probability γ decreases, it is more useful to consider the *relative error* of the estimator, defined here as the standard deviation divided by the actual value

$$\text{Relative error} \equiv \frac{\sqrt{\text{Var}[\hat{\gamma}]}}{\gamma} \approx \frac{\sqrt{\gamma/n}}{\gamma} = \frac{1}{\sqrt{\gamma n}}. \quad (1)$$

The relative error can be interpreted as the root mean square error of the estimator $\hat{\gamma}$ divided by the root mean square error of the trivial estimator that always returns the

value zero. If $\hat{\gamma}$ is approximately normally distributed, the relative error provides a measure of the size of the confidence interval relative to the value being estimated (though “approximately normal” is not typical in the context of rare events). The basic problem is that the relative error increases as the rare-event probability gets smaller. More generally, the required number of replications grows inversely with γ . It also grows inversely with the square of the relative error. Table 1 shows examples of the required time to achieve a given relative error for a specified rare-event probability (assuming simulation speeds of 1 and 1000 replications per second). This illustrates that standard simulation may be impractical for many rare-event problems.

Rare-event simulation is used in many application areas, including finance and insurance where the rare event could be a large financial gain or loss, communication systems where the rare event could be the loss of important information, power systems where the rare event could be a major power outage, network reliability where the rare event could mean that certain nodes in the network are disconnected, nuclear physics where the rare event could be that particles cross a given protection shield, and aviation safety where the rare event could be an aircraft collision. It is also encountered in computer graphics (image synthesis by Monte Carlo methods), computational statistics, computational physics, computational biology, military applications, and so on. The goal is not always to estimate the probability of a rare event. It can be to estimate the mathematical expectation of a random variable that is strongly affected by rare events. But the main ideas and methods are basically the same as when the goal is to estimate the rare-event probability itself.

There are two main approaches for improving the efficiency of rare-event simulations—importance sampling (IS) and splitting. The idea of IS is to change the

Table 1. Computer Time Needed to Achieve a Specified Relative Error (Standard Simulation)

γ	Desired Relative Error			
	1 Replication per Second		1000 Replications per Second	
	100%	10%	100%	10%
10^{-3}	17 min	1 d	1 s	2 min
10^{-5}	1 d	4 mo	2 min	3 h
10^{-7}	4 mo	32 yr	3 h	12 d
10^{-9}	32 yr	3169 yr	12 d	3 yr

underlying sampling distribution so that rare events are more likely. The idea of splitting is to create separate copies of the simulation whenever the simulation gets “close” to the rare event of interest, effectively multiplying promising runs that are more likely to reach the rare event. Splitting is useful for systems that tend to take many incremental steps on the path to the rare event. IS is also useful for systems that tend to take a small number of “catastrophic jumps” to the rare event.

This article does not give a comprehensive literature review but simply offers some suggestions for further reading: The book edited by Rubino and Tuffin [1] contains survey chapters on a wide range of topics and applications in rare-event simulation. Much of the material in this article can be found in more elaborate form in the first three chapters of that book. More extensive coverage and lists of references on IS can be found in Heidelberger [2], Bucklew [3], Juneja and Shahabuddin [4], and Amussen and Glynn [5]. Good introductory references on splitting include Garvels [6] and L’Ecuyer *et al.* [7].

IMPORTANCE SAMPLING

We first consider IS as applied to a static problem. Let X be a random variable in d -dimensional space with density function f . Let $\mathcal{R}$ denote a rare-event set in $\mathbb{R}^d$. Let $\gamma = \Pr\{X \in \mathcal{R}\}$. The standard way to estimate γ using Monte Carlo simulation is to generate independent samples $X_1, \dots, X_n$ from the density f and to estimate γ by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n I_{\mathcal{R}}(X_i), \quad (2)$$

where $I_{\mathcal{R}}(x) = 1$ if $x \in \mathcal{R}$ and 0 otherwise. If γ is very small, the relative error $(\gamma n)^{-1/2}$ of this estimator can be very large. In fact, $\hat{\gamma}$ may turn out to be 0, in which case a confidence interval on γ computed in a standard way would also have zero width.

The idea of IS is to sample values from a different density function g rather than from the original density f . Presumably, the new density g is chosen so that the rare event is more likely to occur. Naturally, some adjustment must be made to correct for sampling from a different density function. The procedure is to generate samples $Y_1, \dots, Y_n$ from the density g and to estimate γ by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n \frac{f(Y_i)}{g(Y_i)} I_{\mathcal{R}}(Y_i) = \frac{1}{n} \sum_{i=1}^n L(Y_i) I_{\mathcal{R}}(Y_i), \quad (3)$$

where $L(y) \equiv f(y)/g(y)$ is the *likelihood ratio* of the two densities. To show that $\hat{\gamma}$ is unbiased, observe that

$$\begin{aligned} E \left[\frac{f(Y_i)}{g(Y_i)} I_{\mathcal{R}}(Y_i) \right] &= \int \frac{f(y)}{g(y)} I_{\mathcal{R}}(y) g(y) dy \\ &= \int I_{\mathcal{R}}(x) f(x) dx = \gamma. \end{aligned}$$

It is assumed that $g(y) > 0$ whenever $f(y) I_{\mathcal{R}}(y) > 0$.

A key objective is to find a change of measure that achieves significantly lower variance for the IS estimator in Equation (3) compared with the standard estimator in Equation (2). In other words, we seek to minimize the variance of $I_{\mathcal{R}}(Y_i) L(Y_i)$. In fact, it is possible to find a change of measure such

that $I_{\mathcal{R}}(Y_i)L(Y_i)$ has zero variance. This is achieved by the *optimal change of measure*

$$g(y) = \frac{1}{\gamma} f(y) I_{\mathcal{R}}(y) \quad \text{and} \quad L(y) = \frac{\gamma}{I_{\mathcal{R}}(y)}.$$

The optimal change of measure g has two key properties: (i) It is zero outside of $\mathcal{R}$, and (ii) it is proportional to the original density f inside of $\mathcal{R}$. The first property implies that $I_{\mathcal{R}}(Y_i)$ (sampled from the density g) is always 1. The second property implies that $L(Y_i) = f(Y_i)/g(Y_i)$ is always the same constant. Together, these two properties imply that $L(Y_i)I_{\mathcal{R}}(Y_i)$ has zero variance.

Practically speaking, finding the optimal change of measure requires knowing the rare-event probability γ , which eliminates the need for simulation in the first place. But the optimal change of measure is still useful in the sense that it provides a guideline for choosing a good change of measure. A good change of measure is mostly concentrated on the rare-event set $\mathcal{R}$ and is roughly proportional to the original density f on this set. Alternatively, $L(y)$ is small and roughly constant when $y \in \mathcal{R}$.

In the following examples, γ is known *a priori*; so, there is no need for simulation. We use these small examples just to illustrate the basic ideas.

Example (See [8]). Let X be an exponential random variable with density $f(x) = \lambda e^{-\lambda x}$, and $\mathcal{R} = \{x | x \geq T\}$, so $\gamma = e^{-\lambda T}$. The optimal change of measure is $g(y) = \lambda e^{-\lambda y} / e^{-\lambda T}$ for $y \geq T$ and $g(y) = 0$ elsewhere. But suppose for the sake of example that we only allow changes of measure that are linear scalings of the original density. In other words, the new sampling density is restricted to be exponential with a different parameter μ , namely, $g(y) = \mu e^{-\mu y}$. The likelihood ratio is

$$L(y) = \frac{\lambda e^{-\lambda y}}{\mu e^{-\mu y}} = (\lambda/\mu) e^{-(\lambda-\mu)y}.$$

Minimizing the variance of the IS estimator $\hat{\gamma}$ in Equation (3) is equivalent to minimizing the second moment of $L(Y_i)I_{\mathcal{R}}(Y_i)$

$$\begin{aligned} E[[L(Y_i)I_{\mathcal{R}}(Y_i)]^2] &= E[L^2(Y_i)I_{\mathcal{R}}(Y_i)] \\ &= \int_T^\infty (\lambda/\mu)^2 e^{-2(\lambda-\mu)y} g(y) dy \\ &= (\lambda^2/\mu) \int_T^\infty e^{-(2\lambda-\mu)y} dy \\ &= \frac{\lambda^2}{\mu(2\lambda-\mu)} e^{-(2\lambda-\mu)T}. \end{aligned} \quad (4)$$

The integral is finite and the equality is true only when $2\lambda - \mu > 0$; so, $\text{Var}[\hat{\gamma}]$ is finite only for $0 < \mu < 2\lambda$. In particular, $\text{Var}[\hat{\gamma}] \rightarrow \infty$ as $\mu \rightarrow 0$. This shows that simply forcing the samples to take on large values (i.e., by letting μ to be a small number) is not necessarily a good idea. Although a small value of μ makes $I_{\mathcal{R}}(Y_i)$ more likely to equal one, it may drastically increase the likelihood function $L(Y_i)$ for some $Y_i \in \mathcal{R}$ by making its denominator $g(Y_i)$ (the new density) much too small, thus increasing the overall variance of $I_{\mathcal{R}}(Y_i)L(Y_i)$. This highlights a key difficulty of IS in general: one must make sure that the probabilities (or densities) of sample realizations that contribute to the expectation being estimated are never reduced too much, because otherwise the estimator may take huge values occasionally and will then have a large variance. In typical multivariate situations, this can be very difficult to take care of.

Figure 1 shows the ratio of the IS-estimator variance and the standard-estimator variance as a function of the change-of-measure parameter μ . The figure shows results for two different problems ($T = 3$ and $T = 5$) with $\lambda = 1$. Intuitively, as might be expected, choosing a value of μ larger than λ (i.e., forcing the samples to take on smaller values) makes the IS estimator worse than standard simulation. Choosing a value of μ smaller than λ (i.e., forcing the samples to take larger values) typically makes the IS estimator better than standard simulation. However, μ cannot be too small, otherwise the variance can be much worse than standard simulation. This illustrates the fact that without being sufficiently careful, it is easy to choose an IS estimator that is much worse than the standard estimator.

It is easily verified that the value of μ that minimizes the second moment in Equation (4)

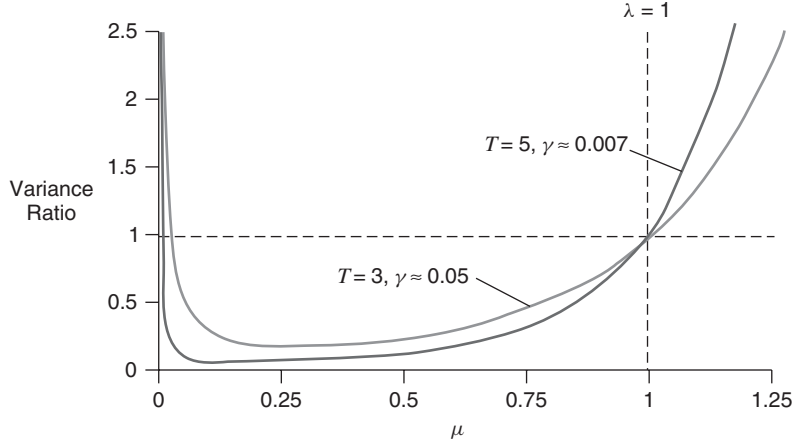


Figure 1. Dependence of variance ratio ($\text{Var}[L(Y_i)I_{\mathcal{R}}(Y_i)]/\text{Var}[I_{\mathcal{R}}(X_i)]$) as a function of μ .

is $\mu^* = \lambda + 1/T - (\lambda^2 + 1/T^2)^{1/2}$. This value is always smaller than λ and $1/T$. Also, $\mu^*T \rightarrow 1$ when $T \rightarrow \infty$. The squared relative error, namely, (Equation (4) $-\gamma^2/\gamma^2$, with the optimal μ is

$$\begin{aligned} \frac{\lambda^2}{\mu^*(2\lambda - \mu^*)} e^{\mu^*T} - 1 &\sim \frac{e}{2} \lambda T - 1 \\ &= \frac{e}{2} \ln(1/\gamma) - 1, \end{aligned}$$

where $u(\cdot) \sim v(\cdot)$ if $u(T)/v(T) \rightarrow 1$ as $T \rightarrow \infty$. Although the squared relative error increases when $\gamma \rightarrow 0$ (or $T \rightarrow \infty$), it only increases at a logarithmic speed in $1/\gamma$ (or linearly in T). In contrast, with crude Monte Carlo, it increases linearly in $1/\gamma$ (or exponentially in T). For example, suppose that $\lambda = 1$ and $T = 20$, so $\gamma = e^{-20}$. With crude Monte Carlo, it would take about $n \geq 100 \cdot (1/\gamma) \approx 5 \times 10^{10}$ replications to get a relative error below 10%. With the given IS scheme, it would take about $n \geq 100 \cdot [(e/2) \ln(1/\gamma) - 1] \approx 2618$ replications.

Very similar examples can be constructed with other distributions, for example a normal or a Poisson distribution for which we want to estimate the (small) probability that the random variable exceeds a given threshold.

With the IS scheme in the previous example, the relative error was still increasing in T , but more slowly than with crude

Monte Carlo. Ideally, we would like to have *bounded relative error*, meaning that the relative error remains bounded as $T \rightarrow \infty$, or more generally as $\gamma \rightarrow 0$. In this case, a fixed number of runs permits one to achieve a desired relative error regardless of how rare the event is. Here, we have a slightly weaker condition called *logarithmic efficiency* (or sometimes *asymptotic optimality*), namely, that

$$\lim_{\gamma \rightarrow 0} \frac{\ln E[\hat{\gamma}^2]}{\ln \gamma^2} = 1. \quad (5)$$

This means that the second moment of the estimator and the square of the mean go to zero at the same exponential rate when $\gamma \rightarrow 0$. This is almost as good as bounded relative error, and it can be obtained in a much broader range of situations. There are also situations where one can achieve *vanishing relative error* under IS; this means that the relative error converges to 0 when $\gamma \rightarrow 0$; so rarer events are easier to estimate [9,10]. These asymptotic properties of the second (relative) moment also generalize to moments of any order [10]. For example, *logarithmic efficiency of order k* means that Equation (5) holds with 2 replaced by k .

Now we consider IS applied to discrete-time Markov chains (DTMCs). Let $\mathcal{R}$ denote a set of rare-event states and let $\mathcal{O}$ denote a set of “other” terminating states. The problem is to estimate the probability γ that a

Markov chain enters $\mathcal{R}$ before $\mathcal{O}$. The idea of IS is to change the transition probabilities to reduce the variance of the rare-event estimator. Specifically, let $p(x, y)$ be the transition probability from x to y in the original chain, and let $p'(x, y)$ be the transition probability from x to y in a modified chain. The only restriction is that $p'(x, y)$ must be positive whenever $p(x, y)$ has a positive contribution to γ . Let $X_0, X_1, X_2, \dots$ and $Y_0, Y_1, Y_2, \dots$ be sets of states sampled from the original and modified chains. Let M and N be stopping times when the original and modified chains first enter either $\mathcal{R}$ or $\mathcal{O}$. The rare-event probability is $\gamma = \Pr\{X_M \in \mathcal{R}\}$.

Standard simulation estimates γ by simulating the original Markov chain to generate multiple samples of $X = I_{\mathcal{R}}(X_M)$. IS simulates the modified Markov chain to generate samples of

$$Y = I_{\mathcal{R}}(Y_N)L(Y_0, Y_1, \dots, Y_N),$$

where the likelihood ratio $L(\cdot)$ of a sample path is

$$L(Y_0, Y_1, \dots, Y_N) = \frac{\prod_{n=1}^N p(Y_{n-1}, Y_n)}{\prod_{n=1}^N p'(Y_{n-1}, Y_n)}. \quad (6)$$

As in the static case, the modified sampling is unbiased: $E[Y] = \gamma$. The main challenge is to find a change of measure so that $\text{Var}[Y] \ll \text{Var}[X]$. As before, it is possible (in theory) to find a change of measure such that $\text{Var}[Y] = 0$. Specifically, the optimal change of measure is

$$p'(x, y) = \frac{\gamma(y)}{\gamma(x)} p(x, y), \quad (7)$$

where $\gamma(x) = \Pr\{X_M \in \mathcal{R} | X_0 = x\}$ is the probability that the original chain enters $\mathcal{R}$ before $\mathcal{O}$, starting in state x . Note that $p'(x, y) = 0$ when $y \in \mathcal{O}$, since $\gamma(y) = 0$ when $y \in \mathcal{O}$. Thus, the optimal change of measure cuts all links to terminating states that are not rare-event states. This forces the chain to eventually hit the rare-event set with probability 1. To verify that $\text{Var}[Y] = 0$

$$\begin{aligned} Y &= I_{\mathcal{R}}(Y_N)L(Y_0, Y_1, \dots, Y_N) \\ &= I_{\mathcal{R}}(Y_N) \frac{\prod_{n=1}^N p(Y_{n-1}, Y_n)}{\prod_{n=1}^N p'(Y_{n-1}, Y_n)} \\ &= I_{\mathcal{R}}(Y_N) \frac{\prod_{n=1}^N p(Y_{n-1}, Y_n) \gamma(Y_{n-1})}{\prod_{n=1}^N p(Y_{n-1}, Y_n) \gamma(Y_n)} \\ &= I_{\mathcal{R}}(Y_N) \frac{\gamma(Y_0)}{\gamma(Y_N)} = \gamma(Y_0). \end{aligned}$$

The last equality follows because $\gamma(Y_N) = I_{\mathcal{R}}(Y_N)$, since Y_N is in a terminating state in either $\mathcal{R}$ or $\mathcal{O}$. Without loss of generality, we can assume that Y_0 is a constant, so $\text{Var}[Y] = \text{Var}[\gamma(Y_0)] = 0$. (If Y_0 is random, define a pseudo-initial state at time -1 that transitions to one of the random initial states at time 0.)

Example. Consider an $M/M/1/K$ queue with arrival rate λ , service rate μ , and maximum system capacity K . The problem is to estimate the probability γ of reaching K customers in the system before returning to the empty state, starting from the empty state. This is equivalent to the gambler's ruin problem in which a gambler makes i.i.d. bets, winning \$1 with probability $p = \lambda/(\lambda + \mu)$ and losing \$1 with probability $q = \mu/(\lambda + \mu)$. The probability $\gamma(n)$ of reaching K in the system before returning to the empty state (or getting K dollars before losing everything), starting with n in the system, has a simple analytic expression [11]

$$\gamma(n) = \frac{1 - (q/p)^n}{1 - (q/p)^K} = \frac{1 - (1/\rho)^n}{1 - (1/\rho)^K}, \quad (8)$$

where $\rho = \lambda/\mu$. Figure 2 shows the transition probabilities of the original chain and a modified chain. From Equation (7), the optimal change of measure is

$$\begin{aligned} p'_n &= \frac{\gamma(n+1)}{\gamma(n)} p = \frac{(1/\rho)^{n+1} - 1}{(1/\rho)^n - 1} p, \\ n &= 1, 2, \dots, K-1. \end{aligned} \quad (9)$$

As $n \rightarrow \infty$, $p'_n \rightarrow (1/\rho)p = q$ (assuming $\rho < 1$). Thus, for large n , the optimal change of measure approximately swaps the arrival and service rates.

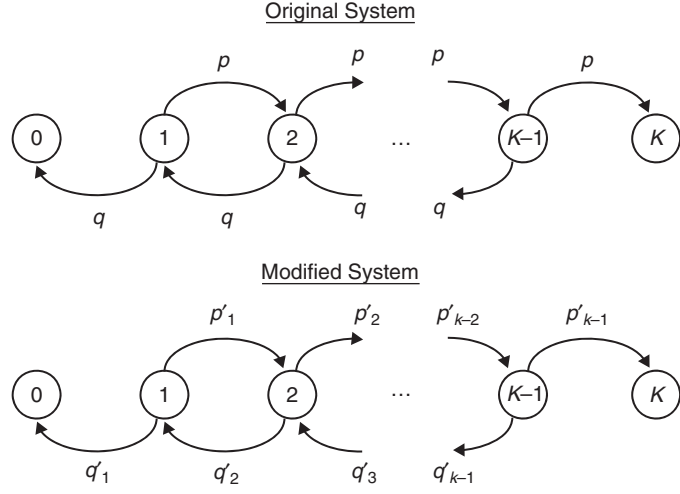


Figure 2. $M/M/1/K$ transition probabilities.

Since the optimal change of measure requires the knowledge of the rare-event function $\gamma(\cdot)$, we may wish to consider a restricted set of measures in which the modified transition probabilities are *state independent*, namely, $p'_n \equiv p'$ for some value p' . Intuitively, a larger value for p' makes the rare event more likely, but if it is too large this can be detrimental. To illustrate, suppose that $p' = 1 - \epsilon$, where $\epsilon > 0$ is some small number. The most likely path to $\mathcal{R}$ is a sequence of upward jumps $Y_0 = 1, Y_1 = 2, \dots, Y_{K-1} = K$. This path occurs with probability $(1 - \epsilon)^{K-1} \approx 1$. The likelihood ratio of this path is $[p/(1 - \epsilon)]^{K-1}$. Occasionally, a sample path may backtrack one state before continuing with upward jumps (e.g., $1, 2, 3, 2, 3, 4, \dots, K$). The likelihood ratio of this path is $qp^K/\epsilon(1 - \epsilon)^K$. In other words, the loop from state 3 to 2 to 3 increases the likelihood ratio by the factor $qp/\epsilon(1 - \epsilon)$. This path occurs with probability $\epsilon(1 - \epsilon)^K$; so it contributes the following term to the *second moment* of the likelihood ratio

$$\left(\frac{qp^K}{\epsilon(1 - \epsilon)^K} \right)^2 \epsilon(1 - \epsilon)^K \approx \frac{q^2 p^{2K}}{\epsilon}.$$

This implies that the variance of the overall estimator $Y = I_{\mathcal{R}}(Y_N)L(Y_0, \dots, Y_N)$ can be made arbitrarily large for small enough ϵ . In summary, although $p' = 1 - \epsilon$ has the positive effect of making the rare event very likely ($I_{\mathcal{R}}(Y_N)$ is usually 1), it has the

negative effect of creating the potential for very large likelihood ratios, which acts to increase the variance of the overall estimator $I_{\mathcal{R}}(Y_N)L(Y_0, \dots, Y_N)$.

A better change of measure makes the rare event very likely but also keeps the likelihood ratio roughly constant over these sample paths. This illustrates why switching the arrival and service rates (i.e., $p' = q$ and $q' = p$) is close to optimal. A loop from n to $n - 1$ to n modifies the likelihood ratio by a factor $qp/q'p' = 1$, leaving the likelihood ratio unchanged. The true optimal change of measure in Equation (9) additionally cuts the link to state 0, and also keeps the property that the likelihood ratio remains unchanged over loops (i.e., $qp/q'_n p'_{n-1} = 1$).

The previous one-dimensional examples are relatively simple in that the rare-event probabilities are known and the optimal change of measure can be found exactly. In practical problems, the optimal change of measure is not known and the Markov chains typically have multidimensional states, which increases the difficulty. To some extent, the one-dimensional examples are deceiving because the sample paths that lead to the rare event are readily identified, which is hardly true in higher dimensional settings. We briefly discuss a few approaches for handling more realistic problems. For more detailed summaries of these methods See [4,8].

A first idea is to make heuristic adjustments to push the model toward the rare event. For instance, in the first example, if one were to make an “educated guess” for μ without the aid of Fig. 1, one could probably choose an IS estimator with a smaller variance than standard simulation, but there is no guarantee. If the system is pushed too hard toward the rare event, then the IS estimator may be much worse than standard simulation. In multivariate state spaces, the variety of trajectories that lead to the rare event increases very fast with the dimension and the variance may explode if the densities of those trajectories are not inflated evenly between them. This is hard to achieve without a good understanding of the system behavior.

A more systematic approach, which dates back to the early days of IS in the 1950s, is to approximate the optimal change of measure via an approximation of $\gamma(\cdot)$ by learning it in an adaptive fashion. For example, the state space of the Markov chain can be discretized into a finite number of regions, and the system can be simulated for a period of time to get an initial estimate for $\gamma(\cdot)$ by assuming a simple form (perhaps a constant function) in each region. Then, a change of measure based on Equation (7) can be applied. The system can be simulated again using the new change of measure, and this process can be repeated iteratively. More generally, one can define a parameterized approximation of $\gamma(\cdot)$ with a few parameters and learn a good set of parameter values, for example, by stochastic approximation. These approaches require one to learn and store an approximation of $\gamma(\cdot)$ over the entire state space, and they hit the curse of dimensionality as soon as the dimension of the state space exceeds a few units. They are also impractical when the state space is already discrete but too large.

Asymptotic approximations of $\gamma(\cdot)$ are sometimes readily available in asymptotic regimes where the rare-event probability converges to 0. Plugging these approximations directly into Equation (7) often works nicely when the rare-event probability is very small. For distributions with an exponentially decaying tail, the resulting IS density $g(x)$ is often the original density $f(x)$

multiplied by an exponential factor $e^{\theta x}/K(\theta)$ for some parameter θ , where $K(\theta)$ is a normalizing factor. This is called *exponential twisting* [4,5].

Consider, for example, a sequence of i.i.d. continuous random variables $X_1, X_2, \dots$, with $E[X_i] < 0$ and finite moment generating function around zero so that X_i is light tailed. Define a random walk $\{S_i, i \geq 0\}$ by $S_0 = 0$ and $S_i = S_{i-1} + X_i$. We want to estimate $\gamma = \Pr\{\sup_{i \geq 0} S_i \geq \ell\}$, the probability that the walk ever exceeds a given constant $\ell > 0$, by simulating it as a Markov chain. This model has several applications. For example, it can be used to obtain the steady-state waiting-time distribution of a $GI/G/1$ queue [12, Section 7.2.3]. Large-deviation theory tells us that when γ is small, $\gamma \approx e^{-\theta \ell}$ for some constant θ that depends on the distribution of X_i . Plugging this approximation into Equation (7) (with probabilities replaced by densities) gives an exponentially twisted density that is typically very effective for small γ . The old density $f(x)$ is multiplied by $e^{\theta x}$ for all x and then renormalized.

Another approach to take when the state space is large is to restrict the change of measure to a specific parametric class. Naturally, the parametric class should be chosen so that it includes good measures to begin with. The problem is to estimate the vector of parameters θ that minimizes the variance (or equivalently the second moment $E[\gamma^2]$) within the parametric class. For a given parameter set θ , estimates of $E[\gamma^2]$ are obtained by sample averages from simulation. Minimization of $E[\gamma^2]$ over θ can be conducted using methods from stochastic optimization. A variant of this is the cross-entropy method which, instead of minimizing the variance, minimizes the cross-entropy distance between the selected change of measure and the zero-variance measure [13].

SPLITTING

The idea of splitting is to “split” (or clone) a simulation run into separate runs whenever it gets “near” the rare event of interest (Fig. 3). These runs share a common history up to the splitting point, but conditional on

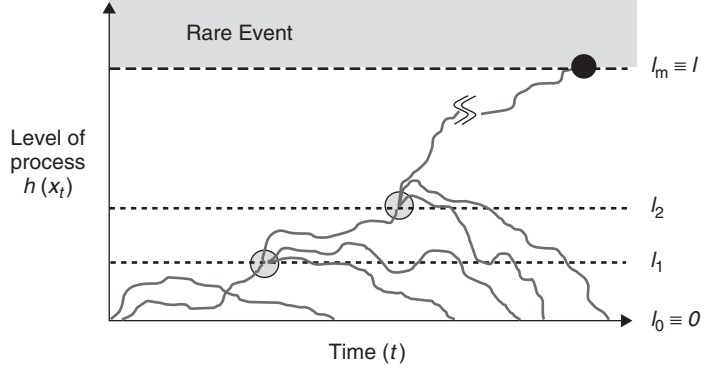


Figure 3. Level splitting.

that history, they evolve independent of each other. In this way, more computer time is spent on promising runs that are close to the rare event. To make this more concrete, let $X \equiv \{X_t, t \geq 0\}$ be a Markov process (possibly multidimensional) with state space χ . Let $h : \chi \rightarrow \mathbb{R}^+$ be a map of the state space to the “level” of the process; $h(\cdot)$ is sometimes called the *importance function*. Figure 3 shows sample paths of $h(X_t)$. The rare-event set $\mathcal{R}$ is defined as the set of states whose level is at least as large as some constant $l > 0$. In other words, $\mathcal{R} \equiv \{x \in \chi : h(x) \geq l\}$. Let T_R be the time that the process first enters $\mathcal{R}$, and let T_S be the time that the process first returns to level 0. Specifically,

$$T_R \equiv \inf \{t > 0 : h(X_t) \geq l\},$$

$$T_S \equiv \inf \{t > 0 : h(X_t) = 0\},$$

where it is assumed that the process has just left level 0 at time $t = 0$ ($h(X_0) > 0$ and $h(X_{0-}) = 0$). The probability to estimate is $\gamma \equiv \Pr\{T_R < T_S\}$. For example, if $h(X_t)$ denotes the number of customers in a queueing system at time t , then $\{T_R < T_S\}$ is the event that the number in the system reaches some critical threshold (say, a buffer overflow) before the system empties.

To implement level splitting, define a sequence of m levels $0 < l_1 < \dots < l_m$, where $l_m \equiv l$. Let $T_j \equiv \inf\{t > 0 : h(X_t) \geq l_j\}$ be the time that the process first crosses l_j ($j = 1, \dots, m$). Let $D_j \equiv \{T_j < T_S\}$ be the event that the process crosses l_j before returning to 0 ($j = 1, \dots, m$). Let $p_j \equiv \Pr\{D_j | D_{j-1}\}$ be the probability that the process crosses

l_j (before returning to 0), given that the process has crossed l_{j-1} (before returning to 0) ($j = 2, \dots, m$). Also, let $p_1 \equiv \Pr\{D_1\}$. Since $D_m \subset D_{m-1} \subset \dots \subset D_1$,

$$\gamma = \Pr\{D_m\} = \Pr\{D_1\}\Pr\{D_2|D_1\}$$

$$\dots \Pr\{D_m|D_{m-1}\} = p_1 p_2 \dots p_m.$$

The basic idea is to estimate each p_j separately, rather than to estimate γ on its own. There are many variations of level splitting, and we describe some of the key ideas here.

Fixed Splitting

In this approach, any simulation run that reaches a new level is split into a fixed number R of independent runs. (The splitting factor can also be generalized so that it is level dependent, R_i .) For example, in Fig. 3, the splitting factor is $R = 3$. A run that down-crosses and up-crosses a level is only split at the first up-crossing. Each run that reaches the rare event comes from a sequence of m splits, each by a factor R . Thus, each hit to the rare event is weighted by the factor $1/R^m$ to produce an unbiased estimator.

An advantage of fixed splitting is that it can be implemented recursively in a depth-first manner. This means that the computer only needs to store, at most, a single system state per level, corresponding to the simulation history of the current run. To implement fixed splitting, the simulation starts at an initial state and proceeds until it reaches either level 1 or level 0. If level 0 is reached, the replication terminates; if level 1 is reached,

the simulation is split into R clones. The first clone is simulated until it reaches either level 2 or level 0. If level 0 is reached, the clone terminates, and the next clone at level 1 is simulated; if level 2 is reached, the simulation is split again, and so forth. This process continues so that all offsprings of a clone are simulated to conclusion before proceeding to the next clone.

One drawback with fixed splitting is that it is sensitive to the choice of the splitting factor R . If R is too high, the total number of runs and the amount of work explode. If R is too low, then few simulations reach the rare event.

Fixed-Effort Splitting

In this approach, a predetermined total number of runs n_i are made at each level. This is different from fixed splitting where the *total* number of runs at each level is random based on the outcomes of runs at lower levels. The main advantage is that there is no need to know a good splitting factor R in advance, and because the total number of runs at each level is fixed, the overall variance is often lower. The drawback is that the computer must store all of the entrance states to a level to serve as the candidate starting states for simulation of the next level. This can lead to memory problems for models with large state spaces.

One way to implement fixed-effort splitting is the following: conduct n_1 runs starting from the initial state and simulate each until the system reaches level 1 or returns to level 0. The end states of the runs that reach level 1 are collected into a set A_1 . These states become the starting states for the next stage of simulation. (A_1 may include duplicate copies of the same state.) Simulation at stage 2 draws n_2 starting states at random, with replacement, from the set A_1 and simulates each run until the system reaches level 2 or returns to level 0. The method proceeds in an analogous manner at higher stages. After that, an unbiased estimator for the rare-event probability is $\hat{\gamma} \equiv \hat{p}_1 \hat{p}_2 \dots \hat{p}_m$, where $\hat{p}_i$ is the number of runs that reach level i divided by n_i .

Fixed-Success Splitting and Fixed Probability of Success

In *fixed-success splitting* [14], at each level, the total number of trajectories that must reach the next level is fixed; independent replications at the current level continue until the fixed number of successes is achieved. According to Amrein and Künsch [15], this approach is often superior to fixed-splitting and fixed-effort approaches. In *fixed probability of success*, the levels are learned adaptively so that the probability of reaching the next level is approximately the same at all levels.

By far the most difficult issue in splitting, when the state space has more than one dimension, is the choice of the importance function $h(\cdot)$. In one dimension, the choice is typically straightforward. For example, for an $M/M/1$ queue, a natural choice is $h(x) = x$, where x is the number of customers in the system. In fact, all increasing functions in x are equivalent, provided one is free to select the location of the levels. The problem becomes more complicated when the state space is multidimensional. To illustrate, consider a two-stage Markovian tandem queue example, taken from Glasserman *et al.* [16]. Let (x_1, x_2) represent the numbers of customers at each station, and let γ be the probability of a buffer overflow at the *second* queue (prior to returning to an empty system). An intuitive choice for the level function is $h(x_1, x_2) = x_2$, since $\mathcal{R}$ is defined purely in terms of x_2 . This turns out to be a bad choice when the first queue is a bottleneck. The reason is that h does not adequately capture *how* the system gets to the rare event. The system typically gets to $\mathcal{R}$ by building up customers at the first queue and then transferring them to the second queue. Thus, a system in state $(x_1 = \text{large}, x_2 = \text{small})$ is “close” to the rare event even though $h(x_1, x_2) = x_2$ is small indicating that the system is far away from $\mathcal{R}$. Thus, h does not adequately favor sample paths that are likely to reach the rare event. A better choice is $h(x_1, x_2) = [x_2 + \min(0, x_2 + x_1 - l)]/2$, where l is the overflow level of the second queue [14,17]. Intuitively, this function is proportional to the minimal number of steps to the rare event (arrivals to the first queue and

transfers to the second queue). In summary, a good choice for the importance function is not necessarily obvious, and a poor choice can increase the variance over crude Monte Carlo. The same type of difficulty occurs when IS is applied to multidimensional systems: the change of measure must favor the sample paths that are representative of how the model reaches the rare event; otherwise the variance can blow up.

One practical issue in splitting is that it can take a long time to simulate from a high level down to level 0. This is inefficient, since a lot of time is spent simulating descending runs that are not likely to rise back up to hit higher levels. The idea of *truncation* is to terminate runs that sink below some threshold under the assumption that they will eventually sink to 0. This reduces computation time but introduces a bias. *Probabilistic truncation* implements this idea in an unbiased way: consider a run that starts at level k and down-crosses level $k - j$, where $j \geq 1$. The run continues with probability $1/r_{kj}$ and is terminated with probability $1 - 1/r_{kj}$, where $r_{kj} \geq 1$ are pre-chosen constants. A run that continues has its weights multiplied by r_{kj} to eliminate the bias. The general idea of killing some trajectories with some probability and inflating their weight when they survive is called *Russian roulette* in the Monte Carlo literature. Other variations include periodic truncation and tag-based truncation. A well-known variant of splitting that incorporates truncation is the RESTART method [18].

To see the kinds of improvements that are possible with splitting, consider a problem in which the levels are chosen so that the probabilities p_i are equal, namely, $p_i = \gamma^{1/m}$. For example, in the $M/M/1/K$ model, choosing evenly spaced levels l_i (e.g., $l_i = ki$ for some constant $k > 0$) yields approximately equal values for p_i (at least for large i). Under a fixed-effort implementation with $n_i \equiv n$ and $p_i = \gamma^{1/m}$, it is relatively straightforward to show [7] that splitting reduces $\text{Var}[\hat{\gamma}]$ by a factor $m^2 \gamma^{1-1/m}$ compared with standard simulation. Taking the derivative with respect to m shows that $m \approx -(\ln \gamma)/2$ achieves the best improvement. For example, if $\gamma = 10^{-9}$, then it is roughly optimal to select $m = 10$ levels yielding a variance

reduction of approximately six orders of magnitude. However, it is not always easy to choose levels so that the p_i are equal. One adaptive approach is to conduct an initial set of simulations to approximate the appropriate levels [6, Section 3.3]. Also, this analysis implicitly ignores the actual time to simulate each level. In practice, it typically takes longer to simulate runs that start at higher levels. This reduces the variance reduction achieved for a fixed computing budget.

Splitting can also be applied without pre-defining any levels for the importance function. A splitting or Russian roulette decision can be made at any step of the chain depending on its current weight and current value of the importance function [7]. Finally, particle filters and sequential Monte Carlo methods, popular in computational statistics, are closely related to fixed-effort splitting; see Andrieu *et al.* [19] and the references given therein.

REFERENCES

1. Rubino G, Tuffin B, editors. Rare event simulation using Monte Carlo methods. Chichester: Wiley; 2009.
2. Heidelberger P. Fast simulation of rare events in queueing and reliability models. ACM Trans Model Comput Simul 1995;5:43–85.
3. Bucklew JA, editor. Introduction to rare event simulation. New York: Springer; 2004.
4. Juneja S, Shahabuddin P. Rare event simulation techniques: an introduction and recent advances. In: Henderson SG, Nelson BL, editors. Handbook in operations research and management science: simulation. Oxford (UK): Elsevier; 2006. pp. 291–350.
5. Asmussen S, Glynn PW. Stochastic simulation. New York: Springer; 2007.
6. Garvels M. The splitting method in rare event simulation [PhD. thesis]. The Netherlands: University of Twente; 2000.
7. L'Ecuyer P, Demers V, Tuffin B. Splitting for rare-event simulation. In: Perrone LF, Wieland FP, Liu J, editors. Proceedings of 2006 Winter Simulation Conference. Piscataway (NJ): IEEE; 2006. pp. 137–148.
8. L'Ecuyer P, Mandjes M, Tuffin B. Importance sampling and rare event simulation. In:

- Rubino G, Tuffin B, editors. Rare event simulation using Monte Carlo methods. Chichester (UK): Wiley; 2009. Chapter 2. pp. 17–38.
9. L'Ecuyer P, Tuffin B. Approximating zero-variance importance sampling in a reliability setting. *Ann Oper Res* 2009. DOI: 10.1007/s10479-009-0532-5.
 10. L'Ecuyer P, Blanchet JH, Tuffin B, *et al.* Asymptotic robustness of estimators in rare-event simulation. *ACM Trans Model Comput Simul* 2010;20: Article 6.
 11. Ross SM. An introduction to probability models. 9th ed. New York: Academic Press; 2007.
 12. Gross D, Shortle J, Thompson J, *et al.* Fundamentals of queueing theory. 4th ed. Hoboken (NJ): Wiley; 2008.
 13. Rubinstein R, Kroese DP. A unified approach to combinatorial optimization, Monte Carlo simulation, and machine learning. Berlin: Springer; 2004.
 14. L'Ecuyer P, LeGland F, Lezaud P, *et al.* Splitting techniques. In: Rubino G, Tuffin B, editors. Rare event simulation using Monte Carlo methods. Chichester (UK): Wiley; 2009. Chapter 3. pp. 39–62.
 15. Amrein M, Künsch H. A variant of importance splitting for rare event estimation: fixed number of successes. *ACM Trans Model Comput Simul* 2011;21: Article 12.
 16. Glasserman P, Heidelberger P, Shahabuddin P, *et al.* A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans Autom Contr* 1998;43:1666–1679.
 17. L'Ecuyer P, Demers V, Tuffin B. Rare-events, splitting, and quasi-Monte Carlo. *ACM Trans Model Comput Simul* 2007;17: Article 9.
 18. Villén-Altamirano M, Villén-Altamirano J. RESTART: a straightforward method for fast simulation of rare events. *Proceedings of the 1994 Winter Simulation Conference*. Piscataway (NJ): IEEE; 1994. pp. 282–289.
 19. Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. *J R Stat Soc Ser B* 2010;72:1–33.

INTRODUCTION TO ROBUST OPTIMIZATION

J. COLE SMITH

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

SHABBIR AHMED

H. Milton School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

OPTIMIZATION UNDER UNCERTAINTY

Mathematical programming models involving uncertain parameters were introduced as early as 1955, with Dantzig's work on linear programming under uncertainty [1]. To illustrate the various approaches for dealing with uncertainty in optimization, we first introduce a generic deterministic optimization problem

$$\min f(\mathbf{x}, \xi) \quad \text{s.t. } \mathbf{x} \in X(\xi), \quad (1)$$

where $\xi \in \mathbb{R}^d$ is a data vector, $\mathbf{x}$ is the decision vector, and where $X(\xi) \subseteq \mathbb{R}^n$ and $f(\cdot, \xi) : \mathbb{R}^n \mapsto \mathbb{R}$ are the feasible region (defined by constraints) and the objective function, respectively, corresponding to the data vector ξ . The goal is to select a decision $\mathbf{x}$ from the feasible set of decisions $X(\xi)$ such that the objective function $f(\mathbf{x}, \xi)$ is minimized. If the data ξ defining f and X are deterministic and completely observable by the decision maker in advance of the optimization process, then one turns to a suite of algorithms dedicated to solving such problems. In most practical settings, however, this assumption does not hold. For instance, it may be impossible to specify future prices, for example, as in the case of portfolio investments [2,3]. Errors may also arise due to numerical instability or measurement inaccuracies [4]. Furthermore, even if exact data has been provided to the decision maker, the implementation

of a continuous-valued solution that involves several digits of precision may be impossible (e.g., as in Ref. 5).

To properly address Equation (1) in the presence of uncertainty on ξ , we need an appropriate model of uncertainty of ξ . There are three common paradigms: (i) In the stochastic programming paradigm, the uncertain parameter ξ is modeled as a random variable ξ with a known distribution. Thus, for a given solution $\mathbf{x}$, feasibility $\mathbf{x} \in X(\xi)$ is a random event and the objective function $f(\mathbf{x}, \xi)$ is a random variable, and stochastic programming models seek a solution $\mathbf{x}$ that is feasible with sufficiently high probability and minimizes some statistic, say expectation, of the random objective. (ii) In contrast, the robust optimization paradigm models the uncertainty in ξ via an uncertainty set $\mathcal{U}$, and seeks a solution $\mathbf{x}$ that is feasible for all possible realizations of $\xi \in \mathcal{U}$ and minimizes the worst-case objective $\max_{\xi \in \mathcal{U}} \{f(\mathbf{x}, \xi)\}$. (iii) A middle ground between the previous two extremes are stochastic programs with partial distributional information. In this setting, a family of distributions on the uncertain parameter ξ is assumed, and a solution that is feasible with high probability and minimizes some statistic over the worst-case distribution from the family is sought. In this article we restrict our focus to paradigm (ii). At first glance, the robust optimization framework may appear overly pessimistic. However, the conservativeness of the approach can be controlled by appropriately constructing the uncertainty set. Much of the research in this area is on constructing appropriate uncertainty sets that

- properly capture the underlying uncertainty and the decision maker's aversion to it; and
- lead to tractable optimization models.

There is a huge body of literature on a variety of robust optimization models. Here, we describe some of the simplest ones for

the case of linear optimization. This article begins in the section titled “Robust Linear Programming Setup” by introducing foundational concepts related to robust linear optimization. In the section titled “Solving Robust LPs”, we discuss robust linear programming optimization fundamentals. We conclude in the section titled “Extensions” with a brief discussion of some extensions to robust linear optimization. This article is meant to be self-contained; however, we refer readers to the foundational works of modern robust optimization theory, pioneered independently by Ben-Tal and Nemirovski [2,4,6,7] and by El-Ghaoui with Lebreton [8] (and with Oustry and Lebreton [9]). A thorough treatment of robust optimization is available in the book [10] and a tutorial treatment is provided in Ref. 11.

ROBUST LINEAR PROGRAMMING SETUP

We consider the case of a linear programming problem, where A is an $m \times n$ matrix:

$$\min \mathbf{c}'\mathbf{x} \quad (2a)$$

$$\text{s.t. } A\mathbf{x} \geq \mathbf{b} \quad (2b)$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}, \quad (2c)$$

where $\mathbf{c}$ and $\mathbf{b}$ have conforming dimensions, and where the problem data values $(\mathbf{c}, A, \mathbf{b})$ are unknown. In fact, we can restrict ourselves without loss of generality to problems in which only the constraint coefficient data A is uncertain, by rewriting the problem above as

$$\min t \quad (3a)$$

$$\text{s.t. } t - \mathbf{c}'\mathbf{x} \geq 0, \quad (3b)$$

$$A\mathbf{x} - \mathbf{b}w \geq 0, \quad (3c)$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}, w = 1. \quad (3d)$$

Note in Equation (3) that the uncertainty appears only in the constraint coefficient data.

Now, suppose that A belongs to the uncertainty set $\mathcal{U}$, and that all other data is treated as certain. The *robust counterpart* of Equation (2) replaces Equation (2b) with

$$\bar{A}\mathbf{x} \geq \mathbf{b} \quad \forall \bar{A} \in \mathcal{U}.$$

As a result, we have a *semi-infinite program*; that is, a mathematical program with a finite number of decision variables but with an infinite number of constraints.

As an illustration, consider the problem

$$\min x_1 + x_2, \quad (4a)$$

$$\text{s.t. } (4 - \Delta_{11})x_1 + (1 - \Delta_{12})x_2 \geq 4, \quad (4b)$$

$$(1 - \Delta_{21})x_1 + 5x_2 \geq 10, \quad (4c)$$

$$-x_1 + x_2 \geq 5, \quad (4d)$$

$$x_1, x_2 \geq 0, \quad (4e)$$

and suppose that $\mathcal{U}$ is the following polytope:

$$\mathcal{U} = \{\Delta_{11}, \Delta_{12}, \Delta_{21} \geq 0 : \Delta_{11} + \Delta_{12} + \Delta_{21} \leq 1\}.$$

When all Δ -values equal zero, we have a nominal feasible region as depicted in Fig. 1. Looking at convex combinations of the points $(\Delta_{11}, \Delta_{12}, \Delta_{21}) = (1, 0, 0)$ and $(0, 1, 0)$, we generate the infinite collection of constraints corresponding to these values as depicted in Fig. 2. Note that Equation (4b) holds for each $\Delta \in \mathcal{U}$ if and only if

$$4x_1 + x_2 + \min_{\Delta \in \mathcal{U}} (-\Delta_{11}x_1 - \Delta_{12}x_2) \geq 4.$$

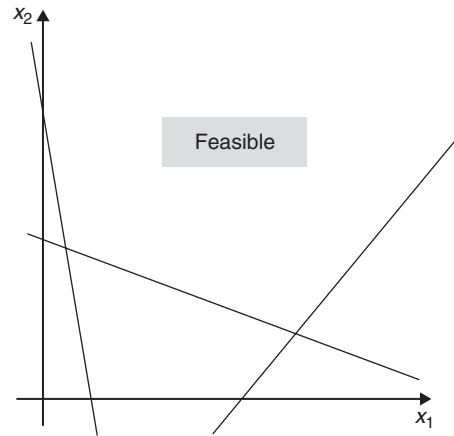


Figure 1. Feasible region for problem (4) in which $\Delta_{11} = \Delta_{12} = \Delta_{21} = 0$.

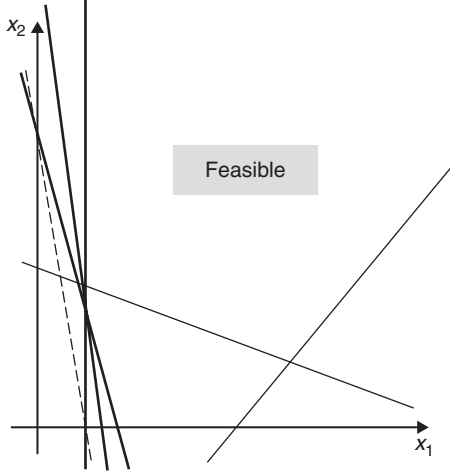


Figure 2. Generation of robust constraints with respect to Equation (4b). The dotted constraint corresponds to the case in which $\Delta_{11} = \Delta_{12} = 0$.

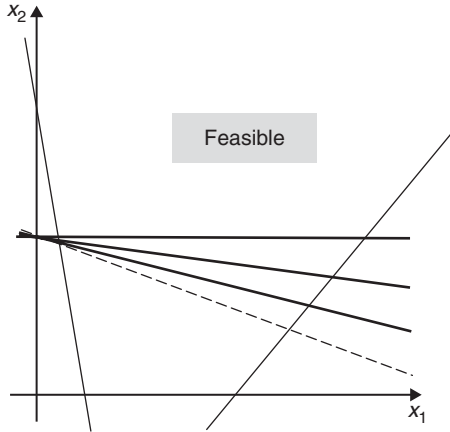


Figure 3. Generation of robust constraints with respect to Equation (4c). The dotted constraint corresponds to the case in which $\Delta_{21} = 0$.

Because $\min_{\Delta \in \mathcal{U}} (-\Delta_{11}x_1 - \Delta_{12}x_2) = \min\{-x_1, -x_2\}$, the robust version of Equation (4b) can be written as the system of two constraints $3x_1 + x_2 \geq 4$ and $4x_1 \geq 4$. Similarly, note that as we consider convex combinations of points between $(\Delta_{11}, \Delta_{12}, \Delta_{21}) = (0, 0, 0)$ and $(0, 0, 1)$, an infinite number of constraints are generated as shown in Fig. 3, all of which are dominated by the constraint

$5x_2 \geq 10$. The robust counterpart in this case replaces Equation (4b) with two constraints $3x_1 + x_2 \geq 4$ and $4x_1 \geq 4$, and replaces Equation (4c) with $5x_2 \geq 10$.

The key technique to solving the semi-infinite robust counterpart is to exploit the structure of $\mathcal{U}$ and reformulate the problem as a finite-dimensional optimization problem (perhaps in higher dimension) so as to make it amenable to standard (deterministic) mathematical programming tools. Before proceeding to discuss these reformulations, we examine several properties that are vital to understanding these problems.

Column-Wise Uncertainty. Suppose that we let $\mathbf{a}_j$ denote the j th column of A , for $j = 1, \dots, n$. Now, suppose that uncertainty in A is column-wise, in the sense that uncertainty in $\mathbf{a}_i$ and $\mathbf{a}_j$ are independent, for each $1 \leq i < j \leq n$. Thus, we write $\mathbf{a}_j \in \mathcal{U}_j$, and $\mathcal{U} = \mathcal{U}_1 \times \dots \times \mathcal{U}_n$. This case is considered in one of the earliest studies on robust optimization by Soyster [12].

Consider any constraint i and, for the sake of simplicity, suppose that $\mathbf{x} \geq 0$. This constraint can be written as

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \geq b_i.$$

Note that because uncertainty is column-wise, variations in $a_{i1}, a_{i2}, \dots, a_{in}$ are independently determined [12]. We can therefore optimize *a priori* to determine $\underline{a}_{ij} = \inf\{\hat{a}_{ij} : \mathbf{a}_j \in \mathcal{U}_j, \hat{a}_{ij} = \hat{a}_{ij}\}$. This gives rise to a simple characterization of the problem's robust counterpart:

$$\underline{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq 0,$$

where $\underline{A}$ is the matrix of entries $\underline{a}_{ij}$.

Equivalence with Row-Wise Uncertainty. Now, let $\mathbf{a}^i$ denote the i th row vector of A , and define $\mathcal{U}_i = \{\mathbf{a}^i : \exists A \in \mathcal{U} \text{ with row } i \text{ of } A = \mathbf{a}^i\}$. One can think of this as the projection of $\mathcal{U}$ onto the set of all possible combinations of constraint i coefficients. Observe that in the robust counterpart, $\mathbf{x}$ must be feasible for *any* $A \in \mathcal{U}$. Therefore, $\mathbf{x}$ is feasible if and only if:

$$(\mathbf{a}^i)' \mathbf{x} \geq b_i \quad \forall \mathbf{a}^i \in \mathcal{U}_i \quad \forall i = 1, \dots, m.$$

The implication is therefore that we can assume row-wise uncertainty of the constraint matrix, in which the distribution of the constraint coefficients are row-wise independent. That is, optimizing over the uncertainty set in which $A \in \mathcal{U}$ is equivalent to optimizing over the uncertainty set in which A belongs to the (larger) set $\mathcal{U}_1 \times \dots \times \mathcal{U}_m$.

Indeed, one can see this from the prior example given in Figs 1–3, in which uncertainty was considered separately for constraints (4b) and (4c). Define $\mathcal{U}_1 = \{\Delta_{11}, \Delta_{12} \geq 0 : \Delta_{11} + \Delta_{12} \leq 1\}$ and $\mathcal{U}_2 = \{\Delta_{21} : 0 \leq \Delta_{21} \leq 1\}$. Then optimizing over $\bar{\mathcal{U}} = \mathcal{U}_1 \times \mathcal{U}_2$ is equivalent to optimizing over $\mathcal{U}$ (even though $\bar{\mathcal{U}} \supset \mathcal{U}$).

Convexity and Closedness of Uncertainty Set. First, note that if $\mathbf{x}$ is feasible with respect to $A^1 \in \mathcal{U}$ and $A^2 \in \mathcal{U}$, then it is also feasible with respect to constraint matrix $\lambda A^1 + (1 - \lambda)A^2$, for $0 \leq \lambda \leq 1$. Therefore, optimizing over uncertainty set $\mathcal{U}$ and over its convex hull are equivalent. Furthermore, optimizing over the closure of $\mathcal{U}$ is equivalent to optimizing over $\mathcal{U}$. Therefore, we can assume that, for robust linear programming problems, the uncertainty set is row-wise uncertain, closed, and convex [5].

SOLVING ROBUST LPs

As mentioned before, the key approach in solving semi-infinite robust optimization problems is to reformulate them as standard mathematical programs by exploiting the structure of the uncertainty set $\mathcal{U}$. In this section, we focus on problems of the form of Equation (2), in which the A -values are uncertain. The standard approach that we take here is to handle robustness by envisioning the problem as a Stackelberg (leader–follower) game: The leader chooses a candidate value $\mathbf{x}$, and the follower attempts to choose an A -matrix in the set $\mathcal{U}$ to make $\mathbf{x}$ infeasible. This type of game is often handled in mathematical programming by dualizing the follower's problem, which allows the leader/follower problem to be combined and solved as a single optimization problem. However, in this setting, the follower's

problem is often easily solvable as we will see below.

In the rest of this section, we consider four types of uncertainty sets, and briefly describe the reformulation of the corresponding robust counterparts into standard mathematical programs.

Interval Uncertainty

In this model, each element a_{ij} of the A matrix ($i = 1, \dots, m$, $j = 1, \dots, n$) is uncertain but belongs to some interval $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$ for all j . Given a value of $\mathbf{x}$, the worst-case scenario is that $a_{ij} = \bar{a}_{ij} - \hat{a}_{ij}$ if $x_j > 0$, and $a_{ij} = \bar{a}_{ij} + \hat{a}_{ij}$ otherwise. (In the context of the Stackelberg game above, the follower chooses a value of a_{ij} within the allowable interval that minimizes $a_{ij}x_j$, noting that the constraints are of the $\geq$ sense.) Letting $\bar{A}$ ($\hat{A}$) be the collection of $\bar{a}_{ij}$ -values ($\hat{a}_{ij}$ -values), we get the following robust counterpart formulation

$$\begin{aligned} \min \quad & \mathbf{c}'\mathbf{x}, \\ \text{s.t.} \quad & \bar{A}\mathbf{x} - \hat{A}|\mathbf{x}| \geq \mathbf{b}, \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \end{aligned}$$

Linearizing $|\mathbf{x}|$ can be done in a number of ways. Here, we define variables $y_j = |x_j|$ for $j = 1, \dots, n$, and enforce constraints $-\mathbf{y} \leq \mathbf{x} \leq \mathbf{y}$. (Since an optimal solution exists in which the y -variables take on their smallest possible values given x , these constraints are sufficient to force $y_j = |x_j|$ at optimality.) We then obtain the following linear program:

$$\min \quad \mathbf{c}'\mathbf{x}, \tag{5a}$$

$$\text{s.t.} \quad \bar{A}\mathbf{x} - \hat{A}\mathbf{y} \geq \mathbf{b}, \tag{5b}$$

$$-\mathbf{y} \leq \mathbf{x} \leq \mathbf{y}, \tag{5c}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \tag{5d}$$

Note here that this is a special case of column-wise uncertainty. If $\mathbf{l} \geq 0$, then no y -variables are required in the formulation: we can remove Equation (5c) and replace Equation (5b) with

$$(\bar{A} - \hat{A})\mathbf{x} \geq \mathbf{b}.$$

It is also important to observe that the uncertainty set $\mathcal{U}$ is polyhedral, and that the robust counterpart to the original problem remains a linear program. We will see that this result holds true in the next two sections as well, which involve more sophisticated types of polyhedral uncertainty.

Budgeted Uncertainty

The case of *budgeted uncertainty* is studied by Bertsimas and Sim [3], and requires slightly more analysis than the interval uncertainty case. We begin by describing the uncertainty set. Let $\Gamma_i \leq n$ be a given integer for all i . Each coefficient a_{ij} of the i -th constraint belongs to an interval $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$, and no more than Γ_i of the coefficients $a_{i1}, \dots, a_{in}$ are allowed to vary simultaneously. Bertsimas and Sim [3] show that such an uncertainty set can be written as

$$a_{ij} = \bar{a}_{ij} + z_{ij}\hat{a}_{ij}, \quad (6)$$

with

$$0 \leq |z| \leq 1 \text{ and } \sum_j |z_{ij}| \leq \Gamma_i \quad \forall i = 1, \dots, m. \quad (7)$$

The “follower” in the Stackelberg analogy chooses, for each row, a set of coefficients that may deviate from their nominal values of $\bar{a}_{ij}$. In the interval uncertainty case, the follower chooses all n coefficients for each constraint, where by contrast, only a “budget” of Γ_i such coefficients can be changed in the budgeted uncertainty case.

The underlying rationale behind this budgeting (and indeed, behind most commonly used uncertainty sets) is that it is unlikely that all data will simultaneously take on their worst possible values (i.e., those values that most tightly constrain the feasible region), even if the “true” uncertainty set is the interval uncertainty set in the section titled “Interval Uncertainty.” Thus, if one considers the case in which only Γ_i out of n values for constraint i take on their worst possible values (assuming that Γ_j and z are integer-valued), then one introduces a slight risk of obtaining an infeasible solution when

the actual data is revealed (when $\Gamma_i < n$), but potentially improves the objective function value of the solution as the problem becomes less-constrained. Indeed, smaller values of Γ_i will cause the optimal objective function value to stay the same or decrease, but will increase the chances that the resulting solution will be infeasible.

In this case, given $\mathbf{x}$, we must account for the fact that the left-hand side is as small as

$$\sum_{j=1}^n \bar{a}_{ij}x_j - \sum_{j=1}^n \hat{a}_{ij}|x_j|z_{ij} \quad \forall i = 1, \dots, m. \quad (8)$$

Hence, the follower in a Stackelberg game—trying to force $\mathbf{x}$ to be infeasible—would solve the following optimization problem.

$$\max \sum_{j=1}^n \hat{a}_{ij}|x_j|z_{ij}, \quad (9a)$$

$$\text{s.t. } \sum_{j=1}^n z_{ij} \leq \Gamma_i, \quad (9b)$$

$$0 \leq z_{ij} \leq 1 \quad \forall j = 1, \dots, n. \quad (9c)$$

Associating dual variable π_i with Equation (9b) and μ_{ij} , $j = 1, \dots, n$, with the upper bounding constraints in Equation (9c), we obtain the following (linear programming) dual formulation of Equation (9):

$$\min \Gamma_i \pi_i + \sum_{j=1}^n \mu_{ij}; \quad (10a)$$

$$\text{s.t. } \pi_i + \mu_{ij} \geq \hat{a}_{ij}|x_j| \quad \forall j = 1, \dots, n; \quad (10b)$$

$$\pi_i \geq 0, \mu_{ij} \geq 0 \quad \forall j = 1, \dots, n. \quad (10c)$$

Therefore, we can replace the term $\sum_{j=1}^n \hat{a}_{ij}|x_j|z_{ij}$ in Equation (8) with the objective function term $\Gamma_i \pi_i + \sum_{j=1}^n \mu_{ij}$ in (10a), constrained by Equation (10b) and Equation (10c). Furthermore, replacing $y_j = |x_j|$ using the same substitution strategy as in the section titled “Interval Uncertainty,” we obtain the following robust counterpart.

$$\begin{aligned}
& \min \quad \mathbf{c}'\mathbf{x}; \\
& \text{s.t.} \quad \bar{\mathbf{a}}^i \mathbf{x} - \left(\Gamma_i \pi_i + \sum_{j=1}^n \mu_{ij} \right) \geq b_i \\
& \quad \quad \quad \forall i = 1, \dots, m; \\
& \quad \quad \quad \pi_i + \mu_{ij} \geq \hat{a}_{ij} y_j \quad \forall j = 1, \dots, n; \\
& \quad \quad \quad -y_j \leq x_j \leq y_j \quad \forall j = 1, \dots, n; \\
& \quad \quad \quad \pi, \mu \geq 0; \\
& \quad \quad \quad \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}.
\end{aligned}$$

Observe that the robust counterpart simultaneously computes both an optimal robust solution $\mathbf{x}$, and a worst-case set of parameter values while preserving the linear structure of the problem.

A vital question addressed by Bertsimas and Sim [3] regards the probability that constraint i is violated, given budgeted uncertainty Γ_i . Assume that in reality, $a_{ij} \in [\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$ for each i and j . Furthermore, assume that the distribution of a_{ij} is symmetric over this interval. Then Bertsimas and Sim [3] show that the probability of violating constraint i is approximately no more than

$$1 - \Phi\left(\frac{\Gamma_i - 1}{\sqrt{n}}\right),$$

where Φ is the cumulative distribution function of a standard normal random variable.

Polyhedral Uncertainty

The previous sections beg the question of whether or not one can still obtain a robust counterpart to the original linear program if the uncertainty set is a general polyhedron. We answer this question in the affirmative in this section. Our uncertainty set in this case is Equation (6) with the z -variables constrained as follows for each $i = 1, \dots, m$:

$$F_i |\mathbf{z}_i| \leq \mathbf{g}_i \quad \text{and} \quad |z_{ij}| \leq 1 \quad \forall j = 1, \dots, n, \quad (11)$$

where $F_i \in \mathbb{R}^{k_i \times n}$ and $\mathbf{g}_i \in \mathbb{R}^{k_i \times 1}$. In this case, we wish to optimize Equation (9a) subject to constraints (11). This problem can be

converted to an equivalent linear program by shifting the absolute values from the z -variables to the x -variables as follows:

$$\max \quad \sum_{j=1}^n \hat{a}_{ij} |x_j| z_{ij}, \quad (12a)$$

$$\text{s.t.} \quad F_i \mathbf{z}_i \leq \mathbf{g}_i, \quad (12b)$$

$$z_{ij} \leq 1 \quad \forall j = 1, \dots, n. \quad (12c)$$

The dual of Equation (12) is given by the following:

$$\begin{aligned}
& \min \quad \mathbf{g}'_i \pi_i + \sum_{j=1}^n \mu_{ij}, \\
& \text{s.t.} \quad F'_i \pi_i + \mu_i \geq \text{diag}(\hat{\mathbf{a}}^i) |\mathbf{x}|, \\
& \quad \quad \pi_i, \mu_i \geq 0.
\end{aligned}$$

Substituting $\sum_{j=1}^n \hat{a}_{ij} |x_j| z_{ij}$ with optimization problem (12) and linearizing the absolute value terms, we get the following (linear programming) robust counterpart.

$$\begin{aligned}
& \min \quad \mathbf{c}'\mathbf{x}; \\
& \text{s.t.} \quad \bar{\mathbf{a}}^i \mathbf{x} - \left(\mathbf{g}'_i \pi_i + \sum_{j=1}^n \mu_{ij} \right) \geq b_i \\
& \quad \quad \quad \forall i = 1, \dots, m; \\
& \quad \quad \quad F'_i \pi_i + \mu_i \geq \text{diag}(\hat{\mathbf{a}}^i) \mathbf{y} \quad \forall i = 1, \dots, m; \\
& \quad \quad \quad -y_j \leq x_j \leq y_j \quad \forall j = 1, \dots, n; \\
& \quad \quad \quad \pi, \mu \geq 0; \\
& \quad \quad \quad \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}.
\end{aligned}$$

Ellipsoidal Uncertainty

The robust optimization approach taken by Ben-Tal and Nemirovski [4] examine ellipsoidal uncertainty sets $\mathcal{U}_i$, for $i = 1, \dots, m$. For simplicity (and without loss of generality after scaling), we can consider hypersphere sets (that no longer contain “box constraints” of the form $|z_{ij}| \leq 1$):

$$\mathcal{U}_i = \{\mathbf{z}_i : \|\mathbf{z}_i\|_2 \leq \Omega_i\}. \quad (13)$$

For each constraint $i = 1, \dots, m$, we guarantee that $\mathbf{x}$ is feasible over the uncertainty set by maximizing $\sum_{j=1}^n \hat{a}_{ij}x_jz_{ij}$ over the z -variables, subject to Equation (13). Thus, one would maximize in the direction $(\hat{\mathbf{a}}_i)'\mathbf{x}$ over Equation (13), and the optimal objective function value is given by

$$\Omega_i \sqrt{\sum_{j=1}^n (\hat{a}_{ij}x_j)^2}. \quad (14)$$

Substituting $(\hat{\mathbf{a}}_i)'\mathbf{x}$ with (14), we obtain the following (nonlinear) second-order cone program:

$$\min \quad \mathbf{c}'\mathbf{x}; \quad (15a)$$

$$\begin{aligned} \text{s.t.} \quad & \bar{\mathbf{a}}^i\mathbf{x} - \Omega_i \sqrt{\sum_{j=1}^n (\hat{a}_{ij}x_j)^2} \geq b_i \\ & \forall i = 1, \dots, m; \end{aligned} \quad (15b)$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}. \quad (15c)$$

Observe that unlike in the case of robust linear programming with budgeted uncertainty, model (15) does not require any auxiliary variables. Ben-Tal and Nemirovski [4] show that the probability of violating constraint i , using the ellipsoidal uncertainty set model, assuming that a_{ij} is symmetrically distributed over the interval $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$, is $\exp(-\Omega_i^2/2)$.

Note that the robust counterpart (Eq. 15) is nonlinear due to the fact that the uncertainty sets are ellipsoidal rather than polyhedral. Furthermore, Ben-Tal and Nemirovski [13] demonstrate a similar behavior in conic uncertainty sets (which generalize all of the uncertainty sets described thus far), and show how more complex uncertainty sets accordingly induce more complex robust counterpart formulations.

EXTENSIONS

To conclude this article, we briefly discuss ongoing research areas in the field of robust optimization.

In the section titled “Solving Robust LPs,” we see that linear programming problems remain linear under their robust counterpart if the uncertainty sets are polyhedral. One may therefore postulate that polynomially solvable mixed-integer programming problems, such as shortest path or assignment problems, remain polynomially solvable given polyhedral uncertainty sets. However, Kouvelis and Yu [14] show that this conjecture is not true. Even under *scenario uncertainty* (in which there are a finite number of possible realizations for stochastic data) with just two scenarios, and uncertainty restricted to the objective function coefficients, the robust shortest path problem can be NP-hard. (Consider two possible sets of cost coefficients, with an objective to minimize the maximum of the two possible objective function values. This problem is isomorphic to the resource-constrained shortest path problem, which is NP-hard.)

Therefore, there is a need to study robust discrete optimization problems using a different set of analytical tools. One approach, taken by Bertsimas and Sim [15], is to identify certain polynomially solvable robust discrete optimization problems. In particular, they provide a polynomial-time algorithm for solving combinatorial optimization problems when uncertainty is limited to the objective function coefficients, uncertainty sets are modeled using the uncertainty budget as described in the section titled “Budgeted Uncertainty,” and the nominal (nonrobust) optimization problem itself is polynomially solvable. Another approach seeks stronger problem formulations and classes of valid inequalities that can be used to improve the solvability of robust counterpart formulation. One such notable approach for NP-hard robust counterparts is given by Atamtürk [16].

Next, the concept of *adjustability* or *recourse* is foundational in stochastic programming, but is an emerging area of robust optimization. After a set of decisions is made before values of the uncertain data is realized, a set of recourse decisions allow a second, reactionary set of decisions to be made after the realization of uncertain

data. Atamtürk and Zhang [17] examine robust network flow problems with recourse. Their paper involves a situation in which initial flows are sent across the network before demands are realized and then, the remaining flows are transmitted after the realization of demands. They extend their results to problems involving discrete decisions on capacity allocation (and thus network design) decisions. Bertsimas and Thiele [18] examine robust multistage inventory optimization problems (see also their tutorial article [11]).

Finally, research in this field is still young, and many more advanced issues pertaining to robust optimization have recently been explored [19–23]. We recommend an on-line resource at <http://www.robustopt.com>, which contains a current bibliography of robust optimization literature, along with Matlab software tools designed to automatically formulate robust counterparts such as those mentioned in this article.

REFERENCES

1. Dantzig G. Linear programming under uncertainty. *Manage Sci* 1955;1(3–4): 197–206.
2. Ben-Tal A, Nemirovski A. Robust solutions to uncertain linear programs. *Oper Res Lett* 1999;25:1–13.
3. Bertsimas D, Sim M. The price of robustness. *Oper Res* 2004;52(1):35–53.
4. Ben-Tal A, Nemirovski A. Robust solutions of linear programming problems contaminated with uncertain data. *Math Program* 2000;88(3):411–424.
5. Nemirovski A. Lectures on robust convex optimization. Available at <http://www2.isye.gatech.edu/~nemirovs/>. 2010.
6. Ben-Tal A, Nemirovski A. Stable truss topology design via semidefinite programming. *SIAM J Optim* 1997;7:991–1016.
7. Ben-Tal A, Nemirovski A. Robust convex optimization. *Math Oper Res* 1998;23: 769–805.
8. El-Ghaoui L, Lebret H. Robust solutions to least-square problems with uncertain data matrices. *SIAM J Matrix Anal Appl* 1997;18: 1035–1064.
9. El-Ghaoui L, Oustry F, Lebret H. Robust solutions to uncertain semidefinite programs. *SIAM J Optim* 1998;9:33–52.
10. Ben-Tal A, El-Ghaoui L, Nemirovski A. *Robust Optimization*. Princeton (NJ): Princeton University Press; 2009.
11. Bertsimas D, Thiele A. Robust and data-driven optimization: modern decision-making under uncertainty. In: Johnson MP, Norman B, Secomandi N, editors. *Tutorials in operations research: models, methods, and applications for innovative decision making*. INFORMS; Volume 2; 2006. pp. 95–122.
12. Soyster AL. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper Res* 1973;21: 1154–1157.
13. Ben-Tal A, Nemirovski A. Robust optimization –methodology and applications. *Math Program Ser B* 2002;92:453–480.
14. Kouvelis P, Yu G. *Robust discrete optimization and its applications*. Norwell (MA): Kluwer Academic Publishers; 1997.
15. Bertsimas D, Sim M. Robust discrete optimization and network flows. *Math Program* 2003;98:49–71.
16. Atamtürk A. Strong formulations of robust mixed 0–1 programming. *Math Program Ser B* 2006;108:235–250.
17. Atamtürk A, Zhang M. Two-stage robust network flow and design under demand uncertainty. *Oper Res* 2007;55(4):662–673.
18. Bertsimas D, Thiele A. A robust optimization approach to inventory theory. *Oper Res* 2006;54(1):150–168.
19. Ben-Tal A, Boyd S, Nemirovski A. Extending the scope of robust optimization: comprehensive robust counterparts of uncertain problems. *Math Program* 2006;107(1–2): 63–89.
20. Bertsimas D, Brown DB. Constructing uncertainty sets for robust linear optimization. *Oper Res* 2009;57(6):1483–1495.
21. Bertsimas D, Pachamanova D, Sim M. Robust linear optimization under general norms. *Oper Res Lett* 2004;32:510–516.
22. Bertsimas D, Sim M. Tractable approximations to robust conic optimization problems. *Math Program* 2006;107(1–2):5–36.
23. Calafiore G, Campi MC. Uncertain convex programs: randomized solutions and confidence levels. *Math Program* 2005;102(1): 25–46.

INTRODUCTION TO SHOP-FLOOR CONTROL

DAMIEN TRENTESAUX

University of Lille Nord de
France, Lille, France

UVHC, TEMPO Research Center,
Valenciennes, France

VITTALDAS V. PRABHU

Penn State University, University
Park, Pennsylvania

INTRODUCTION

A shop floor can be defined as all the production and human resources needed to produce semifinished or finished physical goods. A shop floor is usually organized according to the complexity of the product flow and the flexibility of the resources. For instance, for high variety manufacturing, the shop floor may be organized in job shops when product flows are arbitrary and resources are flexible. In contrast, for high volume manufacturing, a shop floor may be organized as a flow shop or an assembly transfer line. The main objective of a shop-floor control (SFC) system is to keep production execution as close as possible to its plan by taking appropriate corrective actions. In order to meet this objective, SFC system need to measure current execution and compare this feedback with the plan to detect any deviations from the plan as illustrated in Fig. 1.

Some of the main attributes of SFC include

- short-time horizons (e.g., minutes) in comparison with planning (e.g., days or weeks);
- the need to compensate for disturbances and modeling errors in the planning system; and
- the need to interface with a myriad of physical equipments and devices to get feedback.

The combination of the above-mentioned attributes makes optimization of SFC a major challenge in terms of architecting SFC systems.

In this article, the classical centralized approaches and modern distributed approaches to SFC are presented. This is followed by an example of modern SFC to illustrate the manner in SFC systems increasingly encompass manufacturing execution system (MES) and supervisory control and data acquisition (SCADA) levels of enterprise architectures. Some recent developments in feedback control approaches for SFC are also presented along with architectural, functional, technological, and human-centric considerations.

THE CLASSICAL VIEW OF A SHOP-FLOOR CONTROL SYSTEM

The classical view of an SFC consists of an SCADA system of which the main functions are production supervision, data acquisition, and estimation of key performance indicators (KPI) and their transmission to upper decision levels. Such SCADA systems are useful to production managers for tracking customer orders and work in progress (WIP). SCADA systems are integrated into an MES to support other functions such as quality control (SPC, statistical process control) and preventive maintenance. Typically, such integration entails data transfer between an SCADA central host computer with a variety of devices including remote terminal units (RTUs), programmable logic controllers (PLCs), and operator terminals [1]. Therefore, in the classical view of SFC, its inputs are at the level of production planning (e.g., from Materials Resource Planning (MRP) level) and outputs are supervision information regarding WIP, inventory level, machine status, and performance. The resulting architecture of such classical SCADA systems is hierarchical and compliant with the federative Computer Integrated Manufacturing (CIM) concept, being monolithic (central-

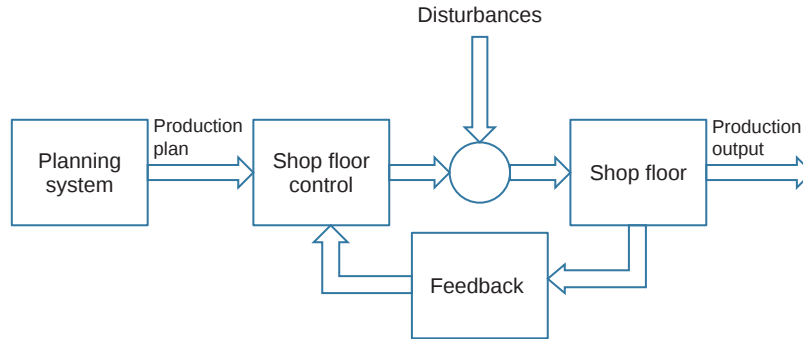


Figure 1. Functioning of shop-floor control.

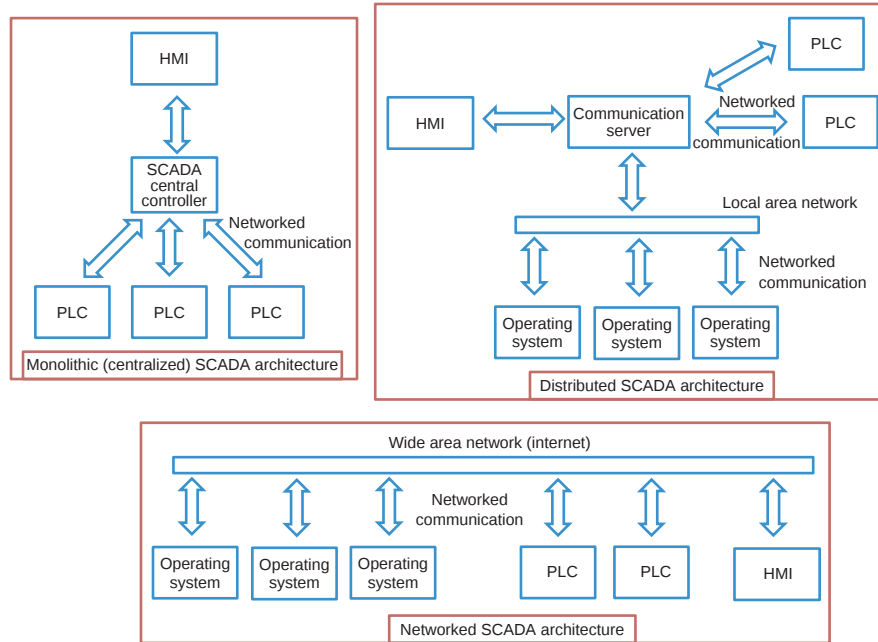


Figure 2. Typical SCADA architectures. *Source:* Adapted from NCS [1].

ized), distributed (using a local area network, LAN), or, more recently, networked (using a wide area network, WAN), as shown in Fig. 2. The latter introduces possible heterarchical relationships among operating systems that will be presented latter in this article.

This classical view of SFC, reducing SFC functions to supervision, low level control, and data acquisition are still considered important in process industries where the continuous and homogeneous aspects of

resources, products, and data enable global and visual supervisory human-machine interface (HMI) to be designed for decision-makers. An important aspect of SCADA systems in such industries concerns the safety of controlled processes. Modern SCADA systems are also concerned with network security, especially when they are part of a critical infrastructure such as power generation. A typical classical view of SFC using SCADA can be found in Dieu [2].

Decision-making in classical SFC using SCADA is predominantly manual (e.g., to prevent from human injuries) and automatic control is limited to physical process control (e.g., Proportional Integral Derivative (PID) control through a PLC). When applied to more complex, heterogeneous production environments, such as discrete product manufacturing systems with random disturbances, the lack of automated decision-making in classical SCADA at higher levels of production is a major short coming. This has led to the need to integrate decision-making capabilities in the control loop shown in Fig. 1. For example, Wysk and Smith [3] proposed a formal functional vision of SFC, addressing the scheduling issue using graph theory, modeling SFC as a kind of discrete event systems. Such functions are not traditionally handled in SCADA systems for the process industry, thus motivating the modern view of SFC systems, which is described in the following section.

THE MODERN VIEW OF SHOP-FLOOR CONTROL SYSTEMS

Nowadays, owing to the increase in the complexity of controlled systems, the lack of midterm range production visibility and the increase in the number and potential consequences of disruptions, it has become very desirable for SFC to have decision-making ability in order to adapt to disturbances [4]. Therefore, SFC is a key interface between systems that plan and schedule production and systems that work in real time. This evolution has been partly influenced by the Japanese way of thinking, with the just-in-time philosophy (pull systems). This philosophy implies a closed loop between production and customers; this is in contrast with the classical MRP way of thinking, which pushes products through the resources on the shop floor. In theory, there is no real need to adapt production in a system like this as all is planned and decisions are made at MRP level; thus, the SFC is only required to supervise production.

This evolution has also been made possible by the technological advances in computing and communication that allow rapid

acquisition of shop-floor status and real-time decision-making. Real time in this context means that the computational results are obtained much faster than the rate of change of shop-floor status; the decisions can thus be executed with high fidelity. Therefore, the term *control* in SFC has now evolved to mean that a higher level feedback loop operates between planning/scheduling level and real-time level, updating information, and adapting planned/scheduled production to the real situation on the shop floor. These SFC systems consider constraints that are not usually considered by planning or scheduling production systems, such as the real states of production resources, capacities, tool management, conveyor system, and inventory levels. Therefore, there is an emerging need for corresponding real-time control algorithms that maintain near-optimal performance at all times.

Consequently, modern SFC approaches integrate SCADA systems as a lower level component coupled to higher level decision-making abilities ensuring a more effective and efficient control of production resources. Typically, such high level decision-making abilities in modern SFC concern scheduling decisions, workload management, and capacity planning and tend to overlap with MES and enterprise resource planning (ERP) levels [5]. This means that traditional low level monitoring (SCADA) had to be adapted to consider this evolution. For example, if Kanban systems are used to control production resources, then its SFC system must ensure that appropriate policies are followed (e.g., no more parts than the number of Kanbans).

Figure 3 illustrates the classical hierarchical breakdown of the time horizon in production planning and control. It is inspired from the ISA-95 international standard for the integration of enterprise and control systems and the IEC 62264 international standard for enterprise-control system integration (based on the ISA-95 standard) that introduces a manufacturing operations management model [6]. In this figure, the way SFC has evolved to encompass more high level decision control loops is highlighted, leading to an approximate functional

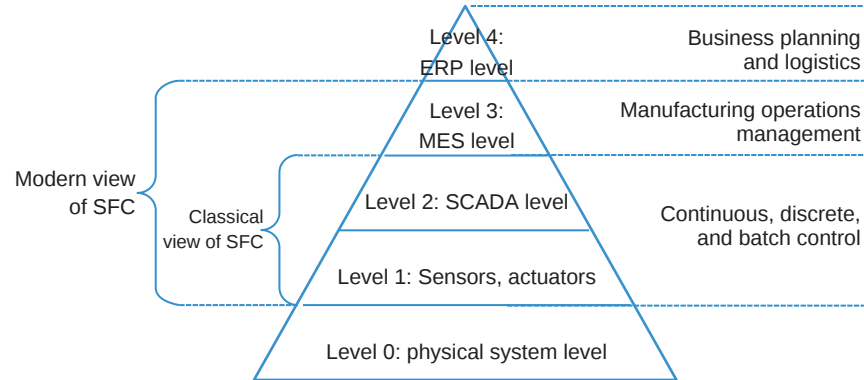


Figure 3. Evolution of SFC. *Source:* Adapted from ISA-95 and IEC-62264.

definition of modern SFC as follows:

$$\text{Modern SFC} = \text{MES} + \text{SCADA}.$$

An important aspect of the modern SFC also concerns the control of resource efficiency. Efficiency is related to the way resources are used to attain production objectives. Efficiency considers, for example, capacity planning; maintenance KPI such as mean time to failure (MTTF), mean time to repair (MTTR), and mean time between failures (MTBF) [7, 8]; and lean/sustainability constraints such as inventory level control and power consumption control. This evolution toward “lean SFC” has been largely motivated by the Toyota way of thinking (Toyota production system, TPS), [9] which gives SFC an important role of maintaining customer satisfaction and internal efficiency. Another illustration concerns the natural integration of kaizen mechanisms within the TPS, aimed to improve KPI through minor but continuous improvement at SFC level. Kaizen decisions are clearly made at SFC level; the local quick response quality control (QRQC) decision mechanism highlights the following aspect: this mechanism basically concerns quality but QRQC decisions impact SFC as operators may modify their own production activity with regard to quality issues. Another feature of modern SFC concerns bottleneck resource management, as proposed by the theory of constraints (TOCs) [10, 11]. According to this approach, the overall effectiveness of a system is essentially guided by

its internal constraints and SFC should pay attention to these bottlenecks, controlling them accurately, ensuring that their input stock is never empty, while approximately controlling the others that are guided by the bottleneck (Drum, Buffer, and Rope). Specific scheduling rules should apply in this context and SFC must handle this aspect. Mixing the just-in-time (JIT) philosophy with the TOC and push methods is also considered in modern SFC. For example, the ConWIP method [11, 12] or demand flow technology (DFT) [13] is bridging the gap between these methods and influences the way modern industrial SFC are designed and used. Meanwhile, these industrial SFC tools are often devoted to specific shop-floor systems, such as line assembly systems within a mass customization context. The general assumption is that demand does not vary too much.

In other types of shop floors with more complex production flows (e.g., flexible manufacturing systems), or where demand varies considerably (highly customized production, niche production, etc.), the number of efficient/effective industrial tools is low and the risk is to reduce SFC to “simple” heuristic decision rules at MES level coupled with an SCADA system [7], thus implying a long-term low visibility. For this kind of SFC, research has focused on compensating this limitation because it is impractical to design standardized solutions. Formal process planning schemas for SFC have been proposed by identifying information requirements that

are unambiguously expressed as a function of the shop-floor activities required to manufacture a product [14]. This enables any generic SFC system to change its actions to flexibly manufacture different parts or the same part using different resources by appropriately modifying the information in its databases and thereby allowing real-time decision-making in the SFC to determine routing and process plans.

Typically, simulation is used to forecast the behavior of SFC, see, for example, [15], because the increasing integration of various functionalities with different time windows, decision variables, and parameters increase the decisional complexity of SFC [5]. The responsiveness of SFC can sometimes be a disadvantage because of the “nervousness” it can cause, which is a well-established issue in MRP systems. The link between SFC and planning layers was recently investigated for developing automated exception analytics systems to monitor and prioritize shop-floor exceptions systematically based on a postoptimality analysis of plans being executed [16].

Other research approaches based on simulation range from capacity planning in ERP to real-time feedback control to compensate for random perturbations [17]. Treating SFC as a continuous variable control problem in which timing of discrete events is controlled has led to a body of work in which controllers can be highly distributed for real-time scheduling [18–20]. This continuous variable SFC approach lends itself to real-time decision-making for routing and process parameters [20, 21]. The resulting dynamics can be analyzed using control theoretic techniques and have been proved to be globally stable for general shop-floor configurations including job-shop-like configurations with dissimilar machines with alternate routings without any restrictive assumptions [22].

AN ILLUSTRATIVE EXAMPLE OF A MODERN SFC SYSTEM

The AIP-Primeca FMS cell is a shop floor located at the University of Valenciennes and Hainaut-Cambr sis (France). This cell was built with the following industrial components:

- A monorail conveyor system with 15 single-direction shuttles [RFID (radio-frequency identification) tagged] and 11 transfer block (TBs) permitting flexible routing inside the cell.
- Seven robotized workstations (WSs) located at specific points on the monorail, which perform automated operations (loading/unloading, assembly, and quality inspection) on the product being manufactured. Loading/unloading consists in loading/unloading a plate on a shuttle, the plate being the support on which the products are made.

Thus, each product is placed on a dedicated shuttle and the set moves from WS to WS. Figure 4 provides an overview of the WSs, their locations, and the conveyor system.

The corresponding SFC architecture, shown in Fig. 5, is based on the “distributed SCADA architecture” shown in Fig. 2 and is augmented with functionalities at MES level. It is composed of the following four levels:

- Levels 0 and 1 are built with mechanical operating systems (St ubli and Kuka robots; Cognex vision inspection; Montratec monorails, transfer gates, and positioning units), PLCs (750-841 Wago controllers programmed with Codesys), stop and go shuttle devices, RFID R/W (Schneider OsiSense XG), and Ethernet switched automation network.
- Level 2 consists in an OPC server (Schneider OFS), several supervisory stations (equipped with Arcinfo PcVue32), a log server (SQL database server), and an Ethernet area operation information network.
- Level 3 is essentially composed of an MES system including scheduling ability optimization (IBM ILOG CPLEX Optimization Studio).

Typically, the problems to be solved by the SFC system in this cell include:

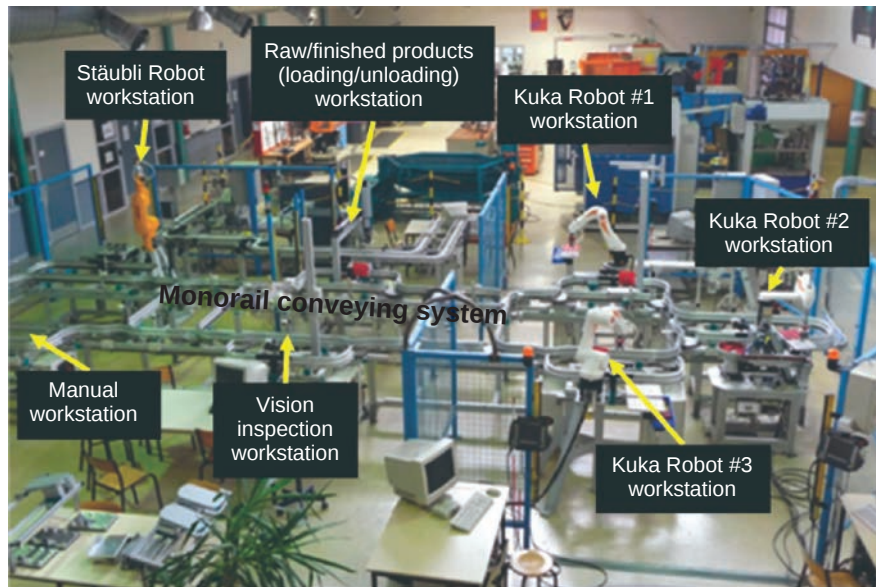


Figure 4. Overview of the AIP cell.

- the scheduling of a limited amount of resources (WSs, shuttles, inventory, and tools) within the time given to make a dynamic set of products;
- the dynamic routing of shuttles containing the WIP;
- the real-time control of WS and transfer gates.

Using this SFC architecture:

- Levels 0 and 1 manage the shuttle routing and WS states in real time.
- Level 2 supervises and monitors the WIP and interfaces levels 1 and 3.
- Level 3 defines the operations to be carried out by WS at levels 0 and 1.

More details about the SFC used in this research can be found in Berger *et al.* [23].

CURRENT RESEARCH IN SHOP-FLOOR CONTROL SYSTEMS

The modern view of an SFC can be seen as a first step toward increased local decision-making capability where events occur and

reactivity is needed. There are several major challenges from a scientific and practical perspective. Some are provided in this section and organized according to architectural, functional, and technological considerations.

Architectural Considerations

Classical hierarchical architectures are rapidly evolving toward more distributed/heterarchical SFC architectures, typically by extending the SCADA architecture. The basic idea is to integrate redundancy in the control decision abilities and horizontal cooperation mechanisms in the control decision process among decisional active or intelligent entities, that is, resources, products, inventories, tools, and so on. The aim is to limit the risks of local controller loss (robustness to failure and ease of restarting) and to enable more “plug and control” approaches making the whole SCADA system more capable of self-reconfiguring as well as self-adapting to changes in the environment and in the production system [24]. This also implies focusing on work at MES level to fine-tune this integration [25].

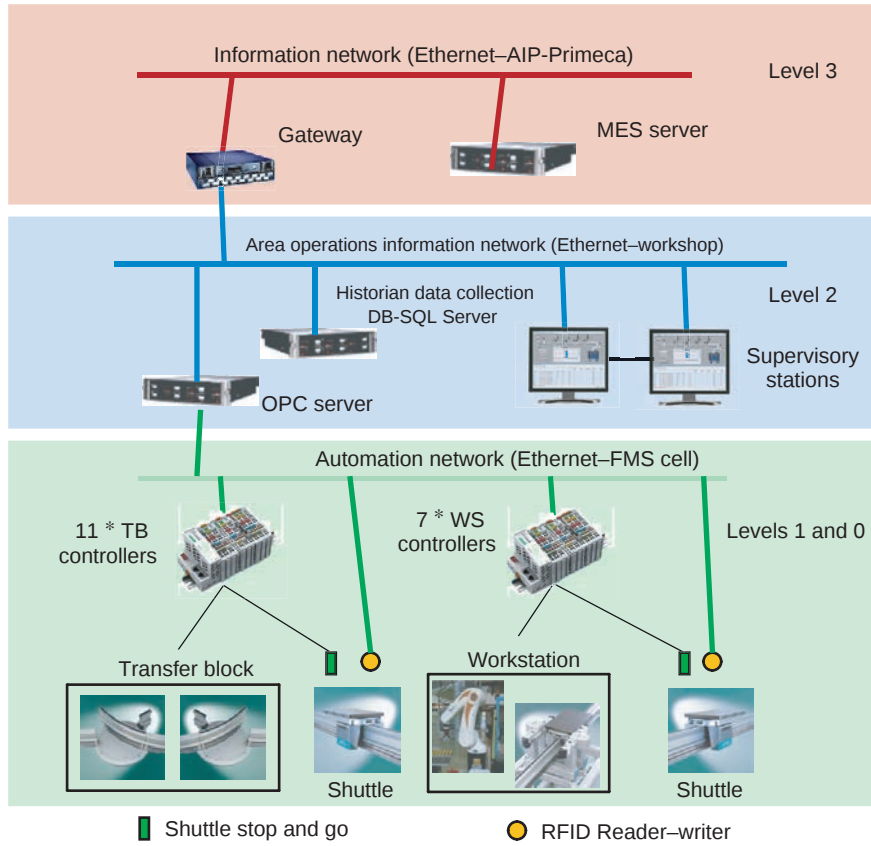


Figure 5. Modern SFC architecture of the AIP cell.

Functional Considerations

The idea is to integrate more functional abilities in SFC to encompass broader time horizons and increased data space for better decisions. This could lead to the integrated cohesive optimization of maintenance, production, and inventory, which are otherwise treated in isolation. The study of robustness is also important nowadays, typically in reactive production scheduling [26]. This issue is increasingly addressed from an operation research point of view. The coupling of optimization and simulation tools in the global context of digital enterprise also seems promising [27, 28].

Integrating intelligence and activeness into entities composing an SFC system (products, resources, inventories, tools, etc.) seems to be another promising research area [29].

From an information management point of view, it would help improve tracking the history of events and use of resources [30]. From an energy management point of view, it would help integrate opportunistic energy gains and improve management of energy costs given the increasing lack of visibility in terms of energy supply (variability of renewable energy power and variability in grid energy costs) [31].

Technological Considerations

Service orientation technology and service-oriented architectures (SOA) is key in the evolution of SFC [32]. Indeed, it enables a “components off the shelves” approach to be used that is required when designing plug and control SFC enabling “plug and produce” production systems [33].

As an illustrative example, the EU FP7 project IMC-AESOP (www.imc-aesop.eu) aims at defining the next generation of SFC systems based on SOA and recent internet technology such as cloud computing and the internet of things (Internet of Things (IoT) [34].

The IoT and ambient intelligence are also technological solutions enabling the improvement of emerging SFC systems and architectures. This evolution will contribute to design more self-adapting, self-organized SFC [35].

Toward More Human-Centered SFC

Human operators know information the SFC system does not know. Typically, such information comes from outside the SFC system itself, and by interaction with the SFC system, operators could help constrain or bound automated decisions according to this knowledge, thus anticipating events. For example, an important incoming order not yet integrated in the production planning, an operator that suspects a production machine will operate defectively (e.g., furtive error), or a possible incoming supplier strike can be considered through such interaction. In addition, despite all the theoretical developments in learning systems, the human operator is still more able of learning, adapting, and reasoning than computerized systems. Human operators can identify repetitive patterns in production, for example, and can use them to anticipate events or to optimize production over broader time windows.

Consequently, keeping operators in the SFC decisional loops shown in Fig. 1 is still relevant. All the research in these different areas (architectural, functional, and technological considerations) could lead to more human-centered and cognitive SFC systems [36] as the knowledge of human operators shall be merged more easily with the knowledge constructed by various active/intelligent interacting entities rather than in purely centralized SFC approaches. The technology currently available enables this smart integration. Now, time has come for researchers to reintegrate human operators in the loop; fully automated shop floors with nobody inside are no longer considered. However, the interaction of emerging SFC

systems with human operators is still poorly addressed by researchers.

Acknowledgment

The authors wish to acknowledge their debt to Thierry Berger for his help in writing this article.

REFERENCES

1. NCS, National Communication System. Supervisory Control and Data Acquisition (SCADA) Systems, Technical Information Bulletin NCS TIB 04-1; 2004.
2. Dieu B. Application of the SCADA system in wastewater treatment plants. *ISA Trans* 2001;40(3):267–281.
3. Wysk RA, Smith JS. A formal functional characterization of shop floor control. *Comput Ind Eng* 1995;28(3):631–643.
4. Morel G, Valckenaers P, Faure J-M, *et al.* Manufacturing plan control challenges and issues. *Control Eng Pract* 2007;15(11):1321–1331.
5. Harjunkski I, Nyström R, Horch A. Integration of scheduling and control—theory or practice? *Comput Chem Eng* 2009;33:1909–1918.
6. IEC, International Electrotechnical Commission. IEC 62264: Enterprise-control system integration—Part 1: models and terminology; March 2003.
7. Kabashkin I. Optimal monitoring strategies. *Wiley Encyclopedia of Operations Research and Management Science*. NJ: Hoboken; 2010.
8. Zio E. Availability analysis: concepts and methods. *Wiley Encyclopedia of Operations Research and Management Science*; 2010.
9. Benton WC. Just-In-Time/Lean production systems. *Wiley Encyclopedia of Operations Research and Management Science*; 2011a.
10. Goldratt EM, Cox J. *The Goal: A Process of Ongoing Improvement*. New York: North River Press; 1986.
11. Benton WC. Push and pull production systems. *Wiley Encyclopedia of Operations Research and Management Science*. NJ: Hoboken; 2011b.
12. Spearman M, Woodruff D, Hopp W. CONWIP: a pull alternative to Kanban. *Int J Prod Res* 1990;28:879–894.
13. Costanza JR. *The quantum leap in speed to market*. Englewood: John Costanza Institute of Technology, Inc.; 1996.

14. Wysk RA, Peters BA, Smith JS. A formal process planning schema for shop floor control. *Eng Des Autom* 1995;1.1:3–19.
15. Son Y-J, Joshi SB, Wysk RA, *et al.* Simulation-based shop floor control. *J Manuf Syst* 2002;21(5):380–394.
16. Shaikh NI, Prabhu VV. Monitoring and prioritising alerts for exception analytics. *Int J Prod Res* 2009;47.10:2785–2804.
17. Duffie NA, Prabhu VV. Real-time distributed scheduling of heterarchical manufacturing systems. *J Manuf Syst* 1994;13(2):94–107.
18. Prabhu VV. Performance of real-time distributed arrival time control in heterarchical manufacturing systems. *IIE Trans* 2000;32:323–331.
19. Hong J, Prabhu VV. Modeling and performance of distributed algorithm for scheduling dissimilar machines with setup. *Int J Prod Res* 2003;41(18):4357–4382.
20. Prabhu VV. Distributed control algorithms for scalable decision-making from sensors-to-suppliers. *Scalable Enterp Syst* 2003a;3:101–159.
21. Cho S, Prabhu VV. Distributed adaptive control of production scheduling and machine capacity. *J Manuf Syst* 2007;26.2:65–74.
22. Prabhu VV. Stability and fault adaptation in distributed control of heterarchical manufacturing job shops. *IEEE Trans Robot Autom* 2003b;19(1):142–147.
23. Berger T, Sallez Y, Valli B, *et al.* Semi-heterarchical allocation and routing processes in FMS control: a stigmergic approach. *J Intell Robot Syst* 2010;58(1):17–45.
24. Trentesaux D. Distributed control of production systems. *Eng Appl Artif Intell* 2009;22(7):971–978.
25. Valckenaers P, Van Brussel H. Holonic manufacturing execution systems. *CIRP Ann - Manuf Technol* 2005;54(1):427–432.
26. Ghezail F, Pierreval H, Hajri-Gabouj S. Analysis of robustness in proactive scheduling: a graphical approach. *Comput Ind Eng* 2010;58:193–198.
27. Pfeiffer A, Kádár B, Monostori L, *et al.* Simulation as one of the core technologies for digital enterprises: assessment of hybrid rescheduling methods. *Int J Comput Integr Manuf* 2008;21:206–214.
28. Monostori L, Erdos G, Kádár B, *et al.* Digital enterprise solution for integrated production planning and control. *Comput Ind* 2010;61:112–126.
29. Sallez Y, Berger T, Deneux D, *et al.* The life-cycle of active and intelligent products: the augmentation concept. *Int J Comput Integr Manuf* 2010;23(10):905–924.
30. Meyer GG, Främling K, Holmström J. Intelligent products: a survey. *Comput Ind* 2009;60:137–148.
31. Prabhu V, Jeon HW, Taisch M. Modeling green factory physics—an analytical approach. *Proceedings of IEEE CASE 2012*; Seoul, Korea; August 2012.
32. Karnouskos S, Colombo AW. Architecting the next generation of service-based SCADA/DCS system of systems. 37th Annual Conference of the IEEE Industrial Electronics Society (IECON 2011); Melbourne, Australia.
33. Onori M, Lohse N, Barata J, *et al.* The IDEAS project: plug & produce at shop-floor level. *Assembly Autom* 2012;32(2):124–134.
34. Karnouskos S, Colombo AW, Bangemann T, *et al.* A SOA-based architecture for empowering future collaborative cloud-based industrial automation. 38th Annual Conference of the IEEE Industrial Electronics Society (IECON 2012); Montréal, Canada; 2012.
35. Zuehlke D. SmartFactory—towards a factory-of-things. *Annu Rev Control* 2010;34: 129–138.
36. Zaeh MF, Reinhart G, Ostgathe M, *et al.* A holistic approach for the cognitive control of production systems. *Adv Eng Inform* 2010;24:300–307.

INTRODUCTION TO STOCHASTIC APPROXIMATION

GEORGE YIN
Department of
Mathematics, Wayne
State University,
Detroit, Michigan

Stochastic approximation (SA) methods were introduced by Robbins and Monro in their pioneering work in 1951 [1]. The original aim was to locate zeros of a nonlinear continuous function. That is, to find the roots $f(x) = 0$ when either the precise form of this function is unknown, or it is too complicated to compute, but one is able to take “noisy” measurements or observations at desired values. The procedure is known as the *Robbins–Monro (RM) algorithm*. The suggested solution method is a stochastic recursive algorithm.

A classical example is to find the appropriate dosage level of a drug, provided only $f(x)$ +noise is available, where x is the level of dosage and $f(x)$ is the probability of success leading to the recovery of the patient at dosage level x . In 1952, Kiefer and Wolfowitz proposed another algorithm that is concerned with the minimization of a real-valued function using only noisy functional measurements [2]. That algorithm is commonly referred to as the *Kiefer–Wolfowitz (KW) algorithm*.

Soon after, researchers realized that SA algorithms can be applied to a wide variety of situations in optimization and in systems with a “control” parameter. Owing to its importance, SA has had a long history and has attracted much attention in the past five decades. Significant progress has been made in the study of such stochastic recursive algorithms. The interesting theoretical issues in the analysis of iteratively defined stochastic processes focus on the basic paradigm of stochastic difference equations.

Much of the development has originated from a wide range of applications in optimization, control theory, economic systems, signal processing, communication theory, learning, pattern classification, neural network, and many other related fields. A number of monographs have been written, each of which has its own distinct features; for instance, the books of Albert and Gardner [3], Wasan [4], Nevel’son and Khasminskii [5], Kushner and Clark [6], Benveniste *et al.* [7], Chen [8], and Kushner and Yin [9].

HISTORICAL DEVELOPMENT

To put things in the historical perspective, we may divide the development of the SA methods into several periods. Since SA is a recursive scheme, convergence and rates of convergence are crucial. In the early development around 1950s and 1960s, mainly basic probabilistic tools and traditional statistical assumptions such as independent and identically distributed noise, together with certain restrictions on the functions such as assuming $f(x)$ to be monotone were used. In the original work [1], convergence in probability was obtained.

Summarizing much of the early historical development, the book [4] included the with probability one (w.p.1) convergence proof for multidimensional problems of Blum, the asymptotic normality study of Sacks, and the work of Fabian, among others. People soon recognized that the SA algorithms are closely related to stochastic dynamic systems. As a representative, Nevelson and Khasminskii’s book [5] treated SA as stochastic processes and dealt with martingale difference type noise. The use of Liapunov function was also brought in, together with recursive estimation.

As time went on, many applications arising in control and optimization forced researchers to reexamine the algorithms in which the noise encountered was correlated.

In the mid 1970s, Ljung studied the problem from a dynamic system point of view [10]. The idea is that in lieu of the discrete recursion, one treats a continuous-time dynamic system given by an ordinary differential equation (ODE). For example, for the root finding problems, once the associated ODE is obtained, the roots being sought become stable points of the associated ODE. Such an idea was further developed in Ref. 6. By combining analysis and probabilistic argument, Kushner and Clark set up a framework by considering asymptotic properties of suitably scaled sequences. The work of Benveniste *et al.* [7] emphasized the close connection of SA and adaptive systems. One of the distinct features is the use of Markovian setting and the treatment of the Poisson equations. In addition, it exploits many identification, estimation, and tracking problems. The book by Chen [8] summarizes the work of using random varying truncation bounds developed by Chen and Zhu; it also provides trajectory subsequence-based convergence proof. The work of Kushner and Yin [9] presents a comprehensive development of the modern theory of SA and recursive stochastic algorithms for both constrained and unconstrained algorithms, with step sizes that either go to zero or are constant and small, and possibly random.

More recently, SA techniques have been employed in the analysis of simulated annealing type of global stochastic optimization problems [11]; emerging applications have also been found in wireless communications (9,12, Chapter 3), discrete optimization [13], financial engineering [14], and so on. For a wide range of connections of SA to simulated annealing, temporal difference algorithms, Markov chain Monte Carlo, experimental design, and so on, and general stochastic search and optimization, the reader is referred to Ref. 15.

ALGORITHMS

RM Algorithm

We begin with the RM algorithms for finding zeros of a nonlinear function. Let $f : \mathbb{R}^r \mapsto \mathbb{R}^r$ be a continuous function. We wish to find

$f(x) = 0$, but only noisy measurements

$$y_n = f(x_n) + \xi_n$$

are available, where $\{\xi_n\}$ denotes a sequence of random noise. Note that n is a positive integer referred to as *discrete time*, representing the iteration number. The algorithm proposed by Robbins and Monro takes the form

$$x_{n+1} = x_n + a_n y_n, \quad (1)$$

where $\{a_n\}$ is a sequence of nonnegative real numbers satisfying $a_n \rightarrow 0$ as $n \rightarrow \infty$, $\sum_n a_n = \infty$, and $\sum_n a_n^2 < \infty$. The sequence $\{a_n\}$ is usually referred to as a *sequence of step sizes* or *gains*. The conditions on the step sizes indicate that they cannot be too small. If they are too small (i.e., $\sum_n a_n < \infty$), then the iterates may not ever converge to the desired value. To see this, take the noise-free case $\xi_n = 0$ and suppose that $f(\cdot)$ is a bounded function. Then

$$\begin{aligned} \left| \sum_{j=0}^{\infty} [x_{j+1} - x_j] \right| &\leq \sum_{j=0}^{\infty} |x_{j+1} - x_j| \\ &\leq \sum_{j=0}^{\infty} a_j |f(x_j)| \leq \kappa. \end{aligned}$$

[Here and throughout the paper, $\kappa > 0$ is used as a generic constant; its value may change for different usage. Thus by our convention, $\kappa + \kappa = \kappa$ and $\kappa \kappa = \kappa$.] The above argument indicates that $\sum_j (x_{j+1} - x_j)$ converges absolutely. On the other hand, by telescoping

$$\sum_{j=0}^n [x_{j+1} - x_j] = x_{n+1} - x_0.$$

Thus $x_n \not\rightarrow x^*$, the true parameter that we are approximating unless x_0 is sufficiently close to x^* .

KW Algorithm

Here, we wish to minimize a real-valued function $f(x) : \mathbb{R}^r \mapsto \mathbb{R}$ based on noise-corrupted observations $F(x, \zeta_n)$ with $EF(x, \zeta_n) = f(x)$,

where $\{\zeta_n\}$ represents the noise. The difficulty is that we know neither the form of $F(\cdot)$ nor $f(\cdot)$. To approximate the minimizer, we use the finite difference approximation to the gradient of $f(x)$. Denote the finite difference interval by $\{c_n\}$ (with $c_n \rightarrow 0$ as $n \rightarrow \infty$). Use x_n to denote the n th estimate of the minimum. Suppose that for each i and each n , we observe

$$y_{n,i} = -\frac{F(x_n + c_n e_i, \zeta_n^+) - F(x_n - c_n e_i, \zeta_n^-)}{2c_n},$$

where $\zeta_n^\pm$ are random noise sequences. Denote $y_n = (y_{n,1}, \dots, y_{n,r})$. Then the approximation algorithm is again given by $x_{n+1} = x_n + a_n y_n$, which is of the same form as that of Equation (1). By introducing

$$\xi_n = (\xi_{n,1}, \dots, \xi_{n,r}), \beta_n = (\beta_{n,1}, \dots, \beta_{n,r}),$$

with

$$\begin{aligned} \xi_{n,i} &= \left[f(x_n + c_n e_i) - F(x_n + c_n e_i, \zeta_{n,i}^+) \right] \\ &\quad - \left[f(x_n - c_n e_i) - F(x_n - c_n e_i, \zeta_{n,i}^-) \right], \\ \beta_{n,i} &= f_{x_i}(x_n) - \frac{f(x_n + c_n e_i) - f(x_n - c_n e_i)}{2c_n}, \end{aligned}$$

where $f_{x_i}(\cdot)$ denotes the gradient of $f(\cdot)$ w.r.t. x . Now the above algorithm can be rewritten as

$$x_{n+1} = x_n - a_n f_x(x_n) + a_n \frac{\xi_n}{2c_n} + a_n \beta_n. \quad (2)$$

In the above, ξ_n represents the noise, β_n denotes the bias, $c_n \rightarrow 0$, and $\sum_n \frac{a_n^2}{c_n^2} < \infty$. We have used two-sided finite difference. One-sided finite difference can also be used. However, in practice, the two-sided finite difference method appears to be more preferable since it has smaller bias. This is easily seen by taking a Taylor expansion of the finite difference quotient in β_n . In the above discussion, we assumed f being smooth. When this condition is violated, the subgradient can be used in lieu of the gradient; see Kushner and Yin [9].

More General Algorithms and Variants

In various applications such as those arising in signal processing and adaptive controls, one often needs to treat stochastic approximation algorithms with noise having a more complex structure of the form

$$x_{n+1} = x_n + a_n f(x_n, \xi_n). \quad (3)$$

Including both Equations (1) and (2) as special cases, such algorithms arise, for example, in “equalization” filters in communication channels, adaptive antenna array processing, and so on (9, Chapter 3). In addition, $f(\cdot)$ can also depend on n :

$$x_{n+1} = x_n + a_n f_n(x_n, \xi_n). \quad (4)$$

To track slight parameter variations, one often uses an algorithm with constant step size of the form

$$x_{n+1} = x_n + \varepsilon f_n(x_n, \xi_n), \quad (5)$$

where $\varepsilon > 0$ is a small parameter. Because of the constant step size used, a pertinent notion of convergence is in the sense of weak convergence of probability measures (9, Chapter 7). Such constant step size algorithms are particularly suitable for tracking slight time-varying parameters. Suppose that H is a constraint set. We require the iterates to be in the set. Write the algorithm as

$$x_{n+1} = \Pi_H(x_n + a_n f(x_n, \xi_n)), \quad (6)$$

where Π_H denotes the projection onto the constraint set H . If the iterate is within the projection region, we simply keep the recursion running; if it is outside the region, we project it back. Define a “reflection” or correction term z_n as

$$a_n z_n = x_{n+1} - x_n - a_n f(x_n, \xi_n),$$

that is, it is the vector of the shortest Euclidean length needed to take $x_n + a_n f(x_n, \xi_n)$ back to the constraint set H if it is

not in H . Using this notation, Equation (6) can be rewritten as

$$x_{n+1} = x_n + a_n f(x_n, \xi_n) + a_n z_n. \quad (7)$$

The constraint set can be as simple as a rectangular region, or as complex as a complicated set bounded by nonlinear functions; see Section 4.3 of Ref. 9 for various possible projection regions.

In addition to the algorithms mentioned above, many other variants have been studied. For example, if one can get the observation y_n but the sequence $\{x_n\}$ emerges in a random manner and is not at one's disposal, then one is in the so-called passive SA framework. Sometimes, we may wish to use constraints but wish to relax the "hard constraints," and allow them to be violated slightly from time to time. Roughly speaking, such constraints are "soft constraints." In addition, as was mentioned earlier, one may also define a sequence of randomly generated truncation bounds. This will enable one to relax the conditions on the growth rates of the function under consideration or the *a priori* knowledge of the constraint set.

CONVERGENCE AND RATES OF CONVERGENCE

Convergence

To study the convergence of the algorithm, we use the ordinary differential equation method. Instead of working with the discrete iteration directly, we take a continuous-time interpolation. To get some insight, we first give some heuristic argument.

Consider Equation (1). Suppose that the function $f(\cdot)$ is continuous. Although rather complex noise processes can be dealt with [9], for simplicity, in this survey, we assume the measurement or observation error is a sequence of independent and identically distributed random variables. Also, for simplicity, assume the iterates $\{x_n\}$ are bounded. The boundedness assumption is not crucial but can simplify the presentation here; a more general treatment is given in

Ref. 9. Choose a small $\Delta > 0$ such that

$$\sum_{j=n}^{n+m_n^\Delta-1} a_j \approx \Delta, \text{ or equivalently}$$

$$m_n^\Delta = \max \left\{ m; \sum_{j=n}^{n+m-1} a_j < \Delta \right\}.$$

Iterating on Equation (1) yields

$$x_{n+m_n^\Delta} = x_n + \sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_j) + \sum_{j=n}^{n+m_n^\Delta-1} a_j \xi_j.$$

For Δ small enough, and for $n \leq j \leq n + m_n$, $f(x_j)$ is "close" to $f(x_n)$ by the continuity. As a result

$$\sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_j) \approx \sum_{j=n}^{n+m_n^\Delta-1} a_j f(x_n) \approx \Delta f(x_n),$$

and

$$x_{n+m_n^\Delta} - x_n \approx \Delta f(x_n) + \text{error}.$$

To see the size of the error, we compute its variance:

$$\begin{aligned} \text{Var}(\text{error}) &= \text{Var} \left(\sum_{j=n}^{n+m_n^\Delta-1} a_j \xi_j \right) \\ &= O \left(\sum_{j=n}^{n+m_n^\Delta-1} a_j^2 \right) = O(\Delta a_n). \end{aligned}$$

Therefore,

$$\frac{x_{n+m_n^\Delta} - x_n}{\Delta} \approx f(x_n) + \text{error with diminishing variance}.$$

Over a small interval, the mean change of the values of the parameter is much more important than that of the noise. The noise is averaged out in the limit, and the asymptotic behavior can be approximated by the ODE

$$\dot{x} = f(x). \quad (8)$$

If a projection algorithm is used, one obtains a projected ODE [9].

To prepare us for the study of the desired asymptotic properties, let us recall the notion of equicontinuity. Use $C([0, \infty) : \mathbb{R}^r)$ to denote the family of $\mathbb{R}^r$ -valued continuous functions defined on $[0, \infty)$. Let $\{f_n\}$ be a sequence of $\mathbb{R}^r$ -valued functions on $[0, \infty)$. It is said to be *equicontinuous* in $C([0, \infty) : \mathbb{R}^r)$, if $\{f_n(0)\}$ is bounded and for each T and $\varepsilon > 0$, there is a $\delta > 0$ such that for all n

$$\sup_{|t-s| \leq \delta, |t| \leq T} |f_n(t) - f_n(s)| \leq \varepsilon. \quad (9)$$

The well-known *Ascoli–Arzelá* Theorem states as follows: Let $\{f_n(\cdot)\}$ be a sequence of functions in $C([0, \infty) : \mathbb{R}^r)$, and let the sequence be equicontinuous. Then there is a subsequence that converges to some continuous limit, uniformly on each bounded interval.

For the SA algorithm, take piecewise linear interpolations and define

$$\begin{aligned} t_n &= \sum_{j=0}^{n-1} a_j, \quad m(t) = \max\{n : t_n \leq t\}, \\ x^0(t_n) &= x_n, \quad x^0(t) = \frac{t_{n+1} - t}{a_n} x_n + \frac{t - t_n}{a_n} x_{n+1}, \\ &\quad t \in (t_n, t_{n+1}). \end{aligned}$$

That is, the interpolation interval is $[t_n, t_{n+1})$. Next, to bring the asymptotic behavior of the process to the foreground, define a shifted sequence by $x^n(t) = x^0(t + t_n)$. Under suitable conditions, it can be shown that $\{x^n(\cdot)\}$ is uniformly bounded and equicontinuous. By the *Ascoli–Arzelá* Theorem, we can extract a convergent subsequence $x^{n_k}(\cdot)$ such that $x^{n_k}(\cdot) \rightarrow x(\cdot)$. Then, we characterize the limit $x(\cdot)$ and prove that it is nothing but the solution of the ODE (Eq. 8). Concerning the limit differential equations, we note that the stationary points of Equation (8) are exactly the roots of $f(\cdot)$ that we are searching for. If x^* is the unique asymptotically stable point of Equation (8), then we can show the iterates $x_n \rightarrow x^*$ w.p.1. If instead of a single point x^* , Z is a set of zeros of $f(\cdot)$ such that Z is asymptotically stable in the sense of Liapunov, then $x_n \rightarrow Z$ w.p.1.

In applications, it is often more convenient to work with piecewise constant interpolations. To use such an interpolation, we

need to modify and extend the definition of the equicontinuity. In (p. 102 of Ref. 9), an extension, known as *equicontinuity* in the extended sense was given. Then a version of the *Ascoli–Arzelá* Theorem for sequences of functions that are equicontinuous in the extended sense can be used, which facilitates the analysis for piecewise constant interpolated sequences.

Rates of Convergence

Suppose that $x_n \rightarrow x^*$ (the true parameter) w.p.1 as $n \rightarrow \infty$. Then one is interested in the rate of convergence. For SA, what do we mean by “rate of convergence?” Consider, Equation (1) for instance. To study the convergence rate, we take a suitably scaled sequence $u_n = (x_n - x^*)/a_n^\alpha$, for some $\alpha > 0$. In case of constant step size algorithm, this is changed to $(x_n - x^*)/\varepsilon^\alpha$. If α is chosen such that u_n converges in distribution to a nontrivial limit, the scaling factor α together with the asymptotic covariance of the scaled sequence gives us the rate of convergence. That is, the scaling α tells us the dependence of the estimation error $x_n - x^*$ on the step size, and the asymptotic covariance is a mean of assessing “goodness” of the approximation.

It turns out that a suitable scaling for Equation (1) is $\sqrt{a_n}$. If a constant step size algorithm is used, the scaling is $\sqrt{\varepsilon}$. To illustrate, consider Equation (1) with $a_n = 1/n$. Assume $f(x) = H(x - x^*) + O(|x - x^*|^2)$ such that $\tilde{H} = H + I/2$ is a stable matrix (i.e., all of its eigenvalues have negative real parts). The rate of convergence is based on local analysis. Using the definition u_n together with Equation (1), we obtain

$$u_{n+1} = u_n + \frac{1}{n} \tilde{H} u_n + \frac{1}{\sqrt{n}} \xi_n + O\left(\frac{1}{n} |u_n|^2\right).$$

The last term above is asymptotically unimportant (i.e., its limit will be 0 in an appropriate sense). Thus we can ignore it. Owing to the presence of $\sqrt{n}$, the noise is not going to be averaged out. We need to deal with $\sum_{k=1}^n \xi_k / \sqrt{k}$. This term is asymptotically equivalent to $\sum_{k=1}^n \xi_k / \sqrt{n}$, which can be further characterized by a normal distributed random variable with covariance

$\Sigma = E\xi_1\xi_1'$ by the well-known central limit theorem.

With more effort, we can even obtain a limit stochastic differential equation. To see this, define a piecewise constant interpolation $u^n(0) = u_n$ and $u^n(t) = u_k$ for $t \in [t_{k+n} - t_n, t_{k+n+1} - t_n)$. It can be shown that for $\ell \geq 1$, $\sum_{k=\ell}^{\ell+m(t)-1} \frac{1}{\sqrt{k}} \xi_k$ converges weakly to a Brownian motion $\tilde{w}(\cdot)$ with covariance Σt . We can further rewrite the $\tilde{w}(\cdot)$ in terms of the standard Brownian motion $w(\cdot)$ as $\tilde{w}(t) = \Sigma^{1/2} w(t)$. Furthermore, it can be shown that $u^n(\cdot)$ converges weakly to the solution of the stochastic differential equation

$$du = \tilde{H}u dt + \Sigma^{1/2} dw.$$

Note that the above stochastic differential equation is linear in u , and, as a result, it has a unique solution for each initial condition u_0 . The limit stochastic differential equation gives a better description of the dynamics of the suitably scaled estimation errors. We refer the reader to Chapter 8 of Ref. 9 for further reading. Denote the asymptotic covariance by S . Then it can be shown that S satisfies the following Liapunov equation:

$$\tilde{H}S + \tilde{H}'S = -\Sigma.$$

This further implies that under suitable conditions, $\sqrt{n}(x_n - x^*)$ converges in distribution to a normal random variable $N(0, S)$, which gives us the desired rate of convergence.

ASYMPTOTIC OPTIMALITY OF PERFORMANCE

One of the most important matters is to improve the performance of SA algorithms. Let us consider KW algorithms. Each step of the KW algorithm uses $2r$ observations (if two-sided finite difference is used). Thus $2r$ steps are needed to get a derivative estimate. An alternative method is to update only one direction at each iteration using a finite difference estimate and to choose the direction randomly at each step. This results in using only two observations at each step. Such an idea appeared in Ref. 6 together with

the associated convergence and rate of convergence results. At that time, it was noted that if the random directions are chosen at the unit sphere, then there is a small advantage as compared to the KW method. The work of Spall [16] indicated that if the random directions are chosen on the unit cube, then the performance is better than the KW algorithms; this is a significant discovery and enables one to treat high dimensional problems effectively. In fact, it was noted in Ref. 16, one can “arbitrarily” distribute perturbations, but the Bernoulli case (unit cube) has special optimality properties and is easier to implement.

Along another line in the late 1980s, Polyak [17] and Ruppert [18] independently discovered that averaging can help to improve the efficiency of SA algorithms. The main idea is to first generate a sequence of rough estimates using large step sizes and then average the resulting estimate to smooth things out. The algorithm is given by

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{n^\gamma} (f(x_n) + \xi_n), \\ \bar{x}_{n+1} &= \bar{x}_n - \frac{1}{n+1} \bar{x}_n + \frac{1}{n+1} x_{n+1}, \end{aligned} \quad (10)$$

where $1/2 < \gamma < 1$. It can be shown that such algorithms are asymptotically optimal in that they have the best scaling α and the smallest variance (9, Chapter 11); see also Yin (20). One may find discussion of the related finite-sample performance of the iterate averaging in Ref. 15 (Section 4.5.3). Moreover, the efficiency issue can also be studied in connection with the Cramér–Rao bounds [5].

The setting of SA can be carried over to infinite-dimensional spaces; for example, Banach spaces and/or Hilbert spaces. In addition to the pure mathematical interest, the motivation of the study stems from the fact that in various optimization problems, the solutions of the problems involve finding the root or the optimum for points not living in Euclidean spaces, but in function spaces. In this article, we focus on finite-dimensional cases. For infinite-dimensional problems, we refer the reader to Ref. 19 and the many references therein.

SUMMARY

To summarize, SA methods have been the subject of an enormous volume of research, both theoretical and applied, for five decades. Owing to the vast amount of literature accumulated, it is very difficult or virtually impossible to provide an exhaustive list of references on stochastic approximation. It is likewise very difficult to give an extensive account of the technical development in a survey paper of this scale. As a result, we chose the road of discussing the main ideas and left most of the technical details aside. Appropriate references have been provided. Our hope is that with the references provided at the end of this article, the reader will be able to pick out references suitable for his/her needs.

Acknowledgments

This author's research was supported in part by the National Science Foundation under DMS-0907753, and in part by the National Security Agency under grant MSPF-068-029.

REFERENCES

1. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22:400–407.
2. Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Ann Math Stat* 1952;23:462–466.
3. Albert AE, Gardner LA. Stochastic approximation and nonlinear regression. Cambridge (MA): MIT Press; 1967.
4. Wasan MT. Stochastic approximation. London: Cambridge Press; 1969.
5. Nevel'son MB, Khasminskii RZ. Stochastic approximation and recursive estimation. Volume 47, Translation of mathematical monographs. Providence (RI): AMS; 1976.
6. Kushner HJ, Clark DS. Stochastic approximation methods for constrained and unconstrained systems. New York: Springer; 1978.
7. Benveniste A, Metivier M, Priouret P. Adaptive algorithms and stochastic approximation. Berlin: Springer; 1990.
8. Chen H-F. Stochastic approximation and its applications. Dordrecht, The Netherlands: Kluwer Academic Press; 2002.
9. Kushner HJ, Yin G. Stochastic approximation and recursive algorithms and applications. 2nd ed. New York: Springer; 2003.
10. Ljung L. Analysis of recursive stochastic algorithms. *IEEE Trans Automat Control* 1977;AC-22:551–575.
11. Yin G. Rates of convergence for a class of global stochastic optimization algorithms. *SIAM J Optim* 1999;10:99–120.
12. Yin G, Krishnamurthy V. Least mean square algorithms with Markov regime switching limit. *IEEE Trans Automat Control* 2005;50:577–593.
13. Yin G, Krishnamurthy V, Ion C. Regime switching stochastic approximation algorithms with application to adaptive discrete stochastic optimization. *SIAM J Optim* 2004;14:1187–1215.
14. Yin G, Liu RH, Zhang Q. Recursive algorithms for stock liquidation: a stochastic optimization approach. *SIAM J Optim* 2002;13:240–263.
15. Spall JC. Introduction to stochastic search and optimization: estimation, simulation, and control. Hoboken (NJ): Wiley; 2003.
16. Spall JC. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans Automat Control* 1992;AC-37:332–341.
17. Polyak BT. New method of stochastic approximation type. *Automat Remote Control* 1990;51:937–946.
18. Ruppert D. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report, No. 781, School of Operations Research & Industrial Engineering, Cornell University; 1988. [see also the chapter Stochastic approximation. In: Ghosh BK, Sen PK, editors. Handbook in sequential analysis. New York: Marcel Dekker; 1991. pp. 503–529.]
19. Berger E. Asymptotic behavior of a class of stochastic approximation procedures. *Probab Theory Relat Fields* 1986;71:517–552.
20. Yin G. Stochastic approximation: theory and applications. In: Kannan D, Lakshmikantham V, editors. Handbook of stochastic analysis and applications. New York: Marcel Dekker; 2002. pp. 577–624.

INTRODUCTION TO THE USE OF LINEAR PROGRAMMING IN STRATEGIC HEALTH HUMAN RESOURCE PLANNING

MARIEL S. LAVIERI
University of Michigan, Ann
Arbor, MI, USA

SANDRA REGAN
The University of Western
Ontario, London, Ontario,
Canada

MARTIN L. PUTERMAN &
PAMELA A. RATNER
University of British Columbia,
Vancouver, British Columbia,
Canada

Effective human resource planning decisions are fundamental to achieving sustainable healthcare systems. The value of health professionals is undeniable, and a shortage of workers significantly affects the number of patients that can receive healthcare, the quality and effectiveness of the services provided, workforce attrition rates, and the length of time patients must wait for such care [1]. Health human resource (HHR) planning goals include identifying and achieving “the optimal number, mix, and distribution of personnel at a cost society is able to afford” [2]. Decision models that integrate elements such as the supply of and demand for providers, along with economic considerations, are crucial for effective HHR planning.

Decisions related to workforce supply have long-term implications. It takes time to educate sufficient numbers of health professionals to replace those who retire. Educating each health professional represents a significant societal financial investment. Admitting too many students is therefore not desirable; it is important that each newly graduated healthcare professional is able to take up a position in the healthcare system. Balancing the number of educational seats with the

in-migration of health professionals, while accounting for the impact of other factors (such as working hours, attrition from the educational programs, and productivity [3]) on the supply of professionals, needs to be considered.

Identifying the “optimal” number and mix of HHR has been repeatedly highlighted as a necessary goal of planning [2, 4]. However, the determination of the “optimal” number and mix of providers (i.e., the number and mix of HHR resources that are able to achieve best outcomes) will vary, and it is dependent on how need and demand for providers is defined. Most methods and models used in practice focus on limited or narrow aspects of planning, and they do not consider the multiple factors and contexts that can influence supply and demand. For example, forecasting models are often produced in isolation of their economic implications and may lead to recommendations that might be believed to be inconsistent with fiscal goals of decision makers. They are perceived as unrealistic because of the constraints in other sectors of the HHR planning system. The models often project estimates for providers holding conditions such as healthcare utilization constant [5]. Further, decisions made today about numbers of education seats, changes in employment positions, or strategies to retain professionals (e.g., decreasing attrition rates) are often examined in isolation.

This article presents an overview of long-term HHR planning models and an introduction to how linear programming may be used to model strategic HHR decisions faced by large-scale healthcare systems. The remainder of this article is organized as follows. This article starts with an introduction to strategic HHR models. The following section provides examples of strategic workforce planning models within the operations research literature. The next section describes in further detail the linear programming approach and key inputs and outputs to consider in such models. We

conclude by summarizing this article and discussing modeling considerations.

HEALTH HUMAN RESOURCE PLANNING MODELS

Given the importance of decisions related to the appropriate supply of professionals to meet demand, various models to estimate the supply and demand have been developed. Approaches to planning HHR have largely focused on physicians and nurses and have been primarily based on supply projections [6]. Historically, provider-to-population ratios have been utilized to project the demand for physicians and nurses. As the collection and availability of data about health professionals have improved, methods to estimate the future supply of health professionals have evolved in their sophistication. Advances in other fields (as well as improvements in computational power) have played a role in the evolution of HHR planning models.

An inventory of key HHR planning models from various jurisdictions across Canada was conducted by Health Canada [7]. It is based on the classification suggested by O'Brien-Pallas *et al.* [6], which distinguishes between three types of models: *supply models*, *demand models*, and *econometric models*. *Supply models* focus on calculating available future resources. They include those proposed in the 1980s by Kazanjian *et al.* [8] and Kazanjian *et al.* [9] as well as models such as [10] and [11] that estimate the number (or utilization) of future health professionals based on descriptions of current workforce characteristics, expected entries, and attrition. More basic supply models (such as Archambault [12]) extrapolate from current conditions based on the growth, and in some cases the cyclical components, of the health workforce. *Demand* (or needs-based) *models* such as [13] and [14] focus on planning the HHRs needed based on general population demographics and health status indicators. Finally, the so-called *econometric approaches* (or resource allocation models), such as Dickson [15], attempt to determine the feasibility of various scenarios based on

the calculation of the associated costs or the application of costs as a constraint.

Although these models provide a starting point to better understand supply and demand for health professionals, none is sufficient on its own to address the current and future challenges of planning the health workforce. Supply models assume that the status quo is appropriate when determining future supply. Demand models concentrate on future health professional requirements, taking supply projections as a given. They do not incorporate the ability of the decision maker to control supply to meet those requirements. They take current supply as a starting point for the estimates, regardless of whether the current supply is sufficient or appropriate. They also assume that current demand is predictive of future demand without considering the changing health status of the population (such as changes in life expectancy, the prevalence of risk factors, or interventions such as self-care or technological advances). Finally, resource allocation approaches determine requirements based on anticipated future costs or financial resources and often do not consider population health needs.

Policy decisions about HHRs can have significant effects on their supply and deployment. An important objective of any effective modeling approach is to identify relevant scenarios and those factors that influence the scenarios. Simulation models are increasingly being used to examine the impact of recruitment and retention strategies on supply and demand [16, 17]. In contrast to how this term is used in operations research [18], simulation here involves projecting the nursing work size based on a proposed plan. On the basis of current demographics (such as workforce entry and attrition), simulation is used to determine the impact of a given plan on population health indicators. This approach assumes that the decision maker is able to derive (in multiple tries) a plan that is both feasible and meets the recognized needs.

EXAMPLES OF RESOURCE PLANNING MODELS WITHIN THE OPERATIONS RESEARCH LITERATURE

Within the operations research context, various approaches have been developed to guide strategic workforce decisions. Among others, models have been applied in the long-term planning of military human resources [19], planning of personnel and determining optimal staffing policies in call centers [20], and industries including chemical [21] and aerospace, automotive, machine tool and, semiconductor [22]. Such models demonstrate the capacity of using operations research to guide workforce planning decisions. However, there continues to be great potential to further incorporate the intricacies of the healthcare system in the long-term human resource models developed.

The approach used will depend on the question addressed. For instance, in modeling, the size and composition of a workforce, newsvendor-type staffing models are used to incorporate stochastic effects such as absences and wastages [23] as well as demand uncertainty [24], whereas forecasting and queuing models provide insights into the relationship among workers available and delays in the system [25], bottlenecks [26], and the need for temporary personnel [27]. Integer and mixed-integer programs are often used to model the tasks performed by each worker, which may involve movement restrictions between workstation groups [28] and restrictions in skill acquisition [29]. Simulation and system dynamics models (such as Vanderby *et al.* [30]) enable the incorporation of feedback loops and other complexities of the system. For higher level planning, dynamic programming (such as [31, 32]) and linear programming (such as [33–37]) may be used to model workforce training, promotion, and recruitment decisions. While dynamic programming (deterministic or stochastic) has the ability to model the exact number of workers in the system, it suffers from the curse of dimensionality. Alternatively, for large healthcare systems, a linear programming approach may be used to gain insights into the system. Notice that a linear program (LP) will not

restrict variables to be integers, yet such an approximation may be appropriate when dealing with large systems.

In the remaining sections, we focus on the linear programming approach but encourage the interested reader to review other approaches described in this section and to select the approach that best fits their modeling needs.

A LINEAR PROGRAMMING APPROACH

A LP is a mathematical tool that enables “the planning of activities to obtain an optimal result, that is, a result that reaches the specified goal best among all feasible alternatives” [38]. Desirable features of an LP are that it:

- Finds the best solution possible to meet given targets and does not rely on judgment or trial and error to find a good solution.
- Takes only seconds to solve, which is ideal for investigating various scenarios and performing sensitivity analyses.
- Has a strong mathematical foundation to support the decision-making process.

Different from the HHR models described, a linear programming approach does not take possible educational plans as inputs to identify the impact on vacancies or population health, rather it allows the decision maker to choose desired minimum targets (which could be determined through various demand or needs-based approaches) and then determine the optimal way of meeting those targets, by providing an optimal education, promotion, and recruitment plan. Such a model combines characteristics of the three types of HHR models described. From the supply models, it considers the current supply and attrition rates to track the number of workers of each age in the workforce. It encompasses demand models by constraining the number of health professionals that are needed in the system to meet population needs. Finally, it compares *all* feasible plans to determine an optimal strategy that is one that achieves a supply–demand balance at the least financial cost to the system and meets the growing

demand for healthcare providers. When a decision maker provides a scenario (i.e., a set of assumptions about the future), a linear programming model provides an optimal plan under that scenario, hence enabling the decision maker to interactively investigate the impact of the assumptions.

An LP has three main components: decision variables, an objective function, and a set of constraints. It determines values for the decision variables so that the objective function is as high or low as possible (depending on whether the decision maker seeks to minimize or maximize the objective) while simultaneously ensuring that all constraints are satisfied. Therefore, rather than relying on multiple tries in search of a strategy that will both meet a set of requirements and be as good as possible, the mathematical structure of the LP ensures this.

Lavieri and Puterman [33] provide an example of the mathematical formulation for such a model. The decision variables in the example are the number of students admitted ($s(.)admitted_j$), the workforce recruited ($n(.)recruited$), and the workforce promoted ($n(.)promoted_j$) in year j of the planning horizon ($1 \leq j \leq N$, where N represents the length of the planning horizon). These decision variables may be used to calculate the total number of students $s(.)_j$ and workers $n(.)_j$ in the system and are nonnegative. As shown in Lavieri and Puterman [33], the decision variables can be modified to incorporate educational programs of different lengths (which is common in HHR planning) as well as different levels of workforce (such as managers and senior managers).

Once the decision variables have been defined, the planner must choose the objective function (which will be used to evaluate the feasible solutions). An objective function often used in HHR planning is to minimize the total cost incurred over the planning horizon. If we account for the training cost associated with educating each student $tsCost$, the recruitment cost associated with hiring a health professional $rn(.)Cost$, the costs associated with promoting each health professional into a new role $tn(.)Cost$, and the annual salary $sn(.)Cost$ paid to healthcare workers in each position, the objective function may

then be written as

$$\min \sum_j tsCost(s(.)_j) + rn(.)Cost(n(.)recruited_j) + tn(.)Cost(n(.)promoted_j) + sn(.)Cost(n(.)_j)$$

To obtain the present value of the cost incurred, a discount factor may be added to the objective function calculations.

Finally, the last component of the LP is the set of constraints to be met. For example, it is important to ensure that minimum quality standards are met. This can be expressed in terms of the minimum number of trained health professionals that must be available and willing to work each year ($n(.)min$). Such a constraint can be written as

$$n(.)_j \geq n(.)min \text{ for } 1 \leq j \leq N$$

where $n(.)min$ may be determined from demand-based models or, in the case of hierarchical models, $n(.)min$ may be a function of the number of workers that need to be supervised.

Furthermore, the assumed flow of health professionals through the system should be well modeled; for example, a student can only be in the fourth year of an educational program if the third year of the program has been completed and there exists an intention to continue. Assume that it takes exactly 1 year to complete a level of the program. If we let $s(k)_j$ be the number of students in the k th year of the educational program ($1 \leq k \leq 4$ for programs that last four years) and let $Pcontinuing(k)$ be the probability that a student in the k th year of the educational program continues to the next year of the program, the above-mentioned example may be written as

$$s(4)_j = s(3)_{j-1}Pcontinuing(3) \text{ for } 1 \leq j \leq N$$

where $s(3)_0$ represents the initial number of students in the third year of the program. For the first level of the program, the total number of students will equal the number of students admitted to the first level of the program ($s(1)admitted_j$):

$$s(1)_j = s(1)admitted_j$$

Some healthcare programs admit students who have completed related studies in other fields into higher levels of the educational program. For instance, if students can be admitted into the third year of the program (as is the case with the RN diploma to Baccalaureate programs in nursing), the flow constraint may be written as

$$s(3)_j = s(2)_{j-1}P_{continuing(2)} + s(3)_{admitted_j} \text{ for } 1 \leq j \leq N$$

Similar flow constraints may be written for all other educational levels. In a similar manner, the number of workers at each level l ($1 \leq l \leq M$, where M is the number of workforce levels modeled) will depend on the number of workers in the previous year ($n(l)_{j-1}$), the number of workers promoted to ($n(l)_{promoted_j}$) and from ($n(l+1)_{promoted_j}$) that level, the number of workers recruited at that level ($n(l)_{recruited_j}$), and the attrition from the profession ($P_{retire(l)}$). Assuming that workers can only be promoted to subsequent levels, this constraint may be written as

$$n(l)_j = n(l)_{j-1}(1 - P_{retire(l)}) + n(l)_{promoted_j} + n(l)_{recruited_j} - n(l+1)_{promoted_j} \text{ for } 1 \leq j \leq N \text{ and } 1 \leq l \leq M$$

where $n(0)_{promoted_j}$ is based on the students in the last year of the program who graduate and are able to practice in the profession and $n(M+1)_{promoted_j}$ should equal 0. Moreover, there might be other restrictions that the decision maker might face such as full-time equivalence calculations of workers in the system (as part-time workers are common in some health professions), limitations on the number of professionals that can be recruited, limitations on the workers that can be promoted, and how quickly educational programs can be expanded. The formulations of such constraints are further discussed in Lavieri and Puterman [33].

Implementing the Linear Programming Approach

We developed such a model on an MS-Excel® platform [34] and used the Frontline solver add-on [39] to obtain solutions. While the size of the model can be appreciated from the screenshot provided, the user-friendly interface and the ease of introducing changes to the model's parameters contribute to its appeal and efficacy.

When opening the tool, the decision maker is first taken to the *Model Setup* tab to specify the general characteristics of the model: the ages modeled (from 18 until 65 in our sample model), the number of periods to model (20 periods in our example), and the levels of workforce (4 years of study and 3 levels of workers in our example).

The next step involves identifying the inputs to the model (initial conditions, probabilities, and costs). The key inputs to the model are summarized in Table 1. Not all quantities will have the same impact on the solution. One attractive feature of a linear programming approach is that even if the decision maker encounters limitations in data accessibility—or is unsure of some assumptions—by formulating the problem as an LP, it is possible to determine how much the solution would change if the objective or constraints were changed or if further constraints were added or deleted. The decision maker might thus place greater emphasis on the accurate collection and calculation of those parameters that have the greatest impact on the decision-making process. Table 1 provides general descriptions of inputs and data sources. The choice of inputs is provider specific and dependent on the scenarios that the decision makers will want to examine. Additional elements or inputs can be identified and are only limited by data availability and the assumptions made about the inputs.

On the basis of these inputs, solving the LP determines a minimum cost plan. Even in the large instance described (with 20 periods, 4 years of study and 3 levels of workers, 126 decision variables, and 7497 bookkeeping variables), the solution is

Table 1. Potential Model Inputs

HHR Element	Potential Inputs	Description	Potential Data Sources
Production	Education costs	Annual cost of funding a health professional education program seat	Government, education programs
Production	Age demographics	Current demographics of students	Education programs, accreditation bodies, and regulatory bodies
Production	Probability of continuing education	Fraction of students that complete each year of the education program	Education programs
Supply	Probability of passing licensure examination	Fraction of graduates that pass the licensure examination	Regulatory bodies
Supply	Probability of leaving the province after graduation	Fraction of health professionals that do not remain in the province after graduation	Education programs and regulatory bodies
Supply	Age distribution	Current demographics of health professionals	Employers, regulatory bodies, and census data
Supply	Attrition rates	Annual probability of permanently leaving the workforce in the province	Regulatory bodies and employers
Deployment	FTE	Full-time equivalents (or other similar measure) of each health professional. For example, full-time positions, hours worked, and number of clients seen per week	Employers and government
Deployment	Workforce ratios	Ratios of professionals in one role to another. For example, family to specialist physicians; private versus public employed pharmacists; and staff nurses to managers	Employers, government, and regulatory bodies
Financial costs	Annual salary	Annual salary or wage paid	Employers (including private sector), unions and professional associations, and government
Financial costs	Recruitment and turnover cost	Incentives, orientation, and other recruitment costs incurred per recruited health professional	Employers
Financial costs	Management education cost	Cost of training to promote a health professional to another role such as management or education	Educational institutions and employers
Need/demand	Demand for provider	These inputs will be provider specific and may include number of visits, hospital utilization, and population healthcare needs	Government, employers, and research literature
Need/demand	Projected population		Census data

obtained in seconds on a PC. Information that can be obtained from such a plan constitutes the outputs of the model. Outputs may be tailored to meet the needs of decision makers. Possible outputs include the number of students and health professionals each year in the system, the fraction of recruits from other healthcare systems versus professionals that are new to their positions and key costs incurred.

As mentioned earlier, the richness of this approach relies on its ability to perform “what-if” analyses to assess the impact of potential changes to the assumptions. Examples of scenarios that could be analyzed in this context include

- Retaining health professionals and providing incentives (e.g., modifying $sn(.)Cost$) that aim at decreasing the attrition rate ($Pretire(.)$) from the profession.
- Changing assumptions regarding demand for health professionals ($n(.)min$).
- Addressing the impact that the global shortage of health professionals might have on the entire system.
- Examining changes in skill mix or role changes within a health professional group (which will also impact $n(.)min$).
- Analyzing the impact of budgetary changes.
- Changing the duration of educational programs.

These scenarios provide examples that could be addressed with a linear programming model. We have shown that it is worthwhile to extend temporary leave if that improves staff retention in the long term. Diminishing attrition rates from educational programs played a greater role in our solutions than increasing the number of years a worker stays in the position before he or she retires. Other policy implications obtained from our models are discussed in Refs [34, 36, 37]. The model itself can be adapted to be used nationally, regionally, or at the institutional level. The inputs and outputs must then be adjusted accordingly.

Modeling Considerations

When designing an LP for HHR planning, there are various issues that must be considered. The planner must first decide the level of detail to be modeled: institutional, provincial, national, or for a subspecialty. The level depends on the questions to be answered. For instance, the decision maker might wish to determine how many students to admit each year to educational programs, whether major recruitment policies of professionals from other regions are necessary, and how to meet the need for upper managerial positions on a regional or national level. Such a framework is more strategic than a model designed to aid in operational decisions (such as hiring, training, and scheduling resources) within a hospital or other health organization.

Furthermore, the planning must be done over the long run to ensure that the decisions made are not only appropriate now but also consider future implications. However, deciding over how many periods to solve the problem is another key issue that needs to be considered. A period is the time interval between decision epochs. Including too many periods makes the model more complex (in terms of the data needed as well as the size of the model), while incorporating too few periods might not represent the problem appropriately, and it might not consider the long-term implications of today's decisions. Under certain conditions, a look-ahead policy (obtained by recursively solving the finite-horizon problem) has been proved to be optimal [35]. We suggest using this type of models on a rolling horizon basis; that is using the solutions for the first periods and then solving the model again as new (or more accurate) data become available. Using sensitivity analysis, it is possible to determine how much the solution for the first periods would change given expected changes in future data.

As the availability of health professionals might be affected by age in terms of attrition or full-time equivalency, the model should track not only the number of professionals each year but also the number in each age range. How wide each age range should be will depend on the level of detail needed to make appropriate decisions and the data

availability. Note that not all age ranges need to be the same size.

In the case that there are multiple ways in which all constraints could be satisfied, the objective function can be used to ensure that no more professionals are educated, promoted, and recruited than are needed in the long run. As described in the previous sections, this is accomplished by setting the objective to minimize the total cost incurred. However, other objective functions may also be used based on the outcomes sought by the user.

The linear programming approach described in this article does not come without limitations. While we have discussed the importance of choosing the length of the planning horizon, as results near the end of the horizon may suffer from end of horizon effects, other aspects need to be considered. For instance, this model is designed to answer aggregate level decisions and hence is based on parameter averages. Greater level of detail may be obtained by considering the distributions of the parameters; stochastic optimization and simulation provide two approaches to do this. In addition, we forecasted demand by multiplying population to nurse ratios by population projections prepared by government agencies. We investigated the sensitivity of the demand forecast by choosing current ratios, the lowest Canada wide ratios, and the most desirable targets based on best practice. The decision maker may wish to consider other needs based approaches to estimate demand. Even with those limitations, our linear programming approach provides useful policy insights and suggests where further effort may be placed to collect more detailed data to address the problem at hand.

CONCLUSIONS

We have presented an introduction to the use of linear programming in strategic HHR planning. Moving beyond static, one-time model approaches to ones that incorporate flexible scenarios provides another tool for decision makers to examine HHR decisions. Recognizing the influence of policy decisions

made in one sector on another and being able to simultaneously model changes can support more evidence-informed HHR planning. The decision maker must be aware of the assumptions made in the model formulation. However, the strength of a linear programming model relies on its ability to perform sensitivity analyses. We believe that a linear programming approach can be very useful to decision makers faced with HHR planning problems and that formulating the problem in such a systematic way might be the key to gaining the necessary insights to make such important decisions. The reader is however cautioned that no one approach should be used in isolation when making HHR planning decisions.

REFERENCES

1. Tomblin Murphy G, O'Brien-Pallas L. How do health human resources policies and practices inhibit change? A plan for the future. In: Forrest P-G, Marchildon GP, McIntosh T, editors. Volume 2, Changing health care in Canada, The Romanow Papers. Toronto: University of Toronto Press; 2004.
2. Kazanjian A, Hebert M, Wood L, *et al.* Regional health human resources planning and management: policies, issues and information requirements. Centre for Health Services and Policy Research: Vancouver; 1999.
3. Association of American Medical Colleges Center for Workforce Studies. The impact of health care reform on the future supply and demand for physicians updated projections through 2025. Washington, DC: Association for American Medical Colleges; 2010.
4. Barer M, Stoddart GL. Toward integrated medical resource policies for Canada. Federal/Provincial/Territorial Conference of Deputy Ministers of Health: Ottawa; 1991.
5. Tomblin Murphy G. Methodological issues in health human resource planning: cataloguing assumptions and controlling for variables in needs-based modeling. *Can J Nurs Res* 2002;33(4):51–70.
6. O'Brien-Pallas L, Baumann A, Donner G, *et al.* Forecasting models for human resources in health care. *J Adv Nurs* 2001;33(1):120–129.
7. Vestimetra International Inc, Groupe de Recherche Interdisciplinaire en Santé Université de Montréal and Health Care and Epidemiology/Medicine University of British

- Columbia. A Pan-Canadian inventory assessment and gap analysis; 2005. Health Canada, Ottawa.
8. Kazanjian A, Brothers K, Wong G. Modeling the supply of nurse labor. Life-cycle activity patterns of registered nurses in one Canadian delivery system. *Med Care* 1986;24(12):1067–1083.
 9. Kazanjian A, Pulcins IR, Kerluke K. A Human Resources Decision Support Model: Nurse Deployment Patterns in One Canadian System. *Hosp Health Serv Adm* 1992;37(3):303–319.
 10. Newton S, Buske L. Physician resource evaluation template: a model for estimating future supply in Canada. *Annals RCPSC* 1998;31(3):145–150.
 11. Abrahams C. In search of the “Right” Ontario postgraduate allocation: the ADIN model for HHR planning. Ontario: Ministry of Health and Long-Term Care; 2005.
 12. Archambault R. New COPS occupational projection methodology; 2000. Applied Research Branch, Human Resources Development Canada, Quebec.
 13. Basu K, Liu J. Registered nurse demand model. HHR modelling working group workshop. Microsimulation modelling and data analysis division health Canada; 2005.
 14. Tomblin Murphy G. Health human resource planning: an examination of relationships among nursing service utilization, and estimate of population health, and overall health outcomes in the province of Ontario [Doctoral dissertation]. University of Toronto; 2005.
 15. Dickson GL. 2002. Forecasting the demand for nurses in New Jersey. Available at <http://www.njccn.org/>. Accessed 2015 Apr 08.
 16. Kephart G, Maaten S, O'Brien-Pallas L, *et al*. Simulation analysis report. Ottawa: Building the Future: An Integrated Strategy for Nursing Human Resources in Canada; 2004.
 17. O'Brien-Pallas L, Duffield C, Tomblin Murphy G, *et al*. Nursing workforce planning: mapping the policy trail. Geneva, Switzerland: International Council of Nurses; 2005.
 18. Law AM, Kelton WD. Simulation modeling and analysis. 3rd ed. New York: McGraw-Hill; 2000.
 19. Martel A, Price W. Stochastic programming applied to human resource planning. *J Oper Res Soc* 1981;32(3):187–196.
 20. Gans N, Zhou Y-P. Managing learning and turnover in employee staffing. *Oper Res* 2002;50(6):991–1006.
 21. Zanakis SH, Maret MW. A Markov chain application to manpower supply planning. *J Oper Res Soc* 1980;31(12):1095–1102.
 22. Anderson EG, Jr. The nonstationary staff-planning problem with business cycle and learning effects. *Manag Sci* 2001;47(6):817–832.
 23. Fry MJ, Magazine MJ, Rao US. Fire-fighter staffing including temporary absences and wastage. *Oper Res* 2006;54(2):353–365.
 24. Davis A, Mehrotra S, Holl J, Daskin MS. Nurse staffing under demand uncertainty to reduce costs and enhance patient safety. *Asia Pac J Oper Res* 2014;31(01):1450005-1–1450005-19.
 25. Véricourt F, Jennings OB. Nurse staffing in medical units: a queueing perspective. *Oper Res* 2011;59(6):1320–1331.
 26. Yankovic N, Green LV. Identifying good nursing levels: a queueing approach. *Oper Res* 2011;59(4):942–955.
 27. Chase V, Cohn AM, Peterson T, Lavieri MS. Predicting emergency department volume using forecasting methods to create a “Surge Response” for non-crisis events. *Acad Emerg Med* 2012;19(5):569–576.
 28. Bard JF, Wan L. Workforce design with movement restrictions between workstation groups. *Manuf Serv Oper Manag* 2008;10:24–42.
 29. Werker GR, Puterman ML. Strategic workforce planning in healthcare under uncertainty. Article under review.
 30. Vanderby SA, Carter MW, Latham T, *et al*. Modeling the future of the Canadian cardiac surgery workforce using system dynamics. *J Oper Res Soc* 2013;65(9):1325–1335.
 31. Balinsky W, Reisman A. Some manpower planning models based on levels of educational attainment. *Manag Sci* 1972;18(12):691–705.
 32. Rao PP. A dynamic programming approach to determine optimal manpower recruitment policies. *J Oper Res Soc* 1990;41(10):983–988.
 33. Lavieri MS, Puterman ML. Optimizing nursing human resource planning in British Columbia. *Health Care Manag Sci* 2009;12:119–128.
 34. Lavieri MS, Regan S, Puterman ML, Ratner PA. Using operations research to plan the British Columbia registered nurses’ workforce. *Health Care Policy* 2008;4(2):e117–e135.
 35. Hu W, Lavieri MS, Toriello A, Liu X. Strategic health workforce planning. Article under review.

36. Schell GJ, Lavieri MS, Li X, Toriello A, Martyn K, Freed G. Strategic modeling of the pediatric nurse practitioner workforce. *Pediatrics* 2015;135(2):298–306.
37. Schell GJ, Lavieri MS, Jankovic F, Li X, Toriello A, Martyn K, Freed G. Strategic modeling of the neonatal nurse practitioner workforce. (Article under review).
38. Hillier FS, Lieberman GJ. Introduction to operations research. 7th ed. Boston, MA: McGraw-Hill; 2001.
39. Frontline Systems, Inc. Optimization software for Excel, .NET, Java, MATLAB. Incline Village, USA; 2007. Available at <http://www.solver.com/>. Accessed 2015 April 29.

INVENTORY INACCURACIES IN SUPPLY CHAINS: HOW CAN RFID IMPROVE THE PERFORMANCE?

YACINE REKIK

Finance and Control Department,
Emlyon Business School, Ecully,
France

INTRODUCTION

Many research workers have alluded to the tremendous payoffs associated with an effective management of operations [1,2]. Companies adopting innovative supply chain management solutions that enhance the value added to customers at a lower cost are improving their competitive advantage. Nowadays, radio-frequency identification (RFID) technology, which is one type of automatic identification and data capture (AIDC) systems currently available, is becoming the basis for such new solutions contributing to a better management of supply chains, in terms of cost reduction and improvement of the customer service level. In most industries, AIDC systems originated with the use of barcode readers, barcode labels, and the universal product code (UPC). In a bar code system, in order to be detected and identified by readers, labels must be positioned precisely. This characteristic, called *line-of-sight positioning requirement*, necessitates human intervention for scanning products and leaves room for errors and inefficiencies. The problem is intensified by the fact that companies rarely have common product codes for specific products or parts [3]. Also, the increasing quantity and variety of information to be managed make the definition of standard transactions for exchanging the gathered data difficult [4]. There is therefore a need for the development of new data capture and product identification systems. Our work is motivated by the recent development of the

RFID technology, which is a system that can potentially perform even better than the bar code system and replace it in the long term [5]. This technology is basically based on the use of wireless RFID tags that carry unique electronic product codes (EPCs) and a set of software components that permit the exchange of data in a supply chain.

Several investigations have attempted to evaluate the benefits of RFID. Most of case-study-type research considers the evaluation of hard (direct) benefits [6]. However, these modeling approaches may not be sufficient for the quantification of some other types of benefits, such as the benefit of having more accurate inventory records.

Atali *et al.* [7] characterize three different kinds of demand streams that result in inventory discrepancy. Some demand streams result in permanent inventory shrinkage (such as theft and damage). Some demand streams are temporary and can be recovered by physical inventory audit and returned to inventory (such as misplacement). The final group of demand stream (such as scanning error) affects only the inventory record and leaves the physical inventory unchanged. A detailed analysis of factors generating errors will be provided later.

The aim of this article is to provide a comprehensive study on the inventory inaccuracy issue and to show the way the RFID technology can improve the performance of supply chains subject to these inaccuracies. The remaining part of this section will be concerned with the main sources causing inventory inaccuracy: we focus on the impact on the performances and briefly detail main investigations performed for each type of source. Then, in the section titled "How to Cope with Inventory Inaccuracy," we provide the different ways that permit us to cope with the inventory inaccuracy issue. For this purpose, we show that RFID technology could be an interesting alternative as proven by many recent investigations. The aim of the section titled "A Framework Enabling Modeling Inaccuracies and the Impact of RFID"

is to present a framework that allows the modeling of an inventory system subject to inaccuracies and to show how we can measure the impact of RFID on such system. Finally, the last section ends the article.

The Inventory Inaccuracy Issue

The standard literature on inventory models has rarely differentiated between the inventory record and the physical inventory. The two have always been considered to be the same, and the main concern was on how, having observed demand and the resulting inventory levels, an inventory manager should determine when and how much to replenish. On the basis of empirical observations, this implicit assumption has been proven wrong. In fact, based on a study done with a leading retailer, Raman [8] reports that out of close to 370,000 Stock-Keeping Unit(SKU) investigated, more than 65% of the inventory records did not match the physical inventory at the store-SKU level. Moreover, 20% of the inventory records differed from the physical stock by six or more items.

A general definition of accuracy includes obtaining the correct value for a measurement at the correct time [9]. According to Iglehart and Morey [10,11], inventory inaccuracy occurs when the system inventory, that is, what, according to the information system (IS), is available, does not match the physical inventory, that is, what is actually available. Various other definitions going in the same sense and measures of inventory accuracy are presented in Refs 12–17. For example, Ernst *et al.* [12] propose using a control chart to monitor the changes in the inventory accuracy. Another definition provided by Bernard [14] considers the percentage (and not the difference) error in the inventory records. Martin and Goodrich [17] define accuracy as the total dollar deviation between the actual dollar value of the inventory and the recorded dollar value of the inventory. As a conclusion of the last provided definitions, we say that an inventory stock is inaccurate when the record stock is not in agreement with the physical stock.

Inventory inaccuracy can be a major obstacle to improvements in firms' performance [18]. While companies have undertaken large

investments to automate and improve their inventory management processes, inventory IS and physical inventory are rarely aligned [8]. Inventory inaccuracy might result from several factors. The aim of this section is to present a comprehensive analysis of the factors generating inventory inaccuracy. On the basis of empirical and qualitative investigations, we try to focus on the order of magnitude of these errors. We also provide the main quantitative investigations for each error type.

Transaction Errors. Transaction errors are unintentional errors occurring during inventory transactions. Some of these transactions happen when counting the inventory, receiving an order, or checking out at the cash register.

Errors when checking out occur if the cash register scans one item twice, rather than each item separately, when a customer is buying two similar (but not identical) items with the same price. This innocuous action by the cash register ensures that the customer pays the correct amount and may even save the customer time as the cash register avoids handling the additional item. However, it generates a discrepancy in the inventory IS. Errors when picking impact inventory records similarly. A warehouse employee can unintentionally ship the wrong quantity of a particular item to a store or even send the wrong item altogether. According to DeHoratius and Raman [11], in an apparel warehouse, it is quite easy for an employee to mistakenly pick a "medium" instead of a "large" garment. Stores typically do not scan merchandise on receipt, allowing such errors to remain invisible. In her PhD dissertation, Sahin [19] provides a comprehensive analysis of inventory systems subject to perturbations in nominal flows. According to the author, the major defects resulting in transaction errors are the following:

- the technical limitations of the bar code system;
- the potential failures stemming from the interaction between inventory operators and the bar code system. Those errors result in (i) errors made

when identifying entities, (ii) errors made when counting, and (iii) errors made when keypunching data.

Concerning the impact of transaction errors, we note that there is some empirical data available in the context of retail stores. Kang and Gershwin [20] report the results of a study conducted by a global retailer in several hundred of its stores. Inventory records were accurate for 51% of SKUs, whereas for 76% of SKUs the deviation between physical inventory and inventory records was within a range of ± 5 units. This means that for close to one in four SKUs, inventory records deviated from physical inventory by six or more units. In another empirical study conducted by DeHoratius and Raman [11], who examined inventory record accuracy at a multibillion dollar retailer, the authors found that the absolute difference between inventory records and physical inventory was on average close to five units. For 15% of SKUs, it was above eight items. This compares with an average target inventory of 14 units per SKU. Average inventory inaccuracy varied considerably by store, with a minimum of 2.4 units per SKU and a maximum of 7.9 units. The same investigation suggests that factors such as higher selling quantity, inventory density, and product variety are associated with higher levels of inventory record inaccuracy.

Among examples of studies that quantify the impact of inventory inaccuracies, the first one is performed by Iglehart and Morey [10], who propose an analytical tool to aid in controlling inventory errors, the objective being to select the type and frequency of counts and to modify the predetermined inventory policy by adding a buffer to compensate for errors so as to minimize the total cost (inventory holding + inspection costs) per unit time. Sandoh and Shimamoto [21] also builds a model that determines the optimal frequency for periodical inventory taking processes within a supermarket, which results from the trade-off between the cost for inventory-taking activities and the cost of investigating the causes of deviations. More recently, the work conducted by K  k and

Shang [18] aims at finding effective inventory replenishment and inspection policies that minimize inventory and inspection costs over a finite horizon. In their framework, the authors represent inventory inaccuracy through random errors that change the physical inventory at the end of each period and errors accumulate until a costly inspection is performed. They propose a near-optimal joint inventory inspection and replenishment heuristic.

Misplacement Errors. Misplacement errors occur when a fraction of the inventory is misplaced; it is not available to meet a customer demand until it is found. According to Chappell [22], there are several sources generating misplacement errors such as (i) consumers picking up products and then putting them down in another location, (ii) clerks not storing products on the correct shelf at the right time, and (iii) clerks losing products in the backroom. From a four-year longitudinal study of 333 stores of a large retailer, Ton and Raman [23] show that increasing product variety and inventory level per product is associated with an increase in misplaced products. The authors also show that increasing misplaced products is associated with a decrease in store sales. According to   amdereli and Swaminathan [24], misplaced inventory is also present in warehouse operations. G.T Interactive, the creator of computer games like Doom II and Driver, suffered from low productivity due to inventory misplacement in the warehouse. Fundamentally what happens in these settings is that the product is misplaced in the supply chain and is unavailable during the sales period but can be retrieved when a cleanup is performed.

Misplaced inventory can be quite large and have a significant impact on the inventory performance. It is reported in Ref. 8 that customers of a "leading retailer" cannot find 16% of the items in the store because of misplacement errors. The consequence is that misplacement errors reduce the profit by 25% for this retailer.

The authors in Ref. 25 model the consequences of random misplacement errors on the retail supply chain by comparing three approaches. In the first approach, the

retailer is unaware of the misplacement and places his orders as if inventory information is perfect. In the second case, the retailer is aware of the errors in his inventory status information, and adjusts his ordering policy accordingly. In the final approach, perfect inventory status information based on RFID technology is assumed. They also provide a critical value of the RFID cost, which makes its deployment cost effective. As an extension of the last investigation, the authors in Ref. 26 consider the impact of misplacement-type errors and the effect of the RFID technology on a decentralized supply chain. In particular, they compare two possible strategies, in which the first deals with the deployment of the RFID technology and the second considers the coordination of the channel by using a modified buy-back contract. Gaukler *et al.* [27] investigate the effects of the RFID technology within the context of the retail supply chain. They build a newsvendor model that takes into account the inefficiency of the replenishment process from the backroom to the shelf in the retail store. Then, based on this model, they examine how the cost of the RFID implementation should be shared among the supply chain actors, and determine coordinating contracts for the RFID-enabled supply chain within a newsvendor framework. Çamdereli and Swaminathan [24] study also the misplacement-type errors in a decentralized supply chain for uniform distributions of demand and error.

Damage and Spoilage. For supermarkets, perishables are the driving force behind the industry's profitability and represent a significant opportunity for improvement, accounting for up to \$200 billion in US sales a year but subjecting firms to losses of up to 15% due to damage and spoilage [28]. Examples of limited lifetime products are drugs and food products. In retail stores, customers can cause damages to products, thereby making them unavailable for sales. Some examples are, tearing a package to try on the contained cloth item, wearing down a shoe by trying it on and walking, erasing software on computers on demonstration, spilling food on clothes, and scratching a car during a test drive [29]. These damages may not be detected by the

inventory manager and as a consequence would cause inventory inaccuracy.

According to Sahin [19], items reach their limit lifetime during storage because of the following:

- Errors when forecasting the customer demand, which may lead to an overestimation of this demand and as consequence a substantial quantity of products not being sold.
- The inability to track accurately the location, condition, and the age of products stored within a facility.

An industry survey performed by the Joint Industry Unsalables Steering Committee [30] provides data on the level of unsalables in the US retail industry. According to the survey results, which are based on responses from over 60 manufacturers and retailers, the cost of unsalable food and grocery products amount to 1% of sales in the United States. Damage is the biggest cause of unsalables with 63% of all unsalables, followed by expired (16%) and discontinued items (12%). The rate of unsalable products can differ by product category: frozen products, for example, reported an unsalable rate of 0.9%, whereas the rate for refrigerated products was 1.7%. Unsalable rates for health and beauty care and general merchandise had even higher unsalable rates (1.9 and 2.2%, respectively), which was attributed to frequent new product introductions, shifts in fashion component, seasonality, and short shelf-life.

Theft. Inventory theft is defined as a combination of employee theft, shoplifting, internal and external theft, vendor fraud, and administrative error. The Efficient Consumer Response (ECR)¹ Europe project on shrinkage subsequently analyzed the causes of stock loss and proposed a systematic and collaborative approach to reducing the phenomenon throughout the supply chain. ECR defines "shrinkage" as the process errors, deceptions, and internal and external thefts. According

¹<http://www.ecr.org/>

to Beck [31], some specific types of internal theft include the following:

- staff stealing goods by either hiding them in their bags or intentionally placing them outside the building for later collection;
- collusion occurring when a staff member collaborates with a customer to steal products. During such an incident, the staff member may not scan the item, or the security personnel may intentionally ignore the offense as it occurs;
- grazing occurring when items stored in the warehouse are consumed by the warehouse staff.

The results of the research carried by ECR Europe have shown that the scale of shrinkage in the fast-moving consumer goods sector is estimated to be €24 billion in 2003 (465 million euros is lost irreparably within the fast-moving consumer goods turnover weekly), which is 2.41% of the whole turnover value of the sector. The process errors present 27% of the whole shrinkage value, 7% deceptions, 28% internal thefts, and 38% external thefts.

On the basis of survey data, internal and external theft, administrative errors, and vendor fraud accounted for an estimated 1.8% of sales in the US retail industry in 2001, costing US retailers \$33 billion [32]. For US supermarkets, the National Supermarket Research Group (NSRG) survey [33] estimates that internal and external theft, receiving errors, damage, accounting errors, and retail pricing errors amount to 2.3% of sales. These figures only take into account the item value, but not any process-related costs (e.g., for handling of damaged items). Kang and Gershwin [20] use simulation to analyze the consequence of inventory inaccuracy. They show that even small undetected losses can lead to important stockouts. They also propose several ways to tackle this problem, including the deployment of RFID technology. The authors in Ref. 34 use also simulation to show the consequences of inventory inaccuracy in a three-stage supply chain. Considering misplaced-type errors, the authors in Ref. 25 show analytically the impact of errors on the inventory decision

strategies and derive a threshold cost value at which RFID deployment would become cost effective. Extending the model of Rekik *et al.* [25,26] consider the case of a decentralized supply chain with one manufacturer and one retailer whose inventory is subject to inaccuracies. They develop the optimal ordering strategies under the centralized and the decentralized scenarios and study the impact of the RFID deployment on such supply chains. Heese [35] studies a supply chain consisting of a Stackelberg leader manufacturer and a retailer with inventory inaccuracy and random demand. Using specific distribution functions, he derives the threshold value of tag cost in order for a firm to adopt RFID. More recently, Rekik *et al.* [36] consider a finite horizon, single-product periodic review store inventory in which inventory records are inaccurate due to theft errors. They analyze the problem of theft in the store by optimizing the holding costs under a service level constraint, and analyze the impact of theft errors and the value of the RFID technology on the inventory system.

Supply Errors: Product Quality, Yield, and Supply Process. When the product quality is low, or a production process has a low yield, or a supply process is unreliable, the physical inventory may not be known and as a consequence may be different from the inventory in the IS [37–39]. Products that are not conforming to quality standards can also make the inventory inaccurate. According to Bensoussan [29], receipts are usually added to the inventory without a full inspection process. The consequence is that the IS may consist of both nondefective products and defective products which are not available for sale.

HOW TO COPE WITH INVENTORY INACCURACY

To cope with inventory inaccuracy, different compensation methods can be used. The analysis conducted by many investigations [19,20,40,41] agreed that (i) making decisions by considering the inventory inaccuracy may be applied to tackle the problem of inventory inaccuracy and (ii) RFID technology

may help to track items through the supply chain. In addition to reducing the inventory inaccuracy, this technology may also help to eliminate the reasons of inventory inaccuracy such as theft.

For the specific case of misplacement errors, inventory managers may apply other methods to tackle the problem of misplaced inventory. For example, in some apparel departments of retailers, there are signs informing customers not to return the product to their original location if they decide not to buy the product after trying it. Some libraries cope with the misplacement problem by putting signs that tell the customers not to reshelve the books after use, while others have designated spaces for returning books taken off the shelf [24].

The analysis conducted by Kang and Gershwin [20] examines various other techniques that inventory managers can use to compensate for the inventory record errors. According to the authors, the compensation techniques can be summarized in the following three points:

- *Verifying Manually the Inventory.* The inventory manager can choose the items in the facility in order to perform a manual, periodic counting. The frequency of counting may depend on various elements, such as the availability of labor and the product characteristics, including the profit margin, sales velocity, and whether the products are highly prone to errors.
- *Performing a Manual Reset of the Inventory Record.* This technique is used if a direct measurement of the physical inventory is not available, and inventory managers can gather and monitor the available data and search for any patterns that may be indicative of the presence of serious inventory error.
- *Performing a Constant Decrement of the Inventory Record.* This technique is performed if the inventory manager is aware of the presence of errors and also knows their stochastic distributions. The inventory manager may decrement the inventory record by the average of the error in each period.

Since the physical value of the error realization at each period is unknown, performing this constant decrement will still not eliminate the error in the inventory record. However, over time, this corrective action can be expected to perform better than leaving the inventory record unadjusted.

In her PhD dissertation, Sahin [19] proposes a set of actions aiming to eliminate errors. These actions can be summarized in the following points:

- reengineering the physical organization of the facility;
- using a new product identification technology that reduces scanning errors;
- using a technology that enables the reduction of theft in the facility;
- using a technology to accurately track the products' sell-by dates;
- performing double receiving and shipment processes;
- improving the actual processes in the facility;
- performing benchmarking analysis and developing personnel awareness building actions that focus on the operational weaknesses.

DeHoratius *et al.* [41] suggest two more ways in which an inventory manager may respond to the inventory inaccuracy problem:

- *Prevention.* Reduce or eliminate the root causes of inventory record inaccuracy through the implementation and execution of process improvement.
- *Correction.* Identify and correct existing inventory record discrepancies through auditing policies.

A FRAMEWORK ENABLING TO MODEL INACCURACIES AND THE IMPACT OF RFID

How to Model Inventory Inaccuracies

A typical supply chain may include three main entities: the manufacturer, the warehouse, and retailers. The manufacturer

produces products, the retailers are the agents selling products to the final customers. Between the manufacturer and the retailers, there may exist a warehouse that acts as an intermediate element that buys products from the manufacturer and resells them to the retailers.

From an inventory control point a view, satisfying customer demands may be different within the retail store and the warehouse. For this reason, two structures may exist:

- *Structure A.* This structure focuses on the retail store context (Fig. 1). The end customers are physically present in the retail store and their demands are handled by the physical on-shelf inventory. The IS does not play a major role in this structure. This structure of the supply chain can be modeled as illustrated in Fig. 2.
- *Structure B.* This structure focuses on the warehouse context (Fig. 3). Customers are not physically present in

the warehouse, and demand satisfaction is met on the basis of the inventory level shown in the IS. On the basis of the demand it receives, the warehouse observes its IS stock and makes a commitment. The commitment is later satisfied depending on the physical available stock. The structure of this supply chain can be modeled as illustrated in Fig. 4.

In the presence of inaccuracies, the two structures do not behave in the same way from an inventory control point of view. We argue that the main difference between the two structures concerns the commitment made by the warehouse under structure B. In structure A, no commitment is made by the retailer since the end customers are physically present in the store.

Remark 1. Note that Internet retailer inventory system can be modeled as structure B since decisions concerning demand satisfaction are based on the IS inventory.

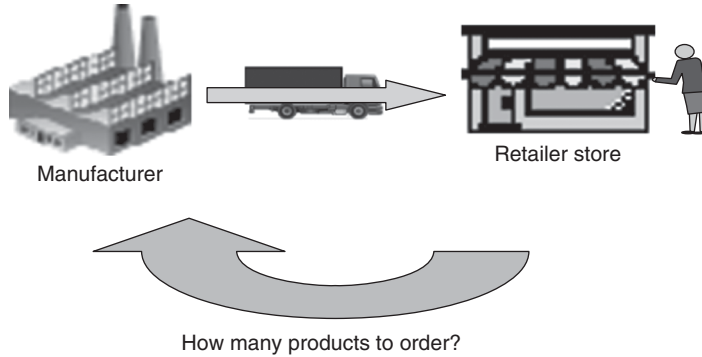


Figure 1. Structure A: the retailer supply chain.

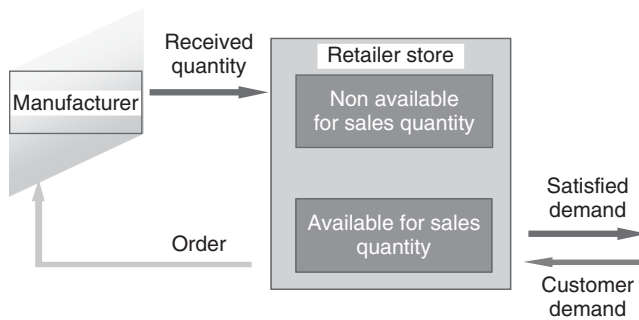


Figure 2. Modeling of structure A.

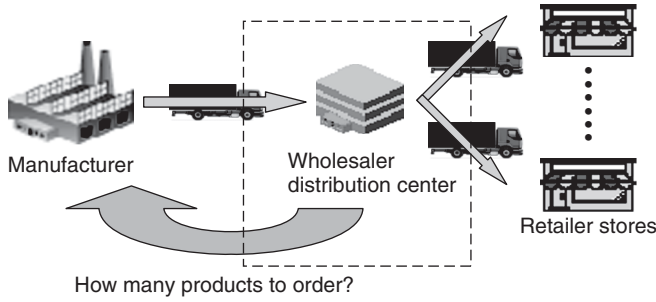


Figure 3. Supply chain B: the warehouse supply chain.

Remark 2. From an inventory control point of view, structure A can be considered as a particular case of structure B, where the available-for-sale quantity in structure A corresponds to Q_{IS} of structure B, assuming that $Q_{IS} \leq Q_{PH}$.

Under structure B, one can consider the following sequence of events: Before the start of the selling season, the warehouse manager orders a quantity from the manufacturer. This quantity is established based on forecast information available to the inventory manager regarding the future demand. The warehouse receives the goods and stores them. Just before the start of the selling season, the inventory manager receives orders from the customers. He compares the cumulative order from all the customers to the quantity seen in the IS. If the cumulative order is less than the IS quantity Q_{IS} , he accepts all the orders. Otherwise, he only accepts orders summing up to Q_{IS} . Later on, the products are shipped from the warehouse and delivered to the customers. All the orders that the inventory manager has committed himself to

should be satisfied, except in the case where the physical inventory is not able to satisfy the committed quantity. Unsatisfied demand during the commitment and unsatisfied commitment are lost since there is no opportunity for replenishment during the selling season.

The inventory manager's decision is to determine the best quantity to order from the supply system before the selling season to satisfy customer's aggregate demand. He faces three risks: (i) risk of having unsold products at the end of the selling season; (ii) risk of shortage; and (iii) risk of not being able to deliver the quantity that he has made a commitment for. Within this framework, the inventory manager faces three types of costs:

- the unit overage cost due to products unsold at the end of the selling season;
- the first type of unit underage cost due to orders rejected by the inventory manager during the commitment;
- the second type of unit underage cost due to orders initially accepted by

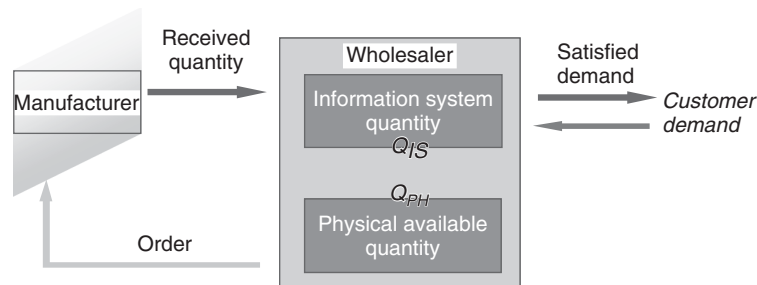


Figure 4. Modeling of supply chain B.

the inventory manager but finally not totally delivered to the customers.

We note that the second type underage cost is the parameter that characterizes the warehouse context. In a retail context, this cost does not exist since the customers are physically present at the store: if the product is not available, the demand is lost, otherwise it is satisfied immediately. The second type underage cost occurs when the customer demand is related to the IS inventory.

To model the impact of inaccuracy errors on the performance of the inventory system, it is important to characterize the error type. In fact, in a general setting, if we let Q the quantity ordered from the supply process, the physical and the IS inventory can respectively be written as follows: $Q_{PH} = \gamma_{PH}Q + \epsilon_{PH}$ and $Q_{IS} = \gamma_{IS}Q + \epsilon_{IS}$, where the couple of random variables $(\gamma_{PH}, \epsilon_{PH})$ $(\gamma_{IS}, \epsilon_{IS})$ characterizes errors on the physical inventory level (IS inventory level). From this general setting, which is called the *mixed error setting*, one can distinguish two particular cases:

- the additive error setting: in this case, $Q_i = Q + \epsilon_i$ where $i \in [PH, IS]$;
- the multiplicative error setting: in this case, $Q_i = \gamma_i Q$ where $i \in [PH, IS]$.

Second, it is also important to discuss how an inventory system subject to inaccuracy problems can be managed. In the presence of errors, one should distinguish between two situations depending on whether or not the inventory manager is aware of the existence of errors. The case where the inventory manager is unaware or simply ignores errors will be referred to as *Situation 1* in the rest of this article. The case where the inventory manager is aware of errors occurring in the inventory system will be referred to as *Situation 2*. For this latter Situation, one can also distinguish between two cases based on the information the inventory manager has about the error parameters. The first case occurs when the inventory manager has statistical information about the error parameter (such as a mean or the distribution of the error). The second case occurs when he has exact

information on the realization of the error. The difference between the two cases makes sense especially in a multi-period framework.

Concerning Situation 2, we notice that an estimation of the error parameters can be realized on the basis of statistical sampling methods as reported by Pergamalis [42], who proposes a methodology for measuring the store's inventory accuracy.

In contrast to Situation 2, Situation 1 is easier to model and optimize. In fact, in Situation 1, the inventory manager acts as if there were no errors. So, his ordering decisions or his replenishment policy coincides simply with the ordering quantity or the replenishment policy of the model without errors.

RFID as a Lever to Tackle the Inaccuracy Problem

RFID technology is a wireless technology that includes RFID tags and readers often linked to an IS through a computer. A tag is basically an antenna, which may be active or passive. An active tag is powered by a battery, while a passive one is not. This affects the performance and price. Because of its characteristics, RFID tags may be read without visual or physical contact. To facilitate physical distribution and inventory management, tags are attached to packaging or unit loads, for instance, transport packaging, pallets, or containers.

RFID technology is developing at a rapid pace, offering companies the opportunity to improve supply chain visibility. In fact, the number of RFID applications in supply chains is increasing steadily, which is indicated by the escalating level of investment in this technology. The Aberdeen Group [43] surveyed over 600 companies using RFID in the last 12 months. According to this study, the average manufacturer's annual RFID budget has grown rapidly, from \$50,000–75,000 in 2006 to \$100,000–\$200,000 in 2007. The study shows that the investments already made greatly improved cycle time, on-time delivery, safety stock, and changeover time.

Several authors highlight the opportunities presented by RFID technology. In the physical distribution field, Jim Wu and

Chen [44] use a simulation study to predict operational efficiency in a distribution center. Thanks to a perfect RFID system (100% read rate), the authors show an inventory reduction of 15%, lowered space use rate of 15%, and elimination of out-of-stock items. The authors in Holmqvist and Stefansson [45] study a mobile RFID system designed to manage arrival inspection and loading of containers. They find both time savings in arrival inspection and lead time reduction in loading, except when there are only a small number of consignments per container. In another study, Jarugumilli and Grasman [46] use invented data to model a vendor-managed inventory (VMI): for an excellent review about VMI systems, the reader is referred to the article titled **Vendor-Managed Inventory** in this encyclopedia (by Michael J. Fry) setting with RFID tags applied to the goods. In Ref. 47, the authors propose a logistics IS to simplify the distribution process with the help of RFID technology. A case study is discussed in applying the proposed IS to solve distribution management problems of a medium-sized logistics service company. By using this RFID-enabled system, the authors show that the overall distribution performance of the company is greatly improved.

On the inventory control side, as stated in the section titled “Introduction,” many recent investigations concerned with inventory inaccuracy issues are motivated by RFID technology and its impact on supply chain performance. Indeed, RFID deployment contributes to the elimination of the sources of inaccuracies through the supply chain:

- Errors resulting from the unreliability of the supply system can be detected thanks to the counting performed when receiving orders. RFID can reduce the errors in receiving via the RFID-enabled process referred to as *electronic proof-of-delivery* (or ePod) [48]. With barcode, mistakes are made by misidentifying the quantity and type of product.
- For food and perishable products, RFID can
 - Facilitate first-in-first-out (FIFO) and pricing management
 - Facilitate the tracking and maintenance of temperature conditions supporting the integrity of the cold chain for some food products
 - Allow getting more accurate information about demand and supply forecasts. Accurate and timely information about stock levels, sales throughput, and production data allow reengineering the overall business. Minimizing buffer and obsolescent stock will lead to lower inventory. Products can be brought closer to the market in the form of mini stocks, once the visibility over those can be guaranteed. Combining product flows from different sources on their way to the receiver becomes faster and more efficient. Overall, this will lead to increased product availability and as such increased sales, which can be even more accelerated by retail applications like automatic cross-selling or product information displays.
- Inaccuracies resulting from shrinkage errors can be reduced by discouraging thieves and by facilitating the location of quantities and timings of theft losses across the supply chain. Loss prevention during transport between sending and receiving parties or even within sites can be achieved by increased transparency and real-time visibility over the goods flow. Basically, the level of tagging (pallet, case, or item) and the frequency of read events determine the degree of loss prevention.
- Products subject to misplacement errors can be detected by deploying RFID readers within shelves and in the backroom.
- Shipping accuracy can be improved. Distribution centers and manufacturers often make mistakes by loading product on the wrong truck. With RFID, the system can send an alert (visible, audible, etc.) to the person loading the truck that a mistake has been made. This alert could save a company money

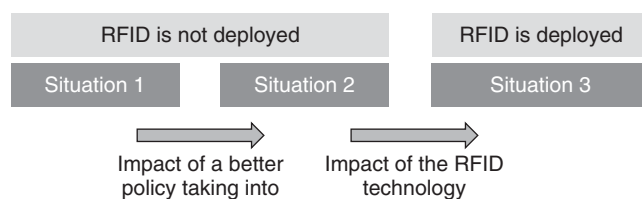


Figure 5. Situations permitting management of the system.

from shipping and reshipping the same items as well as enhance inventory control.

The decision to invest in RFID technology will depend on the trade-off between the cost associated with the deployment of the technology and the savings resulting from investment. Taking into account the variable² costs associated with RFID, that is, the RFID tag cost embedded in each product, we are able to model the RFID-enabled situation. We note that RFID can bring two benefits for the inventory warehouse as follows:

- The visibility provided by this technology. Theoretically, RFID enables tracking and tracing of items in stock and in the pipeline, thereby creating complete inventory visibility, leading to an accurate account of inventory discrepancy. In such cases, the inventory problem under study is nothing but the basic random yield problem.
- The prevention or reduction of the magnitude of some sources of inventory inaccuracy by RFID. In the case of theft errors, by being able to distinguish customer demand and other kinds of demand (theft for example), the inventory manager can act to prevent or discourage the sources of this latter demand.

The RFID-enabled problem, referred to as *Situation 3*, is also a basic inventory problem where the unit purchase (production) cost of the product includes the RFID tag cost.

To summarize, the Fig. 5 presents the different situation that allow modeling of the

impact of inventory inaccuracies and the benefit pertaining to the deployment of the RFID technology.

Comparing Situations 1 and 2 permits getting insights on the benefit of obtaining information about the errors and establishing a better inventory policy taking into account the inaccuracy issue. Comparing Situations 2 and 3 also permits getting insights into the real benefit of RFID and, in particular, answering the key questions:

- Is RFID deployment an economically feasible solution for the warehouse manager?
- If yes, under which tag price is RFID cost effective?

CONCLUSION

Our major focus in this article was to provide a comprehensive study on inventory inaccuracy and the way RFID technology can tackle this issue. We listed the main sources of errors with a focus on their impacts on the supply chain performances and briefly detailed the main research investigations in this field. We proposed a conceptual framework that permits modeling of the inaccuracy errors and the way RFID can contribute. In fact, three modeling situations show the impact of a better inventory policy, taking into account error distribution and the impact of RFID deployment.

REFERENCES

1. Quinn F. The payoff potential in supply chain management in achieving supply chain excellence through technology. Technical Report. San Francisco (CA): Montgomery Research; 1999.

²The introduction of fixed cost can be done by an additional return on investment (ROI) analysis.

2. Mentzer JT. Supply chain management. Thousand Oaks (CA): Sage; 2001.
3. Wilder C. Purchasing analyzer. 1999.
4. Kaörkkaöinen M, Ala-Risku T, Fraömling K. Integrating material and information flows using a distributed peer-to-peer information system. In Collaborative Systems for Production Management. (jagdev H.S., Wortmann J.C., Pels H.J.Eds.), Kluwer Academic Publishers, Boston: 2003. pp. 305–319.
5. Sarma S, Brock DL, Ashton K. The networked physical world. report reference MIT-AUTOID-WH-001. Auto-ID Center; 2000.
6. Alexander K, Gilliam T, Gramling K, *et al.* Focus on the supply chain: applying auto-id within the distribution center. Technical report. MIT, Cambridge: MIT: Auto-ID Center; 2003.
7. Atali A, Lee H, Ozer O. If the inventory manager knew: Value of RFID under imperfect inventory information. Technical report. Stanford: Graduate School of Business, Stanford University; 2009.
8. Raman A, DeHoratius N, Ton Z. Execution: the missing link in retail operations. *Calif Manag Rev* 2001;43:136–152.
9. Schuster EW, Scharfeld TA, Kar P, *et al.* The prospects for improving erp data quality using auto-id. *Cutter IT Journal: Information Technology and the Pursuit of Quality*, 2004.
10. Iglehart D, Morey RC. Inventory systems with imperfect asset information. *Manag Sci* 1972;18(8):388–394.
11. DeHoratius N, Raman A. Inventory record inaccuracy: an empirical analysis. Technical report. Graduate School of Business, University of Chicago; 2004. Working paper, Graduate School of Business, University of Chicago.
12. Ernst R, Guerrero JL, Roshwall A. A quality control approach for monitoring inventory stock levels. *J Oper Res Soc* 1984;44: 1115–1127.
13. Buker DW. Inventory accuracy. Chicago, illinois: Inventory Management Reprints APICS; 1984. pp. 221–223.
14. Bernard A. Cycle counting: the missing link. *Prod Invent Manag* 1985;25(4):27–40.
15. Chopra V. How to set up an effective inventory accuracy program. In: 29th Annual APICS International Conference Proceedings. Chicago, illinois: 1986.
16. Young JB. The limits of cycle counting. In: 29th Annual APICS International Conference Proceedings. Chicago, illinois: 1986.
17. Martin W, Goodrich S. Minimizing sample size for given accuracy in cycle counting. *Prod Invent Manag* 1987;28(4):24–27.
18. Kök AG, Shang KH. Replenishment and inspection policies for systems with inventory record inaccuracy. *Manuf Serv Oper Manag* 2007;9:185–205.
19. Sahin E. A qualitative and quantitative analysis of the impact of Auto ID technology on the performance of supply chains [PhD thesis]. Ecole Centrale Paris; 2004.
20. Kang Y, Gershwin SB. Information inaccuracy in inventory systems - stock loss and stockout. *IEE Trans* 2004;37:843–859.
21. Sandoh H, Shimamoto H. A theoretical study on optimal inventory taking frequency for retailing. *J Retail Consum Serv* 2001;8: 47–52.
22. Chappell G, Durdan D, Gilbert G, *et al.* Auto-id in the box: the value of auto-id technology in retail stores. Auto-ID Center; 2003.
23. Ton Z, Raman A. The effect of product variety and inventory levels on misplaced products at retail stores: a longitudinal study. Technical report. Boston: Harvard Business School; 2004.
24. Çamdereli AZ, Swaminathan JM. Coordination of a supply chain under misplaced inventory. Technical report. North Carolina: Kenan-Flagler Business School; 2009.
25. Rekik Y, Sahin E, Dallery Y. Analysis of the impact of the rfid technology on reducing product misplacement errors at retail stores. *Int J Prod Econ* 2008;112:264–278.
26. Rekik Y, Jemai Z, Sahin E, *et al.* Inventory inaccuracy in the retail supply chain: coordination versus rfid technology. *OR Spectrum* 2007;29:597–626.
27. Gaukler G, Seifert RW, Hausman WH. Item-level rfid in the retail supply chain. *Prod Oper Manage* 2007;16:65–76. Working Paper, Stanford University, Stanford, USA.
28. Ketzenberg ME, Ferguson M. Managing slow moving perishables in the grocery industry. Technical report. College of Business Colorado State University and The College of Management Georgia Institute of Technology; 2006.
29. Bensoussan A, Cakanyildirim M, Sethi SP. Partially observed inventory systems: the case of zero balance walk. Technical report. Richardson: School of Management, University of Texas at Dallas; 2005.
30. Lightburn A. Unsaleables Benchmark Report. Virginia: Joint Industry Unsaleables Steering Committee, Food Distributors International,

- Food Marketing Institute and Grocery Manufacturers of America; 2002.
31. Beck A. How can rfid help to reduce shrinkage. Technical report. Leicester: University of Leicester; 2003.
 32. Hollinger RC, Davis JL. National retail security survey. Florida: Department of Sociology and the Center for Studies in Criminology and Law, University of Florida; 2001.
 33. Supermarket. Supermarket shrink survey. London: National Supermarket Research Group; 2001.
 34. Fleisch E, Tellkamp C. Inventory inaccuracy and supply chain performance: a simulation study of a retail supply chain. *Int J Prod Econ* 2005;95:373–385.
 35. Heese H. Inventory record inaccuracy, double marginalization and rfid adoption. *Prod Oper Manag* 2007;16:542–553.
 36. Rekik Y, Sahin E, Dallery Y. Inventory inaccuracy in retail stores due to theft: An analysis of the benefits of RFID. *Int J Prod Econ* 2009;118:189–198.
 37. Yano CA, Lee HL. Lot sizing with random yields: A review. *Oper Res* 1995;43:311–334.
 38. Inderfurth K. Analytical solution for a single-period production-inventory problem with uniformly distributed yield and demand. *Cent Eur J Oper Res* 2004;12:117–127.
 39. Rekik Y, Sahin E, Dallery Y. A comprehensive analysis of the newsvendor model with unreliable supply. *OR Spectrum* 2007;29: 207–233.
 40. Uckun C, Karaesmen F, Savas S. Investment in improved inventory accuracy in a decentralized supply chain. *Int J Prod Econ* 2008;113:546–566.
 41. DeHoratius N, Mersereau AJ, Schrage L. Retail inventory management when records are inaccurate. *Manuf Serv Oper Manag* 2008;10:257–277.
 42. Pergamalis D. Measurement and checking of the stock accuracy. 2002. Article in www.optimum.gr (www.optimum.gr/Knowledge_Center/articles/).
 43. Klein R. Can rfid deliver the goods?- the manufacturer's visibility into supply and demand. Technical report. Anderson Group; 2007.
 44. Wu Y-C, Chen J-X. RFID application in a cvs distribution centre in taiwan: a simulation study. *Int J Manuf Technol Manag* 2007;1:121–136.
 45. Holmqvist M, Stefansson G. Smart goods' and mobile rfid: a case with innovation from volvo. *J Bus Log* 2006;27:251–272.
 46. Jarugumilli S, Grasman SE. RFID-enabled inventory routing problems. *Int J Manuf Tech Manag* 2007;10:92–105.
 47. Choy KL, Henry CWL, Kwok SK, *et al.* Using radio frequency identification technology in distribution management: a case study on third-party logistics. *Int J Manuf Tech Manag* 2007;10:19–40.
 48. Mason M, Langford S, Supple J, *et al.* Electronic proof of delivery. Technical report. Brussels, Belgium: EPCglobal; 2006.

INVENTORY RECORD INACCURACY IN RETAIL SUPPLY CHAINS

NICOLE DEHORATIUS
Zaragoza Logistics Center,
University of Portland

INTRODUCTION

Inventory record inaccuracy (IRI) is a substantial problem in retail supply chains. Retailers rely on automated decision-support tools that use recorded inventory quantities to replenish store shelves, plan product assortments, and forecast demand. In the presence of IRI, these decision-support tools generate suboptimal solutions and ultimately fail to prevent the very problems they were designed to avoid, namely stockouts and excess inventory.

IRI can compromise automated decision-support tools in several ways. IRI may cause an automated replenishment system to order when ordering is unnecessary or fail to order when it should. “Freezing” of the automatic replenishment system, a scenario in which the retailer has no items on the shelf but a sufficiently positive inventory record so as not to trigger a replenishment order, results in a physical stockout [1]. This stockout persists until the retailer identifies the problem and intervenes by replenishing the inventory and correcting the inventory record. Retailers unable to meet consumer demand due to a stockout run the risk of losing not only the sale of that item but quite possibly the consumer’s entire bundle of goods and perhaps even the consumer’s lifetime purchases should the consumer defect to a competing retailer [2].

IRI may also cause an automated demand forecasting system to misestimate true demand. Take, for example, a demand forecasting system designed to forecast future demand for a particular item in the store. If the inventory record for this item

is incorrectly recorded as positive when the item is actually out of stock on the store floor, the demand forecasting system will observe no sales and may incorrectly attribute this lack of sales to a reduction in consumer demand. The demand forecasting system may lower the recommended stocking quantity for this item as a result. Not only are customers unable to find the product on the shelf today due to the stockout but this reduction in the demand forecast makes it less likely for customer demand to be met in the future. Even after this retailer restocks its shelf, the adjusted demand forecast has reduced the target stocking quantity of this item, ensuring that the impact of today’s IRI persists into the future.

In the sections that follow, we document the extent of the IRI problem in retailing and identify several solutions proposed by operations management researchers. We also describe how discrepancies between the actual and recorded inventory quantity arise within stores, highlighting the role that retail replenishment processes play in generating such discrepancies. In particular, we identify how a single process failure in the distribution center (DC) can generate an inventory discrepancy that cascades across product categories and even across stores. We conclude with a discussion on future research opportunities on this topic, which we believe merits the attention of practitioners and academics alike.

EXTENT OF THE IRI PROBLEM

Recent research finds IRI to be both pervasive and substantial in magnitude in retail supply chains. DeHoratius and Raman [3] find inaccuracies in 65% of the records supporting 37 stores of a leading retailer. Similarly, Kang and Gershwin [1] note inaccuracies in 51% of the inventory records among stores of a different retail firm. Both studies find that the magnitude of IRI differs from store to store and DeHoratius and Raman [3] attribute

these differences to store-specific characteristics such as product variety, defined as the number of unique items in a store,¹ inventory density, defined as the total inventory quantity divided by the retail selling area, and audit frequency, defined as the time between storewide inventory audits. Research conducted in contexts other than retailing finds varying levels of IRI in government supply depots [4,5], pharmaceutical warehouses [6], hospitals [7], and DCs [8,9].²

Several studies explore the financial consequences of IRI in retailing. Kang and Gershwin [1], Fleisch and Tellkamp [11], and Waller *et al.* [12] show that even small discrepancies can lead to substantial lost sales, missed service level targets, and suboptimal retail performance. DeHoratius and Raman [3] estimate the lost revenue caused by IRI to be 1.1% of sales. Raman *et al.* [13] find lost sales due to misplaced inventory reduces profits by 25%. Rather than forgo sales, retailers may choose to carry additional inventory to buffer against the added uncertainty caused by IRI. DeHoratius *et al.* [14] estimate that retailer will incur 20% more inventory holding costs, on average, in order to achieve the same service level as a retailer who attempts to mitigate the IRI problem through other means.

IRI can also negatively impact retail partners (distributors and manufacturers) and undermine the performance of collaborative planning forecasting and replenishment (CPFR) and vendor managed inventory (VMI) initiatives. Angulo *et al.* [15] find IRI significantly impacts vendor performance measures such as fill rate. Sari [16], in measuring the total supply chain inventory cost and retailer's customer service level, finds the impact of IRI on retail supply chains utilizing CPFR and VMI depends on how tightly coupled vendors and retailers are within the supply chain. Camdereli and Swaminathan [17], by deriving the

optimal inventory policy for a retailer with a known proportion of misplaced inventory, show that changes in the proportion of misplaced products impact channel partners differently, making channel coordination more of a challenge.

Several mechanisms for mitigating the negative consequences of IRI exist. In the section that follows, we provide a description of several proposed solution for IRI. We also classify each of these solutions into one of three categories (prevention, correction, or integration) identified by DeHoratius *et al.* [14]. Evidence that the problem of IRI is substantial both operationally and financially and that its impact is not isolated to single stores or even stores within a single chain but rather permeates across firm boundaries within a retail supply chain motivates much of this work.

RECOMMENDED SOLUTIONS TO IRI

IRI is not a problem new to operations researchers. Over the course of the last few decades, researchers have been offering solutions to IRI in numerous contexts. DeHoratius *et al.* [14] identify three categories of solutions typically used to address IRI problems in retailing as follows:

1. *Integration.* Use inventory planning and decision tools robust enough to account for the presence of record inaccuracy;
2. *Correction.* Identify and correct existing inventory record discrepancies through auditing policies;
3. *Prevention.* Reduce or eliminate the root causes of IRI through the implementation and execution of process improvement.

We discuss each category in turn.

Integration

Among the first researchers to address the problem of IRI, Rinehart [4] argues that discrepancies between recorded and actual inventory quantities prevent firms from taking advantage of existing inventory control

¹This study operationalizes product variety in several different ways including the number of unique merchandise categories in a given store.

²See DeHoratius and Ton [10] for a full list of IRI problems in other contexts.

levels of inventory in the presence of IRI may actually be able to learn about their true inventory position. This learning, learning that reduces the uncertainty surrounding the retailer's physical inventory position, provides a retailer with distinct advantages when it comes to making future inventory decisions.

Correction

Prevention

Gaukler *et al.* [28] examine the role RFID plays in improving product availability by reducing the likelihood of in-store, shelf-replenishment process failures. Moreover, Gaukler *et al.* [29] explore the differential benefit RFID technology has among channel members and offer suggestions for improved channel coordination in the presence of RFID. Uçkun *et al.* [30] identify the factors

driving the RFID investment decision among channel partners. These factors include fixed costs, demand variance, and the retailer's profit margin. Focusing on channel partners engaged in VMI, Szmerekovsky and Zhang [31] compare supply chain performance with and without the use of RFID to improve replenishment efficiency.

Prevention, however, requires researchers to understand the root cause of IRI. Discrepancies between recorded and actual inventory quantities arise in numerous ways as actual quantities physically move through the retail supply chain and recorded quantities are automatically updated by the retailer's inventory tracking systems. DeHoratius and Raman [3] identify several sources of IRI including replenishment errors, employee and customer theft, imperfect inventory audits, and incorrect recording of sales and returns. In the section that follows, we focus primarily on replenishment errors and provide an in-depth look at the types of problems that can arise.

RETAIL REPLENISHMENT PROCESSES

Replenishment errors, as the following example reveals, warrant additional attention. Raman *et al.* [13] describe an experimental audit whereby executives of one retail chain, unbeknownst to its store and DC managers, decide to audit a store

before it was opened for business but after it had been stocked. This audit revealed that 29% of its records were incorrect—meaning for 29% of the SKUs in this store of nearly 10,000 SKUs, the on-hand quantity did not match the actual quantity present in the store. This level of IRI could not be explained by customer theft, mis-scanning at the register, poor returns procedures, or untracked interstore transfers, activities often cited as contributing to record inaccuracy, as the store was not yet operational. Instead, this level of inaccuracy can only be attributed to errors in the delivery of product from this retailer's own DC to this store. The accumulation of errors, namely the failure of employees to correctly execute specific processes, within a retail DC can ultimately create record inaccuracies at the retail stores if such errors go unnoticed.

The replenishment of store inventory from retail DCs typically involves five main phases. Figure 1 depicts these phases beginning with the arrival of merchandise from manufacturers and distributors and concluding with the loading of trucks scheduled for store deliveries. These phases include the following:

- *Receiving.* The process of unloading in-bound trucks and verifying actual arrivals against advanced shipping notices and/or purchase orders. This is the entry point of goods from the

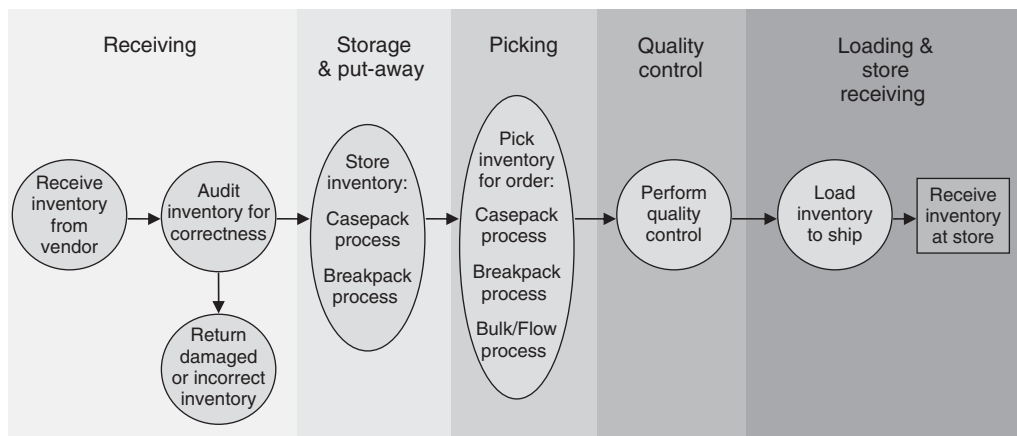


Figure 1. Retail Replenishment Process.

manufacturer or distributor into the DC.

- *Storage and Put-Away.* The process of putting the goods into the place where they will be stored until it is time for them to be picked, Casepack locations and breakpack locations are among the two most common categories of storage types. Casepack locations are for items that are shipped in casepack quantities. Breakpack locations are for items that are shipped in quantities smaller than the casepack size.
- *Picking.* The selection of goods from storage locations. The item and quantity to be picked is itemized on a packing list. The packing list identifies all those items needing to be shipped from the DC to the retail store in a specific time period.
- *Quality control.* The examination and audit of orders picked prior to store shipment.
- *Loading and Store Receiving.* The process of preparing the goods to be shipped as well as packing the truck destined to the retail store. This is the exit point for the goods held within the DC.

Essentially, DC employees receive inventory from manufacturers and distributors, place it in one or more of several storage areas, and ultimately prepare it to be shipped to stores. Shipments from these DCs to retail stores can occur as often as daily or as infrequently as monthly, depending on the demand characteristics of the store. We will explore each of these phases in detail below.

Receiving

The receiving process marks the point of entry for inventory arriving into the DC from the manufacturer. A retailer receives goods through dock doors located in one distinct area of the warehouse that is separate from the outbound shipping area. Receivers are stationed at the dock doors and are tasked to manually examine the in-bound cases, or boxes, stacked onto pallets as the pallets are unloaded from the truck. Their main function is to check for discrepancies between

actual receipts and packing lists, and actual receipts from purchase orders.³ Discrepancies do arise with some frequency [32]. DeHoratius [33] found 8% of the purchase orders arrive in error and require additional work prior to acceptance into the DC.

To conduct this check, receivers sometimes break down “mixed” pallets, pallets that contain cases of different SKUs, and load individual cases onto separate pallets—one for each SKU. This enables them to certify that the expected number of cases for each SKU has arrived. A few cases are randomly opened in order for receivers to check for damage, verify that the appropriate quantity within a case has been received, and determine whether any unauthorized substitutions⁴ have occurred. Once receivers are satisfied with this shipment’s accuracy and completeness, they generate a series of bar code labels, or license plates, to affix to the front of each pallet. These license plates direct the pallet to the appropriate storage location within the DC and will be read by the stock keeper during the next stage of the distribution process.

The DC’s inventory records are updated to reflect these newly received items in the following way. Using a hand-held scanning device, receivers scan the bar code located on the case and then enter the number of cases of that particular SKU they have counted. The warehouse management system contains information regarding the expected number of individual items per case. Thus, this case count triggers an item-level inventory record update in the warehouse management system.

Storage and Put-Away

Once received, these pallets are stacked in a holding zone or staging area on the receiving

³Some companies check receipts against invoices or even against advance shipping notices. The important point is that there is some process whereby what actually arrives at the dock can be verified with what was expected to arrive.

⁴Unauthorized substitution is when a vendor, either mistakenly or knowingly, ships a substitute item for the item ordered.

floor of the DC until stock keepers can put them away into designated storage areas. The put-away process begins with the stock keeper scanning the license plate generated during the receiving process and ends with the stock keeper bringing the goods to their assigned location within one of three primary storage areas. The three primary storage areas include (i) the case-pack area, (ii) the break-pack area, and (iii) the bulk or flow-through area. As their names indicate, each storage area holds goods that will be shipped to stores in different formats. For example, the case-pack area stores goods that will be shipped to stores in full cases, the boxed quantity provided by the manufacturer. The break-pack area stores cases that will be opened, or broken, to allow the hand-picking of individual items to be shipped to stores. As described in the next section, pickers pack select quantities of individual items into totes containing a variety of different items. These totes will make their way to the outbound shipping lanes by way of conveyor belt. The flow-through area is essentially temporary storage for cross-docked products—products that will be shipped to stores within 24 hours of receipt.

Pallets or cases of the same SKU may be stored in multiple locations, each with a unique location number, either out of necessity (the assigned area for a particular SKU is not large enough for all items to be colocated) or by plan. For example, one store may receive case quantities of a particular SKU, and thus this SKU can be found in the case-pack area, while another receives less than case quantities, and thus this SKU can also be found in the break-pack area.

Picking

Picking is the step by which inventory is selected and moved from these storage areas to the shipping lanes dedicated to trucks delivering to particular stores. Pickers are responsible for selecting items from storage areas that will be shipped to stores. Pickers receive “pick lists” to guide their selection process. These pick lists represent the store’s replenishment order as they designate which items and quantity of items

should be selected from DC shelves. The list includes the SKU number, a short description of the product, a product location identifier, and a quantity indicating the exact amount that it to be selected. If a picker is selecting less than case quantities, these items will be selected from shelves that contain open cartons of items from which pickers will select small quantities and the selected items will be placed into a tote that will eventually be sent via a conveyor belt to the appropriate shipping lane for that particular store. For pickers selecting cases, these cases may be sent down the conveyor to the store’s shipping lane. A bar code is placed on each tote and case by the picker to identify which store’s order is contained in that tote or which store’s order is associated with that case.

Quality Control

After being selected and placed onto the conveyor belt, all items pass through quality control prior to being packed and shipped. Quality control often involves a comprehensive automated and a random manual process. The automated process entails a scanner located near the end of the conveyor belt that reads each bar code placed on the outbound totes and cases. This scan is meant to verify the completeness of a store’s order. DC managers turn to the results of this scan to justify that a tote or case destined for a particular store has actually left the facility. In addition, there is a manual audit to supplement this automated process. The manual audit is meant to verify the correctness of the order by actually matching items en route to a store to that store’s pick list for a randomly selected number of outbound cases and totes. Quality control personnel initial the bar code to indicate the order has been checked before allowing the tote or case to proceed to the shipping lanes.⁵

⁵Some retailers require manufacturers to allocate a certain number of cases to specific stores. The manufacturer would then label the cases appropriately (e.g., with a store number) and these cases can then be shipped to stores upon their receipt at the DC. Similarly, retailers may label cases with store specific details upon their receipt and ship them out to stores the same day.

Loading and Store Receiving

Once the totes and cases have cleared quality control, they are automatically sorted by the conveyor belt system into different staging areas near store-specific shipping lanes where the cases and totes are packed and loaded for transport to the retail stores. Packers place the totes and cases onto pallets, shrink-wrap the pallets, and then place the pallet into the appropriate shipping lane. Dockworkers then load the pallets onto the designated truck. Once the truck is loaded, the door is closed and sealed and the items are now en route to the retail store.

At this stage, the DC's and the store's inventory records are updated to reflect this change in inventory ownership. The store's inventory is automatically updated to reflect the receipt of its replenishment order. The DC's inventory is automatically depleted by the amount of items it shipped to fill each of the order's from its affiliated stores. If the DC was unable to completely fill the store order (e.g., the DC was out of stock of a popular item), that item would need to be manually "scratched" or removed from the order to prevent the corruption of both DC and store inventory records.

Most retailers do not scan each individual item into its store inventory but rather rely on the automatic inventory update. The receiving process at the backdoor of a store entails a check to see whether the expected number of pallets or cases has arrived and, perhaps, a partial count of high value items. Some retail policies discourage store employees from conducting a detailed check of store receipts, as only discrepancies above a certain threshold dollar amount qualify a store for possible credit from the DC.

EXECUTION FAILURES IN THE RETAIL REPLENISHMENT PROCESS

The retailer's objective is to ensure that the delivery truck destined for a particular store is filled with the right product in the right quantity as defined by the replenishment order. The expectation is that each replenishment order will be filled accurately and

completely. However, the replenishment process described above may not always proceed as planned. There are numerous opportunities for mistakes to occur (See Fig. 2), each of which may have different consequences for the accuracy of store inventory records. This section describes and categorizes the errors that may occur in the replenishment of store inventory from the retailer's DC.

Sorting Integrity

During the receiving process, when receivers are breaking down mixed pallets and sorting cases in an effort to place like products onto the same pallet, it is possible for receivers to mistakenly misplace a case onto a pallet of dissimilar SKUs. It is very difficult to differentiate cases containing different SKUs from one another quickly since manufacturers' packaging is nearly identical and the markings on bar codes can be difficult to read without the assistance of a hand-held scanner.

In this instance, when a case is misplaced onto a pallet of dissimilar SKUs during the receiving process, this case will be routed to the wrong storage location and, if not discovered, will result in a picker selecting the wrong product and the store receiving the wrong product. Consequently, the store that was supposed to receive a case of SKU A actually receives a case of SKU B and two inventory records are now corrupt. The store's inventory records will be updated to reflect the receipt of an entire case of SKU A but none has been received. Instead, it has received an entire case of SKU B while the recorded quantity of SKU B at that store remains unchanged.

It is difficult to correct a sorting error because it occurs so early in the distribution process. Once a case is on the wrong pallet, the case is sent to storage on this pallet. If the entire pallet gets shipped to the store, the error will be present in the store. The store inventory will be updated with 10 cases of item A when, in reality, 9 cases of item A and 1 case of item B were received. Such errors are difficult to identify without an audit.

Another difficult problem to identify is when a tote or case does not make it to the correct shipping lane from the conveyor belt. This happens when the scanner is unable

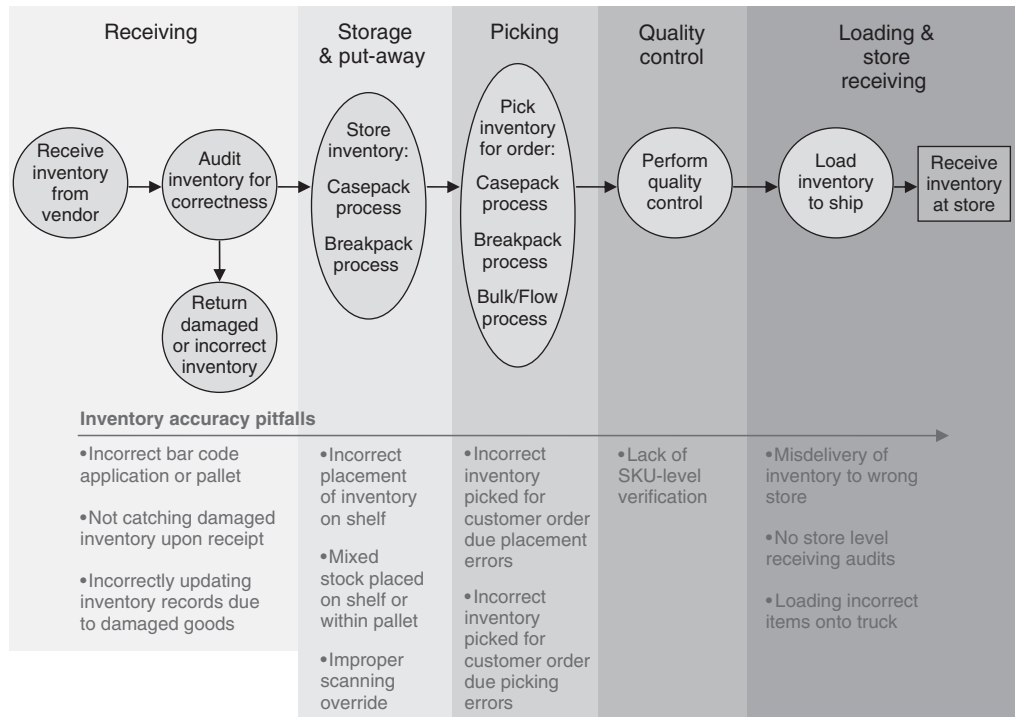


Figure 2. Sample of Errors that Arise During the Retail Distribution Process.

to read the bar code placed on the case or tote by the picker. In this instance, the scanner directs the tote or case to a “jackpot lane.” Employees working this jackpot lane are dedicated problem solvers that need to identify with which store’s order this case or tote is associated. Determining this may be easy when the bar code exist and is still legible. However, the bar code may be missing entirely or be damaged. Thus, it is quite possible that the store-specific pallets get shipped to stores without specific cases or totes. Fixing the inventory record then requires a manual intervention once DC employees identify the items in the jackpot lane.

Slotting Integrity

Slotting integrity refers to the misplacement of product that can occur during the put-away process. Stock keepers can mistakenly place an entire pallet of SKUs in the wrong location. Because it is often difficult to differentiate between shelves and even aisles

within the warehouse, it is possible for DC employees to mistakenly place a pallet on the wrong shelf.

To help limit this problem, some retailers require a stock keeper to scan the pallet license plate and the shelf location to ensure a match. However, it is still possible to scan the correct bar code of the pallet and the location and still place the product in a neighboring slot. Recognizing this, retailers have tried to make some improvements to signage by adding arrows that show whether the location bar code corresponds to the slot above or below the shelf.

Exacerbating the problem of slotting integrity is the problem of product overflow. When there is not enough capacity in the designated location, overflow products are often stacked down on the floor nearby the assigned area. Movement of the product from the floor into the assigned location is not managed by the automated warehousing management system, leaving it prone to error.

The repercussions of slotting errors include an inability to fill a store order due to the misplacement of product within the DC. While the product exists in the warehouse management system, locating it will require additional research and labor. However, an inability to fill a store order will not necessarily result in IRI. Instead, IRI will occur when the slotting error is not identified and a picker selects product from the designated location without checking the SKU number or description. In this scenario, the inventory of two records within one store will be corrupt if the slotting error occurred in the case-pack location and multiple stores if the slotting error occurred in the break-pack locations.

Picking Integrity

Even when product has been stacked and slotted correctly, pickers may pick from the wrong location. According to warehouse managers, it is not uncommon for warehouse employees to be picking from the correct location in terms of shelf number but to be on the incorrect floor. And, even if they are picking from the right location, and the right floor, it is still possible to pick from the wrong bin. Many retailers arrange its aisles similarly to the aisles of their store (i.e., similar products grouped together) and employees are expected to move rapidly back and forth along the aisle, selecting inventory from a range of locations. Differentiating between similar items is difficult when there is only time to quickly scan the product in a particular slot and compare that item to the one on the pick list. Moreover, slots in the break-pack pick locations may become contaminated with stray products. Break-pack items can fall down shelves or slide to an adjacent shelf and such movement often goes unnoticed by warehouse employees since it is extremely difficult to differentiate similar items contained in generic cardboard cartons.

The selection of the wrong item, has repercussions for the inventory of the store. The wrong product will be sent to the store while the record will be updated to reflect the receipt of the product the DC planned to have picked. Consequently, two inventory records

in one store will inaccurately reflect the quantity of that item.

Quantity Integrity

In addition to picking the right product, pickers are expected to pick the right quantity of that product. Employees can miscount, selecting, for example, 15 units of a given product rather than the requested 13 units. And, employees could also confuse items on the pick list. For instance, a pick list that requests 10 units of item A be selected and placed into a tote along with 5 units of item B could be misread as the employee places 5 units of item A into a tote along with 10 units of item B.

“Unit of measure” problems also occur when an employee confuses the quantity noted on the pick list. For example, a pick list could request one case and the picker mistakenly picks one item. This is particularly true in the break-pack area where manufacturer’s packaging must be opened by the picker at the shelf. Manufacturers may have introduced a new SKU which is a two-pack of, say, batteries. The picker mistakenly breaks these two packs into two separate items. In this instance, when the picker selects one unit from this location, one item is taken rather than the one two-pack. Another example where such mistakes can arise is when the manufacturer packs a case of screws and has packed certain quantities of screws within plastic bags that are then shipped in the case. A picker may be asked to pick one unit from a certain location and it is unclear whether that one item means one individual screw or one plastic bag of screws.

Such picking errors result in store-level IRI. The store receives a tote that fails to contain the correct inventory, while its records, upon receiving the tote, have been updated under the assumption of perfect execution by the picker. Quantity errors, unlike other errors, impact only one SKU per store. However, quantity errors such as the ones described also cause problems for the DC inventory management.

Shipping Integrity

Once an item has been picked, it is sent via conveyor belt to the appropriate shipping lane for the store that has ordered that item. At this point in time, DC employees place the totes and cases onto store-specific pallets that will be loaded onto outbound trucks. The shipping lanes are located very close to each other separated only by lines on the floor. The store-specific pallets that are built by these employees may end up on the wrong truck. While this error is often caught as the truck has fixed capacity such that only a certain number of pallets will fit, fixing the errors does take time. Specifically, products need to be off-loaded from the truck in an effort to find the incorrect pallet. If the warehouse identifies the error prior to the release of the final invoice, the products are removed from the truck and ultimately do not cause a data discrepancy. If the error is not observed in advance, a data discrepancy can occur.

Finally, shipping integrity can be compromised by the truck drivers during the unloading process. If the truck has been loaded incorrectly—meaning the pallets are not placed in the truck in the reverse order of the truck drivers route—the driver has to off load the pallets of other stores in order to locate and unload the pallets for the current delivery. It is then the driver's responsibility to get all the totes for the other stores back onto the truck and proceed to the next store. This may not happen. When it does not, store managers may not assist the driver or the DC in locating the inventory. To the store manager, this additional pallet is "free"—meaning they do not incur the cost of goods sold on their store operating budget but they do reap all the reward—sales dollars.

These shipment errors are significant because they not only cause errors in multiple SKUs at one store but they also cause inaccuracies in the inventory of multiple stores. On the other hand, because they involve large dollar amounts, these types of errors are often more readily caught at the store than some of the other errors previously described. Store managers often receive an itemized list of the products scheduled for

delivery and use this list to verify receiving the appropriate number of pallets from the drivers. Yet, there is no mechanism in place to prevent a driver from unloading a pallet at the wrong store. Moreover, few retailers verify, at the item level, whether what was received matches what was expected for several reasons. First, this is an extremely time-consuming task that requires a lot of labor and delays getting the inventory onto the store floor where customers can make a purchase. Second, even if they were to find a discrepancy, many retail-owned DCs have policies in place that discourage the reporting of such discrepancies. Such policies often do not allow the DC to credit stores for shipment discrepancies on items less than a certain threshold. Thus, store employees only verify the quantity of these expensive items while discrepancies in the inventory record for other items will persist until a routine physical inventory count triggers the updating of the recorded inventory.

DOCUMENT INTEGRITY

Receivers are expected to verify that the quantity per case actually reflects the quantity per case noted in the warehouse information system. When the inventory system has outdated information regarding the quantity per case, the inventory record updated at receiving will be inaccurate. Moreover, the picking will be negatively impacted as well. The system generates one label per case on a pick list. If the store order calls for 24 units and the system thinks there are 24 units per case, one label will be generated, a picker will be directed to pick one case, and the store will receive one case. Suppose that the case did not contain 24 units but actually contained only 12 and this was not corrected during the receiving process. The pick list would still say pick one case, one label would be generated, and once the case departs the DC on a truck bound for a store, the inventory records of the store are updated to reflect this case of 24 units. However, in reality, a discrepancy of 12 units exists due to the lack of document integrity.

This problem is exacerbated by special promotions offered by vendors. For example,

HP might run a special promotion on printer ink where they package two of the same SKU together. When they package these inkjet cartridges, they place a total of 12 of these promotional units in the case but one of these promotional units amounts to two units of the actual item. If this promotion unit does not have its own SKU, it is unclear how best to receive these units. Does the case contain 12 or 24? This complexity continues into different stages when pickers are directed to pick one unit of this SKU—does this unit of measure refer to the one promotional unit or to the one individual unit?

Another type of documentation error is the mislabeling either by vendors or by the receivers. It is possible that the bar code on a case received from the manufacturer does not accurately reflect the content of that case. Receivers are expected to open a few of the cases to verify contents but, given that not all cases are checked, such mislabeling can persist. When a case is mislabeled, it will be mistakenly placed on the wrong pallet, compromising the integrity of that pallet. Receivers may also apply erroneous license plates to pallets. Often, these license plates are printed on long strips and receivers are required to manually match the labels with the correct pallets—a process that can be easily circumvented by receivers. When a pallet is mislabeled, ultimately it compromises the integrity of the location such that the inventory system thinks that the product is located on one shelf when it is actually located elsewhere.

When an incorrect bar code has been placed on a pallet during the receiving process, the pallet could be misplaced. The bar code system ultimately relies on the stock keeper to catch mistakes through visual confirmation. However, since the stock keeper is operating under time pressure to put-away stock, the stock keeper can easily misplace stock and perform an override—not catch an incorrect labeling problem or forget to scan the item into its location entirely.

A reliable replenishment process requires that the right item, in the right quantity, with the right documentation is received, stored, and then shipped to stores. However, table 1 summarizes such errors as an obstacle

to reliability across one item, many items, and throughout the chain. Errors that occur at the beginning of the replenishment process, when undiscovered, persist throughout the rest of the replenishment process. Even if errors are uncovered, fixing them often requires manual intervention, which is also prone to error (e.g., entry errors, omissions errors) to correct the inventory management system.

The Impact of Different Error Types

	One SKU	Multiple SKU	One Store	Multiple Stores
Sorting integrity		X	X	
Slotting integrity		X	X	X
Picking integrity		X	X	
Shipping integrity		X		X
Quantity integrity	X		X	
Document integrity		X		X

FUTURE RESEARCH

There exist many unanswered questions regarding IRI. Two that require additional attention include:

1. Why are retail employees more likely to make mistakes in contexts with high product variety and high inventory density, two known drivers of IRI? Past research has identified these factors as exacerbating the record inaccuracy problem. However, this work falls short in explaining why. Instead, DeHoratius and Raman [3] propose that inventory density and product variety are proxies for confusion in the retail context and that such confusion makes it more likely for retail associates to make a mistake. What might a retailer do to mitigate the impact of product variety and inventory density on IRI? Is it feasible to design product packaging in such a way that ensures fewer mistakes in replenishment and sales transactions? Might retailers benefit from designing shelving in stores and DCs that make

it less likely for mistakes to arise? To address some of these questions, researchers will need to employ behavioral operations methods. Such studies of the behavior of individuals and teams within specific contexts would help identify the precise relationship between retail contexts, ones with high variety and high density, and the occurrence of IRI.

2. Retailers continue to question whether the cost of implementing key process improvement techniques aimed at solving IRI can be offset by the potential gains from reducing IRI. Given the dearth of publicly known pilot projects attempting to address IRI through process improvement mechanisms, this question continues to arise. Researchers need to evaluate the performance of different improvement mechanisms in the field.

In addition to these two key questions, we also feel that existing demand forecasting techniques can benefit from incorporating IRI and that the accuracy of selected forecasts depends on the degree to which the inventory record accurately reflects what is physically present in the store.

Among those who have specifically analyzed the problem of IRI [8,9,18,34–35], none provides an in-depth study of the role of human error and, consequently, imperfect execution as a potential source of IRI. Humans, when engaged in any organizational process, will inevitably make mistakes [36,37]. Reason [38] grouped these mistakes into three different categories: knowledge-based errors, rule-based errors, and skill-based slips and lapses. Knowledge-based errors arise from humans attempting to solve a problem that is unfamiliar to them. Rule-based errors arise when humans apply the wrong rule to a situation or elect not to apply an existing rule. Skill-based slips and lapses are essentially failures in execution. A procedure exists but for some reason the existing procedure is not carried out for some routine activity.

Why such errors occur has been a matter of much discussion among human factors and human reliability scholars. In fact, research into the cause of error in organizational processes has resulted in numerous taxonomies to describe how characteristics of the individual, task, team, and organization, for example, impact human performance. Swain and Guttman [39] provide one often referenced taxonomy where internal performance shaping factors (the individual skills, ability, attitudes, and other characteristics a worker brings to the job), external performance shaping factors (situation characteristics, task, equipment, and procedural characteristics), and psychological and physiological stressors affect a worker's performance of tasks within an organizational system. Kim and Jung [40] provide a comprehensive review of these taxonomies and offer a single taxonomy that synthesizes existing work in this area. According to Kim and Jung [40], human performance depends on individual capabilities and characteristics, the physical characteristics of the organizational process, the procedures and tasks required of the individual, and an environmental influence that includes characteristics of the team and culture of the organization.

The impact of human error has been discussed in numerous contexts including hospitals [41], nuclear plants [39,44], chemical plants, aerospace [45], and orienteering [46]. In these contexts, concern over prevention of catastrophic accidents is of primary concern. The possibility of human error contributing to systemwide failure, the consequence of which has impact for human life, sets research in these contexts apart from our own. Research on human error in the DC context is more similar to research on the impact of human error on delivering service quality to both external [37] and internal customers [47,48]. The impact of service failures, like the impact of IRI, is not as readily visible to the outside as the impact of catastrophic failure in these other contexts. Instead, service failures and the presence of IRI both impose excess cost on the organization, but measuring this cost is difficult due the subtlety of tracking changes in customer behavior.

APPENDIX: FIELDWORK ACTIVITIES AND SITE DESCRIPTION

The core of our observations were drawn from data gathered and interviews conducted at two Gamma⁶ DCs, 15 Gamma stores, and corporate headquarters. Gamma Corporation is a large public retailer with annual sales of roughly \$11 billion and more than 1500 stores worldwide. Since Gamma's founding more than 15 years ago, it has grown to employ over 50,000 people and is currently the largest operator of superstores in the world for its specific product category. Despite its reputation as a leader in its field and the perception that Gamma is an advanced user of information technology, inaccurate inventory records are present in each of the 37 stores observed by DeHoratius and Raman [3].

A typical Gamma store is approximately 24,000 sq. ft in size and is served by one of three DCs which range in size from 310,000 to 830,000 sq. ft. Although Gamma stores receive merchandise from their assigned DC, but they also can receive merchandise directly from vendors. This article focuses only on those items shipped to Gamma stores from the retail-owned DC. Previous work on IRI reveals that items shipped from the retail-owned DCs possess the least accurate records, controlling for differences in product category, stores, the cost of an item, and the quantity it has sold throughout the year [3].

Collectively, we have spent more than 12 weeks in the field observing DC processes across a wide variety of retailers in order to develop inductively our insights on distribution process quality and implications of process defects for the management of retail supply chains. Where possible, we recorded our interviews using a hand-held tape recorder and transcribed these tapes immediately following site visits. Otherwise, we took detailed notes by hand. We also, when permitted, took pictures to document our observations. We did not make any attempt to measure the frequency of error occurrence; but rather our intent was to identify and classify the types of errors that arise.

Our objective in conducting this fieldwork was to understand how inaccurate inventory records arise in Gamma stores. Such qualitative methods are appropriate when the phenomenon of interest is not well understood and detailed observations are required for categorization and theory-building [50].

REFERENCES

1. Kang Y, Gershwin S. Information inaccuracy in inventory systems: stock loss and stockout. *IIE Trans* 2005;37:843–859.
2. Gruen TW, Corsten D, Bharadwaj S. Retail out-of-stocks: a worldwide examination of extent, causes, and consumer responses Grocery Manufacturers of America: The Food Marketing Institute, and CIES; Washington DC: 2002.
3. DeHoratius N, Raman A. Inventory record inaccuracy: an empirical analysis. *Manage Sci* 2008;43(4):627–641.
4. Rinehart RF. Effects and causes of discrepancies in supply operations. *Oper Res* 1960;8(4):543–564.
5. Emma CK. Observations on physical inventory and stock record error. Interim Report 1. Mechanicsburg (PA): Department of Navy Supply Systems Command; 1966.
6. Ernst R, Guerrero J, Roshwalb A. A quality control approach for monitoring inventory stock levels. *J Oper Res Soc* 1993;44(11):1115–1127.
7. Opolon D, Schumacher K, Sheffi Y. Impact of imperfect demand recording on service levels: the case of hospital supplies. Proceedings of the 20th Annual POMS Conference. Orlando (FL); 2009.
8. Millet I. A novena to Saint Anthony, or how to find inventory by not looking. *Interfaces* 1994;24:69–75.
9. Sheppard GM, Brown KA. Predicting inventory record-keeping errors with discriminant analysis: a field experiment. *Int J Prod Econ* 1993;32:39–51.
10. DeHoratius N, Ton Z. The role of execution in managing product availability. In: Agrawal N, Smith S, editors. *Retail supply chain management*. New York: Springer; 2008. Chapter 4.
11. Fleisch E, Tellkamp C. Inventory inaccuracy and supply chain performance: a simulation study of retail supply chains. *Int J Prod Econ* 2005;95(3):373–385.

⁶Name disguised at company's request.

12. Waller MA, Nachtmann H, Hunter J. Measuring the impact of inaccurate inventory information on a retail outlet. *Int J Log Manage* 2006;17(3):355–376.
13. Raman A, DeHoratius N, Ton Z. Execution: the missing link in retail operations. *Calif Manage Rev* 2001;43(3):136–152.
14. DeHoratius N, Mersereau A, Schrage L. Retail inventory management when records are inaccurate. *Manuf Serv Oper Manage* 2008;10(2):257–277.
15. Angulo A, Nachtmann H, Waller MA. Supply chain information sharing in a vendor managed inventory partnership. *J Bus Log* 2004;25(1):101–120.
16. Sari K. Inventory inaccuracy and performance of collaborative supply chain practices. *Ind Manage Data Syst* 2008;108(4):495–509.
17. Camdereli AZ, Swaminathan JM. Misplaced inventory and RFID: information and coordination. *Prod Oper Manage* 2005;19(1):1–18.
18. Iglehart DL, Morey RC. Inventory systems with imperfect asset information. *Manage Sci* 1972;18(8):B338–B394.
19. Morey RC. Estimating service level impacts from changes in cycle count, buffer stock, or corrective action. *J Oper Manage* 1985;5(4):411–418.
20. Kök AG, Shang K. Inspection and replenishment policies for systems with inventory record inaccuracy. *Manuf Serv Oper Manage* 2007;9(2):185–205.
21. Atali A, Lee H, Ozer O. If the inventory manager knew: value of visibility and RFID under imperfect inventory information. SSRN Working Paper; 2006.
22. Mersereau AJ. Information-sensitive replenishment when inventory records are inaccurate. Working Paper. University of North Carolina Kenan-Flagler Business School; 2009.
23. Neely PS. A framework for cycle counting. *Prod Inventory Manage* 1983;24:23–32.
24. Brooks RB, Wilson LW. Inventory record accuracy: unleashing the power of cycle counting. Essex Junction: Oliver Wight Publications; 1993.
25. Flores BE, Whybark DC. Implementing multiple criteria ABC analysis. *J Oper Manage* 1987;7:79–85.
26. Morey RC, Dittman DA. Optimal timing of account audits in internal control. *Manage Sci* 1986;32(3):272–282.
27. Rekik Y, Sahin E, Dallery Y. Analysis of the impact of RFID technology on reducing product misplacement errors at the retail stores. *Int J Prod Econ* 2008;112(1):264–278.
28. Gaukler GM, Seifert RW, Hausman WH. Item-level RFID in the retail supply chain. *Prod Oper Manage* 2007;16(1):65–76.
29. Gaukler GM, Ozer O, Hausman WH. Order progress information: improved dynamic emergency ordering policies. *Prod Oper Manage* 2008;17(6):599–613.
30. Uçkun C, Karaesmen F, Sava S. Investment in improved inventory accuracy in a decentralized supply chain. *Int J Prod Econ* 2008;113(2):548.
31. Szmerekovsky J, Zhang J. Coordination and adoption of item-level RFID with vendor managed inventory. *Int J Prod Econ* 2008;114(1):388.
32. Deshpande V, Schwarz B, Raju V. Inventory Management under Product Misidentification/Shipment Errors. Krannert School of Management Working Paper. Purdue University West Lafayette, IN 2009.
33. DeHoratius, Identifying Defects and Rework in the Inventory Management Process Extended Retail Ind J 2006;2(5):24–25.
34. Neeley PS. A framework for cycle counting. *Prod Inv Manage* 1983;24:23–32.
35. Bergman RP. A B count frequency selection for cycle counting support MRP II. *Prod Inv Manage Rev* 1988;8:35–36.
36. Edmondson AC. Learning from mistakes is easier said than done: group and organizational influences on the detection and correction of human error. *J Appl Behav Sci* 1996;32(1):5–28.
37. Stewart DM, Chase RB. The impact of human error on delivering service quality. *Prod Oper Manage* 1999;8(3):240–264.
38. Reason JT. Human error. Cambridge: Cambridge University Press; 1990.
39. Swain AD, Guttman HE. A handbook of human reliability analysis with emphasis on nuclear power plant applications. Washington (DC): US Nuclear Regulatory Commission; 1985. USNRC-Nureg/CR-1278.
40. Kim JW, Jung Wondea. A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. *J Loss Prev Process Ind* 2003;16(6):479–495.
41. Leape L. Out of the darkness: Hospitals begin to take mistakes. *Health Syst Rev* 1996;29(6):21–24.

42. Tucker AL. The impact of operational failures on hospital nurses and their patients. *J Oper Manage* 2004;22(2):151–169.
43. Flynn EA, Barker KN, Gibson JT, *et al.* Impact of interruptions and distractions on dispensing errors in an ambulatory care pharmacy. *Am J Health Syst Pharm* 1999;56: 1319–1325.
44. Perrow C. Normal accidents. Princeton (NJ): Princeton University Press; 1999.
45. Vaughan D. The trickle-down effect: Policy decisions, risky work, and the Challenger tragedy. *Calif Manage Rev*; 1997;39(2): 80–102.
46. Roberto A. Lessons from Everest: The interaction of cognitive bias, psychological safety, and system complexity. *Calif Manage Rev*; 2002;45(1):136–158.
47. Auty S, Long G. Tribal Warfare and gaps affecting internal service quality. *Int J Serv Ind Manage*; 1999;10(1):7–22.
48. Bowen DE, Johnston R. Internal Service recovery: developing a new construct. *Int J Serv Ind Manage*; 1999;10(2):118–131.
49. Strauss A, Corbin J. Basics of qualitative research techniques and procedures for developing grounded theory. Sage Publications; Thousand Oaks(CA): 1998.
50. Yin RK. Case study research: design and methods. Thousand Oaks (CA): Sage Publications; 1994.

IRANIAN OPERATIONS RESEARCH SOCIETY (IORS)

NEZAM MAHDAVI-AMIRI
Department of Mathematical
Sciences, Sharif University of
Technology, Tehran, Iran

Iranian Operations Research Society (IORS) was established in 2005 after the approval of its constitution by the Ministry of Science, Research, and Technology in Tehran, Iran. Shortly afterwards, in July 2005, a meeting of interested scholars was held and the first Executive Council of the Society was formed by election. The Society was officially approved to join IFORS as the 50th national member on 16 December 2009. Currently, the society is being managed by its second executive council to be replaced by the third executive council in May 2010 (the term for the council is two years).

IORS currently has 134 members, mostly composed of college scholars and industry professionals. The scholars are mostly from the departments of mathematics, industrial engineering, and management sciences.

The activities of the Society are summarized below:

1. enhancing scientific and cultural research at national and international levels among the researchers and experts who deal with operations research;
2. cooperating with administrative, scientific, and research agencies in evaluating and reviewing plans and programs related to education and research issues concerning the society's interests;
3. promoting researchers and distinguishing successful researchers by honoring them;
4. rendering educational and research activities;
5. publishing books and international journals;

6. organizing scientific gatherings at national, regional, and international levels by holding regular internal seminars and international conferences.

The principal decision-making body of IORS is its general assembly. There is one yearly regular meeting of the assembly as well as additional emergency meetings, when required. The assembly is formed by the permanent members of IORS and is responsible for upholding, modifying, and amending the constitution. It also has the responsibility to vote for the election of the executive council of IORS (with seven main and two reserved members) and the inspectors (one main and one reserved) for orderly execution of the IORS constitution as well as its management of daily affairs. The main members of the elected executive council then decide upon the President, the Vice-President, and the Treasurer of the council at their first meeting. The main members for the first two executive councils and the main inspector are as follows:

IORS's First Executive Council Members and Inspectors:

- Dr. Korosh Eshghi, Department of Industrial Engineering, Sharif University of Technology, Tehran
- Dr. Hasan Salehi Fathabadi, Department of Mathematics and Computer Science, University of Tehran, Tehran
- Dr. Qolamreza Jahanshahloo, Department of Mathematics and Computer Science, Tarbiat Moallem University, Tehran
- Dr. Ahmad Makoui, Department of Industrial Engineering, Iran University of Science and Technology, Tehran
- Dr. Nezam Mahdavi-Amiri (Vice-President), Department of Mathematical Sciences, Sharif University of Technology, Tehran
- Dr. Azizollah Memariani (President), Department of Industrial Engineering, Tarbiat Modarres University, Tehran

Dr. Mohammad Modarres Yazdi (Treasurer), Department of Industrial Engineering, Sharif University of Technology, Tehran

Dr. Mohammad Reza Hamidi Zadeh (Inspector), Department of Management and Accounting, Shahid Beheshti University, Tehran.

IORS's Second Executive Council Members and the Inspector:

Dr. Farhad Hosseinzadeh Lotfi, Department of Mathematics, Azad University Research Center, Tehran

Dr. Qolamreza Jahanshahloo, Department of Mathematics and Computer Science, Tarbiat Moallem University, Tehran

Dr. Nezam Mahdavi-Amiri (Vice-President), Department of Mathematical Sciences, Sharif University of Technology, Tehran

Dr. Azizollah Memariani (President), Department of Industrial Engineering, Tarbiat Modarres University, Tehran

Dr. Mohammad Modarres Yazdi, Department of Industrial Engineering, Sharif University of Technology, Tehran

Dr. Abbas Seifi, Department of Industrial Engineering, Amirkabir University of Technology, Tehran

Dr. Reza Tavakoli Moghadam (Treasurer), Department of Industrial Engineering, University of Tehran, Tehran

Dr. Mohammad Reza Hamidi Zadeh (Inspector), Department of Management and

Accounting, Shahid Beheshti University, Tehran.

IORS publishes the international scientific journal, *Iranian Journal of Operations Research (IJOR)*. The publication is semianual and was first published in 2009.

IORS also organizes the annual International Conference of Iranian Operations Research Society. The first of these conferences was held in 2008. Below is a list of scientific gatherings from the first three years of this conference:

The first International Conference of Iranian Operations Research Society, Jan. 28–30, 2008, Kish University branch of Sharif University of Technology, Kish Island, Iran.

The second International Conference of Iranian Operations Research Society, May 20–22, 2009, Mazandaran University, Babolsar, Iran.

The third International Conference of Iranian Operations Research Society, May 5–6, 2010, Amirkabir University of Technology, Tehran, Iran.

It is important to note that about 600 papers were submitted to the scientific committee of the third conference, and approximately 200 were selected for oral presentations and 100 were selected as posters.

For more information, the reader may visit the website of IORS at: <http://www.iors.ir>.

JACKSON NETWORKS (OPEN AND CLOSED)

HANS DADUNA
Center of Mathematical Statistics
and Stochastic Processes
Department of Mathematics,
University of Hamburg,
Hamburg, Germany

DESCRIPTION OF JACKSON NETWORKS AND FUNDAMENTAL PROPERTIES

We describe classical queueing network theory, which is the fundamental of the rich class of stochastic network models available today. These networks are built on exponential multiserver queues, see the section titled “Single-Station Queues CTMC Models” in this encyclopedia.

Definition 1 [M/M/s/N: Exponential Multiserver Queues]. This is a service station with $s \geq 1$ service channels and a linear waiting room of length N under (first-come first-served (FCFS) regime, $0 \leq N \leq \infty$). Indistinguishable customers arrive in a Poisson- λ process at the station and request for an amount of work (service time) which is exponentially distributed with mean μ^{-1} . All service times are independent and independent of the arrival stream. If an arriving customer finds free service channels, he selects one of them randomly and service commences immediately. If he finds all channels busy and there is free waiting room, he joins the tail of the waiting line; otherwise, this customer is lost. If a service expires, the customer at the head of the line immediately enters the free service channel and the other customers in the waiting line move one step ahead. Let $X(t)$ denote the number of customers in the system at time $t, t \geq 0$. We call $X(t)$ the queue length ($\equiv$ customers waiting or in service). $X = (X_t : t \geq 0)$ is a Markov process with

state space $\mathbb{N}^J$. Unless otherwise specified, our systems will always have infinite waiting room.

Network Models and Steady-State Behavior

Definition 2 [Jackson Network]. A Jackson network is a network of service stations (nodes) numbered $\{1, 2, \dots, J\} := \tilde{J}$. Station j is a multiserver station with $s_j \geq 1$ service channels and an infinite waiting room under the FCFS regime. Customers in the network are indistinguishable. At node j , there is an external Poisson- λ_j arrival stream, $\lambda_j \geq 0$. Customers at node j request for an amount of work (service time) there which is exponentially distributed with mean $\mu_j > 0$. Service times and inter-arrival times are independent.

Movements of the customers follow a Markovian routing mechanism: A customer on leaving node i visits node j next with probability $r(i, j) \geq 0$, entering node j immediately; with probability $r(i, 0) \geq 0$, this customer leaves the network immediately ($\sum_{j=0}^J r(i, j) = 1, \forall i$). Given departure node i , the customer's routing is independent of the network's history. With $\lambda := \sum_{j=1}^J \lambda_j$ and $r(0, j) := \frac{\lambda_j}{\lambda}$ the matrix $R := (r(i, j) : i, j = 0, 1, \dots, J)$ is assumed to be irreducible.

Let $X_j(t)$ denote the number of customers at node j at time $t \geq 0$, either waiting or in service (local queue length at node j); then $X(t) := (X_j(t) : j = 1, \dots, J)$ is the joint queue length vector of the network at time t . $X = (X(t) : t \geq 0)$ is the joint queue length process of the Jackson network.

Theorem 1 [1]. *The joint queue length process X of the Jackson network is an irreducible Markov process with state space $E = \mathbb{N}^J$. The traffic equation of the network*

$$\eta_j = \lambda_j + \sum_{i=1}^J \eta_i r(i, j), \quad j = 1, \dots, J, \quad (1)$$

has a unique solution, denoted by $\eta = (\eta_1, \dots, \eta_J)$, see Proposition 1 for an interpretation.

X is ergodic if and only if $\eta_j < \mu_j s_j, j = 1, \dots, J$, holds. The unique stationary and limiting distribution π on E of X is given as follows. Let π_j denote the local probability for node j ,

$$\pi_j(n_j) = \begin{cases} K_j^{-1} \left(\frac{\eta_j}{\mu_j} \right)^{n_j} \frac{1}{n_j!}, & \text{if } n_j \leq s_j \\ K_j^{-1} \left(\frac{\eta_j}{\mu_j} \right)^{n_j} \frac{1}{s_j!} \left(\frac{1}{s_j} \right)^{(n_j - s_j)}, & \text{if } n_j \geq s_j, \end{cases}$$

where K_j is the normalization constant. Then π is of product form

$$\pi(n_1, \dots, n_J) = \prod_{j=1}^J \pi_j(n_j), \quad (n_1, \dots, n_J) \in E.$$

Product-Form Equilibrium. The local queue lengths $X_j = (X_j(t) : t \geq 0)$ in equilibrium at a fixed time point are independent. But the processes X_j are not independent. They strongly interact via the moving customers. Jackson proved precisely the following: The one-dimensional marginals in time of the vector-valued stationary Jackson network process have independent coordinates in space.

Note, that something more is behind the product form: In equilibrium, the local queue lengths follow in a first-order description the linear interdependence relation (1), while the random fluctuations around the mean values are independent over space for any fixed time instant. On the other side is computing $\text{cov}(X_i(t), X_j(t+s))$ for $s > 0$, an unsolved problem even for $i = j$.

The form of the local equilibria π_j suggests that, in steady-state, node j acts like an isolated $M/M/s_j/\infty$ with Poisson- η_j arrivals. But, in general, the customer flows in the network are not Poissonian [2]. A survey on customer flow processes is given in Disney and König [3], and further details are in Disney and Kiessler [4].

Definition 3 [Gordon–Newell Network]. Consider the set of multiserver nodes

$\tilde{J} = \{1, \dots, J\}$ as described in Definition 2 without external arrivals. There are $I > 0$ customers cycling according to an irreducible Markov matrix $R = (r(i, j) : i, j = 1, \dots, J)$. Service times are as in the open network and the independence of service times and routing decisions hold as well.

Let $X_j(t)$ denote the number of customers at node j at time $t \geq 0$, waiting or in service (local queue length at j). $X(t) := (X_j(t) : j = 1, \dots, J)$ is the joint queue length vector of the network at time t . $X = (X(t) : t \geq 0)$ is the joint queue length process of the Gordon–Newell network.

Theorem 2 [5,6]. The joint Markovian queue length process $X = (X(t), t \geq 0)$ of the Gordon–Newell network is ergodic with state space $S(I, J) = \{(n_1, \dots, n_J) \in \mathbb{N}^J : n_1 + \dots + n_J = I\}$. Let $\eta = (\eta_1, \dots, \eta_J)$ denote the unique probability solution of the traffic equation

$$\eta = \eta R. \quad (2)$$

The unique stationary and limiting distribution $\pi = \pi(I, J)$ of X on $S(I, J)$ is

$$\pi(I, J)(n_1, \dots, n_J) = G(I, J)^{-1} \prod_{j=1}^J \left\{ \prod_{k=1}^{n_j} \frac{\eta_j}{\mu_j(k)} \right\}, \quad (n_1, \dots, n_J) \in S(I, J), \quad (3)$$

where $\mu_j(n) = \min(n, s_j), n \geq 0$ and $G(I, J)$ is the normalization constant.

Product-Form Equilibrium. The steady-state distribution $\pi(I, J)$ of the Gordon–Newell network process is said to be of product form as well. From Equation (3), the Gordon–Newell network equilibrium looks like being obtained from a Jackson network equilibrium by conditioning on the fixed number of customers. Therefore, $\pi(I, J)$ is negatively associated [7].

Visiting Ratios. The solution of the traffic equation (2) is not unique, but only fixed up to a constant factor. Our choice has the natural interpretation of η_j being the visit ratio of customers at node j . The stochastic solution of

Equation (2) is the steady-state distribution of the customers' migration chain. The choice of the factor for η has no influence on the steady-state distribution π in Equation (3).

Corollary 1 [Networks with State-Dependent Service Intensities]. *Consider a Jackson network or a Gordon–Newell network of the structure described in Definitions 2 and 3, but with single server nodes with state-dependent service intensities which are node (j) and locally state (k_j) dependent: Substitute $\mu_j(n) \leftarrow \min(n, s_j)$. Under ergodicity in the open network and in the closed network, in general, we have steady-state product-form results. Jackson Network with State-Dependent Single Servers.*

$$\pi(n_1, \dots, n_J) = \prod_{j=1}^J \left\{ K_j^{-1} \prod_{k=1}^{n_j} \frac{\eta_j}{\mu_j(k)} \right\},$$

$$(n_1, \dots, n_J) \in \mathbb{N}^J.$$

Gordon–Newell Network with State-Dependent Single Servers.

$$\pi(I, J)(n_1, \dots, n_J) = G(I, J)^{-1} \prod_{j=1}^J \left\{ \prod_{k=1}^{n_j} \frac{\eta_j}{\mu_j(k)} \right\}.$$

$$(n_1, \dots, n_J) \in S(I, J).$$

The network processes do not describe individual customer's behavior. To describe individual customer's behavior with more versatile migration, there are two different (equivalent) methods [8]: Random routing as used in the celebrated BCMP networks [9] and deterministic routing as described by Kelly [10, Chapter 3].

Definition 4 [Mixed Networks with Deterministic Routing]. We consider a network of exponential nodes $\tilde{J} = \{1, \dots, J\}$ according to Definition 1, node j having s_j identical service channels, with FCFS regime, providing service which is exponentially distributed with mean μ_j^{-1} . For a simple description, we assume that all positions where customers may be located are linearly ordered: at node j , positions $1, \dots, s_j$ represent the service channels,

position $s_j + 1$ houses the customer at the head of the queue, and so on.

The network serves a finite population of internal customers numbered $1, \dots, M, M \geq 0$. At any time, internal customer m is of some type $t \in T(m)$. Further, external customers arrive at the network, being of some type $t \in T(0)$. The sets $T(m), m = 0, 1, \dots, M$, are countable and pairwise disjoint. With each type $t \in \bigcup_{m=0}^M T(m)$ is associated the finite route $r(t) = [r(t, 1), \dots, r(t, S(t))]$, which a customer of type t follows. Multiple visits to nodes on a route are allowed.

External customers of type $t \in T(0)$ arrive in a Poisson stream with rate $\lambda_t > 0$, enter node $r(t, 1)$, and after leaving that node immediately jump to node $r(t, 2), \dots$, and being finally served at node $r(t, S(t))$, they depart from the network. Inter-arrival times and service times are independent.

Internal customer m , of type t , travels route $r(t)$, thereafter changes to type $t' \in T(m)$, and travels route $r(t')$ with probability $p_m(t, t')$, where $P_m = (p_m(t, t') : t, t' \in T(m))$ is an irreducible positive recurrent stochastic matrix. Given the previous type t , the type decision is independent of the network's history.

Using type-stage sequences, the state description is as follows: For node j , the local state $e_j \in E_j$ indicates node j to be empty and $z_j = [t_{j1}, s_{j1}; t_{j2}, s_{j2}; \dots; t_{jn_j}, s_{jn_j}] \in E_j$ indicates that there are n_j customers present, on position $p \in \{1, \dots, n_j\}$ resides a customer of type t_{jp} being on stage $s_{jp} \in \{1, \dots, S(t_{jp})\}$ of his route $r(t_{jp})$. E_j is the local state space of node j . Global states are compounded of local states, and the global state space E contains all feasible states

$$z = [z_1, \dots, z_J] \in E \subseteq \prod_{j=1}^J E_j.$$

Theorem 3 [Steady-State Distribution for Mixed Exponential Networks]. *For the mixed exponential network of Definition 4, let $Z = (Z(t) : t \geq 0)$ be the Markovian type-stage process (generalized queue lengths process) with state space E . We assume that Z is irreducible on E .*

For internal customers we define the traffic equations $\alpha_m = \alpha_m P_m, m = 1, \dots, M$

and denote their stochastic solution by $\alpha_m = (\alpha_m(t) : t \in T(m))$. Assume furthermore

$$\forall m = 1, \dots, M : \sum_{t \in T(m)} \alpha_m(t) S(t) < \infty \text{ holds,}$$

$$\text{and } \sum_{t \in T(0)} \lambda_t S(t) < \infty.$$

For $j \in \tilde{J}$, $t \in \bigcup_{m=0}^M T(m)$, $1 \leq s \leq S(t)$, denote

$$\eta_j(t, s) = \begin{cases} \lambda_t & \text{if } t \in T(0) \\ \alpha_m(t) & \text{if } t \in T(m), m \in \{1, \dots, M\}. \end{cases}$$

If Z is ergodic, the unique stationary and limiting (product form !) distribution π of Z on E is for $z = [z_1, \dots, z_J] \in E$ with generic local states $z_j = [t_{j1}, s_{j1}; t_{j2}, s_{j2}; \dots; t_{jn_j}, s_{jn_j}]$, $j = 1, \dots, J$,

$$\pi(z_1, \dots, z_J) = K^{-1} \prod_{j=1}^J \left\{ \prod_{k=1}^{n_j} \left(\frac{\eta_j(t_{jk}, s_{jk})}{\mu_j(k)} \right) \right\}, \quad (4)$$

where $\mu_j(n) = \mu_j \cdot \min(n, s_j)$ and $K < \infty$ is the normalization constant.

Observing Networks at Embedded Time Points-Customer Stationary Distributions

The results on steady-state distributions of the product form presented in the section titled “Network Models and Steady-state Behavior” were extended to more complex networks especially in Baskett *et al.* [9] and Kelly [10]. From these probabilities, the main first-order performance indices of the networks can be obtained, for example, throughputs, mean queue lengths, and via Little’s theorem, mean passage times, see the section titled “Performance Measures and Algorithms” for algorithms and the article titled **Little’s Law and Related Results** in this encyclopedia for details. But it became apparent soon that, for obtaining passage time quantiles, somewhat different distributions were needed. Describing individual customer’s behavior on a route commences with observing “arrival distributions,” which describe the *network’s disposition* from the view of a customer who

enters a station—usually the disposition distributions are not in Equation (4). For open networks, early results of Strauch [11] and Wolff [12,13] were already available. These results stated that “Poisson arrivals see time averages (PASTA),” the celebrated PASTA property; or, as Strauch [11] puts it: When a queue looks the same to an arriving customer as to an observer, differentiating the “time stationary” and the “customer stationary distribution”.

For closed multi-type networks, the problem of determining arrival distributions was solved concurrently by Lavenberg and Reiser [14] and Sevcik and Mittrani [15].

There has been a rapid development of much more general results since then; an early example is given in Melamed [16], where for a general Markovian jump process $(X_t : t \geq 0)$ and an embedded sequence $(T_n : n \in \mathbb{N})$ of stopping times the distribution of X_{T_n+} and of X_{T_n-} are obtained. Melamed applied the results to open Jackson networks of queues, and to their generalization from Baskett *et al.* [9] and Kelly [10]. This approach could be used to prove results like Theorem 4 below.

Several extensions of the framework are obtained by Melamed himself, Whitt [17,18], Wolff [19], König [20], and Schmidt [21], who termed a generalized relation between time-stationary distributions of a process and the distribution at embedded time sequences “Embedded points see time averages” (EPSTA). Deviation of the Poisson property was also investigated under the heading of “Arrivals see time averages” (ASTA) in Melamed and Whitt [18]; Wolff [19] and Green and Melamed [22] proved for suitable systems an ANTI-PASTA property, which states conditions under which ASTA implies that the embedded sequence is Poisson. A similar result for EPSTA can be found in Green and Melamed [21].

On describing the network processes in terms of marked point processes, it is seen that the described results can be interpreted as results on Palm probabilities for the processes, for a survey see Serfozo [23] (Chapter 4). For more details on PASTA and related results, see the article titled **PASTA and Related Results** in this encyclopedia.

Sojourn Time and Passage Time Distributions for Individual Customers

Describing individual customer's behavior on a route commences with observing "arrival distributions," which determines the time-stationary behavior. These arrival distributions are distributions for the dispositions of the network at embedded time points, which in the present framework is often called *customer stationary distributions*. We describe the approach via limiting distributions similar to Sevcik and Mitrani [15]. For the approach via point processes, see Kook and Serfozo [24] and Bremaud *et al.* [25], and in more detail in Daduna and Szekli [26].

Denote by $A(t, s) = (A_n(t, s) : n \in \mathbb{N})$ the sequence of time instances when internal customer m being type t enters stage s of his route $r(t)$. Use a similar notation for the sequence of specific jump instances for external customers of type t . If the network's state just after such time instant $A_n(t, s)$ is $z \in E$, then we know that z shows at node $r(t, s) = j$ on position $n_{r(t, s)} = n_j$ the state description $(t_{jn_j}, s_{jn_j}) = (t, s)$. The disposition $d(z)$ of the other customers with respect to the (t, s) jump is obtained from z by deleting the pair $(t_{jn_j}, s_{jn_j}) = (t, s)$. Note that, if t is an external type, the disposition is a state in E , while if t is an internal type, $d(z) \notin E$. If t is an internal type, we denote the feasible set of dispositions by $E_{-t, s}$, which is a subset of the state space of the network when customer m with $t \in T(m)$ is deleted.

Theorem 4 [Arrival Theorem]. *Suppose the network process $Z = (Z(t) : t \geq 0)$ of the mixed network is ergodic. Let $t \in T(0)$ and $s \in \{1, \dots, S(t)\}$. Then the embedded process $Z^{t, s} = (d(Z(A_n(t, s))) : n \in \mathbb{N})$ is an ergodic Markov chain with unique limiting distribution*

$$\pi^{t, s}(z) := \lim_{n \rightarrow \infty} P(d(Z(A_n(t, s))) = z) = \pi(z), \\ z \in E.$$

Let $t \in T(m)$, $m \in \{1, \dots, M\}$, and $s \in \{1, \dots, S(t)\}$. Then the embedded process $Z^{t, s} = (d(Z(A_n(t, s))) : n \in \mathbb{N})$ is an ergodic

Markov chain with unique limiting distribution

$$\pi^{t, s}(z) := \lim_{n \rightarrow \infty} P(d(Z(A_n(t, s))) = z) \\ = \pi(z') \eta_{r(t, s)}(t, s)^{-1} \\ \times \mu_{r(t, s)}(n_{r(t, s)}(z) + 1) C^{-1}, \quad z \in E_{-t, s}, \quad (5)$$

where $n_{r(t, s)}(z)$ is the queue length of the disposition state at node $r(t, s)$ under z , and $z' \in E$ is obtained from $z \in E_{-t, s}$ by inserting the pair (t, s) at node $r(t, s)$ on position $n_{r(t, s)}(z) + 1$ in state z . C is the new normalization constant.

Limiting Versus Stationary Behavior. From Markov chain properties, it follows that the limiting distributions in Theorem 4 are equilibrium distributions of the embedded Markov chains as well. For more details, generalizations, and references, see the short overview in the section titled "Observing Networks at Embedded Time Points - Customer Stationary Distributions."

For open networks, the result is often loosely stated as follows: In equilibrium, an external arrival observes the other customers in the network in steady-state.

For closed networks, the result is often stated as follows: In equilibrium, a customer in a jump instant observes the other customers distributed according to the steady-state of that network, obtained by deleting himself from the population. This is not always correct [27, p. 144/145]. A consequence of this is the cumbersome notation of the limiting distribution (5).

Departure Theorems. We can write a similar framework for what customers observe at departure instants departing from the network or when jumping inside the network. These are covered by EPSTA and similar properties described in the section titled "Observing Networks at Embedded Time Points - Customer Stationary Distributions."

PASTA. While in the open network case customers entering the network from the exterior experience the consequences of the PASTA property and do the same when departing from the system, the jumps inside

the open network need not be Poissonian. Nevertheless, the dispositions seen by the jumping customer look like a PASTA property: This is visible from the formulae in Equation (5). Note that in nontrivial closed networks there are no Poissonian streams.

An immediate consequence of the Arrival Theorem and the strong Markov property of Z is

Theorem 5 [Asymptotic Sojourn Time Distribution]. *For the n th customer of type t traversing route $r(t)$, denote by $(T_1^{(n)}, \dots, T_{S(t)}^{(n)})$ the vector of his successive sojourn times at the nodes of $r(t)$. Then independently of the initial distribution of the system (with $\Theta_s \geq 0, s = 1, \dots, S(t)$),*

$$\lim_{n \rightarrow \infty} E[\exp(-\Theta_1 T_1^{(n)} - \dots - \Theta_{S(t)} T_{S(t)}^{(n)})] \\ = E_{\pi^{t,1}}[\exp(-\Theta_1 T_1^{(1)} - \dots - \Theta_{S(t)} T_{S(t)}^{(1)})]. \quad (6)$$

$E_{\pi^{t,1}}[\cdot]$ is expectation under $P_{\pi^{t,1}}[\cdot]$, the measure which describes the system conditionally on the first $(t, 1)$ jump at time 0 seeing the other customers' disposition distributed according to $\pi^{t,1}$.

In general, routes need more structure for Equation (6) leading to explicit results. Early results of Reich and Burke for linear systems were proved without explicitly introducing customer types.

Theorem 6. (a) [28,29] *In a linear (tandem) Jackson network of single server nodes, that is, $\lambda_1 > 0, \lambda_2 = \dots = \lambda_J = 0$, and $r(J, 0) = r(i, i+1) = 1, i = 1, \dots, J-1$ with $\lambda_1 < \mu_j, j = 1, \dots, J$, in steady-state the sojourn times of a customer traversing the tandem are independent and*

$$\lim_{n \rightarrow \infty} E[\exp(-\Theta_1 T_1^{(n)} - \dots - \Theta_J T_J^{(n)})] \\ = \prod_{j=1}^J \frac{\mu_j - \lambda_1}{\mu_j - \lambda_1 + \Theta_j}, \quad \Theta_j \geq 0, j = 1, \dots, J.$$

Moreover, with nodes 1 and J being multi-server stations, the independence still holds.

(b) [30] *In a cyclic Gordon–Newell network of single server nodes, that is,*

$r(J, 1) = r(i, i+1) = 1, i = 1, \dots, J-1$, *in steady-state the sojourn times of a customer traversing the cycle equal the limiting joint sojourn time vector, given by* ($\Theta_j \geq 0, j = 1, \dots, J$)

$$\lim_{n \rightarrow \infty} E[\exp(-\Theta_1 T_1^{(n)} - \dots - \Theta_J T_J^{(n)})] \\ = \sum_{(n_1, \dots, n_J) \in S(I-1, J)} \pi(I-1, J)(n_1, \dots, n_J) \\ \times \prod_{j=1}^J \left(\frac{\mu_j}{\mu_j + \Theta_j} \right)^{n_j+1}.$$

If the cycle commences at node 1, then with multiservers at nodes 1, J a similar result holds.

These results were extended to sojourn time computations on so-called *overtake-free* paths in open Jackson networks [8,31] and in closed Gordon–Newell networks [32]. A unification and generalization of these results needed the machinery of the notation invented for formulation of the Arrival Theorem. The main concept invented in Schassberger and Daduna [33] is that of a “quasi overtake-free” path (route). For a survey on this and related topics, see Boxma and Daduna [34] and Daduna [35].

Definition 5 [Quasi Overtake-Free Paths]. Consider a mixed network from Definition 4.

1. For nodes $i, j \in \{1, \dots, J\}$, write $i - (t) \rightarrow j$, if $i = r(t, s), j = r(t, s+1)$ for some $t \in \bigcup_{m=0}^M T(m)$, and $1 \leq s < S(t)$, or if $i = r(t, S(t)), j = r(t', 1)$ for some $m \in \{1, \dots, M\}$ with $t, t' \in T(m)$ and $p_m(t, t') > 0$.
2. For type t a t -relevant path $[j_1, \dots, j_K]$ from node j_1 to node j_K is a sequence of nodes (in general not a path or part of a single route) such that $j_k - (t_k) \rightarrow j_{k+1}, 1 \leq k < K$, holds with

$$t_k \in \bigcup_{m=0}^M T(m) \text{ if } t \in T(0)$$

and

$$t_k \in \bigcup_{m=0, m \neq \bar{m}}^M T(m) \text{ if } t \in T(\bar{m}),$$

$$1 \leq \bar{m} \leq M.$$

3. For $t \in \bigcup_{m=0}^M T(m)$, let $[r(t, u), \dots, r(t, v)]$ be a section of different successive nodes of route $r(t)$, $1 \leq u \leq v \leq S(t)$. This section is overtake-free for type t if every t -relevant path from $r(t, u')$ to $r(t, v')$ includes node $r(t, u' + 1)$, $u \leq u' < v' \leq v$.
4. For a customer of type $t \in \bigcup_{m=0}^M T(m)$, his route $r(t)$ is quasi overtake-free if the following holds: There exist u, v with $1 \leq u \leq v \leq S(t)$ such that the section $[r(t, u), \dots, r(t, v)]$ is overtake-free and

$$s_{r(t,1)} = \dots = s_{r(t,u-1)} = \infty,$$

$$s_{r(t,u+1)} = \dots = s_{r(t,v-1)} = 1,$$

$$s_{r(t,v+1)} = \dots = s_{r(t,S(t))} = \infty.$$

Theorem 7 [Sojourn Time Distribution on Quasi Overtake-Free Paths]. For customer of type t , let his route $r(t)$ be quasi overtake-free. For the n th customer of type t traversing route $r(t)$, let $(T_1^{(n)}, \dots, T_{S(t)}^{(n)})$ be the vector of his successive sojourn times at the nodes of $r(t)$. Then, independent of the initial distribution,

$$\lim_{n \rightarrow \infty} E[\exp(-\Theta_1 T_1^{(n)} - \dots - \Theta_{S(t)} T_{S(t)}^{(n)})]$$

$$\Theta_s \geq 0, s = 1, \dots, S(t),$$

$$= \left(\prod_{s=1}^{S(t)} \frac{\mu_{r(t,s)}}{\mu_{r(t,s)} + \Theta_s} \right)$$

$$\times \sum_{(n_{r(t,1)}, \dots, n_{r(t,S(t))})} \pi^{t,s}(n_{r(t,1)}, \dots, n_{r(t,S(t))})$$

$$\times \prod_{p=u}^v \left(\frac{\mu_{r(t,p)} s_{r(t,p)}}{\mu_{r(t,p)} s_{r(t,p)} + \Theta_p} \right)^{(n_{r(t,p)} + 1 - s_{r(t,p)})_+}, \quad (7)$$

where $(a)_+ = \max(a, 0)$, and $\pi^{t,1}(n_{r(t,1)}, \dots, n_{r(t,S(t))})$ is the probability under $P_{\pi^{t,1}}[\cdot]$, to see at time 0 the arriving type t customer at the

tail of the queue at node $r(t, 1)$, and having queue lengths vector $(n_{r(t,1)}, \dots, n_{r(t,S(t))})$ for the number of other customers present at the nodes of $r(t)$.

The steady-state distribution under $P_{\pi^{t,1}}[\cdot]$ of $(T_1^{(n)}, \dots, T_{S(t)}^{(n)})$ is given by Equation (7) as well.

Quasi overtake-free paths occur, for example, in flow shop systems and transmission lines. Note that any two-station route of successive distinct exponential multiserver queues is quasi overtake-free. Slight (part) generalizations of Theorem 7 are in Kook and Serfozo [24] and Daduna [36].

The problem of computing more passage times on more general routes seems to be challenging as is highlighted by the examples: The Simon–Foley network [37] and Burke’s tandem [38].

PERFORMANCE MEASURES AND ALGORITHMS

The success of Jackson networks in applications is due to having access to steady-state distributions and the fact that the main performance indices of the networks can be directly derived from these. We assume therefore throughout this section that all systems are in steady-state. Although explicit formulas are at hand, computational difficulties occurred when evaluating the explicit formulas. A short survey of the extensive bibliography on algorithmic and computational aspects of queueing network performance analysis is in Kant [39]; more details can be found in Bolch *et al.* [40] and Bruell and Balbo [41].

The most important system-oriented performance index is throughput, while the most important customer-oriented index is sojourn time (travel time) of customers.

Proposition 1. For the open Jackson network, the total arrival rate is $\lambda = \sum_{j=1}^J \lambda_j$. From stationarity, it follows that the overall throughput TH , that is, the mean number of departures from the system per time unit, equals the mean number of arrivals at the system per time unit, so $TH = \lambda$.

The mean number of departures per time unit from node j (node- j throughput) is $TH_j = \eta_j$.

For the Gordon–Newell network, arrival and departure rates at any node are identical in equilibrium. The mean number of departures per time unit from node j (node- j throughput) in the network with $I \geq 1$ customers is

$$\begin{aligned} TH_j(I) &= \sum_{(n_1, \dots, n_J) \in S(I, J)} \pi(n_1, \dots, n_J) \cdot \mu_j(n_j) \\ &= \eta_j \cdot \frac{G(I-1, J)}{G(I, J)}, \end{aligned} \quad (8)$$

where η_j is the customers visit ratio at node j , that is, the probability solution of the traffic equation (2). The (overall) throughput of the network is the sum of the local throughputs

$$TH(I) = \sum_{j=1}^J TH_j(I) = \frac{G(I-1, J)}{G(I, J)}. \quad (9)$$

From Proposition 1, the need for computing normalization constants efficiently is obvious. We demonstrate typical techniques for closed Gordon–Newell networks with single servers with state-independent service rates $\mu_j, j = 1, \dots, J$. From Theorem 2, we have with normalization constant $G(I, J)$

$$\begin{aligned} \pi(I, J)(n_1, \dots, n_J) &= G(I, J)^{-1} \prod_{j=1}^J \left(\frac{\eta_j}{\mu_j} \right)^{n_j}, \\ (n_1, \dots, n_J) &\in S(I, J). \end{aligned} \quad (10)$$

The numerical problems with Equation (10) arise from the complexity of $G(I, J)$. Similar sums occur when computing Gibbs measures in statistical physics, where normalization constants are called *partition functions*. In statistical mechanics, from the partition function many observables of the systems can be computed. A similar program has been developed in network theory.

Theorem 8 [Buzen’s Algorithm [42]]. For a closed Gordon–Newell network with J single server nodes and state-independent service rates $\mu_j, j = 1, \dots, J$

we denote for different population sizes by $1 =: G(0, J), G(1, J), \dots, G(I, J), \dots$ the normalization constants of the respective steady-state distributions $\pi(I, J)$. These normalization constants are computed iteratively as follows. Denote

$$x_j = \frac{\eta_j}{\mu_j}, \quad j = 1, \dots, J,$$

$$\text{and } G(i, j) = \sum_{\substack{(n_1, \dots, n_J) \in \mathbb{N}^J \\ n_1 + n_2 + \dots + n_J = i}} \prod_{k=1}^j x_k^{n_k}$$

$$\text{Then } G(i, 1) = x_1^i, \quad i \geq 0,$$

$$G(0, j) = 1, \quad j \geq 0,$$

$$\text{and } G(i, j) = G(i, j-1) + x_j \cdot G(i-1, j), \\ j \geq 2, i \geq 1.$$

Given $x_j = \frac{\eta_j}{\mu_j}, I, J$, to compute $G(1, J), \dots, G(I, J)$, we need $I \cdot J$ additions and $I \cdot J$ multiplications.

Corollary 2. For a Gordon–Newell network from Theorem 8 with I customers denote by $X = (X_1, \dots, X_J)$ a vector distributed like $\pi(I, J)$, that is, X behaves as the network’s steady state. Then

$$\begin{aligned} P(X_j \geq m_j) &= \left(\frac{\eta_j}{\mu_j} \right)^{m_j} \cdot \frac{G(I-m_j, J)}{G(I, J)}, \\ 0 \leq m_j \leq I, \quad j &= 1, \dots, J, \end{aligned} \quad (11)$$

$$\begin{aligned} P(X_j = m_j) &= \left(\frac{\eta_j}{\mu_j} \right)^{m_j} \cdot \left\{ \frac{G(I-m_j, J)}{G(I, J)} - \frac{\eta_j}{\mu_j} \right. \\ &\quad \times \left. \frac{G(I-m_j-1, J)}{G(I, J)} \right\}, \\ 0 \leq m_j \leq I, \quad j &= 1, \dots, J, \end{aligned}$$

$$E[X_j] = \sum_{m_j=1}^I \left(\frac{\eta_j}{\mu_j} \right)^{m_j} \cdot \frac{G(I-m_j, J)}{G(I, J)}.$$

Theorem 9 [Mean Value Analysis, [43]]. We investigate the Gordon–Newell network with J single server nodes and state-independent service rates μ_j with different population sizes I . Fix $\eta = (\eta_1, \dots, \eta_J)$, the unique probability solution of the traffic equation (2). We denote

- $X(I) = (X_1(I), \dots, X_J(I))$, a vector which has distribution $\pi(I, J)$ from Equation (3). $E[X_j(I)]$ is the respective mean local queue length at node j with population size I .
- $T_j(I)$ a random variable which is distributed according to the stationary distribution of the sojourn time of a customer at node $j, j = 1, \dots, J$. $E[T_j(I)]$ is the respective mean sojourn time at node j .
- $TH(I)$ is the stationary throughput of the network. The MVA is

$$E[X_j(0)] = 0, \quad j = 1, \dots, J,$$

and for $I \geq 1$

$$E[T_j(I)] = \mu_j^{-1} (1 + E[X_j(I-1)]), \quad j = 1, \dots, J,$$

$$TH(I) = \frac{I}{\sum_{j=1, \dots, J} \eta_j E[T_j(I)]} \quad (12)$$

$$E[X_j(I)] = \eta_j TH(I) \cdot E[T_j(I)], \quad j = 1, \dots, J. \quad (13)$$

Equation (13) can be written as $E[X_j(I)] = TH_j(I) \cdot E[T_j(I)], j = 1, \dots, J$. This is Little's equation for each node in the network, for details see El-Taha and Stidham [44]. A similar interpretation can be given to Equation (12) by defining the overall mean sojourn time in a queue of the network to be

$$E[T(I)] := \sum_{j=1, \dots, J} \eta_j E[T_j(I)],$$

which yields $I = TH(I) \cdot E[T(I)]$.

MVA yields recursively the main performance measures without probabilities and normalization constants. These can be obtained via normalization constants from Equation (9) and $G(I, J) = \prod_{k=1}^I TH(K)^{-1}$. From Corollary 2, the probabilities are obtained.

EXTENSION OF THE MODELS: ADDITIONAL FEATURES

Model 1 [State-Dependent Arrivals [6]].

If in the Jackson network of Definition

2 there is a total population of size $k = n_1 + \dots + n_J$ present, the arrival intensity is $\lambda(k) \geq 0$. An arrival enters node j with probability $r(0, j)$. All other properties of the network remain the same. If the joint queue length process is ergodic, then its steady-state distribution is

$$\pi(n_1, \dots, n_J) = G^{-1} \prod_{k=1}^{n_1 + \dots + n_J - 1} \lambda(k) \prod_{j=1}^J \left(\prod_{k=1}^{n_j} \frac{\eta_j}{\mu_j(k)} \right) \quad (n_1, \dots, n_J) \in E, \quad (14)$$

where $\mu_j(k) = \mu_j \min(k, s_j)$ in the case of node being a multiserver with s_j channels, η_j is the solution of the traffic equation (1) with λ_j substituted by $r(0, j)$, and G is the normalization constant.

The result (14) holds as well when μ_j are of general form.

If for some $K_0 : \lambda(k) = 0 \forall k \geq K_0$, Equation (14) remains valid for the finite set of feasible states.

For Jackson networks with different customer types where the arrival intensities depend on the number of types already present, a similar result can be proven [10, Section 3.5].

Model 2 [General Service Disciplines, Nonexponential Service Times [9,10]]. Driven by needs in computer science applications around 1975, several classes of generalizations of Jackson and Gordon–Newell networks were developed which still exhibit product-form steady-state. The additional features embedded into the networks were customer classes and types, type-dependent routing, general service disciplines [10], which encompass especially processor sharing, last-come first-served preemptive resume, infinite servers [9] (which are special cases of so-called *symmetric servers*), and type-dependent and nonexponential service time requests at symmetric servers.

Model 3 [Finite Waiting Rooms [6,45]].

If in the Jackson network of

Definition 2 at some nodes have finite waiting room, it must be specified how blocking of customers due to full waiting space is resolved. Jackson proved that the following rules yield a product-form steady-state again (even for state-dependent arrivals as in Model 1): Whenever a customer on his itinerary encounters a station where all service and waiting places are occupied, he jumps over this station in the direction specified by the routing matrix as if he has been served there (“jump-over protocol,” skipping). If the next destination node is full, the protocol applies again—until he will find a node with waiting or service places available or until he will leave the network.

The same protocol was investigated by Pittel [45] in closed networks, obtaining the product-form steady state again. The jump-over protocol has been rediscovered several times in the literature. If in a Gordon–Newell network routing is reversible, another protocol to handle blocking is available which maintains product-form steady-state: A customer on leaving node i finding his destination node full stays on at node i to obtain another service there (“transmission blocking” $\equiv$ transmission is thought to be not successful), for details, see van Dijk [46] and the references there.

Model 4 [Nonergodic Jackson Networks [47]]. If in the Jackson network of Definition 2 there exist η_j from the solution of the traffic equation (1) with $\eta_j \geq \mu_j$, then the network’s joint queue length process is not ergodic, and no steady-state exists. Goodman and Massey proved for networks of single server stations that in this case the generalized traffic equation

$$\eta_j = \lambda_j + \sum_{i=1}^J \min(\eta_i, \mu_i) r(i, j), \quad j = 1, \dots, J,$$

has a unique solution such that $\eta_j < \mu_j, j \in U \subseteq \{1, \dots, J\}$, $\eta_j \geq \mu_j, j \in U^c \subseteq \{1, \dots, J\}$ that is, nodes in U are “stable” whereas nodes in U^c are not. Then, independent of

the initial distribution

$$\begin{aligned} \pi_U(n_i, i \in U) &:= \lim_{t \rightarrow \infty} P(X_i(t) = n_i; i \in U) \\ &= \prod_{i \in U} \left(1 - \frac{\eta_i}{\mu_i}\right) \left(\frac{\eta_i}{\mu_i}\right)^{n_i}, \\ &\quad (n_i, i \in U) \in \mathbb{N}^{|U|} \end{aligned}$$

$$\begin{aligned} \pi_{U^c}(n_j, j \notin U) &:= \lim_{t \rightarrow \infty} P(X_j(t) = n_j; j \notin U) = 0, \\ &\quad (n_j, j \notin U) \in \mathbb{N}^{|U^c|}. \end{aligned}$$

Model 5 [Networks with Unreliable Servers]. Servers in the Jackson or Gordon–Newell network can be unreliable, that is, they break down randomly and are repaired (usually a random time). This is a problem from the intersection of performance analysis and reliability.

Such problems have found much interest in the literature, but results similar to those of Jackson *et al.* are rare. A folk approach to obtain explicit distributional results is the “reduced workload approximation,” which decouples the performance from the reliability part. For any node i , the portion $\alpha_i \in (0, 1]$ of time the node is functioning (available) is determined often approximately without considering details of the load-dependent service mechanism. Thereafter, the node’s service rate μ_i , say, is reduced to $\alpha_i \cdot \mu_i$. With this reduced service rate for all stations, the network is investigated with all other characteristics unchanged, see for related approximations [48].

The problem with this procedure is that it is not possible to model jointly the availability and performance behavior of the network. Systems with product-form steady state to jointly investigate these quantities in an integrated model are developed in Sauer and Daduna [49]. The overall breakdown-repair processes are modeled as multidimensional birth-death processes, the rates of which may depend on the queue length vector. Because nodes that are down do not accept new customers, regulations for redirecting customers which want to enter a broken-down station are set in force: “jump-over protocol” (skipping) and, in case of reversible routing, “transmission blocking,” see Model 3.

Model 6 [Bottleneck Analysis]. It was already observed by Gordon and Newell that in a closed network under heavy load (large number of customers) bottlenecks occur whenever there are different utilizations η_i/μ_i (we sketch only the single server case of Theorem 2). This means that almost all customers queue up at the nodes with highest utilization. There is an enormous bulk of literature on how to deal with asymptotic analysis of a closed network with fixed number J of nodes and routing behavior when the population size $I \rightarrow \infty$ [50].

More detailed studies are in Malyshev and Yakovlev [51] and Daduna *et al.* [52] where, for example, asymptotic throughput under $J \rightarrow \infty$ and $I \rightarrow \infty$ is investigated, concentrating on the existence of critical customer densities determined by the bottlenecks and queue length behavior without rescaling of time and queue lengths.

An observation of Gordon and Newell [5]: If exactly one bottleneck exists, under $I \rightarrow \infty$ the joint queue length vector of the non-bottleneck nodes converges in distribution to the distribution of the joint queue length vector of an open Jackson network, which is fed by a single Poisson stream originating from the bottleneck.

Model 7 [Infinite Routing Graphs]. The extension of the product-form theory of Jackson and Gordon–Newell networks to routing on infinite graphs needs new techniques from functional analysis of Markov processes as it is needed, for example, for Markov interacting processes which describe particle systems on infinite lattice graphs [53]. Some early contributions from the queueing network point of view are described in Karpelevich *et al.* [54] and the references therein.

Model 8 [Discrete Time Networks]. Building product-form networks in discrete-time has a long history [55], but the construction of suitable models was less successful than in continuous-time. The direct approach by substituting “Poisson arrival streams $\rightarrow$ Bernoulli arrival streams,” “exponential service times $\rightarrow$ geometric service times,” maintaining the routing procedures and the usual independence

assumptions, for single server nodes was successful only for linear networks (closed cycles and open tandems).

A network with general routing was constructed by Walrand [56] with a new class of servers (S-queues with batch service and batch arrivals). A variety of applications can be found in Woodward [57]. For more models and further development of the theory, see Refs 58, 59 and 60.

REFERENCES

1. Jackson JR. Networks of waiting lines. *Oper Res* 1957;5:518–521.
2. Melamed B. Characterisation of Poisson streams in Jackson queueing networks. *Adv Appl Probab* 1979;11:422–438.
3. Disney JR, König D. Queueing networks: a survey of their random processes. *SIAM Rev* 1985;27:335–403.
4. Disney JR, Kiessler PC. Traffic processes in queueing networks: a Markov renewal approach. London: The Johns Hopkins University Press; 1987.
5. Gordon WJ, Newell GF. Closed queueing networks with exponential servers. *Oper Res* 1967;15:254–265.
6. Jackson JR. Jobshop-like queueing systems. *Manage Sci* 1963;10:131–142.
7. Daduna H, Szekli R. On the correlation structure of closed queueing networks. *Commun Stat Stoch Models* 2004;20:1–31.
8. Melamed B. Sojourn times in queueing networks. *Math Oper Res* 1982;7:223–244.
9. Baskett F, Chandy M, Muntz R, *et al.* Open, closed and mixed networks of queues with different classes of customers. *J ACM* 1975;22:248–260.
10. Kelly FP. Reversibility and stochastic networks. Chichester; New York; Brisbane; Toronto: John Wiley & Sons; 1979.
11. Strauch RE. When a queue looks the same to an arriving customer as to an observer. *Manage Sci* 1970;17:140–141.
12. Wolff RW. Poisson arrivals see time averages. Berkely (CA): Department of Operations Research, University of California; 1976.
13. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
14. Lavenberg SS, Reiser M. Stationary state probabilities at arrival instants for closed

- queueing networks with multiple types of customers. *J Appl Probab* 1980;17:1048–1061.
15. Sevcik KC, Mitrani I. The distribution of queueing network states at input and output instants. *J ACM* 1981;28:358–371.
 16. Melamed B. On Markov jump processes embedded at jump epochs and their queueing-theoretic applications. *Math Oper Res* 1982;7:111–128.
 17. Melamed B, Whitt W. On arrivals that see time averages. *Oper Res* 1990;38:156–172.
 18. Melamed B, Whitt W. On arrivals that see time averages: a Martingale approach. *J Appl Probab* 1990;27:376–384.
 19. Wolff RW. A note on PASTA and ANTI-PASTA for continuous-time Markov chains. *Oper Res* 1990;38:176–177.
 20. König D, Schmidt V. Imbedded and non-imbedded stationary characteristics of queueing systems with varying service rate and point processes. *J Appl Probab* 1980;17:753–767.
 21. König D, Schmidt V. EPSTA: the coincidence of time-stationary and customer-stationary distributions. *Queueing Syst Appl* 1989;5:247–264.
 22. Green L, Melamed B. An ANTI-PASTA result for Markovian Systems. *Oper Res* 1990;38:173–175.
 23. Serfozo R. Introduction to stochastic networks. New York: Springer; 1999.
 24. Kook KH, Serfozo RF. Travel and sojourn times in stochastic networks. *Ann Appl Probab* 1993;3:228–252.
 25. Bremaud P, Kannurpatti R, Mazumdar R. Event and time averages: a review. *Adv Appl Probab* 1992;24:377–411.
 26. Daduna H, Szekli R. Conditional job observer property for multitype closed queueing networks. *J Appl Probab* 2002;39:865–881.
 27. Walrand J. An introduction to queueing networks. Englewood Cliffs: Prentice–Hall International; 1988.
 28. Reich E. Note on queues in tandem. *Ann Math Stat* 1963;34:338–341.
 29. Burke PJ. The output process of a stationary M/M/s queueing system. *Ann Math Stat* 1968;39:1144–1152.
 30. Boxma OJ, Kelly FP, Konheim AG. The product form for sojourn time distributions in cyclic exponential queues. *J ACM* 1984;31:128–133.
 31. Walrand J, Varaiya P. Sojourn times and the overtaking condition in Jacksonian networks. *Adv Appl Probab* 1980;12:1000–1018.
 32. Daduna H. Passage times for overtake-free paths in Gordon–Newell networks. *Adv Appl Probab* 1982;14:672–686.
 33. Schassberger R, Daduna H. Sojourn times in queueing networks with multiserver nodes. *J Appl Probab* 1987;24:511–521.
 34. Boxma OJ, Daduna H. Sojourn times in queueing networks. In: Takagi H, editor. Stochastic analysis of computer and communication systems. Amsterdam: IFIP, North–Holland Publishing Co.; 1990. pp. 401–450.
 35. Daduna H. Stochastic networks with product form equilibrium. In: Shanbhag DN, Rao CR, editors. Stochastic processes: theory and methods. Volume 19, Handbook of statistics. Amsterdam: Elsevier Science; 2001. Chapter 11. pp. 309–364.
 36. Daduna H. Sojourn time distributions in non-product form queueing networks. In: Dshalalow J, editor. Frontiers in queueing: models and applications in science and engineering. Boca Raton (FL): CRC Press; 1997. Chapter 7. pp. 197–224.
 37. Simon B, Foley RD. Some results on sojourn times in acyclic Jackson networks. *Manage Sci* 1979;25:1027–1034.
 38. Burke PJ. The dependence of sojourn times in tandem M/M/s queues. *Oper Res* 1969;17:754–755.
 39. Kant K. Steady state analysis of stochastic systems. In: Rao CR, editor. Computational statistics. Volume 9, Handbook of statistics. Amsterdam: North Holland Publishing Co.; 1993. Chapter 2. pp. 17–68.
 40. Bolch G, Greiner S, de Meer H, *et al.* Queueing networks and Markov chains. New York: John Wiley & Sons; 1998.
 41. Bruell SC, Balbo G. Computational algorithms for closed queueing networks. New York: North–Holland Publishing Co.; 1980.
 42. Buzen B. Computational algorithms for closed queueing networks with exponential servers. *Commun ACM* 1973;16:527–531.
 43. Reiser R, Lavenberg SS. Mean value analysis of closed multichain queueing networks. *J ACM* 1980;27:313–322.
 44. El-Taha M, Stidham S Jr. Sample-path analysis of queueing systems. Boston (MA): Kluwer Academic Publisher; 1999.
 45. Pittel B. Closed exponential networks of queues with saturation: the Jackson-type stationary distribution and its asymptotic analysis. *Math Oper Res* 1979;4:357–378.

46. van Dijk NM. Queueing networks and product forms—a systems approach. Chichester: John Wiley & Sons; 1993.
47. Goodman JB, Massey WA. The non-ergodic Jackson network. *J Appl Probab* 1984;21:860–869.
48. Chakka R, Mittrani I. Approximate solutions for open networks with breakdowns and repairs. In: Kelly FP, Zachary S, Ziedins I, editors. *Stochastic networks, theory and applications*. Volume 4, Royal statistical society lecture notes series. Oxford: Clarendon Press; 1996. Chapter 16. pp. 267–280.
49. Sauer C, Daduna H. Availability formulas and performance measures for separable degradable networks. *Econ Qual Contr* 2003;18:165–194.
50. Berger A, Bregman L, Kogan Y. Bottleneck analysis in multiclass closed queueing networks and its application. *Queueing Syst Appl* 1999;31:217–237.
51. Malyshev VA, Yakovlev AV. Condensation in large closed Jackson networks. *Ann Appl Probab* 1996;6:92–115.
52. Daduna H, Pestien V, Ramakrishnan S. Throughput limits from the asymptotic profile of cyclic networks with state-dependent service rates. *Queueing Syst Appl* 2008;58:191–219.
53. Liggett TM. Interacting particle systems. Volume 276, *Grundlehren der mathematischen Wissenschaften*. Berlin: Springer; 1985.
54. Karpelevich FI, Pechersky EA, Suhov YM. Dobrushin’s approach to queueing network. *J Appl Math Stoch Anal* 1996;9:373–397.
55. Hsu J, Burke PJ. Behaviour of tandem buffers with geometric input and markovian output. *IEEE Trans Commun* 1976;24:358–361.
56. Walrand J. A discrete-time queueing network. *J Appl Probab* 1983;20:903–909.
57. Woodward ME. *Communication and computer networks: modelling with discrete-time queues*. Los Alamitos (CA): IEEE Computer Society Press; 1994.
58. Bruneel H, Kim BG. *Discrete-time models for communication systems including ATM*. Boston (MA): Kluwer Academic Publications; 1993.
59. Chao X, Miyazawa M, Pinedo M. *Queueing networks—customers, signals, and product form solutions*. Chichester: John Wiley & Sons; 1999.
60. Daduna H. Queueing networks with discrete time scale: explicit expressions for the steady state behavior of discrete time stochastic networks. Volume 2046, *Lecture notes in computer science*. Berlin: Springer; 2001.

JOB SHOP SCHEDULING

MARC E. POSNER
Integrated Systems Engineering,
Ohio State University,
Columbus, Ohio

In a sequential scheduling problem, jobs must complete a set of tasks or operations. Typically, each task is completed using a specific type of machine or workstation. In the job shop scheduling problem, each job has its own specific set and sequence of tasks that are required for completion. The job shop scheduling problem is one of the three classic sequential multimachine scheduling problems. The requirement regarding the order in which tasks are completed is the distinction between the job shop problem and the other two classic sequential problems, the flow shop and the open shop. In the open shop environment, the tasks for a given job can be completed in any order, and in the flow shop environment, all jobs complete their tasks in the same order.

While numerous variations exist, for the classic job shop problem, the jobs are processed subject to the following conditions:

1. A machine can process only one job at a time.
2. The sequence of machines that a job visits is specified.
3. A job cannot visit a machine more than once.
4. The processing of a job on a machine cannot be interrupted.
5. The input buffer is infinite for each machine.
6. The processing time of a job on a given machine is known.
7. The objective is to minimize the time that is required to complete the last job (the makespan).

The decision for this problem is to specify the order in which the jobs are processed on

each machine. Suppose that there is an idle machine and the next job to be processed by that machine is available. There is no benefit to delaying the processing of that job if the objective is to minimize the makespan, Condition 7. As a result, once the job processing order is decided for each machine, each job is processed as soon as possible, subject to the task order requirements and machine scheduling decisions. Thus, the processing order for each machine determines the exact starting time and completion time of each task.

Many real-world production and service environments can be modeled as job shop scheduling problems. The most typical environments are production shops, where different types of jobs require different treatments. One example is semiconductor wafer manufacturing [1]. Also, repair facilities are usually job shops because different breakdowns require different tools and operations. In the service sector, a common example is the execution of consulting engagements in a management consulting firm.

The job shop scheduling problem is one of the hardest problems in operations research. Almost all versions of this problem are intractable [2]. A 10-machine, 10-job problem was posed by Muth and Thompson [3]. The optimal solution was not verified until Carlier and Pinson [4] solved the problem using a variety of sophisticated mathematical programming techniques and nearly five hours of computer time. While a wide variety of optimization approaches have been considered, none are able to find optimal solutions to large problems. Two such methods are Branch and Bound [5] and Lagrangian Relaxation [6].

Not only is it hard to find optimal solutions, but it is also difficult to generate good quality approximate solutions. The result is that an important area of research has arisen which is based on a procedure that, in the worst case, takes an exponential amount of time (based on the input size) to solve a job shop scheduling problem. This proce-

ture, called the *shifting bottleneck heuristic*, was originally developed by Adams *et al.* [7]. This research area is one of the few in the operations research literature that studies an exponential-time heuristic.

A wide variety of variations on the classic job shop problem have been considered. These variations allow for changes in the objective function, in the constraints and job restrictions, or in the processing approaches. One example is minimizing the work-in-process (WIP) time [5]. By adding due dates to the jobs (suggested completion times), several different objectives have been considered. These include minimizing the sum of lateness [8], minimizing the average cost of early completion (because jobs are held in inventory), late completion, or number of late jobs [9], minimizing the weighted tardiness (Vepsäläinen and Morton), and the maximum tardiness (maximum {job lateness, 0}) [6].

Also, a variety of constraints and restrictions have been considered. These include job deadlines [11], jobs interruption, and later resumption on a machine (preemption) [12], all jobs with unit processing times [13] or with stochastic processing times [14], controlled processing times [15], sequence-dependent setup times [16], machine availability times [17], limited machine buffers [18], insertion of late arriving jobs into the schedule [19], and addition of new tasks for specific jobs [20]. Dauzère-Pérés and Lasserre [21] consider a job shop problem where jobs can be split (lot streaming) to reduce the makespan.

One variation in processing is that a task can be scheduled from a set of flexible machines, which allows for a job to follow multiple routes through the shop [22]. Another variation is considered by Mason

et al. [23] where a job can return to a machine that it previously visited (reentrant job shop).

FORMULATION

To formulate the job shop problem, we first specify the set of problem inputs (parameters). In the classical version of this problem, there are m machines, $M_1, M_2, \dots, M_m$, and n jobs, $1, 2, \dots, n$. Each job $j \in N = \{1, 2, \dots, n\}$ has a set of m tasks, $T_{1j}, T_{2j}, \dots, T_{mj}$, where $T_{ij} = k$ specifies that the k th task for job j is on machine M_i . This vector, $(T_{1j}, T_{2j}, \dots, T_{mj})$, describes the machine processing order for job j . Also, the processing time for job j on machine M_i is denoted by $p_{ij} \geq 0$. The situation where job j is not processed on machine M_i can be modeled by setting $p_{ij} = 0$ and making M_i the first (or last) machine that job j visits. The most common objective is to minimize the makespan.

The problem can be formulated as a mathematical programming problem. The parameters are the machine order and the processing time for each job on each machine. Let

$$y_{ij} =$$

completion time of job j on machine M_i

$$x_{ijk} =$$

$$\begin{cases} 1, \text{ job } j \text{ is processed before job } k \\ \quad \text{on machine } M_i \\ 0, \text{ job } k \text{ is processed before job } j \\ \quad \text{on machine } M_i. \end{cases}$$

One standard mixed integer formulation is

$$\begin{aligned} & \text{minimize } C_{\max} \\ & \text{sub. to } y_{ij} \leq C_{\max}, & T_{ij} = m; i \in \{1, 2, \dots, m\}; j \in N, & (1) \\ & y_{ij} + p_{lj} \leq y_{lj}, & T_{ij} < T_{lj}; i, l \in \{1, 2, \dots, m\}; j \in N, & (2) \\ & y_{ij} + p_{ik} \leq y_{ik} + Lx_{ijk}, & i \in \{1, 2, \dots, m\}; j, k \in N, & (3) \\ & y_{ik} + p_{ij} \leq y_{ij} + L(1 - x_{ijk}), & i, j \in \{1, 2, \dots, m\}; j \in N, & (4) \\ & y_{ij} \geq 0, & i \in \{1, 2, \dots, m\}; j \in N, \\ & x_{ijk} \in \{0, 1\}, & i \in \{1, 2, \dots, m\}; j, k \in N, \end{aligned}$$

where $L \geq \sum_{i=1}^m \sum_{j \in N} p_{ij}$.

Constraint (1) ensures that $C_{\max}$ is no smaller than the completion time of the last task for each job j . The task processing order for each job j is modeled by constraint (2). Constraints (3) and (4) model the requirement that jobs j and k are not processed simultaneously on machine M_i . If $x_{ijk} = 1$, then constraint (3) requires that job j completes before job k starts. Further, constraint (3) holds for any reasonable values of y_{ij} and y_{ik} . The reverse is true if $x_{ijk} = 0$. These types of constraints were first proposed by Manne [24].

An alternative formulation is as a graph problem with two types of directed arcs: conjunctive (standard) and disjunctive [25]. The conjunctive arcs model the processing order for the jobs, while the disjunctive arcs represent the possible job schedules for the machines. An appropriate selection from the set of disjunctive arcs provides a specific machine processing schedule. Then, the remaining disjunctive arcs are removed from the graph. In the resulting graph, the longest path is equal to the makespan corresponding to the machine processing schedule. This concept is illustrated by the following example.

Example 1. There are three machines and three jobs. The machine processing order for Job 1 is M_1, M_2, M_3 , for Job 2 is M_2, M_3, M_1 , and for Job 3 is M_2, M_3 . The processing times are $p_{11} = 3, p_{21} = 5, p_{31} = 1, p_{12} = 2, p_{22} = 4, p_{32} = 4, p_{23} = 4$, and $p_{33} = 2$. The objective is to minimize the makespan.

The disjunctive graph associated with Example 1 is presented in Fig. 1. Each node (i, j) corresponds to the operation for job j on machine M_i . The three solid paths from source S to sink T represent the machine order for each of the jobs. As an example, $S - (2, 2) - (3, 2) - (1, 2) - T$ shows that job 2 is first processed on M_2 , then on M_3 , and finally on M_1 . The sets of dotted arcs form cliques (subgraphs where every pair of nodes is connected by an arc). Each clique corresponds to a machine. Moreover, a selection of a subset of arcs that form a path through the clique is a decision

regarding the job processing order on that machine. The number next to an arc is the length of the arc which corresponds to the processing time of the job at the tail of the arc. For instance, because $p_{23} = 4$, the arcs $((2, 3), (2, 2))$, $((2, 3), (2, 1))$, and $((2, 3), (3, 3))$ all have a distance of four.

Suppose we process the jobs on M_1 in the order $1 \rightarrow 2$, on M_2 in the order $2 \rightarrow 3 \rightarrow 1$, and on M_3 in the order $2 \rightarrow 3 \rightarrow 1$. Then, the resulting graph is shown in Fig. 2. The longest path in this graph is represented by a heavy line. The sum of the distances on this path, which is 12, is the makespan of the schedule corresponding to our job ordering decisions. The processing schedule for Example 1 is presented in the Gantt chart in Fig. 3. Note that for the given job ordering, no job can start earlier without violating some constraint.

AN EASILY SOLVED CASE

One of the few versions of this problem where an optimal solution can be found quickly is when there are only two machines, M_1 and M_2 , and the objective is to minimize the makespan [26]. For this problem, there are four types of jobs:

- A: jobs that are processed only on M_1 ;
- B: jobs that are processed only on M_2 ;
- G: jobs that are first processed on M_1 and then on M_2 ;
- H: jobs that are first processed on M_2 and then on M_1 .

The makespan is minimized when the processing time on the bottleneck machine is minimized. Thus, it is important to start both machines as quickly as possible and to prevent idle time caused by one machine waiting for the other to finish a specific job. Consequently, an optimal solution is to process on M_1 first the G jobs, then the A jobs, and finally the H jobs. On M_2 , the H jobs should be processed first, followed by the B jobs, and finally the G jobs (Fig. 4). The only issue that still needs resolution is how the jobs are processed within a given job type. Since the A jobs are only processed on

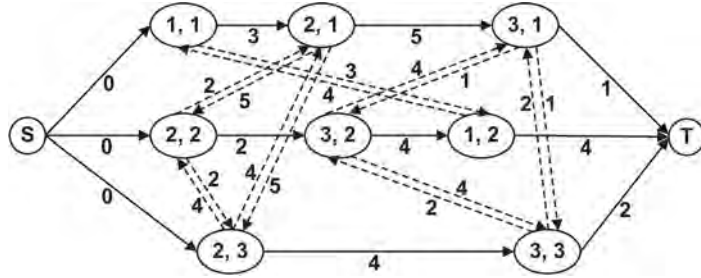


Figure 1. Disjunctive graph for Example 1.

one machine, the order does not affect the completion time. Consequently, any order is optimal. The same is true for the B jobs. However, order may be important for the G and H jobs. As an example, supposed that all jobs were type G . Then, the problem is a two machine flow shop problem, where the optimal solution is to use Johnson's Algorithm [27]. The order generated by Johnson's Algorithm is optimal for the G and H jobs, where the first machine for the H jobs is M_2 .

Another two-machine problem that can be solved quickly is a stochastic version where the processing time of a job on a machine is generated from a unique exponential distribution, and the objective is to minimize the expected completion time [14].

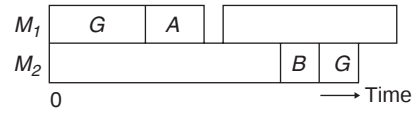


Figure 4. Optimal job order for two-machine makespan problem.

Surprisingly, when a job can be interrupted and then resumed later, called *preemption*, the problem is intractable [28].

HEURISTICS

Because the job shop scheduling problem is so difficult to solve, there is a large literature that examines approximate methods, called

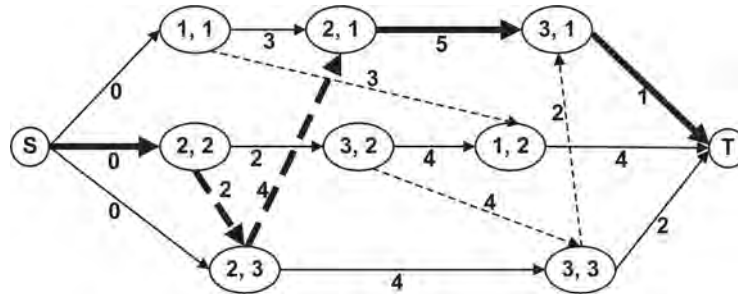


Figure 2. Graph for Example 1 with scheduling decisions.

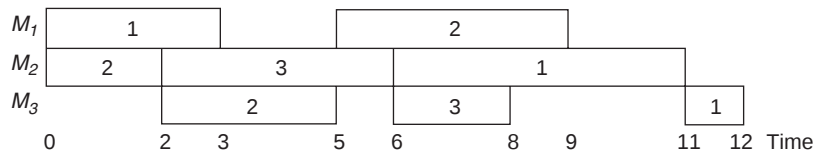


Figure 3. Gantt chart for Example 1.

heuristics, that find good but not necessarily optimal solutions. The most important of these methods is the *shifting bottleneck heuristic* [7]. This heuristic finds the machine that is most likely to become the bottleneck in a current partial schedule. Each job on an unscheduled machine is assigned a release date and due date (desired completion time). Then, each of these one-machine scheduling problems is solved where the objective is to minimize the job tardiness. While this one-machine problem is intractable (see **Single Machine Scheduling**), in general, solutions can be found rapidly. The machine with the maximum tardiness is in some sense, the bottleneck machine. As a result, this machine is scheduled. All of the previously scheduled machines are now rescheduled based on the new bottleneck machine schedule. Then, the process is repeated, each time scheduling one new machine, until all machines are scheduled.

The shifting bottleneck heuristic has been extended to solve other versions of the job shop problem. As examples, Balas *et al.* [29] include sequence-dependent setup times (setup time before processing begins depends on the prior task), release dates, and deadlines, and Ivens and Lambrecht [30] consider release and due dates, assembly and splitting of jobs, overlapping processing times, setup times, transportation times, and multiple machines at a work station. As a final example, Mason *et al.* [23] minimize weighted tardiness where constraints include multiple and batching machines (several tasks are processed simultaneously) at a work center, sequence-dependent setup times, and reentrant flows (a job visits a machine more than once).

While the most important heuristic approach is the shifting bottleneck method, there are several other important research directions. One approach is to approximate the movement of the jobs as a fluid flow [31]. Another approach is to use a randomized algorithm [32]. To improve expected performance, a randomized algorithm makes decisions stochastically.

There is also a large literature on meta-heuristic procedures that solve the job shop scheduling problem. Some examples are

tabu search [33], simulated annealing [34], and genetic algorithms [35]. Aarts *et al.* [36] compare these methods with a threshold acceptance procedure. This is a multistart iterative improvement procedure where a random starting solution is selected. Then, the procedure tries to reduce completion times by repeatedly moving to neighboring solutions that are superior.

For many applied settings, jobs have due dates. Because the decision is to determine the order of the jobs on the machines, a literature on dispatching rules has emerged. These rules provide heuristics that specify the order in which the jobs should be processed on a machine. Most rules are based on processing times, completion time, due dates, or some combination of these measures. One standard rule is shortest remaining processing time (i.e., SPT). When several jobs are available for processing, the rule selects the job with the shortest remaining processing time to be scheduled next. This represents an attempt to minimize the WIP costs. This rule works well when the objective is to minimize the total flow time (completion time minus the start time). Variations on SPT include shortest processing time for the current machine and ratios of total processing time remaining to the number of tasks remaining. When the goal is to meet due dates (desired completion times), other dispatching rules such as minimum slack, work better [37]. Most of these rules either consider variations of the slack time or flow time. Typically, these rules are tested via simulation experiments. Researchers that have studied dispatching rules include [9,38–42].

FUTURE DIRECTIONS

One future direction of job shop scheduling is to improve the current solution techniques both for optimal and heuristic methods. With the wide variety of solution techniques available, there is little knowledge about which method to select for a specific instance. A better understanding of the strengths and weaknesses of each method would improve our ability to find a good or optimal solution. Research will also continue to expand the

generality of the models. We need to be able to solve larger and more realistic problems. Another direction is to more carefully consider robust solutions, solutions that are insensitive to changes in the shop floor. Changes could include new jobs entering the system, the canceling of scheduled jobs, machine breakdowns, and processing time deterioration. A final and perhaps most important direction is to embed the job shop problem into the supply chain. This can be done by integrating the job shop problem with other supply chain activities such as the purchase and storage of raw materials, or the distribution and pricing of final products. An emerging area of research examines supply chain scheduling problems where there are multiple agents. These multiagent problems can be studied where the production facility is a job shop. Examples include models where several customers want to use limited job shop capacity or models where suppliers provide parts to the shop.

REFERENCES

- Ovacik I, Uzsoy R. Decomposition methods for complex factory scheduling problems. Dordrecht: Kluwer Academic; 1997.
- Garey MR, Johnson DS, Sethi R. The complexity of flow shop and job-shop scheduling. *Math Oper Res* 1976;1:117–129.
- Muth JF, Thompson GL, editors. *Industrial scheduling*. Englewood Cliffs (NJ): Prentice Hall; 1963.
- Carlier J, Pinson E. An algorithm for solving the job-shop problem. *Manage Sci* 1989;35:164–176.
- Greenberg H. A branch and bound solution to the general scheduling problem. *Oper Res* 1968;16:353–361.
- Fisher ML. Optimal solution of scheduling problems using lagrange multipliers, Part I. *Oper Res* 1973;21:1114–1127.
- Adams J, Balas E, Zawack D. The shifting bottleneck procedure for job shop scheduling. *Manag Sci* 1988;34:391–401.
- Bertrand JWM. The effect of workload dependent due-dates on job shop performance. *Manag Sci* 1983;29:799–816.
- Chang YL, Sueyoshi T, Sullivan RS. Ranking dispatching rules by data envelopment analysis in a job shop environment. *IIE Trans* 1996;28:631–642.
- Vepsalainen APJ, Morton TE. Priority rules for job shops with weighted tardiness costs. *Manag Sci* 1987;33:1035–1047.
- Balas E, Lancia G, Serafini P, *et al.* Job shop scheduling with deadlines. *J Comb Opt* 1998;1:329–353.
- Sevastianov S, Woeginger GJ. Makespan minimization in preemptive two machine job shops. *Computing* 1994;60:73–79.
- Leighton FT, Maggs B, Rao S. Packet routing and jobshop scheduling in $O(\text{Congestion} + \text{Dilation})$ steps. *Combinatorica* 1994;14:167–186.
- Pinedo M. A note on the two machine job shop with exponential processing times. *Nav Res Logist Q* 1981;28:693–696.
- Jansen K, Mastrolilli M, Solis-Oba R. Approximation schemes for job shop scheduling problems with controllable processing times. *Eur J Oper Res* 2005;167:297–319.
- Brucker P, Thiele O. A branch and bound method for the general-shop scheduling problem with sequence dependent setup times. *OR Spectr* 1996;18:145–161.
- Mauguiere P, Billaut JC, Bouquard JL. New single machine and job-shop scheduling problems with availability constraints. *J Sched* 2005;8:211–231.
- Brucker P, Heitmann S, Hurink J. Job-shop scheduling with limited capacity buffers. *OR Spectr* 2006;28:151–176.
- Werner F, Winkler A. Insertion techniques for the heuristic solution of the job shop problem. *Disc Appl Math* 1995;58:191–211.
- Artigues C, Roubellat F. A polynomial activity insertion algorithm in a multi-resource schedule with cumulative constraints and multiple modes. *Eur J Oper Res* 2000;127:297–316.
- Dauzère-Pérés S, Lasserre JB. Lot streaming in job-shop scheduling. *Oper Res* 1997;45:584–595.
- Brucker P, Schlie R. Job-shop scheduling with multi-purpose machines. *Computing* 1990;45:369–375.
- Mason SJ, Fowler JW, Carlyle WM. A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops. *J Sched* 2002;5:247–262.
- Manne A. On the job shop scheduling problem. *Oper Res* 1960;8:219–223.
- Roy B, Sussmann B. Les problèmes d'ordonnancement avec contraintes disjonctives. Note DS No.9 bis SEMA, Montrouge; 1964.

26. Jackson JR. An extension of Johnson's results on job lot scheduling. *Nav Res Logist Q* 1956;3:201–203.
27. Johnson SM. Optimal two- and three-stage production schedules with setup times included. *Nav Res Logist Q* 1954;1:61–68.
28. Gonzalez T, Sahni S. Flowshop and jobshop schedules. *Oper Res* 1978;26:36–52.
29. Balas E, Simonetti N, Vazacopoulos A. Job shop scheduling with setup times, deadlines and precedence constraints. *J Sched* 2008;11:253–262.
30. Ivens P, Lambrecht M. Extending the shifting bottleneck procedure to real-life applications. *Eur J Oper Res* 1996;90:252–268.
31. Bertsimas D, Gamarnik D. Asymptotically optimal algorithms for job shop scheduling and packet routing. *J Algor* 1999;33:296–318.
32. Shmoys DB, Stein C, Wein J. Improved approximation algorithms for job shop scheduling problems. *SIAM J Comput* 1994;23:617–632.
33. Dell Amico M, Trubian M. Applying tabu search to the job-shop scheduling problem. *Ann Oper Res* 1993;41:231–252.
34. Van Larrhoven PJM, Aarts EHL, Lenstra JK. Job shop scheduling by simulated annealing. *Oper Res* 1992;40:113–125.
35. Bierwirth C. A generalized permutation approach to job shop scheduling with genetic algorithms. *OR Spectr* 1995;17:87–92.
36. Aarts EHL, van Laarhoven PJM, Lenstra JK, *et al.* A computational study of local search algorithms for job shop scheduling. *ORSA J Comput* 1994;6:118–126.
37. Gere WS. Heuristics in job shop scheduling. *Manage Sci* 1966;13:167–190.
38. Jones CH. An economic evaluation of job shop dispatching rules. *Manag Sci* 1973;20:293–307.
39. Panwalkar SS, Iskander W. A survey of scheduling rules. *Oper Res* 1977;25:45–61.
40. Blackstone JH, Philips D, Hogg GL. A state-of-the-art survey of dispatching rules for manufacturing job shop operations. *Int J Prod Res* 1982;20:27–45.
41. Baker KR. Sequencing rules and due-date assignments in a job shop. *Manag Sci* 1984;30:1093–1104.
42. Barman S. The impact of priority rule combinations on lateness and tardiness. *IIE Trans* 1998;30:495–504.

FURTHER READINGS

- Baker KR. Introduction to sequencing and scheduling. New York: Wiley; 1974.
- Brucker P. Scheduling algorithms. Berlin: Springer; 1995.
- Jain AS, Meeran S. Deterministic job-shop scheduling: past, present and future. *Eur J Oper Res* 1999;113:390–434.
- Pinedo M. Scheduling: theory, algorithms, and systems. Englewood Cliffs (NJ): Prentice Hall; 1995.
- Potts CN, Strusevich VA. Fifty years of scheduling: a survey of milestones. *J Oper Res Soc* 2009;60:S41–S68.

JUST-IN-TIME/LEAN PRODUCTION SYSTEMS

W.C. BENTON, JR.
Department of Management Sciences,
Fisher College of Business, The Ohio
State University, Columbus, Ohio

INTRODUCTION

In its simplest form, “the manufacturing process” is a process of material flow. The just-in-time (JIT) philosophy is based on the elimination of capacity and inventory waste and removing non-value-added activities in the manufacturing operation. Specifically, JIT is a manufacturing system in which materials that are pulled through the system are delivered at the precise time to each step in the process, just as they are needed. JIT manufacturing is designed to manage the flow of materials, components, tools, and information. *JIT* production is based on planned elimination of all waste and on continuous improvement. In the past 10 years, JIT has been closely associated with *lean production*. Today, the two terms are used interchangeably simply because they represent highly coordinated repetitive high volume manufacturing systems. An organization driven by JIT philosophy can improve profits and return on investment by reducing inventory levels, reducing variability, improving product quality, reducing production and delivery lead times, and reducing setup costs. With JIT, the entire manufacturing system from purchasing to shop-floor management can be measured and controlled. Therefore, the JIT system is a powerful management tool that could easily determine the success or failure of the manufacturing system. Companies following the JIT philosophy often experience increased efficiency and higher service and product quality. The conceptual framework for the JIT philosophy is shown in Fig. 1.

JIT applies primarily to repetitive manufacturing processes in which the same products and components are produced over and over again. The general idea is to establish efficient flow processes (even when the facility uses a job shop or batch process layout) by linking work centers so that there is an even, balanced flow of materials throughout the entire production process, similar to that found in an assembly line. To accomplish this, an attempt is made to reach the goals of driving all inventory buffers toward zero and achieving the ideal lot size of one unit.

For more than three decades, JIT and material requirement plan (MRP) production systems have followed two independent research streams. However, the performance comparison and compatibility of JIT with MRP systems has led to intensive debates among practitioners as well as researchers (see *Materials Requirements Planning, Push and Pull Production Systems*). As the popularity of JIT motivated by the success of Japanese manufacturing firms has grown, numerous global practitioners initiated complete changeovers from the traditional MRP-based methods to JIT methods. For instance, Harley–Davidson and Hewlett–Packard are two of the most popular examples of JIT success testimonials.

There rarely exists a pure JIT production system in practice. Even at Toyota, the inventor of the JIT system, production smoothing is planned by the master production schedule (MPS), and the MRP is followed on the basis of the MPS, using bill of materials (BOM). This type of manufacturing planning has been adopted by most automobile manufacturers. The integration between MRP and JIT in literature reflects the trend toward a more realistic hybrid-manufacturing environment.

The current shift toward the lean manufacturing environment is one of the major motivations for continuous improvement. There is also a strong need to organize the previous JIT studies and provide directions

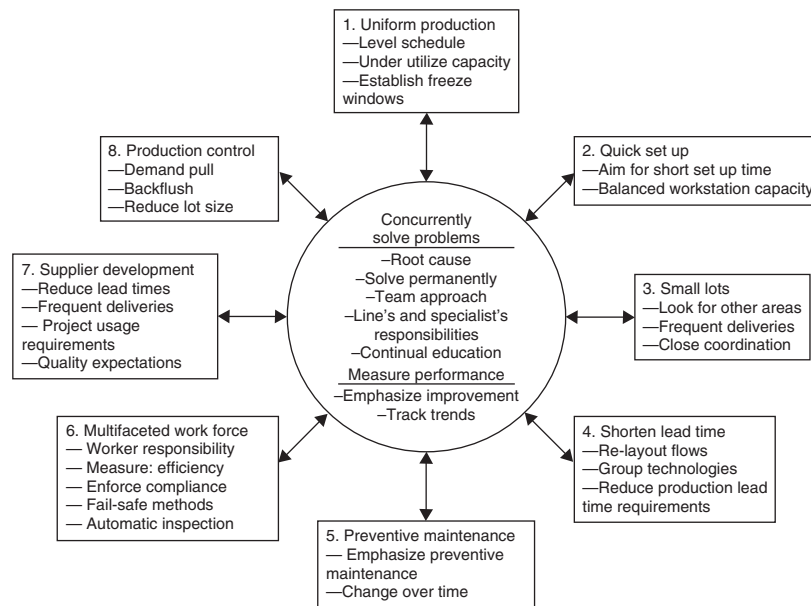


Figure 1. Conceptual framework for JIT.

for innovative JIT implementations and research focus. Since the JIT philosophy is based on planned elimination of all waste and on continuous improvement of productivity, it must encompass essential elements for successful implementation. The following practices are considered as the essential elements for a successful JIT implementation (Fig. 1):

1. *Uniform Production (also known as heijunka).* This creates a uniform load on each work station through constant daily production and producing the same mix of products each day, using a repeating sequence. Meet demand fluctuations through end-item inventory rather than through fluctuations in production level. Use of a stable production schedule also permits the use of *backflushing* to manage inventory: an end-item's bill of materials is periodically exploded to calculate the usage quantities of the various components that were used to make the item, eliminating the need to collect detailed usage information on the shop floor. See the article titled

Materials Requirements Planning in this encyclopedia.

2. *Quick setup Times.* Aim for short setup times—this can be done through better planning, process redesign, and product redesign. A good example of the potential for improved setup times can be found in the airline industry, where Southwest Airline can turn around 30 full capacity flights per day, between Dallas and Houston. Southwest's efficiency is the result of a team effort using the correct aircraft (container) size, and a coordinated, well-rehearsed process.
3. *Small Lot Sizes.* Reducing setup times allows economical production of smaller lots; close cooperation with suppliers is necessary to achieve reductions in order of lot sizes for purchased raw materials and component parts, since this will require more frequent deliveries.
4. *Short Lead Times.* Production lead times can be reduced by moving work stations closer together, applying group technology and cellular manufacturing concepts, reducing queue length, and

improving the coordination and cooperation between downstream processes; delivery lead times can be reduced through close cooperation with supplying organizations, possibly by inducing suppliers to locate closer to the factory.

5. *Preventive Maintenance.* Use machine and worker idle time to maintain equipment and prevent breakdowns.
6. *Multifaceted Work Force.* Workers should be trained to operate several machines, to perform maintenance tasks, and to perform quality inspections. In general, JIT requires teams of competent, empowered employees who have more responsibility for their own work.
7. *Supplier Development.* All defective items must be eliminated, since there are no buffers of excess parts. A *quality at the source (judoka)* program must be implemented to give workers the personal responsibility for the quality of the work they do, and the authority to stop production when something goes wrong. Techniques such as “JIT lights” (to indicate line slowdowns or stoppages) and “tally boards” (to record and analyze causes of production stoppages and slowdowns to facilitate correcting them later) may be used.
8. *Kanban Production Control.* Use a control system such as a *kanban* (card) pull system (or other signaling system) to convey parts between work stations in small quantities (ideally, one unit at a time). In its largest sense, JIT is *not* the same thing as a Kanban system, and a kanban system is not required to implement JIT (some companies have instituted a JIT program along with an MRP system), although JIT is required to implement a Kanban system and the two concepts are frequently equated with one another. The Kanban system is discussed later. See the article titled **Materials Requirements Planning** in this encyclopedia for a kanban, for a comparison of MRP and JIT.

This entry proceeds in the following order. First, the JIT production system is discussed. Then, the kanban production control system is presented. Third, the JIT purchasing framework is evaluated. Next, the role of culture is discussed. Finally, a critical analysis of the JIT philosophy is conducted.

JUST-IN-TIME PRODUCTION SYSTEM

JIT is focused on waste from overproduction, waste of motion, waste of time, waste of movement, waste due to excessive inventory and the waste resulting from poor quality. JIT is Toyota’s manufacturing philosophy to minimize waste, and the JIT production system is controlled by kanban. An example of the kanban production system is presented later in this article. The kanban-controlled JIT production system has been erected on the basis of the premise of minimizing work-in-process inventories (waste) by reducing or eliminating discrete batches. The reduced lot sizes not only contribute to production efficiency and product quality but also reduce the overall costs associated with production in the JIT manufacturing environment. This proposition holds true only when certain conditions are sustained. According to Monden [1], the success of Toyota’s Kanban-controlled production system is supported by smoothing of production, standardization of jobs, reduction of setup times, improvement of activities, design of machine layout, and automation of processes (automation with human touch). In fact, the JIT production system appears most suitable for repetitive manufacturing environments.

Improvements in the kanban-controlled production systems have followed a pragmatic approach, “continuous improvement”. Therefore, success of the JIT production system must be explained in conjunction with continuous improvement, total quality management, and lean manufacturing [1]. Experience and commitment of the workers on the shop floor to continue to improve performance and methods are the major drivers of the JIT production system.

The JIT production system is not a panacea. As with the MRP system, there are

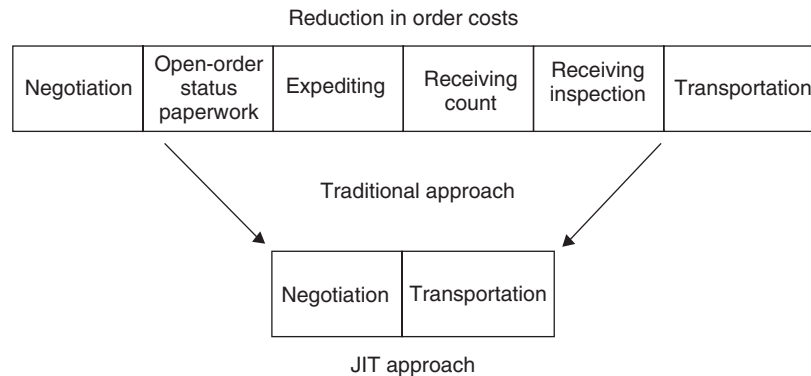


Figure 2. Reduction in order costs.

also operational problems with the JIT system. In fact, there is a list of reasons why the Toyota manufacturing system may not work for all firms. The reasons include cultural differences, geographical dispersion of suppliers, supplier power, different management styles, etc.

The Toyota's manufacturing system has been viewed in different ways by a variety of researchers. The JIT production system is often called *lean production system*, because it uses less of every resource compared with the conventional mass production system [2]. In general, it is well documented that the JIT production system is a "pull" system as opposed to the conventional MRP push system (see **Materials Requirements Planning, Push and Pull Production Systems**). In the MRP/JIT comparison or integration literature, the pull nature of JIT production system has been discussed extensively [3]. Thus, it is only natural to portray the push/pull debates in conjunction with the MRP/JIT studies. One of the most significant changes in manufacturing over the past 15 years has been the purchasing and supply management function. This significant shift has elevated the importance and profile of purchasing function. JIT production requires significant involvement between Tier 1 suppliers and the buying organization. JIT purchasing is discussed in the following section.

JIT PURCHASING

JIT success requires a close relationship with suppliers, who provide multiple deliveries of high-quality materials. Given the expected precision of JIT production systems, there can be no defective raw materials or component parts. Defective parts or late deliveries will cause severe disruptions in JIT production output. JIT requires buying and supplying organizations to work together to achieve consistent delivery of high-quality materials and component parts.

The function of purchasing is to provide a firm with component parts and raw materials. Purchasing must also ensure that high-quality products are provided on time, at a reasonable price. A comparison of critical elements associated with JIT purchasing and traditional purchasing approaches is as follows:

1. **Reduced Order Quantities.** One of the most crucial elements of the JIT system is small lot sizes. Traditionally, long and infrequent production runs have, in the past, been considered beneficial for the overall productivity of a manufacturing organization. However, long production runs sometimes lead to high levels of raw-material and finished-goods inventories. Large setup times have been the primary motivating factor for longer production runs. The JIT concept reduces setup times and the

associated costs by introducing clever changeover techniques and simpler product designs. This permits more frequent production runs and smaller lot sizes. In turn, the JIT purchasing function becomes responsible for more frequent but smaller orders compared to the traditional case. Under the traditional manufacturing system, suppliers, on average, ship enough materials to cover two months of production; since adopting JIT, the lot size has been trimmed to less than three weeks.

If frequent purchase orders are to be a viable option, traditional inventory theory suggests that order costs be minimal for JIT purchasing to be cost effective. Indeed, the JIT philosophy strives to drive the ordering costs down to a bare minimum. A breakdown of the ordering costs associated with conventional purchasing practices is shown in Fig. 2. Under the JIT purchasing approach, the relationship between the supplier and buyer allows the ordering costs to be reduced to simplify the negotiation and transportation costs. All the intermediate costs are expected to be eliminated once the supplier meets the requirements of a Tier 1 supplier. A substantial decrease in the ordering costs permits a greater number of orders to be placed over shorter intervals.

2. *Frequent and "On-Time" Delivery Schedules.* Supplier performance can be measured more accurately under the JIT purchasing approach compared to

the traditional one. In order to obtain small lot sizes for production, the order quantity size needs to be reduced and corresponding delivery schedules need to be made more frequent. The "on-time" windows have been closing systematically over the years. In the pre-JIT days, "on-time" meant anything arriving up to 12 days ahead of the nominal schedule, and as late as six days. On an average, today's JIT buyers set the window at five days early and two days late. Yet, the surprising fact is that on-time deliveries increased from 62 to 79% in spite of the 11-day reduction in delivery window time.

3. *Reduced Lead Times.* To be able to maintain low inventory levels, it is critical that replenishment lead times be as short as possible. The JIT philosophy inherently attempts to reduce lead times for order completions. Under traditional purchasing practices, the lead time is made up of the following components: paperwork lead time, manufacturing time for supplier, transportation lead time, and time spent on receiving and inspection. A comparison between the JIT approach to lead time and the traditional approach is shown in Fig. 3. Under the JIT system, suppliers are usually associated with the company on a long-term basis. Consequently, it is possible for them to reduce paperwork significantly. Also, the supplier may be able to reduce the manufacturing time because of the guaranteed volumes. As a result of building quality

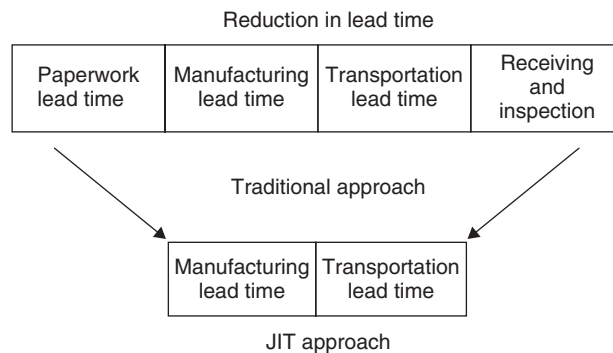


Figure 3. Reduction in lead time.

into the products, the receiving and inspection time spent by the buyer usually decreases. Replenishment lead times also may be reduced by locating a supplier close to the buying firm's plant.

4. *High Quality of Incoming Materials.* Japanese manufacturers attempt to reduce incoming material inspection as much as possible. In order to eliminate the associated receiving inspection costs, a very high emphasis is placed on the quality of incoming materials under the JIT system.

Xerox provides a good example of the unnecessary resources spent on the incoming inspection of goods in its pre-JIT days. All deliveries from suppliers had to pass through Xerox's Webster plant in upstate New York before they could be redirected to the appropriate plants around the United States.

In some instances, component parts from West Coast suppliers were first shipped to the East Coast for inspection and shipped back to the West Coast manufacturing plants. The JIT approach favors material inspection at the supplier plants. Motivation for the suppliers to furnish high-quality materials is cultivated by long-term agreements and strategic relationships between buyers and suppliers. JIT purchasing, therefore, plays a very critical role in improving product quality at the manufacturing source.

5. *Reliable Suppliers.* Since the JIT system does not provide for buffer stocks, unreliable supply, in terms of delivery

time and quality of incoming material, may lead to frequent problems in production. Consequently, reliability of supply is a critical consideration in the selection of JIT suppliers. Since JIT purchasing has gained popularity within the United States, the purchasing function has been preoccupied with trimming the overall supplier base in quest of so-called superior suppliers. Xerox, for example, reduced its supplier base by nearly 4700 over a period of one year [4]. Ford has reduced its supplier base by nearly 40% since 1980. In the US number of superior suppliers is on the rise as JIT promotes the concept of fewer but better suppliers.

A brief summary of the comparison between the traditional and JIT purchasing approaches is shown in Table 1. JIT purchasing systems are much more difficult to manage. If JIT purchasing systems are not properly implemented and managed, the synergies will be lost and the system will collapse. However, significant advantages are realized if these elements are implemented successfully.

JIT BENEFITS

Implementation of JIT assists in its major objectives of reducing waste from overproduction, improving quality of incoming materials and supplier delivery performance, along with reducing lead times and cost of materials. Some of the critical

Table 1. Comparison between Traditional and Just-in-Time (Lean) Purchasing Approaches

	Traditional Purchasing	JIT Purchasing
Order quantities	Based on trade-offs between ordering and carrying costs	Based on small lot sizes for production
Delivery schedules	Infrequent, primarily because of high ordering costs involved	Frequent because of small lot sizes and low ordering costs
Delivery windows	Relatively wide	Very narrow
Delivery lead times	Relatively long and relaxed	Stringent and reduced significantly
Parts quality	Responsibility of the quality function in the buying organization	Responsibility of supplier
Supplier base	Fairly broad	Considerably smaller

JIT advantages for the manufacturer are discussed below.

1. *Reduced Inventory Levels.* JIT purchasing facilitates reduction in inventory levels and the associated inventory holding costs. Asian firms like Toyota have been able to reduce inventory levels to such an extent that their inventory turnover ratios have gone up to over 60 times per year, compared to corresponding ratios of 5 to 8 reported by most American manufacturers. US manufacturers have been trying to reduce inventory levels by using JIT purchasing techniques. National Cash Register (NCR's) Ithaca plant, for instance, has been successful in reducing its number of days of inventory from 110 to 21 within three years of implementing JIT. However, a strong debate exists as to whether or not manufacturers shift the burden of inventories to suppliers using the guise of JIT. This issue will be addressed later. Reduced inventory levels are indeed one of the benefits of the JIT system, which basically stresses continuous improvement and elimination of waste.
2. *Improved Lead-Time Reliability.* Compared to traditional manufacturing approaches, delivery lead times under the JIT system are considerably shorter. Lead-time reliability is usually much better for JIT systems. This implies higher levels of customer service and lower safety stock requirements for the company. Lower levels of safety stock contribute significantly to reduced working capital requirements for the firm.
3. *Scheduling Flexibility.* JIT emphasizes scheduling flexibility by aiming for reduced purchasing lead times and setup times. Such flexibility prevents confusion in the manufacturing plant and offers unique competitive advantages to manufacturing firms since they are capable of adapting to changes in the environment more quickly.
4. *Improved quality and customer satisfaction.* JIT manufacturing results in improved quality and corresponding levels of higher customer satisfaction. Since high-quality products are critical in achieving a competitive advantage in today's global business world, manufacturers gain immensely by implementing the JIT production control system. High-quality incoming materials result in savings associated with reduced rework and scrap.
5. *Reduced Costs of Parts.* As cooperation and relationships between suppliers and manufacturers build up in a JIT system, so do the opportunities to conduct an extensive value analysis and focus on reducing the cost of parts purchased. A comprehensive JIT progress report indicates that supplier costs were reduced by 11% when they adopted the JIT system in cooperation with their customers. Long-term commitments on the part of the manufacturer allow volume purchases, development of supplier learning curves, and overall productivity increases [4].
6. *Constructive Synergies with Suppliers.* A JIT purchasing program involves close technical cooperation with suppliers. This particularly means the cooperation between manufacturing and design engineers. Because of smaller lot sizes and frequent delivery schedules, suppliers are in a position to receive quick feedback regarding any potential manufacturing or design problems. Also, manufacturing is in a position to implement engineering changes quicker because of the reduced inventory levels. The JIT progress report mentioned above indicates that supplier quality improved by 26% since the JIT system was adopted.

It is well documented that JIT reduces the physical inventory level. Reductions in physical inventory will also have a favorable impact on the following:

- Reduced insurance premiums associated with the storage of inventory.

- Reduced inventory holding costs.
- Reduced labor cost in store rooms and material handling costs.
- Reduced clerical and administrative costs.
- Reduced waste from the manufacturing process.
- Reduced obsolescence costs.
- Reduced depreciation of handling and storage equipment.

Each of the cost savings will result in a leaner, more profitable operation.

THE ROLE OF CULTURE

A crucial issue to be considered is the relevance of culture in the successful implementation of the JIT system in a country. Honda's culture and its focus on group-oriented activities are particularly suitable to the implementation of the JIT production control system in that environment. The need to have harmony in organizations provides for better manufacturer-supplier relationships at Toyota and Honda. Moreover, long-term relationships between supplier and manufacturer are the norm of doing business in Japan. Severance of a business relationship between manufacturer and supplier has a strong stigma associated with it, which both manufacturers and suppliers try to avoid as much as possible.

Within the United States, however, such relationships between manufacturers and suppliers are a little more difficult to cultivate. Traditionally, the business firms in the United States are so short-term oriented that they have their immediate interests in mind. Moreover, the level of employee and supplier commitment to the JIT concept is not as uniform and high as it is in Japan. This does not mean that JIT is not a viable concept in the United States. It is not advisable, however, for US firms to blindly emulate the Japanese JIT approach. In fact, US firms should try and tailor JIT to their needs and circumstances. Some firms in the United States have developed and implemented their own version of JIT under different

names, such as lean, ZIPS (zero inventory production systems), MAN (material as needed), and nick of time. A case study on JIT indicates how Hutchinson Technology (a publicly held company that manufactures a variety of products for computer peripheral and military markets) organized and implemented JIT manufacturing [5]. The major difficulties encountered in the process included the inability of purchasing personnel to make the immense cultural transformation and the lack of resources to effect the change properly. Despite these difficulties, their JIT purchasing program has been successful. Greater geographic separation between supplier and manufacturer in the United States is a major impediment to the implementation process according to the one survey [6]. The survey also indicates that JIT does not imply single sourcing in the United States as it does in Japan. This reflects the reality of the marketplace in the United States and efforts of manufacturers to adjust to it as best as they can.

CRITICAL ANALYSIS OF THE JIT CONCEPT

Most of the testimonials published on JIT systems exalt the simplicity inherent in the system processes and procedures. However, the key issue for a firm in the United States is whether it is simple to implement JIT in an existing manufacturing environment. Does JIT really provide the solution to most manufacturing problems in the United States? It is also worth investigating whether traditional purchasing approaches have been outdated in light of JIT production.

JIT came under intense scrutiny when Japanese manufacturers stormed into the US markets and took away a substantial share from the US automobile industry. Subsequent investigations revealed, much to the relief of US manufacturers, that it was not only the work culture in Japan that provided the Japanese an edge but also the JIT approach to manufacturing management. Many US manufacturers have been in strong pursuit of this manufacturing revolution and manufacturing excellence that JIT was portrayed to bring. Several manufacturing firms

that adopted the JIT approach early include General Motors, Hewlett Packard, Ford, and Xerox. Unfortunately, the excitement about the radically new manufacturing approach, coupled with the romantic version of JIT put forward by many, lulled quite a few manufacturers into believing that JIT would bring instantaneous results for their companies. Too many companies turned to JIT looking for a relatively painless financial surgery that would yield substantial short-term benefits. Over the years, these companies have come to realize the tremendous effort and commitment required to make a JIT system run smoothly.

The radical proponents of JIT manufacturing in the United States during the past three decades, the so-called JIT revolutionaries, are to some extent responsible for this initial misunderstanding. The practitioners painted an extremely romantic picture of JIT emphasizing simplicity and efficiency, along with a state of affairs where employee morale would be high and relations between buyers and suppliers would be completely harmonious. They also called for immediate action and changeover to JIT without really considering the possible ramifications of implementation in the United States. Nor did they convey the message that driving obstacles and impediments out of the system would take serious and substantial effort, commitment, and time.

The pragmatic version of JIT put forth by Japanese authors focuses on the details of the production process. Here the emphasis is on identifying impediments to the smooth flow of materials and innovative techniques to overcome those problems. The Japanese perspective clearly stresses the need for careful and slow implementation, following thorough preparation. It takes time to change attitudes of the workforce and nurture long-term relationships with suppliers.

The transition to JIT has not necessarily been a smooth one for many companies in the United States. But this does not imply that switching from a pure MRP system to a JIT or hybrid system was a mistake for most companies. There are two serious drawbacks with the MRP production control system. First, the master production schedule that

drives MRP is based on estimated customer requirements; and second, MRP's production control system utilizes a "push" system for manufacturing goods. That is, the purchasing function places orders for materials in large lot sizes even though the material may not actually be required, and one work station pushes materials to the next regardless of actual production requirements. Changes in demand estimates or forecasts may allow inventory to pile up in the plant. Very frequent adjustments to the master production schedule make the production system extremely nervous and place enormous pressures on purchasing. However, it must be admitted that MRP is an elegant technique for the requirements of exploding materials for production. This system has increasingly become easier to implement with the advent of sophisticated computing technology. It is not surprising to find some Western manufacturers who still utilize MRP for ordering purchased materials but require that delivery schedules be based on the kanban system and lean concepts. For a more comprehensive discussion, see the MRP entry.

Another critical issue for JIT manufacturers is the variability in product demand. The JIT system seems to work best when its smooth production and low inventory requirements are aimed at meeting a relatively stable product demand. However, demand patterns are not stable for all products. In order to induce a relatively stable demand, companies using JIT manufacturing often consolidate their product lines. They emphasize high quality and low cost of the product, but not variety and availability. This suggests that not all marketing strategies are compatible with the JIT system.

Does this mean that JIT, as a concept, is not particularly suited to manufacturers in the United States? The basic concept is as applicable in the United States as it is in Japan. The JIT system does yield substantial benefits where it has been implemented properly. Undoubtedly, proper implementation of JIT is the key to its success. We have seen the advantages JIT offers to a firm from a purchasing point of view. There are significant benefits to other functional areas as well. However, it should be realized that JIT

is not a panacea for all manufacturing problems and scenarios. A comprehensive study by Krajewski *et al.* revealed that the selection of a production or inventory system can be of less importance than the improvement in the manufacturing environment itself [7]. Keeping this in mind, manufacturers in the United States should evaluate the potential benefits of a JIT system from their own perspective, not from that of the romantic visionaries.

CONCLUSION

JIT has changed the way manufacturing is conducted throughout the world. JIT applies primarily to *repetitive manufacturing* processes in which the same products and components are produced over and over again. JIT is driven by a philosophy of eliminating waste and continuous improvement. Waste results from any activity that adds cost without adding value, such as the unnecessary moving of materials, the accumulation of excess inventory, or the use of faulty production methods that create products requiring subsequent rework. JIT improves profits and return on investment by reducing inventory levels, reducing variability, improving product quality, reducing production and delivery lead times, and reducing other costs. In a JIT system, underutilized (excess) capacity is

used instead of buffer inventories to hedge against problems that may arise. Perhaps one of the most important realities of JIT is the fact that supplying organizations have become an important consideration for the manufacturing function, wherein they should be viewed as partners and not adversaries.

REFERENCES

1. Monden Y. Toyota production system. 3rd ed. Institute of industrial engineers. Norcross (GA): Engineering & Management Press; 1998.
2. Womack JP, Jones DT, Roos D. The machine that changed world: The story of lean production. Harper Perennial Boston: Harpercollins publishers; 1990.
3. Benton WC, Shin H. Manufacturing planning and control: The evolution of MRP and JIT integration. *Eur J Oper Res* 1998;110:411–440.
4. Inman AR. Time-based competition: Challenges for industrial purchasing. *Ind Manag* 1992;34(2):31–32.
5. Ray S. Just-in-time purchasing—A case study. *Hosp Manag Q* 1990;12(1):7–12.
6. Freeland JR. A Survey of just-in-time purchasing practices in the United States. *Prod Inventory Manag* 1991;32(2):43–50.
7. Krajewski L, King BE, Ritzman RP, Wong DS. Kanban, MRP, and shaping the manufacturing environment. *Manag Sci* 1987;33:39–57.

KLIMOV'S MODEL

JOSÉ NIÑO-MORA
Department of Statistics,
Carlos III University of
Madrid, Madrid, Spain

FEEDBACK QUEUEING MODELS FOR TIME-SHARED SYSTEMS

The 1960s witnessed a surge of activity in the development and analysis of feedback queueing models of time-shared computer and communication-network systems, following the introduction of the first such model in Ref. 1. For a pioneering analysis of a queue with feedback see Ref. 2. A time-shared system is meant to provide service to multiple users, by allocating processing resources (e.g., a computing processor or a multiaccess communication channel) in such a way that users experience the illusion that they have continuous and simultaneous access to them. In the single-processor case, this is accomplished by allocating the processor at each time to one user but only for a short period of time, termed a *quantum* of service. If the service requirements of a user's submitted job are satisfied within his allotted quantum of service, the user leaves the system. Otherwise, the job is fed back into the system's queue with its service requirements accordingly updated, to await its next turn of server's attention, according to the *scheduling policy* adopted to decide which job to serve next at each time.

A number of such policies were devised at the time [3–6], which give preferential treatment to short (e.g., interactive) jobs at the expense of long (e.g., noninteractive) jobs. Perhaps the simplest is the *round-robin* policy, whereby an incoming job joins the tail of the queue, where queue position determines the order of service (head-of-the-line first) and, after its allotted service quantum, an incomplete job is fed back to the end

of the queue. Another much-studied policy is the *least attained service* (LAS)—also called *foreground-background* (FB)—rule, which incorporates a priority structure by feeding back incomplete jobs to lower priority queues. More analytically tractable idealized versions of such policies, where the processor's capacity is simultaneously shared by multiple jobs, are obtained by letting the quantum size tend to zero, which gives rise to so-called *processor-sharing* policies. See Ref. 7, Chapter 4 for a review and discussion of these and other models and policies. A multiclass queue under a static priority policy with probabilistic feedback and round-robin within each class is analyzed in Ref. 8.

KLIMOV'S MODEL, RESULTS, EXTENSIONS, AND APPLICATIONS

Such early works were mostly concerned with the *performance analysis* of given scheduling policies, bypassing the more ambitious *performance optimization* problem, which is to design a scheduling policy that optimizes a given performance objective. To quote from Ref. 5:

“No mention has been made above about how a designer of a time-shared system should choose among various possible priority rules. This presupposes some sort of cost criterion. Even when this is agreed upon, general conclusions are difficult. As a consequence, most of the literature either avoids this question entirely or presents voluminous graphs and tables.”

In his seminal paper [9], Klimov addressed the problem of finding a scheduling policy for a multiclass M/G/1 queue with general Bernoulli feedback that minimizes the steady-state, mean-weighted holding cost. In such a model, which encompasses and extends previous queueing models of time-shared systems, a single server is to be dynamically allocated to a finite number n of job classes, labeled by $k \in N \triangleq \{1, \dots, n\}$.

Class k jobs arrive at the system as a Poisson stream with rate λ_k , and have class-dependent i.i.d. generally distributed service times with finite mean $1/\mu_k$ and variance σ_k^2 . Arrival and service-time processes are independent across classes. Upon completion of a class k job, it is fed back to continue further processing as a class l job with probability p_{kl} , for $l \in N$, leaving the system with probability $p_{k0} \triangleq 1 - \sum_{l \in N} p_{kl}$.

Such Markovian feedback probabilities are assumed to be such that, in the absence of further arrivals, a job starting out at any class eventually leaves the system with probability 1. This will be the case if the matrix $\mathbf{I} - \mathbf{P}$ is invertible, where $\mathbf{I}$ is the identity matrix and $\mathbf{P} = (p_{kl})_{k,l \in N}$ is the feedback probability matrix. The total arrival rate $\hat{\lambda}_k$ for each class k is given by the unique solution to the traffic equations

$$\hat{\lambda}_l = \lambda_l + \sum_{k \in N} p_{kl} \hat{\lambda}_k, \quad l \in N.$$

Model parameters are assumed to satisfy the stability condition stating that the traffic intensity, or utilization factor, $\rho \triangleq \sum_{k \in N} \rho_k$ be less than unity, where $\rho_k \triangleq \lambda_k / \mu_k$ is the traffic intensity for class k .

During its stay in the system, a class k job incurs linear holding costs at rate c_k per unit time. Server allocation decisions are made according to an *admissible* scheduling policy, which is required to be (i) *work conserving*, that is, it neither allows the server to sit idle while there are jobs waiting, nor does it affect the total service requirement or the exogenous arrival time of any job; (ii) *nonanticipative* or history-dependent, that is, decisions at any given time can only be based on current or past system information, but not on future information such as future arrival times or remaining service requirements of jobs that have not yet departed; (iii) *ergodic*, that is, they induce on system processes, such as the customers' waiting times or the number in system for each class, steady-state distributions with finite means matching the corresponding sample averages; and (iv) *non-preemptive*, that is, the server allocates its full service capacity to jobs one at a time proceeding uninterruptedly to completion of the job at its current class.

The problem addressed by Klimov [9] was to design an admissible scheduling policy that minimizes the steady-state mean holding cost rate. He established the optimality of a static (priority)-index rule, where a numeric index v_k^* is attached to each class k , and jobs are scheduled by awarding higher nonpreemptive priority [10] to classes with larger index values, breaking ties arbitrarily. Jobs within the same class are served on a first-come first-served (FCFS) basis. Such a result extends to a multiclass M/G/1 queue with feedback the earlier result on optimality of a static index rule (the *cμ-rule*) for the special case without feedback, which had been established in Refs 11, (pp. 84–85) and 12. Klimov further furnished in Ref. 9 a one-pass *adaptive-greedy algorithm* that computes the n index values in n steps, requiring the solution of appropriate linear equation systems.

The approach deployed in Ref. 9 to obtain such results was also ground-breaking, as it did not rely on conventional dynamic programming arguments, which do not lend themselves easily to analyze multidimensional semi-Markov system models with unbounded costs per stage, such as Klimov's model. Instead, Klimov [9] introduced the use of *linear programming* (LP) formulations for optimal queueing scheduling problems, which exploit their special structure, thus pioneering what has later been termed the *achievable performance region approach* (see Ref. 13 and the accompanying discussion). By formulating certain flow balance relations as linear equality constraints on system performance measures, Klimov was able to reformulate the optimal dynamic scheduling problem of concern as a tractable LP problem having $O(n^2)$ linear equality constraints on $O(n^2)$ variables. Then, by exploiting the LP problem's special structure and drawing on LP duality theory, he solved analytically both the LP and the scheduling problem in terms of the index values v_k^* referred to above, which emerge in the solution to the dual LP problem. In Ref. 14, Klimov obtained simpler expressions for the index and further results for a special case of the model in Ref. 9. As for interpretation of the Klimov index v_k^* of class k , it is shown in Ref. 14 to measure the maximum reduction

in expected holding cost per unit expected service time that can be achieved on a single job starting out at class k , in a series of processing stages that ends according to a stopping rule. Hence, a remarkable property of the Klimov index is that it does not depend on the exogenous arrival rates λ_k .

A version of Klimov's problem under the discounted cost criterion, which incorporates job completion rewards, allows the server to remain idle while there are jobs waiting, and does not require the stability condition $\rho < 1$ to hold, is addressed in Ref. 15. The optimality of a modified static index rule is established. Besides considering the non-preemptive case, the analysis is extended to the preemptive case, provided service times are exponentially distributed. This paper further discusses application of the model to flexible manufacturing systems, where the production of multiple part types which may require rework must be scheduled on a single machine. The approach in Ref. 15 draws on earlier work [16] on optimal scheduling of a multiclass M/G/1 queue without feedback under the discounted cost criterion.

A stronger sample-path optimality result for a version of the Klimov model with two classes and exponential service times, which allows general arrival processes, is given in Ref. 17. A further extension of the Klimov model under non-Markovian arrival and service processes given by point processes is presented in Ref. 18.

An extension of the Klimov model under general arrival and service processes, along with a proof of asymptotic optimality of the Klimov index rule in heavy traffic, is given in Ref. 19.

The Klimov index policy has been proposed in Ref. 20 as a heuristic scheduling rule for the more complex problem of scheduling a multiclass, multistation Markovian queueing network. It must be emphasized that Rosberg [20] further introduced relaxed LP formulations for scheduling problems on such networks, which extend Klimov's original LP formulation [9] and furnish tractable bounds on optimal performance. That is the true origin of such LP formulations, which instead appear to have been attributed to latter work [21–23].

An approximate analysis of the Klimov index rule for multiclass multistation queueing networks proposed in Ref. 20 is given in Ref. 24, exploiting the structure of the LP formulation referred to above via approximate work conservation laws.

A recent application of the Klimov model and index rule to the optimal dynamic scheduling of wireless error-prone transmissions is given in Ref. 25.

RELATIONS WITH STOCHASTIC SCHEDULING AND BANDIT MODELS

Besides its original motivation in the context of time-sharing queueing systems, insightful connections can be drawn between Klimov's model and models in the areas of stochastic scheduling and bandit problems. Stochastic scheduling [26] problems are concerned with finding the optimal priority order in which a given set of jobs with random processing requirements is to be processed, where typically all jobs under consideration are present at the start. The simplest such model involves sequencing a set of jobs on a single processor in a nonpreemptive fashion, in order to minimize the expected total linear holding cost incurred. Such a problem was first solved in the special deterministic case in Ref. 27, and in the stochastic processing-time case in Ref. 28. In both cases, the optimal sequencing rule is of index type, being consistent with the $c\mu$ -rule referred to above. More complex stochastic scheduling problems that incorporate deterministic (modeling precedence constraints among jobs) or random feedback of jobs into the system, and which are also solved by index rules, are addressed in Refs 29–34, where Bruno [34] deals with the version of Klimov's model without external arrivals. An extension of the latter model and also of Klimov's model which incorporates random bulk arrivals is the undiscounted branching bandit model introduced in Ref. 35.

A closely related although distinct research area which has evolved with substantial autonomy is that of bandit problems [36]; see also the article titled *Multiarmed Bandits and Gittins Index*. The classic *multiarmed bandit problem* concerns the optimal

sequential selection of a project to engage at each time over an infinite horizon among a collection of stochastic projects, where each project is modeled as a bandit process, that is, a binary-action (active or passive) discrete-time discounted Markov decision process which can only change state and earn rewards when active. After being long considered both computationally and analytically intractable, the problem was first solved in a landmark result by Gittins and Jones [37], using a dynamic priority-index rule based on a dynamic allocation index attached to each project as a function of its state, whose roots can be traced back to Refs 38 and 39. See also Ref. 40 for a presentation and discussion of such a celebrated result.

An extension of the Gittins and Jones result is presented in Ref. 41 for the so-called *arm-acquiring bandit* model, which incorporates random bulk project arrivals, and can hence be seen as a discounted discrete-time version of Klimov's model. Such a model is further extended in Ref. 42 (Chapter 14, Section 11) to a continuous-time semi-Markov setting under Poisson arrivals, and in Ref. 43 under generally distributed branching bandit arrival processes.

The relations between such discounted models and their undiscounted (time average) counterparts are explored in Ref. 44. It is shown that, under stability conditions, the optimal index rule for the arm-acquiring bandit model is asymptotically equivalent to that for its "closed" counterpart (i.e., the corresponding model without arrivals) in the undiscounted limit.

A striking difference between the performance objectives in stochastic scheduling and in bandit problems is that, typically, the former seek to minimize a measure of holding costs incurred, whereas the latter strive to maximize a measure of rewards earned on active projects. In Ref. 45, the *multiarmed tax bandit problem* is introduced, which incorporates a discounted waiting cost objective, and the two types of problems (maximize the rewards earned on active projects and minimize the costs incurred by passive projects) are shown to be equivalent for discounted bandit problems via a simple transformation, which allows use of a

Klimov-like algorithm to compute the Gittins index for discounted arm-acquiring bandits, including the discounted version of Klimov's model [15] as a special case. Such a relation between Klimov's and bandit models is further elucidated in Ref. 46 and more generally in Ref. 44, where it is shown that the Klimov index in Ref. 9 is the undiscounted limit of the Gittins index in the corresponding arm-acquiring bandit model. Also, the optimality of the Klimov index rule is established under preemptive policies assuming exponential service time distributions. The relation between classic and tax bandit problems and the Klimov model is further clarified in Ref. 47 via conservation laws, and more recently in Ref. 48, focusing on performance evaluation under the optimal index rule in the undiscounted case.

A modification of Klimov's index algorithm was introduced in Refs 49 and 50 to compute the index for restless bandits introduced in Ref. 51. See also the review paper [52] and the accompanying discussion.

GENERALIZED CONSERVATION LAWS, PERFORMANCE REGION, ADAPTIVE POLICIES, AND PARALLEL SERVERS

The so-called *achievable performance* region approach to stochastic optimization (see Ref. 13 and the accompanying discussion) seeks to address a problem such as Klimov's by reformulating it as an LP problem in the space of performance measures of concern (e.g., mean waiting times for each class), and then obtain an optimal policy from the latter's solution. As pointed out above, such an approach was in fact pioneered by Klimov himself in his original analysis in Ref. 9, where an LP formulation was presented based on flow balance arguments, and the optimal index rule was obtained from the latter's LP solution via duality theory.

A powerful alternative method for obtaining LP formulations for the region of achievable performance of a multiclass queueing system under admissible scheduling policies, which are required to be work conserving (i.e., no work is created or destroyed), is based on exploiting the

fundamental principle of work conservation. This states that, under work-conserving policies, the unfinished work process $U(t)$, measuring the total remaining service time due to jobs present in the system at each time t , is sample-path invariant. For certain models, the work conservation principle yields so-called *work conservation laws*, which are linear constraints on performance measures characterizing, totally or partially, the achievable performance region. The reader is referred to the article titled **Conservation Laws and Related Applications** in this encyclopedia for a comprehensive discussion and review of such a topic. Early work on conservation laws focused on multiclass queues without feedback. The seminal conservation law of Kleinrock [53,54] was gradually developed [55–59] into the framework of so-called *strong conservation laws* [60], which yields a complete characterization of the achievable performance region in the form of (the base of) a polymatroid, a much studied polyhedron in polyhedral combinatorics introduced in Ref. 61. Thus, the optimality of the $c\mu$ -rule for the corresponding dynamic scheduling problems was shown to be a consequence of the optimality of the greedy algorithm for solving LP problems over polymatroids.

An exact polyhedral characterization of the performance region for Klimov's model under ergodic policies, obtained via the principle of work conservation, was first presented in Ref. 62. The linear constraints characterizing the region of achievable mean delays $\bar{W}^\pi$ consist of: an equality constraint of the form

$$\sum_{k \in N} \rho_k^N \bar{W}_k^\pi = g(N),$$

along with a family of inequality constraints of the form

$$\sum_{k \in S} \rho_k^S \bar{W}_k^\pi \geq g(S)$$

for nonempty subsets $S \subset N$ of customer classes, where the ρ_k^N and ρ_k^S are positive workload coefficients. The resulting polytope is no longer the base of a polymatroid, but that of a new type of polytope, a so-called *extended polymatroid*, which also possesses

strong structural and algorithmic properties, as elucidated in Ref. 63. Further extensions of conservation laws under nonergodic policies are introduced in Ref. 62, which along further developments are used in Ref. 64 to obtain adaptive policies for optimizing nonlinear performance objectives in the Klimov model. The design of optimal adaptive policies for the Klimov model had been previously investigated in Refs 65 and 66.

The axiomatic framework of strong conservation laws for multiclass queues without feedback in Ref. 60 is extended in Ref. 47 to the framework of generalized conservation laws, which have the form indicated above for the Klimov model. Further, in Ref. 47 such a framework is deployed to obtain the polyhedral performance region of the more general branching bandit model [35,43], both under average and discounted performance measures, including as a special case the classic multiarmed bandit problem [36]. A new proof is thus obtained via conservation laws and polyhedral LP methods for the optimality of index policies in such classic problems. The reader is referred to Ref. 67, (Chapter 11) for an accessible textbook account of such methods.

In Refs 68 and 69 the properties of generalized conservation laws and extended polymatroids are further reviewed and developed, including connections with submodularity and the efficient algorithmic treatment of linear side constraints.

For the extension of Klimov's model that incorporates parallel servers, the Klimov index rule is no longer optimal. Yet, Glazebrook and Niño-Mora [70] introduced an LP relaxation over an extended polymatroid obtained from approximate generalized conservation laws, which yields closed-form bounds implying the asymptotic optimality of the rule in the heavy traffic regime.

CONSTRAINED OPTIMIZATION

In some applications of Klimov's model, one or more traffic classes (e.g., interactive traffic in a communication network) may demand more stringent quality of service requirements, which can be modeled via

upper bounds on corresponding performance measures such as mean delays. In the case of a multiclass, single-server, discrete-time queue without feedback under the average cost criterion, Nain and Ross [71] addressed the problem of finding an optimal scheduling rule that minimizes mean linear holding costs subject to a single linear performance constraint. They used a Lagrangian approach to establish the optimality of a stationary policy obtained by randomizing at each time period between two static priority policies, with both such policies and the corresponding probabilities remaining fixed over time. Such a result is extended in Ref. 72 to the discrete-time version of Klimov's problem under a single linear performance constraint, both under discounted and average cost criteria.

The problem of finding an optimal scheduling policy for the branching bandit problem under the average criterion subject to linear performance constraints, and hence for the Klimov model with such constraints as a special case, is addressed in Ref. 22 via LP methods, exploiting the extension to branching bandits of Klimov's LP formulation in Ref. 9. Unlike the LP formulation in Ref. 47, which has $2^n - 1$ constraints on n variables, that in Refs 9 and 22 has $O(n^2)$ constraints on $O(n^2)$ variables. The extended polymatroid structure of the performance region for the Klimov and branching bandit models is exploited in Refs 68 and 69 to design and compute optimal policies subject to linear performance constraints.

PERFORMANCE ANALYSIS

An exact performance analysis for a single-class multiclass M/G/1 queue with Bernoulli feedback is given in the seminal work [2], in the form of the distributions for the number in system and the delay. For the multiclass Klimov model, although it is not explicitly mentioned in Ref. 9, the mean number in system (and hence the mean response time) for each class under any static priority policy can be readily obtained from the linear equation system already given in Ref. 9, by setting to zero the appropriate variables

consistently with the priority ordering of concern. Such a fact appears to have been overlooked, as Simon [73] presents a mean delay analysis for Klimov's model—and some extensions involving preemptions, branching and bulk arrivals—assuming a deterministic feedback structure. In Ref. 74, such analyses are extended to obtain the waiting-time distribution. For the discounted version of Klimov's model and more generally for branching bandits, the results in Ref. 47 via generalized conservation laws allow obtaining the mean discounted number in system for each class by solving a system of n linear equations, where n is the number of classes. For a review of performance analysis results for corresponding models without feedback.

THE FLUID KLIMOV MODEL

Researchers have also investigated deterministic continuous-state fluid versions of Klimov's model, within the area of fluid models of multiclass queueing networks. The first such study is Ref. 75, where it is shown that the Klimov index rule emerges as a myopic solution to the corresponding fluid model, which is a global solution under a certain condition. Connors *et al.* [76] reports on the application of the fluid scheduling model in Ref. 75 to a real-world wafer fab. In Ref. 77, optimality of the index rule is obtained via Pontryagin's maximum principle, under an additional stability condition. The achievable region approach via generalized conservation laws used in Ref. 47 for the discrete-state model is deployed in Ref. 78 to analyze the fluid Klimov model. For further developments in the analysis of optimal server allocation in fluid queueing networks, including applications to Klimov's model, see Ref. 79. An analysis of the optimal control of the fluid Klimov model with side constraints, via LP methods, is presented in Ref. 80.

REFERENCES

1. Kleinrock L. Analysis of a time-shared processor. *Nav Res Log Q* 1964;2(1):59–73.
2. Takács L. A single server queue with feedback. *Bell Syst Tech J* 1963;42(2):505–519.

3. Coffman EG, Kleinrock L. Feedback queueing models for time-shared systems. *J Assoc Comput Mach* 1968;15(4):549–576.
4. Schrage L. The queue M/G/1 with feedback to lower priority queues. *Manag Sci* 1969; 13(7):466–471.
5. Wolff RR. Time sharing with priorities. *SIAM J Appl Math* 1970;19(3):566–574.
6. Kleinrock L, Muntz RR. Processor sharing queueing models of mixed scheduling disciplines for time shared systems. *J Assoc Comput Mach* 1972;19(3):464–482.
7. Kleinrock L. Queueing systems. Volume II, Computer Applications. New York: John Wiley & Sons, Inc.; 1976.
8. Chang W. Queues with feedback for time-sharing computer system analysis. *Oper Res* 1968;16(3):613–627.
9. Klimov GP. Time-sharing service systems. I. *Theor Probab Appl* 1974;19(3):532–551.
10. Cobham A. Priority assignment in waiting line problems. *Oper Res* 1954;2(1):70–76.
11. Cox DR, Smith WL. Queues. London: Methuen; 1961.
12. Fife DW. Scheduling with random arrivals and linear loss functions. *Manag Sci* 1965; 11(3):429–437.
13. Dacre M, Glazebrook K, Niño-Mora J. The achievable region approach to the optimal control of stochastic systems. *J R Stat Soc Ser B Stat Methodol* 1999; 61(4):747–791.
14. Klimov GP. Time-sharing service systems. II. *Theory Probab Appl* 1978;23(2):314–321.
15. Tcha DW, Pliska SR. Optimal control of single-server queueing networks and multi-class M/G/1 queues with feedback. *Oper Res* 1977;25(2):248–258.
16. Harrison JM. Dynamic scheduling of a multiclass queue: discount optimality. *Oper Res* 1975;23(2):270–282.
17. Chao X. On Klimov's model with two job classes and exponential processing times. *J Appl Probab* 1993;30(3):716–724.
18. Chao X. Optimal intensity allocation of single-server queueing networks. *Int J Syst Sci* 1994;25(11):1987–2000.
19. Xia CH, Shanthikumar JG. Asymptotic optimal control of multiclass G/G/1 queues with feedback. In: Shanthikumar JG, Yao DD, Zijm WHM, editors. Volume 63, Stochastic modeling and optimization of manufacturing systems and supply chains (International Series in Operations Research & Management Science). Norwell (MA): Kluwer Academic; 2003. pp. 127–142.
20. Rosberg Z. Process scheduling in a computer system. *IEEE Trans Comput* 1985;34(7): 633–645.
21. Bertsimas D, Paschalidis I, Tsitsiklis J. Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann Appl Probab* 1994;4(1):43–75.
22. Bertsimas D, Paschalidis IC, Tsitsiklis JN. Branching bandits and Klimov's problem: achievable region and side constraints. *IEEE Trans Automat Control* 1995;40(12):2063–2075.
23. Kumar S, Kumar PR. Performance bounds for queueing networks and scheduling policies. *IEEE Trans Automat Control* 1994;39(8): 1600–1611.
24. Glazebrook KD, Niño-Mora J. A linear programming approach to stability, optimisation and performance analysis for Markovian multiclass queueing networks. *Ann Oper Res* 1999; 92: 1–18.
25. Huang JW, Bertz RA, Honig ML. Wireless scheduling with hybrid ARQ. *IEEE Trans Wireless Commun* 2005;4(6):2801–2810.
26. Niño-Mora J. Stochastic scheduling. In: Floudas CA, Pardalos PM, editors. Encyclopedia of optimization. 2nd ed. New York: Springer; 2009. pp. 3818–3824.
27. Smith WE. Various optimizers for single-stage production. *Nav Res Log Q* 1956;3(1–2): 59–66.
28. Rothkopf MH. Scheduling with random service times. *Manag Sci* 1966;12(9):707–713.
29. Horn WA. Single-machine job sequencing with treelike precedence ordering and linear delay penalties. *SIAM J Appl Math* 1972;23(2):189–202.
30. Sevcik KC. Scheduling for minimum total loss using service time distributions. *J Assoc Comput Mach* 1974;21(1):66–75.
31. Bruno JL, Hofri M. On scheduling chains of jobs on one processor with limited preemption. *SIAM J Comput* 1975;4(4):478–490.
32. Sidney JB. Decomposition algorithms for single-machine sequencing with precedence relations and deferral costs. *Oper Res* 1975;23(2):283–298.
33. Bruno JL, Coffman EG Jr, Johnson DB. On batch scheduling of jobs with stochastic service and cost structures on a single server. *J Comput Syst Sci* 1976;12(3):319–335.

34. Bruno JL. Sequencing jobs with stochastic task structures on a single machine. *J Assoc Comput Mach* 1976;23(4):655–664.
35. Meilijson I, Weiss G. Multiple feedback at a single-server station. *Stoch Process Appl* 1977;5(2):195–205.
36. Gittins JC. Multi-armed bandit allocation indices. Chichester: Wiley; 1989.
37. Gittins JC, Jones DM. A dynamic allocation index for the sequential design of experiments. In: Gani J, Sarkadi K, Vincze I, editors. *Progress in statistics (European Meeting of Statisticians, Budapest, 1972)*. Amsterdam, The Netherlands: North-Holland Publishing Co.; 1974. pp. 241–266.
38. Bradt RN, Johnson SM, Karlin S. On sequential designs for maximizing the sum of n observations. *Ann Math Stat* 1956;27(4):1060–1074.
39. Bellman R. A problem in the sequential design of experiments. *Sankhya* 1956;16(3–4):221–229.
40. Gittins JC. Bandit processes and dynamic allocation indices. *J R Stat Soc Ser B* 1979;41(2):148–177.
41. Whittle P. Arm-acquiring bandits. *Ann Probab* 1981;9(2):284–292.
42. Whittle P. Volume 1, Optimization over time: dynamic programming and stochastic control. New York: John Wiley & Sons, Inc.; 1982.
43. Weiss G. Branching bandit processes. *Probab Eng Inf Sci* 1988;2:269–278.
44. Lai TL, Ying Z. Open bandit processes and optimal scheduling of queueing networks. *Adv Appl Probab* 1988;20(2):447–472.
45. Varaiya PP, Walrand JC, Buyukkoc C. Extensions of the multiarmed bandit problem: the discounted case. *IEEE Trans Automat Control* 1985;30(5):426–439.
46. Ji C. Sequential scheduling of priority queues and arm-acquiring bandits. *J Appl Math Simul* 1987;1(2):115–135.
47. Bertsimas D, Niño-Mora J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math Oper Res* 1996;21(2):257–306.
48. Whittle P. Tax problems in the undiscounted case. *J Appl Probab* 2005;42(3):754–765.
49. Niño-Mora J. Restless bandits, partial conservation laws and indexability. *Adv Appl Probab* 2001;33(1):76–98.
50. Niño-Mora J. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Math Program* 2002;93(3):361–413.
51. Whittle P. Restless bandits: activity allocation in a changing world. In: Gani J, editor. Volume 25A, A celebration of applied probability, *Journal of Applied Probability*. Sheffield: Applied Probability Trust; 1988. pp. 287–298.
52. Niño-Mora J. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP* 2007;15(2):161–198.
53. Kleinrock L. Communication nets: stochastic message flow and delay. New York: McGraw-Hill; 1964. Reprinted by New York: Dover Publications, Inc.; 1972.
54. Kleinrock L. A conservation law for a wide class of queueing disciplines. *Nav Res Log Q* 1965;12(2):181–192.
55. Coffman EG Jr, Mitrani I. A characterization of waiting time performance realizable by single-server queues. *Oper Res* 1980;28(3):810–821.
56. Gelenbe E, Mitrani I. Analysis and synthesis of computer systems. New York: Academic Press; 1980.
57. Kameda H. Realizable performance vectors of a finite-source queue. *Oper Res* 1984;32(6):1358–1367.
58. Federgruen A, Groenevelt H. M/G/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules. *Manag Sci* 1988;34(9):1121–1138.
59. Federgruen A, Groenevelt H. Characterization and optimization of achievable performance in general queueing systems. *Oper Res* 1988;36(5):733–741.
60. Shanthikumar JG, Yao DD. Multiclass queueing systems: polymatroidal structure and optimal scheduling control. *Oper Res* 1992;40, Suppl 2: S293–S299.
61. Edmonds J. Matroids and the greedy algorithm. *Math Program* 1971;1(1):127–136.
62. Tsoucas P. The region of achievable performance in a model of Klimov. Technical Report RC-16543. Yorktown Heights (NY): IBM Research Division, T. J. Watson Research Center; 1991.
63. Bhattacharya PP, Georgiadis L, Tsoucas P. Extended polymatroids: properties and optimization. In: Balas E, Cornuejols G, Pulleyblank WR, editors. *Proceedings 2nd Conference Integer Programming and Combinatorial Optimization (IPCO II)*. Pittsburgh

- (PA): Mathematical Programming Society, Carnegie Mellon University; 1992. pp. 298–315.
64. Bhattacharya PP, Georgiadis L, Tsoucas P. Problems of adaptive optimization in multiclass M/GI/1 queues with Bernoulli feedback. *Math Oper Res* 1995;20(2):355–380.
 65. Tsoucas P, Walrand J. Optimal adaptive server allocation in a network. *Syst Control Lett* 1986;7(4):323–327.
 66. Makowski AM, Shwartz A. Stochastic approximations and adaptive control of a discrete-time single-server network with random routing. *SIAM J Control Optim* 1992;30(6):1476–1506.
 67. Chen H, Yao DD. Fundamentals of queueing networks: performance, asymptotics, and optimization. New York: Springer; 2001.
 68. Yao DD, Zhang L. Dynamic scheduling of a class of stochastic systems: extended polymatroid, side constraints, and optimality. *Proceedings 36th IEEE Conference on Decision and Control*. New York (NY): IEEE; 1997. pp. 1191–1196.
 69. Yao DD, Zhang L. Stochastic scheduling via polymatroid optimization. In: Yin GG, Zhang Q, editors. Volume 33, *Mathematics of stochastic manufacturing systems, Lectures in Applied Mathematics*. Providence (RI): AMS; 1997. pp. 333–364.
 70. Glazebrook KD, Niño-Mora J. Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance. *Oper Res* 2001; 49(4):609–623.
 71. Nain P, Ross KW. Optimal priority assignment with hard constraint. *IEEE Trans Automat Control* 1986;31(10):883–888.
 72. Makowski AM, Shwartz A. On constrained optimization of the Klimov network and related Markov decision processes. *IEEE Trans Automat Control* 1993;38(2):354–359.
 73. Simon B. Priority queues with feedback. *J Assoc Comput Mach* 1984;31(1):134–149.
 74. Jewkes EM, Buzacott JA. Flow time distributions in a K class M/G/1 priority feedback queue. *Queueing Syst* 1991;8(2):183–202.
 75. Chen H, Yao DD. Dynamic scheduling of a multiclass fluid network. *Oper Res* 1993;41(6): 1104–1115.
 76. Connors D, Feigin G, Yao D. Scheduling semiconductor lines using a fluid network model. *IEEE Trans Rob Autom* 1994;10(2):88–98.
 77. Bäuerle N, Rieder U. Optimal control of single-server fluid networks. *Queueing Syst* 2000;35(1–4):185–200.
 78. Bäuerle N, Stidham S. Conservation laws for single-server fluid networks. *Queueing Syst* 2001;38(2):185–194.
 79. Gajrat S, Hordijk A. On the structure of the optimal server control for fluid networks. *Math Methods Oper Res* 2005;62(1):55–75.
 80. Lu Y, Yao DD. Optimal control of a fluid network with side constraints. *IEEE Trans Automat Control* 2003;48(10):1865–1870.

k -OUT-OF- n SYSTEMS

MING J. ZUO

Department of Mechanical
Engineering, University of
Alberta, Edmonton, Alberta,
Canada

ZHIGANG TIAN

Concordia Institute for Information
Systems Engineering, Concordia
University, Montreal, Quebec,
Canada

In this article, we discuss the k -out-of- n system structure, which is a widely used type of redundancy in both industrial and military systems. First, we investigate the binary k -out-of- n systems, where the system and the components can take only two possible states, completely working or totally failed. The multistate k -out-of- n systems and the weighted multistate k -out-of- n systems are also discussed, where the systems and the components can take multiple discrete states. We limit our discussion of these models for the evaluation of system state distribution given the component state distributions. For topics like order statistics and advanced probabilistic analysis of such systems, readers are referred to Arnold *et al.* [1], Bai and Blasch [2], and Hollander and Pena [3].

BINARY k -OUT-OF- n SYSTEMS

Binary k -out-of- n Systems Concepts

The definition of the binary k -out-of- n system is given as follows:

Definition 1. An n -component system is called a binary k -out-of- n :G system if it is working whenever at least k components are working. An n -component system is called a binary k -out-of- n :F system, if it is failed whenever at least k components are failed.

The k -out-of- n system structure is a very popular type of redundancy in fault-tolerant

systems, with wide applications in both industrial and military systems [4]. For example, it may be possible to drive a car with a V8 engine if only four cylinders are firing. However, if less than four cylinders fire, the automobile cannot be driven. Thus, the functioning of the engine may be represented by a 4-out-of-8:G system. The system is tolerant of failures up to four cylinders for minimal functioning of the engine. In a data processing system with five video displays, a minimum of three displays operable may be sufficient for full data display. In this case, the display subsystem behaves as a 3-out-of-5:G system. In a communications system with three transmitters, the average message load may be such that at least two transmitters must be operational at all times or critical messages may get lost. Thus, the transmission subsystem functions as a 2-out-of-3:G system. Systems with spares may also be represented by the k -out-of- n system model. In the case of an automobile with four tires, for example, usually one additional spare tire is equipped on the vehicle. Thus, the vehicle can be driven as long as at least 4-out-of-5 tires are in good condition [4].

From the definition of binary k -out-of- n :G and k -out-of- n :F systems, it can be seen that redundancy is generally built into a k -out-of- n system, since the value of n is usually larger than the value of k . It can also be observed that both series and parallel systems are special cases of the k -out-of- n system. A series system is equivalent to an n -out-of- n :G system, or a 1-out-of- n :F system. And a parallel system is equivalent to a 1-out-of- n :G system, or an n -out-of- n :F system.

There is an equivalence relationship between the k -out-of- n :G system structure and the k -out-of- n :F system structure. On the basis of the definitions of these two types of systems, a k -out-of- n :G system is equivalent to an $(n - k + 1)$ -out-of- n :F system. Similarly, a k -out-of- n :F system is equivalent to an $(n - k + 1)$ -out-of- n :G system. This means

that provided the systems have the same set of component reliabilities, the reliability of a k -out-of- n :G system is equal to the reliability of an $(n - k + 1)$ -out-of- n :F system and the reliability of a k -out-of- n :F system is equal to the reliability of an $(n - k + 1)$ -out-of- n :G system [4].

Reliability Evaluation of Binary k -out-of- n Systems

In this section, we focus on the methods for reliability evaluation of k -out-of- n :G systems. These methods can also be used to evaluate the reliability of a k -out-of- n :F system, first by converting it to its equivalent k -out-of- n :G system. We discuss the reliability evaluation of k -out-of- n systems with identically and independently distributed (i.i.d.) components, and the reliability evaluation of systems with independent components.

Notation:

- n : number of components in the system;
- k : minimum number of components that must work for the k -out-of- n :G system to work;
- p_i : reliability of component i , $i = 1, 2, \dots, n$;
- p : reliability of each component when all components are i.i.d.;
- q_i : unreliability of component i , $q_i = 1 - p_i$, $i = 1, 2, \dots, n$;
- q : unreliability of each component when all components are i.i.d., $q = 1 - p$;
- $R(k, n)$: reliability of a k -out-of- n :G system or probability that at least k out of the n components are working, where $0 \leq k \leq n$ and both k and n are integers;
- $Q(k, n)$: unreliability of a k -out-of- n :G system or probability that less than k out of the n components are working, where $0 \leq k \leq n$ and both k and n are integers, $Q(k, n) = 1 - R(k, n)$.

Reliability Evaluation of Binary k -out-of- n Systems with i.i.d. Components. In a k -out-of- n :G system with i.i.d. components, the

number of working components follows the binomial distribution with parameters n and p . Thus, we have

$$\begin{aligned} & \Pr(\text{Exactly } i \text{ components work}) \\ &= \binom{n}{i} p^i q^{n-i}, \quad i = 0, 1, 2, \dots, n. \end{aligned} \quad (1)$$

The reliability of the system is equal to the probability that the number of working components is greater than or equal to k .

$$R(k, n) = \sum_{i=k}^n \binom{n}{i} p^i q^{n-i}. \quad (2)$$

Equation (2) is an explicit formula that can be used for reliability evaluation of the k -out-of- n :G system with i.i.d. components.

Reliability Evaluation of Binary k -out-of- n Systems with Independent Components. For systems with independent components, the reliability values of different components can be different. Efficient reliability evaluation algorithms for binary k -out-of- n systems with independent components have been provided by Barlow and Heidtmann [5] and Rushdi [6], as follows:

$$\begin{aligned} R(k, n) &= p_n \cdot R(k-1, n-1) \\ &\quad + q_n \cdot R(k, n-1). \end{aligned} \quad (3)$$

The boundary conditions are

$$\begin{aligned} R(0, n) &= 1, \quad \text{for } n \geq 0, \\ R(k, n) &= 0, \quad \text{for } 0 \leq n < k. \end{aligned} \quad (4)$$

The Binary-Weighted k -out-of- n Systems

Wu and Chen generalized the binary k -out-of- n system models into the binary-weighted k -out-of- n models [7].

Definition 2. In a binary-weighted k -out-of- n system, component i carries a weight of w_i , $w_i > 0$ for $i = 1, 2, \dots, n$. The system works if and only if the total weight of working components is at least k , a prespecified value.

The “weight” here means the contribution of the component. In the binary-weighted system, the component with higher contribution

to the system has a higher “weight.” The component with lower contribution to the system has a lower “weight.” For example, a power generator with a higher output capacity has a larger “weight.”

A recursive equation is provided by Wu and Chen [7]. In the following, $R_w(i, j)$ represents the probability that a system with j components can output a total weight of at least i . Then, $R_w(k, n)$ is the reliability of the weighted k -out-of- n :G system. The following recursive equation can be used for reliability evaluation of such systems:

$$R_w(i, j) = p_j R_w(i - w_j, j - 1) + q_j R_w(i, j - 1), \quad (5)$$

which requires the following boundary conditions:

$$R_w(i, j) = 1, \quad \text{for } i \leq 0, \quad j \geq 0, \quad (6)$$

$$R_w(i, 0) = 0, \quad \text{for } i > 0. \quad (7)$$

When $w_j = 1$ for $1 \leq j \leq n$, the binary-weighted k -out-of- n model is reduced to the regular k -out-of- n system model. Chen and Yang developed a two-stage weighted- k -out-of- n system model with components in common, as well as the reliability evaluation algorithm [8]. Levitin and Lisnianski presented a reliability optimization approach for weighted voting systems [9].

MULTISTATE *k*-OUT-OF-*n* SYSTEMS

Multistate k -out-of- n System Concepts

Many practical components and systems have more than two different performance levels. For example, a power generator in a power station can work at full capacity, which is its nominal capacity, say 10 MW, when there are no failures at all [10]. Certain type of failures can cause the generator to be completely failed, while other failures will lead to the generator working at a reduced capacity say at 4 MW. A component with multiple failure modes can also be considered to be a multistate component [11–13]. Based on the theory of binary coherent systems [14], many system structures, such as series-parallel structure, k -out-of- n structure, and

network structure, have been extended from the binary cases to the multistate cases by allowing the components and the systems to take more than two possible states [4,15–19]. With multistate models, we can deal with reliability issues of engineering systems more accurately, and would be able to answer more operations related questions.

When a component may experience more than two possible states, its state may be represented by a continuous random variable or a discrete random variable. In this article, we adopt the discrete assumption. Each component and the system may experience $M + 1$ possible states represented by $0, 1, 2, \dots, \text{and } M$, where 0 denotes the worst state and M (a positive integer) denotes the best state.

El-Newehi *et al.* [20] defined the first multistate k -out-of- n system model, where the system state was defined as the state of the k th best component. In another word, for the system to be in state j or above, there should be at least k components in state j or above. That is, the k value is the same with respect to all states. Boedigheimer and Kapur [21] defined the multistate k -out-of- n model from the perspective of lower and upper boundary points, and their definition is actually consistent with that by El-Newehi *et al.* Huang *et al.* [22] proposed the generalized multistate k -out-of- n :G system model, where there can be different k values with respect to different states. In the following two definitions, $\phi(\mathbf{x})$ is the structure function of the system, that is, it denotes the state of the system as a function of the component state vector $\mathbf{x}$.

Definition 3. An n -component system is called a generalized multistate k -out-of- n :G system if $\phi(\mathbf{x}) \geq j$ ($1 \leq j \leq M$) whenever there exists an integer value l ($j \leq l \leq M$) such that at least k_l components are in state l or above. In addition, if we have $k_1 \leq k_2 \leq \dots \leq k_M$, the system is called an increasing multistate k -out-of- n :G system and if we have $k_1 \geq k_2 \geq \dots \geq k_M$, the system is called a decreasing multistate k -out-of- n :G system.

Zuo and Tian [23] proposed the definition of the generalized multistate k -out-of- n :F system, which is the mirror image of the

corresponding multistate *k*-out-of-*n*:G system defined in Definition 3.

Definition 4. An *n*-component system is called a generalized multistate *k*-out-of-*n*:F system if $\phi(\mathbf{x}) < j$ ($1 \leq j \leq M$) whenever the states of at least k_l components are below l for all l such that $j \leq l \leq M$.

Reliability Evaluation of the Multistate *k*-out-of-*n* Systems

Notation:

- M : the maximum state level of a multistate system and its components;
- x_i : state of component i , $x_i = j$ if component i is in state j , $0 \leq j \leq M$, $1 \leq i \leq n$;
- $\mathbf{x}$: an n -dimensional vector representing the states of all components, $\mathbf{x} = (x_1, x_2, \dots, x_n)$;
- $\phi(\mathbf{x})$: state of the system, $0 \leq \phi(\mathbf{x}) \leq M$;
- k_j : the k value with respect to level j of a generalized multistate *k*-out-of-*n* system;
- $\mathbf{v}_j$: a minimal cut vector to level j of an increasing multistate *k*-out-of-*n*:F system: $\phi(\mathbf{v}_j) < j$ and $\phi(\mathbf{x}) \geq j$ for all $\mathbf{x} > \mathbf{v}_j$;
- $r_{s,j}$: $\Pr(\phi(\mathbf{x}) = j)$;
- $\mathbf{k}$: the $\mathbf{k}$ vector of a generalized multistate *k*-out-of-*n* system, $\mathbf{k} = (k_1, k_2, \dots, k_M)$;
- $p_{n,j}$: the probability of component n in state j ;
- $\mathbf{k}^j$: the generated $\mathbf{k}$ vector when component n is in state j .

For any generalized multistate *k*-out-of-*n*:G system given in Definition 3, there is an equivalent generalized multistate *k*-out-of-*n*:F system given in Definition 4. The reliability evaluation of a generalized multistate *k*-out-of-*n*:G system can be performed via its equivalent generalized multistate *k*-out-of-*n*:F system.

Before presenting the algorithms, the definition of nominal increasing multistate *k*-out-of-*n*:F system proposed in Ref. 23 needs be introduced.

Definition 5. A nominal increasing multistate *k*-out-of-*n*:F system is the same as an increasing multistate *k*-out-of-*n*:F system except that the probability of a component in all possible states may be less than 1.

The probability for the state of a generalized multistate *k*-out-of-*n*:F system to be below j can be calculated through a generated nominal increasing multistate *k*-out-of-*n*:F system [23]. Thus, the following discussion is focused on the increasing case.

The recursive function in the algorithm is denoted by $Q(m, N, \mathbf{k}, \mathbf{p})$, where $m + 1$ is the number of possible states of the nominal increasing multistate *k*-out-of-*n*:F system, N is the number of components of the nominal increasing multistate *k*-out-of-*n*:F system, $\mathbf{k}$ is the characteristic vector of the nominal increasing multistate *k*-out-of-*n*:F system where $\mathbf{k} = (k_1, k_2, \dots, k_m)$, and $\mathbf{p}$ is the probability vector of the nominal increasing multistate *k*-out-of-*n*:F system where $\mathbf{p} = (p_0, p_1, \dots, p_m)$ [23].

The recursive function $Q(m, N, \mathbf{k}, \mathbf{p})$ is designed to represent the probability for the system to be in state “0” of the nominal increasing multistate *k*-out-of-*n*:F system with $m + 1$ possible states, totally N components, vector $\mathbf{k}$, and probability vector $\mathbf{p}$. Thus, to calculate the probability for the state of a generalized multistate *k*-out-of-*n*:F system to be below j , we can first transform it into a nominal increasing multistate *k*-out-of-*n*:F system to state j and to do the calculation using the recursive algorithm.

The procedure of the recursive algorithms is as follows:

$$Q(m, N, \mathbf{k}, \mathbf{p}) = \sum_{i=k_1}^{k_2-1} \binom{N}{i} p_0^i \cdot \times Q(m-1, N-i, \dot{\mathbf{k}}, \dot{\mathbf{p}}) + Q(m-1, N, \ddot{\mathbf{k}}, \ddot{\mathbf{p}}), \quad (8)$$

where $\dot{\mathbf{k}} = (k_2 - i, k_3 - i, \dots, k_m - i)$, $\dot{\mathbf{p}} = (p_1, p_2, \dots, p_m)$, $\ddot{\mathbf{k}} = (k_2, k_3, \dots, k_m)$, and $\ddot{\mathbf{p}} = (p_0, p_1 + p_2, p_3, \dots, p_m)$. The boundary condition for the recursive algorithm is

$$Q(1, N, \mathbf{k}, \mathbf{p}) = \sum_{i=k_1}^N \binom{N}{i} p_0^i \cdot p_1^{N-i}. \quad (9)$$

From Equations (8) and (9), we can see that this algorithm is actually recursive on the parameter m , not on the number of components N . Therefore, we can apply the recursive algorithm to large systems including a large number of components, without leading to exponential growth of computation time.

The following example is used to illustrate the recursive algorithm for increasing multistate k -out-of- n :F systems described in this section.

Example 1. We consider an increasing multistate k -out-of- n :F system with 10 i.i.d. components and 4 possible states, that is, $n = 10$ and $M = 3$. The $\mathbf{k}$ vector is $\mathbf{k} = (k_1, k_2, k_3) = (3, 6, 8)$. The state distribution of components is $\mathbf{p} = (0.1, 0.3, 0.4, 0.2)$.

To calculate $\Pr(\phi(\mathbf{x}) < 3)$, we get a nominal increasing multistate k -out-of- n :F system to state 3, with $m = 1, N = 10, \mathbf{k} = (k_3) = 8$, and $\mathbf{p} = (p_0 + p_1 + p_2, p_3) = (0.8, 0.2)$. Using the recursive algorithm, the value $\Pr(\phi(\mathbf{x}) < 3)$ is equal to $Q(m, N, \mathbf{k}, \mathbf{p})$, which is 0.6778.

To calculate $\Pr(\phi(\mathbf{x}) < 2)$, we get a nominal increasing multistate k -out-of- n :F system to state 2, with $m = 2, N = 10, \mathbf{k} = (k_2, k_3) = (6, 8)$, and $\mathbf{p} = (p_0 + p_1, p_2, p_3) = (0.4, 0.4, 0.2)$. Using the recursive algorithm, the value $\Pr(\phi(\mathbf{x}) < 2)$ is equal to $Q(m, N, \mathbf{k}, \mathbf{p})$, which is 0.1523.

To calculate $\Pr(\phi(\mathbf{x}) < 1)$, we get a nominal increasing multistate k -out-of- n :F system to state 1, with $m = 3, N = 10, \mathbf{k} = (k_1, k_2, k_3) = (3, 6, 8)$, and $\mathbf{p} = (p_0, p_1, p_2, p_3) = (0.1, 0.3, 0.4, 0.2)$. Using the recursive algorithm, the value $\Pr(\phi(\mathbf{x}) < 1)$ is equal to $Q(m, N, \mathbf{k}, \mathbf{p})$, which is 0.0308.

We can get the probability of the system at each individual state $r_{s,0} = 0.0308, r_{s,1} = 0.1214, r_{s,2} = 0.5255, r_{s,3} = 0.3222$

The algorithm presented above is for evaluating multistate k -out-of- n systems with i.i.d. components [23]. Tian *et al.* extended this algorithm and developed a method for evaluating multistate k -out-of- n

systems with independent components [24,25]. A bounding approach was also proposed for estimating the reliability of multistate k -out-of- n systems in a much shorter time [26]. Amari *et al.* developed a more efficient algorithm for reliability evaluation of multistate k -out-of- n systems with i.i.d. components [27]. Tian *et al.* also proposed another practical model for multistate k -out-of- n systems, and an algorithm for the reliability evaluation [25].

MULTISTATE-WEIGHTED k -OUT-OF- n SYSTEMS

In the binary-weighted k -out-of- n :G system model discussed in the section titled “The Binary-Weighted k -out-of- n Systems,” the weight of the system is the sum of the weights of all working components. When a component is failed, its contribution to the system is zero.

In a multistate context, a component may be in different states. When it is in a different state, it may have a different contribution to the system. When it is completely failed, its contribution to the system is zero. Suppose there are n components in a system, each component may be in $M + 1$ states: $0, 1, 2, \dots, M$, and component i , when in state j , has a weight value of w_{ij} . In the multistate-weighted k -out-of- n :G model to be covered in this section, every component in every possible state has certain contribution to the system’s performance. The formal definition of Model I of the multistate-weighted k -out-of- n :G system is given below [28].

Definition 6. The system is in state j or above if the total weight of all components is equal to or greater than k_j , a prespecified value.

Let ϕ be the structure function of the system representing the state of the system and W the total weight of all components. Then, this definition means $\Pr \phi \geq j = \Pr W \geq k_j$. Since state 0 is the worst state of the system, we have $\Pr \phi \geq 0 = 1$.

According to Definition 3, for the system state to be not lower than a given value j , the

number of components whose states are not lower than j must be at least k_j , a prespecified value. In this definition, the components whose states are below j do not make any contribution for the system to be in state j or above. On the basis of this idea, Li and Zuo define another model of the multistate-weighted k -out-of- n :G system [28]. In the new model (referred to as *Model II* below), for the system to be in state j or above, the sum of the weights of only those components, whose states are in state j or above must be not less than k_j , a prespecified value. The difference between Model I presented above and Model II to be presented next is whether the components whose states are below j are making any contribution for the system to be in state j or above. The formal definition of Model II is given below.

Definition 7. The system is in state j or above if the total weight of the components in state j or above is equal to or greater than k_j , a prespecified value.

Let ϕ be the structure function of the system and W_j be the sum of the weights of the components whose states are j or above. We then have $\Pr \phi \geq j = \Pr W_j \geq k_j$ based on Model II.

Example 2. Consider a conveyor belt set, in which several conveyors work together in parallel. During the life of a conveyor, its state will deteriorate from perfect to failed. When its state degrades, it cannot work under the full load any more. That means corresponding to different states, it has different weights. When we require that in the set, the total drive of all the conveyors to be at least 100 KW to transport coal no matter what the state of each conveyor is, it is an multistate-weighted k -out-of- n system Model I. If we require that in the set, the total drive generated by the conveyors which still can work under full load to be at least 100 KW, it is the multistate-weighted k -out-of- n system Model II.

The reliability evaluation algorithms for the multistate-weighted k -out-of- n system models were presented in Ref. 28. Li and Zuo

also developed an optimal design approach for multistate-weighted k -out-of- n systems based on component design [29].

CONCLUDING REMARKS

The k -out-of- n structure is a very flexible redundancy structure, and has been widely used in many systems. Meanwhile, in many real applications, components and systems might have multiple states. "Multiple states" of a component or system might be interpreted as multiple levels of performance, multiple functions, multiple levels of benefits, or multiple failure modes. The multistate k -out-of- n system models reported in this article enable us to model multiple levels of k -out-of- n requirements of the system. The reliability evaluation algorithms have been presented for efficient performance evaluations of the k -out-of- n system models. A key future work is to do case studies of practical systems with multiple levels of k -out-of- n requirements. It is also important to develop more general models for more general and complex situations.

REFERENCES

1. Arnold BC, Balakrishnan N, Nagaraja HN. A first course in order statistics. New York: John Wiley & Sons Inc.; 1992.
2. Bai L, Blasch E. Two-way handshaking circular sequential k -out-of- n congestion system. IEEE Trans Reliab 2008;57(1):59–70.
3. Hollander M, Pena EA. Dynamic reliability models with conditional proportional hazards. J Lifetime Data 1995;1(4):377–401.
4. Kuo W, Zuo MJ. Optimal reliability modeling: principles and applications. New York: John Wiley & Sons, Inc.; 2003.
5. Barlow RE, Heidtmann KD. Computing k -out-of- n system reliability. IEEE Trans Reliab 1984;33:322–323.
6. Rushdi AM. Utilization of symmetric switching functions in the computation of k -out-of- n system reliability. Microelectron Reliab 1986;26:973–987.
7. Wu JS, Chen RJ. An algorithm for computing the reliability of a weighted- k -out-of- n system. IEEE Trans Reliab 1994;43:327–328.

8. Chen Y, Yang QY. Reliability of two-stage weighted- k -out-of- n systems with components in common. *IEEE Trans Reliab* 2005;54(3):431–440.
9. Levitin G, Lisnianski A. Reliability optimization for weighted voting system. *Reliab Eng Syst Saf* 2001;71:131–138.
10. Lisnianski A, Levitin G. multistate system reliability: assessment, optimization and applications. Singapore: World Scientific; 2003.
11. Elsayed EA. Reliability engineering. New York: Addison Wesley Longman; 1996.
12. Levitin G, Lisnianski A. Structure optimization of multistate system with two failure modes. *Reliab Eng Syst Saf* 2001;72:75–89.
13. Levitin G. Optimal series-parallel topology of multistate system with two failure modes. *Reliab Eng Syst Saf* 2002;77:93–107.
14. Barlow RE, Proschan F. Statistical theory of reliability and life testing. New York: Holt, Rinehart and Winston; 1975.
15. Huang J, Zuo MJ, Fang Z. multistate consecutive- k -out-of- n systems. *IIE Trans* 2003;35(6):527–534.
16. Levitin G. Universal generating function in reliability analysis and optimization. London: Springer-Verlag; 2005.
17. Levitin G, Lisnianski A, Ben-Haim H, *et al.* Redundancy optimization for series-parallel multistate systems. *IEEE Trans Reliab* 1998;47(2):165–172.
18. Natvig B. Two suggestions of how to define a multistate coherent system. *Appl Probab* 1982;14:391–402.
19. Yamamoto H, Zuo MJ, Akiba T, *et al.* Recursive formulas for the reliability of multistate consecutive- k -out-of- n :G systems. *IEEE Trans Reliab* 2006;55(1):98–104.
20. El-Newehi E, Proschan F, Sethuraman J. multistate coherent system. *J Appl Probab* 1978;15:675–688.
21. Boedigheimer RA, Kapur KC. Customer-driven reliability models for multistate coherent systems. *IEEE Trans Reliab* 1994;43:46–50.
22. Huang J, Zuo MJ, Wu YH. Generalized multistate k -out-of- n :G systems. *IEEE Trans Reliab* 2000;49:105–111.
23. Zuo MJ, Tian Z. Performance evaluation for generalized multistate k -out-of- n systems. *IEEE Trans Reliab* 2006;55(2):319–327.
24. Tian Z, Zuo MJ, Yam RCM. Performance evaluation of generalized multistate k -out-of- n systems with independent components. Proceedings of the 4th International Conference on Quality and Reliability; 2005 Aug 9–11; Beijing, China; 2005. pp. 515–520.
25. Tian Z, Zuo MJ, Yam RCM. The multistate k -out-of- n systems and their performance evaluation. *IIE Trans* 2009;41:32–44.
26. Tian Z, Richard RCM, Zuo MJ, *et al.* Reliability bounds for multistate k -out-of- n systems. *IEEE Trans Reliab* 2008;57:303–310.
27. Amari SV, Zuo MJ, Dill G. A fast and robust reliability evaluation algorithm for generalized multistate k -out-of- n systems. *IEEE Trans Reliab* 2009;58:88–97.
28. Li W, Zuo MJ. Reliability evaluation of multistate weighted k -out-of- n systems. *Reliab Eng Syst Saf* 2008;93:160–167.
29. Li W, Zuo MJ. Optimal design of multistate weighted k -out-of- n systems based on component design. *Reliab Eng Syst Saf* 2008;93:1673–1681.

LAGRANGIAN OPTIMIZATION FOR LP

JITAMITRA DESAI
Division of Systems and Engineering
Management, School of
Mechanical and Aerospace
Engineering, Nanyang
Technological University,
Singapore

INTRODUCTION

In many practical applications, there exists a need for repeatedly solving large-scale *linear programming* (LP) problems, especially when such LPs arise as subroutines in a branch-and-bound framework or other polyhedral schemes. It has been amply demonstrated in the literature that solving large-scale linear programs via traditional approaches such as simplex-based algorithms or even interior point methods can sometimes be computationally cumbersome, particularly for cases where the coefficient matrix is dense and ill-conditioned (see Ref. 1, as well as the computational experience in Ref. 2). In such instances, a *Lagrangian optimization* approach can prove to be a very useful tool. Early work related to this approach appeared in Ref. 3, and since then there have been numerous developments in this field, particularly geared toward applying Lagrangian techniques for solving nonlinear and integer programming (IP) problems [1,4–7].

In this article, we focus on developing the Lagrangian optimization approach for solving LP problems. To motivate this methodology, consider the following generic linear program:

$$P: \text{Minimize } \{c^T x \mid Ax \geq b, Cx \geq d, x \geq 0\}, \quad (1)$$

where, following standard notation, $x \in \mathbb{R}_+^n$, $A \in \mathbb{R}^{p \times n}$, $C \in \mathbb{R}^{q \times n}$, $b \in \mathbb{R}^p$, and $d \in \mathbb{R}^q$.

Now, suppose that the set of constraints $Ax \geq b$ are “complicating” constraints, and thus, removing these constraints ensures that this modified problem can be solved fairly easily in comparison to the original Problem P. In a Lagrangian scheme, these complicating constraints are “relaxed” by weighting them with (nonnegative) multipliers $\pi \in \mathbb{R}_+^p$ to obtain the following *Lagrangian relaxation* (LR_π) of Problem P:

$$LR_\pi: \text{Minimize } \{c^T x + \pi^T(b - Ax) \mid Cx \geq d, x \geq 0\}. \quad (2)$$

Remark 1. To be precise, LR_π , as defined in Equation (2), is the *Lagrangian relaxation of Problem P relative to $Ax \geq b$ and a conformable vector $\pi \geq 0$* . For this Lagrangian relaxation to be computationally superior (in comparison to the original Problem P), the remaining constraints $Cx \geq d$ must admit an easy-to-find closed form solution or retain considerable special structure which can be exploited in solving the problem. Examples of such constraints include simple variable upper bounds, flow-balance constraints, block diagonal structures, certain types of inventory constraints, or generalized upper bounding constraints. For detailed examples, we refer the reader to Ref. 5.

In LR_π , notice that the (nonpositive) slack variables of the constraints $Ax \geq b$ have been appended to the objective function of Problem P, weighted with the nonnegative vector π , and moreover, the feasible region has been relaxed by eliminating the complicating constraints. The constraints $Ax \geq b$ are said to have been “dualized,” and the nonnegative weights π used to dualize the complicating constraints are commonly referred to as *Lagrange multipliers*. Given Equation (2), the following propositions hold true trivially.

Proposition 1. *The feasible region of P is a subset of the feasible region of LR_π .*

Proof. The result follows by virtue of relaxing the complicating constraints.

Proposition 2. $v[\text{LR}_{\pi}] \leq v[\text{P}]$, where $v[\cdot]$ denotes the optimal objective function value for any Problem $[\cdot]$.

Proof. Given any $\bar{\pi} \in \mathbb{R}_+^p$, and any solution $\bar{x}$ feasible to Problem P, note that the objective function value of Problem $\text{LR}_{\bar{\pi}}$ is lesser than the objective function value of Problem P; that is,

$$v[\text{LR}_{\bar{\pi}}] \leq c^T \bar{x} + \bar{\pi}^T (b - A\bar{x}) \leq c^T \bar{x}.$$

This inequality, along with Proposition 1, yields the required result.

Using Propositions 1 and 2, we know that for any given $\pi \in \mathbb{R}_+^p$, the optimal solution to Problem LR_{π} provides a valid *lower bound* on Problem P, and this lower bound clearly depends on the value of the Lagrange multipliers $\pi \geq 0$ used in dualizing the complicating constraints. Hence, to obtain the *best* (or *largest*) lower bound for a minimization problem, an appropriate choice of Lagrange multipliers can be realized by *maximizing* LR_{π} over all $\pi \geq 0$, which yields the following *Lagrangian dual* (LD) problem:

$$\text{LD: Maximize } \{\text{LR}_{\pi} \mid \pi \geq 0\}, \quad (3)$$

where LR_{π} is given by Equation (2). It is worth noting here that LR_{π} is a problem in the original variables x , whereas LD is a problem in the so-called “dual space” of the Lagrange multipliers π .

Remark 2. The procedure outlined above for constructing the Lagrangian dual can be easily extended for the case of complicating *equality* constraints as well. If the complicating constraints in Problem P are given by $Ax = b$, then we can equivalently represent this as *two* sets of constraints, $Ax \geq b$ and $-Ax \geq -b$, and associate (conformable) Lagrange multipliers $\lambda \geq 0$ and

$\mu \geq 0$ for respectively dualizing these complicating constraints to obtain:

$$\text{LR}_{\lambda, \mu}: \text{Minimize } \left\{ c^T x + \lambda^T (b - Ax) + \mu^T (-b + Ax) \mid Cx \geq d, x \geq 0 \right\}, \quad (4)$$

which can be rewritten as the equivalent problem:

$$\text{LR}_{\pi}: \text{Minimize } \left\{ c^T x + \pi^T (b - Ax) \mid Cx \geq d, x \geq 0 \right\}, \quad (5)$$

where $\pi \equiv \lambda - \mu$. In this case, π is the difference of two nonnegative quantities and is therefore *unrestricted in sign*. Hence, Lagrange multipliers associated with equality constraints need not be nonnegative for LR_{π} to provide a valid lower bounding relaxation to Problem P.

The remainder of this article is organized as follows: In the section titled “Properties of the Lagrangian Dual”, several propositions and theorems that elucidate the properties of the Lagrangian optimization approach and characterize the Lagrangian dual are presented, followed by an illustrative example. Next, in the section titled “Solution Methodologies For The Lagrangian Dual,” the solution methodologies are described, and finally, the section titled “Extensions and Conclusions” briefly presents some extensions of the Lagrangian approach to integer programs and some concluding remarks.

PROPERTIES OF THE LAGRANGIAN DUAL

In this section, we prove some important results that characterize the properties of the Lagrangian dual and provide some interesting insights that relate the Lagrangian approach to classical LP duality theory.

Definition 1. For any $\pi \geq 0$, an optimal solution to Problem LR_{π} , denoted x_{π} , is called a *Lagrangian solution*.

Theorem 1. Let x_π be a Lagrangian solution corresponding to some $\pi \geq 0$. If x_π satisfies the following conditions:

- (i) $Ax_\pi \geq b$, and
- (ii) $\pi^T(b - Ax_\pi) = 0$,

then x_π is optimal to Problem P.

Proof. As x_π is a Lagrangian solution, we get

$$\begin{aligned} v[\text{LR}_\pi] &= c^T x_\pi + \pi^T(b - Ax_\pi) \\ &\leq c^T x + \pi^T(b - Ax), \quad \forall x \in \{Cx \geq d, x \geq 0\} \\ &\leq c^T x + \pi^T(b - Ax), \\ &\quad \forall x \in \{Ax \geq b, Cx \geq d, x \geq 0\} \\ &\leq \text{minimum} \{c^T x \mid Ax \geq b, Cx \geq d, x \geq 0\} \\ &= v[\text{P}]. \end{aligned}$$

Now, using condition (ii), set $\pi^T(b - Ax_\pi) = 0$ in the above inequality, to get

$$v[\text{LR}_\pi] = c^T x_\pi \leq v[\text{P}].$$

But, from condition (i), we know that x_π is also feasible to $Ax \geq b$, and as Problem P is a minimization problem, this results in

$$v[\text{LR}_\pi] = c^T x_\pi \geq v[\text{P}].$$

Hence proved.

So far, we have proved that a Lagrangian solution, which is feasible to the complicating constraints and satisfies “complementary slackness” conditions is also optimal to Problem P. Continuing in the same vein, Theorem 2 and the subsequent corollaries characterize other important properties of the Lagrangian dual, thereby providing the necessary framework for relating Lagrangian duality with traditional LP duality theory. The main crux of the argument lies in proving that the optimal objective function values of Problem LD and Problem P are equal, and furthermore, the Lagrange multipliers obtained from the optimal solution to Problem LD are indeed the same as the optimal dual variables associated with the complicating constraints

obtained by solving Problem P using standard LP techniques.

Theorem 2. $v[\text{P}] = v[\text{LD}]$.

Proof. Consider,

$$\begin{aligned} v[\text{LD}] &= \text{maximize} \{ \text{LR}_\pi \mid \pi \geq 0 \} \\ &= \text{maximize}_{\pi \geq 0} \left\{ \min \left\{ c^T x + \pi^T(b - Ax) \right. \right. \\ &\quad \left. \left. \mid Cx \geq d, x \geq 0 \right\} \right\}, \\ &= \text{maximize}_{\pi \geq 0} \left\{ \min \left\{ c^T x + \pi^T(b - Ax) \right. \right. \\ &\quad \left. \left. \mid x \in \{x^1, x^2, \dots, x^I\} \right\} \right\}, \end{aligned}$$

where $x^i, \forall i = 1, \dots, I$, are the *extreme points* of $Cx \geq d, x \geq 0$.

Note: For notational ease, we have assumed here that $Cx \geq d, x \geq 0$ is a nonempty *polytope*; that is, a closed and bounded polyhedron, and can be defined purely in terms of its extreme points with no recession directions. A very straightforward extension of the arguments presented here is applicable for the more general case.

Hence,

$$v[\text{LD}] = \text{maximize } v[\text{LR}_\pi] \quad (6a)$$

$$\begin{aligned} \text{subject to : } & v[\text{LR}_\pi] \leq c^T x^i \\ & + \pi^T(b - Ax^i), \quad \forall i = 1, \dots, I; \end{aligned} \quad (6b)$$

$$\pi \geq 0. \quad (6c)$$

Note that Problem (6a)–(6c) is a linear program, and so, associating dual variables u to the constraint (6b) as shown, and taking the regular LP dual, we get

$$v[\text{LD}] = \text{minimize } \sum_{i=1}^I (c^T x^i) u_i \quad (7a)$$

$$\text{subject to : } Ax^i \geq b, \quad \forall i = 1, \dots, I; \quad (7b)$$

$$\sum_{i=1}^I u_i = 1, \quad (7c)$$

$$u \geq 0. \quad (7d)$$

Recognizing that $\sum_{i=1}^I (c^T x^i) u_i \equiv c^T \left(\sum_{i=1}^I u_i x^i \right)$, and setting $x = \sum_{i=1}^I u_i x^i$, that is x is represented as a nonnegative convex combination of the extreme points of $Cx \geq d, x \geq 0$, we can rewrite Problem (7a)–(7d) as

$$\begin{aligned} v[\text{LD}] &= \text{Minimize} \left\{ c^T x \mid Ax \geq b, Cx \geq d, x \geq 0 \right\} \\ &\equiv v[\text{P}]. \end{aligned}$$

To further understand the mathematical properties of the Lagrangian dual, consider the following proposition that characterizes a saddle point optimum.

Definition 2. A point $(\bar{\pi}, \bar{x})$ with $\bar{\pi} \geq 0$ and $\bar{x}$ feasible to Problem $\text{LR}_{\bar{\pi}}$ is a *saddle point* for $\eta(\pi, x) = c^T x + \pi^T(b - Ax)$ if and only if,

$$\begin{aligned} \max_{\pi \geq 0} \eta(\pi, \bar{x}) &\leq \eta(\bar{\pi}, \bar{x}) \leq \min_{\{x \geq 0 \mid Cx \geq d\}} \eta(\bar{\pi}, x). \end{aligned} \quad (8)$$

Proposition 3. If $\bar{\pi} \geq 0$ and $\bar{x}$ is feasible to Problem $\text{LR}_{\bar{\pi}}$, then $(\bar{\pi}, \bar{x})$ is a saddle point for $\eta(\pi, x) = c^T x + \pi^T(b - Ax)$ if and only if

- (i) $\bar{x}$ is a Lagrangian solution to Problem $\text{LR}_{\bar{\pi}}$ at $\pi = \bar{\pi}$,
- (ii) $A\bar{x} \geq b$,
- (iii) $\bar{\pi}^T(b - A\bar{x}) = 0$.

Proof. Let $(\bar{\pi}, \bar{x})$ satisfy conditions (i)–(iii). Then, by definition of a Lagrangian solution, $\eta(\bar{\pi}, \bar{x}) \leq \min_{\{x \geq 0 \mid Cx \geq d\}} \eta(\bar{\pi}, x)$, which proves the right-hand inequality in Equation (8). Now, from condition (iii), $\eta(\bar{\pi}, \bar{x}) = c^T \bar{x}$, and $A\bar{x} \geq b$, which implies

$$\begin{aligned} \eta(\pi, \bar{x}) &= c^T \bar{x} + \pi^T(b - A\bar{x}) \leq c^T \bar{x} \\ &= \eta(\bar{\pi}, \bar{x}), \quad \forall \pi \geq 0, \end{aligned}$$

thereby proving that $(\bar{\pi}, \bar{x})$ is a saddle point.

Conversely, let $(\bar{\pi}, \bar{x})$ be a saddle point for $\eta(\pi, x) = c^T x + \pi^T(b - Ax)$. Then, by definition, $\eta(\bar{\pi}, \bar{x}) \leq \min_{\{x \geq 0 \mid Cx \geq d\}} \eta(\bar{\pi}, x)$, which proves that $\bar{x}$ is a Lagrangian solution to Problem $\text{LR}_{\bar{\pi}}$ at $\pi = \bar{\pi}$. Moreover,

$\max_{\pi \geq 0} \eta(\pi, \bar{x}) \leq \eta(\bar{\pi}, \bar{x})$, which gives $A\bar{x} \geq b$; else, $\max_{\pi \geq 0} \eta(\pi, \bar{x})$ would be unbounded. Finally, for any row index $j \in \{1, \dots, p\}$, where a_j is the j^{th} row of A , if $(a_j)^T \bar{x} > b_j$, then $\bar{\pi}_j = 0$ because $\bar{\pi}$ is a solution to $\max_{\pi \geq 0} \eta(\pi, \bar{x})$, and condition (iii) is satisfied. Hence proved.

Corollary 1. $v[\text{LD}] = v[\text{LR}_{\pi^*}] = v[\text{LR}_{u^*}]$, where π^* is the vector of optimal Lagrange multipliers obtained by solving Problem LD, and u^* is the (partial) optimal dual vector associated with the complicating constraints obtained by solving Problem P.

Corollary 2. The primal–dual pair (u^*, x_{u^*}) , where x_{u^*} is a Lagrangian solution to Problem LR_{π} at $\pi = u^*$, is a saddle point optimum with $v[\text{LR}_{u^*}] = c^T x_{u^*}$.

Corollary 3. Problems P and LD are essentially LP duals of one another.

Here, we have stated several important corollaries that arise from Theorems 1 and 2, but the proofs are fairly straightforward and have been left as an exercise to the reader. For example, Corollary 1 follows as a direct consequence of Theorem 2 and Proposition 3, used in conjunction with each other. Corollaries 1 and 2 also correlate Lagrangian duality to traditional LP duality theory by proving that the optimal Lagrange multipliers obtained by solving Problem LD are indeed the same as the optimal dual variables associated with the complicating constraints, obtained via regular LP duality. Moreover, all of the results presented above not only prove that solving the Lagrangian dual is optimal to the original linear program, but they also present the necessary *optimality conditions* for the Lagrangian dual, and these optimality conditions are used in designing the algorithms presented in the following section.

Next, to showcase the methodology presented above, consider the following illustrative example:

$$\text{P1: Minimize } z \equiv -x_1 - x_2 \quad (9a)$$

$$\text{subject to: } -x_1 + x_2 \geq -1 \quad : u_1, \quad (9b)$$

$$2x_1 - 3x_2 \geq -6 \quad : v_1, \quad (9c)$$

$$\begin{aligned} -4x_1 - 6x_2 &\geq -18 \quad : v_2, (9d) \\ (x_1, x_2) &\geq 0. \end{aligned} \quad (9e)$$

The optimal solution for Problem P1 is $(x_1^*, x_2^*) = (12/5, 7/5)$, and the corresponding dual vector is $(u_1^*, v_1^*, v_2^*) = (1/5, 0, 1/5)$, with optimum objective function value of $v[\text{P1}] = -19/5$. For the purposes of expositing the Lagrangian approach, we let constraint (9b) serve as the complicating constraint, and dualize this constraint by using the nonnegative scalar $\pi \geq 0$ to obtain the following Lagrangian relaxation of P1:

$$\text{LR}_\pi: \text{Minimize} \{-x_1 - x_2 + \pi(-1 + x_1 - x_2) \mid (9c), (9d), \text{ and } (9e)\}. \quad (10)$$

Rewriting the objective function in Equation (10) as $-\pi + (\pi - 1)x_1 - (\pi + 1)x_2$, and with some algebraic analysis, a complete, explicit characterization of the Lagrangian relaxation function is obtained, given by

$$v[\text{LR}_\pi] = \begin{cases} \frac{7\pi}{2} - \frac{9}{2}, & 0 \leq \pi \leq \frac{1}{5} \\ -\frac{11\pi}{4} - \frac{13}{4}, & \frac{1}{5} \leq \pi \leq 5 \\ -3\pi - 2, & \pi \geq 5 \end{cases}. \quad (11)$$

Figure 1 displays $v[\text{LR}_\pi]$ as a function of π (for brevity, shown only for $0 \leq \pi \leq 1$), and the corresponding Lagrangian dual problem is LD: Maximize $\{(11) \mid \pi \geq 0\}$. As seen in the figure, the maximum of this Lagrangian relaxation function and, consequently, the optimum for Problem LD (and Problem P1)

occurs at $\pi^* = 1/5$, with the optimum objective function value of $v[\text{LD}] \equiv v[\text{P1}] = -19/5$. Also, note that the optimal value of the Lagrange multiplier obtained here is indeed the same as the optimal value of the dual variable associated with the complicating constraint (9b); that is, $\pi^* \equiv u_1^* = 1/5$. However, the optimal values of the dual variables corresponding to the remaining constraints as well as the primal solution cannot be obtained via the solution of the Lagrangian dual problem in this case.

So, how can the primal solution vector be recovered from the solution to the Lagrangian dual problem? In order to obtain the optimal values of the decision variables in the space of the original problem, we use the mathematical characterization of a Lagrangian solution provided in Theorem 1 in concert with primal feasibility conditions. From Theorem 1, we know that

$$\pi^*(-1 + x_1 - x_2) = 0 \Rightarrow x_1 = 1 + x_2. \quad (12)$$

Moreover, using Equation (12) along with primal feasibility of constraints (9c) and (9d) yields

$$\left. \begin{aligned} 2x_1 - 3x_2 &\geq -6 \Rightarrow x_2 \leq 8 \\ -4x_1 - 6x_2 &\geq -18 \Rightarrow x_2 \leq 7/5 \end{aligned} \right\} \Rightarrow x_2 \leq 7/5. \quad (13)$$

The objective function of P1 is $z = -x_1 - x_2 \equiv -1 - 2x_2$, and hence, utilizing the primal feasibility bound for x_2 from Equation (13),

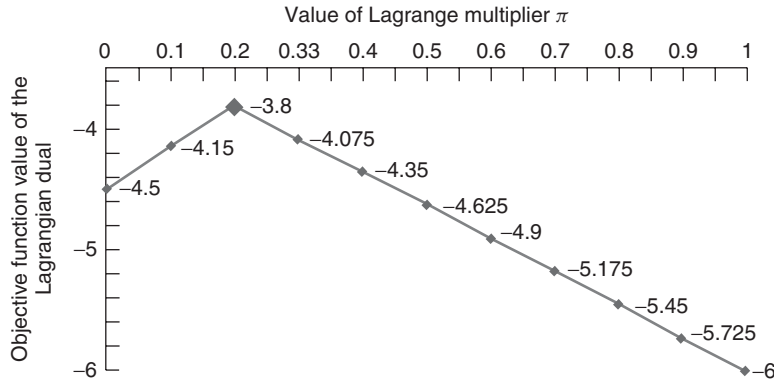


Figure 1. Optimal value of the Lagrangian relaxation problem as a function of the Lagrange multiplier.

we get that the optimal (minimum) objective function value occurs at $x_2^* = 7/5$, and correspondingly, $x_1^* = 1 + x_2^* = 12/5$, which is indeed the primal optimal solution. Also, notice that as the bound $x_2^* = 7/5$ is obtained from Equation (9d), this constraint is tight at optimality.

Remark 3. As seen in Fig. 1, the Lagrangian relaxation function for a minimization problem is a *continuous, piecewise-linear, concave function* (for details, refer to the section titled “Solution Methodologies For The Lagrangian Dual”). Indeed, each of the linear functions in this piecewise-linear concave function can be shown to correspond to an *extreme point* of the feasible region of Problem LR_π ; that is, extreme points of $\{Cx \geq d, x \geq 0\}$. (Alternatively, one can say that each extreme point of the feasible region of LR_π *generates* a unique linear function.) In the example above, the extreme points of Equation (10) are given by $\{(0, 0), (9/2, 0), (3/4, 5/2), (0, 2)\}$, and, for example, the linear function $v[LR_\pi] = \frac{7\pi}{2} - \frac{9}{2}$ corresponds to the extreme point $(9/2, 0)$, that is, for values of $0 \leq \pi \leq 1/5$, the optimum to Problem LR_π occurs at this extreme point. Similarly, the other two linear functions in Equation (10) can be shown to correspond to the extreme points $(3/4, 5/2)$ and $(0, 2)$, respectively. Moreover, based on continuity of the Lagrangian relaxation function, at the optimum value of $\pi^* = 1/5$, the Lagrangian solutions $(9/2, 0)$ and $(3/4, 5/2)$ are both optimal to Problem LR_π . Furthermore, the extreme point $(0, 0)$ is not an optimal solution of Problem LR_π for any value of $\pi \geq 0$, and is therefore not associated with any linear function.

Remark 3 brings to the fore a very important aspect of this Lagrangian approach. As the Lagrangian relaxation function is a continuous, piecewise-linear, concave function, the Lagrangian dual problem which maximizes this function is a “nondifferentiable” optimization problem, and so, the solution to this problem cannot be achieved via classical optimization methods, which mostly rely on differentiability of functions for computing gradients and Hessian matrices. Hence,

a new set of optimization algorithms need to be employed for solving this challenging class of problems.

SOLUTION METHODOLOGIES FOR THE LAGRANGIAN DUAL

Optimization problems having convex, but nondifferentiable objective functions are called *nondifferentiable optimization* (NDO) problems [8]. Within large-scale optimization frameworks, such NDO problems arise in algorithms that cater toward solving min–max problems, dynamic programming, piecewise approximations, and stochastic optimization. One of the important application areas for NDO problems is in Lagrangian duality, and the most prominently featured NDO-based solution methodology for solving the Lagrangian dual problem is the *subgradient optimization* algorithm [8,9]. Other methods that are also commonly used are cutting planes [10], column generation [11,12], and of more recent flavor are bundle methods [13] and volume algorithms [14–16].

Here, we begin by considering the basic version of a subgradient optimization approach for solving Problem LD. To motivate this discussion, note that $v[LR_\pi]$ is an *implicit* function of π , and as seen in the proof of Theorem 2,

$$\begin{aligned} v[LR_\pi] &= \text{minimize } \left\{ c^T x + \pi^T (b - Ax) \right. \\ &\quad \left. | Cx \geq d, x \geq 0 \right\} \\ &= \text{minimize } \left\{ c^T x^i + \pi^T (b - Ax^i) \right\}, \\ &\quad i=1, \dots, p \end{aligned} \tag{14}$$

and so, $v[LR_\pi]$ is the *lower envelope* of a family of affine functions, and is therefore a *concave function* of π . The points where the function is not differentiable are called *breakpoints*, and at these points, the optimal solution of Problem LR_π is not unique. (In the example presented in the section titled “Properties of the Lagrangian Dual”, breakpoints occur at $\pi = 1/5$ and 5 , and at these points, $v[LR_\pi]$ is not only nondifferentiable but also has multiple optimal solutions. At $\pi = 1/5$,

both extreme points (9/2,0) and (3/4,5/2), as well as all points belonging to the facet (9c) are optimal to Problem LR_π).

A piecewise-linear concave function is differentiable almost everywhere except over the set of breakpoints and at these points, as no gradient information is available, *subgradients* prove to be a very useful substitute.

Definition 3. If $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ is a concave function, then for any $x^0 \in \mathbb{R}^n$, $\gamma(x^0)$ is a *subgradient* of $f(x)$ at x^0 if and only if

$$f(x) \leq f(x^0) + \left(\gamma(x^0) \right)^T (x - x^0), \quad \forall x \in \mathbb{R}^n.$$

Definition 4. The set of all subgradients at a point $x^0 \in \mathbb{R}^n$ is called the *subdifferential* of $f(x)$ at x^0 , and is denoted as $\partial f(x^0)$.

Proposition 4. If $\bar{\pi} \geq 0$ is a Lagrange multiplier with corresponding Lagrangian solution $x_{\bar{\pi}}$, then $\gamma \equiv (b - Ax_{\bar{\pi}})$ is a subgradient of $v[LR_{\pi}]$ at $\bar{\pi}$.

Subgradient Algorithm

- Step 1:** Set iteration counter $k = 0$, choose initial Lagrange multipliers π^k , and let η^* be the current guess for the optimal value of Problem LD. Go to Step 2.
- Step 2:** Solve Problem LR_π with $\pi = \pi^k$, to obtain the Lagrangian solution x_{π}^k , with objective function value $\eta_k = v[LR_{\pi^k}]$. Go to Step 3.
- Step 3:** Compute the subgradient of $v[LR_{\pi}]$ at x_{π}^k as $\gamma^k = (b - Ax_{\pi}^k)$. Go to Step 4.
- Step 4:** Let $\pi^{k+1} = \max\{0, \pi^k + \mu_k \gamma^k\}$, where $\mu_k = \frac{\eta^* - \eta_k}{\|\gamma^k\|^2}$ is the *step length* in the direction of the subgradient. If $\|\pi^{k+1} - \pi^k\| \leq \varepsilon$, for some optimality tolerance $\varepsilon > 0$, stop; return π^k as the optimal solution to Problem LD with objective function value $v[LR_{\pi^k}]$. Else, go to Step 5.
- Step 5:** Set $k = k + 1$, and go to Step 2.

Remark 4. In computing the step lengths, the subgradient algorithm uses the unknown optimal value η^* (or *target value*, which lends its name to “target value methods” [15,17]), and a common stumbling block with practical implementations of subgradient algorithms is to overestimate the value of η^* . If the objective function value fails to change for many iterations, then it is likely that η^* has been overestimated, and the corresponding step

Proof. For notational ease, let $\eta_{\pi} = v[LR_{\pi}]$. As $x_{\bar{\pi}}$ is a Lagrangian solution, from Definition 2,

$$\begin{aligned} \eta_{\pi} &\leq c^T x_{\bar{\pi}} + \pi^T (b - Ax_{\bar{\pi}}) \\ &= c^T x_{\bar{\pi}} + \bar{\pi}^T (b - Ax_{\bar{\pi}}) + (\pi - \bar{\pi})^T (b - Ax_{\bar{\pi}}) \\ &= \eta_{\bar{\pi}} + (\pi - \bar{\pi})^T (b - Ax_{\bar{\pi}}). \end{aligned}$$

Setting $\gamma \equiv (b - Ax_{\bar{\pi}})$ completes the proof.

Now, let π^* be the (unknown) optimal Lagrange multipliers with $\eta^* = v[LR_{\pi^*}]$. The subgradient optimization algorithm is an *iterative procedure*, where, at some iteration k , given the current values of the Lagrange multipliers π^k , a step is taken in the direction of the subgradient $\gamma^k = (b - Ax_{\pi}^k)$, using an appropriately determined step length, to determine new values of the multipliers and the process continues until convergence is achieved.

A formal statement of the subgradient algorithm for solving the Lagrangian dual problem is given below.

lengths are too large. Hence, as a “corrector” mechanism, a modified step length is used, given by

$$\pi^{k+1} = \max \left\{ 0, \pi^k + \frac{\xi(\eta^* - \eta_k)}{\|\gamma^k\|^2} \gamma^k \right\}, \quad (15)$$

where, $\xi \in (0, 2)$, and this scalar is reduced whenever there is no improvement for a significant number of iterations.

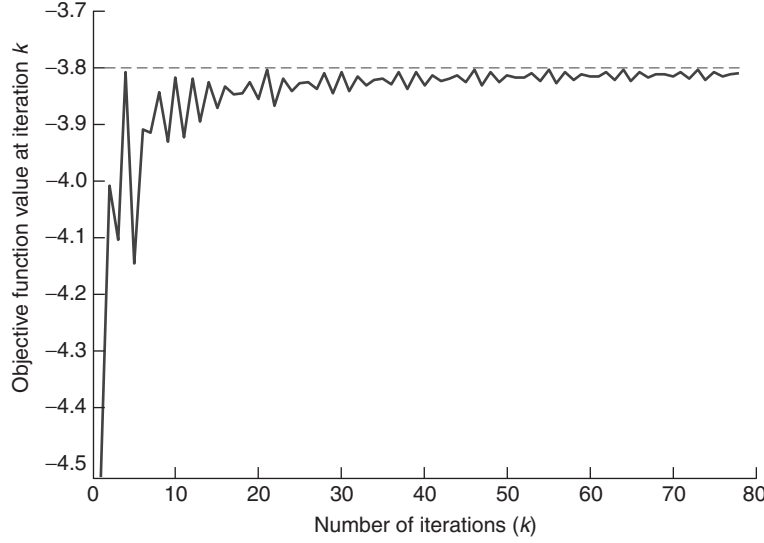


Figure 2. Convergence of the subgradient algorithm as a function of the number of iterations for solving Problem P1.

Given the dual multipliers in Equation (15), it can be shown that the sequence of iterates $\|\pi^k - \pi^*\|$ is monotone nonincreasing and therefore the algorithm produces a convergent sequence of dual multipliers. However, an important theoretical consideration is to determine the values of the step lengths, μ_k , such that not only the dual multipliers converge but also the objective function value converges to the *optimal* solution with a reasonable rate of convergence. A fundamental result in Ref. 9 states that the sequence of iterates $\{\eta_k\}$ converges to η^* if and only if $\{\mu_k\} \rightarrow 0$ and $\sum_{i=0}^{\infty} \mu_k \rightarrow \infty$. The condition $\{\mu_k\} \rightarrow 0$ is necessary to ensure that the step lengths do not continually “overshoot” the optimum, while $\sum_{i=0}^{\infty} \mu_k \rightarrow \infty$ guarantees that the step lengths are large enough to get to π^* from the original estimate of π^0 . As is evident from the foregoing discussion, getting a subgradient algorithm to work in practice requires tweaking the parameters repeatedly to ensure convergence and to make matters more complicated, these parameters are often problem-specific as well. For a detailed proof of convergence, see Refs 9 and 18.

The subgradient algorithm described above was implemented in MATLAB on a computer running Windows XP with

an Intel Core 2 Duo, 2 GHz, 2 MB RAM processor. As seen in the results, beginning with $\pi^0 = 0$, and an initial estimate of the optimum $\eta^* = -5$, the algorithm converged to an ε -optimum in 77 iterations, in 0.594 CPU seconds, with an optimality tolerance of $\varepsilon = 0.0005$. Figure 2 plots the objective function value produced by the subgradient algorithm at iteration k , and as clearly seen in the figure, in the initial stages, the algorithm moves rapidly toward the optimum, and subsequently follows a zigzag pattern before achieving the required tolerance level for convergence. Table 1 records several parameters related to the performance of the subgradient algorithm, applied to Problem P1, for a few select iterations. Notice that the values of the step lengths converge to zero, thereby satisfying the convergence criterion, and the value of the dual multiplier converges to $\pi^* = 1/5$. Also, note that the values of (x_1^k, x_2^k) displayed in the table are the optimal values for the Lagrangian relaxation problem (in the space of the “easy constraints”), and are not optimal to the original primal problem.

The subgradient algorithm outlined above works only in the dual space of the Lagrangian dual problem, and therefore the

Table 1. Results of the Subgradient Algorithm for Solving Problem P1 at Select Iterations

Iteration #	μ_k	π^k	γ^k	x_1^k	x_2^k	η^k
0	0.0408	0	-3.50	4.50	0.00	-4.5000
1	0.0466	0.1429	-3.50	4.50	0.00	-4.0000
2	0.0379	0.3061	2.75	0.75	2.50	-4.0918
3	0.0352	0.2020	2.75	0.75	2.50	-3.8054
4	0.0188	0.1051	-3.50	4.50	0.00	-4.1323
45	0.0031	0.1994	-3.50	4.50	0.00	-3.8021
50	0.0028	0.1959	-3.50	4.50	0.00	-3.8145
76	0.0019	0.1978	2.75	0.75	2.50	-3.8081
77	0.0018	0.2042	-3.50	4.50	0.00	-3.8077

primal solution to Problem P is not recovered by this algorithm. Moreover, as seen in Fig. 2, taking step lengths along the subgradient direction also results in a zigzagging phenomenon that causes the algorithm to slowly crawl toward the optimum. To tackle these shortcomings, a more refined class of subgradient algorithms, namely *primal-dual subgradient algorithms*, has been developed for solving NDO problems. In such methods, a sequence of primal and dual iterates is generated by employing a Lagrangian dual function and a penalty function that satisfies some properties. Consequently, such methods admit both primal and dual optimal solutions, and convergence follows based on narrowing the gap between the penalty function and the Lagrangian dual.

For more details on such primal-dual subgradient methods as well more advanced *conjugate gradient* schemes (that allow for modified gradient directions and improved convergence properties), we refer the reader to Refs 19 and 20. Apart from gradient-based schemes, many other approaches have also been developed for solving the Lagrangian dual problem; for details, see Ref. 7.

EXTENSIONS AND CONCLUSIONS

As discussed earlier (see Remark 3), recognizing that each of the extreme points of the feasible region of LR_π generates a unique linear function, the linkage between the Lagrangian approach and other well-known large-scale LP methods, namely Dantzig-Wolfe decomposition [11,12] and

Benders' decomposition [21] can be further analyzed. Indeed, the astute reader would have already encountered many of the standard elements of decomposition methods in this article, particularly in the proofs of Theorem 2 and Proposition 3, as well as in the subsequent discussion related to continuous, piecewise-linear, concave functions. However, the other articles of this encyclopedia showcase decomposition techniques (see **Dantzig-Wolfe Decomposition** and **Benders Decomposition**) and in the following chapter a detailed discussion relating the Lagrangian dual to these decomposition methods is presented (see **Relationship Among Benders, Dantzig-Wolfe, and Lagrangian Optimization**). Hence, for the remainder of this article, we will proceed in a different direction and focus on a very brief extension of the Lagrangian dual approach geared toward discrete optimization problems.

In so far, while we have focused on using the Lagrangian dual approach for solving LP problems, many of these results and discussions can be extended to the IP case. Indeed, the true strength of the Lagrangian dual approach is realized when solving problems with discrete variables as the bounds provided by the Lagrangian dual are often much tighter than the bounds obtained from the basic LP relaxation. Consequently, there has been extensive research geared toward applying Lagrangian dual schemes to IP problems [1,5-7,22,23].

Now, suppose that in the primal problem, the decision variables are restricted to be integer, that is, IP: Minimize

$\{c^T x | Ax \geq b, Cx \geq d, x \in \mathbb{Z}_+^n\}$. Denoting $\Omega \equiv \{Cx \geq d, x \in \mathbb{Z}_+^n\}$, the Lagrangian relaxation problem for IP relative to the constraints $Ax \geq b$ and a conformable vector $\pi \geq 0$ is given by LR_π : Minimize $\{c^T x + \pi^T(b - Ax) | x \in \Omega\}$, and the corresponding Lagrangian dual problem is LD: Maximize $\{LR_\pi | \pi \geq 0\}$. Akin to the development in the LP case, there are several mathematical properties of the Lagrangian dual formulation that can be proven for the IP case, but the most fundamental result is that the objective function value of the Lagrangian dual is at least as tight as the one obtained from the standard LP relaxation, and furthermore, if the constraint set $\Omega \equiv \{Cx \geq d, x \in \mathbb{Z}_+^n\}$ satisfies the so-called *integrality property*, then the Lagrangian dual produces a bound that is the same as the LP relaxation bound. These two properties are summarized in the theorems below; for proofs, see Refs 1,4,5,7, and 23.

Theorem 3. $v[LP] \leq v[LD] \leq v[IP]$, where LP is the linear programming relaxation of IP.

Theorem 4. If the Lagrangian relaxation problem, LR_π , satisfies the integrality property, that is, $\text{conv}(\Omega) \equiv \{Cx \geq d, x \geq 0\}$, then $v[LP] \equiv v[LD] \leq v[IP]$, where $\text{conv}(\cdot)$ represents the convex hull of the constraints $(\cdot)$.

To gauge the performance of the Lagrangian dual formulation in the case of

integer programs, suppose that we impose integer restrictions on the x variables in Problem P1. In this case, the optimum occurs at $(x_1^*, x_2^*) = (2, 1)$, with optimal objective function value of $v[P1] = -3$. However, solving the Lagrangian dual yields $\pi^* = 1/3$ with optimal objective function value of $v[LD] = -11/3$, with the corresponding primal solution $(x_1^*, x_2^*) = (3, 1)$. The solid and dashed lines in Fig. 3 indicate the performance of the Lagrangian relaxation function of IP and its LP relaxation, respectively, for different values of $\pi \geq 0$ (as before, displayed only for $0 \leq \pi \leq 1$). In this case, as the conditions of Theorem 4 are not satisfied; that is, integrality property does not hold, the Lagrangian dual function yields a tighter bound than the standard LP relaxation. The “duality gap” in this case is $v[P1] - v[LD] = 2/3$. Notice that the duality gap between IP and its corresponding LP relaxation is $v[P1] - v[LP] = 4/5$; hence, the Lagrangian dual formulation provides a tighter bound as compared to the LP relaxation, and this fact can be further verified from Fig. 3.

In this article, we developed the basic methodology for solving large-scale LP problems via a Lagrangian dual approach. Beginning with a formulation of the Lagrangian dual problem, we proved that the optimal objective function value obtained by solving the Lagrangian dual is optimal to the original linear program, and moreover, theorems and propositions that characterized the mathematical properties of the

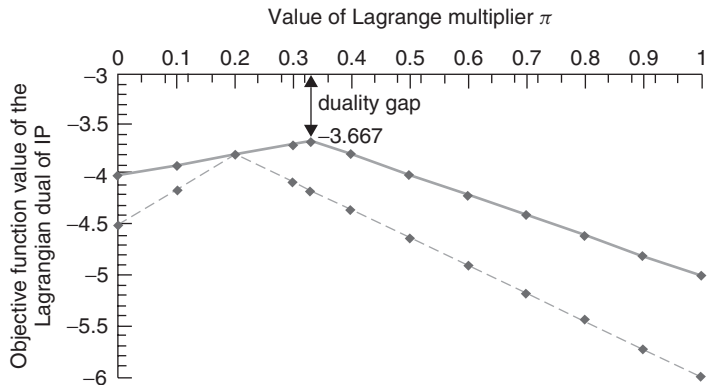


Figure 3. Optimal value of the Lagrangian relaxation for IP as a function of the Lagrange multiplier.

Lagrangian dual to relate this approach to classical LP duality were also included. Recognizing that the Lagrangian dual formulation is a NDO problem, a subgradient algorithm was prescribed for solving this problem, and computational results for an illustrative example were presented. A brief extension of this approach for solving integer programs can also be found in this article.

Alongside a methodological framework, there are several other questions related to the Lagrangian approach that also need to be addressed, primarily the ones arising from a *modeling perspective*. For example, questions such as—what constraints should be dualized in a Lagrangian scheme; does the original problem formulation affect the strength of the Lagrangian dual problem bound; can the Lagrangian relaxation bounds be tightened by incorporating cutting planes—are critical to the success of the Lagrangian optimization approach and need to be answered *prior* to setting up the Lagrangian dual and selecting an appropriate algorithm. For such perspectives, we refer the reader to Ref. 7. In conclusion, the Lagrangian dual approach, when applied judiciously by selecting the appropriate complicating constraints along with a convergent algorithm, can prove to be a very useful tool for solving large-scale linear and IP problems.

As an aside, note that the term *Lagrangian* has also been often used in the optimization literature. An internet search for the terms Lagrangian dual and Lagrangean dual yielded a comparable number of references, and therefore we conclude that both terms have been used interchangeably by the optimization community and represent the same idea presented in this article.

REFERENCES

1. Fisher ML. The Lagrangian relaxation method for solving integer programming problems. *Manage Sci* 1981;27:1–18.
2. Adams WP, Sherali HD. Mixed-integer bilinear programming problems. *Math Program* 1993;59:279–305.
3. Everett H. Generalized Lagrange multiplier method for solving optimum allocation of resources. *Oper Res* 1963;11:399–417.
4. Geoffrion AM. Elements of large-scale mathematical programming. *Manage Sci* 1970;16:652–691.
5. Geoffrion AM. Lagrangian relaxation for integer programming. *Math Program Study* 1974;2:82–114.
6. Geoffrion AM, Marsten R. Integer programming: a framework and state-of-the-art survey. *Manage Sci* 1972;18:465–491.
7. Guignard M. Lagrangean relaxation. *Top* 2003;11(2):151–228.
8. Shor NZ. Minimization methods for nondifferentiable problems. Berlin: Springer; 1985. pp. 116–120.
9. Held M, Wolfe P, Crowder HP. Validation of subgradient optimization. *Math Program* 1974;6:62–88.
10. Kelly JE. The cutting plane method for solving convex programs. *J SIAM* 1960;8:703–712.
11. Dantzig GB, Wolfe P. The decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
12. Dantzig GB, Wolfe P. The decomposition algorithm for linear programs. *Econometrica* 1961;29:767–778.
13. Lemaréchal C, Zowe J. A condensed introduction to bundle methods in nonsmooth optimization. In: Spedicato E, editor. *Algorithms for continuous optimization*. The Netherlands: Kluwer Academic Publishers; pp. 1993;357–382.
14. Barahona F, Anbil R. The volume algorithm: producing primal solutions with a subgradient method. *Math Program* 2000;87:385–399.
15. Lim C, Sherali HD. A trust region target value method for optimizing nondifferentiable Lagrangian duals of linear programs. *Math Methods Oper Res* 2006;34:33–63.
16. Sherali HD, Lim C. Enhancing Lagrangian dual optimization for linear programs by obviating nondifferentiability. *INFORMS J Comput* 2007;19(1):3–13.
17. Sherali HD, Choi G, Tuncbilek CH. A variable target value method for nondifferentiable optimization. *Oper Res Lett* 2000;1:1–8.
18. Kim S, Ahn H. Convergence of a generalized subgradient method for nondifferentiable convex optimization. *Math Program* 1991;50:75–80.
19. Sherali HD, Choi G. Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs. *Oper Res Lett* 1996;19(3):105–113.

20. Camerini PM, Fratta L, Maffoli F. On improving relaxation bounds by modified gradient techniques. *Math Program* 1975;3:26–34.
21. Benders J. Partitioning procedures for solving mixed-variables programming problems. *Numer Math* 1962;4:238–252.
22. Held M, Karp RM. The traveling salesman problem and minimum spanning trees: part II. *Math Program* 1971;1:6–25.
23. Martin RK. Large scale linear and integer optimization: a unified approach. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999.

LAGRANGIAN OPTIMIZATION METHODS FOR NONLINEAR PROGRAMMING

ANGELIA NEDIĆ

Department of Industrial and Enterprise
Systems Engineering, University of Illinois
at Urbana-Champaign, Urbana, Illinois

LAGRANGIAN RELAXATION AND ITS DUALITY THEORY

In a classical constrained optimization, an objective function has to be minimized (or maximized) in the presence of finitely many constraints. Some of the constraints may pose difficulties for solving the problem efficiently, while the removal of these “difficult” constraints results in an optimization problem that can be solved efficiently.

Lagrangian relaxation is one of the basic tools for dealing with difficult constraints in nonlinear constrained optimization problems. Lagrangian relaxation is a technique that “relaxes” the difficult constraints by penalizing (pricing) the constraints and bringing them into the objective. This gives rise to an alternative constrained problem, known as the *dual of the original problem*. While the dual problem bears strong relations with the original problem, it is not necessarily equivalent to the original problem. The duality theory, also known as *Lagrangian duality*, provides conditions under which the dual and the original problem are equivalent.

Lagrangian relaxation and its duality theory have been widely used for the design and analysis of optimization problems arising in vast range of engineering and other applied sciences. Aside from enabling us with framework for dealing with “difficult” constraints, Lagrangian relaxation often provides us with suitable decomposition schemes. Such schemes have been particularly useful when dealing with problems of large size (large

number of variables and/or constraints). More recently, with the advances in wired and wireless communication technology, Lagrangian relaxation is used in the design of algorithms for performance optimization of large systems with distributed information such as sensor networks, Internet, and other communication systems [1–3].

Primal Problem and Its Dual

The power of Lagrangian relaxation lies within constrained optimization problems. A general constrained optimization problem has the following form:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_j(x) \leq 0 \quad j = 1, \dots, m, \\ & && h_i(x) = 0 \quad i = 1, \dots, r, \\ & && x \in X, \end{aligned} \quad (1)$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $g_j: \mathbb{R}^n \rightarrow \mathbb{R}$ for all j , $h_i: \mathbb{R}^n \rightarrow \mathbb{R}$ for all i , and $X \subseteq \mathbb{R}^n$. In general, the functions f , g_j , and h_i need not be defined on the entire $\mathbb{R}^n$, in which case extended-real value functions are considered [4]. To keep the exposition simple, we will not deal with the extended-real value functions.

The problem in Equation (1) is referred to as *primal problem* and its optimal value is denoted by f^* , that is,

$$f^* = \inf_{\substack{g_j(x) \leq 0, j=1, \dots, m \\ h_i(x) = 0, i=1, \dots, r, x \in X}} f(x).$$

A vector x is said to be *primal feasible* if it is feasible for the primal problem, i.e., $g_j(x) \leq 0$ for all j , $h_i(x) = 0$ for all i and $x \in X$. Often, for a more compact notation, the vector functions $g(x)$ and $h(x)$ are defined, respectively, as $g(x) = (g_1(x), \dots, g_m(x))^T$ and $h(x) = (h_1(x), \dots, h_r(x))^T$ for $x \in \mathbb{R}^n$. Throughout the rest of the discussion, for a vector u , the notation $u \leq 0$ (or $u < 0$) denotes the component-wise inequalities, that is, $u_i \leq 0$ (or $u_i < 0$) for all coordinate indices i . Thus, the relations $g(x) \leq 0$ and $h(x) = 0$ in vector notation correspond, respectively, to $g_j(x) \leq 0$ for all j and $h_i(x) = 0$ for all i .

The basic underlying idea of the Lagrangian relaxation is to associate another problem with primal problem (1) and establish the conditions for the equivalence of the two problems. The problem associated with the primal is obtained by penalizing the constraints $g_j(x) \leq 0$ and $h_i(x) = 0$ and bringing them into the objective of the new problem. This is done by assigning a nonnegative scalar μ_j to each inequality constraint $g_j(x) \leq 0$, and assigning a scalar λ_i (unrestricted in sign) to each constraint $h_i(x) = 0$. Then, the functions $\sum_{j=1}^m \mu_j g_j(x)$ and $\sum_{i=1}^r \lambda_i h_i(x)$ are formed, and the objective for the new problem is defined using $f(x) + \sum_{j=1}^m \mu_j g_j(x) + \sum_{i=1}^r \lambda_i h_i(x)$. The problem associated with Equation (1) is given by

$$\begin{aligned} \text{maximize } q(\mu, \lambda) &\triangleq \inf_{x \in X} \left(f(x) + \sum_{j=1}^m \mu_j g_j(x) \right. \\ &\quad \left. + \sum_{i=1}^r \lambda_i (a_i^T x - b_i) \right) \\ \text{subject to } &\mu \geq 0, \quad \mu \in \mathbb{R}^m \\ &\lambda \in \mathbb{R}^r, \end{aligned} \quad (2)$$

where μ and λ are the vectors with components μ_j and λ_i , respectively. This problem is referred to as the *dual problem* associated with the primal problem in Equation (1). Its optimal value is denoted by q^* , that is,

$$q^* = \sup_{\substack{\mu \geq 0, \mu \in \mathbb{R}^m \\ \lambda \in \mathbb{R}^r}} q(\mu, \lambda).$$

Within the Lagrangian relaxation literature, the dual variables μ and λ are known as *Lagrangian multipliers* or *multipliers*. A pair (μ, λ) is said to be *dual feasible* if $\mu \geq 0$. The dual function $q(\mu, \lambda)$ is always *concave*, regardless of the structure of the primal problem (i.e., even when problem (1) is nonconvex, discontinuous, discrete, etc.) [4,5]. While the dual function $q(\mu, \lambda)$ is always concave in $(\mu, \lambda) \in \mathbb{R}^m \times \mathbb{R}^r$, its domain is not necessarily the whole space $(\mu, \lambda) \in \mathbb{R}^m \times \mathbb{R}^r$ [4,5]. Nevertheless, the dual problem is always a concave maximization problem with simple explicit constraints.

The difficulty in solving the dual problem lies in the “implicit” form of the dual function. Specifically, note that the evaluation of the dual function requires one to solve a constrained convex minimization problem entries b_i .

$$q(\mu, \lambda) = \inf_{x \in X} (f(x) + \mu^T g(x) + \lambda^T h(x)). \quad (3)$$

This problem is referred to as *Lagrangian subproblem*. From this perspective, the use of Lagrangian relaxation is the most successful when the solution to Lagrangian subproblem, $\inf_{x \in X} (f(x) + \mu^T g(x) + \lambda^T h(x))$, is available in a closed form or a good approximation can be computed efficiently.

Luckily, arising in various applications, there are many problems for which a closed form solution to Lagrangian subproblem exists [6]. For such problems, the dual problem is often easier to solve than the primal problem. However, to have the guarantee that solving the dual is equivalent to solving the primal, we have to ensure that certain duality relations are in place. These relations are discussed in the following section.

Weak and Strong Duality

At the heart of the duality lies a relation providing a fundamental connection between the optimal values of primal problem (1) and its associated dual problem (2). This relation states that the dual function value $q(\mu, \lambda)$ provides a lower bound for the primal optimal value f^* , that is,

$$q(\mu, \lambda) \leq f^* \quad \text{for all } \mu \in \mathbb{R}^m, \mu \geq 0 \text{ and } \lambda \in \mathbb{R}^r. \quad (4)$$

This relation is often referred to as *lower bounding property* [6]. The relation is implied by the inclusion $\{x \in X \mid g(x) \leq 0, h(x) = 0\} \subseteq X$, as follows:

$$\begin{aligned} q(\mu, \lambda) &= \inf_{x \in X} (f(x) + \mu^T g(x) + \lambda^T h(x)) \\ &\leq \inf_{\substack{x \in X \\ g(x) \leq 0, h(x) = 0}} (f(x) + \mu^T g(x) + \lambda^T h(x)). \end{aligned}$$

Noting that $\mu^T g(x) \leq 0$ and $\lambda^T h(x) = 0$ for any $\mu \geq 0$ and for any $x \in X$ such that $g(x) \leq 0$

and $h(x) = 0$, we have

$$q(\mu, \lambda) \leq \inf_{\substack{x \in X \\ g(x) \leq 0, h(x) = 0}} f(x) = f^*.$$

This bounding property of dual function value $q(\mu, \lambda)$ has been extensively used in the algorithms for integer and mixed integer programming, such as branch-and-bound [4].

The lower bounding property is the key relation that leads to the duality results, known as *weak and strong duality*. Weak duality states that the dual optimal value q^* is never greater than the primal optimal value f^* .

Theorem 1. *Weak duality always holds, that is,*

$$q^* \leq f^*.$$

The result is a direct consequence of the lower bounding property in Equation (4), from which it follows by taking the supremum over all $\mu \in \mathbb{R}^m$ with $\mu \geq 0$ and all $\lambda \in \mathbb{R}^r$. The power of the weak duality relation is in its generality: no assumptions on the structure of the primal problem are needed.

A particularly important case of the weak duality is the case when the primal and the dual optimal values are equal: $q^* = f^*$. This relation is known as the *strong duality* relation, which provides a ground for the equivalence between the primal problem and its dual.

The strong duality does not hold in general. Most of the known strong duality results reside within *convex* constrained optimization problems. A convex constrained optimization problem is a special case of the problem in Equation (1), defined as follows:

Definition 1. The problem in Equation (1) is convex if

- (a) the objective function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex;
- (b) each function $g_j: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex;
- (c) each function $h_i: \mathbb{R}^n \rightarrow \mathbb{R}$ is affine, that is, for each i , we have $h_i(x) = a_i^T x - b_i$ with $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$;
- (d) the set $X \subseteq \mathbb{R}^n$ is closed and convex.

In general, the functions f and g_j need not be convex on the entire $\mathbb{R}^n$, but rather be closed and convex in extended-real value sense [5,7,8].

Strong duality does not hold for any convex problem. In particular, in order to have $q^* = f^*$ for a convex primal problem, some additional conditions are needed. Such conditions are often referred to as *constraint qualifications* [4,5]. The two most widely used conditions are *Slater condition* and *linear constraints*. Slater condition requires that a point $\bar{x} \in X$ exists such that $g_j(\bar{x}) < 0$ for all $j = 1, \dots, m$. Linear condition [4,5,7] basically requires that $X = \mathbb{R}^n$ and all inequality constraints $g_j(x) \leq 0$ are “affine,” that is, $g_j(x) = c_j^T x - d_j$ with a vector $c_j \in \mathbb{R}^n$ and a scalar d_j for each j . Other conditions exist that are reliant on the recession cones of closed convex sets [5,7,9]. The Slater and linear constraint qualifications actually guarantee more than the strong duality relation. These conditions also guarantee the existence of an optimal dual solution, that is, the existence of a multiplier pair (μ^*, λ^*) with $\mu^* \geq 0$ and such that $q(\mu^*, \lambda^*) = q^*$.

Under the Slater condition, we have the following strong duality result.

Theorem 2 [Slater; 5]. *Let the primal problem be convex and let its optimal value f^* be finite, that is, $f^* > -\infty$. Assume also that the Slater condition holds. Then, we have the following:*

- (a) *there is no duality gap, that is, $q^* = f^*$;*
- (b) *the set of dual optimal solutions is nonempty and bounded.*

The results of Theorem 2 hold, except for the boundedness of the optimal dual set, when the Slater condition is replaced with the linear constraint qualification [4,5].

Lagrangian Saddle Point and KKT conditions

Given a primal problem (1), its Lagrangian function is the function involved in the definition of the dual problem, that is, its Lagrangian function is

$$\mathcal{L}(x, \mu, \lambda) = f(x) + \mu^T g(x) + \lambda^T h(x) \quad \text{for} \\ x \in X, \mu \in \mathbb{R}_+^m, \text{ and } \lambda \in \mathbb{R}^r,$$

where $\mathbb{R}_+^m$ is the nonnegative orthant in $\mathbb{R}^m$. Thus, we have

$$q^* = \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \inf_{x \in X} \mathcal{L}(x, \mu, \lambda). \quad (5)$$

Furthermore, we have for any $x \in \mathbb{R}^n$,

$$\sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \mathcal{L}(x, \mu, \lambda) = \begin{cases} f(x) & \text{if } g(x) \leq 0 \text{ and} \\ & h(x) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, it follows

$$\inf_{x \in X} \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \mathcal{L}(x, \mu, \lambda) = \inf_{\substack{x \in X \\ g(x) \leq 0, h(x) = 0}} f(x) = f^*.$$

From the preceding relation and the relation in Equation (5), we see that the weak duality is equivalent to the following inequality for the Lagrangian function:

$$\begin{aligned} & \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \inf_{x \in X} \mathcal{L}(x, \mu, \lambda) \\ & \leq \inf_{x \in X} \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \mathcal{L}(x, \mu, \lambda). \end{aligned}$$

The strong duality corresponds to the case when the preceding inequality holds with equality, that is,

$$\begin{aligned} & \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \inf_{x \in X} \mathcal{L}(x, \mu, \lambda) \\ & = \inf_{x \in X} \sup_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \mathcal{L}(x, \mu, \lambda). \end{aligned}$$

There is a deeper relation between strong duality conditions, and primal and dual optimal solutions of a convex primal problem. To describe this, we need the notion of a saddle point.

Definition 2. A tuple (x^*, μ^*, λ^*) such that $x^* \in X$ and $\mu^* \geq 0$ is a *saddle point of the Lagrangian function* if the following relation holds:

$$\mathcal{L}(x^*, \mu, \lambda) \leq \mathcal{L}(x^*, \mu^*, \lambda^*) \leq \mathcal{L}(x, \mu^*, \lambda^*)$$

for all $x \in X$, $\mu \in \mathbb{R}_+^m$, and $\lambda \in \mathbb{R}^r$.

In other words, (x^*, μ^*, λ^*) is a saddle point of $\mathcal{L}(x, \mu, \lambda)$ if

$$\begin{aligned} & \max_{\mu \in \mathbb{R}_+^m, \lambda \in \mathbb{R}^r} \mathcal{L}(x^*, \mu, \lambda) = \mathcal{L}(x^*, \mu^*, \lambda^*) \\ & = \min_{x \in X} \mathcal{L}(x, \mu^*, \lambda^*) \end{aligned} \quad (6)$$

Note that Lagrangian function is linear in (μ, λ) for every x fixed and, therefore, concave in (μ, λ) . In view of this, the first equality in relation (6) states that (μ^*, λ^*) is a maximizer of the concave function $(\mu, \lambda) \mapsto \mathcal{L}(x^*, \mu, \lambda)$ over the set $\mathbb{R}_+^m \times \mathbb{R}^r$. When the primal problem is convex, its Lagrangian function is convex in x for every fixed $(\mu, \lambda) \in \mathbb{R}_+^m \times \mathbb{R}^r$. Thus, in the convex case, the second equality in relation (6) states that $x^* \in X$ is a minimizer of the convex function $x \mapsto \mathcal{L}(x, \mu^*, \lambda^*)$ over the set X .

The primal/dual optimal solutions of a convex minimization problem can be characterized as follows.

Theorem 3 [Lagrangian Saddle Point; 4,5]. *Let the primal problem be convex. Then, a vector $x^* \in X^*$ is primal optimal and a multiplier pair $(\mu^*, \lambda^*) \in \mathbb{R}_+^m \times \mathbb{R}^r$ is dual optimal if and only if (x^*, μ^*, λ^*) is a saddle point of the Lagrangian function.*

By using the first-order optimality conditions for the “max” problem in relation (6), Theorem (3) can be stated in an equivalent form.

Theorem 4 [4,5]. *Let the primal problem be convex. Then, a vector $x^* \in X^*$ is primal optimal and a multiplier pair $(\mu^*, \lambda^*) \in \mathbb{R}_+^m \times \mathbb{R}^r$ is dual optimal if and only if*

- (a) x^* is primal feasible, that is, $x^* \in X$, $g(x^*) \leq 0$, and $Ax = b$;
- (b) (μ^*, λ^*) is dual feasible, that is, $\mu^* \geq 0$;
- (c) $\mathcal{L}(x^*, \mu^*, \lambda^*) = \min_{x \in X} \mathcal{L}(x, \mu^*, \lambda^*)$;
- (d) complementary slackness holds, that is, $(\mu^*)^T g(x^*) = 0$.

Since we assume that the primal problem is convex, we have $h(x) = Ax - b$, where A is

an $r \times n$ matrix and $b \in \mathbb{R}^r$. When the objective function f and the constraint functions g_j are differentiable, and $X = \mathbb{R}^n$, the condition (c) of Theorem 4 can be equivalently stated as

$$\nabla f(x^*) + \sum_{j=1}^m \mu_j^* \nabla g_j(x^*) + A^T \lambda^* = 0.$$

In this case, Theorem 4 provides necessary and sufficient conditions for primal–dual optimality of tuple (x^*, μ^*, λ^*) . These conditions are commonly known as *Karush–Kuhn–Tucker conditions* (KKT).

ALGORITHMS

The algorithms we discuss here are exclusively for convex (primal) problems. Roughly speaking, there are two types of problems that we often like to solve when dealing with Lagrangian relaxation. One is the problem of determining the saddle points of the Lagrangian function, thus finding both primal and dual optimal solutions. The other problem is the dual problem itself, when we want to determine a dual optimal solution or just the dual optimal value.

Algorithms for Saddle Point Problems

Instead of focusing only on the saddle point of Lagrangian, we consider a generic saddle point problem of a convex–concave function. A generic saddle point problem consists of determining a pair $(x^*, y^*) \in X \times Y$ such that

$$L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*) \quad \text{for all} \\ x \in X \text{ and } y \in Y,$$

where X and Y are closed convex sets in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and $L(x, y) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a convex–concave function (convex in x for every y , and concave in y for every x).

One of the first algorithms proposed for solving such a convex–concave saddle point problem is Arrow–Hurwicz–Uzawa algorithm [10]. This is a gradient-type algorithm that is applicable when the function $L(x, y)$ is differentiable with respect to each of the variables x and y . The algorithm generates a sequence of iterates by taking simultaneous

steps in both variables, as follows: given current iterates x_k and y_k , new iterates are obtained according to the following rule [10]:

$$x_{k+1} = \Pi_X[x_k - \alpha_k \nabla_x L(x_k, y_k)] \\ y_{k+1} = \Pi_Y[y_k - \alpha_k \nabla_y L(x_k, y_k)],$$

where $\alpha_k > 0$ is a stepsize, $\nabla_x L(x_k, y_k)$ is the gradient of L with respect to x evaluated at (x_k, y_k) , and $\nabla_y L(x_k, y_k)$ is the gradient of L with respect to y evaluated at (x_k, y_k) .

There is another popular gradient-based method for saddle point problems when function $L(x, y)$ has Lipschitz continuous gradients $\nabla_x L(x, y)$ and $\nabla_y L(x, y)$. The method is known as *Korpelevich algorithm* [11], also referred to as *extragradient method*.

A convex–concave saddle point problem may also be formulated as a special variational inequality problem, which then may be solved using the algorithms for such problems. For the relation between saddle point problems and variational inequalities, refer to Facchinei and Pang [12] and Konnov [13]. These two references also provide various algorithms for solving variational inequalities such as those arising from saddle point problems. More recently developed algorithms for variational inequalities and saddle point problems can be found in Nemirovski [14], Auslender and Teboulle [15,16], Nesterov [17], and Nedić and Ozdaglar [18]

Algorithms for Dual Problem

The dual problem (2) can be solved provided that the dual function $q(\mu, \lambda)$ can be evaluated (exactly or with a good approximation) at any dual feasible pair (μ, λ) . As discussed earlier, the dual problem (2) is a maximization problem of the concave dual function $q(\mu, \lambda)$ over the convex set $\mathbb{R}_+^m \times \mathbb{R}^r$. Thus, the dual problem is equivalent to convex minimization, and it is often moved to convex minimization framework by setting $F(\cdot) = -q(\cdot)$. Furthermore, the dual function $q(\cdot)$ is typically nondifferentiable [4,5]. For this reason, most of the Lagrangian optimization methods for solving the dual problems have been developed and studied in general form within the

framework of convex nondifferentiable minimization.

In nondifferentiable optimization, a central notion is the notion of a subgradient, which essentially plays the role of a gradient. A definition of a subgradient for a convex function is as follows:

Definition 3 [Subgradient]. A vector $s \in \mathbb{R}^N$ is a subgradient of a convex function $F: \mathbb{R}^N \rightarrow \mathbb{R}$ at the point $\hat{u}$, when the following relation holds:

$$F(\hat{u}) + s^T(u - \hat{u}) \leq F(u) \quad \text{for all } u.$$

The inequality in the subgradient definition basically states that the linear function $\ell(u) = F(\hat{u}) + s^T(u - \hat{u})$, with a fixed $\hat{u}$, underestimates the function $F(u)$ at every point u . Figure 1 gives an illustration of a subgradient of a convex function.

Given a concave function $U(\cdot)$, the function $F(\cdot) = -U(\cdot)$ is convex. Thus, a vector v is a subgradient v of concave $U(u)$ at $u = \hat{u}$ if and only if $s = -v$ is a subgradient of $F(u)$ at $u = \hat{u}$. Formally, a vector v is a subgradient v of concave $U(u)$ at $u = \hat{u}$ if and only if

$$U(\hat{u}) + v^T(u - \hat{u}) \geq U(u) \quad \text{for all } u.$$

In particular, consider the case when the Lagrangian function is given by $\mathcal{L}(x, \mu, \lambda) =$

$f(x) + \mu^T g(x) + \lambda^T Ax$ for $x \in X$. Given a multiplier $(\hat{\mu}, \hat{\lambda})$ such that the dual function value $q(\hat{\mu}, \hat{\lambda})$ is finite, a subgradient of the dual function $q(\mu, \lambda)$ at the point $(\hat{\mu}, \hat{\lambda})$ is given by a vector $(g(\hat{x}), A\hat{x})$, where $\hat{x}$ is a solution of the following problem:

$$\begin{aligned} & \text{minimize } f(x) + \hat{\mu}^T g(x) + \hat{\lambda}^T Ax \\ & \text{subject to } x \in X. \end{aligned}$$

For more details, examples, and a rigorous treatment of subgradients and their properties refer to [4–8, 19, 20].

As mentioned earlier, the methods for solving the dual problems have been studied within the framework of nondifferentiable convex minimization. We adopt this framework for the rest of the discussion.

A subgradient method is one of the simplest algorithms for minimizing a convex nondifferentiable function $F: \mathbb{R}^N \rightarrow \mathbb{R}$ over a convex closed set $\mathcal{X} \subseteq \mathbb{R}^n$. The method takes the following form:

$$u_{k+1} = \Pi_{\mathcal{X}}[u_k - \alpha_k s_k] \quad \text{for all } k \geq 0, \quad (7)$$

where $u_0 \in \mathcal{X}$ is an initial iterate, $\alpha_k > 0$ is a stepsize, and s_k is a subgradient of $F(u)$ at $u = u_k$. The subgradient method of the form (7) has been long studied, and its convergence properties and convergence rate have been

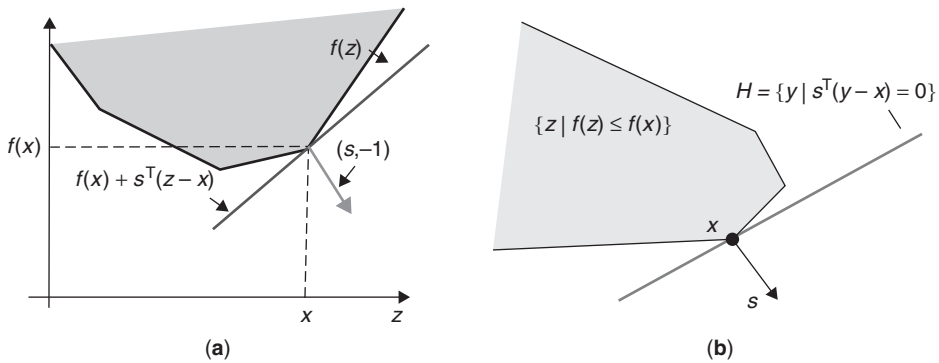


Figure 1. (a) The subgradient property for the epigraph of convex function f : given a subgradient s of f at a point x , the hyperplane in $\mathbb{R}^N \times \mathbb{R}$ with the normal $(s, -1)$ supports the epigraph of a convex function at the point $(x, f(x))$. (b) The subgradient property for the level sets of the function f .

established under various stepsize choices [4,5,8,19,20].

The subgradient method is one of the simplest methods for minimizing a nondifferentiable convex function. Its convergence behavior is very similar to the behavior of gradient methods. In particular, it can be slow just as any gradient algorithm, when the function F has prolonged level sets along some directions. To deal with the slow progress, some modifications of the subgradient method have been considered including those that use space dilation [19].

Unlike gradient methods, the subgradient methods in general *do not have a natural stopping criteria*. Typically, it is run for a pre-specified number of iterations. Also, unlike gradient methods, the subgradient method (7) is *not a descent method*, that is, it does not generate iterates u_k along which the function values $F(u_k)$ necessarily decrease.

There are other subgradient methods that use subgradients to generate descent directions. Some of the descent-based subgradient methods have the same form as in Equation (7), where a descent direction d_k is used in place of a subgradient s_k . The descent direction d_k is obtained by computing several subgradients of F at a given point u_k [4,5, and the references therein].

There is another class of descent-based subgradient algorithms that are more complex. These have been developed starting from the *cutting-plane method* as the basis. The fundamental idea of the cutting-plane method is to generate a sequence of piecewise linear functions F_k , each of which approximates (from below) the given convex function F and the quality of approximation improves as the number k increases. Then, instead of minimizing the function F over the given (closed convex) set $\mathcal{X}$, at each iteration k , the approximate function F_k is minimized over the set $\mathcal{X}$. Formally speaking, at an iteration k , we have the current piecewise linear approximation $F_k(x)$ given by

$$F_k(u) = \max_{0 \leq j \leq k} \left\{ F(u_j) + s_j^T (u - u_j) \right\},$$

where $u_0, u_1, \dots, u_k$ are the iterates generated by the method up to the current time, and s_j is a subgradient of $F(u)$ at $u = u_j$

for all $j = 0, 1, \dots, k$. The next iterate u_{k+1} is determined as a solution to the following minimization problem

$$\begin{aligned} & \text{minimize } F_k(u) \\ & \text{subject to } u \in \mathcal{X}. \end{aligned}$$

The point u_{k+1} is then used to improve the approximate function $F_k(u)$, by computing a subgradient s_{k+1} of $F(u)$ at $u = u_{k+1}$ and, then, defining the approximation $F_{k+1}(u)$ as follows:

$$F_{k+1}(u) = \max \left\{ F_k(u), F(u_{k+1}) + s_{k+1}^T (u - u_{k+1}) \right\}.$$

The new approximate function $F_{k+1}(u)$ is better than the preceding one, $F_k(u)$, in the sense that $F_k(u) \leq F_{k+1}(u)$ for all u . Furthermore, from the subgradient-defining inequality (Definition 3) it can be seen that

$$F_k(u) \leq F(u) \quad \text{for all } u \text{ and } k \geq 0.$$

More discussion on the basic cutting-plane algorithm and its proof can be found in Bertsekas [4,5] (also, see the references therein). The algorithms that build on the cutting-plane method by using analytic center can be found in Kiwiel [21] and Goffin and Vial [22].

More complex descent algorithms that are constructed on the basis of the ideas of the cutting-plane method are known as *bundle methods*. These methods were originally proposed in Kiwiel [23], and further developed in Kiwiel [24], and in Schramm and Zowe [25] where the bundle method is combined with the trust region approach. Bundle methods using a variable metric are discussed in Lemaréchal and Sagastizábal [26], while a quasi-Newton bundle method is given in Mifflin *et al.* [27]. More in-depth coverage of bundle method literature can be found in recent survey papers [22,28].

REFERENCES

1. Johansson B. On distributed optimization in networked systems [PhD thesis]. Stockholm: Royal Institute of Technology; 2008.

2. Palomar DP, Eldar Y. Convex optimization in signal processing and communications. Cambridge: Cambridge University Press; 2010.
3. Srikant R. Mathematics of internet congestion control. Boston (MA): Birkhauser; 2004.
4. Bertsekas DP. Nonlinear programming. Belmont (MA): Athena Scientific; 1999.
5. Bertsekas DP, Nedić A, Ozdaglar AE. Convex analysis and optimization. Belmont (MA): Athena Scientific; 2003.
6. Boyd S, Vandenberghe L. Convex optimization. Cambridge (UK): Cambridge University Press; 2004.
7. Rockafellar RT. Convex analysis. Princeton (NJ): Princeton University Press; 1970.
8. Hiriart-Urruty J-B, Lemaréchal C. Convex analysis and minimization algorithms. Berlin: Springer; 1996.
9. Auslender A, Teboulle M. Asymptotic cones and functions in optimization and variational inequalities. New York: Springer; 2003.
10. Arrow KJ, Hurwicz L, Uzawa H. Studies in linear and non-linear programming. Stanford (CA): Stanford University Press; 1958.
11. Korpelevich GM. The extragradient method for finding saddle points and other problems. Matekon 1977;13:35–49.
12. Facchinei F, Pang J-S. Volume I and II, Finite-dimensional variational inequalities and complementarity problems. New York: Springer; 2003.
13. Konnov IV. Equilibrium models and variational inequalities. Amsterdam: Elsevier; 2007.
14. Nemirovski A. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J Optim 2004;15(1):229–251.
15. Auslender A, Teboulle M. Interior projection-like methods for monotone variational inequalities. Math Program 2005;104:39–68.
16. Auslender A, Teboulle M. Projected subgradient methods with non-Euclidean distances for nondifferentiable convex minimization and variational inequalities. Math Program B 2009;120:27–48.
17. Nesterov Y. Dual extrapolation and its applications to solving variational inequalities and related problems. Math Program 2007;109:319–344.
18. Nedić A, Ozdaglar A. Subgradient methods for saddle-point problems. J Optim Theory Appl 2009;142(1):205–228.
19. Shor NZ. Minimization methods for nondifferentiable functions. Berlin: Springer; 1985. (Translated from Russian by Kiwiel KC, Ruszczyński A).
20. Polyak BT. Introduction to optimization. New York: Optimization Software, Inc.; 1987.
21. Kiwiel KC. Efficiency of the analytic center cutting plane method for convex minimization. SIAM J Optim 1997;7:336–346.
22. Goffin JL, Vial JP. Convex nondifferentiable optimization: a survey focused on the analytic center cutting plane method. Optim Methods Softw 2002;17:805–867.
23. Kiwiel KC. Methods of descent for nondifferentiable optimization. Vienna, Austria: Springer; 1985.
24. Kiwiel KC. Proximity control in bundle methods for convex nondifferentiable minimization. Math Program 1990;46:105–122.
25. Schramm H, Zowe J. A combination of the bundle approach and the trust region concept. Mathematical research 45, advances in mathematical optimization. Berlin: Akademie Verlag; 1988. pp. 196–209.
26. Lemaréchal C, Sagastizábal C. An approach to variable metric bundle methods. Systems modelling and optimization. Volume 197, Lecture notes in control and information sciences. Berlin/Heidelberg: Springer; 1993. pp. 144–162.
27. Mifflin R, Sun D, Qi L. Quasi-newton bundle-type methods for nondifferentiable convex optimization. SIAM J Optim 1998;8:583–603.
28. Mäkelä MM. Survey of bundle methods for nonsmooth optimization. Optim Methods Softw 2002;17(1):1–29.

LARGE DEVIATIONS IN QUEUEING SYSTEMS

AYALVADI J. GANESH
Department of Mathematics,
University of Bristol,
Bristol, UK

BACKGROUND

Queueing theory had its origins in the work of Erlang on telephone networks [1], and has since grown into a branch of operations research with applications in telecommunication and computer networks, manufacturing, service centers, and a number of other areas. In the process, there has been a profusion of mathematical models of queueing networks. The standard approach to obtaining performance measures of queueing systems is to solve for the invariant or equilibrium distribution of the corresponding mathematical model and use that to compute long-run averages of performance measures. However, it turns out that the invariant distribution can only be calculated exactly for a few special queueing models. As a result, practitioners may be faced with a choice between shoehorning the system they are studying into a tractable mathematical model, although the model assumptions are unrealistic for their system, or relying entirely on simulations. Both these choices have their drawbacks. A third alternative is to study asymptotic regimes, in which some “system size” parameter tends to infinity, and in which the model becomes tractable. This approach can be attractive in studying many large systems. Large deviation theory provides one such asymptotic approach.

We begin with a brief review of some widely known results about queueing models. The most classical queueing models are the $M/M/1$ and $M/M/\infty$ queues, in the Kendall notation. (The reader is referred to the encyclopedia articles *The M/G/1 Queue*

and *The M/G/ ∞ Queue*, which include these as special cases.) The queue length process in these two models is a reversible Markov process. This facilitates computation of their invariant distribution, which turns out to be geometric for the $M/M/1$ queue and Poisson for the $M/M/\infty$ queue. The same approach can be used if such queues are connected together in networks with Markovian routing. Such networks are known as *Jackson networks*, and their invariant distribution is of product form. For more details, see the book of Kelly [2]. Unfortunately, this approach only extends to a few special queueing models. Other important examples in this class are $M/GI/1$ -PS and $M/GI/1$ -LCFS-PR queues, with Poisson arrivals, iid service times with general distribution, and a single-server operating either a processor sharing (PS) or last-come-first-served with preemptive resume (LCFS-PR) service discipline, and $M/GI/\infty$ queues. The approach also extends to networks made up of such queues, with Markovian routing between queues. In general, the approach applies to networks that are termed *symmetric queues* [2]. But apart from these few specific models, invariant distributions cannot generally be computed explicitly. In the case of single queues, expressions can be derived for the Laplace transform, or moment-generating function, of the queue length distribution [3], but the techniques involved do not readily generalize to networks.

What can we say more generally about queueing network models? It turns out that, while exact calculations are generally not feasible, results can be obtained for a wide class of models in certain asymptotic regimes. One such regime, which is popular in applications of queueing theory to manufacturing systems, is the heavy-traffic limit. Here, one considers a sequence of systems with, say, the same service time distribution F , but with increasing arrival rate, such that the system is critically loaded in the limit. (By critically loaded, we mean that the rate at which work comes into the system is equal to the rate at

which work can be done.) In this setting, the sequence of queue length processes, suitably scaled, converge to diffusion processes that can be analyzed for a wide range of models.

The heavy-traffic limit is suitable for manufacturing applications, where one typically wishes to operate the system with negligible idling, but it is not appropriate for telecommunication systems. In the latter, it is more common to seek to ensure that blocking or excessive delays are infrequent, even if this requires providing some level of excess capacity. A different asymptotic regime, based on large deviations, is more appropriate in this setting. We give an introduction to large deviations in the next section before we go on to apply it to some queueing models in the following section.

INTRODUCTION TO LARGE DEVIATIONS

Let $X_1, X_2, X_3, \dots$ be a sequence of iid standard normal random variables. Define the partial sums $S_n = \sum_{i=1}^n X_i$. Then S_n/n^α has the normal distribution $N(0, n^{1-2\alpha})$, and a straightforward calculation yields that, for all $\alpha > 1/2$,

$$\begin{aligned} \frac{1}{n^{2\alpha-1}} \log \mathbb{P} \left(\frac{S_n}{n^\alpha} \geq x \right) \\ = \frac{1}{n^{2\alpha-1}} \log \int_x^\infty \sqrt{\frac{n^{2\alpha-1}}{2\pi}} e^{-\frac{n^{2\alpha-1}x^2}{2}} dx \\ \rightarrow \begin{cases} -\frac{x^2}{2}, & x \geq 0, \\ 0, & x < 0. \end{cases} \end{aligned}$$

Let μ_n denote the law of the random variable S_n/n^α . We can then rewrite the above as

$$\frac{1}{n^{2\alpha-1}} \log \mu_n([x, \infty)) \rightarrow -V([x, \infty)),$$

where

$$V([x, \infty)) = \begin{cases} \frac{x^2}{2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The set function V admits a representation as an infimum: $V([x, \infty)) = \inf_{z \in [x, \infty)} I(z)$, where $I(z) = z^2/2$. The above equation says that the

sequence of probability measures μ_n (or random variables S_n/n^α) concentrates around zero at an exponential rate. This is essentially the structure of a large deviation principle, which we now state in a rather general form before considering some concrete examples.

Let E be a Hausdorff space equipped with the topology τ and the corresponding Borel σ -algebra. We say that the sequence of Borel probability measures $\mu_n, n \in \mathbb{N}$ on (E, τ) satisfies a large deviation principle (LDP) with speed a_n and rate function I if there exists a sequence of real numbers a_n such that the inequalities

$$\begin{aligned} -\inf_{x \in B^0} I(x) &\leq \liminf_{n \rightarrow \infty} \frac{1}{a_n} \log \mu_n(B) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{a_n} \log \mu_n(B) \leq -\inf_{x \in \bar{B}} I(x) \end{aligned} \quad (1)$$

hold for all Borel sets B . Here, B^0 and $\bar{B}$ denote the interior and closure of B respectively. We say that the rate function I is good if it has compact level sets.

Examples

1. *Cramer's Theorem.* Let $X_1, X_2, \dots$ be a sequence of iid random variables, with associated partial sums $S_n = X_1 + X_2 + \dots + X_n$. The logarithmic moment-generating function of X_1 is defined as

$$\Lambda(\theta) = \log \mathbb{E} \left[e^{\theta X_1} \right], \quad \theta \in \mathbb{R},$$

and takes values in $\mathbb{R} \cup \{+\infty\}$. The convex conjugate of this function is defined as

$$\Lambda^*(x) = \sup_{\theta \in \mathbb{R}} \theta x - \Lambda(\theta),$$

and is also an extended real-valued function. Under these assumptions, the sequence of sample means S_n/n satisfies an LDP on $\mathbb{R}$ (endowed with the usual metric) with speed n and rate function Λ^* , which is a good rate function if Λ is finite in a neighborhood of the origin.

2. *Gärtner–Ellis Theorem.* Let $Y_1, Y_2, \dots$ be a sequence of random variables, with corresponding logarithmic moment-generating functions $\Lambda_n(\theta) = \log \mathbb{E}[\exp(\theta Y_n)]$, and suppose that the limit

$$\Lambda(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\theta)$$

exists as an extended real number for all $\theta \in \mathbb{R}$. Suppose, too, that Λ is lower semicontinuous, it is finite in a neighborhood of the origin, differentiable in the interior of its domain, and $\Lambda'(\theta)$ tends to infinity as θ tends to any point on the boundary of the domain of Λ . (The domain is defined as the set on which Λ is finite. The last requirement is termed *steepness*.) Under these assumptions, the sequence of random variables Y_n satisfies an LDP on $\mathbb{R}$ with speed n and good rate function Λ^* .

3. *Borovkov’s and Mogulskii’s Theorems.* We now consider a much more abstract setting. Let $X_1, X_2, \dots$ be an iid sequence of random variables as before. For all $t \in [0, 1]$ and $n \in \mathbb{N}$, let $[nt]$ denote the integer part of nt , and define

$$\begin{aligned} \hat{S}_n(t) &= \sum_{k=1}^{[nt]} X_k, \quad \tilde{S}_n(t) = \hat{S}_n(t) \\ &\quad + (nt - [nt]) X_{[nt]+1}. \end{aligned} \quad (2)$$

Thus, $\hat{S}_n(\cdot)$ is a piecewise constant partial sums process, while $\tilde{S}_n(\cdot)$ is its piecewise linear interpolation. Now $\tilde{S}_n(\cdot)$ lives in the space $C([0, 1])$ of continuous functions on $[0, 1]$, and can be viewed as a random element of this space.

Suppose that $\Lambda(\theta) = \log \mathbb{E}[\exp(\theta X_1)]$ is finite for all $\theta \in \mathbb{R}$. Then, the theorem states that the sequence $\tilde{S}_n(\cdot)/n$ satisfies an LDP on $C([0, 1])$ equipped with the topology of uniform convergence, with speed n , and good rate function

$$\mathcal{I}(\phi) = \begin{cases} \int_0^1 \Lambda^*(\phi'(t)) dt, & \phi \in AC_0([0, 1]), \\ +\infty, & \text{otherwise.} \end{cases}$$

Here ϕ' denotes the derivative of ϕ , and $AC_0([0, 1])$ denotes the space of absolutely continuous functions f on $[0, 1]$ with $f(0) = 0$.

This result was first established by Borovkov for real-valued random variables [4] and extended to $\mathbb{R}^d$ -valued random variables by Mogulskii [5].

The motivating example at the start of this section suggests that the Gaussian random variables S_n/n^α satisfy an LDP (though we only established the bounds in Equation (1) for sets of the form $[x, \infty)$ rather than all Borel sets). This can be formally proved using the Gärtner–Ellis theorem, or Cramer’s theorem in the $\alpha = 1$ case.

The theorems above can be readily extended to random variables X_k that take values in $\mathbb{R}^d$ rather than in $\mathbb{R}$. In this case, the logarithmic moment-generating function is an extended real-valued function defined as $\Lambda(\theta) = \log \mathbb{E}[\exp(\theta \cdot X_1)]$ for $\theta \in \mathbb{R}^d$, where $\theta \cdot X_1$ denotes the inner product of θ and X_1 .

The LDP in Borovkov’s and Mogulskii’s theorems is termed as a *functional LDP* or a *sample path LDP* as it is an LDP for the partial sums process in path space/function space. These theorems can be extended to the space $C([0, \infty))$ (of continuous functions taking values in $\mathbb{R}^d$ for $d \geq 1$) equipped with a suitable topology. First, we can define $\tilde{S}_n(t)$ using Equation (2) for all $t \in \mathbb{R}$ and not just for $t \in [0, 1]$. By the law of large numbers for the iid random variables X_i , we have $\tilde{S}_n(t)/t \rightarrow \mathbb{E}[X_1]$ almost surely as $t \rightarrow \infty$, for each n . Hence, if we define $\mu = \mathbb{E}[X_1]$ and let C^μ denote the subset of $C([0, \infty))$ consisting of all functions f satisfying $f(0) = 0$ and $\lim_{x \rightarrow \infty} f(x)/x = \mu$, then $\tilde{S}_n(\cdot) \in C^\mu$ almost surely. Equip C^μ with the topology induced by the scaled uniform norm,

$$\|\phi\| = \sup_{t \geq 0} \left| \frac{\phi(t)}{1+t} \right|,$$

noting that $\|\phi\|$ is finite for all $\phi \in C^\mu$ since $\phi(t)/t$ tends to a finite limit μ as t tends to infinity. It can be shown that C^μ equipped with this topology is a Polish space. The scaled uniform norm has been used in the study of weak convergence, and was first used

in the large deviations context by Deuschel and Stroock [6].

Now, Borovkov's and Mogulskii's theorems can be strengthened to yield an LDP for the sequence of scaled sample paths $\tilde{S}_n(\cdot)$ on the space C^μ , with good rate function

$$\mathcal{I}(\phi) = \begin{cases} \int_0^\infty \Lambda^*(\nabla\phi(t)) dt, & \phi \in AC_0([0, \infty)) \cap C^\mu; \\ +\infty, & \text{otherwise.} \end{cases}$$

The gradient $\nabla\phi$ is just the derivative ϕ' if the underlying random variables are real-valued. We can relax the assumption that the X_k are iid to suitable mixing conditions. The theorem still holds, with the modification that the term $\Lambda^*(\cdot)$ that appears in the rate function is now defined as

$$\begin{aligned} \Lambda^*(x) &= \sup_{\theta \in \mathbb{R}^d} \theta \cdot x - \Lambda(\theta), \\ \Lambda(\theta) &= \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}[\exp(n\theta \cdot \tilde{S}_n(1))]. \end{aligned}$$

The existence of the above limit is part of the assumptions required for the extension to be valid. See Dembo and Zajic [7] for details of the assumptions required to obtain the LDP on $C([0, 1])$, and Ganesh *et al.* [8] for the extension of the LDP to the space C^μ .

In Cramer's theorem, and in the versions of Borovkov's and Mogulskii's theorems that we stated, we scaled the random variables S_n and $\tilde{S}_n$ by n , which is the scaling required to get a law of large numbers or functional law of large numbers, respectively, which has a deterministic limit. Another common scaling involves centering these random variables by subtracting their means, and then scaling by $\sqrt{n}$. This is the central limit scaling (if variances are finite) and yields a stochastic limit, which is either a Gaussian random variable or a Gaussian process. One could instead scale the centered random variable by n^α , for some α between $1/2$ and 1 . (These and more general scalings are considered in the cited papers of Borovkov and Mogulskii.) The resulting sequence of random variables satisfies an LDP with speed smaller than n . Sometimes, the term LDP is reserved for the usual law of large numbers scaling n , and the term moderate deviation principle (MDP) is used for these intermediate scalings

between the law of large numbers and central limit theorem scalings. The distinction is somewhat artificial but not without justification; while the LDP rate function typically depends on the distribution of the random variables involved, the MDP rate function is often quadratic and depends on the random variables only through their variance. To put it another way, fluctuations are usually approximately Gaussian on MDP scales, but there is no such universality on LDP scales.

There are a number of different techniques for establishing LDPs for sequences of random variables, but these are not the subject of this article. The approach that is most relevant to us here is a method for getting new LDPs from old via continuous transformations. It can be seen as an analog of the continuous mapping theorem in weak convergence.

Contraction Principle. Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff topological spaces, and let $f : X \rightarrow Y$ be continuous. Suppose that $(X_n, n \in \mathbb{N})$ satisfy an LDP in $\mathcal{X}$ with speed a_n and good rate function I , and that $Y_n = f(X_n)$. Then, $(Y_n, n \in \mathbb{N})$ satisfy an LDP in $\mathcal{Y}$, with the same speed a_n , and with good rate function

$$J(y) = \inf_{x: f(x)=y} I(x),$$

where the infimum of an empty set is taken to be $+\infty$.

THE SINGLE SERVER QUEUE

We consider a queue operating in discrete time with a single server that serves work in its order of arrival. Let A_n denote the amount of work arriving into the queue in the n th time slot and B_n the amount of work that can be served during this time slot. These may be whole numbers corresponding to numbers of customers, but may more generally be nonnegative real numbers. We assume that $A_n, n \in \mathbb{Z}$ and $B_n, n \in \mathbb{Z}$ are stationary and ergodic stochastic processes.

The results described in this section can be extended to queues operating in continuous time, but we restrict ourselves to the

discrete-time setting for technical simplicity. The discrete-time model could also be justified by considering the queueing process embedded at customer arrival epochs, for instance. Note that we do not assume that interarrival and service times are iid. In many telecommunications applications, there is significant time correlation in these processes, and so the relaxation of independence assumptions is important.

Let W_n denote the amount of work in the queue at the beginning of time slot n . The evolution of W_n is described by Lindley's recursion, which states that

$$W_{n+1} = (W_n - X_n)^+, \text{ where } X_n = A_n - B_n \text{ and } x^+ = \max\{x, 0\}. \quad (3)$$

The recursion simply states that the work in queue at the beginning of time slot $n+1$ is the amount of work that was there at the beginning of time slot n plus the new work that arrived in that time slot, less the amount of work that could have been served, unless more work could have been served than there was in the queue. In the latter case, the queue empties and the amount of work in queue is zero. Implicitly, we think of work as arriving at the beginning of a time slot and being served at the end of the time slot, along with work that is already present.

We assume throughout the following discussion that A_1 and B_1 have finite means, and that $\mathbb{E}[A_1] < \mathbb{E}[B_1]$. This is a standard stability assumption, without which the work in queue tends to infinity over time. Under these assumptions, Loynes [9] showed that Equation (3) has a unique stationary solution. In particular, the solution for W_0 can be written as

$$W_0 = \sup_{n \geq 0} \sum_{k=-n}^{-1} X_k, \quad (4)$$

where an empty sum is taken to be zero by convention. Under the stability assumption, $\mathbb{E}[X_1] < 0$, so the sum on the right-hand side of Equation (4) tends to negative infinity as n tends to infinity, by the law of large numbers. Hence, its supremum is attained at some finite n , almost surely.

Define the process $\tilde{S}_n(t), t \geq 0$ as in Equation (2), with $X_k = A_k - B_k$ as above. Suppose that the process $\tilde{S}_n(\cdot)$ satisfies an LDP on C^μ equipped with the topology induced by the scaled uniform norm, with speed n and with good rate function $\mathcal{I}$. Then, we show that the scaled workload, W_0/n , also satisfies an LDP with the same speed, and identify its rate function under some additional conditions.

The proof of the LDP for W_0/n follows by a straightforward application of the contraction principle. For convenience, we rewrite Equation (4) as

$$W_0 = \sup_{n \geq 0} \sum_{k=1}^n X_k.$$

By the stationarity of the sequence $(X_k, k \in \mathbb{Z})$, the random variable W_0 defined above has the same distribution as the W_0 defined in Equation (4), which is the distribution of the stationary workload in the queue. But the above equation can be rewritten as

$$\frac{W_0}{n} = \sup_{t \geq 0} \tilde{S}_n(t), \quad (5)$$

as the supremum of $\tilde{S}_n$ is necessarily attained at some t such that nt is an integer. Hence, if we can show that the mapping $\phi \mapsto \sup_{t \geq 0} \phi(t)$ is continuous on C^μ , then we can invoke the contraction principle to obtain an LDP for W_0/n .

Recall that $\mu = \mathbb{E}[X_1] < 0$ by the stability assumption. Fix $\phi \in C^\mu$ and $\delta > 0$, and suppose $\psi \in C^\mu$ is such that $\|\phi - \psi\| < \delta$. Now, $\phi(t)/t$ tends to $\mu < 0$ as t tends to infinity, so there is some $T > 0$ such that $\phi(t) < -2\delta$ for all $t > T$. Since $\phi(0) = \psi(0) = 0$ by the assumption that ϕ and ψ are in C^μ , the supremum of ϕ and ψ are both attained on $[0, T]$. But $|\phi(x) - \psi(x)| \leq \delta(1 + T)$ for all $x \in [0, T]$ by the assumption that $\|\phi - \psi\| < \delta$. Hence, if we define $f : C^\mu \rightarrow \mathbb{R}_+$ by $f(\phi) = \sup_{t \geq 0} \phi(t)$, then we have shown that $|f(\phi) - f(\psi)| \leq \delta(1 + T)$ for all ψ such that $\|\phi - \psi\| < \delta$, where T is a constant that depends only on ϕ . This shows that f is continuous.

Since $W_0/n = f(\tilde{S}_n)$, it immediately follows from the contraction principle and the assumed LDP for $\tilde{S}_n$ that W_0/n satisfies an

LDP on $\mathbb{R}_+$ at speed n , and that the rate function for this LDP is given by

$$\begin{aligned} J(x) &= \inf_{\phi \in C^\mu: f(\phi)=x} \mathcal{I}(\phi) \\ &= \inf \left\{ \mathcal{I}(\phi) : \phi \in C^\mu, \sup_{t \geq 0} \phi(t) = x \right\}. \end{aligned} \quad (6)$$

The rate function J is thus given as the solution of a variational problem. If the rate function $\mathcal{I}$ is of the form

$$\mathcal{I}(\phi) = \int_0^\infty \Lambda^*(\phi'(t)) dt,$$

for a convex function Λ^* that takes its minimum value of 0 uniquely at μ (as is the case with the extension of Mogulskii's theorem discussed above), then we can reduce the variational problem in Equation (6) to a finite-dimensional optimization problem. It can be shown, using the convexity of Λ^* and Jensen's inequality, that the infimum in Equation (6) is attained by a ϕ that is piecewise linear, and of the form

$$\phi(t) = \begin{cases} \alpha t, & 0 \leq t \leq T, \\ \mu(t - T), & t > T, \end{cases}$$

for constants $\alpha > 0$ and $T > 0$, which are obtained by solving the finite-dimensional optimization problem stated below. The rate function J can be expressed in terms of its solution, as follows:

$$\begin{aligned} J(x) &= \min_{\alpha, T > 0: \alpha T = x} T \Lambda^*(\alpha) \\ &= x \min_{\alpha > 0} \frac{\Lambda^*(\alpha)}{\alpha}. \end{aligned} \quad (7)$$

Observe that the rate function is linear in x . The LDP implies that $\mathbb{P}(W_0 > w) \approx e^{-wJ(1)}$ for large w . More precisely, the two sides are logarithmically equivalent as w tends to infinity, that is,

$$\lim_{w \rightarrow \infty} \frac{1}{w} \log \mathbb{P}(W_0 > w) = -J(1). \quad (8)$$

The result thus provides quite general conditions under which the distribution of the work in queue has exponential tails, even for arrival and service processes that may show

a considerable degree of time dependence. Conditional on the build-up of a large workload w , the optimization problem in equation (7) also provides information about the most likely trajectory by which the workload built up. It says that the most likely trajectory involved the queue building up at constant rate α for a time period of length Tw , where α and T achieve the minimum for $J(1)$. Thus, the rate of queue build-up α does not depend on w , while the length of time over which the queue builds up scales linearly in w .

The asymptotic regime in Equation (8) is sometimes referred to as the *large buffer asymptotic*, as it describes the probability of the contents of the queue exceeding the capacity of a large buffer. This is one of the regimes of interest in a communication network, where the buffer stores data at the ingress of a network link. Another regime of interest is the so-called many flows regime, where the input (arrival process) is treated as the superposition of a large number of iid arrival processes, and the service capacity scales linearly in this number. The study of this regime is quite similar to that of the large buffer regime. It involves using a sample path LDP for the aggregate arrival process scaled by the number of superposed flows. Thus, the scaling here is over space (number of flows, n) rather than time (aggregate arrivals over a period of length nt), but the approach is otherwise similar and involves the use of the contraction principle. (The map from the arrival process to the queue size is identical, and hence is continuous in the topology induced by the scaled uniform norm.) The main difference is that the rate function for the scaled queue length need not be linear in this regime. The reader is referred to Ganesh *et al.* [8] for details, both for the many flows scaling regime and for moderate deviation scalings.

DISCUSSION

This article aimed to provide an elementary introduction to large deviations theory, and to its application in queueing systems. It only treated the very simple case of a FIFO queue with a single server, albeit with

quite general arrival and service processes. Despite the generality of the driving processes considered, this is still the simplest and most classical queueing model. It would be natural to ask if large deviation techniques permit the treatment of a much wider class of queueing models. In particular, can they be applied to both single and multiclass queues implementing a variety of different service disciplines including processor-sharing and priority-based disciplines? Can they also be applied to networks of queues?

The general answer is that progress has been quite slow and piecemeal. While there is an elegant unifying framework based on establishing the continuity of queueing maps and employing the contraction principle, the details of carrying out these steps pose considerable challenges, which have to be tackled using different techniques in different models. One of the problems is that the queueing maps of interest may not satisfy the strict continuity assumptions described here. In some instances, generalizations of the contraction principle based on Garcia [10] and Puhalskii [11] can be used to obtain an LDP [12]. Another common difficulty is that, while continuity may be established and the existence of an LDP deduced thereby, the variational problem describing the rate function turns out to be intractable. One is then forced to turn to bounds and approximations.

The field continues to provide a rich source of open problems, of both a mathematical and an applied nature, motivated by communication and computer networks, by new scheduling and routing algorithms, and by new application domains such as wireless communications.

REFERENCES

1. Erlang AK. The theory of probabilities and telephone conversations. *Nyt Tidsskr Mat* 1909;B 20:33–39.
2. Kelly FP. Reversibility and stochastic networks. New York: Wiley; 1979.
3. Cohen JW. The single server queue. 2nd ed. Amsterdam: North-Holland Publishing Company; 1982.
4. Borovkov AA. Boundary value problems for random walks and large deviations in function spaces. *Theory Probab Appl* 1967;12: 575–595.
5. Mogulskii AA. Large deviations for the trajectories of multi-dimensional random walks. *Theory Probab Appl* 1976;21:300–315.
6. Deuschel JD, Stroock DW. Large deviations. New York: Academic Press; 1989.
7. Dembo A, Zajic T. Large deviations: from empirical mean and measure to partial sums process. *Stoc Proc Appl* 1995;57:191–224.
8. Ganesh A, O'Connell N, Wischik D. Big queues. Berlin/Heidelberg: Springer; 2004.
9. Loynes RM. The stability of queues with non-independent inter-arrival and service times. *Proc Camb Philos Soc* 1962;58:497–520.
10. Garcia J. An extension of the contraction principle. *J Theor Probab* 2004;17:403–434.
11. Puhalskii A. Large deviations and idempotent probability. London: Chapman and Hall; 2001.
12. Kittipiyakul S, Javidi T, Subramanian VG. Many-sources large deviations for max-weight scheduling. Proceedings of the 46th Annual Allerton Conference. Allerton, Illinois, USA: 2008.

LARGE MARGIN RULE-BASED CLASSIFIERS

TIBÉRIUS O. BONATES
Princeton Consultants, Inc.,
Princeton, New Jersey

INTRODUCTION

In this article, we present an optimization approach for the construction of rule-based classification models, with a focus on logical analysis of data (LAD).

The main idea behind LAD is to create a relevant set of logical rules (called *patterns*) that can collectively be used for classification. Each pattern consists of one or more conditions on the values of features such as “*feature 1 = 0*” or “*feature 3 ≥ 3.2*.” LAD is usually applied to two-class classification problems via a system of weights that assigns a positive weight to patterns that are characteristic of one of the two classes, and a negative weight to patterns that are characteristic of the opposite class. For any given point, its classification according to LAD is given by the total weight obtained by adding up the individual weights of the patterns, whose defining conditions hold on at that point.

Constructing a set of patterns that accurately characterizes the two classes, and properly assigning weights to these patterns in order to obtain a robust classification of unseen points, is crucial for the application of LAD to any given data set. Different algorithms have been proposed for the construction of specific families of patterns [1–4]. Similarly, different approaches have been evaluated for assigning weights to patterns [5,6].

Our algorithm utilizes a clear performance criterion that guides the search for a set of patterns (or rules), reducing the need for a careful calibration of a number of control parameters, as is typically the case with

most classification methods. Moreover, our algorithm automatically adjusts the weights assigned to these patterns while optimizing the performance criterion, accomplishing in a single effort, the distinct tasks of pattern generation and selection of a set of weights defining a robust classifier.

The article is organized as follows: in the section titled “LAD Patterns and Classifiers,” we review the main concepts in LAD. The section titled “Large Margin LAD Classifiers” describes the concept of large margin LAD classifiers and the optimization model, which is the basis of our algorithm. In the section titled “Pricing Subproblem,” we describe the optimization subproblem used to construct patterns as part of our iterative optimization algorithm, which is discussed at length in the section titled “Column Generation Algorithm.” Finally, the section titled “Conclusions” summarizes our results and suggests future research directions.

LAD PATTERNS AND CLASSIFIERS

Consider a binary data set $\Omega \subset \{0, 1\}^n$ consisting of two disjoint sets of positive and negative points, that is, $\Omega = \Omega^+ \cup \Omega^-$, with $\Omega^+ \cap \Omega^- = \emptyset$. While we assume that Ω is a binary data set, the ideas described in this article can be applied to data sets with numerical attributes via the use of a discretization procedure. As part of the standard practice of applying LAD to a data set, a discretization step is carried out according to the method described in Ref. 5. For an overview of some commonly used discretization procedures the reader is referred to Ref. 7.

Patterns

A *pattern* or *rule* is simply a homogeneous subcube of $\{0, 1\}^n$, that is, a subcube having (i) a nonempty intersection with one of the sets Ω^+ or Ω^- , and (ii) an empty intersection with the other set (Ω^- or Ω^+ , respectively). We recall that a subcube consists of those points of the n -cube for which a subset of

the variables is fixed to 0 or 1, while the remaining variables take all the possible 0, 1 values. A pattern P disjoint from Ω^- is called a *positive pattern*, while a pattern N disjoint from Ω^+ is called a *negative pattern*. A point in $\{0, 1\}^n$ is said to be *covered* by a pattern if it belongs to the subcube defined by that pattern.

In the remainder of this article, we will use a relaxed definition of a pattern, in which a positive pattern is allowed to cover some negative points. Similarly, a negative pattern is allowed to cover some positive points. Traditionally, the number of points of the opposite type that a pattern is allowed to cover is controlled by parameters called *homogeneity* or *fuzziness* [1,3,4]. In this article, we will refrain from defining such parameters, since our algorithm does not directly consider these values, but instead implicitly considers *all* possible patterns.

A positive (or negative) pattern can be thought of simply as a realization of a subset of the components in our data set that is characteristic of the Ω^+ (or Ω^-) set. A pattern is typically expressed as a Boolean conjunction such as $P = \prod_{i \in A} x_i \prod_{j \in B} \bar{x}_j$, where x_i is a Boolean literal corresponding to the i th component of the data set, and $\bar{x}_j$ is the associated negated literal. The literal x_i is considered satisfied by a certain point $\omega \in \{0, 1\}^n$ if and only if $\omega_i = 1$. Conversely, $\bar{x}_j$ is satisfied if and only if $\omega_j = 0$. A literal is considered to have a value of 1 if it is satisfied, and a value of 0 otherwise.

Models and Discriminant Functions

An *LAD model* consists of a set M of positive and negative patterns, so that every point in Ω^+ is covered by at least one of the positive patterns in M , and every point in Ω^- is covered by at least one of the negative patterns in M . The construction of an LAD model is a computationally expensive task that typically involves the enumeration of a large set of patterns satisfying a certain property, followed by the selection of a relatively small subset of those [1,8,3].

The fact that only a fraction of the patterns generated is ultimately used for the construction of an LAD model suggests that the approach described above

for model construction may not be the most efficient one. Moreover, the choice of enumerating a specific family of patterns, and the usual selection of a small number of those patterns—which is frequently guided by a greedy criterion—may not be the most appropriate for certain problems, resulting in suboptimal LAD models. In view of this, it is appropriate to say that there is a clear need for a global criterion for choosing an overall “best” LAD model.

Given an LAD model $M = M^+ \cup M^-$ (where M^+ is a set of positive patterns and M^- is a set of negative patterns) and a point $\omega \in \{0, 1\}^n$, the classification of ω is determined by the sign of a so-called *discriminant function* $\Delta : \{0, 1\}^n \rightarrow \mathbb{R}$ associated to the given model. Let $M^+ = \{P_1, \dots, P_{|M^+|}\}$ and $M^- = \{N_1, \dots, N_{|M^-|}\}$, and let us associate to each positive pattern $P_i \in M^+$, a real weight α_i and to each negative pattern $N_j \in M^-$ a real weight β_j . The associated discriminant function on ω is given by

$$\Delta(\omega) = \sum_{P_i \in M^+} \alpha_i P_i(\omega) - \sum_{N_j \in M^-} \beta_j N_j(\omega),$$

where $P_i(\omega)$ equals 1 if ω is covered by P_i , and 0 otherwise (similarly for $N_j(\omega)$). The sign of $\Delta(\omega)$ determines the classification of ω . If $\Delta(\omega) = 0$ the observation is either left unclassified, or is classified according to the more frequent class.

For a given set of patterns, the construction of a discriminant function for classification is usually done either by majority vote among positive and negative patterns (i.e., by assigning equal weights among positive and among negative patterns), or by solving a linear program that provides a set of weights resulting in an optimal separation of the two classes (as described in Ref. 6). In the next section, we will describe a formulation based on the linear program of Ref. 6, which finds an optimal discriminant function defined over the set of all patterns.

LARGE MARGIN LAD CLASSIFIERS

It is important to develop an algorithm that selects a set of patterns and a set of weights

to build an optimal discriminant function according to a certain performance criterion that relates to the generalization of the classifier. A natural criterion is the *separation margin* of the classifier [9,10]. In the context of LAD, the separation margin is simply the difference between the smallest value that the discriminant function takes over the positive points of the training set that are correctly classified and the largest value that the discriminant function takes over the negative points in the training set that are correctly classified. Thus, the separation margin corresponding to a given discriminant function is

$$\min\{\Delta(\omega) : \omega \in \Omega^+, \Delta(\omega) > 0\} \\ - \max\{\Delta(\omega) : \omega \in \Omega^-, \Delta(\omega) < 0\}.$$

By maximizing the separation margin, we expect a robust classification of unseen examples. This expectation has been confirmed in practice in many situations [11–13]. The good performance of boosting and voting-based classifiers has also been shown to be related to the implicit maximization of the separation margin of the resulting classifier [13].

Linear Programming Formulation

In what follows, we show how to formulate the problem of constructing an optimal discriminant function by using a linear programming formulation, similar to the one described in Ref. 6. This linear programming formulation involves a very large number of variables and is shown to be amenable to a solution via column generation [14–17].

Let $\mathbb{P} = \{P_1, \dots, P_{|\mathbb{P}|}\}$ be the set of *all* positive patterns, and let $\mathbb{N} = \{N_1, \dots, N_{|\mathbb{N}|}\}$ be the set of *all* negative patterns. We are interested in solving the following problem:

$$(P) \quad \text{maximize} \quad r + s - C \sum_{\omega \in \Omega} \epsilon_{\omega}$$

subject to:

$$\sum_{i=1}^{|\mathbb{P}|} \alpha_i P_i(\omega) - \sum_{j=1}^{|\mathbb{N}|} \beta_j N_j(\omega) + \epsilon_{\omega} \geq r, \quad \forall \omega \in \Omega^+; \quad (1)$$

$$\sum_{i=1}^{|\mathbb{P}|} \alpha_i P_i(\omega) - \sum_{j=1}^{|\mathbb{N}|} \beta_j N_j(\omega) - \epsilon_{\omega} \leq -s, \\ \forall \omega \in \Omega^-, \quad (2)$$

$$\sum_{i=1}^{|\mathbb{P}|} \alpha_i = \sum_{j=1}^{|\mathbb{N}|} \beta_j = 1,$$

$$r \geq 0, s \geq 0$$

$$\alpha_i \geq 0, i = 1, \dots, |\mathbb{P}|, \beta_j \geq 0, j = 1, \dots, |\mathbb{N}|$$

$$\epsilon_{\omega} \geq 0, \forall \omega \in \Omega,$$

where the values of α_i ($i = 1, \dots, |\mathbb{P}|$) and β_j ($j = 1, \dots, |\mathbb{N}|$) are the weights of the positive and negative patterns, respectively, r represents the positive part of the separation margin, s represents the negative part of the separation margin, and C is a nonnegative penalization parameter that controls how much importance is given to the violations ϵ_{ω} of the separating constraints of Equations (1) and (2).

Owing to the potentially very large total number of patterns, it is highly unlikely that in real-world applications one will encounter a data set, where such a formulation can be directly tackled. Thus, let us consider a subset of positive patterns $\mathcal{P}^0 \subset \mathbb{P}$ and a subset of negative patterns $\mathcal{N}^0 \subset \mathbb{N}$, and let us denote by $I^0 \subset \{1, \dots, |\mathbb{P}|\}$ and $J^0 \subset \{1, \dots, |\mathbb{N}|\}$ the sets of indices corresponding to the elements of $\mathcal{P}^0$ and $\mathcal{N}^0$, respectively. Then, we can write a restricted version of (P) as

$$(RP) \quad \text{maximize} \quad r + s - C \sum_{\omega \in \Omega} \epsilon_{\omega}$$

subject to:

$$r - \sum_{i \in I^0} \alpha_i P_i(\omega) + \sum_{j \in J^0} \beta_j N_j(\omega) - \epsilon_{\omega} \leq 0, \\ \forall \omega \in \Omega^+, \quad (3)$$

$$s + \sum_{i \in I^0} \alpha_i P_i(\omega) - \sum_{j \in J^0} \beta_j N_j(\omega) - \epsilon_{\omega} \leq 0, \\ \forall \omega \in \Omega^-, \quad (4)$$

$$\sum_{i \in I^0} \alpha_i = 1, \quad (5)$$

$$\sum_{j \in J^0} \beta_j = 1, \quad (6)$$

$$r \geq 0, s \geq 0$$

$$\alpha_i \geq 0, \forall i \in I^0, \beta_j \geq 0, \forall j \in J^0,$$

$$\epsilon_\omega \geq 0, \forall \omega \in \Omega.$$

Note that we rearranged constraints (3) and (4) in order to write RP (i.e., the restricted version of P) in a standard maximization form.

It is easy to choose $\mathcal{P}^0$ and $\mathcal{N}^0$ in such a way that RP has a feasible solution. We recall that the *minterm* corresponding to a given point ω is simply the pattern with the largest possible number of literals that is satisfied by ω , that is, it is the pattern given by the following conjunction: $T = \prod_{i:\omega_i=1} x_i \prod_{j:\omega_j=0} \bar{x}_j$. Clearly, T is a pattern that covers ω and does not cover any observation from the opposite part of the data set (since $\Omega^+ \cap \Omega^- = \emptyset$). Let us choose $\mathcal{P}^0$ and $\mathcal{N}^0$ to be the sets of minterms corresponding to the positive and negative observations in the data set, respectively. Then, the vectors α and β are given by $\alpha_i = \frac{1}{|\mathcal{P}^0|}$, $i \in \mathcal{P}^0$ and $\beta_j = \frac{1}{|\mathcal{N}^0|}$, $j \in \mathcal{N}^0$, along with $r = \frac{1}{|\mathcal{P}^0|}$, and $s = \frac{1}{|\mathcal{N}^0|}$ constitute a feasible solution to RP. As we have just showed, constructing a feasible set of patterns to initialize RP is not a difficult task; as a matter of fact, one pattern of each class would suffice.

It is advantageous, however, to find a set of good quality patterns with which to initialize the column generation algorithm as such patterns are likely to be part of the optimal solution and their use in early iterations may significantly speed up the entire procedure. (This should become clearer once we describe the column generation algorithm in a separate section). On the other hand, spending a considerable amount of time to generate patterns of good quality defeats the purpose of running our algorithm to automatically search for the best set of patterns for classifying the training data. In the experiments reported in the section

titled “Column Generation Algorithm,” we adopt a compromise by initializing our algorithm with the set of patterns consisting of single literals $\{x_1, \dots, x_n\}$. This amounts to attempting a linear separation using the original features of the data set. While this is not a sophisticated initial set of patterns, it is more likely that some of these patterns will be used in the optimal classifier than some of the minterms, which are clearly less descriptive of the data set as a whole.

PRICING SUBPROBLEM

In this section, we address the subproblem phase in the column generation algorithm. We formulate this problem as a pseudo-Boolean optimization problem that can also be seen as a weighted version of the *maximum pattern problem* [4,18].

The solution to the problem (RP) provides an optimal discriminant function for the set of patterns $\mathcal{P}^0 \cup \mathcal{N}^0$. However, that discriminant function may not be optimal with respect to the entire set of all patterns. In order to verify global optimality, we need to make sure that there is no pattern that once added to the current set of patterns, allows for an improvement in the value of the objective function of RP.

In the simplex method, one of the commonly used heuristics for choosing a nonbasic variable to enter the current basis is to choose the variable with the best (in the maximization case, the largest) reduced cost [19–22]. Similarly, we use the reduced cost of a specific pattern P to measure the potential improvement in the objective function of RP resulting from the introduction of P into the set of known patterns $\mathcal{P}_0 \cup \mathcal{N}_0$. To verify the optimality of the current solution of RP, we solve a pricing problem that will return a pattern with the best (largest) reduced cost with respect to the current values of the dual variables of RP.

According to the usual optimality criterion of the primal simplex method, if there is no pattern (not in $\mathcal{P}^0 \cup \mathcal{N}^0$) with a nonnegative reduced cost, then the current solution of RP is optimal to P; otherwise, a new pattern

can be added to the set $\mathcal{P}^0 \cup \mathcal{N}^0$ and the algorithm proceeds. Notice here that we allow the introduction of a pattern whose reduced cost is zero. Since the column generation algorithm is equivalent to applying the simplex algorithm to a large formulation, there is a possibility of having degenerate iterations in which no progress is made in terms of the objective function [19–22]. We will refer to a pattern with nonnegative reduced cost at a given iteration as a *candidate pattern*.

Note that since $r + s$ corresponds to the separation margin of the current discriminant function, solving the pricing problem corresponds to asking if the current separation margin can be enlarged by the addition of a pattern. Thus, the column generation algorithm applied to problem P attempts to increase the separation margin by including patterns that make the distinction between the sets Ω^+ and Ω^- more pronounced.

From now on, let us assume that k iterations of the column generation algorithm have been executed. We will refer to the corresponding sets of patterns as $\mathcal{P}^k$ and $\mathcal{N}^k$, and to the corresponding version of problem RP as RP^k . Let λ and μ be the vectors of dual variables associated with constraints (3) and (4), respectively, at an optimal solution of RP^k . Similarly, let θ^+ and θ^- be the dual variables corresponding to constraints (5) and (6). The reduced costs [19–22] corresponding to a positive pattern P_i and a negative pattern N_j are given by $\theta^+ + [\lambda^\top, -\mu^\top]\pi_i$ and $\theta^- + [-\lambda^\top, \mu^\top]v_j$, respectively, where $\pi_i = [P_i(\omega)]_{\omega \in \Omega^+}$ and $v_j = [N_j(\omega)]_{\omega \in \Omega^-}$ are the characteristic vectors of the observations covered by P_i and N_j , respectively.

Formulation

Let us define binary decision variables x_i and x_i^c ($i = 1, \dots, n$) associated with attribute a_i in the following way: (i) $x_i = 1$ and $x_i^c = 0$ if the literal associated to $a_i = 1$ is used in the resulting pattern; (ii) $x_i = 0$ and $x_i^c = 1$ if the literal associated with $a_i = 0$ is present in the resulting pattern. Thus, the positive pricing subproblem associated to optimal vectors of dual variables λ^* and μ^* of RP can be formulated as the following pseudo-Boolean

optimization problem:

$$\begin{aligned} (\text{SP}^+) \quad & \text{maximize} \sum_{\omega \in \Omega^+} \lambda_\omega^* \left(\prod_{i:\omega_i=0} \bar{x}_i \prod_{j:\omega_j=1} \bar{x}_j^c \right) \\ & - \sum_{\omega \in \Omega^-} \mu_\omega^* \left(\prod_{i:\omega_i=0} \bar{x}_i \prod_{j:\omega_j=1} \bar{x}_j^c \right) \end{aligned} \quad (7)$$

subject to:

$$x_j, x_j^c \in \{0, 1\}, \quad j = 1, \dots, n,$$

where each term $\prod_{i:\omega_i=0} \bar{x}_i \prod_{j:\omega_j=1} \bar{x}_j^c$ takes the value 1 whenever ω is covered by the resulting conjunction, and 0 otherwise.

Problem (SP^+) belongs to the family of pseudo-Boolean optimization problems, whose solution was addressed in Ref. 23. In our computational experiments, we utilized the branch-and-bound algorithm described in Ref. 23 to solve (SP^+) .

Note that since the value of the dual variable θ^+ appears as a constant term in the reduced cost formula, it is not required in the formulation of SP^+ . Also, by solving SP^+ , we aim at constructing a positive pattern, and therefore, it is implicit that the column corresponding to such a pattern has a coefficient equal to 1 in the row corresponding to constraint (5) and a coefficient equal to 0 in the row corresponding to constraint (6).

While SP^+ is aimed at finding the best positive candidate pattern at the current iteration, it may be necessary to search for a negative candidate pattern as well. In fact, even if we have already found a positive candidate pattern, it may be advantageous to generate a negative candidate pattern, if there is any. We suggest solving both SP^+ and its negative counterpart SP^- . For the negative pricing problem SP^- , we simply minimize the same objective function as in problem SP^k , as opposed to maximizing it in SP^+ .

The solutions of SP^+ and SP^- provide either: (i) a certificate of optimality of the current discriminant function, in case the objective function value of SP^+ is strictly negative and that of SP^- is strictly positive; or (ii)

-
1. Input: Ω^+, Ω^- , initial sets $\mathcal{P}^0$ and $\mathcal{N}^0$ of patterns. Set $k = 0$.
 2. Solve $(\text{RP}(k))$, with optimal primal and dual solutions $(r^*, s^*, \alpha^*, \beta^*)$ and $(\lambda^*, \mu^*, \theta^{+*}, \theta^{-*})$.
 3. Solve $(\text{SP}(k)^+)$ and $(\text{SP}(k)^-)$, with optimal positive and negative patterns P^* and N^* .
 4. If at least one of the stopping criteria is met, stop.
 5. If P^* or N^* (or both) are candidate patterns, define $\mathcal{P}^{k+1} = \mathcal{P}^k \cup \{P^*\}$ and $\mathcal{N}^{k+1} = \mathcal{N}^k \cup \{N^*\}$ accordingly, and set $k = k + 1$.
 6. Go to Step 2.
-

Figure 1. Pseudocode of the column generation algorithm.

a new positive or negative pattern (or both) to be added to our current set of known patterns. In general, if at iteration k , we generate new patterns $P^* \in \mathbb{P}$ and $N^* \in \mathbb{N}$, then we define the sets of patterns for the next iteration as $\mathcal{P}^{k+1} \leftarrow \mathcal{P}^k \cup \{P^*\}$ and $\mathcal{N}^{k+1} \leftarrow \mathcal{N}^k \cup \{N^*\}$.

COLUMN GENERATION ALGORITHM

The column generation algorithm as described above alternates the solution of problem RP with the solution of SP^+ and SP^- . While problem RP optimizes the discriminant function over the current set of patterns, problems SP verify the optimality of the discriminant function obtained by RP and, if necessary, produce new patterns to improve the current discriminant function. Thus, the iterative process stops at one of the following events: (i) the maximum number of iterations is reached; (ii) no candidate pattern can be found; (iii) no significant improvement of the objective function of RP takes place after a certain number of iterations. For the latter case, we need to define two control parameters: a tolerance value corresponding to the smallest increase in the objective function of RP to be regarded as a significant increase, and the maximum number of iterations to be allowed without a significant increase. These control parameters are important since in practice, it is possible to have long sequences of iterations in which the objective function of RP remains unchanged or changes only slightly.

Although column generation algorithms usually approach the neighborhood of a near-optimal solution in a reasonably small number of iterations [16,21], achieving or proving

optimality may take a very large number of iterations. Typically, toward the end of the algorithm's execution, the graph corresponding to the progress in the objective function value exhibits a long "tail," reflecting the failure in making significant progress during a large number of iterations. This phenomenon is known as the *tailing-off effect* [15–17,20,21]. In our experiments, we used the tolerance parameter equal to 10^{-4} and the maximum number of degenerate iterations equal to 10. The maximum number of iterations allowed for the entire execution of the algorithm was set to 1000, but was never reached in our experiments.

Figure 1 summarizes the algorithm. We use the notation $\text{RP}(k)$, $\text{SP}(k)^+$ and $\text{SP}(k)^-$ to denote the optimization problems solved at iteration k .

Computational Experiments

The algorithm proposed in this study was implemented in C++, using the *MS Visual C++ .NET 1.1* compiler. The linear programming formulations solved as part of the column generation algorithm were solved using the callable library of the Xpress-MP optimizer [24]. The computer used for carrying out the experiments reported here was an *Intel Pentium*[®] 4, 3.4 GHz, with 2 GB of RAM.

Problems SP^+ and SP^- were not necessarily solved to optimality in our experiments. Given the satisfactory performance of the truncated versions of the branch-and-bound algorithm proposed in Ref. 23, we established an explicit limit of 3000 branch-and-bound nodes per problem.

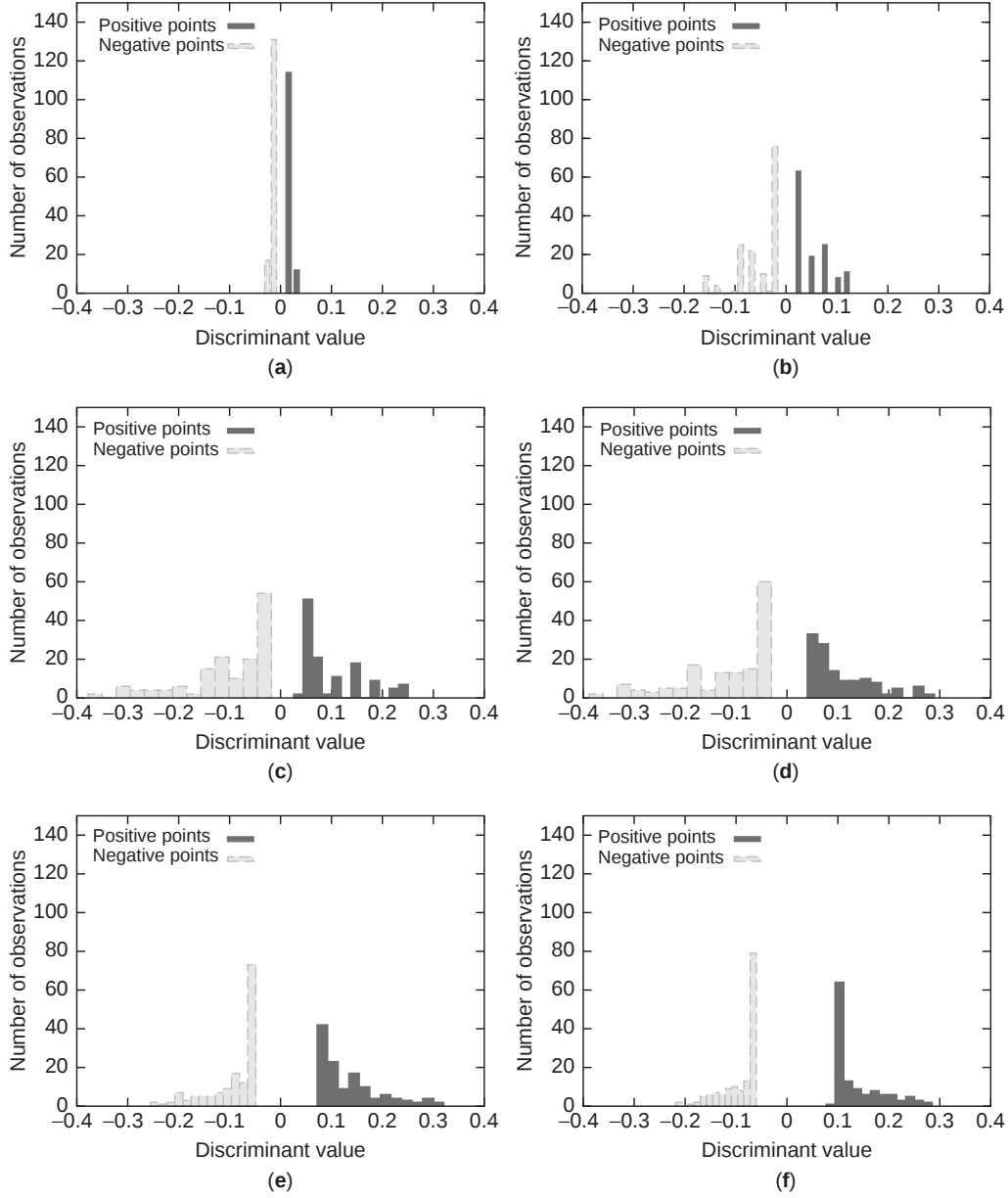


Figure 2. Histograms of observations at different discriminant levels for the “heart” data set: (a) at iteration 2, (b) at iteration 5, (c) at iteration 10, (d) at iteration 20, (e) at iteration 50, and (f) at iteration 72.

In Fig. 2, we show six histograms illustrating a typical example of how the frequencies of values of the discriminant function evolve during the execution of our algorithm. The graphs correspond to the application of our

training algorithm to a randomly extracted sample of 90% of the observations in the “heart” data set, from the UCI Repository [25]. The empty region in the middle of the graphs is the separation margin between

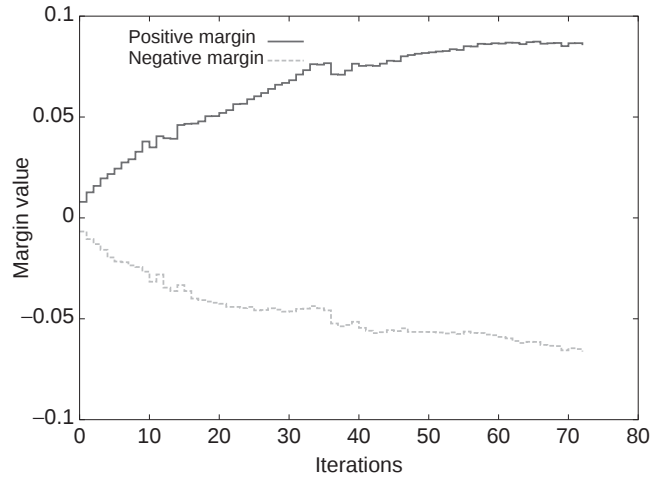


Figure 3. Behavior of the positive and negative margins for the “heart” data set.

positive and negative points. The graphs show how our algorithm iteratively improves the separation margin as it finds candidate patterns and adjusts the weights of the discriminant function accordingly.

Although all observations are correctly separated in each of the iterations, we can see that a large number of observations are on the border of the separating region in early iterations; that is, their corresponding constraints from sets (3) and (4) are binding. The observations on the border of the margin of separation can be seen as “support observations,” just as in the case of support vector machines [10,26].

In Fig. 3, we present the values of the positive and negative margins during the execution of the same experiment on the “heart” data set. It is important to notice that while the individual (positive and negative) margins do not vary monotonically with the number of iterations, the difference between the positive and the negative margin is monotonically nondecreasing with respect to the number of iterations. It can also be seen that roughly after iteration 50, the algorithm stops making significant improvements in the separation margin. This observation helped us to adjust the parameters for early termination that prevent the tailing-off phenomenon and avoid overfitting the discriminant function to the training sample.

Table 1 reports the accuracy of five algorithms implemented in the Weka package, the maximum patterns-based LAD version described in Ref. 4 (termed *CAP-LAD*), and those of LAD models built utilizing the algorithm described in this article (termed *LM-LAD*, for large margin logical analysis of data). The Weka algorithms utilized are the following: SMO, support vector machines; J48, C4.5 decision trees; RF, random forests; MP, multilayer perception; and SL, simple logistic regression.

The data sets used in Table 1 were obtained from the UCI Machine Learning Repository [25]. The reported accuracies were calculated as the average result of one random 10-fold cross-validation experiment with a 95% confidence interval (see Refs 27 and 28 for additional information on standard procedures for cross-validated experiments). In order to apply the LAD algorithms to these data sets, each training set in the cross-validation procedure was discretized [as described in Ref. 5], and the same discretization utilized for the purpose of classifying the test set. We display the highest accuracy for each data set in boldface characters. Moreover, at the bottom of the table, we display the so-called *Borda count* [29] of each algorithm. The Borda count in Table 1 is obtained by ranking the seven algorithms based on their accuracies on a

Table 1. Classification Accuracy and Borda Counts of Weka Algorithms, CAP-LAD and LM-LAD. Each Value is Reported as the Average of one 10-Fold Cross-Validation with a 95% Confidence Interval

Data set	SMO	J48	RF	MP	SL	CAP-LAD	LM-LAD
breast-w	0.965 ± 0.011	0.939 ± 0.012	0.967 ± 0.009	0.956 ± 0.012	0.963 ± 0.013	0.966 ± 0.011	0.942 ± 0.024
credit-a	0.864 ± 0.025	0.856 ± 0.031	0.882 ± 0.027	0.831 ± 0.032	0.873 ± 0.025	0.867 ± 0.025	0.815 ± 0.044
hepatitis	0.772 ± 0.084	0.652 ± 0.086	0.722 ± 0.101	0.727 ± 0.065	0.764 ± 0.090	0.779 ± 0.104	0.738 ± 0.091
krkp	0.996 ± 0.003	0.994 ± 0.003	0.992 ± 0.003	0.993 ± 0.002	0.975 ± 0.005	0.993 ± 0.002	0.962 ± 0.031
boston	0.889 ± 0.028	0.837 ± 0.045	0.875 ± 0.024	0.893 ± 0.031	0.874 ± 0.021	0.855 ± 0.029	0.840 ± 0.045
bupa	0.701 ± 0.045	0.630 ± 0.041	0.731 ± 0.046	0.643 ± 0.020	0.662 ± 0.048	0.734 ± 0.052	0.678 ± 0.034
heart	0.837 ± 0.039	0.799 ± 0.052	0.834 ± 0.051	0.815 ± 0.025	0.834 ± 0.040	0.826 ± 0.032	0.814 ± 0.033
pima	0.727 ± 0.029	0.722 ± 0.026	0.736 ± 0.030	0.726 ± 0.023	0.729 ± 0.031	0.747 ± 0.022	0.682 ± 0.023
sick	0.824 ± 0.027	0.926 ± 0.020	0.832 ± 0.023	0.852 ± 0.049	0.808 ± 0.023	0.825 ± 0.019	0.815 ± 0.041
voting	0.961 ± 0.018	0.960 ± 0.015	0.961 ± 0.016	0.944 ± 0.025	0.961 ± 0.014	0.952 ± 0.021	0.945 ± 0.025
Borda count	52	28	52	35	41	50	20

given data set. The most accurate algorithm is assigned a score of 7, the second best is assigned a score of 6, and so on until a score of 1 is assigned to the algorithm with the worst accuracy on that data set. The Borda count of an algorithm is the sum of these scores over the 10 data sets, and summarizes its comparative performance.

In Table 2, we present the counts of wins, losses, and ties for the pairwise comparison of the seven algorithms. Each pairwise comparison in Table 2 is made with respect to the 95% confidence interval presented in Table 1.

While being superior to some of the other algorithms for some data sets, the performance of the LM-LAD algorithm is, strictly speaking, generally inferior to that of other algorithms on the data sets used in our experiments. The Borda count shown in Table 1 summarizes this observation. However, the accuracies obtained by our algorithm are reasonably close to those of the other algorithms (as can be seen from Table 2). Indeed, as discussed in Ref. 28, the direct comparison of the accuracy of two algorithms in the experimental setup utilized here must be regarded cautiously, while the conclusion that the two accuracies are not statistically different is to be trusted. Moreover, if, for each data set, we compare the accuracy of LM-LAD with the highest accuracy of the other six algorithms, we can see that LM-LAD achieves on average 94.3% of the accuracy of the best algorithm.

It is important to mention that the six-type classification of algorithms used for comparison was extensively calibrated for the data sets used in our experiments. Indeed, most machine learning algorithms would require a careful calibration of their parameters in order to provide accurate cross-validation results as the ones shown in Table 1. On the other hand, the LM-LAD algorithm had its single parameter C fixed at 1 in all experiments reported in Table 1. We performed experiments with different values of the C parameter, reaching the conclusion that it had a minimal effect on the accuracy of the resulting models for the data sets under consideration.

CONCLUSIONS

We described a novel algorithm for constructing a large margin classifier consisting of the combination of LAD patterns (or rules, in a more general sense). Our algorithm is based on a linear programming optimization model, solved via the classical column generation procedure. The most important feature of our algorithm resides in the fact that it requires the calibration of a single control parameter, unlike most machine learning methods, which normally require substantial calibration. The classifiers produced by our algorithm were shown to provide accurate cross-validated results, comparable to those produced by carefully calibrated implementations of standard machine learning algorithms.

While our algorithm relies on a simple linear programming formulation, we believe that the use of different objective functions could further improve the quality of the resulting classifiers. This notion is supported by the fact that our algorithm can also be easily modified to solve regression problems, as described in Ref. 30. Also, a more conservative way of generating patterns by imposing limits on their *homogeneities* and coverages—as is traditionally the case in LAD [1,3,4]—could prove beneficial in terms of accuracy, while incurring the need for additional calibration.

Finally, we would like to remark that our algorithm can be applied in conjunction with any rule-based classification algorithm. Our column generation algorithm can be seen as a high-level procedure for creating a weighted combination of classifiers with the purpose of maximizing the margin of separation. Indeed, the pricing subproblem is the only step in our description, which is directly related to LAD. By replacing the construction of *candidate patterns* with the construction of *candidate rules*, we can use precisely the same framework for creating a classifier comprised of a weighted combination of generic rules.

Acknowledgment

The author is deeply thankful to Peter L. Hammer and Alexander Kogan for helpful discussions on a preliminary version of this

Table 2. Matrix of Wins, Losses, and Ties

	SMO	J48	RF	MP	SL	CAP-LAD
J48	4-1-5					
RF	1-1-8	4-1-5				
MP	0-4-6	2-2-6	0-2-8			
SL	1-1-8	1-2-7	0-2-8	1-2-7		
CAP-LAD	0-1-9	2-1-7	0-0-10	2-1-7	1-0-9	
LM-LAD	0-0-10	0-1-9	0-1-9	0-0-10	0-0-10	0-1-9

text. The author dedicates this manuscript to the memory of Peter L. Hammer, who invented and publicized the main concepts discussed in this text.

REFERENCES

- Alexe G, Hammer PL. Spanned patterns for the logical analysis of data. *Discrete Appl Math* 2006;154(7):1039–1049.
- Alexe G, Alexe S, Hammer PL. Pattern-based clustering and attribute analysis. *Soft Comput* 2006;10:442–452.
- Alexe S, Hammer PL. Accelerated algorithm for pattern detection in logical analysis of data. *Discrete Appl Math* 2006;154(7):1050–1063.
- Bonates TO, Hammer PL, Kogan A. Maximum patterns in datasets. *Discrete Appl Math* 2007;156(6):846–861.
- Boros E, Hammer PL, Ibaraki T, *et al.* Logical analysis of numerical data. *Math Program* 1997;79:163–190.
- Boros E, Hammer PL, Ibaraki T, *et al.* An implementation of logical analysis of data. *IEEE Trans Knowl Data Eng* 2000;12(2):292–306.
- Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: *Machine learning: Proceedings of the 12th International Conference*. Tahoe City (CA): Morgan Kaufmann Publishers, Inc.; 1995. pp. 194–202.
- Alexe G, Alexe S, Hammer PL, *et al.* Comprehensive vs. comprehensible classifiers in logical analysis of data. *Discrete Appl Math* 2008;156(6):870–882.
- Shawe-Taylor J, Cristianini N. *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press; 2004.
- Smola AJ, Bartlett P, Scholkopf B, *et al.* *Advances in large-margin classifiers*. Cambridge (MA): The MIT Press; 2000.
- Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. *Proceedings of the fifth Annual Workshop on Computational Learning Theory*; Pittsburgh (PA): 1992. pp. 144–152.
- Krauth W, Mezard M. Learning algorithms with optimal stability in neural networks. *J Phys A Math General* 1987;20:L745–L752.
- Schapire RE, Freund Y, Bartlett P, *et al.* Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann Stat* 1998;26(5):1651–1686.
- Gilmore PC, Gomory RE. A linear programming approach to the cutting-stock problem. *Oper Res* 1961;9(6):849–859.
- Gilmore PC, Gomory RE. A linear programming approach to the cutting stock problem-Part II. *Oper Res* 1963;11(6):863–888.
- Lubbecke ME, Desrosiers J. Selected topics in column generation. *Oper Res* 2005;53(6):1007–1023.
- Wilhelm WE. A technical review of column generation in integer programming. *Optim Eng* 2001;2(2):159–200.
- Eckstein J, Hammer PL, Liu Y, *et al.* The maximum box problem and its application to data analysis. *Comput Optim Appl* 2002;23(3):285–298.
- Chvatal V. *Linear programming*. New York: Freeman; 1983.
- Lasdon LS. *Optimization theory for large systems*. New York (NY): Dover Publications Inc.; 2002.
- Nazareth JL. *Computer solution of linear programs*. New York: Oxford University Press, Inc.; 1987.
- Wolsey LA. *Integer programming*. New York: John Wiley & Sons, Inc.; 1998.
- Bonates TO, Hammer PL. A branch-and-bound algorithm for a family of pseudo-Boolean optimization problems. Technical Report RRR 21–2005. RUTCOR - Rutgers

- Center for Operations Research, Rutgers University; 2007.
24. Dash Associates. Xpress-Mosel reference manuals and Xpress-optimizer reference manual, release 2004G. 2004.
 25. Newman DJ, Hettich S, Blake CJ, *et al.* UCI Repository of machine learning databases. 2007. Available at <http://www.ics.uci.edu/mllearn/MLRepository.html>.
 26. Scholkopf B, Smola AJ. Learning with kernels. Cambridge (MA): MIT Press; 2002.
 27. Lim T-S, Loh W-Y, Shih Y-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach Learn J* 2000;40:203–228.
 28. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998;10(7):1895–1924.
 29. Borda JC. Memoire sur les elections au scrutin. *Hist Acad Roy des Sci* 1781;2.
 30. Bonates TO, Hammer PL. Pseudo-Boolean regression. Technical Report RRR 3–2007, RUTCOR - Rutgers Center for Operations Research, Rutgers University; 2007.

LATIN–IBERO–AMERICAN ASSOCIATION FOR OPERATIONAL RESEARCH

HÉCTOR CANCELA

Departamento de Investigación Operativa,
Instituto de Computación, Facultad de
Ingeniería, Universidad de la República,
Montevideo, Uruguay

The Latin–Ibero–American Association for Operational Research, usually known as *Asociación Latino–Ibero–Americana de Investigación Operativa* (ALIO) is the association of the national operational research societies in Latin America and the Iberian Peninsula.

The main objectives of ALIO are to promote the progress and application of operational research (OR) in the region, and to support the networking and free exchange of information among academicians, professionals, and decision makers in the region. Also, ALIO plays an important role as the regional grouping for the Latin–American region within IFORS, the International Federation of Operational Research Societies.

ALIO was born in 1982, after an initiative of Nelson Maculan (Brazil), Hugo Scolnik (Argentina), and Andrés Weintraub (Chile), who, meeting at an event in Rio de Janeiro in 1981, foresaw the need to create an organization which would appeal to all researchers working in the OR field in different countries in the region, so that they could work together to strengthen the community and make it grow. In 1982, the association was officially founded, and its initial Executive Committee elected, at the occasion of the first CLAIO, Latin–Ibero–American Operational Research Conference, which has since been organized biennially.

The list of past presidents of ALIO and their terms of office are as follows:

- 1982–1984 Roberto Galvão (Brazil);
- 1984–1986 Héctor Monteverde (Argentina);

- 1986–1988 Andres Weintraub (Chile);
- 1988–1990 Nelson Maculan (Brazil);
- 1990–1992 Hugo Scolnik (Argentina);
- 1992–1994 Celso C. Ribeiro (Brazil);
- 1994–1996 Maximo Bosch (Chile);
- 1996–1998 Irene Loiseau (Argentina);
- 1998–2000 Nelson Maculan (Brazil);
- 2000–2002 Lorena Pradenas (Chile);
- 2002–2004 Horacio H. Yanasse (Brazil);
- 2004–2006 Sira Allende (Cuba);
- 2006–2010 Héctor Cancela (Uruguay).

The main events organized by ALIO are the biennial CLAIO conferences, which have been hosted at the following cities: 1982—Rio de Janeiro, Brazil; 1984—Buenos Aires, Argentina; 1986—Santiago de Chile, Chile; 1988—Rio de Janeiro, Brazil; 1990—Buenos Aires, Argentina; 1992—Ciudad de México, México; 1994—Santiago de Chile, Chile; 1996—Rio de Janeiro, Brazil; 1998—Buenos Aires, Argentina; 2000—México D.F., México; 2002—Concepción, Chile; 2004—La Habana, Cuba; 2006—Montevideo, Uruguay; 2008—Cartagena de Indias, Colombia. Next CLAIO will be held in Buenos Aires, Argentina, in June 2010, together with a joint ALIO–INFORMS international meeting. The CLAIO series has been extremely important to promote regional and world wide synergy and cooperation between researchers of different countries; a measure of its success is the number of participants, which has grown from 12 persons (in the first conference, in 1982) to more than five hundred in the 2008 conference.

Other conferences coorganized by ALIO include the ALIO–EURO Workshop on Applied Combinatorial Optimization Series. The first workshop took place in 1989 at Rio de Janeiro, Brazil; the second one in 1996 at Valparaiso, Chile, 1996; and the following ones (held triennially, and alternating between a European and a Latin–American location) were held at Erice,

Italy (1999); Pucon, Chile (2002); Paris, France (2005); and Buenos Aires, Argentina (2008). A recent joint undertaking by ALIO, INFORMS, and IFORS was the launching of the biennial ALIO/INFORMS/IFORS Teaching Effectiveness Colloquium Workshop series, which has been held in Montevideo, Uruguay (2006) and Cartagena, Colombia (2008).

Between 1986 and 2000, ALIO published a scholarly journal, *Investigación Operativa*, ISSN 1014-8264. This was a venue for publishing high quality research (written in any Latin language or in English). One of the goals of *Investigación Operativa* was to promote the integration, publication, and dissemination of work in OR conducted within the Latin-Ibero-American community. However, the journal also welcomed original, relevant works coming from any other country. The journal contents were selected so as to maintain equilibrium between theoretical and practical work in OR, in all fields of human activity. The journal editors were Nelson Maculan, Celso Carneiro Ribeiro, and José Mario Martínez. Although academically successful, financial issues made it very difficult to continue operations of the journal; after negotiations with IFORS, in 2000, the journal was merged within IFORS flagship journal, *ITOR*, ISSN 0969-6016. *ITOR*'s current general editor is Celso Carneiro Ribeiro, who has taken advantage of his experience as one of the former editors of *Investigación Operativa* to

lead a very successful improvement process in that journal.

Another noteworthy initiative by ALIO is the ELAVIO series, organized since 1994. The ELAVIOs are the Latin-American Operational Research Summer Schools, whose objective is to bring together the graduate students from all countries in the region and give them the opportunity to assist with the short courses, tutorials, discussion panels, and lectures on advanced topics in OR given by senior researchers and scholars. The full list of the ELAVIOs held include Punta de Tralca, Chile (1994); Mendes, Brazil (1995); Bariloche, Argentina (1996); Montevideo, Uruguay (1997); Viña del Mar, Chile (1998); Mendes, Brazil (1999); Habana, Cuba (2000); Viña del Mar, Chile (2001); Vaquerías, Argentina (2003); Montevideo, Uruguay (2004); Villa de Leyva, Colombia (2005); Itaipava, Brazil (2007); Chosica, Peru (2008); and El Fuerte, Mexico (2009). The ELAVIO schools are organized with the sponsorship of IFORS, which makes it possible to exchange students with EURO and CORS, as well as support participants from developing countries. As part of this same agreement, young scholars from ALIO regularly participate in the EURO Summer Institutes (ESIs).

ALIO maintains a homepage at <http://www.dc.uba.ar/alio>, including information about its member societies, current authorities, organized events, and other items of interest to the OR community.

LEARNING WITH DYNAMIC PROGRAMMING

PETER I. FRAZIER
School of Operations Research
and Information Engineering,
Cornell University, Ithaca,
New York

Frequently in operations research and management science we face uncertainty, whether in the form of uncertain demand for goods, uncertain travel times in networks, or uncertainty about which set of parameters best calibrates a simulation model. Often when facing uncertainty we are also offered the opportunity to collect some information that will mitigate uncertainty's effects. On the one hand, collecting information allows us to make better decisions and produce better outcomes. On the other hand, collecting information carries a cost, whether in time or money spent, or in the opportunity cost sacrificed collecting one piece of information when we could have collected another. How can we optimally balance these costs and benefits?

One way to formulate such problems is with Bayesian methods [1], in which we begin with a *prior* probability distribution that describes our subjective belief about how likely each of the many different possible truths are. Dynamic programming (DP) [2] offers a general way to formulate and solve such Bayesian sequential learning problems. This DP formulation provides value in three ways. First, in some cases, the DP formulation allows computing the optimal solution. This is usually only true for smaller problems (e.g., the inventory problem considered in Ref. 3), because the curse of dimensionality [4] prevents computing the solution to larger problems in a practically feasible amount of time. Second, the DP formulation sometimes allows structural results to be shown theoretically, which either provides intuition about the problem and the behavior of the optimal policy as in, for example, Ding *et al.* [5],

or provides a characterization of the optimal policy that may be directly exploited to find an optimal policy, as in sequential hypothesis testing [6]. Third and finally, the DP formulation sometimes suggests useful heuristics (e.g., knowledge-gradient methods [7]) for complex and large-scale learning problems.

The sequential learning problems that we consider here have been studied in several separate communities. Within the operations research community, sequential learning problems, as well as the way in which DP can be used to confront them, were recognized in early work by Howard [8,9]. Within the resulting literature, surveyed in Refs 10 and 11, such sequential learning problems were called *partially observable Markov decision processes* (POMDP). This work on POMDPs was continued in the artificial intelligence and reinforcement learning communities [12,13], where the emphasis is on general purpose complexity results and algorithmic strategies that apply to the whole spectrum of POMDPs. Specific applications tend to come from robotics, game playing, and other problems that are thought to be exemplars for the problems that humans face when interacting in a general way with their environment. More recently, these problems have also been called *Bayes adaptive Markov decision processes* and optimal learning problems [14,15]. Within statistics, a closely related field is sequential design of experiments [16–18]. Here, the emphasis is on rigorous treatment of a class of problems that tend to appear in laboratory and other experimental settings. This field is also closely related to sequential analysis [19]. While much of the literature is written for a specialized research audience, the tutorial [15] introduces this class of problems in an accessible and comprehensive way for the general operations research (OR) audience.

The problems that we consider here are all *sequential*, by which we mean that data is processed as it is collected, and decisions are made on the basis of all available data. Acting sequentially allows adaptation and provides

the possibility of more efficient action. This is in contrast to nonsequential problems, where decisions are made before observing any data at all. The terms “on-line” and “offline” are sometimes used instead of “sequential” and “nonsequential,” although these terms can also refer to differences in reward structure (see the section titled “Multiarmed Bandits”). Sequential methods require a more sophisticated analysis than nonsequential methods, but often provide much better performance.

We begin by describing in detail one sequential learning problem, the Bernoulli ranking and selection (R&S) problem. This problem is simple enough to be described in detail here and to be solved explicitly using DP, but it also contains essential elements observed in all sequential learning problems. Thus, it serves as an excellent example of this class of problems and the way in which DP can be applied toward their solution. We then expand our scope to describe several other problems from operations research and management science to which DP has been fruitfully applied.

THE RANKING AND SELECTION PROBLEM

In this section, we describe an important learning problem in detail, and show how DP may be used to solve it. This problem is the R&S problem, which has a long history beginning with the work of Bechhofer [20,21]. Historically, this problem was most often considered in a non-Bayesian framework (see Ref. 22 for a review), but the Bayesian formulation has grown more prominent in recent decades (see Ref. 23 for a review).

In the R&S problem, we consider the efficient use of experimentation (either simulation based or physical) to determine which of several “alternatives” is best. As an example, suppose that we have several different inventory policies, and we would like to use computer simulation to determine which of these has the highest average profit. We are limited in the number of simulations we can perform by the amount of simulation time that we have available, and we would like to allocate this time to simulations of the different inventory policies. This allocation should

maximize our ability to determine which of the inventory policies is best.

One approach is to run a few simulations of each alternative to roughly estimate which alternatives are likely to be among the best, and then concentrate most of our later simulation effort on just these alternatives. When done intelligently, this allows us to find the best alternative with fewer samples than we could have by spreading our simulation budget equally across the alternatives. The essential question in R&S is how this allocation may be done as efficiently as possible. DP offers an answer.

We formulate the R&S problem formally by supposing that we have a limited budget N of samples to be allocated among k alternatives, and a sampling distribution is associated with each alternative. The R&S literature often assumes that this sampling distribution is normal, but it will be easier to illustrate this problem if we assume that the sampling distribution is Bernoulli. Let θ_x be the parameter of the Bernoulli distribution for alternative x , so θ_x is the probability of observing a “success” from alternative x . This quantity θ_x is unknown, but we will be able to learn it through sampling. Our goal is to use sampling to find the alternative with the largest θ_x .

Time is indexed by n , from $n = 1$ up to $n = N$. At each time $n \leq N$, we choose an alternative $x_n \in \{1, \dots, k\}$ to sample from among the full set of k alternatives. This choice may depend upon all previously observed samples, $(x_m, y_m)_{m < n}$, where x_1 is chosen deterministically. We then observe a sample y_n whose distribution is

$$y_n \mid \theta, x_n \sim \text{Bernoulli}(\theta_{x_n}).$$

This sample is conditionally independent of all other x_m and y_m , $m < n$, given θ and x_n . At the final time N , we select an alternative x_* , and we receive a terminal reward θ_{x_*} . This is the reward that we receive when we implement the alternative x_* in the real world. The decision x_* may depend upon all previously observed samples, $(x_m, y_m)_{m \leq N}$. Given perfect information, we would choose x_* to be equal to $\arg \max_x \theta_x$, but this information is unavailable. Instead, we must base our choice of

x_* upon the sampling information acquired. The crux of the R&S problem is choosing the sampling decisions $x_1, \dots, x_N$ in order to best support our final implementation choice x_* .

Formally, we define a policy π to be a collection of random variables $(x_n)_{1 \leq n \leq N}$ and x_* . We define F_n to be the sigma-algebra generated by the random variables x_m, y_m , $m \leq n$. We enforce the way in which decisions may depend upon previous observations by requiring that $x_n \in \mathcal{F}_n - 1$ for $n \geq 1$ and $x_* \in \mathcal{F}_N$. We define Π to be the set of all such policies π .

Momentarily fixing θ , the expected reward obtained by a policy π is

$$\mathbb{E}^\pi [\theta_{x_*} | \theta], \quad (1)$$

where $\mathbb{E}^\pi$ indicates the expectation taken with respect to policy π . When an expectation depends upon the policy, we write $\mathbb{E}^\pi$, and when it does not we simply write $\mathbb{E}$. Although the only decision that appears in this expression is x_* , the information upon which x_* is based depends critically upon the sampling decisions x_n . Indeed, we will see that choosing x_* given the available information is relatively straightforward, and the difficulty in the R&S problem is choosing x_n , $n \leq N$, i.e., choosing which information to collect.

We would like to choose the policy π that maximizes this quantity (Eq. (1)). The difficulty is that Equation (1) depends upon θ , which is unknown. If we knew θ , we could simply choose $x_* \in \arg \max_x \theta_x$ without sampling at all. Furthermore, different policies may be better for different values of θ . We need to combine the criteria (Eq. (1)) across multiple values of θ and find a policy π that does well according to this combined criterion. One method for combining Equation (1) across multiple values of θ is the Bayesian method. In this method, we suppose that we have a prior probability measure $\mathbb{P}_0$ on θ , and that we wish to maximize the expected reward received when θ is drawn at random according to $\mathbb{P}_0$. This expected reward is $\mathbb{E}[\mathbb{E}^\pi [\theta_{x_*} | \theta] | \theta \sim \mathbb{P}_0] = \mathbb{E}^\pi [\theta_{x_*}]$, and our goal in Bayesian R&S is to find a policy π that solves

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi [\theta_{x_*}]. \quad (2)$$

The prior probability measure $\mathbb{P}_0$ may be interpreted in several ways. One may think of it as expressing our subjective belief about the relative importance of various possible configurations. This first interpretation is the one most commonly understood in Bayesian statistics. Secondly, the prior may be interpreted literally in certain limited cases where θ really has been drawn at random from a known probability distribution. This is the case when one is solving R&S problems on a repeated basis, as one might, for example, when calibrating a single simulation model each day to a new set of data. Third and finally, one may interpret the objective function $\mathbb{E}^\pi [\theta_{x_*}] = \int_{\mathbb{R}^k} \mathbb{E}^\pi [\theta_{x_*} | \theta] \mathbb{P}_0(d\theta)$ as a linear combination of the individual objectives $(\mathbb{E}^\pi [\theta_{x_*} | \theta])_{\theta \in \mathbb{R}^k}$, just as we often treat multi-objective optimization problems by optimizing a linear combination of the objectives. In this interpretation, the choice of $\mathbb{P}_0$ is simply the choice of which linear combination to optimize.

Given a prior distribution $\mathbb{P}_0$ and the information collected up to a time n , we have a posterior distribution $\mathbb{P}_n$. In particular, given a sequence of measurements $x_{1:n} = x_1, \dots, x_n$ and observations $y_{1:n} = y_1, \dots, y_n$, the posterior $\mathbb{P}_n$ is defined as the measure on $\mathbb{R}^k$, $\mathbb{P}_n\{\theta \in du\} = \mathbb{P}_0\{\theta \in du | x_{1:n}, y_{1:n}\}$. We write $\mathbb{E}_n$ to indicate the expectation taken with respect to this posterior distribution. In general, the posterior may be calculated using Bayes' rule,

$$\mathbb{P}_n\{\theta \in du\} \propto \mathbb{P}\{y_{1:n} | x_{1:n}, \theta = u\} \mathbb{P}_0\{\theta \in du\}. \quad (3)$$

We assume that the decisions x_n depend only upon the posterior. One can show that the supremum (Eq. (2)) is unchanged if Π is restricted to this class of policies.

To facilitate computation of the posterior, it is convenient to suppose that the prior has a certain parametric form. If θ_x has a $\beta(a_{0x}, b_{0x})$ distribution under the prior $\mathbb{P}_0$ and is independent of all θ_j , $j \neq x$, then a calculation beginning with Equation (3) shows that θ_x is also beta-distributed under the posterior $\mathbb{P}_n$, and it remains independent of other θ_j . In particular, $\theta_x \sim \beta(a_{nx}, b_{nx})$, where $a_{nx} = a_{0x} + \sum_{m \leq n} 1_{\{x_m = x\}}(1 - y_m)$ is the sum of a_{0x} and the number of

failures seen from alternative x , and $b_{nx} = b_{0x} + \sum_{m \leq n} 1_{\{x_m = x\}} y_m$ is the sum of b_{0x} and the number of successes seen from alternative x . We define

$$\mu_{nx} = \mathbb{E}_n[\theta_x] = b_{nx}/(a_{nx} + b_{nx})$$

to be the expectation of θ_x under the posterior at time n . The explicit expression $b_{nx}/(a_{nx} + b_{nx})$ follows from a standard computation of the expectation of a beta-distributed random variable [24].

Given this beta prior and its associated beta posteriors, we can simplify the objective function $\mathbb{E}^\pi[\theta_{x_*}]$. Using the tower property of conditional expectation and the $\mathcal{F}_N$ -measurability of x_* , we write $\mathbb{E}^\pi[\theta_{x_*}] = \mathbb{E}^\pi[\mathbb{E}_N^\pi[\theta_{x_*}]] = \mathbb{E}^\pi[\mu_{Nx_*}]$. Thus, the stochastic control problem (2) may be rewritten

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi[\theta_{x_*}] = \sup_{\pi \in \Pi} \mathbb{E}^\pi[\mu_{Nx_*}].$$

This problem can be solved using DP. Because $\mu_{Nx} = b_{Nx}/(a_{Nx} + b_{Nx})$ is a function of a_N and b_N , and because $(n, a_n, b_n)_n$ is a Markov process (we see this below in greater detail), the state of the DP may be taken to be (n, a_n, b_n) . We thus define a value function parameterized by n, a_n , and b_n

$$V(n, a_n, b_n) = \sup_{\pi} \mathbb{E}^\pi[\mu_{Nx_*} | a_n, b_n]. \quad (4)$$

At the final time $n = N$, the only decision left to make is x_* . This decision may depend upon all the information available at this time, and, in particular, it may depend upon μ_{Nx} , $x = 1, \dots, k$. Thus, the optimal x_* is any satisfying $x_* \in \arg\max_x \mu_{Nx}$, and the value function at the final time is

$$V(N, a_N, b_N) = \max_x \mu_{Nx}. \quad (5)$$

Given this final value function, we may compute the value function at earlier times $n < N$ using Bellman's recursion,

$$\begin{aligned} V(n, a_n, b_n) &= \max_x \mathbb{E}[V(n+1, a_{n+1}, b_{n+1}) | x_{n+1} \\ &= x, a_n, b_n] \quad \text{for } n < N. \end{aligned} \quad (6)$$

To write this expectation explicitly, we must consider the distribution of a_{n+1} and b_{n+1} given $x_n = x$, a_n , and b_n . If we measure alternative x and observe a success ($y_{n+1} = 1$), then we increment b_{nx} . Instead, if we observe a failure ($y_{n+1} = 0$), then we increment a_{nx} . Thus, letting e_x be the vector of 0s with a single 1 at component x ,

$$\begin{aligned} b_{n+1} &= b_n + y_{n+1}e_x, & a_{n+1} &= a_n \\ & & & + (1 - y_{n+1})e_x, \end{aligned}$$

and the distribution of a_{n+1} and b_{n+1} is determined by the probability of success, $\mathbb{P}_n y_{n+1} = 1$. Given θ_x , we have $y_{n+1} = 1$ with probability θ_x , but θ_x is unknown. We do have a posterior distribution on θ_x , however, which allows us to write

$$\begin{aligned} \mathbb{P}_n \mathbb{E}\{y_{n+1} = 1\} &= \mathbb{E}_n[y_{n+1}] = \mathbb{E}_n[\mathbb{E}_n[y_{n+1} | \theta]] \\ &= \mathbb{E}_n[\theta_{x_{n+1}}] = \mu_{nx_{n+1}}. \end{aligned}$$

We then immediately have the following expression for the distribution of a_{n+1} , b_{n+1}

$$\begin{aligned} \mathbb{P}_n\{b_{n+1} &= b_n + e_x, a_{n+1} = a_n | x_{n+1} = x\} \\ &= \mathbb{P}_n\{y_{n+1} = 1 | x_{n+1} = x\} = \mu_{nx} \\ \mathbb{P}_n\{a_{n+1} &= a_n + e_x, b_{n+1} = b_n | x_{n+1} = x\} \\ &= \mathbb{P}_n\{y_{n+1} = 0 | x_{n+1} = x\} = 1 - \mu_{nx}. \end{aligned}$$

Thus, $(n, a_n, b_n)_{0 \leq n \leq N}$ is a Markov process whose distribution is determined by the measurement policy π . Note that we do not need to include θ into this stochastic process in order for its dynamics to be well defined. This fact is mathematically quite natural but may seem counterintuitive. The Bayesian assumption of a prior on θ has allowed us to define the way in which our posterior belief on θ changes with measurements, without the need to know θ . This is what allows us to formulate the problem as a DP.

Given the conditional distribution of a_{n+1} , b_{n+1} , we can simplify Bellman's recursion

(6) to

$$\begin{aligned}
 & V(n, a_n, b_n) \\
 &= \max_x V(n+1, a_n + e_x, b_n)(1 - \mu_{nx}) \\
 &\quad + V(n+1, a_n, b_n + e_x)\mu_{nx}, \quad \text{for } n < N.
 \end{aligned} \tag{7}$$

Using the terminal condition (5) and Bellman's recursion (7), one can calculate the value function for all possible states n, a_n , and b_n . There are finitely many such states since $a_n - a_0$ and $b_n - b_0$ have nonnegative integer components whose sum must equal n . The value function for $n = 0$ and $a_{0x} = b_{0x} = 1$ for all x is pictured in Fig. 1 for problems with various values of k and N .

Given the value function as computed in this manner, we may compute an optimal policy as one that chooses its decisions to achieve the maximum in Equation (7). That is, an optimal policy π^* is defined by

$$\begin{aligned}
 x_n \in \arg \max_x & V(n+1, a_n + e_x, b_n)(1 - \mu_{nx}) \\
 & + V(n+1, a_n, b_n + e_x)\mu_{nx}.
 \end{aligned}$$

Although computing the value function using Bellman's recursion is straightforward in theory, it is quite difficult practically when the number of states is very large. In our problem, the number of states can be

reduced by removing n from the state, since it is determined by summing the components of a_n and b_n . It can be further reduced by noting the invariance of the value function to permutations of the alternatives. Even with both space-reducing methods, however, the number of states is very large for even moderate values of k and N . This is the so-called "curse of dimensionality," and presents a great challenge to computation. Several approximation techniques have been proposed for addressing this difficulty [4], but the curse of dimensionality still presents a fundamental challenge to DP.

Despite the challenge presented by the curse of dimensionality, the DP solution to the Bernoulli R&S problem provides a number of benefits. First, it provides explicit solutions for small and moderately sized problems. Second, it provides a testing ground on which heuristic algorithms for larger problems may be benchmarked against optimal solutions. Third, it provides a solid theoretical understanding of the problem that may be extended to provide a number of insights into the nature of the problem and the behavior of the value function and associated optimal solution and (see Ref. 25 for similar insights in a sequential normal R&S problem). Software for computing several of the leading algorithms for Bayesian R&S of normal populations is available in the InfoCollection library at 26.

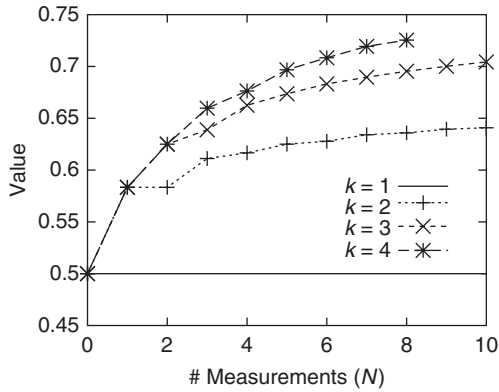


Figure 1. Value function at $n = 0$, with $a_{0x} = b_{0x} = 1$ for Bernoulli R&S as a function of the number of alternatives (k) and the measurement budget (N).

FURTHER APPLICATIONS

In the previous section, we saw how DP can be used to solve one particular sequential learning problem, the fully sequential Bernoulli R&S problem. This DP approach can be applied to nearly any learning problem in which there is some unknown truth (θ in the R&S problem) about which we are learning through our actions while simultaneously collecting rewards whose distribution depends on our actions and the truth. To formulate such problems as DPs, we create a state space consisting of the space of posterior beliefs, and the value function is defined using Bellman's recursion and the predictive distribution for the observation that follows from the current posterior distribution.

In this section, we briefly describe other problems from the literature in which the DP formulation has been useful. In some of these problems, the DP has been solved explicitly. In others, the DP has suggested useful heuristics or provided insights into the structure of the problem. In addition to the problems detailed below, many other sequential learning problems have been approached using DP. These include sequential change detection [27,28], dynamic pricing [29], medical diagnosis [30], and many others (see Ref. 7 Section 1.1 for a list).

Multiarmed Bandits

The multiarmed bandit problem is similar to the R&S problem, except that we receive each y_n as a *reward* rather than merely as an observation. The goal is to maximize the expected discounted total reward received over time

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi \left[\sum_{n=1}^N \gamma^{n-1} y_n \right],$$

where $\gamma \leq 1$ is a discount factor and $N \leq \infty$.

The name “multiarmed bandit” originates in a colorful example in which each alternative represents one arm of a multiarmed slot-machine (slot machines are sometimes called “one-armed bandits”). Pulling arm $x_n \in 1, \dots, k$ results in a payoff whose distribution is fixed but unknown. These distributions can only be learned through experimentation. The goal in this problem is to maximize the expected total value of our earnings over time.

This playful example is a surrogate for a large number of problems in many different fields in which the only way to learn the quality of some action is to actually perform it in the real-world and suffer its corresponding real loss or reward. The optimal behavior in such problems is necessarily different and more cautious than that in R&S, in which one has an experimental test bed available and which does not suffer extreme consequences for testing poor alternatives. This difference in behavior due to the reward structure is sometimes discussed by calling the reward

structures similar to R&S “offline,” and calling those similar to multiarmed bandits “online.” This difference should not be conflated with the difference discussed in the introduction between sequential policies, which make their decisions x_n based upon all of the available data, and nonsequential policies, which make their decisions before seeing any data.

The multiarmed bandit problem can be solved using DP if we assume rewards are conditionally independent given the alternative measured, and if we take the infinite horizon discounted case ($\gamma < 1$ and $N = \infty$). This solution is known as the *Gittins index policy*, and was introduced in Ref. 31. This policy is derived by considering a number of smaller problems, one for each alternative. In each smaller problem, we may pull only a single arm x , and we decide adaptively when to stop. Upon stopping, we are given a fixed retirement reward M . This can be modeled by requiring the policy to satisfy $x_n \in 0, x$ for all n , and defining $\tau = \inf n \geq 0 : x_n = 0$ as the time at which the policy decides to retire. Let Π_x be the space of all such policies. The smaller problem then has a value function V parameterized by the arm x being pulled and the retirement reward M

$$V(P; M, x) = \sup_{\pi \in \Pi_x} \mathbb{E}_n^\pi \left[\sum_{n=1}^{\tau-1} \gamma^{n-1} y_n + \gamma^{\tau-1} M \mid \mathbb{P}_0 = P \right],$$

where P is a probability distribution on the sampling distribution from arm x . In many important cases, it can be parameterized with a few (often one or two) real numbers. We then define a *Gittins index*, $m(P, x)$, to be the smallest retirement reward M such that the optimal policy restricted to pull arm x retires immediately. This is written as

$$m(P, x) = \inf \{ M : V(P; M, x) = M \}.$$

One can then show using DP that the optimal policy for the original multiarmed problem is to pull the arm with the largest Gittins index. One can refer to the original proof in Ref. 31, but more accessible and equally

rigorous proofs may be found in Refs 32 or 33. The resulting optimal policy is

$$x_n \in \arg \max_x m(\mathbb{P}_n, x).$$

To use this optimal policy, we must be able to compute $m(\mathbb{P}_n, x)$, which requires solving and calculating value functions for the smaller problems. This is possible because the state space of this smaller problem is *much* smaller than that for the full multiarmed bandit problem, consisting only of the posterior distribution on a single alternative. In many cases, including Bernoulli rewards with a beta prior on the probability of reward, and normal rewards with known variance and a normal prior on the sampling mean, this state space has only two dimensions. Such small dynamic programs can be solved numerically using standard techniques. Tables of Gittins indices for the beta-Bernoulli and normal-normal cases can be found in Ref. 34, and an easy-to-use analytic approximation for the normal case may be found in Ref. 35. This is in contrast to the full problem under these priors, whose state space has $2k$ dimensions. Computation scales exponentially in the dimension, which makes the smaller problem eminently solvable, while a brute-force numerical solution of the full problem is intractable.

If we relax the assumptions of the Gittins index policy, for example, by taking a finite horizon or correlated y_n , then no tractable optimal solution is known. The proof of the Gittins index policy relies on a decomposition of the problem into single-arm subproblems, which is quite fragile to such changes in assumptions. To solve the problem optimally, one must solve the full DP, which is computationally intractable because of its large state space. Although the optimal policy is unknown, knowledge of the underlying DP formulation and the Gittins index policy has supported the creation and analysis of a number of heuristics. Many of these heuristics are index-based solutions that are not optimal, but have pleasing empirical and theoretical properties. See, for example, heuristics for restless bandit problems [36] and other variants discussed in Section 3

of Ref. 37. DP-motivated heuristics for bandits with finite-horizons and/or correlated rewards are discussed in Refs 38 and 39.

For complete treatments of the classic Bayesian bandit results, see Refs 16,18, or 37. See also the surveys [40] from the economics literature, [41] from the computer science literature, or Chapter 10 of Ref. 4 from the operations research literature. Literature on non-Bayesian formulations of the multiarmed bandit problem in which worst-case regret is minimized is also available. For an introduction, see Chapter 6 of Ref. 42.

Bayesian Global Optimization

Suppose that we have a continuous function f that we would like to optimize over some compact domain $\mathcal{X} \subset \mathbb{R}^d$. The only access to the function is through a black box that tells us the function's value f at points x_n of our choosing, but possibly obscured by noise. We suppose that we have no access to derivative or other information. We further suppose that function evaluation takes a long time (minutes or hours), and so we wish to minimize the number of function evaluations we perform in order to find a "good" solution. Such problems occur quite frequently, for example, when optimizing a complex simulation model, or optimizing temperature and pressure in an industrial chemical-manufacturing process. Time-consuming function evaluation differentiates the problem from other derivative-free global optimization problems for which the goal may be to minimize algorithmic computation time rather than the number of function evaluations, for which algorithms requiring less computation may be desirable.

We may formally pose the problem of finding an algorithm that finds as good a solution as possible in a bounded number of function evaluations as a Bayesian learning problem. This formulation shares substantially with the aforementioned ranking and selection problem, with the critical difference being that we are now searching on a continuous domain for the best point rather than among a finite set of alternatives.

As in the R&S problem, we begin with a prior distribution on the truth, but here the prior is on the function f . Despite the large and complex nature of the space of all

possible continuous functions on a compact domain, the class of Gaussian process priors forms an analytically convenient class of prior distributions on this space. They have been successfully applied to problems in spatial statistics [43] and machine learning [44]. A Gaussian process prior is parameterized by its mean function $\mu : \mathbb{R}^d \mapsto \mathbb{R}$ and its covariance function $\Sigma : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ (Σ is required to be a positive semidefinite function), and may be written $f \sim \text{GP}(\mu, \Sigma)$. With this prior, the problem is stated as

$$\sup_{\pi} \mathbb{E}^{\pi} [f(x_*)], \quad (8)$$

where as in R&S x_* is chosen as well as possible based on the measurements (x_n, y_n) , $n = 1, \dots, N$.

The DP formulation of this problem then proceeds as it did in the R&S problem, but we now have a posterior on the space of functions. One finite-dimensional parameterization of this posterior is simply the set of points and functions evaluations observed so far $(x_m, y_m)_{m \leq n}$. The state space is then quite large, being continuous with a number of dimensions that grows with N . One may also discretize the domain into k points and work with the posterior only on these points. The state space then may be chosen to have a fixed size, but its dimension is still on the order of k^2 . Under either parameterization, the high-dimensionality of the state space prevents computation in all but the simplest cases.

Despite the difficulty in solving the DP, several heuristics that work well empirically may be understood as DP-based approximations to the optimal policy. One such heuristic is the knowledge-gradient (KG) method, which defines

$$\text{KG}_n(x) = \mathbb{E}_n \left[\sup_{x' \in \mathcal{X}} \mu_{n+1}(x') \mid x_{n+1} = x \right],$$

where $\mu_{n+1}(x') = \mathbb{E}_{n+1} f(x')$. The KG method then selects $x_n \in \arg \max_x \text{KG}_n(x)$ as the next point to measure. This is the decision that would be optimal under the DP if we had only a single measurement left to make, that is, if $n = N - 1$. In this sense, the KG method is a one-step, lookahead policy. In

general, computing the KG policy is a difficult computational problem. One approach is to discretize the domain, as in Ref. 45. For truly continuous problems, discretization is an approximation, while for problems with naturally integral decision variables interpolated by a continuous mean function, for example, choosing the reorder point for an inventory policy, this is a natural modification of the BGO problem.

Another heuristic is the expected improvement (EI) method, which is most commonly formulated in the noise-free case. Although this method is often presented without describing the underlying stochastic optimization problem (8) and its DP framework, it can be understood as being motivated by the DP. We consider the noise-free case and restrict our implementation decision to be from among those points that we have already measured. To obtain the policy that is now optimal with one measurement remaining under this restriction, we must replace the supremum over the whole domain $\mathcal{X}$ in the definition of $\text{KG}(x)$ by a maximum over only the points we have measured, $\{x_m : m \leq n + 1\}$. This maximum may be written $\max(f(x_m) : m \leq n + 1) = \max(y_{n+1}, f_n^*)$, where we have used that $\mu_{n+1}(x_m) = f(x_m) = y_m$ for $m \leq n + 1$ in the noise-free case, and then defined $f_n^* = \max_{m \leq n} f(x_m)$ to be the value of the best point observed so far. Subtracting f_n^* , which does not depend on x_{n+1} , provides the expected improvement function

$$\text{EI}(x) = \mathbb{E}_n [(y_{n+1} - f_n^*)^+ \mid x_{n+1} = x].$$

The EI method then recommends that we sample the alternative with the largest $\text{EI}(x)$. This method was introduced in Refs 46–48, building on previous work [49–51]. It was modified to allow measurements with stochastic noise in Ref. 52, and to numerical noise in Ref. 53. For a complete general introduction, see Ref. 54, or for a brief overview see Section 6.3 of Ref. 55.

Thus, both KG and EI methods may be seen as being motivated by the DP. Both are one-step lookahead policies, but under different assumptions on the set of allowed implementation decisions. The KG method

is more difficult to compute, but generalizes more naturally to noisy measurements, and was shown to perform better than EI-based methods in an empirical study [45].

A number of software packages have been written that can calculate Bayesian estimates of the function f using Gaussian process priors. One such software package is DACE (Design and Analysis of Computer Experiments) [56,57]. An up-to-date collection of other software packages is maintained at the website [58], which also lists texts, articles, and other websites relevant to estimation with Gaussian processes.

Software packages for calculating the measurement decisions of DP-based optimization policies are also available. Publicly available code associated with Forrester *et al.* [54] implement a number of prediction methods and sampling algorithms. The SPACE software package [59] implements the efficient global optimization (EGO) algorithm [48] for noise-free optimization, and the commercial software package TOMLAB includes an implementation of the sequential kriging optimization (SKO) algorithm of Ref. 52 for the noisy measurement case. An implementation of the KG algorithm [45] for discretized spaces is available in the MatlabKG package [26].

Sequential Hypothesis Testing

In the sequential hypothesis testing problem, we learn which of the two hypotheses is correct by observing a sequence of samples $y_1, y_2, \dots$. Under one hypothesis, the samples are coming from a density f_0 , and under the other hypothesis they are coming from a different density f_1 . Let $\theta \in \{0, 1\}$ be the correct hypothesis, so f_θ is the true sampling density. We pay a cost c for each sample observed and we may stop observing samples at any time. Let τ be the number of samples observed, and we require that it be a stopping time of the filtration generated by $y_1, y_2, \dots$. Upon stopping, we choose a hypothesis x_* and suffer a loss that is 0 if $\theta = x_*$ and equal to d_θ otherwise. In the Bayesian version of the problem, we begin with a prior probability that $\theta = 1$, and we seek a stopping time τ that minimizes the expected loss $\mathbb{E}[c\tau + d_\theta \mathbb{1}_{\theta \neq x_*}]$.

This problem was introduced and solved in Ref. 6, where the solution was shown to be of the form

$$\tau = \inf \{t \geq 0 : L_t \notin [A, B]\},$$

$$x_* = \mathbb{1}_{\{L_t \geq 0\}},$$

for some levels A and B , where $L_t = \sum_{s \leq t} \log(f_1(y_s)) - \log(f_0(y_s))$ is the log-likelihood at time t . A policy of this form is called a *sequential probability ratio test* (SPRT).

Wald *et al.* [6] show that the SPRT is also optimal in another, non-Bayesian, sense. For any given $\alpha_0 \in (0, 1)$ and $\alpha_1 \in (0, 1)$, the expected number of samples under both hypotheses, $\mathbb{E}[\tau | \theta = 0]$ and $\mathbb{E}[\tau | \theta = 1]$, is minimal among all policies satisfying $\mathbb{P}^\theta x_* \neq \theta | \theta = 0 \leq \alpha_0$ and $\mathbb{P}^\theta x_* \neq \theta | \theta = 1 \leq \alpha_1$. This form of optimality is surprisingly strong because it simultaneously minimizes the expected number of samples taken under both hypotheses. It is known as *Neyman–Pearson optimality* because of its similarity to Neyman–Pearson optimality for nonsequential statistical tests [60]. A modern and accessible treatment of the results of Ref. 6 may be found in Ref. 61. Also see Ref. 62 for a detailed survey of this and many related problems is sequential analysis.

The original motivation for this problem was in reducing the number of rounds of ammunition required to test antiaircraft guns and ordnance during World War II [63], but it or problems similar to it appear whenever we wish to take samples in order to make a decision. One class of similar problems of particular importance appear in the design of clinical trials [64].

Inventory Control

In the classic newsvendor problem [65], we have daily demand for a perishable product, for example, newspapers, and we wish to maximize our expected revenue. Each day we choose a number of units x_n of inventory to buy at a per-unit cost c , and we sell units at a per-unit price of p until we either run out of inventory or meet the days' demand D_n .

If the demand distribution is known, then the classic analysis easily characterizes

the optimal solution. In many problems, however, we do not know the true demand distribution but are instead learning it from sales data. This is particularly true with a new product, or with a product for which the demand distribution is changing over time. When the demand distribution is unknown but fixed, the Bayesian version of this problem may be written

$$\sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{n=0}^N -cx_n + p \min(D_n, x_n) \right]$$

where the demand distribution F is unknown and $D_n|F$ come independently and identically distributed from F . This problem may be solved using DP, where the state of the DP is a parameterization of the posterior belief on F at the current time.

If we observe all of the demand, even if we are unable to meet it, then the DP simplifies and the optimal stocking decision is the same as the greedy stocking decision, $\arg\max_x \mathbb{E}_n[-cx_n + p \min(D_n, x_n)]$. Instead, if observation of demand is censored when we stock out of product, as it often is in practice, then the solution to the DP is more complicated. Ding *et al.* [5] use the DP formulation to show that the optimal policy in the presence of censoring orders at least as much as the greedy policy, and possibly more. An easily computed solution is provided in Ref. 3 for cases with either exponential demand with unknown rate, or Weibull demand with known shape but unknown mean. In these cases, a DP with a two-dimensional state space may be solved using a one-dimensional calculation.

CONCLUSION

We have shown how DP can be used to analyze a number of different learning problems from operations research and management science: ranking and selection, multiarmed bandits, global optimization, sequential hypothesis testing, and inventory control. These are just a few of the interesting and important problems in sequential learning that can be approached

using DP. In some cases, the DP formulation provides an explicit solution, while in others it provides useful heuristics and insight into the problem and the behavior of the optimal policy. Many other important problems remain that we believe can be profitably approached with DP, and we hope that this article inspires readers to apply DP methods to these problems in the future.

REFERENCES

1. Berger JO. Statistical decision theory and Bayesian analysis. 2nd ed. New York: Springer; 1985.
2. Bellman R. The theory of dynamic programming. Bull Am Math Soc 1954;60:503–516.
3. Lariviere M, Porteus E. Stalking information: bayesian inventory management with unobserved lost sales. Manage Sci 1999;45(3):346–363.
4. Powell WB. Approximate dynamic programming: solving the curses of dimensionality. New York: John Wiley & Sons, Inc.; 2007.
5. Ding X, Puterman M, Bisi A. The censored newsvendor and the optimal acquisition of information. Oper Res 2002;50(3):517–527.
6. Wald A, Wolfowitz J. Optimum character of the sequential probability ratio test. Ann Math Stat 1948;19(3):326–339.
7. Frazier P. Knowledge-gradient methods for statistical learning [PhD thesis]. Princeton University; 2009.
8. Howard R. Dynamic programming and Markov processes. MIT press; 1960.
9. Howard R. Information value theory. IEEE Trans Syst Sci Cybern 1966;2(1):22–26.
10. Monahan G. A survey of partially observable Markov decision processes: theory, models, and algorithms. Manage Sci 1982;28:1–16.
11. Lovejoy W. A survey of algorithmic methods for partially observed Markov decision processes. Ann Oper Res 1991;28(1):47–65.
12. Cassandra AR, Kaelbling LP, Littmann ML. Acting optimally in partially observable stochastic domains. Intelligence (AAAI-94). Seattle, Washington: AAAI Press; 1994. pp. 1023–1028.
13. Kaelbling LP, Littmann ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artif Intell 1998;101(1-2):99–134.

14. Duff M. Optimal learning: computational procedures for Bayes-adaptive Markov decision processes [PhD thesis]. University of Massachusetts Amherst; 2002.
15. Powell W, Frazier P. Optimal learning. *Tutorials in operations research: state-of-the-art decision-making tools in the information-intensive age*. Hanover (MD): Informs; 2008.
16. Berry D, Fristedt B. *Bandit problems: sequential allocation of experiments*. Seattle, Washington: London; 1985.
17. DeGroot MH. *Optimal statistical decisions*. New York: McGraw-Hill; 1970.
18. Wetherill G, Glazebrook K. *Sequential methods in statistics*. 3rd ed. London: CRC Press; 1986.
19. Siegmund D. *Sequential analysis: tests and confidence intervals*. New York: Springer; 1985.
20. Bechhofer R. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann Math Stat* 1954;25(1):16–39.
21. Bechhofer R, Kiefer J, Sobel M. *Sequential identification and ranking procedures*. Chicago: University of Chicago Press; 1968.
22. Kim S, Nelson B. *Handbook in operations research and management science: simulation chapter selecting the best system*. Amsterdam: Elsevier; 2006.
23. Chick S. Bayesian ideas and discrete event simulation: why, what and how. In: *Proceedings of the 37th Conference on Winter Simulation*. Winter Simulation Conference. 2006; Monterey (CA). pp. 96–105.
24. Gelman A, Carlin J, Stern H, *et al.* *Bayesian data analysis*. 2nd ed. Boca Raton (FL): CRC Press; 2004.
25. Frazier P, Powell WB, Dayanik S. A knowledge gradient policy for sequential information collection. *SIAM J Control Optim* 2008;47(5):2410–2439.
26. Frazier P. 2009. Available at <http://people.orie.cornell.edu/pfrazier/src.html>.
27. Dayanik S, Goulding C, Poor H. Bayesian sequential change diagnosis. *Math Oper Res* 2008;33(2):475–496.
28. Hadjiliadis O, Poor H. *Quickest detection*. United Kingdom: Cambridge University Press; 2008.
29. Araman V, Caldenty R. Dynamic pricing for perishable products with demand learning. *Oper Res*; 2010;57:1169–1188.
30. Kapoor A, Greiner R. Learning and classifying under hard budgets. 16th European Conference on Machine Learning; 2005 Oct 3–7; Porto, Portugal. New York: Springer-Verlag New York, Inc.; 2005.
31. Gittins JC, Jones DM. A dynamic allocation index for the sequential design of experiments. In: Gani J, editor. *Progress in statistics*. Amsterdam: North-Holland; 1974. pp. 241–266.
32. Whittle P. Multi-armed bandits and the Gittins index. *J R Stat Soc Ser B (Methodological)* 1980;42(2):143–149.
33. Tsitsiklis JN. Asynchronous stochastic approximation and Q-learning. *Mach Learn* 1994;16:185–202.
34. Gittins J. *Multi-armed bandit allocation indices*. New York: John Wiley & Sons, Inc.; 1989.
35. Yao Y. Some results on the Gittins index for a normal reward process, *Lecture Notes-Monograph Series*. Beechwood (OH): Institute of Mathematical Statistics; 2006. pp. 284–294.
36. Whittle P. Restless bandits: activity allocation in a changing world. *J Appl Probab* 1988;25A:287–298.
37. Mahajan A, Teneketzis D. Multi-armed bandit problems. In: Hero AO III, Castanon DA, Cochran D, *et al.*, editors. *Foundations and applications of sensor management*. New York: Springer; 2007.
38. Ryzhov I, Powell W, Frazier P. The knowledge gradient algorithm for a general class of online learning problems. 2009. In press.
39. Ryzhov I, Frazier P. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. 2010;1(1):1629–1638.
40. Bergemann D, Valimaki J. *Bandit problems*. Cowles Foundation Discussion Paper 1551. Yale University; 2006.
41. Szepesvári C. *Reinforcement learning algorithms for mdps*. 2009. Available at <http://webdocs.cs.ualberta.ca/~szepesva/>.
42. Cesa-Bianchi N, Lugosi G. *Prediction, learning, and games*. Cambridge (UK): Cambridge University Press; 2006.
43. Cressie N. *Statistics for spatial data*. revised edition. New York: Wiley Interscience; 1993.
44. Rasmussen C, Williams C. *Gaussian processes for machine learning*. Cambridge (MA): MIT Press; 2006.
45. Frazier P, Powell WB, Dayanik S. The knowledge gradient policy for correlated normal

- beliefs. *INFORMS J Comput* 2009;21(4): 599–613.
46. Schonlau M. Computer experiments and global optimization [PhD thesis]. University of Waterloo; 1997.
 47. Schonlau M, Welch W, Jones D. New developments and applications in experimental design. Volume 34, Global versus local search in constrained optimization of computer models. Seattle (WA): Institute of Mathematical Statistics; 1998. pp. 11–25.
 48. Jones D, Schonlau M, Welch W. Efficient global optimization of expensive black-box functions. *J Glob Optim* 1998;13(4):455–492.
 49. Žilinskas A. Single-step Bayesian search method for an extremum of functions of a single variable. *Cybern Syst Anal* 1975;11(1): 160–166.
 50. Mockus J, Tiesis V, Zilinskas A. The application of Bayesian methods for seeking the extremum. In: Dixon L, Szego G, editors. Volume 2, Towards global optimisation. North Holland, Amsterdam: Elsevier Science Ltd.; 1978. pp. 117–129.
 51. Mockus J. Application of Bayesian approach to numerical methods of global and stochastic optimization. *J Glob Optim* 1994;4(4): 347–365.
 52. Huang D, Allen T, Notz W, *et al.* Sequential kriging optimization using multiple-fidelity evaluations. *Struct Multidiscip Optim* 2006;32(5):369–382.
 53. Forrester A, Keane A, Bressloff N. Design and analysis of “Noisy” computer experiments. *AIAA J* 2006;44(10):2331–2339.
 54. Forrester A, Sobester A, Keane A. Engineering design via surrogate modelling: a practical guide. West Sussex (UK): Wiley; 2008.
 55. Santner T, Willians BW, Notz W. The design and analysis of computer experiments. New York: Springer; 2003.
 56. Lophaven S, Nielsen H, Søndergaard J. Aspects of the Matlab toolbox DACE. Technical Report IMM-REP-2002-13. Lyngby: Technical University of Denmark; 2002a.
 57. Lophaven S, Nielsen H, Søndergaard J. Dace—a matlab kriging toolbox. Technical Report IMM-TR-2002-13. Lyngby: Technical University of Denmark; 2002b.
 58. Rasmussen C. The gaussian process website. 2009. Available at <http://www.gaussianprocess.org/>. Accessed 2010 Apr 16.
 59. Schonlau M. Manual: Space (Stochastic Process Analysis of Computer Experiments). 2001. Available at <http://www.schonlau.net/space.html>.
 60. Bickel P, Doksum K. Mathematical statistics: basic ideas and selected topics. 2nd ed. Prentice Hall; 2007.
 61. Poor H. An introduction to signal detection and estimation. New York: Springer; 1994.
 62. Lai T. Sequential analysis: some classical problems and new challenges. *Stat Sin* 2001;11(2):303–408.
 63. Wallis W. The statistical research group, 1942–1945. *J Am Stat Assoc* 1980;75:320–330.
 64. Jennison C, Turnbull B. Group sequential methods with applications to clinical trials. Boca Raton (FL): CRC Press; 2000.
 65. Porteus E. Foundations of stochastic inventory theory. Stanford (CA): Stanford Business Books; 2002.

LEVEL-DEPENDENT QUASI-BIRTH-AND-DEATH PROCESSES

JEFFREY P. KHAROUFEH

Department of Industrial Engineering,
University of Pittsburgh, Pittsburgh,
Pennsylvania

INTRODUCTION

A quasi-birth-and-death (QBD) process is a bivariate Markov process with state space $S = \{(i, j) : i \geq 0, j = 1, 2, \dots, m\}$ where i is called the *level* of the process, j is called the *phase* of the process, and m is an integer that can be finite or infinite. The process is restricted in level jumps only to its nearest neighbors but is unrestricted in the phase dimension. More precisely, from state $(i, j) \in S$ the process may transition to states of the form (i, k) , $(i - 1, k)$ or $(i + 1, k)$, but not to states of the form $(i \pm n, k)$ where $n \geq 2$. Clearly, the QBD process is an extension of the standard birth-and-death process whose state space consists only of the level i . When the transitions of a QBD process are independent of the level, it is termed a *homogeneous* or *level-independent* QBD process. Otherwise, it is termed an *inhomogeneous* or *level-dependent* quasi-birth-and-death (LDQBD) process. This very general class of models finds wide applicability in the modeling of queueing systems, particularly those with complex arrival and/or service processes such as the $PH/M/\infty$ queue. Several queueing, and non-queueing, applications of LDQBD processes will be discussed in the final section titled "Further Reading and Applications." The level-independent version has been studied extensively by a number of researchers and is given a very lucid treatment in the excellent text by Latouche and Ramaswami [1]. As a prototypical example, the homogeneous QBD process can be used to model

the $M/M/1$ queueing system in a random environment where the level i corresponds to the number of customers in the system, and the phase j is the state of an exogenous environment process that governs the arrival and/or service rates [2]. The class of QBD models, and their associated matrix-analytic solution techniques, provide a powerful set of tools for analyzing a wide variety of stochastic models arising in queueing theory, computer and communications systems, reliability modeling and analysis, inventory systems, and many others. For further information on the homogeneous QBD process, please see ***Level-Independent Quasi-Birth-and-Death Processes***.

The main purpose of the present article is to formally define and describe LDQBD processes whose transitions depend explicitly on the level i . Although their structure, and many of their attributes, mirror those of the level-independent version, their analysis and computational aspects are significantly more complicated. Here, we review both discrete- and continuous-time versions of the LDQBD process. For each type, we discuss necessary and sufficient conditions to establish positive recurrence, describe the limiting distribution, and discuss computational methods proposed for analyzing these models. We begin by describing the discrete-time version in the next section.

DISCRETE-TIME LDQBD PROCESSES

Suppose $\{(X_n, J_n) : n \geq 0\}$ is a discrete-time Markov process on the state space $S = \{(i, j) : i \geq 0, j = 1, 2, \dots, m\}$ where i and j are the level and phase of the process, respectively. For the purposes of this article, we will assume m is finite; however, m may be infinite and/or depend explicitly on the level i (as described by Bright and Taylor [3]). For the sake of notational brevity, let $\{Z_n : n \geq 0\} \equiv \{(X_n, J_n) : n \geq 0\}$. The one-step transition probability matrix of this process possesses the typical block tridiagonal form of a QBD process and is given by

$$P = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & 0 & 0 & \cdots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & \cdots \\ 0 & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & \cdots \\ 0 & 0 & A_2^{(3)} & A_1^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (1)$$

For $i \geq 0$ and $\ell = 0, 1, 2$, $A_\ell^{(i)}$ are m -order non-negative matrices, and the row sums are equal to unity, that is, the row sums of $A^{(i)} \equiv A_2^{(i)} + A_1^{(i)} + A_0^{(i)}$ and $A_1^{(0)} + A_0^{(0)}$ are equal to 1. To interpret the entries of P , we note that for $i \geq 1$, the process transitions from (i, j) to $(i-1, k)$ with probability $[A_2^{(i)}]_{j,k}$, or it transitions to state $(i+1, k)$ with probability $[A_0^{(i)}]_{j,k}$. A transition from (i, j) to (i, k) occurs with probability $[A_1^{(i)}]_{j,k}$. Finally, when $i = 0$, from state $(0, j)$ the process transitions to state $(1, k)$ with probability $[A_0^{(0)}]_{j,k}$, or to state $(0, k)$ with probability $[A_1^{(0)}]_{j,k}$. It should be clear that the rows of P are blocks of order m , and the structure indicates that the process is skip-free in the level in both directions. For the remainder of this section, we assume that P is irreducible.

Positive Recurrence and Limiting Distribution

A natural question is whether the Markov process with transition matrix P of Equation (1) possesses a limiting distribution, and if so, under what conditions does the distribution exist and what is its form? Unfortunately, the extreme generality of the LDQBD process precludes the development of generic tools that can be used to establish such properties as irreducibility, recurrence, and positive recurrence; however, in practice, it is often possible to establish these properties for specific applications using the attributes of the model. Stochastic dominance arguments are also often employed to show that the LDQBD process is stochastically dominated by a simpler process (e.g., a homogeneous QBD process) whose irreducibility and recurrence properties can be more easily established. Our aim here is to review the theoretical conditions for positive recurrence of irreducible LDQBD processes.

Denote the limiting distribution of $\{Z_n : n \geq 0\}$ by a positive row vector π that uniquely solves the system $\pi P = \pi$ and

$\pi e = 1$, where e is a column vector of ones. The vector π is partitioned by levels into subvectors so that

$$\pi = (\pi_0, \pi_1, \pi_2, \dots),$$

and π_i contains the limiting probabilities for states in level i , that is, the states in the set $L_i \equiv \{(i, 1), (i, 2), \dots, (i, m)\}$. Note that π_i has m components for each i . As shown in Chapter 6 of Ref. 1, the vector π_i assumes a *matrix-geometric* form when the process is level-independent, that is, when for $\ell = 0, 1, 2$, $A_\ell^{(i)} = A_\ell$ for every $i \geq 0$. More specifically, for $i \geq 0$, the vector π_i satisfies

$$\pi_i = \pi_0 R^i,$$

where R is a fixed matrix (independent of the level) such that for any pair of phases j and k , $[R]_{j,k}$ is the expected number of visits to state $(1, k)$ before returning to the set L_0 , provided that the process starts in state $(0, j)$. The result is a matrix analog to the scalar geometric distribution. (For a more thorough treatment of the matrix-geometric distribution, please see **Matrix-Geometric Distributions**.) Before stating the necessary and sufficient conditions for positive recurrence of a LDQBD process, we first assume it is irreducible, aperiodic, and positive recurrent in order to characterize the limiting distribution π . The following theorem mirrors Theorem 12.1.1. in Ref. 1 and is stated here without proof.

Theorem 1. *If the LDQBD process with transition probability matrix given by Equation (1) is irreducible, aperiodic, and positive recurrent, then there exist matrices $\{R_i : i \geq 1\}$, such that*

$$\pi_i = \pi_{i-1} R_i, \quad i \geq 1, \quad (2)$$

where the matrix R_i records the expected number of visits to the set L_i between any two visits to the set L_{i-1} , and the sequence $\{R_i\}$ is the minimal nonnegative solution of the set of equations given by

$$R_i = A_0^{(i-1)} + R_i A_1^{(i)} + R_i R_{i+1} A_2^{(i+1)}, \quad i \geq 1.$$

Stated more precisely, $[R_i]_{j,k}$ is equal to the expected number of visits to $(i+1, k)$ before a return to the set $L_0 \cup L_1 \cup \dots \cup L_i$, provided the process starts in state (i, j) . Notice that the expected number of visits depends explicitly on the level i from which we start. To determine conditions for positive recurrence, we next discuss first passage times of $\{Z_n : n \geq 0\}$ and introduce some important matrices that are very similar to those typically used to characterize the behavior of level-independent QBD processes.

Suppose $\{Z_n : n \geq 0\}$ starts in some state (i, j) , that is, in level $i > 0$ and phase j at time 0, and define S_1 as the first return time to level i . That is,

$$S_1 \equiv \inf\{n \geq 1 : Z_n \in L_i | Z_0 = (i, j)\}.$$

Furthermore, adopting the notation in Ref. 1, let τ denote the first passage time of $\{Z_n : n \geq 0\}$ to level $i-1$. Define the following two probabilities:

$$[U_i]_{j,k} = \mathbb{P}(S_1 < \tau, Z_{S_1} = (i, k) | Z_0 = (i, j)), \quad (3)$$

$$[G_i]_{j,k} = \mathbb{P}(\tau < \infty, Z_\tau = (i-1, k) | Z_0 = (i, j)). \quad (4)$$

Note that Equations (3) and (4) bear a striking resemblance to Equations (6.5) and (6.6) of Ref. 1 with one major exception—the probabilities above depend explicitly on the level i whereas they are completely independent of i when the QBD is level independent. We shall see that this dependence significantly complicates the process of computing the limiting distribution of an LDQBD process. The (j, k) th element of the matrix U_i is the probability that, given we start in level i , the process returns to state i before hitting level $i-1$, and the (j, k) th element of G_i denotes the probability that the process hits level $i-1$ in a finite duration of time, given that it starts somewhere in level i . It is not surprising that the sequences of matrices, $\{R_i\}$, $\{G_i\}$, and $\{U_i\}$ are interrelated. In fact, any one of the sequences determines the other two as noted by Theorem 12.1.2 in Ref. 1, which is restated here without proof.

Theorem 2. *Any one of the sequences $\{R_i\}$, $\{G_i\}$, and $\{U_i\}$ determines the other two via the following set of relations:*

$$\begin{aligned} G_i &= (I - U_i)^{-1} A_2^{(i)}, \\ R_i &= A_0^{(i-1)} (I - U_i)^{-1}, \\ U_i &= A_1^{(i)} + A_0^{(i)} G_{i+1}, \\ U_i &= A_1^{(i)} + R_{i+1} A_2^{(i)}. \end{aligned}$$

Note that the same relationships hold for R , G , and U in the level-independent case wherein none of the matrices depend on the level i . For computational purposes, the following result, Theorem 12.1.3 of Ref. 1, shows the explicit relationship between these matrices:

Theorem 3. *For each $i \geq 1$, the matrices R_i , U_i , and G_i verify the following set of equations:*

$$\begin{aligned} G_i &= A_2^{(i)} + A_1^{(i)} G_i + A_0^{(i)} G_{i+1} G_i, \\ R_i &= A_0^{(i-1)} + R_i A_1^{(i)} + R_i R_{i+1} A_2^{(i+1)}, \\ U_i &= A_1^{(i)} + A_0^{(i)} (I - U_{i+1})^{-1} A_2^{(i+1)}. \end{aligned}$$

In light of the above two theorems and Equation (2), it is clear that the condition for positive recurrence can be restated in terms of the matrix G_1 . The following theorem provides the necessary and sufficient condition that ensures positive recurrence of P (see also Theorem 12.1.4 of Ref. 1).

Theorem 4. *The LDQBD process is positive recurrent if and only if there exists a strictly positive solution to the system of equations*

$$\pi_0 = \pi_0 (A_1^{(0)} + A_0^{(0)} G_1) \quad (5)$$

subject to the normalization condition

$$\pi_0 \left(\sum_{n=0}^{\infty} \prod_{i=1}^n R_i \right) \mathbf{e} = 1 \quad (6)$$

where the empty product (when $n = 0$) results in the identity matrix.

We pause here to remark that the matrix G_1 can only be obtained approximately in

practice; therefore, it is difficult to assert positive recurrence using Equations (5) and (6). However, the matrix equation (5) can be replaced by

$$\pi_0 = \pi_0 (A_1^{(0)} + R_1 A_2^{(1)}) \quad (7)$$

due to the equivalence of $A_0^{(0)} G_1$ and $R_1 A_2^{(1)}$. Note that, if $m = \infty$, then we also require that the embedded Markov chain on L_0 be positive recurrent. This condition is obviously met when $m < \infty$. Taken together, Equations (2), (6), and (7) provide the limiting distribution π . It is obvious that we must compute the family of matrices $\{R_i : i \geq 1\}$ to obtain this limiting distribution. By contrast, we need only to compute the matrix R (or G or U) in the level-independent case to completely characterize π . We will review some algorithmic approaches for both the level-independent and level-dependent versions later. First, we describe and review the continuous-time version of the LDQBD process, which parallels the discrete-time case.

CONTINUOUS-TIME LDQBD PROCESSES

The discrete-time LDQBD process has a natural analog in continuous time, which we now describe. Suppose $\{(X(t), J(t)) : t \geq 0\}$ is a continuous-time Markov process on the state space $S = \{(i, j) : i \geq 0, j = 1, 2, \dots, m\}$ where i denotes the level of the process and j denotes the phase. The infinitesimal generator matrix of the continuous-time LDQBD process is of the form

$$Q = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & 0 & 0 & \dots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & \dots \\ 0 & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & \dots \\ 0 & 0 & A_2^{(3)} & A_1^{(3)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (8)$$

For each $i \geq 0$, the diagonal elements of $A_1^{(i)}$ are strictly negative, and the off-diagonal elements of $A_1^{(i)}$ are nonnegative. The matrices $A_0^{(i)}$ and $A_2^{(i)}$ are nonnegative. For $i \geq 1$, the matrix $A^{(i)} = A_0^{(i)} + A_1^{(i)} + A_2^{(i)}$ has zero row sums as does the matrix $A_0^{(0)} + A_1^{(0)}$. The structure of the generator matrix Q reveals that

its transitions, like the discrete-time counterpart, are restricted to nearest neighbors in the levels and unrestricted (in general) across the phase dimension. For $i \geq 1$, the process transitions from (i, j) to $(i-1, k)$ at rate $[A_2^{(i)}]_{j,k}$, it transitions to state $(i+1, k)$ with rate $[A_0^{(i)}]_{j,k}$, or to state (i, k) at rate $[A_1^{(i)}]_{j,k}$. Finally, from state $(0, j)$, the process transitions to $(1, k)$ at rate $[A_0^{(0)}]_{j,k}$, or to state $(0, k)$ with rate $[A_1^{(0)}]_{j,k}$. If the process possesses a limiting distribution, $\pi = (\pi_0, \pi_1, \pi_2, \dots)$, then it must satisfy the system of equations $\pi Q = \mathbf{0}$ and $\pi e = 1$, where $\mathbf{0}$ denotes the zero vector. We next review the necessary and sufficient condition for positive recurrence of the LDQBD process in continuous time and provide an expression for its limiting distribution.

Positive Recurrence and Limiting Distribution

As one might expect, the limiting distribution of the continuous-time version is analogous to its discrete-time counterpart. For the discussion that follows, we assume that $\{(X(t), J(t)) : t \geq 0\}$ with generator matrix Q is irreducible. The following main result shows that the limiting distribution restricted to any one of the levels depends explicitly on π_0 .

Theorem 5. *The LDQBD process with infinitesimal generator matrix given by Equation (8) is positive recurrent if and only if there exists a strictly positive solution to the system of equations*

$$\pi_0 (A_1^{(0)} + R_0 A_2^{(1)}) = \mathbf{0} \quad (9)$$

subject to the normalization condition

$$\pi_0 \left(\sum_{i=0}^{\infty} \prod_{n=0}^{i-1} R_n \right) e = 1. \quad (10)$$

In such a case, the m -order row vector π_i is given by

$$\pi_i = \pi_0 \prod_{n=0}^{i-1} R_n, \quad i \geq 0. \quad (11)$$

In Equations (10) and (11), when $i = 0$, the empty product results in the identity matrix I .

As for the discrete-time version, the limiting distribution is characterized by a level-dependent sequence of matrices $\{R_i : i \geq 0\}$ which, in general, can only be determined numerically. The elements of these matrices can be interpreted as follows. The (j, k) th element of R_i , denoted by $[R_i]_{j,k}$, is the expected sojourn time in state $(i+1, k)$ per unit sojourn in state (i, j) , given that the process started in state (i, j) . It is well known (cf. Bright and Taylor [3]) that the sequence $\{R_i\}$ is the minimal nonnegative solution of the set of equations

$$A_0^{(i)} + R_i A_1^{(i+1)} + R_i (R_{i+1} A_2^{(i+2)}) = 0, \quad i \geq 0. \quad (12)$$

If the number of levels and/or phases is infinite, computing the limiting distribution has two major complications. First, one must truncate the infinite series of Equation (10) in an appropriate manner, and second, the matrices $\{R_i : i \geq 0\}$ must be computed efficiently. The next section provides a brief overview of some of the algorithmic approaches to computing the limiting distribution of the discrete- and continuous-time LDQBD processes.

NUMERICAL ALGORITHMS

Numerical approaches for obtaining the limiting distribution of level-independent QBD processes have been studied fairly extensively in the literature. For both discrete- and continuous-time cases, Neuts [4] provides some methods for determining positive recurrence and computing the limiting distribution. Hajek [5] considered the case of homogeneous QBD processes with a finite number of levels and showed that the limiting distribution vector can be described by the sum of two matrix-geometric terms. For general, level-independent QBD processes, most noteworthy is the logarithmic reduction algorithm developed in 1993 by Latouche and Ramaswami [6], which guarantees quadratic convergence of the solution for the matrix R in level-independent QBD processes. A method for computing the stationary distribution of

finite level-independent QBD processes was contributed by Elhasfi and Molle [7]. Some software tools that implement the most popular algorithms have been contributed by various researchers [8–11].

For the level-dependent case, a few approaches are noteworthy. When the number of levels and the number of phases are both finite, Gaver *et al.* [12] provide stable algorithms for computing the limiting distribution and moments of first passage times in the continuous-time case. These algorithms are similar to the linear iteration schemes developed by Latouche and Ramaswami [13] who show how to use logarithmic reduction to compute the moments of passage times for level-independent processes. Ye and Li [14] developed the “folding algorithm” to compute the limiting distribution of a finite continuous-time LDQBD process. Their efficient algorithm directly solves $\pi Q = 0$ by employing state-space reduction techniques that reduce the complexity induced by level dependence. Li and Cao [15] developed two types of RG-factorizations in order to analyze LDQBD processes with either finitely or infinitely many levels. They use their factorizations to derive the Laplace transforms of the conditional distributions of integral functionals of QBD processes.

When the number of levels and/or phases is infinite, the most widely accepted approach for computing the limiting distribution is due to Bright and Taylor [3,16] who provide an extension of the logarithmic reduction algorithm of Latouche and Ramaswami [6] for the level-dependent case. Define the continuous-time LDQBD process $\{Z(t) : t \geq 0\} \equiv \{(X(t), J(t)) : t \geq 0\}$ on the state space $\{(i, j) : i \geq 0, 1 \leq j \leq M_i\}$ where M_i is the number of phases in level i , which can be infinite. Bright and Taylor [3] suggest truncating the infinite series of Equation (10) at some level K and then renormalizing to compute an approximate subvector of the form

$$\pi_i(K) = \pi_0(K) \prod_{n=0}^{i-1} R_n, \quad i \geq 1,$$

such that $\pi_0(K)$ satisfies Equation (9) and the normalization condition,

$$\pi_0(K) \sum_{i=0}^K \left[\prod_{n=0}^{i-1} R_n \right] \mathbf{e} = 1.$$

The subvectors, $\{\pi_i(K) : i \geq 0\}$, represent an invariant measure for the limiting distribution of all the states at or below level K . Therefore, for any $K \geq 0$, $\pi_i(K)$ is an upper bound for π_i (componentwise), and $\pi_i(K) \rightarrow \pi_i$ as $K \rightarrow \infty$.

In order to compute $\pi_i(K)$ for a chosen truncation level K , Bright and Taylor [3] examine the discrete-time Markov chain embedded at the jump epochs of $\{Z(t) : t \geq 0\}$ to compute the family of matrices $\{R_i : i \geq 0\}$ using a recursive scheme. The following result (Lemma 1 of Ref. 3) is a continuous-time extension of a discrete-time result stated by Ramaswami and Taylor [17].

Lemma 1. *If $\{Z(t) : t \geq 0\}$ is positive recurrent, then for $i \geq 0$ the matrix R_i is given by*

$$R_i = \sum_{\ell=0}^{\infty} U_i^{\ell} \prod_{n=0}^{\ell-1} D_{k+2^{\ell-1}}^{\ell-1-n}, \quad i \geq 0 \quad (13)$$

where, for $i \geq 1$, U_i^{ℓ} and D_i^{ℓ} are $M_{i-1} \times M_{i-1+2^{\ell}}$ and $M_{i-1} \times M_{i-1-2^{\ell}}$ matrices, respectively, that are recursively defined by

$$\begin{aligned} U_i^0 &= A_0^{(i)} \left(-A_1^{(i+1)} \right)^{-1}, \\ D_i^0 &= A_2^{(i)} \left(-A_2^{(i-1)} \right)^{-1}, \\ U_i^{\ell+1} &= U_i^{\ell} U_{i+2^{\ell}}^{\ell} \left[I - U_{i+2^{\ell+1}}^{\ell} D_{i+3 \cdot 2^{\ell}}^{\ell} \right. \\ &\quad \left. - D_{i+2^{\ell+1}}^{\ell} U_{i+2^{\ell}}^{\ell} \right]^{-1}, \\ D_i^{\ell+1} &= D_i^{\ell} D_{i-2^{\ell}}^{\ell} \left[I - U_{i-2^{\ell+1}}^{\ell} D_{i-2^{\ell}}^{\ell} \right. \\ &\quad \left. - D_{i-2^{\ell+1}}^{\ell} U_{i-3 \cdot 2^{\ell}}^{\ell} \right]^{-1}. \end{aligned}$$

The infinite series of Equation (13) can be truncated using a simple scheme (see Algorithms 2 and 3 of Ref. 3). By rearranging the terms in Equation (12), the family of

matrices, $\{R_i : i \geq 0\}$, can be recursively computed by

$$R_i = A_0^{(i)} \left(-A_1^{(i+1)} - R_{i+1} A_2^{(i+2)} \right)^{-1} \quad (14)$$

(assuming the inverse exists) so that Equation (13) need not be computed repeatedly. Algorithm 1 of Ref. 3 can be used to compute $\pi_i(K)$ for a chosen value K . This algorithm is summarized in what follows.

Computing $\pi(K)$ for Given K .

- Step 1: Calculate R_{K-1} using Equation (13).
 Step 2: Recursively calculate $R_{K-2}, R_{K-3}, \dots, R_0$ using Equation (14).
 Step 3: Solve the system of equations

$$\pi_0(K) \left[A_1^{(0)} + R_0 A_2^{(1)} \right] = \mathbf{0}, \quad \pi_0(K) \mathbf{e} = 1.$$

- Step 4: For $i = 1$ to K , set

$$\pi_i(K) = \pi_{i-1}(K) R_{i-1},$$

and normalize $\pi_n(K)$, $n = 0, 1, \dots, i$, so that

$$\sum_{n=0}^i \pi_n(K) \mathbf{e} = 1.$$

Algorithms 2 and 3 of Ref. 3 provide the means by which to truncate the infinite series of Equation (13) to effectively compute R_{K-1} in Step 1.

It is imperative to select an integer K that is large enough to ensure that the limiting probability of being in a state at or above level K is close to zero. Algorithm 4 of Ref. 3 describes a general method for choosing a truncation point that works well in practice. We next review the main steps of this algorithm.

Selecting the Integer K .

- Step 1: Set $K_0 = 0$.
 Step 2: Do the following:
 a. Choose K_1 such that $K_1 > K_0$.
 b. Calculate five terms of Equation (13) for R_{K_1-1} .

- If the five terms are sufficient, go to Step 2(c),
- else repeat Steps 2(a) and 2(b).
- c. Recursively calculate $R_{K_1-2}, R_{K_1-3}, \dots, R_{K_0}$ using Equation (14).
- d. Simultaneously solve

$$\pi_0(K_1) \begin{bmatrix} A_1^{(0)} + R_0 A_2^{(1)} \end{bmatrix} = \mathbf{0}, \quad \pi_0(K_1) \mathbf{e} = 1.$$

- e. For $i = K_0 + 1$ to K_1 , do

$$\pi_i(K_1) = \pi_{i-1}(K_1) R_{i-1},$$

and normalize $\pi_n(K)$, $n = 0, 1, \dots, i$, such that

$$\sum_{n=0}^i \pi_n(K_1) \mathbf{e} = 1.$$

- f. Set $K_0 = K_1$ until $\pi_{K_1}(K_1) \mathbf{e} < \epsilon$.

Step 3: Set $K = K_1$.

Besides this general procedure for determining K and the subvectors $\{\pi_i(K) : i \geq 0\}$, Bright and Taylor [3] also provide algorithms for some special cases including LDQBD processes with a large number of boundary states (Algorithm 5) and those which are stochastically dominated by another LDQBD process (Algorithms 6 and 7). The former is a type of LDQBD process that exhibits level-dependent behavior up to some fixed level $\bar{K}$, and level-independent behavior for each level i such that $i > \bar{K}$ [4].

FURTHER READING AND APPLICATIONS

For extensive coverage of the level-independent QBD process in discrete and continuous time, it is best to first consult the two major texts due to Neuts [4,18] and one due to Latouche and Ramaswami [1]. For a cogent summary of the level-independent case, please see *Level-Independent Quasi-Birth-and-Death Processes*, which provides the main results for limiting distributions and conditions for positive recurrence, as well as a more detailed discussion of algorithmic procedures for computing

the limiting distribution. The best reference for understanding the rudimentary aspects of the LDQBD process is Chapter 12 of the excellent text by Latouche and Ramaswami [1], which considers (primarily) the discrete-time version of the model. As noted above, the two papers by Bright and Taylor [3,16] provide the details of an efficient algorithm for computing the limiting distribution of the infinite state-space, continuous-time model.

Over the years, several extensions have been developed for both the level-independent and LDQBD processes. One of the most important and interesting extensions of level-independent QBD processes is due to Ramaswami [19] who showed connections to the analysis of stochastic fluid queueing models. A stochastic fluid queueing model can be viewed as a QBD process wherein the level assumes values on the nonnegative real line and the phase variable is discrete. The fluid level evolves in a piecewise linear manner depending on the current phase. Subsequent contributions to fluid queues include Ahn and Ramaswami [20] and Bean *et al.* [21,22]. Of particular interest is the recent paper by da Silva Soares and Latouche [23] who analyze level-dependent fluid models within the QBD paradigm. Bean *et al.* [24] provide a complete quasistationary analysis of the discrete-time LDQBD process by assuming that level 0 behaves as an absorbing state. Choi *et al.* [25] provide a generalization of finite LDQBD processes that they call “nested QBD chains,” which entail the addition of a third dimension, the period. They provide the fundamental matrices for these types of chains and prove other interesting properties as well.

The LDQBD process finds wide applicability as a model for systems with complex dynamics. It is most prevalent in the modeling of queueing systems with Markovian arrival process (MAP) input, or for those with phase-type distributions. For example, the LDQBD process is ideal for modeling the $PH/M/\infty$ queue. Some other interesting applications in queueing theory include models described in Refs 26–30. The emergence of the service sector in many economies has sparked a renewed interest in retrial queueing models. These models can help assess the

impact of retrial customers who are initially denied access to service but who retry the system after some random amount of time. Such models are particularly important in large-scale service systems, such as large customer contact centers. Consequently, many new single- and multiserver retrial queueing models have emerged over the past five years. The infinitesimal generator matrix of retrial queues is usually of the LDQBD type because the retrial rate varies with the level (which corresponds to the number of customers in orbit). Some notable LDQBD retrial queueing models can be found in Refs 31–39. In addition to queueing models, LDQBD processes are now emerging in reliability modeling and analysis and inventory systems modeling. Some interesting reliability applications are due to Pérez-Ocón and Montoro-Cazorla [40,41]. Recent inventory models that exploit the LDQBD structure can be found in Refs 42 and 43. The homogeneous QBD process has found wide applicability in the modeling and analysis of computer and communications systems, and Bright and Taylor [3] provide numerous examples of such. Recently, LDQBD processes have been applied as models of real-time wireless code division multiple access (WCDMA) communications systems to estimate performance parameters such as blocking probability and expected delay [44,45].

REFERENCES

1. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling, ASA-SIAM series on statistics and applied probability. Philadelphia (PA): SIAM; 1999.
2. Neuts MF. The M/M/1 queue with randomly varying arrival and service rates. *Opsearch* 1978;15:139–157.
3. Bright LW, Taylor PG. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Commun Stat Stoch Models* 1995;11:497–525.
4. Neuts MF. Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore (MD): The Johns Hopkins University Press; 1981.
5. Hajek B. Birth and death processes on the integers with phases and general boundaries. *J Appl Probab* 1982;19:488–499.
6. Latouche G, Ramaswami V. A logarithmic reduction algorithm for quasi-birth-and-death processes. *J Appl Probab* 1993;30:650–674.
7. Elhafsi EH, Molle M. On the solution to QBD processes with finite state space. *Stoch Anal Appl* 2007;25:763–779.
8. Bini DA, Meini B, Steffé S, *et al.* Structured Markov chains solver: software tools. In: SMC-tools '06: Proceeding from the 2006 Workshop on Tools for Solving Structured Markov Chains. New York: ACM; 2006. pp. 14.
9. Haverkort B, Van Moorsel A, Dijkstra A. MGMtool: a performance analysis tool based on matrix geometric methods. In: Pooley R, Hillston J, editors. *Modelling techniques and tools*. Edinburgh University Press; Edinburgh, UK. 1993. pp. 312–316.
10. Riska A, Smirni E. MAMSolver: a matrix-analytic methods tools. In: Field T, Harrison PG, Bradley J, *et al.*, editors. Volume 2324, *Modelling Techniques and Tools*. 12th International Conference, TOOLS 2002, Lecture Notes in Computer Science. Berlin: Springer; 2002. pp. 205–211.
11. Squillante MS. MAGIC: a computer performance modeling tool based on matrix-geometric techniques. In: Balbo G, Serazzi G, editors. *Computer performance evaluation: modeling techniques and tools*. Amsterdam: North-Holland; 1992. pp. 411–425.
12. Gaver DP, Jacobs PA, Latouche G. Finite birth-and-death models in randomly changing environments. *Adv Appl Probab* 1984;16:715–731.
13. Latouche G, Ramaswami V. Expected passage times in homogeneous quasi-birth-and-death processes. *Commun Stat Stoch Models* 1995;11:103–122.
14. Ye J, Li S. Folding algorithm: a computational method for finite QBD processes with level-dependent transitions. *IEEE Trans Commun* 1994;42(2–4):625–639.
15. Li Q, Cao J. Two types of RG-factorizations of quasi-birth-and-death processes and their applications to stochastic integral functionals. *Stoch Models* 2004;20(3):299–340.
16. Bright LW, Taylor PG. Equilibrium distributions for level-dependent quasi-birth-and-death processes. In: Chakravorthy SR, Alfa AS, editors. *Matrix Analytic Methods in Stochastic Models: Proceedings of the 1st International Conference*. New Jersey: Marcel Dekker; 1997. pp. 359–375.
17. Ramaswami V, Taylor PG. Some properties of the rate operators in level dependent

- quasi-birth-and-death processes with countable number of phases. *Commun Stat Stoch Models* 1996;12:143–164.
18. Neuts MF. Structured stochastic matrices of $M/G/1$ type and their applications. New York: Marcel Dekker; 1989.
 19. Ramaswami V. Matrix analytic methods for stochastic fluid flows. In: Smith D, Hey P, editors. *Teletraffic Engineering in a Competitive World* (Proceedings of the 16th International Teletraffic Congress). Edinburgh: Elsevier Science B.V.; 1999. pp. 1019–1030.
 20. Ahn S, Ramaswami V. Efficient algorithms for transient analysis of stochastic fluid flow models. *J Appl Probab* 2005;42:531–549.
 21. Bean NG, O'Reilly MM, Taylor PG. Hitting probabilities and hitting times for stochastic fluid flows. *Stoch Process Appl* 2005;115:1530–1556.
 22. Bean NG, O'Reilly MM. Performance measures of a multi-layer Markovian fluid model. *Ann Oper Res* 2008;160:99–120.
 23. da Silva Soares A, Latouche G. Fluid queues with level dependent evolution. *Eur J Oper Res* 2009;196:1041–1048.
 24. Bean NG, Pollett PK, Taylor PG. Quasistationary distributions for level-dependent quasi-birth-and-death processes. *Commun Stat Stoch Models* 2000;16(5):511–541.
 25. Choi SH, Kim B, Sohraby K, *et al.* On matrix-geometric solutions of nested QBD chains. *Queueing Syst* 2003;43:5–28.
 26. Alfa AS. Vacation models in discrete time. *Queueing Syst* 2003;44:5–30.
 27. Asmussen S, Pihlsgård M. Transient properties of many-server queues and related QBDs. *Queueing Syst* 2004;46:249–270.
 28. Gómez-Corral A. On a finite-buffer bulk-service queue with disasters. *Math Methods Oper Res* 2005;61:57–84.
 29. He Q, Neuts MF. Two $M/M/1$ queues with transfers of customers. *Queueing Syst* 2002;42:377–400.
 30. Lian Z, Liu X, Liu L. Discriminatory processor sharing queues and the DREB method. *Stoch Models* 2008;24:19–40.
 31. Alfa AS. Discrete-time analysis of $GI/G/1$ system with Bernoulli retrials: an algorithmic approach. *Ann Oper Res* 2006;141:51–66.
 32. Anisimov VV, Artalejo JR. Approximation of multiserver retrial queues by means of generalized truncated models. *Top* 2002;10(1):51–66.
 33. Artalejo JR, Chakravarthy SR. Algorithmic analysis of the $MAP/PH/1$ retrial queue. *Top* 2006;14(2):293–332.
 34. Avram F, Gómez-Corral A. On bulk-service $MAP/PH^{L,N}/1/N$ G -queues with repeated attempts. *Ann Oper Res* 2006;141:109–137.
 35. Breuer L, Dudin A, Klimenok V. A retrial $BMAP/PH/N$ system. *Queueing Syst* 2002;40:433–457.
 36. Kumar BK, Rukmani R, Thangaraj V. An $M/M/c$ retrial queueing system with Bernoulli vacations. *J Syst Sci Syst Eng* 2009;18(2):222–242.
 37. Li Q, Ying Y, Zhao YQ. A $BMAP/G/1$ retrial queue with server subject to breakdowns and repairs. *Ann Oper Res* 2006;141:233–270.
 38. Mushko VV, Jacob MJ, Ramakrishnan KO, *et al.* Multiserver queue with addressed retrials. *Ann Oper Res* 2006;141:283–301.
 39. Shin YW. Multi-server retrial queue with negative customers and disasters. *Queueing Syst* 2007;55:223–237.
 40. Cazorla DM, Pérez-Ocón R. An LDQBD process under degradation, inspection, and two types of repair. *Eur J Oper Res* 2008;190:494–508.
 41. Pérez-Ocón R, Montoro-Cazorla D. A multiple warm standby system with operational and repair times following phase-type distributions. *Eur J Oper Res* 2006;169:178–188.
 42. Deepak TG, Krishnamoorthy A, Narayanan VC, *et al.* Inventory with service time and transfer of customers. *Ann Oper Res* 2008;160:191–213.
 43. Narayanan VC, Deepak TG, Krishnamoorthy A, *et al.* On an (s, S) inventory policy with service time, vacation to server and correlated lead time. *Qual Technol Quant Manage* 2008;5(2):129–143.
 44. Hedge N, Altman E. Capacity of multiservice WCDMA networks with variable GoS. *Wireless Netw* 2006;12:241–253.
 45. Koukoutsidis I, Altman E, Kelif JM. A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system. *Proceedings IWQoS*. Passau, Germany. LNCS 3552; 2005. pp. 327–340.

LEVEL-INDEPENDENT QUASI-BIRTH-AND-DEATH PROCESSES

GUY LATOUCHE

Faculté des sciences, Université
Libre de Bruxelles, Bruxelles,
Belgium

Quasi-birth-and-death (QBD) processes are Markovian processes on a two-dimensional state space and their characteristic feature is that, in one of the dimensions, transitions are allowed only to the nearest values. To fix the ideas, consider that the state space is a strip of the form $S = \{(n, i) : n \geq 0, 1 \leq i \leq m\}$; the process may move from (n, i) to states of the form (n, j) or $(n - 1, j)$ or $(n + 1, j)$, but not to any state of the form $(n \pm k, j)$ where $k \geq 2$. In the first dimension, the process behaves like a birth-and-death process [1], which explains the origin of the name *quasi-birth-and-death* coined by Wallace [2].

The component n in the state (n, i) is usually called the *level* and the component i is called the *phase*. A QBD is *level independent* if the rates at which transitions occur do not depend on the level, although they may depend on the phase. Thus, a level-independent QBD is like an M/M/1 queue [1] where the rates of increase or decrease are constant. Processes for which the transitions do depend on the level are called *level-dependent QBD processes* [1].

Level-independent QBDs furnish one of the best illustrations of the *matrix-analytic method*, a type of mathematical analysis that combines analytic developments—mostly based on probabilistic reasoning and linear algebraic formulation—and algorithmic considerations. QBDs were first studied by Evans [3] and later by Neuts [4,5] who set a major research program in motion.

This article contains the basic material about QBDs: the results and algorithms that are of most immediate use for anyone using a QBD model.

DEFINITIONS

A *discrete-time*, level-independent QBD process is defined as a Markov chain $\{X_t : t \in \mathbb{N}\}$ on the state space $S = \{(n, i) : n \geq 0, 1 \leq i \leq m\}$ where m may be finite or infinite. We order the state space in lexicographic order and we write the transition matrix P as

$$P = \begin{bmatrix} B & A_1 & & & 0 \\ A_{-1} & A_0 & A_1 & & \\ & A_{-1} & A_0 & A_1 & \\ & & A_{-1} & A_0 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}, \quad (1)$$

where A_{-1} , A_0 , A_1 , and B are nonnegative matrices of order m , and the row-sums of P are all equal to 1, which means that the row-sums of $A_{-1} + A_0 + A_1$ and those of $B + A_1$ are also equal to 1.

From the structure of P , one sees that the Markov chain is restricted in its transitions: from the state (n, i) with $n \geq 1$, it may move to $(n - 1, j)$ with probability $(A_{-1})_{ij}$, or to (n, j) with probability $(A_0)_{ij}$, or to $(n + 1, j)$ with probability $(A_1)_{ij}$, independent of n ; for $n = 0$, the feasible transitions are from $(0, i)$ to $(1, j)$ with probability $(A_1)_{ij}$, and to $(0, j)$ with probability B_{ij} .

In many applications, the component n of the state is the number of customers in some queue and for that reason it is called the *level*. The component j is some auxiliary information, such as the environment phase in the M/M/1 queue in a random environment of Neuts [5], or the content of the intermediary buffer in the tandem system of Latouche and Neuts [6], or even a whole network of queues, as in Ramaswami and Taylor [7]. The second component is usually called the *phase*, unless one is dealing with a specific application that requires a more precise terminology. As the examples cited show, the phase may take values in a more or less complicated set of values; the important feature is that this set should be denumerable, so that we may

number its elements from 1 to $m \leq \infty$.

It may happen that B is equal to $A_{-1} + A_0$, but this is not always the case; furthermore, it may be that the system adopts a very different behavior when the main queue is empty, then the transition matrix takes the slightly different form

$$P = \begin{bmatrix} B_0 & B_1 & & 0 \\ B_{-1} & A_0 & A_1 & \\ & A_{-1} & A_0 & A_1 \\ & & A_{-1} & A_0 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}, \quad (2)$$

where B_0 is a square matrix of order m_0 , possibly different from m , in which case B_1 and B_{-1} would be rectangular matrices of appropriate dimensions.

A *continuous-time* level-independent QBD is defined in the same manner. The major difference is that its dynamics is characterized by an infinitesimal generator Q instead of a transition probability matrix P , with

$$Q = \begin{bmatrix} B & A_1 & & 0 \\ A_{-1} & A_0 & A_1 & \\ & A_{-1} & A_0 & A_1 \\ & & A_{-1} & A_0 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}. \quad (3)$$

The structure is the same as Equation (1), the difference being that the diagonal elements of B and of A_0 are strictly negative, that all the off-diagonal elements are nonnegative and that the row-sums of $A_{-1} + A_0 + A_1$ and of $B + A_1$ are zero.

We shall assume throughout this article that the QBD is irreducible: from any state, any other state may be reached after a finite number of transitions. We shall also assume here that the matrix $A = A_{-1} + A_0 + A_1$ is irreducible.

STABILITY CONDITION

A basic question is whether the states of the QBD are positive recurrent, null recurrent, or transient. A simple condition is expressed in terms of the stationary probability vector

α of the matrix A , if it exists. In discrete time, α is the unique solution of the system $\alpha^t A = \alpha^t$, $\alpha^t \mathbf{1} = 1$, where α^t is the transpose of α for any vector α , and $\mathbf{1}$ is a vector of 1's. In continuous time, α is the unique solution of the system $\alpha^t A = 0$, $\alpha^t \mathbf{1} = 1$.

The following property holds both in discrete- and in continuous time.

Theorem 1 [4, Chapter 1; 8, Chapter 7]. *Assume that the QBD is irreducible, that m is finite, and that the matrix A is irreducible, with stationary probability vector α . Define $\mu = \alpha^t A_1 \mathbf{1} - \alpha^t A_{-1} \mathbf{1}$. The QBD is positive recurrent if and only if $\mu < 0$; it is null recurrent if $\mu = 0$, and it is transient if $\mu > 0$.*

Other stability conditions are given in Latouche and Taylor [9]; if m is infinite, or if the matrix A is not irreducible, the conditions for stability are more complex and we refer the reader to Tweedie [10] and to Ref. 8, Chapter 7.

STATIONARY DISTRIBUTION

Discrete Time

We assume in this section that the system is positive recurrent. Quite naturally, we subdivide the stationary probability vector π into smaller vectors $\pi_0, \pi_1, \pi_2, \dots$, of size m , where π_n is the subvector of stationary probabilities for the states $(n, 1), \dots, (n, m)$ in level n . A remarkable fact is that the stationary distribution has a *matrix-geometric* form—the object of the next theorem.

Theorem 2 [4, Chapter 1; 8, Chapter 6]. *If the discrete-time QBD with transition matrix (Eq. 1) is positive recurrent, then there exists a nonnegative matrix R such that*

$$\pi_n^t = \pi_0^t R^n, \quad \text{for } n \geq 0. \quad (4)$$

The vector π_0 is the unique solution of the system

$$\pi_0^t (B + R A_{-1}) = \pi_0^t, \quad \pi_0^t \left(\sum_{k=0}^{\infty} R^k \right) \mathbf{1} = 1. \quad (5)$$

The matrix R is the minimal nonnegative solution of the matrix-quadratic equation

$$X = A_1 + XA_0 + X^2A_{-1} \quad (6)$$

and it has the following interpretation: for each pair of phases i and j , R_{ij} is the conditional expected number of times that the QBD visits the state $(1, j)$ until the next return to any of the states in level 0, given that the QBD is in state $(0, i)$ at time 0.

The stationary distribution is characterized by the subvector π_0 and by the matrix R , and π_0 itself is the solution of a known linear system, once R is known. This explains why a great deal of attention has been given in the literature on QBDs to methods to compute R in as efficient a manner as possible, a topic to which we return in the section titled “Algorithms.”

If the transition matrix is given by Equation (2) instead of Equation (1), we still have that $\pi_n = \pi_{n-1}R$, for $n \geq 2$, where R is the same as in Theorem 2, but the subvectors π_0 and π_1 are now given by the system

$$\begin{bmatrix} \pi_0^t & \pi_1^t \end{bmatrix} \begin{bmatrix} B_0 & B_1 \\ B_{-1} & A_1 + RA_{-1} \end{bmatrix} = \mathbf{0},$$

$$\pi_0^t \mathbf{1} + \pi_1^t \left(\sum_{k=0}^{\infty} R^k \right) \mathbf{1} = 1. \quad (7)$$

Here, π_0 is of size m_0 and π_1 is of size m .

Continuous Time

In continuous time, the results are very much the same, only a few details needing to be changed, because the stationary probability is characterized by the slightly different system $\pi^t Q = \mathbf{0}$, $\pi^t \mathbf{1} = 1$, where Q is the infinitesimal generator of the process.

Theorem 3 [4, Chapter 1; 8, Chapter 6]. Assume that the continuous-time QBD with transition matrix (Eq. 3) is positive recurrent. There exists a nonnegative matrix R such that

$$\pi_n^t = \pi_0^t R^n, \quad \text{for } n \geq 0. \quad (8)$$

The vector π_0 is the unique solution of the system

$$\pi_0^t (B + RA_{-1}) = \mathbf{0}, \quad \pi_0^t \left(\sum_{k=0}^{\infty} R^k \right) \mathbf{1} = 1. \quad (9)$$

The matrix R is the minimal nonnegative solution of the matrix-quadratic equation

$$A_1 + XA_0 + X^2A_{-1} = 0 \quad (10)$$

and it has the following interpretation: for each pair of phases i and j , R_{ij} is the rate at which the QBD visit the state $(1, j)$, per unit of time spent in $(0, i)$.

Observe that if $m < \infty$, the series in Equations (5), (7), and (9) is actually equal to $\sum_{k=0}^{\infty} R^k = (I - R)^{-1}$; if $m = \infty$, the series is the minimal nonnegative solution of the system $X(I - R) = I$.

Furthermore, if $m < \infty$, the convergence of the series indicates that the modulus of all the eigenvalues of R is strictly less than one or, equivalently, that $\text{sp}(R) < 1$, where $\text{sp}(\cdot)$ is the spectral radius. For transient or null-recurrent QBDs, we have $\text{sp}(R) = 1$.

FIRST PASSAGE DOWNWARD

Besides the matrix R , two transition probability matrices play an important role in the theory of QBDs; the first is defined as follows. Let T denote the first passage time to level 0, and define

$$G_{ij} = \Pr[T < \infty, X_T = (0, j) | X_0 = (1, i)], \quad (11)$$

for $1 \leq i, j \leq m$. In words, G_{ij} is the conditional probability that the QBD eventually visits level 0, and that $(0, j)$ is the first state

visited there, given that the process starts in phase i in level 1.

In queueing applications, T would be the length of a busy period and G would indicate whether such a busy period terminates with probability one or not and what is the state of the remainder of the system, upon completion of the busy period.

Theorem 4 [4, Chapter 3; 8, Chapter 8].

In discrete time, the matrix G is the minimal nonnegative solution of the equation

$$X = A_{-1} + A_0X + A_1X^2. \quad (12)$$

In continuous time, G is the minimal nonnegative solution of

$$A_{-1} + A_0X + A_1X^2 = 0. \quad (13)$$

In both cases, G is stochastic if the QBD is recurrent.

The similarity between Equations (6) and (12), and (10) and (13) suggests that there is a strong connection between the matrices R and G . This is made explicit through the introduction of a third matrix.

In *discrete time*, let θ be the first return time to level 1 and define

$$U_{ij} = \Pr[\theta < \infty, \theta < T, X_\theta = (1, j) | X_0 = (1, i)], \quad (14)$$

for $1 \leq i, j \leq m$. In words, U_{ij} is the probability of the following event: starting from $(1, i)$, the QBD eventually returns to level 1 before it visits any state in level 0, and it visits $(1, j)$ upon its return to level 1.

The matrix U is the transition matrix of the transient Markov chain embedded at those epochs when the QBD visits level 1 under taboo of level 0. It is transient because the QBD either eventually visits level 0 or it drifts to infinity and ceases to return to level 1, in both cases the taboo chain stops.

In *continuous time*, we define the censored Markov process of intervals of sojourn in the states of level 1, under taboo of the first visit to level 0, and we define U as the infinitesimal generator of that process.

Theorem 5 [8, Chapter 6]. *In both discrete- and continuous time,*

$$U = A_0 + A_1G. \quad (15)$$

In discrete time, $R = A_1 \sum_{k=0}^{\infty} U^k$, which is equivalent to

$$R = A_1(I - U)^{-1} \quad (16)$$

if $m < \infty$.

In continuous time, $R = A_1 \int_0^{\infty} e^{Ut} dt$, which is equal to

$$R = A_1(-U)^{-1} \quad (17)$$

if $m < \infty$.

ALGORITHMS

We address here the question of actually computing the matrix R ; this requires that m be finite so that calculations are feasible, and for that reason we assume throughout this section that $m < \infty$. Furthermore, as we have seen in the preceding two sections, the equations for discrete- and for continuous-time QBDs are very similar, which is why we restrict our attention to the discrete-time case for the most part.

Early efforts concentrated on directly solving Equation (6) [4, Chapter 1], since then, the focus has shifted to Equation (12). This is due to the fact that the three matrices R , G , and U are linked together by a number of equations, of which the following two are particularly useful:

$$G = (I - U)^{-1}A_{-1} \quad U = A_0 + RA_{-1}. \quad (18)$$

Together with Equations (15) and (16), these equations show that R , G , and U carry the same information and that one may compute any two of them from the third.

The main reason for working with Equation (12) instead of Equation (6) is that G is a stochastic or a substochastic matrix, which imposes stronger constraints on the solution of Equation (12). For that reason, we focus on algorithms to compute G , knowing

that Equations (15) and (16) allow us to easily compute R afterwards. We emphasize, however, that every algorithm designed for G has a twin algorithm which works for R , see Bini *et al.* [11, Section 5.3], Latouche [12], Latouche and Ramaswami [13], Neuts [4], and Ramaswami [14].

Basic Algorithms

It is simplest to use one of the defining equations for G , to start from an appropriate initial value and to proceed by successive substitutions. Thus, using Equation (12), one proves that the sequence

$$\begin{aligned} G_0 &= 0, \\ G_{n+1} &= A_{-1} + A_0 G_n + A_1 G_n^2, \end{aligned} \quad (19)$$

for $n \geq 0$ monotonically converges to G . Moreover, if the QBD under study is recurrent, then G is stochastic and one may stop the computation of the successive approximations as soon as they are sufficiently close to being stochastic.

If the QBD is recurrent there is some advantage in starting the iteration with a stochastic matrix. Some care must be used in choosing G_0 , but the identity matrix is always suitable so that the sequence

$$\begin{aligned} G_0 &= I, \\ G_{n+1} &= A_{-1} + A_0 G_n + A_1 G_n^2, \end{aligned} \quad (20)$$

for $n \geq 0$ also converges to G ; the convergence in this case is not monotone, but the final result is obtained with fewer iterations than with Equation (19).

If we use the pair of Equations (15) and (18), we obtain two linked sequences, one for U and one for G :

$$\begin{aligned} G_0 &= 0, \\ U_n &= A_0 + A_1 G_{n-1}, \\ G_n &= (I - U_n)^{-1} A_{-1}, \end{aligned} \quad (21)$$

for $n \geq 1$. Here, again the sequence $\{G_n\}$ monotonically converges to G , and here also

one may accelerate the procedure by starting with $G_0 = I$.

The sequence (Eq. 21) has two advantages over Equation (19). First, it converges faster, sometimes much faster. Secondly, it is amenable to a very simple interpretation in terms of the dynamic behavior of the QBD: the matrix G_n gives the first passage probabilities from level 1 to level 0 under the additional constraint that the QBD is not allowed to reach up beyond level n . With this interpretation, it is obvious that the sequence should be monotone and convergent.

The convergence rate of Equation (21) happens to be equal to $\text{sp}(R)$ which, as we have seen at the end of the section titled “Stationary Distribution,” is strictly less than 1 for positive recurrent QBDs and equal to 1 in the null recurrent and in the transient cases. One should, therefore, expect that the algorithms require large numbers of iterations when the parameters of the QBD are at the boundary of its stability region.

One may also mention that the algorithms are usually numerically stable; this is due to the fact that one is computing at each step a transition probability matrix for a well-behaved system. Further details are to be found in Ref. 11, Chapter 6, and also in Refs 8, Chapter 8 and 4, Chapter 3.

In continuous time, G is given as

$$G = (-U)^{-1} A_{-1}, \quad (22)$$

instead of Equation (18), and we may combine this equation with Equation (15) to construct the sequence

$$\begin{aligned} G_0 &= 0, \\ U_n &= A_0 + A_1 G_{n-1}, \\ G_n &= (U_n)^{-1} A_{-1}, \end{aligned} \quad (23)$$

for $n \geq 1$; its convergence and stability properties closely mimic those of Equation (21).

Quadratic Algorithms

In time, a number of algorithms have been proposed, which have quadratic convergence. Two popular ones, called the *Logarithmic* and

the *Cyclic reduction* algorithms are nearly identical. In short, they compute quantities similar to the first approximation $(I - A_0 - A_1)^{-1}A_{-1}$ in Equation (21), for a sequence of *new QBDs*, one level of the n th QBD corresponding to 2^n levels of the original one. This gives extremely efficient, quadratically convergent procedures. For cyclic reduction, which is slightly more efficient than logarithmic reduction, the successive approximations $G^{(0)}, G^{(1)}, G^{(2)}, \dots$ are given as

$$G^{(n)} = (I - \widehat{A}_0^{(n)})^{-1}A_{-1}, \quad (24)$$

where

$$\widehat{A}_0^{(0)} = A_0, \quad A_i^{(0)} = A_i, \quad \text{for } i = -1, 0, 1,$$

and

$$\begin{aligned} \widehat{A}_0^{(n+1)} &= \widehat{A}_0^{(n)} + A_1^{(n)}(I - A_0^{(n)})^{-1}A_{-1}^{(n)} \\ A_{-1}^{(n+1)} &= A_{-1}^{(n)}(I - A_0^{(n)})^{-1}A_{-1}^{(n)} \\ A_0^{(n+1)} &= A_0^{(n)} + A_1^{(n)}(I - A_0^{(n)})^{-1}A_{-1}^{(n)} \\ &\quad + A_{-1}^{(n)}(I - A_0^{(n)})^{-1}A_1^{(n)} \\ A_1^{(n+1)} &= A_1^{(n)}(I - A_0^{(n)})^{-1}A_1^{(n)} \end{aligned}$$

for $n \geq 0$.

As each iteration in the sequences (Eqs 19, 21, and 24) requires a few matrix products and the resolution of linear systems of order m , the numerical complexity is $O(m^3)$ per iteration in all cases. The difference in efficiency is determined by the coefficient of m^3 —which is slightly lower for the two linear algorithms—and by the number of iterations—which is much lower for the other two.

Tools

The algorithms above have been the object of tools, some of which are available through the Web [15–18].

Spectral Decomposition

As mentioned above, many authors have proposed other algorithms to determine

the stationary distribution of QBDs; for instance, Grassmann and Heyman [19] developed computational procedures that are similar to Equation (21).

A different class of methods have made use of the fact that the spectral decomposition of R , as well as that of G , can be obtained directly from the polynomial matrix

$$\begin{aligned} \Gamma(x) &= A_{-1} + x(A_0 - I) + x^2A_1 \\ &= (xA_1 + A_0 + A_1G - I)(xI - G) \end{aligned} \quad (25)$$

(in discrete time). The polynomial $\det \Gamma(x)$ has the $2m$ zeros

$$\xi_1 \leq \xi_2 \leq \dots \leq \xi_m \leq \xi_{m+1} \leq \dots \leq \xi_{2m},$$

counting $2m - r$ zeros at infinity if the degree of $\det \Gamma(x)$ is $r < 2m$.

Theorem 6. *For a discrete-time QBD with finitely many phases, under suitable irreducibility assumptions, the roots ξ_1 to ξ_m are the eigenvalues of G and the roots ξ_{m+1} to ξ_{2m} are the reciprocals of the eigenvalues of R . Moreover,*

- $\xi_m = 1 < \xi_{m+1}$ if the QBD is positive recurrent,
- $\xi_m = 1 = \xi_{m+1}$ if the QBD is null recurrent,
- and $\xi_m < \xi_{m+1} = 1$ if the QBD is transient.

A very thorough analysis is found in Gail *et al.* [20] and the references therein, and a summary in Ref. 8, Chapter 9. This is more of theoretical than practical interest, although some authors have made use of it in order to solve specific models, see Mitrani [21], for instance.

Another method based on spectral decomposition has been developed by Akar and Sohraby [22] and named the *invariant subspace* approach. It is shown in Bini *et al.* [23] to be a mirror image of cyclic reduction, after applying a transformation that unfortunately removes all probabilistic significance.

EXAMPLE

We conclude with an illustrative example of a single server queue in a random environment of four phases. At each transition, three events are possible:

- a service completion, with probability p_μ , provided the queue is not empty;
- an arrival, with probability p_λ if the system is in one of the phases 1, 2, or 3, and with probability p_Λ if the system is in phase 4;
- a change of phase with probability p_γ , the phases changing cyclically in the order 1, 2, 3, 4, 1, ...

The transition blocks are given below:

$$A_{-1} = p_\mu I, \quad A_0 = \begin{bmatrix} r_a & p_\gamma & & \\ & r_a & p_\gamma & \\ & & r_a & p_\gamma \\ p_\gamma & & & r_b \end{bmatrix},$$

$$A_1 = \begin{bmatrix} p_\lambda & & & \\ & p_\lambda & & \\ & & p_\lambda & \\ & & & p_\Lambda \end{bmatrix},$$

and $B = A_{-1} + A_0$, where $r_a = 1 - p_\mu - p_\gamma - p_\lambda$, and $r_b = 1 - p_\mu - p_\gamma - p_\Lambda$.

The stationary probability vector of A is uniform and, by Theorem 1, the queue is positive recurrent if $p_\mu > 0.75p_\lambda + 0.25p_\Lambda$ or, equivalently, if the traffic coefficient ρ defined as $(0.75p_\lambda + 0.25p_\Lambda)/p_\mu$ is strictly less than 1.

We have chosen the following values for the parameters: $p_\mu = 0.1$, $p_\gamma = 0.02$, $p_\lambda = 0.02$, and $p_\Lambda = 0.2$, which gives $\rho = 0.65$, a moderately loaded queue. Thus, a cycle through the four phases takes 200 units of time on average and the probability of an arrival increases by a factor of 10 in phase 4, so that the queue is pushed toward high values.

We have used the logarithmic reduction algorithm and asked for a precision of 10^{-8} . The algorithm requires seven iterations and

yields the matrices

$$G = \begin{bmatrix} 0.810 & 0.152 & 0.034 & 0.004 \\ 0.012 & 0.816 & 0.156 & 0.016 \\ 0.052 & 0.037 & 0.833 & 0.078 \\ 0.259 & 0.174 & 0.122 & 0.444 \end{bmatrix}$$

and

$$R = \begin{bmatrix} 0.162 & 0.030 & 0.007 & 0.001 \\ 0.002 & 0.163 & 0.031 & 0.003 \\ 0.010 & 0.007 & 0.167 & 0.016 \\ 0.519 & 0.349 & 0.244 & 0.889 \end{bmatrix}.$$

(The last row of G does not add up to 1 because all round-offs to three decimals happen to be truncations on that row.) Some patterns emerge; for instance, we find from G that if the process starts from $(1, i)$ with $i = 1, 2$, or 3, it is most likely that the queue goes down to level 0 in the same phase, while if we start in phase 4, there is a significant probability that, when the queue has recovered from the arrival spurt, it has had the time to move to another phase.

A comparison of the rows of R indicates, not surprisingly, that the system spends more time in level 1 if it starts from level 0 in phase 4 than if it starts from one of the other three phases. The spectral radius of R is 0.896: this is greater than ρ , which makes one suspect that the queue is not quite as well behaved as the value of ρ seems to indicate. This is clearly shown through a comparison of the QBD to a simple queue with only one phase and arrival probability $p_\lambda^* = 0.75p_\lambda + 0.25p_\Lambda$, so that the two systems are equally loaded, in the long run. In the table below, m_1 and σ_1 are the mean and the standard deviation of the queue length for the QBD and m_2 and σ_2 correspond to the simple queue.

ρ	m_1	σ_1	m_2	σ_2
0.65	5.760	8.430	1.857	2.304

We also plot in Fig. 1 the distribution function of the queue size for the two systems. Both systems have the same probability of being empty but that is where the similarity stops, the QBD having a much longer tailed distribution.

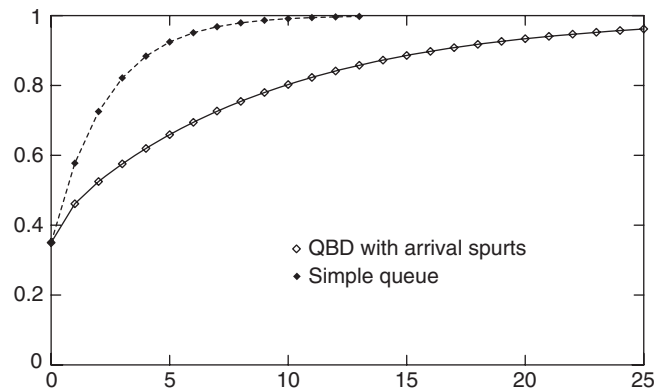


Figure 1. Cumulative probability distribution of a queue in a random environment and of a simple queue, both with traffic coefficient $\rho = 0.65$.

FURTHER READING

The theory of QBDs is extensively covered in Neuts [24,25] and Latouche and Ramaswami [8]. Various extensions are covered there, such as M/G/1- and GI/M/1-type queues, QBDs on finitely many levels, and tree-like processes. Algorithmic issues are discussed in depth in Bini *et al.* [11], see also the article on Markov chains of the M/G/1 type [1]. Chapters on QBDs and their immediate extensions may be found in Asmussen [26], Breuer and Baum [27], Dshalalow [28], and Grassmann [29].

The methods from the theory of QBDs have seen their domain of application greatly extended over the year. One extension that we have mentioned in the introduction is that of level-dependent QBDs. These are processes for which the transition probabilities from level n may depend on n ; they are discussed in this encyclopedia and a good reference is Bright and Taylor [30].

Ramaswami [31] showed that the methods developed for QBDs might be applicable to fluid queues. Fluid queues are two-dimensional stochastic processes for which the level takes continuous values and evolve in a piecewise linear manner. The seminal paper of Ramaswami opened the way to many developments; recent papers on the subject are Ahn and Ramaswami [32], Bean *et al.* [33,34], and da Silva Soares and Latouche [35].

Recently, it has been shown that algorithm for the analysis of Markovian branching processes had much in common with

those for QBDs Bean *et al.* [36] and Hautphenne *et al.* [37].

REFERENCES

1. Cochran JJ, editor. Wiley encyclopedia of operations research and management science. Wiley; 2009.
2. Wallace V. The solution of quasi birth and death processes arising from multiple access computer systems [PhD thesis]. Systems Engineering Laboratory, University of Michigan, 1969.
3. Evans RV. Geometric distribution in some two-dimensional queueing systems. *Oper Res* 1967;15:830–846.
4. Neuts MF. Markov chains with applications in queueing theory, which have a matrix-geometric invariant vector. *Adv Appl Probab* 1978;10:185–212.
5. Neuts MF. The M/M/1 queue with randomly varying arrival and service rates. *Oper Res* 1978;15:139–157.
6. Latouche G, Neuts MF. Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM J Algebraic Discrete Methods* 1980;1:93–106.
7. Ramaswami V, Taylor PG. An operator-analytic approach to product-form networks. *Commun Stat Stoch Models* 1996;12: 121–142.
8. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability. Philadelphia (PA): SIAM; 1999.
9. Latouche G, Taylor PG. Drift conditions for matrix-analytic models. *Math Oper Res* 2003;28:346–360.

10. Tweedie RL. Operator-geometric stationary distributions for Markov chains, with applications to queueing models. *Adv Appl Probab* 1982;14:368–391.
11. Bini DA, Latouche G, Meini B. Numerical methods for structured Markov chains. Numerical mathematics and scientific computation. Oxford: Oxford University Press; 2005.
12. Latouche G. Quasi-birth-and-death processes: beyond stable queues. In: Alfa A, Chakravarthy S, editors. *Advances in matrix analytic methods for stochastic models—proceedings of the 2nd International workshop on matrix-analytic methods*. New Jersey: Notable Publications Inc.; 1998. pp. 79–92.
13. Latouche G, Ramaswami V. A logarithmic reduction algorithm for quasi-birth-and-death processes. *J Appl Probab* 1993;30:650–674.
14. Ramaswami V. A duality theorem for the matrix paradigms in queueing theory. *Commun Stat Stoch Models* 1990;6:151–161.
15. Bini DA, Meini B, Steffe S, *et al.* Structured Markov chains solver: software tools. SMC tools '06: Proceeding from the 2006 workshop on Tools for solving structured Markov chains. New York: ACM; 2006. pp. 14.
16. Haverkort B, Van Moorsel A, Dijkstra A. MGMtool: a performance analysis tool based on matrix geometric methods. In: Pooley R, Hillston J, editors. *Modelling techniques and tools*. Edinburgh University Press; 1993. pp. 312–316.
17. Riska A, Smirni E. MAMSolver: a matrix-analytic methods tools. In: Field T, Harrison PG, Bradley J, *et al.*, editors. Volume 2324. *Modelling techniques and tools*. 12th International conference, TOOLS 2002. Lecture notes in computer science. Berlin: Springer; 2002. pp. 205–211.
18. Squillante MS. MAGIC: a computer performance modeling tool based on matrix-geometric techniques. In: Balbo G, Serazzi G, editors. *Computer performance evaluation: modeling techniques and tools*. Amsterdam: North-Holland; 1992. pp. 411–425.
19. Grassmann WK, Heyman DP. Computation of steady-state probabilities for infinite Markov chains with repeating rows. *ORSA J Comput* 1993;5:292–303.
20. Gail HR, Hantler SL, Taylor BA. Use of characteristic roots to solve infinite state Markov chains. In: Grassmann W, editor. *Computational probability*. Boston: Kluwer; 2000. pp. 205–255.
21. Mitrani I. The spectral expansion solution method for Markov processes on lattice strips. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Raton (FL): CRC Press; 1995. pp. 337–352.
22. Akar N, Sohraby K. An invariant subspace approach in M/G/1 and G/M/1 type Markov chains. *Commun Stat Stoch Mod* 1997;13:381–416.
23. Bini DA, Latouche G, Meini B. Solving matrix polynomial equations arising in queueing problems. *Linear Algebra Appl* 2002;340:225–244.
24. Neuts MF. Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore (MD): The Johns Hopkins University Press; 1981.
25. Neuts MF. Structured stochastic matrices of M/G/1 type and their applications. New York: Marcel Dekker; 1989.
26. Asmussen S. Applied probability and queues. 2nd ed. New York: Springer; 2003.
27. Breuer L, Baum D. An introduction to queueing theory and matrix-analytic methods. Berlin: Springer; 2005.
28. Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Probability and stochastic series. Boca Raton (FL): CRC Press; 1995.
29. Grassmann WK, editor. *Computational probability*. Boston (MA): Kluwer Academic Publisher; 2000.
30. Bright LW, Taylor PG. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Commun Stat Stoch Mod* 1995;11:497–525.
31. Ramaswami V. Matrix analytic methods for stochastic fluid flows. In: Smith D, Hey P, editors. *Teletraffic engineering in a competitive world (Proceedings of the 16th International teletraffic congress)*. Edinburgh: Elsevier Science B.V.; 1999. pp. 1019–1030.
32. Ahn S, Ramaswami V. Efficient algorithms for transient analysis of stochastic fluid flow models. *J Appl Probab* 2005;42:531–549.
33. Bean NG, O'Reilly MM, Taylor PG. Hitting probabilities and hitting times for stochastic fluid flows. *Stoch Proc Appl* 2005;115:1530–1556.
34. Bean NG, O'Reilly MM. Performance measures of a multi-layer Markovian fluid model. *Ann Oper Res* 2008;160:99–120.
35. da Silva Soares A, Latouche G. Fluid queues with level dependent evolution. *Eur J Oper Res* 2009;196:1041–1048.

- 36. Bean NG, Kontoleon N, Taylor PG. Markovian trees: properties and algorithms. *Ann Oper Res* 2008;160:31–50.
- 37. Hautphenne S, Latouche G, Rémiche M-A. Newton's iteration for the extinction probability of a Markovian binary tree. *Linear Algebra Appl* 2008;428:2791–2804.

LIFT-AND-PROJECT INEQUALITIES

QUENTIN LOUVEAUX
Montefiore Institute, University of
Liège, Liège, Belgium

Lift-and-project inequalities were proposed in the early 1990s as a way to strengthen the linear programming (LP) relaxation of a mixed-integer program (MIP). The initial idea was proposed by Lovász and Schrijver [1] when they obtained a strengthened formulation in an extended space by multiplying all constraints by all variables x_i and their complement and to project back to the initial space of variables. A similar but slightly different approach was then later proposed by Sherali and Adams [2]. In this survey, we follow the approach of Balas *et al.* [3] who showed that a simplified version of the Lovász–Schrijver reformulation keeps its main theoretical property. They also showed how the approach can successfully be incorporated in a branch-and-cut solver [4].

The general idea behind lift-and-project is the following: from an initial formulation, we can create a quadratic equivalent formulation by multiplying all inequalities by a binary variable x_j of the problem and its complement $(1 - x_j)$. If we linearize this formulation by introducing a variable $y_i := x_i x_j$, and replacing $x_j^2 = x_j$ (since $x_j \in \{0, 1\}$), we obtain an equivalent formulation in an extended space. This is the *lifting phase*. If we project this extended formulation onto the space of the initial variables, we obtain a strengthened formulation of the initial binary MIP. This is the *projecting phase*.

This article is further subdivided into two parts. The first section presents the lift-and-project technique and shows that it encodes a simple convexification process. We then show that this convexification process can be used sequentially in order to generate the convex hull of all feasible points. Lift-and-project inequalities can be related to other classes of inequalities. We also make this link in this section. The second section focuses on

the different ways to generate lift-and-project inequalities that cut off a fractional point, in particular, starting from the optimal simplex tableau of a linear relaxation. We also present some computational tests that have addressed the strength of the lift-and-project inequalities.

Two other references on the topic are the survey on cutting planes by Cornuéjols [5] and the survey on lift-and-project by Balas and Perregaard [6].

GENERAL THEORY

Before presenting the lift-and-project on its own, we first start with some basics about the projection of polyhedra since this is a building block of the lift-and-project technique.

Definition 1. Let $P \subseteq \mathbb{R}^{p+q}$ be a polyhedron where an element of P is denoted by $(x, y) \in P$, $x \in \mathbb{R}^p, y \in \mathbb{R}^q$. The projection of P on its first p coordinates is defined as

$$\text{proj}_x(P) = \{x \in \mathbb{R}^p \mid \exists y \in \mathbb{R}^q, \text{ with } (x, y) \in P\}.$$

The projection of a polyhedron can be computed using the following result.

Lemma 1. Let $P := \{(x, y) \in \mathbb{R}^{p+q} \mid Ax + Gy \geq b\}$. Let $\{u^t\}_{t \in t}$ be the list of extreme rays of $\{u \in \mathbb{R}^q \mid u^T G = 0, u \geq 0\}$. Then $\text{proj}_x(P) = \{x \in \mathbb{R}^p \mid (u^t)^T Ax \geq (u^t)^T b\}$.

Proof. See any textbook on linear and integer programming; for example, Refs 7–9.

We now turn to lift-and-project. We consider the mixed binary set $K_I = \{x \in \{0, 1\}^n \times \mathbb{R}_+^p \mid Ax \geq b\}$ with $A \in \mathbb{R}^{m \times (n+p)}$, $b \in \mathbb{R}^m$ and its linear relaxation $K = \{x \in \mathbb{R}_+^{n+p} \mid Ax \geq b\}$. We suppose that the constraints $x_i \geq 0, x_i \leq 1, i = 1, \dots, n$ are included in the description $Ax \geq b$. The lift-and-project technique is summarized in Table 1.

Table 1. Lift-and-Project Technique

Step 0	Select a binary variable $x_j, j = 1, \dots, n$.
Step 1	Multiply all constraints by x_j and $(1 - x_j)$ giving the quadratic system $Q_j(K) = \{x \in \mathbb{R}_+^{n+p} \mid x_j(Ax - b) \geq 0, (1 - x_j)(Ax - b) \geq 0\}$.
Step 2	Linearize $Q_j(K)$ by introducing a variable $y_i := x_i x_j, i \neq j$ and replacing x_j^2 by x_j . We obtain the polyhedron $M_j(K)$.
Step 3	Project $M_j(K)$ onto the space of the x -variables. The resulting polyhedron is $P_j(K)$.

Observation 1. $P_j(K) \subseteq K$.

Proof. If we sum up all inequalities defining $Q_j(K)$, we obtain $Ax \geq b$ and therefore summing up all inequalities defining $M_j(K)$, we also obtain $Ax \geq b$. It follows that any point $(x, y) \in M_j(K)$ must satisfy $Ax \geq b$. Hence, $P_j(K) \subseteq K$.

This shows that the lift-and-project technique produces a new valid formulation for the points in K_I . We can also observe that the only operation that takes care of the integrality of the variables is when we replace x_j^2 by x_j . This operation is only valid when $x_j \in \{0, 1\}$. The first theorem of this section is that the operation actually strengthens the formulation in the best possible way considering the fact that only the integrality of x_j is considered.

Theorem 1. $P_j(K) = \text{conv}\{(K \cap \{x_j = 0\}) \cup (K \cap \{x_j = 1\})\}$.

Proof. The proof is taken from Ref. 3.

- (i) $\text{conv}(K \cap \{x : x_j \in \{0, 1\}\}) \subseteq P_j(K)$. Let $\bar{x} \in K \cap \{x : x_j \in \{0, 1\}\}$. Define $\bar{y}_i = \bar{x}_i \bar{x}_j$ for $i \neq j$. Then $(\bar{x}, \bar{y}) \in M_j(K)$ and hence $\bar{x} \in P_j(K)$.
- (ii) $P_j(K) \subseteq \text{conv}(K \cap \{x : x_j \in \{0, 1\}\})$. Assume first $K \cap \{x : x_j = 0\} = \emptyset$. Then $x_j \geq \epsilon$ is valid for K for some $\epsilon > 0$. This implies that $(1 - x_j)(x_j - \epsilon) \geq 0$ is satisfied by any $x \in Q_j(K)$. Replacing x_j^2 by x_j , it follows that $x_j \geq 1$ is valid for $Q_j(K) \cap \{x : x_j^2 = x_j\}$ from which it follows that it is valid for $P_j(K)$. Hence, $P_j(K) \subseteq K \cap \{x : x_j = 1\}$. The case $K \cap \{x : x_j = 1\} = \emptyset$ is similar.

Assume now that $K \cap \{x : x_j = l\} \neq \emptyset, l = 0, 1$. Suppose $\alpha x \geq \beta$ is valid for $\text{conv}(K \cap \{x : x_j \in \{0, 1\}\})$. Since it is valid for $K \cap \{x : x_j = 0\}$, there exists $\lambda \geq 0$ such that $\alpha x + \lambda x_j \geq \beta$ is valid for K . Similarly, there exists $\mu \geq 0$ such that $\alpha x + \mu(1 - x_j) \geq \beta$. Multiplying these two inequalities by $(1 - x_j)$ and x_j respectively, we obtain that $(1 - x_j)(\alpha x + \lambda x_j - \beta) \geq 0$ and $x_j(\alpha x + \mu(1 - x_j) - \beta) \geq 0$ are valid for $Q_j(K)$. Adding these two inequalities yields $\alpha x + (\lambda + \mu)(x_j - x_j^2) - \beta \geq 0$. Hence, $\alpha x \geq \beta$ is valid for $Q_j(K) \cap \{x : x_j^2 = x_j\}$ and therefore it is valid for $P_j(K)$ too.

The nice feature about the lift-and-project technique is that it can be applied sequentially in order to generate $\text{conv}(K_I)$ in a finite number of steps.

Theorem 2. $P_n(P_{n-1}(\dots(P_1(K))\dots)) = \text{conv}(K_I)$.

Proof. The proof is taken from Ref. 5. We proceed by induction. Let $S_t := \{x \in \{0, 1\}^t \times \mathbb{R}^{n-t+p} : Ax \geq b\}$. We want to show that $P_t(\dots(P_1(K))\dots) = \text{conv}(S_t)$. This is true for $t = 1$ as was shown in Theorem 1. Thus, we consider $t \geq 2$ and suppose that it is true for $t - 1$. By the induction hypothesis, we have $P_t(\dots(P_1(K))\dots) = P_t(\text{conv}(S_{t-1}))$ and using again Theorem 1, we have

$$P_t(\dots(P_1(K))\dots) = \text{conv}((\text{conv}(S_{t-1}) \cap \{x_t = 0\}) \cup (\text{conv}(S_{t-1}) \cap \{x_t = 1\})). \quad (1)$$

For any set S that lies entirely on one side of a hyperplane H , we have that $\text{conv}(S) \cap H =$

$\text{conv}(S \cap H)$ [8]. We can therefore rewrite Equation (1) as

$$\begin{aligned} P_t(\dots(P_1(K)\dots) &= \text{conv}(\text{conv}(S_{t-1} \cap \{x_t = 0\}) \\ &\quad \cup (\text{conv}(S_{t-1} \cap \{x_t = 1\}))) \\ &= \text{conv}((S_{t-1} \cap \{x_t = 0\}) \\ &\quad \cup (S_{t-1} \cap \{x_t = 1\})) \\ &= \text{conv}(S_t). \end{aligned}$$

We see that if we apply the lift-and-project technique in sequence, we will eventually generate the integer hull of the initial set. On the other hand, if we apply the lift-and-project technique in “parallel,” we construct the so-called *lift-and-project closure*.

Definition 2. The lift-and-project closure is the set $\cap_{j=1}^n P_j(K)$.

Example. We now present a small example that illustrates the technique. Consider the simple two dimensional polyhedron

$$K = \{x \in \mathbb{R}^2 \mid 0 \leq x_1, x_1 \leq 1, 0 \leq x_2, x_2 \leq 1, \\ 5x_1 - 6x_2 \geq -3, x_1 + 6x_2 \geq 2\},$$

from which we are interested in the description of the integer hull (Fig. 1).

Step 0 We select $j = 1$.

Step 1 We multiply all inequalities by x_1 and $1 - x_1$. We obtain the quadratic system

$$\begin{aligned} Q_1(K) = \{x \in \mathbb{R}^2 \mid x_1^2 \geq 0, \quad x_1^2 \leq x_1, \\ x_1 - x_1^2 \geq 0, \quad x_1 - x_1^2 \leq 1 - x_1, \\ x_1x_2 \geq 0, \quad x_1x_2 \leq x_1, \\ x_2 - x_1x_2 \geq 0, \quad x_2 - x_1x_2 \leq 1 - x_1, \\ 5x_1^2 - 6x_1x_2 \geq -3x_1, \\ 5x_1 - 5x_1^2 - 6x_2 + 6x_1x_2 \geq -3 + 3x_1, \\ x_1^2 + 6x_1x_2 \geq 2x_1, \\ x_1 - x_1^2 + 6x_2 - 6x_1x_2 \geq 2 - 2x_1\}. \end{aligned}$$

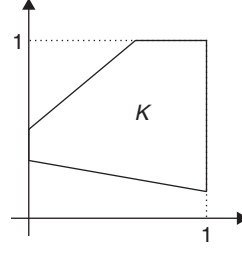


Figure 1. The initial polyhedron K .

Step 2 We replace x_1^2 by x_1 and we linearize x_1x_2 by introducing the variable $y_2 := x_1x_2$. We obtain

$$\begin{aligned} L_1(K) = \{(x, y) \in \mathbb{R}^{2+1} \\ | \quad x_1 \geq 0, y_2 \geq 0, x_1 - y_2 \geq 0, \\ -x_1 \geq -1, x_2 - y_2 \geq 0, -x_1 - x_2 + y_2 \\ \geq -1 \\ 8x_1 - 6y_2 \geq 0, \quad -3x_1 - 6x_2 + 6y_2 \geq -3 \\ -x_1 + 6y_2 \geq 0, \quad 2x_1 + 6x_2 - 6y_2 \geq 2\}. \end{aligned}$$

Step 3 We now need to project out the y_2 variable. To do it, we may consider either a Fourier–Motzkin elimination or use Lemma 1. For the latter, we have to consider vectors $u \in \mathbb{R}_+^{10}$ that are extreme rays of

$$\begin{aligned} Q := \{u \in \mathbb{R}^{10} \mid u_2 - u_3 - u_5 + u_6 - 6u_7 \\ + 6u_8 + 6u_9 - 6u_{10} = 0, u_i \geq 0\}. \end{aligned}$$

The set Q has 18 extreme rays and considering some of them leads to redundant inequalities. If we consider $u_3^1 = 2, u_8^1 = \frac{1}{3}$, we obtain the inequality $x_1 - 2x_2 \geq -1$. Similarly, if we consider $u_8^2 = 1, u_{10}^2 = 1$, we obtain the inequality $x_1 \leq 1$. We can also finally obtain the inequalities $x_1 \geq 0$ and $x_1 + 6x_2 \geq 2$. All the other inequalities are redundant. The resulting polyhedron $P_1(K)$ is represented in Fig. 2. If we compute $P_2(P_1(K))$, we would obtain the integer hull of the initial polyhedron, namely, $P_2(P_1(K)) = \{(1, 1)\} = \text{conv}(K_I)$. The single point of K_I is shown in Fig. 2.

It is possible to make some connections between lift-and-project inequalities and

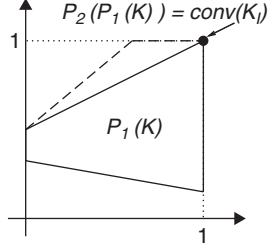


Figure 2. The polyhedron $P_1(K)$ after one iteration of the lift-and-project. After two iterations, we obtain the integer hull.

some other well-known classes of inequalities. Observe indeed that Theorem 1 has several implications that we summarize below.

Observation 2. The class of lift-and-project inequalities is a subset of

- disjunctive inequalities introduced by Balas [10] where the disjunction is $(x_j \leq 0) \vee (x_j \geq 1)$;
- Gomory mixed-integer cuts [11] or mixed-integer rounding [12];
- split cuts [13] where the corresponding split is $\{0 \leq x_j \leq 1\}$;
- intersection cuts [14].

GENERATING LIFT-AND-PROJECT CUTS

In this section, we consider the issue of optimizing a linear function over K_I . This is often done by considering the linear relaxation K and by finding cutting planes that cut off the optimal point of the linear relaxation. A natural question that we may ask is to find the best lift-and-project cut that cuts off an optimal vertex of the linear relaxation. This can be answered by solving a linear program as we explain below. Step 2 of the lift-and-project technique constructs the polyhedron

$$M_j(K) = \{x \in \mathbb{R}^{n+p} \mid Ay - bx_j \geq 0, \\ Ax + bx_j - Ay \geq b, y_j = x_j\}.$$

A simpler representation of M_j can be obtained by getting rid of the y_j variable.

If we denote by $\tilde{A}$ the matrix obtained by removing from A its j^{th} column A_j , we can rewrite $M_j(K)$ as

$$M_j(K) = \{x \in \mathbb{R}^{n+p}, y \in \mathbb{R}^{n+p-1} \mid \tilde{A}y + (A_j - b) \\ x_j \geq 0, Ax + (b - A_j)x_j - \tilde{A}y \geq b\}.$$

By renaming the matrices, we may write the set as

$$M_j(K) = \{x \in \mathbb{R}^{n+p}, y \in \mathbb{R}^{n+p-1} \mid \tilde{C}x + \tilde{A}y \geq 0, \\ \tilde{C}x - \tilde{A}y \geq b\}.$$

A valid lift-and-project inequality is a valid inequality for the set $\text{proj}_x M_j(K)$. It can be computed using Lemma 1 by considering the set $Q = \{(u, v) \in \mathbb{R}_+^m \times \mathbb{R}_+^m \mid u^T \tilde{A} - v^T \tilde{A} = 0, u, v \geq 0\}$. Therefore, an inequality is determined as $(u\tilde{C} + v\tilde{C})x \geq vb$ with $(u, v) \in Q$. If we want to cut off a specific point x^* , it can be modeled as the linear program

$$\begin{aligned} \max \quad & vb - (u\tilde{C} + v\tilde{C})x^*, \\ \text{s.t.} \quad & u^T \tilde{A} - v^T \tilde{A} = 0, \\ & u, v \in \mathbb{R}_+^m. \end{aligned}$$

This is the so-called cut-generating linear programming (henceforth denoted as CGLP) as opposed to the initial program denoted by LP) which needs an additional normalization constraint in order to be bounded. The CGLP has $2m$ variables and $n + p$ constraints and thus, has twice the size of the linear relaxation LP of the initial problem. It is therefore fairly impractical to solve it at each step in order to obtain just one cut. Balas and Perregaard [15] have however, shown that it is possible to mimic simplex pivots in CGLP by performing infeasible simplex pivots in LP. Actually, one simplex pivot in LP is not in one-to-one correspondence with a simplex pivot in CGLP but may correspond to many, often hundreds of (mostly degenerate) pivots in CGLP (see Ref. 16 for details and examples). More specifically, consider a row j in the optimal tableau LP

$$x_j = \bar{a}_{j0} - \sum_{l \in N} \bar{a}_{jl} x_l, \quad (2)$$

where N denotes the set of nonbasic variables. It can be shown that the Gomory mixed-integer cut generated from Equation (2) is a lift-and-project cut from $P_j(K)$. Therefore, there exists a basis of CGLP that corresponds to it. In other words, the optimal tableau of LP can be put in correspondence with a feasible basis of CGLP that represents a Gomory mixed-integer cut generated from Equation (2). Assume now that x_k enters the basis in row i ($i \neq j$) in LP. In the new basis, row j will become row j to which a multiple of row i is added, yielding a row of the type

$$x_j = \tilde{a}_{j0} - \sum_{l \in N} \tilde{a}_{jl} x_l. \quad (3)$$

The Gomory mixed-integer cut generated from Equation (3) is again a lift-and-project cut from $P_j(K)$ and corresponds to a basis of CGLP. Performing a pivot in LP therefore implicitly allows us to move to another basis in CGLP. The question is now how to choose the pivot a_{ik} in order to create a cut that is as strong as possible. There are two choices to be made: the row i and the column k . Improvement in the choice of row i is encoded through the reduced cost of u_i and v_i in CGLP but can be computed directly in LP. The choice of k can also be made from the data in LP in order to obtain a strong cut.

The size of CGLP is roughly twice that of LP and it is no wonder that one basis of CGLP is not in exact one-to-one correspondence to a basis of LP. It turns out that performing one pivot on LP often corresponds to several pivots in CGLP. Hence, it is much more efficient to generate lift-and-project cuts in LP as was also shown by computational evidence [15,16]. For more details on how to compute the reduced costs of u_i and v_i and how to select the variable x_k to pivot in, we refer to the papers by Balas and Perregaard [15] and by Balas and Bonami [16].

We also propose a geometric interpretation of why different bases of LP correspond to different lift-and-project cuts. Recall that a lift-and-project inequality can equivalently be seen as a disjunctive or a split cut using the disjunction $(x_j \leq 0) \vee (x_j \geq 1)$. These cuts can in turn, be interpreted as intersection cuts and are determined by the intersection

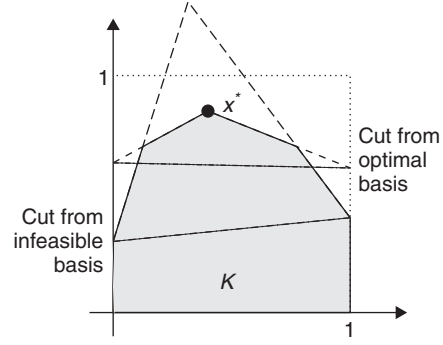


Figure 3. Geometric interpretation of lift-and-project cuts from feasible and infeasible bases.

of the rays of the bases with either $x_j = 0$ or $x_j = 1$. Figure 3 shows an example of two cuts determined from two different bases. The infeasible basis actually generates the stronger cut.

Balas and Jeroslow [17] have proposed a general method to strengthen valid inequalities of MIPs. This can also be applied in general to lift-and-project inequalities. Indeed, when we generate lift-and-project cuts, we only consider the integrality of x_j . We may also want to consider the integrality of other variables $x_i, i \neq j$. This may allow us to decrease their coefficients in the cut. This strengthening has the desirable property that it is extremely easy to incorporate once the CGLP has been solved.

Theorem 3. Consider the lift-and-project cut $\sum_{i=1}^{n+p} \alpha_i x_i \geq \beta$ obtained from the solution (u, v) of CGLP such that

$$\alpha_k = \begin{cases} \max(u^T A_k, v^T A_k) & k \neq j \\ \max(u^T A_j - u_0, v^T A_j + v_0) & k = j, \end{cases}$$

where A_k denotes the k^{th} column of A and $\beta = \min(u^T b, v^T b + u_0)$. Define $m_k = \frac{v^T A_k - u^T A_k}{u_0 + v_0}$ and

$$\bar{\alpha}_k = \begin{cases} \max(u^T A_k + u_0 \lceil m_k \rceil, v^T A_k - v_0 \lfloor m_k \rfloor) & k = 1, \dots, n, k \neq j \\ \max(u^T A_j - u_0, v^T A_j + v_0) & k = j, \\ \max(u^T A_k, v^T A_k) & k = n+1, \dots, p \end{cases}$$

and $\bar{\beta} = \min(u^T b, v^T b + v_0)$. Then, the inequality $\bar{\alpha}x \geq \bar{\beta}$ is valid for K_I and is called a strengthened lift-and-project cut.

Proof. See Ref. 17.

We close this survey article with some numerical tests that have been carried out using lift-and-project inequalities. There exist two commercial implementations of lift-and-project cuts (XPRESS and MOPS) and one open-source implementation Cgl-LandP [18] within the COIN-OR framework [19]. In an initial experiment, Bonami and Minoux [20] experimented on 35 mixed binary problems from the MIPLIB 3 library [21]. They have generated a series of lift-and-project inequalities in order to evaluate the gap closed by this family of cuts. Out of the 35 MIPLIB instances, they have found that the lift-and-project closure closes 37% of the integrality gap on average. This result needs to be moderated by the very large variance of the gap closed. Indeed, in 8 out of 35 instances, the gap closed is lower than 1%, whereas in 4 out of the 35 instances the gap closed is larger than 90%. By using strengthened lift-and-project cuts, the average gap closed increases by 8% to reach 45% gap closed on average. But again on 7 instances, the gap closed is lower than 1% and on 5 instances, the gap closed is larger than 90%.

More recently, Balas and Bonami [16] experimented on MIPLIB instances using the CglLandP cut generator within the COIN-OR framework. In this framework, the lift-and-project cuts are generated as a strengthening of the mixed-integer Gomory cuts by performing infeasible pivots in the optimal LP tableau. They compare the time needed to solve instances using either a cut-and-branch algorithm with mixed-integer Gomory cuts or a cut-and-branch algorithm with strengthened lift-and-project inequalities. For the easy instances, solved in less than 30 s, no major difference can be seen. However for the other instances, the time needed to solve the problems to optimality using the lift-and-project cuts can be drastically reduced compared to an implementation using Gomory cuts only. On average, the instances are solved in less than a third of the time when using the best implementation.

REFERENCES

1. Lovász L, Schrijver A. Cones of matrices and set-functions and 0-1 optimization. *SIAM J Optim* 1991;28:166–190.
2. Sherali H, Adams W. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J Discrete Math* 1990;3:311–430.
3. Balas E, Ceria S, Cornuéjols G. A lift-and-project cutting plane algorithm for mixed 0–1 programs. *Math Program* 1993;58:295–324.
4. Balas E, Ceria S, Cornuéjols G. Mixed 0-1 programming by lift-and-project in a branch-and-cut framework. *Manage Sci* 1996;42:1229–1246.
5. Cornuéjols G. Valid inequalities for mixed integer linear programs. *Math Program B* 2008;112:3–44.
6. Balas E, Perregaard M. Lift-and-project for mixed 0–1 programming: recent progress. *Discrete Appl Math* 2002;123:129–154.
7. Schrijver A. *Theory of linear and integer programming*. New York: Wiley; 1986.
8. Nemhauser G, Wolsey L. *Integer and combinatorial optimization*. New York: Wiley; 1988.
9. Bertsimas D, Weismantel R. *Optimization over integers*. Belmont (MA): Dynamic Ideas; 2005.
10. Balas E. Disjunctive programming: properties of the convex hull of feasible points, GSIA Management Science Report MSRR, 348, 1974: published as invited paper in. *Discrete Appl Math* 1998;89:1–44.
11. Gomory R. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64:275–278.
12. Nemhauser G, Wolsey L. A recursive procedure to generate all cuts for 0–1 mixed integer programs. *Math Program* 1990;46:379–390.
13. Cook W, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program* 1990;47:155–174.
14. Balas E. Intersection cuts - A new type of cutting planes for integer programming. *Oper Res* 1971;19:19–39.
15. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer gomory cuts for 0–1 programming. *Math Program B* 2003;94:221–245.
16. Balas E, Bonami P. Generating lift-and-project cuts from the LP simplex tableau: open source implementation and testing of

- new variants. *Math Program Comput* 2009;1: 165–199.
17. Balas E, Jeroslow R. Strengthening cuts for mixed-integer programs. *Eur J Oper Res* 1980;4:224–234.
 18. CglLandP. Available at <https://projects.coin-or.org/Cgl/wiki/CglLandP>
 19. COIN-OR. Available at <http://www.coin-or.org>
 20. Bonami P, Minoux M. Using rank-1 lift-and-project closures to generate cuts for 0–1 MIPs, a computational investigation. *Discrete Optim* 2005;2:288–307.
 21. Bixby R, Ceria S, McZeal C, *et al.* An updated mixed integer programming library: MIPLIB 3.0. *Optima* 1998;58:12–15.

LIFTING TECHNIQUES FOR MIXED INTEGER PROGRAMMING

JEAN-PHILIPPE P. RICHARD
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

INTRODUCTION

We consider mixed integer programs (MIPs) of the form

$$\max \left\{ f'x + g'y \mid Ax + By \leq d, \right. \\ \left. (x, y) \in \mathbb{Z}^m \times \mathbb{R}^n \right\}, \quad (1)$$

where $A \in \mathbb{Q}^{p \times m}$, $B \in \mathbb{Q}^{p \times n}$, $d \in \mathbb{Q}^p$, $f \in \mathbb{R}^m$, and $g \in \mathbb{R}^n$, and where bounds on the variables are treated as constraints. We denote the feasible region of problem (1) by $\mathcal{F}$. We also define $M = \{1, \dots, m\}$ and $N = \{1, \dots, n\}$. Given a set $M' \subseteq M$ (respectively, $N' \subseteq N$), we use the expression *variables in M'* (respectively, *in N'*) to describe the set of variables x_i for $i \in M'$ (respectively, y_j for $j \in N'$).

In both cutting-plane and branch-and-cut algorithms, an important practical problem is that of deriving strong valid inequalities for $\text{conv}(\mathcal{F})$. A key difficulty in the derivation of such inequalities is that the number of variables is large and therefore a direct geometric analysis of the polyhedral structure of $\text{conv}(\mathcal{F})$ is complex. Lifting is a method that can be used to circumvent this difficulty. It is the process by which a valid inequality for a restriction of the problem obtained by temporarily fixing variables to specific values (a *seed inequality*) is transformed into a valid inequality for the entire problem.

Example 1 [1]. Consider the mixed integer set

$$\tilde{\mathcal{F}} = \{(x_1, x_2, y_1) \in \mathbb{Z}^2 \times \mathbb{R} \mid x_1 + 2x_2 + 2y_1 \leq 3, \\ 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, 0 \leq y_1 \leq 1\}.$$

The set $\tilde{\mathcal{F}}$ is presented in Fig. 1a and its convex hull is depicted in Fig. 1b. It can be verified that

$$\text{conv}(\tilde{\mathcal{F}}) = \{(x_1, x_2, y_1) \in \mathbb{Z}^2 \times \mathbb{R} \mid x_1 + 2x_2 \\ + 2y_1 \leq 3, x_2 + 2y_1 \leq 2, \\ 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1, y_1 \geq 0\}.$$

In particular, adding $x_2 + 2y_1 \leq 2$ to the inequalities defining $\tilde{\mathcal{F}}$ yields a linear description of $\text{conv}(\tilde{\mathcal{F}})$.

We now illustrate how lifting can be used to derive $x_2 + 2y_1 \leq 2$. First, we fix the variable y_1 to 1. The resulting set is $\tilde{\mathcal{F}}_{\{y_1=1\}} = \{(x_1, x_2, y_1) \in \{0, 1\}^2 \times [0, 1] \mid x_1 + 2x_2 \leq 1, y_1 = 1\}$. $\tilde{\mathcal{F}}_{\{y_1=1\}}$ is a slice of the feasible region. The seed inequality

$$x_2 \leq 0 \quad (2)$$

is valid for $\tilde{\mathcal{F}}_{\{y_1=1\}}$. The plane associated with inequality (2) is represented in Fig. 2a, where it can be seen that, although valid for $\tilde{\mathcal{F}}_{\{y_1=1\}}$, inequality (2) is not valid for $\tilde{\mathcal{F}}$. Lifting the variable y_1 is to determine a coefficient β for y_1 for which

$$x_2 + \beta(y_1 - 1) \leq 0 \quad (3)$$

is valid for $\tilde{\mathcal{F}}$. Note that the form $\beta(y_1 - 1)$ in inequality (3) is not innocuous as it guarantees that, when $y_1 = 1$, inequality (3) reduces to seed inequality (2). In this example, it is easy to verify that coefficients β , which ensure that inequality (3) is valid for $\tilde{\mathcal{F}}$ should be sufficiently large so as to guarantee that the point $(0, 1, \frac{1}{2})$ is feasible. This naturally leads to the conclusion that $1 + \beta(\frac{1}{2} - 1) \leq 0$, that is, $\beta \geq 2$. Because we are interested in generating the strongest possible inequality, we choose $\beta = 2$. Substituting β for 2 in inequality (2), we obtain $x_2 + 2y_1 \leq 2$, an inequality that defines a facet of $\text{conv}(\tilde{\mathcal{F}})$. The lifting coefficient β could also have been obtained algebraically as follows. First rewrite inequality (3) as $\beta(y_1 - 1) \leq -x_2$. Then observe that β is a

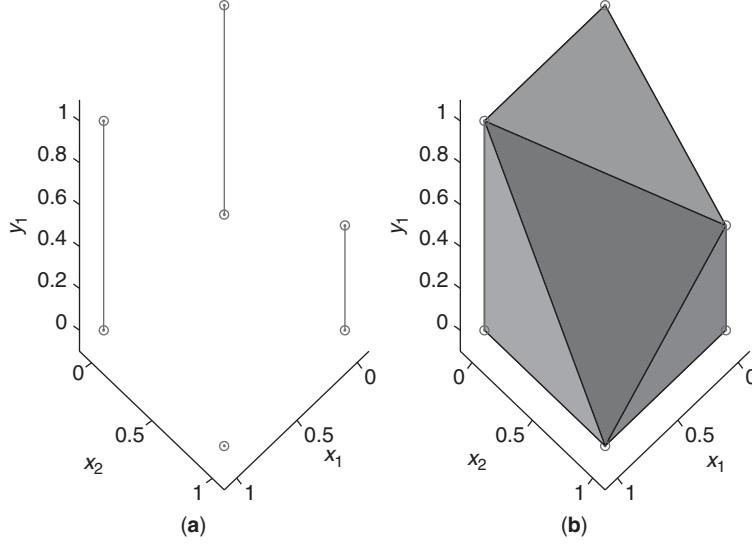


Figure 1. S and $\text{conv}(S)$ in Example 1.

valid lifting coefficient if

$$\beta \geq \max \left\{ \frac{-x_2}{y_1 - 1} \mid x_1 + 2x_2 + 2y_1 \leq 3, \right. \\ \left. x_1 \in \{0, 1\}, x_2 \in \{0, 1\} \text{ and } y_1 \in [0, 1] \right\}. \quad (4)$$

The optimal solution of this problem is $(x_1, x_2, y_1) = (0, 1, \frac{1}{2})$ yielding $\beta \geq 2$. For reasons that will become clear in the section titled “Computing and Recomputing the Lifting Function”, problem (4) is often reformulated as $\max \left\{ \frac{\tilde{\Psi}(2(y_1-1))}{y_1-1} \mid y_1 \in [0, 1] \right\}$, where $\tilde{\Psi}(z) = -\max\{x_2 \mid x_1 + 2x_2 \leq 1 - z, x_1 \in \{0, 1\}, x_2 \in \{0, 1\}\}$ is said to be the lifting function associated with inequality (2).

Example 1 also provides a geometric interpretation of lifting. First observe that the seed inequality (2) intersects the plane $y_1 = 1$ on the thick line represented in Fig. 2a. We refer to this line as *hinge*. Using this terminology, lifting corresponds to rotating the seed inequality plane around its hinge until it becomes valid for $\tilde{\mathcal{F}}$. Such rotation is not always possible as the following example illustrates.

Example 2 [1]. Consider the set $\tilde{\mathcal{F}}$ of Example 1. The inequality $x_1 + x_2 \leq 1$ is

valid for $\tilde{\mathcal{F}}_{\{y_1 = \frac{1}{2}\}}$, where y_1 is fixed to $\frac{1}{2}$.

To perform lifting, we must determine a coefficient β for which $x_1 + x_2 + \beta(y_1 - \frac{1}{2}) \leq 1$ is valid for $\tilde{\mathcal{F}}$. Because $(1, 1, 0)$ belongs to $\tilde{\mathcal{F}}$, it is clear that $\beta \geq 2$. Similarly since $(1, 0, 1)$ belongs to $\tilde{\mathcal{F}}$, $\beta \leq 0$. In other words, no matter how the plane $x_1 + x_2 = 1$ is rotated around its hinge, one of the two feasible points $(1, 0, 1)$ or $(1, 1, 0)$ will be cut off by the lifted inequality: lifting is impossible.

Geometrically, the reason that lifting is impossible in Example 2 is that there are feasible solutions on both sides of the plane $y_1 = \frac{1}{2}$. Further, the plane corresponding to the seed inequality must be rotated in one direction to satisfy the feasible points with $y_1 < \frac{1}{2}$, while it needs to be rotated in the other direction to satisfy the feasible points with $y_1 > \frac{1}{2}$. This does not happen in Example 1 since all feasible points are on one side of the plane $y_1 = 1$.

HISTORY OF LIFTING

The idea of extending linear functions from a linear subspace of a vector space to the vector space itself can be traced back to

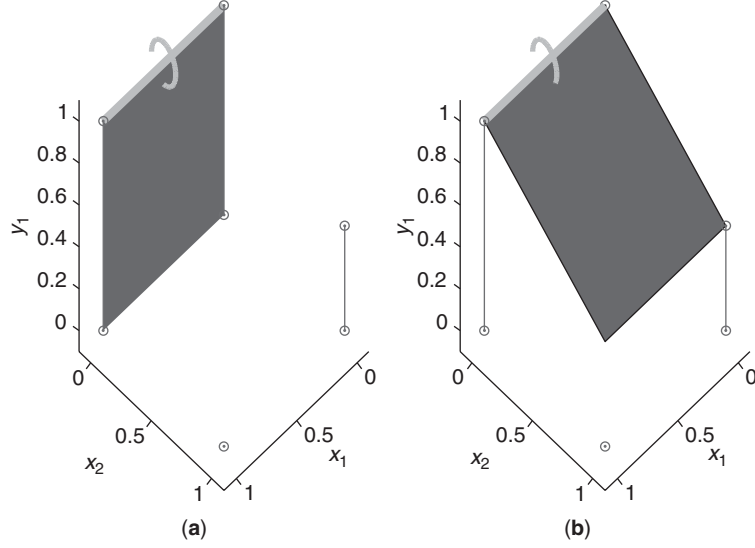


Figure 2. Lifting seed inequality in Example 1.

Hahn–Banach’s extension theorem. For cutting planes in MIP, the ideas underlying lifting were first used by Gomory [2] for integer programs with feasible regions of the form:

$$\text{MG}(G, r) = \left\{ x \in \mathbb{Z}_+^{|G|} \mid \sum_{g \in G} g x_g \equiv r \pmod{1} \right\}, \quad (5)$$

where G is a finite group and $r \in G$. In particular, Gomory proved that strong valid inequalities for $\text{MG}(G, r)$ can be “lifted” into strong valid inequalities for $\text{MG}(G', r)$, where G' is a finite group that contains G as a subgroup [[2], Theorem 19].

In its current acceptance, lifting surfaced in the mid-1970s when it was used to derive facet-defining inequalities for various structured MIPs including knapsack problems, independence system problems, and vertex packing problems [3–8]. Lifting was recognized to be a generic computational tool to generate strong inequalities for 0-1 integer programs in Padberg [9]; see also Zemel [10]. For general unstructured IPs, lifting was formalized in Wolsey [11]. Further, Wolsey [12] introduced the concept of superadditive lifting that is now central to the practical use of lifting. The use of lifted inequalities

was first shown to be beneficial in solving unstructured 0-1 IPs in Crowder *et al.* [13]. Interest in lifting theory for unstructured MIPs was revived by Gu *et al.* [14], where superadditive conditions for efficient lifting were revisited and by Gu *et al.* [15], where the advantages of using lifting in software implementations were demonstrated. Although most early studies focused on lifting integer variables, the lifting of a single continuous variable in a knapsack constraint was studied in Marchand and Wolsey [16]. Sequential lifting for continuous variables was considered in Richard *et al.* [1,17]. Lifted inequalities have now been implemented in all leading commercial branch-and-cut software [18].

Although we center this presentation on lifting concepts that apply to all MIPs, lifting has been used to study many structured problems including 0-1 knapsack problems [5–7,19], variants of knapsack problems [1,16,17,20], knapsack models with side constraints [21–26], independence systems [4,27], set covering models [28], flow models [29,30], and the traveling salesman problem [31,32].

Finally, although initially devised for generating cutting planes in MIP, recent progress has been made in applying lifting to other settings. In particular, de Farias [33]

studied lifting for semicontinuous variables, Richard and Tawarmalani [34] extended lifting to general nonlinear settings, and Atamtürk and Narayanan [35] developed a more explicit lifting theory for conic MIPs. Specific applications of lifting for nonlinear problems also include de Farias *et al.* [36] and Lin and Vandenbussche [37].

FUNDAMENTALS OF LIFTING

We now give a general description of lifting in MIP adapting, restating and generalizing results in the papers [1,9,11,12,14,17,34,38].

Because lifting is typically performed in several rounds in which multiple variables can be lifted simultaneously, we rewrite problem (1) into a form that better captures the sequential nature of lifting. First, we divide the set of integer variables M into $(M_0, M_1, \dots, M_l)$ and the set of continuous variables N into $(N_0, N_1, \dots, N_l)$. We assume that the sets M_k are disjoint (but not necessarily nonempty) for $k = 0, \dots, l$ and that sets N_k are disjoint (but not necessarily nonempty) for $k = 0, \dots, l$. We also require that $|M_k| + |N_k| \geq 1$ for $k = 0, \dots, l$. We use the above notation to signify that, when deriving the lifted inequality, we first derive a seed inequality using only the variables in $M_0 \cup N_0$ and then perform l rounds of lifting, where the k th lifting round only involves variables in $M_k \cup N_k$. Second, we divide the set of constraints into $l + 2$ classes. The first class contains all constraints that relate variables belonging to different lifting rounds. The following $l + 1$ classes are composed of all constraints that relate variables belonging to a single lifting round. We mention that, in general, different *lifting orders*, that is, the way variables are arranged into the sets $M_0, M_1, \dots, M_l$ and $N_0, N_1, \dots, N_l$, result in different lifted inequalities; see the section titled “Superadditive Lifting” for an important exception.

Using the notation defined above, we rewrite problem (1) as

$$\mathcal{F} = \left\{ (x, y) \in \mathbb{Z}^m \times \mathbb{R}^n \mid \begin{array}{l} \sum_{k=0}^l \left(\sum_{i \in M_k} A_i^* x_i \right. \\ \quad \left. + \sum_{j \in N_k} B_j^* y_j \right) \leq d^* \\ \sum_{i \in M_k} A_i^k x_i + \sum_{j \in N_k} B_j^k y_j \\ \quad \leq d^k, \forall k = 0, \dots, l \end{array} \right\}.$$

In the above set, we denote by p^* the number of rows of the matrices A_i^* and B_j^* . In the remainder of the article, we often write x^k to represent the variables x_i for $i \in M_k$ and y^k to represent the variables y_j for $j \in N_k$. We also use the notation $m_k = |M_k|$ and $n_k = |N_k|$ for $k = 0, \dots, l$. We typically write $(x^0, x^1, \dots, x^l; y^0, y^1, \dots, y^l)$ to represent the variables x_i for $i \in M_0 \cup \dots \cup M_l$ and y_j for $j \in N_0 \cup \dots \cup N_l$. Finally, we define $\mathcal{P}^k := \left\{ (x^k, y^k) \in \mathbb{Z}^{m_k} \times \mathbb{R}^{n_k} \mid \sum_{i \in M_k} A_i^k x_i + \sum_{j \in N_k} B_j^k y_j \leq d^k \right\}$.

Next we describe the four key steps of the lifting process: (i) fixing variables, (ii) generating a seed inequality, (iii) computing/recomputing the lifting function, and (iv) deriving the lifting coefficients. Throughout the discussion, we use the following example to illustrate the steps of the lifting procedure.

Example 3. Consider an application where a supplier has a quantity C of a product that can be delivered to customers using n distribution channels. Channel i currently has a capacity of c_i units, but additional capacity can be installed in up to u_i discrete increments of Δ_i for $i = 0, \dots, n - 1$. The MIP formulation corresponding to this application is

$$\hat{\mathcal{F}} = \left\{ (x, y) \in \mathbb{Z}^n \times \mathbb{R}^n \mid \begin{array}{l} \sum_{j=0}^{n-1} y_j \leq C, 0 \leq y_j \\ \leq c_j + \Delta_j x_j, 0 \leq x_j \leq u_j, \forall j = 0, \dots, n - 1 \end{array} \right\}.$$

For this problem, it is clear that $A^* = (0, \dots, 0)$, $B^* = (1, \dots, 1)$, $d^* = C$, $A_j^j = (0, -\Delta_j, -1, 1)'$, $B_j^j = (-1, 1, 0, 0)'$ and $d^j = (0, c_j, 0, u_j)'$ for $j = 0, \dots, n - 1$.

Fixing Variables

The first step of the lifting process is to fix a subset of the variables to specific values. It is customary to fix these variables so that the problem restriction still contains some feasible solutions. Therefore, we assume that the fixing is consistent with a solution $(\bar{x}, \bar{y}) \in \mathcal{F}$ called *fixing point*, that is, variables x^k are

fixed to $\bar{x}^k$ and variables y^k are fixed to $\bar{y}^k$ for $k = 1, \dots, l$. The restriction of the set $\mathcal{F}$ we consider during the q th lifting round is

$$\mathcal{F}^q(\bar{x}, \bar{y}) = \{(x^0, \dots, x^q, y^0, \dots, y^q) \mid (x^0, \dots, x^q, \bar{x}^{q+1}, \dots, \bar{x}^l; y^0, \dots, y^q, \bar{y}^{q+1}, \dots, \bar{y}^l) \in \mathcal{F}\},$$

for $q = 0, \dots, l$. Similarly, we could have defined $\mathcal{F}^q(\bar{x}, \bar{y})$ explicitly as

$$\mathcal{F}^q(\bar{x}, \bar{y}) = \left\{ (x^0, \dots, x^q, y^0, \dots, y^q) \mid \begin{array}{l} \sum_{k=0}^q \left(\sum_{i \in M_k} A_i^* x_i \right. \\ \quad \left. + \sum_{j \in N_k} B_j^* y_j \right) \\ \leq d^*(q) \\ (x^k, y^k) \in \mathcal{P}^k, \\ \forall k = 0, \dots, q \end{array} \right\},$$

where $d^*(q) := d - \sum_{k=q+1}^l \left(\sum_{i \in M_k} A_i^* \bar{x}_i + \sum_{j \in N_k} B_j^* \bar{y}_j \right)$. Note that $\mathcal{F}^0(\bar{x}, \bar{y})$ contains only variables in M_0 and N_0 while $\mathcal{F}^l(\bar{x}, \bar{y}) = \mathcal{F}$.

Whether lifting is possible depends on both the seed inequality and the fixing point $(\bar{x}, \bar{y})$. In practice, $(\bar{x}, \bar{y})$ is often chosen to belong to the boundary of $\text{conv}(\mathcal{F})$, although this choice does not guarantee that lifting is possible; see Example 2. Another popular choice is to select the components of $(\bar{x}, \bar{y})$ to be upper or lower bounds on the corresponding variables. This fixing point ensures that lifting is possible, as the discussion following Example 2 suggests.

When $\mathcal{F}^q(\bar{x}, \bar{y})$ is not full-dimensional, its valid inequalities admit multiple equivalent representations. It then becomes typically difficult to determine what representation ultimately leads to strong lifted inequalities for $\text{conv}(\mathcal{F})$. For this reason, we will assume that lifting rounds and fixing points $(\bar{x}, \bar{y})$ are chosen in such a way that the sets $\mathcal{F}^q(\bar{x}, \bar{y})$ are full-dimensional. This assumption is consistent with the way inequalities are derived in practice.

Example 4. Consider an instance of the problem of Example 3 where $n = 5$, $C = 28$, $c_0 = c_1 = 3$, $c_i = i$ for $i = 2, \dots, 4$, $u_0 = u_1 = \infty$, $u_i = i - 1$ for $i = 2, \dots, 4$ and $\Delta_i = 2$ for $i = 0, \dots, 4$. The solution

$(\bar{x}; \bar{y}) = (0, 0, 1, 2, 3; 0, 3, 4, 7, 10)$ is feasible for $\hat{\mathcal{F}}$. We have

$$\hat{\mathcal{F}}^0(\bar{x}, \bar{y}) = \{(x_0, y_0) \in \mathbb{Z} \times \mathbb{R} \mid y_0 \leq 4, 0 \leq y_0 \leq 3 + 2x_0, x_0 \geq 0\}. \quad (6)$$

Generating a Seed Inequality

The second step of the lifting procedure is to derive a seed inequality

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j \leq \delta \quad (7)$$

that is valid for $\mathcal{F}^0(\bar{x}, \bar{y})$. Seed inequalities are often obtained through simple combinatorial, logical or even lifting arguments. Although the only strict requirement posed on inequality (7) is that it is valid for $\mathcal{F}^0(\bar{x}, \bar{y})$, it is typically chosen to define a face of large dimension of $\text{conv}(\mathcal{F}^0(\bar{x}, \bar{y}))$. This choice is made because strong inequalities remain strong if lifting is performed exactly; see Proposition 6 and ensuing paragraphs. We mention, however, that this choice does not guarantee that lifting is possible; see Example 2. In the remainder of this article, we assume that the face of $\text{conv}(\mathcal{F}^0(\bar{x}, \bar{y}))$ that inequality (7) defines is nonempty.

Example 5. Consider the set $\hat{\mathcal{F}}^0(\bar{x}, \bar{y})$ of Example 4. A linear inequality description of $\text{conv}(\hat{\mathcal{F}}^0(\bar{x}, \bar{y}))$ is given by $y_0 \leq 3 + x_0$, $y_0 \leq 4$, $y_0 \geq 0$, and $x_0 \geq 0$. We therefore choose our seed inequality to be

$$y_0 \leq 3 + x_0. \quad (8)$$

The above first two steps of lifting are performed once. The following two will be repeated l times.

Computing and Recomputing the Lifting Function

Because lifting is a sequential process, we assume here that we already have lifted variables in $M_1 \cup N_1, M_2 \cup N_2, \dots, M_{q-1} \cup N_{q-1}$

for some $q \in \{1, \dots, l\}$ and have generated the lifted seed inequality

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{k=1}^{q-1} \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \leq \delta. \quad (9)$$

We assume that lifting was possible in each of these steps and that inequality (9) is valid for $\mathcal{F}^{q-1}(\bar{x}, \bar{y})$. We are interested in lifting variables x_i for $i \in M_q$ and y_j for $j \in N_q$ into inequality (9), that is, we are interested in finding coefficients α_i for $i \in M_q$ and β_j for $j \in N_q$ for which

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{k=1}^q \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \leq \delta$$

is valid for $\mathcal{F}^q(\bar{x}, \bar{y})$. To this end, we introduce the q th lifting function $\Psi^q : \mathbb{R}^{p^*} \rightarrow \mathbb{R}$ as

$$\begin{aligned} \Psi^q(z) = & \delta \\ & - \max \left\{ \sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j \right. \\ & \left. + \sum_{k=1}^{q-1} \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \right\} \\ \text{s.t. } & \sum_{k=0}^{q-1} \left(\sum_{i \in M_k} A_i^* x_i + \sum_{j \in N_k} B_j^* y_j \right) \\ & + \sum_{k=q}^l \left(\sum_{i \in M_k} A_i^* \bar{x}_i + \sum_{j \in N_k} B_j^* \bar{y}_j \right) \leq d - z \\ & (x^k, y^k) \in \mathcal{P}^k, \forall k = 0, \dots, q-1, \end{aligned}$$

where we set $\Psi^q(z) = \infty$ if the optimization problem is infeasible. This definition is a direct generalization of the function we introduced in Example 1 to reformulate problem (4). For each vector $z \in \mathbb{R}^{p^*}$, the lifting function represents the amount of slack there is in inequality (9) when the variables in

$M_0, \dots, M_{q-1}$ and $N_0, \dots, N_{q-1}$ are allowed to contribute $d^*(q-1) - z$ to the left-hand side of the problem constraints.

Example 6. The lifting function associated with inequality (8) is defined as $\hat{\Psi}^1(z) = 3 - \max\{y_0 - x_0 \mid y_0 \leq 4 - z, 0 \leq y_0 \leq 3 + 2x_0, x_0 \in \mathbb{Z}_+\}$ and is represented in Fig. 3a. It can be verified that $\hat{\Psi}^1(z) = \max\{\lfloor \frac{z}{2} \rfloor, z - \lceil \frac{z}{2} \rceil\}$ when $z \leq 1$, $\hat{\Psi}^1(z) = z - 1$ when $1 \leq z \leq 4$ and $\hat{\Psi}^1(z) = \infty$ when $z > 4$.

Because inequality (9) is valid for $\mathcal{F}^{q-1}(\bar{x}, \bar{y})$, $\Psi^q(z) \geq 0$ for $z \in \mathbb{R}_+^{p^*}$ and $\Psi^q(0) = 0$ if inequality (9) defines a nonempty face of $\text{conv}(\mathcal{F}^{q-1}(\bar{x}, \bar{y}))$. Further, using the fact that the feasible set of the problem defining $\Psi^q(z)$ is a subset of those defining $\Psi^{q+1}(z)$ and $\Psi^q(z - \epsilon)$ for $\epsilon \in \mathbb{R}_+^{p^*}$, $\Psi^q(z)$ is nonincreasing in q and nondecreasing in z .

Proposition 1. For $q \in \{1, \dots, l\}$, $\Psi^q(z) \geq 0$ for $z \in \mathbb{R}_+^{p^*}$. Further, (i) for $q_1, q_2 \in \{1, \dots, l\}$ with $q_1 \leq q_2$, we have $\Psi^{q_1}(z) \geq \Psi^{q_2}(z)$ for all $z \in \mathbb{R}^{p^*}$ and (ii) for $z_1, z_2 \in \mathbb{R}^{p^*}$ with $z_1 \leq z_2$, we have $\Psi^q(z_1) \leq \Psi^q(z_2)$ for $q = 1, \dots, l$.

To conclude our discussion of lifting functions, we mention that if inequality (9) is valid for $\mathcal{F}^{q-1}(\bar{x}, \bar{y})$, then

$$\begin{aligned} & \sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{k=1}^{q-1} \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) \right. \\ & \quad \left. + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \\ & \quad + \Psi^q \left(\sum_{k=q}^r \left(\sum_{i \in M_k} A_i^* (x_i - \bar{x}_i) \right. \right. \\ & \quad \left. \left. + \sum_{j \in N_k} B_j^* (y_j - \bar{y}_j) \right) \right) \leq \delta \quad (10) \end{aligned}$$

is a valid nonlinear inequality for $\mathcal{F}^r(\bar{x}, \bar{y})$ for all $r = q, \dots, l$.

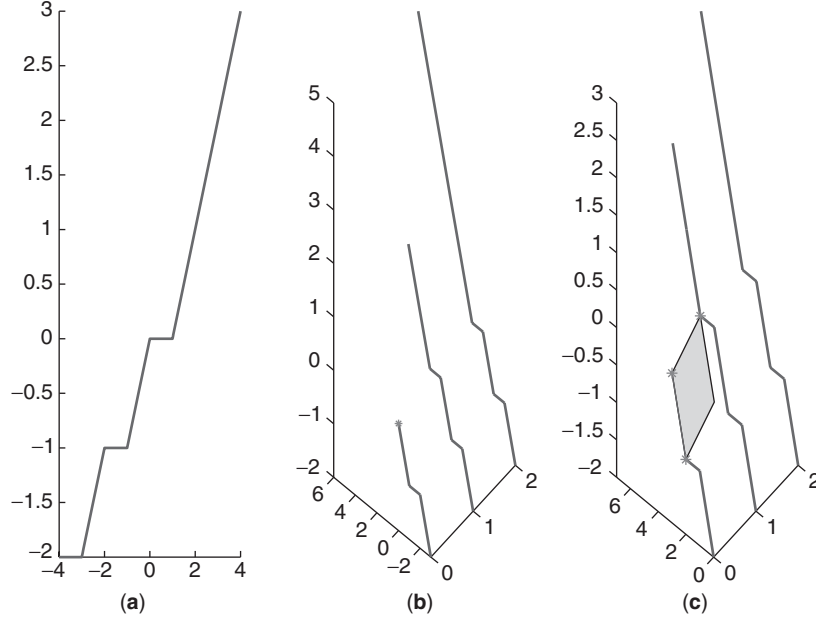


Figure 3. Lifting functions and problems for Example 6.

Deriving Lifting Coefficients

In order for coefficients α_i for $i \in M_q$ and $\beta_j \in N_q$ to yield a valid inequality for $\mathcal{F}^q(\bar{x}, \bar{y})$, these coefficients must be chosen so that

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{k=1}^q \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \leq \delta \quad (11)$$

is satisfied by all points $(x^0, \dots, x^q; y^0, \dots, y^q)$ in $\mathcal{F}^q(\bar{x}, \bar{y})$. This requirement can also be expressed as

$$\begin{aligned} & \sum_{i \in M_q} \alpha_i (x_i - \bar{x}_i) + \sum_{j \in N_q} \beta_j (y_j - \bar{y}_j) \\ & \leq \Psi^q \left(\sum_{i \in M_q} A_i^* (x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^* (y_j - \bar{y}_j) \right) \end{aligned} \quad (12)$$

for all $(x^q, y^q) \in \mathcal{P}^q$. Condition (12) only involves the variables (x^q, y^q) being lifted.

Geometrically, condition (12) requires that the hyperplane with normal vector (α^q, β^q) created in the space of lifted variables (x^q, y^q) lower approximates $\Psi^q \left(\sum_{i \in M_q} A_i^* (x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^* (y_j - \bar{y}_j) \right)$ over $\mathcal{P}^q$. In order to obtain a strong inequality, we should search for coefficients (α^q, β^q) that define a hyperplane that supports this function at as many affinely independent points as possible. It is clear that in general, several choices for (α^q, β^q) are admissible. Because we have established in Proposition 1 that lifting functions Ψ^q are nonincreasing in q , conditions (12) also imply that if we delay the lifting of a variable from one round to the next, its lifting coefficients will be no stronger than those it would have received if it had been lifted earlier.

Example 7. To compute lifting coefficients for (x_1, y_1) we first display the function $\hat{\Psi}^1(y_1 - 3)$ over part of $\hat{\mathcal{P}}^1 = \{(x_1, y_1) \in \mathbb{Z}_+ \times \mathbb{R}_+ \mid y_1 \leq 3 + 2x_1\}$ in Fig. 3b, where the horizontal axes are x_1 and $y_1 - 3$. We must determine coefficients α_1, β_1 for which the plane $\alpha_1 x_1 + \beta_1 (y_1 - 3)$

underestimates $\hat{\Psi}^1(y_1 - 3)$ over $\hat{\mathcal{P}}^1$. Choosing $\alpha_1 = -1$, $\beta_1 = 1$ yields such a plane that passes through the points $(0, 3, 0)$, $(0, 2, -1)$ and $(1, 4, 0)$. This plane is represented in Fig. 3c where the horizontal axes are x_1 and y_1 . The lifted inequality is therefore

$$y_0 + y_1 \leq 6 + x_0 + x_1. \quad (13)$$

We discussed in the section titled “Computing and Recomputing the Lifting Function” that lifting functions can be used to construct valid nonlinear inequalities of the form (10). Requirement (12) can therefore also be interpreted as a condition that allows the linearization of the nonlinear inequality (10) when $r = q$ into valid inequality (11).

PRACTICAL AND COMPUTATIONAL ASPECTS OF LIFTING

The main computational burdens associated with lifting are the computation of $\Psi^q(z)$ for $q = 1, \dots, l$ and the derivation of the associated lifting coefficients. We address these computational aspects next.

About Lifting Functions

Computing Lifting Functions. We introduced lifting functions as a medium by which valid lifting coefficients could be obtained through the study of system (12). We mention first that, although convenient for presentation purposes, the computation of lifting functions is not always necessary in lifting. For instance, assume that binary variable x_r is being lifted as a single variable in the q th round and that $\bar{x}_r = 0$. Then condition (12) reduces to $\alpha_r \leq \Psi^q(A_r)$, a system that can be set up with the solution of the single MIP defining $\Psi^q(A_r)$. In such an instance, computing the lifting function over a large range of z values might neither be necessary nor computationally advantageous. This observation carries over to situations where $\mathcal{P}^q$ is a simple set.

When $\mathcal{P}^q$ is complex, it is common to compute the value of $\Psi^q(z)$ over a subset of $\mathbb{R}^p$ that contains $\left\{ \sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \mid (x^q, y^q) \in \mathcal{P}^q \right\}$

(or the smaller subset that will be described in the section titled “General Techniques for Deriving Lifting Coefficients”). In specific situations, lifting functions can be obtained in closed form either because the seed inequality has few variables or because the structure of the seed inequality is simple; see Marchand and Wolsey [16] for an example. However, since lifting functions may change with every set of variables that is reintroduced, lifting functions Ψ^q are typically not easy to describe in closed form even if Ψ^1 is known in closed form.

A way to derive (and store) lifting functions is to use dynamic programming (DP):

$$\begin{aligned} \Psi^{q+1}(z) = \min_{(x^q, y^q) \in \mathcal{P}^q} & \left\{ \Psi^q \left(z + \sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) \right. \right. \\ & + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \left. \right) - \sum_{i \in M_q} \alpha_i(x_i - \bar{x}_i) \\ & \left. - \sum_{j \in N_q} \beta_j(y_j - \bar{y}_j) \right\}. \end{aligned} \quad (14)$$

When the number of constraints defining $\mathcal{F}$ is small and the coefficients of variables are not large, the recursion (14) or variants can sometimes be used to efficiently obtain the lifting function $\Psi^{q+1}(z)$ from $\Psi^q(z)$. This is, for example, the case when lifting cover inequalities for the 0-1 knapsack polytope [[40], Section II.2.2]. However, in most situations, recurrence (14) does not yield an efficient algorithm for computing lifting functions as its state space might be infinite and even when it is not, a pseudopolynomial amount of computation is typically required to obtain $\Psi^{q+1}(z)$ from $\Psi^q(z)$.

Example 8. For Example 7, we compute

$$\begin{aligned} \hat{\Psi}^2(z) = \min & \{ \hat{\Psi}^1(z + (y_1 - 3)) + x_1 - (y_1 - 3) \mid \\ & 0 \leq y_1 \leq 3 + 2x_1, x_1 \geq 0, x_1 \in \mathbb{Z}, y_1 \in \mathbb{R} \}, \end{aligned}$$

that is, $\hat{\Psi}^2(z) = \max\{\lfloor \frac{z}{2} \rfloor, z - \lceil \frac{z}{2} \rceil\}$ when $z \leq 1$, $\hat{\Psi}^2(z) = z - 1$ when $1 \leq z \leq 7$ and $\hat{\Psi}^2(z) = \infty$ when $z > 7$.

Because computing lifting functions can be time-consuming, it is sometimes advantageous to compute their values approximately. In particular, as long as a lower approximation $\psi^q(z)$ of $\Psi^q(z)$ is obtained, the lifting coefficients derived from the analysis of condition (12), where Ψ^q is replaced with $\psi^q(z)$ will be valid. An option to obtain $\psi^q(z)$ is to solve the linear programming (LP) relaxation of $\Psi^q(z)$ and round the resulting value up if the objective value is known to be integral. Similar ideas were shown to be computationally useful in Crowder *et al.* [13] and Gu *et al.* [15] in the case of 0-1 knapsack sets. LP-based lifting has also been studied in Dey and Richard [41].

Superadditive Lifting. The main computational difficulty with lifting functions is that $\Psi^{q+1}(z)$ and $\Psi^q(z)$ are typically different. For practical purposes, it is therefore interesting to identify conditions under which $\Psi^1(z) = \Psi^2(z) = \dots = \Psi^l(z)$ for all vectors z in an appropriately chosen subset S of $\text{dom}(\Psi^1(z))$. Such conditions were discovered by Wolsey [12]; see also Gu *et al.* [14]. To motivate their introduction, we consider the case where we lift a single binary variable x_r that was fixed at 0 in the q th round of lifting. In this case, the DP recursion (14) takes the form

$$\begin{aligned}\Psi^{q+1}(z) &= \min_{x_r \in \{0,1\}} \{ \Psi^q(z + A_r^* x_r) - \Psi^q(A_r^* x_r) \} \\ &= \min \{ \Psi^q(z), \Psi^q(z + A_r^*) - \Psi^q(A_r^*) \},\end{aligned}$$

since it follows from our discussion in the section titled “Computing Lifting Functions” that $\alpha_r = \Psi(A_r^*)$ is a strong valid lifting coefficient for x_r when $\Psi(A_r^*) < \infty$. In order for $\Psi^{q+1}(z)$ to be equal to $\Psi^q(z)$, it suffices to require that $\Psi^q(z) \leq \Psi^q(z + A_r^*) - \Psi^q(A_r^*)$ for all vectors z . More generally, given a function $\phi : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\pm\infty\}$, we say that ϕ is (S_1, S_2) -superadditive if $\phi(A) + \phi(B) \leq \phi(A + B)$ for all $A \in S_1$, and $B \in S_2$. When $S_1 = S_2$, we say that ϕ is superadditive over S_1 .

Proposition 2. *Define*

$$\begin{aligned}T^q &:= \left\{ \sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \mid \right. \\ &\quad \left. (x^q, y^q) \in \mathcal{P}^q \right\}.\end{aligned}$$

Assume that (i) $T^q \subseteq \text{dom}(\Psi^q)$, (ii) $\Psi^q(z)$ is (T^q, S) -superadditive, and (iii) lifting coefficients (α^q, β^q) for variables (x^q, y^q) exist that satisfy condition (12). Then $\Psi^{q+1}(z) = \Psi^q(z)$ for all $z \in S$.

The proof of Proposition 2 uses recursion (14) to argue that $\Psi^{q+1}(z) \geq \Psi^q(z)$ for all $z \in S$ and Proposition 1 to show that equality holds. Applying Proposition 2 sequentially, we conclude that lifting coefficients for all lifting rounds can be obtained from Ψ^1 if Ψ^1 is superadditive over an appropriate domain.

Proposition 3. Assume that (i) $T^q \subseteq \text{dom}(\Psi^1)$ for all $q = 1, \dots, l$, (ii) Ψ^1 is $(T_q, T_{q+1} + \dots + T_l)$ -superadditive for all $q = 1, \dots, l$ and (iii) lifting coefficients (α^q, β^q) , for variables (x^q, y^q) exist that satisfy condition (12) in which Ψ^q is replaced with Ψ^1 . Then

$$\begin{aligned}\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{k=1}^l \left(\sum_{i \in M_k} \alpha_i (x_i - \bar{x}_i) \right. \\ \left. + \sum_{j \in N_k} \beta_j (y_j - \bar{y}_j) \right) \leq \delta \quad (15)\end{aligned}$$

is a valid inequality for $\text{conv}(\mathcal{F})$.

Proposition 3 can be proven by induction showing that $\Psi^q(z) = \Psi^1(z)$ for all z in $T_q + \dots + T_l$ using Proposition 2. Another way of deriving the result follows by realizing that assumption (ii) implies that

$$\begin{aligned}\sum_{q=1}^l \Psi^1 \left(\sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \right) \\ \leq \Psi^1 \left(\sum_{q=1}^l \left(\sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) \right. \right. \\ \left. \left. + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \right) \right). \quad (16)\end{aligned}$$

Using this observation in inequality (10), where $i = 1$, we obtain the following relaxed

nonlinear inequality

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{q=1}^l \times \Psi^1 \left(\sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \right) \leq \delta$$

that can be linearized into inequality (15) using assumptions (i) and (iii).

In practice, Proposition 3 is often applied by selecting S_1 and S_2 to be small supersets of the sets $\mathcal{T}^1 \cup \dots \cup \mathcal{T}^l$ and $\mathcal{T}^1 + \dots + \mathcal{T}^l$ and then by verifying that Ψ^1 is (S_1, S_2) -superadditive so that assumption (ii) in Proposition 3 is naturally satisfied. In such a situation, the same lifted inequalities will be obtained no matter how lifting rounds are permuted and so lifting is said to be *sequence independent*. This is unlike what happens in general where the values that lifting coefficients take depend on their lifting round: better lifting coefficients are obtained in earlier rounds.

Superadditive conditions typically allow the derivation of strong closed-form inequalities. For instance, sequence-independent lifting has been used by Marchand and Wolsey [16] to derive facet-defining inequalities for a family of mixed integer knapsack sets and by Gu *et al.* [30] to derive valid inequalities for fixed-charge flow models. This concept admits generalizations in nonlinear programming [34,35].

Proving that the given functions are superadditive is cumbersome. However, several generic families of functions have been shown to be superadditive [14,20]. Necessary/sufficient conditions for certain univariate functions to be superadditive are given in Richard [42] and Richard *et al.* [43].

Example 9. The lifting function $\hat{\Psi}^2$ of Example 8 is superadditive over $S = \mathbb{N}^-$. Further $\hat{\mathcal{T}}^i = \{(y_i - (3i - 2)) \mid 0 \leq y_i \leq i + 2x_i, x_i \in \{0, \dots, i - 1\}\} = [-3i + 2, 0]$, for $i = 2, 3, 4$. Since $\hat{\mathcal{T}}^i \subseteq S$ for $i = 2, 3, 4$ and $\hat{\mathcal{T}}^2 + \hat{\mathcal{T}}^3 + \hat{\mathcal{T}}^4 \subseteq S$, we conclude from Proposition 2 that $\hat{\Psi}^i(z) = \hat{\Psi}^2(z)$ for all $z \in \hat{\mathcal{T}}^i$ when $i = 2, 3, 4$ if (α_2, β_2) and (α_3, β_3) exist

that satisfy condition (12) in which Ψ^q is replaced with $\hat{\Psi}^2$.

Approximate Lifting. When lifting functions are not superadditive, lifting can quickly become computationally expensive as it requires lifting functions to be recomputed. Superadditive conditions can be used to alleviate some of these computational difficulties if it is admissible to sacrifice some strength in the lifted inequality.

Proposition 4. Assume that (i) $\phi(z) \leq \Psi^1(z)$ for all $z \in \mathcal{T}_1 + \dots + \mathcal{T}_l$, (ii) $\mathcal{T}^q \subseteq \text{dom}(\phi)$ for all $q = 1, \dots, l$, (iii) ϕ is $(\mathcal{T}_q, \mathcal{T}_{q+1} + \dots + \mathcal{T}_l)$ -superadditive for all $q = 1, \dots, l$ and (iv) lifting coefficients (α^q, β^q) for variables (x^q, y^q) exist that satisfy condition (12) in which Ψ^q is replaced with ϕ . Then inequality (15) is a valid inequality for $\text{conv}(\mathcal{F})$.

The proof of Proposition 4 can be obtained similarly to that of Proposition 3 by showing by induction that $\Psi^q(z) \geq \phi(z)$ for all z in $\mathcal{T}_q + \dots + \mathcal{T}_l$ (note here that Ψ^q is the lifting function of the inequality obtained by using ϕ in the derivation of lifting coefficients (α_k, β_k) for $k = 1, \dots, q - 1$). The result can also be obtained by lower bounding Ψ^1 by ϕ in inequality (10), where $i = 1$ before using the superadditivity of ϕ in a manner similar to inequality (16) to obtain

$$\sum_{i \in M_0} \alpha_i x_i + \sum_{j \in N_0} \beta_j y_j + \sum_{q=1}^l \phi \left(\sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \right) \leq \delta,$$

which can then be linearized using assumptions (ii) and (iv).

Deriving a superadditive approximation of a given lifting function can be difficult. However, for univariate functions, Gu *et al.* [14] proved that such a superadditive approximation exists. Approximate superadditive lifting has been used in various settings where deriving closed-form facet-defining inequalities is difficult. It was first proposed by Gu *et al.* [14] for the cover inequality in 0-1

knapsack polytopes where desirable characteristics of superadditive approximations are described [20,44,45]. We refer the reader to papers [14,26,30] for discussions and applications.

About Lifting Coefficients

General Techniques for Deriving Lifting Coefficients. The problem of determining lifting coefficients is that of finding (α^q, β^q) that satisfy condition (12). In particular, condition (12) imposes a relation between (α^q, β^q) for each point of $\mathcal{P}^q$. Since there can be infinitely many of them, such a derivation is not always simple. One possible way of reducing the number of conditions to consider comes from observing that, in order to ensure that an inequality is valid for $\text{conv}(\mathcal{F}^q(\bar{x}, \bar{y}))$, it suffices to prove that it is satisfied at its extreme points and that its homogeneous counterpart is satisfied at its extreme rays. In particular, when $\text{conv}(\mathcal{F}^q(\bar{x}, \bar{y}))$ is bounded, it suffices to consider condition (12) for those points $(\tilde{x}^q, \tilde{y}^q)$ for which there exists an extreme point of $\text{conv}(\mathcal{F}^q(\bar{x}, \bar{y}))$ of the form $(x^0, \dots, x^{q-1}, \tilde{x}^q; y^0, \dots, y^{q-1}, \tilde{y}^q)$. Since there are finitely many such points, condition (12) defines a polyhedron Λ^q in the space of (α^q, β^q) variables. If Λ^q is empty, lifting is impossible. Otherwise, extreme points of Λ^q yield strong lifted inequalities; see the section titled “On the Strength of the Lifting Procedure”. Therefore, the derivation of lifting coefficients often reduces to finding extreme points of Λ^q . Since describing all extreme points of $\text{conv}(\mathcal{F}^q(\bar{x}, \bar{y}))$ can be difficult, it is common to restrict the number of variables in (M^q, N^q) to be small so that Λ^q is a small-dimensional polyhedron.

Lifting Coefficients from Superadditive Lifting Functions. We mentioned in the section titled “Superadditive Lifting” that lifting exhibits sequence-independent properties when lifting functions are superadditive. Another advantage occurs when deriving lifting coefficients. To illustrate this advantage, assume for the sake of simplicity that there are only integer variables in the definition of $\mathcal{F}$ and that $x^q \geq \bar{x}^q$ for all $x^q \in \mathcal{P}^q$. In this case, if Ψ^q is superadditive

over an appropriate domain, we can write

$$\sum_{i \in M_q} \Psi^q(A_i^*)(x_i - \bar{x}_i) \leq \Psi^q \left(\sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) \right), \quad (17)$$

showing that $\alpha_i = \Psi^q(A_i^*)$ are coefficients that satisfy condition (12). Similarly, the following result can be established, where we make use of Proposition 7.2 of Section II.1.7 of Nemhauser and Wolsey [40].

Proposition 5. *Let $\mathcal{H}^q$ be a mixed integer set of the form $\{(x^q, y^q) \in \mathbb{Z}^{m_q} \times \mathbb{R}^{n_q} \mid l_i \leq x_i \leq u_i \forall i \in M_q \text{ and } \bar{l}_j \leq y_j \leq \bar{u}_j \forall j \in N_q\}$ (where $l_i, u_i, \bar{l}_j$, and $\bar{u}_j$ are possibly infinite) that is such that $\mathcal{P}^q \subseteq \mathcal{H}^q$. Assume that Ψ^q is superadditive over*

$$\mathcal{W}^q := \left\{ \sum_{i \in M_q} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_q} B_j^*(y_j - \bar{y}_j) \mid (x^q, y^q) \in \mathcal{H}^q \right\}$$

and that Ψ^q has directional derivatives in the directions B_j^* and $-B_j^*$ for all $j \in N_q$. Assume also that (α^q, β^q) is a feasible solution to the system

$$\alpha_i \leq \Psi^q(A_i^*) \quad \forall i \in M_q$$

for which $\exists (x^q, y^q) \in \mathcal{P}^q$ with $x_i > \bar{x}_i$ (18)

$$\alpha_i \geq -\Psi^q(-A_i^*) \quad \forall i \in M_q$$

for which $\exists (x^q, y^q) \in \mathcal{P}^q$ with $x_i < \bar{x}_i$ (19)

$$\beta_j \leq \lim_{\lambda \rightarrow 0^+} \frac{\Psi^q(\lambda B_j^*)}{\lambda} \quad \forall j \in N_q$$

for which $\exists (x^q, y^q) \in \mathcal{P}^q$ with $y_j > \bar{y}_j$ (20)

$$\beta_j \geq -\lim_{\lambda \rightarrow 0^+} \frac{\Psi^q(-\lambda B_j^*)}{\lambda} \quad \forall j \in N_q$$

for which $\exists (x^q, y^q) \in \mathcal{P}^q$ with $y_j < \bar{y}_j$, (21)

then (α^q, β^q) satisfies condition (12).

In Proposition 5, $\mathcal{W}^q$ is a superset of $\mathcal{T}^q$ that is typically larger than necessary. Superadditivity of Ψ^q over $\mathcal{W}^q$ allows a derivation of the result in a manner similar to inequality (17).

Further, if Ψ^1 is superadditive over an appropriate subset of $\text{dom}(\Psi^1)$, Propositions 3 and 5 imply that the entire lifted inequality can be derived from Ψ^1 . Observe also that if conditions (18) and (19) (respectively, conditions (20) and (21) must both be imposed simultaneously for some $i \in M_q$ (respectively, $j \in N_q$), then the problem of finding lifting coefficients will likely be infeasible [[38], Corollary 4, [1], Example 4].

Example 10. Since $\hat{\Psi}^2(z) = \hat{\Psi}^3(z) = \hat{\Psi}^4(z)$ are superadditive over $\mathcal{S} = \mathbb{R}^-$, the above result shows that valid lifting coefficients (α_2, β_2) are obtained if $\alpha_2 \geq -\hat{\Psi}^2(-0)$ and $\beta_2 \geq -\lim_{\lambda \rightarrow 0^+} \frac{\hat{\Psi}^2(-\lambda B_2^*)}{\lambda} = 1$. Similarly, $\alpha_3 \geq 0$ and $\alpha_4 \geq 0$ with $\beta_3 \geq 1$ and $\beta_4 \geq 1$. Therefore the lifted inequality is

$$y_0 + y_1 + y_2 + y_3 + y_4 \leq 27 + x_0 + x_1. \quad (22)$$

In the above example, lifting coefficients α_2, α_3 and α_4 are not strongest and can be chosen equal to -1

Sequential Lifting Coefficients. Sequential lifting (i.e., the case where $|M_k| + |N_k| = 1$ for all $k = 1, \dots, q$) has received much attention in the literature. When lifting a single integer or continuous variable, call it z_q , condition (12) reduces to

$$\alpha_q(z_q - \bar{z}_q) \leq \Psi^q(A_q(z_q - \bar{z}_q)) \quad \forall z_q \in \mathcal{P}^q.$$

This condition can be rewritten as $\underline{\alpha}_q \leq \alpha_q \leq \bar{\alpha}_q$, where

$$\begin{aligned} \bar{\alpha}_q &:= \inf \left\{ \frac{\Psi^q(A_q(z_q - \bar{z}_q))}{z_q - \bar{z}_q} \mid z_q > \bar{z}_q \right. \\ &\quad \left. \text{and } z_q \in \mathcal{P}^q \right\} \\ \text{and } \underline{\alpha}_q &:= \sup \left\{ \frac{\Psi^q(A_q(z_q - \bar{z}_q))}{z_q - \bar{z}_q} \mid z_q < \bar{z}_q \right. \\ &\quad \left. \text{and } z_q \in \mathcal{P}^q \right\}. \end{aligned}$$

When $\bar{\alpha}_q < \underline{\alpha}_q$, lifting is impossible. When Ψ^q is superadditive, the result of the above problem reduces to that shown in the section titled “Lifting Coefficients from Superadditive Lifting Functions”. In fact, if the variable z_q is an integer, then the minimum in the definition of $\bar{\alpha}_q$ is achieved when $z_q = \bar{z}_q + 1$ and the maximum in the definition of $\underline{\alpha}_q$ is achieved when $z_q = \bar{z}_q - 1$ (provided that these problems are feasible). Similar results hold for continuous variables [17].

ON THE STRENGTH OF THE LIFTING PROCEDURE

We now derive conditions under which lifting produces facet-defining inequalities for MIPs. First observe that, if a point $(x^0, x^1, \dots, x^{q-1}; y^0, y^1, \dots, y^{q-1})$ of $\text{conv}(\mathcal{F}^{q-1}(\bar{x}, \bar{y}))$ satisfies inequality (9) at equality (we refer to such a point as *tight*), then its expansion $(x^0, \dots, x^{q-1}, \bar{x}^q, \dots, \bar{x}^d; y^0, \dots, y^{q-1}, \bar{y}^q, \dots, \bar{y}^d)$ will be tight in the inequality obtained after all variables are lifted. Therefore, the dimensions of the faces of $\text{conv}(\mathcal{F}^q(\bar{x}, \bar{y}))$ that lifted inequalities induce do not decrease as variables are reintroduced.

Example 11. The points $(0, 3)$ and $(1, 4)$ are tight in the seed inequality (8) and therefore their extensions $(\mathbf{0}, 0, 1, 2, 3; \mathbf{3}, 3, 4, 7, 10)$ and $(\mathbf{1}, 0, 1, 2, 3; \mathbf{4}, 3, 4, 7, 10)$ are tight in the lifted inequality (22).

Further, additional tight points might be added during the lifting process if lifting coefficients are adequately chosen. Remember that the derivation of lifting coefficients corresponds to rotating the plane associated with the seed inequality around its hinge until it becomes valid for the set in which the restrictions that $x^q = \bar{x}^q$ and $y^q = \bar{y}^q$ are relaxed. To obtain a strong lifted inequality, the plane must be rotated so that sufficiently many new feasible points become tight. This can be achieved by selecting (α^q, β^q) in such a way that sufficiently many affinely independent points $(x^q, y^q) \in \mathcal{P}^q$ satisfy inequality (12) at equality.

Example 12. The lifted inequality (13) obtained in Example 7 is facet-defining for $\text{conv}(\hat{\mathcal{F}}^1(\bar{x}, \bar{y}))$. In particular, lifting add the tight points (0, 2) (or (1, 0, 1, 2, 3; 5, 2, 4, 7, 10) in expanded form) and (1, 4) (or (0, 1, 1, 2, 3; 3, 4, 4, 7, 10) in expanded form), which were obtained in Example 7.

More generally, we obtain the following result.

Proposition 6. *Assume that inequality (7) defines a t_0 -dimensional face of $\text{conv}(\mathcal{F}^0(\bar{x}, \bar{y}))$. Assume also that, for each $k = 1, \dots, l$, lifting is possible and the coefficients (α^k, β^k) are chosen in such a way that there exists t_k affinely independent points in $\mathcal{P}_k \setminus \{(\bar{x}^k, \bar{y}^k)\}$ that satisfy inequality (12) at equality, then the resulting lifted inequality will be a face of $\text{conv}(\mathcal{F})$ of dimension at least $\sum_{i=0}^l t_i$.*

Proposition 6 is proven by using, for each k , the t^k points (x^k, y^k) that satisfy inequality (12) at equality to build tight points in the lifted inequality of the form $(\bar{x}^0, \dots, \bar{x}^{k-1}, x^k, \bar{x}^{k+1}, \dots, \bar{x}^l; \bar{y}^0, \dots, \bar{y}^{k-1}, y^k, \bar{y}^{k+1}, \dots, \bar{y}^l)$, where $(\bar{x}^0, \dots, \bar{x}^{k-1}, \bar{y}^0, \dots, \bar{y}^{k-1})$ is an optimal solution to $\Psi^k \left(\sum_{i \in M_k} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_k} B_j^*(y_j - \bar{y}_j) \right)$.

It follows from Proposition 6 that, to obtain strong lifted inequalities, it is desirable to select lifting coefficients so that as many affinely independent points of $\mathcal{P}^l \setminus \{(\bar{x}^l, \bar{y}^l)\}$ as possible satisfy inequality (12) at equality. When such a condition is achieved, we say that lifting is *maximum*.

Exact maximum lifting is often used to generate facet-defining inequalities for MIPs. In particular, if the fixing point $(\bar{x}, \bar{y})$ is such that $\mathcal{F}^k(\bar{x}, \bar{y})$ is full-dimensional for all $k = 0, \dots, l$, if seed inequality (7) is facet-defining for $\text{conv}(\mathcal{F}^0(\bar{x}, \bar{y}))$, if lifting is possible in each lifting round and maximum lifting is performed, then the lifted inequality produced is facet-defining for $\text{conv}(\mathcal{F})$.

Example 13. Consider the lifted inequality (22) obtained in Example 10. This inequality can be shown to define a face of dimension at least 7 for $\text{conv}(\hat{\mathcal{F}})$. In particular, in addition to the points described above,

the lifting of variables (x_2, y_2) added the tight point (1, 0, 1, 2, 3; 4 + ϵ , 3, 4 - ϵ , 7, 10); the lifting of (x_3, y_3) added the tight point (1, 0, 1, 2, 3; 4 + ϵ , 3, 4, 7 - ϵ , 10); and the lifting of (x_4, y_4) added the tight point (1, 0, 1, 2, 3; 4 + ϵ , 3, 4, 7, 10 - ϵ), where ϵ is a sufficiently small positive quantity. These points are easily constructed combining the comment following Proposition 6 and the discussion of the section titled “Sequential Lifting Coefficients”.

Characterizing the strength of an inequality obtained using approximate lifting is more difficult. However, if there is a solution (x^k, y^k) in $\mathcal{P}^k$ that satisfies inequality (12) at equality and that is such that $\Psi^1(z) = \phi(z)$, where $z = \sum_{i \in M_k} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_k} B_j^*(y_j - \bar{y}_j)$, then its extension is tight in the lifted inequality.

Proposition 7. *Assume that inequality (7) is a t_0 -dimensional face of $\text{conv}(\mathcal{F}^0(\bar{x}, \bar{y}))$. Assume also that, for each $k = 1, \dots, l$, lifting is possible and the coefficients (α^k, β^k) are chosen in such a way that there exists t_k affinely independent points (x^k, y^k) in $\mathcal{P}_k \setminus \{(\bar{x}^k, \bar{y}^k)\}$ that satisfy inequality (12) at equality and that satisfy $\Psi^1(z) = \phi(z)$, where $z = \sum_{i \in M_k} A_i^*(x_i - \bar{x}_i) + \sum_{j \in N_k} B_j^*(y_j - \bar{y}_j)$, then the lifted inequality will be a face of $\text{conv}(\mathcal{F})$ of dimension at least $\sum_{i=0}^l t_i$.*

It is clear from Proposition 7 that approximate lifting can produce facet-defining inequalities if the points for which conditions (12) are satisfied at equality correspond to regions where $\phi(z)$ equals $\Psi^1(z)$. When this is not the case, it seems intuitive that inequalities produced should be weaker than those obtained through exact maximum lifting. This intuition is not always verified as the resulting inequalities may still be facet-defining. An interesting such example is presented in Gu *et al.* [14, pp. 124–125] for the case of the 0–1 knapsack polytope.

REFERENCES

1. Richard J-PP, de Farias IR, Nemhauser GL. Lifted inequalities for 0-1 mixed integer programming: basic theory and algorithms. Math Program 2003;98:89–113.

2. Gomory RE. Some polyhedra related to combinatorial problems. *Linear Algebra Appl* 1969;2:341–375.
3. Padberg MW. On the facial structure of set packing polyhedra. *Math Program* 1973; 5:199–215.
4. Nemhauser GL, Trotter LE. Properties of vertex packing and independence system polyhedra. *Math Program* 1974;6:48–61.
5. Balas E. Facets of the knapsack polytope. *Math Program* 1975;8:146–164.
6. Wolsey LA. Facets for a linear inequality in 0–1 variables. *Math Program* 1975;8:165–178.
7. Hammer PL, Johnson EL, Peled UN. Facets of regular 0–1 polytopes. *Math Program* 1975;8:179–206.
8. Balas E, Zemel E. Facets of knapsack polytope from minimal covers. *SIAM J Appl Math* 1978;34:119–148.
9. Padberg MW. A note on zero-one programming. *Oper Res* 1975;23:833–837.
10. Zemel E. Lifting the facets of zero-one polytopes. *Math Program* 1978;15:268–277.
11. Wolsey LA. Facets and strong inequalities for integer programs. *Oper Res* 1976; 24:367–372.
12. Wolsey LA. Valid inequalities and superadditivity for 0/1 integer programs. *Math Oper Res* 1977;2:66–77.
13. Crowder HP, Johnson EL, Padberg MW. Solving large-scale zero-one linear programming problems. *Oper Res* 1983;31:803–834.
14. Gu Z, Nemhauser GL, Savelsbergh MWP. Sequence independent lifting in mixed integer programming. *J Comb Optim* 2000;4: 109–129.
15. Gu Z, Nemhauser GL, Savelsbergh MWP. Lifted cover inequalities for 0–1 integer programs: computation. *INFORMS J Comput* 1998;10:427–437.
16. Marchand H, Wolsey LA. The 0–1 knapsack problem with a single continuous variable. *Math Program* 1999;85:15–33.
17. Richard J-PP, de Farias IR, Nemhauser GL. Lifted inequalities for 0–1 mixed integer programming: superlinear lifting. *Math Program* 2003;98:115–143.
18. Bixby RE, Rothberg EE. Progress in computational mixed integer programming - a look back from the other side of the tipping point. *Ann Oper Res* 2007;149:37–41.
19. Weismantel R. On the 0/1 knapsack polytope. *Math Program* 1997;77:49–68.
20. Atamtürk A. On the facets of the mixed-integer knapsack polyhedron. *Math Program* 2003;98:145–175.
21. Nemhauser GL, Vance PH. Lifted cover facets of the 0–1 knapsack polytope with GUB constraints. *Oper Res Lett* 1994;16:255–263.
22. Park K, Park S. Lifting cover inequalities for the precedence-constrained knapsack problem. *Discrete Appl Math* 1997;72:219–241.
23. Sherali HD, Lee Y. Sequential and simultaneous liftings of minimal cover inequalities for generalized upper bound constrained knapsack polytopes. *SIAM J Discrete Math* 1995;8:133–153.
24. van de Leensel RLMJ, van Hoesel CPM, van de Klundert JJ. Lifting valid inequalities for the precedence constrained knapsack problem. *Math Program* 1999;86:161–185.
25. Zeng B, Richard J-PP. A polyhedral study on 0–1 knapsack problems with disjoint cardinality constraints: facet-defining inequalities by sequential lifting. *Discrete Optim.* In press.
26. Zeng B, Richard J-PP. A polyhedral study on 0–1 knapsack problems with disjoint cardinality constraints: strong valid inequalities by sequence-independent lifting. *Discrete Optim.* In press.
27. Fouilhoux P, Labbé M, Mahjoub AR, *et al.* Generating facets for the independence system polytope. *SIAM J Discrete Math* 2009;23:1484–1506.
28. Nobile P, Sassano A. Facets and lifting procedures for the set covering polytope. *Math Program* 1989;43:111–137.
29. Atamtürk A. Cover and pack inequalities for (mixed) integer programming. *Ann Oper Res* 2005;139:21–38.
30. Gu Z, Nemhauser GL, Savelsbergh MWP. Lifted flow cover inequalities for mixed 0–1 integer programs. *Math Program* 1999;85: 439–467.
31. Balas E, Fischetti M. Lifted cycle inequalities for the asymmetric traveling salesman problem. *Math Oper Res* 1999;24:273–292.
32. Grötschel M, Padberg MW. On the symmetric travelling salesman problem II: lifting theorems and facets. *Math Program* 1979;16:281–302.
33. de Farias IR. Semi-continuous cuts for mixed-integer programming. In: Bienstock D, Nemhauser GL, editors. *Proceedings of the 10th Conference on Integer and Combinatorial Optimization*. Berlin: Springer; 2004. pp. 63–88.

34. Richard J-PP, Tawarmalani M. Lifting inequalities: a framework for generating strong cuts for nonlinear programs. *Math Program* 2010;121:61–104.
35. Atamtürk A, Narayanan V. Lifting for conic mixed-integer programming. *Math Program*. In press.
36. de Farias IR, Johnson EL, Nemhauser GL. Facets of the complementarity knapsack polytope. *Math Oper Res* 2002;27:210–226.
37. Lin T-C, Vandenbussche D. Box-constrained quadratic programs with fixed charge variables. *J Glob Optim* 2008;41:75–102.
38. Atamtürk A. Sequence independent lifting for mixed-integer programming. *Oper Res* 2004;52:487–490.
39. Zemel E. Easily computable facets of the knapsack polytope. *Math Oper Res* 1989;14:760–764.
40. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. New York: Wiley-Interscience; 1988.
41. Dey SS, Richard J-PP. Linear-programming-based lifting and its application to primal cutting-plane algorithms. *INFORMS J Comput* 2009;21:137–150.
42. Richard J-PP. *Lifted Inequalities for 0–1 mixed integer programming* [PhD thesis]. Atlanta (GA): Georgia Institute of Technology; 2002.
43. Richard J-PP, Li Y, Miller LA. Valid inequalities for MIPs and group polyhedra from approximate liftings. *Math Program* 2009;118:253–277.
44. Shebalov S, Klabjan D. Sequence independent lifting for mixed integer programs with variable upper bounds. *Math Program* 2006;105:523–561.
45. Zeng B, Richard J-PP. A framework to derive multidimensional superadditive lifting functions and its applications. In: Fischetti M, Williamson DP, editors. *Proceedings of the 12th Conference on Integer and Combinatorial Optimization*. Berlin: Springer; 2007. pp. 210–224.

LIMIT THEOREMS FOR BRANCHING PROCESSES

PETER JAGERS

SERIK SAGITOV

Department of Mathematical Sciences,
Chalmers University of Technology
and University of Gothenburg,
Gothenburg, Sweden

The classical limit theorems of branching processes show what happens in the long run in a branching population. In principle, they are the same for all types of branching processes, from the simple single-type Galton–Watson process to general processes. (For the corresponding results to hold in multitype frameworks, types should *communicate*, though, in the sense that individuals must be able to get progeny of any type in their line of descent. This requirement may take various technical forms, and in infinite or abstract type spaces further conditions are needed [1].) The limit theorems describe exponential growth of supercritical processes and polynomial growth of critical processes, conditioned not to die out, and the convergence of subcritical populations, conditioned upon nonextinction. In the supercritical case, there are also theorems about the stabilization of population composition. For simple processes, there is only the kin structure to stabilize, but for age-dependent and general processes, the age-distribution becomes an important property, and for multitype processes evolution and possible limits of type distribution is an obvious topic.

GALTON–WATSON PROCESSES

Let $\{Z_n; n = 0, 1, 2, \dots\}$, $Z_0 = 1$ be a Galton–Watson process and X a random variable with the offspring distribution of that process. Write $m = E[X]$ for the reproduction mean, $\sigma^2 = \text{Var}[X]$, and $f(s) = E[s^X]$ for the

reproduction generating function. The logarithmic moment condition $E[X \log X] < \infty$ (which is weaker than $\sigma^2 < \infty$) is referred to as the $x \log x$ -condition.

1. Consider a subcritical process, that is, $m < 1$, satisfying $x \log x$. Then, there is a constant $0 < c \leq 1$, such that $P(Z_n > 0) \sim cm^n$. Further, $\lim_{n \rightarrow \infty} P(Z_n = k \mid Z_n > 0) = a_k$ exists, $\sum_{k=1}^{\infty} a_k = 1$, and the generating function $g(s) = \sum_{k=1}^{\infty} a_k s^k$ satisfies $g \circ f = mg + 1 - m$.
2. If $\{Z_n\}$ is critical (i.e., $m = 1$) and $\sigma^2 < \infty$, then the survival probability $P(Z_n > 0) \sim 2/(\sigma^2 n)$ and the process grows linearly, if conditioned upon survival: $E[Z_n \mid Z_n > 0] \sim \sigma^2 n/2$ and $P(Z_n/n \leq u \mid Z_n > 0) \rightarrow 1 - \exp(-2u/\sigma^2)$, as $n \rightarrow \infty$.
In the case of infinite variance $\sigma^2 = \infty$, a more general limit theorem involves a parameter $\gamma \in (0, 1]$ indicating the heaviness of the distribution tail $P(X > n) \sim cn^{1+\gamma}$. Since $P(Z_n > 0) \sim c'n^{-1/\gamma}$, the population growth conditional on nonextinction becomes polynomial of order $n^{1/\gamma}$.
3. If the process is supercritical, then there is a random variable $W \geq 0$, such that almost surely $Z_n \sim Wm^n$, as $n \rightarrow \infty$. If the $x \log x$ -condition holds, then convergence takes place in the mean as well, $W > 0$ almost surely on the set of nonextinction, and the Laplace transform ϕ of W , $\phi(s) = E[e^{-sW}]$, is the unique solution of the functional equation $\phi(mu) = f \circ \phi(u)$, $\phi'(0) = -1$. If the condition does not hold, $W = 0$ [2,3,14].

SINGLE-TYPE GENERAL PROCESSES

Now let $\{Z_t^X, t \geq 0\}$ denote a general branching process, starting from one newborn ancestor at time $t = 0$, and counted by

the *characteristic* χ . This means that for each individual there is a reproduction point process ξ , and an age-dependent characteristic $\chi \geq 0, \chi(a)$ interpreted as the weight of the individual at age a , so that Z_t^χ is the sum of all individuals' χ -values evaluated at their age at time t . The random characteristic χ is usually bounded and taken to vanish for negative a , so that unborn individuals do not contribute. The ξ of different individuals are supposed to be i.i.d. and so are the characteristics (for further details see Ref. 5). The simplest characteristic would just be 1 while the individual is alive, and 0 otherwise, implying $E[\chi(a)] = 1 - L(a)$, where L stands for the life length distribution. That way the process value reduces to the number of individuals alive. The process counted by that characteristic is usually denoted just $\{Z_t, t \geq 0\}$. Another characteristic would equal 1 if the individual is alive and younger than some given age, yielding the age distribution. In specific applications, the characteristic could refer to kin structure (pedigree) or body mass, number of cousins, and so on.

Write $\mu(a) = E[\xi(a)]$, assume $\mu(0) < 1$, so as to avoid the possibility of instantaneous explosion, and define the *Malthusian parameter* α by

$$\begin{aligned} \hat{\mu}(\alpha) &= \int_0^\infty e^{-\alpha t} \mu(dt) = E[\hat{\xi}(\alpha)] \\ &= E\left[\int_0^\infty e^{-\alpha t} \xi(dt)\right] = 1, \end{aligned}$$

hat denoting the Laplace(–Stieltjes) transform. If we write $m = \mu(\infty)$ for the expected offspring number of an individual during all her life, then m is larger than, smaller than, or equal to 1 together with α having the same relation to 0, that is the process being supercritical, subcritical, or just critical. A crucial entity is

$$\beta = \int_0^\infty u e^{-\alpha u} \mu(u) du,$$

the mean age at child-bearing. In the case $\beta < \infty$ the $x \log x$ -condition takes the form

$$E[\hat{\xi}(\alpha) \log \xi(\infty)] < \infty.$$

The main limit theorems below are formulated for what is called the *nonlattice* case for simplicity. It means that there is no lattice where all events can take place, like the lattice of positive numbers, or the multiples of some distance. The results all have lattice equivalents, covering thus discrete-time branching processes. Technical assumptions (like on measurability and smoothness) are omitted. The discrete-time case is treated in Ref. 4. As mentioned, this article is also limited to single-type populations. General multitype theory can be found in Ref. 1 and 6.

In the Galton–Watson case, $E[Z_n] = m^n$. For Markov branching processes in continuous-time, that is the (very) special case where life spans are exponentially distributed and mothers split into a random number of children at death, the offspring number independent of life span, the expected population size, analogously, grows exactly like $e^{\alpha t}$. For general processes, only an asymptotic such relation holds. Indeed, consider a characteristic which is independent of the past, that is, its value is determined by the lives of an individual and its progeny, and of course the individual's present age. By the renewal theorem [5], for example,

$$\lim_{t \rightarrow \infty} e^{-\alpha t} E[Z_t^\chi] = \int_0^\infty e^{-\alpha u} E[\chi(u)] du / \beta = c(\chi),$$

under minor conditions on the characteristic.

The three classical limit theorems take the form:

1. Consider a subcritical process, that is, $m < 1$, satisfying $x \log x, \beta < \infty$, and having a life length distribution L with $\int_0^\infty t e^{-\alpha t} L(dt) < \infty$. Then, there is an $0 < r \leq 1$, such that $P(Z_t > 0) \sim r e^{\alpha t}$. (Recall that $\alpha < 0$!) Further, $P(Z_t^\chi \leq y | Z_t > 0)$ weakly converges to a proper distribution with mean $c(\chi)/r$, [7].
2. In the critical case $\alpha = 0$, if $\sigma^2 < \infty$ and $1 - L(t) = o(t^{-2})$ as $t \rightarrow \infty$, then $P(Z_t > 0) \sim 2\beta/\sigma^2 t$. The process Z_t^χ divided by $\sigma^2 t/2$ and conditioned on

$Z_t > 0$, has an asymptotic exponential distribution with mean $c(\chi) = \int_0^\infty E[\chi(u)] du / \beta$, [7,8].

The case of heavy tailed life lengths, when $t^2(1 - L(t)) \rightarrow d$ with $0 < d \leq \infty$, brings essentially new limit results as the survival event can be due to just few long-living individuals, [9,10].

3. If, finally, the process is supercritical, then there is a random variable $W \geq 0$, such that almost surely $Z_t^\chi \sim c(\chi)W e^{\alpha t}$, almost surely as $t \rightarrow \infty$, provided the characteristic is independent of the past. (Characteristics that describe history are important for the studies of pedigrees, but here the limit takes a more complex form.) If the $x \log x$ -condition holds, then convergence takes place in the mean as well, and $W > 0$ almost surely on the set of nonextinction.

A consequence of (2) and (3) is that population composition stabilizes under growth. For suitable characteristics χ ,

$$\lim_{t \rightarrow \infty} Z_t^\chi / Z_t = \frac{\int_0^\infty e^{-au} E[\chi(u)] du}{\int_0^\infty e^{-au} (1 - L(u)) du},$$

since the expected value of the characteristic that tells if an individual is alive at age u is precisely $1 - L(u)$. As a linear functional in the characteristics, these ratios define a stable probability measure on a doubly infinite pedigree space (past and future), which describes the history and present composition of populations [6,11,12,13]. Some further general references are the books [15,16,17].

REFERENCES

1. Jagers P. The growth and stabilization of populations. *Stat Sci* 1991;6:269–283.
2. Athreya KB, Ney P. *Branching processes*. Berlin: Springer; 1972.
3. Harris TE. *The theory of branching processes*. Berlin: Springer; 1963.
4. Jagers P, Sagitov S. General branching processes in discrete time as random trees. *Bernoulli* 2008;14(4):949–962.
5. Jagers P. *Branching processes with biological applications*. Chichester: Wiley; 1975.
6. Jagers P, Nerman O. The asymptotic composition of supercritical multi-type branching populations. In: Azéma J, Emery M, Yor M, editors. *Séminaire de Probabilités XXX (Dedicated to P.-A. Meyer and J. Neveu on the occasion of their 60th birthdays)*. Lecture notes in mathematics 1626, Berlin: Springer; 1996, pp. 40–54.
7. Green PJ. Conditional limit theorems for general branching processes. *J Appl Probab* 1977;17:74–88.
8. Sagitov SM. A key limit theorem for critical branching processes. *Stoch Proc Appl* 1995;56:87–100.
9. Sagitov SM. Measure-branching renewal processes. *Stoch Proc Appl* 1994;52:293–307.
10. Vatutin VA. A new limit theorem for the critical Bellman-Harris branching process. *Math USSR-Sbornik* 1980;37:411–423.
11. Nerman O. The stable pedigrees of critical branching populations. *J Appl Probab* 1984;21:447–463.
12. Olofsson P. The $x \log x$ condition for general branching processes. *J Appl Probab* 1998;35:537–544.
13. Olofsson P. Size-biased branching population measures and the multitype $x \log x$ condition. *Bernoulli* 2009;15:1287–1304.
14. Lyons R, Pemantle R, Peres Y. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann Probab* 1995;3:1125–1138.
15. Haccou P, Jagers P, Vatutin VA. *Branching processes: variation, growth, and extinction of populations*. Cambridge: Cambridge University Press; 2005.
16. Sevastyanov BA. *Vetvyashchiesya Protsessy (branching processes)*. Moscow: Nauka; 1971. [Also available in German: *Verzweigungsprozesse*. 1974. Akademie Verlag, Berlin].
17. Mode C. *Multitype branching processes*. Oxford: Elsevier; 1971.

LIMIT THEOREMS FOR MARKOV RENEWAL PROCESSES

SERHAN ZIYA
Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

MARKOV RENEWAL-TYPE EQUATION

Suppose that $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence with kernel $G(\cdot)$ and X_n takes values from the discrete set E . We define $\{X(t), t \geq 0\}$ to be the semi-Markov process generated by this Markov renewal sequence. Specifically, for any $t \geq 0$, we let $X(t) = X_{N(t)}$ where

$$N(t) = \inf\{n \geq 0 : T_n \leq t\}. \quad (1)$$

We define $N(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ to be the Markov renewal process, and $\{M(t), t \geq 0\}$ to be the Markov renewal function generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$. In other words, for any $s \geq 0$, and $k \in E$, $N_k(s)$ is the number of times the semi-Markov process $\{X(t), t \geq 0\}$ entered the state k over $(0, s]$ and $M_{ik}(s)$ is its expectation conditional on the initial state being $i \in E$; that is,

$$M_{ik}(s) = \mathbf{E}[N_k(s) | X_0 = i] \text{ for } i, k \in E \text{ and } s \geq 0.$$

First, we introduce the matrix convolution notation, which will make the presentation significantly simpler. Suppose that $A(t) = [A_{jk}(t)]$ and $B(t) = [B_{jk}(t)]$ are two matrices of functions for which the product $A(t)B(t)$ is defined. Now let

$$\begin{aligned} C_{jk}(t) &= \sum_i \int_0^t A_{ji}(t-u) dB_{ik}(u) \\ &= \sum_i \int_0^t dA_{ji}(u) B_{ik}(t-u). \end{aligned}$$

Then, the matrix $C(t) = [C_{ij}(t)]$ is called the *convolution* of A and B (and written as $C(t) = A * B(t)$).

Now, any equation in the form of

$$R(t) = D(t) + G * R(t), \quad (2)$$

where $D(\cdot)$ is known and $R(\cdot)$ is a functional matrix, which one tries to solve for, is called a Markov renewal-type equation. Such equations are typically obtained by using the Markov renewal argument; that is, by conditioning on X_1 and T_1 . Under certain technical conditions, a closed-form expression for $R(t)$ is known, but this expression is usually not helpful as it requires complete knowledge of the Markov renewal function associated with the Markov renewal sequence. The limiting analysis, which is the subject of this article, is relatively easier.

KEY MARKOV RENEWAL THEOREM

First, we introduce new notation which will be useful in stating the Key Markov Renewal Theorem. Let

$$S_j = \inf\{t \geq T_1 : X(t) = j\}, j \in E;$$

$$\tau_i = \mathbf{E}[T_1 | X(0) = i], i \in E;$$

and

$$\tau_{ij} = \mathbf{E}[S_j | X(0) = i], i, j \in E.$$

Thus, S_j is the time of the first visit to state j at or after T_1 and τ_{ij} is its expectation given that the initial state is i . We also define

$$\theta_{ij} = \mathbf{E}[T_1 | X_0 = i, X_1 = j], i, j \in E;$$

$$a_i^2 = \mathbf{E}[T_1^2 | X_0 = i], i \in E;$$

and

$$a_{ij}^2 = \mathbf{E}[S_j^2 | X_0 = i], i, j \in E.$$

Finally, we define

$$\alpha_{ij} = \lim_{t \rightarrow \infty} \left(M_{ij}(t) - \frac{t}{\tau_{jj}} \right),$$

which can be shown to be equal to Ref. 1.

$$\alpha_{ij} = \begin{cases} \frac{a_{ii}^2 - 2\tau_{ii}^2}{2\tau_{ii}^2} & i = j, \\ \frac{a_{ij}^2 - 2\tau_{ij}\tau_{ji}}{2\tau_{jj}^2} & i \neq j. \end{cases}$$

The Key Markov Renewal Theorem gives an expression for $R(t)$ as $t \rightarrow \infty$. The theorem requires a number of technical conditions to hold. We first state these conditions:

Condition 1. Suppose that $H_j(\cdot)$ is the sojourn time distribution for state j for the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$ and suppose that there exists positive ε and α such that $H_i(\alpha) < 1 - \varepsilon$ for $i \in E$. Note that, as Kulkarni [1] shows, this condition essentially ensures that the process $\{N(t), t \geq 0\}$ as defined in Equation (1) is regular; that is, $N(t) < \infty$ for all $0 \leq t < \infty$ with probability 1.

Condition 2. The semi-Markov process $\{X(t), t \geq 0\}$ is irreducible, aperiodic, positive recurrent with a limiting distribution $\{p_j, j \in E\}$.

Condition 3. The functional matrix $D(t) = [D_{ij}(t)]$ satisfies the following:

- (i) $|D_{ij}(t)| \leq d_j(t) < \infty$ for $i \in E$ and $t \geq 0$,
- (ii) For all $i, j \in E$, $D_{ij}(t)$ can be written as a difference of two nonnegative bounded monotone functions.
- (iii) The limit $d_{ij} = \lim_{t \rightarrow \infty} D_{ij}(t)$ exists for all $i, j \in E$.
- (iv) For all $i, j \in E$, $\int_0^\infty (D_{ij}(t) - d_{ij}) dt < \infty$.
- (v) For all $j \in E$, $\int_0^\infty \sum_{k \in E} \frac{p_k}{\tau_k} D_{kj}(t) dt < \infty$.

If all these three conditions hold, the Key Markov Renewal Theorem says that

$$\lim_{t \rightarrow \infty} R_{ij}(t) = d_{ij} + \sum_{k \in E} \alpha_{ik} d_{kj} + \int_0^\infty \sum_{k \in E} \frac{p_k}{\tau_k} D_{kj}(t) dt. \quad (3)$$

Note that if $d_{ij} = 0$ for all $i, j \in E$, the first two terms in Equation (3) drop and the equation simplifies to

$$\lim_{t \rightarrow \infty} R_{ij}(t) = \int_0^\infty \sum_{k \in E} \frac{p_k}{\tau_k} D_{kj}(t) dt. \quad (4)$$

EXAMPLE: LIMITING BEHAVIOR OF MARKOV REGENERATIVE PROCESSES

In this section, we show how the key Markov renewal theorem can be used to derive the limiting probability distribution for Markov regenerative processes. Suppose that $\{Z(t), t \geq 0\}$ is a Markov regenerative process with an embedded Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$ with kernel $G(\cdot)$ and taking values in a discrete set E .

Define

$$R_{ij}(t) = P\{Z(t) = j | Z(0) = i\} \text{ for } i, j \in E.$$

First, we use a Markov renewal argument to find an expression for $R_{ij}(t)$. Specifically, conditioning on X_1 and T_1 , we get

$$\begin{aligned} P\{Z(t) = j | X_0 = i, X_1 = k, T_1 = s\} \\ = \begin{cases} R_{kj}(t-s), & s \leq t \\ P\{Z(t) = j | X_0 = i, X_1 = k, T_1 = s\}, & s > t \end{cases} \end{aligned}$$

This leads to

$$\begin{aligned} R_{ij}(t) &= \int_0^t \sum_{k \in E} R_{kj}(t-s) dG_{ik}(s) \\ &\quad + \int_t^\infty \sum_{k \in E} P\{Z(t) = j | X_0 = i, X_1 = k, \\ &\quad T_1 = s\} dG_{ik}(s). \end{aligned}$$

Then, we have

$$\begin{aligned} R_{ij}(t) &= P\{Z(t) = j, T_1 > t | X_0 = i\} \\ &\quad + [G * R(t)]_{ij}, \end{aligned} \quad (5)$$

which, in matrix form, can also be written as

$$R(t) = D(t) + G * R(t), \quad (6)$$

where $D(t) = [D_{ij}(t)]$ is defined by

$$D_{ij}(t) = P\{Z(t) = j, T_1 > t | Z(0) = i\} \text{ for } i, j \in E.$$

Note that Equation (6) is a Markov renewal-type equation.

Now, suppose that the sample paths of the Markov regenerative process $\{Z(t), t \geq 0\}$ are right continuous with left limits, and the semi-Markov process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$ is irreducible, aperiodic, and positive recurrent with limiting distribution $[p_j, j \in E]$. One can then verify that all the conditions of the key Markov renewal theorem, as stated in the previous section titled “Key Markov Renewal Theorem,” hold and one can see that $\lim_{t \rightarrow \infty} D_{ij}(t) = 0$ for any $i, j \in E$. Thus, from Equation (4), we get

$$\lim_{t \rightarrow \infty} R_{ij}(t) = \int_0^\infty \sum_{k \in E} \frac{p_k}{\tau_k} D_{kj}(s) ds \text{ for } i, j \in E. \quad (7)$$

Let β_{kj} denote the expected time the Markov regenerative process $\{Z(t), t \geq 0\}$ spends in state j during $(0, T_1)$ given that $X_0 = k$. One can then show that (see Ref. 1 for a complete derivation)

$$\beta_{kj} = \int_0^\infty D_{kj}(t) dt.$$

Plugging this in Equation (7), we find

$$\lim_{t \rightarrow \infty} P\{Z(t) = j | X_0 = i\} = \sum_{k \in E} \frac{p_k}{\tau_k} \beta_{kj} \text{ for } i, j \in E.$$

OTHER SOURCES

For more on the limit theorems for Markov Renewal Processes, we refer the reader to Refs 2 and 1, from which the author of this chapter has also benefited.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. Boca Raton (FL): Chapman & Hall/CRC; 1995.
2. Cinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1975.

LIMIT THEOREMS FOR RENEWAL PROCESSES

NAN LIU

Department of Health Policy and
Management, Mailman School of
Public Health, Columbia
University, New York,
New York

This article describes the *key renewal theorem* (KRT) and *Blackwell's renewal theorem* (BRT). These two limit theorems appear in different forms, but can be shown to be equivalent. For ease of presentation, we follow the terminology used in Karlin and Taylor [1] and refer to them collectively as the *renewal theorem* (RT). The RT is a very useful tool for characterizing the asymptotic behavior of a probabilistic quantity of interest in a *renewal process* (RP). To start, we briefly review renewal processes and the *elementary renewal theorem* (ERT).

An RP is a counting process with independent and identically distributed (iid) inter-event times. Mathematically, it can be defined as follows. Let $S_0 = 0$, and S_n be the occurrence time of the n th event, $n \geq 1$. Assume that

$$0 \leq S_1 \leq S_2 \leq S_3 \leq \dots$$

Define

$$X_n = S_n - S_{n-1}, \quad n \geq 1.$$

Thus $\{X_n, n \geq 1\}$ is a sequence of inter-event times. Next define

$$N(t) = \sup\{n \geq 0 : S_n \leq t\}.$$

Then $N(t)$ is called an RP generated by $\{X_n, n \geq 1\}$ if $\{X_n, n \geq 1\}$ is a sequence of nonnegative iid random variables. We refer the readers to the article titled **Definition and Examples of Renewal Processes** in this encyclopedia for a detailed introduction and more examples of RPs.

Suppose that the inter-event times $\{X_n, n \geq 1\}$ have a common cumulative distribution function (cdf) $G(\cdot)$, that is, $P(X_n \leq t) = G(t)$, $n \geq 1$. Let τ and σ^2 represent the mean and variance of X_1 , respectively. To avoid triviality, we assume that $G(0^-) = 0$ and $G(0^+) = G(0) < 1$, which imply that $\tau > 0$. Let

$$M(t) = E(N(t)), \quad t \geq 0.$$

The function $M(t)$ is called the *renewal function* (RF), which is simply the expected number of events (renewals) up to time t . The ERT states that

$$\lim_{t \rightarrow \infty} \frac{M(t)}{t} = \frac{1}{\tau}, \quad (1)$$

which implies the following asymptotic relation $M(t) \sim t/\mu$ as $t \rightarrow \infty$. As a sample path version of the ERT, we have

$$\lim_{t \rightarrow \infty} \frac{N(t)}{t} = \frac{1}{\tau}, \quad \text{with probability 1,} \quad (2)$$

which is a result from the strong law of large numbers. However, the ERT does not follow from Equation (2) directly since almost sure convergence does not imply convergence of expected values. A more intricate proof is needed to show the ERT. We refer the readers to the article titled **Renewal Function and Renewal-Type Equations** in this encyclopedia for an indepth discussion on the RF and ERT.

Blackwell's renewal theorem, which was proposed by David Blackwell (Professor Emeritus of Statistics at the University of California, Berkeley), can be regarded as a refinement of this asymptotic relation. It asserts that for any $h > 0$

$$M(t+h) - M(t) \rightarrow \frac{h}{\tau}, \quad \text{as } t \rightarrow \infty,$$

under certain mild conditions on $G(\cdot)$. In words, the expected number of renewals during an interval of length h is approxi-

mately h/τ , provided that the process has already run for a long time. This result can be easily verified in a Poisson process, a special case of RPs.

In the next section, we first describe the KRT, which is equivalent to the BRT but of a different form. The KRT provides a convenient setting for the application of RT. Then we formally present the BRT, and briefly discuss its connection with the KRT.

THE RENEWAL THEOREM

The KRT reveals the asymptotic behavior of the solution to a *renewal-type equation*, which has the following form:

$$H(t) = D(t) + \int_0^t H(t-x) dG(x), \quad (3)$$

where $D(\cdot)$ is a known function, $G(\cdot)$ is the cdf of inter-event times, and $H(t)$ is to be determined. Renewal-type equations often arise when using the *renewal argument*, which is a method to derive probabilistic quantities in RPs by conditioning on the occurrence time of the first event, that is, S_1 . We refer the readers to the article titled **Renewal Function and Renewal-Type Equations** in this encyclopedia for an introduction of the renewal-type equation and renewal argument. In this section, we focus on the asymptotic behavior of $H(t)$.

Before presenting the KRT, we need the following definition.

Definition 1 [Periodic Random Variables]. A nondefective random variable X (or its distribution) is called *periodic* (or *arithmetic*, or *lattice*) if there exists a $d > 0$ such that

$$\sum_{k=0}^{\infty} P(X = kd) = 1.$$

The largest value of d for which the above equality holds is called the *period* (or *span*) of the random variable X (or its distribution).

If a random variable X is not periodic, it is called *aperiodic* (or *non-arithmetic*, or *non-lattice*). We refer the readers to Kulkarni [2]

for more examples on periodic and aperiodic random variables. Note that all continuous random variables are aperiodic.

Theorem 1 [Key Renewal Theorem]. Let $H(t)$ be a solution to the renewal-type equation (3). Suppose that $D(t)$ is a monotonic function which is absolutely integrable in the sense that

$$\int_0^{\infty} |D(t)| dt < \infty.$$

(i) If $G(\cdot)$ is aperiodic with mean $\tau > 0$, then

$$\lim_{t \rightarrow \infty} H(t) = \begin{cases} \frac{1}{\tau} \int_0^{\infty} D(t) dt & \text{if } \tau < \infty, \\ 0 & \text{if } \tau = \infty. \end{cases}$$

(ii) If $G(\cdot)$ is periodic with period d and mean $\tau > 0$, then for $0 \leq x < d$,

$$\begin{aligned} & \lim_{k \rightarrow \infty} H(kd + x) \\ &= \begin{cases} \frac{d}{\tau} \sum_{k=0}^{\infty} D(kd + x) & \text{if } \tau < \infty, \\ 0 & \text{if } \tau = \infty. \end{cases} \end{aligned}$$

We refer to Feller [3] for a proof of the KRT. In his proof, Feller assumes that the function $D(t)$ is *directly Riemann integrable*. Feller's condition is technical, and not presented here. The condition we gave above is a simple and sufficient condition for Feller's, and suffices for most of the applications of the KRT. As noted in Tijms [4], a more general condition guaranteeing $D(t)$ to be directly Riemann integrable is that $D(t)$ can be written as a finite sum of monotone, integrable functions.

As mentioned earlier, another form of the KRT is the BRT, which is expressed more explicitly in terms of the renewal function $M(t)$. We present the BRT below, and follow the convention that $h/\tau = 0$ if $\tau = \infty$.

Theorem 2 [Blackwell's Renewal Theorem]. Let $h > 0$ be a fixed number.

(i) If $G(\cdot)$ is aperiodic with mean $\tau > 0$, then

$$\lim_{t \rightarrow \infty} [M(t+h) - M(t)] = \frac{h}{\tau}.$$

(ii) If $G(\cdot)$ is periodic with period d and mean $\tau > 0$, then

$$\lim_{t \rightarrow \infty} [M(t+h) - M(t)] = \frac{h}{\tau},$$

provided that h is a multiple of the period d .

The BRT and the KRT can be shown to be equivalent. The proof of BRT using the KRT can be found in Karlin and Taylor [1] and Kulkarni [2]. As discussed in Karlin and Taylor [1] and Ross [5], the KRT can also be deduced from the BRT by approximating a directly Riemann integrable function with step functions.

The next section presents two classical applications of the KRT. We discuss the limiting distributions of recurrence times (including the current age, excess life, and total life) in an RP, and the asymptotic expansion of a renewal function. We briefly mention how the KRT can be used to find these results. Detailed derivations can be found in Karlin and Taylor [1], Kulkarni [2], and Ross [5].

APPLICATIONS OF THE KEY RENEWAL THEOREM

Limiting Distributions of Recurrence Times

The following three probabilistic quantities related to event times are of particular interest in the study of RPs:

$$\begin{aligned} A(t) &= t - S_{N(t)}, \\ B(t) &= S_{N(t)+1} - t, \\ L(t) &= S_{N(t)+1} - S_{N(t)} \\ &= X_{N(t)+1} = A(t) + B(t), \end{aligned} \quad (4)$$

where $t \geq 0$.

Physically, $A(t)$ represents the time elapsed from the latest event before t to time t , and $B(t)$ represents the time duration between t and the next event immediately after t . Their sum $L(t)$ is the total inter-event time that contains t . The quantities $A(t)$, $B(t)$, and $L(t)$ are called the current age, remaining life (excess

life), and total life at time t , respectively. Sometimes $A(t)$ is also called the backward recurrence time and $B(t)$ is called the forward recurrence time. The stochastic processes $\{A(t), t \geq 0\}$, $\{B(t), t \geq 0\}$, and $\{L(t), t \geq 0\}$ are called the *age process*, *remaining life process*, and *total life process*, respectively.

In this section, we discuss the asymptotic behavior of these three processes. Suppose that $G(\cdot)$ is aperiodic and $\tau < \infty$. Then the following results can be shown by using the KRT:

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbf{P}(A(t) > x) &= \lim_{t \rightarrow \infty} \mathbf{P}(B(t) > x) \\ &= \frac{1}{\tau} \int_x^\infty (1 - G(u)) du, \end{aligned} \quad (5)$$

$$\lim_{t \rightarrow \infty} \mathbf{P}(A(t) > x, B(t) > y) = \frac{1}{\tau} \int_{x+y}^\infty (1 - G(u)) du,$$

and

$$\lim_{t \rightarrow \infty} \mathbf{P}(L(t) > x) = \frac{1}{\tau} \int_x^\infty u dG(u). \quad (6)$$

To facilitate understanding, we present how to derive Equation (5) by applying the KRT, and other results can be shown in a similar way. We start with the derivation of the limiting distribution of $B(t)$. Let $H(t) = \mathbf{P}(B(t) > x)$. Then conditioning on the occurrence time of the first renewal event $X_1 = u$, we have

$$\begin{aligned} \mathbf{P}(B(t) > x | X_1 = u) \\ &= \begin{cases} H(t-u) & \text{if } 0 \leq u \leq t \\ 0 & \text{if } t < u \leq x+t \\ 1 & \text{if } u > x+t. \end{cases} \end{aligned}$$

Unconditioning, we get

$$\begin{aligned} H(t) &= \int_0^\infty \mathbf{P}(B(t) > x | X_1 = u) dG(u) \\ &= \int_0^t H(t-u) dG(u) + \int_{x+t}^\infty dG(u) \\ &= \int_0^t H(t-u) dG(u) + 1 - G(x+t). \end{aligned}$$

Now, for a fixed $x > 0$, $1 - G(x + t)$ is nonnegative and nonincreasing for $t \geq 0$, and

$$\int_0^\infty (1 - G(x + t)) dt \leq \int_0^\infty (1 - G(t)) dt = \tau < \infty.$$

Hence the conditions for the KRT are satisfied, from which it directly follows that

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbf{P}(B(t) > x | X_1 = u) &= \frac{1}{\tau} \int_0^\infty (1 - G(x + t)) dt \\ &= \frac{1}{\tau} \int_x^\infty (1 - G(t)) dt. \end{aligned}$$

To derive the limiting distribution for $A(t)$, first notice that for $t > x$, $\{A(t) > x\} \Leftrightarrow \{\text{No renewals in } [t - x, t]\} \Leftrightarrow \{B(t - x) > x\}$. Thus,

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbf{P}(A(t) > x) &= \lim_{t \rightarrow \infty} \mathbf{P}(B(t - x) > x) \\ &= \lim_{t \rightarrow \infty} \mathbf{P}(B(t) > x). \end{aligned}$$

This completes the derivation of Equation (5).

Applying the KRT, one can further show that

$$\lim_{t \rightarrow \infty} \mathbf{E}(A(t)) = \lim_{t \rightarrow \infty} \mathbf{E}(B(t)) = \frac{\sigma^2 + \tau^2}{2\tau} \quad (7)$$

and

$$\lim_{t \rightarrow \infty} \mathbf{E}(L(t)) = \frac{\sigma^2 + \tau^2}{\tau}. \quad (8)$$

However, we should remind the readers that Equations (7) and (8) cannot be directly deduced from Equations (5) and (6), respectively, since convergence in distribution does not imply convergence of expected values. Hence, one should resort to the KRT for a

proof. See, for example, Kulkarni [2], for a detailed derivation.

Below we introduce an interesting phenomenon called *inspection paradox* in relation to Equations (4) and (8). Recall from Equation (4) that $L(t) = X_{N(t)+1}$, but Equation (8) says that

$$\lim_{t \rightarrow \infty} \mathbf{E}(X_{N(t)+1}) = \frac{\sigma^2 + \tau^2}{\tau} \geq \tau = \mathbf{E}(X_n),$$

that is, for a large t , the inter-event time that covers t is on average longer than a generic inter-event interval. This counterintuitive phenomenon is known as the inspection paradox. In fact, one can further show that, for a fixed $t \geq 0$,

$$\mathbf{P}(X_{N(t)+1} > x) \geq \mathbf{P}(X_1 > x). \quad (9)$$

In words, the length of the renewal interval that contains time t is *stochastically larger* than the first renewal interval. This immediately implies that

$$\mathbf{E}(X_{N(t)+1}) \geq \mathbf{E}(X_1).$$

To understand this paradox, let us first fix an arbitrary point t . We measure the length of the interval containing t . However, it seems plausible that the longer the duration of an interval, the larger the likelihood that it would be experienced and measured. Such samplings are known as *length-biased sampling*, and occur in many sampling situations. Stein and Dattero [6] provide introductory examples to illustrate such sampling bias and the inspection paradox. For a mathematical proof of Equation (9), the readers are referred to Angus [7] and Ross [8].

Now, let us consider a concrete example. Let $\{N(t), t \geq 0\}$ be a Poisson process with mean inter-event time $1/\lambda$. It can be shown that (see, for example, Chapter 5.3 of Karlin and Taylor [1])

$$\mathbf{E}(L(t)) = \frac{1}{\lambda} + \frac{1}{\lambda}(1 - e^{-\lambda t}) \geq \frac{1}{\lambda} = \mathbf{E}(X_1).$$

We see that, as $t \rightarrow \infty$, the mean total life doubles the mean inter-event time!

Asymptotic Expansion of the Renewal Function

Recall that the ERT implies that

$$M(t) = \frac{t}{\tau} + o(t),$$

where $o(t)$ represents a function $f(t)$ such that $f(t)/t \rightarrow 0$ as $t \rightarrow \infty$. We are now interested in refining this asymptotic approximation of $M(t)$. To do this, we study the function $o(t)$, and give the second term of the asymptotic expansion of $M(t)$. Suppose that $\sigma^2 < \infty$ and $G(\cdot)$ is aperiodic. To start, we define

$$H(t) = M(t) - \frac{t}{\tau},$$

and then apply the renewal argument, which leads to a renewal-type equation with $H(t)$ as its solution. From this renewal-type equation, one is able to identify the function $D(t)$, and verify that $D(t)$ can be written as a sum of two monotone functions both of which are absolutely integrable. Thus, the KRT is directly applicable, and it follows that

$$\lim_{t \rightarrow \infty} \left[M(t) - \frac{t}{\tau} \right] = \frac{1}{\tau} \int_0^\infty D(t) dt = \frac{\sigma^2 - \tau^2}{2\tau^2},$$

or equivalently

$$M(t) = \frac{t}{\tau} + \frac{\sigma^2 - \tau^2}{2\tau^2} + o(1), \quad (10)$$

where $o(1)$ is a function of t which approaches 0 as $t \rightarrow \infty$. This expansion gives a more accurate estimate of $M(t)$ for large t , and implies that t/τ overestimates $M(t)$ for large t if the coefficient of variation $\sigma^2/\tau^2 < 1$. Consider a Poisson process, where $\sigma^2 = \tau^2$, and in fact $M(t) = t/\tau$, verifying Equation (10) in this special case of RPs.

FURTHER READINGS

Further details on these two limit theorems for RPs, namely, the KRT and the BRT, can be found in classical textbooks including Karlin and Taylor [1], Kulkarni [2], and Ross [5]. Two seminal papers that prove the BRT and the KRT are Blackwell [9] and Smith [10], respectively. To track the theoretical development of these limit theorems and in general

the renewal theory, we refer the readers to Smith [11].

These limit theorems are very useful tools in the study of stochastic processes. They are used in the proof of those key results in alternating renewal processes and regenerative processes [2,5,12]. These two processes are discussed in the articles titled **Alternating Renewal Processes** and **Regenerative Processes** in this encyclopedia, respectively. The two limit theorems have also been extensively used in the study of queueing theory, inventory control, and warranty management. Literature on these applications is enormous. We provide a few examples in relevance to those applications of the KRT discussed above.

The remaining life in RPs is widely used in stochastic modeling. In queueing theory, it is the remaining service time; in the context of inventory control, it is the undershoot of the reorder point. By considering only those times when the server is busy, the remaining service time seen by an arriving customer in an M/G/1 queue is equivalent to the remaining life in a RP generated by a nonstopping service process [13,14]. Using this result, one can construct a proof for the famous Pollaczek–Khinchine formula, and study the distribution of waiting times in an M/G/1 queue [15]. Karlin [16] discusses the application of renewal theory to the study of inventory control. In particular, he applies the excess life in the study of a periodic review inventory system managed by an (s, S) policy. Green [17] extends the study of limiting distribution of recurrence times. She shows that the asymptotic joint distribution of the remaining life and total life of an subinterval is that of an RP generated by these subintervals only, if attention is restricted to these subintervals.

The RF plays an important role in the cost analysis of warranty management. Many asymptotic results are available for approximating RFs involved in such cost computations. Feller [3] and Cox and Miller [18] provide asymptotic expansions for an RF when inter-event times are periodic. General bounds for an RF are discussed in Lorden [19] and Barlow and Proschan [20]. More recent development of these bounds can be found in

Politis Koutras [21]. For an introduction and review on warranty management and its connection with the renewal theory, the readers are referred to Blischke and Murthy [22] and Thomas and Rao [23].

REFERENCES

1. Karlin S, Taylor HM. A first course in stochastic processes. New York: Academic Press; 1975.
2. Kulkarni VG. Modeling and analysis of stochastic systems. Boca Raton (FL): Chapman & Hall/CRC; 1995.
3. Feller W. Volume I and II, An introduction to probability theory and its applications. New York: John Wiley & Sons, Inc.; 1971.
4. Tijms HC. A first course in stochastic models. Chichester: John Wiley & Sons, Ltd.; 2003.
5. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1996.
6. Stein WE, Dattero R. Sampling bias and the inspection paradox. *Math Mag* 1985;58(2): 96–99.
7. Angus JE. The inspection paradox inequality. *SIAM Rev* 1997;39(1):95–97.
8. Ross SM. The inspection paradox. *Probab Eng Inform Sci* 2003;17(1):47–51.
9. Blackwell D. A renewal theorem. *Duke Math J* 1948;15(1):145–150.
10. Smith WL. Asymptotic renewal theorems. *Proc R Soc Edinb Sect A* 1954;64:9–48.
11. Smith WL. Renewal theory and its ramifications. *J R Stat Soc Ser B* 1958;20(2):243–302.
12. Asmussen S. Applied probability and queues. New York: Springer; 2003.
13. Oliver RM. An alternate derivation of the Pollaczek–Khinchine formula. *Oper Res* 1964;12(1):158–159.
14. Fakinos D. The expected remaining service time in a single server queue. *Oper Res* 1982;30(5):1014–1018.
15. Gross D, Shortle JF, Thompson JM, *et al.* Fundamentals of queueing theory. 4th ed. Hoboken (NJ): John Wiley & Sons, Inc.; 2008.
16. Karlin S. The application of renewal theory to the study of inventory policies. In: Arrow KJ, Karlin S, Scarf HE, editors. Studies in the mathematical theory of inventory and production. Stanford (CA): Stanford University Press; 1958. pp. 270–297.
17. Green LV. A limit theorem on subintervals of interrenewal times. *Oper Res* 1982;30(1):210–216.
18. Cox DR, Miller HD. The theory of stochastic processes. New York: John Wiley & Sons, Inc.; 1977.
19. Lorden G. On excess over the boundary. *Ann Math Stat* 1970;41(2):520–527.
20. Barlow RE, Proschan F. Statistical theory of reliability and life testing: probability models. New York: Holt, Rinehart and Winston; 1975.
21. Politis K, Koutras MV. Some new bounds for the renewal function. *Probab Eng Inform Sci* 2006;20(2):231–250.
22. Blischke WR, Murthy DNP. Warranty cost analysis. New York: Marcel Dekker; 1993.
23. Thomas MU, Rao SS. Warranty economic decision models: a summary and some suggested directions for future research. *Oper Res* 1999;47(6):807–820.

LINEAR PROGRAMMING AND TWO-PERSON ZERO-SUM GAMES

PABLO ZAFRA
Department of Mathematics,
Kean University, Union,
New Jersey

From their inception, the theories underlying linear programming and two-person zero-sum games (also commonly referred to as *matrix games*) have been intertwined, especially with respect to the related primal-dual problems (see the section titled “Linear Programming” in this encyclopedia). The close relationship between these two seemingly disparate areas has been established first in 1947 when von Neumann pointed out that his minimax theorem is equivalent to the concept of duality in linear programming (LP), establishing the transformation of a matrix game into a pair of primal-dual LP problems [1]. Gale *et al.* [2] later illustrated the converse situation of transforming an LP problem into a symmetric game (i.e., a matrix game with skew-symmetric matrix), establishing the full equivalence between the two problems.

However, the relationship between the computational aspects for solving matrix games and linear programs has been one-way; we solve matrix games by transforming them into linear programming problems, but the use of game algorithms in the solution of linear programs is rare. To date, when one wants to solve a matrix game, it is easy and convenient to convert the game to a standard linear programming problem as discussed below, and then solve the resultant problem using a computer code based on the simplex or interior point methods.

TRANSFORMING MATRIX GAMES INTO LINEAR PROGRAMMING PROBLEMS

A two-person zero-sum game is represented by an $(m \times n)$ payoff matrix $A = (a_{ij})$, where

the entry a_{ij} represents the payment of column player P2 to the row player P1, should P1 choose row i as his move and should P2 choose column j as her move. It is clear that P1 would want to select the row that will maximize his potential payoff, while P2 would want to select the column that will minimize her payout. Note that, if a_{ij} is negative, then it is interpreted as P1's payment to P2 for the corresponding pair of strategies. This two-person game is called *zero-sum* because one player's gain is the other player's loss. The rows and columns of A are called *pure* (or *deterministic*) strategies that are available for the two players to choose.

If the maximizing player P1 selects row i , then his payoff is at least $\min_j a_{ij}$, while the minimizing player P2's payout would be at most $\max_i a_{ij}$ should she select column j . For this reason, P1 would select the pure strategy i that would make $\min_j a_{ij}$ as large as possible and, therefore, he is assured of winning at least $\max_i \min_j a_{ij}$. On the other hand, P2 would select the pure strategy j that would make $\max_i a_{ij}$ as small as possible to minimize her losses. Consequently, by doing this, P2 can keep P1 from winning no more than $\min_j \max_i a_{ij}$. It can be shown that $\max_i \min_j a_{ij} \leq \min_j \max_i a_{ij}$, that is, P1's maximin pure strategy is no more than P2's minimax pure strategy. If $\max_i \min_j a_{ij} = \min_j \max_i a_{ij} = v$ for some pair i and j , then the matrix game is said to have a saddlepoint (or equilibrium point) corresponding to the pair of optimal pure strategies i and j with the value of the game given by v . Both players should play the row and column associated with the saddlepoint, which assures P1 to win v and P2 to pay the same amount. As an equilibrium point, given what P2 is playing, any other choice for P1 lowers his opportunity to win at least v and, conversely, given what P1 is playing, any other choice for P2 raises P1's opportunity to win more than v .

However, not all matrix games can be played with an optimal pair of pure strategies. In fact, since the game is antagonistic in nature, fixing a pure strategy for P1 in advance when there is no saddle point is not a good move if P2 can “read” his mind. This is because P2 can always choose a pure strategy that would be favorable to her. In the same manner, if P1 can “read” P2’s mind/move in advance, then P1 can select his pure strategy that would be favorable to him. As a simple illustration that a predictable strategy is bad when there is no saddlepoint, consider the familiar rock-paper-scissors game. If one player fixes his/her move, say “rock,” and the other player can “read” this in advance, then this other player will certainly choose “paper” as his/her move for him/her to win the game. In this situation, both players’ interest is clearly best served if they choose their moves with some amount of randomness.

To do this, they will assign certain probabilities to choosing rows and columns for their strategies. For example, suppose P1 selects row i with probability x_i , and P2 selects column j with probability y_j . With these probabilities as components, the vectors $x = (x_1, \dots, x_m)^T$ and $y = (y_1, \dots, y_n)^T$, where $\sum_i x_i = \sum_j y_j = 1$ and $x_i, y_j \geq 0$, are called *randomized (mixed) strategies* for P1 and P2, respectively. Note that any mixed strategy with one of the components equal to 1 (and all others 0) is a *pure strategy*.

If P1 uses any mixed strategy x and P2 uses any mixed strategy y , the expected payoff to P1 from P2 for this pair of strategies is given by $E(x, y) = \sum_i \sum_j x_i a_{ij} y_j$ [3]. For any mixed strategy x that P1 chooses, P2’s interest is to choose her mixed strategy y that minimizes P1’s expected payoff. In this case, P1 can expect to receive at least $\min_y E(x, y)$, and therefore he would want to select his particular strategy x that will make this as large as possible. That is, his goal is to solve $\max_x \min_y E(x, y)$ for that strategy x . Similarly, for any mixed strategy y that P2 chooses, P1’s interest is to choose a mixed strategy x that maximizes his expected payoff. In this case, P2 can expect not to

lose more than $\max_x E(x, y)$, and therefore she would want to select her particular strategy y that will make this as small as possible. That is, her goal is to solve $\min_y \max_x E(x, y)$ for that strategy y . Von Neuman’s Minimax Theorem [4] states that there exists a pair of optimal mixed strategies (x, y) such that $\max_x \min_y E(x, y) = \min_y \max_x E(x, y)$; that is, the expected payoff for P1 is equal to the expected loss for P2. This common value is considered the value of the game (see **Saddlepoints and von Neumann Minimax Theorem**).

We now formulate the LP equivalence in regards to P1. His goal is to find a mixed strategy $x = (x_1, \dots, x_m)$ where $\sum_i x_i = 1, x_i \geq 0$ that will maximize $\min_y E(x, y)$. It can be shown that, for any given x , $\min_y E(x, y)$ is attained at some pure strategy j that P2 should choose; that is, $\min_y E(x, y) = \min_j \sum_i x_i a_{ij}$ [5]. Therefore, no matter what P2’s pure strategy selection is, P1 is assured of winning at least $\min_j \sum_i x_i a_{ij}$. Let v be a lower bound for each of the j summations, $\sum_i x_i a_{ij} \geq v$. Then P1’s goal is equivalent to maximizing v . We now have the LP problem formulation for P1 as given below (LP1)

$$\begin{aligned}
 & \max v \\
 & \text{s.t.} \\
 & a_{11}x_1 + a_{21}x_2 + \dots + a_{m1}x_m \geq v \quad (j = 1) \\
 & a_{12}x_1 + a_{22}x_2 + \dots + a_{m2}x_m \geq v \quad (j = 2) \\
 & \vdots \\
 & a_{1n}x_1 + a_{2n}x_2 + \dots + a_{mn}x_m \geq v \quad (j = n) \\
 & x_1 + \dots + x_m = 1 \\
 & x_i \geq 0.
 \end{aligned}$$

Note that each inequality corresponds to P2’s pure strategy with the indicated value in parenthesis.

Using analogous arguments above, let us now develop the LP formulation in regard to P2’s point of view. Her goal is to find a mixed strategy $y = (y_1, \dots, y_n)$ where $\sum_j y_j = 1, y_j \geq 0$ that will minimize

$\max_x E(x, y)$. It can also be shown that for any given y , $\max_x E(x, y)$ is attained at some pure strategy i that P1 should choose; that is, $\max_x E(x, y) = \max_i \sum_j a_{ij} y_j$ [5]. Therefore, no matter what P1's pure strategy selection is, P2's cost is no more than $\max_i \sum_j a_{ij} y_j$. Let u be an upper bound for each of the i summations, $\sum_j a_{ij} y_j \leq u$. Then P2's goal is equivalent to minimizing u . We now have the LP problem formulation for P2 as given below (LP2)

$$\begin{aligned} & \min u \\ & \text{s.t.} \\ & a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \leq u \quad (i = 1) \\ & a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \leq u \quad (i = 2) \\ & \vdots \\ & a_{m1}y_1 + a_{m2}y_2 + \cdots + a_{mn}y_n \leq u \quad (i = m) \\ & y_1 + \cdots + y_n = 1 \\ & y_j \geq 0. \end{aligned}$$

Here, each inequality corresponds to P1's pure strategy with the indicated value in parenthesis.

Observe that LP2 is the dual of the LP1 (primal) problem. Since von Neumann's Minimax Theorem guarantees that every matrix game has a solution, the primal-dual problems we formulated are both feasible and have optimal solutions. Suppose LP1 has optimal solution (x^*, v^*) , then it follows that P1's expected payoff is v^* if he selects his optimal mixed strategy x^* . Similarly, if LP2 has optimal solution (y^*, u^*) , then it follows that P2's expected cost is u^* if she selects her optimal mixed strategy y^* . The Duality Theorem of Linear Programming states that the primal-dual problems have the same optimal value for the objective functions [5]; that is, $v^* = u^*$. This remarkable fact that P1's optimal expected payoff of v^* is equal to P2's optimal expected cost of u^* (with the corresponding pair of optimal mixed strategies (x^*, y^*) for P1 and P2, respectively) is precisely what von Neumann's Theorem asserts as well, with the common expected payoff/cost

$v^* = u^*$ as the value of the game. Therefore, the minimax theorem and the duality theorem in LP are indeed equivalent (see **LP Duality and KKT Conditions for LP**).

Example 1. In the rock-paper-scissors game, the payoff matrix is given by

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix},$$

where $i, j = 1$ (rock), 2 (paper), 3 (scissors).

One would quickly realize that any pair of pure strategies (i, j) where $i \neq j$ would not be favorable to the other player. As explained earlier, a predictable move in a game with no saddlepoint such as this one is bad. Let us now show that the rock-paper-scissors game has no saddlepoint. In finding the maximin and minimax pure strategies, we see that

P1's maximin pure strategy:

$$\begin{aligned} & \max_j \{\min_i a_{ij} = -1, \min_j a_{2j} = -1, \\ & \min_j a_{3j} = -1\} = -1 \end{aligned}$$

P2's minimax pure strategy:

$$\begin{aligned} & \min_i \{\max_j a_{i1} = 1, \max_j a_{i2} = 1, \\ & \max_j a_{i3} = 1\} = 1 \end{aligned}$$

Since maximin $\neq$ minimax, the game has no saddlepoint. Therefore, it would be in the best interests of both players to find their optimal mixed strategies instead. To do this, we now formulate the necessary primal-dual LP problems.

Primal LP.

$$\begin{aligned} & \max v \\ & \text{s.t.} \\ & x_2 - x_3 \geq v \\ & -x_1 + x_3 \geq v \\ & x_1 - x_2 \geq v \\ & x_1 + x_2 + x_3 = 1 \\ & x_i \geq 0. \end{aligned}$$

Dual LP.

$$\begin{aligned}
 &\min u \\
 &\text{s.t.} \\
 &\quad -y_2 + y_3 \leq u \\
 &\quad y_1 \quad -y_3 \leq u \\
 &\quad -y_1 + y_2 \leq u \\
 &\quad y_1 + y_2 + y_3 = 1 \\
 &\quad y_j \geq 0.
 \end{aligned}$$

The optimal solutions for the primal-dual problems are obtained with $v^* = u^* = 0$, $x^* = (1/3, 1/3, 1/3)$, and $y^* = (1/3, 1/3, 1/3)$. Note that both players have identical optimal mixed strategies with game value of zero, so that in the long run, by uniformly but randomly mixing his choices, P1's expected payoff will be zero regardless of what P2's chosen strategy is (and vice versa). These characteristics (identical optimal mixed strategies with game value zero) are always true of symmetric games. A symmetric game is a matrix game with skew-symmetric payoff matrix, that is, $A = -A^T$. Observe that the rock-paper-scissors payoff matrix is skew-symmetric.

Example 2. Consider the following payoff matrix:

$$A = \begin{bmatrix} 1 & 5 & 0 & 2 \\ 2 & -1 & 3 & 3 \\ 4 & 2 & -1 & 0 \end{bmatrix}.$$

Let us first find the maximin and minimax pure strategies for both players to determine whether a saddlepoint exists. We see that row player P1 can select among the three pure strategies available, while column player P2 can select among the four pure strategies available, as shown

$$\begin{aligned}
 &\text{P1's maximin pure strategy: } \max_j \{\min_j a_{1j} = 0, \min_j a_{2j} = -1, \min_j a_{3j} = -1\} = 0 \\
 &\text{P2's minimax pure strategy: } \min_i \{\max_i a_{i1} = 4, \max_i a_{i2} = 5, \max_i a_{i3} = 3, \max_i a_{i4} = 3\} = 3.
 \end{aligned}$$

Since there is no saddle point (maximin $\neq$ minimax), let us now formulate the associated primal-dual LP problems to find the optimal mixed strategies.

Primal LP.

$$\begin{aligned}
 &\max v \\
 &\text{s.t.} \\
 &\quad x_1 + 2x_2 + 4x_3 \geq v \\
 &\quad 5x_1 - x_2 + 2x_3 \geq v \\
 &\quad 3x_2 - x_3 \geq v \\
 &\quad 2x_1 + 3x_2 \geq v \\
 &\quad x_1 + x_2 + x_3 = 1 \\
 &\quad x_i \geq 0.
 \end{aligned}$$

Dual LP.

$$\begin{aligned}
 &\min u \\
 &\text{s.t.} \\
 &\quad y_1 + 5y_2 \quad + 2y_4 \leq u \\
 &\quad 2y_1 - y_2 + 3y_3 + 3y_4 \leq u \\
 &\quad 4y_1 + 2y_2 - y_3 \leq u \\
 &\quad y_1 + y_2 + y_3 + y_4 = 1 \\
 &\quad y_j \geq 0.
 \end{aligned}$$

The optimal solutions for the primal-dual problems are obtained with $v^* = u^* = 13/8$, $x^* = (17/40, 11/20, 1/40)$, and $y^* = (3/8, 1/4, 3/8, 0)$. If both players play these optimal mixed strategies, P1 can expect to win, on average, $13/8$ units. P2 under the same assumption can expect to lose, on average, $13/8$ units. Note that $y_4 = 0$ in P2's optimal strategy, which indicates that the respective column (column 4) can be removed from the payoff matrix with no effect on the game. If we actually realized at the start that column 3 dominates over column 4 in our payoff matrix, then we can deduce that P2 would never select column 4 as her strategy. Therefore, we could intuitively discard column 4 immediately before solving the game.

Example 3. Our next example shows that the strength of LP solution comes out when a nondominated strategy is not used by a player. Consider the following payoff matrix:

$$A = \begin{bmatrix} 1 & 5 & 0 & -7 \\ 2 & -1 & 3 & 11 \\ 4 & 2 & -1 & 7 \end{bmatrix}.$$

Note that column 4 has been replaced in Example 2, and as a result, column 3 (or any other column) no longer dominates column 4. However, it turns out from the LP solution below that this “nondominated” strategy is still not used by P2. First, the maximin and minimax pure strategies for both players are calculated below to determine if saddlepoint exists

$$\text{P1's maximin pure strategy: } \max_j \{\min_j a_{1j} = -1, \min_j a_{2j} = -1, \min_j a_{3j} = -1\} = -1$$

$$\text{P2's minimax pure strategy: } \min_i \{\max_i a_{i1} = 4, \max_i a_{i2} = 5, \max_i a_{i3} = 3, \max_i a_{i4} = 12\} = 3$$

Since there is no saddle point (maximin $\neq$ minimax), the associated primal-dual LP problems are formulated below to find the optimal mixed strategies.

Primal LP.

$$\begin{aligned} \max v \\ \text{s.t.} \\ x_1 + 2x_2 + 4x_3 &\geq v \\ 5x_1 - x_2 + 2x_3 &\geq v \\ 3x_2 - x_3 &\geq v \\ -7x_1 + 11x_2 + 7x_3 &\geq v \\ x_1 + x_2 + x_3 &= 1 \\ x_i &\geq 0. \end{aligned}$$

Dual LP.

$$\begin{aligned} \min u \\ \text{s.t.} \\ y_1 + 5y_2 - 7y_4 &\leq u \\ 2y_1 - y_2 + 3y_3 + 11y_4 &\leq u \\ 4y_1 + 2y_2 - y_3 + 7y_4 &\leq u \\ y_1 + y_2 + y_3 + y_4 &= 1 \\ y_j &\geq 0. \end{aligned}$$

The optimal solutions for the primal-dual problems are obtained with $v^* = u^* = 13/8$, $x^* = (17/40, 11/20, 1/40)$, and $y^* = (3/8, 1/4, 3/8, 0)$. These are identical results found in Example 2, which indicates that column 4 could have been discarded at the onset with no effect on the game. However, unlike Example 2, this is not so obvious since none of the other columns dominate column 4. Moreover, as what seems to be a “nondominated” strategy, it is still somewhat surprising that P2 would never play it. The reason is that column 4 is actually dominated by a linear combination of the other columns,

$$\text{namely, } 3A_1 - 2A_2 + A_3 = \begin{bmatrix} -7 \\ 11 \\ 7 \end{bmatrix}, \text{ where}$$

A_j is the j th column of A . This “hidden” domination is nonetheless revealed through the LP solution. To get the coefficients of the linear combination, simply identify the ordered basic variables from the optimal simplex tableau (which in this case are y_1 , y_2 , and y_3) to construct the “basis” matrix B , by using the corresponding columns of the payoff matrix A associated with the ordered basic variables. The product $B^{-1}A_4 = [A_1 \ A_2 \ A_3]^{-1}A_4 = \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}$ will then provide the coefficients.

In general, the product $B^{-1}A_k$ gives the coefficients of the linear combination for any nonbasic variable (whose index is k) in terms of the ordered basic variables.

REDUCTION OF LINEAR PROGRAMMING PROBLEMS INTO MATRIX GAMES

We now consider the converse problem of converting linear programming problems into matrix games. More specifically, we will discuss how to construct the symmetric game that will be associated with a given primal LP problem in three different forms: normal, standard, and mixed. Recall that a symmetric game is a matrix game with skew-symmetric payoff matrix.

Normal LP

We begin by analyzing the following *normal* form of a primal-dual pair of a linear programming problem in matrix-vector notation

Primal LP.

$$\begin{aligned} \min \quad & z = c^T y \\ \text{s.t.} \quad & \\ & Ay \geq b \\ & y \geq 0. \end{aligned}$$

Dual LP.

$$\begin{aligned} \max \quad & w = b^T x \\ \text{s.t.} \quad & \\ & A^T x \leq c \\ & x \geq 0. \end{aligned}$$

In both problems, c is $(n \times 1)$ vector, x is $(m \times 1)$ vector, y is $(n \times 1)$ vector, b is $(m \times 1)$ vector, and A is an $(m \times n)$ general matrix.

The symmetric game equivalent to the primal-dual problems above is given by the $(m+n+1) \times (m+n+1)$ block skew-symmetric payoff matrix [1]

$$S = \begin{bmatrix} 0 & -A & b \\ A^T & 0 & -c \\ -b^T & c^T & 0 \end{bmatrix}.$$

Here, the block matrix 0 indicates a matrix of appropriate dimensions, all of whose components are zero. Observe that S can be

constructed completely from the given primal LP problem. The associated dual problem need not be formulated at all.

Let us now examine how this matrix S is related to the primal-dual LP problems. Let the block $m+n+1$ vector $(x; y; \lambda)^T$ represent the mixed strategy vector for P2. As S is skew-symmetric, the matrix game has, of course, a value of zero and, therefore, P2's goal is to find an optimal strategy that will satisfy the inequality

$$\begin{bmatrix} 0 & -A & b \\ A^T & 0 & -c \\ -b^T & c^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ \lambda \end{bmatrix} \leq 0. \quad (1)$$

Here, the right-hand side is a column zero vector with $m+n+1$ components. Moreover, from the properties of symmetric games, an optimal mixed strategy for P2 is also the optimal strategy for P1.

To see that solving inequality (1) is equivalent to solving the associated primal-dual LP problems, suppose the block vector $(x; y; \lambda)^T$ is an optimal mixed strategy for both players. The optimal primal and dual solutions to the LP can then be calculated as follows:

$$\begin{aligned} y^{\text{opt}} &= y/\lambda \\ x^{\text{opt}} &= x/\lambda \end{aligned} \quad (2)$$

provided, of course, that $\lambda \neq 0$. With these calculations of the LP primal-dual solutions, one can perform the block matrix-vector multiplication in the inequality (1) above, and quickly realize that satisfying the above inequality does indeed correspond to satisfying the primal-dual LP constraints including the dual relationship between the objective functions.

Although a solution to S always exists (as is true of any matrix game), the associated LP, however, may not have any solution. As seen from Equation (2), this happens when $\lambda = 0$ in the optimal mixed strategy solution.

The transformation procedure described above can be adapted easily when the primal-dual LP problems are reversed; namely, when the primal LP being considered is a maximization problem with " $\leq$ " constraints and the dual LP is the corresponding minimization problem with " $\geq$ " constraints. In this

case, the resulting skew-symmetric matrix is given by

$$S_1 = \begin{bmatrix} 0 & A^T & -c \\ -A & 0 & b \\ c^T & -b^T & 0 \end{bmatrix}$$

with the block $m + n + 1$ vector $(y; x; \lambda)^T$ representing the optimal mixed strategy vector for both players. The solutions for the associated primal-dual LP is calculated in an analogous manner as before.

Once the skew-symmetric matrix is formed, an iterative procedure such as Brown's Fictitious Play Method [6] can be used to solve the game, and hence the associated LP (see **Fictitious Play Algorithm**). However, current wisdom dictates that solving an LP using either the simplex method or any of the interior point methods would be vastly superior compared to any game-theoretic computational methods. A hybrid approach might be more effective though, particularly for solving large games. Gass and Zafra [7] described initial experiments with encouraging results in which their Modified Fictitious Play Method can be combined with simplex and interior points to aid in the solution of LP problems.

Example 4. Consider the following LP problem:

$$\begin{aligned} \min z &= 40y_1 + 25y_2 + 10y_3 \\ \text{s.t.} \\ y_1 + 5y_2 &\geq 4 \\ 2y_1 - y_2 + y_3 &\geq 8 \\ y_j &\geq 0. \end{aligned}$$

The associated dual LP is given by

$$\begin{aligned} \max w &= 4x_1 + 8x_2 \\ \text{s.t.} \\ x_1 + 2x_2 &\leq 40 \\ 5x_1 - x_2 &\leq 25 \\ x_2 &\leq 10 \\ x_i &\geq 0. \end{aligned}$$

From the given primal problem, we see that

$$A = \begin{bmatrix} 1 & 5 & 0 \\ 2 & -1 & 1 \end{bmatrix}, b = \begin{bmatrix} 4 \\ 8 \end{bmatrix},$$

and $c = \begin{bmatrix} 40 \\ 25 \\ 10 \end{bmatrix}.$

We use this information to construct the skew-symmetric matrix for the corresponding symmetric game as shown below

$$S = \begin{bmatrix} 0 & 0 & -1 & -5 & 0 & 4 \\ 0 & 0 & -2 & 1 & -1 & 8 \\ 1 & 2 & 0 & 0 & 0 & -40 \\ 5 & -1 & 0 & 0 & 0 & -25 \\ 0 & 1 & 0 & 0 & 0 & -10 \\ -4 & -8 & 10 & 25 & 10 & 0 \end{bmatrix}$$

Standard LP

We describe next how to transform primal LP problems with equality constraints (commonly referred to as *LP in standard form*) into a symmetric game. The basic idea is to first transform the given primal LP problem with minimizing objective function into normal form (i.e., with constraints consisting only of " $\geq$ " inequalities).

Let us now consider the following primal-dual pair of a linear programming problem, with the primal problem given in standard form

Primal LP.

$$\begin{aligned} \min \quad z &= c^T y \\ \text{s.t.} \\ Ay &= b \\ y &\geq 0. \end{aligned}$$

Dual LP.

$$\begin{aligned} \max \quad w &= b^T x \\ \text{s.t.} \\ A^T x &\leq c \\ x, &\text{ unrestricted.} \end{aligned}$$

Here, A, c, b, x , and y have the same dimensions as noted in the “normal” case.

Let us now transform the primal LP into normal form and let $x = x_1 - x_2$. By duality theory of linear programming, we have the equivalent primal-dual pair as follows:

Primal LP.

$$\begin{aligned} \min \quad & z = c^T y \\ \text{s.t.} \quad & \\ & Ay \geq b \\ & -Ay \geq -b \\ & y \geq 0. \end{aligned}$$

Dual LP.

$$\begin{aligned} \max \quad & w = b^T x_1 - b^T x_2 \\ \text{s.t.} \quad & \\ & A^T x_1 - A^T x_2 \leq c \\ & x_1, x_2 \geq 0. \end{aligned}$$

The corresponding $(2m + n + 1) \times (2m + n + 1)$ skew-symmetric game matrix can now be constructed as follows:

$$S_2 = \begin{bmatrix} 0 & 0 & -A & b \\ 0 & 0 & A & -b \\ A^T & -A^T & 0 & -c \\ -b^T & b^T & c^T & 0 \end{bmatrix}.$$

As in S , the block matrix 0 indicates a matrix of appropriate dimensions, all of whose components are zero. Observe that the skew-symmetric matrix is enlarged by additional m rows and m columns due to the “splitting” of the vector x in terms of two vectors x_1 and x_2 . Again, observe that the matrix S_2 can be constructed using only the information from the given primal LP problem.

Now, if the block $2m + n + 1$ vector $(x_1; x_2; y; \lambda)$ is an optimal strategy for S_2 , then the associated LP problem has the optimal primal-dual solution

$$\begin{aligned} y^{\text{opt}} &= y/\lambda \\ x^{\text{opt}} &= (x_1 - x_2)/\lambda \end{aligned}$$

provided again that $\lambda \neq 0$.

Example 5. Consider the primal LP problem in Example 4 where we now change the inequalities in the constraints into equations. The new linear problem is shown below

$$\begin{aligned} \min \quad & z = 40y_1 + 25y_2 + 10y_3 \\ \text{s.t.} \quad & \\ & y_1 + 5y_2 = 4 \\ & 2y_1 - y_2 + y_3 = 8 \\ & y_j \geq 0. \end{aligned}$$

As in Example 4, we see that $A = \begin{bmatrix} 1 & 5 & 0 \\ 2 & -1 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$, and $c = \begin{bmatrix} 40 \\ 25 \\ 10 \end{bmatrix}$. However, the skew-symmetric matrix for this new problem is a larger matrix S_2 as shown

$$S_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & -5 & 0 & 4 \\ 0 & 0 & 0 & 0 & -2 & 1 & -1 & 8 \\ 0 & 0 & 0 & 0 & 1 & 5 & 0 & -4 \\ 0 & 0 & 0 & 0 & 2 & -1 & 1 & -8 \\ 1 & 2 & -1 & -2 & 0 & 0 & 0 & -40 \\ 5 & -1 & -5 & 1 & 0 & 0 & 0 & -25 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & -10 \\ -4 & -8 & 4 & 8 & 40 & 25 & 10 & 0 \end{bmatrix}$$

Mixed LP

Finally, we describe the skew-symmetric matrix associated with minimization LP problems having mixed constraints (equations and inequalities). We apply the same idea introduced in the preceding case and convert the LP into normal form. To start, we express each equality constraints equivalently into two inequalities of the form “ $\geq$ ” and “ $\leq$ ”. In addition, each “ $\leq$ ” constraint is also transformed into “ $\geq$ ” constraint, as was done in the standard LP case. In an analogous manner that S_2 was constructed, the resulting skew-symmetric for the “mixed” LP is given by

$$S_3 = \begin{bmatrix} 0 & 0 & -A & b \\ 0 & 0 & \tilde{A} & -\tilde{b} \\ A^T & -\tilde{A}^T & 0 & -c \\ -b^T & \tilde{b}^T & c^T & 0 \end{bmatrix},$$

where the matrix $\tilde{A}$ and vector $\tilde{b}$ are associated with “ $\leq$ ” inequalities arising from the equations (that were transformed later into “ $\geq$ ”), while the matrix A consists of the elements associated with the regular “ $\geq$ ” inequalities and the original “ $\leq$ ” inequalities (that were transformed later into “ $\geq$ ”). As in S_2 , the block matrix 0 indicates a matrix of appropriate dimensions, all of whose components are zero. Moreover, as in the other previous cases, observe that the matrix S_3 can be constructed using only the information from the given primal LP problem; the corresponding dual problem need not be formulated.

One can quickly notice the similarity of the structures between the matrices S_2 and S_3 . However, on closer inspection, S_3 is more similar to the matrix S of the normal primal LP as can be seen in Example 6 below. This is due to the fact that the mixed constraints were all transformed into “ $\geq$ ” inequalities, resulting thereby in a normal form of the LP problem. The similarity between S and S_3 ends here, because although the optimal mixed strategy in regards to both game matrices have the same form, namely, the block vector $(x; y; \lambda)^T$, the number of components of the vector x in the case of S_3 is not predetermined (recall that in S , x has m components). It depends on how many equations are present in the constraints of the given primal LP problem (each equation adds an additional component to the vector x). To be more specific, if A is $(m \times n)$ and there are k equations among the m constraints, then the vector x will have $m + k$ number of components. Since the number of components of the vector y is known (which is n , of course), then the optimal primal solution to the LP is easily calculated as in S

$$y^{\text{opt}} = y/\lambda,$$

provided again that $\lambda \neq 0$. If needed, the optimal solutions to the corresponding dual LP has to be manually determined, but it is not difficult to do so.

Example 6. Consider the following LP problem with mixed types of constraints:

$$\begin{aligned} \min z &= 40y_1 + 25y_2 + 10y_3 \\ \text{s.t.} \\ y_1 &\quad - y_3 \geq 3 \\ y_1 + 5y_2 &\leq 4 \\ 2y_1 - y_2 + y_3 &= 8 \\ y_j &\geq 0. \end{aligned}$$

Applying the transformation with regard to the equality and “ $\leq$ ” constraints, we obtain the normal form of the primal LP

$$\begin{aligned} \min z &= 40y_1 + 25y_2 + 10y_3 \\ \text{s.t.} \\ y_1 &\quad - y_3 \geq 3 \\ -y_1 - 5y_2 &\geq 4 \\ 2y_1 - y_2 + y_3 &\geq 8 \\ -2y_1 + y_2 - y_3 &\geq -8 \\ y_j &\geq 0. \end{aligned}$$

We see that

$$A = \begin{bmatrix} 1 & 0 & -1 \\ -1 & -5 & 0 \\ 2 & -1 & 1 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} 2 & -1 & 1 \end{bmatrix},$$

$$b = \begin{bmatrix} 3 \\ 4 \\ 8 \end{bmatrix}, \quad \tilde{b} = [8], \quad \text{and } c = \begin{bmatrix} 40 \\ 25 \\ 10 \end{bmatrix}.$$

Therefore, the skew-symmetric matrix for the equivalent game is as given below

$$S_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 & 5 & 0 & 4 \\ 0 & 0 & 0 & 0 & -2 & 1 & -1 & 8 \\ 0 & 0 & 0 & 0 & 2 & -1 & 1 & -8 \\ 1 & -1 & 2 & -2 & 0 & 0 & 0 & -40 \\ 0 & -5 & -1 & 1 & 0 & 0 & 0 & -25 \\ -1 & 0 & 1 & -1 & 0 & 0 & 0 & -10 \\ -3 & -4 & -8 & 8 & 40 & 25 & 10 & 0 \end{bmatrix}$$

For this game matrix, the block vector for the mixed strategy is $(x; y; \lambda)^T = (x_1, x_2, x_3, x_4, y_1, y_2, y_3, \lambda)^T$. Note that the vector x has $4(3 + 1)$ components: one more than

the number of rows of A in the given LP, because there is one equation present in the constraints.

CONCLUDING REMARKS

The equivalence between two-person zero-sum games and linear programming was established in the late 1940s and early 1950s by Dantzig, Gale Kuhn, von Neuman, Tucker and others. When the concept of duality in LP proved to be equivalent to Von Neumann's game theory duality in his Minimax Theorem, which was published earlier in 1928. As discussed in the section titled "Transforming Matrix Games into Linear Programming Problems," this remarkable result stems from the transformation of a matrix game into a pair of primal-dual LP problems. Gale *et al.* [2] later illustrated the converse situation of transforming an LP problem into a symmetric game (i.e., a matrix game with skew-symmetric matrix), establishing the full equivalence between matrix games and LP problems.

Although a number of game-theoretic methods for solving matrix games have been developed early, such as Brown's fictitious play method [6], Brown and von Neumann's differential equation approach [8], and the numerical method of von Neumann [9], it has long been recognized that an efficient way to solve most matrix games is by converting them to equivalent linear programming problems [10]. The resultant LP problem is then solved using a mathematical programming software based on either the simplex or interior point methods. However, faster game-theoretic methods and algorithms are still being developed to solve matrix games directly, such as modifications to the fictitious play method [11,12], fast algorithms to determine whether there is a saddlepoint [13,14], sublinear-time approximation algorithms [15], sequential methods in finding the equilibria [16], iterated technique in finding Nash equilibrium [17], and first-order algorithms that match the complexity of interior-point methods and use less memory [18].

As to the relationship itself between matrix games and linear programming, a

new research trend seems to be focusing on fuzzy linear programs and fuzzy payoffs [19,20]. Interested readers are referred to Nishizaki and Sakawa [21] or Bector and Chandra [22] for an overview of this evolving facet in the study of linear programming and two-person zero-sum games.

REFERENCES

1. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
2. Gale D, Kuhn HW, Tucker AW. On symmetric games. In: Kuhn HW, Tucker AW, editors. Volume 1. Contributions to the theory of games, Annals of mathematics study No. 24. Princeton (NJ): Princeton University Press; 1950. pp. 73–79.
3. Gass SI. Linear programming. 5th ed. New York: McGraw-Hill Book Company; 1985.
4. von Neumann J. Zur theorie der gesellschaftsspiele. Math Ann 1928;100: 295–320.
5. Vanderbei RJ. Linear programming: foundations and extensions. Boston (MA): Springer, Kluwer Academic Publishers; 2008.
6. Brown GW. Iterative solutions of games by fictitious play. In: Koopmans TC, editor. Activity analysis of production and allocation, Cowles commission monograph 13. New York: Wiley; 1951. (Originally published as: Some Notes on Computation of Games Solutions. RAND Report 78. Santa Monica (CA): The RAND Corporation. 1949. pp. 377–380).
7. Gass SI, Zafra PMR. Modified fictitious play for solving matrix games and linear programming problems. Comput Oper Res 1995;22:893–903.
8. Brown GW, von Neumann J. Solutions of games by differential equations. In: Kuhn HW, Tucker AW, editors. Volume 1, Contributions to the theory of games, Annals of Mathematics Study No. 24. Princeton (NJ): Princeton University Press; 1950. pp. 73–79.
9. Von Neumann J. A numerical method to determine optimum strategy. Nav Res Log Q 1954;1:109–115.
10. Szép J, Forgó F. Introduction to the theory of games. Holland: D. Reidel Publishing Company; 1985.
11. Gass SI, Zafra PMR, Qiu Z. Modified fictitious play. Nav Res Log 1996;43:955–970.

12. Washburn A. A new kind of fictitious play. *Nav Res Log* 2001;48:270–280.
13. Llewellyn DC, Tovey C, Trick M. Finding saddlepoints of two-person zero-sum games. *Am Math Mon* 1988;95:912–918.
14. Bienstock D, Chaung F, Fredman M, *et al.* A note on finding a strict saddlepoint. *Am Math Mon* 1991;98:418–419.
15. Grigoriadis MD, Khachiyan LG. A sub-linear—time randomized approximation algorithm for matrix games. *Oper Res Lett* 1995;18:53–58.
16. Miltersen PB, Sorensen TB. Computing proper equilibria of zero-sum games. In: van den Herik HJ, Ciancarini P, Donkers HHLM, editors. *Lecture notes in computer science* 4630. Heidelberg: Springer; 2007. pp. 200–211.
17. Zinkevich M, Bowling M, Burch N. A new algorithm for generating equilibria in massive zero-sum games. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, British Columbia; 2007. pp. 788–793.
18. Gilpin A, Pena J, Sandholm T. First-order algorithm with $O(\ln(1/\varepsilon))$ convergence for ε -equilibrium in two-person zero-sum games. *Proceedings of the AAAI Conference on Artificial Intelligence*. Chicago (IL); 2008. pp. 75–82.
19. Gao Z, Han C, Zhang H, *et al.* Duality in linear programming with fuzzy parameters and two-person zero-sum constrained matrix games with fuzzy payoffs. *Proceedings of the Chinese Control and Decision Conference*. Yantai, China; 2008. pp. 1192–1195.
20. Xu L, Wang X. Fuzzy linear programming to solve two-person zero-sum matrix game with fuzzy payoffs. *International Joint Conference on Computational Sciences and Optimization, CSO, Volume 2*. Sanya, Hainan Island; 2009. pp. 812–815.
21. Nishizaki I, Sakawa M. *Fuzzy and multiobjective games for conflict resolution*. Heidelberg: Springer; 2001.
22. Bector R, Chandra S. *Fuzzy mathematical programming and matrix games*. Berlin: Springer; 2005.

LINEAR PROGRAMMING PROJECTION ALGORITHMS

YURY EVTUSHENKO
ALEXANDER GOLIKOV
Computing Centre of Russian
Academy of Sciences, Moscow,
Russia

Large-scale LP (linear programming) problems usually have more than one solution. Techniques such as the simplex methods and interior point methods make it possible to obtain different solutions in the case of nonuniqueness. For example, the simplex method yields a solution belonging to a vertex of polyhedron. Some variants of the interior point method converge to a solution satisfying the strict complementary slackness condition.

LP projection method is close to the quadratic penalty function method and to the modified Lagrangian function method. This method yields the exact projection of a given point on the solution set of primal LP problem as a result of the single unconstrained maximization of an auxiliary piecewise quadratic concave function for any sufficiently large values of the penalty parameter. A generalized Newton method with a stepsize chosen using Armijo's rule was used for unconstrained maximization. The proof of globally convergent finitely terminating generalized Newton method for piecewise quadratic function was given in Mangasarian [1] and Kanzow *et al.* [2]. LP projection method solves LP problems with a very large ($\approx 10^7$) number of variables and moderate ($\approx 10^5$) number of constraints.

In a similar way, the exact projection of a given point on the solution set of the dual LP problem can be obtained by nonnegative constrained maximization of auxiliary quadratic function for sufficiently large but finite values of the penalty parameter.

Consider the primal linear program in the standard form

$$f_* = \min_{x \in X} c^\top x, X = \{x \in R^n : Ax = b, x \geq 0_n\} \quad (P)$$

together with its dual

$$f_* = \max_{u \in U} b^\top u, U = \{u \in R^m : A^\top u \leq c\}, \quad (D)$$

where $A \in R^{m \times n}$, $c \in R^n$, and $b \in R^m$ are given, x is a primal variable and u is a dual variable, 0_i denotes the i -dimensional zero vector.

Everywhere we assume that the solution set X_* of the primal problem (P) is nonempty, and hence the solution set U_* of the dual problem (D) is also nonempty.

Consider the problems of finding the least 2-norm projection $\hat{x}_*$ of the point $\hat{x}$ on the solution set X_* and the least 2-norm projection $\hat{u}_*$ of the point $\hat{u}$ on the solution set U_* :

$$\frac{1}{2} \|\hat{x}_* - \hat{x}\|^2 = \min_{x \in X_*} \frac{1}{2} \|x - \hat{x}\|^2,$$

$$X_* = \{x \in R^n : Ax = b, c^\top x = f_*, x \geq 0_n\}, \quad (1)$$

$$\frac{1}{2} \|\hat{u}_* - \hat{u}\|^2 = \min_{u \in U_*} \frac{1}{2} \|u - \hat{u}\|^2,$$

$$U_* = \{u \in R^m : A^\top u \leq c, b^\top u = f_*\}. \quad (2)$$

Here, the Euclidian norm of vectors is used, and f_* is an *a priori* unknown optimal value of the objective function of the original LP problems (P) and (D).

Let us introduce the Lagrange functions for these problems:

$$L^1(x, p, \beta) = \frac{1}{2} \|x - \hat{x}\|^2 + p^\top (b - Ax) + \beta(c^\top x - f_*),$$

$$L^2(u, y, \alpha) = \frac{1}{2} \|u - \hat{u}\|^2 + y^\top (A^\top u - c) + \alpha(f_* - b^\top u).$$

Here, $p \in R^m$, $\beta \in R^1$, and $y \in R_+^n$, $\alpha \in R^1$ are Lagrange multipliers for problems (1) and (2), respectively.

The problems dual to problems (1) and (2) have the forms

$$\max_{p \in R^m} \min_{\beta \in R^1} L^1(x, p, \beta), \quad (3)$$

$$\max_{y \in R_+^n} \min_{\alpha \in R^1} L^2(u, y, \alpha). \quad (4)$$

The solutions of the inner minimization problems in Equations (3) and (4) have the following forms, respectively:

$$x = (\hat{x} - A^\top p - \beta c)_+, \quad (5)$$

$$u = \hat{u} + \alpha b - Ay, \quad (6)$$

where a_+ denotes the vector a with all negative components replaced by zeros.

Substituting Equation (5) into the Lagrange function $L^1(x, p, \beta)$ and Equation (6) into the Lagrange function $L^2(u, y, \alpha)$, we obtain the dual functions of problems (3) and (6), respectively:

$$\begin{aligned} \hat{L}^1(p, \beta) &= b^\top p - \frac{1}{2} \|\hat{x} + A^\top p - \beta c\|_+^2 \\ &\quad - \beta f_* + \frac{1}{2} \|\hat{x}\|^2, \\ \hat{L}^2(y, \alpha) &= -c^\top y - \hat{u}^\top (\alpha b - Ay) \\ &\quad - \frac{1}{2} \|\alpha b - Ay\|^2 + \alpha f_*. \end{aligned}$$

The function $\hat{L}^1(p, \beta)$ is concave, piecewise quadratic, and continuously differentiable in variables p and β . The function $\hat{L}^2(y, \alpha)$ is concave, quadratic, and twice continuously differentiable in both variables.

Dual problems (3) and (4) are reduced to the outer maximization problems, respectively:

$$\max_{p \in R^m} \hat{L}^1(p, \beta), \quad (7)$$

$$\max_{y \in R_+^n} \hat{L}^2(y, \alpha). \quad (8)$$

Solving problem (7), we find optimal p and β . Substituting them into Equation (5), we

obtain $\hat{x}_*$, which is the projection of the point $\hat{x}$ on the solution set of the primal LP problem (P).

Solving problem (8), we find optimal y and α . Substituting them into Equation (6), we obtain $\hat{u}_*$, which is the projection of the point $\hat{u}$ on the solution set of the dual LP problem (D).

Unfortunately, maximization problems (7) and (8) contain an *a priori* unknown value f_* , which is the optimal value of the objective function of the original LP problem. However, these problems can be simplified by eliminating this difficulty. For this purpose, the following simplified unconstrained maximization problem is solved instead of problem (7):

$$\begin{aligned} \max_{p \in R^m} S^1(p, \beta, \hat{x}), \text{ where } S^1(p, \beta, \hat{x}) \\ = b^\top p - \frac{1}{2} \|\hat{x} + A^\top p - \beta c\|_+^2. \end{aligned} \quad (9)$$

Here the scalar β is fixed.

Instead of problem (8), the following simplified maximization problem on the positive orthant is solved:

$$\begin{aligned} \max_{y \in R_+^n} S^2(y, \alpha, \hat{u}), \text{ where } S^2(y, \alpha, \hat{u}) \\ = -c^\top y + \hat{u}^\top Ay - \frac{1}{2} \|\alpha b - Ay\|^2. \end{aligned} \quad (10)$$

Here the scalar α is also fixed.

Let us first consider problem (9) and its relation to the primal LP problem (P). Note that, in contrast to problem (1), dual problem (7) has many solutions. Naturally, the question that arises is of finding the minimal value β_* of the Lagrange multiplier β among all solutions of problem (7). Once such β_* is found, one can fix $\beta \geq \beta_*$ in dual problem (7) and maximize the dual function $\hat{L}^1(p, \beta)$ only with respect to variable p , that is, solve problem (9). In this case, the pair $[p, \beta]$ is a solution to problem (7), and the triplet $[\hat{x}_*, p, \beta]$ is a saddle point of problem (1), where the projection $\hat{x}_*$ of the point $\hat{x}$ on the solution set X_* is defined by problem (5).

Theorem 1 [3]. *There exists β_* such that for all $\beta \geq \beta_*$ the unique least 2-norm projection $\hat{x}_*$ of a point $\hat{x}$ on X_* is given by*

$$\hat{x}_* = (\hat{x} + A^\top p(\beta) - \beta c)_+,$$

where $p(\beta)$ is a maximizer of $S^1(p, \beta, \hat{x})$ in problem (9).

Theorem 1 makes it possible to replace problem (7), which contains an *a priori* unknown value f_* , by problem (9), which involves the half-interval $[\beta_*, +\infty]$ instead of this value. This essentially simplifies the calculations. Note that the value β_* may be negative. This occurs when the projection of the point $\hat{x}$ on the solution set X_* coincides with the projection of this point on the feasible set X . The estimation of the threshold value β_* is given in Golikov and Evtushenko [3,5].

The next theorem tells us that a solution to problem (D) can be obtained from the single unconstrained maximization problem (9) if a point $\hat{x}_* \in X_*$ is found according to Theorem 1.

Theorem 2 [3]. *For all $\beta > 0$ and all $\hat{x} = x_* \in X_*$ an exact solution to dual problem (D) is given by $u_* = p(\beta)/\beta$, where $p(\beta)$ is a solution to the unconstrained maximization problem (9).*

To solve the primal and dual LP problems simultaneously, one can use the following iterative process:

$$p_{s+1} \in \arg \max_{p \in R^m} \left\{ b^\top p - \frac{1}{2} \| (x_s + A^\top p - \beta c)_+ \|^2 \right\} \quad (11)$$

$$x_{s+1} = (x_s + A^\top p_{s+1} - \beta c)_+, \quad (12)$$

where x_0 is an arbitrary starting point.

Theorem 3 [3]. *For all $\beta > 0$ and for arbitrary starting point x_0 , the iterative process (11)–(12) converges to $x_* \in X_*$ in a finite number of iterations ω . The formula $u_* = p_{\omega+1}/\beta$ provides an exact solution to the dual problem (D).*

Iterative process (11,12) is finite. It gives an exact solution to the primal problem (P) and the exact solution to the dual problem (D). Note that one can use this method even when one is unaware of the threshold value β_* . However, if the chosen coefficient is less than the threshold value, then this method, in a finite number of steps, yields some solution x_* to the primal problem (P) which is not a projection of the starting point x_0 on the solution set X_* . Note that $x_\omega = x_* \in X_*$ is the projection of the point $x_{\omega-1}$ on X_* .

The subsequent theorems are similar to Theorems 1–3.

Theorem 4 [4]. *There exists α_* such that for all $\alpha \geq \alpha_*$ the unique least 2-norm projection $\hat{u}_*$ of a point $\hat{u}$ on U_* is given by*

$$\hat{u}_* = \hat{u} + \alpha b - Ay(\alpha),$$

where $y(\alpha)$ is a point maximizing $S^2(y, \alpha, \hat{u})$ on R_+^n .

The estimation of the threshold value β_* is given in Evtushenko *et al.* [4].

Theorem 5 [4]. *For all $\alpha > 0$ and all $\hat{u} = u_* \in U_*$ an exact solution to primal problem (P) is given by $u_* = y(\alpha)/\alpha$, where $y(\alpha)$ is a point maximizing $S^2(y, \alpha, u_*)$ on R_+^n .*

To solve the primal and dual LP problems simultaneously, one can use the following iterative process:

$$y_{s+1} \in \arg \max_{y \in R_+^n} \left\{ -c^\top y + u_s^\top Ay - \frac{1}{2} \|\alpha b - Ay\|^2 \right\}, \quad (13)$$

$$u_{s+1} = u_s + \alpha b - Ay_{s+1}. \quad (14)$$

Theorem 6 [4]. *For all $\alpha > 0$ and for arbitrary starting point u_0 , the iterative process (13)–(14) converges to $u_* \in U_*$ in a finite number of iterations v . The formula $x_* = y_{v+1}/\alpha$ provides an exact solution to primal LP problem (P).*

Note that $u_* \in U_*$ is the projection of the point u_{v-1} on the solution set U_* of the problem (D).

The unconstrained maximization in Equations (9) and (11) can be performed by any method, for example, by the conjugate gradient method. However, the generalized Newton method [1] turns out to be much more efficient.

The objective function $S^1(p, \beta, \hat{x})$ of problems (9) is concave, piecewise quadratic, and differentiable. The ordinary Hessian does not exist for this function because the gradient

$$\frac{\partial}{\partial p} S^1(p, \beta, \hat{x}) = b - A(\hat{x} + A^\top p - \beta c)_+.$$

is not differentiable. However, for this function, one can define the generalized Hessian matrix, which is an $m \times m$ symmetric negative semidefinite matrix of the form

$$\frac{\partial^2}{\partial p^2} S^1(p, \beta, \hat{x}) = -AD(z)A^\top.$$

Here, $D(z)$ denotes the $n \times n$ diagonal matrix whose i th diagonal entry z^i is equal to 1 if $(\hat{x} + A^\top p - \beta c)^i > 0$ and z^i is equal to zero if $(\hat{x} + A^\top p - \beta c)^i \leq 0$ ($i = 1, 2, \dots, n$). The generalized Hessian matrix may be singular. Therefore, we use the following generalized Newton method:

$$p_{s+1} = p_s - \mu \left[\frac{\partial^2}{\partial p^2} S^1(p, \beta, \hat{x}) - \delta I_m \right]^{-1} S_p(p, \beta, \hat{x}),$$

where δ is a small positive number (typically $\delta = 10^{-4}$), I_m is the identity matrix of size $m \times m$, and μ is a stepsize chosen using Armijo's rule.

Unfortunately the generalized Newton method cannot be applied to the nonnegativity constrained maximization problem (10) directly. By incorporating the nonnegativity constraint $y \geq 0_n$ into the objective function of problem (10) as a penalty term, we have the unconstrained maximization problem

$$\max_{y \in R^n} \left\{ -c^\top y + \hat{u}^\top A y - \frac{1}{2} \|\alpha b - A y\|^2 - \frac{\tau}{2} \|(-y)_+\|^2 \right\},$$

where $\tau > 0$ is a penalty parameter. In this case, we obtained the optimal solution y only in limit as $\tau \rightarrow +\infty$. This property complicates the computation. There exists a very simple way to overcome this shortcoming.

Let a vector $w \in R^{m+n}$ consist of two vectors $w^\top = [u^\top, v^\top]$, where $u \in R^m, v \in R^n$. Consider the following LP problem

$$\begin{aligned} f_* &= \max_{w \in W} b^\top u, \\ W &= \{u \in R^m, v \in R^n : A^\top u + v = c, v \geq 0_n\}, \end{aligned} \quad (D')$$

which is equivalent to dual problem (D). The solution set of this problem is denoted by $W_* = [U_* \times V_*]$. For a given point $\hat{w}$ we find the least 2-norm projection $\hat{w}_*$ on W_* as a solution to the following minimization problem:

$$\begin{aligned} & \frac{1}{2} \|\hat{u}_* - \hat{u}\|^2 + \frac{1}{2} \|\hat{v}_* - \hat{v}\|^2 \\ &= \min_{w \in W_*} \left\{ \frac{1}{2} \|u - \hat{u}\|^2 + \frac{1}{2} \|v - \hat{v}\|^2 \right\}, \\ W_* &= \{u \in R^m, v \in R^n : A^\top u + v = c, v \geq 0_n, b^\top u = f_*\}. \end{aligned}$$

Using an approach similar to that used above we arrive at the following unconstrained maximization problem:

$$\begin{aligned} & \max_{y \in R^n} S^3(y, \gamma, \hat{w}), \text{ where } S^3(y, \gamma, \hat{w}) \\ &= -c^\top y + \hat{u}^\top A y \\ & \quad - \frac{1}{2} \|\gamma b - A y\|^2 - \frac{1}{2} \|(\hat{v} - y)_+\|^2. \end{aligned} \quad (15)$$

The following theorems hold.

Theorem 7. *There exists γ_* such that for all $\gamma \geq \gamma_*$ the unique least 2-norm projection $\hat{w}_*^\top = [\hat{u}_*^\top, \hat{v}_*^\top]$ of a point $\hat{w}^\top = [\hat{u}^\top, \hat{v}^\top]$ on W_* is given by*

$$\begin{aligned} \hat{u}_* &= \hat{u} + \gamma b - A y(\gamma), \\ \hat{v}_* &= (\hat{v} - y(\gamma))_+, \end{aligned}$$

where $y(\gamma)$ is a solution to the unconstrained maximization problem (15).

Table 1. Comparative Results

$m \times n \times \rho$	Solver	T (s)	Iterations	Δ_1	Δ_2	Δ_3
$500 \times 10^4 \times 1$	LPP (MATLAB)	55.0	12	1.5×10^{-8}	1.8×10^{-12}	1.2×10^{-7}
	BPMPD (interior point)	37.4	23	2.3×10^{-10}	1.8×10^{-11}	1.1×10^{-10}
	MOSEK (interior point)	87.2	6	9.7×10^{-8}	3.8×10^{-9}	1.6×10^{-6}
	CPLEX (interior point)	80.3	11	1.8×10^{-8}	1.1×10^{-7}	0.0
	CPLEX (simplex)	61.8	8308	8.6×10^{-4}	1.9×10^{-10}	7.2×10^{-3}
$3000 \times 10^4 \times 0.01$	LPP (MATLAB)	155.4	11	6.1×10^{-10}	3.4×10^{-13}	3.6×10^{-8}
	BPMPD (interior point)	223.5	14	4.6×10^{-9}	2.9×10^{-10}	3.9×10^{-9}
	MOSEK (interior point)	42.6	4	3.1×10^{-8}	1.2×10^{-8}	3.7×10^{-8}
	CPLEX (interior point)	69.9	5	1.1×10^{-6}	1.3×10^{-7}	0.0
	CPLEX (simplex)	1764.9	6904	3.0×10^{-3}	8.1×10^{-9}	9.3×10^{-2}
$1000 \times (5 \times 10^6) \times 0.01$	LPP (MATLAB)	1007.5	10	3.9×10^{-8}	1.4×10^{-13}	6.1×10^{-7}
$1000 \times 10^5 \times 1$	LPP (MATLAB)	2660.8	8	2.1×10^{-7}	1.4×10^{-12}	7.1×10^{-7}

Theorem 8. For all $\gamma > 0$ and $\hat{w} = w_* \in W_*$ an exact solution to primal problem (P) is given by $x_* = y(\gamma)/\gamma$, where $y(\gamma)$ is a solution to the unconstrained problem (15).

To solve the primal and dual LP problems simultaneously, one can use the following iterative process similarly to the process (13,14):

$$y_{s+1} \in \arg \max_{y \in R^n} \left\{ -c^\top y + u_s^\top A y - \frac{1}{2} \|\gamma b - A y\|^2 - \frac{1}{2} \|(v_s - y)_+\|^2 \right\}, \quad (16)$$

$$u_{s+1} = u_s + \gamma b - A y_{s+1}, \quad (17)$$

$$v_{s+1} = (v_s - y_{s+1})_+. \quad (18)$$

Theorem 9. For all $\gamma > 0$ and for arbitrary starting point w_0 , the iterative process (16–18) converges to $w_* \in W_*$ in a finite number of iterations σ . The formula $x_* = y_{\sigma+1}/\gamma$ gives an exact solution to the primal problem (P).

The proofs of Theorems 7–9 are similar to the proofs of Theorems 1–3, respectively. The goal function $S^3(y, \gamma, \hat{w})$ of unconstrained maximization problem (15) is piecewise quadratic concave function. Therefore, one can define its generalized Hessian, which is $m \times m$ symmetric negative semidefinite matrix:

$$\frac{\partial^2}{\partial y^2} S^3(y, \gamma, \hat{w}) = -A^\top A - D(z).$$

Here, $D(z)$ denotes the $n \times n$ diagonal matrix with diagonal elements z^i , $i = 1, 2, \dots, n$. If $(\hat{v} - y)^i > 0$ then $z^i = 1$, if $(\hat{v} - y)^i \leq 0$ then $z^i = 0$. Now, for solving problem (15) one can use generalized Newton method.

There is an important difference between problems (10) and (15). In the first case, we look for the projection of a given point $\hat{u}$ on U_* , and in the second case we project a point $\hat{W} = [\hat{u}, \hat{v}]$ on the solution set W_* . Let $\hat{u}_*^1$ and $\hat{u}_*^2$ denote the projections of point $\hat{u}$ and $[\hat{u}, \hat{v}]$ in the first and second cases, respectively. Then the following inequality holds:

$$\|\hat{u}_*^1 - \hat{u}\| \leq \|\hat{u}_*^2 - \hat{u}\|.$$

The comparison of LP projection methods (which were implemented in MATLAB) with some well-known commercial and research software packages showed that they are competitive with the simplex and the interior point methods [1,5].

Table 1 presents the results of the test computations obtained using the program LPP, which implements method (11,12) in MATLAB, and other commercial and research packages [5]. Four randomly generated LP problems were solved on a 2.0 GHz Celeron computer with 1 Gb of memory. The following packages were used: BPMPD v. 2.3 (the interior point method) [6], MOSEK v.2.0 (the interior point method) [7], and the popular commercial package CPLEX (v.6.0.1, the interior point and simplex methods).

Table 1 shows the dimensions m and n of the problems, the density ρ of the nonzero

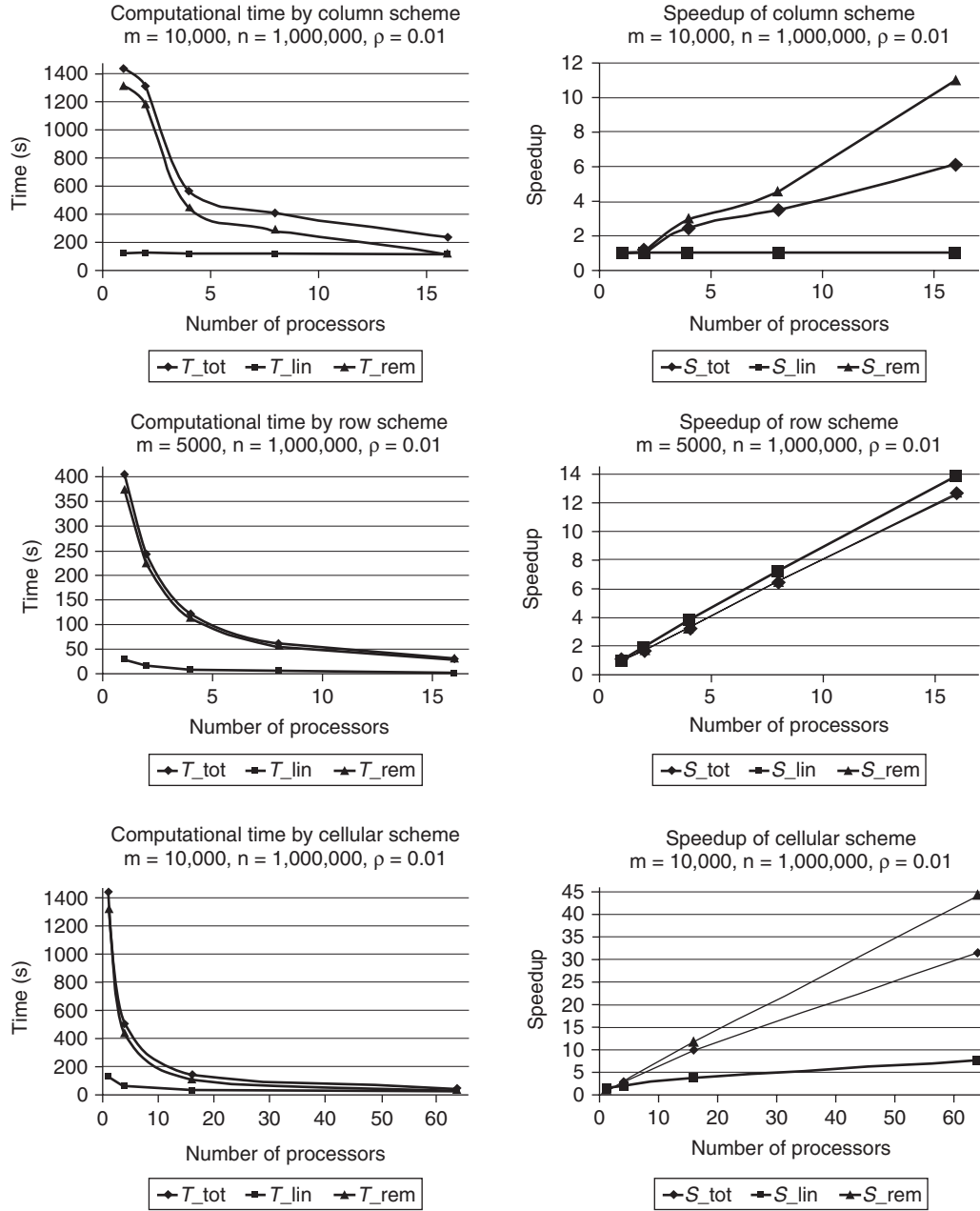


Figure 1. Computational time and speedup diagrams.

entries in the matrix A , the time T needed to solve the LP in seconds, and the number of iterations (for LP projection method (LPP), the total number of systems of linear equations that were solved in the Newton method when problem (11) was solved). Everywhere, parameter β was equal to 1, which exceeded β_* in these problems. Everywhere, $\hat{x} = 0_n$; that is, the normal solution (projection of the origin onto the primal solution set X_*) was sought in primal LP problem.

The l_∞ norms of the residual vectors were calculated:

$$\begin{aligned}\Delta_1 &= \|Ax - b\|_\infty \leq tol, \\ \Delta_2 &= \|(A^\top u - c)_+\|_\infty \leq tol, \\ \Delta_3 &= |c^\top x - b^\top u| \leq tol.\end{aligned}$$

It should be stressed that neither the linear program with 5 million variables, $\rho = 0.01$, $m = 1000$, nor those with 10^5 variables, $\rho = 1$, $m = 1000$, could be successfully solved by any of the codes listed above excepting the proposed LPP, which solved these problems in ≈ 16 and ≈ 44 min, respectively, and gave very high accuracy (less than 7.1×10^{-7}). Nevertheless, LPP code did not give the best results for small and moderate-size problems. As is evident from Table 1, BPMPD solver was superior to the others as applied for the first test problem; MOSEK solver showed the best results in the second test problem. Thus, the examples presented above have clearly demonstrated the principal effectiveness of the method proposed in solving the problems with a very large number ($\approx 10^6$) of nonnegative variables and a moderate ($\approx 10^3$) number of equality type constraints.

It is also evident that the results of LP projection algorithm are highly competitive to those of simplex and interior point methods, at the same time outperforming them in solving large LP problems.

Several parallel versions of the generalized Newton method for solving linear programs based on various data decomposition schemes of matrix A (column, row, and cellular schemes) were implemented [8]. The resulting parallel algorithms were successfully used to solve large-scale LP problems (up to several dozens of millions of variables

and several hundreds of thousands of constraints) for a relatively dense matrix A . The computational experiments were performed on the cluster consisting of two-processor nodes based on 1.6 GHz Intel Itanium 2 processors connected by Myrinet 2000. For example, for LP problem with 1 million variables and 10,000 constraints, the cellular scheme for 144 processors of the cluster accelerated the computations approximately by a factor of 50, and the computation time was 28 s. LP problem with 2 million variables and 200,000 constraints was solved in about 40 min on 80 processors. Another LP problem with 60 million variables and 4000 constraints was solved by column scheme in 140 s on 128 processors.

The computation time for typical test problems and the speedup for the different data decomposition schemes are shown in Fig. 1; where T_{tot} denotes the computation time in seconds, T_{lin} is the time spent on solving systems of linear equations, T_{rem} is the time spent on the other computations. Furthermore, S_{tot} denotes the parallel speedup in solving LP problem, S_{lin} is the parallel speedup in solving linear systems, and S_{rem} is the speedup of the other computations. Parameter β was equal to 100, which exceeded β_* in these problems. Everywhere, $\hat{x} = 0_n$; that is, the normal solution was sought in primal LP problem.

The development of an efficient solver was found to be a very challenging task; for every parallel variant of the algorithm, a reasonable trade-off between the computational scalability and the scalability in terms of memory had to be found. The highest speedup was obtained for the cellular scheme. However, depending on the dimension of the problem and the sparsity structure of the matrix A , the cellular, the row, or the column scheme was optimal.

REFERENCES

1. Mangasarian OL. A Newton method for linear programming. *J Optim Theory Appl* 2004;121(1):1–18.
2. Kanzow C, Qi H, Qi L. On the minimum norm solution of linear programs. *J Optim Theory Appl* 2003;116(2):333–345.

3. Golikov AI, Evtushenko YG. Solution method for large-scale linear programming problems. Dokl Math 2004;70(1):615–619.
4. Golikov AI, Evtushenko YG. Finding the projection of a given point on the set of solutions of a linear programming problems. Proc Steklov Inst Math 2008;2(2 Suppl):S68–S83.
5. Evtushenko YG, Golikov AI, Mollaverdi N. Augmented Lagrangian method for large-scale linear programming problems. Optim Method Softw 2005;7(4–5):515–524.
6. Andersen ED, Andersen KD. The MOSEK interior point optimizer for linear programming: an implementation of homogeneous algorithm. In: Frenk H, Roos K, Terlaky T, *et al.*, editors. High performance optimization. New York: Kluwer Academic Publishers; 2000. pp. 197–232.
7. Mészáros C. The BPMPD interior point solver for convex quadratic programming problems. Optim Method Softw 1999;11(1–4):431–449.
8. Evtushenko YG, Garanzha VA, Golikov AI, *et al.* Parallel implementation of generalized newton method for solving large-scale LP problems. In: Malyshkin V, editor. Parallel computing technologies. Lecture notes in computer science 5698. Berlin, Heidelberg; New York: Springer; 2009. pp. 84–97.

LIPSCHITZ GLOBAL OPTIMIZATION

YAROSLAV D. SERGEYEV
DMITRI E. KVASOV
Dipartimento di Elettronica,
Informatica e Sistemistica,
University of Calabria,
Rende, Italy
Software Department, N.I.
Lobachevsky State
University, Nizhni
Novgorod, Russia

PROBLEM STATEMENT

In various fields of human activity (engineering design, economic models, biology studies, etc.) different decision-making problems arise which can be often modeled by applying optimization concepts and tools. The decision-maker (engineer, physicist, chemist, economist, etc.) frequently wants to find the best combination of the set of parameters (geometrical sizes, electrical and strength characteristics, etc.) describing a particular optimization model which provides the optimum (minimum or maximum) of a suitable objective function, while it satisfies a certain collection of feasibility constraints. The objective function expresses overall system performance, such as profit, utility, risk, error, and so on. The constraints originate from physical, technical, economic, or some other considerations (see Refs 1–12, and the references therein).

The sophistication of the mathematical models describing the objects to be designed, which is the natural consequence of the growing complexity of these objects, significantly complicates the search for the best objective performance parameters. Many practical applications (see Refs 1–9, 11, 13–21, and the references therein) are characterized by optimization problems with several local solutions (typically, their number is unknown

and can be very high). These more complicated problems are usually referred to as *multiextremal* problems [8,10,11,17,22–25].

Nonlinear local optimization methods and theory (described in a vast literature, e.g., Refs 26–37, and the references therein) are not very successful for solving these multiextremal problems. As indicated, for example, in Refs 11 and 23, their limitation is due to the intrinsic multiextremality of the problem formulation and not due to the lack of smoothness or continuity—such properties are often present. It should be also mentioned in this context that in some problems a local solution may just be inappropriate. For example, if the objective function is an indicator of reliability and one is interested in assessing the worst case, then it is necessary to find the best (in this case, the worst) *global* solution [10,38].

Moreover, the objective function and constraints are often *black-box functions*. This means that there are “black-boxes” associated with each of these functions that, with the values of the parameters given in input, return the corresponding values of the functions and nothing is known to the optimizer about the way these values are obtained. Such functions are met in many engineering applications in which observation of the produced function values can be made, but analytical expressions of the functions are not available. For example, the values of the objective function and constraints can be obtained by running some computationally expensive numerical model, by performing a set of experiments, and so on. One may refer, for instance, to various decision-making problems in automatic control and robotics, structural optimization, safety verification problems, engineering design, network and transportation problems, mechanical design, chemistry and molecular biology, economics and finance or data classification (see Refs 1, 4–6, 8–12, 39–50, and the references therein).

To obtain estimates of the global solution related to a finite number of functions

evaluations, some suppositions on the structure of the objective function and constraints (such as continuity, differentiability, convexity, linearity, and so on) should be indicated. These assumptions play a crucial role in the construction of any efficient global search algorithm, able to outperform the simple uniform grid techniques in solving multiextremal problems. In fact, as observed for example in Refs 51–54, if no particular assumptions are made on the objective function and constraints, any finite number of function evaluations cannot guarantee getting close to the global minimum value, since this function may have very high and narrow peaks.

One of the natural and powerful (from both, the theoretical and the applied points of view) assumptions on the global optimization problem is that the objective function and constraints have bounded slopes. In other words, any limited change in the object parameters yields some limited changes in the characteristics of the objective performance. This assumption can be justified by the fact that in technical systems the energy of change is always limited (see the related discussion in Refs 11 and 55). One of the most popular mathematical formulations of this property is the Lipschitz continuity condition, which assumes that the difference (in the sense of a chosen norm) of any two function values is majorized by the difference of the corresponding function arguments, multiplied by a positive factor L . In this case, the function is said to be *Lipschitz* and the corresponding factor L is said to be the *Lipschitz constant*. The problem involving Lipschitz functions (the objective function and constraints) is said to be the *Lipschitz global optimization problem*.

Formally, a general Lipschitz global optimization problem can be stated as follows:

$$f^* = f(x^*) = \min f(x), \quad x \in \Omega \subset \mathbb{R}^N, \quad (1)$$

where Ω is a bounded set defined by the following formulae

$$\Omega = \{x \in D : g_i(x) \leq 0, 1 \leq i \leq m\}, \quad (2)$$

$$D = [a, b] = \left\{x \in \mathbb{R}^N : a(j) \leq x(j) \leq b(j), \right. \\ \left. 1 \leq j \leq N \right\}, \quad a, b \in \mathbb{R}^N, \quad (3)$$

with N being the problem dimension. In Equations (1)–(3), the objective function $f(x)$ and the constraints $g_i(x)$, $1 \leq i \leq m$, are multiextremal, non necessarily differentiable, black-box (and, thus, hard to evaluate) functions that satisfy the Lipschitz condition over the search hyperinterval D :

$$|f(x') - f(x'')| \leq L \|x' - x''\|, \quad x', x'' \in D, \quad (4)$$

$$|g_i(x') - g_i(x'')| \leq L_i \|x' - x''\|, \quad x', x'' \in D, \\ 1 \leq i \leq m, \quad (5)$$

where $\|\cdot\|$ denotes, usually, the Euclidean norm, L and L_i , $1 \leq i \leq m$, are the (unknown) Lipschitz constants such that $0 < L, L_i < \infty$, $1 \leq i \leq m$.

Note also that, due to the multiextremal character of the constraints, the admissible region Ω can consist of disjoint, non-convex subregions. More generally, not all the constraints (and, consequently, the objective function too) are defined over the whole region D : namely, a constraint $g_{i+1}(x)$ (or the objective function $f(x)$) can be defined only over subregions where $g_i(x) \leq 0$, $1 \leq i \leq m$. In this case, the problems (1)–(5) are said to be *with partially defined constraints*. If $m = 0$ in Equation (2), the problem is said to be *box-constrained*.

Such problems are encountered very frequently in practice. Let us refer only to the following examples: general (Lipschitz) nonlinear approximation; solution of nonlinear equations and inequalities; calibration of complex nonlinear system models; black-box systems optimization; optimization of complex hierarchical systems (related, for example, to facility location, mass-service systems), and so on (see Refs 8–11, 17, 25, 56, 57, and the references given therein). Thus, numerous algorithms have been proposed (see, e.g., Refs 8, 11, 12, 17, 23, 52, 55, 56, 58–75, and the references therein) for solving the problems (1)–(5).

Sometimes, the problems (1), (3), and (4) with a differentiable objective function having the Lipschitz (with an unknown Lipschitz

constant) first derivative $f'(x)$ (which could itself be a costly multiextremal black-box function) is included in the same class of Lipschitz global optimization problems (theory and methods for solving this differentiable problem can be found in Refs 8, 10, 11, 70, 76–86).

It should be stressed that not only multidimensional global optimization problems but also one-dimensional ones (in contrast to one-dimensional local optimization problems that have been very well studied in the past) continue to attract attention of many researchers. These happens at least for two reasons. First, there exists a large number of applications where it is necessary to solve such problems [10,11,87–93]. Second, there exist numerous approaches [8,10–12,17,23,55,81,94–96] that enable to generalize to the multidimensional case the methods propose for solving univariate problems.

It is also noteworthy that it is not easy to find suitable ways for managing multiextremal constraints (Eq. 2). For example, the traditional penalty approach (see Refs 17, 23, 26, 36, 97, and the references therein) requires that $f(x)$ and $g_i(x)$, $1 \leq i \leq m$, are defined over the whole search domain D . At first glance it seems that at the regions where a constraint is violated the objective function can be simply filled in with either a big number or the function value at the nearest feasible point [17,23,98]. Unfortunately, in the context of Lipschitz algorithms, incorporating such ideas can lead to extremely high Lipschitz constants, causing degeneration of the methods and nonapplicability of the penalty approach (see the related discussion in Refs 8, 10, 11, 25, 56).

A promising approach called the *index scheme* has been proposed in Refs 99 and 100 (see also Refs 11, 101, 102) to reduce the general constrained problems (1)–(5) to a box-constrained discontinuous problem. An important advantage of the index scheme is that it does not introduce additional variables and/or parameters by opposition to widely-used traditional approaches. It has been recently shown in Ref. 103 that the index scheme can be also successfully used in combination with the

branch-and-bound approach [8,10,23,104] if the Lipschitz constants from Equations (4)–(5) are known *a priori*. An extension of the index approach, originally proposed for information-stochastic Bayesian algorithms, to the framework of geometric algorithms, constructing nondifferentiable discontinuous minorants for the reduced problem, has been also examined [92,105–107].

Therefore, in order to expose the principal ideas of some known approaches to solving the stated problem, the box-constrained Lipschitz global optimization problem (Eqs 1, 3, and 4) will be considered in the next two sections. In particular, these approaches will be distinguished by the mode in which information about the Lipschitz constants is obtained and by the strategy of exploration of the search hyperinterval D from Equation (3).

USAGE OF THE LIPSCHITZ CONDITION

The Lipschitz continuity assumption, being quite realistic for many practical problems (see Refs 8–11, 17, 23, 25, and the references given therein), is also an effective tool for constructing global optimum estimates with a finite number of function evaluations, which allows one to construct global optimization algorithms and to prove their convergence and stability.

In fact, the Lipschitz condition allows us to give its simple geometric interpretation (Fig. 1). Once a one-dimensional Lipschitz (with the Lipschitz constant L) function $f(x)$ has been evaluated at two points x' and x'' , due to Equation (4), the following four inequalities take place over the interval $[x', x'']$

$$\begin{aligned} f(x) &\leq f(x') + L(x - x'), \\ f(x) &\geq f(x') - L(x - x'), x \geq x', \\ f(x) &\leq f(x'') - L(x - x''), \\ f(x) &\geq f(x'') + L(x - x''), x \leq x''. \end{aligned}$$

Thus, the function graph over the interval $[x', x'']$ must lie inside the area limited by four lines which pass through the points $(x', f(x'))$ and $(x'', f(x''))$ with the slopes $\pm L$ (see the light grey parallelogram in Fig. 1).

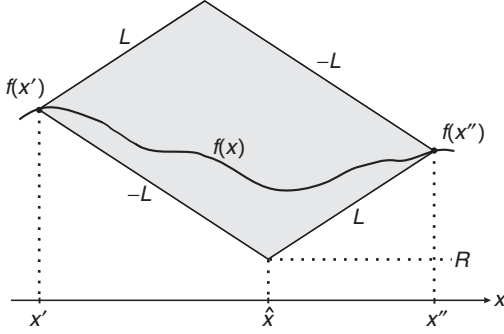


Figure 1. Graphical interpretation of the Lipschitz condition.

If the function $f(x)$ is to be minimized over the interval $[x', x'']$, its lower bound over this interval can be easily calculated. In fact, due to the Lipschitz condition (4), the following inequality is satisfied

$$f(x) \geq F(x),$$

where $F(x)$ is a piecewise linear function over $[x', x'']$

$$F(x) = \max\{f(x') - L(x - x'), f(x'') + L(x - x'')\}, \\ x \in [x', x''].$$

The minimal value of $F(x)$ over $[x', x'']$ is therefore the lower bound of $f(x)$ over this interval. It is calculated as

$$R = R_{[x', x'']} = F(\hat{x}) = \frac{f(x') + f(x'')}{2} - L \frac{x'' - x'}{2} \quad (6)$$

and is obtained at the point

$$\hat{x} = \frac{x' + x''}{2} - \frac{f(x'') - f(x')}{2L} \quad (7)$$

(Fig. 1).

The described geometric interpretation lies in the basis of the Piyavskij–Shubert method (see Refs 68 and 71)—one of the first and well studied algorithms for the one-dimensional Lipschitz global optimization, discussed in a vast literature; e.g., the survey [52] and the references given therein; various modifications of this algorithm have been also proposed in Refs 23, 52, 58, 60, 108–110). As observed, for example in Refs

111–114, this algorithm is “optimal in one step,” that is, the choice of the current evaluation point as in Equation (7) ensures maximum increase of the lower bound of the global minimum value of $f(x)$ with respect to a number of reasonable criteria (see Refs 11, 23, 25, 115 for the related discussions on optimality principles in Lipschitz global optimization).

The Piyavskij–Shubert method belongs to the class of *geometric algorithms*; that is, algorithms that use in their work auxiliary functions to estimate behavior of $f(x)$ over the search region, as just described. This idea has proved to be very fruitful and many algorithms (both one-dimensional and multi-dimensional ones; see Refs 52, 58, 61, 67, 74, 76, 110, 116–118, and the references given therein) based on constructing, updating, and improving auxiliary piecewise minorant functions built using an overestimate of the Lipschitz constant L have been proposed. Close geometric ideas have been used for other classes of optimization problems, as well. For example, various aspects of the so-called α BB-algorithm for solving constrained nonconvex optimization problems with twice-differentiable functions have been examined in Refs 45, 119, 120 (see also Refs 15 and 121). The α BB-algorithm combines a convex lower bounding procedure (based on a parameter α that can be interpreted in a way similar to the Lipschitz constant) within the branch-and-bound framework, whereas nonconvex terms are underestimated through quadratic functions derived from second-order information.

As shown, in Refs 10, 11, 122, there exists a strong relationship between the geometric approach and another possible technique of solving the stated problem—the so-called *information-statistical approach*. This approach originated in Refs 72 and 123 (see also Refs 11 and 55) and, together with Piyavskij–Shubert method, it has consolidated foundations of the Lipschitz global optimization. The main idea of this approach is to apply the theory of random functions to building a mathematical representation of available (certain or uncertain) *a priori* information on the objective function behavior. A systematic approach to the description of some uncertain information on this behavior is to accept that the unknown black-box function to be minimized is a sample of some known random function. Generally, to provide an efficient analytical technique for deriving some estimates of global optimum with a finite number of trials, that is for obtaining by Bayesian reasoning some conditional (with respect to the trials performed) estimations, the random function should have some special structure. Thus, it is possible to deduce the decision rules for performing new trials as some optimal decision functions [11,12,25,42,55,75,124–131].

One of the main questions to be considered in solving the Lipschitz global optimization problem is, how can the Lipschitz constants L and L_i , $1 \leq i \leq m$ be specified? For clarity, let us speak only about the Lipschitz constant L of the objective function $f(x)$ from Equation (1). There are several approaches to specify the Lipschitz constant. For example, it can be given *a priori* [17,23,52,58,66–68,71,74,132,133]. This case is very important from the theoretical viewpoint but is not easily obtained in practice. More practical approaches are based on an adaptive estimation of L in the course of the search. In such a way, algorithms can use either a *global* estimate of the Lipschitz constant [8,11,23,55,65,72,134] valid for the whole region D from Equation (3), or *local* estimates L_i valid only for some subregions $D_i \subseteq D$ [10,11,69,122,135,136]. Estimating local Lipschitz constants during the work of a global optimization algorithm allows

one to significantly accelerate the global search (the importance of the estimation of local Lipschitz constants has also been highlighted in other works; [8,76,137]). Naturally, balancing between local and global information must be performed in an appropriate way [10,11,69] to avoid missing the global solution. Finally, estimates of L can be chosen during the search from a certain set of admissible values from zero to infinity [10,64,98,138–141]. It should be stressed that either the Lipschitz constant is known and an algorithm is constructed correspondingly, or it is not known but there exist sufficiently large values of parameters of the considered algorithm, ensuring its convergence.

Note that the Lipschitz constant has also a significant influence on the convergence speed of numerical algorithms and, therefore, the problem of its specifying is of the great importance for the Lipschitz global optimization [8,10,11,25,52,57,64,69,136,137,142]. An underestimate of the Lipschitz constant L can lead to the loss of the global solution. In contrast, accepting a very high value of L for a concrete objective function means (due to the Lipschitz condition (4)) assuming that the function has a complicated structure with sharp peaks and narrow attraction regions of minimizers within the whole admissible region. Thus, an overestimate of L (if it does not correspond to the real behavior of the objective function) leads to a slow convergence of the algorithm to the global minimizer.

MULTIDIMENSIONAL APPROACHES

Since the uniform grid techniques require a substantial number of evaluation points, they can be successfully applied only if the dimensionality N from Equation (3) is small or the required accuracy is not high. As observed in Refs 8, 11, 17, 143, an adaptive selection of evaluation points $x^1, x^2, \dots, x^k$ usually outperforms the grid technique. Such a selection aims to ensure a higher density of the evaluation points near the global minimizer x^* with respect to the densities in the rest of the search domain D .

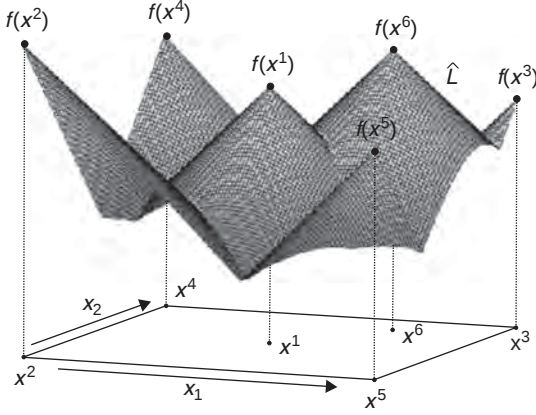


Figure 2. An example of a two-dimensional lower bounding function for an objective function $f(x)$, $x = (x_1, x_2) \in D \subset \mathbb{R}^2$.

One of these adaptive approaches is a direct extension of the one-dimensional geometric algorithms to the multidimensional case. At each iteration $k \geq 1$ of such an algorithm, a function

$$F_k(x) = \max_{1 \leq i \leq k} \{f(x^i) - \hat{L} \|x - x^i\|\}, \quad \hat{L} > L, \quad x \in D, \quad (8)$$

is constructed which is a lower bounding function (minorant) for $f(x)$ (due to the Lipschitz condition (4)), that is,

$$f(x) \geq F_k(x), \quad \forall k \geq 1, \quad x \in D.$$

A global minimizer of the function (8) is determined and chosen as a new evaluation point for $f(x)$, that is,

$$x^{k+1} = \arg \min_{x \in D} F_k(x). \quad (9)$$

A new minorant function (8) is constructed and the process is repeated, until a stopping criterion is satisfied. For example, the algorithm can be stopped when the difference between the current estimate of the minimal function value and the global minimum of $F_k(x)$, which is a lower bound on the global minimum of $f(x)$, becomes smaller than some prescribed value.

Geometrically, the graph of $F_k(x)$ in $\mathbb{R}^{N+1}$ is defined by a set of k intersecting N -dimensional cones bounded by the hyper-interval $D \subset \mathbb{R}^N$, with vertices $(x^i, f(x^i))$,

$1 \leq i \leq k$, and the slope $\hat{L}$. Axes of symmetry of these cones are parallel to the $(N+1)$ -th unit vector and their sections orthogonal to the axes are spheres (Fig. 2).

In the case of one dimension ($N = 1$), there are no serious problems in calculating x^{k+1} from Equation (9). In fact, this is reduced to selecting the minimal value among k values, each of which is easy to compute due to the explicit formula (6). If $N > 1$, then the relations between the location of the next evaluation point and the results of the already performed evaluations become significantly more complicated. Therefore, finding the point x^{k+1} is usually the most time-consuming operation of the multidimensional algorithm, and its complexity increases with the growth of the problem dimension.

In Refs 52 and 68, it is observed that the global minimizer x^{k+1} of the minorant function $F_k(x)$ belongs to a finite set of points; that is, local minima of $F_k(x)$, which can be computed by solving systems of quadratic equations. In Ref. 67, a deep analysis of the algorithm constructing lower bounding functions $F_k(x)$ from Equation (8) is presented and a more efficient way to find their local minima is proposed, which requires solving systems of N linear and one quadratic equations. Algebraic methods aiming to avoid computing solutions of some systems which do not correspond to global minima of $F_k(x)$ are considered in Ref. 144. In Ref. 52,

it is indicated that the number of systems to be examined can be further reduced.

In Refs 74 and 145, cones with spherical base (Eq. 8) are replaced by cones with simplex base. At each iteration, there exists a collection of such inner approximating simplices in $\mathbb{R}^{N+1}$ that bound parts of the graph of the objective function $f(x)$ from below, always including the part containing the global minimum point (x^*, f^*) from Equation (1). Various aspects of generalization of this technique are discussed in Refs 76 and 118.

In Ref. 146, minorant functions similar to Equation (8) are considered. The solution of Equation (9) still remains difficult, the problem of the growth of complexity of Equation (9) is analyzed, and a strategy for dropping elements from the sequence of evaluation points while constructing a minorant function is proposed.

Unfortunately, the difficulty of determining a new evaluation point (Eq. 9) in high dimensions (after executing even relatively small number of function evaluations) has not been yet overcome. It should be noted here that, as observed in Ref. 11, this difficulty is usual for many approaches aiming to decrease the total number of evaluations by adaptively constructing an auxiliary function approximating $f(x)$ over the search domain D . They introduce some selection rule of the type (9) and the auxiliary function $F_k(x)$ used in such a rule often happens to be essentially multiextremal (though its construction may be connected with assumptions quite different from the Lipschitz condition (4)). Moreover, in this case, it is difficult to use adaptive estimates of the global or local Lipschitz constant(s). This can pose a further limitation for the practical application of the considered methods directly extended the one-dimensional Piyavskij–Shubert algorithm to the multidimensional case.

In order to simplify the task of calculating lower bounds of $f(x)$ and determining new evaluation points, many multidimensional methods subdivide the search domain $D = [a, b]$ from Equation (3) into a set of subdomains D_i , obtaining at each iteration k a

partition $\{D^k\}$ of D such that

$$D = \bigcup_{i=1}^M D_i, \quad D_i \cap D_j = \delta(D_i) \cap \delta(D_j), \quad i \neq j, \quad (10)$$

where $M = M(k)$ is the number of subdomains D_i during the k th iteration and $\delta(D_i)$ denotes the boundary of D_i . An approximation of $f(x)$ over each subregion D_i from $\{D^k\}$ is based on the results obtained by evaluating $f(x)$ at some points $x \in D$.

If a partition of D into hyperintervals D_i is used in Equation (10) and the objective function is evaluated, for example, at the central point of every hyperinterval D_i (the so-called *center-sampling partition strategies*—see for example, Refs 61, 64, 66, 98, 137, 139, 147), each cone from Equation (8) can be considered over the corresponding hyperinterval D_i , independently of the other cones. This allows one to avoid the necessity of establishing the intersections of the cones and to simplify the lower bound estimation of $f(x)$ when a new evaluation point is determined by a rule such as Equation (9). The more accurate estimation of the objective function behavior can be achieved when two evaluation points over a hyperinterval are used for constructing an auxiliary function for $f(x)$, for example, at the vertices corresponding to the main diagonal of hyperintervals (the so-called *diagonal partition strategies* [8,10,23,65,83,86,122,141,148]).

In exploring the multidimensional search domain D , various multisection techniques for partitions of hyperintervals have been also studied in the framework of interval analysis [149–154]. More complex partitions based on simplices [22,23,155–158] and auxiliary functions of various nature have also been introduced [76,78,145,159–165]. Several attempts to generalize various partition schemes in a unique theoretical framework have been made [8,23,63,104,166–168].

Another fruitful approach to solving multidimensional Lipschitz global optimization problems consists of extending some efficient one-dimensional algorithms to the multidimensional case. There exist at least three extensions of these algorithms to the multidimensional case: *nested global optimization*

scheme [68,169], *reduction of the dimension by using Peano curves* [11,55], and *diagonal approach* [8,10], which is a special case of the scheme of adaptive partition algorithms (or a more general scheme of the “*divide the best*” algorithms [10,104]).

In the nested optimization scheme [68,169] (see also Refs 11, 22, 55, 95), the minimization over D of the Lipschitz function $f(x)$ from (1) is made using the following nested form

$$\min_{x \in D} f(x) = \min_{a(1) \leq x(1) \leq b(1)} \dots \min_{a(N) \leq x(N) \leq b(N)} f(x(1), \dots, x(N)).$$

This allows one to solve the multidimensional problem (1),(3),(4) by solving a set of univariate problems and, thus, to apply the methods proposed for the one-dimensional optimization.

The second dimensionality reduction approach [11,38,55,171,172] is based on the application of a single-valued Peano-type curve $x(\tau)$ continuously mapping the unit interval $[0, 1]$, $\tau \in [0, 1]$, onto the hyperinterval D . These curves, introduced by Peano and further by Hilbert, pass through any

point of D ; that is, “fill” the hypercube D , and this gave rise to the term *space-filling curves* [173,174] or just *Peano curves*. Particular schemes for the construction of various approximations to Peano curves, their theoretical justification, and the standard routines implementing these algorithms are given [11,55,143,175].

If $x(\tau)$ is Peano curve, then from the continuity (due to Equation (4)) of $f(x)$ it follows that

$$\min_{x \in D} f(x) = \min_{\tau \in [0, 1]} f(x(\tau)), \quad (11)$$

and this reduces the original multidimensional problem (Eqs 1,3,4) to one dimension (Fig. 3). Moreover, it can be proved [11] that if the multidimensional function $f(x)$ from Equation (1) is Lipschitz with a constant L , then the one-dimensional function $f(x(\tau))$, $\tau \in [0, 1]$, satisfies the uniform Hölder condition over $[0, 1]$ with the exponent N^{-1} and the coefficient $2L\sqrt{N+3}$

$$|f(x(\tau')) - f(x(\tau''))| \leq 2L\sqrt{N+3}(|\tau' - \tau''|)^{1/N}, \quad \tau', \tau'' \in [0, 1]. \quad (12)$$

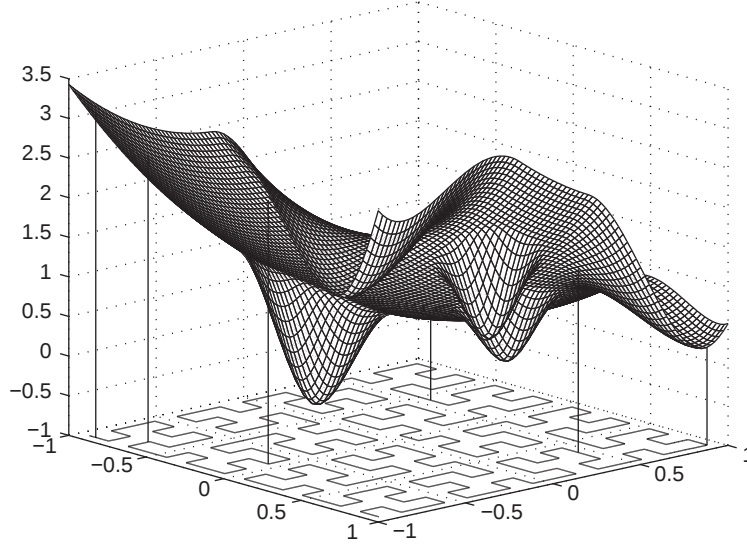


Figure 3. Usage of an approximation of the Peano curve for global optimization: A two-dimensional function from the class [170] has been evaluated at six points belonging to the curve.

Thus, the Lipschitz assumption does not hold for the function $f(x(\tau))$ at the right-hand side of Equation (11), which satisfy the more general Hölder condition (12). However, as shown, [11,55,176–179], many Lipschitz one-dimensional algorithms can be generalized to the case of minimization of Hölder functions too. In the monograph [11], many powerful sequential and parallel global optimization methods based on the usage of Peano curves are proposed and analyzed in depth.

Finally, in the diagonal approach [8,10,148,180] to solving Lipschitz global optimization problems, the initial hyperinterval D from Equation (3) is partitioned into a set of smaller hyperintervals D_i , the objective function is evaluated only at two vertices corresponding to the main diagonal of hyperintervals of the current partition of D , and the results of these evaluations are used to select a hyperinterval for the further subdivision. The diagonal approach has a number of attractive theoretical properties and has proved to be efficient in solving applied problems (see Refs 8, 10, 141, and the references therein).

The extension of one-dimensional global optimization algorithms to the multidimensional case can be performed naturally by means of the diagonal scheme [8,10,65,135,141,148,180,181]. In order to decrease the computational efforts needed to describe the objective function $f(x)$ at every small hyperinterval $D_i \subset D$, $f(x)$ is evaluated only at the vertices a_i and b_i corresponding to the main diagonal of D_i . At every iteration the “merit” of each hyperinterval so far generated is estimated. The higher “merit” of a hyperinterval D_i corresponds to the higher possibility that the global minimizer x^* of $f(x)$ belongs to D_i . The “merit” is measured by a real-valued function R_i called a *characteristic* [10,11,22]. In order to calculate the characteristic R_i of a multidimensional hyperinterval D_i , some one-dimensional characteristics can be used as prototypes. After an appropriate transformation they can be applied to the one-dimensional segment being the main diagonal $[a_i, b_i]$ of the hyperinterval D_i . A hyperinterval having the “best” characteristic (e.g., the biggest one, as in the information approach [11,55,174];

or the smallest one, as in the geometric approach, [10,11,136]) is partitioned by means of a partition operator (*diagonal partition strategy*), and new evaluations are performed at two vertices corresponding to the main diagonal of each generated hyperinterval. The concrete choice of the characteristic R_i and the partition strategy determines the particular diagonal method [8,10,23,64,122,135,139,141,148], and the references therein).

Different exploration techniques based on various diagonal adaptive partition strategies have been analyzed in Ref. 10 (see also Ref. 8). It has been demonstrated in Refs 10 and 148 that partition strategies traditionally used in the framework of the diagonal approach may not fulfill the requirements of computational efficiency because of the execution of many redundant evaluations of the objective function. Such a redundancy slows down significantly the global search in the case of costly functions.

A new diagonal adaptive partition strategy, originally proposed in Ref. 148 (see also Refs 10 and 147), that allows one to avoid this computational redundancy has been thoroughly described in Ref. 10. It can be also viewed as a procedure generating a series of curves similar to Peano space-filling curves—the so-called *adaptive diagonal curves*. However, the order of partition of the adaptive diagonal curves is different within different subdomains of the search domain. A particular diagonal algorithm developed by using the new partition strategy constructs its own series of curves, taking into account properties of the problem to be solved. The curves become denser in the vicinity of the global minimizers of the objective function if the selection of hyperintervals for partitioning is realized appropriately (Fig. (4)). Thus, these curves can be used for reducing the problem dimensionality similarly to Peano curves but in a more efficient way.

In Ref. 10, it is demonstrated theoretically and numerically that the new strategy considerably speeds up the search and also leads to saving computer memory. It is particularly important that its advantages become more pronounced when the problem dimension N increases. Thus, fast one-dimensional

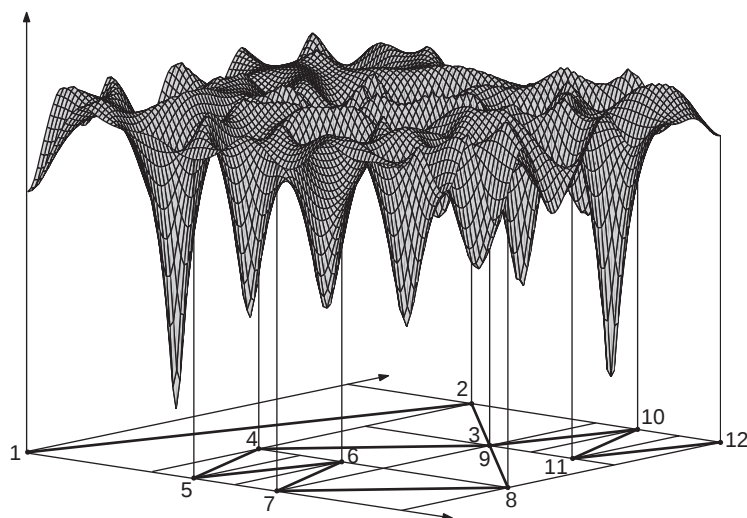


Figure 4. Usage of adaptive diagonal curves (evaluation points from 1 to 12), generated by the efficient diagonal partition strategy, for minimizing a multiextremal two-dimensional objective function from the class [182].

methods can be efficiently extended to the multidimensional case by using this scheme. Moreover, the proposed partition strategy can be subsequently parallelized by for example, the technique from Refs 11 and 183 (where theoretical results regarding non-redundant parallelism have been obtained, as well), allowing one to speed up further.

Finally, it should be emphasized that particular attention should be also paid to the problem of testing global optimization algorithms. The lack of such information as number of local optima, their locations, attraction regions, local and global values, describing global optimization tests taken from real-life applications, creates additional difficulties in verifying validity of the algorithms. Some benchmark classes and related discussions can be found in Refs 4, 10, 11, 17, 25, 59, 170, 184–191.

Acknowledgments

This work was supported by the following grants: Italian PRIN project 20079PLLN7-003, the grants 64729.2010.9 and MK-3473.2010.1 awarded by the President of the Russian Federation for supporting, respectively, the leading research

groups and young researchers, as well as by the Russian Federal Program “Scientists and Educators in Russia of Innovations,” contract number 02.740.11.5018.

REFERENCES

1. Achenie LEK, Gani R, Venkatasubramanian V, editors. Computer aided molecular design: theory and practice. Volume 12, Computer-Aided Chemical Engineering. Amsterdam: Elsevier Publishers; 2002.
2. Audet C, Hansen P, Savard G, editors. Essays and surveys in global optimization (GERAD 25th Anniversary Series). New York: Springer; 2005.
3. Batishchev DI. Search methods of optimal design (in Russian). Moscow: Sovetskoe Radio; 1975.
4. Floudas CA, Pardalos PM, Adjiman CS, *et al.* Handbook of test problems in local and global optimization. Dordrecht: Kluwer Academic Publishers; 1999.
5. Grossmann IE, editor. Global optimization in engineering design. Dordrecht: Kluwer Academic Publishers; 1996.
6. Mockus J. A set of examples of global and discrete optimization: applications of Bayesian heuristic approach. Dordrecht: Kluwer Academic Publishers; 2000.

7. Pardalos PM, Resende MGC, editors. Handbook of applied optimization. New York: Oxford University Press; 2002.
8. Pintér JD. Global optimization in action (Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications). Dordrecht: Kluwer Academic Publishers; 1996.
9. Pintér JD, editor. Global optimization: scientific and engineering case studies. Volume 85, Nonconvex optimization and its applications. Berlin: Springer; 2006.
10. Sergeyev YD, Kvasov DE. Diagonal global optimization methods (in Russian). Moscow: FizMatLit; 2008.
11. Strongin RG, Sergeyev YD. Global optimization with non-convex constraints: sequential and parallel algorithms. Dordrecht: Kluwer Academic Publishers; 2000.
12. Törn A, Žilinskas A. Global optimization. Volume 350, Lecture notes in computer science. Berlin: Springer; 1989.
13. Di Pillo G, Grippo L. An exact penalty function method with global convergence properties for nonlinear programming problems. Math Program 1986;36(1):1–18.
14. Famularo D, Pugliese P, Sergeyev YD. A global optimization technique for fixed-order control design. Int J Syst Sci 2004;35(7):425–434.
15. Floudas CA, Gounaris CE. A review of recent advances in global optimization. J Global Optim 2009;45(1):3–38.
16. Floudas CA, Pardalos PM, editors. Optimization in computational chemistry and molecular biology: local and global approaches. Dordrecht: Kluwer Academic Publishers; 2000.
17. Horst R, Pardalos PM, editors. Volume 1, Handbook of global optimization. Dordrecht: Kluwer Academic Publishers; 1995.
18. Mockus J. Multiextremal problems in engineering design (in Russian). Moscow: Nauka; 1967.
19. Pardalos PM, Romeijn HE, editors. Volume 2, Handbook of global optimization. Dordrecht: Kluwer Academic Publishers; 2002.
20. Strekalovsky AS. Elements of nonconvex optimization (in Russian). Novosibirsk: Nauka; 2003.
21. Tuy H. D.C. optimization and robust global optimization. J Glob Optim 2010. DOI: 10.1007/s10898-009-9475-2.
22. Gorodetsky SY, Grishagin VA. Nonlinear programming and multiextremal optimization. Volume 2, Models and methods of finite-dimensional optimization (in Russian). Nizhni Novgorod, Russia: NNGU Press; 2007.
23. Horst R, Tuy H. Global optimization—deterministic approaches. Berlin: Springer; 1996.
24. Pardalos PM, Rosen JB. Constrained global optimization: algorithms and applications. Volume 268, Springer lecture notes in computer science. New York: Springer; 1987.
25. Zhigljavsky AA, Žilinskas A. Stochastic global optimization. New York: Springer; 2008.
26. Bertsekas DP. Nonlinear programming. Belmont (MA): Athena Scientific; 1999.
27. Conn AR, Scheinberg K, Vicente LN. Introduction to derivative-free optimization. Philadelphia (PA): SIAM; 2009.
28. Clarke FH. Optimization and nonsmooth analysis. New York: John Wiley & Sons, Inc.; 1983. Reprinted by SIAM Publications; 1990.
29. Demyanov VF, Malozemov VN. Introduction to minimax. New York: John Wiley & Sons, Inc.; 1974. (The 2nd English–language edition. Dover Publications; 1990).
30. Demyanov VF, Rubinov AM. Quasidifferential calculus. New York: Optimization Software Inc., Publication Division; 1986.
31. Di Pillo G, Giannessi F, editors. Nonlinear optimization and applications. New York: Plenum Press; 1996.
32. Fletcher R. Practical methods of optimization. New York: John Wiley & Sons, Inc.; 2000.
33. Hiriart-Urruty J-B, Lemaréchal C. Convex analysis and minimization algorithms (Parts I and II). Berlin: Springer; 1996.
34. Kolda TG, Lewis RM, Torczon V. Optimization by direct search: new perspectives on some classical and modern methods. SIAM Rev 2003;45(3):385–482.
35. Mangasarian OL. Nonlinear programming. New York: McGraw–Hill; 1969. Reprinted by SIAM Publications; 1994.
36. Nocedal J, Wright SJ. Numerical optimization. Dordrecht: Springer; 1999.
37. Rockafellar RT. Convex analysis. Princeton (NJ): Princeton University Press; 1970.
38. Strongin RG, Sergeyev YD. Global optimization: fractal approach and non-redundant parallelism. J Glob Optim 2003;27(1):25–50.

39. Björkman M, Holmström K. Global optimization of costly nonconvex functions using radial basis functions. *Optim Eng* 2000;1(4):373–397.
40. Di Serafino D, Gomez S, Milano L, *et al.* A genetic algorithm for a global optimization problem arising in the detection of gravitational waves. *J Glob Optim* 2010;48(1): 41–55.
41. Gutmann H-M. A radial basis function method for global optimization. *J Glob Optim* 2001;19(3):201–227.
42. Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. *J Glob Optim* 1998;13(4): 455–492.
43. Kvasov DE, Menniti D, Pinnarelli A, *et al.* Tuning fuzzy power-system stabilizers in multi-machine systems by global optimization algorithms based on efficient domain partitions. *Int J Electr Power Syst Res* 2008;78(7):1217–1229.
44. Liuzzi G, Lucidi S, Piccialli V, *et al.* A magnetic resonance device designed via global optimization techniques. *Math Program* 2004;101(2):339–364.
45. Maranas CD, Floudas CA. Global minimum potential energy conformations of small molecules. *J Glob Optim* 1994;4(2):135–170.
46. Regis RG, Shoemaker CA. Constrained global optimization of expensive black-box functions using radial basis functions. *J Glob Optim* 2005;31(1):153–171.
47. Sherali HD, Ganesan V. A pseudo-global optimization approach with application to the design of containerships. *J Glob Optim* 2003;26(4):335–360.
48. Vasile M, Summerer L, De Pascale P. Design of Earth–Mars transfer trajectories using evolutionary–branching technique. *Acta Astronaut* 2005;56(8):705–720.
49. Vasile M, Locatelli M. A hybrid multiagent approach for global trajectory optimization. *J Glob Optim* 2009;44(4):461–479.
50. Villemonteix J, Vazquez E, Walter E. An informational approach to the global optimization of expensive-to-evaluate functions. *J Glob Optim* 2009;44(4):509–534.
51. Dixon LCW. Global optima without convexity. Technical Report. Hatfield, England: Numerical Optimization Centre, Hatfield Polytechnic; 1978.
52. Hansen P, Jaumard B. Lipschitz optimization. In: Horst R, Pardalos PM, editors. Volume 1, Handbook of global optimization. Dordrecht: Kluwer Academic Publishers; 1995. pp. 407–493.
53. Rinnooy Kan AHG, Timmer GT. Global optimization. In: Nemhauser GL, Rinnooy Kan AHG, Todd MJ, editors. Volume 1, Handbook of operations research: optimization. Amsterdam: North–Holland Publishing Co.; 1989. pp. 631–662.
54. Stephens CP, Baritomp W. Global optimization requires global information. *J Optim Theory Appl* 1998;96(3):575–588.
55. Strongin RG. Numerical methods in multiextremal problems (Information-Statistical Algorithms) (in Russian). Moscow: Nauka; 1978.
56. Floudas CA, Pardalos PM, editors. (Volumes 6), Encyclopedia of optimization. Dordrecht: Kluwer Academic Publishers; 2001. (The 2nd ed. Springer; 2009).
57. Horst R, Pardalos PM, Thoai NV. Introduction to global optimization. Dordrecht: Kluwer Academic Publishers; 1995. (The 2nd ed. Kluwer Academic Publishers; 2001).
58. Basso P. Iterative methods for the localization of the global maximum. *SIAM J Numer Anal* 1982;19(4):781–792.
59. Dixon LCW, Szegő GP, editors. Volumes 1 and 2, Towards global optimization. Amsterdam: North–Holland; 1975, 1978.
60. Evtushenko YG. Numerical optimization techniques, Translations series in mathematics and engineering. Berlin: Springer; 1985.
61. Galperin EA. The cubic algorithm. *J Math Anal Appl* 1985;112(2):635–640.
62. Gaviano M, Lera D. A global minimization algorithm for Lipschitz functions. *Optim Lett* 2008;2(1):1–13.
63. Horst R, Tuy H. On the convergence of global methods in multiextremal optimization. *J Optim Theory Appl* 1987;54(2):253–271.
64. Jones DR, Perttunen CD, Stuckman BE. Lipschitzian optimization without the Lipschitz constant. *J Optim Theory Appl* 1993;79(1):157–181.
65. Kvasov DE, Sergeyev YD. Multidimensional global optimization algorithm based on adaptive diagonal curves. *Comput Math Math Phys* 2003;43(1):40–56.
66. Meewella CC, Mayne DQ. An algorithm for global optimization of Lipschitz continuous functions. *J Optim Theory Appl* 1988;57(2):307–322.

67. Mladineo RH. An algorithm for finding the global maximum of a multimodal multivariate function. *Math Program* 1986;34(2):188–200.
68. Piyavskij SA. An algorithm for finding the absolute extremum of a function. *USSR Comput Math Math Phys* 1972;12(4):57–67; (In Russian: *Zh Vychisl Mat Mat Fiz* 1972;12(4):888–896).
69. Sergeyev YD. An information global optimization algorithm with local tuning. *SIAM J Optim* 1995;5(4):858–870.
70. Sergeyev YD. Global one-dimensional optimization using smooth auxiliary functions. *Math Program* 1998;81(1):127–146.
71. Shubert BO. A sequential method seeking the global maximum of a function. *SIAM J Numer Anal* 1972;9(3):379–388.
72. Strongin RG. Multiextremal minimization for measurements with interference. *Eng Cybern* 1969;16:105–115.
73. Vasiliev FP. Numerical methods for solving extremum problems (in Russian). Moscow: Nauka; 1988.
74. Wood GR. Multidimensional bisection applied to global optimisation. *Comput Math Appl* 1991;21(6–7):161–172.
75. Žilinskas A. Global Optimization: Axiomatics of statistical models, algorithms, and applications (in Russian). Vilnius: Mokslas; 1986.
76. Baritomba W. Customizing methods for global optimization—A geometric viewpoint. *J Glob Optim* 1993;3(2):193–212.
77. Baritomba W, Cutler A. Accelerations for global optimization covering methods using second derivatives. *J Glob Optim* 1994;4(3):329–341.
78. Breiman L, Cutler A. A deterministic algorithm for global optimization. *Math Program* 1993;58(1–3):179–199.
79. Calvin JM, Žilinskas A. On the convergence of the P-algorithm for one-dimensional global optimization of smooth functions. *J Optim Theory Appl* 1999;102(3):479–495.
80. Evtushenko YG, Malkova VU, Stanevichyus AA. Parallel global optimization of functions of several variables. *Comput Math Math Phys* 2009;49(2):246–260.
81. Floudas CA, Pardalos PM, editors. State of the art in global optimization: computational methods and applications. Dordrecht: Kluwer Academic Publishers; 1996.
82. Gergel VP. A global search algorithm using derivatives. In: Neimark YI, editor. *Systems dynamics and optimization* (in Russian). Nizhni Novgorod, Russia: NNGU Press; 1992. pp. 161–178.
83. Gergel VP. A global optimization algorithm for multivariate function with Lipschitzian first derivatives. *J Glob Optim* 1997;10(3):257–281.
84. Kvasov DE, Sergeyev YD. A univariate global search working with a set of Lipschitz constants for the first derivative. *Optim Lett* 2009;3(2):303–318.
85. Sergeyev YD. A method using local tuning for minimizing functions with Lipschitz derivatives. In: Bomze IM, Csendes T, Horst R, *et al.*, editors. *Developments in global optimization*. Dordrecht: Kluwer Academic Publishers; 1997. pp. 199–216.
86. Sergeyev YD. Multidimensional global optimization using the first derivatives. *Comput Math Math Phys* 1999;39(5):711–720.
87. Casado LG, García I, Sergeyev YD. Interval algorithms for finding the minimal root in a set of multiextremal non-differentiable one-dimensional functions. *SIAM J Sci Comput* 2002;24(2):359–376.
88. Daponte P, Grimaldi D, Molinaro A, *et al.* An algorithm for finding the zero-crossing of time signals with lipschitzian derivatives. *Measurement* 1995;16(1):37–49.
89. Daponte P, Grimaldi D, Molinaro A, *et al.* Fast detection of the first zero-crossing in a measurement signal set. *Measurement* 1996;19(1):29–39.
90. Molinaro A, Sergeyev YD. An efficient algorithm for the zero-crossing detection in digitized measurement signal. *Measurement* 2001;30(3):187–196.
91. Sergeyev YD, Daponte P, Grimaldi D, *et al.* Two methods for solving optimization problems arising in electronic measurements and electrical engineering. *SIAM J Optim* 1999;10(1):1–21.
92. Sergeyev YD. Univariate global optimization with multiextremal non-differentiable constraints without penalty functions. *Comput Optim Appl* 2006;34(2):229–248.
93. van Dam ER, Husslage B, den Hertog D. One-dimensional nested maximin designs. *J Global Optim* 2010;46(2):287–306.
94. Mladineo RH. Convergence rates of a global optimization algorithm. *Math Program* 1992;54(1–3):223–232.
95. Sergeyev YD, Grishagin VA. Parallel asynchronous global search and the nested optimization scheme. *J Comput Anal Appl*

- 2001;3(2):123–145.
96. Zhigljavsky AA. Theory of global random search. Dordrecht: Kluwer Academic Publishers; 1991.
 97. Fiacco AV, McCormick GP. Nonlinear programming: sequential unconstrained minimization techniques. New York: John Wiley & Sons, Inc.; 1968. Reprinted by SIAM Publications; 1990.
 98. Jones DR. The DIRECT global optimization algorithm. In: Floudas CA, Pardalos PM, editors. Volume 1, Encyclopedia of optimization. Dordrecht: Kluwer Academic Publishers; 2001. pp. 431–440.
 99. Strongin RG. Numerical methods for multiextremal nonlinear programming problems with nonconvex constraints. In: Demyanov VF, Pallaschke D, editors. Nondifferentiable optimization: motivations and applications. Proceedings, 1984. Volume 255 of Lecture Notes in Economics and Mathematical Systems. Laxenburg: Springer. IIASA; 1985. pp. 278–282.
 100. Strongin RG, Markin DL. Minimization of multiextremal functions with nonconvex constraints. Cybernetics 1986;22:486–493.
 101. Barkalov KA, Strongin RG. A global optimization technique with an adaptive order of checking for constraints. Comput Math Math Phys 2002;42(9):1289–1300.
 102. Sergeyev YD, Markin DL. An algorithm for solving global optimization problems with nonlinear constraints. J Glob Optim 1995;7(4):407–419.
 103. Sergeyev YD, Famularo D, Pugliese P. Index branch-and-bound algorithm for Lipschitz univariate global optimization with multiextremal constraints. J Glob Optim 2001;21(3):317–341.
 104. Sergeyev YD. On convergence of “Divide the Best” global optimization algorithms. Optimization 1998;44(3):303–325.
 105. Sergeyev YD, Pugliese P, Famularo D. Index information algorithm with local tuning for solving multidimensional global optimization problems with multiextremal constraints. Math Program 2003;96(3):489–512.
 106. Sergeyev YD, Khalaf FMH, Kvasov DE. Univariate algorithms for solving global optimization problems with multiextremal non-differentiable constraints. In: Törn A, Žilinskas J, editors. Models and algorithms for global optimization. New York: Springer; 2007. pp. 123–140.
 107. Sergeyev YD, Kvasov DE, Khalaf FMH. A one-dimensional local tuning algorithm for solving GO problems with partially defined constraints. Optim Lett 2007;1(1):85–99.
 108. Schoen F. On a sequential search strategy in global optimization problems. Calcolo 1982;19:321–334.
 109. Shen Z, Zhu Y. An interval version of Shubert’s iterative method for the localization of the global maximum. Computing 1987;38(3):275–280.
 110. Timonov LN. An algorithm for search of a global extremum. Eng Cybern 1977;15:38–44.
 111. Danilin YuM. Estimation of the efficiency of an absolute-minimum-finding algorithm. USSR Comput Math Math Phys 1971;11(4):261–267.
 112. Ivanov VV. On optimal algorithms for the minimization of functions of certain classes (in Russian). Cybernetics 1972;4:81–94.
 113. Sukharev AG. Optimal strategies of the search for an extremum. USSR Comput Math Math Phys 1971;11(4):119–137.
 114. Sukharev AG. Best sequential search strategies for finding an extremum. USSR Comput Math Math Phys 1972;12(1):39–59.
 115. Sukharev AG. Minimax algorithms in problems of numerical analysis (in Russian). Moscow: Nauka; 1989.
 116. Horst R, Nast M, Thoai NV. New LP bound in multivariate Lipschitz optimization: Theory and applications. J Optim Theory Appl 1995;86(2):369–388.
 117. Norkin VI. On Piyavskij’s method for solving the general global optimization problem. Comput Math Math Phys 1992;32(7):873–886.
 118. Baoping Zhang, Wood GR, Baritompa W. Multidimensional bisection: the performance and the context. J Glob Optim 1993;3(3):337–358.
 119. Adjiman CS, Dallwig S, Floudas CA, *et al.* A global optimization method, α BB, for general twice-differentiable constrained NLPs – I. Theoretical advances. Comput Chem Eng 1998;22(9):1137–1158.
 120. Androulakis IP, Maranas CD, Floudas CA. α BB: a global optimization method for general constrained nonconvex problems. J Glob Optim 1995;7(4):337–363.
 121. Floudas CA, Akrotirianakis IG, Caratzoulas S, *et al.* Global optimization in the 21st century: Advances and challenges. Comput Chem Eng 2005;29(6):1185–1202.

122. Molinaro A, Pizzuti C, Sergeyev YD. Acceleration tools for diagonal information global optimization algorithms. *Comput Optim Appl* 2001;18(1):5–26.
123. Neimark YuI, Strongin RG. The information approach to the problem of search of extrema of functions. *Eng Cybern* 1966;1:17–26.
124. Betrò B. Bayesian methods in global optimization. *J Glob Optim* 1991;1(1):1–14.
125. Calvin JM. A lower bound on convergence rates of nonadaptive algorithms for univariate optimization with noise. *J Glob Optim* 2010;48(1):17–27.
126. Gaviano M, Lera D, Steri AM. A local search method for continuous global optimization. *J Glob Optim* 2010;48(1):73–85.
127. Lera D, Sergeyev YD. An information global minimization algorithm using the local improvement technique. *J Glob Optim* 2010. DOI: 10.1007/s10898-009-9508-x.
128. Mockus J. Bayesian approach to global optimization. Dordrecht: Kluwer Academic Publishers; 1989.
129. Mockus J, Eddy W, Mockus A, *et al.* Bayesian heuristic approach to discrete and global optimization. Dordrecht: Kluwer Academic Publishers; 1996.
130. Žilinskas A. Axiomatic approach to statistical models and their use in multimodal optimization theory. *Math Program* 1982;22(1):104–116.
131. Zhigljavsky AA, Hamilton E. Stopping rules in k -adaptive global random search algorithms. *J Glob Optim* 2010;48(1):87–97.
132. Evtushenko YG. Numerical methods for finding global extrema (Case of a non-uniform mesh). *USSR Comput Math Math Phys* 1971;11(6):38–54.
133. Evtushenko YG, Malkova VU, Stanevichyus AA. Parallelization of the global extremum searching process. *Autom Remote Control* 2007;68(5):787–798.
134. Nefedov VN. Some problems of solving Lipschitzian global optimization problems using the branch-and-bound method. *Comput Math Math Phys* 1992;32(4):433–445.
135. Kvasov DE, Pizzuti C, Sergeyev YD. Local tuning and partition strategies for diagonal GO methods. *Numer Math* 2003;94(1):93–106.
136. Sergeyev YD. A one-dimensional deterministic global minimization algorithm. *Comput Math Math Phys* 1995;35(5):705–717.
137. Meewella CC, Mayne DQ. Efficient domain partitioning algorithms for global optimization of rational and Lipschitz continuous functions. *J Optim Theory Appl* 1989;61(2):247–270.
138. Finkel DE, Kelley CT. Additive scaling and the DIRECT algorithm. *J Glob Optim* 2006;36(4):597–608.
139. Gablonsky JM, Kelley CT. A locally-biased form of the DIRECT algorithm. *J Glob Optim* 2001;21(1):27–37.
140. He J, Watson LT, Ramakrishnan N, *et al.* Dynamic data structures for a direct search algorithm. *Comput Optim Appl* 2002;23(1):5–25.
141. Sergeyev YD, Kvasov DE. Global search based on efficient diagonal partitions and a set of Lipschitz constants. *SIAM J Optim* 2006;16(3):910–937.
142. Wood GR, Baoping Zhang. Estimation of the Lipschitz constant of a function. *J Glob Optim* 1996;8(1):91–103.
143. Strongin RG. Search for global optimum. Volume 2, Mathematics and cybernetics (in Russian). Moscow: Znanie; 1990.
144. Strigul OI. Search for a global extremum in a certain subclass of functions with the Lipschitz condition (in Russian). *Cybernetics* 1985;6:72–76.
145. Wood GR. The bisection method in higher dimensions. *Math Program* 1992;55(1–3):319–337.
146. Mayne DQ, Polak E. Outer approximation algorithm for nondifferentiable optimization problems. *J Optim Theory Appl* 1984;42(1):19–30.
147. Sergeyev YD. Efficient partition of N -dimensional intervals in the framework of one-point-based algorithms. *J Optim Theory Appl* 2005;124(2):503–510.
148. Sergeyev YD. An efficient strategy for adaptive partition of N -dimensional intervals in the framework of diagonal algorithms. *J Optim Theory Appl* 2000;107(1):145–168.
149. Casado LG, García I, Csendes T. A new multisection technique in interval methods for global optimization computing. *Computing* 2000;65(3):263–269.
150. Csallner AE, Csendes T, Markót MCs. Multisection in interval branch-and-bound methods for global optimization—I. Theoretical results. *J Glob Optim* 2000;16(4):371–392.
151. Csendes T, Ratz D. Subdivision direction selection in interval methods for

- global optimization. *SIAM J Numer Anal* 1997;34(3):922–938.
152. Hansen ER, editor. Global optimization using interval analysis. Volume 165, Pure and applied mathematics. New York: M. Dekker; 1992.
 153. Kearfott RB. Rigorous global search: continuous problems. Dordrecht: Kluwer Academic Publishers; 1996.
 154. Ratz D, Csendes T. On the selection of subdivision directions in interval branch-and-bound methods for global optimization. *J Glob Optim* 1995;7(2):183–207.
 155. Clausen J, Žilinskas A. Subdivision, sampling, and initialization strategies for simplicial branch-and-bound in global optimization. *Comput Math Appl* 2002;44(7):943–955.
 156. Gorodetsky SYu. Multiextremal optimization based on domain triangulation (in Russian). *NNGU Bulletin: Math Model Opt Control* 1999;2(21):249–268.
 157. Horst R. On generalized bisection of N -simplices. *Math Comput* 1997;66(218):691–698.
 158. Paulavičius R, Žilinskas J, Grothey A. Investigation of selection strategies in branch-and-bound algorithm with simplicial partitions and combination of Lipschitz bounds. *Optim Lett* 2010;4(2):173–183.
 159. Andramonov MYu, Rubinov AM, Glover BM. Cutting angle methods in global optimization. *Appl Math Lett* 1999;12(3):95–100.
 160. Beliakov G. Cutting angle method—A tool for constrained global optimization. *Optim Methods Softw* 2004;19(2):137–151.
 161. Beliakov G, Ferrer A. Bounded lower subdifferentiability optimization techniques: applications. *J Glob Optim* 2010;47(2):211–231.
 162. Bulatov VP. Methods of embedding-cutting off in problems of mathematical programming. *J Glob Optim* 2010;48(1):3–15.
 163. Fuduli A, Gaudio M, Giallombardo G. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM J Optim* 2004;14(3):743–756.
 164. Kiseleva EM, Stepanchuk T. On the efficiency of a global non-differentiable optimization algorithm based on the method of optimal set partitioning. *J Glob Optim* 2003;25(2):209–235.
 165. Korotchenko AG. An approximately optimal algorithm for finding an extremum for a certain class of functions. *Comput Math Math Phys* 1996;36(5):577–584.
 166. Evtushenko YG, Posypkin MA, Sigal IKh. A framework for parallel large-scale global optimization. *Comput Sci Res Dev* 2009;23(3-4):211–215.
 167. Liuzzi G, Lucidi S, Piccialli V. A partition-based global optimization algorithm. *J Glob Optim* 2010;48(1):113–128.
 168. Tawarmalani M, Sahinidis NV. Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications. Dordrecht: Kluwer Academic Publishers; 2002.
 169. Carr CR, Howe CW. Quantitative decision procedures in management and economic: deterministic theory and applications. New York: McGraw-Hill; 1964.
 170. Gaviano M, Kvasov DE, Lera D, *et al.* Algorithm 829: Software for generation of classes of test functions with known local and global minima for global optimization. *ACM Trans Math Softw* 2003;29(4):469–480.
 171. Butz AR. Space-filling curves and mathematical programming. *Inform Control* 1968;12(4):314–330.
 172. Strongin RG. Algorithms for multiextremal mathematical programming problems employing the set of joint space-filling curves. *J Glob Optim* 1992;2(4):357–378.
 173. Sagan H. Space-filling curves. New York: Springer; 1994.
 174. Strongin RG. Global optimization using space filling. In: Floudas CA, Pardalos PM, editors. Volume 2, Encyclopedia of optimization. Dordrecht: Kluwer Academic Publishers; 2001. pp. 345–350.
 175. Strongin RG, Sergeyev YD. Global multidimensional optimization on parallel computer. *Parallel Comput* 1992;18(11):1259–1273.
 176. Gourdin E, Jaumard B, Ellaia R. Global optimization of Hölder functions. *J Glob Optim* 1996;8(4):323–348.
 177. Lera D, Sergeyev YD. Global minimization algorithms for Hölder functions. *BIT* 2002;42(1):119–133.
 178. Lera D, Sergeyev YD. Lipschitz and Hölder global optimization using space-filling curves. *Appl Numer Math* 2010;60(1–2):115–129.
 179. Vanderbei RJ. Extension of Piyavskii's algorithm to continuous global optimization. *J Glob Optim* 1999;14(2):205–216.
 180. Pintér JD. Convergence qualification of adaptive partition algorithms in global

- optimization. *Math Program* 1992;56(1–3): 343–360.
181. Kvasov DE. Multidimensional Lipschitz global optimization based on efficient diagonal partitions. *4OR Q J Oper Res* 2008;6(4): 403–406.
 182. Grishagin VA. Operating characteristics of some global search algorithms. Volume 7, *Problems of stochastic search*. Riga: Zinatne; 1978. pp. 198–206 (in Russian).
 183. Grishagin VA, Sergeyev YD, Strongin RG. Parallel characteristic algorithms for solving problems of global optimization. *J Glob Optim* 1997;10(2):185–206.
 184. Addis B, Locatelli M. A new class of test functions for global optimization. *J Glob Optim* 2007;38(3):479–501.
 185. Famularo D, Pugliese P, Sergeyev YD. Test problems for Lipschitz univariate global optimization with multiextremal constraints. In: Dzemyda G, Šaltenis V, Žilinskas A, editors. *Stochastic and global optimization*. Dordrecht: Kluwer Academic Publishers; 2002. pp. 93–109.
 186. Gallagher M, Yuan B. A general-purpose tunable landscape generator. *IEEE Trans Evolut Comput* 2006;10(5):590–603.
 187. Lavor C, Maculan N. A function to test methods applied to global minimization of potential energy of molecules. *Numer Algorithms* 2004;35(2–4):287–300.
 188. Locatelli M. On the multilevel structure of global optimization problems. *Comput Optim Appl* 2005;30(1):5–22.
 189. Neumaier A, Shcherbina O, Huyer W *et al.*, A comparison of complete global optimization solvers. *Math Program* 2005; 103(2):335–356.
 190. Schittkowski K. More test examples for nonlinear programming codes. Volume 282, *Lecture notes in economics and mathematical systems*. Berlin: Springer; 1987.
 191. Schoen F. A wide class of test functions for global optimization. *J Glob Optim* 1993;3(2):133–137.

LITTLE'S LAW AND RELATED RESULTS

RONALD W. WOLFF

Department of Industrial Engineering and
Operations Research, University of
California at Berkeley, Berkeley,
California

Queueing theory is about the analysis of mathematical models where entities called *customers* arrive at some facility called the *system*, spend *waiting time* in system, and then depart.

In its early history, methods of analysis were tailored to individual models and varied widely. There were no theorems of any consequence that held across the board from model to model. This changed in 1961 with " $L = \lambda W$ " [1] by John Little, often called *Little's Law* today, and we will sometimes abbreviate as "LL." It, together with extensions, is of fundamental importance for both the foundations of queueing theory and its applications. The literature on this topic since then is enormous. It is impossible to review all of that here. Instead, our goal is to explain LL and important extensions, demonstrate their usefulness, show the connection with some related topics, and provide a literature guide sufficient for further investigation.

LL states that the time average of the number of customers in system is the product of the arrival rate and the customer average of the waiting times.

This article is organized as follows. We present an intuitive explanation and illustrate applications of LL in the section titled "Notation, Explanation, Applications." There are two versions of LL. We prove the sample-path version in the section titled "Sample-Path Proof of Little's Law," treat the stationary version in the section titled "Stationary Version of Little's Law," and extend LL in the section titled "Extension to $H = \lambda G$; Work." We present the equivalence of $H = \lambda G$ to a result in the section titled "A Rate Conservation Law," present another related result in the section titled "A Distributional Little's

Law," and review the literature and other related results in the section titled "Brief History and Literature Review."

NOTATION, EXPLANATION, APPLICATIONS

For $i \geq 1$, let customer C_i arrive at time t_i , have waiting W_i , and depart at $t_i + W_i$, where $t_0 \equiv 0 \leq t_1 \leq t_2 \leq \dots$. For any $t \geq 0$, C_i is *in the system* at t if $t_i \leq t < t_i + W_i$. Define the corresponding indicator, $I_i(t) = 1$ if $t_i \leq t < t_i + W_i$, and $I_i(t) = 0$ otherwise, where $\int_0^\infty I_i(t) dt = W_i$. Let $N(t) = \sum_{i=1}^\infty I_i(t)$, the *number of customers in system* at time t , and $\Lambda(t) = \max\{i : t_i \leq t\}$, the number of arrivals by time t . With $\Lambda(t)$, we rewrite $N(t) = \sum_{i=1}^{\Lambda(t)} I_i(t)$.

From customer data $\{t_i\}$ and $\{W_i\}$ for all i , we can construct $\{N(t)\}$ for all t . We plot $\{\Lambda(t)\}$ as a step function and show some customer waiting-time data in Fig. 1, and the constructed $\{N(t)\}$ in Fig. 2.

In Fig. 1, C_3 arrives after but departs before C_2 . There is no way to deduce this from Figure 2, which tells us only how many customers are in the system, not *which* customers are there. It has less information than Fig. 1.

While waiting times are lengths of time, they are also displayed as rectangles of height 1 in Fig. 2. Thus *length* W_i is also *area* W_i . For each t , $N(t)$ is simply the number of waiting time rectangles that contain point t .

In a (stochastic) queueing model, t_i , W_i , and $N(t)$ are random variables; $\{t_i, i \geq 1\}$, $\{W_i, i \geq 1\}$, and $\{N(t), t \geq 0\}$ are stochastic processes. These random quantities are defined on some sample space. At ω , some arbitrary point in this space, $t_i(\omega)$, $W_i(\omega)$, and $N(t, \omega)$ are numbers, and the collection of stochastic processes is called a *sample path* or *realization*. A simulation of a queueing model generates a sample path. We may also observe a real system as it evolves over time without having a formal model, and view the observed evolution of arrivals, waiting times, and number in system as a sample path.

Usually, assumptions are made such that *stationary versions* of these processes exist, where in particular for these versions, W_i has the same distribution for all i and $N(t)$ has the same distribution for all t . Let W and N be corresponding random variables that have these distributions.

There are two versions of Little's Law. The *sample-path* version relates sample-path averages at some ω in the sample space; the *stationary* version relates the expected values of W and N . The former is easy to understand and so intuitively appealing that, to some, it may not require a proof (we think it does). It also applies to data collected on real systems. The latter is a common way to find these quantities, and is an important tool for the analysis of queueing models. We now present the sample-path version, but omit writing ω explicitly; for example, we write W_i rather than $W_i(\omega)$.

Rectangular area W_i is the contribution to the area under $\{N(t)\}$ that is made by customer C_i . Now consider the area under $\{N(t)\}$ on some interval $(0, T)$, where, as in Figs 1 and 2, T has the property that $N(T) = 0$. For any such T , this area may be

written either as an integral or a sum,

$$\int_0^T N(t) dt = \int_0^T \sum_{i=1}^{\Lambda(t)} I_i(t) dt = \sum \int = \sum_{i=1}^{\Lambda(T)} W_i,$$

where in our figures, $\Lambda(T) = 4$. For any T where $N(T) > 0$, Equation (1) does not hold because a portion of at least one rectangle in the sum contributes to the area under $\{N(t)\}$ after T . In this case, sum – integral $\equiv$ error ≥ 0 . We define long run averages as limits, when they exist, and name them.

$$L = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T N(t) dt, \quad w = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n W_i, \\ \lambda = \lim_{t \rightarrow \infty} \frac{\Lambda(t)}{t}, \quad (1)$$

where L is the average number of customers in the system, w is the average waiting time (of customers in the system), and λ is the arrival rate.

L and w are the two most common performance measures for a queue. We present an intuitive explanation below and a proof in the section titled “Sample-Path Proof of Little's

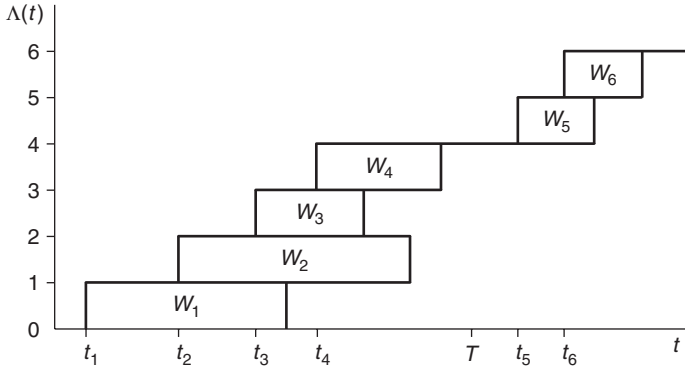


Figure 1. Waiting time data.

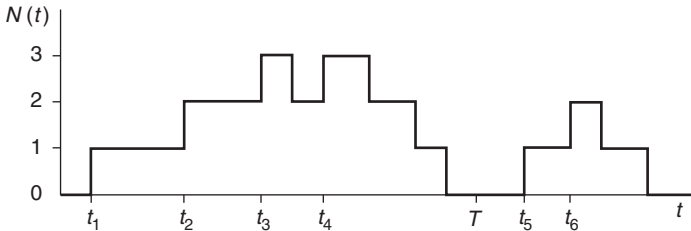


Figure 2. Number of customers in system data.

Law” of this fundamental relation between them (sample-path version):

Theorem 1 [Little’s Law, LL, or $L = \lambda w$]. *If limits λ and w exist and are finite, L exists and is finite, where*

$$L = \lambda w. \quad (2)$$

The basic reason for Equation (3) is Equation (1). For any T (where now $N(T) \geq 0$) we write

$$\begin{aligned} \frac{\int_0^T N(t) dt}{T} &= \frac{\sum_{i=1}^{\Lambda(T)} W_i}{T} - \frac{\text{error}}{T} \\ &= \left(\frac{\Lambda(T)}{T} \right) \left(\frac{\sum_{i=1}^{\Lambda(T)} W_i}{\Lambda(T)} \right) - \frac{\text{error}}{T}. \end{aligned}$$

As T gets large, the first quantity in parenthesis on the right approaches λ , the second approaches w , and the product approaches λw . Hence Theorem 1 holds if the term on the far right approaches 0 as $T \rightarrow \infty$. The error term itself need not get small, but only grows more slowly than T .

Note that Equation (3) holds for some unspecified ω . If limits w and λ hold as constants on a set of ω with probability 1 (w.p.1), Equation (3) holds as a constant w.p.1.

Usually, “ $L = \lambda w$ ” is written “ $L = \lambda W$.” We prefer to use W and N as defined above, in order to write the stationary version as $E(N) = \lambda E(W)$; see the section titled “Stationary version of Little’s Law.” The notation “ L ” has a long history, meaning the average length of a waiting *line*. We now present an application, based on personal experience.

Example 1 [Wine Cellars]. There are about 3000 bottles in my friend Loy’s wine cellar ($L = 3000$). He does not keep records of when each bottle was bought and consumed. He wondered “about how old, on average, are these wines when they are drunk?” He estimates that (with some help) he consumes about 300 bottles per year ($\lambda = 300$). The

answer to his question is $w = L/\lambda = 10$ years. Not quite true. They are on average 10 years older than when he bought them.

When designing my wine cellar, I had to decide its size. I estimated that on average, wines would be held 5 years when they are drunk ($w = 5$ years), and I would drink about 200 bottles per year ($\lambda = 200$). Thus $L = \lambda w = 1000$ bottles, and the capacity of the wine cellar should be somewhat larger to account for fluctuations about the average. (Fluctuations are much less in wine cellars than in a typical queueing model.) It turned out that the wine cellar was too small. What went wrong? Not LL. I had underestimated λ !

Viewing the system as a “black box,” we do not have to know anything about what goes on inside the box; LL holds. In this example, it is natural to think of a collection of wine bottles as *inventory*, and in fact, queueing theory is closely related to what is called *inventory theory*. These theories differ in how we model what goes on inside the box, what aspects of the model may be subject to some control, and what decisions are being contemplated.

In most queueing models, the arrival process is treated as given. At a wine shop, arrivals are scheduled. At a wine cellar, both arrivals and departures are scheduled. We now describe some queueing models of what is inside the box.

In elementary situations, each arrival at a facility has a *service time* to be performed by a server, where the facility has one or more servers. Arrivals, finding all servers busy serving other customers, wait in a queue. While in the system, a customer may spend some time in queue, followed by a service time.

The time that a customer spends in queue before service begins we call *delay in queue* or just *delay* (sometimes called *waiting time in queue*). Thus a customer’s waiting time is the *sum* of the corresponding delay and service time.

Similarly, at any time there will be a *queue length* (number of customers in queue) and a *number of customers in service*, where their sum is the *waiting-line length* (number of customers in system).

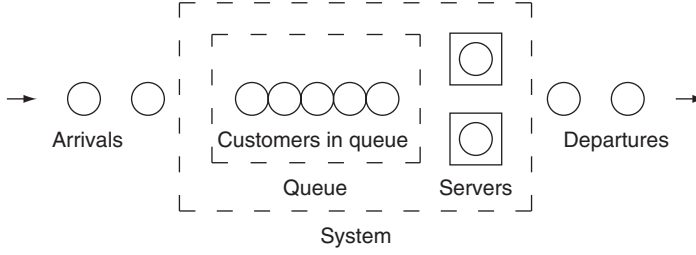


Figure 3. Snapshot of a queueing system.

In Fig. 3, we show a “snapshot” of a system where customers are represented as circles and servers as square boxes. We have five customers in queue, two servers, two customers in service, and a total of seven customers in system.

Let C_i have delay D_i and service time S_i , $i \geq 1$, where $W_i = D_i + S_i$. Similarly, let $N_q(t)$ and $N_s(t)$ be the respective (constructed) number of customers in queue and in service at time t , $t \geq 0$, where $N(t) = N_q(t) + N_s(t)$.

Let the customer average of $\{D_i\}$ and $\{S_i\}$ be d and $E(S)$, respectively, where the notation “ $E(S)$ ” is often used because, in many applications, the service times S_i are independent and identically distributed (i.i.d.) random variables, and their customer average is their expected value. Let the time average of $\{N_q(t)\}$ and $\{N_s(t)\}$ be, respectively, Q and L_s . From LL, we immediately have

$$Q = \lambda d \text{ and } L_s = \lambda E(S). \quad (3)$$

To obtain Equation (4), we did not mention the order of service of customers in queue. In Fig. 3, it would appear to be first-in-first-out (FIFO), also called *first-come-first-served* (FCFS); FIFO is easier to pronounce. Some alternatives are last-in-first-out (LIFO) and shortest job (service time) first (SJF). SJF usually (but not necessarily) reduces both Q and d . Equation (4) holds in each case. We do not have to know what is going on inside the customers-in-queue box.

Is λ in the expression for Q in Equation (4) correct? Arrivals finding an idle server do not join the queue. To investigate, let β be the (long run) fraction of arrivals who find all servers busy and join the queue, so $\beta\lambda$ is the arrival rate of these customers. Let d_β be the average delay of these customers. From LL, $Q = \beta\lambda d_\beta$. The average delay of customers

who do not join the queue is 0, and averaging over both groups, $d = \beta d_\beta + (1 - \beta)0 = \beta d_\beta$, and indeed, $Q = \lambda d$ is correct. For LL, L , λ , and w must be averages over the same stream of customers.

The second equation in Equation (4) also deserves a closer look. It is the average number of customers in service, assuming all are served. This will be true when the number of servers exceeds $\lambda E(S)$, which is also the average number of busy servers. When there is one server, it is the fraction of time that server is busy.

Little’s Law may be applied in several ways. If either L or w is known or can be found by an independent analysis (and we know λ), then we know the other. In other applications, this equation, together with other information, determines both quantities. In a typical queueing *model*, λ is known (is a parameter). In a real system, it is not known; sometimes it is estimated from estimates of L and w . We now illustrate a different application.

Example 2 [Order of Service]. At one time at banks and post offices in the United States, a separate queue formed at each teller (server). These days, a single queue feeds all tellers. (We exclude from consideration special-purpose tellers.) How does changing from multiple queues (MQ) to a single queue (SQ) affect customers? In particular, what is the effect on *average* waiting time?

To consider this question, assume arrival times t_i and service times S_i are fixed. We are changing the *order* of service. (Should there be a positive queue behind one teller when another teller becomes free, we assume that a customer will shift immediately to the idle teller.) With SQ, customers are served FIFO. With MQ, service times *in the order served* are a rearrangement of the S_i .

Suppose customer service times are i.i.d., independent of all else. Service times in the order served will be i.i.d., and the *distribution* of the number of customers in system will not change. That is, $\{N(t), t \geq 0\}$ under MQ is essentially the same process as under SQ. (The number in system processes are said to be *stochastically equivalent*.) Hence time-average L has not changed. As λ has not changed,

$$w = L/\lambda \text{ has not changed.}$$

This does not mean that customers are indifferent to the change. The waiting time *distribution* will change. In fact, from the FIFO property of SQ, it is easily shown that the *variance* of *delay* under SQ is smaller than under MQ, and also under any alternative of this nature, such as an SQ that operates LIFO.

Some queueing models are much more complex. A farmers' market may be viewed as a *queueing network*, where each arriving customer is served at some sequence of stands (stations), and then departs. LL applies to the arrival and departure of customers at the market, and also to the arrival and departure of customers at individual stations. This is an *open network*. In a *closed network*, customers move from station to station, but there are no arrivals to or departures from the system. For closed networks, LL holds at individual stations.

SAMPLE-PATH PROOF OF LITTLE'S LAW

Our intuitive explanation for LL in the section titled "Notation, Explanation, Applications" is the basis for a formal proof:

Proof. Motivated by Figs 1 and 2, we obtain the inequalities in Equation (5), where the sum on the left is over those waiting-time rectangles that have ended by T .

$$\sum_{\{i: t_i + W_i \leq T\}} W_i \leq \int_0^T N(t) dt \leq \sum_{i=1}^{\Lambda(T)} W_i. \quad (4)$$

Now divide Equation (5) by T and let $T \rightarrow \infty$. The right-hand expression has limit λw , which implies the left-hand expression has $\limsup \leq \lambda w$. LL follows if it has limit λw . Toward that end, we need two preliminary results. First, write

$$\frac{W_n}{n} = \frac{1}{n} \sum_{i=1}^n W_i - \left(\frac{n-1}{n}\right) \left(\frac{1}{n-1}\right) \sum_{i=1}^{n-1} W_i.$$

For finite w , this expression has limit $w - w = 0$ as $n \rightarrow \infty$. That is,

$$\lim_{n \rightarrow \infty} W_n/n = 0. \quad (5)$$

Finite λ implies $t_n \rightarrow \infty$ and hence $\Lambda(t_n)/t_n \rightarrow \lambda$ as $n \rightarrow \infty$. If the arrival times are distinct, $\Lambda(t_n) = n$; otherwise, $\Lambda(t_n) \geq n$, and we write

$$\frac{W_n}{t_n} = \frac{W_n}{n} \frac{n}{t_n} \leq \frac{W_n}{n} \frac{\Lambda(t_n)}{t_n}.$$

For finite λ and w and from Equation (6), we have

$$\lim_{n \rightarrow \infty} W_n/t_n = 0. \quad (6)$$

For any $\epsilon > 0$, Equation (7) implies that for some finite m and all $i > m$, $W_i < t_i \epsilon$ and $t_i + W_i < t_i(1 + \epsilon)$. Thus $t_i \leq T/(1 + \epsilon) \implies t_i + W_i \leq T$ for every $i > m$, which gives this *lower bound* on the left-hand expression in Equation (5):

$$\sum_{i=1}^{\Lambda\left(\frac{T}{1+\epsilon}\right)} W_i - \sum_{i=1}^m W_i \leq \sum_{\{i: t_i + W_i \leq T\}} W_i. \quad (7)$$

Now divide Equation (8) by T and let $T \rightarrow \infty$, noting that the second term on the left is a constant. The left-hand side has limit $\lambda w/(1 + \epsilon)$, which implies the right-hand side has $\liminf \geq \lambda w/(1 + \epsilon)$. Because ϵ can be arbitrarily small, this implies the $\liminf \geq \lambda w$. Combining with $\limsup$, the limit is λw , and we have Equation (3).

Some early proofs (but after Little [1]) assumed the system would empty from time to time. Little did not require that, nor do we,

aside from our convention that the system is empty initially (which can be dispensed with). To illustrate, suppose C_i arrives at $t_i = 2i$, with service time $S_i = 3$, $i \geq 1$, at a two-server system. It is easy to see that (draw the figure!) even though the system never empties, the limits are $\lambda = 1/2$, $w = 3$, and $L = \lambda w = 1.5$.

Now, let $\Lambda^d(t)$ be the number of departures by t . When Equation (6) holds, we have Equation (7). By the same method, it is easy to show that

$$\lim_{t \rightarrow \infty} \Lambda^d(t)/t = \lambda, \quad (8)$$

that is, the departure rate is equal to the arrival rate. (We do not require $w < \infty$; see following section.)

Technical Considerations; the Case $w = \infty$

In Figs 1 and 2, $\{W_i\}$ determines $\{N(t)\}$, but not the converse. It also turns out that when finite limits L and $\lambda > 0$ exist, w may not exist; for examples, see Ref. 2 and p. 289 of Ref. 3. In both cases, the waiting time average oscillates, and limit (Eq. 6) does not exist.

Suppose $0 < \lambda < \infty$ and w exist, with $w = \infty$. It would be nice to have $L = \infty$ as well, and this is known to be true for a variety of queueing models.

To deal with this question, suppose Equation (6) still holds, which gives Equations (7) and (8). Now divide Equation (8) by T . For $w = \infty$, the left-hand side $\rightarrow \infty$ as $T \rightarrow \infty$. When $0 < \lambda < \infty$, $w = \infty$, and Equation (6) holds, L is well defined, where

$$L = \infty.$$

For a single-server FIFO queue with finite average service time $E(S)$, where $\rho \equiv \lambda E(S) < 1$, it is easy to show that Equation (6) holds, even when $w = \infty$. However, this is not always true. For example, consider an ∞ -server queue with i.i.d. service times, so the $W_i = S_i$ are i.i.d. For i.i.d. W_i , $w < \infty$ w.p.1 if and only if Equation (6) holds w.p.1; see p. 239 of Ref. 4. The point is not whether $L = \infty$ in this example, but rather to question when it is valid to use Equation (6) to prove it.

STATIONARY VERSION OF LITTLE'S LAW

Queueing models are rarely so simple that either L or w can be found directly as sample-path averages. Instead, we find the distributions of N or W , or enough about them to find their means. This is usually done by a *stationary analysis*.

A substantial portion of the queueing literature is about continuous-time Markov-chain models, sometimes with complex structure. For a queueing network, we let $N_k(t)$ be the number of customers at station k at time t , and $N(t)$ be their sum over k . We replace exponential service with *phase-type* service, and so on. For positive-recurrent chains, stationary N exists, with distribution determined by the balance equations, and $L = E(N)$ w.p.1, even when infinite. In these models, $\lambda > 0$ (in open networks), and $w = L/\lambda$. Sometimes sophisticated techniques are required to determine positive recurrence and to solve for stationary distributions, or their means, or approximations of them, but we have changed the playing field from finding sample-path averages directly.

Sometimes, we find properties of W directly, and from them, properties of N . For a single-server queue with *renewal arrivals* (i.i.d. interarrival times) and i.i.d. service times (the $GI/G/1$ queue), FIFO delays D_i satisfy the recursion

$$D_{i+1} = \max\{D_i + S_i - T_i, 0\}, \quad i \geq 1, \quad (9)$$

where *interarrival time* $T_i = t_{i+1} - t_i$. Assume $0 < \lambda E(S) < 1$. By letting D_i and D_{i+1} in Equation (10) have the same distribution (a stationary analysis), we find approximations and bounds for the mean and other properties of D , W , and N .

Little's Law via Stationary Marked Point Processes

A (one-dimensional) *marked point process* (MPP) is a sequence $\{t_i, k_i\}$, $i \geq 1$, where $0 \leq t_1 \leq t_2 \leq \dots$ are the points, and the k_i are the marks. For queueing theory applications, the t_i are arrival times, so $\{t_i\}$ (the point process) is an arrival process, and for each i , k_i is something associated with arrival i ; in a

single-server queue, for example, it would be the service time of arrival i .

An MPP is called *stationary* MPP (SMPP) if $t_1 > 0$ and $\{\Lambda(t)\}$ has stationary increments, which means that for every t , the distribution of $\Lambda(t+h) - \Lambda(h)$ is independent of h . For an SMPP, it is easily shown that for all t , $E\{\Lambda(t)\} = \lambda t$, where $\lambda > 0$ is the arrival rate. For renewal arrivals, where $T_i = t_{i+1} - t_i$, $i \geq 1$, have distribution function F and mean $1/\lambda$, we get the corresponding SMPP by letting t_1 have *equilibrium* density function $f_e(u) = \lambda[1 - F(u)]$, $u \geq 0$.

$\{\Lambda(t)\}$ is *time* stationary. For some $\{\Lambda(t)\}$, we would like to determine the distribution of *event* (or *customer*) stationary $\{T_i\}$, and the converse. For renewal arrivals, this is easy, but in general, where the T_i are dependent, this is not the case. One of the major achievements of point-process theory was to *invert* (determine one from the other) the distributions of $\{\Lambda(t)\}$ and $\{T_i\}$ (the latter is called the *Palm* distribution), where the one we start with is stationary and ergodic. Originally, this was for *simple* arrival processes, where this means $T_i > 0$. Later, this was extended to allow $P(T_i = 0) > 0$ (batch arrivals).

We also have marks. Starting with stationary $\{T_i, W_i\}$, the corresponding $\{\Lambda(t), N(t)\}$ was constructed in Ref. 5 (and in earlier work referenced there), where $\{N(t)\}$ is stationary, and the stationary version of Little's Law

$$E(N) = \lambda E(W), \quad (10)$$

was shown, even when infinite, without appealing to the sample-path version.

Describing how the inversion or this construction is done is beyond our scope. In addition to Franken *et al.* [5], we cite Daley and Vere-Jones [6] for point-process theory, Sigman [7] for an empirical-inversion

approach, Miyazawa *et al.* [8] for more on Palm theory, and all four for a historical account and references.

Equation (11) appears to be very general. For example, we can view waiting times in some model as though they were generated by an ∞ -server queue, and set $W_i = S_i$, where the S_i are stationary and ergodic. However, this is somewhat misleading because a queueing *model* usually would consist of an arrival process, service requirements, and details about the service facility. Stationary $\{W_i\}$ has to be *constructed*. This is easy to do for a single-server queue, but for, say, a network of multiserver stations, we do not know how to do this, or even whether such a $\{W_i\}$ exists, without having some structure, such as regeneration points (in a general sense), or a Markov process that is Harris recurrent.

We now present a deterministic example where customer- and time-stationary processes are easily constructed independently.

Example 3. Consider a single-server queue with $t_1 = 1$, subsequent interarrival times that alternate, 1, 4, 1, 4, ... (so $\lambda = 1/2.5 = 0.4$), and constant service times $S_i = 2$. We plot $\{N(t)\}$ in Fig. 4.

When $T_i = 1$, C_i finds the system empty, and $W_i = 2$. Similarly, when $T_i = 4$, $W_i = 3$. These events alternate. We have jointly stationary $\{T_i, W_i\}$, where events $\{T_i = 1, W_i = 2\}$ and $\{T_i = 4, W_i = 3\}$ each has probability 1/2.

It is easy to see (as fractions of time) that stationary N has distribution

$$\begin{aligned} P(N = 0) &= P(N = 2) = 0.2, \text{ and} \\ P(N = 1) &= 0.6, \end{aligned} \quad (11)$$

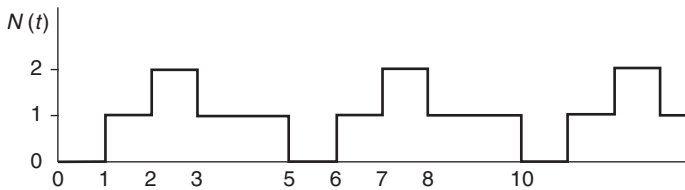


Figure 4. Constructing stationary versions.

with $E(N) = 1.0$. From the distribution of $\{T_i, W_i\}$, $E(W) = 2.5$, and (of course!) $E(N) = \lambda E(W) = (0.4)(2.5)$.

Constructing $\{\Lambda(t)\}$ with stationary increments: As the arrival process is deterministic, its stationary version is completely determined by where the initial point $t_0 = 0$ falls. We call interarrival intervals either *long* (length 4) or *short* (length 1). If t_0 falls in a long, the time until the first arrival is uniformly distributed on $(0, 4)$, with subsequent interarrival times $1, 4, 1, 4, \dots$. If it falls in a short, the time until the first arrival is uniformly distributed on $(0, 1)$, with subsequent interarrival times $4, 1, 4, 1, \dots$. The probability that t_0 falls in a long is 0.8, which is the fraction of the real line that is covered by long intervals.

Selecting $t_0 = 0$ in this manner, we get waiting-time sequence $2, 3, 2, 3, \dots$ with probability 0.8 (t_0 falls in a long), and $3, 2, 3, 2, \dots$ otherwise. $\{W_i\}$ is *not* stationary. Except in special cases such as renewal arrivals, it is not possible for $\{N(t)\}$ and $\{W_i\}$ to be stationary simultaneously on the same sample space.

EXTENSION TO $H = \lambda G$; WORK

For Little's Law, the C_i (and unspecified system behavior) generate $\{W_i\}$ and $\{N(t)\}$, where C_i contributes W_i to the area under $\{N(t)\}$. We now present an extension, where the C_i may generate other discrete- and continuous-time processes that are related in the same way. We consider only the sample-path version, at sample point ω , and again omit writing ω explicitly.

For each C_i , we associate function $f_i(t)$, $t \geq 0$, where $\int_0^\infty |f_i(t)| dt < \infty$, and for some finite $l_i > 0$, $f_i(t) = 0$ for $t \notin [t_i, t_i + l_i)$. Now define

$$G_i = \int_0^\infty f_i(t) dt, \quad i \geq 1,$$

$$\text{and } H(t) = \sum_{i=1}^\infty f_i(t), \quad t \geq 0.$$

When $f_i(t) \geq 0$, G_i is the area under $\{H(t)\}$ contributed by C_i . Often, $l_i = W_i$. We define

customer and time averages, and λ as defined before

$$\bar{G} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n G_i, \quad \text{and } \bar{H} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T H(t) dt.$$

The following is an important extension of LL:

Theorem 2 $[\bar{H} = \lambda \bar{G}]$. *If limits λ and $\bar{G}$ exist and are finite, and technical condition (TC): $l_i/t_i \rightarrow 0$ as $i \rightarrow \infty$, $\bar{H}$ exists, where*

$$\bar{H} = \lambda \bar{G}. \quad (12)$$

In the literature, the usual notation is " $H = \lambda G$," but we prefer to reserve H and G for the stationary version. For LL, $l_i = W_i$, and TC is implied by $w < \infty$. Without TC, some of the area contributed by C_i may occur so far in the future that, as i increases, some of it is "lost" in the limit, where, possibly

$$\bar{H} < \lambda \bar{G},$$

or $\bar{H}$ may not exist. However, TC is not necessary. For a necessary and sufficient condition, see Ref. 9, p. 178. The easily proven version here is widely applicable.

The same issue arises with LL when we have *multiple visits*; that is, the same customer visits the system many times, and we define W_i as the sum of C_i 's waiting times on all visits. Let C_i depart (for the last time) at $t_i + l_i$. If w is finite but TC does not hold, it is possible to have

$$L < \lambda w.$$

The so-called *counterexamples* to Little's Law have been constructed in this manner.

We sketch a proof of $\bar{H} = \lambda \bar{G}$, which is almost identical to that for LL. We assume $f_i(t) \geq 0$. When this is not the case, split the functions into their positive and negative parts, $f_i^+(t) = \max\{f_i(t), 0\}$ and $f_i^-(t) = -\min\{f_i(t), 0\}$, $t \geq 0$. Go through the steps below for the positive parts and the negative parts separately, and combine the results.

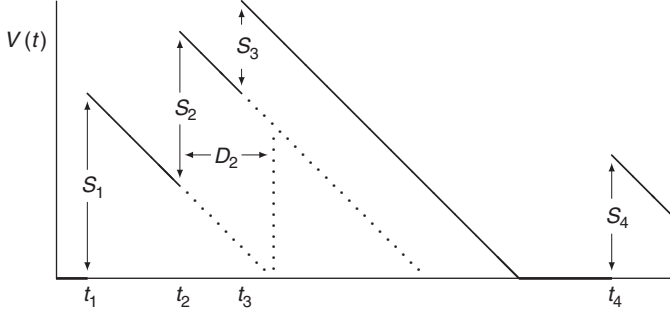


Figure 5. Work in system for a single-server queue.

When $f_i(t) \geq 0$, the bounds on the integral below are immediate:

$$\sum_{\{i: t_i + l_i \leq T\}} G_i \leq \int_0^T H(t) dt \leq \sum_{i=1}^{\Lambda(T)} G_i. \quad (13)$$

Now divide Equation (14) by T and let $T \rightarrow \infty$. The right-hand expression has limit $\lambda \bar{G}$, and the left-hand expression has $\limsup \leq \lambda \bar{G}$. To complete the proof, we obtain a lower bound on the left-hand expression in Equation (14). We already have $l_n/t_n \rightarrow 0$, and following the argument for Equation (7), we get $l_n/t_n \rightarrow 0$, as $n \rightarrow \infty$. Following the argument for Equation (8), we have the lower bound

$$\sum_{i=1}^{\Lambda\left(\frac{T}{1+\epsilon}\right)} G_i - \sum_{i=1}^m G_i \leq \sum_{\{i: t_i + l_i \leq T\}} G_i.$$

The remaining steps are the same as those for Equation (3), and we have proven Equation (13).

When $\lambda > 0$ and $\bar{G} = \infty$, and we still have $\text{TC}, \bar{H} = \infty$.

Work in System

After Little's Law, the most important applications of Equation (13) involve various representations of process $\{V(t)\}$, where $V(t)$, called *work in system* or just *work* at time $t \geq 0$, is the sum of the remaining service times of all customers in system at time t .

With the notation in the section titled "Notation, Explanation, and Applications," suppose we have a *single-server* queue with arrival times t_i and service times S_i .

$V(t)$ jumps by S_i at t_i , and where positive, decreases with slope -1 between jumps, because the remaining service time of any customer in service decreases at that rate. A typical sample path of $V(t)$ is shown as the heavy horizontal (where $V(t) = 0$) or slanted lines in Fig. 5.

When customers are served FIFO, the diagonal dotted lines show when the first two customers finish service. The vertical dotted line shows when C_2 enters service, and the time between t_2 and that line is D_2 . Similarly, D_3 (not shown) is easy to represent. The contribution of customer C_i to the area under $\{V(t)\}$ is a trapezoid that has a rectangular piece with area $D_i S_i$, and a triangular piece with area $S_i^2/2$. So for this example, $G_i = D_i S_i + S_i^2/2$, and from Equation (13)

$$E(V) = \lambda \left[\overline{DS} + \overline{S^2}/2 \right]. \quad (14)$$

To be consistent with our notation, " $E(V)$ " should be " $\bar{V}$." From the stationary version, however, it is also the mean of a stationary V . The *form* of Equation (15) does not require FIFO order of service, but only that customers are served to completion, without interruption. For example, it holds under LIFO and SJF. (While each customer still generates a trapezoid, the D_i would change. If we had LIFO in Fig. 5, D_3 would be the time between t_3 and the vertical dotted line.) Note that TC is the same as Equation (6) here, and is satisfied when $w < \infty$.

When we make additional assumptions, Equation (15) can be written in different ways. For example, when service times are i.i.d. and independent of the arrival process, the D_i and S_i are independent and

$\overline{DS} = dE(S)$; Equation (15) becomes

$$E(V) = \rho d + \lambda E(S^2)/2, \quad (15)$$

where $\rho = \lambda E(S)$. This includes FIFO and LIFO, but *not* SJF. When we also have Poisson arrivals (now an $M/G/1$ queue) and FIFO, $d = E(V)$ (from PASTA, see Ref. 10), and from Equation (16), we get

$$d = \lambda E(S^2)/2(1 - \rho), \quad (16)$$

which is the average delay in queue for this model. (As in Example 2, d is the same under LIFO; only the *argument* for Equation (17) requires FIFO.) When arrivals are not Poisson, and we can show either $d > E(V)$ or $d < E(V)$, then from Equation (16), we get a *bound* on d .

Equation (15) holds for multiserver queues, but the sample paths of $\{V(t)\}$ would change. When service times are i.i.d., Brumelle [11] used these ideas to obtain an important lower bound on d_c , the average delay in queue for a c -server system, in terms of the average delay for a corresponding fast single server; that is, a server with service times S_i/c .

Sometimes a queue operates under rules that give preferential treatment to some customers or classes of customers over others. This may permit interrupting service on one customer in order to serve another. Suppose we have *work conservation*, which means that the total time in service of every customer is rule invariant. For a *single-server* queue, this implies that the sample paths of $\{V(t)\}$, as in Fig. 5, are the same for all rules, as is $E(V)$. This fact plays an important role in the analysis of these rules. For many examples, see Ref. 3.

A RATE CONSERVATION LAW (RCL)

As with Little's Law and its extensions, the rate conservation law (RCL) by Miyazawa is of fundamental importance in queueing theory. We cite Miyazawa [12], a review article; his publications in this area go back to 1983. The theory of *level crossings* [13] is a special case.

As conceived by Miyazawa, RCL is a relation between the expected values of time-stationary and arrival-stationary quantities, as is the case for Little's Law in Equation (11). Sigman [14] showed that the sample-path version of RCL is equivalent to the corresponding version of $\overline{H} = \lambda \overline{G}$.

However, the starting points for these results (geometrically-appealing areas in one case, jumps and derivatives in the other) are quite different. In an application, one is often much easier and more intuitive to use than the other.

In this brief introduction, we present only the sample-path version of RCL.

Consider a sample path $\{Y(t, \omega), t \geq 0\}$ of some stochastic process at sample point ω . As before, we drop ω , writing $Y(t) = Y(t, \omega)$, and let $\mathbf{Y} = \{Y(t)\}$. We assume $\mathbf{Y}$ is right-differentiable for all t , and define

$$Y'(t) = \lim_{h \downarrow 0} [Y(t+h) - Y(t)]/h,$$

where we assume that $\mathbf{Y}'$ is right-continuous and has left-hand limits. Now $\mathbf{Y}$ may have discontinuities. Let "events" occur at points in an underlying point process (events could be arrivals), $0 < t_1 < t_2 < \dots$, where $r(t)$ is the number of points that occur by t , such that each discontinuity of $\mathbf{Y}$ occurs at one of these points. (We don't require $\mathbf{Y}$ to be discontinuous at every such point.) Define

$$-J_i = Y(t_i+) - Y(t_i-),$$

the size of the i th jump.

Then we have

$$\int_0^t Y'(u) du = Y(t) - Y(0) + \sum_{i=1}^{r(t)} J_i, \quad (17)$$

which simply means that the *change* in $\mathbf{Y}$ between 0 and t is the sum of the change while continuous plus the sum of the jumps.

Define the limits, when they exist, event rate λ , event average $\overline{J}$, and time average $\overline{Y'}$. Now divide Equation (18) by t and let $t \rightarrow \infty$. The following has been shown:

Theorem 3 [Sample-Path RCL]. When λ and $\bar{J}$ are finite, and $Y(t)/t \rightarrow 0$ as $t \rightarrow \infty$,

$$\bar{Y}' = \lambda \bar{J}. \quad (18)$$

(The stationary-version notation is to write $\bar{Y}' = E(Y')$ and $\bar{J} = E(J)$.)

It is easier to obtain Equation (15) via Equation (13) (it is immediate), rather than via RCL. On the other hand, for finding an expression for $P(V > v)$, for stationary work in system V , RCL is the better tool. See p. 664 of Ref. 14.

A DISTRIBUTIONAL LITTLE'S LAW

The stationary version of Little's Law raises this question: Is there a similar relation between the distributions of N and W , a *distributional* Little's Law?

In Example 2, we compared a bank under different rules of operation. Under both rules, the distribution of N is the same, but the distribution of W is different. There is no hope for a distributional law in a general setting. Nevertheless, we derive a law in the restricted setting of Haji and Newell [15]. For this result, further restricted to Poisson arrivals, with applications, see Keilson and Servi [16].

Let $t_0 = 0$ be a random time point, where $N(0) = N$ is stationary, and $\{\Lambda(t), -\infty < t < \infty\}$, has stationary increments. Number customers backward in time, where C_i occurs at $-t_i$, $0 \leq t_1 \leq t_2 \leq \dots$. So C_1 is the most recent arrival. We make the following assumptions:

- (a) Customers depart FIFO from the system,
- (b) $\{W_i\}$ is stationary, and for every i ,
- (c) W_i is independent of the arrival process after C_i arrives; in particular, it is independent of t_i , which is "how long ago" that customer arrived.

Thus, customer i is in the system at time $t_0 = 0$ if and only if $\{W_i \geq t_i\}$, and because of (a), these are the same events:

$$\{N \geq i\} = \{W_i \geq t_i\}, i \geq 1,$$

where (c) W_i and t_i are independent, and (b) $P(N \geq i) = P(W \geq t_i)$, where W is an arbitrary stationary waiting time, independent of every t_i . The event $\{W \geq t_i\}$ means that at least i arrivals occur in interval $[-W, 0]$, where W is independent of $\{\Lambda(t)\}$, and the probability of this event does not depend on the interval location. We have

Theorem 4 [A Distributional Little's Law, (DLL)]. Under assumptions (a)–(c),

$$N \text{ has the same distribution as } \Lambda(W), \quad (19)$$

where $\{\Lambda(t)\}$ and W are independent.

We now introduce some new quantities to compare this result with another. For a queueing system, let N_a and N_d be, respectively, the stationary number of customers in system *found by an arrival*, and *left behind by a departure*. It is easily shown that, in general, N_a and N_d have the same distribution. On the other hand, it is usually the case that N_a and N have different distributions and means. For simplicity, assume interarrival times $T_i > 0$ (no batches). Let $\{\Lambda_c(t)\}$ be the *customer-stationary* arrival process, which begins with an arrival at $t_0 = 0$, and $\Lambda_c(t)$ is the number of arrivals in interval $(0, t]$. When we have assumption (a), N_d is the number of arrivals during the departing customer's waiting time, and when we also have renewal arrivals (Poisson arrivals is a special case),

$$N_d \text{ has the same distribution as } \Lambda_c(W), \quad (20)$$

where $\{\Lambda_c(t)\}$ and W are independent. While Equation (21) is similar to Equation (20), it is not as useful. Usually, $E\{\Lambda_c(t)\} \neq \lambda t$ and $E(N_d) = E\{\Lambda_c(W)\} \neq \lambda E(W)$.

To illustrate, consider a single-server queue, where $t_i = 10i$ and $S_i = 9$ for all i (a $D/D/1$ queue). Stationary waiting time $W_i = 9$, a constant. As fractions of time, it is easy to see that N is either 0 or 1, with distribution

$$P(N = 0) = 0.1 \text{ and } P(N = 1) = 0.9. \quad (21)$$

We have $\lambda = 0.1$ and $E(N) = \lambda E(W)$, but $N_a = N_d = \Lambda_c(W) = 0$. For $\{\Lambda(t)\}$, t_1 (backward in time) is uniformly distributed on $(0, 10)$. C_1 is in the system at $t_0 = 0$ if and only if $\{t_1 \leq 9\}$, and $\Lambda(W) = 1$. $\Lambda(W)$ has distribution (Eq. 22); DLL holds. This is a special case of a FIFO $GI/G/1$ queue, where DLL also holds.

Confusion between $\{\Lambda(t)\}$ and $\{\Lambda_c(t)\}$ may be partly responsible for the following *incorrect* “intuitive” and “elementary” explanation of the stationary version of Little’s Law that has appeared several times in the literature: (i) $E(N_a) = E(N_d)$ and (ii) $E(N_d) = \lambda E(W)$.

While (i) is true, there are several problems with this argument. Implicit in (ii) is FIFO, as in (a) in Theorem 4. That is merely a restriction. More serious is that (ii) is often false, as in the elementary example above. Also serious is the implicit equating of N_a and N . We are left with

$$L = E(N) \neq E(N_a) = E(N_d) \neq \lambda E(W). \quad (22)$$

For the $M/G/1$ queue, both inequalities in Equation (23) are equalities, the first from PASTA (N and N_a have the same distribution); for the second, a Poisson process has stationary and independent increments ($\{\Lambda(t)\}$ and $\{\Lambda_c(t)\}$ are equivalent). In this case, the explanation is correct, but it is neither intuitive nor elementary.

Now return to DLL. When does it hold? Condition (a) holds for single-server queues when customers in queue are served FIFO. It holds for a tandem (series) arrangement of single-server queues. It holds for multiserver queues under FIFO and constant service. Looking only at the queue, there is a corresponding relation between the number of customers in queue and the delay in queue. It holds for multiserver queues under FIFO, with a general service distribution. There are other examples for priority classes in what are called *priority queues*.

Condition (c) appears to require renewal arrivals, which includes the Poisson process, but not (say) the typical arrival process at the second station of a tandem queue. There is an exception to this requirement when there are an infinite number of servers. DLL does not hold in Example 3.

The DLL is very useful when it applies, but unlike the ubiquitous LL, the conditions for it to hold are rather restrictive.

BRIEF HISTORY AND LITERATURE REVIEW

Little’s Law was believed to be true long before 1961. In p. 75 of Ref. 17, Morse noted that it holds for every model he was aware of, and challenged someone to either come up with a general proof or a counterexample. Unfortunately, this pre-1961 status inhibited the use of LL as a tool in analysis of queueing models.

We have emphasized the sample-path version because it is the easiest to understand and prove, and it holds in situations when there may be no stationary version. Viewed this way, it is a conservation law (of area). Furthermore, it is rare in an application to use the full stationary version in Equation (11). Instead, a stationary analysis is performed to determine properties of *one* of the stationary distributions sufficient to find, approximate, or make other statements about its mean. The sample-path version then gives us the other.

At first glance, Little’s formulation seems to be sample-path, but it is not, as it relies on stochastic properties to obtain limits. There also is a flaw in the formulation, as noted in Ref. 18 and elsewhere. In 1967, Jewell [19] proved LL for (classically) regenerative processes where the system empties from time to time, and a new cycle begins. Thus over a cycle, Equation (1) is exact! (Shortly thereafter, several authors either proved LL or gave intuitive explanations for systems that empty periodically.) What Jewell showed may be viewed as a sample-path result and may have provided intuition for what followed. However, his conditions are very restrictive. Stidham has the first sample-path proof [20], and in 1974 [2], a proof similar to the one here.

What is now usually called $H = \lambda G$ has a similar history. Brumelle [18] proved it in 1971 in a stochastic setting similar to Little. He also obtained the important Equation (15). Heyman and Stidham [21] has a sample-path proof in 1980 similar to the one presented here. Brumelle’s result is not the

equivalent stationary version (Eq. 11) in the section titled "Stationary Version of little's Law." See the references there.

There are too many related results to discuss in the space we have. Here are three. For $H = \lambda G$, a customer's contribution is an integral, so it accumulates gradually. This has been extended to allow "lumps" at certain times. Often, a model is simulated to estimate (say) Q or d , and direct (straightforward) estimators of each, $\hat{Q}$ or $\hat{d}$, are available. From LL, an *indirect* estimator of (say) Q is $\lambda \hat{d}$. The statistical efficiency of indirect estimators was investigated in Law [22], and by others later. Finally, there are also central-limit-theorem versions of these results.

For more on these and other results, many other references, and discussion of the literature, see Refs 9, 23, and 24.

REFERENCES

1. Little JDC. A proof of the queueing formula: $L = \lambda W$. *Oper Res* 1961;9:383–387.
2. Stidham S Jr. A last word on $L = \lambda W$. *Oper Res* 1974;22:417–421.
3. Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice-Hall; 1989.
4. Loève M. Probability theory. 3rd ed. Princeton (NJ): Van Nostrand; 1963.
5. Franken P, König D, Arndt U, *et al.* Queues and point processes. New York: John Wiley & Sons, Inc.; 1982.
6. Daley DJ, Vere-Jones D. An introduction to the theory of point processes. New York: Springer-Verlag; 1988.
7. Sigman K. Stationary marked point processes. New York: Chapman & Hall; 1995.
8. Miyazawa M, Nieuwenhuis G, Sigman K. Palm theory for random time changes. *J Appl Math Stoch Anal* 2001;14:55–74.
9. El-Taha M, Stidham S Jr. Sample-path analysis of queueing systems. Boston (MA): Kluwer Academic Publishers; 1999.
10. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
11. Brumelle S. Some inequalities for parallel-server queues. *Oper Res* 1971;19:402–413.
12. Miyazawa M. Rate conservation laws: a survey. *Queueing Syst Theor Appl* 1994;15:1–58.
13. Brill P, Posner M. Level crossings in point processes applied to queues: single-server case. *Oper Res* 1977;25:662–674.
14. Sigman K. A note on a sample-path rate conservation law and its relationship with $H = \lambda G$. *Adv Appl Probab* 1991;23:662–665.
15. Haji R, Newell GF. A relation between stationary queue and waiting time distributions. *J Appl Probab* 1971;8:617–620.
16. Keilson J, Servi LD. A distributional form of little's law. *Oper Res Lett* 1988;7:223–227.
17. Morse PM. Queues, inventories and maintenance. New York: John Wiley & Sons, Inc.; 1958.
18. Brumelle S. On the relation between customer and time averages in queues. *J Appl Probab* 1971;8:508–520.
19. Jewell WS. A simple proof of $L = \lambda W$. *Oper Res* 1967;15:1109–1116.
20. Stidham S Jr. $L = \lambda W$: a discounted analogue and a new proof. *Oper Res* 1972;20:1115–1126.
21. Heyman DP, Stidham S Jr. The relation between customer and time averages in queues. *Oper Res* 1980;28:983–994.
22. Law AM. Efficient estimators for simulated queueing systems. *Manage Sci* 1975;22:30–41.
23. Stidham S Jr, El-Taha M. Sample-path techniques in queueing theory. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Raton (FL): CRC Press; 1995. pp. 119–166.
24. Whitt W. A review of $L = \lambda W$ and extensions. *Queueing Syst* 1991;9:235–268.

LOAD-SHARING SYSTEMS

ISHA DEWAN
Indian Statistical Institute, New
Delhi, India

UTTARA V. NAIK-NIMBALKAR
University of Pune, Pune,
Maharashtra, India

In Ref. 1, the authors refer to an article in the New York Times by Glanz and Lipton [2] pertaining to the collapse of the twin towers of the World Trade Center on September 11, 2001. The World Trade Center had its weight resting mainly on interior columns and also on the pinstripe columns at the exterior. A support system at the top of the building connected these columns at the facade to the core. When the airplanes hit the top of the building, the support system at the top shifted the load on the damaged columns to the undamaged columns. Because of this load sharing, the towers did not collapse immediately after they were hit by the airplanes. This saved thousands of lives.

In Ref. 3, the authors refer to the power crisis in the city of Auckland, New Zealand. At the beginning of 1998, the city received electricity from a single supplier via four power cables. Two of them were 40 years old. One cable failed on January 20 due to extremely hot and dry conditions, the second failed on February 9. This led to increased load on the remaining cables, which too failed on February 19 and 20, leaving large parts of the city without power.

In general, in any multicomponent system, the failure of one component can affect the working of the remaining components. But the effect varies according to the load-sharing rule, which tells us how the “load” of the failed component is distributed among the other surviving components of the system. An *equal load-sharing* system is subject to a constant load. Once a component fails, the system load is spread equally among the surviving components. A common example is a parallel system where the k component

lifetimes are independent and identically distributed (i.i.d.) random variables with constant failure rate θ . When the first component fails, then each of the surviving components has failure rate $\frac{k}{k-1}\theta$, on second failure the remaining $k-2$ components have failure rate $\frac{k}{k-2}\theta$, and so on. Under the *monotone load-sharing* rule the load on the surviving components increases as other components fail, that is, the greater the number of failures, the larger is the load on the surviving components. Under *local load sharing*, the effect of the failure of a component is relatively more on certain components (usually the neighbors) and is generally less for components placed at a greater distance from the surviving component.

Sometimes the failure of the first component shortens the residual life of the remaining components. This can lead to an error in the estimation of the system reliability. In load-sharing models, the failure rate of the components depends on whether the other components in the system are functioning or not.

The earliest work on load-sharing models is due to Daniels [4] and Rosen [5]. They observed that yarns and cables in a bundle fail only when the last fiber (or wire) in the bundle breaks. A bundle of fibers can be considered as a parallel system (see also ***Parallel Configurations***) subject to a constant tensile load. After one fiber breaks yarn bundles or untwisted cables tend to spread the stress load uniformly on the remaining unbroken fibers. This is the equal load-share rule under which the load of the failed component is distributed equally among the remaining working components.

Apart from textile industry, such model arises in manufacturing where a part can be considered failed only when the entire set of welded joints that holds the part together fails. However, the failure of one or two joints can increase the stress on the remaining joints.

In the life testing of roller bearings, two test bearings are run simultaneously in a

given machine. The failure of the first bearing leads to exfoliation of material from the bearing rolling paths. This could affect the oil supply shared by the two test bearings with metallic particles. This results in the early failure of the second bearing. Sometimes the failure is not immediately detected but the vibration level increases because of the first failure. This could damage the rolling path of the second bearing.

Another illustration discussed by McCool [6] is for the ink-jet printer elements. These units contain about 400 heating elements. These elements dispense ink which is used for character formation and their failure leads to poor printing quality. One normally tests only till the first heater on each of the elements fails as the first failure affects failure of the subsequent heaters.

This phenomenon can also be seen on airplanes. A small airplane can function on one jet engine, whereas two engines are required for a big airplane. In order to increase reliability, small airplanes have two functioning engines and big airplanes have four functioning engines. The extra engine shares the load. This helps in reducing the failure rate of the engines.

In all the above examples, the failure of the first component adversely affects the system performance. However, in some cases, the failure of the first component has a beneficial effect on the system performance. For example, the detection of a bug in a software can help in the detection of other bugs. Once a critical fault in the software has been detected, it can help in finding other bugs that had not yet been undetected.

Drummond *et al.* [7] carried out a study in a vertebrate species showing that selective deaths due to food shortage result in surviving offspring receiving an increased share of an undiminished food supply. They observed littermates of the domestic rabbit *Oryctolagus cuniculus* and found that after individual baby rabbits died, the total daily milk weight obtained by the litter continued to be the same. Hence the surviving baby rabbits consumed more milk and showed greater growth.

Kim and Kvam [1] observed that such models also arise in sampling techniques. Suppose the total resources allocated toward

finding a finite set of items is fixed. Once one item is detected, resources can be redistributed to finding the remaining items.

Murthy and Nguyen [8] and Murthy and Wilson [9] studied a two-component parallel system (see also **Parallel Configurations**) where the failure of the first component leads to increased failure of the second component with probability p and does not effect the failure of the second component with probability $1 - p$. This is analogous to the Brown and Proschan (1983) model for repairable systems with imperfect repair.

Bebbington *et al.* [3] considered the general setup where the failure of the first component does not effect the hazard function of the other components immediately, but affects its conditional time to failure. They obtained nonparametric estimators of the failure rate function and the mean residual life function under this general setup.

Physicists have been studying load-sharing models under the study of fiber bundles. For some recent work, see Pride and Toussaint (2002), Bhattacharyya, Pradhan and Chakrabarti (2003) and Kang, Jo, Choi, Choi and Yoon (2007).

MODELS AND STATISTICAL INFERENCE

Here, we discuss several parametric and semiparametric models proposed for studying load-sharing systems over the last four decades, and discuss the related problems of estimation and testing of parameters.

Gross *et al.* [10] observed that two-organ subsystems in human body typically exhibit load-sharing phenomenon. If one organ fails, the surviving organ is usually subject to higher failure rate. If a patient gets his kidney removed because of some illness, then the second kidney shows a higher failure rate. However, if a kidney is removed because of an accident, then the second kidney may not exhibit an increased failure rate. The authors develop a survival distribution for such two-organ systems. They assume that if one organ fails, then the other organ has a higher failure rate. However, failure rates of both components are assumed to be constant.

Gross *et al.* [10] considered the following setup. The failure rate of each organ initially is a constant λ_0 . After the failure of one organ, the failure rate of the other organ is a constant $\lambda_1 \geq \lambda_0$. When $\lambda_1 \neq 2\lambda_0$, then the probability that the individual survives beyond time t is

$$S(t) = e^{-2\lambda_0 t} + \frac{2\lambda_0}{\lambda_1 - 2\lambda_0} (e^{-2\lambda_0 t} - e^{-\lambda_1 t}), \quad t \geq 0. \quad (1)$$

When $\lambda_1 = 2\lambda_0$, the density function of failure time is the convolution of two densities with the same parameter. Further, if $\lambda_1 = \lambda_0 = \lambda$ (say), it means that the failure of one organ does not affect the functioning of the other organ. In this case,

$$S(t) = e^{-\lambda t} (2 - e^{-\lambda t}), \quad t \geq 0, \quad \lambda > 0. \quad (2)$$

The authors obtained maximum likelihood estimators (MLEs) for the case $\lambda_1 \geq \lambda_0$. Since the MLEs cannot be expressed in a closed form, they used the Newton–Raphson method and the method of scoring to get iterative solutions. They concluded that both the methods gave similar estimates and the average number of iterations required were 3.2 for Newton–Raphson method and 4.0 for the scoring method. The method of moment estimate of λ_1 is biased (see also **Estimating Failure Rates and Hazard Functions**).

An m -out-of- k load-sharing system is considered in Refs 11–14. A general model is specified by considering arbitrary continuous distribution functions $F_1, F_2, \dots, F_k$. Before a failure, the k failure times are assumed to be i.i.d. with F_1 as the common distribution function. The first-component failure time is then the minimum of these k i.i.d. random variables. Upon the $(i-1)$ th ($i \geq 2$) component failure, conditionally given the $(i-1)$ th failure time z , the remaining $k-i+1$ working components are assumed to be i.i.d. random variables with the common distribution function

$$\frac{F_i(\cdot) - F_i(z)}{1 - F_i(z)}, \quad 2 \leq i \leq k. \quad (3)$$

The next component failure time is the minimum in the sample of the $k-i+1$ i.i.d.

random variables with distribution function given in Equation (3).

As a particular case of the above model, they modeled F_i by

$$F_i(t) = 1 - (1 - F(t))^{\alpha_i}, \quad 1 \leq i \leq k, \quad (4)$$

where F is an absolutely continuous distribution function and $\alpha_1, \alpha_2, \dots, \alpha_k$, are positive real numbers. With this model, after the $(i-1)$ th ($i \geq 2$) component failure, the hazard rate of the working components is $\alpha_i f / (1 - F)$, where f denotes the density function of F . Before any failure, the hazard rate is $\alpha_1 f / (1 - F)$.

For an exponential distribution F , the reliability of an m -out-of- k load-sharing system defined by Equations (3) and (4) is studied in Ref. 15.

The MLEs for model parameters $\alpha_1, \alpha_2, \dots$ for a known distribution F are obtained in Refs 12 and 13. Lifetimes of all the $k-m+1$ components that fail till the system fails are observed sequentially. These ordered observations are called *sequential order statistics*. Let $\{x_{i(j)}; 1 \leq j \leq k-m+1, 1 \leq i \leq n\}$ be the observed sequential failure times of components of n identical and independent systems. Then the likelihood is

$$\begin{aligned} L(\alpha_j, j = 1, \dots, k-m+1; F) &= \left(\frac{k!}{(m-1)!} \right)^n \left(\prod_{j=1}^{k-m+1} \alpha_j \right)^n \\ &\quad \left(\prod_{i=1}^n \prod_{j=1}^{k-m} (1 - F(x_{i(j)}))^{p_j} f(x_{i(j)}) \right) \\ &\quad \times \prod_{i=1}^n (1 - F(x_{i(k-m+1)}))^{p_{k-m+1}} \\ &\quad f(x_{i(k-m+1)}), \end{aligned} \quad (5)$$

where $p_j = (k-j+1)\alpha_j - (k-j)\alpha_{j+1} - 1$, $1 \leq j \leq k-m$ and $p_{k-m+1} = m\alpha_{k-m+1} - 1$.

The MLEs of $\alpha_1, \alpha_2, \dots, \alpha_{k-m+1}$ are

$$\hat{\alpha}_1 = -\frac{n}{k} (\log \prod_{i=1}^n (1 - F(x_{i(1)})))^{-1},$$

and

$$\hat{\alpha}_j = -\frac{n}{k-j+1} \left(\log \prod_{i=1}^n \frac{1-F(x_{i(j)})}{1-F(x_{i(j-1)})} \right)^{-1},$$

$$2 \leq j \leq k-m+1.$$

The following properties of the corresponding MLEs, also denoted by $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-m+1}$, are established in Ref. 13.

Theorem 1.

- (i) *The MLEs $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-m+1}$ are jointly independent.*
- (ii) *$\hat{\alpha}_j$ has an inverted gamma distribution with parameters n and $(1/(n\alpha_j))$, $1 \leq j \leq k-m+1$.*
- (iii) *$\hat{\alpha}_j \rightarrow \alpha_j$, a.s. as $n \rightarrow \infty$, $1 \leq j \leq k-m+1$.*
- (iv) *$\sqrt{n}(\hat{\alpha}_j/\alpha_j - 1)$ converges in distribution as $n \rightarrow \infty$ to a standard normal random variable, $1 \leq j \leq k-m+1$.*

For the model specified by Equations (3) and (4), three test procedures are suggested in Ref. 14 for testing the hypothesis

$$H_0 : \alpha_1 = \dots = \alpha_{k-m+1}$$

against the alternative

$$H_1 : \exists i \neq j, i, j \in \{1, 2, \dots, k-m+1\},$$

such that $\alpha_i \neq \alpha_j$.

The null hypothesis H_0 corresponds to an m -out-of- k system with i.i.d. components, that is, the failure of a component does not affect the failure of the other components. The failure rate of each component in this case remains $\alpha_1 f(t)/(1-F(t))$. The alternative H_1 corresponds to what the authors refer to as a *sequential m -out-of- k system*, where failure of a component does affect the failure rate of the surviving components.

Test I: Likelihood Ratio Test Let

$$Q = \frac{\sup_{\alpha_1=\dots=\alpha_{k-m+1}} L(\alpha_j, j=1, \dots, k-m+1; F)}{\sup_{\alpha_1, \dots, \alpha_{k-m+1}} L(\alpha_j, j=1, \dots, k-m+1; F)},$$

where $L(\alpha_j, j=1, \dots, k-m+1; F)$ is as in Equation (5). The hypothesis H_0 is rejected at level of significance α if $Q^{n(k-m+1)} \leq z_\alpha$. The authors provide values for the quantiles z_α for $\alpha = 0.01$ and $\alpha = 0.05$, for $r = 2, 3$ and 5 and $n = 1, 2, \dots, 10$. They suggest using chi-square approximations for large sample sizes.

Let $\hat{\beta}_j = n/\hat{\alpha}_j$, $1 \leq j \leq k-m+1$, where the $\hat{\alpha}_j$'s are the MLEs of α_j 's. Then, from Theorem 1, it follows that under the null hypothesis, the random variables $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{k-m+1}$ are independently and identically gamma distributed with parameters n and α_1^{-1} . The next two tests are based on this result.

Test II: Hartley's Test Let $\hat{\beta}_{(1)} \leq \hat{\beta}_{(2)} \leq \dots \leq \hat{\beta}_{(k-m+1)}$ be the order statistics corresponding to $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{k-m+1}$. The test procedure suggested is to reject H_0 , if the ratio $\hat{\beta}_{(1)}/\hat{\beta}_{(k-m+1)} \leq z_\alpha$, where z_α is determined by $P_{H_0}(\hat{\beta}_{(1)}/\hat{\beta}_{(k-m+1)} \leq z_\alpha) = \alpha$. The authors provide values of z_α for $\alpha = 0.01$ and $\alpha = 0.05$, $k-m+1 = 2, 3$, and 5 and $n = 1, 2, \dots, 10$.

Test III: The third test for the above hypothesis is based on the following statistic

$$B = \hat{\beta}_1 / \left(\sum_{j=1}^{k-m+1} \hat{\beta}_j \right),$$

which under H_0 has a Beta distribution on (0,1) with parameters n and $(k-m)n$. The hypothesis H_0 is rejected at level α if either

$$B \leq z_{\alpha/2} \text{ or } B > z_{1-\alpha/2},$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are determined by

$$P(B \leq z_{\alpha/2} | H_0) = \alpha/2 \quad \text{and}$$

$$P(B > z_{1-\alpha/2} | H_0) = \alpha/2.$$

Kim and Kvam [1] considered k -component parallel system (see also **Parallel Configurations**). Initially, the components

are identically distributed with failure rate θ . After the first failure, the failure rate of the remaining $(k-1)$ components changes to $\gamma_1\theta, \gamma_1 > 0$. After the second failure, the failure rate of remaining $k-2$ components changes to $\gamma_2\theta$, and so on. Suppose there are n such systems. Let X_{ij} denote the lifetime of the j th component in the i th parallel system. Let

$$T_{ij} = X_{ij} - X_{i,j-1}, i = 1, 2, \dots, n, j = 1, \dots, k$$

denote the time between the j th failure and $(j-1)$ th failure of the i th system. Then, the likelihood function of the i th system is

$$\begin{aligned} L_i(\theta, \underline{\gamma} \mid t_{i1}, t_{i2}, \dots, t_{ik}) \\ = k! \theta^k \prod_{j=1}^{k-1} \gamma_j \exp \left(-\theta \sum_{j=1}^k (k-j+1) \gamma_{j-1} t_{ij} \right), \\ \text{with } \gamma_0 \equiv 1. \end{aligned} \quad (6)$$

The likelihood function based on n samples is,

$$\begin{aligned} L(\theta, \underline{\gamma} \mid \underline{T}) &= (k!)^n \theta^{nk} \left(\prod_{j=1}^{k-1} \gamma_j \right)^n \\ &\times \exp \left(-\theta \sum_{i=1}^n \sum_{j=1}^k (k-j+1) \gamma_{j-1} t_{ij} \right) \end{aligned} \quad (7)$$

where

$$\underline{T} = \{t_{ij}, 1 \leq j \leq k, 1 \leq i \leq n\}, \gamma > 0.$$

Then, the MLEs $(\hat{\theta}, \hat{\gamma})$ are solutions of the equations

$$\theta = \frac{nk}{\sum_{i=1}^n \sum_{j=1}^k (k-j+1) \gamma_{j-1} t_{ij}} \quad (8)$$

and

$$\begin{aligned} \gamma_{j-1} - \frac{\sum_{i=1}^n \sum_{j=1}^k (k-j+1) \gamma_{j-1} t_{ij}}{K \sum_{i=1}^n (k-j+1) t_{ij}} = 0, \\ j = 2, 3, \dots, k. \end{aligned} \quad (9)$$

Theorem 2. As $n \rightarrow \infty$, $\sqrt{n} \{(\hat{\theta}, \hat{\underline{\gamma}})' - (\theta, \underline{\gamma})'\}$ converges to a k -parameter Gaussian distribution with mean $\underline{Q}_k$ and covariance matrix Σ , where

$$\Sigma = \begin{pmatrix} \theta^2 & -\theta \underline{\gamma}' \\ -\theta \underline{\gamma}', & D(\underline{\gamma}^2) + \underline{\gamma} \underline{\gamma}' \end{pmatrix}$$

and $D(\underline{\gamma}^2)$ is the diagonal matrix $(\gamma_1^2, \dots, \gamma_{k-1}^2)$.

If the components' lifetimes were i.i.d. exponential variables, the asymptotic variance for $\hat{\theta}$ would be θ^2/k . In case of parallel systems, this is k times larger in size.

As discussed earlier, under equal load sharing, a constant total system load is divided equally among all working components. In this case, $\gamma_i = \frac{k}{k-i}$, $i = 1, \dots, k-1$. This generates another set of i.i.d. exponential data.

Under monotone load-share rule, a component failure results in increase in the workload of other components, and hence the failure rate might increase. Suppose $1 \leq \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_k$. The conditional likelihood between the $(j-1)$ th and j th failure is given as

$$L_j(\alpha_j) = (k-j+1) \alpha_j \exp(-(k-j+1) t_j \alpha_j)$$

where $\alpha_j = \gamma_{j-1} \theta$, $j = 1, \dots, k$,

where $\alpha_j = \gamma_{j-1} \theta$, $j = 1, \dots, k$. Then $\underline{\alpha} = (\alpha_1, \dots, \alpha_k)$, with $\alpha_1 = \theta$ is isotonic if and only if γ is isotonic.

Then Kim and Kvam [1] obtained MLEs of the parameters subject to $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_k$. If one were to test $H_0 : \gamma_1 = \gamma_2 = \gamma_{k-1}$ against the alternative that all γ s are not equal, then,

one would use likelihood ratio-based techniques for hypotheses testing. One could also find confidence intervals for any combination of the failure rate θ and load-sharing parameters $\underline{\gamma}$. For large samples, the approximate $(1 - \alpha)$ confidence ellipsoid for $(\theta, \underline{\gamma})$ is

$$(\hat{\underline{\beta}} - \underline{\beta})' I_0^{-1} (\hat{\underline{\beta}} - \underline{\beta}) \leq \chi_{k, \alpha}^2. \quad (10)$$

where $\hat{\underline{\beta}}$ is the MLE of $\underline{\beta} = (\theta, \underline{\gamma})$. I_0 is the information matrix.

However, if the alternative indicates monotone load sharing $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{k-1}$, one can consider a likelihood ratio statistic where the MLEs are found using isotonic regression. In this case, under H_0 , the test statistic would have a chi-square distribution (a mixture of chi-square distributions). For details, see Ref. 1.

Extension of these likelihood-based techniques to Weibull, lognormal, and normal lifetimes would be difficult to handle analytically.

However, McCool [6] proposed a test procedure to test the hypothesis that the time to the second failure is not affected by the occurrence of the first failure. He assumes that the failure time follows a two-parameter Weibull distribution.

$$\bar{F}(t) = \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right], \quad t \geq 0. \quad (11)$$

The scale parameter η has the same units as those in which lifetime is measured. It was observed that for a given value of the scale parameter, the variance decreases with increase in the value of β .

Suppose for each system we observe the first two failure times denoted by $X_{i(1)}$ and $X_{i(2)}$, respectively. Let $Z_i = \frac{X_{i(2)}}{X_{i(1)}}$, $i = 1, 2, \dots, n$.

Then on the basis of a random sample of size n from Equation (11), the MLE of the shape parameter β is the solution of the following nonlinear equation:

$$\left(\sum_{i=1}^n t_i^{\hat{\beta}_1} \right) \left(\frac{1}{\hat{\beta}_1} + \frac{1}{n} \sum_{i=1}^n \ln t_i \right) - \sum_{i=1}^n t_i^{\hat{\beta}_1} \ln t_i = 0. \quad (12)$$

The density of the ratio Z is given by

$$f(z) = \frac{n(n-1)\beta z_j^{\beta-1}}{[1 + (n-1)z^\beta]^2}, \quad z > 0. \quad (13)$$

Then, the MLE of the shape parameter β based on a random sample of size n from Equation (13) is the solution of the following equation:

$$\frac{1}{\hat{\beta}_2} + \frac{1}{n} \sum_{i=1}^n \ln z_i - 2 \frac{(k-1)}{n} \times \sum_{i=1}^n \frac{z_i \hat{\beta}_2 \ln z_i}{[1 + (k-1)z_i \hat{\beta}_2]} = 0. \quad (14)$$

If the null hypothesis is not true then, $\hat{\beta}_2$ obtained in Equation (14) will be larger than $\hat{\beta}_1$ obtained in Equation (12) because of reduced variability.

Under H_0 , the distribution of $LS_1 = \frac{\hat{\beta}_2}{\hat{\beta}_1}$ depends only on the sample size n and number of units in each system k . Hence it is distribution free.

McCool [6] obtained the percentiles for the distribution of LS_1 for $k = 2$ and He also found the power of the test when the underlying distribution is Weibull with the same shape parameter β , but the scale parameter has been reduced by a factor c .

Kvam and Pena [16] generalized the ideas of Kim and Kvam [1] with constant failure rate θ to arbitrary failure rate $r(x)$. They considered a system with k identical components and consider the equal load-sharing rule. Until the first-component failure, the failure rate of each of the k components of the system equals $r(x)$, called the *baseline failure rate* (see also **Estimating Failure Rates and Hazard Functions**). Upon the first failure, the failure rate of the $k-1$ remaining components becomes $\gamma_1 r(x)$. After the second failure, the failure rate of the $k-2$ surviving components changes to $\gamma_2 r(x)$, and so on. The failure rate of the last remaining component is $\gamma_{k-1} r(x)$. The main aim of their paper is to estimate the baseline component failure time distribution $F(\cdot)$ and the hazard function $R(\cdot)$ corresponding to the failure rate $r(x)$. The

data consist of observations on component failure times of n independent and identical systems over an observation period $[0, \tau]$, where τ could be random. For $i = 1, 2, \dots$, let $X_{i(1)} < X_{i(2)} < \dots < \tau$ be the successive component failure times for the i th system. They use the counting process theory to obtain the estimators and establish their asymptotic properties. Let $N_i(t) = \sum_{j=1}^k I(X_{i(j)} \leq t)$ and $Y_i(t) = (k - N_i(t-))I(\tau \geq t)$. The likelihood based on the observed data $\{(N_i(t), Y_i(t), 0 \leq t \leq \tau); i = 1, 2, \dots, n\}$ is given by

$$L(R(\cdot), \gamma) = \left\{ \prod_{i=1}^n \prod_{0 \leq u \leq \tau} [Y_i(u) \gamma [N_i(u-)] dR(u)] dN_i(u) \right\} \times \exp \left\{ - \int_0^\tau Y_i(u) \gamma [N_i(u-)] dR(u) \right\}, \quad (15)$$

where the second product denotes the product integral and

$$\gamma [N_i(u)] = \sum_{j=0}^{k-1} \gamma_j I(N_i(u) = j).$$

For a fixed γ , the generalized Nelson–Aalen estimator for R is then given by

$$\hat{R}(s, \gamma) = \int_0^s \frac{J(u) dN(u)}{\sum_{i=1}^n Y_i(u) \gamma [N_i(u-)]}. \quad (16)$$

For unknown γ , the profile likelihood for γ is obtained by substituting $\hat{R}(s, \gamma)$ for R in Equation (15), which is

$$L_p(s; \gamma) = \prod_{i=1}^n \prod_{0 \leq u \leq s} \left[\frac{Y_i(u) \gamma [N_i(u-)]}{\sum_{l=1}^n [Y_l(u) \gamma [N_l(u-)]]} \right] N_i(u). \quad (17)$$

This profile likelihood is maximized with respect to γ to obtain its estimator $\hat{\gamma}$, which is then plugged in $\hat{R}(s, \gamma)$ of Equation (16) to obtain the semiparametric estimator $\hat{R}(s)$ of $R(s)$. The estimator of the survival function $\bar{F}$

is $\hat{\bar{F}}(s) = \prod_{0 \leq u \leq s} [1 - \hat{R}(du)]$. The asymptotic properties of the estimators established in Ref. 16 are stated in the following theorem. First, some notations are introduced.

$$Q_{i,j}(t) = (Y_i(t) I(N_i(t-) = j), j = 0, 1, \dots, k-1; i = 1, 2, \dots, n,$$

$$Q_i(t) = (Q_{i,0}(t), \dots, Q_{i,k-1}(t))', i = 1, 2, \dots, n.$$

$$Q(t) = \left(\sum_{i=1}^n Q_{i,0}(t), \sum_{i=1}^n Q_{i,1}(t), \dots, \sum_{i=1}^n Q_{i,k-1}(t) \right)'$$

$$q(t) = (E(Q_{1,0}(t)), E(Q_{1,1}(t)), \dots, E(Q_{1,k-1}(t)))$$

and

$$\rho(t; \gamma) = E \left[\sum_{i=1}^n \gamma * Q_i(t) \right] / E[\gamma' Q(t)] \\ = \gamma * q(t) \times (\gamma' q(t))^{-1}.$$

Theorem 3. If $\{(N_i(\cdot), Y_i(\cdot), i = 1, 2, \dots, n)\}$ are i.i.d and $\inf_{0 \leq u \leq \tau} \sum_{j=0}^{k-1} (k-j) \gamma_j P(N_1(u-) = j) > 0$, then as $n \rightarrow \infty$,

- (i) $\hat{\gamma} \rightarrow \gamma$ in probability and $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, \Sigma(\tau, \gamma))$, where $\Sigma(\tau, \gamma) = D(\gamma) \Upsilon(\tau, \gamma)^{-1} D(\gamma)$ and

$$\Upsilon(\tau, \gamma) = \int_0^\tau [D(\rho(u; \gamma)) - \rho(u; \gamma) \rho(u; \gamma)'] \gamma' q(u) dR(u),$$

with $D(\eta)$ representing a diagonal matrix with diagonal elements η .

- (ii) Further, if $\gamma' q(\tau) > 0$, then

$$\{\sqrt{n}(\hat{R}(s) - R(s)) : 0 \leq s \leq \tau\}$$

converges weakly to a zero-mean Gaussian process with variance function

$$K(s; \gamma) = \int_0^s [\gamma' q(u)]^{-1} dR(u) + \varrho(s; \gamma)' \times [\Upsilon(\tau; \gamma)]^{-1} \varrho(s; \gamma), \quad (18)$$

where $\varrho(s; \gamma) = \int_0^s \rho(u; \gamma) dR(u)$.

PROBABILISTIC WORK

Here, we summarize some probabilistic results pertaining to load-sharing systems. Daniels [4] showed that the strength of a bundle of parallel fibers is asymptotically normally distributed. It is assumed that when a free load is applied to the bundle all the fibers extend by an equal amount and for each fiber, there is a definite extension and load at which it breaks. Suppose the load on a bundle of n parallel fibers is increased gradually from 0 to its final value, at first it is distributed between the n fibers and equilibrium may be reached without any fiber giving way. However, if the load is great enough, one of the fibers may break and the load is redistributed among the remaining $n - 1$ fibers, each bearing a larger load than before and so being more likely to break. In this manner, a number of fibers may break successively until either a point of equilibrium is reached where the surviving fibers have sufficient strength to carry the final load between them, or no such point is reached and all the n threads ultimately break, in which case, the bundle is broken. The minimum load beyond which all the fibers of a bundle break is called the *strength* of the bundle.

We consider below the situation from Ref. 4 where all fibers have the same load-extension ratio and each fiber bears an equal share of the load. Let the breaking loads or strengths $X_1, \dots, X_n$ of the n fibers of the bundle be i.i.d. random variables with common probability density $f(x)$ and c.d.f. $F(x)$. Let S denote the total load applied to the bundle and let the corresponding load on each of the r out of n surviving fibers at equilibrium be s . It is assumed that r satisfies the relation

$$r = n(1 - F(s)).$$

Since each fiber bears an equal share of the load,

$$S = rs = ns[1 - F(s)],$$

and the (expected) breaking load S_b of the bundle occurs when $s = s_b$ satisfies

$$\frac{d}{ds_b} \{S_b[1 - F(S_b)]\} = 0$$

and gives S its greatest possible value.

Let $X_{(1)}, \dots, X_{(n)}$ denote the ordered strengths of the n fibers. If T denotes the strength of the bundle, then under the equal load-share assumption,

$$B_n(S) = P[T \leq S] = P[nX_{(1)} \leq S, (n-1)X_{(2)} \leq S, \dots, X_{(n)} \leq S].$$

Thus

$$B_n(S) = n! \int_0^{S/n} f(x_1) dx_1 \int_{x_1}^{S/(n-1)} f(x_2) dx_2 \cdots \times \int_{x_{n-2}}^{S/2} f(x_{n-1}) dx_{n-1} \int_{x_{n-1}}^S f(x_n) dx_n.$$

Daniels established that if $s(1 - F(s)) \rightarrow 0$ as $s \rightarrow \infty$, then as $n \rightarrow \infty$, $B_n(S)$ tends to a normal distribution with mean $S_b = ns_b[1 - F(s_b)]$ and variance $s_b^2 n F(s_b)[(1 - F(s_b))]$.

Smith [17] extended his results to a series parallel model (see also **Systems in Series** and **Parallel-Series and Series-Parallel Systems**) consisting of a long chains of bundles put together in series.

Coleman [18] found the mean time to ultimate failure of a bundle of parallel fibers when the number of fibers becomes large. Birnbaum and Saunders [19] derived the life-time distributions of the materials. Phoenix [20] showed that the system failure is asymptotically normally distributed as the number of components become large. All above results are derived for the equal load-sharing model.

Lynch [21] characterized some relationships between the failure rate and the load-share rule. Durham and Lynch [22] showed that when components have independent exponential lifetimes, the system lifetime has a threshold representation of the form $\frac{S}{\theta}$, where S and θ are independent random variables with S being a gamma variable with scale 1 and varying shape parameter, and θ is interpreted as the random environmental stress on the system with a given density.

Harlow *et al.* [23] justified the use of a Weibull distribution for the strength of a single fiber. They also showed that a Weibull distribution with different scale and shape parameters is appropriate for the lower tail distribution of a parallel system subject to some load-sharing rule. They have also considered a recursive formula for computing the system life distribution under monotone load-share rule. But the formula is very complicated when there are a large number of components in the system. A new algorithm for studying the system distribution is given in Ref. 22. This algorithm is based on graphical techniques. It also gives an insight into the effect of the load-sharing rule on the system lifetime.

Schechner [24] considered stochastic characteristics of a system under the linear breakdown rule, that is, the failure rate of a component at any time t is a linear function of the load applied at time t . He showed that if the load rule is nondecreasing, then the failure time has the increasing failure rate (IFR) property. Further, if components have independent exponential lifetimes, then the system lifetime has an asymptotically normal distribution when the number of components is large.

Tierney [25] considered a composite material consisting of a set of stiff parallel fibers embedded in a flexible bonding matrix. The time of composite failure depends on the load applied and the properties of individual fibers. He derived two approximations to the distribution of time to failure under local load sharing. In the first case, the entire distribution is approximated by uniformly small absolute errors. In the second case, the lower tail is approximated by a small relative error. For more probabilistic results under local load sharing see Harlow and Phoenix (1982), Tierney (1982) and Kuo and Phoenix (1987).

A dynamic modeling approach for the reliability of a coherent system, which accommodates the fact that the failure of a set of components may result in a change in the structure function and change the distribution function of the lifetimes of the surviving components, is discussed in Hollander and Pena (1995).

DATA SETS

In this section, we describe some data sets from the literature useful for illustration of load-share models.

Example 1. Data on 50 male and 50 female litters, each of three rats are reported in Ref. 26. One rat in each litter was drug-treated and the other two served as control animals. The records are either on the week of tumor appearance or the week of death. The death of a littermate may affect the lifetime of the surviving animals. If one deletes the litters with tumor deaths and also the litters in which either both the control animals were sacrificed or died at the end of 104 weeks, the time at which the study was either ended or censored, there are 48 male litters and 22 female litters whose exact death times are known. In case of the male litters, a nonparametric test procedure proposed in Ref. 27 does not reject the null hypothesis H_0 that the first death does not affect the second death, whereas in case of the female litters, it rejects H_0 even at 1% level of significance in favor of the alternative that death of a littermate increases the residual life of the surviving mate.

Example 2. McCool [6] has reported data on ball bearings. The data consist of first and second failure times of bearings on a given unit (system) for six such units. Based on his test (LS_1) described above in the section titled "Models And Statistical Inference," he concludes that there was an influence of the first failure on the second failure. The same conclusion is reached in Ref. 27.

Example 3. In Ref. 16, the authors have discussed life testing of plasma display devices (PDPs). PDP degradation is measured in luminosity and failure is declared when luminosity decreases below a given threshold. The measurements are recorded by sensors spaced evenly across the surface of the PDP. The aim was to determine whether sudden pixel failures indicated by one sensor would affect the degradation at other parts of the PDP. The number of sensors can be considered to be equivalent to the number of components. The authors present simulated

data with sample size $n = 20$, number of sensors $k = 3$ and lifetimes for the three sensors for each of the 20 devices. They conclude that there is a slight indication of an increased failure rate after the first observed failure.

Example 4. Data on failure times of ventilating tubes inserted in ears of children with chronic otitis media with effusion (OME) are reported in Ref. 28. Otitis media or inflammation of the middle ear is a common childhood illness, the most common form being OME. The authors carried out a study to determine possible effect of a medical treatment in the prolonging of tube life. Censored survival times of tubes are observed for ears with treatment and ears without the treatment (controls) forming two samples. For comparing the two samples, they propose a test that incorporates the fact that the survival times of tubes in the left and right ears of the same child may not be independent. Failure of the tube in one ear may increase the failure rate of the tube in the other ear. The data consist of censored failure times of tubes in each ear for 41 children with the treatment and 39 without the treatment after placement of the tubes.

Example 5. Reliability Edge Home [29] reports a data set obtained from a life test on 18 parallel systems consisting of two motors operating continuously. For each system, the failure times of both the motors along with their identification labels are given.

CONCLUSIONS AND FUTURE WORK

Herein, we have reviewed the existing literature. The emphasis has been on the statistical models proposed in the literature and associated inference problems. There are several papers available on the probabilistic aspects of load-sharing systems. All related references have been listed for further reading.

The authors along with Deshpande in Ref. 27 have studied a model for k -out-of- m load-share systems that incorporates the changes in the performance of the surviving

components due to the failure of a component in the spirit of dynamic reliability systems. These models, for various choices of the initial distribution function F , give us families of distributions that incorporate load-sharing ideas. We have considered two-component parallel load-share systems and some observations, schemes, and identifiability issues under them. We have constructed a general semiparametric bivariate family of distributions that explicitly models this phenomenon through proportional conditional hazards, and have given estimates for the constant of proportionality. We have proposed a nonparametric test for the hypothesis that the failures take place independently versus first failure affects the failure rates of the surviving components, which may be used with any continuous distribution for the component lifetimes and have obtained estimates of the power of the test, and we observe that it is quite high even for moderately distant alternatives. We are also looking at nonparametric estimation of the hazard rates of the proposed model in the presence of covariates and at other nonparametric test procedures.

Acknowledgments

We thank the referees for their suggestions.

REFERENCES

1. Kim H, Kvam PH. Reliability estimation based on system data with an unknown load share rule. *Lifetime Data Anal* 2004;10:83–94.
2. Glanz B, Lipton E. The height of ambition: in the epic story of how the trade towers rose, their fall was foretold. *New York Times* 2002;8:32–63.
3. Bebbington M, Lai CD, Zitikis R. Reliability of modules with load sharing components. *J Appl Math Decision Sci* 2007. article id 43565. pp. 18 DOI:10.1155/2007/43565.
4. Daniels HE. The statistical theory of the strength of bundles of threads. I *Proc R Soc Lond A* 1945;183:404–435.
5. Rosen BW. Tensile failure of fibrous composites. *AIAA J* 1964;2:1985–1991.
6. McCool JI. Testing for dependency of failure times in life testing. *Technometrics* 2006;48:41–48.

7. Drummond H, Vazquez E, Sanchez-Colon S, *et al.* Competition for milk in the domestic rabbit: survivors benefit from littermate deaths. *Ethology* 2000;106:511–526.
8. Murthy DNP, Nguyen DG. Study of two component system with failure interaction. *Naval Res Logist Quarterly* 1985;32:239–247.
9. Murthy DNP, Wilson RJ. Parameter estimation in multi-component systems with failure interaction. *Appl Stoch Models Data Anal* 1994;10:47–60.
10. Gross AJ, Clark VA, Liu V. Estimation of survival parameters when one of two organs must function for survival. *Biometrics* 1971;27:369–377.
11. Kamps U. A concept of generalized order statistics, Stuttgart: Teubner; 1995.
12. Cramer E, Kamps U. Sequential order statistics and k -out-of- n systems with sequentially adjusted failure rates. *Ann Inst Statist Math* 1996;48:535–549.
13. Cramer E, Kamps U. Maximum likelihood estimation with different k -out-of- n systems. In: Kahle W, von Collani E, Franz J, *et al.* eds. *Advances in stochastic models for reliability, quality and safety*. Boston: Birkhäuser; 1998. pp. 101–111.
14. Cramer E, Kamps U. Sequential k -out-of- n systems. In: Balakrishnan N, Rao CR, editors. *Handbook of statistics*, North Holland: Elsevier Science; 2001. Vol. 20. pp. 301–372.
15. Scheuer EM. Reliability of an m -out-of- n system when component failure induces higher failure rates in survivors. *IEEE Trans Reliability* 1988;37:73–74.
16. Kvam PH, Pena EA. Estimating load sharing properties in a dynamic reliability system. *J Am Statist Assoc* 2005;100:263–272.
17. Smith RL. The asymptotic distribution of the strength of a series-parallel system with equal load sharing. *Ann Probab* 1982;10:137–171.
18. Coleman BD. Statistics and time dependence of mathematical breakdowns in fibres. *J Appl Phys* 1958;29:968–983.
19. Birnbaum ZW, Saunders SC. A statistical model for life-length of materials. *J Am Statist Assoc* 1958;53:151–160.
20. Phoenix SL. The asymptotic time to failure of a mechanical system of parallel members. *Siam J Appl Math* 1978;34:227–246.
21. Lynch JD. On the joint distribution of component failures for monotone load-sharing systems. *J Statist Plann Inference* 78 1999;78:13–21.
22. Durham SD, Lynch JD. A threshold representation for the strength distribution of a complex load sharing system. *J Statist Plann Inference* 2000;83:25–46.
23. Harlow DG, Smith RL, Taylor HM. Lower tail analysis of the distribution of the strength of load sharing systems. *J Appl Probab* 1983;20:358–367.
24. Schechner Z. A load sharing model: the linear breakdown rule. *Naval Res Logist Quarterly* 1984;31:137–144.
25. Tierney L. A probability model for the time to fatigue failure of a fibrous composite with local load sharing. *Stoch Proc Appl* 1984;18:139–152.
26. Mantel N, Bohidar NR, Ciminera JL. Mantel-Haenszel analyses of litter-matched time to response data, with modifications for recovery of interlitter information. *Cancer Res* 1977;37:3863–3868.
27. Deshpande JV, Dewan I, Naik-Nimbalkar UV. Family of distributions to model k -out-of- m :G load sharing systems. 2010;140(6):1441–1451. Available at dx.doi.org/10.1016/j.jspi.2009.12.005
28. Le CT, Lindgren BR. Duration of ventilating tubes: a test for comparing two clustered samples of censored data. *Biometrics* 1996;52:328–334.
29. Reliability Edge Home (2003). Quarter 2, {3}, (www.reliasoft.com)

FURTHER READING

- Beutner E. Nonparametric inference for sequential k -out-of- n systems. *Ann Inst Stat Math* 2008;60:605–626.
- Bhattacharyya P, Pradhan S, Chakrabarti BK. Phase transition in fiber bundle models with recursive dynamics. *Phys Rev* 2003;57:046–122.
- Brown M, Proschan F. Imperfect Repair. *J Appl Probab* 1983;20:851–859.
- Chong EKP, Ramadge PJ. Optimal load sharing in soft real time systems using likelihood ratio. *J Optim Th Appl* 1994;82:23–48.
- Harlow DG, Phoenix SL. Probability distributions for the strength of fibrous materials under local load sharing I: two-level failure and edge effects. *Adv App Probab* 1982;14:68–94.
- Harlow DG, Phoenix SL. The chain of bundles probability model for the strength of fibrous materials I: analysis and conjectures. *J Composite Mater* 1978;12:195–214.

- Hollander M, Pena E. Dynamic reliability models with conditional proportional hazards. *Lifetime Data Anal* 1995;1:377–401.
- Kang H, Jo J, Choi MY, *et al.* Noise effects on the health status in a dynamic failure model for living organisms. *J Phys* 2007;40:3319–3328.
- Kuo CC, Phoenix SL. Recursions and limit theorems for the strength and lifetime distributions of a fibrous composite. *J Appl Probab* 1987;24:137–159.
- Kvam PH, Pena EA. Estimating load sharing properties in a dynamic reliability system. <http://www.stat.sc.edu/~pena/TechReports/KvamPena2003.pdf> 2003.
- Lee S, Durham S, Lynch J. On the calculation of the reliability of general load sharing systems. *J Appl Probab* 1995;32:777–792.
- Lin H, Chen K, Wang R. A multivariate exponential shared-load model. *IEEE Trans Reliability* 1993;42:165–171.
- Liu H. Reliability of a load-sharing k -out-of- n :G system: non i.i.d. components with arbitrary distributions. *IEEE Trans Reliability* 1998;47:279–284.
- Phoenix SL. The asymptotic distribution for the time to failure of a fiber bundle. *Adv Appl Probab* 1979;11:153–187.
- Phoenix SL, Taylor HM. The asymptotic strength distribution of a general fiber bundle. *Adv Appl Probab* 1973;5:200–216.
- Pride SR, Toussaint R. Thermodynamics of fiber bundles. *Phys A* 2002;312:159–171.
- Tierney L. Asymptotic bounds on the time to fatigue failure of bundles of fibers under local load sharing. *Adv Appl Probab* 1982;14:95–121.

LOCATION (HOTELLING) GAMES AND APPLICATIONS

STEFFEN BRENNER
Department of International
Economics and Management,
Copenhagen Business School,
Frederiksberg, Denmark

INTRODUCTION

The location of a production facility or a retail outlet may frequently be an important determinant of economic success, for example, because resources or customers are within easy reach, or because transportation to and from the facility is efficient to organize. Even though most of economic research abstracts from the exact location of economic activities, during the past decades, the analysis of spatial location patterns has become a well-developed research stream transcending into fields of regional economics, industrial organization, and international trade. One of the dominating paradigms of studying spatial competition was provided by Hotelling's [1] pioneering study.

Hotelling's original model contains two firms that plan to set up shop in a street of finite length. His main proposition is that firms agglomerate in the center of the street. Interestingly, pointing out an error in the argument regarding the equilibrium and slightly modifying the model led to the opposite result of extreme differentiation [2]. Also in reality, we observe agglomeration and differentiation. The former occurs, for example, in the form of shopping malls where retail businesses benefit from economies of local concentration. In contrast, differentiation tends to prevail when it comes to supermarkets that typically do not to operate next to each other. Hence, differentiation patterns seem to depend on further determinants upon which we will elaborate in this article.

Competition in Hotelling games is local, that is, firms compete for customers with their direct neighbors along the spatial domain. Commonly, the equilibrium results of Hotelling games are driven by a basic trade-off between two opposite effects of relocation: the positive (short-run) demand effect, which leads firms to locate close to their direct neighbors in order to attract a high demand, and the negative (long-run) price effect, which results from the increased local competition. The relative strength of those effects decides upon which equilibrium pattern is established: either the maximum, the minimum, or an in-between differentiation solution (given an equilibrium solution exists).

Hotelling's model was reviewed and modified many times such that, now, a huge body of literature exists from which we can draw. Reviewing the literature shows that the relative strength of the centripetal and the centrifugal forces in the model depends on a variety of features of the considered market, in particular on how the price competition is set up, demand elasticity, incentives to collocate, distribution of customers, uncertainty, and number of firms and dimensions. Note that the model of spatial competition easily mutates into a model of product differentiation. Instead of interpreting the domain of the model as space, we consider it as a dimension in the product characteristics space. For example, the product ice cream may differ with respect to the fat content. Varieties of ice cream may range from low fat to very fat, and just as a firm may choose an address on the street for its shop, it may choose the exact fat content in its ice cream.

The organization of the survey is as follows. In the section titled "Location Models with a Price Setting Stage", the reference model will be presented. The section titled "Determinants of Differentiation" surveys the literature, providing the reader with a discussion of the determinants of differentiation. The final section concludes the article.

PURE LOCATION GAMES

Most of the models in the spirit of Ref. 1 allow the players to make two choices: the location from which products are sold, and the price these firms charge for them. However, a series of influential papers have abstracted away from the choice of the price, and looked at whether location equilibria exist, and how they can be characterized when the location has no impact on the pricing. In a simple interval model with two firms, under general conditions, one would find that firms choose the same locations in equilibrium. Eaton and Lipsey [3] show that this pattern is not supported anymore when the number of firms is higher than two, and when other restrictive assumptions are abandoned. Eaton and Lipsey [3] furthermore conjecture that no pure Nash equilibrium in locations exists when the number of firms is at least three. Shaked [4] actually shows that such equilibria, indeed, do not exist in the plane if the number of firms equals three. Okabe and Aoyagi [5] analytically show that there is a marked difference in the location patterns if one compares one- and two-dimensional spaces. While under some conditions, one-dimensional domains are associated with a low degree of spatial differentiation, the distance between the players' locations substantially increases in the two-dimensional space.

LOCATION MODELS WITH A PRICE-SETTING STAGE

As mentioned already, one may interpret the model along a spatial or a product characteristics dimension. In order to avoid confusion, we will henceforth stick to the interpretation on the domain of product differentiation. There are n firms producing a good at zero marginal cost. The products are horizontally differentiated. We consider a domain of product characteristics, which is projected onto the unit interval, that is, each value from the $[0, 1]$ -interval represents a realization of an ordered set of product characteristics. In some cases we allow for a unit circumference representation. There is an infinite number of

customers i whose preferences for the product characteristics w_i are distributed along the unit interval with distribution $F(w_i)$.

The model is a two-stage game of complete information. In the first stage, firms $i = 1, \dots, n$ choose product characteristics $\bar{x} = (x_1, x_2, \dots, x_n) \in [0, 1]^n$ simultaneously before posting prices $\bar{p} = (p_1, p_2, \dots, p_n) \in R^n$ in the second stage of the game. This reflects the idea that producing a product requires a long-term commitment while setting the price can be done instantaneously. Without loss of generality, we assume $0 \leq x_1 \leq \dots \leq x_n \leq 1$. Finally, customers decide whether to buy and if so which product to buy. This decision is determined by the preference w_i of the customer, the available products, and their respective prices according to the following indirect utility function

$$u(k_j, w_j, p_i, x_i) = k_j - p_i - d(x_i - w_j),$$

where p_i is the price charged for the product of firm i , $d()$ represents any distance function, which is monotonously increasing. This function measures the disutility, which is related to the difference between the amount of the characteristics of the most preferred product w_j and the considered product x_j . These disutility costs are referred to as *transport costs* in this article. An outside option is accounted for by reservation price $k_j > 0.5$. We assume that the number assigned to a firm corresponds to its rank in the spatial ordering, that is, firm 1 is at the extreme left and firm n is at the extreme right. We focus on Nash equilibria as the only equilibrium concept.

Let us derive the equilibria in the simple case when the number of firms is two and transport costs are quadratic. That is, $d()$ is a quadratic function [2]. We first establish the equilibrium of the second stage of the game where prices are chosen. One typically starts by computing the demand for each firm by determining the location $\hat{w}$ of the customer who is indifferent to buying from either firm. Specifically, from $u(k, \hat{w}, p_1, x_1) = u(k, \hat{w}, p_2, x_2)$ we obtain

$$p_1 + (x_1 - \hat{w})^2 = p_2 + (x_2 - \hat{w})^2,$$

and after some transformations, $\hat{w} = \frac{p_2 - p_1}{2(x_2 - x_1)} + \frac{x_1 + x_2}{2}$ yields. All the demand

located to the left of $\hat{w}$ is served by firm 1, and the remaining demand is served by firm 2. To be more exact, since $\hat{w}$ may lie outside the unit interval, the demand for firm 1 is equal to $D_1 = \max\{0, \min\{\hat{w}, 1\}\}$. For simplicity, we assume that $\hat{w}$ is inside the interval's boundaries. Taking the first derivative of the profit functions of both firms with respect to the firm's own price and solving for it yields the following price-response functions:

$$p_1(p_2) = \frac{p_2 + (x_2 - x_1)(x_1 + x_2)}{2}$$

$$p_2(p_1) = \frac{p_1 + (x_2 - x_1)(2 - x_1 - x_2)}{2}$$

From these two equations, we can easily infer the prices in equilibrium by inserting one expression into the other one.

$$p_1^* = \frac{(x_2 - x_1)(2 + x_1 + x_2)}{3}$$

$$p_2^* = \frac{(x_2 - x_1)(4 - x_1 - x_2)}{3}.$$

Having solved the second stage of the game, we may now turn to the first stage. Anticipating the price setting behavior just derived, we are now ready to compute the best response function regarding locations by computing the first-order condition on profits $\pi_1(x_1, x_2)$ as follows:

$$\begin{aligned} & \frac{\partial \pi_1(x_1, x_2)}{\partial x_1} \\ &= \frac{\partial [p_1^*(x_1, x_2) \hat{w}(x_1, x_2, p_1^*(x_1, x_2), p_2^*(x_1, x_2))]}{\partial x_1}, \\ & \frac{\partial \pi_2(x_1, x_2)}{\partial x_2} \\ &= \frac{\partial [p_2^*(x_1, x_2)(1 - \hat{w}(x_1, x_2, p_1^*(x_1, x_2), p_2^*(x_1, x_2)))]}{\partial x_2}. \end{aligned}$$

With $\hat{w}(x_1, x_2, p_1^*(x_1, x_2), p_2^*(x_1, x_2)) = \frac{2+x_1+x_2}{6}$, we obtain

$$\frac{\partial \pi_1(x_1, x_2)}{\partial x_1} = -\frac{(2 + 3x_1 - x_2)(2 + x_1 + x_2)}{18},$$

$$\frac{\partial \pi_2(x_1, x_2)}{\partial x_2} = \frac{(4 + x_1 - 3x_2)(4 - x_1 - x_2)}{18}.$$

It is easy to see that the first derivative is negative for firm 1 and positive for firm 2, provided that $x_i \in [0, 1]$. Hence, irrespective of the location of the other player, firm 1 would always benefit from moving to the left, and firm two would benefit from moving to the right. As a result, $x_1^* = 0$ and $x_2^* = 1$. At the equilibrium prices, we obtain $p_1^* = p_2^* = 1$.

DETERMINANTS OF DIFFERENTIATION

As we will see below, the considered models allow for a great variety of equilibrium outcomes in prices and locations, which range from minimum to maximum differentiation, and from soft to tough price competition. We show how differentiation is affected by a variety of parameters of the market structure game. We start with the determinants of price competition. Note that for a subset of model parameters, the games are plagued with the nonexistence of equilibria.

Price Competition

The competition in prices is affected by two model features: the transportation costs and the general pricing strategy of the firms. Let us first concentrate on the impact of transportation costs on price competition. The specific way as to how these costs are modeled plays a central role in research on product differentiation. Furthermore, existence of price equilibrium and thus the existence of subgame perfect equilibrium for the whole game depend on transport costs. Transport costs influence the price competition between two firms by the number of customers, which a firm is able to withdraw from (lose to) its neighbor by decreasing (increasing) its price by one unit. We refer to this concept as the *degree of price competition*. More formally, we use the first derivative of the demand function with respect to the price as a measure for the degree of price competition.

Table 1 lists the papers on duopoly Hotelling models with uniformly distributed customers and inelastic demand across different specification of the transport costs. The first five articles are devoted to the interval model and the last two papers consider the circumference domain. The

Table 1. Studies on Horizontal Product Differentiation and Transport Costs

Article	Domain	Cost Function	Existence of Price Equilibrium	Equilibrium Locations	Degree of Price Competition
Hotelling [1]	Interval	$d(x) = ax$ $a > 0$	No	—	$\frac{1}{2a}$
d'Aspremont <i>et al.</i> [2]	Interval	$d(x) = bx^2$	Yes	Maximum differentiation	$\frac{1}{2bx}$
Economides [6]	Interval	$d(x) = bx^\alpha$ $b > 0$ $1 \leq \alpha \leq 2$	If $1.26 < \alpha < \frac{5}{3}$ If $\frac{5}{3} < \alpha \leq 2$ Not otherwise	Interior Maximum differentiation —	$\frac{1}{ab(y_1^{\alpha-1} + y_2^{\alpha-1})}$ where $y_i =$ $ D_1(p) - x_i $
Gabszewicz and Thisse [7], Anderson [8]	Interval	$d(x) =$ $ax + bx^2$	No	—	$\frac{1}{2a+2bx}$
Katz [9]	Circular	$d(x) = ax$ $a > 0$	No	—	$\frac{1}{a}$
de Frutos <i>et al.</i> [10]	Circular	$d(x) = ax +$ $bx^2, a \geq 0$ $b \in \mathbb{R}$	If $a = 0$ or if $a = -b$ Not otherwise	Maximum differentiation Maximum differentiation —	$\frac{1}{a+bx}$

fourth column indicates whether a price equilibrium exists for this class of cost functions. In the class of interval models, an equilibrium on the price setting stage only exists for quadratic and a subset of convex nonlinear cost functions considered by Economides [6]. However, convexity is not sufficient to ensure the existence of the price equilibrium (see fifth and sixth row). Rather, it is the non-quasi-concavity of the profit function that prevents the equilibrium existence. One may suspect that the degree of price competition is related to the existence of a price equilibrium. Both cases, which provide the existence of an equilibrium, are associated with an infinitely high degree of price competition if firms are very close (see last column of Table 1). In contrast, in the setup examined by Gabszewicz and Thisse [7] and Anderson [8] price competition has an upper bound even if firms locate at the same position. However, as proved by Anderson [8], configurations in which firms are located very closely and in which firms are located

very distant imply an equilibrium. For a region in-between, the equilibrium breaks down. Hence, there is no clear relationship between the degree of price competition and the existence of a price equilibrium for the interval model.

Unfortunately, previous analysis is restricted to rather simple functional forms of the transport costs. Although it would be desirable to have some theorems on the conditions of nonexistence of the price equilibrium, establishing them represents a hopeless task ([7], p. 167). Thus, for a great variety of transport functions no price equilibrium and hence, no subgame perfect equilibrium exist.

Furthermore, the degree of price competition is not linked to the existence of price equilibria via a simple relationship. In the circumference model, the price equilibrium does not exist in the case of linear transport costs [9] but exists if costs are linear quadratic [10]. Hence, both convex and concave functions may allow price equilibria.

Concentrating on the cases where equilibrium existence is of no concern, we see that the degree of price competition and the pattern of differentiation are related in the way intuitively expected [6]. Hence, given that a price equilibrium exists for the duopoly Hotelling model with uniformly distributed consumer preferences on the unit interval, then the higher the degree of price competition the stronger is the strategic differentiation. It is worth mentioning that differentiation does not range from minimum to maximum differentiation but rather from in-between to maximum differentiation. From this result, minimum differentiation does not seem to be a likely outcome. We will see below that minimum differentiation can still be obtained by relaxing other model parameters. For the circular model, de Frutos *et al.* [10] show that the principle of maximum differentiation is dominating because firms would always like to differentiate unless linear transport costs are assumed. In the previously mentioned papers, transport costs were taken to be exogenous. However, several examples like PCs designed to fulfill a wide range of functions or cosmetic products (“all in one”) suggest that firms try to reduce the disutility to consumers incurred by the distance of their preferred good and the offered good measured in the space of product characteristics. von Ungern-Sternberg [11] and Hendel and de Figueiredo [12] tackle this problem for the circular domain. In their model, firms first choose to enter and then choose a location. Subsequently, they set transport costs and prices simultaneously [11] or sequentially [12]. If the transport costs are close to zero, the general-purpose of the product is said to be high. Otherwise, the product design is said to be focused. Transport costs are assumed to be linear with a variable slope parameter. In the simultaneous choice model, firms set transport costs at a rather low level (lower than socially optimal), which increases price competition. In the sequential setup of Ref. 12, however, the choice of the focus introduces a strategic effect on pricing. The impact strategic focusing has on price competition is not straightforward. In the duopoly case where changing the focus is

costless, firms have an incentive to increase their focus in order to relax price competition. For the case $n > 2$ this relation does not hold anymore. Here firms are interacting with two different neighbors, resulting in externalities of any change in prices. Consequently, firms increase the general purpose of their products and set prices at marginal costs. Given that entry only occurs at fixed cost there is no equilibrium with more than two firms. Hence, despite free entry, in equilibrium there are only two firms in the market. However, if changing the focus is associated with costs the results change for $n > 2$. In this case, the focus will be increased by the firms leading to positive prices and hence, possible oligopoly equilibria.

A final note is due regarding the plausibility of different functional forms of transport costs. Lancaster ([13], Chapter 2) argues that in contrast to quadratic costs, linear transport costs are inconsistent with plausibly shaped indifference curves.

Pricing Strategies

The lion’s share of papers on Hotelling models assumes mill pricing, that is, firms sell the products at an fob (free on board) price, which is not related to the location of the consumers. Another possibility would be to charge spatially discriminatory prices. Here, for example, firms charge lower prices from customers who are more distant. Applied to the domain of product differentiation there exist two possible interpretations. The obvious one is that firms compensate the disutility customers incur from buying a product whose characteristics are not aligned to their preferences by adjusting the price accordingly. That is, firms would charge different prices depending on the location of customers in the characteristics space. Another interpretation is that firms redesign the goods to be sold (which is assumed to be done at zero cost), pay the cost of transport, and charge a uniform total price. Both price strategies can be observed in reality: discriminatory and nondiscriminatory price-setting. An industry where the two forms occur simultaneously is the book retail industry. Books may be either bought in the book shop or by

one of the numerous mail-order or internet book retailers at a similar compound price (including transportation).

The impact of different price regimes on spatial competition was considered in Refs 14–19. From the competition point of view, the impact of the choice between mill pricing and discriminatory pricing is not straightforward. First, the opportunity to discriminate on prices increases the firm's flexibility. A number of articles show the profit enhancing effect of monopolistic price discrimination [20]. However, in the short-run, firms would prefer mill pricing. This is because mill pricing offers the opportunity to commit to a set of prices while in discriminatory pricing it is possible to cut prices at each location separately. Gabszewicz and Thisse [21] point out that under discriminatory pricing, there is a separate Bertrand price game at each location of the interval. Hence, price competition is tougher under discriminatory pricing resulting in lower equilibrium prices [16]. Moreover, it was shown for a general set of conditions that if prices and price regimes are chosen simultaneously, discrimination occurs as an equilibrium. Additionally, under more rigid assumptions the sequential subgame of first choosing the price regime and then posting a price simultaneously delivers the same result [16].

Tabuchi [18] analyzes the case in which marginal costs of production differ between discriminatory and mill pricing. Mill pricing yields as an equilibrium only if marginal costs of discriminatory pricing are much larger than marginal costs of mill pricing. However, taking into account the long-run effect of discriminatory pricing on entry, fewer firms enter the market. The higher concentration of firms is accompanied by higher profits [17]. Consequently, discriminatory pricing can be applied as an effective deterrent of entry. However, if entry is blockaded exogenously, incumbent firms would generate greater profits using mill pricing.

Zhang [19] examines the effect of price-matching policies on the Hotelling game. If a firm follows a price-matching policy, it is committed to match or beat the prices of its competitors. The game has three stages:

(i) location choice, (ii) decision about the price-matching policy, and (iii) setting the price. Here again, an instrument that seems very competitive leads to a softer price competition via the strategic effect of entry deterrence. Consequently, even though we assume quadratic transport costs, in equilibrium, the locations follow the minimum differentiation principle.

Elasticity of Demand

The standard Hotelling model makes the rather unrealistic assumption that demand is inelastic. Relaxing this assumption may have two effects. First, equilibrium prices can be expected to decrease if demand elasticity is sufficiently high because then it is worthwhile for the firms to attract new demand by decreasing their prices. Second, price competition between the firms may be cooled because the firms' focus will become increasingly local. Since transport costs are paid by the customers, the real prices are lower for customers close to the firm, who are easily attracted even if demand elasticity is high. Consequently, this would lower the incentives to differentiate to the maximum.

Several papers contribute to model generalizations in this respect. Smithies ([22], p. 432) investigated a setting in which the consumers are individually endowed with a linear inverse demand function. He argued informally that firms would locate "closer to the center than to the quartiles." Customers in the hinterland of each firm are not necessarily attracted by the respective firms if demand is elastic. Salop [23] concentrates on the price stage given symmetric locations of firms on the circumference and shows the existence of the price equilibrium if there is an outside good. Böckem [24] models a continuum of consumers whose reservation price $k \in [0, 1]$ is distributed uniformly over the interval. Assuming a duopoly with quadratic transport costs she showed that a price equilibrium exists, and further that firms locate between maximum and minimum differentiation. The numeric solution of equilibrium locations is $[0.272, 0.728]$. Indeed, it is the effect that it pays for the firms to be close to the demand, which attracts them toward the market center.

Hinloopen and Marrewijk [25] elaborate on a similar setup where transport costs are linear and the reservation price is constant across consumers. Given that the reservation price is sufficiently high, the original Hotelling result holds in which no price equilibrium exists. If the reservation price is low, firms become local monopolists, and a continuum of equilibrium locations yield including maximum and intermediate differentiation. However, reservation prices in-between imply symmetric equilibrium locations where the distance between the firms is between one-fourth and half of the market. For a range of reservation prices there exists a negative relationship between this value and the amount of differentiation. Hence, given a price equilibrium exists for the duopoly Hotelling model with uniformly distributed consumer preferences on the unit interval then the higher the elasticity of demand the less firms will differentiate.

Collusion

The games usually considered in spatial competition assume noncooperative behavior of firms. However, through frequent interaction firms might partially build up a cooperative attitude toward their competitors resulting in collusive agreements. Firms may collude on both locations and prices. In the duopoly examined by d'Aspremont *et al.* [2] this would lead to profit maximizing (and socially optimal) locations at the quartiles. Jehiel [26] and Friedman and Thisse [27], however, argue that firms will likely collude on prices only since locations are fixed once for all. Following the collusive agreement implies bearing the risk of being cheated. If a rival deviates from the collusive outcome, revenge could be taken only via prices, which may not be credible. For the case that firms are allowed to collude only in prices, Jehiel [26] and Friedman and Thisse [27] consider a game of location choice in the first stage followed by an infinitely repeated price game. While Jehiel applies Nash bargaining within the cartel, Friedman and Thisse use a profit-sharing rule corresponding to the respective ratio of profits without collusion to determine the

collusive outcome. Without money transfers between competitors, both firms locate at the market center. This result is to be expected since through collusion, the threat of fierce price competition preventing the firms to approach each other disappears. Moreover, taking the same position has the advantage that the threat of cutting the price has the greatest effect on the rivals' profits and thus, making it more probable for collusion to sustain.

The Number of Firms

Let us now address the question of how a firm's choices of location and price is affected by the number of firms. In general, the analytical tractability of Hotelling models decreases with the number of firms. Hence, there is only a limited number of contributions to the multiple firms model. Examining the model on the circular domain, Salop [23] proves the existence of the price stage equilibrium given an equidistant location setting. However, an equilibrium does not exist for every feasible subgame. Again, as in the duopoly, this problem can be circumvented by assuming quadratic transport cost. This assures the existence of a price equilibrium in every subgame and further, the equilibrium existence of the whole game [28]. The locations in this equilibrium can be interpreted as following the principle of maximum differentiation. That is, firms locate equidistantly in order to maximize the minimum distance to their direct neighbors. Hence, in the circular domain, the number of firms does not have a significant effect on the relative amount of differentiation.

In the linear market, the nonexistence of a price equilibrium of Hotelling's original model could be remedied by allowing more than two firms to enter [28]. However, as there are strong incentives for the firms to agglomerate at the market center, a location equilibrium could not be established. Agglomeration is not stable since it is associated with Bertrand competition and deviating increases profits. In (nonequilibrium) equidistant configurations the corner firms are provided with some kind of market power, which enables them to charge

higher prices than their competitors. This is attributable to the fact that corner firms only have rivals on one side and hence, price competition is lower than for inside firms. The price structure is U-shaped, that is, prices decrease toward the center firms.

A similar setup with quadratic transport costs was analyzed by Brenner [29]. He shows that a price equilibrium exists for every feasible subgame and further, that the principle of maximum differentiation does not carry over to the multiple firms case. For games with up to nine players, equilibria are calculated, which are characterized by a U-shaped price structure. Moreover, corner firms locate inside the interval, which manifests their market power. However, if more than three firms are in the market, the equilibrium differentiation pattern does not much deviate from the socially optimal pattern.

In summary, allowing more than two firms to compete has the following effects. The price competition is not endangered although it seems to be softened. This effect arises because there exists an externality of a price change. By responding to an action of one neighbor, a firm also has to take into account the response of the second neighbor. Consequently, the maximum differentiation equilibrium in the linear market with quadratic transport costs is destroyed. Furthermore, given a market boundary the corner firms enjoy a market power from possessing a hinterland, which is reflected in higher prices and higher profits than their competitors if $n > 3$.

Customer Distributions

In the standard model, customers are distributed uniformly over the unit interval or the circumference. However, in the domain of product characteristics one usually finds customers' preferences clustered around some brands while in a geographic space, there is often an agglomeration of inhabitants at the business center of a town. Intuitively, we might expect that differentiation decreases as the density at the interval center increases, while price competition gets fiercer. In the

extreme case where all the density is concentrated in one point, the Bertrand paradigm results.

Several attempts have been made to relax the uniform distribution assumption. Shilony [30] has proven that the problem of equilibrium nonexistence in Hotelling's original model could not be solved by allowing more general density functions. Interestingly, the locations that provide price equilibria still imply a tendency to agglomerate. Assuming quadratic transport costs, Neven [31] has shown that for concave symmetric distributions a price equilibrium exists and further, that the whole game has an equilibrium. For distributions that are not too concave maximum differentiation holds. From a certain degree of concavity onward, however, firms choose inside locations, which are not more than one-eighth away from the interval boundaries.

In their frequently cited paper, Caplin and Nalebuff [32] have demonstrated that price equilibria exist for the broader class of log-concave distributions. This finding has led to the question of whether equilibrium existence holds for the whole game. Tabuchi and Thisse [33] found that symmetric locational equilibria do not exist for the subclass of triangular density functions. It was suspected that the nondifferentiability of the density at the median point would be the reason for this.

The most general approach to this problem so far was presented by Anderson *et al.* [34]. Considering a setup of quadratic transport costs and unrestricted locations on the real line, they have shown for the class of log-concave distributions that symmetric location equilibria exist only if the distribution of consumers is not too concave at the center. Otherwise, asymmetric equilibria may appear given the distribution not being too asymmetric. By positioning asymmetrically, firms shift the marginal consumer away from the very competitive region of high density to a more remote area in order to relax the competition.

Uncertainty

Uncertainty by the firms about an auxiliary dimension of product characteristics was

introduced by Rhee *et al.* [35]. In their model of linear transport costs, a variety of possible outcomes exist. If consumers exhibit sufficient heterogeneity along the unobservable attribute, minimum differentiation results. In contrast, the level of differentiation increases along the observable attribute as the uncertainty about the other dimension decreases. Moreover, increased uncertainty makes the competition more and more irrelevant leading to increasing prices.

Multiple Dimensions

Adding further spatial dimensions may influence the outcome of the game [36,37]. Neven and Thisse [36] demonstrate that maximum differentiation prevails at one dimension while agglomeration is observed at the other. Along which dimension the distance is maximized depends on the relative length of those domains. For the case of a two-dimensional plane of horizontal product differentiation (a square) minimum differentiation results along one and maximum differentiation along the other dimension.

This finding gives rise to the question of whether introducing further dimensions along which firms are able to differentiate from rivals leads to more or less differentiation. Irmen and Thisse [38] show that along all but one dimension firms agglomerate. Dimensions are weighted differently according to their relative importance. The dimension which is weighted maximally is used to differentiate. Even for the three-firms case it could be shown that maximum differentiation is never an equilibrium.

CONCLUSIONS

This survey reveals that differentiation—either in geographic or product space—depends delicately on parameters of the market structure. With respect to the degree of price competition, the existing literature suggests that it becomes tougher

- as discriminatory pricing or price-matching is available as a strategy,
- as the demand elasticity becomes lower,
- when fewer firms are competing,
- when the density of consumers is less concentrated in the center of the market,
- as the firm's uncertainty increases regarding the heterogeneity of consumers' preferences of a second dimension of product characteristics,
- if advertising tools are available that shift consumers toward the advertised product.

Furthermore, price competition is eliminated if firms are able to collude in prices. If the number of dimensions increases, price competition seems to become more and more irrelevant since differentiation occurs only along one single dimension. Endogenous transport costs, however, are inconclusive with respect to price competition. The degree of price competition affects the location choices via the strategic effect. If price competition becomes tougher, differentiation is fostered, and vice versa. Furthermore, there are effects of the model parameters, which directly influence profits of the location choice. In particular, differentiation is increased if the density of customers becomes more concentrated. Furthermore, there is a tendency to agglomerate when collusion is permitted in order to punish a deviating rival maximally.

One of the points of concern about the Hotelling literature is the strong imbalance between the number of theoretical papers on the one side and the number of empirical and experimental papers on the other. Notable exceptions are the empirical studies in Refs 39–41. The former two estimate a highly localized discrete-choice model in a space of several dimensions of product characteristics. The latter discriminate between theories of localized and global competition. Simplified versions of the Hotelling game were subject to experimental research. The impact of collusion in the location choice stage was examined by Brown-Kruse *et al.* [42] and Brown-Kruse and Schenk [43]. In laboratory experiments, individuals have chosen

- as the transport costs become more convex,

to locate at the center if communication was possible. Otherwise, there was a tendency toward the profit maximizing point at the quartiles. Huck *et al.* [44] investigated how individuals behave in a four-players game without price competition. In the Nash equilibrium, two players locate at each of the quartiles. The experimental results, however, suggest that because of best-response dynamics there is a tendency to locate at the center of the market.

REFERENCES

- Hotelling H. Stability in competition. *Econ J* 1929;3:41–57.
- d'Aspremont CJ, Gabszewicz J, Thisse J-F. On Hotelling's stability in competition. *Econometrica* 1979;47:1145–1150.
- Eaton B, Lipsey R. The principle of minimum differentiation reconsidered: some new developments in the theory of spatial competition. *Rev Econ Stud* 1975;42:27–50.
- Shaked A. Non-existence of equilibrium for the 2-dimensional 3-firm location problem. *Rev Econ Stud* 1975;42:51–56.
- Okabe A, Aoyagi M. Existence of equilibrium configurations of competitive firms on an infinite 2-dimensional space. *J Urban Econ* 1991;29:349–370.
- Economides N. Minimal and maximal product differentiation in Hotelling's duopoly. *Econ Lett* 1986;21:67–71.
- Gabszewicz JJ, Thisse J-F. On the nature of competition with differentiated products. *Econ J* 1986;96:160–172.
- Anderson SP. Equilibrium existence in the linear model of spatial competition. *Economica* 1988;55:479–491.
- Kats A. More on Hotelling's stability in competition. *Int J Ind Organ* 1995;13:89–93.
- de Frutos MA, Hamoudi H, Jarque X. Equilibrium existence in the circle model with linear quadratic transport cost. *Reg Sci Urban Econ* 1999;29:605–615.
- von Ungern-Sternberg T. Monopolistic competition and general purpose products. *Rev Econ Stud* 1988;55:231–246.
- Hendel I, de Figueiredo JN. Product differentiation and endogeneous disutility. *Int J Ind Organ* 1997;16:63–79.
- Lancaster K. Variety, equity, and efficiency. New York: Columbia University Press; 1979.
- Hoover EM. Spatial price discrimination. *Rev Econ Stud* 1937;4:182–191.
- Eaton BC, Wooders MH. Sophisticated entry in a model of spatial competition. *RAND J Econ* 1985;16:282–297.
- Thisse J-F, Vives X. On the strategic choice of spatial price policy. *Am Econ Rev* 1988;78:122–137.
- Norman G, Thisse J-F. Product variety and welfare under discriminatory and mill pricing policies. *Econ J* 1996;106:76–91.
- Tabuchi T. Pricing policy in spatial competition. *Reg Sci Urban Econ* 1999;29:617–631.
- Zhang ZJ. Price matching policy and the principle of minimum differentiation. *J Ind Econ* 1995;43:287–299.
- Tirole J. The theory of industrial organization. Cambridge (MA): MIT Press; 1990.
- Gabszewicz JJ, Thisse J-F. Location. In: Aumann RJ, Hart S, editors. *Handbook of game theory*. Amsterdam: Elsevier Science Publishers; 1992. pp. 282–304.
- Smithies A. Optimum location in spatial competition. *J Polit Econ* 1941;49:423–439.
- Salop S. Monopolistic competition with outside goods. *Bell J Econ* 1979;10:141–156.
- Böckem S. A generalized model of horizontal product differentiation. *J Ind Econ* 1994;42:287–298.
- Hinloopen J, Marrewijk CV. On the limits and possibilities of the principle of minimum differentiation. *Int J Ind Organ* 1999;17:735–750.
- Jehiel P. Product differentiation and price collusion. *Int J Ind Organ* 1992;10:633–643.
- Friedman JW, Thisse J-F. Partial collusion fosters minimum product differentiation. *RAND J Econ* 1993;24:631–645.
- Economides N. Hotelling's "Main Street" with more than two competitors. *J Reg Sci* 1993;33(3):303–319.
- Brenner S. Hotelling games with three, four, and more players. *J Reg Sci* 2005;45(4): 851–864.
- Shilony Y. Hotelling's competition with general customer distributions. *Econ Lett* 1981;8:39–45.
- Neven DJ. On Hotelling's competition with non-uniform customer distributions. *Econ Lett* 1986;21(2):121–126.
- Caplin A, Nalebuff B. Aggregation and imperfect competition: on existence of equilibrium. *Econometrica* 1991;59:25–59.

33. Tabuchi T, Thisse J-F. Asymmetric equilibria in spatial competition. *Int J Ind Organ* 1995;13:213–227.
34. Anderson SP, Goeree JK, Ramer R. Location, location, location. *J Econ Theory* 1997;77:102–127.
35. Rhee B-D, de Palma A, Fornell C, *et al.* Restoring the principle of minimum differentiation in product positioning. *J Econ Manage Strategy* 1992;1:475–506.
36. Neven D, Thisse J-F. On quality and variety competition. In: Gabszewicz JJ, Richard J, Wolsey L, editors. *Economic decision making: games, econometrics and optimisation. contributions in honour of J. Drèze*. Amsterdam: Elsevier Science Publishers; 1990.
37. Tabuchi T. Two-stage two-dimensional spatial competition between two firms. *Reg Sci Urban Econ* 1994;24:207–227.
38. Irmen A, Thisse J-F. Competition in multi-characteristics spaces: Hotelling was almost right. *J Econ Theory* 1998;78:76–102.
39. Feenstra RC, Levinsohn JA. Estimating markups and market conduct with multidimensional product attributes. *Rev Econ Stud* 1995;62:19–52.
40. Thomas L, Weigelt K. Product location choice and firm capabilities: Evidence from the U.S. automobile industry. *Strateg Manage J* 2000;21:897–909.
41. Pinkse J, Slade ME, Brett C. Spatial price competition: a semiparametric approach. *Econometrica* 2002;70(3):1111–1153.
42. Brown-Kruse J, Cronshaw MB, Schenk DJ. Theory and experiments on spatial competition. *Econ Inq* 1993;31:139–165.
43. Brown-Kruse J, Schenk DJ. Location, cooperation and communication: an experimental examination. *Int J Ind Organ* 2000;18:59–80.
44. Huck S, Müller W, Vriend N. The East End, the West End, and King's Cross: on clustering in the four-player Hotelling game. *Econ Inq* 2002;40(2):231–240.

LOT-SIZING

WILCO VAN DEN HEUVEL
Econometric Institute, Erasmus
University Rotterdam,
Rotterdam, The Netherlands

INTRODUCTION

Lot-sizing is one of the key planning activities involved in production and inventory management. Every firm that faces customer demand needs to decide when to produce (or order) and in what quantities, in order to satisfy the demand. Typically, firms incur fixed costs for every time production takes place, and holding costs for carrying items from a period to the next period. In general, if production takes place frequently, average fixed costs are relatively high, while holding costs are relatively low. On the other hand, in the case of infrequent production and higher production quantities, average fixed costs are relatively low, while holding costs are relatively high. The objective of a lot-sizing model is to determine an optimal production plan, that is, the timing and level of production such that demand is satisfied and costs are minimized.

The term *lot-sizing* is usually used in the case of deterministic demand in the literature. The economic order quantity (EOQ) model is probably the most famous lot-sizing model in production and inventory management [1]. In this infinite and continuous-time model, we are faced with a constant demand rate that has to be satisfied by placing orders. For every order placed, there is a fixed order cost, while holding costs are incurred for keeping items in stock. The objective is to find the order plan that minimizes the sum of these costs.

This article mainly focuses on the discrete counterpart of the EOQ model: the dynamic lot-sizing (DLS) or the economic lot-sizing model. Both names are used in the literature and can be traced back from the titles of

the two seminal papers “Dynamic version of the economic lot size model” [2] and “Programming of economic lot sizes” [3]. The former paper proposes a polynomial-time solution approach, while the latter paper provides a mathematical formulation for the problem with multiple items. In this article, we use the name *dynamic lot-sizing (DLS)* model as is used most in the literature (based on the number of hits on Google Scholar).

In contrast to the EOQ model, the DLS model has a discrete and finite time horizon. Moreover, “dynamic” refers to the dynamic nature of the demands, that is, the demand quantities may vary over time but they are deterministic (known in advance). The DLS model can be applied both in a production and in an ordering setting. In the first case, demands are satisfied by production, while in the latter case, demands are satisfied by ordering items from an external supplier. We use the term *procurement* to cover both settings.

The remainder of this article is organized as follows. We start the next section with a brief description of the different cost components in lot-sizing models. In the section titled “Continuous-Time Models,” we consider two continuous-time lot-sizing models: the EOQ model and the joint replenishment problem. The following sections deal with the DLS model. In the section titled “The Dynamic Lot-Sizing Model: Formulations,” we present a basic mathematical formulation and discuss two alternative formulations; in the section titled “Solution Methods for the DLS Model,” we give the solution approach proposed in Ref. 2; and in the section titled “Extensions of the DLS Model,” we provide a list of possible extensions of the DLS model. We end the article with some concluding remarks.

COST COMPONENTS IN LOT-SIZING

As already mentioned, a lot-sizing model tries to find a procurement plan that minimizes the total cost. The costs in

lot-sizing models typically consist of the following cost components: fixed procurement cost, unit variable procurement cost, and unit holding cost. In the case of the DLS model, all cost components are allowed to vary over the time periods.

The fixed costs in lot-sizing models do not depend on the procured quantity. In an ordering setting, the fixed cost may consist of the administrative cost, such as typing of orders, receiving orders, handling invoices, and so on. In a production setting, the cost may consist of the labor cost of a mechanic to setup or to change the settings of a machine.

Furthermore, in an ordering setting, the unit variable procurement costs involve the unit purchasing cost, with possibly the addition of freight and handling cost. For producers, the unit production cost consists of the amount of money spent on producing a single item. For example, this includes the cost of the required raw materials and the labor cost to assemble an item. As for the fixed cost, it is assumed that the unit variable costs do not depend on the procurement quantity. For example, this means that we do not consider quantity discounts in the basic models.

Finally, the unit holding costs consist of the cost to keep an item in stock. This includes the cost of running the warehouse, handling items, damage, insurance, and so on. However, the largest portion of the holding cost often consists of the opportunity cost of the capital invested in the inventory. Instead of investing the money in the purchased or produced items, a company could have earned an alternative return on the investment.

It should be clear from the above discussion that, in practice, the precise figures for the different cost components are difficult to calculate or estimate in general. One reason is that they often do not appear explicitly in the accounting systems of most organizations. For a more detailed discussion on the different cost components, we refer to (4, Section 3.6.1).

CONTINUOUS-TIME MODELS

In this section, we consider two continuous-time lot-sizing models. The first model deals

with a single-item problem, while the second model deals with the procurement of multiple items.

Economic Order Quantity Model

The EOQ model dates back to 1913 [1]. In this model, there is a continuous demand rate over an infinite planning horizon. The demand needs to be satisfied by placing orders. For each order placed, there is a fixed order cost and holding cost is incurred for keeping items in stock. To formulate the EOQ model, we use the following notation:

- D : demand rate (in items per unit time);
- K : fixed order cost (in \$ per order);
- h : holding cost (in \$ per item per unit time).

We now determine the average cost per unit time if the order quantity is equal to Q . This means that the time between orders, or the cycle time, is $T = Q/D$ time units and hence the average order cost is $K/T = KD/Q$ per unit time. Furthermore, the inventory is equal to Q items just after the replenishment and there are no items in stock just before the next replenishment. The inventory level over time is depicted in Fig. 1. Thus, the average inventory equals $\frac{1}{2}Q$ with average holding cost $\frac{1}{2}hQ$. So the average total cost per unit time is equal to

$$KD/Q + \frac{1}{2}hQ. \quad (1)$$

Since Equation (1) is a convex function in Q , it is sufficient to use the first-order condition to find the minimum. It follows that the optimal order quantity Q^* is equal to

$$Q^* = \sqrt{\frac{2KD}{h}}$$

and so the optimal cycle time T^* equals $T^* = Q^*/D = \sqrt{2K/(Dh)}$. The total cost in an optimal solution equals $\sqrt{2KDh}$. The optimal order quantity satisfies the property that $KD/Q^* = \frac{1}{2}hQ^* = \sqrt{\frac{1}{2}KDh}$, which means that the average order cost and average holding cost are perfectly balanced in an optimal solution.

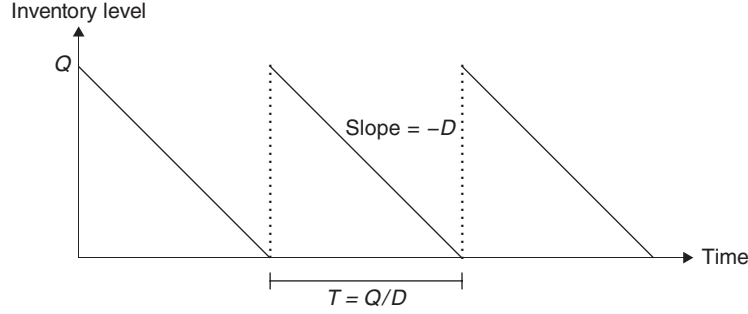


Figure 1. Inventory level in the EOQ model.

We note that variable order cost does not affect an optimal solution and hence need not be taken into account in the EOQ model: if the unit variable order cost equals p , then the variable procurement cost equals pQ per order and hence $pQ/T = pD$ per unit time. Since this is just a constant, it does not affect the value of Q that minimizes Equation (1). Finally, because the total cost curve is relatively flat in the area of the optimal order quantity, the EOQ model is a robust model. So as long as the parameter estimates are good (but not perfect), the cost will be very close to the minimum total cost.

Joint Replenishment Problem

Instead of a single item, a firm often needs to determine a procurement plan for a collection of items. Economies of scale occur if several items are procured at the same time. This generalization of the EOQ model is called the *joint replenishment problem (JRP)*. Assume that we have a JRP for n items. There is a fixed cost K_i if item i is ordered and a fixed cost K_0 if any item is procured. Furthermore, we apply the same notation as in the previous section with an additional index i that corresponds to item $i = 1, \dots, n$. Clearly, we save on the fixed cost K_0 if two or more items are procured simultaneously.

In contrast to the EOQ model, the optimal order quantity for an item i does not have to be equal at every replenishment point in the JRP [5]. However, to simplify the problem, many authors apply a policy with equal cycle times and hence equal order quantities for every item. Let T be the base cycle time and let $m_i T$ ($m_i \in \mathbb{N}$) be the cycle time of item

i . So at every m_i th multiple of T , item i is procured. To find the optimal base cycle time T and the multipliers m_i , we need to solve the model

$$\begin{aligned} \min \quad & K_0/T + \sum_{i=1}^n \left(\frac{K_i}{m_i T} + \frac{1}{2} h_i m_i T D_i \right) \\ \text{s.t.} \quad & T > 0 \\ & m_i \in \mathbb{N} \quad i = 1, \dots, n. \end{aligned}$$

Using calculus, it is easy to find the optimal T for given values m_i and conversely, it is easy to find the values m_i for a given T . However, it is not easy to find the optimal T and m_i simultaneously. In fact, it is shown in Ref. 6 that the JRP is $\mathcal{NP}$ -hard. A lot of research has been done to find optimal or near-optimal solutions for the above model. For a review on solution approaches, we refer to Refs 7 and 8.

THE DYNAMIC LOT-SIZING MODEL: FORMULATIONS

In contrast to continuous-time models, the DLS model has a finite and discrete-time horizon. Assume that the model horizon consists of T time periods. We use the following notation for the parameters in periods $t = 1, \dots, T$:

- d_t : demand in period t ,
- K_t : fixed procurement cost in period t ;
- p_t : unit variable procurement cost in period t ;
- h_t : unit holding cost in period t .

It is assumed that all costs are nonnegative. Furthermore, we let $d_{s,t} = \sum_{i=s}^t d_i$ denote the cumulative demand from period s up to period t with $1 \leq s \leq t \leq T$.

In the next section, we give a basic formulation of the DLS model. In the sections titled “The Network or the Shortest Path Formulation” and “The Facility Location Formulation,” we discuss two alternative mathematical formulations that are frequently used in the literature: the network or the shortest path (SP) formulation and the facility location formulation. The advantage of these formulations over the basic formulation is that they are tight. That is, the solution of the linear programming (LP) relaxation is integer. When solving extensions of the DLS model, such as the problem with multiple items and capacity constraints (see the section titled “Extensions of the DLS Model”), the alternative formulations are useful because they often provide better lower bounds.

Basic Mathematical Formulation

In the basic formulation, we use the following decision variables for periods $t = 1, \dots, T$

- x_t : procurement quantity in period t ;
- y_t : is 1 if period t is a procurement period and 0 otherwise;
- I_t : ending inventory in period t .

Then the DLS model can be formulated as the mixed integer programming (MIP) model

$$\begin{aligned} \min \quad & \sum_{t=1}^T (K_t y_t + p_t x_t + h_t I_t) \\ \text{s.t.} \quad & I_{t-1} + x_t = d_t + I_t \quad t = 1, \dots, T \end{aligned}$$

$$\begin{aligned} x_t &\leq d_t y_t & t = 1, \dots, T \\ x_t, I_t &\geq 0 & t = 1, \dots, T \\ y_t &\in \{0, 1\} & t = 1, \dots, T \\ I_0 &= 0. \end{aligned}$$

The first set of constraints are the inventory balance or demand constraints. Demand is satisfied from procurement or inventory from the previous period, while remaining items build up the current inventory. The second set of constraints are the setup forcing constraints, which, together with the fourth set of constraints, ensure that the setup variable is forced to 1 if the procurement quantity is positive. Or alternatively, if there is no setup ($y_t = 0$), then there will not be procurement. Note that the procurement quantity will never exceed the demand from period t to the final period T ($d_{t,T}$), as costs are assumed to be nonnegative. The third set of constraints makes sure that the procurement quantity and inventory are nonnegative. Finally, we assume that the starting inventory is equal to zero.

In fact, the model can be represented as the minimum cost flow network (see also **Minimum Cost Flows**) depicted in Fig. 2. There is one source node 0 with procurement arcs $(0, t)$ to every period t . Furthermore, there is an arc $(t, t+1)$ that carries inventory from period t to $t+1$.

The Network or the Shortest Path Formulation

The network or the shortest path (SP) formulation [9] utilizes the fact that there exists an optimal solution satisfying the zero-inventory ordering (ZIO) property (see the

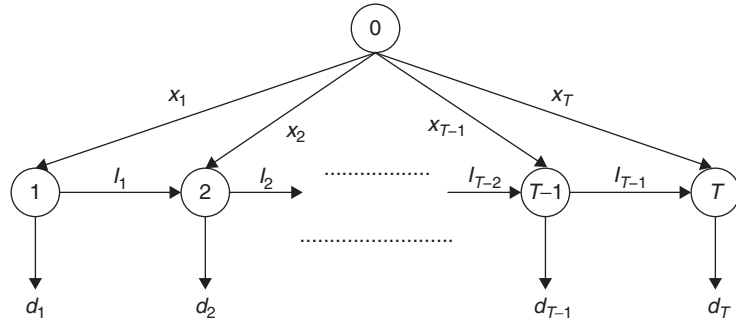


Figure 2. The DLS problem represented as a network flow problem.

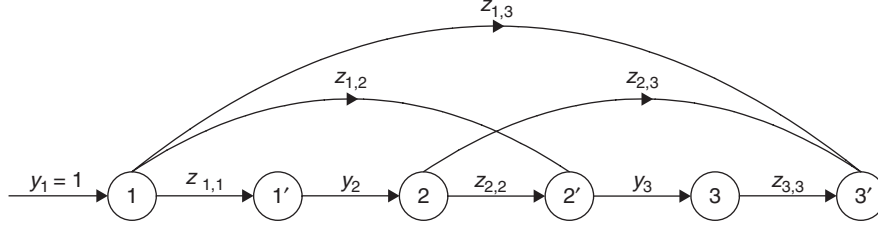


Figure 3. The DLS problem represented as a shortest path problem.

section titled “Solution Methods for the DLS Model” for more details). Such a solution has the property that in a procurement period τ , demands $d_{\tau,t}$ are procured for some $t \geq \tau$. In the formulation, the binary variable $z_{\tau,t}$ is equal to 1 if period τ exactly satisfies the demands for periods $\tau, \dots, t$. The associated variable procurement and holding costs are equal to

$$C_{s,t} = p_s d_{s,t} + \sum_{\tau=s}^{t-1} h_{\tau} d_{\tau+1,t}.$$

Again using the setup variables y_t , the SP formulation is given by

$$\begin{aligned} \min \quad & \sum_{\tau=1}^T (K_{\tau} y_{\tau} + \sum_{t=\tau}^T C_{\tau,t} z_{\tau,t}) \\ \text{s.t.} \quad & \sum_{t=1, \dots, T} z_{1,t} = 1, \\ & \sum_{\tau=1}^t z_{\tau,t} = \sum_{\tau=t+1}^T z_{\tau+1,\tau} \quad t = 1, \dots, T-1 \\ & \sum_{t=\tau}^T z_{\tau,t} \leq y_{\tau} \quad \tau = 1, \dots, T \\ & y_{\tau}, z_{\tau,t} \in \{0, 1\} \quad \tau = 1, \dots, T; \\ & \quad \quad \quad t = \tau, \dots, T. \end{aligned}$$

The first constraint ensures that period 1 is a procurement period. The second constraint set models the requirement that if demand d_t is the last demand satisfied by procurement in some period τ , then period $t+1$ should also be a procurement period. The third set of constraints makes sure that y_{τ} is set to 1 if period τ is a procurement period. We note that the variables $z_{\tau,t}$ will become continuous

in the extended models (see the section titled “Extensions of the DLS Model”), because the ZIO property does not hold anymore.

Figure 3 graphically represents the SP formulation for a three-period problem. The arcs in the figure correspond to the variables y_t and $z_{\tau,t}$ with associated cost K_t and $C_{\tau,t}$, respectively. The optimal cost can be determined by finding the shortest path in the network. As mentioned earlier, the LP relaxation of this formulation provides an integer solution.

The Facility Location Formulation

Krarup and Bilde [10] propose a formulation using the variables $x_{\tau,t}$, where $x_{\tau,t}$ is the procured quantity in period τ to satisfy demand in period t . Furthermore, we let $c_{\tau,t} = p_{\tau} + \sum_{j=\tau}^{t-1} h_j$ denote the variable cost (procurement plus holding cost) to satisfy a unit demand in period t by procurement in period τ . If we again use the setup variables y_t , then the following model is called the *facility location* (FL) formulation:

$$\begin{aligned} \min \quad & \sum_{\tau=1}^T \left(y_{\tau} + \sum_{t=\tau}^T c_{\tau,t} x_{\tau,t} \right) \\ \text{s.t.} \quad & \sum_{\tau=1}^t x_{\tau,t} = d_t \quad t = 1, \dots, T \\ & x_{\tau,t} \leq d_t y_{\tau} \quad \tau = 1, \dots, T; \quad t = i, \dots, T \\ & x_{\tau,t} \geq 0 \quad \tau = 1, \dots, T; \quad t = \tau, \dots, T \\ & y_t \in \{0, 1\} \quad t = 1, \dots, T. \end{aligned}$$

The first constraint set ensures that every demand is satisfied, while the second set of constraints is a set of setup forcing constraints.

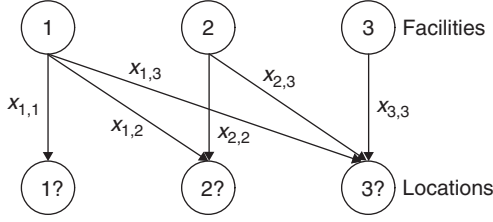


Figure 4. The DLS problem represented as a facility location problem.

Figure 4 graphically represents this formulation as a facility location problem (hence the name). The FL problem is a problem in location theory (see also the section titled “Location Analysis” in this encyclopedia) where demands of a set of clients need to be satisfied from a set of facilities. Demand of a client can only be satisfied from a facility if the facility is open. There is a fixed cost to open a facility and a variable transportation cost to satisfy a unit demand of a client by some facility.

If we associate a facility τ and a client t for every time period of the DLS problem, then $x_{\tau,t}$ is the demand of facility t satisfied by facility τ . Furthermore, the setup variable $y_\tau = 1$ ($y_\tau = 0$) corresponds to an open (closed) facility. Note that a facility τ can only satisfy the demand of some period t with $t \geq \tau$ and hence it is a special case of an FL problem. It is shown in Ref. 10 that the LP relaxation of this special FL problem provides an integer solution. This property does not hold for the general FL problem.

SOLUTION METHODS FOR THE DLS MODEL

One way of solving the DLS model is by applying a commercial solver to one of the formulations of the previous section. However, another solution approach is to develop a specialized algorithm. Wagner and Whitin [2] propose a solution method that is based on an important structural property of an optimal solution: the ZIO property. This property states that there is no procurement in a period if there is inventory available at the start of the period, or, more formally, $I_{t-1}x_t = 0$ for

$t = 1, \dots, T$. Zangwill [11] proves that this property holds by using the network flow representation of Fig. 2. He shows that there exists an optimal solution in a single source network with concave costs, such that every node has at most one incoming arc with positive flow. Applying this to the network in Fig. 2 implies that the flow on one of the incoming arcs of node t (I_{t-1} or x_t) is zero. The ZIO property implies that if t is a procurement period, then $x_t = d_{t,k}$ for some $k = t, \dots, T$. In other words, a procurement period always satisfies a consecutive number of demands starting at period t . This property is used to develop a dynamic programming (see also the section titled “Dynamic Programming” in this encyclopedia) algorithm.

To this end, let $C_{s,t}$ be the total cost to satisfy demand in periods $s, \dots, t$ by procurement in period s , that is,

$$C_{s,t} = K_s + p_s d_{s,t} + \sum_{i=s}^{t-1} h_i d_{i+1,t},$$

where the terms correspond to the setup, procurement, and holding costs (we define the empty sum to be 0). Furthermore, let F_t be the minimal cost to satisfy demands in periods $1, \dots, t$. So F_t is the optimal cost of the problem when considering only periods $1, \dots, t$. Then the optimal cost can be determined from the following recursive equation:

$$F_t = \min_{\tau=1, \dots, t} \{F_{\tau-1} + C_{\tau,t}\} \text{ for } t = 1, \dots, T. \quad (2)$$

In the recursion formula, period τ is the last procurement period in the t -period problem. The optimal cost for periods $1, \dots, \tau - 1$ has already been determined and equals $F_{\tau-1}$, while $C_{\tau,t}$ is the cost to satisfy demand in periods $\tau, \dots, t$ by definition. Because we are searching for the minimum cost procurement plan, we take the minimum over all periods $\tau = 1, \dots, t$. The recursion is initialized by $F_0 = 0$ and the optimal cost is equal to F_T .

It is not difficult to see that the running time of this algorithm is $\mathcal{O}(T^2)$. More than 30 years after the Wagner–Whitin algorithm,

faster algorithms were independently developed by Federgruen *et al.* [12,13] and Aggarwal and Park [14]. The algorithms solve the DLS model in a running time of $\mathcal{O}(T \log T)$. In the special case of nonspeculative motives, that is, $p_t + h_t \geq p_{t+1}$, the running time even reduces to $\mathcal{O}(T)$.

EXTENSIONS OF THE DLS MODEL

In practice, lot-sizing problems are often more complex than the DLS problem. Therefore, many extensions have been proposed in the literature to capture reality better. In this section, we discuss some of the extensions and show how these extensions can be incorporated in the formulation using the variables of the basic formulation. We consider the following three categories of extensions: demand, operational, and multiple items. We do not discuss solution approaches to solve the more complex lot-sizing models, because this is beyond the scope of this article.

EXTENSIONS ON DEMAND

Backlogging. If demand is not immediately met but satisfied in a later period, then the demand is satisfied by backlogging [11]. This means that demand in a period can now be satisfied by procurement, inventory, or backlogging. So backlogging can be considered as negative inventory. If I_t^- is the amount backlogged in period t , then the inventory balance constraints change to

$$x_t + I_{t-1} + I_t^- = d_t + I_t + I_{t-1}^- \quad \text{for } t = 1, \dots, T.$$

In the network flow representation of Fig. 2, backlogging can be represented by arcs from node t to $t-1$ for $t = 2, \dots, T$.

Stockouts. A stockout occurs if demand is not satisfied or backlogged but lost [15]. Some models allow for partial stockouts, which means that a portion of the demand is satisfied. Let l_t be the

amount of lost-sales in period t . Then partial stockouts are imposed by the constraints

$$\begin{aligned} I_{t-1} + x_t + l_t &= d_t + I_t \\ &\text{for } t = 1, \dots, T, \\ 0 \leq l_t &\leq d_t \quad \text{for } t = 1, \dots, T. \end{aligned}$$

The variables l_t are included in the objective function to account for the stockout costs. For example, the coefficient of l_t in the objective function may represent the selling price of the item.

Time Windows. Sometimes demand does not have to be satisfied in a single time period, but an earliest and latest delivery period are specified [16]. Let D_j ($j = 1, \dots, M$) be the demand that needs to be satisfied in the time window $[b_j, e_j] = \{t = 1, \dots, T : b_j \leq t \leq e_j\}$. Note that the number of time windows M equals at most $\frac{1}{2}T(T+1)$ because $1 \leq b_j \leq e_j \leq T$. We need some additional variables to formulate the time window constraints. Let $w_{t,j}$, $j = 1, \dots, M$ and $b_j \leq t \leq e_j$, be the procurement quantity in period t to satisfy demand D_j . Then the constraints

$$\sum_{t=b_j}^{e_j} w_{t,j} = D_j \quad \text{for } j = 1, \dots, M$$

together with the modified demand constraints

$$I_{t-1} + x_t = \sum_{j: b_j \leq t \leq e_j} w_{t,j} + I_t \quad \text{for } t = 1, \dots, T,$$

ensure that demand is met in the specified time window.

Another application of time windows is considered in Ref. 17, where demand D_j needs to be satisfied in period e_t by procurement in the time window $[b_j, e_j]$. This is called *lot-sizing with delivery time windows*. Again, let $w_{t,j}$, $j = 1, \dots, M$ and $b_j \leq t \leq e_j$ be the procurement amount in period t to satisfy D_j .

Furthermore, let $\Delta_t = \sum_{j:e_j=t} D_j$ be the total demand that needs to be satisfied in period t . To formulate the DLS model with delivery time windows, the demand constraints change to

$$\begin{aligned} x_t + I_{t-1} &= \Delta_t + I_t, & \text{for } t = 1, \dots, T, \\ \sum_{t=b_j}^{e_j} w_{t,j} &= D_j, & \text{for } j = 1, \dots, M. \end{aligned}$$

and we need the constraints

$$x_t = \sum_{j:b_t \leq j \leq e_t} w_{t,j} \text{ for } t = 1, \dots, T$$

to link the x_t and $w_{t,j}$ variables.

Operational Extensions

Procurement Capacity. In the DLS model, it is implicitly assumed that the procurement capacity is infinite. However, the capacity in a period (such as the number of man-hours) is limited in practice. This leads to the additional constraints

$$x_t \leq C_t \text{ for } t = 1, \dots, T,$$

with C_t the procurement capacity in period t . Because of this constraint, the well-known ZIO property does not hold anymore and the structure of an optimal solution becomes more complicated. In fact, in the case of time-varying capacities, the problem becomes $\mathcal{NP}$ -hard [18]. The computational complexity of the capacitated DLS model is studied in more detail in Ref. 19. Complexity results are derived for families of problem instances based on the characteristics of the cost parameters and capacities.

Lower and Upper Bounds on Inventory.

The warehouse where the items are stored also has a finite capacity, which leads to an upper bound on the inventory in a period [20]. On the other hand, if a firm applies a (deterministic) lot-sizing model in a stochastic demand setting, lower bounds on inventory may be a heuristic way to set safety

stocks. This can be captured in the constraints

$$L_t \leq I_t \leq U_t \text{ for } t = 1, \dots, T,$$

with L_t (U_t) the lower (upper) bound on inventory in period t .

Multiple Items

Multiple Items. Instead of determining the timing and procurement quantities for a single item, multiple items are often involved in procurement planning. This can be modeled by introducing the additional index i for the parameters and decision variables. So the variables x_t^i , y_t^i , and I_t^i denote the procurement quantity, setup decision, and inventory of item i in period t , respectively. The capacity constraints change to

$$\sum_{i=1}^n x_t^i \leq C_t \text{ for } t = 1, \dots, T,$$

if the items share the same capacity. If only one item can be produced in a period, then the constraints

$$\sum_{i=1}^n y_t^i \leq 1 \text{ for } t = 1, \dots, T,$$

need to be added, where n is the number of items. This type of model is called a *small-bucket model*. This is in contrast to a *big-bucket model*, where all items can be produced in a period. Within the small-bucket models, we distinguish two types of models: (i) if production occurs, it must be at full capacity (this is called the *discrete lot-sizing and scheduling problem*) and (ii) if production occurs, it can be at any level within the capacity limits.

Startups. A startup or changeover occurs if a machine needs to be setup and has not been setup for the item in the previous period. So there is no startup if

the same item is produced in two consecutive periods. Additional setup cost or setup time may be associated with a startup. Let $z_t^i = 1$ denote a startup for item i in period t and $z_t^i = 0$ otherwise. If the z_t^i variables appear with a positive coefficient in the objective function, then the constraints

$$z_t^i \geq y_t^i - y_{t-1}^i \text{ for } t = 1, \dots, T; \\ i = 1, \dots, n,$$

model the startup. Clearly, startups can also be considered for the single-item version of the problem [21].

Setup Times. Setting up a machine may not only induce cost but may also consume part of the production capacity [3,22]. For example, capacity is lost due to cleaning, changing machine settings, inspection, testing, and so on. To account for setup times, the capacity constraints change to

$$\sum_{i=1}^n (s_t^i y_t^i + v_t^i x_t^i) \leq C_t \text{ for } t = 1, \dots, T,$$

with s_t^i the capacity of setting up the machine for item i and v_t^i the variable capacity consumption of item i in period t .

It is not difficult to see that the extensions in the sections “Extensions on Demand” and “Operational Extensions” can be easily generalized to the multi-item case.

CONCLUDING REMARKS

In the previous section, we considered several extensions of the DLS model and the consequences to the basic formulation. Clearly, the extended DLS models are more difficult to solve than the DLS model. Some practical lot-sizing problems can be effectively solved by mixed integer programming via tight formulations and valid inequalities [23,24]. To solve more complicated DLS models, several other advanced techniques are proposed in the literature. These techniques include,

for example, Dantzig–Wolfe decomposition, Lagrange relaxation, and column generation (see also the section titled “Large-Scale Optimization” in this encyclopedia), or metaheuristics (see also the section titled “Heuristics and Meta-Heuristics” in this encyclopedia) such as tabu search, simulated annealing, and genetic algorithms [25].

A drawback of lot-sizing models is that they are not able to capture uncertainty in the problem parameters. For example, the demand may not be known over the planning horizon or the capacity may be uncertain because of technical breakdowns. Therefore, one should be careful when applying a lot-sizing model and ascertain whether the assumptions of the model are a good representation of reality. If there is considerable uncertainty, one should apply a stochastic inventory model (see the section titled “Inventory Management and Control” in this encyclopedia). It should be mentioned that recently a stochastic lot-sizing model has been proposed, where stochasticity is captured in a scenario tree [26]. Each level of the scenario tree corresponds to a period t and each scenario occurs with some probability. The objective is to find a procurement plan that minimizes the expected lot-sizing costs.

Finally, the number of possible extensions of the DLS model is much larger than the ones presented in this article. For a comprehensive list of possible extensions and corresponding references (more than 240), we refer to Ref. 27.

REFERENCES

1. Harris FW. How many parts to make at once. *Factory, Mag Manage* 1913;10:135–136.
2. Wagner HM, Whitin TM. Dynamic version of the economic lot size model. *Manage Sci* 1958;5:89–96.
3. Manne AS. Programming of economic lot sizes. *Manage Sci* 1958;4(2):115–135.
4. Silver EA, Pyke DF, Petersen R. *Inventory management and production planning and scheduling*, 3rd ed. New York: Wiley; 1998.
5. Andres FM, Emmons H. On the optimal packaging frequency of products jointly replenished. *Manage Sci* 1976;22(10):1165–1166.
6. Arkin E, Joneja D, Roundy R. Computational complexity of uncapacitated multi-echelon

- production planning problems. *Oper Res Lett* 1989;8:61–66.
7. Goyal SK, Satir AT. Joint replenishment inventory control: deterministic and stochastic models. *Eur J Oper Res* 1989;38:2–13.
 8. Khouja M, Goyal S. A review of the joint replenishment problem literature: 1989–2005. *Eur J Oper Res* 2008;186:1–16.
 9. Eppen GD, Martin RK. Solving multi-item capacitated lot-sizing problems using variable redefinition. *Oper Res* 1987; 35(6):832–848.
 10. Krarup J, Bilde O. Plant location, set covering and economic lot size: an $O(mn)$ -algorithm for structured problems. In: Collatz L, Wetterling W editors, *Optimierung bei graphentheoretischen und ganzzahligen problemen*. Basel: Birkhäuser; 1977. pp. 155–180.
 11. Zangwill WI. A backlogging model and multi-echelon model of a dynamic economic lot size production system - a network approach. *Manage Sci* 1969; 15:506–527.
 12. Federgruen A, Tzur M. A simple forward algorithm to solve general dynamic lot sizing models with n periods in $O(n \log n)$ or $O(n)$ time. *Manage Sci* 1991; 37:909–925.
 13. Wagelmans APM, van Hoesel CPM, Kolen A. Economic lot sizing: an $O(n \log n)$ algorithm that runs in linear time in the Wagner-Whitin case. *Oper Res* 1992;40:S145–S156.
 14. Aggarwal A, Park JK. Improved algorithms for economic lot-size problems. *Oper Res* 1993;14:549–571.
 15. Sandbothe RA, Thompson GL. A forward algorithm for the capacitated lot size model with stockouts. *Oper Res* 1990; 38(3):474–486.
 16. Lee C-Y, Çetinkaya S, Wagelmans APM. A dynamic lot size model with demand time windows. *Manage Sci* 2001; 47(10):1384–1395.
 17. Wolsey LA. Lot-sizing with production and delivery time windows. *Math Program, Series A* 2006;107:471–489.
 18. Florian M, Lenstra JK, Rinnooy Kan AHG. Deterministic production planning: algorithms and complexity. *Manage Sci* 1980;26:669–679.
 19. Bitran GR, Yanasse HH. Computational complexity of the capacitated lot size problem. *Manage Sci* 1982; 28:1174–1186.
 20. Love F. Bounded production and inventory models with piecewise concave costs. *Manage Sci* 1973;20:313–318.
 21. Karmarkar US, Kekre S, Kekre S. The dynamic lot-sizing problem with startup and reservation costs. *Oper Res* 1987;35(3):389–398.
 22. Trigeiro WW, Thomas LJ, McClain JO. Capacitated lot sizing with setup times. *Manage Sci* 1989;35(3):353–366.
 23. Belvaux G, Wolsey LA. *bc—prod*: A specialized branch-and-cut system for lot-sizing problems. *Manage Sci* 2000; 46(5):724–738.
 24. Belvaux G, Wolsey LA. Modelling practical lot-sizing problems as mixed-integer programs. *Manage Sci* 2001; 47(7):993–1007.
 25. Jans R, Degraeve Z. Meta-heuristics for dynamic lot sizing: a review and comparison of solution approaches. *Eur J Oper Res* 2007;177:1855–1875.
 26. Guan Y, Miller AJ. Polynomial-time algorithms for stochastic uncapacitated lot-sizing problems. *Oper Res* 2008; 56(5):1172–1183.
 27. Jans R, Degraeve Z. Modeling industrial lot sizing problems: a review. *Int J Prod Res* 2008;46(6):1619–1643.

FURTHER READING

- Bahl HC, Ritzman LP, Gupta JND. Determining lot sizes and resource requirements: a review. *Oper Res* 1987;35(3):329–345.
- Brahimi N, Dauzere-Peres S, Najid NM, Nordli A. Single item lot sizing problems. *Eur J Oper Res* 2006;168(1):1–16.
- De Bodt MA, Gelders LF, van Wassenhove LN. Lot sizing under dynamic demand conditions: a review. *Eng Costs Prod Econ* 1984;8:165–187.
- Drexel A, Kimms A. Lot sizing and scheduling - survey and extensions. *Eur J Oper Res* 1997;99:221–235.
- Karimi B, Fatemi Ghomi SMT, Wilson JM. The capacitated lot sizing problem: a review of models and algorithms. *Omega* 2003;31:365–378.
- Kuik R, Salomon M, van Wassenhove LN. Batching decisions: structure and models. *Eur J Oper Res* 1994;18:243–263.

LOVÁSZ-SCHRIJVER REFORMULATION

MADHUR TULSIANI
Institute for Advanced Study,
Princeton, New Jersey

A common technique for solving a 0–1 integer program is to work with its linear programming (LP) relaxation, and then relate the optimum of the relaxation to that of the integer program. Consider an integer program with a feasible region described by the m inequalities of the form $\mathbf{a}_i \cdot \mathbf{x} - b_i \geq 0$ in variables $x_1, \dots, x_n \in \{0, 1\}$. In formulating a convex relaxation of this problem, one would like to optimize over the convex hull of the integral solutions to the above program (since for a linear objective function, the optimum over the integral solutions would be equal to the optimum over their convex hull). Most LP relaxations are obtained by simply relaxing the domain of the variables to be $[0, 1]$ instead of $\{0, 1\}$. However, in that case the resulting polytope may contain points that are outside this convex hull.

Starting from the works of Gomory [1] and Chvátal [2], a large amount of research has been done on developing methods that generate additional constraints valid for the integral solutions of the problem, to obtain tighter and tighter convex relaxations. Lovász and Schrijver [3] describe a powerful technique for generating additional inequalities. Given a program as above, let $\mathcal{P}$ denote the feasible polytope for the linear program obtained by relaxing the domain of the variables to be $[0, 1]$. Using the fact that any integer solution must satisfy $x_j^2 = x_j$ for all $j \in \{1, \dots, n\}$, it is easy to check that for $\lambda_{ij}, \sigma_{ij} \geq 0$, and $\mu_j \in \mathbb{R}$, any quadratic inequality of the form

$$\sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \lambda_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) x_j \\ + \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \sigma_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) (1 - x_j)$$

$$+ \sum_{1 \leq j \leq n} \mu_j (x_j^2 - x_j) \geq 0 \quad (1)$$

is also valid for any integer solution to the program. The Lovász–Schrijver reformulation automatically adds any *linear* inequalities which can be derived as above, to the linear program describing $\mathcal{P}$. They also define a stronger version that adds even more constraints giving a semidefinite program (SDP).

Thinking of the strengthening as an operator applied to a polytope, we can also apply it iteratively, yielding a sequence of stronger and stronger relaxations, often referred to as *hierarchies* of linear and semidefinite programs. We describe two such hierarchies defined by Lovász and Schrijver, known as the *LS* (linear) and *LS₊* (semidefinite) hierarchies. These hierarchies have been studied in different contexts, such as algorithmic analysis, computational complexity, and proof complexity, and somewhat different ways of looking at them were considered in each of these. The rest of this article is devoted to describing some of these viewpoints.

DESCRIPTION OF THE HIERARCHIES

To describe these hierarchies, it will be more convenient to work with *convex cones* rather than convex subsets of $[0, 1]^n$. Recall that a *cone* is a subset K of $\mathbb{R}^d$ such that if $\mathbf{x}, \mathbf{y} \in K$ and $\alpha, \beta \geq 0$ then $\alpha\mathbf{x} + \beta\mathbf{y} \in K$. If we are interested in a convex set $\mathcal{P} \subseteq [0, 1]^n$ (which might be the feasible region of our starting convex relaxation), we first convert it into a cone in $\mathbb{R}^{n+1}$. We define

$$\text{cone}(\mathcal{P}) = \{ \mathbf{y} = (\lambda, \lambda x_1, \dots, \lambda x_n) \mid \lambda \geq 0, \\ (x_1, \dots, x_n) \in \mathcal{P} \}.$$

It is easy to check that all the quadratic constraints of the form described in Equation (1) would follow if we could impose constraints of the form $x_j^2 = x_j$ for each $j \in \{1, \dots, n\}$. However, this constraint is neither linear nor convex. We thus impose linear constraints

implied by this, by introducing auxiliary variables $Y_{i,j}$, which are supposed to be equal to the product $x_i x_j$ when $\mathbf{y} = (1, x_1, \dots, x_n)$. We require that $Y_{i,i} = y_i$. Also, if Y_i denotes the i th row of the matrix Y , then it must be true that $(Y_i/y_i)_{1\dots n} \in \mathcal{P}$ when $y_i \neq 0$, where by $(Y_i/y_i)_{1\dots n}$ we denote the vector consisting of the last n coordinates of (Y_i/y_i) . This is equivalent to saying that $Y_i \in \text{cone}(\mathcal{P})$. Similarly, $y_i \neq y_0$ must imply $((Y_0 - Y_i)/(y_0 - y_i))_{1\dots n} \in \mathcal{P}$ and hence $Y_0 - Y_i \in \text{cone}(\mathcal{P})$. Lovász and Schrijver use this to define two operators N and N_+ , such that for a starting cone K , $N_+(K) \subseteq N(K) \subseteq K$.

Definition 1. For a cone $K \subseteq \mathbb{R}^{n+1}$ we define the set $N(K)$ (also a cone in $\mathbb{R}^{n+1}$) as follows: a vector $\mathbf{y} = (y_0, \dots, y_n) \in \mathbb{R}^{n+1}$ is in $N(K)$ iff there is a matrix $Y \in \mathbb{R}^{(n+1) \times (n+1)}$ such that

1. Y is symmetric.
2. For every $i \in \{0, 1, \dots, n\}$, $Y_{0,i} = Y_{i,i} = y_i$.
3. Each row Y_i is an element of K .
4. Each vector $Y_0 - Y_i$ is an element of K .

In such a case, Y is called the *protection matrix* of $\mathbf{y}$. If, in addition, Y is positive semidefinite, then $\mathbf{y} \in N_+(K)$. We define $N^0(K)$ and $N_+^0(K)$ as K , and $N^t(K)$ ($N_+^t(K)$) as $N(N^{t-1}(K))$ (respectively, $N_+(N_+^{t-1}(K))$).

Remark 1. Lovász and Schrijver also define a third operator which they denote by N_0 , which is weaker than the ones described here. It is defined in a manner similar to the N and N_+ operators, except that the protection matrix Y is not required to be symmetric. However, we shall not discuss this here.

If $\mathbf{y} = (\lambda, \lambda x_1, \dots, \lambda x_n) \in \mathbb{R}^{n+1}$ for $x_1, \dots, x_n \in \{0, 1\}$, then we can set $Y_{i,j} = \lambda \cdot x_i x_j$. Such a matrix Y is clearly positive semidefinite, and it satisfies $Y_{i,i} = \lambda x_i^2 = \lambda x_i = y_i$. Also $Y_i = x_i \cdot \mathbf{y}$ and $Y_0 - Y_i = (1 - x_i) \cdot \mathbf{y}$. Hence they lie in the same cone as $\mathbf{y}$.

In the case when $\mathcal{P} \in \mathbb{R}^n$ is a polytope defined by linear constraints and $\text{cone}(\mathcal{P})$ is the corresponding cone, we often use

the notation $N^t(\mathcal{P})$ and $N_+^t(\mathcal{P})$ to denote $N^t \text{cone}(\mathcal{P}) \cap \{y_0 = 1\}$ and $N_+^t \text{cone}(\mathcal{P}) \cap \{y_0 = 1\}$, respectively. The following claim characterizes $N(\mathcal{P})$ and $N_+(\mathcal{P})$ in terms of the additional inequalities imposed by the reformulation.

Claim 1. Let $\mathcal{P} \in \mathbb{R}^n$ be a polytope defined by the constraints $\mathbf{a}_i \cdot \mathbf{x} - b_i \geq 0$ for $i \in \{1, \dots, m\}$. Then for a vector $\mathbf{x} \in \mathbb{R}^n$,

1. $\mathbf{x} \in N(\mathcal{P})$ iff it satisfies all linear equalities $\mathbf{c} \cdot \mathbf{x} - d \geq 0$ which can be derived as

$$\begin{aligned} \mathbf{c} \cdot \mathbf{x} - d = & \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \lambda_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) x_j \\ & + \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \sigma_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) (1 - x_j) \\ & + \sum_{1 \leq j \leq n} \mu_j (x_j^2 - x_j) \end{aligned}$$

for $\lambda_{ij}, \sigma_{ij} \geq 0$ and $\mu_j \in \mathbb{R}$.

2. $\mathbf{x} \in N_+(\mathcal{P})$ iff it satisfies all linear equalities $\mathbf{c} \cdot \mathbf{x} - d \geq 0$ which can be derived as

$$\begin{aligned} \mathbf{c} \cdot \mathbf{x} - d = & \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \lambda_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) x_j \\ & + \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \sigma_{ij} (\mathbf{a}_i \cdot \mathbf{x} - b_i) (1 - x_j) \\ & + \sum_{1 \leq j \leq n} \mu_j (x_j^2 - x_j) + \sum_k (\mathbf{g}_k \cdot \mathbf{x} - h_k)^2 \end{aligned}$$

for $\lambda_{ij}, \sigma_{ij} \geq 0$, $\mu_j, h_k \in \mathbb{R}$ and $\mathbf{g}_k \in \mathbb{R}^n$.

The characterization in terms of the inequalities is particularly useful when reasoning about these hierarchies as proof systems for (say) proving that a certain integer program has no integral solutions. The rules for deriving the inequalities characterizing $N(\mathcal{P})$ from those characterizing $\mathcal{P}$ can then be thought of as the derivation rules of this proof system. One is then interested in the smallest t such that $N^t(\mathcal{P})$ is empty (called the *rank* of the

proof system) and in the minimum number of inequalities needed (called the *size* of the proof) to derive a convex system of inequalities that has no feasible solution. We discuss two different ways of viewing Definition 1 below.

The Prover–Adversary Game

It is sometimes convenient to think of the LS and LS_+ hierarchies in terms of a game between two parties called a *prover* and an *adversary*. This formulation was first used by Buresh-Oppenheim *et al.* [4] who used it to prove lower bounds on the LS_+ procedure as a proof system. For example, the following is a formulation that is convenient to use for proving that a certain vector belongs to $N_+^t(K)$ for a cone $K \subseteq \mathbb{R}^{n+1}$. The formulation remains the same for talking about $N(K)$, except that we omit the positive semidefiniteness constraint in the matrix below.

A *prover* is an algorithm that, on an input vector $(y_0, \dots, y_n)$, either fails or outputs a matrix $Y \in \mathbb{R}^{(n+1) \times (n+1)}$ and a set of vectors $O \subseteq \mathbb{R}^{n+1}$ such that

1. Y is symmetric and positive semidefinite.
2. $Y_{i,i} = Y_{0,i} = y_i$.
3. Each vector Y_i and $Y_0 - Y_i$ is a nonnegative linear combination of vectors of O .
4. Each element of O is in K .

Now, consider the following game played by a prover against another party called the *adversary*. We start from a vector $\mathbf{y} = (y_0, \dots, y_n)$, and the prover, on input $\mathbf{y}$, outputs Y and O as before. Then the adversary chooses a vector $\mathbf{z} \in O$, and the prover, on input $\mathbf{z}$, outputs a matrix Y' and a set O' , and so on. The adversary *wins* when the prover fails to output a matrix and vectors with the required properties.

Lemma 1. *Suppose that there is a prover such that, starting from a vector $\mathbf{y} \in K$, every adversary strategy requires at least $t + 1$ moves to win. Then $\mathbf{y} \in N^t(K)$.*

Proof. We proceed by induction on t , with $t = 0$ being the simple base case. Suppose that,

for every adversary, it takes at least $t + 1$ moves to win, and let Y and O be the outputs of the prover on input $\mathbf{y}$. Then, for every element $\mathbf{z} \in O$, and every prover strategy, it takes at least t moves to win starting from $\mathbf{z}$. By the inductive hypothesis, each element of O is in $N_+^{t-1}(K)$, and since $N_+^{t-1}(K)$ is closed under nonnegative linear combinations, the vectors Y_i and $Y_0 - Y_i$ are all in $N_+^{t-1}(K)$, and so Y is a protection matrix that shows that $\mathbf{y}$ is in $N_+^t(K)$.

While the framework is simply a restatement of the conditions defining $N(K)$ and $N_+(K)$, it turns out to be surprisingly convenient to think of many arguments in terms of strategies for this game. Also, note that we added the set O in the description of the conditions. In many proofs, it is convenient to think of O as a set of vectors satisfying some structural invariant. Often it may be easier to exhibit matrices Y for vectors satisfying some invariants, than for arbitrary vectors in K .

The View in Terms of Local Probability Distributions

Yet another view of the conditions which has been useful in reasoning about optimization problems is the interpretation of solutions as “locally conditioned probability distributions.” Let $\mathbf{x} \in \mathcal{P}$ be expressible as $\mathbf{x} = \sum_i \lambda_i \mathbf{z}^{(i)}$ where $\sum_i \lambda_i = 1$ and $\forall i, \mathbf{z}^{(i)} \in \mathcal{P} \cap \{0, 1\}^n, \lambda_i \geq 0$. Then, we can consider a random variable $\mathbf{z}$ which takes value $\mathbf{z}^{(i)}$ with probability λ_i . For $j \in \{1, \dots, n\}$, the numbers x_j then are equal to $\mathbf{P}[z_j = 1]$, that is, the marginals of this distribution.

To prove that $\mathbf{x} \in \mathcal{P}$, we then require a matrix $Y \in \mathbb{R}^{n+1}$ which satisfies the conditions stated in the section titled “The Prover–Adversary Game”. For each $\mathbf{z}^{(i)}$ such a matrix $Y^{(i)}$ can be given as $Y^{(i)} = (1, \mathbf{z}^{(i)})(1, \mathbf{z}^{(i)})^T$ where $(1, \mathbf{z}^{(i)}) \in \mathbb{R}^{n+1}$. The matrix Y for $\mathbf{x}$ can then be exhibited as $Y = \sum_i \lambda_i Y^{(i)}$, where each entry $Y_{ij} = \mathbf{P}[(z_i = 1) \wedge (z_j = 1)]$. Arguing that the vector $Y_i \in \text{cone}(\mathcal{P})$ is then equivalent to arguing that the vector $\mathbf{x}^{(i,1)} \in \mathbb{R}^n$ with

$$x_j^{(i,1)} = Y_{ij}/x_i = \mathbf{P}[z_j = 1 | z_i = 1]$$

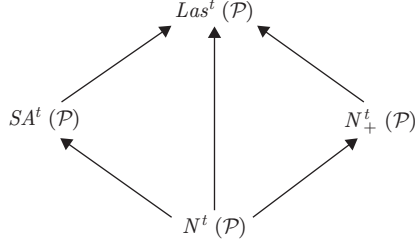


Figure 1. A comparison of the hierarchies.

is in $\mathcal{P}$. Similarly, $Y_0 - Y_i \in \text{cone}(\mathcal{P})$ is equivalent to proving that the vector $\mathbf{x}^{(i,0)}$ with coordinates $x_j^{(i,0)} = (Y_{0j} - Y_{ij})/(1 - x_i) = \mathbf{P}[z_j = 1 | z_i = 0]$ is in $\mathcal{P}$.

Thus, the conditions for membership in $N(\mathcal{P})$ and $N_+(\mathcal{P})$ can be thought of as conditions for checking that a vector $\mathbf{x}$ indeed represents the marginals of a distribution over $\mathcal{P} \cap \{0, 1\}^n$. The test is the existence of conditional distributions, conditioned on each coordinate i being either 0 or 1. Similarly, exhibiting membership in $N^t(\mathcal{P})$ then corresponds to exhibiting distributions after conditioning on any t variables instead of just one.

Combining this with the prover–adversary framework, we can think that at each step, the adversary conditions a particular variable to be 0 or 1. The task of the prover is then to provide a matrix corresponding to the conditional distribution. We must emphasize that this is *not an exact characterization* of the hierarchies, but this intuition has been quite useful in arguing about their performance on optimization problems, as discussed in the section titled “Hardness Results for Optimization Problems.”

Relation to Sherali–Adams and Lasserre Hierarchies

Hierarchies of relaxations similar to the ones by Lovász and Schrijver were also defined by Sherali and Adams [5] and Lasserre [6]. For a starting linear program with feasible region $\mathcal{P}$, let $\text{SA}^t(\mathcal{P})$ and $\text{Las}^t(\mathcal{P})$ denote the feasible regions of the relaxations obtained at the t th level of these hierarchies. The set $\text{SA}^t(\mathcal{P})$ can be described as the projection of the feasible region of a linear program in

$n^{O(t)}$ variables, while $\text{Las}^t(\mathcal{P})$ is the projection of the feasible set for a semidefinite program in $n^{O(t)}$ dimensions.

It is also known that $\text{SA}^t(\mathcal{P}) \subseteq N^t(\mathcal{P})$ and $\text{Las}^t(\mathcal{P}) \subseteq N_+^t(\mathcal{P})$ and thus the Sherali–Adams and Lasserre hierarchies provide tighter relaxations, respectively, than the LS and LS_+ hierarchies. Figure 1 shows the relationships between these different hierarchies, with an arrow $A \rightarrow B$ indicating that $A \supseteq B$ and thus B is tighter than A . A more detailed comparison can be found in the excellent survey by Laurent [7].

ALGORITHMIC ASPECTS AND APPLICATIONS

We first note that if K is defined by a linear program with m constraints in the $n + 1$ variables $y_0, \dots, y_n$, then $N(K)$ is defined by a linear program with $O(mn)$ constraints in the $O(n^2)$ variables corresponding to the entries of the matrix Y . Similarly, $N_+(K)$ is defined by a semidefinite program in Y with $O(mn)$ constraints. More generally, Lovász and Schrijver show the following.

Theorem 1. *Given a weak separation oracle for the cone K , the weak separation problem for $N(K)$ and $N_+(K)$ can be solved in polynomial time.*

Hence, whenever it is possible to easily optimize over the cone K , one can also optimize over the cone $N(K)$ (or $N_+(K)$) with a polynomial overhead. Using the above, it is also easy to deduce that weak separation oracles for $N^t(K)$ and $N_+^t(K)$ can be implemented in time $n^{O(t)}$, given one for K running in polynomial time. For a convex region $\mathcal{P} \subseteq \mathbb{R}^n$, let $K_I \subseteq \mathbb{R}^{n+1}$ denote the cone generated by the integral vectors $\{(1, \mathbf{x}) | \mathbf{x} \in \mathcal{P} \cap \{0, 1\}^n\}$ and let $K = \text{cone}(\mathcal{P})$. Then, it is known that n applications of $N(\cdot)$ are sufficient to converge to K_I .

Theorem 2. $N^n(K) = N_+^n(K) = K_I$.

However, note that n applications take time $n^{O(n)}$, which is super-exponential in n . For general cones, this bound is known to be tight for both N and N_+ operators, as was shown by

Cook and Dash [8] and also by Goemans and Tunçel [9]. Typically, the interesting question is whether, for a given problem, optimizing over $N^t(K)$ or $N_+^t(K)$ for a small t can give some nontrivial approximation guarantee.

Application to Independent Set

As an example, we consider the constraints obtained by applying the $N(\cdot)$ and $N_+(\cdot)$ operators to the feasible set of the LP relaxation for maximum independent set, which is the problem of finding the largest subset of vertices in a graph not containing an edge (also known as the *maximum stable set* problem). Let $G = (V, E)$ be a graph on n vertices. Then, the linear programming relaxation for the maximum independent set problem has variables $x_1, \dots, x_n \in [0, 1]$ and constraints $x_i + x_j \leq 1$ for all $(i, j) \in E$. We denote the feasible region of these constraints as $\text{IS}(G)$. The objective is to maximize the sum $\sum_i x_i$.

The set cone $\text{IS}(G)$ is then given by the inequalities $0 \leq y_i \leq y_0$ for all $i \in \{1, \dots, n\}$ and $y_i + y_j \leq y_0$ for all $(i, j) \in E$. We are interested in the sets given by $N(\text{IS}(G))$ and $N_+(\text{IS}(G))$.

There are also various additional constraints that are satisfied by an integer solution to the independent set problem. Lovász and Schrijver consider the so-called *odd-cycle constraints* and *clique constraints* and show that they are implied by one application of $N(\cdot)$ and $N_+(\cdot)$, respectively. The former type of constraints say that if C is a set of vertices forming a cycle of odd length, then $\sum_{i \in C} x_i \leq (|C| - 1)/2$. The latter require that if B is a set of vertices forming a clique, then $\sum_{i \in B} x_i \leq 1$. We give a proof below for the sake of illustration.

Claim 2. *For a graph $G = (V, E)$, let $\mathbf{x} = (x_1, \dots, x_n)$ be a vector in $\text{IS}(G)$. Then*

1. *If $\mathbf{x} \in N(\text{IS}(G))$, then for every odd cycle C in G , $\sum_{i \in C} x_i \leq (|C| - 1)/2$.*
2. *If $\mathbf{x} \in N_+(\text{IS}(G))$, then for every clique B in G , $\sum_{i \in B} x_i \leq 1$.*

Proof. For this part, it is convenient to think of $N(\text{IS}(G))$ in terms of the rules for deriving

new inequalities by a combination of old ones, as given in Claim 1. Let $C = \{i_1, \dots, i_k\}$ be a cycle of odd length in the graph. Since $i_1, \dots, i_{k-1}$ can be covered by $(k-1)/2$ edges, adding their inequalities we have $x_{i_1} + \dots + x_{i_{k-1}} \leq (k-1)/2$. Similarly, $x_{i_2} + \dots + x_{i_{k-2}} \leq (k-3)/2$. Using these and Claim 1, we get that $(x_1, \dots, x_n)$ must satisfy

$$\begin{aligned} & \left(\sum_{j=1}^{k-1} x_{i_j} - \frac{(k-1)}{2} \right) (1 - x_{i_k}) \\ & + \left[\left(\sum_{j=2}^{k-2} x_{i_j} - \frac{(k-3)}{2} \right) + (x_{i_1} + x_{i_k} - 1) \right. \\ & \quad \left. + (x_{i_{k-1}} + x_{i_k} - 1) \right] x_{i_k} \\ & - 2(x_{i_k}^2 - x_{i_k}) \leq 0 \end{aligned}$$

Simplifying the above equation yields $\sum_{j=1}^k x_{i_j} \leq (k-1)/2$.

For the second part, it is more convenient to think of $\mathbf{x} \in N_+(\text{IS}(G))$ in terms of the protection matrix Y certifying that $\mathbf{y} = (1, x_1, \dots, x_n) \in N(\text{cone}(\text{IS}(G)))$, as in Definition 1. Since the matrix must be positive semidefinite, we consider its Gram decomposition, say given by the vectors $\mathbf{v}_0, \dots, \mathbf{v}_n$. The conditions $Y_{i,0} = Y_{i,i} = y_i$ imply that $|\mathbf{v}_0|^2 = 1$ and $\langle \mathbf{v}_0, \mathbf{v}_i \rangle = \|\mathbf{v}_i\|^2 = x_i$.

If (i, j) is an edge in the graph, then since Y_i must be in the cone $\text{IS}(G)$, we get $Y_{i,i} + Y_{i,j} \leq Y_{i,0}$ and hence $Y_{i,j} \leq 0$. However, since a vector in the cone must have nonnegative entries, we must have $Y_{i,j} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for any edge (i, j) . Thus, all the vectors corresponding to vertices in a clique must be orthogonal. By Pythagoras' theorem,

$$\begin{aligned} \sum_{i \in B, |\mathbf{v}_i| > 0} \frac{\langle \mathbf{v}_0, \mathbf{v}_i \rangle^2}{|\mathbf{v}_i|^2} & \leq |\mathbf{v}_0|^2 \implies \sum_{i \in B, x_i > 0} \frac{x_i^2}{x_i} \leq 1 \\ & \implies \sum_{i \in B} x_i \leq 1. \end{aligned}$$

The graphs for which the clique constraints exactly characterize the convex hull of 0–1 independent sets are called *perfect* graphs. Hence, the above equation shows that

optimizing over $N_+(\text{IS}(G))$, one finds exactly the size of the maximum independent set for a perfect graph G . Similarly, by optimizing over $N(\text{IS}(G))$, one finds the maximum independent set in graphs for which the odd-cycle constraints characterize the 0–1 independent sets. Such graphs are known as *h-perfect*.

Deriving the SDP relaxation for Sparsest Cut

Alekhovich *et al.* [10] sketch how the well-known semidefinite programming relaxation for sparsest cut by Arora *et al.* [11] can be derived by three levels of the LS_+ hierarchy starting from a basic linear program. We give a proof below showing that in fact, one level of the LS_+ hierarchy already derives the relaxation in Arora *et al.* [11].

Sparsest cut is the problem of finding a cut $(S, \bar{S})$ minimizing $|E(S, \bar{S})|/|S|$, the ratio of the number of outgoing edges to the size of S . In the discussion below, we instead describe the argument for the c -balanced separator problem, which is known to be closely related to sparsest cut [11–13]. Instead of asking for the sparsest cut of any size, the c -balanced separator problem asks for the sparsest cut with at least cn vertices on each side of the cut. For a graph $G = (V, E)$ on n vertices, the linear program involves $n + \binom{n}{2}$ variables: $x_i \in [0, 1]$ for each vertex i , representing which side of the cut it is, and $d_{ij} \in [0, 1]$ for every pair $\{i, j\}$ of vertices representing the distance between them (which in an integral solution is 0 if they are on the same side and 1 otherwise).

The constraint $\sum d_{ij} \geq c(1 - c)n^2$ imposes the condition (on integer solutions) that the number of vertices on each side of the cut is at least cn . Note that the program mentioned in Fig. 2 is quite weak (in fact its integrality gap, as defined in the section titled “Hardness Results for Optimization Problems” is infinite). This can be seen by using the solution $x_i = 1/2$ for all i and $d_{ij} = 0$ when $(i, j) \in E$ and $d_{ij} = 1$ otherwise. This is a feasible solution with objective value 0 for any graph in which $|E| \leq \binom{n}{2} - c(1 - c)n^2$, while the integer optimum may be arbitrarily large. Adding the inequality $d_{ij} + d_{jk} \geq d_{ik}$ for every triple i, j, k makes it at least as strong as the linear program used by Leighton and

Rao [12,13], which was shown to achieve an $O(\log n)$ approximation. However, even without this added inequality, one application of the N_+ operator derives the relaxation in Arora *et al.* [11], which achieves an $O(\sqrt{\log n})$ approximation.

Arora *et al.* [11] use a semidefinite program with $n \times n$ semidefinite matrix Y whose Gram decomposition is given by the vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. A solution to their program would be a feasible solution to the above LP if we take $x_i = |\mathbf{v}_i|^2$ and $d_{ij} = |\mathbf{v}_i - \mathbf{v}_j|^2$. They also impose the additional inequality that $|\mathbf{v}_i - \mathbf{v}_j|^2 + |\mathbf{v}_j - \mathbf{v}_k|^2 \geq |\mathbf{v}_i - \mathbf{v}_k|^2$. We argue below that this inequality is derived by one application of N_+ .

Claim 3. *Let $(1, \mathbf{x}, \mathbf{d}) \in N_+(\text{cone}(BS(G)))$ and let Y be its protection matrix with the Gram decomposition $\{\mathbf{v}_0, \{\mathbf{v}_i\}_{i \in V}, \{\mathbf{v}_{ij}\}_{\{i,j\} \in \binom{V}{2}}\}$. Then the vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ satisfy for all i, j, k*

$$d_{ij} = |\mathbf{v}_i - \mathbf{v}_j|^2 \quad \text{and} \quad |\mathbf{v}_i - \mathbf{v}_j|^2 + |\mathbf{v}_j - \mathbf{v}_k|^2 \geq |\mathbf{v}_i - \mathbf{v}_k|^2.$$

Proof. If $x_1, \dots, x_n$ were indeed an integer solution, we would have $d_{ij} = (x_i - x_j)^2$, which is 1, when exactly one of x_i, x_j is 1. To derive this for an integer solution, we use $d_{ij} \leq (1 - x_i) + (1 - x_j)$ and $d_{ij} \leq x_i + x_j$ to get $d_{ij}x_j \leq (2 - x_i - x_j)x_j$ and $d_{ij}(1 - x_j) \leq (x_i + x_j)(1 - x_j)$. Adding these two, and using $x_i^2 = x_i$ and $x_j^2 = x_j$, we get $d_{ij} \leq (x_i - x_j)^2$. Similarly, starting with the inequalities $d_{ij} \geq x_i - x_j$ and $d_{ij} \geq x_j - x_i$, we can get $d_{ij} \geq (x_i - x_j)^2$.

The proof essentially simulates the “linearized” version of this reasoning, replacing a quadratic term $x_i x_j$ by the entry $Y_{i,j}$ of the matrix. We also use $Y_{i,jk}$ to refer to the entry of the matrix corresponding to variables x_i and d_{jk} . Since $Y_j \in \text{cone}(BS(G))$ and $Y_0 - Y_j \in \text{cone}(BS(G))$, we get

$$Y_{j,ij} \leq 2Y_{j,0} - Y_{j,i} - Y_{j,j} \quad \text{and} \\ Y_{0,ij} - Y_{j,ij} \leq (Y_{0,i} - Y_{j,i}) + (Y_{0,j} - Y_{j,j}).$$

Again, adding the two and using $Y_{i,i} = Y_{0,i}$ and $Y_{j,j} = Y_{0,j}$ we get

$$d_{ij} = Y_{0,ij} \leq Y_{i,i} + Y_{j,j} - 2Y_{i,j} = |\mathbf{v}_i - \mathbf{v}_j|^2.$$

$$\begin{array}{ll}
\text{minimize} & \sum_{(i,j) \in E} d_{ij} \\
\text{subject to} & d_{ij} \leq x_i + x_j \quad \forall \{i, j\} \\
& d_{ij} \leq (1 - x_i) + (1 - x_j) \quad \forall \{i, j\} \\
& d_{ij} \geq x_i - x_j \quad \forall \{i, j\} \\
& d_{ij} \geq x_j - x_i \quad \forall \{i, j\} \\
& \sum_{\{i,j\} \in \binom{V}{2}} d_{ij} \geq c(1-c)n^2 \\
& x_i, d_{ij} \in [0, 1]
\end{array}$$

Figure 2. LP relaxation for balanced separator.

The proof of the inequality in the other direction is similar.

For the second part, we again note for an integer solution, $d_{ij} = (x_i - x_j)^2 = x_i(1 - x_j) + x_j(1 - x_i)$. Hence, it would suffice to prove $x_j[(1 - x_i) + (1 - x_k)] + (1 - x_j)[x_i + x_k] \geq d_{ik}$. However, this follows by using $x_j[(1 - x_i) + (1 - x_k) - d_{ik}] \geq 0$ and $(1 - x_j)[x_i + x_k - d_{ik}] \geq 0$. For the matrix version, we again use $Y_j, Y_0 - Y_j \in \text{cone}(\text{BS}(G))$ to get

$$\begin{aligned}
2Y_{j,0} - Y_{j,i} - Y_{j,k} &\geq Y_{j,ik} \quad \text{and} \\
(Y_{0,i} - Y_{j,i}) + (Y_{0,k} - Y_{j,k}) &\geq (Y_{0,ik} - Y_{j,ik}).
\end{aligned}$$

Adding the two and replacing every entry $Y_{i,j}$ by $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ we get $|\mathbf{v}_i - \mathbf{v}_j|^2 + |\mathbf{v}_j - \mathbf{v}_k|^2 \geq |\mathbf{v}_i - \mathbf{v}_k|^2$.

Finding Dense Subgraphs in a Graph

Both the above arguments start with a polytope $\mathcal{P}$ and only consider $N(\mathcal{P})$ and $N_+(\mathcal{P})$. Tighter relaxations, which are still efficient, can be obtained by considering $N^t(\mathcal{P})$ and $N_+^t(\mathcal{P})$, where $t = O(1)$ or even when $t = \omega(1)$ is a slowly growing function of n . However, not many arguments are known which use these, as the sets $N^t(\mathcal{P})$ or $N_+^t(\mathcal{P})$ are often difficult to understand for large t .

Very recently, an application for the densest k -subgraph problem was obtained by Bhaskara *et al.* [14]. Given a graph G and a size parameter k , the problem requires finding a subgraph of k vertices in G , with the maximum number of edges. The previously

best known approximation factor for this problem was $O(n^{0.3159})$ [15,16]. The authors achieve an approximation factor of $O(n^{1/4+\epsilon})$ using $N^{O(1/\epsilon)}(\cdot)$.

Starting with a basic linear program, they strengthen it to a linear program implied by $O(1/\epsilon)$ applications of $N(\cdot)$. The strengthening uses intuition of the solution as a distribution over dense subgraphs. The constraints require that when one conditions $O(1/\epsilon)$ vertices to be in the subgraph, they all must have large expected degree according to the conditional distribution. The full algorithm, however, is significantly more involved and uses various other combinatorial arguments in addition to the LP.

HARDNESS RESULTS FOR OPTIMIZATION PROBLEMS

As discussed in the algorithmic applications, for hard optimization problems, one is often interested in how well the relaxation obtained by $N^t(\mathcal{P})$ approximates the value of the integer program. For a maximization problem, if FRAC denotes the value of a relaxation, and OPT denotes the combinatorial optimum for the optimization problem (i.e., the maximum objective value for an integer solution), then we define the *integrality gap* of the relaxation as the *supremum* of the ratio FRAC/OPT over all instances of the problem. For a minimization problem, we take it to be the inverse ratio. When the integrality gap varies with the

size of the problem, we take the supremum over all instances of size n and express it as a function of n . Note that according to our definition, the integrality gap is always larger than 1, and the closer it is to 1, the better is the performance of the convex relaxation.

From the perspective of complexity theory, one can view the hierarchies of programs defined by the N and N_+ operators as restricted models of computation, with the number of applications of these operators as a resource. This also corresponds naturally to time of computation in this model as optimizing over $N^t(\mathcal{P})$ takes time $n^{O(t)}$. One is then interested in the trade-off between the level of the relaxation in the hierarchy (denoted by t) and the integrality gap of the relaxation. A lower bound in this computational model corresponds to showing that the integrality gap remains large even for the relaxations obtained at very high levels of the hierarchy.

Integrality Gaps for Vertex Cover

Minimum Vertex Cover, the problem of finding the smallest subset of vertices in a graph such that every edge is incident on some vertex in the subset, is perhaps the most studied problem with regard to integrality gaps in the Lovász–Schrijver hierarchy. The study of integrality gaps in this model was initiated by the works of Arora *et al.* [17,18]. They showed that the integrality gap remains at least $2 - \epsilon$ even after $\Omega(\log n)$ levels of the hierarchy. Since the integrality gap can be easily shown to be *at most* 2 even for the starting linear relaxation, this showed that even the use of time $n^{O(\log n)}$ in this model yields no significant improvement. The results were later improved to $3/2$ for $\Omega(\log^2 n)$ levels by Tzourakis [19] and to an optimal lower bound of $2 - \epsilon$ for $\Omega(n)$ levels by Schoenebeck *et al.* [20]. The last result even rules out exponential time algorithms (note that $\Omega(n/\log n)$ levels would correspond to $2^{\Omega(n)}$ time) in this model.

These results work implicitly or explicitly with the “locally conditioned distributions” interpretation discussed in the section titled “The View in Terms of Local Probability Distributions.” One considers sparse random

graphs that have no cycles of size less than $\Omega(\log n)$ so that any subgraph with $O(\log n)$ vertices is a tree. One then starts with a solution that has fractional value $1/2 + \epsilon$ on every vertex. The move of the adversary then corresponds to selecting a vertex where the solution has a fractional value, and fixing it to 1 or 0, that is, conditioning it to be in or out of the vertex cover. As long as the adversary conditions on $O(\log n)$ vertices, the set of conditioned vertices form a tree, restricted to which there is an *actual distribution of vertex covers* with marginal values $1/2 + \epsilon$. Hence, one can use this to produce the required conditional distribution. We remark that this is just an intuition for the proof in Arora *et al.* [18]. The actual proof proceeds by looking at the duals of the linear programs obtained by the LS hierarchy and involves significantly more work. The result in Schoenebeck *et al.* [20] for $\Omega(n)$ levels uses an explicit version of the above intuitive argument together with a more careful use of the sparsity of these graphs.

Integrality gaps for the semidefinite version of this hierarchy have been somewhat harder to prove. It was shown by Goemans and Kleinberg [21] that the integrality gap for the relaxation obtained by one application of N_+ (which is equivalent to the ϑ -function of Lovász) is at least $2 - \epsilon$. The result was strengthened by Charikar [22], who showed that the same gap holds even when the relaxation is augmented with “triangle inequalities” discussed in the section titled “Deriving the SDP Relaxation for Sparsest Cut.” The gap was extended to $\Omega(\sqrt{\log n}/\log \log n)$ levels of the LS_+ hierarchy by Georgiou *et al.* [23]. Interestingly, all these results for LS_+ were proven for the same family of graphs, inspired by a paper of Frankl and Rödl [24]. It is an interesting problem to construct an alternate family which is also an integrality gap instance, or to extend the above results to even $\Omega(\log n)$ levels.

Somewhat incomparable to the above results, integrality gaps of factors $7/6$ for $\Omega(n)$ levels and 1.36 for $n^{\Omega(1)}$ levels of the semidefinite hierarchy were obtained by Schoenebeck *et al.* [25] and Tulsiani [26]. These results also differ from the ones discussed above in that they do not directly

exhibit a family of integrality gap instances for vertex cover. Instead, they start with an integrality gap instance for a constraint satisfaction problem, and proceed by using a reduction from the constraint satisfaction problem to vertex cover.

Results for Constraint Satisfaction Problems

For constraint satisfaction problems with three or more variables in each constraint, strong integrality gaps have been shown for the Lovász–Schrijver semidefinite (and hence also the linear) hierarchy. For these problems, one studies how well convex relaxations approximate the maximum number of constraints that can be satisfied by any assignment to the variables.

Optimal integrality gaps of a factor $2^k/(2^k - 1)$ for K–SAT with $k \geq 5$ and $\Omega(n)$ levels were shown by Buresh-Oppenheim *et al.* [4]. These were later extended to the important remaining case of 3–SAT by Alekhnovich *et al.* [10], who also proved strong lower bounds for approximating minimum vertex cover in hypergraphs. $\Omega(n)$ level integrality gaps for other constraint satisfaction problems also follow from the corresponding results in the Lasserre hierarchy by Schoenebeck [27] and Tulsiani [26]. All these results crucially rely on the ideas of expansion from proof complexity, explained in the section titled “Proof Complexity.”

Results for Other Problems

The performance of the LS hierarchy for the traveling salesman problem was considered by Cook and Dash [8] and by Cheung [28]. A basic LP relaxation for the problem with n cities involves $n(n - 1)/2$ variables and is due to Dantzig, Fulkerson, and Johnson. This relaxation is known to have integrality gap of at least $4/3$. Cook and Dash showed that exactly characterizing the integer optimum requires at least $\lceil n/8 \rceil$ levels of the LS_+ hierarchy. They also exhibited an inequality valid for the integer solutions, such that any LS_+ -derivation of it requires to use at least $2^{n/8}/n^4$ intermediate inequalities. Cheung showed that the integrality gap of the relaxation remains at least $4/3$ after any constant number of levels in the LS hierarchy.

Feige and Krauthgamer [29] studied how well the programs in the Lovász–Schrijver hierarchy do in determining the size of the largest clique in random graphs. Let $G_{n,1/2}$ denote the probabilistic model for generating a random graph by picking each pair of vertices to be an edge independently with probability $1/2$. They were motivated by the the problem of distinguishing a random graph in $G_{n,1/2}$ from the one in which a clique of size k ($\approx \sqrt{n}$) has been planted, which has been suggested as the basis of a cryptographic scheme [30,31].

They showed that the value of the relaxation for maximum clique obtained by N_+^t is almost surely between $\sqrt{n/(2 + \epsilon)^{t+1}}$ and $4\sqrt{n/(2 - \epsilon)^{t+1}}$ for any $\epsilon \in (0, 2)$ and $t = o(\log n)$. Hence, their result shows that one can use $N_+^t(\cdot)$ to find planted cliques of size greater than $4\sqrt{n/(2 - \epsilon)^t}$, but not much smaller than that.

Also, it is well known [32] that the size of the largest clique in a random graph generated according to $G_{n,1/2}$ is almost surely equal to $(2 + o(1))\log_2 n$. It was shown by Lovász and Schrijver [3] that if largest clique in G has size $\omega(G)$, then the value of the relaxation obtained by $t = \omega(G)$ levels of the LS_+ hierarchy is exactly $\omega(G)$. Hence, while the relaxation obtained by $(2 + o(1))\log n$ levels of the hierarchy is almost surely tight, any relaxation obtained by $t = o(\log n)$ levels has a large integrality gap.

PROOF COMPLEXITY

Proof complexity studies the efficiency with which the proof of a given proposition can be derived by a proof system. As discussed earlier, the Lovász–Schrijver hierarchies can be viewed as a proof system, where one starts with a given set of inequalities defining a convex region and derives new ones using the rules in Claim 1. This proof system is known to be at least as powerful as the well-known resolution proof system [33].

Consider a formula φ in conjunctive normal form (CNF), the feasibility of which can be phrased as an integer program in n variables. If $\mathcal{P}$ is the feasible polytope for an LP relaxation of this program, and if φ is unsatisfiable, then we must have $\mathcal{P} \cap \{0, 1\}^n = \emptyset$.

Hence, t applications of $N(\cdot)$ will *prove* that φ is unsatisfiable, if $N^t(\mathcal{P}) = \emptyset$. The smallest t for which $N^t(\mathcal{P})$ (or $N_+^t(\mathcal{P}) = \emptyset$) is known as the N -rank (respectively N_+ -rank) of φ . For proving a lower bound on proof complexity, one is then interested in exhibiting a family of formulas with large (typically $\Omega(n)$) rank.

The techniques developed for proving lower bounds in proof complexity have also been useful in reasoning about general constraint satisfaction problems, and in proving hardness of approximating various optimization problems (using convex relaxations), as discussed in the section titled “Hardness Results for Optimization Problems.”

Expansion and Rank Lower Bounds for Formulas

An important technique for proving lower bounds in proof complexity has been the use of *expansion* in formulas. It was used for proving that if φ is an unsatisfiable 3-CNF formula in n variables, with each set of s clauses (for say $s \leq n/1000$) involving at least (say) $3s/2$ variables, then any proof of the unsatisfiability of φ in the *resolution* proof system must have exponential size [34,35].

Rank lower bounds for LS-proof systems were first proved by Grigoriev *et al.* [36]. Buresh-Oppenheim *et al.* [4] were the first to use expansion arguments in the context of Lovász–Schrijver proof systems. They also proved lower bounds for various optimization problems, which were later extended by Alekhovich *et al.* [10]. Both these papers phrase their arguments in terms of the prover–adversary game discussed in the section titled “The Prover–Adversary Game.”

In their argument, they start with a vector with a fractional value for each variable and prove that the vector is in $N_+^t(\mathcal{P})$ for $t = \Omega(n)$. A move of the adversary in their setting amounts to fixing one of the fractional variables to 1 or 0 (i.e., true or false) and the prover is supposed to provide a fractional assignment consistent with the fixing, which is still in the polytope of feasible solutions. They then obtain the set O by fixing *additional* variables such that the desired fractional vector is a convex combination of

the assignments in O . The additional variables are fixed to maintain the invariant that if one considers the formula restricted only to the variables that have not been fixed to 0 or 1, then the formula is still expanding (in the sense that a set of s clauses will contain at least $3s/2$ unfixed variables). Expansion essentially means that even when $O(n)$ variables are fixed, most clauses still have a large number of unfixed variables, whose value can be modified suitably to satisfy the constraints of the convex program.

Lower Bounds on Size of Proofs

A finer measure than the *rank* of a proof that φ is unsatisfiable is the *size* of the proof. For the Lovász–Schrijver proof system, where one proves unsatisfiability by progressively deriving new inequalities until one gets an obvious contradiction like $0 = 1$, the size may be thought of as the *number* of inequalities that one needs to derive before obtaining a contradiction. A lower bound of t on the rank implies that one will need to use at least $n^{\Omega(t)}$ inequalities, if using *all the inequalities describing* $N^t(\mathcal{P})$. On the other hand, a size lower bound of $n^{\Omega(t)}$ would rule out *any way* of using $n^{O(t)}$ inequalities to prove unsatisfiability. We remark that a more accurate definition would involve the sum of the sizes of the coefficients in all the inequalities used. However, the lower bounds described here also hold for the definition involving number of inequalities.

Dash [37] provided exponential size lower bounds for a version of the Lovász–Schrijver proof systems, which uses an operator weaker than the N and N_+ operators. For “cutting-plane” proof systems based on the cut procedures of Gomory [1] and Chvátal [2], exponential lower bounds on the size of proofs were proved by Pudlák [38]. Kojevnikov and Itsykson [39] later proved exponential lower bounds for “tree-like” proofs in the Lovász–Schrijver proof system, using expansion-based arguments. One says that a given proof is tree-like when each derived inequality is used at most once for another derivation, that is, the derived inequalities form a tree, where the parent is derived using the children.

Pitassi and Segerlind [40] later derived a general trade-off between the rank and size of tree-like proofs in the LS and LS_+ proof systems. They proved that if such a proof has size S and rank t , then one must have $t = \Omega(\sqrt{n \log S})$. Using a rank lower bound of $t = \Omega(n)$, this implies a lower bound of $2^{\Omega(n)}$ on S . However, they also showed that this trade-off did not hold for general (non-tree like) proofs in these proof systems.

Size lower bounds for general proofs in the LS and LS_+ can now be derived via an indirect route, by combining the results of Pitassi and Segerlind with those of Schoenebeck [27]. Although the trade-off of Pitassi and Segerlind did not work for general proofs in the LS-proof system, they showed (extending an argument of Clegg *et al.* [41]) that it *does* hold for general proofs in the proof systems corresponding to the Sherali–Adams and Lasserre hierarchies. This means that an $\Omega(n)$ lower bound on the rank of Lasserre proofs would imply an exponential lower bound on the size in the Lasserre proof system; and hence also on the size of *general* proofs in the LS and LS_+ proof systems (each LS and LS_+ proof can be simulated by a Lasserre proof of the same size). Such rank lower bounds for Lasserre proofs were provided by Schoenebeck [27], thus also proving the required size bounds.

CONCLUSION

The hierarchies of Lovász and Schrijver are an interesting computational model, and can be seen to capture many known approximation algorithms within the first few levels. Perhaps the most interesting problem about these is finding algorithmic applications of programs at the higher levels of these hierarchies. For problems such as sparsest cut and approximate graph coloring, the known lower bounds are far from tight, and it is an intriguing possibility that programs at higher levels can provide a better approximation guarantee for these problems.

In the context of lower bounds, it still remains to settle the integrality gap for the much investigated minimum vertex cover problem in the LS_+ hierarchy. Also, it is interesting to investigate the integrality gaps for

other optimization problems such as sparsest cut and unique games.

Acknowledgments

I thank Sanjeev Arora and Pavel Hrubes for helpful comments and references. I also thank the anonymous referees for suggesting various corrections and improvements to the previous version of this manuscript, and for pointing out that the LP relaxation in Fig. 2 has infinite integrality gap. This work was supported by NSF grant CCF-0832797.

REFERENCES

1. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64:275–278.
2. Chvátal V. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete Math* 1973;4:305–337.
3. Lovász L, Schrijver A. Cones of matrices and set-functions and 0–1 optimization. *SIAM J Optim* 1991;1(12):166–190.
4. Buresh-Oppenheim J, Galesi N, Hoory S, *et al.* Rank bounds and integrality gaps for cutting planes procedures. *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*; Cambridge (MA); 2003;pp. 318–327.
5. Sherali HD, Adams WP. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J Discrete Math* 1990;3(3):411–430.
6. Lasserre JB. An explicit exact SDP relaxation for nonlinear 0–1 programs. *Proceedings of the 8th Conference on Integer Programming and Combinatorial Optimization*. London: Springer; 2001. pp. 293–303.
7. Laurent M. A comparison of the Sherali–Adams, Lovász–Schrijver and Lasserre relaxations for 0–1 programming. *Math Oper Res* 2003;28:470–496.
8. Cook W, Dash S. On the matrix-cut rank of polyhedra. *Math Oper Res* 2001;26(1):19–30.
9. Goemans MX, Tunçel L. When does the positive semidefiniteness constraint help in lifting procedures? *Math Oper Res* 2001;26(4):796–815.
10. Alekhnovich M, Arora S, Tourlakis I. Towards strong nonapproximability results in the Lovász–Schrijver hierarchy. *Proceedings of*

- the 37th ACM Symposium on Theory of Computing; Baltimore (MD); 2005. pp. 294–303.
11. Arora S, Rao S, Vazirani U. Expander flows and a $\sqrt{\log n}$ -approximation to sparsest cut. Proceedings of the 36th ACM Symposium on Theory of Computing; Chicago (IL); 2004.
 12. Leighton FT, Rao S. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. Proceedings of the 29th IEEE Symposium on Foundations of Computer Science; White Plains (NY); 1988. pp. 422–431.
 13. Leighton FT, Rao S. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J ACM* 1999;46:787–832.
 14. Bhaskara A, Charikar M, Chlamtac E, *et al.* Detecting high log-densities – an $O(n^{1/4})$ approximation for densest k -subgraph. Proceedings of the 42nd ACM Symposium on Theory of Computing; Cambridge (MA); 2010.
 15. Feige U, Kortsarz G, Peleg D. The dense k -subgraph problem. *Algorithmica* 2001;29(3): 410–421.
 16. Goldstein D, Langberg M. The dense k subgraph problem. 2009. (arXiv:math/0912.5327).
 17. Arora S, Bollobás B, Lovász L. Proving integrality gaps without knowing the linear program. Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science; Vancouver, BC; 2002. pp. 313–322.
 18. Arora S, Bollobás B, Lovász L, *et al.* Proving integrality gaps without knowing the linear program. *Theory Comput* 2006;2(2):19–51.
 19. Tzourakis I. New lower bounds for vertex cover in the Lovasz-Schrijver hierarchy. Proceedings of the 21st IEEE Conference on Computational Complexity; Prague, Czech Republic; 2006.
 20. Schoenebeck G, Trevisan L, Tulsiani M. Tight integrality gaps for Lovasz-Schrijver LP relaxations of vertex cover and max cut. Proceedings of the 39th ACM Symposium on Theory of Computing; San Diego (CA); 2007.
 21. Kleinberg JM, Goemans MX. The Lovász Theta function and a semidefinite programming relaxation of vertex cover. *SIAM J Discrete Math* 1998;11:196–204.
 22. Charikar M. On semidefinite programming relaxations for graph coloring and vertex cover. Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms; San Francisco (CA); 2002. pp. 616–620.
 23. Georgiou K, Magen A, Pitassi T, *et al.* Integrality gaps of $2 - o(1)$ for vertex cover SDPs in the Lovász-Schrijver hierarchy. Proceedings of the 48th IEEE Symposium on Foundations of Computer Science; Providence, RI; 2007. pp. 702–712.
 24. Frankl P, Rodl V. Forbidden intersections. *Trans Am Math Soc* 1987;300(1):259–286.
 25. Schoenebeck G, Trevisan L, Tulsiani M. A linear round lower bound for Lovász-Schrijver SDP relaxations of vertex cover. Proceedings of the 22nd IEEE Conference on Computational Complexity; San Diego (CA); 2007.
 26. Tulsiani M. CSP gaps and reductions in the Lasserre hierarchy. Proceedings of the 41st ACM Symposium on Theory of Computing; Bethesda (MD); 2009.
 27. Schoenebeck G. Linear level Lasserre lower bounds for certain k -csp. Proceedings of the 49th IEEE Symposium on Foundations of Computer Science; Philadelphia (PA); 2008.
 28. Cheung KKH. On Lovász-Schrijver lift-and-project procedures on the Dantzig-Fulkerson-Johnson relaxation of the TSP. *SIAM J Optim* 2005;16(2):380–399.
 29. Feige U, Krauthgamer R. The probable value of the Lovász–Schrijver relaxations for maximum independent set. *SIAM J Comput* 2003;32(2):345–370.
 30. Jerrum M. Large cliques elude the metropolis process. *Random Struct Algor* 1992;3(4):347–360.
 31. Kučera L. Expected complexity of graph partitioning problems. *Discrete Appl Math* 1995;57(2-3):193–212.
 32. Alon N, Spencer J. *The Probabilistic method*. Wiley-Interscience Series; 2008;
 33. Cook W, Coullard CR, Turán G. On the complexity of cutting-plane proofs. *Discrete Appl Math* 1987;18(1):25–38.
 34. Ben-Sasson E, Wigderson A. Short proofs are narrow: resolution made simple. *J ACM* 2001;48(2):149–169.
 35. Ben-Sasson E. Expansion in proof complexity [PhD. thesis]. Hebrew University. 2001.
 36. Grigoriev D, Hirsch EA, Pasechnik DV. Complexity of semi-algebraic proofs. Symposium on Theoretical Aspects of Computer Science; 2002. pp. 419–430.
 37. Dash S. An exponential lower bound on the length of some classes of branch-and-cut proofs. *Integer Programming and Combinatorial Optimization*; 2002. pp. 145–160.

- 38. Pudlák P. Lower bounds for resolution and cutting plane proofs and monotone computations. *J Symb Log* 1997;62(3):981–998.
- 39. Kojevnikov A, Itsykson D. Lower bounds of static lovász-schrijver calculus proofs for tseitin tautologies. *International Colloquium on Automata, Languages and Programming* (1); Venice, Italy; 2006. pp. 323–334.
- 40. Pitassi T, Segerlind N. Exponential lower bounds and integrality gaps for tree-like Lovász-Schrijver procedures. *Symposium on Discrete Algorithms*; New York (NY); 2009. pp. 355–364.
- 41. Clegg M, Edmonds J, Impagliazzo R. Using the Groebner basis algorithm to find proofs of unsatisfiability. *Proceedings of the 28th ACM Symposium on Theory of Computing*; Philadelphia (PA); 1996. pp. 174–183.

LP DUALITY AND KKT CONDITIONS FOR LP

JENS CLAUSEN
DTU Management,
Technical University of
Denmark, Lyngby,
Denmark

Duality is one of the most fundamental concepts in connection with linear programming (LP) and provides the basis for better understanding of LP models and their results and for algorithm construction in linear and in integer programming. Duality in LP was first introduced by von Neumann in Ref. 1 and can be defined in two ways: The *mathematical definition approach* and the *multiplier approach*. Duality can furthermore be motivated in an economical setting. In most textbooks, as for example [2–4], duality is introduced combining the first and third of these approaches.

Here, we first define duality using the multiplier approach for a specific set of LPs. Although not used explicitly to define duality, the approach is implicitly described in Ref. 5 under the heading Lagrangean relaxation. A similar approach is used in Ref. 6, and for non-linear programming (NLP) and integer programming in Refs 7 and 8.

Then we state both the Weak and the Strong Duality Theorems formulated first in Ref. 9. The mathematical definition approach is used to define the dual to any given LP and we show the natural correspondence with the multiplier approach.

The Complementary Slackness Conditions for optimality of a pair of solutions to a primal LP and its corresponding dual are stated and proved, and finally the relation between these and the KKT conditions for optimality, known from NLP, is demonstrated, thus showing the intimate relationship between duality in LP and KKT-conditions for NLPs.

DERIVING THE DUAL PROBLEM

The standard form of an LP, which is commonly used when introducing the Simplex method for solving LP problems, is a problem with n nonnegative variables and m equality constraints:

$$\begin{array}{ll} \min & cx \\ & Ax = b \quad (P) \\ & x \geq 0. \end{array}$$

Here, c is an $1 \times n$ matrix, A is an $m \times n$ matrix, b is an $m \times 1$ matrix of reals, and x is an $n \times 1$ vector of real variables. This form is the outset for our presentation of duality.

A very common and general approach to problem solution is to transform the problem at hand to a similar problem, which is possibly easier to solve. In the LP case, this raises the question of which LPs are easily solvable. The answer is immediate: an LP consisting only of an objective function and nonnegativity constraints on the variables. If any coefficient in the objective function is negative, the solution is unbounded from below (in the following termed *equal to* $-\infty$) with the corresponding variable unbounded (termed *equal to* ∞) and all other variables equal to 0. Otherwise, it is 0 with all variables equal to 0.

Transforming our initial problem by deleting all constraints does not make much sense, that is, the transformed problem is simply too far from the original one. We therefore reintroduce each constraint in the objective function through a term, which consists of the difference between the right-hand side and the left-hand side of the constraint, $b_i - A_i x$, multiplied with a multiplier y_i . We hence introduce m multipliers $y_1, \dots, y_m$ —one for each constraint of the original problem.

Regardless of the sign of these multipliers, the optimal value of the transformed problem now constitutes a lower bound for the value of the original problem. For any feasible solution to the original problem, the new terms in the objective function vanish (since

the constraints of the original problem are equality constraints) and the original value results. Thus the minimum is to be found over a range including all values of feasible solutions and most likely many other values; hence, the minimum is less than or equal to the minimum for the original LP.

Since we are interested in information about our original problem, a natural question is: Which multipliers provide the best possible lower bound, and how is this bound related to the optimal value of the original problem? The answer can be found by solving the following maximization problem:

$$\max_{y \in \mathcal{R}^m} \left\{ \min_{x \in \mathcal{R}_+^n} \{ cx + y(b - Ax) \} \right\}$$

In the inner minimization problem, the y 's are constants and hence yb is a constant. Rearranging the terms in the objective function, we arrive at an LP problem with no constraints but the nonnegativity of the variables:

$$\max_{y \in \mathcal{R}^m} \left\{ \min_{x \in \mathcal{R}_+^n} \{ (c - yA)x + yb \} \right\}$$

If any coefficient $(c - yA)_j$ in the objective function is less than 0, we know that the lower bound becomes $-\infty$, which provides no new knowledge. In the process of deriving the best lower bound, we therefore require that

$$\forall j \in 1, \dots, m : (c - yA)_j \geq 0 \Leftrightarrow yA_j \leq c_j$$

Since we have imposed no bounds on our multipliers, the problem of finding the best multipliers providing the tightest lower bound for our original problem becomes

$$\begin{array}{ll} \max & yb \\ & yA \leq c \quad (DP) \\ & y \text{ free} \end{array}$$

As shown in the following section, the best lower bound surprisingly turns out to equal the optimal value of the primal objective function. The step in the process, where the constraints are included in the objective function, is called a *relaxation*—the constraints of the transformed problem are weaker and

the objective function is pointwise below the original one for all feasible solutions.

The original problem is termed the *primal* LP problem as indicated by (P) and the problem of finding the best lower bound is termed the *dual* LP problem as indicated by (DP) . Together, these are called a *pair of primal and dual* LP problems. Note the following crucial points: For each constraint in the primal problem, there is a variable in the dual problem, and for each variable in the primal problem, there is a constraint in the dual. When the primal problem is a minimization problem, the dual is a maximization problem.

We further illustrate the process and the interplay between relational operators of constraints respective sign constraints for variables of the primal problem and sign constraints respective relational operators of constraints in the dual problem by summarizing the process for a maximization problem in non-negative variables with less than or equal constraints:

$$\begin{array}{ll} \max & cx \\ & Ax \leq b \\ & x \geq 0 \end{array} \quad \mapsto \quad \begin{array}{ll} \min_{y \in \mathcal{R}_+^m} & \left\{ \max_{x \in \mathcal{R}_+^n} \{ cx + y(b - Ax) \} \right\} \\ \min_{y \in \mathcal{R}_+^m} & \left\{ \max_{x \in \mathcal{R}_+^n} \{ (c - yA)x + yb \} \right\} \end{array} \quad \mapsto \quad \begin{array}{ll} \min & yb \\ & yA \geq c \\ & y \geq 0. \end{array}$$

Since the primal problem is one of maximization, we search for a tightest upper bound. To provide an upper bound, the new objective function has to be pointwise larger than or equal to the original one for each feasible solution; hence, each multiplier must be nonnegative. The outer objective function is to be minimized, and each coefficient in the inner maximization problem must be nonpositive in order to provide a bounded solution to this.

Before we address the general construction of a dual problem from a given primal, we state the Weak and the Strong Duality theorem for the particular case just developed.

THE DUALITY THEOREMS

Theorem 1 [The Weak Duality Theorem].

Consider a given pair of primal and dual LP problems $\min\{cx \mid Ax = b, x \geq 0\}$ and $\max\{yb \mid yA \leq c, y \text{ free}\}$. The value yb of any dual feasible solution y is less than or equal to the value cx of any primal feasible solution x .

We have already proven the theorem during the construction process by noting that for any given set of multipliers, the value of the transformed problem is a lower bound for the value of the primal problem.

The Weak Duality Theorem shows that if the primal problem is feasible and unbounded, the dual problem must be infeasible and vice versa. However, this leaves open the possibility that both the primal and dual problems are infeasible, and such examples do exist.

Theorem 2 [The Strong Duality Theorem].

If feasible solutions to both the primal problem $\min\{cx \mid Ax = b, x \geq 0\}$ and the dual problem $\max\{yb \mid yA \leq c, y \text{ free}\}$ exist, then each of these has an optimum solution and the values of these solutions are equal.

Our proof of the Duality Theorem proceeds in the traditional way. We identify a set of dual variables—multipliers—which satisfy the dual constraints and give a dual objective function value equal to the optimal primal value.

Assuming that P and DP have feasible solutions implies that P can be neither infeasible nor unbounded. Hence an optimal basis B and a corresponding optimal basic solution x_B for P exists. The coefficient matrix A , the vector of objective coefficients c , and the vector of variables x can be rearranged and denoted $A = BN$, $c = (c_B c_N)$, and $x = (x_B x_N)$. With this notation, the optimal basic solution take the values $x_B = B^{-1}b$, $x_N = 0$. The vector $y_B = c_B B^{-1}$ is called the *dual solution* or the set of *simplex multipliers* corresponding to B . If the reduced costs of the simplex tableau are calculated using y_B as multipliers:

$$\bar{c} = c - y_B A,$$

then as required by the Simplex method, $\bar{c}_j$ equals 0 for all basic variables and $\bar{c}_j$ is non-negative for all nonbasic variables. Hence,

$$y_B A \leq c$$

holds, showing that y_B is a feasible dual solution. The value of this solution is $c_B B^{-1}b$, which is exactly the same as the primal objective value: $(c_B c_N)(x_B x_N) = c_B B^{-1}b$.

The dual solution y_B corresponding to the optimal basis B is often termed the vector of *shadow prices*. The terminology stems from problems, in which one wants to maximize the profit of selling a set of products produced from a set of raw materials by, for example, blending. This leads to a maximization problem in nonnegative variables with constraints being of the type “less than or equal,” each constraint limiting the use of some raw materials by the available amount. The optimal value of the dual variable for such a constraint indicates the break-even price for the corresponding raw material, that is, the price at which the cost of buying more raw material equals the increase in profit resulting from using the extra raw material. Hence, if not all the available units are used, the corresponding dual optimal value must be 0.

GENERAL DEFINITION OF DUALITY

Consider now the situation, where a general LP with m constraints and n variables is given:

$$\begin{array}{llll} \text{min/max} & cx & & \\ & Ax & \geq, =, \leq & b \\ & x & \geq, \leq, \text{free} & 0. \end{array}$$

The objective function may be minimize or maximize, for each constraint the primal relational operator may be $\leq$, $\geq$, or $=$, and for each variable, the primal sign constraint may be ≥ 0 , ≤ 0 , or *free* (indicating no constraint).

The dual problem is then

$$\begin{array}{llll} \text{max/min} & yb & & \\ & yA & \geq, =, \leq & c \\ & y & \geq, \leq, \text{free} & 0. \end{array}$$

Table 1. The Interplay Between Relational Operators in the Primal LP and Sign Constraints of the Variables in the Dual LP

Primal: Min		Dual: Max	
(a) i th constraint	$\geq$	i th variable	≥ 0
(b) i th constraint	$\leq$	i th variable	≤ 0
(c) i th constraint	$=$	i th variable	free
(d) j th variable	≥ 0	j th constraint	$\leq$
(e) j th variable	≤ 0	j th constraint	$\geq$
(f) j th variable	free	j th constraint	$=$

Primal: Max		Dual: Min	
(a) i th constraint	$\geq$	i th variable	≤ 0
(b) i th constraint	$\leq$	i th variable	≥ 0
(c) i th constraint	$=$	i th variable	free
(d) j th variable	≥ 0	j th constraint	$\geq$
(e) j th variable	≤ 0	j th constraint	$\leq$
(f) j th variable	free	j th constraint	$=$

The interplay between primal relational operators and the corresponding dual sign constraints and vice versa can be derived by “best bound” arguments as in the section titled “Deriving the Dual Problem” and is given in Table 1.

In the general case, the Weak Duality Theorem follows from relaxation arguments exactly, as previously shown. The Strong Duality Theorem is slightly more complicated to prove - the given problem is transformed to standard form using slack and/or surplus variables, and it is demonstrated that these two equivalent problems have the same dual problem. The theorem now follows from the construction applied for the standard form.

OPTIMALITY CONDITIONS FOR PAIRS OF LPs—COMPLEMENTARY SLACKNESS

When developing iterative solution algorithms for the optimization problems, it is crucial to identify optimality conditions—such conditions enable an algorithm to stop iterating when optimality is achieved. One obvious condition when solving an LP is that a primal feasible solution x and a dual feasible solution y with equal values are identified. This is a consequence of

the Weak Duality Theorem. The *Complementary Slackness Conditions*—termed the *CS-conditions*—constitute another set of optimality conditions for a given pair of dual LPs. We first formulate these for a specific type of pair of dual problems.

Theorem 3. *Consider a pair of dual LP-problems, P , and DP :*

$$\begin{array}{ll} \max & cx \\ & Ax \leq b \quad (P) \\ & x \geq 0 \end{array} \quad \begin{array}{ll} \min & yb \\ & yA \geq c \quad (DP) \\ & y \geq 0 \end{array}$$

and a pair $\bar{x}, \bar{y}$ of feasible solutions to P and DP , respectively. $\bar{x}$ and $\bar{y}$ are optimal solutions to P and DP , respectively, if and only if, they satisfy the complementary slackness conditions:

$$\begin{aligned} \forall i \in \{1, \dots, m\} : \bar{y}_i(b_i - A_i\bar{x}) &= 0 \\ \forall j \in \{1, \dots, n\} : (\bar{y}A_j - c_j)\bar{x}_j &= 0 \end{aligned}$$

The terms $(b_i - A_i\bar{x})$ and $(\bar{y}A_j - c_j)$ are called the i th primal slack and j th dual slack, respectively.

Before commenting on the general form of the CS-conditions, we prove the theorem. Both primal and dual variables are non-negative. Furthermore, for feasible solutions, $b_i - A_i x \geq 0$ and $yA_j - c_j \geq 0$ hold. Therefore, each term $y_i(b_i - A_i x)$ and $(yA_j - c_j)x_j$ is nonnegative, and the CS-conditions can be stated in compact notation as

$$\bar{y}(b - A\bar{x}) = 0 \wedge (\bar{y}A - c)\bar{x} = 0.$$

Suppose that a pair $\bar{x}, \bar{y}$ of feasible solutions satisfy the CS-conditions. Now:

$$c\bar{x} = (\bar{y}A)\bar{x} = \bar{y}(A\bar{x}) = \bar{y}b,$$

where the first and third equality follow from the CS-conditions. Hence $\bar{x}$ and $\bar{y}$ are feasible

primal and dual solutions with the same value, and by the Weak Duality Theorem, both must be optimal.

Suppose now $\bar{x}$ and $\bar{y}$ are feasible optimal solutions implying that $c\bar{x} = \bar{y}b$. Thus,

$$c\bar{x} \leq (\bar{y}A)\bar{x} = \bar{y}(A\bar{x}) \leq \bar{y}b = c\bar{x}.$$

Here, the first inequality is due to the feasibility of $\bar{y}$ and the nonnegativity of $\bar{x}$, the second to the feasibility of $\bar{x}$ and nonnegativity of $\bar{y}$, and the last equation to the optimality of both. Hence, equality must hold throughout the expression implying that the CS-conditions are satisfied.

The CS-conditions can be shown to be necessary and sufficient optimality conditions for a pair of primal and dual feasible solutions, no matter which form the relational operators and sign constraints take in the problems at hand. Note specifically the situation when the primal problem is in standard form: minimization of a problem in nonnegative variables and with equality constraints. In this case, the CS-conditions

$$\forall i \in \{1, \dots, m\} : \bar{y}_i(b_i - A_i\bar{x}) = 0,$$

are vacuously fulfilled for a pair of feasible solutions $\bar{x}, \bar{y}$. Noting that for a set of basic feasible solutions, the reduced costs are calculated as $\bar{c} = c - c_B B^{-1}A$, the remaining CS-conditions can be formulated as

$$\forall j \in \{1, \dots, n\} : \bar{c}_j x_j = 0$$

Since only basic variables are allowed to take positive values in a basic solution, and all nonbasic reduced costs are 0, these CS-conditions are also satisfied by any basic solution generated when the problem at hand is solved iteratively by the simplex method. The ratio test ensures that the basic solution of each iteration is feasible. The iterations stop when all reduced costs are nonnegative, which is equivalent to that $y_B = c_B B^{-1}$ constitute a dual feasible solution. Hence the iterations maintain primal feasibility of the basic solution and the CS-conditions, and

strive for feasibility of the dual solution corresponding to the basis.

KKT CONDITIONS FOR AN LP

The KKT conditions [9] are optimality conditions for NLPs. Since LPs are special cases of NLPs these conditions apply also in the linear case. However, as will become clear, the KKT-conditions for optimality are in the linear case the CS-conditions just stated. We start by stating the conditions, and then we interpret these in a linear setting.

We consider the following problem in n variables with m inequality and p equality constraints. All functions are assumed to be differentiable:

$$\begin{aligned} \max \quad & f(x) \\ g(x) \quad & \leq b \\ h(x) \quad & = c \\ x \quad & \in R^n \end{aligned}$$

A given $x^o \in R^n$ satisfies the KKT-conditions if and only if:

1. x^o is feasible, and
2. there exist $\lambda_j, j = 1, \dots, m$ and $\mu_k, k = 1, \dots, p$ such that

$$\begin{aligned} \nabla f(x^o) &= \sum_{j=1}^m \lambda_j \nabla g_j(x^o) + \sum_{k=1}^p \mu_k \nabla h_k(x^o) \\ \lambda_j(b_j - g_j(x^o)) &= 0, \quad j = 1, \dots, m \\ \lambda_j &\geq 0, \quad j = 1, \dots, m. \end{aligned}$$

The operator ∇ indicates partial derivation $\nabla g_j(x) = \partial(g_j(x))/\partial x_1, \dots, \partial(g_j(x))/\partial x_i, \dots, \partial(g_j(x))/\partial x_n$. Condition (2) can be rephrased as follows: There exist vectors $\lambda \in R^m$ and $\mu \in R^p$ such that

$$\begin{aligned} \nabla f(x^o) &= \lambda \nabla g(x^o) + \mu \nabla h(x^o) \\ \lambda(b - g(x^o)) &= 0 \\ \lambda &\geq 0. \end{aligned}$$

If the functions $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are all linear, we may consider the problem at hand as the primal problem in a pair of dual problems and state it together with its dual

$$\begin{array}{llll}
\max & cx & \min & \lambda b_1 + \mu b_2 \\
& A_1 x & \leq & b_1 \\
(P) & A_2 x & = & b_2 \\
& x & \text{free} & \\
\end{array}
\quad
\begin{array}{ll}
(\lambda, \mu) \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} & = c \\
\lambda & \geq 0 \\
\mu & \text{free}
\end{array}
\quad (DP)$$

Writing out the KKT-conditions for a given $x^o \in R^n$ using that the j th partial derivative of a linear function $a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n$ is just a_{ij} , these state that

1. x^o is feasible, and
2. there exist $\lambda_j, j = 1, \dots, m$ and $\mu_k, k = 1, \dots, p$ such that

$$\begin{aligned}
c &= \sum_{j=1}^m \lambda_j A_{1j} + \sum_{k=1}^p \mu_k A_{2k}. \\
\lambda_j (b_j - A_{1j} x^o) &= 0, \quad j = 1, \dots, m \\
\lambda_j &\geq 0, \quad j = 1, \dots, m.
\end{aligned}$$

If we rephrase (2), we get that there exist vectors $\lambda \in R^m$ and $\mu \in R^p$ such that

$$\begin{aligned}
c &= \lambda A_1 + \mu A_2 \\
\lambda (b - A_1 x^o) &= 0 \\
\lambda &\geq 0.
\end{aligned}$$

However, considering x^o as a primal solution and λ, μ as a dual solution, condition (1) states primal feasibility, the first part and third part of condition (2) dual feasibility, and the second part of condition (2) that x^o, λ , and μ satisfy the CS-conditions. Note that both the CS-conditions regarding the primal constraints $A_2 x = b_2$ and the dual constraints $c = \lambda A_1 + \mu A_2$ are vacuously fulfilled—all feasible solutions leave no slack in any of these constraints.

Hence, the statement for NPLs that “under certain assumptions on the functions $f(\cdot), g(\cdot)$, and $h(\cdot)$, the KKT-conditions are both necessary and sufficient for optimality of a point $x^o \in R^n$ ” in the linear case reduces to “a primal feasible solution is optimal if and only if a dual feasible solution exists

such that the pair of solutions satisfy the CS-conditions.”

OPTIMALITY CONDITIONS REVISITED - GEOMETRY

Through an example, we now illustrate the CS- and KKT-conditions for LP optimality from a geometric point of view. Consider the following LP-problem in two nonnegative variables:

$$\begin{aligned}
\max & 7x_1 + 11x_2 \\
& -2x_1 + x_2 \leq 1 \\
& x_1 + x_2 \leq 7 \\
& x_2 \leq 2.5 \\
& x_1, x_2 \geq 0
\end{aligned}$$

The feasible region is shown in Fig. 1, and the optimal solution is $A = (x_1, x_2) = (9/2, 5/2)$.

An alternative formulation of the problem is as a maximization problem with only $\geq$ -constraints and free variables:

$$\begin{aligned}
\max & 7x_1 + 11x_2 \\
& 2x_1 - x_2 \geq -1 \\
& -x_1 - x_2 \geq -7 \\
& -x_2 \geq -2.5 \\
& x_1 \geq 0 \\
& x_2 \geq 0 \\
& x_1, x_2 \text{ free}
\end{aligned}$$

The general form with m constraints and n variables is

$$\begin{array}{ll}
\max & cx \\
& Ax \geq b \\
& x \text{ free}
\end{array}$$

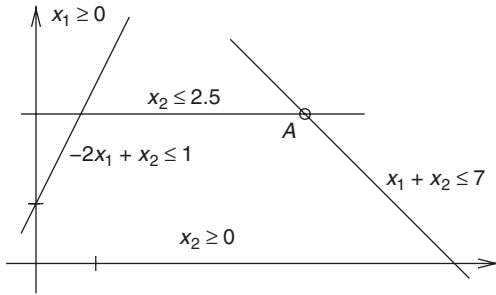


Figure 1. The feasible region and optimal solution for an LP in 2 dimensions.

with the dual problem

$$\begin{aligned} \min \quad & yb \\ \text{subject to} \quad & yA = c \\ & y \leq 0 \end{aligned}$$

Hence the dual problem to the alternative formulation is

$$\begin{aligned} \min \quad & -y_1 - 7y_2 - 2.5y_3 + 0y_4 + 0y_5 \\ \text{subject to} \quad & 2y_1 - y_2 + y_4 = 7 \\ & -y_1 - y_2 - y_3 + y_5 = 11 \\ & y_1, y_2, \dots, y_5 \text{ free} \end{aligned}$$

A given x^o is now according to the KKT-conditions an optimal solution *if and only if*

1. x^o is a feasible solution to P, and
2. there exists a vector y^o satisfying the following:

$$\begin{aligned} y^o A &= c & (1) \\ y^o (b - Ax^o) &= 0 & (2) \\ y^o &\leq 0 & (3) \end{aligned}$$

Considering (1) and (3), these conditions indicate that y^o is feasible for the dual problem, and (2) is just CS-conditions for x^o, y^o .

Geometrically, the vector c is the *gradient* of the objective function cx , that is, that direction for a move in R^n , for which the objective function *increases* most rapidly. The feasible region $S = \{x | Ax \geq b\}$ for the problem is defined by the functions $a_{11}x_1 + \dots + a_{1n}x_n - b_1, \dots, a_{m1}x_1 + \dots + a_{mn}x_n - b_m$, which must

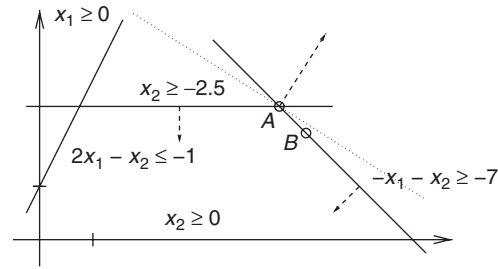


Figure 2. The KKT-optimality conditions for an LP in two dimensions.

all have values greater than or equal to 0 in a given x for this to be feasible. The gradients $(a_{11}, \dots, a_{1n}), \dots, (a_{m1}, \dots, a_{mn})$ for these functions viewed as vectors are directed *into* the feasible region S ; see Fig. 2. Condition (1) and (3) now state that at an optimal point of S —an extreme point Q —the gradient of the objective function must be a *nonpositive linear combination* of the gradients of the constraints, while condition (2) states that only *binding* constraint may have nonzero coefficients in this linear combination. The situation for the example with $Q = A = (9/2, 5/2)$ is illustrated in Fig. 2. Intuitively, if just one of the gradients of the linear combination has a positive coefficient, the current solution is not optimal. This is demonstrated by the existence of a positive reduced cost.

One interesting final point to mention is that the KKT-conditions may be satisfied also by optimal points, which are not extreme points of the feasible region. Consider, for example, the previous example but now with the objective function $x_1 + x_2$. In this case, all points on the line connecting $(9/2, 5/2)$ and $(7, 0)$ are optimal, for example, $B = (5, 2)$. The only binding constraint for this point is $-x_1 - x_2 + 7 \geq 0$. Choosing the dual variables as $(y_1, y_2, y_3, y_4, y_5) = (0, -1, 0, 0, 0)$, the optimality of $(5, 2)$ is verified by the KKT-conditions as well as the CS-conditions.

SUMMARY

We have presented the theory of duality in LP in a rather untraditional setting with the goal of providing a thorough understanding of the origin of the dual problem for any

given primal problem. With this approach, the use of relaxation and duality in integer programming algorithms such as Branch and Bound and the general KKT-conditions for optimality in NLP become natural extensions of the LP duality theory.

REFERENCES

1. von Neumann J. On a maximization problem. Manuscript. Princeton (NJ): Institute for Advanced Study; 1947.
2. Taha HA. Operations research: and introduction. New Jersey: Pearson Prentice Hall; 2007.
3. Hillier FS, Lieberman GJ. Introduction to operations research. New York: Mc Graw Hill; 2010.
4. Schrijver A. Theory of linear and integer programming. New York: John Wiley & Sons; 1986.
5. Dantzig GB. Linear programming and extensions. New Jersey: Princeton University Press; 1963.
6. Vanderbei RJ. Linear programming: foundations and extensions. New Jersey: Kluwer; 1996.
7. Geoffrion AM. Duality in non-linear programming: a simplified applications-oriented approach. SIAM Rev 1971;13(1):1–37.
8. Geoffrion AM. Lagrangean relaxation for integer programming. Math Prog Study 1974;2: 82–114.
9. Gale D, Kuhn HW, Tucker AW. Linear programming and the theory of games. In: Koopmans TC, editor. Activity analysis of production and allocation. New York: John Wiley & Sons; 1951.

MACBETH (MEASURING ATTRACTIVENESS BY A CATEGORICAL BASED EVALUATION TECHNIQUE)

CARLOS A BANA E COSTA
Centre for Management Studies of
IST, Technical University of
Lisbon, Lisbon, Portugal

JEAN-MARIE DE CORTE
JEAN-CLAUDE VANSNICK
Centre de Recherche Waroqué,
University of Mons, Mons, Belgium

BACKGROUND OF MACBETH

Measuring the attractiveness of options through a category-based evaluation technique is the goal of MACBETH. Its key distinction from other value-measurement procedures, such as numerical direct rating, is that MACBETH uses qualitative judgments of differences in attractiveness in order to generate value scores for the options.

To facilitate comparison of options, two at a time, MACBETH introduces seven qualitative categories of difference in attractiveness: Is there no difference (indifference), or is the difference very weak, weak, moderate, strong, very strong, or extreme? Judgmental disagreement or hesitation between two or more consecutive categories, except indifference, is also allowed. MACBETH relies on such a pairwise comparison questioning mode. Each time that a qualitative judgment is elicited, the consistency of all the judgments thereto made by the respondent is verified and suggestions are offered to resolve inconsistencies if they arise, as detailed in the section titled “Dealing with Inconsistency while Eliciting Qualitative Judgments.”

MACBETH then derives scores for the options from the consistent set of judgments. These scores constitute the basic MACBETH numerical scale (v), which the respondent should subsequently validate (see the section titled “The Basic MACBETH Scale”). To validate the scale, the respondent should compare the sizes of the intervals between the

proposed scores and adjust them if necessary, one at a time within a range compatible with the judgments elicited. If further adjustment is needed, the respondent may even revise judgments previously provided. The aim of this process is to quantify the relative attractiveness of the options on an interval measurement scale (see the section titled “Constructing an Interval Scale”).

In a multiple criteria evaluation context, scoring the options on an interval scale within each criterion is important because it permits one to meaningfully take a weighted average of each option's scores on the criteria. This overall score measures the relative attractiveness of the option across all the criteria. The weights of the criteria can also be derived by applying the MACBETH procedure to a set of reference, hypothetical, options.

MACBETH uses mathematical programming to test consistency, find suggestions to resolve inconsistencies if they arise, derive the basic MACBETH scale for a set of consistent judgments, and find the range within which each proposed score can be adjusted. All technical details are given in the article On the Mathematical Foundations of MACBETH [1], which integrates and updates all developments since the initial formulation in Ref. 2. The MACBETH procedure is implemented in the M-MACBETH software application (available at www.m-macbeth.com), which also includes functionalities to help structure criteria and to interactively analyze the sensitivity and robustness of the model's results in the face of several sources of uncertainty (as detailed in Refs 3 and 4).

DEALING WITH INCONSISTENCY WHILE ELICITING QUALITATIVE JUDGMENTS

Suppose that one wants to apply MACBETH to help in evaluating four options A, B, C, D, on a given criterion. In the process of pairwise comparing the options, assume the respondent started by judging A to be weakly more attractive than B, then B to be more attractive than C but hesitated between a very weak or a weak difference in attractiveness, and considered B to be moderately more

	A	B	C	D
A	No	Weak	?	?
B		No	Vweak-weak	Moderate
C	?		No	Moderate
D	?			No

(a)

Figure 1. Consistent (a) and inconsistent (b) sets of judgments. It is of note that “?” in a entry means that there is no information about the difference in attractiveness between the corresponding two options; “vweak-weak” is an abbreviation of “very weak or weak” difference in attractiveness.

	A	B	C	D
A	No	Weak	Strong	?
B		No	Vweak-weak	Moderate
C			No	Moderate
D	?			No

(b)

attractive than D, and considered C also to be moderately more attractive than D. These judgments were used to populate the upper triangular part of the matrix in Fig. 1a. As each judgment was entered into the matrix, its consistency with the judgments already inserted into the matrix was checked.

In MACBETH, consistency means that, from a set of elicited judgments, it is possible to derive scores for the options such that equally attractive options get the same score, an option more attractive than another one gets a greater score, and if the difference in attractiveness between two options (strong, for example) is greater than the difference in attractiveness between two other options

(moderate, for example), then the difference between the scores for the first two options should be greater than the difference between the scores for the other two.

The set of MACBETH judgments in the matrix of Fig. 1a is consistent. However, when the respondent next judged A to be strongly more attractive than C, the set of judgments elicited (Fig. 1b) became inconsistent. This is because the judgments elicited impose that $v(C) - v(D) > v(A) - v(B)$ and $v(A) - v(C) > v(B) - v(D)$, which lead, by summation, to an impossibility: $v(A) - v(D) > v(A) - v(D)$. This conflict between qualitative judgments is easy to see in the graph of Fig. 2: indeed, the length of path A–C–D should be equal to the length of path A–B–D, but this is impossible with the judgments elicited.

Faced with an inconsistent matrix of judgments, the M-MACBETH software recognizes the problem and is able to identify the minimum number of judgments that should be modified to achieve consistency; moreover, it gives users suggestions for such modifications. In the example, consistency is achieved if one of the four suggestions shown in Fig. 3 is accepted: move (A, B) or (B, D)

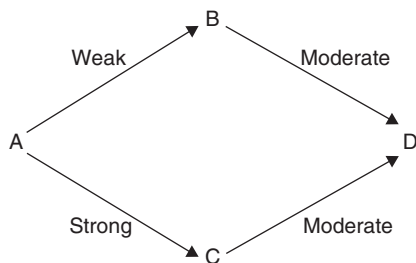


Figure 2. Graph of inconsistent judgments.

	A	B	C	D
A	No	↑ Weak	↓ Strong	?
B		No	Vweak-weak	↑ Moderate
C			No	↓ Moderate
D	?			No

Figure 3. MACBETH suggestions to resolve inconsistency.

up one category, or (A, C) or (C, D) down one category.

Suppose that when alerted to the inconsistency and presented with the suggestions to resolve it, the respondent decided to revise the judgment between A and C from strong to moderate. Finally, the last judgment was elicited, extreme difference of attractiveness between A and D, completing the consistent matrix of judgments of Fig. 4.

The number of qualitative judgments elicited can range from a maximum of $n(n-1)/2$ (6 judgments for $n=4$ options, as in Fig. 4), when all pairwise comparisons are made, to a minimum acceptable number of $n-1$ judgments, as when comparing only one option with all of the other $n-1$ or each of two consecutive options (as done in Ref. 5). However, it is recommended that some additional judgments be elicited. Belton and Stewart [6, p. 173] present an example in which $2n-3$ judgments are made, filling in only the two first diagonals as in Fig. 1b, whereas Bana e Costa and Chagas [7] fill in the border of the upper triangular portion of the matrix, for a total of $3(n-2)$ pairwise comparisons.

THE BASIC MACBETH SCALE

Let X be a finite set of $n > 2$ (actual or hypothetical) options, among which x^+ and x^- are such that x^+ is at least as attractive as any other option of X and x^- is at most as attractive as any other option of X . Assume that a set of pairwise comparison judgments has been expressed by a respondent in terms of the seven MACBETH categories ($C_k, k = 0, \dots, 6$) of difference in attractiveness: “null” (C_0), “very weak” (C_1), “weak” (C_2), “moderate” (C_3), “strong” (C_4), “very strong” (C_5), and “extreme” (C_6). A judgment between two options x and y of X such that x is at least as attractive as y will be denoted by $(x, y) \in C_k (k = 0, \dots, 6)$ when the difference in attractiveness between x and y is expressed by the single category C_k and $(x, y) \in C_i \cup \dots \cup C_s (i, s = 1, \dots, 6 \text{ with } i < s)$, when judgmental disagreement or hesitation among the categories C_i to C_s is expressed.

The basic MACBETH scale corresponding to the set of judgments expressed by the respondent can be obtained by solving the following linear programming problem, where

	A	B	C	D
A	No	Weak	Moderate	Extreme
B		No	Vweak-weak	Moderate
C			No	Moderate
D				No

Figure 4. MACBETH matrix of judgments of difference in attractiveness.

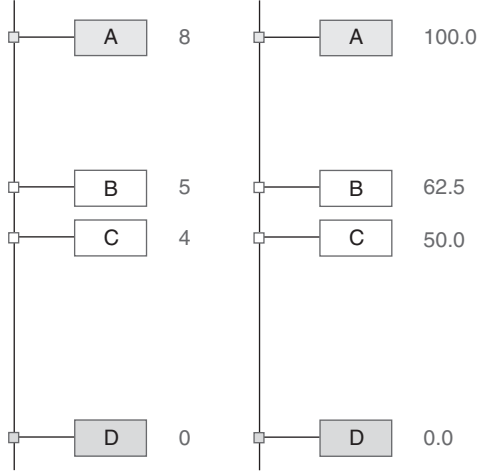


Figure 5. Basic and anchored MACBETH scales.

$v(x)$ is the score assigned to option x :

$$\text{Min}[v(x^+) - v(x^-)]$$

s. t.

- (1) $v(x^-) = 0$ (arbitrary assignment),
- (2) $\forall(x, y) \in C_0 : v(x) - v(y) = 0$,
- (3) $\forall(x, y) \in C_i \cup \dots \cup C_s$
with $i, s \in \{1, 2, 3, 4, 5, 6\}$
and $i \leq s : v(x) - v(y) \geq i$,
- (4) $\forall(x, y) \in C_i \cup \dots \cup C_s$ and
 $\forall(w, z) \in C_{i'} \cup \dots \cup C_{s'}$
with $i, s, i', s' \in \{1, 2, 3, 4, 5, 6\}, i \leq s, i' \leq s'$,
and $i > s' :$
 $v(x) - v(y) \geq v(w) - v(z) + i - s'$.

When this linear program is infeasible, the set of judgments is inconsistent. When it is feasible, the optimal solution may not be unique. If multiple solutions exist, there is more than one possible score for at least one option $x \in X \setminus \{x^-, x^+\}$, in which case their average is taken to ensure the uniqueness of the basic MACBETH scale (see details in Ref. 1).

The basic MACBETH scale for the set of judgments in Fig. 4 is $v(A) = 8, v(B) =$

$5, v(C) = 4$, and $v(D) = 0$. Note that this scale is equivalent to the scale anchored on $v(x^+) = 100$ and $v(x^-) = 0$, which is displayed graphically by default in the M-MACBETH software (see Fig. 5).

It is worthwhile to point out that the above linear program corresponds to the LP-MACBETH presented in Ref. 1. If none of the judgments are expressed by more than one category (i.e., there is no hesitation between two categories for any of the judgments elicited), the problem takes the form presented in Ref. 8 by writing constraints (3) and (4) above simply as $\forall(x, y) \in C_k$ and $\forall(w, z) \in C_{k'}$ with $k, k' \in \{0, 1, 2, 3, 4, 5, 6\}$ and $k > k'$:

$$v(x) - v(y) \geq v(w) - v(z) + k - k'.$$

CONSTRUCTING AN INTERVAL SCALE

Deriving a MACBETH scale from a consistent set of judgments is not an end in itself but a path in the learning process of measuring attractiveness in an interval scale: relative differences of scores should reflect the relative differences in attractiveness that the respondent feels between the respective options; thus, the ratio between any two positive differences of scores should represent the proportion between the respective differences in attractiveness. This is facilitated by inviting the respondent to observe the scale axis and compare score intervals. For example, one can observe in Fig. 5 that the intervals between A and C and C and D are equal and validate if the respondent feels that the difference in attractiveness between A and C is the same as that between C and D. Similarly, one can validate if the difference of attractiveness between A and B is three times greater than that between B and C. If it is found that distances between options in the scale axis do not adequately represent the respective differences in attractiveness, then at least one score should be adjusted.

As exemplified in Fig. 6, when the user selects an option in the axis, the software shows the range within which the respondent can adjust the score of the selected

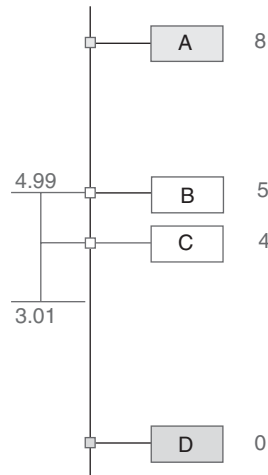


Figure 6. Range of variation of a score.

option, keeping the scores of the remaining options unchanged, without violating the consistency of the matrix of judgments. For example, as shown in Fig. 7, the lower bound 3.01 for the score of C is due to the weak and moderate judgments of difference in attractiveness between A and B and between C and D, respectively, implying that $v(C) - v(D) > v(A) - v(B)$, and so, with the fixed scores $v(A) = 8$, $v(B) = 5$, and $v(D) = 0$, one must have $v(C) > 3$. Obviously, if the respondent feels that the intervals between A and B and between C and D should be equal,

	A	B	C	D	Current scale
A	No 0.00	Weak 3.00	Moderate 4.99	Extreme 8.00	8.00
B		No 0.00	Vweak-weak 1.99	Moderate 5.00	5.00
C			No 0.00	Moderate 3.01	3.01
D				No 0.00	0.00

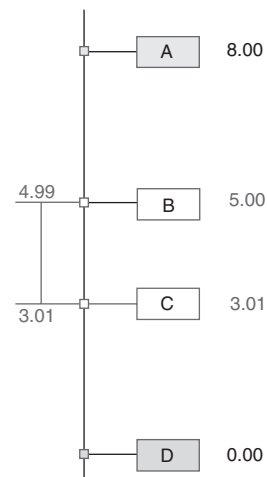


Figure 7. Judgments defining a bound for a score.

the set of judgments can be revised and the process restarted until the respondent feels that the final scores are such that relative differences between them reflect the respective relative differences in attractiveness.

CONCLUSION

One of the most crucial and difficult stages in making multicriteria evaluations is expressing clearly and consistently one's judgments about the relative value or attractiveness of options. When called in to organizations to facilitate this process, decision analysts can use numerical or nonnumerical value-elicitation techniques, such as the numerically based direct rating technique [9] or *measuring attractiveness by a categorical-based evaluation technique (MACBETH)*.

As described, the MACBETH value-elicitation procedure is composed of an input stage aimed at eliciting a consistent set of qualitative (non-numerical) pairwise comparison judgments of difference in attractiveness, and an output stage aimed at constructing an interval value scale from the set of judgments, which numerically measures the relative attractiveness of options for the person or group that made the judgments.

A wide variety of real-world cases using MACBETH and its software is referred to in

Ref. 1; some of them are described in the operational research and management science literature, such as the multicriteria socio-technical processes, combining social aspects of decision conferencing [10] with the technical components of MACBETH, to help groups evaluate bids in international call for tenders [11,12], prioritize public investments, [8,13,14] or allocate resources [15].

REFERENCES

1. Bana e Costa CA, De Corte JM, Vansnick JC. On the mathematical foundations of MACBETH. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys. New York: Springer; 2005. pp. 409–442.
2. Bana e Costa CA, Vansnick JC. MACBETH—an interactive path towards the construction of cardinal value functions. *Int Trans Oper Res* 1994;1(4):489–500.
3. Bana e Costa CA, De Corte JM, Vansnick JC. MACBETH, LSE OR Working Paper 03.56. London: London School of Economics; 2003.
4. Bana e Costa CA, De Corte JM, Vansnick JC. M-MACBETH Version 1.1 User's Guide. 2005. Available at <http://www.m-macbeth.com/downloads.html#guide>
5. Clivillé V, Berrah L, Maurris G. Quantitative expression of aggregation of performance measurements based on MACBETH multicriteria method. *Int J Prod Econ* 2007;105(1):171–189. DOI: 10.1016/j.ijpe.2006.03.002.
6. Belton V, Stewart TJ. Multiple criteria decision analysis: an integrated approach. Boston (MA): Kluwer Academic Publishers; 2002.
7. Bana e Costa CA, Chagas MP. A career choice problem: an example of how to use MACBETH to build a quantitative value model based on qualitative value judgments. *Eur J Oper Res* 2004;153(2):323–331. DOI: 10.1016/S0377-2217(03)00155-3.
8. Bana e Costa CA, Vansnick JC. The MACBETH approach: basic ideas, software, and an application. In: Meskens N, Roubens M, editors. *Advances in decision analysis*. Dordrecht: Kluwer Academic Publishers; 1999. pp. 131–157.
9. von Winterfeldt D, Edwards W. *Decision analysis and behavioural research*. New York: Cambridge University Press; 1986.
10. Phillips LD, Bana e Costa CA. Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing. *Ann Oper Res* 2007;154(1):51–68. DOI: 10.1007/s10479-007-0183-3.
11. Bana e Costa CA, Corrêa EC, De Corte JM, *et al.* Facilitating bid evaluation in public call for tenders: a socio-technical approach. *OMEGA* 2002;30(3):227–242. DOI: 10.1016/S0305-0483(02)00029-4.
12. Bana e Costa CA, Lourenço JC, Chagas MP, *et al.* Development of reusable bid evaluation models for the Portuguese Electric Transmission Company. *Decis Anal* 2008;5(1):22–42. DOI: 10.1287/deca.1080.0104.
13. Bana e Costa CA, Fernandes TG, Correia PVD. Prioritisation of public investments in social infra-structures using multicriteria value analysis and decision conferencing: A case-study. *Int Trans Oper Res* 2006;13(4):279–297. DOI: 10.1111/j.1475–3995.2006.00549.x.
14. Bana e Costa CA, Oliveira CS, Vieira V. Prioritization of bridges and tunnels in earthquake risk mitigation using multicriteria decision analysis: application to Lisbon. *OMEGA* 2008;36(3):442–450. DOI: 10.1016/j.omega.2006.05.008.
15. Bana e Costa CA, Oliveira RC. Assigning priorities for maintenance, repair and refurbishment in managing a municipal housing stock. *Eur J Oper Res* 2002;138(2):380–391. DOI: 10.1016/S0377-2217(01)00253-3.

MAINTENANCE MANAGEMENT

KIT-NAM FRANCIS LEUNG
Department of Management Sciences,
City University of Hong Kong,
Kowloon Tong, Hong Kong

PRELUDE

The mathematical theory of reliability has grown out of the demands of modern technology, and particularly the demand for effective maintenance and use of management science (MS), also called *operational/operations research* (OR), as is evident from complex military systems experience in World War II [1]. One of the first areas of reliability to be approached with any mathematical sophistication was the area of machine maintenance [2]. The techniques used to solve these problems grew out of the successful experiences of Erlang and Palm, among others, in solving telephone trunk problems. The earliest attempts to justify the Poisson distribution as the input distribution of calls to a telephone trunk also laid the basis for using the exponential distribution as the failure law of complex equipment [3].

Since then, the field of preventive maintenance for plant and buildings and the analytically related field of medical screenings have received considerable attention and interest because of the increasing complexity of machinery, building designs and screening programs, and the necessity for such systems to perform their function optimally in terms of profit (or its negation loss), cost, downtime (or its complement availability), and reliability (which can be used to measure the degree of risk or safety). Stochastic modeling and statistical analysis have underpinned much of the MS methodology used for attempting to model and predict the consequences to and behavior of a system (otherwise called a *piece of equipment*—these terms are used interchangeably in this article) when maintenance strategies are applied. Informative accounts

and brief reviews of mathematical techniques applied to maintenance are given in the article titled ***Reviews of Maintenance Literature and Models*** which can be regarded as an accompaniment of the current article.

The purpose of this article is to introduce to engineers, healthcare workers, and others, who have the responsibility for decision making in the general area of equipment maintenance, the potential benefits that can accrue to an organization by an examination of maintenance problems from an MS viewpoint. MS is concerned, to a large degree, with the construction of mathematical models of problem situations whereby alternative actions can be assessed, and the best identified.

This article proceeds by examining “single-unit” systems such as ball bearings, light bulbs, and radio tubes or “complex” systems (regarded as a single entity) such as engines, gearboxes, and TV sets, where it makes sense to assume that only one type of failure can be experienced. Models for profit, cost, downtime, and reliability consequences, given maintenance options, are then formulated along with the necessary statistical analysis and estimation techniques when the lifetime distribution is completely or partially known or completely unknown. The term “lifetime” is defined as follows: Consider a human being, a piece of equipment, an operating system, or a component of such a system. Suppose that the system under consideration is subject to random failure such that the time to failure is random or that the system deteriorates and is no longer capable of functioning properly after a random amount of time. In either case, we refer to such a length of time as the lifetime of the system.

INDUSTRIAL MAINTENANCE PROBLEMS

Industry has clearly recognized the importance of modeling and decision making in maintenance. Maintenance is now not viewed by many as a marginal activity because the

effects of poor maintenance of a plant will become manifest in the medium to long term; however, those who seek to profit in the short term do not concern themselves with effective maintenance—production is all. Consequently, the direct benefit of effective maintenance is difficult to measure; “the true value of maintenance is given in terms of events that do not occur (like a machine not breaking down), which makes the assessment of the value of these measures at best speculative.” Thus, the gains attributable to maintenance are seldom measured either directly through properly designed experiments or by monitoring the consequences for comparison with previous operating data. However, the consequences of changes in maintenance practice are measurable, or at least estimable, through the development of validated modeling. Modelers should work closely with maintenance engineers, particularly on problem recognition and data collection.

Reliable maintenance-related data, which are necessary input for model application, are also difficult to obtain, and records are often incomplete. The modeling of condition-based maintenance is a case in point here. In order for monitored variables relating to condition to be useful in anticipating failures, it is necessary to observe such variables up to and on failure. However, failures are rare, as regular monitoring invariably leads to preventive maintenance being carried out before failure. Thus, while the qualitative effect of such condition monitoring can be seen (fewer failures), the efficiency of such monitoring is difficult to assess (many systems may be overmaintained with components replaced prematurely by conservative maintenance engineers). Again, if the consequences of condition monitoring are to be measured, then a valid model of the monitoring process must be developed. When objective maintenance records are incomplete, the subjective estimation techniques may be considered for modeling maintenance in general [4,5].

IMPORTANCE OF MAINTENANCE

The importance of the maintenance function and maintenance management has grown

over the years. In industry, the reasons for this are as follows:

1. It has been pointed out by Christer [6] that management problems of competitiveness and profitability are more critical because of the high cost of plant, the loss of skilled cheap labor, and competitive pressures on profit margins. Competing in a world market for manufactured goods requires highly automated equipment perceived as leading to the most sustained profit. Expensive high volume production units are required to be available for profitable production, that is, downtime should be minimal. As the existing number of such production units increases, the importance of the task of plant management is evident and will become increasingly so if profit margins are squeezed by competition. Downtime must be avoided, but not at any cost.
2. It has also been pointed out by Christer [6] that customer and marketing demands for products of high quality and high reliability often means reducing human involvement in manufacturing, that is, using automated equipment and robotics. Quality and reliability constraints upon the product then impose in turn corresponding quality and reliability constraints upon the plant producing the product. Management systems such as Just-in-Time and other production control and inventory reduction techniques imply tighter delivery constraints, which again require plant to be reliable, that is, available when needed.
3. As technological advances produce more and more complex systems that are extremely expensive to build and the consequences of their unreliable behavior have become severe in terms of cost, effort, lives, and so on, the interest in maintaining system function has become very important.
4. Widespread mechanization and automation has reduced the number

of production personnel and has increased the capital employed in production equipment and civil structures. As a result, the fraction of employees working in the maintenance area has grown, and so has the fraction of maintenance spending on the total operational costs.

MAINTENANCE AREA

The main industries considered include the oil and chemical industry, railways, transport companies, airlines, steelworks, production lines (e.g., canning lines), and discrete part manufacturers. Also considered are the public sector, for example, electricity generation, defense and infrastructure (e.g., roads), and the health sector, for example, chronic diseases and tumors.

It will assist us conceptually if for the present we restrict attention to the operation and maintenance of plant such as industrial manufacturing equipment, vehicle fleets, power generation facilities, and transport systems.

MAINTENANCE PERSONNEL

While organizations may know their annual expenditure on maintenance, they will not know whether this sum is too high, too low, or about right when viewed against consequence variables in the absence of modeling. For example, we can address the modeling problem in a particular context using monotone processes such as arithmetic, geometric, and arithmetico-geometric processes [7,8].

Maintenance personnel needs to address an inherently stochastic deterioration and failure process. Also, it is very difficult to justify the maintenance budget with its generally unknown contribution to company profits. Therefore, maintenance is often seen as a cost function only, with all the associated negative implications.

Thinking on maintenance should start in the design phase of systems. The type of equipment, the level of redundancy, and the accessibility will strongly affect the maintainability. When purchasing

systems, managers should take future maintenance costs into account. To cover the costs over all phases of systems, the life-cycle-costing concept was introduced. The maintenance concept [9] describes what events (e.g., fault and failure) trigger what type of maintenance (e.g., inspection, repair, replacement); this can be determined both after the design phase and in the operations phase. Once a system is in operation, maintenance has to be planned. By planning, we mean the determination of the execution moments of maintenance activities in accordance with other (e.g., production) plans (e.g., planning shutdowns of major units), the work preparation, and determination of the required maintenance capacity.

MAINTENANCE CONCEPT

Essentially, the maintenance concept [9] is the set of rules prescribing what maintenance is required and how demand for it is activated for a system. The introduction of large, complex systems in production, the growing awareness of the natural environment, and increasing labor costs imply the introduction of more sophisticated maintenance concepts such as reliability-centered maintenance reaching to qualitative maintenance concepts.

Maintenance is the totality of activities intended to retain a system in, or restore it to, a state in which it can perform its required function. In this definition, "retaining in" and "restoring to" correspond to preventive maintenance or condition monitoring and corrective maintenance (which is carried out upon failure and thus called *breakdown maintenance* as well). Failure denotes the transition of a system to the state in which it is incapable of fulfilling its function. The number of failures during actual operation should be decreased to as few as possible by means of preventive maintenance. In particular, as the system ages, it needs more frequent maintenance activities to the extent that costs and circumstance permit.

We are concerned with a system liable to (costly) failure. Failure is taken here to

mean a breakdown or catastrophic event, after which the system is unusable until repaired or replaced. It may also be simply deterioration to a state such that the repair can no longer be postponed. Operational definitions of failure and defectiveness used in the plant concerned are usually adopted for modeling purposes. The judgment that a system is defective is made by maintenance technicians or engineers.

Preventive maintenance (PM) is some activity carried out at (equal or unequal) intervals with the intention of reducing or eliminating the number of failures occurring or of reducing the consequences of failure in terms of, say, downtime or operating cost. This sort of activity aims at retaining part of a technical system in the operable state; this compares with alternative corrective maintenance which is carried out upon failure and aims to restore the failed part of the system to the operable state.

PM has been described as the most difficult activity to model in the field of maintenance [10]. The complexity of modeling PM stems from the difficulty of quantifying the effects of performing PM at different intervals. The traditional method for modeling PM is to model the relationship between the PM interval and the operating cost per unit time or the system availability.

Also, a great deal of the literature has been devoted to condition-monitoring models. In these models, it is assumed that there exists an indicator whose value is related to the condition of the unit. One of the successful models in this field is delay-time analysis [11]. Condition monitoring focuses on techniques that predict failures using information on the actual state (e.g., lubricant debris analysis, vibration monitoring) of equipment.

Reliability-centered maintenance (RCM) directs maintenance efforts at those parts and units where reliability is critical [12,13]. It can be regarded as the more qualitative approach to maintenance where optimization models are the quantitative approach. RCM is probably the global analysis method which is most frequently used for identifying a cost-effective maintenance program for a plant [14]. The goal of RCM is to identify the optimal maintenance for each component, taking

safety requirements, maintenance costs, and failure costs into account. No doubt the RCM methodology gives a practical and structured approach for each component. However, there is no sound foundation for claiming that the maintenance strategy derived from the RCM approach is in any sense optimal [15].

Total productive maintenance (TPM), another qualitative approach, centers on solving maintenance problems using quality circles. Nakajima [16–18] has defined TPM as an innovative approach to maintenance that optimizes equipment effectiveness, eliminates breakdowns, and promotes autonomous maintenance by operators through day-to-day activities involving the total workforce. Thus, TPM is not a specific maintenance policy; it is rather a culture, a philosophy, and a new attitude toward maintenance. The salient features of TPM are the involvement of operators in carrying out autonomous maintenance by participating in cleaning, lubrication, house-keeping, minor repair, adjustments, and so on. The benefits of TPM can be very tangible. Through implementation of TPM, organizations have separately been able to increase the production volume, and reduce downtime and rate of defective products. TPM also offers various intangible benefits such as fostering team work, increasing morale and safety, and nurturing the workforce.

MANAGEMENT SCIENCE (MS) MODELING

Yet, it is very difficult to answer the main questions faced by maintenance personnel: whether its output is produced effectively in terms of its contribution to company profits and efficiently in terms of manpower and materials employed.

In theory, maintenance personnel facing these problems could have benefited from the advent of an area of MS called *maintenance optimization*. Basically, a maintenance optimization model is a mathematical model in which both costs and benefits of maintenance are quantified and in which an optimal balance between both is obtained.

Here, we define maintenance optimization models as mathematical models whose aim is

to find the optimal balance between the costs and benefits of maintenance, while taking all kinds of constraints into account. In almost all cases, maintenance benefits consist of savings on costs which would be incurred otherwise (e.g., less failure costs).

In general, applications of maintenance optimization models cover the following four aspects:

1. a description of a technical system, its function, and its importance;
2. a modeling of the deterioration of the system in time and possible consequences for the system;
3. a description of the available information about the system and the actions open to management; and
4. an objective function and an optimization technique which helps in finding the best balance.

The maintenance objectives can be summarized under the three headings as follows:

1. ensuring system function,
2. ensuring system life (e.g., civil structures like buildings, roads, dam), and
3. ensuring safety (e.g., airplanes, nuclear plants, and chemical plants).

For production equipment, sustaining system availability in a cost-effective manner should be the prime maintenance objective. Here, maintenance has to provide the right (but not maximum) reliability, availability, efficiency, and capability (i.e., producing at the right quality) of production systems, in accordance with the need for these characteristics. In principle, it is possible to give an economic value to the maintenance results and a cost-balance can be done.

Maintenance optimization models have numerous benefits and advantages. Examples are as follows:

1. different maintenance policies can be evaluated and compared with respect to cost-effectiveness and reliability characteristics;

2. insight can be obtained on the structure of optimal policies, like the existence of an optimal control-limit policy;
3. models can assist in defining the maintenance concept—how often to inspect or to maintain; and
4. models can be of help in determining effective and efficient overall scheduling plans, taking all constraints into account.

MAINTENANCE MODELING

Decisions associated with managing the maintenance function could be divided into engineering decisions and operational decisions. The former relate to the choice of what engineering actions to actually take and when to take them (i.e., determine the maintenance concept itself), while the latter relate to the operations aspect of implementing decided engineering actions (e.g., scheduling shutdown, managing manpower, logistics, and inventory control); hence influencing the efficiency of implementing a maintenance concept.

By maintenance model, we mean a model specifically built for the maintenance function and which, unlike the other models used in the management of maintenance, actually addresses decision makers relating to the engineering issues and consequences as opposed to operational choices [19].

A maintenance model is a device that enables one to relate a decision variable to its consequence in a maintenance concept. Data modeling is well understood, especially Weibull modeling [20], and reliability-based modeling is still a very novel concept among engineers; especially the modeling of engineering aspects of maintenance. Failure data will be modeled, but not maintenance decision problems (e.g., Weibull analysis is applied to a set of reliability data collected under a particular maintenance practice).

The problem here is the difficulty of relating maintenance expenditure to production output, quality, and cost. Overcoming this requires maintenance modeling, which

relates, for example, the following maintenance decision variables to outputs or returns of interest.

1. Cost-effective maintenance for production is possible if the consequences upon production of maintenance activity are both recognized and known quantitatively. Understanding such a complex interaction requires a maintenance modeling exercise. The notion of actually modeling maintenance is novel. Routinely collected data are usually inadequate for the maintenance modeling task, and that additional information will usually need to be obtained from experienced maintenance engineers in order to construct maintenance models.
2. The degree of interest one has in the reliability of a system and the standard of reliability to be achieved are closely coupled to the consequences of unreliable behavior. Improving reliability will, in general, cost more, but the achievement of reliability usually saves money and sometimes saves lives. Accordingly, the need to maintain an "economic balance" determines the level of reliability, to which one should aspire in the design/maintenance of components and systems. Instead of asking of a system, "Is it reliable?" one should ask, "Is it reliable enough?" Finding an answer to this question obviously requires the quantification of reliability, which is achieved by using the theory of probability and statistics.

The relationship between reliability, cost, and maintenance requirements is important. Examples of high-reliability systems are all around us, from aircraft systems, electric power generating stations and chemical plants to communication systems, and computer systems.

Improving the reliability of a system costs money and is economically justified if the decrease in the expected cost of failure is at least equal to the expected cost of the improvement in reliability. In the past, only high-technology areas in

aerospace and defense were the recipients of reliability improvement efforts. However, as we enter the twenty-first century, high technology is permeating almost all aspects of life. The need to maintain high levels of productivity and competitiveness in the international marketplace indicates that efforts at improving reliability can have a broad, positive impact in manufacturing and service industries as well. Once an improvement in reliability is shown to be justified economically if the data inputs are realistic, the question of whether it is too expensive becomes a moot point. In all our discussions, we should not overlook the ever-increasing cost of liability.

The options available for improving reliability obviously depend on the problem at hand. In most cases, the options can be grouped into four basic categories:

1. providing redundancy,
2. designing environmental capability,
3. performing environmental testing (detailed studies by manufacturers about the failures of their products resulted in better designs, with less failures as a result), and
4. selecting maintenance strategy and servicing.

In practice, a combination of the four options would be considered, and some acted upon if felt appropriate. The nature of the mix will depend on the ratio of incremental benefit to incremental cost as a function of cost for each of the four options, the role played by risk, and budgetary considerations. Within specific cases, the definitions of "incremental benefit" and "incremental cost" are often nontrivial.

There are two aspects to modeling maintenance, the first being to seek to reduce the number of faults arising and their consequence, and the second the subsequent maintenance concept as an aid to decision making. To reduce the number of defects arising, it is necessary to identify the cause of defects and possible means of removing or reducing the cause. This is a key part in forming a maintenance concept [9] and

is directly associated with snapshot modeling [21]. When there are no data to be fitted, subjective information from engineers is of value and can be obtained by administering questionnaires [22]. At any maintenance intervention, there is a wealth of information potentially available, both objective (e.g., what has failed, what is being replaced) and subjective (e.g., what caused the defect and whether the occurrence is preventable). A systematic method of doing this has been called *snapshot modeling*.

Maintenance optimization models, embedded in decision-support systems, provide an objective and quantitative way of decision making in which the expected consequences of not maintaining, or maintaining in a different way, become clear. Although the number of real applications of maintenance models published in the literature is relatively small, these case studies show that mathematical models are a good means of achieving both effective and efficient maintenance. The factors that may have hampered the applications are discussed in Dekker [23], Scarf [24], and Dekker and Scarf [25].

MEDICAL MAINTENANCE

Someone who has worked on both industrial and medical maintenance problems will be struck by the differing preoccupations of workers in the two fields. Healthcare workers are very interested in ensuring that their models fit the real world, and in fact their distrust of models is such that, even after formulating the models, they will still wish to test the efficacy of screening in a particular age range directly by a medical trial. However, although they will evaluate different policies, they do not have the concept of an optimal policy. In industrial maintenance modeling, in contrast, there is great interest in finding optimal policies for many seemingly arbitrary models, but often little interest indeed in comparing models with data. It is seldom indeed that there is any attempt to show by direct statistical methods that a move to optimize maintenance has produced a net benefit.

Periodic or nonperiodic medical screening (or mass medical screening)—the deliberate examination of a single individual (or a large group of individuals) in a search for disease at its earliest stages—is a logical extension of the role of preventive medicine. In the medical screening problem, we seek the vector of ages at which people should be screened in order to minimize a suitable cost function. There is a close analogy with industrial inspection, where a machine is inspected for defects (that can give rise to failures) several times during its service life. In the industrial case, defective components can be replaced or repaired, hence averting a costly failure. One medical analog of such a defect is a tumor, which once detected at screening can be treated, [26] and the journal *Health Care Management Science*. The rationale of screening is that the prognosis is better when treating malicious tumors. In industrial inspection, a machine would probably be inspected for all possible defects simultaneously. In medical screening, however, inspection merely attempts to detect one particular type of defect. This means that mortality from other causes also occurs, and must be modeled.

If it is established that a particular screening procedure is medically desirable, there remains the practical necessity of determining whether or not the costs involved make the procedure worthwhile. Clearly, in these days of scarce resources, it is necessary to search for screening policies that will make the procedures as efficient as possible. The aim is to choose a screening policy that minimizes, say, the total cost of screening. Many variables, some of them stochastic in nature, influence the efficiency of a screening program. One difficulty here is that, in general, it is not possible to express the risks, costs, and benefits of a screening procedure in the same common and easily understood measure, for example, monetary units. The testing expense includes easily quantifiable economic costs such as those of the labor and materials needed to administer the testing. However, there can also be other important cost components which are more difficult to quantify; for example, in the case of a human population subject to medical

screening, the cost of side effect and that of false positives or false negatives, which entail both emotional distress and the need to do unnecessary follow-up testing and even the risk of physical harm to the testee, for example, the cumulative effect of X-ray exposure. We agree that the usual cost-benefit analysis is unlikely to suffice in this very complex situation involving pain, suffering, and human life. Nevertheless, some quantitative guides are essential for making good decisions about screening procedures. Suitable mathematical models, if available, would help the medical decision makers confronted with this complex situation.

Many of the techniques and approaches of industrial maintenance management mentioned in the sections titled "Industrial Maintenance Problems," "Importance of Maintenance," "Maintenance Area," "Maintenance Personnel," "Maintenance Concept," "Management Science (MS) Modeling," and "Maintenance Modeling" are relevant to medical screening problems.

CONCLUSIONS

Research into production technologies should consider maintenance. The significance and complexity of maintenance in modern production and manufacturing systems indicate that maintenance should be viewed as an increasingly important area of study and research.

In the follow-up article "Reviews of Maintenance Literature and Models", we review some basic maintenance models such as age replacement and block replacement. In addition, we list a series of survey papers, handbooks, and monographs on maintenance research dating from 1960 to 2010.

REFERENCES

1. Waddington CH. OR in world war II: OR against the U-boat. London: Elek Science; 1973.
2. Khintchine AY. Mathematisches über die Erwartung von einen' öffentlicher Schalter. Matem. Sbornik 1932.
3. Drenick RF. The failure law of complex system. J Soc Ind Appl Math 1960;8:680-690.
4. Wang WB, Jia X. An empirical Bayesian based approach to delay time inspection model parameters estimation using both subjective and objective data. Qual Reliab Eng Int 2007;23:95-105.
5. Wang WB, Zhang W. An asset residual life prediction model based on expert judgments. Eur J Oper Res 2008;188:496-505.
6. Christer AH. A possible future for industrial maintenance within India. Technical report. Centre of Operational Research and Applied Statistics, Salford University; 1995. pp. 1-26.
7. Lam Y. The geometric process and its applications. Singapore: World Scientific; 2007.
8. Leung KNF. Arithmetic and geometric processes. Pham H, editor. Handbook of engineering statistics. Germany: Springer; 2006. Chapter 49. pp. 931-955.
9. Gits CW. Design of maintenance concepts. Int J Prod Econ 1992;24:217-226.
10. Christer AH, Waller WM. An operational research approach to planned maintenance: modeling planned maintenance for a vehicle fleet. J Oper Res Soc 1984;35:967-984.
11. Christer AH. Developments in delay-time analysis for modeling plant maintenance. J Oper Res Soc 1999;50:1120-1137.
12. Moubray J. Reliability-centered maintenance. 2nd ed. Singapore: Butterworth-Heinemann; 1997.
13. Rausand M. Reliability-centered maintenance. Reliab Eng Syst Saf 1998;60:121-132.
14. Jia X, Hao J. Reliability-centered maintenance analysis for a complex military system. IMA J Math Appl Bus Ind 1995;6:135-143.
15. Jia X, Christer AH. A prototype cost model of functional check decisions in reliability-centered maintenance. J Oper Res Soc 2002;53:1380-1384.
16. Nakajima S. TPM: challenge to the improvement of productivity by small group activities. Mainten Manage Int 1986;6:73-83.
17. Nakajima S. Introduction to TPM: total productive maintenance. Cambridge (MA): Productivity Press; 1988.
18. Nakajima S. TPM development program: implementing total productive maintenance. Cambridge (MA): Productivity Press; 1989.
19. Baker RD, Christer AH. Review of delay-time operational research modeling of engineering aspects of maintenance. Eur J Oper Res 1994;73:407-422.
20. Murthy DNP, Xie M, Jiang RY. Weibull models. Hoboken (NJ): Wiley; 2004.

21. Christer AH, Whitelaw J. An operational research approach to breakdown maintenance: problem recognition. *J Oper Res Soc* 1983;34:1041–1052.
22. Wang WB. Subjective estimation of the delay-time distribution in maintenance modeling. *Eur J Oper Res* 1997;99:516–529.
23. Dekker R. Applications of maintenance optimization models: a review and analysis. *Reliab Eng Syst Saf* 1996;51:229–240.
24. Dekker R, Scarf PA. On the impact of optimization models in maintenance decision making: a state of the art. *Reliab Eng Syst Saf* 1998;60:111–119.
25. Scarf PA. On the application of mathematical models in maintenance. *Eur J Oper Res* 1997;99:493–506.
26. Yakovlev AY, Tsodikov AD. Stochastic models of tumor latency and their biostatistical applications. Singapore: World Scientific; 1996.

MANAGEMENT OF NATURAL GAS STORAGE ASSETS

DEREK WANG
Desautels Faculty of
Management, McGill
University

In North America and Europe, natural gas storage plays an essential role in managing the mismatch between energy supply and demand. Consider the United States as an example [1]: daily production of natural gas is relatively constant at 52 billion cubic feet per day (Bcf/d), whereas consumption exhibits significant seasonal variations: 55 Bcf/d during the nonheating season (April through October) and 70 Bcf/d during the heating season (November through March). The seasonal supply–demand imbalance makes it necessary to build underground natural gas storage facilities throughout the United States, in particular, in the Gulf production area and the North East and Midwest consumption areas.

There are three main types of underground natural gas storage facilities [2], (i) depleted oil/gas reservoirs, (ii) aquifers, and (iii) salt caverns. There are 410 underground storage sites in the United States, with a total storage capacity of about 8.8 trillion cubic feet (Tcf) in 2011 [3]. The Federal Energy Regulatory Commission (FERC) Order 636 in 1992 requires that storage facilities operate on an open-access basis such that storage owners can only control enough capacity to maintain system integrity and the rest capacity must be made available for lease to third parties. In such a deregulated market, the owners of storage facilities include pipeline companies, local distribution companies, and independent storage service providers. Deregulation opens the storage market to competition and expands the use of storage beyond its traditional role of supply–demand balancing. Today, storage operators can move gas into and out of storage to capitalize on gas price variations.

Physical constraints need to be considered in managing natural gas storage assets. The injection and withdrawal capacities of natural gas are governed by the Bernoulli's law and, hence, depend on the pressure differential between the storage and the pump/pipeline. The more the natural gas is in the storage, the higher the pressure in the reservoir, hence the slower the maximum injection rate and the faster the maximum withdrawal rate. For the ease of implementation in reality, industry practitioners approximate the real injection/withdrawal capacities with piece-wise constant capacities, which are called *ratchets* as they ratchet up and down with inventory. For a typical natural gas depleted reservoir storage facility [4, p. 456], the maximum injection rate is initially constant and then declines steadily in response to the higher pressure within the gas storage as working inventory builds up. A similar but reverse trend is observed for the withdrawal rate. It takes about four months to fill up or deplete the storage and eight or nine months to complete a cycle. Salt caverns are relatively small but provide the highest withdrawal and injection rates relative to their storage volume. Aquifers lie somewhere in between. The depleted gas/oil reservoirs hold more than 80% of the total capacity in the United States and are mainly used to meet seasonal demand variations and capture seasonal price differentials. Owing to their high deliverability, the salt caverns are widely used as peak-load storage to meet demand variations and exploit profit opportunities in the time scale of weeks, days, and even hours.

Storing and releasing natural gas also involve operational frictions. For example, the pumps of the storage facility use some of the gas as fuel [5]. If q units are to be added to the storage, the firm needs to purchase $(1 + \alpha)q$ units; if q units are withdrawn from the storage, a fraction βq will be lost and $(1 - \beta)q$ can be sold. In addition to the volume losses, the firm also incurs a variable

cost of $c_\alpha q$ when q units are stored and a cost of $c_\beta q$ when q units are withdrawn. These costs can cover the use of pumps and other equipment. For depleted reservoirs, the injection loss rate α is typically between 0% and 3%, the withdrawal loss rate β is between 0% and 2%, and the variable operating costs c_α and c_β are below \$0.02 per million British thermal unit (mmBtu).

Storage assets owners and third-party lease contract holders have great incentives to optimize the utilization of storage facilities to maximize their profit. The value of storage assets primarily arises from the natural gas summer-to-winter price differentials. Reflecting the seasonal supply-demand imbalance, the natural gas futures market on the New York Mercantile Exchange (NYMEX) prices summer contracts at a significant discount in relative to the winter contracts. Figure 1 shows the natural gas futures prices observed on the first trading day of March in recent years. The seasonal price pattern allows a storage operator to capture the summer-to-winter price spread by injecting gas in the summer and withdrawing in the winter. In March, the

firm (owner or leaseholder of a natural gas storage asset) can decide an injection-and-withdrawal schedule from April through to the next March by taking long or short positions on futures contracts. Once these futures positions are determined (e.g., by solving a static profit maximization problem subject to certain physical constraints discussed in detail in the next section), the firm essentially locks in a risk-free profit. This value is referred to as the *intrinsic value of the storage asset*.

Optimization of storage operations is considerably more challenging than solving for intrinsic value, because it involves multiple interacting real options, that is, options to store or withdraw within capacity limits in every period. References 6 and 7 provide the theoretical background of real options. A crucial distinction of applying real option approach to storage valuation from its general application is the physical constraints constraining the procuring/selling decisions. Explicit analysis of storage operations is possible when physical constraints are ignored [8]. Analytical valuation of the general storage options typically does not exist because

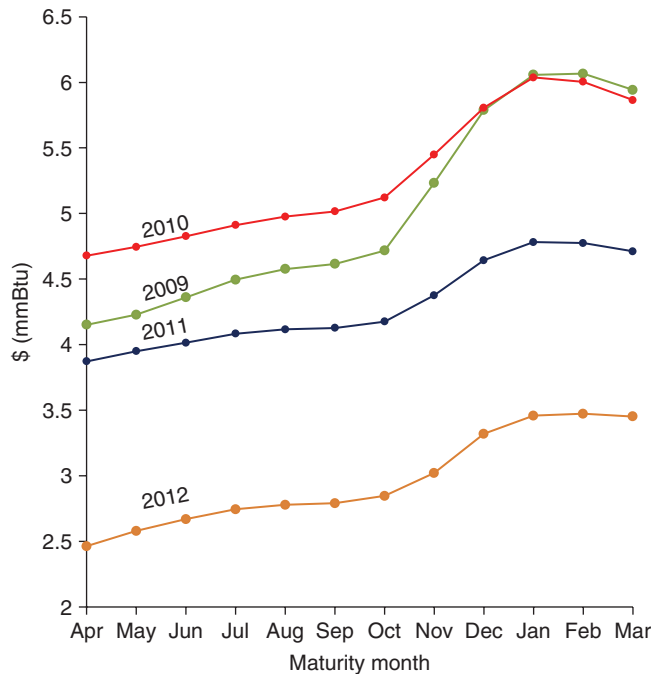


Figure 1. Natural Gas Forward Curve on the First Trading Day of March (2009–2012).

of the injection and withdrawal constraints. The section titled “Exact Model of Storage Operations” outlines the management of gas storage under the real option framework. In general, finding the optimal storage policy is analytically and computationally challenging. Consequently, heuristic methods have been developed in practice and studied in academia and are discussed in the section titled “Heuristic Policies.” Theoretical and numerical techniques have been proposed to improve heuristics and are summarized in the section titled “Beyond Heuristics.” A brief discussion of potential future works is presented in the section titled “Future Directions.”

EXACT MODEL OF THE STORAGE OPERATIONS

The valuation under the real option framework can be formulated as a continuous-time stochastic control problem [9–12] or discrete-time stochastic dynamic program [4,10,13,14].

We first explain the discrete-time models. At the beginning of period t , the futures price for delivery in period t is maturing, denoted as f_{tt} . The firm sees the maturity price and other futures prices f_{ts} ($s > t$) and decides the quantity to purchase or sell at price f_{tt} . The settled amount is then stored into or withdrawn from the storage over the entire period. Under the standard no-arbitrage assumption, the futures prices are martingales under an equivalent martingale measure Q . The firm can adjust futures positions any time before they expire. As futures prices are martingales, those adjustments do not affect the expected value of the energy storage because the expected marked-to-market cash flow is zero under Q -measure. Owing to the operational frictions, consider $\tilde{f}_{ts} = e^{-R(s-t)}[(1-\beta)f_{ts} - c_\beta]$ as the forward selling price and $\tilde{f}_{ts}^b = e^{-R(s-t)}[(1+\alpha)f_{ts} + c_\alpha]$ as the forward buying price of inventory discounted to the first period. It is straightforward to verify that both $\tilde{f}_{ts}$ and $\tilde{f}_{ts}^b$ are martingales. Let $\tilde{f}_t = (\tilde{f}_{ts} : s \geq t)$ denote the discounted forward prices observed at the beginning of period t . Let y_t be the ending

inventory in period t , which is decided by the firm at the beginning of period t . The storage valuation problem can be written as:

$$V_t(x_t, \tilde{f}_t) = \max_{y_t \in [\underline{y}(x_t), \bar{y}(x_t)]} r(y_t - x_t, \tilde{f}_{tt}) + E_t^Q[V_{t+1}(y_t, \tilde{f}_{t+1})],$$

where $r(q, \tilde{f}_{tt})$ is the one period revenue function; E_t^Q denotes the expectation under the equivalent martingale measure Q conditional on information until period t ; and y_t is bounded between the inventory-dependent capacities $\underline{y}(x_t) = x_t + \underline{\lambda}(x_t)$ and $\bar{y}(x_t) = x_t + \lambda(x_t)$. Here, $\lambda(x_t) \geq 0$ models injection capacity and $\underline{\lambda}(x_t) \leq 0$ models withdrawal capacity. In the last period, the firm sells as much as possible to maximize profit, $V_N(x_N, \tilde{f}_{NN}) = -\tilde{f}_{NN} \lambda(x_N) - \underline{y}(x_N)p$, where $p \geq 0$ denotes the penalty per unit of inventory at the end of period N , and p is realized in period N and may depend on the market price [9,12,15].

Continuous-time representations of gas storage valuation as a Markov decision process are described in Refs 9–12. Let $\pi(t, x_t, p_t, q_t)$ denote the instantaneous profit at t , where the inventory level is x_t , p_t is the gas price under Q -measure and follows Markov continuous-time stochastic process, and the injection/withdrawal rate q_t is bounded between injection/withdrawal capacities. Then, the expected profit-to-go at t under some admissible policy u is

$$V_t^u(x_t, p_t) = \int_t^T e^{-R(s-t)} \pi(s, x_s^u, p_s, q_s^u) ds + e^{-R(T-t)} V_T^u(x_T^u, p_T),$$

where the inventory evolves according to $dx_t^u = -q_t^u dt$, and $V_T^u(x_T^u, p_T)$ is the terminal value of inventory. The objective is to find an admissible policy to maximize the profit

$$V_t(x_t, p_t) = \sup_u V_t^u(x_t, p_t).$$

We comment that continuous-time and discrete-time formulations can be used for both spot and futures market valuations. In

the discrete-time model, the formulation continues to hold for spot market valuation, with f_{tt} interpreted as the spot price in period t and f_{ts} interpreted as the forecast in period t for the price in period s . The continuous-time model can be interpreted as valuation through either spot or futures market, as p_t can represent the spot price or the price of the near-month forward contract. Therefore, Carmona and Ludkovski [12] state, “given a variety of quoted gas prices (spot, balance-of-the-month, futures, etc.), we remain agnostic about the precise interpretation.”

Price Models

Price dynamics are crucial in the valuation problem, and a broad range of price models have been used. Driftless geometric Brownian motions are widely used to model the futures prices under the equivalent martingale measure [4,14,16,17]. In an n -factor model, f_{ts} , the price of contract with maturity s , evolves according to

$$\frac{df_{ts}}{f_{ts}} = \sum_{i=1}^n \sigma_i(t,s) dW_t^i,$$

where W_t^i is an independent standard Brownian motion and the corresponding volatility $\sigma_i(t,s)$ is a function of time and maturity. The spot price is usually modeled as a mean-reverting process. Jumps are often incorporated into the mean-reverting dynamics to model price spikes observed in the natural gas markets [9–11]. Therefore, the general form of spot price is

$$dp_t = \mu(t,p_t)dt + \sum_{i=1}^n \sigma_i(t,p_t)dW_t^i + \sum_{j=1}^k \delta_j(t,p_t)dN_t^j,$$

where N_t^j is an independent Poisson process and δ_j is the jump size. Some articles do not explicitly specify the processes followed by futures prices but derive futures prices from the spot price process [17,18].

Two salient features of natural gas price are convenience yield and seasonality. Convenience yield, defined as the premium of holding the physical commodity over the futures contract, can be incorporated into the spot price drift term $\mu(t,p_t)$ as in Ref. 18. Seasonality in futures market is reflected by the shape of the forward curve and the time-dependent volatility $\sigma_i(t,s)$. Seasonality in spot price can be captured by $\sigma_i(t,s)$ and seasonally adjusted mean-reverting level in the drift term $\mu(t,p_t)$ [13,16,17].

Carmona and Ludkovski [12] abstract from specific price models by treating the price as a general multidimensional Markov process. They demonstrate that their simulation approach is able to deal with complex price processes such as mean-reverting process with jumps. In general, high dimensional models more precisely capture price dynamics but are more difficult to estimate and require more computational efforts. Further discussions of gas price modeling can be found in Ref. 19.

Structure of Optimal Policy

Structural properties of the optimal policy have been documented. Hodges [8] formulates the storage operation as an infinite horizon optimal control problem without operational frictions. The optimal control is to fill up the storage as soon as the price falls below a critical price p^* and sell stock as soon as price rises above p^* . References 14 and 20 show that the optimal policy is characterized by two stage- and price-dependent base-stock targets: if inventory falls between the two targets, it is optimal to do nothing; otherwise, the firm should inject or withdraw to bring the inventory as close to the nearest target as possible. In the continuous-time framework [12], the optimal policy has a bang-bang structure. The control policy has only three choices: inject at the maximum rate, withdraw at the maximum rate, or do nothing. A sufficient condition for the above structural result is that the storing (releasing) capacity function (e.g., ratchets) is concave (convex) in inventory. The dependence of operation policy on price process has also been investigated [10]. If the price is stochastically increasing, a mild assumption

holding for many price processes, then, under higher price, the operator should inject less or withdraw more.

Numerical Methods

Practitioners and researchers have resorted to numerical methods for solving the exact storage operations problem. Three computational methods are used: numerical partial differential equation techniques [9,11,12], binomial/trinomial trees [21,22], and Monte Carlo simulation [12,13]. Least squares Monte Carlo method (approximation of the expected value function with ordinary least square regression) to value storage assets can incorporate very general physical constraint [13]. The regression may become slow when the full spectrum of spot and forward prices is considered. In the optimal switching approach [12], the model can accommodate general price models and allows the decision maker to make frequent trading decisions. Tree-based methods can use one- or two-factor price models, and three-factor model has never been implemented so far.

HEURISTIC POLICIES

Significant computational burden arises in real applications of the exact stochastic dynamic program where the optimization problem involves multidimensional futures prices evolving in a larger state space over a longer horizon. As stated in Ref. 14, “the number of maturities N associated with natural gas storage contracts is at least twelve, so that model is computationally intractable because of its high-dimensional state space.” Consequently, heuristic methods have been developed in practice and studied in academia.

Practitioners typically use two heuristic strategies to value seasonal energy storage: the rolling intrinsic (RI) approach and the rolling basket of spread options approach [23–26]. They are also referred as *reoptimized intrinsic value strategy* and *reoptimized linear program strategy*, respectively, in Ref. 14. Both policies are well received by the industry practitioners in the operation of low deliverability storage because of

their simple logic and ease of implementation. However, the efficacy and accuracy of these heuristics have not been established for high deliverability storage for which operations are scheduled more frequently on daily or hourly basis. The results in this section are all derived for the operation of low deliverability storage.

Rolling Intrinsic Policy

To define the RI policy, we first define the intrinsic policy, a policy that decides in the first period the actions to be performed in each of the remaining periods. The intrinsic policy is found by solving an optimization problem using only the forward prices seen in the first period. The corresponding value is called the *intrinsic value*. The RI policy is pseudodynamic: Every day, the firm solves a static intrinsic value optimization problem to decide the futures positions and reoptimizes the futures positions in each period by solving the intrinsic valuation problem using the updated forward prices. The corresponding value is referred to as the *RI value*.

The RI policy can be mathematically formulated as a dynamic programming problem [4]. First, given the discounted forward prices $\tilde{f}_t$, the intrinsic value of the storage is determined by:

$$\begin{aligned} V_N^I(x_N, \tilde{f}_t) &= -\tilde{f}_{tN} \underline{\lambda}(x_N) - \underline{y}(x_N) E_t^Q[p], \\ V_s^I(x_s, \tilde{f}_t) &= \max_{y_t \in [y(x_t), \bar{y}(x_t)]} r(y_s - x_s, \tilde{f}_{ts}) \\ &\quad + V_{s+1}^I(y_s, \tilde{f}_t), \quad t \leq s < N. \end{aligned}$$

Let $y_t^\dagger$ be the futures position on the maturing contract $\tilde{f}_{tt}$ in period t , obtained from the above deterministic intrinsic value optimization. Computing $y_t^\dagger$ requires no assumptions on the stochastic evolution of the forward curve. In RI policy, the firm solves the above deterministic optimization problem in every period with updated forward curve $\tilde{f}_t$ and adjusts the futures positions accordingly. The RI value of the storage is then determined by:

$$V_N^{\text{RI}}(x_N, \tilde{f}_N) = V_N^I(x_N, \tilde{f}_N),$$

$$V_s^{\text{RI}}(x_t, \tilde{f}_t) = r(y_t^\dagger - x_t, \tilde{f}_{tt}) + E_t^Q V_{s+1}^{\text{RI}}(y_t^\dagger, \tilde{f}_{t+1}),$$

$$1 \leq t < N.$$

The RI policy solves a deterministic optimization problem every period and may miss potential option values rising from the stochastic evolution of the forward curve. Further discussion of missing option value is presented in the section titled “When Are the Heuristics Suboptimal” in this article.

Rolling Basket of Spread Options Approach

The spread options approach benchmarks the natural gas storage value using a portfolio of calendar spread options [14,26]. According to the NYMEX rule, “buying a call (put) on the natural gas calendar spread options contract will represent a long (short) position in the closer natural gas futures contract month and a short (long) position in the longer-dated natural gas futures contract month.” In practice, injection and withdraw operations always appear in pairs: If one unit of natural gas is injected into the storage, it will be withdrawn later at some date. The operation of injecting one unit of natural gas at time t at price f_{tt} , and then withdrawing it later at time s ($s > t$) at price f_{ts} resembles the exercise of a European call spread option. The real profit of such an operation is random because of the stochastic futures prices, but the expected present value of such an operation can be approximated by the price of the spread option P_{ts} . P_{ts} can be numerically computed or approximated using Kirk’s formula.

Suppose now is day 1, and as the owner of the storage assets, we have decided a certain injection/withdrawal schedule. Then, the schedule can be thought of as a basket of spread options. To maximize the expected profit under this schedule, we only need to find the optimal portfolio with the highest value. Let q_{ts} ($s > t$) be the quantity of the spread option P_{ts} in the portfolio. Then, the optimal portfolio can be found as follows:

$$\max_{q_{ts}} \sum_{t < s} P_{ts} q_{ts}$$

$$\text{s.t. } q_{ts} \geq 0,$$

where q_{ts} subject to physical constraints.

The physical constraints ensure that the injection/withdrawal operations represented by the portfolio are compatible with the physical limits of the storage. The number of variables depends on planning horizon. For the standard April-start/March-end contract, there are 66 potential spread options to trade. The above optimization yields the day 1 intrinsic value of the storage. The portfolio is dynamically adjusted every period by resolving the above optimization problem conditional on updated forward curve information, that is, the portfolio of spread options is getting rolled. The above portfolio optimization problem reduces to linear optimization under constant injection/withdrawal capacities. Integer linear programming is needed to deal with nonlinear capacities.

Performance of Heuristics

The RI policy has near-optimal performance in many circumstances [14] but can significantly underperform in some cases [4]. Lai *et al.* [14] used an approximate dynamic programming approach coupled with information relaxation to derive upper bounds on value of storage with constant capacities. They assess the performance of different approaches under either 12-month or 24-month contract terms starting from four different dates in 2006. Using a two-factor price model, Wu *et al.* [4] examined RI valuation in each of the nine years from 2001 to 2009 under different ratchets and discount rates. The RI valuation has a median value loss of 0.59% compared with the optimal valuation. For 15% of the instances, the value loss of RI policy is at least 1% (1% value loss is roughly \$0.1 million for a 10-trillion Btu storage under \$1 per million Btu seasonal price spread). The maximum value loss is 9.23%.

In practice, people often find that spread options approach yields higher value than the intrinsic approach, because the former captures the option value. But, when the two strategies are “rolled” or reoptimized frequently, their performance is empirically the same in terms of the expected value [23,24]. Gray and Khandelwal [24, p. 4] state,

“we have found empirically that, in general, the rolling intrinsic value is equal to the rolling basket value.” This is also confirmed by Ref. 14, which finds that both strategies capture more than 97% of the upper bound on average.

In the RI approach, the funding risk of daily mark-to-market exposure can be significant. A large marked-to-market loss might require a posting of collateral. Although this would ultimately be mitigated by the execution of the physical hedge, funding capital might be at risk, possibly resulting in early termination. In the basket spreads, daily marked-to-market profits and losses are part of the game, as both must be born out in a true dynamic hedging scenario. In terms of value distribution, the RI has a guaranteed lower bound equal to the initial intrinsic value, thereby creating a “truncated” distribution. This is because it only rebalances the futures positions on favorable price movements.

When Are the Heuristics Suboptimal?

In order to improve the operations, it is critically important to understand why the heuristic decisions deviate from the optimal decisions. Several situations can be identified. These situations serve as “red flags” for practitioners to consider their operations before applying the RI strategy.

Some situation has a flavor similar to the value of waiting found in real options applications. For example, during the peak selling season, if all future prices are expected to be lower than the current price, RI policy is to sell as much as possible. But, waiting may be beneficial because the firm has the flexibility of scheduling operations to capture the expected maximum of future prices, which may be higher than the current price. Waiting for a month means that the firm has one less month to schedule the remaining operations, and it would have to sell inventory at whatever market price. Waiting may also forgo the option of counter-seasonal operations (e.g., buying in the selling season). On top of the value derived from seasonal price spread, the option of counter-seasonal operations provides additional value from within-season price fluctuations. Thus, striking a balance between the value of waiting

and the value of potential counter-seasonal operations is necessary.

When the storage can be sold later in the season and the current price is the lowest compared to the expected future prices, waiting for higher prices might seem to be optimal, as prescribed by the RI strategy. Contrary to the intuition, the optimal strategy may involve selling inventory at the lowest price. In fact, if the firm sells one unit at the current low price, it will sell one less unit at the lower of the two future prices (avoiding the adverse price), which may be even lower, in expectation, than the current price.

BEYOND THE HEURISTICS

Heuristic methods in last section are developed based on practitioners’ intuitions and experiences, rather than derived from real options theory, because of the theoretical challenges in analyzing storage options. In recent years, several approaches have been developed by academia to close the gap between the heuristics and optimality. At \$5 per mmBtu, filling a typical 10-Bcf depleted reservoir storage facility costs \$50 million. With a summer-to-winter spread of \$1 per mmBtu, the storage is expected to make a profit of at least \$10 million. Even small improvement in operations can bring significant benefit to the storage owners and lease contract holders.

Upper Bounds of Storage Value

Upper bounds on the storage value can be derived with approximate dynamic programming and benchmark various policies when exactly solving for the optimal storage value is not feasible under high dimensional price models [14]. In this approach, the intractable high dimensional stochastic dynamic program is transformed into a low dimensional dynamic program through information relaxation, where the decision maker is given anticipative information on price evolution and the information renders the program simpler to solve. The approach is quite general and can accommodate different price models and long planning

horizons (24-month contract term is used in Ref. 14). The approximate dynamic programming method consumes more CPU time than RI or rolling spread options approach but can generate very tight and effective upper bounds. The results confirm that RI and rolling spread options methods are nearly optimal for spring, summer, and fall scenarios. The gap between the RI/rolling spread options and the upper bound is wider for those winter scenarios.

Price-Adjusted Rolling Intrinsic

The idea of the price-adjusted rolling intrinsic (PARI) policy is to adjust the forward curve to bring the adjusted intrinsic values closer to the option values and then apply the RI policy on the adjusted forward curve to improve the RI decision [4]. It can be analytically proved that there exists a set of price adjustments under which the adjusted intrinsic values equal the option values in a three-period model under general price distributions and capacity functions. The same logics are then adapted to N -period problems. While computing PARI decision requires assumptions on the stochastic evolution of the forward curve, the price adjustment can be done without solving a stochastic dynamic program, and hence the PARI policy remains computationally tractable.

FUTURE DIRECTIONS

There are several untapped issues on the management of gas storage. First, owing to the illiquidity of the spot market, perfect hedging does not exist. Hence, advanced risk objectives may be taken into account to complement the existing models of expected value maximization. Moreover, the proliferation of gas-fired power plants has spurred the rapid growth of high deliverability storage capacity in recent years. It is of significant value to develop heuristic methods for storage with daily and hourly reoptimization, where existing heuristics may fail because of heavy computational burden and lack of financial instruments on a more granular time scale. Finally, existing models typically make a simplifying assumption that the capacities

are constant to help simplify analysis. "This assumption, widespread in the industry, is very convenient," as commented in Ref. 5, p. 436, and may be a good approximation when the storage level is within a certain range. However, the importance of real level-dependent capacities has been recognized in Ref. 23, p. 1: "Very often, realistic issues such as bid/ask spread, discounting and, level-dependent rates (ratchets) and hard inventory level constraints are ignored or assumed away . . . Critically, storage valuation must be consistent with the physical constraints of the storage field." It is of both practical and academic value to understand how inventory-dependent capacities impact storage operations and storage value.

REFERENCES

1. Energy Information Administration. 2004. How EIA estimates natural gas production. http://www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/ngprod/ngprod.pdf. Accessed 2013 April 15.
2. Federal Energy Regulatory Commission. 2004. Underground natural gas storage. <http://www.ferc.gov/EventCalendar/Files/20041020081349-final-gs-report.pdf>. Accessed 2013 April 15.
3. Energy Information Administration. 2013. Underground natural gas storage capacity. http://www.eia.gov/dnav/ng/ng_stor_cap_dcunus_a.htm. Accessed 2013 Aug 29.
4. Wu O, Wang D, Qin Z. Seasonal energy storage operations with limited flexibility: the price-adjusted rolling intrinsic policy. *Manuf Serv Oper Mgmt* 2012;14(3):455–471.
5. Maragos S. Valuation of the operational flexibility of natural gas storage reservoirs. In: Ronn EI, editor. *Real Options and Energy Management*. London: Risk Books; 2002. pp. 431–456.
6. Dixit A, Pindyck R. *Investment under Uncertainty*. Princeton, NJ: Princeton University Press; 1994.
7. Schwartz E, Trigeorgis L. *Real Options and Investment under Uncertainty*. Cambridge, MA: MIT Press; 2001.
8. Hodges, S. 2004. The value of a storage facility. Working paper 2004. University of Warwick, Coventry, UK.
9. Chen Z, Forsyth P. A semi-Lagrangian approach for natural gas storage valuation

- and optimal operation. *SIAM J Sci Comput* 2007;30(1):339–368.
10. Kaminski V, Feng Y, Pang Z. 2008. Value, trading strategies and financial investment of natural gas storage assets. Northern Finance Association Conference 2008, Kananaskis, AB, Canada. <http://www.northernfinance.org/2008/papers/75.pdf>. Accessed 2013 April 15.
 11. Thompson M, Davison M, Rasmussen H. Natural gas storage valuation and optimization: a real options application. *Nav Res Logist* 2009;56(3):226–238.
 12. Carmona R, Ludkovski M. Valuation of energy storage: an optimal switching approach. *Quant Finan* 2010;10(4):359–374.
 13. Boogert A, de Jong C. Gas storage valuation using a Monte Carlo method. *J Derivatives* 2008;15(3):81–98.
 14. Lai G, Margot F, Secomandi N. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Oper Res* 2010;58(3):564–582.
 15. Buurma C. Inventories could trump cold for natural-gas prices. *Wall Street J* 2010 January 25.
 16. Kjaer M, Ronn E. Valuation of natural gas storage facility. *J Energy Markets* 2008;1(4):3–22.
 17. Manoliu M, Stathis T. Energy futures prices: term structure models with Kalman filter estimation. *Appl MathFinan* 2002;9(1):21–43.
 18. Schwartz E. The stochastic behavior of commodity prices: implications for valuation and hedging. *J Finance* 1997;52(3):923–973.
 19. Clewlow L, Strickland C. *Energy Derivatives: Pricing and Risk Management*. London: Lacima Publications; 2000.
 20. Secomandi N. Optimal commodity trading with a capacitated storage asset. *Mgmt Sci* 2010;56(3):449–467.
 21. Manoliu M. Storage options valuation using multilevel trees and calendar spreads. *Int J Theoret Appl Finan* 2004;7(4):425–464.
 22. Parsons, C. 2007. Valuing commodity storage contracts: a two-factor tree approach. WTM Energy Software LLC. Available at <http://ngstoragemodel.com/>. Accessed 2014 Jan 10.
 23. Gray J, Khandelwal P. Towards a realistic gas storage model. *Commodities Now* 2004 June, 1–4.
 24. Gray J, Khandelwal P. Realistic natural gas storage model II: trading strategies. *Commodities Now* 2004 September, 1–5.
 25. Eydeland A, Wolyniec K. *Energy and Power Risk Management*. New Jersey: John Wiley & Sons; 2003.
 26. Byers J. Commodity storage valuation: a linear optimization based on traded instruments. *Energy Econ* 2006;28:275–287.

MANAGEMENT SCIENCE/ OPERATIONS RESEARCH SOCIETY OF MALAYSIA

ILIAS MAMAT
Operations Research Society of
Malaysia, Permodalan Nasional
Berhad, Kuala Lumpur,
Malaysia

ORIGINS OF MSORSM

The Management Science/Operations Research Society of Malaysia (MSORSM) [1] was registered as a professional society under Section 7, Societies Act of Malaysia 1966, on July 19, 1986. The society was inaugurated on October 5, 1986. MSORSM was the 36th national society to become a member of the International Federation of Operational Research Societies (IFORS), the 9th national society to become a member of the Association of Asian Pacific Operational Research Societies (APORS), and is a member of the Confederation of Scientific and Technological Associations of Malaysia (COSTAM).

THE MISSION, STRATEGIES, AND ACTIVITIES OF MSORSM

The mission of MSORSM embodies several goals as follows:

- to be a well recognized body representing, promoting, and advancing management science/operations research;
- to play the lead role for the advancement of a progressive and well informed management science/operations research community;
- to promote and develop management science/operations research as an interdisciplinary field;

- to foster the application, wherever appropriate, in the widest possible exchange of information and ideas on related subjects; and
- to generate a group of actively participative communities for the advancement of the discipline of management science/operations research both within and outside Malaysia.

MSORSM works to achieve these objects through several strategies, including

- positioning into key committees that involve management science/operations research in the industry;
- advancing the practices of management science/operations research through seminars, workshops, discussions, and publications for the well-being of society at large;
- establishing academic and industry linkages for the application of management science/operations research tools.

MSORSM's strategies are in turn implemented through several activities, including

- to organize national and international conferences, seminars, IFORS workshops, professional talks, training modules, linkages, and collaborations with operations research societies worldwide, and linkages with industries;
- to publish newsletters, periodicals, journals, books, circulars, and information leaflets after due permission from relevant authorities;
- to give awards in recognition of outstanding academic or professional achievement in management science/operations research. These awards include Student Achievement Awards

in the following categories:

- PhD Thesis
- Master Thesis
- Master Dissertation
- Bachelor Project
- to raise funds and to accept gifts, donations of bequests with a view to furthering, either directly or indirectly, the objectives of the society.

ADMINISTRATION OF MSORSM

MSORSM is governed by an executive committee that consists of the President, Immediate Past President, Vice President, Honorary Secretary, Honorary Treasurer, two Honorary Auditors, eight at-large Council members, and a Patron. In 2009/2010 these offices are held by

- Patron: Dr. Halim Shafie
- President: Dr. Ilias Mamat
- Immediate Past President: Dr. Han Chun Kwong
- Vice President: Dr. Adnan Ahmad
- Honorary Treasurer: Mr. Liew Swee Liang

- Honorary Secretary: Dr. Adibah Shuib
- Honorary Auditors: Dr. Mohd. Sahar Sawiran and Associate Professor Hasni Hashim
- At-Large Council Members: Dr. Razidah Ismail, Mr. Amiruddin Jumaat, Associate Professor Adam Baharum, Dr. Mohd. Omar, Mr. Muhamad Shahbani Abu Bakar, Dr. Isahak Kassim, Dr. Amirudin Abd. Wahab, and Dr. Muhammad Rozi Malim

Past Presidents of MSORSM (with their terms in office) include

- 1986–1987: Dr. Han Chun Kwong
- 1988–1994: Dr. Othman Yeop Abdullah
- 1994–1995: Mr. Rodzlan Akib Abu Bakar
- 1995–1997: Dr. Han Chun Kwong
- 1997–2001: Dr. Halim Shafie
- 2001–2008: Dr. Han Chun Kwong

REFERENCES

1. MSORSM Constitution. Reg 3935/86. selangor (Incorporated under Societies Act 1966).

MANAGING A PORTFOLIO OF RISKS

JEFFREY M. KEISLER

Department of Management Science
and Information Systems, Boston
College of Management, University
of Massachusetts, Boston,
Massachusetts

IGOR I. LINKOV

Environmental Laboratory,
U.S. Army Engineer Research and
Development Center, Vicksburg,
Mississippi

INTRODUCTION

Risk of harm and loss are present in many aspects of modern life, and as society and technology become more complex, there are ever more risks to manage. This article is aimed at readers who already know about managing risk, and describes issues that arise in managing a *portfolio* of risks. The terms risk and portfolio have many definitions, so working definitions are useful. The Society for Risk Analysis gives the following definition of risk: "The potential for realization of unwanted, adverse consequences to human life, health, property, or the environment; estimation of risk is usually based on the expected value of the conditional probability of the event occurring times the consequence of the event given that it has occurred." In this article, a single risk is taken to mean a possible event associated with a resulting consequence. A *portfolio* refers to a set of objects that are considered together. For purposes of discussion, there are several distinctions that may be useful. A portfolio of risks is a portfolio of possible events that may result in a portfolio of consequences of concern. There may be a portfolio of at-risk assets affected by events, thereby resulting in consequences. Importantly, the risk manager chooses a portfolio of decisions aimed at reducing risks

(and which may contribute to and complicate the portfolio of consequences).

This positioning of the article has several implications. First, while the article references problems in risk analysis and risk management that are generic and not specific to portfolio problems. It does not explore them in detail; for example, processes for involving stakeholders, and methods for assessing probabilities. Second, because it is about managing rather than just analyzing, decisions are relevant. Following Pate-Cornell and Dillon-Merrill [1], who describe the way that decision analysis overlaps with risk analysis and helps connect it to action, this article incorporates a decision-analytic perspective, rather than a purely risk-analytic perspective. Third, defining risk in terms of negative consequences, the article focuses on portfolio problems aimed at reducing risk rather than the many portfolio problems that merely involve risk (although the latter provide lessons applicable to the former). Finally, the range of portfolio and risk problems and methods found in practice is broad, and their intersection is not always well documented. Therefore, this article surveys broadly what the authors view as the relevant terrain and the general guidelines for navigating it.

Commonly, risk management is approached as efforts directed at single risks or ad hoc groupings of efforts directed at risks. Neither of these treatments explicitly considers possible interactions among projects or risks when risks stem from multiple sources. Whether a set of objects is considered explicitly as a portfolio is a choice made by whoever is considering them. Considering objects as a portfolio in this sense typically involves at least some additional effort in understanding relations involving multiple objects (risk, decision, and so on). Risks should be considered as a portfolio when risk from multiple sources must be managed in a coordinated manner. The need for coordination depends on the decisions under consideration: if the optimal decision that

Table 1. Examples of Portfolios of Risks and Risky Portfolios

Problem\ Application	Elements of Portfolio Directly Affected by Decisions	Nature of Risk	Interactions between Elements
NASA unmanned flight	Design elements	Component failure leading to system failure	Causation, shared vulnerabilities
Oil and gas	Wells, seismic tests	Failure to find oil or gas leads to lost investment	Correlation in well performance, dynamic information
Infrastructure	Structures (repair and improvement)	Failures of any part of system lead to a variety of consequences	Shared vulnerabilities, causation
Business products	Individual products, project development phases	Loss of investment, disappointment	Production and market synergy or dissynergy
Real estate	Structures, land	Small- and large-scale economic, social, and environmental conditions affect value	Virtuous and vicious cycles, supply/demand

influences one risk could change depending on decisions or events that relate to another risk, then ideally those risks would be managed as portfolio.

Portfolios of risks arise in many domains, examples of which are summarized in Table 1. In real estate, a portfolio of properties faces the risk that one or more properties may lose value [2]. These losses may correlate with common initiating events such as local economic conditions or interest rates [3]. For business product portfolios, there is a risk that a new product will fail and the money invested in its development will be lost [4]. Of course, this is a risk that can be minimized by simply not launching any new products, and in this sense, it may be better viewed as a risky portfolio of opportunities that involve risk rather than as a portfolio of risks. In the oil and gas industry, there are portfolios of opportunities involving risk in drilling expensive, unsuccessful wells at one site that correlate with those of other sites [5]. Additionally, investments in information technology can be considered as a portfolio, more so than other corporate infrastructure investments since information technology shares both financial and nonfinancial resources in a coordinated manner [6]. Critical infrastructure within a region or

of a certain type, such as dams, chemical plants, and power plants, may be viewed as portfolios of assets with an associated portfolio of risks (due to failure) and associated investments that can also be viewed as portfolios. Purely financial assets (stocks and bonds) are commonly considered as portfolios with associated risks. Large-scale engineering projects, such as those undertaken by NASA [7] may be viewed as portfolios of risks (involving multiple physical components and multiple program activities). Environmental risks may be regarded as a portfolio if shared hazards, pathways, or locations exist. In health care, a portfolio of diseases may result in a portfolio of negative health consequences (pain, death, economic costs), and can be mediated by a portfolio of treatments.

The remainder of this article will focus on approaches to portfolio risk management and best practices associated with portfolio risk management. The following sections begin by reviewing key portfolio methods and how they can contribute to a quality decision process. Continuing along the same lines, a comparison of the strengths and weaknesses of the different approaches together with a method of combining approaches is developed next. The final section illustrates risk portfolio management concepts using the

example of managing a portfolio of infrastructure risks.

APPROACHES TO PORTFOLIOS

It is useful to first consider four primary approaches to managing risk portfolios: project portfolio management (PPM), modern portfolio theory (MPT), risk analysis and risk management, and systems modeling. These approaches have different strengths, and the manager who is aware of the range of possibilities can and does select among features of each of them (or other methods) as needed to apply an integrated approach to risk portfolio management. For this discussion, it helps to characterize the approaches in terms of the generic steps of a quality decision process. According to Howard [8], a decision process has six elements on which its quality depends: framing, alternatives, information, values, logic, and implementation. These will be highlighted as they arise in the discussion of each approach.

Framing seeks to define what is to be decided and why. This often means delineating the current decision from previously established policies and future downstream decisions. Within risk portfolio management, framing comprises defining which risks are to be included within the portfolio and clarifying any portfolio constraints. The generation of alternatives ultimately creates the decision problem. A good set of alternatives is creative but feasible; in the context of risk portfolio management, alternatives are the many means of reducing the risks (either the likelihood or the magnitude of negative consequences) within the frame. Information, ideally high quality information which is relevant, correct, and complete, is necessary to make predictions. Risk portfolio management information typically pertains to individual risks. The values stakeholders and decision makers ascribe to the range of outcomes must then be determined. A logical synthesis of all of the above results in the identification and selection of the preferred alternatives. Finally, implementation of the selected alternative contributes to the realization of a desirable outcome.

The rest of this section sketches the four portfolio approaches in some more detail, together with some of their major variations. The goal here is to orient the reader but not to cover in depth all the issues that arise in practice, so many details have been left out. The last two sections consider these approaches as part of an integrated approach to risk portfolio management.

Project Portfolio Management

PPM refers to an approach that selects which projects from a portfolio of candidates should be funded in the face of a budget constraint, often utilizing decision analysis methods. In the simplest case, there are no significant interactions between projects other than competition for funds. Projects are ranked in terms of a productivity index measuring their payoff per unit of resource used. When total cost is plotted against total value for each portfolio, then the cumulative cost and value for the projects in order of productivity form an efficient frontier for the set of possible portfolios. This approach has been widely applied in industry for R&D projects, especially in the pharmaceutical industry [9], and within the government for army-base closures [10]. In its simplest form, PPM finds the optimal portfolio solving $\text{Max}_{F_i} V = \sum V_i F_i$ s.t. $\sum F_i \leq B$, where F_i denotes the funds allocated to project i , V_i denotes the value (or expected value) of project i and B denotes the available budget. At its most complex, PPM relies on sophisticated optimization tools to choose a strategy of coordinated and in some cases, dynamic decisions subject to maximization of some value measure on the distribution over possible outcomes subject to constraints describing resources, allowable interactions between actions, allowable levels on certain outcomes, laws of nature, and many other properties of a situation. Decision-support software may be useful in finding solutions.

PPM may be appropriate for managing risks under the following conditions: the only (or main) interaction between risks is that they compete for resources; the objective is to minimize the relevant measure of risk, expected loss (compared to some baseline); there is a budget constraint for investing

in risk reduction; for each risk there is a corresponding project to reduce that risk. Here, expected loss is the sum over all risks of the product of the probability that the negative outcome associated with the risk occurs and the magnitude of the outcome. The risks thus map to the project funding decisions. The alternatives in the portfolio decision problem are the levels at which to fund (if at all) efforts toward reducing each risk. PPM prioritizes the projects that reduce risk. This approach is used for business investments with some uncertainty, notably R&D. Typically, one project affects one risk. The valuation metric is typically expected dollar value or expected net present value; with portfolios of risks it may be natural to think instead of expected loss.

Consequences can take many forms, and it may be unnatural to consider only financial loss. Multicriteria methods [11,12] in PPM allow for project values (and expected values) across a range of objectives to be converted to a single score and expected value can still be used.

The portfolio is defined to include as its elements the possible activities under the funding authority, so different levels can view the portfolio differently. For example, in the pharmaceutical industry, there may be a division focusing on heart disease and the manager of this division views the portfolio as the products and compounds under development. At a higher level of the organization, an executive may think about a portfolio of business lines; for example, cancer therapies, psychiatric drugs, medical devices, generic drugs, and medical supplies.

Framing for the problem involves determining the set elements that are candidates to receive funding. The primary information needed is the cost, probability of failure, and the loss given failure (alternatively, the probability of success and the gain given success) associated with each project. If needed, interactions between projects are considered explicitly; for example, two projects may be able to share the cost of a common item, or two projects may be mutually exclusive. The decision rule that logically synthesizes all of these inputs is to select the portfolio that maximizes the measure of value, subject to the

budget constraint. If there are no interactions among risks other than their competition for funds, a good decision rule may be as simple as ranking projects in terms of a productivity index; for example, value/cost, and funding them in order until the budget is exhausted.

Portfolio Risk, Variance, and Nonlinear Loss Functions. One step in complexity beyond basic PPM is when outcomes for the elements of the portfolio interact in terms of a loss function, meaning the total loss depends on some measure of the sum of the realized consequences of the risks. If the combined loss from two risks combined is equal to the sum of the loss from each risk alone, then the *expected* loss of the two risks is equal to the sum of the expected loss for each risk alone. With a more general nonlinear utility or loss function, this is not the case. One risk might be acceptable if there are no others, but if there is already a good possibility of ending up with a significant loss, the prospect of further loss may render an additional risk unacceptable. In such cases, it is particularly important to understand the loss function/risk tolerance as there may be heuristics to manage risks based on it [13,14]. For example, such a loss function can make risk diversification strategies attractive when they reduce the likelihood of extreme downsides. This feature complicates the portfolio management approach in two ways: a loss (or utility) function must be assessed, and computation of the expected loss minimizing portfolio has a nonlinear element that may require more advanced algorithms.

One key difference between portfolios of risk-reduction efforts and portfolios of, say, R&D projects is the way we would think about variance. For the former, the expected payoff of an investment to remove a risk is equal to the expected loss avoided. This is the probability of the negative consequence occurring in the absence of the investment multiplied by the impact of that negative consequence if it occurs. In a sense, the project is “successful” and adds value in the counterfactual situation where the negative would have occurred were it not for the investment, and otherwise it adds no value. In contrast,

an R&D is successful and adds value if the technology works. If the project is not undertaken, there is no chance of it working. The variance for the value of a project in either case is the $x^2 p(1-p)$ where p is the probability of the investment being a “success” and x is the “gain” in the case of a success.

With portfolio variance, there is a key difference. In the simple case of uncorrelated elements, the variance of the value of the portfolio is the sum of the variance of all its uncertain elements. R&D managers are concerned with the variance of the payoff of R&D but with a risk portfolio, stakeholders are concerned with the variance of the results of the negative consequences. If no projects are funded, each risk contributes its variance to the total; the total variance of the risk portfolio is thus its original variance less the variance of the projects. If a risk-adjusted bang-for-the-buck ranking is used, it may be necessary to add an adjustment factor for the projects in a risk portfolio, instead of subtracting it as in an R&D portfolio. In some cases, this implies that a high variance portfolio of risk mitigation activities is desirable. This seeming paradox is due to a definition of value, whereby the project value is negatively correlated with the value of the underlying portfolio of assets.

Interactions within a Project Portfolio.

Beyond interacting due to a nonlinear loss function, risks can interact in more specific ways: risks may interact dynamically (whereby one risk leads to another).¹ Risks may interact with respect to multiple criteria so that multiattribute utility functions [16] might be suitable, especially if negative outcomes in one dimension make negative outcomes in another dimension more costly. There may be cost synergies or dissynergies or more general nonlinear relationships between efforts and risk reduction. The logical synthesis of the project level information must produce a picture of the portfolio level risk. Advanced techniques in PPM can

accommodate some such considerations by expanding value calculations for specific projects but at some point, these interactions become too numerous and complex to deal with on a project-to-project basis, and other methods are suitable for characterizing the situation and making decisions.

Interactions also present a major challenge for selecting actions within portfolio management (whether PPM or other approaches) because there can be a vast range of possible strategies involving a large number of decision variables. Combinatorial optimization (often using integer, binary, and logic programming methods) and constrained optimization methods have been successful in this regard. Contingent portfolio programming [17] has been applied to risky projects with multistage decisions and with multiple objectives incorporating risk aversion. Most computational optimization methods require well-specified objective functions; objectives can be identified and agreed upon. Robust portfolio modeling [18] can be useful where multiple objectives apply but where, as is common in portfolio problems, it is hard for the range of stakeholders to specify an exact objective function. There is quite a substantial literature on multicriteria optimization methods [19].

The example at the end of this article illustrates different types of interactions between individual risks and projects. If there is rich enough interaction between them, enumerative approaches may be impractical. When there are well-defined correlations between many risks and where it is desirable to reduce the variance of the sum of the risk outcomes, financial theory described in the next section can be useful in characterizing the portfolio.

Modern Portfolio Theory

MPT was first formalized in several well-known articles [20–23] and is now standard in finance textbooks, for example in Ref. 24. The crown jewels of MPT are elegant models resulting in simple formulas for investing under the assumption of efficient markets.

MPT is the nearly universal basis for management involving portfolios of risky financial assets, including stocks, bonds, and cash.

¹It may be possible in such cases to characterize and calculate outcome distributions based on these interactions using precedence matrices [15].

Individual assets such as stocks have uncertain returns with both upside potential and downside risk. An investment portfolio consists of a number of different assets purchased. It is common to denote as w_i the weight of asset i in the portfolio. The return on an asset i , r_i , is the percentage by which its value changes over a period of time (e.g., annual return). Arithmetic shows that the return on a portfolio is the weighted sum of the returns of the assets in it. The *expected* return on a portfolio is likewise the weighted sum of the expected returns of its assets.

Due to the diminishing marginal utility of wealth (as well due to psychology and the planning benefits of predictability), investors are generally risk averse. In market portfolios, this means that they would be willing to give up some expected return to reduce the spread of outcomes (typically measured as variance). Investors thus typically aim to achieve the maximum return for a given level of risk, or alternatively, to have the minimum risk associated with their expected level of return. For such investors, diversification is a free lunch. Variance can be reduced without affecting average return. If asset returns are uncorrelated, the portfolio variance is the sum of the product of each individual asset's variance and the square of the asset's weight in the portfolio. Then portfolio variance can be reduced arbitrarily close to zero by having many assets each with a weight in the portfolio of close to zero.

The fact that returns on assets are correlated, makes this theory more interesting. Many investment risks are essentially uncorrelated; for example, the fortunes of one company in one industry may depend on the occurrence of an unpredictable technical breakthrough, while the fortunes of another may depend on the status of international relations. The insurance aspect of pooling poorly correlated risks is the motivation for having mutual funds. Individual mutual fund investors need not be concerned with the risk associated with specific companies. Stock prices, however, are often correlated. In a strong economy, for example, most stocks tend to go up while in a weak economy they do not. The correlations are even stronger between stocks of similar

companies in similar industries. Variances and covariances of individual asset returns are estimated using a mix of judgment and sophisticated econometric methods but unfortunately, MPT models require correlations and uncertainty looking forward to the future, but automated methods can only mine historic data. From these estimates, the variance of portfolio return can be calculated based on the weights in different assets and the matrix of covariances between the assets.

As a mathematical problem, it is theoretically possible to find the variance minimizing mix of investments for any level of expected return, that is, the efficient risk-return frontier. With a substantial number of potential investments, however, finding this frontier could be a combinatorially large or even impractical problem. Fortunately, mathematical simplification is possible. The market as a whole is the sum of all assets weighted by the proportion that they represent of the market's total value. The Nobel prize-earning references above show how useful this fact is. For the purpose of constructing an efficient portfolio, it is often sufficient to summarize all the asset correlations in terms of the correlation between each asset's return and the total market return. The asset's correlation with the market as a whole is its β value (because the theory derives from statistical regression model coefficients), and the portfolio itself has a β value equal to the weighted sum of the β values of its elements. Thus correlation coefficients are a widely used and theoretically supported model simplification of multivariate probability distributions that would be difficult to assess directly.

This gives rise to the distinction between systematic and nonsystematic (sometimes called *idiosyncratic*) risk. If the market is efficient (as is often assumed in theory), there is no way to hold stocks that give the same return as the market without also at least having the risk of the overall market. That risk, which cannot be diversified away, is called *systematic risk*. In contrast, a portfolio on the efficient frontier would have small proportions of many different stocks so that the idiosyncratic risk of each asset is diversified away.

But the investor can do better than holding a mix of risky assets. A diversified index of the market as a whole is on the efficient frontier. But the investor can (in theory) also hold a “risk-free” investment, one with guaranteed return. US Treasury bills have sometimes been assumed to be risk-free in this sense. For any level of risk the investor might wish to accept, some percentage β of funds can be invested in the market portfolio and the remainder invested in the risk-free asset leaves the portfolio to leave that risk. The variance of this investment mix is β times the variance of the market portfolio. In an efficient market, the expected return on this portfolio is $(1 - \beta)r_f + \beta r_m$, where r_f is the rate of return on the risk-free rate and r_m is the return on the market. The set of possible risk-return points obtainable by choosing different levels of the risk-free investment is called the *capital allocation line*, and for each level of risk, the corresponding point on this line has return greater than or equal to the best portfolio that does not contain the risk-free asset. The value of β represents the leverage of the portfolio and could range from less than 0 (for a portfolio which actually moves opposite to the market; i.e., one in which stock is sold short in order to purchase the risk-free investment) to more than one (a leveraged portfolio). The risk appetite of the total pool of investors determines just how high r_m must be in order for there to be an equal number of buyers and sellers of risk, and MPT says that current market prices will adjust so that expected r_m is at the right level.

Finding optimal financial portfolios is by now a well- and continually studied problem. Simple algorithms often suffice, although professional investors such as hedge funds certainly have more complex and sophisticated models. These models may have constraints or value measures involving risk statistics other than standard deviation; for example value-at-risk, which gives the amount of loss associated with a given percentile. Financial engineering tools extend MPT concepts as well as options theory [25] to manage risk and return to sculpt a relatively

desirable probability distribution by diversifying across assets, diluting risks, and creating or obtaining various insurance mechanisms defined in terms of future asset prices.

Arbitrage pricing theory, a fairly recent development, allows the portfolio to be further characterized in terms of multiple “ β ” factors, each corresponding to a different systematic and nondiversifiable risk (e.g., inflation risk, real estate risk, and classic stock/bond market risk), with associated market equilibrium prices for each type of risk. This matches up nicely with applications of multicriteria decision analysis (MCDA) in finance [26,27], where alternatives are evaluated in terms of a multicriteria model and the criteria correspond to sensitivity to different kinds of risks.

MPT-based approaches can be adapted to more general decisions involving risks. Smith and Nau [28] and others have shown how in decision analysis, some risks may be treated as market-based and assigned a market-derived price while others risks must remain internal. Decision theorists [29] have identified utility functions with intuitively desirable properties that can also be characterized as simple functions of risk and return. In particular, Bell argues that $u(w) = w - be^{-cw}$ is an especially useful form when risk takes the form of a normally distributed increment to initial wealth. Such a function based on mean and variance provides a bridge from the methods and statistics utilized by financial theory (as described above) to the concerns of loss-sensitive decision making. Because utility is usually decreasing in risk, the idea of efficient risk-return frontiers may also apply to nonfinancial portfolios of actions, and some aspects of the MPT conceptual framework have been used to describe such situations [30].

To put this in terms of the decision quality framework, in MPT the frame that defines what is to be included in the portfolio is primarily the notion of correlation, that is, potential investments are considered as parts of a portfolio to the extent that the investor is concerned about their correlation or lack thereof. Risks and decisions involve available investment instruments. Where the ultimate values may be correlated,

where the level of correlation matters, and where there is some way to estimate the correlations, it is useful to consider these possible investments as the elements of the portfolio. The alternatives are best described in terms of how much of each risky entity to hold and in what overall mix. MPT typically values the portfolio solely in terms of risk (e.g., variance) and return (e.g., expected value, or expected rate of growth), and optimization formulas and algorithms logically synthesize the information with an aim to maximizing return for a given level of risk.

Risk Analysis and Risk Management

Risk analysis involves “detailed examination including risk assessment, risk evaluation, and risk management alternatives, performed to understand the nature of unwanted, negative consequences to human life, health, property, or the environment; an analytical process to provide information regarding undesirable events; the process of quantification of the probabilities and expected consequences for identified risks” [31], and *risk assessment* is “the process of establishing information regarding levels of a risk and/or levels of risk for an individual, group, society, or the environment.” Risk assessment uses a set of tools to incorporate technical information and expert judgment about the structure, likelihood, and magnitude of consequences from sets of related events. Ideally, it yields precise understanding of what drives risks and generates defensible probability distributions over outcomes from those risks under various scenarios. *Risk evaluation* (following the definition above) should aggregate information from the assessments associated with the portfolio of individual risks to develop a risk profile (e.g., a probability distribution) for the portfolio. It then relates the risk profile to the situation at hand (e.g., levels of background risk, importance of different potential losses associated with negative outcomes in various dimension). *Risk management* is the process of judging the desirability of potential consequences and selecting action alternatives, often in response to the findings of risk assessment and risk evaluation. It uses a loose set

of tools, practices, and processes, both qualitative and quantitative, that may be used with or without risk assessment. Risk management helps to ensure that there is a reasonable plan to reduce risks and that this plan is executed. The reader should note that the labels of these different activities are not always interpreted precisely and that there is overlap with other activity labels such as risk characterization and risk integration.

Risk assessment and risk management have been separated in many regulatory and planning applications. Risk assessment is considered to be a scientific evaluation process performed by scientists while risk management is associated with policy and regulatory process and is performed by regulators or planners. For example, the Environmental Protection Agency [32] specifically states that “risk assessment provides ‘INFORMATION’ on potential health or ecological risks, and risk management is the ‘ACTION’ taken based on consideration of that and other information” including, among others, economic, social, and political factors as well as stakeholder values.

In practice, this artificial separation can result in a disconnect between risk assessment and risk management. For example, the framing of problems in risk analysis is usually limited to identifying population and resources at risk and not risk management alternatives. It is generally done using conceptual models (influence diagrams visualizing relationship among stressors and resources at risk). The goal of risk assessment is not to prioritize risk management alternatives, but rather to assess risks resulting from multiple stressors and inform risk managers about these risks. Alternatives are not explicitly considered as a part of risk assessment; instead risk models are used to assess residual risks associated with implementing risk management alternatives that are developed outside of the risk analysis process.

The focus for risk assessment is collecting information and developing models to assess risk associated with potential stressors. Risk assessments usually include steps of hazard identification, exposure assessment, effects assessment, and risk characterization. Hazards are usually identified utilizing

historical data, reconnaissance site survey, preliminary models, and other available information based on professional judgment of risk assessors involved in the process. In risk assessment conducted for regulatory compliance (e.g., Superfund sites in the United States), this step also involves detailed consultations with stakeholders and regulators. Exposure assessment includes evaluation of the degree to which different population groups, environmental receptors, and resources are affected by stressors while effects assessment quantifies the level of harm resulting from exposure to stressors. Risk characterization combines results of exposure assessments and effects assessments, and compares resulting risk to regulatory benchmarks or other criteria deemed to be acceptable. Estimates and judgments related to uncertainty are part of the process, but these are rarely integrated explicitly and quantitatively during risk characterization.

Risk assessment informs risk management. Treatment of risk assessment process as “science” and risk management process as “policy” has resulted in very limited use of quantitative tools at risk management stage of risk analysis. Risk management alternatives are usually generated based on professional judgment and residual risks associated with their implementation is assessed using developed risk models. Values and other factors are incorporated in an ad hoc manner and decision analysis tools for value assessment and for logical synthesis are rarely used. Alternative selection is extensively vetted through stakeholder consultations. Implementation of selected alternatives usually requires monitoring to confirm expected reduction of risks. In many cases, though, expected risk reduction is not achievable and risk assessment needs to be repeated to meet regulatory needs [33].

In general, risk assessment treats individual risk independently and passes this information for consideration by risk managers. Sometimes simple integrated metrics are used (e.g., hazard index is developed as sum of hazard quotients for individual chemicals). In contrast to the simple portfolio problem, where the only decision variable

available to management is allocating funds to projects, risk management has a richer set of alternatives about how to manage each of the risks such as *mitigation*, *amelioration*, *diversification*, *risk sharing* (in which case, we may also be concerned with incentive effects, as in Ref. 15 as well as *establishing reserves* [34], *insuring*, or *obtaining information*, but the framework for quantitative assessment is not developed. In many cases, a single risk driver can be identified since risks, more so than project portfolios, tend to involve low probability, large magnitude events; in some cases, we can essentially ignore their interaction because of this while in other cases, their interaction is plausible enough for us to have to worry about disastrous consequences of multiple large magnitude events that overload our capacity to deal with them. Assessment of low probability events and of correlations between those events thus becomes important. If we plot resources used against the measure of risk, we obtain a risk-reduction trajectory (similar to the Pareto frontier for project portfolios).

Systems Modeling and Design

Interactions among a set of risks may be too complex and dynamic to practically capture with the approaches above. In that case, the portfolio of risks can be viewed as a system [1]. *Risk-based design* approaches are a more powerful tool for characterizing the problem structure. Dillon and Pate-Cornell describe a system for programmatic risk analysis where risks can be characterized by interactions among subsystems and components. Originally developed for NASA, their methodology divides optimization within a single project into four steps. The first identifies all viable alternatives, developing a minimum cost and residual budget for each. Next, a project-specific residual budget curve is developed which allocates the residual budget between system reinforcement and project reserves using probabilistic risk assessment. Problem scenarios are then identified together with their likelihood of developing a managerial risk profile. Finally, the components are optimized together, with risk thresholds serving as constraints, to determine if an optimal project exists.

A variation on this approach accounts for project interactions within a project portfolio. Prospects for projects dependent on previous outcomes are calculated, conditional on those previous outcomes, to form an expected project cost. Initial projects incur a cost penalty, the difference between the nonconditional and conditional cost of the later projects, for disrupting later projects. Optimization occurs to minimize the second project expected cost and to minimize the cost penalty. The framework was originally developed for sequential projects directly affected by a project failure, such as an initial space orbiter mission followed by a lander mission [35]. However, projects occurring in parallel may also be considered. Parallel risks are represented in the projects' respective fault trees; a secondary tree can be developed to show the interaction among all projects. For a range of circumstances, assessments are performed to estimate the probability of the system being in a state of failure. The tree methodology treats each node as a random variable, with arrows showing probabilistic dependence. The final node shows the probability of project failure. This approach models the interaction among projects, information acquisition possibilities, and the whole range of scenarios affecting systems, with probabilistic risk assessment.

Risk-based design incorporates risk analysis methods within the much broader discipline of *systems engineering* [36]. Systems engineering defines a system as a set of interdependent components (this set itself could be considered as a portfolio), aligned for a common objective or objectives. It is used to manage all aspects of complex projects and, therefore, could encompass an even broader notion of portfolio and of risk.

Events and consequences are quite often embedded within a system; for example, a nuclear waste remediation system [37]. Thus, it is common to bring some systems modeling tools into any of the aforementioned approaches. In even more complex settings, *dynamic systems modeling* is useful; for example understanding feedback loops may be necessary in order to anticipate emergent events. For example, it may be

that decision makers failed to anticipate the financial crisis that emerged in 2008 because they failed to take such dynamics into consideration. The global climate is an even larger system for which large dynamic models are used to gain insight about risks. The decision variables in such large systems may be a portfolio of interventions that modify feedback loops, but these are not associated in any simple manner with specific initiating events, assets or consequences. Such modeling approaches go well beyond the problem of managing a portfolio of risks and are well-detailed elsewhere [38]. Overall, managers can choose from a wide range of approaches to craft an appropriate approach to managing risk portfolios with attention to the specific challenges of the situation they face.

STRENGTHS AND WEAKNESSES OF, CONNECTIONS BETWEEN, AND CHOOSING AMONG METHODS

This section aids in selecting the right risk portfolio approach by comparing the relative strengths of the different methods and considering the type of situation in which those strengths are needed. It is helpful to start the discussion with Table 2, which summarizes (the authors' view on) the differences between the approaches described above as well as the characteristics of an integrated approach to risk portfolio management.

The different approaches have contrasting strengths and weaknesses. PPM identifies efficient allocation of resources, such as with dynamic models involving multiple resources. It tends to foster an organizational culture of efficiency too, and may explicitly encourage risk taking more than other approaches. On the other hand, although it is possible to account for project interactions, it is not simple to do so. The approach works simply when straight expected value is the decision criterion, but this may not be the case with risk portfolios.

MPT methods are quantitatively tight so that they can be used on a large scale where appropriate markets and mechanisms exist to allow for precise application of controls.

Table 2. Comparison of Approaches

Process	Project Portfolio Management	Modern Portfolio Theory	Simple Risk Analysis and Risk Management	Systems Modeling	Integrated Approaches to Risk Portfolio Management
11 Framing Factors determining what is treated as a portfolio Constraining factors	Funding authority and cycle Budget and other resources	Fungible/tradable market assets Ratios, funds available	Potential events and scenarios Acceptable risk levels, budget	Interactions between projects Acceptable risk levels, budget	Flexible Resources and performance requirements
Alternatives	Funding for each project	Percentage mix of assets	Tactics at project level	Allocating funds between projects/tasks	Look at many dimensions
Information and interactions	Required resources, probability, and value of success, <i>ad hoc</i> interactions	Return correlations and distributions, prices	Failure trees	Models of system and components	Decision-oriented interactions of various types
Values	Expected value or utility/MCDA	Risk and expected return	Acceptability based on probability of failure and magnitude of loss	Acceptability based on probability of failure and magnitude of loss	Multicriteria/utility function selected for good risk properties
Logical synthesis	Optimization to maximize portfolio-wide productivity index	Compute theoretically optimal balanced portfolios	Simulation, scenario planning	Assign optimal distribution of funding to maximize objective	Compute as needed
Implementation	Project budgets and milestones	Trades and purchases	Processes, monitoring, and responsibility	Budget development	One or more of other approaches

However, as the worldwide financial crisis of 2009 demonstrates, it tends to miss system dynamics and if conditions have changed, it may use the wrong assumptions (e.g., volatility is higher than assumed) and its decisions may not be robust in the face of such problems. Furthermore, it has not been adapted much to use beyond pure profit measures. Risk analysis and management deal especially well with uncertainty about events, as well as acknowledging the existence of ignorance. It adapts well to a system view and its processes (in theory) accommodate multiple stakeholders (as there commonly are in portfolio settings). On the other hand, risk management tools such as red–yellow–green classifications provide only a coarse prioritization. Simplistic risk management practices can be hard to translate to a specific implementation; they can lead to very (overly) conservative practices (e.g., where any probability at all of an event moves it out of the green category into yellow), and may lack sufficient quantification to identify the most optimal plan. In order to trade-off risks against benefits, MCDA approaches can again be useful.

In choosing the right tools, practitioners should match the strengths and weaknesses of different aspects of each approach to the needs of the situation. For example, certain established approaches combine two or more methods. In terms of risk measures, *value-at-risk* [39] combines from risk analysis and finance to summarize and communicate risk, making it suitable for managing certain day-to-day operations although as a management tool, it does not capture well problems involving system failure and very low probability events. *Risk-based design* combines risk analysis and systems analysis/design, and it is useful for capturing complex interactions between risks and providing a productive approach to managing risks. Programmatic risk analysis [40] combines elements of decision analysis and risk analysis in a customized approach to managing related risks. Decision analysis and risk management techniques can be combined to consider alternatives involving mitigation, amelioration, allocation, and acceptance of risks. Finally,

MCDA has been combined with risk analysis [41] to translate the range of consequences into a basis for decisions.

In an integrated approach, portfolio problems should be classified according to which tools are appropriate and then, those tools should be used. The example in the following section illustrates this idea by describing a portfolio of risks (dams and levees) and discussing how they might be managed with an integrated approach and the other approaches explained above.

RISK PORTFOLIO MANAGEMENT: APPLICATION EXAMPLE

This section uses an example from the United States Army Corps of Engineers (USACE) in order to illustrate in a specific context the general steps for managing a risk portfolio and to discuss how, within an integrated approach, different methods can be fitted to a particular problem. We note that there is no single representative portfolio in terms of the specific methods that should be used to analyze and manage it, but the idea of actively selecting methods for managing each of the various dimensions of the problem is common across portfolios. USACE and related government agencies manage thousands of dams and levees. Within the dam and levee portfolio, these agencies are responsible for allocating limited budgets to reduce the risk of failure. However, if risks between projects correlate, or if losses are not constant functions, traditional methods of allocating funding may not maximize risk reductions. Portfolio methods must then be utilized to ensure the greatest possible risk reduction in these complex circumstances. Differing elements of the previously outlined risk approaches apply in the context of dams and levees.

The methods used to support decisions about the portfolio should match the degree and type of complexities in the risks, decisions, and consequences of the portfolio and their interactions; for example, correlations among risks and nonlinear loss functions. Risk correlations occur when the likelihood and the impact of failures for different structures are not independent. Put similarly, the

loss due to two related failures may cause much greater damage than the sum of the damage caused if both failed separately. For example, a 10-ft rise in water level may be more than twice as costly as a 5-ft rise or the failure of an upstream dam may precipitate the failure of a downstream dam. Correlations thus account for the tendency of issues to change with the surrounding events. Failures may be correlated with internal or external stressors. Structures with a common design or other common characteristics could fail disproportionately in response to a specific stressor. In particular, climate change could increase the likelihood that a stressor will affect a large number of structures (sea level rise may affect structures in low-lying areas and changing precipitation patterns could have similar effects on structures in similar locations). Therefore, considerations such as diversification, system design, and dynamics could be important and could incorporate qualitative, quantitative or integrated measures of performance [42,43]. Correlations add complexity to the management of projects by instilling relationships with surrounding projects and events. Nonconstant loss functions also add complexity, as it is less intuitive to convert losses to an equivalent single dimensional value, such as monetary cost.

Managing dam and levee risks can be accomplished through both infrastructure improvements and policies/practices. Infrastructure improvements are projects that require substantial amounts of funding to put into place new hardware. These address risks to many types of structures and facilities, various groups of people, and risks from chains of events. Infrastructure improvements also ameliorate consequences ranging from loss of life to economic loss and social displacement to environmental damage. Within dam and levee management, there is a substantial but limited budget for infrastructure improvements.

Limited budgets for infrastructure improvements stimulate interest in other ways of achieving risk reduction with less financial expenditure, typically through policies and practices. Policies and practices involve zoning, organizations, and planning

for both, standard operations and contingencies. This complicated problem is surely better viewed as a portfolio than as a set of independent risks. What is done with regard to one risk can influence other risks. For example, contingency planning to open flood gates during high water events could cause downstream flooding, having the effect of exchanging dam failure risk (probability) for human risk (level of harm). A combination of policy methods and infrastructure improvements will typically be involved in most risk-reduction strategies.

Several courses of action are available to the USACE given the aforementioned considerations. The worst practice might be to fix each dam after it fails. Alternatively, a plan might be announced to perform all proposed repairs on all structures, but that plan would be unrealistic because it would likely exceed any available budget and some work would remain unfinished. Historically, engineers for each dam would write a justification for funding agencies about why they need money, and decisions about allocations would then be made through a rather unstructured political process. But if it is taken as given that the societal objective ought to be to minimize damage, that approach also has great shortcomings. USACE has recently taken steps toward more systematic decisions about these risks [44]. Dams are graded on engineering criteria (e.g., cracks, voids, spilling, concrete deterioration) and impact criteria (i.e., magnitude of consequences: potential hectares flooded, people killed, etc.) as in Ref. 45. The product of these two scores produces a prioritized list of dams (high, medium, low risk). Investment is made in the highest risk structures first. This approach imposes discipline on the process, but it may not be sufficient to allocate the available budget so as to minimize society's expected loss. It does not explicitly consider cost, so optimization algorithms cannot be applied.

PPM methods that value risk mitigation efforts can improve the effectiveness of resource allocation beyond simpler approaches that prioritize risks in terms of probability of failure or severity of failure. Where losses may be correlated, notions from financial portfolio management, MPT can be

applied. With coastal project levees, there is reason to expect losses to be correlated along various measures and returns on investments to be correlated with respect to those measures. With dams, losses should correlate with rainfall and upstream dam failures. Notions from MPT have been applied to these problems [30]. Small changes are more easily accommodated than large shifts, so there is reason for risk aversion in this kind of portfolio. This would be true—to different extents—with respect to shipping capacity, human displacement, ecological damage, etc. With the utilization of a multiattribute utility function (ideally of a form that can be expressed in terms of risk and return), one can describe portfolios with respect to each of the measures.

The purpose of the following case discussion is to illustrate how the generic steps might translate given a realistic situation, as well as to illustrate how a computational model can be incorporated. The computational model used here is a relatively simple spreadsheet. For more complex applications, a more complex spreadsheet, and a spreadsheet with Monte Carlo simulation [46] and optimization tools included in an add-in (e.g., Risk Solver by Frontline Systems, Decision Tools Suite by Palisade, or Crystal Ball Decision Optimizer by Oracle), or dedicated modeling software could be suitable.

The authors' list of generic steps (contributing to various elements of decision quality):

- Define what is in the portfolio and defining constraints (Framing)
- Define consequences of concern (Setting up to obtain information that informs values)
- Identify initiating events and emergent risks (Information)
- Structure and characterize individual risks (Information)
- Map risks to decisions about steps to reduce risk (Alternatives)
- Characterize impact of steps to reduce risks (Information bridging from Alternatives)
- Characterize cost/resources of steps (Information bridging from Alternatives)

- Characterize interactions between risks (Information bridging to Logical Synthesis)
- Calculate portfolio risk profile (Logical Synthesis of Alternatives and Information)
- Calculate portfolio risk metrics/characterizing portfolio risk (Value)
- Choose a plan (Logical Synthesis of Frame, Alternatives, Information, and Value)

In selecting the exact approach, it is useful to apply an underlying concept of decision quality. In each of the six dimensions, a 100% quality is defined as the point at which additional improvement stops being worth the additional effort required to achieve it. What this point is depends on the situation. So, while no dimension should be forgotten, it is possible to go beyond this point by doing too much analysis on some aspect of the decision. Thus, modeling structures and tools should be used that support the level of detail appropriate (generally, this is still determined subjectively) for the problem in each dimension.

The following example, using a simulated portfolio (using random number formulas in a spreadsheet to generate the parameters for each structure's condition, size, cost of repair, potential impact of failure, etc.) demonstrates how the generic steps could apply in case of USACE's management of dam and levee risks.

Define What is in the Portfolio

For this example, the portfolio set will be the dams and levees on the east coast. This selection is based partly on the charge and expertise of the USACE, but also on commonalities, concerns, consequences, and budgets shared across the projects. Also, it is unnecessary to expand the portfolio beyond this. If the goal is to manage and mitigate risk at dams and levees, including sewage processing systems or railways will not contribute to this goal. Though a greater portfolio would allow for greater resource sharing and more subtle analyses of risk, it will still contribute little to the ultimate goal of dam and levee risk reduction.

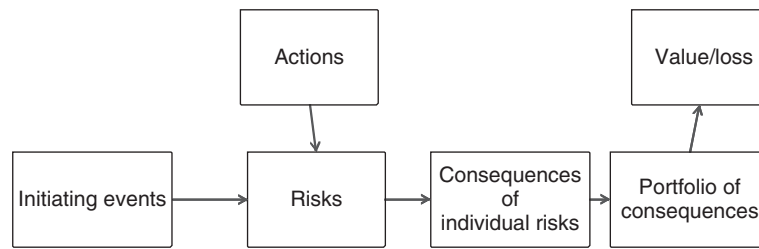


Figure 1. Drivers of portfolio value.

Define Consequences of Concern

If a dam or levee fails in part or completely, the immediate consequence is water release (in reality, there are multiple failure modes some of which have more severe consequences than others). That itself is not a problem, though downstream that water could destroy cities, homes, and habitat. The main consequences can be classified as economic, social, or environmental. Any of the structures failing would have roughly similar consequences although they would differ in magnitude. For other portfolios, it is possible only one description, financial, might be the only concern.

Identify Initiating Events and Emergent Risks

Negative consequences for dams and levees could be caused either by structural failure or operation out of specification. Operation out of specifications might result from stressed conditions such as changes in weather patterns (e.g., rainstorms). Internal causes could result from design flaws, maintenance problems, and age. In this portfolio, similar fault trees are assumed to simplify the analysis.

The logic in this simple model (Fig. 1 is that initiating events may occur (not shown in model), actions may be taken to reduce risks, outcomes for individual risks then result, and the sum of all risk outcomes determines the portfolio loss.

Structuring and Characterizing Individual Risks

In the context of the example, there must be a determination about how to judge the severity of risk. Various approaches are applicable. A simple approach might be to label dams

as low, medium, or high priority and levees as structurally sound, in need of repairs, or deficient. A second approach might be to prioritize structures based on their size or age. Yet another approach would be to classify dams and levees by the size of their downstream cities. For a PPM approach, if intuitive prioritization is not a fine enough process, one can be more precise and specify a probability of failure to capture both the condition of and the stress on a given dam as well as a numerical value for the consequence of failure, based on either a single value or multiple values such as number of displaced people and the business assets lost. In other cases, detailed fault trees might be useful. In either case, if failure probabilities are assigned, they must be defined with respect to a time period. Regardless of the method, with portfolios, it is desirable to be an honest broker by using a consistent method for all elements of the portfolio.

In Table 3, the fourth column gives the probability of failure of each dam, while the magnitude of failure consequences is shown in the sixth column.

Map Risks to Decisions about Steps to Reduce Risk

To manage the portfolio of risks, it is necessary to map the set of risks to a set of decisions. For the example the simplest mapping method was utilized, assuming that there is a single project that will completely eliminate risks for a given dam or levee. A more complicated and realistic mapping might include multiple decisions such as diverting water in addition to repairing the structure. In these more complicated situations, the model

would require entries for the different decisions and their incremental changes to probabilities and magnitudes. This could mean a much longer list of decisions with logical constraints on the decisions and mapping the interactions among decisions.

In Fig. 2, it is assumed that each project corresponds to a single reduction decision. As such, a full listing of decisions is not developed.

Characterize Impact of Steps to Reduce Risks

Steps to reduce risks can decrease the overall probability of failure, the probability of failure given an initiating event, or the consequences of failure. The anticipated decrease can then be compared against the “no action” case. The decrease in expected loss for each risk step is calculated as the difference between the original expected loss and the expected loss after the step is taken. In Table 3, it is assumed that each step reduced the probability of failure to zero. A wider selection of possible risk-reduction steps can also be accommodated. In a spreadsheet, this requires several extra columns and some additional calculations to find the impact of each step and the overall risk for any portfolio plan.

Characterize Cost/Resources of Steps

Effective management of portfolios of risks involves allocating resources effectively. For dams and levees in the example, the repair budget is the only resource under

consideration; estimating the cost associated with a given repair for a given structure is a standard engineering task. In other cases, there may be multiple resources that are limited and cannot easily be exchanged for one another; for example in a public health emergency, there may be a fixed amount of hospital beds, medical personnel and supplies, and no way to get more of one resource by giving up (or even selling) another. The amount required of each of these resources by a risk-reduction step should then be estimated.

Characterize Interactions between Risks

Characterizing risk interactions is a key and complex step. There may be interactions in terms of the likelihood of events (e.g., causation), the consequences of events, or the effectiveness of actions taken that reduce either likelihood or magnitude of consequences of two or more risks.

In the basic numerical example, it is assumed for simplicity that there are no interactions between the risks. For dams and levees, this could be true if each structure were unique and located in an entirely different geographic region. In more complex situations, risks can interact in multiple ways. Examples include the following scenarios:

- If one structure fails, structures downstream are at greater risk of failure. Sometimes this can be accommodated

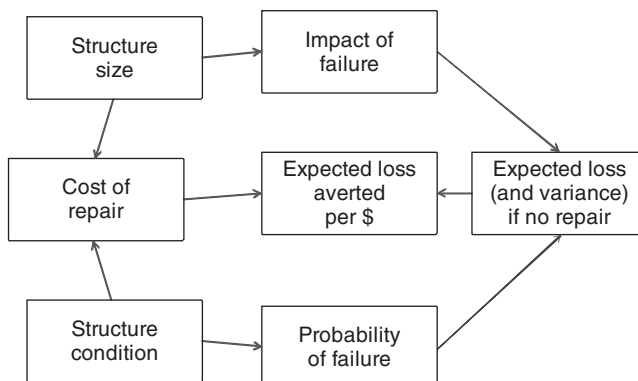


Figure 2. Logic of spreadsheet model.

Table 3. A Simulated Portfolio of Risks in a Spreadsheet Model

Structure Index	Structure Size	Problem Level	Prob of Failure	Cost to Repair	Impact of Failure	Expected Loss	Variance of Loss	Exp. Loss Averted per \$	Rank	Funded?
1	3.27	4	0.13	0.56	3.57	0.47	1.46	0.04	18	0
2	10.35	3	0.49	2.15	19.72	9.65	97.18	3.95	1	1
3	6.98	2	0.15	2.96	13.21	2.01	22.55	0.29	9	0
4	4.32	2	0.31	1.39	4.28	1.31	3.89	0.24	10	0
5	8.30	2	0.41	0.85	8.91	3.69	19.25	1.37	3	1
6	8.78	3	0.32	2.90	6.22	1.98	8.39	0.64	4	1
7	2.26	4	0.11	0.57	2.53	0.28	0.62	0.09	15	0
8	9.96	3	0.40	3.81	3.57	1.41	3.05	0.23	11	0
9	8.16	2	0.05	0.94	28.96	1.48	40.64	0.42	6	1
10	1.71	3	0.25	0.39	1.17	0.29	0.25	0.07	17	0
11	10.04	1	0.15	0.91	6.06	0.91	4.71	0.41	7	1
12	5.31	3	0.46	1.15	1.21	0.56	0.36	0.11	14	0
13	1.87	4	0.00	0.30	1.98	0.01	0.01	0.00	22	0
14	3.64	2	0.04	1.07	1.24	0.05	0.06	0.01	20	0
15	1.97	2	0.10	0.54	1.20	0.13	0.13	0.04	19	0
16	5.65	2	0.27	0.62	2.23	0.61	0.98	0.20	12	0
17	2.32	3	0.03	0.60	1.36	0.04	0.06	0.01	21	0
18	7.52	3	0.36	1.21	14.89	5.43	51.37	1.56	2	1
19	5.29	3	0.00	0.70	1.15	0.00	0.00	0.00	24	0
20	6.66	4	0.01	0.87	23.95	0.20	4.66	0.09	16	0
21	4.39	2	0.00	0.44	3.05	0.00	0.01	0.00	23	0
22	4.54	4	0.38	0.98	9.14	3.47	19.69	0.56	5	1
23	5.07	1	0.28	1.24	7.09	1.97	10.08	0.39	8	0
24	8.87	4	0.00	1.00	61.93	0.00	0.11	0.00	25	0
25	11.08	2	0.17	2.83	3.13	0.52	1.36	0.18	13	0

in a spreadsheet using a precedence matrix.

- If one structure fails, it could mean that other structures with the same design are also more likely to fail. Such risk could be incorporated into a spreadsheet using joint probability or correlation tables.
- If there are dams on two rivers that join into a single river downstream, the damage caused by two dam failures together could be worse than the sum of damage caused by each one failing alone. The two dams could be modeled as a single system with multiple failure modes. The interactions of parallel or serial elements of such systems are simply modeled using logic functions.
- Some of the initiating events that contribute to failure (e.g., particularly heavy rainfall over a wide region) will affect multiple structures. This could be incorporated into the model by assigning probabilities to different initiating events and then entering probability of failure conditional on the set of initiating events. The total probability of failure would end up being the result of a calculation rather than a directly entered value.
- If the consequences of failure are uncertain, consequences could be correlated across risks. For example, a general increase in population and urban development could increase the potential cost of flooding everywhere. Such consequences could be correlated within the model.
- There are synergies in risk-reduction actions, for example if a dam and a nearby levee are fixed, it might be cheaper to do them both together than the sum of their costs if considered independently, because construction equipment will not have to be transported.

Calculating the Portfolio Risk Profile

Once all the individual risks, their interactions, and the steps to reduce the risks are

characterized, it is meaningful to look at the overall range of outcomes. A typical format for the risk profile is a probability distribution over loss. If loss is multidimensional, the risk profile could take the form of a joint probability distribution. These probability distributions can be calculated, though in general they are often simulated. In some cases, statistics (especially mean and variance, as in financial portfolios) for the distribution suffice. For example, if the family of distribution is known or if the expected loss can be calculated directly from the statistics (e.g., with a linear loss function, expected portfolio loss is by definition the mean of the distribution). These statistics can often be calculated analytically, saving the effort of simulation. Whether this works, depends on the risk measures and loss functions that are applied in the following step. Profiles may also be intertemporal, capturing risk at multiple points in time [47,48]. In the example, the expected loss across the whole portfolio, and thus the decrease in expected loss associated with a portfolio of risk-reduction activities, is simply the sum of the expected losses associated with each risk. If projects are independent, the variance of the total portfolio loss which can be calculated once assuming no reduction efforts and again, with a portfolio of reduction efforts, is similarly equal to the sum of the variance of the loss for each risk $i, p_i(1 - p_i)\text{loss}_i^2$. In cases with moderately complex correlations, variance for portfolios with correlated elements can also be calculated formulaically.

Calculating Portfolio Risk Metrics

From the risk profile, one can calculate summary statistics: the sufficient statistics above and possibly other quantities to be calculated from the profile that will feed into evaluation of plans.

For the dam and levee example, economic, societal, and environmental costs are all of concern. A total (multicriteria) damage measure could be calculated as a weighted sum of, say, hectares of habitat destroyed, number of people dislocated, and dollars of infrastructure lost. There are various methods in risk and decision analysis for estimating such trade-off weights. The marginal harm to the

area would become steeper as the magnitude of loss increases, as it becomes harder for regions to recover, so some nonlinear utility (or loss) function is appropriate.

For illustrative purposes, the numerical example here uses only a single damage measure, and a simple risk-averse utility function (for illustrative purposes only) over just one metric in terms of mean and variance where utility $u = 100 - \text{expected loss} - \text{variance}/10$. Thus, if all risk were eliminated, utility would be 100, and it is potentially unbounded on the downside.

In some situations similar to this one, there are specific guidelines to plan with regard to the worst case that has a 1/100 chance of occurring. That would be similar to a value-at-risk measure, which has been used in financial planning. The “worst case” would be the first percentile of simulation output. Alternatively, there may be some limit above which the loss becomes catastrophic, for example so much habitat disappears that coastal species go extinct. In that case, the probability that the total habitat loss exceeds some threshold amount could also be a simulation output.

It is reasonable to apply multiple criteria to the basic spreadsheet using a simple formula so that the total loss for the portfolio is a weighted sum of the loss function for economic, social, and environmental factors. More sophisticated multiattribute utility functions are still relatively simple to compute with formulas that take attribute measures as inputs but using them requires careful encoding of value judgments.

An appropriate utility function for decision-making uncertainty must reflect the decision maker’s/stakeholder’s preferences as follows: the expected utility for one risk profile is higher than that for another if and only if the decision maker prefers the first risk profile to the second. In risk situations, there may be no single clear decision maker and hence no single utility function can be entirely appropriate [49]. In that case, simpler and/or more robust measures could be used such as outranking-based multicriteria decision analysis methods, or analysis using a range of utility function parameters. Recommendations may not vary much over

the range of possible weightings. If they do, these results can still be informative for some further political process needed to ultimately choose priorities.

Choosing a Risk Management Plan

Once the portfolio level measure can be calculated for any plan, the optimal plan must be selected. The gold-standard approach is mathematical programming or optimization about which there are many textbooks [50] to find the values of decision variables that maximize or minimize an objective function subject to a set of constraints. For risk portfolios, the objective could be to minimize the expected value of the loss function or to minimize the probability that loss will exceed some level. The objective could involve one dimension or multiple dimensions. Decision variables typically allocate resources to activities or investments, but could also include scheduling and timing of efforts. Constraints typically include the available resources, but can also include policy constraints such as the probability of losses exceeding x cannot be greater than $y\%$. Constrained and combinatorial optimization methods range from the relatively simple to the highly sophisticated, requiring extensive data collection for inputs and extensive efforts at formulation. If optimization proves difficult, the choices are to simplify the model or to bring in a professional with special expertise in optimization methods.

For the simple dam case as developed here that ignores interactions between risks, optimization is relatively straightforward and does not require search algorithms. In the basic case where the objective is simple expected loss and the constraint is a monetary budget, it works to rank projects in terms of a productivity index: expected loss avoided per dollar of investment and to select the top-ranking projects up to the point that the budget is expended. The same method works if the loss associated with a failure is defined as the weighted sum of losses in each of the three dimensions, in which case the expected value of that weighted sum is used

Table 4. Portfolio Results under Different Prioritization Rules									
Criterion for Funding	Do Nothing	Arbitrary Choice	Structure Size	Problem Level	Probability of Failure	Cost to Repair	Impact of Failure	Expected Loss	Expected Loss Averted per Dollar
Cumulative cost	0.00	9.86	9.70	9.77	8.94	9.26	9.13	8.14	9.94
Cumulative value	0.00	11.64	12.50	10.06	18.78	11.57	18.77	24.26	26.61
Expected loss	36.47	24.83	23.97	26.41	17.68	24.89	17.70	12.21	9.85
Final variance	290.89	203.55	184.59	210.77	151.35	198.40	74.38	80.84	49.65
Expected loss averted per dollar	0.00	1.18	1.29	1.03	2.10	1.25	2.06	2.98	2.68
Utility = $100 - E(\text{Loss}) - \text{Var}/10$	34.44	54.81	57.57	52.51	67.18	55.27	74.86	79.70	85.18
Which projects are funded?									
1	0	0	0	0	0	1	0	0	0
2	0	0	1	0	1	0	1	1	1
3	0	0	0	0	0	0	1	1	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	1	1	1	0	1	1
6	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	1	0	0	0
8	0	0	1	0	1	0	0	0	0
9	0	0	0	1	0	1	1	0	1
10	0	0	0	0	0	1	0	0	0
11	0	0	1	1	0	1	0	0	1
12	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	1	0	0	0
14	0	0	0	1	0	0	0	0	0
15	0	0	0	1	0	1	0	0	0
16	0	0	0	0	0	1	0	0	0
17	0	1	0	0	0	1	0	0	0
18	0	1	0	0	0	0	1	1	1
19	0	1	0	0	0	1	0	0	0
20	0	1	0	0	0	1	1	0	0
21	0	1	0	0	0	1	0	0	0
22	0	1	0	0	1	1	0	1	1
23	0	1	0	1	0	0	0	0	0
24	0	1	0	0	0	0	1	0	0
25	0	1	1	1	0	0	0	0	0

as the numerator of the productivity index.² If the utility function uses expected loss and variance only, a heuristic that modifies the investment's productivity to account for its impact on variance may suffice. Otherwise, a fairly simple binary search method will find the optimal portfolio.

Table 4 shows the expected loss, variance, and the other calculated measures under different prioritization rules. The columns labeled "do nothing" and "arbitrary choice" are included to give a baseline of what happens when there is no strategy, for the purpose of comparing the impact of the other strategies. Funding choices change with different rules, and as rules better approximate the maximization of portfolio utility, funding choices tend to converge toward what a gold-standard decision analysis would recommend. It is less helpful to rank choices on the basis of benefit-related characteristics that are correlated with cost in predictable ways. For example, structure size and cost are correlated, so there is no benefit to choosing either large structures—whose repair cost tends to be high, or low-cost repairs—which tend to impact only smaller structures.

Ranking projects in terms of the impact of failure or the probability of failure leads to a substantially better portfolio; ranking in terms of expected loss combines both of these factors and leads to still more improvement, and best of all by a significant margin it uses the productivity index, expected loss averted per unit of cost. For all of these better rankings, there is not a great difference in which projects are funded. Ranking with multiple criteria is not shown here but the

implementation and implications are similar. If projects share initiating events, correlated responses, or causal interdependencies, the projects would start with a similar model with additional interactions structured to calculate the consequences for the set of individual risks. Simulation results could then be used to estimate the distribution of outcomes by which potential portfolios are valued. Note that correlations would typically be entered in the simulation tool itself, for example in specifying distributions for a cell and if necessary specifying correlations between various cells. It is possible to do this within a spreadsheet, for example by using a common random number as input to formulas for different cells, but such an approach is often unwieldy.

SUMMARY

This article has described what makes a set of risks worth considering as a portfolio and how decisions about managing that portfolio of risks can be structured and executed. Some sort of quantitative modeling is almost always needed. Significant gains come just from recognizing portfolio management as an optimization problem, that is minimizing the expected value of the loss function by choice of risk management decisions subject to a resource constraint. Depending on the characteristics of the portfolio of risks and the potential actions to reduce them, different modeling approaches can be used. Simpler approaches have the benefit of taking far less time to perform and explain, but if they are too simple, they ignore important interactions between risks. This can lead to underestimating the remaining portfolio risk or overlooking ways to eliminate more risk with a fixed budget, or otherwise getting the wrong answer. The specific theoretical approaches described in this article are well documented elsewhere in risk analysis, finance, and operations research. The reader should use this article as a guideline for identifying an approach and the steps to plan for it, and to access the referenced literature for in-depth assessment and modeling guidelines.

²Although appealing in its simplicity, the use of productivity indices can have substantial limitations. A single productivity number can more easily reflect cost efficiency than cost effectiveness, and may be insufficiently sensitive to context in cases where effectiveness differs from efficiency. Examples include dealing with uncertain costs, multiple resource constraints, multistage decision processes, synergies or dissynergies across expenditures, large discrete expenditures that must be coordinated within a fixed budget, or nonfixed budgets.

This article has focused on assessment and modeling approaches for managing risk portfolios. There are certainly other related issues to keep in mind. To implement decisions, for example, requires a variety of managerial and relationship skills. It is surely helpful to have transparency in the process so that stakeholders understand the process and believe it to be fair and valid. The process itself may involve oversight, quality controls, and participation of key experts and decision makers. It can be helpful to question the robustness of models. This means asking what happens to the decisions if assumptions about key relationships or parameter values are wrong. Sensitivity analysis techniques are a good starting point for this. Such steps can both, protect against oversights and errors in analysis and create some of the trust required for implementation. Finally, implementation of efforts to reduce risk often involves handing over of responsibility to engineering and project management professionals. Careful definition and documentation of exactly which alternatives affect which risks, to what degrees, at which times, and with which resources will facilitate a smooth hand over.

Although explicitly managing a set of risks as a portfolio is more complicated than managing them merely as a series of individual risks, it is often still worth the additional effort. In such diverse domains as financial management, product development, and military procurement, portfolio management tools have been explicitly adopted even to the point where there is a high level job with the title "portfolio manager." Risk managers have long been managing implicit portfolios of risks, and they can productively incorporate a variety of portfolio management techniques.

Acknowledgments

We would like to thank Alex Tkachuk, Laure Canis, and Dr. Lambert for advice and discussions. We would also like to thank the assistance of Drew Loney. This effort was sponsored in part by the Civil Works Basic Research Program by the US Army Engineer Research and Development Center.

Permission was granted by the USACE Chief of Engineers to publish this material.

REFERENCES

1. Pate-Cornell ME, Dillon RL. The respective roles of risk and decision analyses in decision support. *Decis Anal* 2006;3(4):1–13.
2. Lee S, Stevenson S. Real estate portfolio construction and estimation risk. *J Property Invest Finance* 2005;23(3):234–253.
3. Wooldridge SC. Balancing capital and condition: an emerging approach to facility investment strategy [Doctoral Thesis]. Massachusetts Institute of Technology; 2002.
4. Menke M. Managing R&D for competitive advantage. *Res Tech Manag* 1997;40(6):40–42.
5. Skaf M. Portfolio management in an upstream oil & gas organization. *Interfaces* 1999;29(6):84–104.
6. Kauffman RJ, Sougstad R. Risk management of contract portfolios in IT services: the profit-at-risk approach. *J Manag Inform Syst* 2008;25(1):17–48.
7. Botkin BL. Applying financial portfolio analysis to government program portfolios [Masters thesis]. Naval Post-Graduate School, Jun. 2007.
8. Howard RA. Decision analysis: practice and promise. *Manag Sci* 1988;34(6):679–695.
9. Sharpe PT, Keelin T. How SmithKline Beecham makes better resource-allocation decisions. *Harv Bus Rev* 1998;76(2):45–57.
10. Ewing PL, Tarantino W, Parnell GS. Use of decision analysis in the army base realignment and closure (BRAC) 2005 military value analysis. *Decis Anal* 2006;3(1):33–49.
11. Belton V, Stewart T. Multiple criteria decision analysis: an integrated approach. Springer, 2001.
12. Zeleny M. Multiple criteria decision making. New York: McGraw-Hill; 1982.
13. Delqu   P. Interpretation of the risk tolerance coefficient in terms of maximum acceptable loss. *Decis Anal* 2008;5(1):5–9.
14. Ringuest JL, Graves SB, Case RH. Mean–Gini analysis in R&D portfolio selection. *Eur J Oper Res* 2004;154(1):157–169.
15. Keisler J, Buehring W, McLaughlin P, *et al.* Allocating contractor risks in the Hanford waste cleanup. *Interfaces* 2004;34(3):180–190.

16. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: Wiley; 1976.
17. Gustafsson J, Salo A. Contingent portfolio programming for the management of risky projects. *Oper Res* 2005;53(6):946–956.
18. Liesiö J, Mild P, Salo A. Preference programming for robust portfolio modeling and project selection. *Eur J Oper Res* 2007; 181(3):1488–1505.
19. Ehrgott M, Gandibleux X. Multiple criteria optimization: State of the art annotated bibliographic surveys. Kluwer Academic Publisher; 2002.
20. Markowitz HM. Portfolio selection. *J Health Care Finance* 1952;7(1):77–91.
21. Sharpe WF. Capital asset prices: a theory of market equilibrium under conditions of risk. *J Health Care Finance* 1964;19(3):425–442.
22. Sharpe WF. A simplified model for portfolio analysis. *Manag Sci* 1963;9(2):277–293.
23. Modigliani F, Miller MH. The cost of capital, corporation finance and the theory of investment. *Am Econ Rev* 1958;48(3): 261–297.
24. Brealey A, Myers SC. Principles of corporate finance. 7th ed. New York: McGraw Hill; 2003.
25. Black F, Scholes M. The pricing of options and corporate liabilities. *J Polit Econ* 1973; 81(3):637–654.
26. Hallerbach W, Spronk J. A multi-dimensional framework for financial-economic decisions. *J Multi-Criteria Decis Anal* 2002; 11(4–5): 111–124.
27. Hallerbach W, Spronk J. The relevance of MCDM for financial decisions. *J Multi-Criteria Decis Anal* 2002;11(4–5):187–195.
28. Smith JE, Nau RF. Valuing risky projects: option pricing theory and decision analysis. *Manag Sci* 1995;41(5):795–816.
29. Bell DE. Risk, return, and utility. *Manage Sci* 1995;41(1):23–30.
30. Aerts JCH, Botzen W, van der Veen A, *et al.* Dealing with uncertainty in flood management through diversification. *Ecol Soc* 2008;13(1):Article 41. Available at <http://www.ecologyandsociety.org/vol13/iss1/art41/>. Accessed 2009 May 8.
31. Society for Risk Analysis. Available at http://sra.org/resources_glossary_p-r.php. Accessed 2009 Feb 8.
32. Environmental Protection Agency, website. Available at <http://www.epa.gov/risk/basicinformation.htm>. Accessed 2009 Feb 8.
33. Linkov I, Kiker G, Wenning R. Environmental security in harbors and coastal areas: management using comparative risk assessment and multi-criteria decision analysis. Amsterdam: Springer; 2007.
34. Dillon RL, Pate-Cornell ME, Guikema SD. Programmatic risk analysis for critical engineering systems under tight resource constraints. *Oper Res* 2003;51(3):354–370.
35. Pate-Cornell ME, Dillon RL, Guikema SD. On the limitations of redundancies in the improvement of system reliability: the case of the Mars Rovers. *Risk Anal* 2004;24(6):1423–1436.
36. Sage AP. Systems engineering. New York: Wiley; 1992.
37. Parnell GS, Driscoll PJ, Henderson DL. Decision making in systems engineering and management. New York: Wiley; 2008.
38. Richardson G, Pugh A. Introduction to system dynamics modeling. Waltham (MA): Pegasus Communications; 1981.
39. Duffie D, Pan J. An overview of value at risk. *J Derivatives* 1997;4(3):7–49.
40. Dillon RL, Pate-Cornell ME. APRAM: an advanced programmatic risk analysis method. *Int J Tech Pol Manag* 2001;1(1):47–65.
41. Linkov I, Satterstrom FK, Kiker G, *et al.* From comparative risk assessment to multi-criteria decision analysis and adaptive management: recent developments and applications. *Environ Int* 2006;32:1072–1093.
42. Lambert JH, Peterson KA, Joshi NN. Synthesis of quantitative and qualitative evidence for risk-based analysis of highway projects. *Accid Anal Prev* 2006;38:925–935.
43. Lambert JH, Joshi NN. Diversification modeling for risk management of safety and security programs. Proceedings of the Ninth International Conference on Probabilistic Safety Assessment and Management (IAP-SAM). Hong Kong; 2008.
44. Department of the Army, U.S. Army corps of engineers, screening for portfolio risk analysis (SPRA) guidelines for dams. Circular, No. 1110-2-xxx, Dec. 2007.
45. Federal Emergency Management Agency, FEMA: Risk prioritization tool for dams users manual, FEMA P-713CD (CD-ROM). Jun. 2008.
46. Law A, Kelton WD. Simulation modeling and analysis. New York: McGraw Hill; 1999.
47. Mauney DA. Using financial/decision analysis risk optimization to prioritize maintenance expenditures, justify maintenance spending

- and maximise the NPV savings of the maintenance function. Structural integrity associates technical paper. Available at <http://www.structint.com/tekbrefs/t99018r0.pdf>. Accessed 2009 May 8.
48. Stummer C, Heidenberger K. Interactive R&D portfolio analysis with project interdependencies and time profiles of multiple objectives. *IEEE Trans Eng Manag* 2003;50:175–183.
 49. Arrow KJ. A difficulty in the concept of social welfare. *J Polit Econ* 1950;58(4):328–346.
 50. Hillier F, Lieberman G. Introduction to operations research. New York: McGraw-Hill; 2005.

MANAGING CORPORATE MOBILE VOICE EXPENSES: PLAN CHOICE, POOLING OPTIMIZATION, AND CHARGEBACK

SAMUEL S. CHIU
Department of Management
Science and Engineering,
Stanford University,
Stanford, California

INTRODUCTION

Mobility communication management spans the spectrum of user choice, device procurement/inventory/replacement, IT security, corporate policy compliance, usage monitoring, expense optimization, and billing. Today, there are many vendors providing a subset or the full spectrum of such mobility communication management services to corporate customers, a service collectively (perhaps not quite accurately) known as TEM, *telecom expense management*. The focus of this article is on mobile voice expense optimization. A (stand-alone) mobile voice subscription plan, which is readily available to a consumer, can generally be characterized by three parameters: monthly recurring charge, monthly (peak) minute allowance, and per minute overage charge, denoted by the triplet (F, B, δ) , a fixed charge, a block of minute allowance, and δ per minute if the monthly peak usage exceeds B . An (stand-alone) individual user has to decide which plan to choose: too large a base plan will result in unused wasted minutes, while too small a plan will incur expensive overage charges. Similar to family plans, a corporation has different pooling options. A pooling plan allows subscribers in the same pool to share a common pool of minutes. An overage results when the collective usage exceeds the common pool of minutes. We examine the mathematics behind pooling and how to create the best

pooling arrangement minimizing mobile voice expenses. Depending on the options chosen (or often negotiated between a corporation and its mobile carrier), the mathematics can become quite challenging. It turns out that the most complex combinatorial optimization pooling problem can be efficiently and effectively solved using a column generation technique. We provide case studies to illustrate the potential (and actual savings) for the stand-alone plan choice decision as well as the collective pooling optimization; typically resulting in 20–30% of annual savings, which translate into tens of millions of dollars for large corporations. A very important, often a stumbling block to implement a pooling solution, aspect of pooling is *chargeback*—how to allocate monthly pooled expenses to individual accounts (which often belong to different units in an organization). We share with you a simple and equitable chargeback scheme which is well received and readily implemented by corporate clients.

Column generation techniques likely trace its origin to Ford and Fulkerson [1] and Dantzig and Wolfe [2], while Gilmore and Gomory [3] specialize this decomposition principle to the cutting stock problem. Barnhart [4] and Moretti *et al.* [5] provide excellent literature background on this subject. This article focuses on the application of this technique to a real-world problem with its main contribution in problem formulation, turning a seemingly intractable combinatorial optimization program into a manageable problem, which results in substantial savings, both in absolute and percentage terms.

We first introduce notations to define our optimization problem: the input parameters, available data and, most importantly, formulating an objective function. Once we are able to evaluate the objective function of one's decision, we turn to the issue of optimizing. In one instance of our optimization problem, it becomes necessary to reformulate the problem to allow for a tractable

solution approach. We illustrate our problems with many examples, concluding with an actual case study illustrating the benefit of optimization.

PRELIMINARIES: INDIVIDUAL USAGE PATTERNS AND SUBSCRIPTION PLANS

We define various terms and parameters in this section.

USER SPECIFIC PARAMETERS:

- N number of users
- i user index, $1 \leq i \leq N$
- $\mathbf{m}_i^m$ minute usage for user i in month m , a random variable
- $\mathbf{m}_i$ $\mathbf{m}_i^m$ is independent and identically distributed (iid) with distribution $\mathbf{m}_i$
- μ_i monthly mean (peak) minute usage for user i , expectation of $\mathbf{m}_i$
- σ_i standard deviation, SD, of $\mathbf{m}_i$

SUBSCRIPTION PLAN SPECIFIC PARAMETERS:

- K number of available plans
- k plan index, $1 \leq k \leq K$

- F_k monthly fixed cost (also known as MRC, *monthly recurring cost*) for plan k
- B_k monthly peak minute allowance for plan k
- δ_k per minute overage charge when monthly usage is above B_k for plan k

A usage pattern (or profile) is represented generally by a distribution of the random variable $\mathbf{m}$, with mean-SD pair (μ, σ) . Plan k is represented by the triplet of plan parameters (F_k, B_k, σ_k) .

It is instructive to graphically view the total (monthly) cost of each of the six plans as a function of usage, superimposed by several usage patterns (Figure 1). For illustration simplicity, we use a normal density function to represent a usage pattern. It turns out that the issue of negative usage (of a normal random variable) can be explained away; this will be addressed after the derivation of Equation (2).

- Each plan has a fixed MRC up to its monthly allowance. Usage above the allowance will incur a linear cost shown as the slope of the two-part tariff.

Table 1 shows four usage patterns

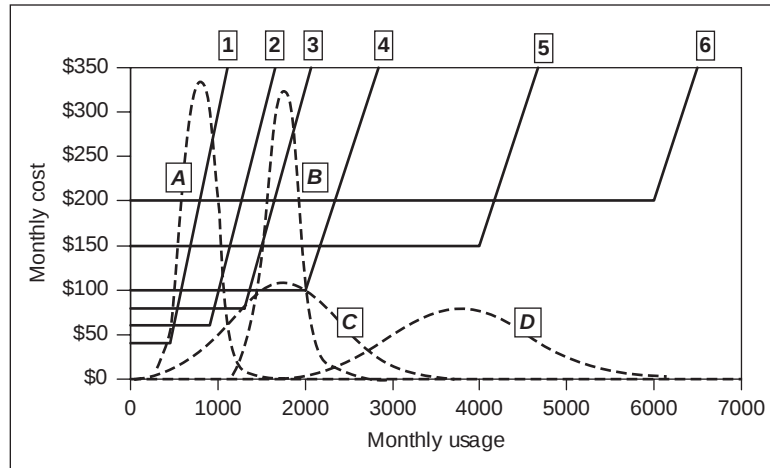


Figure 1. Plan and usage profile.

Table 1. Usage Pattern/Profile (μ, σ)

	μ	σ
A	800	190
B	1750	200
C	1750	600
D	3800	800

- Usage patterns *B* and *C* consume the same peak minutes on the average while pattern *C* has a higher usage variation. It is, perhaps, intuitive that usage *C* may benefit from subscription to a higher plan (e.g., Plan 5) since this user routinely exceeds the allowance of Plan 4 even though its average usage is below Plan 4's allowance. Usage pattern *B* rarely exceeds the monthly allowance of Plan 4, thus the appropriate (best) subscription is likely to be Plan 4.

We define $\text{Cost}[m; (F, B, \delta)]$ as the monthly cost when m minutes are consumed under a plan with parameters (F, B, δ) . We note that the usage m has the distribution of $\mathbf{m}$.

$$\text{Cost}[m; (F, B, \delta)] = F + \delta \max(0, m - B). \quad (1)$$

MODEL FORMULATION

Modeling Usage Distribution and Cost Function

The following observations can be made after the examination of many individual usage patterns.

- Unlike personal users, corporate liable account users do not *watch* their minutes toward the end of the month: to be more frugal if current usage is near (or has already exceeded) the allowance.
- There are slight seasonal fluctuations over the course of a year (e.g., Christmas and Thanksgiving), but they are not

significant—again quite distinct from a personal account.

- Most users follow a rather stable usage pattern when they stay on the same job with similar responsibility.
- The usages of different users can be reasonably assumed to be independent.
- It is generally, but not always, true that the higher the average usage, the higher the variability (i.e., standard deviation of usage).

Given these observations, it is appropriate to use expected cost as the minimizing objective when a user is to choose the best plan (least cost over the long run). In order to evaluate the expected monthly cost of a user, one needs to choose a usage distribution of $\mathbf{m}$. There are many choices of distributions with parameters chosen to fit usage statistics (e.g., the mean and standard deviation of monthly usage). We have considered various distributions (e.g., normal, triangular, and trapezoid) and found that the exact choice of distribution generally does not impact the optimal plan selection, while the mean and standard deviation of usage play an important role. We eventually choose the normal distribution to represent a usage pattern by matching its mean and standard deviation from historical usage data.

Computing Usage Mean and Standard Deviation

For each user, we compute the mean and standard deviation of monthly usage if enough history is available. The mean and standard deviation will be updated each month to incorporate new usage data. A new user's statistics have to be estimated, either by a three-point triangular distribution (minimum, most likely, and maximum usage, via questionnaire) or using the *average* of the user base.

Computing the Average Monthly Cost

The following formulae are used to compute the monthly usage cost for each subscriber, based on their usage statistics (or usage profile) and plan parameters. It is an exercise in

integration (by parts) using the normal distribution and the two-part tariff of Equation (1).

$$\begin{aligned}
C[(\mu, \sigma); (F, B, \delta)] &= \int_m \text{Cost}[m; (F, B, \delta)] f_m(m; \mu, \sigma) dm \\
&= \int_{-\infty}^{\infty} F f_m(m) dm + \delta \int_{m=B}^{\infty} (m - B) f_m(m; \mu, \sigma) dm \\
&= F + \frac{\delta \sigma}{\sqrt{2\pi}} e^{-0.5z^2} - \delta \sigma z [1 - \Phi(z)]. \quad (2) \\
z &= \frac{B - \mu}{\sigma}, \\
f_m(m; \mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{m - \mu}{\sigma} \right)^2}, \\
-\infty &< m < \infty, \\
\Phi(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-0.5x^2} dx.
\end{aligned}$$

To derive Equation (2) regardless of usage a flat MRC will be charged (in addition to overage, if any) accounting for the first term in Equation (2). Therefore, *negative* value of the normal random variable impacts the calculation only in the probability that usage will be below the monthly allowance B and not on the exact distribution of the chosen random variable. As discussed earlier, we have considered various distributions (e.g., normal, triangular, trapezoid) and found that the exact choice of distribution generally does not impact the optimal plan selection, while the mean and

standard deviation of usage play an important role.

RIGHT SIZING

Find the plan that will minimize a usage pattern (μ_i, σ_i) with plan choices (F_j, B_j, δ_j) , and $1 \leq k \leq K$.

$$\begin{aligned}
\text{Min}_{1 \leq k \leq K} C[(\mu_i, \sigma_i); (F_k, B_k, \delta_k)] \\
= C[(\mu_i, \sigma_i); (F_{k_i^*}, B_{k_i^*}, \delta_{k_i^*})] \equiv C_i^*. \quad (3)
\end{aligned}$$

In words, plan k_i^* is the expected cost minimizing plan for user i , leading to optimal expected cost C_i^* .

It is our experience that the optimal plan is robust in the choice of distributions (normal, triangular, trapezoid) when we match a user's mean and standard deviation of minute usage with historical data.

Because of incentive conflicts, the plan (originally) recommended by a carrier's sales representative is (almost) always not the least cost option for the user. The savings due to right sizing are in the 10–25% range.

A Numerical Example

Table 2 shows an example of the right-sizing calculation with seven users and six available plans. The expected monthly cost is calculated using Equation (2), showing the impact of the user profile pair (μ, σ) on the optimal plan choice: a smaller average usage does

Table 2. Right-Sizing Examples

		Available Plans						Best Plan
		1	2	3	4	5	6	
Usage Profile		F (\$)	B	δ (\$)	F (\$)	B	δ (\$)	
μ	σ							
325	190	\$53.10	\$60.02	\$79.99	\$99.99	\$149.99	\$199.99	1
800	190	\$198.59	\$74.41	\$80.08	\$99.99	\$149.99	\$199.99	2
800	400	\$216.41	\$105.81	\$87.07	\$100.04	\$149.99	\$199.99	3
1750	200	\$624.99	\$399.99	\$237.79	\$103.03	\$149.99	\$199.99	4
1750	900	\$638.45	\$433.29	\$299.80	\$174.33	\$150.53	\$199.99	5
3800	1000	\$1,547.54	\$1,220.21	\$955.69	\$644.27	\$242.06	\$201.46	6
5000	250	\$2,087.49	\$1,699.99	\$1,374.99	\$999.99	\$449.99	\$199.99	6

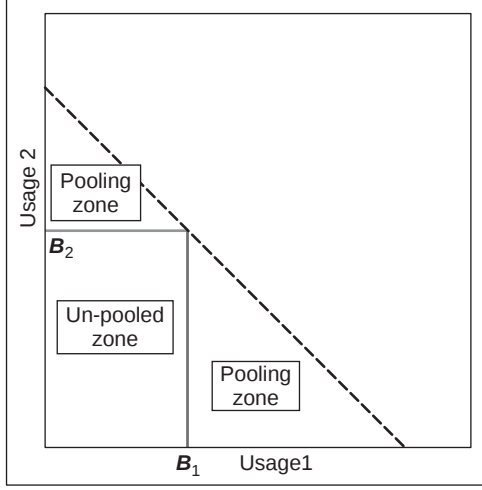


Figure 2. Pooling benefit zones.

not always favor a plan with smaller minute allowance.

POOLING CONCEPT

A pooling plan allows subscribers in the same pool to share a common pool of minutes. An overage results when the collective usage exceeds the shared pool of minutes. Figure 2 shows the benefit of pooling for two users with respective allowances B_1 and B_2 . When the monthly usage (m_1, m_2) is in the pooling zone, there will be no overage charges in a pooled environment as the pooled allowance equals $B_1 + B_2$. On the other hand, overage will incur in these two pooling zones under separate individual subscription plans.

We examine two pooling schemes in this article: total pooling and bucket pooling.

TOTAL POOLING

The specifics for total pooling are:

- N users $\{(\mu_i, \sigma_i)\}_{i=1}^N$ and K plans $\{(F_k, B_k, \delta_k)\}_{k=1}^K$.
- N plans are to be selected for the pool: number of plans = number of users.

- There will be a monthly pooling fee, f , for each user in the pool, considered as a feature fee.
- The total usage $\mathbf{M} = \sum_{i=1}^N \mathbf{m}_i$ assumes a normal distribution with mean and standard deviation computed as $(\mu_M, \sigma_M) = \left(\sum_{i=1}^N \mu_i, \sqrt{\sum_{i=1}^N \sigma_i^2} \right)$.
- The pool will collectively subscribe to n_k plans of type k , with pooled parameters defined as $(F^P, B^P, \delta^P) \equiv \left(\sum_{k=1}^K n_k F_k, \sum_{k=1}^K n_k B_k, \frac{1}{N} \sum_{k=1}^K n_k \delta^k \right)$.
- The constraint on $\{n_k\}_{k=1}^K$ is $\sum_{k=1}^K n_k = N$.

The total pooling optimization problem becomes

$$C_{TP}^* \equiv Nf + \underset{\{n_k\}_{k=1}^K, \sum_{k=1}^K n_k = N}{\text{Min}} C[(\mu_M, \sigma_M); (F^P, B^P, \delta^P)]. \quad (4)$$

We will not provide details to solving this integer programming problem, except that the optimal pool cost (including the monthly feature fee Nf) is substantially below the sum of the individually optimized plans—showing the benefit of pooling. We delay our numerical example presentation till the last section where we provide details of a case study.

CHARGEBACK

If a user is in a plan by himself, he/she (or his/her department) will be responsible for all its charges. In a pooled environment, how should the collective cost be charged back to each user account? For example, if the pool incurs a cost of \$10,000 and a (peak) usage of 100,000 min, should a user pay proportionally to his/her usage, at \$0.1/min? The answer is no. Such a proportional chargeback scheme would imply that everyone will be paying for his/her average usage over the long run, which does not differentiate between users with different degree of usage variations. Consider two usage profiles (μ, σ_1) and (μ, σ_2) with $\sigma_1 \gg \sigma_2$. Results of right sizing would likely return a larger subscription plan for user 1 providing a larger cushion to avoid

likely overage charge. It is arguably fair to charge user 1 more even when these two users consume the same number of minutes in a given month. The following chargeback scheme is proposed and implemented:

- Right size user j resulting in optimal plan k_j^* with $(F_{k_j^*}, B_{k_j^*}, \delta_{k_j^*})$.
- If user j consumes m_j minutes, use Equation (1) to calculate his/her cost as if he/she were subscribing to his/her optimal stand-alone plan: $C_j = \text{Cost}[m_j; (F_{k_j^*}, B_{k_j^*}, \delta_{k_j^*})] = F_{k_j^*} + \delta_{k_j^*} \max(0, m - B_{k_j^*})$.
- The total “artificial” charge is: $C = \sum_{j=1}^N C_j$. However, the actual invoice to the pool, denoted as PC, should be lower than C ; resulting in a total pool benefit of $PB = PC - C$. We note that there is no guarantee that PB will be positive, but we have not experienced otherwise in many years of pooling.
- PB will be distributed back to the user proportional to their spend: $PB_j = \frac{C_j}{C} PB$.
- Each user will be paying less than subscription to the best stand-alone plan.

This chargeback scheme is well received by corporations because it shows the benefit of pooling when each user is paying less than what he/she can do optimally on his/her own. There are other variations on this chargeback scheme, which we will not detail here.

BUCKET POOLING

This is by far the most interesting pooling scheme, which naturally uses the column

generation technique. Our contribution is mainly in the formulation of the optimization problem, from a seemingly intractable combinatorial optimization to a manageable problem with pleasing interpretation. We will first specify the rules of bucket pooling.

- There are L “bucket” plans, each with a triplet of parameters. These plans are usually larger in size (in the monthly charges as well as allowances).
- One can place any number of users in one of these bucket plans for a per user fee g . The total feature fee equals g times the number of users in a plan, minus one—the *primary* user of each bucket is not charged a feature fee.
- Users in a bucket will share the minute allowance for that bucket, with overage being charged as a per-minute cost when all bucket minute allowances are collectively exhausted.
- Every user is assigned to exactly one bucket, shared or unshared.

A moment’s reflection should convince the readers that this is a rather challenging combinatorial optimization problem. Table 3 shows an example with partial plan assignments to highlight its complexity. There are 1000 users to be assigned to five available bucket plans, each with its respective plan parameters. For examples, users 5 and 7 share the #1 plan, while user 8 is all by herself subscribing to plan #2. Once each user is assigned to a plan, we use the evaluation formula (2), with additional line charge fees, to compute the expected cost of such a user grouping; and to eventually seek the optimal grouping. We will not know *a priori* how

Table 3. Sample Bucket Pooling Assignments

	Monthly (\$)	Minutes Allowance	Overage (\$/min)	Additional Line Fee (\$)	Assigning Users	Total Line Fee (\$)
Plan 1	89.99	900	0.40	5.00	5, 7	5.00
Plan 1	89.99	900	0.35	5.00	3, 12, 28	10.00
Plan 2	99.99	1200	0.35	5.00	8	0.00
Plan 3	134.99	2000	0.35	5.00	11, 25	5.00
Plan 3	134.99	2000	0.35	5.00	6, 13, 20	10.00
Plan 4	179.99	3000	0.35	5.00	10, 15, 17	10.00
Plan 5	250.00	5000	0.30	5.00	1, 2, 4, 9	15.00

many plans (and of what kind) will be used and how many (and who) will be assigned to each plan.

To reduce such assignment complexity, we seek to classify all users into a small number (between 10 and 20) of *types* (indexed by t), each characterized by its common mean and standard deviation pair (μ_t, σ_t) . We now say a few words about our clustering effort since the focus of this article is on the application of column generations technique. Clustering users into types is, perhaps, the most important contribution/observation of this article because this formulation allows an efficient algorithm to solve this complex combinatorial optimization problem with demonstrable benefit in a real-world application.

The eventual optimal “bucketing” will tell us how many users of what type will be assigned to what plan.

We view our clustering effort as a location–allocation problem on a two-dimensional Euclidean (μ, σ) space:

- Locate T points on the (μ, σ) plane, each with coordinate (μ_t, σ_t) . View these T points as centroids.
- Allocate each of the N users, each with its own mean and standard deviation pair to its closest centroid (using the Euclidean distance) to form N clusters.
- Find the location of the new centroid for each of the clusters by minimizing the sum of the distances of the cluster members to the “optimal” centroid.
- Repeat until there is no change in the locations of the centroids in the location step, reaching a fixed point.

The following notes are in order:

- The location-allocation procedure is a heuristic: the terminating fixed point depends on how one chooses the initial T anchoring centroids. Since the procedure is relatively fast, we examine several starting points and choose the least cost one.
- We use the Weiszfeld algorithm to solve the location problem for each of the clusters [6–9].
- Figure 3 is a scattered plot showing the diversity of these user profiles (μ, σ) for more than 15,000 users. Note that there are really no natural clusters dividing these users.
- Our intent is to establish “good enough” clusters, where the (μ_t, σ_t) centroid coordinates provide diversity in their values to cover the usage profiles (range) of all the N users. The centroids (μ_t, σ_t) of these T clusters cover

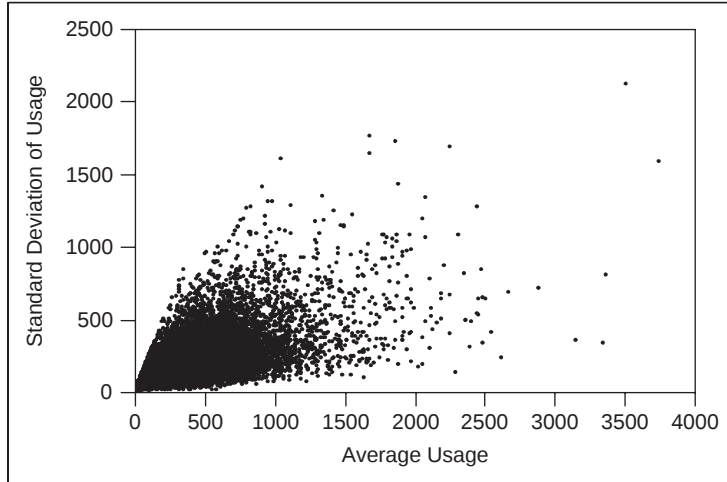


Figure 3. Diversity of usage profiles (μ, σ) .

large/medium/small values in terms of the mean and standard deviation of monthly usage.

It turns out the bucket pooling option is not as cost effective for a corporation, both in terms of savings and the ease to optimize/manage/monitor. The total pooling option is simpler to optimize/monitor/manage and it benefits more from the mathematics of probability: the mean of the sum (of independent random variables) grows linearly with the number of users, while the standard deviation of the sum grows as the square root (of the number of users). The cushion (or buffer) needed to avoid overage charges in a pooling environment is much less (on a per user basis) for the total pooling option.

We now introduce the following definitions and notations:

- N Total number of users to be classified into T types.
- (μ_t, σ_t) Usage mean and standard deviation of user type t , $1 \leq t \leq T$, which is the centroid location of each cluster.
- n_t Number of users of type t , with $\sum_{t=1}^T n_t = N$.
- L Number of available bucket plans.
- l Plan index, $1 \leq l \leq L$.
- (F_l, B_l, σ_l) Plan parameter for bucket plan l .
- V^j User (column) vector, $V^j = \{v_{tj}\}_{t=1}^T$, a column vector with each vector element v_{tj} representing the number of type t users in this user vector V^j . For example, $(3, 0, 0, 1, 0, 0)$ has three users of type 1 and one user of type 4 to share a single bucket plan when there are six user types. A user vector will be assigned to the least (expected) cost bucket plan.
- Y_{V^j} Number of users in V^j , $Y_{V^j} = \sum_{t=1}^T v_{tj}$.
- (μ^j, σ^j) Mean and standard deviation of usage for user vector V^j . $\mu^j = \sum_{t=1}^T v_{tj} \mu_t$, $\sigma^j = \sqrt{\sum_{t=1}^T v_{tj} \sigma_t^2}$.

C_j^l Expected cost of user vector V^j under bucket plan l .

$$C_j^l = (Y_{V^j} - 1)g + C[(\mu^j, \sigma^j); (F_l, B_l, \sigma_l)],$$

using Equation (2). (5)

C_j Least cost for user vector V^j ,

$$C_j = \min_{1 \leq l \leq L} C_j^l. \quad (6)$$

Note that we have assumed the independence of usage of each user in calculating the mean and standard deviation of minute usage for a user vector. Suppose we are able to generate all user vectors $\{V^j\}_{j \in J}$, we can succinctly write our decision problem as follows.

Bucket pooling optimization (BPO) problem:

$$C_{BP}^* = \text{Min} \sum_{j \in J} C_j x_j$$

$$\text{BPO : s.t. } \sum_{j \in J} v_{tj} x_j \geq n_t, \quad t = 1, 2, \dots, T$$

$$x_j \text{ are nonnegative integers, } j \in J. \quad (7)$$

Each user vector V^j represents a candidate bucket (choosing the least cost plan with a bucket cost C_j) consisting of users possibly from all the user types. The optimal solution to BPO specifies that x_j^* copies of user vector V^j will be purchased to cover all the users. The objective function coefficient represents the best cost value for that particular user vector selected from all bucket plan choices. The inequality constraint t requires that all type t users are assigned to a bucket. A user vector is termed *active* when $x_j^* > 0$.

This problem formulation is identical to the (zero-dimensional) cutting stock problem. It is zero-dimensional because there is no constraint(s) on how many users can be packed into a bucket plan. Instead, the cost coefficient, using Equation (6), will price out a user vector if this user vector is too expensive. We note that this is a linear integer programming problem since the cost coefficient absorbs all information about a user

vector into a constant C_j . The challenge is to generate all the (distinct) user vectors. We will use BPO-LP to represent the LP relaxation of Equation (7) without the integrality constraint, and labeled as (7-LP). We make the following observations about this optimization problem:

- In this formulation, the plan parameters are not present. Given a user vector V^j , the least cost plan will be selected using Equations (5) and (6) resulting in the cost coefficient C_j .
- The number of such user vectors is, putting it mildly large. Therefore, it is not practical (or feasible) to generate all.
- Since there are only T constraints, the number of *active* user vectors will be small (about T if we solve our optimization problem as an LP relaxation, BPO-LP).
- Column generation technique is an obvious candidate for solving this problem, given that only a very small number of user vectors will be active. Column generation considers only the user vectors (or columns) that are attractive candidates to become active.

Solving the LP relaxation of BPO as in (7-LP): LP-BPO

We denote the optimal objective value of LP-BPO as C_{LP-BP}^* and note that $C_{LP-BP}^* \leq C_{BP}^*$ because LP-BPO is a relaxation of BPO, without the integrality constraint.

The optimality condition for LP-BPO requires that the reduced cost for each x_j to be nonnegative:

$$C_j - \sum_{t=1}^T \lambda_t v_{tj} \geq 0, \quad \text{for all } j \in J, \quad (8)$$

$\Lambda = \{\lambda_t\}_{t=1}^T$ are the shadow prices associated with $\sum_{j \in J} v_{tj} x_j \geq n_t, 1 \leq t \leq T$, the constraint of LP-BPO.

In the spirit of the simplex method, the column generation technique can be verbally expressed as follows:

Start with a basic feasible solution of LP-BPO.

- We examine the reduced cost to identify non-basic variables as a candidate to enter the basis.
- If all reduced costs are nonnegative, we have an optimal (LP)solution to LP-BPO.
- If the reduced cost of some non-basic variables are negative. We pivot them to arrive at a new basic feasible solution.
- Repeat.

Instead of working with the full set of user vectors, the column generation technique seeks to generate attractive columns iteratively by solving a series of restricted master problems (LP-BPO without all the user vectors) and subproblems (each generating an attractive user vector with negative reduced cost, if one exists) using the (optimal) dual variables from the current restricted master problem.

Restricted Master Problem

Suppose we only have a subset of the user vectors $S \subseteq J$, we define a restriction of (LP-BPO) as follows, which will be referred to as a *restricted master problem*:

$$C_{LP-BP}^*(S) = \text{Min} \sum_{j \in S} C_j x_j$$

LP – BPO(S) :

$$\text{s.t.} \quad \sum_{j \in S} v_{tj} x_j \geq n_t, \quad t = 1, 2, \dots, T. \quad (9)$$

We note that $C_{LP-BP}^* \leq C_{LP-BP}^*(S)$ since LP-BPO(S) is a restriction of LP-BPO with less columns (user vectors) being considered in its solution. The optimal dual variables of LP-BPO(S), denoted as $\Lambda(S) = \{\lambda_t(S)\}_{t=1}^T$, will be used to generate additional columns to be used in a new (augmented) restricted master problem.

Subproblem. Consider input dual variables $\Lambda(S) = \{\lambda_t(S)\}_{t=1}^T$, optimal output $U^*(S) = \{u_t^*(S)\}_{t=1}^T$.

$U^*(S)$ is a candidate user vector (column) to be included in an augmented restricted master. In words, a subproblem asks: if there exists a user vector with a negative reduced

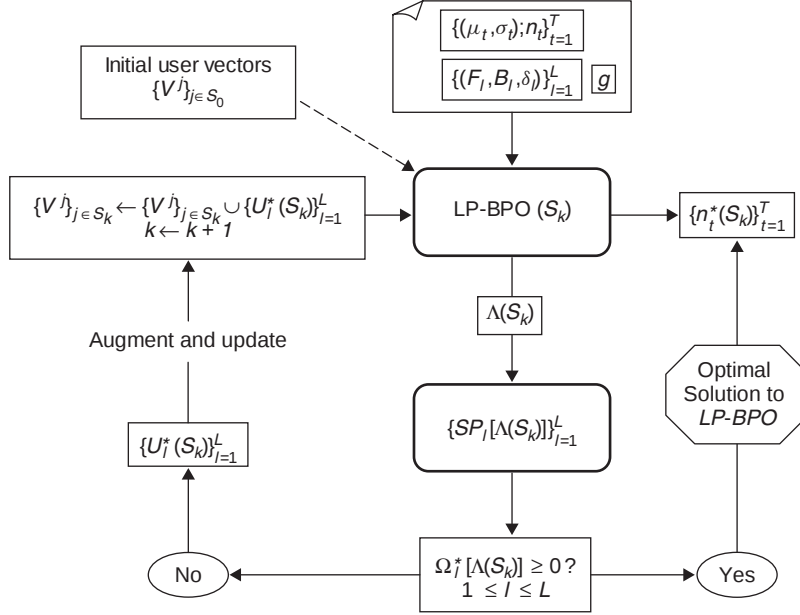


Figure 4. Flow chart: column generation to solve LP-BPO.

cost relative to the current basic feasible solution of the restrictive master problem? We note that each subproblem is associated with bucket plan l , with $l = 1, 2, \dots, L$. The optimal objective value of each subproblem represents the smallest reduced cost achievable using the current basic feasible solution in the restricted master LP.

$SP_l[\Lambda(S)] :$

$\Omega_l^*[\Lambda(S)]$

$$= \min_{\{u_t(S)\}_{t=1}^T} C_l[U(S)] - \sum_{t=1}^T \lambda_t(S) u_t(S)$$

$\{u_t(S)\}_{t=1}^T$ nonnegative integers,

$$C_l[U(S)] = [M - 1]g$$

$$+ C[(\mu, \sigma); (F_l, B_l, \delta_l)],$$

$$M = \sum_{t=1}^T u_t(S),$$

$$(\mu, \sigma) = \left(\sum_{t=1}^T u_t(S) \mu_t, \sqrt{\sum_{t=1}^T u_t(S) \sigma_t^2} \right). \quad (10)$$

Table 4. Bucket Pooling Example: User Types

User Type	μ	σ	# of Users
1	180	306	200
2	250	225	150
3	700	625	120
4	750	375	100
5	1000	500	150
6	2500	500	230
7	2500	1250	50

The cost $C_l[U(S)]$ consists of the feature fee $[M - 1]g$ and the usage associated cost $C[(\mu, \sigma); (F_l, B_l, \delta_l)]$ evaluated using Equation (2). M is the number of users in this particular user vector (or bucket assignment).

If $\Omega_l^*[\Lambda(S)] \geq 0, \forall 1 \leq l \leq L$, we have reached an optimal solution for LP-BPO—the reduced cost for all (explicit and implicit) user vectors are positive. On the other hand, we introduce new user vectors to the restricted master problem, user vectors with negative reduced costs, denoted by $\{U_l^*(S_k)\}_{l=1}^L$. If $\Omega_l^*[\Lambda(S)] \geq 0$ for a particular l , we will simply leave out bucket plan l in $\{U_l^*(S_k)\}_{l=1}^L$.

Figure 4 shows the flow of our column generation procedure to solve LP-BPO, with

index k represents the progress of the iterative algorithm:

A NUMERICAL ILLUSTRATION FOR BUCKET POOLING

We illustrate the column generation technique applied to our bucket pooling optimization problem, with problem parameters depicted in Tables 4 and 5. In this example, there are seven different user types ($T = 7$) and four bucket plans ($L = 4$).

An initial set of user vectors is used as the starting columns in the restricted master problem: at least seven linearly independent column vectors. The dual solution of the restricted master will be used to generate “attractive” user vectors to construct an augmented master problem in the next iteration. We choose to solve four different subproblems, examining each of the four bucket plan types to see if a user vector can be generated with negative reduced cost. Table 6 shows the progress of our column generation procedure. We now narrate the progress of the iterative procedure.

The algorithm starts (iteration #0) with seven user (column) vectors and it generates optimal dual variables for the seven (primal) constraints (one for each of the user types). Each constraint corresponds to satisfying the number of users in each user type, that is, all users have to be assigned to some bucket plan. The restricted master (and its dual) is an LP relaxation of the integer problem. The optimal objective value is \$70,944. Using the optimal dual in each of the (plan type) subproblem, we generate two additional columns (user vectors), bringing the number of columns to nine with which a new restricted master (iteration #1) is solved. Subsequent iterations bring the total number of columns to 10, 11, 13, 15, ... and eventually to 21. At the 9th iteration, no new columns are generated in the subproblems (the optimal value of all subproblems are nonnegative), we declare an optimal solution for the original bucket pooling optimization problem, in its LP relaxation.

Table 7 shows all the columns vectors used in the iterations. Only 7 of the 21 columns will be active in the optimal LP solution of the master problem. Note that the initial set

Table 5. Bucket Plans Parameters

Plan Type	Fixed, Cost, F (\$)	Allowance, B	Overage, σ (\$)
A	350	9500	0.25
B	225	6500	0.25
C	160	3500	0.25
D	75	1500	0.25

Table 6. Bucket Pooling: Restricted Master Program Progress

Iteration #	0	1	2	3	4	5	6	7	8	9: Optimal
Objective (\$)	70,944	65,306	55,012	54,547	54,301	53,395	53,161	53,062	53,062	53,062
# of Columns in Restricted Master	7	9	10	11	13	15	17	19	20	21
Shadow Prices	20.885	14.562	14.562	14.562	14.562	14.245	14.245	14.245	14.239	14.238
	14.961	14.961	14.961	14.961	14.362	14.362	14.362	14.362	14.362	14.362
	45.252	45.252	45.252	41.376	41.376	41.376	41.376	41.338	41.342	41.343
	39.353	39.353	39.353	39.353	39.353	30.822	37.978	37.978	37.978	37.978
	83.496	53.282	53.282	53.282	51.370	51.409	51.409	50.781	50.783	50.784
	161.061	161.061	110.849	110.849	111.486	111.473	107.895	107.895	107.895	107.895
	111.779	114.940	140.047	140.047	139.728	139.893	137.344	137.344	137.344	137.344

Table 7. Bucket Pooling: Columns Generated by Iteration

	Initial User Vectors							1 st		2 nd	3 rd	4 th		5 th		6 th		7 th		8 th	9 th
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
	C	A	A	B	A	C	A	A	B	A	A	A	B	A	B	A	B	A	B	A	A
	10	0	0	0	1	0	1	47	0	0	0	1	0	43	0	0	0	7	2	42	11
User Vector	0	37	0	0	5	0	0	0	0	0	0	0	24	0	0	0	3	0	0	0	0
	0	0	13	0	0	0	0	0	0	0	11	0	0	0	0	0	0	9	0	0	8
	0	0	0	8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	5	0	0
	0	0	0	0	10	0	0	0	6	1	0	8	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	1	0	0	3	0	0	0	0	2	0	2	0	0	0	0
	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0

Table 8. Bucket Pooling: LP Solution Summary

			User Vector # and Best Plan									Round-Up Solution	
μ	σ	User Type	13 B	15 B	16 A	17 B	18 A	19 B	21 A	# Users of Each Type	# Users Served	Over Served	
180	306	1	0	0	0	0	7	2	11	200	205	5	
250	80	2	24	0	0	3	0	0	0	150	165	15	
700	625	3	0	0	0	0	9	0	8	120	125	5	
750	375	4	0	1	0	0	0	0	0	100	100	0	
1000	500	5	0	0	0	0	0	5	0	150	150	0	
2500	500	6	0	2	0	2	0	0	0	230	230	0	
2500	1250	7	0	0	3	0	0	0	0	50	51	1	
			Various Solution Quantities, by User Vector										
LP Solution			4.375	100	16.66667	15	4.651163	30	9.767442				
Round-Up			5	100	17	15	5	30	10				
Round-Down+			4	100	16	15	4	30	9				
Cost of Each Vector(\$)			344.70	253.77	412.03	258.88	471.76	282.39	487.36	Solution (\$)			
LP Cost (\$)			1,508.04	25,376.79	6,867.22	3,883.15	2,194.21	8,471.84	4,760.31	Infeasible	53,062		
Round-Up (\$)			1,723.48	25,376.79	7,004.56	3,883.15	2,358.78	8,471.84	4,873.65	Feasible	53,692		
Round-Down (\$)			1,378.78	25,376.79	6,592.53	3,883.15	1,887.02	8,471.84	4,386.28	Infeasible	51,976		
Fixed Cost (\$)			225.00	225.00	350.00	225.00	350.00	225.00	350.00				
Allowance			6500	6500	9500	6500	9500	6500	9500				
Overage (\$/min)			0.25	0.25	0.25	0.25	0.25	0.25	0.25				
Feature Fee (\$)			5.00	5.00	5.00	5.00	5.00	5.00	5.00				
User Vector μ			6000	5750	7500	5750	7560	5360	7580				
User Vector σ			391.92	800.39	2165.06	720.56	2042.32	1198.86	2038.38				
z Value			1.2758	0.9370	0.9238	1.0409	0.9499	0.9509	0.9419				

of user vectors are linearly independent, otherwise, the initial restricted master problem may be infeasible.

Table 8 shows various statistics of the optimal solution of BPO-LP. As an example, user vector #17 consists of three users of type 2 and two users of type 6, with a total of five users. This user vector has an expected usage of 5750 min and a standard deviation of usage 720.56 min.

$$\begin{aligned}\mu^{j=17} &= \sum_{t=1}^T v_{tj} \mu_t \\ &= 3 \times 250 + 2 \times 2500 = 5750. \\ \sigma^{j=17} &= \sqrt{\sum_{t=1}^T v_{tj} \sigma_t^2} \\ &= \sqrt{3 \times 80^2 + 2 \times 500^2} = 720.56.\end{aligned}$$

The usage profile of this user vector $(\mu^{17}, \sigma^{17}) = (5750, 720.56)$, will be evaluated under bucket plan B, with plan parameters $(F_B, B_B, \delta_B) = (225, 6500, 0.25)$, using Equation (2). The cost of this user vector is calculated to be \$258.88, which includes an additional feature fee of \$20 (four additional users at \$5 each). In all, 15 such plans will be used costing $15 \times 258.88 = \$3883.15$ —the cost to serve 15 bucket plans of type B for user vector #17. Note that in the LP solution some of the bucket plans are fractional. For example, user vector #18 will be in the optimal solution using bucket plan B, 4.65 times.

Table 8 also shows the round-up (from the fractional LP solution) and a round-down solution. The round-up solution will be feasible, but it will serve more than the required users. For examples, a total of 205 users of type 1 will be included in the round-up solution while only 200 are required; for user type 2, fifteen (slack) more users are included in the round-up solution; five more for user type 3 and one more for user type 7. The LP solution has an optimal objective value of \$53,062, which is a lower bound of the optimal integer problem—but it is not feasible. The round-up solution has a value of \$53,962, which is \$630 over the lower bound of the LP solution, or no more than 1.19% over the

Table 9. Bucket Pooling: Round-Down LP Solution Deficit

User Type	# of Users of each type	Round-Down Solution	
		# Users Served	Additional Need
1	200	187	13
2	150	141	9
3	120	108	12
4	100	100	0
5	150	150	0
6	230	230	0
7	50	48	2

optimal integer solution. The last row of this table shows the z -value of each user vector: how many standard deviations (of minute usage) is the minute allowance above the mean usage. For example, we again examine user vector #17, subscribed to bucket plan B.

$$z_{17} = \frac{(B_B - \mu_{17})}{\sigma_{17}} = 1.0409. \quad (11)$$

We can view the z -value as a “cushion” for each user vector–bucket type combination, showing the margin (measured in the unit of standard deviation) that the bucket allowance is providing, above and beyond the expected usage (of the user vector). This cushion ranges from 0.92 to 1.28. The cushion is used to avoid overage charge, which costs \$0.25/min.

The round-down solution is not feasible where 36 users are not provided for Table 9 shows the deficiency: 13 type 1 users, 9 type 2, 12 type 3, and 2 type 7 users still need to be provided for. However, one can augment the round-down solution to provide a feasible integer solution.

We can then solve the much smaller residual problem: find the best user vectors and their respective least cost bucket plans to provide service to these 36 users.

Table 10 shows the details to augment the round-down solution so that the neglected 36 users will be covered, in the Round-Down⁺ row. Three more user vectors are generated. The resultant cost of the augmented round-down solution equals \$53,148, which is only 0.16% above the LP optimal solution—a slight, but nontrivial improvement over the round-up solution (where there

Table 10. Bucket Pooling: Augmenting the Round-Down LP Solution

	User Vector # and Best Plan										Round-Down ⁺ Solution		
	13	15	16	17	18	19	21	Augmenting Round-Down Solution			# Users of Each Type	# Users Served	Over Served
User Type	B	B	A	B	A	B	A	A	C	A			
1	0	0	0	0	7	2	11	13	0	0	200	200	0
2	24	0	0	3	0	0	0	0	9	0	150	150	0
3	0	0	0	0	9	0	8	0	0	12	120	120	0
4	0	1	0	0	0	0	0	0	0	0	100	100	0
5	0	0	0	0	0	5	0	0	0	0	150	150	0
6	0	2	0	2	0	0	0	0	0	0	230	230	0
7	0	0	3	0	0	0	0	2	0	0	50	50	0
LP Solution	4.375	100	16.66667	15	4.651163	30	9.767442						
Round-Up	5	100	17	15	5	30	10						
Round-Down+	4	100	16	15	4	30	9	1	1	1			
Cost of Each Vector(\$)	344.70	253.77	412.03	258.88	471.76	282.39	487.36	460.46	200.00	510.72		Solution (\$)	Maximum % Deviation
LP Cost (\$)	1,508	25,377	6,867	3,883	2,194	8,472	4,760				Infeasible	53,062	
Round-Up (\$)	1,723	25,377	7,005	3,883	2,359	8,472	4,874				Feasible	53,692	1.1886%
Round-Down+ (\$)	1,379	25,377	6,593	3,883	1,887	8,472	4,386	460	200	511	Feasible	53,148	0.1621%
Fixed Cost (\$)	225.00	225.00	350.00	225.00	350.00	225.00	350.00	350.00	160.00	350.00			
Allowance	6500	6500	9500	6500	9500	6500	9500	9500	3500	9500			
Overage (\$/min)	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25			
Feature Fee (\$)	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00			
User Vector <i>m</i>	6000	5750	7500	5750	7560	5360	7580	7340	2250	8400			
User Vector <i>s</i>	391.92	800.39	2165.06	720.56	2042.32	1198.86	2038.38	2083.81	240.00	2165.06			
z Value	1.2758	0.9370	0.9238	1.0409	0.9499	0.9509	0.9419	1.0366	5.2083	0.5081			

Table 11. Total Pooling Case Study: User Profiles

	μ	σ	σ/μ
Maximum	3746.50	2126.10	2.45
Minimum	20.00	6.16	0.03
Average	458.53	216.83	0.60

were slacks: serving more users than what is demanded).

We now present a case study to illustrate the benefit of total pooling.

TOTAL POOLING CASE STUDY

This case study involves 15,619 users with actual usage history, varying from six months to well over two years of usage data for each user. Figure 3 is a scattered plot showing the diversity of these user profiles (μ_j, σ_j).

Table 11 contains some statistics about these users. The largest monthly (average) usage is about 3750 min, while the smallest average user consumes 20 min/month, with the overall mean usage of 458.53 min/user. The variation of usage (in terms of the standard deviation of usage) range from a maximum of 2126–6.16 min. We also provide statistics of coefficient of variation of each user, defined as the ratio of σ over μ , which ranges from 0.03 to 2.45.

There are six plans available for this group of users, detailed in Table 12. This table also shows the result of our right-sizing operation using Equation (2), which finds the least average cost plan individually for each of

the 15,619 users. Many users are assigned the smallest plan, while two large users subscribe to Plan 6. Before right-sizing, most users subscribed to Plan 3 (the default plan), while large users pick larger plans.

The pre right-sizing monthly cost averaged about \$1,100,000. Right sizing alone reduces monthly cost by 17% to \$938,971. Typical savings from right-sizing range from 15% to over 25%. Note that the figure of \$938,971 is not the actual monthly cost. Rather, it is computed using Equation (2) as an average cost for each user using individually optimized plan.

Total pooling incurs an additional per user feature fee of \$4 (\$5 list price with a negotiated 20% feature fee discount, which is typical). This feature fee alone adds more than \$60,000 to this pool of users. The mean and standard deviation of pooled usage is computed to be:

$$(\mu_M, \sigma_M) = \left(\sum_{i=1}^N \mu_i, \sqrt{\sum_{i=1}^N \sigma_i^2} \right) \\ = (7,161,846, 33,588).$$

The following optimization problem will be solved to find the optimal number of plans (of each plan type), $\{n_k\}_{k=1}^K$, for this user pool, using Equation (4). Note that we have used the weighted average of all the overage rates (weighted by the number of plans of each type in the optimal pool) to represent the overage rate for the pool. It turns out the optimal plan selection is not sensitive to the

Table 12. Total Pooling Case Study: Right-Sizing Results

	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6
F (\$)	39.99	59.99	79.99	99.99	149.99	199.99
B	450	900	1300	2000	4000	6000
δ (\$)	0.45	0.40	0.35	0.30	0.30	0.30
# Subscribed	7107	6201	1481	729	99	2

Table 13. Total Pooling Case Study: Optimal Pooling Plans

	Plan 1	Plan 2	Plan 3	Plan 4	Plan 5	Plan 6
Right Sizing	7107	6201	1481	729	99	2
Total Pooling	15,586	0	0	0	0	33

Table 14. Total Pooling Case Study: Pooling Benefits

# of Users	15,619	Total Average Minute Usage in the Pool =		7,161,846			
	Subscription and Usage (\$)	Feature Fee (\$)	Total Monthly (\$)	% Saving from Previous Step	% Saving from Preoptimization	Cost per User (\$)	Cost per Average Pool Minute (\$)
Pre Optimizing	1,100,000	0	1,100,000	N/A	N/A	70.43	0.15
Right Sizing	938,971	0	938,971	17		60.12	0.13
Total Pooling	630,343	62,476	692,819	36	37	44.36	0.10

pool overage rate.

$$C_{TP}^* \equiv Nf + \min_{\{n_k\}_{k=1}^K} C \left[(\mu_M, \sigma_M); (F^P, B^P, \delta^P) \right], \text{ with } (F^P, B^P, \sigma^P) \\ \equiv \left(\sum_{k=1}^K n_k F_k, \sum_{k=1}^K n_k B_k, \frac{1}{N} \sum_{k=1}^K n_k \delta_k \right).$$

The resultant pooled solution is shown in Table 13, where only the largest and the smallest plans are active in the optimal solution.

Table 14 shows relevant cost savings statistics for this case study. Note that each user is charged a feature fee of \$4 a month for the pooling privilege.

The savings from right sizing comes from two factors. The first factor is the misalignment of incentives when a sales representative recommends subscription plans to the users with his income being dependent on commission. Many small users subscribed to Plan 3, which is the default plan for this company. Many new users are “advised” to choose a “large enough” plan, just to be safe. Once the plan assignment is made, there is no review process to adjust—which is the second factor. A typical application of total pooling (from pre-optimization) usually results in percentage savings ranging between 20 and 30%.

CONCLUDING REMARKS

Mobile communication has become a necessity in today’s corporate world. This article examines the minimization of mobile voice expenses, starting with individually optimized plan for each user to rather complex plan sharing configurations. The challenge is to determine what is to be optimized since voice (minute) usage usually varies from one month to the next. A casual corporate mobile phone user typically subscribes to a plan recommended by a sales representative motivated by commission, which is not incentive compatible. Right-sizing an individual plan serves two purposes. First, it prescribes

the least cost plan (over the long run) should an individual stay on a stand-alone plan. Secondly, right-sizing provides a benchmark against which the benefit of pooling can be quantitatively measured.

Total pooling is relatively straightforward as all users share a common pool of minutes. The usage profile of the pool (in terms of the average total pooled usage and its standard deviation) can be readily determined. The optimization is to choose the right mix of available plans so that the total (expected) pool cost can be determined.

We argue that chargeback of the total pool cost to each individual user should not be proportional to the actual usage of a user, as a user with more usage variation contributes to a more expensive plan mix configurations. The right-sizing solution provides a benchmark so that the savings (due to pooling) can be equitably distributed to the users. A case study (with more than 15,000 users) is presented to show the savings due to right-sizing and then to total pooling. Typical savings can range between 20 and 30%. In our case study, the saving is in the high 30%. Many Fortune 100 companies incur tens of millions in annual mobile communication expenses. Quantitative optimization can certainly contribute to attractive savings.

Bucket pooling assigns users to share in various bucket plans (each with its own plan parameters), whose complexity defies a solution. We reformulate the problem by clustering users into T types, each with a common usage profile (e.g., 1000 users into seven types). The resultant optimization becomes a cutting stock problem with many columns (cutting pattern, or usage vector in our environment), which will be generated as needed in a master–subproblem iteration. The master problem will generate dual variables, being passed to the subproblem. A subproblem will determine if a better column can be generated. Optimality is established if the subproblem fails to generate a better column. We provide a numerical example to show the column iteration procedure as well as a process to round (up or down) the fractional LP solution to a feasible (near) optimal integer solution, without fraction of 1% to the LP solution.

Many features are to be developed (perhaps not as mathematically interesting) to provide a complete expense management system for mobile communication services (voice and data). These features include: data plan optimization, continuous expense monitoring to detect exception, real-time alerts to inform users of less expensive alternatives (e.g., corporate conference bridge, overseas roaming), and corporate policy compliance (abuse of use). Operations research techniques can definitely be productively applied in the mobile communication expenses management space.

Acknowledgments

The author wishes to thank an anonymous reviewer for the comments and providing more details to the data description/processing of the case study.

REFERENCES

1. Ford LR, Fulkerson DR. A suggested computation for multi-commodity flows. *Manage Sci* 1958;5:97–101.
2. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
3. Gilmore PC, Gomory RE. A linear programming approach to the cutting stock problem. *Oper Res* 1960;9:849–859.
4. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. *Oper Res* 1998;46:316–329.
5. Moretti AC, de Salles Neto LL. Nonlinear cutting stock problem model to minimize the number of different patterns and objects. *Comput Appl Math* 2008;27(1):61–78.
6. Kolliopoulos SG, Rao S. A nearly linear-time approximation scheme for the Euclidean k -median problem. *SIAM J Comput* 2007;37:757–782.
7. Li Y. A Newton Acceleration of the Weiszfeld algorithm for minimizing the sum of Euclidean distances. *Comput Optim Appl* 1998;10(3):219–242.
8. Vardi Y, Zhang CH. A modified Weiszfeld algorithm for the Fermat–Weber location problem. *Math Program* 2001;90:559–566.
9. Weiszfeld E. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematics Journal* 1937;43:355–386.

MANAGING PERISHABLE INVENTORY

OPHER BARON
Rotman School of Management,
University of Toronto, Toronto,
Ontario, Canada

INTRODUCTION

This article discusses inventory management of perishable items, including the differences between inventory models that consider items' perishability and the standard models that ignore this issue. We hope that this article will help fostering the interest of students and researchers in inventory control of perishable items, and ultimately help improving the inventory control of perishable items in practice.

To emphasize the importance of appropriately managing inventory of perishables, note that Ref. 1 estimates the total cost of unsalable merchandise of suppliers to drug-stores and supermarkets in 2005 by \$2.05 billion. This cost goes directly to the bottom line of many of these companies. Moreover, higher clock speed in many industries results in shorter and shorter life-cycles of many products. The inventory control of such products might become similar to the one of perishable items as well after some period, when a newer version is presented, the value of old items declines significantly. One important difference between items with short life-cycle and perishables is that in some cases, the firm controls the introduction of the new technology. Still, much of the insight derived for managing perishable inventory may be valid for products with short life cycles.

Most of the traditional inventory models analyzed in the literature (e.g., EOQ, (Q, r) , and (S, s) policies) assume an infinite shelf-life for items (see *Multiechelon Multiproduct Inventory Management*). (An exception

is the newsvendor model discussed in the section titled "Extending the Newsvendor Model".) Thus, when the shelf-life of the items stored is finite, traditional control policies may not be sufficient. For this reason, special purpose models that focus on the management of inventory of perishable items were developed. Our purpose is to highlight the reasons for these differences and enable further investigation of the models developed for perishable items by providing relevant references. A more technical discussion of the techniques employed in the analysis of perishable items is given in the article titled *Mathematical Models for Perishable Inventory Control* in this encyclopedia.

This article continues as follows: the section titled "Modeling Inventory Problems for Perishable Items" highlights the differences between inventory models for perishables and models that ignore perishability. The section titled "Taxonomy of Inventory Models for Perishable Items with Infinite Horizon" provides a taxonomy of models for controlling inventory of perishables based on Ref. 2, and lists some of the basic results of such models. An important observation from this section is that effective heuristics are required for the management of perishable items. Thus, the section titled "Developing a Heuristic Control Policy" demonstrates the development of such a heuristic based on Ref. 3, and finally, the section titled "Future Research Directions" lists probable directions for future research on the management of perishable items.

MODELING INVENTORY PROBLEMS FOR PERISHABLE ITEMS

The analysis of inventory systems is primarily focused on the tactical question of which inventory control policies to use and the operational questions of when and how much inventory to order. By and large, these are the main questions for managing the inventory of perishable items as well. This section reminds the reader of the assumptions used

in modeling standard inventory systems for nonperished items and then lists the special features of models for perishables. It concludes by generalizing the newsvendor model to consider items with finite shelf-life. In the process, we present relevant notation.

Inventory Modeling for Nonperishables

In inventory models for nonperishables, time is measured either in a discrete (i.e., periodic) or continuous fashion. The choice of time measurement is often related to the review period (periodic or continuous) and to the model horizon, T , that may be a single period, finite, or infinite horizon. The replenishment lead time, L , may be assumed to be zero, deterministic, or stochastic (with or without order crossings). The customers arrival rate per time period, λ , may be deterministic or stochastic, each individual's demand, D , may be deterministic or stochastic, and D may be discrete or continuous. A typical assumption is that the arrival process is Poisson with rate λ ; $D = 1$ (see *Poisson Process and its Generalizations*). Finally, when there are no items on the shelf, that is there are shortages, demand is typically assumed to be backlogged or lost.

The costs considered in the management of standard inventory systems are per unit procurement cost, c ; per unit selling price, p ; order setup cost, $K \geq 0$; and holding cost per unit per time period, $h > 0$. Another cost is that of shortages. The shortage costs can include a one-time cost per shortage, $K_b \geq 0$ if shortages are backlogged, or $K_l \geq 0$ when shortages result in lost sales, and a cost per unit per unit time, $c_b \geq 0$, in cases of backlog, or a cost per unit, $c_l \geq 0$, in the case of lost sales.

Special Features of Models of Perishables

In addition to the above list of characteristics, modeling of inventory for perishable items also requires a characterization of the time to perishability, L_p . This time to perishability may be either deterministic or stochastic. Moreover, when orders arrive in batches, all items in a batch may share the same time to perishability, or each item may have its own shelf-life. An alternative assumption is that

the inventory of good items is deteriorating with time at some rate, as in Ref. 4.

The two most common models for perishability are outdatedness due to reaching expiry date (e.g., food items or medicine) and sudden perishability due to disaster (e.g., spoilage because of extreme weather conditions). The perishability due to outdatedness is typically modeled as a deterministic time to perishability and the perishability due to disaster is typically modeled as an exponential (or its discrete counterpart, geometric) time to perishability. This is because the memory-less property of these distributions often results in more tractable models.

The treatment of lead time in the management of inventory for perishable items is also not trivial. One difficulty is that the items might perish during the delivery time. This might be addressed by assuming that the supplier supplies fresh items upon their delivery and changes their lead time accordingly. For example, when the lead time is deterministic and items' shelf-life is exponential, we can assume that delivered items are fresh by modeling the lead time as a geometric random variable.

In the example above, changing the lead time to support the assumption that supplied items are fresh complicates the model because it introduces another uncertainty in the model; namely, the corrected lead time becomes uncertain. To avoid complicating the model, much of the literature on perishable items assumes that supplied items (after the lead time) are fresh.

Another complication is that when lead times are long relative to the items' shelf-lives, there might be several orders out at the same time. Then, orders might cross each other. This potentially affects the shelf-life distribution of delivered items. (Order crossing typically increases the complexity of traditional models as well.)

A general difficulty caused by the presence of lead time in a stochastic environment is that shelf-life of items in stock depends on which batch they arrived at. Thus, the controller should keep track not only of the quantity of the items on shelf but also the length of time different items are on the shelf; that is, their age and its distribution.

Another challenge in the analysis of perishable items is that items that have not yet perished may have different remaining shelf-life. (Either because items have a unique shelf-life, or because batches with common shelf-lives arrived in different periods.) Thus, the information required to completely characterize the on-hand and on-order inventory includes not only the quantity of inventory but also the remaining shelf-lives of each unit in the inventory. This increase in information requirement makes the analysis of multiechelon supply chains for perishable products especially challenging (because even for standard items, such an analysis often relies on dynamic programming and suffers from the curse of dimensionality). Therefore, some important theoretical contributions developed in the analysis of perishable items address the information required to characterize the inventory level process. (See the discussion in the section titled “Models Without Order Setup Cost or Lead Times” and in the article titled **Mathematical Models for Perishable Inventory Control** in this encyclopedia).

Another relevant decision in managing perishables when items of several different ages coexist, is items dispatching to fulfill orders. In such cases, it might not be optimal to dispatch products in a first-in-first-out fashion. For an extreme case, consider the following example.

Example 1. A last-in-first-out dispatching policy would decrease the effect of perishability when an item’s shelf-life follows a new worse-than-used distribution. (Let $F_{L_p}(s)$ denote the cumulative distribution function of the shelf-life. Then, if

$$1 - F_{L_p}(s + t) \geq (1 - F_{L_p}(s))(1 - F_{L_p}(t)) \\ \forall s \geq 0, \forall t \geq 0,$$

we say that L_p has a new worse-than-used distribution.)

While perishable items with a new worse-than-used shelf-life are hard to find in practice, this example highlights the difficulty of characterizing the optimal dispatching policy. To substantially complicate the exact

analysis of the optimal control policy for perishables, it is enough that a newly arriving batch may have a shorter shelf-life than that of existing items. Moreover, in cases when customers may control the items’ dispatching they may prefer fresher items (consider items such as milk), causing the actual dispatching to differ from first-in-first-out.

Extending the Newsvendor Model

The newsvendor model discussed in the article titled **Newsvendor Models** in this encyclopedia is an exception in the standard inventory literature, because it considers perishability. This model considers the order quantity required to maximize the profit over a single period when the demand over the period is uncertain. In addition to the costs listed above, the newsvendor model also considers the salvage cost of perishable items c_s , which can be negative or positive.

The newsvendor model can also be used to model perishable items that can be ordered only once. Now, we model the length of the single period (i.e., the lifetime) as uncertain, but there are no further procurement opportunities. In this case, the uncertainty in the demand depends on both, the uncertain demand per time period and the uncertain length of the selling season. A similar idea was first investigated in Ref. 5, which in contrast to the newsvendor model, allows several ordering opportunities before the perishability time.

To express the uncertainty in demand in these settings, we assume that the shelf-life L_p is a discrete random variable with a z -transform $Z_{L_p}(z) = \sum_{i=0}^{\infty} z^i \Pr(L_p = i)$ and that the demand in time period i is independent and identically distributed, D_i , with a probability density function (pdf), $f_{D_i}(x)$, and a moment-generating function, $L_{D_i}^*(\alpha) = \int_{-\infty}^{\infty} e^{\alpha x} f_{D_i}(x) dx$. Then, the horizon considered by the planner is $T = L_p$, the overall demand during the period, D_T , is a random sum of random variables with a moment-generating function $L_{D_T}^*$:

$$L_{D_T}^*(\alpha) = E(L_{D_i}^*(\alpha)^{L_p}) = Z_{L_p}(L_{D_i}^*(\alpha)).$$

Example 2. If $L_p \sim \text{Poisson}(p)$ and $D_i \sim \text{Normal}(\mu, \sigma^2)$ and are i.i.d, Then $Z_{L_p}(z) = e^{p(z-1)}$, $L_{D_i}^*(\alpha) = e^{(\mu\alpha + (\sigma\alpha)^2/2)}$ and the moment-generating function of overall demand over the period is

$$L_{D_T}^*(\alpha) = e^{p(\exp(\mu\alpha + (\sigma\alpha)^2/2) - 1)}.$$

Because the moment-generating function fully characterizes the demand distribution over the products' random shelf-life, the standard solution of the newsvendor model can be implemented to maximize the profit even when the valuable shelf-life of the items is uncertain.

TAXONOMY OF INVENTORY MODELS FOR PERISHABLE ITEMS WITH INFINITE HORIZON

As shown above, the newsvendor model is useful in the inventory management of perishable items that can only be ordered once. However, many perishable items such as food and medicines are consumed over a much longer horizon; then a single order is not practical. Finite horizon models were addressed in the literature almost solely in the discrete review settings. In contrast, infinite horizon models were considered in both discrete and continuous review settings. Below, we describe the primary models, assumptions, and results in the inventory literature on perishable items. We follow the classification of Ref. 2 in the section titled "Taxonomy of Inventory Models for Perishable Items with Infinite Horizon" because it is both thorough and recent. The interested reader is encouraged to read Ref. 2.

The infinite horizon models for perishables include a cost, c_p , per perished item. This cost replaces the salvage value of perished items in the newsvendor model. It is also possible to include a fixed cost, K_p , in cases where batches of items perish; however, most of the models do not include the latter cost.

A more detailed discussion of the state-of-the-art results for each of these models is given in the article titled *Mathematical*

Models for Perishable Inventory Control in this encyclopedia.

Discrete Review Systems

Models with No Fixed Order Setup Cost and No Lead Time. The simplest model is one in which items' shelf-life equals the review period. Then, different periods are independent and the standard newsvendor solution is optimal in each period. The second simplest model, in which items' shelf-life equals two review periods, is already much more interesting. This model was pioneered in Ref. 6, where it was shown that the optimal control policy is state dependent.

That the optimal control policy is state dependent, is somewhat surprising, because the optimality of the base stock control policy is easily established in discrete review models for nonperishables without order setup cost. Moreover, this result is also important because it implies that for models in this category, the simple base stock-level control is not optimal, in sharp contrast to the optimality of this control for nonperishables.

Thus, for perishable items, the optimal inventory control requires solving a dynamic programming problem. The difficulty in finding and implementing the optimal state-dependent controls motivated researchers to search for effective heuristics. For example, Broadheim *et al.* [7] suggested a heuristic that is based only on the age of the newest item in the system and the simulation-based study in Ref. 8 considered several heuristics including one that keeps the total number of items in the system fixed, ignoring their ages.

Models with Fixed Ordering Cost but no Lead Time. The research on this category was pioneered in Ref. 9, which highlights the complexity of the optimal control structure for these models and suggests heuristic for control of the system based on the (S, s) model.

A nice observation for this model of the Poisson demand with backlogs case is that the optimal reorder level is always at or below -1 . This result, which was established in Ref. 10, echoes the one in Ref. 11 for the continuous review system. The proof follows

because—with no lead time—higher reorder points would increase holding cost without reducing the backlog. (This result also holds in the case where customer's arrival process is Poisson and each customer requests some integer quantity of the product.)

Models with Lead Time. The inclusion of positive lead time leads to much more complicated models, as the number of states required for the dynamic program increases. With lead time, one needs to also keep track of the age of ordered items. This model is considered in Ref. 12, where methods to solve it are suggested. Recently, Berk [13] analyzed the (Q, r) periodic review inventory model for perishables with lead time by using the effective shelf-life distribution and assuming lost sales.

Discussion. The assumption of zero lead time in the case of a discrete review system is often justified if the lead time is shorter than the review period. Therefore, the analysis of such models (with periodic review and no lead time) is valuable. The zero lead time assumption is not as meaningful for continuous review models or when the lead time is long and makes reviewing the items at such intervals too expensive.

From an analysis point of view, treating perishable items requires consideration of the age distribution of the on-hand inventory in both ordering and dispatching of perishable items. Dispatching is also important because, typically (and in contrast to Example 1), the value of older items in stock is lower than that of the newer ones. The challenges in both ordering and dispatching of perishables causes the traditional optimal policies for nonperishable items to be suboptimal for perishables.

It turns out that the optimal control policies for perishable items are fairly hard to characterize. This difficulty encouraged researchers to develop effective heuristics for the periodic control of perishable items in the absence of lead time. However, no such heuristics were developed for models with long lead time (Williams and Patuwo [12] give an exact analysis of several relevant cases and suggest guidelines for developing such heuristics).

Continuous Review Systems

Models without Order Setup Cost or Lead Times. This category was motivated to model blood banks. It was originated by Graves [14], who assumed that the items are continuously produced, perish after a deterministic time, and that demand follows a compound Poisson process with either a single unit or an exponential demand at each arrival. A large body of work within this category was done by Perry *et al.* in Ref. 15 and references therein; thus Karasesman *et al.* [2] named it the *Perry model*.

An interesting observation that is used in the analysis of the Perry model is that its performance can often be characterized based on knowledge of the virtual death process, that is, the time until the next perishability [16].

The main assumption made in many of the models within this category is that good items arrive one by one independently of the decision making. Thus, in contrast to standard inventory models, the controller does not decide on when and how much to order. Still, the controller might affect the arrival rate of good items by advertising, for example.

Because of this lack of control on the input, most work on the Perry model focused on performance analysis rather than on finding optimal controls.

Models without Order Setup Cost but with Positive Lead Time. The first work on continuous review models without setup cost but with lead time is given in Ref. 17 where, as in many subsequent works, the performance of an $(S - 1, S)$ control policy is investigated. However the optimality of this control policy is not established, even when time to perishability and interarrival times are exponential. In fact, in view of the fact that base stock is not the optimal policy in the corresponding discrete review model, even in the absence of lead time, this control policy is probably not optimal for the continuous review models either. While some study of problems with fixed lead time and time to perishability has been pursued [18], the theoretical results about these applicable models are far from being complete.

Models with Order Setup Cost. Models with order setup cost can be portioned into ones with a fixed or variable order size, corresponding to the (Q, R) or the (s, S) standard inventory models. Both types of models can be analyzed as cost minimization problems, possibly subject to some service level constraint.

For fixed batch size problems, two policies were considered: the standard (Q, R) and the (Q, R, T) policy, where orders of a batch are triggered by either the time passed (since the inventory level was Q) being T , or by the inventory level falling below R [19].

The vast majority of work allowing for variable order size assumes no lead time, as in Ref. 11. In Ref. 11, it is established that with Poisson demand the (S, s) is the optimal policy when shortages are either backlogged or result in lost sales, and also that $s < -1$ or $s = 0$ for these models, respectively. (The insight behind the proof on the optimal reorder point is similar to the one for the corresponding discrete review model discussed in the section titled “Models with Fixed Ordering Cost but no Lead Time.”)

Recently, Gurlur and Ozkaya [20] considered a zero lead time with backlog (s, S) policy for perishables. They used sums and integrations of relevant distribution functions to express the expected cost rate function for their model and developed a heuristic for the positive lead time case.

Discussion. To summarize, stochastic analysis is the main tool used in the investigation of continuous review models for perishable items. This analysis is not trivial and the characterization of the optimal control policies is more complicated than in the periodic review models. Therefore, there is a need to find effective control policies for these models as well, especially in the presence of lead time.

There is still a place for developing efficient heuristics for the control of continuous review policies for perishables with lead time. Both the heuristic reported in Ref. 20 and the one for the no lead time case from Ref. 3, which is also reproduced in the next section, seem good starting points for finding such heuristics.

DEVELOPING A HEURISTIC CONTROL POLICY

As discussed above, there is a need for developing effective control policies for both, continuous and discrete review models of perishable items. Below, we discuss such a heuristic based on Ref. 3. Consider a continuous review model with no lead time, where the customers arrival process is Poisson with rate λ and each customer requests a continuous random quantity D of items with a mean demand size of $E(D)$. The costs considered are the order setup cost, K , holding cost per item per time period, h , and cost of perishable items, c_p . Let us focus on an (S, s) control policy and allow no backlog, which is legitimate due to the zero lead time assumption. Thus, $s = 0$ is optimal and due to the Poisson arrival process, the times where the inventory level is raised to S are renewal epochs (see **Definition and Examples of Renewal Processes**, properties of renewal processes)

In the standard $(S, 0)$ model without perishable items, the end of the inventory cycle, denoted by T_S is caused only due to the arrival of a demand. Of course, T_S is a random variable that depends on S and on the demand arrival process. But when items are perishable, cycles can also end due to perishability, at time L_p , which is the random variable describing the time to perishability of a new batch. To emphasize that the model focuses on perishable items, we denote the time to end the cycle by $\tau = \min(T_S, L_p)$ and the corresponding control policy by (S, τ) , as in Ref. 3.

Figure 1 shows a sample path of inventory over more than two inventory cycles. The first cycle ends due to the demand, as in a standard $(S, 0)$ model; then $\tau = T_{S1} \leq L_{p1}$. The second cycle ends due to perishability; then $\tau = L_{p2} < T_{S2}$.

Let $V(t)$ denote the inventory level at time $t \in [0, \tau)$ from the beginning of the cycle. Thus, $\mathbf{V} = \{V(t) : t \geq 0\}$ is a regenerative process such that $V(0) = S$. We let $V(\tau)$ denote the inventory level at the end of the cycle and use E as the expected value operator (both $E(V)$ and $E(V(\tau))$ are well defined because $\mathbf{V}$ is a regenerative process). Then, the long-run

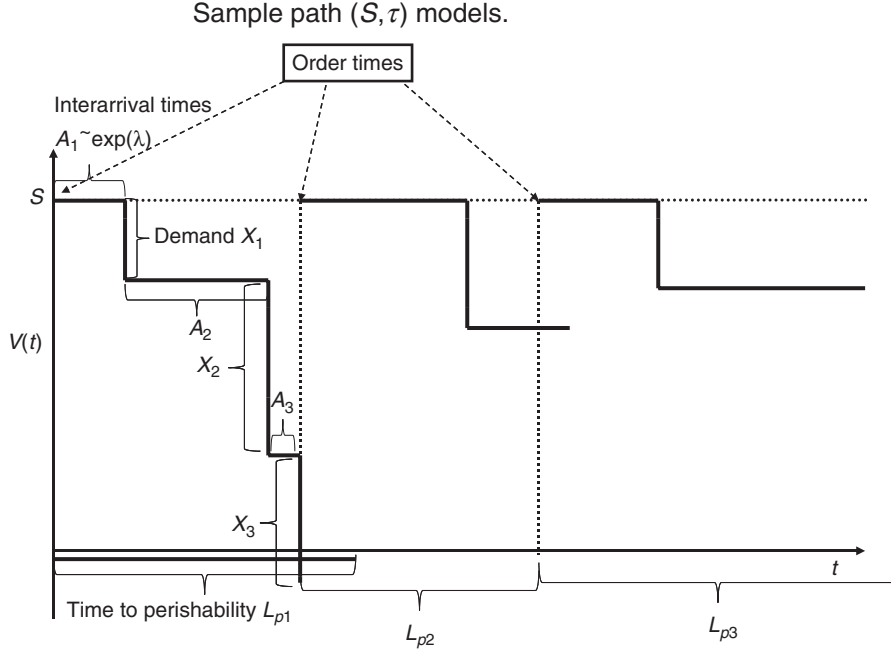


Figure 1. Sample path (S, τ) models.

average cost $C(S)$ is

$$C(S) = \frac{K + c_p E(V(\tau)I\{\tau < T_s\})}{E(\tau)} + h E(V), \quad (1)$$

where $I\{\cdot\}$ is the indicator function with a value 1, if the event $\{\cdot\}$ occurs and 0 otherwise.

To find a heuristic control policy, we replace the stochastic demand assumption with a deterministic demand with a rate $\lambda E(D)$. Then, starting at the reorder level S for any time t before the end of the cycle, we have

$$V(t) = S - t \lambda E(D) \quad t \in [0, \tau].$$

Therefore, if items do not perish within a cycle its length is $T_s = S/(\lambda E(D))$. Letting $f_{L_p}(t)$ denote the pdf of L_p , the expected cycle length is given by

$$\begin{aligned} E(\tau) &= E(\min\{L_p, T_s\}) \\ &= \int_0^{S/(\lambda E(D))} t f_{L_p}(t) dt \\ &\quad + \frac{S}{\lambda E(D)} \Pr\left(L_p > \frac{S}{\lambda E(D)}\right). \end{aligned} \quad (2)$$

Since for every $t < \tau$ we have $V(t) = S - t \lambda E(D)$ and, as in the EOQ model, the average inventory resulting from an inventory triangle of height Y is $Y/2$, we have

$$\begin{aligned} E(V) &= \int_0^{S/(\lambda E(D))} \left(S - t \lambda E(D) + \frac{t \lambda E(D)}{2}\right) \\ &\quad f_{L_p}(t) dt + \frac{S}{2} \Pr\left(L_p > \frac{S}{\lambda E(D)}\right) \end{aligned} \quad (3)$$

and

$$E(V(\tau)I\{\tau < T_s\}) = \int_0^{S/(\lambda E(D))} (S - t \lambda E(D)) f_{L_p}(t) dt. \quad (4)$$

Remark. This heuristic assumes a continuous and deterministic demand process, thus new demand arrives immediately after an old batch perishes. In practice, when there is no lead time, ordering the next batch after items perish can be postponed until the next customer's arrival. That is, the order at the end of the second cycle in Fig. 1 can be postponed until the first arrival in the third cycle. Such

postponement would increase $E(\tau)$ and can be incorporated into the heuristic by adding the expected time to an arrival to the expected cycle length whenever cycles end due to perishability. In contrast, when the demand size is continuous, it is likely that when the cycle ends due to demand there is some demand that should be satisfied from the next batch. Then when no backlog is allowed, a similar postponement is not feasible for cycles that end due to demand. In any case, postponing orders when cycles end due to perishability might not be beneficial when lead time is positive. Thus, in the example given below, we ignore such postponement and use $E(\tau)$ as given in Equation (2).

Example 3. The heuristic when $L_p \sim \exp(\xi)$: Assuming $L_p \sim \exp(\xi)$ using Equations (2)–(4), we get the approximated control problem:

$$\min_S C^h(S) = \frac{K + \pi \left(S - \frac{\lambda E(D) \left(1 - e^{-\xi \frac{S}{\lambda E(D)}} \right)}{\xi} \right)}{\frac{1 - e^{-\xi \frac{S}{\lambda E(D)}}}{\xi}} + h \left(S - \frac{\lambda E(D) \left(1 - e^{-\xi \frac{S}{\lambda E(D)}} \right)}{2\xi} \right), \quad (5)$$

while in general there is no closed form solution to problem (5), it is easily solved numerically.

This heuristic, of solving Equation (5) was investigated in Ref. 3, for cases in which the time to perishability is exponential or deterministic, arrivals follow a Poisson process, and customers' demand is either a unit or exponentially distributed. For these cases, Baron *et al.* [3] expressed the optimal order up to level, that is, the one that minimizes Equation (1) denoted by S^* , and compared it to the order up to the level based on solving Equation (5), S^H . For these cases, the cost error of the heuristic is relatively small

(around 2% on average). Thus, Baron *et al.* [3] suggested that this heuristic be used for cases where S^* cannot be found.

FUTURE RESEARCH DIRECTIONS

Below, we discuss several research directions and explain their importance. Many of these extensions have attracted significant research for standard items, but in our opinion, they are not addressed well enough for perishable items.

The most important research direction for managing inventory of perishables is to develop effective heuristics for inventory control of such items. Some heuristic control policies were developed earlier, such as by Nahmias [21], who discusses heuristics for the deterministic shelf-life case. However, developing effective heuristics for additional settings is still required. Such settings should include both continuous and periodic review systems and focus on cases with positive lead times. A closely related research direction is to provide theoretical guarantees on the performance of different heuristics, similar to the established efficiency of power of two policies for standard inventory items [22].

The standard assumption in the literature on continuously reviewed perishable items is that the planning horizon is infinite. Investigating the inventory control of perishables for finite horizons in such settings also deserves more research attention.

Four other issues that have garnered a lot of attention for standard inventory items are their management (i) in multiechelon settings, (ii) when aiming to coordinate the supply chain, (iii) for multiple items, and (iv) in the presence of competition (see **Business Process Outsourcing** multiechelon, multiple items and **Supply Chain Coordination**). An important feature in investigating these issues is internal transshipment. Many models addressing nonperishable items ignore internal shipments, because the cost of internal shipments might offset their benefit. However, because transshipments of perishable items may reduce the proportion of perished items, the value of internal transshipments for perishable items is higher

than for nonperishables. Thus, allowing internal transshipments is preferable in models that consider multiple sites.

Initial steps at investigating the above issues for supply chains of perishable items were taken (Section 4 in Ref. 2). For example, for the supply of blood in multiechelon settings (Ref. 23 and references within), consider both, a rotation (or recycling) policy and a retention one. In a rotation system at the end of each period (e.g., day) the supplier gathers all unsold (and not yet perished) items from the retailers and then supply them at the beginning of the next period; in a retention system, supplied items are left with the retailers until they are demanded or perished. However, the literature addressing issues (i)–(iv) for perishables typically considered fairly restrictive settings. Therefore, there is room for addressing these issues and developing effective management policies for managing the flow and allocation of perishable items in supply chains.

Note that to properly address these issues, investigation of the strategic planning of supply-chain networks is required for perishable items. Appropriate models for supply-chain networks require as input, the inventory control policy used and their performances (see *Supply Chain Coordination*). For example, locating warehouses and distribution centers should consider the costs implied by different supply-chain configurations. However, in the absence of efficient methods to manage the inventory within the supply chain and to estimate the implied inventory-related costs, designing the network of the supply chain might be of limited value.

Another model that is well established for standard items is joint manufacturing, storing, and allocation of products for different customer types. This is also relevant for perishable items. An example is the choice made at a central depot that serves several different retailers with different importance. Then, an allocation of items may need to consider performance measures as seen by the different customer types. Such measures of service levels could be on-stock availability of “fresh enough” items, fill rates, and so on. In fact, inventory control of perishable items

subject to service level constraints even for a single customer type has won little attention. Such constraints could be quite specific in defining appropriate service levels in relation to the residual life of items sold. Customers might not be satisfied with only a high on-shelf availability, but may also demand that available items are fresh enough. Moreover, in some cases different customer types prefer different levels of freshness that require different storage processes. For example, the shelf-life of tomatoes is longer if they are kept on the vine, refrigerated (that may also reduce their flavor), or turned into a tomato paste, tomato juice, or ketchup. Thus, it is likely that formulating problems with different customer types would be very application dependent. Still, in view of the importance of these applications, for example in the health-care industry, this is a topic deserving further research.

The above discussion of different service levels and storage requirements for perishable items raises another question; namely, the effective analysis of the logistics required for the delivery of perishable items. Logistics planning for perishable items involves capacity, storage, production, and transportation decisions as well as their effect on items’ shelf-life. For example, some fruits are delivered to North America in ships, with part of their ripening occurring during this long and often uncertain transportation window. Once ripened, such fruits are often stored at cold temperatures to increase their shelf-life. Thus, managing the logistics process required to bring valuable items to final customers is not simple, and deserves further attention. Again, such research is likely to be application dependent.

An important issue that is unique for perishable items is the dispatching policy which may be complex even for a single customer type. In addition to the complexity highlighted by Example 1, the dispatching policy can also affect the demand. For example, consumers of food items such as milk, often check the “best before” date of items on the shelf. Thus, demand for a specific item depends not only on its availability on the shelf and its age, but also on the age and shelf availability of items of different ages. That is, there

are substitutability effects among items that differ by age. In addition, in this setting, the controller can only affect rather than dictate the dispatching policy. This is a mirror problem to the models in the section titled “Models Without Order Setup Cost or Lead Times,” where the controller can only affect rather than dictate the arrival of new items. Again, initial works on this subject exist (Section 5.2 of Ref. 2), but there is still plenty of opportunity for research in more realistic settings.

As the last but not the least research direction, we list the following simplifying assumptions that are often made in works on perishable items. Most of the references above assume that (i) demand is known (in case of stochastic demand its distribution is known), (ii) there is no seasonality, (iii) demand is independent of pricing, shelf space allocation, and inventory level, (iv) order yield is perfect, and (v) there is no substitutability of items (both within items of different ages and among different products). For nonperishable items, all of these assumptions have been relaxed to some extent. However, the vast majority of papers on managing inventory of perishables make these assumptions. (Several initial attempts at relaxing some of these assumptions are referenced in Section 5 of Ref. 2.)

REFERENCES

1. Grocery manufacturers of America. 2006 Unsaleables benchmark report. 2006.
2. Karaesmen I, Scheller-Wolf A, Deniz B. Managing Perishable and Aging Inventories: Review and Future Research Directions. Kempf K, Keskinocak A, Uzsoy P, editors. Handbook of Production Planning. Kluwer Academic Publishers; 2009. To appear.
3. Baron O, Berman O, Perry D. Stochastic (τ, S) models for managing inventory of perishable items. Working paper. Rotman School of Management, University of Toronto; 2008.
4. Rajan A, Steniberg R. Dynamic pricing and ordering decisions by a monopolist. *Manag Sci* 1992;38:240–262.
5. Brown GW, Lu JY, Wolfson RJ. Dynamic modeling of inventories subject to obsolescence. *Manag Sci* 1964;11(1):51–63.
6. Nahmias S, Pierskalla W. Optimal ordering policies for a product that perishes in two periods subject to stochastic demand. *Nav Res Logist Q* 1973;20:207–229.
7. Broadheim E, Derman C, Prastacos GP. On the evaluation of a class of inventory policies for perishable products such as blood. *Manag Sci* 1975;22:1320–1325.
8. Nahmias S. A comparison of alternative approximations for ordering perishable inventory. *INFOR* 1975;13:175–184.
9. Nahmias S. The fixed charged perishable inventory problem. *Oper Res* 1978;26:441–481.
10. Lian Z, Liu L. A discrete-time model for perishable inventory systems. *Ann Oper Res* 1999;87:103–116.
11. Weiss H. Optimal ordering policies for continuous review perishable inventory models. *Oper Res* 1980;28:365–374.
12. Williams CL, Patuwo BE. A perishable inventory model with positive order lead times. *Euro J Oper Res* 1999;116:352–373.
13. Berk E, Gurler U. Analysis of the (Q, r) inventory model for perishables with positive lead times and lost sales. *Oper Res* 2008;56(5):1238–1246.
14. Graves S. The application of queuing theory to continuous perishable inventory system. *Manag Sci* 1982;28:400–406.
15. Nahmias S, Perry D, Stadje W. Actuarial valuation of perishable inventory systems. *Probab Eng Infor Sci* 2004;18:219–232.
16. Kaspi H, Perry D. Inventory system of perishable commodities. *Adv Appl Probab* 1983;15:674–685.
17. Pal M. The $(s-I, S)$ inventory model for deteriorating items with exponential lead time. *Calcutta Stat Assoc Bull* 1989;38:83–91.
18. Perry D, Posner M. An $(S-I, S)$ inventory system with fixed shelf-life and constant lead times. *Oper Res* 1998;46:65–71.
19. Tekin E, Gurler U, Berk E. Age based vs stock-level control policies for a perishable inventory system. *Eur J Oper Res* 2001;134:309–329.
20. Gurlur U, Ozkaya BY. Analysis of (s, S) policy for perishables with a random shelf-life. *IIE Trans* 2008;410:759–781.
21. Nahmias S. Perishable inventory theory: a review. *Oper Res* 1982;30(4):680–708.
22. Roundy R. A 98% effective integer ratio lot-sizing rule for a one warehouse multi-retailer, system. *Manag Sci* 1985;31:1416–1430.
23. Prastacos GP. Allocation of a perishable product inventory. *Oper Res* 1981;29:95–107

MANAGING PRODUCT INTRODUCTIONS AND TRANSITIONS

ÖZLEM BİLGİNER

FERYAL ERHUN

Department of Management Science and
Engineering, Stanford University,
Stanford, California

INTRODUCTION

Reducing time-to-market is commonly advocated as a source of competitive advantage. As a result, companies introduce new products more frequently than ever. For example, lifecycles are under a year for most electronic products [1]. In the semiconductor industry, a new processor is introduced every 18–24 months [2]. By launching new products faster than competitors, firms can rapidly respond to customer feedback, establish new standards, improve brand recognition, quickly introduce technical innovations, and capture a bigger share of the market, resulting in higher profits. Therefore, fast-paced product launches potentially bring tremendous value to firms. However, these launches also pose enormous challenges:

Companies that expand their product and service varieties now face a new set of problems: accurate demand forecasts, controlling production and inventory costs, and providing high quality delivery performance. (...) [I]ncreasing product lines can also decrease the efficiency of manufacturing processes and distribution systems. [3]

The period of product transitions is often the most vulnerable time for a company. Studies show that approximately 60% of new products fail in the marketplace [4]. Moreover, many of those products that make it in the market face rough transitions. As a result, product transitions can be costly due to excessive research and development (R&D) investment, marketing, and advertising costs of the product launch. In addition, there are

operational challenges that must be managed effectively, the failure of which may be detrimental. In the worst case, mismanaged product transitions may even lead to a company failure (e.g., Osborne Computers [1]).

This article serves as an introduction to product introductions and transitions, and as a survey of the literature on this important phase in a product's lifecycle. We define *product introduction* as the process of introducing a product, without considering the interactions with previous and future generations. *Product transition*, on the other hand, covers the introduction of the new product *and* the displacement of the previous generation [1]. In the next section, we provide a brief background of product introductions and transitions, and discuss supply and demand factors that need to be considered for a successful product launch. The dynamics of product launches depend heavily on how demand is generated. Of the various theories which explain demand generation, new product diffusion theory is among the most commonly used; this is discussed in the section titled "Diffusion Models." We then review the literature on product introductions in the following section and on product transitions in the section titled "Diffusion Models for Product Transitions." We conclude with potential research directions in the last section.

There is extensive literature on product feature selection and product line design, as well as on the product development structure of a company. Our scope in this article does not include these topics. Similarly, a stream of research in operations management literature considers pricing at the *end* of the product lifecycle; we do not consider the pricing of the old product unless it is in conjunction with the new product; hence, we omit a review of this stream as well.

PRODUCT INTRODUCTIONS AND TRANSITIONS

Launching new products frequently results in more product transitions (rollovers), and thus, situations where companies have to manage multiple product generations

simultaneously. Billington *et al.* [1] classify product rollover strategies as either *solo-product rolls* or *dual-product rolls*. In the solo-product roll strategy (Fig. 1a), at any given point of time, only one product is marketed; that is, the company aims to clear all of the old product inventory before introducing the new product. This strategy is simpler than the dual-product roll from the viewpoint of the company; however, planning has to be more precise because the old product has to sell out exactly at the time of the introduction of the new product in order to prevent a gap or an overlap. In the dual-product roll strategy (Fig. 1b), on the other hand, the old and new products overlap in the market for a period of time. As such, timing of the introduction is less of an issue; yet, products interact during the transition phase and managing this interaction can be challenging due to issues such as capacity management and cannibalization between products.

There are many supply and demand factors that companies should consider during product launches. We discuss several of these factors below:

- *Demand Forecasts.* Marketing indicators may have errors, especially at the time of a product launch. Companies may prepare themselves for a lower or a higher forecast than the original demand. As a result of poor sales forecasts, product launches may suffer from inadequate capacity decisions.

For example, after months of supply problems, Nintendo decided to increase the production of their popular console, Wii, by increasing production in one of their plants [5]. At the opposite end of the spectrum, the introduction of the Segway scooter suffered from overly optimistic sales forecasts. The Segway, which was released as a product that would revolutionize transportation, sold fewer than ten thousand units in the 2 years after its introduction. This was far short of the anticipated demand of 50,000 to 100,000 units [6], and the Segway failed to recoup its R&D expenses.

- *Pricing.* Pricing has a major impact on product launches. Eight months after the release of Intel's Pentium 4 processor, sales of the new product had not met expectations, in part, due to cannibalization by the Pentium III processor, which had a better value-price combination for consumers [2]. Product launches can be managed using price as a control mechanism by understanding the underlying dynamics of demand generation and by taking advantage of consumer heterogeneity. Proper pricing of a new product is a difficult balancing act. Higher prices will target only a minority of consumers (for whom the perceived value is higher than the price), may lead to slower adoption of the new product, and may result in excess inventory and capacity. Lower

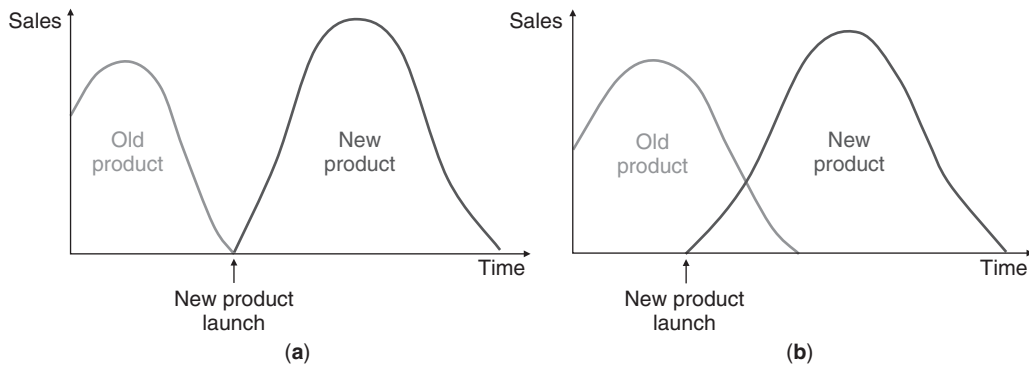


Figure 1. Product rollover strategies: (a) solo-product roll strategy and (b) dual-product roll strategy.

prices will increase the potential market and may lead to faster adoption, but may increase the possibility of running out of stock due to accelerated sales. The initial price of a new product also affects the price trajectory of the product over its lifecycle. During transitions, the price of a new product impacts the demand of both the new and existing products. Therefore, pricing the products correctly enables a company to control the sales of (and hence the revenue from) the products not only during product launches but also over the lifecycles of existing and future product generations.

- *Timing.* The introductory timing of a new product should be carefully chosen. If the new product is introduced later, a longer R&D cycle may result in a better product. However, the company may suffer due to late entry and lose market share to its competitors (e.g., the qualification process in the semiconductor industry that we describe in more detail in the section titled “Timing of New Product Introductions;” see our discussion on Ref. 7). On the other hand, if the product is introduced too early, the market is captured, but there may be technical risks due to product imperfections or a product which lacks desired attributes (e.g., the problems of Microsoft with its platform Windows Vista). Additional concerns should be considered: the time gap between generations should be optimized to balance the trade-offs between product cannibalization and the age of the product platform. An introduction done too soon may lead to excessive cannibalization of the old product, and an introduction done too late may cause customers to switch to other platforms, decreasing brand loyalty.
- *Capacity Decisions.* One critical decision prior to a new product launch is the production capacity decision. It is difficult to determine the optimal capacity to build at the time of a product launch due to supply and demand uncertainties. If the capacity is set too high, the firm will

have to pay for the idle capacity. If the capacity is set too low, establish a parallel structure with the previous sentence the capacity investment cost will be lower, but the firm will hold inventory to manage demand surges during the introduction (which increases the inventory holding cost) and/or will suffer backorders and lost sales (which may hamper the product’s acceptance in the market). These trade-offs should be considered when determining the optimal capacity to build. Capacity allocation among new and existing products further complicates the trade-offs in capacity decisions. Interactions between products (relative product capabilities, relative prices, relative demand, etc.) should also be carefully analyzed. In 1992, based on its initial forecasts, GM allocated half of the capacity of its Detroit-Hamtramck plant to the redesigned Seville, Cadillac, and Eldorado models and the remaining capacity to Buicks and Oldsmobiles. However, demand for Seville and Eldorados quickly exceeded supply, leading to the loss of thousands of potential customers. GM eventually allocated the whole plant to those two models, but by then “the damage had already been done” [8]. In this example, GM was able to reallocate capacity. However, capacity may not always be interchangeable, which limits the possibility of reallocation based on supply and demand dynamics.

In such a fast-moving environment, to stay competitive, companies are actively seeking new ways to make their supply chains more responsive. For example, companies may invest in better forecasting techniques (such as information sharing with supply chain partners (see *Information Sharing in Supply Chains*) or collaborative planning and forecasting) to increase their knowledge of the sources of uncertainty. Alternatively, companies may undertake various initiatives to shorten product development processes in order to be flexible in the face of uncertainty.

For example, clothing manufacturer Benetton employed the concept of postponement to be more responsive to market conditions, and Intel Corporation aims to stay one product generation ahead of its competitors to maintain industry leadership [2]. However, even with extensive planning, contingency strategies that enable adjustments to initial decisions may be necessary due to market dynamics and supply and demand uncertainties. Therefore, any strategy that addresses new product launches should include the ability to learn, as well as the ability to set transition decisions at the beginning (i.e., primary strategies) and to adjust these decisions later (i.e., contingency strategies). This will enable a firm to observe supply and demand signals, revise its understanding of the transitional markets based on these signals, update market knowledge accordingly, and thus react effectively to the dynamics and uncertainties of supply and demand.

Until this point, we have reviewed the problems concerning product launches and some current practices of facilitating these launches in industry. In the following sections, we will review the related academic literature. We will start with demand generation models, specifically diffusion models, as a basic demand forecasting tool (see the section titled “Forecasting Techniques” in this encyclopedia). We will then review pricing and timing decisions in new product introductions and transitions. For new product introductions, we will also provide a brief review on supply restrictions and capacity selection. To the best of our knowledge, there are no equivalent studies of new product transitions.

DIFFUSION MODELS

One of the basic models that govern demand generation is diffusion. Rogers defines the diffusion of an innovation as “the process by which an innovation is communicated among the members of a social system” [9]. Thus, diffusion models focus on “communication channels, which are the means by which information about an innovation is transmitted to or within the social system” [10]. There are several diffusion models used in the literature; we provide a brief introduction to

the Bass model [11], which is one of the most widely used models in marketing.

The Bass diffusion model is a simple “epidemic” model, which represents the probability of a purchase at t given that no purchase has yet been made as a linear function of previous purchases [11]. To express the model mathematically, let $f(t)$ be the likelihood of adoption at time t , $F(t)$ the cumulative fraction of adopters at time t , p the coefficient of innovation, and q the coefficient of imitation. Then, the likelihood of purchase at time t given that no purchase has yet been made can be expressed as

$$\frac{f(t)}{1 - F(t)} = p + qF(t), \quad (1)$$

where $f(t) = dF(t)/dt$. The Bass model assumes that potential adopters of an innovation consist of two groups: the first term on the right-hand side of Equation (1) captures *innovators* and the second term captures *imitators*. Innovators are influenced *externally*, that is, by mass-media communication, and their purchases are independent of the previous purchasers (i.e., the first term is independent of $F(t)$) while imitators are influenced *internally*, that is, by word-of-mouth through previous purchasers (i.e., the second term is a function of $F(t)$).

Defining m as the market potential, $S(t) := mF(t)$ as the cumulative sales at time t , and $s(t) := dS(t)/dt = mf(t)$ as the instantaneous sales at time t , we can express the Bass equation in terms of sales as follows:

$$\begin{aligned} \frac{s(t)}{m - S(t)} &= p + \frac{q}{m}S(t) \Rightarrow \\ s(t) &= \left[p + \frac{q}{m}S(t) \right] (m - S(t)). \end{aligned} \quad (2)$$

Equation (2) is a differential equation that can be solved to provide $S(t)$ in closed form. Figure 2 displays the adoption of a new product as per the Bass model. Note that the Bass diffusion model assumes “first-purchase,” that is, there are no repeat buyers and purchase volume per buyer is one unit. Furthermore, the model assumes that the market potential m and diffusion coefficients p and q are constant over time, and that the adoption does not depend

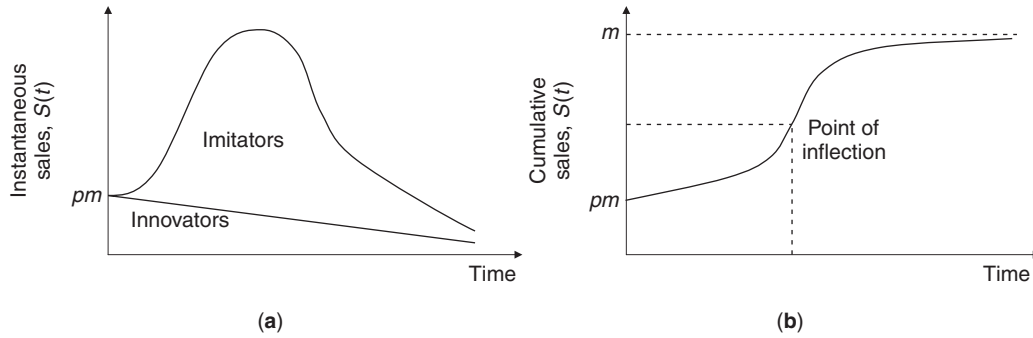


Figure 2. Bass new product diffusion model: (a) instantaneous sales and (b) cumulative sales. [Source: Ref. 10]

on any marketing-mix variables like price and advertisement. The formulation in Equations (1) and (2) is for a single product lifecycle with infinite product availability and considers neither the phenomenon of diffusion in overlapping product lifecycles nor the concept of supply constraints.

Due to the limitations mentioned above, the Bass model can provide limited normative guidance. However, the Bass model and its various extensions have proven to be quite robust in practice. According to Bass, “the most important use of the model for forecasting purposes is the forecast made prior to product launch when there are no sales data upon which to base the forecast” [12]. The estimates of the Bass model parameters for different products and services have been developed using large databases for different industries. These estimates enable analysis of empirical studies using the Bass model; see Table 1 for a sample based on the work of Lilien *et al.* [13]. Lilien *et al.* report a low innovation coefficient $p = 0.000$ and a high imitation coefficient $q = 0.534$ for ultrasound imaging, which means that this product was slow to penetrate the market at first, but as the word-of-mouth effect increased, market penetration sped up. On the other hand, camcorders have a relatively higher $p = 0.044$ and a relatively lower $q = 0.304$; they penetrated the market fast and then over time the penetration slowed down. Actually, ultrasound imaging devices penetrated to only 20% of the market in the first 10 years after their introduction, while camcorders penetrated to 75%. In the next 10 years, both markets were almost fully penetrated.

MODELS FOR PRODUCT INTRODUCTIONS

In this section and the next, we provide a brief review of product introductions and product transitions in marketing and operations management literature. Our review is not comprehensive; instead, our goal is to provide a general understanding of this vast literature. In surveying the literature, we do not limit ourselves to diffusion models; in fact, some of the included models have general time-dependent demand functions.

Pricing of New Product Introductions

As mentioned earlier, one of the major drawbacks of the Bass model is its exclusion of price. To remedy this drawback, Robinson and Lakhani [14] modify Equation (2) as follows:

$$s(t) = \left[p + \frac{q}{m} S(t) \right] (m - S(t)) e^{-kPr(t)}, \quad (3)$$

where $Pr(t)$ is the price at time t and k is a price coefficient. According to Robinson and Lakhani, product price affects the likelihood of product adoption at any given time t but not the potential market size m . That is, only the adoption probability $f(t)$ is impacted by price. Using Equation (3), the authors study the optimal price sequence of a new product over the product's lifecycle. Their numerical results (obtained through simulation) indicate that, with a positive discount rate and experience-curve-based costs, cost-plus pricing performs poorly for a “rapidly evolving business.” The optimal price scheme should

Table 1. Sample p and q Values for Different Products

Product	Period of Analysis	p	q
<i>Electrical appliances</i>			
Blender	1949–1979	0.000	0.260
Can opener	1961–1979	0.050	0.126
Clothes dryer	1950–1979	0.009	0.143
Clothes washer	1923–1971	0.016	0.049
Dishwasher	1949–1979	0.000	0.213
Electric coffee maker	1955–1979	0.042	0.103
Freezer	1949–1979	0.019	0.000
Microwave oven	1972–1990	0.002	0.357
Refrigerator	1926–1979	0.025	0.126
<i>Medical equipment</i>			
CT scanners (50–99 beds)	1980–1993	0.044	0.350
CT scanners (>100 beds)	1974–1993	0.036	0.268
Ultrasound imaging	1965–1978	0.000	0.534
<i>Consumer electronics</i>			
Cable television	1981–1994	0.100	0.060
Camcorder	1986–1996	0.044	0.304
CD player	1986–1996	0.055	0.378
Cellular telephone	1986–1996	0.008	0.421
Cordless telephone	1984–1996	0.004	0.338
Home PC	1982–1988	0.121	0.281
Telephone answering device	1984–1996	0.025	0.406
Television, black and white	1949–1979	0.108	0.231
Television, color	1965–1979	0.059	0.146

[Source: Ref. 13].

be dynamic and aggressive: the company should be ready to use penetration pricing (i.e., setting a relatively low initial price to penetrate the market, then increasing the price over time and decreasing it toward the end of the product lifecycle) and to potentially incur losses during the initial phases of the product launch. Robinson and Lakhani's work initiated a stream of research which aims to identify optimal price paths under different settings; we refer the interested reader to Section 4 of Ref. 15 for a brief review.

The impact of competition has also been analyzed in the pricing of new product introductions. Clarke and Dolan [16] study different pricing strategies in a duopoly. The authors compare myopic pricing (i.e., setting prices to maximize only the current period's profit), skimming pricing (i.e., setting relatively higher prices at the time of introduction and lowering them over time), and penetration pricing via simulation; they show that myopic pricing may yield highly suboptimal profits. Dockner and

Jorgensen [17] incorporate competition into the Bass model and analyze the optimal pricing strategies of competing firms where production costs decline with learning. In their model, the sales function depends on cumulative sales (as in the Bass model) and on prices of all competing firms. The authors also consider special cases of the demand function and determine the conditions under which the optimal price path is decreasing or increasing. Eliashberg and Steinberg [18] analyze competition between two firms with asymmetric cost structures and production modes. The first firm has a convex production cost and can hold inventory, whereas the second firm has a linear cost and cannot hold inventory. The firms compete on prices, production rates and, in the case of the first firm, inventory levels. The authors show that the first firm holds a positive amount of inventory at the beginning of the period, and stops holding inventory after a while. Prices first increase and then decrease.

Timing of New Product Introductions

In timing new product introductions, the main trade-off is between product quality, price, and time-to-market. Kalish and Lilien [19] are among the first to study this trade-off. The authors modify the Bass diffusion model to incorporate the effect of prices on market size as well as advertising-dependent diffusion parameters to determine the best time to enter a market. A premature entry results in a negative word-of-mouth effect due to a low quality product. On the other hand, being late to market results in lost sales. Kalish and Lilien do not provide general optimization results; instead, they provide a solution for a specific application where the objective is to maximize installed bases; that is, the number of current users of the products.

Bayus [20] focuses on the effects of competition on the timing problem and considers two scenarios. In the first scenario, the competitor is already in the market. Through numerical analysis, the author shows that it is optimal to accelerate product development and launch a low-quality product if the competitor's quality level is low, the time horizon is short, and product development costs are low. It is optimal to accelerate product development and launch a high quality product if sales are high and product development costs are low. In the second scenario, the competitor is also developing the product and the question is whether to be the first to market or not. The author numerically shows that it is optimal to accelerate product development and launch a high-quality product if the time horizon is long, sales are high, and product development costs are low. It is never optimal to be the first to the market with a low quality product in this case.

Recently, Özer and Uncu [7] have analyzed a similar problem. In their framework, a supplier decides when to finish design and launch his product to apply for a seller's product qualification. The problem has two stages. In the first stage, the supplier applies for a qualification. There are some trade-offs that the supplier faces. Entering early results in larger market share, higher production costs, and lower chances of getting qualified. Entering late results in lower market share, lower production cost, and higher

chances of getting qualified because waiting improves the manufacturing processes. If the supplier becomes *qualified*, then in the second stage, he determines the production quantities. Both, the qualification process and the demand are assumed to be uncertain. Even if the supplier decides to apply for qualification, he may not be qualified by the seller and, in turn, will not be able to "enter" the market. Using a dynamic programming approach (see articles in the section titled "Fundamentals" in this encyclopedia, for a review of the fundamentals of dynamic programming), the authors show that for the first stage, a threshold policy is optimal: the supplier decides to apply for qualification if his rank among the other suppliers is above some threshold. The authors also prove that, in the second stage, the optimal production policy is a state-dependent base-stock policy.

Supply Restrictions and Capacity Selection during New Product Introductions

Another drawback of the Bass model is that it does not model supply restrictions. Jain *et al.* [21] attempt to fill this gap by modifying Equation (2) as

$$\begin{aligned}\frac{dA(t)}{dt} &= \left(p + \frac{q_1}{m}A(t) + \frac{q_2}{m}N(t) \right) \\ &\quad (m - A(t) - N(t)) - c(t)A(t), \\ \frac{dN(t)}{dt} &= c(t)A(t),\end{aligned}$$

where $N(t)$ is the cumulative sales as in the Bass model, $A(t)$ is the number of waiting adopters, and $c(t)$ is the supply rate. The authors apply their model to data from the diffusion of new telephones in Israel, and find that the model is a good fit.

With similar models, Ho *et al.* [22] and Kumar and Swaminathan [23] study the supply-restricted diffusion problem to characterize the optimal time-to-market and capacity decisions of a firm. Ho *et al.* find that a myopic sales plan is optimal; that is, the firm should not delay demand fulfillment. The authors show that the firm may substitute capacity with inventory by delaying the product launch and building up inventory during that period. Kumar and Swaminathan, on the other hand, argue

that a myopic sales plan is not necessarily optimal. Via numerical analysis, the authors show that a build-up type heuristic, where the firm does not sell at the beginning but builds inventory instead, is optimal when unmet demand is lost and performs well when demand is backordered. It is surprising that these two papers derive such different results given that their models are quite similar. Kumar and Swaminathan argue that “although the models considered in Ho *et al.* (2002) and [our model] are different, it is unlikely that the differences in the models alone account for the contrasting results. Resolving this issue is the subject of future work.”

DIFFUSION MODELS FOR PRODUCT TRANSITIONS

The Bass model does not incorporate the interactions among existing products. Norton and Bass [24] relax the assumption of a single product and study successive generations of innovation using a cumulative sales growth model that accounts for cannibalization/substitution between two product generations. If $S_1(t)$ and $S_2(t)$ are the installed bases of the first- and second-generation products at time t , then

$$\begin{aligned} S_1(t) &= m_1 F_1(t) - m_1 F_1(t) F_2(t), \\ S_2(t) &= m_2 F_2(t) + m_1 F_1(t) F_2(t), \end{aligned}$$

where m_1 is the market size of the first-generation product (old product) and m_2 is the incremental market that exists only for the new product and cannot be satisfied by the old product. The fractions of adoption for each generation, $F_1(t)$ and $F_2(t)$, are given by the Bass model solution, and $F_2(t) = 0$ for all t before the introduction time of the second-generation product. The term $m_1 F_1(t) F_2(t)$ represents the cannibalization (or substitution) effect. The second-generation product will cannibalize the sales of the first-generation product and, over time, the installed base of the first-generation product will fall to zero. As such, this model formulation is more applicable to the repeat purchase of technology generations such as the transition from black-and-white to color

televisions, where all customers eventually move to the new product.

In the following sections, we review models that provide some normative guidance on pricing and timing issues during product transitions.

Pricing of New Product Transitions

The paper by Norton and Bass [24] does not account for marketing-mix variables such as prices. Padmanabhan and Bass [25] fill this gap by deriving the optimal pricing policy of a monopolist marketing successive generations of a product. Their model assumes that price affects sales at any instant but not the market size. The authors present various conditions when the optimal pricing of old and new generations would monotonically increase or decrease. They also present the optimal pricing policy when two generations are launched by independent producers. Padmanabhan and Bass’s results demonstrate that the integrated producer offers a lower price for the first-generation product than the independent producer and that both producers charge the same price for the second-generation product.

Lim and Tang [26] model product rollover strategies to determine the optimal entry time of the new product and the optimal pricing of the new and old products. The authors consider both solo- and dual-product rollover strategies, two-way cannibalization, and a linear price-demand relationship. As a consequence of this linearity, they are able to find the optimal decision variables in closed form, and they provide various insights on the dependency of optimal prices on parameters. They also determine the optimal time to introduce the new product and the optimal time to phase out the old product. Finally, the authors compare the two rollover strategies and provide conditions under which solo-product rollover is preferable to dual-product rollover.

Padmanabhan and Bass [25] and Lim and Tang [26] both assume that demand is known and that the firm does not hold any inventory. Li and Graves [27] relax these assumptions and study the optimal pricing decisions of two product generations under inventory considerations and uncertain demand. They restrict

the new product price to be constant and assume that the old product price can change dynamically. Due to the inherent complexities of changing prices dynamically, they propose a fixed-price policy and a heuristic to approximate that policy. The authors show that with excessive amounts of inventory, the performance of the heuristic approaches the optimal fixed-price policy.

Recently, Bilginer and Erhun [28] study the pricing problem where sales are represented as

$$\begin{aligned}s_1(t) &= m_1 f_1(t) - m_1 f_1(t) \gamma, \\ s_2(t) &= m_2 f_2(t) + m_1 f_1(t) \gamma.\end{aligned}$$

In the above equations, m_i is a price-dependent market size function for product generation $i = 1, 2$, $f_i(t)$ is the sales probability (this is a generalization of the Bass model $f(t)$), and γ is the price-dependent cannibalization coefficient. This coefficient is given by $\gamma = Be^{-\frac{Pr_2}{Pr_1}}$ where B is a number between 0 and 1 which represents the compatibility of the successive generations and Pr_i is the price of product generation $i = 1, 2$. The authors provide bounds on the optimal prices, and also propose two simple heuristics to find those prices. The first heuristic sets the prices ignoring cannibalization and the second one sets the old product's price assuming no cannibalization but optimizes the new product's price based on the old product's price. The authors show that both heuristics perform well. The second heuristic's maximum deviation from the optimal prices is only 4.75% and the average deviation is around 1% for the cases considered. Bilginer and Erhun further extend their model to the case where products are marketed by independent firms to analyze the effects of competition on prices. When there is no cannibalization, competition does not affect the prices. If cannibalization is price-independent, the old product's price increases under competition, while the new product's price may increase or decrease based on the parameters. If cannibalization is price-dependent, their numerical analysis shows that, in general, both prices fall with competition.

Timing of New Product Transitions

Determining the best time to introduce a new product to market is an important decision to make during product transitions. One of the earliest papers that model transition timing is Wilson and Norton [29]. In this paper, the authors study the optimal entry timing for a product line extension using a diffusion framework. In their model, the existing product has a higher revenue margin than the new product and the new product expands the market; nevertheless, the two products cannibalize each other. The authors show that, most of the time, it is optimal to either introduce the new product as early as possible or to not introduce it at all, depending on the demand parameters, market size, and relative product margin. They also show that in the remaining cases, it is optimal to introduce earlier rather than later. Mahajan and Muller [30] extend Wilson and Norton's model by allowing the marketing of more than two products at once and using general profit margins. Their results are very similar to Wilson and Norton's: it is optimal to either introduce a new product right away or wait until the old product matures. They apply their model to data obtained for IBM mainframes, and conclude that the time-to-market of two product families was too long.

Another stream of literature on new product launch timing treats the evolution of the current technology in the market as stochastic (e.g., the time between technology improvements and/or the amount of technology improvement may be unknown); see *Stochastic Dynamic Programming Models and Applications* for an introduction to stochastic models. Consider a firm that develops products based on available technologies and that must schedule its own product offerings based on its knowledge of a potential technology roadmap. Given the technology roadmap, the firm must determine when to replace its existing product and which technology to adopt. Balcer and Lippman [31] model the time between technology improvements as well as the amount of each improvement as stochastic. In order to adopt the latest technology, the firm needs to incur a fixed cost in addition to variable costs.

The firm's revenue is proportional to the technology level of the product. The authors show that a threshold policy is optimal; that is, it is optimal to adopt the latest technology if the technology gap (the difference between the latest technology in the market and the current technology of the product) exceeds a threshold. Paulson Gjerde *et al.* [32] extend Balcer and Lippman's model by assuming that the technology of the product is represented by a vector. The decision variables of the firm are the amount of innovation to invest in each of these dimensions. With a focus on when should firms make incremental versus frontier innovation, the authors characterize the conditions under which it is optimal to innovate to the level of the latest technology. Krankel *et al.* [33] extend these two models by assuming that the demand is diffusion type. (This paper also extends [29,30] by assuming a stochastic technology improvement process.) Using a dynamic programming formulation, the authors show that the optimal policy is a threshold policy, in which the threshold depends on the cumulative sales of the firm and the technological R&D level of the new product. Recently, Liu and Özer [34] have shown that under a faster technological pace, firms' expected total profits decrease unless their fixed costs are sufficiently low. Thus, the authors conclude that firms should try to lower their fixed costs of product development in order to stay profitable.

Several studies in the literature analyze the factors driving time between transitions. Souza *et al.* [35] study how factors such as "clockspeed" (the industry's speed of new product introductions), product development cost, production cost, inventory cost, and competition affect a firm's optimal new product introduction time and product quality decisions. They model this problem as an infinite-horizon Markov decision process (see the section titled "Introduction to MDPs" in this encyclopedia for an introduction to Markov decision processes) where the uncertainty derives from the product demand. Via an extensive numerical study and statistical analysis, the authors show that the optimal pace of new product introductions is mainly influenced by industry clockspeed. On the

other hand, the optimal product quality primarily depends on internal factors, such as the relative costs of introducing new products with incremental versus more substantial improvements. Similarly, Druehl *et al.* [36] analyze the factors that affect the time between product transitions. They assume that revenue margins decline exponentially over time and that the product development cost is U-shaped. The authors use the Norton and Bass [24] demand model, so the new product expands the market, but cannibalizes the old product's demand. Through extensive numerical analysis, Druehl *et al.* statistically show that the higher the diffusion rate, the higher the pace of product introductions. The authors also show that product development cost, market growth rate resulting from the new product, and the rate of margin decline highly impact the new product development pace.

CONCLUSION

Companies are under great pressure to serve their customers better, faster, and cheaper. This fast-paced environment requires agile and efficient supply chains that adapt well to changes in product characteristics and to an ever-changing market environment. Companies are launching new products more frequently and the pressures for success are greater than ever. In this article we reviewed the tools developed in industry and in academic literature to mitigate the problems encountered during the product introduction and transition phases of companies. We propose several directions for future research:

- With the exception of a few papers, the majority of the literature on new product introductions and transitions is focused on the organizational structure of the firm. There are many opportunities for the application of operations research and management techniques and tools. Inventory management, capacity planning and allocation during the new product introduction and transitions, and supply chain design for the new product are some of the unexplored areas of research.

- As we discussed in the first two sections, there are many supply and demand uncertainties burdening product launches. Surprisingly, almost all of the models in the literature ignore these uncertainties. There is a big opportunity to study these uncertainties systematically. The literature would also benefit from frameworks (similar to Refs 1 or 37) that combine many different supply and demand aspects of product launches to provide guidelines for successful product introductions and transitions.
- There is a need for better understanding of product transitions. Compared to the literature on product introductions, the literature on product transitions is still limited. For example, we were unable to identify any papers that analyze capacity allocation among multiple generations. Similarly, we did not encounter any work on using marketing-mix variables to alleviate capacity problems. Supply constraints in product transitions is also an open area. A better understanding of product diffusions and their extensions for product transitions would be valuable.
- A contractual partnership between tiers of a supply chain may alleviate some of the problems during the product introduction and transitions. To the best of our knowledge, the only paper that incorporates the interactions among multiple tiers of a supply chain into their framework is Li and Gao [38]. Li and Gao consider a periodic-review inventory system consisting of a manufacturer and a retailer, where the manufacturer introduces new products using a solo-roll strategy. The authors consider two cases, one with information sharing between the parties and one without, and devise the form of optimal contracts between the manufacturer and the retailer. They show that information sharing is beneficial for both parties when the inventory system is coordinated; however, this may not hold if the inventory system is not coordinated.

A deeper look at the contractual agreements and the interactions among supply chain partners during new product introductions and transitions is a potentially fruitful research direction.

- The Bass diffusion model easily becomes complicated when extensions are considered. This makes analysis quite tedious and many times, numerical solutions are the only viable outcomes. In this sense, future research should seek simple yet elegant ways to extend existing models in order to move forward and study the issues we have raised above. Practice calls for a better understanding of product transitions, especially under various demand and supply uncertainties. Demand generation models that enable such expansions would be especially welcome.

REFERENCES

1. Billington C, Lee HL, Tang CS. Successful strategies for product rollovers. *Sloan Manag Rev* 1998;39(3):23–30.
2. Erhun F, Hopman J, Lee HL, *et al.* Intel Corporation: product transitions and demand generation. Stanford Global Supply Chain Management Forum Case, Case No: GS-43 and SGSCMF-004-2204. Mar 2 2005a.
3. Ho T-H, Tang CS. Product variety management: research advances. Norwell (MA): Kluwer Academic Publishers; 1998.
4. Lynn GS, Reilly RR. Blockbusters: the five keys to developing great new products. Harper Business New York (NY): 2002: 237.
5. GameSpot. Nintendo: Wii have a supply problem. CNet News.com, May 7 2007.
6. Kawamoto D. Segway sales haven't transported maker. CNet News.com, Sep 29 2003.
7. Özer Ö, Uncu O. An integrated framework to optimize stochastic dynamic qualification timing and production decisions. Working paper. New York (NY): Columbia University; 2008.
8. Fisher ML, Hammond JH, Obermeyer WR, *et al.* Making supply meet demand in an uncertain world. *Harv Bus Rev* 1994;72(3): 83–93.

9. Rogers EM. Diffusion of innovations. 5th ed. New York (NY): The Free Press; 2003.
10. Mahajan V, Muller E, Bass FM. New product diffusion models in marketing: A review and directions for research. *J Mark* 1990;54(1): 1–26.
11. Bass FM. A new product growth for model consumer durables. *Manag Sci* 1969;15(1): 215–227.
12. Bass FM. Comments on “A new product growth for model consumer durables”: the Bass model. *Manag Sci* 2004;50(12): 1833–1840.
13. Lilien GL, Rangaswamy A, Van den Bulte C. Diffusion models: managerial applications and software. In: Mahajan V, Muller E, Wind Y, editors. New-product diffusion models. New York (NY): Springer; 2000. Chapter 12. pp. 295–336.
14. Robinson B, Lakhani C. Dynamic price models for new-product planning. *Manag Sci* 1975;21(10):1113–1122.
15. Rao VR. Pricing models in marketing. In: Eliashberg J, Lilien GL, editors. Volume 5, Handbooks in OR & MS, Marketing. Amsterdam, The Netherlands: Elsevier Science Publishers; 1993. Chapter 11. pp. 517–552.
16. Clarke DG, Dolan RJ. A simulation analysis of alternative pricing strategies for dynamic environments. *J Bus* 1984;57(1): S179–S200.
17. Dockner E, Jorgensen S. Optimal pricing strategies for new products in dynamic oligopolies. *Mark Sci* 1988;7(4):315–334.
18. Eliashberg J, Steinberg R. Competitive strategies for two firms with asymmetric production cost structures. *Manag Sci* 1991;37(11): 1452–1473.
19. Kalish S, Lilien GL. A market entry timing model for new technologies. *Manag Sci* 1986;32(2):194–205.
20. Bayus BL. Speed-to-market and new product performance trade-offs. *J Prod Innov Manag* 1997;14(6):485–497.
21. Jain D, Mahajan V, Muller E. Innovation diffusion in the presence of supply restrictions. *Mark Sci* 1991;10(1):83–90.
22. Ho T-H, Savin S, Terwiesch C. Managing demand and sales dynamics in new product diffusion under supply constraint. *Manag Sci* 2002;48(2):187–206.
23. Kumar S, Swaminathan JM. Diffusion of innovations under supply constraints. *Oper Res* 2003;51(6):866–879.
24. Norton JA, Bass FM. A diffusion theory model of adoption and substitution for successive generations of high-technology products. *Manag Sci* 1987;33(9):1069–1086.
25. Padmanabhan V, Bass FM. Optimal pricing of successive generations of product advances. *Int J Res Mark* 1993;10(2):185–207.
26. Lim WS, Tang CS. Optimal product rollover strategies. *Eur J Oper Res* 2006;174(2): 905–922.
27. Li H, Graves SC. Pricing decisions during inter-generational product transition. Working paper. Tempe (AZ) and Cambridge (MA): Arizona State University and Massachusetts Institute of Technology; 2007.
28. Bilginer O, Erhun F. Pricing product transitions. Working paper. Stanford (CA): Stanford University; 2009.
29. Wilson LO, Norton JA. Optimal entry timing for a product line extension. *Mark Sci* 1989;8(1):1–17.
30. Mahajan V, Muller E. Timing, diffusion, and substitution of successive generations of technological innovations: the IBM main-frame case. *Technol Forecast Soc Change* 1996;51(2):109–132.
31. Balcer Y, Lippman SA. Technological expectations and adoption of improved technology. *J Econ Theor* 1984;34(2):292–318.
32. Paulson Gjerde KA, Slotnick SA, Sobel MJ. New product innovation with multiple features and technology constraints. *Manag Sci* 2002;48(10):1268–1284.
33. Krankel RM, Duenyas I, Kapuscinski R. Timing successive product introductions with demand diffusion and stochastic technology improvement. *M&SOM* 2006;8(2):119–135.
34. Liu H, Özer Ö. Managing a product family under stochastic technological changes. *Int J Prod Econ* 2009;122(2):567–580.
35. Souza GC, Bayus BL, Wagner HM. New-product strategy and industry clockspeed. *Manag Sci* 2004;50(4):537–549.
36. Druehl CT, Schmidt GM, Souza GC. The optimal pace of product updates. *Eur J Oper Res* 2009;192(2):621–633.
37. Erhun F, Gonçalves P, Hopman J. The art of managing new product transitions. *MIT Sloan Manag Rev* 2007;48(3):73–80.
38. Li Z, Gao L. The effects of sharing upstream information on product rollover. *Prod Oper Manag* 2008;17(5):522–531.

MANAGING R&D AND RISKY PROJECTS

LOUIS ANTHONY COX, JR.
Cox Associates and Department
of Mathematical and
Statistical Sciences,
University of Colorado,
Denver, Colorado
Department of Risk Analysis,
University of Colorado,
Denver, Colorado

INTRODUCTION

Research and development (R&D) projects are inherently risky. Investors in R&D projects are often uncertain about the probability of eventual technical success, the costs and time needed to succeed, competitor activities, and market acceptance and financial rewards if a technical success occurs. Quantitative risk models can help to characterize and manage R&D risks and to answer such practical risk management decision questions as the following:

- *R&D Project Scheduling and Management.* How to order the activities within a project, decision rules for when to abandon a project.
- *R&D Project Selection.* Which projects to fund.
- *R&D Investment Strategy.* How intensely or quickly to pursue R&D.
- *R&D and Intellectual Property.* How to manage the results of R&D (e.g., through licensing and patent systems) to create incentives that will benefit society.

This article surveys methods for quantifying and managing R&D risks.

MINIMIZING R&D PROJECT RISKS BY OPTIMALLY ORDERING ACTIVITIES

A very simple model of an R&D project divides it into multiple activities, each having a known cost for attempting it and a known probability that it can be successfully completed if it is attempted. An R&D manager may wish to minimize the expected cost to either successfully complete the whole project or to learn (by trying activities and learning that they fail) that it cannot be completed. The problem of sequencing the project activities to achieve this goal can be solved easily in certain cases.

Example: Optimal Sequencing of Activities in a Risky Project.

Setting: Suppose that an R&D project can be completed successfully only by successfully completing both of two activities, A and B. The cost to attempt activity A is $c_A = 1$ cost unit; the cost to attempt B is $c_B = 2$ cost units; and the conditional probabilities of successfully completing these activities, if they are attempted, are $p_A = 0.8$ and $p_B = 0.5$, respectively. The two activities may be attempted in either order. The value of successfully completing the project (i.e., of successfully completing both A and B) is $V = 10$ benefit units (on a scale commensurable with the cost units).

Problem: (i) What investment strategy maximizes the expected value of investment in this set of R&D activities? (ii) What is the expected value of this R&D project?

Solution: (i) Attempting A first, and then B if A succeeds, yields an expected net value of

$$\begin{aligned} & (\text{cost to attempt A}) + (\text{probability A succeeds}) \\ & \times (\text{cost to attempt B after A succeeds} \\ & + \text{probability that B succeeds} \\ & \times \text{value obtained if B and A succeed}) \end{aligned}$$

$$\begin{aligned}
&= -c_A + p_A \times (-c_B + p_B V) \\
&= -1 + 0.8 \times (-2 + 0.5 \times 10) = 1.4 \text{ units.}
\end{aligned}$$

Symmetrically, attempting B first, and then A if B succeeds, yields an expected net value of

$$\begin{aligned}
&-c_B + p_B \times (-c_A + p_A V) \\
&= -2 + 0.5 \times (-1 + 0.8 \times 10) = 1.5 \text{ units.}
\end{aligned}$$

Since 1.5 is the higher value (and is positive), the optimal strategy is to attempt B before A, and the expected monetary value of the R&D project (if the optimal strategy is used) is 1.5.

More generally, trying A before B yields a higher expected value than trying B before A if and only if

$$\begin{aligned}
&-c_A + p_A(-c_B + p_B \times V) > -c_B \\
&+ p_B(-c_A + p_A V), \text{ or, equivalently, if} \\
&c_B(1 - p_A) > c_A(1 - p_B).
\end{aligned}$$

That is, for a project with positive expected value, *activities should be attempted in order of decreasing failure-probability-per-unit-cost ratio* (i.e., in order of decreasing $c_i/(1 - p_i)$, where $i = A$ or B in this example); this decision rule holds for any number of activities. (Projects with negative expected value should not be attempted at all, if the goal is to maximize expected value.)

The preceding example is especially simple because completing the project required all of the activities in it to be completed. The only decision was the order in which to attempt them. More generally, there may be several alternative ways to successfully complete an R&D project. A subset of activities is called a *path set* if completing all of the activities in it suffices to successfully complete the project. A *minimal path set* is one in which no proper subset of activities is a path set. The project succeeds if and only if all of the activities in at least one minimal path set are successfully completed. If minimal path sets are disjoint, then an optimal (expected value-maximizing) investment strategy is

to attempt alternative minimal path sets in decreasing order of success-probability-per-unit-expected cost (where the success probability for a minimal path set is the product of the success probabilities of the activities in it and the expected cost of the minimal path set is calculated assuming that activities *within* each minimal path set are sequenced in decreasing order of failure-probability-per-unit-cost ratio, as in the preceding example). Investment in a project should cease as soon as no minimal path sets with positive expected value remain.

These ideas can be extended to projects in which activities cannot be attempted in any arbitrary order. Suppose that precedence constraints among activities restrict the order in which they may be attempted. It may be necessary to complete some before others can be attempted. Suppose that these constraints are represented by some acyclic graph, where each edge represents an activity and each activity can be attempted if and only if all of its predecessors have been successfully completed. The project succeeds and its benefit is obtained if and only if at least one path is successfully completed from a starting activity (one with no predecessors) to an ending activity (one with no successors). In this quite general framework, the optimal (expected utility-maximizing) investment strategy for a decision maker with linear or exponential utility can still be found by assigning to each activity a number called its *index* (by generalizing the failure-probability-per-unit-cost ratio in the preceding example) that can be computed quickly from the cost to attempt, success probability, and precedence constraint data. The optimal decision rule is to *attempt activities in order of decreasing index until either success is achieved or no activities with positive indices remain* [1]. Many other extensions of index policies (also called *Gittins index* policies) have been developed for situations with discount rates, randomly evolving opportunity sets, and relatively sophisticated models of risky projects, such as Markov decision processes and multiarm bandit decision problems [2,3].

OPTIMIZING R&D PROJECT PORTFOLIOS

Optimally sequencing the activities within an R&D project can maximize the expected utility of investing in it, by minimizing the expected time or cost to either complete it, obtaining its benefits, or to discover that this cannot be done, if such is the case. But even after such within-project optimization, the problem of choosing a subset of available R&D projects to fund—often from many proposals—remains. Quantitative approaches to optimizing investment in proposed risky projects to maximize resulting net benefits over time include the following:

- *Combinatorial Optimization Models.* These models search for combinations of select and reject decisions to maximize the net value of the selected portfolio of projects, given budget constraints and possible interactions and dependencies among the projects. Large-scale nonlinear 0–1 programming and mixed integer programming models have been developed for this purpose and are practical with current optimization technology and computational resources [4,5].
- *Improving an Existing Portfolio.* Given any proposed portfolio of risky projects, optimization algorithms can seek to build a modified portfolio with preferred risk-return characteristics by adding or deleting projects and modeling how such additions and deletions shift the predicted (simulated) probability distribution of net return from the entire project portfolio, taking into account interactions among projects [6].
- *Continuous Dynamic Optimization.* In continuous optimization models, explicit optimal investment formulas can be developed and interpreted as equating appropriately defined marginal returns across investments in different projects. This leads to dynamic resource allocation strategies for projects that may have statistically dependent, dynamically evolving market payoffs, assuming that partial investment in a project can generate partial returns [7].

In general, optimization of R&D project portfolios when there are strong interactions among projects (e.g., due to shared subsets of activities or interdependencies among their technical success probabilities and/or market values if successfully completed) provides a rich set of challenging problems for sophisticated, computationally intensive, optimization-based approaches to risk management. Promising approaches include *simulation-optimization* and *robust* and *multicriteria optimization methods* for portfolio selection, together with financial evaluation models that recognize the real option value of risky projects that can be abandoned if they become unattractive. The pharmaceutical and energy industries have developed and applied many such models [8].

As a practical matter, R&D portfolio optimization models can require potentially burdensome input data on project investment opportunities and their interactions (e.g., on overlapping tasks or substitute or complement effects for benefits of different projects), especially since the number of possible interactions grows exponentially with the number of projects. To ease this burden, some recent work has explored the use of approximate (fuzzy) descriptions of potential future cash flows in assessing project portfolio values in a *real options* framework, leading to a fuzzy mixed integer programming model for optimization of R&D portfolios [5].

OWNER VERSUS MANAGER INCENTIVES IN R&D PROJECT MANAGEMENT

In the simplest case of a single minimal path set and no precedence constraints, expected value is maximized by attempting the “most risky” or “most difficult” activities first, that is, those with the highest failure-probability-per-unit-cost ratios. This strategy minimizes the expected cost of discovering that success is impossible, if such is the case. However, rewarding a manager for successfully completed activities may create an incentive to postpone high-risk activities as long as possible, increasing the manager’s expected reward but increasing the expected cost of unsuccessful projects. Projects that should

be abandoned may be continued too long if managers postpone attempting high-risk activities until ones that are more likely to succeed have been completed. Conversely, managers may abandon projects too soon and/or may invest too little effort in making them succeed (from the owner's perspective) when effort and success probability are the managers' private information.

Such incentive problems (which are instances of moral hazard and principal-agent conflict problems) can be largely overcome by evaluating partially completed R&D projects using a *real options* framework and then compensating R&D managers according to these evaluations. More specifically, compensating R&D managers using appropriate adjusted residual income measures (with capital charge rates that are contingent on manager reports of estimated profitability, recognizing that these may be distorted by managers in an attempt to obtain greater compensation) allows an effective separation of R&D investment decisions by owners and effort allocation decisions by R&D managers. Such compensation rules optimally balance incentives to invest in risky R&D against incentives to abandon unfavorable projects in some specific models of owner and manager information and incentives [9].

OPTIMAL INTENSITY OF INVESTMENT IN COMPETITIVE R&D PROJECTS

When a company undertakes R&D to obtain a competitive advantage, time pressure from competition (especially in a "winner takes all" patent and intellectual property environment) may make it worthwhile to invest simultaneously in multiple alternative approaches (e.g., in multiple minimal path sets). This stochastically reduces the time, but stochastically increases the cumulative expenditure, until either success is achieved or no attractive paths forward remain.

Exactly how intensely a company should invest in R&D to maximize expected profit depends on how intensely competitors invest. Hence, quantitative models of optimal investment intensity are often formulated as differential games in continuous time, with

the competing firms deciding at each moment how intensely to invest in R&D based on information available so far. In such competitive R&D models, details of R&D project activities are typically suppressed, and R&D projects are represented by relatively simple models such as functions showing how the cumulative probability (or, equivalently, the hazard function) for successful completion of a project varies with the cumulative amount invested. For example, in the simplest case, each increment of cost or effort creates a fixed probability of successful completion (e.g., successfully finding a new process, drug, or so on); if it does not, then the next increment of cost or effort produces the same probability of success, and so forth. In this case, the probability of succeeding within a budget of t units of cost or effort is just $1 - \exp(-pt)$, where p reflects the success probability per unit of cost or effort. (In more general models, if it is assumed that the most promising approaches are tried first, then p is a decreasing function of t .) In these models, incremental investment purchases incremental success probability, and more rapid expenditures hasten the time either until success occurs, it becomes optimal to abandon the project, or no investment opportunity remains. Typical conclusions from such models are as follows:

- Competitive rivalry increases R&D spending by each competitor when the benefits of the first breakthrough are captured by the firm that achieves it [10].
- If patent protection is incomplete (e.g., reflecting positive externalities from increased knowledge) and/or there are opportunities for licensing so that the benefits of innovation may be shared even by noninnovators, then the effects of competition on the overall pace of innovation are ambiguous. Only some firms may invest in R&D, and social welfare may be either increased or decreased compared to the "winner takes all" situation [11].

Such strategic models of competitive R&D investment and resulting project value complement real options approaches that

treat competitor actions as exogenously specified and then optimize a firm's abandonment decisions (i.e., choice of when and whether to exercise the "option" to abandon). Such models typically assume that remaining cost-to-complete is uncertain and evolves according to a random process as R&D proceeds [12].

CONCLUSIONS

Project portfolio selection, project management, and incentive design problems pose challenging practical applications for quantitative risk management methods. The stakes for effective risk management methods in R&D and other risky projects are high, as the same set of risky project opportunities can lead to either profits or losses, depending on how well the risks are managed through project selection and management.

Although sophisticated quantitative methods for modeling and managing R&D risks (see *R&D Risk Management*) have been developed, as summarized in this article and pursued further in the references, applying such methods in practice, with realistically limited and uncertain data and estimates of potential future R&D costs and benefits continues to be challenging. Yet, it is of great importance in areas such as pharmaceuticals, high technology, and energy, where portfolios of risky projects (see *Managing A Portfolio of Risks*) often drive company value. Therefore, further development of practical methods for R&D risk characterization and risk management decision making will probably remain an active and rewarding area of both theoretical and applied risk management for the foreseeable future.

REFERENCES

1. Denardo EV, Rothblum UG, van der Heyden L. Index policies for stochastic search in a forest with an application to R&D project management. *Math Oper Res* 2004;29(1):162–181. DOI: 10.1287/moor.1030.0072.
2. Tsitsiklis JN. A short proof of the Gittins index theorem. *Ann Appl Probab* 1994;4(1):194–199.
3. Kaspi H, Mandelbaum A. Multi-armed bandits in discrete and continuous time. *Ann Appl Probab* 1998;8(4):1270–1290.
4. Ghasemzadeh F, Archer N, Iyogun P. A zero-one model for project portfolio selection and scheduling. *J Oper Res Soc* 1999;50(7):745–755.
5. Carlsson C, Fullér R, Heikkilä M, Majlender P. A fuzzy approach to R&D project portfolio selection. *Int J Approx Reason* 2007;44(2):93–105.
6. Ringuest JL, Graves SB, Case RH. Conditional stochastic dominance in R&D portfolio selection. *IEEE Trans Eng Manage* 2000;47(4):478–484.
7. Loch CH, Kavadias S. Dynamic portfolio selection of NPd programs using marginal returns. *Manage Sci* 2002;48:1227–1241.
8. Blau GE, Pekny JF, Varma VA, Bunch PR. Managing a portfolio of interdependent new product candidates in the pharmaceutical industry. *J Prod Innov Manage* 2004;21(4):227–245.
9. Pfeiffer T, Schneider G. Residual income-based compensation plans for controlling investment decisions under sequential private information. *Manage Sci* 2007;53(3):495–507.
10. Dockner EJ, Feichtinger G, Mehlman A. Dynamic R&D competition with memory. *J Evol Econ* 1993;3(2):145–152.
11. Mukherjee A. R&D, licensing and patent protection. 2002. Available at <http://citeseer.ist.psu.edu/545430.html>. Accessed 2007 Apr 15.
12. Schwartz ES. Patents and R&D as real options. *Econ Notes* 2004;33(1):23–54.

MANUFACTURING FACILITY DESIGN AND LAYOUT

SUNDERESH S. HERAGU
BANU Y. EKREN
Department of Industrial Engineering,
University of Louisville, Louisville,
Kentucky

INTRODUCTION¹

Structures that are thousands of years old, for example, the Egyptian pyramids, the ruins of the city of Pompeii in modern day Italy, and Harappa and Mohenjo-Daro civilizations of the Indus Valley, suggest that the layout problem has been considered by facility designers for thousands of years. The industrial revolution offers some examples of the importance of layout in designing factories. Henry Ford's assembly line is perhaps the first and significant example of factory layout used as one of the means to achieve higher levels of productivity and efficiency. Although we can cite numerous examples to illustrate the importance of facility planners and analysts placed on facility logistics, the study of facility logistics problems via sophisticated mathematical models did not occur until the introduction of the quadratic assignment problem (QAP) in the mid-1950s by Koopmans and Beckmann [2].

Manufacturing facilities can be broadly defined as buildings where men, material, and machines come together for manufacturing a product or providing a service [3]. It is important to manage a facility in such a way that several, often conflicting, objectives are satisfied. These objectives include producing a product or providing a service at minimum cost, high quality, and in a timely manner; minimizing use of

materials and natural resources; facilitating employee communication and supervision, and so on. To achieve these objectives, a good understanding and analysis of the underlying decision problems are important and necessary.

It is estimated that roughly 8% of the US gross national product has been spent annually since 1955 on new facilities [4]. In addition, expense for modification of previously purchased facilities is also significantly high. Tompkins *et al.* [4] claim that 20–50% of the total operating costs are related to material handling costs. Thus, if we develop an effective facilities plan, not only can these costs be reduced by at least 10–30%, but productivity can be increased as well.

In recent years, there has been a marked increase in the number and types of automated systems. This trend has greatly increased the complexity of system design problems. Designers and users of automated systems have therefore developed new tools to cope with their complexity.

Manufacturing or service system design involves solving a number of design and planning problems arranged in a hierarchy. Some examples are determining the products to be manufactured, services to be provided, manufacturing or service processes to be used, number and types of manufacturing or service equipment capable of performing the required operations, and so on. In the case of manufacturing systems, preliminary process plan development, determining the tooling and fixture requirements, layout of manufacturing cells and machines, material handling methods to be used, and number and types of specific material handling devices capable of performing the required material handling moves are some of the other important design questions that need to be addressed.

Design problems are typically addressed once in 3–5 years or more because it is costly to alter a design decision frequently. For example, once a plant is built in a certain geographic area, it is difficult to justify

¹Some of the material in this section has been reproduced from [3] with permission.

moving the plant to another location within 1 or 2 years of the first decision because the fixed costs would not have been fully recovered within this time period. For example, UPS built its air hub in Louisville, Kentucky, in 1980 to handle packages and letters for domestic and international shipment. When it wanted to integrate heavy freight into the network in 2006, it found that expanding at the same location was more cost effective than moving all its package delivery and heavy freight business to a new facility in another location. Thus, UPS undertook and recently completed a one-million square foot expansion in its Louisville facility. It is thus important to understand the long-term impact of design decisions and consider manufacturing or service activities that are to take place not only in the short run but also in the long run.

FACILITY LAYOUT

An important consideration in the design of both manufacturing and service systems is the physical arrangement or layout of departments. Studies on manufacturing systems indicate that 30–75% of the cost of a product is related to material handling expenses [5]. Tompkins *et al.* [4] mention that material handling activities account for 20–50% of a manufacturing company's total operating budget. Thus, if the departments are arranged optimally, manufacturers can reduce product costs and significantly enhance their competitive position. An arrangement is optimal if no other arrangement can be better with regard to the chosen criterion. In practice, besides minimizing the cost involved in moving goods or personnel between departments, many more factors such as the ones below need to be considered in optimizing the arrangement of departments:

- facilitating smooth flow of people and material by minimizing or eliminating congestion;
- utilizing the available space effectively and efficiently;
- facilitating communication and effective supervision; and

- providing a safe and pleasant environment for employees and visitors.

When designing a facility, the following constraints must be considered:

- Some department pairs must be located adjacently for safety reasons. For example, because of fire hazards, the forging and heat-treatment stations must be next to each other even if relatively few parts flow between them.
- Some department pairs must be in nonadjacent locations. Because the welding station generates sparks that could possibly ignite flammable solvents in the painting station, the two stations should be located as far apart as possible, although there may be a high interaction between them.
- Often, federal, state, and local governmental regulations and hazard insurance company requirements may force departments to be located in specific locations or numbers within the facility. For example, fire-fighting facilities must be strategically located in certain positions, a building of a certain size must have a minimum number of fire exits, and so on. The Occupational Health and Safety Administration (OSHA) requires separate rest rooms for men and women.

Before a facility design problem is addressed, its location must be determined. In practice, there is a multitude of factors that impact location decisions. These include taxes, proximity to source of raw materials, construction cost, transportation system, and so on. Models for determining the location of one or more facilities in a continuous or discrete space are available in the literature (see for example, [3,6]). In this article, we focus on manufacturing facility layout problem rather than its location.

Even though many companies may realize the importance of developing efficient layouts, many of them—especially the smaller ones—do not take this problem seriously. They think that it is too time consuming and difficult to collect and generate data in the format required for problem solving.

BASIC TYPES OF LAYOUT

The first step in the design of a manufacturing facility is to determine the general flow pattern for material, parts, and work-in-process (WIP) inventory through the system. *Flow pattern* refers to the overall pattern in which the product flows from beginning to end. It includes the flow patterns of the product through various stages of processing—from raw material (at the receiving stage) through semifinished product (at the fabrication stage) to the finished product (at the assembly stage). The second step is determining the *type* of layout to be used. We discuss five types of layouts. Although we focus on manufacturing systems layouts, these five layout types are also seen in nonmanufacturing (service) facilities [30].

Product Layout

The product layout is also known as *flow-line layout*, *production-line layout*, *assembly-line layout*, and *layout by product*. In a product layout, the machines and workstations are arranged in the sequence corresponding to the sequence of operations undergone by the product. For example, if a product undergoes milling, drilling, assembly, and packing operations in sequence, the vertical milling machines, drilling machines, assembly equipment, and packing equipment must be arranged one after the other in a line. This type of layout is mostly used by companies that manufacture a single or few items in large quantities (e.g., an automotive assembly plant). The most important advantage of this layout is reduced material handling

effort and time, minimal in-process inventory, and efficient planning and control. A major disadvantage is the relative inflexibility of the layout to changes in product mix or volume. Product mix includes the types of products that need to be produced and the volume represents the (annual) demand for the product. Therefore, this type of layout is not suitable for companies that frequently change production volume or variety.

Process Layout

The process layout is also known by other names such as *layout by process* or *jobshop layout*. In a process layout, machines and work stations are arranged on the basis of the processes they perform. For example, all milling machines are placed together in one department, all turning machines are placed together in another, and so on. A process layout is shown in Fig. 1. Each geometric shape represents a particular machine type; for example, circles represent drilling machines, squares represent milling machines, triangles represent pressing equipment, and so on. In a process layout, machines that perform the same process are grouped together in their own department.

The process layout is useful for companies that manufacture a wide variety of products or process a wide variety of jobs, in small quantities. The process layout is flexible and it allows workers to become experts in a particular process. The disadvantages of this layout are the increased material handling costs, long travel distances, high inventory, long cycle times, complexity in planning and control, and decreased productivity.

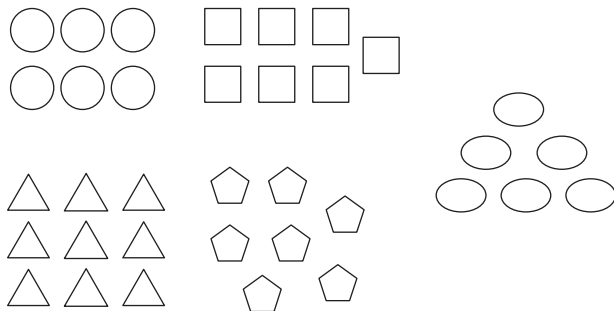


Figure 1. A process layout.

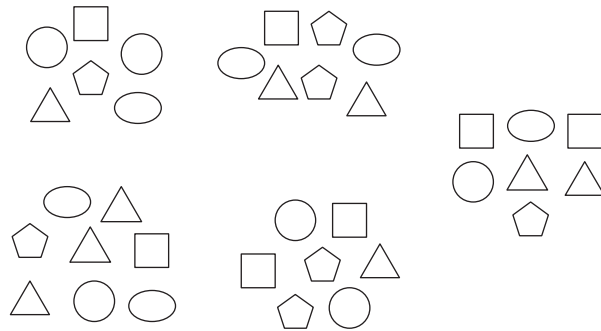


Figure 2. A GT-based or cellular manufacturing layout.

Fixed Position Layout

Fixed position layout is also referred to as *project layout*. In a fixed position layout, the product does not move from one location to another. Instead, the processes and equipment move to the product's location to perform the necessary operations. This layout is usually adopted when the manufactured or processed product is bulky and cannot be transported easily. Ship and aircraft manufacturing and dam, road, and house construction are some examples that employ this type of layout. The advantages of fixed position layout are that the likelihood of damaging the product and the cost of moving it are reduced. On the other hand, the cost of moving equipment to and from the work area is significantly high and the equipment utilization is low.

Group Technology–Based Layout

Group technology (GT) is a manufacturing philosophy in which similar parts are grouped together to maximize production efficiencies. Companies making a large

number of parts often manufacture these on a large number of machines. Managing this complex system as a whole is often difficult and inefficient. If such a large system can be divided into independent and smaller subsystems, managing all the subsystems separately is much easier and cost effective than attempting to manage the complex system as a whole [7]. In order to identify the subsystems, planners must first determine the sets of machines (machine cells) which can process corresponding families of parts (part families) in such a way that all or most of the operations required by any part in a part family are almost entirely processed by a subset of the machines in the corresponding machine cell. This type of layout is also known as *cellular* or *cellular manufacturing* layout.

A GT-based layout is illustrated in Fig. 2. The difference between a GT and process layout is that machines in a cell are not necessarily similar in their processing capabilities. The advantages of GT are reduced traffic congestion, material handling costs, WIP inventory, production lead times, and other forms of waste.

Hybrid Layout

As and when companies expand their production range and capabilities, so does machine variety. Very soon these companies find that the layout type used previously does not meet their needs entirely. Thus, they may choose to have a GT layout for the new set of machines added to the facility which already has a production-line layout. Figure 3 shows a sample hybrid layout that has characteristics of a process, GT, and product layout.

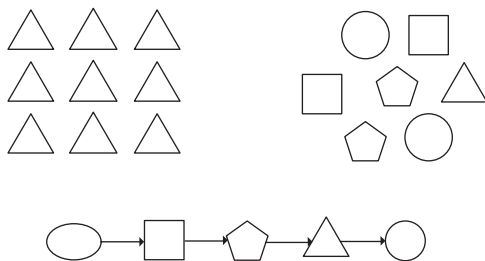


Figure 3. A hybrid layout.

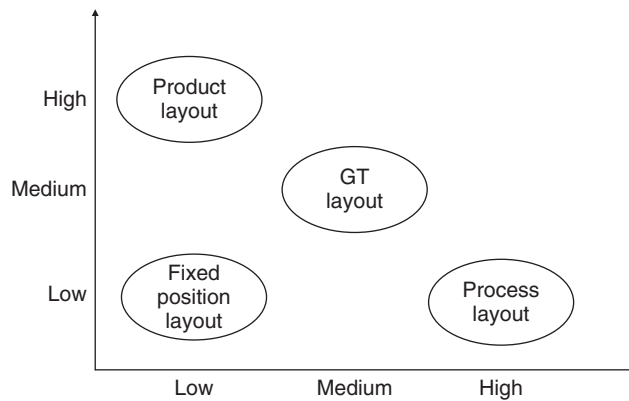


Figure 4. Layout classification based on volume-variety. (Source: Adapted from [4].)

The layout type which is suitable for a facility depending on product volume-variety can be summarized as in Fig. 4.

is further developed to include budget and time for the new layout establishment.

SYSTEMATIC LAYOUT PLANNING

Systematic layout planning (SLP) is a simple, step-by-step approach in the design of facilities. Even today, it is widely used to develop layouts. It consists of four steps [8].

Step 1—Determination of the location where departments will be established. In this step, the locations of the facilities or departments are determined.

Step 2—Overall layout establishment. In this step, the flow of materials between departments, space requirement for each department, and closeness requirements of departments are determined. Constraints such as safety and budget are also considered. Up to five alternate layouts are developed in this step. Then, the best layout is chosen according to designer preferences.

Step 3—Detailed layout plan development. In this step a detailed layout including the location of each machine, their supporting stations such as inspection station, and other facilities such as restroom and locker rooms are determined.

Step 4—The selected layout installation. In this step, the selected layout

Additional information on the SLP is provided in Refs 3, 4, 9.

NEXT-GENERATION FACILITY LAYOUTS

The discussion about traditional layouts in the section titled “Facility Layout” did not focus on minimizing inventory, shortening cycle times, or increasing flexibility. The focus was primarily on factors such as minimization of (loaded) material handling trips or costs, facilitating supervision or communication between personnel, and so on. Moreover, the GT and product layouts are designed assuming that changes in product mix and volume are relatively stable over a long period, say 3–5 years.

Today’s factories face constant changes in production requirements, and the production environment is thus highly dynamic. Existing layouts are not fully capable of meeting such changes or operating efficiently in such environments. Benjaafar *et al.* [10] point out the need for a new generation of facility layouts that are more flexible, modular, and easy to reconfigure. Other researchers such as Askin *et al.* [11], Benjaafar and Sheikhzadeh [12], Irani and Huang [13], Kochhar and Heragu [14], Montreuil [15], National Research Council [16], and Yang and Peters [17] have also pointed out the need for alternate types of next-generation factory layouts.

The metric for evaluating layouts has also changed in today's factories. The "robustness" of a layout over long time periods (during which there is a significant evolution relative to the product mix and volume) is of importance. Layouts that can be configured and reconfigured as frequently as necessary to meet the changing product mix and volume are also of importance. Modern layouts must lower inventories, decrease production lead times, and offer greater flexibility for product customization, in addition to minimizing the traditional measures such as minimizing loaded material handling trips.

Robust layouts are those that remain relatively efficient over relatively long periods from a material-handling standpoint even though the product mix and volume may have shifted dramatically from one planning period to the next. Robustness is typically achieved by duplicating key processes or machines at strategic locations within the production facility.

Reconfigurable layouts are those that can be changed relatively easily each time there is a moderate or heavy shift in product mix and volume. The assumption is that the costs of physically changing the layout from one configuration to another is relatively less expensive and time consuming and thus the benefits of a new layout for a given product mix and volume can be more fully realized [1,18]. Of course, an implicit assumption here is that the production processes are relatively

light-weight, a trend that is being seen in many process industries.

To develop reconfigurable manufacturing systems (RMS) with new methodologies and equipment, the Engineering Research Center for Reconfigurable Manufacturing Systems (ERC/RMS) at the University of Michigan (<http://erc.engin.umich.edu/>) focuses on building machines that can be quickly adjusted for changes in product mix or volumes. For example, machines can be quickly upgraded by adding spindles, axes, tool magazines, or controllers [19] so that facilities may use one machine for much of the processing without incurring any material handling movement or cost. Because a machine can be configured for different mixes and volumes easily, changes in production requirements will not affect the layout significantly.

Holographic layout is an alternative to process layout for systems operating in highly volatile environments. In this type of layout, subdepartments or machines are distributed throughout the shop floor. Duplication of departments or machines throughout the factory may also allow the facility to respond to future fluctuations quickly (see Fig. 5). Because the machines are dispersed throughout the building, it is possible to reduce material handling costs somewhat using holographic layouts. To design this layout, the planner should decide how to disaggregate and duplicate the subdepartments or machines throughout plant [13,20].

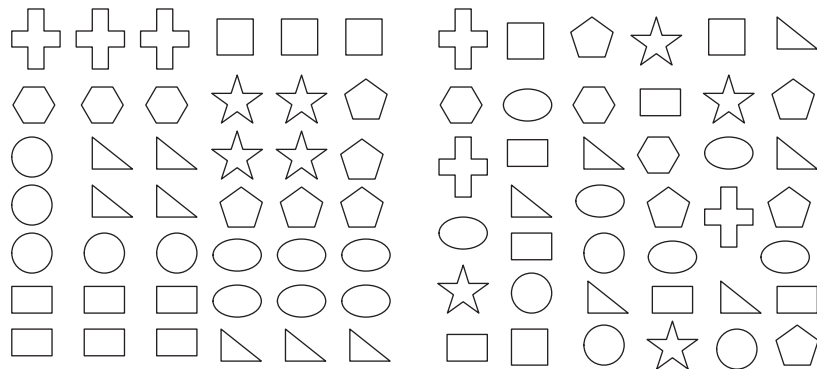


Figure 5. Holographic layout. (a) Partially distributed layout. (b) Entirely distributed layout. (Source: Adapted from [10].)

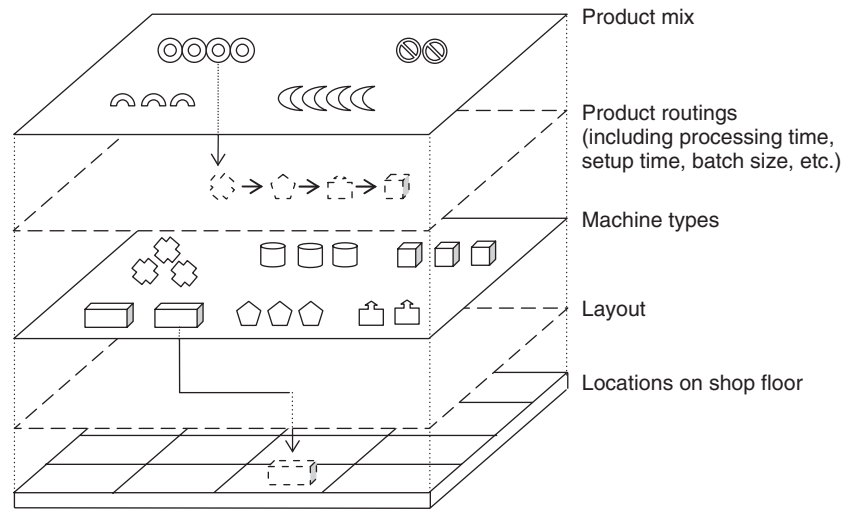


Figure 6. Facility layout problem reference model. (Source: [1].)

Fractal layouts can be helpful to a company that handles products requiring short runs or products having short life cycles. Temporary cells are formed by joining sub-departments dedicated to a particular product line or job order. This cell can be switched with a new one when a product is phased out or when a customer order is completed and a new one arrives. Drolet [21] studied such virtual cells. Lahmar and Benjaafar [22] found that distributed configurations can be effective in handling production growth and contraction.

LAYOUT DESIGN AND ANALYSIS²

To put a layout problem into context, the entities and activities relevant to the layout problem can be organized into a layered model as shown in Fig. 6. The reference model has three physical layers: product mix, machine types, and locations on the shop floor. Each available machine belongs to a machine type. The number of locations on the shop floor is equal to the total number of machines. The two logical layers of the

reference model denote the design activities involved. The process planning problem maps each product to a sequence of machine types. Its output (product routings) includes other production data such as processing time, setup time, and tooling information. The layout problem is to find a one-to-one mapping from machines on the machine types layer to locations on the shop floor. The third logical layer (not shown) pertains to the scheduling problem. It finalizes the machine types in a product routing to specific machines on specific shop floor locations, and coordinates the timing, sequencing, and prioritizing of all work orders assigned to one machine.

The solution of the layout problem is determined by entities and activities in other layers of the reference model. Let us define the collection of these entities and activities (i.e., product mix, product routings, machines, and locations on shop floor) as the context of a layout problem. As long as the context is fixed, theoretically speaking, there exists an optimum layout for this context. By definition, the facility layout problem is a simple assignment of n machines to n locations on the shop floor. What makes the facility layout problem difficult to solve is the large combinatorial search space, especially when n is large (possibly $n!$ feasible solutions if there are no special restrictions on the locations

²Some of the material in this section has been reproduced from [1] with permission.

of specific machines or relative location of a subset of machines) and the construction of a score function that incorporates various business considerations to evaluate the goodness of a layout.

Koopmans and Beckman [2] first presented the QAP in the context of macrolevel facility location problems, but it has been extensively used to model microlevel factory layout problems. Since the 1960s, there has been a significant volume of research focused on solving the QAP, hence the layout problem, optimally. Assuming the cost to transport a unit load of material through a unit distance is known between each pair of facility locations, the objective of a QAP is to find optimal assignment of all the facilities to locations so that the overall cost of transporting material between each pair of facility locations is minimized. Ever since the QAP was adopted for the microlevel facility logistical problems, there have been numerous attempts to solve it optimally. But because the QAP is known to be NP-complete, optimal algorithms cannot solve problems with more than 15 or 20 facilities. Real-world layout problems, however, are much larger, and numerous attempts have been made to solve the QAP heuristically. Several sources including Kusiak and Heragu [23], Meller and Gau [24,25], and others have surveyed these optimal and heuristic algorithms. A major drawback of adapting the QAP for the layout problem (and hence the techniques developed to solve the QAP-based layout problem) is that the QAP assumes the locations of facilities to be known *a priori*, implicitly assuming that the facilities are of the same area and shape. Because this assumption does not reflect reality—not all facilities will have the same shape and area—Heragu [26] and others have introduced more realistic modeling representations of the layout problem and appropriate techniques to solve them.

Another important deterministic solution procedure in facility layout problem uses space-filling curves. Bozer *et al.* [27] show how to use it in facility design problem by extending a layout algorithm—CRAFT—for solution of multiple floors layout. Using ideas from drawing space-filling curves, they come

up with a more efficient exchange of two departments than is possible via two-way or three-way exchanges. Bozer *et al.* [27] define the new algorithm as MULTIPLE (MULTI-floor Plant Layout Evaluation). They illustrate how space-filling curves and flexible department areas allow MULTIPLE to increase the number of exchanges within a floor and the number of exchanges across floors without splitting departments as well as how MULTIPLE controls department shapes.

The static plant layout (SPLP) minimizes the total material handling costs by assigning departments to various locations. It is usually formulated as a QAP. However, Rosenblatt [28] studied dynamic plant layout and developed optimal procedure for this problem on the basis of dynamic programming formulation. In this study, because solving a problem with a large number of departments may be computationally prohibitive, he also recommends heuristic approaches.

All of the above-mentioned layout models and techniques treat the problem in a deterministic manner. For example, a major assumption made by traditional layout models and algorithms is that the material flows between each pair of facilities is known well into the future (say, for a period of 5 years or more) *and* that these values do not change. Current realities simply do not allow us to look at these values as being static. The dynamic nature of today's manufacturing and service activities does not allow us to predict material flows that far into the future. Often, we are fortunate if we can predict material flows for the upcoming planning period. Moreover, as pointed out by Heragu and Kochhar [29], two changes taking place in the manufacturing design and manufacturing process technologies enable us to envision layouts that can be changed relatively frequently for relatively low cost.

1. Because we are making products increasingly with light-weight composites that can be engineered to have all the desired tensile strength, and other mechanical properties, machines or material handling systems processing or transporting these composite

products need not be heavy or require elaborate foundations.

2. Noncontact manufacturing processes such as electron beam welding and laser cutting also suggest that we do not need heavy machines with elaborate foundations. In fact, Heragu and Kochhar [29] have proposed light-weight machines with wheels and mounted on tracks that can be clamped in any desired location in a grid of tracks on a factory floor permitting their movement from one location to another as frequently as production changes warrant.

As discussed in Heragu and Kochhar [29], composites are the primary choice for many discrete manufactured components. Aluminum composites, for instance, can now replace cast iron parts, and phenolics are replacing aluminum parts [30,31]. Not only are these light but they can also be engineered to have excellent mechanical properties such as hardness, heat resistance, tensile strength, and vibration absorption. The last property permits machine tool designers to design functionally equivalent but lighter tools that do not require an elaborate foundation, making them easily movable. Nonabrasive manufacturing process technology, such as laser cutting, electron beam hardening, and molecular nanotechnology, also supports machine tool designers' quest for making light-weight machining equipment [32]. Permanent magnetic chucks that facilitate quick mounting and dismounting of tools are being developed [33]. In fact, these chucks do not magnetize the cutting tool, carry their own energy source, and do not obstruct machining. Once again, these features in and of themselves support rapid equipment reconfiguration. This trend is likely to continue well into the next two decades. In fact, through a workshop and a delphi survey, the committee on Visionary Manufacturing Challenges for 2020 has identified adaptable processes and equipment and reconfiguration of manufacturing operations as two key enabling technologies that will help companies overcome two of the six grand challenges or fundamental goals to remain

productive and profitable in the year 2020 [16]. These grand challenges are to *achieve concurrency in all operations* and to *reconfigure manufacturing enterprises rapidly in response to changing needs and opportunities*.

Numerous examples where facility layouts are modified on a frequent basis, sometimes every few months, have been cited in Benjaafar *et al.* [11]. Northern Telecom, in one of its manufacturing facilities facing constant product design changes, employs conveyor-mounted work cells that can be readily relocated just before a scheduled production/assembly change [34]. A primary advantage of reconfiguring a layout when warranted by changes in product mix and volume is that material handling cost can be minimized because equipment can be reconfigured to suit the new production mix and volume. Of course, this cost must more than offset the cost of moving equipment from its current location to a new one.

Thus, given the new reality that layouts are likely to change very frequently, possibly every few months rather than years and that, at best, we only have knowledge of the production activities during the upcoming planning period, we need to develop a layout only for the next planning period. In addition, due to the short term life of a given layout and availability of production data for this time period, it is possible to consider optimizing operational performance measures such as minimizing part cycle times and WIP inventory.

Notice that it is relatively easy to get detailed data on material flows, machine setup, process and transfer batching, and others relative to the production activities in the next period as opposed to manufacturing activities for the next 5 years. Although the required data are available, it is well known that these data are not static. For example, the time to transport a load from one machine to another is a function of where the material handling device (assigned to transfer that load) is located at the time the material transfer request comes in. Even if we were to assume that the device is always at the same known location, the travel time is determined by congestion and other factors and is highly variable.

Thus, it makes sense to develop stochastic models for layout analysis. Furthermore, because many factors—deterministic as well as stochastic—have an impact on determining the suitability of a layout, we need tools that can design and analyze a layout with respect to static design criteria *and* stochastic operational performance criteria. Developing layouts on the basis of static design criteria alone is inadequate and, in some cases, could lead to undesirable consequences. For example, Benjaafar [11] has shown that a layout that minimizes material transfer costs when considering only loaded trips (static design criteria) may have much higher WIP inventory (stochastic operational performance measure) than another layout that may not minimize (loaded) material transfer costs. In fact, it is quite possible that a layout that minimizes loaded material transfer costs might be infeasible in the sense that it leads to an infinite WIP accumulation at one or more machines.

To summarize, the potential to frequently alter layouts, therefore, in a sense, transforms the modern layout problem from a strategic problem in which only long-term material handling costs are considered to a tactical problem in which operational performance measures such as reduction of product flow times, WIP inventories, and maximizing throughput rate are considered in addition to material handling and machine relocation costs when changing from one layout configuration to the next.

Meng *et al.* [1] have proposed a three-phase procedure for the design and analysis of next-generation factory layouts. In the first phase, multiple, alternate layouts that perform well with respect to static design criteria are generated using available algorithms for facility layout. In the second phase, each alternate layout is evaluated with respect to several stochastic operational performance measures. The manufacturing system performance analyzer (MPA) developed for analytically evaluating the performance of a layout is an extension of Whitt's [35] queueing network analyzer (QNA) [36,37]. Because MPA also incorporates the factors such as empty travel time of the material handling device, setup, machine

failures, and batching, it is a more realistic, comprehensive, and accurate tool. MPA not only determines the performance measures analytically but also provides simulation results for the same performance measures using ProModel software.

The analysis from phases 1 and 2 are combined to rank-order a set of three or so layouts that perform well with respect to the user-weighted design and operational performance criteria.

SUMMARY

Because of continuous changes in product mix, volume, and manufacturing requirements, next-generation facility layouts that are robust for multiple production periods or scenarios or are flexible and therefore can be reconfigured with minimal effort to meet the production requirement changes are preferred. When evaluating the candidate layouts, their performance relative to stochastic performance measures must also be considered. For example, WIP inventory level and cycle time evaluations, in addition to static material handling costs make a layout analysis more comprehensive. As a performance evaluation tool of facility layouts, MPA is a powerful tool for reconfigurable layout problems.

REFERENCES

1. Meng G, Heragu SS, Zijm WHM. Reconfigurable layout problem. *Int J Prod Res* 2004; 42(22):4709–4729.
2. Koopmans TC, Beckman M. Assignment problems and the location of economic activities. *Econometrica* 1957;25(1):53–76.
3. Heragu SS. *Facilities design*. 3rd ed. Clermont (FL): CRC Press; 2008.
4. Tompkins JA, White JA, Bozer YA, Tanchoco JMA. *Facilities planning*. 4th ed. New York: John Wiley & Sons, Inc.; 2010.
5. Sule DR. *Manufacturing facilities: location, planning and design*. Boston (MA): PWS Kent; 1991.
6. Daskin ML. *Network and discrete location: models, algorithms, and applications*. New York: Wiley; 1995.

7. Heragu SS, Zijm WHM, van Ommeren JCW. An open queueing network approach to evaluating cellular and jobshop layouts. Technical report. Troy (NY): DSES Department, Rensselaer Polytechnic Institute; 2000.
8. Muther R. Systematic layout planning. New York: Van Nostrand Reinhold Company; 1973.
9. Francis RL, McGinnis LF, White JA. Facility layout and location: an analytical approach. New Jersey: Prentice Hall; 1992.
10. Benjaafar S, Heragu SS, Irani SA. Next generation factory layouts: research challenges and recent progress. *Interfaces* 2002;32(6):58–76.
11. Askin RG, Lundgren NH, Ciarallo F. A material flow based evaluation of layout alternatives for agile manufacturing. In: Graves RJ, McGinnis LF, Medeiros DJ, *et al.* editors. *Progress in material handling research*. Ann Arbor (MI): Braun-Brumfield; 1997. pp. 71–90.
12. Benjaafar S, Sheikhzadeh M. Design of flexible plant layouts. *IIE Trans* 2000; 32(4):309–322.
13. Irani SA, Huang H. Custom design of facility layouts for multi-product facilities using layout modules. *IEEE Trans Rob Autom* 2000;16:259–267.
14. Kochhar JS, Heragu SS. Facility layout design in a changing environment. *Int J Prod Res* 1999;37(11):2429–2446.
15. Montreuil B. Fractal layout organization for job shop environments. *Int J Prod Res* 1999;37(3):501–521.
16. National Research Council. Visionary manufacturing challenges for 2020. Washington (DC): National Academy Press; 1998.
17. Yang T, Peters BA. Flexible machine layout design for dynamic and uncertain production environments. *Eur J Oper Res* 1998; 108(1):49–64.
18. Heragu SS, Zijm WHM. A framework for designing reconfigurable facilities. Technical report. Troy (NY): DSES Department, Rensselaer Polytechnic Institute; 2000.
19. Koren Y, Jovane F, Moriwaki T, *et al.* Reconfigurable manufacturing systems. *Ann CIRP* 1999;48:527–539.
20. Montreuil B, Venkatadri U. Strategic interpolative design of dynamic manufacturing systems layout. *Manage Sci* 1991;37(6):682–694.
21. Drolet JR. Scheduling virtual cellular manufacturing systems [PhD dissertation]. West Lafayette (IN): School of Industrial Engineering, Purdue University; 1989.
22. Lahmar M, Benjaafar S. Design of dynamic distributed layouts. Working paper. Minneapolis (MN): Department of Mechanical Engineering, University of Minnesota; 2002.
23. Kusiak A, Heragu SS. The facility layout problem. *Eur J Oper Res* 1987;29(3):229–251.
24. Meller R, Gau KY. The facility layout problem: Recent and emerging trends and perspectives. *J. Manuf Syst* 1996a;15(5):351–366.
25. Meller R, Gau KY. Facility layout objective functions and robust layouts. *Int J Prod Res* 1996b;34(10):2727–2742.
26. Heragu SS. Recent models and techniques for the layout problem. *Eur J Oper Res* 1992; 57(2):136–144.
27. Bozer YA, Meller RD, Erlebacher SJ. An improvement-type layout algorithm for single and multiple-floor facilities. *Manage Sci* 1994;40(7):918–932.
28. Rosenblatt MJ. The dynamics of plant layout. *Manage Sci* 1984;32(1):76–86.
29. Heragu SS, Kochhar JS. Material handling issues in adaptive manufacturing systems. The materials handling engineering division 75th anniversary commemorative volume. New York: ASME; 1994. pp. 9–13.
30. Arimond J, Ayles WR. Phenolics creep up on engine applications. *Adv Mater Proc* 1993; 143(6):20.
31. Fujine M, Kaneko T, Okijima J. Aluminium composites replace cast iron. *Adv Mater Processes* 1993;143:34–36.
32. Asari M. Electron beam hardening system. *Adv Mater Proc* 1993;143(6):30–31.
33. American Machinist (online edition). U.S. builder enters small lathe market. 137. 10. www.americanmachinist.com. 1993.
34. Assembly Magazine (online edition). Norstart custom telephones rely on flexible conveyor systems. May issue. Retrieved 2002 Jun 2. www.assemblymag.com. 1996.
35. Whitt W. Performance of the queueing network analyzer. *Bell Syst Tech J* 1983; 62(9):2817–2843.
36. Meng G. Open queueing network performance analyzer for reconfigurable manufacturing systems [PhD thesis]. Troy (NJ): Department of Decision Science and Engineering System, Rensselaer Polytechnic Institute; 2002.
37. Meng G, Heragu SS. Batch size modeling in a multi-item, discrete manufacturing system via an open queueing network. *IIE Trans* 2004;36(8):1–11.

MARKOV AND HIDDEN MARKOV MODELS

REFIK SOYER

Department of Decision Sciences,
The George Washington
University, Washington, DC

INTRODUCTION AND OVERVIEW

Stochastic process models are fundamental for assessing the reliability of items that operate under a dynamic environment. As pointed out by Singpurwalla [1], since the dynamic environment causes changes in the physics of failure, use of stochastic processes provides flexibility in describing failure characteristics under these environments. Earlier uses of stochastic process models include Mercer [2] who modeled item wear and Klein [3] who used a Markov chain (MC) model to describe stochastic deterioration of a system. Gaver [4] is the first who proposed modeling the failure rate as a stochastic process and introduced the notion of a *randomly changing environment*.

Assessment of reliability of an item/system as a function of the stochastic process governing the environment was considered by Çinlar [5–7], and the notion further developed by Çinlar and Özekici [8] who considered models of stochastic dependence caused by a random environment. The model of Çinlar and Özekici [8] is studied further in Çinlar *et al.* [9]. Related developments are those by Lindley and Singpurwalla [10], and by Singpurwalla and Youngren [11]. Özekici [12] analyzed the optimal maintenance problem of a single-component device operating in a random environment. In all these cases, the environment is characterized by a stochastic process model; see Singpurwalla [1], for a survey. Recent work in stochastic hazard processes can be found in Özekici [13].

Note that a stochastic process is a collection of random variables that are

indexed by a parameter space T . In our setup the parameter space T will be time which could be discrete, say $T = I = (0, 1, 2, \dots)$, or continuous, say $T = \mathbb{R}^+ = (t; 0 \leq t < \infty)$. In our notation, the collection of random variables $X_0, X_1, \dots, X_n, \dots$, will denote a discrete-time stochastic process, $\{X_n; n \in I\}$, whereas the collection $\{Y_t; t \in \mathbb{R}^+\}$ denotes a continuous-time stochastic process. The state space E of the stochastic process, that is, the possible set of values that either X_n or Y_t can take, can be discrete or continuous. When E is discrete, the process $\{X_n; n \in I\}$ is said to be a discrete-state, discrete-time stochastic process, and $\{Y_t; t \in \mathbb{R}^+\}$ is a discrete-state, continuous-time process. Similarly, when E is continuous the process $\{Y_t; t \in \mathbb{R}^+\}$ is referred to as a continuous-state, continuous-time process.

Of the several stochastic process models that have been considered by the above authors, the simplest is the Markov process (MP), but many stochastic models used in reliability analysis belong to the Markov class. Important examples include discrete and continuous-time Markov chains (CTMCs) and diffusion processes. The section titled “Markov Processes” gives an overview of Markov processes and discusses some of the widely used examples. Most of the material in this section can be found in Çinlar [14], Karlin and Taylor [15], or Ross [16].

Stochastic processes that are governed by MPs, which cannot be observed, are referred to as *hidden Markov models* (HMMs) [17]. Such processes have found applications in areas such as speech recognition [18], signal processing [19], biology and medicine [20], environmental sciences [21], software engineering [22–24] and have been of interest to statisticians [25–28] because of the difficult inferential issues that they pose. An excellent book on the HMMs is by Elliot *et al.* [29]. The HMMs will be discussed in the section titled “Hidden Markov Models”. An application of the HMMs to software reliability analysis is presented in the section titled “Application of HMMs in Software Reliability”.

MARKOV PROCESSES

The stochastic process $\{Z_t; t \in T\}$ is said to be an MP if for any $t, s \in T$ and for any set of states $A \in E$,

$$P(Z_{t+s} \in A | Z_u; u \leq t) = P(Z_{t+s} \in A | Z_t). \quad (1)$$

The above relationship is referred to as the *Markov property*. It simply states that given the present the future of the process is independent of the past. If Equation (1) is independent of t , then the MP is said to be time-homogeneous.

A continuous-time MP, Z_t , with continuous state space E is called a *diffusion process*. Diffusion processes are used to describe item wear and system degradation over time [30]. Brownian motion is the most commonly used diffusion process for modeling such phenomena. An important class of MPs are processes with *independent increments*. A continuous-time stochastic process Z_t will have independent increments if for all $t_1 < t_2 < \dots < t_n$ the random variables Z_{t_1} , $Z_{t_2} - Z_{t_1}, \dots, Z_{t_n} - Z_{t_{n-1}}$ are independent [16]. Furthermore, the process will have *stationary increments* if the distribution of $(Z_{t+s} - Z_t)$ does not depend on t . It is easy to see that every stochastic process Z_t with independent increments is an MP. Brownian motion process has independent increments.

Another process with the independent increments property is the gamma process [31]. As noted by Singpurwalla [1] and Park and Padgett [30], the Gaussian assumption of increments in the Brownian motion makes it an unattractive choice for modeling wear and degradation since the process is not monotonically increasing. One way to alleviate this is to use a geometric Brownian motion or a gamma process as suggested in Park and Padgett [30]. As pointed by the authors, the gamma process is more desirable than the geometric Brownian motion since it is always positive and increasing.

The original use of gamma process for describing deterioration goes back to Abdel-Hameed [32]. A gamma process Z_t has the following properties:

1. $Z_0 = 0$;
2. Z_t has independent increments;
3. $(Z_t - Z_s)$ has a gamma distribution $G[(a_t - a_s), b]$ with shape parameter $(a_t - a_s) > 0$ and scale parameter $b > 0$ for all $t > s$.

An excellent review of gamma processes and their use in maintenance is given in van Noortwijk [33].

Markov Chains

An MP with discrete state space E is called a *Markov chain*. If the process is a discrete-time process, that is, if $T = I$ then the process is called a *discrete-time Markov chain* or more commonly a *Markov chain*. More specifically, a stochastic process $\{X_n; n \in I\}$ is said to be an MC, with state space E , if for all $n, m \in I$, and $(x_1, x_2, \dots) \in E$,

$$\begin{aligned} P(X_{n+m} = x_{n+m} | X_n = x_n, \dots, X_0 \\ = x_0) = P(X_{n+m} = x_{n+m} | X_n = x_n). \end{aligned} \quad (2)$$

The MC is said to be time-homogeneous, if the above relationship is independent of n , that is,

$$P(X_{n+m} = j | X_n = i) = P_m(i, j),$$

where $P_m(i, j)$ is called the *m-step transition probability* of the MC. Also, $P(i, j) = P_1(i, j)$ is called the *transition probability* from state i to state j of the time-homogeneous MC, that is,

$$\begin{aligned} P(X_{n+1} = j | X_n = i, \dots, X_0) \\ = P(X_{n+1} = j | X_n = i) = P(i, j). \end{aligned}$$

The matrix $\mathbf{P}$, whose (i, j) th element is $P(i, j)$ is called the *transition probability matrix* of the MC. Note that $P(i, j) \geq 0$, for any $i, j \in E$ and $\sum_{j \in E} P(i, j) = 1$.

A continuous-time discrete state space MP, Y_t , is called a *continuous-time Markov chain*. Our discussion of CTMCs in the Section titled “Continuous-Time Markov Chains” is based on Ref. 17.

Continuous-Time Markov Chains

The stochastic process $\{Y_t; t \in \mathbb{R}^+\}$ is said to be a CTMC with discrete state space E , if for any $j \in E$, and $t, s \in \mathbb{R}^+$

$$P(Y_{t+s} = j | Y_u; u \leq t) = P(Y_{t+s} = j | Y_t). \quad (3)$$

If the above relationship does not depend on t , then the CTMC is said to be *time-homogeneous*, and we write,

$$P(Y_{t+s} = j | Y_t = i) = P_s(i, j).$$

where $P_s(i, j)$ is called the *transition function* of the CTMC and it has the following properties:

1. $P_s(i, j) \geq 0$
2. $\sum_{j \in E} P_s(i, j) = 1$
3. $P_{r+s}(i, j) = \sum_{k \in E} P_r(i, k) P_s(k, j)$.

The relationship (3) above is called the *Chapman—Kolmogorov equation*.

A CTMC with

$$P(Y_{t+s} = j | Y_t = i) = \begin{cases} P_s(i, j) = 0, & \text{if } j < i \\ \frac{e^{-\lambda s} (\lambda s)^{j-i}}{(j-i)!}, & \text{if } j \geq i, \end{cases} \quad (4)$$

is known as a homogeneous *Poisson process* for some constant $\lambda > 0$.

Associated with any CTMC, is a discrete parameter, discrete-state stochastic process $\{X_n; n \in I\}$, where X_n can be constructed from the process $\{Y_t; t \in \mathbb{R}^+\}$ as follows:

$$X_n = Y_{T_n}, \text{ for } T_n \leq t < T_{n+1}, \quad n = 0, 1, 2, \dots, \quad (5)$$

where $0 \equiv T_0 < T_1 < \dots < T_n < T_{n+1} < \dots$ are the times at which the CTMC changes states. The process $\{X_n; n \in I\}$ keeps track of the various states that the process $\{Y_t; t \in \mathbb{R}^+\}$ takes over time starting with the initial state Y_{T_0} at time 0. The process $\{X_n; n \in I\}$ is said to be *engendered* by the process $\{Y_t; t \in \mathbb{R}^+\}$ and has certain properties.

Result 1. Suppose that $\{Y_t; t \in \mathbb{R}^+\}$ is a CTMC with state space E . Let $\{X_n; n \in I\}$ be the corresponding discrete parameter process that is engendered by the above CTMC via the tracking scheme mentioned before. Then, $\{X_n; n \in I\}$ is an MC with some transition matrix $\mathbf{P}$, where $P(i, j) \geq 0$, for $i \neq j$, $P(i, i) = 0$, and $\sum_{j \in E} P(i, j) = 1$.

Thus, the Markov property of the $\{Y_t; t \in \mathbb{R}^+\}$ process is inherited by its engendered $\{X_n; n \in I\}$ process. The process $\{X_n; n \in I\}$ is called the *embedded chain* of the CTMC.

Result 2. For any $u \geq 0$, and a constant $\mu(i) > 0$,

$$P(T_{n+1} - T_n \geq u \mid X_{n+1} = j, X_n = i) = e^{-\mu(i)u}.$$

Thus the sojourn times in a CTMC have an exponential distribution whose scale parameter depends only on the current state i ; it does not depend on the next state j to which X_n is going to make a transition. Thus μ needs only be indexed by i .

Result 3. For any $u_i \geq 0$, $i = 1, 2, \dots$, and $\mu(i_j) > 0$, $j = 0, 1, 2, \dots$

$$\begin{aligned} P(T_1 - T_0 \geq u_1, T_2 - T_1 \geq u_2, \dots, \\ T_n - T_{n-1} \geq u_n \mid X_0 = i_0, \dots, X_n = i_n) \\ = \exp(-\mu(i_0)u_1) \exp(-\mu(i_1)u_2) \dots \\ \exp(-\mu(i_{n-1})u_n). \end{aligned}$$

That is, the sojourn times of a CTMC are, conditional on $X_0, \dots, X_n$, independently, but not identically exponentially distributed. The scale parameters of the exponential distributions depend only on the current states of the X_i 's.

The next result pertains to the joint distribution of the sojourn time in a state, and the state to which the CTMC is going to make a transition to.

Result 4. For any $i, j \in E$, and $u \geq 0$

$$\begin{aligned} P(X_{n+1} = j, T_{n+1} - T_n \geq u \mid X_0, \dots, \\ X_n, T_0, \dots, T_n) \\ = P(X_{n+1} = j, T_{n+1} - T_n \geq u \mid X_n = i) \\ = P(i, j) e^{-\mu(i)u}. \end{aligned}$$

Thus for a CTMC, the said joint distribution depends only on the current state of the process i , via the $\mu(i)$, and the transition probability matrix $\mathbf{P}$ of the embedded chain. Furthermore, the transition to a particular state is independent of the time spent in the current state.

In principle, we should be able to obtain results parallel to those of the four results for processes other than the Markov. However, the attractive properties of inheriting the Markovian feature, the exponentiality and the (conditional) independence of the sojourn times, and the (conditional) independence between the sojourn time and the state transitioned to, may be lost; see Limnios [34] for *semi-Markov processes*. Finally, since homogeneous Poisson processes with the parameter $\lambda > 0$ is a special case of the CTMC, our Results 2, 3, and 4 simplify with all the $u(\cdot)$'s replaced by a common λ . Consequently, here, conditioning on the current state does not matter. Furthermore, in the case of a Poisson process $P(i, j) = 1$, only if $j = i + 1$; otherwise it is zero.

The Generator of CTMC and Markovian Analysis

Analogous to the notion of a transition probability matrix of a MC, is the notion of a transition rate matrix of a CTMC. Specifically, the matrix $\mathbf{P}_s$, whose (i, j) th element is $P_s(i, j)$ is of relevance for studying several characteristics of a CTMC. Recall that $P_s(i, j)$ is the probability that the CTMC transitions from state i to state j in a time interval of length s .

Result 5. For any $i, j \in E$, $P_s(i, j)$ is differentiable, and $\lim_{s \rightarrow 0} \frac{d}{ds} P_s(i, j) = A(i, j) = \mu(i)[P(i, j) - I(i, j)]$, where $I(i, j) = 1$, if $j = i$, and is zero otherwise. Thus,

$$A(i, j) = \begin{cases} -\mu(i), & \text{if } j = i \\ \mu(i)P(i, j), & \text{if } j \neq i. \end{cases}$$

Consequently, if $P_s(i, j)$ is known for every $i, j \in E$, then $A(i, j)$ is known, and from there we know $\mu(i)$ and $P(i, j)$. Thus a knowledge of the transition probability matrix $\mathbf{P}_s$ of the CTMC enables us to obtain the transition probability matrix $\mathbf{P}$ of the embedded chain,

and also, the $\mu(i)$'s, the scale parameters of the distributions of the sojourn times of the process. Conversely, if E is a finite set of say m elements, then a knowledge of $\mu(1), \dots, \mu(m)$, and $\mathbf{P}$ enables us to obtain $\mathbf{A}$, the matrix whose (i, j) th element is $A(i, j)$. Once $\mathbf{A}$ is known, we may obtain the matrix $\mathbf{P}_s$ via the *matrix-exponential* relationship

$$\mathbf{P}_s = e^{s\mathbf{A}} = \sum_{k=0}^{\infty} \frac{s^k \mathbf{A}^k}{k!};$$

where the summation is a consequence of the Taylor series expansion of $e^{s\mathbf{A}}$. The above development is known as a *Markovian analysis*. The matrix $\mathbf{A}$ is called the *generator* of the CTMC.

HIDDEN MARKOV MODELS

As previously pointed out, HMMs are stochastic processes that are governed by MPs, which cannot be observed. Examples of HMMs include Markov modulated Bernoulli processes (MMBPs) [27,35], where the governing or modulating process is an MC and Markov modulated Poisson processes (MMPPs) [36] where the modulating process is a CTMC. It is important to note that here we use the term HMM or a Markov modulated process in a more general manner to include any MP as a governing process. Sometimes in the literature, the term HMM is used to refer only to discrete-time processes governed by MCs [28], whereas the term Markov modulated process is reserved for the case where the governing process is a CTMC [37].

In reliability modeling, the HMMs are sometimes referred to as *random environment models* as in Özekici and Soyer [38]. In Özekici and Soyer [38] the item/system in question is assumed to operate in a randomly changing environment depicted by $\{Y_t; t \in \mathbb{R}^+\}$, where Y_t is the state of the environment at time t . The environmental process Y is an MP with some state space E , which is assumed to be discrete to simplify the notation. Note that if the environmental process is assumed to be a semi-MP, then the resulting processes will be hidden semi-Markov models as considered by Limnios [34] and Özekici and Soyer [39].

In what follows, we assume that the environmental process Y is an MP and we consider the MMPP and the MMBP. As noted in Singpurwalla *et al.* [17], MMPP is a special case of the *Markov renewal process* (MRP) of Çinlar [5] as well as the *doubly stochastic Poisson process* of Kingman [40], which is also known as the *Cox process*.

Markov Modulated Poisson Model

Let N be a modulated Poisson process such that N_t depicts the total number of arrivals until time t . The modulation is done via an environmental process Y with a discrete state space E where Y_t represents the state of the environment at time t . The rate of arrivals at time t is $\lambda(Y_t)$ for some arrival rate vector λ defined on E . We suppose that while the environment is at state i arrivals occur according to an ordinary Poisson process with rate $\lambda(i)$. To be more precise,

$$P(N_t = k | Y) = \frac{e^{-H_t} H_t^k}{k!}, \quad (6)$$

where

$$H_t = \int_0^t \lambda(Y_s) ds, \quad (7)$$

for all $k = 0, 1, \dots$, and $t \geq 0$.

It follows from the above that, given Y , N is a nonstationary Poisson process with mean value function $E(N_t | Y) = H_t$. Defining T to be the arrival time process so that T_n is the time of the n th arrival, we have the conditional interarrival time distribution

$$P(T_{n+1} - T_n > t | Y, T_n) = e^{-(H_{T_n+t} - H_{T_n})}. \quad (8)$$

The modulated process reduces to the ordinary Poisson process with rate λ if the arrival rate vector is $\lambda(i) = \lambda$ independent of the environment for all i . In this case, $H_t = \lambda t$ deterministically.

The arrival process N can be studied via the additive functional H of Y . In particular, Equations (6) and (8) directly yield

$$P(N_t = k) = E \left(\frac{e^{-H_t} H_t^k}{k!} \right) \quad (9)$$

and

$$P(T_{n+1} - T_n > t | T_n) = E(e^{-(H_{T_n+t} - H_{T_n})}). \quad (10)$$

Therefore, the probability law of H , thus, that of the environmental process Y , will play an important role in our analysis of N and T .

We assume that Y is an MP with an embedded MC X and jump times S . In other words, X_n is the n th state visited by the environmental process and S_n is the time of this visit such that,

$$Y_t = X_n \text{ whenever } S_n \leq t < S_{n+1}. \quad (11)$$

If $\mathbf{P}$ is the transition matrix of the embedded MC X , and μ is the vector of jump rate, then transition function of the Y process is given by

$$P(Y_t = j | Y_0 = i) = P(i, j)(1 - e^{-\mu(i)t}). \quad (12)$$

More precisely, if the process Y is in some state i , then it stays there for an exponentially distributed amount of time with rate $\mu(i)$ and then jumps to some other state j with probability $P(i, j)$. It also follows that

$$F_i(t) = P(S_1 \leq t | Y_0 = i) = 1 - e^{-\mu(i)t} \quad (13)$$

and the generator of the MP Y is given by the matrix

$$A(i, j) = \begin{cases} -\mu(i), & \text{if } j = i \\ \mu(i)P(i, j), & \text{if } j \neq i. \end{cases} \quad (14)$$

Let P_t denote the transition function of Y given by Equation (12), that is,

$$P_t(i, j) = P(Y_t = j | Y_0 = i). \quad (15)$$

Following Özekici and Soyer [39], we define another process Y^λ such that

$$Y_t^\lambda = \begin{cases} Y_t, & \text{if } t < T_1, \\ \Delta, & \text{if } t \geq T_1, \end{cases} \quad (16)$$

where T_1 is the time of the first arrival. While the environment is in state i , the time of arrival has the exponential distribution with rate $\lambda(i)$. It is clear that Y^λ is also an MP on the extended state space $E_\Delta = E \cup \{\Delta\}$ and it is obtained by “stopping” the MP Y as soon

as an arrival occurs. Here, Δ is an absorbing state where the process is dumped to as soon as it is stopped. The transition matrix of the embedded MC is now extended as

$$P_\lambda(i, j) = \begin{cases} \frac{\mu(i)}{\mu(i) + \lambda(i)} P(i, j), & \text{if } i, j \in E \\ \frac{\lambda(i)}{\mu(i) + \lambda(i)}, & \text{if } i \in E, j = \Delta \\ 1, & \text{if } i, j = \Delta \end{cases} \quad (17)$$

and the transition rate vector is

$$\mu_\lambda(i) = \begin{cases} \mu(i) + \lambda(i), & \text{if } i \in E \\ 0, & \text{if } i = \Delta. \end{cases} \quad (18)$$

If we let the matrix $A_\lambda(i, j) = \mu_\lambda(i)(P_\lambda(i, j) - I(i, j))$ denote the generator of Y^λ , then it is well known that the transition function $P_t^\lambda(i, j) = P(Y_t^\lambda = j | Y_0^\lambda = i)$ for all $i, j \in E_\Delta$ is given by the matrix-exponential solution

$$P_t^\lambda = e^{A_\lambda t} = \sum_{n=0}^{+\infty} \frac{t^n}{n!} A_\lambda^n. \quad (19)$$

A further simplification is obtained by noting that

$$A_\lambda(i, j) = A(i, j) - \Lambda(i, j), \quad (20)$$

for all $i, j \in E$ where Λ is a diagonal matrix defined as

$$\Lambda(i, j) = \begin{cases} \lambda(i), & \text{if } j = i \\ 0, & \text{if } j \neq i. \end{cases} \quad (21)$$

Since $A_\lambda(\Delta, j) = 0$ and $A_\lambda(i, \Delta) = \lambda(i)$ for all $i \in E$ and $j \in E_\Delta$, we can rewrite Equation (19) as

$$P_t^\lambda(i, j) = e^{A_\lambda t}(i, j) = e^{(A - \Lambda)t}(i, j) \quad (22)$$

for all $i, j \in E$.

Now, note that our construction of Y^λ implies

$$T_1 = \inf\{t \geq 0; Y_t^\lambda = \Delta\} \quad (23)$$

and T_1 is the first-passage time to the absorbing state Δ . So, it has a phase-type distribution and, in particular,

$$\begin{aligned} P_i(T_1 > t) &= P_i(Y_t^\lambda \in E) = \sum_{j \in E} P_t^\lambda(i, j) \\ &= \sum_{j \in E} e^{(A - \Lambda)t}(i, j). \end{aligned} \quad (24)$$

Note that in reliability applications where a device fails exponentially with a failure rate that depends on the randomly changing environment, Equation (24) gives the survival function. In this case, the mean time to failure is another quantity of interest. Using the Markov property, it can be computed by solving the system of linear equations

$$E_i(T_1) = \mu_\lambda(i)^{-1} + \sum_{j \in E} P_\lambda(i, j) E_j(T_1), \quad (25)$$

for $i \in E$ so that the explicit solution is

$$E_i(T_1) = \sum_{j \in E} [I - P_\lambda]^{-1}(i, j) \mu_\lambda(j)^{-1}. \quad (26)$$

Following Özekici and Soyer [39] an ergodic analysis can be developed. Suppose that both X and Y are ergodic processes with limiting distributions $\nu(j) = \lim_{n \rightarrow +\infty} P(X_n = j)$ and $\pi(j) = \lim_{t \rightarrow +\infty} P(Y_t = j)$. This implies that ν is the unique solution of $\nu = \nu \mathbf{P}$ with the normalizing condition $\sum_{i \in E} \nu(i) = 1$. It can be shown that

$$\pi(j) = \frac{\nu(j)/\mu(j)}{\sum_{k \in E} (\nu(k)/\mu(k))}. \quad (27)$$

Using Theorem 1 of Özekici and Soyer [39], we can obtain

$$\lim_{t \rightarrow +\infty} E_i(N_t - \hat{\lambda}t) = \sum_{j \in E} \pi(j) \left[\frac{\hat{\lambda} - \lambda(j)}{\mu(j)} \right] \quad (28)$$

and

$$\lim_{t \rightarrow +\infty} \frac{E_i(N_t)}{t} = \hat{\lambda} = \sum_{j \in E} \pi(j) \lambda(j). \quad (29)$$

It follows from Fischer and Meier-Hellstern [36], the expected number of arrivals until time t is

$$E_i(N_t) = \hat{\lambda}t + \sum_{j \in E} \left([e^{At} - I] [A + \Pi]^{-1} \right)(i, j) \lambda(j) \quad (30)$$

where $\Pi(i, j) = \pi(j)$.

Markov Modulated Bernoulli Process

We now consider discrete-time models for systems observed periodically at discrete time points. The system survives each period with a probability that depends on the state of the prevailing environment in that period. Since each period ends with a failure or survival, one can model this system as a Bernoulli process where the success probability is modulated by the environmental process. Using this setup with a Markovian environmental process, Özekici [35] focuses on probabilistic modeling and provides a complete transient and ergodic analysis. We assume throughout the following discussion the sequence of environmental states $\{Y_t; t \in I\}$ is a MC with some transition matrix $\mathbf{P}$ on a discrete state space E .

Consider a system observed periodically at times $t = 1, 2, \dots$, and the state of the system at time t is described by a Bernoulli random variable

$$X_t = \begin{cases} 1, & \text{if system is not functioning at time } t \\ 0, & \text{if system is functioning at time } t. \end{cases}$$

Given that the environment is in some state i at time t , the probability of failure in the period is

$$P(X_t = 1 | Y_t = i) = \pi(i) \quad (31)$$

for some $0 \leq \pi(i) \leq 1$. The states of the system at different points in time constitute a Bernoulli process $X = \{X_t; t = 1, 2, \dots\}$ where the success probability is a function of the environmental process Y .

Given the environmental process Y , the random quantities $X_1, X_2, \dots$, represent a conditionally independent sequence, that is,

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | Y) \\ = \prod_{k=1}^n P(X_k = x_k | Y). \end{aligned} \quad (32)$$

In the above setup, the reliability of the system is modulated by the environmental process Y , which is assumed to be an MP and thus the model is referred to as the *Markov*

modulated Bernoulli process (MMBP). If the system fails in a period, then it is replaced immediately by an identical one at the beginning of the next period. It may be possible to think of the environmental process Y as a random mission process such that Y_t is the t th mission to be performed. The success and failure probabilities depend on the mission itself. If the device fails during a mission, then the next mission will be performed by a new and identical device.

If we denote the lifetime of the system by L , then the conditional life distribution is

$$\begin{aligned} P(L = m | Y) \\ = \begin{cases} \pi(Y_1), & \text{if } m = 1 \\ \pi(Y_m) \prod_{j=1}^{m-1} (1 - \pi(Y_j)), & \text{if } m \geq 2. \end{cases} \end{aligned} \quad (33)$$

Note that if $\pi(i) = \pi$ for all $i \in E$, that is, the system reliability is independent of the environment, then Equation (33) is simply the geometric distribution $P(L = m | Y) = \pi(1 - \pi)^{m-1}$. We can also write

$$\begin{aligned} P(L > m | Y) &= (1 - \pi(Y_1))(1 - \pi(Y_2)) \dots \\ &\quad (1 - \pi(Y_m)), \end{aligned} \quad (34)$$

for $m \geq 1$.

We represent the initial state of the MC by Y_1 , rather than Y_0 , as it is customarily done in the literature, so that it represents the first environment that the system operates in. Thus, most of our analysis and results will be conditional on the initial state Y_1 of the MC. Therefore, for any event A and random variable Z we set $P_i(A) = P(A | Y_1 = i)$ and $E_i(Z) = P(Z | Y_1 = i)$ to express the conditioning on the initial state.

As pointed out by Özekici [35], the life distribution satisfies the recursive expression

$$\begin{aligned} P_i(L > m + 1) \\ = (1 - \pi(i)) \sum_{j \in E} P(i, j) P_j(L > m) \end{aligned} \quad (35)$$

with the obvious boundary condition $P_i(L > 0) = 1$. The survival probabilities can be explicitly computed via

$$P_i(L > m) = \sum_{j \in E} Q_0^m(i, j), \quad (36)$$

where $Q_0(i, j) = (1 - \pi(i))P(i, j)$. Using Equation (36), the conditional expected lifetime can be obtained as

$$E_i(L) = \sum_{m=0}^{+\infty} \sum_{j \in E} Q_0^m(i, j) = \sum_{j \in E} R_0(i, j), \quad (37)$$

where $R_0(i, j) = \sum_{m=0}^{+\infty} Q_0^m(i, j) = (I - Q_0)^{-1}(i, j)$ is the potential matrix corresponding to Q_0 .

APPLICATION OF HMMS IN SOFTWARE RELIABILITY

The notion of an *operational profile* was introduced by Musa *et al.* [41]. An operation is an externally initiated task performed by a system “as built”. The operational profile of any software describes how users employ the system. It is a quantitative and probabilistic characterization of how a system will be used. An operational profile is defined as a set of operations and the probabilities of their occurrence [42].

Optimal testing problems involving operational profiles are discussed in detail in Özekici and Soyer [22] and Özekici *et al.* [43]. Once testing is completed, the software is released to the users. This is done in an uncontrolled setting and the sequence of operations as well as their durations are now random. This operational process or the environmental process now modulates the parameters of the reliability model and plays a crucial role in software reliability assessment. Now, the environmental state Y_t at time t represents the operation performed by the user. The analysis of the software failure process obviously depends on the stochastic structure of the operational process.

In Özekici and Soyer [44], Y is assumed to be an MP. Briefly, this means that the sequence of operations performed is an MC and the amount of time spent on each operation is exponentially distributed. More precisely, we let X_n denote the n th operation that the system performs and T_n be the time at which the n th operation starts. Following our development, X is an MC with some transition matrix

$$P(i, j) = P(X_{n+1} = j | X_n = i) \quad (38)$$

and

$$P(T_{n+1} - T_n > t | X_n = i) = e^{-\mu(i)t}, \quad (39)$$

so that the duration of the n th operation is exponentially distributed with rate $\mu(i)$ if this operation is i . The probabilistic structure of the operational process is given by the generator $A(i, j) = \mu(i)(P(i, j) - I(i, j))$, where I is the identity matrix.

An overview of software failure models is presented in Singpurwalla and Soyer [45]. Perhaps the most important aspect of these models is related to the stochastic structure of the underlying failure process. This could be a “times-between-failures” model, which assumes that the times between successive failures follow a specific distribution whose parameters depend on the number of faults remaining in the program after the most recent failure. One of the most celebrated failure models in this group is that of Jelinski and Moranda [46] where the basic assumption is that there are a fixed number of initial faults in the software and each fault causes failures according to a Poisson process with the same failure rate. After each failure, the fault causing the failure is detected and removed with certainty so that the total number of faults in the software is decreased by one. In the present setting, the time to failure distribution for each fault in the software is exponentially distributed with parameter $\lambda(k)$ during operation k and this results in an extension of the Jelinski—Moranda model.

In dealing with software reliability, one is interested in the number of faults N_t remaining in the software at time t . Then, N_0 is the initial number of faults and the process $\{N_t; t \in \mathbb{R}^+\}$ depicts the stochastic evolution of the number of faults. If there is perfect debugging, then N decreases as time goes on, eventually to diminish to zero. Defining the bivariate process $Z_t = (Y_t, N_t)$, it follows that $Z = (Y, N)$ is an MP with discrete state space $F = E \times \{0, 1, 2, \dots\}$. As pointed out in Özekici and Soyer [44], this follows by noting that Y is an MP and N is a process that decreases by 1 after an exponential amount of time with a rate that

depends only on the state of Y . In particular, if the current state of Z is (i, n) for any $n > 0$, then the next state is either (j, n) with rate $\mu(i)P(i, j)$ or $(i, n-1)$ with rate $n\lambda(i)$. If $n = 0$, then the next state is $(j, 0)$ with rate $\mu(j)$. Note that 0 is an absorbing state for N .

This implies that the sojourn in state (i, n) is exponentially distributed with rate

$$\beta(i, n) = \mu(i) + n\lambda(i) \quad (40)$$

and the generator Q of Z is

$$Q((i, n), (j, m)) = \begin{cases} -(\mu(i) + n\lambda(i)), & j = i, m = n \\ \mu(i)P(i, j), & j \neq i, m = n \\ n\lambda(i), & j = i, m = n - 1 \end{cases} \quad (41)$$

Reliability is defined as the probability of failure free operation for a specified time. We will denote this by the function

$$\begin{aligned} R(i, n, t) &= P(L > t | Y_0 = i, N_0 = n) \\ &= P(N_t = n | Y_0 = i, N_0 = n), \end{aligned} \quad (42)$$

defined for all $(i, n) \in F$ and $t \geq 0$. Note that this is equal to the probability that there will be no arrivals until time t in a MMPP with intensity function $\hat{\lambda}(t) = n\lambda(Y_t)$. Thus, following Fischer and Meier-Hellstern [36], we can obtain the explicit formula

$$R(i, n, t) = \sum_{j \in E} \left[e^{(A - n\Lambda)t} \right]_{ij}, \quad (43)$$

where

$$e^{(A - n\Lambda)t} = \sum_{k=0}^{+\infty} \frac{t^k}{k!} (A - n\Lambda)^k \quad (44)$$

is the exponential matrix and $\Lambda(i, j) = \lambda(i)I(i, j)$.

REFERENCES

1. Singpurwalla ND. Survival in dynamic environments. *Stat Sci* 1995;10:86–103.
2. Mercer A. Some simple wear-dependent renewal processes. *J R Stat Soc Ser B* 1961;23:368–376.
3. Klein M. Inspection-maintenance-replacement schedules under Markovian deterioration. *Manage Sci* 1962;9:25–32.
4. Gaver DP. Random hazard in reliability problems. *Technometrics* 1963;5:211–226.
5. Çinlar E. Markov renewal theory. *Adv Appl Probab* 1969;1:123–187.
6. Çinlar E. Markov additive processes: I. *Z Wahrscheinlichkeitstheorie verw Geb* 1972;24:85–93.
7. Çinlar E. Markov additive processes: II. *Z Wahrscheinlichkeitstheorie verw Geb* 1972;24:95–121.
8. Çinlar E, Özekici S. Reliability of complex devices in random environments. *Probab Eng Inform Sci* 1987;1:97–115.
9. Çinlar E, Shaked M, Shanthikumar JG. On lifetimes influenced by a common environment. *Stoch Processes Appl* 1989;33:347–359.
10. Lindley DV, Singpurwalla ND. Multivariate distributions for the lifetimes of components of a system sharing a common environment. *J Appl Probab* 1986;23:418–431.
11. Singpurwalla ND, Youngren MA. Multivariate distributions induced by dynamic environments. *Scand J Stat* 1993;20:251–261.
12. Özekici S. Optimal maintenance policies in random environments. *Eur J Oper Res* 1995;82:283–294.
13. Özekici S. Stochastic hazard process. In: Cochran JJ, editor. *Wiley Encyclopedia of Operations Research and Management Science*. New York: Wiley; 2010, in this volume.
14. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
15. Karlin S, Taylor HM. A first course in stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
16. Ross SM. Stochastic processes. New York: Wiley; 1996.
17. Singpurwalla ND, Mazzuchi TA, Özekici S, Soyer R. Stochastic process models for reliability in dynamic environments. In: Rao CR, Khattree R, editors. *Volume 22, Handbook of statistics: statistics in industry*. Amsterdam: Elsevier; 2003. pp. 1109–1129.
18. Rabiner LR, Juang B-H. Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans Acoust Speech* 1985;33:1404–1412.
19. Hock M, Soyer R. A Bayesian approach to signal analysis of pulse trains. In: Colosimo BM, del Castillo E, editors. *Bayesian monitoring*.

- control and optimization. New York: CRC; 2006. pp. 215–243.
20. Guttorp P, Golinelli D, Abkowitz J. Bayesian inference in hidden stochastic two-compartment model for feline hematopoiesis. *Math Med Biol* 2006;23:153–172.
 21. Hughes JP, Bellone E, Guttorp P. A hidden Markov model for relating synoptic scale patterns to precipitation amounts. *Clim Res* 2000;15:1–12.
 22. Özekici S, Soyer R. Bayesian testing strategies for software with an operational profile. *Nav Res Log* 2001;48:747–763.
 23. Durand JB, Gaudoin O. Software reliability modelling and prediction with hidden Markov chains. *Stat Model* 2006;5:75–93.
 24. Ruggeri F, Soyer R. Advances in Bayesian software reliability modeling. In: Bedford T, editor. *Advances in mathematical modeling for reliability*. Amsterdam: IOS Press; 2008. pp. 165–176.
 25. Celeux G, Robert CP, Diebolt J. Bayesian estimation of hidden Markov chains: a stochastic implementation. *Stat Probab Lett* 1993;16:77–83.
 26. Scott SL. Bayesian methods for hidden Markov models. *J Am Stat Assoc* 2002;97:337–351.
 27. Özekici S, Soyer R. Bayesian analysis of Markov modulated Bernoulli processes. *Math Methods Oper Res* 2003;57:125–140.
 28. Cappe O, Moulines E, Ryden T. *Inference in hidden Markov models*. New York: Springer; 2005.
 29. Elliott RJ, Aggoun L, Moore JB. *Hidden markov models: estimation and control*. New York: Springer; 1994.
 30. Park C, Padgett WJ. Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. *Lifetime Data Anal* 2005;11:511–527.
 31. Berman M. Inhomogeneous and modulated gamma processes. *Biometrika* 1981;68:143–152.
 32. Abdel-Hameed M. A gamma wear process. *IEEE Trans Reliab* 1975;24:152–153.
 33. van Noortwijk JM. A survey of the application of gamma processes in maintenance. *Reliab Eng Syst Saf* 2009;94:2–21.
 34. Limnios N. Semi-Markov processes and hidden models. In: Cochran JJ, editor. *Wiley Encyclopedia of Operations Research and Management Science*. New York: Wiley; 2010, in this volume.
 35. Özekici S. Markov modulated Bernoulli process. *Math Methods Oper Res* 1997;45:311–324.
 36. Fischer W, Meier-Hellstern K. The Markov-modulated Poisson process cookbook. *Perform Eval* 1992;18:149–171.
 37. Rydén T. Parameter estimation for Markov modulated Poisson processes. *Stoch Models* 1994;10:795–829.
 38. Özekici S, Soyer R. Reliability modeling and analysis under random environments. In: Soyer R, Mazzuchi TA, Singpurwalla ND, editors. *Mathematical reliability: an expository perspective*. Boston (MA): Kluwer; 2004. pp. 249–273.
 39. Özekici S, Soyer R. Semi-Markov modulated Poisson process: Probabilistic and statistical analysis. *Math Methods Oper Res* 2006;64:125–144.
 40. Kingman JFC. On doubly stochastic Poisson processes. *Proc Camb Philos Soc* 1964;60:923–960.
 41. Iannino A, Musa JD, Okumoto K. *Software reliability: measurement, prediction, application*. New York: McGraw-Hill; 1987.
 42. Musa JD. Operational profiles in software reliability engineering. *IEEE Softw* 1993;10:14–32.
 43. Özekici S, Altınel IK, Özçelikyürek S. Testing of software with an operational profile. *Nav Res Log* 2000;47:620–634.
 44. Özekici S, Soyer R. Reliability of software with an operational profile. *Eur J Oper Res* 2003;149:459–474.
 45. Singpurwalla ND, Soyer R. Assessing the reliability of software: an overview. In: Özekici S, editor. *Volume F154, Reliability and maintenance of complex systems, NATO ASI series*. Berlin: Springer; 1996. pp. 345–367.
 46. Jelinski Z, Moranda P. Software reliability research. In: Freiberg W, editor. *Statistical computer performance evaluation*. New York: Academic Press; 1972. pp. 465–484.

MARKOV CHAINS OF THE M/G/1-TYPE

DARIO ANDREA BINI
BEATRICE MEINI
Dipartimento di Matematica,
Università di Pisa, Pisa, Italy

M/G/1-type Markov chains are a generalization of M/G/1 Markov chains obtained by replacing the one-dimensional state space of the latter with a bidimensional state space. Thus, the transition matrix of an M/G/1-type Markov chain is a block generalization of the transition matrix of an M/G/1 Markov chain.

M/G/1-type Markov chains [1] have been introduced and analyzed by Marcel F. Neuts, and model a large number of queueing problems from telecommunications, engineering, and applied sciences. The specific features of M/G/1-type Markov chains and the strong structure of the transition matrix have been further exploited by many authors. Several interesting results have been obtained both from the theoretical and computational points of view along with the design of highly efficient algorithms for computing the steady-state vector and other measures of interest. The rich mathematical structure of M/G/1-type Markov chains still provides several interesting theoretical problems and computational challenges.

The analysis of M/G/1-type Markov chains, performed through matrix analytic methods, provides very effective theoretical and computational tools to treat some structured Markov chains. This methodology provides also a systematic way of approaching problems related to both transient and steady-state results for a wide class of structured Markov renewal processes [2]. The algorithms obtained in this framework, endowed with very high efficiency, can be described both with the language of applied probability and of numerical analysis. The different points of view enable one to better

understand the deep features and the elegance of these algorithms [3].

In this article, we report the main theoretical properties and the up-to-date solution algorithms concerning discrete-time M/G/1-type Markov chains. In the section titled “Definitions” we recall the definitions and assumptions used in the sequel of the article. In the section titled “Positive Recurrence,” the necessary and sufficient conditions for positive recurrence are given. The section titled “Steady-State Vector” reports the main properties of the steady-state vector. The section titled “Transformation into QBD” describes a general reduction of an M/G/1-type Markov chain to a QBD (Quasi birth-death process). The section titled “Numerical Methods” is devoted to numerical methods for solving M/G/1-type Markov chains. We describe the classical approach of fixed-point iteration and fast methods based on the doubling strategy and acceleration techniques together with a description of the currently available software. The last section reports on further reading.

DEFINITIONS

Let $\mathbb{N} = \{0, 1, 2, \dots\}$ be the set of natural numbers, $\mathbb{N}^+$ the set of positive natural numbers, and $S_m = \{1, 2, \dots, m\}$. An M/G/1-type Markov chain is a two-dimensional process (X_n, φ_n) on the state space $E = (\{0\} \times S_{m_0}) \cup (\mathbb{N}^+ \times S_m)$, with transition matrix

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \dots \\ C & A_0 & A_1 & A_2 & \ddots \\ & A_{-1} & A_0 & A_1 & \ddots \\ & & A_{-1} & A_0 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}, \quad (1)$$

where B_0 is an $m_0 \times m_0$ matrix, $B_i, i \geq 1$ are $m_0 \times m$ matrices, C is an $m \times m_0$ matrix, and $A_i, i \geq -1$ are $m \times m$ matrices such that P

is stochastic. Each state of the Markov chain is formed by a pair (i, j) where $i \in \mathbb{N}^+$ and $j \in S_m$, or $i = 0$ and $j \in S_{m_0}$; moreover, $\mathbb{N}$ is called *level set* and S_m, S_{m_0} are called *phase sets*.

In particular, one has

$$(B_0)_{jj'} = P[X_{n+1}=0, \varphi_{n+1}=j' | X_n=0, \varphi_n=j], \\ j, j' \in S_{m_0},$$

$$(C)_{jj'} = P[X_{n+1}=0, \varphi_{n+1}=j' | X_n=1, \varphi_n=j], \\ j \in S_m, \quad j' \in S_{m_0},$$

$$(A_k)_{jj'} = P[X_{n+1}=i+k, \varphi_{n+1}=j' | X_n=i, \\ \varphi_n=j], \quad (i, k) \neq (1, -1), \\ k \geq -1, \quad i \geq 1, \quad j, j' \in S_m,$$

$$(B_k)_{jj'} = P[X_{n+1}=k, \varphi_{n+1}=j' | X_n=0, \varphi_n=j], \\ k \geq 1, \quad j \in S_{m_0}, j' \in S_m.$$

In other words, if the Markov chain is in the nonboundary level $i \geq 1$, then A_{-1} gives the transition probabilities down by one level and A_k , for $k \geq 0$, gives the transition probabilities up by k levels, while B_k gives the transition probabilities up by k levels, starting from the boundary level 0.

Observe that P is an upper block Hessenberg matrix; that is, the blocks of P below the subdiagonal are zero. This structure implies that starting in level k at time n , the process may move to any higher level at time $n+1$, but it may not move in one step beyond $k-1$ toward lower levels.

Another feature of P is that, except for the blocks in the first block row and block column, the block entries of P are constant along the block diagonals. A matrix with this structure is called *block Toeplitz*. The block Toeplitz structure implies that the transition probabilities from $X_n = i$ to $X_{n+1} = j$ depend on $j-i$ only and not specifically on i and j (for i and j greater than or equal to 1).

Similarly to the discrete case, a continuous-time Markov process on the state space E with a generator matrix Q having the same block structure as P in Equation (1) is called a *continuous-time M/G/1-type* Markov process.

By using the uniformization technique [4, Section 2.8], one can easily transform the

continuous-time process defined by Q into a discrete-time M/G/1-type Markov chain with transition matrix $P = \frac{1}{c}Q + I$, where c is such that $c \geq |q_{i,i}|$, for any i , where $q_{i,i}$ are the diagonal entries of Q . In this way, theory and algorithms valid for discrete-time M/G/1-type Markov chains can be applied to the case of continuous-time Markov chains.

Throughout the article, we denote by $\mathbf{1}$ a vector with all components equal to 1 whose dimension is determined by the context.

An Example

A simple example of an M/G/1-type Markov chain comes from the model of a buffer in a telecommunications system [3]. The buffer content follows the evolution equation

$$X_{n+1} = \max(X_n + \alpha_n - 1, 0),$$

where X_n is the buffer occupancy at time n and α_n is the number of new packets that arrive during the n th transmission interval. We assume that the packet lengths are fixed and have a length of one time slot. The α_n 's depend on various factors such as the number of active customers or the congestion of the network. These external factors often are modeled as a Markov chain $\{\varphi_n\}$ on some state space S_m , and one lets the distribution of α_n depend on φ_n .

Under this assumption, the two-dimensional process $\{(X_n, \varphi_n)\}$ is a Markov chain on the state space $E = \mathbb{N} \times S$ with the transition matrix P of Equation (1), where $B_i = A_i$, $i \geq 1$, $C = A_{-1}$, $B_0 = A_{-1} + A_0$, with

$$(A_k)_{i,j} = P[\alpha_n = k+1, \varphi_{n+1}=j | \varphi_n=i] \\ \text{for } k \geq -1, i, j \in S_m,$$

being the probability that $k+1$ packets arrive during a packet transmission interval, and that the next phase is j , given that the current one is i .

A detailed analysis of other examples of this paradigm can be found in Refs 5 and 6.

Non-Skip-Free Markov Chains

A generalization that falls in the class of M/G/1-type Markov chains is the class of non-skip-free M/G/1-type Markov chains where

the number of boundary levels is $N \geq 1$. In this case, the transition matrix is given by

$$P = \begin{bmatrix} B_{0,0} & B_{0,1} & B_{0,2} & \dots \\ B_{1,0} & B_{1,1} & B_{1,2} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ B_{N-1,0} & B_{N-1,1} & B_{N-1,2} & \dots \\ A_{-N} & A_{-N+1} & A_{-N+2} & \dots \\ & A_{-N} & A_{-N+1} & \ddots \\ & & A_{-N} & \ddots \\ 0 & & & \ddots \end{bmatrix},$$

where the blocks $B_{i,j}$ and A_i have size $m \times m$. We may view this matrix as having the structure (Eq. 1) with blocks of size Nm , corresponding to what we might call “macro levels”:

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & \dots \\ A_{-1} & A_0 & A_1 & \dots \\ & A_{-1} & A_0 & \ddots \\ 0 & & \ddots & \ddots \end{bmatrix},$$

where

$$A_i = \begin{bmatrix} A_{iN} & A_{iN+1} & \dots & A_{iN+N-1} \\ A_{iN-1} & A_{iN} & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{iN+1} \\ A_{iN-N+1} & \dots & A_{iN-1} & A_{iN} \end{bmatrix},$$

for $i = -1, 0, 1, \dots$, where we assume that $A_j = 0$ if $j < -N$, and

$$B_i = \begin{bmatrix} B_{0,Ni} & B_{0,Ni+1} & \dots & B_{0,Ni+N-1} \\ B_{1,Ni} & B_{1,Ni+1} & \dots & B_{1,Ni+N-1} \\ \vdots & \vdots & \ddots & \vdots \\ B_{N-1,Ni} & B_{N-1,Ni+1} & \dots & B_{N-1,Ni+N-1} \end{bmatrix},$$

$i \geq 0$.

An instance of a non-skip-free Markov chain is given by an M/G/1 model with batch services where at most a batch of size $N < +\infty$ is taken into service with probability p_k for a batch of size k .

Assumptions

The following assumptions on the M/G/1-type Markov chain with transition matrix (Eq. 1) will be useful.

Condition 1. The Markov chain with transition matrix (Eq. 1) is irreducible and aperiodic.

Condition 2. In addition to being stochastic, the following matrix is irreducible:

$$A = \sum_{i=-1}^{+\infty} A_i.$$

The above assumption is not restrictive. In fact, if A is reducible, it is possible to consider A in its reduced form. This way, the problem can be split into a set of smaller irreducible problems. Details on this strategy can be found in Section 3.5 of Ref. 1 and in Refs 7, 8.

Consider the Markov chain defined on the set of states $\mathbb{Z} \times S_m$, where $\mathbb{Z}$ is the set of all integers, with transition matrix

$$P' = \begin{bmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & A_0 & A_1 & A_2 & A_3 & \ddots \\ & A_{-1} & A_0 & A_1 & A_2 & \ddots \\ & & A_{-1} & A_0 & A_1 & \ddots \\ & & & A_{-1} & A_0 & \ddots \\ 0 & & & & \ddots & \ddots \end{bmatrix}; \quad (2)$$

that is, we remove the boundary at level 0 and allow the Markov chain to move freely to any negative level, as well as to the positive ones.

Condition 3. The Markov chain on $\mathbb{Z} \times S_m$ with transition matrix (Eq. 2) has only one final class $\mathbb{Z} \times S$, where $S \subseteq S_m$. Every other state is on a path to the final class.

A final class is a set of states S such that there is no path going out of S , and for any pair of states $i, j \in S$ there is a path connecting i to j . We refer the reader to Ref. 3 (Chapter 4) for a discussion on the above assumptions.

POSITIVE RECURRENCE

The peculiar structure of the matrix P enables one to provide simple conditions for the positive recurrence of the Markov chain. For Markov chains, recurrence properties are governed by almost sure returns to the states, and for M/G/1-type chains, studying returns involves considering the first entrance to lower level states.

Here and hereafter we assume that Condition 1 holds.

Assume that $C = A_{-1}$ and define θ as the first return time to the level 0:

$$\theta = \min\{n \geq 0 : X_n = 0\}$$

and consider the matrix $G = (G_{j,j'})$, with entries

$$G_{j,j'} = P[\theta < \infty, \varphi_\theta = j' | X_0 = 1, \varphi_0 = j]$$

being the probability that, starting from the state $(1, j)$ in level 1, the process reaches down to the level 0 in a finite time, and that $(0, j')$ is the first state visited in level 0.

Observe that [3] the matrix G is the matrix of first passage probabilities from level $k + 1$ to level k , independently of $k \geq 0$. The matrix G is strictly related to the matrix equation

$$X = \sum_{i=-1}^{+\infty} A_i X^{i+1}, \quad (3)$$

where the unknown X is an $m \times m$ matrix, as stated by the following result [3, Theorem 4.3]:

Theorem 1. *The matrix G is the minimal nonnegative solution of Equation (3), that is, G satisfies Equation (3) and $G \leq Y$ for any other nonnegative solution Y of Equation (3). If the Markov chain is recurrent, then G is stochastic; otherwise, it is substochastic with $G\mathbf{1} \leq \mathbf{1}$, $G\mathbf{1} \neq \mathbf{1}$.*

The following result [3, Theorem 4.7] gives the necessary and sufficient conditions for positive recurrence:

Theorem 2. *Assume that Conditions 1 and 2 hold and that $\sum_{i=-1}^{+\infty} (i+1)A_i < +\infty$, so*

that we may define the vector $\mathbf{a} = \sum_{i=-1}^{+\infty} iA_i\mathbf{1}$; moreover, set

$$\mu = \boldsymbol{\alpha}^T \mathbf{a}, \quad (4)$$

where $\boldsymbol{\alpha}$ is the vector such that $\boldsymbol{\alpha}^T A = \boldsymbol{\alpha}^T$, $\boldsymbol{\alpha}^T \mathbf{1} = 1$, for $A = \sum_{i=-1}^{+\infty} A_i$. Then the Markov chain with transition matrix (Eq. 1) is recurrent if and only if $\mu \leq 0$ and transient if and only if $\mu > 0$. It is positive recurrent if and only if $\mu < 0$ and $\mathbf{b} = \sum_{i=1}^{+\infty} iB_i\mathbf{1} < \infty$.

For the proof of the above result we refer the reader to Ref. 9, Proposition 4.6, Ref. 1, Theorem 1.3.1, or Ref. 10, Corollary 3. Theorem 2 is simple to interpret. Assume that the state is (n, i) at some time t ; at time $t + 1$, it will be (n', j) , with n' and j random. The component a_i of the vector $\mathbf{a}$ is the expectation of the jump size $n' - n$, when the phase is i , and μ is obtained by averaging over the phases with the probability distribution $\boldsymbol{\alpha}$. If $\mu < 0$, then the jumps are negative on average and the process is attracted to the level 0 at the bottom of the state space. The number μ defined in Equation (4) is called *drift* of the Markov chain.

Conditions for positive recurrence can be given also in terms of the location of the zeros of the function $a(z) = \det(zI - A(z))$, as proved in Ref. 10 (see also Ref. 3, Theorem 4.9).

Theorem 3. *Assume that Condition 2 holds, and that $\mathbf{a}$ is finite. Define*

$$\kappa = \max\{k : z^{-m/k} a(z^{1/k}) \text{ is a (single valued) function in } |z| \leq 1\}.$$

Then the following properties hold:

- *if $\mu < 0$, then $a(z)$ has $m - \kappa$ zeros in the open unit disk, and κ simple zeros on the unit circle at the κ th roots of 1;*
- *if $\mu = 0$ and $\sum_{i=-1}^{+\infty} (i+1)^2 A_i$ is finite, then $a(z)$ has $m - \kappa$ zeros in the open unit disk, and κ zeros of multiplicity 2 on the unit circle at the κ th roots of 1;*
- *if $\mu > 0$, then $a(z)$ has m zeros in the open unit disk, and κ simple zeros on the unit circle at the κ th roots of 1.*

The eigenvalues of G are a special subset of the zeros of $a(z)$, as shown in Ref. 11 (see also 3, Theorem 4.10)

Theorem 4. *Under the hypotheses of Theorem 3 the eigenvalues of G are:*

- the zeros of $a(z)$ in the closed unit disk if $\mu < 0$;
- the zeros of $a(z)$ in the open unit disk and the κ th roots of 1 if $\mu = 0$;
- the zeros of $a(z)$ in the open unit disk if $\mu > 0$.

Observe that, if $G\mathbf{u} = \lambda\mathbf{u}$, then $(\lambda I - A(\lambda))\mathbf{u} = 0$. If λ is the spectral radius of G , then under Condition 2 there exists a positive vector $\mathbf{u}$ such that $G\mathbf{u} = \lambda\mathbf{u}$ (see Ref. 1, Lemma 2.3.1).

Reducing Transient to Positive Recurrent Markov Chains

In developing the theory, it is customary to restrict oneself to positive recurrent Markov chains. An extension to the transient case is enabled by the following result which shows that, to any transient Markov chain of the M/G/1-type, one can associate a closely related chain that is positive recurrent.

More precisely, the following property derives from the results of Refs 12–14.

Theorem 5. *Assume that Condition 2 holds and that the Markov chain with transition matrix (Eq. 1) is transient. Let ξ be the spectral radius of G , and $\mathbf{u} = (u_i)$ a positive vector such that $G\mathbf{u} = \xi\mathbf{u}$ where G is the minimal nonnegative solution of Equation (3). Define the following matrices*

$$C_i = \xi^i D^{-1} A_i D, \quad i = -1, 0, 1, 2, \dots,$$

for $D = \text{Diag}(\mathbf{u})$ being the diagonal matrix with diagonal entries $u_1, u_2, \dots, u_m$. Then C_i are nonnegative matrices such that $\sum_{i=-1}^{+\infty} C_i$ is stochastic, $H = \xi^{-1} D^{-1} G D$ is the minimal nonnegative solution of the matrix equation

$$X = \sum_{i=-1}^{+\infty} C_i X^{i+1},$$

and the spectral radius of H is 1. Moreover, the Markov chain defined by the transition matrix P of Equation (1) where the blocks A_i are replaced with the blocks C_i , is positive recurrent if $\sum_{i=1}^{+\infty} i B_i$ is finite.

From GI/M/1 to M/G/1

Another major structural class of Markov chains besides the M/G/1-type is the set of Markov chains of the GI/M/1-type. There exists a set of duality results [15] that allow GI/M/1-type chains to be studied in terms of closely related M/G/1-type chains. These results are useful algorithmically also in that certain key quantities like the R matrix (see Refs 3, 16, 17) can be directly expressed in terms of the better behaved G matrix of the associated M/G/1-type chain.

Let

$$P = \begin{bmatrix} B_0 & A_1 & & & 0 \\ B_{-1} & A_0 & A_1 & & \\ B_{-2} & A_{-1} & A_0 & A_1 & \\ B_{-3} & A_{-2} & A_{-1} & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix} \quad (5)$$

be the transition matrix of a GI/M/1-type Markov chain, where we assume that $\sum_{i=-1}^{+\infty} A_{-i}$ is irreducible and stochastic.

Let R be the minimal nonnegative solution of the matrix equation

$$X = \sum_{i=-1}^{+\infty} X^{i+1} A_{-i}.$$

Define $D = \text{diag}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the strictly positive stationary probability vector of A and define $\tilde{A}_i = D^{-1} A_{-i}^T D$, for $i = -1, 0, 1, \dots$. Note that the matrix $\tilde{A} = \sum_{i=-1}^{+\infty} \tilde{A}_i$ is stochastic and has the same stationary probability vector $\boldsymbol{\alpha}$ as A . Then $G = D^{-1} R^T D$ is the minimal nonnegative solution of

$$X = \sum_{i=-1}^{+\infty} \tilde{A}_i X^{i+1}, \quad (6)$$

which is the characteristic equation of the M/G/1-type Markov chain with

transition matrix

$$P = \begin{bmatrix} \tilde{B}_0 & \tilde{B}_1 & \tilde{B}_2 & \tilde{B}_3 & \dots \\ \tilde{A}_{-1} & \tilde{A}_0 & \tilde{A}_1 & \tilde{A}_2 & \dots \\ & \tilde{A}_{-1} & \tilde{A}_0 & \tilde{A}_1 & \ddots \\ & & \tilde{A}_{-1} & \tilde{A}_0 & \ddots \\ 0 & & & \ddots & \ddots \end{bmatrix}, \quad (7)$$

where $\tilde{B}_i = D^{-1}B_{-i}^T D$, for $i = 0, 1, 2, \dots$. Furthermore, it is easy to see that the drift $\tilde{\mu} = \alpha^T \sum_{i=-1}^{+\infty} (i+1)\tilde{A}_i \mathbf{1}$ of the new Markov chain is equal to the drift δ of the Markov chain (5). In this way, a positive recurrent (transient) GI/M/1-type Markov chain is equivalent to a transient (positive recurrent) M/G/1-type Markov chain.

STEADY-STATE VECTOR

Throughout this section, we assume that Conditions 1 and 2 hold and that the Markov chain with transition matrix (Eq. 1) is positive recurrent. We denote by $\pi \in \mathbb{R}^N$ the stationary probability vector, that is, the unique solution of the system

$$\pi^T = \pi^T P, \quad \pi^T \mathbf{1} = 1,$$

and we partition it as $\pi = (\pi_n)_{n=0,1,\dots}$, with $\pi_0 = (\pi_{0,j})_{j=1,\dots,m_0}$, $\pi_n = (\pi_{n,j})_{j=1,\dots,m}$, $n \geq 1$, and $\pi_{n,j} = \lim_{t \rightarrow +\infty} P[X_t = n, \varphi_t = j | X_0 = n', \varphi_0 = j']$, independently of n' and j' .

The next theorem gives a recursive expression for the steady-state vector, known as *Ramaswami's formula*. This formula has been proved for the first time in Ref. 18 (see also Refs 1, Theorem 3.2.5 and 3, Theorem 4.4) by using probabilistic arguments, and interpreted in terms of algebraic properties in Refs 19 and 20 (see also 3, Section 4.5)

Theorem 6. *Let*

$$\begin{aligned} A_n^* &= \sum_{i=n}^{+\infty} A_i G^{i-n}, \\ B_n^* &= \sum_{i=n}^{+\infty} B_i G^{i-n}, \quad \text{for } n \geq 0. \end{aligned} \quad (8)$$

Then $I - A_0^*$ is nonsingular,

$$\begin{aligned} \pi_n^T &= \left(\pi_0^T B_n^* + \sum_{i=1}^{n-1} \pi_i^T A_{n-i}^* \right) (I - A_0^*)^{-1}, \\ &\text{for } n \geq 1, \end{aligned} \quad (9)$$

and π_0 is the unique solution of the system

$$\begin{aligned} \pi_0^T B_0^* &= \pi_0^T, \\ \pi_0^T \mathbf{b} - \mu \pi_0^T \mathbf{1} + \pi_0^T (I - B) W \mathbf{a} &= -\mu, \end{aligned}$$

where $B = \sum_{n=0}^{+\infty} B_n$ and $W = (I - A + \mathbf{1} \mathbf{a}^T)^{-1} - \mathbf{e} \mathbf{a}^T$.

The following result [3, Theorem 4.13], relating the generating functions of the sequences $\{A_i\}$ and $\{A_i^*\}$, is the basis of the algebraic interpretation of Ramaswami's formula:

Theorem 7. *Under the hypotheses of Theorem 3, the Laurent matrix power series $I - S(z)$, $S(z) = \sum_{i=-1}^{+\infty} z^i A_i$, has the following factorization*

$$\begin{aligned} I - S(z) &= U(z) L(z), \quad |z| = 1, \\ U(z) &= I - \sum_{i=0}^{+\infty} z^i A_i^*, \quad L(z) = I - z^{-1} G, \end{aligned}$$

where A_i^* , for $i \geq 0$, are defined in Equation (8). Moreover,

- if $\mu < 0$ then $U(z)$ is nonsingular for $|z| = 1$, $L(z)$ is singular for $z = \omega_\kappa^i$, $i = 0, \dots, \kappa - 1$, and the coefficients of $L(z)^{-1}$ are uniformly bounded in norm by a constant;
- if $\mu = 0$ and $\sum_{i=-1}^{+\infty} (i+1)^2 A_i$ is finite, then both $L(z)$ and $U(z)$ are singular for $z = \omega_\kappa^i$, $i = 0, \dots, \kappa - 1$ and the coefficients of $L(z)^{-1}$ are uniformly bounded in norm by a constant;
- if $\mu > 0$ then $L(z)$ is nonsingular for $|z| = 1$, $U(z)$ is singular for $z = \omega_\kappa^i$, $i = 0, \dots, \kappa - 1$.

Qualitative properties of π , like the exponential decay of its components, have been proved in Refs 21 and 22.

TRANSFORMATION INTO QBD

An M/G/1-type Markov chain can be transformed into a suitable QBD process having an infinite number of phases as shown in Ref. 23. For notational simplicity, we assume that $m_0 = m$.

In the transition matrix (Eq. 1), we replace the simple transitions of the form $(n, i) \rightarrow (n + k, j)$ by a more elaborate virtual procedure, whereby a new phase j is chosen first, as well as a number k of steps upward, and only then the level is increased, one unit at a time, until level $n + k$ is finally reached. Graphically, we have

$$\begin{aligned} (n, i) &\rightarrow (n; k, j) \rightarrow (n + 1; k - 1, j) \cdots \\ &\rightarrow (n + k - 1; 1, j) \rightarrow (n + k, j). \end{aligned}$$

At each transition of this virtual sequence, one keeps track of the number of steps that remain until the final level is reached. In order to have homogeneous notation, we shall write (n, i) as $(n; 0, i)$.

We define in this way a three-dimensional process $\{X_t, K_t, \varphi_t\}_{t=0,1,\dots}$ on the state space $\mathbb{N} \times \mathbb{N} \times S_m$ for which the pair (K_t, φ_t) constitutes the phase. We also write that K_t is the *sublevel*. The transition matrix of the new process is

$$P = \begin{bmatrix} \mathcal{B} & \mathcal{A}_1 & & 0 \\ \mathcal{A}_{-1} & \mathcal{A}_0 & \mathcal{A}_1 & \\ & \mathcal{A}_{-1} & \mathcal{A}_0 & \ddots \\ 0 & & \ddots & \ddots \end{bmatrix}, \quad (10)$$

with

$$\begin{aligned} \mathcal{A}_{-1} &= \begin{bmatrix} \mathcal{A}_{-1} & 0 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \\ \mathcal{A}_0 &= \begin{bmatrix} \mathcal{A}_0 & \mathcal{A}_1 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \\ \mathcal{A}_1 &= \begin{bmatrix} 0 & 0 \\ I & 0 \\ & I & \ddots \\ 0 & & \ddots \end{bmatrix}, \end{aligned}$$

and

$$\mathcal{B} = \begin{bmatrix} B_0 & B_1 & \cdots \\ 0 & 0 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

The following result relates the QBD 10 to the M/G/1-type Markov chain 1 (see Ref. 3, Theorem 5.21).

Theorem 8. *The minimal nonnegative solution $\mathcal{G}$ of the matrix equation*

$$\mathcal{X} = \mathcal{A}_{-1} + \mathcal{A}_0 \mathcal{X} + \mathcal{A}_1 \mathcal{X}^2$$

associated with the QBD (Eq. 10) is

$$\mathcal{G} = \begin{bmatrix} G & 0 & 0 & \cdots \\ G^2 & 0 & 0 & \cdots \\ G^3 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

NUMERICAL METHODS

Once the minimal nonnegative solution G to the matrix equation (3) has been computed, the invariant probability vector of an M/G/1-type Markov chain can be obtained by means of formula (9).

In certain applications, the blocks A_i may have some special structure which is inherited by G . A meaningful case is when $\mathcal{A}_{-1}$ is a rank-one matrix, that is, $\mathcal{A}_{-1} = \mathbf{u}\mathbf{v}^T$ for some vectors $\mathbf{u}, \mathbf{v}$ [8,7]. It is possible to prove that in the positive recurrent case $G = \frac{1}{v^T \mathbf{1}} \mathbf{1}\mathbf{v}^T$, that is G has rank one. In general, an explicit formula for G cannot be given and numerical methods for its computation must be provided.

Therefore in this section we focus our attention on the computation of G .

Fixed-Point Iterations

In this section, we report results concerning the approximation of the minimal nonnegative solution of Equation (3) by means of fixed-point iterations. For more details, we refer the reader to Ref. 3, Chapter 6 and to Refs 24–28.

A natural way for generating a matrix sequence converging to a solution of Equation

(3) is given by the following fixed-point iteration known in the literature as *natural iteration*:

$$X_{k+1} = \sum_{i=-1}^{+\infty} A_i X_k^{i+1}, \quad k = 0, 1, \dots \quad (11)$$

starting from a given initial approximation X_0 .

Indeed, if the sequence $\{X_k\}$ has a limit X^* , then X^* is a solution of Equation (3).

By rewriting the original Equation (3) in the form

$$X = (I - A_0)^{-1} \left(A_{-1} + \sum_{i=1}^{+\infty} A_i X^{i+1} \right),$$

one obtains the so-called *traditional iteration*

$$X_{k+1} = (I - A_0)^{-1} \left(A_{-1} + \sum_{i=1}^{+\infty} A_i X_k^{i+1} \right), \quad k = 0, 1, \dots \quad (12)$$

In a similar way, one obtains the iteration

$$X_{k+1} = \left(I - \sum_{i=0}^{+\infty} A_i X_k^i \right)^{-1} A_{-1}, \quad k = 0, 1, \dots, \quad (13)$$

which is known as *U-based iteration*. The name is motivated by the fact that the matrix $\sum_{i=0}^{+\infty} A_i G^i$, which appears in the right-hand side of Equation (13) with $X_k = G$, is customarily denoted by U .

The following convergence results hold:

Theorem 9. *For $X_0 = 0$ the iterations (11), (12), (13), converge monotonically to G . Denoting X_k^N , X_k^T and X_k^U the sequences generated by the natural, traditional and U-based iterations, respectively, it holds*

$$X_k^N \leq X_k^T \leq X_k^U,$$

where the inequalities are meant component-wise. If $X_0 = I$, the sequences generated by the iterations (11), (12), (13), converge to a stochastic solution G^* of Equation (3). In the recurrent case $G^* = G$.

It has been proved in Ref. 28 that in the recurrent case each one of the three sequences generated with $X_0 = I$ converges to G faster than the corresponding sequence obtained with $X_0 = 0$. That is, the error reduction per step is larger for $X_0 = I$.

The iterations based on Equations (11), (12), and (13) have a linear convergence in the transient or in the positive recurrent case. That is, for any norm there exists, independent of the norm, the limit $\limsup_k \|G - X_k\|^{1/k} = \theta$, and $0 < \theta < 1$ holds. In other words, the error $\|G - X_k\|$ converges to zero exponentially as θ^k . In the null-recurrent case with $X_0 = 0$, it holds $\theta = 1$, which means that the error $\|G - X_k\|$ does not tend to zero exponentially and convergence slows down much. This kind of convergence is called *sublinear*.

A quadratically convergent iteration, that is, such that $\limsup_k \|G - X_k\|^{1/2^k} < 1$, can be obtained by applying Newton's method to the equation

$$X - \left(I - \sum_{i=0}^{+\infty} A_i X^i \right) A_{-1} = 0,$$

which generates the sequence

$$X_{k+1} = X_k - H_k, \quad k = 0, 1, 2, \dots, \quad (14)$$

where H_k solves the linear matrix equation in the unknown H

$$\begin{aligned} H - U_k \left(\sum_{i=1}^{+\infty} A_i \right) \left(\sum_{j=0}^{i-1} X_k^j H X_k^{i-1-j} \right) U_k A_{-1} \\ = X_k - U_k A_{-1}, \quad U_k = \left(I - \sum_{i=0}^{+\infty} A_i X_k^i \right)^{-1}. \end{aligned}$$

More details on this method can be found in Refs 24, 27, 29.

In the positive recurrent case, if $0 \leq X_0 \leq G$, then the sequence X_k generated by 14 converges monotonically to G with quadratic convergence speed.

Even though the speed of convergence is higher, the cost of a single step of Newton's iteration is much larger than the cost of a single step of one of the three iterations (11), (12), and (13). This is due to the large cost needed for computing the matrix H_k .

Doubling Methods

Doubling methods have the nice feature to provide the quadratic convergence speed while still maintaining a low cost per step.

Cyclic reduction (CR) is one of the most efficient doubling algorithms. It is based on a technique introduced in the 1970s by Hockney [30] and Buzbee *et al.* [31] for solving special linear systems and specialized to M/G/1-type Markov chains in Refs 3, Chapter 7, 12, 32, 33. In the context of Markov chains, the idea on which CR is based is to divide the set of levels into two groups: the ones with odd indices and the ones with even indices, and to apply censoring. The new Markov chain obtained in this way is still of the M/G/1-type so that the process can be applied cyclically. CR can be viewed as an extension to M/G/1-type Markov chains of the logarithmic reduction algorithm by Latouche and Ramaswami [34], designed for the solution of QBDs. Comparisons and relationships between CR and logarithmic reduction can be found in Refs 3, Chapter 7, 12, 35.

A natural way for rigorously describing CR is to introduce the matrix power series $A(z) = \sum_{i=-1}^{+\infty} z^{i+1} A_i$. Starting from $A^{(0)}(z) = A(z)$ and $\hat{A}^{(0)}(z) = \sum_{i=0}^{+\infty} z^i A_i$, CR generates the following two sequences of matrix power series $A^{(n)}(z) = \sum_{i=-1}^{+\infty} z^{i+1} A_i^{(n)}$, and $\hat{A}^{(n)}(z) = \sum_{i=0}^{+\infty} z^i \hat{A}_i^{(n)}$ recursively defined by

$$\begin{aligned} A^{(n+1)}(z) &= z A_{\text{odd}}^{(n)}(z) + A_{\text{even}}^{(n)}(z) (I \\ &\quad - A_{\text{odd}}^{(n)}(z))^{-1} A_{\text{even}}^{(n)}(z), \\ \hat{A}^{(n+1)}(z) &= \hat{A}_{\text{even}}^{(n)}(z) + \hat{A}_{\text{odd}}^{(n)}(z) (I \\ &\quad - A_{\text{odd}}^{(n)}(z))^{-1} A_{\text{even}}^{(n)}(z), \end{aligned} \quad (15)$$

where for a generic matrix power series $F(z) = \sum_{i=0}^{+\infty} z^i F_i$ we define

$$\begin{aligned} F_{\text{even}}(z) &= \frac{1}{2} (F(\sqrt{z}) + F(-\sqrt{z})) = \sum_{i=0}^{+\infty} z^i F_{2i}, \\ F_{\text{odd}}(z) &= \frac{1}{2\sqrt{z}} (F(\sqrt{z}) - F(-\sqrt{z})) = \sum_{i=0}^{+\infty} z^i F_{2i+1}. \end{aligned}$$

We have the following properties [3]:

Theorem 10. *Assume that Conditions 1 and 3 hold. Moreover, if the Markov chain*

with transition matrix (Eq. 1) is positive recurrent assume also that $A(z)$ is analytic in a disk of radius greater than 1 and that there exists a zero of $a(z) = \det(zI - A(z))$ of modulus greater than 1. If the Markov chain is positive recurrent or transient, then for the matrix power series $A^{(n)}(z)$ and $\hat{A}^{(n)}(z)$ generated by CR one has:

1. *the matrices $A_i^{(n)}$, $\hat{A}_{i+1}^{(n)}$, $i = -1, 0, 1, \dots$, are nonnegative and such that $A^{(n)}(1)$, $A_{-1} + \hat{A}^{(n)}(1)$ are stochastic;*
2. *the matrices $I - \hat{A}_0^{(n)}$ are nonsingular for $n = 0, 1, 2, \dots$ and such that*

$$G = (I - \hat{A}_0^{(n)})^{-1} \times \left(A_{-1} + \sum_{i=1}^{+\infty} \hat{A}_i^{(n)} G^{i \cdot 2^n + 1} \right);$$

3. *for any matrix norm $\|\cdot\|$ there exists a constant γ such that*

$$\begin{aligned} \|G - G_n\| &\leq \gamma \sigma^{-2^n}, \\ G_n &= (I - \hat{A}_0^{(n)})^{-1} A_{-1} \end{aligned}$$

for any σ such that $1 < \sigma < \xi$, where in the positive recurrent case ξ is the smallest real number greater than 1 such that $a(\xi) = 0$, while in the transient case $\xi = \theta^{-1}$ with θ being the largest real number less than 1 such that $a(\xi) = 0$.

The convergence of G_n to G described in part 3 of the above theorem is said to be quadratic with rate ξ^{-1} . In the case of null-recurrent Markov chains, part 1 of the above theorem still holds but convergence turns linear.

If the M/G/1-type Markov chain reduces to a QBD then the cyclic reduction step can be implemented by means of a few matrix operations without involving matrix power series. In this case, CR shares common properties with the algorithm of logarithmic reduction of Ref. 34.

For general M/G/1-type Markov chains, the functional formulation (15) of CR enables one to provide an effective implementation of each step by means of the evaluation/interpolation technique at the

set of Fourier points. More details on this algorithm can be found in Refs 3, Chapter 7 and 33. With this technique, the matrix power series are numerically approximated by matrix polynomials, and fast polynomial arithmetic based on the fast Fourier transform (FFT) is applied. Owing to the quadratic convergence, the matrix power series generated at the generic step n of CR are well approximated by polynomials whose degree decreases to zero in a few steps.

A different implementation, given in matrix form, but still based on FFT, can be found in Ref. 32. Advanced software implementing CR and other effective algorithms can be found in Refs 36 and 37.

An alternative approach is to apply CR to the QBD process obtained by means of the transformation described in the section titled "Transformation into QBD." Indeed, as already pointed out, applying CR to a QBD does not require the matrix power series machinery; on the other hand, managing with blocks of infinite size makes the implementation of the CR step more tricky. A discussion on this approach can be found in Refs 3, Section 9.3 and 38.

A different doubling method, described in Refs 39 and 40 and improved in Refs 3 and 41, is based on the successive solution of the sequence of block Hessenberg systems

$$\begin{bmatrix} I - A_0 & -A_1 & \dots & -A_{n-1} \\ -A_{-1} & I - A_0 & \ddots & \vdots \\ & \ddots & \ddots & -A_1 \\ 0 & & -A_{-1} & I - A_0 \end{bmatrix} \begin{bmatrix} X_1^{(n)} \\ X_2^{(n)} \\ \vdots \\ X_n^{(n)} \end{bmatrix} = \begin{bmatrix} A_{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (16)$$

Indeed, it is possible to prove that for any norm $\|\cdot\|$ there exists a constant $\gamma > 0$ such that

$$\|G - X_1^{(n)}\| \leq \gamma \sigma^{-n}$$

for any σ such that $1 < \sigma < \xi$, where ξ is defined in Theorem 10.

The computation of $X_1^{(n)}$ can be performed by means of a doubling technique by solving the system (Eq. 16) for $n = 2^k$, with $k = 1, 2, \dots$. In fact, the solution of Equation (16) with $n = 2^k$ can be easily computed from the solution of Equation (16) with $n = 2^{k-1}$. The numerical performances of this technique are generally inferior to those of CR.

Other doubling methods, based on the computation of an invariant subspace of a suitable matrix, have been introduced in Ref. 42. These algorithms are quadratically convergent but may suffer from numerical instabilities. The reader can find more details on this class of algorithms in Refs 3 [Section 8.6], 35, 43.

Acceleration Techniques

In the case of null-recurrent and close to null-recurrent Markov chains, the convergence speed of the algorithms described in the previous sections slows down significantly. A way to overcome this drawback is to apply the *shift technique* introduced in Ref. 44 and improved in Ref. 45.

The idea is to replace the original matrix equation with a new one

$$X = \sum_{i=-1}^{+\infty} \tilde{A}_i X^{i+1} \quad (17)$$

for suitable matrices $\tilde{A}_i$ such that the convergence of iterative methods of the previous two sections applied to the new equation is faster and the solution of the original matrix equation can be easily computed from the solution of the new one.

More precisely, we have the following result [3, Theorems 3.32, 3.33].

Theorem 11. *Assume that the Markov chain defined by the transition matrix (Eq. 1) is recurrent. Let $\mathbf{u}$ be any vector such that $\mathbf{u}^T \mathbf{e} = 1$ and define $\tilde{A}_{-1} = A_{-1} - A_{-1}Q$, $\tilde{A}_i = A_i + (I - \sum_{j=-1}^i A_j)Q$, $i = 0, 1, \dots$, with $Q = \mathbf{1u}^T$. Then the matrix equation (17) has a unique solution $\tilde{G}$ with spectral radius less than 1. Moreover, it holds $G = \tilde{G} + Q$.*

It is possible to prove that cyclic reduction as well as the doubling method based

on Equation (16) applied to Equation (17) converges quadratically with rate η/ξ , for a suitable $0 < \eta < 1$, in the positive recurrent case, and with rate $1/\xi$ in the null-recurrent case.

Also, the fixed-point iterations applied to the new Equation (17) increase their convergence.

Sometimes it is useful to shift more than one eigenvalue. The technique obtained in this way is known as *multiple shift*.

Software

Some software packages have been designed containing the implementation of the most effective methods for solving M/G/1-type Markov chains. A Matlab Toolbox called *SMCSolver* [36,37], is available at Ref. 46. It covers M/G/1-type, GI/M/1-type, QBD, and non-skip-free Markov chains. A Fortran version with a graphical interface is available at Ref. 47.

This tool computes the matrix G (or R) of M/G/1-type, GI/M/1-type, and non-skip-free Markov chains as well as its steady-state probability vector. It includes implementations of the following algorithms: cyclic reduction, fixed-point iterations, invariant subspace, and Ramaswami reduction to a QBD. The steady-state vector is computed by means of the Ramaswami formula.

Another Matlab Toolbox, called *Q-MAM*, is based on *SMCSolver* and available at [46]. It performs further computations like the queue length, waiting time, and delay distribution of various queueing systems of infinite size. It includes, amongst others, implementations of the following queueing models both in discrete and continuous-time: PH/PH/1, MAP/MAP/1, MAP/M/c, MAP/D/c, RAP/RAP/1, MMAP[K]/PH[K]/1, and others.

Another package, called *MAMSolver* and available at Ref. 48, provides solutions for both continuous- and discrete-time Markov chains of QBD, GI/M/1, and M/G/1-types. It relies on the Matlab version of *SMCSolver*, computes the matrix G by means of cyclic reduction and logarithmic reduction, and uses an alternate algorithm to the Ramaswami formula called *ETAQA* [49] for computing π .

FURTHER READING

There is a vast literature concerning M/G/1-type Markov chains and the algorithms for their solution. The pioneering monograph of M.F. Neuts [1] provides an important reference in this regard.

Two interesting (even though not recent) surveys are Refs 50 and 51. The book [4], even though concerning QBDs, contains basic definitions and properties of M/G/1-type Markov chains. An interesting tutorial that gives the Markov renewal framework and extends the theory to the infinite-dimensional case is Ref. 52.

Important collections of theoretical results, algorithm, and applications can be found in the proceedings of a series of conferences called *Matrix Analytic Methods in Stochastic Models* [53–58]. Other interesting sources of results are the Proceedings of the Dagstuhl Seminar “Numerical Methods for Structured Markov Chains” [59] and the special issue of *Stochastic Models* [60] in honor of M.F. Neuts. Algorithmic issues are developed in the book [3], and a more probabilistic oriented approach can be found in the book [61].

REFERENCES

1. Neuts MF. Structured stochastic matrices of M/G/1 type and their applications. Volume 5, Probability: pure and applied. New York: Marcel Dekker Inc.; 1989.
2. Lucantoni DM, Choudhury GL, Whitt W. The transient BMAP/G/1 queue. *Commun Stat Stoch Models* 1994;10(1):145–182.
3. Bini DA, Latouche G, Meini B. Numerical methods for structured Markov chains. Numerical Mathematics and Scientific Computation. Oxford University Press. New York; Oxford Science Publications; 2005.
4. Latouche G, Ramaswami V. Introduction to Matrix analytic methods in stochastic modeling, ASA-SIAM Series on Statistics and Applied Probability. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM); 1999.
5. Ramaswami V. The N/G/1 queue and its detailed analysis. *Adv Appl Probab* 1980;12(1):222–261.

6. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Commun Stat Stoch Models* 1991;7(1):1–46.
7. Ramaswami V, Lucantoni DM. Algorithmic analysis of a dynamic priority queue. Volume II, *Applied probability—computer science: the interface*; (Boca Raton (FL) 1981), volume 3 of *Program Computer Science*. Boston (MA): Birkhäuser Boston; 1982. pp. 157–206. With discussion by Eckberg AE.
8. Chandramouli Y, Neuts MF, Ramaswami V. A queueing model for meteor burst packet communication systems. *IEEE Trans Commun* 1989;37(10):1024–1030.
9. Asmussen S. *Applied probability and queues*. New York: Wiley; 1987.
10. Gail HR, Hantler SL, Taylor BA. Spectral analysis of $M/G/1$ and $G/M/1$ type Markov chains. *Adv Appl Probab* 1996;28(1):114–165.
11. Gail HR, Hantler SL, Taylor BA. Solutions of the basic Matrix equation for $M/G/1$ and $G/M/1$ type Markov chains. *Commun Stat Stoch Models* 1994;10(1):1–43.
12. Bini DA, Meini B, Ramaswami V. A probabilistic interpretation of Cyclic Reduction and its relationships with Logarithmic Reduction. *Calcolo* 2008;45(3):207–216.
13. Bright L. *Matrix-analytic methods in applied probability* [PhD thesis]. University of Adelaide; 1996.
14. Van Houdt B, Taylor PG. A new dual relationship between Markov chains of $GI/M/1$ and $M/G/1$ type. *Adv Appl Probab* 2010;42(1):210–225.
15. Ramaswami V. A duality theorem for the Matrix paradigms in queueing theory. *Stoch Models* 1990;6:151–161.
16. Neuts MF. Volume 2, *Matrix-geometric solutions in stochastic models: an algorithmic approach* Johns Hopkins series in the mathematical sciences. Baltimore (MD): Johns Hopkins University Press; 1981.
17. Latouche G, Ramaswami V. *Introduction to Matrix analytic methods in stochastic modeling*. Philadelphia (PA): SIAM; 1999.
18. Ramaswami V. A stable recursion for the steady state vector in Markov chains of $M/G/1$ type. *Commun Stat Stoch Models* 1988;4(1):183–188.
19. Schellhaas H. On Ramaswami's algorithm for the computation of the steady state vector in Markov chains of $M/G/1$ -type. *Commun Stat Stoch Models* 1990;6(3):541–550.
20. Meini B. An improved FFT-based version of Ramaswami's formula. *Stoch Models* 1997;13(2):223–238.
21. Falkenberg E. On the asymptotic behaviour of the stationary distribution of Markov chains of $M/G/1$ -type. *Commun Stat Stoch Models* 1994;10(1):75–97.
22. Gail HR, Hantler SL, Taylor BA. Use of characteristics roots for solving infinite state Markov chains. In: Grassmann WK, editor. *Computational probability*. Boston (MA): Kluwer Academic Publishers; 2000. pp. 205–255.
23. Ramaswami V. The generality of Quasi Birth-and-Death processes. In: Alfa AS, Chakravarthy S, editors. *Advances in Matrix analytic methods for stochastic models*. New Jersey: Notable Publications; 1998.
24. Ramaswami V. Nonlinear Matrix equations in applied probability—solution techniques and open problems. *SIAM Rev* 1988;30(2):256–263.
25. Latouche G. *Algorithms for infinite Markov chains with repeating columns*. Volume 48, *Linear algebra, Markov chains, and queueing models* (Minneapolis (MN); 1992), IMA Volumes in Mathematics and its Applications. New York: Springer; 1993.
26. Latouche G. Algorithms for evaluating the Matrix G in Markov chains of $PH/G/1$ type. *Cah Cent'Étud Rech Opér* 1994;36:251–258. Hommage à Simone Huyberegts.
27. Latouche G. Newton's iteration for non-linear equations in Markov chains. *IMA J Numer Anal* 1994;14(4):583–598.
28. Meini B. New convergence results on functional iteration techniques for the numerical solution of $M/G/1$ type Markov chains. *Numer Math* 1997;78(1):39–58.
29. Neuts MF. The Markov renewal branching process. *Proceedings of the Conference on Mathematical Methodology in the Theory of Queues*. New York: Springer Verlag; 1974. pp. 1–21.
30. Hockney RW. A fast direct solution of Poisson's equation using Fourier analysis. *J Assoc Comput Mach* 1965;12:95–113.
31. Buzbee BL, Golub GH, Nielson CW. On direct methods for solving Poisson's equations. *SIAM J Numer Anal* 1970;7:627–656.
32. Bini D, Meini B. On the solution of a non-linear Matrix equation arising in queueing problems. *SIAM J Matrix Anal Appl* 1996;17(4):906–926.

33. Bini DA, Meini B. Improved cyclic reduction for solving queueing problems. *Numer Algorithms* 1997;15(1):57–74.
34. Latouche G, Ramaswami V. A logarithmic reduction algorithm for quasi-birth-death processes. *J Appl Probab* 1993;30(3):650–674.
35. Bini DA, Latouche G, Meini B. Solving matrix polynomial equations arising in queueing problems. *Linear Algebra Appl* 2002;340:225–244.
36. Bini DA, Meini B, Steffé S, *et al.* Structured Markov chains solver: algorithms. *Proceedings of ValueTools*. Pisa: ACM; 2006.
37. Bini DA, Meini B, Steffé S, *et al.* Structured Markov chains solver: software tools. *Proceedings of ValueTools*. Pisa: ACM; 2006.
38. Bini DA, Meini B, Ramaswami V. Analyzing M/G/1 paradigms through QBDs: the role of the block structure in computing the matrix G . In: Latouche G, Taylor P, editors. *Advances in algorithmic methods for stochastic models*, *Proceedings of the 3rd Conference on Matrix Analytic Methods*. New Jersey: Notable Publications Inc.; 2000. pp. 73–86.
39. Latouche G, Stewart G. Numerical methods for M/G/1 type queues. In: Stewart WJ, editor. *Computations with Markov Chains*. Boston (MA): Kluwer Academic Publishers; 1995. pp. 571–581.
40. Stewart GW. On the solution of block Hessenberg systems. *Numer Linear Algebra Appl* 1995;2(3):287–296.
41. Bini DA, Meini B. Inverting block Toeplitz matrices in block Hessenberg form by means of displacement operators: application to queueing problems. *Linear Algebra Appl* 1998;272:1–16.
42. Akar N, Sohraby K. An invariant subspace approach in $M/G/1$ and $G/M/1$ type Markov chains. *Commun Stat Stochastic Models* 1997;13(3):381–416.
43. Meini B. Solving QBD problems: the cyclic reduction algorithm versus the invariant subspace method. *Adv Perform Anal* 1998;1:215–225.
44. He C, Meini B, Rhee NH, *et al.* A quadratically convergent Bernoulli-like algorithm for solving matrix polynomial equations in Markov chains. *Electron Trans Numer Anal (electronic)* 2004;17:151–167.
45. Bini DA, Gemignani L, Meini B. Solving certain matrix equations by means of Toeplitz computations: algorithms and applications. Volume 323, *Fast algorithms for structured matrices: theory and applications* (South Hadley (MA); 2001), *Contemporary Mathematics*. Providence (RI): American Mathematical Society; 2003. pp. 151–167.
46. Van Houdt B. Software SMCSolver. 2006. Available at <http://ftp.win.ua.ac.be/pub/pats/tools>.
47. Bini DA, Meini B, Steffé S. SMCSolver, Fortran 90 version. 2006. Available at <http://bezout.dm.unipi.it/SMCSolver>.
48. Riska A, Zhang Q, Zhang EZ, *et al.* Software MAMSolver. Available at <http://www.cs.wm.edu/MAMSolver/index.html>.
49. Riska A, Smirni E. ETAQA solutions for infinite markov processes with repetitive structure. *INFORMS J Comput* 2007;19(2):215–288.
50. Neuts MF. Matrix-analytic methods in queueing theory. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Radon (FL): CRC Press; 1995. Chapter 10.
51. Grassmann W, Stanford D. Matrix-analytic methods. In: Grassmann WK, editor. *Computational probability*. Boston (MA): Kluwer Academic Publishers; 2000.
52. Ramaswami V. Matrix analytic methods: a tutorial overview with some extensions and new results. Volume 183, *Matrix-analytic methods in stochastic models* (Flint, MI), *Lecture Notes in Pure and Applied Mathematics*. New York: Dekker; 1997. pp. 261–296.
53. Chakravorthy SR, Alfa AS, editors. Volume 183, *Matrix analytic methods in stochastic models*, *Lectures Notes in Pure and Applied Mathematics*. New York: Marcel Dekker; 1996.
54. Chakravorthy SR, Alfa AS, editors. *Proceedings of the 2nd International Workshop on Matrix Analytic Methods*, Winnipeg. Neshanic Station (NJ): Notable Publications; 1998.
55. Latouche G, Taylor P, editors. *Advances in algorithmic methods for stochastic models*, *Proceedings of the 3rd Conference on Matrix Analytic Methods*, Leuven. New Jersey: Notable Publications; 2000.
56. Latouche G, Taylor P, editors. *Proceedings of 4th International Conference on Matrix-analytic Methods in Stochastic Models*. River Edge (NJ): World Scientific Publishing; 2002.
57. Bini DA, Latouche G, Meini B. In: *Proceedings of 5th International Conference on Matrix-analytic Methods in stochastic models*, Volume 21(2–3). Pisa; 2005.

58. He Q-M, Takine T, Taylor P, *et al.*, Proceedings of the 6th International Conference on Matrix-analytic methods in stochastic models. Ann Oper Res; 2008;160(1).
59. Bini DA, Meini B, Ramaswami V, *et al.* 07461 abstracts collection-numerical methods for structured Markov chains. In: Bini DA, Meini B, Ramaswami V, *et al.*, Eds. Numerical methods for structured Markov chains, 07461 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI); 2008.
60. O'Cinneide CA, editor. (Special issue in honor of Marcel F. Neuts). Stat Stoch Models; 1998;14:1–2.
61. Breuer L, Baum D. An introduction to queueing theory and matrix-analytic methods. Dordrecht: Springer; 2005.

MARKOV REGENERATIVE PROCESSES

YASEMIN SERIN
Industrial Engineering Department,
Middle East Technical University,
Ankara, Turkey

INTRODUCTION

Markov regenerative processes (MRGPs) are continuous-time stochastic processes with more general conditions of regeneration than regenerative processes. Some authors refer to these processes as *semi-regenerative processes* [1]. The limiting lengths of some queues with general distributions can be obtained using the results related to MRGPs. This article benefitted from Çinlar [1] and is largely based on Kulkarni [2]. The proofs of the theorems below can be found in Kulkarni [2].

DEFINITIONS AND TRANSIENT BEHAVIOR

A Markov renewal sequence (MRS) is a generalization of a renewal sequence and a MRGP is a generalization of a regenerative process. The first is a discrete-time process and the second is a continuous-time process; in both the future depends on the past only through the state at the last “regeneration.”

Definition. The process $\{(X_n, S_n), n = 0, 1, 2, \dots\}$ is called a *Markov renewal sequence* (MRS) with a discrete state space S if for $S_0 = 0, S_{n+1} > S_n$,

$$\begin{aligned} P(X_{n+1} = j, S_{n+1} - S_n \leq x | X_n \\ = i, S_n, X_{n-1}, S_{n-1}, \dots, X_0, S_0) \\ = P(X_{n+1} = j, S_{n+1} - S_n \leq x | X_n = i) \\ = P(X_1 = j, S_1 \leq x | X_0 = i) \end{aligned}$$

for all $n = 0, 1, \dots$

The matrix G is defined by $G(y) = [G_{ij}(y)]$ where

$$G_{ij}(y) = P(X_1 = j, S_1 \leq y | X_0 = i) \text{ for } i, j \in S$$

is called the *kernel of the MRS*. An MRS is well defined by its initial distribution and its kernel.

Let $\{X_n, n = 0, 1, 2, \dots\}$ be a discrete-time stochastic process with a countable state space S . Let Y_n be a sequence of random variables taking values on the nonnegative real line, $S_n = \sum_{k=1}^n Y_k$ and $N(t) = \max\{n : S_n \leq t\}$.

Definition. The continuous-time process $\{X(t), t \geq 0\}$ defined by $X(t) = X_{N(t)}$ is called a *semi-Markov process* (SMP) if the sequence $\{(X_n, Y_{n+1}), n = 0, 1, \dots\}$ satisfies

$$\begin{aligned} P(X_{n+1} = j, Y_{n+1} < y | X_n \\ = i, Y_n, X_{n-1}, Y_{n-1}, \dots, X_1, Y_1, X_0) \\ = P(X_{n+1} = j, Y_{n+1} < y | X_n = i) \\ = G_{ij}(y) \end{aligned}$$

for all $i, j \in S, n = 0, 1, \dots$ and $y \geq 0$.

Definition. A stochastic process $\{Z(t), t \geq 0\}$ with a countable state space S is called a *Markov regenerative process* (MRGP) if there is an (embedded) MRS $\{(X_n, S_n), n = 0, 1, 2, \dots\}$ such that finite-dimensional distributions of the processes

$$\begin{aligned} \{Z(t + S_n), t \geq 0 | X_n = i, Z(u) \text{ for } 0 \leq u < S_n\} \\ \text{and } \{Z(t), t \geq 0 | X_0 = i\} \end{aligned}$$

are identical.

An alternative definition can be given as follows. A stochastic process with a countable state space S is called a *Markov regenerative process* if for all $n = 0, 1, 2, \dots, 0 \leq t_1 \leq t_2 \leq \dots \leq t_n$ and $A_1 \not\subset A_2 \not\subset \dots \not\subset A_n \subset S$,

there exists an S_1 such that $P(S_1 > 0) > 0$, $P(S_1 < \infty) = 1$ and

$$\begin{aligned} P(Z(S_1 + t_1) \in A_1, Z(S_1 + t_2) \in A_2, \dots, \\ Z(S_1 + t_n) \in A_n | Z(S_1), Z(u) \\ \text{for } 0 \leq u < S_1) \\ = P(Z(t_1) \in A_1, Z(t_2) \in A_2, \dots, \\ Z(t_n) \in A_n | Z(0)). \end{aligned}$$

The future probabilities of an MRGP after a transition point S_n depend on the past only through the state at S_n . Continuous-time Markov chains, SMPs, and regenerative processes are special cases of MRGPs.

Example [M/G/1 Queue]. The arrivals to a queue are according to a Poisson process with rate λ and the iid service times have a general distribution with cdf $B(\cdot)$ with mean r . Let $Z(t)$ be the number of customers in the system at time t and let S_n be the n th departure time, $X_n = Z(S_n +)$ be the number of customers left in the system just after the n th departure, Y_n be the n th service time, and $N(t)$ be the number of departures from the queue in t time units. The process defined by $X(t) = X_{N(t)}$ is an SMP with

$$G_{ij}(y) = \begin{cases} \int_0^y \frac{e^{-\lambda x} (\lambda x)^{j-i-1}}{(j-i-1)!} dB(x), & \text{if } i > 0, j \geq i-1 \\ \int_0^y \frac{e^{-\lambda x} (\lambda x)^j}{j!} dB(x), & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $Z(t)$ satisfies the definition of MRGP.

Figure 1 demonstrates a sample path of an M/G/1 queue as an MRGP $Z(t)$, the corresponding sample paths of MRS (X_n, S_n) and the SMP $X(t)$. While the SMP $X(t)$ is in state k , the MRGP $Z(t)$ can visit other states such as $j \neq k$.

The next theorem gives a Markov renewal type equation for

$$P_{ij}(t) = P(Z(t) = j | X_0 = i) \text{ for } i, j \in S.$$

Theorem 1. Let $\{Z(t), t \geq 0\}$ be an MRGP with state space S and let $\{(X_n, S_n), n =$

$0, 1, \dots\}$ be the embedded MRS with state space S and kernel G . Then $P_{ij}(t)$ satisfies the Markov renewal type equation

$$\begin{aligned} P_{ij}(t) &= P(Z(t) = j, S_1 > t | X_0 = i) \\ &+ \sum_{k \in S} \int_0^t P_{kj}(t-u) dG_{ik}(u). \end{aligned}$$

LIMITING PROBABILITIES

The transient probabilities are difficult to obtain in most cases. The limiting behavior of an MRGP is given in the following theorem.

Theorem 2. Let $\{Z(t), t \geq 0\}$ be an MRGP with state space $S = \{0, 1, 2, \dots\}$ and let $\{(X_n, S_n), n = 0, 1, \dots\}$ be the embedded MRS with kernel G . Let

$$\begin{aligned} \mu_i &= E(S_1 | X_0 = i), \\ \alpha_{kj} &= E(\text{time spent by } Z(t) \text{ in state } j \text{ for } 0 \leq t < S_1 | X_0 = k). \end{aligned}$$

Suppose that

- (i) the sample paths of $\{Z(t), t \geq 0\}$ are right continuous with left limits,
 - (ii) the SMP $X(t) = X_{N(t)}$ is irreducible, aperiodic and positive recurrent, and
 - (iii) π is a positive solution to $\pi = \pi G(\infty)$.
- Then

$$\lim_{t \rightarrow \infty} P_{ij}(t) = \frac{\sum_{k=0}^{\infty} \pi_k \alpha_{kj}}{\sum_{k=0}^{\infty} \pi_k \mu_k}.$$

The limit result of Theorem 2 can be demonstrated on Fig. 1 as follows:

$$P(Z(t) = j) = \sum_{k=0}^{\infty} P(Z(t) = j, X(S_{N(t)}) = k)$$

$$= \sum_{k=0}^{\infty} P(Z(t) = j | X(S_{N(t)}) = k)$$

$$\times P(X(S_{N(t)}) = k) \text{ and}$$

$$\begin{aligned} \lim_{t \rightarrow \infty} P(Z(t) = j) &= \sum_{k=0}^{\infty} \lim_{t \rightarrow \infty} P(Z(t) \\ &= j | X(S_{N(t)}) = k) \lim_{t \rightarrow \infty} P(X(S_{N(t)}) = k). \end{aligned}$$

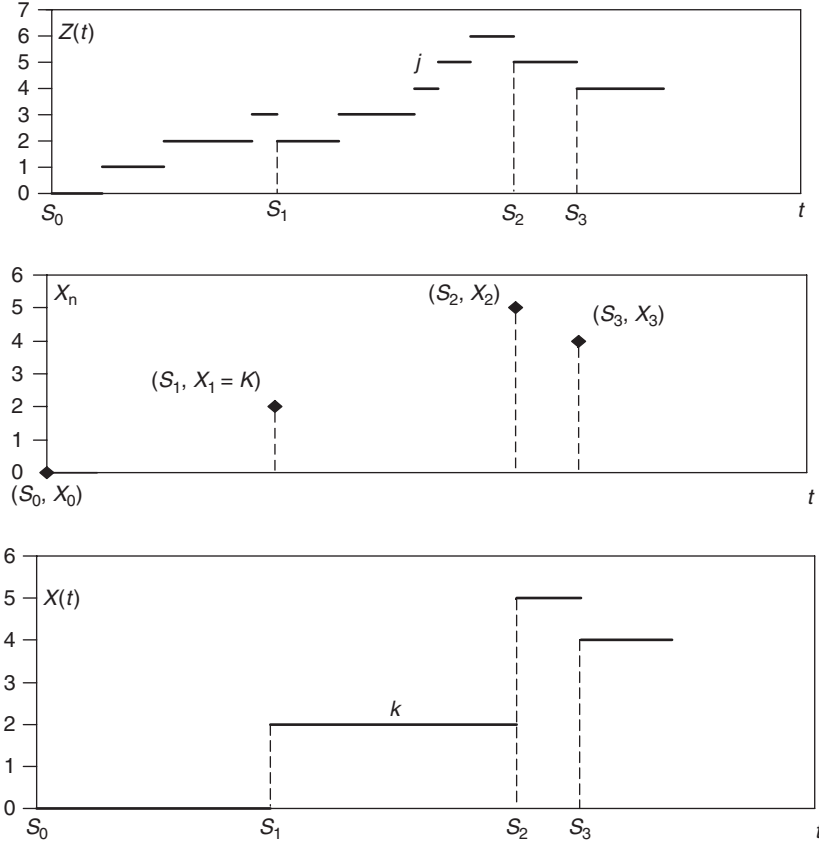


Figure 1. Markov regenerative process $Z(t)$, Markov renewal sequence (X_n, S_n) , and semi-Markov process $X(t)$.

The summand of the right side of the equation intuitively is, in the long run, the fraction of the time MRGP is in state j when the embedded SMP is in state k times the fraction of the time the SMP is in state k . From the limiting probabilities of an SMP from Kulkarni [2]

$$\lim_{t \rightarrow \infty} P(X(t) = k) = \frac{\pi_k \mu_k}{\sum_{i=0}^{\infty} \pi_i \mu_i}$$

and

$$\lim_{t \rightarrow \infty} P(Z(t) = j | X(S_{N(t)}) = k) = \frac{\alpha_{kj}}{\mu_k},$$

and the result in Theorem 2 follows.

Example [A Series System]. Consider an equipment with N components. The equipment

works as long as all the components work. Component i works for an *exponential*(λ_i) units of time and its repair takes R_i units of time with cdf $F_i(r)$ and mean r_i . There is a single repairman and components do not fail when one is being repaired. Let $S_0 = 0, S_1, S_2, \dots$ be the successive failure times and X_n be the component causing n th failure. $S_{n+1} - S_n$ consists of a repair and a failure-free time. Then the process $\{(X_n, S_n), n = 0, 1, 2, \dots\}$ is a Markov renewal process with kernel

$$G_{ij}(y) = \int_0^y \lambda_j e^{-u \sum_{i=1}^N \lambda_i} F_i(y - u) du$$

for $i, j = 1, 2, \dots, N$. Then the process $\{Z(t), t \geq 0\}$ defined by

$$Z(t) = \begin{cases} 0, & \text{if the equipment is} \\ & \text{working at time } t \\ i, & \text{if the component } i \text{ is under} \\ & \text{repair at time } t \end{cases}$$

is an MRGP. Note that the MRS or the SMP have the state space $\{1, 2, \dots, N\}$, while the MRGP has the state space $\{0, 1, 2, \dots, N\}$. Then for $i, j = 1, 2, \dots, N$,

$$G_{ij}(\infty) = \frac{\lambda_j}{\sum_{k=1}^N \lambda_k}.$$

For $j = 1, 2, \dots, N$, the limiting probabilities of the embedded MRS are

$$\pi_j = \frac{\lambda_j}{\sum_{k=1}^N \lambda_k},$$

and for $k = 1, 2, \dots, N$, the expected durations are

$$\alpha_{k0} = \frac{1}{\sum_{k=1}^N \lambda_k}$$

$$\alpha_{kk} = r_k$$

$$\alpha_{kj} = 0 \text{ otherwise}$$

$$\mu_k = r_k + \frac{1}{\sum_{j=1}^N \lambda_j}.$$

Then from Theorem 2, the probability that the system is working in the long run is

$$\lim_{t \rightarrow \infty} P_{i0}(t) = \frac{1}{1 + \sum_{k=1}^N r_k \lambda_k}$$

and the probability that the component j is under repair in the long run is

$$\lim_{t \rightarrow \infty} P_{ij}(t) = \frac{r_j \lambda_j}{1 + \sum_{i=1}^N r_k \lambda_k}.$$

REFERENCES

1. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
2. Kulkarni VG. Modelling and analysis of stochastic systems. New York: Chapman Hall; 1995.

MARKOV RENEWAL FUNCTION AND MARKOV RENEWAL-TYPE EQUATIONS

SERHAN ZIYA

Department of Statistics and Operations
Research, University of North Carolina,
Chapel Hill, North Carolina

MARKOV RENEWAL FUNCTION

Suppose that $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence with a corresponding kernel $G(t)$. For $n \geq 0$, X_n takes values from the discrete set E , $\{X(t), t \geq 0\}$ is the semi-Markov process and $\mathbf{N}(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ is the Markov renewal process generated by this Markov renewal sequence. For $t \geq 0$, let

$$M_{jk}(t) = \mathbf{E}[N_k(t) | X_0 = j], j, k \in E.$$

Thus, $M_{jk}(t)$ is the expected number of transitions of $\{X(t), t \geq 0\}$ into state k over $(0, t]$ given that the initial state is j . Let $M(t) = [M_{jk}(t)]_{j,k \in E}$ denote the matrix of these expectations. Then, the matrix $M(t)$ is called the *Markov renewal function* generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$.

Before we move further, we first introduce the matrix convolution notation, which will make the presentation significantly simpler. Suppose that $A(t) = [A_{jk}(t)]$ and $B(t) = [B_{jk}(t)]$ are two matrices of functions so that the product $A(t)B(t)$ is defined. Now, let

$$\begin{aligned} C_{jk}(t) &= \sum_i \int_0^t A_{ji}(t-u) dB_{ik}(u) \\ &= \sum_i \int_0^t dA_{ji}(u) B_{ik}(t-u). \end{aligned}$$

Then, the matrix $C(t) = [C_{ij}(t)]$ is called the *convolution* of A and B (and written as $C(t) = A * B(t)$).

In the analysis of Markov renewal processes, one frequently uses “Markov renewal arguments,” which is the idea of first obtaining an expression for whatever is of interest

by conditioning on the first sojourn time T_1 and the first state the Markov renewal sequence visits, X_1 . We next use the Markov renewal argument to determine an expression for the Markov renewal equation.

By conditioning on X_1 and T_1 , we obtain

$$\begin{aligned} \mathbf{E}[N_k(t) | X_0 = j, X_1 = i, T_1 = s] \\ = \begin{cases} \delta_{ik} + M_{ik}(t-s), & s \leq t, \\ 0, & s > t, \end{cases} \end{aligned}$$

where $\delta_{ik} = 1$ if $i = k$ and zero otherwise.

Using this equation, one can show that

$$M_{jk}(t) = \int_0^t dG_{jk}(s) + \sum_{i \in E} \int_0^t M_{ik}(t-s) dG_{ji}(s),$$

which, in matrix form, can also be written as

$$M(t) = G(t) + G * M(t). \quad (1)$$

Equation (1) is called the *Markov renewal equation*.

Now, suppose that $H_j(\cdot)$ is the sojourn time distribution for state j for the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$ and suppose that there exists positive ε and α such that $H_i(\alpha) < 1 - \varepsilon$ for $i \in E$. This condition essentially ensures the regularity of the process by preventing sojourn times to be consistently too small so that up to any point in time t , the total number of renewals is always finite. To be more precise, as Kulkarni [1] shows, under this condition, the process $\{N(t), t \geq 0\}$ where $N(t) = \max\{n \geq 0 : T_n \leq t\}$ is regular, that is, $N(t) < \infty$ for all $0 \leq t < \infty$ with probability 1.

Under the same regularity condition, by repeatedly using the Markov renewal equation (1), one can show that

$$M(t) = \sum_{i=1}^{\infty} G^{*i}(t) \text{ for } t \geq 0,$$

where $G^{*i}(\cdot)$ is the i -fold convolution of the matrix $G(\cdot)$ with itself.

MARKOV RENEWAL-TYPE EQUATIONS AND THEIR SOLUTION

In the previous section, we used a Markov renewal argument to derive an expression for the Markov renewal function and obtained the Markov renewal equation (1). In general, when we use the Markov renewal argument, we usually end up with the following generalized version of Equation (1), which we call Markov renewal-type equation

$$R(t) = D(t) + G * R(t).$$

In this equation, typically, $D(t)$ and $G(t)$ are known and the objective is to determine $R(t)$. Under certain conditions on $D(t)$ and the regularity condition on $G(t)$, which we described in the previous section, it is known that there exists a unique solution given by

$$R(t) = D(t) + M * D(t), \quad (2)$$

where $M(t)$ is the Markov renewal function associated with kernel $G(\cdot)$. (For a formal statement of this result, see Kulkarni [1]).

Even though Equation (2) gives the solution for $R(\cdot)$, it is usually not used for computational purposes. It is, however, crucial in establishing the key Markov-renewal theorem, which gives the expression for the limiting behavior of $R(t)$, and is one of the most fundamental results in Markov renewal theory.

OTHER SOURCES

For more on Markov renewal functions and Markov renewal-type equations, we refer the reader to Cinlar [2] and Kulkarni [1], from which the author of this article has also benefited.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. Boca Raton (FL): Chapman & Hall/CRC; 1995.
2. Cinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1975.

MARKOV RENEWAL PROCESSES

SERHAN ZIYA

Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

Markov renewal processes are a generalization of renewal processes. While a renewal process is a counting process that keeps track of the number of occurrences of a certain type of event associated with a stochastic process, a Markov renewal process is a vector-valued process that keeps a separate count of these events depending on the state of the stochastic process at the times these events occur. For example, while a renewal process would simply keep track of the number of arrivals at a queueing system, a Markov renewal process could help keep a separate count of the arrivals depending on the queue lengths they observe.

MARKOV RENEWAL SEQUENCES

In order to clearly define Markov renewal processes, it helps to introduce Markov renewal sequences first. Consider a sequence of bivariate random variables $\{(X_n, T_n), n \geq 0\}$ where $T_0 = 0$, $T_n \leq T_{n+1}$, and X_n takes values in a discrete set E for all $n \geq 0$. Suppose that the following condition holds for all $n \geq 0$:

$$\begin{aligned} P\{X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i, T_n, X_{n-1}, \\ T_{n-1}, \dots, X_0, T_0\} \\ = P\{X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i\} \\ = P\{X_1 = j, T_1 \leq t | X_0 = i\}. \end{aligned} \quad (1)$$

Then, the stochastic process $\{(X_n, T_n), n \geq 0\}$ is said to be a Markov renewal sequence. Thus, a Markov renewal sequence satisfies the Markov property at each $n \geq 0$ since by Condition (1), the future stochastic evolution of the process depends on the past only through the system state X_n at n .

Markov renewal sequences form the basic structure of some of the important classes of stochastic processes such as semi-Markov processes and Markov regenerative processes. A continuous-time stochastic process with a countable state space and piecewise constant, right-continuous sample paths is called a *semi-Markov process* if the process $\{(X_n, T_n), n \geq 0\}$ where T_n is the time of the n th jump epoch for the process and X_n is the system state right after T_n is a Markov renewal sequence. In other words, for a continuous-time stochastic process to be a semi-Markov process, a Markov renewal sequence must arise at times where the process makes jumps. If there exists a Markov renewal sequence embedded in a continuous-time stochastic process (not necessarily at jump points but perhaps at some other carefully chosen instances), the stochastic process is said to be a Markov regenerative process.

We next discuss how to describe a Markov renewal sequence. Suppose that $\alpha_i = P\{X_0 = i\}$ is the probability that the initial state is i and $G_{ij}(t) = P\{X_1 = j, T_1 \leq t | X_0 = i\}$. Then, it can be easily shown that the Markov renewal sequence is completely characterized by the initial distribution $\alpha = [\alpha_i]_{i \in E}$ and the matrix $G(y) = [G_{ij}(y)]_{i,j \in E}$, which is called the *kernel of the Markov renewal sequence*.

Since the sequence $\{(X_n, T_n), n \geq 0\}$ satisfies the Markov property at all $n \geq 0$, it is also easy to see that the process $\{X_n, n \geq 0\}$ is a discrete-time Markov chain with the same initial distribution as the Markov renewal sequence and transition probability matrix P defined by $P = G(\infty)$ with the exception that $P_{ii} = 1$ whenever $\sum_{j \in E} G_{ij}(\infty) = 0$ for some $i \in E$. On the other hand, any discrete-time Markov chain can be seen as a Markov renewal sequence where $T_n = n$ for all $n \geq 0$.

MARKOV RENEWAL PROCESSES

Suppose that $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence as defined in the section titled "Markov Renewal Sequences" and let $\{N(t), t \geq 0\}$ be a counting process defined as

$$N(t) = \sup \{n \geq 0 : T_n \leq t\}.$$

We say that the process $\{N(t), t \geq 0\}$ is regular if $N(t) < \infty$ with probability 1 for all $0 \leq t < \infty$. Now, for $i \in E$, define

$$F_i(t) = P\{T_1 \leq t \mid X_0 = i\} = \sum_{j \in E} G_{ij}(t).$$

Then, Kulkarni [1] shows that one sufficient condition for $\{N(t), t \geq 0\}$ to be regular is that there exists positive ε and δ such that $F_i(\delta) < 1 - \varepsilon$ for all $i \in E$. Note that this condition essentially prevents T_n 's take very small values with probability 1.

Assuming that $\{N(t), t \geq 0\}$ is regular, we define a new process

$$X(t) = X_{N(t)} \text{ for } t \geq 0.$$

Then, $\{X(t), t \geq 0\}$ is a semi-Markov process and it is said to be the semi-Markov process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$.

Now, define $N_k(t)$ to be the number of times the semi-Markov process visits state k over the interval $(0, t]$. Then, the stochastic vector process $\mathbf{N}(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ is said to be the Markov renewal process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$.

Remark. The reader should note that the use of the terms Markov renewal sequence and Markov renewal process is not consistent in the literature. The definitions of these terms given in this article follow those in Kulkarni [1]. The definition of Markov renewal process given in Heyman and Sobel [2] is the same as that in Kulkarni [1], however, Cinlar [3] and Kihlbas [4] use the term Markov renewal process to refer to the process defined as Markov renewal sequence in this article and in Kulkarni [1].

EXAMPLES

In this section, we provide a number of examples where Markov renewal sequences and Markov renewal processes appear.

Example 1. Suppose that $\{X(t), t \geq 0\}$ is a continuous-time Markov chain. Let T_n denote the n th time the Markov chain makes a jump and define X_n to be the system state right after the n th jump, that is, $X_n = X(T_n+)$. Then, since in a continuous-time Markov chain, Markov property holds at all times, it also holds at jump epochs, and thus we conclude that $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence. Clearly, the continuous-time Markov chain $\{X(t), t \geq 0\}$ is also the semi-Markov process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$. Therefore, if we define $N_k(t)$ to be the number of times the Markov chain visits state k over $(0, t]$, we can see that the vector stochastic process $\mathbf{N}(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ is a Markov renewal process.

Example 2. Suppose that $\{Q(t), t \geq 0\}$ is the queue length process for the G/M/1 queue. Let $H(\cdot)$ denote the cumulative distribution function for the interarrival times and μ denote the exponential service rate.

Define T_n to be the arrival time of the n th customer and $X_n = Q(T_n-)$. Thus, X_n is the queue length the n th arriving customer sees. Now, since service times are exponentially distributed, at arrival points the queue length process has the Markov property, that is, its future evolution depends on the past only through the queue lengths at those times. Thus, $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence and its kernel can be expressed as

$$G_{ij}(t) = \begin{cases} \int_0^t \frac{e^{-\mu s} (\mu s)^{i+1-j}}{(i+1-j)!} dH(s), & 1 \leq j \leq i+1, \\ H(t) - \sum_{k=1}^{i+1} G_{ik}(t), & j = 0. \end{cases}$$

Let $N_k(t)$ denote the number of arrivals over $(0, t]$ who find k customers at the time of their arrival. One can see that $N_k(t)$ is also the number of times the semi-Markov process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$ visits state k over $(0, t]$. Thus, $\mathbf{N}(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ is a Markov renewal process.

Example 3. Suppose that a new machine has a random lifetime with a cumulative distribution function $A(\cdot)$. Once the machine fails, it is repaired, and the repaired machine becomes as good as new. The repair time depends on the amount of time the machine has been operational since the last repair. If the machine remains operational less than T units of time before it fails, then the repair time distribution is given by $B_1(\cdot)$. Otherwise, the distribution is given by $B_2(\cdot)$. Let $X(t)$ denote the state of the machine at time t . If $X(t) = 0$, that means the machine is operational at time t . If $X(t) = i$ where $i \in \{1, 2\}$, that means the machine is going through a repair of type i at time t . Let X_n denote the state of the machine after the n th jump of the process $\{X(t), t \geq 0\}$ and T_n denote the time of the n th jump. Then $\{(X_n, T_n), n \geq 0\}$ is a Markov renewal sequence with the following kernel.

$$G(t) = \begin{bmatrix} 0 & A(t) & 0 \\ B_1(t) & 0 & 0 \\ B_2(t) & 0 & 0 \end{bmatrix} \text{ if } t \leq T$$

and

$$G(t) = \begin{bmatrix} 0 & A(T) & A(t) - A(T) \\ B_1(t) & 0 & 0 \\ B_2(t) & 0 & 0 \end{bmatrix} \text{ if } t > T.$$

Note also that $\{X(t), t \geq 0\}$ is a semi-Markov process and, in fact, the semi-Markov process generated by the Markov renewal sequence $\{(X_n, T_n), n \geq 0\}$.

Now, let $N_0(t)$ denote the number of times the machine is repaired and $N_i(t)$ denote the number of times the machine has started a type i repair over the time interval $(0, t]$. Then, $N_k(t)$ is the number of times the semi-Markov process $\{X(t), t \geq 0\}$ visited state k over $(0, t]$ and thus $\mathbf{N}(t) = \{[N_k(t)]_{k \in E}, t \geq 0\}$ is a Markov renewal process.

OTHER SOURCES

For more on Markov renewal processes, we refer the reader to Cinlar [3], Heyman and Sobel [2], Kohlas [4], and Kulkarni [1], from which the author of this article has also benefited. Cinlar [3] and Kulkarni [1], in particular, provide a comprehensive coverage of Markov renewal process including key results that are typically used in the analysis of these processes.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. Boca Raton (FL): Chapman & Hall/CRC; 1995.
2. Heyman DP, Sobel MJ. Volume I, Stochastic models in operations research. New York: McGraw-Hill, Inc.; 1982.
3. Cinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1975.
4. Kohlas J. Stochastic methods of operations research. Cambridge, Great Britain: Cambridge University Press; 1982.

MARKOVIAN ARRIVAL PROCESSES

SRINIVAS R. CHAKRAVARTHY
Department of Industrial and
Manufacturing Engineering,
Kettering University, Flint,
Michigan

INTRODUCTION

Every one has seen a point process in action in one form or the other in day-to-day activities. For example, on a regular basis we all visit a grocery store or a post office or a bank or a restaurant or a drive-through place such as a fast food restaurant or a coffee place. Each customer can be thought of as a “point” arriving to the “place of interest” with some “work” to be done. These points arrive at random times (in some cases in batches) and the associated amounts of work can be characterized by another set of random variables. Mathematically, a point process $\{X_n\}$ is a sequence of random variables that may represent epochs at which certain events occur. Associated with these epochs are other random variables, $\{Y_n\}$ that contain further information about the events. Each pair (X_n, Y_n) is said to be a *marked point*, and the sequence $\{(X_n, Y_n)\}$ is called a *marked point process* (MPP). For example, in a drive-through coffee house, the quantity X_n will represent the time of the n th car (since the last car) to the place where the order is given, and Y_n may stand for the amount (in dollars) of the n th order or the time the n th order is picked up (at the next window) from the time it was placed. It is possible that $\{Y_n\}$ could be a vector process. So, going back to the earlier example, the first component of Y_n can be taken as the amount of the order and the second component is the time to pick up the order.

Why do we need to study a point process? Modeling plays an important role in practice. Models are representations of objects, processes, or anything that we wish

to describe or whose patterns of behavior we wish to analyze. They enable us to understand the behavior of complex systems. Modeling is an iterative process involving several stages. They can be summarized as follows:

- *Formulation Stage.* Study the problem; propose a model.
- *Solution Stage.* Choose appropriate tools to bring out essential characteristics of the model.
- *Validation Stage.* Verify the model and propose modification, if any.

There are basically two types of modeling: deterministic modeling and stochastic modeling. The *deterministic models* include topics such as mathematical programming (linear and nonlinear programming), scheduling, and sequencing. The *stochastic modeling* (also referred to as *probabilistic modeling*) covers areas such as Markov processes, queueing theory, reliability theory, warranty analysis, inventory, and probabilistic scheduling. Note that *statistical modeling* that deals with the study of uncertainty and variability is subsumed in stochastic modeling.

Going back to our drive-through coffee house example, a prudent management would be interested to see how the arrivals occur (including the pattern of arrivals) so as to schedule its workforce as well as to upgrade its service facility for continuous improvement. The formulation stage of the modeling involves a good understanding of the point process describing the underlying arrivals.

MAP IN CONTINUOUS TIME

Point processes can be studied in discrete time or in continuous time, depending on the nature of applications in which these processes are used. First, we will discuss Markovian arrival process (MAP) in continuous time by highlighting the building blocks for this process.

Poisson Process

One of the most celebrated point processes in continuous time is the classical Poisson process (see **Poisson Process and its Generalizations**). Since this process is a building block for the MAP, we will briefly revisit the Poisson process. Suppose that the interarrival times of the cars to the drive-through coffee house are independent and identically distributed with the common probability density function, $f(t)$, and the cumulative distribution function, $F(t)$, given by

$$f(t) = \lambda e^{-\lambda t}, t \geq 0, F(t) = 1 - e^{-\lambda t}, t \geq 0, \quad (1)$$

where $\lambda > 0$ and gives the average rate of cars arriving per unit of time.

The counting process $\{N(t), t \geq 0\}$ giving the number of cars arriving at the place of order in the drive-through coffee house during $(0, t]$ plays an important role, notably, in stochastic models. Denoting by $p_n(t) = P(N(t) = n)$ it can easily be verified that the following differential-difference equations are satisfied (see **Poisson Process and its Generalizations**):

$$\begin{aligned} p'_0(t) &= -\lambda p_0(t), \\ p'_n(t) &= -\lambda p_n(t) + \lambda p_{n-1}(t), \quad n = 1, 2, 3, \dots \end{aligned} \quad (2)$$

subject to the initial conditions

$$p_0(0) = 1 \text{ and } p_n(0) = 0, n = 1, 2, 3, \dots \quad (3)$$

Equations (2) and (3) can be solved to give an explicit solution for $p_n(t)$ as

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n = 0, 1, 2, \dots \quad (4)$$

The renewal function, $H(t) = E[N(t)]$, and its density are given by

$$H(t) = \lambda t \text{ and } h(t) = H'(t) = \lambda. \quad (5)$$

Why Phase-Type Distributions?

In practice, the qualitative aspects of stochastic models are studied mostly through

well-planned computation. Computing is both an art and science. It has become an integral and important part of the analysis of any stochastic model. Those days are gone when one had the need to have *explicit* analytical solutions. Algorithmic or computational probability plays a major role in the analysis and implementation of stochastic models in practice. Computational aspects are not a luxury any more! Modern day development of algorithmic probability is mainly due to Neuts [1].

Poisson processes and exponential distributions have very nice mathematical properties that make queueing models with these as arrival processes or service-time distributions very attractive and tractable. However, these assumptions are highly restrictive in application. Thus, seeing this limitation, Neuts [1] first developed the theory of phase-type (PH) distributions (see **Phase-Type (PH) Distributions**) and related point processes. In stochastic modeling, PH distributions lend themselves naturally to algorithmic implementation. They have nice closure properties and a related matrix formalism that make them attractive for use in practice (see **Phase-Type (PH) Distributions**).

PH distributions are natural generalizations of exponential distributions and have found tremendous applications among others in queueing, reliability, inventory, and warranty models. Furthermore, these distributions require simple matrix formalism not only in the analysis but also in algorithmic implementation. PH distributions can easily be incorporated into undergraduate curriculum as students are exposed to matrix theory and basic probability in almost all engineering curricula.

Like Poisson process, PH-renewal process forms the (second) building block in the development of MAP. For the sake of continuity, we will only highlight key aspects of PH distributions and refer the reader to article titled **Phase-Type (PH) Distributions** of this handbook for complete details.

Continuous Phase-Type Distributions. Suppose that $\{Y_t\}$ is continuous-time Markov chain (CTMC) on $\{1, 2, \dots, m, m+1\}$ with m transient states $1, 2, \dots, m$ and an absorbing state $m+1$. For details on CTMC, we refer

the reader to the section titled “Continuous-Time Markov Chains (CTMCs)” of this handbook. The generator of the Markov chain (MC) is of the form

$$\tilde{Q} = \begin{bmatrix} S & \mathbf{S}^0 \\ \mathbf{0} & 0 \end{bmatrix}, \quad (6)$$

where S is an $m \times m$ matrix, $\mathbf{S}^0$ is a column vector of order m such that $\mathbf{S}\mathbf{e} + \mathbf{S}^0 = \mathbf{0}$. Assume that the matrix $\mathbf{S}\mathbf{e} + \mathbf{S}^0$ is irreducible. Let $(\beta_1, \dots, \beta_m, \beta_{m+1}) = (\beta, \beta_{m+1})$ be a probability vector giving the initial probabilities. That is, $P(Y_0 = i) = \beta_i, 1 \leq i \leq m+1$. Suppose we start the CTMC in one of the m transient states. We are interested in finding the time until absorption into the absorbing state.

Let X denote the time *until* absorption into state $m+1$. Then X is a continuous random variable on $[0, \infty)$ and its probability density and cumulative probability distribution functions are given by (we assume that $\beta_{m+1} = 0$ which is the case in most applications)

$$\begin{aligned} f(t) &= \beta e^{\mathbf{S}t} \mathbf{S}^0, \quad t \geq 0, \\ F(t) &= P(X \leq t) = 1 - \beta e^{\mathbf{S}t} \mathbf{e}, \quad t \geq 0. \end{aligned} \quad (7)$$

In this case, we say that X follows a PH distribution with representation (β, S) of order m and denote this by $X \sim \text{PH}(\beta, S)$ of order m . The mean and variance of X are given by

$$\mu_X = \beta(-S)^{-1} \mathbf{e} \text{ and } \sigma_X^2 = 2\beta S^{-2} \mathbf{e} - \mu_X^2. \quad (8)$$

In Equation (7) and also in the sequel, we will be using the exponential matrix frequently. Exponential matrix is defined as $e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} = I + A + \frac{A^2}{2!} + \dots$. Some well-known distributions such as exponential, (generalized) Erlang, and hyperexponential are special cases of PH distributions (see **Phase-Type (PH) Distributions**). There are some interesting closure properties associated with the PH distributions and can be seen in Ref. 1 (also see **Phase-Type (PH) Distributions**).

Going back to our drive-through coffee house example, suppose that the interarrival times of successive cars to the drive-through coffee house are independent and identically

distributed with the common probability density function $f(t)$ as given in Equation (7). That is, we are assuming the interarrival times to be of the phase type with representation (β, S) of order m . The underlying CTMC is governed by the generator given in Equation (6). Let $\{N(t), t \geq 0\}$ denote the associated counting process, giving the number of cars arriving at the place of order in the drive-through coffee house during $(0, t]$.

Denoting by $P(n, t) = \{P_{ij}(n, t)\}$, for $n \geq 0, t \geq 0$, the (i, j) th element of $m \times m$ matrices is defined as

$$P_{ij}(n, t) = P(N(t) = n, Y_t = j | N(0) = 0, Y_0 = i), \quad (9)$$

for $n \geq 0, t \geq 0, 1 \leq i, j \leq m$. These matrices satisfy the following differential equations (Ref. 1 and see **Phase-Type (PH) Distributions**) where we assume that $\beta_{m+1} = 0$ which is the case in most applications. That is, $F(0) = 0$.

$$\begin{aligned} P'(0, t) &= SP(0, t) = P(0, t)S, \\ P'(n, t) &= SP(n, t) + \mathbf{S}^0 \beta P(n-1, t) \\ &= P(n, t)S + P(n-1, t) \mathbf{S}^0 \beta, \end{aligned} \quad (10)$$

with the initial conditions $P(n, 0) = 0$ for $n \geq 1$, and $P(0, 0) = I$, an identity matrix of order m .

The computation of the matrices $\{P(n, t): \text{for } n \geq 0, t \geq 0\}$ can be carried out efficiently and will be shown in a more general context when dealing with MAP.

Note that the PH-renewal process is related to the irreducible MC obtained by instantaneously restarting the absorbing MC (whose generator is as given in Eq. 6) with the same initial probability vector β every time an absorption occurs. Thus, the PH-renewal process is related to the irreducible MC with generator

$$Q^* = S + \mathbf{S}^0 \beta. \quad (11)$$

The stationary probability vector, π^* of Q^* satisfying $\pi^* Q^* = \mathbf{0}, \pi^* \mathbf{e} = 1$, is given by

$$\pi^* = \frac{1}{\mu_X} \beta(-S)^{-1}, \quad (12)$$

where μ_X is the mean of X and is given in Equation (8).

The renewal function, $H(t) = E[N(t)]$, and its density for the PH-renewal process are given by the following equation (Ref. 1 and also see **Phase-Type (PH) Distributions**):

$$H(t) = \frac{1}{\mu_X} \left\{ t + \frac{\sigma_X^2 + \mu_X^2}{2\mu_X} + \beta [e\pi^* - e^{Q^*t}] S^{-1} \mathbf{e} \right\}$$

and

$$h(t) = H'(t) = \beta e^{Q^*t} \mathbf{S}^0. \quad (13)$$

Why Markovian Arrival Process?

In practice, we come across many situations where the arrival processes are not necessarily renewal processes. For example, consider a queueing network consisting of, say, 2 nodes. The output of node 1 will form the input to node 2. If we consider non-Poisson arrivals, say, to the first node, the output process will not necessarily be a renewal process. Also, in production line problems where arrivals from different sources form input to an assembly system, the arrival process may not necessarily be a renewal process. So, how do we model these point processes? The answer is to use MAP.

History of MAP. Earlier, we talked about how important it is to have algorithmic aspects built into the analysis and implementation of any stochastic model in practice. During the 1970s, with a genuine concern for the usefulness of stochastic models in practice, Neuts developed algorithmic probability and matrix-analytic methods (see **Matrix Analytic Method: Overview and History**): In the first stage of that development, he introduced PH distributions. In the second stage, Neuts developed matrix geometric methods in queueing and popularized the algorithmic methods as a way to study stochastic models. Then he introduced MAP, a versatile point process that not only generalized the PH-renewal process but also provided a way to model correlated arrivals that had largely been ignored in the literature.

What is a Markovian Arrival Process?

Consider an irreducible CTMC with m transient states. At the end of a sojourn in state i , that is exponentially distributed with parameter λ_i , there are two possibilities. The first possibility corresponds to an “event” (or an arrival) and the CTMC can visit any of the m transient states including the state from which this event occurred with probability p_{ij} . The second possibility corresponds to no arrival and the CTMC can visit any of the $m - 1$ transient states (all m except i) with probability q_{ij} . Thus, the CTMC can go from state i to state i only through an arrival. Define matrices $D_0 = (d_{ij}^{(0)})$ and $D_1 = (d_{ij}^{(1)})$ such that $d_{ii}^{(0)} = -\lambda_i$, $1 \leq i \leq m$; $d_{ij}^{(0)} = \lambda_i q_{ij}$, $j \neq i$, $1 \leq i, j \leq m$; $d_{ij}^{(1)} = \lambda_i p_{ij}$, $1 \leq i, j \leq m$, with $\sum_j p_{ij} + \sum_{j \neq i} q_{ij} = 1$, for $1 \leq i \leq m$. By assuming D_0 to be a nonsingular matrix, the successive times between “events” (or arrivals) will be finite with probability one and the process will not terminate. A pictorial description of this process is given in Fig. 1.

Thus, a MAP is described by the two parameter matrices (D_0, D_1) of order m such that the transitions corresponding to no arrivals are governed by D_0 and the transitions corresponding to arrivals are governed by D_1 . The underlying CTMC has the generator given by $Q = D_0 + D_1$. This representation of MAP is a special case of batch Markovian arrival process (BMAP) (see **Batch Markovian Arrival Processes (BMAP)**) with batch size equal to 1. This

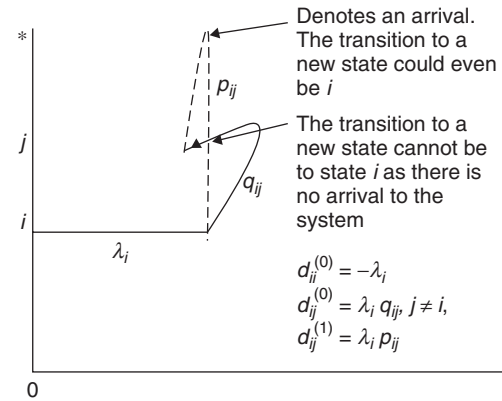


Figure 1. Description of MAP.

BMAP was originally introduced by Neuts [2] as a *versatile Markovian point process* (VMPP) in 1979. The original description of the VMPP was quite involved with heavy notation due to several different types of arrivals that are to be accounted explicitly. Lucantoni, Meier-Hellstern, and Neuts [3] introduced the terms MAP and BMAP to describe the VMPP with much simpler notation. It has now become a standard notation in describing MAP and BMAP. The reason for using the term BMAP rather than the originally coined term VMPP is that, at the time, it was thought that the VMPP is a special class of BMAP. Later, it was realized that the VMPP and BMAP are equivalent processes [4].

The point process described by the MAP is a special class of semi-Markov processes with transition probability matrix given by

$$\int_0^t e^{D_0 x} D_1 dx = [I - e^{D_0 t}](-D_0)^{-1} D_1. \quad (14)$$

Let π be the steady state probability vector of the generator Q governing the underlying CTMC satisfying

$$\pi Q = \mathbf{0}, \pi \mathbf{e} = 1. \quad (15)$$

Let α be the initial probability vector of the underlying CTMC with generator Q . We can choose α in a variety of ways to model different scenarios. For example, if we want the time origin to coincide with an arbitrary arrival point, we can take $\alpha = c\pi D_1$, where c is the normalizing constant to make the vector α a probability vector. Other scenarios include the time origin to be the end of an interval during which there are at least k arrivals, the point at which the system is in specific state such as when the busy period ends or busy period begins. The most interesting case is the one where we get the stationary version of the MAP by $\alpha = \pi$.

The fundamental rate (the rate of arrivals per unit of time), λ , is given by

$$\lambda = \pi D_1 \mathbf{e}. \quad (16)$$

Often, in model comparisons it is convenient to select the timescale of the MAP so

that λ has a certain value. This is accomplished by multiplying the parameter matrices D_0 and D_1 by the appropriate common constant.

The stationary MAPs are dense in the family of all stationary MPPs [5]. Thus, MAPs can be used to represent a wide range of applications in point processes.

Special Cases of MAP. A number of well-known and useful processes are obtained as special cases of MAP. Some examples are given below:

1. *Poisson Process.* Suppose that the interarrival times are exponentially distributed with parameter η . Then in this case, we have $m = 1$, $D_0 = -\eta, D_1 = \eta$.
2. *PH-Renewal Process.* Suppose that the interarrival times are of PH with representation (β, S) of order m . Then this is a particular case of MAP with $D_0 = S, D_1 = S^0 \beta$.
3. *Markov-Modulated Poisson Process.* This is a doubly stochastic Poisson process in which the instantaneous arrival rate depends on an auxiliary Markov process. Suppose that the Markov process has the generator Q of order m . When that Markov process is in state i , arrivals occur at rate λ_i . Then, the MMPP with (single) arrivals is a MAP with $D_0 = Q - \Delta(\lambda), D_1 = \Delta(\lambda)$, where $\Delta(\lambda)$, is a diagonal matrix whose i th diagonal entry is given by λ_i . Thus, in this MAP arrival, transitions do not change the current state of the underlying CTMC. Note that a two-state MMPP in which only one diagonal of $\Delta(\lambda)$ is positive is the *interrupted Poisson process* (IPP).
4. *Markov-Switched Poisson Process.* Suppose we consider a point process that consists of geometric runs of intervals that are exponentially distributed with rates depending on the state of the underlying discrete-time Markov chain (DTMC) with m transient states and with transition probability matrix P . This process is a MAP with $D_0 = -\Delta(\lambda)$

and $D_1 = \Delta(\lambda)P$, where $\Delta(\lambda)$ is a diagonal matrix whose i th diagonal entry is given by λ_i . In this special case of MAP, only arrivals can change the current state of the underlying CTMC.

5. *PH Semi-Markov Process.* Consider a semi-Markov process $\{(X_n, Y_n) : n \geq 0\}$

with PH interval distributions; the conditional distribution of X_n given $Y_n = i$ is a PH distribution with representation $(\beta(i), S(i))$ of order m_i and $P(Y_{n+1} = j | Y_n = i) = p_{ij}$, $1 \leq i, j \leq r$. This is a MAP with parameter matrices D_0 and D_1 that are given by

$$D_0 = \begin{bmatrix} S(1) & 0 & \cdots & 0 \\ 0 & S(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S(r) \end{bmatrix}, D_1 = \begin{bmatrix} p_{11}\mathbf{S}^0(1)\beta(1) & p_{12}\mathbf{S}^0(1)\beta(2) & \cdots & p_{1r}\mathbf{S}^0(1)\beta(r) \\ p_{21}\mathbf{S}^0(2)\beta(1) & p_{22}\mathbf{S}^0(2)\beta(2) & \cdots & p_{2r}\mathbf{S}^0(2)\beta(r) \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1}\mathbf{S}^0(r)\beta(1) & p_{r2}\mathbf{S}^0(r)\beta(2) & \cdots & p_{rr}\mathbf{S}^0(r)\beta(r) \end{bmatrix}.$$

The Probability Functions of MAP. Suppose that Z denotes the time to the first arrival in a MAP with representation (D_0, D_1) of order m . The cumulative probability distribution function (CDF) and the probability density function of Z are given by

$$F(t) = \alpha[I - e^{D_0 t}](-D_0)^{-1}D_1\mathbf{e}, \text{ for } t \geq 0, \\ f(t) = \alpha e^{D_0 t}D_1\mathbf{e}, \text{ for } t \geq 0. \quad (17)$$

By taking $\alpha = \pi$, we get the stationary version for $F(\cdot)$. By taking $\alpha = (\pi D_1 \mathbf{e})^{-1} \pi D_1$, $F(\cdot)$ is the CDF of two successive arrivals in the stationary version.

Measures of MAP. Let μ_Z and σ_Z^2 denote respectively the mean and variance of Z . Then we have

$$\mu_Z = \frac{\pi D_1 (-D_0)^{-1} \mathbf{e}}{\pi D_1 \mathbf{e}} = \frac{1}{\pi D_1 \mathbf{e}} = \frac{1}{\lambda} \text{ and} \\ \sigma_Z^2 = \frac{2}{\lambda} \pi (-D_0)^{-1} \mathbf{e} - \frac{1}{\lambda^2}. \quad (18)$$

The Laplace–Stieltjes transform (LST) of the joint distribution of the first n intervals between arrivals is given by

$$\Phi(s_1, \dots, s_n) \\ = \alpha (s_1 I - D_0)^{-1} D_1 \dots (s_n I - D_0)^{-1} D_1 \mathbf{e}. \quad (19)$$

By taking $\alpha = (\pi D_1 \mathbf{e})^{-1} \pi D_1$, the serial correlation, $\rho(Z_1, Z_n)$, of the first, Z_1 , and the

n th, Z_n , interarrival times, is given by

$$\rho(Z_1, Z_n) \\ = \frac{\frac{1}{\lambda} \pi [(-D_0)^{-1} D_1]^{n-1} (-D_0)^{-1} \mathbf{e} - \frac{1}{\lambda^2}}{\frac{2}{\lambda} \pi (-D_0)^{-1} \mathbf{e} - \frac{1}{\lambda^2}}. \quad (20)$$

Generally speaking, in most applications one is interested in correlation between successive intervals. That is, we are interested in $\rho(Z_1, Z_2)$, of Z_1 and Z_2 .

Counting Process of MAP. Let $\{N(t), t \geq 0\}$ denote the associated counting process in the MAP with representation (D_0, D_1) of order m . Suppose that $\{Y_t\}$ denotes the state of the underlying CTMC with generator $Q = D_0 + D_1$. Denoting by $P(n, t) = \{P_{ij}(n, t)\}$, for $n \geq 0, t \geq 0$, the $m \times m$ matrices' (i, j) th element is defined as

$$P_{ij}(n, t) = P(N(t) = n, Y_t = j | N(0) = 0, Y_0 = i), \quad (21)$$

for $n \geq 0, t \geq 0, 1 \leq i, j \leq m$. These matrices satisfy the following differential equations [6]:

$$P'(0, t) = P(0, t)D_0, \\ P'(n, t) = P(n, t)D_0 + P(n-1, t)D_1, \quad (22)$$

with the initial conditions $P(n, 0) = 0$ for $n \geq 1$, and $P(0, 0) = I$, an identity matrix of order m .

The MAP can be viewed as a two-dimensional Markov process $\{(N(t), Y_t)\}$ on

the state space $\{(i, j) : i \geq 0, 1 \leq j \leq m\}$ with the infinitesimal generator given by

$$\hat{Q} = \begin{bmatrix} D_0 & D_1 & & & \\ & D_0 & D_1 & & \\ & & D_0 & D_1 & \\ & & & D_0 & \ddots \\ & & & & D_0 & \ddots \end{bmatrix}. \quad (23)$$

The exponential of the matrix in Equation (23) plays an important role and is given by

$$e^{\hat{Q}t} = \begin{bmatrix} P(0,t) & P(1,t) & P(2,t) & P(3,t) & \cdots \\ & P(0,t) & P(1,t) & P(2,t) & \cdots \\ & & P(0,t) & P(1,t) & \cdots \\ & & & P(0,t) & \cdots \\ & & & & \ddots \end{bmatrix}. \quad (24)$$

An efficient algorithm based on the *uniformization method* for computing the matrices in Equation (24) is discussed in Ref. 7. Jensen [8] first described the uniformization method, also called the method of randomization. It is a powerful method in the sense that (i) it allows us to interpret a continuous-time Markov process as a DTMC in which the constant times between any two transitions are replaced by independent exponential variables with the same parameter; (ii) it avoids the use of differential equations to evaluate the transition matrix.

Computation of $P(n, t)$. Defining

$$\theta \geq \max_i \{(-D_0)_{ii}\}, K = I + \frac{1}{\theta} \hat{Q},$$

$$C_0 = I + \frac{1}{\theta} D_0, \text{ and } C_1 = \frac{1}{\theta} D_1, \quad (25)$$

the uniformization method allows the computation of $e^{\hat{Q}t}$ in terms of a partial sum of the series [7]:

$$e^{\hat{Q}t} = e^{-\theta t} e^{K\theta t} = \sum_{r=0}^{\infty} e^{-\theta t} \frac{(\theta t)^r}{r!} K^r. \quad (26)$$

Since $\hat{Q} = \theta(K - I)$ (Eq. 25), using the form of $\hat{Q}$ as given in Equation (23), we see that

$$K = \begin{bmatrix} C_0 & C_1 & & & \\ & C_0 & C_1 & & \\ & & C_0 & C_1 & \\ & & & C_0 & \ddots \\ & & & & C_0 & \ddots \end{bmatrix}. \quad (27)$$

The form of the matrix K allows one to compute the powers of K as follows:

$$K^r = \begin{bmatrix} V(0,r) & V(1,r) & V(2,r) & V(3,r) & \cdots \\ & V(0,r) & V(1,r) & V(2,r) & \cdots \\ & & V(0,r) & V(1,r) & \cdots \\ & & & V(0,r) & \cdots \\ & & & & \ddots \end{bmatrix}, \quad r \geq 1, \quad (28)$$

with the convention that $K^0 = I$, an identity matrix of infinite dimension. From the structure of K given in Equation (27), it can readily be verified that $V(i, r) = 0$ for all $i > r$.

Using Equations (24), (26) and (27), the matrices $P(n, t)$ can be written in terms of $V(n, r)$ as

$$P(n, t) = \sum_{r=n}^{\infty} a_r V(n, r), \quad (29)$$

where $\{a_r\}$ is the probability mass function of Poisson with parameter θt . That is,

$$a_r = e^{-\theta t} \frac{(\theta t)^r}{r!}, r \geq 0. \quad (30)$$

The matrices $V(k, r)$ are computed recursively as follows:

Steps Involved in the Computation of the Matrices $V(k, r)$

Step 0. For a given, small number $\varepsilon > 0$, find n^* such that $\sum_{r=n^*+1}^{\infty} a_r < \varepsilon$. That is, n^* is the smallest index for which the sum of the tail probabilities of the Poisson with parameter θt is less than the prespecified number ε .

Step 1. Set $n = 0, V(0, 0) = I, P(0, t) \leftarrow a_0 V(0, 0)$

Step 2. $V(n, 0) \leftarrow 0, P(n, t) \leftarrow 0$, for $n \geq 1$

Step 3. For $1 \leq k \leq n^*$ and $0 \leq i \leq k$, do

$$V(i, k) = \sum_{j=\max\{0, i-(k-1)\}}^{\min\{i, 1\}} V(i-j, k-1) C_j$$

and $P(i, t) \leftarrow P(i, t) + \alpha_k V(i, k)$.

In any computational scheme, one has to have internal accuracy checks to make sure the computations are producing the desired results up to the degree of accuracy sought. In the current scheme, one can compute the exponential matrix, e^{Qt} using the uniformization method outlined earlier. Compare this matrix to $\sum_{i=0}^{n^*} P(i, t)$. These two matrices should be close (component-wise) to each other to the level of accuracy used. Another accuracy check is to compute the first two factorial moments matrices, $M_1(t)$ and $M_2(t)$, of the counting process $\{N(t)\}$ directly without the knowledge of $P(n, t)$ and compare them to the ones obtained from the computed values of $P(n, t)$ using the fact that $M_1(t) \cong \sum_{i=0}^{n^*} iP(i, t)$ and $M_2(t) \cong \sum_{i=0}^{n^*} i^2 P(i, t)$. Subject to the accuracy level, the results should agree. For direct computations of $M_1(t)$ and $M_2(t)$ we refer to Neuts [6].

The Palm Function of MAP. The Palm function, $H(t) = \alpha E[N(t)] \mathbf{e}$, and its density for the MAP process are given by [6]

$$H(t) = \lambda t + \alpha [I - e^{Qt}] (\mathbf{e} \pi - Q)^{-1} D_1 \mathbf{e},$$

$$h(t) = H'(t) = \lambda + \alpha (-e^{Qt} Q) (\mathbf{e} \pi - Q)^{-1} D_1 \mathbf{e},$$
(31)

where the initial probability vector, α , of the MAP can be appropriately chosen. Here, we take $\alpha = (\pi D_1 \mathbf{e})^{-1} \pi D_1$, so as to start the MAP at time origin with an arbitrary arrival. Note that when $\alpha = \pi$, the Palm function and its density are given by $H(t) = \lambda t$ and $h(t) = \lambda$ since $\pi e^{Qt} = \pi$ for all t .

Now we will consider the following five different MAPs, all of which have the same mean of 1. However, each one of them is qualitatively different. The first three labeled ERLANG, EXPON, and HYPEXP are all renewal processes, and their standard deviations are 0.44721, 1.0, and 2.24472, respectively; the other two labeled MAP-NC

and MAP-PC are correlated processes such that the first one has a correlation of -0.48891 and the second has 0.48891 as its correlation. Both of these correlated processes have the same standard deviation of 1.40952.

ERLANG. This is Erlang of order 5 with parameter 5. Recall that for this case D_0 and D_1 are given by

$$D_0 = \begin{bmatrix} -5 & 5 & 0 & 0 & 0 \\ 0 & -5 & 5 & 0 & 0 \\ 0 & 0 & -5 & 5 & 0 \\ 0 & 0 & 0 & -5 & 5 \\ 0 & 0 & 0 & 0 & -5 \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

EXPON. This is exponential with parameter 1. Here, $D_0 = -1$ and $D_1 = 1$.

HYPEXP. This is the mixture of two exponentials with mixing probabilities 0.9 and 0.1, and with parameters 1.9 and 0.19. Here,

$$D_0 = \begin{bmatrix} -1.9 & 0 \\ 0 & -0.19 \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 1.710 & 0.190 \\ 0.171 & 0.019 \end{bmatrix}.$$

MAP-NC. Here, we take D_0 and D_1 to be

$$D_0 = \begin{bmatrix} -1.00243 & 1.00243 & 0 \\ 0 & -1.00243 & 0 \\ 0 & 0 & -225.797 \end{bmatrix},$$

$$D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.99241 \\ 223.539 & 0 & 2.258 \end{bmatrix}.$$

MAP-PC. Here we take D_0 and D_1 to be

$$D_0 = \begin{bmatrix} -1.00243 & 1.00243 & 0 \\ 0 & -1.00243 & 0 \\ 0 & 0 & -225.797 \end{bmatrix},$$

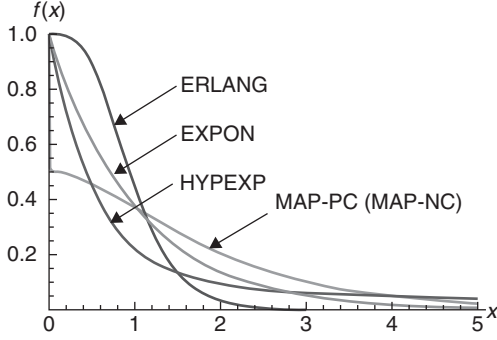


Figure 2. Probability density functions of the MAPs.

$$D_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.99241 & 0 & 0.01002 \\ 2.258 & 0 & 223.539 \end{bmatrix}.$$

For these five MAPs, we will first plot the probability density functions. First note that the probability density functions of MAP-NC and MAP-PC are identical since the row sums of D_1 are the same for both. These graphs, given in Fig. 2, were generated using Mathematica 6.0.

A quick look at Fig. 2 reveals the qualitative nature of these MAPs even though the means are all the same. Also, the density functions of the two MAPs that have nonzero correlation are identical due to the fact that the row sums of the matrix D_1 are the same for these two processes. However, when one looks at the other features of these processes (discussed below), the difference in these two processes is noticeably clear.

The Palm function and its density for these five MAPs are displayed in Figs. 3 and 4. These were obtained using the FORTRAN code that was kindly supplied to the author by its developer, Dr. Jian-Min Li.

When looking at Fig. 3 we notice that for the three renewal processes (namely, the ones labeled, ERLANG, EXPON, and HYPExp), for every fixed t , $\ln(H(t))$ appears to increase with increasing variability. However looking at all five MAPs we cannot say the same. For example, the MAP labeled MAP-PC has a smaller variance compared to HYPExp, and yet it has the largest value for $\ln(H(t))$ for

every t among all five MAPs. A similar observation holds good when looking at $\ln(h(t))$ plot as shown in Fig. 4.

Properties of MAP. Suppose that there are n assembly lines each of which is producing items that are sent to a common area for further processing (like inspecting and packing). Let the times between the products arriving to the common area from assembly line i be modeled as a MAP with parameter matrices $(D_0(i), D_1(i))$ of order m_i . Then the times between arrivals of items to the common area (irrespective of where the items are produced) is again a MAP with parameter matrices (C_0, C_1) of order $m_1, m_2, \dots, m_n$, where

$$C_0 = D_0(1) \oplus D_0(2) \oplus \dots \oplus D_0(m),$$

$$C_1 = D_1(1) \oplus D_1(2) \oplus \dots \oplus D_1(m),$$

where the symbol $\oplus$ denotes the Kronecker sum. Given two matrices A of dimension p and B of dimension q , the Kronecker sum $A \oplus B$ of dimension pq is defined as $A \otimes I + I \otimes B$, where the symbol $\otimes$ denotes the Kronecker product and I is an identity matrix of appropriate dimension. The (i, j) th block element of the matrix $A \otimes B$ obtained as Kronecker product of A and B is given by $(a_{ij}B)$. For details on the Kronecker product and sum, we refer the reader to Ref. 9.

The output or departure process of queues plays an important role, especially in the context of network of queues. For example, in a network of queues (queues in series) the input to station 2 is nothing but the output from station 1, input to station 3 is output from station 2, and so on. The output process of many finite capacity queueing models with Markovian structure can be modeled as a MAP or batch MAP depending on the nature of services (see for example, the references listed as 72, 76, 101, 114, 115, and 116 in Ref. 10). But these suffer from the curse of dimensionality when the buffer sizes are large.

Descriptors for MAP. As mentioned earlier, one of the distinct features of a MAP process is its ability to model a variety of processes in practice, and hence MAP plays

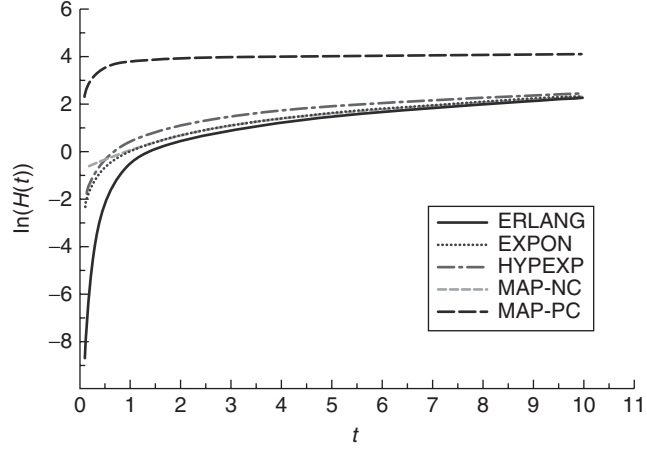


Figure 3. $\ln(H(t))$ for five different MAPs.

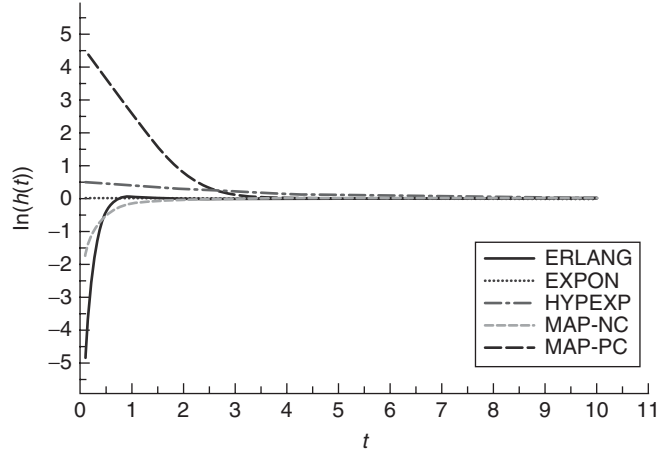


Figure 4. $\ln(h(t))$ for five different MAPs.

a significant role in stochastic modeling. A natural question is how do we qualitatively compare different MAP processes? A number of descriptors have been proposed in the literature. Some of these include (i) the lower order moments that are commonly used to characterize the interarrival times (these do not capture the dependencies among the interarrival times); (ii) the dispersion curve; (iii) the peakedness; (iv) the autocorrelation of the interarrival times; (v) the index of dispersions for intervals (IDI); (vi) the index of dispersions for the counts (IDC); and (vii) the index of burstiness.

Note that the behavior of queueing models, especially when MAP is used as input process, cannot be accurately predicted based

on the first- and second-order properties of the counting process associated with the arrival process. Here, we will discuss only one descriptor, namely, the IDC and refer to the survey paper [10] on MAP and BMAP for references to other descriptors.

Index of Dispersions for the Counts of a MAP. Let $V(N(t))$ denote the variance of the counting process $N(t)$. This function for the stationary version (namely, when $\alpha = \pi$) is given by [11]

$$\begin{aligned} V(N(t)) = & [\lambda - 2\lambda^2 + 2\pi D_1(\mathbf{e}\pi - Q)^{-1}D_1\mathbf{e}] \\ & \times t - 2\pi D_1(\mathbf{e}\pi - Q)^{-1}(I - e^{Qt}) \\ & \times (\mathbf{e}\pi - Q)^{-1}D_1\mathbf{e}. \end{aligned} \quad (32)$$

The index, $ID(t)$, of dispersions for the counts of a MAP with representation (D_0, D_1) of order m defined as the ratio of the variance

$$ID(t) = \frac{V(N(t))}{H(t)} = \frac{[\lambda - 2\lambda^2 + 2\pi D_1(\mathbf{e}\pi - Q)^{-1}D_1\mathbf{e}]t - 2\pi D_1(\mathbf{e}\pi - Q)^{-1}(I - e^{Qt})(\mathbf{e}\pi - Q)^{-1}D_1\mathbf{e}}{\lambda t}. \quad (33)$$

As $t \rightarrow \infty$, the limiting index of dispersion for counts approaches a finite value and is given by

$$ID(\infty) = \lim_{t \rightarrow \infty} ID(t) = 1 - 2\lambda + \frac{2\pi D_1(\mathbf{e}\pi - Q)^{-1}D_1\mathbf{e}}{\lambda}. \quad (34)$$

Using the same five MAPs, we now display the $\ln(ID(t))$ as a function of t in Fig. 5. Observe the noticeable differences in the five MAPs under consideration and the role of the correlation (especially the positively correlated one) played in this measure as well.

Examples. We will illustrate the versatility of the MAP with a few additional examples here. First, we would like to highlight a few general observations useful in stochastic modeling. In order to compare different models (for different scenarios) properly, one has to keep the underlying characteristics of the random variable the same. For example, let $\{X_n\}$ denote the sequence of the interarrival times of a point process with a common probability function to a queueing system.

function to the Palm function and is given by (for the stationary version by taking $\alpha = \pi$)

In order to compare different arrival processes on key performance measures such as throughput or the mean time spent in the system, we would like to fix the mean and possibly the variance of X . Fixing the mean to take a specific value is very simple and requires only scaling the random variable accordingly. However, to require that this random variable also take a specific value for the variance may need more work than simply scaling it. One of the challenging aspects dealing with MAP is the number of parameters that need to be manipulated to have the required mean, variance, and possibly the correlation values.

Generally speaking, for renewal processes (used as interarrival times for a queueing system) the key measures such as throughput and the mean waiting time in the system are monotonic (either increasing or decreasing) as the variance of the interarrival times (fixing the mean to be the same for all arrival processes) increases. However, when we look at correlated arrival processes such as monotonic properties appear to not hold good. We will illustrate a few of these through the versatility of the MAP here.

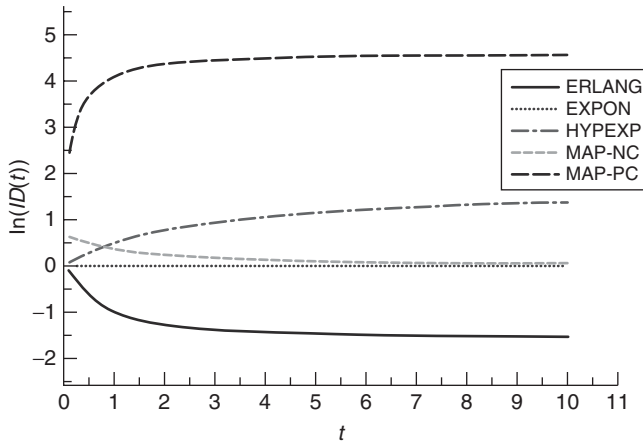


Figure 5. $\ln(ID(t))$ for five different MAPs.

Table 1. Selected Performance Measures as Functions of Input Parameters

α	p_1	p_2	Standard Deviation	Correlation	$ID(\infty)$	μ_{WTQA}	$P(\text{WAIT})$
2.00500	0.01	0.01	1.21966	-0.65224	0.50752	4.97248	0.72549
2.11120	0.11	0.01	1.16753	-0.57489	0.52947	5.04940	0.72697
2.22916	0.21	0.01	1.11290	-0.49805	0.54546	5.10572	0.72777
2.36125	0.31	0.01	1.05539	-0.42190	0.55441	5.13703	0.72780
2.51006	0.41	0.01	0.99450	-0.34670	0.55497	5.13658	0.72686
2.67899	0.51	0.01	0.92956	-0.27290	0.54542	5.09386	0.72461
2.87239	0.61	0.01	0.85965	-0.20116	0.52356	5.00637	0.72103
3.09602	0.71	0.01	0.78345	-0.13277	0.48647	4.85660	0.71543
3.35754	0.81	0.01	0.69891	-0.07030	0.43027	4.63265	0.70734
3.66750	0.91	0.01	0.60251	-0.01991	0.34963	4.31008	0.69556
0.04990	0.01	0.99	8.91064	0.00002	79.40205	357.38352	0.98898
0.05433	0.11	0.99	8.62285	0.07976	87.51089	395.06202	0.98993
0.05988	0.21	0.99	8.29377	0.15931	96.19822	429.28290	0.99018
0.06700	0.31	0.99	7.91780	0.23898	105.49734	470.31410	0.99070
0.07650	0.41	0.99	7.48308	0.31855	115.43222	520.27358	0.99140
0.08980	0.51	0.99	6.97320	0.39789	126.01865	561.76144	0.99125
0.10975	0.61	0.99	6.36593	0.47699	137.17974	612.19609	0.99122
0.14300	0.71	0.99	5.62236	0.55551	148.67829	664.61806	0.99079
0.20950	0.81	0.99	4.66808	0.63263	159.62166	728.05052	0.98986
0.40900	0.91	0.99	3.31014	0.70355	165.13209	741.66354	0.98403

First we generate a MAP with representation (D_0, D_1) of order $m = m_1 + m_2$, using two PH distributions with representations (α, T) of order m_1 and (β, S) of order m_2 as follows. Take

$$D_0 = \begin{bmatrix} T & \\ & S \end{bmatrix},$$

$$D_1 = \begin{bmatrix} p_1 \mathbf{T}^0 \alpha & (1-p_1) \mathbf{T}^0 \beta \\ (1-p_2) \mathbf{S}^0 \alpha & p_2 \mathbf{S}^0 \beta \end{bmatrix}, \quad (35)$$

where p_1 and p_2 are two parameters such that $0 \leq p_1, p_2 \leq 1$ and one can play with these parameters to arrive at close to desired values for the correlation coefficient. It should be noted that for medium to large values for the correlation, the values of m_1 and m_2 may have to be large.

For this set of examples, we choose (α, T) to be Erlang of order 4 with parameter a and (β, S) to be exponential with parameter $100a$. To get MAP with negative correlation, we fix $p_2 = 0.01$ and vary p_1 from 0.01 to 0.91 with increments of 0.1; to get MAP with positive correlation, we fix $p_2 = 0.91$ and vary p_1 from 0.01 to 0.91. The values of a are chosen so that the mean for all these MAPs are set

at 1. In Table 1, we list the values of a , p_1 , and p_2 along with the standard deviations, correlations, along with (i) the index of dispersions for the count as t approaches infinity (Eq. 34), (ii) the mean waiting time in the queue of an arriving customer, μ_{WTQA} , to a MAP/M/5 (a 5-server queueing model in which arrivals occur according to a MAP and homogeneous exponential servers with a mean service time of 4.5), and (iii) P (an arrival has to wait before getting into service). We refer the reader to article titled *The G/G/s Queue* in this encyclopedia for details on multiserver queueing model.

A striking observation of the table is the fact that in the case of positively correlated MAPs, the measure $ID(\infty)$ appears to increase as the correlation coefficient increases. Note that the same cannot be said by ignoring the correlation and looking at $ID(\infty)$ as a function of the standard deviation. The mean waiting time in the queue of an arriving customer, μ_{WTQA} as well as the probability that an arriving customer has to wait to get into service, appear to increase as the correlation increases (even though the corresponding standard deviation is decreasing). This clearly demonstrates the key role played by correlation (especially the

positive one) in stochastic modeling; this is largely ignored in the literature.

Simulation of MAP. Simulating a MAP with representation (D_0, D_1) of order m is carried out as follows: For the sake of simplicity in describing the simulation process, we will write these parameter matrices as given below:

$$D_0 = \begin{bmatrix} -\lambda_1 & \lambda_1 q_{12} & \cdots & \lambda_1 q_{1m} \\ \lambda_2 q_{21} & -\lambda_2 & \cdots & \lambda_2 q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m q_{m1} & \lambda_m q_{m2} & \cdots & -\lambda_m \end{bmatrix},$$

$$D_1 = \begin{bmatrix} \lambda_1 p_{11} & \lambda_1 p_{12} & \cdots & \lambda_1 p_{1m} \\ \lambda_2 p_{21} & \lambda_2 p_{22} & \cdots & \lambda_2 p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m p_{m1} & \lambda_m p_{m2} & \cdots & \lambda_m p_{mm} \end{bmatrix}. \quad (36)$$

Step 0. Choose an initial probability vector, $\alpha = (\alpha_1, \dots, \alpha_m)$ for the underlying CTMC with generator $Q = D_0 + D_1$. If one is interested in simulating a stationary version, then solve for the steady state probability of Q .

Step 1. Choose a random sample from the set $\{1, 2, \dots, m\}$ with probabilities given by the components of α , generate the starting state. Let this state be denoted by i . The MAP will be in this state for an exponential amount of time with parameter λ_i . Choose a random sample from this exponential distribution. This sample determines the sojourn time of the MAP in this state during this visit.

Step 2. Choose a random sample from the discrete probability function on the set $\{1^*, 2^*, \dots, (i-1)^*, (i+1)^*, \dots, m^*, 1, 2, \dots, m\}$ with probabilities $\{q_{i1}, q_{i2}, \dots, q_{i,i-1}, q_{i,i+1}, \dots, q_{im}, p_{i1}, p_{i2}, \dots, p_{im}\}$. Let the state chosen be j .

Step 3. If j belongs to the set $\{1^*, 2^*, \dots, (i-1)^*, (i+1)^*, \dots, m^*\}$, then the MAP made a transition without an arrival and the MAP spends in state j ($j \neq i$) for an exponential amount of time with parameter λ_j . Otherwise,

there is an arrival to the system and the MAP will be in phase j , $1 \leq j \leq m$, for an exponential amount of time with parameter λ_j .

Step 4. If the desired number of samples from the MAP is taken or the simulation run time is expired, stop. Otherwise, go to step 2 by replacing state i with state visited in step 3.

Using ARENA, a powerful simulation tool developed by Rockwell Software Inc., we simulated the MAP-PC using the steps outlined in steps 0 through 4 above. The simulation was run for 100,000 units and the inter-arrival times of the simulated values are plotted along with their autocorrelation function in Fig. 6. This autocorrelation function was generated using MINITAB. One can see the closeness of the actual and fitted density functions as well as the correlations (look at the autocorrelation of lag 1 in Fig. 6b to compare with the theoretical value of 0.48891).

MAP IN DISCRETE TIME

Recall that the development of MAP in continuous time started from generalizing to PH-renewal process and on to MAP. The development of MAP in discrete time starts from generalizing the classical geometric process with discrete PH-renewal process (see **Phase-Type (PH) Distributions**).

Discrete MAP (D-MAP)

Consider an m -state Markov renewal process in discrete time $\{(X_n, J_n), n \geq 0\}$ in which each transition results in an arrival. The transition probabilities are given by

$$P\{J_n = j, Y_n = k, X_n = r | J_{n-1} = i\} \\ = [(D_0)^r D_1]_{ij}, \text{ for } 1 \leq i, j \leq m, r \geq 0.$$

Let $D = D_0 + D_1$ denote the transition probability matrix of the underlying DTMC of the D-MAP. For details on DTMC, we refer the reader to the section titled "Discrete-Time Markov Chains (DTMCs)" of

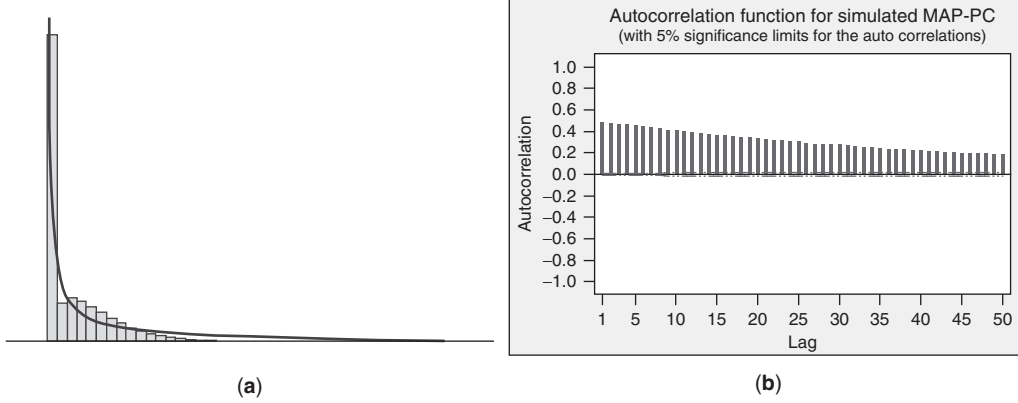


Figure 6. Simulated MAP-PC process: (a) fitted density function and (b) autocorrelation function.

this handbook. The (i, j) th entry of D_0 gives the conditional probability that in the next time unit, the MAP goes to state j with *no* arrivals given that the current state is in i . The (i, j) th entry of D_1 gives the conditional probability that in the next time unit the MAP goes to state j with an arrival given that the current state is in i . In order for a MAP to have an arrival with probability 1, it is necessary and sufficient that the matrix $(I - D_0)$ is nonsingular.

Thus, MAP in discrete time is described by the parameter matrices (D_0, D_1) of order m , whose underlying DTMC is governed by the transition probability matrix, D . Let δ be the stationary probability vector of the MC with transition probability matrix D . That is, δ is the unique (positive) probability vector satisfying

$$\delta D = \delta, \delta \mathbf{e} = 1. \quad (37)$$

Let α be the initial probability vector of the underlying DTMC governing the MAP. Then, similar to continuous-time MAP, one can appropriately choose α to model the time origin to be (i) an arbitrary arrival point; (ii) the end of an interval during which there are at least k arrivals; (iii) the point at which the system is in a specific state such as the end or beginning of a busy period. The most interesting case is the one in which we get stationary version of the MAP by $\alpha = \delta$.

The constant $\lambda = \delta D_1 \mathbf{e}$, referred to as the fundamental rate, gives the expected number of arrivals per unit of time in the stationary version of the MAP.

By taking $\alpha = \pi$, we get the stationary version for $F(\cdot)$. By taking $\alpha = (\pi D_1 \mathbf{e})^{-1} \pi D_1$, $F(\cdot)$ is the CDF of two successive arrivals in the stationary version.

Let X denote the time to the first arrival in a discrete MAP with representation (D_0, D_1) of order m . Then, the probability mass function of X is given by

$$P(X = n) = \alpha D_0^{n-1} D_1 \mathbf{e}, n \geq 1, \quad (38)$$

with $P(X = 0) = 1 - \alpha \mathbf{e}$. In most applications $\alpha \mathbf{e} = 1$ and so $P(X = 0) = 0$.

Let $\Phi(z_1, z_2)$ denote the joint probability-generating function of the first two intervals, say, X_1 and X_2 , in the MAP with initial probability vector α . Then, it is easy to verify that

$$\begin{aligned} \Phi(z_1, z_2) &= \\ z_1 z_2 \alpha (I - z_1 D_0)^{-1} D_1 (I - z_2 D_0)^{-1} D_1 \mathbf{e}, \\ 0 &\leq z_1, z_2 \leq 1. \end{aligned} \quad (39)$$

From Equation (39) one can easily obtain the means, variances, and the correlation of X_1 and X_2 , as follows.

$$\begin{aligned}
E(X_1) &= \alpha(I - D_0)^{-1}\mathbf{e}, E(X_2) = \alpha(I - D_0)^{-1}D_1(I - D_0)^{-1}\mathbf{e}, \\
V(X_1) &= 2\alpha(I - D_0)^{-2}D_0\mathbf{e} - [E(X_1)]^2, \\
V(X_2) &= 2\alpha(I - D_0)^{-1}D_1(I - D_0)^{-2}D_0\mathbf{e} - [E(X_2)]^2, \\
\rho(X_1, X_2) &= \frac{\alpha(I - D_0)^{-1}\mathbf{e} + \alpha(I - D_0)^{-2}D_1(I - D_0)^{-1}D_0\mathbf{e} - E(X_1)E(X_2)}{\sqrt{V(X_1)V(X_2)}}.
\end{aligned} \tag{40}$$

Like in the continuous-time MAP, discrete-time MAP can be described as a two-dimensional Markov process and the role of generator given in Equation (23) is played by the transition probability matrix with the same structure. The details are omitted.

The Counting Process of D-MAP

Let $\{N(t), t = 0, 1, 2, \dots\}$ denote the counting process associated with the discrete MAP with parameter matrices (D_0, D_1) of order m . Denoting by $p_{ij}(n, t) = P(N(t) = n, J(t) = j | N(0) = 0, J(0) = i)$ where $J(t)$ denotes the state of the DTMC with transition probability matrix D . Let $P(n, t)$ denote the matrix whose (i, j) th element is given by $p_{ij}(n, t)$. It is easy to verify that

$$\begin{aligned}
P(0, t+1) &= P(0, t)D_0, \quad t \geq 0, \\
P(n, t+1) &= P(n, t)D_0 + P(n-1, t)D_1, \\
n &\geq 1, t \geq 0.
\end{aligned} \tag{41}$$

The matrix probability-generating function, $P^*(z, t)$, of the process $\{P(n, t)\}$ is then given by

$$\begin{aligned}
P^*(z, t) &= \sum_{n=0}^{\infty} z^n P(n, t) \\
&= (D_0 + zD_1)^t, \\
0 &\leq z \leq 1, t \geq 0.
\end{aligned} \tag{42}$$

First, note that $P(n, t) = 0$, for $n > t$ since we are dealing with single arrivals and these occur only at discrete time points. From Equation (42) either through identifying like coefficients or through differentiation, one can derive an expression for $P(n, t)$ in terms of the parameter matrices. It can be verified that

$$\begin{aligned}
P(n, t) &= \sum_{\substack{i_1 + i_2 + \dots + i_t = n \\ i_r = 0/1, 1 \leq r \leq t}} \\
&D_1^{i_1} D_0^{1-i_1} D_1^{i_2} D_0^{1-i_2} \dots D_1^{i_t} D_0^{1-i_t}, \quad t \geq n \geq 0.
\end{aligned} \tag{43}$$

Let $M(t) = E[N(t)]$. First note that $M(0) = 0$. From Equation (41), we get the following difference equation:

$$M(t+1) = M(t)D + D^t D_1, \quad t \geq 0, \tag{44}$$

where we use the fact that $\sum_{n=0}^{\infty} P(n, t) = P^*(1, t) = D^t$. The difference equation in Equation (44) admits an explicit solution and is given by

$$M(t) = \sum_{i=0}^{t-1} D^i D_1 D^{t-1-i}, \quad t \geq 1. \tag{45}$$

For the computation of $M(t)$ it is recommended to use the recursive equation given in Equation (44) rather than the explicit expression in Equation (45). The Palm function $H(t) = \alpha M(t)\mathbf{e}$ can be evaluated depending on how the initial probability vector, α , of the MAP is chosen. If $\alpha = \delta$, the Palm function using Equation (41) reduces, as is to be expected, to $H(t) = \lambda t$. The details are omitted.

Special Cases of D-MAP

By appropriately choosing the parameter matrices (D_0, D_1) of the discrete MAP, we can get some well-known processes. Some commonly used ones are summarized here.

Geometric Process

Here $m = 1, D_0 = p, D_1 = 1 - p$, for $0 < p < 1$.

Batch PH-Renewal Process. Suppose that the interarrival times in a point process follow a discrete PH distribution with representation (β, T) of order m (see **Phase-Type (PH) Distributions**). Then this point process is a MAP with $D_0 = T$ and $D_1 = T^0\beta$.

Discrete PH Semi-Markov Process. Consider a semi-Markov process

$$D_0 = \begin{bmatrix} T(1) & 0 & \cdots & 0 \\ 0 & T(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T(r) \end{bmatrix}, D_1 = \begin{bmatrix} p_{11}\mathbf{T}^0(1)\beta(1) & p_{12}\mathbf{T}^0(1)\beta(2) & \cdots & p_{1r}\mathbf{T}^0(1)\beta(r) \\ p_{21}\mathbf{T}^0(2)\beta(1) & p_{22}\mathbf{T}^0(2)\beta(2) & \cdots & p_{2r}\mathbf{T}^0(2)\beta(r) \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1}\mathbf{T}^0(r)\beta(1) & p_{r2}\mathbf{T}^0(r)\beta(2) & \cdots & p_{rr}\mathbf{T}^0(r)\beta(r) \end{bmatrix}.$$

Markov-Modulated Bernoulli Process (MMBP). This is a discrete version of the MMPP. Suppose that P is the transition probability matrix of the MC that governs an auxiliary process generating the arrivals. When this MC is in state i , arrivals occur with probability τ_i and with probability $(1 - \tau_i)$ no arrivals occur. Then, MMBP arrivals are a MAP with $D_0 = \Delta(e - \tau)P$ and $D_1 = \Delta(\tau)P$, where $\Delta(r)$ is a diagonal matrix whose i th diagonal entry is given by the i th component of the vector $\mathbf{r}$.

Discrete-Time Platoon Arrival Process. A platoon process is one in which arrivals

$\{(X_n, Y_n) : n \geq 0\}$ with discrete PH distributions for the interarrivals; the conditional distribution of X_n given $Y_n = i$ is a discrete PH distribution with representation $(\beta(i), T(i))$ of order m_i and $P(Y_{n+1} = j | Y_n = i) = p_{ij}$, $1 \leq i, j \leq r$. This is a MAP with parameter matrices D_0 and D_1 given by

occur in short intervals followed by long intervals of no arrivals. This type of process is common in systems such as transportation and communications. Let the intraplatoon intervals and the interplatoon intervals be discrete PH distributions with representations $(\beta(1), T(1))$ of order m_1 and $(\beta(2), T(2))$ of order m_2 , respectively. Suppose that the number of arrivals in a platoon follows a discrete PH distribution with representation $(\beta(3), T(3))$ of order m_3 . Let $\beta_0(3) = 1 - \beta(3)e$. This platoon process is a MAP with parameter matrices D_0 and D_1 given by

$$D_0 = \begin{bmatrix} T(2) & 0 \\ 0 & I \otimes T(1) \end{bmatrix}, D_1 = \begin{bmatrix} \beta_0(3)\mathbf{T}^0(2)\beta(2) & \beta(3) \otimes \mathbf{T}^0(2)\beta(1) \\ \mathbf{T}^0(3) \otimes \mathbf{T}^0(1)\beta(2) & T(3) \otimes \mathbf{T}^0(1)\beta(1) \end{bmatrix}.$$

CONCLUSION

MAP is truly a versatile point process that has found tremendous practical applications in production and manufacturing, computer and communications, and transportation systems. MAP enables one to capture the realistic assumptions as much as possible and provide solutions that practitioners can implement. Invariably models involving MAP need to have algorithmic aspects built into them to make them useful for

practitioners. Hence, it is imperative that researchers who want to get into learning and using MAP should also incorporate the algorithmic aspects as part of their routine analysis. For additional information on MAP we refer the reader to review papers [4,10], books [1,6,12,13], and the paper by Ramaswami [14] dealing with the versatile point process as input to a single server queue with general service times. For latest developments in MAP, we refer the reader to the series of *Proceedings of the International*

Conferences on Matrix-Analytic Methods in Stochastic Models [15–20].

REFERENCES

1. Neuts MF. Matrix-geometric solutions in stochastic models - an algorithmic approach. New York: Dover Publications; 1995 (originally published by Johns Hopkins University Press; 1981).
2. Neuts MF. A versatile Markovian point process. *J Appl Probab* 1979;16:764–779.
3. Lucantoni D, Meier-Hellstern KS, Neuts MF. A single-server queue with server vacations and a class of nonrenewal arrival processes. *Adv Appl Probab* 1990;22:676–705.
4. Lucantoni D. New results on the single server queue with a batch Markovian arrival process. *Stoch Models* 1991;7:1–46.
5. Asmussen S, Koole G. Marked point processes as limits of Markovian arrivals streams. *J Appl Probab* 1993;30:365–372.
6. Neuts MF. Structured stochastic matrices of M/G/1 type and their 1 applications. NY: Marcel Dekker; 1989.
7. Neuts MF, Li JM. An algorithm for the $P(n, t)$ matrices of a continuous BMAP. In: Chakravorthy SR, Alfa AS, editors. Matrix-analytic methods in stochastic models. NY: Marcel Dekker; 1996. pp. 7–19.
8. Jensen A. Markoff chains as an aid in the study of Markoff processes. *Skand Aktuari-etidskr* 1953;36:87–91.
9. Marcus M, Minc H. A survey of matrix theory and matrix inequalities. Boston (MA): Allyn and Bacon; 1964.
10. Chakravorthy SR. The batch Markovian arrival process: a review and future work. In: Krishnamoorthy A, *et al.*, editors. Advances in probability theory and stochastic processes. NJ: Notable Publications, Inc.; 2001. pp. 21–49.
11. Narayana S, Neuts MF. The first two moments matrices of the counts for the Markovian arrival process. *Commun Stat Stoch Models* 1992;8(3):459–477.
12. Neuts MF. Algorithmic probability: a collection of problems. NY: Chapman and Hall; 1995.
13. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. Philadelphia (PA): SIAM; 1999.
14. Ramaswami V. The N/G/1 queue and its detailed analysis. *Adv Appl Probab* 1980;12: 222–261.
15. Chakravorthy SR, Alfa AS, editors. Matrix-analytic methods in stochastic models. NY: Marcel Dekker; 1996.
16. Alfa AS, Chakravorthy SR, editors. Advances in matrix analytic methods for stochastic models. NJ: Notable Publications; 1998.
17. Latouche G, Taylor P. Advances in algorithmic methods for stochastic models. NJ: Notable Publications; 2000.
18. Latouche G, Taylor P. Matrix-analytic methods: theory and applications. Singapore: World Scientific publishers; 2002.
19. Bini DA, Latouche G, Meini B, editors. *Stoch Models* 2005;21(2–3).
20. He QM, Takine T, Taylor P, Zhang H. Special issue on matrix-analytic methods in stochastic models. *Ann Oper Res* 2008;160.

MASS CUSTOMIZATION

ALI K. PARLAKTÜRK
JAYASHANKAR M. SWAMINATHAN
Kenan–Flagler Business School,
University of North Carolina,
Chapel Hill, North Carolina

WHAT IS MASS CUSTOMIZATION?

Mass customization can be defined as producing goods and services to meet individual customer's needs with near mass production efficiency [1]. This term was coined to parallel mass production that relied on producing large volumes of identical products like the Ford Model T line. The fundamental difference between mass production and mass customization is in the amount of product variety that is enabled by the process. Pine popularized this concept in the 1990s [2]. Over the years, it has been popular to claim that a firm provides high degree of variety in offerings, therefore the term *mass customization* has been overused.

Customization is everywhere—it is the degree of customization and the way it is offered that may be different. Automotive firms allow limited customization of individual cars through bundled options like air conditioning and leather are offered typically with an automatic transmission; Burger King offers “Have it your way” burgers that enable customers to selectively customize ingredients in their burgers; Dell computers and other electronics manufacturers successfully adapted customization to enable customers to pick and choose the key modules (such as memory, disk drive, chipset) that are associated with the computer; Sun Microsystems Java applets as well as the more popular Apple iPod and iPhones take this customization to the next level wherein the hardware is identical, but the software components (applets or iTunes files) are chosen by the customer according to their choice; the Build-A-Bear experience wherein the customer

cocreates the bear and clothing of choice is another example of customers actively involved in the customization process. There are three types of mass customization.

- *Cosmetic.* A popular approach is to customize aesthetic design of the product such as color, label, packaging, and so on. For example, M&M's (www.mymms.com) can be customized to have a personalized message, picture, and packaging. Cosmetic customization is especially popular in apparel and footwear. One example is Nike's custom-colored sneakers, which also allows the customer to imprint a name on the shoe. Nike's customization initiative is called NikeID. Customized sales accounted for 20% of the sneakers purchased from Nike.com in 2002 [3]. Another example is Zazzle offering shirts, shoes, hats, ties, and so on, with custom colors, embroidery, as well as imprinted images [4].
- *Physical Fit.* Customized physical fit can be achieved by individually customizing measurements of the product for each customer. Examples include mi adidas, Land's End, and Andersen Windows. In the mi adidas example, width, length, and insole of the shoe are customized based on the the 3D scan of the customer's foot [5]. Interestingly, an analysis of the order data shows that more than 40% of the orders are placed with different sizes for the right and left shoe [6]. Land's End, which sells custom shirts and pants, does not use 3D scanners; it asks customers to take their own measurements along with other questions about their body shape and then it relies on an algorithm fitting software developed by Archetype Solutions that finds the closest match in the extensive database of typical sizes [7]. Likewise, Andersen Windows can build a window to fit any size. Offering customized physical fit becomes more challenging

when it involves more spatial dimensions [8]. Note that customized physical fit in apparel and footwear usually requires customization along three dimensions. A number of companies either gave up their customization initiatives or went bankrupt. Examples include Levi's original spin initiative and Customatix [9], which went bankrupt selling customized sneakers.

- *Features.* Firms can also customize functional features of the product. These firms take advantage of the modular product architecture and allow customer to choose from many options for each module enabling vast number of choices for the end product. For example, Dell builds custom laptops where the customer gets to choose features like memory, hard drive capacity, CPU speed, and multimedia options. Similarly, carmakers let customer choose from various features, which can be added to the car on the basis of the customer's preferences.

Note that these three types of customization are not mutually exclusive.¹ A firm can customize its product in more than one way. For example, in addition to customized features Dell provides cosmetic customization by allowing personalized laptop covers with many different color and pattern options. Similarly, mi adidas also offers cosmetic customization allowing personalized colors and imprints on its sneakers in addition to customized physical fit. While many of the examples above are from the business to consumer segment, a large number of examples are available from the business to business segment as well. For example, Xilinx's field programmable gate arrays (FPGA) allowed customers to burn and change the circuit on the chip much later in the process. Similarly, Dell created customization in their sales process for individual business customers through customized web pages as well as staff (more examples in Ref. 13).

¹Also note that other taxonomies for mass customization are possible; for example, see Refs 10–12.

BENEFITS OF MASS CUSTOMIZATION

A firm's ability to offer mass customized products leads to a number of marketing and operational benefits.

Pricing Advantage

Mass customization creates value for customers by providing products that better match their individual preferences. This allows charging a price premium for custom products. For example, Nike is able to add 10–20% price premium for its customized sneakers that allows for custom colors and personalized imprint [3]. Adidas offers a higher degree of customization, it allows customized fit in addition to customized colors and imprinted names, and this in turn enables charging a higher 30–50% price premium [6]. In general, products that require matching physical dimensions often allow a higher price premium than products that customize just on design patterns. A customer survey on customized footwear shows that more than 50% of customers are willing to accept a price premium of 20–30% on average for custom shoes [14]. Furthermore, because mass customization allows a vast number of product versions, it enables implementing a finer price discrimination. For example, Time 121 (www.121time.com) produces custom-made Swiss watches with many modular customization options. It uses a modular pricing approach, pricing each component differently, which allows creating a product that matches more closely each customer's willingness to pay.

Mass customization also helps a firm avoid competing solely on the basis of price. Offering individually customized products enables a firm differentiate itself from its competition. A firm can offer unique attributes that cannot be directly compared to standard product alternatives (see Ref. 15 for a discussion of competition between firms following mass customization and mass production). In particular, mass customization can be a successful strategy against low-cost overseas manufacturers, as they have limited customization capabilities due to the distance from the market. Indeed, mass customization is recognized as a strategy followed by a

variety of domestic industries to fight the outsourcing of production to low-cost overseas manufacturers [16–18]. The US furniture industry is one example: US manufacturers are more successful against imports in market sectors where they offer more customization [19]. Other examples include the EU apparel industry [16] and footwear manufacturing in Finland [20].

Customer Loyalty

Customization is information intense in that a customer needs to communicate her preferences to the service provider. The firm gets into a learning relationship [21] with its customers, which increases the revenue from each customer through repeat purchases and related sales. Mass customization helps establish stronger relations with customers and gain customer loyalty. For example, when ordering a custom shirt from Lands' End [22], a customer needs to specify many physical measurements in addition to the preferred style choices. Similarly, in the case of mi adidas, the customer's feet need to get scanned to achieve customized fit. This is in addition to customers describing their style choices. In both examples, the customers can reorder with ease from the same company. Even when a customer prefers a shirt of different style, Lands' End can easily pull up the measurements from the previous order. In contrast, if the customer wants to order from a different company, she again needs to go through the process of supplying information for product customization. With mass customization, the customer invests effort to understand the firm's customization offerings and communicate her preferences, which serve as sunk costs that increase the customer's *switching costs* [23]. Furthermore, the stronger relationship established and higher levels of information acquired provide more opportunity for related sales. Thus mass customization helps build customer loyalty. The most recent example of Apple iPhone is a good one. While the sleek looks and the touch screen got the initial attention of the customer base, most of the iPhone customers rave about the application features that does small tasks such as fetching

weather, rating restaurants, finding bank information, and teaching small sentences in new languages. More importantly, no two iPhones are likely to have the same application set—demonstrating the fact that each customer's need is unique and the platform provides an opportunity for completely customization. The net result is that iPhone customers are more likely to be loyal customers.

The customer becomes coproducer of the product with mass customization. This experience itself can create additional value on top of the value of a custom product. For example, the Build-A-Bear workshop lets its customers create their custom teddy bears and other stuffed toy animals. First, the customer selects her unstuffed bear and then brings it to the stuffing machine where she gets to watch her chosen stuffing material put into the bear. This includes the memorable heart ceremony where the customer gets to choose a heart, make a special wish and place it inside her stuffed bear. The customer also gets to choose from various outfits and accessories to personalize the stuffed bear to her liking. Another example is create your own food style restaurants such as Fire and Ice (www.fire-ice.com) where the customer gets to choose the ingredients and sauces and a chef cooks the food to customer's liking. In both of these examples, the customers attain value from the customization process in addition to the end product. Note that because the customer is the coproducer of the product, he/she will have a higher tolerance and is more likely to like the product.

Better Market Understanding

In the early nineties, when Dell's computer business showed strong growth, one of the main reasons that was cited for the growth was the loyalty shown by their business customers whose needs were met more closely by Dell in comparison with its competitors. Dell accomplished this by creating individually customized web pages for each of their clients. Such an effort enabled Dell to get a better understanding of customer needs and preferences. Letting each customer design the product to her liking allows the

firm observe a more accurate reflection of customer preferences. By aggregating these observations the firm can generate better market research information without incurring any traditional market research cost. This is especially valuable for firms that also sell standard products. Indeed, most firms that offer custom products also sell standard products and they need to decide the design of products in their standard product line. In these examples, mass customization can make it possible to improve the design of standard products to better match the needs of customers. One firm that took advantage of these synergies between mass production and customization is P&G. Through its subsidiary Reflect, P&G offered customized cosmetics where customers created their own cosmetic line, mixing and matching various options like colors, scents, and skin-care preferences.² Reflect had sold nearly 10 million customized items since its launch [24]. Matching the customized order specifications with sociodemographic profile of customers and their feedback or change of specifications after the sale provided invaluable market information, which can then be used to improve the standard product offerings [23].

Better market understanding resulting from mass customization can also be very helpful for identifying trends in the marketplace and responding to them sooner. For example, when many of Dell's customers choose large-capacity hard disks in their customized configurations, Dell understands that the demand for large-capacity hard disks will go up and it can strike deals with its suppliers before the competitors notice this trend.

Reduced Inventories

An important operational benefit of mass customization is that it allows operating make-to-order as opposed make-to-stock. Because customers receive individually customized products, they are willing to

wait to some extent for their custom orders. A classic example in this regard is the European Delivery Program of BMW, where the long wait for delivery is offset by the fact that the car is built according to specifications and the customer gets a chance to pick it up from Europe.³ Industry analysts estimate that if the US car makers were to operate on a build-to-order basis that would translate to cost savings of \$3600 for each car and \$65–80 billion per year for the entire industry [26]. Avoiding carrying inventory helps avoid costs associated with inventories. For example, Dell [27] operates with less than 4 h of work in process inventory and practically no finished-goods inventory. This leads to significant savings in inventory-holding costs; furthermore, this also eliminates the forecasting risks, which are especially important for technology and fashion products. Retailers sell more than one-third of their inventory at discounted prices mainly either because of having too many items or because of stocking unpopular items [28]. Furthermore, inventory of technology products are prone to devaluation and obsolescence. The make-to-order model also has a cash flow advantage as goods are sold and at least partially paid for before they are produced. Moreover, the fact that the customers tolerate waiting for a customized product allows a firm to sell the product without even touching it as in Lands' End; in this case, the product is produced and distributed by third parties [22].

As discussed above, there are some important benefits of mass customization. However, mass customization is not for everyone. Zipkin points out that mass customization imposes special requirements for information elicitation, process flexibility,

²P&G decided to close Reflect in June 2005 and continue its mass customization initiative under its own brands such as Cover Girl [24].

³However, one needs to be careful with customization delays. A survey of manufacturing firms [25] that adopted or were considering mass customization found that delay was viewed as an important shortcoming of mass customization. For some products, customers may not have too much tolerance for delay, indeed this is one of the main reasons that few US consumers (7% in 2000) order custom cars [26].

and logistics, and many product markets are not attractive for mass customization [8]. Companies like Mattel and Levi Strauss have experimented with mass customization and abandoned these initiatives, and a number of companies went bankrupt selling customized products. Pine [2] identifies conditions under which mass customization is attractive vis-a-vis mass production. Mendelson and Parlakturk [15,29] characterize some operational and market characteristics that make mass customization attractive.

ENABLERS OF MASS CUSTOMIZATION

One of the key issues related to mass customization is that in most situations it leads to greater variance in operations and difficulties in predictability [30]. This inability to predict well leads to worse performance and hence leads to additional cost. While many firms tend to increase the price for a customized product or service, sometimes operating costs outweigh any benefits that a firm may have gained in pricing a highly customized product or service. This is one of the main reasons that mass customization in its extreme form is not widely prevalent in industry [30]. It is important to address this challenge in order to successfully implement mass customization.

Standardization Techniques

In mass customization, variability could be introduced in different ways. The most common types of variability that are associated include variability in demand both in terms of quantity and timing of orders, variability in terms of processing times since each order is different, and variability in customer knowledge and ability to participate in the cocreation of products and services. These variabilities make it difficult to operate efficiently. One way to address this is through standardization techniques. Swaminathan [30] describes four common standardization techniques, namely, process standardization, product standardization, part standardization, and procurement standardization.

Process Standardization. Process standardization (also called *postponement* or *vanilla box approach*) refers to standardizing the initial steps in the process delivery and stocking inventory at that point. Such inventory is utilized to customize individual orders that come in. The key trade-off in postponement is the ability to respond quickly and effectively to individual customization needs while optimizing the amount of additional investments in the form of semifinished inventory; Refs 31–33 provide an extensive survey of research and results on postponement while Venkatesh and Swaminathan [13] provide a rich set of real success stories and managerial implications of adopting postponement.

Product Standardization. Product standardization refers to reducing the amount of products that are actually available for the customer to buy even though on paper it might appear that everything can be customized. This way the firm is able to reduce the amount of internal variability that is created. In such situations, the firm should either price their unique products higher or quote a longer than normal lead time for delivery. This is also called *reduction* by Frei [34] in that the menu of offerings is reduced.

Part Standardization. Part standardization refers to standardizing components in the product platform; therefore, while providing high degree of variety and customization, the firm could still benefit from economies of scale on these basic components. Clearly, adopting part standardization has its negative implications as well. It is important that a customer does not perceive that level of commonality between its premium and low-quality product. However, if done right, the economies of scale and the associated innovation on parts could significantly improve the overall product and lead to higher sustained quality in such products [35]. Many successful firms such as Honda and Sony utilize part and platform standardization as an integral part of their operations strategy.

Procurement Standardization. Procurement standardization refers to creating

synergies across various components and services a firm might be buying. This is often overlooked in large organizations that grow into very independent smaller organizations that have separate procurement arms. For example, GM, in the early nineties, went through a significant change in its procurement strategy by making it more centralized and by standardizing parts and procurement across many of its product lines. This enabled the firm to cater to multiple segments in the market while obtaining synergies in procurement.

Technology Solutions

Firms can also take advantage of various technology solutions to overcome the challenges of mass customization. An important enabler is flexible manufacturing systems, which reduces the trade-off between variety and productivity. Mass customization requires the production to switch between different product versions rapidly and frequently; flexible manufacturing systems make this possible without significantly increasing the costs. Firms offering custom-fit footwear and apparel such as mi adidas and Brooks Brothers use 3D scanner technology to elicit customers' measurements, while Land's End relies on software technology running a series of algorithms that can identify a person's body size by taking just a few of his or her measurements along with questions about his or her body shape and then running them against a huge database of typical sizes. This helps create a custom fit for each customer. Companies like Amazon are able to make personalized offers taking advantage of data-mining capabilities. Amazon's recommendation engines process vast amount of information collected not only from the targeted customer but also from other customers with similar profiles. Another strategy is to provide customization through customized software while keeping the hardware identical, as in the examples of iPod and iPhones. Here, the product is decoupled into hardware and software and customization is achieved by customizing the software, which is relatively easier. In these examples, customers can add software components and content to customize the

functionality of the product on the basis of their preferences.

Involving Customers

Finally, another way to manage mass customization is to cocreate the product or service along with the customer [36]. This provides an opportunity to utilize customer labor, which can result in lower costs for the firm. For example, in the case of Fire and Ice restaurants, the customer fills a bowl with her selected vegetables, meats, and sauces, which is then cooked to the customer's liking. Here, the customer prepares the meal herself by adding the ingredients and bringing it to the centralized grill and after it is cooked, the customer carries food to her table. All these enable the restaurant avoid much of the needed labor. From the perspective of the customer, she is getting a highly customized product, while from the perspective of the restaurant, it is operating highly standardized processes, that is, replenishing ingredient bins and cooking the food brought by customers. This is possible because the customer is an integral part of the customization. Zazzle also takes advantage of greater customer involvement. Zazzle provides a platform that allows its customers to design their custom line of merchandise including T-shirts, ties, hats, mugs, shoes, and so on. The customer decides what to include in her shop (called a *gallery*) and she can customize the products by adding custom images and texts. The customer then gets a commission when another customer orders a product from her gallery. Zazzle does the order fulfillment. In this case, tasks with higher variability are done by customers whereas tasks performed by Zazzle.com have less variability. The customers design the products to be included in their stores, for other customers to buy. The products differ only in the images imprinted on them, so the production can be standardized to a great extent. Similar examples include Cafepress and Threadless.

OR/MS MODELS OF MASS CUSTOMIZATION

Mass customization is attracting increasing interest in the OR/MS literature

[15,29,37–40]. In the following, we review some of the analytical models used in these studies. Mass customization aims to eliminate each customer's sacrifice from her ideal choice by providing individually customized products. This is captured through the classical Hotelling model [41] where differences in customers' tastes are represented by their locations on the unit interval. Each product type also maps to a point on the unit interval. When a type- θ customer buys product type ζ at price p , her utility is equal to

$$U(\theta, \zeta, p) = w - p - r|\theta - \zeta|, \quad \theta, \zeta \in [0, 1], \quad (1)$$

where the reservation price $w > 0$ is the customer's willingness to pay for her ideal product and $r > 0$ is the intensity of customer preference. Mass customization enables a firm eliminate the customer's disutility by adjusting its product type to match the customer type. This framework allows studying the competition between standard and custom products [15,39,40]: firms selling standard products choose fixed locations on the unit interval (i.e., their product configurations), whereas firms selling a custom product can match any point on the interval.

There are important differences between the operations of traditional and customizing firms. A traditional firm usually carries inventory and fulfills customer demand from stock. In contrast, a customizing firm does not carry finished-goods inventory because it makes to order. Mendelson and Parlakturk [15], Xia and Rajagopalan [40], and Alptekinoglu and Corbett [42] add a delay cost to Equation (1), which is proportional to the delay customer experiences before getting her customized product. Mendelson and Parlakturk [15], Alptekinoglu and Corbett [42] model make-to-order production as an $M/M/1$ queue with higher congestion resulting in longer delays for the customers. Mendelson and Parlakturk [15], Jiang and Lee [38], and Alptekinoglu and Corbett [42] also include holding costs for standard products, which reflect the inventory cost advantage of customized products.

While the above models of mass customization are very important for developing

insights on the problem and developing strategies around product offering, yet another of the set of OR/MS models can be useful for practical operations for mass customization. These are two-stage stochastic models where, in the first stage, the firm may decide on the variants of the products to be offered and the inventory of semifinished products or parts to be stocked, while in the second stage, the firm decides how to allocate the inventory and capacity to meet the demand for customized products. While Swaminathan and Tayur [32] and Lee and Tang [43] provide details on specific models where inventory is stored as semifinished products, a much broader set of such stochastic models of this kind that are related to mass customization are discussed in Refs 33 and 44.

CONCLUSIONS

In summary, there is a growing demand for products that better match customers' individual preferences. With mass customization, the firms aim to provide each customer with his/her ideal product rather than proliferating the marketplace with product versions and hoping that customers find products that are close enough to their ideal choices. Firms mainly offer three types of customization: (i) Firms can customize appearance of the product involving colors, labeling, and so on; (ii) they can customize physical fit of the product such as customizing width, length, and insole of a sneaker; or (iii) finally, firms can customize functional features of the product adding on options that the customer chooses. There are a number of benefits that makes mass customization attractive. This includes ability to charge price premiums, establishment of stronger relations with customers, better market understanding as a result of directly observing customers' preferences, and cost savings due to operating without finished-goods inventories. While mass customization results in many benefits, it leads to some challenges due to the variability it induces to a firm's operations. To overcome these challenges, a firm can use standardization techniques such as product, part,

and procurement standardization. It can use technology enablers like flexible manufacturing systems, 3D scanners, and software technologies and finally, it can take advantage of greater customer involvement, which gives the opportunity to pass some of the task complexity on to customers.

REFERENCES

1. Tseng M, Jiao J. Mass customization. In: Salvendy G, editor. *Handbook of industrial engineering*. 3rd ed. New York: Wiley; 2001. pp. 684–709.
2. Pine BJ. *Mass customization: the new frontier in business competition*. Boston (MA): Harvard Business School Press; 1993.
3. Swartz J. Thanks to net, consumers customizing more. *USA Today* 2002. Oct 29.
4. Graham J. Zazzle aims to dazzle with on-demand merchandise. *USA Today* 2008. March 4.
5. Kamenev M. Adidas' high tech footwear. *Business Week* 2006. Nov 3.
6. Piller FT, Muller M. A new marketing approach to mass customisation. *Int J Comput Integr Manuf* 2004;17(7):583–593.
7. Schlosser J. Cashing in on the new world of me. *Fortune* 2004;150(12):244–250.
8. Zipkin P. The limits of mass customization. *MIT Sloan Man Rev* 2001;42(3):81–88.
9. Crawford KA. Customizing for the masses. *Forbes* 2000. Oct 16.
10. Gilmore J, Pine BJ. The four faces of mass customization. *Harv Bus Rev* 1997;75(1):91–101.
11. Silveira GD, Borenstein D, Fogliatto FS. Mass customization: Literature review and research directions. *Int J Prod Econ* 2001;72:1–13.
12. Piller FT, Stotko CM. Mass customization: four approaches to deliver customized products and services with mass production efficiency. In: Durrani TS, editor. *Proceedings of the IEEE International Engineering Management Conference (IEMC-2002)*. Cambridge: Cambridge University; 2002. pp. 773–778.
13. Venkatesh S, Swaminathan JM. Managing product variety through postponement: concept and applications. In: Harrison T, Lee HL, Neale J, editors. *The practice of supply chain management: where theory and applications converge*. New York: Springer; 2005. pp. 139–156.
14. Piller FT. The market for customized footwear in Europe, market demand and consumers' preferences. Technical report, Euroshoe project report. 2002. pp. 92–98.
15. Mendelson H, Parlakturk AK. Product-line competition: customization vs. proliferation. *Manage Sci* 2008;54(12):2039–2053.
16. Keenan M, Saritas O, Kroener I. A dying industry - or not? The future of the European textiles and clothing industry. *Foresight* 2004;6(5):313–322.
17. Schuler A, Buehlmann U. Identifying future competitive business strategies for the U.S. residential wood furniture industry: Benchmarking and paradigm shifts. Technical report. Northeastern Research Station: United States Department of Agriculture, Forest Service; 2003. Report NE-304.
18. Karnes CL, Karnes LP. Ross controls: A case study in mass customization. *Prod Invent Manage J* 2000;41(3):1–4.
19. Lihra T, Buehlmann U, Beauregard R. Mass customization of wood furniture: literature review and application potential. Working Paper. Canada: CENTOR, University Laval; 2005.
20. Sievanen M, Peltonen L. Mass customising footwear: the left foot company case. *Int J Mass Customisation* 2006;1(4):480–491.
21. Peppers D, Rogers M. *Enterprise one to one*. New York: Doubleday; 1997.
22. Piccoli G, Bass B, Ives B. Custom-made apparel at Lands' End. *MIS Q Exec* 2003; 2(2):74–84.
23. Piller FT, Moeslein K, Stotko CM. Does mass customization pay? An economic approach to evaluate customer integration. *Prod Plan Control* 2004;15(4):435–444.
24. Women's wear daily. P&G closes door on custom beauty internet site reflect. 2005. Jul 1.
25. Ahlstrom P, Westbrook R. Implications of mass customization for operations management. *Int J Oper Manage* 1999;19(3): 262–274.
26. Agrawal M, Kumaresh TV, Mercer GA. The false promise of mass customization. *McKinsey Q* 2001;38(3):62–71.
27. Dell M, Fredman C. *Direct from Dell: strategies that revolutionized an industry*. New York: HarperCollins Publishers; 2000.
28. Friend SC, Walker PH. Welcome to the new world of merchandising. *Harv Bus Rev* 2001;79(10):133–141.
29. Mendelson H, Parlakturk AK. Competitive customization. *M&SOM* 2008;10(3):377–390.

30. Swaminathan JM. Enabling customization using standardized operations. *Calif Manage Rev* 2001;43(3):125–135.
31. Lee H, Billington C. Materials management in decentralized supply chains. *Oper Res* 1993;41(5):835–847.
32. Swaminathan JM, Tayur SR. Managing broader product lines through delayed differentiation using vanilla boxes. *Manage Sci* 1998;44(12):S161–S172.
33. Swaminathan JM, Lee HL. Design for postponement. In: Graves SC, de Kok AG, editors. *OR/MS handbook on supply chain management: design, coordination and operations*. Amsterdam: Elsevier Publishers; 2003. pp. 199–228.
34. Frei F. Breaking the trade-off between efficiency and services. *Harv Bus Rev* 2006; 84(11):92–101.
35. Heese HS, Swaminathan JM. Product line design with component commonality and design investments. *Manuf Serv Oper Manage* 2006;8(2):206–219.
36. Prahalad C, Ramaswamy V. *The future of competition: co-creating unique value with customers*. Cambridge (MA): Harvard Business School Press; 2004.
37. Dewan R, Jing B, Seidmann A. Product customization and price competition on the Internet. *Manage Sci* 2003;49(8):1055–1070.
38. Jiang K, Lee HL, Seifert RW. Satisfying customer preferences via mass customization and mass production. *IIE Trans* 2006;38:25–38.
39. Alptekinoglu A, Corbett CJ. Mass customization versus mass production: Variety and price competition. *M&SOM* 2008;10(2):204–217.
40. Xia N, Rajagopalan S. Standard versus custom products: variety, leadtime and price competition. Working paper. 2006.
41. Hotelling H. Stability in competition. *Econ J Nepal* 1929;39(1):41–57.
42. Alptekinoglu A, Corbett CJ. Leadtime - variety tradeoff in product differentiation. Working Paper. University of Florida; 2007.
43. Lee HL, Tang C. Modelling the costs and benefits of delayed product differentiation. *Manage Sci* 1997;43(1):40–55.
44. Song J-S, Zipkin P. Supply chain operations: assemble-to-order systems. In: Graves SC, de Kok AG, editors. *OR/MS handbook on supply chain management: design, coordination and operations*. Amsterdam: Elsevier Publishers; 2003.

MATERIALS REQUIREMENTS PLANNING

W.C. BENTON, JR.
Department of Management Sciences,
Fisher College of Business, The Ohio
State University, Columbus, Ohio

INTRODUCTION

Materials requirements planning (MRP) was developed specifically to assist companies in managing dependent demand inventory and scheduling replenishment orders. Specifically, MRP is a computerized *information system* that creates planned orders for both manufactured and purchased materials. The logic of the MRP (dependent demand) approach is significantly different from reorder point (ROP) (independent demand) approach. MRP has been shown to be beneficial to a variety of manufacturing companies.

In order to manage the various types of inventories for a manufacturing operation, attributes of the items must first be analyzed in terms of cost, lead time, past usage, and the nature of demand. The nature of demand is perhaps the most important attribute. The nature of demand can be either independent or dependent. *Independent* demand is unrelated to the demand for other items. In other words, an independent item must be forecasted independently. *Dependent* demand is directly derived from the demand for another inventoried item. In manufacturing firms, demand for raw materials, components parts, and subassemblies is dependent on the demand for the final item. Thus, dependent demand should not be forecasted independently. As an example, a completed automobile is an independent demand item that is forecasted. However, we know that each automobile requires a product structure for which it would make no sense to forecast the demand for wheels

independently, simply because the wheels are dependent and their demand is derived from that for the automobile. Inventory management for dependent items is usually managed by MRP. In distribution firms, demand is usually *independent*. The order quantities for each inventory item should be forecasted separately. Stock replenishment in independent demand systems is usually determined using statistical inventory control or order-point systems. The characteristics and examples of the types of demand and the types of inventory control systems are given in the text that follows.

Types of Demand

1. Independent demand characteristics and examples:
 - automobiles, televisions, and computers
 - demand often occurs at constant rate
 - effects inventories that are managed separately from other items (e.g., finished goods made-to-stock)
 - demand is managed based on forecasts.
2. Dependent demand characteristics and examples:
 - most raw materials, components, and subassemblies
 - demand is managed based on plans [e.g., master production schedule (MPS)]
 - effects inventories that are managed to support production of other items (e.g., component parts)
 - demand often occurs in lumps.

Types of Manufacturing Control Systems

Reorder Point Inventory Control System. The ROP system is based on independent demand concept. Independent demand inventory is not directly related to the internal manufacturing operation. A flow diagram of an independent demand system is given in Fig. 1.

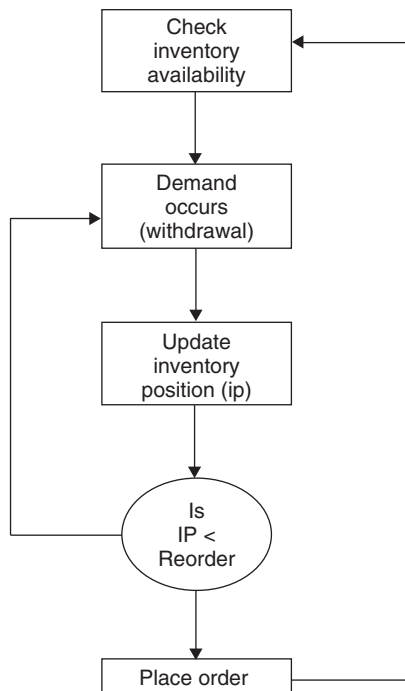


Figure 1. Flow diagram for the reorder point inventory system.

ROP systems are driven by historical data to forecast future inventory demand. The forecast assumes that past data are representative of future demand. As shown in Fig. 1, if at any time an item's inventory level falls below some predetermined level, either additional inventory is ordered or new production orders are released in fixed order quantities (FOQ). Although most early ROP systems were manual, automated ROP systems soon followed when commercial computer use increased in the 1960s.

Evolution from ROP Inventory Control to MRP Systems. During the early 1970s computerized MRP systems began to replace ROP systems as the preferred manufacturing control system. MRP systems presented a clear advantage in that they offered a forward-looking, demand-based approach for planning the manufacture of products and the ordering of inventory. This approach allowed lumpy inventory levels experienced under ROP's FOQ capabilities to be smoothed

and managed more effectively under MRP's more precise time-phased order-generation capabilities. In addition, MRP systems introduced basic computerized production reporting tools that could be used to evaluate the viability of the master schedule against projected materials demand.

MRP is a computer-driven information system that translates end item demand from the master production schedule into time-phased requirements for raw materials, component parts, and subassemblies, working backward from the due date using lead times, lot sizes, and other relevant production information. The MRP system is different from the ROP manufacturing planning and control. Prior to the introduction of MRP in the early 1970s, variations of the ROP system were used to plan and control production inventories. The MRP system makes it possible to construct a time-phased requirement record for any part number. The MRP logic also allows for detailed capacity planning models. Developing the requirements plan (materials and capacity) is basically an iterative process, where the planning is executed level by level. For example, planning for a tractor would determine requirements for tires, transmission, bucket size, and so on. However, planning for tires has to be done after the planning for wheels. Suppose the company planned to build 40 tractors and each tractor requires four wheels, then 160 wheels will be required. If 100 wheels are currently on hand and 20 wheels have already been ordered, only 40 more wheels must be ordered to complete the 40 tractors. An illustration of the MRP lot-sizing approach for time-phased requirements is given in the next section.

MRP LOT SIZING

The general lot-sizing problem for time-phased requirements for a component part involves converting the requirements over the planning horizon (the number of periods into the future for which there are requirements) into planned orders by batching the requirements into lots. The specific method for determining the order quantities for a component part is given by a lot-sizing procedure. In many cases, the lot-sizing procedure

Table 1. Net Requirements for 12 Periods

Period	1	2	3	4	5	6	7	8	9	10	11	12
Net requirements	80	100	124	50	50	100	125	125	100	100	50	100

Order cost = \$92, inventory carrying cost = \$0.5/period/unit.

performance is based on the minimization of the sum of ordering and inventory carrying costs subject to meeting all requirements for each period. An ordering cost is incurred for each planned order placed. A carrying cost is charged against the ending inventory balance in each period. The sum of these two cost components is the total inventory cost.

Lot-sizing procedures evaluate orders that cover the requirements for one or more periods. For example, in Table 1, a minimum of 80 units must be ordered for period 1 (to meet that period's requirement), but other alternatives would include an order of 180 (for the first two periods' requirements), 304 (first three periods' requirements), and so on.

In order to avoid being out of stock, there must be an order received in the first period in which there is a requirement not covered by inventory (net requirement). In order to fill the first net requirement, the firm will incur the cost of ordering (all costs associated with placing and receiving an order). It may be less expensive to combine the first net requirement with another requirement and hold the additional units in inventory until they are needed rather than pay for another order. Additional net requirements for periods beyond the first should be included as long as the cumulative units, times the number of periods to be held, times the cost to hold a unit for one period, are less than or equal to the order cost. When this is no longer true, placing another order is less costly. In general, this is the criterion used to establish the purchase order quantity for the lot-sizing procedures. The MRP concept is illustrated in the example briefly below. A more comprehensive discussion is given in the article *Lot-Sizing*.

For components of assembled products, the demands are not usually constant per unit time, and depletion is anything but gradual. Inventory depletion for component parts tends to occur in discrete "lumps" because of the lot sizing of the final product.

Requirements dependent on the final product are usually discontinuous and lumpy since requirements for these components depend on when the product is built. In some periods, there may be a few or no component requirements, and in the next, a requirement for many will occur. As an example, Table 2 shows a case in which customer demand is fairly uniform but, because of the build schedules, the requirements for the components are "lumpy." The build schedule shows periods of zero requirements before a requirement of 50 component parts is encountered. This requirement sequence, very common to component parts, is not handled well with the non-time-phased ROP techniques.

With the ROP system, the production control manager does not plan for future requirements but reacts to the current situation, which may require expediting to prevent a materials shortage. In these situations, expediting is used as a substitute for planning future requirements. An improved basis for planning to meet future needs would be to extend the requirement information throughout enough time periods to cover the entire manufacturing lead time. This provides the time-phased requirement information that is the basis for MRP systems.

MRP systems utilize substantially better information on future requirements than is possible by the non-time-phased ROP systems. MRP systems are helpful for

Table 2. An Example for Lumpy Requirements

Periods	1	2	3	4	5	6	7	8
Projected customer demand	15	15	15	10	15	10	10	10
Build schedule				50				50
Component requirements			50				50	

On hand = 50 units, safety stock = 5 units, lead time = 1 period.

companies with assembled products that have component requirements dependent on the final product. The system provides information to better determine the quantity and timing of component parts and purchase orders than is possible with the non-time-phased ROP systems.

MRP BASICS

The MRP concept provides the basis for projecting future inventories in a manufacturing operation. MRP can help improve the traditional non-time-phased order-point system because it allows the operating manager to plan requirements (raw material, component parts) to meet the final assembly schedule. In other words, MRP provides a plan for component and subassembly availability that allows certain end products to be scheduled for final assembly in the future. Once a firm's final assembly schedule has been determined and the product bills of materials have been finalized, it is possible to precisely calculate the future materials needs for the

final assembly schedule. The product bill of materials for a given finished product can be broken down, or "exploded," and extended for all component parts to obtain that product's exact requirements for each component part. The MPS sets out an aggregate plan for production. MRP translates that aggregate plan into an extremely detailed plan. An overview of the MRP system is illustrated in Fig. 2.

MRP Inputs

The MRP inputs include the MPS, the bill of materials (BOM) file, and the inventory records file.

1. *Master Production Schedule (MPS)*. The MPS is the anticipated build schedule that is derived from forecasted demand and actual customer-booked orders. The MPS is the exact time-phased build schedule needed to satisfy booked orders and forecasted demand.
2. *Bill of Materials (BOM)*. The BOM is a list of all of the raw materials, parts, subassemblies, and assemblies needed

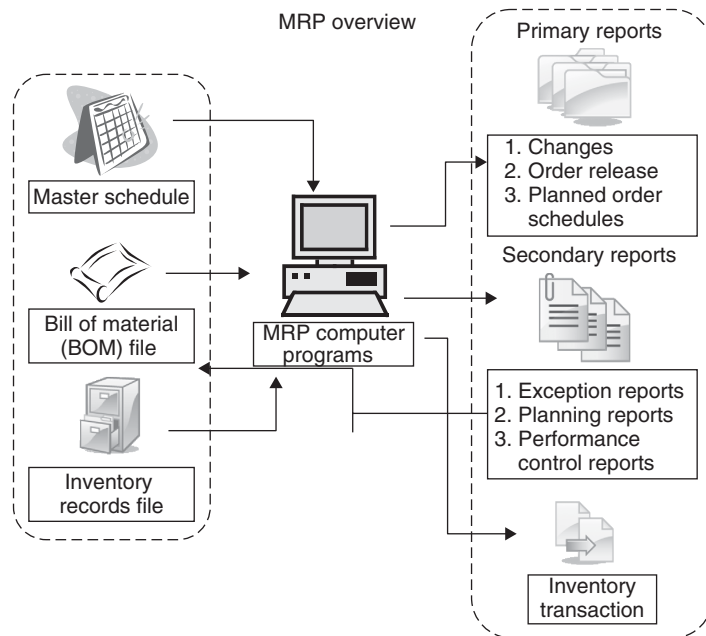


Figure 2. An overview of MRP.

to produce one unit of a product. The computerized MRP system processes the BOM file from left to right starting with the highest or zero level. Sub-assemblies and parts that go directly into the assembly of the finished product are called *level 1 components*. The parts and components that go into the assembly items are level 2, and so on. The MRP explosion involves determining the quantity and lead time for each item required to satisfy the item at that level. The results of the BOM explosion process are a time-phased requirement for specific quantities of each item necessary to assemble the finished product.

3. *Inventory Records File*. The inventory records file contains specific information on the quantity of each item on hand, on order, and committed to use in future time periods. The MRP gross to net processing logic determines the quantity available for use in a given period, and if enough are available to meet the order requirements, it commits these for use during the time period by updating the inventory record file. If there is not enough of the item available, the system includes the item order information (e.g., lot size, lead time, and scheduled receipts). A planned order release is then initiated.

MRP Outputs

The MRP outputs are the order action report, the open order report, and the planned order release report. The order action report shows which orders are to be released during the current time period and which orders are to be canceled. The open order report shows the orders that are to be expedited. The planned order release report is the time-phased plan for orders to be released in future periods. The secondary output reports are performance-control reports, planning reports, and exception reports.

The Basic MRP Record

The basic MRP record is given in Table 3. The explanation for each row is as follows:

Table 3. Time-Phased MRP Record

	Week											
	1	2	3	4	5	6	7	8	9	10	11	12
<i>J750 Engine Assembly Master Schedule</i>												
Quantity												
<i>Turbine housing requirements</i>												
Gross requirements												
Scheduled receipts ^a												
Projected available ^b												
Net requirements												
Planned order release												

^aMeasured at the end of each week.

^bReceived at the beginning of each week.

- *Gross Requirements (GRs)*. The gross requirement (GR) is the total quantity needed for an item on a weekly basis. It is assumed that the quantity will be available on Monday morning for the week in question.
- *Scheduled Receipts (SRs)*. The scheduled receipts (SRs) are orders that have already been launched, either as a shop order or purchased item. Scheduled receipts are important because they represent a commitment of resources. It should be assumed that the exact quantity and timing of these orders are certain and that the orders will be delivered in the expected period.
- *Projected Available (PA)*. The projected available (PA) is the inventory of the component at the end of the week.
- *Net Requirements (NR)*. The net requirements are the amount of the inventory item needed for the week after the GRs have been netted against available inventory and SRs.
- *Planned Order Releases (PORs)*. The planned order releases (PORs) are the quantities of net requirements that are planned to be ordered in the beginning of the period, taking into account lot sizes and time-phasing. Unlike SRs, PORs are not firm orders. The PORs are the primary outputs of the MRP system, as they represent the expected timing and quantities of future purchases and production plans.

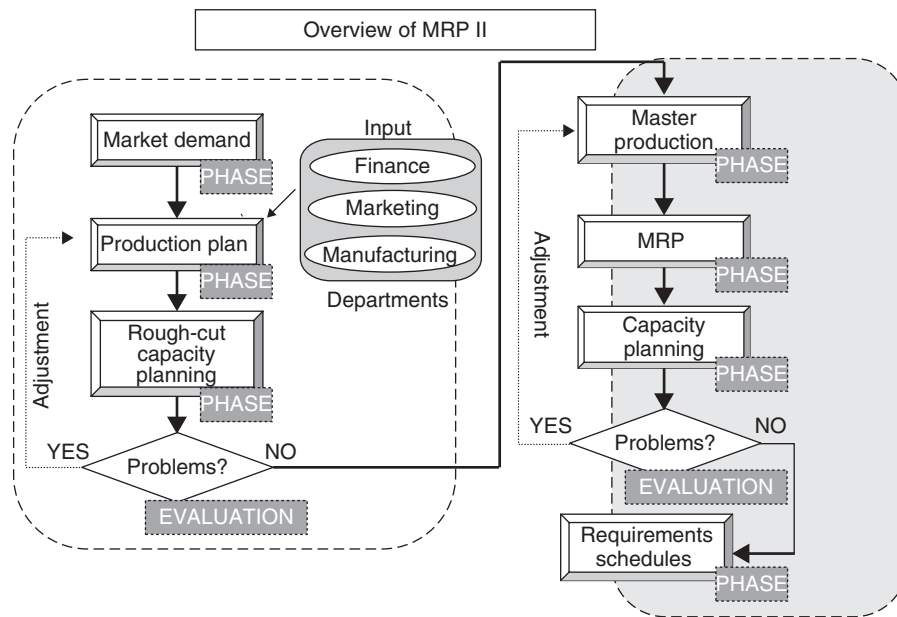


Figure 3. An overview of MRP II.

Manufacturing Resource Planning II (MRPII)

MRP II involves integrating the computerized MRP systems with financial planning, marketing, engineering, purchasing, and other functional areas. There are a number of MRP II software packages available for implementation. An overview of an MRP II system is shown in Fig. 3.

EXAMPLE FOR MRP¹

The production control manager at Point Clear, Inc. decided to implement a time-phased requirements planning (MRP) system. In May, he had attended a short course on world-class manufacturing. He now needed to convince the plant manager that requirements planning could work for Point Clear's jet engine assembly operation. The production control manager decided that a simple example was needed to illustrate his ideas.

He first prepared a master schedule for one of the engine types produced by Point

Table 4. J750 Engine Master Schedule

	Week											
	1	2	3	4	5	6	7	8	9	10	11	12
Quantity	75	25	35	50	0	45	100	50	0	40	10	80

Clear—the J750 turbo engine. The resulting master schedule is shown in Table 4. This 12-week schedule shows the number of units of the J750 engines that had to be assembled. He also decided to consider two “A” items to represent the projected requirements for the J750. The two “A” items were representative of the many other components. These two components, the turbine housing and the 562 turbine assemblies, are shown in the product structure diagram in Fig. 4. The production control manager noted that the turbine auxiliary power unit is purchased from a tier one supplier and is shipped to the main engine assembly plant. The manufacturing stages that are involved in producing a J750 involve the engine assembly department and the auxiliary power unit.

The manufacturing and purchase lead times required to produce the turbine housing and the turbine assembly components

¹Abbreviated from core study in Benton, 2010.

J750 Engine product structure

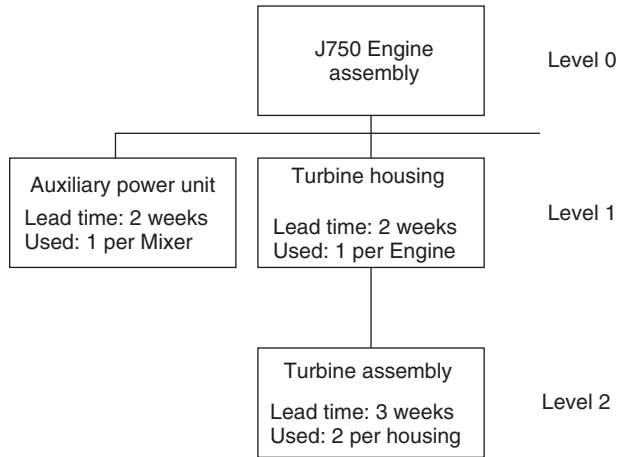


Figure 4. J750 engine product structure.

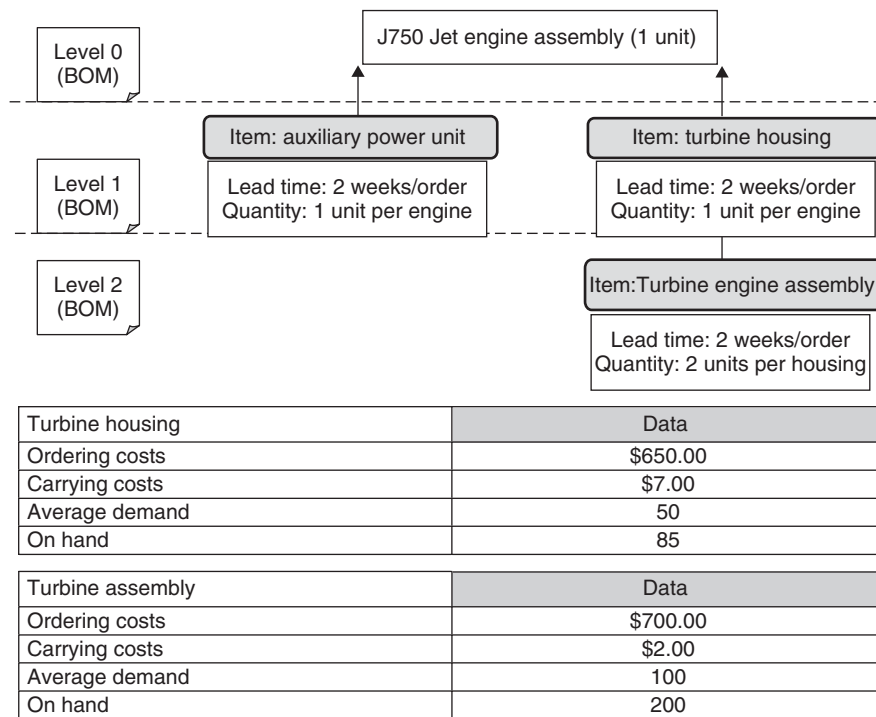


Figure 5. Point clear BOM and economic data.

are also indicated in Fig. 4. Note that two weeks are required to fabricate the turbine housings and that all of the turbine housings must be delivered to the engine assembly plant before Monday morning of the week

in which they are to be used. It takes three weeks to produce one lot of turbine assemblies, and all of the assemblies that are needed for the production of turbine housings for a specific week must be delivered to the

subassembly department stockroom before Monday morning of that week. The BOM and economic data for the Point Clear case example are shown in Fig. 5.

In preparing the MRP example, the following assumptions are made:

1. Eighty-five turbine housings are on hand at the beginning of week 1 and 25 turbine housings are currently on order to be delivered at the start of week 2.
2. Two hundred turbine assemblies are on hand at the start of week 1 and 120 are scheduled for completion at the beginning of week 2.
3. Point Clear is currently ordering using the lot-for-lot (L4L) method. The period-order-quantity (POQ) method will be compared with the L4L ordering method.
3. Twenty-five turbine housings are scheduled to be received in week 2.
4. Two hundred turbine assemblies are on hand at the beginning of week 1.
5. One hundred and twenty turbine assemblies are scheduled to be received in week 2.
6. The BOM is as given in Fig. 5.
7. Costs for each assembly are as given in Fig. 5.

The following are the alternatives:

- Order using L4L (see Table 5).
- Order using the POQ method (see Table 6)

See the lot-sizing entry on page_ for a more comprehensive discussion on lot sizing.

Solution to the Point Clear engine company MRP Example

The data pertaining to Point Clear are as follows:

1. There are no capacity issues.
2. Eighty-five turbine housings are on hand at the beginning of week 1.

LIMITATIONS OF MRP

At a glance, MRP appears to be technically robust. However, the implementation of MRP is fraught with many problems. The source of the failures of MRP can be traced to both organizational and behavioral factors. The primary factors responsible for MRP failure are: (i) the lack of top management commitment; (ii) the failure to realize that MRP

Table 5. Lot-for-Lot

	Week											
	1	2	3	4	5	6	7	8	9	10	11	12
<i>Turbine housing record^a</i>												
Gross requirements	75	25	35	50	0	45	100	50	0	40	10	80
Projected available ^b	10	10	10									
Scheduled receipts ^c		25										
Net requirements				50		45	100	50		40	10	80
Planned order release	25	50		45	100	50		40	10	80		
<i>Turbine assembly record^d</i>												
Gross requirements	50	100		90	200	100		80	20	160		
Projected available ^b	150	170	170	80								
Scheduled receipts ^c		120										
Net requirements					120	100		80	20	100		
Planned order release		120	100		80	20	100					

^aOrdering costs = \$5200.00; Holding costs = \$270.00; Total costs = \$5470.00.

^bMeasured at the end of each week.

^cReceived at the beginning of the week.

^dOrdering costs = \$3500.00; Holding costs = \$1140.00; Total costs = \$4640.00

Total cost for lot-for-lot = (turbine housing + housing assembly) = \$10110.00

Table 6. Period-Order-Quantity Solution^a

	Week											
	1	2	3	4	5	6	7	8	9	10	11	12
<i>Turbine housing record</i> (EOQ = 96.36 POQ = EOQ/average demand = 1.92) ^b												
Gross requirements	75	25	35	50	0	45	100	50	0	40	10	80
Projected available ^c 85	10	10	50				50			10		
Scheduled receipts ^d		25										
Net requirements			25	50		45	100	50		40	10	80
Planned order release	75			45	150			50		80		
<i>Turbine assembly record</i> (EOQ = 264.67 POQ = EOQ/average demand = 2.64) ^e												
Gross requirements	150			90	300			100		160		
Projected available ^c	200	50	50	300								
Scheduled receipts ^d		120										
Net requirements				40								
Planned order release	340				100		160					

^aTotal cost for lot-for-lot = (turbine housing + housing assembly) = \$7460.00

^bOrdering costs = \$3250.00; Holding costs = \$910.00; Ordering costs = \$4160.00

^cMeasured at the end of each week.

^dReceived at the beginning of the week.

^eOrdering costs = \$2100.00; Holding costs = \$1200.00; Total costs = \$3300.00

Total cost for lot-for-lot = (turbine housing + housing assembly) = \$7460.00

is merely a computer program that requires accurate input data, and (iii) that there is a need to integrate MRP with a detailed shop floor control system, such as JIT. See the JIT and Push and Pull entries given in the article **Lot-Sizing**.

Some of the specific limitations of MRP are given below:

1. MRP software assumes that lead times are certain or fixed. It is well known that for most manufacturing environments lead times are uncertain for internal and external reasons. There are also many sources of manufacturing lead time. Some of the more common sources of manufacturing lead time are: queue time, setup time, run time, wait time, and move time.
2. MRP software assumes that all products are produced on a predetermined BOM structure. In some manufacturing, the most efficient production sequence may not be consistent with the bottom-up BOM structure.
3. In many cases it is difficult to make engineering and part changes within the software.

4. MRP software schedules the shop floor on the basis of the BOM, ignoring efficient shop floor control issues.
5. MRP software assumes infinite shop floor capacity.
6. If the MRP software does not consider capacity constraints, the master scheduler must balance capacity by hand each time a change is made to the master schedule. In addition, the complexities of resolving capacity constraints become more complex as the number of master schedulers increase.
7. MRP software does not consider "floating bottlenecks." Depending on changes to the master production schedule, bottlenecks are intractable.

As can be seen from this discussion, MRP is not a stand-alone system. MRP must be integrated with top management and shop floor control if it is to be an effective manufacturing tool.

CONCLUSION

MRP is concerned with the overall logistics of the flow of materials in order to satisfy demand schedules for finished products.

The MRP system coordinates all of the information on the state and activities of the purchasing, production control, and inventory control areas, in order to determine the target, quantity, and time when all materials flow in each of these areas.

REFERENCES

1. Benton WC. Purchasing and supply chain management. New York: McGraw Hill-Irwin; 2010.

MATHEMATICAL MODELS FOR PERISHABLE INVENTORY CONTROL

STEVEN NAHMIA
Operations and MIS Department,
Santa Clara University, Santa
Clara, California

INTRODUCTION

There are thousands of papers describing mathematical models of inventory systems, but a relatively small number consider items which have limited life. However, such items play an important part in everyday life. Virtually all foodstuffs as well as most pharmaceuticals have a limited lifetime. Initial interest in these models was sparked by bloodbanking applications, but the scope of applications is far greater.¹

It is important to understand the different kinds of perishability that have been studied. Fixed life perishability means that the useful lifetime of the item is known in advance to be a fixed number of periods, say m . Exponential decay means that a fixed fraction of the inventory is lost each period. Exponential decay has also been referred to as *age independent perishability*. From the author's experience, few real problems are accurately described by exponential decay, while fixed life perishability has wide-ranging applications. For that reason, we will restrict our attention to fixed life perishability in this article.

One way of classifying inventory models is by the review period. Periodic review means that inventory levels are reviewed at predetermined discrete points in time. Continuous review means that the inventory level is known at all times. Most of the

literature on ordering policies for perishables assumes periodic review. We will also briefly review the major developments under continuous review, although this problem is largely unsolved.

We assume the following definition. A perishable item is one that has constant utility up until an expiration date (which may be known or uncertain), at which point the utility drops to zero. In general, new orders are assumed to be of fresh goods only. What makes perishable inventory problems so difficult to analyze is the fact that on-hand inventory must be age differentiated. In classical inventory analysis, the state of the system is simply the amount of inventory on hand. In that setting, items are indistinguishable. However, in the perishable inventory case, one must distinguish between items of different ages. As items age, their value to the system decreases, since there is the likelihood that they will outdate before being able to satisfy the increase in demand. This results in a multidimensional state space, which severely complicates the analysis.

Another key distinction in inventory models is between deterministic versus stochastic (random) demand. While there are some interesting deterministic problems, the most interesting and challenging problems arise under stochastic demand. For that reason we will focus on that case in this article.

The purpose of this article is to provide the reader with an overview of the basic theory and methods for modeling perishable inventory problems subject to periodic review and uncertain demand at a single installation. Hopefully, the reader will come away with an appreciation of the basic tools required for the analysis, and the important issues in the field. This introduction is not intended to be an up-to-date review of the literature.

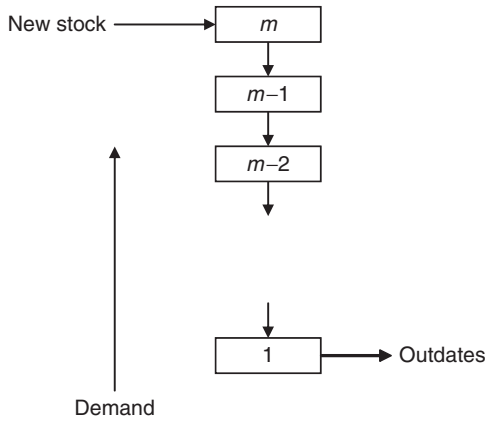
THE STATE SPACE

Assume that all orders are for fresh items. If an item is not consumed by demand during a

¹Please note that Opher Baron's piece appearing in this volume also discusses control of perishables. His article is more expository in nature and focuses more on continuous review than periodic review.

period, it remains in inventory, but ages by one period. It is obvious, in this context, that the most cost efficient issuing policy is FIFO (i.e., oldest first). This assumes that the issuing policy is under the control of the producer, not the consumer. When consumers have the opportunity to choose the issuing policy, they will typically choose LIFO (i.e., newest first). For example, in a grocery store, if all of the milk is put on display, consumers will compare expiration dates, and pick the newest cartons. On the other hand, the person stocking the shelves would display only the oldest items, thereby forcing consumers to accept the oldest stock. Very little work has been done on the LIFO issuing policy, so we will assume FIFO unless otherwise stated.

The state space is described in the figure below.



The number in each bin represents the remaining lifetime in periods of the stock in that bin. At the end of each period, remaining stock in each bin is moved to the bin below, and any remaining stock in the last bin must be discarded as outdates.²

Hence it follows that we can represent the state of the system at the start of a planning period before any new orders are placed as $(x_{m-1}, x_{m-2}, \dots, x_1)$, where the subscript

²An outdate is an item that can no longer be used to satisfy demand, either because it is at or past its expiration date, or because its utility has dropped below an acceptable level. An example of the second case is spoiled produce.

is the remaining useful life of the product. We will use y to represent new stock rather than x_m to distinguish between the state variable and the decision variable. Note that we assume zero order lead time and the following sequence of events: at the start of each period, one observes the state vector of on-hand inventories of each age level. One then orders y units of fresh stock which are delivered immediately. Subsequently, the demand in the period is realized. If demand is less than the total stock on hand, it is satisfied by issuing the on-hand stock in FIFO order. If demand exceeds the total stock on hand, the excess demand is either backordered or lost.

Two different approaches to costing have been suggested. One method is to use conventional costing, so that in the current period one pays for ordering, holding and stockouts, and the outdating that occurs in the current period. The outdating in the current period, which is $E[\max(x_1 - D, 0)]$, is independent of the current decision variable, y . Hence, using conventional costing, the effect of outdating will not be seen for m periods, when the current order is exposed to outdating. This was the approach taken by Fries [1].

Nahmias [2] adopted a different approach. His approach was based on computing the expected amount of outdating of the current order, y , m periods into the future, and placing that expectation into the one period cost function. In this way, one can construct a meaningful one period model that includes outdating, and hence incorporate the perishability aspects of the problem into the current decision without requiring m periods of a dynamic programming iteration. As we will see, a side benefit of this approach is that it leads to effective approximations. The downside of this approach is that it significantly complicates the analysis.

DETERMINING OPTIMAL POLICIES

Define the following sequence of random variables:

$$R_1 = y + \sum_{i=1}^{m-1} x_i,$$

$$R_2 = \left[y + \sum_{i=2}^{m-1} x_i - (D_1 - x_1)^+ \right]^+,$$

$$R_3 = \left[y + \sum_{i=3}^{m-1} x_i - (D_2 + (D_1 - x_1)^+ - x_2)^+ \right]^+,$$

and so on.

Then R_i represents the amount of the current on-hand inventory i periods into the future, assuming that we start with $\mathbf{x}$ and order y . Each value of R_i is a random variable, since it is a function of the future demands. Note that the use of the imbedded $+$ functions are necessary to keep track of units lost due to outdating.

To simplify the notation, define the following sequence of random variables recursively:

$$B_0 = 0$$

$$B_1 = (D_1 - x_1)^+$$

$$\vdots$$

$$B_j = (D_j + B_{j-1} - x_j)^+, \text{ for } 1 \leq j \leq m-1.$$

Interpret the random variable B_j as the total unsatisfied demand in period j after depleting the on-hand inventory that would have outdated in period j . It follows that

$$R_m = [y - (D_m + B_{m-1})]^+$$

represents the amount of the current order y that outdates in m periods.

The goal of the analysis is to compute the expected value of R_m and incorporate this into the one period model.

Define $G_n(t; \mathbf{w}_{n-1}) = P(D_n + B_{n-1} \leq t)$, where $\mathbf{w}_i = (x_i, x_{i-1}, \dots, x_1)$. Note that $\mathbf{w}_{m-1} = \mathbf{x}$. One may interpret the random variable $D_n + B_{n-1}$ as the total demand faced by the fresh stock arriving in period n , and $G_n(t; \mathbf{w}_{n-1})$ as the CDF of this random variable.

The following results establish the form of the outdating cost function.

Theorem 1. $G_n(t; \mathbf{w}_{n-1}) = \int_0^t G_{n-1}(v + x_{n-1}; \mathbf{w}_{n-2}) f(t - v) dv$.

Theorem 2. $E(R_m) = \int_0^y G_m(t; \mathbf{x}) dt$.

The term $E(R_m)$ represents the expected amount of the current order, y , which will outdate m periods into the future. By deriving an explicit penalty for outdating of the current order, one can define a one period expected cost function that balances all the relevant costs in the problem, thus avoiding the need to iterate through m periods of a dynamic programming recursion to see these effects.

Whether one uses current costs only or incorporates future outdate costs, the dynamic programming recursion has the form

$$C_n(\mathbf{x}) = \min_{y \geq 0} \left\{ L(\mathbf{x}, y) + \alpha \int_0^\infty C_{n-1}(\mathbf{s}(y, \mathbf{x}, t)) f(t) dt \right\}.$$

The transfer function $\mathbf{s}(y, \mathbf{x}, t)$ has the following form:

$$s_i(y, \mathbf{x}, t) = \left[x_{i+1} - \left(t - \sum_{j=1}^i x_j \right)^+ \right]^+,$$

for $1 \leq i \leq m-2$

and

$$s_{m-1}(y, \mathbf{x}, t) = \begin{cases} y - (t - x)^+, & \text{if excess demand is backordered,} \\ [y - (t - x)^+]^+, & \text{if excess demand is lost,} \end{cases}$$

where $x = \sum_{i=1}^{m-1} x_i$.

The one period expected cost function, $L(\mathbf{x}, y)$, may or may not include the term $E(R_m)$ depending on whether one uses Nahmias' or Fries' formulation. The object of the analysis is to show that the term in brackets is convex in the decision variable, y . The proof requires a multistep induction argument. In addition, one can show that the optimal ordering policy function when i periods remain in the planning horizon, say $y_i(\mathbf{x})$, possesses the following important property:

$$-1 \leq y_n^{(1)}(\mathbf{x}) \leq y_n^{(2)}(\mathbf{x}) \leq \dots \leq y_n^{(m-1)}(\mathbf{x}) < 0.$$

Interpret $y_n^{(i)}(\mathbf{x})$ as the partial derivative of y_n with respect to the i th component of the state vector, $\mathbf{x}$. This means that the order cost function is a decreasing function of each component of the state variable, and furthermore, it decreases more quickly in newer stock than older stock. That is, newer stock has more value to the system.

Dynamic programming is not a practical solution technique for lifetimes of more than two or three periods. As the dimension of the state variable grows, the calculations required grows geometrically (this is known as *Bellman's "curse of dimensionality"*). Hence, approximations are of particular interest.

MYOPIC APPROXIMATIONS

In order to eliminate the curse of dimensionality, it is desirable to find approximations that depend only on the total inventory on hand $x = \sum_{i=1}^{m-1} x_i$. There are several approaches to finding simple approximations. In a simulation study, Nahmias [3] compared several different forms of approximations and determined that the standard S type policy (also known as a *critical number policy*, or an *order-up-to policy*) performed well. Since it is known that the S type policy is optimal for the nonperishable problem, this would seem to be a natural choice.

Given that one has identified the form of the approximate policy, the next step would be to identify the best approximate policy from that class of policies. Finding optimal S type policies was explored by Van Zyl [4] for the case of $m = 2$, and Cohen [5] for the case of $m > 2$. Van Zyl was able to obtain a simple expression for the stationary distribution of the system state, but Cohen could not. However, the case $m = 2$ can easily be solved for an optimal solution, as the dynamic program only has a single dimensional state variable, so Van Zyl's results are of little practical value. And since Cohen could not obtain an explicit expression for the stationary distribution of the state vector, it would appear that this approach leads to pretty much a dead end.

Optimal S policies could conceivably be found by computer simulation. However,

simulations can also be quite tedious for larger values of m . Hence, approximate, and easily computed S policies are of interest. The first to explore this possibility was Nahmias [6]. His approach was somewhat unconventional. He developed approximations of the one period outdating function, $G_m(t; \mathbf{x})$, and the one period transfer function, $s(y, \mathbf{x}, t)$, that were functions of the sum x . Using these approximations, the original problem was transformed into one in which S policies were optimal. Computational tests indicated that the method resulted in effective approximations.

Chazan and Gal [7] developed a different approach to finding an approximate policy. Their method was based on the observation that if demand is bounded by S/m , then it is easy to compute the expected outdating each period under FIFO issuing. In fact, it will be simply $S/m - \mu$. Since the outdating will decrease if demand is larger, this provides an upper bound on the expected outdating each period. They also obtain a lower bound as $S/m - E(\overline{D_T})$, where $E(\overline{D_T})$ is the average demand over m periods truncated S/m .

Although Chazan and Gal do not consider the problem of optimizing the value of S , it turns out that their bounds can be applied to same methods used by Nahmias [6] to construct an approximation. Subsequent tests showed that the Chazan and Gal bounds provided a more accurate approximation than did the Nahmias bounds, and further, their approximation seemed to improve as m increased, while Nahmias' got worse as m increased. By combining the methodology of Nahmias [6] with the bounds of Chazan and Gal [7], one obtains a very efficient and easy way to compute S type approximation for ordering perishable inventory.

The bounds obtained by Chazan and Gal can be expressed respectively as $S/m - a(S)$ and $S/m - b(S)$ where

$$a(S) = \int_0^{S/m} mx f^{m*}(mx) dx + (z/k)(1 - F^{m*}(S))$$

$$b(S) = \int_0^{S/m} xf(x) dx + (S/m)(1 - F(S/m)).$$

Since we have no reason to believe that either the lower or upper bound is more

accurate, a reasonable approximation for the expected outdating based on these bounds is simply the average of the upper and lower bounds. That means we would approximate the outdating function as

$$O(z) = z/m - 0.5(a(z) + b(z)).$$

Using this bound, one can show that an approximate myopic policy is the minimizing value of $W(z)$, given by

$$W(z) = c(1 - \alpha) + L(z) + (\alpha^{m-1}\theta + \alpha c)O(z),$$

where c is the marginal order cost and $L(z)$ is the expected one period holding and penalty cost function. The minimizing value of $W(z)$ is easy to find. The approximation is to order up to the value of z^* that minimizes $W(z)$. This approach should provide very accurate S policy approximations for ordering perishables. (Details of the theoretical basis for this approximation will be presented in more detail in an upcoming monograph by this author.)

FIXED ORDER QUANTITY APPROXIMATION

Consider a system where a fixed quantity of fresh product is added to the inventory each period. This approach was suggested by Brodheim *et al.* [8] to model the operation of a blood bank. In the context of blood banking, a fixed replenishment policy assumption might be reasonable if the bank has a relatively constant pool of donors. However, fixed order quantity policies are likely to be inferior to S policies in terms of cost performance, since they do not respond to demand. Since an S policy orders exactly the previous period's demand, it is responsive to highs and lows in the pattern of demand realizations.

However, the analysis is significantly simplified under a fixed order quantity policy. Suppose q units are brought into the system each period. Then the age distribution of on-hand inventory will be of the form $(q, q, q, \dots, r, 0, 0, \dots, 0)$ where $0 \leq r \leq q$. Notice that this means that knowing only the total stock on hand uniquely defines

the state vector. If the product lifetime is m periods, and the total inventory on hand starting any period is i , then every age category will contain exactly q units up until age category $[i/m]$, which will contain exactly $r = i - [i/m]q$ units. Interpret $[x]$ as the largest integer less than or equal to x .

This means that knowing only the total inventory on hand provides a complete description of the system state, and the total starting stock forms a Markov chain of one dimension. This obviates the need to keep track of the age distribution of stock, and significantly simplifies the analysis. The range of possible values of state is 0 to qm . Since under reasonable assumptions, the resulting Markov chain is recurrent, there exists a stationary distribution $\pi = (\pi_1, \pi_2, \dots, \pi_{qm})$ satisfying $\pi P = \pi$, where P is the one step transition matrix for the chain.

Even though the state is only one dimensional, the number of states is likely to be large for most applications. In the blood banking application, $m = 21$, and for a value of q of 100 the resulting chain has 2100 states. This is certainly not beyond current computing power, but would be, at best, unwieldy. For that reason, approximations are considered, which provide reasonably tight bounds on the probability of shortage and the expected daily outdating. The authors do not consider the problem of optimizing the value of q . While potentially useful in the right environment, this approach is of limited interest (as noted above) due to the fact that a fixed order quantity policy does not respond to demand fluctuations. In the blood banking context, a sudden surge in demand brought on by a catastrophe would certainly be a cause to seek larger supplies of fresh blood.

RANDOM LIFETIMES

An important extension of the basic model is to the case of uncertainty in the product lifetime, which could be an accurate description of obsolescence.³ Such a model

³Obsolescence means that the environment changes and the item is no longer of use. It differs

might also be a more accurate descriptor of fresh food than assuming fixed life. An issue that arises when dealing with lifetime uncertainty is that of crossovers. That is, is it more accurate to assume that items outdate in the same order that they enter inventory or not?

This is precisely the same issue that arises in the standard dynamic inventory model with lead time uncertainty. That is, if orders are placed periods i and j , where $i < j$, should one assume that the order placed in period i arrives prior to that of the order placed in period j ? If so, successive lead times must be dependent random variables. This presents a challenge for modeling, as one rarely finds stochastic models incorporating dependency. This problem was solved very elegantly by Kaplan [9]. Kaplan's approach was adapted by Nahmias [10] to the perishable inventory problem.

The idea is to define a set of surrogate random variables that are independent, and build the distribution of order arrival from the surrogate variables. Let $A_1, A_2, \dots$ be a sequence of independent and identically distributed random variables defined on the discrete set $\{1, 2, \dots, m\}$, where m is interpreted as the maximum lifetime. Note that these surrogate random variables are not successive lifetimes of orders. The assumption is that if the event $(A_n = k)$ is realized in any period, then one assumes that all of the on-hand stock k periods or older outdates in the current period. If no stock is on hand of age k after satisfying demand, no outdating occurs.

This framework can be incorporated into both the analysis of optimal policies and approximate policies. Computation of an optimal policy still requires solution of an $m - 1$ dimensional dynamic program, so again approximations are of critical importance.

We will not go into the details here, but one can show that the expected outdating of the current order y can be expressed as

$$\sum_{k=1}^m r_k \sum_{j_1=2}^m \dots \sum_{j_{k-1}=k}^m \prod_{i=1}^{k-1} P_{j_i} \times \int_{v_k}^{y+v_k} G_k(u, v(k-1)) du,$$

where $P_i = P\{A_n = i\}$, $r_k = \sum_{i=1}^k P_i$, and G_k is the expected outdating function. While this looks very daunting, it can be simplified significantly. We will not present details here, but if one defines

$$q_k = \sum_{k=1}^m r_k \sum_{j_1=2}^m \dots \sum_{j_{k-1}=k}^m \prod_{i=1}^{k-1} P_{j_i},$$

then one can show that $q_k = r_k \prod_{i=1}^{k-1} (1 - r_i)$.

Again, one may adopt the Nahmias' [6] approach to the Chazan and Gal [7] bounds to obtain a myopic one period cost of function of the form

$$W(z) = c(1 - \alpha)z + L(z) + \sum_{k=1}^m q_k (\alpha^{k-1}\theta + \alpha c) \mu_k(z).$$

Since $W(z)$ is quasi-convex, the minimizing value occurs where the first derivative is zero. This results in z^* satisfying

$$c(1 - \alpha) + L'(z) + \sum_{k=1}^m q_k (\alpha^{k-1}\theta + \alpha c) \left\{ \frac{F(z/k) + F^{k*}(z)}{2k} \right\} = 0.$$

One then orders up to z^* each period. Comparisons of this approximation with the optimal policy for lifetimes of 2 and 3 periods, and for a variety of cost and demand parameters, shows close agreement between optimal and approximate policies. One would hope that this would signal that these approximations would also be accurate for longer lifetimes, although such tests have not been done.

INCLUSION OF A SET-UP COST

All of the work reviewed thus far on the periodic review perishable inventory problem

from perishability in that one cannot predict in advance exactly when an item will become obsolete in most applications.

assumes that there is no fixed cost (set-up cost) for placing an order. Fixed order costs can rarely be ignored in practice. In rare occasions, when deliveries are prescheduled on a daily or weekly basis, fixed order costs are sunk costs. However, it is more common that fixed costs determine the optimal frequency of orders, and cannot be ignored.

The results of this section are based on Nahmias [11]. Both optimal and approximate ordering policies are considered. The optimal policy is derived only for the one period problem,⁴ but it is likely the multiperiod version possesses the same policy structure. The optimal policy requires calculation of two functions of the starting state, $s(\mathbf{x})$ and $y(\mathbf{x})$, with $s(\mathbf{x}) \leq y(\mathbf{x})$. These functions determine the optimal policy as follows: if $s(\mathbf{x}) > 0$, the optimal policy is to place an order for $y(\mathbf{x})$, and if $s(\mathbf{x}) \leq 0$, then no order should be placed. As in the case of no setup cost, $y(\mathbf{x})$ minimizes the expected one period cost function, $B(\mathbf{x}, y)$. (That is, $B(x, y(\mathbf{x})) = \min_{y \geq 0} B(\mathbf{x}, y)$.) The function $s(\mathbf{x})$ solves $B(\mathbf{x}, s(\mathbf{x})) = B(\mathbf{x}, y(\mathbf{x})) + K$, and $s(\mathbf{x}) \leq y(\mathbf{x})$, where K is the fixed set-up cost. Since $B(\mathbf{x}, y)$ is convex in y , these equations uniquely define $s(\mathbf{x})$. Note that Nahmias only derived first period results, but given the fact that the structure of the optimal policy is the same for the first period problem as the multiperiod problem when there is no setup cost present, it is reasonable to speculate that this is the case here as well.

Except for product lifetimes of two or three periods, the optimal policy is too complex to be practical. Hence, approximations are of particular interest. The natural form of an approximate policy is (s, S) , since this policy is known to be optimal for the nonperishable problem, and is easy to implement in practice. (An (s, S) policy is one in which an order is placed up to S if and only if the starting inventory position is less than s .) The approach to constructing an approximation follows closely the method discussed above for the case of $K = 0$. That is, the expected outdate function is approximated by a function of the total on-hand inventory

using either the Nahmias or Chazan and Gal bounds, and this approximation is then used to construct a surrogate problem whose optimal solution is used to approximate the original problem. The surrogate problem has only a single dimensional state variable, and is solved by dynamic programming.

Extensive computations reported in Nahmias [11] indicate that the approximation generally gives costs within 1% of the optimal, although because of the difficulty of computing an optimal policy, the comparisons were done for $m = 2$ only. However, the results would seem to indicate that this is a reasonable approach for approximating dynamic perishable inventory problem with a fixed order cost.⁵

MULTIPRODUCT MODELS OF PERISHABLES

Several applications in blood banking give rise to several multiproduct perishable inventory models. The earliest of which we are aware is that developed by Nahmias and Pierskalla [12]. The model was motivated by the operation of a blood bank that stores both fresh and frozen blood. Since frozen red blood cells have a much longer lifetime than fresh blood, the authors assumed that one product had a finite lifetime and the other an infinite lifetime. The model allows for one way substitution, that is, the infinite life time product can be substituted for the finite lifetime product (but not vice versa).⁶

The model developed is a direct extension of the approach applied by Nahmias [2]. Assume that at the beginning of each planning period prior to the realization of demand, orders for both product 1 (the perishable product) and product 2 (the nonperishable product) must be placed. There is a single demand source that depletes first

⁴That is, one planning period. The product lifetime is still assumed to be m periods.

⁵Note that in some cases, such as supermarkets, fixed costs may be unimportant as many items may be ordered simultaneously. However, there still remain many applications in which fixed costs cannot be ignored.

⁶The interested reader should refer to the Rodney Parker piece in this volume for a discussion of multiproduct inventory control for durable goods.

from product 1 according to FIFO, and then product 2 until the demand is either satisfied or backordered. (Note: in the context of blood banking, it is probably more accurate to assume that excess demand is lost rather than backordered. This would mean that excess demand is made up by emergency shipments from other blood banks. The backorder assumption is not crucial to the analysis.)

Costs are charged separately for holding at h_1 for product 1 and h_2 for product 2. Outdating of product 1 is charged at θ and there is a penalty cost of p for unsatisfied demand. Let $\mathbf{x}$ be the vector of on-hand inventories of the perishable product, $x = \sum_{i=1}^{m-1} x_i$, and y the size of the order of fresh stock. In addition, define x^2 as the on-hand inventory of product 2 at the start of any period, and let z be the on-hand inventory of product 2 after ordering in any period, so that the order quantity of product 2 is $z - x^2$. Then the one period expected cost function has the form:

$$\begin{aligned} L(\mathbf{x}, y, z) = & c_1 y + h_1 \int_0^{x+y} (x+y-t)f(t) dt \\ & + \theta \int_0^y G_m(u; \mathbf{x}) du + c_2(z - x^2) \\ & + h_2 z F(x+y) \\ & + h_2 \int_{x+y}^{x+y+z} (x+y+z-t)f(t) dt \\ & + p \int_{x+y+z}^{\infty} (t - x - y - z)f(t) dt. \end{aligned}$$

Assume the following relationships among the cost parameters:

- (i) $0 \leq h_2 \leq h_1$,
- (ii) $0 < c_1 < c_2$,
- (iii) $p > (1 - \alpha)c_2$,
- (iv) $0 \leq (1 - \alpha)(c_2 - c_1) + (h_2 - h_1) < \theta$.

The rationale for these assumptions is the following. Assumption (i) says that the cost of storing the nonperishable product is at least as large as the cost of storing the perishable product. Assumption (ii) says that the cost of acquiring the nonperishable product exceeds the cost of acquiring the perishable product.

Assumption (iii) is common for single product inventory systems. It says that the unit cost of excess demand exceeds the one period discounted cost of acquiring a new nonperishable product. This guarantees that it is not more economical to incur stockouts than purchase new stock—otherwise there would be no motivation to operate the system. The expression in the final case corresponds to the cost of acquiring one unit of product 2 and salvaging it one period later minus the cost of acquiring one unit of product 1 and salvaging it one period later plus the difference in the holding costs. If this term were negative, one would never order product 1. If this term exceeded the outdate cost, it would never be optimal to order product 2.

The analysis proceeds by defining the functional equations based on the expanded state variable $(\mathbf{x}, x^2)$. The authors proceed to show that the optimal policy for the case $m = 2$ can be described by regions in the (x, x^2) plane. Because of perishability, the optimal policy has the unusual property that when one is in an ordering region, the optimal policy calls for ordering an amount that does not take one to the boundary of the ordering region, but to a point in the interior of the region.

In his review, Nahmias [13], showed how the approximation techniques discussed in the previous section could be adapted to this problem. This approximation was never tested, but is likely to be accurate, based on results for the single product problem.

Another multiproduct perishable inventory model, which was suggested by blood banking applications, was explored by Deuermeyer [14]. Consider two production processes, A and B. Process A produces two products (1 and 2), while Process B produces only product 2. Both products are perishable, and both face independent stochastic demands. In the blood banking application, product 1 is platelets (a blood component) and product 2 is whole blood. Process A separates some of the whole blood into platelets and packed red cells, while Process B just produces whole blood.

Suppose that the lifetime of product 1 is m periods and the lifetime of product 2 is l periods. The state vector consists of the

starting stocks of each product at each age level, represented as $(\mathbf{x}, \mathbf{w})$ having dimension $(m - 1) + (l - 1)$. The decision variable is the pair (a, b) , which corresponds to the production quantities for Process A and Process B respectively. Assume that a units of Type A production yields $\eta_1 a$ of product 1 and $\eta_2 a$ of product 2, while b units of Type B production yield b units of Product 2.

Deuermeyer develops a one period model using the expected future outdating of both products in the one period cost function as we did in the section titled “Determining Optimal Policies.” This makes it likely that the structure of the optimal one period policy is the same as the structure of the optimal multiperiod policy. (Unfortunately, due to the complexity of the problem, it was possible to obtain first period results only.)

AGE DIFFERENTIATED DEMAND STREAMS

In a recent study, Deniz *et al.* [15] treat the case where customers demand items of different ages. This situation arises, for example, in the blood banking context where newer blood is required for certain critical surgeries. The authors consider only the case of $m = 2$, so that the age differentiated demand is for either fresh stock or one period old stock. They note that this framework includes both LIFO and FIFO issuing policies.

This study does not treat optimal policies, but considers two heuristic policies. They are labeled TIS and NIS. The TIS policy is what we previously referred to as an *order-up-to-policy* or an *S policy*. The NIS policy is what we previously referred to as *fixed order quantity policy*: one orders S units of new stock independent of the inventory position.

A unique aspect of this study is the assumptions made regarding upward and downward substitution. First, suppose there are given fractions of customers denoted by $0 \leq \pi_D \leq 1$ and $0 \leq \pi_U \leq 1$ who are willing to accept downward and upward substitution respectively. Downward substitution means that a new product is sold to a customer who requests old stock, and upward substitution is the opposite. The authors examine the following four scenarios: no substitution, downward substitution only, upward

substitution only, and both downward and upward substitution. What makes this study interesting and unique is the consideration of four distinct issuing policies. The vast majority of the perishable inventory literature assumes FIFO issuing; and a few studies assume LIFO issuing. The primary focus of this research is to determine under what conditions one of these four issuing rules is preferred given the replenishment policy. In particular, they delineate 10 scenarios involving relationships among the costs and substitution parameters resulting in a preferred issuing policy. We refer the interested reader to their work for further details.

LIFO INVENTORY SYSTEMS

Most of the mathematical models for ordering perishable inventory assume that demands are filled according to FIFO. That is, one depletes the oldest inventory first. From the point of view of the manufacturer, FIFO is the most efficient issuing policy. However, in many circumstances, the decision is not up to the manufacturer, but up to the consumer. One example of this is age differentiated demand streams discussed in the previous section. Other examples are products such as fresh milk that are stamped with an expiration date. Consumers seek the product with the latest expiration date. In some circumstances, paperback books, music CDs, and film DVDs follow LIFO issuing.⁷ Customers are generally much more interested in newer rather than older products. This is typically the case for a video rental service, such as Netflix. Customers queue up for new films, while tens of thousands of older films are rarely requested.

To our knowledge, the only analysis of ordering policies in the context of LIFO issuing is due to Cohen and Pekelman [16]. Using the notation developed earlier, the state of the system after placing an order can be

⁷These are examples of products that are subject to obsolescence rather than perishability. However, the effect is the same: at some point these items are no longer demanded by the consumer.

represented as $(y, x_{m-1}, x_{m-2}, \dots, x_1)$. Under LIFO issuing, the one period transfer function is given by

$$s_i(y, \mathbf{x}, t) = \left[x_{i+1} - (t - y - \sum_{j=i+2}^{m-1} x_j)^+ \right]^+,$$

for $1 \leq i \leq m-2$

and $s_{m-1}(y, \mathbf{x}, t) = (y - t)^+$ under lost sales.

(Note: $s_{m-1}(y, \mathbf{x}, t) = y + \sum_{j=1}^{m-1} x_j - t$, if $t > y + \sum_{j=1}^{m-1} x_j$ under backordering. However Cohen and Pekelman only considered the lost sales case.)

The analysis in Cohen and Pekelman does not proceed along the lines of prior work. That is, they do not develop the dynamic programming functional equations in order to analyze the properties of an optimal ordering policy. Rather, they assume fixed amounts of fresh stock that enter the system, and proceed to analyze the underlying stochastic processes arising from the age distribution process. In fact, to our knowledge, no one has attempted to discover the properties of an optimal policy under LIFO issuing.

The authors develop an analysis of the demand flow process and show how this relates to a ladder height process, which can be posed as a Weiner–Hopf integral equation. However, they obtain a solution to this equation only for the case $m = \infty$ and exponentially distributed demand, which does not shed much light on the perishable case or the general demand case. For $m < \infty$, explicit results are obtained for the case $m = 1$ only and shortage and outdate bounds are obtained for the more general case. It is unclear what the value of these bounds would be for constructing order policies.

Obviously, the LIFO problem has difficulties that the FIFO problem does not. Perhaps, this is why the problem has remained largely unsolved, and is an area of research opportunity.

INVENTORY DEPLETION MANAGEMENT

The inventory depletion management problem is the following. Items are stored in a stockpile for the purpose of serving a demand

in the field. Items typically age at different rates in the stockpile and in the field. The stockpile may or may not be replenished, and there may be different costs to replenish the stockpile and make emergency replacements in the field. The vast majority of the analysis on this problem is to determine conditions under which either FIFO (i.e., issuing the oldest item next) or LIFO (issuing the newest item next) is optimal.

How does this problem relate to the problem of ordering perishable inventory? The problems are similar in that they both deal with optimal management of perishable inventory. But that is where the similarity ends. In the basic perishable inventory problem, items are assumed to be of uniform utility to the field irregardless of their ages at issue, as long as they have not expired in inventory. This means that from the inventory management point of view, it is optimal to issue the items in FIFO order. The classic inventory depletion problem is static in most cases: most of the research assumes a fixed stockpile of items with no opportunity for replenishment. There is also no external demand, other than the “field.” Items are simply issued on a one-at-a-time basis as they die in the field. For the ordering problem, once items are “issued” (i.e., leave inventory to satisfy demand) they leave the system. One is not concerned with how long these items last in the field. In the inventory depletion management problem, the key driver is the relative rates of aging in the stockpile and in the field.

Hence there is little crossover between the inventory depletion problem, and that of ordering perishable inventory. It is clear that there is a large body of real problems in which one must determine replenishment policies for perishables. On the other hand, the inventory depletion problem is more specialized, and would appear to have fewer applications in the real-world.

Irregardless, there is a fairly large body of research on the depletion problem, which is why we review it in detail here. The first study of the depletion problem was put forth by Greenwood [17]. Greenwood’s interest in the problem clearly arose from military applications. As he notes in his introduction, the

prevailing thinking of the time was that FIFO was the preferred issuing order, but that LIFO might be a better choice in many cases. Greenwood notes that the basic trade-off is that under FIFO, issuing will be occurring frequently, but fewer items will expire in the stockpile; while under LIFO, one will observe longer field lives, but many items might expire in storage. Greenwood's basic observation is that if the field life function is linearly decreasing with slope one, both LIFO and FIFO are optimal, but if the field life function is convex, LIFO is optimal and if it is concave FIFO is optimal. He notes that when the field life function is convex and decreasing, FIFO requires substantially more replenishment of the stockpile. Formal proofs of these results were not provided.

Later, Derman and Klein [18] consider a rigorous analysis of a more narrowly defined problem. Virtually all of the research that followed is based on Derman and Klein's formulation, which is the following. Assume a fixed stockpile of n items of varying ages, say $(S_1, S_2, \dots, S_n)$. Items are issued one-at-a-time to the field upon expiration of the previously issued item. An item issued at age S has a field life of $L(S)$, where $L(S)$ is a known function (later extensions would consider the case, where field life is a random variable). The objective is to determine the sequence in which to issue the items so as to maximize the total field life (or total expected field life) of the stockpile.

Consider the following example. Many portable electronic items, such as digital cameras, mp3 players, and flashlights use disposable batteries. As batteries fail, they are replaced. Batteries age in storage, but much more slowly than while in use. Given a stockpile of batteries of varying ages, in what order should the batteries be issued to maximize the total useful lifetime of the stockpile?

As noted above, in the case of the standard perishable inventory ordering problem, the issue of field life never arises. Items are assumed to age at unit rate while in storage, but once issued to meet demand, they leave the system. In the basic perishable inventory model, a new item with remaining lifetime m has the same utility to the buyer as an

item with one unit of time remaining in its lifetime. To this writer's knowledge, the only study which considers both ordering and issuing of perishable items is the unpublished work of Veinott [19], but only deterministic demand was considered in this study. Treating both ordering and issuing of perishable inventory under general stochastic demand would be very difficult.

Consider the field life function, which defines the relative rates of aging in the stockpile and in the field. The simplest assumption is that items age at the same rate in the field as in the stockpile. This would correspond to a field life function of the form $L(S) = m - S$, where m is the useful lifetime of a new item. Note that in this case, we would require that $L(S) = 0$ for $S > m$. As we will see, this is a trivial case in which all issuing policies are optimal. However, if items age at different rates in the stockpile and in the field, the field life function will not have unit slope. Suppose, for example, that items in the stockpile are refrigerated and age at exactly half the rate of items in the field. In this case, the appropriate field life function is $L(S) = 0.5(m - S)$, where m should be interpreted as the lifetime of a new item that remains in the stockpile until it expires. We would expect that many real cases can be expressed in the form $L(S) = b(m - S)$, where the slope term b indicates the relative rate of aging of items in the stockpile and the field. In virtually all of the real-world applications, we can envision $0 \leq b \leq 1$, implying that items age more quickly in the field than in the stockpile.

As noted earlier, Derman and Klein defined the problem more narrowly than Greenwood. This allowed them to develop rigorous proofs of their results for more general forms of $L(S)$. Their article was significant in several respects. One was their problem definition. More importantly, they discovered what would become the standard approach for analyzing the issuing problem, namely, an induction argument on n , the number of items in the stockpile. Derman and Klein's main result is: if (i) $L(S)$ is a positive convex function and (ii) LIFO is optimal when $n = 2$, then LIFO is optimal when $n \geq 3$.

Zehna [20] points out that one must qualify the assumptions of this theorem. It is

untrue, for example, if $L(S)$ is an increasing function, or is not monotone. As noted earlier, this is unlikely to occur in the real-world, as it implies that items improve in the field. He also points out that the assumption made by Derman and Klein that issuing a single item is never optimal may, in fact, not be true in all circumstances. Zehna provides a proof of the slightly modified theorem that does not require this assumption. It is:

Theorem 3. *If $L(S)$ is a convex, nonincreasing function and LIFO is optimal for $n = 2$, then LIFO is optimal for all $n \geq 2$.*

For this result to be useful, one still needs to determine those convex functions for which LIFO is optimal when $n = 2$. Two examples presented by Derman and Klein are (i) $L(S) = a/(b + S)$, for $a > 0, b \geq 0$ and (ii) $L(S) = ce^{-ks}$ ($c, k > 0$). (Note that the first case was misprinted in the original as $L(S) = (a/b + S)$.) In both cases, the field life function is monotonically decreasing, so they do correspond to perishable items. Both functions are also nonlinear.

The first extension of the Derman and Klein work was due to Lieberman [21], who showed that if $L'(S) \geq -1$ and LIFO is optimal for $n = 2$, then LIFO is optimal for $n > 2$; and, if FIFO is optimal for $n = 2$, then FIFO is optimal for $n > 2$.

Zehna also provided several extensions of Derman and Klein's and Lieberman's results. He considers cases, where $L'(S) < -1$ and shows that LIFO is optimal in both cases where L is convex or concave. However, this is an unlikely case, as it implies that items age faster in stockpile than in the field. Zehna was also the first to consider stochastic field life functions, but obtained few results for this case.

As noted earlier, Derman and Klein's original work sparked a series of articles. In the opinion of this writer, most of these are of limited practical interest because of Derman and Klein's earlier result that, if the form of the field life function is known, all inventory problems can be solved as assignment problems. One of these minor extensions was due to Bomberger [22], who extended Derman

and Klein's results based on properties of the inverse function $L^{-1}(S)$.

Eilon [23] focused on the problem of whether LIFO or FIFO is optimal for the case $n = 2$. This is of interest since most of the earlier work used the Derman and Klein induction argument, which requires one to assume that LIFO or FIFO is optimal for $n = 2$. One interesting idea was that if one expressed the field life function as a power series expansion, one can find conditions for the optimality of LIFO or FIFO for two items based on these expansions. In a later work, Eilon [24] focused on the problem of providing closed form expressions for the total field life of a stockpile under several circumstances. He considered cases of identical items, and nonidentical items. He also considered the case where items flow into the stockpile at a constant rate. The replenishments could be of identical or nonidentical items.

Pierskalla [25] made several important contributions to the issuing problem. He extended the earlier work to include multiple demands on the stockpile, inclusion of penalty costs for issuing, and S -shaped field life functions. Several researchers considered the extension to the case of random field lives, but these results required strong assumptions. These studies include Pierskalla [26], Nahmias [27], and Albright [28].

CONTINUOUS REVIEW

Continuous review means that the state of the inventory system is known at all times, as opposed to periodic review, where the state of the system is known only at discrete points in time labeled periods. Surprisingly, very little is known about optimal ordering policies for fixed life perishables under continuous review. The difficulty is that when an order lead time is present, aging is applied only to the units on hand, and not to the units on order. Furthermore, since units may be ordered at any point in time, there is no limit to the number of different orders that comprise the on-hand inventory at any point in time. Hence, the vector describing the on-hand inventory of each age level has an unlimited number of dimensions.

One way to circumvent this problem is to assume zero order lead time. It is the belief of the author that the simultaneous assumptions of continuous review and zero order lead time are unrealistic, and it is likely that these models have little practical value. Note that there is a big difference between assuming zero order lead time in periodic review systems and in continuous review systems. In the former case, it only means that the lead time is less than a review period, as orders are assumed to be placed at the beginning of the planning period, and are assumed to arrive at the end of the planning period. In the case of continuous review, it means that orders arrive instantaneously. For that reason, the policies obtained are not likely to be useful in a practical setting. However, they may provide a theoretical stepping stone to the more realistic problem with positive lead times.

The first to explore this approach was Weiss [29]. Weiss assumed that demands are generated by a stationary Poisson process with fixed rate λ . His main result is that, in the case of lost sales, the form of the optimal policy is to either never order, or to order to a fixed level S when the inventory level drops to zero. Notice that the issue of perishability never really comes up, as order cycles are completely independent. One simply waits until the inventory level drops to zero (whether through demand or outdating), and replenishes to a fixed level [much as one does in the simple economic order quantity (EOQ) model].

Weiss' results were generalized by Liu and Lian [30]. They assumed full backordering of demand and generalized from a Poisson demand process to a stationary renewal process. They showed that the form of the optimal policy is (s, S) with $s \leq -1$, and provide explicit formulas for computing the optimal policy parameters. Note that the form of the policy (namely, that one only orders when the system is backordered), depends heavily on the assumption that order lead times are zero. It is unlikely that one would use such a policy in the real-world, where order lead times are always positive. Hence, it would appear that these results do not provide much insight into the general continuous review perishable inventory problem.

It is worth noting that the primary driver of the continuous review heuristic (Q, R) inventory models that form the basis for most commercial inventory control systems is uncertainty of demand over the replenishment lead time. Were lead times reduced to zero, optimal order quantities would collapse to the simple EOQ formula and reorder levels would be zero. (See Nahmias [31], Chapter 5)

When there is a positive lead time for ordering, the form of an optimal order policy for perishables under continuous review is not known. Part of the problem is defining the state of the system. As noted above, the state could be an infinite-dimensional vector of all previous orders placed and their ages. Given the complexity of the continuous review problem, a reasonable starting point is to find the best policy from a prespecified class of policies.

There is a large stream of research that exploits the analogy between queues with impatient customers and the perishable inventory problem. Consider a queueing system in which customers are willing to wait no more than m units of time in the queue for service. If we identify the queue with the on-hand inventory, the arrival process with the replenishment process, and the service process with the demand process, the analogy is clear. However, queueing models tend to be descriptive rather than prescriptive, as are inventory models. It is not clear that they shed much light on optimal ordering policies. It is rare that replenishments occur according to a random process. One real-world situation where this assumption is reasonable is in blood bank management. Blood is replenished via donations that do occur randomly on a one-at-a-time basis. The blood bank might have some control over the rate of replenishment via appeals. Hence, a queueing model in which the customer arrival rate is a control variable could be an accurate description of a blood bank.

Nahmias [32] appears to be the first to consider such a model. He assumes a simple $M/M/1$ queue with fixed customer impatience and considers optimization of the arrival rate to balance shortage and outdating costs. Graves [33] was the first to notice that the virtual waiting time process

could be used to derive the process for the age of the oldest unit in inventory. Although he does not explicitly show how, this approach could be used to optimize the arrival rate under certain conditions.

There is a large stream of research on queueing-based models, such as that carried out by Kaspi and Perry [34] and extensions, which we will not review here. These appear to be essentially descriptive, and have not yet been exploited to obtain optimal order policies.

One for One ($S-1, S$) Policies

The only rigorous analysis of a continuous review perishable inventory system with positive order lead time of which we are aware is Schmidt and Nahmias [35] (and an extension by Perry and Posner [36]). They assumed a so-called $(S-1, S)$ policy in which the inventory position (stock on hand plus stock on order) is maintained at a fixed level S . Assume that demands occur one-at-a-time, which would be the case, for example, if demands were generated by a stationary Poisson process. For nonperishables, an $(S-1, S)$ policy places an order for one unit at each occurrence of a demand. With perishability, orders are placed at both the occurrence of demands and outdating. $(S-1, S)$ policies are optimal for very high value items, and are common in military resupply and repair systems. A common maintenance strategy for high value equipment is to perform maintenance when the equipment fails or at fixed intervals, whichever comes first. Since unplanned failures occur at random, these can be modeled as a stochastic process, and may be labeled the demand process. If an item reaches age m and has not failed (i.e., been demanded), it is taken out of the system for routine maintenance, and thus “outdates.” Hence, the extension of traditional $(S-1, S)$ policies to the case of fixed life perishable inventories can be useful for modeling maintenance systems.

As noted above, what makes the problem difficult is the interaction of perishability and the order lead time. Assume that demands, (i.e., failures), occur one-at-a-time completely at random according to a stationary Poisson process with rate λ . Costs are charged in the

usual way against proportional ordering at c per unit, holding at h per unit held per unit time, outdating at θ per unit, and lost sales at p per unit of unsatisfied demand. Assume a positive order lead time of τ .

The analysis proceeds by analyzing the multidimensional stochastic process $\xi(t) = (\xi_1(t), \xi_2(t), \dots, \xi_S(t))$ defined as the amount of time elapsed since the last S orders were placed. Given the stationary distribution of this process, one can derive the steady-state distribution of the on-hand inventory at a random point in time. From this, one can derive the steady-state distribution of the on-hand inventory at a random point in time, say $(P_0, P_1, \dots, P_S)$, and use that to obtain an explicit expression for the expected cost rate as a function of S .

Perry and Posner [36] developed the following extension of this model. Suppose that customers arriving to the system, when it is out of stock, are willing to wait a random amount of time Y for the next arrival of a unit. The rationale for this extension is that one can keep track of the times of orders, so one can determine the instance of the next arrival. If this is imminent, it makes sense that a customer would wait rather than leave the system unsatisfied. While this extension is interesting, and could be potentially useful in some circumstances, the authors do not provide any calculations to compare the results of their model to those of Schmidt and Nahmias [35]. Until that is done, there is no way to tell if their extension leads to policies that differ significantly from those computed assuming customers do not wait when the system is out of stock.

Optimal (Q, r) Policies with Positive Lead Time

Short of finding optimal policies, a common method in inventory control is to determine the best policy from a class of policies. This is the approach taken by Berk and Gurler [37]. It is well known that for many nonperishable continuous review systems, the optimal policy is a (Q, r) policy. That is, when the inventory position hits r , an order for size Q is placed. Because the form of the optimal policy is not known, it would seem natural to consider the best (Q, r) policy in the perishable inventory setting. In their

analysis, the authors treat separately those cases where there is only single outstanding order permitted, and those where multiple outstanding orders are permitted. They also assume throughout that demands are generated from a stationary Poisson process, thus allowing them to make use of the memoryless property of the exponential interarrival time distribution.

Because items can perish, a slight modification of the standard policy is required. In nonperishable continuous review systems, one is guaranteed that an order can be placed at the instant the inventory level (or inventory position when more than one outstanding order is allowed) hits the reorder point, r . However, if only a single order is in inventory, the entire order will perish at the same instant, thus dropping the inventory level immediately to zero. Hence the policy must be modified to: one orders Q units whenever the inventory level hits r or zero, whichever comes first.

This modification points out the inadequacy of this policy, however. Consider a blood bank storing a rare form of blood (e.g., AB negative). Suppose that the lead time for replenishment is 3 weeks owing to the difficulty of finding donors with this blood type. Also, suppose that there is a substantial amount of AB negative blood on hand due to outdate in 1 day. According to this policy, the blood bank would sit on their current supply until it outdates the next day, dropping their inventory to zero. During the 3-week lead time, there would be no AB negative blood available, precipitating a crisis situation. (Note that this problem does not arise in the model considered by Schmidt and Nahmias [35], since all orders are for size one only, so the on-hand inventory level can only drop by one unit at a time independent of whether the drop is due to outdating or filling demand.)

Clearly, perishability fundamentally changes the form of the optimal policy. The blood bank seeing that their supply of this rare blood type would soon expire would take steps to replenish their inventory well in advance of the outdating. Hence, this type of (Q, r) policy does not make sense in this context. Before one can consider effective heuristics for this problem, a better

understanding of the nature of the optimal policy is required.

Another Potential Approach

In order to make the continuous review problem tractable, it would be valuable if we were able to find a single dimensional surrogate variable that conveys the key information about the on-hand inventory. Note that simply considering the total number of units on-hand irrespective of the age distribution of these units is inadequate, as pointed out in the example above. We propose that one should base a replenishment rule based on the expected remaining lifetime of the on-hand inventory. In the queueing context, this corresponds to the expected time to clear the current queue irrespective of future arrivals. Unfortunately, the calculation of this surrogate measure appears to be quite difficult, but could ultimately lead to an effective control mechanism. (i.e., one would wait until the expected lifetime of the on-hand inventory dropped to a critical level, and then place an order for Q units.)

THE APPLICATION OF PERISHABLE INVENTORY THEORY

This introduction to perishable inventories has focused on the development of the mathematical models that constitute the underlying theory. To what extent have these (or other) mathematical models been used in the real-world? The answer does not appear to be very much.

Throughout this article, it was emphasized that food management is perhaps the most significant potential area of application. Grocery store sales alone account for billions of dollars daily. A modest \$50 weekly grocery bill per person translates into an annual expenditure of \$750 billion in US alone—and this excludes restaurant receipts. Clearly, the potential for savings in food management is enormous. Unfortunately, not a great deal seems to be known about how grocery chains manage their inventories. Efficient inventory management provides a powerful competitive edge in this industry, making the large chains very tight lipped about their procedures.

We also noted that much of the research in perishable inventory theory was motivated by blood bank management. There is substantial literature on the application of analytical methods to blood banking. A good summary can be found in Prastacos [38]. The blood banking problem is much more complex than the relatively simple structures considered here. There are eight major blood types, whose frequencies vary from 38% (O+) to 0.5% (AB-). Blood is collected in units (one pint per donor) at a collection site such as a regional blood center, hospital blood bank, or mobile unit. Blood requests are issued at the hospital blood bank level, where blood is typed, stored, and issued to meet transfusion requests. A unique aspect of blood banking is cross matching. Donor and receiver blood (and possibly organs) must be tested prior to transfusion for compatibility.

This process can take several days. Another important aspect of the problem is the difference between amounts requested and amounts used. Surgeons will typically build in safety stock when they request blood. That is, in order to insure that they do not run out during surgery, they will request more than they expect to use. This feature was not included in any of the mathematical models we reviewed. Hence, a key part of blood bank management is determining the conditional distribution of usage given request size. Several researchers have attacked this problem (refer to the discussion in Prastacos [38] for developments in this area). Hence, a mathematical model of blood banking would have to include these complex features.

Radioactive fuels and pharmaceuticals are two products that exhibit continuous exponential decay. Inventory management of radioactive drugs was considered by Emmons [39] and Nahmias and Levine [40]. Emmons considered the relationship between the nucleotide generator and the products that are derived from that generator (commonly referred to as *parent/daughter relationship*). Although the degree of radioactivity does follow exponential decay, the analysis employed by Emmons did not rely on any of the exponential decay models of inventory management subject to exponential decay that appeared in the literature. Nahmias

and Levine also considered inventory management of radioactive pharmaceuticals, but were interested in the cost/benefit trade-offs of two isotopes of iodine used in the treatment of certain types of tumors. Again, their analysis did not rely upon traditional inventory control models.

REFERENCES

1. Fries BE. Optimal order policy for a perishable commodity with fixed lifetime. *Oper Res* 1975;23:46–61.
2. Nahmias S. Optimal ordering for perishable inventory—II. *Oper Res* 1975a;23:735–749.
3. Nahmias S. A comparison of alternative approximations for ordering perishable inventory. *INFOR* 1975b;13:175–184.
4. Van Zyl GJ. Inventory control for perishable commodities [Unpublished doctoral dissertation], University of North Carolina, Chapel Hill, 1964.
5. Cohen M. Analysis of single critical order policies in perishable inventory theory. *Oper Res* 1976;24:726–741.
6. Nahmias S. Myopic approximations for the perishable inventory problem. *Manag Sci* 1976;22:1002–1008.
7. Chazan D, Gal S. A Markovian model for a perishable product inventory. *Manag Sci* 1977;23:512–521.
8. Brodheim E, Derman C, Prastacos G. On the evaluation of a class of inventory policies for perishable products such as blood. *Manag Sci* 1975;21:1320–1325.
9. Kaplan RS. A dynamic inventory model with stochastic leadtimes. *Manag Sci* 1969;16:491–507.
10. Nahmias S. On ordering perishable inventory when both the demand and lifetime are random. *Manag Sci* 1977;24:82–90.
11. Nahmias S. The fixed charge perishable inventory problem. *Oper Res* 1978;26:464–481.
12. Nahmias S, Pierskalla WP. A two-product perishable/nonperishable inventory problem. *SIAM J Appl Math* 1976;30:483–500.
13. Nahmias S. Perishable inventory theory: a review. *Oper Res* 1982a;30:680–708.
14. Deuermeyer B. A multi-type production system for perishable inventories. *Oper Res* 1979;27:935–943.
15. Deniz B, Karaesmen A, Scheller-Wolf I. Managing inventories of perishable goods:

- the effect of substitution. Unpublished manuscript, 2007, p. 34.
16. Cohen M, Pekelman D. LIFO inventory systems. *Manag Sci* 1978;24:1150–1162.
 17. Greenwood JA. Issue priority. *Nav Res Logist Q* 1955;2:251–268.
 18. Derman C, Klein M. Inventory depletion management. *Manag Sci* 1958;4:450–456.
 19. Veinott AF, Jr. Optimal ordering issuing and disposal of inventory with known demand [Unpublished doctoral dissertation], Columbia University, New York, 1960.
 20. Zehna P. Inventory depletion policies. In: Arrow KJ, Karlin S, Scarf HE, editors. *Studies in applied probability and management science*. Stanford (CA): Stanford University Press; 1962. pp. 263–287.
 21. Lieberman GJ. LIFO vs. FIFO in inventory depletion management. *Manag Sci* 1958;5:102–105.
 22. Bomberger EE. Optimal inventory depletion policies. *Manag Sci* 1961;7:294–303.
 23. Eilon S. FIFO and LIFO policies in inventory management. *Manag Sci* 1961;7:304–315.
 24. Eilon S. Obsolescence of commodities which are subject to deterioration in store. *Manag Sci* 1963;9:623–642.
 25. Pierskalla W. Optimal issuing policies in inventory management. *Manag Sci* 1967a;13:395–412.
 26. Pierskalla W. Inventory depletion management with stochastic field life functions. *Manag Sci* 1967b;13:877–886.
 27. Nahmias S. Inventory depletion management when the field life is random. *Manag Sci* 1974;20:1276–1283.
 28. Albright SC. Optimal stock depletion policies with stochastic field lives. *Manag Sci* 1976;22:852–857.
 29. Weiss HJ. Optimal ordering policies for continuous review perishable inventory models. *Oper Res* 1980;28:365–374.
 30. Liu L, Lian Z. (s, S) Continuous review models for products with fixed lifetimes. *Oper Res* 1999;47:150–158.
 31. Nahmias S. *Production and operations analysis*. 6th ed. Burr Ridge (IL): McGraw Hill/Irwin; 2009. p. 789.
 32. Nahmias S. Queueing models for controlling perishable inventories. In: Chikan A, editor. *The economics of management of inventories. Proceedings of the First International Conference on Inventories*. Amsterdam: Elsevier; 1982b.
 33. Graves SC. The application of queueing theory to continuous perishable inventory systems. *Manag Sci* 1982;28:400–406.
 34. Kaspi H. and Perry D. Inventory Systems of Perishable Commodities. *Adv. in App. Prob.* 1983;15:674–685.
 35. Schmidt CP, Nahmias S. $(S-1, S)$ Policies for perishable inventory. *Manag Sci* 1985;31:719–728.
 36. Perry D, Posner MJM. An $(S-1, S)$ inventory system with fixed shelf life and constant lead times. *Oper Res* 1998;46 (Suppl. #3):S65–S83.
 37. Berk E, Gurler U. Exact analysis of the (Q, r) inventory model for perishables with positive lead times and lost sales. Unpublished manuscript, 2004, p. 40.
 38. Prastacos GP. Blood inventory management: an overview of theory and practice. *Manag Sci* 1984;30:777–787.
 39. Emmons H. A replenishment model for radioactive nuclide generators. *Manag Sci* 1968;14:263–274.
 40. Nahmias S, Levine G. Inventory management and cost benefit analysis: a case study with G. Levine. *Omega* 1979;7:321–332.

MATHEMATICAL PROGRAMMING APPROACHES TO THE TRAVELING SALESMAN PROBLEM

ADAM N. LETCHFORD
Department of Management
Science, Lancaster University,
Lancaster, UK

ANDREA LODI
DEIS, Università di Bologna,
Bologna, Italy

The *Traveling salesman problem* (TSP) is probably the best known combinatorial optimization problem. Informally speaking, one is given a set of cities, along with the cost of traveling between each pair of cities, and one wishes to find a tour of all the cities of minimum cost. The TSP has applications, not only in OR/MS, but in many other fields. In fact, there are no fewer than four books devoted to it [1–4].

The TSP is notoriously hard to solve, being one of Karp's original $\mathcal{NP}$ -complete problems [5]. Nevertheless, and surprisingly, large-scale instances arising in practice can be often solved to proven optimality (or near-optimality) by sophisticated mathematical programming techniques. This article gives an introduction to these techniques.

The article is structured as follows. We begin by reviewing some of the basic mathematical programming formulations of the TSP. We then discuss polyhedral theory, which is used to derive stronger formulations. Next, we discuss separation routines, which are used to strengthen formulations iteratively. Finally, we discuss exact algorithms for the TSP based on such strengthened formulations.

We remark that some other effective techniques for tackling the TSP exist. For an introduction to *combinatorial* TSP algorithms (i.e., algorithms that work with graph-theoretic relaxations rather than linear programming relaxations), see the article titled **Combinatorial Traveling Salesman Problem Algorithms** in this

encyclopedia. For heuristic and metaheuristic approaches (which give good, but not necessarily optimal solutions), see the article titled **Heuristics for the Traveling Salesman Problem** in this encyclopedia.

Throughout this article, we distinguish between the *symmetric* TSP (STSP), in which the cost of traveling from city A to city B is the same as the cost of traveling in the reverse direction, and the *asymmetric* TSP (ATSP), in which these costs are permitted to be different.

FORMULATIONS

Combinatorial optimization problems can often be formulated in several different ways—see the article titled **Formulating Good MILP Models** in this encyclopedia. In the case of the STSP, the standard formulation is the following, due to Dantzig *et al.* [6]. We are given a complete undirected graph G , with node set V and edge set E . For each $e \in E$, we let c_e denote the cost of traversing edge e . For any node set $S \subset V$, we let $\delta(S)$ denote the set of edges with exactly one end-node in S . Next, for each edge $e \in E$, we define a binary variable x_e , taking the value 1 if and only if edge e is in the tour. Finally, for any set of edges $F \subset E$, we let $x(F)$ denote $\sum_{e \in F} x_e$. The STSP is then equivalent to the following zero–one linear program (0–1 LP):

$$\begin{aligned} \min \quad & \sum_{e \in E} c_e x_e \\ \text{s.t.} \quad & x(\delta(\{i\})) = 2 \quad (\forall i \in V) \end{aligned} \quad (1)$$

$$x(\delta(S)) \geq 2 \quad \left(\forall S \subset V: 2 \leq |S| \leq \frac{|V|}{2} \right) \quad (2)$$

$$x \in \{0, 1\}^{|E|}. \quad (3)$$

The equations (1), called *degree equations*, simply express the fact that the salesman must arrive at and depart from each city. The inequalities (2), called *subtour elimination constraints* (SECs), ensure that the tour is connected.

The ATSP can be formulated in a similar way. Now we have a complete directed graph G , with node set V and arc set A . For each $a \in A$, we let c_a denote the cost of traversing arc a . For any set $S \subset V$, we let $\delta^+(S)$ denote the set of arcs going from S to $V \setminus S$, and $\delta^-(S)$ denote the set of arcs going in the reverse direction. For each arc $a \in A$, we have a binary variable x_a . The ATSP is then equivalent to the following 0–1 LP:

$$\min \sum_{a \in A} c_a x_a$$

$$\text{s.t. } x(\delta^+(\{i\})) = 1 \quad (\forall i \in V) \quad (4)$$

$$x(\delta^-(\{i\})) = 1 \quad (\forall i \in V) \quad (5)$$

$$x(\delta^+(S)) \geq 1 \quad \left(\forall S \subset V: 2 \leq |S| \leq \frac{|V|}{2} \right) \quad (6)$$

$$x \in \{0, 1\}^{|A|}. \quad (7)$$

The equations (4) and (5) are called *out-degree* and *in-degree* equations, respectively. The inequalities (6) are again called *SECs*.

Note that, in both formulations, the SECs are exponential in number. It is in fact possible to formulate the STSP and ATSP as 0–1 LPs in which the number of constraints is polynomial, provided that one is willing to introduce a polynomial number of additional variables. An extensive survey of such alternative formulations was recently given by Öncan *et al.* [7]. Nevertheless, the formulations given here have been found to be very useful in practice, and are therefore the most commonly used. The large number of SECs is not a serious problem, since one can generate the SECs dynamically as cutting planes (see **Branch and Cut**).

POLYHEDRAL THEORY

In this section, we review some of the known polyhedral results for the TSP. We cover the symmetric case first, and then the asymmetric case. We assume that the reader is already familiar with the polyhedral approach to combinatorial optimization problems, which is covered in the article titled **Basic Polyhedral Theory** in this encyclopedia.

The Symmetric Traveling Salesman Polytope

The convex hull in $\mathcal{R}^{|E|}$ of solutions to the system (1)–(3), for a given number n of nodes, is called the *symmetric traveling salesman polytope of order n* , and is denoted by $\text{STSP}(n)$. The first systematic study of $\text{STSP}(n)$ was performed by Grötschel and Padberg [8], who proved the following results:

- The degree equation (1) give a complete and nonredundant description of the affine hull of $\text{STSP}(n)$.
- The nonnegativity inequalities $x_e \geq 0$, for all $e \in E$, define facets of $\text{STSP}(n)$ for $n \geq 4$.
- The SECs (Eq. 2) define facets of $\text{STSP}(n)$ for $n \geq 4$.
- Let H be a subset of V with $3 \leq |H| \leq n - 2$ and let $F \subset \delta(H)$ be a set of node-disjoint edges with $|F| \geq 3$ and odd. The *blossom* inequality $x(\delta(H) \setminus F) \geq x(F) - |F| + 1$ defines a facet of $\text{STSP}(n)$.

Grötschel and Padberg [8] also proved a “lifting” theorem that enables one to convert simple facets into more complex facets. When lifting is applied to the blossom inequalities, one obtains the following *comb* inequalities:

$$x(\delta(H)) + \sum_{j=1}^t x(\delta(T_j)) \geq 3t + 1.$$

Here, $t \geq 3$ is an odd integer, H is a proper subset of V (called the *handle*), and $T_1, \dots, T_t$ are disjoint node-sets (called the *teeth*), such that $T_j \cap H \neq \emptyset$ and $T_j \setminus H \neq \emptyset$ for $j = 1, \dots, t$.

The comb inequalities are undoubtedly a very important class of inequalities for the STSP. They are used in all branch-and-cut algorithms of which we are aware.

Since the publication of Ref. 8, the polytope $\text{STSP}(n)$ has been studied in great depth. Several generalizations of the comb inequalities are known, with exotic names like *clique-tree*, *star*, *hyperstar*, *bipartition*, and *binested* inequalities. There are also several other classes of facet-inducing inequalities known, such as *chain*, *ladder*, *crown*, and *hypohamiltonian* inequalities. For the sake of space, we do not review these developments here, and

refer the reader to the excellent surveys by Grötschel and Padberg [9], Jünger *et al.* [10], and Naddef [11]. We will, however, mention some of the inequalities again, in subsequent sections.

The Asymmetric Traveling Salesman Polytope

The convex hull in $\mathcal{R}^{|A|}$ of solutions to the system (4)–(7), again for a given n , is called the *asymmetric traveling salesman polytope of order n* , and is denoted by $\text{ATSP}(n)$. A survey of early work on $\text{ATSP}(n)$ was given by Grötschel and Padberg [9]. Some of the key early results are as follows:

- The out- and in-degree equations (4)–(5) define the affine hull of $\text{ATSP}(n)$, and all but one of them (arbitrarily chosen) are linearly independent.
- The nonnegativity inequalities $x_a \geq 0$, for all $a \in A$, define facets of $\text{ATSP}(n)$ for $n \geq 4$.
- Any valid inequality for $\text{STSP}(n)$ can be converted into a valid inequality for $\text{ATSP}(n)$, by simply replacing x_e with $x_{ij} + x_{ji}$ for all edges $e = (i, j) \in E$. The resulting inequalities are called *symmetric*.
- The SECs (Eq. 6), which are symmetric, define facets of $\text{ATSP}(n)$ for $n \geq 4$.

Grötschel and Padberg [9] also list various *asymmetric* inequalities. Among them, we mention the so-called D_k^+ and D_k^- inequalities, which have proven to be fairly useful in practice. The D_k^+ inequalities take the form:

$$x_{i_1 i_k} + \sum_{h=2}^k x_{i_h i_{h-1}} + 2 \sum_{h=2}^{k-1} x_{i_1 i_h} + \sum_{h=3}^{k-1} x(\{i_2, \dots, i_{h-1}\}, i_h) \leq k-1, \quad (8)$$

where $(i_1, \dots, i_k)$ is any sequence of $k \in \{3, \dots, n-1\}$ distinct nodes. The D_k^- inequalities are derived from the D_k^+ inequalities by simply swapping the left-hand side coefficients of the two arcs (i, j) and (j, i) , for all $i, j \in V$, $i < j$. This is a perfectly general operation, called *transposition* in Ref. 9.

Another important class of inequalities, which turns out to be very useful in practice, was introduced by Balas [12]. Two distinct arcs (i, j) and (u, v) are called *incompatible* if $i = u$, or $j = v$, or $i = v$ and $j = u$; *compatible* otherwise. A *closed alternating trail* (CAT) is a sequence $T = \{a_1, \dots, a_t\}$ of t distinct arcs such that, for $k = 1, \dots, t$, arc a_k is incompatible with arcs a_{k-1} and a_{k+1} , and compatible with all other arcs in T (with $a_0 := a_t$ and $a_{t+1} := a_1$). Now, call a node v a *source* if $|\delta^+(v) \cap T| = 2$, and a *sink* if $|\delta^-(v) \cap T| = 2$. Note that a node can be a source and a sink simultaneously. Let Q be the set of arcs $(i, j) \in A \setminus T$ such that i is a source and j is a sink. Given a CAT T with t odd, the *odd CAT inequality* $x(T \cup Q) \leq \lfloor t/2 \rfloor$ is valid for $\text{ATSP}(n)$. Balas [12] proved that odd CAT inequalities define facets of $\text{ATSP}(n)$ under mild conditions.

There exist several other articles on $\text{ATSP}(n)$. Again, for the sake of brevity, we do not review them all here, and refer the reader to the excellent survey [13].

SEPARATION ROUTINES

In order to use a class of valid inequalities as cutting planes, one needs a *separation routine*, that is, a routine for detecting when an inequality in that class is violated by a given LP solution x^* (see **Branch and Cut**). As usual, we distinguish between *exact* separation routines, which find a violated inequality in that class whenever one exists, and *heuristic* separation routines, which may fail to find a violated inequality.

Separation Routines for the STSP

Several positive results are known concerning the complexity of exact separation for various inequalities:

- As noted by Hong [14], SEC separation amounts to a minimum weight cut problem, and is therefore solvable in polynomial time. (The current fastest deterministic minimum weight cut algorithm is that of Nagamochi *et al.* [15].)
- The separation problem for the blossom inequalities can be solved in polynomial

time (Padberg and Rao [16]). The current fastest algorithm is that of Letchford *et al.* [17].

- There exists a polynomial-time separation algorithm for a class of inequalities that includes all blossom inequalities and a large subclass of the comb inequalities (Fleischer *et al.* [18]).
- There is a polynomial-time algorithm that generates a violated valid inequality whenever a comb inequality is violated by 1 (Caprara *et al.* [19]).
- There exists a polynomial-time separation algorithm for a class of inequalities that includes all comb inequalities with a fixed handle H (Caprara and Letchford [20]).
- The separation problem for bipartition inequalities (and therefore comb and clique-tree inequalities) can be solved in polynomial time, provided the number of teeth is bounded by a constant (Carr [21]).
- More generally, there exists a polynomial-time separation algorithm for any class of inequalities that can be obtained by “lifting” an inequality that defines a facet of STSP(n), for constant n (Carr [22]).

Further positive results are known for the case in which the *support graph* is planar. (The support graph is the graph induced by the edges with positive x^* -value.) Fleischer and Tardos [23] found an $\mathcal{O}(n^2 \log n)$ algorithm that produces a violated comb inequality whenever a comb inequality is violated by 1, and Letchford [24] gave an $\mathcal{O}(n^3)$ exact separation algorithm for a generalization of the comb inequalities, called *domino-parity* inequalities. Further results for the planar case can be found in Ref. 25.

In addition to these exact separation results, there exists a wide range of separation heuristics. For example:

- Crowder and Padberg [26] presented a fast heuristic for SECs, that is based on the idea of iteratively “shrinking” the support graph by identifying certain pairs of nodes.
- Grötschel and Holland [27] proposed a heuristic for comb separation, in which one first shrinks the support graph, and then calls an exact blossom separation algorithm on the shrunk support graph.
- Padberg and Rinaldi [28] presented several separation heuristics for SECs, blossom, comb, and clique-tree inequalities, along with some new conditions under which shrinking can be safely performed.
- Naddef and Thienel [29] presented new heuristics for comb, clique-tree, path, bipartition, and ladder inequalities. They are based on shrinking and local search.
- Applegate *et al.* [1] analysed in depth certain data structures that can be used to speed up various SEC, blossom, and comb heuristics.
- Boyd *et al.* [30] and Cook *et al.* [31] presented heuristics for domino-parity inequalities, based on shrinking the support graph to make it planar, and then running the algorithm of Letchford [24].

Finally, we mention two “nonstandard” heuristics for separation: the “small instance” approach of Christof and Reinelt [32] and the “local cuts” approach of Applegate *et al.* [1]. Both of them begin by shrinking the support graph down to a tiny size. The small instance approach then searches through a list of inequality “templates,” whereas the local cuts approach solves a series of tiny STSP instances to generate a cutting plane.

Separation Routines for the ATSP

In the section titled “The Asymmetric Traveling Salesman Polytope,” we mentioned that every valid inequality for STSP(n) can be converted into a valid inequality for ATSP(n), simply by replacing x_e with $x_{ij} + x_{ji}$ for all edges $e = (i, j) \in E$. This implies that every separation algorithm for the STSP can be used, as a “black box,” for the ATSP as well. To this end, given the ATSP (fractional) point $x^* \in [0, 1]^{|A|}$, one first defines the undirected counterpart $\tilde{x} \in [0, 1]^{|E|}$ by means of the transformation

$$\tilde{x}_e := x_{ij}^* + x_{ji}^* \quad \text{for all edges } e = (i, j) \in E,$$

and then applies the STSP separation algorithm to $\tilde{x}$. On return, the most violated STSP inequality detected is transformed into its ATSP counterpart, both inequalities having the same degree of violation.

The above approach to separation can be used only for *symmetric* ATSP inequalities. The only article of which we are aware that addresses separation for *asymmetric* ATSP inequalities is Fischetti and Toth [33]. These authors presented an exact separation algorithm for D_k^+ and D_k^- inequalities, based on partial enumeration of suitable node sequences $(i_1, \dots, i_k)$, for $k = 2, \dots, n - 1$. To prune the search, upper bounds on the violation of the inequalities are used. The authors also presented a heuristic separation algorithm for odd CAT inequalities, based on the computation of minimum weight odd circuits in an auxiliary “incompatibility graph.” This separation algorithm can be viewed as a specialized version of a scheme proposed by Caprara and Fischetti [34] for the separation of a subclass of Chvátal-Gomory cuts for general integer programming problems.

EXACT TSP ALGORITHMS

In this section, we review the main exact algorithms for the TSP that are based on the use of strong valid inequalities as cutting planes. We assume that the reader is already familiar with the basic ideas of branch-and-bound and branch-and-cut, which are covered in the articles titled **Branch-and-Bound Algorithms** and **Branch and Cut**, respectively in this encyclopedia.

Exact Algorithms for the STSP

The first article to propose the use of linear programming to solve the STSP was Dantzig *et al.* [6] in 1954. The authors computed an optimal tour through 49 cities in the United States, an impressive achievement at the time. The cutting planes, mainly SECs, were identified by eye. Moreover, the authors used a limited form of branching, even though branch-and-bound had not been invented at the time.

In the 1970s and 1980s, several articles appeared in which larger TSP instances were solved by the same approach, but with the addition of other kinds of cutting planes, such as blossom and comb inequalities. A good example is Grötschel [35], which details the solution of a 120-city problem.

The next step was the development of what is now called the *cut-and-branch* approach, in which a cutting plane algorithm is used to create a strong LP relaxation, which is then fed into a branch-and-bound solver. An early example can be found in Crowder and Padberg [26]. This approach was pushed to its limits by Grötschel and Holland [27], who solved instances with up to around 600 cities.

A disadvantage of the cut-and-branch approach is that the integer solution obtained might violate some SECs, in which case they must be added to the formulation, and branch-and-bound must be repeated. The natural solution to this is to run separation algorithms at every node of the branch-and-bound tree. This is the *branch-and-cut* approach of Padberg and Rinaldi [36]. Using this approach, Padberg and Rinaldi solved instances with up to around 2000 cities, using SECs and blossom, comb, and clique-tree inequalities.

The branch-and-cut approach has been further refined, for example, by Naddef and Thienel [29], Applegate *et al.* [1], and Cook *et al.* [31]. The software of Applegate *et al.*, called Concorde, has been made freely available on the web [37]. It is capable of solving instances with tens of thousands of cities to proven optimality. According to Cook *et al.* [31], the domino-parity inequalities and the local cuts are the most effective cutting planes for large instances.

Exact Algorithms for the ATSP

Only two articles concerned with branch-and-cut algorithms for the ATSP have been published: Fischetti and Toth [33] and Fischetti *et al.* [38].

The algorithm of Fischetti and Toth [33] is based closely on the Padberg and Rinaldi [36] framework. There are, however, some important details:

1. A sophisticated scheme is used for pricing (column generation), which exploits the fact that a directed tour cannot select arcs with negative reduced cost in an arbitrary way.
2. Asymmetric inequalities— D_k^+ , D_k^- and odd CAT inequalities—are used in addition to symmetric ones.
3. Rounds of violated inequalities are added when possible, rather than only the most violated one.
4. A primal heuristic, based on detecting Hamiltonian cycles in the (directed) support graph, is extensively used.

The resulting algorithm solved to optimality most of the ATSP instances in the literature.

Fischetti *et al.* [38] proposed some enhancements to the Fischetti–Toth algorithm. In particular, the branching variable in Ref. 33 was selected among those closest to 0.5 (a classical criterion) while in Ref. 38, priority for branching was given to the variables that have been persistently fractional in the last (consecutive) LP optimal solutions at any branching node.

Fischetti *et al.* [38] also tested a rather different approach, in which the ATSP instance is transformed to an STSP instance, which is then solved by the STSP solver Concorde mentioned in the previous section. They tested two different transformations: the one due to Karp [5], in which an ATSP instance with n nodes is converted to an STSP instance with $3n$ nodes, and the one due to Jonker and Volgenant [39], in which the resulting STSP instance has only $2n$ nodes. The computational results in Ref. 38 show that this alternative approach is competitive with (although not superior to) the approach in Ref. 33.

We close this discussion by pointing out an important difference between the STSP and the ATSP. In the case of the STSP, branch-and-cut is currently the undisputed best approach for solving large-scale instances. This is not so with the ATSP: for some instances, such as randomly generated instances with no correlation between c_{ij} and c_{ji} , combinatorial algorithms based on *assignment relaxations* can perform as

well or even better. This is discussed in more detail in the article titled **Combinatorial Traveling Salesman Problem Algorithms** in this encyclopedia.

REFERENCES

1. Applegate D, Bixby RE, Chvátal V, *et al.* The traveling salesman problem: a computational study. Princeton (NJ): Princeton University Press; 2007.
2. Gutin G, Punnen AP, editors. The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002.
3. Lawler EL, Lenstra JK, Rinnooy Kan AHG, *et al.*, editors. The Traveling Salesman Problem. Chichester: Wiley; 1985.
4. Reinelt G. The traveling salesman: computational solutions for TSP applications. Berlin: Springer; 1994.
5. Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, editors. Complexity of computer computations. New York: Plenum; 1972. pp. 85–103.
6. Dantzig GB, Fulkerson DR, Johnson SM. Solution of a large-scale traveling salesman problem. *Oper Res* 1954;2:393–410.
7. Öncan T, Altinel IK, Laporte G. A comparative analysis of several asymmetric traveling salesman problem formulations. *Comput Oper Res* 2009;36:637–654.
8. Grötschel M, Padberg MW. On the symmetric travelling salesman problem (parts I and II). *Math Program* 1979;16:265–302.
9. Grötschel M, Padberg MW. Polyhedral theory. In: Lawler EL, *et al.*, editors. The Traveling Salesman Problem. Chichester: Wiley; 1985. pp. 251–305.
10. Jünger M, Reinelt G, Rinaldi G. The traveling salesman problem. In: Ball M, Magnanti T, Monma C, editors. Network models. Amsterdam: Elsevier; 1995. pp. 225–330.
11. Naddef D. Polyhedral theory and branch-and-cut algorithms for the TSP. In: Gutin G, Punnen A, editors. The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002. pp. 29–114.
12. Balas E. The asymmetric assignment problem and some new facets of the traveling salesman polytope on a directed graph. *SIAM J Discr Math* 1989;2:425–451.
13. Balas E, Fischetti M. Polyhedral theory for the asymmetric traveling salesman problem. In: Gutin G, Punnen A, editors.

- The traveling salesman problem and its variations. Dordrecht: Kluwer; 2002. pp. 117–168.
14. Hong S. A linear programming approach for the traveling salesman problem [PhD thesis]. Baltimore (MD): Johns Hopkins University; 1972.
15. Nagamochi H, Ono T, Ibaraki T. Implementing an efficient minimum capacity cut algorithm. *Math Program* 1994;67:325–341.
16. Padberg MW, Rao MR. Odd minimum cut-sets and b -matchings. *Math Oper Res* 1982;7: 67–80.
17. Letchford AN, Reinelt G, Theis DO. Odd minimum cut-sets and b -matchings revisited. *SIAM J Discr Math* 2008;22:1480–1487.
18. Fleischer LK, Letchford AN, Lodi A. Polynomial-time separation of a superclass of simple comb inequalities. *Math Oper Res* 2006;31:696–713.
19. Caprara A, Fischetti M, Letchford AN. On the separation of maximally violated mod- k cuts. *Math Program* 2000;87:37–56.
20. Caprara A, Letchford AN. On the separation of split cuts and related inequalities. *Math Program* 2003;94:279–294.
21. Carr RD. Separating clique tree and bipartition inequalities having a fixed number of handles and teeth in polynomial time. *Math Oper Res* 1997;22:257–265.
22. Carr RD. Separation algorithms for classes of STSP inequalities arising from a new STSP relaxation. *Math Oper Res* 2004;29:80–91.
23. Fleischer LK, Tardos E. Separating maximally violated comb inequalities in planar graphs. *Math Oper Res* 1999;24:130–148.
24. Letchford AN. Separating a superclass of comb inequalities in planar graphs. *Math Oper Res* 2000;25:443–454.
25. Letchford AN, Pearson NA. Exploiting planarity in separation routines for the symmetric traveling salesman problem. *Discr Opt* 2008;5:220–230.
26. Crowder H, Padberg MW. Solving large-scale symmetric traveling salesman problems to optimality. *Mgmt Sci* 1980;26:495–509.
27. Grötschel M, Holland O. Solution of large-scale symmetric traveling salesman problems. *Math Program* 1991;51:141–202.
28. Padberg MW, Rinaldi G. Facet identification for the symmetric traveling salesman polytope. *Math Program* 1990;47:219–257.
29. Naddef D, Thienel S. Efficient separation routines for the symmetric traveling salesman problem (parts I and II). *Math Program* 2002;92:237–283.
30. Boyd SC, Cockburn S, Vela D. On the domino-parity inequalities for the TSP. *Math Program* 2007;110:501–519.
31. Cook W, Espinoza D, Goycoolea M. Computing with domino-parity inequalities for the TSP. *INFORMS J Comput* 2007;19:356–365.
32. Christof T, Reinelt G. Combinatorial optimization and small polytopes. *TOP* 1996;4:1–53.
33. Fischetti M, Toth P. A polyhedral approach to the asymmetric traveling salesman problem. *Manag Sci* 1997;43:1520–1536.
34. Caprara A, Fischetti M. $\{0, \frac{1}{2}\}$ -Chvátal-Gomory cuts. *Math Program* 1996;74: 221–236.
35. Grötschel M. On the symmetric travelling salesman problem: solution of a 120-city problem. *Math Program Study* 1980;12:61–77.
36. Padberg MW, Rinaldi G. A branch-and-cut algorithm for the resolution of large-scale symmetric travelling salesman problems. *SIAM Rev* 1991;33:60–100.
37. Concorde TSP Solver. Available at <http://www.tsp.gatech.edu/concorde.html>.
38. Fischetti M, Lodi A, Toth P. Exact methods for the asymmetric traveling salesman problem. In: Gutin G, Punnen A, editors. *The traveling salesman problem and its variations*. Dordrecht: Kluwer; 2002. pp. 169–205.
39. Jonker R, Volgenant T. Transforming asymmetric into symmetric traveling salesman problems. *Oper Res Lett* 1983;2:161–163.
40. Reinelt G. TSPLIB – a traveling salesman problem library. *ORSA J Comput* 1991;3: 376–384.

MATRIX ANALYTIC METHOD: OVERVIEW AND HISTORY

ATTAHIRU S. ALFA
Department of Electrical &
Computer Engineering,
University of Manitoba,
Winnipeg, Canada

VAIDYANATHAN RAMASWAMI
AT&T Labs Research, Florham
Park, New Jersey

In applied probability, “matrix analytic methods” is the name given to a body of unified algorithmic techniques exploiting probabilistic structure and based on nonlinear matrix iterations for analyzing large classes of structured Markov chains and Markov renewal processes. Although there have been sporadic uses of matrix-based methods in the literature [1–4], the credit for initiating a systematic theory with wide scope and leading to reliable algorithms must go to Marcel F. Neuts whose two papers [5,6], which respectively analyzed two major classes of models known as the $M/G/1$ and the $GI/M/1$ types, marked the clear beginnings of that theory. The approach has since been embellished with many generalizations, faster and better algorithms, and more importantly many credible applications to practical problems. This is a brief overview of the subject with a historical perspective concentrating primarily on major milestones along its growth. For a detailed treatment, we refer to the following texts: Neuts [7,8]; Latouche and Ramaswami [9]; and Bini *et al.* [10]. The beginner may find Latouche and Ramaswami [9] to be the most accessible. Two useful tutorials on the topic are Ramaswami [11] and Alfa [12].

THE BASIC STRUCTURES

Matrix analytic methods deal with two-dimensional, countable state space Markov renewal processes of which the first coordinate has a skip-free structure in at least one direction. Denoting a general state by (i, j) ,

$i \in N = \{0, 1, \dots\}$ and $1 \leq j \leq m < \infty$ and grouping states into “levels” $\mathbf{i} = \{(i, 1), \dots, (i, m)\}$, there are three structures that are covered:

- The **M/G/1** type: Here any one-step transition from a level $\mathbf{i}$ can occur only into levels $\mathbf{i} - \mathbf{1}$ and above; in other words, the model is skip-free downward in levels.
- The **GI/M/1** type: Here any one-step transition from a level $\mathbf{i}$ is restricted to levels up to $\mathbf{i} + \mathbf{1}$; in other words, the model is skip-free upward in levels.
- Quasi-Birth-and-Death (QBD): Here one-step transitions from $\mathbf{i}$ are restricted to levels $\mathbf{i} - \mathbf{1}, \mathbf{i}, \mathbf{i} + \mathbf{1}$; clearly, such a process is skip-free in levels in both directions.

In each of these cases, it is further assumed that the transition structure has a spatial homogeneity; specifically, except for the boundary level $\mathbf{0}$, the transition probabilities from level $\mathbf{i}$ to $\mathbf{k}$ depend on the values of i and k only through the difference $i - k$. Thus, in the case of Markov chains, the three models above correspond respectively to block partitioned transition matrices of the forms given below.

$$\begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \dots \\ A_0 & A_1 & A_2 & A_3 & \dots \\ 0 & A_0 & A_1 & A_2 & \dots \\ 0 & 0 & A_0 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\begin{bmatrix} B_0 & A_0 & 0 & 0 & \dots \\ B_1 & A_1 & A_0 & 0 & \dots \\ B_2 & A_2 & A_1 & A_0 & \dots \\ B_3 & A_3 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\begin{bmatrix} B_0 & A_0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & \dots \\ 0 & 0 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The names come from their scalar versions, which correspond respectively to the Markov chains in the ordinary $M/G/1$, $GI/M/1$, and birth–death models. In the Markov renewal case, the transition kernels inherit the above structures with convolutions replacing matrix multiplication. Also, continuous time models have an analogous treatment [9].

Some comments are immediately in order:

- (i) The coordinate j itself could be a vector of states; thus, the process may actually be multidimensional; (ii) extensions could be made trivially for processes like multiserver queues wherein the underlying structure holds only after a finite number of levels [9]; (iii) the restriction of the second coordinate to a finite set was necessitated only by early methodology based on Perron Frobenius results for finite dimensional matrices and is not really needed.

THE $M/G/1$ TYPE

We assume all models under consideration to be irreducible. It is well known [13] for the ordinary $M/G/1$ queue that the key quantity to be analyzed is the busy period. Denoting by g_k the probability that starting in 1 the Markov chain makes its first visit to 0 at step k , a simple conditioning argument shows that $g^*(z) = \sum z^k g_k$ must satisfy the equation $g^*(z) = \sum_{n=0}^{\infty} z a_n [g^*(z)]^n$, and furthermore that in the ergodic case the eventual probability of visiting 0 which is $g = g^*(1)$ is 1 and the mean time to visit 0 which is $g^{*(1)}$ is finite. The classical approach, however, treated the equation as $g^*(z) = z a^*(g^*(z))$, where $a^*(\cdot)$ is the generating function of $\{a_i\}$ and proceeded based on complex analysis via Rouché’s theorem. Once $g^*(z)$ is determined, one can immediately determine the return time distribution to level 0 to be given by $\sum_{n=0}^{\infty} z b_n [g^*(z)]^n$. These enabled a detailed analysis in terms of $g^*(z)$. In the ergodic case one can also get a simple recursion due to P.J. Burke for steady state probabilities by rearranging the steady state equations slightly ([8], pp. 16–17).

The approach in Neuts’ articles [5,14] can be considered to be a direct generalization of these ideas. For the first passage time from

level 1 to level 0, the natural quantity to be considered is the conditional probability $g_{ij}(k)$ that, given the initial state is $(1,i)$, the first visit to 0 occurs at step k through a visit to the specific state $(0,j)$; the associated matrix generating function $G^*(z) = \sum_{k=1}^{\infty} z^k G(k)$, where $G(k)$ is the $m \times m$ matrix of elements $g_{ij}(k)$. With $G = G^*(1)$ yielding the probabilities of eventual entry into 0, we get the equations

$$G^*(z) = z \sum_{n=0}^{\infty} A_n [G^*(z)]^n, \quad G = \sum_{n=0}^{\infty} A_n G^n. \quad (1)$$

From this, it would follow easily that the return times to level 0 are governed by the generating function $B^*(z) = \sum_{n=0}^{\infty} z B_n [G^*(z)]^n$. In the Markov chain case, it was advocated to first compute G , then compute $B^*(1)$ from which the steady state vector $\mathbf{x}_0$ for level 0 could be computed up to a scalar multiple, and then use a Gauss–Seidel iteration scheme for higher steady state probabilities; see Neuts [8] for a detailed treatment.

With regard to the equations in (1), two different approaches were taken: (i) Perron–Frobenius theory for nonnegative matrices; (ii) fixed point theorems from functional analysis. It is somewhat interesting to note that the latter already showed that the finiteness of the second coordinate was really not germane to the structure [11,15].

For the computation of G and various moments of first passage times, and so forth, the recommended method was one of iterations based on a bootstrap. Thus for G , one could implement the iterations $G \leftarrow \sum A_n G^n$ starting with the zero matrix. Neuts showed that these iterates are monotonically increasing to the required solution. Several variants of the basic equation were considered, as also the Newton scheme. Neuts demonstrated that the Newton scheme, though quadratically convergent, is computationally too intensive to yield gains in computing time. Later, Ramaswami [16] developed several monotone sub-Newton schemes, but no major breakthrough really occurred for a long time. The present state of the art is to use the logarithmic and cyclic reduction algorithms; refer to the section titled “Algorithms.”

Some notable early applications of the theory can be found in the articles of Neuts [14,17], Ramaswami [18], Lucantoni [19], and others. In his work [14] Neuts revisited the semi-Markov queue and showed that an analysis not incurring any of the unwanted and unverifiable technical assumptions such as those of [1,20] could indeed be made. In Ref. 17, he demonstrated how for a variety of classical queues and other applied probability models, complex variable methods based on Rouché's theorem could be avoided in favor of stable algorithms. Ramaswami [18] generalized the standard $M/G/1$ queue to the case of a batch Markov arrival process, and provided a unified approach with a substantial matrix generalization of the celebrated Pollaczek–Khinchin formula for the stationary waiting time distribution. Lucantoni [19] cast Ramaswami's results in a more convenient notation and added a few additional results of his own. Chaudhury *et al.* [21] provided a detailed transient analysis of this model. Besides these, a large set of examples of the $M/G/1$ paradigm have been analyzed, and many papers in journals like those of the IEEE have appeared applying them to queueing models for computer and communication systems. Even the early monograph of Neuts [8] has numerous examples falling under the matrix $M/G/1$ paradigm.

Early approaches [1,2] had in some sense considered generalizing the classical approach based on the scalar equation $g^*(z) = za^*(g^*(z))$ to matrix-based equations. But with the Rouché's theorem-based approach as guide, they had, instead of the equations in (1), considered the scalar equation in u given by $\det[uI - zA^*(u)] = 0$, which one can quickly identify as the equation for the eigenvalues of $G^*(z)$. Unfortunately, those attempts entailed the difficult questions related to uniqueness of eigenvalues and availability of enough independent eigenvectors one inevitably encounters in spectral decomposition [22].

Returning to Neuts' matrix analytic approach, the literature on the paradigm also included many methodological advances. These include the following: (i) forging connections of the approach with the theory of ladder heights and Wiener–Hopf methods,

Asmussen being a main driver; see Asmussen [23] and references therein; (ii) probabilistic interpretation of various algorithms, notable being the work of Latouche [24] which provided an interesting and useful interpretation of a linear bootstrap algorithm for G ; (iii) Ramaswami's development of a recursive scheme avoiding the arduous Gauss–Seidel iterations for steady state probabilities [25]; (iv) uniformization-based schemes that avoid tedious matrix numerical integrations for $\{A_n\}$, and so on, in various models, Lucantoni and Ramaswami [26].

In this context, it is worth noting that the phase-type distributions introduced by Neuts [27] (see Ref. 7, Chapter 2) and related point process models like the batch Markov arrival process (BMAP) [28] (see also Refs 19 and 9, Chapter 3) provided a tractable set of general models engineers could use and helped to propel the use of matrix analytic methods in diverse applications related to computer and communication engineering as well as to other areas like insurance risk.

THE GI/M/1 TYPE

The classical ergodic $GI/M/1$ type Markov chain is skip-free upward and is known to have a geometric steady state distribution $x_n = x_0 r^n$, where r is the unique solution in $(0, 1)$ of the equation $r = \sum_{n=0}^{\infty} r^n a_n$. As Neuts was able to generalize the $M/G/1$ type to the matrix case, it was natural for him to consider the matrix analog of the $GI/M/1$ type. Neuts [6] demonstrated that the resulting model exemplified by the second matrix structure in the displayed paradigms has a steady state distribution of a matrix-geometric form $\mathbf{x}_n = \mathbf{x}_0 R^n$, where R is the minimal nonnegative solution of the matrix equation

$$R = \sum_{n=0}^{\infty} R^n A_n, \quad (2)$$

and that the vector $\mathbf{x}_0$ could be determined up to a scalar multiple from the first partitioned steady state equation, which reduces to $\mathbf{x}_0 = \mathbf{x}_0 \sum_{n=0}^{\infty} R^n B_n$. The approach of Neuts [6] was purely *ad hoc*, and the mathematical details built on a skillful use of Perron–Frobenius

theory. Neuts quickly followed through with another article [29], providing an interpretation for the matrix R as comprising of the expected number of visits to $(n+1, j)$ by the Markov chain in a return time to level $\mathbf{n}$ starting at (n, i) . Soon a monograph of Neuts [7] emerged providing a detailed account of the theory and numerous example applications. The monograph caught the attention of the applied probability community in a significant way establishing matrix analytic methods as a major set of tools in applied probability.

The quantity R was extended to a kernel $R(z, s)$ related to the transient analysis of the models by Ramaswami [22], who also showed it to satisfy the nonlinear matrix equation

$$R(z, s) = z \sum_{n=0}^{\infty} [R(z, s)]^n A_n(s), \quad (3)$$

and to have the interpretation as the transform of a taboo Markov renewal function yielding the expected number of visits to a higher level at all time points during a return time to the current level; we have, $R = R(1, 0)$. With the power of hindsight, one can say that this article paved the way later on for the transient analysis of many models of the matrix-geometric type, including of stochastic fluid flows. A later article of Ramaswami [11] developed an entirely probabilistic approach to the matrix-geometric results based on the elementary theory of terminating renewal processes, thereby also removing the shackle of finiteness of the second coordinate. An important result pertaining to the infinite dimensional extension is the work of Ramaswami and Taylor [30] demonstrating the celebrated product form result for Jackson networks as a structural result in an operator-geometric characterization.

As with the $M/G/1$ paradigm, the literature abounds in numerous exemplifications of the $GI/M/1$ type structure and its value to applications. Among the theoretical developments, a particularly important one is the characterization by Asmussen (see Ref. 31 and references there) and by Ramaswami [32] using entirely different arguments, of the stationary waiting time distributions in many phase type queues as being of phase

type. There has also been work on several special cases where the matrix R (or the matrix G in the case of the $M/G/1$ type) could be obtained explicitly or shown to have a simplified structure that leads to more efficient computations; for example, Ramaswami and Lucantoni [33], Latouche and Ramaswami [34,35].

To the list of theoretical developments, we should add Ramaswami's duality results [36], which helped to demonstrate a unifying theme between the $M/G/1$ and $GI/M/1$ type models via time reversal. The close resemblance of the nonlinear equations for $G^*(z)$ and $R^*(z)$ begged a result that would connect them, but had eluded researchers for long. The simple intuition behind Ramaswami's finding, as noted in Ref. 37, is that a first passage path from $\mathbf{n} + \mathbf{1}$ to $\mathbf{n}$ in an $M/G/1$ type model when reversed in time gives a taboo upward path from $\mathbf{n}$ to $\mathbf{n} + \mathbf{1}$ avoiding level $\mathbf{n}$ in the time reversal, which is a model of the $GI/M/1$ type. In addition to unifying the results on the two major paradigms, this duality result has been found to be of algorithmic value. Unlike G , the matrix R is not substochastic or stochastic; the corresponding nonlinear equation is usually not a contraction either. Thus iterative methods for R tend to be more difficult than those for G . A simple way to avoid problems is to compute the G matrix of the time reversed dual and then evaluate the needed R matrix using a simple duality transformation.

QBDs AND FLUID FLOWS

The final structure in the set of three paradigms, the QBD, falls in the intersection of the $M/G/1$ and the $GI/M/1$ paradigms and is a favorite of Markov modelers in engineering. The $M/M/1$ queue in a random environment [7] is an example. For QBD's, both $G^*(z)$ and $R^*(z)$ exist, and furthermore the equations governing them reduce to matrix quadratic equations. The QBD thus becomes a special case of the Neuts' paradigms. We note here that matrix-geometric steady state distributions had been obtained earlier for a special case of the QBD by Evans [3], and for the

general case of the QBD in the unpublished dissertation of Wallace [4]; unfortunately, neither authors had developed the approach to its full potential.

Though the QBD is a special case, it has a major importance of its own. Ramaswami [38] demonstrates that the QBD is not only a special case of the $M/G/1$ and the $GI/M/1$ types, but also its theory subsumes those of the more general paradigms. The key to this is a demonstration that, with a suitable embedding in a process with an enlarged state space, the more general paradigms can be realized within a QBD and analyzed via the QBD.

Another reason for the importance of the QBD arises from its connection to Markovian stochastic fluid flows. Asmussen [39] developed a phase type characterization for the steady state distribution of stochastic fluid flow models along with a set of algorithms resembling the linear iterations; his approach, however, was not based on the matrix analytic theory of Neuts. However, Ramaswami [40] made a connection of stochastic fluid flow models to discrete time QBDs and developed quadratically convergent algorithms based on the QBD for steady state results. Ahn and Ramaswami [41] developed this idea further to obtain a complete transient analysis and also developed a powerful quadratically convergent algorithm [42]. This work has generated much interest, and many researchers—Badescu, Bean, Latouche, O'Reilly, Remiche, Soares, Stanford, Taylor, Takine, *et al.*—have since made many additional contributions. Albeit in a different context, for continuous level problems, Sengupta [43] appears to be the first to see a connection with matrix-geometric type models.

ALGORITHMS

We noted that early attempts at numerics were via bootstrapping on the nonlinear matrix equations. The technique provided only linear convergence and could converge very slowly as exemplified for instance in Daigle and Lucantoni [44]. Hence, the need for efficient algorithms was pressing but not

much progress could be made since familiar accelerations of numerical analysis like the Newton scheme either did not yield good results for computational time or were violating the constraints needed such as non-negativity of the solutions.

A major breakthrough occurred in the field with the logarithmic reduction algorithm of Latouche and Ramaswami [45]. The idea was a simple one and had been proposed by Ramaswami in answer to the Daigle and Lucantoni problem [44] and tried out successfully: “If one knew R^2 in a QBD, one could determine R from a linear equation; so, get R^2 which has a smaller spectral radius (it is when the spectral radius of R is near 1 that one has problems of convergence); and we can get R^2 by considering the embedded QBD on levels $\{0, 2, 4, \dots\}$. Indeed, this idea could be iterated a few times until an equation with a nice behavior obtains.” However, no formal analysis of the method was done for long.

During a visit of Latouche to Bellcore, Latouche and Ramaswami worked together and demonstrated in an article [45] that analyzing the sequence of embedded QBDs on levels $\{0, 2^n, 2 \times 2^n, 3 \times 2^n, \dots\}$ does indeed lead to an explicit expression for G , and that in turn provides an approximation with quadratic convergence. Enabling the proof were two different pieces of earlier work of the authors: analysis of the linear algorithm by Latouche [24] and a duality result of Ramaswami and Neuts [46], the former providing the probabilistic interpretation for the steps and the latter automatically implying quadratic convergence of the scheme. The name of the algorithm arises from the fact that while the linear algorithm covers in n steps exactly n levels up in a return time to a given level, the present algorithm covers 2^n levels above; thus the reduction in time in covering excursions to higher levels is reduced by a logarithmic factor. For mathematical details of the algorithm, we refer the reader to Latouche and Ramaswami [9, 45]. The algorithm has been analyzed by Ye [47] who has established its numerical stability and also made a valuable addition to find a solution to a (very rare) problem encountered by the method.

In the form in which it was derived in the article [45], it was not easy to generalize the algorithmic steps of the logarithmic reduction algorithm to the $M/G/1$ type. However, two able numerical analysts Bini and Meini from Pisa quickly established connections of the logarithmic reduction algorithm to procedures called *cyclic reduction* and *even-odd reduction* in linear algebra and developed, in a series of articles (see Refs 48 and 49), a quadratically convergent algorithm for the more general $M/G/1$ type of models. The basic tool of their analysis is Schur complementation for Toeplitz systems. Recent papers of Hunt [50] and Bini *et al.* [51] show, however, that the Bini–Meini method is equivalent to a logarithmic reduction with a slightly different set of successive filtering of the levels into the sets of levels $\{0, 1, 1 + 2^n, 1 + 2 \times 2^n, 1 + 3 \times 2^n, \dots\}$. We hasten to note, however, that the contributions of Bini and Meini extend far beyond the algorithm and cover numerous enhancements to it, including efficient implementations using fast Fourier techniques. For a detailed discussion and extensive bibliography, we refer the reader to the recent monograph by Bini *et al.* [10].

OTHER DEVELOPMENTS AND EXTENSIONS

The above discussion covers the core theory and algorithms, but the contributors and contributions to matrix analytic methods are much more numerous than what we have been able to cite in a brief discussion. The objective of this section is to highlight some other major players and developments in the field with due apologies for omissions inadvertent or necessitated by brevity. For specific work not cited in the bibliography, see the referenced texts. More recent ones may be found easily through a web search.

Finite state models. The QBD with a finite number of levels is exceptional in that it leads to a characterization of its steady state vector as a sum of two matrix-geometric terms [52]. Besides this, other models rarely result in interesting near closed form formulae. However, the basic ideas of matrix analytic methods can be employed efficaciously as shown

for instance by Gaver, Jacobs and Latouche. One may also develop efficient algorithms based on the idea of even-odd reduction for steady state probabilities and first passage times. Details and references may be found in the book of Latouche and Ramaswami ([9], Chapter 10). In this context, we wish to note that the Russian school led by Professor P. Bocharov had made much progress on finite state space models along the lines of matrix analytic methods, but the work unfortunately went unnoticed outside the former Soviet Union.

Spatial inhomogeneity. The basic ideas extend to processes without spatial homogeneity, but at the expense of complexity in that each level has its own R or G matrix [11]. While this makes a general theory difficult, major inroads have been made by Taylor and his colleagues, with Bright and Taylor [53] being a noteworthy beginning for systematic efforts.

Non-skip-free models. Moran’s dam considered in Neuts [17] is an early example of a model that does not have a structure amenable to matrix analytic methods, but could be turned into one by a clever reblocking of the states. The literature has a number of examples where such reblocking is employed, and among them we draw particular attention to the work of Gail *et al.* [54], which attempts to unify a lot of special cases.

Tree structured models. Underlying matrix analytic methods is the idea of a “neighbor” class of states and connectivity restrictions among them. Thus, the basic algorithms could be extended to instances where levels are not restricted to be integers but could be more general like nodes in a tree. Such models arise, for example, in queues with the LIFO discipline. Yet, the extension is not routine in that additional structures are obtained that significantly simplify computations. Yeung and Sengupta [55] and Takine *et al.* [56] are credited for this important discovery, and two often cited articles are Yeung and Alfa [57] and He and Alfa [58]. The teams Bini, Meini, and Latouche and Blondia and Van Houdt have provided efficient algorithms geared to these specific structures.

Retrial queues. A renewed interest due to call centers has spawned off much recent work on retrial queues with orbiting customers. There are two early articles in this area using matrix analytic methods [59,60]. For many others, see the various proceedings of the Conferences on Retrial Queues organized by Artalejo and Krishnamoorthy [61].

Discrete time models. Many practical queueing systems operate in discrete time, and discrete time queues present opportunities for exploring additional structures and nuances. Indeed, the early applications of matrix analytic methods were to a set of discrete time queues; see the papers of Neuts with Heimann [62] and Dafermos [63] in the Naval Research Logistic Quarterly. Discrete time queues continue to be an active area of research with some notable work in this area by Alfa and his colleagues.

Related methods. Interconnections with established methods of mathematics abound. Thus, earlier spectral methods based on eigenvalues, and others, have new interpretations in terms of the matrix analytic structure; these have been explored in depth by many researchers like Hantler and Taylor. A prescient work in this area is that of Neuts [64] related to the importance of the largest eigenvalue of the matrix R as a major determinant of path behavior, and the highly used “equivalent bandwidth” [65] for packet queues based on fluid flow models can now be directly related to that through the connection of the fluid model to QBDs. Researchers like Grassman and Heyman have established connections with the theory of censoring for Markov chains, while Akar and Sohraby have approached matrix analytic methods via an invariant subspace approach. Lipsky approached the Markov chain problems here through a purely linear algebraic approach. These connections add greater luster to the area and provide additional perspectives.

Real-world applications. Matrix analytic methods have been applied to many real-world problems in computer performance and communications engineering, and related articles may be found notably in *Performance Evaluation*, *Queueing Systems*, and the IEEE journals. New applications are found

in insurance risk theory and mathematical finance. These are too numerous to list individually, and we cite only a few simple examples here [66–68] to highlight the continuing relevance of the methods to modern day engineering.

Numerical analysis. Matrix analytic methods have gained much from classical applied mathematics and linear algebra. It is therefore a delight to see the field returning the favor by way of new tools. Examples of these are the extensions by Bini *et al.*, of the cyclic reduction methods for solving matrix polynomial equations efficiently [69], as well as the emerging new algorithms by Guo, Higham, Bini, and others (see Ref. 70 for an example) for algebraic Riccati equations. The latter is a take-off from Ramaswami’s work on fluid flow models that essentially involves turning a Riccati equation into a matrix quadratic.

Software. A major effort has been launched by the researchers Bini, Steffé, and Meini in Pisa and Van Houdt in Antwerp to develop downloadable software for the key algorithms underlying matrix analytic methods. For details on these, refer to [71].

Playing a significant role in the rapid development of matrix analytic methods has been a series of international MAM conferences starting with the one organized by Chakravathy and Alfa in 1995. So far, a set of six conferences have taken place, and their proceedings contain many interesting papers. These conferences have enabled the community of researchers to interact frequently, disseminate results faster, and exchange ideas.

What has been summarized here is the achievement of a large community of colleagues who have generously shared their ideas and research. We dedicate this note to them in deep friendship and gratitude.

REFERENCES

1. Çinlar E. Queues with semi-Markovian arrivals. *J Appl Probab* 1967;4:365–379.
2. Conolly BW. The busy period in relation to the single server queueing system with general independent arrivals and Erlangian service time. *J R Stat Soc [Ser B]* 1969;22:89–96.

3. Evans RV. Geometric distribution in some two-dimensional queueing systems. *Oper Res* 1967;15:830–846.
4. Wallace V. The solution of quasi birth and death processes arising from multiple access computer systems [PhD thesis]. Ann Arbor (MI): Systems Engineering Laboratory, University of Michigan; 1969.
5. Neuts MF. The Markov renewal branching process. *Proceedings of the Conference on Mathematical Methodology in the Theory of Queues*. Kalamazoo (MI): Springer; 1974. pp. 1–21.
6. Neuts MF. Markov chains with applications in queueing theory, which have a matrix-geometric invariant vector. *Adv Appl Probab* 1978;10:185–212.
7. Neuts MF. Matrix-geometric solutions in stochastic models, an Algorithmic approach. Baltimore (MD): The Johns Hopkins University Press; 1981.
8. Neuts MF. Structured stochastic matrices of M/G/1 type and their applications. New York: Marcel Dekker; 1989.
9. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modelling. Philadelphia (PA): SIAM & ASA; 1999.
10. Bini D, Latouche G, Meini B. Numerical methods for structured Markov chains. Oxford, (UK): Oxford University Press; 2005.
11. Ramaswami V. Matrix analytic methods: a tutorial overview with some extensions and new results. In: Chakravorthy SR, Alfa AS, editors. *Matrix analytic methods in stochastic models*. New York: Marcel Dekker; 1997. pp. 261–295.
12. Alfa AS. Discrete time queues and matrix analytic methods. *TOP* 2002;10:147–210.
13. Takács L. Introduction to the theory of queues. New York: Oxford University Press; 1962.
14. Neuts MF. Some explicit formulas for the steady state behavior of the queue with semi-Markovian service time. *Adv Appl Probab* 1977;9:141–157.
15. Tweedie RL. Operator-geometric stationary distribution for Markov chains, with applications to queueing models. *Adv Appl Probab* 1982;14:368–391.
16. Ramaswami V. Nonlinear matrix equations in applied probability - solution techniques and open problems. *SIAM Rev* 1988;30:256–263.
17. Neuts MF. Queues solvable without Rouché's theorem. *Oper Res* 1978;27:767–781.
18. Ramaswami V. The N/G/1 queue and its detailed analysis. *Adv Appl Probab* 1980;12:222–261.
19. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Stoch Models* 1991;7(1):1–46.
20. Neuts MF. The single server queue with Poisson input and semi-Markov service times. *J Appl Probab* 1966;3:202–230.
21. Choudhury GL, Lucantoni DM, Whitt W. The transient BMAP/G/1 queue. *Stoch Models* 1994;10:145–182.
22. Ramaswami V. The busy period of queues which have a matrix geometric steady state probability vector. *Opsearch* 1982;19:238–261.
23. Asmussen S. *Applied probability and queues*. 2nd ed. New York: Springer; 2003.
24. Latouche G. Algorithms for infinite Markov chains with repeating columns. IMA workshop on linear algebra, markov chains and queueing models. Minnesota: University of Minnesota; 1992.
25. Ramaswami V. A stable recursion for the steady state vector in Markov chains of M/G/1 type. *Stoch Models* 1988;4:183–188.
26. Lucantoni DM, Ramaswami V. Efficient algorithms for solving the non-linear matrix equations arising in phase type queues. *Stoch Models* 1985;1:29–52.
27. Neuts MF. Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin*. Belgium: University of Louvain; 1975. pp. 173–206.
28. Neuts MF. A versatile Markovian point process. *J Appl Probab* 1979;16:764–779.
29. Neuts MF. The probabilistic significance of the rate matrix in matrix-geometric invariant vectors. *J Appl Probab* 1980;17:291–296.
30. Ramaswami V, Taylor PG. An operator-analytic approach to product-form networks. *Stoch Models* 1996;12:121–142.
31. Asmussen S. Phase type representation of waiting times. In: Cohen JW, Pack CD, editors. *Queueing, performance & control in ATM*. Amsterdam: North Holland Publishing Co.; 1991. pp. 157–159.
32. Ramaswami V. From the matrix-geometric to the matrix-exponential. *QUESTA* 1990;6:229–260.
33. Ramaswami V, Lucantoni DM. Algorithmic analysis of a dynamic priority queue. In: Disney RL, Ott TJ, editors. *Volume II, Applied probability - computer science: the interface*. New York: Birkhauser; 1981. pp. 157–206.

34. Latouche G, Ramaswami V. A general class of Markov processes with explicit matrix-geometric solutions. *OR Spektrum* 1986;8:209–218.
35. Latouche G, Ramaswami V. The $PH/PH/1$ queue at epochs of queue size changes. *QUESTA* 1997;25:97–114.
36. Ramaswami V. A duality theorem for the matrix paradigms in queueing theory. *Stoch Models* 1990;6:151–161.
37. Asmussen S, Ramaswami V. Probabilistic interpretations of some duality results for the matrix paradigms in queueing theory. *Stoch Models* 1990;6:715–734.
38. Ramaswami V. The generality of the quasi birth and death process. In: Alfa AS, Chakravarthy S, editors. *Advances in matrix analytic methods for stochastic models*. New Jersey: Notable Publications, Inc.; 1998.
39. Asmussen S. Stationary waiting time distributions for fluid flow models with or without Brownian noise. *Stoch Models* 1995;11:1–20.
40. Ramaswami V. Matrix analytic methods for stochastic fluid flows. In: Key P, Smith D, editors. *Teletraffic engineering in a competitive world, ITC-16 Proceedings*. Edinburgh, Scotland: Elsevier; 1999. pp. 1019–1030.
41. Ahn S, Ramaswami V. Transient analysis of fluid flow models via stochastic coupling to a queue. *Stoch Models* 2004;20:71–102.
42. Ahn S, Ramaswami V. Efficient algorithms for transient analysis of stochastic fluid flow models. *J Appl Probab* 2005;42(2):531–549.
43. Sengupta B. Markov processes whose steady state distribution is matrix exponential with an application to the $GI/G/1$ queue. *Adv Appl Probab* 1989;21:159–180.
44. Daigle JN, Lucantoni DM. Queueing systems having phase dependent arrival and service rates. In: Stewart WJ, editors. *Numerical solutions of Markov chains*. New York: Marcel Dekker; 1991. pp. 161–202.
45. Latouche G, Ramaswami V. A logarithmic reduction algorithm for quasi birth and death processes. *J Appl Probab* 1993;30:650–674.
46. Ramaswami V, Neuts MF. A duality theorem for phase type queues. *Ann Probab* 1980;17:498–514.
47. Ye Q. On Latouche-Ramaswami's logarithmic reduction algorithm for quasi-birth-and-death processes. *Stoch Models* 2002;18:449–467.
48. Bini D, Meini B. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J Matrix Anal Appl* 1996;17:906–926.
49. Bini D, Meini B. Improved cyclic reduction for solving queueing problems. *Numer Algorithms* 1997;15:57–74.
50. Hunt E. Determining the fundamental matrix of a block $M/G/1$ Markov chain. In: Osaki RJ, Faddy MJ, editors. *Stochastic models in engineering, technology & management*. Queensland, Australia: University of Queensland; 1999. 446–455.
51. Bini D, Meini B, Ramaswami V. A probabilistic interpretation of cyclic reduction and its relationships with logarithmic reduction. *CALCOLO* 2008;45(3):207–216.
52. Hajek B. Birth and death processes on the integers with phases and general boundaries. *J Appl Probab* 1982;19:488–499.
53. Bright L, Taylor PG. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stoch Models* 1995;11:497–525.
54. Gail HR, Hantler SL, Taylor BA. Non-skip-free $M/G/1$ and $G/M/1$ type Markov chains. *Adv Appl Probab* 1997;29:733–758.
55. Yeung RW, Sengupta B. Matrix product-form solutions for Markov chains with a tree structure. *Adv Appl Probab* 1994;26:965–987.
56. Takine T, Sengupta B, Yeung RW. A generalization of the matrix $M/G/1$ paradigm for Markov chains with a tree structure. *Stoch Models* 1995;11:411–421.
57. Yeung RW, Alfa AS. The quasi-birth-and-death type Markov chain with a tree structure. *Stoch Models* 1999;15:639–659.
58. He Q-M, Alfa AS. The $MMAP[K]/PH[K]/1/LCFS-GPR$ queue and its variants. In: Latouche G, Taylor PG, editors. *Advances in algorithmic methods for stochastic models*. New Jersey: Notable Publications Inc.; 2000.
59. Diamond J, Alfa AS. Analysis of $M/PH/1$ retrial queues. *Stoch Models* 1995;11(3):447–470.
60. Diamond J, Alfa AS. The $MAP/PH/1$ retrial queue. *Stoch Models* 1998;14(5):1151–1177.
61. Artalejo JR, Krishnamoorthy A, editors. *Advances in Stochastic Modelling*. New Jersey: Notable Publication Inc.; 2002.
62. Heimann D, Neuts MF. The single server queue in discrete time: Numerical Analysis IV. *Nav Res Logist Q* 1973;20:753–766.
63. Dafermos S, Neuts MF. A single server queue in discrete time. *Cahiers du Centre Recherche Operationnelle* 1971;13:23–40.
64. Neuts MF. The caudal characteristic of queues. *Adv Appl Probab* 1986;18:535–557.

65. Elwalid AI, Mitra D. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans Netw* 1993;1(3): 329–343.
66. Foh CH, Meini B, Wdrowski B, *et al.* Modeling and performance evaluation of GPRS. *Proceedings of IEEE VTC*. Rhodes, Greece; 2001;
67. Ramaswami V, *et al.* Assuring emergency services access: providing dial tone in the presence of long holding time internet dial-up calls, 2004 INFORMS Wagner Prize finalist paper. *Interfaces* 2005;35:411–422.
68. Squillante MS, Nelson RD. Analysis of task migration in shared-memory multiprocessor scheduling. *ACM Sigmetrics Performance Evaluation Review* 1991;19:143–155.
69. Bini D. Solving quadratic matrix equations and factoring polynomials: new fixed point iterations based on Schur complements and Toeplitz matrices. *Numer Linear Algebra Appl* 2005;12:181–189.
70. Guo C, Higham NJ. Iterative solution of non-symmetric Riccati equation. *SIAM J Matrix Anal Appl* 2007;29(2):396–412.
71. Bini D, Meini B, Steffé S, *et al.* Structured Markov chains solver: software tools. *ACM International Conference Proceedings Series, Volume 201, Proceedings from the 2006 Workshop on tools for Solving Structured Markov Chains*. Pisa, Italy; 2006.

MATRIX-GEOMETRIC DISTRIBUTIONS

DIETER BAUM
Department IV, Subdepartment
of Computer Science,
University of Trier, Trier,
Germany

INTRODUCTION

In this article, matrix-geometric distributions are introduced by emphasizing the significance of structure. We use capital letters for matrices in the following, for example, I_k for the identity (or unit) matrix of order k , O for the null matrix (of any order), and bold face capital letters for matrices in block form. Row vectors are typed as bold face lowercase letters, in particular $\mathbf{o} = (0, 0, \dots)$, $\mathbf{e} = (1, 1, \dots)$, where subscripts may indicate lengths. Transpose vectors or matrices are marked by a superscript T .

Let ξ denote a discrete random variable defined on a lattice $\{x_{nj}\}_{n \in \mathbb{N}_0, j \in \{1, \dots, m\}}$, the rows of which form a countable family $\{\mathbf{x}_n\}_{n \in \mathbb{N}_0}$ of m -dimensional real vectors $\mathbf{x}_n = (x_{n1}, \dots, x_{nm})$ (m may be infinite). Set $\mathbb{P}(\xi = x_{nj}) = \pi_{nj}$, $\boldsymbol{\pi}_n = (\pi_{n1}, \dots, \pi_{nm})$. The distribution of ξ is said to be of *matrix-geometric type* if there is some constant square matrix R of order m such that

$$\boldsymbol{\pi}_{n+s} = \boldsymbol{\beta} \cdot R^n \quad \forall n \geq 0, \quad 0 \leq s < \infty \quad (1)$$

($\boldsymbol{\beta} \in \mathbb{R}^m$ a constant vector). Obviously, for $m = 1$ and $s = 0$ Equation (1) resembles the customary expression $\mathbb{P}(\xi = n) = (1 - \rho) \cdot \rho^n$ for a geometric random variable with parameter ρ . Within the framework of vector state Markov models, that is, $m > 1$, the lattice is made up by pairs $x_{nj} = (n, j)$, $n \in \mathbb{N}_0$, $j \in \{1, \dots, m\} = E$. The number n in (n, j) —or, occasionally, the entire *macrostate* $\mathbf{n} = ((n, 1), \dots, (n, m))$ —is termed *level*. An irreducible recurrent Markov process with an

invariant vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ satisfying Equation (1) is called *matrix-geometric Markov process*, and the techniques to solve for this distribution by constructing the matrix R are classified under the general appellation *matrix-geometric methods*. Correspondingly, stochastic models defining a matrix-geometric Markov process are called *matrix-geometric models* or *models of matrix-geometric structure*.

In what follows, we use the term *characteristic matrix* to designate the transition probability matrix P of some discrete time Markov process (Markov chain), or rather the generator matrix Q in the continuous time case. A vector state Markov process X is called *skip-free upward in levels* if the elements $w_{(n_1, i); (n_2, j)}$ ($n_1, n_2 \in \mathbb{N}_0, i, j \in E$) of its characteristic matrix W satisfy $w_{(n, i); (n+k, j)} = 0 \quad \forall k > 1$, and is called *spatially homogeneous* from ℓ on if $w_{(n_1+k, i); (n_2+k, j)}$ does not depend on k for $n_1 \geq \ell, n_2 \geq \ell$. The grouping of state probabilities π_{nj} for $n \geq 1$ as $\boldsymbol{\pi}_n = (\pi_{n1}, \dots, \pi_{nm})$ induces a partitioning of W , converting it into a block matrix $\mathbf{W}$. Assume that X is spatially homogeneous from $\ell = 2$ on and irreducible recurrent (as well as aperiodic in the discrete time case). Then, as we shall show, the matrix-geometric solution (Eq. 1) is a consequence from the lower block Hessenberg structure of $\mathbf{W}$ which results from the skip-free upward property

$$\mathbf{W} = \begin{bmatrix} B_{00} & B_{01} & O & O & O & \dots \\ B_{10} & B_{11} & A_0 & O & O & \dots \\ B_{20} & B_{21} & A_1 & A_0 & O & \dots \\ B_{30} & B_{31} & A_2 & A_1 & A_0 & \dots \\ B_{40} & B_{41} & A_3 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (2)$$

Equation (2) represents the main structure of interest. The matrices $B_{n_1 n_2}$ ($n_1 \in \mathbb{N}_0$, $n_2 \in \{0, 1\}$) are called *boundary matrices*. B_{00} is a square matrix of order $m_0 \in E$, B_{01} an $m_0 \times m$ matrix, and for $n_1 \geq 1$ the $B_{n_1 n_2}$ represent $m \times m_0$ matrices for

$n_2 = 0$, and $m \times m$ matrices for $n_2 = 1$. The corresponding states are called *boundary states*, all other *repeating states*. Since Equation (2) resembles the transition matrix of the embedded Markov chain at arrival instants of a $GI/M/1$ queueing model, it is termed *canonical $GI/M/1$ form*, and W is said to be of *$GI/M/1$ type* (the *repetitive upper Hessenberg structure* of $M/G/1$ type matrices does not lead to a matrix-geometric solution).

We mention three special examples:

1. The simple $M/M/1$ queueing system in equilibrium, leading to a continuous time Markov process with geometric steady-state distribution. Let N_t denote the number of customers resident at time t , and λ and μ the mean arrival and service rates, respectively. Then, given that $\rho = \frac{\lambda}{\mu} < 1$, $\lim_{t \rightarrow \infty} \mathbb{P}(N_t = n) = \pi_n = \beta \cdot \rho^n \quad \forall n \geq 0$ with $\beta = \pi_0$. State transitions take place between neighboring states only, and so the infinitesimal generator Q of the process $\{N_t : t \geq 0\}$ possesses a tri-diagonal structure with “repeating rows”, describing a birth-and-death queueing system. It is this tri-diagonal repetitive structure (resembling a lower Hessenberg form for $\ell = 1$) that lends to the geometric stationary distribution. More generally, given an irreducible recurrent continuous time Markov process on a denumerable state space with generator

$$Q = \begin{bmatrix} b_0 & a_0 & 0 & 0 & 0 & \dots \\ a_2 & a_1 & a_0 & 0 & 0 & \dots \\ 0 & a_2 & a_1 & a_0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

the limiting state probabilities satisfy $\pi_n = \pi_{n-1} \frac{a_0}{a_2} = \pi_0 \left(\frac{a_0}{a_2} \right)^n$ for $n \geq 1$, and the process is ergodic if and only if $a_0 < a_2$. This result, although easily obtained from $\pi Q = \mathbf{o}$ with $\pi = (\pi_0, \pi_1, \dots)$ and $Q\mathbf{e}^T = \mathbf{o}^T$, can be proven by using purely probabilistic arguments from the theory of terminating renewal processes [1].

2. The $M/PH/1$ queueing system as an example for a continuous time matrix-geometric model with $m > 1$. Service periods are times until absorption in a finite state Markov process $\{J_t : t \geq 0\}$ with transient states $1, \dots, m$ and absorbing state $m+1$. Let α describe the “ m -portion” of the initial probability vector $(\alpha_1, \dots, \alpha_m, \alpha_{m+1}) = (\alpha, \alpha_{m+1})$, and $S = [S_{ij}]$ the $m \times m$ matrix of state transition rates among transient states, such that $S_{ii} < 0$, $S_{ij} \geq 0$ for $i \neq j$, and $S\mathbf{e}^T = -\eta^T$, η^T the column vector of absorption rates; then (α, S) is the representation of the phase type distribution (see **Phase-Type (PH) Distributions**). Absorption occurs with probability 1 from any transient state $i \in \{1, \dots, m\}$ if and only if S is nonsingular, and the system assumes equilibrium provided the product of mean service time $\mathbb{E}[H]$ and mean arrival rate λ satisfies $\mathbb{E}[H] \cdot \lambda = \alpha(-S^{-1})\mathbf{e}^T \cdot \lambda < 1$. Given that no group services occur and each service duration is > 0 , we have that $\alpha_{m+1} = 0$. The underlying stochastic process $\{(N_t, J_t) : t \geq 0\}$, where N_t stands for the number of customers resident in the system at time t , is a homogeneous ergodic Markov process on the state space $\{0\} \cup \{(n, j) : n \geq 1, j \in \{1, \dots, m\}\}$ whose infinitesimal generator Q in block form reads

$$Q = \begin{bmatrix} B_{00} & B_{01} & \mathbf{o} & \mathbf{o} & \mathbf{o} & \dots \\ B_{10} & A_1 & A_0 & O & O & \dots \\ \mathbf{o}^T & A_2 & A_1 & A_0 & O & \dots \\ \mathbf{o}^T & O & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}. \quad (3)$$

B_{00} equals the scalar $-\lambda$, B_{01} is the row vector $\lambda\alpha$, B_{10} is the column vector η^T , and the A_i are $m \times m$ matrices given by $A_2 = \eta^T\alpha$, $A_1 = S - \lambda I_m$, and $A_0 = \lambda I_m$. From $\pi Q = \mathbf{o}$, by virtue of the repetitive triangular structure of Q , the following relations result: $\pi_1 \eta^T = \lambda \pi_0$, $\pi_0 \lambda \alpha + \pi_1 A_1 + \pi_2 A_2 = \mathbf{o}$, and $\pi_{n-1} A_0 + \pi_n A_1 + \pi_{n+1} A_2 = \mathbf{o} \quad \forall n \geq 2$.

Exploiting global balance, that is, equating the flow *out* of any macrostate $\mathbf{n} = ((n, 1), \dots, (n, m))$ due to a service completion to the flow *into* the same macrostate caused by an arrival, one is led to $\pi_{n+1}\eta^T = \lambda\pi_n\mathbf{e}^T$ for $n \geq 1$, and this yields $\pi_n(I_m - \mathbf{e}^T\alpha - \lambda^{-1}S) = \pi_{n-1}$. The matrix $\mathbf{e}^T\alpha + \lambda^{-1}S = M$ is indecomposable [2], and $I_m - M$ can easily be proven to be regular. Hence, Equation (1) holds with $s = 1$, $\beta = \pi_0 \cdot \alpha$ and $R = (I_m - \mathbf{e}^T\alpha - \lambda^{-1}S)^{-1}$. By rule of vector-matrix multiplication, Equation (1) implies

$$A_0 + RA_1 + R^2A_2 = O. \quad (4)$$

The matrix R is called *rate matrix*. Two-dimensional Markov processes with block tri-diagonal characteristic matrix have been coined *Quasi birth-and-death* processes (QBDs) by Wallace [3]. Their structure combines the $GI/M/1$ and $M/G/1$ types. The solution (1) for QBDs was first obtained by Evans [4]. The Markov process $\{J_t : t \geq 0\}$ with state space $E = \{1, \dots, m\}$ is termed *phase process*, its states $j \in E$ are called *phases* or *interlevel components*.

3. The $GI/PH/1$ queueing system with an arbitrary interarrival time distribution $F(\cdot)$ that satisfies $\int_0^\infty \tau dF(\tau) < \infty$ and $\lim_{t \downarrow 0} F(t) = 0$ (a generalization of the previous example). The primary ingredient here is the probability transition matrix P of the embedded Markov chain at arrival instants. According to the state grouping, P assumes the lower block Hessenberg form

$$P = \begin{bmatrix} B_0 & A_0 & O & O & O & \dots \\ B_1 & A_1 & A_0 & O & O & \dots \\ B_2 & A_2 & A_1 & A_0 & O & \dots \\ B_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

which leads to the solution (1) for the steady-state probabilities at arrival instants, where R satisfies $\sum_{n=0}^\infty R^n A_n = R$. Once π is known, the

steady-state probabilities at arbitrary time can be computed by applying the key renewal theorem (see Ref. 5 for details).

The investigation of infinite stochastic matrices whose stationary probability distributions assume the matrix-geometric form has been pioneered by Neuts [5–7] and his school. An approach based on terminating renewal processes (as outlined below) originates from the work of Ramaswami [8]. Essentially, in this context, continuous time Markov processes and discrete time Markov processes (Markov chains) can be handled alike. For that reason we may concentrate mainly on the discrete time case.

PROBABILISTIC DERIVATION

To begin with, consider an irreducible aperiodic and spatially homogeneous Markov chain $X = \{(N_v, J_v) : v \in \mathbb{N}_0\}$ on the state space $\mathbb{N}_0 \times E$ whose transition matrix $P = [p_{(n_1, i); (n_2, j)}]_{n_1, n_2 \in \mathbb{N}_0, i, j \in E}$ possesses the *skip-free upward* property, in which case $P = W$ is a lower block Hessenberg matrix as given in Equation (2). Let $p_{(r, i); (s, j)}^{(k)} = \mathbb{P}((N_k, J_k) = (s, j) \mid (N_0, J_0) = (r, i))$ denote the k -step transition probabilities, $P_{r; s}^{(k)} = [p_{(r, i); (s, j)}^{(k)}]_{i, j \in E}$, the corresponding matrices with respect to phases. For $k = 1$ the superscript is omitted. The k -step taboo transition probability from (n, i) to $(n + s, j)$ under taboo of (i.e., avoiding in between) the set of macrostates $T = \{\mathbf{o}, \dots, \mathbf{n}\}$ is designated as $TP_{(n, i); (n+s, j)}^{(k)}$, and its matrix as $TP_{n; n+s}^{(k)}$ ($s \in \mathbb{N}_0$). Owing to the skip-free upward property, the last macrostate in T before visiting $\mathbf{n} + 1$ must be $\mathbf{n}$; therefore, $TP_{(n, i); (n+s, j)}^{(k)} = nP_{(n, i); (n+s, j)}^{(k)}$. By conditioning on the last visit in T before reaching $\mathbf{n} + 1$ at step k , the following is obtained

$$\begin{aligned} P_{(0, i); (n+1, j)}^{(k)} &= \sum_{\ell=1}^m \sum_{h=1}^m \sum_{v=0}^{k-1} P_{(0, i); (n, \ell)}^{(k-1-v)} \cdot P_{(n, \ell); (n+1, h)} \\ &\quad \cdot nP_{(n+1, h); (n+1, j)}^{(v)} \\ P_{0; n+1}^{(k)} &= \sum_{v=0}^{k-1} P_{0; n}^{(k-1-v)} \cdot (P_{n; n+1} \cdot nP_{n+1; n+1}^{(v)}) \end{aligned}$$

$$=: \sum_{v=0}^{k-1} H_{k-(v+1)} \cdot {}_n P_{n;n+1}^{(v+1)} \quad (5)$$

(matrices H_r defined by $H_r = P_{0;n}^{(r)}$). Subsequent epochs of visits to state $(n+1, j)$ after start in (n, ℓ) define, for any fixed n , a delayed renewal process (with some delay time distribution G), which is terminating since X is irreducible. For $n \geq 2$, all these renewal processes are indistinguishable due to the proposed spatial homogeneity, so we may speak of *the* renewal process for $n \geq 2$ and ℓ, j fixed. Since ${}_n p_{(n,\ell);(n+1,j)}^{(v)} =: r_v(n) = r_v$ gives the probability that v is a renewal point, the sequence $\{r_v\}_{v \in \mathbb{N}_0}$ constitutes the delayed renewal density of this process, and $R_k = \sum_{\ell=0}^{\infty} (G * F^{*\ell})_k = \sum_{v=1}^k r_v$ its renewal function, where F_v denotes the distribution function of interrenewal times. As is well known [1,9,10], for an unknown sequence $u = \{u_v\}_{v \in \mathbb{N}_0}$ and a known sequence $\{H_v\}_{v \in \mathbb{N}_0}$, an equation of the form $u_v = H_v + (F * u)_v$ represents a renewal equation whose solution is given by $u_k = \sum_{v=0}^k H_{k-v} \cdot r_v$. Thus, Equation (5) is the solution of a renewal equation. For a terminating renewal process, that is, if $\lim_{v \rightarrow \infty} F_v = F_{\infty} < \infty$, the limiting behavior of this solution takes the form $\lim_{v \rightarrow \infty} H_v \cdot \lim_{k \rightarrow \infty} R_k$ (cf. Çinlar [9]). Consequently, since the k -step probabilities $p_{(0,i);(n+1,j)}^{(k)}$ assume a limit as $k \rightarrow \infty$ (independently from the initial state) Equation (5) yields

$$\pi_{n+1} = \pi_n \cdot R = \pi_1 \cdot R^n \quad \forall n \geq 1, \quad (6)$$

where $R = R_{\infty} = \sum_{v=0}^{\infty} P_{n;n+1} \cdot {}_n P_{n+1;n+1}^{(v)} = \sum_{v=1}^{\infty} {}_n P_{n;n+1}^{(v)}$. It should be noted that similar inferences apply to spatially nonhomogeneous (*level-dependent*) chains, in which case level-dependent rate matrices replace R . Entry R_{ij} of R is the expected total number of visits to state $(n+1, j)$ before the chain returns to $T = \{\mathbf{o}, \dots, \mathbf{n}\}$, given that it started in (n, i) .

The z -transform of v -step taboo transition probability matrices ${}_n P_{n;n+1}^{(v)}$ avoiding $T = \{\mathbf{o}, \dots, \mathbf{n}\}$ is given by $R(z) = \sum_{v=1}^{\infty} r_v \cdot z^v = \sum_{v=1}^{\infty} {}_n P_{n;n+1}^{(v)} \cdot z^v$. Accordingly, the z -transform of the v -step transition probability matrices from $\mathbf{o}$ to $\mathbf{n}$ reads $X_n(z) = \sum_{v=1}^{\infty} P_{0;n}^{(v)} \cdot z^v$. Similar expressions

are valid also for *transient* spatially homogeneous chains, leading to $X_{n+1}(z) = X_n(z) \cdot R(z)$. Observing $P_{n+1+k;n+1} = A_k$ for $n \geq 1$, and ${}_n P_{n;n+1}^{(v)} = \delta_{v1} P_{n;n+1} + (1 - \delta_{v1}) \sum_{k=0}^{\infty} {}_n P_{n;n+1+k}^{(v-1)} P_{n+1+k;n+1}$, where $\delta_{v\mu}$ is the Kronecker symbol, the following expressions are obtained (see Refs. 1 and 8 for details):

$$R(z) = z \sum_{k=0}^{\infty} [R(z)]^k A_k, \quad R = \sum_{k=0}^{\infty} R^k A_k. \quad (7)$$

Note that the result for $z = 1$ in Equation (7) as well as the relation $\pi_n = \sum_{k=0}^{\infty} \pi_{n-1+k} A_k \quad \forall n \geq 2$ are also consequences of Equation (6) when inserted in $\pi P = \pi$. For QBDs, the matrices A_k , for $k \geq 3$, have to be replaced by the null matrix O . For an ergodic chain, the sum $\sum_{n=0}^{\infty} \pi_n = \pi_0 + \pi_1 (I_m + (I_m - R)^{-1}) = \mathbf{p}$ is the steady-state probability vector of the phase process. If $m < \infty$, the inverse $(I_m - R)^{-1} = \sum_{n=0}^{\infty} R^n$ exists, and the spectral radius $sp(R)$ of R is strictly less than 1, which is equivalent to ergodicity [11]; see also [12]. The case $m = \infty$ can be handled by more intricate but, nevertheless, similar techniques (see Ramaswami [8]). In that case the inverse $(I_m - R)^{-1}$ is defined by $(I_m - R)^{-1} = \sum_{n=0}^{\infty} R^n$ (see Seneta [13] for relevant issues on infinite matrices).

The case of a *continuous time* Markov process with spatially homogeneous lower block Hessenberg generator $[q(r,i);(s,j)]_{r,s \in \mathbb{N}_0, i,j \in E}$ (skip-free upwards in levels) can be attributed to the discrete time case by either use of uniformization if $\sup\{|q(r,i);(r,i)| : r \in \mathbb{N}_0, i \in E\} < \infty$, or inspection of the appertaining discrete time Markov chain embedded at transition epochs. This leads to completely analogous formulas: Equation (5), providing Equation (6), has to be replaced by $P_{0;n+1}(t) = \int_0^t (P_{0;n}(t-\tau) \cdot P_{n;n+1}) \cdot {}_n P_{n+1;n+1}(\tau) d\tau$. Instead of z -transforms the Laplace–Stieltjes transforms apply, that is, $X_{n+1}(s) = \int_0^{\infty} e^{-st} P_{0;n+1}(t) dt$ for $Re(s) \geq 0$, resulting in $X_{n+1}(s) = X_n(s) R(s)$ for $n \geq 1$ where $R(s) = P_{n;n+1} \int_0^{\infty} e^{-st} {}_n P_{n+1;n+1}(t) dt$ and $sR(s) = \sum_{k=0}^{\infty} R(s)^k A_k$. The rate matrix

$R = R(0)$ for a continuous time process has to be computed from

$$\sum_{k=0}^{\infty} R^k A_k = O. \quad (8)$$

Examples are queueing systems with bulk services in a random environment [14].

ERGODICITY CONDITIONS

Let $X = \{(N_v, J_v) : v \in \mathbb{N}_0\}$ be the considered irreducible aperiodic Markov chain with state space $\mathbb{N}_0 \times \{1, \dots, m\}$ whose transition matrix $\mathbf{P}$ in block form possesses the skip-free upward property and is spatially homogeneous as in Equation (2). Assume that $m < \infty$, and that the matrix $A = \sum_{k=0}^{\infty} A_k$ is irreducible and stochastic such that it can be interpreted as the transition matrix of some irreducible finite state Markov chain. Let $\psi = (\psi_1, \dots, \psi_m)$ with $\psi_j > 0$ satisfy $\psi A = \psi$ and $\psi e^T = 1$, and define the vector β by $\beta = \sum_{k=1}^{\infty} k A_k e^T$. Then the quantity $\vartheta_X = 1 - \psi \beta^T$ is called *average drift of the chain X*. The following statements hold [8]:

$$\begin{aligned} \vartheta_X > 0 &\implies X \text{ is transient,} \\ \vartheta_X = 0 &\implies X \text{ is null recurrent,} \\ \vartheta_X < 0 &\implies X \text{ is ergodic.} \end{aligned} \quad (9)$$

If $Y = \{(N_t, J_t) : t \geq 0\}$ is a continuous time Markov process on the state space $\mathbb{N}_0 \times \{1, \dots, m\}$, m finite, whose generator $\mathbf{Q}$ in block form is a lower block Hessenberg matrix, the *average drift* of this process is defined as $\vartheta_Y = -\psi \beta^T$, where again $A = \sum_{k=0}^{\infty} A_k$ is assumed to be irreducible and stochastic, and the positive vector ψ satisfies $\psi A = \mathbf{o}$ and $\psi e^T = 1$. In this case again Equation (9), with X replaced by Y , characterizes the ergodicity behavior of the process. Equation (9) reflects the customary conditions for stability of queueing systems: Note that $-\psi \beta^T = \psi (A_0 e^T - \sum_{k=2}^{\infty} (k-1) A_k e^T)$, and consider, for example, the $M/M/s$ queueing system, in which case $m = 1$, $A_0 = \lambda$ (the arrival rate), $A_2 = s\mu$ (the total service rate), and $A_k = 0$ for $k > 2$; then, since $\psi > \mathbf{0}$, $\vartheta_Y < 0$ means $\lambda < s\mu$, the familiar stability condition.

COMPUTATIONAL ASPECTS

In order to solve Equation (1) for π_{n+s} , one has to compute the solution R of Equation (7) in the discrete time case, or rather of Equation (8) in the continuous time case. We concentrate on the discrete time case. Once R has been determined, the computation of the boundary probabilities (assuming $s = 2$) requires to solve

$$\begin{aligned} (\pi_0, \pi_1) \begin{bmatrix} B_{00} & B_{01} \\ \sum_{k=1}^{\infty} R^{k-1} B_{k0} & \sum_{k=1}^{\infty} R^{k-1} B_{k1} \end{bmatrix} \\ = (\pi_0, \pi_1) \end{aligned} \quad (10)$$

subject to $\pi e^T = \pi_0 e_{m_0}^T + \pi_1 \sum_{n=0}^{\infty} R^n e_m^T = \pi_0 e_{m_0}^T + \pi_1 (I_m - R)^{-1} e_m^T = 1$ (the normalization condition).

Simple iteration schemes for solving Equation (7) have first been reported in Ref. 5. Essentially, they are of the form $R^{(0)} = 0$; $R^{(v+1)} = \sum_{k=0}^{\infty} R^{(v)k} A_k$ or, slightly modified, based on the two equations $R = A_0(I_m - U)^{-1}$ and $U = \sum_{k=0}^{\infty} R^k A_{k+1}$, yielding $R^{(0)} = 0$; $R^{(v+1)} = A_0(I_m - A_1)^{-1} + \sum_{k=2}^{\infty} R^{(v)k} A_k(I_m - A_1)^{-1}$ for $v \geq 0$. Note that $I_m - A_1$ is nonsingular since A_1 is substochastic. Matrices R and U are the minimal nonnegative solutions of the above equations (see Theorem 3.4 and Corollary 3.3 in Ref. 8), and if $A = \sum_{k=0}^{\infty} A_k$ is irreducible (as is the case in most applications) the iterates $R^{(v)}$ form a monotonically nondecreasing sequence converging linearly to R [5]. Entry U_{ij} in U is the taboo probability for eventually returning to level $\mathbf{n}$ in (n, j) after start in (n, i) under taboo of lower levels. The speed of convergence of such schemes can become extremely slow as has been demonstrated by Ramaswami [15] and Daigle and Lucantoni [16].

For QBDs, assuming $m < \infty$, let G_{ij} denote the probability that starting from $(n+1, i)$ the chain reaches level $\mathbf{n}$ for the first time in (n, j) . Then, given that the QBD is recurrent and $A = \sum_{k=0}^{\infty} A_k$ is irreducible, G is the minimal nonnegative *stochastic* solution of $G = A_2 + A_1 G + A_0 G^2$. This fact, together with $U = A_1 + R A_2$, gives rise to a stable recursive procedure for calculating G , starting

with $G^{(0)} = I_m$, which yields $U = A_1 + A_0G$ and $R = A_0(I_m - U)^{-1}$ and is considered the most efficient *linear* algorithm for QBDs [1] (complexity is $\approx \frac{14}{3}K \cdot m^3$, K the number of performed iterations). The iterates $U^{(v)}$ and $G^{(v)}$ are substochastic and stochastic, respectively. $G^{(v)}$ is the taboo probability matrix for the events that, starting from some level $\mathbf{n}$, the chain eventually visits level $\mathbf{n} - 1$ under taboo of levels $\mathbf{n} + v$ and above, where at each new iteration the chain can visit one higher level. Based on that, a significant improvement resulted from Ramaswami's idea to compute R^2 by considering the embedded QBD on levels 2^k ($k = 0, 1, \dots$) and then determining R from a linear equation. This led to the quadratically convergent and numerically stable *logarithmic reduction algorithm* (LRA) for QBDs for computing the matrix G , published by Latouche and Ramaswami in 1993 [17] (also compare Ref. 1 or 18). The number of iterations necessary for attaining the same accuracy for G as in the linear algorithm with K steps is $\lfloor \log_2 K \rfloor$.

$G = A_2 + A_1G + A_0G^2$ is the “QBD-version” of the matrix equation appertaining to the $M/G/1$ type counterpart of the considered $GI/M/1$ type model. QBDs represent the bridge between both canonical types. Let C_k denote the block entries of the spatially homogeneous part in an $M/G/1$ type characteristic matrix of upper block Hessenberg form ($k \in \mathbb{N}_0$), such that $G = \sum_{k=0}^{\infty} C_k G^k$. As for QBDs, G (the *fundamental period matrix*) is the matrix of probabilities G_{ij} that, starting in (n, i) , the process eventually reaches level $\mathbf{n} - 1$ in state $(n - 1, j)$ avoiding levels below $\mathbf{n}$ in intermediate steps. Whereas R , in the $GI/M/1$ type case, neither is substochastic nor stochastic, G is stochastic and, equivalently, has a spectral radius equal to 1 provided that $C = \sum_{k=0}^{\infty} C_k$ is irreducible and the chain is recurrent [19,20]. Exploiting these properties, Bini and Meini [21] developed a quadratically convergent and numerical stable algorithm for the computation of G by extending a technique of successive state reduction—known as *cyclic reduction* or *even-odd reduction* in linear algebra [22]—to infinite block matrices of $M/G/1$ type (for a convincing description of the copiousness of this technique the reader

should consult Ref. 23). The algorithm is similar to the LRA when restricted to QBDs, and has been advanced subsequently in several steps [24–26]. Its significance for matrix-geometric methods lies in the fact that it can as well be utilized to compute R due to an important duality result of Ramaswami [27]. Let ${}_n P_{n;n+r}^{(k)}(t)$ denote the k -step probability matrix for transitions from level $\mathbf{n}$ to level $\mathbf{n} + \mathbf{r}$ in $(0, t]$ under taboo of levels $\leq \mathbf{n}$, and ${}_n P_{n-r;n}^{(k)}(t)$ the corresponding matrix regarding transitions from level $\mathbf{n}$ to level $\mathbf{n} - \mathbf{r}$ for the $M/G/1$ paradigm (under taboo of levels $\geq \mathbf{n}$, $n \geq r$ in the spatially homogeneous part). Then the following double transforms are defined for $|z| \leq 1$, $Re(s) \geq 0$: $R^{[r]}(z, s) = \sum_{k=0}^{\infty} \int_0^{\infty} z^k e^{-st} {}_n P_{n;n+r}^{(k)}(dt)$, $G^{[r]}(z, s) = \sum_{k=0}^{\infty} \int_0^{\infty} z^k e^{-st} {}_n P_{n-r;n}^{(k)}(dt)$. With $R^{[r]}(z, s) = [R(z, s)]^r$, $G^{[r]}(z, s) = [G(z, s)]^r$, and $R^{[1]}(z, s) = R(z, s)$, $G^{[1]}(z, s) = G(z, s)$, these transform expressions satisfy the nonlinear matrix equations $R(z, s) = z \sum_{k=0}^{\infty} C_k^*(s) [G(z, s)]^k$ [5,20,28]. The $A_k^*(s)$ and $C_k^*(s)$ are the respective Laplace–Stieltjes transforms of the kernel entries $A_k(t)$ and $C_k(t)$ that belong to the Markov renewal processes (MRPs) with interrenewal times equal to interarrival times in the $GI/M/1$ type model, and inter-service-completion times in the $M/G/1$ type model. $A_k(\infty)$ equals A_k in Equation (2), and $C_k(\infty) = C_k$ in its upper Hessenberg counterpart. Using these constructs, Ramaswami [27] proved the following theorem: Assume that $A = \sum_{k=0}^{\infty} A_k$ is irreducible, $\psi = (\psi_1, \dots, \psi_m)$ is its positive invariant vector, and Δ is the diagonal matrix with diagonal entries $\psi_1, \dots, \psi_m$. Then, setting $C_k^*(s) = \Delta^{-1} [A_k^*(s)]^T \Delta$, the characteristic matrix of the time reversed dual $M/G/1$ type MRP has block entries $C_k = C_k^*(0)$ and is related to the considered matrix of the $GI/M/1$ type MRP by

$$R(z, s) = \Delta^{-1} [G(z, s)]^T \Delta.$$

$C = \sum_{k=0}^{\infty} C_k$ is irreducible. $R = R(1, 0)$ and $G = G(1, 0)$ yield $R = \Delta^{-1} G^T \Delta$. Consequently, any algorithm for evaluating G , in particular those developed by Bini, Meini and colleagues, provide solutions for R (as well as moments) with same efficiency.

FURTHER READING

Neuts, giving prominence to the role of structure, promoted algorithmic probability as “the study of stochastic models with a genuine added concern for algorithmic feasibility” [5,29]. In Ref. 30 Neuts provides an important survey. Textbook introductions can be found, for example, in Nelson [14] and Kant [10]. In Ref. 31, by enlarging the state space and decomposing transitions from (n, i) to $(n + k, j)$ into sequences of single step transitions in the larger space, Ramaswami showed the generality of QBDs comprising $GI/M/1$ -type processes (same techniques apply to $M/G/1$ type systems). An overview on applications of QBDs to computer performance analysis is given in Ref. 32. In Ref. 33, parallel-server scheduling models are analyzed, and in Ref. 34, the behavior of a synchronization queue with Markovian arrival process is modeled as a finite state QBD. Systems with continuously varying levels can be related to QBDs, too, as shown by Ramaswami [35]. He established a connection by introducing a stochastic coupling of a Markov modulated fluid flow model (MMFF) with a queue of QBD characteristics and developed, on this basis, quadratically convergent algorithms similar to those mentioned in the previous section. Subsequently, much progress has been achieved with respect to the steady-state and transient behavior of a wide class of MMFF's as documented in a series of papers by Ahn and Ramaswami [36–39]. Efficient algorithms for the computation of busy period transform expressions are presented in Ref. 39. A concise survey on these results is given in Ref. 40. Additionally, first passage times in fluid models have been analyzed by Ramaswami [41] via QBD techniques. The first generalization of $GI/M/1$ type models to systems with continuously varying level, but, was introduced in 1989 by Sengupta [42], who considered a bivariate Markov process $\{(N_t, J_t) : t \geq 0\}$ whose level component can increase linearly or can have downward jump discontinuities; stationary distributions assume a matrix-exponential form in this case. In Ref. 43, Sengupta showed for the discrete time

process $\{(N_k, J_k) : k \in \mathbb{N}_0\}$ that the marginal steady-state distribution of N_k is discrete phase type. In Ref. 44, considering queues with (in case dependent) semi-Markov arrival and service processes, Sengupta obtained the matrix-exponential form for waiting times and proved that queue length distributions are matrix-geometric if arrival and service processes are independent. Generalizations of the $GI/M/1$ paradigm are handled in Ref. 45, where the level variable can take values in a tree structure, and in Ref. 46, where the phase variables are defined over a general measure space. In Ref. 47, Latouche considers invariant measures of matrix-geometric form. Multi-server retrial queues, modeled as continuous time $GI/M/1$ -type processes, were analyzed by Artalejo *et al.* [48]. An extensive survey on iterative methods for the computation of rate matrices is given in Ref. 49. The textbook [50] deals with numerical methods for the analysis of structured Markov chains as developed by Bini, Meini, and colleagues (compare also the first textbook on numerical evaluation of Markov chains published by Stewart [18]). In Ref. 51, a software package based on the above mentioned methods is provided. Recently, Bini *et al.* contributed careful comparisons between cyclic and logarithmic reduction [52].

REFERENCES

1. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. Philadelphia (PA): SIAM; 1999.
2. Marcus M, Minc H. A survey of matrix theory and matrix inequalities. New York: Dover Publications, Inc.; 1992.
3. Wallace V. The solution of quasi birth and death processes arising from multiple access computer systems [PhD thesis]. (Tech. Rept. No. 07742-6-T), Systems Engineering Laboratory, University of Michigan, 1969.
4. Evans RV. Geometric distribution in some two-dimensional queueing systems. *Oper Res* 1967;15:830–846.
5. Neuts MF. Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore (MD): The John Hopkins University Press; 1981. Also published by New York: Dover Publications; 1994.

6. Neuts MF. Markov chains with applications in queueing theory, which have a matrix-geometric invariant vector. *Adv Appl Probab* 1978;15:707–714.
7. Neuts MF. The probabilistic significance of the rate matrix in matrix-geometric invariant vectors. *J Appl Probab* 1980;17:291–296.
8. Ramaswami V. Matrix analytic methods: a tutorial overview with some extensions and new results. In: Chakravarthi SR, Alfa AS, editors. *Matrix analytic methods in stochastic models*. New York; Basel; Hong Kong: Marcel Dekker, Inc.; 1997. pp. 261–296.
9. Çinlar E. *Introduction to stochastic processes*. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1975.
10. Kant K, Srinivasan MM. *Introduction to computer system performance evaluation*. New York: McGraw-Hill, Inc.; 1992.
11. Taussky O. A remark concerning the roots of the finite segments of the Hilbert matrix. *Q J Math Oxf Ser* 1949;20(2):80–85.
12. Householder AS. *The theory of matrices in numerical analysis*. New York: Dover Publications, Inc.; 1964.
13. Seneta E. *Non-negative matrices and Markov chains*. New York: Springer; 1981.
14. Nelson R. *Probability, stochastic processes, and queueing theory*. New York, Berlin, Heidelberg, London, Paris: Springer; 1995.
15. Ramaswami V. Nonlinear matrix equations in applied probability—solution techniques and open problems. *SIAM Rev* 1988;30(2):256–263.
16. Daigle JN, Lucantoni DM. Queueing systems having phase dependent arrival and service rates. In: Stewart WJ, editor. *Numerical solutions of Markov chains*. New York: Marcel Dekker, Inc.; 1991. pp. 161–202.
17. Latouche G, Ramaswami V. A logarithmic reduction algorithm for quasi-birth-and-death processes. *J Appl Probab* 1993;30:650–674.
18. Stewart WJ. *Introduction to the numerical solution of Markov chains*. New Jersey: Princeton University Press; 1994.
19. Neuts MF. The fundamental period of the queue with Markov modulated arrivals. *Probability, statistics, and mathematics: papers in honor of Samuel Karlin*. New York: Academic Press; 1989. pp. 187–200.
20. Neuts MF. *Structured stochastic matrices of M/G/1 type and their applications*. New York: Marcel Dekker; 1989.
21. Bini D, Meini B. On the cyclic reduction applied to a class of Toeplitz-like matrices arising in queueing problems. *Proceedings 2nd International Workshop on Numerical Solution of Markov Chains*; 1995; Raleigh (NC). pp. 21–38.
22. Buzbee BL, Golub GH, Nielson CW. On direct methods for solving Poisson's equation. *SIAM J Numer Anal* 1970;7:627–656.
23. Bini D, Meini B. The cyclic reduction algorithm: from Poisson equation to stochastic processes and beyond. *Numer Algorithms*. Springer; 2009;51(1):23–60. Special issue dedicated to Gene Golub.
24. Bini D, Meini B. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J Matrix Anal Appl* 1996;17:906–926.
25. Bini D, Meini B. Improved cyclic reduction for solving queueing problems. *Numer Algorithms* 1997;15:57–74.
26. Meini B. An improved FFT-based version of Ramaswami's formula. *Stoch Models* 1997;13(2):223–238.
27. Ramaswami V. A duality theorem for the matrix paradigms in queueing theory. *Stoch Models* 1990;6(1):151–161.
28. Ramaswami V. The busy period of queues which have a matrix-geometric steady state probability vector. *Oper Res* 1982; 19:238–261.
29. Neuts MF. Some promising directions in algorithmic probability. In: Alfa AS, Chakravarthi SR, editors. *Advances in Matrix Analytic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 1998. pp. 429–434.
30. Neuts MF. *Matrix-analytic methods in queueing theory*. In: Dshalalow JH, editor. *Advances in Queueing*. Boca Raton (NY); London; Tokyo: CRC Press; 1995. pp. 265–292.
31. Ramaswami V. The generality of quasi-Birth-and-Death processes. In: Alfa AS, Chakravarthi SR, editors. *Advances in Matrix Analytic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 1998. pp. 93–113.
32. Squillante MS. Matrix-analytic methods: applications, results and software tools. In: Latouche G, Taylor P, editors. *Advances in Algorithmic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 2000. pp. 351–374.
33. Squillante MS. Matrix-analytic methods in stochastic parallel-server scheduling models. In: Alfa AS, Chakravarthi SR, editors. *Advances in Matrix Analytic Methods for*

- Stochastic Models. New Jersey: Notable Publications, Inc.; 1998. pp. 311–340.
34. Takahashi M, Takahashi Y. Synchronization queue with two inputs and finite buffers. In: Latouche G, Taylor P, editors. *Advances in Algorithmic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 2000. pp. 375–390.
 35. Ramaswami V. Matrix analytic methods for stochastic fluid flows. In: Key P, Smith D, editors. *Teletraffic Engineering in a Competitive World, ITC - 16 Proceedings*. New York: Elsevier; 1999. pp. 1019–1030.
 36. Ahn S, Jeon J, Ramaswami V. Steady state analysis of finite fluid flow models using finite QBD's. *Queueing Syst* 2005;49:223–259.
 37. Ahn S, Ramaswami V. Fluid flow models and queues—a connection by stochastic coupling. *Stoch Models* 2003;19(3):325–348.
 38. Ahn S, Ramaswami V. Transient analysis of fluid flow models via stochastic coupling. *Stoch Models* 2004;20(1):71–102.
 39. Ahn S, Ramaswami V. Efficient algorithms for transient analysis of stochastic fluid flow models. *J Appl Probab* 2005;42(2):531–549.
 40. Ahn S, Ramaswami V. Matrix-geometric algorithms for stochastic fluid flows. Volume 180, *SMCtools'06, ACM International Conference Proceedings Series*. Pisa: ACM; 2006.
 41. Ramaswami V. Passage times in fluid models with applications to risk processes. *Methodol Comput Appl Probab* 2006;8(4):497–515.
 42. Sengupta B. Markov processes whose steady state distribution is matrix-exponential with an application to the $GI/PH/1$ queue. *Adv Appl Probab* 1989;21:159–180.
 43. Sengupta B. Phase-type representations for matrix-geometric solutions. *Stoch Models* 1990;6(1):163–167.
 44. Sengupta B. The semi-Markovian queue: theory and applications. *Stoch Models* 1990; 6(3):383–413.
 45. Sengupta B, Takine T, Yeung RW. Matrix Analytic Methods for Markov chains on a tree structure. In: Chakravarthi SR, Alfa AS, editors. *Matrix Analytic Methods in Stochastic Models*. New York; Basel; Hong Kong: Marcel Dekker, Inc.; 1997. pp. 296–311.
 46. Tweedie RL. Operator-geometric stationary distributions for Markov chains, with applications to queueing models. *Adv Appl Probab* 1982;14:368–391.
 47. Latouche G. Quasi-Birth-and-Death processes: Beyond stable queues. In: Alfa AS, Chakravarthi SR, editors. *Advances in Matrix Analytic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 1998. pp. 79–92.
 48. Artalejo JR, Gomez-Corral A, Neuts MF. Numerical analysis of multiserver retrieval queues under a full access policy. In: Latouche G, Taylor P, editors. *Advances in Algorithmic Methods for Stochastic Models*. New Jersey: Notable Publications, Inc.; 2000. pp. 1–19.
 49. Latouche G. Algorithms for infinite of Markov chains with repeating columns. In: Meyer CD, Plemmons RJ, editors. *Linear Algebra, Markov Chains and Queueing Models*. New York: Springer-Verlag; 1993. pp. 231–265.
 50. Bini D, Latouche G, Meini B. *Numerical methods for structured Markov-chains*. Oxford: Oxford University Press; 2005.
 51. Bini D, Meini B, Steffe S, *et al.* Structured Markov chain solver: tool extension. *Workshop SMCtools*. Pisa: ACM; 2009.
 52. Bini D, Meini B, Ramaswami V. A probabilistic interpretation of cyclic reduction and its relationships with logarithmic reduction. *CALCOLO* 2008;45(3):207–216.

MAXIMUM CLIQUE, MAXIMUM INDEPENDENT SET, AND GRAPH COLORING PROBLEMS

JEFF PATTILLO

Department of Mathematics, Texas A&M University, College Station, Texas

SERGIY BUTENKO

Department of Industrial and Systems Engineering, Texas A&M University, College Station, Texas

The maximum clique, maximum independent set, graph coloring, and minimum clique partitioning problems are classical problems in combinatorial optimization. Owing to their important role in several theoretical fields and applicability in a wide variety of practical settings, these problems have been extensively studied from different perspectives by mathematicians, computer scientists, operations researchers, engineers, biologists, and social scientists. This article surveys only a fraction of these research developments with the aim to provide a brief introduction to these problems for the operations research community. The reader is referred to surveys [1,2] for far more extensive information about each problem.

According to WordNet dictionary [3], clique is defined as “an exclusive circle of people with a common purpose.” Cliques, as described in the dictionary definition, represent a natural object of interest for social and behavioral sciences. Thus, it is not surprising that the first mentioning of this term in graph-theoretic context is attributed to researchers in social network analysis: in their 1949 paper [4], Luce and Perry used complete subgraphs to model social cliques, which are defined as groups of people that know (are friends of) all other people in the group. It should be noted that the complete subgraphs had been studied by mathematicians even before the term *clique* was introduced to graph theory.

Graph coloring problems have been around even longer, dating back to the

famous four-color conjecture first attacked in the nineteenth century [5,6], stating that regions of any simple planar map can be colored with four colors so that any two adjacent regions have different colors. The proof of this seemingly simple statement required the efforts of several generations of mathematicians and involvement of a computer [7,8].

Clique-like structures frequently arise in many other applications, where one is interested in detecting large groups of elements that are all closely related to each other in some sense. Depending on the application of interest, such structures are often referred to as *clusters*, *modules*, *complexes*, or *cohesive subgroups*. If the elements in the application of interest are represented as vertices (nodes) and the relationships between the elements are represented as edges (links, arcs), then clusters can be naturally modeled as cliques in this graph-theoretic representation. The specific application areas of the maximum clique and the graph coloring problems include information retrieval, signal transmission analysis, classification theory, economics, scheduling, experimental design, and computer vision [1,9–14]. The reader is referred to Butenko and Wilhelm [15] for a survey of applications of clique detection models in biochemistry and genomics. Applications of cliques and coloring in telecommunications are discussed in Balasundaram and Butenko [16].

In this article, we seek to give an introduction to these problems, their basic properties, as well as a brief summary of the most popular methods of solving the problems and the capabilities and limitations of these methods. We also describe some results concerning the maximum clique and the graph coloring problems in uniform random graphs (See also ***Network Theory: Concepts and Applications, Domination Problems, Clique Relaxations***).

DEFINITIONS AND NOTATIONS

Given a simple undirected graph $G = (V, E)$, where $V = \{1, \dots, n\}$ is the vertex set and $E \subseteq V \times V$ is its edge set, we denote by $G(S) = (S, E \cap S \times S)$, the subgraph induced by a subset of vertices S , and we use the notation $\bar{G}$ to represent the complement graph, that is $\bar{G} = (V, \bar{E})$ where $\bar{E} = \{(i, j) | i, j \in V, i \neq j \text{ and } (i, j) \notin E\}$. Vertices i and j are called *adjacent* if $(i, j) \in E$. The neighborhood $N(i)$ of a vertex i is the set of vertices adjacent to i , $N(i) = \{j \in V : (i, j) \in E\}$. By A_G we denote the *adjacency matrix* of G , which is an $n \times n$ matrix $A_G = [a_{ij}]_{i,j=1}^n$ such that $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ otherwise. The degree $\deg(i)$ of i in G is the size $|N(i)|$ of the neighborhood of i in G . We will denote by $\delta(G)$ and $\Delta(G)$ the smallest and the largest degree of a vertex in G , respectively.

A graph $G = (V, E)$ is called *complete* if all its vertices are adjacent, that is, if $\forall i, j \in V, i \neq j$, we have $(i, j) \in E$. By definition, a *clique* C is a subset of vertices such that $G(C)$ is complete. An *independent set* in G is a subset of vertices I such that the induced subgraph $G(I)$ is edgeless. A clique (independent set) is called *maximal* if it is not a subset of a larger clique (independent set) in G , and it is called *maximum* if there is no larger clique (independent set) in the graph. Given a simple undirected graph G , the *maximum clique problem* (*maximum independent set problem*) asks for a maximum clique (independent set) in G . The *clique number*, which is traditionally denoted as $\omega(G)$, is the number of vertices in the largest clique in G . The independence number $\alpha(G)$ is defined analogously. It is easy to see that C is a clique in G if and only if C is an independent set in $\bar{G}$; hence, $\omega(G) = \alpha(\bar{G})$. Owing to this close relation between the two concepts, we will use the notions of cliques and independent sets interchangeably in this article, since all the facts stated for a clique in G are true for an independent set in $\bar{G}$, and vice versa.

A *proper k -coloring* of the vertices of G is an assignment of colors (e.g., $\{1, 2, \dots, k\}$) to the vertices of G so that no two adjacent vertices get the same color. A vertex k -coloring can also be thought of as a mapping $f : V \rightarrow \{1, 2, \dots, k\}$ such that $f(v_i) \neq f(v_j)$ if

$(v_i, v_j) \in E$. A graph G is called *k -colorable* if there exists a proper k -coloring of its vertices. The *graph coloring problem* is to find a proper coloring of its vertices with the smallest possible number of colors. The *chromatic number* $\chi(G)$ of G is the smallest integer k for which G is k -colorable.

A k -coloring partitions V into k different *color classes*, where each member of the class has the same color. To have the same color, the members of each class must be pairwise nonadjacent, which by definition makes them an independent set. Similarly, one can define a *k -clique partitioning* of G , which is a partitioning of the set of vertices of G into k disjoint subsets $C_1, \dots, C_k$, each of which is a clique. The *minimum clique partitioning problem* asks for a k -clique partitioning of G using the smallest possible number k of cliques, which is denoted by $\bar{\chi}(G)$ and is called the *clique partitioning number*. Note that a proper k -coloring of G is a k -clique partitioning of $\bar{G}$, thus, $\chi(\bar{G}) = \bar{\chi}(G)$.

The maximum clique, maximum independent set, vertex coloring, and minimum clique partitioning problems are all known to be NP-hard [17]. Moreover, these problems are known to be hard to approximate. Arora *et al.* [18,19] showed that the maximum clique size cannot be approximated within a factor of n^ϵ for any $\epsilon > 0$ unless $P = NP$. Similarly, the graph coloring problem is not approximable within $n^{1-\epsilon}$ for any $\epsilon > 0$ [20].

MATHEMATICAL PROGRAMMING FORMULATIONS

The Maximum Clique Problem

The maximum clique problem can be formulated as the following integer program:

$$\begin{aligned} \max \quad & \sum_{i=1}^n x_i \\ \text{s.t.} \quad & x_i + x_j \leq 1, \quad \forall (i, j) \in \bar{E} \\ & x_i \in \{0, 1\}, \quad i = 1, \dots, n. \end{aligned} \quad (1)$$

In this *edge* formulation any feasible solution x defines a clique C in G as follows: $C = \{i \in V : x_i = 1\}$. The linear relaxation of this formulation is significant as well. Nemhauser

and Trotter [21] found that if a variable x_i in any optimal solution to the linear relaxation had the value 1, then $x_i = 1$ in at least one optimal solution to the above formulation. When some such variable $x_i = 1$, this clearly helps cut down the search space for an optimal clique.

Shor [22] modified the edge formulation to obtain an equivalent nonlinear problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^n x_i \\ \text{s.t.} \quad & x_i x_j = 0, \quad \forall (i, j) \in \bar{E} \\ & x_i^2 - x_i = 0, \quad i = 1, \dots, n. \end{aligned} \quad (2)$$

He used this formulation to obtain dual estimates of good quality.

Motzkin and Straus [23] related $\omega(G)$ to the global optimal value of a certain quadratic function over the standard simplex. Before we state this classical formulation, some additional notations are in order. For a set of vertices C we use the notation x^C to denote the *characteristic vector* of C , which is a vector that satisfies $x_i^C = 1/|C|$ if $i \in C$ and $x_i^C = 0$, otherwise, for $i \in \{1, \dots, n\}$. By S we will denote the standard simplex in $\mathbb{R}^n$:

$$S = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x_i \geq 0, i = 1, \dots, n \right\}.$$

Let A_G be the adjacency matrix of a graph G , and for $x \in \mathbb{R}^n$ let $g(x) = x^T A_G x$. Then the global optimal solution x^* to $\max_{x \in S} g(x)$ is related to the clique number by the formula

$$\omega(G) = \frac{1}{1 - g(x^*)}.$$

The following generalization of the Motzkin–Straus theorem was introduced by Bomze [24] to guarantee a one-to-one correspondence between maximal cliques and local maximizers of the quadratic program. Let $f(x) = x^T A_G x + \frac{1}{2} x^T x$.

Theorem 1. *Let C be a subset of vertices of a graph G , and let x^C be its characteristic vector. Then the following statements hold:*

- *C is a maximum clique of G if and only if x^S is a global maximizer of the function f over the simplex S . In this case, $\omega(G) = \frac{1}{2(1-f(x^C))}$.*
- *C is a maximal clique of G if and only if x^C is a local maximizer of f over S .*
- *All local (and hence global) maximizers x of g over S are strict, and of the form $x = x^C$ for some $C \subseteq V$.*

The maximum independent set problem can be equivalently formulated using the following mathematical programs over the unit hypercube in $\mathbb{R}^n$ [25,26]:

$$\alpha(G) = \max_{x \in [0,1]^n} \sum_{i \in V} x_i - \sum_{(i,j) \in E} x_i x_j. \quad (3)$$

$$\alpha(G) = \max_{x \in [0,1]^n} \sum_{i \in V} x_i \prod_{j \in N(i)} (1 - x_j). \quad (4)$$

$$\alpha(G) = \max_{x \in [0,1]^n} \sum_{i \in V} \frac{x_i}{1 + \sum_{j \in N(i)} x_j}. \quad (5)$$

In each of these formulations, the feasible region can be replaced with $\{0, 1\}^n$ without changing the optimal objective value. Several other similar formulations can be found in Harant [27].

Recently, Martins [28] proposed discretized formulations for the maximum clique problem and reported tighter upper bounds based on linear programming relaxations than those for the known formulations on many benchmark graphs.

Formulations of the Graph Coloring Problem

The following is perhaps the most straightforward integer programming formulation of the graph coloring problem. Let r be an upper bound on the chromatic number of G . We have:

$$\begin{aligned} \chi(G) = \min \quad & \sum_{k=1}^r y_k \\ \text{s.t.} \quad & \sum_{k=1}^r x_{ik} = 1, \quad \forall i \in V \end{aligned}$$

$$\begin{aligned}
x_{ik} + x_{jk} &\leq 1, \quad \forall (i, j) \in E, \\
k &= 1, \dots, r \\
y_k &\geq x_{ik}, \quad \forall i \in V, \quad k = 1, \dots, r \\
y_k, x_{ik} &\in \{0, 1\} \quad \forall i \in V, \\
k &= 1, \dots, r.
\end{aligned} \tag{6}$$

In this formulation, the variable y_k indicates whether the k th from the r available colors is used, and $x_{ik} = 1$ if and only if the k th color is used for vertex i . The first set of constraints ensures that exactly one color is used for each vertex, and the remaining two sets of constraints guarantee that the coloring is proper and that all used colors are counted. Thus, any feasible solution induces a proper coloring by assigning each vertex i the unique color k for which $x_{ik} = 1$. The optimal objective function value gives the chromatic number of the graph. Note that the formulation remains valid if the variables $y_k, k = 1, \dots, r$ are not required to be integers.

The following integer programming formulation has the optimal objective equal to zero if and only if the graph has a proper r -coloring:

$$\begin{aligned}
\min \quad & \sum_{k=1}^r \sum_{(i,j) \in E} x_{ik} x_{jk} \\
\text{s.t.} \quad & \sum_{k=1}^r x_{jk} = 1, \quad j = 1, \dots, n \\
& x_{jk} \in \{0, 1\}, \quad j = 1, \dots, n, \quad k = 1, \dots, r.
\end{aligned} \tag{7}$$

This formulation has a quadratic objective function, but its constraint set is much simpler than in the previous formulation.

Let $\mathcal{I}$ be the set of all maximal by inclusion independent sets of $G = (V, E)$. For each $I \in \mathcal{I}$, let a binary variable x_I be such that $x_I = 1$, if and only if all vertices in I use the same color in a proper coloring of G . Then we have the following *set covering* formulation of the graph coloring problem [29]:

$$\chi(G) = \min \sum_{I \in \mathcal{I}} x_I$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{I: i \in I} x_I \geq 1 \quad \forall i \in V \\
& x_I \in \{0, 1\}, \quad I \in \mathcal{I}.
\end{aligned} \tag{8}$$

BOUNDS

Since the number of vertices in any color class defined by a proper coloring of G cannot exceed $\alpha(G)$, we obtain the following inequality:

$$\chi(G) \geq \frac{|V|}{\alpha(G)}. \tag{9}$$

The Caro–Wei bound on the independence number is expressed in terms of vertex degrees as follows [30,31]:

$$\alpha(G) \geq \sum_{i \in V} \frac{1}{\deg(i) + 1}. \tag{10}$$

This bound is sharp if and only if each connected component of G is a clique. Note that the Caro–Wei bound follows from formulation (5) by setting all variables to 1.

In 1967, Wilf [32] showed that

$$\omega(G) \leq \rho(A_G) + 1, \tag{11}$$

where $\rho(A_G)$ is the spectral radius of the adjacency matrix of G (the largest eigenvalue of A_G).

Amin and Hakimi [33] proved that

$$\omega(G) \leq N_{-1} + 1 < n - N_0 + 1, \tag{12}$$

where N_{-1} is the number of eigenvalues of A_G that do not exceed -1 , and N_0 is the number of zero eigenvalues. This bound is sharp if G is a complete multipartite graph.

Applying a greedy coloring yields a simple upper bound on the chromatic number:

$$\chi(G) \leq \Delta(G) + 1. \tag{13}$$

Brook’s theorem [34] slightly improves this bound for connected graphs: for a connected, simple graph G that is not a complete graph or an odd cycle, we have

$$\chi(G) \leq \Delta(G). \tag{14}$$

Szekeres and Wilf [35] generalized the bound 13 by proving that for any function $\lambda(G)$ satisfying two properties, (i) $G' \subset G \Rightarrow \lambda(G') \leq \lambda(G)$ and (ii) $\lambda(G) \geq \delta(G)$, we have

$$\chi(G) \leq \lambda(G) + 1.$$

Reed [36] conjectured that for every graph G ,

$$\chi(G) \leq \frac{1}{2}(\Delta(G) + \omega(G)) + 1. \quad (15)$$

The clique number $\omega(G)$ and chromatic number $\chi(G)$ satisfy the inequality $\omega(G) \leq \chi(G)$, since every vertex in the largest clique is adjacent to every other vertex in this clique, and thus requires a different color. What is remarkable is that there exists a polynomially computable graph function $\vartheta(G)$, known as the Lovász ϑ -function [37], that always satisfies the *sandwich theorem* [38]:

$$\omega(G) \leq \vartheta(\overline{G}) \leq \chi(G) \quad (16)$$

Hence, $\vartheta(\overline{G})$ provides a polynomially computable upper bound on $\omega(G)$ and lower bound on $\chi(G)$.

A graph G , such that for any subset of vertices $S \subseteq V$ the corresponding induced subgraph has equal clique and chromatic numbers, $\omega(G(S)) = \chi(G(S))$, is called *perfect*. Recently, Chudnovsky *et al.* [39] proved the strong perfect graph theorem (first conjectured by Claude Berge in 1960) stating that a graph is perfect if and only if it contains no odd hole (that is, a chordless cycle of length at least four) and no odd antihole (a complement of a hole). Hence, checking whether a graph G is perfect can be done in polynomial time. On the other hand, Busygin and Pasechnik [40] proved that recognizing whether $\chi(G) - \omega(G) = 0$ is *NP*-hard, implying that any polynomially computable parameter that lies between $\omega(G)$ and $\chi(G)$ will provide the “provably best” upper bound for the independence number in the sense that no other polynomially computable bound can improve this bound whenever it can be improved. In particular, the Lovász theta is one such bound.

COMPUTING CLIQUES AND COLORINGS

Exact Solutions

The early algorithms for the maximum clique problem were motivated by various applications and, because of the nature of the considered applications, aimed to enumerate all maximal cliques in the graph rather than just finding a single maximum clique. In 1957, Harary and Ross [41] published the first algorithm for enumerating all maximal cliques in a graph, which was motivated by an application in sociometry. Their method first reduces the problem on general graphs to a special case with graphs having at most three maximal cliques, and then solves the problem for this special case. Many other algorithms for enumerating all maximal cliques followed [42–45].

The progress in computing technologies in the 1960s motivated the development of many new enumerative algorithms. Bron and Kerbosch [46] proposed a backtracking method that requires only polynomial storage space and excludes the possibility of computing the same clique twice. Various versions of this algorithm are still used in many different applications, in particular, in computational biology, where it was found very effective for problems related to matching three-dimensional molecular structures [47]. Tomita *et al.* [48] developed a modification of this approach, which has the time complexity of $O(3^{n/3})$. This complexity is cannot be improved for enumerative algorithms, since there are graphs with $3^{n/3}$ maximal cliques [49]. More recently, Tomita *et al.* [50] presented a depth-first search algorithm for generating all maximal cliques of an undirected graph, which uses the same pruning rules as the Bron–Kerbosch algorithm, and also has $O(3^{n/3})$ worst-case time complexity. They report encouraging results of computational experiments.

Many exact algorithms for the maximum clique and related problems have been proposed in the literature, most of which can be viewed as variations of the branch-and-bound technique. Some of the algorithms were designed with worst-case time complexity in mind, while others emphasize the practical performance. In 1977, Tarjan

and Trojanowski [51] proposed a recursive algorithm for the maximum independent set problem with the time complexity of $O(2^{n/3})$. Later, Robson [52] improved this result to obtain the time complexity of $O(2^{0.276n})$. In 2001, Robson [53] further brought down the best-known worst-case complexity to $O(2^{n/4})$.

As for more practical algorithms, Balas and Yu [11] used an interesting implementation of the implicit enumeration, capable of computing maximum cliques in graphs with up to 400 vertices and 30,000 edges. Carraghan and Pardalos [54] proposed an alternative, simpler approach examining vertices in the order corresponding to the nondecreasing order of their degrees within the implicit enumeration framework. This method proved to be extremely effective, especially for sparse graphs and was selected as a benchmark for comparing different algorithms in The Second DIMACS Implementation Challenge [55]. Östergård [56] proposed a branch-and-bound algorithm that employs a vertex ordering based on approximate coloring. This algorithm showed an excellent performance on random graphs and DIMACS benchmark instances and is considered the state-of-the-art general-purpose exact algorithm for the maximum clique problem by many researchers. Tomita and Kameda [57] used similar ideas in their branch-and-bound algorithm, which also uses approximate coloring along with an appropriate sorting of vertices. They report superior computational results on graphs with up to 15,000 vertices, including instances arising in image processing, design of quantum circuits, and bioinformatics.

As we have seen in the previous section, the maximum clique and graph coloring problems can be formulated as integer or continuous nonconvex programs. Modern integer programming solvers typically combine branch-and-bound strategies with cutting-plane methods, efficient preprocessing schemes, including fast heuristics, and sophisticated decomposition techniques in order to find an exact solution, and may be quite effective even with their default settings when used with an appropriate integer programming formulation. However, the experience shows

that developing specialized strategies based on a detailed polyhedral study of the specific problem helps to speed up the computations and tackle larger problem instances. Several attempts of using algorithms based on the polyhedral properties of the maximum clique problem have been reported in the literature. For example, Bourjolly *et al.* [58] propose a column generation algorithm for the maximum stable set problem, and Rossi and Smriglio [59] reports encouraging computational results with a branch-and-cut algorithm for the same problem. While these approaches show some promise and may still be improved in the future, at the moment they are, generally speaking, outperformed by the above-mentioned branch-and-bound strategies in practice.

The exact algorithms for the graph coloring follow the same general line of development as for the maximum clique problem. Brown [60] proposed a branch-and-bound algorithm that uses a sequential greedy coloring heuristic (SCGH) described in the next section. Several improvements were suggested later [55,61,62]. Mehrotra and Trick [29] developed a column generation approach for graph coloring. Herrmann and Hertz [63] attempt to compute the chromatic number of a graph G by first determining a critical subgraph, which is the smallest induced subgraph of G that has the same chromatic number. They report very encouraging computational results. Lucet *et al.* [64] proposed yet another algorithm for the graph coloring problem based exactly on a linear decomposition of the graph. This method is exponential with respect to the linear width of the input graph, but only linear with respect to its number of vertices. The method was used to successfully solve some of the problem instances for the first time.

Approximation Algorithms and Heuristics

Owing to the inapproximability of the maximum clique and graph coloring problems mentioned above, the approximation algorithms for these problems, normally, have nonconstant performance ratio and are of mostly theoretical value. For example,

a greedy algorithm that builds a maximal independent set by recursively adding a minimum degree vertex and removing its neighbors has an approximation ratio of $(\Delta(G) + 2)/3$ [65]. Johnson [66] presented an algorithm for graph coloring with a performance guarantee of $O(n \log n)$.

Heuristics for the maximum clique and graph coloring problems usually have little theoretical justification or performance bounds. Thus evaluations of which heuristics tend to work best in which situations are mostly based on experimentation. The heuristics listed below are among the most popular and successful.

Greedy Construction Heuristics. Greedy heuristics for the maximum clique problem typically work in one of two ways: either adding in vertices into a partial clique until no more can be inserted or deleting vertices from a set of vertices that is not a clique until it induces a complete subgraph. The rule governing which vertex to add or delete next is usually based upon simple local information, such as the degree of the vertex. Heuristics such as this tend to run extremely fast. For examples of greedy heuristics, see Johnson [67], Kopf and Ruhe [68], and Tomita *et al.* [69].

The most commonly implemented greedy heuristic for the coloring problem is called the *sequential greedy coloring heuristic (SGCH)*. The first step of this algorithm is to specify some vertex ordering. After that each vertex is colored in the specified order in a greedy method, using the smallest feasible color. There will always be an ordering that will lead to an optimal coloring. To see this, suppose the color classes from an optimal coloring are $\{S_1, \dots, S_k\}$. If SGCH is applied to an ordering where all of S_1 is colored, then all of S_2 , then all of S_3 , and so on, it is easy to see that, while the same exact coloring may not be obtained, it will certainly not use more than k colors. Because of this fact, heuristics with greedy coloring tend to concentrate heavily on the order in which the vertices are colored.

Ordering schemes based on vertex degrees, such as largest first [70], smallest last [71], and dynamic ordering schemes that

do not preorder vertices [61, 72] have all been well studied. For research into how one can move gradually from an initial ordering to an optimal ordering, see Biggs [73], Chams *et al.* [74], and White [75].

Local Search Heuristics. The definition of the neighborhood is a critical element of a local search algorithm. The most common neighborhoods for the maximum clique and the graph coloring problems are the so-called *exchange neighborhoods*. An exchange neighborhood for the maximum clique problem is defined as follows: given a clique representing the current solution, its neighbors are obtained by removing one or more vertices from the clique and adding at least as many new vertices to the clique whenever possible. For graph coloring, the exchange is typically performed between different color classes. Namely, given a proper coloring, its neighbors are obtained by moving a certain number of vertices from one color class to another in an attempt to eventually reduce some of the color classes to empty set, thus obtaining a better coloring. Oftentimes, exchanges resulting in infeasible solutions are allowed with a penalty added to the objective for the infeasibility. This is done to diversify the search to increase the chances of identifying a high quality solution.

In Culberson [76], an alternative iterative improvement heuristic for the coloring problem is proposed. The paper notes that if you take a coloring $\{S_1, \dots, S_r\}$, where the S_i 's are the color classes, then coloring an ordering of the form $(S_{i_1}, \dots, S_{i_r})$ by a greedy heuristic, so that each color class is colored entirely before the next class, can only improve upon the previous coloring. This observation is independent of the ordering of the vertices within each class, as well as the order in which the classes appear, so in Culberson [76] various ordering schemes of the classes are explored. Typically a mixture of the schemes is best because it allows the algorithm to search a "larger" neighborhood of potential solutions.

While traditional local search methods are typically capable of improving on the solutions generated by construction heuristics, they are known for their tendency

to often getting trapped in local optima of rather poor quality. Several strategies commonly used to overcome this drawback are briefly reviewed in the next section. Various other heuristics can be found in Johnson and Trick [55] and Caramia and Dell’Olmo [77].

Metaheuristics. While many metaheuristic strategies, such as neural networks, genetic algorithms, and evolutionary computing have been applied to the maximum clique and graph coloring problems, few offer very good results [1]. Both problems do, however, surrender good results to the same two metaheuristic methods: simulated annealing and tabu search.

For different implementations of simulated annealing to solve the maximum clique problem, see Aarts and Korst [78] and Homer and Peinado [79]. It is interesting to note that it has been shown in Jerrum [80] that, theoretically, simulated annealing should perform poorly for the clique problem, but the results in Homer and Peinado [79] conclude just the opposite. Simulated annealing actually outperforms most other heuristics on the clique problem in experiments in that paper. Pure versions of simulated annealing for the coloring problem, along with implementations where it is combined with greedy heuristics, exist in Chams *et al.* [74] and Johnson *et al.* [81]. It seems to outperform most greedy heuristics both in time and quality of the solution on random graphs of up to 1000 vertices.

The implementations of tabu search that have performed well for the maximum clique problem can be found in Battiti and Protasi [82], Gendreau *et al.* [83], and Soriano and Gendreau [84,85]. For implementations of tabu search for the graph coloring problem, see de Werra [86], Hertz [87], and Hertz and de Werra [88]. For results comparing not only simulated annealing and tabu search but also greedy heuristics on random graphs of up to 500 vertices and varying densities, see Klimowicz and Kubale [89].

Heuristics based on Continuous Optimization. The Motzkin–Straus formulation yields another potential way of solving the max

clique problem, shifting it from a discrete to a continuous setting. Gibbons *et al.* [90] relaxed the Motzkin–Straus formulation to where it became a problem of optimizing a quadratic over a sphere, which is solvable in polynomial time. The drawback is that solutions may become infeasible for the original problem and hence have to be rounded. Their algorithm performed well on DIMACS benchmark graphs, as recorded in Gibbons *et al.* [90].

Others have taken advantage of the fact that the Motzkin–Straus formulation happens to be a *replicator equation*. A replicator equation, appropriately named because it behaves like the fitness of a population as it replicates, is an equation that as each new solution is plugged in recursively, always increases. With enough starts in random locations, one hopes that recursively plugging in solutions to the replicator equation will end in discovering the global optimum. For a discussion on algorithms to take advantage of the Motzkin–Straus formulation as a replicator equation, including how to remove the need for iteration in the procedure and empirical analysis, see Bomze *et al.* [91] and Bomze and Stix [92]. Results on benchmark graphs are reported in Johnson and Trick [55].

Burer *et al.* [93] studied rank-one and rank-two relaxations of the Lovász semidefinite program [94] and obtained two continuous formulations for the maximum independent set problem. They used these formulations to develop new heuristics for finding large independent sets. Several other heuristics for the maximum independent set problem based on continuous optimization techniques have been proposed in the literature [25,95–97].

As for the graph coloring problem, Karger *et al.* [98] used semidefinite programming to approximate graph coloring. More recently, Dukanovic and Rendl [99] used information provided by near-optimal solutions of a semidefinite programming formulation for the Lovász ϑ -function to generate heuristic solutions for the graph coloring problem. On the basis of the results of numerical experiments, they suggest that this approach could be useful for coloring medium-sized instances.

CLIQUE AND CHROMATIC NUMBERS IN UNIFORM RANDOM GRAPHS

In this section we provide some results concerning the asymptotic behavior of clique and chromatic numbers in *random graphs*. Erdős and Rényi [100–102] founded the random graph theory by introducing several models of *uniform* random graphs. Two of the most popular such models are $G(n, m)$ and $\mathcal{G}(n, p)$ [103], the first of which assigns the same probability to all graphs with n vertices and m edges, while in the second model each pair of vertices is connected by an edge randomly and independently with probability p .

Studying the asymptotic behavior of various graph properties may provide reasonable theoretical estimates of graph invariants for massive graph instances that cannot be tackled by a computer. We say that *almost every* (a.e.) graph has property Q , or the property Q holds *asymptotically almost surely* (a.a.s.), if the probability that a random graph on n vertices has property Q tends to 1 as $n \rightarrow \infty$. Erdős and Rényi observed that many graph properties either hold or do not hold for almost every graph.

Clique Number

Assume that $0 < p < 1$ is fixed. Then instead of the sequence of spaces $\{\mathcal{G}(n, p), n \geq 1\}$ one can work with the single probability space $\mathcal{G}(\mathbf{N}, p)$ containing graphs on $\mathbf{N}$ with the edges chosen independently with probability p . For a graph $G \in \mathcal{G}(\mathbf{N}, p)$ we denote by G_n the subgraph of G induced by the first n vertices $\{1, 2, \dots, n\}$. Then the sequence $\omega(G_n)$ appears to be almost completely determined for a.e. $G \in \mathcal{G}(\mathbf{N}, p)$.

Bollobás and Erdős [104] proved that if $p = p(n)$ satisfies $n^{-\epsilon} < p \leq c$, for every ϵ and some $c < 1$, then there exists a function $cl : \mathbf{N} \rightarrow \mathbf{N}$ such that a.a.s.

$$cl(n) \leq \omega(G_n) \leq cl(n) + 1,$$

that is, the clique number is asymptotically distributed on at most two values. The sequence $cl(n)$ appears to be close

to $l_0(n) = 2 \log_{1/p} n + O(\log \log n)$: for a.e. $G \in \mathcal{G}(\mathbf{N}, p)$ if n is large enough then

$$\begin{aligned} |l_0(n) - 2 \log \log n / \log n| &\leq \omega(G_n) \\ &\leq \lfloor l_0(n) + 2 \log \log n / \log n \rfloor \end{aligned}$$

and

$$\begin{aligned} |\omega(G_n) - 2 \log_{1/p} n + 2 \log_{1/p} \log_{1/p} n \\ - 2 \log_{1/p}(e/2) - 1| &< \frac{3}{2}. \end{aligned}$$

Frieze [105] and Janson *et al.* [106] extended these results by showing that for $\epsilon > 0$ there exists a constant c_ϵ , such that for $\frac{c_\epsilon}{n} \leq p(n) \leq \log^{-2} n$ a.a.s.

$$\begin{aligned} |2 \log_{1/p} n - 2 \log_{1/p} \log_{1/p} n + 2 \log_{1/p}(e/2) \\ + 1 - \epsilon/p| &\leq \omega(G_n) \leq |2 \log_{1/p} n \\ - 2 \log_{1/p} \log_{1/p} n \\ + 2 \log_{1/p}(e/2) + 1 + \epsilon/p|. \end{aligned}$$

Chromatic Number

The problem of coloring random graphs has been studied by several researchers Bollobás [107], Alon and Krivelevich [108], Grimmett and McDiarmid [109], Sós and Straus [110], and Łuczak [111]. Łuczak [111] improved the previous results about the concentration of $\chi(G(n, p))$, proving that for every sequence $p = p(n)$ such that $p \leq n^{-6/7}$ there is a function $ch(n)$ such that a.a.s.

$$ch(n) \leq \chi(G(n, p)) \leq ch(n) + 1.$$

Alon and Krivelevich [108] have shown that for any positive constant γ the chromatic number of a uniform random graph $G(n, p)$, where $p = n^{\frac{1}{2}-\gamma}$, is a.a.s. distributed among two consecutive values. Moreover, a proper choice of $p(n)$ may result in a one-point distribution. The function $ch(n)$ is difficult to find in general, however, it can be characterized in some cases. In particular, Janson *et al.* [106] proved that there exists a constant c_0 such that for any $p = p(n)$ satisfying $\frac{c_0}{n} \leq p \leq \log^{-7} n$ a.a.s.

$$\frac{np}{2 \log np - 2 \log \log np + 1} \leq \chi(G(n, p))$$

$$\leq \frac{np}{2 \log np - 40 \log \log np}.$$

If p is constant, we have the following estimate:

$$\chi(G(n, p)) = \frac{n}{2 \log_b n - 2 \log_b \log_b n + O_c(1)},$$

where $b = 1/(1 - p)$.

REFERENCES

1. Bomze IM, Budinich M, Pardalos PM, *et al.* The maximum clique problem. In: Du D-Z, Pardalos PM, editors. Handbook of combinatorial optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999. pp. 1–74.
2. Pardalos PM, Mavridou T, Xue J. The graph coloring problem: a bibliographic survey. In: Du D-Z, Pardalos PM, editors. Volume 2, Handbook of combinatorial optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1998. pp. 331–395.
3. Miller GA. WordNet. Princeton University; 2009. Available at <http://wordnet.princeton.edu>. Accessed in 2010.
4. Luce RD, Perry AD. A method of matrix analysis of group structure. *Psychometrika* 1949;14:95–116.
5. Kempe AB. On the geographical problem of four colors. *Am J Math* 1879;2:193–200.
6. Cayley A. On the colourings of maps. *Proc R Geogr Soc* 1879;1:259–261.
7. Wilson R. Four colours suffice. Princeton (NJ): Princeton University Press; 2002.
8. Gonthier G. Formal proof—the four-color theorem. *Not Am Math Soc* 2008;55:1382–1393.
9. Abello J, Pardalos PM, Resende MGC. On maximum clique problems in very large graphs. In: Abello J, Vitter J, editors. Volume 50, External memory algorithms and visualization, DIMACS Series on Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 1999. pp. 119–130.
10. Avondo-Bodeno G. Economic applications of the theory of graphs. New York: Gordon and Breach Science Publishers; 1962.
11. Balas E, Yu C. Finding a maximum clique in an arbitrary graph. *SIAM J Comput* 1986;15:1054–1068.
12. Berge C. The theory of graphs and its applications. London: Methuen; 1962.
13. Deo N. Graph theory with applications to engineering and computer science. Englewood Cliffs (NJ): Prentice-Hall; 1974.
14. Trotter LE Jr. Solution characteristics and algorithms for the vertex packing problem. Technical Report 168. Ithaca (NY): Department of Operations Research, Cornell University; 1973.
15. Butenko S, Wilhelm W. Clique-detection models in computational biochemistry and genomics. *Eur J Oper Res* 2006;173:1–17.
16. Balasundaram B, Butenko S. Graph domination, coloring and cliques in telecommunications. In: Resende MGC, Pardalos PM, editors. Handbook of optimization in telecommunications. New York: Springer Science + Business Media; 2006. pp. 865–890.
17. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-completeness. New York: W.H. Freeman and Company; 1979.
18. Arora S, Lund C, Motwani R, *et al.* Proof verification and hardness of approximation problems. *J ACM* 1998;45:501–555.
19. Arora S, Safra S. Probabilistic checking of proofs: a new characterization of NP. *J ACM* 1998;45(1):70–122.
20. Feige U, Kilian J. Zero knowledge and the chromatic number. *J Comput Syst Sci* 1998;57:187–199.
21. Nemhauser GL, Trotter LE. Vertex packing: structural properties and algorithms. *Math Program* 1975;8:232–248.
22. Shor NZ. Dual quadratic estimates in polynomial and Boolean programming. *Ann Oper Res* 1990;25:163–168.
23. Motzkin TS, Straus EG. Maxima for graphs and a new proof of a theorem of Turán. *Can J Math* 1965;17:533–540.
24. Bomze IM. Evolution towards the maximum clique. *J Glob Optim* 1997;10:143–164.
25. Abello J, Butenko S, Pardalos P, *et al.* Finding independent sets in a graph using continuous multivariable polynomial formulations. *J Glob Optim* 2001;21:111–137.
26. Balasundaram B, Butenko S. On a polynomial fractional formulation for independence number of a graph. *J Glob Optim* 2006;35:405–421.

27. Harant J. Some news about the independence number of a graph. *Discussiones Math Graph Theory* 2000;20:71–79.
28. Martins P. Extended and discretized formulations for the maximum clique problem. *Comput Oper Res* 2010;37:1348–1358.
29. Mehrotra A, Trick MA. A column generation approach for graph coloring. *INFORMS J Comput* 1996;8(4):344–354.
30. Caro Y. New results on the independence number. Technical report. Israel: Tel-Aviv University; 1979.
31. Wei VK. A lower bound on the stability number of a simple graph. Technical Report TM 81-11217-9. Murray Hill (NJ): Bell Laboratories; 1981.
32. Wilf HS. The eigenvalues of a graph and its chromatic number. *J Lond Math Soc* 1967;42:330–332.
33. Amin AT, Hakimi SL. Upper bounds on the order of a clique of a graph. *SIAM J Appl Math* 1972;22:569–573.
34. Brooks RL. On coloring the nodes of a network. *Proc Camb Philos Soc* 1941;37:194–197.
35. Szekeres G, Wilf HS. An inequality for the chromatic number of a graph. *J Comb Theory* 1968;4:1–3.
36. Reed B. ω , δ , and χ . *J Graph Theory* 1998;27:177–212.
37. Lovász L. On the shannon capacity of a graph. *IEEE Trans Inform Theory* 1979;25:1–7.
38. Knuth DE. The sandwich theorem. *Electron J Comb* 1994;1:A1.
39. Chudnovsky M, Robertson N, Seymour PD, *et al.* The strong perfect graph theorem. *Ann Math* 2006;164:51–229.
40. Busygin S, Pasechnik DV. On NP-hardness of the clique partition – independence number gap recognition and related problems. *Discrete Math* 2006;304:460–463.
41. Harary F, Ross IC. A procedure for clique detection using the group matrix. *Sociometry* 1957;20:205–215.
42. Paull MC, Unger SH. Minimizing the number of states in incompletely specified sequential switching functions. *IRE Trans Electr Comput* 1959;EC-8:356–367.
43. Marcus PM. Derivation of maximal compatibles using Boolean algebra. *IBM J Res Dev* 1964;8:537–538.
44. Bonner RE. On some clustering techniques. *IBM J Res Dev* 1964;8:22–32.
45. Bednarek AR, Taulbee OE. On maximal chains. *Roum Math Pres et Appl* 1966;11:23–25.
46. Bron C, Kerbosch J. Algorithm 457: finding all cliques on an undirected graph. *Commun ACM* 1973;16:575–577.
47. Gardiner EJ, Artymiuk PJ, Willett P. Clique-detection algorithms for matching three-dimensional molecular structures. *J Mol Graphics Model* 1998;15:245–253.
48. Tomita E, Tanaka A, Takahashi H. The worst-time complexity for finding all the cliques. Technical Report UEC-TR-C5. Tokyo: University of Electro-Communications; 1988.
49. Moon JW, Moser L. On cliques in graphs. *Isr J Math* 1965;3:23–28.
50. Tomita E, Tanaka A, Takahashi H. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor Comput Sci* 2006;363(1):28–42.
51. Tarjan RE, Trojanowski AE. Finding a maximum independent set. *SIAM J Comput* 1977;6:537–546.
52. Robson JM. Algorithms for maximum independent sets. *J Algorithms* 1986;7:425–440.
53. Robson JM. Finding a maximum independent set in time $o(2^{n/4})$. Technical Report 1251-01. Bordeaux, France: LaBRI, Université Bordeaux; 2001.
54. Carraghan R, Pardalos P. An exact algorithm for the maximum clique problem. *Oper Res Lett* 1990;9:375–382.
55. Johnson DS, Trick MA, editors. Volume 26, Cliques, coloring, and satisfiability: second DIMACS implementation challenge, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 1996.
56. Östergård PRJ. A fast algorithm for the maximum clique problem. *Discrete Appl Math* 2002;120:197–207.
57. Tomita E, Kameda T. An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments. *J Glob Optim* 2007;37(1):95–111.
58. Bourjolly J-M, Laporte G, Mercure H. A combinatorial column generation algorithm for the maximum stable set problem. *Oper Res Lett* 1997;20(1):21–29.

59. Rossi F, Smriglio S. A branch-and-cut algorithm for the maximum cardinality stable set problem. *Oper Res Lett* 2001;28(2):63–74.
60. Brown JR. Chromatic scheduling and the chromatic number problem. *Manage Sci* 1972;19:456–463.
61. Brélaz D. New methods to color the vertices of a graph. *Commun ACM* 1979;22(4):251–256.
62. Korman SM. The graph-colouring problem. In: Christofides N, Mingozzi A, Toth P, *et al.*, editors. *Combinatorial optimization*. New York: Wiley; 1979. pp. 211–235.
63. Herrmann F, Hertz A. Finding the chromatic number by means of critical graphs. *J Exp Algorithmics* 2002;7:10.
64. Lucet C, Mendes F, Moukrim A. An exact method for graph coloring. *Comput Oper Res* 2006;33(8):2189–2207.
65. Halldórsson MM, Radhakrishnan J. Greed is good: Approximating independent sets in sparse and bounded-degree graphs. *Algorithmica* 1997;18:145–163.
66. Johnson DS. Worst-case behavior of graph-coloring algorithms. *Proceedings of 5th Southeastern Conference on Combinatorics, Graph Theory and Computing*. Winnipeg; 1974. pp. 513–528.
67. Johnson DS. Approximation algorithms for combinatorial problems. *J Comput Syst Sci* 1974;9:256–278.
68. Kopf R, Ruhe G. A computational study of the weighted independent set problem for general graphs. *Found Control Eng* 1987;12:167–180.
69. Tomita E, Tanaka A, Takahashi H. Two algorithms for finding a near-maximum clique. Technical Report UEC-TR-C1. Tokyo: University of Electro-Communications; 1988.
70. Welsh DJA, Powell MB. An upper bound for the chromatic number of a graph and its applications to timetabling problems. *Comput J* 1967;10:85–86.
71. Matula DW, Marble G, Isaacson JD. Graph coloring algorithms. *Graph theory and computing*. New York: Academic Press, Inc.; 1972. pp. 109–122.
72. Leighton FT. A graph colouring algorithm for large scheduling problems. *J Res Nat Bur Stand* 1979;84:489–496.
73. Biggs N. Some heuristics for graph coloring. In: Nelson R, Wilson R, editors. *Graph colorings*, Pitman Research Notes in Mathematics. New York: Wiley; 1990. pp. 87–96.
74. Chams M, Hertz A, de Werra D. Some experiments with simulated annealing for coloring graphs. *Eur J Oper Res* 1987;32:260–266.
75. White AT. *Graphs, groups and surfaces*. Amsterdam: Elsevier Science Publishers B.V.; 1984.
76. Culberson JC. Iterated greedy graph coloring and the difficulty landscape. Technical Report TR 92-07. Edmonton: University of Alberta Dept of Computer Science; 1992.
77. Caramia M, Dell’Olmo P. Coloring graphs by iterated local search traversing feasible and infeasible solutions. *Discrete Appl Math* 2008;156(2):201–217.
78. Aarts E, Korst J. *Simulated annealing and boltzmann machines*. Chichester: John Wiley & Sons, Inc.; 1989.
79. Homer S, Peinado M. Experiments with polynomial-time clique approximation algorithms on very large graphs. In: Johnson DS, Trick MA, editors. Volume 26, *Cliques, coloring, and satisfiability: Second DIMACS implementation challenge*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 1996. pp. 147–167.
80. Jerrum M. Large cliques elude the metropolis process. *Random Struct Algorithms* 1992;3:347–359.
81. Johnson DS, Aragon CR, McGeoh LA, *et al.* Optimization by simulated annealing: An experimental evaluation, part ii, graph coloring and number partitioning. *Oper Res* 1991;39:378–406.
82. Battiti R, Protasi M. Reactive local search for the maximum clique problem. *Algorithmica* 2001;29:610–637.
83. Gendreau A, Salvail L, Soriano P. Solving the maximum clique problem using a tabu search approach. *Ann Oper Res* 1993;41:385–403.
84. Soriano P, Gendreau M. Tabu search algorithms for the maximum clique problem. In: Johnson DS, Trick MA, editors. Volume 26, *Cliques, coloring, and satisfiability: second DIMACS implementation challenge*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence (RI): American Mathematical Society; 1996. pp. 221–242.
85. Soriano P, Gendreau M. Diversification strategies in tabu search algorithms for the maximum clique problem. *Ann Oper Res* 1996;63:189–207.

86. de Werra D. Heuristics for graph coloring. *Computing* 1990;7:191–208.
87. Hertz A. Cosine: a new graph coloring algorithm. *Oper Res Lett* 1991;10:411–415.
88. Hertz A, de Werra D. Using tabu search techniques for graph coloring. *Computing* 1987;39:345–351.
89. Klimowicz D, Kubale M. Graph coloring by tabu search and simulated annealing. *Arch Control Sci* 1993;2:41–54.
90. Gibbons LE, Hearn DW, Pardalos PM. A continuous based heuristic for the maximum clique problem. In: Johnson DS, Trick MA, editors. *Cliques, coloring, and satisfiability: second DIMACS challenge, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 26. Providence (RI): American Mathematical Society; 1996. pp. 103–124.
91. Bomze IM, Pelillo M, Giacomini R. Evolutionary approach to the maximum clique problem: empirical evidence on a larger scale. In: Bomze IM, Csendes T, Horst R, editors. *Developments of global optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1997. pp. 95–108.
92. Bomze IM, Stix V. Genetical engineering via negative fitness: evolutionary dynamics for global optimization. *Ann Oper Res* 1999;89: 279–318.
93. Burer S, Monteiro RDC, Zhang Y. Maximum stable set formulations and heuristics based on continuous optimization. *Math Program* 2002;94:137–166.
94. Grötschel M, Lovász L, Schrijver A. *Geometric algorithms and combinatorial optimization*. 2nd ed. Berlin: Springer; 1993.
95. Busygin S, Butenko S, Pardalos PM. A heuristic for the maximum independent set problem based on optimization of a quadratic over a sphere. *J Comb Optim* 2002;6:287–297.
96. Balasundaram B, Butenko S. On a polynomial fractional formulation for independence number of a graph. *J Glob Optim* 2006;35(3):405–421.
97. Busygin S. A new trust region technique for the maximum weight clique problem. *Discrete Appl Math* 2006;154:2080–2096.
98. Karger D, Motwani R, Sudan M. Approximate graph coloring by semidefinite programming. *J ACM* 1998;45(2):246–265.
99. Dukanovic I, Rendl F. A semidefinite programming-based heuristic for graph coloring. *Discrete Appl Math* 2008;156(2):180–189.
100. Erdős P, Rényi A. On random graphs. *Publ Math* 1959;6:290–297.
101. Erdős P, Rényi A. On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 1960;5:17–61.
102. Erdős P, Rényi A. On the strength of connectedness of a random graph. *Acta Math Acad Sci Hung* 1961;12:261–267.
103. Bollobás B. *Random graphs*. New York: Academic Press; 1985.
104. Bollobás B, Erdős P. Cliques in random graphs. *Math Proc Camb Phil Soc* 1976;80: 419–427.
105. Frieze A. On the independence number of random graphs. *Discrete Math* 1990;81:171–175.
106. Janson S, Łuczak T, Ruciński A. *Random graphs*. New York: John Wiley & Sons, Inc.; 2000.
107. Bollobás B. The chromatic number of random graphs. *Combinatorica* 1988;8:49–56.
108. Alon N, Krivelevich M. The concentration of the chromatic number of random graphs. *Combinatorica* 1997;17:303–313.
109. Grimmett GR, McDiarmid CJH. On coloring random graphs. *Math Proc Camb Philos Soc* 1975;77:313–324.
110. Sós VT, Straus EG. Extremal of functions on graphs with applications to graphs and hypergraphs. *J Comb Theory B* 1982;32:246–257.
111. Łuczak T. A note on the sharp concentration of the chromatic number of random graphs. *Combinatorica* 1991;11:295–297.

MAXIMUM FLOW ALGORITHMS

ZEYNEP SARGUT
Department of Business
Administration, Izmir
University, Izmir, Turkey

RAVINDRA K. AHUJA
Innovative Scheduling, Inc., and
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

The *maximum flow problem* seeks the maximum possible flow in a capacitated network from a specified source node to a specified sink node without exceeding the capacity of any arc while conserving the flow at the intermediate nodes. The dual of this problem is the *minimum cut problem*, which seeks a set of arcs with the smallest total capacity whose removal separates the source node and the sink node.

The maximum flow and minimum cut problems arise in a variety of application settings as diverse as manufacturing, communication systems, distribution planning, and scheduling. Moreover, we come across these problems as subproblems in the solution of more difficult network optimization problems. The maximum flow problem directly occurs in problems such as machine scheduling, the assignment of computer modules to computer processors, the rounding of census data to retain the confidentiality of individual households, and tanker scheduling. In this article, we study the maximum flow and minimum cut problems by introducing the underlying theory and algorithms. Ahuja *et al.* [1] contains a wealth of additional material that expands this discussion.

We provide the representative selection problem as an example of a maximum flow problem. Consider a town that has r residents

$R_1, R_2, \dots, R_r$; q clubs $C_1, C_2, \dots, C_q$; and p political parties $P_1, P_2, \dots, P_p$. Each resident is a member of at least one club and can belong to exactly one political party. Each club must nominate one of its members to represent it in the town's governing council so that the number of council members belonging to the political party P_k is at most u_k . Is it possible to find such a "balanced" council?

We consider a problem with $r = 7, q = 4, p = 3$ and formulate it as a maximum flow problem in Fig. 1. The network also contains a source node s and a sink node t . It contains an arc (s, C_i) for each node C_i denoting a club, an arc (C_i, R_j) whenever the resident R_j is a member of the club C_i , and an arc (R_j, P_k) if the resident R_j belongs to the political party P_k . Finally, we add an arc (P_k, t) for each $k = 1, 2, 3$ of capacity u_k ; all other arcs have unit capacity. The number over an arc represents the arc capacity.

We next find a maximum flow value in this network. If the maximum flow value equals q , then the town has a balanced council; otherwise, it does not. This is because any flow of value q in the network corresponds to a balanced council.

LINEAR PROGRAMMING FORMULATION

Let $G = (N, A)$ be a *directed network* defined by a set N of n nodes and a set A of m directed arcs. We refer to nodes i and j as *end points* of arc (i, j) . A *directed path* $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k$ is a set of arcs $(i_1, i_2), (i_2, i_3), \dots, (i_{k-1}, i_k)$. Each arc (i, j) has an associated *capacity* u_{ij} denoting the maximum possible amount of flow on this arc. We assume that each arc capacity u_{ij} is an integer, and let $U = \max\{u_{ij} : (i, j) \in A\}$. The network has two distinguished nodes, a *source node* s and a *sink node* t .

The maximum flow problem is to find the maximum flow from the source node s to the sink node t that satisfies the arc capacities and mass balance constraints at all nodes. We can state the problem formally as follows:

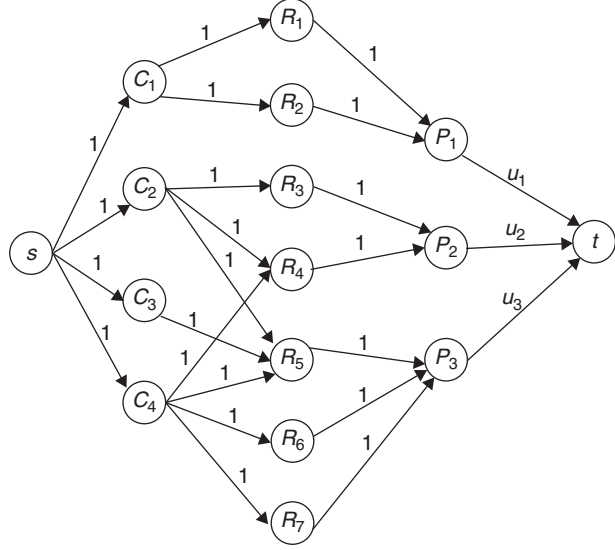


Figure 1. System of distinct representatives.

Maximize v

subject to

$$\sum_{\{j:(i,j) \in A\}} x_{ij} - \sum_{\{j:(j,i) \in A\}} x_{ji} = \begin{cases} v & \text{for } i = s \\ 0 & \text{for } i \neq s \text{ or } t \\ -v & \text{for } i = t, \end{cases} \quad (1a)$$

$$0 \leq x_{ij} \leq u_{ij} \text{ for all } (i,j) \in A. \quad (1b)$$

We refer to a vector $x = \{x_{ij} : (i,j) \in A\}$ satisfying Equations (1a) and (1b) as a *flow*, and the corresponding value of the scalar variable v as the *value* of the flow. We refer to the constraints (1a) as the *mass balance constraints*, and we refer to the constraints (1b) as the *flow bound constraints*.

The maximum flow problem is a special type of minimum cost flow problem and can be formulated as a minimum cost flow problem by adding an arc from node t to s , and setting the capacity of this arc to infinity (if it does not exist).

In examining the maximum flow problem, we impose the following assumptions.

Assumption 1. Whenever the network contains $\text{arc}(i,j)$, then it also contains $\text{arc}(j,i)$.

This assumption is satisfied by adding the missing arcs with zero capacities.

Assumption 2. The network is directed.

If the network contains an undirected arc (i,j) with capacity u_{ij} , it means that this arc permits flow from node i to node j and also from node j to node i . The total flow (from node i to node j , plus from node j to node i) has an upper bound u_{ij} . It can be shown that the undirected problem can be reduced to the directed problem with comparable size and capacity values [2].

Assumption 3. The lower bounds on the arc flows are zero.

Sometimes the flow vector x might be required to satisfy lower bound constraints imposed upon the arc flows; that is, for each arc $(i,j) \in A$, we impose the condition $x_{ij} \geq l_{ij}$. Whereas the maximum flow problem with zero lower bounds always has a feasible solution (since the zero flow is feasible), the problem with nonnegative lower bounds could be infeasible (Fig. 2). This problem does not have a feasible solution because arc $(s, 2)$ must carry at least five units of flow into node 2, and arc $(2, t)$ can take out at most



Figure 2. A maximum flow problem with no feasible solution.

four units of flow; therefore, we can never satisfy the mass balance constraint of node 2.

Any maximum flow algorithm for problems with nonnegative lower bounds has two phases: (i) to determine whether the problem is feasible or not, and (ii) if feasible, to establish a maximum flow. It can be shown that each phase essentially reduces to solving a maximum flow problem with zero lower bounds [2].

Assumption 4. *There is a single source and a single sink.*

If the problem is to find the maximum flow from the multiple sources to the multiple sinks, we can create a combined source that is connected with infinite capacity arcs to all source nodes and a combined sink that is connected with infinite capacity arcs to all sink nodes.

Assumption 5. *Algorithms whose complexity bounds involve U , assume integrality of the arc capacities.*

There are two cases for the arc capacities: real valued and integral.

DEFINITIONS

s-t path

An *s-t path* is defined as a path starting from node s and ends at node t .

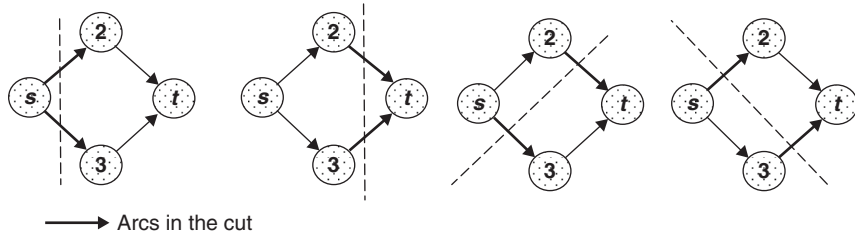


Figure 3. All $s-t$ cuts for a given network.

Cut

A *cut* $[S, \bar{S}]$ partitions the node set N into two nonempty subsets, S and $\bar{S} = N - S$. It consists of all arcs with one end point in S and the other in $\bar{S}$. We refer to the arcs directed from S to $\bar{S}$, denoted by $(S, \bar{S})$, as *forward arcs* in the cut. The cut $[S, \bar{S}]$ is called an *s-t cut* if $s \in S$ and $t \in \bar{S}$. We define the *capacity of the cut* $[S, \bar{S}]$, denoted by $u[S, \bar{S}]$, as $\sum_{(i,j) \in (S, \bar{S})} u_{ij}$ (note that only forward arcs are included). A *minimum cut* in G is an $s-t$ cut of minimum capacity (Fig. 3).

Max-Flow Min-Cut Theorem

The value of the maximum flow from node s to node t in a capacitated network equals the value of the minimum capacity $s-t$ cut.

After determining a maximum flow, a minimum cut can be obtained by a small amount of postprocessing. The minimum-cut problem is the dual of the maximum-flow problem.

Residual Network

The concept of *residual network* plays a central role in the development of maximum flow algorithms. Given a flow x , the residual capacity r_{ij} of any arc $(i, j) \in A$ is equal to $u_{ij} - x_{ij} + x_{ji}$. The residual capacity r_{ij} has two components: (i) $u_{ij} - x_{ij}$, the unused capacity of arc (i, j) , and (ii) x_{ji} , the current flow on arc (j, i) , which we can cancel to increase the flow from node i to node j . We refer to the network $G(x)$ consisting of the arcs with positive residual capacities as the *residual network with respect to the flow x* (Fig. 4).

Distance Labels

A *distance function* $d: N \rightarrow \mathbb{Z}^+ \cup \{0\}$ with respect to the residual capacities r_{ij} is a function from the set of nodes to the set of nonnegative integers. We refer to $d(i)$ as the *distance label* of node i , and distance labels are *valid* with respect to a flow x if they satisfy the following conditions:

$$d(t) = 0; \quad (2a)$$

$$d(i) \leq d(j) + 1$$

for every arc (i,j) in the residual

$$\text{network } G(x). \quad (2b)$$

We say that an arc (i,j) in the residual network is *admissible* if it satisfies the condition that $d(i) = d(j) + 1$; otherwise, we refer to that arc as *inadmissible*. We also refer to a path from node s to node t consisting entirely of admissible arcs as an *admissible path*. An *admissible path* is the *shortest path* (in terms of number of arcs) from the source to the sink.

Complexity

We say that an algorithm is a *polynomial-time algorithm* if its worst case running time is bounded above by a polynomial function of the input size parameters. For a maximum flow problem, the input size parameters are n, m , and $\log U$ (the number of bits needed to specify the largest arc capacity). We refer to a maximum flow algorithm as a

pseudopolynomial-time algorithm if its worst case running time is bounded by a polynomial function of n, m , and U . $O(nmU)$ denotes that the running time is bounded above by some constant multiple of $n m U$ and $\Omega(nmU)$ denotes that the running time is bounded below by some constant multiple of $n m U$. For many years, researchers attempted to find an algorithm faster than the decomposition barrier $\Omega(nm)$. This is a natural lower bound on algorithms that require flow decomposition and on algorithms that augment flow on one arc of a path at a time. Cheriyan *et al.* [3] beat this barrier for dense graphs, while Goldberg and Rao [4] further improve this lower bound.

AUGMENTING PATH ALGORITHMS

The first generation algorithms to solve maximum flow problems are augmenting path algorithms. Let x be a feasible flow in the network G . We refer to a directed path from the source to the sink in the residual network $G(x)$ as an *augmenting path*.

We define the residual capacity $\delta(P)$ of an augmenting path P as the maximum amount of flow that can be sent along it; $\delta(P) = \min\{r_{ij} : (i,j) \in P\}$. Since the residual capacity of each arc in the residual network is strictly positive, the residual capacity of an augmenting path is strictly positive. Sending δ units of additional flow along an augmenting path decreases the residual capacity of each arc (i,j) in the path by δ units and increases the

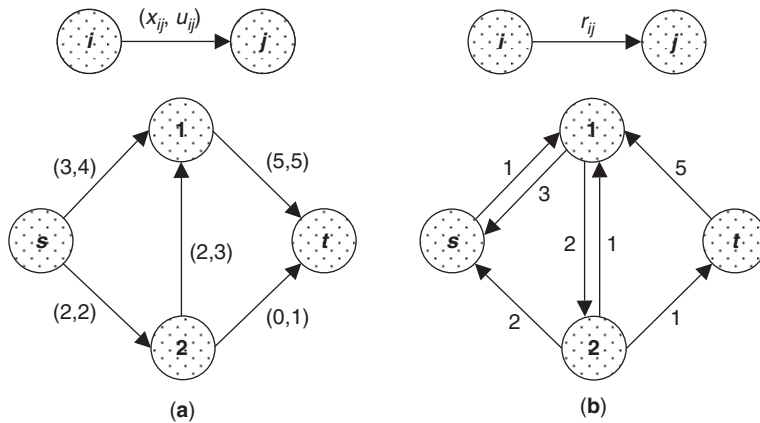


Figure 4. (a) Original network with arc capacities and a flow x ; (b) the residual network $G(x)$.

residual capacity of the reverse arc (j, i) by δ units.

The generic augmenting path algorithm [5,6] identifies augmenting paths in $G(x)$ and sends flow on these paths until the network contains no such path. We state the algorithm below.

Algorithm. Generic augmenting path algorithm.

```

begin
     $x := 0$ ;
    while  $G(x)$  contains an augmenting
    path do
        begin
            identify an augmenting path  $P$ ;
            augment  $x$  by  $\min \{r_{ij} : (i, j) \in P\}$  units
            along  $P$  and update  $G(x)$ ;
        end
    end

```

When the algorithm terminates, the final residual capacities r_{ij} can be used to construct a maximum flow. Since $r_{ij} = u_{ij} - x_{ij} + x_{ji}$, the arc flows satisfy the equality $x_{ij} - x_{ji} = u_{ij} - r_{ij}$. If $u_{ij} > r_{ij}$, we can set $x_{ij} = u_{ij} - r_{ij}$ and $x_{ji} = 0$; otherwise, we set $x_{ij} = 0$ and $x_{ji} = r_{ij} - u_{ij}$.

The performance of the algorithm depends on (i) the number of augment iterations, and (ii) the time to identify an augmenting path. Both numbers depend on which augmenting path we choose at each iteration. The further discussion on which augmenting paths to choose appears later in the section. We can identify an augmenting path P in $G(x)$ by using a graph search algorithm. A graph search algorithm starts at node s and progressively finds all nodes that are reachable using directed paths. Most search algorithms run in $O(m)$ time, and either identify an augmenting path or conclude that $G(x)$ contains no augmenting paths; the latter happens when the sink node is not reachable from the source node.

In each iteration, the algorithm augments a positive integer amount of flow from the source node to the sink node. To give a bound on the number of iterations, we compute a bound on the maximum flow value. By definition, U denotes the largest arc capacity, and so the capacity of the cut $[\{s\}, N - \{s\}]$

is at most nU . Since the value of any flow can never exceed the capacity of any cut in the network and the algorithm sends at least one unit of flow at each iteration, we obtain a bound of nU on the number of iterations. Consequently, the running time of the algorithm is $O(nmU)$, which is a pseudopolynomial-time bound.

Shortest Augmenting Path Algorithm

The shortest augmenting path algorithm [7] always augments flow along an s - t path in the residual network that has the fewest number of arcs. Since the minimum distance from any node i to the sink node t is monotonically nondecreasing over all augmentations, we reduce the average time per augmentation to $O(n)$.

The shortest augmenting path algorithm proceeds by augmenting flows along *admissible paths*. It constructs an admissible path, incrementally maintaining a partially admissible path, that is, a path from s to a node i consisting solely of admissible arcs, and iteratively performs advance or retreat operations from the last node of the partially admissible path, which we refer to as the *current node*. If the current node i has (i.e., is incident to) an admissible arc (i, j) , then we perform an advance operation and add arc (i, j) to the partial admissible path; otherwise, we perform a retreat operation and backtrack one arc. We repeat these operations until the partial admissible path reaches the sink node, at which time we augment $\min \{r_{ij} : (i, j) \in P\}$ units of flow. We repeat this process until the flow is maximum. We present the algorithm below. Note that $\text{pred}(i)$ is defined as the predecessor of node i on the current path.

Algorithm. Shortest augmenting path algorithm.

```

begin
     $x := 0$ , obtain the exact distance labels  $d$ ;
    current node  $:= s$ ;
    while  $d(s) < n$  do
        begin
            if current node has an outgoing
            admissible arc
                then advance (current node)
            else retreat (current node);
        end
    end

```

```

        if current node =  $t$  then augment;
    end
end
procedure advance ( $i$ )
begin
    let  $(i, j)$  be an admissible arc
     $\text{pred}(j) := i$  and  $i := j$ ;
end
procedure retreat ( $i$ )
begin
     $d(i) := \min\{d(j)+1 : (i, j) \in A \text{ and } r_{ij} > 0\}$ ;
    if  $i \neq s$  then  $i := \text{pred}(i)$ ;
end
procedure augment
begin
    using the predecessor indices identify an
    augmenting path  $P$ ;
    augment  $x$  by  $\min\{r_{ij} : (i, j) \in P\}$  units
    along  $P$ ;
end

```

This algorithm is easy to implement in practice. The resulting algorithms identify at most $O(nm)$ augmenting paths, and this bound is tight; that is, in particular examples, these algorithms perform $O(nm)$ augmentations. The only way to improve the running time of the shortest augmenting path algorithm is to perform fewer computations per augmentation. It can be shown that the use of a sophisticated data structure, called *dynamic trees*, reduces the average time for each augmentation from $O(n)$ to $O(\log n)$ [8]. This implementation of the shortest augmenting path algorithm runs in $O(nm \log n)$ time.

Capacity-Scaling Augmenting Path Algorithm

We start with explaining the *maximum capacity augmenting path algorithm*. The maximum capacity augmenting path algorithm, which was originally developed by Edmonds and Karp [9], always augments flow only along the path with the maximum residual capacity. Let x be any flow, v be its flow value, and v^* be the maximum flow value. Flow decomposition theory shows that in the residual network $G(x)$ we can find m or fewer directed paths from the source to the sink whose residual capacities sum to $(v^* - v)$. Therefore, the maximum

capacity augmenting path has a residual capacity of at least $(v^* - v)/m$. Now consider a sequence of $2m$ consecutive maximum capacity augmentations starting with the flow x . Suppose that v' is the resulting flow value. If in each of these augmentations we augment at least $(v^* - v)/2m$ units of flow, then we will establish a maximum flow within $2m$ or fewer iterations. However, if one of these $2m$ augmentations carries less than $(v^* - v)/2m$ units of flow, $(v^* - v')$ is at most $(v^* - v)/2$ [[1], p. 67]. Since the residual capacity of any augmenting path is between 1 and $2U$, the flow must be maximum in $O(m \log U)$ iterations.

The maximum capacity augmentation algorithm reduces the number of iterations in the generic augmenting path algorithm from $O(nU)$ to $O(m \log U)$. However, the algorithm performs more computations per iteration, since it needs to identify not just any augmenting path but the one with the maximum residual capacity.

We shall now present a variation of the maximum capacity augmenting path algorithm [7] that does not perform more computations per iteration than the generic augmenting path algorithm, and yet establishes a maximum flow within $O(m \log U)$ iterations. Since this algorithm scales the arc capacities implicitly, we refer to it as the *capacity-scaling* algorithm.

We augment flow along a path with a sufficiently large residual capacity instead of a path with the maximum augmenting capacity. We can obtain a path with a sufficiently large residual capacity fairly easily in $O(m)$ time. To define the capacity-scaling algorithm, let us introduce the parameter Δ and define the Δ -residual network as a network containing arcs whose residual capacity is at least Δ with respect to a given flow x . Let $G(x, \Delta)$ denote the Δ -residual network. Note that $G(x, 1) = G(x)$ and $G(x, \Delta)$ is a subgraph of $G(x)$. The algorithm is given below.

Algorithm. Capacity-scaling algorithm.

```

begin
     $x := 0$ ;  $\Delta := 2^{\lfloor \log U \rfloor}$ ;
    while  $\Delta \geq 1$  do
        begin
            If  $G(x, \Delta)$  contains a path from node  $s$ 

```

```

    to node  $t$  then
    begin
        identify a path  $P$  in  $G(x, \Delta)$ ;
        augment  $x$  by  $\min\{r_{ij} : (i, j) \in P\}$ 
        units along  $P$ ;
        update  $G(x, \Delta)$ ;
    end
     $\Delta := \Delta/2$ ;
end
end

```

The capacity-scaling algorithm solves the maximum flow problem within $O(m \log U)$ iterations and runs in $O(m^2 \log U)$ time [9]. Moreover, Ahuja and Orlin [7] improve the running time to $O(nm \log U)$ by using the shortest augmenting path as a subroutine in the step where an augmenting path is identified.

BLOCKING FLOW ALGORITHMS

Instead of identifying augmenting paths one by one, Dinic [10] suggested identifying a set of augmenting paths at each iteration.

Let $d(i)$ be the length of the path with fewest number of arcs from node i to node t . A *layered network* is the subgraph of $G(x)$ that only contains the admissible arcs, that is, (i, j) with $d(i) = d(j) + 1$ (Fig. 5).

A *blocking flow* is the maximum flow in the layered network, and it includes at most m augmenting paths. The flow will be maximum after $n - 1$ blocking flow operations. A blocking flow can be identified in $O(nm)$ time

by using depth-first search or $O(m \log n)$ time by employing dynamic tree structure [8]. The total running times of these algorithms are $O(n^2m)$ and $O(nm \log n)$, respectively. The algorithm is formally given below.

Algorithm. Blocking flow algorithm.

```

begin
     $x := 0$ ;
    while sink is reachable from the source do
    begin
        identify the current layered network;
        find the corresponding blocking flow  $f$ ;
        augment  $x$  by  $f$ ;
        update the residual network;
    end
end

```

Binary Blocking Flow Algorithm

In classical blocking flow algorithms, each arc has a unit length. Goldberg and Rao [4] introduce the blocking flow algorithm with residual capacity-based arc lengths. Let us denote the length of arc (i, j) as

$$l_{ij} = \begin{cases} 0 & \text{if } 2\Delta \leq r_{ij} < 3\Delta, d(i) = d(j), r_{ji} \geq 3\Delta \\ l_{ij} & \text{otherwise,} \end{cases}$$

where

$$l_{ij} = \begin{cases} 0 & \text{if } r_{ij} \geq 3\Delta \\ 1 & \text{otherwise.} \end{cases}$$

The algorithm consists of $O(n \log U)$ Δ -scaling phases. $[S_k, T_k]$ is a canonical cut if $S_k = \{v \in N; d(v) \geq k\}$, $T_k = N - S_k$. We

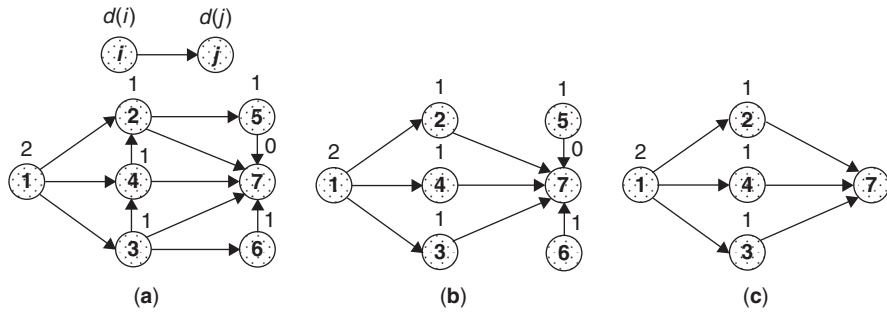


Figure 5. Forming layered network: (a) residual network; (b) corresponding layered network; (c) layered network after deleting redundant arcs.

denote $[\bar{S}, \bar{T}]$ as the canonical cut with the smallest capacity and $u[\bar{S}, \bar{T}]$ as the capacity of this cut. d_l and $d_{l'}$ are the distance labels in terms of the length functions l and l' , respectively. F is the upper bound on the flow value. The algorithm is given below.

Algorithm. Binary blocking flow algorithm.

```

begin
  current cut :=  $\{s\}, \{N - s\}, F := nU$ ,
   $\Delta := \min\{\lceil F/m^{1/2} \rceil, \lceil F/n^{2/3} \rceil\}$ 
  while  $\Delta \geq 1$  do
    begin
      while  $u[\bar{S}, \bar{T}] > F/2$  do
        begin
          compute  $l'$ ;
          contract strongly connected
            components induced by
            zero-length arcs;
          find a flow of value  $\Delta$  or a
            blocking flow;
          update the residual
            network;
          compute  $l$  and  $d_l$ ;
        end
       $F := u[\bar{S}, \bar{T}]$ , current cut :=  $[\bar{S}, \bar{T}]$ ,
       $\Delta := \min\{\lceil F/m^{1/2} \rceil, \lceil F/n^{2/3} \rceil\}$ 
    end
  end

```

Contracting the strongly connected components is a very important step for the algorithm, since the shortest paths on the residual graph may have admissible cycles because of the zero-length arcs. The algorithm contracts the nodes that are strongly connected with zero-length arcs. Afterward, the flow is routed inside the contracted nodes. The binary blocking flow algorithm runs in $O(\min\{n^{2/3}, \sqrt{m}\}m \log(n^2/m) \log U)$ time.

PREFLOW-PUSH ALGORITHMS

Another class of algorithms for solving the maximum flow problem, known as *preflow-push* algorithms, works with individual arcs rather than paths as compared to augmenting path algorithms. Sending a flow of δ units along a path of k arcs decomposes into k basic operations of sending a flow of δ units along

each of the arcs of the path. We shall refer to each of these basic operations as a *push*.

While a path augmentation maintains conservation of flow at all nodes, the preflow-push algorithms violate conservation of flow at all steps except at the very end, and instead maintain a “preflow” at each iteration. A *preflow* is a vector x satisfying the flow bound constraints (1b) and the following relaxation of the mass balance constraints (1a):

$$\sum_{(j,i) \in A} x_{ji} - \sum_{(i,j) \in A} x_{ij} \geq 0, \text{ for all } i \in N - \{s, t\}.$$

Each element of a preflow vector is either a nonnegative real number or $+\infty$. For a given preflow x , we define the *excess* for each node $i \in N - \{s, t\}$ as follows:

$$e(i) = \sum_{(j,i) \in A} x_{ji} - \sum_{(i,j) \in A} x_{ij}.$$

We refer to a node with positive excess as an *active node*. We adopt the convention that the source and sink nodes are never active. In a preflow-push algorithm, the presence of an active node indicates that the solution is infeasible. Consequently, the basic operation in this algorithm is to select an active node i and try to remove the excess by pushing flow out of it. If we simply push flow to an adjacent node in an arbitrary manner and the other nodes do the same, then it is conceivable that some nodes would continue pushing flow among themselves, resulting in an infinite loop. Ultimately, since we want to send the flow to the sink node, it seems reasonable for an active node to push flow to another node that is “closer” to the sink. If all nodes maintain this rule, then the algorithm could never encounter an infinite loop. The concept of distance labels defined next allows us to implement this algorithmic strategy.

The preflow-push algorithms maintain a *valid distance label* for each node in the network. The basic operation in the preflow-push algorithm is to select an active node i and try to remove the excess by pushing flow to a node with a smaller distance label. If node i has an admissible arc (i, j) , then $d(j) = d(i) - 1$ and the algorithm sends flow on admissible

arcs to relieve the node's excess. If node i has no admissible arc, then the algorithm increments the distance label of node i so that node i has an admissible arc. The algorithm terminates when the network contains no active nodes; that is, excess resides only at the source and sink nodes.

A push of δ units from node i to node j decreases both the excess $e(i)$ of node i and the residual r_{ij} of arc (i, j) by δ units, and increases both $e(j)$ and r_{ji} by δ units. If $\delta = r_{ij}$, the push on arc (i, j) is *saturating*. Otherwise, if $\delta = e(i)$, then the push is *nonsaturating*. The algorithm is given below.

Algorithm. Generic preflow-push algorithm.

```

begin
   $x := 0, d(i) := 0$  for all  $i \in N$ ;
   $x_{sj} := u_{sj}, e(j) = u_{sj}$  for all arc  $(s, j) \in A$ ,
   $e(s) := 0, d(s) := n$ ;
  while  $G(x)$  contains an active node do
    begin
      select any active node  $k$ ;
      if the network contains an admissible
        arc  $(k, j)$  then
        push  $\min\{e(k), r_{kj}\}$  units from  $k$  to  $j$ ;
      else
        replace  $d(k)$  by  $\min\{d(j) + 1 : (k, j) \in A \text{ and } r_{kj} > 0\}$ ;
      end
    end
  end

```

The complexity of the algorithm depends on how we choose an active node. It is possible to show that the preflow-push algorithm maintains valid distance labels at all steps of the algorithm and increases the distance label of any node at most $2n$ times. If we choose any active node, the algorithm performs $O(nm)$ saturating pushes and $O(n^2m)$ nonsaturating pushes. Therefore, the algorithm runs in $O(n^2m)$ time [11].

FIFO Preflow-Push Algorithm

The first-in-first-out (FIFO) preflow-push algorithm [12] examines active nodes in *first-in-first-out* order. The algorithm maintains the set of active nodes as a queue. It selects the node i in front of the queue, performs pushes from this node, and adds new active

nodes/relabelled nodes to the end of queue. The algorithm terminates when the queue is empty.

The first phase consists of those node examinations for the nodes that become active during the preprocess operation. The second phase consists of node examinations of all the nodes that are in the queue after the algorithm has examined the nodes in the first phase. Similarly, the third phase consists of the node examinations of all the nodes that are in the queue after the algorithm has examined the nodes in the second phase, and so on.

It can be shown that the algorithm performs at most $2n^2 + n$ phases. Each phase examines any node at most once, and each node examination performs at most one nonsaturating push. Therefore, a bound of $2n^2 + n$ on the total number of phases would imply a bound of $O(n^3)$ on the number of nonsaturating pushes. This result would also imply that the FIFO preflow-push algorithm runs in $O(n^3)$ time, because the bottleneck operation in the generic preflow-push algorithm is the number of nonsaturating pushes.

Highest Label Preflow-Push Algorithm

The highest label preflow-push algorithm always pushes flow from an active node with the highest distance label. Let $h^* = \max\{d(i) : i \text{ is active}\}$. The algorithm first examines nodes with distance labels equal to h^* and pushes flow to nodes with distance labels equal to $h^* - 1$. In turn, these nodes push flow to nodes with distance labels equal to $h^* - 2$, and so on, until either the algorithm relabels a node or it exhausts all the active nodes. When it relabels a node, the algorithm repeats the same process. Note that if the algorithm does not relabel any node during n consecutive node examinations, then all the excess reaches the sink (or the source) and the algorithm terminates. Since the algorithm performs at most $2n^2$ relabel operations, we immediately obtain a bound of $O(n^3)$ on the number of node examinations. Since each node examination entails at most one nonsaturating push, the highest label preflow-push algorithm performs $O(n^3)$ nonsaturating pushes.

We use the following data structure to select a node with the highest distance label without spending too much effort. For each $k = 1, 2, \dots, 2n - 1$, we maintain the list

$$\text{LIST}(k) = \{i : i \text{ is active and } d(i) = k\}$$

in the form of either linked stacks or linked queues (see Ref. 13, pp. 200–204, for information on stacks and queues). We define a variable *level*, which is an upper bound on the highest value of k for which $\text{LIST}(k)$ is nonempty. In order to determine a node with the highest distance label, we examine the lists $\text{LIST}(\text{level})$, $\text{LIST}(\text{level}-1)$, $\dots$, until we find a nonempty list, say $\text{LIST}(p)$. We set the level equal to p and select any node in $\text{LIST}(p)$. Moreover, while examining a node, if its distance label increases, then we set the level equal to the new distance label of the node. Observe that the total increase in level is at most $2n^2$, and so the total decrease is at most $2n^2 + n$. Consequently, scanning the lists $\text{LIST}(\text{level})$, $\text{LIST}(\text{level}-1)$, $\dots$ in order to find the first nonempty list is not a bottleneck operation.

Cheriyān and Maneshwari [14] show that the number of nonsaturating pushes is $O(n^2\sqrt{m})$. The highest label preflow-push algorithm is currently the most efficient method for solving the maximum flow problem in practice, because it performs the least number of nonsaturating pushes. Cherkassy and Goldberg [15] exhibit a family of problems for which the running time of all implementations they consider is quadratic. They show that the highest label preflow-push algorithm with global and gap relabeling heuristics is the fastest algorithm for all problem classes that they studied.

Excess-Scaling Algorithm

Excess-scaling algorithm is another version of the generic preflow-push algorithm that pushes preflows only from the active nodes with large excess. The algorithm is given below.

Algorithm. Excess-scaling algorithm.

begin

$x := 0, d(s) := n, d(t) := 0, d(i) := 1$
for $i \in N - \{s, t\}$;

set x_{sj} and $e(j)$ to u_{sj} for all arcs $(s, j) \in A$,
and $\Delta := 2^{\lceil \log U \rceil}$;
while $\Delta \geq 1$ **do**
 begin
 if $G(x, \Delta)$ contains an active node with
 excess more than $\Delta/2$ **then**
 begin
 select node i with excess more than
 $\Delta/2$ and the smallest distance label;
 if the network contains an admissible
 arc (k, j) **then**
 push $\min\{e(k), r_{kj}, \Delta - e(j)\}$ units
 of flow from k to j ;
 else
 update $d(k) = \min\{d(j) + 1 :$
 $(k, j) \in A \text{ and } r_{kj} > 0\}$
 end
 $\Delta := \Delta/2$;
 end
 end

The excess-scaling algorithm runs in $O(nm + n^2 \log U)$ time [16]. Ahuja *et al.* [17] obtain the improved version by (i) increasing the scale factor from 2 to k (Δ becomes Δ/k in the next phase), and (ii) choosing the excess node with the highest distance label. This version runs in $O(nm \log(n/m\sqrt{\log U}))$ time.

OTHER ALGORITHMS

In this article, we briefly mention the other approaches to solve maximum flow problems.

Malhotra, Kumar, and Maheshwari (MPM) Algorithm

This algorithm relies on the idea of selecting the node with the lowest flow potential, $p_x(i)$, that is, the maximum amount of flow that can be forced through a node.

$$p_x(i) = \min \left\{ \sum_{(j,i) \in A} (u_{ji} - x_{ji}), \sum_{(i,k) \in A} (u_{ik} - x_{ik}) \right\}.$$

Algorithm. MPM algorithm.

begin

$x := 0$;
while there are nodes in the
network **do**
 begin
 select the node k with minimum

```

        flow potential;
        send  $p_x(k)$  units of flow from node  $k$ 
            toward the node  $s$  and  $t$ ;
        augment  $x$  by  $p_x(k)$ , update  $G(x)$ ;
        delete all saturated arcs, node  $k$ , and
            the nodes with only incoming or
            outgoing arcs;
    end
end

```

Since at the last step in the while loop at least one node will be deleted, this loop will be executed at most n times.

Maximum Adjacency Ordering Algorithm

The ordering of nodes $v_1, v_2, \dots, v_n$ is called maximum adjacency (MA) ordering if for all i , v_i is the node in $N - \{v_1, v_2, \dots, v_{i-1}\}$ that has the largest sum of the residual capacities of the arcs to the nodes $v_1, v_2, \dots, v_{i-1}$.

MA ordering algorithm is proposed by Fujishige [18] and runs in $O(n(m + n \log n) \log n U)$ time and the scaling version runs in $O(nm \log U)$ time. Shioura [19] shows that the algorithm is not strongly polynomial by giving a real-valued instance for which the algorithm does not terminate. Matsuoka and Fujishige [20] propose an improved version of Fujishige's algorithm using preflows. While the theoretical bound is the same Fujishige [18], the algorithm is faster in practice.

Arc-Balancing Algorithm

Tarjan *et al.* [21] present a round-robin arc-balancing algorithm that computes a maximum flow in $O(n^2 m \log(nU))$ time. Although this algorithm is slower than other known algorithms, it is extremely simple. This algorithm maintains a *pseudoflow* rather than a preflow. Let x be a pseudoflow then $x_{ij} = -x_{ji}$ and $x_{ij} \leq u_{ij}$ for all arcs $(i, j) \in A$. Moreover, the balance at node i can be calculated as

$$b(i) = \sum_{(j,i) \in A} x_{ji},$$

where $b(i)$ can take both positive (excess) and negative (deficit) values.

Algorithm. Arc-balancing algorithm.

```

begin
    Construct an initial pseudoflow by
        saturating every arc out of  $s$  and every
        arc into  $t$ , and assigning zero flow to
        every other arc.
    while (sum of positive balances  $> 1$ ) and
        (sum of negative balances  $< -1$ ) do
        begin
            choose an arc  $(v, w)$  with  $b(v)$ 
                 $> b(w)$ ,  $w \neq s$ , and  $v \neq t$ ;
            increase the flow on  $(v, w)$  by
                 $\min\{r_{vw}, e(v)/2 - e(w)/2\}$ ;
        end
    end
end

```

Hochbaum's Pseudoflow Algorithm

Hochbaum [22] proposes a pseudoflow algorithm that runs in $O(mn \log n)$ time. Chandran and Hochbaum [23] show that the highest label pseudoflow implementation is faster than the preflow-push counterpart in most of the instances across different problem families. Hochbaum and Orlin [24] show that the highest label version of the pseudoflow algorithm has complexities $O(mn \log n^2/m)$ and $O(n^3)$ with and without the use of dynamic trees, respectively.

MAXIMUM FLOW ON SPECIAL GRAPHS

Frequently, it is possible to solve flow problems on special networks more efficiently than those defined on general capacity networks. In this section, we identify three special networks.

Flows in Unit Capacity Networks

Some combinatorial problems are naturally formulated as 0–1 optimization models. When viewed as flow problems, these models yield networks whose arc capacities are all 1, which are referred to as *unit capacity networks*. In this section, we describe *unit capacity maximum flow algorithm* [7].

In a unit capacity network, the maximum flow value is at most n since the capacity of the s – t cut $[s, S - s]$ is at most n . Therefore, general augmenting path algorithm determines a maximum flow within n

augmentations and requires $O(nm)$ effort. The shortest augmenting path algorithm also solves this problem in $O(nm)$ time since its bottleneck operation, which is the augmentation step, requires $O(nm)$ time instead of $O(n^2m)$ time. The unit capacity maximum flow algorithm that we describe is a hybrid version of these two algorithms

requires only $O(\min\{n^{2/3}m, m^{3/2}\})$ time, which is consistently better than the $O(nm)$ bound of either algorithm by itself. This algorithm runs in $O(n^{1/2}m)$ time for a simple network. A unit capacity network is called a *simple network* if either the number of incoming arcs or the number of outgoing arcs of each node is at most one.

Algorithm. Unit capacity maximum flow algorithm.

begin

 apply the shortest augmenting path algorithm until $d(s) \geq \min\{2n^{2/3}, m^{1/2}\}$;
 then apply the general augmenting path algorithm until establishing the maximum flow;

end

Flows in Bipartite Networks

A bipartite network is a network $G = (N, A)$ with a node set N partitioned into two subsets, N_1 and N_2 , so that for every arc $(i, j) \in A$, either $i \in N_1$ and $j \in N_2$, or $i \in N_2$ and $j \in N_1$. We often represent a bipartite network using the notation $G = (N_1 \cup N_2, A)$. Let $n_1 = |N_1|$ and $n_2 = |N_2|$.

In this section, we describe an adaptation of the preflow-push algorithms that we considered in the previous discussion in this article. The worst case behaviors of these special-purpose algorithms are similar to those of the original algorithms if the node sets N_1 and N_2 are of comparable size. Without any loss of generality, we

assume that $n_1 \leq n_2$ and the source node is in N_2 .

As an example of the type of results that we obtain, we show the specialization of the generic preflow-push algorithm. We refer to this algorithm as the *bipartite preflow-push algorithm*. The running time is $O(n_1^2m)$ time and if $n_1 < n_2$ then the new implementation is considerably faster than the original algorithm [24]. The ideas that we consider also apply in a straightforward manner to FIFO, highest label, preflow-push, and excess-scaling algorithms and yield algorithms with improved worst case complexity. We just include the push/relabel procedure since the remaining parts of the algorithms are the same.

Algorithm.

procedure *bipartite push/relabel*(i);

begin

if the residual network contains an admissible arc (i, j) **then**

if the residual network contains an admissible arc (j, k) **then**

 push $\delta = \min\{e(i), r_{ij}, r_{jk}\}$ units of flow over the path $i \rightarrow j \rightarrow k$ and augment x by δ ;

else replace $d(j)$ by $\min\{d(k) + 1 : (j, k) \in A \text{ and } r_{jk} > 0\}$;

else replace $d(i)$ by $\min\{d(j) + 1 : (i, j) \in A \text{ and } r_{ij} > 0\}$;

end

Flows in Planar Networks

Planar networks can be drawn on the plane so that no two arcs intersect each other. Itai and Shiloach [25] present an algorithm that needs $O(n \log n)$ computations for undirected planar networks.

This algorithm first constructs the dual network and solves the shortest path problem over it. The shortest path in the dual network yields a minimum cut in the original network and the shortest path distances yield a maximum flow. Borradaile and Klein [26] recently

Table 1. Summary of the Algorithms

Augmenting path algorithms		# Augment iterations	Time to find an augmenting path	Best running time
Generic [5,6]		$O(nU)$	$O(m)$	$O(nmU)$
Shortest augmenting path [7,8]		$O(nm)$	$O(\log n)$	$O(nm \log n)$
Capacity scaling [7,9]		$O(m \log U)$	$O(n)$	Dynamic tree implementation $O(nm \log U)$
Blocking flow algorithms		# Blocking flows	Time to find a blocking flow	Best running time
Blocking flow [10]		$O(n)$	$O(nm)$	$O(n^2m)$
Sleator and Tarjan [8]		$O(n)$	$O(m \log n)$	$O(nm \log n)$
Binary blocking flow [4]		$O(\min(n^{2/3}, \sqrt{m}) \log U)$	$O\left(m \log \frac{n^2}{m}\right)$	Dynamic tree implementation $O(\min(n^{2/3}, \sqrt{m})m \log \frac{n^2}{m} \log U)$
Preflow-push algorithms		Notes	# Nonsaturating pushes	Best running time
Generic [11]		The number of saturating pushes is $O(nm)$	$O(n^2m)$	$O(n^2m)$
13	FIFO [12]	Chooses the active nodes in first-in-first-out order	$O(n^3)$	$O(n^3)$
	Highest label [11] with dynamic trees	Chooses the active node with the highest label	$O(n^3)$	$O(nm \log(n^2/m))$
	Highest label [14]	Chooses the active node with the highest label	$O(n^2 \sqrt{m})$	$O(n^2 \sqrt{m})$
Excess scaling [16]		Chooses the large excess node with the smallest distance label	$O(n^2 \log U)$	$O(nm + n^2 \log U)$
Excess scaling with dynamic trees [17]		Extension of [16]	Wave scaling:	Wave scaling:
		Uses a scale factor (k) greater than 2	$O(n^2 \sqrt{\log U})$	$O(nm + n^2 \sqrt{\log U})$
		Chooses the large excess node with the highest distance label	Stack scaling: $O\left(\frac{n^2 \log U}{\log \log U}\right)$	Stack scaling: $O\left(nm + \frac{n^2 \log U}{\log \log U}\right)$
				Dynamic tree implementation runs in $O(nm \log(n \sqrt{\log U}/m + 2))$

(continued overleaf)

Table 1. (Continued)		
Other algorithms	Notes	Best running time
MPM [27]	—	$O(n^3)$
MA ordering [18]	Sorts nodes according to maximum adjacency Capacity scaling version Practically better than [11,27]	$O(n(m + n \log n) \log nU)$ $O(nm \log U)$ Preflow version [20] $O(n(m + n \log n) \log nU)$
Arc-balancing [21]	Uses pseudoflows	$O(n^2m \log (nU))$
Hochbaum's [22]	Uses pseudoflows	$O(mn \log n)$ Highest label [24] $O(mn \log n^2/m)$ Highest label with dynamic tree [24] $O(n^3)$

introduced an $O(n \log n)$ algorithm for the planar directed networks. A planar network with a source node s and a sink node t is called s - t planar if both nodes lie at the boundary of the unbounded face. There are also special algorithms for s - t planar networks.

SUMMARY OF ALGORITHMS

The majority of the best theoretical time bounds for maximum flow problems are based on the preflow-push type of algorithms: The algorithm proposed by Goldberg and Tarjan [11] runs in $O(nm \log(n^2/m))$ time; Ahuja *et al.* [17] runs in $O(nm \log(n/m\sqrt{\log U} + 2))$ time. Cheriyan *et al.* [3] have a random arc selection rule and run in $O(n^3/\log n)$ time and $O(nm + (n \log n)^2)$ time with high probability. King *et al.* [28] present the deterministic version of Cheriyan *et al.* [3] that runs in $O(nm + n^2 + \varepsilon)$ time for any constant $\varepsilon > 0$. The best preflow-push algorithms currently outperform the best augmenting path algorithms in theory as well as in practice (see, for example, Ahuja *et al.* [12]). Chandran and Hochbaum [23] show that the pseudoflow algorithms run faster than the preflow-push algorithms for most of the problem instances.

The presented algorithms for general networks are summarized in Table 1. For each algorithm, we reported the references, some highlights, and the decomposition of the running time. The algorithms with more than one references are improved over time; therefore we only report the best running time.

REFERENCES

- Ahuja RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms, and applications. Englewood Cliffs (NJ): Prentice Hall; 1993.
- Ford LR, Fulkerson DR. Flows in networks. Princeton (NJ): Princeton University Press; 1962.
- Cheriyan J, Hagerup T, Mehlhorn K. Can a maximum flow be computed in $O(nm)$ time. Proceedings of the 7th International Colloquium on Automata, Languages and Programming. Warwick, Warwick University. 1990; pp. 235–248.
- Goldberg AV, Rao S. Beyond the flow decomposition barrier. J ACM 1998;45(5):783–797. DOI: 10.1145/290179.290181.
- Ford LR, Fulkerson DR. Maximal flow through a network. Can J Math 1956; 8:399–404.
- Elias P, Feinstein A, Shannon CE. Note on maximum flow through a network. IRE Trans Inf Theory 1956;IT-2:117–119.
- Ahuja RK, Orlin JB. Distance-directed augmenting path algorithms for maximum flow and parametric maximum flow problems. Nav Res Log Q 1991;38:413–430.
- Sleator DD, Tarjan RE. A data structure for dynamic trees. J Comput Syst Sci 1983;26(3):362–391. DOI: 10.1016/0022-0000(83)90006-5.
- Edmonds J, Karp RM. Theoretical improvements in algorithmic efficiency for network flow problems. J ACM 1972;19:248–264.
- Dinic EA. Algorithm for solution of a problem of maximum flow in networks with power estimation. Sov Math Dokl 1970;11:1277–1280.
- Goldberg AV, Tarjan RE. A new approach to the maximum flow problem. Proceedings of the 19th ACM Symposium on the Theory of Computing. Berkeley (CA). 1986. pp. 136–146. DOI: 10.1145/12130.12144.
- Ahuja RK, Kodialam M, Mishra AK, *et al.* Computational investigations of maximum flow algorithms. Eur J Oper Res 1997;97:509–542. DOI: 10.1016/S0377-2217(96)00269-X.
- Cormen TH, Leiserson CE, Rivest RL, *et al.* Introduction to algorithms. 2nd ed. New York: McGraw-Hill; 2001.
- Cheriyan J, Maheshwari SN. Analysis of preflow-push algorithms for maximum network flow. SIAM J Comput 1989;18:1057–1086.
- Cherkassky BV, Goldberg AV. On implementing the push—relabel method for the maximum flow problem. Algorithmica 1997;19:390–410. DOI: 10.1007/PL00009180.
- Ahuja RK, Orlin JB. A fast and simple algorithm for the maximum flow problem. Oper Res 1989;37:748–759. DOI: 10.1287/opre.37.5.748.
- Ahuja RK, Orlin JB, Tarjan RE. Improved time bounds for the maximum flow problem. SIAM J Comput 1989;18:939–954.
- Fujishige S. A maximum flow algorithm using MA ordering. Oper Res Lett 2003;31: 176–178.

19. Shioura A. The MA-ordering max-flow algorithm is not strongly polynomial for directed networks. *Oper Res Lett* 2004;32(1):31–35. DOI: 10.1016/S0167-6377(03)00070-1.
20. Matsuoka Y, Fujishige S. Practical efficiency of maximum flow algorithms using MA orderings and preflows. *J Oper Res Soc Jpn* 2005;48(4):297–307.
21. Tarjan R, Ward J, Zhang B, *et al.* Balancing applied to maximum network flow problems. In: Azar Y, Erlebach T, editors. Volume 4168, ESA 2006. LNCS. Heidelberg: Springer; 2006. pp. 612–623.
22. Hochbaum DS. The pseudoflow algorithm: a new algorithm for the maximum flow problem. *Oper Res* 2008;56(4):992–1009.
23. Chandran BG, Hochbaum DS. A computational study of the pseudoflow and push-relabel algorithms for the maximum flow problem. *Oper Res* 2009;57(2):358–376. DOI: 10.1287/opre.1080.0572.
24. Hochbaum DS, Orlin JB. The pseudoflow algorithm in $O(mn \log(n^2/m))$ and $O(n^3)$. Berkeley: University of California; 2007.
25. Itai A, Shiloach Y. Maximum flow in planar networks. *SIAM J Comput* 1979;8:135–150.
26. Borradaile G, Klein P. An $O(n \log n)$ algorithm for maximum *st*-flow in a directed planar graph. *J ACM* 2009;56(2):1–30. DOI: 10.1145/1502793.1502798.
27. Malhotra VM, Pramodh Kumar M, Maneshwari SN. An $O(V^3)$ algorithm for finding maximum flows in networks. *Inf Process Lett* 1978;7:277–278.
28. King V, Rao S, Tarjan RE. A faster deterministic maximum flow algorithm. *Proceedings of the 3rd ACM-SIAM Symposium on Discrete Algorithms*. 1992. pp. 157–164. DOI: 10.1006/jagm.1994.1044.

MDP BASIC COMPONENTS

JONATHAN PATRICK
Telfer School of Management,
University of Ottawa,
Ottawa, Ontario, Canada

THE BASICS OF MARKOV DECISION PROCESSES

The article entitled, *Historical Development of MDP Theory*, represented an overview of how Markov decision process (MDP) theory evolved over the last century. From those initial forays at such places as the RAND Corporation in Santa Monica, California, MDP models have found applications in any number of research areas including medical decision making [1,2], health care management [3,4], forestry [5], fisheries [6] and business [7–9]. The types of problems solved include scheduling [4], capacity planning [5], revenue management [7–9], supply chain and policy analysis.

THE FIVE COMPONENTS OF AN MDP MODEL

MDP models are mathematical methods for solving sequential decision problems where the decision taken today impacts what actions are possible tomorrow. MDP models are divided into numerous categories depending on the length of the decision process (infinite or finite horizon) and whether decisions are taken at regular intervals (discrete time models) or at random times, the length of which is triggered by some event (continuous time models). A description of the five major components of an MDP model: decision epochs, states, actions, transition probabilities, and rewards (or costs) is provided below. They allow the modeler to formulate a wide array of sequential decision problems that can then be solved to find the

optimal decision policy on the basis of the optimization methods developed in other articles within this encyclopedia.

Decision Epochs

The simplest form of an MDP is a *finite horizon, discrete time* model. In such a model, a finite number of decisions are taken at fixed, regular intervals. The time at which a decision is taken is called a *decision epoch*. In this article, the letter N is used to represent the number of decision epochs. The article entitled *Finite Horizon MDPs* deals with the methods for solving these *finite horizon discrete time* problems. If N is allowed to go to infinity so that there are an infinite number of decision epochs, then the methods for solving these problems become more complex but the basic form of the MDP as described in this article remains essentially the same. (See the articles entitled *Total Expected Discounted Reward Criterion*, *Total Expected Reward Criterion*, and *Average Reward Criterion* for methods for solving *infinite horizon discrete time* problems.) One of the most well-known discrete time MDP models involves solving the inventory management problem. Many variations of this problem are possible, but to illustrate the basic MDP concepts, consider a simple version. In this problem, a manager is in charge of managing the inventory of a store that sells a single product type. The store earns revenue by selling the product to customers. Demand for the product is assumed to be variable. The inventory level decreases when the items are sold to customers, and backorders are not allowed; that is, if there is insufficient inventory to meet demand then the excess demand is lost. The manager replenishes the inventory by deciding when to order new inventory from a single supplier and how much to order. There is a cost associated with placing an order and a cost associated with holding inventory. The manager's goal is to make inventory decisions in a way that maximizes the store's net profit. Thus, decisions are made at discrete intervals

(each day at 9 a.m. for instance) and could potentially be framed as a finite horizon MDP (i.e., if stocking for a season) or as an infinite horizon MDP (if running a business for the foreseeable future).

Alternatively, there are other applications where it does not make sense to assume that decisions are made at regular intervals. Suppose, for instance, that a manager is in charge of insuring that a queue does not build up at a given workstation in a factory. It is assumed that the rate at which products are serviced is under his control. (He may be able to increase the speed of service or else introduce additional servers.) Usually there is a cost associated with increasing the rate of service and a cost associated with each customer waiting in the queue. Thus, there is a balance to be struck between keeping the queue small and not incurring too much cost for running at a speedier service rate. In such a case, it would make sense to reevaluate after each new product arrival and after each new service completion. But these are random events and thus the decision epochs do not occur at fixed intervals. Such applications are modeled as *continuous time MDPs* (see the article entitled *Continuous-time MDPs*).

States. Clearly, in order to solve a decision problem, the model must have some means of tracking the information necessary for making an informed decision. For instance, if one is trying to solve the inventory management problem then it is paramount that the current inventory level be known each time a decision is made. The *state* of the system at decision epoch t , often denoted by s_t , is the current value of all relevant information required in order to make an informed decision. We need only concern ourselves with information that can change from decision epoch to decision epoch as unchanging information can be left as input to the model. The *state space*, S_t , is the set of all possible values of the state variable(s) at time t . Thus, in the inventory management problem, the state space might be $S_t = 0, 1, \dots, M$, where M is the maximum number of items that the store can hold. (If backorders are allowed, then the state space would expand to include the negative integers in order to represent a backlog of demand.)

In this example, the state space is a finite set of integers. However, in other instances, the state space might be a continuous range (i.e., the level of water in a reservoir) or an infinite set of discrete numbers (i.e., clients waiting for service in a system with an infinite buffer).

More complicated states and state spaces that involve more than one state variable are also possible. For instance, in advanced appointment scheduling, the state space must include the number of available appointment slots in each day of the booking horizon as well as the number of clients waiting to be booked. Such a complicated state space quickly runs into the *curse of dimensionality* that renders the optimization methods developed in the articles mentioned earlier intractable. This curse refers to the fact that as the dimensions in the state vector increases, the size of the state space grows dramatically. Traditional MDP solutions present a look-up table that gives the optimal action to take in every state and at every decision epoch. If the state space is too large, this is simply not feasible. The article entitled *Large-Scale MDPs* deals with methods for approximating a solution to these larger and more complex MDP models.

Actions. The third component of an MDP model is the set of *actions* available to the user. Clearly, what actions are available will most often depend on the state of the system. For instance, if the inventory level is at $M-1$ then the only action available to the manager is not to order any new items or to order only one. The *action space* at decision epoch t is the set of available actions for the current state, s_t , and is usually denoted by $A_t(s_t)$. The action chosen at decision epoch t is denoted by a_t . Again, the action space need not be a finite set of integer values as in the inventory example. It could equally be a continuous range (i.e., how much additional water to remove from a reservoir) or an infinite set of discrete numbers (i.e., the number of waiting clients to refuse service in a system with an infinite buffer).

Transition Probabilities. The fourth component required for solving a sequential decision problem is how the system evolves

from one decision epoch to the next. These are called the *transition probabilities*. In the inventory management problem, the state at the next decision epoch will equal the current inventory level plus any new orderings minus any demand that arrives between now and the next decision epoch. As no backorders are allowed (by assumption), the new inventory level will be zero if demand exceeds the current inventory plus newly ordered items. Notice the further assumption that there is no lag time in delivery. While including a lag time is quite possible, it complicates the model and thus makes it less useful for illustrative purposes. The value of the next state can be written formally as

$$s_{t+1} = \max\{s_t + a_t - D_t, 0\},$$

where D_t is a random variable representing the demand that arrives between decision epochs t and $t + 1$. The above equation is an example of a state transformation function. The transition probabilities would therefore depend on the arrival distribution of new demand as well as the current state and action. They are often denoted by $p_t(s_{t+1}|s_t, a_t)$, which represents the probability of being in state s_{t+1} at decision epoch $t + 1$ given that at decision epoch t the system was in state s_t and action a_t was taken.

Rewards or Costs. Finally, if a decision process is to be optimized then there clearly needs to be some form of *reward* or cost that determines what actions are better than others. This is the fifth and final piece of an MDP model. In the most general setting, these rewards might depend on the current state of the system, the action taken and even the state of the system at the next decision epoch and are often denoted by $r_t(s_t, a_t, s_{t+1})$. Returning once again to the inventory example, the reward function should include a holding cost, $h()$, for storing inventory, an ordering cost, $O()$, and a revenue function, $f()$, for any sales.

Formally, this can be written as

$$r_t(s_t, a_t, s_{t+1}) = -O(a_t) - h(s_t + a_t) + f(s_t + a_t - s_{t+1}^+)$$

A typical ordering cost function would include a fixed cost for any size order placed as well as a variable cost as the size of the order increases. Since the state of the system at the next decision epoch is generally not known in advance, the typical MDP works with the expected reward based on the current state and action. This is given by

$$r_t(s_t, a_t) = \sum_{s_{t+1} \in S} r(s_t, a_t, s_{t+1}) p_t(s_{t+1}|s_t, a_t).$$

In a finite horizon model, a salvage function, $g(s_N)$, denotes the value of being in state s_N at the end of the planning horizon. It is assumed that there is no action to be taken at the final decision epoch. Most inventory management problems are more reasonably modeled as infinite horizon problems; however, there are some instances (seasonal products) where a finite horizon is more logical.

The reward structure is somewhat more complicated in continuous time models as the length of time between decision epochs will vary (possibly impacting on the reward or cost). The reader is referred to the article entitled *Continuous-Time MDPs* for a discussion of the differences.

A Markov process is said to be *stationary* if both the reward functions and the transition probability matrices are the same for every decision epoch; that is $r_t(s_t, a_t) = r(s_t, a_t)$ and $p_t(s_{t+1}|s_t, a_t) = p(s_{t+1}|s_t, a_t)$ for all states s_{t+1} and s_t and all actions a_t . If this is not the case, then the Markov process is said to be *non-stationary*. Infinite horizon models assume that the Markov process is stationary.

DECISION RULES AND POLICIES

The above five components formally define an MDP model. The next step is to determine the optimal decision policy. A *decision rule*, d_t , is defined as a function that dictates what decision to take for each possible state at decision epoch t . Decision rules may

be *history dependent* if the decision taken depends on the entire history up to that decision epoch or they may be *Markovian* if they depend only on the current state of the system. Decision rules may also be *deterministic* (if they dictate only one given action for each state) or they may be *randomized* (if they give a probability distribution over the set of possible actions that is used to randomly determine the appropriate action). History dependent policies are extremely difficult to analyze (largely because the state vector quickly becomes too large for tractability if the entire history needs to be included) and can often be avoided by properly structuring the state space. Moreover, under mild assumptions, it can be shown [10] that there exists a deterministic decision rule that does at least as well as any randomized one. Thus, the focus is commonly on deterministic Markovian decision rules, $d_t : S_t \rightarrow A_t(s_t)$, that map each possible state to a specific feasible action.

A policy, $\pi = (d_1, d_2, d_3, \dots, d_N)$, is a set of decision rules (one for every decision epoch) that dictates what action to take for any given state and at any given decision epoch. A policy is *Markovian* if the decision rule for each decision epoch is *Markovian* and *deterministic* if each decision rule is *deterministic*. Since it is possible, under mild assumptions, to show that there always exists an optimal policy that is Markovian and deterministic (given the objectives discussed below) [10], the rest of the article restricts discussion to this subset of all possible policies.

A policy is said to be *stationary* if, for each decision epoch, the same decision rule is used [i.e., $\pi = (d, d, d, \dots, d)$]. In finite horizon models, the optimal policy is rarely stationary as the best action invariably changes as the end of the horizon approaches. Consider again the inventory example. As the season for a given product approaches its end, it is clearly advantageous to carry less inventory so as not to be left at the end of the season with significant inventory that usually is sold off for a reduced price or may even have to be absorbed as a loss. However, in the infinite horizon setting, it is intuitively reasonable to assume that a stationary policy will in fact be optimal as there is no

substantive difference between the decision process at decision epochs t and $t + 1$ provided the horizon is infinite and the rewards and transitions are independent of t . Thus, for infinite horizon models, the set of stationary policies plays an important role.

DEFINING THE APPROPRIATE GOAL

The natural goal is to find the policy, π^* , that gives the greatest total reward over the length of the horizon. Formally, the total reward of a policy π over a finite horizon of length N and starting at state s is defined as

$$v_N^\pi(s) = E_s^\pi \left\{ \sum_{t=1}^N r_t(X_t, Y_t) + r_N(X_N) \right\},$$

where X_t is a sequence of random variables representing the progression of the state and Y_t is a sequence of random variables representing the actions taken at each decision epoch. Actions are determined using the policy π . v_N^π is called the *value function* of the policy π . (In the infinite horizon setting, the subscript N is dropped.) Ideally, one would like to find a policy π^* such that

$$v_N^{\pi^*}(s) \geq v_N^\pi(s),$$

for all starting states s and for all alternative policies π . This would seem to be a fairly daunting task, especially if the planning horizon is long. The genius of MDP theory was to realize that such a problem could be solved one decision epoch at a time rather than trying to solve the whole decision problem at once. The process is called backward induction (see the article entitled *Finite Horizon MDPs*) as it starts in the last period in which there is a decision, period $N-1$, and moves back to period 1.

However, in the infinite horizon setting, the total expected reward often turns out to be problematic due to the fact that the sum of rewards is now infinite and therefore

1. may not even exist;
2. if it does exist, may have no policy that achieves the maximum; and

3. may be infinite for a number of different policies that are not all equally good policies. (For instance, it would be infinite both for a policy that yields a reward of \$1 for every decision epoch and one that yields \$100 for every decision epoch.)

One of two solutions is generally employed when the total expected reward is inappropriate. The first solution introduces a discount factor that causes rewards to be valued less highly the further into the future they are realized and finds the policy that gives the greatest total *discounted* reward over the infinite horizon (see the article entitled *Total Expected Discounted Reward Criterion*). This objective has the advantage of being guaranteed to exist and favors policies with upfront rewards. The second solution is to find a policy with the greatest *average* reward per decision epoch instead (see the article entitled *Average Reward Criterion*). Which objective is employed depends on what best suits the application.

THE FORM OF AN OPTIMAL POLICY

One of the benefits of MDP theory is that, in some instances, it is possible to prove that the optimal policy has a certain form independent of the actual parameter values. For instance, in the inventory management problem with no backlogs and over an infinite horizon, it has been proven that the optimal policy is an (s, S) policy. Such a policy places an order as soon as the inventory level drops below s and orders enough to make the post-decision inventory equal to S . This form of the policy holds true for a whole range of holding and ordering costs as well as transition probabilities. Such structural results provide useful insight into a wide variety of sequential decision problems, are often easy to implement, and can even help reduce the computational burden associated with finding an optimal policy. Such results are discussed in the article entitled *Structured MDP Policies*.

CONCLUSION

The above formulation of an MDP model allows for significant flexibility and therefore covers a broad range of potential instances of sequential decision problems. From scheduling patients in health care, to planning capacity in airlines or managing inventory in retail, the potential of MDP theory to improve practice is undeniable. Other articles in this encyclopedia will introduce the reader to a number of optimization methods for solving finite or infinite horizon, discrete or continuous time models. Further, the reader will be introduced to methods for dealing with instances where the state information is only an approximation of the true state (see the article entitled *Partially Observable MDPs*) and with more complex methodologies that seek to overcome some of the persistent limitations of the standard methods for solving MDPs (see the article entitled *Large Scale MDPs*). For more in-depth studies of MDP theory, the reader is referred to Richard Bellman's classic book *Dynamic Programming* [11] as well as more recent works by Puterman [10], Bertsekas [12] and Feinberg [13].

REFERENCES

1. Schechter S, Bailey M, Schaefer A, *et al.* The optimal time to initiate HIV therapy under ordered health states. *Oper Res* 2008;56(1):20–33.
2. Alagoz O, Maillart L, Schaefer A, *et al.* The optimal timing of living donor liver transplantation. *Manage Sci* 2004;50(10):1420–1430.
3. Green L, Savin S, Wang B. Managing patient demand in a diagnostic medical facility. *Oper Res* 2006;54(1):11–25.
4. Kolesar P. A Markovian model for hospital admission scheduling. *Manage Sci* 1970;16(6):384–396.
5. Buongiorno J. Generalization of Faustmann's formula for stochastic forest growth and prices with Markov decision process models. *Forest Sci* 2001;47(4):466–474.
6. Lane DE. A partially observable model of decision making by fishermen. *Oper Res* 1989;37:240–254.

7. Bertsimas D, Popescu I. Revenue management in a dynamic network environment. *Transp Sci* 2003;37(3):257–277.
8. Bertsimas D, Shioda R. Restaurant revenue management. *Oper Res* 2003;51(3):472–486.
9. Brumelle S, Walczak D. Dynamic airline revenue management with multiple semi-Markov demand. *Oper Res* 2003;51(1):137–148.
10. Puterman ML. Markov decision processes. New York: John Wiley & Sons, Inc.; 1994. p. 649.
11. Bellman R. Dynamic programming. Princeton (NJ): Princeton University Press; 1957. p. 342.
12. Bertsekas D. Dynamic programming and optimal control. 3rd ed. Belmont (MA): Athena Scientific; 2000. p. 558.
13. Feinberg E, Shwartz A. Handbook of Markov decision processes. New York: Kluwer; 2002. p. 580.

MEASURES OF RISK EQUITY

L. ROBIN KELLER

YITONG WANG

The Paul Merage School of
Business, University of
California Irvine, Irvine,
California

TIANJUN FENG

Fudan University, School of
Management, Shanghai,
China

This article covers measures of fairness or equity for situations involving risk, as represented with a probability distribution over outcomes, which often include adverse health or safety outcomes. For an introduction to fairness, the reader is referred to Keller *et al.* [1]. Such risk equity measures may be used in societal utility functions to identify preferred policies from the perspective of what is good for society as a whole. The aim of this article is to present enough detail on this topic so that a person can understand the basic concepts and find further references to get more details; this is not a complete compilation of all the research in this field.

Suppose two policies are being considered to improve public safety, and only one policy can be funded.

Safety Policy Scenario

Policy A: Install emergency phones along roadsides; this will save 10 lives annually.

Policy B: Install a flood warning system, with a 50% chance of saving 20 lives annually (when flood occurs and assume one flood only in that year) and a 50% chance of saving no lives (when no flood occurs).

One way of evaluating such alternative societal risk policy options might be to calculate expected fatalities, then choose the

policy with the maximum number of lives saved. However, in this case, the expected number of lives saved is 10 for both policies.

To go beyond using the objective of maximizing expected number of lives saved to make this choice, one could construct a societal utility function over the number of lives. The utility of policy A is $u(A) = u(10 \text{ lives saved})$. The utility of policy B is $u(B) = 0.50u(20 \text{ lives saved}) + 0.50u(0 \text{ lives saved})$. A risk-averse utility function would favor the choice of policy A, with the sure number of 10 lives saved. A risk-prone utility function would favor the choice of policy B.

While specifying a risk-averse or risk-prone utility function in this scenario will lead to a clear choice in this case, more generally one may wish to consider how these fatalities are distributed among the members or groups of society. In policy A, 10 lives are saved each year, from separate roadside incidents. In policy B, if there is a flood during a year, 20 lives will be saved from that one flood, and in nonflood years, no lives will be saved.

Keeney and Winkler [2] point out that

In evaluating public risks, it may be useful to separate three issues:

1. The undesirability of fatalities, aside from equity considerations.
2. *Ex post* equity, meaning the equity associated with the fatalities that actually occur.
3. *Ex ante* equity, meaning the equity of the process and risks which eventually lead to the fatalities.

See the following sections titled “*Ex Ante* Equity,” “*Ex Post* Equity,” “Catastrophe Avoidance,” and “Envy-Free Allocation,” for details. The section titled “Decision Models Containing Equity Measures” presents equity models and applications. For other papers on the use of utility functions to represent preferences regarding societal equity, see Keeney [3–5], Broome [6], Fishburn [7], and Sarin [8].

EQUITY AXIOMS

Individual Equity

Suppose there are two people at potential risk. (We often begin to examine equity measures by simplified scenarios involving just two people, and then extend the concepts to n people.) The concept of *individual equity* is that *society is indifferent if the two people switch their risk of being the sole person to die and everything else stays the same.*

This preference is contained in Fishburn's [7] *individual equity axiom E1*. To set up the notation for his axioms, which are stated for just two people, denoted as 1 and 2, he first identifies the four basic consequences and the chances of each:

- p chance that
(Person 1 lives, Person 2 dies),
- q chance that
(Person 1 dies, Person 2 lives),
- r chance that (Person 1 lives,
Person 2 lives) = both live, and
- $1 - (p + q + r) = \#$ chance that
(Person 1 dies, Person 2 dies) = both die.

We will arbitrarily scale the societal utility function with two end points: Suppose that the certainty (i.e., the probability $r = 1$) that both live has utility 2 and the certainty (i.e., the probability $1 - (p + q + r) = 1$) that both die has utility 0. Denote the utility of person 1 living alone as a_1 and the utility of person 2 living alone as a_2 , with both a_1 and a_2 between 0 and 2.

Having set the utility of both people dying as 0 (the minimum possible utility) and the utility of both people living as 2 (the maximum possible utility), there are only two remaining utilities to specify. They are

a_1 = the utility of person 1 living and
person 2 dying

a_2 = the utility of person 1 dying and
person 2 living.

These utilities (a_1 and a_2) might be equal to each other or unequal. They might both

equal 1, or both be under 1 or above 1, and so on. Different equity axioms lead to specific possible ranges of values for the utilities of these two intermediate outcomes.

For each decision scenario, we want to represent the societal utility for a policy which will result in the four basic consequences with some probability of each. This societal utility can be conveniently written, using the probabilities of the four basic consequences, and suppressing the detailed description of the consequences, as

$$u(p, q, r, 1 - (p + q + r)).$$

(We will use this short-hand notation whenever we present Fishburn's axioms in the sections on equity axioms.)

So, we compute the vonNeumann–Morgenstern expected utility of a policy as follows:

$$\begin{aligned} u(\text{policy}) &= u(p, q, r, 1 - (p + q + r)) \\ &= pu(\text{Person 1 lives, Person 2 dies}) \\ &\quad + qu(\text{Person 1 dies, Person 2 lives}) \\ &\quad + ru(\text{Person 1 lives, Person 2 lives}) \\ &\quad + (1 - (p + q + r))u(\text{Person 1 dies,} \\ &\quad \text{Person 2 dies}) \\ &= p(a_1) + q(a_2) + r(2) + (1 - (p + q + r))0 \\ &= a_1p + a_2q + 2r \end{aligned}$$

Different equity axioms result in different possible values of $u(\text{Person 1 lives, Person 2 dies}) = a_1$ and $u(\text{Person 1 dies, Person 2 lives}) = a_2$.

Using our notation, the assumption of *individual equity* is that

$$u(p, q, r, \#) = u(q, p, r, \#),$$

where $\#$ represents the probability of the fourth consequence = $1 -$ the sum of the first three probabilities, which, on the left side of the equation is $1 - (p + q + r)$.

Under this individual equity axiom, the utility of the left-hand side is $a_1p + a_2q + 2r$. The utility of the right-hand side is $a_1q + a_2p + 2r$. They are equal, so

$$a_1p + a_2q + 2r = a_1q + a_2p + 2r.$$

Subtracting $2r$ from both sides,

$$a_1p + a_2q = a_1q + a_2p.$$

Rearranging, we get

$$a_1(p - q) = a_2(p - q),$$

and thus

$$a_1 = a_2.$$

Ex Ante Equity

Ex ante equity can be seen “before the fact” of a risk event being resolved. A person who prefers *ex ante* equity will choose an equal distribution of risk, where risk is defined by the probability of an individual becoming a fatality, over an unequal distribution.

Consider the scenario below.

Ex Ante Equity Scenario

Alternative A: 50% chance that Person 1 lives, Person 2 dies

50% chance that Person 1 dies, Person 2 lives.

Alternative B: Person 1 lives,
Person 2 dies.

In Alternative B, one person dies for sure, and we know that it will be person 2. In Alternative A, one person dies for sure, also. So, if a societal decision maker wishes to use a von Neumann–Morgenstern utility function defined over number of fatalities (only), then the model will be indifferent between the two alternatives.

Before the uncertainty in Alternative A is resolved, each person has an equal 0.5 chance of death, so they share in facing the risk. Alternative B has an unequal distribution, with person 1 having a probability of 0 of death and person 2 having a probability of 1. A decision maker who prefers Alternative A, because it “is fairer since each person has an equal chance of surviving,” is demonstrating a preference for *ex ante* equity.

This seemingly reasonable assumption has been debated in the literature and

has caused some to conclude that the decision analytic approach (when considering only the number of fatalities rather than the distribution of risk) cannot satisfactorily address this equity issue; see Broome [6].

Risk-Sharing Equity. To generalize the notion of *ex ante* equity from the scenario above, recall Fishburn’s notation introduced in the section titled “Individual Equity,” where p and q refer to the probabilities:

p chance that (Person 1 lives, Person 2 dies)

q chance that (Person 1 dies, Person 2 lives).

According to Fishburn’s [7] *risk-sharing equity axiom E2*,

if $p + q = p^* + q^*$ and $|p - q| < |p^* - q^*|$ then

$$u(p, q, r, \#) > u(p^*, q^*, r, \#).$$

In the scenario above, $p = q = 0.5$ in Alternative A, and in Alternative B we have $p^* = 1$ and $q^* = 0$, so according to the risk-sharing equity axiom, $u(A) > u(B)$.

Independence Risk Equity. Next, we consider a preference for equal *independent* probabilities of death. The *independent risk equity axiom E3* by Fishburn [7] considers the two person’s *ex ante* independent probabilities of living (rather than dying), and states it is better if the probabilities are closer together (rather than farther apart). Let α be the probability that person 1 lives and β be the probability that person 2 lives. Each can live alone or with the other person.

Consider Option 1 with the following probabilities and outcomes:

$\alpha(1 - \beta)$ p chance that Person 1 lives,

Person 2 dies,

$(1 - \alpha)\beta$ q chance that Person 1 dies,

Person 2 lives,
 $\alpha\beta$ r chance that Person 1 lives,
 Person 2 lives, i.e. both live, or
 $\#$ $1 - (p + q + r)$ chance that
 Person 1 dies, Person 2 dies, i.e.
 both die.

Person 1 can live alone with the probability $\alpha(1 - \beta)$, or can live with Person 2 with the probability $\alpha\beta$. So, person 1 has a probability of living $= \alpha = \alpha(1 - \beta) + \alpha\beta$ and Person 2 has a β probability of living, where $\beta = (1 - \alpha)\beta + \alpha\beta$.

Option 2 has the same structure, but α and β are replaced with α^* and β^* .

Independence Risk Equity Axiom.

If $\alpha + \beta = \alpha^* + \beta^*$ and
 $|\alpha - \beta| < |\alpha^* - \beta^*|$, then
 $u(\text{Option 1}) = u(\alpha(1 - \beta), (1 - \alpha)\beta, \alpha\beta, \#)$
 $> u(\text{Option 2}) =$
 $u(\alpha^*(1 - \beta^*), (1 - \alpha^*)\beta^*, \alpha^*\beta^*, \#).$

So, a more even division of the unconditional probabilities of living is preferred. This is the same as Keeney's [5] risk equity assumption under independence between the unconditional events "1 dies" and "2 dies." Under this axiom, $u(\text{Person 1 lives, Person 2 dies}) = a_1 = u(\text{Person 1 dies, Person 2 lives}) = a_2 < 1$.

Consider the scenario below given by Keller and Sarin [9] to Americans and by Bian and Keller [10] to Chinese.

Serum Distribution Scenario. There are 100 islanders who are susceptible to a specific fatal disease which has recently appeared on the mainland. Scientists have identified a kind of serum which has the potential of protecting people from contracting the disease. *Unfortunately, there is not enough serum available to give all the susceptible islanders a high enough dose to successfully prevent the disease.* Action must be taken immediately to protect the public health.

All susceptible people must be injected with the serum within 24 hours, or each will have a 15% chance of contracting the disease and eventually dying. There is no time to acquire more serum. There are only 3000 milligrams of the serum available. As the public health officer, it is your job to choose between the following options. Circle your choice.

- A. Give the same low dose of 30 milligrams of serum to all 100 susceptible islanders. 50 of those susceptible are northerners, 50 are southerners. Each susceptible person will have an independent 10% chance of dying. The expected number of deaths is 10. (Considered more fair and chosen by most Americans and Chinese)
- B. Divide up the available serum among the 50 northerners who are susceptible to the disease. Thus, these people will receive a higher 60 milligram dose. Each of the 50 will now have an independent 5% chance of dying. Since the 50 susceptible southerners will receive none of the serum, each will still have a 15% chance of dying by contracting this disease. The expected number of deaths is 10.

Preference for equal independent probabilities of death in option A reveals a preference for independence risk equity. It is clear from the survey (in which half the subjects reported which option was fairer and the other half reported their chosen action) that the subjects' preference for option A has resulted from it being a fairer choice; however, technically speaking, such a choice can be explained by assuming that their utility functions defined over number of fatalities are risk prone. See the section titled "Catastrophe Avoidance" for a discussion of utility function shapes.

Measure of Ex Ante Equity Φ . There are different ways to measure *ex ante* equity quantitatively. A specific measure of *ex ante* equity Φ is one that penalizes unevenness across groups in terms of the average probability of death $\bar{p}_i$ for a person in group i , $i = 1$ to m [9]. Presuming there are roughly equal numbers of people in each group, a special

form for Φ is defined by

$$\Phi = - \sum_{i=1}^m |\bar{p}_i - \bar{p}|,$$

where the average of the probabilities of death for the m groups is $\bar{p} = \sum_{i=1}^m \bar{p}_i / m$. For the simple *ex ante* equity scenario above at the beginning of the section titled “*Ex Ante Equity*,” there are two groups, each having one person in it.

For Alternative A,

$$\Phi = -(|0.5 - 0.5| + |0.5 - 0.5|) = 0,$$

For Alternative B,

$$\Phi = -(|1 - 0.5| + |0 - 0.5|) = -1.$$

So, Alternative A has the higher (and presumably better) *ex ante* equity. See the section titled “Fair-Risk Model, Incorporating *Ex Post* and *Ex Ante* Equity Preferences” for a societal utility function which includes Φ .

For another example of preference for *ex ante* equity, see the rescuer at risk scenario in ***Fairness and Equity in Societal Decision Analysis***.

Ex Post Equity

A person who prefers *ex post* equity will choose an equal *ex post* distribution of final outcomes for the affected people, instead of an unequal distribution [7].

Consider the scenario below. First note that, in terms of *ex ante* equity, both alternatives are equal, since both lead to an *ex ante* probability of 0.5 that Person 1 dies and 0.5 that Person 2 dies. Thus, a decision maker cannot distinguish between the two alternatives in terms of *ex ante* equity. Now examine the alternatives below in terms of *ex post* equity.

Ex Post Equity Scenario

Alternative C: 50% chance that Person 1 lives, Person 2 lives
50% chance that Person 1 dies, Person 2 dies

Alternative A: 50% chance that Person 1 lives, Person 2 dies
50% chance that Person 1 dies, Person 2 lives

After the uncertainty is resolved in Alternative C, the *ex post* outcome is either both people will be dead or both people will be alive. In Alternative A, the *ex post* outcome will be that one person is alive and one is dead. So, Alternative C will be chosen by a person preferring *ex post* equity.

Common-Fate Equity. Common-fate equity is a typical kind of *ex post* equity. Note that in Alternative C above, after the fact of the risk event occurring (or not) the two people will share a common fate (either both living or both dying), no matter what happens. In Alternative A, they will have different fates, no matter what happens. In many situations, people will prefer to share a common fate. But, in some situations, such as when top executives of a company or parents of young children take air flights, they may take different flights to avoid the common fate of dying in the same airplane crash. In both of these situations, they are thinking of their responsibilities to their company or family.

Fishburn’s [7] *common-fate equity* axiom E4 states

$$u(p - \delta, q - \delta, r + \delta, \#) > u(p, q, r, \#).$$

In this axiom, it is an improvement to move the probability up that the two people share common fates (which occurs in consequence 3: both living, and consequence 4: both dying) and move the probability down that they have unequal fates (in consequence 1: only Person 1 lives and consequence 2: only Person 2 lives). Under this axiom, $u(\text{Person 1 lives, Person 2 dies}) + u(\text{Person 1 dies, Person 2 lives}) = a_1 + a_2 < 2$.

Gajdos *et al.* [11] present a new method for modeling preference for (or against) what they call “shared destinies,” which weakens a necessary and sufficient condition by Fishburn and Straffin [12] for two societal risk distributions to be judged to be indifferent whenever their associated distributions

of risk of death for individuals and for the number of fatalities are the same.

For an example of common-fate equity, see the miner location scenario discussed in Keller and Sarin [9]; see also Section 3 in ***Impossibility Theorems and Voting Paradoxes in Collective Choice Theory***.

Also, in the section titled “Catastrophe Avoidance,” Alternative D has common-fate equity, with either all 100 dying or all 100 living.

Common-Fate Independence. Fishburn’s [7] *common-fate independence* axiom E5 states

$$\begin{aligned} \text{If } p + r = p^* + r^* \text{ and } q + r = q^* + r^*, \\ \text{then } u(p, q, r, \#) = u(p^*, q^*, r^*, \#). \end{aligned}$$

In this axiom, it does not matter how a person dies, as long as the probability of dying remains the same. As long as a person’s probability of dying is constant, the probability can be split in any division between dying alone and dying together. This is a different kind of *ex post* equity. Under this axiom, $u(\text{Person 1 lives, Person 2 dies}) + u(\text{Person 1 dies, Person 2 lives}) = a_1 + a_2 = 2$.

Ex Post Equity Measure θ . As one example of an *ex post* equity measure, Keller and Sarin [9] suggest the following equity measure, θ , which depends on the distribution of the number of fatalities among the m groups. Suppose there are n_i fatalities in group i , $i = 1$ to m . Then

$$\theta = \sum_{n_m=0}^{N_m} \sum_{n_{m-1}=0}^{N_{m-1}} \dots \sum_{n_1=0}^{N_1} \pi(n_1, n_2, \dots, n_m),$$

where π is the joint probability mass function that can be computed from the estimates of risks to members in each group. N_i is the maximum possible number of fatalities in group i , $i = 1$ to m .

If $\bar{n}$ is the expected number of fatalities per group and each group experiences exactly $\bar{n}$ fatalities, there is no *ex post* inequity, from the perspective of a comparison across groups. So, any deviation from $\bar{n}$ is an

indication of *ex post* inequity, especially when the size of the population in each group is approximately the same or is substantially larger than the number of fatalities actually experienced. In this case, Keller and Sarin [9] define the *ex post* equity of a specified number of fatalities n_i in each group i as

$$\theta(n_1, n_2, \dots, n_m) = - \sum_{i=1}^m (n_i - \bar{n})^2.$$

For the simple *ex post* equity scenario above at the beginning of the section titled “*Ex Post* Equity,” there are two groups ($i = 1$ and 2), each having one person in it. The data to compute θ are organized in the following table for a generic policy alternative.

In each row in the table, multiply the probability from column 1 times $\theta(n_1, n_2)$ in the last column, then sum up the four products to get the overall *ex post* equity measure θ .

In the *ex post* equity scenario,

Alternative C has $p = 0$, $q = 0$, $r = 0.5$ and $\# = 0.5$, so $\theta = 0.5(0) + 0.5(0) = 0$.

Alternative A has $p = 0.5$, $q = 0.5$, $r = 0$ and $\# = 0$, so $\theta = 0.5(-0.5) + 0.5(-0.5) = -0.5$.

Higher numbers indicate better performance in terms of *ex post* equity, so Alternative C has the best *ex post* equity, with $\theta = 0$. One can interpret this as zero inequity. More negative numbers show greater *ex post* inequity. See the section below on fair-risk models for a societal utility function which includes θ .

Catastrophe Avoidance

Merely calculating expected number of lives lost and choosing protective actions that minimize the number of expected lives lost can lead to choices that might not be agreed upon by members of society. By observation of the nightly TV news we see that the public often reacts more to disasters when 100 people are lost all at once rather than 100 less newsworthy smaller events when one person at a time dies. It is as if the disutility of 100 deaths at once is greater than 100 times the disutility of 1 death.

Probability of Specific Number of Deaths in Group 1 and Group 2 $\pi(n_1, n_2)$	Group 1 Fate of Person 1	Number of Fatalities in Group 1 n_1	Group 2 Fate of Person 2	Number of Fatalities in Group 2 n_2	Average Number of Fatalities $\bar{n}$	$\theta(n_1, n_2)$
p chance	Person 1 lives	0	Person 2 dies	1	0.5	$-((0 - 0.5)^2 + (1 - 0.5)^2) = -0.5$
q chance	Person 1 dies	1	Person 2 lives	0	0.5	$-((1 - 0.5)^2 + (0 - 0.5)^2) = -0.5$
r chance	Person 1 lives	0	Person 2 lives	0	0	$-((0 - 0)^2 + (0 - 0)^2) = 0$
$1 - (p + q + r) = \#$ chance	Person 1 dies	1	Person 2 dies	1	1	$-((1 - 1)^2 + (1 - 1)^2) = 0$

According to Keeney's [5] catastrophe-avoidance condition and Fishburn's [7] *catastrophe-avoidance preference axiom E6*,

$$u(p, q, 1 - p - q, 0) > u(0, 0, 1 - (p + q)/2, \#).$$

In this axiom, it is better for at least one person to survive (just Person 1 in consequence 1 or just Person 2 in consequence 2) than to have both die when the expected number of deaths is equal. Under this axiom, $u(\text{Person 1 lives, Person 2 dies}) = a_1 > 1$ and $u(\text{Person 1 dies, Person 2 lives}) = a_2 > 1$.

Consider the scenario below given by Keller and Sarin [9] to Americans and by Bian and Keller [10] to Chinese. In this scenario, both alternatives have an expected loss of one life, so in terms of expected number of fatalities, they are equal.

Serum-Producing Scenario.

One hundred islanders were born highly susceptible to contracting a fatal disease. Recently, it was discovered that the presence of a naturally occurring noxious gas led to this condition and the gas has been eradicated. However, there is still some chance of the islanders contracting the disease and thus

dying. You could decide to give an injection to all 100 islanders. This injection will prevent everyone from contracting the disease. However, the serum for the injection can only be obtained from the blood of a person who has artificially been made to contract the fatal disease. The serum cannot be obtained from a person who has naturally contracted the disease, so you cannot just wait to see if one person contracts the disease and then make the serum from the sick person's blood. If one islander is sacrificed by being made to contract the disease, enough serum will be obtained to eliminate the risk of death to the remaining 99 islanders. If nothing is done, there is a 1% chance of an epidemic breaking out in which all 100 islanders will contract the disease and thus die. There is a 99% chance that no epidemic will break out, so all 100 islanders will live.

The two options are summarized below. Circle your choice/the option that is fairer.

Alternative D: Do nothing, and thus take a 1% chance of all 100 islanders dying (*Considered fairer by most Americans and Chinese, and chosen by most Americans*).

Alternative E: Sacrifice one islander (*Chosen by most Chinese*).

Alternative D was seen as fairer by most Chinese and Americans. It was chosen as an action by most Americans. It has the chance of the catastrophe of all 100 dying, yet each person has an equal 1% chance of dying. Thus, there is *ex ante* equity and *ex post* equity. There is also common-fate equity, since in one outcome all 100 live and in the other outcome all 100 die.¹

Alternative E avoids any risk of catastrophe, but involves one specific person dying for sure. Most Chinese chose Alternative E, so they preferred *catastrophe avoidance*. It may be that the Chinese cultural value of collectivism led Chinese to protect the 100 person group from all dying, even if one individual had to be sacrificed. In contrast, more Americans may have been led by the American cultural value of individualism to choose Alternative D.

Keeney [4,5] shows that a preference for more equitable distributions of risk, like in the “Do nothing” Alternative D, implies a risk-prone attitude and that catastrophe avoidance, like in the “Sacrifice 1 islander” Alternative E, reveals a risk-averse attitude. Table 1 provides examples of utility functions over number of fatalities which are risk prone, risk neutral, or risk averse. The scale is from 0 to 1, but can be rescaled from -1 to 0, if decision makers prefer to use negative numbers for outcomes of deaths. (See the section titled “Utility Function Demonstrating a Preference for *Ex Post* Equity” for another risk-prone utility function.)

Envy-Free Allocation

Imagine you give some cookies to two children. When one child whines “That’s not fair,” what is meant by this complaint? It could mean that she *envies* the cookie allocation given to the other child, in comparison to her own cookie allocation.

The concept of an *envy-free allocation* means that neither child wants to switch his allocation with the other child.

An envy-free cookie allocation might be attained by following a process where the first child divides the cookies and the second child chooses which of two piles of cookies to take, leaving the rejected pile for the child who divided them.

Envy-Free Axiom.

The chosen allocation should be envy free. An allocation (of risks and/or benefits) between two people is envy free if neither person would want to switch his/her allocation with the other person.

Keller and Sarin [9] discuss the idea of an envy-free allocation of risks and benefits as a way to resolve controversial facility siting problems. See the section titled “Envy-Free Allocation Model.”

DECISION MODELS CONTAINING EQUITY MEASURES

One purpose of identifying equity axioms is to examine what equity principles are satisfied in different existing decision models. A related purpose is to build a decision model based upon the desired set of equity axioms. This section provides a few examples of decision models incorporating a variety of equity preferences and applications. For more models see papers cited earlier and Refs 15–29 for additional references on societal equity.

Utility Function Demonstrating a Preference for *Ex Post* Equity

If societal decision makers prefer more equitable distributions of *ex post* risk, then they may choose to evaluate each individual at risk identically to all others [2,4]. A utility function over the number of fatalities can sufficiently summarize preference for this concept of *ex post* equity of public risk. One specific utility function form examined in

¹This scenario’s context is a modification of one presented in Hammerton *et al.* [13]. Fischhoff [14] used the same probabilities and outcomes to illustrate framing effects and only 29% chose the sure-loss option E. Thus, subjects generally preferred the more equitable option D in a civil defense context also.

Table 1. Examples of Utility Functions Over Number of Fatalities

Utility in Serum-Producing Scenario	Risk Prone: Prefer D (Do Nothing) over E (Sacrifice One Islander)	Risk Neutral: Indifferent between D (Do Nothing) and E (Sacrifice One Islander)	Risk Averse: Prefer E (Sacrifice One Islander) Over D (Do Nothing)
Utility of 0 deaths	1.000	1.000	1.000
Utility of 1 death	0.950	0.990	0.995
Utility of 100 deaths	0.000	0.000	0.000

Keeney [4] is a constantly risk-prone utility function $u(y)$, over the number of fatalities y :

$$u(y) = (1/d)[(1 - d)^y - 1], 0 < d < 1, \\ y = 0, 1, 2, \dots, N,$$

where $u(0 \text{ fatalities}) = 0$ and $u(1 \text{ fatality}) = -1$ to set the origin and scale. Figure 1 below shows the utility over number of fatalities y when the curvature parameter $d = 0.01$. In this convex utility curve with $d = 0.01$, $u(1) = -1$, $u(10 \text{ fatalities}) = -9.56$, $u(20 \text{ fatalities}) = -18.21$, $u(100 \text{ fatalities}) = -64.40$, and $u(200 \text{ fatalities}) = -86.60$. For smaller values of d , the curve is more linear. For $d = 0.0000001$, the utility curve is nearly linear. Also since $\lim_{y \rightarrow +\infty} u(y) = -1/d$, the negative reciprocal of d can be interpreted as the lower bound on the utility due to the societal impact of a tragedy where y is large. Alternatively, d can also be interpreted as the ratio of the societal utility of the first involuntary risk fatality to that of a large tragedy.

To include preference over fatalities, the utility function should be monotonically decreasing since fewer fatalities are always preferred; To include preference for *ex post* equity, the utility function should be risk prone (convex) since a lottery with an expected number of fatalities should be preferred to a sure consequence with the same fatalities which has less spread over society members. The utility function above is both monotonically decreasing and convex. Thus the utility function above addresses both attitudes toward fatalities and *ex post* equity without separating them into individual components.

Fair-Risk Model, Incorporating *Ex Post* and *Ex Ante* Equity Preferences

Keller and Sarin [9] present a simple weighted additive model that includes the number of fatalities, *ex ante* equity of the distribution of the risks, and the *ex post* equity of final consequences. (Keeney and Winkler [2] propose a similar weighted additive structure, in their Equation (3).) Suppose the N_i members in each of m groups, indexed by i , have an independent probability p_i of becoming a fatality. The *fair-risk model* is

$$K_Y U_Y(y) + K_\theta \theta + K_\Phi \Phi,$$

where U_Y is the utility function defined over number of fatalities y , θ is the *ex post* equity measure defined in the respective section, Φ is the *ex ante* equity measure defined in its respective section, and K_Y, K_θ , and K_Φ are scaling constants that reflect trade-offs among the three attributes. A simple utility function U_Y for total number of fatalities is the negative of the expected number of fatalities.

Examination of Antarctica Operations in Terms of Equity to Different Groups

Broder and Keller [17] provide examples of equity measures applied to Antarctica operations. They created two stylized scenarios, based on interviews with US Antarctica operations managers at the US National Science Foundation and reviews of planning documents. In scenario 1, the *status quo* situation was represented, with different fatality risks from a tourist airplane crash or a fatal emergency in Antarctica operations for 3000 tourists, 50 rescue workers, and 950

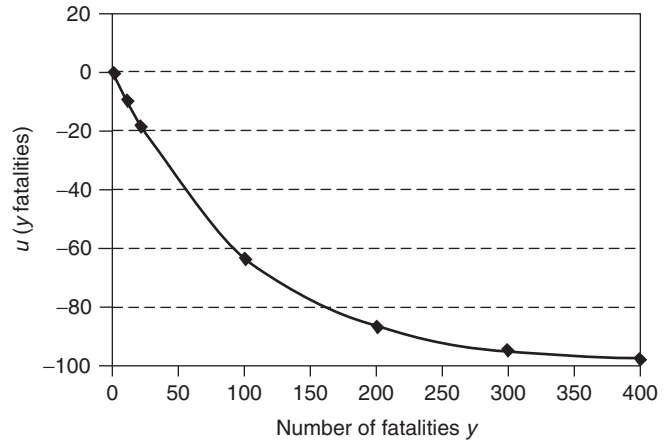


Figure 1. Utility function over number of fatalities when $d = 0.01$.

other Antarctica workers. In scenario 2, a streamlining plan then under consideration was represented in a stylized example where the other Antarctic workers were decreased to 900, and the probabilities and numbers of deaths in groups changed a bit. They then numerically calculated the following equity measures, based on formulae proposed in Fishburn and Sarin [28]:

- individual risk (*ex ante*) inequity;
- group risk (*ex post*) inequity;
 - intergroup inequity (nonuniformity of expected fatality rates across groups);
 - within-group risk inequity;
 - total group risk inequity, that is, combining intergroup and within-group inequity measures,
 - social outcome inequity, assuming common-fate preferences;
- dispersive inequity, for example, risks to scientists versus support staff compared with risks to governmental versus nongovernmental workers, when there are overlapping memberships in these groups (since there are governmental scientists, nongovernmental scientists, governmental support staff, and non-governmental support staff).

Envy-Free Allocation Model

Keller and Sarin [9] presented the following envy-free risk benefit model, which incorporates the concept of an envy-free allocation, as described in the corresponding section. This model also includes information on both the risks and the benefits received by each party. For scenarios involving risks and benefits, see Keller and Sarin's [9] scenario 6 and the scenarios in Keller and Sarin [15].

The model argues that an allocation of benefits b_i and risks r_i to each group i expressed as $(b_1, r_1; b_2, r_2; \dots; b_m, r_m)$ is fair if $u_i(b_i, r_i) \geq u_i(b_j, r_j)$ for $i = 1$ to m , $j = 1$ to m . Then if we assume that a decision maker has a utility function u_D , which is defined over the utility functions u_i of each of the $i = 1$ to m groups, the following mathematical program will give you the fair allocation:

Envy-Free Risk-Benefit Model

Maximize $u_D(u_1, u_2, \dots, u_m)$

Subject to

$$u_i(b_i, r_i) \geq u_i(b_j, r_j), i = 1 \text{ to } m, \\ j = 1 \text{ to } m$$

(envy-free allocations)

$$\sum_{i=1}^m b_i = 1 \quad (\text{all benefits allocated})$$

$$\sum_{i=1}^m r_i = 1 \quad (\text{all risks allocated})$$

$$b_i, r_i \geq 0, i = 1 \text{ to } m$$

(nonnegative variables).

In the above model u_D can be regarded as a social welfare function. u_D increases in group i 's utility u_i . Group i 's utility u_i increases in the benefits b_i and decreases in the risks r_i .

SUMMARY

This article has provided an introduction to a number of *ex ante* and *ex post* equity concepts. When guiding societal policy making, such concepts will often be part of the discussion. They can be dealt with qualitatively in policy discussions, or be formally incorporated into preference models for guiding decision making. Some examples of societal utility function models are discussed in the section titled "Decision Models Containing Equity Measures." See Refs 15–29 for additional references on societal equity, and the article titled ***Fairness and Equity in Societal Decision Analysis***.

REFERENCES

1. Keller LR, Feng T, Wang Y. Fairness and equity in societal decision analysis. Encyclopedia of Operations Research/Management Science (EORMS). Hoboken (NJ): Wiley; 2010. In press.
2. Keeney RL, Winkler RL. Evaluating decision strategies for equity of public risks. Oper Res 1985;33(5):955–970.
3. Keeney RL. Evaluating alternatives involving potential fatalities. Oper Res 1980;28(1):188–205.
4. Keeney RL. Utility functions for equity and public risk. Manage Sci 1980;26(4):345–353.
5. Keeney RL. Equity and public risk. Oper Res 1980;28(3):527–534.
6. Broome J. Equity in risk bearing. Oper Res 1982;30(2):412–414.
7. Fishburn PC. Equity axioms for public risks. Oper Res 1984;32(4):901–908.
8. Sarin RK. Measuring equity in public risk. Oper Res 1985;33(1):210–217.
9. Keller LR, Sarin RK. Equity in social risk: some empirical observations. Risk Anal 1988;8(1):135–146.
10. Bian WQ, Keller LR. Chinese and Americans agree on what is fair, but disagree on what is best in societal decisions affecting health and safety risks. Risk Anal 1999;19:439–452.
11. Gajdos T, Weymark JA, Zoli C. Shared destinies and the measurement of social risk equity. Ann Oper Res 2010;176(1):409–424.
12. Fishburn PC, Straffin PD. Equity considerations in public risks evaluation. Oper Res 1989;37(2):229–239.
13. Hammerton M, Jones-Lee MW, Abbott V. Equity and public risk: some empirical results. Oper Res 1982;30(1):203–207.
14. Fischhoff B. Predicting frames. J Exp Psychol Learn Mem Cogn 1983;9(1):103–116.
15. Keller LR, Sarin RK. Fair processes for societal decisions involving distributional inequalities. Risk Anal 1995;15(1):49–59.
16. Bian WQ, Keller LR. Patterns of fairness judgments in North America and the People's Republic of China. J Consum Psychol 1999;8(3):301–320.
17. Broder I, Keller LR. Fairness of distribution of risks with applications to Antarctica. In: Mellers B, Baron J, editors. Psychological perspectives on justice: theory and application. Cambridge University Press; 1993. Chapter 14. pp. 292–312.
18. Harvey C. Decision analysis models for social attitudes toward inequity. Manage Sci 1985;21(10):1199–1212.
19. Harvey C. Preference functions for catastrophe and risk inequity. Large Scale Syst 1985;8(2):131–146.
20. Kirkwood C. Pareto optimality and equity in social decision analysis. IEEE Trans Syst Man Cybern 1979;9(2):89–91.
21. Bodily S. Analysis of risks to life and limb. Oper Res 1980;28(1):156–175.
22. Harrison D Jr. Distributional objectives in health and safety regulation. In: Ferguson A, LeVeen EP, editors. The benefits of health and safety regulation. Ballinger; Cambridge (MA): 1981. pp. 177–201.
23. Vaupel J. On the benefits of health and safety regulation. In: Ferguson A, LeVeen EP, editors. The benefits of health and safety regulation. Cambridge (MA): Ballinger; 1981. pp. 1–22.

24. Diamond P. Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: comment. *J Polit Econ* 1967;75(5):765–766.
25. Ulph A. The role of ex ante and ex post decisions in the valuation of life. *J Public Econ* 1982;18(2):265–276.
26. Blackorby C, Donaldson D. Can risk-benefit analysis provide consistent policy evaluations of projects involving loss of life? *Econ J* 1986;96(383):758–773.
27. Zeckhauser R, Shepard DS. Principles for saving and valuing lives. In: Ferguson A, LeVeen EP, editors. *The benefits of health and safety regulation*. Cambridge (MA): Ballinger; 1981. pp. 91–130.
28. Fishburn PC, Sarin RK. Dispersive equity and social risk. *Manage Sci* 1991;37(7):751–769.
29. Fishburn PC, Sarin RK. Fairness and social risk I: unaggregated analyses. *Manage Sci* 1994;40(9):1174–1188.

MEMETIC ALGORITHMS

PABLO MOSCATO

REGINA BERRETTA

Centre for Bioinformatics, Biomarker
Discovery and Information-based
Medicine, School of Electrical
Engineering and Computer Science,
University of Newcastle, Callaghan,
New South Wales, Australia

CARLOS COTTA

Departamento de Lenguajes y Ciencias de
la Computación, Escuela Técnica
Superior de Ingeniería Informática,
Universidad de Málaga, Málaga,
Spain

BRIEF HISTORICAL NOTES

Two decades ago, a technical report from a high-performance computing group suggested that we were at a turning point in the application of mathematical programming solutions to challenging optimization problems [1]. This same manuscript also introduced the denomination of “*memetic algorithms*” (MAs). The conclusions obtained were based on the computational experience obtained with Norman and Moscato’s first MA [2]. The results were obtained with the first computer code implemented by Norman in November 1988, who skillfully ported it, in just a matter of days, to two different types of parallel computers. This first program was then thoroughly tested and examined by Moscato during most of 1989 using different variants of the basic initial MA scheme on a number of instances of the traveling salesman problem (TSP) [1]. That technical report presented a qualitative analysis of other hybrid metaheuristics published at the time. It was suggested that MAs were the natural unifying scheme, and the report also predicted the usefulness of this approach and theorized on the reasons for its success.

A common misconception remains 20 years later; some researchers still believe

that MAs evolved from hybrid genetic algorithms (GAs), a new (or at least incipient in 1988) type of strategy for combinatorial and nonlinear optimization. This is certainly not true. MAs actually evolved from the principle that the pragmatic use of different problem-specific algorithmic strategies for a given computational search problem can be synergistically used. The particular target hardware for these algorithms was heterogeneously distributed computing systems and concurrent/parallel computer systems. Moscato recognized the *computer game tournaments* [2,3] used to study *the evolution of cooperation* as his influence [4,5]. As a consequence, from the beginning, MAs were defined as a *competitive and cooperative approach for complex optimization search* [6], where software agents compete for CPU time and can achieve *superlinear speedup* [7]. From the very beginning, the emphasis was on combinatorial optimization problems, as it was perceived that a steady progress on algorithmic solvers for well-known test case scenarios, like those provided by the traveling salesman or the quadratic assignment problem instances, would lead to a clear recognition of the power of these new hybrid algorithms.

By 1989, algorithmic developers in the field of evolutionary computation were subjected to unnecessary restrictions by the *genetic algorithm dogma*, and were constrained to publishing only methodologies that were highly linked to biological evolution. Moscato then proposed that, perhaps, ideas from *cultural evolution* could be a more compelling way of introducing the liberating quest of MAs. In turn, it seemed closer to the discussion of the evolutionary dynamics cultural processes, in which the correct utilization of problem domain knowledge is of paramount importance, which is also one of the key elements of MAs. The term *memetic*, derived from Richard Dawkins’ neologism *meme*, was the perfect match [7]. Since we have *hybridization* of algorithmic techniques as a key ingredient

of MAs, it was not entirely unexpected that at the beginning of the 1990s there was a surge of evolutionary algorithms that used this term, with Lawrence Davis being one of the most stoic supporters of the necessary changes required to be competitive with state-of-the-art algorithmics [8–10]. It was not Moscato's purpose to eliminate an existing dogma by generating a new one. Instead, he aimed at noting more comprehensive forms of evolutionary systems that gain information by creation and destruction processes but employ long periods of individual optimization of designs.

In the early 1990s, researchers who were working on MAs, if they were willing to publish their results, most likely needed to show that their methods were improving existing state-of-the-art heuristics and metaheuristics. Consequently, a benchmark was to show that the cooperative and competitive aspects of their hybrid strategies were producing better results than when acting individually. This resulted in a large number of papers where the MAs were presented as *hybrid evolutionary algorithms*, *hybrid genetic algorithms*, *hybrid metaheuristics*, *genetic local search*, and so on. Reviewers required that these hybrids should outperform both the basic population-based strategies for optimization as well as known metaheuristics like *simulated annealing* (SA) or *tabu search* (TS). This has resulted, during the next 20 years, in a stream of publications that combine these methods as individual optimizers for each agent with population-based techniques, resulting in very robust optimization methods.

MAs have pioneered and championed hybridization of algorithmic techniques, and anticipated many existing methodologies in practice today, which are converging toward this paradigm. We note that the early references to MAs [1] already make use of individual search optimization techniques, which may be different for each agent, use other metaheuristics (i.e., SA, TS and some forms of *variable neighborhood search*), and control the solution population to avoid premature loss of diversity via mechanisms that employ a certain structure of the population (e.g., toroidal [1] and tree [7]).

MAs were also the first metaheuristic that clearly championed the need and relevance of using mathematically literate computing techniques, incorporating *anytime algorithms*, *heuristics*, *approximation algorithms*, *local search techniques*, and *truncated anytime complete algorithms*, with population-based approaches. With a new generation of researchers, after having proved the case, and also due to innovations in the area of exact algorithms, new MAs that include powerful adaptive systems are being developed, while retaining a preference for the design of *parameter-free* MAs, which are very robust, easily implementable, computational solutions for challenging problems.

BASIC ELEMENTS OF A MEMETIC ALGORITHM

All MAs are characterized by the use of (i) a population of agents that use search/optimization methods over a set of solutions (instead of a single optimizing agent); (ii) at least one form of a *competitive* mechanism that produces some sort of selective pressure, not driven entirely by the objective function value of the solutions already found; and (iii) at least one type of *cooperative* procedure that allows information exchange between agents. The search/optimization processes used by each agent do not necessarily need to be the same. It is also not necessary that an individual agent should have “exhausted” its exploration to be ready to recombine solutions. These types of strategies were already presented in the earliest implementations of MAs and are not a novel concept [1]. Another aspect normally present in MAs is that the “wall-clock” time used during exploration by individual agents tends to be much greater than the time employed during the exchange of information. As such, MAs have gained popularity as being ideal for MIMD parallel computer systems. It is increasingly the case that MAs are being used as a metaheuristic template and are becoming the focal point of convergence of several metaheuristic techniques, where these principles of cooperation and competition are used interspersed in many ways.

Solution, Representation, and Fitness/Guiding Functions

One of the first elements highlighted by Moscato [1] as a key for the development of MAs is the choice of a suitable *representation* of solutions of the computational problem at hand. Moscato insisted on the key role that the *correlation of local optima* had on the performance of a MA [1,11], and used the Kauffman's NK-landscapes, the TSP, the quadratic assignment problem, and the problem of finding the optimal conformation of proteins as key examples where the correlation of local optima is in some way evident. In other cases, it can be somewhat enforced by the choice of a particular "*configuration space*" [1,11]. Consequently, we need to define these terms more formally. Readers can then reflect on how to approach a new problem with MAs by choosing a proper configuration space. We note, however, that we have more or less simplified the discussion in what follows, as different agents could be using different configuration spaces, yet they can be exchanging information about the solutions found using some common specification.

In several MAs, it is customary to use a single search space S and we refer to the elements $s \in S$ of this space as *configurations*. The search in this *configuration space* is linked to a search on associated sets that represent solutions of the problem at hand, and consequently some definitions are necessary to formalize these concepts. We assume here that the task is to solve a computational problem P , which has a domain set of *instances* denoted I_P (the "inputs"). For each instance $x \in I_P$, we have an associated set called $\text{solp}(x)$, which represents the *feasible* solutions for problem P given instance x (the "output" we need to compute). The set $\text{solp}(x)$ is also known as the set of *acceptable* or *valid* solutions. Generating these valid solutions is generally not hard (formally, computationally tractable for many problems of interest). Our algorithm must, given an instance $x \in I_P$, return at least one element y from a set $\text{ansp}(x) \supseteq \text{solp}(x)$, that is, a potential (hopefully feasible) solution to the problem. Computational problems can then be classified as follows: *decision* problems (the task

is to determine whether the set $\text{solp}(x)$ is *empty or not*); *satisfaction/search problems* (the task is to find *one* solution $y \in \text{solp}(x)$); *enumeration* problems (the task is to find *all* solutions in $\text{solp}(x)$); *counting* problems (the task is to find *how many* solutions exist in $\text{solp}(x)$); and *optimization* problems (the task is to find a solution in $\text{solp}(x)$, maximizing or minimizing a given function). The latter is the group for which MAs have been mostly used as heuristic *anytime* algorithms. There are, however, MAs that guarantee to deliver the optimal solution, and, in that case, following the standard use of the term in artificial intelligence, we call them *complete MAs*.

An optimization problem is considered *solved* by an MA if it guarantees that it always finds a certain feasible solution $y \in \text{solp}(x)$ or gives an indication that no such feasible solution exists. A Boolean *feasibility* function $\text{feasiblep}(x, y)$ may sometimes be required. The function reports TRUE whether a given solution $y \in \text{ansp}(x)$ is acceptable for an instance $x \in I_P$ of a computational problem P , that is, checking if $y \in \text{solp}(x)$. An MA is said to *solve problem* P if it can fulfill this condition for any given instance $x \in I_P$. There is no doubt that this is a broad definition, hence a more restrictive characterization for our problems of interest is necessary. This characterization is provided by restricting ourselves to *combinatorial optimization* problems, in which the cardinality of $\text{solp}(x)$ is finite for each instance $x \in I_P$, and—as for any optimization problem—a partial order $<_P$ is defined over solutions, so that preferences among them can be established. This partial order is typically defined by means of an objective function that numerically quantifies the quality of a solution (see also the section titled "Multiobjective Memetic Algorithms" for a generalization of this scenario). An instance $x \in I_P$ of a combinatorial optimization problem P is solved by finding the best solution $y^* \in \text{solp}(x)$, that is, finding a solution y^* such that no other solution $y <_P y^*$ exists if $\text{solp}(x)$ is not empty. Note that the optimal solution may not be unique.

We now need to define what a *valid search space* is. Let $S_P(x)$ be a set defined for problem

P whose elements have the following properties: (i) each element $s \in S_P(x)$ represents an answer in $\text{ansp}(x)$ and (ii) *at least one optimal* element y^* of $\text{solp}(x)$ is represented by one element in $S_P(x)$. The elements of $S_P(x)$ are named *configurations* (note that some configurations in the search space explored by MAs may actually correspond to infeasible solutions). For simplicity, we just write S to refer to $S_P(x)$ when x and P are clear from the context. People using biologically inspired metaphors like to call $S_P(x)$ the *genotype space* and $\text{ansp}(x)$ the *phenotype space*, hence we appropriately refer to $g : S_P(x) \rightarrow \text{ansp}(x)$ as the *growth function*.

Designers of MAs generally take inspired advantages of these arbitrary mathematical constructions and, given a problem P , the set $\text{solp}(x)$ is not necessarily the same as $\text{ansp}(x)$. For instance, the use of a *growth function* $g : S_P(x) \rightarrow \text{ansp}(x)$ would allow that a known and useful greedy constructive algorithm may be deterministically generating a solution from a configuration in $S_P(x)$. In addition, each agent in the population may be using a different $S_P(x)$, and it is common to have MAs in which each agent is concurrently optimizing a pool of configurations, with a set of different algorithmic techniques. It is part of the craftsmanship of MAs to be well versed in different algorithmic techniques in order to produce hybrids that bring the best elements among them and synergize to achieve better results in reduced times, exploiting the possibility of *superlinear speed up*, thus making the approach a practical choice. Almost invariably, we have noticed that *correlation of local optima* [1] is always present in good MA implementations [1,11–13].

Individual, Local, and Population-Based Processes

Local search algorithms maintain a single configuration at a time; the maintenance of $k \geq 2$ configurations is an obvious generalization. We call *population-based* search algorithms those search techniques that use multiple concurrent configurations during the search process. Most MAs, either explicitly or implicitly, use a population-based search [11].

Given a certain search space $S = S_P(x)$, an objective function, and a particular search engine determining which solutions are reachable from which other solutions, the search process can be adequately modeled as traversing a so-called *fitness landscape* [13–15]. The latter aspect, namely, the connectivity of the search space largely determines the dynamics of the search algorithm. In local search, a widely used method in MAs, these connectedness relationships are dependent on a *neighborhood function* $N_P(s, x) : S \rightarrow 2^S$. This function assigns to each element $s \in S$ a set $N(s) \subseteq S$ of neighboring configurations of s . The set $N_P(s, x)$ is called the *neighborhood* of s and each element in $N_P(s, x)$ is called a *neighbor* of s . In local search agents, we iteratively explore $N_P(s, x)$, starting with a configuration $s_0 \in S$ (generated at random or constructed by some other algorithm) until a termination criterion is met (realization of a number of iterations, no improvement in the last m iterations, the probability of being at a local optimum, etc. [16]). In MAs, the generation of these new configurations $s_0 \in S$ is due to processes that combine features present in other configurations, by *recombination* processes (see section titled “Recombination and Mutation”). Other types of exact algorithms are used for individual optimization and recombination algorithms. We describe in the section titled “Adaptive Memetic Algorithms,” some of the exact methods that are currently being used. It is not a design requirement of MAs that recombination should proceed only if the solutions to be recombined are local optima; in fact, even the first MAs explicitly avoided this unnecessary restriction [1,11].

Random diversification of the search process is necessary, and MAs employ a variety of procedures (called *mutations* in the context of GAs) to achieve this diversification. These randomizations may be simple or complex “chain of transitions” [17]. Early on, it was clear that even a finite size population would lose diversity, and *restart* mechanisms were proposed in MAs to generate new configurations, iterating to find even better regions of configuration space.

Another method to increase diversification is to introduce a *population structure*

[1,7,18–20], which in turn prevents some recombinations from taking place. Moscato and Norman [6] introduced these structures in their first MA. This methodology has been used by several authors [5,9], but it is yet underexploited in its full potential.

Recombination and Mutation

Recombination operators are algorithms that allow the generation of new restart configurations for the individual search processes in MAs. In MAs, “recombination” is not linked to a particular “biological” metaphor, that is, it allows the use of “multiparents” ($k > 2$), the use of previously generated solutions, constructive algorithms, and so on. The use of clever recombination algorithms, which synergize with the individual search optimizers, allows a more efficient exploration of fitness landscape and it is part of the design of good MA implementations. Recombination can be defined as a process in which a set S_{par} of n configurations (informally referred to as “parents”) is manipulated to create a set S_{desc} of new configurations. The creation of these new configurations involves the identification and combination of features extracted from the parents. Important properties of recombination operators are *respect* (offspring carry features common to both parents), *assortment* (offspring can potentially carry any combination of compatible features from the parents), and *transmission* (offspring exclusively include

features carried by any of the parents) [21,22]. A recombination operator is said to be *blind* if it has no other input than S_{par} , that is, it does not use any information from the problem instance. This definition is certainly very restrictive, and hence is sometimes relaxed to allow the recombination operator to use information regarding the problem constraints (to construct feasible descendants), and possibly the fitness values of configurations $y \in S_{\text{par}}$ (to bias the generation of descendants toward the best parents). A typical example of a blind recombination operator is the classical *Uniform crossover* [23]. In MAs, researchers tend to refer to those blind recombination algorithms as “crossover” that have low computational complexity. Currently, MAs are increasingly using recombination operators that exploit problem knowledge well [18–20,24] (commonly referred to as *heuristic* or *hybrid* recombinations). In these operators, both problem- and instance-specific information is employed to create new solutions. Examples that use problem domain information in different ways are *dynamically optimal recombination* (DOR) [25], *patching-by-forma completion* [22], and *edge assembly crossover* (EAX) [26], which has an analogy with the one proposed in Ref. 11.

A Very Simple Memetic Algorithm Template

A generic template for a MA is shown in Algorithm 1.

Algorithm 1. MA-Based Search Engine

```

begin
  Initialize pop using GenerateInitialPopulationOfAgents()
  Repeat
    newpop ← GenerateNewPopulationOfAgents(pop)
    pop ← UpdatePopulationOfAgents(pop)
    If pop has converged then
      pop ← RestartPopulationOfAgents(pop)
    Endif
  Until TerminationCriterion()
end

```

The `GenerateNewPopulationOfAgents` procedure is at the core of MAs. Essentially, this procedure can be seen as the search engine of a MA. New solutions are generated

at this step, which in general involves the exchange of pieces of information gathered in previous optimization steps. It is an increasingly more common strategy to develop MAs

in which the agents use different strategies (*blend heuristics* [7]). We also have MAs in which an *algorithmic portfolio* [27] is employed by the agents to optimize the configurations they have assigned. New individual optimization strategies can also be generated in a process that evolves not only solutions but solutions methods (i.e., coevolutionary MAs). The `UpdatePopulationOfAgents` procedure involves the optimization of current configurations being explored and the selection of a set of solutions to be further contemplated as necessary to continue the search process. The `GenerateInitialPopulationOfAgents` procedure generates an initial set of $|pop| \geq 2$ agents, each initially having at least one configuration of the search space and a set of algorithmic strategies to use. In some cases, the initial configurations are chosen at random or via more elaborate strategies (typically by application of individual optimization methods or constructive heuristics [28]).

Several criteria can be used to evaluate the loss of diversity of a population, which then triggers the request to restart the population in other regions of configuration space. There are also different types of termination criteria used, via simple computations such as attaining a certain limit on the total CPU time (or number of iterations), attainment of a maximum number of iterations without improvement, or a maximum number of population restarts. In many papers, these criteria have been generally dictated by the need for fair comparisons against population-based methods that do not employ exchange of information between agents. Consequently they are, together with smart restart procedures, fertile ground for further exploration of methodologies.

Criteria for Designing an Effective Memetic Algorithm

There is no single approach for the design of effective MAs, but there are generic rules that have worked well in practice. These *design heuristics* have collectively found good-performing strategies in several problem domains. It is, however, impossible to guarantee that the same strategy that is working well with one particular problem

will work well for all types of problems. It is in the instantiation of the previous template with precise problem-dependent components where the skills of MA design are applied (and where the use of problem domain information becomes essential).

We mentioned before that the representation of solutions in terms of “meaningful information units” is essential. Several measures have been used as a proxy for this “meaningfulness” property, such as the variability of the objective function for solutions carrying a common set of features [22,29], or the interaction (nonadditive effects on the objective function) among these information units [22,29], among others. In essence, these measures try to capture the fact that the fitness landscape has to exhibit a certain regularity that can be exploited by the search algorithm. In addition to that design requirement (recall that the fitness landscape is given by the choice of representation and operators), there should also be a good synergy between the individual optimization strategies of the MA and the type of recombination operators used. We often note the case put forward by Freisleben and Merz [30] for the TSP in which the authors first noted that a highly intensive local search procedure (Lin–Kernighan; [31]) produced local optimums that closely resemble each other, thus the global optimum (which is one of them) should be searched with a synergistic local search strategy. For this reason, they introduced a *distance-preserving* recombination operator that generates new solutions whose distance (measured in number of different edges between a pair of tours) from the parents is the same as the distance between the parents themselves. It is customary in MAs to first experiment with the individual optimization search strategies of the agents and, when some experience has been gained, design a recombination operator that works as a complement to the individual search processes.

Selection Strategies and Population Structure

The selection of optimization strategies for the agents is vital, along with adequate parameterizations of the basic procedures. It

is often the case that a simple parameterization may help in diversifying the search processes when necessary. Designers should also have other considerations such as when it is proper to stop individual optimization for a given agent, how often it should be applied, to which agents it should be applied, how to select which solutions should have “deeper” search processes, and so on [32,33]. Although these complex rules have shown to provide clear benefits, other population management strategies are also found in the literature. The use of strategies to control the loss of diversity of the population was always a concern in MAs since their foundation. Algorithmic strategies that aim to directly address this issue, via control mechanisms, are currently being proposed since the population sizes of MAs are smaller than in other evolutionary computation methods (e.g., “population management” [34–37]). Other strategies include *random selection*, *fitness-based selection* (only the best individuals are subject to local improvement), and the *stratified approach*. In this case, the population is sorted and divided into n levels, with n being the number of local search applications, and one individual per level is randomly selected [38].

There are several criteria for selecting which individuals will be subject to individual optimization. Some of them restrict this by the use of *population-structured MA* [18,39–42], where good solutions evolve within a tree-structured population. Since 1992 [7], Moscato and his collaborators have been experimenting, with a population structure-based on a ternary tree [18–20,41–46] for an effective, yet undirected and static, mechanism to control a premature loss of diversity. A comprehensive computational study in 2004, on the Asymmetric TSP, showed the benefits of introducing this strategy [39]; we expect that more research into this aspect of MAs is guaranteed. We also note that early MAs were also based on population structures [1,9,11,47,48] and this is yet a field with untapped potential as not many researchers have fully explored it.

In the case where the computation of the guiding function is costly, some researchers

have developed fast approximate models of the true guiding function to help accelerate their MAs [49–53]. Other strategies include allowing that not every configuration be subject to optimization steps (so-called *partial Lamarckism* by some authors [54–56]). While these algorithmic decisions tend to be problem dependent, with time, most researchers in MAs move toward the simplest, parameter-free techniques, or they establish *adaptive* and *self-adaptive* mechanisms to learn-on-the-fly the most appropriate parameter setting. We discuss this issue in the section titled “Adaptive Memetic Algorithms.”

ALGORITHMIC EXTENSIONS OF MEMETIC ALGORITHMS

Complete Memetic Algorithms

Complete MAs can be approached from both a sequential and an intertwined perspective. *Sequential approaches* are those in which the execution of an MA is followed by an exact method, for example, *Branch-and-Cut* (see Ref. 57 for an example). *Concurrent approaches* also exist; Denzinger and Offerman [58] combine genetic algorithms and *Branch-and-Bound* techniques within a parallel multiagent system. Cooperation of these algorithms can also be found in Refs 59 and 60, in which the exact technique provides partial promising solutions, and the metaheuristic is in charge of returning an improved bound. *Beam search* and full-fledged MAs can be found in Refs 61–63. Exact algorithms, such as *Branch-and-Bound* [25] or *Dynamic Programming* [64], among others, have provided powerful recombination algorithms (recall the section titled “Recombination and Mutation”).

This trend is perhaps not so new; using exact techniques to solve some subproblems provided by evolutionary algorithms dates back to 1995 [59,60]. Fernandes and Lourenco [65] also give a large list of references regarding local search/exact hybrids. Also, recently, Puchinger and Raidl [66] classified methods of this kind in which algorithmic combinations are *collaborative* (sequential or intertwined/concurrent execution) or *integrative* (one technique works

inside the other one, as a subordinate). Employing an exact technique to explore neighborhoods is indeed an example of an integrative approach. Further examples are the use of Branch-and-Bound in the decoding process [67] of a genetic algorithm or the use of evolutionary computation for the strategic guidance of Branch-and-Bound [68] (exemplifying the use of a metaheuristic approach within an exact method).

In spite of the fact that they use exact techniques along with metaheuristics, most of these hybrid algorithms are not complete, as they do not guarantee an optimal solution. However, they can be readily modified to achieve completeness. Exceptions exist, like French *et al.*'s [69] proposal of combining an integer-programming Branch-and-Bound approach with genetic algorithms for MAX-SAT, and the famous CONCORDE package for the TSP developed in the past decade due to the combined talents of Applegate, Bixby, Cook, and Chvatal [70], which is regarded by many as the most complex MA ever built for an optimization problem. In their approach, a local search algorithm (chained Lin–Kernighan) produces a collection of tours, followed by the DOR method called *tour merging*. This scheme has shown to produce a nonoptimal tour only 0.0002% above the proved optimal tour for 13,509 cities. Even today, one of the best local search methods for the TSP (Helsgaun's LKH code) now implements multiparent recombination, showing once again the increasing importance of this scheme (again referred to as *tour merging* in LKH-2 Version 2.0 [71]).

Multiobjective Memetic Algorithms

Multiobjective problems are frequent in real-world applications that present multiple, partially conflicting objectives. The notion of Pareto-dominance is a key: given two solutions $s, s' \in \text{sol}_P(x)$, s is said to dominate s' if it is better than s' in at least one of the objectives, and it is no worse in the remaining ones. This induces a partial order $<_P$, and the resulting collection of optimal solutions is termed the optimal *Pareto front*, or the *optimal nondominated front*. We refer readers to Refs 72–75 for more information on this topic. Multiobjective MAs adapt

local search processes in different ways [76], typically *Pareto-based* [77] and *scalarizing* approaches [78,79]. Several multiobjective MAs (MOMAs) exist like those developed by Ishibuchi and Murata [80,81] and Jaszkiewicz [82,83]. They employ random scalarization each time a local search is to be used. A single-objective local search could be used to optimize individual objectives [84]. We refer to Ref. 85 for an approach which, although originally designed for a biobjective capacitated arc routing problem (CARP), it is nevertheless, highly competitive on instances originating from single-objective CARP problems. Such studies allow researchers to have a better understanding of the power of these methods. Ad hoc mating strategies based on the particular weights chosen at each local search invocation (whereby the solutions to be recombined are picked according to these weights) are used as well. A related approach, including the on-line adjustment of scalarizing weights, is followed by Guo *et al.* [78,86,87]. On the other hand, an MA based on PAES (Pareto Archived Evolution Strategy) was defined by Knowles and Corne [88,89]. More recently, an MOMA based on a basic particle swarm optimization (PSO) has been defined by Liu *et al.* [90,91]. In this algorithm, an archive of nondominated solutions is maintained and randomly sampled to obtain reference points for particles. A different approach is used by Schuetze *et al.* [92] for numerical-optimization problems. The continuous nature of solution variables allows using their values for computing search directions. This fact is exploited in their local search procedure (HCS for Hill Climber with Sidestep) for directing the search toward specific regions (e.g., along the Pareto front) when required.

Adaptive Memetic Algorithms

In MAs, as in other novel metaheuristics converging to this paradigm, we need to customize individual search strategies for the particular problem [93]. *Parametric* adaptations exist, for instance, like modifications of operator application rates (see Refs 8 and 94); this type of adaptive strategy also exists in MAs [95–98]. The *meta-Lamarckian learning* [99] strategy works at an algorithmic

Table 1. Applications in Production Planning and Management Science^a

Manufacturing	Assembly line	[114–116]
	Flexible manufacturing	[117–120]
	Lot sizing	[20]
	Multitool milling, cutting, and scheduling problem in printed circuit board	[67,121,122]
	Supply chain	[123,124]
Production planning, scheduling, and management science	Temporal planning and multiobjective dynamic location	[125,126]
	Flowshop scheduling	[40,90,127–138]
	Job-shop	[139–148]
	Parallel machine scheduling	[131,149,150]
	Project scheduling, integrated production-distribution, assembly sequence planning, and assembly line balancing	[151–155]
	Single machine scheduling	[131,156]
Timetabling	Driver scheduling	[157]
	Examination	[158]
	Rostering	[159–162]
	Sport league	[163]
	Train	[164]
	University course	[165–169]

^aCheck also Ref. 113.**Table 2. Applications in Combinatorial Optimization**

Binary and set problems	Binary quadratic programming	[170]
	Knapsack problem	[60,86,87,171,172]
	Low autocorrelation sequences	[173]
	Max-SAT	[174,175]
	Set covering	[83]
Graph-based problems	Crossdock optimization	[176,177]
	Graph coloring	[178]
	Graph matching	[179]
	Hamiltonian cycle	[180]
	Maximum cut	[181]
	Quadratic assignment	[182,183]
	Routing problems	[34,35,37,155,184–197]
	Spanning tree	[198,199]
	Steiner tree	[57]
Constrained optimization	Traveling salesman problem	[39,200–207]
	Golomb ruler	[208,209]
	Social golfer	[210]
	Maximum density still life	[61,64]

level by considering the setting in which the MA has a collection of local search operators available, and how the selection of the particular operator(s) to be applied to a specific solution can be done on the basis of past performance of the operator, or on the basis of the similarity of the solution to previous successful cases of operator application. Some analogies with *hyperheuristics* [100,101] exist.

Self-adaptation refers to the use of mechanisms that evolve the search algorithms during the run of a MA [102]. In Ref. 103, it was already envisioned that MAs should and would evolve in at least two levels and two time scales. In a short-time scale, a set of agents would be searching in the search space associated with the problem. The long-time scale would *adapt all the*

Table 3. Applications in Electronics, Telecommunications, and Engineering

Electronics	Analog circuit design	[211]
	Circuit partitioning	[212]
	Electromagnetism	[49,213,214]
	Filter design	[215]
	VLSI design	[44,216,217]
Engineering	Chemical kinetics	[218,219]
	Crystallography	[56]
	Drive design	[220,221]
	Power systems	[41,222]
	Structural optimization and robust airfoil shape optimization	[223,224]
Computer science	System modeling	[225,226]
	Code optimization, web document query optimization, and multiprocessor scheduling	[227–229]
	Information forensics	[230]
	Information theory	[231]
	Software engineering	[232]
Telecommunications	Antenna design	[233–237]
	Mobile networks	[238,239]
	P2P networks	[240–242]
	Wavelength assignment	[243]
	Wireless networks	[244,245]

Table 4. Applications in Machine Learning and Knowledge Discovery

Data mining and knowledge discovery	Image analysis	[246–250]
	Clustering	[251–255]
	Feature selection	[256,257]
	Pattern recognition, fuzzy rule base extraction	[258–261]
Machine learning	Decision trees	[262]
	Inductive learning	[263]
	Neural networks	[264–271]

Table 5. Applications in Bioinformatics and Biomedical Research

Phylogeny	Phylogenetic inference and reconstruction	[62,272–274]
	Consensus tree	[275]
Microarrays	Biclustering	[276]
	Feature selection	[277–279]
	Gene ordering	[41,45]
Sequence analysis	Shortest common supersequence	[54,63]
	DNA sequencing	[280]
Protein science	Sequence assignment	[281]
	Structure comparison	[104]
	Structure prediction	[282–286]
Systems biology	Gene regulatory networks and PCR product primer design	[287–289]
	Cell models	[290]
Biomedicine	Drug therapy design and 3D reconstruction of forensic objects	[291–293]

algorithms associated with the agents (i.e., recombination, individual search strategies, etc.). Examples of this methodology are the *multimemetic algorithms*, in which each solution carries a gene that indicates which local search has to be applied on it and other more complex methodologies currently under intensive development in MAs [104–108].

APPLICATIONS OF MEMETIC ALGORITHMS

We refer to applications produced before 2004 by consulting Refs 109–112. References are listed as belonging to five major areas: production planning and management science (Table 1), traditional combinatorial optimization (Table 2), electronics, engineering, and telecommunications (Table 3), machine learning and knowledge discovery (Table 4), and bioinformatics and biomedical research (Table 5). As mentioned before, we have tried to be illustrative rather than exhaustive, noting some selected references from these well-known application areas.

There are several applications of MAs in economics, for example, in portfolio optimization [294,295], risk analysis [296], optimal winner determination problem [297], time series forecasting [298], and labour-market delineation [299].

CONCLUSIONS

MAs have shown an impressive record of success in solving an ample variety of problems, owing to its pragmatic approach to optimization. The generally complex structure of MAs, blending together elements from diverse metaheuristics, makes them harder to analyze from a theoretical point of view though. Subsequently, the past decade has witnessed a remarkable increase in our theoretical understanding of MA-specific issues such as the relevance of local search depth and intensity [33,300], or the complexity of provably good multiparent recombination operators [301]. This understanding of the underlying theoretical issues will foster in upcoming years, and should lead to a more

principled approach to MA design for specific problems.

We face a future in which computing will be pervasive, and MAs can flourish in this scenario. Indeed, the sheer computational power and the distributed nature of computing resources will provide an opportunity for increasingly complex MA models envisioned in the past and only recently explored in practice, such as multilevel MAs in which evolution takes place at both the solution and the heuristic level [104–108] or large-scale cooperative models involving exact techniques. Increasingly complex solvers will be required to tackle the increasingly complex optimization problems that will appear in the future, and MAs will play a paramount role here.

REFERENCES

1. Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms. Technical report: Caltech concurrent computation program. Pasadena (CA): California Institute of Technology; 1989.
2. Moscato P, Norman MG. A memetic approach for the traveling salesman problem: implementation of a computational ecology for combinatorial optimization on message-passing systems. Parallel computing and transputer applications. Barcelona: IOS Press, Amsterdam; 1992.
3. Hofstadter DR. Computer tournaments of the prisoners-dilemma suggest how cooperation evolves. *Sci Am* 1983;248(5):16–23.
4. Axelrod R, Hamilton WD. The evolution of cooperation. *Science* 1981;211(4489):1390–1396.
5. Nakamaru M, Nogami H, Iwasa Y. Score-dependent fertility model for the evolution of cooperation in a lattice. *J Theor Biol* 1998;194(1):101–124.
6. Norman MG, Moscato P. A competitive and cooperative approach to complex combinatorial search. Technical report: Caltech concurrent computation program. Pasadena; 1989.
7. Moscato P, Tinetti F. Blending heuristics with a population-based approach: a memetic algorithm for the traveling salesman problem. La Plata: Universidad Nacional de La Plata; 1992.

8. Davis L. Handbook of genetic algorithms. New York: Van Nostrand Reinhold Computer Library; 1991.
9. Gorges-Schleuter M. Explicit parallelism of genetic algorithms through population structures. Parallel problem solving from nature. Dortmund, Germany; 1991. pp. 150–159.
10. Mühlenbein H. Evolution in time and space—the parallel genetic algorithm. Foundations of genetic algorithms. Morgan Kaufmann Publishers; 1991. pp. 316–337.
11. Moscato P. An introduction to population approaches for optimization and hierarchical objective functions: the role of tabu search. *Ann Oper Res* 1993;41(1–4):85–121.
12. Merz P. Advanced fitness landscape analysis and the performance of memetic algorithms. *Evol Comput* 2004;12(3):303–325.
13. Merz P, Freisleben B. Fitness landscapes and memetic algorithm design. New ideas in optimization. McGraw-Hill; 1999. pp. 245–260.
14. Kauffman SA, Levin S. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 1987;128:11–45.
15. Kauffman SA. The origins of order: self-organization and selection in evolution. New York: Oxford University Press; 1993.
16. Cotta C, Alba E, Troya JM. Stochastic reverse hillclimbing and iterated local search. Proceedings of the 1999 Congress on Evolutionary Computation. IEEE; 1999. pp. 1558–1565.
17. Radcliffe NJ, Surry PD. Formal memetic algorithms. Evolutionary computing: AISB workshop. Berlin: Springer; 1994. pp. 1–16.
18. Berretta R, Cotta C, Moscato P. Enhancing the performance of memetic algorithms by using a matching-based recombination algorithm: results on the number partitioning problem. *Metaheuristics: computer-decision making*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2003. pp. 65–90.
19. Berretta R, Moscato P. The number partitioning problem: an open challenge for evolutionary computation? New ideas in optimization. McGraw-Hill; 1999. pp. 261–278.
20. Berretta R, Rodrigues LF. A memetic algorithm for a multistage capacitated lot-sizing problem. *Int J Prod Econ* 2004;87(1):67–81.
21. Radcliffe NJ. The algebra of genetic algorithms. *Ann Math Artif Intell* 1994;10:339–384.
22. Radcliffe NJ, Surry PD. Fitness variance of formae and performance prediction. Proceedings of the 3rd Workshop on Foundations of Genetic Algorithms. Morgan Kaufmann; 1994. pp. 51–72.
23. Syswerda G. Uniform crossover in genetic algorithms. Proceedings of the 3rd International Conference on Genetic Algorithms. Morgan Kaufmann; 1989. pp. 2–9.
24. Cotta C, Troya JM. On the influence of the representation granularity in heuristic forma recombination. *ACM Symposium on Applied Computing* 2000. ACM Press; 2000. pp. 433–439.
25. Cotta C, Troya JM. Embedding branch and bound within evolutionary algorithms. *Appl Intell* 2003;18(2):137–153.
26. Nagata Y, Kobayashi S. Edge assembly crossover: a high-power genetic algorithm for the traveling salesman problem. Proceedings of the 7th International Conference on Genetic Algorithms. Morgan Kaufmann; 1997. pp. 450–457.
27. Gomes CP, Selman B. Algorithm portfolios. *Artif Intell* 2001;126(1–2):43–62.
28. Surry PD, Radcliffe NJ. Inoculation to initialise evolutionary search. *Evolutionary computing: AISB workshop*. Springer; 1996. pp. 269–285.
29. Davidor Y. Epistasis variance: suitability of a representation to genetic algorithms. *Complex Syst* 1990;4(4):369–383.
30. Freisleben B, Merz P. A genetic local search algorithm for solving symmetric and asymmetric traveling salesman problems. Proceedings of the 1996 IEEE International Conference on Evolutionary Computation. Nagoya: IEEE Press; 1996. pp. 616–621.
31. Lin S, Kernighan B. An effective heuristic algorithm for the traveling salesman problem. *Oper Res* 1973;21:498–516.
32. Krasnogor N, Smith J. Memetic algorithms: the polynomial local search complexity theory perspective. *J Math Model Algorithm* 2008;7(1):3–24.
33. Sudholt D. Memetic algorithms with variable-depth search to overcome local optima. *GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. ACM Press; 2008. pp. 787–794.
34. Boudia M, Prins C, Reghioui M. An effective memetic algorithm with population management for the split delivery vehicle routing problem. *Hybrid metaheuristics* 2007. Springer; 2007. pp. 16–30.

35. Prodhon C, Prins C. A memetic algorithm with population management (MA|PM) for the periodic location-routing problem. *Hybrid Metaheuristics 2008*. Springer; 2008. pp. 43–57.
36. Sörensen K, Sevaux M. MA|PM: memetic algorithms with population management. *Comput OR* 2006;33:1214–1225.
37. Prins C, Prodhon C, Calvo RW. A memetic algorithm with population management (MA|PM) for the capacitated location-routing problem. *Evolutionary computation in combinatorial optimization*. Springer; 2006. pp. 183–194.
38. Nguyen QH, Ong Y-S, Krasnogor N. A study on the design issues of memetic algorithm. *2007 IEEE Congress on Evolutionary Computation*. IEEE Press; 2007. pp. 2390–2397.
39. Buriol L, Franca PM, Moscato P. A new memetic algorithm for the asymmetric traveling salesman problem. *J Heuristics* 2004;10(5):483–506.
40. Franca PM, Mendes AS, Moscato P. A memetic algorithm for the total tardiness single machine scheduling problem. *Eur J Oper Res* 2001;132:224–242.
41. Mendes A, Cotta C, Garcia V, *et al.* Gene ordering in microarray data using parallel memetic algorithms. *Proceedings of the 2005 International Conference on Parallel Processing Workshops*. IEEE Press; 2005. pp. 604–611.
42. Mendes AS, Franca PM, Moscato P. Fitness landscapes for the total tardiness single machine scheduling problem. *Neural Netw World* 2002;2(2):165–180.
43. Mendes A, Franca PM, Lyra C, *et al.* Capacitor placement in large-sized radial distribution networks. *International NAISO Symposium on Engineering of Intelligent Systems*. Malaga; 2005. pp. 496–502.
44. Mendes A, Linhares A. A multiple-population evolutionary approach to gate matrix layout. *Int J Syst Sci* 2004;35(1):13–23.
45. Moscato P, Mendes A, Berretta R. Benchmarking a memetic algorithm for ordering microarray data. *Biosystems* 2007;88(1–2):56–75.
46. Mendes AS, Muller FM, Franca PM, *et al.* Comparing meta-heuristic approaches for parallel machine scheduling problems with sequence-dependent setup times. In: *Proceedings of the 15th International Conference on CAD/CAM Robotics and Factories of the Future*; Aguas de Lindoia, Brazil. 1999.
47. Gorges-Schleuter M. ASPARAGOS: an asynchronous parallel genetic optimization strategy. *Proceedings of the 3rd International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers; 1989. pp. 422–427.
48. Gorges-Schleuter M. Genetic algorithms and population structures—a massively parallel algorithm. Germany: University of Dortmund; 1991.
49. Guimaraes FG, Campelo F, Igarashi H, *et al.* Optimization of cost functions using evolutionary algorithms with local learning and local search. *IEEE Trans Magn* 2007;43(4):1641–1644.
50. Lim D, Ong Y-S, Jin Y, *et al.* A study on metamodeling techniques, ensembles, and multi-surrogates in evolutionary computation. *GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*. ACM Press; 2007. pp. 1288–1295.
51. Wanner E, Guimares F, Takahashi R, *et al.* Multiobjective memetic algorithms with quadratic approximation-based local search for expensive optimization in electromagnetics. *IEEE Trans Magn* 2008;44(6):1126–1129.
52. Wanner EF, Guimaraes FG, Takahashi RHC, *et al.* Local search with quadratic approximations into memetic algorithms for optimization with multiple criteria. *Evol Comput* 2008;16(2):185–224.
53. Zhou Z, Ong Y-S, Lim M-H, *et al.* Memetic algorithm using multi-surrogates for computationally expensive optimization problems. *Soft Comput* 2007;11(10):957–971.
54. Cotta C. Memetic algorithms with partial Lamarckism for the shortest common supersequence problem. *Artificial intelligence and knowledge engineering applications: a bioinspired approach*. Springer; 2005. pp. 84–91.
55. Houck C, Joines JA, Kay MG, *et al.* Empirical investigation of the benefits of partial Lamarckianism. *Evol Comput* 1997;5(1):31–60.
56. Paszkowicz W. Properties of a genetic algorithm extended by a random self-learning operator and asymmetric mutations: a convergence study for a task of powder-pattern indexing. *Anal Chim Acta* 2006;566(1):81–98.

57. Klau GW, Ljubić I, Moser A, Mutzel P, Neuner P, *et al.* Combining a memetic algorithm with integer programming to solve the prize-collecting Steiner tree problem. GECCO 04: Gen Evol Comput Conf 2004; 3102(Part 1):1304–1315.
58. Denzinger J, Offermann T. On cooperation between evolutionary algorithms and other search paradigms. 6th International Conference on Evolutionary Computation. IEEE Press; 1999. pp. 2317–2324.
59. Cotta C, Aldana JF, Nebro AJ, *et al.* Hybridizing genetic algorithms with branch and bound techniques for the resolution of the TSP. Artificial neural nets and genetic algorithms 2. Springer; 1995. pp. 277–280.
60. Gallardo JE, Cotta C, Fernandez AJ. Solving the multidimensional knapsack problem using an evolutionary algorithm hybridized with branch and bound. Artificial intelligence and knowledge engineering applications: a bioinspired approach. Springer; 2005. pp. 21–30.
61. Gallardo JE, Cotta C, Fernandez AJ. A multi-level memetic/exact hybrid algorithm for the still life problem. Parallel problem solving from nature IX. Springer; 2006. pp. 212–221.
62. Gallardo JE, Cotta C, Fernández AJ. Reconstructing phylogenies with memetic algorithms and branch-and-bound. Analysis of biological data: a soft computing approach. World Scientific; 2007. pp. 59–84.
63. Gallardo JE, Cotta C, Fernández AJ. On the hybridization of memetic algorithms with branch-and-bound techniques. IEEE Trans Syst Man Cybern Part B 2007;37(1):77–83.
64. Gallardo JE, Cotta C, Fernandez AJ. A memetic algorithm with bucket elimination for the still life problem. Evolutionary computation in combinatorial optimization. Springer; 2006. pp. 73–84.
65. Fernandes S, Lourenco H. Hybrids combining local search heuristics with exact algorithms. V congreso español sobre meta-heurísticas, algoritmos evolutivos y bioinspirados. Las Palmas, Spain; 2007. pp. 269–274.
66. Puchinger J, Raidl GR. Combining metaheuristics and exact algorithms in combinatorial optimization: a survey and classification. Artificial intelligence and knowledge engineering applications: a bioinspired approach. Springer; 2005. pp. 41–53.
67. Puchinger J, Raidl GR, Koller G. Solving a real-world glass cutting problem. 4th European Conference on Evolutionary Computation in Combinatorial Optimization. Springer; 2004. pp. 165–176.
68. Kostikas K, Fragakis C. Genetic programming applied to mixed integer programming. 7th European Conference on Genetic Programming. Springer; 2004. pp. 113–124.
69. French AP, Robinson AC, Wilson JM. Using a hybrid genetic-algorithm/branch and bound approach to solve feasibility and optimization integer programming problems. J Heuristics 2001;7(6):551–564.
70. Applegate DL, Bixby RE, Chvátal V, *et al.* The traveling salesman problem: a computational study (Princeton series in applied mathematics). Princeton University Press; 2007.
71. Helsgaun K. LKH-2.0. Available at <http://akira.ruc.dk/~keld/research/LKH/>. Accessed 2010 Jul 1
72. Coello Coello CA, Lamont GB. Applications of multi-objective evolutionary algorithms. New York: World Scientific; 2004.
73. Coello Coello CA, Van Veldhuizen DA, Lamont GB. Evolutionary algorithms for solving multi-objective problems. Genetic algorithms and evolutionary computation. Kluwer Publishers; 2002.
74. Deb K. Multi-objective optimization using evolutionary algorithms. Chichester: John Wiley & Sons, Ltd.; 2001.
75. Zitzler E, Kaumanns M, Bleuler S. A tutorial on evolutionary multiobjective optimization. Metaheuristics for multiobjective optimization. Springer; 2004.
76. Knowles J, Corne D. Memetic algorithms for multiobjective optimization: issues, methods and prospects. Recent advances in memetic algorithms. Springer; 2005. pp. 313–352.
77. Knowles J, Corne D. Approximating the non-dominated front using the pareto archived evolution strategy. Evol Comput 2000;8(2): 149–172.
78. Guo XP, Yang GK, Wu ZM. A hybrid self-adjusted memetic algorithm for multi-objective optimization. 4th Mexican International Conference on Artificial Intelligence. Springer; 2005. pp. 663–672.
79. Ulungu EL, Teghem J, Fortemps P, *et al.* MOSA method: a tool for solving multi-objective combinatorial optimization problems. J Multi Criteria Decis Anal 1999;8(4): 221–236.
80. Ishibuchi H, Murata T. Multi-objective genetic local search algorithm. 1996

- International Conference on Evolutionary Computation. IEEE Press. pp. 119–124.
81. Ishibuchi H, Murata T. Multi-objective genetic local search algorithm and its application to flowshop scheduling. *IEEE Trans Syst Man Cybern* 1998;28(3):392–403.
 82. Jaszkiwicz A. Genetic local search for multi-objective combinatorial optimization. *Eur J Oper Res* 2002;137(1):50–71.
 83. Jaszkiwicz A. A comparative study of multiple-objective metaheuristics on the bi-objective set covering problem and the Pareto memetic algorithm. *Ann Oper Res* 2004; 131(1–4):135–158.
 84. Ishibuchi H, Hitotsuyanagi Y, Tsukamoto N, *et al.* Use of heuristic local search for single-objective optimization in multiobjective memetic algorithms. *Parallel problem solving from nature X*. Springer; 2008. pp. 743–752.
 85. Lacomme P, Prins C, Sevaux M. A genetic algorithm for a bi-objective capacitated arc routing problem. *Comput Oper Res* 2006; 33(12):3473–3493.
 86. Guo XP, Wu ZM, Yang GK. A hybrid adaptive multi-objective memetic algorithm for 0/1 knapsack problem. *AI 2005: advances in artificial intelligence*. Springer; 2005. pp. 176–185.
 87. Guo XP, Yang GK, Wu ZM, *et al.* A hybrid fine-timed multi-objective memetic algorithm. *IEICE Trans Fund Electron Commun Comput Sci* 2006;E89A(3):790–797.
 88. Knowles JD, Corne DW. M-PAES: a memetic algorithm for multiobjective optimization. *Proceedings of the 2000 Congress on Evolutionary Computation (CEC00)*. IEEE Press; 2000. pp. 325–332.
 89. Knowles JD, Corne DW. A comparison of diverse approaches to memetic multiobjective combinatorial optimization. In: *Proceedings of the 2000 Genetic and Evolutionary Computation Conference Workshop Program*. 2000. pp. 103–108.
 90. Li BB, Wang L, Liu B. An effective PSO-based hybrid algorithm for multiobjective permutation flow shop scheduling. *IEEE Trans Syst Man Cybern Part B* 2008;38(4):818–831.
 91. Liu D, Tan KC, Goh CK, *et al.* A multiobjective memetic algorithm based on particle swarm optimization. *IEEE Trans Syst Man Cybern Part B* 2007;37(1):42–50.
 92. Schuetze O, Sanchez G, Coello Coello CA. A new memetic strategy for the numerical treatment of multi-objective optimization problems. *GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. ACM Press; 2008. pp. 705–712.
 93. Cotta C, Sevaux M, Sörensen K. Adaptive and multilevel metaheuristics. *Studies in computational intelligence*. Springer; 2008.
 94. Smith JE. Self-adaptation in evolutionary algorithms for combinatorial optimization. *Adaptive and multilevel metaheuristics*. Springer; 2008. pp. 31–57.
 95. Bambha NK, Bhattacharyya SS, Teich J, *et al.* Systematic integration of parameterized local search into evolutionary algorithms. *IEEE Trans Evol Comput* 2004;8(2):137–155.
 96. Lozano M, Herrera F, Krasnogor N, *et al.* Real-coded memetic algorithms with crossover hill-climbing. *Evol Comput* 2004; 12(3):273–302.
 97. Molina D, Herrera F, Lozano M. Adaptive local search parameters for real-coded memetic algorithms. *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*. IEEE Press; 2005. pp. 888–895.
 98. Molina D, Lozano M, Herrera F. Memetic algorithms for intense continuous local search methods. *Hybrid metaheuristics 2008*. Springer; 2008. pp. 58–71.
 99. Ong Y-S, Keane AJ. Meta-Lamarckian learning in memetic algorithms. *IEEE Trans Evol Comput* 2004;8(2):99–110.
 100. Chakhlevitch K, Cowling P. Hyperheuristics: recent developments. *Adaptive and multilevel metaheuristics*. Springer; 2008. pp. 3–29.
 101. Cowling P, Kendall G, Soubeiga E. A hyperheuristic approach to schedule a sales submit. *PATAT 2000*. Springer; 2008. pp. 176–190.
 102. Ong Y-S, Lim M-H, Zhu N, *et al.* Classification of adaptive memetic algorithms: a comparative study. *IEEE Trans Syst Man Cybern Part B* 2006;36(1):141–152.
 103. Moscato P. *Memetic algorithms: a short introduction*. New ideas in optimization. McGraw-Hill; 1999. pp. 219–234.
 104. Krasnogor N. Self generating metaheuristics in bioinformatics: the proteins structure comparison case. *Genet Program Evol Mach* 2004;5(2):181–201.
 105. Krasnogor N, Blackburne BP, Burke EK, *et al.* Multimeme algorithms for protein structure prediction. *Parallel problem*

- solving from nature VII. Springer; 2002. pp. 769–778.
106. Krasnogor N, Gustafson SM. A study on the use of “self-generation” in memetic algorithms. *Nat Comput* 2004;3(1):53–76.
 107. Smith JE. Co-evolution of memetic algorithms: initial investigations. Parallel problem solving from nature VII. Springer; 2002. pp. 537–548.
 108. Smith JE. Coevolving memetic algorithms: a review and progress report. *IEEE Trans Syst Man Cybern Part B* 2007;37(1):6–17.
 109. Moscato P, Cotta C. A gentle introduction to memetic algorithms. *Handbook of metaheuristics*. Boston (MA): Kluwer Academic Publishers; 2003. pp. 105–144.
 110. Moscato P, Mendes A, Cotta C. Memetic algorithms. New optimization techniques in engineering. Berlin Heidelberg: Springer; 2004. pp. 53–85.
 111. Moscato P, Cotta C. Memetic algorithms. In: Gonzalez T, editor. *Handbook of approximation algorithms and metaheuristics*. Boca Raton, FL: Taylor & Francis; 2006.
 112. Hart WE, Krasnogor N, Smith JE. Volume 166, Recent advances in memetic algorithms. Springer; 2005.
 113. Cotta C, Fernández AJ. Memetic algorithms in planning, scheduling, and timetabling. *Evolutionary scheduling*. Springer; 2007. pp. 1–30.
 114. Rabbani M, Rahimi-Vahed A, Torabi SA. Real options approach for a mixed-model assembly line sequencing problem. *Int J Adv Manuf Technol* 2008;37(11–12):1209–1219.
 115. Tavakkoli-Moghaddam R, Rahimi-Vahed AR. A memetic algorithm for multi-criteria sequencing problem for a mixed-model assembly line in a JIT production system. 2006 IEEE Congress on Evolutionary Computation (CEC'2006). IEEE; 2006. pp. 10350–10355.
 116. Tseng HE, Wang WP, Shih HY. Using memetic algorithms with guided local search to solve assembly sequence planning. *Expert Syst Appl* 2007;33(2):451–467.
 117. Amaya JE, Cotta C, Fernández AJ. A memetic algorithm for the tool switching problem. *Hybrid metaheuristics 2008*. Springer; 2008. pp. 190–202.
 118. Chen J-H, Chen J-H. Multi-objective memetic approach for flexible process sequencing problems. *GECCO-2008 late-breaking papers*. ACM Press; 2008. pp. 2123–2128.
 119. Muruganandam A, Prabhakaran G, Asokan P, *et al.* A memetic algorithm approach to the cell formation problem. *Int J Adv Manuf Tech* 2005;25(9–10):988–997.
 120. Tavakkoli-Moghaddam R, Safaei N, Babakhani M. Solving a dynamic cell formation problem with machine cost and alternative process plan by memetic algorithms. *International symposium on stochastic algorithms: foundations and applications*. LNCS; 2005.
 121. Baskar N, Asokan P, Saravanan R, *et al.* Selection of optimal machining parameters for multi-tool milling operations using a memetic algorithm. *J Mater Process Tech* 2006;174(1–3):239–249.
 122. Neammanee P, Reodecha M. A memetic algorithm-based heuristic for a scheduling problem in printed circuit board assembly. *Comput Ind Eng* 2009;56(1):294–305.
 123. Yeh W-C. An efficient memetic algorithm for the multi-stage supply chain network problem. *Int J Adv Manuf Tech* 2006; 29(7–8):803–813.
 124. Rabbani M, Layegh J, Ebrahim RM. Determination of number of kanbans in a supply chain system via Memetic algorithm. *Adv Eng Softw* 2009;40(6):431–437.
 125. Schoenauer M, Saveant P, Vidal V. Divide-and-evolve: a new memetic scheme for domain-independent temporal planning. *Evolutionary computation in combinatorial optimization*. Springer; 2006. pp. 247–260.
 126. Dias J, Captivo ME, Climaco J. A memetic algorithm for multi-objective dynamic location problems. *J Glob Optimiz* 2008;42(2): 221–253.
 127. Franca PM, Gupta JND, Mendes AS, *et al.* Evolutionary algorithms for scheduling a flowshop manufacturing cell with sequence dependent family setups. *Comput Ind Eng* 2005;48:491–506.
 128. Franca PM, Tin G, Buriol LS. Genetic algorithms for the no-wait flowshop sequencing problem with time restrictions. *Int J Prod Res* 2006;44(5):939–957.
 129. Liu B, Wang L, Jin Y. An effective PSO-based memetic algorithm for flow shop scheduling. *IEEE Trans Syst Man Cybern Part B* 2007;37(1):18–27.
 130. Liu B, Wang L, Jin Y-H. An effective hybrid particle swarm optimization for no-wait flow shop scheduling. *Int J Adv Manuf Tech* 2007;31(9–10):1001–1011.

131. Moscato P, Mendes A, Cotta C. Scheduling and production & control. New optimization techniques in engineering. Springer; 2004. pp. 655–680.
132. Pan Q-K, Wang L, Qian B. A novel multi-objective particle swarm optimization algorithm for no-wait flow shop scheduling problems. *J Eng Manuf* 2008;222(4):519–539.
133. Sevaux M, Jouglet A, Oduz C. MLS + CP for the Hybrid flowshop scheduling problem. In: Workshop on the Combination of Metaheuristic and Local Search with Constraint Programming Techniques; 2005.
134. Sevaux M, Jouglet A, Oduz C. Combining constraint programming and memetic algorithm for the hybrid flowshop scheduling problem. In: ORBEL 19th Annual Conference of the SOGESCI-BVWB; Louvain-la-Neuve, Belgium; 2005.
135. Qian B, Wang L, Huang DX, *et al.* Multi-objective no-wait flow-shop scheduling with a memetic algorithm based on differential evolution. *Soft Comput* 2009;13(8–9):847–869.
136. Qian B, Wang L, Huang DX, *et al.* An effective hybrid DE-based algorithm for flow shop scheduling with limited buffers. *Int J Prod Res* 2009;47(1):1–24.
137. Zobolas GI, Tarantilis CD, Ioannou G. Minimizing makespan in permutation flow shop scheduling problems using a hybrid metaheuristic algorithm. *Comput Oper Res* 2009;36(4):1249–1267.
138. Layegh J, Jolai F, Amalnik MS. A memetic algorithm for minimizing the total weighted completion time on a single machine under step-deterioration. *Adv Eng Softw* 2009;40(10):1074–1077.
139. Caumond A, Lacomme P, Tchernev N. A memetic algorithm for the job-shop with time-lags. *Comput OR* 2008;35(7):2331–2356.
140. González MA, Vela CR, Sierra MR, *et al.* Comparing schedule generation schemes in memetic algorithms for the job shop scheduling problem with sequence dependent setup times. 5th Mexican International Conference on Artificial Intelligence. Springer; 2006. pp. 472–482.
141. Gonzalez MA, Vela CR, Varela R. Scheduling with memetic algorithms over the spaces of semi-active and active schedules. *Artificial intelligence and soft computing*. Springer; 2006. pp. 370–379.
142. González-Rodríguez I, Vela CR, Puente J. A memetic approach to fuzzy job shop based on expectation model. In: 2007 IEEE International Conference on Fuzzy Systems. London, UK; 2007. pp. 1–6.
143. Qian B, Wang L, Huang DX, *et al.* Scheduling multi-objective job shops using a memetic algorithm based on differential evolution. *Int J Adv Manuf Tech* 2008;35:9–10.
144. Varela R, Puente J, Vela CR. Some issues in chromosome codification for scheduling with genetic algorithms. Planning, scheduling and constraint satisfaction: from theory to practice. IOS Press; 2005. pp. 1–10.
145. Varela R, Serrano D, Sierra M. New codification schemas for scheduling with genetic algorithms. *Artificial intelligence and knowledge engineering applications: a bioinspired approach*. Springer; 2005. pp. 11–20.
146. Yang J-H, Sun L, Lee HP, *et al.* Clonal selection based memetic algorithm for job shop scheduling problems. *J Bionic Eng* 2008;5(2):111–119.
147. Chiang TC, Fu LC. A rule-centric memetic algorithm to minimize the number of tardy jobs in the job shop. *Int J Prod Res* 2008;46(24):6913–6931.
148. Ho NB, Tay JC. Solving multiple-objective flexible job shop problems by evolution and local search. *IEEE Trans Syst Man Cybern Part C-Appl Rev* 2008;38(5):674–685.
149. Khafa F, Duran B. Parallel memetic algorithms for independent job scheduling in computational grids. Recent advances in evolutionary computation for combinatorial optimization. Springer; 2008. pp. 219–239.
150. Garcia VJ, Franca PM, Mendes A, *et al.* A parallel memetic algorithm applied to the total tardiness machine scheduling problem. In: 20th International Parallel and Distributed Processing Symposium. Rhodes, Greece; 2006.
151. Chen AHL, Chyu C-C. A memetic algorithm for maximizing net present value in resource-constrained project scheduling problem. 2008 IEEE World Congress on Computational Intelligence. IEEE Press; 2008. pp. 2401–2408.
152. Boudia M, Prins C. A memetic algorithm with dynamic population management for an integrated production-distribution problem. *Eur J Oper Res* 2009;195(3):703–715.
153. Tavakkoli-Moghaddam R, Safaei N, Sassani F. A memetic algorithm for the flexible flow line scheduling problem with processor blocking. *Comput Oper Res* 2009;36(2):402–414.

154. Tseng HE, Chen MH, Chang CC, *et al.* Hybrid evolutionary multi-objective algorithms for integrating assembly sequence planning and assembly line balancing. *Int J Prod Res* 2008;46(21):5951–5977.
155. Geismar HN, Laporte G, Lei L, *et al.* The integrated production and transportation scheduling problem for a product with a short lifespan. *Inform J Comput* 2008;20(1):21–33.
156. Maheswaran R, Ponnambalam SG, Aravindan C. A meta-heuristic approach to single machine scheduling problems. *Int J Adv Manuf Tech* 2005;25(7–8):772–776.
157. Li J, Kwan RSK. A self adjusting algorithm for driver scheduling. *J Heuristics* 2005;11(4):351–367.
158. Petrovic S, Patel V, Yang Y. Examination timetabling with fuzzy constraints. *Practice and theory of automated timetabling V*. Springer; 2005. pp. 313–333.
159. Aickelin U, White P. Building better nurse scheduling algorithms. *Ann Oper Res* 2004; 128:159–177.
160. Burke EK, De Causmaecker P, van den Berghe G. Novel metaheuristic approaches to nurse rostering problems in Belgian hospitals. *Handbook of scheduling: algorithms, models, and performance analysis*. Chapman Hall/CRC Press; 2004. pp. 44.1–44.18.
161. Özcan E. Memetic algorithms for nurse rostering. *Computer and information sciences - ISCIS 2005, 20th International Symposium (ISCIS)*. Springer; 2005. pp. 482–492.
162. Beddoe G, Petrovic S, Li JP. A hybrid meta-heuristic case-based reasoning system for nurse rostering. *J Sched* 2009;12(2):99–119.
163. Schönberger J, Mattfeld DC, Kopfer H. Memetic algorithm timetabling for non-commercial sport leagues. *Eur J Oper Res* 2004;153:102–116.
164. Semet Y, Schoenauer M. An efficient memetic, permutation-based evolutionary algorithm for real-world train timetabling. *Proceedings of the 2005 Congress on Evolutionary Computation*. IEEE Press; 2005. pp. 2752–2759.
165. Lewis R, Paechter B. Finding feasible timetables using group-based operators. *IEEE Trans Evol Comput* 2007;11(3):397–413.
166. Petrovic S, Burke EK. University timetabling. In: Leung J, editor. *Handbook of scheduling: algorithms, models, and performance analysis*. Chapman Hall/CRC Press; 2004.
167. Rossi-Doria O, Paechter B. A memetic algorithm for university course timetabling. *Combinatorial optimisation 2004 book of abstracts*. Lancaster University; 2004. pp. 56.
168. Tesfaldet BT. Automated lecture timetabling using a memetic algorithm. *Asia Pacif J Oper Res* 2008;25(4):451–475.
169. Jat SN, Yang S. A memetic algorithm for the university course timetabling problem. *Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence*. Washington (DC): IEEE Computer Society; 2008.
170. Merz P, Katayama K. Memetic algorithms for the unconstrained binary quadratic programming problem. *Biosystems* 2004; 78(1–3):99–118.
171. Gallardo JE, Cotta C, Fernandez AJ. A hybrid model of evolutionary algorithms and branch-and-bound for combinatorial optimization problems. *2005 Congress on Evolutionary Computation*. IEEE Press; 2005. pp. 2248–2254.
172. Puchinger J, Raidl GR, Pferschy U. The core concept for the Multidimensional Knapsack Problem. *Evolutionary computation in combinatorial optimization*. Springer; 2006. pp. 195–208.
173. Gallardo JE, Cotta C, Fernández AJ. A memetic algorithm for the low autocorrelation binary sequence problem. *GECCO '07: proceedings of the 9th annual conference on genetic and evolutionary computation conference*. ACM Press; 2007. pp. 1226–1233.
174. Borschbach M, Exeler A. A Tabu history driven crossover operator design for memetic algorithm applied to Max-2SAT-problems. In: *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. Atlanta (GA): 2008. pp. 605–606.
175. Qasem M, Prugel-Bennett A. Complexity of Max-SAT using stochastic algorithms. *GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. Seattle (WA): ACM Press; 2008. pp. 615–616.
176. Aickelin U, Adewunmi A. Simulation optimization of the crossdock door assignment problem. In: *UK Operational Research Society Simulation Workshop 2006 (SW 2006)*; 2006.
177. Lim A, Rodrigues B, Zhu Y. Airport gate scheduling with time windows. *Artif Intell Rev* 2005;24(1):5–31.

178. Cosmin D, Hao J-K, Kuntz P. Diversity control and multi-parent recombination for evolutionary graph coloring. In: Cowling CCaP, editor. *Evolutionary computation in combinatorial optimization*. Tübingen: Springer; 2009. pp. 121–132.
179. Bärecke T, Detyniecki M. Memetic algorithms for inexact graph matching. 2007 IEEE Congress on Evolutionary Computation. IEEE Press; 2007. pp. 4238–4245.
180. Chen XS, Lim MH, Wunsch DC II. A memetic algorithm configured via a problem solving environment for the hamiltonian cycle problems. 2007 IEEE Congress on Evolutionary Computation. IEEE Press; 2007. pp. 2766–2773.
181. Wang J. A memetic algorithm with genetic particle swarm optimization and neural network for maximum cut problems. *International conference on life system modeling and simulation*. Springer; 2007. pp. 297–306.
182. Drezner Z. Extensive experiments with hybrid genetic algorithms for the solution of the quadratic assignment problem. *Comput Oper Res* 2008;35(3):717–736.
183. Tang J, Lim MH, Ong Y-S, *et al.* Parallel memetic algorithm with selective local search for large scale quadratic assignment problems. *Int J Innov Comput Inf Control* 2006;2(6):1399–1416.
184. Bouly H, Dang D-C, Moukrim A. A memetic algorithm for the team orienteering problem. *Applications of evolutionary computing*. Springer; 2008. pp. 649–658.
185. Creput JC, Koukam A. The memetic self-organizing map approach to the vehicle routing problem. *Soft Comput* 2008;12(11):1125–1141.
186. Cruz-Chavez MA, Díaz-Parra O, Juárez-Romero D, *et al.* Memetic algorithm based on a constraint satisfaction technique for VRPTW. 9th Artificial Intelligence and Soft Computing Conference. Springer; 2008. pp. 376–387.
187. Fleury G, Lacomme P, Prins C. Evolutionary algorithms for stochastic arc routing problems. *Applications of evolutionary computing*. Springer; 2004. pp. 501–512.
188. Kubiak M, Wesolek P. Accelerating local search in a memetic algorithm for the capacitated vehicle routing problem. *Evolutionary computation in combinatorial optimization*. Springer; 2007. pp. 96–107.
189. Lacomme P, Prins C, Ramdane-Cherif W. Competitive memetic algorithms for arc routing problems. *Ann Oper Res* 2004; 131(1–4):159–185.
190. Lacomme P, Prins C, Ramdane-Cherif W. Evolutionary algorithms for periodic arc routing problems. *Eur J Oper Res* 2005;165(2):535–553.
191. Tavakkoli-Moghaddam R, Saremi AR, Ziaee MS. A memetic algorithm for a vehicle routing problem with backhauls. *Appl Math Comput* 2006;181(2):1049–1060.
192. Tricoire F. Vehicle and personnel routing optimization in the service sector: application to water distribution and treatment. *4OR-A Q J Oper Res* 2007;5(2):165–168.
193. Labadi N, Prins C, Reghioui M. A memetic algorithm for the vehicle routing problem with time windows. *Rairo-Oper Res* 2008;42(3):415–431.
194. Mendoza JE, Medaglia AL, Velasco N. An evolutionary-based decision support system for vehicle routing: the case of a public utility. *Decis Support Syst* 2009;46(3):730–742.
195. Fallahi AE, Prins C, Calvo RW. A memetic algorithm and a tabu search for the multi-compartment vehicle routing problem. *Comput Oper Res* 2008;35(5):1725–1741.
196. Velasco N, Castagliola P, Dejax P, *et al.* A memetic algorithm for a pick-up and delivery problem by helicopter. *Bio-inspired algorithms for the vehicle routing problem*. Berlin Heidelberg: Springer; 2008.
197. Zak J, Jaszkievicz A, Redmer A. Multiple criteria optimization method for the vehicle assignment problem in a bus transportation company. *J Adv Transp* 2009;43(2):203–243.
198. Fischer T, Merz P. A memetic algorithm for the optimum communication spanning tree problem. *Hybrid metaheuristics 2007*. Springer; 2007. pp. 170–184.
199. Rocha DAM, Goldbarg EFG, Goldbarg MC. A memetic algorithm for the biobjective minimum spanning tree problem. *Evolutionary computation in combinatorial optimization*. Springer; 2006. pp. 222–233.
200. Liu Y-H. A memetic algorithm for the probabilistic traveling salesman problem. 2008 IEEE World Congress on Computational Intelligence. IEEE Press; 2008. pp. 146–152.
201. Nguyen HD, Yoshihara I, Yamamori K, *et al.* Implementation of an effective hybrid GA for large-scale traveling salesman problems. *IEEE Trans Syst Man Cybern Part B* 2007;37(1):92–99.
202. Wang Y, Qin J. A memetic-clustering-based evolution strategy for traveling salesman

- problems. 2nd International Conference on Rough Sets and Knowledge Technology. Springer; 2007. pp. 260–266.
203. Bontoux B, Artigues C, Feillet D. Memetic algorithm with a large neighborhood crossover operator for the generalized traveling salesman problem. EU/MEeting. Universite de Tolouss; 2008.
 204. Creput JC, Koukam A. A memetic neural network for the Euclidean traveling salesman problem. Neurocomputing 2009;72(4–6):1250–1264.
 205. Gutin G, Karapetyan D, Krasnogor N. Memetic algorithm for the generalized asymmetric traveling salesman problem. Computational Intelligence (SCI). Berlin Heidelberg: Springer; 2008.
 206. Jaskiewicz A, Zielniewicz P. Pareto memetic algorithm with path relinking for bi-objective traveling salesperson problem. Eur J Oper Res 2009;193(3):885–890.
 207. Samanlioglu F, Ferrell WG, Kurz ME. A memetic random-key genetic algorithm for a symmetric multi-objective traveling salesman problem. Comput Ind Eng 2008; 55(2):439–449.
 208. Cotta C, Dotu I, Fernandez AJ, *et al.* A memetic approach to Golomb rulers. Parallel problem solving from nature IX. Springer; 2006. pp. 252–261.
 209. Cotta C, Fernandez A. A hybrid GRASP - evolutionary algorithm approach to golomb ruler search. Parallel problem solving from nature VIII. Springer; 2004. pp. 481–490.
 210. Cotta C, Dotu I, Fernandez AJ. Scheduling social golfers with memetic evolutionary programming. Hybrid metaheuristic 2006. Springer; 2006. pp. 150–161.
 211. Dantas MJP, Brito LC, de Carvalho PH. Multi-objective Memetic Algorithm applied to the automated synthesis of analog circuits, in Advances in Artificial Intelligence. Springer; 2006. pp. 258–267.
 212. Coe S, Areibi S, Moussa M. A hardware Memetic accelerator for VLSI circuit partitioning. Comput Electr Eng 2007;33(4): 233–248.
 213. Caorsi S, Massa A, Pastorino M, *et al.* Detection of PEC elliptic cylinders by a memetic algorithm using real data. Microw Opt Technol Lett 2004;43(4):271–273.
 214. Pastorino M. Stochastic optimization methods applied to microwave imaging: A review. IEEE Trans Antenn Propag 2007; 55(3, Part 1): 538–548.
 215. Tagawa K, Matsuoka M. Optimum design of surface acoustic wave filters based on the Taguchi's quality engineering with a memetic algorithm. Parallel problem solving from nature IX. Springer; 2006. pp. 292–301.
 216. Areibi S, Yang Z. Effective memetic algorithms for VLSI design = genetic algorithms plus local search plus multi-level clustering. Evol Comput 2004;12(3):327–353.
 217. Tang M, Yao X. A memetic algorithm for VLSI floorplanning. IEEE Trans Syst Man Cybern Part B 2007;37(1):62–69.
 218. Kononova AV, Hughes KJ, Pourkashanian M, *et al.* Fitness diversity based adaptive memetic algorithm for solving inverse problems of chemical kinetics. 2007 IEEE Congress on Evolutionary Computation. IEEE Press; 2007. pp. 2366–2373.
 219. Kononova AV, Ingham DB, Pourkashanian M. Simple scheduled memetic algorithm for inverse problems in higher dimensions: application to chemical kinetics. IEEE Congress on Evolutionary Computation, 2008. CEC 2008. Hong Kong: IEEE; 2008. pp. 3905–3912.
 220. Caponio A, Cascella GL, Neri F, *et al.* A fast adaptive memetic algorithm for online and offline control design of PMSM drives. IEEE Trans Syst Man Cybern Part B 2007;37(1):28–41.
 221. Caponio A, Neri F, Cascella GL, *et al.* Application of memetic differential evolution frameworks to PMSM drive design. 2008 IEEE World Congress on Computational Intelligence. Hong Kong: IEEE Press; 2008. pp. 2113–2120.
 222. Carrano EG, Souza BB, Neto OM, *et al.* An immune inspired memetic algorithm for power distribution system design under load evolution uncertainties. IEEE Congress on Evolutionary Computation. Hong Kong: IEEE Press; 2008. pp. 3251–3257.
 223. Kaveh A, Shahrouzi M. Graph theoretical implementation of memetic algorithms in structural optimization of frame bracing layouts. Eng Comput 2008;25(1–2):55–85.
 224. Song W. Multiobjective memetic algorithm and its application in robust airfoil shape optimization. Multi-objective memetic algorithms. Berlin Heidelberg: Springer; 2009.
 225. Ahmad R, Jamaluddin H, Hussain MA. Application of memetic algorithm in modelling discrete-time multivariable dynamics systems. Mech Syst Signal Process 2008; 22(7):1595–1609.

226. Tenne Y, Armfield SW. A memetic algorithm using a trust-region derivative-free optimization with quadratic modelling for optimization of expensive and noisy black-box functions. *Evolutionary computation in dynamic and uncertain environments*. Springer; 2007. pp. 389–415.
227. Özcan E, Onbasioglu E. Memetic algorithms for parallel code optimization. *Int J Parallel Program* 2007;35(1):33–61.
228. Wang Z, Sun X, Zhang D. Web document query optimization based on memetic algorithm. *IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*. Washington (DC): IEEE Computer Society; 2008.
229. Goh CK, Teoh EJ, Tan KC. A hybrid evolutionary approach for heterogeneous multiprocessor scheduling. *Soft Comput* 2009; 13(8–9):833–846.
230. Sheng W, Howells G, Fairhurst M, *et al.* A memetic fingerprint matching algorithm. *IEEE Trans Inf Forensics Secur* 2007; 2(3, Part 1): 402–412.
231. Cotta C. Scatter search and memetic approaches to the error correcting code problem. *Evolutionary computation in combinatorial optimization*. Springer; 2004. pp. 51–60.
232. Arcuri A, Yao X. A memetic algorithm for test data generation of object-oriented software. 2007 IEEE Congress on Evolutionary Computation. IEEE Press; 2007. pp. 2048–2055.
233. Hsu C-H, Chou P-H, Shyr W-J, *et al.* Optimal radiation pattern design of adaptive linear array antenna by phase and amplitude perturbations using memetic algorithms. *Int J Innov Comput Inf Control* 2007;3(5):1273–1287.
234. Hsu C-H, Shyr W-J. Optimizing linear adaptive broadside array antenna by amplitude-position perturbations using memetic algorithms. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Springer; 2005. pp. 568–574.
235. Hsu C-H, Shyr W-J. Memetic algorithms for optimizing adaptive linear array patterns by phase-position perturbations. *Circuits Syst Signal Process* 2005;24(4):327–341.
236. Hsu C-H, Shyr W-J, Chen C-H. Adaptive pattern nulling design of linear array antenna by phase-only perturbations using memetic algorithms. 1st International Conference on Innovative Computing, Information and Control. IEEE Computer Society; 2006. pp. 308–311.
237. Hsu CH. Downlink mimo-sdma optimization of smart antennas by phase-amplitude perturbations based on memetic algorithms for wireless and mobile communication systems. *Int J Innov Comput Inf Control* 2009; 5(2):443–459.
238. Karaoglu B, Topcuoglu H, Gürgen F. Evolutionary algorithms for location area management. *Applications of evolutionary computing*. Springer; 2005. pp. 175–184.
239. Quintero A, Pierre S. On the design of large-scale cellular mobile networks using multipopulation memetic algorithms. *Engineering evolutionary intelligent systems*. Springer; 2008. pp. 353–377.
240. Merz P, Wolf S. Evolutionary local search for designing peer-to-peer overlay topologies based on minimum routing cost spanning trees. *Parallel problem solving from nature IX*. Springer; 2006. pp. 272–281.
241. Neri F, Kotilainen N, Vapa M. An adaptive global-local memetic algorithm to discover resources in P2P networks. *Applications of evolutionary computing*. Springer; 2007. pp. 61–70.
242. Neri F, Kotilainen N, Vapa M. A memetic-neural approach to discover resources in P2P networks. *Recent advances in evolutionary computation for combinatorial optimization*. Springer; 2008. pp. 113–129.
243. Fischer T, Bauer K, Merz P. Distributed memetic algorithm for the routing and wavelength problem. *Parallel problem solving from nature X*. Springer; 2008. pp. 879–888.
244. Kim S-S, Smith AE, Lee J-H. A memetic algorithm for channel assignment in wireless FDMA systems. *Comput OR* 2007;34(6): 1842–1856.
245. Konstantinidis A, Yang K, Chen H-H, *et al.* Energy-aware topology control for wireless sensor networks using memetic algorithms. *Comput Commun* 2007;30(14–15): 2753–2764.
246. Cordon O, Damas S, Santamaria J. A scatter search algorithm for the 3D image registration problem. *Parallel problem solving from nature VIII*. Springer; 2004. pp. 471–480.
247. di Gesu V, Bosco GL, Millonzi F, *et al.* A memetic algorithm for binary image

- reconstruction. *Combinatorial Image Analysis*; 2008. pp. 384–395.
248. di Gesù V, Bosco GL, Millonzi F, *et al.* Discrete tomography reconstruction through a new memetic algorithm. *Applications of evolutionary computing*. Springer; 2008. pp. 347–352.
249. Fernandez E, Graña M, Ruiz-Cabello J. An instantaneous memetic algorithm for illumination correction. *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*. Portland (OR): IEEE Press; 2004. pp. 1105–1110.
250. Pastorino M, Caorsi S, Massa A, *et al.* Reconstruction algorithms for electromagnetic imaging. *IEEE Trans Instrum Meas* 2004;53(3):692–699.
251. Do A-D, Cho SY. Memetic algorithm based fuzzy clustering. *2007 IEEE Congress on Evolutionary Computation*. IEEE Press; 2007. pp. 2398–2404.
252. Inostroza-Ponta M, Berretta R, Mendes A, *et al.* An automatic graph layout procedure to visualize correlated data. In: *IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream. Artificial Intelligence in Theory and Practice*. Santiago, Chile; 2006. pp. 179–188.
253. Inostroza-Ponta M, Mendes A, Berretta R, *et al.* An integrated QAP-based approach to visualize patterns of gene expression similarity. *GECCO 04: Genetic and Evolutionary Computation Conference*. Seattle (WA): 2007.
254. Capp A, Inostroza-Ponta M, Bill D, *et al.* Is there more than one proctitis syndrome? A revisitation using data from the TROG 96.01 trial. *Radiother Oncol* 2009;90(3):400–407.
255. Sheng W, Liu X, Fairhurst MC. A memetic algorithm for simultaneously clustering and feature selection. *IEEE Trans Knowl Data Eng* 2008;20(7).
256. Sheng W, Liu X, Fairhurst M. A niching memetic algorithm for simultaneous clustering and feature selection. *IEEE Trans Knowl Data Eng* 2008;20(7):868–879.
257. Zhu Z, Ong Y-S, Dash M. Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Trans Syst Man Cybern Part B* 2007;37(1):70–76.
258. Garcia S, Cano JR, Herrera F. A memetic algorithm for evolutionary prototype selection: a scaling up approach. *Pattern Recognit* 2008;41(8):2693–2709.
259. Gal L, Botzheim J, Koczy LT. Improvements to the bacterial memetic algorithm used for fuzzy rule base extraction. In: *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. Istanbul; 2008.
260. Gál L, Botzheim J, Kóczy LT. Modified bacterial memetic algorithm used for fuzzy rule base extraction. *International Conference on Soft Computing as Transdisciplinary Science and Technology*. France ACM: Cergy-Pontoise; 2008.
261. Tirronen V, Neri F, Karkkainen T, *et al.* An enhanced memetic differential evolution in filter design for defect detection in paper production. *Evol Comput* 2008;16(4):529–555.
262. Kretowski MA. Memetic algorithm for global induction of decision trees. *34th Conference on Current Trends in Theory and Practice of Computer Science*. Springer; 2008. pp. 531–540.
263. Divina F. Hybrid genetic relational search for inductive learning. Amsterdam, The Netherlands: Department of Computer Science, Vrije Universiteit; 2004.
264. Delgado M, Cuellar MP, Pegalajar MC. Multiobjective hybrid optimization and training of recurrent neural networks. *IEEE Trans Syst Man Cybern Part B* 2008;38(2):381–403.
265. Delgado M, Pegalajar MC, Cuellar MP. Memetic evolutionary training for recurrent neural networks: an application to time-series prediction. *Expert Syst* 2006;23(2):99–115.
266. Guillén A, Pomares H, González J, *et al.* Parallel multi-objective memetic RBFNNs design and feature selection for function approximation problems. *9th International Work-Conference on Artificial Neural Networks*. Springer; 2007. pp. 341–350.
267. Hervás C, Silva M. Memetic algorithms-based artificial multiplicative neural models selection for resolving multi-component mixtures based on dynamic responses. *Chemomet Intell Lab Syst* 2007;85(2):232–242.
268. Liu B, Wang L, Jin Y, *et al.* Designing neural networks using PSO-based memetic algorithm. *4th International Symposium on Neural Networks*. Springer; 2007. pp. 219–224.
269. Martínez-Estudillo FJ, Hervás-Martínez C, Martínez-Estudillo AC, *et al.* Memetic algorithms to product-unit neural networks

- for regression. 8th International Work-Conference on Artificial Neural Networks. Springer; 2005. pp. 83–90.
270. Neruda R, Slusny S. Variants of memetic and hybrid learning of perceptron networks. 18th International Workshop on Database and Expert Systems Applications. IEEE Computer Society; 2007. pp. 158–162.
 271. Togelius J, Schaul T, Schmidhuber J, *et al.* Countering poisonous inputs with memetic neuroevolution. Parallel problem solving from nature X. Springer; 2008. pp. 610–619.
 272. Cotta C. Scatter search with path relinking for phylogenetic inference. *Eur J Oper Res* 2005;169(2):520–532.
 273. Williams TL, Smith ML. The role of diverse populations in phylogenetic analysis. GECCO 2006: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation. ACM Press; 2006. pp. 287–294.
 274. Richer JM, Goëffon A, Hao JK. A memetic algorithm for phylogenetic reconstruction with maximum parsimony. In: Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Tübingen, Germany; 2009. pp. 164–175.
 275. Pirkwieser S, Raidl GR. Finding consensus trees by evolutionary, variable neighborhood search, and hybrid algorithms. GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. ACM Press; 2008. pp. 323–330.
 276. Palacios P, Pelta D, Blanco A. Obtaining biclusters in microarrays with population-based heuristics. Applications of evolutionary computing. Springer; 2006. pp. 115–126.
 277. Cox M, Bowden N, Moscato P, *et al.* Memetic algorithms as a new method to interpret gene expression profiles in multiple sclerosis. *Mult Scler* 2007;13, Suppl 2: S205.
 278. Zhu Z, Ong Y-S. Memetic algorithms for feature selection on microarray data. 4th International Symposium on Neural Networks. Springer; 2007. pp. 1327–1335.
 279. Zhu Z, Ong Y-S, Dash M. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognit* 2007;40(11): 3236–3248.
 280. Dorrnsoro B, Alba E, Luque G, *et al.* A self-adaptive cellular memetic algorithm for the DNA fragment assembly problem. 2008 IEEE World Congress on Computational Intelligence. IEEE Press; 2008. pp. 2656–2663.
 281. Volk J, Herrmann T, Wuthrich K. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *J Biomol NMR* 2008;41(3):127–138.
 282. Bazzoli A, Tettamanzi AGB. A memetic algorithm for protein structure prediction in a 3D-lattice HP model. Applications of evolutionary computing. Springer; 2004. pp. 1–10.
 283. Cotta C. Hybrid evolutionary algorithms for protein structure prediction in the HPNX model. Computational intelligence, theory and applications. Springer; 2004. pp. 525–534.
 284. Oakley MT, Barthel D, Bykov Y, *et al.* Search strategies in structural bioinformatics. *Curr Protein Pept Sci* 2008;9(3):260–274.
 285. Zhao XC. Advances on protein folding simulations based on the lattice HP models with natural computing. *Appl Soft Comput* 2008;8(2):1029–1040.
 286. Santos EE, Santos JE. Effective computational reuse for energy evaluations in protein folding. *Int J Artif Intell Tools* 2006;15(5):725–739.
 287. Noman N, Iba H. Inferring Gene regulatory networks using differential evolution with local search heuristics. *IEEE/ACM Trans Comput Biol Bioinformatics* 2007;4(4): 634–647.
 288. Spieth C, Streichert F, Supper J, *et al.* Feedback memetic algorithms for modeling gene regulatory networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2005). IEEE Press; 2005. pp. 61–67.
 289. Yang CH, Cheng YH, Chuang LY, *et al.* Specific PCR product primer design using memetic algorithm. *Biotechnol Prog* 2009;25(3):745–753.
 290. Romero-Campero FJ, Cao H, Camara M, *et al.* Structure and parameter estimation for cell systems biology models. GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. ACM Press; 2008. pp. 331–338.
 291. Neri F, Toivanen J, Cascella GL, *et al.* An adaptive multimeme algorithm for designing HIV multidrug therapies. *IEEE/ACM Trans Comput Biol Bioinformatics* 2007;4(2):264–278.

292. Tse S-M, Liang Y, Leung K-S, *et al.* A memetic algorithm for multiple-drug cancer chemotherapy schedule optimization. *IEEE Trans Syst Man Cybern Part B* 2007; 37(1):84–91.
293. Santamaria J, Cordon O, Damas S, *et al.* Performance evaluation of memetic approaches in 3D reconstruction of forensic objects. *Soft Comput* 2009;13(8–9):883–904.
294. Lumanpauw E, Pasquier M, Quek C. MNFS-FPM: a novel memetic neuro-fuzzy system based financial portfolio management. 2007 IEEE Congress on Evolutionary Computation. IEEE Press; 2007. pp. 2554–2561.
295. Aranha C, Iba H. Application of a memetic algorithm to the portfolio optimization problem. In: 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence. Auckland, New Zealand; 2008.
296. Maringer DG. Finding the relevant risk factors for asset pricing. *Comput Stat Data Anal* 2004;47(2):339–352.
297. Boughaci D, Benhamou B, Drias H. A memetic algorithm for the optimal winner determination problem. *Soft Comput* 2009; 13(8–9):905–917.
298. Chiam SC, Tan KC, Mamun AM. A memetic model of evolutionary PSO for computational finance applications. *Expert Syst Appl* 2009;36(2):3695–3711.
299. Florez-Revuelta F, Casado-Diaz JM, Martinez-Bernabeu L, *et al.* A memetic algorithm for the delineation of local labour markets. *Parallel problem solving from nature X*. Springer; 2008. pp. 1011–1020.
300. Sudholt D. The impact of parameterization in memetic evolutionary algorithms. *Theor Comput Sci* 2009;410(26):2511–2528.
301. Cotta C, Moscato P. The parameterized complexity of multiparent recombination. *Proceedings of the 6th Metaheuristic International Conference*. Vienna: Universitt Wien; 2005.

METAHEURISTICS FOR STOCHASTIC PROBLEMS

LARS MAGNUS HVATTUM
EYSTEIN FREDRIK ESBENSEN

Department of Industrial Economics and
Technology Management, Norwegian
University of Science and
Technology, Trondheim, Norway

The story of metaheuristics is one of success, especially when considering their application to computationally intractable optimization problems. At the same time, the understanding that many real-world problems are inherently stochastic has led to research on how this stochasticity should be modeled and handled to provide useful support for decision makers. Metaheuristics are natural candidates when stochastic problems are considered, given their history of solving difficult optimization problems. In this article, we assume that the reader is familiar with the basics of metaheuristics and has some general understanding of stochastic problems. Good introductions to metaheuristics can be found in Refs. 1–3.

Stochastic optimization problems are simply problems where some element of randomness is involved. This includes a wide variety of problems, and the numerous ways of modeling such problems reflect this variety. Among the most typical models are *stochastic programming problems* (SPPs) [4] and *Markov decision processes* (MDPs) [5], and exact solution methods for SPPs and MDPs have been studied by many researchers. For some stochastic problems, typically arising from *chance-constrained programs* (CCPs) or in robust optimization [6], deterministic equivalents exist that allow the problem to be handled as a normal deterministic problem. When using metaheuristics, one relies less on a proper mathematical formulation of the problem at hand; a formal model may be used to describe the problem, but is not a requirement. As the burden of proving optimality is relaxed,

a heuristic solution method has additional degrees of freedom when representing the reality and presenting solutions to the decision maker. The problem can then readily be documented with nonstandard models, such as eclectic models combining SPPs and MDPs [7]. Sometimes when applying metaheuristics a formalized model is omitted, especially if the model would be too intricate to interpret and too complex to solve using standard exact methods, and instead, the problem is described verbally. The field of *simulation optimization* [8] is appropriate when a simulation model can be built for a problem that is otherwise too complex to model mathematically. Multiobjective models often arise naturally in stochastic settings, either as a method to simultaneously consider solution quality, robustness, and flexibility [9], or to consider a trade-off between violating soft constraints [10].

Considering the popularity of metaheuristics and the increasing focus on stochastic problems, there is no surprise that there is already a decent amount of relevant publications to be found. It is not within the scope of this article to cover the entire field, and readers searching for further references can consider a recent survey on metaheuristics for stochastic combinatorial problems [11] and a 2005 survey [12] on evolutionary optimization in uncertain environments. We will focus on *how metaheuristics must be adapted when applied to stochastic problems*: what are the main issues that need to be considered when using metaheuristics to solve stochastic problems, as opposed to deterministic problems.

The field of metaheuristics now comprises a large number of methods and variations. Combined with the trend of making hybrid methods, it is increasingly difficult to classify metaheuristics in any meaningful way. Nevertheless, we choose to divide our discussion of how to apply metaheuristics to stochastic problems into three parts: First, we discuss local search-based metaheuristics and how these can be adapted to handle stochastic problems in the section titled

“Local Search-Based Metaheuristics”. Next, construction-based metaheuristics are examined in the second section. Then, we look at population-based metaheuristics in the subsequent section. The techniques discussed in any of these three parts should, however, be considered as potential improvements for methods that are described in the other parts; there are still ample opportunities to benefit from combining ideas of different camps of metaheuristics researchers. In the section titled “Conclusions” we present our concluding remarks.

LOCAL SEARCH-BASED METAHEURISTICS

The foundation of local search is the ability to evaluate a huge number of candidate solutions generated by applying neighborhood operators to a current solution. For deterministic problems, evaluating an arbitrary solution can usually be accomplished in polynomial time, and evaluating a neighboring solution can often be done in constant time when proper data structures are used. For stochastic problems, the outcome of one variable might be dependent on the outcome of another variable in such a way that the exact evaluation of a solution will take an exponential amount of time [13]. It is sometimes possible to do better within a local search. An example of this is Refs 14 and 15 where the evaluation of a single solution of one particular stochastic problem is reduced from an exponential number to $O(n^3)$ operations. Furthermore, it is shown how the complete neighborhood of $O(n^2)$ neighbors can also be evaluated using only $O(n^3)$ operations. However, as illustrated by Bianchi and Campbell [16], finding efficient ways to explore neighborhoods, and making sure that the resulting formulas are correct, is challenging.

When an exact evaluation of a neighboring solution is impossible, either due to the computational effort required or due to the lack of an explicit evaluation function altogether, there are two main alternatives [11]: to use an ad hoc approximation of the objective function, or to use sampling to estimate the value of the solution. For both alternatives there is a trade-off between precision

and speed, considering the computational burden of different ad hoc approximations or the number of samples used in estimation. In Ref. 17 two different proxies to replace an exact move evaluation are tested, and the most simplifying approximation seems to give the best results after a specific fixed time limit, at least for the particular problem instances considered. The use of sampling usually implies that the solution must be evaluated from scratch, but Ref. 18 gives an interesting example where sampling is used within delta evaluations. In delta evaluations, one evaluates only the changes in the solution that result from making a move, so that only a partial solution is evaluated. While efficient exact evaluation techniques often rely on the order in which neighbors are examined and require that the full neighborhood is explored [15,16], the use of sample-based delta evaluations allows efficient use of neighborhood reduction techniques.

Tabu Search

An early use of tabu search (TS) for a stochastic problem is shown in Ref. 19. Since the exact evaluation of a move is too time consuming, it is proposed to use a proxy for the objective function. It is not required that the proxy is a good approximation of the objective function, but rather that it shows a strong correlation with the change in the objective when performing a move. It was found important to compute the true value of the objective function for each of the five best neighbors determined by the proxy when selecting a move.

In Ref. 20, a very simple TS is coupled with a simulation model that is used to calculate the objective function. The only mechanisms included are tabu tenure and the standard aspiration criterion of accepting a new best solution. At each iteration, the complete neighborhood is explored and the best neighbor is selected as the new solution. Another application using simulation to evaluate moves is for determining buffer sizes in production lines [21]. Knowledge of the problem domain, realizing the importance of bottlenecks in the production line, is used to create a candidate list strategy where only a subset of the neighborhood is evaluated at each iteration. The size of the subset of

evaluated moves increases when improving moves are made and decreases otherwise, to speed up the search. When no improvements are found, different neighbors are considered by allowing bigger changes to be made to the solution through each move.

Allowing infeasible intermediate solutions during the search is sometimes beneficial in TS [22]. This was also found to be the case in Ref. 23, where a TS was designed for solving CCPs. Simulation was used to find approximate probabilities that each constraint was violated and two alternative move evaluation functions were tested: allowing infeasible solutions with a penalty gave better results than avoiding infeasible solutions altogether.

An alternative tabu criterion is suggested in Ref. 17: the cost evaluation of a non-tabu move is skipped with a low probability, whereas the cost evaluation of a tabu move is skipped with a high probability. An evaluated non-tabu move is accepted if an improvement is made, whereas an evaluated tabu move is accepted only if it leads to a new best solution. The approach is unfortunately not compared with a conventional TS implementation.

Simulated Annealing

Two techniques that can be used in simulated annealing (SA) for problems involving solutions evaluated through simulations are an acceptance criterion based on confidence intervals [24] and a constant temperature [25]. The acceptance criterion implies that the probability of accepting a nonimproving move is larger when the sample size is smaller. A constant temperature greater than zero gives the search freedom to aggressively move around the search space and locate good solutions quickly, although the Markov chain associated with the search may not converge to the set of global optimal solutions. With a combination of the two techniques, it can be shown that the most often visited solution in an SA converges almost surely to the optimal solution [26].

If the noise in the objective function evaluation can be reduced gradually during the search, for example by increasing the sample size, convergence proofs even for the current solution remain valid [27]. This

means that less effort can be used to estimate objective functions early in the search, and that the accuracy should be increased when the temperature is decreased. In Ref. 28 the error in sampling is used instead of temperature, which gives very similar behavior if the error has a normal distribution: the effective temperature corresponds to the standard deviation of the estimated change in objective function value for the move in question. Increasing the number of samples corresponds to reducing the temperature, but it was observed that a smoother annealing could be obtained by directly controlling the effective temperature.

The case of solving multiobjective stochastic combinatorial optimization problems using SA was investigated in Ref. 29. The SA follows an implementation for deterministic problems, but uses estimation based on sampling whenever an evaluation of the objective function is required. To spend more effort in crucial moments, three different sample sizes are used, depending on whether a solution is considered for inclusion in the approximation of the pareto front, for move acceptance, or for comparing with other solutions for which the local search is performed in parallel. Inclusion in the approximated pareto front is only considered when the temperature has fallen below a certain threshold.

Iterated Local Search

Iterated local search (ILS) is used in Ref. 18 for the *probabilistic traveling salesman problem* (PTSP). The ILS is based on sample-based delta evaluations, but otherwise follows a standard implementation. The sampled demand realizations used in the delta evaluation are kept throughout the search. Sampling the realizations anew after each perturbation did not yield significantly different results. For some problem instances, the variance of the cost estimates in the delta evaluations is high. Instead of just increasing the number of samples from the offset, Balaprakash *et al.* [30] examine an adaptive sample size as well as the use of importance sampling. The adaptive sampling size involves adding samples until a student's t-test shows significance or an upper limit on the number of samples is reached.

Importance sampling is a general variance reduction technique. It entails sampling from a biased distribution, making events with very low probability appear more often, and then correcting the result with likelihood ratios. The latter proves valuable for problem instances where some of the customers has only a small probability of appearing.

Variable Neighborhood Search

A variable neighborhood search (VNS) for stochastic combinatorial optimization problems is developed in Ref. 31. The focus is on problems where the objective function has a random component and is difficult or even impossible to compute. The basic VNS method is followed, but is extended by using sample average estimates as an estimator of the true objective function value for any given solution. An additional step is added to compare the current solution with the best solution found so far, again using sample average estimates but with a potentially larger sample size. A convergence proof is given, although as is typically the case for metaheuristics, a realistic implementation would not adhere to the assumptions of the proof. An important point is that the sample size used when challenging the best solution found should increase for every round of the search.

CONSTRUCTION-BASED METAHEURISTICS

While both local search-based metaheuristics and evolutionary metaheuristics have received much attention, construction-based metaheuristics have to some degree been overlooked. Constructions can conveniently allow diversification, at the expense of intensification, which may be an advantage in certain complex search spaces that can arise in stochastic problems.

Ant Colony Optimization

An ant system-based algorithm for infinite horizon discounted reward MDPs was described in Ref. 32. An ant will construct a policy by finding a walk in the following directed construction graph: each state of the Markov process is a node in the construction

graph, the out-going arcs from a node correspond to the possible actions to take when present in that state. All arcs going from one node will enter one other node, based on an arbitrary sequence of states. When the ant reaches the final node in the graph, it will have selected one out-going arc from each of the nodes corresponding to each of the states, thus representing a complete policy. An aspect of this ant system-based algorithm that is special to the application to MDPs is that of generating an elite policy for each generation of ants. This elite policy is a combination of the previous elite policy and all ants in the previously completed generation. By using the concept of policy switching [33], the elite policy is guaranteed to be the best policy found so far and to improve monotonically. This gives a basis for a probabilistic convergence proof, but the method has not yet been tested empirically. Another untested ant system-method, for MDPs with unknown models, can be found in Ref. 34.

Empirical results for ant colony optimization (ACO) on stochastic problems are mostly limited to the PTSP. One of the major issues that have been the focus of research is whether analytical computations or empirical estimation of the solution quality is more efficient [35]. Currently, the use of empirical estimation techniques has the upper hand. However, the use of ACO for this problem relies heavily on the use of local search, and progress in local search techniques [18,30] is closely related to the efficacy of estimation-based ACO.

For problems involving simulations or sampling to determine solution quality, an ant colony-method was outlined in Ref. 36. An interesting challenge is to compute the visibility values that are used by the ants in addition to the deposited pheromone when creating walks in the construction graph. As certain variables depend on random influence, it is suggested that the visibility values can be based either on taking expected values of the variables or on a sample of the random variables that changes for every round. In the tests performed, however, visibility values were not used. A round winner is determined based on evaluating all walks generated using a single random

sample. When the round winner is compared to the current best solution, a more involved test is used based on a larger set of samples. Both the round winner and the current best solution are reinforced through pheromone increments, although convergence proofs have been found only for the case when the round winner is not reinforced. Several alternatives for determining the current best solution were examined in Ref. 37, such as using a single sample, using several samples as in Ref. 36, and using a series of nonparametric statistical tests in a procedure called F-Race. This procedure is a general method for comparison of a number of candidates and selecting the best. Here, it was used to determine the best solution among all the solutions produced at the current iteration and the previous best solution. The use of F-Race seems to dominate both the use of a large iteration-dependent number of samples as well as the use of an adaptive number of samples based on parametric tests. However, when the variance of the solution value is low, it seems best to use the strategy of using only a single sample to determine the best solution.

In Ref. 29, ACO is used to solve multiobjective stochastic combinatorial optimization problems, where the objective function values are found through sample-based estimation. Round winners are determined by comparing the weighted sum of each objective function, using a single scenario. A global best solution subject to a particular random weighting of the objective functions is determined among round winners by averaging the weighted sum of objective functions over a number of scenarios. Global best solutions are then considered for inclusion in the approximation of the pareto front, based on an evaluation over a third set of scenarios.

Greedy Randomized Adaptive Search Procedure

The *stochastic inventory routing problem* (SIRP) has been formulated as an infinite horizon discounted reward MDP. A method based on Greedy Randomized Adaptive Search Procedure (GRASP) was developed for the SIRP in Ref. 38. The main idea was that the stochastics and dynamics

of this particular infinite MDP could be approximated by a finite scenario tree. A scenario tree would be created for each state encountered, and by controlling the width and height of the tree, a balance would be found between the computational effort required, the precision of the representation of the stochasticity, and the avoidance of end-of-horizon effects. Both the action space and the state space of the problem are large, so all proposed solution methods must be evaluated using simulations.

In the typical GRASP implementation one is allowed to fix any one decision at each step of the construction, which would imply that actions at every node of the tree could be modified at each step of the construction. Since the scenario tree consists of a large number of nodes, it is costly to reevaluate decisions at all nodes affected by the last decision made. An alternative that was shown effective in Ref. 38 was to perform the GRASP in a top-down fashion on the scenario tree, thus first fixing all decisions at the root node before continuing recursively down the tree. Of importance when applying this strategy is that the GRASP must include learning strategies [39], to make sure that information based on the actions selected at the bottom of the tree will eventually influence decisions at higher nodes. The usual conduct of GRASP is that once a variable has been assigned a value in the construction, that value will remain fixed for the remainder of the construction. A softer construction step was adopted where a value assigned could still be increased at later steps of the construction, which slightly reduced the consequences of each construction step, especially when the recursive top-down construction was not in force. The GRASP was also adapted to solve single scenarios and used in a progressive hedging-based decomposition method [40].

POPULATION-BASED METAHEURISTICS

Evolutionary algorithms are perhaps the most widely employed class of metaheuristics for stochastic problems. A survey [12] on evolutionary optimization in uncertain

environments pinpointed many of the critical issues to consider when using population-based metaheuristics: techniques to handle noisy or approximate evaluation (fitness) functions, to search for robust solutions, and to tackle dynamic problems. A thorough survey of approximations in fitness evaluations can be found in Ref. 41.

Genetic Algorithm

Three different techniques to overcome noisy evaluations were discussed in Ref. 42. First, the obvious choice is to take the average value over several samples. Second, it is possible to increase the population size used by the GA. This can be based on models for calculating the probability of selecting the correct building blocks, which in turn depends on both the size of the population and the different sources of noise. Obviously, increasing the population and increasing the number of samples per solution both increases the overall computational effort, but there are no definite results as to how one can best balance the two techniques. The third option discussed is called inheritance of rescaled mutations, which was examined in the context of evolution strategies (ES): big mutations are made, the best change is selected based on evaluations of the big mutations, and the best mutation determines the direction of change, although only a smaller mutation is actually recorded for the next iteration. This alternative seems to be applicable only to continuous search spaces.

An example of applying ES to a continuous problem is in manufacturing, where the decision variables are the scheduled start times of each operation, the duration of each operation is stochastic, and each candidate solution is evaluated using simulations [43]. ES was compared to SA and found to converge slower, but to better solutions. The idea that determining the difference in rank is easier than determining the difference in value was then used to improve the speed of the ES [44], by first using a small sample size to remove the worst new individuals from further consideration before doing further selection based on a large sample size.

A method called *evolutionary policy iteration*, based on genetic algorithmss (GAs), has

been proposed for infinite horizon discounted reward MDPs [33]. Instead of the usual parent selection, subsets of arbitrary size are generated from the current population and are used to generate new solutions. Each subset results in one new solution based on a mechanism called policy switching. Two levels of mutations are used: a local mutation that results in small changes to the solution and a global mutation that gives larger changes. By also creating a special elite solution from each population, which by using the policy switching mechanism is guaranteed to be at least as good as any solution from the population, the evolutionary policy iteration method can be shown to converge to the optimal solution when the number of iterations goes to infinity.

An example of a distributed GA that operates in parallel on a network of workstations is given in Ref. 45. From a modeling perspective, the stochastic labor scheduling problem that is studied differs from the deterministic version simply by the number of labor requirement constraints per time interval, with one constraint for each possible realization of labor demand. The GA in question allows some infeasible solutions in the population by adding a penalty term to the fitness evaluation, but no experiments are shown to assess the effect of this choice with respect to the increased number of constraints arising due to the representation of the stochasticity.

A stochastic one-machine scheduling problem is studied in Ref. 46. The authors distinguish between quality robustness and solution robustness, with the former being a measure of robustness for the objective function when changes occur in the problem data, and the latter being robustness with respect to the amount of changes in the solution upon reoptimization after new information is revealed. It is argued that in some cases it is desirable to find solutions that most likely will remain mostly unchanged upon reoptimization. The problem is solved by a GA using a multiobjective fitness function, consisting of a weighted sum of the solution robustness and the quality robustness. Adding a local search operator to the GA appeared to overemphasize solution quality,

and even a slightly unfavorable stochastic outcome could necessitate important modifications of the resulting job-shop schedule. Therefore, the authors argue that local search should not be performed when dealing with the stochastic variant of the problem, although it remains unclear whether a suitably modified local search would be able to efficiently overcome this issue.

When GA is used in simulation optimization, there is a question whether one can decide *a priori* whether some simulations are unnecessary. Using domain knowledge seems to be a reasonable way to discard potential solutions before evaluating them with a costly simulation. In Ref. 47 a neural network was trained during the execution of the GA and after some initialization period, was used to filter out solutions with less potential.

The number of samples used to evaluate the fitness represents a trade-off between precision and time consumption. However, an increase in the precision of the fitness evaluations is perhaps not necessary for all solutions encountered, but only for the best ones [48]. One alternative is to gradually increase the number of samples used, and to reevaluate fitness when an increase is made. Thus, generating samples during the optimization can be better than fixing a set of scenarios before performing any optimization, especially when the number of potential scenarios is large [49]. Another suggestion is to use only one sample in fitness evaluations but to recalculate the fitness using a different sample every generation. The fitness calculations are then assumed to reflect the fluctuations created by the stochasticity [50], and the solution most frequently appearing in the population is assumed to be a good one, although Yoshimoto and Yamaguchi showed that a further examination of the set of the most frequent solutions can give improved results [51].

The use of multiple types of cross-over operators was examined in Ref. 52, and the resulting diversified cross-over operator was shown to be superior to the use of any single standard cross-over operator for the *heterogeneous PTSP*.

Scatter Search

Compared to GAs, the evolutionary meta-heuristic scatter search (SS) has received little attention. However, even when standard implementations are used as a black-box, the method can be successful: in the study by Yang *et al.* [53] it seemed better than a response surface method when considering two responses simultaneously. Using a standard implementation, Ref. 54 focused on variance reduction techniques, such as common random numbers and antithetic variates, and on dynamically adjusting the number of simulations performed to reduce the computational effort required to reach certain confidence interval widths for solution evaluations.

As an improvement method can often be time consuming for stochastic problems, Liu [55] suggested that the improvement method should be dropped unless the solution to be improved has a sufficiently high quality relative to the best solution in the reference set. Using approximate evaluations can help decrease the computational burden, and even when an improvement method based on local search is time consuming, it can be worthwhile to combine several different neighborhood operators [56].

Several alternatives to the standard SS implementation are proposed in Ref. 57. They consider problems where the objective function is given by a simulation process, although noisy evaluations are not considered, and where some constraints are given as differential equations describing system dynamics. Infeasible solutions are allowed but penalized based on the maximum percentage of violation. In a standard SS, building the initial reference set requires that all solutions in the original pool are evaluated using costly simulations. The suggested alternative is to build the initial reference set by only adding solutions that maximize diversity, after initially including three solutions where all variables are either at their lower bounds, their upper bounds, or at the midpoints of their bounds. This alternative saves computational effort at the expense of a reference set containing fewer high quality solutions. An interesting extension is made to the reference set update

mechanism: the standard update can lead to reference sets with very similar solutions. This could be particularly true for problems with a flat solution landscape, where one has many solutions with relatively similar objective function values which can often be the case for stochastic problems. The extension relies on a distance filter and an aspiration criterion: solutions are only added to the reference set if they can replace a worse solution and if they are sufficiently different from the rest of the reference set. They are, however, always added if they represent a new best solution found. Another change is that the reference set is not only rebuilt when no combined solution qualifies to enter, but also if the reference set becomes too homogeneous, either in terms of distances or objective function values.

Particle Swarm Optimization

Most work using particle swarm optimization (PSO) for stochastic problems seems to apply more or less standard implementations, for example Ref. 58, although the calculation of the fitness of particles may not be readily available in closed form [59]. An interesting example is given in Ref. 60, where an adaptive PSO is used both for generating a multistage scenario tree and then for solving the optimization problem based on the scenario tree. Although the use of PSO to generate scenario trees is not compared to other scenario tree generation methods, it does illustrate how metaheuristics can be used also when determining how the stochasticity can be represented, and not just to solve the problem once the stochasticity has been defined.

The ability of standard PSO heuristics to cope with noise was examined in Ref. 61. In PSO, noise affects every iteration when the solution of a particle is compared to the previous best solution of the particle, as well as when the overall best solution is determined. The authors conclude that noise can lead to stagnation, where better solutions are not accepted as such due to random noise, or to a search that is misguided by solutions that appear to be better than they are. Furthermore, the resulting stagnation can not be removed by parameter optimization alone,

but requires reduction of noise, for example through averaging over multiple samples.

CONCLUSIONS

In this article we have tried to illustrate a selection of issues that arise when using metaheuristics to solve stochastic optimization problems. Some metaheuristics have received attention in numerous publications, for example GAs, which have mostly been applied to problems with continuous variables [12,41]. Other metaheuristics have been tested on stochastic problems in only a few cases, or never at all.

A major concern when implementing metaheuristics for stochastic problems is that of efficiently evaluating solutions. This is sometimes accomplished through problem-specific analysis to provide quick, exact evaluations, sometimes by providing ad hoc approximations or proxies to generate evaluations, and sometimes by using filters to avoid evaluating unpromising solutions. Often the evaluation of solutions is handled by averaging over a set of samples. The sampling method is facilitated by recent developments in electronics: processor manufacturers pack more computing cores in each chip instead of reducing the sizes of the transistors as a primary means of increasing processing power. Parallelization of metaheuristics for stochastic problems therefore seems promising, in particular, when the evaluation of a given solution requires multiple simulations that can be performed on separate processor cores.

One lesson from research on heuristic search methods is that domain-specific knowledge is important [62], and that methods that easily allow the incorporation of such knowledge can have an advantage. There are two ways to include knowledge in a search: it can be specified *a priori* in the design of the search, for example where neighborhood structures are composed so that good moves will lead to promising parts of the search space. Alternatively, it can be formed *a posteriori* based on the search history, with metaheuristic components that analyze and learn from the attributes of

visited solutions. Most of the adaptations of metaheuristics for stochastic problems have focused on components that represent the former type of knowledge: finding efficient methods to evaluate solutions, showing how the use of multiple neighborhoods or cross-over operators can facilitate a better search, or tuning parameters such as population size in GAs to better handle problems with noise. Efficient inclusion of *a priori* knowledge is an important step toward successful metaheuristics for stochastic problems, but learning from search history may also prove beneficial: ACO achieves this to some extent through pheromone updates, and the GRASP in Ref. 38 includes learning as a central mechanism. Other examples include adaptively changing the number of samples used to evaluate solutions based on statistical tests, and the use of artificial intelligence techniques to estimate objective function values. Learning mechanisms that use search history to provide a balance between intensification and diversification [1,63] could be important: knowledge about the structure of stochastic problems, with solution quality depending on the interaction of many stochastic parameters, may be difficult to incorporate *a priori* in the search, but components using learning could potentially be used to find and exploit such structures.

A topic that deserves more attention is the coupling of exact and heuristic methods. For example, in Ref. 64 a GA was used as a component of an exact algorithm to solve finite-horizon partially observed MDPs. On the other hand, in Ref. 7, a GA used an exact algorithm as a subsolver, with dynamic programming being used to find the optimal policy for several MDPs to evaluate the solutions generated by the GA.

Compared to exact methods, metaheuristics have the advantage of being able to handle bigger problem instances and offer a significant reduction in computation times [11]. However, there are still open questions as to how one can best apply metaheuristics to stochastic problems. As for deterministic problems, the choice of solution method is dependent on characteristics of the problem,

but it is difficult to determine which method is most appropriate for any given problem.

Acknowledgments

The authors thank the anonymous referees for suggesting several improvements to the manuscript. This work was supported by the projects *DESIMAL* and *Coping with Risk in Maritime Logistics*, both partially funded by the Research Council of Norway.

REFERENCES

1. Blum C, Roli A. Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Comput Surv* 2003;35(3):268–308.
2. Glover F, Kochenberger GA, editors. Volume 57, *Handbook of metaheuristics*, International Series in Operations Research & Management Science. Boston (MA): Kluwer; 2003.
3. Reeves CR, editor. *Modern heuristic techniques for combinatorial problems*. Oxford: Blackwell Scientific Publishing; 1993.
4. Birge JR, Louveaux FV. *Introduction to stochastic programming*, Springer Series in Operations Research. New York: Springer; 1997.
5. Puterman ML. *Markov decision processes*. New York: Wiley; 1994.
6. Moreno Pérez JA, Melián Batista B, Laguna M. Scatter search based metaheuristic for robust optimization of the deploying of “dwdm” technology on optimal networks with survivability. *Yugosl J Oper Res* 2005;15(1): 65–77.
7. Yokoyama M, Lewis HW III. Optimization of the stochastic dynamic production problem by a genetic algorithm. *Comput Oper Res* 2003;30:1831–1849.
8. Fu MC. Optimization for simulation: theory vs. practice. *INFORMS J Comput* 2002;14: 192–215.
9. Sörensen K. Investigation of practical, robust and flexible decisions for facility location problems using tabu search and simulation. *J Oper Res Soc* 2008;59:624–636.
10. Fleury G, Gourgand M, Lacomme P. Metaheuristics for the stochastic hoist scheduling problem (SHSP). *Int J Prod Res* 2001;39(15): 3419–3457.

11. Bianchi L, Dorigo M, Gambardella LM, *et al.* A survey on metaheuristics for stochastic combinatorial optimization. *Nat Comput.* 2009;8(2):239–287.
12. Jin Y, Branke J. Evolutionary optimization in uncertain environments—a survey. *IEEE Trans Evol Comput* 2005;9(3):303–317.
13. Dyer M, Stougie L. Computational complexity of stochastic programming problems. *Math Program Ser A* 2006;106(3):423–432.
14. Beraldi P, Ghiani G, Laporte G, *et al.* Efficient neighborhood search for the probabilistic pickup and delivery travelling salesman problem. *Networks* 2005;45(4):195–198.
15. Beraldi P, Ghiani G, Musmanno R, *et al.* Efficient neighborhood search for the probabilistic multi-vehicle pickup and delivery problem. *Asia Pac J Oper Res.* In press.
16. Bianchi L, Campbell AM. Extension of the 2-p-opt and 1-shift algorithms to the heterogeneous probabilistic traveling salesman problem. *Eur J Oper Res* 2007;176:131–144.
17. Bianchi L, Birattari M, Chiarandini M, *et al.* Metaheuristics for the vehicle routing problem with stochastic demands. In: Yao X, Burke E, Lozano JA, *et al.*, editors. Volume 3242, Proceedings of the 8th International Conference on Parallel Problem Solving from Nature, Lecture Notes in Computer Science. Berlin: Springer; 2004. pp. 450–460.
18. Birattari M, Balaprakash P, Stützle T, *et al.* Estimation-based local search for stochastic combinatorial optimization using delta evaluations: a case study in the probabilistic traveling salesman problem. *INFORMS J Comput* 2008;20(4):644–658.
19. Gendreau M, Laporte G, Séguin R. A tabu search heuristic for the vehicle routing problem with stochastic demands and customers. *Oper Res* 1996;44(3):469–477.
20. Dengiz B, Alabas C. Simulation optimization using tabu search. In: Joines JA, Barton RR, Kang K, *et al.*, editors. Proceedings of the 2000 winter simulation conference. Piscataway (NJ): IEEE Press; 2000. pp. 805–810.
21. Lutz CM, Davis KR, Sun M. Determining buffer location and size in production lines using tabu search. *Eur J Oper Res* 1998;106:301–316.
22. Glover F. Infeasible/feasible search trajectories and directional rounding in integer programming. *J Heuristics* 2007;13(6):505–542.
23. Aringhieri R. Solving chance-constrained programs combining tabu search and simulation. In: Ribeiro CC, Martins SL, editors. Volume 3059, Proceedings of the 3rd International Workshop on Experimental and Efficient Algorithms, Lecture Notes in Computer Science. Berlin: Springer; 2004. pp. 30–41.
24. Alkhamis TM, Ahmed MA, Tuan VK. Simulated annealing for discrete optimization with estimation. *Eur J Oper Res* 1999;116:530–544.
25. Alfarei MH, Andradóttir S. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Manage Sci* 1999;45(5):748–764.
26. Alkhamis TM, Ahmed MA. Simulation-based optimization using simulated annealing with confidence intervals. In: Ingalls RG, Rosetti MD, Smith JS, *et al.*, editors. Proceedings of the 2004 Winter Simulation Conference. Piscataway (NJ): IEEE Press; 2004. pp. 514–518.
27. Gutjahr WJ, Pflug G. Simulated annealing for noisy cost functions. *J Glob Optimiz* 1996;8:1–13.
28. Bowler NE, Fink TMA, Ball RC. Characterization of the probabilistic traveling salesman problem. *Phys Rev E* 2003;68(3):036703.
29. Gutjahr WJ. Two metaheuristics for multiobjective stochastic combinatorial optimization. Volume 3777, Stochastic algorithms: foundations and applications, Lecture Notes in Computer Science. Berlin: Springer; 2005. pp. 116–125.
30. Balaprakash P, Birattari M, Stützle T, *et al.* Adaptive sample size and importance sampling in the estimation-based local search for the probabilistic traveling salesman problem. *Eur J Oper Res.* 16 November 2009;199(1):98–110.
31. Gutjahr WJ, Katzensteiner S, Reiter P. A VNS algorithm for noisy problems and its application to project portfolio analysis. In: Hromkovic J, Kralovic R, Nunkesser M, *et al.*, editors. Volume 4665, Stochastic algorithms: foundations and applications, Lecture Notes in Computer Science. Heidelberg: Springer; 2007. pp. 93–104.
32. Chang HS, Gutjahr WJ, Yang J, *et al.* An ant system approach to Markov decision processes. Volume 4, Proceedings of the 23rd American Control Conference. Piscataway (NJ): IEEE Press; 2004. pp. 3820–3825.
33. Chang HS, Lee H-G, Fu MC, *et al.* Evolutionary policy iteration for solving markov decision processes. *IEEE Trans Autom Control* 2005;50(11):1804–1808.
34. Chang HS. An ant system based exploration-exploitation for reinforcement learning.

- Proceedings of the IEEE Conference on Systems, Man, and Cybernetics. Piscataway (NJ): IEEE Press; 2004. pp. 3805–3810.
35. Balaprakash P, Birattari M, Stützle T, *et al.* Ant colony optimization and estimation-based local search for the probabilistic traveling salesman problem. Technical Report TR/IRDIA/2008-020. Bruxelles, Belgium: Université Libre de Bruxelles; 2008.
 36. Gutjahr WJ. S-ACO: an ant-based approach to combinatorial optimization under uncertainty. In: Dorigo M, Birattari M, Blum C, *et al.*, editors. Volume 3172, Ant colony optimization and swarm intelligence, 4th international workshop, ANTS 2004, Lecture Notes in Computer Science. Berlin: Springer; 2004. pp. 238–249.
 37. Birattari M, Balaprakash P, Dorigo M. The ACO/F-RACE algorithm for combinatorial optimization under uncertainty. In: Doerner KF, Gendreau M, Greistorfer P, *et al.*, editors. Volume 39, Metaheuristics: progress in complex systems optimization, Operations Research/Computer Science Interfaces Series. Berlin (NY): Springer; 2007. pp. 189–203.
 38. Hvattum LM, Løkketangen A, Laporte G. Scenario tree based heuristics for stochastic inventory routing problems. *INFORMS J Comput* 2009;21:268–285.
 39. Glover F. Multi-start and strategic oscillation methods –principles to exploit adaptive memory. In: Laguna M, González-Velarde JL, editors. OR computing tools for modeling, optimization and simulation: interfaces in computer science and operations research. Boston (MA): Kluwer; 2000. pp. 1–24.
 40. Hvattum LM, Løkketangen A. Using scenario trees and progressive hedging for stochastic inventory routing problems. *J Heuristics*. 2009;15:527–557.
 41. Jin Y. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Comput* 2005;9(1):3–12.
 42. Beyer H-G. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. *Comput Meth Appl Mech Eng* 2000;186:239–267.
 43. Earl CF, Hicks C, Song D-P. Planning complex engineer-to-order products. In: Gogu G, Coutellier D, Chedmail P, *et al.*, editors. Recent advances in integrated design and manufacturing in mechanical engineering. Dordrecht: Kluwer; 2003. pp. 463–472.
 44. Song D-P, Hicks C, Earl CF. An ordinal optimization based evolution strategy to schedule complex make-to-order products. *Int J Prod Res* 2006;44(22):4877–4895.
 45. Easton FF, Mansour N. A distributed genetic algorithm for deterministic and stochastic labor scheduling problems. *Eur J Oper Res* 1999;118(3):505–523.
 46. Sevaux M, Sörensen K. A genetic algorithm for robust schedules in a one-machine environment with ready times and due dates. *4OR - Q J Belg Fr Ital Oper Res Soc* 2004; 2(2):129–148.
 47. Zeng Q, Yang Z. Integrating simulation and optimization to schedule loading operations in container terminals. *Comput Oper Res* 2009;36:1935–1944.
 48. Stagge P. Averaging efficiently in the presence of noise. In: Eiben AE, Bäck T, Schoenauer M, *et al.*, editors. Volume 1498, Proceedings of the 5th International Conference on Parallel Problem Solving from Nature, Lecture Notes in Computer Science. Berlin: Springer; 1998. pp. 188–200.
 49. Wang K-J, Wang S-M, Chen J-C. A resource portfolio planning model using sampling-based stochastic programming and genetic algorithm. *Eur J Oper Res* 2008;184: 327–340.
 50. Yoshimoto Y. A genetic algorithm approach to solving stochastic job-shop scheduling problems. *Int Trans Oper Res* 2002;9:479–495.
 51. Yoshimoto Y, Yamaguchi R. A genetic algorithm and the Monte Carlo method for stochastic job-shop scheduling. *Int Trans Oper Res* 2003;10:577–596.
 52. Liu Y-H, Jou R-C, Wang C-C, *et al.* An evolutionary algorithm with diversified crossover operator for the heterogeneous probabilistic TSP. In: Carbonell JG, Siekmann J, editors. Volume 4617, Modeling decisions for artificial intelligence, 4th international conference, Lecture Notes in Computer Science. Berlin: Springer; 2007. pp. 351–360.
 53. Yang T, Kuo Y, Chou P. Solving a multiresponse simulation problem using a dual-response system and scatter search method. *Simul Model Pract Theor* 2005;13:356–369.
 54. Anderson NP, Evans GW, Biles WE. Simulation optimization of logistics systems through the use of variance reduction techniques and criterion models. *Eng Optimiz* 2006;38(4):441–460.
 55. Liu Y-H. A hybrid scatter search for the probabilistic traveling salesman problem. *Comput Oper Res* 2007;34:2949–2963.

56. Liu Y-H. Diversified local search strategy under scatter search framework for the probabilistic traveling salesman problem. *Eur J Oper Res* 2008;191:332–346.
57. Egea J, Rodríguez M, Banga J, *et al.* Scatter search for chemical and bio-process optimization. *J Glob Optimiz* 2007;37(3):481–503.
58. Wang L, Singh C. Stochastic economic emission load dispatch through a modified particle swarm optimization algorithm. *Elec Power Syst Res* 2008;78(8):1466–1476.
59. Brodersen O, Schumann M. Optimizing a stochastic warehouse using particle swarm optimization. *Proceedings of the 2nd International Conference on Innovative Computing, Information and Control*. Piscataway (NJ): IEEE Press; 2007. pp. 449–452.
60. Pappala VS, Erlich I. Management of distributed generation units under stochastic load demands using particle swarm optimization. *Power engineering society general meeting*. Piscataway (NJ): IEEE Press; 2007. pp. 1–7.
61. Bartz-Beielstein T, Blum D, Branke J. Particle swarm optimization and sequential sampling in noisy environments. In: Doerner KF, Gendreau M, Greistorfer P, *et al.*, editors. Volume 39, *Metaheuristics: progress in complex systems optimization*, Operations Research/Computer Science Interfaces Series. Berlin (NY): Springer; 2007. pp. 261–273.
62. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997;1:67–82.
63. Andradóttir S, Prudius AA. Balanced explorative and exploitative search with estimation for simulation optimization. *INFORMS J Comput* 2009;21(2):193–208.
64. Lin AZ-Z, Bean JC, White CC III. A hybrid genetic/optimization algorithm for finite-horizon, partially observed Markov decision processes. *INFORMS J Comput* 2004;16(1): 27–38.

METHODS FOR LARGE-SCALE UNCONSTRAINED OPTIMIZATION

GIOVANNI FASANO
Dipartimento di Matematica
Applicata, Università Ca' Foscari,
Venezia, Italy

The general large-scale unconstrained optimization problem can be formulated as follows:

$$\min_{x \in \mathbb{R}^n} f(x), \quad (1)$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is a real-valued objective function, and the parameter n is *large*. Observe that in general the minimization in problem (1) refers to the search for a *local minimizer* of $f(x)$. Standard assumptions ensure that the problem (1) admits solution. In particular, if the level set $\mathcal{L}_0 = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is compact and $f(x)$ is continuous on $\mathcal{L}_0$, then by the Weierstrass theorem (1) has a solution. Moreover, if the objective function is coercive, that is, $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ (which usually holds in real applications), then all the level sets of $f(x)$ are compact.

The definition of “large n ” is a little vague for the problem (1), in that it does not specify exactly a range of values and is straightforwardly machine dependent. Indeed, broadly speaking, n may be considered large on a particular machine, if standard optimization techniques for problem (1) become progressively inadequate when n increases, so that *ad hoc* methods must be adopted. Anyway, 10^6 – 10^7 may be considered a reference value for n , on a serial machine, in order to consider problem (1) a large-scale continuous optimization problem. Real-life applications increasingly yield complex models that require efficient techniques (see the Nonlinear Optimization Models collected by Robert Vanderbei on <http://www.orfe.princeton.edu/~rvdb/amp1/nlmodels>). This has led in the last

decades to a growing interest for large-scale optimization, so that the latter topic can be considered by this time a mature field of research.

Optimization methods for small- and medium-scale problems [1–4] are usually hardly adaptable to large-scale problems. Indeed, though they may be still effective in most of the cases, on problems where n is large they often prove not to be efficient enough (i.e., they are unable to solve the problems using a reasonable workload).

As for most of the algorithms for continuous unconstrained optimization, methods for large values of n claim for suitable search directions and stepsizes along them, too. In particular, the choice of the search directions is responsible for the *efficiency* of the methods (i.e., the rate of convergence), while a proper choice of the steplength ensures the *effectiveness*. When n is large, iterative methods typically show a reduced computational burden with respect to direct methods. They generate a sequence of iterates $\{x_k\}$ approaching a solution, which usually follows one of the patterns:

1. $x_{k+1} = x_k + \alpha_k d_k$ in case convergence to a stationary point x^* is sought, which satisfies the first-order necessary optimality conditions;
2. $x_{k+1} = x_k + \phi_d(\alpha_k) d_k + \phi_u(\alpha_k) u_k$ in case convergence to a stationary point x^* is sought, which satisfies the second-order necessary optimality conditions;

where $d_k \in \mathbb{R}^n$ is a search direction that approximates a Newton-type direction at x_k , $u_k \in \mathbb{R}^n$ is a search direction that takes into account the local curvature of the objective function at x_k , $\phi_d(\cdot)$ and $\phi_u(\cdot)$ are polynomial functions of degree at most two, and α_k is a suitable stepsize. The schemes in the classes 1. and 2. may typically be differentiated on the basis of the information on $f(x)$, which is available at the iterate x_k . In particular, if at x_k we can simply use $f(x_k)$, $\nabla f(x_k)$, and a

partial information on $\nabla^2 f(x_k)$, then we will be able to generate $\{x_k\}$ according to 1.

On the other hand, if in addition we get information on the eigenpair associated with the smallest eigenvalue of the Hessian matrix $\nabla^2 f(x_k)$, we can generate the sequence $\{x_k\}$ according to 2. Observe that the last statement does not imply that a full knowledge (and possibly the storage) of $\nabla^2 f(x_k)$ is required (which is commonly considered cumbersome). When n increases, though in principle the entries of the Hessian matrix could be computed by finite differences, iterative methods commonly need to perform only the product *Hessian* $\times$ *vector*, which is computationally cheaper. Alternatively, in case $f(x)$ is explicitly defined and is not a *black box*, automatic differentiation [5] is often adopted to provide information on $\nabla^2 f(x_k)$.

GLOBAL CONVERGENCE AND GRADIENT METHODS

A very general convergence result may be proved for the iterative scheme 1. Indeed, introducing the *forcing function* $\sigma : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, such that $\lim_{k \rightarrow \infty} \sigma(w_k) = 0$ implies $\lim_{k \rightarrow \infty} w_k = 0$, we have [6]:

Proposition 1. *Let $\{x_k\}$ be the sequence generated by the scheme 1. starting from $x_0 \in \mathbb{R}^n$. Suppose the level set $\mathcal{L}_0$ is compact and for any k we have $d_k \neq 0$ if $\nabla f(x_k) \neq 0$, with $f(x_{k+1}) \leq f(x_k)$. Assume that*

1. *we have*

$$\lim_{k \rightarrow \infty} \frac{\nabla f(x_k)^T d_k}{\|d_k\|} = 0,$$

2. *for any k for which $d_k \neq 0$*

$$\frac{|\nabla f(x_k)^T d_k|}{\|d_k\|} \geq \sigma(\|\nabla f(x_k)\|).$$

Then, either there exists a finite index $\hat{k}$ such that $\nabla f(x_{\hat{k}}) = 0$, or the infinite sequence $\{x_k\}$ satisfies

- $x_k \in \mathcal{L}_0$ for any k ;

- the sequence $\{f(x_k)\}$ converges;
- for any infinite subsequence $\mathcal{K}$ of indices

$$\lim_{k \rightarrow \infty, k \in \mathcal{K}} \|\nabla f(x_k)\| = 0.$$

A very simple class of iterative methods for large-scale unconstrained optimization, which satisfy the hypotheses of Proposition 1, is given by the so-called *gradient methods* [4,7], which may be specified in the form

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \quad (2)$$

with $D_k \in \mathbb{R}^{n \times n}$ positive definite for any k . Observe that whenever $\nabla f(x_k) \neq 0$ the search direction $d_k = -D_k \nabla f(x_k)$ in iteration (2) is of descent at the nonstationary point x_k , and satisfies statement 2. of Proposition 1. Moreover, an Armijo-type linesearch procedure [8] should be adopted to compute α_k in iteration (2) in order to satisfy also statement 1. of Proposition 1. Of course, when $D_k = I$, for any k , then the iteration (2) reduces to the standard *steepest descent method*, which is equivalent to minimize the first-order local model of the objective function. The steepest descent method yields a linear convergence rate, and it is not scale invariant under a coordinate transformation. In case the evaluation of $f(x)$ is expensive (e.g., for several simulation problems [9]), a constant value for the stepsize α_k may be adopted in place of the linesearch procedure. On the other hand, a computationally efficient choice for α_k and D_k was suggested in Barzilai and Borwein [10] and Raydan [11], where $D_k = I$, for any k , and

$$\alpha_k = \frac{\|x_k - x_{k-1}\|^2}{(x_k - x_{k-1})^T [\nabla f(x_k) - \nabla f(x_{k-1})]}.$$

Finally, when $D_k = [\nabla^2 f(x_k)]^{-1}$ and $\alpha_k = 1$ the iteration (2) reduces to *Newton's method*, which corresponds to minimize the quadratic local model of $f(x)$ at x_k . The latter technique is appealing thanks to its quadratic rate of convergence and the scale invariance. However, when $\nabla^2 f(x_k)$ is dense and n is large, it may not be a suitable choice. In the latter case, fully computing $[\nabla^2 f(x_k)]^{-1}$ by means of a Cholesky factorization or a symmetric indefinite factorization requires nearly $1/3n^3$

floating-point operations, while for sparse Hessians the workload is proportional to the sparsity pattern.

INEXACT AND TRUNCATED NEWTON METHODS

The application of Newton's method requires one to compute the direction $d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$, which is equivalent to solving *Newton's equation*

$$\nabla^2 f(x_k) d = -\nabla f(x_k). \quad (3)$$

When the iterate x_k is still "far" from the stationary point x^* , an accurate solution of Equation (3) may be unjustified; thus, when n is large the solution of Equation (3) may be obtained by an iterative procedure. In addition, in order to preserve the global convergence of Newton's method, a suitable choice of α_k must be computed (a globalization technique), either based on a *linesearch approach* [1,12,13] or a *trust-region approach* [14,15]. On these guidelines, for large values of n , *inexact Newton method* have been proposed

Linesearch-based Truncated Newton scheme

```

Set  $x_0 \in \mathbb{R}^n$ 
Set  $\eta_k \in [0, 1)$  for any  $k$ , with  $\{\eta_k\} \rightarrow 0$ 
OUTER ITERATIONS
for  $k = 0, 1, \dots$ 
    Compute  $\nabla f(x_k)$ ; if  $\|\nabla f(x_k)\|$  is small then STOP
    INNER ITERATIONS
    Compute  $d_k$  which approximately solves Equation (3)
    and satisfies the truncation rule
         $\|\nabla^2 f(x_k) d_k + \nabla f(x_k)\| \leq \eta_k \|\nabla f(x_k)\|$ 
    Compute the steplength  $\alpha_k$  by an Armijo-type linesearch scheme
    Update  $x_{k+1} = x_k + \alpha_k d_k$ 
endfor
```

Observe that at the outer iteration k a second-order expansion of the function $f(x)$ is considered, whose stationary point (if any) is detected by solving Equation (3). The number of inner iterations, which are necessary to approximately solve Newton's equation (3), depends on the current parameter η_k . Thus, a fine (and more expensive) solution of Equation (3) is required only when $k \rightarrow \infty$. The truncation rule (4) is often replaced in practice by more efficient conditions [22,23].

in the literature [16]. The underpinning idea of these schemes is that we can balance the computational burden of solving Equation (3) and the accuracy of the solution obtained. In particular, from Dembo *et al.* [16], if the approximate solution d_k of Equation (3) is computed, such that (truncation rule)

$$\lim_{k \rightarrow \infty} \frac{\|\nabla^2 f(x_k) d_k + \nabla f(x_k)\|}{\|\nabla f(x_k)\|} = 0 \quad (4)$$

holds, then the sequence of iterates $\{x_k\}$ is *superlinearly convergent* to a stationary point x^* . Broadly speaking, the condition (4) states that when k increases, the gradient of the quadratic local model of $f(x)$ at x_k approaches zero "more quickly" than the gradient $\nabla f(x_k)$ of the objective function. When n is large, iterative methods such as the *conjugate gradient* (CG) may be used to compute an approximate solution d_k of Equation (3), such that Equation (4) holds. The latter philosophy is based on *truncated Newton methods* [17–19], which proved to be a very efficient tool [20,21]. An example of truncated Newton method, based on a linesearch approach to ensure the global convergence, is the following:

The Hessian matrix in Equation (3) may be possibly indefinite. Thus, additional safeguard is required for the choice of the iterative solver, since for instance a negative curvature direction for function $f(x)$ at x_k may be detected, or a pivot breakdown may occur with the CG method. The Lanczos process and suitable extensions of the CG method (namely, the planar CG) may successfully be used as alternatives to the CG [24–28]. Furthermore, finite differences

have also been used in the past, in order to approximately compute the information on the Hessian matrix, within a truncated Newton method (see TNPack in Schlick and Fogelson [29]).

Three other relevant issues have been studied in the last two decades, in order to drastically improve the performance of truncated Newton methods: the *preconditioning strategies* to solve Newton's equation, the exploitation of *negative curvature directions* of the objective function, and the introduction of *nonmonotone globalization* schemes. The first issue studies efficient preconditioners of the matrix $\nabla^2 f(x_k)$, for the specific case of large n , when the Hessian is either positive definite or indefinite [14, 30–32]. The second issue needs a specific treatment when n is large, since it deals with the careful choice for the iterative solver of Newton's equation. Moreover, an accurate exploration of the negative curvatures of $f(x)$ is essential in order to prove the convergence of truncated methods toward solutions that satisfy the second-order optimality conditions [33, 34]. The latter task is pursued by computing the pair of promising search directions (d_k, u_k) at the outer iteration k [20, 25, 35–37], by means of efficient iterative techniques. Roughly speaking, d_k contains information on the local convexity of $f(x)$ at x_k , while u_k conveys information on the local nonconvexity of $f(x)$, and approximates the eigenvector corresponding to the minimum negative eigenvalue of $\nabla^2 f(x_k)$. Then, in order to guarantee the convergence to second-order points, the two directions may be combined to produce the next iterate, as in (see pattern 2 mentioned in the introductory text.)

$$x_{k+1} = x_k + \phi_d(\alpha_k)d_k + \phi_u(\alpha_k)u_k,$$

where α_k is chosen by a proper linesearch procedure. The introduction of nonmonotone globalization schemes, in truncated Newton methods, was first proposed in Grippo *et al.* [38] (see also Grippo *et al.* [13], Lucidi *et al.* [25] and Ferris *et al.* [35]). Considering the standard Armijo linesearch scheme $f(x_k + \alpha d_k) \leq f(x_k) + \gamma \alpha \nabla f(x_k)^T d_k$, where $\gamma \in (0, 1/2)$, the main idea behind a nonmonotone approach generalizes the mono-

tone decreasing of $f(x_k)$, as in the scheme

$$f(x_k + \alpha d_k) \leq \max_{0 \leq i \leq M} \{f(x_{k-i})\} + \gamma \alpha \nabla f(x_k)^T d_k,$$

where the integer M indicates the “memory” of the scheme. The nonmonotone approach proves to be particularly efficient in the case of highly nonlinear functions, where enforcing the monotonic decrease of $f(x)$ may cause the algorithms to be trapped within narrow valleys.

Global convergence of truncated Newton methods may be proved either using a linesearch framework, or adopting a *trust-region* approach [14]. A trust-region scheme provides the new iterate $x_{k+1} = x_k + d_k$ at step k (i.e., now d_k includes both the search direction and the steplength), by solving the constrained subproblem

$$\begin{aligned} \min_d \quad & Q_k(d) \\ \text{s.t.} \quad & \|\Lambda_k d\| \leq \Delta_k, \end{aligned} \quad (5)$$

where $Q_k(d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$ and Λ_k is a scaling matrix (possibly $\Lambda_k = I$, for any k), which may be regarded as introducing an implicit preconditioning. The basic trust-region approach adaptively assesses the trust-region radius Δ_k in problem (5) with the following idea: as long as the solution d_k of problem (5) yields a value for ρ_k close to 1, where

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{Q_k(0) - Q_k(d_k)}, \quad (6)$$

$Q_k(d)$ is a “trusted” local model of $f(x)$ and the radius Δ_{k+1} may be possibly larger than Δ_k . On the contrary, if ρ_k in relation (6) is relatively small, then the model $Q_k(d)$ is not reliable and the radius Δ_{k+1} at the next iteration satisfies $\Delta_{k+1} < \Delta_k$.

When n is large, the trust-region approach requires both a strategy to assess the parameter Δ_k and an efficient procedure to approximately solve problem (5). Among the first techniques to solve problem (5), the *dogleg* schemes proposed to investigate the solution over a suitable one-dimensional arc [39, 40].

In Byrd *et al.* [41], a pair of feasible directions for problem (5) is preliminarily

computed, then the minimization of $Q_k(d)$ is performed over the two-dimensional manifold associated with the latter vectors. Another approach to solve problem (5) imposes the KKT conditions of problem (5), and solves a resulting perturbation of the Newton equation for $f(x)$ [42,43].

Since the previous approaches may recur to direct solvers (e.g., Cholesky factorization), when n is large, a different strategy is suggested in Steihaug [44] and Toint [45]. Here, CG is used to minimize $f(x)$; then, a piecewise path connecting the feasible iterates generated by CG is considered, and can be followed up to the boundary of the trust region. Alternatively, both the steepest descent direction (moving to the Cauchy point¹ on the boundary of the trust region) and possibly a negative curvature direction (in case $\nabla^2 f(x_k)$ is indefinite) are investigated. An example of a robust and reliable truncated Newton method, in a trust region rather than a linesearch framework, is given by the package LANCELOT [21]. Trust-region methods have several advantages, including the fact that they show strong convergence properties [14]. However, in case for several values of k the Hessian matrix $\nabla^2 f(x_k)$ is indefinite, these methods may be inefficient, inasmuch as they can stop the inner iterations prematurely, when approximately minimizing $Q_k(d)$. These drawbacks have been addressed and partially overcome in Gould *et al.* [26], by using the Lanczos process for the solution of large symmetric and indefinite linear systems, and exploiting its relation with the CG (see also Stoer [46]).

NONLINEAR CONJUGATE GRADIENT METHODS

The CG used for the minimization of a strictly convex quadratic function has been extended to the minimization of nonlinear functions, by suitably modifying the computation of its

coefficients [47–49]. In particular, suppose that at step k the directions $\{p_1, \dots, p_k\}$ were generated by the Nonlinear Conjugate Gradient (NCG). Then, the steplength α_k is chosen to satisfy the standard first Wolfe linesearch condition (minimization along the direction p_k)

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \gamma \alpha_k \nabla f(x_k)^T p_k, \quad \gamma \in (0, 1/2) \quad (7)$$

and either one of the following additional relations (Wolfe conditions)

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq \theta \nabla f(x_k)^T p_k, \quad \theta \in (\gamma, 1) \quad (8)$$

$$|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq \theta |\nabla f(x_k)^T p_k|, \quad \theta \in (\gamma, 1) \quad (9)$$

(theoretical reasons may impose the more restrictive condition $\theta \in (\gamma, 1/2)$). Furthermore, in the NCG, it is possible to adopt different formulae for the computation of the coefficient β_k used in the iteration $p_k = -\nabla f(x_k) + \beta_k p_{k-1}$. Very standard formulae for β_k in the literature are [48]

Fletcher–Reeves:

$$\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2},$$

Polak–Ribiere:

$$\frac{[\nabla f(x_k) - \nabla f(x_{k-1})]^T \nabla f(x_k)}{\|\nabla f(x_{k-1})\|^2},$$

Hestenes–Stiefel:

$$\frac{[\nabla f(x_k) - \nabla f(x_{k-1})]^T \nabla f(x_k)}{[\nabla f(x_k) - \nabla f(x_{k-1})]^T p_{k-1}}. \quad (10)$$

We report below a general scheme for the NCG. We remark that, apart from relations (10), several other choices for the coefficient β_k have been studied in the literature [48].

¹The Cauchy point is the minimizer of $Q_k(d)$ along the direction $-\nabla f(x_k)$, subject to the trust-region bound.

Nonlinear Conjugate Gradient (NCG) method

Step 0: Choose $x_0 \in \mathbb{R}^n$, set $k = 0$

Step 1: Compute $\nabla f(x_k)$. If $\nabla f(x_k) = 0$ then STOP

Step 2: If $k \neq 0$ compute β_k as in relations (10). Compute the direction

$$p_k = \begin{cases} -\nabla f(x_k) & k = 0 \\ -\nabla f(x_k) + \beta_k p_{k-1} & k \geq 1 \end{cases}$$

Step 3: Choose α_k such that relation (7) is satisfied, along with either relation (8) or (9)

Step 4: Set $x_{k+1} = x_k + \alpha_k p_k$, $k = k + 1$ and go to **Step 1**

Unlike the CG, the NCG does not generally retain global convergence properties, because of the loss of conjugacy caused by the nonlinearity of $f(x)$. In order to enforce global convergence, a periodic restart may be imposed, either when $k > n$ or if a test on the conjugacy loss, like

$$|\nabla f(x_k)^T \nabla f(x_{k-1})| > \sigma \|\nabla f(x_{k-1})\|^2, \\ \sigma \in (0, 1)$$

is satisfied. However, when n is large, the latter choice is impracticable (very rarely adopted) and proved to be inefficient. Using the formula of β_k , proposed by Fletcher–Reeves, in case α_k is computed by an exact linesearch procedure [50], or by an Armijo-type linesearch based on the strong Wolfe condition (9) with $\theta < 1/2$ [2,51], the global converge is retained and $\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$. In Gilbert and Nocedal [49], the global convergence was also proved for the case $|\beta_k| \leq \beta_k^{FR}$, where β_k^{FR} refers to the formula in relations (10) proposed by Fletcher and Reeves.

Though the choice of β_k^{PR} made by Polak–Ribiere (see relations (10)) proved to be computationally more efficient than β_k^{FR} [2,48], convergence properties for the NCG with β_k^{PR} have always been more difficult to investigate. In particular, the global convergence was proved under strong convexity assumption of $f(x)$, using $\beta_k = \max\{0, \beta_k^{PR}\}$, with an inexact linesearch procedure which also satisfies condition (9). Only recently [52],

by using a more sophisticated linesearch procedure, the global convergence of NCG using β_k^{PR} was proved (in the sense that $\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$), without the strong convexity assumption on $f(x)$. An extension of the NCG, which uses the Polak–Ribiere value β_k^{PR} , is implemented in the routine VA14, of the Harwell Subroutine Library [53].

QUASI-NEWTON METHODS AND PARTIALLY SEPARABLE FUNCTIONS

Quasi-Newton methods are a powerful tool for large-scale optimization. They are used to solve Newton's equation, without using the second-order derivatives. In particular, these methods generate the sequence $\{x_k\}$ by iteratively solving the modified Newton's equation

$$B_k d = -\nabla f(x_k),$$

with B_k positive definite

and computing $x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k)$, where B_k in some sense *approximates* the Hessian matrix $\nabla^2 f(x_k)$ of $f(x)$. More specifically, B_{k+1} is the solution of the subproblem

$$\begin{aligned} \min_B \quad & \|B - B_k\|_F \\ \text{s.t.} \quad & B = B^T \\ & B s_k = y_k, \end{aligned}$$

where $\|\cdot\|_F$ indicates the Frobenius norm, $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$,

so that no second-order derivatives are required. Several quasi-Newton schemes have been developed, using approximations of either $\nabla^2 f(x_k)$ or $[\nabla^2 f(x_k)]^{-1}$ [2,12,48,54,55]. One of the most effective, in small-scale and medium-scale optimization, is BFGS (*Broyden–Fletcher–Goldfarb–Shanno*). Here, a sequence $\{H_k\}$ of approximations of $[\nabla^2 f(x_k)]^{-1}$ is generated, with H_k being positive definite, by solving the subproblem

$$\begin{aligned} \min_H \quad & \|H - H_k\|_F \\ \text{s.t.} \quad & H = H^T \\ & s_k = Hy_k, \end{aligned}$$

so that $x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k)$. Moreover, the global convergence properties can be proved by computing the steplength α_k via an Armijo-type linesearch procedure, satisfying also condition (9). It can be easily proved [2] that the matrices $\{H_k\}$ satisfy the relation

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T,$$

where $V_k = I - \rho_k y_k s_k^T$ and $\rho_k = 1/y_k^T s_k$, so that the pairs $\{(s_k, y_k)\}$ are necessary and sufficient to generate the whole sequence $\{H_k\}$. We remark that an interesting relationship between NCG and BFGS can be stated (see Nocedal and Wright [2] or Pytlak [48]), both for quadratic and general nonlinear functions.

In a large-scale setting, the storage of V_k cannot be proposed; Thus, the formula to update H_{k+1} is suitably modified by using just the most m recent pairs (s_i, y_i) , $i = 1, \dots, m$. On this guideline, the matrix H_k is computed as

$$\begin{aligned} H_k = & (V_{k-1}^T \dots V_{k-m}^T) H_k^0 (V_{k-m} \dots V_{k-1}) \\ & + \rho_{k-m} (V_{k-1}^T \dots V_{k-m+1}^T) \\ & \quad s_{k-m} s_{k-m}^T (V_{k-m+1} \dots V_{k-1}) \\ & + \rho_{k-m+1} (V_{k-1}^T \dots V_{k-m+2}^T) \\ & \quad s_{k-m+1} s_{k-m+1}^T (V_{k-m+2} \dots V_{k-1}) \\ & + \dots \\ & + \rho_{k-1} s_{k-1} s_{k-1}^T, \end{aligned}$$

and a practical choice of H_k^0 is often $H_k^0 = \gamma_k I$, with $\gamma_k = s_{k-1}^T y_{k-1} / y_{k-1}^T y_{k-1}$. The storage of

H_k now requires only the pairs (s_i, y_i) , $i \leq m$, so that it can be defined by setting the parameter m . The resulting new sequence $\{H_k\}$ yields a very efficient method called *L-BFGS* [56–58] (where “L” stands for *limited memory*).

Observe that in practice very small values of m are used (say $3 \leq m \leq 8$) and the L-BFGS method often may be competitive with truncated Newton methods or the NCG [59]. In addition, thanks to the simplified structure of H_k , iterative Krylov-based methods may be easily applied when performing the product $H_k \nabla f(x_k)$ in the iteration $x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k)$. In order to cope with problems of illconditioning of the Hessian matrix, a combined approach between L-BFGS and the discrete Newton method was studied in Byrd *et al.* [60]. The L-BFGS method was coded and is available in the routine `VA15`, of the Harwell Subroutine Library [53]. The memory storage of the NCG implemented in the routine `VA14` is $6n$, while the memory storage of the L-BFGS method implemented in the routine `VA15` is $2nm + 4n$.

When the scale n of problem (1) is large, also another strategy may be adopted to tackle a minimum point: namely, *partially separable functions* can be exploited. Suppose $f(x)$ is a partially separable function, that is, it can be written in the form

$$f(x) = \sum_{i=1}^p f_i(x), \quad (11)$$

where each of the functions $f_1(x), \dots, f_p(x)$ depends just on a subset of the unknowns x . Evidently, by Equation (11) the gradient and the Hessian matrix of $f(x)$ are given by

$$\nabla f(x) = \sum_{i=1}^p \nabla f_i(x); \quad \nabla^2 f(x) = \sum_{i=1}^p \nabla^2 f_i(x).$$

Thus, drawing our inspiration from the quasi-Newton methods, let B be a quasi-Newton approximation of $\nabla^2 f(x)$, while B_i , $i = 1, \dots, p$, is a quasi-Newton approximation of $\nabla^2 f_i(x)$. Then, numerical experiences proved that $\sum_{i=1}^p B_i$ is often a better approximation of $\nabla^2 f(x)$ than B , as long as a satisfactory quasi-Newton approximation of each $\nabla^2 f_i(x)$ is given.

The AMPL modeling language [61] contains a specific tool, which is able to compute the approximation B_i of $\nabla^2 f_i(x)$ by automatic differentiation [5], so that a sparse representation of $\nabla^2 f(x)$ is available. For other references on this topic see also Griewank and Toint [62,63], Malmedy and Toint [64] and the references therein.

MULTIGRID OPTIMIZATION

Optimization-based multigrid methods are a class of algorithms which have had a fast development in the last decade [65–67], since efficient methods for the optimization of large-scale nonlinear systems governed by differential equations are sought. In particular, let us consider the problem

$$\begin{aligned} \min_x \quad & f[x, u(x)] \\ \text{s.t.} \quad & S[x, u(x)] = 0, \end{aligned} \quad (12)$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, and the variables $u = u(x)$, with $u : \mathbb{R}^n \rightarrow \mathbb{R}^m$, are commonly called *state unknowns* and are implicitly defined by the system of partial differential equations (PDE) $S[x, u(x)] = 0$ (*state equations*).

Optimization-based multigrid methods for solving problem (12) consider a family of optimization problems, associated with problem (12), each corresponding to a different discretization of the system $S[x, u(x)] = 0$. Intuitively speaking, the workload to solve problem (12) by choosing a finer discretization (*grid*) is larger than that by choosing a coarser grid. These methods use the computation on a coarser grid (which implies the solution of unconstrained optimization subproblems), in order to improve an approximate solution of problem (12) on a finer grid.

A very general algorithm for the latter purpose is MG/Opt, which is described in Nash [66]. This algorithm does not apply to a specific family of grids; moreover, it relies on the use of optimization models along with a linesearch procedure, in order to guarantee the global convergence of the method. Further details and a numerical experience for multigrid methods are given in Lewis and Nash [67] and the references therein.

Similar ideas for unconstrained optimization problems have been investigated in Toint *et al.* [68] and Gratton *et al.* [69], in order to solve very large-scale trust-region subproblems. Here, some convergence issues related to solutions satisfying both first- and second-order optimality conditions were studied.

REFERENCES

1. Fletcher R. An overview of unconstrained optimization. In: Spedicato E, editor. Algorithms for continuous optimization. The state of the art. The Netherlands: Kluwer Academic Publishers; 1994. pp. 109–143.
2. Nocedal J, Wright S. Numerical optimization. 2nd ed. Springer series in operations research and financial engineering. New York: Springer; 2006.
3. Nocedal J. Large scale unconstrained optimization. In: Watson GA, Duff I, editors. The state of the art in numerical analysis. Oxford: Oxford University Press; 1997. pp. 311–338.
4. Bertsekas DP. Nonlinear programming. 2nd ed.. Belmont (MA): Athena Scientific; 1999.
5. Griewank A. Automatic differentiation. Philadelphia: SIAM; 2000.
6. Orthegea JM. Introduction to parallel and vector solution of linear systems (frontiers In Computer Science). New York: Plenum Press; 1988.
7. Bertsekas DP. Constrained optimization and Lagrange multiplier methods. Belmont (MA): Athena Scientific; 1996.
8. Armijo L. Minimization of functions having continuous partial derivatives. Pac J Math 1966;3:1–3.
9. Alexandrov NM, Hussaini MY, editors. Multidisciplinary design optimization - state of the art. In: Proceedings of the ICASE/NASA Langley Workshop on Multidisciplinary Design Optimization. SIAM Proceedings Series. Philadelphia: SIAM; 1997.
10. Barzilai J, Borwein JM. Two point step size gradient method. IMA J Numer Anal 1988;8:141–148.
11. Raydan M. The Barzilai and Borwein gradient method for large scale unconstrained minimization problems. SIAM J Optim 1997;7:26–33.

12. Fletcher R. Practical methods of optimization. New York: John Wiley & Sons; 1987.
13. Grippo L, Lampariello F, Lucidi S. A class of nonmonotone stabilization methods in unconstrained optimization. *Numer Math* 1991;59:779–805.
14. Conn AR, Gould NIM, Toint PhL. Trust region methods. Philadelphia: SIAM; 2000.
15. Shultz GA, Schnabel RB, Byrd RH. A family of trust-region-based algorithms for unconstrained minimization. *SIAM J Numer Anal* 1985;22:47–67.
16. Dembo R, Eisenstat S, Steihaug T. Inexact Newton methods. *SIAM J Numer Anal* 1982;19:400–408.
17. Dembo RS, Steihaug T. Truncated-Newton algorithms for large-scale unconstrained optimization. *Math Program* 1983;26:190–212.
18. Nash S. A survey of Truncated-Newton methods. *J Comput Appl Math* 2000;124:45–59.
19. Dembo R, Steihaug T. Truncated-Newton algorithms for large-scale unconstrained optimization. *Math Program* 1983;26:190–212.
20. Gould NIM, Lucidi S, Roma M, *et al.* Exploiting negative curvature directions in linesearch methods for unconstrained optimization. *Optim Methods Softw* 2000;14:75–98.
21. Conn AR, Gould NIM, Toint PhL. LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A). Heidelberg, Berlin: Springer; 1992.
22. Nash S, Sofer A. Assessing a search direction within a Truncated-Newton method. *Oper Res Lett* 1990;9:219–221.
23. Fasano G, Lucidi S. A nonmonotone truncated Newton-Krylov method exploiting negative curvature directions, for large scale unconstrained optimization. *Optim Lett* 2009;3:521–535.
24. Nash S. Newton-type minimization via the Lanczos method. *SIAM J Numer Anal* 1984;21:770–788.
25. Lucidi S, Rochetich F, Roma M. Curvilinear stabilization techniques for Truncated Newton methods in large scale unconstrained Optimization. *SIAM J Optim* 1998;8:916–939.
26. Gould NIM, Lucidi S, Roma M, *et al.* Solving the trust-region subproblem using the Lanczos method. *SIAM J Optim* 1999;9:504–525.
27. Fasano G. Planar-conjugate gradient algorithm for large-scale unconstrained optimization, part 1: theory. *J Optim Theory Appl* 2005;125:523–541.
28. Fasano G. Planar-conjugate gradient algorithm for large-scale unconstrained optimization, part 2: application. *J Optim Theory Appl* 2005;125:543–558.
29. Schlick T, Fogelson A. TNPACK - a truncated Newton package for large-scale problems: I. Algorithm and usage. *ACM Trans Math Softw* 1992;18:46–70.
30. Nash S. Preconditioning of truncated-Newton methods. *SIAM J Sci Stat Comput* 1985;6:599–616.
31. Morales JL, Nocedal J. Automatic preconditioning by limited memory quasi-Newton updating. *SIAM J Optim* 2000;10:1079–1096.
32. Roma M. A dynamic scaling based preconditioning for truncated Newton methods in large scale unconstrained optimization. *Optim Methods Softw* 2005;20:693–713.
33. Moré JJ, Sorensen DC. On the use of directions of negative curvature in a modified Newton method. *Math Program* 1979;16:1–20.
34. Fasano G, Roma M. Iterative computation of negative curvature directions in large scale optimization. *Comput Optim Appl* 2007;38:81–104.
35. Ferris MC, Lucidi S, Roma M. Nonmonotone curvilinear linesearch methods for unconstrained optimization. *Comput Optim Appl* 1996;6:117–136.
36. Goldfarb D. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Math Program* 1980;18:31–40.
37. Olivares A, Moguerza JM, Prieto FJ. Non-convex optimization using negative curvature within a modified linesearch. *Eur J Oper Res* 2008;189:706–722.
38. Grippo L, Lampariello F, Lucidi S. A truncated Newton method with nonmonotone linesearch for unconstrained optimization. *J Optim Theory Appl* 1989;60:401–419.
39. Powell MJD. A new algorithm for unconstrained optimization. In: Mangasarian O, Ritter K, editors. *Nonlinear programming*. New York: Academic Press; 1970. pp. 31–65.
40. Dennis J, Mei H. Two new unconstrained optimization algorithms which use function and gradient values. *J Optim Theory Appl* 1979;28:453–482.
41. Byrd R, Schnabel R, Schultz G. An approximate solution of the trust region problem by minimization over two-dimensional subspaces. *Math Program* 1988;40:247–263.

42. Sorensen DC. Newton's method with a model trust-region modification. *SIAM J Sci Stat Comput* 1982;19:409–427.
43. Moré JJ, Sorensen DC. Computing a trust region step. *SIAM J Sci Stat Comput* 1983;4:553–572.
44. Steihaug T. The conjugate gradient method and trust regions in large-scale optimization. *SIAM J Numer Anal* 1983;20:626–637.
45. Toint PhL. Towards an efficient sparsity exploiting Newton method for minimization. In: Duff IS, editor. *Sparse matrices and their uses*. London: Academic Press; 1981. pp. 57–88.
46. Stoer J. Solution of large linear systems of equations by conjugate gradient type methods. In: Bachem A, Grottschel M, Korte B, editors. *Mathematical programming. The state of the art*. Berlin/Heidelberg: Springer; 1983. pp. 504–565.
47. Hestenes M. *Conjugate direction methods in optimization*. New York: Springer; 1980.
48. Pytlak R. *Conjugate gradient algorithms in nonconvex optimization*. Berlin/Heidelberg: Springer; 2009.
49. Gilbert J, Nocedal J. Global convergence properties of conjugate gradient methods for optimization. *SIAM J Optim* 1992;2:21–42.
50. Zoutendijk G. Nonlinear programming computational methods. In: Abadie J, editor. *Integer and nonlinear programming*. Amsterdam: North-Holland Publishing Co.; 1970. pp. 37–86.
51. Al-Baali M. Descent property and global convergence of the Fletcher-Reeves method with inexact line search. *IMA J Numer Anal* 1985;5:121–124.
52. Grippo L, Lucidi S. A globally convergent version of the Polak-Ribiere conjugate gradient method. *Math Program* 1997;78:375–391.
53. Harwell Subroutine Library. *A catalogue of subroutines*. Harwell, England: AEA Technology; 1998.
54. Davidon WC. Variable metric method for minimization. *SIAM J Optim* 1991;1:1–17.
55. Broyden CG. The convergence of a class of double-rank minimization algorithms, part I and II. *J Inst Math Appl* 1970;6:76–90, 222–231.
56. Nocedal J. Updating Quasi-Newton matrices with limited storage. *Math Comput* 1980;35:773–782.
57. Al-Baali M. Extra-updates criterion for limited memory BFGS algorithm for large scale nonlinear optimization. *J Complex* 2002;18:557–572.
58. Burdakov OP, Martinez JM, Pilotta EA. A limited-memory multipoint symmetric secant method for bound constrained optimization. *Ann Oper Res* 2002;117:51–70.
59. Nash S, Nocedal J. A numerical study of a limited memory BFGS method and the truncated Newton method for large scale optimization. *SIAM J Optim* 1991;1:358–372.
60. Byrd R, Nocedal J, Zhu C. Towards a discrete Newton method with memory for large-scale optimization. In: Di Pillo G, Giannessi F, editors. *Nonlinear optimization and applications*. New York: Plenum Publishing; 1995. pp. 1–12.
61. Fourer R, Gay DM, Kernighan BW. *AMPL: a modeling language for mathematical programming*. South San Francisco (CA): The Scientific Press; 1993.
62. Griewank A, Toint PhL. Partitioned variable metric updates for large structured optimization problems. *Numer Math* 1982;39:119–137.
63. Griewank A, Toint PhL. Local convergence analysis of partitioned quasi-Newton updates. *Numer Math* 1982;39:429–448.
64. Malmedy V, Toint PhL. Approximating Hessians in multilevel unconstrained optimization. Report 08/19. FUNDP - University of Namur; 2008.
65. Lewis RM, Nash SG. A multigrid approach to the optimization of systems governed by differential equations. *AIAA Paper* 2000–4890. Reston (VA): American Institute of Aeronautics and Astronautics; 2000.
66. Nash SG. A multigrid approach to discretized optimization problems. *J Comput Appl Math* 2000;14:99–116.
67. Lewis RM, Nash SG. Model problems for the multigrid optimization of systems governed by differential equations. *SIAM J Sci Comput* 2005;26:1811–1837.
68. Toint PhL, Tomanos D, Weber-Mendonca M. A multilevel algorithm for solving the trust-region subproblem. *Optim Methods Softw* 2009;24:299–311.
69. Gratton S, Mouffe M, Sartenauer A, *et al.* Numerical experience with a recursive trust-region method for multilevel nonlinear optimization. *Optim Methods Softw* 2010;25:359–386.

MILITARY OPERATIONS RESEARCH SOCIETY (MORS)

TRENA C. LILLY
The Johns Hopkins
University Applied
Physics Laboratory,
Laurel, MD

The Military Operations Research Society (MORS) is a professional society dedicated to promoting excellence in analysis within the national security community through the advancement and application of the interdisciplinary field of operations research to national security issues.

MORS has served the US Department of Defense analytic community for over 40 years. Today, the society also serves other parts of the government concerned with national security matters. Under the sponsorship of the army, navy, air force, marine corps, Office of the Secretary of Defense, the Joint Staff, and the Department of Homeland Security, the objective of MORS is to enhance the quality and effectiveness of operations research (OR) as applied to national security issues.

MORS vision is to “become the recognized leader in advancing the national security analytic community through the advancement and application of the interdisciplinary field of operations research to national security issues, being responsive to our constituents, enabling collaboration and development opportunities, and expanding our membership and disciplines, while maintaining our profession’s heritage.” This vision encompasses all aspects of national security including not only the military but also Homeland Security and the other agencies of government—as well as allies of the United States.

Members of the society include a cross section of professional defense analysts, operators, and managers from government, industry, and academia. Their involvement fosters professional interchange within the

military operations research community, the sharing of insights and information on challenging national security issues, and specific support to decision makers in the many organizations and agencies that address national defense. MORS provides an array of meetings and publications to expose its members to the broad spectrum of thought by distinguished analysts and senior members of our community. In particular, the society provides a unique environment in which classified presentations and discussions can take place with joint service participation and peer criticism from the full range of students, theoreticians, practitioners, and users of military analysis. Throughout its activities, the society promotes the highest standards of professional methodology, individual excellence, and ethical conduct.

BACKGROUND

The first Military Operations Research Symposium, sponsored by the Office of Naval Research (ONR), Pasadena, was held at Corona Naval Ordnance Laboratory, Corona, California, in August 1957. The subject of the meeting was Air Defense, and 83 scientists attended. This and subsequent early meetings were oriented to fulfill the needs of the Operations Research Community on the West Coast. Starting with the Eighth MORS, the symposia became nationally oriented joint service meetings. The first national symposium was the Ninth MORS, held at Fort Monroe, Virginia, in April 1962. From the First through the 10th MORS, there was no formal organization to stage the meetings. They were conducted by ONR–Pasadena with the help of a volunteer steering committee.

Beginning with the eleventh MORS, ONR–Washington assumed supervision and hired a contractor to perform the work in cooperation with a volunteer

executive committee. This arrangement continued until April 1966 during which MORS was incorporated under the laws of Virginia.

Since its incorporation in 1966, MORS has been led by a Board of Directors. The Board of Directors consists of 28 voting members and 2 nonvoting members: the Chief Executive Officer and the Immediate Past President. The board is also led by an Executive Council consisting of the President, the President-Elect, the Immediate Past President, the Vice President for Financial Management, the Vice President for Member and Societal Services, the Secretary of the Society, and the Chief Executive Officer.

Figure 1 shows the PHALANX Newsletter announcement of MORS incorporation.

Table 1 shows the Past Presidents of the Military Operations Research Society.

RECOGNIZING EXCELLENCE AND LONG-TERM CONTRIBUTIONS

MORS is proud of its heritage, as it closely reflects the heritage of operations research in the United States. Since it was incorporated, MORS has recognized and continues to recognize individuals whose contributions continue to promote excellence in the field of OR.

Fellows of the Society

MORS recognizes those individuals who have made long-lasting and significant contributions to the society as *Fellows of the Society*. Fellows are elected for life. Fellows were first elected in 1989 and are elected annually at the December Board of Directors meeting. These individuals represent many of the great minds in the military and national security OR field.

Awards

In direct support of the society's purpose to enhance the quality and effectiveness of national security analysis, MORS annually

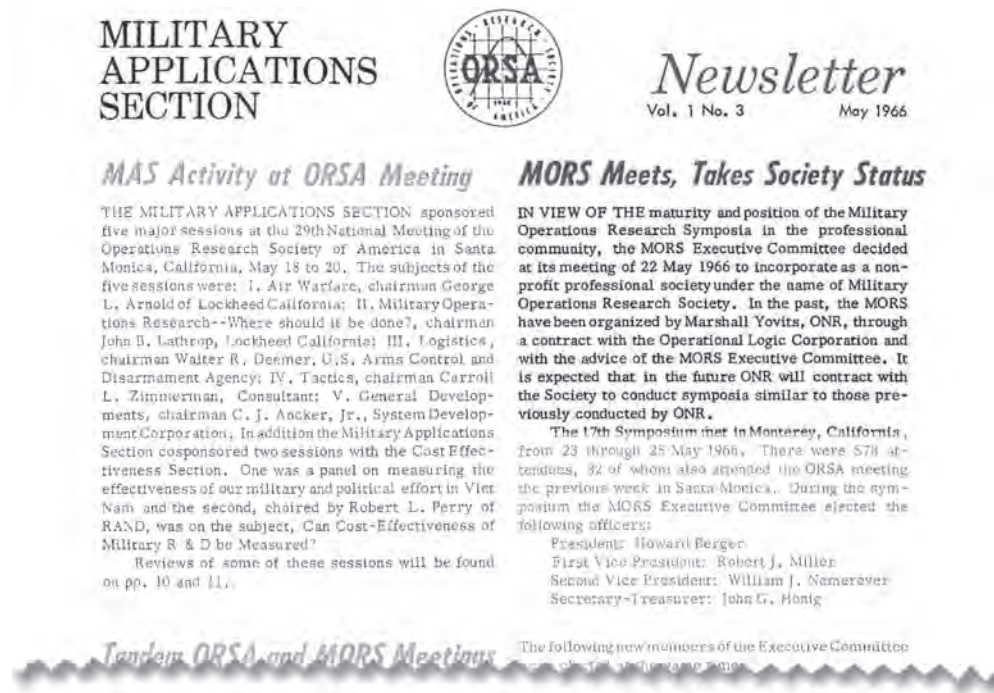


Figure 1. PHALANX article highlighting MORS incorporation.

Table 1. MORS Past Presidents

Lewis Leake (1966)	Richard E. Garvey Jr. (1986–87)
Howard M. Berger (1966–67)	George H. Dimon, Jr., FS (1987–88)
Arthur Stein (1967–68)	Dr. Kleber S. Masterson, Jr. (1988–89)
Dr. John Honig (1968–69)	Edward C. Brady, FS (1989–90)
Dr. Seth Bonder (1969–1970)	Mary G. B. Pace, FS (1990–91)
Dr. Jack R. Borsting, FS (1970–71)	Vernon M. Bettencourt, Jr, FS (1991–92)
Robert Stevens (1971–72)	E. B. Vandiver III, FS (1992–93)
Kenneth Yudowitch (1972–73)	Dr. Gregory S. Parnell, FS (1993–94)
Clayton Thomas (1973–74)	Brian R. McEnany, FS (1994–95)
John K. Walker, Jr (1974–75)	Christine A. Fossett, FS (1995–96)
Dr. Marion R. Bryson, FS (1975–76)	Frederick E. Hartman, FS (1996–97)
Stephen A. Murtaugh, FS (1976–77)	Dr. Jerry A. Kotchka, FS (1997–98)
David E. Spencer, FS (1977–78)	Dennis R. Baer, FS (1998–99)
Prof. David A. Schradly, FS (1978–79)	Dr. Robert S. Sheldon, FS (1999–2000)
Jack Englund (1979–80)	Dr. Roy E. Rice, FS (2000–2001)
Charles E. Woods (1980–81)	Dr. Thomas L. Allen, FS (2001–2002)
Amoretta M. Hoeber (1981–82)	Edward A. Smyth, FS (2002–2003)
Dr. Marion L. Williams, FS (1982–83)	Dr. Willie McFadden II, FS (2003–2004)
James N. Bexfield, FS (1983–85)	Dr. Andrew G. Loerch, FS (2004–2005)
Prof. Wayne P. Hughes, FS (1985–86)	Dr. Suzanne M. Beers, FS (2005–2006)
Patrick J. McKenna, FS (2006–2007)	John F. Keane, FS (2007–2008)
Dr. Michael J. Kwinn, Jr. (2008–2009)	Kirk A. Michealson (2009–2010)
Terrance J. McKearney (2010–2011)	Trena C. Lilly (2011–2012)
Michael W. Garrambone (2012–2013)	

Abbreviation: FS, Fellow of the Society.

presents awards that recognize outstanding individuals in the profession as well as the outstanding work by individuals and teams. These awards include the Wanner Award, the Thomas Award, the Walker Award, and the Hughes Award.

The Wanner Award. Named after Vance R. Wanner, a respected professional who provided enduring value to the military operations research community, this award recognizes individuals for long-term contributions to the profession. The award gives evidence of MORS' belief that OR represents a symbiosis of people and systems, not merely a set of techniques for manipulating data and evaluating results. The award emphasizes that leadership in military OR includes more than mastery of a set of analytic tools. This award, therefore, seeks to identify and recognize members of the profession who have advanced beyond mere excellence in individual achievement and have expanded the application of military OR and raised its standards. This award also gives consideration to the management and leadership of military OR— its direction and control, quality,

applicability, and timeliness. The award is given with the hope that it will inspire members of the military OR community to broaden their scope of participation in the advancement of the profession.



Clayton J. Thomas Award. The Clayton J. Thomas Award recognizes individuals for long-term technical contributions to enhance the analytical capabilities of the profession. This award recognizes outstanding individuals for consistent, sustained technical

contributions to improve the analytical underpinnings of the military OR profession.



John K. Walker, Jr. Award. The John K. Walker, Jr. Award is named in recognition of John K. Walker, Jr., a respected professional, *PHALANX* Editor for 12 years and *PHALANX* Editor Emeritus for 7 years. By his long-term involvement with the MORS, Mr. Walker gave much of enduring value to the military OR community. This award recognizes the author(s) of the technical article judged to be the best such article published in *PHALANX*, the Bulletin of Military Operations Research, during the previous calendar year. This award also emphasizes that *PHALANX* promotes communication among practitioners of military OR that is timely, important, and interesting to the *PHALANX* audience.



Wayne P. Hughes, Jr. Award. The Wayne Pilo Hughes, Jr. Award is named after a former member of the MORS Board of Directors, the 20th MORS President (1985–1986), and a MORS Fellow of the Society (FS) of the First Fellow's Class of 1989. Wayne Hughes served as a distinguished commander, a military analyst, and a Service school educator. Throughout his career, he contributed as a monograph editor and Naval Tactics book author and served as mentor to an enormously long list of Junior Analysts. This award recognizes young analysts who are already making an impact on OR through outstanding publication(s) highlighting their analyses, development of significant new analysis, or the use of research results for solving military problems.



Prizes. MORS annually presents awards in direct support of the society's purpose to enhance the quality and effectiveness of national security analysis through the application of military OR techniques and the improvement of its set of analytical tools. MORS offers the David Rist Prize and the Richard H. Barchi Prize, as well as the Graduate Research Prizes.

David Rist Prize. The David Rist Prize, named after a former member of the MORS Board of Directors, was first awarded in 1965 and is presented to the best study or

other operations research-based efforts, for example, analyses and methodology improvements that influenced major decisions or processes. The objective of the David Rist Prize is to recognize the practical benefits sound OR can have on real-life decision making as well as to provide recognition to OR practitioners whose study/project recommendations were implemented, in order to reward them and to encourage a high degree of excellence in military OR.



Richard H. Barchi Prize. First awarded in 1983, the *Richard H. Barchi Prize* is named for Commander Richard H. Barchi, United States Navy (USN), a former member of the MORS Board of Directors. The objective of the Richard H. Barchi Prize is to provide recognition to authors of the best paper presented at the MORS' annual Symposium, in order to encourage a high degree of excellence in military OR. To be eligible, papers must report on original analysis of methodology developed by the authors, be relevant to problems of military planning, operations, or management and must represent OR or systems analysis in the broadest sense. The papers must be clear, logical, and coherent in structure and should report on completed analysis.



Graduate Research Prize. A Graduate Research Prize for OR students is presented for each graduating class at the US Naval Postgraduate School and the US Air Force Institute of Technology. The prize consists of a \$200 honorarium and a plaque. The prize at the Naval Postgraduate School has been named the Military Operations Research Society Stephen A. Tisdale Prize in honor of LCDR Stephen A. Tisdale, a winner of the Prize who was killed in a military aircraft accident in 1991.

The Graduate Research Prize emphasizes that progress in military OR depends on continuous improvement to a kit of analytic tools. It therefore seeks to identify and recognize members of the profession who have excelled in individual achievement, have expanded the application of military OR techniques and have improved its set of analytical tools. It also gives consideration to the development of the analytical underpinnings of military OR. The award is given to inspire members of the military OR community to continue technical, in-depth participation in the advancement of the profession by emphasizing the analytical foundations of the profession.

MORS MEETINGS

MORS Symposium

For over 45 years, the annual MORS Symposium has been a premier opportunity

MILITARY APPLICATIONS SECTION



Newsletter
Vol. 1 No. 3 May 1966

MAS Activity at ORSA Meeting

THE MILITARY APPLICATIONS SECTION sponsored five major sessions at the 29th National Meeting of the Operations Research Society of America in Santa Monica, California, May 18 to 20. The subjects of the five sessions were: I. Air Warfare, chairman George L. Arnold of Lockheed California; II. Military Operations Research--Where should it be done?, chairman John B. Lathrop, Lockheed California; III. Logistics, chairman Walter R. Deemer, U.S. Arms Control and Disarmament Agency; IV. Tactics, chairman Carroll L. Zimmerman, Consultant; V. General Developments, chairman C. J. Ancker, Jr., System Development Corporation. In addition the Military Applications Section cosponsored two sessions with the Cost Effectiveness Section. One was a panel on measuring the effectiveness of our military and political effort in Viet Nam and the second, chaired by Robert L. Perry of RAND, was on the subject, Can Cost-Effectiveness of Military R & D be Measured?

Reviews of some of these sessions will be found on pp. 10 and 11.

MORS Meets, Takes Society Status

IN VIEW OF THE maturity and position of the Military Operations Research Symposia in the professional community, the MORS Executive Committee decided at its meeting of 22 May 1966 to incorporate as a non-profit professional society under the name of Military Operations Research Society. In the past, the MORS have been organized by Marshall Yovits, ONR, through a contract with the Operational Logic Corporation and with the advice of the MORS Executive Committee. It is expected that in the future ONR will contract with the Society to conduct symposia similar to those previously conducted by ONR.

The 17th Symposium met in Monterey, California, from 23 through 25 May 1966. There were 578 attendees, 82 of whom also attended the ORSA meeting the previous week in Santa Monica. During the symposium the MORS Executive Committee elected the following officers:

President: Howard Berger
First Vice President: Robert J. Miller
Second Vice President: William J. Nemerever
Secretary-Treasurer: John G. Honig

Tandem ORSA and MORS Meetings

THE NATIONAL MEETINGS of the Operations Research Society of America and of the Military Operations Research Society were held back-to-back on the West Coast this May. The ORSA meeting was held at Santa Monica, May 18-20, ending on a Friday. The MORS was held at Monterey, California, May 23-25, starting on a Monday. Many persons attended both meetings. The intervening weekend was used to good advantage by some for a pleasant drive up the coast to Monterey, or for a happy time at San Francisco. Eighty two of the 578 MORS attendees had also attended the ORSA meeting.

Planning is now well underway for a similar arrangement on the East Coast. The 30th National Meeting of ORSA will be held October 17-19 in Durham, North Carolina. The chairman is George E. Nicholson, Jr., of the University of North Carolina. MAS will

See *Tandem*, p. 2

The following new members of the Executive Committee were elected at the same time:

Lloyd F. Bell
Alfred Blumstein
Seth Bonder
Melvin Dresher
Joseph H. Engel
Murray Greyson
Laurence A. Harvey
Alexander L. Slafkosky

Under the new By-Laws of the Society, Lewis A. Leake will serve as the Prior Past President on the Executive Council along with the newly elected officers.

Now that MORS has emerged as a full-fledged professional society the Executive Committee feels that it is an appropriate time to review the operation of the symposia and to consider what long-range relationships should exist between ORSA and MORS. In order

See *Society*, p. 3

Figure 2. One of the first PHALANX publications.

for the national security community to exchange information, examine research and discuss critical national security topics. Held in notable locations, the Symposium gathers over a thousand OR

professionals from military, government, industry and academic ranks to share best practices and take advantage of peer-to-peer networking in a supportive environment.



Figure 3. First PHALANX published in 1974 in conjunction with MORS.

Special Meetings

MORS Special Meetings are convened five or more times a year to investigate critical national security topics. The purpose of the meetings is to explore methodology, examine an issue and develop potential recommendations. Participants have a unique, high

level opportunity to explore the leading edge with their peers and influence the course of research.

Education and Professional Development Colloquium

The MORS Education and Professional Development Colloquium brings together

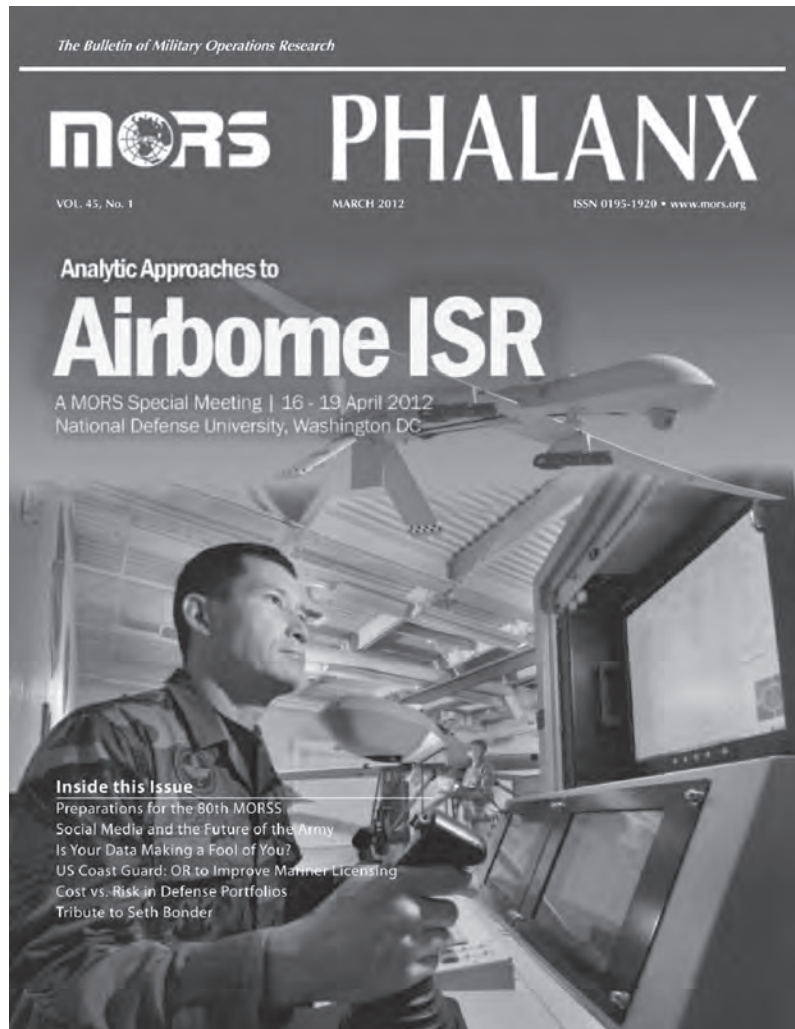


Figure 4. PHALANX 2012.

students, academics and professionals from the OR community to meet, share and learn about current trends in OR education and professional development, as well as new research emerging from OR programs at the academies, universities and colleges.

The colloquium helps to prepare the next generation of national security analysts for the challenges of the profession with the guidance of experienced analysts.

PUBLICATIONS

MORS publications provide sources of unique and vital information and insight on the business and science of OR. These publications include PHALANX, the MORS Journal, Books, and unclassified reports from Symposia and Special Meetings. MORS publications provide a foundational “body of knowledge,” as well as information on trends and the leading edge of research.

Military Operations Research

A Journal of the Military Operations Research Society



Figure 5. MOR Journal.

PHALANX

Beginning in 1974, *PHALANX*, the Bulletin of Military Operations Research was published quarterly by the Military Applications Section of the Institute for Operations Research and Management Sciences (INFORMS/MAS) jointly with MORS. *PHALANX* provides a cross section of important current research, meetings summaries, MORS news and informative oral histories.

Figure 2 shows the first *PHALANX* publication.

Figure 3 shows the first *PHALANX* published in 1974 in conjunction with MORS.

Figure 4 shows the current edition of *PHALANX*.

Military Operations Research (MORS) Journal

Military Operations Research is a peer-reviewed journal of high academic quality.

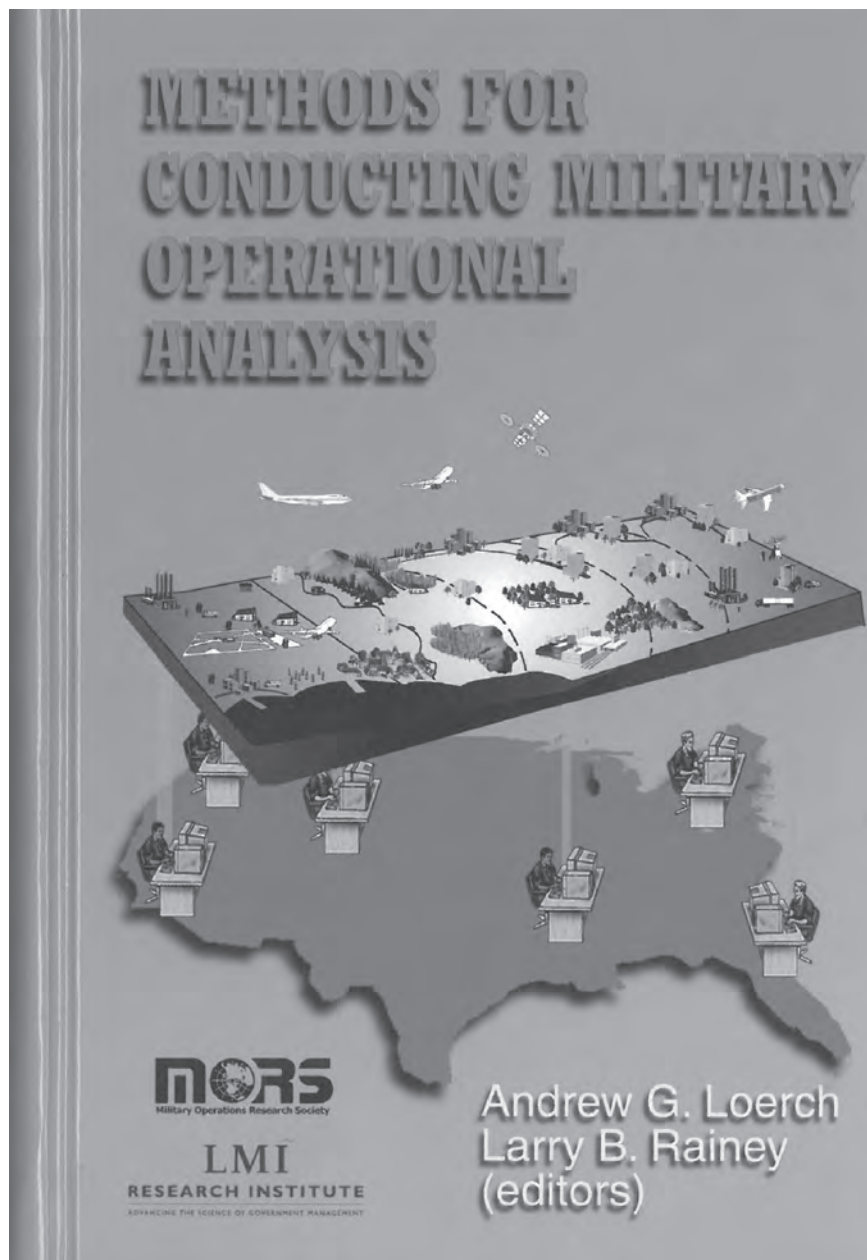


Figure 6. MORS Publication “Methods for Conducting Military Operational Analysis”.

The MORS Journal publishes articles that describe OR methodologies and theories used in key military applications. Of particular interest are papers that present case studies showing innovative OR applications,

applications of OR to major policy issues, introduce interesting new problem areas, highlight educational issues, and document the history of military OR. Figure 5 shows the MOR Journal.

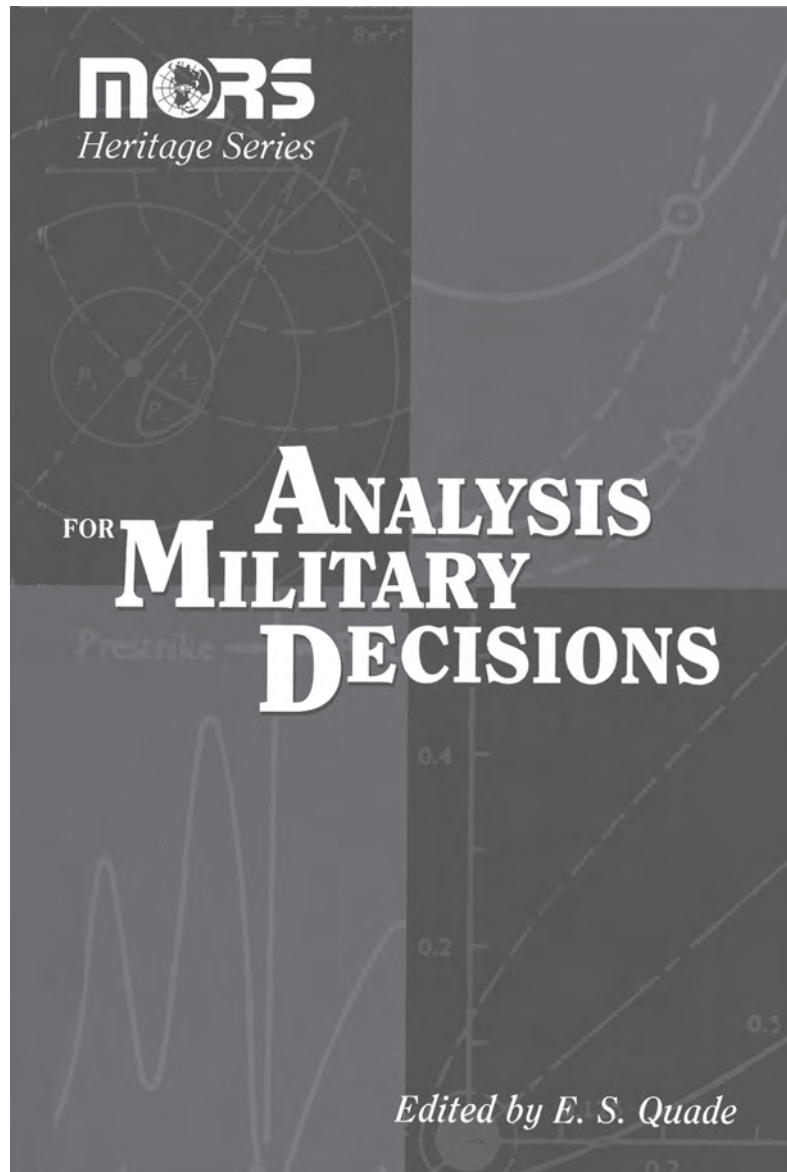


Figure 7. MORS Heritage Series Publication “Analysis for Military Decisions”.

MORS Books

MORS books are a valuable source of reference material on OR methods, models and techniques, as well as informative and useful historic accounts.

Figure 6 highlights the MORS publication “*Methods for Conducting Military Operational Analysis*”.

Figure 7 highlights the MORS Heritage Series publication “*Analysis for Military Decisions*”.



Figure 8. MORS Publication “Military Modeling for Decision Making”.

Figure 8 highlights the MORS publication “*Military Modeling for Decision Making*”.

SUMMARY

The Military Operations Research Society (MORS) will continue to support the national security analytic community through the advancement and application of the interdisciplinary field of OR to national security issues, by being responsive to its constituents, enabling collaboration and professional development opportunities, and expanding its membership and disciplines, while maintaining the profession’s heritage.

MILP SOFTWARE

JEFFREY T. LINDEROTH
Department of Industrial and
Systems Engineering,
University of
Wisconsin-Madison, Madison,
Wisconsin

ANDREA LODI
DEIS, Università di Bologna,
Bologna, Italy

INTRODUCTION

This article discusses software for solving a general *mixed-integer linear program* (MILP) in the form

$$\min\{c^T x : Ax \geq b, x \geq 0, x_j \in \mathbb{Z} \forall j \in I\}. \quad (1)$$

The algorithmic approach relies on the iterative solution, through general-purpose techniques, of the *linear programming* (LP) relaxation

$$\min\{c^T x : Ax \geq b, x \geq 0\}, \quad (2)$$

that is, the same as problem (1) above but the integrality requirement on the x variables in the set I has been dropped. We denote an optimal solution of problem (2) as x^* . The reason for dropping such constraints is that MILP is NP-hard while LP is polynomially solvable and general-purpose techniques for its solution are efficient in practice.

The articles titled ***Simplex-Based LP Solvers*** and ***Interior-Point Linear Programming Solvers*** cover state-of-the-art LP techniques and solvers, while in the present article we concentrate on the basic characteristics and components of current—commercial and noncommercial—MILP solvers.

Roughly speaking, using the LP computation as a tool, MILP solvers integrate

the *branch-and-bound* and the *cutting-plane* algorithms through variations of the general *branch-and-cut* scheme proposed by Padberg and Rinaldi [1] in the context of the *traveling salesman problem* (TSP). The articles titled ***Mathematical Programming Approaches to the Traveling Salesman Problem*** and ***Branch and Cut*** discuss, in detail, approaches for the TSP and the branch-and-cut approach in its full generality, respectively. Nevertheless, with the aim of giving a quick overview and fixing the notation, we briefly discuss in the following both the branch-and-bound and the cutting-plane algorithms.

The Branch-and-Bound Algorithm

In its basic version the branch-and-bound algorithm [2] iteratively partitions the solution space into sub-MILPs (the children nodes), which have the same theoretical complexity of the originating MILP (the father node or the root node if it is the initial MILP). Usually, for MILP solvers, the branching creates two children by rounding of the solution of the LP relaxation value of a fractional variable, say x_j , constrained to be integral

$$x_j \leq \lfloor x_j^* \rfloor \quad \vee \quad x_j \geq \lfloor x_j^* \rfloor + 1. \quad (3)$$

The two children above are often referred to as *left* (or “down”) branch and *right* (or “up”) branch, respectively. On each of the sub-MILPs the integrality requirement on the variables $x_j, \forall j \in I$ is relaxed and the LP relaxation is solved. Despite the theoretical complexity, the sub-MILPs become smaller and smaller because of the partition mechanism (basically some of the decisions are taken) and eventually the LP relaxation is directly integral for all variables in I . In addition, the LP relaxation is solved at every node to decide if the node itself is worthwhile to be further partitioned: if the LP relaxation value is already not smaller than the best feasible solution encountered so far, called

incumbent, the node can safely be fathomed because none of its children will yield a better solution than the incumbent. Finally, a node is also fathomed if its LP relaxation is infeasible.

The Cutting-Plane Algorithm

Any MILP can be solved without branching by simply finding its “right” LP description; more precisely, the *convex hull* of its (mixed-)integer solutions [3]. In order to do that, one has to iteratively solve the so-called *separation problem*

Given a feasible solution x^* of the LP relaxation (2) which is not feasible for the MILP (1), find a linear inequality $\alpha^T x \geq \alpha_0$ which is valid for system (1), that is, satisfied by all feasible solutions $\bar{x}$ of the system (1), while it is violated by x^* , that is, $\alpha^T x^* < \alpha_0$.

Any inequality solving the separation problem is called a *cutting plane* (or a *cut*, for short) and has the effect of tightening the LP relaxation to better approximate the convex hull.

Gomory [3] has given an algorithm that converges in a finite number of iterations for pure *integer linear programming* (ILP)¹, with integer data. Such an algorithm solves the separation problem above in an efficient and elegant manner in the special case in which x^* is an optimal basis of the LP relaxation. No algorithm of this kind is known for MILPs.

The idea behind integrating the two algorithms above is that LP relaxations (2) do not naturally well approximate, in general, the convex hull of (mixed-)integer solutions of MILPs (1). Thus, some extra work to devise a better approximation by tightening any relaxation with additional linear inequalities (cutting planes) increases the chances that fewer nodes in the search tree are needed. On the other hand, pure cutting-plane algorithms show, in general, a slow convergence and the addition of too many cuts can lead to very large LPs, which in turn present

numerical difficulties for the solvers. The branch-and-cut algorithm has been proven to be very effective initially for combinatorial optimization problems (like TSP) with special-purpose cuts based on a polyhedral analysis and later on in the general MILP context.

This article is split into two parts. In the section titled “Basic Components of MILP Solvers,” we discuss the basic components of nowadays MILP solvers. In the section titled “MILP Software,” we list the available, commercial and noncommercial, solvers with special emphasis to their flexibility in terms of being usable within different modeling and development environments.

Some of the solvers that are discussed in the above-mentioned section have their own modeling environment and modeling language. Moreover, a recent trend for MILP software producers has been adding additional capabilities to the solvers, like that of solving some special classes of (mixed-integer) nonlinear programs (MINLPs). Although modeling languages and the possibility of solving MINLPs within an MILP enumerative framework are very interesting topics, they are not discussed in detail in this article. The reader is referred to the article titled for modeling environments and languages and to the articles **Conic Optimization Software**; **Software for Nonlinearly Constrained Optimization**; and **MINLP Solver Software** for NLP and MINLP software.

BASIC COMPONENTS OF MILP SOLVERS

Nowadays MILP solvers incorporate key ideas developed during the first 50 years of integer programming. In the next sections, we discuss the main ingredients in a concise way. For more details, the reader is referred to Achterberg [4] and Lodi [5] and to many articles in this encyclopedia.

Presolving

In the presolving (often called *preprocessing*) phase, the solver tries to detect certain changes in the input that will probably lead to a better performance of the solution process.

¹ILPs are the special case of MILPs where all variables belong to I , that is, are constrained to be integers.

This is generally done without “changing” the set of optimal solutions of the problem at hand² and it affects two main situations.

On the one side, it is often the case that MILP models can be improved with respect to their given formulation³ by either removing redundant information (variables or constraints) or strengthening the variable bounds generally by exploiting the integrality of variables in I . Modern MILP solvers have the capability of cleaning up the models so as to create a presolved instance associated with the original one on which the MILP technology is then applied. The advantage is twofold: first the LP relaxation becomes smaller (then, generally, quicker to solve), and second such relaxations become stronger, that is, better approximating the convex hull of (mixed-)integer solutions. Finally, more sophisticated presolve mechanisms are also able to discover important implications and substructures that might be of fundamental importance later on in the computation for both branching purposes and cutting-plane generation.

For a detailed overview of presolving techniques, the reader is referred to Savelsbergh [7] and Martin [8].

Cutting-Plane Generation

The importance of cutting planes has been already pointed out in the introduction: without iteratively strengthening the approximation of the convex hull of (mixed-)integer solutions provided by the LP relaxation, current enumerative schemes would fail in solving difficult MILP instances.

The arsenal of separation algorithms has been continuously enlarged over the years. A large group of cuts, with strong relationship among each other, includes

Chvátal–Gomory cuts, Gomory mixed-integer cuts, mixed-integer rounding cuts, $\{0, \frac{1}{2}\}$ cuts, lift-and-project cuts, and split cuts. Essentially, all these inequalities are obtained by applying a disjunctive argument, exploiting the integrality, on an integer or mixed-integer set given by a single constraint generally derived by aggregating many others. This group of cuts is presented in a brilliant and unified way by Cornuéjols [9].

Among the more “combinatorial” cutting planes, that is, those whose derivation is not directly associated with disjunctions on the integer variables, certainly the most used and useful ones are clique and cover inequalities. Cover inequalities played a fundamental role in pioneering MILP solvers [10,11].

These classes of inequalities are discussed in the sections titled “Cutting Planes” and “Automatic Convexification” in this encyclopedia.

Sophisticated Branching Strategies

The branching mechanism introduced in the introduction section requires to take two independent and important decisions at any step: node and variable selections.

Node Selection. This is very classical: one extreme is the so-called *best-bound first* strategy in which one always considers the most promising node, that is, the one with the smallest LP value, while the other extreme is *depth first* where one goes deeper in the tree and starts backtracking only once a node is fathomed, that is, it is either (mixed-)integer feasible, or LP infeasible or it has a lower bound not better (smaller) than the incumbent. The pros and cons of each strategy are well known: the former explores a less number of nodes but generally maintains a larger tree in terms of memory while the latter might need to explore a very large number of nodes. Modern software packages, typically, by default employ a hybrid of these two search techniques.

Variable Selection. The variable selection problem is the one of the methods to decide how to partition the current node, that is, on which variable to branch on in order to create the two children. For a long time, a

²In fact, the set of optimal solutions might change due to presolve, for example, in the case of symmetry breaking reductions; see Margot [6] and the article titled *Heuristics in Mixed Integer Programming*.

³This is especially true for models originated from real-world applications and created by using modeling languages.

classical choice has been branching on the most fractional variable, that is, in the 0–1 case the closest to 0.5. This rule has been computationally shown [12] to be worse than a random selection. However, it is of course cheap to evaluate. In order to devise stronger criteria one has to do much more work. The extreme is the so-called *strong branching* technique [13], in which at any node one has to tentatively branch on each fractional variable and select the one on which the increase in the bound on the left branch times the one on the right branch is maximum. In such a version strong branching is clearly impractical but its computational effort can be limited by (i) defining a small candidate set of variables to branch on and (ii) heuristically solving each LP by limiting *a priori* the number of simplex pivots to be performed. Other sophisticated techniques are *pseudocost branching* [14] and, the recently introduced *reliability branching* [12].

Primal Heuristics

Although primal heuristics have been part of the arsenal of the MILP solvers, almost since the beginning, only recently have they become a crucial component. This is likely in response to user demands. Solving an MILP to optimality might be less important in application than providing “good” solutions within short computing times.

The article titled *Heuristics in Mixed Integer Programming* is devoted to heuristics for mixed-integer programming (MIP), and two main classes of primal heuristics are considered.

On the one hand, at the root node and often at the internal nodes of the search tree, the solvers apply heuristics with the aim of converting a solution of the LP relaxation into an integer feasible solution. This is obtained through *rounding* an LP solution either component by component (without any backtracking or additional LP solve) or by *diving*, that is, rounding one or more variables and solving the LP relaxation again, iteratively. The most successful algorithm falling in this category is the *feasibility pump* heuristic proposed (for 0–1 MILPs) by Fischetti *et al.* [15].

On the other hand, once one or more feasible solutions are available, *improvement* heuristics are executed to improve the quality of the current incumbent solution. These heuristics are usually local search methods and we have recently seen a variety of algorithms which solve sub-MILPs for exploring the neighborhood of the incumbent or of a set of solutions. When these sub-MILPs are solved in a general-purpose manner, that is, through a nested call to an MILP solver, this is referred to as *MIPping* [16]. The most successful algorithms falling in this category are *local branching* by Fischetti and Lodi [17] and *relaxation induced neighborhood search* by Danna *et al.* [18].

Parallel Implementation

The branch-and-bound algorithm is a natural one to parallelize, as nodes of the search tree may be independently processed. A long body of research and implementations of parallel branch-and-bound are outlined in the survey [19]. The two types of parallel integer programming research can be loosely categorized based on the type of parallel computing architecture used. Distributed-memory architectures rely on message passing to communicate results of the algorithm. Shared-memory computers communicate information among CPUs by reading from and writing to a common memory pool.

Eckstein [20,21] was among the first to write an industrial strength parallel branch-and-bound code for general MILP. Over the past few years, there has been an emergence of *multicore* computer architectures, such that nearly all modern CPUs contain more than one unit on which computation may be performed. MILP software has followed this trend, so that many of the below surveyed packages have the ability to run multiple threads of computation, thus making use of multiple cores. For many applications, an important characteristic of the branch-and-bound algorithm is to be able to produce solutions that are *reproducible*. In parallel branch-and-bound, it is well known that the *order* in which node computations are completed can have a significant impact on performance, and often

lead to anomalous behavior [22]. An (often) undesirable consequence of this is that users can run the same instance, with the same parameter settings, and achieve very different results in terms of nodes evaluated and CPU time. To combat this undesirable behavior, modern (shared-memory-based) MILP software has introduced appropriate synchronization points in the algorithm to ensure reproducible behavior in a parallel environment. Some overhead is introduced by these synchronization mechanisms.

MILP SOFTWARE

In this section we concentrate on MILP software by first presenting a historical perspective and briefly discussing the issue of the user interface. Then, we review the main characteristics of both commercial and non-commercial MILP software packages.

Historical Perspective

The earliest commercial grade implementations of branch-and-bound algorithms for solving MILP were done by Beale and Small [23]. For a historical perspective of the development and features of early MILP software packages, such as UMPIRE, SCICONIC, and MPSX/370, the reader is referred to the article of Forrest and Tomlin [24]. These early MILP software packages developed many effective branching and node selection techniques, some of which are still in use today. A major step forward in MILP solution technology occurred in the works of Crowder *et al.* [10] and Van Roy and Wolsey [11]. These papers introduced many of the preprocessing- and structure-specific cutting-plane techniques in use today. Another important work from a computational perspective was by Balas *et al.* [25], which demonstrated that the inequalities of Gomory [26] could be effectively used as cutting planes in a branch-and-cut procedure.

Modern MILP codes synthesized many of the ideas present in the literature, taking the best points of these works and engineering them into commercial-grade software packages. The article of Bixby and Rothberg [27] offers a nice historical exposition

of how MILP software improved by orders of magnitude over a period of a decade, starting in the late 1990s.

User Interfaces

An important aspect of the design of software for solving MILPs is the user interface. The range of purposes for MILP software is quite large, so it stands to reason that the number of user interface types is also large. In general, users may wish to solve MILPs using the solver as a “black box,” they may wish to call the software from a third party package, or they may wish to embed the solver into custom applications, which would require software to have a callable library. Often, the user may wish to adapt certain aspects of the algorithm. For instance, the user may wish to provide a custom branching rule or problem-specific valid inequalities. The customization is generally accomplished either through the use of language callback functions or through an abstract interface wherein the user must derive certain base classes and override default implementations for the desired functions. Nearly all of the below surveyed software packages have stand-alone executables, callable libraries, and allow the user to customize the branch-and-bound algorithm to varying degrees.

MILP Software Packages

Any review of software features is inherently limited by the temporal nature of software itself. In fact, algorithmic advances in the field of computational integer programming, many previously described in this article, can quickly render obsolete MILP software that was once state of the art. Here, we give a brief overview of what are the most widely used MILP software packages at the time of publication. Hans Mittelmann has for many years run independent benchmarks of MILP software, and been publishing the results. In general, the commercial software significantly outperforms noncommercial software on the instances in the Mittelmann suite, but it is difficult to draw strong conclusions about the relative performance of different commercial systems. Results of these benchmarks can be found

at <http://plato.asu.edu/ftp/milpf.html> and <http://plato.asu.edu/ftp/milpc.html>. Many of these solvers are available for use free of charge on the NEOS server website <http://www-neos.mcs.anl.gov>.

Commercial Software Packages. Unsurprisingly, exact implementation details of specific commercial MILP software packages are not known, as revealing such details would be a competitive disadvantage. In this article, information from the various software user manuals was used to compile features specific to each package. Further, it is impossible to detail all features of a software package, so we focus on those features that are, in the opinion of the authors, particularly noteworthy. Many of the commercial software packages listed here have free or limited cost licensing options available for academic users.

CPLEX

Version: 12.2

Website: <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/>

Interfaces: C, C++, Java, .NET, Matlab, Python, Microsoft Excel

CPLEX is a commercial MILP software package owned and distributed by IBM. The branch-and-cut algorithm contains a full suite of presolving techniques, cutting planes, search strategies, and heuristic techniques for finding feasible solutions. A special search algorithm in CPLEX, called *dynamic search*, can be used instead of traditional branch-and-cut. CPLEX also includes a *mipemphasis* metaparameter that adjusts many algorithmic parameters simultaneously to help users achieve a high-level goal. CPLEX contains facilities for automatically tuning MILP algorithm parameters for a particular class of instances. CPLEX may parallelize the branch-and-bound search using multiple threads in either a deterministic or nondeterministic manner. A final notable feature of CPLEX is its ability to enumerate multiple solutions that are close to

optimality and store them in a *solution pool*.

Gurobi

Version: 3.0

Website: www.gurobi.com

Interfaces: C, C++, Java, Python, .NET, Matlab

Gurobi Optimizer contains a relatively new MILP solver that was built from scratch to exploit modern multicore processing technology. Gurobi contains all advanced algorithmic features, including a variety of cutting planes, heuristics, and search techniques. Notable features of Gurobi include an *MIPFocus* metaparameter that simultaneously controls many algorithmic parameters to achieve the desired goal. Gurobi may execute the branch-and-bound algorithm using multiple computational threads. The parallel search is done in a deterministic manner. The Gurobi interactive shell is implemented as a set of extensions to the Python shell. Gurobi is also available “on demand” using the Amazon Elastic Compute Cloud.

LINDO

Version: 6.1

Website: www.lindo.com

Interfaces: C, Visual Basic, Matlab, Ox

LINDO Systems offers an MILP solver as part of its LINDO API. The MILP solver offers 15 different types of cutting planes, 6 different node selection rules, and a variety of preprocessing techniques and branching rules.

Mosek

Version: 6.0

Website: www.mosek.com

Interfaces: C, C++, Java, .NET, Python

MOSEK ApS is a company specializing in generic mathematical optimization software, and their suite of software includes a solver for MILP. The solver has 6 node selection rules, 11 types of cutting planes, the feasibility pump, and other heuristics, as well

as advanced branching methodologies such as strong branching. Mosek is available for use, through a GAMS interface, on the NEOS server.

XPRESS-MP

Version: 7.0

Version: <http://www.fico.com/en/Products/DMTools/xpress-overview/Pages/Xpress-Optimizer.aspx>

Interfaces: C, C++, Java, .NET, VBA

FICO Xpress Optimization Suite 7 offers a branch-and-cut-based software for solving MILP. XPRESS-MP offers many advanced cutting planes, including an effective implementation of lift-and-project cuts [28]. Modern preprocessing, heuristic, and branching techniques are also available. A unique feature of XPRESS-MP is that it offers an option to branch into general (split) disjunctions, or to search for special structures on which to branch. XPRESS-MP has features to enumerate multiple feasible solutions, and also to tune algorithmic parameters. XPRESS-MP is available for use, with three different input formats, on the NEOS server.

Noncommercial MILP Packages. Here, we give a cursory description of the most popular noncommercial MILP software packages. The paper of Linderoth and Ralphs [29] contains a more in-depth treatment of noncommercial software packages.

BLIS

License: Common Public License

Version: 0.91

Website: <https://projects.coin-or.org/CHiPPS>

Language: C++

BLIS is an open-source MILP solver available as part of the COIN-OR project. It is built on top of the COIN-OR high-performance parallel search (CHiPPS) framework, so it has been designed to run on distributed-memory platforms (one of the few available MILP software packages with this feature). Linear programs are solved using the COIN-OR

linear programming (Clp) solver, and cutting planes from the COIN cut generation library (CGL) are used.

CBC

License: Common Public License

Version: 2.5

Website: <https://projects.coin-or.org/Cbc>

Language: C++

CBC is an open-source MILP solver distributed under the COIN-OR project. It is built from many COIN components, including Clp and CGL. It is available both as a library and a stand-alone solver. CBC may be parallelized using shared-memory parallelism, and there are a large number of examples demonstrating its use. CBC is available for use on the NEOS server.

GLPK

License: GNU General Public License (GPL)

Version: 4.44

Website: <http://www.gnu.org/software/glpk/>

Language: C

GLPK is the GNU LP kit, a set of subroutines comprising a callable library and black box solver for MILP. The software distinguishes itself by being available under the GNU GPL and through the large number of community-built interfaces available for its use. GLPK is available for use, though a GAMS interface, on the NEOS server.

lp_solve

License: GNU lesser general public license (LGPL)

Version: 5.5

Website: <http://lpsolve.sourceforge.net/5.5/>

Language: C

The software lp_solve is an open-source linear and integer programming solver. There are a large number of interfaces available through which lp_solve can be used, including Java, AMPL, MATLAB, O-Matrix,

Sysquake, Scilab, Octave, Freemat, Euler, Python, Sage, PHP, and R.

MINTO

License: Given as library only

Version: 3.1

Website: <http://coral.ie.lehigh.edu/minto/>

Language: C

MINTO (Mixed INTeget Optimizer) is a black box solver and solver framework for MILP. MINTO is available only in library form. MINTO relies on external software to solve the LP relaxations that arise during the algorithm. Primary development of the software was done in the 1990s. The code was noteworthy at that time for a dynamic preprocessing engine [30] and effective lifted cover inequalities [31]. MINTO is available for use, through an AMPL interface, on the NEOS server.

SCIP

License: ZIB Academic License

Version: 1.2

Website: <http://scip.zib.de/>

Language: C

SCIP is an MILP software package developed and distributed by a team of researchers at Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB). SCIP is also a framework for constraint integer programming and branch-cut-and-price, allowing the user significant control of the algorithm. Development of SCIP was awarded the Beale Orchard Hayes Prize in 2009 [32]. Current benchmarks indicate that SCIP is likely the fastest noncommercial MILP solver. Other references for SCIP include the work of Achterberg [33] and Achterberg *et al.* [34]. SCIP is available for use, through a variety of interfaces, on the NEOS server.

SYMPHONY

License: Common Public License

Version: 5.2

Website: <http://www.coin-or.org/SYMPHONY/index.htm>

Language: C

SYMPHONY is a black box solver and callable library for MILPs. The core solution methodology of SYMPHONY is a customizable branch, cut, and price algorithm that can be executed sequentially or in parallel [35]. SYMPHONY calls on several other open-source libraries for specific functionality, including COIN-OR's CGL and Clp. There are several unique features of SYMPHONY that are worth mentioning. First, SYMPHONY contains implementation of an algorithm for solving bicriteria MILPs and methods for approximating the set of Pareto outcomes [36]. SYMPHONY also contains functions for local sensitivity analysis [37]. Second, SYMPHONY has the capability to warm start the branch-and-bound process from a previously calculated branch-and-bound tree, even after modifying the problem data. The work by Ralphs and Guzelsoy [38] gives an overview of the SYMPHONY callable library. SYMPHONY is available for use via MPS files on the NEOS server.

CONCLUSIONS

As decision processes become increasingly complex, the need for formal tools to aid in both planning and operation of the underlying system becomes more important. Many disciplines have embraced the use of optimization as a central tool to enhance the quality of decisions made. A large fraction of the models used in commercial settings are MILPs. Powerful, reliable, and flexible software systems (both commercial and noncommercial) are a key component of this broad-based acceptance. The field of MILP and computational aspects therein remain an active area of research. Given the high commercial penetration of MILP planning models, undoubtedly future research advances will find themselves implemented in software.

REFERENCES

1. Padberg MW, Rinaldi G. A branch and cut algorithm for the resolution of large-scale symmetric traveling salesmen problems. *SIAM Rev* 1991;33:60–100.

2. Land AH, Doig AG. An automatic method of solving discrete programming problems. *Econometrica* 1960;28:497–520.
3. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64:275–278.
4. Achterberg T. Constraint integer programming [PhD thesis]. Berlin: ZIB; 2007.
5. Lodi A. MIP computation. In: Jünger M, Liebling TM, Naddef D, Nemhauser GL, Pulleyblank WR, Reinelt G, Rinaldi G, Wolsey LA, editors. 50 years of integer programming 1958-2008. Heidelberg: Springer; 2009. pp. 619–645.
6. Margot F. Symmetry in integer linear programming. In: Jünger M, Liebling TM, Naddef D, Nemhauser GL, Pulleyblank WR, Reinelt G, Rinaldi G, Wolsey LA, editors. 50 years of integer programming 1958-2008. Heidelberg: Springer; 2009. pp. 647–686.
7. Savelsbergh MPW. Preprocessing and probing techniques for mixed integer programming problems. *ORSA J Comput* 1994;6:445–454.
8. Martin A. General mixed integer programming: computational issues for branch-and-cut algorithms. In: Jünger M, Naddef D, editors. Computational combinatorial optimization. Volume 2241, Lecture notes in computer science. Berlin/Heidelberg: Springer; 2001. pp. 1–25.
9. Cornuéjols G. Valid inequalities for mixed integer linear programs. *Math Program* 2008;112:3–44.
10. Crowder H, Johnson E, Padberg MW. Solving large scale zero-one linear programming problem. *Oper Res* 1983;31:803–834.
11. Van Roy TJ, Wolsey LA. Solving mixed integer programming problems using automatic reformulation. *Oper Res* 1987;35:45–57.
12. Achterberg T, Koch T, Martin A. Branching roles revisited. *Oper Res Lett* 2005;33:42–54.
13. Linderoth JT, Savelsbergh MWP. A computational study of search strategies for mixed integer programming. *INFORMS J Comput* 1999;11:173–187.
14. Benichou M, Gauthier JM, Girodet P, *et al.* Experiments in mixed-integer programming. *Math Program* 1971;1:76–94.
15. Fischetti M, Glover F, Lodi A. The feasibility pump. *Math Program* 2005;104:91–104.
16. Fischetti M, Lodi A, Salvagnin D. Just MIP it! In: Maniezzo V, Stützle T, Voss S, editors. MATHEURISTICS: hybridizing metaheuristics and mathematical programming. Operations research/computer science interfaces series. Heidelberg: Springer; 2009. pp. 39–70.
17. Fischetti M, Lodi A. Local branching. *Math Program* 2002;98:23–47.
18. Danna E, Rothberg E, Le Pape C. Exploiting relaxation induced neighborhoods to improve MIP solutions. *Math Program* 2005;102:71–90.
19. Gendron B, Crainic TG. Parallel branch and bound algorithms: survey and synthesis. *Oper Res* 1994;42:1042–1066.
20. Eckstein J. Parallel branch-and-bound algorithms for general mixed integer programming on the CM-5. *SIAM J Optim* 1994;4:794–814.
21. Eckstein J. Parallel branch-and-bound methods for mixed integer programming. *SIAM News* 1994;27:12–15.
22. Lai TH, Sahni S. Anomalies in parallel branch and bound algorithms. *Proceedings of the 1983 International Conference on Parallel Processing*; Columbus (OH). 1983. pp. 183–190.
23. Beale EML, Small RE. Mixed integer programming by a branch and bound method. In: Kalenich WH, editor. Volume 2, *Proceedings IFIP Congress 65*. Washington (DC): Spartan press and London: Macmillan; 1965. pp. 450–451.
24. Forrest JJH, Tomlin JA. Branch and bound, integer and non-integer programming. *Ann Oper Res* 2007;149:81–87.
25. Balas E, Ceria S, Cornuéjols G, *et al.* Gomory cuts revisited. *Oper Res Lett* 1996;19:1–9.
26. Gomory RE. An algorithm for the mixed integer problem. Technical Report RM-2597. The Rand Corporation; 1960.
27. Bixby R, Rothberg E. Progress in computational mixed integer programming – a look back from the other side of the tipping point. *Ann Oper Res* 2007;149:37–41.
28. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0-1 programming. *Math Program* 2003;94:221–245.
29. Linderoth JT, Ralphs TK. Noncommercial software for mixed-integer linear programming. In: Karlof JK, editor. *Integer programming: theory and practice*. Operations research series. CRC Press; 2005. pp. 253–303.
30. Atamtürk A, Nemhauser G, Savelsbergh MWP. Conflict graphs in solving integer

- programming problems. *Eur J Oper Res* 2000;121:40–55.
31. Gu Z, Nemhauser GL, Savelsbergh MWP. Cover inequalities for 0-1 linear programs: computation. *INFORMS J Comput* 1998;10:427–437.
 32. Achterberg T. SCIP: solving constraint integer programs. *Math Program Comput* 2009;1(1): 1–41.
 33. Achterberg T. Constraint integer programming [PhD thesis]. Berlin: Technischen Universität Berlin; 2007.
 34. Achterberg T, Berthold T, Koch T, *et al.* Constraint integer programming: a new approach to integrate CP and MIP. In: Perron L, Trick MA, editors. *Integration of AI and OR techniques in constraint programming for combinatorial optimization problems*. Volume 5015, *Lecture notes in computer science*. Heidelberg: Springer; 2008. pp. 6–20.
 35. Ralphs TK, Ladányi L, Saltzman MJ. Parallel branch, cut, and price for large-scale discrete optimization. *Math Program* 2003;98:253–280.
 36. Ralphs T, Saltzman M, Wiecek M. An improved algorithm for biobjective integer programming. *Ann Oper Res* 2006;147:43–70.
 37. Schrage L, Wolsey LA. Sensitivity analysis for branch and bound linear programming. *Oper Res* 1985;33:1008–1023.
 38. Ralphs TK, Guzelsoy M. The SYMPHONY callable library for mixed integer programming. *The Proceedings of the Ninth Conference of the INFORMS Computing Society*; Annapolis (MD). 2005. pp. 61–71.

MINIMUM COST FLOWS

BALACHANDRAN VAIDYANATHAN
Operations Research, FedEx Express,
Memphis, Tennessee

RAVINDRA K. AHUJA
Innovative Scheduling, Inc., and
Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

INTRODUCTION

Network flows is a problem domain that lies on the cusp between several fields of inquiry, including applied mathematics, computer science, engineering, management, and operations research [1]. The minimum cost flow problem is a fundamental network flow problem. The objective of this problem is to send units of a good that reside at one or more points in a network (sources or supplies) with arc capacities, to one or more other points in the network (sinks or demands), while incurring minimum cost. All other network flow problems such as the shortest path problem, maximum flow problem, assignment problem, and transportation problem can be viewed as special cases of the minimum cost flow problem. In this article, we define the minimum cost flow problem, outline some of its applications, and describe three algorithmic approaches for solving it, namely (i) the cycle canceling algorithm; (ii) the successive shortest path algorithm; and (iii) the network simplex algorithm. These algorithms are iterative improvement algorithms that start with a flow and improve it until optimality conditions are met.

Let $G = (N, A)$ be a directed network with a cost per unit flow c_{ij} and a capacity u_{ij} associated with every arc $(i, j) \in A$. For each node $i \in N$, $b(i)$ indicates its supply or demand depending upon whether $b(i) > 0$ (supply) or $b(i) < 0$ (demand). Let decision variable x_{ij} denote the flow on arc $(i, j) \in A$.

Then, the minimum cost flow problem is mathematically formulated below.

$$\begin{aligned} \text{Minimize } z = & \sum_{(i,j) \in A} c_{ij} x_{ij} \\ & \sum_{\{j:(i,j) \in A\}} x_{ij} - \sum_{\{j:(j,i) \in A\}} x_{ji} = b(i) \\ & \text{for all } i \in N \\ & 0 \leq x_{ij} \leq u_{ij}, \text{ for all } (i,j) \in A. \end{aligned}$$

The first set of constraints ensures the balance of flows at each node and is called the *flow balance constraint*. The second set of constraints establishes lower and upper bounds on the flow on each arc and is called the *flow bound constraint*. Let C denote the largest magnitude of any arc cost and U denote the largest magnitude of any supply/demand or finite arc capacities. We assume that the lower bounds on arc flows are all zero. Let n denote the number of nodes and m the number of arcs. We assume that all data (cost, supply/demand, and capacity) is integral, the network is directed, $\sum_{i \in N} b(i) = 0$, and all arc costs are nonnegative. All these assumptions can be made without loss of generality, because a minimum cost flow problem which violates any of these conditions can be transformed to the standard form [1]. Our algorithms rely on the concept of residual networks. The residual network $G(x)$ corresponding to a flow x is defined as follows. We replace each arc $(i, j) \in A$ by two arcs (i, j) and (j, i) . The arc (i, j) has cost c_{ij} and residual capacity $r_{ij} = u_{ij} - x_{ij}$, and the arc (j, i) has cost $c_{ji} = -c_{ij}$ and residual capacity $r_{ji} = x_{ij}$. $G(x)$ only contains arcs with positive residual capacity.

The minimum cost flow problem has been extensively studied by researchers. Jewell [2], Iri [3], and Busaker and Gowen [4] independently developed the successive shortest path algorithm. Their algorithm solves the minimum cost flow problem as a sequence of shortest path problems with arbitrary arc lengths. Subsequently, Tomizava [5] and Edmonds and Karp [6] observed that if the

computations use node potentials, then it is possible to implement these algorithms so that the shortest path problems have nonnegative arc lengths; this makes the successive shortest path algorithm more efficient. Ford and Fulkerson [7] developed the primal-dual algorithm, and Klein [8] developed the cycle canceling algorithm for the minimum cost flow problem. Three polynomial-time implementations of the cycle canceling algorithm are due to (i) Ref. 9, which modifies an algorithm by Weintraub [10] and augments flow along (negative) cycles with the maximum possible improvement; (ii) Ref. 11, which augments flow along minimum mean cost (negative) cycles; and (iii) Ref. 12, which augments flow along minimum ratio cycles.

The minimum cost flow problem can also be viewed as a special case of the linear programming problem, and therefore the simplex algorithm for linear programming can be used to solve it. The simplex algorithm for the linear programming problem was specialized to develop the network simplex algorithm for the minimum cost flow problem. Efficient manipulation of spanning trees is crucial to the efficacy of the network simplex algorithm. Johnson [13] suggested the first tree-manipulating data structure for implementing the network simplex algorithm for the minimum cost flow problem. Degeneracy is another problem that was encountered while using the network simplex algorithm. Experience with solving minimum cost flow problems established that for certain classes of problems, more than 90% of the pivots in the network simplex algorithm can be degenerate. Cunningham [14], and Barr *et al.* [15,16] proposed the strongly feasible spanning-tree basis that led to improvements on this front. Computational experiences showed that maintaining a strongly feasible spanning-tree basis substantially reduces the number of degenerate pivots. On the theoretical front, Orlin [17] showed that for integer data, an implementation of the network simplex algorithm that maintains a strongly feasible basis performs $O(nmCU)$ pivots when used with any arbitrary pricing strategy, and $O(nmC \log(mCU))$ pivots when

used with Dantzig's pricing strategy. A polynomial-time implementation remained elusive to researchers until Orlin [18] developed an $O(\min(n^2m \log nC, n^2m^2 \log n))$ time algorithm.

The first strongly polynomial-time minimum cost flow algorithm is attributed to Tardos [19]. Subsequently, several researchers [11,20–26] developed other strongly polynomial-time algorithms. Currently, the fastest strongly polynomial-time algorithm is due to Orlin [24]. His enhanced capacity scaling algorithm solves the minimum cost flow problem as a sequence of $O(\min(m \log U, m \log n))$ shortest path problems. The best known theoretical time bound for the minimum cost flow problem is $O(\min\{nm \log(n^2/m \log(nC)), nm(\log \log U) \log(nC), (m \log n)(m + n \log n)\})$; the three bounds in this expression are due to Goldberg and Tarjan [23], Ahuja *et al.* [27], and Orlin [24], respectively.

The rest of this article is organized as follows. In the section titled “Applications,” we identify some applications of the minimum cost flow problem; in the subsequent section, we discuss the optimality conditions; and finally in the section titled “Minimum Cost Flow Algorithms,” we describe algorithms to solve the minimum cost flow problem.

APPLICATIONS

Minimum cost flow problems have been extensively applied to solve problems in diverse areas such as transportation, communications, manufacturing, defense, finance, and health care. In this section we describe some of these applications in four different areas and we refer the interested reader to the book by Ahuja *et al.* [1] for a comprehensive review of minimum cost flow applications.

Production Planning Application

While planning for production at a plant, it is preferable to produce in periods when the production costs are lower. On the other hand, this may involve carrying inventory and incurring inventory storage costs. The production planning (or lot-sizing) problem

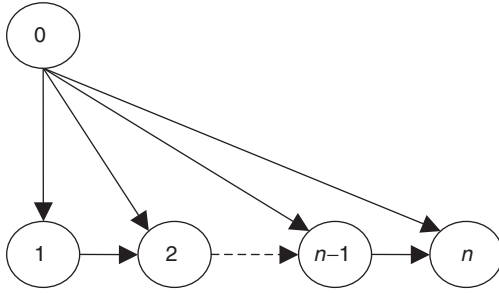


Figure 1. Production planning application.

determines the optimal production schedule by balancing all these costs. The production planning problem with linear production and inventory carrying costs can be formulated as a minimum cost flow problem.

Let the demand of a single item be $d(i)$ in each of n periods $i = 1, 2, \dots, n$. The demand in period i is fulfilled by either producing x_{0i} in that period and/or by drawing upon inventory $x_{i-1,i}$ from the previous period. Figure 1 shows the minimum cost network for a problem with four periods. Node 0 denotes the source node and has an associated supply of $\sum_{i=1}^n d(i)$, and all other nodes i represent period i and have an associated demand $d(i)$. Node 0 is connected to each demand node i using *production arcs* $(0, i)$; flow on production arcs represent the quantity produced in period i . Each arc $(i-1, i)$ for $i = 2, 3, \dots, n$ is an *inventory arc*; flow on inventory arcs represent the inventory carried from one period to the next. Let c_{ij} be the cost per unit flow on arc (i, j) , and u_{ij} be its capacity. Then, the cost c_{0i} represents the unit cost of production in period i and $c_{i-1,i}$ represents the unit holding cost from period $i-1$ to period i . The capacity u_{0i} represents the production capacity in period i and $u_{i-1,i}$ denotes the inventory carrying capacity from period $i-1$ to i . It is easy to see that the constructed minimum cost flow problem can be used to solve the linear cost dynamic lot-sizing problem.

Teleprocessing Network Design Application

Centralized teleprocessing networks contain several geographically dispersed terminals that need to communicate with a central processing unit (CPU). Each terminal can be connected to the CPU either directly or

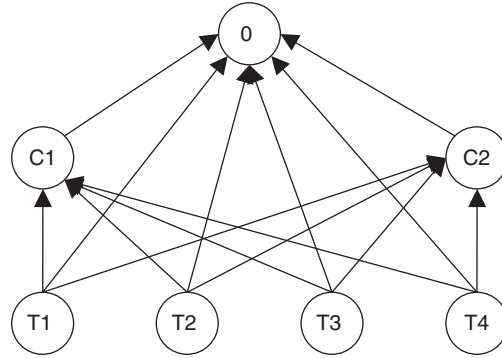


Figure 2. Teleprocessing network application.

through a concentrator. A concentrator is capable of merging the data flows from several terminals and sending them to the CPU, and is linked to the CPU by a dedicated high-speed line. Each terminal connection to a concentrator or the CPU has an associated cost of construction. Given the capacity of each concentrator, the teleprocessing network design problem is to determine the minimum cost network configuration that connects each terminal to the CPU.

Let the set of concentrators in the system be C and the set of terminals be T . Each concentrator can serve at most K terminals. Let c_{j0} be the cost of directly connecting terminal j to the CPU and c_{ji} be the cost of connecting terminal j to concentrator i . The minimum cost flow network is constructed as follows. We construct node 0 to represent the CPU. For each concentrator i , we construct a node in the network. We also construct one node for each terminal j . Each terminal j is connected to each concentrator i using arcs (j, i) with cost c_{ji} . Each terminal j is also connected directly to the CPU using arc $(j, 0)$ with cost c_{j0} . Each concentrator i is connected to the CPU using arc $(i, 0)$ with capacity K and zero cost. All the terminals in the network are assigned a supply of unit, and the CPU is assigned a demand of $|T|$ units. Figure 2 shows the minimum cost flow network for an example with four terminals (T1, T2, T3, T4), two concentrators (C1, C2), and one CPU. By solving the constructed minimum cost flow problem, we can identify the minimum cost set of links to construct or the optimal network configuration.

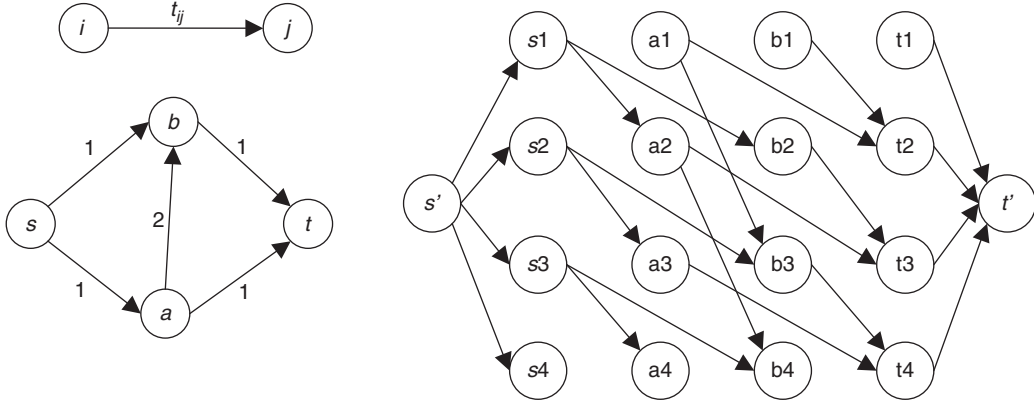


Figure 3. Building evacuation application.

Building Evacuation Application

One of the important considerations while designing a building is the time taken to evacuate in an emergency situation like an earthquake, fire, or bomb threat. Architects typically need to develop an evacuation plan and assess the evacuation time for a particular design. We show how this problem can be modeled as a minimum cost flow problem.

The evacuation network for a building is constructed as follows. The nodes represent locations in the building such as offices, hallways, elevators, staircases, and exits. The arcs represent connections between two locations. The locations in the building that house a significant number of people are the source nodes in the network and the exits are the sinks. The supply of each node is an estimate of the number of people that are likely to be at that location. Each arc has a capacity that represents the maximum number of people that can simultaneously use it at any point in time. We can transform the problem into one with a single source and single sink by using the standard technique of constructing a super-source and a super-sink. In order to correctly account for the travel time on each arc, we consider multiple copies of each node at various points in time using the concept of *time-expanded replica* of a network. For each node i in the original network, we construct p copies $i_1, i_2, \dots, i_p$ representing the node at various times. We also construct arcs (i_k, j_l) to represent a passage that connects location

i and location j if $l - k$ is equal to the travel time on passage (i, j) . This arc represents the potential movement of a person from node i to node j in time $l - k$. The minimum time required to evacuate the building is the smallest index r satisfying the property that the maximum flow from source nodes $s_1, s_2, \dots, s_r$ to the sink nodes $t_1, t_2, \dots, t_r$ is at least a prespecified number of people or the buildings capacity. In Fig. 3, we illustrate the underlying network for a small instance with four locations, travel times as indicated in the figure, and then time expanded over four periods. This problem can be solved using a minimum cost flow algorithm. We refer the interested reader to Ref. 28 for more details.

Distribution Application

Consider a company that needs to manufacture and ship a set of different product models from plants to retailers. Each plant p has an associated capacity of production for each model and each retailer r has a known demand for each model. The company wants to determine how many products of each type should be produced at each plant and the shipping pattern between plants and retailers that minimizes the overall cost of production and transportation.

This problem can be formulated as a minimum cost flow problem. The network contains four types of nodes: (i) plant nodes representing the various plants; (ii) plant-model nodes representing a model made at a plant; (iii) retailer-model nodes representing

the products required by each retailer; (iv) retailer nodes corresponding to the various retailers. The network contains three types of arcs: (i) production arcs connect the plant nodes to the plant-model nodes and the cost of flow on this arc is unit cost of production of the associated model at the plant. The production arcs have an upper bound that represents the capacity of production of the associated model at the plant; (ii) Transportation arcs connect plant-model nodes to retailer-model nodes. The cost of the arc is the total cost of shipping one unit between the associated plant and retailer. (iii) Demand arcs connect the retailer-model nodes and the retailer nodes. The demand arcs have zero cost and a lower bound that represents the demand of the product at that retailer.

The minimum cost flow network is completed by (i) creating a production super-node and connecting it to the plant nodes with arcs of zero cost and infinite capacity; (ii) creating a retailer super-node and connecting the retailer nodes to it with arcs of zero cost and infinite capacity; (iii) creating a feedback arc to connect the retailer super-node to the production super-node. These augmentations ensure that the constructed problem is a minimum cost flow problem in its standard form. In Fig. 4, we give the network for a problem with two plants, two models, and two retailers. Note that the supply and demands of all nodes in the network are zero. Such a problem is a special case of a minimum cost flow problem and is called a *minimum cost circulation problem*. It can be solved using any standard minimum cost flow algorithm.

OPTIMALITY CONDITIONS

Optimality conditions provide us with a simple method to validate whether a feasible solution to a problem is optimal. They also motivate algorithms to solve the underlying optimization problem. Algorithms for the minimum cost flow problem typically start with a flow (or pseudoflow) and iteratively improve it until the optimality conditions for the minimum cost flow is met. In this section, we describe two different yet equivalent optimality conditions for the minimum cost flow problem: (i) negative cycle optimality

conditions; and (ii) reduced cost optimality conditions. In the next section, we describe algorithms which are based on the optimality conditions.

Negative Cycle Optimality Condition

The negative cycle optimality condition is very intuitive in nature and easy to understand. The negative cycle optimality condition states that *a feasible solution to the minimum cost flow problem is optimal if and only if the associated residual network does not contain a negative cost directed cycle*. The if direction is easily established since if the residual network contains a negative cost cycle, then by sending a positive flow on the cycle we obtain a feasible solution with a smaller cost. We refer the interested reader to Ref. 1, p. 307 for a complete proof.

Reduced Cost Optimality Condition

The reduced cost optimality conditions rely on the concept of node potentials. Let $\pi(i)$ be a real number associated with each node $i \in N$. We call $\pi(i)$ the potential of node i . Using these node potentials, we define the reduced cost of arc (i, j) as $c_{ij}(\pi) = c_{ij} - \pi(i) + \pi(j)$. Due to the way in which the reduced costs are defined, the following properties are immediate: (i) for any directed path P between node k and l , $\sum_{(i,j) \in P} c_{ij}(\pi) = \sum_{(i,j) \in P} c_{ij} - \pi(k) + \pi(l)$; (ii) for any directed cycle W , $\sum_{(i,j) \in W} c_{ij}(\pi) = \sum_{(i,j) \in W} c_{ij}$. This property implies that using the reduced costs instead of the actual costs does not change the shortest path between any pair of nodes. Also, if W is a negative cycle with respect to c_{ij} as arc costs, then it is also a negative cycle with respect to $c_{ij}(\pi)$.

The reduced cost optimality conditions are equivalent to the negative cycle optimality conditions. The reduced cost optimality conditions state that *a feasible solution to the minimum cost flow problem is optimal if and only if for some set of node potentials, the reduced cost of all arcs in the residual network are nonnegative*. The proof for correctness is straightforward. Suppose a solution satisfies the reduced cost optimality conditions, then using property (ii) the residual network does not contain a negative cost cycle; hence, the solution satisfies the negative

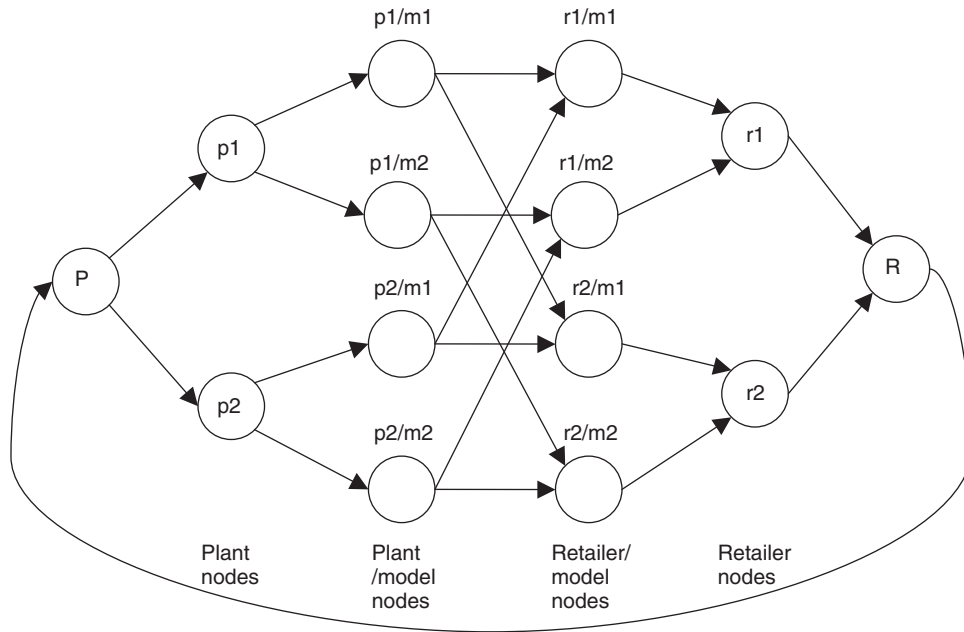


Figure 4. A distribution application.

cycle optimality condition and is optimal. Conversely, suppose a solution is optimal, then the residual network does not have negative cost cycles. Accordingly, the shortest path distances d from node 1 to all other nodes are well defined. It can be shown that by setting $\pi = -d$, we obtain the desired set of node potentials. We refer to the set of node potentials that satisfy the conditions $c_{ij}(\pi) \geq 0$ for all (i,j) in $G(x)$ as the *optimal node potentials*.

MINIMUM COST FLOW ALGORITHMS

In this section, we discuss three fundamental algorithms to solve the minimum cost flow

problem. All these algorithms start with an initial solution and perform a set of operations repeated until flow feasibility and/or optimality conditions are satisfied.

Cycle Canceling Algorithm

The cycle canceling algorithm is a direct consequence of the negative cycle optimality conditions. This algorithm starts with a feasible flow x and at every iteration it improves the current feasible solution by augmenting flow on a negative cost cycle in $G(x)$. The algorithm terminates when it cannot find any negative cost cycle in $G(x)$. We now formally define the cycle canceling algorithm.

```

algorithm cycle canceling;
begin
  establish a feasible flow  $x$  in the network;
  while  $G(x)$  contains a negative cycle do
    begin
      identify a negative cycle  $W$ ;
      compute  $f = \min\{r_{ij} : (i,j) \in W\}$ ;
      augment  $f$  units of flow in the cycle  $W$  and update  $G(x)$ ;
    end
  end

```

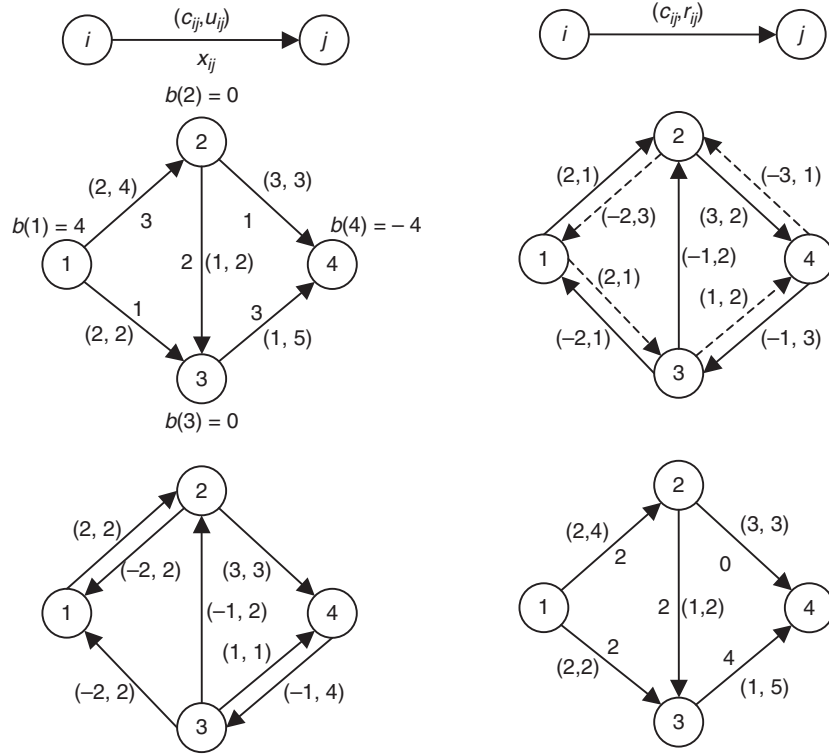


Figure 5. Illustration of the cycle canceling algorithm: (a) initial solution with $z = 16$; (b) negative cost cycle 1-3-4-2-1 in $G(x)$; (c) $G(x)$ after flow augmentation of one unit (contains no negative cost cycle); and (d) optimal solution with $z = 14$.

An interesting corollary of the cycle canceling algorithm is that it proves that when the arc capacities and supply/demands are integers, then the minimum cost flow problem always has an integer optimal flow. Our description of the algorithm is generic in the sense that when there are multiple negative cost cycles, we do not specify the criteria to choose one. Different rules for selecting negative cycles produce different versions of the algorithm, each with different running times.

Consider the running time of the generic version of the cycle canceling algorithm. An initial feasible flow x can be determined by solving a maximum flow problem. Also, each negative cost cycle can be identified in $O(nm)$ time using the label-correcting algorithm for the shortest path problem. Note that mCU is an upper bound on the cost of the initial feasible solution. Also, mCU is a lower bound

on the optimal solution cost. Since each iteration reduces the solution cost by at least one unit (all data is integral), the algorithm runs through $O(mCU)$ iterations. Hence, the algorithm runs in $O(nm^2CU)$ time.

In Fig. 5, we give an example for the cycle canceling algorithm. We start with a feasible solution in Fig. 5(a) with objective function $z = 16$. Next, we identify the residual network corresponding to this solution in Fig. 5(b). This residual network has a negative cost cycle 1-3-4-2-1 with cost -2 (indicated in dotted lines). We augment one unit of flow on this cycle to obtain a better solution with $z = 14$. Next, we update the residual network as shown in Fig. 5(c). The residual network no longer contains a negative cost cycle and the algorithm accordingly terminates. We use the residual network in Fig. 5(c) to obtain the actual optimal flows in Fig. 5(d).

Successive Shortest Path Algorithm

The successive shortest path algorithm is different from the cycle canceling algorithm in the sense that the algorithm maintains a solution that satisfies the optimality conditions and strives for feasibility (the cycle canceling algorithm maintains feasibility and strives for optimality). The algorithm always maintains a solution x that satisfies the flow bound constraints, but may violate the flow balance constraints. We call such a solution a *pseudoflow*. The algorithm starts with zero pseudoflow and repeatedly sends flow on successive shortest paths in $G(x)$ until all the demands have been met. It can be shown that by sending flows on successive shortest paths in the residual network, the algorithm

```

algorithm successive shortest path;
begin
  set  $x = 0$ ;
  while  $x$  is not a flow do
    begin
      select an excess node  $s$  and a deficit node  $t$ ;
      identify a shortest path  $P$  from node  $s$  to node  $t$  in  $G(x)$ ;
      augment maximum possible flow on  $P$  and update  $G(x)$ ;
    end
  end

```

We now consider the running time of the algorithm. Let U be an upper bound on the largest supply of a node. Since each iteration of the algorithm augments at least one unit of flow, the number of flow augmentations is $O(nU)$. Also, every iteration of the algorithm involves solving the shortest path problem; let this take $S(n, m, C)$ time. All other steps in the while loop can be performed in $O(n)$ time. Accordingly, the running time of the algorithm is $O(n U S(n, m, C))$. Note that during the algorithm, the cost of arcs in $G(x)$ can become negative which makes the shortest path problem more difficult to solve than one with nonnegative arc lengths. However, it is possible to use the node potentials to maintain nonnegative arc lengths, which allows us to solve the shortest path problems more efficiently. We refer the interested reader to Ref. 1 for further details.

In Fig. 6, we give an example for the successive shortest path algorithm. We start with a pseudoflow $x = 0$. The shortest path in

always maintains a pseudoflow that satisfies the negative cycle optimality condition (the residual network does not contain a negative cost cycle). Accordingly, whenever the pseudoflow satisfies all the flow balance conditions and becomes a feasible flow, feasibility and optimality are achieved and the algorithm terminates.

For any pseudoflow x , we define the *imbalance* of node i as $e(i) = b(i) - \sum_{j:(i,j) \in A} x_{ij} + \sum_{j:(j,i) \in A} x_{ji}$. If $e(i) > 0$, we refer to $e(i)$ as the *excess* of node i and if $e(i) < 0$, we call $e(i)$ the *deficit* of node i . The residual network $G(x)$ for a pseudoflow x is defined in a similar manner to the residual network for a flow. The successive shortest path algorithm is defined below.

$G(x)$ from the excess node to the deficit node is 1-3-4 with cost 3 as shown in Fig. 6(b). We augment 2 units of flow on 1-3-4 and update the residual network as shown in Fig. 6(c). The shortest path in $G(x)$ from the excess node to the deficit node is 1-2-3-4. We augment 2 units of flow on 1-2-3-4 to obtain the residual network shown in Fig. 6(d). At this point, the excess and deficit are zero and solution is a flow. Accordingly, the algorithm terminates.

Network Simplex Algorithm

The simplex algorithm for solving linear programming problems is perhaps the most powerful algorithm devised for solving constrained optimization problems. Due to the special constraint structure (unimodularity) of the minimum cost flow problem, all extreme point solutions are integral and the problem can be solved as a linear programming problem. However, using the

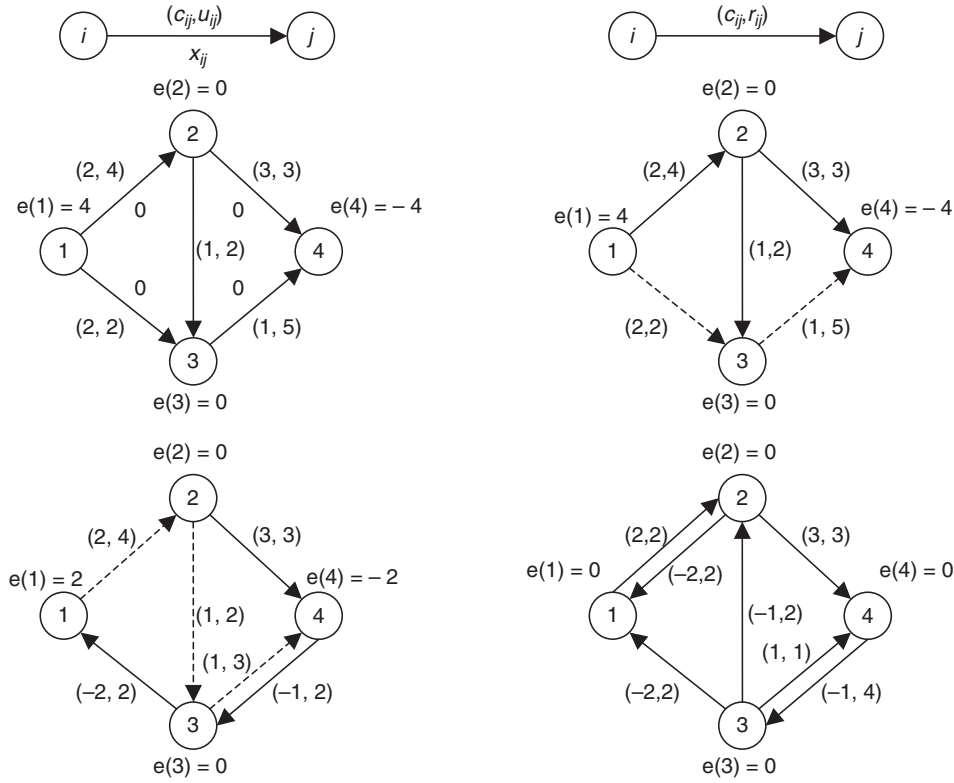


Figure 6. Illustration of the successive shortest path algorithm: (a) initial solution with $x = 0$, $z = 0$; (b) shortest path 1-3-4 in $G(x)$; (c) $G(x)$ after augmentation of 2 units on 1-3-4; and (d) $G(x)$ after augmentation of 2 units on 1-2-3-4.

generalized simplex method to solve the minimum cost flow problem does not lead to competitive running times when compared to previously discussed combinatorial approaches. In this section, we describe a specialized network adaptation of the simplex method that exploits the network property to solve the minimum cost flow problem efficiently. The network simplex algorithm allows us to perform all the steps in the simplex algorithm directly on the network.

The central concept in the *network simplex algorithm* is the notion of spanning-tree solutions.

In a spanning-tree solution, the flows on all arcs which are not in the spanning tree are fixed to either zero or the arc's flow capacity (*non-tree arcs*). Once this is done, the flow on all other arcs in the spanning tree can

be uniquely determined (*tree arcs*). It can be shown that the minimum cost flow problem has an optimal spanning-tree solution, and this optimal solution can be determined by moving from one feasible spanning-tree solution to the next by introducing one new non-tree arc into the spanning tree in the place of one tree arc. The successive spanning trees correspond to the basic feasible (or extreme point) solutions of linear programming, and moving from one spanning tree to the next is essentially a pivot operation in the simplex algorithm.

A spanning-tree solution partitions the arc set A into three subsets: (i) the arcs in the spanning tree, say T ; (ii) the non-tree arcs whose flow is zero, say L ; and (iii) the non-tree arcs whose flow is at the upper bound, say U . We refer to (T, L, U) as the *spanning-tree structure*. Given a

spanning-tree structure, we can determine the corresponding spanning-tree solution as follows. Set the flows on all arcs $(i, j) \in L$ to zero, the flow on all arcs $(i, j) \in U$ to u_{ij} , and then use the flow balance equations to determine the flow on all arcs in T . A spanning-tree structure is feasible if the corresponding spanning-tree solution satisfies all the arc flow bounds. If all the tree arcs in the spanning-tree solution are free arcs, then the solution is called *nondegenerate*; otherwise the solution is *degenerate*. The following theorem states a sufficient condition for a spanning-tree structure to be optimal. This theorem is analogous to the complementary slackness optimality conditions of linear programming.

A spanning-tree structure (T, L, U) is an optimal spanning-tree structure of the minimum cost flow problem if it is feasible and for some choice of node potentials π , the arc reduced costs $c_{ij}(\pi)$ satisfy the following conditions: (a) $c_{ij}(\pi) = 0$, for all $(i, j) \in T$; (b) $c_{ij}(\pi) \geq 0$, for all $(i, j) \in L$; and (c) $c_{ij}(\pi) \leq 0$ for all $(i, j) \in U$.

The network simplex algorithm maintains a feasible spanning-tree structure and moves from one spanning-tree structure to the next

until it finds the optimal spanning-tree structure. In each step, the algorithm adds one non-tree arc violating its optimality condition to the spanning tree in place of one of its current tree arcs. A non-tree arc violates its optimality condition if (i) its flow is at the lower bound and the reduced cost is negative; or (ii) its flow is at the upper bound and the reduced cost is positive. The algorithm (i) adds the arc to the spanning tree, creating a negative cycle; (ii) sends the maximum possible flow in this cycle until the flow on at least one arc in the cycle reaches its lower or upper bound; and (iii) drops an arc whose flow has reached its lower or upper bound from the tree, giving us a new spanning-tree structure. This operation of moving from one spanning-tree structure to another is known as a *pivot operation*, and the two spanning-tree structures obtained in consecutive iterations are called *adjacent spanning-tree structures*. The algorithm terminates when all non-tree arcs satisfy the optimality condition (the optimality of tree arcs is maintained throughout the algorithm). The algorithm is formally defined below.

algorithm *network simplex*;

begin

 determine an initial feasible tree structure (T, L, U) ;

 let x be the flow and π the node potentials associated with this tree structure;

while some non-tree arc violates the optimality conditions **do**

begin

 select an entering non-tree arc (k, l) violating its optimality condition;

 add arc (k, l) to the tree and determine the leaving arc (p, q) ;

 perform a tree update and update the solutions x and π ;

end;

end

We now describe the computation of node potentials π for a given spanning-tree structure (T, L, U) . Since the addition of a constant to all node potentials does not alter the reduced cost of any arc ($c_{ij}(\pi) = c_{ij} + \pi(i) - \pi(j)$), we assume without loss of generality that $\pi(1) = 0$. During each iteration of the algorithm, we compute node potentials that preserve the optimality condition for all tree arcs, $c_{ij}(\pi) = c_{ij} + \pi(i) - \pi(j) = 0$ for all $(i, j) \in T$. In order to achieve this, we start from node 1 and fan out such that whenever we

encounter a node k , we have already determined the potential of its predecessor. Arc flows are computed starting from a leaf node of the spanning-tree structure and moving toward the root while computing flows on arcs on the way.

In Fig. 7, we give an example for the network simplex algorithm. The non-tree arcs are dotted and the tree arcs are regular. We start with the initial solution shown in Fig. 7(a) and determine the node potentials and reduced costs. Non-tree arc $(1, 3)$ has flow

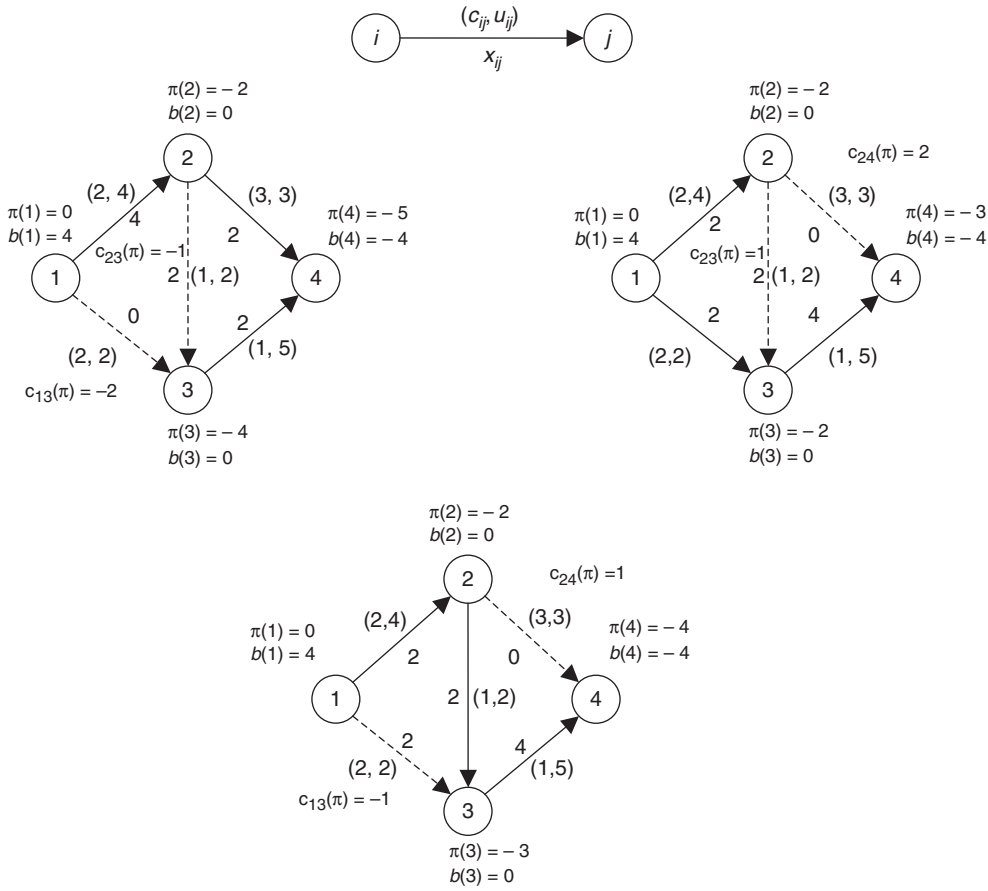


Figure 7. Illustration of the network simplex algorithm: (a) initial spanning-tree solution with $z = 18$; (b) spanning-tree solution with $z = 14$ obtained after (1,3) enters and (2,4) leaves; and (c) optimal spanning-tree solution with $z = 14$ obtained after (2,3) enters and (1,3) leaves.

at its lower bound and a negative reduced cost; we choose it as the entering arc. When (1, 3) enters, it creates a unique cycle 1-3-4-2-1; (1, 3) enters the tree and (2, 4) leaves the tree. The new solution and spanning-tree structure are shown in Fig. 7(b). We again compute the node potentials and reduced costs. Non-tree arc (2, 3) has flow at its upper bound and a positive reduced cost; we choose it as the entering arc. When (2, 3) enters, it creates the unique cycle 1-3-2-1; (2, 3) enters the tree and (1, 3) leaves the tree. The new solution and spanning-tree structure are shown in Fig. 7(c). We once again compute the node potentials and the reduced costs. Both non-tree arcs (1, 3) and

(2, 4) satisfy the optimality conditions and the algorithm terminates.

The network simplex algorithm moves from one spanning-tree structure to the next until it obtains a solution that satisfies the optimality conditions. If each pivot operation is nondegenerate, then since any network has a finite number of spanning-tree structures and every spanning-tree structure has a unique associated cost, the network simplex algorithm will encounter any spanning-tree structure at most once and will terminate finitely. However, degenerate pivots pose a difficulty due to *cycling*; that is, the algorithm goes through an infinite repetitive sequence of degenerate pivots. Computational studies

have shown that as many as 90% of the pivot operations in commonly encountered networks can be degenerate. Fortunately, it can be shown that by maintaining a special type of spanning tree called a *strongly feasible spanning tree*, the network simplex algorithm terminates finitely; moreover, it runs faster in practice as well.

REFERENCES

1. Ahuja RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms, and applications. Englewood Cliffs (NJ): Prentice Hall; 1993.
2. Jewell WS. Optimal flow through networks. Interim Technical Report No. 8. Cambridge (MA): Operations Research Center, MIT; 1958.
3. Iri M. A new method of solving transportation-network problems. J Oper Res Soc Jpn 1960; 3:27–87.
4. Busaker RG, Gowen PJ. A procedure for determining minimal-cost network flow patterns. O.R.O. Technical Report No. 15. Baltimore (MD): Operational Research Office, John Hopkins University; 1961.
5. Tomizawa N. On some techniques useful for solution of transportation network problems. Networks 1972;1:173–194.
6. Edmonds J, Karp RM. Theoretical improvements in algorithmic efficiency for network flow problems. J ACM 1972;19:248–264.
7. Ford LR, Fulkerson DR. Flows in networks. Princeton (NJ): Princeton University Press; 1962.
8. Klein M. A primal method for minimal cost flows. Manage Sci 1967;14:205–220.
9. Barahona F, Tardos E. Note on Weintraub's minimum cost circulation algorithm. SIAM J Comput 1989;18:579–583.
10. Weintraub A. A primal algorithm to solve network flow problems with convex costs. Manage Sci 1974;21:87–97.
11. Goldberg AV, Tarjan RE. Finding minimum-cost circulations by canceling negative cycles. In: Proceedings of the 20th ACM Symposium on the Theory of Computing. Chicago (IL). pp. 388–397. (Full paper in J ACM 1988;36:873–886).
12. Wallacher C, Zimmermann U. A combinatorial interior point method for network flow problems. Math Program 1992;56:321–335.
13. Johnson EL. Networks and basic solutions. Oper Res 1966;14:619–624.
14. Cunningham WH. A network simplex method. Math Program 1976;11:105–116.
15. Barr R, Glover F, Klingman D. The alternative path basis algorithm for the assignment problem. Math Program 1977;12:1–13.
16. Barr R, Glover F, Klingman D. Generalized alternating path algorithms for transportation problems. Eur J Oper Res 1978;2: 137–144.
17. Orlin JB. On the simplex algorithm for networks and generalized networks. Math Program Study 1985;24:166–178.
18. Orlin JB. A polynomial time primal simplex algorithm for minimum cost flows. Math Program 1997;78:109–129.
19. Tardos E. A strongly polynomial minimum cost circulation algorithm. Combinatorica 1985;5:247–255.
20. Orlin JB. Genuinely polynomial simplex and non-simplex algorithms for the minimum cost flow problem. Technical Report No. 1615-84. Cambridge (MA): Sloan School of Management, MIT; 1984.
21. Fujishige S. An $O(\underline{m}^3 \log n)$ capacity-rounding algorithm for the minimum cost circulation problem: a dual framework of Tardos' algorithm. Math Program 1986;35:298–309.
22. Galil Z, Tardos E. An $O(n^2(m + n \log n) \log n)$ min-cost flow algorithm. J ACM 1987;35: 374–386.
23. Goldberg AV, Tarjan RE. Solving minimum cost flow problem by successive approximation. Math Oper Res 1990;15:430–466.
24. Orlin JB. A faster strongly polynomial minimum cost flow algorithm. Oper Res 1993;41:377–387.
25. Ervolina TR, McCormick ST. Two strongly polynomial cut canceling algorithms for minimum cost network flow. Discrete Appl Math 1993;46:133–165.
26. Sockalingam PT, Ahuja RK, Orlin JB. New polynomial-time cycle-canceling algorithms for minimum-cost flows. Networks 2000; 36:53–63.
27. Ahuja RK, Goldberg AV, Orlin JB, *et al.* Finding minimum-cost flows by double scaling. Math Program 1992;53:243–266.
28. Chalmet LG, Francis RL, Saunders PB. Network models for building evacuation. Manag Sci 1982;28:86–105.

MINIMUM PREDICTION ERROR MODELS AND CAUSAL RELATIONS BETWEEN MULTIPLE TIME SERIES

JICONG ZHANG

PETROS XANTHOPOULOS

VERA TOMAINO

PANOS M. PARDALOS

Department of Industrial and Systems
Engineering and Biomedical
Engineering, University of Florida,
Center for Applied Optimization,
Gainesville, Florida

JUI-HONG CHIEN

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

INTRODUCTION

A prediction is a statement that a particular event will occur in the future, which can refer to the estimation of unknown situations in time series, cross-sectional, or longitudinal data. Prediction implies at least two factors: importance and difficulty [1]. Risk and uncertainty are crucial to prediction; thus it is important to minimize the prediction errors. There are no specifically assumed statistical distributions in the prediction models in this article.

In the first part of this article, we discuss two prediction models: the least square estimation (LSE) estimation model (section titled “LSE Model”), and the minimum mean square error (MMSE) model (section titled “MMSE Model Prediction”). Both of these two models aim at minimizing the square errors in their respective senses: (i) sum of squares, and (ii) mean square. LSE is known as a *linear deterministic model* that can be used to obtain approximate solutions of overdetermined systems. The application in the section titled “LSE Model” is to predict a limited-length time series $\{Z(i)\}$ from another limited-length time series $\{X(i)\}$, which can

be generalized to infinite-length time series case. In the section titled “Basic LSE Model”, by minimizing the sum of error squares in time series prediction, the *principle of orthogonality* is obtained as a necessary condition of reaching the global minimum error. It is further shown that the energy of the observed time series $\{Z(i)\}$ is the summation of the energy of the LSE estimation time series $\{Z_{\text{LSE estimation}}(i)\}$ and the energy of the LSE process time series $\{e(t)\}$. In the section titled “Another Approach to Understanding the LSE Model”, another way of understanding the LSE model is exhibited: a linear projection onto the column space of the data matrix $\mathbf{A}$, where the columns of $\mathbf{A}$ are composed of the time-delay vectors of the time series $\{X(i)\}$. In the section titled “MMSE Model Prediction”, we switch to a stochastic model: MMSE, which is also linear, to minimize the mean square error in prediction. We further show that the MMSE model is a special case of autoregressive moving average (ARMA)/(autoregressive integrated moving average) (ARIMA) models.

In the second part of this article (section titled “Causal Relations between Two, Three, and Multiple Variables in Prediction”), the causal relations in two or multiple time series are discussed in comparison, in statistical senses. In the section titled “Two-Variable (Time Series) Models” [2], we discuss (i) the simple causal model and (ii) the instantaneous causal model, between two time series $\{X_t\}$ and $\{Y_t\}$. We assume that the time series are wide-sense stationary and purely nondeterministic. In the simple causal model, the previous states of $\{X_t\}$ and $\{Y_t\}$ are used, to predict the present states of $\{X_t\}$ and $\{Y_t\}$, respectively. In the instantaneous causal model, the previous states of $\{X_t\}$, as well as the previous and present states of $\{Y_t\}$, are used to predict the present state of $\{X_t\}$; and vice versa. Though this is in stochastic sense, the *principle of orthogonality* (discussed in the section titled “LSE Model”) is still kept for minimizing prediction error. In the section titled “Three-Variable

(Time Series) Models” [2], the three-variable models are discussed briefly in frequency formulation. In the section titled “Multiple-Variable Models” [3], causal relations between one group of time series and the other group of time series are studied in frequency domain. The underlying *principle of orthogonality* is kept for error minimization.

Finally, we briefly discuss other prediction methods, that is, the Box–Jenkins Prediction.

LSE MODEL

Basic LSE Model

At the beginning of this article, the LSE model is presented.

In a causal system, consider two time series $X(i)$ and $Z(i)$ ($i = 0, 1, 2, 3, \dots$), where $X(i), X(i-1), X(i-2), \dots, X(i-M+1)$ are the unobserved underlying variables and influence the observed time series $Z(i)$ in a linear way, expressed as following

$$Z(i) = \sum_{k=0}^{M-1} w'_k X(i-k) + e'(i), \quad (1)$$

where w'_k ($k = 0, 1, 2, \dots, M-1$) are parameters of the LSE model and $e'(i)$ is the model error. The measurement error $e'(i)$ is unobservable and is used to count the model's inaccuracy.

Let

$$Y(i) \triangleq \sum_{k=0}^{M-1} w'_k X(i-k), \quad (2)$$

$$e'(i) \triangleq Z(i) - Y(i). \quad (3)$$

The measurement error process $e'(i)$ in the LSE model is assumed white with zero expectation and the same variance:

$$E(e'(i)) = 0, \quad \text{for all } i, \quad (4)$$

and

$$E(e'(i)e'(k)) = \sigma^2 I(i-k), \quad (5)$$

where

$$I(i-k) = \begin{cases} 1 & \text{if } i-k \neq 0 \\ 0 & \text{if } i-k = 0. \end{cases}$$

In the LSE model, given the observed data, we select/design the tap weights w'_k ($k = 0, 1, 2, \dots, M-1$) to minimize the sum of error squares (viewed as error energy). The objective function is as following:

$$\Theta(w'_0, w'_1, \dots, w'_{M-1}) = \sum_{i=s}^t [e'(i)]^2, \quad (6)$$

where s and t ($s \leq t$) are the index limits at which the error minimization occurs. Once the tap weights w'_k ($k = 0, 1, 2, \dots, M-1$) are selected, they are fixed as parameters in the interval $s \leq i \leq t$. Without losing generality, suppose index range of the observed data is $[1, N]$, and set $s = M, t = N$. Then, we get the observed input data matrix [4,5]:

$$\begin{bmatrix} X(M) & X(M+1) & \dots & X(N) \\ X(M-1) & X(M) & & X(N-1) \\ \vdots & & \ddots & \vdots \\ X(1) & X(2) & \dots & X(N-M+1) \end{bmatrix},$$

where the arrays above, from the first column (M) to the last column (N), correspond to the unobserved input variable vectors. Hence, the cost function in Equation (6) becomes

$$\Theta(w'_0, w'_1, \dots, w'_{M-1}) = \sum_{i=M}^N [e'(i)]^2, \quad (7)$$

Assume that $\Theta(w'_0, w'_1, \dots, w'_{M-1})$ is differentiable, then the necessary condition for minimizing $\Theta(w'_0, w'_1, \dots, w'_{M-1})$ is

$$\nabla \Theta = [0, 0, \dots, 0]^T \quad (M \text{ dimensional}). \quad (8)$$

Applying Equation (7) to Equation (8), we get

$$\begin{aligned} & \left[-2 \sum_{i=M}^N X(i-0)e'_{\min}(i), \right. \\ & \quad -2 \sum_{i=M}^N X(i-1)e'_{\min}(i), \dots, \\ & \quad \left. -2 \sum_{i=M}^N X(i-(M-1))e'_{\min}(i) \right]^T \\ & = [0, 0, \dots, 0]^T \quad (M \text{ dimensional}). \end{aligned} \quad (9)$$

Therefore, a necessary condition for $\Theta(w'_0, w'_1, \dots, w'_{M-1})$ to reach its global minimum is:

$$\sum_{i=M}^N X(i-k)e'_{\min}(i) = 0, \quad k = 0, 1, \dots, M-1. \quad (10)$$

Let the tap weights $w'_0, w'_1, \dots, w'_{M-1}$ that are optimized to operate the LSE condition have the special values $w_0^{\min}, w_1^{\min}, \dots, w_{M-1}^{\min}$. Then, we define $Y^{\min}(i)$ as

$$Y^{\min}(i) \triangleq \sum_{k=0}^{M-1} w_k^{\min} X(i-k). \quad (11)$$

Multiplying both sides of Equation (10) by $w_k^{\min}$ and then adding the results over the value of index k in the range $0, 1, 2, \dots, (M-1)$, we get

$$\sum_{k=0}^{M-1} w_k^{\min} \sum_{i=M}^N X(i-k)e'_{\min}(i) = 0. \quad (12)$$

Interchanging the order of summation, we get

$$\sum_{i=M}^N \left[\sum_{k=0}^{M-1} w_k^{\min} X(i-k) \right] e'_{\min}(i) = 0. \quad (13)$$

Using Equation (11) to substitute the inside summation item in Equation (13), we get

$$\sum_{i=M}^N Y_i^{\min} e'_{\min}(i) = 0. \quad (14)$$

Equation (14) is a necessary condition for global minimization of $\Theta(w'_0, w'_1, \dots, w'_{M-1})$, which is called the *principle of orthogonality*. In Equation (14), $Y_i^{\min}$ (defined in Equation (11)) is considered the least square (LS) estimate of the desired (observed) response $Z(i)$. Hence, $Y_i^{\min}$ can be written as $\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})$, then Equation (14) can be reformulated as

$$\sum_{i=M}^N \hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1}) e'_{\min}(i) = 0. \quad (15)$$

Taking the time average of the left-hand side of Equation (15), we find that the cross-correlation of two time series, $\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})$ and $e_{\min}^*(i)$ is 0. So, the LSE estimate of the observed response $Z(i)$ represented by $\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})$, and the LSE model error process $e_{\min}^*(i)$ are orthogonal over time.

$$\begin{aligned} \underbrace{Z(i)}_{\text{observed response}} &= \underbrace{\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})}_{\text{estimate of observed response}} \\ &+ \underbrace{e'_{\min}(i)}_{\text{LSE model estimation error}} \end{aligned} \quad (16)$$

If we define

$$\begin{aligned} E_z &= \sum_{i=M}^N |Z(i)|^2 \\ E_{\text{LSE est}} &= \sum_{i=M}^N |\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})|^2 \\ E_{\text{LSE error}} &= \sum_{i=M}^N |e_{\min}^*(i)|^2 \end{aligned}$$

by using the *principle of orthogonality* (Eq. 15), we get

$$E_z = E_{\text{LSE estimation}} + E_{\text{LSE error}}. \quad (17)$$

Equation (17) shows that the energy of the observed time series $Z(i)$ is the summation of the energy of the LSE estimation time series, and the energy of the LSE process time series [24]. All the three items in Equation (17) are nonnegative.

Another Approach to Understanding the LSE Model

Since $Y_i^{\min}$ in Equation (11) is also written as $\hat{Z}^{\min}(i|X_i, X_{i-1}, \dots, X_{i-M+1})$ in Equation (16), according to Equations (11) and 16, we get

$$e_{\min}^*(i) = Z(i) - \sum_{k=0}^{M-1} w_k^{\min} X(i-k). \quad (18)$$

Substituting Equation (18) into Equation (14), we get

$$\sum_{i=M}^N X(i-k) \left[Z(i) - \sum_{h=0}^{M-1} w_h^{\min} X(i-h) \right] = 0, \quad k = 0, 1, 2, \dots, M-1. \quad (19)$$

$$\Phi \triangleq \begin{bmatrix} \Phi(0,0) & \Phi(1,0) & \dots & \Phi(M-1,0) \\ \Phi(0,1) & \Phi(1,1) & \dots & \Phi(M-1,1) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(0,M-1) & \Phi(1,M-1) & \dots & \Phi(M-1,M-1) \end{bmatrix}. \quad (24)$$

Let the vector β (M by 1) be

$$\beta \triangleq [\beta(0) \quad \beta(-1) \quad \dots \quad \beta(-M+1)]^T. \quad (25)$$

Let the vector $w^{\min}$ (M by 1) be

$$w^{\min} \triangleq [w_0^{\min} \quad w_1^{\min} \quad \dots \quad w_{M-1}^{\min}]^T. \quad (26)$$

Then, Equation (23) can be rewritten as

$$\beta = \Phi w^{\min}. \quad (27)$$

Assuming Φ is nonsingular, then

$$w^{\min} = \Phi^{-1} \beta. \quad (28)$$

Reorganizing we get

$$\begin{aligned} & \sum_{i=M}^N X(i-k)Z(i) \\ &= \sum_{h=0}^{M-1} \left[w_h^{\min} \sum_{i=M}^N X(i-k)X(i-h) \right], \\ & k = 0, 1, 2, \dots, M-1. \end{aligned} \quad (20)$$

Here we define

$$\beta(-k) \triangleq \sum_{i=M}^N X(i-k)Z(i), \quad 0 \leq k \leq M-1. \quad (21)$$

$$\begin{aligned} \phi(h, k) &\triangleq \sum_{i=M}^N X(i-k)X(i-h), \\ 0 \leq h \leq M-1, \quad 0 \leq k \leq M-1. \end{aligned} \quad (22)$$

Then Equation (20) can be rewritten as

$$\begin{aligned} \beta(-k) &= \sum_{h=0}^{M-1} [w_h^{\min} \phi(h, k)], \\ k &= 0, 1, 2, \dots, M-1. \end{aligned} \quad (23)$$

Let the M by M matrix Φ be

Equation (28) is the design of a linear LSE estimation, where Φ^{-1} is the inverse of the time-average cross-correlation matrix and β is the cross-correlation vector between unobserved time series $\mathbf{X}$ and observed time series $\mathbf{Z}$. When noise is assumed to be white Gaussian distributed, its counterpart filter design is the Wiener-Hopf Equation.

According to Equations (22) and (24), Φ can be decomposed into

$$\Phi \triangleq \mathbf{A}^T \mathbf{A}, \quad (29)$$

where

$$\mathbf{A} \triangleq \begin{bmatrix} X(M) & X(M+1) & \cdots & X(N) \\ X(M-1) & X(M) & \cdots & X(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ X(1) & X(2) & \cdots & X(N-M+1) \end{bmatrix}. \quad (30)$$

$$\text{Let } \mathbf{Z} \triangleq [Z(M) \ Z(M+1) \ \dots \ Z(N)]^T; \quad (31)$$

then, according to Equations (30) and (31) and the definition of β in Equations (21) and (25):

$$\beta \triangleq \mathbf{A}^T \mathbf{Z}. \quad (32)$$

Equation (27) can be rewritten as the following using the decomposition expression in Equation (29)

$$\beta = \mathbf{A}^T \mathbf{A} \mathbf{w}^{\min}. \quad (33)$$

Integrating Equations (32) and (33)

$$\mathbf{A}^T \mathbf{Z} = \mathbf{A}^T \mathbf{A} \mathbf{w}^{\min}. \quad (34)$$

Therefore,

$$\mathbf{w}^{\min} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Z}. \quad (35)$$

$$\text{Let } \mathbf{Y} = \hat{\mathbf{Z}}^{\min}$$

$$\triangleq [\hat{Z}^{\min}(M) \ \hat{Z}^{\min}(M+1) \ \dots \ \hat{Z}^{\min}(N)]^T, \quad (36)$$

Substituting Equation (35) into Equation (11), we get

$$\mathbf{Y} = \hat{\mathbf{Z}}^{\min} \triangleq \mathbf{A} \mathbf{w}^{\min} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Z}. \quad (37)$$

It is worth noting that $\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is the projection operator onto the linear space spanned by the columns of the data matrix $\mathbf{A}$.

Hence, the LSE estimation vector, which is expressed by $\hat{\mathbf{Z}}^{\min}$ or $\mathbf{Y}$, is the orthogonal projection of observable variable vector $\mathbf{Z}$ onto the linear space spanned by the columns of the data matrix $\mathbf{A}$. Suppose $M < N - M + 1$ (Fig. 1) [6]. If data matrix $\mathbf{A}$ is full rank, then its column space is an M -dimensional subspace of the $(N - M + 1)$ -dimensional full space. Assume that the data matrix $\mathbf{A}$ is already known with no uncertainty.

In practice, $\mathbf{A}$ could be unknown. In this case, a set of training data with limited length is used to estimate the LSE parameters. Some further properties of LSE are investigated by other literatures [7–10]. In an alternative way of thinking, $X(t)$ can be considered as another time series observed. Then, we interpret the problem as follows: the present states of time series $Z(t)$ can be predicted by the present and previous states of another time series $X(t)$, using the LSE model.

MMSE MODEL PREDICTION

Next, we introduce the MMSE prediction, described by the conditional expectation [11,12]:

$$\hat{X}_p(q) = E[X_{p+q} | X_p, X_{p-1}, \dots]. \quad (38)$$

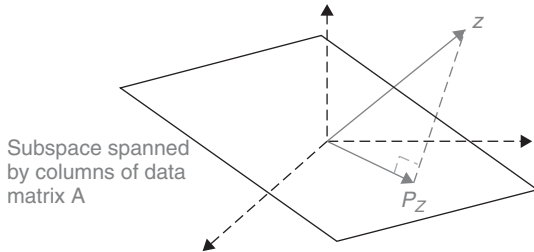


Figure 1. The LSE prediction can be considered as the observed variable vector $\mathbf{Z}$ projected from the $(N - M + 1)$ -dimensional whole space onto the M -dimensional subspace spanned by columns of data matrix $\mathbf{A}$.

If we want a linear-form prediction, $\hat{X}_p(q)$ should have the following form [12,13]:

$$\hat{X}_p(q) = \sum_{i=0}^{p-1} w_i X_{p-i}. \quad (39)$$

Suppose we identify a particular model for a given time series, we need to estimate the model parameters and wish to compute forecasts from the fitted model (M_f) instead of the true model (M) (which we do not know). If we use a quadratic loss function, the best way to compute a forecast is to choose $\hat{X}_p(h)$ ($h = 1, 2, \dots, q-1$) to be the conditional expected value of $\hat{X}(p+h)$ on the model, and use information I_N available at time N .

$$\hat{X}_p(q) = E[X_{p+q} | M_f, I_N] \quad (40)$$

Then, the MMSE forecast with MA model of possibly infinite order would be [14]:

$$\hat{X}_p(q) = \sum_{i=q}^{\infty} w_i Z_{p+q-i}, \quad (41)$$

where the future values are replaced by zero.

More generally, the MMSE forecast from an ARMA model [15] can be computed by

$$\begin{aligned} \hat{X}_p(q) = & \sum_{i=1}^{q-1} \Theta_i \hat{X}_{p+q-i} \\ & + \sum_{i=q}^K \Theta_i X_{p+q-i} + \sum_{i=q}^J w_i Z_{p+q-i}, \end{aligned} \quad (42)$$

where the future value of Z is replaced by zero and the future values of X is replaced by the conditional expectation $\hat{X}$ [14].

Similarly, if we use the ARIMA model [15], then the MMSE forecast would be

$$\begin{aligned} & \hat{X}_p(q) - \hat{X}_p(q-1) \\ &= \sum_{i=1}^{q-2} \Theta_i (\hat{X}_{p+q-i} - \hat{X}_{p+q-1-i}) \\ &+ \Theta_{q-1} (\hat{X}_{p+1} - X_p) \\ &+ \sum_{i=q}^K \Theta_i (X_{p+q-i} - X_{p+q-1-i}) \\ &+ \sum_{t=q}^J w_t Z_{p+q-t}, \end{aligned} \quad (43)$$

where the future value of Z is replaced by zero and the future values of X are replaced by the conditional expectation of $\hat{X}$.

ARMA and ARIMA models can convert to each other, although generally ARMA models are used for accumulative value forecast, while ARIMA models are used for differential value forecast. For special cases, for example, ARIMA (0, 1, 1), it is a simple exponential smoothing. Recursive calculation is commonly used for the q -step MMSE forecast at time p using ARMA or ARIMA models.

CAUSAL RELATIONS BETWEEN TWO, THREE AND MULTIPLE VARIABLES IN PREDICTION

For time series prediction, in the section titled “LSE Model”, we discuss a deterministic model: LSE; in the section titled “MMSE Model Prediction”, we discuss a stochastic model: MMSE. Both the LSE and MMSE models aim at minimizing the prediction errors, in their respective senses. In this section, we continue to discuss on a wide-sense stationary, purely nondeterministic time series, for a closely related issue: the causal relations between multiple time series. We start from two-variable (time series) and three-variable (time series) models.

Two-Variable (Time Series) Models

Let X_t and Y_t be two wide-sense stationary, purely nondeterministic time series with zero

means. The *simple causal model* [2] is

$$X_t = \sum_{j=1}^m a_j X_{t-j} + \sum_{j=1}^m b_j Y_{t-j} + \varepsilon_t, \quad (44)$$

$$Y_t = \sum_{j=1}^m c_j X_{t-j} + \sum_{j=1}^m d_j Y_{t-j} + \eta_t, \quad (45)$$

where ε_t and η_t are the two zero-mean serially uncorrelated noise processes (time series) with $E(\varepsilon_t, \eta_s) = 0$ for any t and s . Since ε_t is serially uncorrelated, for any $p \neq q$, $E(\varepsilon_p, \varepsilon_q) = 0$. Since η_t is serially uncorrelated, for any $p \neq q$, $E(\eta_p, \eta_q) = 0$.

In the *simple causal model*, the previous states of $\{X_t\}$ and $\{Y_t\}$ are used to predict the present states of $\{X_t\}$ and $\{Y_t\}$, respectively.

A more general model—*instantaneous causal model* [2]—is

$$X_t = \sum_{j=1}^m a_j X_{t-j} + \sum_{j=0}^m b_j Y_{t-j} + \varepsilon_t, \quad (46)$$

$$Y_t = \sum_{j=0}^m c_j X_{t-j} + \sum_{j=1}^m d_j Y_{t-j} + \eta_t. \quad (47)$$

In the *instantaneous causal model*, the previous states of $\{X_t\}$, as well as the previous and present states of $\{Y_t\}$, are used to predict the present state of $\{X_t\}$. Similarly, the previous states of $\{Y_t\}$, as well as the previous and present states of $\{X_t\}$, are used to predict the present state of $\{Y_t\}$.

Comparing Equations (44) and (46), if the instantaneous causality is occurring ($b_0 \neq 0$), then the knowledge of the present state of $\{Y_t\}$ will improve the prediction for the present state of $\{X_t\}$, and $\text{var}(\varepsilon_t)$ will decrease. Similarly, comparing 45 and 47, if the instantaneous causality is occurring ($c_0 \neq 0$), then the knowledge of the present state of $\{X_t\}$ will improve the prediction for the present state of $\{Y_t\}$, and $\text{var}(\eta_t)$ will decrease.

To obtain the frequency formulation of the *simple causal model*, we can rewrite Equations (44) and (45) by using the time shift (delay) operator U , where $UX_t = X_{t-1}$

$$X_t = a(U)X_t + b(U)Y_t + \varepsilon_t, \quad (48)$$

$$Y_t = c(U)X_t + d(U)Y_t + \eta_t, \quad (49)$$

where $COEF(U)$ ($COEF = a, b, c, d$) are power series in U with the coefficient of U^0 being zero.

Next, we use *Cramer representation* of the series to get [2]:

$$X_t = \int_{-\pi}^{\pi} e^{it\omega} dZ_x(\omega), \quad (50)$$

$$Y_t = \int_{-\pi}^{\pi} e^{it\omega} dZ_y(\omega). \quad (51)$$

(*Cramer representation* is widely used in spectrum estimation for wide-sense stationary process; similar to the role of *Fourier transform* in time-frequency analysis in deterministic time series.)

Because k units of time shift (delay) (represented by U^k) in the *time domain* are equivalent to being multiplied by $e^{-i\omega k}$ in the *frequency domain*, and also noting that *Cramer representation* is a linear transformation, we get:

$$a(U)X_t = \int_{-\pi}^{\pi} e^{it\omega} a(e^{-i\omega}) dZ_x(\omega), \quad (52)$$

$$b(U)Y_t = \int_{-\pi}^{\pi} e^{it\omega} b(e^{-i\omega}) dZ_y(\omega), \quad (53)$$

$$c(U)X_t = \int_{-\pi}^{\pi} e^{it\omega} c(e^{-i\omega}) dZ_x(\omega), \quad (54)$$

$$d(U)Y_t = \int_{-\pi}^{\pi} e^{it\omega} d(e^{-i\omega}) dZ_y(\omega). \quad (55)$$

Integrating Equations (50)–(55) into Equations (48) and (49), we get (for every t)

$$\int_{-\pi}^{\pi} e^{it\omega} [(1 - a(e^{-i\omega})) dZ_x(\omega) - b(e^{-i\omega}) dZ_y(\omega) - dZ_\varepsilon(\omega)] = 0, \quad (56)$$

$$\int_{-\pi}^{\pi} e^{it\omega} [(-c(e^{-i\omega})) dZ_x(\omega) + (1 - d(e^{-i\omega})) dZ_y(\omega) - dZ_\eta(\omega)] = 0. \quad (57)$$

Because Equations (56) and (57) hold for every t , it further implies that

$$A(\omega) \begin{bmatrix} dZ_x(\omega) \\ dZ_y(\omega) \end{bmatrix} = \begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\eta(\omega) \end{bmatrix}, \quad (58)$$

where $A(\omega) \triangleq \begin{bmatrix} 1 - a(e^{-i\omega}) & -b(e^{-i\omega}) \\ -c(e^{-i\omega}) & 1 - d(e^{-i\omega}) \end{bmatrix}$
for every $\omega \in [-\pi, \pi]$. Therefore,

$$\begin{bmatrix} dZ_x(\omega) \\ dZ_y(\omega) \end{bmatrix} = A(\omega)^{-1} \begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\eta(\omega) \end{bmatrix} \quad (59)$$

Then, we obtain the spectral and cross-spectral matrices:

$$\begin{aligned} & \begin{bmatrix} f_x(\omega) & Cr(\omega) \\ Cr^*(\omega) & f_y(\omega) \end{bmatrix} \\ & \triangleq E \begin{bmatrix} dZ_x(\omega) \\ dZ_y(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_x(\omega)} & \overline{dZ_y(\omega)} \end{bmatrix} \\ & = A(\omega)^{-1} E \left(\begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\eta(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_\varepsilon(\omega)} & \overline{dZ_\eta(\omega)} \end{bmatrix} \right) \\ & \quad \times (A(\omega)^*)^{-1} \\ & = A(\omega)^{-1} \begin{bmatrix} \sigma_\varepsilon(\omega)^2 & 0 \\ 0 & \sigma_\eta(\omega)^2 \end{bmatrix} (A(\omega)^T)^{-1}, \end{aligned} \quad (60)$$

where

$$\begin{aligned} f_x(\omega) &= \frac{1}{\Delta(\omega)} (|1 - d(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |b(e^{-i\omega})|^2 \sigma_\eta(\omega)^2), \\ f_y(\omega) &= \frac{1}{\Delta(\omega)} (|c(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |1 - a(e^{-i\omega})|^2 \sigma_\eta(\omega)^2, Cr(\omega) \\ & \triangleq C_1(\omega) + C_2(\omega) \\ & = \frac{1}{\Delta(\omega)} (1 - d(e^{-i\omega})) \overline{c(e^{-i\omega})} \sigma_\varepsilon(\omega)^2 \\ & \quad + \frac{1}{\Delta(\omega)} (1 - \overline{a(e^{-i\omega})}) b(e^{-i\omega}) \sigma_\eta(\omega)^2, \\ \Delta(\omega) &= |(1 - a(e^{-i\omega})) (1 - d(e^{-i\omega})) \\ & \quad - b(e^{-i\omega}) c(e^{-i\omega})|^2. \end{aligned}$$

When we study the causality between the two time series $\{X_t\}$ and $\{Y_t\}$, if $\{Y_t\}$ is NOT causing $\{X_t\}$, then $b(U) = 0$, $b(e^{-i\omega}) = 0$ for all ω , thus $C_2(\omega) = 0$; similarly, if $\{X_t\}$ is NOT causing $\{Y_t\}$, then $c(U) = 0$, $c(e^{-i\omega}) = 0$ for all ω , thus $C_1(\omega) = 0$. Therefore, the cross spectrum $Cr(\omega)$ can be decomposed into $C_1(\omega)$ and $C_2(\omega)$, representing “ $\{X_t\}$ causing $\{Y_t\}$ ” and “ $\{Y_t\}$ causing $\{X_t\}$ ”, respectively, that is,

$C_1(\omega)$ and $C_2(\omega)$ can be considered as two arms of the feedback mechanism.

In Equations (61) and (62), we define two measures of *strength of causality*: representing “ $\{X_t\}$ to $\{Y_t\}$ ” and “ $\{Y_t\}$ to $\{X_t\}$,” respectively:

$$\begin{aligned} C_{\mathbf{xy}}(\omega) &\triangleq \frac{|C_1(\omega)|^2}{f_x(\omega)f_y(\omega)} \\ &= \frac{|1 - d(e^{-i\omega})|^2 |c(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2}{(|1 - d(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |b(e^{-i\omega})|^2 \sigma_\eta(\omega)^2) (|c(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 + \\ & \quad |1 - a(e^{-i\omega})|^2 \sigma_\eta(\omega)^2)} \end{aligned} \quad (61)$$

$$\begin{aligned} C_{\mathbf{yx}}(\omega) &\triangleq \frac{|C_2(\omega)|^2}{f_x(\omega)f_y(\omega)} \\ &= \frac{|1 - a(e^{-i\omega})|^2 |b(e^{-i\omega})|^2 \sigma_\eta(\omega)^2}{(|1 - d(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |b(e^{-i\omega})|^2 \sigma_\eta(\omega)^2) (|c(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |1 - a(e^{-i\omega})|^2 \sigma_\eta(\omega)^2)} \end{aligned} \quad (62)$$

In Equations (63) and (64), the *causal phase* diagrams are further defined, to study the phase lag against frequency cross spectrum:

$$\phi_{\mathbf{xy}}(\omega) \triangleq \text{phase of } C_1(\omega), \quad (63)$$

$$\phi_{\mathbf{yx}}(\omega) \triangleq \text{phase of } C_2(\omega). \quad (64)$$

Similar to the frequency formulation of the *simple causal model* (Eqs 44 and 45), we consider the frequency formulation of the *instantaneous causal model* (Eqs 46 and 47). *Spectral and cross-spectral matrices* can be obtained similar to what is given by Equation (60) as follow:

$$\begin{aligned} & \begin{bmatrix} f_x(\omega) & Cr(\omega) \\ Cr^*(\omega) & f_y(\omega) \end{bmatrix} \\ & \triangleq E \begin{bmatrix} dZ_x(\omega) \\ dZ_y(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_x(\omega)} & \overline{dZ_y(\omega)} \end{bmatrix} \\ & = A(\omega)^{-1} E \end{aligned}$$

$$\begin{aligned} & \times \left(\begin{bmatrix} dZ_\varepsilon(\omega) \\ dZ_\eta(\omega) \end{bmatrix} \begin{bmatrix} \overline{dZ_\varepsilon(\omega)} & \overline{dZ_\eta(\omega)} \end{bmatrix} \right) \\ & \times (A(\omega)^{*T})^{-1} = A(\omega)^{-1} \begin{bmatrix} \sigma_\varepsilon(\omega)^2 & 0 \\ 0 & \sigma_\eta(\omega)^2 \end{bmatrix} \\ & \times (A(\omega)^T)^{-1}, \end{aligned} \quad (65)$$

where

$$\begin{aligned} f_x(\omega) &= \frac{1}{\Delta(\omega)} (|1 - d(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |b(e^{-i\omega})|^2 \sigma_\eta(\omega)^2) \\ f_y(\omega) &= \frac{1}{\Delta(\omega)} (|c(e^{-i\omega})|^2 \sigma_\varepsilon(\omega)^2 \\ & \quad + |1 - a(e^{-i\omega})|^2 \sigma_\eta(\omega)^2), Cr(\omega) \\ & \triangleq C_1(\omega) + C_2(\omega) \\ &= \frac{1}{\Delta(\omega)} (1 - d(e^{-i\omega})) \overline{(c(e^{-i\omega}) - c_0)} \sigma_\varepsilon(\omega)^2 \\ & \quad + \frac{1}{\Delta(\omega)} (1 - \overline{a(e^{-i\omega})}) (b(e^{-i\omega}) \\ & \quad - b_0) \sigma_\eta(\omega)^2, \\ \Delta(\omega) &= |(1 - a(e^{-i\omega}))(1 - d(e^{-i\omega})) \\ & \quad - (b(e^{-i\omega}) - b_0)(c(e^{-i\omega}) - c_0)|^2 \end{aligned}$$

(Note: b_0 and c_0 are the coefficients of instant causality of $\{Y_t\}$ in Equation (46) and $\{X_t\}$ in Equation (47), respectively)

Summary: the spectral methods can be applied to the analysis of causal relations between two time series $\{X_t\}$ and $\{Y_t\}$. Furthermore, the spectral methods in the frequency domain are robust in their interpretation, compared with the equation models in the time domain.

Three-Variable (Time Series) Models

Extending the two-variable model (*simple causal model*) in Equations (48) and (49) to

three-variable model [2], we get

$$X_t = a_1(U)X_t + b_1(U)Y_t + c_1(U)Z_t + \varepsilon_{1,t}, \quad (66)$$

$$Y_t = a_2(U)X_t + b_2(U)Y_t + c_2(U)Z_t + \varepsilon_{2,t}, \quad (67)$$

$$Z_t = a_3(U)X_t + b_3(U)Y_t + c_3(U)Z_t + \varepsilon_{3,t}. \quad (68)$$

Here we keep using the time shift operator U ($UX_t = X_{t-1}$). Since there is no instantaneous causality, the coefficients of U^0 are always zero.

Applying the frequency formulation (using *Cramer representation*) to the three-variable model, the spectral and cross-spectral matrices are obtained:

$$\begin{aligned} & \begin{bmatrix} f_x(\omega) & C_r^{xy}(\omega) & C_r^{xz}(\omega) \\ C_r^{xy*}(\omega) & f_y(\omega) & C_r^{yz}(\omega) \\ C_r^{xz*}(\omega) & C_r^{yz*}(\omega) & f_z(\omega) \end{bmatrix} \\ &= E \begin{bmatrix} dZ_x(\omega) \\ dZ_y(\omega) \\ dZ_z(\omega) \end{bmatrix} \\ & \quad \times \begin{bmatrix} \overline{dZ_x(\omega)} & \overline{dZ_y(\omega)} & \overline{dZ_z(\omega)} \end{bmatrix} \\ &= A(\omega)^{-1} E \begin{bmatrix} dZ_{\varepsilon 1}(\omega) \\ dZ_{\varepsilon 2}(\omega) \\ dZ_{\varepsilon 3}(\omega) \end{bmatrix} \\ & \quad \times \begin{bmatrix} \overline{dZ_{\varepsilon 1}(\omega)} & \overline{dZ_{\varepsilon 2}(\omega)} & \overline{dZ_{\varepsilon 3}(\omega)} \end{bmatrix} \\ & \quad \times (A(\omega)^{*T})^{-1} \\ &= A(\omega)^{-1} \begin{bmatrix} \sigma_1(\omega)^2 & 0 & 0 \\ 0 & \sigma_2(\omega)^2 & 0 \\ 0 & 0 & \sigma_3(\omega)^2 \end{bmatrix} \\ & \quad \times (A(\omega)^{*T})^{-1}, \end{aligned} \quad (69)$$

where

$$A = \begin{bmatrix} 1 - a_1(e^{-i\omega}) & -b_1(e^{-i\omega}) & -c_1(e^{-i\omega}) \\ -a_2(e^{-i\omega}) & 1 - b_2(e^{-i\omega}) & -c_2(e^{-i\omega}) \\ -a_3(e^{-i\omega}) & -b_3(e^{-i\omega}) & 1 - c_3(e^{-i\omega}) \end{bmatrix}$$

Similar to the calculation in the two-variable model, the power spectrum of X_t ,

Y_t , and Z_t , as well as the cross spectrum: $C_r^{xy}(\omega)$, $C_r^{xz}(\omega)$, and $C_r^{yz}(\omega)$, can be calculated in the three-variable model.

Further, the *Partial cross spectrum* between X_t and Y_t given Z_t is defined as $C_r^{xy,z}(\omega) \triangleq C_r^{xy}(\omega) - C_r^{xz}(\omega)f_z(\omega)^{-1}C_r^{zy}(\omega)$, where $C_r^{xy}(\omega)$, $C_r^{xz}(\omega)$, $C_r^{zy}(\omega)$, $f_z(\omega)$ are derived from Equation (69). When *partial cross spectrum* is considered, more useful results arise.

$$\begin{aligned} C_r^{xy,z}(\omega) &= \frac{1}{\Delta_Z(\omega)} \sigma_1^2 \sigma_2^2 b_3(e^{-i\omega}) a_3(e^{-i\omega}) \\ &\quad - \frac{1}{\Delta_Z(\omega)} \sigma_1^2 \sigma_2^2 (b_2(e^{-i\omega}) - 1) a_2(e^{-i\omega}) \\ &\quad - \frac{1}{\Delta_Z(\omega)} \sigma_2^2 \sigma_3^2 b_1(e^{-i\omega}) (a_1(e^{-i\omega}) - 1) \\ &\triangleq C_{r1}^{xy,z}(\omega) + C_{r2}^{xy,z}(\omega) + C_{r3}^{xy,z}(\omega), \end{aligned} \quad (70)$$

where

$$\begin{aligned} \Delta_Z(\omega) &= \sigma_1^2 |(b_2(e^{-i\omega}) - 1)(c_3(e^{-i\omega}) - 1) \\ &\quad - c_2(e^{-i\omega})b_3(e^{-i\omega})|^2 \\ &\quad + \sigma_2^2 |c_1(e^{-i\omega})b_3(e^{-i\omega}) \\ &\quad - b_1(e^{-i\omega})(c_3(e^{-i\omega}) - 1)|^2 \\ &\quad + \sigma_3^2 |b_1(e^{-i\omega})c_2(e^{-i\omega}) \\ &\quad - c_1(e^{-i\omega})(b_2(e^{-i\omega}) - 1)|^2. \end{aligned}$$

As shown in Equation (70), the partial cross spectrum between X_t and Y_t can be decomposed into $C_{r1}^{xy,z}(\omega)$, $C_{r2}^{xy,z}(\omega)$, and $C_{r3}^{xy,z}(\omega)$, where the first component represents the relation between X_t and Y_t through Z_t , and the second and third components together represent generalizations of the causal cross spectra (between X_t and Y_t), which come from the two-variable case.

Multiple-Variable Models

Generalizing the two- and three-variable causal relation models to multiple-variable models brings in-depth insights into the causal relation and feedback problem [3]. In this section, causal relations between one group of time series and another group of time series are studied. The discussion in this section is still based on wide-sense stationary, purely nondeterministic time series.

Let $\mathbf{V}_t : m$ by 1 be a zero-mean wide-sense stationary, purely nondeterministic

multiple time series. And suppose that $\mathbf{V}_t$ has been partitioned into two sub-vectors: $\mathbf{V}_t^T = [\mathbf{X}_t^T, \mathbf{Y}_t^T]$, $\mathbf{X}_t : k$ by 1 and $\mathbf{Y}_t : l$ by 1, $m = k + l$. We try to find out the causal relations as well as the measurement of linear dependence and feedback between $\mathbf{X}_t : k$ by 1 and $\mathbf{Y}_t : l$ by 1.

By assuming the existence of the spectral density matrix $S_{\mathbf{V}}(\omega)$ at all frequencies $\omega \in [-\pi, \pi]$, we have a corresponding partition of the spectral density matrix $S_{\mathbf{V}}(\omega)$:

$$S_{\mathbf{V}}(\omega) = \begin{bmatrix} S_{\mathbf{X}}(\omega) & S_{\mathbf{X}\mathbf{Y}}(\omega) \\ S_{\mathbf{Y}\mathbf{X}}(\omega) & S_{\mathbf{Y}}(\omega) \end{bmatrix}. \quad (71)$$

We assume that there is a moving average representation of $\mathbf{V}_t : m$ by 1 as follows:

$$\mathbf{V}_t = \sum_{r=0}^{\infty} \mathbf{A}_r \boldsymbol{\varepsilon}_{t-r}, \quad E(\boldsymbol{\varepsilon}_t) = 0, \text{var}(\boldsymbol{\varepsilon}_t) = \gamma, \quad (72)$$

where $\boldsymbol{\varepsilon}_{t-r}$ is serially uncorrelated; for any $\sum_{r=0}^{\infty} \mathbf{A}_r \mathbf{Z}^r = 0$, $|\mathbf{Z}| \geq 1$ holds; and $\sum_{r=0}^{\infty} \|\mathbf{A}_r\|^2 < \infty$. ($\|P\|$ is the square root of the largest eigenvalue of $P'P$; $|P|$ is the square root of the determinant of $P'P$). The existence of moving average representation is equivalent to the existence of the spectral density matrix $S_{\mathbf{V}}(\omega)$ at almost all frequencies $\omega \in [-\pi, \pi]$ [16]. Furthermore, a lower bound on the mean-squared error of one-step-ahead minimum mean γ is given by using $S_{\mathbf{V}}(\omega)$:

$$\ln |\gamma| = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |S_{\mathbf{V}}(\omega)| d\omega. \quad (73)$$

If there exists a constant $\text{CONST} \geq 1$ such that for almost all $\omega \in [-\pi, \pi]$, [17]

$$\text{CONST}^{-1} I_n \leq S_{\mathbf{V}}(\omega) \leq \text{CONST} I_n. \quad (74)$$

(This means that both $(\text{CONST} I_n - S_{\mathbf{V}}(\omega))$ and $(S_{\mathbf{V}}(\omega) - \text{CONST}^{-1} I_n)$ are positive semidefinite, and that the spectral density matrix is bounded uniformly away from zero almost everywhere in $[-\pi, \pi]$, which is weaker than the condition that $\mathbf{V}_t$ is an ARMA process of finite orders with invertible AR and MA parts.)

Then, the relationship in Equation (70) can be inverted so that $\mathbf{V}_t$ become a function of $\{\mathbf{V}_{t-r}, r \geq 1\}$ and $\boldsymbol{\varepsilon}_t$.

$$\mathbf{V}_t = \sum_{r=1}^{\infty} B_r \mathbf{V}_{t-r} + \boldsymbol{\varepsilon}_t, \quad E(\boldsymbol{\varepsilon}_t) = 0, \text{var}(\boldsymbol{\varepsilon}_t) = \gamma. \quad (75)$$

Predict from its Own Past. Considering that $\mathbf{V}_t$ has been partitioned into two sub-vectors $\mathbf{V}_t^T = [\mathbf{X}_t^T, \mathbf{Y}_t^T]$, if the condition in Equation (73) holds, then each of $\mathbf{X}_t$ and $\mathbf{Y}_t$ has an autoregressive (AR) representation denoted by

$$\mathbf{X}_t = \sum_{r=1}^{\infty} \mathbf{E}1_r \mathbf{X}_{t-r} + \boldsymbol{\eta}1_t, \quad E(\boldsymbol{\eta}1_t) = 0, \text{var}(\boldsymbol{\eta}1_t) = \Gamma 1; \quad (76)$$

$$\mathbf{Y}_t = \sum_{r=1}^{\infty} \mathbf{G}1_r \mathbf{Y}_{t-r} + \boldsymbol{\sigma}1_t, \quad E(\boldsymbol{\sigma}1_t) = 0, \text{var}(\boldsymbol{\sigma}1_t) = \Xi 1; \quad (77)$$

where $\boldsymbol{\eta}1_t$ is one-step-ahead error when $\mathbf{X}_t$ is predicted from its own past $\{\mathbf{X}_{t-r}, r \geq 1\}$; similarly, $\boldsymbol{\sigma}1_t$ is one-step-ahead error when $\mathbf{Y}_t$ is predicted from its own past $\{\mathbf{Y}_{t-r}, r \geq 1\}$. Recall Equations (15), (16) and (17) (the *principle of orthogonality*) in the LSE model in the section titled “Basic LSE Model”. By generalizing the *principle of orthogonality* to infinite orders, to minimize the one-step prediction error, $\boldsymbol{\eta}1_t$ is orthogonal (uncorrelated) to $\sum_{r=1}^{\infty} \mathbf{E}1_r \mathbf{X}_{t-r}$ (the projection of $\mathbf{X}_t$ on its own past), and $\boldsymbol{\sigma}1_t$ is orthogonal (uncorrelated) to $\sum_{r=1}^{\infty} \mathbf{G}1_r \mathbf{Y}_{t-r}$ (the projection of $\mathbf{Y}_t$ on its own past). Thus, $\boldsymbol{\eta}1_t$ and $\boldsymbol{\sigma}1_t$ are by themselves serially uncorrelated. But $\boldsymbol{\eta}1_t$ and $\boldsymbol{\sigma}1_t$ may be correlated with each other contemporaneously and at various leads and lags.

Predict from Its Own Past and the Other’s Past. A further step to take is the projection of $\mathbf{X}_t$ on its own past and the past of $\mathbf{Y}_t$, and the projection of $\mathbf{Y}_t$ on its own past and the

past of $\mathbf{X}_t$, as follows:

$$\mathbf{X}_t = \sum_{r=1}^{\infty} \mathbf{E}2_r \mathbf{X}_{t-r} + \sum_{r=1}^{\infty} \mathbf{F}2_r \mathbf{Y}_{t-r} + \boldsymbol{\eta}2_t, \quad E(\boldsymbol{\eta}2_t) = 0, \text{var}(\boldsymbol{\eta}2_t) = \Gamma 2; \quad (78)$$

$$\mathbf{Y}_t = \sum_{r=1}^{\infty} \mathbf{H}2_r \mathbf{X}_{t-r} + \sum_{r=1}^{\infty} \mathbf{G}2_r \mathbf{Y}_{t-r} + \boldsymbol{\sigma}2_t, \quad E(\boldsymbol{\sigma}2_t) = 0, \text{var}(\boldsymbol{\sigma}2_t) = \Xi 2. \quad (79)$$

By generalizing the *principle of orthogonality* in the section titled “Basic LSE Model” (LSE model) to infinite orders, $\boldsymbol{\eta}2_t$ is orthogonal (uncorrelated) to $\sum_{r=1}^{\infty} \mathbf{E}2_r \mathbf{X}_{t-r} + \sum_{r=1}^{\infty} \mathbf{F}2_r \mathbf{Y}_{t-r}$ (the projection of $\mathbf{X}_t$ on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$), and $\boldsymbol{\sigma}2_t$ is orthogonal (uncorrelated) to $\sum_{r=1}^{\infty} \mathbf{G}2_r \mathbf{Y}_{t-r} + \sum_{r=1}^{\infty} \mathbf{H}2_r \mathbf{X}_{t-r}$ (the projection of $\mathbf{Y}_t$ on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$), to minimize the one-step prediction error. Thus, $\boldsymbol{\eta}2_t$ and $\boldsymbol{\sigma}2_t$ are by themselves serially uncorrelated. But $\boldsymbol{\eta}2_t$ and $\boldsymbol{\sigma}2_t$ may be correlated with each other contemporaneously and at various leads and lags.

The linear prediction model in Equations (77) and (78) is another expression of the linear prediction model in Equation (74). The two expressions are the same intrinsically, considering $\mathbf{V}_t^T = [\mathbf{X}_t^T, \mathbf{Y}_t^T]$.

Therefore, $\boldsymbol{\varepsilon}_t^T = [\boldsymbol{\eta}2_t^T, \boldsymbol{\sigma}2_t^T]$ and we have the partition of γ as follows:

$$\begin{aligned} \gamma &= \text{var}(\boldsymbol{\varepsilon}_t) \\ &= \text{var} \begin{pmatrix} \boldsymbol{\eta}2_t \\ \boldsymbol{\sigma}2_t \end{pmatrix} \\ &= \begin{bmatrix} \Gamma 2 & \text{Cov}_{xy} \\ \text{Cov}_{xy}' & \Xi 2 \end{bmatrix}. \end{aligned} \quad (80)$$

Predict from Its Own Past and the Other’s Present and Past. A third step to take is to pre-multiply the following matrix on Equations (77) and (78)

$$\begin{bmatrix} I_k & -\text{Cov}_{xy} \Xi 2^{-1} \\ -\text{Cov}_{xy}' \Gamma 2^{-1} & I_l \end{bmatrix}$$

Then, the first k equations can be considered as a new linear system of $\mathbf{X}_t$, projected on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 0\}$;

similarly, the next l equations can be considered a new linear system of $\mathbf{Y}_t$, projected on $\{\mathbf{X}_{t-r}, r \geq 0\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$, as follow:

$$\mathbf{X}_t = \sum_{r=1}^{\infty} \mathbf{E}\mathbf{3}_r \mathbf{X}_{t-r} + \sum_{r=0}^{\infty} \mathbf{F}\mathbf{3}_r \mathbf{Y}_{t-r} + \eta\mathbf{3}_t,$$

$$E(\eta\mathbf{3}_t) = 0, \text{var}(\eta\mathbf{3}_t) = \mathbf{\Gamma}\mathbf{3}; \quad (81)$$

$$\mathbf{Y}_t = \sum_{r=0}^{\infty} \mathbf{H}\mathbf{3}_r \mathbf{X}_{t-r} + \sum_{r=1}^{\infty} \mathbf{G}\mathbf{3}_r \mathbf{Y}_{t-r} + \sigma\mathbf{3}_t,$$

$$E(\sigma\mathbf{3}_t) = 0, \text{var}(\sigma\mathbf{3}_t) = \mathbf{E}\mathbf{3}; \quad (82)$$

where

$$\eta\mathbf{3}_t = \eta\mathbf{2}_t - \text{Cov}_{xy} \mathbf{\Xi} \mathbf{2}^{-1} \sigma\mathbf{2}_t, \quad (83)$$

$$\sigma\mathbf{3}_t = \sigma\mathbf{2}_t - \text{Cov}'_{xy} \mathbf{\Gamma} \mathbf{2}^{-1} \eta\mathbf{2}_t. \quad (84)$$

(Hence, $E(\eta\mathbf{3}_t) = E(\eta\mathbf{2}_t) - \text{Cov}_{xy} \mathbf{\Xi} \mathbf{2}^{-1} E(\sigma\mathbf{2}_t) = 0 - \text{Cov}_{xy} \mathbf{\Xi} \mathbf{2}^{-1} \mathbf{0} = 0$, and similarly one can prove that $E(\sigma\mathbf{3}_t) = 0$.)

Note that each of $\eta\mathbf{2}_t$ and $\sigma\mathbf{2}_t$ is uncorrelated (orthogonal) to both $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$ as already discussed above. According to Equations (82) and (83), each of $\eta\mathbf{3}_t$ and $\sigma\mathbf{3}_t$ is uncorrelated to both $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$.

According to Equation (82), $\eta\mathbf{3}_t$ is also uncorrelated to $\sigma\mathbf{2}_t$: $\text{Cov}(\eta\mathbf{3}_t, \sigma\mathbf{2}_t) = E(\eta\mathbf{3}_t \sigma\mathbf{2}_t^T) = E((\eta\mathbf{2}_t - \text{Cov}_{xy} \mathbf{\Xi} \mathbf{2}^{-1} \sigma\mathbf{2}_t) \sigma\mathbf{2}_t^T) = \text{Cov}_{xy} - \text{Cov}_{xy} \mathbf{\Xi} \mathbf{2}^{-1} \mathbf{\Xi} \mathbf{2} = 0$.

Hence, $\eta\mathbf{3}_t$ is uncorrelated to all $\sigma\mathbf{2}_t$, $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$. According to Equation (78), $\eta\mathbf{3}_t$ is uncorrelated to $\mathbf{Y}_t$. Thus, $\eta\mathbf{3}_t$ is uncorrelated to $\{\mathbf{X}_{t-r}, r \geq 1\}$, and $\{\mathbf{Y}_{t-r}, r \geq 0\}$. Therefore, Equation (80) is a linear (orthogonal) projection of $\mathbf{X}_t$ on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 0\}$. Similarly, according to Equation (83), $\sigma\mathbf{3}_t$ is also uncorrelated to $\eta\mathbf{2}_t$: $\text{Cov}(\sigma\mathbf{3}_t, \eta\mathbf{2}_t) = E(\sigma\mathbf{3}_t \eta\mathbf{2}_t^T) = E((\sigma\mathbf{2}_t - \text{Cov}'_{xy} \mathbf{\Gamma} \mathbf{2}^{-1} \eta\mathbf{2}_t) \eta\mathbf{2}_t^T) = \text{Cov}'_{xy} - \text{Cov}'_{xy} \mathbf{\Gamma} \mathbf{2}^{-1} \mathbf{\Gamma} \mathbf{2} = 0$.

Hence, $\sigma\mathbf{3}_t$ is uncorrelated to all $\eta\mathbf{2}_t$, $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$. According to Eq. 77, $\eta\mathbf{3}_t$ is uncorrelated to $\mathbf{X}_t$. Thus, $\sigma\mathbf{3}_t$ is uncorrelated to $\{\mathbf{X}_{t-r}, r \geq 0\}$, and $\{\mathbf{Y}_{t-r}, r \geq 1\}$. Therefore, Equation (81) is a linear (orthogonal) projection of $\mathbf{Y}_t$ on $\{\mathbf{X}_{t-r}, r \geq 0\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$.

Predict from Its Own Past and the Other's Temporally Complete Information.

A final step is to consider the linear projection of $\mathbf{X}_t$ on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_{t-r}, \text{integer } r \in (-\infty, \infty)\}$, and the linear projection of $\mathbf{Y}_t$ on $\{\mathbf{X}_{t-r}, \text{integer } r \in (-\infty, \infty)\}$ and $\{\mathbf{Y}_{t-r}, r \geq 1\}$.

Recall the partition of the spectral density matrix $\mathbf{S}_V(\omega)$ in Equation (70). Let

$$\tilde{L}(\omega) = \mathbf{S}_{\mathbf{X}\mathbf{Y}}(\omega) \mathbf{S}_{\mathbf{Y}}(\omega)^{-1} \quad \text{for all } \omega \in [-\pi, \pi] \quad (85)$$

(Assume that Equation (73) is true; then, the spectral density matrix exists and is bounded uniformly away from zero almost everywhere in $[-\pi, \pi]$) Use inverse Fourier transform of $\tilde{L}(\omega)$:

$$L_r = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{L}(\omega) e^{-i\omega r} d\omega. \quad (86)$$

Let

$$\mathbf{W}_t = \mathbf{X}_t - \sum_{r=-\infty}^{\infty} L_r \mathbf{Y}_{t-r}. \quad (87)$$

Then, from the spectral representation of $\mathbf{V}$ (Eq. 70), it is be proved that $\mathbf{W}_t$ is uncorrelated (orthogonal) with all $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$.

Therefore, if we rewrite Equation (86) in $\mathbf{X}_t = \sum_{r=-\infty}^{\infty} L_r \mathbf{Y}_{t-r} + \mathbf{W}_t$, then Equation (86) can be considered as an expression of the linear projection of $\mathbf{X}_t$ on $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$.

From Equations (70) and (86), it can be further shown that

$$\mathbf{S}_W(\omega) = \mathbf{S}_X(\omega) - \mathbf{S}_{\mathbf{X}\mathbf{Y}}(\omega) \mathbf{S}_{\mathbf{Y}}(\omega)^{-1} \mathbf{S}_{\mathbf{Y}\mathbf{X}}(\omega) \quad \text{for all } \omega \in [-\pi, \pi]. \quad (88)$$

Considering the assumption in Equation (73): because $\mathbf{S}_W(\omega)$ consists of the first k rows and columns of $\mathbf{S}_V(\omega)$, and Equation (73) holds, we also have

$$\text{CONST}^{-1} I_k \leq \mathbf{S}_W(\omega) \leq \text{CONST} I_k. \quad (89)$$

(For some constant CONST and for all $\omega \in [-\pi, \pi]$ in the above equation)

Therefore, $\mathbf{W}_t$ has an autoregressive representation, as follows:

$$\mathbf{W}_t = \sum_{r=1}^{\infty} D_r \mathbf{W}_{t-r} + \boldsymbol{\eta}_t, \\ E(\boldsymbol{\eta}_t) = 0, \text{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Gamma} \mathbf{4}, \quad (90)$$

where $(\boldsymbol{\eta}_t)$ is uncorrelated (orthogonal) with $\{\mathbf{W}_{t-r}, r \geq 1\}$.

Using $\mathbf{X}_t = \sum_{r=-\infty}^{\infty} L_r \mathbf{Y}_{t-r} + \mathbf{W}_t$ from Equation (86) to substitute all $\{\mathbf{W}_{t-r}, r \geq 1\}$ in Equation (89), after reorganizing, we get

$$\mathbf{X}_t = \sum_{r=1}^{\infty} \mathbf{E} \mathbf{4}_r \mathbf{X}_{t-r} + \sum_{r=-\infty}^{\infty} \mathbf{F} \mathbf{4}_r \mathbf{Y}_{t-r} + \boldsymbol{\eta}_t, \\ E(\boldsymbol{\eta}_t) = 0, \text{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Gamma} \mathbf{4}. \quad (91)$$

By Equation (89), $\boldsymbol{\eta}_t$ is a linear function of $\{\mathbf{W}_{t-r}, r \geq 1\}$, and since any $\{\mathbf{W}_r, \text{integer } r \in (-\infty, \infty)\}$ is uncorrelated (orthogonal) with all $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$, therefore, $\boldsymbol{\eta}_t$ is uncorrelated with all $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$.

According to Equation (86), $\mathbf{X}_s$ is a linear function of $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$ and $\mathbf{W}_s$ for any integer s . Also, $\boldsymbol{\eta}_t$ is uncorrelated (orthogonal) with $\{\mathbf{W}_{t-r}, r \geq 1\}$, as already discussed above. Therefore, $\boldsymbol{\eta}_t$ is uncorrelated with $\{\mathbf{X}_{t-r}, r \geq 1\}$.

Hence, Equation (90) is the linear projection of $\mathbf{X}_t$ on $\{\mathbf{X}_{t-r}, r \geq 1\}$ and $\{\mathbf{Y}_r, \text{integer } r \in (-\infty, \infty)\}$.

Similarly (symmetrically), Equation (91) is the linear projection of $\mathbf{Y}_t$ on $\{\mathbf{Y}_{t-r}, r \geq 1\}$ and $\{\mathbf{X}_r, \text{integer } r \in (-\infty, \infty)\}$.

$$\mathbf{Y}_t = \sum_{r=-\infty}^{\infty} \mathbf{H} \mathbf{4}_r \mathbf{X}_{t-r} + \sum_{r=1}^{\infty} \mathbf{G} \mathbf{4}_r \mathbf{Y}_{t-r} + \boldsymbol{\eta}_t, \\ E(\boldsymbol{\sigma}_t) = 0, \text{var}(\boldsymbol{\sigma}_t) = \boldsymbol{\Xi} \mathbf{4}. \quad (92)$$

It is worth mentioning that all the prediction models in the section titled “Causal Relations between Two, Three, and Multiple Variables in Prediction” are based on the assumption that the time series analyzed are a zero-mean wide-sense stationary, and purely nondeterministic. The analysis could be extended to some nonstationary and

purely nondeterministic time series, and similar autoregressive representations can be obtained.

OTHER PREDICTION METHODS

To handle nonstationarity and seasonality, the prediction process needs to select a suitable model for a given time series. The prediction based on the more general class of ARIMA or seasonal autoregressive integrated moving average (SARIMA) models is called the *Box–Jenkins prediction* [14]. The model involves an iterative procedure with: (i) formulating, (ii) fitting, (iii) checking and adjusting.

Depending on how the first few observations are treated, there exist different prediction procedures for fitting ARIMA models (e.g., maximum likelihood and conditional least square). But only for short series (not more than several hundred points), the choice of procedures is important and makes a significant difference. It is shown that for short series even when asymptotically unbiased, parameter estimates are likely to be biased [18]. Furthermore, different software packages, using different estimation routines, can produce model parameter estimates with nontrivial differences [19]. Hence, it is a wise choice to use the software which exactly specifies what estimation procedures are adopted.

The way that trend is removed before applying the Box–Jenkins approach can be vital for nonstationary time series. Especially, the order of differencing can be crucial if differencing is applied. Alternative methods of removing trend, prior to fitting an ARIMA model, may lead to better forecasting [20].

CONCLUDING REMARKS

In this article, we discuss the *rolling method*, which is a projection based on past performances, and which is periodically updated on a regular schedule to incorporate data (e.g., LSE). Especially, in the LSE model, the orthogonality principle is discussed, which

is widely used in different prediction models. There are also derivative models of LSE, which intrinsically keep the same orthogonality principle, that is, the total least squares (TLS), the robust least squares (RLS), and the ordinary least squares (OLS).

We also discuss the causal models, which are under the assumption that it is possible to identify the underlying factors that might influence the variable that is being predicted (e.g., ARMA; ARIMA; and Granger causality). The Granger causality basically is studying the causal relations between/among two or multiple time series, and is widely used in neuroscience, cognitive research, and economics. The predicting power of Granger causality models can be explored if all the time series can be considered as in one system. We further categorized and discussed Granger causality against two time series, three time series, and more generally, two groups of time series. In each of those cases, there are some results, which can benefit the prediction.

There are other categories of prediction methods that are not included here in this article, like the ensemble prediction, simulation methods, or the judgmental methods incorporating intuitive judgments, opinions, and subjective probability estimates (e.g., composite prediction; Delphi method; and scenario building) [21–23]. Different models have their respective limitations and advantages in difference scenarios.

REFERENCES

1. Stevenson H, editor. *Do lunch or be lunch*. Boston (MA): Harvard Business School Press; 1998.
2. Granger CW. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969; 37(3): 424–438.
3. John G. Measurement of linear dependence and feedback between multiple time series. *J Am Stat Assoc* 1982; 77(378): 304–313.
4. Makhoul J. Linear prediction: a tutorial review. *Proc IEEE* 1975; 63: 561–580.
5. Markel JD, Gray AH Jr. *Linear prediction of speech*. New York: Springer; 1976.
6. Stewart GW. *Introduction to matrix computations*. New York: Academic Press; 1973.
7. Miller KS. *Complex stochastic processes: an introduction to theory and application*. Reading (MA): Addison-Wesley; 1974.
8. Goodwin GC, Payne RL. *Dynamic system identification: experiment design and data analysis*. New York: Academic Press; 1977.
9. Hanson RJ. *Solving least squares problems*. SIAM; 1995.
10. Hayes MH. *Statistical digital signal processing and modeling*. John Wiley & Sons Inc.; 2008.
11. Hamilton JD. *Time series analysis*. Princeton (NJ): Princeton University Press; 1994. (1.5, 3.1, 3.4, 4.1, 5.2, 5.4)
12. Priestley MB. *Spectral analysis and time series*. London: Academic Press; 1981. Chapter 1–5, 10.
13. Whittle P. *Prediction and Regulation*. 2nd ed Minneapolis (MN): University of Minnesota Press; 1983. Chapter 4. Revised.
14. Box GEP, Jenkins GM, Reinsel GC. *Time series analysis, forecasting and control*. 3rd ed. Englewood Cliffs (NJ): Prentice-Hall; 1994. (Chapter 1–5, 7, 8).
15. Shumway RH, Stoffer DS. *Time series analysis and its applications*. Springer; 2000. Chapter 2.
16. Doob J. *Stochastic process*. New York: John Wiley; 1953.
17. Rozanov. 1967.
18. Ansley CF, Newbold P. Finite sample properties of estimators for autoregressive moving average models. *J Econom* 1980; 13: 159–183.
19. Newbold P, Agiakloglou C, Miller J. Adventures with ARIMA software. *Int J Forecast* 1994; 10: 573–581.
20. Makridakis S, Hibon M. ARMA models and the Box-Jenkins methodology. *J Forecast* 1997; 16: 147–163.
21. Chatfield C. *Time-series forecasting*. Chapman & Hall/CRC Press; 2001.
22. Makhoul J. Stable and efficient lattice methods for linear prediction. *IEEE Trans Acoust Speech Signal Process* 1977; ASSP-25: 423–428.
23. Gonzalez RC, Woods RE. *Digital image processing*. Prentice Hall; 2007. Chapter 5.
24. Simon H. *Adaptive filter theory*. New York: Prentice Hall Press; 2002.

MINIMUM SPANNING TREES

ASHISH K. NEMANI

Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

RAVINDRA K. AHUJA

Innovative Scheduling, Inc., and
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

Most of the well-known network-based combinatorial optimization problems (traveling salesman, shortest paths, matching, etc.) have spanning trees as their core structure. Several exact and approximate algorithms for these problems repeatedly solve minimum spanning tree (MST) instances. The fast and efficient polynomial-time algorithms existing for MST prove its effectiveness as a subroutine in solving these problems. Given an undirected graph $G = (N, A)$ with $n = |N|$ nodes and $m = |A|$ arcs and a weight w_{ij} associated with each arc $(i, j) \in A$, a spanning tree (T) is a connected acyclic subgraph that spans all the nodes. An MST (M) is the spanning tree of G with the minimum total weight of all constituent arcs. Figure 1(a) gives an example of an MST having a total weight of 55 represented by bold arcs (A_M). In this article, without loss of generality, we assume that all arc weights are distinct and as in all discussed properties and algorithms, ties can be broken arbitrarily.

The MST problem was first formally defined by Borůvka [1] in 1926 and since then it has drawn the attention of researchers. It arises in many applications either directly or indirectly. Some direct applications include designing computer and communication networks, power and leased-line telephone networks, piping in a flow network, wiring connections, and so on. In these problems, we aim to ensure a

viable path between each pair of nodes so that the total cost is minimized. Indirect applications include the all-pairs shortest path problem, network reliability, clustering and classification problems, and so on. In these applications, we either model the problem as an MST or try to find a substructure where its properties are utilized. We briefly explain some of these applications in the next section.

APPLICATIONS

Designing Physical Systems

Designing a physical system can be very complex, involving the interplay between multiple factors such as system performance, cost, and available technology. The major criterion in designing a network is usually either to physically connect geographically dispersed components, or to provide a channel for communication between the users. In most of these settings, the system needs to have the simplest possible connection without including any redundancies. This simple connection has the structure of a spanning tree, and it arises in numerous problem scenarios including the building of certain types of highways, computer networks, leased-line telephone networks, railroads, the installation of cable television lines or high-voltage electrical power transmission lines. Some typical problem settings are listed below in which spanning tree finds its application; however, more customized and problem-specific models can be developed for each scenario:

- when connecting terminals in cabling the panels of electrical equipment, the design should utilize the shortest possible length of wire;
- when constructing a pipeline network to connect a number of cities using the smallest possible length of pipeline;
- when linking isolated places in remote regions that are connected by roads,

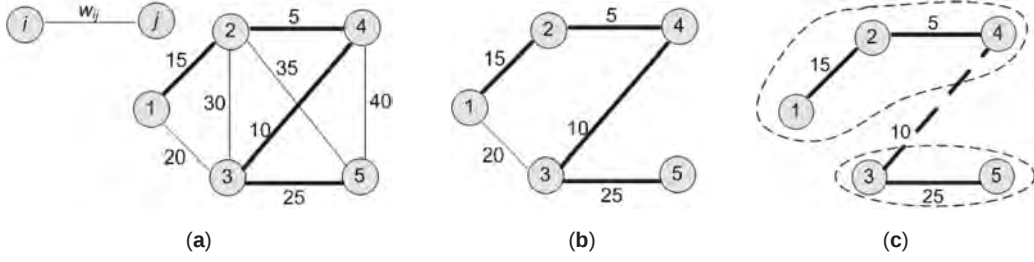


Figure 1. Illustrating properties of spanning trees: (a) MST with $n - 1$ arcs; (b) adding arc $(1, 3)$ to the spanning tree forms a unique cycle $1-2-4-3-1$; (c) deleting arc $(3, 4)$ forms the cut $[S, \bar{S}]$ with $S = [1, 2, 4]$.

but not yet by telephone lines, we aim to determine along which stretches of roads we should place the telephone lines so that all areas are connected using the least possible total road miles;

- when constructing a digital computer system composed of high frequency circuitry, it is important to minimize the lengths of the wires between different components to reduce both capacitance and delay line effects and since the system must connect all the components, it reduces to a spanning tree problem;
- when connecting a number of computer sites by high speed lines, each line is available for leasing at a certain monthly cost, and we wish to determine a configuration that connects all the sites at the least possible cost.

Min-Max Paths

In a graph $G = (N, A)$ with arc weights w_{ij} , we define the value of a path P between a pair $[p, q]$ of nodes as the maximum weight arc in P . The min-max path problem determines the minimum value path from node p to node q . There are several real-life instances in which we solve all-pairs min-max path problems, determining the min-max paths for each pair of nodes. In designing the tracks or roads, one of the most critical factors is the change in elevation. In mountainous regions, this becomes even more important when determining the alignment of tracks. The change in elevation is very costly, as it increases the wear and tear associated with

braking. Besides, only trains with enough traction power can run in the regions with very high elevations. Therefore, we would like to find the paths between each pair of train stations that minimize the maximum elevation change. There are several other examples in which we solve the all-pairs min-max path problems: (i) in designing the trajectory of a spacecraft, we try to keep the maximum temperature to which its surface is exposed as low as possible; (ii) in designing wheelchair-accessible paths, we try to minimize the maximum ascent path segments; and (iii) in traveling through a desert, we want to minimize the maximum stretch between rest areas.

The all-pairs min-max paths problems can be easily solved by simply solving an MST instance. Let M be an MST of G , and P denote the unique path in M between a node pair $[p, q]$ with (k, l) as the maximum weight arc. Therefore, the value of path P is w_{kl} . Deleting the arc (k, l) divides the nodes into two subsets forming a cut $[S, \bar{S}]$. For example, in Fig. 1(c), deleting the arc $(3, 4)$ generates the cut $[S, \bar{S}]$, where $S = \{1, 2, 4\}$ and $\bar{S} = \{3, 5\}$. According to the strong cut property of an MST (discussed in the section titled “Properties of Minimum Spanning Trees”), the arc (k, l) is the minimum weight arc among all arcs in the cut $[S, \bar{S}]$. Then, consider any path P' between the node pair $[k, l]$ in G . It must contain an arc in $[S, \bar{S}]$ and, therefore, its value is at least w_{kl} . Because the path P in M is of value w_{kl} , it is the min-max path. This observation proves that the unique path between any pair $[k, l]$ of nodes is the min-max path in G .

Optimal Message Passing

An intelligence service agency has n agents located in an unfriendly country. Each agent knows a group of other agents and regularly passes messages to them using a special procedure. Each message between the agent pair (k, l) will fall into the hands of opponents with the probability of p_{kl} . These messages are very confidential and need to be passed with the minimum possible probability of being intercepted.

We can easily transform this problem into an MST instance. Consider each agent as a node and the medium of the message passing as an arc in the resulting graph G' . We need to determine a spanning tree T in G' that minimizes the probability of interception given by the expression $\{1 - \prod_{(k,l) \in T} (1 - p_{kl})\}$. Alternatively, we find a tree T that maximizes $\{\prod_{(k,l) \in T} (1 - p_{kl})\}$, or equivalently that maximizes $\{\sum_{(k,l) \in T} \log(1 - p_{kl})\}$. Therefore, it reduces to the problem of finding an MST in G' with an arc of weight $(-\log(1 - p_{kl}))$ between each node pair $[k, l]$. We add the negative sign to make the arc weights positive.

Clustering

Clustering is to partition a set of data into subsets or cluster such that the data points within a particular cluster should be more closely related to each other than the data points not in that cluster. Cluster analysis is important in a variety of disciplines such as medical and manufacturing that rely on empirical investigations. Suppose we need to find a partition of a set of n points in two-dimensional Euclidean space into clusters. Kruskal's algorithm explained in the following section is one of the most popular algorithms to solve this problem. As explained later, at each intermediate step, Kruskal's algorithm maintains a collection of node-disjoint trees and adds arcs in nondecreasing order of their lengths. The nodes spanned by these trees can be regarded as different clusters, which often provide excellent solutions for the clustering problem. We can also start from the MST and delete k arcs to find $k + 1$ clusters. The best partition can be found by simple visualization or by

defining an appropriate objective function. A good choice of an objective function depends on the underlying features of a particular clustering application. This analysis is not limited to points in two-dimensional space, and can be easily extended to multidimensional space by defining interpoint distances appropriately.

PROPERTIES OF MINIMUM SPANNING TREES

Some properties of spanning trees that are frequently used in the development of algorithms can be easily observed in Fig. 1: (i) each spanning tree has exactly $n - 1$ arcs; (ii) for every non-tree arc $(i, j) \in A/A_M$, the spanning tree T contains a unique path from node i to node j . Adding the arc (i, j) creates a cycle with this unique path; and (iii) deleting any arc (i, j) of the spanning tree partitions the graph G into two subsets.

These observations give rise to the two most important properties of MST, the building blocks of most of the algorithms: (i) strong cycle property and (ii) strong cut property. These properties are also referred to as *optimality conditions* for MST.

Theorem 1 [Strong Cut Property]. *A spanning tree M is an MST if and only if every MST arc $(i, j) \in A_M$ is the minimum weight arc in the cut formed by deleting the arc (i, j) .*

Proof. It is easy to show that every MST should satisfy this property. For, if $w_{ij} > w_{kl}$ and arc (k, l) is contained in the cut formed by deleting the arc (i, j) from M , then introducing arc (k, l) in place of arc (i, j) would create an MST of total weight less than M , contradicting the optimality of M . In Fig. 1(c), arc $(3, 4)$ has the minimum weight among all arcs in the cut formed by deleting it.

We now prove that, any spanning tree T^* , satisfying the property, is an MST. Suppose M is an MST and $T^* \neq M$. Therefore, T^* must contain an arc (i, j) that is not in M . Deleting

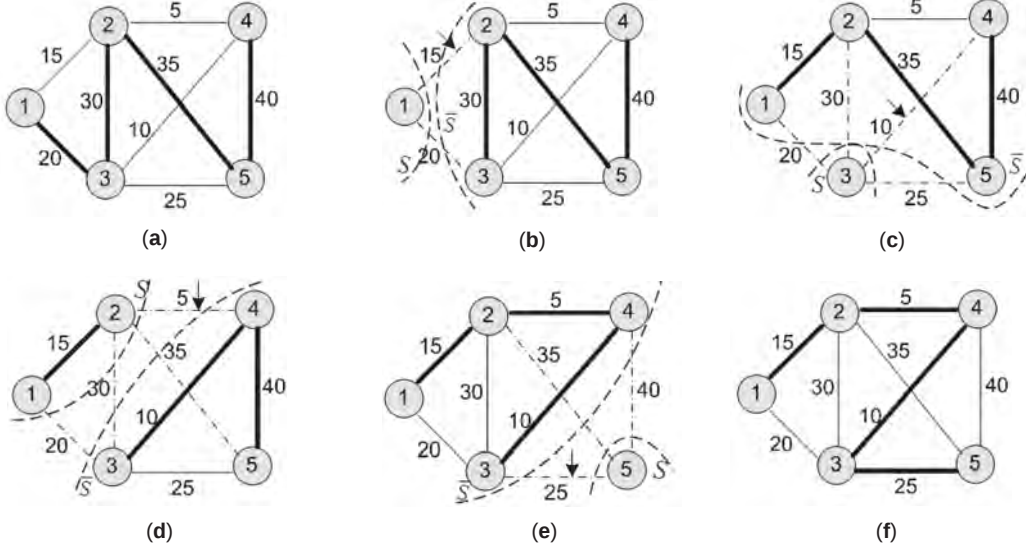


Figure 2. Finding MST using the strong cut property.

arc (i, j) from T^* forms the cut $[S, \bar{S}]$. Now, we add the arc (i, j) in M creating a unique cycle containing an arc (k, l) other than arc (i, j) , with $k \in S$ and $l \in \bar{S}$. As T^* satisfies strong cut property, $w_{ij} \leq w_{kl}$. However, M is an MST, so $w_{ij} \geq w_{kl}$, otherwise we can easily reduce the total weight by replacing the arc (k, l) by the arc (i, j) . Now, we introduce the arc (k, l) in T^* in place of the arc (i, j) , increasing the similarity between T^* and M by one more arc. By repeatedly replacing the uncommon arcs in T^* , we can transform it into the MST M , showing that T^* is also an MST. This completes the proof.

The strong cut property is utilized in deriving a simple algorithm that can transform any spanning tree T into an MST in a finite number of steps. We satisfy the property by investigating each cut one at a time, and appropriately deleting and adding arcs into the tree. Figure 2(a)–(f) illustrates the transformation of an arbitrary tree to an MST. The highlighted arcs represent the spanning tree at each stage and the arrow points to the minimum weight arc in the cut $[S, \bar{S}]$. Figure 2(a) shows a spanning tree of weight

125 and in the next four iterations, the arc with minimum weight in the cut being examined replaces the arc violating the strong cut property. We reach the MST of weight 55 in a finite number of steps, but the running time is not very attractive. However, it builds the foundation of Kruskal's algorithm [2], discussed in the following section, which improves the complexity by systematically analyzing each cut.

Theorem 2 [Strong Cycle Property]. *A spanning tree M is an MST if and only if every non-tree arc $(i, j) \in A/A_M$ is the maximum weight arc in the cycle, uniquely created by introducing the arc (i, j) in M .*

The proof is similar to Theorem 1, and details can be found in Ref. 3.

We can also derive a very simple algorithm using the strong cycle property. We start from an arbitrary spanning tree, and then verify the strong cycle property for each cycle created by adding a non-tree arc to the spanning tree. If any cycle does not satisfy the property, we replace the existing arc with the lower weight arc present in the cycle and thereby reduce the weight of the tree. This illustration is shown in Fig. 3(a)–(f) starting with the

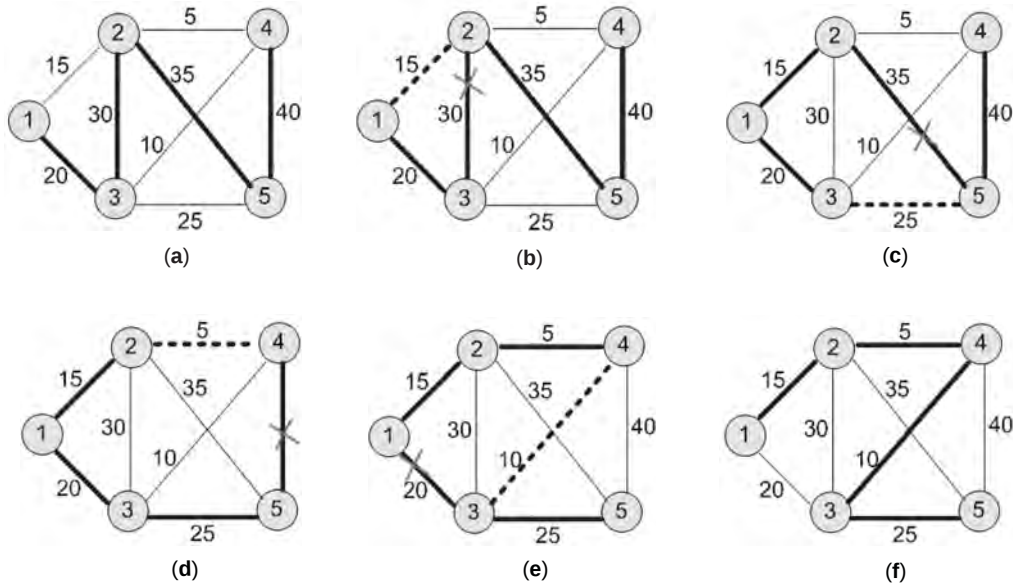


Figure 3. Finding MST using the strong cycle property.

same spanning tree considered in Fig. 2(a). In the next four iterations, we create a cycle by adding a non-tree arc and then reducing the total weight by deleting the arc of maximum weight among all arcs in the cycle. Although we finally achieve the MST, the steps of creating cycles and deleting maximum weight arcs are not time efficient. Prim [4] utilizes these properties to develop an efficient algorithm, which is discussed in a subsequent section.

We have described two fundamental properties that form the basis of most of the MST algorithms. In this article, we present two classical algorithms in detail, with an overview of recent advances: (i) Kruskal's algorithm and (ii) Prim's algorithm.

KRUSKAL'S ALGORITHM (1956)

Kruskal's algorithm is based on the strong cycle property of MSTs. We generate an MST from scratch by maintaining a forest of trees at each step. The basic algorithm is given below:

Step 1. Sort all arcs in the nondecreasing order of their weights.

Step 2. Initialize the list of arcs, *CURRENT*, which contains the current set of arcs in MST to *NULL*.

Step 3. Examine arcs in their sorted order one at a time; add an arc to the *CURRENT* list if it does not create a cycle with the arcs already present in it. Examining the arcs in their sorted order helps in maintaining the strong cycle property.

Step 4. Terminate the algorithm when the cardinality of the list becomes $n - 1$; that is $|CURRENT| = n - 1$.

We illustrate Kruskal's algorithm in Fig. 4 by running it on the example given in Fig. 1(a). We also explain how the strong cycle property forms the base of the algorithm. As given in Step 1, we first sort all arcs in the nondecreasing order of their weights: $\{(2, 4), (3, 4), (1, 2), (1, 3), (3, 5), (2, 3), (3, 4), \text{ and } (4, 5)\}$. In Fig. 4(a) there are five disconnected sets, each containing one node. In the first three iterations, the algorithm adds the arcs $(2, 4)$, $(3, 4)$, and $(1, 2)$

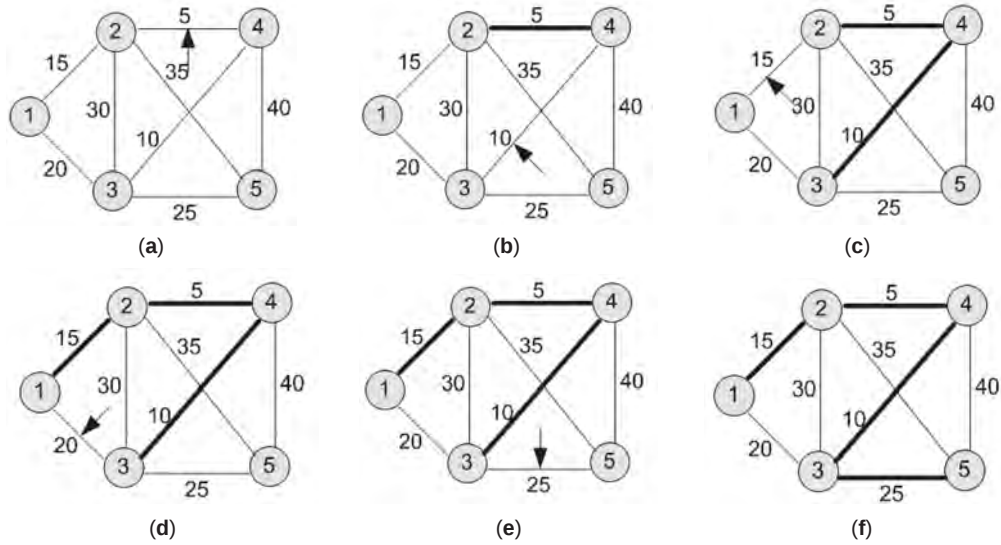


Figure 4. Illustration of Kruskal's algorithm.

to *CURRENT*, respectively, as shown in Fig. 4(b)–(d). As we examine them in sorted order, each newly added arc has a lower weight than the maximum weight arc among all the possible cycles, which contain nodes that it is joining. For example, in Fig. 4(b), arc (2, 4) has a weight lower than the maximum weight arc in any cycle containing nodes 2 and 4. In the next iteration, the algorithm examines and discards arc (1, 3) as it creates a cycle. It can be observed that adding arc (1, 3) violates the strong cycle property, as it has the maximum weight in the cycle 1-2-4-3. Finally, the arc (3, 5) is added to *CURRENT* and the algorithm terminates as the cardinality becomes four. It can be observed that at each stage the strong cycle property is maintained. This forms the basis of the proof of correctness.

The correctness of Kruskal's algorithm follows from the fact that at each step we include the minimum weight arc in the tree, while discarding those arcs that create a unique cycle. As arcs are examined in the nondecreasing order of their weights, any discarded arc has a maximum weight among all arcs already present in the cycle formed by its addition.

Complexity Analysis

Kruskal's algorithm maintains a forest of trees in *CURRENT* at each step. For example, the set *CURRENT* that corresponds to Fig. 2(c) comprises three trees containing nodes {1}, {2, 3, 4}, and {5}. At each step, a new arc is put in *CURRENT* if it connects distinct trees (i.e., no cycle). Hence, the algorithm requires repetitive applications of two basic operations, commonly known as *union-find* operations [5]. The *find* operation returns the lists (trees) to which an arc belongs, and the *union* operation merges the two lists (trees). Several data structures have been proposed for these operations having different complexities.

We first discuss the implementation of Kruskal's algorithm, which runs in $O(m + n \log n)$ time, besides taking $O(m \log m) = O(m \log n^2) = O(m \log n)$ time for sorting all arcs. We maintain a set of singly linked lists, $T_1, T_2, \dots$, each containing the nodes of the tree in the forest maintained by the algorithm. We also maintain the cardinality and the last node of each list L in $card(L)$, and $last(L)$, respectively. We associate an index $first(i)$ for each node i , which records the first node of the list

containing the node i . For example, the list T_2 stores the second tree of nodes $\{2, 3, 4\}$ in Fig. 2(c): $\text{card}(T_2) = 3$; $\text{last}(T_2) = 4$; and $\text{first}(2) = \text{first}(3) = \text{first}(4) = 2$.

Each *find* operation takes $O(1)$ time using this data structure. To check any arc (k, l) , we compare $\text{first}(k)$ and $\text{first}(l)$. If they are equal, k and l belong to the same tree and the arc (k, l) cannot be added; otherwise, they belong to different trees. In a worst-case scenario, all arcs are checked, requiring a total of $O(m)$ time. If the arc (k, l) is to be added, we need to merge two trees: $T_1 = \text{find}(k)$ and $T_2 = \text{find}(l)$. Without a loss of generality, assume that $\text{card}(T_1) > \text{card}(T_2)$. While merging, we put the larger list T_1 first and append the smaller list T_2 with the last node [determined by $\text{last}()$] of T_1 . Thereafter, we update the (i) cardinality of the new list requiring $O(1)$ time; (ii) last node of the new list requiring $O(1)$ time; and (iii) first-node-index of all nodes that were originally contained in T_2 requiring $O(\text{card}(T_2))$. Hence, each union operation takes $O(h)$ time, where $h = \text{card}(T_2)$. It can be easily proven by induction that the total time taken by all union operations to create a list of size n in the worst case is $O(n \log n)$. Therefore, Kruskal's algorithm takes $O(m)$ time for all *find* operations, $O(n \log n)$ time for all *union* operations, and $O(m \log n)$ for sorting all arcs summing up to $O(m + (m + n) \log n) = O(m \log n)$.

With the development of new data structures, several authors have significantly improved the running time of the classical or modified Kruskal's algorithm [6]. The Fliter–Kruskal algorithm [7] reduces the running time to $O(m + n \log n \log m/n)$, which is very close to linear for a not very sparse graph. We briefly discuss some more algorithms in the section titled “Recent Algorithms.” In the next section, we describe another extensively studied algorithm, which was first developed by Jarnik [8] in 1930, and later independently by Prim [4] in 1957, and by Dijkstra [9] in 1959.

Therefore, it is known as Jarnik–Prim's, Prim's, or Dijkstra–Jarnik–Prim's (DJP) algorithm.

PRIM'S ALGORITHM (1930/1957/1959)

Prim's algorithm is based on the strong cut property of MSTs. Similar to Kruskal's algorithm, we build an MST starting from a single node and adding one arc at a time. However, in Prim's algorithm, a single tree T spanning on a subset S of nodes is maintained instead of a forest. The basic algorithm is given below:

Step 1. Initialize a set S and the current spanning tree T with node 1 (arbitrarily chosen). Initialize set $\bar{S}$ by inserting all nodes except $\{1\}$ in it, that is, $S = \{1\}$ and $\bar{S} = N \setminus \{1\}$.

Step 2. Find the arc (i, j) with the minimum weight among all arcs in the cut $[S, \bar{S}]$ such that $i \in S$ and $j \in \bar{S}$. Add arc (i, j) in tree T and move node j from set $\bar{S}$ to set S . This step directly follows from the strong cut property of an MST.

Step 3. If $|S| = n$, T is an MST, stop; otherwise, go to Step 2.

We illustrate Prim's algorithm in Fig. 5 on the same example, shown in Fig. 1(a), which was used earlier to explain Kruskal's algorithm. As given in Step 1, set $S = \{1\}$ and $\bar{S} = \{2, 3, 4, 5\}$. The cut set $[S, \bar{S}]$ contains two arcs $(1, 2)$ and $(1, 3)$ [see Fig. 5(b)]. The algorithm selects the arc having the least weight, that is arc $(1, 2)$. Now, set $S = \{1, 2\}$, T contains the arc $(1, 2)$, and the cut set $[S, \bar{S}]$ comprises arcs $(1, 3)$, $(2, 3)$, $(2, 4)$, and $(2, 5)$. In the next step, the algorithm selects the arc $(2, 4)$ because it has the minimum weight and updates all sets [see Fig. 2(c)]. In the next two iterations, as shown in Fig. 2(d) and (e), the algorithm chooses the arcs $(3, 4)$ and $(3, 5)$, respectively. The final MST produced

by the algorithm is shown in Fig. 2(f). It can be easily observed that each newly added arc has the minimum weight among all arcs in the cut that was obtained by deleting that arc. For example, in Fig. 5(c), arc (2, 4) has the lowest weight among all arcs, $\{(1, 3), (2, 3), (2, 5), \text{ and } (2, 4)\}$, joining the cut $\{1, 2\}$ and $\{3, 4, 5\}$.

The correctness of the algorithm can be shown through induction by proving that at each step the spanning tree T is an MST of nodes present in the set S . At the start, T contains only one node, so it is a trivial MST. Assume that T is an MST just before adding a new arc (k, l) , where $k \in S$ and $l \in \bar{S}$, and a final MST (M) is consistent with T so far (i.e., A_T is a subset of A_M). If the arc (k, l) is in M , new T is obviously an MST; otherwise, introducing it in M forms a unique cycle W with some nodes of S and some nodes of $\bar{S}$. The cycle W must contain one arc (p, q) connecting set S and $\bar{S}$ other than arc (k, l) . From Prim's algorithm, we know that $w_{kl} \leq w_{pq}$, and because M is an MST we have $w_{kl} \geq w_{pq}$, together implying that $w_{kl} = w_{pq}$. Therefore, T is a tree with exactly the same total weight as M containing the same set

of nodes. It proves the inductive hypothesis showing the correctness of Prim's algorithm.

Complexity Analysis

Prim's algorithm performs $n - 1$ iterations of Step 2 of the basic algorithm, each adding one arc at a time to the current tree until it becomes an MST containing $n - 1$ arcs. In each iteration, the selected arc has the minimum weight in the cut $[S, \bar{S}]$. A simple yet expensive way of selecting the minimum weight arc is to scan the entire set of arcs in $O(m)$ time. Therefore, the total running time of this algorithm is $O(nm)$. We can substantially improve the running time by efficiently choosing the entering arc (minimum weight arc) using various types of *heaps* (a specialized tree-based data structure), such as the binary heap and Fibonacci heap. A binary heap (Fig. 6) is created using a binary tree with two main properties: (i) it is almost a complete binary tree, that is all levels of the tree except the last one must be fully filled, and if the last level is not complete then the nodes of that level are filled from left to right and (ii) each node's value is greater than or equal to that of each of its children. We briefly

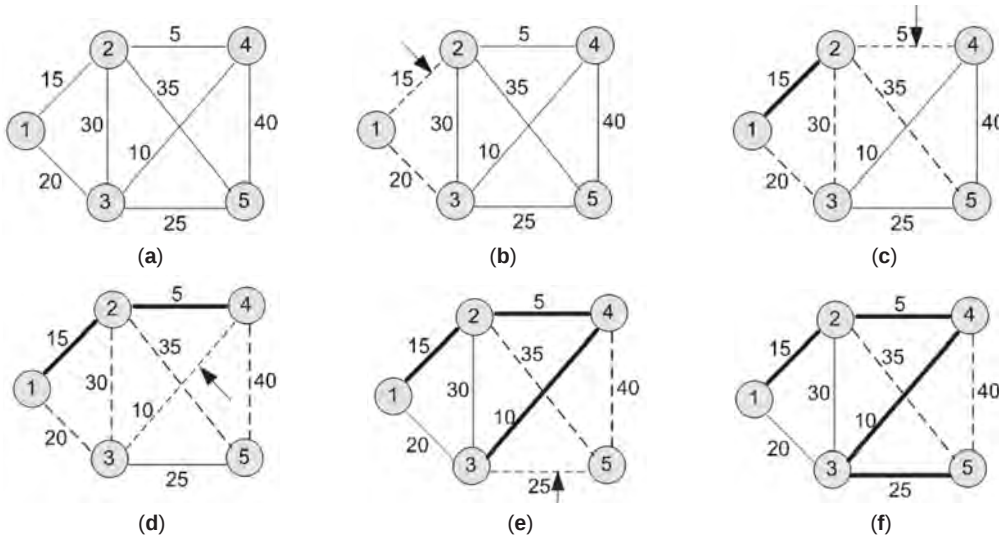


Figure 5. Illustration of Prim's algorithm.

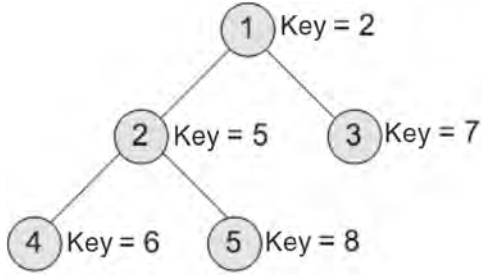


Figure 6. A binary heap with five elements.

explain the operations supported by a heap containing a collection H of objects and their use in Prim's algorithm (Fig. 7).

create-heap(H): Creates an empty heap H .

find-min(H): Returns an object of minimum key from H . Each object corresponds to a node j and its key equals the distance label $d(j)$. It facilitates choosing the minimum weight arc in the cut

$[S, \bar{S}]$, which is the bottleneck element of Prim's algorithm. The distance label $d(j)$ represents the minimum weight of all arcs in the cut $[S, \bar{S}]$ incident to any node j in $\bar{S}$ from any node in S , that is $d(j) = \min\{w_{ij} : (i, j) \in [S, \bar{S}]\}$. For example, in Fig. 5(d), $d(3) = 10$ as among all arcs in the cut $[S, \bar{S}]$ incident to node 3, $\{(1, 3), (2, 3), \text{ and } (4, 3)\}$, the arc $(4, 3)$ has the minimum weight of 10.

insert(H): Inserts a new object i with a predefined key into H .

decrease-key($i, value, H$): Reduces the key of object i to $value$. As new nodes are inserted into set S , the distance labels may decrease due to a new set of arcs in the cut $[S, \bar{S}]$. For example, in Fig. 5(c), $d(3) = 20$ and after inserting node 4 in S , $d(3)$ decreases to 10.

delete-min(i, H): Deletes an object i with minimum key from H . It corresponds to

```

algorithm Prim
begin
    create heap( );
    for each  $j \in N - \{1\}$  do  $d(j) := C + 1$ ; //  $C$  is the maximum arc weight in graph  $G$ .
    set  $d(1) := 0$ ; and  $pred(1) := 0$ ;
    for each  $j \in N$  do insert( $j$ );
     $T := \emptyset$ ;
    while  $|T| < (n - 1)$  do
        begin
            find min( $i$ );
            delete min( $i$ );
             $T := T \cup (pred(i), i)$ ;
            For each  $(i, j) \in A(i)$  with  $j \notin T$  do
                If  $d(j) > w_{ij}$  then
                    begin
                         $d(j) := w_{ij}$ ;
                         $pred(j) := i$ ;
                        decrease key( $j, w_{ij}$ );
                    end;
            end;
         $T$  is an MST.
    end;

```

Figure 7. Prim's algorithm based on heap data structure.

Table 1. Running Time of Prim’s Algorithm with Various Data Structures

Data Structure	Running Time
Adjacency matrix [6]	$O(n^2)$
Binary heap,	$O(m \log n)$
Adjacency list [6]	
d -heap [6]	$O(m \log_d n)$, with $d = \max\{2, m/n\}$
Fibonacci heap [6]	$O(m + n \log n)$
Johnson’s data	$O(m \log \log C)$
structure [10]	
Thorup <i>et al.</i> [11]	$O(m + n \log \log \min(n, C))$

the transfer of a node from set S to set $\bar{S}$. At each step, H contains all elements of set $\bar{S}$.

It is clear that Prim’s algorithm performs insert, find-min, and delete-min operations at most n times and the decrease-key operation at most m times. The time bound for each operation, and consequently the total running time of the algorithm, varies with the type of heap. A binary heap supports insertion in $O(\log n)$, find-min in $O(1)$, and delete-min in $O(\log n)$; thereby taking a total time of $O(n + 2n \log n) = O(n \log n)$. At most, $O(m \log n)$ time is consumed in m decrease-key operations each requiring $O(\log n)$ time. Therefore, implementing Prim’s algorithm using a binary heap takes $O(n \log n + m \log n) = O(m \log n)$ time. Table 1 gives the running time of the algorithm with various types of heaps.

RECENT ALGORITHMS

The criticality of an MST—due to its direct and indirect applications in almost all network flow problems—has always inspired researchers to find faster algorithms. With the development of novel data structures, new implementation schemes are being constructed. The primary aim of these schemes is to improve the asymptotic running time of the algorithm. We list the recent advances

Table 2. Recent Advances in the MST Algorithms

Algorithms	Running Time
Yao [16]	$O(m \log \log n)$
Cheriton and Tarjan [17]	$O(m \log \log n)$
Karger [18]	$O(n \log n + m)$
Fredman and Tarjan [19]	$O(m \beta(m, n))^a$
Gabow <i>et al.</i> [20]	$O(m \log \beta(m, n))$
Osipov <i>et al.</i> [7]	$O(m + n \log n \log m/n)$
Chazelle [21]	$O(m \alpha(m, n))^b$
Karger <i>et al.</i> [15]	$O(\min(n^2, m \log n));$ average: $O(m)$
Pettie and Ramachandran [22]	Optimal but unknown
Fredman and Willard [23]	$O(m + n \log n / \log \log n)$

^a $\beta(m, n) = \min(i : \log^{(i)} n \leq m/n)$, and $\log^0 n = n$.

^b $\alpha(m, n)$, inverse Ackermann’s function [24].

in Table 2. The details of these algorithms can be found in corresponding papers and with a summarized analysis in Refs 12, 13, and 14. The randomized algorithm by Tarjan *et al.* [15] finds the MST within linear time with very high probability of $1 - \exp(-\Omega(m))$; however, its worst-case complexity is $O(\min(n^2, m \log n))$, which is same as Borůvka’s algorithm [1].

SUMMARY

The MST problems are generally recognized as one of the first combinatorial optimization problems studied specifically from an algorithmic perspective. Its diverse set of direct and indirect applications has always encouraged researchers to improve the preexisting algorithms. Several variants of MST are being extensively studied; some of which are NP-complete, while for others polynomial-time algorithms exist. For example, (i) *k*-minimum spanning tree problem: to find a tree that spans any set of k nodes and has a minimum total weight; (ii) *capacitated MST problem*: to find an MST honoring a set of capacity constraints; (iii) *Euclidean MST problem*: to find an MST in a graph with arc weights corresponding to the Euclidean

distance between nodes; and (iv) *p-smallest spanning tree problem*: to find a set of p spanning trees such that no spanning tree outside this set has a lesser weight. The first two problems are NP-complete, while polynomial-time algorithms exist for the rest.

In this article, we have discussed two central properties of MST, which are at the core of all existing algorithms: strong cut and strong cycle properties. We have also described two classical algorithms based on these properties: Kruskal's algorithm (strong cycle) and Prim's algorithm (strong cut). We have also listed a set of leading contributions in this area, which have reached very close to linear time complexity. We have emphasized, with some examples, that running times of the algorithms can be improved significantly by using advanced data structures.

Even after 80 years of research on the MSTs, it remains an open problem to determine whether or not there exists a deterministic algorithm with a linear time bound for a graph with general arc weights. Pettie and Ramachandran [22] have developed a provably optimal deterministic algorithm, but its complexity is still unknown. There are also some research opportunities in developing parallelizable algorithms, as most of the existing ones are not very effective.

REFERENCES

1. Borůvka O. Ojstém problému minimálním [About a certain minimal problem]. *Práce Moravské Přírodovědecké Společnosti* 1926;3:37–58 (in Czech).
2. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc Am Math Soc* 1956;7:48–50.
3. Ahuja RK, Magnanti TL, Orlin JB. Minimum spanning trees. In: *Network flows: theory, algorithms, and applications*. New Jersey: Prentice Hall; 1993. pp. 510–530.
4. Prim RC. Shortest connection networks and some generalizations. *Bell Syst Tech J* 1957;36:1389–1401.
5. Cormen TH, Leiserson CE, Rivest RL, Stein C. Data structures for disjoint sets. In: *Introduction to algorithms*, 2nd ed. Cambridge (MA): MIT Press; New York: McGraw-Hill; 2001. pp. 498–524.
6. Tarjan RE. Minimum spanning trees. In: *Data structures and network algorithms*. CBMS-NSF regional conference series in applied mathematics. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2001. pp. 71–81.
7. Osipov V, Sanders P, Singler J. The Filter-Kruskal minimum spanning tree algorithm. *Proceedings of the Workshop on Algorithm Engineering and Experiments*, New York, USA; 2009. pp. 52–61.
8. Jarník V. Ojstém problému minimálním [About a certain minimal problem]. *Práce Moravské Přírodovědecké Společnosti* 1930;6: 57–63 (in Czech).
9. Dijkstra EW. A note on two problems in connection with graphs. *Numer Math* 1959;1:269–271.
10. Johnson DB. A priority queue in which initialization and queue operations take $O(\log \log D)$ time. *Math Syst Theor* 1982;15:295–309.
11. Thorup M. Integer priority queues with decrease key in constant time and the single source shortest paths problem. *J Comput Syst Sci* 2004;69:330–353.
12. Bazlamacci CF, Hindi KS. Minimum weight spanning tree algorithms A survey and empirical study. *Comput Oper Res* 2001;28: 767–785.
13. Graham RL, Hell P. On the history of the minimum spanning tree problem. *Ann History Comput* 1985;7:43–57.
14. Pettie S, Ramachandran V. On the shortest path and minimum spanning tree problems. *Doctoral Dissertation*, The University of Texas at Austin. 2003. pp. 96–98.
15. Karger DR, Klein PN, Tarjan RE. A randomized linear-time algorithm to find minimum spanning trees. *J ACM* 1995;42:321–328.
16. Yao A. An $O(|E| \log \log |V|)$ algorithm for finding minimum spanning trees. *Inform Proc Lett* 1975;4:21–23.
17. Cheriton D, Tarjan RE. Finding minimum spanning trees. *SIAM J Comput* 1976;5: 724–742.
18. Karger DR. Random sampling in matroids, with applications to graph connectivity and minimum spanning trees. In: *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science*. Los Alamitos, California: IEEE Computer Society Press; 1993. pp. 84–93.
19. Fredman ML, Tarjan RE. Fibonacci heaps and their uses in improved network optimization algorithms. *J ACM* 1987;34:596–615.

20. Gabow HN, Galil Z, Spencer T, Tarjan RE. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica* 1985;6:109–122.
21. Chazelle B. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *J ACM* 2000;47:1028–1047.
22. Pettie S, Ramachandran V. An optimal minimum spanning tree algorithm. *J ACM* 2002;49:16–34.
23. Fredman MN, Willard DE. Transdichotomous algorithms for minimum spanning trees and shortest paths. *J Comput Syst Sci* 1994;48:533–551.
24. Tarjan RE. Efficiency of a good but not linear set-union algorithm. *J ACM* 1975;22:215–225.

MINLP SOLVER SOFTWARE

MICHAEL R. BUSSIECK
GAMS Development Corporation,
Washington, DC

STEFAN VIGERSKE
Department of Mathematics,
Humboldt–Universität zu Berlin,
Berlin, Germany

INTRODUCTION

The general form of an MINLP is

$$\begin{array}{ll} \text{minimize} & f(x,y) \\ \text{subject to} & g(x,y) \leq 0 \\ & x \in X \\ & y \in Y \text{ integer.} \end{array} \quad (\text{P})$$

The function $f: \mathbb{R}^{n+s} \rightarrow \mathbb{R}$ is a possibly nonlinear objective function and $g: \mathbb{R}^{n+s} \rightarrow \mathbb{R}^m$ a possibly nonlinear constraint function. Most algorithms require the functions f and g to be continuous and differentiable, but some may even allow for singular discontinuities. The variables x and y are the decision variables, where y is required to be integer valued. The sets $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^s$ are bounding-box-type restrictions on the variables. In addition to integer requirements on variables, other kinds of discrete constraints are also commonly used. These are, for example, special-ordered-set constraints [only one (SOS type 1) or two consecutive (SOS type 2) variables in an (ordered) set are allowed to be nonzero] [1], semicontinuous variables (the variable is allowed to take either the value zero or a value above some bound), semiinteger variables (like semicontinuous variables, but with an additional integer restriction), and indicator variables (a binary variable indicates whether a certain set of constraints has to be enforced). In all cases, it is possible to reformulate such constraints into a standard form by introducing additional

variables and linear constraints. The purely continuous case ($s = 0$) is not considered here; see the article titled **Software for Nonlinearly Constrained Optimization** in this encyclopedia for an overview on NLP Software.

Computational tractability depends significantly on whether the functions $f(x,y)$ and $g(x,y)$ are convex or not; In this article, we say an MINLP is *convex* if both $f(x,y)$ and $g(x,y)$ are convex over $X \times Y$. Otherwise the MINLP is said to be *nonconvex*. Note that some solvers for convex MINLPs can also be applied under less strict notions of convexity, for example, to the case where the set defined by the constraints $g(x,y) \leq 0$ is convex, or where the objective function and constraints are only pseudoconvex [2]. (A differentiable function $h: X \rightarrow \mathbb{R}$ is *pseudoconvex* on a convex set $X \subseteq \mathbb{R}^n$ if for every $x, y \in X$ with $h(x) < h(y)$, it follows that $\langle \nabla h(y), x - y \rangle < 0$. An important property of a pseudoconvex function is the convexity of its level-sets.)

HISTORY

To the best of our knowledge, the earliest commercial software package that could solve MINLP problems was SCICONIC in the mid-1970s [3–5]. Rather than handling nonlinearities directly, linked special-ordered-set variables provided a mechanism to represent low dimensional nonlinear terms by a piecewise linear approximation and thus allowed to use mixed-integer linear programming (MIP) to obtain solutions to an approximation of the MINLP. In the mid-1980s, Grossmann and Kocis developed DICOPT, a general purpose algorithm for convex MINLP, based on the outer approximation method [6]. Since then, a number of academic and commercial codes for convex MINLP have emerged, either based on outer approximation using MIP relaxations [6], an integration of outer approximation into a linear programming (LP) relaxation based branch and cut [7],

or nonlinear programming (NLP) relaxation based branch-and-bound algorithms [8]. For the global solution of nonconvex MINLP, the first general purpose solvers were ALPHABB, BARON, and GLOP, all based on convexification techniques for nonconvex constraints [9–12]. See also the section titled “Algorithms” for a small discussion of MINLP algorithms.

GROUPINGS

Embedded versus Independent

Due to the high complexity of MINLP and the wide range of applications that can be modeled as MINLPs, it is sometimes desirable to customize the MINLP solver for a specific application in order to achieve good computational performance [13–15]. Further, MINLP solvers are often built by combining LP, MIP, and NLP solvers. These are two main reasons for tightly integrating some MINLP solvers into modeling systems (general systems like AIMMS [16], AMPL [17], and GAMS [18] or vendor specific systems like FICO XPRESS-MOSEL [19], LINGO [20], and OPL [21]). For example, the AIMMS outer approximation solver (AOA) allows modifications of its algorithm by the user. Further, the solvers DICOPT and SBB are exclusively available for GAMS users, since they revert to MIP and NLP solvers in the GAMS system for the solution of subproblems. Also for an efficient use of the solver OQNLP, it is preferable to use one of the GAMS NLP solvers.

On the other side, there are many solvers that can be used independent of a modeling system, even though they may still require the presence of an LP, MIP, or NLP solver plugin. However, often these “independent” solvers are also used within a modeling system, since the modeling system typically provides evaluators for nonlinear functions, gradients, and Hessians and gives easy access to algebraic information about the problem.

Extending MIP versus Extending NLP versus Starting from Scratch

MINLP solvers are seldom developed completely from scratch. In many cases, an

MIP or NLP solver builds the basis for an extension toward MINLP. Solvers which can be categorized as extending an MIP solver toward handling of nonlinear objectives and constraints are BONMIN, COUENNE, CPLEX, FICO XPRESS-OPTIMIZER, FILMINT, LINDOBB, MOSEK, and SCIP. On the other hand, solvers where an NLP solver was extended to handle integrality restrictions are BNB, FICO XPRESS-SLP, FMINCONSET, KNITRO, MILANO, MINLP_BB, MISQP, OQNLP, and SBB.

Finally, there is a group of solvers which are more-or-less developed from scratch, but which may solve LP, MIP, NLP, or MINLP subproblems. In this category, we have ALPHABB, ALPHAECF, AOA, BARON, DICOPT, LAGO, LINDOGLOBAL, and MIDACO.

Algorithms

Algorithms for solving MINLPs are often build by combining algorithms from linear programming, integer programming, and nonlinear programming, for example, branch-and-bound, outer approximation, local search, and global optimization. We refer to the sections titled “Fundamental Techniques,” “Nonlinear Programming (NLP) and Global Optimization (GO),” and “Models and Algorithms” in this encyclopedia for an introduction into these topics.

Most of the solvers implement one (or several) of three algorithmic ideas to tackle MINLPs. First, there are branch-and-bound solvers that use NLP relaxations such as ALPHABB, BNB, BONMIN (in B-BB mode), CPLEX, FICO XPRESS-OPTIMIZER, FICO XPRESS-SLP (in “SLP within MIP” mode), FMINCONSET, KNITRO, LINDOBB, MILANO, MINLP_BB, MOSEK, and SBB. For all these solvers except ALPHABB, the NLP relaxation is obtained by relaxing the integrality restriction in (P), thus, it may be a nonconvex NLP. Since the NLP solver used to solve the NLP relaxation usually ensures only local optimal solutions, these solvers work as heuristics in case of a nonconvex MINLP. For the solver ALPHABB, however, a convex NLP relaxation is generated by using convex underestimators for the functions $f(x,y)$ and

$g(x,y)$ in (P). This solver can therefore be applied also to nonconvex MINLPs.

As an alternative to relaxing integrality restrictions and keeping nonlinear constraints, some solvers keep the integrality constraints and instead replace the nonlinear functions $f(x,y)$ and $g(x,y)$ by a linear relaxation. In an outer-approximation algorithm [6,22], a relaxation is obtained by using gradient-based linearizations of $f(x,y)$ or $g(x,y)$ at solution points of NLP subproblems. The resulting MIP relaxation is then solved by an MIP solver. Solvers in this class are AOA, BONMIN (in B-OA mode), DICOPT, and FICO XPRESS-SLP (in “MIP within SLP” mode). Since gradient-based linearizations yield an outer-approximation only for convex MINLPs, these solvers are only applicable for convex MINLPs. In contrast to outer-approximation based algorithms, an extended cutting-plane algorithm solves a sequence of MIP relaxations which encapsulate optimal solutions of (P) by cutting-planes and supports of $f(x,y)$ rather than outer-approximating the whole feasible region of (P) [23]. This algorithm is implemented by the solver ALPHAECP, that can be applied to convex as well as pseudoconvex MINLPs.

A third class of solvers are those which integrate the linearization of $f(x,y)$ and $g(x,y)$ into the branch and cut process [7]. Thus, here an LP relaxation is successively solved, new linearizations of $f(x,y)$ and $g(x,y)$ are generated to improve the relaxation, and integrality constraints are enforced by branching on the y variables. Solvers which use gradient-based linearizations are AOA, BONMIN (in B-QG mode), and FILMINT.

Since the use of gradient-based linearizations in a branch and cut algorithm ensures global solutions only for convex MINLPs, solvers for nonconvex MINLPs use convexification techniques to compute linear underestimators of a nonconvex function. However, the additional convexification step may require to branch also on continuous variables in nonconvex terms (so called *spatial* branching). Such a branch and cut algorithm is implemented by

BARON, COUENNE, LAGO, LINDOGLOBAL, and SCIP.

The remaining solvers implement a different methodology. BONMIN (in B-Hyb mode) alternates between LP and NLP relaxations during one branch-and-bound process. MISQP integrates the handling of integrality restrictions into the solution of a nonlinear program via sequential quadratic programming, that is, it ensures that $f(x,y)$ and $g(x,y)$ are only evaluated at points where y is integral. MIDACO applies an extended ant colony optimization method and can use MISQP as a local solver. Finally, OQNLP applies a randomized approach by sampling starting points and fixings of integer variables for the solution of NLP subproblems.

Capabilities

Not every solver accepts general MINLPs as input. Solvers that currently handle only MINLPs, where the objective function and constraints are quadratic (so-called *MIQCPs*) or second order cone (SOC) programs are CPLEX, FICO XPRESS-OPTIMIZER, MOSEK, and SCIP. All solvers support convex quadratic functions. Further, nonconvex quadratic functions that involve only binary variables are supported by CPLEX and FICO XPRESS-OPTIMIZER. Quadratic constraints that permit a SOC representation are supported by CPLEX. SOC constraints are supported by MOSEK. SCIP supports quadratic constraints in any form.

Solvers that guarantee global optimal solutions for convex general MINLPs are ALPHAECP, AOA, BNB, BONMIN, DICOPT, FICO XPRESS-SLP, FILMINT, FMINCONSET, KNITRO, LAGO, LINDOBB, MILANO, MINLP_BB, and SBB. In case of a nonconvex MINLP, these solvers can still be used as a heuristic. Especially branch-and-bound based algorithms that use NLPs for bounding, often find good solutions also for nonconvex problems, while pure outer approximation based algorithms may easily run into infeasible LP or MIP relaxations due to wrong cutting-planes. Note, that ALPHAECP ensures global optimal solutions also for pseudoconvex MINLPs.

Solvers that also guarantee global optimality for nonconvex general MINLPs require an algebraic representation of the functions $f(x,y)$ and $g(x,y)$ for the computation of convex underestimators. That is, for each function a representation as a composition of basic arithmetic operations and functions (addition, multiplication, power, exponential, trigonometric, etc.) on constants and variables need to be provided. The solvers ALPHABB, BARON, COUENNE, and LINDOGLOBAL belong into this category.

MIDACO, MISQP and OQNLP can handle general MINLPs, but do not guarantee global optimality even on convex problems.

MINLP SOLVERS

In the following, we briefly discuss individual solvers for MINLPs. We have excluded solvers from this list that are clearly no longer available (e.g., SCICONIC). The solvers listed below have different levels of reliability and activity with respect to development and maintenance. Wide availability through modeling systems and other popular software indicates that a solver has reached a decent level of maturity. Hence, in this list, we mention availability (e.g., open source, standalone binary, interfaces to general modeling systems) in addition to a solver's developer, capability, and algorithmic details. Table 1 summarizes the list of solvers and indicates the availability via AIMMS, AMPL, GAMS, and the NEOS server for each solver [43].

ALPHABB (α -Branch-and-Bound) [9,29]. This solver has been developed by the research group of C. A. Floudas at the Computer-Aided Systems Laboratory of Princeton University. It is available to this group and to their collaborators.

ALPHABB can be applied for convex and nonconvex MINLPs. It implements a branch-and-bound algorithm that utilizes convex NLPs for bounding. Convex envelopes and tight convexifications are obtained for specially structured nonconvex terms (e.g., bilinear, trilinear, multilinear, univariate concave, edge concave,

generalized polynomials, and fractional), and α -convex underestimators for general twice continuously differentiable functions. The latter are determined by adding a nonpositive convex function to the original nonconvex function such that the Hessian of the sum is guaranteed to be positive semidefinite (PSD) [51]. Various interval arithmetic based techniques for estimating rigorous bounds on the minimal eigenvalue of the Hessian of the original nonconvex function are available.

ALPHAIECP (α -Extended Cutting-Plane) [2,30]. This solver has been developed by the research group of T. Westerlund at the Process Design and Systems Engineering Laboratory of the Åbo Akademi University, Finland. It is available as a commercial solver within GAMS.

ALPHAIECP ensures global optimal solutions for convex and pseudoconvex MINLPs. It generates and successively improves an MIP outer approximation of a neighborhood of the set of optimal solutions of (P) and can solve NLP subproblems to find feasible solutions early. The MIP is here refined by linearizing nonlinear constraints at solutions of the MIP outer approximation. By shifting hyperplanes, pseudoconvex functions can also be handled.

AOA (AIMMS Outer Approximation) [16]. This solver has been developed by Paragon Decision Technology. AOA is available as an "open solver" inside AIMMS. The open solver approach allows the user to customize the algorithm for a specific application.

AOA ensures global optimal solutions only for convex MINLPs. It generates and successively improves an MIP outer approximation of (P) and can solve NLP subproblems to find feasible solutions early. In contrast to ALPHAIECP, AOA constructs an MIP outer approximation of the feasible region of (P) by linearizing nonlinear functions in solutions of NLP subproblems [6]. Since for a nonconvex constraint such a linearization may not be valid, the MIP relaxation is modified such that the corresponding hyperplane is allowed to move away from its support point.

Table 1. An Overview on Solvers for MINLP. The First Row Indicates Whether a Solver Accepts Only Problems with Quadratic Functions (MIQCPs) or General Nonlinear Functions (MINLPs)

	Solver	Literature	AIMMS	AMPL	GAMS	NEOS	URL
MIQCP	CPLEX	[24]	✓	✓	✓	–	http://www-01.ibm.com/software/integration/optimization/cplex
	FICO Xpress-Optimizer	[25]	✓	✓	✓	–	http://www.fico.com/xpress
	MOSEK	[26]	✓	✓	✓	–	http://www.mosek.com
General	SCIP	[27,28]	–	–	✓	✓	http://scip.zib.de
	ALPHA BB	[9,29]	–	–	–	–	http://titan.princeton.edu
	ALPHA ECP	[2,30]	–	–	✓	✓	http://www.abo.fi/~twesterl
MINLP	AOA	[16]	✓	–	–	–	http://www.aimms.com
	BARON	[12,31]	✓	–	✓	✓	http://archimedes.cheme.cmu.edu/baron/baron.html
	BNB	[32]	–	–	–	–	http://www.mathworks.com/matlabcentral/fileexchange/95
	BONMIN	[33]	–	✓	✓	✓	https://projects.coin-or.org/Bonmin
	COUENNE	[34]	–	✓	✓	✓	https://projects.coin-or.org/Couenne
	DICOPT	[18,35]	–	–	✓	✓	http://www.gams.com/solvers
	FICO Xpress-SLP	[36]	–	–	–	–	http://www.fico.com/xpress
	FILMINT	[37]	–	✓	–	✓	http://www-neos.mcs.anl.gov/neos/solvers/minco:FilMINT/AMPL.html
	FMINCONSET	[38]	–	–	–	–	http://www.mathworks.com/matlabcentral/fileexchange/96
	KNITRO	[39]	✓	✓	✓	✓	http://www.ziena.com
	LaGO	[40]	–	✓	✓ ^a	–	https://projects.coin-or.org/LaGO
	LINDO BB	[41]	–	–	✓	–	http://www.lindo.com
	LINDO GLOBAL	[42,44]	–	–	✓	✓	http://www.lindo.com
	LOGMIP	[45]	–	–	✓	–	http://www.logmip.ceride.gov.ar
	MIDACO	[46,47]	–	–	–	–	http://www.midaco-solver.com
	MILANO	[48]	–	–	–	–	http://www.pages.drexel.edu/~hvb22/milano
	MINLP_BB	[8]	–	✓	–	✓	http://wiki.mcs.anl.gov/leyffer/index.php/Sven_Leyffer's_Software#MINLPBB
	MISQP	[49]	–	–	–	–	http://www.math.uni-bayreuth.de/~kschittkowski/MISQP.pdf
	OQNLP	[18,50]	–	–	✓	✓	http://www.gams.com/solvers
	SBB	[18]	–	–	✓	✓	http://www.gams.com/solvers

^aInterface for MINLPs available, but not included in GAMS distribution.

Recently, also a branch-and-bound algorithm that utilizes LPs for bounding [7] has been added to AOA.

BARON (Branch-And-Reduce Optimization Navigator) [12,31]. This solver was originally developed by the group of N.V. Sahinidis at the University of Illinois at Urbana-Champaign and is currently developed by N.V. Sahinidis at Carnegie Mellon University and M. Tawarmalani at Purdue University. It is available as a commercial solver within AIMMS and GAMS.

BARON can be applied to convex and nonconvex MINLPs. It implements a spatial branch-and-bound algorithm that utilizes LPs for bounding. The linear outer-approximation is based on a reformulation of (P) that it constructed (by adding auxiliary variables) in a way that it contains only nonconvex terms for which a convex underestimator (or concave overestimator) is known. The algorithm is enhanced by using advanced box reduction techniques and new convexification techniques for multilinear terms [52]. Further, BARON is able to use NLP relaxations for bounding [44], even though this option is not encouraged.

BNB (Branch and Bound) [32]. This solver has been developed by K. Kuipers of the Department of Applied Physics at the University of Groningen. It is available as MATLAB [53] source.

BNB ensures global optimal solutions for convex MINLPs. It implements a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLPs are solved by the MATLAB Optimization Toolbox routine `FMINCON`.

BONMIN (Basic Open-source Nonlinear Mixed Integer Programming) [33]. This open-source solver has been developed primarily by P. Bonami in a cooperation of Carnegie Mellon University and IBM Research. It is available in source code and as standalone binaries from COIN-OR (Computational Infrastructure for Operations Research) [54], has an AMPL interface, and is distributed as a free solver within GAMS.

BONMIN ensures global optimal solutions only for convex MINLPs. It implements (at

least) four algorithms: B-OA is an outer-approximation algorithm that generates and successively improves an MIP outer approximation of (P) [6], B-QG is a branch-and-bound algorithm that utilizes LPs for bounding [7], B-BB is a branch-and-bound algorithm that utilizes NLPs for bounding [8], and B-Hyb is a hybrid of B-QG and B-BB, which alternates between LP and NLP relaxations for bounding. BONMIN is implemented on top of the MIP solver CBC [55] and can use FILTERSQP [56] and IPOPT [57] as NLP solvers.

COUENNE (Convex Over and Under ENvelopes for Nonlinear Estimation) [34]. This open-source solver has been developed primarily by P. Belotti, originally in cooperation of Carnegie Mellon University and IBM Research, and now at Lehigh University. It is available in source code and as standalone binaries from COIN-OR, has an AMPL interface, and is distributed as a free solver within GAMS.

COUENNE ensures global optimal solutions for convex and nonconvex MINLPs. It implements a spatial branch-and-bound algorithm that utilizes LPs for bounding. Similar to BARON, the linear outer-approximation is generated from a reformulation of (P). The algorithm is enhanced by box reduction techniques and disjunctive cuts [58]. COUENNE is implemented on top of BONMIN.

CPLEX [24]. This solver has been developed by CPLEX Optimization, Inc. (later acquired by ILOG and recently acquired by IBM). It is available as standalone binaries and as a component in many modeling systems.

CPLEX can solve convex MIQCPs. For models that only have binary variables in the potentially indefinite quadratic matrices, CPLEX automatically reformulates the problem to an equivalent MIQCP with PSD matrices. It implements a branch-and-bound algorithm that utilizes LPs or QCPs for bounding.

DICOPT (Discrete and Continuous Optimizer) [18,35]. This solver has been developed by the research group of I. E. Grossmann at the Engineering Research Design Center at Carnegie Mellon University. It is available as a commercial solver within GAMS.

DICOPT ensures global optimal solutions for convex MINLPs. Starting with the NLP relaxation (obtained from (P) by relaxing the integer requirement on y), it alternates between solving MIP outer approximations and NLP subproblems of (P) to compute lower and upper bounds [6]. To accommodate nonconvex MINLPs, nonlinear equality constraints are relaxed by replacing them with inequalities where the linearizations of the nonlinear functions are allowed to move away from their support point by the use of slack variables and through an augmented penalty function in the MIP relaxation. Since in this case valid lower bounds cannot be obtained, the termination is based on lack of improvement in the objective of the NLP subproblem.

FICO XPRESS-OPTIMIZER [25]. This solver has been developed by Dash Optimization (later acquired by FICO). It is available as standalone binaries and as a component in many modeling systems.

FICO XPRESS-OPTIMIZER can solve convex MIQCPs. For models that only have binary variables in the potentially indefinite quadratic matrices, FICO XPRESS-OPTIMIZER automatically reformulates the problem to an equivalent MIQCP with PSD matrices. It implements a branch-and-bound algorithm that utilizes QCPs for bounding.

FICO XPRESS-SLP [36]. This solver has been developed by Dash Optimization (later acquired by FICO). It is available as standalone binaries and as a FICO XPRESS-MOSEL module [19].

FICO XPRESS-SLP ensures global optimal solutions for convex MINLPs. It implements three algorithms: The (default) “SLP within MIP” variant is a branch-and-bound algorithm that utilizes NLPs for bounding [8]. The NLP subproblems are solved by successive linear programming (SLP). Solving MIPs as subproblems of the SLP algorithm leads to the “MIP within SLP” variant. It is comparable with an MIP relaxation based outer-approximation algorithm [6]. A third variant (“SLP then MIP”) solves first an NLP relaxation (by SLP), then an MIP relaxation, and

finally an NLP subproblem to obtain a feasible solution to (P) [36]. Also to accommodate nonconvex constraints, in all variants, the hyperplanes obtained from gradient-based linearizations in SLP are allowed to move away from their support point.

FILMINT (Filter-Mixed Integer Optimizer) [37]. This solver has been developed by the research groups of S. Leyffer at the Laboratory for Advanced Numerical Simulations of Argonne National Laboratory and J. T. Linderoth at the Department of Industrial and Systems Engineering of Lehigh University. It provides an AMPL interface.

FILMINT ensures global optimal solutions only for convex MINLPs. It implements a branch-and-bound algorithm that utilizes LPs for bounding [7]. FILMINT is implemented on top of the MIP solver MINTO [59] and the NLP solver FILTERSQP [56].

FMINCONSET [38]. This solver had been developed by I. Solberg at the Department of Engineering Cybernetics of the University of Trondheim (now NTNU). It is available as MATLAB source.

FMINCONSET ensures global optimal solutions for convex MINLPs. It implements a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLPs are solved by the MATLAB Optimization Toolbox routine FMINCON.

KNITRO [39]. This solver has been developed by Ziena Optimization, Inc. It is available as standalone binary and as a component in many modeling systems.

KNITRO ensures global optimal solutions for convex MINLPs. MINLPs are solved by branch-and-bound, where both linear or nonlinear problems can be used for the bounding step [7,8].

LAGO (Lagrangian Global Optimizer) [40]. This open-source solver had been developed by the research group of I. Nowak at the Department of Mathematics of Humboldt University, Berlin. It is available in source code from COIN-OR and provides AMPL and GAMS interfaces.

LAGO ensures global optimal solutions for convex MINLPs and nonconvex MIQCPs. It implements a spatial branch-and-bound algorithm utilizing a linear relaxation for the bounding step. The relaxation is obtained by linearizing convex functions, underestimating quadratic nonconvex functions, and approximating nonconvex nonquadratic functions by quadratic ones.

LINDOBB (LINDO Systems Mixed Integer Nonlinear Solver) [41]. This solver has been developed by LINDO Systems, Inc. It is available within the LINDO environment and as a commercial solver within GAMS.

LINDOBB ensures global optimal solutions for convex MINLPs. It was the first commercially available solver implementing a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLP relaxations are solved by CONOPT [18,60].

LINDOGLOBAL (LINDO Systems Global Solver) [42,61]. This solver has been developed by LINDO Systems, Inc. It is available within the LINDO environment [41], LINGO [20], *What'sBest!* [62], and as a commercial solver within GAMS.

LINDOGLOBAL ensures global optimal solutions for convex and nonconvex MINLPs. It implements a branch and cut algorithm that utilizes LPs for bounding. Branching is performed for subproblems that are not provably infeasible and where nonconvex constraints are present or the LP relaxation has a fractional solution. Thus, in difference to other spatial branch-and-bound algorithm, branching on a continuous variable especially targets on achieving convexity of (nonlinear) subproblems. LINDOGLOBAL can also handle some nonsmooth or discontinuous functions like $\text{abs}(x)$, $\text{floor}(x)$, and $\text{max}(x, y)$.

MIDACO (Mixed Integer Distributed Ant Colony Optimization) [46,47]. This solver has been developed by M. Schlüter at the Theoretical and Computational Optimization Group of the University of Birmingham. It works as a library with Matlab, C/C++, and

Fortran interfaces and is available from the author on request.

MIDACO can be applied to convex and nonconvex MINLPs. It implements an extended ant colony search method based on an oracle penalty function and can be combined with MISQP as solver for local searches in (P). It targets applications where the problem formulation is unknown ($f(x, y)$ and $g(x, y)$ are black-box functions) or involves critical properties like non-convexities, discontinuities, flat spots, or stochastic distortions. Further, MIDACO can exploit distributed computer architectures by parallelizing function evaluation calls.

MILANO (Mixed-Integer Linear and Nonlinear Optimizer) [48]. This solver is developed by H. Y. Benson at the Department of Decision Sciences of Drexel University. It is still in development and available as MATLAB source.

MILANO ensures global optimal solutions for convex MINLPs. It implements a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLPs are solved by LOQO [63], where special emphasis is put on how to warmstart this interior-point solver.

MINLP_BB (Mixed Integer Nonlinear Programming Branch-and-Bound) [8]. This solver had been developed by R. Fletcher and S. Leyffer at the University of Dundee. It provides an AMPL interface and is available for MATLAB via the TOMLAB Optimization Environment [64].

MINLP_BB ensures global optimal solutions for convex MINLPs. It implements a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLPs are solved by FILTERSQP.

MISQP (Mixed Integer Sequential Quadratic Programming) [49]. This solver has been developed by the research group of K. Schittkowski at the Department of Computer Science of the University of Bayreuth. It works as a standalone library with a Fortran interface.

MISQP can be applied to convex and non-convex MINLPs, but assumes that the values

of $f(x, y)$ and $g(x, y)$ do not change drastically as a function of y . MISQP implements a modified sequential quadratic programming (SQP) method, where functions are only evaluated at points (x, y) with y integer. It targets applications where the evaluation of $f(x, y)$ or $g(x, y)$ may be expensive.

MOSEK [26]. This solver has been developed by MOSEK ApS. It is available as a standalone binary, has AMPL and MATLAB interfaces, and is distributed as a commercial solver within AIMMS and GAMS.

MOSEK can be applied to convex MIQCPs and to mixed-integer conic programs. It implements a branch-and-bound method that utilizes QCPs or SOC programs for bounding [7].

OQNLP (OptQuest Nonlinear Programming) [18,50]. This solver has been jointly developed by OptTek Systems, Inc. and Optimal Methods, Inc. It is available as a standalone library, for MATLAB via the TOMLAB Optimization Environment, and is distributed as a commercial solver within GAMS.

OQNLP is a heuristic that can be applied to any MINLP. It implements a multistart scatter search algorithm which solves NLP subproblems with fixed discrete variables.

SBB (Simple Branch-and-Bound) [18]. This solver had been developed by ARKI Consulting and Development A/S. It is available as a commercial solver within GAMS.

SBB ensures global optimal solutions for convex MINLPs. It implements a branch-and-bound algorithm utilizing nonlinear relaxations for the bounding step [8]. The NLP relaxations are solved by one (or several) of the NLP solvers available with GAMS. Using the GAMS Branch-Cut-and-Heuristic facility [14], SBB allows the user to implement a model-specific heuristic in the GAMS language.

SCIP (Solving Constraint Integer Programs) [27,28]. This solver has been developed by the Optimization Department at the Zuse Institute, Berlin. For academic institutions, it is available in source code and as standalone binary, and is distributed within GAMS.

SCIP ensures global optimal solutions for convex and nonconvex MIQCPs. It implements a spatial branch-and-bound algorithm utilizing a linear relaxation for the bounding step. The relaxation is obtained by linearizing convex or convexified quadratic functions.

OUTLOOK AND SUMMARY

Combining discrete and nonlinear optimization results in a rich modeling paradigm applicable to many real-world optimization problems. At the same time, mixed integer nonlinear programming represents a theoretically and computationally challenging problem class and hence provides many interesting research opportunities. Software for solving MINLP models facilitates cooperation between research and application and explains the popularity and increased level of activity around MINLP.

While state-of-the-art MIP solvers typically implement advanced automatic reformulation and preprocessing algorithms, such techniques are less commonly available in MINLP solvers, and in a limited form. Therefore, the modeler's choice of problem formulation is still very important when solving an MINLP. However, software for guided automatic model reformulations and relaxations has recently been developed. LOGMIP [45], one of the first systems available, translates an MINLP with disjunctions into a standard MINLP by applying bigM and convex hull reformulations [65]. More recently, frameworks like GAMS/EMP (Extended Mathematical Programming) [66] and ROSE (Reformulation/Optimization Software Engine) [67] provide a growing toolbox for reformulating MINLPs. Other recent activities, like LIBMC [68] and parts of MINOTAUR (Mixed-Integer Nonconvex Optimization Toolbox—Algorithms, Underestimators, Relaxations) [69] focus on (convex) relaxations for (nonconvex) MINLP.

Another important area is the collection and dissemination of MINLP models. Instance collections like MACMINLP [70] and MINLPLIB [71] provide valuable test cases for solver developers. The new Cyber-Infrastructure for MINLP [72] features a

growing library of problems with high level model descriptions, reformulations, and problem instances.

In this article we have given a concise description of the state-of-the-art in MINLP solvers and have established several groupings with respect to various features of the software. We hope that these groupings and the individual descriptions give sufficient information to guide the selection of the best solver for a particular MINLP problem.

Acknowledgments

We thank Oliver Bastert, Pietro Belotti, Zsolt Csizmadia, Steve Dirkse, Arne Drud, Christodoulos Floudas, Ignacio Grossmann, Marcel Hunting, Ed Klotz, Nikolaos Sahinidis, Martin Schlüter, Linus Schrage, and Tapio Westerlund for their invaluable comments. The second author was supported by the DFG Research Center MATHEON *Mathematics for key technologies* in Berlin.

REFERENCES

1. Beale EML, Tomlin JA. Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables. London: Tavistock Publishing; 1970. pp. 447–454.
2. Westerlund T, Pörn R. Solving pseudoconvex mixed integer optimization problems by cutting plane techniques. *Optim Eng* 2002;3:253–280.
3. Beale EML. Branch and bound methods for numerical optimization of non-convex functions. In: Barritt MM, Wishart D, editors. *COMPSTAT 80 Proceedings in Computational Statistics*. Vienna: Physica-Verlag; 1980. pp. 11–20.
4. Forrest JJH, Tomlin JA. Branch and bound, integer, and non-integer programming. *Ann Oper Res* 2007;149(1):81–87.
5. SCICON Ltd. *SCICONIC user guide version 1.40*. Milton Keynes: Scicon Ltd.; 1989.
6. Duran MA, Grossmann IE. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Math Program* 1986;36:307–339.
7. Quesada L, Grossmann IE. An LP/NLP based branch and bound algorithm for convex MINLP optimization problems. *Comput Chem Eng* 1992;16:937–947.
8. Leyffer S. Integrating SQP and branch-and-bound for mixed integer nonlinear programming. *Comput Optim Appl* 2001;18:295–309.
9. Androulakis IP, Maranas CD, Floudas CA. α BB: a global optimization method for general constrained nonconvex problems. *J Glob Optim* 1995;7:337–363.
10. Sahinidis NV. BARON: a general purpose global optimization software package. *J Glob Optim* 1996;8(2):201–205.
11. Smith EMB, Pantelides CC. A symbolic reformulation/spatial branch-and-bound algorithm for the global optimization of nonconvex MINLPs. *Comput Chem Eng* 1999;23:457–478.
12. Tawarmalani M, Sahinidis NV. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*. Kluwer Academic Publishers; 2002.
13. Bragalli C, D'Ambrosio C, Lee J, *et al.* *Practical water network design by MINLPs*. Technical Report OR-09-21. Bologna, Italy: University of Bologna; 2009.
14. Bussieck MR. *Introduction to GAMS branch-and-cut facility*. Technical report. Washington, DC: GAMS Development Corporation; 2003. Available at <http://www.gams.com/docs/bch.htm>.
15. Farkas T, Czuczai B, Rev E, *et al.* New MINLP model and modified outer approximation algorithm for distillation column synthesis. *Ind Eng Chem Res* 2008;47:3088–3103.
16. Roelofs M, Bisschop J. *AIMMS 3.9—the language reference*. Haarlem, The Netherlands: Paragon Decision Technology B.V.; 2009.
17. Fourer R, Gay DM, Kernighan BW. *AMPL: a modeling language for mathematical programming*. Duxbury Press, Brooks/Cole Publishing Company; 1993.
18. GAMS Development Corporation. *GAMS—the solver manuals*. Washington (DC): GAMS Development Corporation; 2009.
19. FICO. *Xpress-Mosel 7.0*. 2009. Available at <http://www.fico.com/xpress>.
20. Schrage L. *Optimization modeling with LINGO*. Lindo Systems, Inc.; 2008. Available at <http://www.lindo.com>.
21. Dong T. *Efficient modeling with the IBM ILOG OPL-CPLEX development bundles*. 2009. Available at <http://www-01.ibm.com/software/integration/optimization/cplex-dev-bundles/library>.

22. Fletcher R, Leyffer S. Solving mixed integer nonlinear programs by outer approximation. *Math Program* 1994;66(3):327–349.
23. Westerlund T, Pettersson F. An extended cutting plane method for solving convex MINLP problems. *Comput Chem Eng* 1995;21:131–136.
24. Czyzyk J, Mesnier M, Moré J. The NEOS server. *IEEE J Comput Sci Eng* 1998;5:68–75. Available at <http://neos.mcs.anl.gov>.
25. IBM. CPLEX. Available at <http://www.ibm.com/software/integration/optimization/~cplex>.
26. FICO. Xpress-optimizer reference manual, 20.0 edition. 2009. Available at <http://www.fico.com/xpress>.
27. MOSEK Corporation. The MOSEK optimization tools manual, 6.0 edition. 2009. Available at <http://www.mosek.com>.
28. Achterberg T. SCIP: solving constraint integer programs. *Math Program Comput* 2009;1(1):1–41.
29. Berthold T, Heinz S, Vigerske S. Extending a CIP framework to solve MIQCPs. Technical Report 09-23. Berlin, Germany: Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB); 2009.
30. Adjiman CS, Androulakis IP, Floudas CA. Global optimization of mixed-integer nonlinear problems. *J Am Inst Chem Eng* 2000;46:1769–1797.
31. Westerlund T, Lundquist K. Alpha-ECP, version 5.04. An interactive MINLP-solver based on the extended cutting plane method. Technical Report 01-178-A. Åbo, Finland: Process Design Laboratory, Åbo Akademi University; 2003. Available at <http://www.abo.fi/~twesterl/A-ECPManual.pdf>.
32. Tawarmalani M, Sahinidis NV. Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. *Math Program* 2004;99:563–591.
33. Kuipers K. bnb. Available at <http://www.mathworks.com/matlabcentral/fileexchange/95>.
34. Bonami P, Biegler LT, Conn AR, *et al.* An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optim* 2008;5:186–204.
35. Belotti P, Lee J, Liberti L, *et al.* Branching and bounds tightening techniques for non-convex MINLP. *Optim Methods Softw* 2009;24(45):597–634.
36. Kocis GR, Grossmann IE. Computational experience with DICOPT: solving MINLP problems in process systems engineering. *Comput Chem Eng* 1989;13:307–315.
37. FICO. Xpress-SLP program reference manual, 1.41 edition. 2008. Available at <http://www.fico.com/xpress>.
38. Abhishek K, Leyffer S, Linderoth JT. FilMINT: an outer-approximation-based solver for nonlinear mixed integer programs. Technical Report by INFORMS Journal on Computing, Argonne (IL): Argonne National Laboratory, Mathematics and Computer Science Division; 2006. In press.
39. Solberg I. fminconset. Available at <http://www.mathworks.com/matlabcentral/fileexchange/96>.
40. Byrd RH, Nocedal J, Waltz RA. KNITRO: an integrated package for nonlinear optimization. In: di Pillo G, Roma M, editors. Large-scale nonlinear optimization. Springer; 2006. pp. 35–59.
41. Nowak I, Vigerske S. LaGO: a (heuristic) branch and cut algorithm for non-convex MINLPs. *Cent Eur J Oper Res* 2008;16(2):127–138.
42. Lindo Systems, Inc.. Lindo API 6.1.2010. 2009. Available at <http://www.lindo.com>.
43. Gau C, Schrage L. Implementation and testing of a branch-and-bound based method for deterministic global optimization: operations research applications. In: Floudas CA, Pardalos PM, editors. *Frontiers in global optimization. Nonconvex optimization and its applications*. Springer; 2003. pp. 145–164.
44. Ghildyal V, Sahinidis NV. Solving global optimization problems with BARON. In: Migdalas A, Pardalos PM, Värbrand P, editors. Volume 53, *From local to global optimization, Nonconvex optimization and its applications*. Springer; 2001. Chapter 10. pp. 205–230.
45. Vecchietti A, Grossmann IE. LOGMIP: a disjunctive 0-1 nonlinear optimizer for process system models. *Comput Chem Eng* 1999;23:555–565.
46. Schlüter M, Egea JA, Banga JR. Extended ant colony optimization for non-convex mixed integer nonlinear programming. *Comput Oper Res* 2009;36(7):2217–2229.
47. Schlüter M, Gerdts M. The oracle penalty method. *J Glob Optim* 2009;47(2):293–325.
48. Benson HY. Mixed integer nonlinear programming using interior-point methods. *Optimization Methods and Software*. 2010. Available at <http://www.pages>.

- drexel.edu/~hvb22/minlp.pdf.
49. Exler O, Schittkowski K. A trust region SQP algorithm for mixed-integer nonlinear programming. *Optim Lett* 2007;1:269–280.
 50. Ugray Z, Lasdon L, Plummer J, *et al.* Scatter search and local NLP solvers: a multistart framework for global optimization. *INFORMS J Comput* 2007;19(3):328–340.
 51. Adjiman CS, Floudas CA. Rigorous convex underestimators for general twice-differentiable problems. *J Glob Optim* 1996;9:23–40.
 52. Bao X, Sahinidis NV, Tawarmalani M. Multiterm polyhedral relaxations for non-convex, quadratically-constrained quadratic programs. *Optim Methods Softw* 2009;24:485–504.
 53. The Math Works. MATLAB user's guide. 2009. Available at <http://www.mathworks.com>.
 54. Lougee-Heimer R. The common optimization interface for operations research. *IBM J Res Dev* 2003;47(1):57–66. Available at <http://www.coin-or.org>.
 55. Forrest JJH. COIN-OR branch and cut. Available at <http://projects.coin-or.org/Cbc>.
 56. Fletcher R, Leyffer S. Nonlinear programming without a penalty function. *Math Program* 2002;91:239–270.
 57. Wächter A, Biegler LT. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Math Program* 2006;106(1):25–57. Available at <http://projects.coin-or.org/Ipopt>.
 58. Belotti P. Disjunctive cuts for non-convex MINLP. Technical report. Bethlehem (PA): Lehigh University. 2009. Available at <http://coral.ie.lehigh.edu/belotti/papers/~belotti-disj-MINLP.pdf>.
 59. Nemhauser GL, Savelsbergh MWP, Sigismondi GS. MINTO, a mixed integer optimizer. *Oper Res Lett* 1994;15:47–58.
 60. Drud A. CONOPT—a large-scale GRG code. *INFORMS J Comput* 1992;6:207–216.
 61. Lin Y, Schrage L. The global solver in the LINDO API. *Optim Methods Softw* 2009;24(45):657–668.
 62. Lindo Systems, Inc. What's Best! 10.0. 2009. Available at <http://www.lindo.com>.
 63. Vanderbei RJ, Shanno DF. An interior-point algorithm for nonconvex nonlinear programming. *Comput Optim Appl* 1999;13:231–252.
 64. Holmström K. The tomlab optimization environment in matlab. *Adv Model Optim* 1999;1:47–69. Available at <http://tomopt.com/tomlab>.
 65. Grossmann IE, Lee S. Generalized disjunctive programming: nonlinear convex hull relaxation and algorithms. *Comput Optim Appl* 2003;26(1):83–100.
 66. Ferris MC, Dirkse SP, Jagla J-H, *et al.* An extended mathematical programming framework. *Comput Chem Eng* 2009;33(12):1973–1982. (FOCAPO 2008 - Selected Papers from the Fifth International Conference on Foundations of Computer-Aided Process Operations.)
 67. Liberti L, Cafieri S, Tarissan F. Reformulations in mathematical programming: a computational approach. In: Abraham A, Hasanien A-E, Siarry P, editors. Volume 203, Global optimization. Studies in computational intelligence. New York: Springer; 2009. pp. 153–234.
 68. Mitsos A, Chachuat B, Barton PI. McCormick-based relaxations of algorithms. *SIAM J Optim* 2009;20(2):573–601.
 69. Leyffer S, Linderoth J, Munson T. MINOTAUR: Mixed-Integer Nonconvex Optimization Toolbox—Algorithms, Underestimators, Relaxations.
 70. Leyffer S. MacMINLP. Available at <http://wiki.mcs.anl.gov/leyffer/index.php/MacMINLP>.
 71. Bussieck MR, Drud AS, Meeraus A. MINLPLib—a collection of test models for mixed-integer nonlinear programming. *INFORMS J Comput* 2003;15(1):114–119.
 72. CMU-IBM. Cyber-Infrastructure for MINLP. Available at <http://www.minlp.org>.

MIXING SETS

LAURENCE A. WOLSEY
Center for Operations Research and
Econometrics (CORE) and
Département d'Ingénierie
Mathématique (INMA), Université
Catholique de Louvain
Louvain-la-Neuve, Belgium

A mixing set is a set of the form

$$X^{\text{MIX}} = \{(s, z) \in \mathbb{R}^1 \times \mathbb{Z}^n : \\ s + z_t \geq b_t \text{ for } t = 1, \dots, n\}. \quad (1)$$

Note that the variables s, z are unrestricted in sign, though often in practice, there will be nonnegativity or other bound constraints on the variables. The special case when s is nonnegative will be denoted by X_+^{MIX} .

Mixing sets first arose in the study of lot-sizing problems [1], but were defined explicitly by Günlük and Pochet [2], and have since been generalized in a variety of ways. The majority of applications are to lot-sizing problems, but they arise also in single arc network models, as well as in robust and stochastic integer programming. In addition, attempts are under way to use such sets to generate cutting planes for general mixed integer programs (MIPs).

Below we first describe a few examples in which such sets arise and then we develop the basic theory of such sets: in particular, tight and compact extended formulations whose projection gives the convex hull of the set X^{MIX} ; these can be used to optimize over such sets by linear programming (LP) or network flows; a valid inequality description of $\text{conv}(X^{\text{MIX}})$ and a polynomial separation algorithm that allows one to use these inequalities in a cutting plane algorithm. Next, we present several generalizations of mixing sets and indicate briefly some of the important results concerning the convex hulls of these sets.

SOME APPLICATIONS OF MIXING SETS

Here we present four problems, whose formulations naturally lead to mixing sets.

A Discrete Version of the Newsvendor Problem. Consider a product that can be sold during one season in arbitrary (real-valued) amounts. The demand and future production cost are uncertain. Specifically, it is estimated that with probability p_t , there will be a demand d_t and a production cost of q_t per batch of C units during the coming season, where $t = 1, \dots, n$. The other option is to produce the goods in advance at a unit cost of h . The problem is to minimize the total expected cost given that all demands must be satisfied.

Letting s be the amount produced in advance and z_t the number of batches produced during the season under scenario t , one obtains the MIP,

$$\min \left\{ hs + \sum_{t=1}^n p_t q_t z_t : s + C z_t \geq d_t \right. \\ \left. \text{for } t = 1, \dots, n, s \in \mathbb{R}_+^1, z \in \mathbb{Z}_+^n \right\}.$$

Note that after dividing through by C , the feasible region is a mixing set with nonnegativity constraints on the s and z variables.

Lot-Sizing: The Wagner–Whitin Relaxation. In constant capacity lot-sizing, one is given deterministic (known) demands d_t for periods 1 to n . In each period, one can produce any amount between 0 and the capacity C . The production costs in each period t are p_t per unit plus a fixed setup or order cost q_t if the production is nonzero. In addition, there is per unit end of period storage cost h_t . The problem is to satisfy all the demands at a minimum cost.

Letting x_t be the production in period t , $y_t \in \{0, 1\}$ a variable that must take value 1

if $x_t > 0$ and s_t the stock at the end of period t , one obtains the MIP

$$\min \sum_{t=1}^n p_t x_t + \sum_{t=0}^n h_t s_t + \sum_{t=1}^n q_t y_t \quad (2)$$

$$s_{t-1} + x_t = d_t + s_t \quad \text{for } t = 1, \dots, n \quad (3)$$

$$x_t \leq C y_t \quad \text{for } t = 1, \dots, n \quad (4)$$

$$x \in \mathbb{R}_+^n, s \in \mathbb{R}_+^{n+1}, y \in \{0, 1\}^n. \quad (5)$$

The Wagner–Whitin relaxation is obtained by summing the balance constraints (Eq. 3) for periods k up to t , using Equation (4) to replace x_u by its upper bound $C y_u$ and by replacing s_t by its lower bound of zero. The resulting feasible region is

$$\left\{ (s, y) \in \mathbb{R}_+^{n+1} \times \{0, 1\}^n : s_{k-1} + C \sum_{u=k}^t y_u \geq \sum_{u=k}^t d_u \text{ for } 1 \leq k \leq t \leq n \right\}.$$

For fixed k , taking $\sigma = s_{k-1}/C$, $z_t = \sum_{u=k}^t y_u$, $b_t = (\sum_{u=k}^t d_u)/C$, one obtains

$$\sigma + z_t \geq b_t \text{ for } k \leq t \leq n$$

$$0 \leq z_t - z_{t-1} \leq 1$$

$$\sigma \in \mathbb{R}_+^1, z \in \mathbb{Z}^{n-k+1},$$

which is a mixing set with additional constraints on the integer variables of a very special form. So the Wagner–Whitin relaxation can be viewed as the intersection of n mixing sets with common integer variables.

Probabilistic (Chance-Constrained) Programming. Let $\tilde{\zeta} \in \mathbb{R}_+^d$ be a random vector taking a finite number of discrete values $b^k \in \mathbb{R}_+^d$ with probability p_k for $k = 1, \dots, K$, and suppose that there is a probabilistic constraint that some vector $s \in \mathbb{R}_+^d$ satisfies $s \geq \tilde{\zeta}$ with probability at least $1 - \tau$. Setting $z_k = 0$, if $s \geq b^k$ and $z_k = 1$ otherwise, one

obtains the MIP constraints

$$s_i \geq b_i^k (1 - z_k) \text{ for } k = 1, \dots, K, \\ i = 1, \dots, d \quad (6)$$

$$\sum_{k=1}^K p_k (1 - z_k) \geq 1 - \tau \quad (7)$$

$$s \in \mathbb{R}_+^d, \quad z \in \{0, 1\}^K. \quad (8)$$

With $M \geq \max_{i,k} b_i^k$ an arbitrary large value, constraint (6) can be replaced by

$$s_i + M z_k \geq b_i^k \text{ for } k = 1, \dots, K, i = 1, \dots, d. \quad (9)$$

Now the set defined by Equations (8) and (9) is the intersection of d mixing sets with common 0–1 variables.

Mixing Sets and General MIPs. Suppose that one has a mixed integer set $\{(x, y) \in \mathbb{Z}_+^n \times \mathbb{R}_+^p : Ax + Gy \geq b\}$. One possible relaxation is the set

$$\sum_{j=1}^n \lceil a_{ij} \rceil x_j + \sum_{j=1}^p (\max_i g_{ij}^+) y_j \geq b_i \quad \forall i,$$

where w^+ demotes $\max(w, 0)$. Taking $s = \sum_{j=1}^p (\max_i g_{ij}^+) y_j$, and $z_i = \sum_{j=1}^n \lceil a_{ij} \rceil x_j$, one clearly has a mixing set.

THE PROPERTIES OF MIXING SETS

The Simple Case $n = 2$

When $n = 2$, a simple change of variable $\sigma = s + z_1 - b_1$ allows one to rewrite X^{MIX} as the simple mixed integer set

$$X^{\text{MIP}} = \{(\sigma, y) \in \mathbb{R}_+^1 \times \mathbb{Z}^1 : \sigma + y \geq d\};$$

namely, one substitutes $s = \sigma - z_1 + b_1$ in $s + z_2 \geq b_2$ giving $\sigma + (z_2 - z_1) \geq (b_2 - b_1)$ and then sets $y = z_2 - z_1$ and $d = b_2 - b_1$.

X^{MIP} is a well-studied set. When $d \notin \mathbb{Z}^1$ so that $f_d = d - \lfloor d \rfloor > 0$, addition of the valid inequality

$$s \geq f_d (\lfloor d \rfloor + 1 - y), \quad (10)$$

called the *mixed integer rounding (MIR)* inequality, is sufficient to give a complete description of $\text{conv}(X^{\text{MIP}})$.

The General Case

We consider again the set X^{MIX} with $n \geq 3$ and set $N = \{1, \dots, n\}$. First we present some basic results, an outline of the proofs, along with an example.

Observation 1.

- (i) $\text{conv}(X^{\text{MIX}})$ contains the line $(-1, \mathbf{1})$ and the extreme rays $(1, \mathbf{0})$ and $(0, \mathbf{e}_i)$ for $i \in N$, where $\mathbf{e}_i$ is the i th unit vector, $\mathbf{0} = (0, \dots, 0)$ and $\mathbf{1} = (1, \dots, 1)$.
- (ii) Let $f_t = b_t - \lfloor b_t \rfloor$ for $t \in N$. In an extremal solution of $\text{conv}(X^{\text{MIX}})$, $s \equiv f_t \pmod{1}$ for some t . Specifically, there are the n extremal solutions $(f_t, \lfloor b_1 - f_t \rfloor, \dots, \lfloor b_n - f_t \rfloor)$ for $t \in N$.

Extended Formulations. We say that a polyhedron $Q \subset \mathbb{R}^{n+p}$ is a tight extended formulation for a set $X \subset \mathbb{R}^n$ if $\text{conv}(X) = \text{proj}_x(Q)$, where $\text{proj}_x(Q) \equiv \{x \in \mathbb{R}^n : \exists y \in \mathbb{R}^p \text{ with } (x, y) \in Q\}$.

Theorem 1. *A tight, extended formulation for $\text{conv}(X^{\text{MIX}})$ is given by*

$$\begin{aligned} s - \sum_{t=1}^n f_t \delta_t - \mu &= 0, \\ \mu + \sum_{if_i \geq f_t} \delta_i + z_t &\geq \lfloor b_t \rfloor + 1 \quad \text{for } t \in N \\ \sum_{i=1}^n \delta_i &= 1 \end{aligned}$$

$$\begin{array}{ccccccccccccc}
s & -\mu & -0.9\delta_1 & -0.7\delta_2 & -0.6\delta_3 & & & & & & & & = & 0 \\
& \mu & +\delta_1 & & & & & +z_1 & & & & & \geq & 3 \\
& \mu & +\delta_1 & +\delta_2 & & & & & +z_2 & & & & \geq & 1 \\
& \mu & +\delta_1 & +\delta_2 & +\delta_3 & & & & & +z_3 & & & \geq & -1 \\
& \mu & +\delta_1 & +\delta_2 & +\delta_3 & +\delta_4 & & & & & +z_4 & & \geq & 3 \\
& & \delta_1 & +\delta_2 & +\delta_3 & +\delta_4 & & & & & & & = & 1 \\
s & \in & \mathbb{R}^1, & z & \in & \mathbb{R}^4, & \mu & \in & \mathbb{R}^1, & \delta & \in & \mathbb{R}_+^4.
\end{array}$$

$$s \in \mathbb{R}^1, z \in \mathbb{R}^n, \delta \in \mathbb{R}_+^n, \mu \in \mathbb{R}^1.$$

Proof. Based on (ii) of Observation 1, we take μ to be the integer part of s and $\delta_t = 1$ if $s = f_t \pmod{1}$. Next one observes that after substituting for s , the set

$$= \left\{ (z, \mu, \delta) \in \mathbb{Z}^n \times \mathbb{Z}^1 \times \mathbb{Z}_+^n : \mu + \sum_{i=1}^n f_i \delta_i + z_t \geq b_t \quad \text{for } t \in N, \sum_{t=1}^n \delta_t = 1 \right\}$$

Finally, one observes that the matrix in the z, μ , and δ variables is totally unimodular and the right-hand side vector is an integer, so the integrality of the z, μ , and δ can be dropped.

Example. Consider an instance of X^{MIX} with $n = 4$:

$$\begin{aligned} s + z_1 &\geq 2.9, \\ s + z_2 &\geq 0.7, \\ s + z_3 &\geq -1.4, \\ s + z_4 &\geq 2.0, \\ s &\in \mathbb{R}^1, z \in \mathbb{Z}^4. \end{aligned}$$

The corresponding extended formulation is

Now we show how this inequality can be derived from the extended formulation. We have that

Observation 2 [Nonnegativity of s]. Note that the constraint $s \geq 0$ can be modeled as part of a mixing set by adding the constraints $s + z_{n+1} \geq 0, z_{n+1} = 0$. Now if $T = \{i_1, \dots, i_t, n+1\}$ in Theorem 3, one obtains the additional inequality

$$s \geq \sum_{j=1}^t (f_{i_j} - f_{i_{j+1}})(\lfloor b_{i_j} \rfloor + 1 - z_{i_j}), \quad (17)$$

in which the final term of inequality (16) has disappeared. The terms in these inequalities can be seen as a mixing of the simple MIR inequalities (10). Hence the name mixing sets and mixing inequalities.

It is not difficult to see that the separation problem for the mixing inequalities (16) can be solved in $O(n \log n)$; see Pochet and Wolsey [3].

Observation 3 [Mixing with “ $\leq$ ” Constraints]. The set with “ $\leq$ ” constraints, $X^{\leq} = \{(\sigma, y) \in \mathbb{R}^1 \times \mathbb{Z}^n : \sigma + y_t \leq c_t \text{ for } t \in N\}$ is equivalent to a standard mixing set. It suffices to take $s = -\sigma, z_t = -y_t$, and $b_t = -c_t$ for all t . Note that in this case, it is easy to add an upper bound $s \leq u$ by adding the constraints $\sigma + y_{n+1} \leq u, y_{n+1} = 0$.

NETWORK DUAL GENERALIZATIONS OF MIXING SETS

Here, we consider several extensions of mixing sets and then present a network dual model or bipartite vertex cover model of which the extensions are special cases.

The Double Mixing Set and Bounds on s . This is the set

$$X^{\text{DM}} = \{(s, z) \in \mathbb{R}^1 \times \mathbb{Z}^n : s + z_t \geq b_t \text{ for } t \in N_1, s + z_t \leq c_t \text{ for } t \in N_2\},$$

where $N_1 \cup N_2 = N$. Note that X^{DM} is the intersection of two mixing sets $X^{\text{MIX}}(b)$ and $X^{\leq}(c) = \{(s, z) \in \mathbb{R}^1 \times \mathbb{Z}^n : s + z_t \leq c_t \text{ for } t \in$

$N_2\}$ with the same continuous variable. Conforti *et al.* [4] show

$$\text{conv}(X^{\text{DM}}) = \text{conv}(X^{\text{MIX}}(b)) \cap \text{conv}(X^{\leq}(c)) \\ \bigcap_{(i,j) \in N_1 \times N_2} \{z : z_i - z_j \geq \lceil b_i - c_j \rceil\}.$$

Note that this includes the very special case of a mixing set with both lower and upper bounds on the continuous variable s .

The Continuous Mixing Set. This is the set

$$X^{\text{CM}} = \{(s, r, z) \in \mathbb{R}^1 \times \mathbb{R}_+^n \times \mathbb{Z}^n : \\ s + r_t + z_t \geq b_t \text{ for } t \in N\}.$$

This set was first studied in the context of lot-sizing. The constant capacity Wagner–Whitin relaxation of the problem with backlogging is the set

$$X^{\text{WW-CC-B}} \\ = \left\{ (s, r, y) \in \mathbb{R}_+^1 \times \mathbb{R}_+^n \times \{0, 1\}^n : s_{k-1} \right. \\ \left. + \sum_{u=k}^t y_u + r_t \geq \sum_{u=1}^t d_u \text{ for } 1 \leq k \leq t \leq n \right\}.$$

Fixing k and setting $x_t = \sum_{u=k}^t y_u$, one sees that this set is an intersection of n continuous mixing sets.

A first extended formulation for $\text{conv}(X^{\text{CM}})$ with $O(n^2)$ variables and constraints was given in Ref. 5. Later Van Vyve in Ref. 6 presented an extended formulation with $O(n)$ variables and $O(n^2)$ constraints. He also showed that every facet-defining inequality is a “cycle” inequality. Consider the complete directed graph D on n nodes, and let C be a directed cycle in D . Let γ_C denote the number of arcs $(i, j) \in C$ such that $i > j$. Then the cycle inequality is the following inequality

$$\gamma_C s + \sum_{i \in V(C)} r_i + \sum_{(i,j) \in C: i < j} (f_i - f_j)(z_i - \lfloor b_i \rfloor) \\ + \sum_{(i,j) \in C: i > j} (1 + f_i - f_j)(z_i - \lfloor b_i \rfloor) \geq \sum_{(i,j) \in C: i > j} f_j.$$

He also showed that the separation algorithm could be solved by finding a negative cost cycle in an n -node weighted directed graph.

The Mixing Set with Flows. This is the set

$$X^{\text{MF}} = \{(s, z, x) \in \mathbb{R}^1 \times \mathbb{R}_+^n \times \mathbb{Z}^n : \\ s + z_t \geq b_t, z_t \leq x_t \text{ for } t \in N\}.$$

This set can be used to model a slightly more general version of the discrete newsvendor problem described above in which one can produce an amount z_t that is not a multiple of the batch size C , but one then pays both a per unit cost and a fixed batch cost.

Consider the family of n mixing sets

$$X_k^{\text{MF}} = \{(\sigma_k, x_{k+1}, \dots, x_n) \in \mathbb{R}_+^1 \times \mathbb{Z}_+^{n-k} : \\ \sigma_k + x_{k+t} \geq b_t - b_k \text{ for } t = k+1, \dots, n\}$$

for $k = 0, \dots, n-1$, where $x_0 = b_0 = 0$.

Conforti *et al.* showed in Ref. 7 that

$$\text{conv}(X^{\text{MF}}) = \text{proj}_{s,x,z}(\cap_{k=0}^{n-1} \text{conv}(X_k^{\text{MF}})) \cap \{(\sigma, x) : \\ \sigma_k = s + x_k - b_k \forall k\} \cap \{(z, x) : \\ z_t \leq x_t \text{ for } t \in N\}.$$

The Continuous Mixing Set with Flows. Combining the previous two models, one obtains the continuous mixing set with flows

$$X^{\text{CMF}} = \{(s, r, z, x) \in \mathbb{R}^1 \times \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{Z}^n : \\ s + r_t + z_t \geq b_t, z_t \leq x_t \text{ for } t \in N\}.$$

An extended formulation is given in Ref. 8. Recently, in Ref. 4, Conforti *et al.* consider the family of sets

$$X_Q^{\text{CMF}} = \{(s, r, z, x) \in \mathbb{R}^1 \times \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{Z}^n : \\ s + r_t + z_t \geq b_t \text{ for } t \in N, z_t \geq 0 \\ \text{for } t \in Q, z_t \leq x_t \text{ for } t \in N \setminus Q\},$$

where $Q \subseteq N$.

They show that

$$\text{conv}(X^{\text{CMF}}) = \bigcap_{Q \subseteq N} \text{conv}(X_Q^{\text{CMF}}),$$

and show that the inequalities needed to describe $\text{conv}(X_Q^{\text{CMF}})$ closely resemble the cycle inequalities describing the continuous mixing set, and the separation problem is again a negative cycle problem.

The Enlarged Mixing Set. Given additional values u_t for $t \in N$, this is the set

$$X^{\text{EM}} = \{(s, v, z) \in \mathbb{R}_+^1 \times \mathbb{R}_+^n \times \mathbb{Z}^n : \\ s + \sum_{i=1}^t v_i + z_t \geq b_t, v_t \leq u_t \text{ for } t \in N\}.$$

Such sets can be used to model lot-sizing problems with sales in which the demand in each period is a fixed amount d_t plus a variable amount v_t lying between 0 and u_t [9]. Given $Q \subseteq N$, consider the set

$$X_Q^{\text{EM}} = \{(s + \sum_{i \in N \setminus Q} v_i, z) \in \mathbb{R}_+^1 \times \mathbb{Z}^n : \\ s + \sum_{i \in N \setminus Q} v_i + z_t \geq b_t - \sum_{i \in Q: i \leq t} u_i \text{ for } t \in N\}.$$

This is again a mixing set and it is shown in Ref. 10 that

$$\text{conv}(X^{\text{EM}}) = \bigcap_{Q \subseteq N} \text{conv}(X_Q^{\text{EM}}) \cap \{v \in \mathbb{R}_+^n : v \leq u\}.$$

The optimization problem over such sets is polynomially solvable because the objective function allows one to restrict oneself to the intersection of just $n+1$ of the 2^n mixing sets. However, no polynomial time combinatorial separation algorithm is known.

The Network Dual and Bipartite Vertex Cover Sets

Given a directed graph $D = (V, A)$, $b \in \mathbb{Q}^{|A|}$ and a partition (I, L) of V , the network dual set is the set

$$X^{\text{ND}} = \{x \in \mathbb{R}^{|V|} : x_i - x_j \geq b_{ij} \\ \text{for } (i, j) \in A, x_i \in \mathbb{Z}^1 \text{ for } i \in I, \}$$

introduced in Ref. 11. Here I is the set of indices of the integer variables and L that of the continuous variables. It is not difficult to see that most of the sets described above can be rewritten as network dual sets. As

an example, for the continuous mixing set, it suffices to make the change of variables $\rho_t = s + r_t$ and $y_t = -x_t$. Then $s + r_t + x_t \geq b_t$ becomes $\rho_t - y_t \geq b_t$, and $r_t \geq 0$ becomes $\rho_t - s \geq 0$.

As for the mixing set, a crucial object is the set of fractional values encountered by the continuous variables $x_i, i \in L$ in extremal solutions. Suppose that the K possible values are $1 > f_1 \geq \dots \geq f_K \geq 0$. As for the extended μ formulation of $\text{conv}(X^{\text{MIX}})$, we define the μ variables as follows:

$$\begin{aligned}\mu_k^i &= \lfloor x_i \rfloor \text{ if } x_i < \lfloor x_i \rfloor + f_k \text{ and} \\ \mu_k^i &= \lfloor x_i \rfloor + 1 \text{ if } x_i \geq \lfloor x_i \rfloor + f_k.\end{aligned}$$

Now one obtains an extended formulation by reformulating each constraint $x_i - x_j \geq b_{ij}$ separately

$$x_i = \sum_{\ell=0}^K \mu_\ell^i (f_\ell - f_{\ell+1}) \quad i \in V \quad (18)$$

$$\mu_K^i - \mu_0^i = 1, \quad i \in V \quad (19)$$

$$\mu_k^i - \mu_{k-1}^i \geq 0, \quad 1 \leq k \leq K, \quad i \in V \quad (20)$$

$$\mu_k^i - \mu_{q_k^{ij}}^j \geq \beta_k^{ij} \quad 1 \leq k \leq K, (i, j) \in A \quad (21)$$

$$x_i = \mu_k^i \text{ for } k = 0, 1, \dots, K, i \in I, \quad (22)$$

where $f_{ij} = b_{ij} - \lfloor b_{ij} \rfloor$, and q_k^{ij} and β_k^{ij} are calculated as follows:

$$\begin{aligned}\text{if } f_{ij} - f_{k+1} > 0, \text{ then } q_k^{ij} &= \max\{t : f_t + f_{ij} - f_{k+1} > 1\} \text{ and } \beta_k^{ij} = \lfloor b_{ij} \rfloor + 1, \\ \text{if } f_{ij} - f_{k+1} \leq 0, \text{ then } q_k^{ij} &= \max\{t : f_t + f_{ij} - f_{k+1} > 0\} \text{ and } \beta_k^{ij} = \lfloor b_{ij} \rfloor.\end{aligned}$$

Observing again that the corresponding matrix is TU, one has the following theorem.

Theorem 4. *The polyhedron (18)–(22) is an extended formulation for $\text{conv}(X^{\text{ND}})$.*

Note that even after replacement of μ_k^i by x_i for all $i \in I$, the resulting system still has $K|L|$ additional variables and $K|\tilde{A}|$ constraints of the form (21), where $\tilde{A} = \{(i, j) \in A : i, j \in L\}$.

Corollary 1. *If K is polynomial in the input size of $D = (V, A)$ and b , linear programming over the extended formulation provides a polynomial algorithm for optimization over X^{ND} .*

It is natural to ask when K is of polynomial size. For most of the lot-sizing examples and for the continuous mixing set with flows, it can be shown that $K = O(n^2)$. More generally, it is the structure of the “continuous arc” subgraph $\tilde{D} = (V, \tilde{A})$ that is important. In fact, when the underlying undirected graph of D is acyclic, one can show that K is of polynomial size.

We now consider a second model. Given a bipartite graph $G = (U, V, E)$, $b_e \in \mathbb{Q}^{|E|}$ and a set $I \subset U_1 \cup U_2$, the (mixed integer) bipartite vertex cover set is the set

$$\begin{aligned}X^{\text{BVC}} &= \{x \in \mathbb{R}^{U \cup V} : x_i + x_j \geq b_e \\ &\text{for } e \in E, x_i \in \mathbb{Z}^1 i \in I\}.\end{aligned}$$

Conforti *et al.* [12] first examined the case in which the b_e is an half-integer and gave a complete description of the facet-defining inequalities of $\text{conv}(X^{\text{BVC}})$. In Ref. 4, the network dual and bipartite vertex cover models are shown to be equivalent, and the structure of the facets is studied when the corresponding “continuous edge” subgraph is a tree. It is easy to see that X^{BVC} can be transformed into X^{ND} just by taking $y_j = -x_j$ for all $j \in U_2$. To go in the other direction, one can construct a bipartite graph $G = (V, \bar{V}, E)$, where $\bar{V}$ is a copy of V representing the variables $\bar{x}_i = -x_i$ for $i \in V$, $E = \{(i, j) \in A\} \cup \{(i, i) : i \in V\}$ and $b_e = b_{ij}$ for $(i, j) \in A$, and for (i, i) the constraints are $x_i + \bar{x}_i \geq 0$. Now the face of $X^{\text{BVC}} \cap \{(x, \bar{x}) : x_i + \bar{x}_i = 0 \text{ } i \in V\}$ gives X^{ND} .

Further Generalizations of Mixing Sets

The Divisible Capacity Mixing Set. This is the set

$$\begin{aligned}X^{\text{DCM}} &= \{(s, z) \in \mathbb{R}^1 \times \mathbb{Z}^n : s + C_t z_t \geq b_t \text{ for } t \in N\},\end{aligned}$$

where $C_1|C_2|\dots|C_n$. This set was first studied by Van Vyve [13] as an abstraction of a lot-sizing problem with constant upper and lower

bounds on production. In Ref. 14, an extended formulation, inequalities, and a separation algorithm are given for the special case in which the C_i only take two distinct values and all the variables are nonnegative. Later, de Farias and Zhao [15] presented an extended formulation for $\text{conv}(X^{\text{DCM}})$ and an alternative formulation was produced by Di Summa [16] which could also be extended to deal with bounds on the s variable. A very different formulation is given in Ref. 17. Very recently, Conforti and Zambelli [18] have shown how, by a unimodular transformation, the problem becomes simple to solve and also leads to an extended formulation that is a shortest path problem. However, this approach does not allow one to handle bounds on any of the variables.

The Generalized Mixing Set. When the C_i are not divisible, one obtains the generalized mixing set

$$X^{\text{GMIX}} = \{(s, z) \in \mathbb{R}^1 \times \mathbb{Z}^n : s + C_t z_t \geq b_t \text{ for } t \in N\}.$$

This problem has been shown to be *NP*-hard by Eisenbrand and Rothvoss [19], based on a reduction from a problem of diophantine approximation.

The Divisible Capacity Discrete Lot-Sizing Set.

$$X^{\text{DCLS}} = \left\{ (s_0, y) \in \mathbb{R}_+^1 \times \{0, 1\}^{Kn} : \right. \\ \left. s_0 + \sum_{u=1}^t \sum_{k=1}^K C_k y_u^k \geq b_t \text{ for } t \in N \right\}.$$

Results for this model would provide interesting relaxations for lot-sizing problems with different production alternatives. When $K = 1$, this is just the discrete version of the Wagner–Whitin lot-sizing set presented in the section titled “The Properties of Mixing Sets.” For $K = 2$, it is closely related to the continuous mixing set, but nothing is known for $K \geq 3$.

REMARKS AND OPEN QUESTIONS

Related to the attempt to construct mixing sets as relaxations of general mixed integer sets, Dash and Günlük [20] define the mixing closure of a set and show that it is polyhedral. They also consider the more standard split rank, and show that the mixing inequality (16) has rank at most $n - 1$. Dey [21], on the other hand, has demonstrated a lower bound on the split rank of $\log n$. Another attempt to use mixing sets to obtain valid inequalities for general MIPs is explored by Dey and Wolsey [22], based on lifting a mixing inequality into a valid inequality for any two (or more) row-basic LP tables. The computational effectiveness of using mixing sets and inequalities for general mixed integer sets is still to be investigated.

One significant open question is the complexity of the optimization problem over network dual or bipartite vertex cover sets, even when restricted to certain classes of digraphs. Another concerns the complete structure of the facet-defining inequalities in the case where the underlying “continuous arc” subgraph is a tree.

There are also questions concerning algorithms for the optimization and separation problems for the mixing set and its generalizations. For the optimization problem, one knows that the dual of the extended formulation is a network flow problem, so one might hope to find specialized network flow algorithms.

More generally, it is natural to ask whether there are other simple mixed integer sets with such a rich structure.

REFERENCES

1. Pochet Y, Wolsey LA. Lot-sizing with constant batches: Formulation and valid inequalities. *Math Oper Res* 1993;18:767–785.
2. Günlük O, Pochet Y. Mixing mixed integer inequalities. *Math Program* 2001;90:429–457.
3. Pochet Y, Wolsey LA. *Production planning by mixed integer programming*. New York: Springer; 2006.

4. Conforti M, Wolsey LA, Zambelli G. Projecting an extended formulation for mixed-integer covers on bipartite graphs. University of Padua; 2009.
5. Miller A, Wolsey LA. Tight formulations for some simple MIPs and convex objective IPs. *Math Program B* 2003;98:73–88.
6. Van Vyve M. The continuous mixing polyhedron. *Math Oper Res* 2005;30:441–452.
7. Conforti M, Di Summa M, Wolsey LA. The mixing set with flows. *SIAM J Discrete Math* 2007;29:396–407.
8. Conforti M, Di Summa M, Wolsey LA. The intersection of continuous mixing polyhedra and the continuous mixing polyhedron with flows. In: Fischetti M, Williamson DP, editors. *Proceedings of IPCO XII*. LNCS 4513. Berlin, Heidelberg: Springer; 2007.1–14.
9. Loparic M, Pochet Y, Wolsey LA. The uncapacitated lot-sizing problem with sales and safety stocks. *Math Program* 2001;89:487–504.
10. Conforti M, Di Summa M, Wolsey LA. Enlarged mixing sets. manuscript in preparation, University of Padua; 2010.
11. Conforti M, Di Summa M, Eisenbrand F, *et al.* Network formulations of mixed integer programs. *Math Oper Res* 2009;34:194–209.
12. Conforti M, Gerards B, Zambelli G. Mixed-integer vertex covers in bipartite graphs. In: Fischetti M, Williamson DP, editors. *Proceedings of IPCO XII*, LNCS 4513. Berlin, Heidelberg: Springer; 2007. pp. 324–336
13. Van Vyve M. A solution approach of production planning problems based on compact formulations for single-item lot-sizing models. PhD thesis, Université catholique de Louvain; 2003.
14. Constantino MM, Miller AJ, Van Vyve M. Mixing MIR inequalities with two divisible coefficients. *Math Program* 2010; 00:10.1007/s10107–009–0266–9.
15. de Farias IR, Zhao M. The mixing-mir set with divisible capacities. *Math Program* 2008;115:73–103.
16. Di Summa M. The mixing set with divisible capacities. Université catholique de Louvain: Louvain-la-Neuve: CORE; 2007. Working paper.
17. Conforti M, Di Summa M, Wolsey LA. The mixing set with divisible capacities. In: Panconesi A, Lodi A, Rinaldi G, editors. *Proceedings of IPCO XIII*. LNCS Berlin, Heidelberg: Springer; 2008.435–449.
18. Conforti M, Zambelli G. The mixing set with divisible capacities. *Oper Res Lett* 2009;00:doi:10.1016/j.orl.2009.07.001.
19. Eisenbrand F, Rothvoß T. New hardness results for diophantine approximation. In: *Proceedings of the 12th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems - APPROX 2009*, 2009.
20. Dash S, Günlük O. On mixing inequalities: rank, closure and cutting plane proofs. Yorktown Heights: IBM; 2008. Working paper.
21. Dey SS. A note on the split rank of intersection cuts. Discussion Paper DP 30/2008, CORE, 2008.
22. Dey SS, Wolsey LA. Lifting group inequalities and an application to mixing inequalities. Discussion Paper DP 44/2009, CORE, 2009.

MODEL-BASED FORECASTING

INGRIDA RADZIUKYNIENE
NIKITA BOYKO
PANOS PARDALOS
Department of Industrial and Systems
Engineering and Biomedical
Engineering, University of Florida,
Center for Applied Optimization,
Gainesville, Florida

INTRODUCTION

A forecasting model is a mathematical representation of the forecasted process. It is important to distinguish a model from formula or forecasting method: A method is just a rule or formula for computing a forecast. The latter may or may not depend on a model. The goal of the article is to introduce the most widely used forecasting model to the reader, provide a classification of existing models, and offer a guideline for further study of this area.

The forecasting is based on the data from previous time points, that is, the values of one or more additional time series variables, called *predictor* or *explanatory variables*. These variables are used by the model to forecast the output. The selection (or identification) of the model is performed based on a partial understanding of the forecasting phenomenon and/or scientific intuition. Then, when selected, the goal is to tune the model parameters to fit the data in a certain sense. Lastly, it is important to ensure that the model adequately describes the process using statistical validation techniques.

The section titled “Linear Models” provides an introduction to linear models, that is, models which are described by linear equations. We start from the Autoregressive Integrated Moving Average (ARIMA). The autoregressive part (AR) determines the forecast based on the previous time series values (regressors). The moving average (MA)

term determines the noise of the model. The Box–Jenkins family of models is an example of integrated AR and MA models. State-space models provide forecasting based on the input and an internal state of the model, which is switched over time according to certain transition rules. Bayesian forecasting is done based on some prior knowledge of the model that is iteratively updated when new data are received.

The nonlinear models use nonlinear functions explaining the underlying phenomena. In many cases, they extend corresponding linear models allowing the forecasts to describe a wider set of processes. Nonlinearity does however provide extra computational and theoretical challenges. The section titled “Nonlinear Models,” introduces some popular models of the nonlinear family such as nonlinear dynamic regression models, nonlinear MA models, nonlinear autoregressive (NLAR) models, smooth transition autoregressive (STAR) models, and Markov-switching models.

LINEAR MODELS

In this section, we will consider the linear models for producing forecasts of time series data. Even though the time series contains some nonlinearities, linear forecasts can be considered as a good benchmark or can compete with nonlinear ones [1]. As there is a vast array of linear models, we will focus on the extensive class of ARIMA, state-space, and Bayesian-based forecasting models.

ARIMA Models

The ARIMA class of models is an important forecasting tool, and forms the basis for many fundamental ideas in time series analysis. The ARIMA approach to time series forecasting dominated the statistical literature in the 1970s and early 1980s, and attracted much attention in control engineering. The studies on the characteristics and performance of these models are presented in Refs 2–5.

Autoregressive Processes. The most popular approach for modeling univariate time series is the AR model. A process y_t is said to be an AR process if

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + z_1. \quad (1)$$

where ϕ are the parameters of the model and z_1 is a random variable. The value of p designates the order of the AR model. An AR model is simply a linear regression of the current value of the series against one or more previous values. The simplest AR process is the first-order case denoted by AR(1), and is given as follows:

$$y_t = \phi_1 y_{t-1} + z_t. \quad (2)$$

If $\phi_1 = 1$, we obtain the nonstationary, random walk model. However, if $\phi_1 < 1$, it can be shown that the process is stationary, with autocorrelation function (ac.f.) given by $\rho_k = \phi^k$ for $k = 0, 1, 2, \dots$

Moving Average Processes. Another common approach for modeling univariate time series models is the MA model. The process $\{y_t\}$ is said to be an MA process of order q , if it has the following structure:

$$y_t = z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q}. \quad (3)$$

where $\theta_1 \dots \theta_q$ are the parameters of the model and the $z_t, z_{t-1}, \dots$ are the error terms.

Conceptually, the MA model is a linear regression of the current value of the series against the white noise (or random shocks) of one or more of the previous values of the series. The random shocks at each point $\{z_t\}$ are assumed to come from the same distribution, typically a normal distribution. Unlike AR models, the error terms are not observable and therefore, fitting the model is a more challenging task. Normally, iterative nonlinear fitting procedures are used instead of least-squares to restore the parameters [6]. The simplest example of an MA process is the first-order case, denoted as MA(1), given by

$$y_t = z_t + \theta_1 z_{t-1}. \quad (4)$$

It can be shown that this process is stationary for all values of θ , with an ac.f. given by

$$\rho_k = \begin{cases} 1, & k = 0, \\ \theta/(1 + \theta^2), & k = 1, \\ 0, & k > 1. \end{cases} \quad (5)$$

Box–Jenkins Models. Two statisticians Box and Jenkins proposed to combine AR and MA models, which resulted in the procedure that is known today as *Box–Jenkins forecasting* [2,7]. The notation ARMA(p, q) refers to the model with p AR terms and q MA terms:

$$y_t = c + z_t + \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=1}^q \theta_i z_{t-i}, \quad (6)$$

where c is a constant and z_t is white noise.

Accurate forecasting depends on finding a suitable model for a given time series. For this, the Box–Jenkins approach uses an iterative three-stage procedure:

1. *Model Identification and Plausible Model Selection.* This ensures that the variables are stationary and decide which (if any) AR or MA component should be used in the model.
2. *Fitting the Model.* One uses econometric computation algorithms to obtain coefficients which best fit the selected ARIMA model. The most widely used methods are maximum likelihood estimation or nonlinear least-squares estimation.
3. *Checking.* One tests whether the estimated model fulfils the specifications of a stationary univariate process. If it is necessary, a better model is built.

The most general Box–Jenkins model includes difference operators to cope with nonstationarity, AR terms, MA terms, seasonal difference operators, seasonal AR terms, and seasonal MA terms.

State-Space Model

The state-space model is defined by two equations: an observation equation that describes how the hidden state or latent process is observed and a state equation that defines the evolution of the process through time.

In state-space, the signal at time t is taken to be a linear combination of a set of variables (state variables), which constitute the state vector at time t . Denoting the number of state variables by m , and the $(m \times 1)$ state vector by θ_t , we can write

$$y_t = h_t^T \theta_t + n_t, \quad (7)$$

where h_t is assumed to be a known $(m \times 1)$ vector, and n_t denotes the observation error, assumed to have a zero mean.

The set of state variables may be defined as the smallest possible subset of system variables that can represent the future behavior of the system and is completely determined by the present values of the state variables. Thus, the future is linearly independent of past values. This means that the state vector possesses a Markovian property: the latest value is all that is needed for making predictions.

It may not be possible to directly observe all (or even any) of the elements of the state vector, θ_t , but it may be reasonable to make assumptions about how the state vector changes over time. A key assumption of the linear state-space model is that the state vector evolves according to the equation

$$\theta_t = G_t \theta_{t-1} + \omega_t, \quad (8)$$

where the $(m \times m)$ matrix G_t is assumed to be known and ω_t denotes an m -vector of disturbance having a zero mean. The general form of the univariate state-space model is composed of two equations (Eqs 7 and 8). Equation (7) is known as the *observation* (or *measurement*) equation, while Equation (8) is called the *transition* (or *system*) equation.

Point forecasts can readily be obtained for the general state-space model by applying the *Kalman filter* [8]. This procedure requires knowledge of the most recent state vector and

the exact value of the current state vector

$$\hat{y}_N(h) = h_{N+h}^T \theta_{N+h}, \quad (9)$$

where h_{N+h} is assumed to be known, and $\theta_{N+h} = G_{N+h} G_{N+h-1} \dots G_{N+1} \theta_N$, if future values of G are also known. In practice, the exact value of θ_N is not known and has to be estimated from data up to time N .

Equation (9) is then replaced by an estimate $\hat{y}_N(h) = h_{N+h}^T \hat{\theta}_{N+h}$. In the special case, where h_t and G_t are known constant functions, say h and G , then the forecast is given by

$$\hat{y}_N(h) = h^T G^h \hat{\theta}_N. \quad (10)$$

Assuming the structure of the model is known, the forecasts depend on obtaining strong estimates of the current state vector θ_N . A basic property of state-space models is that the latter quantity can easily be obtained by updating the formula of the Kalman filter as each new observation becomes available [6]. The Kalman filter will also give estimates of the variance–covariance matrix of $\hat{\theta}$ to indicate the likely size of estimation errors.

The Kalman filter has two distinct phases: *prediction* and *update*. The prediction phase uses the state estimate from the previous time step ($t-1$) to get the forecast θ_t at the current time step. If we denote the resulting predicted state by $\hat{\theta}_{t|t-1}$, then according to Equation (8), it can be calculated as follows:

$$\hat{\theta}_{t|t-1} = G_t \hat{\theta}_{t-1}. \quad (11)$$

It can be shown that the variance–covariance matrix of this estimate is given by

$$P_{t|t-1} = G_t P_{t-1} G_t^T + W_t, \quad (12)$$

where W_t is the known variance–covariance matrix.

In the update phase, when the new observation y_t at the current time step t becomes available, the prediction is refined and a new, presumably more accurate, state estimate is obtained again for the current time step using the following equation:

$$\hat{\theta}_t = \hat{\theta}_{t|t-1} + K_t e_t \quad (13)$$

and

$$P_t = P_{t|t-1} - K_t h_t^T P_{t|t-1}, \quad (14)$$

where

$$e_t = y_t - h_t^T \hat{\theta}_{t|t-1} \quad (15)$$

is the prediction error at time t , and

$$K_t = \frac{P_{t|t-1} h_t}{h_t^T P_{t|t-1} h_t + \sigma_n^2} \quad (16)$$

is called the *Kalman gain matrix*.

Bayesian Forecasting

The Bayesian forecasting procedure is based on the dynamic linear model. This model is based on the idea of Kalman filtering and is useful to predict time series data if the generating process changes over time. In an AR model, it would be necessary to make new identifications and estimations of the model if something changed, whereas in a dynamic linear model these changes are included automatically.

Although it looks like a state-space model, the forecasts are based on a Bayesian rationale, which involves updating prior estimates of model parameters to get posterior estimates. The set of recurrence relationships is particularly useful for short series, where the presence of some prior information [6,9] has been confirmed.

Bayesian methods of forecasting are based on two simple principles:

1. *Explicit Formulation.* Formal probability statements as (i) future event of interest and (ii) relevant events observed at the same time as the decisions are used to express all assumptions.
2. *Relevant Conditioning.* The distribution of future events is conditional upon observed relevant events which have to be used.

In the Bayesian method, the information about point estimate θ is contained not in a single point estimate but in the posterior density $p(\theta|y)$, and so prediction is

correspondingly based on averaging $p(z|y, \theta)$ over this posterior density [10]. In general $p(z|y, \theta) = p(z|\theta)$, that is, predictions do not depend on the observations when θ is given. Therefore, in case of discrete θ , the predicted or replicated data z given the observed data y is equal

$$p(z|y) = \sum_{\theta} p(z|\theta) p(\theta|y) \quad (17)$$

and for continuous θ case, the predicted data z is an integral over the product $p(z|\theta) p(\theta|y)$. In the sampling approach, with iterations $t = B + 1, \dots, B + T$ after convergence, this involves iteration-specific samples of $z(t)$ from the same likelihood form used for $p(y|\theta)$, given the sampled value $\theta(t)$.

NONLINEAR MODELS

In practice, many stationary phenomena can be described or at least approximated by stationary linear time series models. However, many nonlinear phenomena cannot be described adequately by linear time series models, unless superfluous parameters are involved [11]. In recent years, the interest in nonlinear time series models has steadily increased, especially in the field of empirical economics. This has yielded a huge amount of nonlinear models, and it is not possible to present all of them in this section. Here, we introduce a number of well-known nonlinear forecasting models excluding nonparametric ones and discuss their properties.

There are two possibilities to produce forecasts from nonlinear models. Sometimes it is possible to use analytical formulas as in linear models. In many other cases, forecasts more than one period ahead have to be generated numerically. We focus on the nonlinear MA and AR models, the Markov-switching or hidden Markov regression model, and a few others.

Nonlinear Dynamic Regression Model

A general nonlinear dynamic model with a noise component can be defined as follows:

$$y_t = f(z_t; \theta) + \varepsilon_t, \quad (18)$$

where $z_t = (w_t', x_t')$ is a vector of explanatory variables, $w_t = (1, y_{t-1}, \dots, y_{t-p})'$, and the vector of strongly exogenous variables $x_t = (x_{1t}, \dots, x_{kt})'$. Furthermore, $\varepsilon_t \sim \text{iid}(0, \sigma^2)$. There are many popular special cases of this model applied in forecasting; these can be found in Ref. 12.

Nonlinear Moving Average Models

A general nonlinear MA model of order q may be defined as follows:

$$y_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}, \theta) + \varepsilon_t, \quad (19)$$

where $\{\varepsilon_t\} \sim \text{iid}(0, \sigma^2)$. The common characteristic of such models is that if the model is invertible, forecasts from it for more than q steps ahead equal the unconditional mean of y_t . In case of unknown invertibility conditions, the models cannot be used for forecasting. When nonlinear MA models are linear in parameters, forecasting is easy in the sense that no numerical techniques are required for the several steps ahead forecasts. Wecker [13] proposed the asymmetric moving average (asMA) model, which has the following form:

$$y_t = \mu + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \sum_{j=1}^q \psi_j I(\varepsilon_{t-j} > 0) \varepsilon_{t-j} + \varepsilon_t, \quad (20)$$

where $I(\varepsilon_{t-j} > 0) = 1$, when $\varepsilon_{t-j} > 0$ and zero otherwise, and $\{\varepsilon_t\} \sim \text{iid}(0, \sigma^2)$. This model has the property that the effects of a positive shock and a negative shock of the same sizes on y_t are not symmetric when $\psi \neq 0$ for at least one $j, j = 1, \dots, q$.

Nonlinear Autoregressive Models

The NLAR model is a natural extension of the linear AR model. It attracted considerable interest in the 1990s. More details on this model can be found in Refs 14–16.

The general NLAR model of order p has the following form:

$$y_t = k(y_{t-1}, y_{t-2}, \dots, y_{t-p}; \theta) + \varepsilon_t, \quad (21)$$

where $\{\varepsilon_t\}$ is a sequence of $\text{iid}(0, \sigma^2)$ variables.

Guo *et al.* have introduced a multistep prediction procedure for NLAR models based on empirical distribution in Ref. 1. Kapetanios has investigated the features of AR processes including the long memory behavior [17]. The relationship between the stationary marginal tail and the innovation's tail probability has been studied in Ref. 18. The research on the stability of NLAR models are presented in Ref. 19.

Switching Regression and Threshold Autoregressive Model. The standard switching regression (SR) model is piecewise linear and is defined as follows:

$$y_t = \sum_{j=1}^{r+1} (\phi_j' z_t + \varepsilon_{jt}) I(c_{j-1} < s_t \leq c_j), \quad (22)$$

where $z_t = (w_t', x_t')$ is defined as before, s_t is a switching variable, c_j are threshold parameters, and $c_0 = -\infty, c_{r+1} = +\infty$. In addition, $\{\varepsilon_{jt}\} \sim \text{iid}(0, \sigma_j^2)$ and $j = 1, \dots, r$. It can be noticed that Equation (22) is a piecewise linear model with generally unknown switch-points.

The threshold autoregressive (TAR) model was introduced by Tong [20]. The first-order TAR has the following form:

$$x_t = \sum_{j=1}^{r+1} (\phi_j' z_t + \varepsilon_{jt}) I(c_{j-1} < s_t \leq c_j). \quad (23)$$

When x_t is absent and $s_t = y_{t-d}, d > 0$, the Equation (22) becomes the self-exciting threshold autoregressive (SETAR) model. The class of SETAR models has been widely employed and their forecasting performance and statistical characteristics have been examined in Refs 21 and 22. This class is typically applied as an extension of AR models allowing a higher degree of flexibility in model parameters through regime-switching behavior [23]. The SETAR model is a useful tool for predicting future values, assuming that when the behavior of the series changes, the series enters a different regime.

The simple first-order SETAR model has the following expression:

$$y_t = \phi_{11} y_{t-1} I(y_{t-1} \leq c) + \phi_{12} y_{t-1} I(y_{t-1} > c) + \varepsilon_t, \quad (24)$$

where $I(A)$ is an indicator function: $I(A) = 1$ when event A occurs and zero otherwise. The necessary and sufficient condition for this SETAR process to be geometrically ergodic is $\phi_{11} < 1$, $\phi_{12} < 1$, and $\phi_{11}\phi_{12} < 1$. Normally, for higher-order models only sufficient conditions exist and these conditions are quite restrictive.

Suppose we have data up to time N and that y_N happens to be larger than threshold c . It is obvious from the model that $\hat{y}_N(h) = E[y_{N+h} | \text{data to time } N] = \phi_{12} y_N$, and so the one-step-ahead forecast is easy to find.

However forecasting two and more steps ahead becomes more complicated, since it depends on whether the next observation exceeds the threshold. Thus, we need to account for the expectation of future errors that are affected by the threshold. Practitioners use a “deterministic” rule, in which future errors are assumed to have a zero value. In case of the above model, if the observation is smaller than the threshold c , we take

$$\hat{y}_N(2) = \phi_{11} \hat{y}_N(1) = \phi_{11} \phi_{12} y_N. \quad (25)$$

However, this is only an approximation to the true conditional expectation namely, $E[y_{N+2} | \text{data to time } N]$. It is possible to write the conditional expectation analytically for a finite fixed delay d , but in general the forecast cannot be found analytically.

Smooth Transition Autoregressive Model

The STAR model was developed and presented by Chan and Tong [24]. The survey on STAR model and its modification can be found in Refs 25 and 26. These models are usually applied to allow a higher degree of flexibility in model parameters through a smooth transition.

A k -dimensional 2-regime STAR(1) can be defined as

$$y_t = (\phi_{1,0} + \phi_{1,1} y_{t-1} + \cdots + \phi_{1,p} y_{t-p} (1 - G(s_t; \gamma, c)) + (\phi_{2,0} + \phi_{2,1} y_{t-1} + \cdots + \phi_{2,p} y_{t-p} G(s_t; \gamma, c) + \varepsilon_t), \quad (26)$$

where $y_t = (y_{1t}, \dots, y_{kt})'$ is a $k \times 1$ vector time series, $\phi_{i,0}$, $i = 1, 2$ are $k \times 1$ vectors, $\phi_{i,j}$, $i = 1, 2, j = 1, \dots, p$ are $k \times k$ matrices.

There are several extensions of the STAR model, for example, STAR model with exponential or logistic transition function (Exponential STAR or Logistic STAR), models allowing for multiple regimes (MRSTAR), time-varying properties in conjunction with regime-switching behavior (TVSTAR), and modeling several times series jointly. More details on these models can be found in Refs 25 and 27–29.

Markov-Switching Model

Another way of addressing the nonlinear nature of the forecasted process is the Markov-switching model. The Markov process, when embedded in the model, allows for switching between regimes over time.

It is said that a stochastic process satisfies the Markovian property, if the probability of being in a particular state depends on the state directly before and not on older states. Transitions between states happen according to fixed probabilities.

The regime-switching model has several sets of parameters. Each set corresponds to the regime of the system (the state of an embedded Markov process) at a certain moment of time. The Markov-switching model can be defined by the following equation:

$$y_t = \sum_{j=1}^r \alpha'_j z_t I(s_t = j) + \varepsilon_t, \quad (27)$$

where $\{s_t\}$ follows a Markov chain. If the order equals one, the conditional probability of the event $s_t = i$ given s_{t-k} , $k = 1, 2, \dots$, is

only dependent on s_{t-1} and equals

$$Pr s_t = i | s_{t-1} = j = p_{ij}, i, j = 1, \dots, r, \quad (28)$$

such that $\sum_i i = 1' p_{ij} = 1$. The transition probabilities p_{ij} are unknown and have to be estimated from the data. The error process ε_t is often assumed to not be dependent on the “regime” or the value of s_t , but the model may be generalized to incorporate that possibility.

Markov-switching models can be applied when the data can be conveniently thought of as having been generated by a model with different regimes such that the regime changes do not have an observable or quantifiable cause [30]. They may also be used when data on the switching variable is not available and no suitable proxy can be found. This is one of the reasons why Markov-switching models have been fitted to interest rate series, where changes in monetary policy have been a motivation for adopting this approach. Modeling asymmetries in macroeconomic series has, as in the case of SETAR and STAR models, been another area of application.

CONCLUSIONS

Model-based forecasting is the identification of the model behind the process of interest, fitting the parameters of the model according to the training set, and performing statistical test validating the selected model with respect to the test set.

This article provides a brief introduction to several popular forecasting models representing a large family of forecasting approaches. We refer the reader to the exhaustive guidelines to the models presented here, as well as many others in Refs 6, 9, and 31–33.

REFERENCES

1. Guo M, Bai Z, An HZ. Multistep prediction for nonlinear autoregressive models based on empirical distributions. *Stat Sinica* 1999;9:559–570.
2. Box GEP, Jenkins GM. Time series analysis: forecasting and control. Revised edition. San Francisco (CA): Holden-Day; 1976.
3. Brillinger DR. Time series data analysis and theory. Expanded edition. New York: McGraw-Hill, Inc.; 1981.
4. Chatfield C. The analysis of time series: an introduction. New York: Chapman and Hall; 1975.
5. Shittu OI, Yaya OS. Measuring forecast performance of ARMA and ARFIMA Models: an application to US Dollar/UK Pound foreign exchange rate. *Eur J Sci Res* 2009;32(2):167–176.
6. Chatfield C. Time series forecasting. London: Chapman & Hall/CRC; 2000.
7. Pankratz A. Forecasting with univariate Box–Jenkins models: concepts and cases. New York: John Wiley & Sons, Inc.; 1983.
8. Kalman RE. A new approach to linear filtering and prediction problems. *Trans ASME–J Basic Eng [Ser D]* 1960;82:35–45.
9. Pole A, West M, Harrison J. Applied Bayesian forecasting and time series analysis. New York: Chapman & Hall/CRC; 1994.
10. Congdon P. Introduction: the Bayesian method, its benefits and implementation. In: Shewhart WA, Wilks SS, editors. Bayesian statistical modelling, Wiley series in probability and statistics. 2nd ed. Chichester, West Sussex: John Wiley & Sons, Inc.; 2006.
11. Guo M, Tseng YK. A comparison between linear and nonlinear forecasts for nonlinear ar models. *J Forecast* 1997;16:491–508.
12. Anderson HM, Vahid F. Predicting the probability of a recession with nonlinear autoregressive leading-indicator models. *Macroecon Dyn* 2001;5:482–505.
13. Wecker WE. Asymmetric time series. *J Am Stat Assoc* 1981;76:16–121.
14. Tiao GC, Tsay RS. Some advance in nonlinear and adaptive modeling in time series. *J Forecast* 1994;13:109–131.
15. Tjostheim D. Nonlinear time series and Markov chains. *Adv Appl Probab* 1990;22:587–611.
16. Tong H. Nonlinear time series: a dynamic system approach. New York: Oxford University Press; 1990.
17. Kapetanios G. Nonlinear autoregressive models and long memory. *Econ Lett* 2006;91:360–368.
18. Jin Y, An H. Nonlinear autoregressive models with heavy-tailed innovation. *China Ser A Math* 2005;48(3):333–3340.

19. Meitz M, Saikkonen P. Stability of nonlinear AR-Garch models. *J Time Ser Anal* 2008;29(3):453–475.
20. Tong H. On a threshold model. In: Chen CH, editor. *Pattern recognition and signal processing*. Amsterdam: Kluwer; 1978.
21. Chan WS, Ng M-W. Robustness of alternative nonlinearity tests for SETAR models. *J Forecast* 2004;23:215–2231.
22. Baharumshah AZ. Forecasting performance of exponential smooth transition autoregressive exchange rate models. *Open Econ Rev* 2006;17:235–251.
23. Clements MP, Franses PH, Smith J, *et al.* Asymptotic and bootstrap prediction regions for vector autoregression. *J Forecast* 2003;22(5):359–375.
24. Chan KS, Tong H. On estimating thresholds in autoregressive models. *J Time Ser Anal* 1986;7:178–190.
25. Dijk D, Terasvirta T, Franses PH. Smooth transition autoregressive models—a survey of recent developments. *Economet Rev* 2002;21(1):1–47.
26. Granger CWJ, Terasvirta T. A simple nonlinear time series model with misleading linear properties. *Econ Lett* 1999;62:161–1165.
27. Granger CWJ, Teräsvirta T. *Modeling nonlinear economic relationships* (advanced texts in econometrics). USA: Oxford University Press; 1993. pp. 198.
28. Chen Y-T. Discriminating between competing STAR models. *Econ Lett* 2003;79:161–167.
29. Escribano A, Jorda O. Testing nonlinearity: decision rules for selecting between logistic and exponential STAR models. *Span Econ Rev* 2001;3:193–209.
30. Terasvirta T. No. 598: Forecasting economic variables with nonlinear models. SSE/EFI Working Paper Series in Economics and Finance. Stockholm: Department of Economic Statistics, Stockholm School of Economics; 2005 Dec 29.
31. Box G, Jenkins GM, Reinsel G. *Time series analysis: forecasting and control*. Upper Saddle River (NJ): Prentice Hall PTR; 1994.
32. Chatfield C. *The analysis of time series: an introduction*. 6th ed. London: Chapman & Hall/CRC; 2003.
33. Tsay RS. *Analysis of financial time series* (Wiley series in probability and statistics). New Jersey: Wiley-Interscience; 2005.

MODELING AND FORECASTING BY MANIFOLD LEARNING

XUELEI (SHERRY) NI
Department of Mathematics
and Statistics, Kennesaw
State University,
Northwest Kennesaw,
Georgia

It is now common to encounter data sets with hundreds or thousands of attributes, instead of the handful ones one had a few decades ago. A key step to the success of data analysis is to model the data with fewer attributes, while capturing the main characteristics after the reduction of dimensionality. Manifold-based learning is a new and promising approach in nonparametric dimension reduction. In this article, we review state-of-the-art developments in manifold learning, as well as its interesting application in forecasting.

We start with recalling the definition of a manifold. In mathematics, a manifold is a topological space that is locally Euclidean (i.e., around every point, there is a neighborhood that is topologically identical with the open unit ball in $\mathbb{R}^d$). As an example, in a one-dimensional manifold, every point has a neighborhood that is topologically identical with a line segment. A line is a 1D manifold, so is a curve. Similarly in a two-dimensional manifold, every point has a neighborhood that is like a disk. Examples include a plane, a Möbius band, and the sphere of the Earth. Locally at every point on the surface of the Earth, we see a plane (2D), but it actually curves back on itself and is globally a 3D sphere instead of a plane (i.e., 2D manifold in a 3D space).

Manifolds offer a powerful framework for dimension reduction. The key idea of dimension reduction is to find the most succinct low-dimensional structure that is embedded in a higher dimensional space. Many high-dimensional data sets can be modeled as sets of points lying close to a low-dimensional manifold. Given a set of data

points $x_1, x_2, \dots, x_N \in \mathbb{R}^D$, we can assume that they are sampled from a manifold with noise, that is,

$$x_i = \phi(y_i) + \varepsilon_i, \quad i = 1, \dots, N, \quad (1)$$

where $y_i \in \mathbb{R}^d$, d is usually far less than the original dimension D ($d \ll D$), and ε_i is noise. Function $\phi : C(\subset \mathbb{R}^d) \rightarrow \mathbb{R}^D$ represents the embedding of a d -dimensional manifold M in a D -dimensional space. Here C is a compact subset of $\mathbb{R}^d$ with open interior. Integer d is also called the *intrinsic dimension*. Manifold-based methodologies offer ways to find the embedded low-dimensional feature vectors, y_i , from the high-dimensional data points, x_i . Some methods also analyze how to reconstruct ϕ from x_i 's.

Manifolds also offer powerful tools in forecasting. The following simple example illustrates the key idea. Figure 1 plots a simulated time series data $x_t \in \mathbb{R}$, $t = 1, 2, \dots, 256$. There is no obvious rule to estimate the future value. However, in Fig. 2, where x_t versus x_{t-1} is plotted, it shows that all the 2D points, $\tau_t = (x_{t-1}, x_t)^T$, $t = 2, \dots, 256$, are around the solid curve (a one-dimensional manifold). That is,

$$\tau_t = \phi(y_t) + \varepsilon_t, \quad t = 2, \dots, 256,$$

where $y_t \in \mathbb{R}$ is one-dimensional. If we can reconstruct the manifold, ϕ , that is, if we can find the method to construct $(x_{t-1}, x_t)^T$, we can use this manifold structure to make one-step forecasting: $x_t = f(x_{t-1})$. Note here f is the relationship between x_t and x_{t-1} found from the reconstruction.

In general, we consider the forecasting framework:

$$x_t = f(x_{t-1}, \dots, x_{t-p}), \quad (2)$$

that is, predicting the value at the t th occasion via the previous p observed values. We have two challenges: (i) x_t can be high dimensional; (ii) the function f may not be in any known functional class. Challenge (i) can be addressed by dimension reduction. Evidently

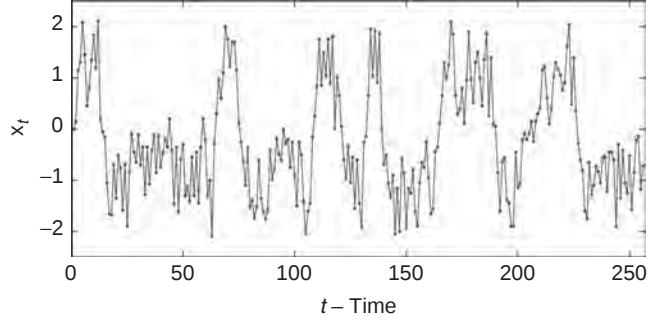


Figure 1. A time series data set.

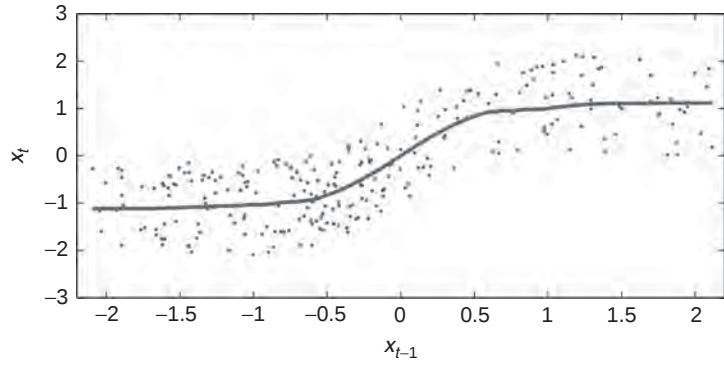


Figure 2. x_t versus x_{t-1} .

in some cases, nonlinear dimension reduction such as manifold-based methods can work well. It is not straightforward that challenge (ii) can be resolved via manifold-based methods as well. Note that if relation (Eq. 2) holds, then vectors $(x_{t-p}, \dots, x_t)$, for a range of t , reside on a low-dimensional manifold. Hence, determining f is equivalent to finding this manifold. After knowing the manifold, we can utilize reconstruction to predict according to Equation (2).

From the above, it is evident that we have two major tasks: finding the manifold from the high-dimension data, and reconstructing based on a learned manifold model. Note that there are many distinct manifold-based algorithms. It is important to understand their properties. We survey the existing manifold learning algorithms, and discuss their reconstructions. Their applications in forecasting are relatively underdeveloped. There is not much to be surveyed at this point, except for a few works such as Ref. 1. The promises of manifold-based forecasting have been realized by more and more researchers and practitioners.

This article is organized as follows. The section titled “Review of Manifold Learning Methods” reviews existing manifold learning methods, focusing on the dimensionality reduction. The methods surveyed include principal components analysis (PCA), multidimensional scaling, locally linear embedding (LLE), isometric feature mapping (ISOMAP), Laplacian eigenmaps, Hessian eigenmaps, and local tangent space alignment (LTSA). The section titled “Reconstruction” summarizes the reconstruction, which is a significant step for forecasting. The section titled “A Synthetic Example” gives a simple simulated example. The section titled “Another Framework” discusses another framework of utilizing manifold learning in forecasting for high dimensional data. Finally, the section titled “Conclusion” concludes the article.

REVIEW OF MANIFOLD LEARNING METHODS

On the basis of the discussion above, the first major task for manifold-based learning is to

find a coordinate system that characterizes the set of data points in a low-dimensional space. Given Model (1), the task is to find the estimation of the unknown lower-dimensional feature vectors y_i 's from the x_i 's. Note that x_i 's are the data points in $\mathfrak{R}^D$, while y_i 's are the points in $\mathfrak{R}^d$. With $d < D$ this task performs dimension reduction of the data points.

In this section, we review some manifold learning techniques, with the focus on dimension reduction. Our goal is to expose the reader to the issues involved and to describe some of the advanced approaches. We begin with the classical methods: PCA [2] and multidimensional scaling (MDS) [3–6] and then, the two fundamental nonlinear manifold searching methods: LLE [7,8] and ISOMAP [9,10]. LLE and ISOMAP appeared in the same issue of *Science* and inspired many other nonlinear manifold learning methods. We will provide a quick overview of three developments: Laplacian eigenmaps [11,12], Hessian eigenmaps (HLE, Hessian locally linear embedding) [13], and LTSA [14].

Owing to space, there are many other techniques not discussed here, such as principal curves and surfaces, nonlinear PCA, independent components analysis, self-organizing maps [15], generative topographic mapping (GTM) [16], and charting [17]. Descriptions of these techniques and additional references can be found in surveys [18,19]. The second survey [19] also provides applications of manifold learning methods on denoising, curve clustering, image detection, and the localization of sensor networks.

Principal Component Analysis (PCA)

PCA has been widely used for dimension reduction in a number of fields. A comprehensive discussion of PCA can be found in Ref. 2. PCA is closely related to singular value decomposition (SVD) [20]. These two techniques are the most classical methods in dimensional reduction. The key idea of PCA is to find the low-dimensional linear subspace that captures as much of the variability as possible.

Suppose $\mathbf{X} \in \mathfrak{R}^D$ is a random vector, whose variance–covariance matrix is Ω , that

is, $\text{Cov}(\mathbf{X}) = E\{[\mathbf{X} - E(\mathbf{X})][\mathbf{X} - E(\mathbf{X})]^T\} = \Omega$. Let $0 \leq \lambda_D \leq \lambda_{D-1} \leq \dots \leq \lambda_1$ be the ordered eigenvalues of Ω , and $U = [u_1, u_2, \dots, u_D]$ be the orthonormal matrix of eigenvectors, where u_i is the eigenvector corresponding to the i th largest eigenvalue λ_i .

PCA makes the following transformation $\mathbf{X} \rightarrow \mathbf{Y}$:

$$\mathbf{Y} = [u_1, \dots, u_d]^T \mathbf{X} = \begin{pmatrix} u_1^T \mathbf{X} \\ u_2^T \mathbf{X} \\ \vdots \\ u_d^T \mathbf{X} \end{pmatrix}, \text{ where } d \leq D.$$

The new vector $\mathbf{Y}$ is in a lower-dimensional space, $\mathfrak{R}^d$. It is provable that $\lambda_1, \lambda_2, \dots$, and λ_d are the variances of the transformed random variables $u_1^T \mathbf{X}, u_2^T \mathbf{X}, \dots$, and $u_d^T \mathbf{X}$. And these new variables keep the greatest possible proportion of the variation in the data. Consequently, they are called *principal components*.

PCA gives a natural dimension reduction. Consider an extreme case: if all the data lie in a low-dimensional linear subspace of a very high-dimensional space, PCA will find such a linear subspace, because the variations in the directions that are orthogonal to the embedded linear subspace will be equal to zero.

An evident disadvantage of PCA is that the embedded subspace has to be linear. For example, if the data are located on a circle in 3D, PCA will not be able to identify such a structure.

Multidimensional Scaling (MDS)

MDS is also often used for dimensionality reduction. MDS refers to a group of methods that have found a wide range of applications. Although a number of variations have been proposed, the key idea is the same: find a mapping to a low-dimensional space such that the pairwise distances between the observed points are preserved as much as possible. MDS starts from a distance (or dissimilarity) matrix, whose entry d_{lm} is the distance between the l th and m th objects. Let d'_{lm} denote the distance between the same two objects after they have been transformed

to the low-dimensional space. Classical multidimensional scaling (CMDS, that is, the model with only one distance matrix) tries to make the mapping by solving the following optimization problem:

$$\min \sum_{l \neq m} [d'_{lm} - d_{lm}]^2.$$

The objective function is also called *stress* in the MDS literature. The above stress is named *raw stress*, which is the most elementary loss function for MDS. Some other stress definitions include:

- Normalized stress:

$$\frac{\sum_{l \neq m} [d'_{lm} - d_{lm}]^2}{\sum_{l \neq m} d_{lm}^2};$$

- Kruskal's stress:

$$\frac{\sum_{l \neq m} [d'_{lm} - d_{lm}]^2}{\sum_{l \neq m} d_{lm}'^2};$$

- Weighted normalized stress:

$$\frac{\sum_{l \neq m} w_{lm} [d'_{lm} - d_{lm}]^2}{\sum_{l \neq m} w_{lm} d_{lm}'^2}; \text{ and so on.}$$

More techniques can be found in Ref. 3.

Those aforementioned stresses are all for the metric MDS, where the distance measure is quantitative. An intuitive example of the metric CMDS is to recover the relative positions of cities from the intercity mileages [21] (also see the example in Ref. 22). Soon after the introduction of metric MDS, non-metric MDS was proposed [4,6], in which, the data could be categorical so that the distance measure is at the ordinal level. In such case, a monotone transformation of the observed original distance, d_{lm} was applied to reproduce the general rank ordering of distances between the objects in the analysis.

For example, the generalized Kruskal stress becomes the following:

$$\frac{\sum_{l \neq m} [d'_{lm} - f(d_{lm})]^2}{\sum_{l \neq m} d_{lm}'^2},$$

where f is a monotone increasing function. For any fixed set of objects, f is specified. Again, we refer technical details to Refs 3,4, and 6.

We will not discuss the details of any specific MDS algorithm, except to say that the typical approach is to initially assign objects to lower-dimensional points in some manner and then try to modify the points to reduce the stress. In most existing MDS algorithms, a linear subspace is still the ultimate result.

MDS is a very useful tool when the interpoint distances need to be preserved. In ISOMAP, which is a method that will be described later, MDS is applied to geodesic distances. The above results in a nonlinear dimension reduction method. We will give more details in section titled "ISOMAP".

Locally Linear Embedding (LLE)

Given a set of data points $x_1, x_2, \dots, x_N \in \mathbb{R}^D$, LLE analyzes overlapping local neighborhoods in order to determine the local structure. The local structure is then preserved in the estimation of the lower-dimensional feature vectors $y_1, y_2, \dots, y_N \in \mathbb{R}^d$. The LLE algorithm is given below [7].

1. Find the k nearest neighbors of each data point based on Euclidean distance. Let N_i denote the set of indices of the k nearest neighbors of x_i .
2. Express each data point, x_i , as a linear combination of the k nearest neighbors. That is, find w such that the following objective function is minimized:

$$E(w) = \sum_i \left\| x_i - \sum_{j \in N_i} w_{ij} x_j \right\|^2,$$

where $\|\cdot\|$ is the l_2 norm and $\sum_{j \in N_i} w_{ij} = 1$.

It can be shown that the above can be solved as a least-squares problem. The weight w_{ij} indicates the contribution of the j th data point to the representation of the i th data point.

3. Find the low-dimensional points, y_i , which minimize the following objective function:

$$\Phi(y) = \sum_i \left\| y_i - \sum_{j \in N_i} w_{ij} y_j \right\|^2.$$

This step performs the actual dimensionality reduction. Note that the low-dimensional points, y_i , preserve the local convex representation, that is, the weights, w_{ij} , obtained from the last step are used in the construction of y_i . It can be shown that with some additional conditions, which make the problem well defined, the minimization task can be accomplished by solving a sparse $N \times N$ eigenvector problem. More specifically, the d eigenvectors associated with the d smallest nonzero eigenvalues provide an ordered set of orthogonal coordinates centered on the origin. We refer interested readers to Ref. 7 for the details of this formulation.

The LLE authors suggest that $k-b$ trees can be used to compute the k nearest neighbors efficiently [23]. The sparse eigenvector problem can be solved by fast algorithms as well [24]. One disadvantage of LLE is that it implicitly assumes that the manifold is convex. The methods that will be described later can overcome such a disadvantage.

ISOMETRIC FEATURE MAPPING (ISOMAP)

ISOMAP [9] is another nonlinear dimension reduction method. It can be viewed as an extension of metric MDS, by replacing the Euclidean distance with another type of distance. MDS and PCA are not good at dimensionality reduction when the points have a complicated, nonlinear relationship. However, ISOMAP can handle such data sets. The method works as follows:

1. Find the nearest neighbors of each data point $x_i \in \mathbb{R}^D, i = 1, 2, \dots, N$, based on Euclidean distance. Let N_i denote the set of indices of the nearest neighbors of x_i . There are two methods to find the nearest neighbors: either take the k nearest points of each point x_i , where k is a parameter (k nearest neighbors); or take all the points that are no more than a specified distance (ε) away from x_i (ε neighborhoods).
2. Redefine the distances between points to be the geodesic distance, which is the shortest curve that is on the manifold and connects the two points. The geodesic distances can be approximated as follows: First, construct a graph in which each point is a node, and two nodes x_i and x_j are connected if and only if $i \in N_j$ or $j \in N_i$. Then define the distance between x_i and x_j to be the sum of the arc lengths of the shortest chain connecting x_i and x_j . The distance is called a *graphical distance*.
3. Generate the low-dimensional points by calling a metric MDS on the new distance matrix.

The two kinds of neighborhoods in Step 1 have pros and cons. The k nearest neighbors has the advantage that it does not tend to disconnected graphs and the parameter k is easier to choose. However, it is less geometrically intuitive. The other choice, ε neighborhoods, is geometrically motivated, but often leads to graphs with several disconnected parts, and it is not obvious on how to choose ε .

The shortest chain (i.e., the graphic distance) in Step 2 can be computed via dynamic programming [25]. Bernstein *et al.* [26] show that the graphical distance is in some sense a good surrogate of the geodesic distance. Note that a graphical distance is computable from data, while geodesic distance is not observable.

Laplacian Eigenmaps

Laplacian eigenmaps are proposed in Refs 11 and 12. This work is a member of the ISOMAP-LLE family of algorithms, which build a graph incorporating neighborhood

information of the data set and compute for a low-dimensional representation of the data set that optimally preserves local neighborhood information in a certain sense. The justification of the Laplacian eigenmaps comes from the role of the Laplace–Beltrami operator and graph Laplacian, which are well known to specialists in spectral graph theory [27]. We first describe the algorithm for discrete data. Its justification will follow.

Given $x_1, x_2, \dots, x_N \in \mathbb{R}^D$, Laplacian eigenmaps looks for the embedded low-dimensional vectors $y_1, y_2, \dots, y_N \in \mathbb{R}^d$ by the procedure as follows:

1. Find the nearest neighbors of each data point, and use N_i to denote the set of indices of the nearest neighbors of x_i .
2. Similar to the second step in ISOMAP, construct a weighted graph in which each point is a node, and two nodes x_i and x_j are connected by an edge if and only if $i \in N_j$ or $j \in N_i$. For any pair of connected nodes i and j , assign following weight for the edge (heat kernel):

$$W_{ij} = \exp \left\{ -\frac{\|x_i - x_j\|^2}{t} \right\}, \quad (3)$$

where $\|\cdot\|$ is the l_2 norm, $t \in \mathbb{R}$ is a parameter.

3. Compute the eigenvalues and eigenvectors for the generalized eigenvector problem:

$$Lf = \lambda Gf, \quad (4)$$

where $f \in \mathbb{R}^N$ is eigenvector, $\lambda \in \mathbb{R}$ is eigenvalue, G denotes the diagonal weight matrix such that $G_{ii} = \sum_j W_{ji}$. Suppose W is the symmetric matrix with entries W_{ij} , then $L = G - W$ is the Laplacian matrix (graph Laplacian). Let $f_0, f_1, \dots, f_{N-1}$ be the solution vectors with corresponding eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{N-1}$; that is,

$$\begin{aligned} Lf_0 &= \lambda_0 Df_0, \\ Lf_1 &= \lambda_1 Df_1, \\ &\vdots \\ Lf_{N-1} &= \lambda_{N-1} Df_{N-1}. \end{aligned}$$

4. Leave out the first eigenvector f_0 , which corresponds to eigenvalue 0. Use the next d eigenvectors for the maps:

$$x_i \rightarrow (f_1(i), f_2(i), \dots, f_d(i))^T.$$

The edge weight in Step 2 has an alternative choice: $W_{ij} = 1$ if nodes i and j are connected. This simplification avoids the necessity of choosing parameter t . The weight (Eq. 3) is motivated by heat flow, which is intimately related to the Laplace–Beltrami operator on differentiable functions on a manifold. The Laplace–Beltrami operator is the justification of the algorithm in the continuum.

An intuitive justification for solving the eigenvalue and eigenvector problem (Eq. 4) is to consider a “good” mapping minimizing the following objective,

$$\sum_{i,j} \|y_i - y_j\|^2 W_{ij} = \text{tr}(Y^T L Y). \quad (5)$$

Note that based on the definition of edge weight, W_{ij} is higher if nodes i and j are closer. Hence, the objective function (Eq. 5) punishes heavier if neighboring points x_i and x_j are mapped far apart. Therefore, minimizing Equation (5) keeps the “closeness” of the neighborhood. It is shown in Ref. 12 that the solution of Equation (5) provided by the matrix of eigenvectors corresponds to the lowest eigenvalues of the generalized eigenvalue problem (Eq. 4).

Belkin and Niyogi also explained in Ref. 12 that in the continuum, the Laplace–Beltrami operator on manifolds ($\mathcal{L}$) is analogous to the Laplacian of a graph (L). Chung [27] serves as a good reference. And, in order to optimally preserve locality on average, in the continuum, Ref. 12 shows that it leads to solving the following:

$$\underset{\|f\|_{L^2(M)}=1}{\text{argmin}} \int_M \|\nabla f(x)\|^2 = \underset{\|f\|_{L^2(M)}=1}{\text{argmin}} \int_M \mathcal{L}(f)f,$$

where the integral is taken with respect to the standard measure on a Riemannian manifold M , which is isometrically embedded in $\mathbb{R}^D$. f is a map from the manifold M to $\mathbb{R}$, which is twice differentiable. Note that minimizing

$\int_M \|\nabla f(x)\|^2$ corresponds directly to minimizing Equation (5) on a graph. The spectrum of L on a compact manifold M is known to be discrete. The rest of the dimension reduction is identical with the approach in the discrete case.

Hessian Eigenmaps (HLE)

The connection between spectral theory and dimension reduction, which is established in Laplacian eigenmaps, is very inspiring. This connection leads to the creation of Hessian eigenmaps (HLE) [13]. While all the previous methods do not give good results for nonconvex manifolds, Hessian eigenmaps overcomes the convexity limitation.

The conceptual framework of HLE in the continuum can be viewed as a modification of the Laplacian eigenmaps. Recall that in Laplacian eigenmaps, the following functional is considered:

$$\int_M \mathcal{L}(f) f.$$

In HLE, the above functional is replaced with

$$H(f) = \int_M \|H_f(m)\|_F^2 dm,$$

where M is the manifold, $m \in M$, $H_f(m)$ is the Hessian of function f , which is defined by orthogonal coordinates on the tangent plane of M , $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Therefore, $\mathcal{H}(f)$ measures the average, over the data manifold M , of the Frobenius norm of the Hessian H of f . Donoho and Grimes [13] proved that $\mathcal{H}$ has $(d+1)$ -dimensional null-space, consisting of the constant function and d -dimensional spaces spanned by the original isometric coordinates. This result establishes the fundamental correctness of HLE, while no similar result is found for Laplacian eigenmaps.

The algorithm proposed in Ref. 13 is based on a discrete estimation of tangent coordinates, the approximation to the tangent Hessian H , and the empirical version of the operator $\mathcal{H}$. The algorithm requires to solve N separate k -by- k eigen problems, where N is the sample size, and k is the neighborhood size. Additionally, similar to the LLE

algorithm, HLE finally needs to solve a single N -by- N sparse eigen problem. We omit the details here and refer the readers to the original article for details.

Local Tangent Space Alignment (LTSA)

LTSA [14] proposed by Zhang and Zha draws inspiration from the pioneering work in LLE. LTSA learns the local geometry of the manifold by constructing a local tangent space for each data point, which is similar to the first steps of HLE. However in HLE, the local tangent spaces are used to approximate Hessian, while in LTSA, the tangent subspaces are aligned together to give the global coordinates of the feature vectors. Given N, D -dimensional data points $x_1, x_2, \dots, x_N$, LTSA works as follows:

1. Identify k nearest neighbors for each data point $x_i, i = 1, 2, \dots, N$. As usual, let N_i denote the set of indices of the nearest neighbors of x_i . Let $X_i = [x_{i_1}, \dots, x_{i_k}]$ be a $D \times k$ matrix consisting of x_i 's nearest neighbors including x_i .
2. Obtain recentered neighborhoods. That is, for each data point, compute the $D \times k$ matrix $X^i = [x_{i_1} - \bar{x}_i, \dots, x_{i_k} - \bar{x}_i]$, where $\bar{x}_i$ is the average of all points in N_i . In matrix format, $X^i = X_i \left(I - \frac{1}{k} \mathbf{1}\mathbf{1}^T\right)$, where $\mathbf{1} \in \mathbb{R}^k$ is an all-one vector.
3. Find the d -dimensional affine subspace approximation of all the points in X_i by using the local tangent subspace. Specifically, suppose $\theta_{i_j} \in \mathbb{R}^d$ is the local parameterization for $x_{i_j}, i_j \in N_i$. The local parameterization matrix $\Theta_i = [\theta_{i_1}, \dots, \theta_{i_k}] \in \mathbb{R}^{d \times k}$ for the points in the neighborhood X_i is found by solving the following:

$$\begin{aligned} & \operatorname{argmin}_{\Theta_i, Q_i} \|X^i - Q_i \Theta_i\|_F^2 \\ & = \operatorname{argmin}_{\Theta_i, Q_i} \sum_{i_j \in N_i} \|(x_{i_j} - \bar{x}_i) - Q_i \theta_{i_j}\|_2^2, \end{aligned} \quad (6)$$

where $Q_i \in \mathbb{R}^{D \times d}$ forms an orthonormal basis of the tangent subspace, that is, $Q_i^T Q_i = I_d$. The solution of Q_i and Θ_i can be found by the SVD of X^i .

4. Compute the global feature coordinates $y_i \in \mathbb{R}^d, i = 1, 2, \dots, N$, by solving the following:

$$\underset{Y, L_i}{\operatorname{argmin}} \sum_{i=1}^N \|Y^i - L_i \Theta_i\|_F^2, \quad (7)$$

where $\|\cdot\|_F$ stands for the Frobenius norm of a matrix, $Y^i = [y_{i1} - \bar{y}_i, y_{i2} - \bar{y}_i, \dots, y_{iK} - \bar{y}_i]$ is the recentered matrix of y_i 's inside neighborhood N_i . The optimal L_i is given by $L_i = Y^i \Theta_i^+$. The solution for Y is provided by solving a sparse $N \times N$ eigenvector problem. Again, we omit the details here and refer the readers to Ref. 14 for details and fast implementations.

LTSA and LLE share the same structure: a local parameterization is first constructed and then a global alignment is computed. The difference is how the local information is summarized and preserved. There is another method similar to LTSA, charting [17], which is not equally well-developed mathematically.

RECONSTRUCTION

Given a new feature vector in the embedded low-dimensional space, the reconstruction method is used to find its counterpart in the high-dimensional space. That is, learning the mapping ϕ in Model (1). We have seen that reconstruction accuracy is critical for the application of manifold learning in the prediction. However, there are a limited number of reconstruction methods in the literature. For a specific linear manifold, the reconstruction can be easily made by PCA. For a nonlinear manifold, LLE reconstruction, which is derived in the similar manner as LLE, is introduced in Ref. 7. LTSA reconstruction and nonparametric regression reconstruction are introduced in Ref. 14. The summaries of these reconstruction methods are provided in the following.

PCA Reconstruction

Retrieving the old data is straightforward if the PCA transformation is originally used for dimension reduction. Recall that the PCA transform is $\mathbf{Y} = [u_1, \dots, u_d]^T \mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^D$ is the original high-dimensional random point, $\mathbf{Y} \in \mathbb{R}^d, d \leq D$, is the lower-dimensional feature vector, $U_d = [u_1, u_2, \dots, u_d]$ contains the first d columns of the eigenvector matrix U . Recall that U is orthonormal, that is, $U^T U = I_D$. If we take all the eigenvectors in the transformation, that is, $\mathbf{Y} = U^T \mathbf{X}$, then getting the original data back is very easy: $\mathbf{X} = U \mathbf{Y}$. This formula also applies to the case when we do not have all the eigenvectors in the transformation ($d < D$). That is, we make the return trip through $\hat{\mathbf{X}} = U_d \mathbf{Y}$. The retrieved data $\hat{\mathbf{X}}$ has lost some information because $U_d U_d^T \neq I_D$. But since the PCA transformation preserves the greatest proportion of variance along the principle eigenvectors, the information lost is the variation in the other components (the other eigenvectors that were left out).

LLE Reconstruction

LLE reconstruction, which is introduced together with the LLE dimension reduction in Ref. 7, is derived in a similar manner as LLE. Given low-dimensional feature vectors $y_i \in \mathbb{R}^d, i = 1, 2, \dots, N$. Denote the new low-dimensional feature vector as y_0 . LLE reconstructs its high-dimensional correspondence x_0 as follows:

1. Identify the k nearest neighbors of y_0 among $y_1, y_2, \dots, y_N$. Let N_0 denote the neighborhood indices.
2. Express y_0 as a linear combination of the k nearest neighbors. That is, the following objective function is minimized.

$$E(w) = \left\| y_0 - \sum_{j \in N_0} w_j y_j \right\|^2, \quad (8)$$

where $\|\cdot\|$ is the l_2 norm and $\sum_{j \in N_0} w_j = 1$.

3. x_0 is constructed by $\hat{x}_0 = \sum_{j \in N_0} w_j x_j$.

Recall that LLE preserves the local convex representation (w_j 's) in the dimension reduction, the same structure is used in the reconstruction. Solving Equation (8) is equivalent to solving a linear system of equations. However, when $k > d$, the coefficient matrix associated with the system of linear equations is singular, which means that the solution is not unique. Saul and Rowe [28] propose to solve this issue by adding an identity matrix multiplied with a small constant to the coefficient matrix.

LTSA Reconstruction

The derivation of the LTSA algorithm does not consider nonlinear dimension reduction in isolation as merely constructing a nonlinear projection, but rather to combine it with the process of the reconstruction of the nonlinear manifold. Therefore, the LTSA reconstruction procedure appears along with the dimension reduction procedure in Ref. 14. In general, when the low-dimensional coordinates y_i are available, the LTSA reconstruction retrieves the high-dimensional data as follows:

1. Given $y_0 \in \mathbb{M}^d$, identify its closest point among $y_1, y_2, \dots, y_N$. Suppose y_i is the nearest neighbor.

2. Calculate the local tangent parameterization of y_0 :

$$\theta_0 = L_i^{-1}(y_0 - \bar{y}_i),$$

where $\bar{y}_i$ is the average of all y_{i_k} 's in the neighborhood N_i , L_i is the optimal alignment matrix that solves Equation (7).

3. Construct x_0 by $\hat{x}_0 = \bar{x}_i + Q_i \theta_0$, where Q_i solves Equation (6).

As seen from the aforementioned procedure, the LTSA reconstruction is exactly the reverse of LTSA dimension reduction. The errors of the reconstructed manifold depend on the sample errors, the local tangent subspace reconstruction errors, and the alignment errors. Zhang and Zha [14] prove that this dependence is linear.

Other Reconstructions

Many other reconstruction methods can be proposed. For example, in MDS, paired distance is preserved. We can reconstruct the original data by keeping those distances too. LLE reconstruction can be used after a HLLE dimension reduction. In Ref. 14, a nonparametric regression reconstruction is proposed. Specifically, once y_i are computed for each of the data points x_i , some nonparametric regression methods such as local polynomial regression can be applied to $\{(y_i, x_i)\}_{i=1}^N$ to

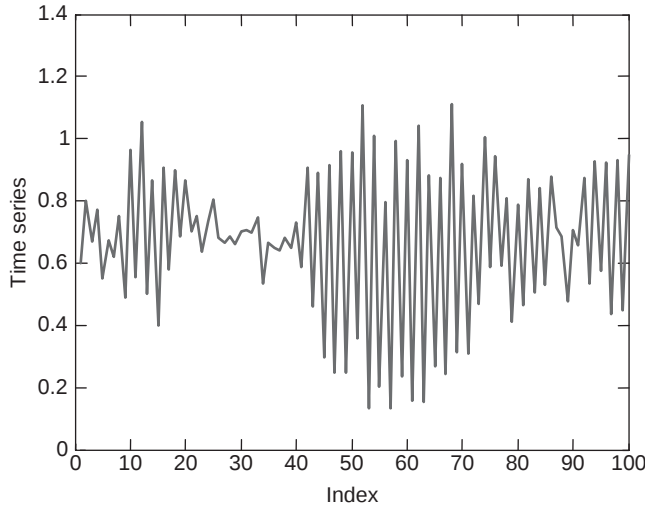


Figure 3. First 100 data points in the generated time series.

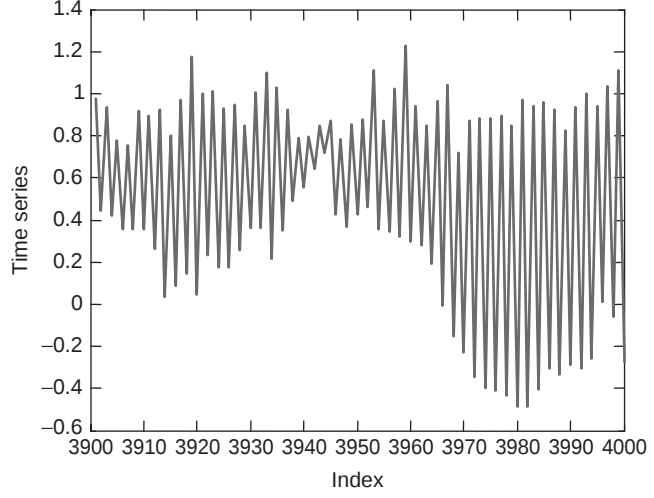


Figure 4. Last 100 data points in the generated time series.

construct the manifold underlying the set of points x_i . That is, we can assume that the manifold function ϕ in Equation (1) is locally polynomial, and we construct ϕ separately in each neighborhood.

A SYNTHETIC EXAMPLE

For illustration purpose, we consider a length-4000 time series, which follows the

following rule: $x_i = \frac{x_{i-2}}{\sqrt{x_{i-2}^2 + x_{i-1}^2}} + \varepsilon_i$, where $i = 3, \dots, n$, $\varepsilon_i \sim \text{Normal}(0, 0.01)$. The first and last 100 data points are plotted in Figs 3 and 4.

Evidently, from x_{i-1} and x_{i-2} , there is a fundamental rule to determine the mean value of x_i . However, such a rule is hidden in the data. If we plot x_{i+1} against x_i , we observe a circle as in the following Fig. 5. Note the value of x_{i+1} cannot be uniquely determined

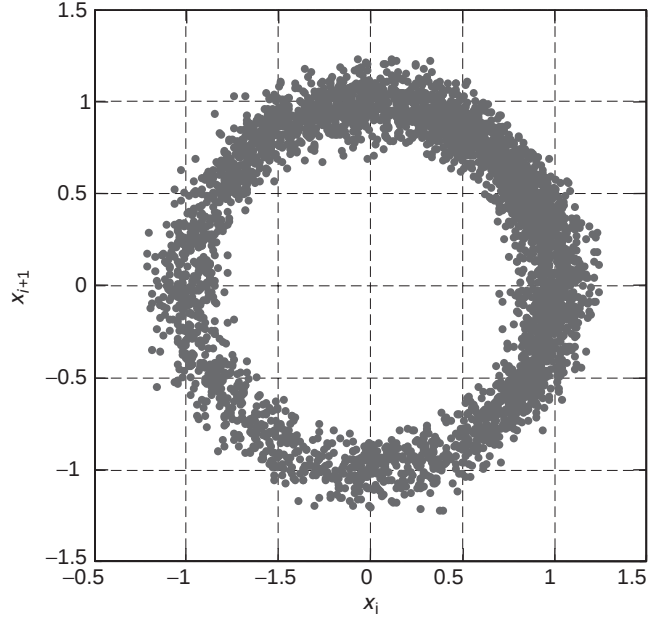


Figure 5. Plot of points against from the generated time series.

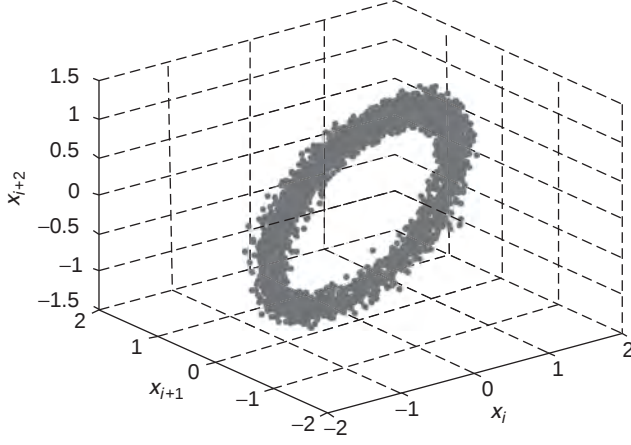


Figure 6. Plot of three consecutive observation points from the time series.

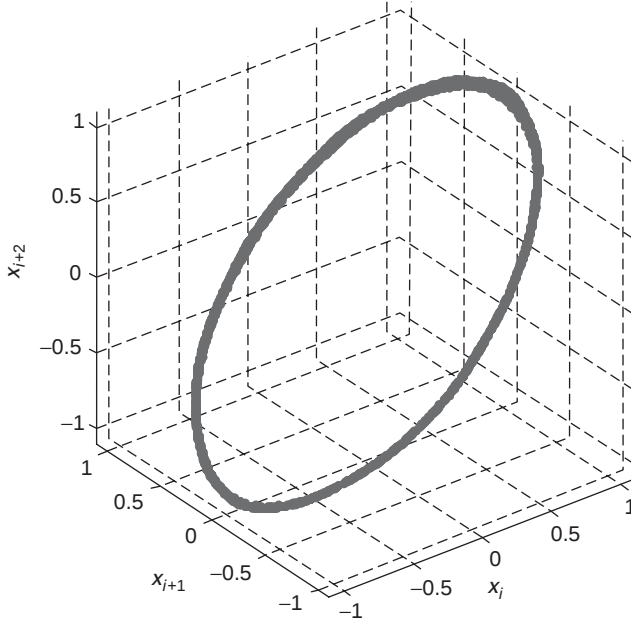


Figure 7. Illustration of the learned model based on the three-tuple data points from the time series.

by x_i , hence three consecutive observations, (x_i, x_{i+1}, x_{i+2}) , must be considered.

The 3D plot of points (x_i, x_{i+1}, x_{i+2}) is displayed in Fig. 6. The embedded underlying rule can be seen vaguely. However, we need to compute for the underlying rule from the data. Note that it is too opportunistic to assume any parametric form to the underlying rule. Hence a manifold-based learning method is favored, because it requires little presumption.

Applying a manifold-based learning method—we utilized the Local linear projection (LLP) method in Huo and Chen [22]—the underlying structure is learned to be nearly a 3D circle, which is plotted in Fig. 7. This is also consistent with the way the time series is generated.

Given the learned model, one can implement a prediction scheme on this time series. We omit the details of forecasting here, because it is evident from the context.

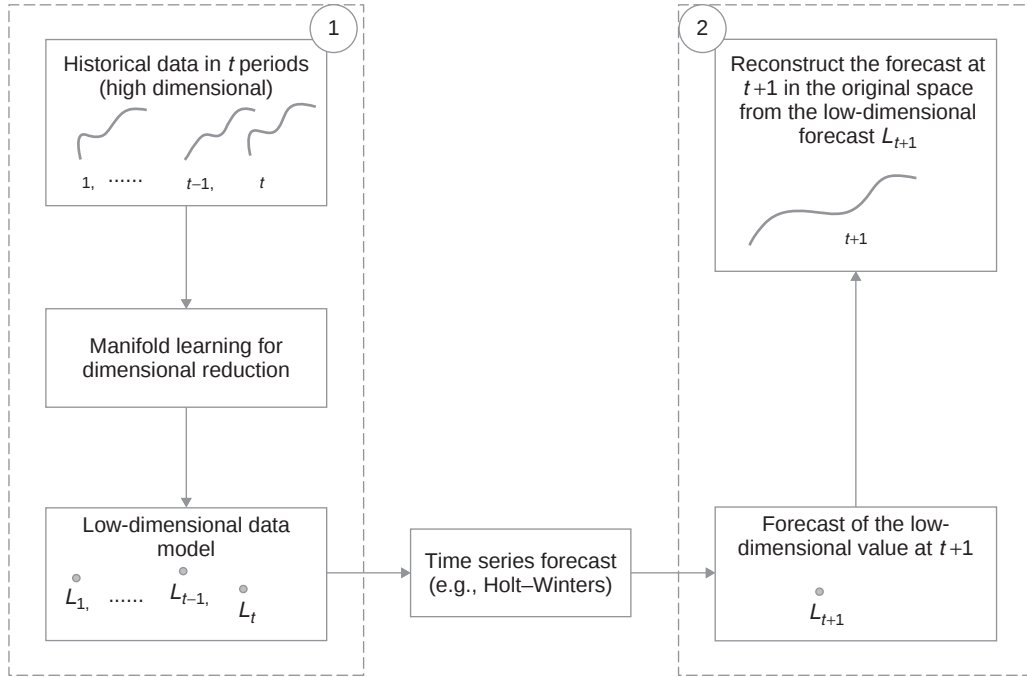


Figure 8. The conceptual flowchart of modeling and forecasting with manifold learning from Ref. 1.

ANOTHER FRAMEWORK

There are not many manifold-based forecasting models in the literature. Recently, Chen *et al.* applied LLE and LLE reconstruction in predicting the electricity price curve [1]. However, their framework is different with the procedure we explained in the introduction and the previous section. We review their idea in the following.

The time series data (electricity price curve) in Ref. 1 is high dimensional. Each data point contains daily or weekly price changes. Modeling the daily or weekly price curve as a whole can provide great advantages for forecasting in a longer horizon from days to weeks. Chen *et al.* [1] use manifold learning methods to analyze the intrinsic structures of the high-dimensional price curves in the low-dimensional space (dimension reduction), and then make the forecasting on the low-dimensional representation. Finally, the low-dimensional prediction is mapped back to the original

high-dimensional space by manifold reconstruction. Figure 8 illustrates the conceptual flowchart of their modeling approach.

It can be seen that the forecasting in the above model is still the classical parametric or nonparametric models such as autoregressive integrated moving average model (ARIMA) or artificial neural networks. The idea here is different from the illustration in Figs 1 and 2. However, the two procedures we have surveyed, dimension reduction and manifold reconstruction are still very important in this framework.

In general, the task of analytically forecasting the dynamics of high-dimensional time series is daunting. The above framework provides a useful strategy for the analysis. The real-world forecasting is difficult also because of the irregular format of the prediction function f [2]. We hope our framework in the introduction can provide new insight into the problem and stimulate a new direction for work in manifold-based forecasting.

CONCLUSION

Manifold-based forecasting gives a promising alternative to existing forecasting approaches. To correctly utilize it, modelers must understand different applicable domains for different algorithms. We did a survey on some widely known manifold-based learning algorithms. We also reviewed their corresponding reconstruction strategies. We hope to draw more attention to this field of research from both methodologists and practitioners.

REFERENCES

1. Chen J, Deng SJ, Huo X. Electricity price curve modeling and forecasting by manifold learning. *IEEE Trans Power Syst* 2008;23(3):877–888.
2. Jolliffe IT. *Principle component analysis*. 2nd ed. Springer Verlag; 2002.
3. Borg I, Groenen P. *Modern multidimensional scaling: theory and applications*. New York: Springer-Verlag; 1997.
4. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 1964;29:1–27.
5. Kruskal JB, Wish M. *Multidimensional scaling*. In: Uslaner EM, editor. *The sage series on quantitative applications in the social sciences*. Beverly Hills: Sage Publications; 1978.
6. Shepard RN. The analysis of proximities: multidimensional scaling with an unknown distance function: I & II. *Psychometrika* 1962;27:125–140, 219–246.
7. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;290:2323–2326.
8. Roweis ST, Saul LK. LLE code page. Available at <http://www.cs.toronto.edu/~roweis/lle/code.html>. 2001.
9. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;290:2319–2323.
10. Tenenbaum JB, de Silva V, Langford JC. ISOMAP webpage; 2000. Available at <http://waldron.stanford.edu/~isomap/>.
11. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Dietterich TG, Becker S, Ghahramani Z, editors. *Volume 14, Advances in neural information processing systems* (NIPS). Cambridge: MIT Press; 2001. pp. 585–591.
12. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 2003;15(6):1373–1396.
13. Donoho DL, Grimes CE. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Arts Sci* 2003;100:5591–5596.
14. Young G, Householder AS. Discussion of a set of points in terms of their mutual distances. *Psychometrika* 1938;3:19–22.
15. Kohonen T. *Self-organizing maps*. 3rd ed. New York: Springer; 1995, 1997, 2001.
16. Bishop CM, Svensen M, Williams CKI. GTM: The generative topographic mapping. *Neural Comput* 1998;10(1):215–234.
17. Brand M. *Charting a manifold*. Volume 15, *Proceedings, neural information processing systems*. Cambridge: Mitsubishi Electric Research Lab: MIT Press; 2003. TR-2003-13 March 2003. Presented at NIPS-15, 2002 Dec. Available at <http://www.merl.com>.
18. Fodor IK. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, LLNL. 2002.
19. Huo X, Ni SX, Smith AK. A survey of manifold-based learning methods. Volume 6, *Recent advances in data mining of enterprise data: algorithms and applications*. Series on computers and operations research. Hackensack: World Scientific Publishing; 2007.
20. Demmel JW. *Applied numerical linear algebra*. Philadelphia: SIAM Press; 1997.
21. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika* 1952;17:401–419.
22. Huo X, Chen J. Local linear projection (LLP). First IEEE workshop on genomic signal processing and statistics (GENSIPS); October; Raleigh (NC). 2002. Available at <http://www.gensips.gatech.edu/proceedings/>.
23. Friedman JH, Bentley JL, Finkel RA. An algorithm for finding best matches in logarithmic expected time. *ACM Trans Math Softw* 1977;3(3):290–226.
24. Bai Z, Demmel J, Dongarra J, *et al.* *Templates for the solution of algebraic eigenvalue problems: a practical guide*. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2000.
25. Dijkstra EW. A note on two problems in connection with graphs. *Numer Math* 1959;1:269–271.

26. Bernstein M, de Silva V, Langford JC, *et al.* Graph approximations to geodesics on embedded manifolds. Technical report. Stanford: Stanford University; 2000.
27. Chung FRK. Volume 92, Spectral graph theory: regional conference series in mathematics Providence. 1997.
28. Saul LK, Roweis ST. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J Mach Learn Res* 2003;4: 119–155.
29. Fukumizu K, Bach FR, Jordan MI. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J Mach Learn Res* 2004;5:73–79.
30. Zhang Z, Zha H. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Sci Comput* 2004;26 (1):313–338.

MODELING UNCERTAINTY IN OPTIMIZATION PROBLEMS

ANDY PHILPOTT

Department of Engineering Science,
University of Auckland, Auckland,
New Zealand

INTRODUCTION

This expository article discusses approaches for modeling optimization problems that involve uncertainty. The emphasis of the article is on motivation and intuition rather than technical completeness. By virtue of its length, the article is inevitably incomplete, and the reader is advised to look at other articles in this encyclopedia to explore the richness of modeling approaches that this article can only hint at.

How to model real decision problems using optimization models depends on the exact form of the problem being faced, and so it is difficult to give a general description of modeling techniques without some context in which to view the techniques. For this reason, we have chosen to illustrate the modeling approaches we discuss by placing them in the context of optimization of some models of electricity generation.

Optimization problems that involve uncertainty fall into several broad classes that admit different modeling approaches. A good starting point is to consider any system with random effects that can be simulated. As an example consider a simple unit commitment problem in which the owner of a thermal generation unit and a wind farm must decide ahead of time whether to start up their thermal unit to meet a known future demand. If the wind blows enough then there will be sufficient power to meet the load, and so the cost incurred by the startup will be unnecessary. On the other hand, if the wind does not blow, and the unit is not started then the owner will incur some penalty of

not supplying the demand. The decision in this case is a simple binary choice, and our model should enable the decision-maker to make the correct choice.

The natural question to ask at this point is “how do we determine the correct choice?” Since there are only two alternatives, one can investigate all the implications of making each choice and compare them. We suppose here, and throughout the rest of the article that the best choice would be clear if there were no uncertainty in the model. In other words, if we had a perfect forecast of wind then our choice would be to start the thermal plant only if the cost of doing so were smaller than the cost of shortage that the startup avoids. (Of course there are settings in which we might consider other criteria such as environmental impact, in which case we seek the solution to a multicriteria optimization problem, but we shall not be concerned with this aspect in this article, and the reader should consult, for example, Figueira *et al.* [1].)

If a perfect forecast of wind is not available, then one is faced with the problem of determining the best choice under uncertainty. A good question to ask at this point is “can we construct or estimate a probability distribution for uncertain wind events?” for example, from historical data and/or meteorological models. If so, then we say that the uncertainty is *probabilistic*. (It is well known that economists like to distinguish between *risk* and *uncertainty*, where risk involves events having probability distributions, and uncertainty is a situation for which we have no previous data or experience so that it is impossible to model statistically, but since we use the word “risk” in a different sense, we call these probabilistic and nonprobabilistic uncertainty.)

Let us first consider the case of probabilistic uncertainty. The distribution of this might take the form of some well-known distribution from a parametric family (such as a lognormal distribution) or it might be empirically determined (e.g., from a Monte Carlo simulation). The existence of such

a distribution means in principle that the probability distribution of total cost of either switching on the unit, or keeping it off can be computed. Armed with these distributions the decision-maker can express a preference for one to determine the best decision. The message here is that optimization under uncertainty with known probability distributions is about choosing between a (possibly infinite) set of probability distributions of total cost. Furthermore, as we discuss in this article, the existence of probability distributions allows us to use probability theory to explore mathematically different ways of making this choice. We shall discuss this more fully in the section titled “Optimization Modeling with Probability.”

How does one handle nonprobabilistic uncertainty? On one hand the decision-maker might be faced with a collection of data points with insufficient a priori knowledge to confidently determine a probability distribution from which these data are drawn. In our unit commitment example, we may have only 10 observations of previous wind outcomes. In this context the modeling choices become less obvious. On one hand, the collected data could be viewed as an empirical discrete distribution. In our example, this would give rise to 10 wind outcomes (or *scenarios*) each with probability 0.1. Even if they were not evenly weighted, at least we would want to investigate the candidate decisions in each of the 10 scenarios to see what the cost would have been in these scenarios. The decision-maker can then choose the vector of 10 outcomes that is preferred. How to do this soundly is less obvious when there is no probability distribution, but there is a rich literature on making these choices according to different criteria. We shall return to this in the section titled “Optimization modeling without probability.”

The final form of uncertainty the decision-maker might be confronting is that for which there is no probability distribution, or reliable data, or even a coherent understanding. Uncertain events might occur that are not even foreseen—the so-called *black swans* (the probability of a swan being white was 1 until black swans were discovered in Western

Australia) that have been popularized in the recent book by Nicolas Taleb [2] that warns of the dangers of quantitative modeling in finance. However, faced with such comprehensive uncertainty, there is little one can do from an operations research perspective. This does not mean that one should be entirely solipsistic—stress testing candidate decisions under some outlandish scenarios can yield valuable insights even those might be supposed to occur with miniscule probability. Nevertheless, the black swan will, by definition, always elude the analyst.

The remainder of this article is laid out as follows. The next section gives an overview of probabilistic modeling in optimization problems. The section titled “Optimization Modeling without Probability” deals with the treatment of uncertainty when probability distributions are not available. The final section “Bibliographical Notes” provides some brief bibliographical references.

OPTIMIZATION MODELING WITH PROBABILITY

Probabilistic optimization models take advantage of the fact that probability distributions governing the data are known or can be estimated. In this section, we will discuss how to model these with reference to problem examples drawn from electricity generation.

To begin the discussion we recall the unit commitment problem discussed in the section titled “Introduction,” and formulate this mathematically. For simplicity, assume that there is a single hour of demand d MW to be met. Suppose, the thermal unit has a capacity g MW, a startup cost of $\$h$, a running cost of $\$/\text{MWh}$, and the supply of wind power is the random variable $S(\omega)$, where $\omega \in \Omega$ denotes a random wind event for the hour in question. Finally, suppose that any shortages of electricity cost $\$/\text{MWh}$. Let

$$z = \begin{cases} 1, & \text{if the unit is started} \\ 0, & \text{if the unit is not started} \end{cases},$$

and suppose $X(\omega)$ is the (random) power generated by the unit. Observe that in any

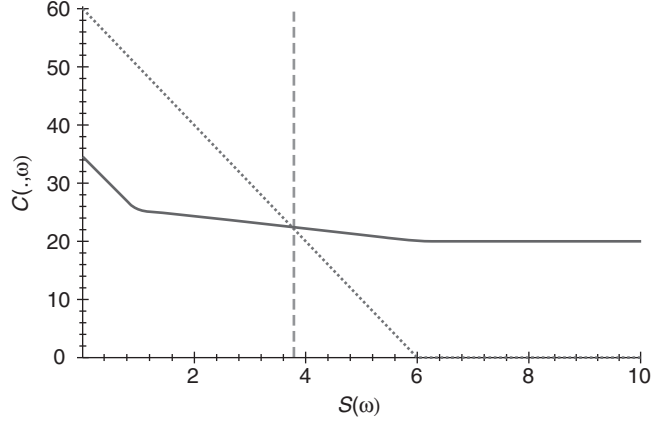


Figure 1. Plot of $C(0, \omega)$ (dotted) and $C(1, \omega)$ (solid) for unit commitment example.

outcome ω , the optimal choice of $X(\omega)$ is

$$X(\omega) = z \min \{g, \max(0, d - S(\omega))\}.$$

Then the cost of decision z in outcome ω is

$$C(z, \omega) = hz + cz \min \{g, \max(0, d - S(\omega))\} + s \max(0, d - S(\omega) - gz).$$

As discussed in the section titled “Introduction,” a decision-maker can compare $C(0, \omega)$ and $C(1, \omega)$ for every realization of ω and choose which decision they prefer. As an illustration, these functions are plotted in Fig. 1 for $h = 20, c = 1, g = 5, d = 6$, and $s = 10$ over the range $S(\omega)$ from 0 to 10.

They show that not committing the unit is a costly decision when wind generation is low, but is inexpensive when wind becomes plentiful. Conversely, committing the unit incurs an expense but saves on shortage cost, except when $S(\omega)$ is less than 1. From these plots alone it is not clear which is the better decision.

If a probability distribution is known for wind outcomes ω , then it is possible to analyze the choices in more detail. For example, if there is zero probability of observing wind generation less than $\frac{34}{9}$ MWh, then it is less costly to leave the unit off, as this policy is less costly with probability 1. This approach can be explored further by attempting to order the probability distributions that emerge from different decisions. The theory of *stochastic dominance* gives

a methodology for doing this. A random variable C_1 is said to *stochastically dominate* random variable C_2 to *first-order* if for every x , $\Pr(C_1 \leq x) \leq \Pr(C_2 \leq x)$. When minimizing cost we might seek a policy that gives rise to a (random) objective function that is dominated by those of every alternative policy.

In the unit commitment problem suppose that $S(\omega)$ has a uniform distribution on $[0, 10]$, then $C(0, \omega)$ has distribution function

$$\Pr(C(0, \omega) \leq x) = \begin{cases} 0, & x \leq 0 \\ 0.4 + 0.01x, & 0 < x \leq 60 \\ 1, & 60 < x \end{cases}$$

and $C(1, \omega)$ has distribution function

$$\Pr(C(1, \omega) \leq x) = \begin{cases} 0, & x \leq 20 \\ -1.6 + 0.1x, & 20 < x \leq 25 \\ 0.65 + 0.01x, & 25 < x \leq 35 \\ 1, & 35 < x. \end{cases}$$

as shown in Fig. 2.

It is easy to see that neither $C(0, \omega)$ nor $C(1, \omega)$ is first-order stochastically dominant. This example illustrates that this is a very strong form of optimality, and often precludes finding a suitable solution. (Second- and higher-order stochastic dominance relations can be defined that admit more possible solutions; we shall not discuss these here but refer the reader to Shapiro *et al.* [3].)

As an alternative to stochastic dominance, we might constrain the policies that we are prepared to accept to those that have a high

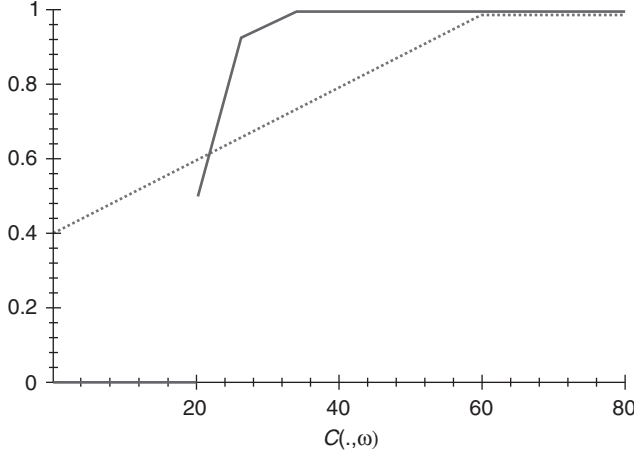


Figure 2. Cumulative distribution functions of $C(0, \omega)$ (dotted) and $C(1, \omega)$ (solid).

cost only with low probability. This is an example of a *chance constraint*, which takes the form

$$\Pr(C(z, \omega) \leq \tau) \geq 1 - \alpha.$$

For example, if the probability distribution of wind outcomes $S(\omega)$ is uniform on $[0, 10]$, then we could constrain z to satisfy

$$\Pr(C(z, \omega) \leq 30) \geq 0.9.$$

Since $\Pr(C(0, \omega) > 30) = 0.3$, and $\Pr(C(1, \omega) > 30) = 0.05$, this constraint is satisfied by $z = 1$, but not $z = 0$.

The choice of a policy becomes more obvious in settings in which decisions are made repeatedly in essentially the same circumstances, and the objective is to come up with a decision that will perform well on average. In our unit commitment example this would be the case if the commitment decision was being repeated every day or every few hours. In this setting the goal is to find some policy that minimizes the expectation $\mathbb{E}[C(z, \omega)]$ of $C(z, \omega)$. This objective is justified by the law of large numbers for infinitely repeated trials which implies that a policy that minimizes expected costs will result in the minimum long-run average cost. In the example with a uniform distribution of wind outcomes, $\mathbb{E}[C(1, \omega)] = 22.25$ and $\mathbb{E}[C(0, \omega)] = 18$, and so the minimum expected cost policy is to keep the unit off.

The mathematical expectation operator is an example of a *risk measure*, a function $\rho(Z)$ from a space of random variables to $\mathbb{R}$. It assigns a score to any random variable Z that can then be ranked against the scores of others. In the example $Z_0 = C(0, \omega)$, and $Z_1 = C(1, \omega)$, and $\rho(Z_0) = 18$ is preferable to $\rho(Z_1) = 22.25$. When modeling optimization problems it is often crucial that the risk measure be convex and monotone, so that efficient numerical procedures can be applied to solve the problem. More generally, a risk measure is called *coherent* if it has the following properties.

1. It is convex, that is, if Z_1 and Z_2 are two random variables and $t \in [0, 1]$, then

$$\begin{aligned} \rho(tZ_1 + (1-t)Z_2) \\ \leq t\rho(Z_1) + (1-t)\rho(Z_2). \end{aligned}$$

2. It is monotone, that is, if Z_1 and Z_2 are two random variables such that with probability 1 $Z_2 \geq Z_1$, then $\rho(Z_2) \geq \rho(Z_1)$.
3. It is positively homogeneous, that is, if Z is a random variable and $t > 0$, then $\rho(tZ) = t\rho(Z)$.
4. It has the (translation equivariance) property: $\rho(Z + a) = \rho(Z) + a$ for any $a \in \mathbb{R}$.

Properties (1) and (2) ensure that if $C(\cdot, \omega)$ is convex for almost every ω , then the function $f(z) := \rho(C(z, \omega))$ is also convex.

As mentioned above, the expected cost is an appropriate risk measure for an optimization problem that is solved repeatedly as variations of cost around the expected cost will balance out in the long run. For a solution that is to be used only once, we might seek to control this random variation, especially if it produces high costs. One way of doing this is by adding a chance constraint of the form

$$\Pr(C(z, \omega) \leq \tau) \geq 1 - \alpha. \quad (1)$$

Unfortunately this is not convex in general, and so does not correspond to a coherent risk measure.

The Constraint (1) can be approximated by

$$\text{CVaR}_{1-\alpha}(C(z, \omega)) \leq \tau,$$

where CVaR is called the *conditional value at risk*, defined to be

$$\text{CVaR}_{1-\alpha}(Z) := \inf_{t \in \mathbb{R}} \left\{ t + \alpha^{-1} \mathbb{E} [\max\{0, Z - t\}] \right\}.$$

It can be shown that the coherent measure CVaR is the tightest safe convex approximation to Equation (1). In the example, the CVaR of each unit commitment decision at level 0.95 is respectively

$$\begin{aligned} & \text{CVaR}_{0.95}(C(0, \omega)) \\ &= \inf_{t \in \mathbb{R}} \left\{ t + 20 \mathbb{E} [\max\{0, C(0, \omega) - t\}] \right\} \\ &= \inf_{t \geq 0} \left\{ -11t + 360 + \frac{1}{10}t^2 \right\} = \frac{115}{2}, \\ & \text{CVaR}_{0.95}(C(1, \omega)) \\ &= \inf_{t \in \mathbb{R}} \left\{ t + 20 \mathbb{E} [\max\{0, C(1, \omega) - t\}] \right\} \\ &= \inf_{t \geq 25} \left\{ -6t + \frac{245}{2} + \frac{1}{10}t^2 \right\} = \frac{65}{2}, \end{aligned}$$

so with this risk measure there is less risk in running the unit. Observe in this example that the underlying (discrete) optimization problem is not convex, and so the use of a coherent risk measure here does not retain

convexity as it would if applied to a linear program.

Finally, we mention the classical approach to modeling risk that can be traced back to Bernoulli [4] that uses utility and disutility functions as risk measures. (Here we speak of disutility when minimizing and utility when maximizing). In a minimization problem like our unit commitment problem, the risk measure $\rho(Z)$ is defined to be $\mathbb{E}[u(Z)]$ for some disutility function u . Typically, utility functions are chosen to be concave and disutility functions are convex. This represents aversion to risk on the part of decision-makers. To illustrate this, consider the unit commitment example with the disutility function $u(z) = z^2$. Then

$$\mathbb{E}[u(C(0, \omega))] = 720, \quad \mathbb{E}[u(C(1, \omega))] = 505,$$

so with this risk measure the decision-maker would prefer to run the unit.

The unit commitment example we have been studying involves choosing from two possible decisions, to run the thermal unit or not to run it. Implicit in each of these choices is an operating level $X(\omega)$ for the plant under each wind realization. Because of its cost structure, our model can assume this to be the minimum level required to meet demand minus wind generation if this does not exceed the capacity of the unit.

In many circumstances this implicit optimization is not possible, and for each unit commitment decision we must solve a dispatch problem to determine the optimal level at which to run each plant. This leads to the widely studied *two-stage stochastic program with recourse*, where the first stage of the problem determines the unit commitment and the second stage, or recourse problem, determines the optimal generation level in each scenario given this unit commitment decision.

In general, the classical two-stage linear stochastic programming problem can be formulated as

$$\min_{x \in X} \{c^T x + \mathbb{E}[Q(x, \xi)]\}, \quad (2)$$

where $Q(x, \xi)$ is the optimal value of the second-stage problem

$$\min_y q^\top y \text{ subject to } Tx + Wy \leq h. \quad (3)$$

Here, x is the first-stage decision vector that must lie in some set X , $y \in \mathbb{R}^m$ is the second-stage decision vector and $\xi = (q, T, W, h)$ contains the data of the second-stage problem. In this formulation, at the first-stage we have to make a “here-and-now” decision x before the realization of the uncertain data ξ , viewed as a random vector, is known. At the second stage, after a realization of ξ becomes available, we optimize our behavior by solving an appropriate optimization problem.

An important practical question is whether the formulated problem can be solved numerically. In that respect the standard approach is to assume that random vector ξ has a finite number of possible realizations or scenarios, say $\xi_1, \dots, \xi_K$, with respective (positive) probabilities $p_1, \dots, p_K$. Then the expectation can be written as the summation

$$\mathbb{E}[Q(x, \xi)] = \sum_{k=1}^K p_k Q(x, \xi_k), \quad (4)$$

and, moreover, the two-stage problem (2) and (3) can be formulated as one optimization problem

$$\begin{aligned} \min_{x, y_1, \dots, y_K} \quad & c^\top x + \sum_{k=1}^K q_k^\top y_k \\ \text{s.t.} \quad & x \in X, T_k x + W_k y_k \leq h_k, k = 1, \dots, K. \end{aligned} \quad (5)$$

In the above formulation (5), we make one copy y_k , of the second-stage decision vector, for every scenario $\xi_k = (q_k, T_k, W_k, h_k)$. By solving Equation (5), we obtain an optimal solution $\bar{x}$ of the first-stage problem and optimal solutions $\bar{y}_k$ of the second-stage problem for each scenario ξ_k , $k = 1, \dots, K$. Given $\bar{x}$, each $\bar{y}_k$ gives an optimal second-stage decision corresponding to a realization $\xi = \xi_k$ of the respective scenario.

In practice this recourse structure exists at many stages, and the two-stage model is an

approximation of a more complicated process in which observations of uncertain parameters occur over time, and decisions are made that adapt to these observations. In mathematical terms, the random variables form a *stochastic process*, and the decisions to be made must be measurable with respect to the filtration defined by this process.

As an example of a problem with such a structure consider the problem of controlling a reservoir with maximum volume a and initial storage r , that is used for hydroelectricity generation in conjunction with a thermal unit to minimize the expected cost of meeting demand over some planning horizon $t = 1, 2, \dots, T$. For simplicity, we assume that the thermal unit has zero startup cost but capacity g MW, and a running cost of $\$c/\text{MWh}$. We also assume that (known) electricity demand is $d(t)$, $t = 1, 2, \dots, T$, and the supply of wind power is zero. As before suppose that any shortages of electricity cost $\$s/\text{MWh}$.

The new feature of our model is a random inflow sequence $I(t)$, $t = 1, 2, \dots, T$, that may be stored for electricity generation through a turbine with capacity b MW. This gives the following model:

MSP:

$$\begin{aligned} \min \quad & \mathbb{E} \left[\sum_{t=1}^T cX(t) + sY(t) \right] \\ \text{s.t.} \quad & X(t) + U(t) + Y(t) = d, \\ & t = 1, 2, \dots, T, \\ & R(t) = R(t-1) - \beta U(t) - V(t) + I(t), \\ & t = 1, 2, \dots, T, \\ & 0 \leq R(t) \leq a, \quad R(0) = r, \\ & t = 1, 2, \dots, T, \\ & 0 \leq X(t) \leq g, \quad t = 1, 2, \dots, T, \\ & 0 \leq U(t) \leq b, \quad t = 1, 2, \dots, T, \\ & V(t) \geq 0, \quad t = 1, 2, \dots, T. \end{aligned}$$

Here, $X(t)$ is the thermal generation, $U(t)$ is hydro generation, $Y(t)$ is unserved demand, $V(t)$ is spill, and $R(t)$ is reservoir storage at the end of period t , all of which are random variables. The parameter $\beta \text{ m}^3/\text{MWh}$

converts the electric power generated in a period into the corresponding volume of water released.

This general formulation MSP can be interpreted in many ways depending on how the random inflows $I(t)$ are modeled. A possible model could consider all T inflows to be revealed to the decision-maker at once, who plans a (possibly different) least cost electricity generation schedule for each realization of inflows and then computes its cost. This *wait-and-see* version of the problem is of course unrealistically optimistic, and will give policies that will be unable to be implemented unless perfect foresight is available. In contrast, a *here-and-now* version of the problem might seek a fixed plan of thermal generation decisions $x(t)$ that must be feasible for all realizations of $I(t)$ and minimizes $\sum_{t=1}^T cx(t) + \mathbb{E}[\sum_{t=1}^T sY(t)]$. Because of its inflexibility, this model is likely to result in expensive policies.

Between these extremes there are many models that incorporate flexibility in the policies. The modeling choice then becomes a trade-off between realism and computational tractability. If $I(t)$ follows a general stochastic process then MSP becomes a multistage stochastic linear program. Traditional stochastic programming methodologies for solving these models use *scenario-trees*, each of which approximates the stochastic process by one defined on a finite set of outcomes in each stage leading to a large-scale mathematical programming problem. If either the number of stages or dimension of the random variable becomes large, then scenario-tree approximations can be quite poor, and hence this approach has serious limitations for many realistic-sized problems.

On the other hand, when $I(t)$ is stage-wise independent or is a Markov process, MSP can be treated as a Markov decision problem that can be attacked using methods of dynamic programming. For the reservoir problem, we define $Q_{t-1}(r)$ to be the minimum expected thermal and shortage cost incurred over the period $t, t+1, \dots, T$, if the reservoir storage at the end of period $t-1$ is r , and an optimal policy is followed. In the case where $I(t)$ is stage-wise independent this gives the

recursion:

$$Q_{t-1}(r) = \text{DP}(t):$$

$$\begin{aligned} \min \quad & cx(t) + sy(t) + \mathbb{E}[Q_t(R(t))] \\ \text{s.t.} \quad & x(t) + u(t) + y(t) = d, \\ & R(t) = r - \beta u(t) - V(t) + I(t), \\ & 0 \leq R(t) \leq a, \\ & 0 \leq x(t) \leq g, \\ & 0 \leq u(t) \leq b, \\ & V(t) \geq 0. \end{aligned}$$

Solving such a recursion in general is possible only if the state variable r has low dimension.

We conclude here by discussing the advantages of modeling a multistage decision problem as a dynamic program rather than a multistage stochastic program. As mentioned above scenario-tree models become computationally intractable as the number of stages and random outcomes grows. In contrast, dynamic programming can admit many stages, but suffers from a curse of dimensionality in the state variables. Hence, there are computational advantages when the state space is of low dimension.

A second advantage of dynamic programming is that it gives a state-feedback policy, rather than an open-loop set of actions adapted to the filtration of the stochastic process. This makes simulation much easier. To simulate a multistage stochastic programming solution in a rolling horizon implementation requires a model like MSP to be generated and solved as each stage unfolds, a considerable computational effort considering that $T-1$ such solves are required for a single realization of random outcomes over T stages. In contrast, dynamic programming, given $Q_t(\cdot)$ would solve $\text{DP}(t)$ at each stage in such a simulation, a much less intensive effort.

OPTIMIZATION MODELING WITHOUT PROBABILITY

In this section, we give a brief discussion of optimization under uncertainty without

using probability distributions. The principle in this setting is to make decisions that are somehow *robust* in the face of variations in data. Different modeling approaches are available here, and we give a brief overview, returning to the unit commitment problem discussed in the section titled “Optimization Modeling with Probability” as an illustration. Recall the plot of the cost outcomes of the alternative decisions, shown in Fig. 1.

The classical *minimax* criterion used in modeling without probability seeks to minimize the worst case. In the unit commitment example, the worst-case cost is 60 for $z = 0$ and 35 for $z = 1$, hence $z = 1$ is the preferred solution under a minimax objective. Observe that this would still be the minimax solution even if $C(0, \omega)$ was 36 when $S(\omega) = 0$, and zero for $S(\omega) > 0$, which would make the choice of $z = 1$ an overly cautious solution.

An alternative criterion is to minimize maximum regret. Here *regret* is defined to be the difference in cost between the decision taken and the optimal decision with hindsight. The maximum regret for the decision $z = 1$ in the example is 20, which occurs when $S(\omega) \geq 6$. The maximum regret for the decision $z = 0$ in the example is 25, which occurs when $S(\omega) \leq 1$. Thus we should choose $z = 1$. Observe that if $C(0, \omega)$ was 36 when $S(\omega) = 0$, and zero for $S(\omega) > 0$ then the maximum regret for the decision $z = 0$ in the example is 1, which makes it preferred to $z = 1$. This instance does however illustrate a paradox with this approach. If $C(0, \omega)$ was 56 when $S(\omega) = 0$, and zero for $S(\omega) > 0$ then the maximum regret for the decision $z = 0$ is 21, which makes $z = 1$ preferred again, even though it is a worse decision in every outcome except when $S(\omega) = 0$. This highlights the advantages of a probabilistic model that can quantify the trade-offs between different outcomes.

Since the minimax criterion is widely recognized as being too conservative, there have been a number of efforts to make it less. *Robust optimization* applies the minimax criterion to a subset of possible outcomes called an *uncertainty set*. Formally robust optimization seeks the solution to the semi-infinite

program

$$\begin{aligned} \text{RP: } \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x, u) \leq 0, \quad u \in U, \end{aligned}$$

where the functional dependence of g on u and uncertainty set U is specially constructed so that this problem can be solved using an efficient algorithm.

In the unit-commitment example, one might choose

$$U_1 = \{\omega \mid 2 \leq S(\omega) \leq 8\},$$

and solve, for example,

$$\begin{aligned} \text{RP: } \min_{x,z} \quad & x \\ \text{s.t.} \quad & C(z, \omega) \leq x, \quad \omega \in U_1. \end{aligned}$$

This has solution $z = 1$. On the other hand choosing

$$U_2 = \{\omega \mid 4 \leq S(\omega) \leq 10\},$$

gives optimal solution $z = 0$.

There has been considerable discussion in the literature on the relative merits of robust optimization and chance-constrained programming. The latter approach as part of the optimization seeks the best set U of outcomes with probability $1 - \alpha$ for which RP gives the least cost, and hence from a modeling approach it is richer. Thus, if we assume a uniform distribution in the above example, the chance-constrained problem

$$\begin{aligned} \text{CP: } \min_{x,z} \quad & x \\ \text{s.t.} \quad & \Pr(C(z, \omega) \leq x) \geq 0.6, \end{aligned}$$

has optimal solution $z = 0$, corresponding to the optimal choice of uncertainty set U_2 .

On the other hand, the optimal set U may be poorly behaved (making even RP for this U a difficult problem to solve) and so a chance-constrained model might prove impossible to solve for instances with lots of variables. In this case, robust optimization provides an alternative modeling approach.

BIBLIOGRAPHICAL NOTES

The concept of two-stage linear stochastic programming with recourse was introduced in Beale [5] and Dantzig [6]. For the theory and applications of two-stage and multistage stochastic programming, see the monographs [3,7–10]. Chance-constrained problems were introduced in Charnes *et al.* [11]. For a thorough discussion and development of that concept we may refer to Prékopa [8]. An introductory treatment is available in the on-line tutorial by Henrion [12].

The first mathematical discussion of risk aversion (using utility functions) is often attributed to Bernoulli [4]. The concept of utility was formally defined and expounded by von Neumann and Morgenstern [13]. Coherent risk measures were first introduced by Artzner *et al.* [14]. The approach of using CVaR for approximating chance constraints is due to Rockafellar and Uryasev [15].

Robust optimization models using ellipsoidal uncertainty sets were introduced by Ben Tal and Nemirovski [16], who showed how to formulate these as second-order cone problems. An alternative approach is due to Bertsimas and Sim [17].

Acknowledgments

Much of this article is drawn from the material in the on-line tutorial article [18]. The author would like to acknowledge the contributions of Alexander Shapiro to this work.

REFERENCES

1. Figueira JR, Greco S, Ehrgott M. Multiple criteria decision analysis: state of the art surveys. Volume 78, International series in operations research & management science. New York: Springer; 2005.
2. Taleb N. The Black Swan: the impact of the highly improbable. New York: Random House; 2007.
3. Shapiro A, Dentcheva D, Ruszczyński A. Lectures on stochastic programming: modeling and theory. Philadelphia: SIAM; 2009.
4. Bernoulli D. Exposition of a new theory on the measurement of risk, 1738 (Translated by L. Sommer). *Econometrica* 1954;22(1):22–36.
5. Beale EML. On minimizing a convex function subject to linear inequalities. *J R Stat Soc Ser B* 1955;17:173–184.
6. Dantzig GB. Linear programming under uncertainty. *Manage Sci* 1955;1:197–206.
7. Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
8. Prékopa A. Stochastic programming. Dordrecht, Boston: Kluwer Academic Publisher; 1995.
9. Ruszczyński A, Shapiro A, editors. Stochastic programming. Volume 10, Handbook in OR & MS. Amsterdam: North-Holland Publishing Company; 2003.
10. Ziemba WT, Wallace SW. Applications of stochastic programming. Philadelphia: SIAM; 2006.
11. Charnes A, Cooper WW, Symonds GH. Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Manage Sci* 1958;4:235–263.
12. Henrion R. Introduction to chance-constrained programming. <http://www.stoprog.org>. Accessed in 2010.
13. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ) Princeton University Press; 1944.
14. Artzner P, Delbaen F, Eber J-M, *et al.* Coherent measures of risk. *Math Finance* 1999;9: 203–228.
15. Rockafellar RT, Uryasev SP. Optimization of conditional value-at-risk. *J Risk* 2000;2: 21–41.
16. Ben Tal A, Nemirovski A. Robust convex optimization. *Math Oper Res* 1998;23(4):769–805.
17. Bertsimas D, Sim M. The price of robustness. *Oper Res* 2004;52(1):35–53.
18. Shapiro A, Philpott AB. A tutorial on stochastic programming. <http://www.stoprog.org>. Accessed in 2010.

MODELS AND BASIC PROPERTIES

RENÉ HENRION
Weierstrass Institute,
Berlin, Germany

Probabilistically Constrained Programming (also known as *Chance-Constrained Programming*) belongs to the major approaches for dealing with random parameters in optimization problems. Typical areas of application are engineering and finance, where uncertainties like product demand, meteorological or demographic conditions, currency exchange rates and so on enter the inequalities describing the proper working of a system under consideration. The main difficulty of such models is due to the fact that decisions have to be taken prior to the observation of the random variables (“here and now”). A naive way to circumvent this difficulty would consist in replacing the random variables by their expectations estimated from historical data. However, then the resulting decisions would severely lack robustness. This could result in frequent violations of constraints when randomness becomes manifest. On the other hand, in a here-and-now situation, one can hardly find a decision which would definitely exclude later constraint violation. And if so, such decisions typically turn out to involve unreasonably high costs, just in order to hedge against extremely rare events.

Sometimes, such constraint violation can be balanced afterward by some compensating decisions taken in a second stage. For instance, taking recourse to pumped storage plants or buying energy on the liberalized market is an option for power-generating companies that are faced with unforeseen peaks of electrical load. As long as the costs of compensating decisions are known, these may be considered as a penalization for constraint violation. This idea leads to the important class of *two-stage* or *multistage stochastic programs* (see the sections titled “Two-Stage Stochastic Programming” and

“Multistage Stochastic Programming” in this encyclopedia).

In many applications, however, compensations simply do not exist (e.g., for safety-relevant restrictions in technical equipment) or cannot be modeled by costs in a sufficiently precise way. In such circumstances, one would rather insist on decisions guaranteeing feasibility “as much as possible”. This loose term refers once more to the fact that constraint violation can almost never be avoided because of unexpected extreme events. On the other hand, when knowing or approximating the distribution of the random parameter, it makes sense to call decisions feasible (in a stochastic meaning) whenever they are feasible with high probability, that is, only a low percentage of realizations of the random parameter leads to constraint violation under this fixed decision. A generic way to express such a *probabilistic* or *chance constraint* as an inequality is

$$P(h(x, \xi) \geq 0) \geq p. \quad (1)$$

Here, x and ξ are decision and random vectors respectively, and “ $h(x, \xi) \geq 0$ ” refers to a finite system of random inequalities and P is a probability measure. The value $p \in [0, 1]$ is called the probability level, and it is chosen by the decision maker in order to model the safety requirements. Sometimes, the probability level is strictly fixed from the very beginning (e.g., $p = 0.95, 0.99$ and so on). In other situations, the decision maker may have only a vague idea of a properly chosen level. Of course, he or she is aware that higher values of p lead to fewer feasible decisions x in Equation (1), and hence to optimal solutions at higher costs. Fortunately, it turns out that usually p can be increased over quite a wide range without affecting greatly the optimal value of some problem, until it closely approaches 1 and a strong increase in costs becomes evident. In this way, models with chance constraints can also give a hint for a good compromise between costs and safety.

Among the numerous applications of chance-constrained programming, one may find areas such as water resource management, energy production, medical and military science, circuit manufacturing, chemical engineering, telecommunications, reliability theory, and finance (for an overview see Ref. 1). For a comprehensive introduction into theory, models, and algorithms of probabilistically constrained programming we refer to Prékopa [2].

MODELS

Formally, the probabilistic constraint of Equation (1) can be understood as a single inequality

$$\alpha(x) \geq p,$$

where

$$\alpha(x) := P(h_j(x, \xi) \geq 0 \quad (j = 1, \dots, m)) \quad (2)$$

as it appears in optimization problems. However, it is by no means obvious to see how algorithmically important properties of α follow from those of the random constraint mapping h and of the distribution of ξ . The complicated interplay between these two ingredients gives rise to a variety of models leading to the application of different mathematical tools.

Joint and Individual Probabilistic Constraints

The probabilistic constraint as presented in Equations (1) or (2) is also called the *joint type* because it evaluates the probability of a whole random inequality system to hold true. Another possibility of treating the random inequalities would consist in transforming each one of them individually into a probabilistic constraint:

$$\alpha_j(x) := P(h_j(x, \xi) \geq 0) \geq p_j \quad (j = 1, \dots, m). \quad (3)$$

Here, it is possible to choose different probability levels p_j for each constraint. At first glance, the individual model looks more complicated than the joint one because the number of probabilistic constraints has increased now from 1 to m . It turns out, however, that in

certain situations one may benefit a lot from the individual character of the constraints in that they may avoid the consideration of multivariate random distributions as required for joint probabilistic constraints (see below). On the other hand, it is not this potential algorithmic simplification which should decide on the choice of joint or individual constraints but rather the right modeling point of view in a concrete application. As an illustration, one might think of a problem where the random vector represents a certain discrete stochastic process (e.g., some cash flow). With the individual model, a decision would be declared feasible whenever it guarantees, that at each individual time step some level has to be satisfied with a certain probability. In general, however, one is interested in satisfying this level with a given probability over the whole time horizon, which is a much more restrictive requirement. One can state the following relation between both types of constraint sets (where all probability levels are assumed to be equal in the individual model):

$$\overline{M}_p \supseteq M_p \supseteq \overline{M}_{1-(1-p)/m},$$

where

$$M_p := \{x | \mathbb{P}(h_j(x, \xi) \geq 0 \quad (j = 1, \dots, m)) \geq p\}$$

$$\overline{M}_p := \{x | \mathbb{P}(h_j(x, \xi) \geq 0) \geq p \quad (j = 1, \dots, m)\}.$$

In other words, each joint probabilistic constraint can be approximated from below and above by the corresponding set of individual probabilistic constraints upon adjusting the probability levels appropriately. This observation could be exploited in order to approximate programs with the more complicated joint constraints by solutions to programs with the simpler individual constraints. However, the gap between the lower and upper bound of the resulting optimal values can be large.

Random Right-Hand Side

Many practical applications (such as demand satisfaction) lead to probabilistic constraints with random right-hand side, hence $h(x, \xi) = g(x) - \xi$. The decoupling of decision

and random vector then allows rewriting of Equation (1) as

$$F_{\xi}(g(x)) \geq p, \quad (4)$$

where F_{ξ} refers to the (multivariate) distribution function of the random vector ξ . In this way, the generally nonexplicit mapping α in Equation (2) now gets a concrete shape as the composition $F_{\xi} \circ g$ of a distribution function and a fully explicit mapping of g . Though the computation of multivariate distribution function remains a challenging task, this setting offers far more efficient treatment than the general case. In particular, nondegenerate multivariate normal distributions are well investigated from the theoretical and algorithmic viewpoint [2–4]. For the computation of other multivariate distribution functions, such as gamma, Dirichlet, or exponential, we direct the reader to Refs 5–7.

The positive effect of the individual model (Eq. 3) for probabilistic constraints is revealed in combination with random right-hand sides. Indeed, one has the following equivalences for all $j = 1, \dots, m$:

$$\begin{aligned} \alpha_j(x) \geq p_j &\iff P(g_j(x) \geq \xi_j) \geq p_j \\ &\iff F_{\xi_j}(g_j(x)) \geq p_j \iff g_j(x) \geq q_{p_j}^{(j)}. \end{aligned} \quad (5)$$

Here, $q_{p_j}^{(j)}$ are the p_j -quantiles of the one-dimensional distribution functions F_{ξ_j} . Since quantiles of one-dimensional distribution functions are easily determined numerically, the individual probabilistic constraints on the left-hand side boil down to the right-hand side inequality system which is exactly of the same type as the original random inequality system $g_j(x) \geq \xi_j$. For instance, if g was a linear mapping, then the model of individual probabilistic constraints leads to linear inequalities again, which are easily incorporated into a linear programming problem.

Another situation in which advantage can be taken from a model with random right-hand side even for the more demanding case of joint probabilistic constraints as in Equation (2) is the following: if the random vector ξ has independent components, then

the calculation of α reduces to a product of one dimensional distribution:

$$\alpha(x) = F_{\xi_1}(g_1(x)) \cdots F_{\xi_m}(g_m(x)).$$

Although the constraint $\alpha(x) \geq p$ cannot be further simplified to an explicit constraint involving just the g_j (as it was the case for individual chance constraints), one may benefit again from the fact that one-dimensional distributions are easy to calculate.

Linear Probabilistic Constraints

As far as the structure of the mapping h in Equation (1) is concerned, the most typical case in practice occurs when h is affine linear in ξ in which case one speaks of linear chance constraints. There are two main situations to be distinguished then. The first and most important, h may be affine linear in the couple of variables (x, ξ) implying the probabilistic constraint (Eq. 1) takes the form

$$P(Ax + B\xi \geq c) \geq p, \quad (6)$$

where A, B, c are matrices/vectors of appropriate sizes. Typical applications of this model are reservoirs (water/money/energy etc.) with a discrete random inflow process ξ , the accumulation of which over time is described by means of B . A could similarly describe the accumulation of a discrete controlled outflow process but more generally A and B could also reflect network topologies in problems involving several reservoirs. Evidently, unless B is the identity, Equation (6) is not of the random right-hand-side type because ξ undergoes a linear transformation. Of course, one could formally put $\eta := B\xi$ such that the transformed random vector would yield the desired type. However, this is meaningful only if one knows how to also transform the associated distribution parameters. This is indeed possible when ξ obeys a multivariate normal or, more generally, an elliptically symmetric distribution. Therefore, for these distributions, Equation (6) can be considered to have random right-hand side. In particular, when passing from the joint model (Eq. 6) to the individual one in the sense of Equation (3),

then, as mentioned above, one arrives at a simple, explicit system of linear inequalities

$$\langle a_j, x \rangle \geq c_j + q_{p_j}^{(j)},$$

where the $q_{p_j}^{(j)}$ refer to the p_j -quantiles of the one-dimensional distribution functions of $\langle b_j, \xi \rangle$.

Coming back to the model with joint probabilistic constraints (Eq. 6), the following peculiarity is of interest: If B happens to be surjective and ξ follows a nondegenerate multivariate normal distribution, then so will the transformed random vector $B\xi$. In many applications, however, surjectivity of B can fail (e.g., when B describes demand satisfaction in the nodes of a network). Then, the normal distribution of $B\xi$ will be degenerate regardless of the nondegeneracy of the distribution of ξ .

The second instance of a linear probabilistic constraint allows a coupling of random and decision vectors while keeping affine linearity in ξ . The most prominent model is induced from a system of linear inequalities with random coefficient matrix:

$$P(\Xi x \geq c) \geq p. \quad (7)$$

Note that this probabilistic constraint is easily rewritten in our standard form (1) upon putting

$$h(x, \xi) := M(x)\xi - c; \quad \xi := \text{vec } \Xi;$$

$$M(x) := \begin{pmatrix} x^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x^T \end{pmatrix},$$

where $\text{vec } \Xi$ refers to the row-wise vectorization of Ξ . Such models occur in problems where certain raw materials with uncertain concentrations are mixed (e.g., coffee blends or assets of a portfolio). Now, the coupling of x and ξ prevents a straightforward reduction of the probabilistic constraint (Eq. 7) to the case of random right-hand side. Nevertheless, it is still possible, for the aforementioned elliptically symmetric distributions, to rewrite Equation (7) in a form amenable to the calculus of multivariate distribution functions. To illustrate this in case ξ has a

normal distribution, assume that $\mathbb{E}\xi = \mu$ and $\text{Cov}(\xi, \xi) = \Sigma$. Then Σ is built up from blocks $\Sigma_{i,j} = \text{Cov}(\xi_i, \xi_j)$ (with ξ_i the i th row of Ξ) and one can easily show that

$$P(\Xi x \geq c) \geq p \iff F^{0, \Sigma(x)}(M(x)\mu - c) \geq p \quad (8)$$

Here, $F^{0, \Sigma(x)}$ refers to the distribution function of a multivariate normal distribution with zero mean vector and covariance matrix

$$\Sigma(x) = M(x)\Sigma M^T(x).$$

This means that for each fixed x the probability $P(\Xi x \geq c)$ can be calculated via the distribution function of a normal distribution, quite similar to what was the case for the random right-hand side model (Eq. 4). The difference to the latter consists in the fact that now the parameters of the distribution are not fixed but depend on x . This difficulty complicates the calculus of gradients to the probability function in the course of an optimization algorithm. Again, things simplify in a model with individual probabilistic constraints because—as a special case of the preceding observation—one has that

$$P(\langle \xi_i, x \rangle \geq c) \geq p_i \iff$$

$$F^{0, \sigma(x)}(\langle x, \mu_i \rangle - c_i) \geq p_i$$

$$(i = 1, \dots, m),$$

where now $F^{0, \sigma(x)}$ is the distribution function of the one-dimensional normal distribution with zero mean and variance $\sigma(x) = x^T \Sigma_{i,i} x$. Thanks to the one-dimensionality of the normal distribution, the usual standardization argument leads to the equivalence

$$P(\langle \xi_i, x \rangle \geq c) \geq p_i \iff$$

$$F^{0,1}(\sigma(x)^{-1/2}(\langle x, \mu_i \rangle - c_i)) \geq p_i$$

$$(i = 1, \dots, m),$$

where now the distribution function no longer depends on x . Passing to the quantiles q_{p_i} of the standard normal distribution, one arrives

again at a fully explicit description of the individual constraints:

$$\begin{aligned} P(\langle \xi_i, x \rangle \geq c) &\geq p_i \iff \\ q_{p_i} \sqrt{x^T \Sigma_{i,i} x} - \langle x, \mu_i \rangle + c_i &\leq 0 \\ (i = 1, \dots, m). \end{aligned} \quad (9)$$

This final form allows once more the passing of the individual probabilistic constraints directly to a nonlinear optimization solver.

Continuous and Discrete Distributions

In the models presented so far, only continuous distributions have been discussed.

The reason is that discrete distribution appears less favorable in the numerical treatment of probabilistic constraints than they are in expectation-based models of stochastic optimization. For instance, distribution functions of discrete distributions are discontinuous and lack concavity properties that may be hidden in continuous ones. Recent investigations, however, seem to justify an increasing interest in discrete distributions for solving probabilistically constrained programs. The key issue is finding the so-called p -efficient points [8] of the distribution function F_ξ of ξ . These are points z such that $F_\xi(z) \geq p$ and the relations $F_\xi(y) \geq p$, $y \leq z$ (partial order of vectors) imply that $y = z$. One easily observes that all the information about the p -level set of F_ξ is contained in these points because

$$\{y | F_\xi(y) \geq p\} = \bigcup_{z \in E} (z + \mathbb{R}_+^s),$$

where E is the set of p -efficient points and $\mathbb{R}_+^s$ is the positive orthant in the space of the random vector. In the case of ξ having integer-valued components and $p \in (0, 1)$, E is a finite set [Theorem 1 in Ref. 9]. Algorithms for enumerating or generating p -efficient points are described, for instance, in Refs 9 and 10.

STRUCTURAL PROPERTIES

In order to solve probabilistically constrained programs it is essential to know structural properties (e.g., continuity, differentiability,

concavity) of the function α defining the crucial inequality $\alpha(x) \geq p$ in Equation (2). The main challenge in establishing such properties consists in understanding the interaction between the random inequality system $h(x, \xi) \geq 0$ and the distribution of ξ . In simple cases, the effect is as expected. For instance, if the components h_i of h are upper semicontinuous (in both variables simultaneously), then α is upper semicontinuous as well and, as a consequence, the feasible set defined by the probabilistic constraint $\alpha(x) \geq p$ is closed, a minimum requirement in the context of optimization problems. This conclusion holds true regardless of any assumption about the distribution of ξ . Things become already much less obvious if (full) continuity is considered. Then, even the best distribution of ξ and the simplest type of h alone cannot guarantee the continuity of α . As an illustration, consider Equation (2), where ξ has a one-dimensional standard normal distribution and $h(x, \xi) = Qx + L\xi - b$ is affine linear with

$$(Q|L) = \left(\begin{array}{cc|c} 2 & 1 & -1 \\ -1 & 1 & 0 \end{array} \right); \quad b = \begin{pmatrix} 0 \\ -0.5 \end{pmatrix}.$$

Figure 1 shows the resulting probability function α which is evidently discontinuous.

It turns out that to the continuity of h one has to add the condition

$$\mathbb{P}(h_i(x, \xi) = 0) = 0 \quad \forall x \forall i,$$

which is violated in the example of the figure. Analogous unexpected phenomena occur in the characterization of Lipschitz continuity or differentiability of α which in the case of random right-hand side hinges upon the corresponding property of the distribution function F_ξ of ξ (4). For instance, it is obvious that for one-dimensional distribution functions Lipschitz continuity will follow from the boundedness of the underlying density, as soon as the latter exists. This conclusion does no longer hold true for dimensions larger than one. Then, it is not boundedness of the density itself but rather of the marginal densities which matters. On the other hand, the latter property can be easily verified for the class of so-called quasi-concave distributions. Similarly, the fact that

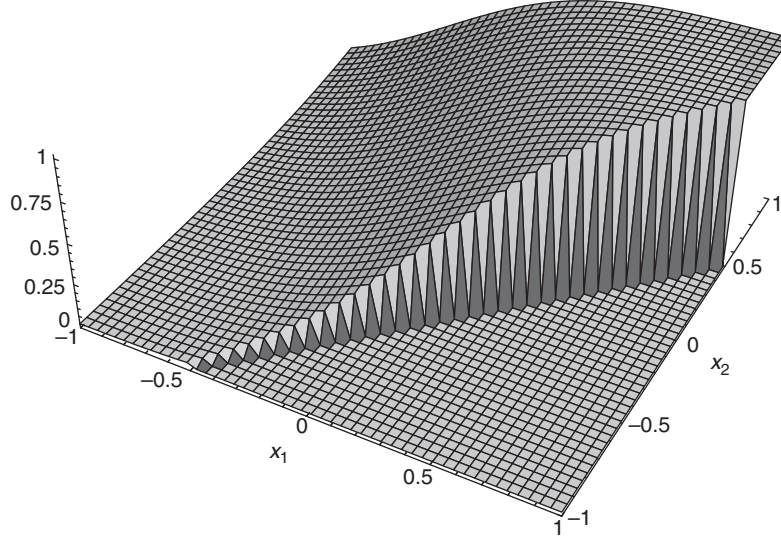


Figure 1. Example for a discontinuous probability function under smooth input data.

one-dimensional distribution functions are differentiable if they possess a continuous density does not carry over to larger dimensions. These difficulties are related with the calculus of improper integrals for densities with unbounded domain.

One of the most important features in numerical optimization is convexity of the set of feasible decisions. First of all, this is of interest in the case of random right-hand side (section titled “Random Right-Hand Side”). The question of convexity is very easy to answer if individual probabilistic constraints are considered. Given Equation (3), one has to check concavity of all the α_j (due to the inequality sign ‘ $\geq$ ’). Now, Equation (5) confirms that this concavity reduces to the concavity (or even just quasi-concavity) of the components g_j of g , which is easily checked as g is given explicitly. On the other hand, the individual model may be unjustified in many situations (see discussion in the section titled “Joint and Individual Probabilistic Constraints”). Things become more involved in the case of joint probabilistic constraints, because then convexity/concavity properties of multivariate distribution functions come into play. According to Equation (4), one would have to check concavity of the composition $F_\xi \circ g$.

Since distribution functions are increasing (with reference to the partial order), concavity of this composition would follow easily from concavity of the components g_j (which is easy to check) and from concavity of F_ξ . Unfortunately, being bounded from below and above but not constant, distribution functions are never concave (or convex). On the other hand, suitable transformations could do the job. For instance, upon passing in Equation (4) to the equivalent (by monotonicity of $\log$) inequality $\log F_\xi(g(x)) \geq p$, one realizes that concavity of $\log F_\xi$ would suffice. Checking log-concavity of F_ξ from the definitions is generally impossible due to a lack of explicit formulae for multivariate distribution functions. However, by a deep result due to Prékopa [11], log-concavity of distribution functions is equivalent with log-concavity of the underlying densities (if they exist). As one has typically access to explicit formulae for the densities, this offers a comfortable way to verify the desired property for the distribution functions. Given this, it turns out that a large class of multivariate distribution functions [including multivariate normal, Dirichlet, Gamma, uniform etc., refer [2]] shares the property of being log-concave. This allows us to apply convex optimization tools to

(joint) probabilistic constraints with random right-hand side whenever the random vector ξ has such distribution and the components g_j are concave (e.g., if g is an affine linear mapping).

The situation is again more complicated even for linear probabilistic constraints outside the setting of random right-hand side. Indeed for a probabilistic constraint of the type in Equation (7), no general convexity result is known even if ξ has a multivariate normal distribution. Again, individual probabilistic constraints are much easier to analyze. Indeed, having a look at Equation (9) and observing that $\sqrt{x^T \Sigma_{i,i} x}$ represents a norm, this inequality defines for arbitrarily fixed i a convex feasible set provided that $q_{p_i} \geq 0$ or, in other words, provided that $p_i \geq 0.5$. This observation is a classical result due to Van de Panne/Popp/Kataoka. As convex sets are stable under intersection, the whole system of inequalities (Eq. 9) is convex if $p_i \geq 0.5$ for all indices i . This restriction is not harmful because in probabilistic programming one is usually interested in probability levels close to 1. Generalizations to the case of joint probabilistic constraints of the type in Equation (7) are known up to now only for particular cases. For instance, when Ξ has a multivariate normal distribution and all rows of ξ are pairwise independent (whereas entries within each row may correlate) then the set of feasible decisions x defined by Equation (7) is convex for sufficiently large probability level p (12).

For a detailed analysis of convexity (in the individual model) and compactness (in the individual and joint model) of the feasible set defined by Equation (1) under elliptically symmetric distributions of ξ , we refer the reader to Ref. 13.

REFERENCES

1. Prékopa A. Probabilistic programming. In: Ruszczyński A, Shapiro A, editors. Volume 10, Stochastic programming: handbooks in

- operations research and management science. Amsterdam: Elsevier; 2003, Chapter 5.
2. Prékopa A. Stochastic programming. Dordrecht: Kluwer; 1995.
3. Gassmann HI, Deák I, Szántai T. Computing multivariate normal probabilities: a new look. J Comput Graph Stat 2002;11:920–949.
4. Genz A. Numerical computation of the multivariate normal probabilities. J Comput Graph Stat 1992;1:141–150.
5. Szántai T. Evaluation of a special multivariate Gamma distribution. Math Program Study 1996;27:1–16.
6. Gouda A, Szántai T. New sampling techniques for calculation of Dirichlet probabilities. Cent Eur J Oper Res 2004;12:389–403.
7. Olieman N, van Putten B. Estimation method of multivariate exponential probabilities based on a simplex coordinates transform. J Stat Comput Simul 2010;80:355–361.
8. Prékopa A. Dual method for a one-stage stochastic programming problem with random RHS, obeying a discrete probability distribution. Z Oper Res 1990;34:441–461.
9. Dentcheva D, Prékopa A, Ruszczyński A. Concavity and efficient points for discrete distributions in stochastic programming. Math Program 2000;89:55–79.
10. Prékopa A, Vizvári B, Badics T. Programming under probabilistic constraint with discrete random variables. In: Giannessi F, Rapcsák T, editors. New trends in mathematical programming. Dordrecht: Kluwer Academic Publishers; 1997. pp. 235–257.
11. Prékopa A. On logarithmic concave measures and functions. Acta Sci Math (Szeged) 1973;34:335–343.
12. Henrion R, Strugarek C. Convexity of chance constraints with independent random variables. Comput Optim Appl 2008;41:263–276.
13. Henrion R. Structural properties of linear probabilistic constraints. Optimization 2007;56:425–440.

MONTE CARLO SIMULATION AS AN AID FOR DECIDING AMONG TREATMENT OPTIONS

STEVEN SHECHTER

Operations and Logistics Division,
Sauder School of Business,
University of British Columbia,
Vancouver, British Columbia,
Canada

Few decisions are as important and complex as choosing among treatment options. While randomized controlled trials (RCTs) are often considered the gold standard for deciding among treatment alternatives, they may be too expensive, risky, or time consuming to inform decision making in the near term. Yet decisions must go forward. And although physician experience is undoubtedly a key consideration, it is also important to synthesize available data into a critical analysis of the various options under consideration. This paper describes one increasingly popular technique for doing this, which contributes to the spirit of “evidence-based medicine:” Monte Carlo simulation (MCS) modeling.

The last two decades have seen significant increases in both computational power and the options available for treating various diseases. Though seemingly unrelated, these two evolutions have been wed over the years through the application of MCS modeling for evaluating treatment alternatives. By projecting the stochastic progression of a disease and the effects of various treatments over longer time periods, MCS models provide a useful methodology for evaluating and comparing treatment policies. For example, a MCS model has been used to evaluate the effects of different treatment start times for HIV patients [1]. Additionally, one may attach costs to the treatments and events that take place during the patient’s simulated lifetime, so as to evaluate the cost-effectiveness of the treatments alternatives. Examples include

the cost-effectiveness analysis (CEA) of HIV screening [2] and colorectal cancer screening [3]. The CEA that results by considering societal costs (and benefits) in these models is a useful tool for governments and insurers to decide which treatments should be covered and which should not. This is essential in an age of increasing health-care costs and limited health-care budgets.

Consistent with the recent focus on “patient-centered care” [4,5], MCS models may also be adapted to model the decision of a specific patient, rather than an “average” patient from a cohort. For example, some patients elicit different preferences for being in the same health conditions as well as different preferences for the value of time. MCS models can easily handle these considerations by incorporating patient-based utilities for being in various health states as well as discount rates for future outcomes. These often get summarized in one output measure used in the medical decision making literature for evaluating treatment benefits—quality-adjusted life years (QALYs) [6,7]. In other words, one treatment may result in a rather different estimate of QALYs for two different patients with the same health condition, due to their different valuations of quality of life. Moreover, if a particular patient’s disease progression does not seem characteristic of the cohort of patients upon whom the models were based (e.g., the patient under consideration may be a “rapid progressor”), the MCS model can be altered to reflect this and analyzed under that setting. The implication is that for some patients, a model may recommend one treatment while for other patients it may recommend a different treatment.

THE MONTE CARLO APPROACH

As disease progression is often a complex stochastic process, long-term patient outcomes are well suited to evaluation via simulation. These models work by using

(pseudo) random numbers to generate random changes to patient health (e.g., laboratory values or disease status) in a way that is consistent with a probability distribution representing the possible changes (for more on random number and random variable generation, we refer the reader to simulation textbooks such as Refs 8 and 9). These probability distributions, in turn, are based on available data and/or clinically reasonable assumptions. By simulating many sample paths of a patient's lifetime and then averaging outcomes, one may obtain estimates of several output measures of interest such as expected quality-adjusted lifetime and lifetime costs. As in traditional statistical sampling, the accuracy and precision of these estimates improves as more patient lifetimes are simulated.

Often, the underlying stochastic process is modeled as a discrete-time, discrete-state, Markov reward process [10–12]. For example, one might model monthly transitions between discrete health states (which may include categorized continuous variables) and accrue state-based benefits (e.g., quality-adjusted time in that health state) and/or costs over the patient's lifetime. Under restrictive assumptions (e.g., transition probabilities and rewards do not depend on time), one can apply matrix algebra to obtain the exact solution to a set of equations, which yield expected total outcomes. However, realistic models of lifetime patient outcomes must consider time-varying factors, such as age-based mortality rates. One way of dealing with this discussed in the medical decision making literature is referred to as the *Markov cohort simulation* approach [10,12], in which some arbitrary-sized cohort (e.g., 10,000 patients) is filtered at each time step according to the appropriate state transition probabilities, and members of each filtered subgroup accrue benefits and costs according to which state they moved. Because patient models have the absorbing death state, all patients get filtered to this state in the limit. The cohort simulation approach terminates when the number of patients still alive becomes sufficiently small, and then total rewards are tallied and divided by the original cohort

size. This approach is tantamount to working out the limit of a non-time-homogeneous Markov reward process, which does not have a straightforward matrix algebra solution as in the simple, time-homogeneous setting described before.

The other approach for evaluating outcomes discussed in the literature is MCS [10,12]: individual patients go one-by-one through the model, with each patient's state-trajectory until death determined by appropriate draws from the transition probabilities at each time period (e.g., patient health might be updated every month). Because of the probabilistic progression represented through these probabilities, two patients going through the model may have very different recorded outcomes (e.g., one simulated patient may die within a year, while another simulated patient with the same starting condition and disease dynamics may survive five more years). By running several patients through the simulation model and averaging results, the MCS approach provides an estimate of expected outcomes. The individual patient histories can also be used to understand other measures of the distribution of possible outcomes, such as variance, percentiles, and overall shape of the distribution (e.g., with a histogram). Moreover, by having a record of each patient's history, MCS allows for the easy calculation of several possible outcomes of interest, in addition to the expected lifetime calculation typically sought with the cohort simulation approach. For example, in a model of HIV treatment, one can use MCS to estimate the time until an AIDS-defining illness occurs, the time spent on different treatment regimens, or the probability of dying from HIV versus comorbid diseases [1]. Note that the MCS method is not exact: it provides a sample-based estimate, which itself is subject to variability (typically described through a confidence interval). However, this sampling variability vanishes to zero as the number of simulated patients approaches infinity, and by the law of large numbers, the sample mean converges to the distribution mean. We discuss statistical output analysis more below.

Besides allowing for a better understanding of the distribution of outcomes, MCS has other advantages that arise from its considerable flexibility. For example, it allows for easy consideration of transition probabilities that depend on patient histories (such as previous laboratory test readings, prior heart attacks, etc.). Technically, Markov models have state transitions that are independent of past history. While one can get around this by expanding the state definition to include some past history, this quickly expands the state space, making solution via matrix algebra or cohort simulation rather cumbersome. MCS is also very useful for establishing the face validity of a model: by observing generated patient histories, one can ascertain whether the model appears to be valid or not. For example, if a patient's simulated trajectory appears clinically implausible (or impossible), then the model should be refined to ensure that a similar trajectory cannot occur again.

Another advantage of MCS is that it no longer requires the traditional Markov process restrictions of discrete states and time steps. For example, one might use MCS to sample variability around a regression model [1] or a differential equations model [13] representing a continuous change in a patient variable over time, rather than break the variable down into arbitrary categories. Also, one might have an MCS model that allows a patient to remain in the same state for an amount of time that follows some distribution, rather than try to break that period of time up into smaller, discrete time periods. For example, in a model of end-stage liver disease and the effects of donated liver allocation policies, one might update patients' natural histories periodically in the model (e.g., monthly, so that one has up-to-date information in case an organ becomes available) [14], but then model the posttransplantation outcomes by just drawing the random time until either the patient's organ fails (and he/she needs to relist for a transplant), or the patient dies [15].

Even if a simple Markov process is used as a baseline model, many researchers will still apply MCS to conduct probabilistic sensitivity analyses (PSA) [16]. In general,

sensitivity analysis (SA) is an important step of model building: because one cannot be 100% certain of the accuracy of model inputs, SA allows one to understand how variation of the inputs affects variation of the outputs. The idea is that where outputs are sensitive to variation in inputs, one should collect more data to improve the accuracy of the inputs. While one may take a one-parameter-at-a-time approach to conducting SA, PSA considers simultaneous uncertainty (represented by a probability distribution over a range of values) across several parameters. As this is too complicated for exact analysis, MCS is used to estimate expected outcomes as well as proportions of the time that one treatment option yields better outcomes than the others.

Note that we distinguish MCS from traditional discrete-event simulation (DES) models, even though both use computer-generated random numbers to drive probabilistic outcomes. Our reason for doing so stems from the fact that simulation modeling is used widely in health care, and the applications typically fall into one of two types. The first is the patient-level disease modeling that we have described so far and refer to as MCS. The other applications involve models of health-care systems, in which patients join queues and compete for limited resources. A classic example is a simulation model of an emergency room, which considers the arrival pattern of patients, along with the available nurses, physicians, and beds, to estimate outcomes such as average patient wait time and resource utilization. For an overview of these and other similar types of process simulations in health care, we refer the reader to Ref. 17. These types of simulations are more traditionally referred to as *DES* (it should be noted, however, that some authors use "DES" in the way we use "MCS"). Common to both types of health-care simulations is a complex stochastic process which is not amenable to exact analytical solution, and for which a decision maker wants to evaluate various "what if" scenarios. In MCS, the stochastic processes are generally changes to patient health, while in DES, the uncertainty arises from randomness in patient arrival

times, service times (e.g., surgery duration), and resource availability (e.g., absenteeism and machine break down). The intended decision makers using MCS are patients, their physicians, and organizations tasked with deciding which treatments to pay for or not; users of DES are typically health-care managers and administrators in charge of the efficient operation of a health-care unit or system. Typically, MCS does not involve queueing: it assumes that the patient has access to whichever treatment is under consideration, and this is independent of other patients' path of care. However, some models do combine elements of MCS (the stochastic progression of patient health) with DES (patients competing for resources). For example, in the model of liver allocation described previously [15], MCS is used to stochastically evolve the health of patients waiting with end-stage liver disease, and DES is used to generate arrivals of patients and liver donations at random, discrete points in time. For a further discussion of modeling patient health progression within a DES framework, we refer the reader to Ref. 18.

STATISTICAL ANALYSIS OF SIMULATION OUTPUT

Randomness in inputs leads to randomness in outputs; therefore, a proper statistical analysis should be performed on an appropriate number of simulation replications. Whereas a replication in a DES model may represent what occurs at a clinic over the course of a day, a replication of an MCS model represents a particular patient's journey, from the starting conditions (e.g., age and state of disease) until death. Variability in disease progression and response to treatment may lead to different clinical events for different patients, resulting in different lifetimes, quality-adjusted lifetimes, and costs. MCS may be viewed as a computational method for sampling from an unknown distribution (e.g., the distribution of patient lifetimes under treatment A), and as with any statistical sample, we are interested in both point estimates of

outcomes of interest, as well as confidence intervals around these.

Suppose one wants to estimate the expected patient lifetime under a given treatment policy, with a 95% confidence interval half length of at most h years. A two-step heuristic for determining the appropriate number of patient replications is as follows: (i) run some moderately sized number, n_1 , of patients through the model, and find the sample standard deviation, s , of their lifetimes; (ii) run $n_2 = (1.96 \times s/h)^2 - n_1$ additional patients through the model and form the sample average and confidence interval for the $n_1 + n_2$ patient replications. If the resulting confidence interval is still not to the desired precision, then additional patients can be run through the model [9].

More commonly, the purpose of a MCS model is to compare patient outcomes under two or more different treatment policies. When comparing two treatment policies A and B, one can simulate n_A patients under A, n_B patients under B, and then perform an appropriate two-sample statistical analysis on the differences in outcomes between the two treatments. In fact, one can pair outputs from an identical number of replications of A and B and form the pair-wise differences, thus transforming the two-sample analysis into a standard one-sample analysis on the differences. This paired approach is especially useful when applying the variance reduction technique of common random numbers (CRN), which may yield similar precision in estimates but with many fewer replications (a discussion of CRN as applied to MCS is given in Refs 19 and 20). The idea behind CRN is that the random numbers used to generate events in the replication of one policy should also be used to generate events in the replication of another policy, so that one policy does not appear better than the other just because it received a more favorable stream of random numbers. For example, if the probability of death in a particular month under policy A is p_A and under policy B is p_B , then to simulate whether a patient dies that month under a given replication of A and B, the CRN approach would generate an instance of $u \sim \text{Uniform}[0,1]$, and if $u \leq p_A$, the patient

dies under policy A and if that same $u \leq p_B$, the patient dies under policy B. In other words, CRN avoids the replication giving a low Uniform[0,1] number to A but a higher one to B (which is more favorable to survival). The key to CRN is properly synchronizing the random number streams between policies so that they are used for similar purposes. Whereas this can be difficult to achieve for general DES models due to entities interacting with each other, MCS models are well suited to being able to achieve perfect or near-perfect synchronization, resulting in significant computational savings [19].

In comparing multiple treatment policies (say M), different perspectives may be taken. For example, if one policy is considered the status quo, then $M - 1$ confidence intervals may be generated (using the Bonferroni procedure), comparing the status quo policy to each other policy. If none is considered a baseline policy, then one can perform an all-pairwise comparison among the $M(M - 1)/2$ pairs. Ranking and selection procedures make statements about which policy is either the best or which subset of policies contains the best. For example, certain algorithms allow one to determine how many additional replications should be performed to be 95% confident that the policy with the best sample average is actually the one with the best true expected value. Other procedures allow one to determine how many of the top policies (according to the sample averages) should be included to achieve 95% confidence that the resulting subset contains the policy with the best expected value. This technique is sometimes performed as a first screening step to identify just a few of many policies as candidates for the best. For a further discussion of these methods, we refer the reader to Ref. 8.

CHOICE OF SOFTWARE

There are a wide variety of software options available for developing MCS models, ranging from general purpose languages (such as C or Java) to decision-analytic software packages with MCS capabilities (such as

TreeAge or Crystal Ball) to DES programs (such as Arena or Simul8). Modelers should spend time up front making a careful decision on software selection; otherwise, they may find that they have invested considerable time and money developing a model in one software language only to find that the model has become too complex to be adequately handled by that package. The choice of software may depend, among other things, on the following: (i) The degree of modeling flexibility one desires—general purpose languages having the most flexibility. (ii) The degree of computer programming experience one has—simulation-specific packages generally have nice graphical user interfaces which do not require programming expertise, as opposed to the line-by-line coding of general programming languages. (iii) Whether or not animation is important—animation is rarely used in MCS models of disease progression and treatment, but is often desired with DES models of patient flow through a system. DES packages have nice animation capabilities whereas, it is extremely burdensome to create an animation when working with a general programming language. (iv) Execution speed—for a given model, the run time of several patient replications will generally be much faster when executed from within a general programming language than from within a specialty simulation program. (v) Price—software products vary widely in the purchase cost, and this is where one would have to look into the features of each product, technical support, and so on, in making a decision.

CHALLENGES IN MONTE CARLO SIMULATION MODELING

Most of the challenges in MCS modeling center on the validation process. This is the means by which one gains confidence that the insights and conclusions of the modeling results would be similar to actual results if the same decisions were made in practice. Therefore, as with any modeling effort, the availability of useful data is essential. This may require prospectively collecting the data, obtaining retrospective data, or referring to

available literature (and perhaps combining published results into a meta-analysis). One should also be cognizant of what treatment policies were in place and whether they changed over the timeline of patients' longitudinal data, as this could affect the interpretation of patient health progression over time.

As mentioned above, having a clinician look at sample trajectories generated by the MCS to ascertain their plausibility is an important part of the validation process. After ensuring that the model does not generate any clinically absurd behavior, the validation process continues by comparing simulated output measures with available actual output measures. For example, one may compare the model with actual outcomes such as average lifetime, five-year survival probability, number of hospitalizations, time on therapy, and so on. Graphical and tabular comparisons are often used in these contexts. While traditional statistical hypothesis testing may be done to compare actual and simulated patient outcomes, one should keep in mind that a simulation model is necessarily an abstraction of reality; therefore, one should not strive to confirm a null hypothesis of equality of means, which is surely false [8]. In fact, with computer simulation, one can easily increase the number of patient replications to a point where the null hypothesis will be rejected. Instead, confidence intervals are more useful, as they shift the focus from whether there is a statistical significance of differences, to whether there is a clinically relevant difference between actual and simulated outcomes. Finally, SA is an important validation step, as it helps determine where more data should be collected or which assumptions should be carefully reexamined.

Time also presents validation challenges, in two different respects. First, a significant amount of time may pass from problem definition, to data collection and analysis, to model construction, and to model validation. Therefore, one should recheck that the model assumptions and data used near the beginning of the project timeline are still relevant near the end. Second, a key issue in

modeling long-term patient outcomes is the fact that medical research and development often leads to significant changes in the treatments available to patients over the course of their lifetimes. For example, in just the past 15 years, the standard for treating HIV has evolved from patients taking a single available drug (AZT) to them taking three different drugs (out of several available) simultaneously. This has led to dramatic improvements in patient survival. Therefore, an MCS model of HIV progression constructed just before new classes of antiretroviral drugs were developed (in 1995) would estimate radically different patient outcomes compared to a model constructed after this time. Because forecasting new treatment development is difficult at best, this remains a significant modeling challenge for the research community.

CONCLUSIONS

MCS takes advantage of computational power to estimate a variety of patient outcomes that may be too difficult to estimate otherwise. It should be viewed as one piece of a medical decision making process, rather than as a substitute for traditional clinical approaches to decision making. If randomized clinical trials are possible to run with a reduced set of treatment arms, then MCS provides a useful platform for narrowing down a multitude of potential treatment arms to consider in an RCT. If a clinical trial is simply not feasible to run, then an MCS model, in combination with clinical expertise and oversight, may recommend a best course of treatment for patients. Furthermore, one may incorporate costs into these models so as to determine the cost-effectiveness of the various treatment alternatives. That these models can impact policy recommendations is evidenced by a study that suggested it is cost-effective to screen nearly all adults for HIV, not just those at high risk for contracting the disease [2]. This influenced new CDC guidelines in 2006 that recommended routine screening for all people 13–64 years old [21]. Also,

MCS was recently used to support the development of new colorectal cancer screening guidelines in the United States [3,22].

FURTHER READING

In addition to all the references noted, we refer the reader to Ref. 23 for an overview of various modeling approaches to medical decision making. For a discussion of choosing one decision modeling approach over another, we refer the reader to Ref. 24. Recent developments in decision-analytic modeling in health care (including a discussion of patient simulation models), are discussed in Ref. 25. An annotated bibliography of simulation modeling in health care (covering both operational and medical decision making examples) is presented in Ref. 26. For more details on computer simulation in general, the reader is referred to various textbooks on this topic [8,9]. While these texts focus on traditional DES, they contain material that applies just as well to MCS (for example, detailed discussions of random number generation, validation, and statistical analysis).

Acknowledgment

The author thanks Adelle Bueno and Nick Bansback for their help in providing useful references.

REFERENCES

1. Braithwaite RS, Roberts MS, Chang CCH, *et al.* Influence of alternative thresholds for initiating HIV treatment on quality adjusted life expectancy: a decision model. *Ann Intern Med* 2008;148(3):178–185.
2. Paltiel AD, Walensky RP, Schackman BR, *et al.* Expanded HIV screening in the United States: effect on clinical outcomes, HIV transmission, and costs. *Ann Intern Med* 2006;145(11):797–806.
3. Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, *et al.* Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. preventive services task force. *Ann Intern Med* 2008;149(9):659–669.
4. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA* 2008;300(4):436–438.
5. Laine C, Davidoff F. Patient-centered medicine: a professional evolution. *JAMA* 1996;275(2):152–156.
6. Gold MR, Patrick DL, Torrance GW, *et al.* Identifying and valuing outcomes. In: Gold MR, Siegel JE, Russell LB, *et al.*, editors. *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. pp. 82–134.
7. Hunink MGM, Glasziou PP, Siegel JE, *et al.* *Decision making in health and medicine: integrating evidence and values*. Cambridge: Cambridge University Press; 2001.
8. Law AM, Kelton WD. *Simulation modeling and analysis*. 3rd ed. New York: McGraw-Hill; 2000.
9. Banks J, Carson J, Nelson BL, *et al.* *Discrete-event system simulation*. 4th ed. Englewood Cliffs (NJ): Prentice Hall; 2004.
10. Beck JR, Pauker SG. The Markov process in medical prognosis. *Med Decis Making* 1983;3(4):419–458.
11. Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics* 1998;13(4):397–409.
12. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making* 1993;13(4):322–339.
13. Schlessinger L, Eddy DM. Archimedes: a new model for simulating health care systems—the mathematical formulation. *J Biomed Inform* 2002;35(1):37–50.
14. Alagoz O, Bryce CL, Shechter S, *et al.* Incorporating biological natural history in simulation models: empirical estimates of the progression of end-stage liver disease. *Med Decis Making* 2005;25(6):620–632.
15. Shechter SM, Bryce CL, Alagoz O, *et al.* A clinically based discrete-event simulation of end-stage liver disease and the organ allocation process. *Med Decis Making* 2005;25(2):199–209.
16. Dobilet P, Begg CB, Weinstein MC, *et al.* Probabilistic sensitivity analysis using Monte Carlo simulation: a practical approach. *Med Decis Making* 1985;5(2):157–177.
17. Jun JB, Jacobson SH, Swisher JR. Application of discrete-event simulation in health care clinics: a survey. *J Oper Res Soc* 1999;50(2):109–123.
18. Caro JJ. Pharmacoeconomic analyses using discrete event simulation. *Pharmacoeconomics* 2005;23(4):323–332.

19. Shechter SM, Schaefer AJ, Braithwaite RS, *et al.* Increasing the efficiency of Monte Carlo cohort simulations with variance reduction techniques. *Med Decis Making* 2006;26(5):550–553.
20. Stout NK, Goldie SJ. Keeping the noise down: common random numbers for disease simulation modeling. *Health Care Manag Sci* 2008;11:399–406.
21. http://publichealth.yale.edu/news/dec06/paltiel_study.html, 2009.
22. <http://cisnet.cancer.gov/colorectal/highlights/zauber2008.html>, 2009.
23. Roberts MS, Sonnenberg FA. In: Chapman GB, Sonnenberg FA, editors. *Decision making in health care: theory, psychology, and applications*. New York: Cambridge University Press; 1999.
24. Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies. *Health Econ* 2006;15:1295–1310.
25. Weinstein MC. Recent developments in decision-analytic modelling for economic evaluation. *Pharmacoeconomics* 2006;24(11):1043–1053.
26. Klein RW, Dittus RS, Roberts SD, *et al.* Simulation modeling and health-care decision making. *Med Decis Making* 1993;13(4):347–354.

MONTE CARLO SIMULATION FOR QUANTITATIVE HEALTH RISK ANALYSIS

ALISON CULLEN
Evans School of Public Affairs,
University of Washington,
Seattle, Washington

INTRODUCTION AND BACKGROUND

Managing risky opportunities, scenarios, and decisions is one of the most interesting and challenging aspects of human life. But meaningful characterization of risks is elusive because it relies on information that is incomplete, imperfect, and/or variable across time, space, and populations. And consequently informed decision making may be compromised. Probabilistic tools such as Monte Carlo simulation have evolved in which uncertainty and variability are characterized and carried through a process of generating probabilistic estimates of risk. The goal of this article is to introduce basic tools and techniques, as well as to provide pointers to other sections of this encyclopedia and some of the more common extensions practiced in the field. In order to preserve a focus, we restrict our treatment to the assessment of human health risk, while acknowledging that many other fields including finance, engineering, and law, can and do adapt these tools to a range of purposes extending far beyond those covered here. This article begins with an introduction to human health risk assessment. Following that introduction we explore the concepts of uncertainty and variability, the basics of Monte Carlo simulation, the sources of information used in this approach, probabilistic representation of risk model inputs, propagation of uncertainty and variability, and the identification of significant contributors to output variance.

The history of probabilistic risk assessment encompasses many fields of application including military decision making, engineering, medicine, and finance. These continue to be active fields of endeavor today. We find examples such as the use of random sampling techniques to solve challenging deterministic equations in the Manhattan Project at Los Alamos [1–3], applications in regulatory policy development such as the Reactor Safety Study with its eventual role in the development of probabilistic safety objectives for power plants [4] and the much contested application of value at risk models for financial risk management in which the unquantified tail risk accounts for a small but important fraction of the total risk [5]. In climate science, researchers confront output from a multimodel ensemble of atmosphere–ocean general circulation models and observations, introducing an overlay of model uncertainty, which joins the ever present uncertainty and variability, related to inputs such as future emission scenarios and climate sensitivity [6]. In medicine, we find cases in which interindividual differences between people throw into stark relief the distinction between clinical case management and epidemiological approaches to risk across populations [7], driven by the same concepts we confront in health risk analysis, our focus for this article.

Human health risk analysis produces estimates, ranges, and distributions related to the magnitude of adverse health effects resulting from man-made and naturally occurring health hazards. Decisions about how best to remove, reduce, or manage these risks are sometimes made by individuals on their own behalf. But these decisions are also made by regulators and other decision makers on behalf of subpopulations, large and small, aware and unaware, close to the source of risk or far removed from it. Decision making is complicated by limitations in our understanding of the physical, chemical, and

biological processes that mediate between hazards and people, limitations in the information and models available, and basic gaps in knowledge at every step of this chain, which result in our being uncertain about the magnitude and distribution of risk. Further complexity stems from the very real differences in exposure and sensitivity, which lead to very different likelihoods of negative health effects across populations, time, and space. These differences are referred to as *variability*.

More specifically, health risk is estimated as the product of exposure and dose/response factors and is expressed as the probability of a health outcome. The term *exposure* represents the level, duration, and frequency of human contact with an agent of risk and the route by which that contact occurs, for example, dermally, by inhalation, or by ingestion. Sources of exposure may be uncertain and/or exhibit variability over time, space, climate conditions and seasons, actions that mitigate or exacerbate the interaction of people and risk agents, and human behaviors and time activity patterns. Dose/response or exposure/response factors estimate the likelihood of a health outcome, that is, a response, given a specified degree of exposure. Dose/response relationships are uncertain and variable across individual populations, over genetic makeup, lifetimes, disease status, environmental triggers, and combinations of health threats. If major decisions are to be adequately supported, it is important to characterize and address both uncertainty and variability at an appropriate level [8–11].

TOOLS FOR ADDRESSING UNCERTAINTY AND VARIABILITY—SIMPLE TO COMPLEX

Addressing uncertainty and variability in a risk analysis takes many forms. At the simplest, it may refer to a qualitative discussion of data gaps and differences between human subpopulations or between different risk models themselves. Or it may refer to the acknowledgment of a basic and profound lack of insight into a future source of risk, a population at risk, or

the conditions under which the population and the risk source interact. These gaps can be handled by defining and running scenarios in which pivotal uncertainties are fixed with specific assumptions, and by multimodel analyses that grapple with model uncertainty by comparisons among model structures. Traditionally, when single values are assigned to risk model inputs they are selected to be protective of health, and in combination these can result in a “cascade of conservatism” far beyond the level of protectiveness that may have been intended originally [12,13]. At the other end of the spectrum, addressing uncertainty and variability, we find highly sophisticated and complex mathematical treatments based on statistical and/or artificial intelligence tools for estimating risk. Between these extremes, we find analytical methods exploiting mathematically exact solutions whose application rests on knowledge of the moments of the uncertain risk model inputs, and also numerical simulation methods such as Monte Carlo analysis, which rely on probabilistic descriptions and random number generators to carry uncertainty through the process. With the rise of desktop computing accompanied by vast leaps in computational power, speed, and ease, Monte Carlo methods have gained steadily in the field of risk analysis, while analytical solutions and approximations are increasingly rare for these applications. Detailed information about simulation approaches, in general, appears in the section titled “Simulation Modeling and Analysis”, with Monte Carlo Method appearing as the focus of the section titled “Monte Carlo Method” in this encyclopedia.

In order to justify the additional time, effort, or expense of extra layers of analysis it is necessary to consider the consequences of failing to pursue it. Where the cost of alternatives or post hoc remediation is high, there may be an argument for more sophisticated analysis. Likewise, where a rigorous comparison of options is needed, where the direction of additional scientific research will be decided, or where the stakes are potentially very high for any reason—it is easier to justify an additional focus on analysis.

MONTE CARLO SIMULATION FOR DEALING WITH UNCERTAINTY AND VARIABILITY

Monte Carlo simulation (and probabilistic analysis more generally) is neither the simplest nor the most complex approach to dealing with uncertainty and variability, as outlined at the beginning of this article. Its application has become very common in some areas of risk analysis, while it remains virtually nonexistent in others [14]. In very basic terms, *Monte Carlo simulation* refers to an approach in which probability and/or frequency distributions (histograms) are used to represent the components of risk models and are combined statistically to generate distributed estimates of risk. This is an alternative to deterministic approaches in which a single number is assigned to each input in a risk calculation, which is in turn carried out for a single context or scenario, resulting in a single number estimate of risk. By propagating representations of uncertainty and/or variability in inputs, assumptions, and decision contexts, through risk models, it is possible to generate a probabilistic description of risk which can open up a world of insight into related decisions.

While variability and uncertainty may be addressed with similar tools and techniques their roles in the decision process are markedly different. “Uncertainty forces decision makers to judge how probable it is that risks will be overestimated or underestimated for every member of the exposed population, whereas variability forces them to cope with the certainty that different individuals will be subjected to risks both above and below any reference point one chooses” [15]; see also Ref. 16. And the differences extend into the steps that follow analysis as well. Complex analysis and follow-up research have the potential to reduce uncertainty about an estimate of risk; however, more information is unlikely to reduce variability, since it reflects true and inherent differences over time, space, and individuals or populations. Further, most inputs that inform risk models carry both uncertainty and variability with them.

For example, the concentration of particulate matter in ambient air varies with location, averaging time, time of day, season and patterns of industrial, transportation, and individual behavior. The measurement of air pollution is an imperfect science; thus even with assumptions in place that fix all of the previously mentioned sources of variability exactly, we would still be uncertain about the concentration of particulate matter in ambient air due to instrument and analytic error, randomness, and other causes. To help distinguish the sources of variance that reflect true differences from those that reflect a lack of understanding, some analysts and decision makers prefer multidimensional analyses in which uncertainty and variability are represented in separate dimensions. These analyses result in multidimensional output, for example, probability surfaces rather than one-dimensional probability distributions of risk. For example, there may be variability in how much of a particular contaminant different individuals in the population are exposed to, as well as uncertainty about how dangerous the contaminant is to individual people who are exposed. Multidimensional analyses allow a simultaneous look at both of these aspects of risk and sometimes present opportunities to gauge whether uncertainty or variability is dominant in a particular instance.

Decision makers value analysis of both uncertainty and variability because it provides information that is useful for the communication, interpretation, or justification of a decision. It may provide a measure of the degree to which an estimate is conservative or a decision equitable. It may also reveal the relative impact of individual sources of uncertainty and variability on the decision at hand.

SOURCES OF INFORMATION—OBJECTIVE AND SUBJECTIVE

Sources of information for risk analyses include data, model results, expert judgment, and combinations of these. They span both objective and subjective categories as well. Perhaps not surprisingly there

is overlap between these; for example, judgment influences how, why, and what empirical data are collected, as well as how modeling efforts are approached and interpreted. Still it may be convenient to think in terms of objective and subjective information sources. And the development of probabilistic descriptions of model inputs necessarily depends on the nature, quantity, and quality of available information.

Where data exist it is possible to generate histograms representing past measurements. An example of this is the extensive database of measurements for factors such as fish consumption rates, time-activity patterns, and water ingestion rates published in the *Exposure Factors Handbook* [17]. Where insight into the processes that govern particular inputs is present, it is possible to develop probability distributions based on theoretical understanding. An example of this is the use of lognormal distributions to represent concentrations of environmental contaminants resulting from the knowledge that sequential dilution processes lead to lognormally distributed quantities [18,19].

Judgment in the form of expert elicitation is useful for generating distributions to represent quantities for which data exist but may be plagued by inherent systematic biases related to the timing, population, or conditions through which they were collected, for example, estimates of the increase in lung cancer associated with a $1\mu\text{g}/\text{m}^3$ increase in the concentration of fine particulate matter in air [20]. The elicitation of subjective probabilities and correcting for bias are the subject of articles *Eliciting Subjective Probabilities from Individuals and Reducing Biases*, *Eliciting Subjective Probability Distributions from Groups*, *Multivariate Elicitation: Association, Copulae, and Graphical Models*.

PROBABILISTIC REPRESENTATIONS OF UNCERTAINTY AND VARIABILITY

In Monte Carlo simulation, probability distributions link the possible values of a random variable, or ranges of possible values, with probabilities. A probability distribution

may be expressed as a probability density function (pdf) or a cumulative distribution function (cdf). For a continuous random variable, the area under a pdf represents the relative likelihood with which the specified range of values may be obtained. The cdf represents a cumulative total of the probability of occurrence of any value up to a specified magnitude of the random variable. For discrete random variables, the probability distribution function represents the probability of the random variable assuming any discrete value, for example, a histogram of historical measurements. For additional information about this area of probability and statistics the reader is referred to Hahn and Shapiro [21] or Benjamin and Cornell [22]. For examples of common distributional families and forms used in probabilistic human health risk analysis the reader is referred to Cullen and Frey [11].

We rarely have abundant, high quality, and relevant data for every quantity that goes into a risk model. For human health risk, inputs and models' absolute relevance would demand that the entire decision context and scenario of interest be reflected in the information used to represent inputs. This would encompass the averaging time, population demographics and characteristics, exposure scenario, season and climate, and other particulars. Obviously it is rare to have measurements that are perfectly relevant from one decision scenario to another. There are multiple means of addressing this lack of congruence between data sets and the needs of analysis. Sometimes it is possible to combine multiple data sets in order to develop a more generalizable distribution for a risk model input. At other times it is necessary to shift the first moment of an available input distribution, or to increase its variance using expert judgment, in order to reflect differences between the scenario in which the measurements were collected and the scenario or decision now at hand.

A more formal approach to combining multiple sources of information involves the use of Bayes' theorem to generate distributions, which reflect an expert analyst's belief or state of knowledge about an

uncertain quantity, or which combine the beliefs of multiple experts [23]. Aggregation of information from multiple experts is the subject of the section titled “Aggregating Individual Knowledge, Beliefs, and Preferences” in this encyclopedia. A particularly powerful application of Bayes’ theorem allows prior distributions based on subjective knowledge about model inputs to be refined or “updated” with information from other sources such as empirical measurements. In this process the relative impact of the empirical information will depend on how narrowly the prior describes the quantity, that is, how much certainty is reflected in the subjective statement of belief and certainty. With a broad prior distribution, implying a less informed analysis, even a small number of directly applicable measurements may have a huge impact on the resulting or posterior distribution. By contrast with a narrow prior distribution as the starting point, implying more certainty, one would need to mathematically incorporate many measurements demonstrating a high degree of consistency to have a noticeable impact on the shape of the posterior [24]. Further, Bayesian approaches may be applied to dealing with model uncertainty as in the context of multiple climate models mentioned earlier [6].

CORRELATION AND SIMULATION

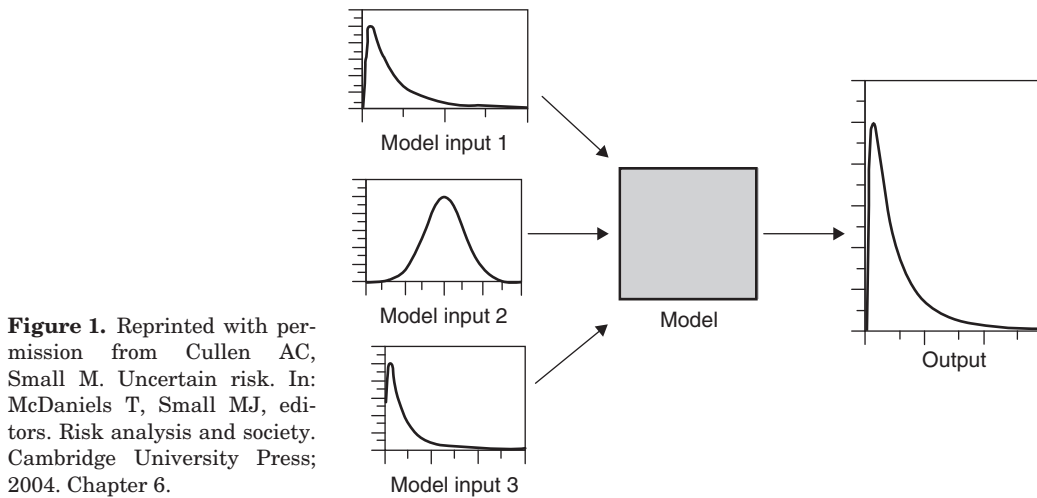
Monte Carlo simulation has been criticized on occasion for assuming independence among model inputs. It is important to note that a tool does not assume anything; all assumptions in the analysis are made by a person (whether deliberately or not). Not surprisingly model inputs that come together in a risk model are often correlated among themselves. For example, individuals whose body weights fall in the low percentiles of the human body distribution are unlikely to have high percentile rates of food consumption. Monte Carlo simulation software tools are equipped to handle correlation between inputs. But, of course, this requires that analysts be aware of correlation and able to estimate its magnitude reasonably well. Also

the correlation must have the potential to significantly affect the results, so as to justify the effort of representing it. Smith *et al.* [25] studied the potential impact of neglecting correlations and present basic screening parameters to identify cases where it is important to treat these explicitly.

PROPAGATION OF UNCERTAINTY AND VARIABILITY

Numerical simulation techniques for propagating uncertainty and variability through models whose inputs have been defined probabilistically have been in use for decades. The most common technique is Monte Carlo analysis; however, this is just one of several related approaches including Latin hypercube simulation (LHS), importance sampling, and others. In Monte Carlo risk simulation a risk model is run repeatedly, each time with one possible value substituted for each uncertain or variable input. The value of risk associated with the specified set of input values is calculated and the output stored. This process is repeated 100, 1000, 10 000, or more times, while accumulating the output values, that is, the estimates of risk [10,21] (Fig. 1). The number of iterations may be selected based on the desired properties of the output distribution—such as stability of particular percentiles of interest—or dictated by resource constraints. The name Monte Carlo stems from the chance process by which the values of the uncertain and/or variable quantities are assigned for each run of the model. As with certain gambling games the values are random draws from the probability distributions representing each quantity. For simple Monte Carlo simulation the values are drawn randomly and according to the probability associated with each via the assigned probability distribution. As mentioned above, there are decision contexts in which the separate treatment of uncertainty and variability is warranted and valuable. The result then may take the form of a probability surface, rather than a one-dimensional distribution.

Many inputs to risk models are both uncertain and variable, although usually



one dimension or the other is dominant [11]. For two-dimensional analyses probability distributions may represent uncertainty across statistical descriptions of variability, such as those based on frequencies in random samples from a population. Kaplan and Garrick [26] defined risk as a triplet composed of scenarios, probabilities, and consequences. In health risk analysis, scenarios are defined such that variability inherent in biological, chemical, physical, and social systems is fixed. Probabilities are then employed to represent uncertainty. Consequences are outcomes, for example, potential health effects, costs, and so on. In developing two-dimensional input distributions, the parameters of the distribution of probability in one of the dimensions, such as that representing variability, are themselves distributed. This implies that both the mean and the variance of the distribution of the model input are also represented by distributions, according to the associated level of uncertainty. Variability and uncertainty are then propagated separately to produce two-dimensional output, in the form of a risk surface with variability in one dimension and uncertainty in the other [27].

Monte Carlo simulation relies on a random number generator to produce a set of random draws from a uniform distribution with bounds 0 and 1. These random draws

from the uniform distribution are then transformed to represent draws from other distributional families such as normals, lognormals, betas, and gammas. Most analysts rely on available software packages to achieve these steps and generate input values for the Monte Carlo simulation, for example, @Risk or Crystal Ball, which appear as overlays on spreadsheet programs on personal computers. In contrast with standard spreadsheets that can accommodate an individual number in a cell representing an input to a model, for example, body weight in an exposure calculation, these programs allow users to specify a distribution to represent each quantity by encoding the distribution parameters in the cell, for example, the geometric mean and geometric standard deviation defining a lognormal distribution of body weight. When an analyst launches a Monte Carlo simulation the software generates one random draw from each input distribution needed for the risk model, makes a deterministic calculation of risk dictated by this set of inputs, and then saves the result as one point in the output distribution. This process is repeated over and over until the desired number of points has been generated for the output distribution.

A common alternative to Monte Carlo simulation is LHS [28]. LHS addresses the

shortcoming of random draws taken from highly skewed or highly variable input distributions for the purpose of estimating risk. In these cases, a huge number of random draws would be required to ensure with any reasonable confidence that meaningful representation from the extremes or tails of input distributions had been achieved. LHS addresses this issue by using random draws from ranges of equal probability in the input distribution, rather than relying on simple random draws. Thus, LHS ensures that values from the whole range of each input distribution will be sampled, while still requiring that these draws be made according to the probability density as assigned. This stratified sampling approach, in some cases, increases the efficiency of sampling and reduces the number of iterations required in an analysis. However, it is worth noting that LHS does not always outperform Monte Carlo analysis, for example, when input-output relationships in risk models are nonadditive and/or nonmonotonically nonlinear [29]. LHS is an available option in most probabilistic simulation packages such as @Risk, Crystal Ball, and Analytica, with more sophisticated alternatives increasingly included as routine options in software as well.

The output distribution resulting from the Monte Carlo process, based on multiple individual calculations of risk, itself often becomes the subject of further statistical treatment. Although the output is generated by a simulation exercise, it may be analyzed with the same tools and techniques that would be applied to a data set of empirically or experimentally derived values. Analysis of the output distribution allows the analyst to determine percentiles of interest, especially upper end and lower end percentiles, which may be relevant to decision making. In addition, analysis of the output distribution is quite useful for identifying significant contributors to the overall variance in risk. For example, decisions about how to define the residential duration of local anglers used in assessments of polychlorinated biphenyls (PCB) exposure via fish consumption were shown to be highly significant in studies of communities living near the Hudson

river [30]. Finally, analyses of significant contributors to overall variance may take the form of value of information analysis, which serves as a precursor to planning directions for future research [23]. Value of information analysis is discussed in detail in the section titled "Value of Information and Option Values" in this encyclopedia.

Similarly, multidimensional output is well suited to additional interpretation through statistical analysis. For example, there may be specific percentiles of interest at the high or low end of the variability dimension. High end percentiles of variability often represent highly susceptible or highly exposed subpopulations. If output is two-dimensional the results carry additional information about uncertainty regarding specific subpopulations, represented by particular percentiles of variability. A detailed application from Cullen and Frey [11] related to community exposure to PCBs near a Superfund site illustrates one strength of this approach, gauging the relative importance of uncertainty and variability in exposure estimates. In this context they find that irreducible variability associated with interhousehold differences in exposure, such as lifestyle and diet choices, dominates variance in exposure estimates, as compared to measurement error, which is a function of the science, equipment, and laboratory techniques used in the measurement of contaminant concentrations.

IDENTIFYING SIGNIFICANT CONTRIBUTORS TO OUTPUT VARIABILITY

Analysts and decision makers often share a strong interest in identifying the most significant contributors to variability in a probabilistic risk assessment, even if they do not anticipate being able to do much to narrow the overall variance in the result. Their motivation to do sensitivity analysis is sometimes linked to the desire to know which uncertainties are driving the results, and consequently to gauge one's level of comfort with that knowledge. Alternatively, the sensitivity analysis may be cast as a comparison of alternative model structures to

measure the impact of making a particular selection on the decision at hand [31,33]. Inputs may be significant because of the way in which they enter the risk model, for example, as an exponent, or because of the degree of variance they exhibit themselves. In some cases it may be of more interest to identify inputs that have a strong influence on a particular percentile of the output distribution, for example, the 95th percentile. To get a sense of the likelihood of individual inputs to drive variance in the output before carrying out a simulation, analysts can look at the relative magnitudes of input variances or coefficients of variation, where models are composed of linear combinations of inputs.

After simulation, or where linearity does not hold, there are additional options. Analysts may look at scatter plots of individual inputs against the risk output; check the sample or rank correlation coefficient between a simulated input and the simulated output; or use techniques from multivariate linear regression including partial correlation coefficients, and rank regression coefficients [10,32]. A detailed development of sensitivity analysis methods is presented in the section titled "Communicating, Interpreting, and Understanding Decision Analysis Results" in this encyclopedia.

SUMMARY

Probabilistic simulation approaches for quantitative health risk analysis allow analysts and decision makers to be explicit about how they are addressing uncertainty. They allow a full and open exploration of the impact of decision about the treatment of uncertainty and variability on estimated distributions of risk across time, space, and populations. These tools also encourage a look into the future, with the consideration of options for additional data collection and future opportunities for developing alternative approaches for reducing risk. Finally, sound probabilistic simulation opens up the possibility for communication about

the limitations of results and any associated issues of generalizability.

REFERENCES

1. Hammersley JM, Handscomb DC. Monte Carlo methods. London, New York: Methuen and Co., Ltd., John Wiley and Sons.
2. Ulam S. Adventures of a mathematician. New York: Scribners; 1976.
3. Rugen P, Callahan B. An overview of Monte Carlo: a fifty year perspective. *Hum Ecol Risk Assess* 1996;2:671–680.
4. NRC. Reactor safety study. 1975. Available at <http://en.wikipedia.org/wiki/WASH-1400>. Accessed 2010.
5. Jorion P. Value at risk: the new benchmark for managing financial risk. 3rd ed. United states: McGraw-Hill; 2006. ISBN 978–0071464956.
6. Tebaldi C, Smith R, Nychka D, *et al.* Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J Clim* 2005;18:1524–1540.
7. Shields P. Understanding population and individual risk assessment: the case of polychlorinated biphenyls. *Cancer Epidemiol Biomarkers Prev* 2006;15:830–839.
8. Bogen KT, Spear RC. Integrating uncertainty and interindividual variability in environmental risk assessment. *Risk Anal* 1987;7:427–436.
9. Bogen KT. Uncertainty in environmental health risk assessment. New York: Garland Publishing, Inc.; 1990.
10. Morgan MG, Henrion M. Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. New York: Cambridge University Press; 1990.
11. Cullen AC, Frey HC. Use of probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs. New York: Plenum Press; 1999.
12. Finkel AM. Is risk assessment really too conservative? Revising the revisionists. *Columbia J Environ Law* 1989;14:427–467.
13. Cullen AC. Measures of conservatism in probabilistic risk assessment. *Risk Anal* 1994;14:389–393.
14. U.S. Environmental Protection Agency (USEPA). Using probabilistic methods

- to enhance the role of risk analysis in decision making with case study examples. 2009. Available at <http://www.epa.gov/osa/raf/prawhitepaper/index.htm>. Accessed 2010.
15. National Research Council. Science and judgment in risk assessment. Washington (DC): National Academy Press; 1994.
 16. Hoffman FO, Hammonds JS. Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Anal* 1994;14:707–712.
 17. U.S. Environmental Protection Agency (USEPA). National center for environmental assessment. Exposure factors handbook. Washington (DC): NTIS; 2009. PB98-124217. Available at <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=20563>.
 18. Ott W. A physical explanation of the log-normality of pollutant concentrations. *J Air Waste Manage Assoc* 1990;40:1378–1383.
 19. Ott W. Environmental statistics and data analysis. Boca Raton (FL): Lewis Publishers; 1995.
 20. Industrial Economics. Expanded expert judgment assessment of the concentration-response relationship between PM_{2.5} exposure and mortality. Prepared for Office of Air Quality Planning and Standards. Research Triangle Park (NC): U.S. Environmental Protection Agency; 2006.
 21. Hahn GJ, Shapiro SS. Statistical models in engineering. New York: Wiley Classics Library, John Wiley and Sons; 1967.
 22. Benjamin JR, Cornell CA. Probability and statistics and decisions for civil engineers. New York: McGraw-Hill; 1970.
 23. Raiffa H, Schlaifer RO. Applied statistical decision theory. Cambridge (MA): Harvard University Press; 1961.
 24. Raiffa H. Decision analysis: introductory lectures on choice under uncertainty. Reading (MA): Addison Wesley; 1968.
 25. Smith AE, Ryan PB, Evans JS. The effect of neglecting correlations when propagating uncertainty and estimating the population distribution of risk. *Risk Anal* 1992;12:467–474.
 26. Kaplan S, Garrick BJ. On the quantitative definition of risk. *Risk Anal* 1981;1(1):11–27.
 27. Frey HC, Rhodes DS. Characterizing, simulating and analyzing variability and uncertainty: an illustration of methods using an air toxics emissions example. *Hum Ecol Risk Assess* 1996;2:762–797.
 28. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from computer code. *Technometrics* 1979;21(2):239–225.
 29. Homma T, Saltelli A. Sensitivity analysis of model output: performance of the Sobol'quasi random sequence generator for the integration of the modified Hora and Iman importance measure. *J Nucl Sci Technol* 1995;32:1164–1173.
 30. U.S. Environmental Protection Agency (USEPA). Executive summary human health risk assessment: upper hudson river. 1999. Available at <http://www.epa.gov/hudson/hhraexsum.htm>. Accessed 2010.
 31. Frey HC, Patil SR. Identification and review of sensitivity analysis methods. *Risk Anal* 2002;22:553–578.
 32. Iman RL, Helton JC. An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Anal* 1988;8:71–90.
 33. Cullen AC. The sensitivity of Monte Carlo simulation results to model assumptions: the case of municipal solid waste combustor risk assessment. *J Air Waste Manage Assoc* 1995;45:538–546.
 34. Cullen AC, Small M. Uncertain risk. In: McDaniels T, Small MJ, editors. Risk analysis and society. Cambridge (UK): Cambridge University Press; 2004. Chapter 6.

MULTIARMED BANDITS AND GITTINS INDEX

RICHARD WEBER
Department of Industrial
Engineering, Universidad de
Chile, Santiago, Chile
Statistical Laboratory, University
of Cambridge, Cambridge, UK

THE MULTIARMED BANDIT PROBLEM

A multiarmed bandit is a slot-machine, or fruit-machine, with multiple arms. The arms differ in the distributions of rewards that they pay when pulled. An important special instance is when arm i is a so-called *Bernoulli bandit*, with parameter p_i . Such an arm pays \$1 with probability p_i , and \$0 with probability $1 - p_i$; this happens independently each time the arm is pulled. If there are n such arms, and a gambler knows the values of $p_1, \dots, p_n$, then he maximizes his expected reward by always pulling the arm of maximum p_i . However, if he does not know these parameters, then he must choose each successive arm that he pulls on the basis of the information he has obtained from previous pulls. His aim is to maximize some measure of the sequence of rewards he obtains.

Robbins Multiarmed Bandit Problem

Suppose that under some policy, the arm that is pulled on pull t is denoted i_t , and r_{i_t} denotes the reward obtained. In a version of the multiarmed bandit problem (MABP) due to Robbins [1], we define the *regret* after T pulls, as

$$\rho = T \max\{p_1, \dots, p_n\} - \sum_{t=1}^T r_{i_t},$$

and seek a strategy such that the average regret of ρ/T tends to 0 as T tends to infinity, for all values of the unknown parameters

$p_1, \dots, p_n$. This is possible with any sensible strategy that pulls each arm infinitely often, although some such policies are better than others in terms of minimizing a secondary criterion, such as the worst-case expected regret, $\sup_{p_1, \dots, p_n} E[\rho]$.

Gittins Multiarmed Bandit Problem

This article focuses upon a second formulation of the problem in which the aim is to maximize the expected total discounted reward over the infinite horizon

$$E \left[\sum_{t=1}^{\infty} r_{i_t} \beta^t \mid W_0 \right],$$

where β is a *discount parameter*, $0 < \beta \leq 1$, and W_0 is some initial information, such as prior distributions for the unknown parameters $p_1, \dots, p_n$.

Modeling generally, we formulate the MABP as a Markov decision problem in discrete time. Each arm i has a state $x_i(t)$, which is known at time t , ($t = 0, 1, 2, \dots$). The state of the decision problem is the vector of states $x(t) = (x_1(t), \dots, x_n(t))$. If at time t the gambler pulls arm i , then he obtains a reward of $r_i(x_i(t))$. The state of arm i moves to a new state y , with probability $P_i(x_i, y)$, where x and y are members of the state space of arm i , say E_i . The states of all other arms are unchanged and no rewards are obtained from them. In choosing which of the arms to pull at time t , the decision maker knows $x(t)$, and all the data of the problem; that is, he knows, for all i , the functions $r_i(\cdot)$ and $P_i(\cdot, \cdot)$. Given a discount parameter β , he wishes to maximize the expected discounted sum of rewards. Let π denote a policy, and let i_t denote the arm that is pulled at time t under this policy. In the language of Markov decision problems, we wish to find the value function

$$V(x) = \sup_{\pi} E \left[\sum_{t=0}^{\infty} r_{i_t}(x_{i_t}(t)) \beta^t \mid x(0) = x \right],$$

where the supremum is taken over all policies π that are realizable (or nonanticipatory), in the sense that i_t depends only on the problem data and $x(t)$, not on any information which can be known only after time t .

Setup in this way, the MABP is an infinite-horizon discounted-reward Markov decision problem. It therefore has a deterministic, stationary, Markov optimal policy. The natural way to address it is by its stochastic dynamic programming equation of

$$V(x) = \max_{i: i \in \{1, \dots, n\}} \left\{ r_i(x) + \beta \sum_{y \in E_i} P_i(x_i, y) \times V(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) \right\}. \quad (1)$$

In essence then, the MABP is about controlling the evolution of n independent reward-producing Markov processes. At each instant (in discrete time) exactly one of the processes is *continued* (and reward from it is obtained), while all the other processes are *frozen*. The continued process can change state, but the frozen processes do not change state. We suppose that reward is produced only from the continued process (although this assumption may be relaxed, as we see later.)

With respect to its usefulness in solving the MABP, Equation (1) suffers from the *curse of dimensionality*, which is typical in dynamic programming. Suppose, for example, that the state space of arm i is finite and of size $k_i = |E_i|$, $k_i \geq 2$. In principle, a solution to Equation (1) could be found by the policy improvement algorithm, or by solving a linear program. However, the vector state $x = (x_1, \dots, x_n)$ evolves on the state space $E = E_1 \times \dots \times E_n$, of size $k = |E| = \prod_i k_i$. So the number of stationary policies is n^k , where $k > 2^n$; also, the number of variables in a linear program would be $O(2^n)$. So the solution of Equation (1) is computationally infeasible even for moderate-sized k and n .

THE GITTINS INDEX

Remarkably, the MABP that has been described above can be solved by an *index policy*. That is, we can compute a number

(called an *index*), separately for each arm, such that the optimal policy is always to pull the arm having the currently greatest index. As an introduction, we illustrate the idea of an index policy with an example from job scheduling.

Single Machine Scheduling. Suppose that n jobs are to be processed successively on one machine. Job i has a known processing times t_i , assumed to be a positive integer. On completion of job i , a positive reward r_i is obtained. If job 1 is processed first, and job 2 is processed immediately after it, then the sum of discounted rewards obtained from these two jobs is $r_1\beta^{t_1} + r_2\beta^{t_1+t_2}$. If the processing order of the jobs is interchanged, we obtain $r_2\beta^{t_2} + r_1\beta^{t_1+t_2}$. A little algebra shows that the first ordering has the greater reward if $r_1\beta^{t_1}/(1 - \beta^{t_1}) > r_2\beta^{t_2}/(1 - \beta^{t_2})$. Using this idea, it is not hard to see that the total discounted reward obtained from the n jobs is maximized by processing them in decreasing order of indices, computed as $v_i = r_i\beta^{t_i}/(1 - \beta^{t_i})$.

The appropriate index for the MABP is in the same spirit, but more complicated. The key result for the MABP is the following.

Gittins Index Theorem. *The multiarmed bandit problem, as setup above, is solved by always pulling the arm having the greatest Gittins index, which is defined for arm i as*

$$G_i(x_i) = \sup_{\tau} \frac{E \left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) \mid x_i(0) = x_i \right]}{E \left[\sum_{t=0}^{\tau-1} \beta^t \mid x_i(0) = x_i \right]}, \quad (2)$$

where τ is a stopping time constrained to take a value in the set $\{1, 2, \dots\}$.

By a “stopping time” τ we mean a time that can be recognized when it occurs. In fact, it can be shown that τ attains the supremum when $\tau = \min\{t : G_i(x_i(t)) \leq G_i(x_i(0)), \tau > 0\}$; that is, τ is the first time at which the process reaches a state in which the Gittins index is no greater than it was initially.

Examining Equation (2), we see that the Gittins index is the maximal possible

quotient of a numerator that is “expected total discounted *reward* over τ steps,” and denominator that is “expected total discounted *time* over τ steps,” where τ is at least one step. Notice that the Gittins index can be computed for all states of arm i as a function only of the data $r_i(\cdot)$ and $P_i(\cdot, \cdot)$. That is, it can be computed without knowing anything about the other arms. The curse of dimensionality disappears. If arm i moves on a finite state space of size k_i , then its Gittins indices (one for each of the k_i states) can be computed in time $O(k_i^3)$ [2,3].

In the job scheduling example just considered, the optimal stopping time on the right-hand side of Equation (2) is $\tau = t_i$, the numerator is $r_i \beta^{t_i}$ and the denominator is $1 + \beta + \dots + \beta^{t_i-1} = (1 - \beta^{t_i})/(1 - \beta)$. Thus, $G_i = r_i \beta^{t_i} (1 - \beta)/(1 - \beta^{t_i}) = (1 - \beta) v_i$.

The index theorem above is due to Gittins and Jones, who had obtained it by 1970, and presented it in 1972 [4]. They called $G_i(x_i)$ a *dynamic allocation index*, but it is now widely known as the Gittins index. The result became better known with Gittins’ paper of 1979 [5]. The idea that an optimal policy might have an index characterization was not new, since it had been known for some very simple scheduling problems since the 1950s [6]. In the 1970s, a deeper understanding of the optimality of index policies for more complicated problems was advancing rapidly. Researchers such as Klimov [7], Sevcik [8], Harrison [9], Tcha and Pliska [10], and Meilijson and Weiss [11] were proving results about the optimality of index policies for scheduling problems in various forms of multiclass queues, and finding indices similar to Gittins indices.

Nonetheless, the solution of the MABP in the general setting that has been described above impressed many experts as surprising and beautiful. In the forward to Gittins’ book [12], Whittle describes a colleague of high repute, asking another colleague “What would you say if you were told that the multi-armed bandit problem had been solved?” The reply was “Sir, the multi-armed bandit problem is not of such a nature that it can be solved.”

Subsequent to Gittins’ pioneering work, many researchers have sought to reprove

the index theorem by different means, to exploit it and to understand how it might be generalized. Today, the multiarmed bandit model and Gittins indices are widely used by researchers in computer science, economics, engineering, operations research, and statistics. Let us make some brief notes on other characterizations of the Gittins index and its associated policy.

Calibration. An alternative characterization of $G_i(x_i)$ is

$$G_i(x_i) = \sup \left\{ g : \frac{g}{1 - \beta} \leq \sup_{\tau > 0} E \left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) + \beta^\tau \frac{g}{1 - \beta} \mid x_i(0) = x_i \right] \right\}.$$

That is, we may consider a two-armed bandit problem that is posed for arm i and one other *calibrating* arm which pays out a known reward g at each pull. The Gittins index of arm i is the value of g for which it is optimal to pull either arm initially. Notice that once we decide to switch from pulling arm i to pulling the known arm, at time τ , then information about arm i does not change and so it would be optimal to stick with pulling the known arm ever after.

Forwards Induction. If we put $\tau = 1$ on the right-hand side of Equation (2), then it evaluates to $E r_i(x_i(t))$. If we use this as an index for choosing between projects, we have what is called a *myopic* or *one-step-look-ahead* policy. We may view the Gittins index policy as generalizing the idea of a one-step-look-ahead policy, since it looks-ahead by some optimal time τ , so as to maximize, on the right-hand side of Equation (2), a measure of the rate at which reward can be accrued. This defines a so-called *forwards induction policy*. Similarly, one can define a forwards induction policy for any Markov decision process. Either the one-step-look-ahead-policy or a forwards induction policy needs to be optimal, but they can provide guidance about heuristics.

Proofs of Optimality

Alternative proofs of the Gittins index theorem have helped to reveal its beauty and initiate further developments; for example, for heuristic approaches to problems that do not quite fit the assumptions of the MABP. We mention four proof methodologies: (i) interchange arguments [2,4,13–15], (ii) a linear program posed over an achievable region for a performance measure [16], (iii) an explicit formula for the value function that can be shown to satisfy the dynamic programming equation [17], and (iv) an intuitive method based on the notion of a “fair charge” [18]. We sketch the main ideas of (i)–(iii), and essentially the full proof of (iv).

The idea of an “interchange argument,” mentioned as (i), has already been illustrated in the single machine scheduling problem described above. However, great care is needed to make an interchange argument work for the MABP. In particular, one has to find a clever way to interchange τ_1 successive pulls of arm 1 followed with τ_2 successive pulls of arm 2, choosing τ_1 and τ_2 to be special stopping times.

Method (ii) recasts the MABP as a linear program. One defines z_{ij} as the expected total discounted total time spent pulling arm i on occasions that it is in state j . The MABP can be formulated as a linear program in variables z_{ij} , with objective function of $\sum_i \sum_{j \in E_i} r_i(j) z_{ij}$, and subject to certain constraints on the z_{ij} . For instance, since one arm must be pulled at each time, one of these constraints is $\sum_{i=1}^n \sum_{j \in E_i} z_{ij} = 1 + \beta + \beta^2 + \dots = 1/(1 - \beta)$. It turns out that the resulting special linear program can be solved by computing its optimal dual variables by way of a greedy algorithm, which also calculates the Gittins index values. This proof method resonates with work of Klimov [7], in which essentially the same greedy algorithm is used to compute indices for an index policy. The greedy algorithm works because the achievable region is an *extended polymatroid* [16].

In method (iii) the standard setup is augmented by giving the decision maker a *resignation option*, such that he may any time choose to take a lump sum M , and then pull

no further arms. Let $V(x, M)$ be the value function in this augmented problem when $x = (x_1, \dots, x_n)$ is the initial vector of states of the arms. Suppose that there is an upper bound R on the value of the total discounted reward, and M is such that resignation in a finite time is certain. By demonstrating that it solves the relevant dynamic programming equation, Whittle [17] showed that $V(x, M)$ can be expressed as

$$V(x, M) = R - \int_M^R \prod_i \frac{\partial \phi_i(x_i, m)}{\partial m},$$

where $\phi_i(x_i, m)$ is the maximal expected discounted reward in the one-armed bandit problem of arm i on its own, initially in state x_i , and augmented with the resignation option.

Finally, we turn to (iv) and present a proof in the following few sentences. Consider a one-armed bandit problem in which only arm i is available. Let us define the “fair charge,” $\phi_i(x_i)$, as the maximum amount that a gambler would be willing to pay for each further pull of arm i in order to be permitted to pull it at least one more time, but then being free to quit whenever he likes after the first pull. We next define the “prevailing charge” for arm i at time t as $g_i(t) = \min_{s \leq t} \phi_i(x_i(s))$. So $g_i(t)$ actually depends on $x_i(0), \dots, x_i(t)$ (which we omit from its argument for convenience). Note that $g_i(t)$ is a nonincreasing function of t and its value depends only on the states through which arm i evolves. The proof of the index theorem is completed by verifying the following facts, each of which is almost immediate. (i) If, in the MABP with n available arms, the gambler not only collects rewards but also must pay the prevailing charge of each arm as he pulls it, then he cannot do better than break even (in terms of expected value). (ii) He maximizes the expected discounted sum of the prevailing charges that he pays by always playing the arm with the greatest prevailing charge. (iii) Using this strategy he also breaks even; so this strategy, of pulling an arm having greatest $g_i(x_i)$, maximizes the expected discounted sum of the rewards he can obtain from the arms. By simple algebra, one can check that $g_i(x_i) = G_i(x_i)$.

Calculation of the Gittins Index

We now turn to the problem of computing the Gittins index value for each possible state of a given arm i . The input to this calculation is the data for arm i of $r_i(\cdot)$ and $P_i(\cdot, \cdot)$. If the state space of an arm i is finite, say $E_i = \{1, \dots, k_i\}$, then the Gittins indices can be computed in an iterative fashion. First, we find the state of greatest index, say 1 such that $1 = \arg \max_j r_i(j)$. Having found this state, we can next find the state of second-greatest index. If this is state j , then $G_i(j)$ is computed in Equation (2) by taking τ to be the first time that the state is not 1. This means that the second-best state is the state j , which maximizes

$$\frac{E[r_i(j) + \beta r_i(1) + \dots + \beta^{\tau-1} r_i(1)]}{E[1 + \beta + \dots + \beta^{\tau-1}]},$$

where τ is the time at which, having started at $x_i(0) = j$, we have $x_i(\tau) \neq 1$. One can continue in this manner, successively finding states and their Gittins indices, in decreasing order of their indices [2,15]. This method is foreshadowed in the way Klimov calculated indices for his problem [7], and the greedy algorithm in Bertsimas and Niño Mora [16] is similar. For further ways of calculating indices see also Niño Mora [3], Katehakis and Veinott [19], and Sonin [20].

If the state space of an arm is infinite, as in the case of the Bernoulli bandit, there may be no finite calculation by which to determine the Gittins indices for all states. In this circumstance, we can approximate the Gittins index using something like the value iteration algorithm. Essentially, one solves a problem of maximizing right-hand side of Equation (2), subject to $\tau \leq N$, where N is large.

APPLICATIONS

Clinical Trials

The MABP has its history in a model of clinical trials in which a decision maker is trying to discover which of n drugs has the greatest probability of providing successful treatment. As explained in the beginning of this article, there is an unknown

probability p_i with which drug i successfully treats a patient, where p_i has some known prior distribution. The posterior distribution of p_i is updated as independent observations of successes and failures of the drug are observed. For example, if the prior distribution of p_i is uniform on $[0, 1]$ and s successes and f failures are observed, then the posterior distribution of p_i is a beta distribution, with density proportional to $p^s(1-p)^f$, $0 \leq p \leq 1$.

Patients are treated one at a time. The decision maker knows the number of successes and failures that has been observed for each drug so far. He must choose a drug with which to treat the next patient. In the short term, if he cared only about the next patient, he would use the drug that has the greatest success probability. This is one for which the posterior mean is greatest, this being $(s_i + 1)/(s_i + f_i + 2)$. This is the *myopic policy*, which places all emphasis on exploitation. However, an optimal policy must also pay attention to exploration.

It is instructive to look at some numbers. Suppose that in 8 treatments with drug i we have seen 4 successes and 4 failures, and that in 40 treatments with drug j we have seen 20 successes and 20 failures. Then the posterior means of p_i and p_j are both 0.5. From the tables in Gittins [12] given for $\beta = 0.8$, we read that the Gittins indices for drugs i and j are 0.5431 and 0.5114, respectively. This means that if we were faced with a two-armed bandit problem consisting of just drugs i and j , then we would strictly prefer the next treatment to be with drug i , even though the posterior means of p_i and p_j are the same. This is because there is more to be learned by experimentation with drug i than with j . Similarly, if we are faced with a two-armed bandit problem consisting of drug i and one other drug k , which has known success probability p_k , then we should strictly prefer to make the next treatment with drug i if $p_k < 0.5431$. The fact that the Gittins index exceeds 0.5, reflects the fact that there is potential to learn that $p_i > 0.5$. If $\beta = 0.9$ then the Gittins index for drug i is 0.5676. This is greater than its value for $\beta = 0.8$, since future rewards count more as β becomes closer to 1.

As we have already noted, the fact that the state space of a Bernoulli bandit is infinite means that there is no method of calculating the Gittins indices exactly. The tables in Gittins [12] were produced by a method of approximation in which the stopping time τ in Equation (2) was constrained to be no more than N , with $N = 300$ being found sufficient to estimate the indices to four significant figures.

It is worth remarking that an elementary version of this problem was solved by Bellman [21], for the special case in which there are just two drugs and the success probability of one of them is known, and by Bradt *et al.* [22] for undiscounted rewards and a finite horizon. These authors elucidated indices characterizing optimal policies (but without actually calling them indices). They presaged the idea, described above, in which an arm with unknown p_i is calibrated against an arm with known p_k such that one is indifferent (in a two-armed bandit problem) as regards initially pulling arm i or k .

Target Processes

The Gittins index can also be used to solve a problem in which the objective is to minimize costs up until the first time that one of the arms pays out a jackpot. An example is the game of “golf with n balls,” described in Dumitriu *et al.* [23]. Imagine a golfer is playing with n balls, each of which rests in a different location. The golfer is to play one ball at a time (incurring cost as he does so, and with a known probability distributions determining how balls move). His objective is to minimize the expected cost incurred until he first puts one of the balls into the hole. We can make this a standard MABP by imagining that once a ball goes in the hole it is in a state that it can be played with 0 cost thereafter. So the Gittins index theory can answer the question: Which ball should the golfer hit next?

More practically, a company might be engaged in pharmaceutical research and wishing to discover a drug that achieves a target level of effectiveness, T . Suppose that there are n families of similar compounds. On successive trials of compounds from family i , the observed effectiveness is

$X_i(1), X_i(2), \dots$, where these are independent and identically distributed according to an exponential distribution with parameter θ_i . The value of θ_i is unknown, but *a priori* it is exponentially distributed with known mean μ_i . It is desired to minimize the expected number of trials until first obtaining a reward that exceeds T . Thus, after having made a first observation $x_i = X_i(1)$, the posterior distribution of θ_i is proportional to $x_i \exp(-x_i \theta_i) \mu_i \exp(-\mu_i \theta_i)$. See Gittins [12] for more details, and tables of index values for such problems.

Job Scheduling

The theory of MABP and the Gittins index theorem for discrete-time Markov reward processes can be extended to bandits which evolve in continuous-time as semi-Markov processes. A continuous-time setting is more appropriate when we wish to model the processing of jobs.

Suppose we are given n jobs, and that job i has an unknown processing time, which may be modeled as a random variable with distribution function F_i and density f_i . The only reward that can be obtained from a job i is obtained upon its completion, and this is r_i . We wish to maximize the expected sum of discounted rewards obtained by processing the jobs on a single machine. It turns out that if job i has been processed for time x_i and has not yet completed, then its Gittins index at that point is

$$G_i(x_i) = \sup_{\tau > 0} \frac{r_i \int_0^\tau e^{-\gamma t} f_i(x_i + t) dt}{\int_0^\tau e^{-\gamma t} [1 - F_i(x_i + t)] dt}, \quad (3)$$

where $e^{-\gamma t}$ ($\gamma > 0$) plays an analogous role to β^t .

In some special cases, Equation (3) is easy to evaluate. The *completion rate* of job i (which is also called the hazard rate or failure rate) is defined as $\rho_i(x) = f_i(x)/(1 - F_i(x))$. This has the interpretation that if having been processed for time x the job is not yet complete, then $\rho(x)\delta + o(\delta)$ is the probability that it does complete when given a further small amount of processing δ . If the processing times of all jobs have

completion rates that are decreasing in x , then $G_i(x_i) = r_i \rho_i(x_i)$, so it is optimal to always process the job of greatest value of $r_i \rho_i(x_i)$, where x_i is the amount of processing job i has received thus far. Similarly, if all jobs have increasing completion rates, then the optimal τ in Equation (3) is ∞ . In this case, as $\gamma \rightarrow 0$, we find $G_i(x_i) \rightarrow r_i/E(\text{remaining processing time} | x_i)$.

Tax Problems

In a variation of the job scheduling problem, one is charged a tax (or holding cost) c_i , so long as job i is not complete. If the job were never to be completed, then the total discounted cost so paid would be $\int_0^\infty c_i e^{-\gamma t} dt = c_i/\gamma$. So an alternative way to compute the true expected total holding cost is to imagine that we must pay c_i/γ at time 0 for every job, but then receive a “refund” as reward $r_i = c_i/\gamma$ at the time when job i completes. Thus, tax problems can be solved as MABPs. By a similar construction, one can generalize the standard MABP by permitting state-dependent rewards to be obtained from frozen arms. See the work of Varaiya *et al.* [15], who also describe an algorithm for computing indices, similar to Tsitsiklis [2].

If $\mu_i = 1/E(\text{remaining processing time} | x_i = 0)$ (as is the case when job i has a processing time that is exponentially distributed with parameter μ_i) then as $\gamma \rightarrow 0$, we find $G_i(x_i) \rightarrow c_i \mu_i$, the famous so-called ‘ $c\mu$ rule’. Cost is minimized by allocating the processing to the job of greatest value of $c_i \mu_i$.

Much more could be said about how the theory of the Gittins index can be used to solve scheduling and stochastic scheduling problems. It can be used to prove that a $c\mu$ rule provides optimal for scheduling in a multiclass $M/G/1$ queue. An idea of “arm-acquiring bandits” can be used to give a solution to Klimov’s problem about scheduling in such a queue when there is also Markovian feedback [7]. The theory is also useful in discussing stochastic scheduling problems in which there are precedence constraints between jobs, and in evaluating the suboptimality of index policies with switching costs [24,25]. For more on scheduling and stochastic scheduling problems see Pinedo [26].

GENERALIZATIONS

The Gittins index provides the solution to the MABP, under the modeling assumptions posed for it above. However, an index policy is not optimal, in general, if any of the following are true: (i) the time horizon is finite rather than infinite, (ii) the discount rate depends on the arm that is pulled, or (iii) more than one arm can be pulled at the same time.

There is one interesting example of a stochastic job scheduling problem in which an index policy is optimal in the case of (iii). Suppose n jobs have unknown processing times which are *a priori* distributed independently as exponential random variables with known parameters $\lambda_1, \dots, \lambda_n$. The jobs are to be processed using $m (\geq 2)$ identical machines operating in parallel. Then the expected flow sum of completion times (called the flow time) is minimized by giving priority to short jobs (largest λ_i), and the expected completion time of the last job (called the makespan) is minimized by giving priority to long jobs (smallest λ_i). This can be generalized to some other distributions [27]. However, such a result is exceptional. Most problems that include aspects of (i), (ii), or (iii) are NP-hard.

Restless Bandits

A restless bandit is a generalization of a one-armed bandit, in which we relax the assumption that the state of an arm is frozen when it is not pulled. We suppose that there are two controls, called “active” and “passive,” which can be applied to each arm. The rewards and state evolution differ under the active and passive controls. This model has proved popular in a number of application areas, where typically active/passive controls can correspond to running a system fast/slow, or monitored/unmonitored, or to workers being in states of laboring/resting.

The restless bandits model is due to Whittle [28], who suggested a heuristic index policy for the multiarmed restless bandit problem that can be described as follows. Consider a one-armed restless bandit and imagine that we provide a subsidy λ each time that a bandit in the passive state is pulled. Taking our objective to be the maximization of time-average reward obtained

from this one restless bandit, it is reasonable to imagine that as λ increases from $-\infty$ to ∞ the set of states in which it is optimal to take the passive action increases monotonically, from the empty set to the set of all states.

If this monotonicity in λ holds for bandit i (said to be indexable), then in state x_i , we can define the index $W_i(x_i)$ as the value of λ such that one is indifferent between applying the passive and active controls in state x_i . As with the Gittins index, this “Whittle index” for bandit i is a function only of the data for bandit i . If the objective is to maximize time-average reward under a constraint that the active control be applied to exactly m ($< n$) of the arms at each instant, then a heuristic is to apply the active control to the m bandits of greatest $W_i(x_i)$. In the case that $m = 1$ and the passive control is the same as freezing, then this is the same as the Gittins index policy. If $m > 1$ and the passive action is not freezing, then the Whittle index policy may not be optimal, but many researchers find that the heuristic works well [29].

REFERENCES

1. Robbins H. Some aspects of the sequential design of experiments. *Bull Am Math Soc* 1952;58(5):527–535.
2. Tsitsiklis JN. A short proof of the Gittins index theorem. *Adv Appl Probab* 1994;4(1):194–199.
3. Niño Mora J. A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. *INFORMS J Comput* 2007;19:596–606.
4. Gittins JC, Jones DM. A dynamic allocation index for the sequential design of experiments. In: Gani J, editor. *Progress in statistics*. Amsterdam: North-Holland Publishing Co.; 1974. pp. 241–266. (Read at the 1972 European Meeting of Statisticians, Budapest.)
5. Gittins JC. Bandit processes and dynamic allocation indices (with discussion). *J R Stat Soc B* 1979;41:148–177.
6. Smith WE. Various optimizers for single-stage production. *Nav Res Log Q* 1956;1–2:59–63.
7. Klimov GP. Time-sharing service systems I. *Theory Probab Appl* 1974;19:532–551.
8. Sevcik KC. Scheduling for minimum total loss using service time distributions. *J ACM* 1974;21(1):66–75.
9. Harrison JM. Dynamic scheduling of a multiclass queue: discount optimality. *Oper Res* 1975;23(2):270–282.
10. Tcha DW, Pliska SR. Optimal control of single-server queuing networks and multi-class $M/G/1$ queues with feedback. *Oper Res* 1977;25(2):248–258.
11. Meilijson I, Weiss G. Multiple feedback at a single-server station. *Stoc Proc Appl* 1977;5(2):195–205.
12. Gittins JC. *Multi-armed bandit allocation indices*. New York: John Wiley & Sons, Inc.; 1989.
13. Glazebrook KD. Stoppable families of alternative bandit processes. *J Appl Probab* 1979;16:843–854.
14. Nash P. A generalized bandit problem. *J R Stat Soc B* 1980;42(2):165–169.
15. Varaiya P, Walrand J, Buyukkoc C. Extensions of the multiarmed bandit problem: the discounted case. *IEEE Trans Automat Control* 1985;AC-30:426–439.
16. Bertsimas D, Niño Mora J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math Oper Res* 1996;21(2):257–306.
17. Whittle P. Multi-armed bandits and the Gittins index. *J R Stat Soc B* 1980;42:143–149.
18. Weber RR. On the Gittins index for multiarmed bandits. *Ann Appl Probab* 1992;2:1024–1–33.
19. Katehakis MN, Veinott AF. The multi-armed bandit problem: decomposition and computation. *Math Oper Res* 1987;12:262–268.
20. Sonin IM. A generalized Gittins index for a Markov chain and its recursive calculation. *Stat Probab Lett* 2008;78: 1526–1533.
21. Bellman RE. A problem in the sequential design of experiments. *Sankhya Ser A* 1956;30:221–252.
22. Bratt RN, Johnson SM, Karlin S. On sequential designs for maximizing the sum of n observations. *Ann Math Stat* 1956;27(4):1060–1074.
23. Dumitriu I, Tetali P, Winkler P. On playing golf with two balls. *SIAM J Discrete Math* 2003;16(4):604–615.
24. Glazebrook KD. Stochastic scheduling and forwards induction. *Discrete Appl Math* 1995;57(2–3):145–165.

25. Glazebrook KD, Minty R. A generalized Gittins index for a class of multiarmed bandits with general resource requirements. *Math Oper Res* 2009;34(1):26–44.
26. Pinedo ML. *Scheduling: theory, algorithms and systems*. 3rd ed. Springer; New York 2008.
27. Weber RR. Scheduling jobs with stochastic processing requirement on parallel machines to minimise makespan or flowtime. *J Appl Probab* 1982;19:167–182.
28. Whittle P. Restless bandits: activity allocation in a changing world. In: Gani J, editor. *A celebration of applied probability*, Applied probability special volume 25A. Applied Probability Trust; Sheffield (UK) 1988. pp. 287–298.
29. Weber RR, Weiss G. On an index policy for restless bandits. *J Appl Probab* 1990;27:637–648.

MULTICLASS QUEUEING NETWORK MODELS

JOHN J. HASENBEIN
Department of Mechanical
Engineering, University of Texas
at Austin, Austin, Texas

INTRODUCTION

A multiclass queueing model is a network with a number of stations, each of which process multiple classes of customers. These networks are a generalization of the classic queueing network models of Jackson and Kelly, among others. In this section, we give an informal description of these models and a more formal mathematical description is presented in the next section. To begin, suppose the network consists of J processing stations. Each station could have multiple identical servers, but for simplicity, we focus on the case in which each station consists of a single server with an unlimited buffer for customers waiting to be processed at the station. Customers in the network are partitioned into K classes, with each class being associated with a fixed station. Customers within a class are homogeneous in terms of their processing requirements and routing requirements, after service. In particular, the service requirements for customers at a station are determined only by the class of the customer. A simple special case is when the service times for customers of class ℓ form a sequence of i.i.d. random variables. Furthermore, although customers can be routed in a random manner after service at a station, all customers of class ℓ are routed using the same mechanism after they complete their processing at the station associated with that class. We assume that customers of type ℓ arrive exogenously according to some counting process, are routed through the network for processing, and then depart the network. The discussion focuses on the case when

all customers depart eventually with probability 1 (w.p. 1). Such networks are called *open*. When $J = K$ the network is a single-class network, that is, all customers waiting for service at a given station are homogeneous. When $K > J$ at least one station serves more than one class of customers, thus giving rise to the term *multiclass*. In summary, we describe here *open multiclass queueing networks* (OMQNs) with single-server stations. Generalizations to this model are mentioned in the section titled “Further Topics”, but the most important features of this class of networks are contained in the model described in this article.

Perhaps, the best way to see the utility of OMQN models is to compare them to the classic open Jackson network described in the article ***Jackson Networks (Open and Closed)***. The OMQN extends Jackson’s model in a number of ways. First, in a Jackson network, service and interarrival distributions are exponential. In an OMQN, this restriction is removed. Second, in a Jackson network, all customers at a station have homogeneous service and routing requirements, whereas this need not be true in an OMQN.

In a Jackson network, the routing requirement is perhaps more restrictive than it initially appears. To see this, consider a simple network with two stations, A and B. Every customer needs to undergo service at station A, then station B, and the station A, after which it departs the system. Any network with this routing mechanism is not a Jackson network because the customers at station A are not homogeneous with respect to their routing. In particular, after service at station A, some customers will go to station B and some will depart the system, depending on whether or not the customer has undergone an earlier service at station A. Therefore, many relatively simple networks with exponential service and interarrival times are outside of the Jackson network paradigm, simply because of the routing requirements. The OMQN model is

useful because it allows for quite general, complex routing configurations.

Multiclass queueing network models have applications in manufacturing, telecommunications, supply chain, and service systems. There has also been a significant amount of theoretical work on these models, starting mostly in the late 1980s. Although many papers contributed to the overall development of these models, the first widely accepted canonical description of these models appeared in a paper by Harrison [1]. In the 1990s, researchers initiated several significant research streams devoted to analyzing OMQNs. One major stream has been devoted to studying the stability behavior, especially via the fluid model. A second stream has investigated heavy traffic, or diffusion approximations. Finally, much work has been devoted to finding optimal or near optimal scheduling policies (see **Conservation Laws and Related Applications**). Since several other articles cover specific topics in analysis and control of OMQNs, this article is primarily restricted to introducing the modeling framework.

FORMAL MODEL DESCRIPTION

In this section, we provide the formal mathematical description of an OMQN. Our notation is consistent with the primary references in the literature, which include Refs 2–4. An OMQN consists of J single-server stations and K classes of customers, where $K \geq J$. Stations are labeled $j = 1, \dots, J$ and classes are labeled $k = 1, \dots, K$. Each class k is associated with a unique station $s(k)$. The set of classes served at station j is denoted by $C(j)$.

We first describe three primitive cumulative processes. The number of customers that arrive from outside the network to class ℓ up to time t is given by $E_\ell(t)$. For each integer $n > 0$, $V_\ell(n)$ is the total service requirement for the first n class ℓ customers. This service requirement is given in units of time. Next, for each integer $n > 0$, $\Phi_\ell^k(n)$ counts the total number of class k jobs that go to class ℓ , among the first n class k jobs that complete service at station $s(k)$. We let $\Phi^k(n)$ be the associated K -dimensional vector. These three

processes are often referred to as the *primitive cumulatives*. The primitive cumulatives embody all of the randomness in the queueing dynamics and the other performance processes of the network can be derived (in a deterministic manner) from these processes.

In most modeling contexts, it is useful to assume that the processes above have properties that are related to the strong law of large numbers. In particular, we assume that w.p. 1,

$$\lim_{t \rightarrow \infty} \frac{E_k(t)}{t} = \alpha_k, \quad (1)$$

$$\lim_{n \rightarrow \infty} \frac{V_k(n)}{n} = m_k, \quad (2)$$

$$\lim_{n \rightarrow \infty} \frac{\Phi_\ell^k(n)}{n} = P_{k,\ell}, \quad (3)$$

for $k, \ell \in \{1, \dots, K\}$. For example, if inter-arrival times of class ℓ customers are i.i.d. random variables with mean $1/\alpha_k$, then Equation (1) holds w.p. 1 by the strong law of large numbers for renewal processes. Then, one interprets α_ℓ as the long-run arrival rate for class ℓ customers. Similarly, if processing times for customers of type ℓ are i.i.d. random variables with mean m_ℓ , then Equation (2) holds w.p. 1 by the usual strong law of large numbers. Finally, if the network uses *Bernoulli routing*, then Equation (3) also holds w.p. 1 by the strong law of large numbers. However, it should be noted that the assumptions made in Equations (1–3) allow for much more general arrival, service, and routing processes.

The K -dimensional column vectors of arrival rates and mean service times are given by $\alpha = (\alpha_1, \dots, \alpha_K)'$ and $m = (m_1, \dots, m_K)'$, respectively. We also define $\mu_\ell = 1/m_\ell$ to be the service rate for class ℓ customers. The $K \times K$ matrix $P = [P_{k,\ell}]$ is called the *routing matrix* for the network. Note that we allow customers to be routed back to the same class at which they just completed service. After a customer of class k completes service, the customer is either routed to a station or leaves the network. In the case of *Bernoulli routing*, we interpret

$P_{k,\ell}$ as the probability that a class k customer becomes a class ℓ customer (at station $s(\ell)$) after service. The probability the customer leaves the network is then

$$1 - \sum_{\ell=1}^K P_{k,\ell}.$$

Recall that we assume that the network is *open*, which means that all customers leave the network w.p. 1. This is equivalent to assuming that

$$R := I + P' + (P')^2 + \dots$$

is finite, which is also equivalent to assuming that $(I - P')$ is invertible. In this case, we have $R = (I - P')^{-1}$ (Lemma 2.1, Chapter 6 in Ref. 5). Finally, in the sections below, we let $Q_k(t)$ denote the number of class k customers present in the network at time t . Such customers may either be waiting in line or in service at station $s(k)$.

Special Cases

Certain subclasses of OMQNs have been studied in the literature. We have already mentioned the case when $J = K$, in which the network is single-class network or *generalized Jackson network*. If, in addition to having $J = K$, all service and interarrival distributions are exponential, and the routing is Bernoulli, then the network is a classical Jackson network [6,7]. *Kelly networks* are also special cases of OMQNs, see the related encyclopedia article for more details.

Some other special cases rely in a more detailed way on the structure of the routing matrix P . The first definition is adapted from Chen and Yao [8], which studies networks with an acyclic transfer mechanism.

Let

$$\mathcal{I}(k) \equiv \{i = 1, \dots, K : P_{i,\ell_1} P_{\ell_1,\ell_2} \dots P_{\ell_n,k} > 0 \\ \text{for some } \ell_1, \dots, \ell_n\}$$

that is, $\mathcal{I}(k)$ is the set of classes from which class k is reachable. An OMQN is said to be an *acyclic transfer mechanism network* (ACTN) if $k \notin \mathcal{I}(k)$ for all classes $k = 1, \dots, K$. Essentially, customers in an ACTN cannot

pass through any given buffer more than once.

An OMQN is a *multitype network* (MTN) if

- $P_{\ell,k}$ is 0 or 1 for every $\ell, k \in \{1, \dots, K\}$.
- For every class k , there exists at most one class ℓ , such that $P_{\ell,k} > 0$.
- For every class k , $\alpha_k > 0$ if there are no classes ℓ , such that $P_{\ell,k} > 0$.

In other words, an MTN can be thought of a network which processes a number of products (or customer types) each of which traverses a deterministic route through the network. An MTN with only one type is called a *reentrant line*. The classes in a reentrant line can be numbered such that $P_{k,\ell} = 1$ if $\ell = k + 1$ and for which $P_{k,\ell} = 0$ otherwise. In other words, all customers follow one deterministic route through the stations. See the article for some examples of results which hold for these special cases.

SCHEDULING POLICIES

The aspect of control of OMQNs is discussed in this section. In particular, what has not yet been described is the order of service provided to customers at each station. A rule which determines this order of service at all times is called a *scheduling policy*. In other contexts, such rules may be called *service disciplines* or *dispatch policies*. Scheduling policies may be *idling* or *nonidling*. When the network is operating under a nonidling policy, each station must devote its full capacity to processing customers, whenever customers are present at the station. In most cases, it is more efficient to use a nonidling discipline, but this depends highly on the structure of the network and the performance goals. For simplicity of exposition, we restrict ourselves to nonidling disciplines here.

In general, scheduling policies may be of the processing sharing type, that is, the server at each station is allowed to simultaneous process several jobs, sharing its total capacity in some manner among the jobs. For mathematical tractability, much of the work in OMQNs is restricted to the case in which policies are *head-of-the-line* (HL).

Under an HL policies, at most one job from each class in $C(j)$ can be in process simultaneously at station j . Note that “in process” also refers to jobs that have been partially processed and are then preempted temporarily by other jobs. This restriction is important in both fluid and diffusion approximations to OMQNs. Under non-HL policies, the state descriptor for the network process is more complex, often requiring a so-called measure-valued process [9].

Two standard policies that have been studied extensively in the literature on single-station systems are FIFO (first-in-first-out) and LIFO (last-in-first-out) (see **Queueing Disciplines** for further discussion). Note that the preemptive version of LIFO is not an HL policy. Generally speaking, in the network setting neither of these policies is appealing. In fact, FIFO, has been shown to have very poor performance in OMQNs [10,11]. However, a reasonable modification of FIFO is the FISFO (first-in-system-first-out) policy. Under FISFO, customers are given a time stamp when they enter the system and carry this time stamp throughout their sojourn in the network. When a station becomes free, it chooses the customer with the earliest time stamp. One can argue that FISFO is the appropriate generalization of single station inspired FIFO policies, when working in the network setting. Furthermore, FISFO does not exhibit the undesirable properties of FIFO in this setting [12].

One class of policies that is important in the study of OMQNs is the set of *static buffer priority* (SBP) policies. Under such a policy, all the classes served at station j are given a fixed ranking. When the station becomes free, it serves the job at the HL in the highest ranking class in which customers are present. It is usually assumed that within each class, customers are lined up in FIFO order. The SBP policies are subdivided into those that preempt a low-ranking customer when a higher-ranking customer arrives, and those that allow the low-ranking customer to complete service. Furthermore, preemptive SBP policies may require preempted jobs to begin service anew, or to continue

service at the point at which it was interrupted (these are called *preemptive resume* policies). Special cases of SBP policies can be defined under certain network topologies. For example, in a reentrant line, two policies of interest are FBFS (first-buffer-first-served) and LBFS (last-buffer-first-served). Under FBFS, those classes with lower indices have higher rankings. Under LBFS, those classes with higher indices have higher rankings. Both of these policies have good properties with respect to stabilizing OMQNs [13].

Another class of policies of interest are HL processor-sharing policies, which have applications in computer systems modeling. The class of generalized head-of-the-line processor-sharing (GHLPS) policies is parameterized by a proportion vector $\beta = (\beta_1, \dots, \beta_K) > 0$. At each instant, the station shares its capacity among all the customers at HL in any nonempty class, where the proportion of capacity devoted to k customers at time t is given by

$$\frac{\beta_k \cdot 1\{Q_k(t) > 0\}}{\beta_1 \cdot 1\{Q_1(t) > 0\} + \dots + \beta_K \cdot 1\{Q_K(t) > 0\}},$$

where $1\{A\}$ is an indicator function for event A . A related class of policies are the generalized HL proportional processor-sharing policies. Again, these policies are parameterized by a positive vector $\beta = (\beta_1, \dots, \beta_K) > 0$, now called the *weight vector*. Similar to the GHLPS policies, at each instant, the station shares its capacity among customers at HL in each nonempty class. However, now the proportion dedicated to class k at time t is weighted by the number of customers in the class:

$$\frac{\beta_k \cdot Q_k(t)}{\beta_1 \cdot Q_1(t) + \dots + \beta_K \cdot Q_K(t)}.$$

TRAFFIC EQUATIONS

In OMQNs an important quantity is the *total arrival rate* λ_ℓ , for each class $\ell \in \{1, \dots, K\}$. For class ℓ , λ_ℓ is the long-run rate at which customers of class ℓ arrive at buffer ℓ as a result of internal routing and exogenous

arrivals, assuming that the network is stable (a notion to be discussed shortly). The total arrival rate is given by the *traffic equations*:

$$\lambda_\ell = \alpha_\ell + \sum_{k=1}^K \lambda_k P_{k,\ell}. \quad (4)$$

These equations can be written more compactly in vector form: $\lambda = \alpha + P'\lambda$, where $\lambda = (\lambda_1, \dots, \lambda_K)'$. The classic traffic equations for a Jackson network have the same form as those presented above in Equation (4). Since we assumed that P corresponds to an open network, these equations have a unique solution given by $\lambda = R\alpha$.

The total arrival rates can then be used to define the traffic intensity at each station $j \in \{1, \dots, J\}$:

$$\rho_j = \sum_{k \in C(j)} m_k \lambda_k.$$

One can interpret ρ_j as the average amount of work that arrives at station j per unit time, in the long run.

The traffic intensity plays a role in stability analysis of queueing networks. Roughly speaking, a network is stable if the number of customers in the system does not grow without bound over time. Various more precise notions of stability are discussed in the article. For certain definitions of stability, the *usual traffic conditions* must hold for the network to be stable:

$$\rho_j < 1, \quad \text{for all } j = 1, \dots, J.$$

In other cases, it is sufficient for a weaker form of the usual traffic conditions to hold:

$$\rho_j \leq 1, \quad \text{for all } j = 1, \dots, J.$$

QUEUEING NETWORK EQUATIONS

In order to precisely describe the evolution of multiclass queueing networks, we present the equations governing the dynamics of a network. To facilitate the mathematical description we introduce several new processes.

Let $A_\ell(t)$ denote the total numbers of class ℓ arrivals (internal and external) up to time t . $T_\ell(t)$ is the total amount of time that server $s(\ell)$ has spent serving class ℓ jobs and $D_\ell(t)$ is the total number of jobs that have completed processing up to time t . Finally, let $W_j(t)$ be the *immediate workload process* for station j . In particular, $W_j(t)$ represents the total amount of time required to finish all the work currently at the station. We let $A(t)$, $T(t)$, and $D(t)$ denote the natural K -dimensional column vectors and $W(t)$ denote the J -dimensional column vector of the workload processes. Note that one can think of $\{T(t), t \geq 0\}$ as the network control, that is, given the primitive cumulatives and the K -dimensional process $T(\cdot)$, the queueing dynamics are completely determined, as will be seen in the equations below. Furthermore, the scheduling policies discussed in the section titled “Scheduling Policies” determine the effort that is allocated to each class at each time instant, and thus a properly defined scheduling policy should uniquely determine $T(\cdot)$ for a given sample path of the primitive cumulatives.

We are now ready to present the queueing network equations:

$$A(t) = E(t) + \sum_{k=1}^K \Phi^k(D_k(t)), \quad (5)$$

$$Q(t) = Q(0) + A(t) - D(t), \quad (6)$$

$$W(t) = CV(A(t) + Q(0)) - CT(t). \quad (7)$$

Above, C is the $J \times K$ *constituency matrix* whose elements are

$$C_{j,k} = \begin{cases} 1 & \text{if } k \in C(j), \\ 0 & \text{otherwise.} \end{cases}$$

The first equation expresses the fact the total arrivals are equal to the internal arrivals plus external arrivals. Equation (6) is simply an input–output equation. The last equation precisely defines the notion of immediate workload discussed above in the following sense. The term $A(t) + Q(0)$ embodies the number of jobs that have arrived at each

station (or that were present at time 0) in the network up to time t . Applying the operator $V(\cdot)$ “converts” jobs into units of work. The constituency matrix C sums the work from the job classes according to stations. Thus, each component of $CV(A(t) + Q(0))$ gives the amount of work presented to each station up to time t . Each component in the second term in Equation (7) is the amount of time each station has worked during $[0, t]$. Therefore, the difference of the terms gives the current workload at each station.

Equations (5–7) represent the dynamics of the network under control policy $\{T(t), t \geq 0\}$. For $T(\cdot)$ to be *feasible* it must satisfy: $CT(t) \leq t$ for all $t \geq 0$. (Here, the vector inequality is interpreted componentwise). This requirement simply says no station can devote more than 100% of its time to processing customers.

If, in addition, one wishes to restrict the control to nonidling scheduling policies, then the following additional equations must be satisfied:

$$CT(t) + Y(t) = et \quad (8)$$

$$\begin{aligned} Y_j(t) \text{ can increase only when } W_j(t) &= 0, \\ j &= 1, \dots, J, \end{aligned} \quad (9)$$

where e is the column vector of all 1s and $Y_j(\cdot)$ is the cumulative idleness process at station j (note that idleness “accumulates” whenever a station works at less than full capacity). Since the $Y_j(\cdot)$ are continuous, Equation (9) can be rewritten as follows:

$$\int_0^\infty W_j(t) \, dY_j(t) = 0.$$

This representation is useful when developing approximations (such as fluid models) related to OMQNs.

Equations (5–9) constitute the nonidling queueing network equations. Yet another equation is added when the policy under consideration is HL. HL policies must also satisfy

$$V(D(t)) \leq T(t) \leq V(D(t) + e). \quad (10)$$

To understand this equation, consider a particular class k . Under any service policy (HL

or not), for each fixed t we must have $T_k(t) \geq V_k(D_k(t))$, since the right-hand side is the amount of time that was needed to process the $D_k(t)$ class k jobs, which were processed in $[0, t]$. Next, at any fixed time t , job $D_k(t) + 1$ has not yet departed station $s(k)$. Under an HL policy, only this job will receive service next, hence $T_k \leq V_k(D_k(t) + 1)$. This need not be true under a non-HL policy such as the processor-sharing discipline in which every class k job receives service simultaneously.

However, Equations (5–10) are still very general in that a large class of policies satisfy these equations. When a particular scheduling policy is considered, more equations need to be added to Equations (5–10). We consider two examples below.

FIFO Multiclass Queueing Networks

In an OMQN operating under the FIFO-scheduling policy, jobs at each station are served in the order in which they arrived at the station. This requires the imposition of additional queueing equations:

$$\begin{aligned} D_k(t + W_j(t)) &= Q_k(0) + A_k(t), \\ k &= 1, \dots, K, \end{aligned} \quad (11)$$

for all $t \geq 0$. It may not be immediately obvious that Equation (11) is equivalent to one’s intuitive notion of the FIFO policy. To gain some intuition, consider the case when $Q_k(0) = 0$, in which case, the equation is $D_k(t + W_j(t)) = A_k(t)$. Since $W_j(t)$ is exactly the amount of time required to process the class k jobs present at time t , this expression implies that the processing time during $[t, t + W_j(t)]$ was spent only on jobs that were present at time t . In particular, later-arriving jobs were not processed ahead of those job present at time t .

Thus Equations (5–11) constitute the queueing networks equations for a FIFO OMQN. However, these equations do not uniquely determine the evolution of the network without additional initial conditions, that is, it is not enough to only specify $Q(0)$. We must also give the initial order of the customers at each station. Equivalently, one

can specify the following initial part of the departure process from each class:

$$\{D_k(t) \text{ for } t \leq W_j(0), \quad k = 1, \dots, K\}.$$

Thus Equations (5–11) plus these initial conditions uniquely determine the dynamics of the network, for a given sample path of the primitive processes.

Static Buffer Priority Policies

As mentioned in the section titled “Scheduling Policies”, the set of SBP policies is an important class of policies in which the classes at the station are ranked in a fixed order. Under an SBP policy, additional network equations are again needed to specify the network dynamics. There are several variations of SBP policies. For simplicity, we focus on preemptive resume SBP. Now for any class k let $\Theta(k)$ be the set of classes with the same or higher priority than class k . Then we define the following auxiliary processes:

$$Q_k^+(t) = \sum_{\ell \in \{C(s(k)) \cap \Theta(k)\}} Q_\ell(t)$$

$$\text{and } T_k^+(t) = \sum_{\ell \in \{C(s(k)) \cap \Theta(k)\}} T_\ell(t).$$

An SBP policy with a given ranking can be characterized as follows:

$t - T_k^+(t)$ can increase only when

$$Q_k^+(t) = 0 \quad k = 1, \dots, K. \quad (12)$$

Essentially, Equation (12) indicates that for a fixed class k , the idleness process for class k jobs and higher priority jobs can increase only when no such jobs are present. Put another way, if class k or higher priority jobs are present, then 100% of processing time is devoted to those jobs. Equations (5–10) and Equation (12) constitute the SBP queueing networks equations. The initial data for an SBP policy is simply the initial queue length vector $Q(0)$.

FURTHER TOPICS

Additional details on the framework discussed in this article can be found in Refs 2–4. We have not discussed the case where the network is *closed*. In such a network, a finite number of jobs circulate forever in the network [14–17]. Obviously, other enhancements to the basic framework are possible. For example, some researchers have considered OMQNs with setups, batching, or parallel servers [18,19]. Another important extension is to so-called flexible server networks, which are discussed in the article. A further generalization is described. Several other articles discuss analysis and control of OMQNs, including ***Conservation Laws and Related Applications***, and ***Performance Bounds in Queueing Networks***.

REFERENCES

1. Harrison JM. Brownian models of queueing networks with heterogeneous customer populations. In: Fleming W, Lions P.L, editors. Proceedings of the IMA Workshop on Stochastic Differential Systems. New York: Springer; 1988;10:147–186.
2. Bramson M. Stability of queueing networks. Berlin Heidelberg: Springer; 2008.
3. Chen H, Yao D. Fundamentals of queueing networks. New York: Springer; 2001.
4. Dai JG. Stability of fluid and stochastic processing networks. MaPhySto miscellanea publication, No. 9. Aarhus, Denmark: Centre for Mathematical Physics and Stochastics; 1999.
5. Berman A, Plemmons RJ. Nonnegative matrices in the mathematical sciences. New York: Academic Press; 1979.
6. Jackson JR. Networks of waiting lines. Oper Res 1957;5:518–521.
7. Jackson JR. Jobshop-like queueing systems. Manage Sci 1963;10:131–142.
8. Chen H, Yao D. Stable priority disciplines for multiclass networks. In: Paul Glasserman KS, Yao D, editors. Proceedings of Workshop on Stochastic Networks: Stability and Rare Events. Columbia University, New York: Springer-Verlag; 1996.
9. Gromoll HC, Puha A, Williams RJ. The fluid limit of a heavily loaded processor sharing queue. Ann Appl Probab 2002;12:797–859.

10. Bramson M. Instability of FIFO queueing networks. *Ann Appl Probab* 1994;4:414–431.
11. Seidman TI. ‘First come, first served’ can be unstable! *IEEE Trans Automat Control* 1994;39:2166–2171.
12. Bramson M. Stability of earliest-due-date, first-served queueing networks. *Queueing Syst: Theory Appl* 2001;39(1): 79–102.
13. Dai JG, Weiss G. Stability and instability of fluid models for re-entrant lines. *Math Oper Res* 1996;21:115–134.
14. Baskett F, Chandy KM, Muntz RR, *et al.* Open, closed and mixed networks of queues with different classes of customers. *J ACM* 1975;22:248–260.
15. Harrison JM, Nguyen V. Some badly behaved closed queueing networks. In: Kelly FP, Williams RJ, editors, *Stochastic networks*, volume 71, of *The IMA volumes in mathematics and its applications*. New York: Springer; 1995. pp. 117–124.
16. Harrison JM, Wein LM. Scheduling networks of queues: heavy traffic analysis of a two-station closed networks. *Oper Res* 1990;38:1052–1064.
17. Harrison JM, Williams RJ, Chen H. Brownian models of closed queueing networks with homogeneous customer populations. *Stochastics* 1990;29:37–74.
18. Dai JG, Jennings OB. Stabilizing queueing networks with setups. *Math Oper Res* 2004;29(4): 891–922.
19. Dai JG, Li C. Stabilizing batch processing networks. *Oper Res* 2003;51:123–136.

MULTICOMMODITY FLOWS

BALACHANDRAN VAIDYANATHAN
Operations Research, FedEx Express,
Memphis, Tennessee

RAVINDRA K. AHUJA
Innovative Scheduling, Inc., and
Department of Industrial and
Systems Engineering, University
of Florida, Gainesville, Florida

INTRODUCTION

In many applications, several physical commodities share the same network, yet each is governed by its own network flow constraints. For example, in telecommunications applications, telephone calls between specific node pairs in an underlying telephone network each define a separate commodity. However, all these commodities share common telephone line resources. In this article, we study the *multicommodity network flow problem*, which can be used to model such problems. Let $G(N, A)$ denote the underlying network with node set N and arc set A . Each arc has a capacity u_{ij} that restricts the total flow of all commodities on it. The objective of the problem is to send several commodities that reside at one or more points in a network (sources or supplies) to one or more points on the network (sinks or demands), incurring minimum cost. The arc capacities bind the flows of all commodities together and complicate the problem. If the common arc capacity constraints are relaxed, the multicommodity flow problem can be solved as several independent minimum cost flow problems.

Let K be the set of all commodities, decision variable x_{ij}^k be the flow of commodity $k \in K$ on arc (i, j) , x^k be the flow vector of commodity $k \in K$, c^k be the cost vector for commodity $k \in K$, b^k be the supply/demand vector for commodity $k \in K$, u_{ij}^k be the maximum flow of commodity $k \in K$ on arc $(i, j) \in A$, and u_{ij} be the total capacity of arc $(i, j) \in A$.

Then, the multicommodity flow problem is formulated as follows:

$$\begin{aligned} & \text{Minimize } \sum_{k \in K} c^k x^k \\ & \sum_{k \in K} x_{ij}^k \leq u_{ij}, \text{ for all } (i, j) \in A \\ & \sum_{\{j: (i, j) \in A\}} x_{ij}^k - \sum_{\{j: (j, i) \in A\}} x_{ji}^k = b_i^k, \text{ for all } i \in N, k \in K \\ & 0 \leq x_{ij}^k \leq u_{ij}^k, \text{ for all } (i, j) \in A, k \in K. \end{aligned}$$

The first set of constraints (*bundle constraints*) link the different commodities by restricting the total flow on each arc by its capacity. The second set of constraints ensures the flow balance of each commodity at each node in the network, and the third set of constraints imposes individual bounds on the flow of each commodity on each arc. The integrality of solutions is one very important distinguishing feature between single and multicommodity flow problems. One particularly beneficial feature of single-commodity network flow problems (or minimum cost flow problems) is that they always have integer solutions when the supply/demand and capacity data are integer-valued. For multicommodity flow problems, however, this is not the case [1].

The three basic approaches for solving multicommodity flow problems are (i) price-directive decomposition algorithm [2–8]; (ii) resource-directive decomposition algorithm [5, 9, 10]; and (iii) basis partitioning [11–13]. Ford and Fulkerson [14] and Tomlin [15] first suggested the column generation approach. The first of these papers was the forerunner to the general [16] decomposition procedure of mathematical programming. The excellent survey papers by Assad [17] and Kennington [18] describe all these algorithms and several standard properties of multicommodity flow problems. The book by Kennington and Helgason [12] and the doctoral dissertation by Schneur [19] are other valuable references on this

topic. Other algorithms for the multicommodity flow problem are due to Barnhart [20], Barnhart and Sheffi [21], Farvolden *et al.* [22], Frangioni and Gallo [23], and Detlefsen and Wallace [24]. Schneur [19] and Schneur and Orlin [25] studied scaling techniques for the multicommodity flow problem. Barnhart *et al.* [26] developed a column generation model and an integer programming based branch-and-price-and-cut algorithm for integral multicommodity flow problems. Brunetta *et al.* [27] investigated polyhedral approaches and branch-and-cut algorithms for the integral version.

Multicommodity network flow models have been widely applied in several domains. Some interesting applications are due to the following [5, 28–40].

The general integral version of the multicommodity problem is NP-complete. In this article, we consider the multicommodity flow problem under the assumption that the flow variables can be fractional. This problem is a linear programming problem, and therefore polynomially solvable. While this is a simplifying assumption, techniques to solve the nonintegral version give us a basic understanding and often serve as building blocks for algorithms for the integral version. For example, the branch-and-price-and-cut algorithm for integral multicommodity flow problems is based on the column generation algorithm for the nonintegral version. The rest of this article is organized as follows. In the section titled “Applications,” we identify some applications of the multicommodity flow problem; in the section titled “Optimality Conditions,” we discuss the optimality conditions; and finally in the section titled “Multicommodity Flow Algorithms,” we describe algorithms to solve the problem.

APPLICATIONS

Multicommodity flow problems arise in a wide variety of application contexts where (i) several physically distinct commodities flow over a common network and (ii) a single commodity flows over the network, but it has multiple points of origin and destination that need to send the commodities to each other. We now introduce both kinds of applications.

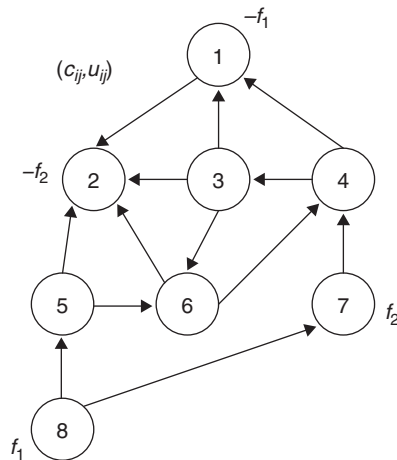


Figure 1. Multicommodity flow problem.

Communication Networks. Nodes represent the origin and destination points for messages, and arcs represent the transmission lines. The messages between different pairs of nodes represent the different commodities, and the supply and demand for each commodity is the number of messages that must be sent between its origin and destination. Each transmission line has a fixed capacity that is shared by all the messages that are routed through it. The problem of determining the minimum cost routing of messages is a multicommodity flow problem. Figure 1 gives an illustration of a problem with two commodities: commodity 1 containing f_1 messages that must be sent from node 8 to node 1, and commodity 2 containing f_2 messages that must be sent from node 7 to node 2.

Railroad Networks. In the railroad context, yards and junction points represent the nodes, and the tracks represent arcs in the network. Since the system incurs various costs for different goods, we divide traffic demand into separate classes. Each commodity in this network corresponds to a particular class of demand between a particular origin–destination pair. The bundle capacity of each arc is the number of railcars that we can load on the trains scheduled to be

dispatched on that arc (over some period of time). The decision problem is to meet the demand of cars at the minimum possible operating cost.

Distribution Networks. In distribution systems planning, we wish to distribute multiple products from plants to retailers using a fleet of trucks or railcars with a variety of railheads and warehouses. The products define the commodities of the multicommodity flow problem, and the joint capacities of the plants, warehouses, rail yards, and the shipping lanes define the bundle constraints. These applications may have bundle constraints imposed on the nodes (warehouses/plants) as well as arcs; such bundle constraints can be handled by the multicommodity flow framework using a node-splitting transformation.

OPTIMALITY CONDITIONS

Since the multicommodity flow problem is a linear program, we can use linear programming optimality conditions to characterize optimal solutions to the problem. The dual has two types of variables: a *price* w_{ij} corresponding to each arc's bundle constraint, and a node potential $\pi^k(i)$ corresponding to commodity k 's flow balance constraint at node i (we assume that the flow variables x_{ij}^k are unbounded). Using the dual variables, we define reduced cost $c_{ij}^k(\pi)$ of arc (i, j) with respect to commodity k as $c_{ij}^k(\pi) = c_{ij}^k - \pi^k(i) + \pi^k(j) + w_{ij}$.

Next, we write the dual of the multicommodity flow problem

$$\begin{aligned} & \text{Maximize } - \sum_{(i,j) \in A} u_{ij} w_{ij} + \sum_{k \in K} b^k \pi^k \\ & c_{ij}^k(\pi) = c_{ij}^k - \pi^k(i) + \pi^k(j) + w_{ij} \geq 0, \\ & \quad \text{for all } (i, j) \in A, k \in K \\ & w_{ij} \geq 0, \text{ for all } (i, j) \in A. \end{aligned}$$

The optimality conditions for linear programming are called the *complementary slackness (optimality) conditions* and state that a primal feasible solution x and a dual

feasible solution are optimal to the respective problems, if and only if the product of each primal variable and the slack in the corresponding dual constraint is zero. The complementary slackness conditions for the primal-dual pair of the multicommodity flow problem are as follows.

Multicommodity Flow Complementary Slackness Conditions. *The commodity flows x_{ij}^k are optimal for the multicommodity flow problem with each $u_{ij}^k = +\infty$ if and only if they are feasible and for some choice of arc prices $w_{ij} \geq 0$ and node potentials $\pi^k(i)$, the reduced costs and arc flows satisfy the following complementary slackness conditions:*

- (a) $w_{ij} \left(\sum_{k \in K} x_{ij}^k - u_{ij} \right) = 0$ for all $(i, j) \in A$;
- (b) $c_{ij}^k(\pi) = c_{ij}^k - \pi^k(i) + \pi^k(j) + w_{ij} \geq 0$, for all $(i, j) \in A, k \in K$;
- (c) $c_{ij}^k(\pi) x_{ij}^k = 0$ for all $(i, j) \in A, k \in K$.

In the next section, we describe algorithms that use the complementary slackness conditions to solve the multicommodity flow problem.

MULTICOMMODITY FLOW ALGORITHMS

In this section, we describe two techniques to solve the multicommodity flow problem, namely, Lagrangian relaxation and Dantzig-Wolfe Decomposition.

Lagrangian Relaxation

Lagrangian relaxation is a general solution strategy for solving mathematical programs that permit us to decompose problems to exploit their special structure. It works by moving hard constraints to the objective function and penalizing their violation. For the multicommodity flow problem, the complicating constraints are the bundle constraints. We move these constraints to the objective function by associating nonnegative Lagrangian multipliers w_{ij} with the bundle constraints and creating the

following Lagrangian subproblem.

$$\begin{aligned}
L(w) = & \text{minimize } \sum_{k \in K} \sum_{(i,j) \in A} c_{ij}^k x_{ij}^k \\
& + \sum_{(i,j) \in A} w_{ij} \left(\sum_{k \in K} x_{ij}^k - u_{ij} \right) \\
= & \sum_{k \in K} \sum_{(i,j) \in A} (c_{ij}^k + w_{ij}) x_{ij}^k - \sum_{(i,j) \in A} w_{ij} u_{ij} \\
& \sum_{\{j:(i,j) \in A\}} x_{ij}^k - \sum_{\{j:(j,i) \in A\}} x_{ji}^k = b_i^k, \text{ for all } i \in N, k \in K \\
& x_{ij}^k \geq 0, \text{ for all } (i,j) \in A, k \in K.
\end{aligned}$$

The second term in the objective function is a constant for a given set of Lagrangian multipliers and can be ignored. Also, since the bundling constraints that link the different commodity flows are relaxed, the Lagrangian subproblem can be solved efficiently as k independent, minimum-cost flow problems. For any value of Lagrangian multipliers, it can be shown that $L(w)$ is a lower bound to the optimal objective function of the original problem. The tightest possible lower bound is obtained by solving the Lagrangian multiplier problem $L^* = \max_w L(w)$. The Lagrangian multipliers w for which $L(w) = L^*$ are called the *optimal Lagrangian multipliers*. The following theorem, stated without proof, is the basis for Lagrangian relaxation used to solve the multicommodity flow problem.

Theorem 1. *Suppose we apply the Lagrangian Relaxation technique to a linear programming problem, the optimal value L^* of the Lagrangian multiplier problem equals the optimal objective function value of the linear programming problem.*

Lagrangian relaxation is a price-directive decomposition method to solve the multicommodity flow problem. From Theorem 1, it follows that by solving the Lagrangian multiplier problem, we obtain the optimal solution of the multicommodity flow problem. Subgradient optimization is a commonly used technique to solve the Lagrangian multiplier problem. This method iteratively does the following: (i) given a set of Lagrangian multipliers (w), it solves the multicommodity flow

problem as $|K|$ independent minimum cost flow problems with cost coefficients $c_{ij}^k + w_{ij}$; and (ii) updates the multipliers based on the solution obtained in (i). It can be shown that, by choosing a suitable step size to update the multipliers, the subgradient method terminates in a finite number of steps to the optimal multipliers. Lagrangian relaxation is a popular method to solve the multicommodity flow problem owing to its ease of implementation and empirical performance.

Dantzig-Wolfe Decomposition

To simplify our discussion, we consider a special case of the multicommodity flow problem: we assume that each commodity k has a single source node s^k and a single sink node t^k , and a flow requirement of d^k units between these source and sink nodes. We also assume that there are no flow bounds on the individual commodities other than the bundle constraints. Therefore, for each commodity k , the subproblem constraints $\sum_{\{j:(i,j) \in A\}} x_{ij}^k - \sum_{\{j:(j,i) \in A\}} x_{ji}^k = b_i^k, x_{ij}^k \geq 0$ define a shortest path problem. We also assume that the cost of every cycle in the network is nonnegative for each commodity. Hence, any potential optimal solution can be represented as the sum of flows on directed paths.

Path Flow Reformulation. For each commodity k , let P_k denote the set of all directed paths from the source node s^k to the sink node t^k in the underlying network. In the path flow formulation, we consider one decision variable $f(P)$ for the flow on each path $P \in P_k$ for commodity k . Let $d_{ij}(P)$ be 1 if arc (i,j) is contained in path P , and 0 otherwise. Then, the flow decomposition theorem of network flows states that we can always decompose the optimal arc flow x_{ij}^k into path flows $f(P)$ such that $x_{ij}^k = \sum_{P \in P_k} f(P) d_{ij}(P)$. Hence, the multicommodity flow problem can be reformulated as follows:

$$\begin{aligned}
& \text{Minimize } \sum_{k \in K} \sum_{P \in P_k} c^k(P) f(P) \\
& \sum_{k \in K} \sum_{P \in P_k} d_{ij}(P) f(P) \leq u_{ij}, \text{ for all } (i,j) \in A
\end{aligned}$$

$$\sum_{P \in P_k} f(P) = d^k, \text{ for all } k \in K$$

$$f(P) \geq 0, \text{ for all } P \in P_k, k \in K.$$

The path flow formulation has a simple structure. It has one constraint that bounds the flow on each arc in the network, and it has one constraint for each commodity that ensures the demand for that commodity is satisfied. The path-based formulation implicitly takes care of the flow balance constraints for each commodity at every node in the network. For a network with n nodes, m arcs, and K commodities, this formulation contains $m + K$ constraints in addition to the nonnegativity constraints on path flows.

Optimality Conditions. Since the path flow formulation contains one bundle constraint for each arc and one demand constraint for every commodity, the dual linear program has a dual variable w_{ij} for each arc and another dual variable r^k for each commodity $k \in K$. With respect to these dual variables, the reduced cost for each path flow variable $f(P)$ is $C_P(w, r) = c^k(P) + \sum_{(i,j) \in P} w_{ij} - r^k$. The complementary slackness optimality conditions for the linear programming problem take the following form.

Path Flow Complementary Slackness Optimality Conditions. The commodity path flows $f(P)$ are optimal if and only if we can find (non-negative) arc prices w_{ij} and commodity price r^k so that the reduced costs and arc flows satisfy the following complementary slackness conditions:

- (a) $w_{ij}(\sum_{k \in K} \sum_{P \in P_k} d_{ij}(P)f(P) - u_{ij}) = 0$, for all $(i, j) \in A$;
- (b) $C_P(w, r) \geq 0$ for all $P \in P_k, k \in K$;
- (c) $C_P(w, r)f(P) = 0$ for all $P \in P_k, k \in K$.

The Dantzig–Wolfe decomposition method works in the following manner. Imagine that K different decision makers, as well as one “coordinator,” are solving the K -commodity flow problem. The coordinator’s job is to solve the path formulation of the problem, which we refer to as the “master” or “coordinating”

problem. In general, the coordinator has on hand only a subset of the columns of the master problem. Since the coordinator can, at best, solve the linear program as restricted to this subset of columns, we refer to this smaller linear program as *the restricted master problem*. Each of these K decision makers plays a special role in solving the problem. The K decision makers, with guidance from the coordinator in the form of arc prices, generate entering variables or columns, with the k th decision maker generating the columns of the master problem corresponding to the k th commodity.

The coordinator solves the restricted master problem to optimality using any linear programming technique, such as the simplex algorithm, and then determines whether the solution to the restricted master is optimal for the original problem, or if another column has a negative reduced cost and can enter the basis. The coordinator broadcasts the optimal set of simplex multipliers (or prices) for the restricted master problem: w_{ij} corresponding to each arc $(i, j) \in A$ and r^k corresponding to each commodity $k \in K$.

Once the coordinator has broadcast the prices, the decision maker for commodity k determines the least cost way of shipping d^k units from the source node s^k to the sink node t^k of commodity k , assuming that each arc (i, j) has an associated toll of w_{ij} in addition to its arc cost c_{ij}^k . If the cost of the shortest path is less than r^k , it implies that the corresponding reduced cost $C_P(w, r)$ is negative. The k th decision maker will report this path to the coordinator as an improving path, and the decision variable corresponding to this path becomes a candidate to enter the basis. If the cost of this path equals r^k , then the k th decision maker will not report anything to the coordinator. The cost of the path can never be more than r^k because the current solution is optimal to the restricted master problem, and the coordinator is already using some path of cost r^k for the k th commodity in the current solution. The algorithm terminates when no decision maker finds a candidate path to enter the basis (complementary slackness conditions are satisfied).

To dynamically generate entering variables or columns, we solve a shortest path problem; hence, the Dantzig–Wolfe decomposition method is an example of the column generation method for solving linear programming problems. Also, notice that the K subproblems correspond to the Lagrangian relaxation problem with an additional multiplier cost of w_{ij} imposed upon each arc (i, j) . Consequently, we could view the coordinator as setting the Lagrangian multipliers and solving the Lagrangian multiplier problem. In fact, Dantzig–Wolfe decomposition is an efficient method for solving the Lagrangian multiplier problem if we measure efficiency by the number of iterations an algorithm performs.

Unfortunately, in applying Dantzig–Wolfe decomposition, the coordinator must solve a linear program with $m + K$ constraints at each iteration. This method for updating simplex multipliers is far more time consuming than updating the multipliers using subgradient optimization. Hence, the Dantzig–Wolfe decomposition method has generally not proven to be an efficient method for solving the multicommodity flow problem. Nevertheless, Dantzig–Wolfe decomposition has one important advantage that distinguishes it from other Lagrangian relaxation-based algorithms. It can be shown that the solution to the subproblems provides us with a lower bound on the optimal value of the problem. Consequently, at each step we also have a bound on how far the current feasible solution is from optimal. Therefore, we can terminate the algorithm at any step not only with a feasible solution but also with a guarantee of how far that solution is from optimality.

REFERENCES

- Ahuja RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms, and applications. New Jersey: Prentice Hall; 1993.
- Cremeans JE, Smith RA, Tyndall GR. Optimal multicommodity network flows with resource allocation. *Nav Res Log Q* 1970;17:269–280.
- Swoveland C. Decomposition algorithms for the multi-commodity distribution problem. Working Paper No. 184. Los Angeles (CA): Western Management Science Institute, University of California; 1971.
- Chen H, Dewald CG. A generalized chain labeling algorithm for solving multicommodity flow problems. *Comput Oper Res* 1974;1:437–465.
- Assad AA. Models for rail transportation. *Transp Res* 1980;14A:205–220.
- Mamer JW, McBride RD. A decomposition-based pricing procedure for large-scale linear programs: An application to the linear multicommodity flow problem. *Manage Sci* 2000;46:693–709.
- Holmberg K, Yuan D. A multicommodity network-flow problem with side constraints on paths solved by column generation. *INFORMS J Comput* 2003;15:42–57.
- Larsson T, Yuan D. An augmented Lagrangian algorithm for large scale multicommodity routing. *Comput Optim Appl* 2004;27:187–215.
- Geoffrion AM, Graves GW. Multicommodity distribution system design by Bender's decomposition. *Manage Sci* 1974;20:822–844.
- Kennington JL, Shalaby M. An effective subgradient procedure for minimal cost multicommodity flow problems. *Manage Sci* 1977;23:994–1004.
- Graves GW, McBride RD. The factorization approach to large scale linear programming. *Math Program* 1976;10:91–110.
- Kennington JL, Helgason RV. Algorithms for network programming. New York: Wiley-Interscience; 1980.
- Castro J, Nabona N. An implementation of linear and nonlinear multicommodity network flows. *Eur J Oper Res* 1996;92:37–53.
- Ford LR, Fulkerson DR. A suggested computation for maximal multicommodity network flow. *Manage Sci* 1958;5:97–101.
- Tomlin JA. A linear programming model for the assignment of traffic. Proceedings of the 3rd Conference of the Australian Road Research Board 3. 1966. pp. 263–271.
- Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;8:101–111.
- Assad AA. Multicommodity network flows - A survey. *Networks* 1978;8:37–91.
- Kennington JL. Survey of linear cost multicommodity network flows. *Oper Res* 1978;26:209–236.
- Schneur R. Scaling algorithms for multicommodity flow problems and network

- flow problems with side constraints [PhD dissertation]. Cambridge (MA): Department of Civil Engineering, MIT; 1991.
20. Barnhart C. Dual-ascent methods for large-scale multi-commodity flow problems. *Nav Res Log* 1993;40:305–324.
 21. Barnhart C, Sheffi Y. A network-based primal-dual heuristic for the solution of multicommodity network flow problems. *Transp Sci* 1993;27:102–117.
 22. Farvolden JM, Powell WB, Lustig IJ. A primal partitioning solution for the arc-chain formulation of a multicommodity network flow problem. *Oper Res* 1993;41:669–693.
 23. Frangioni A, Gallo G. A bundle type dual-ascent approach to linear multicommodity min-cost flow problems. *INFORMS J Comput* 1999;11:370–393.
 24. Detlefsen NK, Wallace SW. The simplex algorithm for multicommodity networks. *Networks* 2002;39:15–28.
 25. Schneur R, Orlin JB. A scaling algorithm for multicommodity flow problems. *Oper Res* 1998;46:231–246.
 26. Barnhart C, Hane CA, Vance PH. Using branch-and-price-and-cut to solve origin-destination integer multicommodity flow problems. *Oper Res* 2000;48:318–326.
 27. Brunetta L, Conforti M, Fischetti M. A polyhedral approach to an integer multicommodity flow problem. *Discrete Appl Math* 2000;101:13–36.
 28. Golden BL. A minimum cost multicommodity network flow problem concerning imports and exports. *Networks* 1975;5:331–356.
 29. Crainic T, Ferland JA, Rousseau JM. A tactical planning model for rail freight transportation. *Transp Sci* 1984;18:165–184.
 30. Bellmore M, Bennington G, Lubore S. A multivehicle tanker scheduling problem. *Transp Sci* 1971;5:36–47.
 31. Bodin LD, Golden BL, Schuster AD, *et al.* A model for the blockings of trains. *Transp Res* 1980;14B:115–120.
 32. Korte B 1988. Applications of combinatorial optimization. Technical Report No. 88541-OR. Bonn: Institute für Ökonometrie und Operations Research, Nassestrasse 2, D-5300.
 33. Gautier A, Granot F. Forest management: a multicommodity flow formulation and sensitivity analysis. *Manage Sci* 1995;41:1654–1688.
 34. Ouaja W, Richards B. A hybrid multicommodity routing algorithm for traffic engineering. *Networks* 2004;43:125–140.
 35. Ahuja RK, Liu J, Orlin JB, *et al.* Solving real-life locomotive scheduling problems. *Transp Sci* 2005;39:503–517.
 36. Vaidyanathan B, Jha KC, Ahuja RK. Multicommodity network flow approach to the railroad crew scheduling problem. *IBM J Res Dev* 2007;51:325–344.
 37. Vaidyanathan B, Ahuja RK, Liu J, *et al.* Real-life locomotive planning: new formulations and computational results. *Transp Res Part B* 2008a;42:147–168.
 38. Vaidyanathan B, Ahuja RK, Orlin JB. The locomotive routing problem. *Transp Sci* 2008b;42:492–507.
 39. Kumar A, Vaidyanathan B, Ahuja RK. Railroad locomotive scheduling. In: Floudas CA, Pardalos PM, editors. *Encyclopedia of optimization*. Dordrecht, The Netherlands: Springer; 2009a. pp. 3236–3245.
 40. Kumar A, Vaidyanathan B, Jha KC, *et al.* Railroad crew scheduling. In: Floudas CA, Pardalos PM, editors. *Encyclopedia of optimization*. Dordrecht, The Netherlands: Springer; 2009b. pp. 3227–3236.

MULTIECHELON MULTIPRODUCT INVENTORY MANAGEMENT

RODNEY P. PARKER
The University of Chicago, Chicago,
Illinois

INTRODUCTION

The management of inventory when facing uncertain demand is a fundamental operational challenge and has amongst the largest pools of research in operations research. Specifically, the analysis of multiechelon systems, where there is a physical flow of goods from one installation to another, ultimately satisfying uncertain market demand, has been an active research area from the 1960s until the present day. The typical trade-off is to reduce the costs associated with holding inventory, while being responsive in providing goods to end customers in the face of uncertain demand. In this article, we initially summarize the optimal inventory policies for single installations and then discuss various multiechelon inventory systems.

SINGLE INSTALLATION INVENTORY MODELS

Appreciating the fundamentals of single installation models will facilitate the understanding of multiechelon inventory systems considered in the next section. Most inventory research utilizes discrete-time models (known as *periodic review* models), while the remainder considers a continuous-time basis (known as *continuous review* models), with the exception being the newsvendor model which considers inventory management in a single period (see ***Newsvendor Models***). We discuss about both in this section but focus on discrete-time models in subsequent sections.

Discrete-Time Single Installation

Considered the most fundamental inventory model, we examine a single installation where the objective is the optimal management of a single durable product in the face of unknown demand over multiple (discrete) time periods. Some common assumptions dealing with such models include (i) the demand probability distribution is known, (ii) the cost associated with holding physical inventory is a convex function in the inventory amount, (iii) the cost associated with unmet demand is a convex function in the demand shortfall, (iv) a specific treatment of customers whose demand is unmet. Common treatments under assumption (v) include customers' unmet demand is backlogged (where the amount of unsatisfied demand is carried into the following time period and satisfied then or thereafter), lost (where the unsatisfied demand in the current period disappears entirely), or partially backlogged and partially lost (where some proportion of the unsatisfied demand will be backlogged while the remainder will be lost).

Let us introduce some notation. Let $z^+ = \max(z, 0)$. Let x_t be the amount of inventory at the beginning of period t , a_t be the amount ordered in period t , D_t be the (unknown) demand in period t , $H(x)$ and $P(x)$ be the cost of holding or not satisfying demand for finishing inventory x , $f(x) = x^+ - \gamma(-x)^+$ be the amount of inventory/backlog in the following period if the current period ends with x and $\gamma \in [0, 1]$ is the proportion of x that will be backlogged (under this regime, a negative beginning inventory position represents that amount of backlog), and $\alpha \in [0, 1]$ be the single-period discount factor. A common assumption is to apply unit holding (h) and stockout (p) costs: $H(x) = hx^+$ and $P(x) = p(-x)^+$. We will use these unit holding and stockout costs. Let the objective be to minimize the discounted expected costs (over $T < \infty$ periods):

$$V_t(x_t) = \min_{a_t, a_{t+1}, \dots, a_T \geq 0} \mathbb{E} \left[\sum_{i=t}^T \alpha^{i-t} \times (h(x_i + a_i - D_i)^+ + p(D_i - x_i - a_i)^+) \right] \quad (1)$$

$$= \min_{a_t \geq 0} \mathbb{E} [h(x_t + a_t - D_t)^+ + p(D_t - x_t - a_t)^+ + \alpha V_{t+1}(f(x_t + a_t - D_t))] \quad (2)$$

$$V_{T+1}(\cdot) = 0, \quad (3)$$

where Equation (2) is the recursive Bellman representation of the model (see the sections titled “Dynamic Programming” and “Markov Decision Processes (MDPs)” in this encyclopedia). A common variable translation is to define $y_t := x_t + a_t$ as the new decision variable, representing the amount of inventory after the order has arrived but before demand is subtracted.

$$V_t(x_t) = \min_{y_t \geq x_t} \mathbb{E} [h(y_t - D_t)^+ + p(D_t - y_t)^+ + \alpha V_{t+1}(f(y_t - D_t))] \quad (4)$$

It is straightforward to show the operand of Equation (4) is convex in y_t in time period t . Let z_t be a quantity that minimizes the operand of Equation (4) in period t . Thus, the most desirable action in period t is to order an amount such that $y_t = z_t$. This will be possible if period t 's beginning inventory position (x_t) is lower than z_t , resulting in an ordered quantity $a_t = z_t - x_t$ (equivalently, $y_t = z_t$). However, if the beginning inventory is greater than z_t , the optimal decision is to order nothing, $a_t = 0$ or $y_t = x_t$. Thus, y_t is known as the *order up-to* quantity. Since the minimizer, z_t , is independent of the initial inventory x_t , the optimal target is known as the *base-stock* level. These statements are true in any (general) period t and constitutes the optimal inventory policy known as the *base-stock* inventory policy. The implementation of this policy requires a comparison of the inventory position at the beginning of the period (x_t) and the target stocking level (z_t), resulting in a different amount (a_t) being ordered in each period.

Now, consider an additional complication where the firm incurs a fixed cost, K , whenever an order is placed ($a_t > 0$). It should be clear that the operand is no longer convex. However, Scarf [1] showed that it is K -convex, basically a concept where a function is convex within a margin of K (a single-variable function g is K -convex if $g(x) + \lambda(g(x) - g(x - v))/v \leq g(x + \lambda) + K$ for all x and positive λ, v). Intuition should suggest that incurring an additional fixed ordering cost (independent of the quantity ordered) should lessen the frequency of ordering. The optimal policy again is a base-stock policy (with an optimal base-stock level, S_t) but an order is placed *only* when the beginning inventory level (x_t) is below $s_t (< S_t)$, a quantity known as the *trigger level*. This is known as an (s, S) policy and is optimal for systems with full backlogging or full lost sales (Heyman and Sobel [2], Chapter 7). In short, the manager will place an order (and incur the fixed cost K) only when there is a sufficient ordering quantity to warrant it. When an order is placed (when $x_t < s_t$), $S_t - x_t$ is ordered and always at least $S_t - s_t$ is ordered.

A notable omission heretofore has been the inclusion of delivery lead times. When $\gamma = 1$ (i.e., full backlogging), lead times that are a known integer multiple of the period can be included while preserving the structure of the above optimal inventory policies, as the materials in transit to the installation may be incorporated into the decision variable, simplifying the analysis. The effect of a larger lead time is for the installation to stock up to a higher level. For values of $\gamma < 1$, this optimal policy structure is destroyed and the analysis is vastly more complicated as the individual quantities in transit to the installation must be tracked within the state space. Indeed, Zipkin [3] indicates that a simple base-stock policy performs poorly compared to numerous suboptimal heuristics under a lost sales assumption ($\gamma = 0$). Kaplan [4] incorporates stochastic lead times under an assumption that orders placed with the supplier will not “cross,” that is, an order placed in period t will not arrive after an order placed in period $t' > t$. Under this condition, Kaplan [4] finds a base-stock inventory policy is optimal.

Another extension is the application of production capacity constraints in the periodic review model (this applies only when $K = 0$). Federgruen and Zipkin [5,6] show that when there is a capacity limit C on the order quantity in each period, the optimal policy is *modified base-stock*, where it is optimal to order up to a target z_t in period t , *if possible*. The last caveat applies when the beginning inventory quantity $x_t < z_t - C$. Under such circumstances $a_t = C$, the most of which can be processed in a period, thus reaching the closest feasible inventory position to z_t . Nonstationary parameters are permitted in the model with a base-stock policy remaining optimal. Also, Markov-modulated demand may also be incorporated, thus permitting demands with a degree of autocorrelation, also resulting in optimal base-stock policies.

Continuous-Time Single Installation

A commonly used policy when the manager continually assesses the inventory levels (rather than examining the inventory periodically) is to place an order of a fixed quantity (Q), whenever the inventory level falls below an ordering trigger level (r). A reasonable (and optimal for the deterministic system) approximation for the order quantity is the Economic Order Quantity (EOQ), $Q = \sqrt{\frac{2\bar{D}K}{h}}$, where $\bar{D}$ is the mean annual demand and h is the annualized unit holding cost. Define the random variable D_{LT} as the demand during the lead time. The ordering trigger level is $r = \mu + z\sigma$, where $\mu = E[D_{LT}]$, σ is the standard deviation of D_{LT} , and z is the distance from the mean of D_{LT} to the trigger level, expressed in standard deviations. z is typically set by considering the equation, $\Pr(D_{LT} > r) = \delta$, where δ is the proportion of periods in which we expect customers to experience a stockout (this is typically called a *type 1 service level*). The quantity δ is typically quite small in most retail scenarios: $\delta = 0.05, 0.01, 0.005$. The quantity $z\sigma$ is known as the *safety stock* and is the additional stock embedded in the system to compensate for the demand uncertainty. It is straightforward to see that if there is no uncertainty, there are no safety stocks and

then $r = \mu$. The above method is straightforward to calculate (and widely taught) but not optimal. To find the optimal (Q, r) parameters, it requires the recursive solution of two equations: $Q = \sqrt{\frac{2\bar{D}(K+pE(D_{LT}-r)^+)}{h}}$ and $\Pr(D_{LT} \geq r) = \frac{Qh}{p\bar{D}}$, where p is the unit shortage cost. Hadley and Whitin [7] indicate that the differences between the backlogging and lost sales cases for the values of Q and r will be small in practice.

In practical terms, the periodic review model requires fewer “inspection” costs to assess the level of inventories compared with the continuous review. As the name suggests, the periodic review measurement of inventory could be done on a daily, weekly, or monthly basis, for example, whereas inventory level must be constantly assessed under continuous review. While the latter may sound arduous, this is typically achieved using automated systems, especially for continuous quantities such as liquids (e.g., oil, sugar) or simply using the checkout scanner to deduct quantities from a virtual inventory counter, vastly reducing the assessment costs, although such scanning may be a source of potentially large error rates in the virtual inventory (see the section titled “Retail Applications” in this encyclopedia; DeHoratius and Raman [8]). Perhaps a more important consideration under this continuous-review optimal policy is that a constant quantity (Q) is ordered whenever an order is placed, which perhaps provides a practical simplicity when considering pallet sizes, truck capacities, and other considerations. In practice, the continuous review policy suggests that an order be placed at the precise time when the inventory level falls below r , which may not necessarily be convenient, given practical organizational routines. The periodic review optimal policy, however, recommends a (potentially) different amount whenever an order is placed, but the orders are placed precisely at the time when the inventory is reviewed; this will be in every period when $K = 0$ for the simple base-stock policy but may not be in every period when $K > 0$ under the (s, S) policy. Such periodic review

policies frequently dovetail with organizational norms such as “place an order every Monday.”

MULTIECHELON INVENTORY SYSTEMS

A common convention we adopt hereafter is to refer to “upstream” and “downstream” (analogizing the flow of water in a river) as directions within the channel, with the former referring to the source of the material and the latter referring to the destination (the market) of the material.

Serial Systems

The simplest multiechelon system is a serial system consisting of N installations, where each installation has (at most) a single upstream and a single downstream installation. Let installation 1 be the one closest to the market and installation N be the one most upstream closest to the “outside supplier.” The market demand is uncertain although the probability distribution is assumed known, as before. The system incurs a stockout cost for not satisfying demand. Unsatisfied demand arises when the realized demand exceeds the available inventory at installation 1 and will be carried into the following period as a backlog. Holding costs are assessed for physical inventory remaining at each installation at the end of each period. With the objective of minimizing the system-wide expected discounted costs, each installation obtains an order from its immediate upstream supplying installation; the amount delivered to installation $j < N$ in period t , a_t^j , cannot exceed installation $j + 1$ ’s inventory at the beginning of period t , x_t^{j+1} . It is assumed that installation N has no availability constraints from the “outside supplier.” Clearly, the flow of material is from upstream to downstream (installation j to installation $j + 1$) and the flow of orders is from downstream to upstream. Assume unit holding and stockout costs, where h_i is the cost of holding a unit of stock at installation i for one period and p is the unit stockout cost. We assume that $h_1 > h_2 > \dots > h_N$. We denote vectors as $\tilde{x} = (x^1, x^2, \dots, x^N)$. The

expected discounted cost is

$$V_t(\tilde{x}_t) = \min_{\tilde{a}_t} \mathbb{E} \left[h_1(x_t^1 + a_t^1 - D_t)^+ + \sum_{j=2}^N h_j(x_t^j + a_t^j - a_t^{j-1}) + p(D_t - x_t^1 - a_t^1)^+ + \alpha V_{t+1}(\tilde{x}_t + \tilde{a}_t - \tilde{D}_t) \right]$$

$$\text{s.t. } a_t^1, a_t^2, \dots, a_t^N \geq 0 \quad (5)$$

$$a_t^j \leq x_t^{j+1}, j < N \quad (6)$$

where constraints (6) state that any order quantity at an installation cannot exceed the inventory available at the immediate upstream installation at the beginning of the period. It should be noted that $x_t^j \geq 0$ for $j > 1$ but x_t^1 could be positive (reflecting physical inventory) or negative (reflecting a backlog). This formulation assumes the delivery lead time between installations is a single time period (for notational convenience), but the results of this section extend to situations where there are lead times, which are an integer multiple of the period. Clark and Scarf [9] are responsible for the analysis of this model and demonstrated the optimal inventory policy (Federgruen and Zipkin [10] showed the finite horizon model’s optimal policy extends to the infinite time horizon). This involved counting inventory in a slightly different way. The proposed counting concept known as *echelon* inventory where echelon i ’s inventory is $X_t^i := \sum_{j=1}^i x_t^j$. In short, echelon inventory is counting all the inventory at an installation and anywhere downstream of this installation which provides a natural ordering of the echelon inventories, $X_t^i \leq X_t^{i+1}$ for $i < N$. As before, backlogging of unmet demand is assumed and an installation 1 backlog is recorded as negative inventory, that is, $X_t^1 = x_t^1 < 0$. Echelon i inventory level, X_t^i , will be negative if the sum of all the downstream physical inventory, $\sum_{j=2}^i x_t^j > 0$, is less than the installation 1 backlog, $x_t^1 < 0$. There is no loss of information in the mapping from installation inventory variables to echelon inventory variables. Similar to the single installation inventory problem, the decision variables

can be modified to the order up-to levels but in echelon inventory variables: $Y_t^i = X_t^i + a_t^i = \sum_{j=1}^i x_t^j + a_t^i$. The value function becomes

$$V_t(\tilde{X}_t) = \min_{\tilde{Y}_t} \mathbb{E} \left[h_1(Y_t^1 - D_t)^+ + \sum_{j=2}^N h_j(Y_t^j - Y_t^{j-1}) + p(D_t - Y_t^1)^+ + \alpha V_{t+1}(\tilde{Y}_t - \tilde{D}_t) \right] \quad (7)$$

$$\text{s.t. } Y_t^j \geq X_t^j, j \leq N \quad (8)$$

$$Y_t^j \leq X_t^{j+1}, j < N. \quad (9)$$

Clark and Scarf [9] demonstrated that this model's objective function is convex in echelon up-to decision variables, thus establishing that an *echelon base-stock* inventory policy is optimal. The importance of this cannot be understated since there is a relationship of differently shaped supply chains to the serial system, as described in the following sections. The interpretation of this policy is that the system manager observes the echelon inventory levels at the beginning of a period (captured in the state space) and for those echelon inventory levels below the optimal echelon base-stock levels, order that difference if such a quantity is available at the immediate upstream installation; if insufficient inventory resides at the upstream installation, order as much as possible. Constraints (8) state that the echelon up to amount must exceed the initial echelon inventory while constraints (9) state the amount ordered cannot exceed the inventory available at the immediate upstream installation. Clark and Scarf [9] show the operand of Equation (7) is *convex* and *separable* in each echelon inventory variable. Since there is a natural ordering of the decision variables arising from their definition, the target minimizing echelon inventory levels, Z^i , are likewise ordered. This could be thought of a *nested* ordering of inventory where echelon i ordering up to optimal Z^i has a total of Z^i units of inventory

within $i - 1$ periods of the market (similarly, if we insert delivery lead times of L^i between the installations $i + 1$ and i , then echelon i has Z^i units within $\sum_{j=1}^{i-1} L^j$ periods of the market). Thus, this *nested* arrangement (i.e., $Z^1 \leq Z^2 \leq \dots \leq Z^N$) adds a layer of protection where, for example, if there is a large demand realization, which dramatically reduces the installation 1 inventory, all the upstream installations observe this (since they are measuring inventory in echelon terms) and immediately order quantities to compensate for this rather than wait for rippling orders to flow up the channel. Additionally, the system has a natural tension of wishing to have as much inventory close to the market as possible for sale while wishing to store the inventory further upstream to incur lower holding costs. Thus, the optimal $\tilde{Z}$ balances the echelon holding costs and the stockout cost for unmet demand at installation 1.

Solving a dynamic program with an N -dimensional state space, even one with a convex value function, can be challenging; Bellman [11] coined the phrase “curse of dimensionality” to describe this situation. Usefully, Clark and Scarf [9] additionally proved the *separability* of the system's value function, meaning it may be decomposed into N single-dimension convex dynamic programs with the echelon inventory level as the state space, vastly reducing the computational effort needed to determine the optimal echelon base-stock levels. This reduces the search for optimal echelon stocking levels to a line search for each individual dynamic program, each of which is convex and independent of the other echelons. In doing this, Clark and Scarf [9] were able to discern an insight that upon this decomposition, each installation absorbs an *induced penalty cost function*, which is a convex cost function applied when the echelon has insufficient inventory to enable the downstream echelon to reach its desired stocking level. In short, the lack of inventory at an installation increases the channel costs and this decomposed single-variable value function compensates by incurring the induced penalty cost.

Linear costs associated with “purchasing” goods by one installation from another may

be included and the echelon base-stock policy remains optimal. A fixed cost (independent of the purchased quantity) may be incurred by installation N when buying from the outside supplier. Installation N will now optimally follow an (s, S) policy (*à la* Scarf [1]), while all the other echelons follow the echelon base-stock policy as previously described. Inserting fixed ordering costs for any echelon $j < N$ creates problems and destroys the optimal policy structure described above and only approximate techniques exist, although Chen and Zheng [12] establish well-performing bounds on such problems. Shang and Song [13] derive upper and lower bounds on the optimal base-stock levels and Gallego and Zipkin [14] derive multiple heuristic methods for this system, usefully connecting the discrete-time and continuous-time versions of this model. Chen and Zheng [15] establish lower bounds on the minimum costs for serial, assembly, and single-warehouse distribution systems. Muharremoglu and Tsitsiklis [16] describe effective algorithms for solving the optimal problem using the single-unit decomposition approach.

Cachon and Zipkin [17] consider a decentralized version of the serial system operated under echelon and installation equilibrium policies, finding system-wide costs increase while equilibrium stocking levels decrease, relative to the integrated channel considered above (see the section titled “Game Theory” in this encyclopedia). The optimal inventory policy of the general N -stage serial system subject to production capacity constraints is unknown, although Parker and Kapuscinski [18] find the optimal *modified echelon base-stock* policy for the $N = 2$ system where the capacity levels are ordered, $C_1 \leq C_2$. This policy dictates that optimal echelon base-stock levels exist and both installations will attempt to reach these levels if possible, although C_1 acts to limit the entire system and the higher installation will never stock any more than C_1 , since the lower installation can never process any more than this amount in a single period.

Assembly Systems

Assembly systems combine multiple parts into a final assembled product. In this section, the optimal inventory policies in assembly systems facing uncertain market demand are described. Such systems are commonplace in practice, where parts and modules from adjacent installations are combined into sub-assemblies, which are themselves combined into a final assembled product. The product’s bill of materials (BOM) specifies the quantity and type of part and sequence of assembly which constitutes the final assembled product. Analyzing the management of inventory within an assembly system facing uncertain demand and with deterministic stationary arbitrary delivery lead times (assumed to be an integer multiple of the period) between installations is considered here.

As before, we consider an objective of minimizing the discounted expected system-wide costs. For the purposes of illustration, we will consider a system where each installation stocks a single part (without loss of generality) which when combined with parts of other installations, results in a single assembly or subassembly. Rosling [19] demonstrated that under a “balanced inventory” condition, the inventory operation in any general assembly system can be reduced to an *equivalent serial system*, thus being able to utilize the optimal policy for a serial system (echelon base-stock) described above. The basic approach is for the system to account for potentially differing lead times for installations supplying an immediately downstream assembling installation. Each of these supplying installations will stock such that it can provide the same quantity to the assembler at the same time. Rosling [19] found that the *echelon lead times*, defined as the sum of all the lead times between the installation in question and the final assembler, determines the order of the installations in an equivalent serial system. The lead times in the equivalent serial system will simply be the differences of the echelon lead times from the assembly system. The intuition behind this is simply to find an equivalent distance from the market in terms of lead time. Once this transformation to the series system is achieved, the optimal echelon base-stock levels can be found using

the Clark and Scarf [9] approach, which will serve as the echelon base-stock levels in the assembly system.

Distribution Systems

Distribution systems consist of installations where there is at most a single upstream installation and potentially multiple downstream installations. In practice, they typically involve the movement of a finished good to warehouses and retail outlets, where they are sold to the market. When facing unknown demand, the optimal inventory policy is not known for general distribution systems, but some results are known for specific but commonly observed distribution systems. Most of the known results are for a single warehouse supplying parallel retailers, the focus of our attention in this section.

The echelon base-stock policies previously described also play a role in distribution systems. In the two-echelon system consisting of a single warehouse and multiple retailers, stocking decisions need to be made at all installations, but additionally it is possible that *allocation* decisions need to be made when scarce warehouse inventory needs to be distributed across needy retailers. Optimal results can be found when there are no differences between the retailers in terms of demands, lead times, and holding and stock-out costs, and these optimal policies can perform well when there are differences between the retailers. Federgruen [20] distinguishes between the models when inventory is held at the warehouse and when it is not. When no inventory is stocked at the warehouse, the total amount drawn by the retailers will equal the quantity ordered by the warehouse (in a previous period), but this is not necessarily so when some remnant stock may be left at the warehouse. For each of these warehouse with/without inventory models, Federgruen [20] describes a “relaxed” and a “restricted” model to provide lower and upper bounds of the optimal policy. The relaxed model for the warehouse without inventory model consists of allowing a negative value of the allocation to particular retailers, enabling a vast simplification of the formulation so that the sum of the individual retailer inventories replaces the individual quantities in the state

space. Allowing negative allocations amounts to assuming transshipment between retailers can be carried out (essentially) costlessly. When the retailers’ cost parameters are identical, the myopic (Newsvendor) solution is then optimal for the retailers and an echelon base-stock decision (simple base-stock or (s, S) if there is a fixed ordering cost) is optimal for the warehouse. The restricted approach consists of applying a specific but parsimonious order policy, such as following an echelon base-stock policy for the warehouse when only ordering in certain periods (e.g., every m periods). Similarly, particular allocation policies have been examined including the myopic (each retailer solves the Newsvendor) and cyclic allocation policies. Knowing that the optimal policy lies between these relaxed and restricted models on a system-wide cost basis, the gap between these approximations can give an idea of the degree of suboptimality of following these approaches, given that finding the true optimal policy is impractical. For the warehouse with inventory model, the relaxed approach (removing the nonnegativity constraint for retailer orders) with myopic retailer allocation results in an even better solution than for the warehouse without inventory model. Again, the relaxation and restriction approach with appropriate allocation policies in conjunction with echelon base-stock policies tends to perform well.

FURTHER SOURCES

Zipkin [21], Porteus [22], and Axsäter [23] serve as excellent references for inventory theory under stochastic demand. Excellent summaries of the literature exist in Federgruen [20], Axsäter [24], and Porteus [25].

REFERENCES

1. Scarf H. The optimality of (s, S) policies in the dynamic inventory problem. In: Arrow K, Karlin S, Suppes P, editors. *Mathematical methods in the social sciences*. Palo Alto (CA): Stanford University Press; 1960.
2. Heyman DP, Sobel MJ. *Stochastic models in operations research*. Volume II, New York (NY): McGraw-Hill; 1984.

3. Zipkin PH. Old and new methods for lost-sales inventory systems. *Oper Res* 2008; 56(5):1256–1263.
4. Kaplan R. A dynamic inventory model with stochastic lead times. *Manage Sci* 1970; 16:491–507.
5. Federgruen A, Zipkin PH. An inventory model with limited production capacity and uncertain demands I: the average-cost criterion. *Math Oper Res* 1986a;11(2):193–207.
6. Federgruen A, Zipkin PH. An inventory model with limited production capacity and uncertain demands II: the discounted-cost criterion. *Math Oper Res* 1986b;11(2):208–215.
7. Hadley G, Whitin TM. *Analysis of inventory systems*. Englewood Cliffs (NJ): Prentice-Hall; 1963.
8. DeHoratius N, Raman A. Inventory record inaccuracy: an empirical analysis. *Manage Sci* 2008;54(4):627–641.
9. Clark AJ, Scarf H. Optimal policies for a multi-echelon inventory problem. *Manage Sci* 1960;6:475–490.
10. Federgruen A, Zipkin PH. Computational issues in an infinite-horizon multiechelon inventory model. *Oper Res* 1984;32:818–836.
11. Bellman R. *Adaptive control processes: a guided tour*. Princeton (NJ): Princeton University Press; 1961.
12. Chen F, Zheng YS. Evaluating echelon stock (R, nQ) policies in serial production/inventory systems with stochastic demand. *Manage Sci* 1994a;40(10):1262–1275.
13. Shang KH, Song J-S. Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Manage Sci* 2003;49(5):618–638.
14. Gallego G, Zipkin PH. Stock positioning and performance estimation in serial production-transportation systems. *Manuf Serv Oper Manage* 1999;1(1):77–88.
15. Chen F, Zheng YS. Lower bounds for multi-echelon stochastic inventory systems. *Manage Sci* 1994b;40(11):1426–1443.
16. Muharremoglu A, Tsitsiklis JN. A single-unit decomposition approach to multiechelon inventory systems. *Oper Res* 2008; 56(5):1089–1103.
17. Cachon GP, Zipkin PH. Competitive and cooperative inventory policies in a two-stage supply chain. *Manage Sci* 1999;45(7):936–953.
18. Parker RP, Kapuscinski R. Optimal policies for a capacitated two-echelon inventory system. *Oper Res* 2004;52(5):739–755.
19. Rosling K. Optimal inventory policies for assembly systems under random demands. *Oper Res* 1989;37:565–579.
20. Federgruen A. Centralized planning models. In: Graves SC, Rinnooy Kan AHG, Zipkin PH, editors. *Handbooks in operations research and management science: logistics of production and inventory*. Amsterdam: Elsevier; 1993.
21. Zipkin PH. *Foundations of inventory management*. New York (NY): McGraw Hill; 2000.
22. Porteus EL. *Foundations of stochastic inventory theory*. Palo Alto (CA): Stanford University Press; 2002.
23. Axsater S. *Inventory control*. Norwell (MA): Kluwer; 2000.
24. Axsater S. Serial and distribution inventory systems. In: de Kok AG, Graves SC, editors. *Handbooks in operations research and management science: supply chain management: design, coordination, and operations*. Amsterdam: Elsevier; 2003.
25. Porteus EL. Stochastic inventory theory. In: Heyman DP, Sobel MJ, editors. *Handbooks in operations research and management science: stochastic models*. Amsterdam: Elsevier; 1990.

MULTIMETHODOLOGY

JOHN MINGERS
Kent Business School,
University of Kent,
Canterbury, Kent, UK

Operations research developed during and after World War II. It was based on adopting a scientific approach toward decision making in organizations, and many specific techniques, usually quantitative or mathematical, were developed. However, by the 1970s and 1980s, practitioners had found that there were many situations, or perhaps aspects of situations, that could not be adequately represented by mathematical or computer-based modeling [1]. This led to the development of several new “softer” methods that took the human or subjective dimension of problem situations seriously, while still being rigorous and systemic. Examples of such methods are soft systems methodology (SSM) [2], cognitive mapping and journey making [3] (see *Cognitive Mapping and Strategic Options Development and Analysis (SODA)*), strategic assumption surfacing and testing (SAST) [4], strategic choice approach (SCA) [5], and critical systems heuristics (CSH) [6]. These are sometimes called *problem-structuring methods* [7] (see *Problem Structuring for Multicriteria Decision Analysis Interventions*) or more recently *qualitative OR*.

However, this plethora of methods, both hard and soft, led to difficulties in knowing which was the most appropriate one to use in a particular situation. To try and help with this issue, Jackson and Keys [8] developed a framework of “problem contexts”—the *system of systems methodologies*. This proposed that there were two primary dimensions to be considered in looking at the context of a problem—the degree of agreement among participants about the nature of the problem and the objectives to be pursued, and the degree of complexity of the problem. The first

dimension was split into three possibilities: *unitary* situations, where there was a high degree of consensus among participants; *pluralistic* situations, where there were different viewpoints or stakeholder interests but a negotiated agreement was seen as possible; and *coercive* situations, where there were significant disagreements that could only be resolved by the exercise of power by one party over another. The second dimension was split into *simple* and *complex*, although in reality this is really a continuum. Simple problems are reasonably well defined and the nature of the problem is clear, although there could be a high degree of combinatorial difficulty. Complex problems are hard to define, difficult to bound, and resist analysis. They have also been called *wicked problems* by Rittel and Weber [9].

This yielded six cells into which OR methods or methodologies could be placed, thus giving guidance on which one should be used in a particular context. For example, traditional, hard, mathematical OR models were most suitable in simple unitary contexts, where the nature of the problem was clear and there was agreement regarding the objectives. System dynamics was a possibility in complex unitary contexts involving dynamic, feedback-driven situations. SSM was seen as appropriate for complex pluralistic contexts in which there were significant differences in viewpoints. This framework was later developed into a methodology known as *total systems intervention* (TSI) [10].

Although a step forward, the framework had several limitations [11]. In particular, it assumed that problems could be clearly identified as being of a particular type; that methods fitted neatly into the cells; and therefore a practitioner just needed to choose an appropriate method to deal with their particular type of problem. While it may sometimes be the case that a very well-defined problem requires a clear and unique method of solution, it can be argued that generally problem situations within real-world organizations are complex with

several dimensions—technical, social, and personal. Also, projects go through several phases—from initial discovery and problem shaping, through analysis, to eventual implementation—and different methods may be helpful at certain stages and not others. Finally, methods themselves may be used in different ways—a system dynamics model, for example, could be seen as a model of objective reality, or a model of an individual's personal mental model.

This leads to the idea of using multiple methods—*multimethodology* [12,13]—that is, deliberately seeking to combine together a range of methods, perhaps both soft and hard, in order to match the richness of the problem situation and to deal effectively with the different stages of a project. In doing this, practice was ahead of theory: a survey of SSM users [14] unexpectedly found SSM routinely combined with other methods, and Ormerod [15–17] has documented a number of sophisticated multimethodology interventions from the early 1990s.

ARGUMENTS FOR MULTIMETHODOLOGY

Adopting a particular method or methodology is like viewing the world through an instrument such as a telescope, an X-ray machine, or an electron microscope. Each reveals certain aspects of the situation but is completely blind to others. Although they may be pointing at the same place, each instrument produces a different, and sometimes seemingly incompatible, representation. Thus, in adopting only one method one is inevitably gaining only a limited view of a particular problem situation. For example, mathematical programming (see *An Introduction to Linear Programming*) focuses on factors that can be quantified and relationships (often assumed to be linear) between them; discrete event simulation (see *Introduction to Discrete-Event Simulation*) looks at the frequency of occurrence of particular events; SSM models notional systems of human activity within a social context; and critical systems heuristics (CSH) draws attention to the power of experts in defining boundaries and problems.

It has been suggested [18] that there are three fundamental dimensions of importance—the material world, the social world, and the personal world. Thus, any real-world situation into which we are intervening or researching will be a complex interaction of substantively different elements. There will be aspects that are relatively hard and observer-independent, particularly material and physical processes, which we can observe and model. There will be aspects that are socially constituted, dependent on particular cultures, social practices, languages, and power structures, which we must come to share and participate in. Finally, there will be aspects that are individual, such as beliefs, values, fears, and emotions, which we must try to express and understand. An effective intervention should therefore combine methods that address all three domains.

The second argument is that an OR project is not a discrete event but a process that has phases or different types of activities predominating at different times. Particular methodologies and techniques are more useful for some phases than others and so a combination of approaches may be necessary to provide a comprehensive outcome. The following four phases can be identified:

- *Appreciation* of the situation as experienced by the practitioners involved and expressed by participants in the situation. This involves an initial identification of the concerns to be addressed; conceptualization and design of the study; and the production of basic data using methods such as observation, interviews, experiments, surveys, or qualitative approaches.
- *Analysis* of the information produced so as to be able to understand and explain why the situation is as it is. What are the underlying structures and constraints that shape the situation? This involves analysis methods appropriate to the goal(s) of the intervention and the information produced in the first stage.
- *Assessment* of possible explanations and of potential changes to the situation. This involves the generation of

alternative arrangements or courses of action and an evaluation of their desirability and feasibility.

- *Action* to bring about changes, if necessary or desired, or to report on and disseminate the results if the project is simply research.

Put crudely, these phases cover: What is happening? Why it is happening? How could the situation be changed? And, what shall we do? At the beginning of an intervention, especially for an agent from outside the situation, the primary concern is to gain as rich an appreciation of the situation as possible. Note that this cannot be an “observer-independent” view of the situation “as it really is.” It is conditioned by the researchers’ previous experiences and their access to the situation. The next activity is to begin to analyze why the situation is as it appears, to understand the history that has generated it, and the particular structure of relations and constraints that maintain it. Next, in cases where changes to the situation are sought, consideration must be given to ways in which the situation could be changed. This means focusing attention away from how things are, and considering the extent to which the structures and constraints can be changed within the general limitations of the intervention. Finally, action must be undertaken that will effectively bring about agreed changes. It should be emphasized that these activities are not discrete stages that are enacted one by one; rather, they are aspects of the intervention that need to be considered throughout, although their relative importance will differ as the project progresses.

It is clear that the wide variety of methods and techniques available do not all perform equally well in all these activities. Some brief examples are as follows: collecting data, administering questionnaires and surveys, developing rich pictures and cognitive maps, and employing the 12 CSH questions, all these contribute toward finding out about the different aspects of a particular situation (Appreciation), whereas building simulation or mathematical models, constructing root definitions and conceptual models

(SSM), using role-playing and gaming, or undertaking participant observation help in understanding why the situation is as it is (Analysis) and in evaluating other possibilities (Assessment).

MULTIMETHODOLOGY IN ACTION

Multimethodology, when in use, is a creative process of design based on competence in a range of methods. Each project or intervention is seen as a unique situation (although it has features in common with others) for which a particular combination of methods, or parts of methods, needs to be constructed. This is an on-going process throughout the project, as events occur and the situation evolves. The general process can be described as follows:

Reflection:

- *review* the current situation;
- *determine* which areas of the problem situation currently need to be addressed.

Design:

- *understand* what methods or techniques could possibly be useful;
- *choose* the most appropriate method to use in relation to the project context as a whole.

There are several general approaches to combining together methods and methodologies, and more guidance, both for practical interventions and for research, can be found in Mingers [19], Mingers [20], Jackson [21], and Midgley [22]. The idea of mixed methods is not just confined to OR but is now also common in social science research in general [23,24]

Munro and Mingers [25] undertook a survey of examples of multimethodology in practice, which showed that the examples were common and successful. Drawing from this survey, and from other sources, we can highlight particular combinations of methods that have been shown to work well together.

- Cognitive mapping is a general approach to capturing peoples' thinking about particular complex issues and as such is compatible with many other methods. It is particularly synergistic with systems dynamics (SD), as a map can be converted into an SD model; it is also often used in the early stages of SSM to enhance the appreciation of the problem situation; and it has been used with Delphi, scenarios, and strategic choice.
- SSM itself is very flexible and can be used to structure the whole intervention. It is often used as the dominant method augmented by other techniques. It has been used extensively in information systems development [26], both as a "front-end" to harder, structured systems analysis techniques, and as the controlling method throughout the systems development.
- Strategic choice can also be combined with SSM, particularly in the later stages of an intervention when decisions about, and commitments to, action are being negotiated.
- The viable systems model (VSM) [27] is useful when there is organizational change as it focuses attention on the necessary structural and communication features of an effective organization. It has regularly been used with SSM and CSH.
- It is also interesting to see techniques being used out of their usual setting, such as a hard technique being used in a soft way. For example, quantitative models such as mathematical programming or system dynamics (SD) are usually assumed to be objective, that is, models of *independent reality*. However, it is possible to see them as subjective, as models of *particular concepts or views of possible realities*. They can then be used within, say, SSM, to help articulate a debate among rival positions.

EXAMPLES

To illustrate the approach, we note two examples that mix together hard and soft approaches. The first is a combination of mathematical modeling, in this case data envelopment analysis (DEA), and SSM [28]. The project was to develop a set of indicators to be used in evaluating the performance of the 80 research institutes of the Chinese Academy of Sciences (CAS). The initial plan was to use DEA to compare the performance of the institutes. However, it proved very difficult to get agreement about the real purpose of the evaluation and which indicators to use. At this point, SSM was brought in to facilitate a debate between the different participants, and provide a structured method for generating appropriate indicators. The combination was found to be very successful and has now been developed into a general methodology for performance evaluation in public sector organizations [29].

In the second example, cognitive mapping was combined with SD and statistical analysis in order to model how delays had occurred within a major construction project—the Channel Tunnel [30]. A contractor was suing a client for the extra costs caused by late design approvals and specification changes. A system dynamics model was required that could be used in court to support the claim and quantify the extent of the costs. The complexity of the project and the various different groups and professions involved made it very difficult to begin the construction of an SD model. So the project began by using cognitive mapping to capture the different viewpoints of the senior staff. A single group map was agreed upon with those concerned. This group map was maintained and updated throughout the project forming invaluable documentation "behind" the SD model. This map was then converted into an influence diagram by picking out the main (98) feedback loops. These were often interrelated in clusters.

The model was decomposed into two main areas: the design and acquisition-of-resources processes. From these, the SD models were developed, first structurally and

then in a quantified form. Inevitably, there were many problems with data availability and reliability. Much spreadsheet work was involved. Often inconsistencies were discovered. These had to be resolved by further discussion and development of both the cognitive map and the SD model. The SD model became very large (over 350 variables) and rather opaque, yet it needed to be clearly understood by a judge and barristers. This was helped by the COPE model (software for cognitive mapping), which was much easier to understand (as it was text based) but still mapped onto the SD one. The final result was developed through a constant interaction between the two. In the end, neither model actually appeared in court as the case was settled by agreement after a detailed audit of the models by both sides.

REFERENCES

1. Ackoff R. The future of operational research is past. *J Oper Res Soc* 1979;30:93–104.
2. Checkland P, Scholes J. *Soft systems methodology in action*. Chichester: Wiley; 1990.
3. Eden C, Ackerman F. *Making strategy: the journey of strategic management*. London: Sage; 1998.
4. Mason R, Mitroff I. *Challenging strategic planning assumptions*. New York: Wiley; 1981.
5. Friend J, Hickling A. *Planning under pressure: the strategic choice approach*. 3rd ed. Oxford: Pergamon; 2004.
6. Ulrich W. *Critical heuristics of social planning: a new approach to practical philosophy*. 2nd ed. Chichester: Wiley; 1994.
7. Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited*. Chichester: Wiley; 2001.
8. Jackson M, Keys P. Towards a system of system methodologies. *J Oper Res Soc* 1984;35(6): 473–486.
9. Rittel H, Webber M. Dilemmas in a general theory of planning. *Policy Sci* 1973;4: 155–169.
10. Flood R, Jackson M. *Creative problem solving: total systems intervention*. London: Wiley; 1991.
11. Mingers J. Towards critical pluralism. In: Mingers J, Gill A, editors. *Multimethodology: theory and practice of combining management science methodologies*. Chichester: Wiley; 1997. pp. 407–440.
12. Mingers J, Brocklesby J. Multimethodology: towards a framework for mixing methodologies. *Omega* 1997;25(5):489–509.
13. Mingers J, Gill A, editors. *Multimethodology: theory and practice of combining management science methodologies*. Chichester: Wiley; 1997.
14. Mingers J, Taylor S. The use of soft systems methodology in practice. *J Oper Res Soc* 1992;43(4):321–332.
15. Ormerod R. Putting soft OR methods to work: information systems strategy development at Sainsbury's. *J Oper Res Soc* 1995;46(3): 277–293.
16. Ormerod R. On the nature of OR - entering the fray. *J Oper Res Soc* 1996;47(1):1–17.
17. Ormerod R. Information systems strategy development at Sainsbury's supermarkets using "Soft" OR. *Interfaces* 1996;26(1): 102–130.
18. Mingers J. Multi-paradigm multimethodology. In: Mingers J, Gill A, editors. *Multimethodology: theory and practice of combining management science methodologies*. Chichester: Wiley; 1997. pp. 1–20.
19. Mingers J. Variety is the spice of life: combining soft and hard OR/MS methods. *Int Trans Oper Res* 2000;7:673–691.
20. Mingers J. *Realising systems thinking: knowledge and action in management science*. New York: Springer; 2006.
21. Jackson M. *Systems thinking: creative holism for managers*. Chichester: Wiley; 2003.
22. Midgley G. *Systemic intervention: philosophy, methodology, and practice*. New York: Kluwer. Plenum; 2000.
23. Tashakkori A, Teddlie C. *Mixed methodology: combining qualitative and quantitative approaches*. London: SAGE Publications; 1998.
24. Tashakkori A, Teddlie C. *Handbook of mixed methods in social and behavioural research*. Thousand Oaks (CA): Sage; 2003.
25. Munro I, Mingers J. The use of multimethodology in practice - results of a survey of practitioners. *J Oper Res Soc* 2002;59(4):369–378.
26. Stowell F, editor. *Information systems provision: the contribution of soft systems methodology*. London: McGraw Hill; 1995.

27. Espejo R, Harnden R. The viable systems model: interpretations and applications of stafford Beer's VSM. Chichester: Wiley; 1989.
28. Mingers J, Liu W, Meng W. Using SSM to structure the identification of inputs and outputs in DEA. *J Oper Res Soc* 2009;60: 168–179.
29. Liu W, Cheng Z, Mingers J, *et al.* A methodology for developing performance indicators for public sector organisations. *Public Money and Management*; 2010. In press.
30. Ackerman F, Eden C, Williams T. Modeling for litigation: mixing qualitative and quantitative approaches. *Interfaces* 1997;27(2):48–65.

MULTISTAGE (STOCHASTIC) GAMES

NAGARAJAN KRISHNAMURTHY
Chennai Mathematical Institute,
Chennai, India

THIRUVENKATACHARI PARTHASARATHY
Indian Statistical Institute, Chennai,
India

Shapley introduced the concept of stochastic games in his seminal paper [1] in 1953. In a stochastic game, players play noncooperatively and simultaneously in stages (discrete points in time). Shapley defined two-person zero-sum (TPZS) discounted finite stochastic games as a generalization of matrix games or two-person zero-sum (TPZS) games. Refer the section titled “Two-Person Zero-Sum (TPZS) Games” for TPZS Games. A TPZS discounted finite stochastic game consists of a finite number of states, finite sets of actions of the players, a payoff matrix in each state, and transition probabilities from each state to every other state, for every pair of actions of the players. As in the case of TPZS games, one of the players chooses rows and the other player chooses columns. We shall assume that the row chooser is the maximizer and the column chooser is the minimizer. Given a starting state, the players simultaneously choose actions resulting in an immediate payoff (the corresponding entry in the payoff matrix) that is paid to the row chooser by the column chooser and the game moves to a new state depending on the transition probabilities. Now, the players choose actions in the new state, resulting in a payoff in that state, and so on. At each time period, these payoffs are successively discounted by a factor $\beta \in [0, 1)$ and these discounted payoffs are accumulated over the infinite horizon. Starting at different states, we obtain different accumulated discounted payoffs. The aim of the row chooser is to maximize these accumulated discounted payoffs, and the aim of the column chooser is to minimize the same. Strategies of players are probability

distributions over their action sets, at each time period.

Shapley originally defined discounted stochastic games as stochastic games with positive stop probabilities. That is, given any state, the probability that the game stops is positive. It follows that the game stops with a probability 1 after a finite number of steps. It is easy to see that “discounted stochastic games” and “stochastic games with positive stop probabilities” are equivalent. Gillette [2,3] introduced stochastic games with zero stop probabilities or equivalently, stochastic games with undiscounted (or limiting average) payoffs. Undiscounted payoffs are *lim sup* or *lim inf* of accumulated average payoffs over the infinite run.

As we discussed above, zero-sum stochastic games can be considered as a generalization of TPZS games. Nonzero-sum stochastic games are a generalization of one shot noncooperative strategic-form (NCSF) games. Concepts of von Neumann’s minimax value (see *Saddlepoints and von Neumann Minimax Theorem*) and Nash equilibria [4] (see *Nash Equilibrium (Pure and Mixed)*) can be extended to stochastic games as well. Stochastic games may also be considered as a generalization of Markov decision processes (MDPs) (see the section titled “Markov Decision Process (MDP)” in this encyclopedia) with two or more players whose payoffs and the transition probabilities depend on the choice of actions of the players. Stochastic games are a generalization of repeated games as well. Repeated games are stochastic games with just one state.

Following is an example of a zero-sum stochastic game.

Example 1. The following game has two states s_1 and s_2 . Both players have two actions in state s_1 . In s_2 , player 1 (the row player/ maximizer) has three actions and player 2 (the column player/ minimizer) has two actions. The payoffs and transition probabilities are given below.

$$s_1 : \begin{bmatrix} 2 & 1 \\ (\frac{1}{4}, \frac{3}{4}) & (0, 1) \\ 0 & 3 \\ (\frac{1}{4}, \frac{3}{4}) & (0, 1) \end{bmatrix}, s_2 : \begin{bmatrix} 0 & 2 \\ (\frac{1}{2}, \frac{1}{2}) & (\frac{1}{2}, \frac{1}{2}) \\ 2 & 1 \\ (1, 0) & (1, 0) \\ 3 & 0 \\ (0, 1) & (0, 1) \end{bmatrix}.$$

In each row, the entries on the top are the payoffs. The entries below the payoffs (enclosed in parentheses) are the transition probabilities to states s_1 and s_2 , respectively. For example, in state s_1 , if the players choose row 1 and column 1, the row chooser gets an immediate payoff of 2 from the column chooser. The game remains in state s_1 with probability $\frac{1}{4}$ and transitions to state s_2 with probability $\frac{3}{4}$.

We formally define two-player finite stochastic games in the following section.

TWO-PERSON FINITE STOCHASTIC GAMES

In this section, we define TPZS and two person nonzero-sum finite stochastic games with discounted as well as undiscounted payoffs. We define stationary strategies, Markov strategies, and behavioral (or history-dependent) strategies. For zero-sum stochastic games, we define optimal strategies and optimal value, and we state Shapley's theorem for the discounted case. We define Nash equilibrium strategies and payoffs for the nonzero-sum case.

Definition 1 [Two-player, finite stochastic game]. A two-player, finite state space, finite action space stochastic game consists of

1. Two players P_1 and P_2 . We shall sometimes refer to them just as players 1 and 2.
2. A finite, nonempty set of states, $S = \{1, 2, \dots, N\}$.
3. For each state $s \in S$, finite, nonempty sets $A_1(s) = \{1, 2, \dots, m_1(s)\}$ and $A_2(s) = \{1, 2, \dots, m_2(s)\}$ of actions for players P_1 and P_2 , respectively. Without loss of generality, we may assume $A_1(s) = A_1$

and $A_2(s) = A_2$, $\forall s \in S$. (This assumption is only for ease of analysis and may be dropped).

4. Immediate rewards $r_1(s, i, j)$ for player 1 and $r_2(s, i, j)$ for player 2, where $s \in S, i \in A_1, j \in A_2$, when the game is in state s and the players choose actions i and j , respectively. If $r_2 = -r_1$, then, we have a zero-sum game. Otherwise, the game is a nonzero-sum game. We will use r in place of r_1 in case of zero-sum games. We will denote the matrix of immediate rewards in state s by $R_1(s)$ and $R_2(s)$ for players 1 and 2, respectively (and by $R(s)$ in case of zero-sum games).
5. Transition probabilities $(q(s'|s, i, j) : (s, s') \in S \times S, i \in A_1, j \in A_2)$ where $q(s'|s, i, j)$ is the probability of transition from state s to state s' given that players 1 and 2 choose actions $i \in A_1, j \in A_2$, respectively. These transition probabilities constitute the "law of motion" of the game. We will use $q(s, i, j)$ to denote the corresponding probability distribution. Given $i \in A_1, j \in A_2$, we denote the $N \times N$ transition matrix by $Q(i, j)$.

The game proceeds as follows. Given a starting state $s_0 \in S$, the players simultaneously choose actions $i_0 \in A_1$ and $j_0 \in A_2$, resulting in payoffs of $r_1(s_0, i_0, j_0)$ and $r_2(s_0, i_0, j_0)$ to players 1 and 2, respectively. The game moves to a new state s_1 according to the law of motion $q(s_0, i_0, j_0)$, and the players choose actions $i_1 \in A_1$ and $j_1 \in A_2$ resulting in payoffs of $r_1(s_1, i_1, j_1)$ and $r_2(s_1, i_1, j_1)$, and so on.

In general, strategies can depend on complete histories of the game until the current stage. Such strategies are called *behavioral strategies*. *Markov (or memoryless) strategies* are behavioral strategies that depend only on the current state and the current stage (or time period). *Stationary strategies* are a simpler class of strategies, which depend only on the current state s and not on how s was reached. In other words, stationary strategies are time-independent Markov strategies. For

discounted finite stochastic games, it suffices to look at stationary strategies.

For player 1, a stationary strategy f is a function from S to P_{A_1} , where S is the state space and P_{A_1} is the set of probability distributions on player 1's action set A_1 . Similarly, we define a stationary strategy g for player 2 as a function from S to P_{A_2} . Alternatively, we can look at f as an N -tuple of probability distributions from P_{A_1} , one distribution per state. In other words, $f \in P_{A_1}^N$ where N is the number of states. Similarly $g \in P_{A_2}^N$. *Pure stationary strategies* are simply a set of actions, one per state.

Remark. Strategies for the players determine the distribution of the stochastic process of states and actions.

In the following definitions, $r_t^{(1)}(s_0, f, g)$ and $r_t^{(2)}(s_0, f, g)$ are the expected immediate rewards at the t th stage, to players P_1 and P_2 , respectively.

Definition 2 [β -discounted payoffs]. In the nonzero-sum case, given an initial state s_0 , a pair of stationary strategies (f, g) of players 1 and 2, and a discount factor $\beta \in (0, 1)$, we define β -discounted payoffs as follows:

$$I_\beta^{(1)}(f, g)(s_0) = \sum_{t=0}^{\infty} \beta^t r_t^{(1)}(s_0, f, g) \text{ for } P_1 \text{ and} \quad (1)$$

$$I_\beta^{(2)}(f, g)(s_0) = \sum_{t=0}^{\infty} \beta^t r_t^{(2)}(s_0, f, g) \text{ for } P_2. \quad (2)$$

In the zero-sum case,

$$I_\beta^{(1)}(f, g)(s_0) = -I_\beta^{(2)}(f, g)(s_0). \quad (3)$$

(We shall denote $I_\beta^{(1)}$ by I_β in this case).

Definition 3 [Undiscounted payoffs].

In the nonzero-sum case, given an initial state s_0 , and a pair of stationary strategies (f, g) , of players 1 and 2, we define the

undiscounted (or limiting average) payoffs as follows:

$$\Phi^{(1)}(f, g)(s_0) = \liminf_{T \uparrow \infty} \left[\left(\frac{1}{T+1} \right) \times \sum_{t=0}^T r_t^{(1)}(s_0, f, g) \right] \text{ for } P_1 \text{ and} \quad (4)$$

$$\Phi^{(2)}(f, g)(s_0) = \liminf_{T \uparrow \infty} \left[\left(\frac{1}{T+1} \right) \times \sum_{t=0}^T r_t^{(2)}(s_0, f, g) \right] \text{ for } P_2. \quad (5)$$

In the zero-sum case,

$$\Phi^{(1)}(f, g)(s_0) = -\Phi^{(2)}(f, g)(s_0). \quad (6)$$

(We shall denote $\Phi^{(1)}$ by Φ in this case).

We write $\Gamma_\beta = (S, A_1, A_2, r_1, r_2, q, \beta)$ ($\Gamma = (S, A_1, A_2, r_1, r_2, q)$) to denote a β -discounted (undiscounted) nonzero-sum stochastic game with set of states S , sets of actions A_1 and A_2 , and rewards r_1 and r_2 as defined above. Whenever $r_1 = -r_2$ ($=r$, say), we have a zero-sum game.

Definition 4 [Optimal strategies and optimal value]. A pair of stationary strategies (f^*, g^*) is optimal in the zero-sum discounted case, if for all $s \in S$

$$I_\beta(f, g^*)(s) \leq I_\beta(f^*, g^*)(s) \leq I_\beta(f^*, g)(s) \quad \forall f \in P_{A_1}^N, \forall g \in P_{A_2}^N. \quad (7)$$

(assuming that player 1 is the maximizer and player 2 is the minimizer).

In other words, $I_\beta(f^*, g^*)(s) = \inf_g [I_\beta(f^*, g)(s)] = \sup_f [I_\beta(f, g^*)(s)] \forall s \in S$.

Shapley [1] proved the existence and uniqueness (across all pairs of optimal strategies (f^*, g^*)) of the optimal value vector $I_\beta(f^*, g^*)$, denoted by v_β . That is, when the game starts at state $s \in S$, $v_\beta(s)$ is unique and this is true for all $s \in S$. That is,

$$\begin{aligned} v_\beta(s) &= I_\beta(f^*, g^*)(s) = \sup_f \inf_g [I_\beta(f, g)(s)] \\ &= \inf_g \sup_f [I_\beta(f, g)(s)] \forall s \in S. \end{aligned}$$

Note that there may be different pairs of optimal strategies (f^*, g^*) .

A pair of stationary strategies (f^*, g^*) is optimal for the undiscounted zero-sum game, if for all $s \in S$

$$\Phi(f, g^*)(s) \leq \Phi(f^*, g^*)(s) \leq \Phi(f^*, g)(s) \quad \forall f \in P_{A_1}^N, g \in P_{A_2}^N. \quad (8)$$

(assuming player 1 is the maximizer and player 2 is the minimizer).

Mertens and Neyman [5] proved the existence and uniqueness of the optimal value vector in the undiscounted case. Both players may not have optimal strategies though, as we shall see soon. We write v to denote the optimal value vector Φ of the undiscounted stochastic game, which is a function of the initial state s . That is, for all $s \in S$, $\Phi(s) = \sup_f \inf_g [\Phi(f, g)(s)] = \inf_g \sup_f [\Phi(f, g)(s)]$.

Definition 5 [Nash equilibrium]. A pair of stationary strategies (f^*, g^*) constitutes a Nash equilibrium in the discounted case if $\forall s \in S$

$$I_\beta^{(1)}(f^*, g^*)(s) \geq I_\beta^{(1)}(f, g^*)(s) \text{ for each } f \in P_{A_1}^N \text{ and} \quad (9)$$

$$I_\beta^{(2)}(f^*, g^*)(s) \leq I_\beta^{(2)}(f^*, g)(s) \text{ for each } g \in P_{A_2}^N. \quad (10)$$

(assuming that both players want to maximize their payoffs).

Similarly, in the undiscounted case, a pair of stationary strategies (f^*, g^*) constitutes a Nash equilibrium, if $\forall s \in S$

$$\Phi^{(1)}(f^*, g^*)(s) \geq \Phi^{(1)}(f, g^*)(s) \text{ for each } f \in P_{A_1}^N \text{ and} \quad (11)$$

$$\Phi^{(2)}(f^*, g^*)(s) \leq \Phi^{(2)}(f^*, g)(s) \text{ for each } g \in P_{A_2}^N. \quad (12)$$

(assuming that both players want to maximize their payoffs).

[We can extend these definitions to the n -player case too ($n > 2$)].

Now, we discuss Shapley's theorem [1] regarding the existence of optimal value and optimal stationary strategies in zero-sum discounted finite stochastic games. Informally, the value of the stochastic game starting at state s is the value of the matrix game obtained by adding to the immediate payoffs at s , the optimal expected discounted rewards from the next time period onwards. This leads to a system of equations, solving which we obtain the value vector. We formally state Shapley's theorem below.

Theorem 1 [Shapley [1]] . A two-player zero-sum discounted stochastic game $\Gamma_\beta = (S, A_1, A_2, r, q, \beta)$ has an optimal value vector v_β which is the unique solution of the following system of equations

$$v(s) = \text{val}[R(s, v)]$$

for all $s \in S$, where $R(s, v)$ is the auxiliary matrix game with (i, j) th entry $r(s, i, j) + \beta \sum_{s' \in S} q(s' | s, i, j) v(s')$ and $\text{val}[R(s, v)]$ is the minimax value of the matrix game.

For each state $s \in S$, if $(f^*(s), g^*(s))$ is a pair of optimal strategies of the matrix game $R(s, v_\beta)$, then (f^*, g^*) is a pair of optimal strategies for the discounted stochastic game Γ_β , where $f^* = (f^*(1), f^*(2), \dots, f^*(N))$ and $g^* = (g^*(1), g^*(2), \dots, g^*(N))$.

The proof of Shapley's theorem follows from the fact that $L : R^N \rightarrow R^N$ defined by $L(v)(s) = \text{val}[R(s, v)]$ for all $v \in R^N, s \in S$, is a contraction map (as $\beta \in [0, 1)$). Hence, by Banach's fixed point theorem, L has a unique fixed point. The unique solution of $L(v) = v$ gives the optimal value vector v_β of the stochastic game. We implicitly use the fact that a strategy is optimal versus all strategies if it is optimal versus stationary strategies.

In the following example, we illustrate the use of Shapley's theorem in solving a zero-sum discounted stochastic game.

Example 2. The game has two states s_1 and s_2 . Both players have two actions in s_1

and one action each in s_2 . The payoffs and transition probabilities are given below:

$$s_1 : \begin{bmatrix} 2 & 0 \\ (0, 1) & (0, 1) \\ 0 & 2 \\ (1, 0) & (0, 1) \end{bmatrix}, \quad s_2 : \begin{bmatrix} 0 \\ (0, 1) \end{bmatrix}.$$

Here, we represent the (i, j) th entry in the matrix corresponding to state s_k as follows:

$$\begin{bmatrix} r(s_k, i, j) \\ (q(s_1|s_k, i, j), q(s_2|s_k, i, j)) \end{bmatrix}.$$

Using Shapley's theorem, optimal value of this stochastic game starting at state s_1 is

$$v_\beta(1) = \text{val} \begin{bmatrix} 2 + \beta v_\beta(2) & 0 + \beta v_\beta(2) \\ 0 + \beta v_\beta(1) & 2 + \beta v_\beta(2) \end{bmatrix}. \quad (13)$$

For brevity, we shall write v_s in place of $v_\beta(s)$. Let $\beta = \frac{1}{2}$ be the discount factor. Then

$$v_1 = \text{val} \begin{bmatrix} 2 + \frac{1}{2}v_2 & \frac{1}{2}v_2 \\ \frac{1}{2}v_1 & 2 + \frac{1}{2}v_2 \end{bmatrix}. \quad (14)$$

In state s_2 , both players have no choice but to play their only actions available. Further, s_2 is an absorbing state. In other words, once the game reaches s_2 , it remains in s_2 forever. Therefore, value of the game starting at state s_2 is $v_2 = 0$.

It is easy to check that the auxiliary matrix game corresponding to state 1 has no pair of pure optimal strategies. (That is, the players cannot play optimally by choosing actions with probability 1, and they must choose mixed strategies, or probability distributions over their respective action sets). Therefore, we can use Kaplansky's theorem [6] to solve the game.

We state Kaplansky's theorem below and omit the proof.

Theorem 2 [Kaplansky [6]]. *Given a 2×2 matrix game*

$$M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

that does not have a pure optimal strategy pair, the value of the game is

$$\text{val}(M) = (m_{11}m_{22} - m_{21}m_{12}) / (m_{11} + m_{22} - m_{12} - m_{21}).$$

Using Kaplansky's theorem, we get

$$v_1 = 4 / (4 - \frac{1}{2}v_1)$$

solving which we get

$$v_1 = 4 \pm 2\sqrt{2}.$$

$v_1 = 4 + 2\sqrt{2}$ is not possible as $v_1 \leq 4$ (since, at each stage of the game, even if player 1 gets the maximum possible payoff of 2, $v_1 = 2 + \beta v_1$ which gives $v_1 = 4$).

Therefore, $v_1 = 4 - 2\sqrt{2}$.

For nonzero-sum n -person (or multi player) discounted stochastic games, existence of equilibrium strategies was shown by Fink [7], Takahashi [8], Rogers [9], and Sobel [10]. We state the theorem below for the two-person case, and the proof uses Kakutani's fixed point theorem.

Theorem 3. *Every two-person nonzero-sum discounted stochastic game has a pair of stationary strategies (f^*, g^*) of the players that constitutes a Nash equilibrium. Assuming that the players are maximizers, for each $s \in S$ and for all strategies f of P_1 and g of P_2*

$$I_\beta^{(1)}(f^*, g^*)(s) = \sup_f [I_\beta(f, g^*)(s)] \text{ and}$$

$$I_\beta^{(2)}(f^*, g^*)(s) = \sup_g [I_\beta(f^*, g)(s)].$$

As in the case of MDPs (refer articles in the section titled "Markov Decision Processes" in this encyclopedia), analyzing the undiscounted case (see articles in the section titled "Average Reward Criterion" in this encyclopedia) is more challenging than the discounted case (see articles in the section titled "Total Expected Discounted Reward Criterion" in this encyclopedia) for stochastic games as well. Gillette [3] gave an example

of an undiscounted zero-sum stochastic game, where one of the players does not have a stationary optimal strategy. In 1968, Blackwell and Ferguson [11] showed that this game (they call it “the big match”) does not admit an optimal strategy for one of the players, even when behavioral strategies are allowed. We state this example below and refer the reader to the paper by Blackwell and Ferguson [11] for details.

Example 3 [The big match (with undiscounted payoffs)]. The game has three states s_1 , s_2 , and s_3 . States s_2 and s_3 are absorbing, that is, once the game reaches these states, it remains there forever. The payoffs and transition probabilities are given below.

$$s_1 : \begin{bmatrix} 1 & 0 \\ (1, 0, 0) & (1, 0, 0) \\ 0 & 1 \\ (0, 1, 0) & (0, 0, 1) \end{bmatrix},$$

$$s_2 : \begin{bmatrix} 0 \\ (0, 1, 0) \end{bmatrix},$$

$$s_3 : \begin{bmatrix} 1 \\ (0, 0, 1) \end{bmatrix}.$$

Player 1 tries to maximize the undiscounted payoff and player 2 tries to minimize the same.

It is easy to see that $v_2 = 0$ and $v_3 = 1$.

Further, $v_1 = \frac{1}{2}$ and in state 1, it is optimal for player 2 to play column 1 with probability $\frac{1}{2}$ and column 2 with probability $\frac{1}{2}$. However, player 1 does not have exact optimal strategies. (But for every $\epsilon > 0$, player 1 has a behavioral strategy that fetches her or him $(\frac{1}{2} - \epsilon)$).

Though optimal strategies may not exist for both players, every zero-sum undiscounted stochastic game has a value as shown by Mertens and Neyman [5]. We state this theorem below.

Theorem 4 [Mertens and Neyman [5]]. *Every zero-sum undiscounted stochastic game has a (unique optimal) value. That is,*

$$v(s) = \Phi(s) = \sup_f \inf_g [\Phi(f, g)(s)] \\ = \inf_g \sup_f [\Phi(f, g)(s)].$$

Further, the players have ϵ -optimal strategies (not necessarily stationary). That is, for all $\epsilon > 0$, there exists a pair of strategies $(f_\epsilon^, g_\epsilon^*)$ of players 1 and 2, such that for all $s \in S$ and for strategies f and g of players 1 and 2, respectively,*

$$\Phi(f, g_\epsilon^*)(s) - \epsilon \leq \Phi(s) \\ = v(s) \leq \Phi(f_\epsilon^*, g)(s) + \epsilon.$$

The proof of the above theorem uses a result of Bewley and Kohlberg [12], and we refer the reader to Refs 5 and 12 for details.

On the one hand, significant research has been done in proving such existence results. Such existence results have also been extended to stochastic games with infinite state space and infinite action space for the discounted zero-sum and nonzero-sum cases as well as to the undiscounted zero-sum case. (Refer Maitra and Parthasarathy [13], Sobel [14]). Maitra and Sudderth [15–18] proposed an alternative proof of the existence of value in the finite, undiscounted, zero-sum case, which extends to the case when the state space is uncountable as well. For nonzero-sum undiscounted finite stochastic games, Vieille [19,20] proved the existence of ϵ -equilibrium strategies for the two-player case. The problem remains open for undiscounted stochastic games with three or more players.

On the other hand, computing the optimal value and optimal strategies (in the zero-sum case), and a Nash equilibrium (in the nonzero-sum case) has triggered a flurry of research activity, computational as well as theoretical. In general, solving a discounted stochastic game may not be as easy as in Example 2 above, either. Solving the system of equations as a result of Shapley’s theorem does give us the value vector but this system of equations

may not be easy to solve. However, we can compute approximate Nash equilibria by using iterative Newton–Raphson type procedures (Pollatschek and Avi-Itzhak [21]). The basis of such procedures is the fact that Shapley’s theorem can be viewed as computing the fixed point of the operator $L : R^N \rightarrow R^N$ defined by $L(v)(s) = \text{val}[R(s, v)]$ as we discussed earlier.

Even in Example 2 above, as v_1 is irrational, it is not possible to design a finite arithmetic step exact algorithm. However, some classes of stochastic games are guaranteed to have (at least) one rational solution, given rational inputs, and this is the focus of the next section.

ORDERFIELD PROPERTY OF STOCHASTIC GAMES

In 1950, Weyl [22] showed that matrix games possess the orderfield property. In other words, given payoffs from an ordered field, there exists a pair of optimal strategies whose coordinates lie in the same ordered field. It follows that the optimal value lies in the same ordered field as well. Bimatrix games (two-person NCSF games) also have the orderfield property when restricted to the rational field. Nash [4], in his seminal paper in 1951, gave an example of a three-player noncooperative game with rational payoffs but a unique irrational equilibrium. Unlike matrix games, stochastic games may not possess the orderfield property even in the discounted zero-sum case as pointed out by Shapley [1]. Parthasarathy and Raghavan [23] provide an explicit example. Example 2 above is another example of such a game that does not possess the orderfield property.

In this article, whenever we talk of the orderfield property, we restrict ourselves to the field of rationals.

Definition 6 [Stochastic game with the orderfield property]. A zero-sum stochastic game with rational inputs [i.e., rational payoffs, rational transition probabilities, and a rational discount factor (in case of discounted stochastic games)] is said to possess the orderfield property if it has a pair

of rational optimal strategies (i.e., optimal strategies whose coordinates are rational). It follows that the value of the game is rational too.

A nonzero-sum stochastic game with rational inputs is said to possess the orderfield property if it has a pair of Nash equilibrium strategies whose coordinates are rational. It follows that the corresponding equilibrium payoffs are rational as well.

Some classes of stochastic games have been shown to satisfy the orderfield property owing to their special structures. As at least one rational solution is guaranteed for these games, there is hope for finding a finite arithmetic-step algorithm to solve them. Researchers have been successful in designing algorithms for some of these classes. Refer Filar *et al.* [24], Nowak and Raghavan [25], Raghavan and Syed [26,27], and Vrieze [28] for some of these algorithms. On the other hand, there are classes where the problem is open. For some classes of stochastic games, though algorithms for solving them are known, search is on for efficient algorithms to solve them. For example, there is no efficient algorithm (yet) to solve switching control (SC) stochastic games.

We describe some classes of stochastic games that are known to possess the orderfield property. We refer the reader to a survey by Raghavan and Filar [29], a survey by Mohan *et al.* [30] and a survey by Raghavan [31] for more details on these classes and on algorithms for solving some of them.

Definition 7. Stochastic games with perfect information (Shapley [1]) are stochastic games in which in every state, the action space of (at least) one of the players is a singleton. In other words, we can partition the set of states S into S_1 and S_2 , where S_1 is the set of player 1 states (where player 2 has just one action and hence no choice) and S_2 is the set of player 2 states (where player 1 has just one action).

Definition 8. One player control stochastic games (Parthasarathy and Raghavan [23]) are stochastic games in which only one of

the players controls the transitions. For example, when player 2 controls transitions, $q(s'|s, i, j) = q(s'|s, j) \forall i \in A_1, j \in A_2, s, s' \in S$.

Definition 9. Simple stochastic games (SSG) (Condon [32]) are stochastic games with reachability objectives and no immediate rewards. There are two special absorbing states, the 0-sink and the 1-sink. Player 1 (the maximizer) wins if the 1-sink is reached, player 2 wins otherwise. Leaving out these special sink states, we can partition the remaining states into S_1 , S_2 , and S_3 , where S_1 is the set of player 1 states, (where player 2 has just one action and hence no choice), S_2 is the set of player 2 states (where player 1 has just one action), and S_3 is the set of nature (or average) states where the players do not have a choice and nature chooses the next state according to some probability distribution.

Remark. SSGs form a subclass of perfect-information stochastic games where the transition probabilities are 0 or 1 for all action-pairs in all states, S_1 and S_2 are player 1 and player 2 states, respectively, and in S_3 , both players have just one action.

Definition 10. Switching control (SC) stochastic games (Filar [33]) are games where the law of motion is controlled by player 1 alone when the game is played in a certain subset of states and by player 2 alone when the game is played in other states. In other words, a switching control game is a stochastic game in which the set of states are partitioned into sets S_1 and S_2 where the transition probabilities are given by

$$q(s'|s, i, j) = q(s'|s, i), \quad \text{for all } s' \in S, s \in S_1, \\ i \in A_1, j \in A_2$$

$$q(s'|s, i, j) = q(s'|s, j), \quad \text{for all } s' \in S, \\ s \in S_2, j \in A_2, i \in A_1$$

Remark. Switching control stochastic games are a superclass of SSGs,

perfect-information stochastic games, and one-player control stochastic games.

Definition 11. Separable reward-state independent transition (SER-SIT) (Parthasarathy *et al.* [34]) games are stochastic games in which

1. The rewards can be written as the sum of a function that depends on the state alone and another function that depends on the actions alone. That is, $r(s, i, j) = c(s) + a(i, j)$ for all $s \in S, i \in A_1, j \in A_2$, for the zero-sum case. For the nonzero-sum case, $r_1(s, i, j) = c(s) + a(i, j)$ and $r_2(s, i, j) = d(s) + b(i, j)$ for P_1 and P_2 , respectively, for all states $s \in S$, and actions $i \in A_1, j \in A_2$.
2. The transitions are independent of the state from which the game transitions. That is, $q(s'|s, i, j) = q(s'|i, j)$ for all $i \in A_1, j \in A_2, s, s' \in S$.

Definition 12. Additive reward additive transition (ARAT) (Raghavan *et al.* [35]) games are stochastic games where

1. The reward function can be written as the sum of two functions, one depending on player 1 and the other on player 2. For the zero-sum case, $r(s, i, j) = r'(s, i) + r''(s, j)$, for all $i \in A_1, j \in A_2, s \in S$. We can similarly write down r_1 and r_2 in the case of nonzero-sum games.
2. The transition probabilities can be written as a sum of two functions, one depending on player 1 and the other on player 2. That is, $q(s'|s, i, j) = q_1(s'|s, i) + q_2(s'|s, j)$, $i \in A_1, j \in A_2, s, s' \in S$.

There are interesting examples of stochastic games that have the orderfield property and do not fall into any of the known classes. The big match with discounted payoffs is one such game.

Example 4. The big match with discounted payoffs has the orderfield property:

$$s_1 : \begin{bmatrix} 1 & 0 \\ (1, 0, 0) & (1, 0, 0) \\ 0 & 1 \\ (0, 1, 0) & (0, 0, 1) \end{bmatrix}, \quad s_2 : \begin{bmatrix} 0 \\ (0, 1, 0) \end{bmatrix},$$

$$s_3 : \begin{bmatrix} 1 \\ (0, 0, 1) \end{bmatrix}.$$

Using Shapley's theorem, the value of the stochastic game starting at state s_1 is

$$v_1 = \text{val} \begin{bmatrix} 1 + \beta v_1 & \beta v_1 \\ 0 & 1 + \frac{\beta}{1-\beta} \end{bmatrix}. \quad (15)$$

This auxiliary matrix game is completely mixed. Using Kaplansky's theorem, we get

$$v_1 = \frac{1}{2(1-\beta)} \text{ which is rational when } \beta \text{ is rational.}$$

Now, we show that there exists a pair of rational optimal strategies as well. Assuming the column chooser is the minimizer, $(\frac{1}{2}, \frac{1}{2})$ is optimal for the column chooser.

Now, let $(z, 1-z)$ be an optimal strategy for the row chooser. Then,

$$\begin{bmatrix} z & 1-z \end{bmatrix} \begin{bmatrix} 1 + \beta v_1 & \beta v_1 \\ 0 & 1 + \frac{\beta}{1-\beta} \end{bmatrix} = \begin{bmatrix} v_1 & v_1 \end{bmatrix}.$$

Using $v_1 = \frac{1}{2(1-\beta)}$, we get $z = \frac{1}{2-\beta}$.

In other words, $(\frac{1}{2-\beta}, \frac{1-\beta}{2-\beta})$ is optimal for the row chooser and this strategy is rational whenever β is rational.

In the next section, we discuss algorithms for solving some of the above classes of stochastic games.

ALGORITHMS FOR SOME CLASSES OF STOCHASTIC GAMES

Different algorithms have been proposed for solving classes of stochastic games. Some of these algorithms reduce these stochastic games to a linear program (LP) (refer the section titled "Linear Programming (LP)" in this encyclopedia), some of them reduce them to matrix or bimatrix games, some of them use policy-improvement techniques similar to those for MDPs (refer **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**), and others use linear complementarity problems (LCP). (For LCPs, refer to the book by Cottle *et al.* [36]). For related results on the orderfield property of stochastic games and algorithms for various classes, apart from the citations listed above, we refer the reader to Flesch *et al.* [37], Mohan *et al.* [38], Schultz [39], Syed [40], Thuijsman and Raghavan [41], and Vrieze *et al.* [42].

LP Formulation of Zero-Sum Discounted One Player Control Stochastic Games

Let player 2 be the controlling player. In other words, $q(s'|s, i, j) = q(s'|s, j)$ for all $i \in A_1, j \in A_2, s, s' \in S$.

Let β be the discount factor.

Without loss of generality, we may assume that the payoffs are nonnegative (hence, $v_\beta(s) \geq 0$ for all $s \in S$).

Solving this game is equivalent to solving the following LP as shown by Parthasarathy and Raghavan [23]:

$$\max \sum_{s=1}^N v_\beta(s) \quad (16)$$

subject to

$$\sum_{i \in A_1} (r(s, i, j) + \beta \sum_{s' \in S} q(s'|s, j) v_\beta(s')) f_i(s) \geq v_\beta(s)$$

$$\forall j \in A_2, \forall s \in S. \quad (17)$$

$$\sum_{i \in A_1} f_i(s) = 1 \quad \forall s \in S \quad (18)$$

$$f_i(s), v_\beta(s) \geq 0 \quad \forall s \in S. \quad (19)$$

Here, $f_i(s)$ is the probability with which player 1 plays action i in state s .

For zero-sum undiscounted one player control stochastic games, Vrieze [28] proposed an LP formulation. Zero-sum discounted as well as undiscounted SER-SIT games can be reduced to matrix games (refer Sobel [43] and Parthasarathy *et al.* [34]) (and matrix games can be reduced to LPs).

Solving Two-Player Nonzero-Sum SER-SIT Games via Reduction to Bimatrix Games

Two-player nonzero-sum discounted and undiscounted SER-SIT games can be solved by reducing them to bimatrix games (Sobel [43], Parthasarathy *et al.* [34]). We discuss the discounted case below.

Let Γ be the β -discounted SER-SIT game with transitions $q(s'|i, j)$ and rewards $r_1(s, i, j) = c(s) + a(i, j)$ and $r_2(s, i, j) = d(s) + b(i, j)$ for P_1 and P_2 , respectively.

Construct the bimatrix game (A, B) where A and B are the following $m_1 \times m_2$ matrices:

$$A = [a(i, j) + \beta \sum_{s' \in S} q(s'|i, j)c(s')], B = [b(i, j) + \beta \sum_{s' \in S} q(s'|i, j)d(s')]. \quad (20)$$

If (x^*, y^*) is a pair of equilibrium strategies for the bimatrix game (A, B) , then $((x^*, x^*, \dots, x^*), (y^*, y^*, \dots, y^*))$ is a pair of equilibrium strategies for the β -discounted SER-SIT game Γ .

For two-player nonzero-sum one player control stochastic games, Nowak and Raghavan [25] proposed solving the discounted case and the irreducible undiscounted case via bimatrix games. (By irreducible we mean that the transition matrix induced by any pure stationary strategy of the controlling player is irreducible).

A Policy-Improvement Algorithm for Solving Zero-Sum Discounted Perfect Information Stochastic Games

Given a zero-sum β -discounted perfect-information stochastic game Γ , Raghavan and Syed [27] proposed the following algorithm to find a pair of pure stationary optimal strategies (f^*, g^*) . The algorithm is a policy-improvement algorithm similar to

that of Blackwell [44] for MDPs. MDPs can be thought of as one-person perfect-information stochastic games and as the player has the same objective (to maximize reward, say) across all states, it is intuitively clear why policy improvement works. In the case of TPZS perfect-information stochastic games, the players have conflicting objectives: one wants to maximize in some states and the other wants to minimize in other states. Even in this case, as Raghavan and Syed [27] prove, policy improvement works. The basis of the proof is Shapley's [1] theorem that perfect-information stochastic games have pure stationary equilibria. Hence, we only need to find a pair of pure stationary strategies (f^*, g^*) such that $I_\beta(f, g^*) \leq I_\beta(f^*, g^*) = v_s \leq I_\beta(f^*, g)$ for all pure strategies f and g of players 1 and 2, respectively.

The outline of the algorithm is as follows. The players start with an arbitrary stationary strategy pair (f_0, g_0) and generate a sequence of improvements via lexicographic search. That is, P_1 first looks for an improvement in state 1. Given P_2 's g_0 , let the best response of P_1 be f_1 . (Note that f_1 may be the same as f_0 . Also note that, at this step of the algorithm, P_1 is allowed to change his/her strategy only in state 1). Then P_2 looks for improvements across all states, starting at state 1. Let g_1 be P_2 's best response against P_1 's f_1 . Now, P_1 tries to improve his/her payoff by changing his/her strategy in state 1. Once, P_1 makes an improvement, the ball is back in P_2 's court. If f_1 and g_1 are the best responses to each other (when P_1 is allowed a move only in state 1), P_1 moves to state 2, and so on. If no such improvements exist, then clearly we have obtained a saddle point (f^*, g^*) . The algorithm terminates, because each player has a finite number of stationary strategy pairs and no cycling occurs (that is, the above algorithm does not visit the same strategy pair (f_i, g_j) twice). For details of the proof, refer Ref. 27.

For an algorithm for a class of multi-player stochastic games, refer Ref. 45. In the next section, we present some interesting results for the advanced reader. These results assume familiarity with real analysis and measure theory, and we just outline the

results and proofs. Readers without sufficient background or readers who are not interested in these topics may skip the following section.

STOCHASTIC GAMES WITH COUNTABLE STATE SPACE AND BOREL PAYOFFS

We start off with Gale–Stewart games [46] and Blackwell games [47], and Martin’s results [48,49] on the determinacy of these games. These results can be used to prove the existence of value in TPZS stochastic games with countable state space and Borel-measurable payoff function (Maitra and Sudderth [50,51], Martin [49]).

Borel Determinacy of Gale–Stewart Games and Blackwell Games

Definition 13 [Gale–Stewart games [46]]. A Gale–Stewart game is a TPZS game of perfect information played over the infinite horizon. A_1 and A_2 are arbitrary finite, nonempty sets of actions for players P_1 and P_2 , respectively. One of the players starts the game and the game is played sequentially. P_1 and P_2 alternately choose actions from A_1 and A_2 , respectively, after observing previous actions. Without loss of generality, let P_1 be the player to move first. P_2 observes P_1 ’s action, say $i_1 \in A_1$ and chooses $j_1 \in A_2$. P_1 now observes j_1 (and already knows i_1) and chooses i_2 , and so on. Let the sequence of actions chosen be $i_1, j_1, i_2, j_2, \dots$. Let $W \subset (A_1 \times A_2)^\infty$ be the winning set for P_1 . That is, P_1 wins if $(i_1, j_1, i_2, j_2, \dots) \in W$, and P_2 wins otherwise.

Example 5. Both the players can choose either 0 or 1. That is, $A_1 = A_2 = \{0, 1\}$. Player 1 chooses 0 or 1 in the first step. Player 2 observes this move and chooses 0 or 1. Player 1 observes this and picks 0 or 1, and so on. Let the sequence of bits chosen by the players be $b_1, b_2, \dots$. The payoff to player 1 is defined as $\sum_i (b_i/2^i)$, which is a real number in $[0, 1]$. Now, given a subset $W \subseteq [0, 1]$, player 1 wins if his/her payoff is in W , and player 2 wins otherwise.

Theorem 5 [Martin [48]]. Let A_1 and A_2 be arbitrary finite, nonempty sets of actions for players P_1 and P_2 , respectively. If W is a Borel subset of $(A_1 \times A_2)^\infty$, then the zero-sum perfect-information game as defined above is determined. That is, either P_1 has a winning strategy or P_2 has a winning strategy.

In Example 5 above, if $W = [0, 1/2]$, player 1 always wins. (P_1 can always choose 0). Given $W = [1/4, 3/4]$, P_2 always wins. (If $b_1 = 1$, P_2 chooses $b_2 = 1$ so that P_1 ’s payoff $> 3/4$ and if $b_1 = 0$, P_2 chooses $b_2 = 0$ so that P_1 ’s payoff $< 1/4$).

Now, we look at Martin’s result for Blackwell games, which are TPZS games of imperfect information.

Definition 14 [Blackwell games [47]]. Blackwell defined the following class of TPZS games of imperfect information played over the infinite horizon. Let $Z = A_1 \times A_2$, where A_1 is the set of actions of player 1 and A_2 is the set of actions of player 2. In step 1, players P_1 and P_2 simultaneously pick actions i_1 and j_1 . Let $z_1 = (i_1, j_1)$. In step 2, the players simultaneously pick $z_2 = (i_2, j_2)$, and so on. Let $f : Z^\infty \rightarrow R$ be the payoff to player 1.

Theorem 6 [Martin [49]]. Blackwell games with bounded Borel-measurable payoff functions are determined.

Value of Stochastic Games with Borel-Measurable Payoffs

Maitra and Sudderth [51] discuss the existence of value in TPZS stochastic games with countable state space S , and finite sets of actions A_1 and A_2 of players P_1 and P_2 , respectively. We define transition probabilities q as defined in the section titled “Two-Person Finite Stochastic Games.”

We fix the starting state s_0 . The game, then, proceeds as follows. In step 1, the players play $i_1 \in A_1$ and $j_1 \in B_1$ and the game transitions to state s_1 , say. We denote this by $z_1 = (i_1, j_1, s_1)$. $z_1 \in Z$ where $Z = A_1 \times A_2 \times S$. Similarly, in step 2 we have $z_2 = (i_2, j_2, s_2)$. $z_1, z_2, \dots$ define the histories. The players

may play pure strategies as described above or may play mixed strategies, say σ for P_1 and τ for P_2 . (σ and τ are behavioral strategies, that is, at each stage of the game, the strategy of each player depends on the history up to that stage. In other words, $\sigma = (\sigma_1, \sigma_2, \sigma_3, \dots)$, $\tau = (\tau_1, \tau_2, \tau_3, \dots)$ where σ_1 and τ_1 are the strategies of players 1 and 2, respectively, on day 1, σ_2 and τ_2 are the strategies of players 1 and 2, respectively, on day 2, given their strategies on day 1, and so on, the strategies on the t th day depending on the history of all days up to $t - 1$).

Let $H = (A_1 \times A_2 \times S) \cdot H^\infty = (A_1 \times A_2 \times S)^\infty$ is the set of all histories. Let the payoff function f be a bounded Borel-measurable function defined on H^∞ , where the sets A , B , and S are given the discrete topology and H^∞ is given the product topology. Without loss of generality, assume $f : H^\infty \rightarrow [0, 1]$. The payoff to player 1 is the expected value of f , denoted $E_{\sigma, \tau}(f)$. We denote the resulting stochastic game by Γ_f .

Theorem 7. *The stochastic game Γ_f has a value.*

Before we discuss the outline of the proof, we note the following. Clearly, Γ_f subsumes the case of discounted and undiscounted stochastic games. We stress that only the existence of a value for Γ_f is guaranteed and not the existence of optimal strategies for both the players.

The outline of the proof is as follows. Given Γ_f , we construct an equivalent perfect-information game.

Let $v \in [0, 1]$.

P_1 chooses a function $\phi_0 : H \rightarrow [0, 1]$ such that $v_{\phi_0}(s_0) \geq v$, where $v_{\phi_0}(s_0)$ is the value of the game, where $\phi_0(i, j, s)$ is the expected payoff to player 1 when the players choose actions i_1 and j_1 , respectively, and s is the next state which is chosen with probability $q(s|s_0, i, j)$. Now P_2 chooses $h_1 = (i_1, j_1, s_1) \in H$. P_1 now chooses $\phi_1 : H \rightarrow [0, 1]$ such that $v_{\phi_1}(s_1) \geq \phi_0(h_1)$, and so on.

P_1 wins if $f(h_1, h_2, \dots) \geq \liminf_{T \uparrow \infty} [\phi_{T-1}(h_T)]$.

We denote this perfect-information game by G_v .

We can apply Martin's determinacy theorem for perfect-information games (theorem above) as the winning set of P_1 is Borel.

Let $u \in [0, 1]$, $u \leq v$. Any winning strategy for P_1 in G_v is winning for P_1 in G_u as well.

Let $v_{\sup} = \sup\{v | v \in [0, 1] \text{ such that } P_1 \text{ has a winning strategy in } G_v\}$. Then P_1 has a winning strategy for all $v \in [0, v_{\sup})$ and P_2 has a winning strategy for $(v_{\sup}, 1]$.

The value of the stochastic game Γ_f is $v_{\sup}$, and it suffices to show that the lower value is at least $v_{\sup}$ and the upper value is at most $v_{\sup}$.

Maitra and Sudderth [51] show that whenever player 1 has a winning strategy in G_v , player 1 has a strategy σ in Γ_f such that $E_{\sigma, \tau}(f) \geq v$ for all strategies τ of player 2. Further, whenever player 2 has a winning strategy in G_v , player 2 has a strategy τ such that $E_{\sigma, \tau}(f) \leq v + \epsilon$, ($\epsilon > 0$), for all strategies σ of player 1. Interested readers may refer to Ref. 51 for details of the proof.

As discussed above, Martin's theorem implies the existence of value and ϵ -optimal strategies for TPZS stochastic games with countable state space and finite sets of actions, but the existence of ϵ -equilibria is an open problem when the game is nonzero-sum.

APPLICATIONS OF STOCHASTIC GAMES

Stochastic games have been (and are being) widely applied to various fields. In this section, we briefly discuss some of these applications.

Altman [52] discusses the applications of stochastic games to queues. (Refer the section titled "Queueing Theory and Queueing Networks" in this encyclopedia for details regarding queues.) We discuss a scenario that occurs in communication networks. Given a network, the traffic assignment problem involves assigning different packets to paths in the network. BCMP (Baskett, Chandy, Muntz, Palacios) networks are a class of product-form queueing networks (refer the section titled "Product-Form Queueing Networks" in this encyclopedia) that satisfy assumptions of the traffic assignment problem. Given various nodes in the network where packets may

be queued (or various paths that the packets may take), the problem of where to queue (or which route to take) can be modeled as a stochastic game. The arriving packets are the players, and at each time instant, the state of each node (or path) may be defined by the current queue length at that node (or traffic density in that path). A player at a particular node (or path) may not know the precise congestion state of the other nodes (paths). However, given the initial state and the arrival rate, the joint distribution of the congestion state as a function of the routing policy can be computed. Refer to Ref. 52 for details.

Now, we summarize an application of stochastic games to inventory control proposed by Avsar and Gursoy [53]. In their model, there are two retailers who compete for the substitutable demand of different products. Both the retailers know the substitution rates beforehand. The demand for each product is assumed to be random, and the demand distributions are known to the retailers too. At each time instant, the retailers decide their orders depending on their current inventory levels. The payoff function of each retailer includes their corresponding revenue, holding cost, and ordering cost. The authors model this substitutable product inventory problem over the infinite horizon as a stochastic game and prove that if the ordering cost is linear, there exists a unique pair of stationary base stock strategies that constitute a Nash equilibrium. Refer to Ref. 53 for details.

Other applications of stochastic games include an application to network security by Lye and Wing [54], the traveling inspector model (Filar [55]), applications to economics (Amir [56,57]), and a more recent application to power control by Altman *et al.* [58]. Kardes [59] discusses an application of stochastic games to a homeland security decision problem. We refer the reader to the respective papers for details.

SOME RECENT WORK

Apart from some of the recent applications we discussed in the previous section, there are

some recent works on theoretical as well as computational aspects of stochastic games. For example, Brazdil *et al.* [60] show that every continuous-time stochastic game with reachability objective has a value. For solving some subclasses of such games, such as finite uniform games, they propose algorithms and discuss the complexity. Neogy *et al.* [61] discuss an algorithm to solve a mixture class of stochastic games consisting of SC and ARAT states, a class which was shown to possess the orderfield property by Sinha [62]. Krishnamurthy *et al.* [63] discuss the communication complexity of some classes of nonzero-sum stochastic games. They discuss the problem of determining the existence of pure Nash equilibria when each player knows only his or her payoffs and not that of the opponent. We refer the reader to Ref. 63 for details. Other recent works on stochastic games include the results of Ganzfried and Sandholm [64] on solving multiplayer stochastic games of imperfect information, and results of Krishnamurthy *et al.* [65] on sufficient conditions for mixtures of stochastic games to possess the orderfield property.

We also refer the reader to the additional reading list for more details on stochastic games—theoretical aspects, computational aspects, as well as applications.

Acknowledgments

The authors would like to thank the Indian Statistical Institute, Chennai and the Chennai Mathematical Institute for their support. The second author would like to thank the Indian National Science Academy (INSA) for financial support under the Senior Scientist Scheme. The authors would also like to thank Dr. G. Ravindran for useful discussions. The authors would also like to thank an anonymous referee for useful suggestions and the editorial team of the Wiley Encyclopedia of Operations Research and the Management Sciences (EORMS) for all their help.

REFERENCES

1. Shapley L. Stochastic games. *Proc Natl Acad Sci* 1953;39:1095–1100.

2. Gillette D. Representable infinite games [PhD thesis]. Berkeley: University of California; 1953.
3. Gillette D. Stochastic games with zero-stop probabilities, contributions to the theory of games. In: Dresher M, Tucker AW, Wolfe P, editors. Volume III, Annals of Mathematics Studies 39. Princeton (NJ): Princeton University Press; 1957. pp. 179–188.
4. Nash JF. Non-cooperative Games. *Ann Math* 1951;54:286–295.
5. Mertens JF, Neyman A. Stochastic games. *Int J Game Theor* 1981;10:53–66.
6. Kaplansky I. A contribution to von Neumann's theory of games. *Ann Math* 1945;46(3):474–479.
7. Fink AM. Equilibrium in a stochastic n-person game. *J Sci Hiroshima Univ Ser A* 1964;28:89–93.
8. Takahashi M. Equilibrium points of stochastic, noncooperative n-person games. *J Sci Hiroshima Univ Ser A* 1964;28:95–99.
9. Rogers, PD. Non-zero sum stochastic games [PhD thesis]. Report ORC 69-8. Berkeley: University of California; 1979.
10. Sobel MJ. Noncooperative stochastic games. *Ann Math Stat* 1971;42:1930–1935.
11. Blackwell D, Ferguson TS. The big match. *Ann Math Stat* 1968;39:159–168.
12. Bewley T, Kohlberg E. The asymptotic theory of stochastic games. *Math Oper Res* 1976;1:197–208.
13. Maitra A, Parthasarathy T. On stochastic games. *J Opt Theor Appl* 1970;5:289–300.
14. Sobel MJ. Continuous stochastic games. *J Appl Probab* 1973;10:497–504.
15. Maitra AP, Sudderth W. An operator solution of stochastic games. *Israel J Math* 1992;78:33–49.
16. Maitra AP, Sudderth W. Finitely additive and measurable stochastic games. *Int J Game Theor* 1993;22:201–223.
17. Maitra AP, Sudderth W. Borel stochastic games with limsup payoff. *Ann Probab* 1993;21:861–885.
18. Maitra AP, Sudderth W. Discrete gambling and stochastic games. New York: Springer; 1996.
19. Vieille N. Two-player stochastic games I: a reduction. *Israel J Math* 2000;119:55–91.
20. Vieille N. Two-player stochastic games II: the case of recursive games. *Israel J Math* 2000;119:93–126.
21. Pollatschek MA, Avi-Itzhak B. Algorithms for stochastic games with geometrical interpretation. *Manag Sci* 1969;15(7):399–415.
22. Weyl H. Elementary proof of a minimax theorem due to von neumann. In: Kuhn HW, Tucker AW, editors. Volume I, Contributions to the theory of games, Annals of Mathematics Studies 24. Princeton (NJ): Princeton University Press; 1950. pp. 19–25.
23. Parthasarathy T, Raghavan TES. An orderfield property for stochastic games when one player controls transition probabilities. *J Optim Theor Appl* 1981;33(3):375–392.
24. Filar JA, Schultz T, Thuijsman F, *et al.* Nonlinear programming and stationary equilibria in Stochastic Games. *Math Program* 1991;50:227–237.
25. Nowak AS, Raghavan TES. A finite step algorithm via a bimatrix game to a single controller non-zero sum stochastic game. *Math Program* 1993;59:249–259.
26. Raghavan TES, Syed Z. Computing stationary nash equilibria of undiscounted single-controller stochastic games. *Math Oper Res* 2002;27(2):384–400.
27. Raghavan TES, Syed Z. A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. *Math Program Ser A* 2003;95:513–532.
28. Vrieze OJ. Linear programming and undiscounted stochastic game in which one player controls transitions. *OR Spektrum* 1981;3:29–35.
29. Raghavan TES, Filar JA. Algorithms for stochastic games - A survey. *Z Oper Res* 1991;35:437–472.
30. Mohan SR, Neogy SK, Parthasarathy T. Pivoting algorithms for some classes of stochastic games: a survey. *Int Game Theor Rev* 2001;3(2 & 3):253–281.
31. Raghavan TES. Finite-step algorithms for single controller and perfect information stochastic games. In: Neyman A, Sorin S, editors. Volume 570, Stochastic games and applications, NATO Science Series: C. Dordrecht, The Netherlands: Kluwer Academic Publishers Group; 2003. pp. 227–251.
32. Condon A. The complexity of stochastic games. *Info Comput* 1992;96:203–224.
33. Filar JA. Orderfield property of stochastic games when the player who controls the transition changes from state to state. *J Optim Theor Appl* 1981;34:505–517.
34. Parthasarathy T, Tijs SJ, Vrieze OJ. Stochastic games with state independent transitions

- and separable rewards. *Lect Notes Econ Math Syst* 1984;226:262–271.
35. Raghavan TES, Tijs SJ, Vrieze OJ. Stochastic games with additive rewards and additive transitions. *J Optim Theor Appl* 1986;47:451–464.
 36. Cottle RW, Pang JS, Stone RE. *The linear complementarity problem*. New York: Academic Press; 1992.
 37. Flesch J, Thuijsman F, Vrieze OJ. Stochastic games with additive transitions. *Eur J Oper Res* 2007;179:483–497.
 38. Mohan SR, Neogy SK, Parthasarathy T, *et al.* Vertical linear complementarity and discounted zero-sum stochastic games with ARAT structure. *Math Program Ser A* 1999;86(3):637–648.
 39. Schultz TA. Linear complementarity and discounted switching controller stochastic games. *J Optim Theor Appl* 1992;73(1):89–99.
 40. Syed Z. Algorithms for stochastic games and related topics [PhD thesis]. Chicago: University of Illinois; 1999.
 41. Thuijsman F, Raghavan TES. Perfect information stochastic games and related classes. *Int J Game Theor* 1997;26:403–408.
 42. Vrieze OJ, Tijs SH, Raghavan TES, *et al.* A finite algorithm for the switching control stochastic game. *OR Spektrum* 1983;5(1):15–24.
 43. Sobel MJ. Myopic solutions of markov decision processes and stochastic games. *Oper Res* 1981;29:995–1009.
 44. Blackwell D. Discrete dynamic programming. *Ann Math Stat* 1962;33:719–726.
 45. Mohan SR, Neogy SK, Parthasarathy T. Linear complementarity and discounted polystochastic game when one player controls transitions. In: Ferris MC, Pang J-S, editors. *Proceedings of the International Conference on Complementarity Problems*. Philadelphia (PA): SIAM; 1997. pp. 284–294.
 46. Gale D, Stewart FM. Infinite games with perfect information, contributions to the theory of games. In: Kuhn HW, Tucker AW, editors. *Volume II, Annals of Mathematics Studies* 28. Princeton (NJ): Princeton University Press; 1953. pp. 245–266.
 47. Blackwell D. Infinite G_δ games with imperfect information. *Zastos Mat* 1969;10:99–101.
 48. Martin DA. Borel determinacy. *Ann Math* 1975;102(2):363–371.
 49. Martin DA. The determinacy of Blackwell games. *J Symbolic Logic* 1998;63:1565–1581.
 50. Maitra AP, Sudderth W. Finitely additive stochastic games with borel-measurable payoffs. *Int J Game Theor* 1998;27:257–267.
 51. Maitra AP, Sudderth W. Stochastic games with borel payoffs. In: Neyman A, Sorin S, editors. *Volume 570, Stochastic games and applications*, NATO Science Series: C. Dordrecht, The Netherlands: Kluwer Academic Publishers Group; 2003. pp. 367–373.
 52. Altman E. Applications of dynamic games in queues. *Volume 7, Advances in dynamic games*. Boston (MA), Basel, Berlin: Birkhauser; 2005. pp. 309–342.
 53. Avsar MZ, Gursoy MB. Inventory control under substitutable demand: a stochastic game application. *Volume 49, Naval research logistics*. Hoboken (NJ): Wiley Periodicals, Inc.; 2002. pp. 359–375.
 54. Lye KW, Wing JM. Game strategies in network security. *Int J Inf Secur* 2005;4:71–86.
 55. Filar JA. Player aggregation in the traveling inspector model. *IEEE Trans Automat Contr* 1985;AC-30:723–729.
 56. Amir R. Stochastic games in economics: the lattice theoretic approach. In: Neyman A, Sorin S, editors. *Volume 570, Stochastic games and applications*, NATO Science Series: C. Dordrecht, The Netherlands: Kluwer Academic Publishers Group; 2003. pp. 443–453.
 57. Amir R. Stochastic games in economics and related fields: an overview. In: Neyman A, Sorin S, editors. *Volume 570, Stochastic games and applications*, NATO Science Series: C. Dordrecht, The Netherlands: Kluwer Academic Publishers Group; 2003. pp. 455–470.
 58. Altman E, Kamble V, Silva A. Stochastic games with one step delay sharing information pattern with application to power control. *Proceedings of the International Conference on Game Theory for Networks (GameNets)*. Istanbul, Turkey, IEEE Press, Piscataway (NJ): IEEE Xplore; 2009. pp. 124–129.
 59. Kardes E. The Use of Stochastic Games on a Homeland Security Problem. CREATE (Center for Risk and Economic Analysis of Terrorism Events) Report. 2008.
 60. Brazdil T, Forejt V, Krcal J, *et al.* Continuous-Time Stochastic Games with Time-Bounded Reachability. *Proceedings Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*. In: Kannan R, Kumar KN, editors. *Leibniz International*

- Proceedings in Informatics (LIPIcs). Germany: Schloss Dagstuhl - Leibniz-Zentrum für Informatik; 2009. pp. 61–72.
61. Neogy SK, Das AK, Sinha S, *et al.* On a mixture class of stochastic game with ordered field property. In: Neogy SK, Bapat RB, Das AK, *et al.*, editors. Volume 1, Mathematical programming and game theory for decision making, ISI platinum jubilee series on statistical science and interdisciplinary research. Singapore: World Scientific; 2008. pp. 451–477.
 62. Sinha S. A contribution to the theory of stochastic games [PhD thesis]. New Delhi: Indian Statistical Institute; 1989.
 63. Krishnamurthy N, Parthasarathy T, Ravindran G. Communication complexity of stochastic games. Proceedings of the International Conference on Game Theory for Networks (GameNets). Istanbul, Turkey, IEEE Press, Piscataway, (NJ): IEEE Xplore; 2009. pp. 411–417.
 64. Ganzfried S, Sandholm T. Computing equilibria in multiplayer stochastic games of imperfect information. Proceedings 21st International Joint Conference on Artificial Intelligence. Pasadena (CA): AAAI Publications; 2009. pp. 140–146.
 65. Krishnamurthy N, Parthasarathy T, Ravindran G. Orderfield property of mixtures of stochastic games. Sankhya: Indian J Stat 2010;72-A:246–275. (Also available online at sankhya.isical.ac.in).

ADDITIONAL READING

- Filar JA, Vrieze OJ. Competitive Markov decision processes. New York: Springer, New York, Inc.; 1997.
- Mertens JF. Stochastic games. Chapter 47, In: Aumann RJ, Hart S, editors. Volume 3, Handbook of game theory with economic applications. Amsterdam, The Netherlands: Elsevier; 2002. pp. 1809–1832.
- Neyman A, Sorin S, editors. Volume 570, Stochastic games and applications, NATO Science Series C. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2003.
- Raghavan TES, Ferguson TS, Parthasarathy T, *et al.*, editors. Stochastic games and related topics: In Honor of Professor L. S. Shapley. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1990.

MULTISTATE SYSTEM RELIABILITY

BENT NATVIG
Department of Mathematics,
University of Oslo, Oslo,
Norway

One inherent weakness of traditional reliability theory (see **Coherent Systems**) is that the system and the components are always described just as functioning or failed. The first attempts to replace this by a theory for multistate systems of multistate components were done in the late 1970s in Refs 1–3. This was followed up by independent work in Refs 4–6 giving proper definitions of a multistate monotone system (MMS) and of multistate coherent systems (MCS) and also of minimal path and cut vectors. Furthermore, in Ref. 7 upper and lower bounds for the availabilities and unavailabilities, to any level, in a fixed time interval were arrived at for MMSs based on corresponding information on the multistate components. These were assumed to be maintained and interdependent. Such bounds are of great interest when trying to predict the performance process of the system, noting that exact expressions are obtainable just for trivial systems. Hence, by the mid-1980s, the basic multistate reliability theory was established. A review of the early development in this area is given in Ref. 8. Recently, probabilistic modeling of partial monitoring of components with applications to preventive system maintenance has been extended in Ref. 9 to MMSs of multistate components.

The theory was applied in Ref. 10 to an offshore electrical power generation system for two nearby oilrigs, where the amounts of power that may possibly be supplied to the two oilrigs are considered as system states. This application is also used to illustrate the theory in Ref. 9. In Ref. 11 the theory was applied to the Norwegian offshore gas pipeline network in the North Sea, as of the end of the 1980s, transporting gas to Emden in Germany. The system state depends on the

amount of gas actually delivered, and also to some extent on the amount of gas compressed mainly by the compressor component closest to Emden. Also, rather recently, the first book [12] on multistate system reliability analysis and optimization appeared. The book also contains many examples of application of reliability assessment and optimization methods to real engineering problems. A recent review of the area is given in Ref. 13 on which the present one is based.

BASIC CONCEPTS AND BASIC BOUNDS

Let $S = \{0, 1, \dots, M\}$ be the set of states of the system; the $M + 1$ states representing successive levels of performance ranging from the perfect functioning level M down to the complete failure level 0. Furthermore, let $C = \{1, \dots, n\}$ be the set of components and S_i ($i = 1, \dots, n$) the set of states of the i th component. We claim $\{0, M\} \subseteq S_i \subseteq S$. Hence, the states 0 and M are chosen to represent the end points of a performance scale that might be used for both the system and its components. Let x_i ($i = 1, \dots, n$) denote the state or performance level of the i th component and $\mathbf{x} = (x_1, \dots, x_n)$. It is assumed that the state, ϕ , of the system is given by the structure function $\phi = \phi(\mathbf{x})$. For the following type of multistate systems, a series of results can be derived.

Definition 1. A system is an *MMS* if its structure ϕ satisfies:

- (i) $\phi(\mathbf{x})$ is nondecreasing in each argument
- (ii) $\phi(\mathbf{0}) = 0$ and $\phi(\mathbf{M}) = M$ ($\mathbf{0} = (0, \dots, 0)$, $\mathbf{M} = (M, \dots, M)$).

The first assumption roughly says that improving one of the components cannot harm the system, whereas the second says that if all components are in the complete failure (perfect functioning) state, then the system is in the complete failure (perfect functioning) state.

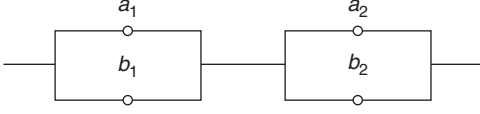


Figure 1. A simple network.

Table 1. States of the Simple Network System of Fig. 1

Component 2	3	0	2	3
	1	0	1	2
	0	0	0	0
		0	1	3
Component 1				

As a simple example of an MMS consider the network depicted in Fig. 1. Here, component 1 is the parallel module of the branches a_1 and b_1 and component 2 is the parallel module of the branches a_2 and b_2 . Let for $i = 1, 2$ $x_i = 0$ if no branches work, 1 if one branch works, and 3 if two branches work. The states of the system are given in Table 1.

Note, for instance, that the state 1 is critical both for each component and the system as a whole in the sense that the failing of a branch leads to the 0 state. In binary theory the functioning state comprises the states $\{1, 2, 3\}$ and hence only a rough description of the system's performance is possible. It is not hard to see that the structure function is given by

$$\phi(\mathbf{x}) = x_1 x_2 - I(x_1 x_2 = 3) - 6I(x_1 x_2 = 9),$$

where $I(\cdot)$ is the indicator function.

We now impose some further restrictions on the structure function ϕ . The following notation is needed:

$$\begin{aligned} (\cdot, \mathbf{x}) &= (x_1, \dots, x_{i-1}, \cdot, x_{i+1}, \dots, x_n), \\ S_{i,j}^0 &= S_i \cap \{0, \dots, j-1\}, \quad \text{and} \\ S_{i,j}^1 &= S_i \cap \{j, \dots, M\}. \end{aligned}$$

Definition 2. Consider an MCS with structure function ϕ satisfying

- (i) $\min_{1 \leq i \leq n} x_i \leq \phi(\mathbf{x}) \leq \max_{1 \leq i \leq n} x_i$.
If in addition $\forall i \in \{1, \dots, n\}, \forall j \in \{1, \dots, M\}, \exists (\cdot, \mathbf{x})$ such that

- (ii) $\phi(k_i, \mathbf{x}) \geq j, \phi(\ell_i, \mathbf{x}) < j, \forall k \in S_{i,j}^1, \forall \ell \in S_{i,j}^0$, we have a *multistate strongly coherent system (MSCS)*,
(iii) $\phi(k_i, \mathbf{x}) > \phi(\ell_i, \mathbf{x}) \forall k \in S_{i,j}^1, \forall \ell \in S_{i,j}^0$, we have an *MCS*,
(iv) $\phi(M_i, \mathbf{x}) > \phi(0_i, \mathbf{x})$, we have a *multistate weakly coherent system (MWCS)*.

When $M = 1$, all reduce to the established binary coherent system (BCS) (see **Coherent Systems**). The structure function $\min_{1 \leq i \leq n} x_i (\max_{1 \leq i \leq n} x_i)$ is often denoted the multistate series (parallel) structure.

Now choose $j \in \{1, \dots, M\}$ and let the states $S_{i,j}^0, (S_{i,j}^1)$ correspond to the failure (functioning) state for the i th component if a binary approach had been applied. Condition (ii) above means that for all components i and any level j , there shall exist a combination of the states of the other components, $(\cdot, \mathbf{x})$, such that if the i th component is in the binary failure (functioning) state, the system itself is in the corresponding binary failure (functioning) state. Roughly speaking, modifying [6], condition (ii) says that every level of each component is relevant to the same level of the system, condition (iii) says that every level of each component is relevant to the system, whereas condition (iv) simply says that every component is relevant to the system.

For a BCS, one can prove the following practically very useful principle: Redundancy at the component level is superior to redundancy at the system level except for a parallel system where it makes no difference. Assuming $S_i = S$ ($i = 1, \dots, n$) this is also true for an MCS, but not for an MWCS.

We now mention a special type of an MSCS. Introduce the indicators ($j = 1, \dots, M$), $I_j(x_i) = 1(0)$ if $x_i \geq j$ ($x_i < j$), and the indicator vector $\mathbf{I}_j(\mathbf{x}) = (I_j(x_1), \dots, I_j(x_n))$.

Definition 3. An MMS is said to be a *binary type multistate monotone system (BTMMS)* iff there exist binary monotone structures $\phi_j, j = 1, \dots, M$ such that its structure function ϕ satisfies $\phi(\mathbf{x}) \geq j \Leftrightarrow \phi_j(\mathbf{I}_j(\mathbf{x})) = 1$ for all $j \in \{1, \dots, M\}$ and all $\mathbf{x}$. If the binary monotone structure functions are

coherent, we have a *binary type multistate strongly coherent system (BTMSCS)*.

Choose again $j \in \{1, \dots, M\}$ and let the states $S_{i,j}^0(S_{i,j}^1)$ correspond to the failure (functioning) state for the i th component if a binary approach is applied. By the definition above, ϕ_j will, from the binary states of the components, uniquely determine the corresponding binary state of the system.

In what follows $\mathbf{y} < \mathbf{x}$ means $y_i \leq x_i$ for $i = 1, \dots, n$, and $y_i < x_i$ for some i .

Definition 4. Let ϕ be the structure function of an MMS and let $j \in \{1, \dots, M\}$. A vector $\mathbf{x}$ is said to be a *minimal path (cut) vector to level j* iff $\phi(\mathbf{x}) \geq j$ and $\phi(\mathbf{y}) < j$ for all $\mathbf{y} < \mathbf{x}$ ($\phi(\mathbf{x}) < j$ and $\phi(\mathbf{y}) \geq j$ for all $\mathbf{y} > \mathbf{x}$).

We now illustrate Definition 4 by finding the minimal path and cut vectors to all levels for the simple network system given in Fig. 1. From Table 1, it is easy to come up with the minimal path vectors and minimal cut vectors given in Tables 2 and 3, respectively. Note that the same vector may be a minimal cut vector to more than one level. This is not true for the minimal path vectors.

Definition 5. The *performance process of the i th component* ($i = 1, \dots, n$) is a stochastic process $\{X_i(t), t \in [0, \infty)\}$, where

Table 2. Minimal Path Vectors for the Simple Network System of Fig. 1

Level	Component 1	Component 2
1	1	1
2	1	3
2	3	1
3	3	3

Table 3. Minimal Cut Vectors for the Simple Network System of Fig. 1

Levels	Component 1	Component 2
1, 2	0	3
1, 2	3	0
2	1	1
3	1	3
3	3	1

for each fixed $t \in [0, \infty)$ $X_i(t)$ is a random variable which takes values in S_i . The *joint performance process for the components* $\{\mathbf{X}(t), t \in [0, \infty)\} = \{(X_1(t), \dots, X_n(t)), t \in [0, \infty)\}$ is the corresponding vector stochastic process. The *performance process of an MMS* with structure function ϕ is a stochastic process $\{\phi(\mathbf{X}(t)), t \in [0, \infty)\}$, where for each fixed $t \in [0, \infty)$, $\phi(\mathbf{X}(t))$ is a random variable which takes values in S .

Definition 6. The performance processes $\{X_i(t), t \in [0, \infty)\}$, $i = 1, \dots, n$ are *independent* in the time interval I iff, for any integer m and $\{t_1, \dots, t_m\} \subset I$ the random vectors $\{X_1(t_1), \dots, X_1(t_m)\}, \dots, \{X_n(t_1), \dots, X_n(t_m)\}$ are independent.

Definition 7. Let $j \in \{1, \dots, M\}$. The *availability*, $p_\phi^{j(I)}$ and the *unavailability*, $q_\phi^{j(I)}$ to level j in the time interval I for an MMS with structure function ϕ are given by

$$p_\phi^{j(I)} = P[\phi(\mathbf{X}(s)) \geq j \quad \forall s \in I]$$

$$q_\phi^{j(I)} = P[\phi(\mathbf{X}(s)) < j \quad \forall s \in I].$$

The corresponding *availability* and *unavailability* for the i th component to level j in the time interval I are denoted respectively by $p_i^{j(I)}$ and $q_i^{j(I)}$ for $i = 1, \dots, n$.

Note that $p_\phi^{j(I)} + q_\phi^{j(I)} \leq 1$, with equality for the case $I = [t, t]$.

As an example of the bounds for $p_\phi^{j(I)}$ and $q_\phi^{j(I)}$ given in Ref. 7, without any assumption according to Ref. 14 that each of the performance processes of the components is associated in I , we give the following theorem by first introducing the $n \times M$ matrices

$$\mathbf{P}_\phi^{(I)} = \{p_i^{j(I)}\}_{\substack{i=1,\dots,n \\ j=1,\dots,M}}$$

$$\mathbf{Q}_\phi^{(I)} = \{q_i^{j(I)}\}_{\substack{i=1,\dots,n \\ j=1,\dots,M}}.$$

Theorem 1. Let (C, ϕ) be an MMS with the marginal performance processes of its components being independent in I . Furthermore,

for $j \in \{1, \dots, M\}$ let $\mathbf{y}_k^j = (y_{1k}^j, \dots, y_{n_k}^j)$, $k = 1, \dots, n^j$ ($\mathbf{z}_k^j = (z_{1k}^j, \dots, z_{n_k}^j)$, $k = 1, \dots, m^j$) be its minimal path (cut) vectors to level j . Define

$$\ell_\phi^{j(I)}(\mathbf{P}_\phi^{(I)}) = \max_{1 \leq k \leq n^j} \prod_{i=1}^n p_i^{y_{ik}^j(I)}$$

$$\bar{\ell}_\phi^{j(I)}(\mathbf{Q}_\phi^{(I)}) = \max_{1 \leq k \leq m^j} \prod_{i=1}^n q_i^{z_{ik}^j(I)}$$

$$\ell_\phi^{**j(I)}(\mathbf{P}_\phi^{(I)}) = \prod_{k=1}^{m^j} \prod_{i=1}^n p_i^{z_{ik}^j(I)}$$

$$\bar{\ell}_\phi^{**j(I)}(\mathbf{Q}_\phi^{(I)}) = \prod_{k=1}^{n^j} \prod_{i=1}^n q_i^{y_{ik}^j(I)}$$

$$L_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}) = \max[\ell_\phi^{j(I)}(\mathbf{P}_\phi^{(I)}), \ell_\phi^{**j(I)}(\mathbf{P}_\phi^{(I)})]$$

$$\bar{L}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) = \max[\bar{\ell}_\phi^{j(I)}(\mathbf{Q}_\phi^{(I)}), \bar{\ell}_\phi^{**j(I)}(\mathbf{Q}_\phi^{(I)})]$$

$$B_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}) = \max_{j \leq k \leq M} [L_\phi^{*k(I)}(\mathbf{P}_\phi^{(I)})]$$

$$\bar{B}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) = \max_{1 \leq k \leq j} [\bar{L}_\phi^{*k(I)}(\mathbf{Q}_\phi^{(I)})].$$

Then

$$\begin{aligned} L_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}) &\leq B_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}) \leq p_\phi^{j(I)} \\ &\leq \inf_{t \in I} [1 - \bar{B}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I, t)})] \\ &\leq 1 - \bar{B}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) \leq 1 - \bar{L}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) \\ \bar{L}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) &\leq \bar{B}_\phi^{*j(I)}(\mathbf{Q}_\phi^{(I)}) \\ &\leq q_\phi^{j(I)} \leq \inf_{t \in I} [1 - B_\phi^{*j(I)}(\mathbf{P}_\phi^{(I, t)})] \\ &\leq 1 - B_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}) \leq 1 - L_\phi^{*j(I)}(\mathbf{P}_\phi^{(I)}). \end{aligned}$$

Here, $\prod_{i=1}^n a_i \stackrel{\text{def}}{=} 1 - \prod_{i=1}^n (1 - a_i)$. By specializing $M = 1$ and $I = [t, t]$ the bounds reduce to the familiar ones from binary theory as given in Ref. 15. It should be admitted that the upper bounds are poor if $p_\phi^{j(I)} + q_\phi^{j(I)}$ is not close to 1.

We now apply the theory to the simple network system given in Fig. 1. We consider the time interval $I = [t_1, t_2]$ and assume that the marginal performance processes of the two components are independent in $[0, t_2]$ and also that the two branches of each component fail and are repaired/replaced

independently of each other. All branches have the same instantaneous failure rate λ and repair/replacement rate μ .

Applying Tables 2 and 3, noting that $p_i^{2(I)} = p_i^{3(I)}$, the lower bounds are given by

$$\begin{aligned} \ell_\phi^{1(I)}(\mathbf{P}_\phi^{(I)}) &= \ell_\phi^{**1(I)}(\mathbf{P}_\phi^{(I)}) = L_\phi^{*1(I)}(\mathbf{P}_\phi^{(I)}) \\ &= B_\phi^{*1(I)}(\mathbf{P}_\phi^{(I)}) = (p_i^{1(I)})^2 = p_\phi^{1(I)} \\ \ell_\phi^{2(I)}(\mathbf{P}_\phi^{(I)}) &= p_i^{1(I)} p_i^{3(I)} \leq p_i^{1(I)} p_i^{3(I)} \\ &\quad \max[1, p_i^{1(I)}(2 - p_i^{3(I)})] \\ &= L_\phi^{*2(I)}(\mathbf{P}_\phi^{(I)}) = B_\phi^{*2(I)}(\mathbf{P}_\phi^{(I)}) \leq p_\phi^{2(I)} \\ \ell_\phi^{3(I)}(\mathbf{P}_\phi^{(I)}) &= \ell_\phi^{**3(I)}(\mathbf{P}_\phi^{(I)}) = L_\phi^{*3(I)}(\mathbf{P}_\phi^{(I)}) \\ &= B_\phi^{*3(I)}(\mathbf{P}_\phi^{(I)}) = (p_i^{3(I)})^2 = p_\phi^{3(I)}. \end{aligned}$$

Noting that $q_i^{2(I)} = q_i^{3(I)}$ the upper bounds are given by

$$\begin{aligned} 1 - \bar{\ell}_\phi^{1(I)}(\mathbf{Q}_\phi^{(I)}) &= 1 - q_i^{1(I)} \geq (1 - q_i^{1(I)})^2 \\ &= 1 - \bar{L}_\phi^{*1(I)}(\mathbf{Q}_\phi^{(I)}) = 1 - \bar{B}_\phi^{*1(I)}(\mathbf{Q}_\phi^{(I)}) \\ 1 - \bar{\ell}_\phi^{2(I)}(\mathbf{Q}_\phi^{(I)}) &= 1 - \max[q_i^{1(I)}, (q_i^{3(I)})^2] \\ &\geq 1 - \max[q_i^{1(I)}, (q_i^{1(I)} + q_i^{3(I)} - q_i^{1(I)} q_i^{3(I)})^2] \\ &= 1 - \bar{L}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)}) \geq 1 - \bar{B}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)}) \\ &= \min[(1 - q_i^{1(I)})^2, 1 - (q_i^{1(I)} + q_i^{3(I)} - q_i^{1(I)} q_i^{3(I)})^2] \\ 1 - \bar{\ell}_\phi^{3(I)}(\mathbf{Q}_\phi^{(I)}) &= 1 - q_i^{3(I)} \geq (1 - q_i^{3(I)})^2 = 1 - \bar{L}_\phi^{*3(I)}(\mathbf{Q}_\phi^{(I)}) \\ &= 1 - \bar{B}_\phi^{*3(I)}(\mathbf{Q}_\phi^{(I)}). \end{aligned}$$

In Table 4, we give numerical values of the bounds for all combinations of $\lambda = 0.001, 0.01; \mu = 0.01$, and $I = [100, 110], [100, 200], [1000, 1100]$. The lower bounds given are representatives of the different ones listed above. Hence, for level 1, $p_\phi^{1(I)}$ is given, for level 2, $\ell_\phi^{2(I)}(\mathbf{P}_\phi^{(I)})$ and $L_\phi^{*2(I)}(\mathbf{P}_\phi^{(I)})$ are given, whereas for level 3, $p_\phi^{3(I)}$ is given. Similarly, the upper bounds given are representatives of the different ones listed above. Hence, for level 1,

Table 4. Bounds for the Availabilities of the Simple Network System of Fig. 1

	Level 1		Level 2		Level 3	
	Lower	Upper	Lower	Upper	Lower	Upper
$\lambda = 0.001$	0.9903	0.9970	0.8607	0.9886	0.7481	0.8932
$\mu = 0.01$		0.9940	0.9723	0.9880		0.7978
$I = [100, 110]$				0.9880		
$\lambda = 0.001$	0.9667	0.9995	0.7103	0.9979	0.5219	0.9543
$\mu = 0.01$		0.9990	0.8923	0.9979		0.9108
$I = [100, 200]$				0.9979		
$\lambda = 0.001$	0.9522	0.9989	0.6603	0.9954	0.4578	0.9320
$\mu = 0.01$		0.9978	0.8526	0.9952		0.8686
$I = [1000, 1100]$				0.9952		
$\lambda = 0.01$	0.5862	0.8470	0.2020	0.6012	0.0696	0.3685
$\mu = 0.01$		0.7174	0.2685	0.5268		0.1358
$I = [100, 110]$				0.5268		
$\lambda = 0.01$	0.2024	0.9747	0.0196	0.8705	0.0019	0.6401
$\mu = 0.01$		0.9500	0.0196	0.8585		0.4097
$I = [100, 200]$				0.8585		
$\lambda = 0.01$	0.1651	0.9662	0.0137	0.8349	0.0011	0.5937
$\mu = 0.01$		0.9335	0.0137	0.8182		0.3525
$I = [1000, 1100]$				0.8182		

$1 - \bar{\ell}_\phi^{1(I)}(\mathbf{Q}_\phi^{(I)})$ and $1 - \bar{L}_\phi^{*1(I)}(\mathbf{Q}_\phi^{(I)})$ are given, for level 2, $1 - \bar{\ell}_\phi^{2(I)}(\mathbf{Q}_\phi^{(I)})$, $1 - \bar{L}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)})$, and $1 - \bar{B}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)})$ are given, whereas for level 3, $1 - \bar{\ell}_\phi^{3(I)}(\mathbf{Q}_\phi^{(I)})$ and $1 - \bar{L}_\phi^{*3(I)}(\mathbf{Q}_\phi^{(I)})$ are given.

As we see from Table 4, the bounds for this simple network system are amazingly good. Only the upper bounds for the availability to level 2 are getting really bad; that is, for $\lambda = \mu = 0.01$ and $I = [100, 200]$, $[1000, 1100]$. In these cases, we expect more fluctuations between the states of the two components and hence of the system during I . Since our bounds are based on component availabilities and unavailabilities, a lot of information is lost in these cases, and the bounds are hence poor.

Note that we always have $1 - \bar{L}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)}) = 1 - \bar{B}_\phi^{*2(I)}(\mathbf{Q}_\phi^{(I)})$ in Table 4. In fact, we have been unable to find values of λ , μ , and I such that this equality does not hold. For the special case $t_1 = t_2 = \infty$, it is not too hard to show this analytically.

AN OFFSHORE ELECTRICAL POWER GENERATION SYSTEM

The purpose of the offshore electrical power generation system considered in Refs 9 and

10, depicted in Fig. 2, is to supply two nearby oilrigs with electrical power. Both oilrigs have their own main generation, represented by equivalent generators A_1 and A_3 each having a capacity of 50 MW. In addition, oilrig 1 has a standby generator A_2 that is switched into the network in case of outage of A_1 or A_3 . A_2 also has a capacity of 50 MW. The control unit, U , continuously supervises the supply from each of the generators with automatic control of the switches. If, for instance, the supply from A_3 to oilrig 2 is not sufficient, whereas the supply from A_1 to oilrig 1 is sufficient, U can activate A_2 to supply oilrig 2 with electrical power through the standby subsea cables L .

The components to be considered here are A_1, A_2, A_3, U , and L . We let the perfect functioning level M equal 4 and let the set of states of all components be $\{0, 2, 4\}$. For A_1, A_2 , and A_3 these states are interpreted as

- 0: The generator cannot supply any power;
- 2: The generator can supply maximum 25 MW;
- 4: The generator can supply maximum 50 MW.

Note that as an approximation for these generators, we have chosen to describe their

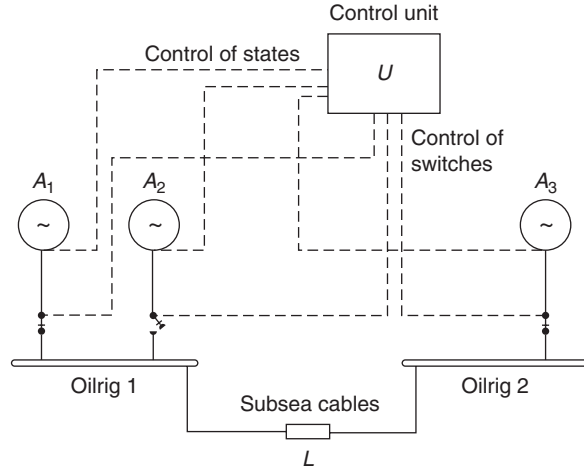


Figure 2. Outline of an offshore electrical power generation system.

supply capacity on a discrete scale of three points. The supply capacity is not a measure of the actual amount of power delivered at a fixed point of time.

The control unit U has the states

- 0: U will by mistake switch off the main generators A_1 and A_3 without switching A_2 on;
- 2: U will not switch A_2 on when needed;
- 4: U is functioning perfectly.

The subsea cables L are actually assumed to be constructed as double cables transferring half of the power through each simple cable. This leads to the following states of L

- 0: No power is transferred;
- 2: 50% of the power is transferred;
- 4: 100% of the power is transferred.

Let us now for simplicity assume that the mechanism that distributes the power from A_2 to platform 1 or 2 is working perfectly, transferring excess power from A_2 to platform 2 if platform 1 is ensured a delivery corresponding to state 4. Now let $\phi(A_1, A_2, A_3, U, L)$ = the amount of power that can be supplied to platform 2. In addition to the states taken by A_1, A_2, A_3 , ϕ can also take the following states

- 1: The amount of power that can be supplied is maximum 12.5 MW;

- 3: The amount of power that can be supplied is maximum 37.5 MW.

Number the components A_1, A_2, A_3, U, L successively 1, 2, 3, 4, 5. Then a little thought leads to

$$\phi(x) = I(x_4 > 0) \min(x_3 + \max(x_1 + x_2 I(x_4 = 4) - 4, 0) x_5 / 4, 4),$$

noting that $\max(x_1 + x_2 I(x_4 = 4) - 4, 0)$ is just the excess power from A_2 which one tries to transfer to platform 2. This is, obviously, an MMS.

FORTHCOMING WORK

In a forthcoming book on multistate reliability theory with applications [16], an in depth treatment of the material presented here is given. Also, a series of new results are developed. For instance, some generalizations of bounds for the availabilities and unavailabilities, to any level, in a fixed time interval given in Ref. 7 have been established. Furthermore, the theory for Bayesian estimation of system reliability, as presented in Ref. 17 for binary systems, has been extended to multistate systems. Finally, a theory for measures of component importance in non-repairable and repairable MSCSs have been developed with accompanying advanced discrete simulation methods. This generalizes recent work given in Refs 18 and 19 for BCSs.

REFERENCES

1. Barlow RE, Wu AS. Coherent systems with multistate components. *Math Oper Res* 1978;4:275–281.
2. El-Newehi E, Proschan F, Sethuraman J. Multistate coherent systems. *J Appl Probab* 1978;15:675–688.
3. Ross S. Multivalued state component reliability systems. *Ann Probab* 1979;7:379–383.
4. Griffith W. Multistate reliability models. *J Appl Probab* 1980;17:735–744.
5. Natvig B. Two suggestions of how to define a multistate coherent system. *Adv Appl Probab* 1982;14:434–455.
6. Block HW, Savits TS. A decomposition for multistate monotone systems. *J Appl Probab* 1982;19:391–402.
7. Funnemark E, Natvig B. Bounds for the availabilities in a fixed time interval for multistate monotone systems. *Adv Appl Probab* 1985;17:638–665.
8. Natvig B. Multistate coherent systems. In: Johnson NL, Kotz S, Read CB, editors. *Encyclopedia of statistical sciences* 5. New York: John Wiley & Sons, Inc.; 1985. pp. 732–735.
9. Gåsemyr J, Natvig B. Probabilistic modelling of monitoring and maintenance of multistate monotone systems with dependent components. *Methodol Comput Appl Probab* 2005;7:63–78.
10. Natvig B, Sørmo S, Holen AT, *et al.* Multistate reliability theory - a case study. *Adv Appl Probab* 1986;18:921–932.
11. Natvig B, Mørch HW. An application of multistate reliability theory to an offshore gas pipeline network. *Int J Reliab Qual Saf Eng* 2003;10:361–381.
12. Lisnianski A, Levitin G. *Multi-state system reliability*. London: World Scientific; 2003.
13. Natvig B. Multistate reliability theory. In: Ruggeri F, Kenett R, Faltin F, editors. *Encyclopedia of statistics in quality and reliability*. New York: John Wiley & Sons, Inc.; 2007. pp. 1160–1164.
14. Natvig B. Strict and exact bounds for the availabilities in a fixed time interval for multistate monotone systems. *Scand J Stat* 1993;20:171–175.
15. Barlow RE, Proschan F. *Statistical theory of reliability and life testing*. Probability models. New York: Holt, Rinehart and Winston; 1975.
16. Natvig B. *Multistate systems reliability theory with applications*. New York: John Wiley & Sons, Inc.; 2011.
17. Natvig B, Eide H. Bayesian estimation of system reliability. *Scand J Stat* 1987;14: 319–327.
18. Natvig B, Gåsemyr J. New results on the Barlow-Proschan and Natvig measures of component importance in nonrepairable and repairable systems. *Methodol Comput Appl Probab* 2009;7:603–620.
19. Natvig B, Eide KA, Gåsemyr J, *et al.* Simulation based analysis and an application to an offshore oil and gas production system of the Natvig measures of component importance in repairable systems. *Reliab Eng Syst Saf* 2009;94:1629–1638.

MULTIVARIATE ELICITATION: ASSOCIATION, COPULAE, AND GRAPHICAL MODELS

ALIREZA DANESHKHAH
TIM BEDFORD
Department of Management
Science, University of
Strathclyde, Glasgow, UK

INTRODUCTION

In the context of decision analysis, elicitation is the process of translating someone's beliefs about some uncertain quantities into a probability distribution. Elicitation is an important subject: it has an important role to play in every application where the data do not make the prior beliefs of the decision maker insignificant.

In practice, we often wish to obtain a joint distribution from the expert for two or more uncertain quantities. It is clear that even for two variables the elicitation of a joint distribution is more complex than that of the marginal distributions discussed in O'Hagan *et al.* [1] and references therein. For more than two variables, the task becomes rapidly even more complex. For instance, for three variables we need to consider probabilities such as $P(X \leq x; Y \leq y; Z \leq z)$. Unfortunately, despite the recent work, it is difficult to find research guiding us in such an elicitation. We focus on the elicitation of a single expert's beliefs about uncertain quantities. To avoid convoluted language, we let the expert be female.

The task becomes much simpler in the special case where the variables are independent. In this case, the elicitation issue would reduce to the elicitation of marginal distributions of the variables separately, and hence we do not need to think about joint distributions. However, in general, we cannot expect independence to hold in practical elicitation tasks. The uncertain variables about

which we wish to elicit judgments are all connected by arising in the same problem, and all falling in the area of expertise of the expert in question, and this is likely to imply that they are connected also in the sense that acquiring knowledge about any one variable would affect the expert's judgment about the others. There are techniques that can be used to transform a problem to make use of partial or complete independence [1–3]. However, these are practically useful only in limited circumstances—there is, therefore, substantial interest in the problem of eliciting a joint probability distribution for two or more variables. In this review, we present the approaches to elicit the models with complex dependency structures between uncertain quantities by using copulae, vines, and other relevant graphical tools.

The outline of the present article is as follows. A number of graphical models in probability theory such as a copula, vine, and Bayesian network have been developed to capture dependencies between random variables. In the section titled “Joint Distribution and Copula,” we show how a multivariate distribution can be constructed using a copula. The expert judgments required to elicit the various marginal distributions and dependence information, in the form of correlations, conditional rank correlations, and so on, are the input to these models. In the section titled “Measures of Bivariate Association,” we review the methods to elicit unconditional and conditional rank correlations. We present the minimum information algorithm as a much more flexible method defining copulae given some constraints specified by an expert, in the section titled “Expert Assessment of Minimally Informative Copula.” A vine is another way to model dependencies between uncertain quantities and can be used to build a higher dimensional distribution from two-dimensional copulae. This model, and ways to elicit conditional rank correlation required to build a joint distribution, is presented in the section titled “Elicitation Multivariate Distribution by Vines.”

JOINT DISTRIBUTION AND COPULA

The construction of a probabilistic model is a key step in many areas of analysis, including decision risk and uncertainty analysis. This can be done by constructing a joint distribution based on marginal and conditional distributions associated with the uncertain quantities. A disadvantage with this approach is that the required number of probability assessments can grow exponentially with the number of variables. The use of a copula to construct joint distributions and pairwise correlations to incorporate dependence among the variables has become very popular. The approach enables us to use an expert's subjective judgments of marginal distributions and correlations. Clemen and Reilly [4] used the copula that underlies the multivariate normal distribution, to construct a complete copula-based joint distribution using assessed rank-order correlations and marginal distributions. They reduced the number of required assessments and relaxed the need to search for conditional independence. The copula approach relies fundamentally on the ability of experts to reliably assess correlations, which may be difficult.

The essence of the copula approach is that a joint distribution, $F(x_1, \dots, x_n)$, of random variables, $X_1, \dots, X_n$, can be expressed as a function of the marginal distributions, $F_1(x_1), \dots, F_n(x_n)$:

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)),$$

where $C(\cdot)$ is a joint distribution with uniform marginals called the *copula*. In addition to this, if each $F_i(\cdot)$ is a continuous distribution function (CDF) then $C(\cdot)$ is unique [4]. If we assume that each F_i is differentiable, we can rewrite the joint density

$$f(x_1, \dots, x_n) = f_1(x_1) \times \dots \times f_n(x_n) \times c(F_1(x_1), \dots, F_n(x_n)) \quad (1)$$

where $f_i(\cdot)$ is the density corresponding to $F_i(\cdot)$ and $c = \partial^n C / (\partial F_1 \dots \partial F_n)$, which is called the *copula density*.

Using this copula-based representation, we can break the elicitation of a joint

probability distribution into smaller, more manageable tasks: the individual marginal distributions and the copula are separately elicited. The copula is a model of the dependence between variables. There are many standard functional forms assumed for copulae.

The multivariate Gaussian copula, which is the unique copula arising from the above expression when we take any multivariate Gaussian distribution F , is a well-known and flexible copula form that can be fully specified through judgments about correlation between the variables. We present the ways to elicit multivariate Gaussian copula parameters indirectly by eliciting the conditional and unconditional rank correlations.

To understand the multivariate Gaussian copula, Clemen and Reilly [4] examine its relationship with the multivariate normal distribution that is typically parameterized in terms of Pearson product-moment correlations. Then, for each element of $\mathbf{R}^*$, the corresponding product-moment correlation r_{ij} for the multivariate normal is calculated. If $\mathbf{R}^*$ is made up of τ_{ij} (Kendall's tau) or made of ρ_{ij} (Spearman's correlation coefficient), Clemen and Reilly [4] present the formulas corresponding to the relationships between (r_{ij}, τ_{ij}) and (r_{ij}, ρ_{ij}) , respectively, as

$$r_{ij} = \sin(\pi \tau_{ij}/2) \quad r_{ij} = \sin(\pi \rho_{ij}/6). \quad (2)$$

We are now able to construct the matrix $\mathbf{R}$ with elements r_{ij} . Note that this has to be a nonnegative semidefinite matrix, and that not all choices of parameters will give such a matrix with these properties (we return to this issue later). Clemen and Reilly [4] then present a joint density based on the multivariate Gaussian copula as

$$\begin{aligned} f(x_1, \dots, x_n | \mathbf{R}^*) &= f_1(x_1) \times \dots \times f_n(x_n) \times |\mathbf{R}|^{-\frac{1}{2}} \\ &\times \exp\{-(\Phi^{-1}[F_1(x_1)], \dots, \Phi^{-1}[F_n(x_n)]) \\ &\times (\mathbf{R}^{-1} - \mathbf{I}) \times (\Phi^{-1}[F_1(x_1)], \dots, \\ &\Phi^{-1}[F_n(x_n)])^T / 2\}, \end{aligned} \quad (3)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard Gaussian

distribution, $\mathbf{R}$ is a $n \times n$ correlation matrix in terms of r_{ij} , and $\mathbf{I}$ is the $n \times n$ identity matrix.

Clemen and Reilly [4] argued that the multivariate Gaussian copula is flexible enough for specifying joint distributions in most applications. However, if a joint distribution is constructed using a multivariate Gaussian copula, then the dependence structure between variables will preserve many of the properties of the multivariate Gaussian distribution, such as constant conditional rank correlation. As we see below, it is possible to use more general constructions to obtain a multivariate distribution.

MEASURES OF BIVARIATE ASSOCIATION

The use of the multivariate normal copula requires the assessment of correlations and marginal probabilities. The assessment of correlations would appear to be a more difficult task than assessing probabilities [1]. Although such judgments are inherently difficult, there are some approaches available to elicit measures of bivariate association. In this section, we briefly describe some approaches that can be used to elicit the copula parameters.

It is reasonable to use statistical techniques based on the expert's familiarity with statistical concepts related to correlation. For instance, an expert might be able to make a judgment regarding the percentage of variance explained, R^2 , that would result from regressing one variable on another. Another possibility is to have the expert view several scatter-plots representing different levels of correlation and select one that is consistent with the expert's belief about the strength of the relationship between the variables [5].

The second approach is called *probability of concordance*. We assess conditional or joint probabilities and relate them to the required measure of dependence. The concordance probability P_C for two independent draws (X_1, Y_1) and (X_2, Y_2) is defined as

$$\begin{aligned} P_C &= \Pr[(X_1 \leq X_2, Y_1 \leq Y_2) \text{ or} \\ &\quad (X_1 > X_2, Y_1 > Y_2)] \\ &= \Pr(X_1 \leq X_2, Y_1 \leq Y_2). \end{aligned}$$

The probability of concordance is simply related to Kendall's τ by the expression $\tau = 2P_C - 1$. This method is recommended for the situation in which a frequency or cross-sectional interpretation is reasonable [3,4].

The method, known as *conditional fractile estimates*, requires conditional estimates and uses these to derive Spearman's ρ . Given random variables X_1 and X_2 , the standard nonparametric regression representation is $E[F_1(x_1) | x_2] = \rho_{X_1 X_2} [F_2(x_2) - 0.5] + 0.5$. In this method, the analyst might have the expert make several conditional estimates and use a least-squares approach to estimate the Spearman's $\rho_{X_1 X_2}$. This method is closely related to a technique called *predictive assessment* [6,7].

One can use a combination of these methods to increase the accuracy in assessing association between two variables. Winkler and Clemen [8] studied gains in accuracy in assessing correlations by averaging different assessments from a single expert and/or from multiple experts. They asked 90 subjects to assess correlations for five pairs of variables using the following methods: strength of relationship; correlation; conditional fractile; concordance probability; joint probability; and conditional probability (see O'Hagan *et al.* [1] and Clemen *et al.* [9] for details of these methods). In the end, they conclude that asking for a correlation is simply a reasonable method, the strength of relationship is simple and also performs quite well, but the method performance based on joint probabilities was the worst in terms of both accuracy and ease of use. However, the accuracy of the better methods mentioned above is not desirable. To improve the accuracy, they propose to use multiple experts or multiple methods in terms of a promising strategy. Their analyses in terms of using averages of correlation assessments from multiple experts and/or multiple methods show a considerable increase in accuracy as they increase the number of experts or methods (see *Eliciting Subjective Probability Distributions from Groups* for details of combination of multiple experts' assessments). As a result, the variability of assessment accuracy decreases significantly

by increasing the number of experts or methods, which leads to a large risk reduction (see Clemen *et al.* [8] for details).

The *association maps* proposed by Gosling [10] are a simple visual tool for eliciting correlations used in building a copula. The basic idea behind this map is to place dots on a two-dimensional shape such that the distance between the points symbolizes the level of association. In other words, the closer the points, the higher the correlation. Although this map gives an impression of the correlation between variables, no numerical values can be obtained to help the analyst to specify the copula parameters or correlation coefficients required to construct the copula-based multivariate distribution.

EXPERT ASSESSMENT OF MINIMALLY INFORMATIVE COPULA

We introduce a methodology based on the minimum information principle that provides a much more flexible approach to defining copulae. Any quantity (or a group of quantities) depending on the two variables in question can be used through *expected values* specified by an expert. The system can give guidance on the range of values for each expected value compatible with the specifications already made by the expert. The method uses a D_1AD_2 algorithm to build the copula that minimizes the information function, given the constraints. This minimum information technique [11] is used in conjunction with expert elicitation of observables to define a copula that represents the decision makers uncertainty about the joint distribution of two random variables.

Bedford and Meeuwissen [11] applied a so-called *DAD* algorithm to produce discretized minimally informative copula with given rank correlation. The same approach can be used whenever we wish to specify the expectation of any symmetric function of U and V . In order to have asymmetric specifications we need to use the more general approach provided in Borwein *et al.* [12]. It states that if A is a positive square matrix (called a *kernel*), then we can find row vectors D_1 and D_2 such that the cell-wise product of

$D_1^T D_2$ and A is doubly stochastic. The theory also exists for continuous functions and, indeed in much more generality.

Suppose there are two random variables X and Y , with cumulative distribution functions F_X and F_Y , respectively. These are the variables of interest that we would like to associate by introducing constraints based on some knowledge about functions of these variables. Suppose there are k of these functions, namely, $h'_1(X, Y), h'_2(X, Y), \dots, h'_k(X, Y)$, and that the expert wishes to specify mean values $\alpha_1, \dots, \alpha_k$ for all these functions, respectively. We can find corresponding functions of the copula variables U and V , defined by $h_1(U, V) = h'_1(F_1^{-1}(U); F_2^{-1}(V))$, and so on, and clearly these should also have the specified expectations $\alpha_1, \dots, \alpha_k$. We form the kernel

$$A(u, v) = \exp(\lambda_1 h_1(u, v) + \dots + \lambda_k h_k(u, v)), \quad (4)$$

where u and v denote the realizations of U and V , respectively.

For practical implementations, we have to discretize the set of (u, v) values such that the whole domain of the copula is covered. This means that the kernel A described above becomes a two-dimensional matrix A and that we seek matrices D_1 and D_2 such that

$$P = D_1 A D_2 \quad (5)$$

is a doubly stochastic matrix P , that is, a discretized copula density. For each vector $(\lambda_1, \dots, \lambda_k)$, we can use the $D_1 A D_2$ algorithm to generate a unique joint density with uniform marginals. This copula gives the vector of functions $(h_1, \dots, h_k)$ an expected value vector which we call $\phi(\lambda_1, \dots, \lambda_k)$. Now, general theory [12] says that this copula is always the unique minimum information copula (with respect to the uniform distribution) giving the expected value vector $\phi(\lambda_1, \dots, \lambda_k)$. Furthermore, the mapping ϕ maps $\mathcal{R}^k$ onto the set of achievable expected value vectors. This set of possible expected value vectors is a convex set, but little else can be said about it in general. The computation details can be found in Bedford [13], Lewandowski [14], and Bedford and Daneshkhah [15].

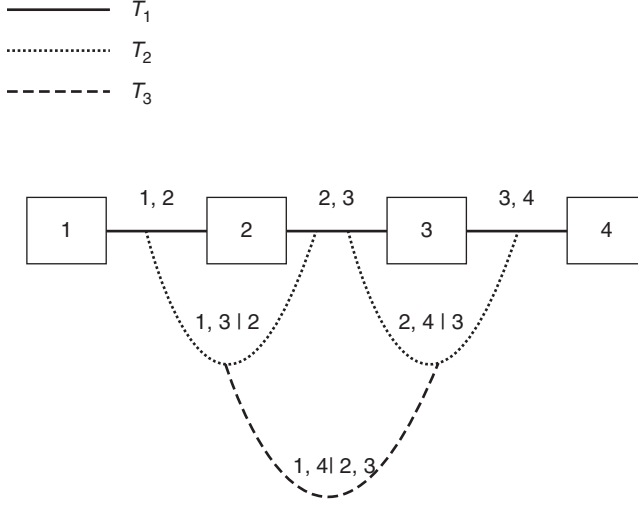


Figure 1. A regular vine with four elements.

ELICITATION MULTIVARIATE DISTRIBUTION BY VINES

It is clear now that copulae represent a natural tool for modeling multivariate distribution. A recent generalization of copulae is that of vines [16,17] in which a multivariate distribution is built from bivariate and conditional bivariate distributions. For example, a so-called *D-vine* for four variables can be constructed with specifications of the sequence $F_1, F_2, F_3, F_4, F_{12}, F_{23}, F_{34}, F_{13|2}, F_{24|3}$, and $F_{14|2,3}$. The graph associated with this D-vine is shown in Fig. 1. Thus, graphically, a vine on n variables is a nested set of trees, where the edges of the j th tree become the nodes of the $(j+1)$ th tree for $j = 1, \dots, n-1$. A *regular vine* on n variables is a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node [2,16,18]. In vines, each edge in the regular vine may be associated with a conditional rank correlation. The conditional rank correlation of X and Y given Z is $r_{X,Y|Z} = r_{X_z,Y_z}$, where (X_z, Y_z) has the distribution of $(X, Y) | Z = z$.

All assignments of (conditional) rank correlations to edges of a vine are consistent, in the sense that a joint distribution exists with those correlations. This situation differs from that when directly specifying pairwise correlations as in the Gaussian copula method discussed above, as in that case there

are algebraic constraints. Any copula with an invertible conditional CDF may be used. In order to specify a joint distribution for a vine, marginal distributions and conditional rank correlations must be specified (or more generally, conditional copulae). We review a procedure for eliciting conditional rank correlations (originally presented in Morales *et al.* [19] for the Bayesian network shown in Fig. 2. This method can be easily used and generalized to other Bayesian belief networks (BBNs) and vines. To elicit $r_{4,3}$, the following question should be asked to the expert:

Q₁. Suppose that the variable X_3 was observed above its q th quantile. What is the probability that also X_4 will be observed above its q th quantile?

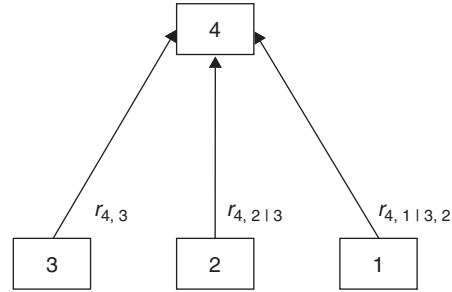


Figure 2. The representation of a Bayesian network with four variables.

The question requires an expert's estimate of $P_1 = P(F_4(X_4) > q \mid F_3(X_3) > q)$ (these sort of elicitation questions have been addressed in O'Hagan *et al.* [1]). The relationship between P_1 and $r_{4,3}$ depends in general on the (conditional) copula; the case of the normal and minimum information copulae is discussed above. To calculate the exceedance probability P_1 , one can numerically integrate the bivariate normal density $\phi(x_3, x_4, \rho_{4,3})$ over the region corresponding to the quantile's exceedance region $[\Phi^{-1}(q), \infty)^2$.

$$P_1 = \frac{1}{1-q} \int_{\Phi^{-1}(q)}^{\infty} \int_{\Phi^{-1}(q)}^{\infty} \phi(x_3, x_4, \rho_{4,3}) dx_3 dx_4, \quad (6)$$

where Φ^{-1} stands for the inverse standard normal CDF. Equation (2) can then be used to fit the value of the normal copula parameter ρ , given the expert estimate of P_1 .

The normal copula parameter ρ can then be found in such a way that it satisfies the expert's conditional probability statement P_1 and then be transformed into the corresponding rank correlation by using Equation (2). It was shown that the choice of copula has a strong impact on the conditional probability when $q \neq 0.5$ [19]. In addition, it was suggested to use $q = 0.5$ in combination with the normal copula (as presented above) to elicit the (conditional) rank correlation.

The following questions, Q_2 and Q_3 , should be asked to the expert in order to assess $r_{4,2|3}$ and $r_{4,1|3,2}$, in terms of the probabilities P_2 and P_3 .

- Q_2 . Suppose that not only variable X_3 but also X_2 was observed above their medians. Now, what is the probability that also X_4 will be observed above its median value?
- Q_3 . Suppose that not only variable X_3 but also X_2 and X_1 were observed above their medians. Now, what is the probability that also X_4 will be observed above its median value?

$$\begin{aligned} P_2 &= P(F_{X_4}(X_4) > 0.5 \mid F_{X_3}(X_3) > 0.5, \\ &\quad F_{X_2}(X_2) > 0.5) \\ P_3 &= P(F_{X_4}(X_4) \geq 0.5 \mid F_{X_3}(X_3) \geq 0.5, \\ &\quad F_{X_2}(X_2) \geq 0.5, F_{X_1}(X_1) \geq 0.5). \end{aligned}$$

A similar method is used to estimate P_2 and P_3 in terms of the multivariate normals [15,19].

In summary, the conditional probability technique can be considered as an aid to elicit conditional and unconditional rank correlations. The dependence relations can be revealed through the successive conditional probability assessments regarding the given nested constraints. To compute these constraints efficiently, so as to support the elicitation, the Gaussian copula is recommended by Morales *et al.* [19]. The same copula is suggested by Clemen and Reilly [4] and Gosling [10] in their studies to elicit the correlations. If the Gaussian copula is not desired then a range of other copulae, including the minimum information copulae mentioned above, can be used. However, the above elicitation procedure only specifies one parameter; so, additional questions would be needed if a copula with more parameters was to be used. For any one-parameter family of copulae, it is possible to precompute the relationship between the parameter and the probability being elicited in the question above. Hence, the analyst is not actually restricted to the Gaussian copula in practice.

REFERENCES

1. O'Hagan A, Buck C, Daneshkhah A, *et al.* Uncertain judgements: eliciting expert probabilities. Chichester: Wiley; 2006.
2. Kurowicka D, Cooke R. Uncertainty analysis with high dimensional dependence modelling. Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence. Chichester: Wiley; 2006. pp. 334–341.
3. Daneshkhah A, Oakley JE. Multivariate elicitation: association, copulae and graphical models. In: Böcker K, editor. Rethinking risk measurement and reporting uncertainty: Bayesian analysis and expert judgement. London: Risk Books Home; 2010. pp. 155–180.

4. Clemen RT, Reilly T. Correlations and copulae for decision and risk analysis *Manage Sci* 1999;45:208–224.
5. Gokhale DV, Press SJ. Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J Roy Stat Soc A* 1982;145:237–249.
6. Winkler RL, Smith W, Kulkarni R. Adaptive forecasting models based on predictive distributions. *Manage Sci* 1978;24:977–986.
7. Kadane JB, Dickey JM, Winkler RL, *et al.* Interactive elicitation of opinion for a normal linear model. *J Am Stat Assoc* 1980;75:845–854.
8. Winkler RL, Clemen RT. Multiple experts vs. multiple methods: combining correlation assessments. *Decis Anal* 2004;1(3):167–176.
9. Clemen RT, Fischer GW, Winkler LR. Assessing dependence: some experimental results. *Manage Sci* 2000;46(8):1100–1115.
10. Gosling JP. Eliciting correlations through association maps. Technical report. Sand Hutton: DEFRA, Central Science Laboratory; Submitted to Management Science, 2009. Available at <http://www.jpgosling.co.uk/Pub/AssMaps.pdf>.
11. Bedford T, Meeuwissen A. Minimally informative distributions with given rank correlation for use in uncertainty analysis. *J Stat Comput Simul* 1997;57(1–4):143–174.
12. Borwein J, Lewis A, Nussbaum R. Entropy minimization, DAD problems, and doubly stochastic kernels. *J Funct Anal* 1994;123:264–307.
13. Bedford T. Interactive expert assignment of minimally-informative copulae. Proceedings of the 3rd International Conference on Mathematical methods in reliability; Trondheim: 2006. The paper is also available at WWW URL:<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.8352>.
14. Lewandowski D. High dimensional dependence: copulae, sensitivity, sampling [PhD. thesis]. Delft University; 2008.
15. Bedford T, Daneshkhah A. Approximating multivariate distributions with vines. Working paper number 2010-02, Management Science department: University of Strathclyde; Submitted to Operations Research. 2010.
16. Bedford T, Cooke RM. Vines - a new graphical model for dependent random variables. *Ann Stat* 2002;30(4):1031–1068.
17. Bedford T, Cooke RM. Probability density decomposition for conditionally dependent random variables modeled by vines. *Ann Math Artif Intell* 2001;32:245–268.
18. Cooke RM. Markov and entropy properties of tree and vine dependent variables. Proceedings of the ASA Section on Bayesian Statistical Science. Alexandria (VA): American Statistical Association; 1997. pp. 166–175.
19. Morales O, Kurowicka D, Roelen A. Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliab Eng Syst Saf* 2008;93:699–710.

MULTIVARIATE INPUT MODELING

BAHAR BILLER
Tepper School of Business,
Carnegie Mellon University,
Pittsburgh, Pennsylvania

INTRODUCTION

As large-scale discrete-event stochastic simulation turns into a tool that is used routinely for the design and analysis of service and manufacturing systems, it becomes critical to provide accurate and automated multivariate input-modeling support. Examples of multivariate inputs include the processing times of work pieces across several work centers, the product demands and the country exchange rates of global supply chains, the annual returns on stock and bond investments in portfolios, and the medical characteristics of organ-transplant donors and recipients. Multivariate input modeling helps to choose an appropriate mathematical representation of these phenomena. However, representing a multivariate input process is a challenge in today's use of simulation software packages because of the need to capture the interdependencies among the many components of the multivariate input process. If the underlying input process could be represented as a sequence of independent random variables having an identical distribution belonging to one of the standard families (e.g., beta, exponential, gamma, lognormal, normal, triangular, uniform, or Weibull), then development of an input model would be a simple task. When such simple models apply, there are a number of software packages that support automated input modeling, including ExpertFit (Averill M. Law and Associates, Inc.), the Arena Input Analyzer (Rockwell Software Inc.), Stat::Fit (Geer Mountain Software Corporation), and BestFit (Palisade Corporation). Good reviews of these models are available in Refs 1 and 2. Unfortunately, simple input models often fall short

of what is needed in a large-scale discrete-event stochastic simulation because the input processes are not inherently independent, either in time sequence or with respect to other input processes in the simulation. Furthermore, the limited shapes represented by the standard families of distributions are not flexible enough to represent most characteristics of the multivariate inputs. We can find a large number of input-modeling problems for which simple models based on independent and identically distributed random variables are inadequate, for example, interarrival times of file accesses in a computer system [3,4], medical characteristics of organ-transplant donors and recipients [5], and returns on different asset classes [6]. Therefore, it is important to develop input models that can represent multivariate input processes with a wide variety of distributional shapes and dependence structures.

A recent review of the methods of developing multivariate input models is available in Ref. 7. In this document, we review the recent advances in multivariate input modeling with focus on the development of high-dimensional input models that lead to fast sampling algorithms for driving large-scale discrete-event stochastic simulations. We specifically consider simulation input models with continuous-valued components and classify them into two types. The first type corresponds to a finite number of dependent random variables, jointly called a *random vector*. The second type captures temporal dependence that arises over a countably infinite sequence of random variables and is studied in terms of time series. More specifically, a k -variate random vector $\mathbf{X} = (X_1, X_2, \dots, X_k)'$ denotes a collection of k random components, each of which is a real-valued random variable. A k -variate time series $\{\mathbf{X}_t = (X_{1,t}, X_{2,t}, \dots, X_{k,t})'; t = 1, 2, \dots\}$, on the other hand, denotes a sequence of random vectors observed at time $t = 1, 2, \dots$. We associate with the random vector, a probability distribution function, called the *joint distribution function*, in the space $\mathcal{R}^k$.

The joint distribution completely defines the stochastic properties of $\mathbf{X}$, and hence it is often used for understanding the effect of the dependence between the components of $\mathbf{X}$ on system performance measures. On the other hand, the time series process $\{\mathbf{X}_t; t = 1, 2, \dots\}$ is a stochastic process that can be studied using the probability distribution, where it imposes on the space of all possible path realizations. Such a probability distribution completely describes the stochastic properties of $\{\mathbf{X}_t; t = 1, 2, \dots\}$. However, time series are often specified in terms of the probability distributions of the components $X_{i,t}$, $i = 1, 2, \dots, k$ and the autocorrelation structure of the sequence, that is, the sequence of correlations between the various elements of $\{X_{i,t}; i = 1, 2, \dots, k, t = 1, 2, \dots\}$. Most of our presentation in this document focuses on capturing the joint distributional properties of random vectors. However, we also provide a brief description of the extension of these models to represent stationary multivariate time series.

We discuss the multivariate input models developed for random vectors in the section titled “Input Modeling for Random Vectors” and the extension of multivariate input modeling to time series in the section titled “Input Modeling for Multivariate Time Series.” The discussion focuses on modeling high-dimensional processes via the use of flexible multivariate density models and presenting fast sampling algorithms well suited for driving large-scale discrete-event stochastic simulations. We conclude with promising areas for multivariate input-modeling research in the concluding section.

INPUT MODELING FOR RANDOM VECTORS

The multivariate normal distribution is the most frequently used distribution in multivariate input modeling. However, it possesses a number of limitations that have motivated the recent advances in multivariate input modeling. We first describe the multivariate normal distribution in the next section and then present the multivariate input models in the order of

increasing flexibility they provide to the simulation user. Specifically, we present the normal-to-anything (NORTA) distribution, the multivariate Johnson translation system, the multivariate Johnson-marginal-based distribution, and the multivariate copula-based distribution in the respective sections. It is important to note that the NORTA distribution is a special case of the multivariate copula-based distribution. The focus of the multivariate input-modeling research has been initially on the NORTA distribution, while it has recently switched to the use of copula theory for extending NORTA to capture a wide variety of dependence structures. Therefore, we discuss the development of NORTA and copula-based distributions in separate sections.

The Multivariate Normal Distribution

A k -dimensional random vector $\mathbf{X} = (X_1, X_2, \dots, X_k)'$ with mean vector $\boldsymbol{\mu}_{\mathbf{X}}$ (i.e., $(E[X_1], E[X_2], \dots, E[X_k])'$) and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$ (i.e., $[\text{Cov}(X_i, X_j); i, j = 1, 2, \dots, k]$) is said to have a nonsingular multivariate normal distribution [8], in symbols $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, if (i) $\boldsymbol{\Sigma}_{\mathbf{X}}$ is positive definite, and (ii) the density function of $\mathbf{X}$ is of the form

$$f(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}_{\mathbf{X}}|^{1/2}} \times \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})' \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \right\}.$$

We can easily generate random variates from this density function to drive stochastic simulations with multivariate normal inputs: (i) We find the Cholesky decomposition $\mathbf{P}$ of $\boldsymbol{\Sigma}_{\mathbf{X}}$. (ii) We generate k independent random variates v_i , $i = 1, 2, \dots, k$ from the standard normal distribution and construct vector $\mathbf{v} = (v_1, v_2, \dots, v_k)'$. (iii) We set $\mathbf{z} = \mathbf{P}\mathbf{v}$ and modify the components of vector $\mathbf{z} = (z_1, z_2, \dots, z_k)'$ via $x_i = \mu_i + \sigma_i z_i$ for the i th component of the multivariate input process to have mean $E[X_i]$ and standard deviation $\sqrt{\text{Cov}(X_i, X_i)}$.

However, there are two important limitations to the use of multivariate normal distribution for input modeling, especially

in large-scale stochastic simulations: (i) The marginal distribution of each component of the input model with the multivariate normal distribution is normal. (ii) Since the interactions among the components is represented in a correlation matrix, it is not possible to capture a wide variety of dependence structures. The first limitation forces the simulation user to assume unimodal, symmetric distributional shapes with a coefficient of kurtosis that is equal to three for all the components of the multivariate input model. In sections titled “NORTA Distribution,” “Multivariate Johnson Translation System,” and “Multivariate Johnson-Marginal-Based Distribution,” we present the multivariate input models that provide the simulation user the flexibility of choosing a different marginal distribution for each component of the multivariate process. The second limitation is, however, associated with the modeling of the dependencies among the components via correlations. Accounting for dependence via correlation is, in fact, a practical compromise made in multivariate input modeling. The use of correlation is often justified by the fact that making it possible for simulation users to incorporate dependence via correlation, while limited, is substantially better than the practice of ignoring dependence. However, correlation is not sufficient to describe the dependence structures that arise in situations, where extreme component realizations occur together; for example, in exchange rates that are among the input processes driving large-scale global supply chain simulations [9]. Recent research has gone beyond the use of correlation as a dependence measure and applied the copula theory that allows the consideration of dependencies between input processes in a general way and the construction of flexible density models that represent a wide variety of dependence structures. We review these copula-based multivariate distributions in the section titled “Multivariate Copula-Based Distribution.”

NORTA Distribution

A general method for obtaining random vector $\mathbf{X} = (X_1, X_2, \dots, X_k)'$ with component

marginal distributions F_{X_i} , $i = 1, 2, \dots, k$ and positive definite correlation matrix $\Sigma_{\mathbf{X}}$ specified via Pearson product-moment correlations $\rho_{\mathbf{X}}(i, j) = \text{Corr}(X_i, X_j)$, $i, j = 1, 2, \dots, k$ is the transformation-based model described in Ref. 10. The central idea behind the development of this model is to transform a standard multivariate normal vector into the desired random vector, which is referred to as having a *NORTA distribution*. Specifically, the approach is to let

$$\mathbf{X} = \left(F_{X_1}^{-1}[\Phi(Z_1)], F_{X_2}^{-1}[\Phi(Z_2)], \dots, \right. \\ \left. \times F_{X_k}^{-1}[\Phi(Z_k)] \right)',$$

where Φ is the cumulative distribution function (cdf) of the standard normal random variable and the base vector $\mathbf{Z} = (Z_1, Z_2, \dots, Z_k)'$ is a standard k -dimensional normal vector with correlation matrix $\Sigma_{\mathbf{Z}}$ composed of Pearson product-moment correlations $\rho_{\mathbf{Z}}(i, j) = \text{Corr}(Z_i, Z_j)$, $i, j = 1, 2, \dots, k$. This modeling approach works for any marginal distribution, although $F_{X_i}^{-1}$, $i = 1, 2, \dots, k$ may have to be evaluated by approximate numerical methods, when there are no exact closed-form expressions.

The challenge associated with the construction of the NORTA distribution is to find a $\Sigma_{\mathbf{Z}}$ that implies the desired correlation matrix $\Sigma_{\mathbf{X}}$ for $\mathbf{X}$. This requires the solution of a total of $k(k-1)/2$ individual correlation-matching problems of the following form, where ϑ_{ρ} corresponds to the standard bivariate normal probability density function (pdf) with correlation ρ :

$$\begin{aligned} \rho_{\mathbf{X}}(i, j) &= \text{Corr}\left(F_{X_i}^{-1}[\Phi(Z_i)], F_{X_j}^{-1}[\Phi(Z_j)]\right) \\ &= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_i)] F_{X_j}^{-1}[\Phi(z_j)] \vartheta_{\rho_{\mathbf{Z}}(i, j)} \\ &\quad \times (z_i, z_j) dz_i dz_j - \mu_i \mu_j}{\sigma_i \sigma_j} \\ &= c_{ij}(\rho_{\mathbf{Z}}(i, j)). \end{aligned} \quad (1)$$

Solving the correlation-matching problems of this form might be a difficult task when the Pearson product-moment correlations are used. Fortunately, the corresponding

function has the properties that allows the implementation of efficient numerical search procedures to find $\rho_{\mathbf{Z}}(i, j)$ within a predetermined precision. Good references for numerical search procedures exploiting these properties are Cario and Nelson [11], Chen [12], and Biller and Nelson [13]. However, when the rank correlation (i.e., the correlation between $F_i(X_i)$ and $F_j(X_j)$) is used, solving a correlation-matching problem is an easy task because the rank correlation-matching problem reduces to $\rho_{\mathbf{X}}(i, j) = (6/\pi) \sin^{-1}[\rho_{\mathbf{Z}}(i, j)/2]$ [14].

The NORTA distribution allows the use of the standard theory for the fast generation of random vectors with given marginal distributions F_{X_i} , $i = 1, 2, \dots, k$ and correlation matrix $\Sigma_{\mathbf{X}}$: (i) We solve $k(k-1)/2$ correlation-matching problems and identify the base correlation matrix $\Sigma_{\mathbf{Z}}$. (ii) We generate realizations from a k -variate standard normal vector of the required length, say n . To obtain the sequence of independent normal vectors, $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$, we first choose k independent univariate standard normal variates $v_{1,t}, v_{2,t}, \dots, v_{k,t}$ for $t = 1, 2, \dots, n$ and then multiply $\mathbf{v}_t = (v_{1,t}, v_{2,t}, \dots, v_{k,t})'$, $t = 1, 2, \dots, n$ by a $(k \times k)$ matrix $\mathbf{P}$ for which $\mathbf{P}\mathbf{P}' = \Sigma_{\mathbf{Z}}$; that is, $\mathbf{Z}_t = \mathbf{P} \mathbf{v}_t$, $t = 1, 2, \dots, n$. (iii) For $i = 1, 2, \dots, k$, we apply the transformation $X_{i,t} = F_{X_i}^{-1}[\Phi(Z_{i,t})]$, $t = 1, 2, \dots, n$ to ensure that the i th component has marginal distribution F_{X_i} .

However, the base correlation matrix $\Sigma_{\mathbf{Z}}$ identified after the solution of the correlation-matching problems might not always be positive semidefinite, violating one of the necessary conditions for a matrix to be a multivariate normal correlation matrix. This problem was postulated in Li and Hammond [15] and was proved to exist in Ghosh and Henderson [16]. Ghosh and Henderson [17] report that the failure of this transformation-based method is relatively rare in moderate dimensions and that the method fails when the correlations lie either on the boundary or in close proximity to the set of correlations achievable for the specified marginals of the process. However, they note that the problem becomes increasingly evident as the dimension of the underlying input process increases.

Considering the flexibility the NORTA distribution provides for multivariate input modeling, a natural question to ask is how to overcome the challenge of obtaining a nonpositive semidefinite base correlation matrix. A close look at the existing literature reveals several solutions to this problem. The first one is to apply the procedures proposed by Lurie and Goldberg [18] and Ghosh and Henderson [17] to find a symmetric, positive definite approximation of the base correlation matrix. The hope is that a multivariate process constructed in this manner has a correlation structure close to the desired one, because the function c_{ij} defined in Equation (1) is known to be continuous under fairly general conditions. Another solution is to use a procedure that does not need to determine the correlation matrix of the multivariate normal random vector. We provide such a procedure in the section titled “Multivariate Johnson-Marginal-Based Distribution,” while we present the solution of using a copula-vine specification for multivariate input modeling in the section titled “Multivariate Copula-Based Distribution.”

Multivariate Johnson Translation System

Here, our objective is to describe the construction of multivariate input models that are particularly easy to use and flexible enough to capture distributional characteristics such as bimodality and heavy tails. Therefore, in this section, we focus on the Johnson translation system [19] that has become one of the most popular flexible system of distributions used in simulation applications in recent years [13,20,21]. We refer the reader to Ref. 7 for other flexible families of distributions that have been given particular attention in the literature.

An input random variable X having a marginal distribution from the Johnson translation system is defined by a cdf of the form

$$F_X(x) = \Phi \left\{ \gamma + \delta g \left[\frac{x - \xi}{\lambda} \right] \right\},$$

where γ and δ are the shape parameters, ξ is a location parameter, λ is a scale parameter, and $g(\cdot)$ is one of the following transformations:

$$g(y) = \begin{cases} \log(y), & \text{for the } S_L \\ & \text{(lognormal) family,} \\ \log(y + \sqrt{y^2 + 1}), & \text{for the } S_U \\ & \text{(unbounded) family,} \\ \log\left(\frac{y}{1-y}\right), & \text{for the } S_B \\ & \text{(bounded) family,} \\ y, & \text{for the } S_N \\ & \text{(normal) family.} \end{cases} \quad (2)$$

There is a unique family, that is, choice of g , for each feasible combination of the coefficient of skewness and the coefficient of kurtosis that are not necessarily equal to zero and three, contrary to the case in the section titled “Multivariate Normal Distribution.” The range of X depends on the family of interest and within each family, a distribution is completely specified by the values of parameters ξ , λ , γ , and δ . Distributional properties of X represented by the Johnson translation system can be found in Refs 19 and 22.

The use of the NORTA distribution with components from the Johnson translation system coincides with the multivariate distribution introduced in Ref. 23 using the following k -variate normalizing translation

$$\mathbf{Z} = \boldsymbol{\gamma} + \delta_g(\boldsymbol{\lambda}^{-1}(\mathbf{X} - \boldsymbol{\xi})) \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma}_Z). \quad (3)$$

In this representation, $\mathbf{Z} = (Z_1, Z_2, \dots, Z_k)'$ is the k -variate standard normal random vector with the $(k \times k)$ covariance matrix $\boldsymbol{\Sigma}_Z$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_k)'$ and $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_k)'$ are the k -variate vectors of shape and location parameters, $\boldsymbol{\delta} = \text{diag}(\delta_1, \delta_2, \dots, \delta_k)'$ and $\boldsymbol{\lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)'$ are the diagonal matrices, whose entries are the shape and scale parameters, and the transformation $g(\cdot)$ is defined as $g(y_1, y_2, \dots, y_k) = (g_1(y_1), g_2(y_2), \dots, g_k(y_k))$, where $g_i(y_i)$, $i = 1, 2, \dots, k$ are defined as in Equation (2). This characterization ensures that the marginal distribution of each component of the multivariate input model is a univariate Johnson distribution.

A random vector with the multivariate Johnson translation system is easy to sample using the normalizing translation defined in Equation (1) once the correlation matrix $\boldsymbol{\Sigma}_Z$ and the marginal distribution

parameters g , $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\xi}$ are given. We simply use the generation procedure presented in the section titled “NORTA Distribution” for the NORTA distribution with $X_{i,t} = F_{X_i}^{-1}[\Phi(Z_{i,t})] = \xi_i + \lambda_i g_i^{-1}[(Z_{i,t} - \gamma_i)/\delta_i]$, $i = 1, 2, \dots, k$, avoiding the need to evaluate $\Phi(\cdot)$. Thus, the Johnson translation system is a particularly good choice for high-dimensional input modeling.

Multivariate Johnson-Marginal-Based Distribution

An alternative method for representing random vector $\mathbf{X}$ with mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$ is to use the multivariate distribution suggested by Stanfield *et al.* [20] for multivariate input modeling. The distinguishing features of this distribution are that its construction does not require the determination of a correlation matrix for a base normal random vector and the component random variables have marginal distributions from the Johnson translation system.

More specifically, we first define a random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)'$, whose components are independent standardized Johnson variates, that is, $E(Y_i) = 0$ and $\text{Var}(Y_i) = 1$ and $\text{Cov}(Y_i, Y_j) = 0$ for $i, j = 1, 2, \dots, k$. The approach of Stanfield *et al.* [20] is to let $L_X = \boldsymbol{\Sigma}_X^{1/2}$ correspond to the lower triangular Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\sigma}_X$ be a diagonal matrix whose entries are the standard deviations of the components of $\mathbf{X}$. Then, random vector $\mathbf{X} = \boldsymbol{\mu}_X + \boldsymbol{\sigma}_X L_X \mathbf{Y}$ has mean $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$. Expressions that relate the skewness and kurtosis of $\mathbf{X}$ to those of $\mathbf{Y}$ are available in Ref. 20. Thus, we can start with the right parameter values for $\mathbf{Y}$ to obtain the given first four moments of $\mathbf{X}$. The computational effort needed for finding the right parameters is relatively less. Hence, it is an attractive generation method for obtaining samples with given mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$ via the utilization of the relation between $\mathbf{X}$ and $\mathbf{Y}$. The resulting data-generation procedure has been shown to always match the first three marginal moments of the components and the correlation matrix of $\mathbf{X}$. However, the

approach may fail to match exactly the fourth marginal moments of the components.

Multivariate Copula-Based Distribution

Despite the recent advances in representing a wide variety of distributional shapes for the components of the multivariate input processes, the input models fall short in representing the dependence structures of simulation inputs assuming extreme realizations together. Example inputs include the returns of financial assets and the exchange rates of global supply chains. Motivated by this particular limitation of the multivariate input models, a different approach to multivariate input modeling for the last few years, has been to utilize the copula theory [24]. Specifically, a k -variate copula is a function that links together the k univariate distribution functions to form a k -variate distribution function. Sklar [25] shows that every joint distribution H_k of k components X_i , $i = 1, 2, \dots, k$ with marginal cdfs F_{X_i} , $i = 1, 2, \dots, k$ can be written as

$$H_k(x_1, x_2, \dots, x_k) = C_k(F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_k}(x_k)), \quad (4)$$

where C_k is a copula that is uniquely defined for continuous marginal cdfs F_{X_i} , $i = 1, 2, \dots, k$. Thus, copula C_k can be interpreted as the dependence structure of H_k .

The main advantage of using a copula for multivariate input modeling is that it remains invariant under strictly increasing transformations of the component random variables, simplifying the random vector generation. To sample a random vector $\mathbf{X} = (X_1, X_2, \dots, X_k)'$ with joint distribution H_k as defined in Equation (4), we first generate a multivariate uniform vector $\mathbf{U} = (U_1, U_2, \dots, U_k)'$ from copula C_k and then set the random vector $\mathbf{X}$ to be $(F_{X_1}^{-1}(U_1), F_{X_2}^{-1}(U_2), \dots, F_{X_k}^{-1}(U_k))'$. Thus, copulas provide an easy method to model and generate random vectors as they represent the dependence between the components of the random vector, independent of the marginal distributions.

This kind of dependence modeling is becoming increasingly popular in cases

where correlation falls short of capturing the complex interactions and interdependencies among the simulation inputs. It is important to note that the use of a normal copula for dependence modeling reduces the multivariate distribution of this section to the NORTA distribution as well as the multivariate distributions of the sections titled “Multivariate Johnson Translation System” and “Multivariate Johnson-Marginal-Based Distribution” under certain assumptions about the component marginal distributions. However, a close look at the existing literature reveals numerous parametric families of bivariate copulas that emphasize different joint distributional properties [24,26]. While these parametric families can be coupled with arbitrary marginal distributions to form bivariate input processes with a wide variety of dependence structures, the selection of a multivariate copula appropriate for multivariate input modeling is a challenging problem, limited by the small number of copula parameters that are not sufficient for representing all of the interactions among the components of the multivariate input process.

However, the multivariate copula resulting from the application of the copula-vine specification overcomes this challenge, providing the simulation user full flexibility not only in achieving a wide variety of distributional shapes for the components, but also in admitting various dependence structures among the components. The copula-vine specification introduced by Cooke and his coauthors [27–30] are described in detail in Ref. 31. Specifically, a vine associated with a k -variate random vector is a nested set of $k - 2$ spanning trees where the edges of tree j are the nodes of tree $j + 1$, starting with a tree on a graph with the k components as nodes. A regular vine on the k components is a vine in which two edges in tree j are joined by an edge in $j + 1$ only if they share a common node in the j th tree for $j = 1, 2, \dots, k - 2$. Thus, there are a total of $k(k - 1)/2$ edges in a regular vine and the representation of each edge with a different distribution leads to the construction of a k -variate distribution.

Cooke [27] provides a generation procedure that is easy to implement for the fast

sampling of a random vector from the regular vine. Sampling starts with k independent uniform samples and proceeds by traversing the vine in a specific order and applying successive inversions of the conditional distributions derived from the bivariate copulas of each edge [7]. The joint distribution of the random vector generated by this procedure is unique. Cooke [27] demonstrates how to implement this procedure by associating minimum information copulas with the vine edges, while Bedford and Cooke [29] associate elliptical copulas with the edges of the vine. The use of elliptical copulas has the advantage that the conditional correlations of the edges are related by a set of recursive equations, to the set of correlations induced between the components of the random vector generated by the vine-traversing procedure. Therefore, a set of conditional rank correlations of a regular vine can be derived for any desired correlation matrix and the method can be used for matching a given $k \times k$ rank correlation matrix to uniform marginals. Additionally, conditional distributions of elliptical copulas can be derived in explicit form and are very easy to invert. Thus, the generation procedure involves applying $k(k-1)/2$ of these inverses to k independent uniform variates, and it is fast and efficient. This procedure can match any feasible correlation matrix with uniform marginals in the trivariate case [30]. The method also comes very close to achieving any feasible correlation matrix in higher dimensions. And recently, Biller [32] has associated Archimedean copulas, which have the ability to capture different levels of (tail) dependencies in the lower-quadrant tail and upper-quadrant tail of the joint distribution, with the edges of the vine. This leads to the development of a fast sampling algorithm for multivariate input processes whose dependence structures exhibit positive tail dependencies.

INPUT MODELING FOR MULTIVARIATE TIME SERIES

Much of the previous work on multivariate time series input modeling is based on univariate and bivariate time series models.

A detailed review of these models can be found in Ref. 7. In this document, the focus is on the representation of high-dimensional time series input processes via the use of flexible multivariate density models that work for a wide variety of distributional properties and dependence structures. Such comprehensive support for multivariate time series input modeling comes from the vector-autoregressive-to-anything (VARTA) input model developed for k -variate p th-order stationary time series [13]. This model is superior to prior stochastic input models because of its ability to characterize dependencies both in time sequence and with respect to other input processes in the simulation. It also represents the key features of the components through the Johnson translation system.

Specifically, the p th-order VARTA process defines a stationary, k -variate time series $\{X_{i,t}; t = 0, 1, 2, \dots\}$, $i = 1, 2, \dots, k$ with marginal cdfs F_{X_i} , $i = 1, 2, \dots, k$ and product-moment correlations $\rho_{\mathbf{X}}(i, j, h) = \text{Corr}(X_{i,t}, X_{j,t-h})$, $h = 0, 1, 2, \dots, p$, $i, j = 1, 2, \dots, k$. It is defined by an inverse transformation from a standardized normal vector autoregressive process $\{\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, \dots, Z_{k,t})'; t = 0, 1, 2, \dots\}$ of order p with representation $\mathbf{Z}_t = \sum_{h=1}^p \alpha_h \mathbf{Z}_{t-h} + \mathbf{Y}_t$, $t = 0, 1, 2, \dots$, where α_h , $h = 1, 2, \dots, p$ are k -variate autoregressive coefficient matrices and $\{\mathbf{Y}_t = (Y_{1,t}, Y_{2,t}, \dots, Y_{k,t})'; t = 0, 1, 2, \dots\}$ is a sequence of independent and identically distributed k -variate normal random vectors, each of which has a zero mean vector and a k -variate covariance matrix denoted by $\Sigma_{\mathbf{Y}}$. The selection of $\Sigma_{\mathbf{Y}}$ as $\Sigma_{\mathbf{Z}}(0) - \sum_{h=1}^p \alpha_h \Sigma'_{\mathbf{Z}}(h)$, where $\Sigma_{\mathbf{Z}}(h)$ is the lag- h autocorrelation matrix, insures the standard normality of the base component time series $\{Z_{i,t}; t = 0, 1, 2, \dots\}$, $i = 1, 2, \dots, k$.

The major challenge in the construction of the VARTA process is to select the base correlations $\rho_{\mathbf{Z}}(i, j, h)$, $i, j = 1, 2, \dots, k$, $h = 0, 1, 2, \dots, p$ that gives the prespecified input correlations $\rho_{\mathbf{X}}(i, j, h)$, $i, j = 1, 2, \dots, k$, $h = 0, 1, 2, \dots, p$ by solving $pk^2 + k(k-1)/2$ correlation-matching problems of the form in Equation (1) with X_i replaced by $X_{i,t}$ and X_j replaced by $X_{j,t-h}$. Notice that the VARTA distribution reduces to the NORTA distribution

for $p = 0$. Therefore, the numerical search procedures of the section titled “NORTA Distribution” can also be used for solving the correlation-matching problems of the VARTA distribution. Furthermore, we can easily generate multivariate time series of length n with component marginal distributions F_{X_i} , $i = 1, 2, \dots, k$ and input autocorrelation matrices $\Sigma_{\mathbf{X}}(h) = [\rho_{\mathbf{X}}(i, j, h); i, j = 1, 2, \dots, k, h = 0, 1, 2, \dots, p]$: (i) We solve the correlation-matching problem for the base autocorrelation matrix $\Sigma_{\mathbf{Z}}(h)$ that would match the prespecified input autocorrelation matrix $\Sigma_{\mathbf{X}}(h)$ for $h = 0, 1, 2, \dots, p$. (ii) We obtain base process parameters $\alpha_1, \alpha_2, \dots, \alpha_p$ from $\Sigma_{\mathbf{Z}}(h)$, $h = 0, 1, 2, \dots, p$ by using the multivariate Yule–Walker equations [33]. Then, we compute covariance matrix $\Sigma_{\mathbf{Y}}$ as equivalent to $\Sigma_{\mathbf{Z}}(0) - \sum_{h=1}^p \alpha_h \Sigma_{\mathbf{Z}}'(h)$. (iii) We obtain starting values for $\mathbf{z}_{-p+1}, \mathbf{z}_{-p+2}, \dots, \mathbf{z}_0$ using $\Sigma_{\mathbf{Z}}(h)$, $h = 0, 1, \dots, p$. Additionally, we generate a series of normal vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ using $\Sigma_{\mathbf{Y}}$, so that, we can generate time series $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n$ recursively from $\mathbf{z}_t = \sum_{h=1}^p \alpha_h \mathbf{z}_{t-h} + \mathbf{y}_t$ for $t = 1, 2, \dots, n$. (iv) Finally, we transform the generated base time series $\{z_{i,t}; i = 1, 2, \dots, k, t = 1, 2, \dots, n\}$ with standard normal marginal distributions into the desired process $\{x_{i,t}; i = 1, 2, \dots, k, t = 1, 2, \dots, n\}$ with cdfs F_{X_i} , $i = 1, 2, \dots, k$ by implementing the inverse transformation $x_{i,t} = F_{X_i}^{-1}[\Phi(z_{i,t})]$ for $i = 1, 2, \dots, k$ and $t = 1, 2, \dots, n$.

However, after solving the correlation-matching problems, the resulting base autocorrelation matrix might not be positive definite and/or the resulting autoregressive coefficients might violate the stability condition, that is, the roots of the reverse characteristics polynomial $|\mathbf{I}_{k \times k} - \alpha_1 \mathbf{z} - \alpha_2 \mathbf{z}^2 - \dots - \alpha_p \mathbf{z}^p| = 0$ might not lie outside of the unit circle in the complex plane ($\mathbf{I}_{k \times k}$ is the $k \times k$ identity matrix), and/or the resulting covariance matrix $\Sigma_{\mathbf{Y}}$ might not be positive definite. Therefore, there exist sets of marginals with feasible input autocorrelation structures that are not representable by the VARTA transformation. However, the copula-based multivariate time series input model of Biller [32] includes the VARTA model as a special case and works for sets of marginals with feasible

autocorrelation structures that are not representable by the VARTA transformation.

CONCLUSION

Since simulation inputs form the core of every stochastic simulation, it is important to develop multivariate input models that lead to improved system representations for accurate results and better decisions. Research on multivariate input modeling for stochastic simulations is far from being complete. There are many interesting problems that promise both methodological and practical impacts.

Most of the multivariate input-modeling research focuses on the representation of stationary input processes. However, the joint distributional properties of multivariate time series processes often vary with time. Therefore, it is critical to develop multivariate input models that represent multivariate time series input processes with possibly many components having time-varying distributional properties. It is also important that the resulting models apply to a wide variety of dependence structures and lead to the development of fast sampling algorithms for the input models to drive large-scale discrete-event stochastic simulations.

The presentation in this document has focused on the modeling and variate-generation aspect of the multivariate input modeling. Another important aspect of input modeling is to fit models to historical data. A close look at the existing literature indicates a lack of statistically valid fitting algorithms well suited for large-scale discrete-event stochastic simulations. Development of algorithms that produces estimates with good statistical properties is a promising area of research.

REFERENCES

1. Vincent S. Input data analysis. In: Banks J, editor. Handbook of simulation. New York: John Wiley & Sons, Inc.; 1998. pp. 55–91.
2. Law AM. Simulation modeling and analysis. New York: McGraw-Hill; 2007.

3. Ware PP, Page TW, Nelson BL. Automatic modeling of file system workloads using two-level arrival processes. *ACM Trans Model Comput Simul* 1998;8:305–330.
4. Livny M, Melamed B, Tsiolis AK. The impact of autocorrelation on queuing systems. *Manag Sci* 1993;39:322–339.
5. Pritsker AAB, Martin DL, Reust JS, *et al.* Organ transplantation policy evaluation. In: *Proceedings of the 1995 Winter Simulation Conference*. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 1995. pp. 1314–1323.
6. Biller B, Foster J. A copula-based multivariate input model for financial time series. Working Paper, Pittsburgh (PA): Tepper School of Business, Carnegie Mellon University; 2009.
7. Biller B, Ghosh S. Multivariate input processes. In: Nelson BL, Henderson SG, editors. *Handbooks in operations research and management science: simulation*. Amsterdam: Elsevier Science; 2006.
8. Tong YL. *The multivariate normal distribution*. New York: Springer; 1990.
9. Kouvelis P, Su P. *The structure of global supply chains: research monograph for book series on foundations and trends in technology, information, and operations management*. Boston-Delft: NOW; 2007.
10. Cario MC, Nelson BL. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Northwestern University, Evanston (IL): Working Paper, Department of Industrial Engineering and Management Sciences; 1997.
11. Cario MC, Nelson BL. Numerical methods for fitting and simulating autoregressive-to-everything processes. *INFORMS J Comput* 1998;10:72–81.
12. Chen H. Initialization for NORTA: Generation of random vectors with specified marginals and correlations. *INFORMS J Comput* 2001;13:312–331.
13. Biller B, Nelson BL. Modeling and generating multivariate time series input processes using a vector autoregressive technique. *ACM Trans Model Comput Simul* 2003;13:211–237.
14. Kruskal W. Ordinal measures of association. *J Am Stat Assoc* 1958;B53:814–861.
15. Li ST, Hammond JL. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Trans Syst Man Cybern* 1975;5:557–561.
16. Ghosh S, Henderson SG. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Oper Res* 2002;50:820–834.
17. Ghosh S, Henderson SG. Behavior of NORTA method for correlated random vector generation as the dimension increases. *ACM Trans Model Comput Simul* 2003;13:276–294.
18. Lurie PM, Goldberg MS. An approximate method for sampling correlated random variables from partially-specified distributions. *Manag Sci* 1998;44:203–218.
19. Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika* 1949a;36:149–176.
20. Stanfield PM, Wilson JR, Mirka GA, *et al.* Multivariate input modeling with Johnson distributions. In: *Proceedings of the 1996 Winter Simulation Conference*. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 1996. pp. 1457–1464.
21. Mirka GA, Glasscock NF, Stanfield PM, *et al.* An empirical approach to characterizing trunk muscle coactivation using simulation input modeling techniques. *J Biomech* 2000;33:1701–1704.
22. Johnson ME. *Multivariate statistical simulation*. New York: Wiley; 1987.
23. Johnson NL. Bivariate distributions based on simple translation systems. *Biometrika* 1949b;36:297–304.
24. Nelsen RB. *An introduction to copulas*. New York: Springer; 1999.
25. Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ Inst Stat Univ Paris* 1959;8:229–231.
26. Joe H. *Multivariate models and dependence concepts*. London: Chapman & Hall; 1997.
27. Cooke R. Markov and entropy properties of tree and vines-dependent variables. *Proceedings of the ASA Section of Bayesian Statistical Science*. Anaheim (CA); 1997.
28. Bedford T, Cooke RM. Probability density decomposition for conditionally dependent random variables modeled by vines. *Ann Math Artif Intell* 2001;32:245–268.
29. Bedford T, Cooke RM. Vines—a new graphical model for dependent random variables. *Ann Stat* 2002;30:1031–1068.
30. Kurowicka D, Cooke R. A parameterization of positive definite matrices in terms of

- partial correlation vines. *Linear Algebra Appl* 2003;372:225–251.
31. Kurowicka D, Cooke R. Uncertainty analysis with high dimensional dependence modeling: Wiley series in probability and statistics. Berlin, Germany: Wiley & Sons; 2006.
32. Biller B. Copula-based multivariate input models for stochastic simulation. *Oper Res* 2009;57:878–892.
33. Lütkepohl H. Introduction to multiple time series analysis. New York: Springer; 1993.

MUSIC AND OPERATIONS RESEARCH

ELAINE CHEW
Viterbi School of Engineering,
University of Southern California,
Los Angeles, California

CHRISTOPHER RAPHAEL
School of Informatics and Computing,
Indiana University, Bloomington,
Indiana

INTRODUCTION

Music listening and music making are important human experiences that enrich our lives and stimulate our brains. Advances in modeling and computing technologies are opening doors to new ways to create and understand music. Devices for interacting with music, ranging from mobile phones to video games, pervade our lives. The need for efficient ways to represent, organize, manipulate, and access digital music media has never been greater. In response to this growing need, the number of researchers and quantity of publications in music information research has grown exponentially over the past decade. Some recently founded conferences and societies are listed in Ref. 1. Operations research (OR) has much to offer the music research community; modeling techniques such as dynamic programming (DP), network optimization, and linear and nonlinear programming are already employed in music research. Much more can be done and improved. It is our goal to survey some advances at the intersection of music and OR, as a starting point for the reader to delve further into this exciting new field.

Music, broadly considered to be organized sound, lends itself readily to numeric representation. Sound is produced by vibrations, often at quantifiable frequencies. When these frequencies exist, they give rise to pitched sound. A sound with a perceived pitch is called a tone. A tone of a given duration can

be represented by a note. Like many human experiences, music exists in time, and unfolds over time. A sequence of notes form a melody, which is a single musical voice; the sequence of note durations gives the rhythm of the melody. A complete musical composition can be regarded as a layered collection of musical voices. Vertical collections of synchronous sounds form chords or harmonies; a sequence of harmonies forms a chord progression.

The human brain, when exposed to music, learns to perceive more abstract musical structures such as key and meter. A musical sequence often generates the perception of key and meter, unless a composer explicitly avoids these constructs. Key refers to the pitch set used in a composition; it also establishes a hierarchy of perceived stability amongst the pitches. Articulation of the most stable of these pitches gives rise to a sense of musical closure, either of a phrase or larger section of a piece. Meter refers to the time structure of the music, with strong and weak beats occurring over an underlying grid of regular pulses. Meter, amongst other things, is how we can tell the difference between a waltz and a tango. Abstract structures such as key and meter also evolve over time, and can themselves be modeled as sequences.

Creating new music, whether by improvisation or composition, can be considered the mapping of a sequence of decisions over a space of possible choices, constrained by aesthetics, common practice, or arbitrary rules. A performance, too, can be considered the result of a series of decisions on variables such as fingering, loudness, timbre, and deviations from the prescribed rhythm, constrained somewhat but not entirely by the composer's markings in the score, and the musical structure.

With the above prelude, it should not come as a surprise that music analysis, music perception and cognition, music composition and improvisation, and expressive performance, can be modeled mathematically and computationally as optimization problems. With the proliferation of and

widespread access to digital music, automatic ways to quantify and describe musical properties have become indispensable tools for managing enormous amounts of digital music.

The gamut of OR techniques that have been applied to music information research range from the traditional optimization areas of linear and nonlinear programming, DP, combinatorial optimization, network flows, and Markov models, to machine learning and hidden Markov models (HMMs). The first part of this article gives a brief overview of a range of mathematical problems and models employed in music. The sequential nature of music makes it particularly amenable to DP models. The second part of this article delves into the use of DP across a number of musical applications.

MUSIC REPRESENTATION

Mathematical and computational models of musical problems must first begin with a mathematical representation of music. Music information researchers categorize music representations into *symbolic* and *audio* data. Symbolic representations describe music through possibly layered or synchronous note sequences, represented in terms of discrete note attributes such as pitch, position, and duration. Examples of symbolic representations include text-based human-readable formats such as ***kern* (*kern.humdrum.net*), *MuseData* (*www.musedata.org*), *Guido* (*www.informatik.tu-darmstadt.de/AFS/GUIDO*), and *Music XML* (*www.recordare.com/xml.html*), as well as the more free-form musical instrument digital interface (MIDI) format which contains event-based information such as pitch number, onset time, onset velocity, and offset time. Guido, MusicXML and other representations are reviewed in Ref. 2. Symbolic information has the advantage of being discrete, thus accurately representing our abstract understanding of pitch, onset position, and note length.

Audio data refers to music sound files. Analysis of audio information is typically preceded by a spectral decomposition of the

signal. Researchers who start from the audio music signal often tackle problems similar to those who work from symbolic information, with the exception that audio signals present the added challenge of recovering the notes, chords, rhythms, or other discrete musical attributes. The mapping of audio to symbolic information (such as the score) is itself an interesting optimization problem that will be discussed in the latter half of this article.

Pitches, Chords, and Keys

Apart from straightforward numerical quantification of a musical property, representations can be designed to abstract patterns from the basic numeric quantities, or to model perceived relations between the entities. For instance, a straightforward pitch representation, such as MIDI, labels each pitch activated by a lever on a piano keyboard with a single number. Pitches can be grouped into equivalence classes using modulo 12 arithmetic so that all pitches whose frequencies are related by factors of 2^n , where $n \in \mathbb{Z}$, are grouped under the same class (i.e., all the different notes related by an octave). Numerous articles exist that explore the group theoretic properties of these pitch and interval classes [3–5].

The neo-Riemannian framework [6,7] employed by music theorists to chart chord progressions, consists of a network of pitches in two dimensional space—the *tonnetz* (tone network)—in which cliques are triads (three-note chords), and neighboring triads share two common pitches. Recent work by Callender *et al.* generalizes this space to all possible chords in non-Euclidean space [8,9].

Music psychologists have proposed spatial representations of pitches that reflect their perceived relations, for example Ref. 10. A historical review can be found in Ref. 11. Krumhansl's pitch, chord, and key maps [12] are obtained by applying multidimensional scaling to experimental data. The experimentally obtained pitch profiles serve as prototype templates for key finding. DP approaches to key finding will be discussed in the latter part of this article. Toiviainen [13] uses self-organizing maps to track and visualize the evolution of key in Krumhansl's key space.

Cognitive scientists Longuet-Higgins and Steedman independently discovered the harmonic network, equivalent to the *tonnetz*, and used it to automate the task of key finding [14]. Inspired by this work and von Neumann's interior point method, [15], Chew used the interior of the three-dimensional configuration of the harmonic network to represent pitches, chords, and keys. The calibration of this "spiral array" is a nonlinear constraint satisfaction problem. The model is used to build more robust key finding algorithms [16,17]. Chew and François use an exponentially decaying function to track long-term (key) and short-term (chord) tonal contexts in the spiral array, in their interactive tonal visualization system, MuSA.RT [18].

In the related realm of convex optimization, Kassakian and Wessel [19] propose a solution method for computing positions in Euclidean space for new musical entities, citing examples for key and melody. Suppose we have a set of data points in Euclidean space, where distance corresponds to dissimilarity. Given a new data point and dissimilarity measures to the existing points, the nonconvex problem is one of finding a position for the new point such that it minimizes the sum of squared errors between desired dissimilarities and achieved distances. The problem's dual with Lagrangian relaxation is convex, and the dual of the dual, also convex, is solved using semidefinite programming.

Melodies

We now consider mathematical models for note sequences. A sequence of notes forms a melody. Melodies can be segmented into phrases, and both melodies and phrases can contain identifiable repeating patterns called motifs. The prototypical example of a motif is the opening four notes of Beethoven's Fifth Symphony. The motif repeats throughout the composition, both locally and globally, in some combination of the following guises: verbatim, transposed, or inverted. Because melodies behave very much like sequences of text, any model for representing, indexing, analyzing, comparing, and generating text strings, such as Markov models, suffix trees, and edit distance, can also be applied to

melodic sequences. The fundamental building blocks of these models tend to be the basic units of pitch numbers or intervals, and durations or duration ratios.

Melody analysis algorithms often cannot be easily dissociated from the representation. Buteau has proposed topological spaces for motivic analysis [20], and Lartillot network-based representations for pattern extraction [21]. Combinatorial explosion remains a challenge for these techniques. Conklin and Anagnostopoulou propose a viewpoints method for discovering recurrent patterns in segmented melodies [22]. Inspired by linguistics models, Lerdahl and Jackendoff use tree structures to model the perceptual groupings of not only pitches, but also rhythm and meter [23]. Rens Bod analyzes melodic phrase structure with the linguistically oriented probabilistic context-free grammars [24]. Other generative models are covered in later sections on composition and improvisation.

Rhythm and Meter

At the most basic level, the time structure of a note can be represented as an onset with some duration. Musical rhythm can be indicated with a sequence of numbers representing onsets or durations. It can also be represented as a sequence of duration ratios. Toussaint [25] uses geometric shapes inscribed in a circle to represent musical rhythms, similar to the approach taken by Schell [26] to represent chords.

Durations or onset sequences give rise to beats (regular pulses) and meter (periodic groupings and accent patterns that persist over time). For example, marches tend to be in duple meters, typically having four beats in a measure or bar, with accents on the first and third beats, the first being stronger than the third, and waltzes in triple meter, having three beats to a measure. The rate of beats is called tempo, and is a parameter that is manipulated in expressive performance, which will be addressed in a later section.

In Lerdahl and Jackendoff [23], we have an example of a model for musical rhythm and meter that parses adjacent notes and builds a tree structure. Other models, such as Volk's inner metric analysis [27] and Smith

and Honing's wavelets approach [28] consider events over multiple time scales to uncover hierarchical metric patterns.

COMPOSITION AND IMPROVISATION

Composition can be expressed in terms of decision variables, constraints, and objective functions. This section surveys some mathematical models for musical problems in composition.

Chord Progressions: A Shortest Path Problem

A typical task in composition is the assignment of chords to harmonize a melody. A rudimentary criterion of a successful composition is that transitions between chords must be smooth; that is, the interval between successive notes in a "voice" should be as small as possible. This is sometimes referred to as parsimonious voice leading.

Consider the problem of assigning accompanying chords to the notes of a melody. For each of the N notes in the melody, one can enumerate all plausible accompanying chords. Each chord is represented by a node while an arc connects each node at time n (coincident with the n th note) with each node at time $n + 1$. A cost is now associated with each arc that reflects the distance between the adjacent chords in terms of voice leading. The problem of determining the accompanying chords with smooth voice leading becomes one of minimizing the total cost of the arcs traversed from time 1 through N . Linear programming or DP are both valid solution methods for this shortest path problem. We shall elaborate on DP as a solution technique across multiple musical problems in the latter part of this article.

Distance minimization, as the only criterion, can lead to trivial or uninteresting solutions. Constraints offer an effective way to capture more aspects of creative composition. For example, using constraints, we could require that the melody note be a member of the chord accompanying it, or that accompanying one of its neighbors (to model what are called passing tones and suspensions); or, we could have a model that prescribes accompanying chords, but not their inversions.

Constraints in Computer-Assisted Composition

Constraints present a natural way to model the composition process. Sometimes the constraints or rules are decided by common practice, sometimes by the desired auditory effect, and sometimes they are arbitrarily created by a composer to help define and shape the composition. To quote from Stravinsky [29, p. 152],

"My freedom consists in my moving about within the narrow frame that I have assigned myself for each one of my undertakings. . . . The more constraints one imposes, the more one frees one's self of the chains that shackle the spirit."

The numbers and types of musical constraints are limited only by one's imagination. In the preceding section, we mentioned some examples of constraints in harmonizing a melody. A survey of constraint-based models for harmonization can be found in Ref. 30. Truchet and Codognet present a catalog of fourteen constraint satisfaction problems in computer-assisted composition that are solved using adaptive search [31]. Some of these problems are described below.

Network Problems in Composition

Apart from minimizing voice leading distance in chord transitions, sometimes composers wish to constrain chord progressions in ways to achieve a slightly different auditory effect.

The following two examples are given in Ref. 31. A composer may wish to find an ordering of a sequence of chords so that any two consecutive chords have a fixed number of common tones, or some maximum or minimum number of common tones. The appropriate network can be constructed with the chords as nodes, and having an arc between any two nodes that satisfy the common tones criterion or criteria. The problem then becomes one of finding a Hamiltonian path through the graph, which can be solved using the traveling salesman problem algorithms.

Another composer may wish to find an ordering that maximizes the number of common tones between adjacent chords. Again, let the chords be nodes, and the arc between

each pair of nodes having a cost reflecting the number of common tones between the chords. The problem becomes a traveling salesman problem when one negates the costs.

In 2004, composer Tom Johnson (personal communication) posed the following problem to his mathematician friends: for all six-note chords from a nine-note scale, find the largest set of chords that have m tones in common with each other. Suppose that each chord is a node, and an arc connects each pair of chords having m common tones. The problem can be re-framed as one of finding the maximal clique in this graph. More mathematical puzzles for composition, and their musical solutions can be found at www.editions75.com.

Constraints on Melody and Rhythm

The fugue subjects in Bach's *48 Preludes and Fugues* are perfect examples of melodies that satisfy stringent compositional constraints. The fugue subjects are designed so that some combination of multiple instances of the subject, either in its original form, transposed (played higher or lower), inverted, reversed (in retrograde), augmented (stretched out in time), or diminished (squished in time), whether simultaneously or staggered, sound harmonious together.

Some compositional techniques, such as *serialism*, seek musical sequences that explicitly avoid repetition. A recent example of such a problem arises in Ref. 31 as follows: find permutations of the numbers $\{1, \dots, n\}$ such that the difference between consecutive numbers is never repeated, that is no interval is repeated in the sequence.

Examples of constraints one could impose on rhythm include the following: finding parallel streams of duration patterns so that no two notes entering simultaneously have the same durations; or, creating streams of durations consisting of pairs of consecutive numbers in the Fibonacci sequence such that the sum of durations in each stream is a constant.

Generating Music: Markov Models

There is a long history of composers using formal mathematical methods to generate music. The methods employed include

probability (such as in Mozart's dice game), stochastic processes, game theory [32,33], and chaos theory [34]. Markov models are the traditional generators of choice for emulating style [35,36], while a newer model uses a data structure called a factor oracle to learn patterns from musical sequences [37].

Markov chains have formed the backbone of a number of music generation systems. The states can be any combination of musical entities, such as pitches, chords, durations, pitches with durations (notes), and so on. Consider the concrete example of learning from a sample melody. For each note, one can calculate the probability that it is followed by any other note to create a first-order Markov chain. Generalizing from this concept, one could construct an n th order Markov model, where $n > 1$, from any music data stream.

Markov models are particularly suited to generating music in a given style, as captured by the transition probabilities. They have been used to analyze and generate Palestrina style counterpoint [38], and to produce style-specific accompaniment [36]. Markov models, of fixed or variable order, have also been used in human-machine improvisation systems to generate music in the style of some user-provided musical input [35,37]. The systems build Markov models from the input, then generate more music by traversing the graph according to the assigned probabilities.

MUSIC ANALYSIS

Music analysis is the process of breaking music down into smaller parts, such as chords and keys or melody streams, so as to better understand its content. Computational music analysis involves the automatic extraction of music structures such as those described in the section on music representations. The analysis of basic musical structures serves as a prelude to the understanding of more complex musical forms and gestures, some of which will be addressed in the section on expressive performance.

For every structure that exists, the automatic recognition of that structure—such as chord recognition, key finding, rhythm estimation, and meter induction—forms a

computational problem of interest. The representation of a music structure is often closely linked to algorithms for recognizing it. Consequently, as we introduced each music representation earlier, we also touched upon algorithms for automatic assignment of corresponding labels.

Pitch Structure Analysis: Keys, Chords, and Melodies

Key induction is the problem of determining the key of a music stream. We briefly mentioned some algorithms for key induction [12, 14, 16] in the section titled “Pitches, Chords, and Keys.” Other solution methods, including Refs 12 and 39 are reviewed in the section titled “Pitch Structure: Key and Chord Analysis.” The problems of segmenting a musical piece into regions of the same key and chord labeling will also be discussed in the section on pitch structure.

Moving from vertical (synchronous) to horizontal (sequential) pitch structures, the next paragraphs will review analysis methods for melodies.

An important problem in music information retrieval is the computing of music similarity, an example for which is melodic similarity. A typical solution to this problem is the computing of the edit distance, which can be solved using DP. In Ref. 40, Typke *et al.* propose the use of the Earth Mover’s Distance (EMD) to measure melodic similarity. The calculation of the EMD can be formulated as a linear programming problem, which can be solved using the simplex method. This EMD method extends to polyphonic music.

Another interesting problem in the analysis of note streams is that of voice separation, sometimes known as *stream segregation*. In the symbolic domain, voice separation takes music having multiple stream layers and separates them into their component voices in accordance to some subset of the principles of auditory streaming [41], such as voices tend not to cross, and voices tend to move by stepwise motion. Optimization-related algorithms for voice separation include Temperley’s rule-based approach using DP [42], Kilian and Hoos’ local search

[43], and Chew and Wu’s contig mapping algorithm [44].

Sequence alignment problems will be addressed in the section titled “Score Alignment.”

Time Structure Analysis: Meter, Tempo, and Rhythm

Given a rhythm, a sequence of onset times or of durations, algorithms for determining the periodicity and accent patterns of musical beats, that is meter induction, include Refs 27 and 28, mentioned in the section titled Rhythm and Meter. In Ref. 45, Klapuri *et al.* use an HMM for the tatum (a term coined by Bilmes [46], referring to the smallest unit of musical time needed to represent all events as multiples of that unit), beat, and measure. Rhythm estimation will be addressed in the section titled “Time Structure: Rhythm Estimation.”

In expressive performance, tempo (and loudness) variation can be used to indicate, and guide the perception of, groupings. Musical expression is also designed to communicate and elicit emotions from the listener. For example, a fast tempo engenders excitement; music perceived to be sad is typically slow. Thus, a problem of interest to the analysis of expressive performance, is that of beat and tempo (beat rate) tracking. A number of algorithms have been proposed for beat and tempo extraction. For example, Dixon [47] uses multiple hypothesis search while Goto [48] uses autocorrelation and cross-correlation to determine tempo. Reference 47 also provides a broad review of the literature.

Using a tempo-loudness space, Goebel *et al.* [49] employ self-organizing maps to analyze expressive gestures. Widmer and Tobudic apply machine learning to learn tempo and dynamic (loudness) strategies at different timescales [50]. Because tempo and/or dynamics vary to indicate phrase groupings [51], a performer’s phrasing strategies can be directly inferred from tempo and dynamic data. Chuan and Chew [52] use DP to automatically extract phrases from tempo time series, and Cheng and Chew [53] propose quantitative measures for evaluating phrasing strategies.

DYNAMIC PROGRAMMING

Given the sequential nature of so many aspects of music, it is not surprising that DP has many musical applications. In essentially all of these applications, the musical sequences are viewed as paths through a state graph, in which costs (or scores) are assigned to the arcs in the graph. In such a case, it is well known that DP can be used to find the least costly (or best scoring) path through the graph with computation proportional to the number of arcs in the graph.

It is important to understand the fundamental limitation on the types of objective functions that can be optimized with DP: as the cost of a path is expressed as the sum of arc costs traversed, the objective function is composed as a sum of terms depending on the values of adjacent pairs of sequence elements. That is, DP can optimize functions depending on *local* characteristics of the sequence, but not *global* ones. Fortunately, there are a great number of musical problems whose objectives fit naturally into this paradigm.

Fingering Problems

Piano fingering offers a simple yet interesting example of a musical DP problem. Suppose we are given a sequence of musical notes which are to be played by the right hand of the pianist. To each of the notes we will assign a finger, $\{1, \dots, 5\}$, with 1 representing the thumb, 2 the index finger, and so on. Thus, each possible sequence with elements in $\{1, \dots, 5\}$ constitutes a fingering of the passage.

The plausibility of each possible fingering can be captured by assigning penalties to each pair of adjacent fingering assignments. Simply by playing two notes at the same time with different finger pair combinations, we see that some are more natural than others, due to stretches involved, the unique role of the thumb, as well as other considerations. Thus, one can assign a cost to each fingering of each pair of neighboring notes, and optimize the sum of costs using DP.

This DP technique was first proposed by Parncutt [54], and later extended by Jacobs [55]. An *“in vivo”* view of piano fingering

involves two simultaneous hands, and must take into account vertical (chords) as well as horizontal (counterpoint) structures, the use of pedals, finger exchanges, and the desired kinds of articulation and phrasing goals [56]. While the most realistic version of the piano fingering problem remains open, the basic DP approach shows promise while allowing for many variations and improvements.

Other interesting examples of fingering problems are found in string instruments such as the guitar and violin. For these instruments, a note is played by choosing a string, a hand position or fret, and the finger to be used, thus giving a wide variety of possible choices for each note. As with the piano, a reasonable measure of fingering quality can be captured by considering only the fingering decisions associated with each pair of adjacent notes. In particular, the penalty function should discourage changes in hand position, large stretches, along with other undesirable traits, as described in Ref. 57 for the case of the guitar. Approaches having a similar spirit could also be applied to breathing on wind instruments as well as bowing for string instruments.

Pitch Structure: Key and Chord Analysis

Key analysis labels a piece of music with a corresponding key or collection of keys. The most simple-minded version of the key finding problem views a musical piece as a “bag of notes” without regard for the notes’ order, duration, simultaneity, and so on. Usually, the notes are further simplified to pitch classes (octave equivalence classes). For these music data, we wish to assign the most plausible key from the 24 possible major and minor keys. Krumhansl and Schmuckler [12] and others have characterized a key in terms of its *pitch profile*—the importance of the pitch classes in a given key, or the proportions with which they occur. Such a profile recognizes a preference for the more fundamental notes, such as the tonic and the scale tones, over other pitch classes. Thus, one could view key estimation as choosing the key whose profile best fits the empirical distribution of pitch classes. Various measures of agreement are used between the empirical distribution of pitch classes

and the pitch profile such as correlation [12], inner product [58], or Kullback-Leibler divergence [39].

However, it is common for music to visit several different keys as it evolves. A more realistic description of the key analysis problem partitions the music into a collection of key-labeled regions [59], say consecutive measures of music. The best explanation of the pitch data would be found by picking a different key for each measure of music. However, this conflicts with our strong preference for key interpretations that do not modulate (change) too much. A compromise approach creates an objective function that measures both, the agreement between the sequence of key labels and the actual musical data, penalizing key changes while rewarding agreement between key and data. Such an objective function fits perfectly into the DP formulation. Any change of key can be seen as a difference between an adjacent pair of key labels, while the data score depends only on the individual labels given to each measure. Thus the objective function is easy to optimize using DP and often results in thoroughly plausible explanations of key activity, as in Ref. 39.

The problem of chord labeling is similar to the problem of key finding. Chords are vertical pitch structures that exist on a more local timescale than key. Because chords change much faster than keys, the problem of segmentation for labeling by region becomes even more important for chords than it is for keys. The sequential nature of data for segmentation makes it particularly amenable to DP approaches. Pardo and Birmingham employ a graph search approach using DP [60], Temperley and Sleater also use DP, but in a rule-based approach [42], and Stoddard and Raphael [61] use an HMM with simultaneous key and chord recognition. In all of these cases, the essential approach is similar to that of key recognition. One difference is that chord knowledge offers more information about pitch profile than does key knowledge, so the chord data models discriminate better between hypotheses. However, this is offset by the greater number of chords that must be considered. Temperley [42] gives an

interesting discussion of the relative merits of these different approaches.

Time Structure: Rhythm Estimation

There are several other aspects of symbolic musical analysis naturally suited to the DP framework. One of the most interesting involves inferences about musical rhythm, made from time sequence (onset) data. Score-writing programs, the musical equivalent of word processors, offer a context in which rhythm recognition would clearly be useful. Here, one would like to play an electronic keyboard, or other MIDI-controlled instrument, into the computer, allowing the program to notate the intended rhythm of the events.

The rhythm estimation problem turns out to be much harder than one might suspect. While one can always choose a rhythm and tempo to approximate an onset sequence to arbitrary precision, the result will almost never be the interpretation the human would choose. Rather, one seeks interpretations that balance faithfulness to the data with *musical plausibility*. This latter trait stems not only from the human preference for simplicity, favoring interpretations of durations as simple fractions of the basic beat, but also from deeper issues of consistency when representing different instances of the same musical figure. A further complication results when tempo changes over time. In this case, one cannot make rhythm assignments without knowing the local tempo, and vice versa, leading to a “chicken-and-egg” problem.

Several authors, such as Cemgil and Kappen [62], Temperley [42], and Raphael [63], address this rhythm estimation problem in various ways. Temperley [42] presents the most straightforward DP-based approach, recovering a sequence of music measure lengths, thus allowing for changing tempo, while ascribing the best rhythmic interpretation of the collection of onset times that lie in the hypothesized measure. Thus, the state in the DP recursion is the time range of the most recently processed measure.

Raphael [63] presents a probabilistic formulation of the problem in terms of two unobserved processes, a Gaussian tempo process

and a discrete process of the musical onset positions of the notes. The computational approach then finds the most likely configuration of both continuous (tempo) and discrete (rhythm) variables using an extension of DP to hybrid state space. Cemgil [62] presents a more sophisticated model than either of the above, not lending itself to DP approaches. Rather, Markov-chain Monte Carlo (MCMC) methods are explored for the actual recognition of the sequence data.

Score Alignment

The previous discussion focused on symbolic music analysis. DP also has applications in audio analysis. The most well-known application of DP here is the *score alignment* problem [64], in which one seeks a correspondence between a symbolic musical score and an audio performance of that score. In essence, one simply wants to know where the score notes occur in the audio. This problem has many possible uses, including the generation of large music audio databases, facilitating quantitative study of musical performance, performance critique, editing of digital music audio, and the coordination of other media, such as animation, with prerecorded music.

Score alignment begins by dividing the audio into a sequence of “frames,” or local snapshots of sounds, as in the frames of a movie. We also can simplify our view of the score, regarding it as a sequence of single notes that will be played in the audio, disregarding any information about notated rhythm. We now seek to label each audio frame with a note from the score, with the constraint that notes must appear in the order prescribed by the score. For instance, if we have three notes in the score, a possible labeling would be $1^i 2^j 3^k$ where $i, j, k \in \mathbb{Z}_+$. One can build a state graph that obeys this constraint simply by allowing each of the possible notes to appear at each frame (level) of the graph, while only allowing a state to remain in the current note or move on to the next note.

A data cost measures the degree to which a particular frame of audio agrees with a particular note in the music score. Most current approaches rely on Fourier analysis. Since pitched musical notes have approximately

periodic signals, we expect to see spectral energy concentrated at frequencies that are integral multiples of the fundamental frequency. There are many possible ways of employing this knowledge to create a data cost for the labeled frame of audio. The cost for an entire path through the state graph can now be constructed as the sum of the costs for each individual frame, thus the arc cost depends only on the state the arc moves to.

Having constructed a graph with arc costs, it is straightforward to use DP to find the minimal cost path through the graph, thus labeling each frame of data with a position in the music score. In fact, this basic idea can be extended to polyphonic music alignment, by viewing the composition as a sequence of chords, rather than single notes. Variations on this basic approach have formed the basis for a number of score alignment efforts, including Refs 65–68.

The HMM has become a popular modeling tool for score alignment researchers in recent years [69–71]. While the approach previously described can be recast as an HMM in which the optimal DP state sequence becomes the maximum a posteriori state sequence, the HMM framework brings other powerful ideas. In this version the audio data becomes the observable part of a hidden Markov chain representing the changing music score position. For instance, a real-time alignment can be constructed using the filtered distribution on the hidden state, while the most likely onset time for a given note can be computed with the forward–backward algorithm.

REFERENCES

1. Chew E. Math & Music - the perfect match. *OR/MS Today* 2008;35(3):26–31.
2. Hewlett WB, Selfridge-Field E. The virtual score: representation, retrieval, restoration. Volume 12, Computing in musicology. Cambridge (MA): MIT Press; 2001.
3. Clampitt D. Pairwise well-formed scales: structural and transformational properties [PhD dissertation]. New York: SUNY Buffalo; 1997.
4. Douthett J, Krantz R. Maximally even sets and configurations: common threads in mathematics, physics, and music. *J Combin Optim* 2007;14(4):385–410.

5. Wild J. Pairwise well-formed scales and a bestiary of animals on the hexagonal lattice. *Math Comput Music Springer Commun Comput Inf Sci* 2009;38:273–285.
6. Lewin D. Generalized musical intervals and transformations. New Haven (CT): Yale University Press; 1987.
7. Cohn R. An introduction to neo-riemannian theory: a survey and historical perspective. *J Music Theor* 1998;42(2):167–180.
8. Callender C, Quinn I, Tymoczko D. Generalized voice-leading spaces. *Science* 2008;320(5874):346–348.
9. Tymoczko D. The geometry of musical chord. *Science* 2006;313(5783):72–74.
10. Shepard RN. Geometrical approximations to the structure of musical pitch. *Psychol Rev* 1982;89:305–333.
11. Krumhansl CL. The geometry of musical structure: a brief introduction and history. *Comput Entertain* 2005;3(4):1–14.
12. Krumhansl CL. Cognitive foundations of musical pitch. New York: Oxford University Press; 1990.
13. Toivainen P. Visualization of tonal content with self-organizing maps and self-similarity matrices. *ACM Comput Entertain* 2005;3(4):1–10.
14. Longuet-Higgins HC, Steedman MJ. On interpreting bach. *Mach Intell* 1971;6:221–241.
15. Dantzig GB. Converting a converging algorithm into a polynomially bounded algorithm. Report SOL 91-5. Stanford University: System Optimization Lab; 1991.
16. Chew E. Towards a mathematical model of tonality [PhD dissertation]. Cambridge (MA): MIT; 2000.
17. Chew E. Dantzig's indirect contribution to music research: how the von Neumann center of gravity algorithm influenced the center of effect generator key finding algorithm. *INFORMS Comput Soc Newsl Spring* 2008;21–25.
18. Chew E, François ARJ. Interactive multi-scale visualizations of tonal evolution in MuSART Opus 2. *ACM Comput Entertain* 2005;3(4):1–16.
19. Kassakian P, Wessel D. Optimal positioning in low-dimensional control spaces using convex optimization. *Proc Int Comput Music Conf* 2005;31:379–382.
20. Buteau C, Viperman J. Representations of motivic spaces of a score in OpenMusic. *J Math Music* 2008;2(2):61–79.
21. Lartillot O. Multi-dimensional motivic pattern extraction founded on adaptive redundancy filtering. *J New Music Res* 2005;34(4):375–393.
22. Conklin D, Anagnostopoulou C. Segmental pattern discovery in music. *INFORMS J Comput* 2006;18(3):285–293.
23. Lerdahl F, Jackendoff R. A generative theory of tonal music. Cambridge (MA): MIT Press; 1983.
24. Bod R. Memory-based models for music analysis: challenging the gestalt principle. *J New Music Res* 2002;31:27–36.
25. Toussaint G. Computational geometric aspects of rhythm, melody, and voice-leading. *Comput Geom Theor Appl* 2009;43(1):2–22.
26. Schell D. Optimality in musical melodies and harmonic progressions: the travelling musician. *Eur J Oper Res* 2002;140(3):354–372.
27. Volk A. Persistence and change: local and global components of metre induction using inner metric analysis. *J Math Music* 2008;2(2):99–115.
28. Smith LM, Honing H. Time-frequency representation of musical rhythm by continuous wavelets. *J Math Music* 2008;2(2):81–97.
29. Strauss JN. Stravinsky the serialist. In: Cross I, editor. *The cambridge companion to stravinsky*. Cambridge: Cambridge University Press; 2003. pp. 149–173.
30. Pachet F, Roy P. Musical harmonization with constraints: a survey. *Constraints* 2004;6(1):7–19.
31. Truchet C, Codognet P. Musical constraint satisfaction problems solved with adaptive search. *Soft Comput* 2004;8:633–640.
32. Xenakis I. Formalized music: thought and mathematics in music. Hillsdale (NY): Pendragon Press; 2001.
33. Xenakis I. Arts/Sciences: alloys. Hillsdale (NY): Pendragon Press; 1985.
34. Dabby DS. Musical variations from a chaotic mapping. *Chaos Interdiscip J Nonlin Sci* 1996;6(2):95–107.
35. Pachet F. The continuator: musical interaction with style. *J New Music Res* 2003;32(3):333–341.
36. Chuan C-H, Chew E. A hybrid system for automatic generation of style-specific accompaniment. In: *Proceedings of the 4th International Joint Workshop on Computational Creativity*. London, UK: 2007. pp. 57–64.
37. Assayag G, Dubnov S. Using factor oracles for machine improvisation. *Soft Comput* 2004;8(9):604–610.

38. Farbood M, Schoner B. Analysis and synthesis of palestrina-style counterpoint using markov chains. In: Proceedings of the International Computer Music Conference. La Habana, Cuba: 2001. p. 27.
39. Temperley D. Music and probability. Cambridge (MA): MIT Press; 2007.
40. Typke R, Giannopoulos P, Veltkamp RC, *et al.* Using transportation distances for measuring melodic similarity. In: Proceedings of the International Conference on Music Information Retrieval. Baltimore, Maryland: 2003.
41. Huron D. Tone and voice: a derivation of the rules of voice-leading from perceptual principles. *Music Percept* 2001;19(1):1–64.
42. Temperley D. The cognition of basic musical structures. Cambridge (MA): MIT Press; 2001.
43. Kilian J, Hoos H. Voice separation: a local optimisation approach. In: Proceedings of the International Conference on Music Information Retrieval, Paris, France: 2002. pp. 39–46.
44. Chew E, Wu X. Separating voices in polyphonic music: a contig mapping approach. Volume 3310, *Computer Music modeling and retrieval*. LNCS., Springer; 2005. pp. 1–20.
45. Klapuri AP, Eronen AJ, Astola JT. Analysis of the meter of acoustic musical signals. *IEEE Trans Audio Speech Lang Process* 2006;14(1):342–355.
46. Bilmes J. Timing is of the essence: perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm [Master's thesis]. Cambridge (MA): MIT; 1993.
47. Dixon S. Automatic extraction of tempo and beat from expressive performances. *J New Music Res* 2001;30(1):39–58.
48. Goto M. An audio-based real-time beat tracking system for music with or without drum-sounds. *J New Music Res* 2001;30(2):159–171.
49. Goebel W, Pampalk E, Widmer G. Exploring expressive performance trajectories: six famous pianists play chopin pieces. In: Proceedings of the International Conference on Music Perception and Cognition. Evanston, Illinois: 2004. pp. 505–509.
50. Widmer G, Tobudic A. Playing mozart by analogy: learning multi-level timing and dynamics strategies. *J New Music Res* 2003;32(3): 259–268.
51. Palmer C. Music performance. *Annu Rev Psychol* 1997;48:115–138.
52. Chuan C-H, Chew E. A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. Proceedings of the International Conference on Music Information Retrieval. Vienna, Austria: 2007. pp. 305–308.
53. Cheng E, Chew E. Quantitative analysis of phrasing strategies in expressive performance: computational methods and analysis of performances of unaccompanied bach for solo violin. *J New Music Res* 2008;37(4):325–338.
54. Parncutt R, Sloboda J, Clarke E, *et al.* An ergonomic model of keyboard fingering for melodic fragments. *Music Percept* 1997;14(2):341–382.
55. Jacobs JP. Refinements to the ergonomic model for keyboard fingering of Parncutt, Sloboda, Clarke, Raekallio, and Desain. *Music Percept* 2001;18(4):505–511.
56. Bamberger J. The musical significance of Beethoven's fingerings in the piano sonatas. *Music Forum* 1976;4:237–280.
57. Radicioni D, Lombardo V. Guitar fingering for music performance. In: Proceedings of the International Computer Music Conference, Barcelona, Spain: Volume 31. 2005. pp. 527–530.
58. Gabura J. Music style analysis by computer. In: Lincoln HB, editor. *The computer and music*. Ithaca (NY): Cornell University Press; 1970. pp. 223–276.
59. Chew E. The spiral array: an algorithm for determining key boundaries. Volume 2445, *Music and Artificial Intelligence - Second International Conference*. Heidelberg, Germany: Springer LNCS/LNAI. 2002. pp. 18–31.
60. Pardo B, Birmingham WP. Algorithms for chordal analysis. *Comput Music J* 2002;26(2):27–49.
61. Raphael C, Stoddard J. Functional harmonic analysis using probabilistic models. *Comput Music J* 2004;28(3):45–52.
62. Cemgil AT, Kappen HJ. Monte Carlo methods for tempo tracking and rhythm quantization. *J Artif Intell Res* 2003;18:45–81.
63. Raphael C. A hybrid graphical model for rhythmic parsing. *Artif Intell.* 2002;137(1–2):217–238.
64. Dannenberg R, Raphael C. Music score alignment and computer accompaniment. *Commun ACM* 2006;49(8):38–43.

65. Dannenberg R. An on-line algorithm for real-time accompaniment. Proc Int Comput Music Conf Miami, Florida: 1984;10:193–198.
66. Grubb L, Dannenberg D. A stochastic method of tracking a vocal performer. Proc Int Comput Music Conf Thessaloniki, Greece: 1997;23:301–308.
67. Arifi V, Clausen M, Kurth F, *et al.* Automatic synchronization of musical data: a mathematics approach. Comput Musicol 2004;13:9–33.
68. Turetsky R, Ellis D. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In: Proceedings of the International Conference on Music Information Retrieval, Baltimore, Maryland: 2003. pp. 135–141.
69. Raphael C. Automatic segmentation of acoustic musical signals using hidden Markov model. IEEE Trans on PAMI 1999;21(4): 360–370.
70. Schwarz D, Cont A, Schnell N. From Boulez to Ballads: training IRCAM's score follower. In: Proceedings of International Computer Music Conference. Barcelona, Spain: 2005.
71. Loscos A, Cano P, Bonada J. Score-performance matching using HMM. In: Proceedings of the International Computer Music Conference, Beijing, China: Volume 25. 1999. pp. 441–444.

NASH EQUILIBRIUM (PURE AND MIXED)

MICHAEL A. JONES
Mathematical Reviews, American
Mathematical Society, Ann Arbor,
Michigan

Von Neumann and Morgenstern [1] introduced two distinct, but related approaches to the theory of games. The first approach consisted of two-player, zero-sum games in which the players act simultaneously and independently of one another and the two players' interests are at odds with one another (so that player 2's utility for any outcome of the game is the opposite of player 1's utility for the same outcome). For these games, von Neumann and Morgenstern [1] used the Minimax Theorem [2] (also referred to as the *Fundamental Theorem of Game Theory*—see **Saddlepoints and von Neumann Minimax Theorem**) to determine players' optimal strategies, in which each player's strategy is optimal given the other player's strategy. The second approach consisted of cooperative games in which players can form coalitions and make binding agreements, as well as communicate with one another before and during the game. They analyzed n -player, nonzero sum games by introducing an $(n + 1)$ st player to turn the game into a zero-sum, cooperative game. (See **Cooperative Games with Transferable Utility** and **Cooperative Game Theory with Nontransferable Utility**). Von Neumann and Morgenstern's methods for cooperative games failed to extend optimal strategies to games in which players act simultaneously and independently of one another. These games are called *noncooperative strategic form* (or normal form) games. (Games in which players make their moves sequentially are considered in the sections titled "Noncooperative Extensive-Form (NCEF) Games of Perfect Information" and "Noncooperative Extensive-Form (NCEF) Games of Imperfect Information" in this encyclopedia.)

In Refs 3 and 4, Nash generalized Von Neumann's minimax strategies for strictly competitive, two-player zero-sum games to n -player, noncooperative strategic form games. Like Von Neumann's minimax strategies, Nash's solution concept, now known as the *Nash equilibrium*, allows for mixed strategies, in which players randomize over possible actions. Such a set of strategies (one for each player) is a Nash equilibrium if no player has an incentive to deviate from its strategy when all other players use their strategies in the set. The Nash equilibrium solution concept lends itself to multiple interpretations. If players were instructed to play their portions of a Nash equilibrium strategy, no player would have an incentive not to follow the instructions. Hence, a Nash equilibrium is stable, requiring no binding agreements, as those present in cooperative games. A Nash equilibrium can also be viewed as the stable point of an evolutionary process, as used in evolutionary biology. Holt and Roth [5] discussed the above two interpretations of the Nash equilibrium and its impact on game theory and the social sciences. A Nash equilibrium may also be viewed as the players' optimal responses to their beliefs about the other players' strategies. The variety of useful interpretations of Nash's work was in part the reason why he was awarded the Nobel Prize in Economics in 1994 (along with Harsanyi and Selten).

The next section begins with the definition of noncooperative strategic form games. A discussion of mixed strategies and Von Neumann–Morgenstern utility functions is followed by Nash's theorem of the existence of equilibria (in possibly mixed strategies). Examples in this section highlight canonical two-player games (in which each player has two possible actions) and demonstrate that more than one Nash equilibrium may exist for noncooperative strategic form games. The section concludes with a comparison of Nash equilibria for nonzero and zero-sum games, as well as a stylized view of Nash's work which applies a decomposition of a

two-player, noncooperative strategic form game into a zero-sum game and an equal-payoff game.

Because of the possible multiplicity of Nash equilibria, the number of equilibria in generic noncooperative strategic form games, the construction of games with a unique, pre-specified equilibrium, and the calculation of equilibria are considered in the next section. The article concludes with a description of the attempts to refine the Nash equilibrium concept for noncooperative strategic form games, including perfect and proper equilibria, as well as evolutionary approaches to refinement.

NONCOOPERATIVE STRATEGIC FORM GAMES AND NASH EQUILIBRIA

In a noncooperative strategic form game, players simultaneously and independently select actions or make moves in the game. Each player has a utility or payoff function that defines the player's preferences over the possible outcomes, where each player's goal is to maximize his payoffs, taking into consideration the actions selected by the other players. A formal definition of a noncooperative strategic form game appears below.

Definition 1. An n -player, noncooperative strategic-form game G consists of

- a finite set of players $N = \{1, 2, \dots, n\}$;
- for each player $i \in N$, a nonempty set of possible actions A_i where $A = \times_{i \in N} A_i$ is the set of action profiles;
- for each player $i \in N$, a payoff or utility function $u_i : A \rightarrow \mathbb{R}$ that characterizes player i 's preferences over the action profiles so that player i prefers outcome a to b if and only if $u_i(a) > u_i(b)$;

and is denoted by $G = \langle N, A, u \rangle$ where $u = (u_1, \dots, u_n) : \times_{i \in N} A \rightarrow \mathbb{R}^n$.

If each player's action set A_i is finite, then the game is finite. This article will focus on finite action sets, but will consider a two-player example with an infinite action set.

(For other infinite action spaces, see *Differential Games*). Game theorists have a collection of easy-to-understand games that are part of the vernacular and tradition of game theory because the types of strategic interaction described in these games readily occur in more complex games. These games typically consist of two players, each with a two-element action set. To introduce terminology, a general version of such a game is considered below.

Example 1 [(2 × 2) Bimatrix Games]. A (2×2) bimatrix form game consists of two players. Player 1 (the row player) has action set $A_1 = \{r_1, r_2\}$ where r_i is referred to as row i . Player 2 (the column player) has action set $A_2 = \{c_1, c_2\}$ where c_j is referred to as column j . Because the pair of payoffs for each row and column pair of actions r_i, c_j , $u((r_i, c_j)) = (u_1(r_i, c_j), u_2(r_i, c_j))$, may be arranged in a table or matrix (as in Fig. 1), these games are referred to as *bimatrix games*, or (2×2) bimatrix games in the event that $|A_1| = |A_2| = 2$; two-player, zero-sum games are referred to as *matrix games* (see the section titled “Two-Person Zero-Sum (TPZS) Games” in this encyclopedia).

Each player is aware of all players' possible actions and the associated payoffs, but is unaware of the actions selected by the other players and unaware of the processes used to select the actions. For a bimatrix game, each player is aware of the structure of the bimatrix and its payoffs, aware that the other players are aware, and so on; this is referred to as a *game with perfect information*.

Pure Nash Equilibria

Nash generalized the solution concept of Von Neumann and Morgenstern by considering

		Player 2	
		c_1	c_2
Player 1	r_1	$u_1(r_1, c_1), u_2(r_1, c_1)$	$u_1(r_1, c_2), u_2(r_1, c_2)$
	r_2	$u_1(r_2, c_1), u_2(r_2, c_1)$	$u_1(r_2, c_2), u_2(r_2, c_2)$

Figure 1. General (2×2) bimatrix for a two-player, noncooperative strategic form game.

not just how an individual player should play, but how each player's behavior is a necessary part of an equilibrium profile. The notation (a_i, a_{-i}^*) represents an action profile in which player i plays a_i and all other players play their portion of the action profile $a^* = (a_1^*, \dots, a_n^*)$. Consequently, it follows that $a^* = (a_i^*, a_{-i}^*)$. Below is a definition for an action profile to be a Nash equilibrium.

Definition 2. A Nash equilibrium of a noncooperative strategic form game $G = \langle N, A, u \rangle$ is an action profile $a^* \in A$ such that for every player $i \in N$,

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \text{ for all } a_i \in A_i.$$

In words, no player i has an action a_i that she prefers to a_i^* when all other players play their respective actions in the profile $a^* = (a_i^*, a_{-i}^*)$.

The action profile a^* has a self-fulfilling property: if all players play their portion of a^* , then no player has an incentive to play anything else. Although compelling from a theoretical viewpoint, the question remained whether or not such equilibria modeled reality. In an effort to determine whether or not the von Neumann and Morgenstern [1] and Nash [3,4] solution concepts were successful in predicting behavior, Flood [6] conducted a series of experiments—part psychology—that formed a foundation for experimental economics and game theory. The most well-known of these was conducted with Dresher [6, Experiment 5: A noncooperative pair]. Albert Tucker later provided a story for this game that came to be known as the *Prisoners' Dilemma*. A stylized story in the spirit of Tucker is given below.

Example 2 [The Prisoners' Dilemma].

Two suspects are found with a bag of money in front of a recently robbed bank. The suspects are taken into custody and separated

so that they do not have time to corroborate a story. Because the police do not have enough evidence to convict the suspects on bank robbery (only possessing the money), the police simultaneously offer the two “prisoners” the following plea bargains. If only one prisoner confesses that the other was the mastermind, then he will receive probation and serve no time while the other will serve a long sentence. If both confess, then they will split the sentence for being coconspirators. If neither confesses, then both prisoners will serve minimal time for the lesser charge of possessing the money. The action of confessing and not confessing are represented by D and C in the Prisoners' Dilemma bimatrix in Fig. 2, respectively, where D stands for “defect” against the other prisoner and C stands for “cooperate” with the other prisoner. The payoffs appear in Fig. 2; recall that the payoffs represent utility and not the amount of time of the sentence.

A short analysis determines that it is in the best interest of each prisoner to defect (play D), regardless of what the other prisoner does. That is, $u_1((D, \cdot)) > u_1((C, \cdot))$ and $u_2((\cdot, D)) > u_2((\cdot, C))$ so that D strictly dominates C . The unique Nash equilibrium is (D, D) . This game is confounding because the Nash equilibrium is inefficient: both players would do better if they were to cooperate and both play C , resulting in (C, C) for which $1 = u_i((C, C)) > u_i((D, D)) = 0$ for both i . Because no agreement can force the players to play (C, C) , both players prefer to play their dominant strategies D , resulting in the Nash equilibrium outcome.

The Prisoners' Dilemma has been used to model diverse strategic situations from nuclear war to trade negotiations. Poundstone [7] provided a popular account of the early history of game theory, the experiment of Flood and Dresher and the ubiquity

	C	D		F	B		H	T		c_1	c_2	
C	1,1	-1,2		F	2,1	0,0	H	1,-1	-1,1	r_1	0,0	1,1
D	2,-1	0,0		B	0,0	1,2	T	-1,1	1,-1	r_2	-1,-1	1,1
Prisoners' Dilemma			Battle of the Sexes			Matching Pennies			Nongeneric Game			

Figure 2. Bimatrix games used in the examples.

of the Prisoners' Dilemma as a model of strategic interaction. Nash's comments about the experiment can be found in Ref. 6 (footnote, p. 24). The players did not play the Nash equilibrium in the experiment. However, the experiment was not a single play of the game. The experiment involved the same two players playing the game repeatedly. Axelrod [8] described a well-known account of a tournament in which players submitted strategies to play the repeated Prisoners' Dilemma. The tournament's winning strategy (garnering the highest utility over games against all other strategies) was tit-for-tat, in which a player plays C initially and continues to do so, unless her opponent defects in round k in which case she responds by playing D in round $k + 1$.

While the Prisoners' Dilemma has a unique Nash equilibrium, the next canonical game demonstrates that there may be more than one Nash equilibrium and that it may be impossible to predetermine which equilibrium is more likely to occur. The following game is referred to as the *Battle of the Sexes* because of the supporting story of a husband and wife (the treatment below is an adaptation of Luce and Raiffa [9, p. 90]). More recent retellings are politically correct (e.g., the equivalent Bach or Stravinsky proposed in Ref. 10).

Example 3 [Battle of the Sexes]. A husband (player 1) and wife (player 2) are to meet up for a night out. Each can go either to the prize fight (F) or the ballet (B). As Luce and Raiffa [9, p. 90] note, "following the usual cultural stereotypes, the man prefers the fight while the woman the ballet; however, to both it is more important that they go out together than that each see the preferred entertainment." Because of a mix-up (and in a time before cell phones), the players are unable to communicate their intentions of which event to attend. The payoffs for the game are given in Fig. 2. There are two Nash equilibrium action profiles: (F, F) and (B, B) with asymmetric payoffs $(2, 1)$ and $(1, 2)$, respectively. It is impossible to presume that either outcome is more likely to occur.

Nash equilibria are stable points, suggesting that players will remain at the equilibrium if starting at the point. Besides the lack of a predictive ability, there is a more troubling occurrence for our current formulation of a Nash equilibrium. Not only do there exist games with multiple equilibria, but there exist games with no equilibrium action profiles (these are referred to as *pure Nash equilibria* below). The following example is a zero-sum game and was used by Von Neumann and Morgenstern [1, p. 94 (Fig. 12); p. 143] in describing the intuition behind mixed strategies (see **Saddlepoints and von Neumann Minimax Theorem**). That same intuition is necessary for nonzero sum, noncooperative strategic form games.

Example 4 [Matching Pennies]. For this bimatrix game (see Fig. 2), let $A_1 = A_2 = \{H, T\}$ where H represents "heads" and T represents "tails." Players 1 and 2 simultaneously each select an action by placing a penny flat on a table (the side facing upward represents the player's strategy). If the coins agree (both H or both T), then (player 1) gets both pennies—and, $u_1((H, H)) = u_1((T, T)) = 1$ while $u_2((H, H)) = u_2((T, T)) = -1$. If the coins do not agree (one is T and the other is H), then player 2 gets both pennies—and, $u_2((T, H)) = u_2((H, T)) = 1$ while $u_1((T, H)) = u_1((H, T)) = -1$. Notice that this game is a zero-sum game (see **Saddlepoints and von Neumann Minimax Theorem**).

Consider the set of four action profiles $A = \{(H, H), (H, T), (T, H), (T, T)\}$. The coins cannot match in equilibrium. This follows because player 2 has an incentive to switch her coin if they do match. However, the coins cannot be mismatched in equilibrium either, as Player 1 will have an incentive to switch his coin so that they agree. Hence, none of the four action profiles is a Nash equilibrium.

In the previous example, players were limited to selecting an action. Allowing players to randomize over their action set provides a richer source of strategies for which there is always a Nash equilibrium (in possibly mixed strategies). However, there are some questions about how mixed strategies should be

interpreted. We will discuss these questions and will return to Matching Pennies in the next section.

Mixed Strategies and Expected Utility

Because a single action may be exploitable (such as playing H in Matching Pennies), a mixed strategy allows a player to select randomly an action according to a probability distribution. It is easy to extend mathematically the definition of a strategy to include lotteries over actions. However, in practice, people do not make decisions according to the outcome of such a lottery (and would find it hard to implement the selection of an action without the help of an external device such as a computer or a coin to flip).

To resolve these types of concerns for how players randomize over their actions, Aumann and Brandenburger [11] viewed a Nash equilibrium as a collection of optimal responses to the players' beliefs over the other players' strategies. In this framework, each player (e.g., a player in Matching Pennies) realizes that his opponent is drawn from a population of possible game players. Each player in the population of game players plays a single action, but the player chooses her behavior to optimize over her belief about the probability distribution of her opponents and consequently, the probability distribution of the opponents' actions—which is just a mixed strategy profile σ_{-i} . Further interpretations of mixed strategies (and mixed strategy equilibria) are considered later in this section (and can be found in Binmore's collection of essays [12]).

Definition 3. A mixed strategy for player i is a probability distribution $\sigma_i = (\sigma_{i,1}, \sigma_{i,2}, \dots, \sigma_{i,m_i})$ over $A_i = \{a_{i,1}, a_{i,2}, \dots, a_{i,m_i}\}$ where $\sigma_{i,j} \geq 0$ is the probability that player i places on action $a_{i,j}$ so that $\sum_{j=1}^{m_i} \sigma_{i,j} = 1$. Let Σ_i be the set of all mixed strategies of player i . A mixed strategy profile $\sigma = (\sigma_1, \dots, \sigma_n)$ is a vector of players' mixed strategies.

The set of mixed strategies contains the set of pure strategies (equivalent to the set of action profiles); hence, a pure strategy for

player i is a mixed strategy in which $\sigma_i = (\sigma_{i,1}, \dots, \sigma_{i,m_i})$ has $\sigma_{i,j} = 1$ for some j and $\sigma_{i,k} = 0$ for all $k \neq j$. To simplify notation in (2×2) bimatrix games, let the mixed strategy profile $((p, 1-p), (q, 1-q))$ —in which player 1 places probabilities p and $1-p$ on r_1 and r_2 , respectively, and player 2 places probabilities q and $1-q$ on c_1 and c_2 , respectively—be represented by (p, q) . For a mixed strategy profile σ and an action profile $a \in A$, let $\sigma_j(a)$ be the probability that player j places on its portion a_j of the action profile a . It follows that $\sigma_j(a) = \sigma_{j,k}$ when player j 's component of a is $a_{j,k}$. Hence, for $a = (a_1, \dots, a_n) \in A$, $\prod_{j=1}^n \sigma_j(a)$ is the probability that action a is played under σ .

As part of the foundation for game theory, Von Neumann and Morgenstern [1] described how to account for the uncertainties of players' actions in mixed strategies by considering the players' utilities of lotteries (probability distributions) over outcomes. Payoffs indicating preferences of action profiles had to be extended to preferences of lotteries over action profiles. Von Neumann and Morgenstern [1] provided a set of axioms so that a player's preferences of lotteries could be described by the expected value of a (Von Neumann–Morgenstern) utility function. In what follows, all utility functions are assumed to be Von Neumann–Morgenstern utility functions. And, for consistency, assume that the payoffs in the previous examples are described by such utility functions.

Definition 4. The expected utility for player i under the mixed strategy profile σ is

$$u_i(\sigma) = \sum_{a \in A} \left(\prod_{j=1}^n \sigma_j(a) \right) u_i(a).$$

Using words to describe Definition 4, player i 's expected utility is the sum of i 's utility under every action profile weighted by the probability that the particular action profile occurs. If all players use a pure strategy under σ' , then, for one $a' \in A$, the product $\prod_{j=1}^n \sigma_j(a') = 1$ while for all other $a \in A$, $\prod_{j=1}^n \sigma_j(a) = 0$. It follows that

$$u_i(\sigma') = \sum_{a \in A} \left(\prod_{j=1}^n \sigma_j'(a) \right) u_i(a) = u_i(a').$$

Hence, the expected utility function for a pure strategy profile returns the utility function that describes preferences over actions in Definition 1.

Mixed Strategy Nash Equilibria

By allowing players to use mixed strategies, players' payoffs had to be defined in terms of Von Neumann–Morgenstern utilities. It follows that the definition of a Nash equilibrium must be redefined, as well.

Definition 5. A mixed strategy σ^* for a non-cooperative strategic form game $\langle N, A, u \rangle$ is a Nash equilibrium if, for all players i ,

$$\begin{aligned} u_i(\sigma_i^*, \sigma_{-i}^*) &\geq u_i(\sigma_i, \sigma_{-i}^*) \\ &\text{for all } \sigma_i \in \Sigma_i, \text{ or, equivalently,} \\ u_i(\sigma_i^*, \sigma_{-i}^*) &\geq u_i(s_i, \sigma_{-i}^*) \\ &\text{for all pure strategies } s_i. \end{aligned}$$

The first of the two equivalent inequalities mimics the requirements for an action profile to be a Nash equilibrium from Definition 2. However, it is the second inequality that is preferred and easier to apply. To see that these are equivalent, suppose that the first inequality holds. Because a pure strategy is by default a mixed strategy, then the second inequality is satisfied as well. Now suppose that the second inequality holds but that the first does not. Then at least one player i has a mixed strategy profile τ_i in which his expected payoff is greater than under σ_i^* . Because $u_i(\tau_i, \sigma_{-i}^*)$ is a convex combination (from the probability distribution τ_i over the pure strategies) of the payoffs for player i under pure strategies, then at least one of the pure strategies must have a greater payoff for player i than $u_i(\sigma_i^*, \sigma_{-i}^*)$, contradicting the second inequality of Definition 5.

For a Nash equilibrium σ^* , the support S_i of σ_i^* is the subset of A_i for which σ_i^* places positive probability. An important consequence of the definition of a mixed strategy

Nash equilibrium is that the expected utility of playing the pure strategy s_i that places probability 1 on $a_i \in S_i$ versus σ_{-i}^* is equal to the expected utility under σ_i^* . If it were less, then player i would do better to decrease the weight on a_i to 0. If it were more, then player i would do better to increase the weight on a_i to 1. In particular, this shows that positive probability is never placed on a dominated strategy. In later sections, the payoff under an action that receives positive probability in a mixed strategy Nash equilibrium is revisited in the discussion of an algebraic proof of the existence of a Nash equilibrium, as well as in describing how to calculate equilibria.

As a criticism of mixed strategies, realize that the outcome of a play of a game (often referred to as a *single-shot game*) is some action profile $a \in A$. Even when players play a mixed strategy Nash equilibrium σ^* , the outcome a of their mixed strategy may not be a Nash equilibrium action profile—and most likely will not be. The fact that $u_i(s_i, \sigma_{-i}^*) = u_i(\sigma_i^*, \sigma_{-i}^*)$ for a pure strategy s_i that places positive probability on the support of σ_i^* is now problematic. If the player receives the same expected payoff using any pure strategy on any action in the support of the mixed strategy Nash equilibrium, then there is impetus for a player to play one of these pure strategies instead. For that matter, any probability distribution of the support of the mixed strategy Nash equilibrium yields that same expected payoff! Besides Aumann and Brandenburger's description [11] of mixed strategies in terms of beliefs, others have considered reformulations of mixed strategies in an effort to justify their use. For example, in what is known as the *purification problem*, Harsanyi [13] interpreted mixed strategies as the result of limiting (pure) strategies of perturbed games based on players' knowledge of the payoffs, in which each player knows his own payoffs but not his opponents' payoffs.

Example 5 [Matching Pennies—Revisited]. Recall the payoffs for Matching Pennies in Fig. 2 and the discussion of the game in Example 4. Assume that player 2 plays H more frequently than T in the mixed strategy $(q, 1 - q)$. Player 1's expected payoff for playing H is $u_1((1, q)) = q + (-1)(1 - q) =$

$2q - 1$ while his expected payoff under T is $u_1((0, q)) = -1q + 1(1 - q) = 1 - 2q$. Because $q > 1/2$, player 1 can exploit player 2's mixed strategy by always playing H (or $p = 1$). The situation is reversed if player 2 uses $q < 1/2$ —that is, player 1 can exploit player 2's strategy by playing T with certainty. The only value of q for which player 1 cannot exploit player 2's strategy is $q = 1/2$. This can be seen by equating the payoffs that player 1 would receive under H and T or $2q - 1 = 1 - 2q$ and solving for q . The asymmetric payoffs suggest that $p = q = 1/2$ is a mixed strategy Nash equilibrium. This is verified because $\sigma^* = (\sigma_1^*, \sigma_2^*) = ((1/2, 1/2))$ (where the first $1/2$ is p and the second $1/2$ is q) satisfies

$$0 = u_i(\sigma^*) \geq u_i(s_i, \sigma_{-i}^*) \text{ for all pure strategies } s_i \in \Sigma_i \text{ for both players } i.$$

Intuition present in Matching Pennies was necessary to make an educated guess about a mixed strategy Nash equilibrium. That mixed strategy Nash equilibria exist in general requires a more systematic approach. We will apply the idea behind Nash's proof [3] of the existence of Nash equilibria to determine all the Nash equilibria for different (2×2) bimatrix games. This idea requires the definition of a correspondence or point-to-set mapping f where $f : \Sigma \rightrightarrows \Omega$ satisfies $f(\sigma) \subseteq \Omega$ for all $\sigma \in \Sigma$.

Definition 6. The best-response correspondence for player i in the game $G = \langle N, A, u \rangle$ is the point-to-set correspondence $BR_i : \times_{j \neq i} \Sigma_j \rightrightarrows \Sigma_i$ where

$$BR_i(\sigma_{-i}) = \{\sigma_i \mid \sigma_i \in \Sigma_i \text{ such that } u_i(\sigma_i, \sigma_{-i}) \geq u_i(\tau_i, \sigma_{-i}) \text{ for all } \tau_i \in \Sigma_i\}.$$

The best-response correspondence for player i returns the set of strategies that maximizes player i 's payoff when the other players play σ_{-i} . Nash's insight was to apply a fixed point theorem to the best-response correspondence

$$BR = (BR_1, \dots, BR_n) : \times_{i \in N} \Sigma_i \rightrightarrows \times_{i \in N} \Sigma_i.$$

A mixed strategy profile σ^* is a fixed point of the best response correspondence BR (so that $BR(\sigma^*) = \sigma^*$) when no player i can increase his expected payoff by deviating from σ_i^* when the other players play σ_{-i}^* . But this is exactly the definition of a mixed strategy Nash equilibrium (Definition 5).

Theorem 1 [Nash [3]]. *Every finite non-cooperative strategic form game has a mixed strategy equilibrium.*

Nash [3] demonstrated that a fixed point theorem of Kakutani was applicable to the best-response correspondences of finite non-cooperative strategic form games. Nash also provided a different, more involved proof of the existence of Nash equilibria for finite, n -player games using the Brouwer fixed point theorem in Ref. 4 that required defining an appropriate continuous function, a fixed point of which is also a Nash equilibrium. The following version of Kakutani's theorem is adapted from Ref. 14, which refers to upper hemi-continuity as the closed graph property because the set of (σ, τ) where $\tau \in f(\sigma)$ is closed. Upper hemi-continuity is also referred to as *upper semicontinuity* (including in Ref. 15).

Theorem 2 [Kakutani [15]]. *Let Σ be a compact, convex, nonempty subset of $\mathbb{R}^n$. Let f be a point-to-set correspondence $f : \Sigma \rightrightarrows \Sigma$ such that*

1. $f(\sigma)$ is nonempty for all σ ,
2. $f(\sigma)$ is convex for all σ ,
3. f is upper hemi-continuous (i.e., if σ_n converges to σ^* , τ_n converges to τ^* such that $\tau_n \in f(\sigma_n)$, then $\tau^* \in f(\sigma^*)$).

Then, there exists at least one $\sigma^ \in \Sigma$ such that $\sigma^* \in f(\sigma^*)$.*

The following sketch of the application of Kakutani's theorem follows Fudenberg and Tirole's proof [14]. Further, Myerson [16] gives a lengthy discussion of the application of Kakutani's fixed point theorem to prove Nash's theorem.

Proof. (Sketch for Theorem 1; Nash [3]) To apply Kakutani's theorem, define $\Sigma = \times_{i \in N} \Sigma_i$. Because each Σ_i is an $(m_i - 1)$ -simplex (that is, nonnegative vectors in $\mathbb{R}^{m_i}$ whose entries sum to 1), then Σ is compact, convex, and nonempty. The set $BR(\sigma)$ is nonempty because each player's payoff function is linear (and hence continuous) over the compact simplex of mixed strategies, and continuous functions attain their extrema on compact sets. The set $BR(\sigma)$ is convex if for any $\sigma', \sigma'' \in BR(\sigma)$, then $\alpha\sigma' + (1 - \alpha)\sigma''$ is also in $BR(\sigma)$. This follows due to the definition of best response and expected utility for each player i because $u_i(\sigma'_i, \sigma_{-i}) = u_i(\sigma''_i, \sigma_{-i}) = \alpha u_i(\sigma'_i, \sigma_{-i}) + (1 - \alpha)u_i(\sigma''_i, \sigma_{-i}) = u_i(\alpha\sigma' + (1 - \alpha)\sigma'', \sigma_{-i})$. The upper hemi-continuity of BR can be proved by contradiction and relies on the continuity of the u_i 's. Hence, BR satisfies the conditions of Kakutani's theorem—there exists a fixed point of BR , which is a Nash equilibrium.

To demonstrate the application of Kakutani's theorem with best-response correspondences from a (2×2) bimatrix game, we return to the Battle of Sexes.

Example 6 [Battle of the Sexes—Revisited]. Recall the bimatrix form of the Battle of the Sexes from Fig. 2. Let (p, q) represent a mixed strategy profile in which player 1 places probability p on F and player 2 places probability q on F . The best-response correspondence for player 1 is derived as follows. For player 2 playing q , player 1 prefers to play F when the expected value under F is greater

than under B , that is, $u_1((1, q)) > u_1((0, q))$ or $2q > 1 - q$ which reduces to $q > 1/3$. Hence, player 1's best response to $q > 1/3$ is to play F with probability 1 or $p = 1$. It follows that player 1 prefers B when $q < 1/3$ and is indifferent between any mixture of F and B when $q = 1/3$. This information gives us the best-response correspondence for player 1. A similar approach can be used to determine the best-response correspondence of player 2.

$$BR_1(q) = \begin{cases} \{0\} & \text{if } q < 1/3 \\ [0, 1] & \text{if } q = 1/3 \\ \{1\} & \text{if } q > 1/3 \end{cases}$$

$$BR_2(p) = \begin{cases} \{0\} & \text{if } p < 2/3 \\ [0, 1] & \text{if } p = 2/3 \\ \{1\} & \text{if } p > 2/3 \end{cases}$$

Since $p \in [0, 1] = \Sigma_1$ and $q \in [0, 1] = \Sigma_2$, then $\Sigma = \Sigma_1 \times \Sigma_2$ is the set of mixed strategies for the Battle of the Sexes. The graph of player 1's best-response correspondence $\{(x, q) \mid x \in BR_1(q) \text{ for all } q\}$ appears in the left of Fig. 3. Similarly, the graph of player 2's best-response correspondence $\{(p, y) \mid y \in BR_2(p) \text{ for all } p\}$ appears in the middle of Fig. 3. The intersection of these two sets is $\{(0, 0), (1, 1), (2/3, 1/3)\}$ and appears in the right of Fig. 3. For $\sigma^* = (\sigma_1^*, \sigma_2^*) \in \{(0, 0), (1, 1), (2/3, 1/3)\}$, it follows that $\sigma_1^* \in BR_1(\sigma_2^*)$ and $\sigma_2^* \in BR_2(\sigma_1^*)$. Hence, each σ^* is a Nash equilibrium.

Two of the three Nash equilibria for the Battle of the Sexes were determined previously in Example 3. Naturally, the mixed

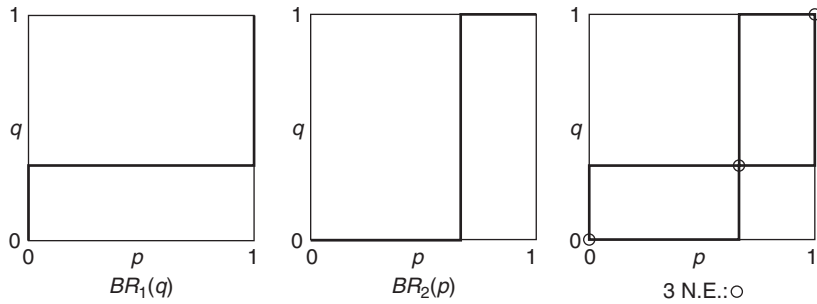


Figure 3. Best-response correspondences for player 1 (left) and player 2 (middle). Graphed together (right), their intersections yield three Nash equilibria in the Battle of the Sexes (Example 6).

strategy Nash equilibrium was not considered before. Example 6 demonstrates that a game may have both pure and mixed strategy Nash equilibria. Matching Pennies demonstrates that there may only be a single Nash equilibrium in mixed strategies, while the Prisoners' Dilemma does the same for pure strategies. The following example shows that a game need not have only a finite number of equilibria.

Example 7 [Infinite Number of Equilibria]. Consider the nongeneric game in Fig. 2. Again, let mixed strategy profiles be denoted by (p, q) . The best-response correspondences are

$$BR_1(q) = \begin{cases} [0, 1] & \text{if } q = 0 \\ \{1\} & \text{if } q > 0. \end{cases} \quad BR_2(p) = \{0\},$$

which are represented as graphs in Fig. 4. It follows that $\{(p, 0) \mid p \in [0, 1]\}$ are all Nash equilibria. Because player 1 weakly prefers r_1 to r_2 regardless of the value of q , the Nash equilibrium $(1, 0)$ seems more likely to occur in practice.

Besides showing that there may be an infinite number of equilibria, the previous example suggests that perhaps other secondary characteristics may be useful to narrow further the set of Nash equilibria. Referring to the game in Example 7 as “nongeneric” is not an accident. When the payoffs for games are drawn from continuous distributions over intervals of real numbers, finite noncooperative strategic form games

have a finite number of equilibria except for an exceptional set of payoffs whose closure has Lebesgue measure zero. A property that is satisfied for all games up to a set of measure zero is referred to as *generic*. Most proofs that generic games have a finite number of equilibria were provided en route to establishing an algorithm to construct or to approximate equilibria. However, Govindan and Wilson [17] provided a direct proof. The number of equilibria is considered below in the section titled “Counting, Constructing, and Calculating Nash Equilibria.” Refining the set of Nash equilibria by considering additional characteristics is discussed in the section titled “Refinements of Nash Equilibria for Strategic Form Games.”

The following game demonstrates that additional requirements are necessary to ensure that there exists a Nash equilibrium when the action sets of the players are not finite.

Example 8 [Race to the Bottom]. For a two-player noncooperative strategic form game, let $A_1 = A_2 = \mathbb{Q}^+$ (the positive rational numbers). For positive rational numbers q_1 and q_2 , the payoffs in the zero-sum game are described by the utility functions

$$\begin{aligned} u_1((q_1, q_2)) &= -u_2((q_1, q_2)) \\ &= \begin{cases} 1 & \text{if } q_1 < q_2 \\ 0 & \text{if } q_1 = q_2 \\ -1 & \text{if } q_1 > q_2 \end{cases} . \end{aligned}$$

Suppose that player 2's mixed strategy σ_2 places positive probability on a finite set

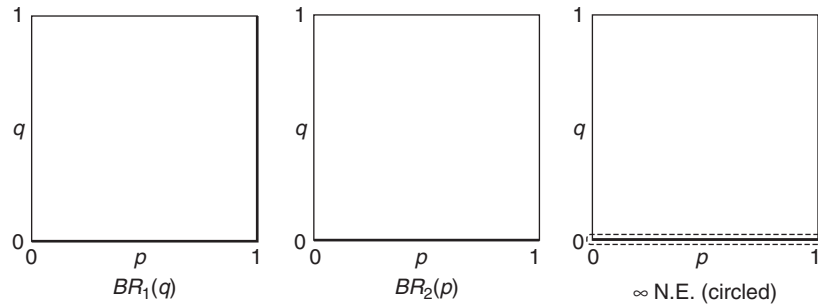


Figure 4. Best-response correspondences for player 1 (left) and player 2 (middle). Graphed together (right), their intersections (circled) yield infinitely many Nash equilibria for the nongeneric game (Example 7).

$Q_2 \subset \mathbb{Q}^+$. Then, player 1's best-response to σ_2 , $BR_1(\sigma_2)$, is the set of mixed strategies over $(0, q_*) \cap \mathbb{Q}^+$ where q_* is the minimum rational number in Q_2 . However, if player 2's mixed strategy has infinite support with no minimum value, then player 1 has no best response. For example, if σ'_2 places probability $1/2^n$ on $1/2^n$ for positive integers n , then $BR_1(\sigma'_2) = \emptyset$. Each player prefers to use rational numbers closer and closer to zero. Because there is no smallest positive rational number, then there is no Nash equilibrium.

Noncooperative strategic form games with infinite action spaces are covered in the articles titled in **Differential Games** in this encyclopedia. Below we examine some properties of Nash equilibria for zero-sum games and consider a stylized view of Nash equilibria for two-player, noncooperative strategic form games.

Nash Equilibria for Zero-Sum Games and a Decomposition of Bimatrix Games

As remarked by Nash in Refs 3, 4, his theorem reduces to the Minimax Theorem [2] when applied to zero-sum or matrix games. Not only do minimax/maximin strategies for zero-sum games satisfy the aforementioned properties of Nash equilibria, but additional properties as well. These properties are reviewed and compared to Nash equilibria for nonzero sum games below.

For a two-player, zero-sum game, the entries $(u, -u)$ of the bimatrix sum to zero. If $u > 0$, then Player 1 receives u and Player 2 loses u (which can be thought of as Player 2 paying Player 1); this relationship of who pays whom changes if $u < 0$. A strategy σ_1^* for player 1 is a maximin strategy if it maximizes the minimum payoff of player 1 (over player 2's strategies) while a strategy σ_2^* for player 2 is a minimax strategy if it minimizes the maximum payoff that player 2 pays to player 1 (over player 1's strategies). In the Minimax Theorem, Von Neumann [2] showed that minimax and maximin strategies always exist for zero-sum games with finite actions; see the article titled **Saddlepoints and von Neumann Minimax Theorem** in this encyclopedia.

Both σ_1^* and σ_2^* may be viewed as security strategies (see Ref. 18 for misconceptions about the optimality of minimax strategies) for which a value v is associated so that player 1 can be guaranteed a payoff of at least v and in which player 2 can be guaranteed to lose no more than v . The payoff v to player 1 under σ^* is referred to as the *value of the game*. Of course, $\sigma^* = (\sigma_1^*, \sigma_2^*)$ is a Nash equilibrium. As with Nash equilibria for nonzero sum games, there may be more than one Nash equilibrium for a zero-sum game. However, each Nash equilibrium strategy profile yields the same value so that a matrix game's value is well-defined. This is in sharp contrast to nonzero sum games in which the payoffs to the players depends on which equilibrium is played (as in the Battle of the Sexes, Example 6). For zero-sum games, there is no coordination problem of determining which equilibrium should be played, as all equilibria result in the same value. Further, each pair of maximin and minimax strategies combine to form a Nash equilibrium. For these reasons, Nash equilibrium strategies are referred to as *exchangeable*. See the section titled "Two-Person Zero-Sum (TPZS) Games" in this encyclopedia for more information.

For finite, bimatrix games, the existence of Nash equilibria may be viewed as an extension of the equilibrium concept from a subspace of zero-sum games to a larger space. A bimatrix game can be decomposed into the sum of an equal-payoff symmetric game (in which the payoff for each player is the same) and a zero-sum game. These form the two extremes in which players payoffs are completely aligned and completely opposed, respectively. The discussion will be limited to (2×2) bimatrix games but extends to $(m \times n)$ bimatrix games. The decomposition is presented below. This decomposition was used by Nash [19] in developing his two-person bargaining solution using players' threat strategies (a description of which appeared in Shubik [20])—see the article titled **Cooperative Games with Transferable Utility** in this encyclopedia.

Example 9 [(2 × 2) Bimatrix Games—Revisited]. Let (a, b, c, d, e, f, g, h) represent the utilities of the bimatrix game in Fig. 5.

$$\begin{array}{|c|c|} \hline a,b & c,d \\ \hline e,f & g,h \\ \hline \end{array} = \begin{array}{|c|c|} \hline \frac{a+b}{2}, \frac{a+b}{2} & \frac{c+d}{2}, \frac{c+d}{2} \\ \hline \frac{e+f}{2}, \frac{e+f}{2} & \frac{g+h}{2}, \frac{g+h}{2} \\ \hline \end{array} + \begin{array}{|c|c|} \hline \frac{a-b}{2}, \frac{b-a}{2} & \frac{c-d}{2}, \frac{d-c}{2} \\ \hline \frac{e-f}{2}, \frac{f-e}{2} & \frac{g-h}{2}, \frac{h-g}{2} \\ \hline \end{array}$$

(2 × 2) Bimatrix game
Equal-payoff game
Zero-sum game

Figure 5. Decomposition of a general (2 × 2) bimatrix into zero-sum and equal-payoff symmetric games.

The space of all (2 × 2) bimatrix games is of dimension eight—one dimension for each of the eight parameters in (a, b, c, d, e, f, g, h) . Figure 5 demonstrates a decomposition of a general bimatrix form game into a component in the equal-payoffs subspace and a component in the zero-sum game subspace. These orthogonal subspaces each have dimension four.

An equal-payoff game is reminiscent of a decision theory problem, as both players have the same incentives. (And, in fact, it is easy to predict an outcome when there is a single entry in the bimatrix with the largest value. However, in such a case, there may be more than one Nash equilibrium, even if there is a single, largest entry.) Quint and Shubik [21] referred to these equal-payoff games as coordination games and gave bounds on the number of Nash equilibria for such games. The nonzero sum, noncooperative strategic form game also has a zero-sum game component. A conceptual view of Nash's theorem for bimatrix games is that it proves the existence of equilibria for the extremes cases of no conflict and total conflict games, and everywhere in between.

CLASSIFICATION OF EQUILIBRIA FOR NON-COOPERATIVE STRATEGIC FORM GAMES

After Nash [3,4] provided the proof of the existence of equilibria for nonzero sum, finite noncooperative strategic form games, the question turned to characterizing Nash equilibria. Ideally there would be a single Nash equilibrium. However, as demonstrated in the previous examples, there may be more than one equilibrium, even infinitely many. Further, pure and mixed strategies may exist

concurrently, or separately. To determine what may happen, researchers focused on counting, constructing, and calculating equilibria. These results are considered below.

For the instances in which there are more than one Nash equilibrium, the question of whether any of the equilibria may be eliminated (as infinitely many of the equilibria in Example 7 were viewed as specious) by considering additional qualifications is known as *equilibrium refinement*. The section on equilibrium refinement and evolutionary game theory does not attempt to be comprehensive, but to provide a sample of the types of results that are possible.

Counting, Constructing, and Calculating Nash Equilibria

For a finite, noncooperative strategic form game, let $m_i = |A_i|$ for all $i \in N$. From the discussion following the definition of a mixed strategy Nash equilibrium (Definition 5), it follows that $u_i(\sigma_i^*, \sigma_{-i}^*) = u_i(s_i, \sigma_{-i}^*)$ where s_i places probability 1 on an action a_i in the support S_i of σ_i^* . Reverse engineering this fact, one can determine Nash equilibria by solving systems of equations that are based on the support of a possible mixed strategy Nash equilibrium. For each player, there are $2^{m_i} - 1$ possible sets of actions that may be the support of a Nash equilibrium. Considering the possible supports for all players, there are $\prod_{i=1}^n (2^{m_i} - 1)$ systems of equations to solve. If a solution to a system of equations yields a probability distribution over each player's support, then the result is a Nash equilibrium. If each player i 's support is A_i for a mixed strategy profile, then the profile is totally mixed. For two players, the resulting linear equations have at most one solution, so there is only one totally mixed Nash equilibrium strategy profile (unless the game is

nongeneric). For more than two players, the equations are nonlinear and can have a multitude of solutions. McKelvey and McLennan [22] provided a 10-player game with more than a million totally mixed Nash equilibria where each player has two strategies. Datta [23] described the complexity of the nongeneric case using algebraic geometry.

Nash's two proofs of the existence of equilibria for finite, n -player games [3,4] were based on fixed point theorems. Just as the Minimax Theorem is an algebraic theorem—Weyl [24] and Gale *et al.* [25] provided algebraic proofs of the existence of equilibria for finite, two-player zero-sum games—so, too, is the existence of Nash equilibria: solving for the weights of the probability distributions in the aforementioned systems of equations. Lemke and Howson [26] were the first to prove the algebraic existence of equilibria for two-player, finite strategic form games by a constructive method related to linear programming; they proved that generic bimatrix games have an odd number of equilibria (and so there cannot be zero equilibria). Rosenmüller [27] and Wilson [28] extended the Lemke–Howson approach to prove that there are an odd number of Nash equilibria in generic n -player, finite strategic form games. Harsanyi [29] provided an alternate proof, while Govindan and Wilson's proof [17] was not the consequence of an algorithm to compute equilibria.

If a game only has a single Nash equilibrium, then there is no need to evaluate or to compare the equilibrium. Hence, some research focused on how likely it is for there to be a unique Nash equilibrium. More generally, research concentrated on determining the distribution of the number of equilibria, constructing games with a specific equilibrium, and bounding the number of equilibria.

Goldberg *et al.* [30] focused on the likelihood that a two-player, noncooperative strategic form game has a single, pure strategy Nash equilibrium and showed that this probability converges to $1 - 1/e$ as the numbers of actions of the players increase without bound. This result was extended by Dresher [31] for games with more than two players, again showing that the probability of there being a single, pure strategy Nash

equilibrium converges to $1 - 1/e$ as two or more of the numbers of actions of the players increase without bound. For n -player noncooperative strategic form games with a countable number of actions, Powers [32] proved that the distribution of the number of pure strategy Nash equilibria approaches the Poisson distribution with mean 1 as the numbers of strategies of two or more players go to infinity; Stanford [33] provided a simpler proof.

Additional work has focused on the uniqueness of a mixed strategy Nash equilibrium. For bimatrix games, Kreps [34], Heuer [35], and Quintas [36] constructed games that would have a specified mixed strategy profile as its unique Nash equilibrium. A mixed strategy profile (σ_1^*, σ_2^*) can be the unique Nash equilibrium if the support of σ_1^* and σ_2^* is the same size. Kreps [37] determined conditions for the uniqueness of totally mixed strategies in n -player games, as well as constructed payoffs for a game with a given unique, totally mixed Nash equilibrium.

Quint and Shubik [38] showed that for any odd integer m between 1 and $2n - 1$, there is an $n \times n$ bimatrix game with exactly m Nash equilibria—their proof relies on an existence result on weighted majority cooperative games—and conjectured that a generic $m \times m$ bimatrix game could have at most $2^m - 1$ Nash equilibria. Von Stengel [39] provided a counterexample for $m = 6$ with 75 Nash equilibria and showed how to extend the construction for $m > 6$. Quint and Shubik [21] proved that their conjectured bound of a maximum of $2^m - 1$ Nash equilibria does hold for a class of coordination games (the equal-payoff games of the section titled “Nash Equilibria for Zero-Sum Games and a Decomposition of Bimatrix Games”).

Von Stengel [39] also computed new lower bounds for the maximum number of equilibria in an $(m \times m)$ bimatrix game. McLennan and Berg [40] determined the expected number, asymptotically in the number of actions, of Nash equilibria for bimatrix games under the assumption that the players' payoffs are identically and independently distributed normal random variables with mean zero. Under different probabilistic assumptions, McLennan [41] addressed the question of the expected number of Nash equilibria

for n -player games. For more on equilibrium computation, see von Stengel's survey [42].

Brown [43] proposed a simple iterative method for approximating the solutions of zero-sum, two-player games that Robinson [44] proved converges to a Nash equilibrium of the game. This type of fictitious play was a precursor to evolutionary biological and evolutionary dynamics approaches to game theory. Evolutionary stability is viewed as an equilibrium refinement below. The Nash equilibria for modest-sized noncooperative strategic form games can be calculated through Gambit [45], a freely available game theory solver. The application of Gambit, as well as code for the computer algebra systems Mathematica and Maple, to compute Nash equilibria is demonstrated in the game theory book by Barron [46].

Refinements of Nash Equilibria for Strategic Form Games

From the examples of (2×2) bimatrix games, there may be more than one Nash equilibrium for a finite, noncooperative strategic form game. The idea behind equilibrium refinement is to consider additional criteria to disqualify some of the Nash equilibria. An important criterion is whether or not the refinements always exist.

Selten [47] introduced perfect (also called *trembling-hand perfect*) equilibria, which he showed always exist for finite, noncooperative strategic form games, as a refinement of the set of Nash equilibria. A perfect equilibrium may be viewed as the limit of a totally mixed strategy profile for which a minimum positive probability must be placed on each action. This minimum weight can be viewed as a tremble or mistake by a player. Hence, a Nash equilibrium is a perfect equilibrium if it is stable or robust in the presence of small trembles or perturbations. Three formulations of perfect equilibria are proved to be equivalent in the text by Fudenberg and Tirole [14, p. 351]. One of these equivalent formulations is to define perfect equilibria as the limit of Nash equilibria in some sequence of constrained games. All three formulations define perfect equilibria as the limit of *some* sequence of totally mixed strategies, as opposed to *all*

such sequences. Perfect equilibria may also be defined for extensive form games and are related to sequential equilibria, a refinement of Nash equilibria for extensive-form games. For refinements of extensive form games, see the sections titled "Non-Cooperative Extensive-Form (NCEF) Games of Perfect Information" and "Non-Cooperative Extensive-Form (NCEF) Games of Imperfect Information" in this encyclopedia.

Myerson [48] introduced proper equilibria as a refinement of perfect equilibria. In this case, trembles are adjusted according to how severe the mistake would be (according to the payoff function). A minor mistake is more likely to be made than a costly mistake; more specifically, a player's n th-best actions are played with at most ε probability of her $(n - 1)$ st-best actions. In this sense, trembles are rational. Myerson showed that the set of proper equilibria is a nonempty subset of perfect equilibria, which makes sense as the types of trembles are more restricted than under perfect equilibria and provide a sequence for limit definitions of perfect equilibria. Myerson [48] showed that proper equilibria always exist for finite noncooperative strategic form games. Halpern [49] characterized the proper equilibrium refinement for noncooperative strategic form games in terms of nonstandard probability. For more information on refinements, see other articles in this encyclopedia.

Perfect and proper equilibria may be viewed as robust against certain types of mutations of the strategies. Evolutionary game theory studies evolutionary processes that are governed by a mutation mechanism as well as a selection mechanism that rewards successful strategies in a next generation (see Weibull [50, p. 69]). The mutation mechanism is embodied by notions of evolutionary stability, while the selection mechanism is governed by dynamic processes that update the population levels. For these games, a mixed strategy represents the percentages of the population that play the associated pure strategies. Members of the populations are matched at random to play the game, often restricted to two species or populations. In some cases, the players within a population are playing against

one another and have the same ordered action sets so that the resulting $(n \times n)$ bimatrix game has symmetric payoffs where $u_1(r_i, c_j) = u_2(r_j, c_i)$ for all $i, j \in \{1, \dots, n\}$. However, evolutionary game theory is not limited to symmetric, bimatrix games.

Biologist Maynard Smith [51] introduced evolutionary stable strategies (ESSs) based on the ability of a population (of game players using the same pure strategy) to survive against a mutation. Because the biological underpinnings require specific relationships between payoffs, an ESS is not well-defined for all noncooperative strategic form games. An ESS (if one exists) is by definition a Nash equilibrium, but if there is another best-reply (like the pure strategies in the support of a mixed strategy Nash equilibrium) strategy, then the ESS must receive a higher payoff against the other strategy than the other strategy does against itself. This prevents the new strategy from gaining a foothold in the population. To give a sample of the relationship between an ESS and other topics considered in this article, if the set of ESS for a finite, noncooperative strategic form game is finite and there is a totally mixed ESS, then the set consists only of this totally mixed strategy. For more information on ESSs, see the text by Hofbauer and Sigmund [52].

Another possibly empty refinement of Nash equilibria is an equilibrium evolutionary stable (EES) set. Weibull [50, p. 168] described an EES set as “a minimal closed set of Nash equilibria such that no small-scale invasion of equilibrium entrants can lead the population out of the set.” Equilibrium entrants introduce a type of rational mutation. For bimatrix games, Swinkels [53] showed that a nonempty EES set contains a proper equilibrium.

Selection processes reward successful pure strategies with a greater payoff that allows them to reproduce (or replicate) more successfully in the population by assuming that the strategy is passed on to offspring, thereby increasing the percentage of the population playing the strategy. The most common selection or dynamic process is called *replicator dynamics*, introduced by Taylor and Jonker [54], in which reproduction takes place continuously over time and

the dynamics are represented by a system of differential equations. For a population playing an evolutionary game against other members of the population, the mixed strategy profile $p = (p_1, \dots, p_n)$ describes the distribution of a population so that p_i of the population plays the pure strategy s_i that places probability 1 on action a_i . Each p_i is viewed as a function of time t so that the population evolves according to

$$\frac{dp_i(t)}{dt} = [u_i(s_i, p) - u_i(p, p)]p_i(t),$$

where the difference $u_i(s_i, p) - u_i(p, p)$ indicates how well the pure strategy s_i does in payoff compared to the population p on average. When the difference is positive, then the population playing s_i grows because its derivative is positive. When the difference is negative, then the population playing s_i decreases because its derivative is negative. It is not difficult to see that a Nash equilibrium p^* is a stationary point of the system of differential equations. This follows because if s_i is part of the support of p^* then $u_i(s_i, p^*) - u_i(p^*, p^*) = 0$ and if s_i is not part of the support then $p_i^* = 0$; in both cases, $dp_i(t)/dt = 0$ for all i .

Notice that stationary points are not necessarily Nash equilibria; any pure strategy is a stationary point for the replicator dynamics without mutation described above. Other selection processes exist and incorporate mutation into the differential (or difference) equations that govern the dynamics. The pertinent question is which equilibria of the game also have appealing dynamics properties. In this way, the evolutionary process refines the Nash equilibria. For example, if a totally mixed strategy of a finite, symmetric bimatrix game is a Lyapunov stable stationary state of the associated replicator dynamics, then it is also a Nash equilibrium. For additional information on evolutionary game theory, see Weibull [50].

This article has had two principal purposes. The first was to develop the idea of pure and mixed Nash equilibria for noncooperative strategic form games, including the analysis of canonical games (the Prisoners' Dilemma, Matching Pennies, and the Battle of the Sexes) to demonstrate the range of

possible equilibrium behavior. The second purpose was to summarize subsequent research regarding the number and computability of Nash equilibrium and attempts to refine Nash's equilibrium concept. Two comprehensive surveys on strategic equilibrium concepts are by Van Damme [55] and Hillas and Kohlberg [56]. These include discussions of the Nash equilibrium concept and mixed strategies. For a survey on bimatrix games, see Raghavan [57].

REFERENCES

1. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton University Press; 1944.
2. Von Neumann J. Zur theorie der gesellschaftsspiele. *Math Ann* 1928;100:295–320.
3. Nash JF Jr. Equilibrium point in n -person games. *Proc Nat Acad Sci USA* 1950;36(1):48–49.
4. Nash JF Jr. Non-cooperative games. *Ann Math* 1951;54(2):286–295.
5. Holt CA, Roth AE. The Nash equilibrium: a perspective. *Proc Nat Acad Sci USA* 2004;101(12):3999–4002.
6. Flood MM. Some experimental games. *RAND RM*-789-1. 1952;44
7. Poundstone W. The Prisoner's Dilemma. New York: Doubleday; 1992.
8. Axelrod RM. The evolution of cooperation. New York: Basic Books; 1984.
9. Luce RD, Raiffa H. Games and decisions. New York: John Wiley & Sons, Inc.; 1957.
10. Osborne MJ, Rubinstein A. A course in game theory. Cambridge (MA): M.I.T. Press; 1994.
11. Aumann R, Brandenburger A. Epistemic conditions for Nash equilibrium. *Econometrica* 1995;63(5):1161–1180. Stable URL: <http://www.jstor.org/stable/2171725>.
12. Binmore K. Essays on the foundations of game theory. Cambridge (MA): Basil Blackwell Inc.; 1990.
13. Harsanyi JC. Games with randomly disturbed payoffs: a new rationale for mixed strategy equilibrium points. *Int J Game Theor* 1973;2(1):1–23. DOI: 10.1007/BF01737554.
14. Fudenberg D, Tirole J. Game theory. Cambridge (MA): M.I.T. Press; 1991.
15. Kakutani S. A generalization of Brouwer's fixed point theorem. *Duke Math J* 1941;8(3):457–459. DOI: 10.1215/S0012-7094-41-00838-4.
16. Myerson RB. Game theory: analysis of conflict. Cambridge (MA): Harvard University Press; 1991.
17. Govindan S, Wilson RB. Direct proofs of generic finiteness of Nash equilibrium outcomes. *Econometrica* 2001;69(3):765–769. DOI: 10.1111/1468-0262.00213.
18. Spruill ML, Lyon HL. Misconceptions on optimality in two-person, zero-sum games. *Decis Sci* 1972;3(3):124–127. DOI: 10.1111/j.1540-5915.1972.tb00552.x.
19. Nash JF Jr. Two-person cooperative games. *Econometrica* 1953;21(1):128–140.
20. Shubik M. Game theory in the social sciences. Cambridge (MA): M.I.T. Press; 1982.
21. Quint T, Shubik M. A bound on the number of Nash equilibria in a coordination game. *Econ Lett* 2002;77(3):323–327. DOI: 10.1016/S0165-1765(02)00143-X.
22. McKelvey RD, McLennan A. The maximal number of regular totally mixed Nash equilibria. *J Econ Theor* 1997;72(2):411–425. DOI: 10.1006/jeth.1996.2214.
23. Datta RS. Universality of Nash equilibria. *Math Oper Res* 2003;28(3):424–432.
24. Weyl H. Elementary proof of a minimax theorem due to von Neumann. In: Kuhn HW, Tucker AW, editors. Volume 1, Contributions to the theory of games, Annals of mathematics study 24. Princeton (NJ): Princeton University Press; 1950. pp. 81–88.
25. Gale D, Kuhn HW, Tucker AW. On symmetric games. In: Kuhn HW, Tucker AW, editors. Volume 1, Contributions to the theory of games, Annals of mathematics study 24. Princeton (NJ): Princeton University Press; 1950. pp. 19–25.
26. Lemke CE, Howson JT Jr. Equilibrium points of bimatrix games. *J Soc Ind Appl Math* 1964;12(2):413–423. <http://dx.doi.org/10.1137/0112033>.
27. Rosenmüller J. On a generalization of the Lemke-Howson algorithm to noncooperative N -person games. *SIAM J Appl Math* 1971;21:73–79.
28. Wilson RB. Computing equilibria on n -person games. *SIAM J Appl Math* 1971;21(1):80–87.
29. Harsanyi JC. Oddness of the number of equilibrium points: a new proof. *Int J Game Theor* 1973;2(1):235–250. DOI: 10.1007/BF-01737572.

30. Goldberg K, Goldman AJ, Newman M. The probability of an equilibrium point. *J Res Natl Bur Stand Sect B* 1968;72B:93–101.
31. Dresher M. Probability of a pure equilibrium point in n -person games. *J Comb Theor* 1970; 8:134–145.
32. Powers IY. Limiting distributions of the number of pure strategy Nash equilibria in N -person games. *Int J Game Theor* 1990;19(3): 277–286. DOI: 10.1007/BF01755478.
33. Stanford W. A note on the probability of k pure Nash equilibria in matrix games. *Games Econ Behav* 1995;9(2):238–246. DOI: 10.1006/game.1995.1019.
34. Kreps VL. Bimatrix games with unique equilibrium points. *Int J Game Theor* 1974;3(2): 115–118. DOI: 10.1007/BF01766397.
35. Heuer GA. Uniqueness of equilibrium points in bimatrix games. *Int J Game Theor* 1979;8(1):13–25. DOI: 10.1007/BF01763049.
36. Quintas LG. Constructing bimatrix games with unique equilibrium points. *Math Soc Sci* 1988;15(1):61–72. DOI: 10.1016/0165-4896(88)90029-7.
37. Kreps VL. Finite N -person non-cooperative games with unique equilibrium points. *Int J Game Theor* 1981;10(3-4):125–129. DOI: 10.1007/BF01755957.
38. Quint T, Shubik M. A theorem on the number of Nash equilibria in a bimatrix game. *Int J Game Theor* 1997;26(3):353–359. DOI: 10.1007/s001820050038.
39. von Stengel B. New maximal numbers of equilibria in bimatrix games. *Discrete Comp Geom* 1999;21(4):557–568. DOI: 10.1007/PL00009438.
40. McLennan A, Berg J. Asymptotic expected number of Nash equilibria of two-player normal form games. *Games Econ Behav* 2005;51(2):264–295. DOI: 10.1016/j.geb.2004.10.008.
41. McLennan A. The expected number of Nash equilibria of a normal form game. *Econometrica* 2005;73(1):141–174. DOI: 10.1111/j.1468-0262.2005.00567.x.
42. von Stengel B. Computing equilibria for two-person games. In: Aumann Robert, Hart Sergiu, editors. Volume 3, Handbook of game theory with economic applications. Amsterdam, Netherlands: Elsevier Science; 2002. pp. 1723–1759.
43. Brown GW. Iterative solution of games by fictitious play. Activity analysis of production and allocation. Cowles commission monograph no 13. New York: John Wiley & Sons, Inc.; 1951. pp. 374–376.
44. Robinson J. An iterative method of solving a game. *Ann Math* 1951;54(2):296–301.
45. McKelvey RD, McLennan AM, Turocy TL. Gambit: Software Tools for Game Theory. Version 0.2007.01.30. 2007. <http://www.gambit-project.org>.
46. Barron EN. Game theory: an Introduction. Hoboken (NJ): John Wiley & Sons, Inc.; 2008.
47. Selten R. Reexamination of the perfectness concept for equilibrium points in extensive form games. *Int J Game Theor* 1975;4(1): 25–55. DOI: 10.1007/BF01766400.
48. Myerson RB. Refinements of the Nash equilibrium concept. *Int J Game Theor* 1978;7(2): 73–80. DOI: 10.1007/BF01753236.
49. Halpern JY. A nonstandard characterization of sequential equilibrium, perfect equilibrium, and proper equilibrium. *Int J Game Theor* 2009;38(1):34–49. DOI: 10.1007/s00182-008-0139-0.
50. Weibull JW. Evolutionary game theor. Cambridge (MA): M.I.T. Press; 1995.
51. Maynard Smith J. The theory of games and the evolution of animal conflicts. *J Theor Biol* 1974;47(1):209–221. DOI: 10.1016/0022-5193(74)90110-6.
52. Hofbauer J, Sigmund K. Evolutionary games and population dynamics. New York: Cambridge University Press; 1998.
53. Swinkels JM. Evolutionary stability with equilibrium entrants. *J Econ Theor* 1992; 57(2):306–332. DOI: 10.1016/0022-0531(92)90038-J.
54. Taylor P, Jonker L. Evolutionary stable strategies and game dynamics. *Math Biosci* 1978;40(1-2):145–156. DOI: 10.1016/0025-5564(78)90077-9.
55. Van Damme E. Strategic equilibrium. In: Aumann Robert, Hart Sergiu, editors. Volume 3, Handbook of game theory with economic applications. Amsterdam, Netherlands: Elsevier Science; 2002. pp. 1521–1596.
56. Hillas J, Kohlberg E. Foundations of strategic equilibrium. In: Aumann R, Hart S, editors. Volume 3, Handbook of game theory with economic applications. Amsterdam, Netherlands: Elsevier Science; 2002. pp. 1597–1663.
57. Raghavan TES. Non-zero-sum two-person games. In: Aumann R, Hart S, editors. Volume 3, Handbook of game theory with economic applications. Amsterdam, Netherlands: Elsevier Science; 2002. pp. 1687–1721.

FURTHER READING

Aliprantis CD, Chakrabarti SK. Games and Decision Making. New York, NY: Oxford University Press, Inc; 1999.

Kuhn HW, editor. Classics in game theory. Princeton, NJ: Princeton University Press; 1997. (This is a compendium of previously published articles including [3,4], and [44].).

Peters H. Game theory: A multi-leveled approach. Heidelberg, Germany: Springer; 2008.

Rapaport A. Two-person game theory. (Republication of 1966, University of Michigan Press, Ann Arbor, MI edition). Mineola, NY: Dover Publications, Inc.; 1999.

Straffin PD. Game theory and strategy. Washington, DC: Mathematical Association of America; 1996.

NETWORK RELIABILITY PERFORMANCE METRICS

SALVATORE D. ANTONIO
Department of Technology,
University of Naples
“Parthenope”, Naples, Italy

INTRODUCTION

This section illustrates the existing definitions for network performance and reliability metrics proposed by relevant standardization bodies such as IETF (Internet Engineering Task Force) or ITU-T. Two different definitions are provided by the two bodies for each metric, but it is worth noting that the different definitions of similar metrics are in most cases compatible, that is, semantically equivalent or easily convertible into one another. These metrics can be measured by using two alternative approaches: passive measurements and active measurements.

Active measurement techniques inject traffic into the network and evaluate traffic metrics by monitoring the injected traffic or the response to the injected traffic. Active test traffic may perturb other traffic already present on the network, the so-called background traffic; therefore, scheduling and volume of the injected traffic must be carefully configured.

Passive measurements provide information about traffic in the observed network by capturing all or a selected subset of the traffic flow packets traversing a monitoring point. Since no test traffic is generated, passive measurements can only be applied when the traffic of interest is already present on the network. The deployment of monitoring probes in the network can be performed in different ways, depending on the metrics of interest and on the communication technology; for example, via a physical line splitter, via a normal client connection in broadcast networks, or via a dedicated monitoring port on a switch or router.

NETWORK PERFORMANCE METRICS

One-Way Delay

Delay is used to measure the expected time for a network traffic packet to traverse the network from one host to another one.

IETF RFC 2679 [1] defines the one-way delay metric as the difference between the time when the destination host received the last bit of the test packet and the time when the source host sent the first bit of that packet.

RFC 2679 also distinguishes between a “singleton analytic metric” and a “sample.” The singleton is introduced to measure a single observation of one-way delay, while the sample is used to measure a sequence of singleton delays measured at times taken from a Poisson process. Based on these samples, several statistics are defined such as one-way-delay percentile, one-way-delay median, one-way-delay minimum, and one-way-delay inverse percentile.

If the test packet fails to arrive within a reasonable period of time, the one-way delay is taken to be undefined (informally, infinite).

The ITU-T document Y.1540 [2] defines the one-way delay metric as the time, $(t_2 - t_1)$ between the occurrence of two corresponding packet reference events, ingress event at time t_1 and egress event at time t_2 , where $(t_2 > t_1)$ and $(t_2 - t_1) \leq T_{\max}$. A packet transfer event occurs when the following conditions are met:

- packet crosses a measurement point;
- standard procedures applied to the packet verify that the header checksum is valid;
- the source and destination address fields within the packet header represent the addresses of the expected source host and destination host.

The two definitions from IETF and ITU-T for one-way delay can be considered compatible, as for both definitions the relevant events are (i) the time just before the packet is being put

on the wire and (ii) the time after complete reception of the packet at the destination.

To measure one-way metrics, synchronization of both transmission ends of the measurement device needs to be ensured. This synchronization may be done by GPS (global positioning system) clocks when the two ends are distant. When ends are colocated, synchronization may be done directly by the analyzer.

Round-Trip Delay

Round-trip delay is used to measure the expected time for network interaction between two hosts on a network; conceptually, it is equivalent to the sum of the two one-way delays, each experienced in one direction between two hosts.

Active measurement of round-trip delay as defined by the IETF in RFC 2681 [3] requires the observation of test packets transmitted in both directions between two endpoints across a network: a “source” host which sends the first packet and a “destination” host, which receives the first test packet and sends a test packet back to the source in reply. The round-trip delay is then calculated as the difference between the time at which the reply is received at the source and the time at which the original test packet was sent.

The procedure for passive measurement of round-trip delay is similar to that for active measurement: a packet is sent from a source to a destination, that packet causes the destination to send the packet back to the source, and the packets are identifiable at the source in order to correlate each packet of the round trip and calculate a delay. There are two potential architectures here: one utilizing a source observation point (OP) placed topologically close to the source of traffic, and one utilizing an additional destination OP placed topologically close to the destination of traffic.

Delay Variation

Delay variation is used to measure the variation in a sequence of delay values over time.

In RFC 3393 [4], the packet delay variation is presented as the difference between the one-way delays of two selected packets within a stream of packets across the

network paths. The definition relies on the introduction of a selection function F defining unambiguously the two packets from the stream selected for the metric.

Examples of a selection function are as follows:

- consecutive packets within the specified interval;
- packets with specified indices within the specified interval;
- packets with the minimum and maximum one-way delays within the specified interval;
- packets with specified indices from the set of all defined (i.e., noninfinite) one-way delay packets within the specified interval.

In Recommendation Y.1540, packet delay variation is defined based on the observations of corresponding packet arrivals at ingress and egress measurement points. More precisely, the end-to-end packet delay variation (v_k) for a packet k between a source host and a destination host is the difference between the absolute packet transfer delay (x_k) of the packet and a defined reference packet transfer delay, $d_{1,2}$, between the same measurement points: $v_k = x_k - d_{1,2}$.

The reference packet transfer delay, $d_{1,2}$, between the source host and the destination host can be the absolute packet transfer delay experienced by the first packet between the two measurement points or any other fixed packet delay.

Therefore, the delay variation of an individual packet is naturally defined as the difference between the actual delay experienced by that packet and a reference delay.

Recommendation Y.1540 also presents two alternative methods for summarizing the delay variation experienced by a population of packets:

A first method is to prespecify a delay variation interval and then observe the percentage of individual packet delay variations that fall inside and outside of that interval.

A second method consists of selecting upper and lower quantiles of the delay variation distribution and then measuring

the distance between those quantiles. It is suggested in Y.1540 to select the 99.9th percentile and then 0th percentile (i.e., the minimum), make measurements, and observe the difference between the delay variation values at these two quantiles for each interval.

Packet Loss

Packet loss denotes the expected probability that a transmitted packet will reach its destination and is usually expressed as a percentage. It is a general metric for a network's packet delivery reliability. The IETF and ITU-T approaches are generally compatible, given assumptions about the measurement methods in use.

RFC 2680 [5] defines the one-way packet loss metric as follows: the one-way packet loss of a packet from a source host to a destination host is 0 if the source host sent the first bit of the packet to the destination host at time T and the destination host received that packet. The one-way packet loss of a packet from a source host to a destination host is equal to 1 if the source host sent the first bit of the packet to the destination host at time T and the destination host did not receive that packet.

RFC 2680 also relates the definition of one-way packet loss to one-way delay. More precisely, it is stated that the one-way packet loss is equal to 0 exactly when the one-way delay assumes a finite value, whereas an infinite value of the one-way delay corresponds to a one-way packet loss equal to 1. Furthermore, the document outlines the need for the measurement methodology to include a way to distinguish between a packet loss and a very large, but finite, delay. As with delay measurements, the threshold time to consider a late packet lost should be reported as part of the measurement results.

Recommendation Y.1540 defines the packet loss ratio as the ratio of total lost packet outcomes to total transmitted packets in a population of interest.

One-way active packet loss measurements always include the sending of test packets from a sender to a packet receiver in a predefined sending schedule and/or pattern; the receiver will then count these packets.

Usually, the test packets contain a specific content including sequence numbers and timestamp so that analysis by the receiver can be more detailed.

Passive measurements of one-way packet loss are only possible through multipoint measurement, or by observing traffic from which packet loss can be deduced from higher layer header fields. In both cases, it is important to ensure that all of the packets of interest do pass the observation point(s). In the first approach, routers or measurement probes along the path of the traffic of interest are instructed to filter and count the packets. Periodically, the probes export counters per flow and differences along the path can be compared to compute and localize the packet loss.

The second approach is possible with a single point measurement, but only for protocols whose packet headers have certain fields allowing the deduction of skipped, lost, or retransmitted packets.

NETWORK RELIABILITY METRICS

Availability

In general, the availability of a system as a function of time, $A(t)$, can be defined as the probability that the system is operational at the instant of time, t . If the limit of this function exists as t goes to infinity, it expresses the expected fraction of time that the system is available to perform useful computations.

ITU-T defines the service availability metric in ITU-T Recommendation Y. 1540 [2]. Such metric relies on the notion of packet loss ratio, measured as a percentage of lost packets across an interval of time. If the packet loss percentage is below a certain defined level during this time interval, then the service is called available in this time slot, else it is unavailable. On the basis of the availability results (yes/no) for consecutive time intervals, the document defines the percentage of (un)availability as follows:

- *Percent Service Unavailability.* The percentage of total scheduled service time

that is categorized as unavailable using the service availability function.

- *Percent Service Availability.* The percentage of total scheduled service time that is classified as available using the service availability function: percent service availability = $100 - \text{percent service unavailability}$.

IETF defines the concept of connectivity as a measure of the expected probability that one host can reach another and then, as an attribute affecting the service availability. RFC 2678 [6] defines metrics which allow to define the level of connectivity between two hosts A_1 and A_2 . It starts with the definition of an analytic metric, called instantaneous unidirectional connectivity, to define one-way connectivity at one single moment in time. This metric is extended to introduce the concept of bidirectional connectivity (for one single moment in time), and later to connectivity within a certain period of time. Each of the mentioned metrics produces a Boolean value which states whether the desired level of connectivity has been achieved or not. A methodology for estimating the last defined metric is sketched in the document. This methodology consists in using randomly distributed start times for sending the probe packets within a selected interval of time. The document also shows how to apply this testing methodology for checking TCP-based connectivity.

Reliability

The reliability of a system as a function of time, $R(t)$, is the conditional probability that the system has survived the interval $[0, t]$, given that the system was operational at time $t = 0$.

Reliability in this context is used to describe systems in which repair cannot take place; for example, systems in which the computer is serving a critical function and cannot be lost even for the duration of a repair, or systems where the repair is prohibitively expensive.

The reliability function $R(t)$ holds the following properties:

- It is a monotonically decreasing function.
- $R(t = 0) = 1$ (the component is assumed to work correctly in the beginning).
- For $R(t = \text{TTF}) = 0$ (the component will eventually fail), where TTF is the time to failure.

Unreliability, Probability of Failure

Unreliability (also referred to as the probability of failure) $F(t)$, is a property related to reliability by this law:

$$R(t) + F(t) = 1$$

$F(t)$ is the cumulative distribution function (CDF) of the TTF random variable and represents the probability that the system fails in the time interval $[0, t]$. The parameter to be managed is the mean or expected value, the mean time between failure (MTBF).

Time to Repair

Time to repair (TTR) is the amount of time needed to restore system's correct service delivery after a failure has occurred. The parameter to be managed is the mean or expected value, the mean time to repair (MTTR). This metric represents the average time required to repair a failed component or device.

Maintainability and Maintenance

The maintainability CDF, $M(t)$, is defined as the probability of performing a successful repair action within a given time period. In other words, maintainability measures the ease and speed with which a system can be restored after a failure has occurred, or a maintenance plan may extend the lifetime of a system by interceding at a point before the TTF to replace components and thus, reset the TTF. For a maintenance plan to be effective, the time to maintain should be less than TTR (including any time to analyze the cause of a fault and to identify the required repair). If a system fails, a time variable may be introduced to measure the time needed to analyse the failure and repair the system.

Coverage

The coverage, c , is the parameter that measures the fault tolerance of a system. It is defined as the conditional probability that given the existence of a failure in the operational system, the system is able to recover and continue information processing with no permanent loss of essential information; that is, $c = P(\text{system recovers} \mid \text{system fails})$.

REFERENCES

1. IETF RFC 2679: "A One-way Delay Metric for IPPM". Almes G., Kalidindi S., Zekauskas M. September 1999.
2. ITU-T Recommendation Y. 1540: "Internet protocol data communication service-IP packet transfer and availability performance parameters".
3. IETF RFC 2681: "A Round-trip Delay Metric for IPPM". Almes G., Kalidindi S., Zekauskas M. September 1999.
4. IETF RFC 3393: "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)". Demichelis C., Chimento P. November 2002.
5. IETF RFC 2680: "A One-way Packet Loss Metric for IPPM". Almes G., Kalidindi S., Zekauskas M. September 1999.
6. IETF RFC 2678: "IPPM Metrics for Measuring Connectivity". Mahdavi J., Paxson V. September 1999.

NETWORK THEORY: CONCEPTS AND APPLICATIONS

ALBERTO GARCIA-DIAZ
Department of Industrial and
Information Engineering, The
University of Tennessee,
Knoxville, Tennessee

DON T. PHILLIPS
Department of Industrial and
Systems Engineering, Texas A&M
University, College Station, Texas

ILLYA HICKS
Department of Computational and
Applied Mathematics, Rice
University, Houston, Texas

INTRODUCTION

A modern society may be viewed as a system of networks for transportation, communication, and distribution of energy, goods, and services. Network analysis can be used in the design, improvement, and rationalization of these subsystems. The origins of *graph theory* and *networks analysis* are old and diverse. In 1736, Leonhard Euler wrote the first article on a topic related to graph theory. The solution to the Königsberg's bridge problem considered in his article involves several basic concepts of graph theory. A succinct description of the problem follows.

One of the cities in Prussia was Königsberg. In this city there is an island called Kneiphof, surrounded by the river Pregel, as shown in Fig. 1. In this figure, A represents an island and B, C, and D represent land areas. The branches of the river are crossed by seven bridges a, b, c, d, e, f, and g. Is it possible to arrange a route such that each bridge is crossed exactly once? In 1736, at the age of 29, Euler modeled this problem using nodes to represent land areas and arcs to represent bridges. In his famous paper "The solution of a problem relating to the geometry of

position" (translated from German) [1], Euler demonstrated that this problem had no solution and indicated the required conditions for the solution to exist. Euler's procedure for solving the Königsberg bridge problem can be found in several bibliographical references, such as those by Biggs, Lloyd, and Wilson [2], Newman [3], and Scientific American [4].

The following milestones in the history of network analysis mark important developments since the beginning through the 1960s:

1. Euler, 1736: the Königsberg bridge problem
2. Kirchhoff, 1847: law of conservation of electrical current
3. Maxwell, 1864: electromagnetic theory applications
4. König, 1936: first book
5. Hitchcock, 1941: economic aspects of distribution problems
6. Koopmans, 1947: optimal utilization of transportation systems. This work marks a clear separation between network analysis and graph theory
7. Between 1950 and 1965: much emphasis was placed on the development of algorithms for linear networks
8. More recently: a strong effort has been concentrated on generalized networks, multicommodity networks, and nonlinear costs

In 1956, Orden [5] proposed a generalization of the transportation problem (TP) known as the *transshipment problem*. At approximately the same time, the maximum-flow problem and the minimum-cost flow problem were formulated and investigated by Ford and Fulkerson. From 1950 through 1965 much activity was directed toward developing *primal simplex methods* and *primal-dual methods*. The primal simplex specializations began with the work of Dantzig [6] and culminated with the paper by Johnson [7]. The primal-dual methods originated with Kuhn's Hungarian algorithm

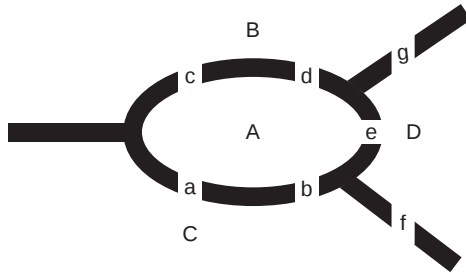


Figure 1. Königsberg's seven bridges.

for the assignment problem (AP) [8], and culminated with the *out-of-kilter algorithm* by Fulkerson [9]. In 1975, the Royal Swedish Academy of Science awarded the Nobel Prize in Economic Science equally to Professor L. Kantorovich of the USSR and Professor T. C. Koopmans of the US. Both these individuals are associated with some of the first papers describing network-flow problems as we know them today.

Most of the work since the mid-1960s has involved the efficient implementation of basic techniques and their extension to (i) linear networks with arc flow gain/loss factors, (ii) linear multicommodity networks, (iii) linear networks with side constraints, and (iv) network problems with nonlinear convex cost functions. New methods invented in the 1980s challenged the old ones. Two of these methods originally proposed by Bertsekas are called *relaxation and auction* [10,11].

Network models have some distinct characteristics that have resulted in a remarkable growth in both industry and business applications. The following are perhaps the most important of these characteristics:

1. Ability to model complex systems
2. Mechanistic procedure
3. Identification of significant operation features
4. Data requirements
5. Starting point for further analysis

Furthermore, network-based solution procedures have some clear advantages over other more general models, including the following:

1. Accuracy
2. Acceptability

3. Computational efficiency
4. Effectiveness when used as components of optimization procedures

Network models have been successfully applied in a wide range of situations. Most of these models can be classified according to the following types of problems:

1. Warehousing and distribution
2. Project planning and scheduling
3. Equipment replacement
4. Cost control
5. Traffic studies
6. Transportation systems
7. Queueing analysis
8. Assembly line balancing
9. Inventory control
10. Manpower allocation (and many others)

Definitions, Notations, and Symbolism

A *node* is a point where flow is originated, relayed, or terminated. If the *net flow* (defined as flow out minus flow in) is strictly positive the node is a *source*, if it is equal to zero the node is an *intermediate node*, and if it is strictly negative the node is a *terminal*. An *arc* is a generic channel for transmission of flow from a given node to another one directly connected from it. The *orientation* of the arc is represented by an arrow connecting the start node to the end node of the arc. There are three types of arcs (also called branches): *undirected*, *directed*, and *bidirected*, depending on the specification of the direction along which flow can move along the arcs. An undirected arc has no direction, while a directed arc has a unique direction specified in advance. A bidirected arc actually has two directions specified in advance. To represent networks graphically, we use circles to indicate nodes, lines to indicate arcs, and arrowheads to indicate arc orientation. We also use the notation (i,j) to refer to the directed arc leading from node i to node j . Figure 2a shows an undirected arc, Fig. 2b a directed arc, and Fig. 2c a bidirected arc.

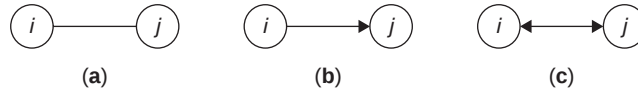


Figure 2. Node-arc representations: (a) undirected arc; (b) directed arc; (c) bidirected arc.

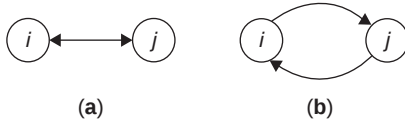


Figure 3. Representation of a bidirected arc by two directed arcs.

For most practical purposes, there is no necessity to distinguish between undirected and bidirected arcs. Further, it is evident that a bidirected arc between two arbitrary nodes i and j can be represented by two directed arcs (i, j) and (j, i) , as shown in Fig. 3.

If an arc is traversed by flow in the same direction of the orientation, the arc is said to be a *forward* arc. If it is traversed in a direction opposite to its orientation, the arc is said to be a *reverse* arc. When the flow along an arc is constant, the arc is referred to as a *pure* arc; otherwise, it is referred to as a *generalized* arc. Generalized arcs allow *gains* and *losses* on arc flows. In these arcs the flow at the end node is equal to the flow at the start node multiplied by a positive constant known as the gain/loss factor.

A *network* is a collection of nodes and arcs, with one or more quantitative elements associated with each arc or node. Networks can be *pure* or *generalized*, depending on the values of the gain/loss factors. A *graph* $G = (\mathbf{N}, \tau)$ is the pair consisting of a set $\mathbf{N}$ and function τ mapping $\mathbf{N}$ into $\mathbf{N}$. Whenever possible, the elements of set $\mathbf{N}$ will be represented by points in the plane and if x and y are two points such that $y \in \tau x$, then they will be joined by a continuous line with an arrow head pointing from x to y . Hence, an element of $\mathbf{N}$ is called a point or vertex or node of the graph while the pair (x, y) with $y \in \tau x$ is called an arc of the graph. The set of arcs of the graph is denoted by $\mathbf{A}$ and instead of $G = (\mathbf{N}, \tau)$ we can write $G = (\mathbf{N}, \mathbf{A})$. An illustration of a network is given in

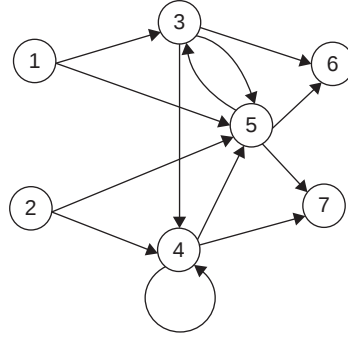


Figure 4. Illustrative network.

Fig. 4. Nodes 1 and 2 are source nodes and nodes 6 and 7 are sink nodes. Here, $\mathbf{N} = \{1, 2, 3, 4, 5, 6, 7\}$. Also, $\tau_1 = \{3, 5\}$, $\tau_2 = \{4, 5\}$, $\tau_3 = \{4, 5, 6\}$, $\tau_4 = \{4, 5, 7\}$, $\tau_5 = \{3, 6, 7\}$, $\tau_6 = \tau_7 = \emptyset$. Therefore, $\mathbf{A} = \{(1, 3), (1, 5), (2, 4), (2, 5), \dots, (5, 7)\}$.

A *chain* is a connected sequence of all forward (or all reverse) arcs. An example of a chain is the sequence of nodes 1-2-3-4-5-6 in Fig. 5a. A *path* is a connected sequence of forward and reverse arcs. An example of a path is the sequence 1-4-3-6 in Fig. 5a. A *circuit* is a chain such that the first and last nodes coincide. An example of a circuit is the sequence 3-4-5-3 in Fig. 5a.

A *cycle* is a path such that the first and last nodes coincide. An example of a cycle is the sequence 2-3-4-2 in Fig. 5b. A *self-loop* is a one-arc cycle (circuit). An example of a self-loop is shown in Fig. 4 for node 4. Often in the network literature we can find that the terms *chain* and *circuit* are defined as we have defined the terms *path* and *cycle*, respectively, and vice versa.

As mentioned in the introduction of this article, in his article on the Königsberg bridge problem, Euler replaced the map shown in Fig. 1 by a graph in which nodes represent land areas and arcs represent bridges, as shown in Fig. 6. Using this figure, the

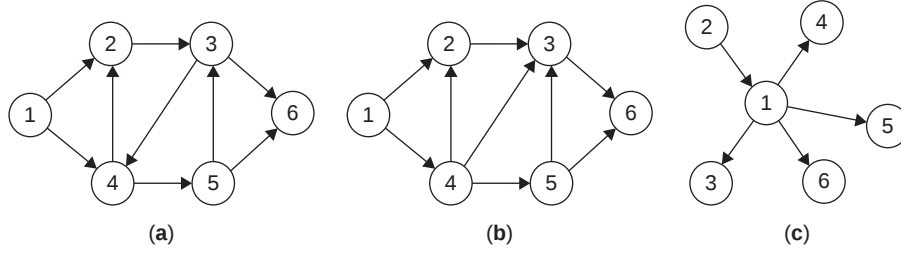


Figure 5. (a) Network with circuits; (b) network without circuits; (c) acyclic network.

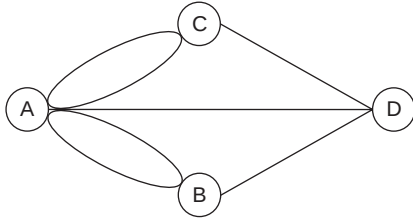


Figure 6. Königsberg bridge problem graph.

object of the Königsberg bridge problem can be restated as finding a path that contains each arc of the graph exactly once. A path of this kind is now known as an *Eulerian path*.

A *connected* graph is one such that at least one path exists between every pair of nodes. A *tree* is a finite connected graph with no cycles and possessing at least two nodes. A *spanning tree* is a tree which contains *all* nodes of a given graph. A finite graph is an *arborescence* with root x if: (i) every node different from x is the terminal node of a single arc, (ii) x is not the terminal node of any arc, (iii) there are no circuits.

Figure 7a shows an undirected network, and Fig. 7b–c shows a spanning tree and a nonspanning tree, respectively, for that network. Figure 7d shows a directed network, and Fig. 7e–f shows an arborescence and an arborescent spanning tree, respectively, for the network. Figure 7g–h shows a tree and a spanning tree, respectively, for the network shown in Fig. 7d.

Other graph theory concepts (where the orientation of arcs is undirected) that may prove useful are the following. A *subgraph* of a graph $(\mathbf{N}, \mathbf{A})$ is a pair $(\mathbf{N}', \mathbf{A}')$ where $\mathbf{N}'$ is a subset of $\mathbf{N}$ and $\mathbf{A}'$ is a subset of

$\mathbf{A}$, where $(i, j) \in \mathbf{A}'$ only if both i and j are elements of $\mathbf{N}'$. A *clique* of a graph $(\mathbf{N}, \mathbf{A})$ is a subgraph $(\mathbf{N}', \mathbf{A}')$ where for all $i, j \in \mathbf{N}'$ we have $(i, j) \in \mathbf{A}'$. Furthermore, given a graph $(\mathbf{N}, \mathbf{A})$, a subgraph $(\mathbf{N}', \mathbf{A}')$ where every node of $\mathbf{N}'$ has exactly one arc of $\mathbf{A}'$ associated with it is called a *matching*. Finally, a graph is considered *bipartite* if it contains no odd cycles.

Matrix Representation of Networks

Let $G = (\mathbf{N}, \mathbf{A})$ be a directed network. The *adjacency matrix* summarizes the *connectivity* of the network by specifying which nodes are directly connected by arcs. It is defined as $\mathbf{X} = [x_{ij}]$, where $x_{ij} = 1$ if (i, j) exists, and 0 otherwise. The *node-arc-incidence matrix* also summarizes the connectivity of the network by specifying the position of each node with respect to the arcs of the set $\mathbf{A}$. It is defined as $\mathbf{Z} = [z_{ik}]$, where $z_{ik} = +1$ if node i is the starting node of arc a_k ; $z_{ik} = -1$ if node i is the ending node of arc a_k ; $z_{ik} = 0$, otherwise. As an illustration, the network shown in Fig. 8 is considered. For this network, $\mathbf{N} = \{1, 2, 3, 4, 5, 6\}$ and $\mathbf{A} = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9\}$. Also, the adjacency matrix is shown below:

$$\mathbf{x} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Again, for the network of Fig. 8, the node-arc-incidence matrix is shown below:

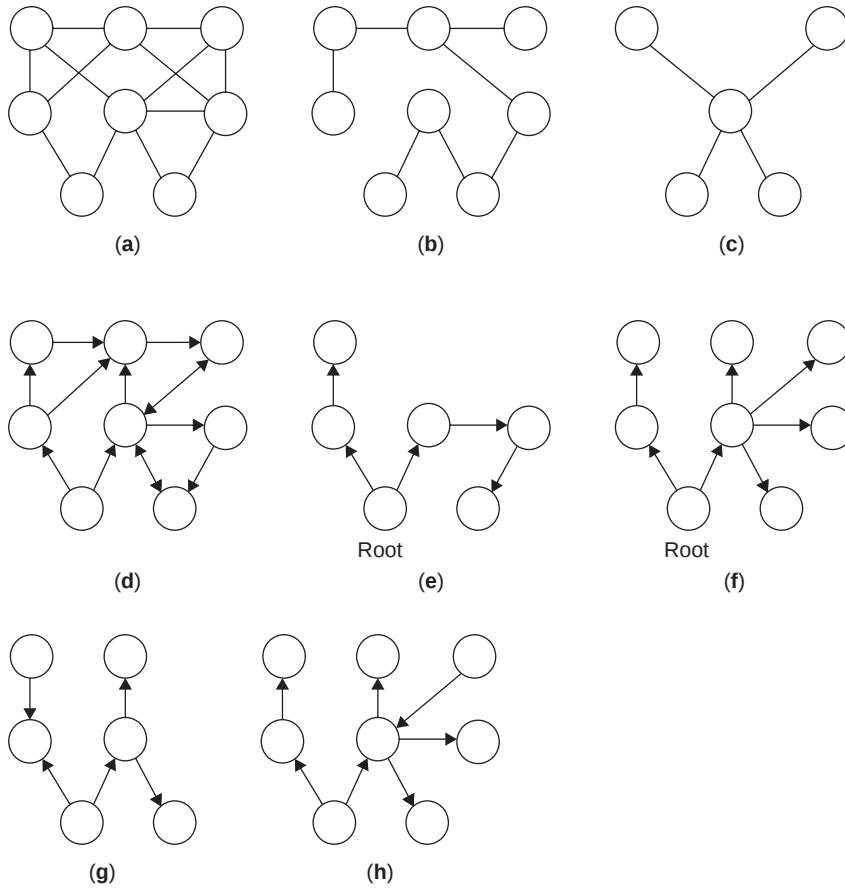


Figure 7. Trees and arborescences.

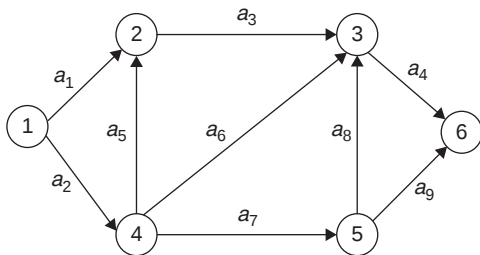


Figure 8. Directed network consisting of 6 nodes and 9 arcs.

$$\mathbf{z} = \begin{bmatrix} +1 & +1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & +1 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & +1 & +1 & +1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & +1 & +1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

A network can be represented more effectively using two *connectivity lists*: (i) nodes at the beginning of directed arcs going into nodes $1, 2, 3, \dots, n$, (ii) number of directed arcs into nodes $1, 2, 3, \dots, n$. Alternatively, the first list shows nodes at the end of arcs coming from nodes $1, 2, 3, \dots, n$, and the second list indicates the number of arcs from each node. The information in the distance (or cost or capacity) matrix can be more efficiently stored in a third list when using connectivity lists. Consider again the network given in Fig. 8. This network can be represented in terms of two lists: (i) nodes at the beginning of directed arcs going into nodes $1, 2, 3, \dots, n$, (ii) number of directed arcs into nodes $1, 2, 3, \dots, n$. These two lists are shown below:

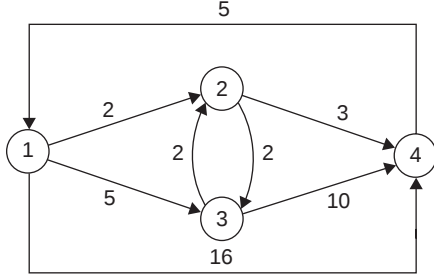


Figure 9. Network with arc lengths.

List 1:

-1	4	2	4	5	1	4	3	5
----	---	---	---	---	---	---	---	---

List 2:

0	2	3	1	1	2
---	---	---	---	---	---

The *cost* or *distance matrix* summarizes the natural limitations and capabilities of the network (system). Let $G = (\mathbf{N}, \mathbf{A})$ be a directed network. The *cost* or *distance matrix* is defined by $\mathbf{C} = [c_{ij}]$, where c_{ij} is the parameter associated with each arc (i, j) in $\mathbf{A}$. As an example, the network shown in Fig. 9 will be considered.

For this network the distance matrix $\mathbf{C}$ is given below:

$$\mathbf{C} = \begin{bmatrix} 0 & 2 & 5 & 16 \\ \infty & 0 & 2 & 3 \\ \infty & 2 & 0 & 10 \\ 5 & \infty & \infty & 0 \end{bmatrix}$$

The information in the distance (or cost or capacity) matrix can be more efficiently stored in a third list when using the *forward-star representation*. As an illustration, the connectivity and arc length information for the network shown in Fig. 9 can be represented as follows:

List 1:

4	1	3	1	2	1	2	3
---	---	---	---	---	---	---	---

List 2:

1	2	2	3
---	---	---	---

List 3:

5	2	2	5	2	16	3	10
---	---	---	---	---	----	---	----

Conservation of Flow

First we consider a pure network. Let s, t be source and terminal nodes, α_i the set of all directed arcs starting at node i , and β_i the set of all arcs ending at node i . Furthermore, let

c_{ij} be the capacity of (i, j) . Flow conservation at node i means that the net flow at this node is equal to zero. The flow conservation is formulated in Equation (1):

$$\sum_{j \in \alpha_i} f_{ij} - \sum_{j \in \beta_i} f_{ji} = 0 \quad i \neq s, t \quad (1)$$

where, for feasibility, the condition $0 \leq f_{ij} \leq c_{ij}$, must hold for all (i, j) .

In the case of a generalized network the flow conservation condition for node i is formulated as shown in Equation (2), where a_{ji} is the *gain/loss factor* of arc (j, i) :

$$\sum_{j \in \alpha_i} f_{ij} - \sum_{j \in \beta_i} a_{ji} f_{ji} = 0 \quad i \neq s, t \quad (2)$$

It can be easily verified that the coefficient matrix corresponding to the flow conservation constraints is the $\mathbf{Z}$ matrix (node-arc-incidence matrix). To illustrate this, the *pure* network shown in Fig. 10a will be considered.

The numbers associated with the arcs of the network in Figure 10a are the *gain/loss* factors, which by definition are all equal to 1. Figure 10b shows a *generalized* network with gain/loss factors given next to each arc. The coefficient matrix corresponding to the *pure* network given in Fig. 10a is

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & 0 & -1 \end{bmatrix}$$

Similarly, the coefficient matrix corresponding to the *generalized* network shown in Fig. 10b is

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ -3 & 0 & 1 & 1 & -1/4 & 0 \\ 0 & -6 & -2 & 0 & 1 & 1 \\ 0 & 0 & 0 & -2/3 & 0 & -1/3 \end{bmatrix}$$

If we now assume that node 1 has a supply of 5 units of flow and node 2 has a demand of one unit in the network shown in Fig. 10b, the flow at the end of path 1-3-2-4 is equal to $[5(6)(1/4) - 1](2/3) = 13/3$. Moreover, the flow conservation condition for node 3 is $f_{32} + f_{34} - 2f_{23} - 6f_{13} = 0$.

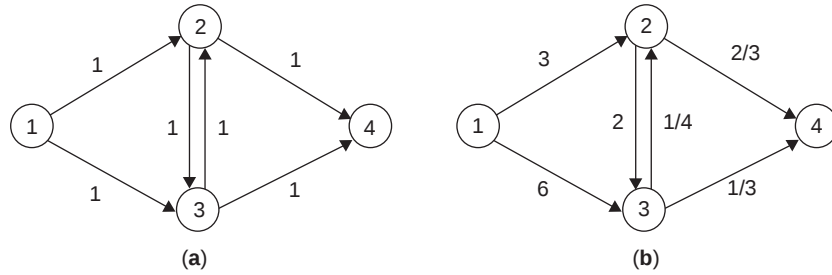


Figure 10. (a) Pure network; (b) Generalized network.

Cut and Cut Value

Let us for now consider a network with 0 lower bounds on arc flows. A *cut* is defined as a *set of arcs* whose removal will *not* allow any flow to be transmitted from a source s to a terminal t . As an illustration, in the network shown in Fig. 11, it is noted that a cut disconnecting node 1 from node 4 consists of arcs (1,3), (2,3), and (2,4). This cut has a capacity or value equal to 15.

More generally, the value of a cut between two nodes s and t on a directed pure network with *positive* lower bounds L_{ij} and upper bounds U_{ij} on arc flows is defined in Equation (3):

$$C = \sum_{i \in N_1} \sum_{j \in N_2} U_{ij} - \sum_{i \in N_1} \sum_{j \in N_2} L_{ji} \quad (3)$$

In the definition given in Equation (3), the set of nodes is partitioned into two sets N_1 and N_2 and, as illustrated in Fig. 12, such that $s \in N_1$ and $t \in N_2$.

Example. The definition of cut-set value given in Equation (3) will be illustrated using the network of Fig. 13, where each arc has two parameters $[L_{ij}, U_{ij}]$ such that $L_{ij} \leq f_{ij} \leq U_{ij}$.

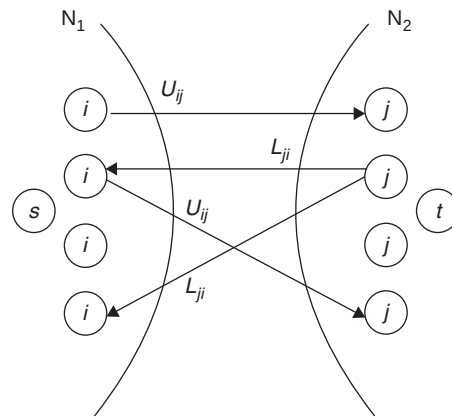
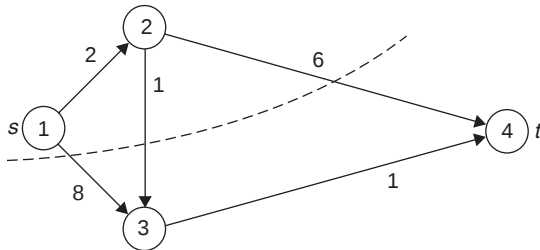


Figure 12. Illustration of cut-set value.

As can be seen in this figure, the nodes are *partitioned* into two sets $N_1 = \{1, 2\}$ and $N_2 = \{3, 4, 5\}$. It is desired to verify that the cut shown is a minimum cut whose value is equal to the maximum flow from node $s = 1$ to node $t = 5$. For the given cut, according to Equation (3), the cut value is equal to

$$\sum_{i \in N_1} \sum_{j \in N_2} U_{ij} - \sum_{i \in N_1} \sum_{j \in N_2} L_{ji} = 5 + 4 - 4 = 5$$

Figure 11. A cut between nodes 1 and 4.

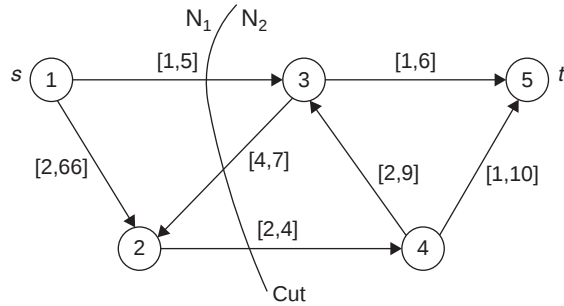


Figure 13. Minimum-cut illustration.

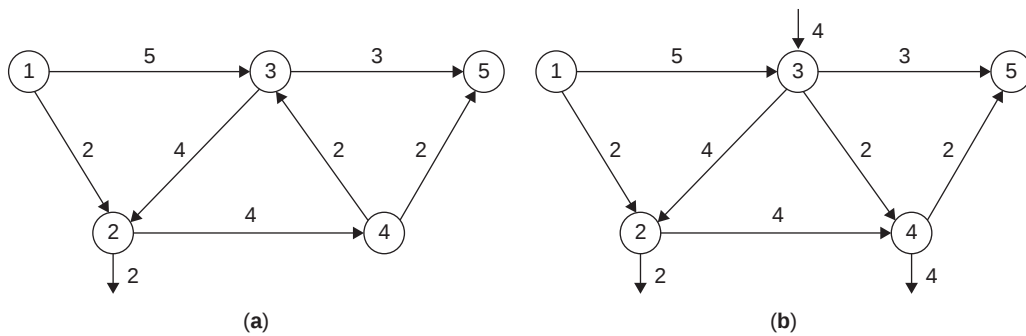


Figure 14. Maximum-flow illustrations.

By inspection, it is possible to determine that the maximum flow is equal to 5. Therefore, the cut shown in Fig. 13 is a minimum cut. The corresponding arc flows resulting in a maximum flow of 5 being delivered to node 5 are shown in Fig. 14a. Note that for the feasibility of the maximum flow, node 2 can be a terminal node with a demand equal to 2. Indeed, node 2 can be viewed as a terminal node with a flow demand equal to at least 2 and at most 66.

To further illustrate the concept under study, let us assume that instead of arc (4,3) in Fig. 13 we have arc (3,4) with parameters [11,12]. In this case, the minimum cut is still the same as in Fig. 13, and the arc flows are as in Fig. 14b. However, for the feasibility of the maximum flow it is necessary to make further assumptions. Specifically, node 3 becomes a source with supply equal to 4 and node 4 becomes a terminal with demand equal to 4, as shown in Fig. 14b. There are other illustrations resulting in the same values for the minimum cut and the maximum flow in Fig. 14.

DETERMINISTIC NETWORK MODELS

This section presents two fundamentally important flow models and several applications of these models. The summary presented here is based on the work by Phillips and Garcia-Diaz [13]. The network models are well known as the Maximum-Flow Model and the Minimum-Cost Model. They will be presented in the section titled "Maximum-Flow and Minimum-Cost Flow Models". The applications are summarized in the section titled "Applications of Network Models".

Maximum-Flow and Minimum-Cost Flow Models

Maximum-Flow Model. Let $G = (\mathbf{N}, \mathbf{A})$ be directed and capacitated, $\mathbf{N} = \{1, 2, \dots, n\}$. Source is node 1; terminal is node n . Also, U_{ij} is the capacity of arc (i, j) and it is assumed that there is an infinite availability of flow at the source node. It is desired to find the

maximum flow shipped from 1 to n through the arcs in $\mathbf{A}$. Let f_{ij} be the flow from i to j along arc (i,j) , and let V be the amount of flow into the terminal (or out of the source). As before, α_i is the set of all directed arcs starting at node i , and β_i is the set of all arcs ending at node i . The linear programming (LP) formulation for this problem is given by Equations (4)–(6):

$$\begin{aligned} & \text{Maximize } V \\ & \text{subject to} \end{aligned} \quad (4)$$

$$\sum_{j \in \alpha_i} f_{ij} - \sum_{j \in \beta_i} f_{ji} = \begin{cases} V & i = 1 \\ 0 & i \neq 1, n \\ -V & i = n \end{cases} \quad (5)$$

$$0 \leq f_{ij} \leq U_{ij} \quad (i,j) \in \mathbf{A} \quad (6)$$

Minimum-Cost Flow Model. Let $G = (\mathbf{N}, \mathbf{A})$ be directed and capacitated. As before, α_i is the set of all directed arcs starting at node i , and β_i is the set of all arcs ending at node i . Let b_i be the flow supply at node i . Let c_{ij} be the per-unit cost of flow along (i,j) . Also, let L_{ij} be a lower bound on the flow on arc (i,j) . The model is shown in Equations (7)–(9):

$$\text{Minimize } \sum_{(i,j) \in \mathbf{A}} c_{ij} f_{ij} \quad (7)$$

subject to

$$\sum_{j \in \alpha_i} f_{ij} - \sum_{j \in \beta_i} f_{ji} = b_i \quad i \in \mathbf{N} \quad (8)$$

$$L_{ij} \leq f_{ij} \leq U_{ij} \quad (i,j) \in \mathbf{A} \quad (9)$$

where b_i is positive for source nodes, zero for intermediate (transshipment) nodes, and negative for terminal nodes.

Maximum-Flow Minimum-Cut Theorem. A *minimum cut* is one having minimum-cut value among all possible cuts between nodes s and t . The *Maximum-Flow Minimum-Cut Theorem* [14] specifies that the *maximum* flow from s to t is equal to the capacity of the *minimum* cut between s and t . As an illustration of this result, it can be verified that the maximum flow that can be shipped from node s to node t in the network of Fig. 15 is

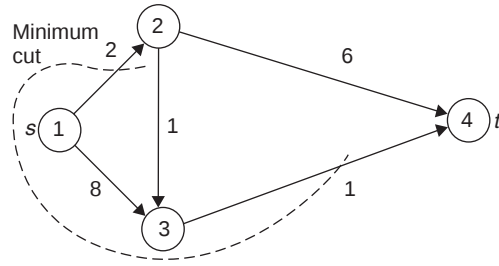


Figure 15. Minimum cut for a directed network.

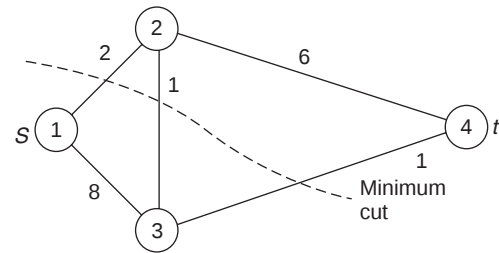


Figure 16. Minimum-cut for an undirected network.

equal to 3. In this figure, the numbers shown on the arcs are the corresponding flow capacities in the specified directions. The minimum cut consists of arcs (1,2) and (3,4), and has a capacity equal to 3 units.

Finally, if the network is undirected, the maximum flow from node s to node t would be equal to 4, since the minimum cut in this case consists of arcs (1,2), (2,3), and (3,4), as shown in Fig. 16.

Minimum-Cost Flow Model for Generalized Networks. At this time we want to emphasize the difference between pure and generalized networks. The network whose formulation is shown in Equations (7)–(9) is referred to as a *pure* network because of the constancy of its flow along the arcs. In many applications, however, the flow entering an arc is not equal to the flow exiting the same arc. This flow may actually be equal to the initial flow multiplied by a positive constant known as the *gain/loss factor*. When the multiplier is larger than 1 it is called a gain factor; otherwise, it is called a loss factor.

Let $G = (\mathbf{N}, \mathbf{A})$ be a directed and capacitated generalized network. The flow entering arc (i,j) is represented by f_{ij} and the

gain/loss factor for this arc is represented by a_{ij} . Furthermore, let c_{ij} , L_{ij} and U_{ij} be the per-unit cost, lower bound, and upper bound, respectively, for the initial flow on arc (i, j) . Also, let b_i be the net flow supply at node i . As before, α_i is the set of all directed arcs starting at node i , and β_i is the set of all arcs ending at node i . The model shown in Equations (7)–(9) can be reformulated as in Equations (10)–(12).

$$\text{Minimize } \sum_{(i,j) \in \mathbf{A}} c_{ij} f_{ij} \quad (10)$$

subject to

$$\sum_{j \in \alpha_i} f_{ij} - \sum_{j \in \beta_i} a_{ji} f_{ji} = b_i \quad i \in \mathbf{N} \quad (11)$$

$$L_{ij} \leq f_{ij} \leq U_{ij} \quad (i, j) \in \mathbf{A} \quad (12)$$

where b_i is positive for source nodes, 0 for intermediate (transshipment) nodes, and negative for terminal nodes.

Applications of Network Models

The applications presented in this section can be classified into two broad categories. The first category includes applications where the solution to a problem is achieved by selecting a chain or cycle that optimizes a suitable objective function. In these problems each arc is interpreted as a directed link that allows the movement of one unit of flow from the start node to the end node of the arc. The arc parameter or *length* is a generic cost of using the arc. Usually, these models are referred to as *distance networks*. In the second category, a more general interpretation of an arc is used, namely, that of a device or channel used for shipping multiple units of flow. In this case, the arc parameter represents the maximum amount of flow per unit time that can be shipped along the arc at any time. This parameter is called the *capacity* of the arc, and the network models are referred to as *capacitated flow networks*.

Equipment Replacement.

Problem Definition. It is desired to find an optimal machine replacement policy for an n -year planning horizon under the assumption

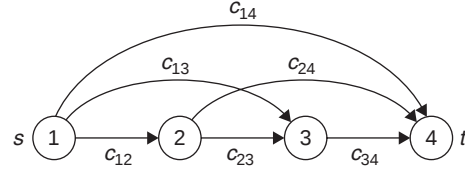


Figure 17. Equipment replacement network.

that the machine can be replaced instead of repaired at the beginning of each year.

Network Representation. Nodes represent beginning of periods (years); arcs represent periods of useful service life (replacement alternatives); arc parameters represent costs of replacement alternatives. This cost is equal to replacement cost plus operating cost minus salvage value. The network consists of $n + 1$ nodes. A chain from $s = 1$ to $t = n + 1$ represents a replacement policy.

As an illustration, we consider a planning period of $n = 3$ years. Here $s = 1$ and $t = 4$. The optimal solution is given by the shortest *chain* from node 1 to node 4. The set of all replacement plans can be represented by the chains in the network in Fig. 17. For example, the chain $(1,3) - (3,4)$ represents the policy of buying a machine at the beginning of period 1, keeping it for 2 years and selling it at the end of period 2 (beginning of period 3). A new machine is then purchased at the beginning of period 3 and sold at the end of the same period. The total cost of this policy is $c_{13} + c_{34}$.

Network Solution. The optimal solution corresponds to the shortest *chain* from node 1 to node t . In the absence of *negative* circuits (one with total length being negative), the shortest chain can be easily found using Dijkstra's algorithm [13].

Project Planning.

Problem Definition. A project is defined as a collection of activities and events. An activity is any undertaking that consumes time and other resources. An event is a well-defined occurrence in time. An *Activity Network* is a representation of two particular aspects of a project: (i) precedence

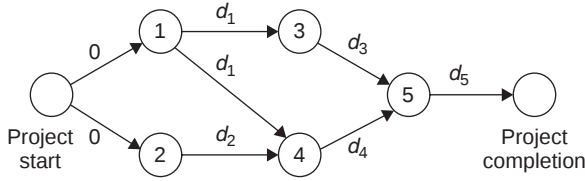


Figure 18. Activity-on node network.

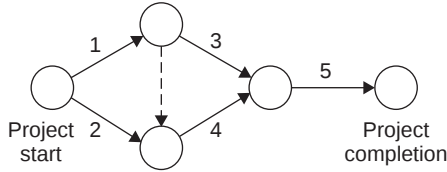


Figure 19. Activity-on arc network.

relationships among the activities, (ii) duration of each activity. The precedence relationships come about from technological and other considerations. The precedence relationships are transitive, nonreflexive, and nonsymmetric.

Network Representation. Let $G(\mathbf{N}, \mathbf{A})$ be an activity network satisfying three conditions: (i) one source, (ii) one terminal, and (iii) at most one arc to directly connect two nodes. There are two representations: **A-on-N** (Activity-on Node) and **A-on-A** (Activity-on Arc). These are illustrated in Figs 18 and 19 for a case involving five activities. In Fig. 19 the dashed line represents a *dummy* activity (duration equal to 0) created in order to avoid both arcs 1 and 2 joining exactly the same two nodes.

Network Solution. For a *node* to be *realized*, all *arcs* ending at the node must be *completed*. Additionally, for an arc to be *started* its *beginning node* must be realized. Under these rules, it can be verified that the *critical* activities and events (those that would delay the project if they are delayed themselves) correspond to the *longest* path from the source to the terminal of the network. In this case finding the longest path is straightforward because the network is *circuitless*. When the network has *positive* circuits (one with total length being positive), the longest path problem becomes very hard to solve.

The Caterer Problem.

Problem Definition. A caterer knows that he will require d_j clean napkins on each of n successive days, $j = 1, 2, \dots, n$. He can meet his needs by purchasing new napkins and by using napkins previously laundered. The laundry has two kinds of service, quick and slow. A napkin sent for slow service is available 2 days later, whereas a napkin sent to the fast service is available 1 day later. New napkins from the store cost a cents each, quick laundry service is c cents per napkin, and slow service b cents per napkin, $b < c$. How does the caterer meet his requirements at minimum cost?

Linear Programming Model: Case 1. Let p_j, s_j, q_j , and h_j be the number of napkins purchased, sent to slow service, sent to quick service, and held over on day j , respectively. The LP model for $n = 3$ is shown in Equations (13)–(16).

$$\text{Minimize } \sum_{j=1}^3 (ap_j + cq_j + bs_j) \quad (13)$$

subject to

$$p_j + s_{j-2} + q_{j-1} \geq d_j \quad j = 1, 2, 3$$

$$\text{(demand for clean napkins)} \quad (14)$$

$$s_j + q_j + h_j - h_{j-1} \leq d_j \quad j = 1, 2, 3$$

$$\text{(supply of soiled napkins)} \quad (15)$$

$$h_j, p_j, s_j, q_j \geq 0 \quad j = 1, 2, 3 \quad (16)$$

Revised Linear Programming Model: Case 2. This revision is done to allow excess purchasing of new napkins. The LP formulation is shown in Equations (17)–(20).

$$\text{Minimize } \sum_{j=1}^3 (ap_j + cq_j + bs_j) \quad (17)$$

subject to

$$p_j + s_{j-2} + q_{j-1} + E_{j-1} - E_j \geq d_j \quad j = 1, 2, 3$$

(demand for clean napkins) (18)

$$s_j + q_j + h_j - h_{j-1} \leq d_j \quad j = 1, 2, 3$$

(supply of soiled napkins) (19)

$$h_j, p_j, s_j, q_j \geq 0 \quad j = 1, 2, 3$$

(20)

Network Representation. Let us consider Case 1 again, assuming that $a_j = a, c_j = c$, and $b_j = b$, for the appropriate values of the index j . Furthermore, let $d_1 = 10, d_2 = 20$, and $d_3 = 5$. In the model shown below, the first three constraints are for clean napkins. The next three constraints are for dirty napkins. In these constraints the net flow is defined as flow out minus flow in. Moreover, the net flow is negative for terminal nodes and positive for source nodes. The seventh constraint is for network circulation, and the last one represents the nonnegativity condition for arc flows.

$$\begin{aligned} \text{Minimize } & a(p_1 + p_2 + p_3) + c(q_1 + q_2) + bs_1 \\ \text{subject to } & -p_1 = -10 \\ & -p_2 - q_1 = -20 \\ & -p_3 - q_2 - s_1 = -5 \\ & s_1 + q_1 + h_1 = 10 \\ & q_2 + h_2 - h_1 = 20 \\ & h_3 - h_2 = 5 \\ & p_1 + p_2 + p_3 - h_3 = 0 \\ & p_1, p_2, p_3, q_1, q_2, s_1, h_1, h_2, h_3 \geq 0 \end{aligned}$$

A network for Case 1 with $n = 3$ is given in Fig. 20, where the variables are indicated adjacent to their respective arcs and the supplies (positive values) and demands (negative values) are given adjacent to the nodes. The unique aspect of this problem is that the demand for clean napkins d_j also works as a supply of dirty napkins at the end of each period.

The corresponding *circulation networks* (a network with only *intermediate* nodes) for

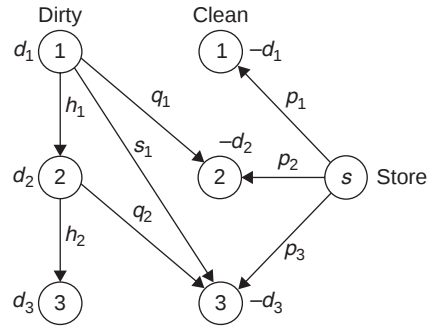


Figure 20. Network for sample caterer problem.

Case 1 and Case 2 are shown in Fig. 21a and b.

In Fig. 21 each arc has a triplet of values $[U_{ij}, L_{ij}, c_{ij}]$. These symbols represent the upper bound on flow, the lower bound on flow, and the per-unit flow cost, respectively. The return arc $(3, s)$ is used to make sure that the total number of soiled napkins at the end of the planning period is equal to the total number of purchased napkins. All arcs not yet considered, should have triplets of values $[\infty, 0, 0]$ assigned to them.

Network Solution. The optimal solution corresponds to the flow resulting in minimum cost. As indicated before, the demand for clean napkins works as a supply of dirty napkins at the end of each period. To accomplish this, each arc in Fig. 21a directed from node j in the column of clean napkins to node j in the column of used napkins has *both* its lower and upper bounds set equal to d_j . Additionally, the per-unit flow cost is set equal to zero. Note that the given network does not allow holding new napkins for future days. Case 2 is a reformulation that allows this to happen by including the arcs in Fig. 21b corresponding to vertical arrows in the column of new napkins.

Employment Scheduling.

Problem Definition. A contract maintenance firm provides and supervises semiskilled manpower for major overhauls of chemical processing equipment. A standard job frequently requires a thousand or more men and may extend from 1 week to a few months.

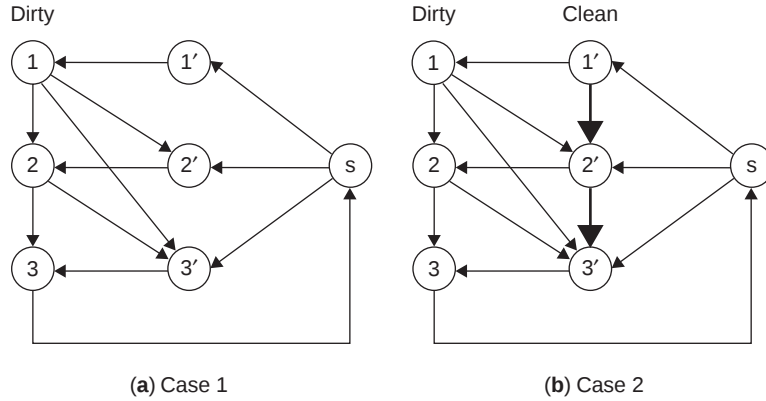


Figure 21. Circulation networks for caterer problem.

Usually, the plant site is not located in a major metropolitan area, so workmen must be transported for several miles. The total labor expenses can be classified as follows: (i) operating expenses = transportation + food + housing + wages; (ii) cost not depending on how long the crew remains on site = recruiting + briefing + transportation. The demand for labor is deterministic but not necessarily uniform.

LP Formulation. Consider a planning horizon consisting of periods $1, 2, \dots, n-1$; let us use the following notation: x_{ij} = number of employees beginning employment at the start of period i and terminated at the start of period j ($1 \leq i \leq j \leq n$), c_{ij} = total cost per employee associated with x_{ij} , D_j = demand for labor in period j , S_j = excess not used in period j . The corresponding *min-cost formulation* is shown in Equations (21)–(24).

$$\text{Minimize } \sum_{r=1}^{n-1} \sum_{t=r+1}^n c_{rt} x_{rt} \quad (21)$$

subject to

$$\sum_{r=1}^j \sum_{t=j+1}^n x_{rt} - S_j = D_j \quad j = 1, 2, \dots, n-1 \quad (22)$$

$$S_j \geq 0 \quad j = 1, 2, \dots, n-1 \quad (23)$$

$$x_{ij} \geq 0 \quad 1 \leq i \leq j \leq n \quad (24)$$

In the above LP formulation, the double summation of constraint (22) represents the number hired by the beginning of period j and terminated after the beginning of period $j+1$.

Network Representation. This problem does not have an obvious network interpretation and it is only after some manipulation of the mathematical model that the network representation becomes evident. The procedure will be illustrated with the following example for $n = 4$. The corresponding constraints are given below:

$$\begin{array}{rcl} x_{12} + x_{13} + x_{14} & -s_1 & = D_1 \\ x_{13} + x_{14} + x_{23} + x_{24} & -s_2 & = D_2 \\ x_{14} & + x_{24} + x_{34} & -s_3 = D_3 \end{array}$$

By subtracting the first equation from the second one, the second equation from the third one and multiplying the third equation by -1 , we obtain

$$\begin{array}{rcl} x_{12} + x_{13} + x_{14} & -s_1 & = D_1 \\ -x_{12} & + x_{23} + x_{24} & + s_1 - s_2 = D_2 - D_1 \\ -x_{13} & -x_{23} & + x_{34} + s_2 - s_3 = D_3 - D_2 \\ -x_{14} & -x_{24} - x_{34} & + s_3 = -D_3. \end{array}$$

An examination of this set of equations reveals that each variable appears exactly twice in the equations, once with a $+1$ coefficient and once with a -1 coefficient. The network for this example is given in Fig. 22a, where the variables are shown next to the arcs and the supplies are shown next to the nodes. In this figure, nodes represent

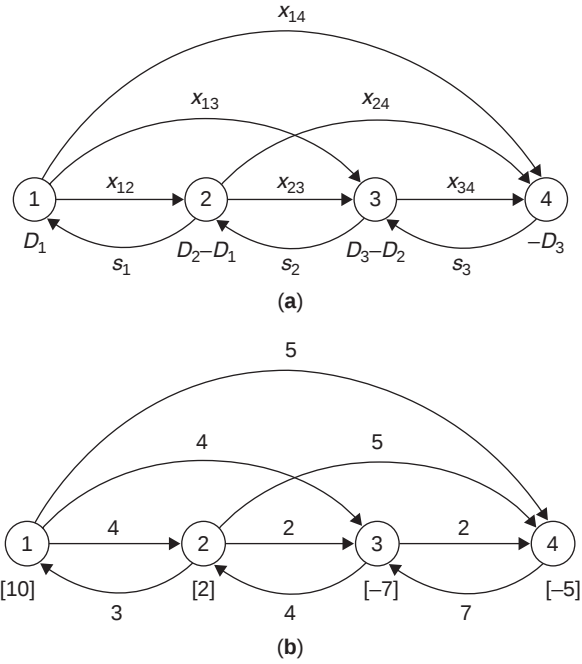


Figure 22. (a) Employment scheduling network. (b) Illustration of employment scheduling network.

Table 1. Interpretation of Network Flows in Fig. 22b

Start of Period	Hired	Terminated	Excess	Working 1 Period	Working 2 Periods	Working 3 Periods	Available
1	13	0	3	4	4	5	13
2	7	4	4	2	5	—	16
3	2	6	7	2	—	—	12
4	—	12	—	—	—	—	—

the beginning of each period. Arcs correspond to all possible employment strategies. Any arc (i, j) such that $j > i$ is used to model the number of workers hired at the beginning of period i and terminated at the beginning of period j . Additionally, any arc (j, i) such that $i < j$ is used to model the excess personnel at the beginning of period i . The parameters for the complete formulation of the problem are node demands and arc costs. The optimal solution corresponds to the *minimum-cost flow* on the network.

An illustration of the model being discussed is considered now. Assume that $D_1 = 10, D_2 = 12, D_3 = 5$. A network with *feasible* flows for the above demand conditions is shown in Fig. 22b. The corresponding

interpretation of flows is summarized in Table 1.

Fleet Allocation

Problem Definition. A problem that often arises is the scheduling and routing of vehicles to accomplish the shipment of some commodity from points of supply to points of demand. There are two important versions of this problem. In Case 1 it is desired to maximize total profit. Case 2 focuses on the minimization of the fleet size. This can be accomplished by minimizing the number of idle ship days. In both cases it is desired to accomplish the stated objective under

the constraints that shipments must be delivered by their due dates.

Network Representation. A timescale is placed below the network. In addition to representing a source and a terminal, nodes represent port/date combinations. One node is defined for each date and place a shipment is delivered, and for each time and place a ship starts loading. Arcs indicate all possible trips with and without cargo. All arcs emanating from the source correspond to the first use of a ship. Additionally, all arcs ending at the terminal represent the removing of ships from service. For Case 1, the numbers shown next to each arc are costs per unit of flow (a unit of flow represents one ship). It is noted that profits are considered here as negative costs. All arcs have infinite flow capacities except those arcs representing shipments, which have capacities equal to 1. In the example shown in Fig. 23 these arcs having unit capacities are (A,C) with A at time 0, (A,C) with A at time 5, (B,D) with B at time 0, and (B,C) with B at time 4. Each chain from the source node to the terminal node represents a feasible schedule for a ship.

Network Solution. The optimal solution corresponds to the distribution of flow (number of ships used) having minimal cost. If a fleet of size M is used, the minimum-cost flow with a total of M units of flow will give the optimal ship schedules and also determine which shipments should be carried. If the fleet size is to be selected to maximize profit (Case 1), the problem can be solved parametrically (setting $M = 1, 2, \dots$) to determine the profit profile as the number of ships increased. For Case 2 the notation used is as follows:

- c_{ij} number of *idle ship days* associated with arc (i,j)
- f_{ij} flow (ships) on arc (i,j)
- L_{ij} minimum number of ships on arc $(i,j) = 1$ for *shipment arcs* and 0 otherwise
- U_{ij} maximum number of ships on arc $(i,j) = 1$ for *shipment arcs* and ∞ otherwise

A convenient way to represent Case 1 and Case 2 is using a *circulation network* with a *return arc* connecting the *terminal* to the

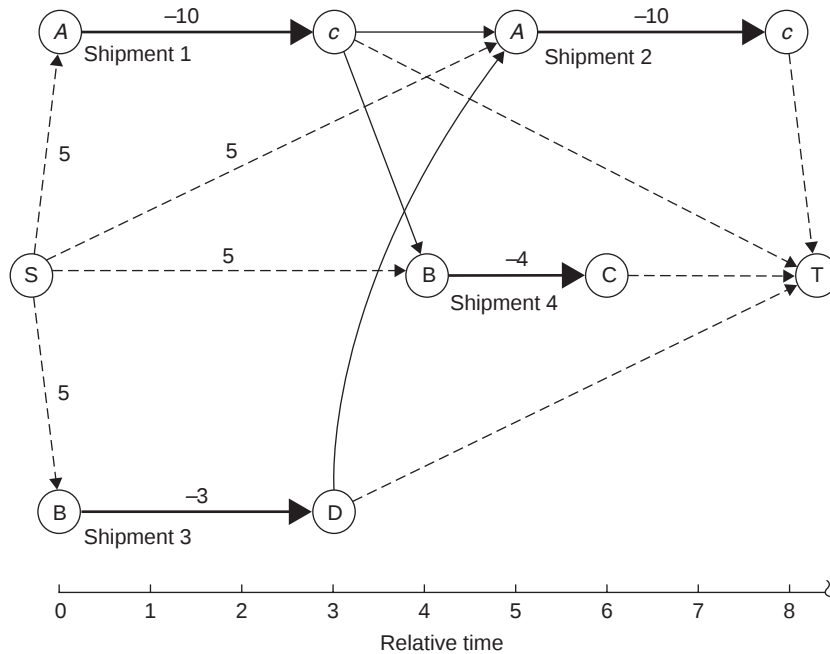


Figure 23. Fleet allocation network.

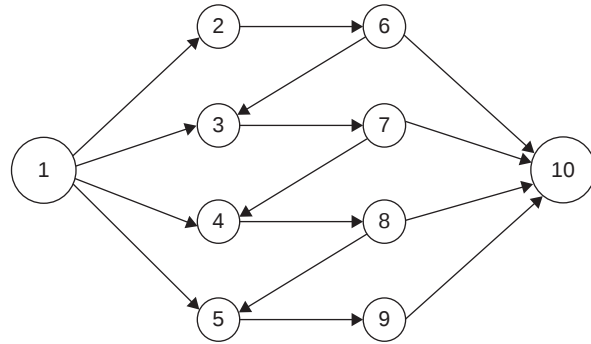


Figure 24. Fuel allocation network.

source. In Case 1, the lower and upper bounds on flow on the return arc are both equal to M and the cost per unit of flow is 0; alternatively, for Case 2, the lower bound is 0 and the upper bound is set equal to ∞ , with per-unit flow cost equal to ∞ .

Aircraft Fuel Allocation.

Problem Definition. Given the route of a single aircraft, it is desired to find the minimum-cost fuel purchasing policy under the existence of price differentials among the cities in the route, as well as burn-off penalties associated with excess (beyond minimum requirements) fuel carried on the aircraft. It is also assumed that fuel availability conditions for the cities, and fuel requirements for the flights are known.

Network Representation. The fundamental difference between this model and the previous model is that the fuel not used is subject to a reduction.

Nodes

Cities

Flights between cities

Arcs

Buying fuel in each city

Carrying fuel in aircraft

Using fuel for flight requirements

Carrying excess fuel after flights

Parameters

Minimum flow

Maximum flow

Per-unit flow cost

Burn-off penalty factor

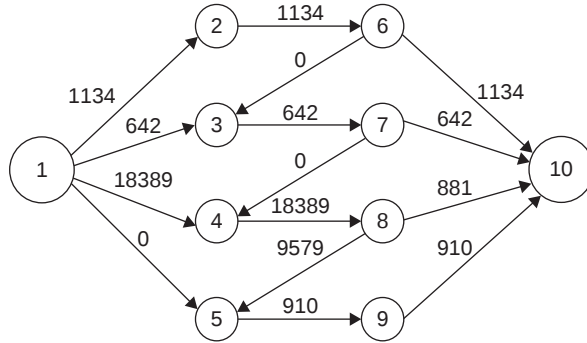
Figure 24 shows a network for a four-city route. Node 1 is the source node. The intermediate nodes represent cities and flights between cities. Nodes 2, 3, 4, and 5 represent the four cities A, B, C, and D, respectively. Node 6 represents the flight from city A to city B. Node 7 represents the flight from city B to city C. Node 8 represents the flight from city C to city D. Node 9 represents the flight from city D to city A. Node 10 is the terminal node. Furthermore, the arcs are classified into four categories: (i) arcs coming out of the source represent buying fuel in each city, (ii) arcs (2,6), (3,7), (4,8), and (5,9) represent carrying fuel in the aircraft, (iii) arcs (6,3), (7,4), and (8,5) represent carrying excess fuel after flights, (iv) arcs into the terminal represent the use of fuel for flight requirements.

The arc parameters for any arc are: minimum flow, maximum flow, per-unit flow cost, and the burn-off penalty factor. These parameters can be listed for arc (i,j) as $[L_{ij}, U_{ij}, c_{ij}, A_{ij}]$. The arcs associated with excess fuel left at the end of a flight are generalized because of the burn-off losses. The arcs representing the fuel carried in the aircraft tank can be eliminated, if desired, since the fuel in the tank can be calculated by subtracting the amount used from the total available before a flight. An illustration of the model shown in Fig. 24 follows. Table 2 summarizes the data for this example.

Network Solution. This problem can be formulated as a *generalized minimum-cost*

Table 2. Fuel Allocation Example

Purchasing Strategies	Fuel Carrying Strategies	Excess Fuel Strategies	Requirements Strategies
(1,2) [0,1500,0.48,1]	(2,6) [0,2000,0,1]	(6,3) [0,2000,0,0.93]	(6,10) [1134,1134,0,1]
(1,3) [0,1000,0.49,1]	(3,7) [0,2000,0,1]	(7,4) [0,2000,0,0.96]	(7,10) [642,642,0,1]
(1,4) [0,2000,0.46,1]	(4,8) [0,2000,0,1]	(8,5) [0,2000,0,0.95]	(8,10) [881,881,0,1]
(1,5) [0,1000,0.75,1]	(5,9) [0,2000,0,1]	—	(9,10) [910,910,0,1]


Figure 25. Optimal flows for fuel allocation network.

network problem whose LP model was shown in Equations (10)–(12). In this generalized network with losses, flow arriving at the end of an arc is less than the flow entering at the beginning of an arc. This class of network lends itself to a simplified and powerful representation of the *burn-off* penalties associated with excess fuel carried on an aircraft.

The optimal flows for the numerical example are shown in Fig. 25. The total cost of the fuel purchasing policy shown is \$1704.8. More details on this application are available in the article by Garcia-Diaz [15].

LINEAR PROGRAMMING AND NETWORKS

Totally Unimodular Matrix

A *unimodular* matrix is a square, integer matrix with determinant equal to 0, +1, or −1. An integer matrix (not necessarily square) is *totally unimodular* if every square, nonsingular submatrix of it is unimodular. The following statements are equivalent:

1. $\mathbf{A}$ is totally unimodular
2. $\mathbf{A}^T$ is totally unimodular
3. $(\mathbf{A}, \mathbf{I})$ is totally unimodular
4. A matrix obtained by deleting a row or column of $\mathbf{A}$ is totally unimodular

5. A matrix obtained by multiplying a row or column of $\mathbf{A}$ by -1 is totally unimodular
6. A matrix obtained by interchanging two rows or two columns of $\mathbf{A}$ is totally unimodular
7. A matrix obtained by duplicating rows or columns of $\mathbf{A}$ is totally unimodular
8. A matrix obtained by a *pivot operation* on $\mathbf{A}$ is totally unimodular.

A pivot operation is exactly the same operation in the Gaussian elimination technique resulting in a *column* with an entry equal to 1 in the position of the pivot and all other entries equal to 0.

Fundamental Results

The following theorems are stated without proof:

Theorem 1. If a matrix $\mathbf{D}$ is totally unimodular, then every basic solution of $\mathbf{D}\mathbf{f} = \mathbf{b}$, $\mathbf{f} \geq \mathbf{0}$, where $\mathbf{b}$ is an integer vector, is integer.

Theorem 2. $\mathbf{D}$ is totally unimodular if and only if $\mathbf{D}\mathbf{f} \leq \mathbf{b}$, $\mathbf{f} \geq \mathbf{0}$, where $\mathbf{b}$ is an integer vector, has all basic solutions integer.

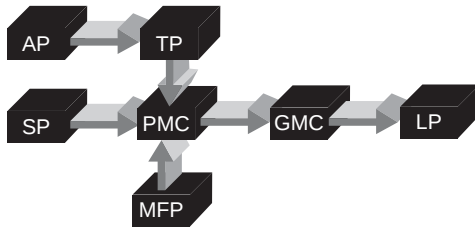


Figure 26. Network models as special cases of LP model.

Theorem 3 [12]. An integer matrix $\mathbf{A} = [a_{ij}]$ with $a_{ij} = 0, 1, -1$, for all i and all j , is totally unimodular if: (i) No more than two nonzero elements appear in each column. (ii) The rows can be partitioned into two sets $\mathbf{E}_1$ and $\mathbf{E}_2$ such that (a) if a column contains two nonzero values with the same sign, one value is in each of the sets; (b) if a column contains two nonzero elements of opposite sign, both elements are in the same set.

The pure minimum-cost flow model formulated in Equations (7)–(9) can be reformulated using matrix notation as minimizing $\mathbf{Cf}$ subject to $\mathbf{Df} \leq \mathbf{b}$ and $\mathbf{f} \geq \mathbf{0}$, where $\mathbf{b}$ is an integer vector and $\mathbf{D}$ consists of matrices $\mathbf{Z}$ and $\mathbf{I}$. Since $\mathbf{Z}$ is totally unimodular (by Theorem 3), then $\mathbf{D}$ is totally unimodular. For this reason, optimal flows are integer when $\mathbf{b}$ is integer. However, in the case of a generalized network, the LP model formulated in Equations (10)–(12) does not generally satisfy the total unimodularity condition. Therefore, flows are no longer guaranteed to be integer even when the right-hand sides of the constraints are integer.

Special Cases

Now we focus on several well-known network problems that can be formulated as special cases of the model formulated in Equations (7)–(9). The problems are known as follows:

1. Transportation Problem (TP)
2. Assignment Problem (AP)
3. Maximum-Flow Problem (MFP)
4. Shortest-Path Problem (SPP)

Figure 26 illustrates the relationships among these problems. In this figure the problem mentioned at the *beginning* of an arrow is a *special case* of the problem shown at the *end* of the same arrow.

REFERENCES

1. Euler L. Solutio problematis Ad geometriam situs pertinentis [The solution of a problem relating to the geometry of position]. In: Translated into English, Biggs NL, Lloyd EK, Wilson RJ, editors. Graph theory: 1736–1936. Oxford: Clarendon Press; 1976. pp. 3–11.
2. Biggs NL, Lloyd EK, Wilson RJ. Graph theory: 1736–1936. Oxford: Clarendon Press; 1976.
3. Newman JR. The world of mathematics. New York: Simon and Schuster; 1956.
4. Scientific American. Königsberg Bridges 1953; 189:66–70.
5. Orden A. The transshipment problem. Manage Sci 1956;2(3):276–285.
6. Dantzig G. Application of the simplex method to a transportation problem. In: Koopmans TC, editor. Activity analysis of production and allocation. New York: John Wiley & Sons, Inc.; 1951.
7. Johnson EL. Networks and basic solutions. Oper Res 1966;14:619–623.
8. Kuhn HW. The hungarian method for the assignment problem. Nav Res Log Q 1955;2: 83–97.
9. Fulkerson DR. An out-of-kilter method for solving minimum flow cost problems. SIAM J Appl Math 1961;9:18–27.
10. Bertsekas DP. Linear network optimization. Cambridge (MA): The MIT Press; 1991.
11. Bertsekas DP. Network optimization: continuous and discrete models. Nashua, NH; Athena Scientific; 1998.
12. Heller L, Tomkins CB. An extension of a theorem of dantzig's. In: Kuhn H, Tucker AW, editors. Linear inequalities and related systems. Princeton (NJ): Princeton University Press; 1956.
13. Phillips DT, Garcia-Diaz A. Fundamentals of network analysis. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1981.
14. Dantzig G, Fulkerson DR. On the min-cut max-flow theorem of networks. Linear inequalities and related systems. Princeton (NJ): Princeton University Press; 1956.

15. Garcia-Diaz A. A network approach to airline fuel allocation problems. *Ann Soc Log Eng* 1990;2(1):39–53.

SUGGESTED READING

Ahuja RK, Magnanti TL, Orlin JB. *Network flows: theory, algorithms, and applications*. Englewood Cliffs (NJ): Prentice-Hall; 1993.

Bazarrar M, Jarvis JJ. *Linear programming and network flows*. New York: John Wiley & Sons, Inc.; 1978.

Bradley GH. A survey of deterministic networks. *AIIE Trans* 1975;7:222–234.

Busacker RG, Saaty T. *Finite graphs and networks*. New York: McGraw-Hill Book Company; 1965.

Charnes A, Cooper WW. Volumes 1 and 2, *Management models and industrial applications of linear programming*. New York: John Wiley & Sons, Inc.; 1961.

Elmaghraby S. *Some network models in management science*. New York: Springer; 1970.

Evans JR, Minieka E. *Optimization algorithms for networks and graphs*. 2nd ed. New York: Marcel Dekker, Inc.; 1992.

Ford LR, Fulkerson DR. *Flows in networks*. Princeton (NJ): Princeton University Press; 1962.

Frank H, Frisch IT. *Communication, transmission, and transportation networks*. Reading (MA): Addison-Wesley Publishing Co., Inc.; 1971.

Glover F, Klingman D, Phillips NV. *Network models in optimization and their applications in practice*. New York: John Wiley & Sons, Inc.; 1992.

Hitchcock FL. The distribution of a product from several sources to numerous localities. *J Math Phys* 1941;20:224–230.

Hu TC. *Integer programming and network flows*. Reading (MA): AddisonWesley Publishing Co., Inc.; 1969.

Iri M. *Network flow, transportation and scheduling*. New York: Academic Press, Inc.; 1969.

Jensen PA, Barnes W. *Network flow programming*. New York: John Wiley & Sons, Inc.; 1980.

Koopmans TC. Optimum utilization of the transportation system. In: *Proceedings of the International Statistical Conference*. Washington (DC); 1947.

Magnanti TL, Golden BL. *Transportation planning: network models and their implementation*. Working Paper 77-008. University of Maryland, General Research Board, Faculty Research Award; 1977.

Minieka E. *Optimization algorithms for networks and graphs*. New York: Marcel Dekker, Inc.; 1978.

Pritsker AAB, Happ WW. GERT: PART I—fundamentals. *J Ind Eng* 1966;17(5):267.

Whitehouse GE. *Systems analysis and design using network techniques*. Englewood Cliffs (NJ): Prentice-Hall, Inc.; 1973.

NETWORK-BASED DATA MINING: OPERATIONS RESEARCH TECHNIQUES AND APPLICATIONS

VLADIMIR BOGINSKI

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

Research and Engineering Education
Facility (UF-REEF), University of
Florida, Shalimar, Florida

INTRODUCTION

The process of studying real-life complex systems often deals with large data sets arising in diverse applications including government and military systems, telecommunications, biotechnology, medicine, finance, astrophysics, ecology, geographical information systems, etc. [1,2]. Understanding the structural properties of a certain data set is in many cases the task of crucial importance. To extract useful information from the data, one often needs to apply advanced mathematical modeling techniques to summarize, visualize, and process the information contained in a data set. *Data mining* problems deal with multiple aspects of information retrieval from data sets, and a variety of techniques has been developed to address different types of tasks arising in this area [3].

During the past several years, *operations research (OR)* / *mathematical programming approaches* have received increased attention in the context of data mining problems [4]. Along with classical data mining techniques (e.g., support vector machines, artificial neural networks, decision trees, and K-means/K-median clustering), many recent studies deal with *network (graph)-based representations* of large real-world data sets and utilize graph-theoretic and network optimization concepts to extract useful and nontrivial information from these data sets. In this approach, a data set is represented

as a *graph (network)* with certain attributes associated with its vertices and edges.

Studying the structure of a graph representing a data set is often important for understanding the internal properties of the application it represents, as well as for improving storage organization and information retrieval. One can visualize a graph as a set of dots and links connecting them, which often makes this representation convenient and easily understandable.

The main concepts of graph theory were founded several centuries ago, and many network optimization algorithms have been developed since then. However, the trend of applying graph models to represent various real-life massive data sets has started relatively recently. Graph theory and related OR techniques are quickly becoming a field with a strong practical impact. The expansion of graph-theoretical approaches in various applications gave birth to the terms *graph practice* and *graph engineering* [5]. Graph (network)-based data mining techniques have gained significant popularity in recent years. Various aspects of this emerging field are addressed in Cook and Holder [6] and Washio and Motoda [7]. The term *network science* referring to a broad variety of network-based modeling approaches has also been widely used recently. In Alderson [8], the potential relevance between “network science” and OR is discussed. Although potential applications of OR in this broad field can be diverse, network-based data mining is among the areas where OR-based techniques can provide valuable contributions.

Network-based models allow one to extract information from real-world data sets using various standard concepts from graph theory. In many cases, one can investigate specific properties of a data set by detecting special formations in the corresponding graph, for instance, *connected components*, *spanning trees*, *cliques*, and *independent sets*. In particular, cliques and independent sets can be used for solving the important *clustering* problem arising in data mining,

which essentially represents partitioning the set of elements of a certain data set into a number of subsets (clusters) of objects according to some *similarity* (or *dissimilarity*) criterion. These concepts are associated with a number of network optimization problems that are discussed below.

Another aspect of investigating network models of real-world data sets is studying the global organization and structural properties of these networks, which are reflected by *degree distributions* of the constructed graphs. The degree distribution represents the large-scale pattern of connections in the graph, which reflects the global properties of the data set. One of the important results discovered during the last several years is the observation that many graphs representing the data sets from diverse areas (Internet, telecommunications, biology, and sociology) obey the *power-law* model [9,45]. The fact that graphs representing completely different data sets have a similar well-defined power-law structure has been widely reflected in the literature [2,5,10–13]. It indicates that global organization and evolution of data sets arising in various spheres of life follow similar laws and patterns. This fact served as a motivation to introduce the concept of “self-organized networks.” The information about the degree distribution of a graph that represents a certain data set can be significant in the data mining context, since it may allow one to analytically predict the size of certain types of clusters in this graph (e.g., connected components, cliques, and other structures). In particular, the size of connected components (clusters) in very large power-law graphs has been studied in Aiello *et al.* [9].

Further in this article, we address the main issues and concepts arising in network-based data mining. Specifically, we will describe the process of representing a data set as a graph using appropriate similarity (proximity) measures, which serves as a basis for applying graph-theoretic and network optimization techniques. These concepts and techniques are discussed later in the article. Finally, we give several examples of popular application areas of network-based data mining.

SIMILARITY MEASURES USED IN GRAPH-BASED DATA MINING

As indicated above, the essence of network-based data mining is representing a data set under consideration as a network (graph) where vertices and edges are constructed according to certain principles that may depend on the structure and origin of the data set. The vertices of the graph usually represent data records (objects), whereas the edges that connect the vertices are constructed using an appropriate *similarity criterion*. That is, two data objects (vertices) are connected if they are “similar enough” to each other. However, there are a lot of different ways to quantify the degree of similarity. In this section, we specifically address possible types of *similarity measures* (also referred to as *proximity measures*) between data objects that can be used for constructing the corresponding graph representations of data sets.

In a standard setup of many data mining problems, one deals with a data set of N data records that can be represented by vectors $x^i = (x_1^i, x_2^i, \dots, x_n^i)$, $i = 1, \dots, N$, where every vector x^i has n components that are referred to as the *features*, or *attributes*, of the data record. The similarity measure between any two data records would be defined by the corresponding values of their attributes. The following similarity measures are commonly used in practice.

- *Distance Metrics.* These can generally be expressed in the form

$$d(x, y) = \left(\sum_{j=k}^n |x_k - y_k|^r \right)^{(1/r)},$$

where r is a parameter that defines different types of metrics. The most commonly used distance metrics are L_1 norm for $r = 1$ (Manhattan distance), L_2 norm for $r = 2$ (Euclidian distance), and L_∞ norm for $r = \infty$ (supremum distance).

- *Cosine Similarity.* This measure is often used to characterize text document similarity (e.g., documents represented as vectors with attributes denoting the

frequency of each word in the considered document) and can be expressed as

$$\cos(x, y) = \frac{xy}{||x|| ||y||},$$

where xy is a scalar product of vectors x and y .

- *Correlation.* The standard correlation measure is commonly used to characterize the similarity between different types of data, for instance, the time series data, where the attributes represent the values of a certain parameter for different time moments. A well-known example where the correlation measure can be effectively used is the analysis of stock market data, where high correlation would mean high degree of similarity between a given pair of stocks.

Other definitions of similarity (proximity) measures are also used in practice [3]. For instance, simple matching coefficient, Jaccard coefficient, and extended Jaccard coefficient can be efficiently used to measure the similarity between binary data objects. A generalization of Euclidian distance referred to as *Mahalanobis distance* can be applied to measure similarity when data attributes are correlated. Moreover, besides defining the similarity between data objects, in some cases, it is also needed to quantify the *dissimilarity*. The aforementioned proximity measures can be also used to define the pairwise dissimilarity; therefore, all further discussion is applicable to clustering applications based on either similarity or dissimilarity. Without loss of generality, we will mostly use the term “similarity” throughout the article.

Determining an appropriate similarity measure is often a nontrivial task; however, if a suitable similarity measure is determined for a given data set, this gives one the information about pairwise proximities for all pairs of data objects. On the basis of this information, one can easily construct the $N \times N$ proximity matrix and then use the proximity matrix to create a graph with N vertices and $N(N - 1)/2$ weighted edges,

where the weights $w_{ij} = d(x^i, x^j)$ are equal to the corresponding proximity measures between data objects x^i and x^j . Note that this simple procedure would produce a complete weighted graph (with all possible edges), which can be challenging to analyze from both theoretical and computational perspectives. To reduce the edge density of the obtained graph, one can use an appropriate *sparsification* procedure that can substantially reduce the number of edges in the graph while still preserving valuable information about the data set. To achieve this, the following simple principles can be used:

- *k-Nearest Neighbor Approach.* For every vertex i , keep only the edges that connect it to its k nearest neighbors, that is, preserve k edges (i, j) with the smallest values of w_{ij} (as determined by an appropriate proximity measure) and delete the remaining edges.
- *Threshold Approach.* For every edge (i, j) , delete it if the corresponding similarity measure w_{ij} is below (or above) a specified threshold.

Applying one or both of these approaches can, in many cases, eliminate most of the edges in the constructed graph and make sure that the remaining edges preserve the most meaningful similarity relationships between the data objects in the considered data set. It should be noted that the similarity measures between each pair of vertices can be further updated to incorporate the idea of *shared nearest neighbors (SNN)*. That is, if two vertices i and j are both connected to certain number of other vertices, the weight of the edge (i, j) is increased by the number of these shared neighbors. This idea can be used for constructing the *SNN similarity graph*. In particular the concept of SNN graph has been used in the Jarvis–Patrick clustering algorithm, which essentially performs clustering on the considered data set by identifying connected components in the corresponding SNN graph [14].

In general, after the graph representing a data set has been constructed and sparsified, one can use a variety of graph-theoretic concepts and techniques to extract useful

information about the considered data set and divide the data set into a number of meaningful clusters according to the specified similarity criterion. In the next section, we concentrate on relevant concepts, definitions, and optimization problems arising in this analysis.

BASIC CONCEPTS FROM GRAPH THEORY AND THEIR DATA MINING INTERPRETATION

To facilitate further discussion, we present several formal basic definitions and notations from graph theory and discuss the interpretation of the introduced concepts from the perspective of data mining and information retrieval.

Let $G = (V, E)$ be an undirected graph with the set of N vertices V and the set of edges $E = \{(i, j) : i, j \in V\}$. Directed graphs, where the head and tail of each edge are specified, are considered in some applications. The concept of a *multigraph* is also sometimes used. A multigraph is a graph where multiple edges connecting a given pair of vertices may exist. One of the important characteristics of a graph is its *edge density*: the ratio of the number of edges in the graph to the maximum possible number of edges. As indicated in the previous section, one often needs to eliminate a substantial number of edges from a graph representing a real-life data set, so that the edge density is reduced and only the most meaningful connections are preserved.

Connected Components and Degree Distributions

The graph $G = (V, E)$ is *connected* if there is a path from any vertex to any vertex in the set V . If the graph is disconnected, it can be decomposed into several connected subgraphs, which are referred to as the *connected components* of G . In many clustering techniques, distinct connected components are interpreted as distinct *clusters* in the corresponding data set. It should be noted that although using connected components for clustering is widely used in a variety of applications, there are alternative graph-theoretic

concepts that can be utilized in clustering problems. These concepts are discussed in further sections.

The *degree* of a vertex is the number of edges emanating from it. For every integer k , one can calculate the number of vertices $n(k)$ with a degree equal to k , and then get the estimate of the probability that a vertex has the degree k as $P(k) = n(k)/n$, where n is the total number of vertices. The function $P(k)$ is referred to as the *degree distribution* of the graph. In the case of a directed graph, the concept of degree distribution is generalized: one can distinguish the distribution of *in-degrees* and *out-degrees*, which deal with the number of edges ending at and starting from a vertex, respectively.

Degree distribution is an important characteristic of a data set represented by a graph. It reflects the overall pattern of connections in the graph, which in many cases reflects the global properties of the data set that this graph represents. As mentioned above, many real-world graphs representing the data sets coming from diverse areas (Internet, telecommunications, finance, biology, and sociology) have degree distributions that follow the *power-law* model, which states that $P(k) \propto k^{-\gamma}$. Equivalently, one can represent it as $\log P \propto -\gamma \log k$, which demonstrates that this distribution forms a straight line in the logarithmic scale, and the slope of this line equals the value of the parameter γ .

An important characteristic of the power-law model is its *scale-free* property. This property implies that the power-law structure of a certain network should not depend on the size of the network. Clearly, real-world networks dynamically grow over time, therefore, the growth process of these networks should obey certain rules in order to satisfy the scale-free property. The necessary properties of the evolution of the real-world networks are *growth* and *preferential attachment* [12]. The first property implies the obvious fact that the size of these networks grows continuously (i.e., new vertices are added to a network, which means that new elements are added to the corresponding data set). The second property represents the idea that new vertices are more likely to be connected to old vertices with high degrees. It is intuitively

clear that these principles characterize the evolution of many real-world complex networks and massive data sets they represent.

From another perspective, some properties of graphs that follow the power-law model (and the corresponding data sets) can be predicted theoretically. Interesting properties of power-law graphs were studied using the theoretical *power-law random graph model* [9,13]. Among these results, one can mention the existence of a giant connected component (i.e., a giant cluster that contains $\Theta(N)$ vertices) in a power-law graph with $\gamma < \gamma_0 \approx 3.47875$, and the fact that all connected components (clusters) are small ($o(N)$ vertices) otherwise¹. The emergence of a giant connected component (or, a giant connected cluster) at the point $\gamma_0 \approx 3.47875$ is often referred to as a *phase transition*.

The size of connected components of the graph may provide useful information about the structure of the corresponding data set, since, as indicated above, the connected components would normally represent clusters of “similar” objects. In many applications, decomposing the graph into a set of connected components can provide a reasonable solution to the clustering problem. From the computational perspective, identifying all connected components in a graph is a *polynomially solvable* problem, which is especially important when the considered graphs and data sets are very large.

However, in some situations, clusters based on connected components can be extremely large (e.g., comparable with the size of the whole graph, as indicated above), and/or they may not be “tight enough,” that is, the degree of connectivity within a cluster may not be sufficient to make reliable conclusions regarding the similarity of the data objects. This motivates one to look for substantially more “robust” clusters that have specific properties of their connectivity patterns. The corresponding graph structures that address these issues are discussed in the following sections.

¹These results are valid *asymptotically almost surely* (a.a.s.), which means that the probability that a given property takes place tends to 1 as the number of vertices N goes to infinity.

Cliques and Independent Sets

Given a subset $S \subseteq V$, we denote $G(S)$ as the subgraph induced by S . A subset $C \subseteq V$ is a *clique* if $G(C)$ is a complete graph (i.e., it has all possible edges). One of the most well-known problems in combinatorial optimization, the maximum clique problem, is to find the largest clique in a graph [46].

An *independent set* is a subset $I \subseteq V$ such that the subgraph $G(I)$ has no edges. The maximum independent set problem can be easily reformulated as the maximum clique problem in the *complementary* graph $\overline{G}(V, \overline{E})$, defined as follows. If an edge $(i, j) \in E$, then $(i, j) \notin \overline{E}$; and if $(i, j) \notin E$, then $(i, j) \in \overline{E}$. A maximum clique in $\overline{G}$ is a maximum independent set in G , so the maximum clique and maximum independent set problems can be easily transformed to each other and are essentially equivalent.

Clearly, locating cliques and/or independent sets in a graph representing a data set provides important information in terms of identifying completely connected (or, on the contrary, completely disconnected) clusters. Therefore, cliques would naturally represent very *dense* clusters of similar objects. On the contrary, independent sets can be treated as groups of objects that differ from every other object in the group. This information may be important in some applications. In particular, it is often useful to find a *maximum clique or independent set* in the graph, since it would give the maximum possible size of the groups of “similar” or “different” objects.

Although finding large cliques in a graph representing a data set may be useful in terms of identifying large tightly connected clusters, an important computational disadvantage of this approach is that the maximum clique problem (as well as the maximum independent set problem) is known to be *NP-hard* [15]. Moreover, it turns out that these problems are difficult to approximate [16,17]. This makes these problems especially challenging in large graphs.

Clustering via Clique Partitioning

The problem of locating cliques and independent sets in a graph can be naturally extended to finding an optimal *partition*

of a graph into a minimum number of distinct cliques or independent sets. These problems are referred to as *minimum clique partition* and *graph coloring*, respectively. Pardalos *et al.* [18] give various mathematical programming formulations of these problems. Clearly, as in the case of maximum clique and maximum independent set problems, minimum clique partition and graph coloring are reduced to each other by considering the complimentary graph, and both of these problems are NP-hard [15]. Solving these problems for graphs representing real-life data sets is important from a data mining perspective; especially for addressing the clustering problem.

Since the data elements assigned to the same cluster should be similar to each other, the goal of clustering is achieved by finding a clique partition of the graph, and the number of clusters will equal the number of cliques in the partition.

Similar arguments hold for the case of the graph coloring problem which should be solved when a data set needs to be decomposed into the clusters of “different” objects (i.e., each object in a cluster is different from all other objects in the same cluster), that can be represented as independent sets in the corresponding graph. The number of independent sets in the optimal partition is referred to as the *chromatic number* of the graph.

Using Clique Relaxations for Graph-Based Clustering

Although cliques and independent sets address the issue of “robust” tightly connected clusters, a significant practical drawback of these structures is that they are often *too restrictive*, that is, all pairs of vertices need to be connected (disconnected). On the contrary, as mentioned above, a simple connectivity requirement may not be restrictive enough to guarantee sufficient tightness of a cluster. In an attempt to provide a trade-off between the clique-based and sparse connected component-based clusters, the concepts of *clique relaxations* have been introduced.

For instance, instead of cliques and independent sets, one can consider *quasi-cliques*

and *quasi-independent sets*, and partition the graph on this basis. Quasi-cliques are subgraphs that are *dense enough* (i.e., they have a sufficiently high edge density, but the edge density does not have to be one as in the case of cliques). Therefore, it is often reasonable to relate clusters to quasi-cliques, since they still represent sufficiently dense clusters of similar objects [19,20]. Obviously, in the case of partitioning a data set into clusters of “different” objects, one can use quasi-independent sets (i.e., subgraphs that are sparse enough) to define these clusters.

Other types of clique relaxations have also been introduced. Many of these definitions come from the studies of social networks; however, these concepts clearly have important data mining interpretations. The main idea behind these concepts is to “relax” certain properties of a clique while still maintaining sufficient connectivity and robustness characteristics of the obtained network structures. Note that in many cases the size of these clique relaxations is substantially larger than the size of cliques, which provides a significant advantage in situations when a *large tightly connected cluster* needs to be identified. In addition, clusters defined by cliques are sensitive to potential errors in data sets, since one missing edge (that may be due to a collection/measurement error) violates the whole clique structure, whereas clusters defined by clique relaxations may help avoid this drawback.

More specifically, there are three main directions for possible relaxations of the clique definition as follows:

1. *Density-Based Relaxations*: relaxing the requirement for the edge density of a clique to be 1: *quasi-cliques* (γ - dense subgraphs) [21].
2. *Degree-Based Relaxations*: relaxing the requirement that the degree of each node in a clique of size q is $q - 1$: *k-plexes* [22].
3. *Path (Diameter)-Based Relaxations*: relaxing the requirement that the length of the path between any two nodes in a clique is 1: *k-cliques* [23] and *k-clubs* [24].

A *quasi-clique* (also referred to as a γ -dense subgraph) is a subgraph that has the edge density of at least γ , where $\gamma \in (0, 1]$. Clearly, a quasi-clique becomes a clique if $\gamma = 1$. A *k-plex* is a subgraph in which the degree of each node is at least $q - k$ (assuming that q is the number of nodes in this subgraph). A *k-clique* is a subgraph where the length of the path between any two nodes is at most k (note that other nodes in this path are *not required* to belong to the *k-clique*), whereas a *k-club* is a subgraph that has a *diameter* of at most k (note that in this definition all the nodes in the paths that connect any pair of nodes within a *k-club* should also belong to this *k-club*). Obviously, for $k = 1$, a *k-plex*, a *k-clique*, or a *k-club* would also be a clique.

Although these definitions are rather straightforward and intuitively clear, mathematically rigorous studies on related optimization aspects (e.g., mathematical programming formulations for finding the largest clique relaxations in graphs) have started to appear only within the past few years. The most compact known linear 0-1 formulation for the maximum quasi-clique problem with $O(N)$ variables and constraints has been developed by Veremyev *et al.* [25]. The maximum *k-plex* problem has been addressed by Balasundaram *et al.* [26], where they provide the most compact formulation with N variables and constraints. The maximum *k-clique* and *k-club* problems and their extensions are discussed in Balasundaram *et al.* [27] and Veremyev and Boginski [28].

Utilizing clique relaxations in the context of clustering in data mining appears to be a promising approach with attractive characteristics, since it provides sufficient flexibility in terms of the tightness and size of the obtained clusters.

EXAMPLES OF REAL-WORLD APPLICATIONS

The aforementioned network-based data mining techniques have been used in a variety of real-world applications during recent years. Although the length of this article does not allow one to describe all the applications in detail, we briefly mention

several types of networks and data sets that have been addressed in the recent literature.

Biological Networks

Nowadays, representing complex biological systems as networks and studying the properties of these networks has become a very popular tool in computational biology and biomedical applications. For instance, an overview of graph-based molecular data mining approaches is presented in Fischer and Meinl [29]. A subject of particular research interest are protein-protein interaction networks that have been experimentally constructed for many known organisms. Many biological functions involve interactions between proteins at different levels, including signal transduction in cells (i.e., conversion of one kind of a signal to another inside a cell, which may play an important role in biological processes, including disease development), formation of protein complexes (i.e., stable over time structures involving multiple proteins), brief interactions between proteins involving the processes of modification of one protein by another, etc. For instance, the protein interaction network of *Saccharomyces cerevisiae*, as well as many other networks of this type, has been experimentally showed to have a *power-law structure*. Network-based clustering techniques utilizing clique relaxations (e.g., *k-clubs*) have been applied to studying these networks [27,30]. Gene expression data is also often represented in a network format, and quasi-clique-based clusters can be identified in these networks [31]. Moreover, brain dynamics data (e.g., EEG recordings from multiple electrodes located at different brain sites) can be naturally represented as a network. For instance, in Prokopyev *et al.* [32], a network-based approach of modeling brain data and identifying connected components and cliques in the considered graphs was applied to enhance the capability to identify possible epileptic seizures based on EEG time series recordings (appropriate statistical measures were used to define the similarity between EEG time series). Overall, network-based data mining approaches are receiving increased attention in biology and biomedicine.

Telecommunication/Information Exchange Networks

Large-scale information exchange networks arise in a great variety of areas, including large-scale wireless communication/sensor networks, the Internet/World Wide Web, telephone traffic networks, etc. The Internet, the World-Wide Web, and the telephone call networks (also referred to as *call graphs*) were experimentally shown to have a power-law structure. Studying the structural properties and the global organization of these networks has been addressed in recent literature. Graph theory has been applied for web search [33,34], web mining [35,36], and other problems arising in the Internet and World Wide Web. In an interesting experimental study of the world wide web, Broder *et al.* [37] used two Altavista crawls, each with about 200 million pages and 1.5 billion links and identified certain properties of the connected components of this graph. Studying the structural properties of the Internet is an exciting area of ongoing research [38]. The overall structural characteristics of the Internet are sometimes referred to as *robust yet fragile* [39], which implies the overall robustness with respect to random disruptions/errors, but potential vulnerability to targeted attacks on “hubs” (i.e., high-degree vertices).

Another well-known experimental study of a large communication network was conducted on the “call graph” representing daily telephone traffic data from AT&T telephone billing records [40]. The call graph is also very large, and the considered instance contained 53,767,087 vertices and over 170 million edges. The size of cliques and quasi-cliques in this graph were investigated, and it turned out that the size of the largest clique did not exceed 32, and the size of the largest quasi-clique with $\gamma \geq 0.5$ did not exceed 96. Since the size of the considered graphs is extremely large, the problems of identifying tight clusters, such as cliques and quasi-cliques, cannot be solved exactly due to computational challenges. Therefore, appropriate heuristic algorithms need to be applied to tackle these problems. For instance, in Abello *et al.* [40], the greedy

randomized adaptive search procedure (GRASP) [41,42] was successfully used.

Financial Networks

Although not as naturally as in the case of social and communication network data, stock market data can also be represented as a network. Let every traded stock be represented by a vertex in a network, and a given pair of vertices will be connected by an edge if the corresponding stocks exhibit a similar behavior over a certain period (this similarity can be measured in terms of correlation between the corresponding time series, and a relatively high required threshold can be set, so that an edge would be in place if this threshold is exceeded). Boginski *et al.* have previously done extensive work on studying the behavior of the entire US stock market and modeling this complex system as a large-scale network (referred to as the *market graph*) [43,44]. It was experimentally verified that these networks also have the power-law structure. Moreover, the authors rigorously studied and interpreted cliques/independent sets and clique relaxations in these networks were in terms of stock classification and choosing robust diversified portfolios. Interestingly, due to the sparse structure of this graph, the maximum clique problem could be solved exactly using an appropriate preprocessing procedure, despite the fact that the original graph contained over 6000 vertices. This study provided an innovative approach to the analysis of the structural properties of the stock market, as well as its evolution over time. In particular, it turned out that the size of the largest cluster (clique) in this graph was steadily and substantially increasing over time, which indicated that more and more stocks exhibit similar behavior in the modern financial markets.

CONCLUDING REMARKS

Network-based data mining and related OR-based techniques are promising in a variety of applications, which has been demonstrated by multiple recent studies. In this article, we briefly discussed graph-theoretic concepts and problems that are useful from both data

mining and OR perspectives; however, the research in this area is far from complete. In the coming years, we expect that this methodology will continue to be developed and utilized in new exciting applications.

REFERENCES

1. Abello J, Pardalos PM, Resende MGC, editors. Handbook of massive data sets. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002.
2. Boginski V, Butenko S, Pardalos PM. Modeling and optimization in massive graphs. In: Pardalos PM, Wolkowicz H, editors. Novel approaches to hard discrete optimization. Providence (RI): American Mathematical Society; 2003. pp. 17–39.
3. Tan P-N, Steingach M, Kumar V. Introduction to data mining. Boston (MA): Addison-Wesley; 2006.
4. Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: formulations and challenges. *INFORMS J Comput* 1999;11(3):217–238.
5. Hayes B. Graph theory in practice. *Am Sci* 2000;88:9–13(Part I),104–109 (Part II).
6. Cook DJ, Holder LB. Graph-based data mining. *IEEE Intell Syst* 2000;15(2):32–41.
7. Washio T, Motoda H. State of the art of graph-based data mining. *SIGKDD Explor Newsl* 2003;5(1):59–68.
8. Alderson D. Catching the “Network Science” Bug: insight and opportunity for the operations researcher. *Oper Res* 2008;56:1047–1065.
9. Aiello W, Chung F, Lu L. Random evolution in massive graphs. In: Abello J, Pardalos P, Resende M, editors. Handbook on massive data sets. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2002.
10. Albert R, Barabasi A-L. Statistical mechanics of complex networks. *Rev Mode Phys* 2002;74:47–97.
11. Barabasi A-L. *Linked*. New York: Perseus Publishing; 2002.
12. Barabasi A-L, Albert R. Emergence of scaling in random networks. *Science* 1999;286:509–511.
13. Chung F, Lu L. Complex graphs and networks. CBMS Lecture Series. Providence (RI): American Mathematical Society; 2006.
14. Jarvis RA, Patrick EA. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans Comput* 1973;C-22(11):1025–1034.
15. Garey MR, Johnson DS. Computers and intractability: A guide to the theory of NP-completeness. New York: Freeman; 1979.
16. Arora S, Safra S. Approximating clique is NP-complete. Proceedings of the 33rd IEEE Symposium on Foundations on Computer Science; 1992 Oct 24–27; Pittsburg (PA). 1992. pp. 2–13.
17. Håstad J. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Math* 1999;182:105–142.
18. Pardalos PM, Mavridou T, Xue J. The graph coloring problem: a bibliographic survey. In: Du D-Z, Pardalos PM, editors. Volume 2, Handbook of combinatorial optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1998. pp. 331–395.
19. Jiang D, Pei J. Mining frequent cross-graph quasi-cliques. *ACM Trans Knowl Discov Data* 2009;2(4):1–42.
20. Zeng Z, Wang J, Zhou L, Karypis G. Coherent closed quasi-clique discovery from large dense graph databases. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York (NY): ACM; 2006. pp. 797–802.
21. Abello J, Resende MGC, Sudarsky S. Massive quasi-clique detection. Lecture notes in computer science, LATIN 2002: theoretical informatics. Heidelberg, Germany: Springer; 2002. pp. 598–612.
22. Seidman SB, Foster BL. A graph theoretic generalization of the clique concept. *J Math Sociol* 1978;6:139–154.
23. Luce RD. Connectivity and generalized cliques in sociometric group structure. *Psychometrika* 1950;15:169–190.
24. Mokken RJ. Cliques, clubs and clans. *Qual Quant* 1979;13:161–173.
25. Veremyev A, Boginski V, Krokhmal P, Jeffcoat D. Asymptotic behavior of cliques relaxations in random graphs. 2010. Working paper.
26. Balasundaram B, Butenko S, Hicks I. Clique relaxations in social network analysis: the maximum k-plex problem. *Oper Res* 2010. In press.
27. Balasundaram B, Butenko S, Trukhanov S. Novel approaches for analyzing biological networks. *J Comb Optim* 2005;10:23–39.
28. Veremyev A, Boginski V. Identifying large robust network clusters via new compact formulations of maximum k-club problems. 2010. Working paper.

29. Fischer I, Meinl T. Graph based molecular data mining - an overview. Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics. Piscataway (NJ): IEEE; 2004. pp. 4578–4582.
30. Balasundaram B, Butenko S. Network clustering. In: Junker BH, Schreiber F, editors. Analysis of biological networks. Hoboken (NJ): Wiley; 2008. pp. 113–138.
31. Pei J, Jiang D, Zhang A. Mining cross-graph quasi-cliques in gene expression and protein interaction data. Proceedings of the 21st International Conference on Data Engineering. Los Alamitos (CA): IEEE Computer Society; 2005. pp. 353–354.
32. Prokopyev OA, Boginski V, Chaovalitwongse W, *et al.* Network-based techniques in EEG data analysis and epileptic brain modeling. In: Pardalos PM, Boginski V, Vazacopoulos A, editors. Data mining in biomedicine. New York (NY): Springer; 2007. pp. 559–573.
33. Kleinberg J. Authoritative sources in a hyperlinked environment. *J ACM* 1999;46:604–632.
34. Brin S, Page L. The anatomy of a large scale hypertextual web search engine. Proceedings of the 7th World Wide Web Conference. Amsterdam, The Netherlands: Elsevier; 1998. pp. 107–117.
35. Mendelzon A, Mihaila G, Milo T. Querying the world wide web. *J Digit Libr* 1997;1:68–88.
36. Mendelzon A, Wood P. Finding regular simple paths in graph databases. *SIAM J Comp* 1995;24:1235–1258.
37. Broder A, Kumar R, Maghoul F, *et al.* Graph structure in the Web. *Comput Netw* 2000;33:309–320.
38. Willinger W, Alderson D, Doyle JC. Mathematics and the internet: a source of enormous confusion and great potential. *Not Am Math Soc* 2009;56(5):286–299.
39. Doyle JC, Alderson DL, Li L, *et al.* The “robust yet fragile” nature of the internet. *Proc Nat Acad Sci* 2005;102(41):14497–14502.
40. Abello J, Pardalos PM, Resende MGC. On maximum clique problems in very large graphs. Volume 50, DIMACS series. Providence (RI): American Mathematical Society; 1999. pp. 119–130.
41. Feo TA, Resende MGC. Greedy randomized adaptive search procedures. *J Glob Optim* 1995;6:109–133.
42. Feo TA, Resende MGC. A greedy randomized adaptive search procedure for maximum independent set. *Oper Res* 1994;42:860–878.
43. Boginski V, Butenko S, Pardalos PM. Mining market data: a network approach. *Comput Oper Res*. 2006;33:3171–3184. (Special Issue on Oper Res Data Min.)
44. Boginski V, Butenko S, Pardalos PM. Statistical analysis of financial networks. *Comput Stat Data Anal* 2005;48(2):431–443.
45. Aiello W, Chung F, Lu L. A random graph model for power law graphs. *Exp Math* 2001; 10:53–66.
46. Bomze IM, Budinich M, Pardalos PM, Pelillo M. The maximum clique problem. In: Du D-Z, Pardalos PM, editors. Handbook of combinatorial optimization. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1999. pp. 1–74.

NEUROECONOMICS AND GAME THEORY

MIRRE STALLEN
Rotterdam School of
Management, Erasmus
University Rotterdam,
Rotterdam The Netherlands

Donders Institute for Brain,
Cognition and Behavior,
Radboud University
Nijmegen, Nijmegen,
The Netherlands

ALAN G. SANFEY
Donders Institute for Brain,
Cognition and Behavior,
Radboud University
Nijmegen, Nijmegen,
The Netherlands
Behavioral Science Institute,
Radboud University
Nijmegen, Nijmegen,
The Netherlands

Traditionally, studies of decision making have focused on situations in which people choose between outcomes that only affect themselves, such as playing a lottery where choices are described in terms of outcomes and probabilities and players have to pick the lottery that they prefer. However, many of our decisions in daily life are made in a social context in which the outcome of a decision has consequences not only for one's self but for others as well, and where we need to consider the desires and values of others before deciding.

Game theory is a collection of models for understanding such instances of interactive decision making [1]. These models make clear mathematical predictions about behavior in social situations by describing what strategies decision makers converge on as they try to maximize their payoff. To study complex social processes, such as trust, cooperation, and reputation, game theory offers a rich variety of well-designed multiplayer games. Recently, a new interdisciplinary field

has emerged, popularly known as *neuroeconomics*, which aims to better understand how people make decisions by integrating research from several different fields (for a detailed overview of the field see the handbook on neuroeconomics by Glimcher and colleagues [2]). As part of this neuroeconomic approach, researchers have begun to investigate the neural correlates of social decision making using some of these games.

In this article, we first provide a brief introduction as to how neuroeconomic studies are typically conducted. Following this, we describe the main findings of current neuroeconomic studies as they relate to game theory, and we conclude with an outline of potential future research directions in this field.

NEUROECONOMICS

Researchers in neuroeconomics have sought to investigate economic decision making in human subjects by combining a variety of methods from experimental economics, psychology, and neuroscience. These methods include (i) behavioral studies that examine how manipulation of the decision environment can affect choice, reaction time, skin conductance, or eye movements; (ii) studies with brain lesion patients that examine the consequences of abnormal brain function on decision making; (iii) experiments applying repetitive transcranial manipulation (rTMS) to temporarily disrupt activity within the brain; (iv) electroencephalography (EEG) or magnetoencephalography (MEG) studies, which measure at the scalp the electrical signals of neuronal firing; (v) pharmacological research to examine the effects of drug administration and neurotransmitters; (vi) genetic studies looking at the correlation between individual differences in the expression of certain genes and behavior; (vii) research using positron emission topography (PET) to investigate the functioning of brain structures by the use of radioactive isotopes; and (viii) functional magnetic resonance imaging (fMRI) experiments that

allow for the relatively direct measurement of real-time neural activity.

Over the last decade, the latter of these methods, fMRI, has emerged as the dominant technique in neuroeconomic studies. This method provides a noninvasive, indirect measure of the neural activity that occurs during decision making by assessing regional changes in the level of blood oxygenation. Typical fMRI experiments have a temporal resolution in the order of 1 s and a spatial resolution of 2–3 mm, which means that use of fMRI allows for the measurement of transient cognitive events, and also that relatively small structures within the brain can be imaged. A technical constraint of the method is that the blood oxygen level dependent response (BOLD) signal is rather noisy, requiring usually many trials to average the signal to a stable level. This has implications for the type of decision that can be studied, and as a consequence most game theoretic experiments undertaken with fMRI require many repetitions of the same game. This in turn means that genuine “single-shot” games (i.e., where players play only one game with only one partner) are typically not well suited to the MRI environment.

CURRENT RESEARCH

Recent research in neuroeconomics has begun to investigate the processes underlying social decision making using game theoretic paradigms. In contrast to standard behavioral studies in game theory, the neuroeconomic approach allows for the discrimination and modeling of processes that are hard to separate at the behavioral level. The combination of game theoretic models with the on-line measurement of brain activity during instances of decision making offers the promise of increasing our understanding of the processes and factors that are involved in social decision making.

An additional benefit of the neuroeconomic approach is that it has the potential to inform game theory models. This can be helpful, since actual decision behavior in game theory tasks often deviates from the models’ predictions. Game theory models assume that

players in a social context interact rationally, with full information, and that they make purely self-interested decisions. However, several decades of laboratory experiments by psychologists and economists have found that players rarely behave according to these strict game theoretic strategies [3]. Instead, players often appear to care about the welfare of others, and in addition value factors like reciprocity and fairness. Identification of the neural mechanisms underlying social interaction in game theory paradigms may result in a more precise characterization of behavior and, subsequently, game theory models may be adapted to better fit decision-making behavior.

Although neuroeconomics is still a relatively young discipline, there have already been many studies conducted that have used game theoretic approaches to gain insight into the neural basis of social decision making [4,5]). The current findings can be usefully summarized in three themes: (i) reasoning about the intentions of others, (ii) social reward, and (iii) the role of emotions in decision making, each of which is expanded upon in the remainder of this section.

Reasoning about the Intentions of Others

One focus of game theory is to model reciprocal exchange, which has been studied extensively in the laboratory with games like the trust game [6–9] and the prisoner’s dilemma game [10–12]. A trust game is played with two players. One is termed the *investor*, who is endowed with a particular monetary amount, and the other is termed the *trustee*. The investor is instructed to choose how much of her endowment she would like to pass on to the trustee and how much she will keep for herself. Once this decision is made, the experimenter multiplies the transferred amount of money by some factor (usually 3 or 4) and passes it to the trustee. At this point the trustee has the option to return the money to the investor or to keep all the money to herself. If the trustee decides to return the money, thereby honoring the trust of the investor, both players finish the game with more money than they originally had. However, if the trustee decides to keep the money and thus does not repay the trust,

the investor ends up with a loss and the trustee is the only one who benefits.

As indicated above, game theory assumes that a decision maker acts in a rational and self-interested manner, and therefore the classical economic prediction would be that a trustee in the trust game would never repay the trust of the investor. The investor, in turn, will realize this and should therefore never invest in the trustee in the first place. However, although game theory predicts no transactions to occur, behavioral experiments have consistently shown that players in the trust game do in fact engage in social interaction: the majority of investors send money to the trustees and most of the time this trust is reciprocated with payments back to the investor [3].

In a standard laboratory prisoner's dilemma game, game theoretic predictions are also usually violated. The prisoner's dilemma game is quite similar to the trust game, except that in the prisoner's dilemma game both players simultaneously choose whether to cooperate with the other or not, without knowing their partner's choice. The payoff for both players depends on the interaction of their choices, such that a player's individual earnings are highest when he or she defects and the other cooperates, with the cooperating player receiving only a small amount. Mutual cooperation, in contrast, is rewarding for both players, although individual earnings are lower than during unilateral defection. Game theory would predict that in the prisoner's dilemma game both players should decide to defect, but again, this appears to be not in accordance with the experimental data, as behavioral experiments typically find mutual cooperation occurring about half of the time. Given that standard game theoretic models are often inaccurate at describing individual decision making, how can neuroeconomic studies shed light on why human behavior appears at odds with classical economic predictions?

One potentially interesting contribution may come from how the brain perceives and computes the intentions of others, suggesting that we value not only the monetary outcome of a social interaction, but are also concerned with the message

our partner conveys by his or her behavior [13,14]. Several neuroimaging studies have shown a consistent network of areas to be involved during reciprocal exchange in both the trust game and prisoner's dilemma game [11,15,16]. For example, one of the earliest neuroeconomic studies on the trust game found that activity in the paracingulate cortex was higher when subjects played the game with another human being as opposed to a computer opponent [16] (Fig. 1a). Increased activation in this same region was also found in a following MRI study using the trust game [15]. In this study, trust building was examined by having several pairs of strangers play a nonanonymous multiround trust game, while alternating their roles as trustor and trustee. Both players were each in a separate MRI scanner, with both brains scanned simultaneously during the game. Using within- and between-brain analyses, the results showed that the paracingulate cortex was particularly active during the first phase of the multiround game for players that reciprocated their partner's decision to trust. In later stages of the game, the activity in this region decreased for this group of reciprocators. This result suggested that the paracingulate cortex is important during the early trust building phases of a relationship and that it plays less of a role during the maintenance stage of trust.

Other areas that have been consistently associated with reciprocal exchange are the posterior superior temporal sulcus and the temporal-parietal junction (Fig. 1b). Together with the paracingulate cortex, these regions have been strongly implicated in the cognitive capacity of perspective taking, an ability often termed as *theory of mind* [17,18]. The concept of theory of mind is defined as the understanding that others have of their own intentions and individual perspective of the world, and that these may differ from your own. The finding that regions related to theory of mind are involved in reciprocal exchange games suggests that players in the trust game and prisoner's dilemma game attempt to infer the strategy of their opponent and subsequently act on the belief they have about their partner's intentions. Therefore, if we perceive that a partner treats

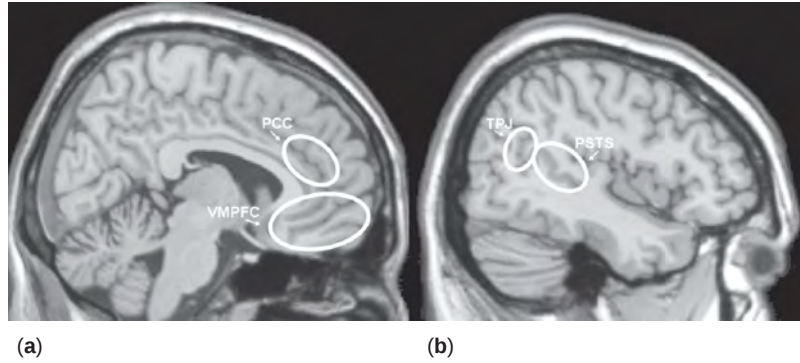


Figure 1. Brain regions involved in decision making: (a) PCC, paracingulate cortex; VMPFC, ventromedial prefrontal cortex; $x = 46$; (b) TPJ, temporal-parietal junction; PSTS, posterior superior temporal sulcus; $x = 7$.

us badly when they have the opportunity to do otherwise, we may punish that person. However, if we believe that our partner had no choice in their behavior, we may decide to be more lenient [19]. Furthermore, these regions also seem to play a role in people's ability to determine whether the actions of others are meaningful. Many neuroimaging studies have found that game interactions with human partners produce stronger activity in theory of mind regions than identical interactions with purported computer partners [16,20,21].

Interestingly, areas important for theory of mind may also be involved in the decision to behave selfishly or altruistically. A study on altruism found that activity in the posterior superior temporal cortex correlated with individual's tendency to engage in helpful behavior [22]. Of course, in addition to the cortical network related to theory of mind, other areas may be relevant for reasoning about the intentions of others. For example, the cingulate cortex seems to represent information about which player in a trust game is responsible for a specific outcome (themselves or their partner), as its activity correlates with the agent-responsibility ("me" or "not me") of an action [23]. Future research is expected to increase our understanding on the neural mechanism of theory of mind even further by investigating what additional areas may be involved in the network as well. Furthermore, our knowledge of the specific

functions of these regions is still limited. This situation will improve as more studies in this domain will be conducted.

Social Reward

In addition to the involvement of brain structures related to theory of mind, research on the trust game and prisoner's dilemma game have indicated another area that is critical to social decision making: the striatum [7,10,11,24] (Fig. 2a). The striatum is located in the center of the brain and is a major input station of midbrain dopamine neurons that play a primary role in the processing of rewards. Single-cell recordings in nonhuman primates have demonstrated that dopamine neurons in the striatum track reward prediction errors, with unexpected rewards increasing the firing rate of dopamine cells and the omission of expected rewards leading to a decrease in the firing rate [25,26]. This reinforcement-learning mechanism is thought to be important for the learning of reward values of stimuli in the environment, thereby allowing for the improvement of choices over time.

In parallel to these primate studies, the striatum in humans is known to be involved in reward processing. Using neuroimaging and PET, researchers have discovered that this area responds not only to primary rewards such as liquids, foods, and sexual stimuli [27–30] but also to money [31,32] or more abstract rewards like reputation or

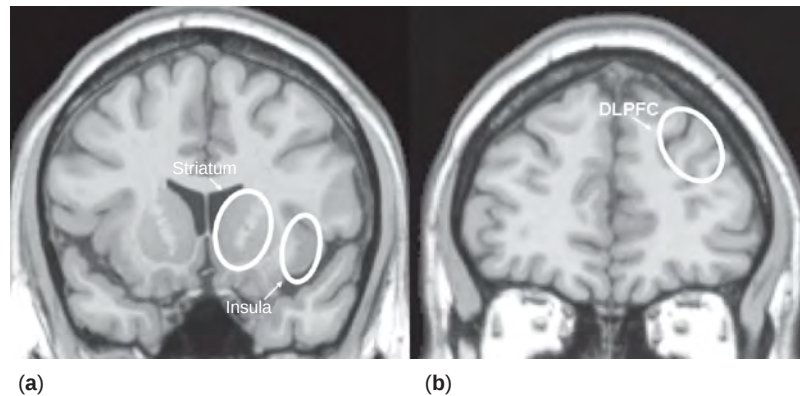


Figure 2. Brain regions involved in decision making. Locations indicated only in the right hemisphere. (a) Striatum; anterior insula; $y = 14$. (b) DLPFC, dorsolateral prefrontal cortex; $y = 45$.

status [33,34]. Importantly, reward-related activity has also been associated with social decision-making behavior. Several neuroimaging studies have demonstrated that this area responds to the decision of one's partner to either reciprocate or not reciprocate cooperation [7,10,11,24]. For example, in the prisoner's dilemma game, the striatum is activated when cooperation is reciprocated and deactivated if not reciprocated [10,11].

In addition, the striatum appears to update the trustworthiness of a partner, as activity in this area has been associated with increased cooperation or reciprocity in subsequent rounds of a prisoner's dilemma [10] or trust game [7]. A further study using the trust game paradigm showed that reward prediction errors in the striatum are significantly dependent on prior experiences, such as the moral reputation of players [24]. Providing investors with a fictional profile of the moral reputation of their partners (either morally good, bad, or neutral) before the start of a trust game resulted in reduced striatal activity when people interacted with morally good or bad partners, but not with neutral partners. Presumably this modulation of striatal activity was absent when players interacted with neutral partners because the behavior of the neutral partners was more unpredictable, as no prior information was given on their reputation. This finding

demonstrates that prior beliefs may influence the processing of social feedback information, thereby guiding future cooperation.

Reward-related activity in social decision making has been associated not only with positive, mutual cooperation. A study using PET showed that signals in the striatum were also enhanced when players punished free riders (nonreciprocators that benefit from others in a trust game), suggesting that players may derive satisfaction from punishing defectors [35]. In addition, studies have reported activity in reward networks when participants engaged in charitable donations at a personal cost to themselves [36,37]. For example, a recent study [38] provided participants with an endowed amount of money and then observed them using fMRI while they engaged in a variety of donation situations. The researchers found that reward areas were activated when participants observed a donation to a charity and, importantly, that this activation was enhanced when the donation was made voluntarily.

These neuroeconomic results are important for game theoretic approaches, as they show that regions underlying social decision making overlap with classical reward circuitry, thereby suggesting that cooperative and reciprocal behavior may be experienced directly as rewarding by players, independent of any monetary reward. This provides valuable evidence that social motivations are

indeed important when engaging in social decision making, and that we need to consider rewards other than money when constructing models of behavior in these contexts. Furthermore, the results on charitable donations are intriguing as they have relevant implications for decision making at a societal level. The fact that mandatory transfers to a charity elicit reward-related activity suggests that even taxation may produce satisfaction in certain situations.

The Role of Emotions

In addition to research on strategizing and social reward, another fruitful direction within neuroeconomics is the growing insight that emotions play a crucial role in social decision making. This idea is in contrast to classical models of economic decision making, which assume that decisions are the results of a rational and deliberate evaluation of the available choice options. However, although these models have provided a strong foundation for the development of central decision-making theories, more recent research showed that this rationality assumption is often violated [39,40].

Early work in this domain showed that patients with prefrontal brain damage and associated emotional deficits performed worse on a gambling task than healthy participants [41]. The area that was damaged in these patients was the ventromedial prefrontal cortex, a region that later was found to be essential for the activation and successful integration of emotion-related memories in the anticipation of future consequences of actions [42] (Fig. 1a). Building from this pioneering research, researchers have continued to discover that emotions play a vital role in determining decisions.

A standard game that is frequently used to illustrate that emotions have an effect on decision making is the ultimatum game. In this well-studied task, two players are required to split a sum of money that is provided by the experimenter. One of the players is deemed the proposer, and the other the responder. The proposer is instructed to specify the division of the money between the two players, and is allowed to make any split he or she wants. The responder can then accept

or reject the offer; accepting means the money is split as proposed and rejecting means that neither player receives anything. Since game theory predicts people are motivated purely by self-interest, the standard game theoretic solution for the responder is to accept any amount that is offered. The proposer should anticipate this and therefore offer the responder the smallest amount possible. However, a considerable amount of behavioral research showed that this prediction is at odds with observed behavior in the laboratory and the field: proposers generally propose an equal split and responders reject offers that are 20% of the total amount or less [43]. This very robust finding indicates that people's behavior in the ultimatum game appears not to be motivated solely by financial self-interest.

Neuroeconomic studies have attempted to specify the processes underlying the reactions of responders to unfair offers in the ultimatum game. An early study by Sanfey and colleagues [44] found that activity in the anterior insula correlated with the degree of unfairness of an offer (Fig. 2a). In addition, this region was more active for offers made by a human partner as opposed to that by a computer partner. In terms of the responder's decision, activation in this area predicted the player's decision to either accept or reject the offer, with a higher level of activity for rejections than for acceptances. The activation of the insula in the ultimatum game is particularly interesting in light of its suggested association with negative emotional states. Previous studies showed that activation in this area is consistently seen in studies to pain and disgust, and to autonomic arousal [45,46]. On the basis of these findings, anterior insula activity in the ultimatum game can be interpreted as a reflection of the responder's negative emotional state to an unfair offer.

Another brain region that was engaged during unfair offers in the ultimatum game was the right dorsolateral prefrontal cortex, a region that, in contrast to the anterior insula, usually has been associated with cognitive processes, such as the implementation of goal-directed behavior (Fig. 2b). Activation in this area was greater than activation in the insula when responders accepted unfair

offers, and less when they rejected these offers. To investigate the function of the dorsolateral prefrontal cortex in the decision to accept or reject an unfair offer, activity in this brain region has been disrupted using TMS [47,48]. After receiving TMS at the right dorsolateral prefrontal cortex, the acceptance rate of unfair offers increased, suggesting that the dorsolateral prefrontal cortex plays a causal role in the implementation of fairness-related behavior.

The finding of separable neural correlates for decisions based on emotion (rejections of an unfair offer) and deliberation (acceptance of an unfair offer) suggests that there is no unitary system at the neural level that is in control of decision making. Indeed, several neuroimaging studies showed that emotional processes seem to reliably engage a different set of brain structures than cognitive processes [49,50]. This distinction provides some support for dual-process theories, which hypothesize that decision making is the result of an interaction between an automatic, affective system and a more controlled, deliberative system. However, explaining decision making in terms of the interaction between different subsystems violates traditional economic theory, as standard economic models typically describe human behavior as being governed by a unitary process. Future economic models of decision making could usefully take into account the neuroeconomic evidence for the differential processing of emotional and deliberative information.

THE FUTURE

Many of our most important decisions are made in a social context. In the preceding sections we presented an overview of how neuroeconomics has contributed thus far to the understanding of social decision-making behavior. These findings have demonstrated that the brain has the ability to rapidly recognize and act on the intentions of others, that key regions involved in reward-based learning also underlie choices in social settings, and that emotions play a crucial role in social decision-making situations. Despite these advances, the current state of knowledge regarding the mechanisms underlying

social decision making is still rather limited. However, owing to the explosion of research in this field, and to continual refinement and advances in neuroscientific methodology, this area of research offers much promise in the years to come.

One potentially fruitful avenue of research is the investigation of how parameters of decision processes are represented in the brain using a formal modeling approach [51,52]. By connecting the parameters of formal (game theoretic) models with neural activity, a more precise characterization of the mechanisms underlying decision making can be described. In turn, by examining whether correlates of game theory parameters are present or absent in the brain, this approach may provide additional constraints on models that aim to predict social decision-making behavior.

Future studies are also likely to benefit from advancements in neuroscientific methods. For instance, fMRI measurement techniques are being improved continuously. The traditional, and still most common, approach to investigate brain function is the localization of function to specific isolated brain areas. However, a more recent way of examining neural function is to look at the structural and functional connectivity between brain areas [53]. Structural connectivity methods, such as diffusion tensor imaging, aim at inferring with the anatomical connectivity between brain regions by tracking bundles of white matter fibers linking cortical areas. Functional methods like Granger causality mapping or dynamic causal modeling investigate the effective relationship between brain areas by modeling the direct or indirect influence of one brain area over another (possibly more spatially remote) brain area. Currently, these brain connectivity methods are becoming increasingly popular and neuroeconomic data generated by this alternative approach can yield useful insights into the neural networks that are relevant for decision making.

A final interesting direction for future research is to investigate the role of hormonal and genetic factors in decision making. For instance, oxytocin, a peptide that functions both as a neurotransmitter and as a hormone,

appears to modulate a broad profile of human social behaviors. In a trust game, intranasal administration of this peptide increased the amount of money transferred by the investor [8]. A related neuroimaging study showed that this change in trusting behavior was associated with a modulation of activity in the neural systems mediating both fear and reward processing [54]. In addition, higher levels of oxytocin have been associated with more generous offers toward strangers in an ultimatum game [55,56]. Another hormone potentially of importance for social decision making is testosterone. A recent study showed that men high in testosterone were more likely to reject relatively low offers in an ultimatum game [57]. Subsequent research also examined whether testosterone levels affect the decision of the proposer. However, as is often the case in early studies contrasting findings have been reported, as one study found that artificially raised testosterone levels decreased the generosity of proposers [58], whereas another study found opposite results [59].

Researchers have also begun to examine whether genetic variation within people can explain individual differences in social decision-making behavior. Results from a twin study of the heritability of ultimatum game behavior showed that more than 40% of the variation of responder's rejection rate was explained by genetic effects. Furthermore, studies of oxytocin and vasopressin—a peptide with a structure very similar to that of oxytocin—have shown that the genetic polymorphisms for oxytocin and vasopressin receptors are associated with the allocation of funds in economic games that measure altruistic behavior [60–62].

Clearly, the future looks promising for a neuroeconomic approach of social decision making. Exploration of new research domains and the development of more advanced techniques offer fruitful opportunities for the study that combines the mathematical models of game theory with techniques of modern neuroscience. This approach will benefit the predictive accuracy of economic models by constraining their parameters based on neural substrates as

well as further our knowledge of how the brain functions.

REFERENCES

1. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton: Princeton University Press; 1947.
2. Glimcher PW, Camerer CF, Fehr E, *et al.*, editors. Neuroeconomics: decision making and the brain. London: Academic Press; 2009.
3. Camerer CF. Behavioral game theory—experiments in strategic interaction. Princeton: Princeton University Press; 2003.
4. Krueger F, Grafman J, McCabe K. Neural correlates of economic game playing. *Philos Trans R Soc Lond B Biol Sci* 2008; 363(1511):3859–3874.
5. Sanfey AG. Social decision-making: insights from game theory and neuroscience. *Science* 2007;318(5850):598–602.
6. Berg J, Dickhaut J, McCabe K. Risk preference instability across institutions: a dilemma. *Proc Natl Acad Sci USA* 2005; 102(11):4209–4214.
7. King-Casas B, Tomlin D, Anen C, *et al.* Getting to know you: reputation and trust in a two-person economic exchange. *Science* 2005;308(5718):78–83.
8. Kosfeld M, Heinrichs M, Zak PJ, *et al.* Oxytocin increases trust in humans. *Nature* 2005;435(7042):673–676.
9. Zak PJ, Kurzban R, Matzner WT. Oxytocin is associated with human trustworthiness. *Horm Behav* 2005;48(5):522–527.
10. Rilling J, Gutman D, Zeh T, *et al.* A neural basis for social cooperation. *Neuron* 2002;35(2):395–405.
11. Rilling JK, Sanfey AG, Aronson JA, *et al.* Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport* 2004;15(16):2539–2254.
12. Sally D. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Ration Soc* 1995;7(1):58–59.
13. Boudreau C, McCubbins MD, Coulson S. Knowing when to trust others: an ERP study of decision making after receiving information from unknown people. *Soc Cogn Affect Neurosci* 2009;4(1):23–34.
14. Rudoy JD, Paller KA. Who can you trust? Behavioral and neural differences between

- perceptual and memory-based influences. *Front Hum Neurosci* 2009;3:16.
15. Krueger F, McCabe K, Moll J, *et al.* Neural correlates of trust. *Proc Natl Acad Sci USA* 2007;104(50):20084–20089.
 16. McCabe K, Houser D, Ryan L, *et al.* A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci USA* 2001;98(20):11832–11835.
 17. Frith CD, Frith U. How we predict what other people are going to do. *Brain Res* 2006;1079(1):36–46.
 18. Saxe R. Uniquely human social cognition. *Curr Opin Neurobiol* 2006;16(2):235–239.
 19. Fehr E, Gächter S. Altruistic punishment in humans. *Nature* 2002;415(6868):137–140.
 20. Gallagher HL, Jack AI, Roepstorff A, *et al.* Imaging the intentional stance in a competitive game. *Neuroimage* 2002;16(3 Pt 1):814–821.
 21. Rilling JK, Sanfey AG, Aronson JA, *et al.* The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* 2004;22(4):1694–1703.
 22. Tankersley D, Stowe CJ, Huettel SA. Altruism is associated with an increased neural response to agency. *Nat Neurosci* 2007;10(2):150–151.
 23. Tomlin D, Kayali MA, King-Casas B, *et al.* Agent-specific responses in the cingulate cortex during economic exchanges. *Science* 2006;312(5776):1047–1050.
 24. Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat Neurosci* 2005;8(11):1611–1618.
 25. Schultz W. Behavioral theories and the neurophysiology of reward. *Ann Rev Psychol* 2006;57(1):87–115.
 26. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. *Science* 1997;275(5306):1593–1599.
 27. O'Doherty JP, Dayan P, Friston K, *et al.* Temporal difference models and reward-related learning in the human brain. *Neuron* 2003;38(2):329–337.
 28. Berns GS, McClure SM, Pagnoni G, *et al.* Predictability modulates human brain response to reward. *J Neurosci* 2001;21(8):2793–2798.
 29. Arnow BA, Desmond JE, Banner LL, *et al.* Brain activation and sexual arousal in healthy, heterosexual males. *Brain* 2002;125(5):1014–1023.
 30. Redouté J, Stoléru S, Grégoire MC, *et al.* Brain processing of visual sexual stimuli in human males. *Hum Brain Mapp* 2000;11(3):162–177.
 31. Knutson B, Westdorp A, Kaiser E, *et al.* fMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* 2000;12(1):20–27.
 32. Breiter HC, Aharon I, Kahneman D, *et al.* Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 2001;30(2):619–639.
 33. Izuma K, Saito DN, Sadato N. Processing of social and monetary rewards in the human striatum. *Neuron* 2008;58(2):284–294.
 34. Zink CF, Tong Y, Chen Q, *et al.* Know your place: neural processing of social hierarchy in humans. *Neuron* 2008;58(2):273–283.
 35. de Quervain DJ, Fischbacher U, Treyer V, *et al.* The neural basis of altruistic punishment. *Science* 2004;305(5688):1254–1258.
 36. Moll J, Krueger F, Zahn R, *et al.* Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc Natl Acad Sci USA* 2006;103(42):15623–15628.
 37. Hare TA, Camerer CF, Knoepfle DT, *et al.* Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *J Neurosci* 2010;30(2):583–590.
 38. Harbaugh WT, Mayr U, Burghart DR. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 2007;316(5831):1622–1625.
 39. Kahneman D, Slovic P, Tversky A. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press; 1982.
 40. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science* 1974;185(4157):1124–1131.
 41. Bechara A, Damasio H, Tranel D, *et al.* Deciding advantageously before knowing the advantageous strategy. *Science* 1997;275(5304):1293–1295.
 42. Bechara A, Damasio H, Tranel D, *et al.* The Iowa gambling task and the somatic marker hypothesis: some questions and answers. *Trends Cogn Sci* 2005;9(4):159–162; discussion 162–154.
 43. Güth W, Schmittberger R, Schwarze B. An experimental analysis of ultimatum bargaining. *J Econ Behav Organ* 1982;3(4):367–388.

44. Sanfey AG, Rilling JK, Aronson JA, *et al.* The neural basis of economic decision-making in the Ultimatum Game. *Science* 2003;300(5626):1755–1758.
45. Damasio AR, Grabowski TJ, Bechara A, *et al.* Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat Neurosci* 2000;3(10):1049–1056.
46. Calder AJ, Lawrence AD, Young AW. Neuropsychology of fear and loathing. *Nat Rev Neurosci* 2001;2(5):352–363.
47. van 't Wout M, Kahn RS, Sanfey AG, *et al.* Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport* 2005;16(16):1849–1852.
48. Knoch D, Pascual-Leone A, Meyer K, *et al.* Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 2006;314(5800):829–832.
49. Dalglish T. The emotional brain. *Nat Rev Neurosci* 2004;5(7):583–589.
50. Sanfey AG, Chang LJ. Multiple systems in decision making. *Ann N Y Acad Sci* 2008;1128:53–62.
51. van Winden F, Stallen M, Ridderinkhof KR. On the nature, modeling, and neural bases of social ties. *Adv Health Econ Health Serv Res*. 2008;20:125–159.
52. Beer JS, Stallen M, Lombardo MV, *et al.* The quadruple process model approach to examining the neural underpinnings of prejudice. *Neuroimage* 2008;43(4):775–783.
53. Guye M, Bartolomei F, Ranjeva JP. Imaging structural and functional connectivity: towards a unified definition of human brain organization. *Curr Opin Neurol* 2008;21(4):393–403.
54. Baumgartner T, Heinrichs M, Vonlanthen A, *et al.* Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 2008;58(4):639–650.
55. Barraza JA, Zak PJ. Empathy toward strangers triggers oxytocin release and subsequent generosity. *Ann N Y Acad Sci* 2009;1167:182–189.
56. Zak PJ, Stanton AA, Ahmadi S. Oxytocin increases generosity in humans. *PLoS One* 2007;2(11):e1128.
57. Burnham TC. High-testosterone men reject low ultimatum game offers. *Proc Biol Sci* 2007;274(1623):2327–2330.
58. Zak PJ, Kurzban R, Ahmadi S, *et al.* Testosterone administration decreases generosity in the ultimatum game. *PLoS One* 2009;4(12):e8330.
59. Eisenegger C, Naef M, Snozzi R, *et al.* Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature* 2009;463(7279):356–359.
60. Israel S, Lerer E, Shalev I, *et al.* Molecular genetic studies of the arginine vasopressin 1a receptor (AVPR1a) and the oxytocin receptor (OXTR) in human behaviour: from autism to altruism with some notes in between. *Prog Brain Res* 2008;170:435–449.
61. Israel S, Lerer E, Shalev I, *et al.* The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. *PLoS One* 2009;4(5):e5535.
62. Knafo A, Israel S, Darvasi A, *et al.* Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes Brain Behav* 2008;7(3):266–275.

NEUROECONOMICS INSIGHTS FOR DECISION ANALYSIS

KAISA HYTÖNEN

Rotterdam School of Management,
Erasmus University Rotterdam,
Rotterdam, The Netherlands

Donders Institute for Brain,
Cognition and Behavior, Radboud
University Nijmegen, Nijmegen,
The Netherlands

ALAN G. SANFEY

Donders Institute for Brain,
Cognition and Behavior, Radboud
University Nijmegen, Nijmegen,
The Netherlands

Behavioral Science Institute,
Radboud University Nijmegen,
Nijmegen, The Netherlands

In recent years, neuroeconomics has emerged as an exciting interdisciplinary field with the stated aim of combining knowledge from economics, psychology, and neuroscience in order to increase the understanding of decision-making behavior. The combination of these different disciplines has provided an opportunity to begin specifying the brain basis of decision-making, as well as informing theoretical models of decision-making. In addition, there is now the growing awareness that many of the findings from the field can potentially have implications for more practical decision analysis situations.

According to the classical economic perspective [1], decision makers approach choice situations by analyzing the attainable outcomes and their associated probabilities. The well-known expected utility model of decision making under risk is a good formulation of this behavior, with this model based on an axiomatic foundation that reflects a rationality assumption on the part of the decision maker [2]. Despite the model's theoretical elegance, it has been often demonstrated that it does not provide an adequate description of typical decision making, and neuroeconomic studies are beginning to demonstrate

how a better understanding of the neural processes involved can help explain these discrepancies.

In this article, we will first provide some brief details about typical methods used in neuroeconomic studies, before examining how knowledge about the valuation system in the human brain can yield insight into how decisions are made, and how this knowledge can in turn help build better models of decision and choice. We will then discuss a particularly well-known behavioral anomaly, the framing effect, and how understanding of the brain's emotional mechanisms can lead to better models of this behavior, before we conclude with some general observations about the relevance of this new discipline for decision analysis.

METHODS

One important approach to neuroeconomics has been to utilize the varied set of methods developed by neuroscientists to examine higher-level cognitive processes. These methods vary widely both in their use and in terms of what questions they can answer, but together they provide a highly flexible set of tools for examining the neural substrates of decision making. Broadly speaking, they can be divided into two types, those that observe the normal functioning brain and those that examine perturbations in these normal functions.

The most frequently used methods measure the degree of "activation" of various brain regions while people are making choices and decisions. For instance, the commonly used method of functional magnetic resonance imaging (fMRI), (see Ref. 3 for information on the method) leverages the physiological fact that changes in neural blood flow leads to corresponding changes in local magnetic properties of the brain, which in turn can be detected by an MRI machine. These blood flow changes are thought to be directly related to regional neural activity, thus providing a measure which correlates

with neural firing. The majority of the results we discuss in this article are based on fMRI studies; however, other noninvasive measuring techniques are used by neuroeconomists, such as those that measure electrical (EEG) and magnetic field (MEG) changes related to brain activity.

A different approach has been to examine the functioning of the disrupted brain to make inferences about normal behavior. Indeed, the study of decision making processes in the human brain was largely motivated by behavioral deficits that were apparent in patients with brain damage. For example, brain lesions in the frontal lobes of the brain appear to be associated with uncontrolled behavior: patients take excessive risks regardless of their intellectual capabilities or other demographic or psychological factors. The lack of an appropriate skin conductance response to emotional stimuli in this patient group set forth the idea that brain damage to the frontal lobe interfered with behavior through inappropriate integration of emotional knowledge in the decision-making process [4,5], thus providing an important initial clue as to how the brain's organization related to decision-making processes. In addition to examining the behavioral deficits of patients with lesions, there are also methods to temporarily disrupt neural processing using electrical and magnetic signals in healthy participants, such as Transcranial Magnetic Stimulation (TMS).

PREFERENCES AND VALUATION

Reward Processing in the Brain

The functioning of the human brain and the nervous system is based on millions of brain cells called neurons. In practical terms the brain is a vast network of neurons that communicate with each other. While structurally, the human brain can be divided into separate regions, some of which appear to be functionally specific, there is also much overlap between these regions.

Reward processing in the human brain is closely related to the neurotransmitter dopamine. Dopamine is produced in the brainstem and then projected to multiple

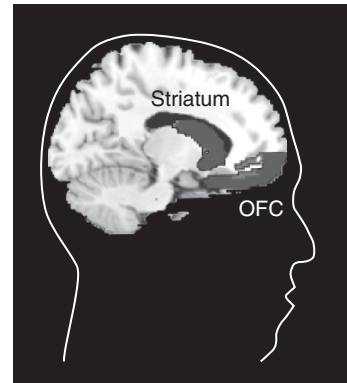


Figure 1. The striatum is located in the basal ganglia at the center of the brain. The orbitofrontal cortex (OFC) is a subsection of the frontal cortex that is specialized in valuation. The figure indicates the approximate anatomical locations of the regions of interest.

areas in the brain. Areas of the brain such as the striatum and orbitofrontal cortex (OFC) receive input from these dopaminergic neurons and are thought to process the rewarding value of stimuli (Fig. 1). Additionally, ventromedial prefrontal cortex (VMPFC) is often mentioned in this context as well. Since the anatomical definitions of medial OFC and VMPFC are somewhat overlapping, and indeed often used in parallel to refer to the same location, we adopt here for simplicity the notation of using OFC when we refer to a prefrontal region that reacts to rewards (medial OFC and VMPFC) and punishments (lateral OFC) [6].

The target regions of dopamine neurons have been shown to react to both primary and secondary rewards. Striatum activity increases with the receipt of primary rewards such as water and juice [7], and OFC activity reflects the rewarding value of stimulus in multiple different modalities like taste and olfaction [8]. In addition to primary rewards, both the striatum and OFC react to the receipt of secondary rewards such as financial incentives [9–12]. OFC is also sensitive to more subjective features of the stimuli, and is involved in the calculation of the hedonic value of rewards [6,13]. OFC even tracks modulations in the hedonic value that are driven by contextual information. In one study where participants evaluated

wine, the perceived pleasantness of the wine and the OFC activity varied as a function of retail price, though the actual wine consumed remained the same: high price increased the pleasantness estimation both in the brain and in the behavioral measures [14]. Of course, in terms of rational models, the consumption experience of identical wines should not depend on the surrounding context (price), but this research demonstrates that the brain's computations are not fully in accordance with a classical utility calculation process.

Anticipation and Evaluation of Future Rewards

In order to choose the option that provides the highest level of satisfaction, decision makers need to anticipate the rewarding value of each outcome in the relevant choice set. As this is clearly an important issue to economic decision making, anticipatory reward mechanisms were among the first topics that were widely researched in neuroeconomics. The research to date indicates that the reward circuitry in the brain demonstrates increased activation not only when rewards are obtained but also during the anticipation of rewarding outcomes [8,15,16]. Importantly, these anticipatory patterns of activation are not just correlates of the subjective value of the reward, but also have an influence on future economic choices themselves. Negative anticipatory affect during decision making is related to an increase in risk aversion, and positive anticipatory affect in contrast is associated with an increasing risk seeking attitude [17].

In addition to anticipatory reward processing, decision makers in risky choice scenarios must also take into account the probability of each outcome option in order to reach an overall estimate of the attractiveness of the prospects. In a typical experiment, participants encounter a series of uncertain prospects in turn, after which the prospects are resolved, revealing the gain or loss to the participant. As an example, in one task participants view eight cards face-down in random order, one of which is the target card. The task of the participants is to place a bet on the cards. If participants place the bet on the target card they win the bet, otherwise

they lose an equal amount of money from their total earnings. The magnitude of the outcome is manipulated by providing the participants with either 1 or 5 Euros to bet with. The probability of receiving the reward is manipulated by allowing the participants to place the bet on either just one card (low probability) or on the corners of four adjacent cards (high probability). By using this paradigm, Yacubian and colleagues showed that both the striatum and OFC encode anticipation of probabilistic rewards [18]. The findings of this experiment indicate that both the reward magnitude and also the associated probabilities are directly encoded in the valuation network itself: the striatum and OFC had higher activation levels in the high bet and high probability conditions than in the low bet and low probability conditions, respectively. Further, the results of this experiment show that the reward circuit also provides a measure for the overall desirability of a risky prospect as it encodes the expected value of a prospect, integrating the value and probability information for rewarding stimuli.

Other fMRI experiments that have studied the valuation of risky prospects have reported similar findings in reward circuitry for manipulations of outcome magnitudes and probabilities as well as for the overall expected value of the prospect [19–23]. Thus, the findings suggest that the evaluation of risky prospects is not solely performed by higher-level “rational” calculations in the brain but instead both the magnitude of the outcomes and probability of receiving them can be processed via the basic reward mechanisms of the brain.

As prior research indicates, people do not react linearly to anticipated outcomes and their probabilities. The very early work of Bernoulli showed that outcomes are processed in a subjective manner rather than relying on the objective numerical value [1]. Typically the utility associated with one additional unit decreases as the outcome level becomes higher. Intriguingly, the activation in the striatum to anticipated monetary gains appears to have the same property: the incremental additions to the strength of the striatum activity

become smaller as the magnitude of reward increases linearly [24]. This property of the striatal response is in line with the decreasing marginal utility effect which is well-characterized at the behavioral level. This finding implies that the striatum might be involved in the calculation of subjective utility of the reward rather than simply reflecting the absolute reward value. Similarly, to reward magnitude, the striatum does not react linearly to probabilities but instead it overweights small probabilities and underweights large probabilities [25]. Similar nonlinear weighting pattern of probabilities is also apparent in the behavior of decision makers [26]. These results imply that the behavioral deviations from linearity may have their roots in the properties of the reward system, both in terms of the sensitivities of the valuation and probabilistic nature of rewards and punishments.

Thus far, we have discussed the anticipatory reactions related to the expectation of financial incentives. However, it is also the case that this mechanism may underlie decisions of a nonfinancial nature, suggesting that the basic reward system may be a mechanism for most, if not all, of the typical day-to-day decisions we take. For instance, the striatum has been shown to reflect the preference for a variety of consumer products [27,28], as well as encoding the anticipatory value of hedonic outcomes [29]. The striatum activity also reflects modulations in the expected hedonic value in a similar fashion to the way the OFC encodes the influence of price on the experienced pleasantness of wines. In an fMRI study, participants evaluated the expected hedonic value of several potential vacations, after which participants were forced to make a choice between two options of similar desirability to them [29]. Behaviorally, a commitment to one of these previously equally valued options increased the estimated hedonic value of the chosen option, and decreased the valuation of the unselected one. The change in the expected pleasure of the two vacation destinations was also reflected in the striatum, with higher activity here for the chosen option than for the rejected one. Another study indicated that the subjective hedonic expectations are

indeed influenced by the basic reward processing mechanisms, with a pharmacological manipulation which increases dopamine levels, also increasing hedonic expectations for future life events [30].

Reference Dependence of Valuation

Normative decision-making models, such as the expected utility model, assume that valuation is done in respect to the total amount of wealth. One of the main insights of prospect theory [26], a descriptive theory of decision-making under risk, is that decision makers tend to value outcomes in a reference dependent manner. In accordance with this theoretical model, the brain seems to process gains and losses in a similar way. The most dramatic example of reference dependence is that the anticipated and experienced gains and losses activate different brain networks. For instance, the calculation of the expected value of a prospect has been suggested to be performed by the striatum when calculating the expectation of gains [18,22] and by more affective brain structures (amygdala) when calculating the expectation of losses [18]. In addition, the anticipation of a negative outcome has been linked to a brain structure (insula) that is often associated with processing of negative experiences such as pain [17], whereas anticipation of rewards is reflected in the striatum and OFC activity.

The reference dependence of striatum activity has been implied by multiple studies. In one early neuroeconomics study, participants played lotteries while undergoing fMRI. Two types of lotteries were used, one consisting of positive and zero outcome options, and one with negative and zero outcome options. When people received the zero outcome from the negative lottery (a relative gain), striatum activity was higher than when the zero outcome was received from the positive lottery (a relative loss) [31]. Generally, the striatum seems to process the outcomes with respect to subjective expectation [18,32], though some parts of the striatum have recently been shown to reflect reference independent calculation together with some subregions of OFC [33].

In addition to performing calculations with respect to other possible rewards, Tom

and colleagues have shown that the striatum induces nonlinear reactions to gains and losses in a task where participants made choices accepting or rejecting mixed gambles [34]. The researchers found a common network, including the striatum and OFC, that was activated for gains and deactivated for losses. These areas also showed a pattern of loss aversion, namely, that the slope of the deactivation for losses was greater than the slope of activation for gains in a majority of participants. Therefore, the striatum again demonstrates some of the properties of descriptive models of risky choice, showing loss aversion rather than the patterns predicted by classical expected utility maximization. Also of interest was that during decision making, no brain areas were found that were specifically activated by evaluation of increasing losses, in contrast to other studies that have reported increased activity in emotional regions, such as amygdala and insula, for losses [17,18]. One possible explanation for this difference is that the Tom *et al.* study restricted analysis purely to the decision-making phase (decision utility), and excluded anticipatory effects (anticipated utility) by only resolving the lotteries after the fMRI scanning.

Besides the set of possible outcomes, other contexts also influence the valuation process, such as social aspects [35] and numerical representations of financial gains [36]. Fliessbach *et al.* [35] scanned two participants simultaneously with fMRI while they performed a simple visual counting task. Participants received monetary rewards of varying magnitudes for correct performance, whereas incorrect answers had no financial consequences, and participants were informed of the performance and rewards of both players. Of interest were trials where both were correct but one of the participants received a higher reward than the other. Across reward levels, participants' striatum activity was higher when they received more than their partner relative to when they received less than the other player, demonstrating that participants evaluated their outcome relative to that of their partner, as opposed to purely evaluating the payoff.

Summary and Implications

Studies in neuroeconomics have indicated multiple properties in the valuation network of the brain that either support or contradict the assumption of rational evaluation of risky prospects. One important property of the reward circuitry is that it appears to anticipate the utility of future outcomes, providing the decision maker with necessary information for comparing the possible outcome options and selecting the most attractive course of action. Further, this circuit also informs the decision-maker of the desirability of risky prospects by integrating the subjective value and probability of expected outcomes, thus reflecting the overall attractiveness of risky alternatives. These results suggest that the valuation of risky prospects is done in the basic reward circuitry of the brain in accordance with the principles of expected utility maximization. However, when the outcome values and probabilities are manipulated, the reward circuitry does not always behave as predicted by expected utility maximization. For instance, the probabilities are represented in a nonlinear fashion in the valuation process; reward circuitry processes outcomes as gains and losses relative to a reference point; and even arbitrary contexts such as the rewards of others and the numerical presentation of the financial rewards, influence the valuation of outcomes. None of these properties meet the requirements of expected utility maximization but they are reflected in the behavior of decision makers and accounted for in descriptive models of decision making under risk [26].

In sum, the brain evidence indicates that both the evaluation of the attractiveness of risky prospects and the behavioral biases observed in these estimations, are reflected in the reward circuitry of the brain. One implication of these findings is that the evaluation of risky prospects is not controlled by higher-level deliberation, but that instead prospects are valued via more fundamental mechanisms, which also potentially explains the well-observed behavioral deviations from utility maximization.

When interpreting neuroscience findings it is important to take into account the so-called "reverse inference" fallacy, which

refers to the difficulty of inferring mental states of the decision-maker based on the activated brain regions, as one specific brain region can be involved in multiple processes. While there is a growing amount of evidence implicating striatum and OFC in the processing of rewards, it should not be concluded that these areas are reward-meters, where activity can be read off and taken as a metric of positive or negative valence. Despite this caveat, however, the investigation of how the brain computes and processes rewards and punishments offers a very useful window into basic decision making.

FRAMING EFFECT AND THE BRAIN

We have discussed above how behavioral findings such as loss aversion and reference dependent valuation can be observed at the neural level, and here we extend that by discussing another well-characterized behavioral anomaly, the framing effect [37]. The framing effect is a phenomenon where decision makers display different preferences in choice situations which are identical other than for largely irrelevant contextual effects. The effect is often described as resulting from the general tendency of decision makers to be risk averse in the gain domain and risk seeking in the loss domain. For example, imagine the following two scenarios in Fig. 2. In the first scenario you are given \$50 and then asked to make a choice between two options:

keep \$20 of the initial endowment or play a lottery where you have 40% chance of keeping the whole endowment and 60% chance of losing everything. In the second scenario, after you first received the initial endowment of \$50, you are asked to choose: either lose \$30 or participate in a lottery identical to that of the first scenario. Importantly, both of these scenarios are equivalent—they provide a choice between gaining \$20 for sure and having a 40% chance of receiving \$50. However, when people answer these questions separately, they tend to prefer the sure option in the gain domain (scenario 1: risk averse) and the lottery in the loss domain (scenario 2: risk seeking). Thus, the framing of the choice options influences the risk attitude of the decision maker and causes preference reversals. This violates the invariance principle of normative decision-making models which holds that the preference structure is independent of the phrasing of the choice options. The framing effect can be observed in multiple types of decisions and even in participant groups, such as financial planners, who are dealing with risk on a daily basis [38]. As a practical example, when people consider selling their stock investments during an economic slump, it makes a difference whether they think about the money they can save or the money they will lose in case of selling. When people think about the money they could save by selling the stocks, they often prefer the safe option of selling the

	Initial endowment	Choice	Choice in respect to total earnings
Scenario 1	\$50	Option A: keep \$20 Option B: 40% keep all 60% lose all	Option A: \$20 Option B: 40% \$50 60% \$0
Scenario 2	\$50	Option A: lose \$30 Option B: 40% keep all 60% lose all	Option A: $\\$50 - \\$30 = \\$20$ Option B: 40% \$50 60% \$0

Figure 2. Example scenarios for the framing effect. The dashed lines indicate typical choices that people make in line with the framing effect.

stocks and keeping the money (risk averse in gain domain), but when they think about the realized losses related to the sale, they prefer more risky courses of action (risk seeking in loss domain).

One possible explanation for this phenomenon is that the presence of risk in the choice situations elicits emotional reactions which intrude in the decision-making process [39]. Behavioral research suggests that, increasing the emotional salience of the choice situation strengthens nonlinearities in the valuation process [40] which leads to further deviations from risk neutrality [41], thereby strengthening the framing effects. This hypothesis of emotional involvement in framing effects was tested by De Martino *et al.* in an fMRI study, where participants made choices similar to those described above [42]. Choices where participants adhered to the standard bias (safe choice in the gain domain and lottery choice in the loss domain) were compared to those situations where participants chose in the opposite direction (lottery choice in the gain domain, and safe choice in the loss domain). This comparison showed activity in an area related to valuation and processing of emotional stimuli (amygdala; Fig. 3), supporting the hypothesis that emotional processes contribute to the framing effects. The researchers also examined which brain mechanisms were

related to resisting the framing effect, and found activity in a region that is commonly involved in conflict detection and cognitive control (anterior cingulate cortex; Fig. 3). This implies that resistance to the framing bias requires detection of conflict between the rational course of action and an emotional tendency, and regulatory control of the emotional mechanisms to override the bias.

Framing effects are not equally strong in all individuals. In a follow-up study, the researchers were interested in these differences between individuals [43]. They hypothesized that genetic variation might be able to explain these differences, and examined variation in a gene 5-HTTLPR, which has previously been shown to modulate amygdala and anterior cingulate function. The results showed that people with the short allele form of the gene demonstrated strong behavioral framing effects, which were also reflected in the activation in amygdala. In contrast, the people with the long allele variant showed a weaker behavioral effect, which was not reflected in the pattern of amygdala activity. By performing a connectivity analysis between brain regions, the researchers were able to show that the coupling between anterior cingulate and the amygdala increased when participants with long alleles were countering the framing effect. This increase in the connectivity did

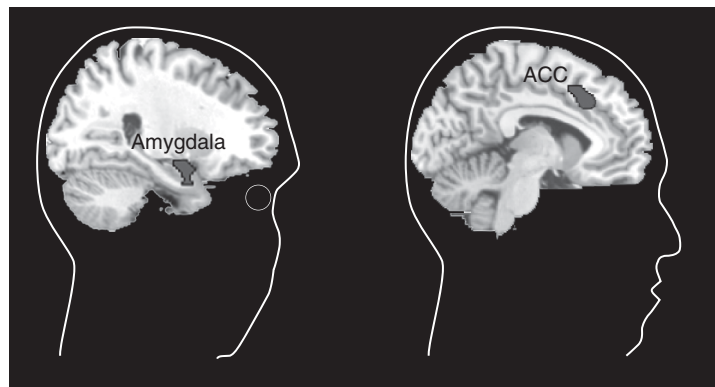


Figure 3. Framing effect is related to activity in amygdala and resisting the frame involves anterior cingulate cortex (ACC) activity. The picture on the right is a slice of the brain that is cut along the midline of the head. The picture on the left is a parallel slice taken from the brain at the level of eyes. These pictures indicate the anatomical locations of the regions of interest, and do not represent data from the actual measurement.

not occur in the subjects with short alleles. On the basis of the results, the researchers claim that the participant group with the long alleles has more efficient dynamic regulatory control over the emotional amygdala reactions than the group with short alleles, which further leads to reduced behavioral framing effects. Thus, the results indicate both the role of emotional reactions and the regulation of those reactions in enhancing and controlling of behavioral framing effects, respectively, and how individual variability in the control mechanisms influences the sensitivity to the framing effects.

Summary and Implications

The fMRI experiments outlined here indicate the role of affective neural processes in driving the behavioral framing effects, and the importance of cognitive control in decreasing them. These findings imply that even when the decision situation does not have any clear emotional context, the presence of risky prospects can activate emotional brain regions. The effect of this emotional activation differs in positive and negative domains: in a positive frame involvement of emotional brain regions increases risk aversion, and in a negative frame the emotional activation leads to risk seeking attitude. Additionally, the susceptibility of individuals to framing effects differs, and these differences can at least partly be explained by genetic variation. These findings suggest that it might be possible to decrease framing effects either by decreasing the emotional reaction induced by risky prospects or by increasing the cognitive control of the decision scenario.

CONCLUSIONS

In this article, we have outlined the importance of both reward circuitry and the balance of emotional and cognitive control in the evaluation of risky prospects, and we have described how these mechanisms can be related to decision biases that have previously been observed in the behavior of decision makers. As this existing research suggests, neuroscience studies can provide insights in the decision-making processes,

and give potentially valuable information for further development of economical models. Further, a better understanding of the neural architecture underlying behavioral biases may provide opportunities to minimize the errors associated with these processes.

This work also suggests that different biases may have different fundamental causes and so might differ in how easy they are to overcome. For example, biases associated with reward mechanisms might be more persistent than the biases that arise from the imbalance of emotional and control signals. Another important role for the striatum is as part of the learning mechanism in the brain [44], which may imply that the biases that are reflected in the striatum activity are more difficult to overcome. One conception of the role of the striatum is that it is not engaged in the processing of rewards and punishments *per se*, but rather that it tracks so-called *reward prediction error*, that is, the difference between what we expect and what we receive [32], thus computing gains and losses in respect to the reference point of expected outcome. If there is a bias in the calculation of the prediction error signals, these biases are continuously reinforced in the behavior, complicating efforts to overcome the resulting behavioral biases. For example, in a loss-aversion account, if outcomes that are worse than expected are followed by an exaggerated reward prediction error signal, this will lead to the learning of loss-averse behavioral patterns for choice situations involving losses.

Despite this however, research indicates that either neural or intentional control mechanisms have the ability to reduce the behavioral biases in certain conditions [45,46]. This may be because cognitive mechanisms can overcome more affective reactions that also enhance the biases. For instance, in the case of loss aversion, there is evidence that points toward the role of emotions in enhancing the bias [28,47], and that the decision maker can reduce the influence of this emotional reaction in their decision making process by intentionally choosing to use an emotion regulation strategy [46]. As discussed earlier in this article, framing effects are also driven by emotional mechanisms. The tendency of decision makers to be

risk averse in gain domains and risk seeking in loss domains appears to be related to increased emotional brain reactions, whereas resisting the biases involves use of cognitive control mechanisms. Those individuals who have better control over their affective responses show smaller framing effects than other people. Thus, these results also indicate the importance of regulation of emotional signals in reducing biases in the decision-making process, and suggest that it is possible that the framing effects in risky decision making could also be reduced by exerting intentional control over the emotional reactions with existing emotion regulation strategies. Overall, the role of emotional reactions in inducing decision biases, and the possibilities to reduce these influences with simple instructions to the decision maker, are important aspects to take into account when designing decision support tools.

An important future step for neuroeconomics is to increase the understanding of individual variability in the sensitivity to decision biases. For developing the most efficient methods to reduce and control the decision biases on an individual level, it is important to learn more about individual differences in choice situations, which will lead to better understanding of the brain mechanisms central to decision making biases. In order to achieve a more complete picture of the neural calculations on an individual level, more attention should be paid to the functioning of larger brain networks involved in the decision making, and in particular, how the different parts of the network communicate with each other. Some initial steps toward this goal have already been taken, for instance, by categorizing people in different groups based on their genotype. Though neuroeconomics is a young field, it offers enormous potential for the better specification of decision-making behavior, and as such promises to open up interesting new avenues for decision analysis in the near future.

REFERENCES

1. Bernoulli D. Exposition of a new theory on the measurement of risk (Sommer L, Trans.). *Econometrica* 1954/1738;22(1):23–36.
2. von Neumann J, Morgenstern O. *Theory of games and economic behavior*. 2nd ed. Princeton (NJ): Princeton University Press; 1947.
3. Huettel SA, Song AW, McCarthy G. *Functional magnetic resonance imaging*. 2nd ed. Sunderland: Sinauer Associates; 2009.
4. Damasio A. *Descartes' error: emotion, reason, and the human brain*. New York: Putman; 1994.
5. Bechara A, Damasio H, Tranel D, *et al.* Deciding advantageously before knowing the advantageous strategy. *Science* 1997; 275(5304):1293–1295.
6. Kringelbach ML. The human orbitofrontal cortex: Linking reward to hedonic experience. *Nat Rev Neurosci* 2005;6(9):691–702.
7. Berns GS, McClure SM, Pagnoni G, *et al.* Predictability modulates human brain response to reward. *J Neurosci* 2001;21(8):2793–2798.
8. O'Doherty JP. Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr Opin Neurobiol* 2004;14(6):769–776.
9. Thut G, Schultz W, Roelcke U, *et al.* Activation of the human brain by monetary reward. *Neuroreport* 1997;8(5):1225–1228.
10. Knutson B, Westdorp A, Kaiser E, *et al.* fMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* 2000;12(1):20–27.
11. O'Doherty J, Kringelbach ML, Rolls ET, *et al.* Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 2001;4(1):95–102.
12. Delgado MR, Locke HM, Stenger VA, *et al.* Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cogn Affect Behav Neurosci* 2003;3(1):27–38.
13. de Araujo IET, Rolls ET, Kringelbach ML, *et al.* Taste-olfactory convergence, and the representation of the pleasantness of flavour, in the human brain. *Eur J Neurosci* 2003;18(7): 2059–2068.
14. Plassmann H, O'Doherty J, Shiv B, *et al.* Marketing actions can modulate neural representations of experienced pleasantness. *Proc Natl Acad Sci U S A* 2008;105(3):1050–1054.
15. Knutson B, Adams CM, Fong GW, *et al.* Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci* 2001;21(16):RC159.
16. Knutson B, Greer SM. Anticipatory affect: neural correlates and consequences for choice.

- Philos Trans R Soc Lond B Biol Sci 2008; 363(1511):3771–3786.
17. Kuhnen CM, Knutson B. The neural basis of financial risk taking. *Neuron* 2005;47(5): 763–770.
 18. Yacubian J, Glascher J, Schroeder K, *et al.* Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. *J Neurosci* 2006;26(37): 9530–9537.
 19. Knutson B, Taylor J, Kaufman M, *et al.* Distributed neural representation of expected value. *J Neurosci* 2005;25(19):4806–4812.
 20. Abler B, Walter H, Erk S, *et al.* Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage* 2006;31(2):790–795.
 21. Preusschoff K, Bossaerts P, Quartz SR. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* 2006;51(3):381–390.
 22. Tobler PN, O'Doherty JP, Dolan RJ, *et al.* Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiol* 2007;97(2):1621–1632.
 23. Yacubian J, Sommer T, Schroeder K, *et al.* Subregions of the ventral striatum show preferential coding of reward magnitude and probability. *Neuroimage* 2007;38(3):557–563.
 24. Pine A, Seymour B, Roiser JP, *et al.* Encoding of marginal utility across time in the human brain. *J Neurosci* 2009;29(30):9575–9581.
 25. Hsu M, Krajbich I, Zhao C, *et al.* Neural response to reward anticipation under risk is nonlinear in probabilities. *J Neurosci* 2009;29(7):2231–2237.
 26. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
 27. Knutson B, Rick S, Wimmer GE, *et al.* Neural predictors of purchases. *Neuron* 2007;53(1): 147–156.
 28. Knutson B, Wimmer GE, Rick S, *et al.* Neural antecedents of the endowment effect. *Neuron* 2008;58(5):814–822.
 29. Sharot T, De Martino B, Dolan RJ. How choice reveals and shapes expected hedonic outcome. *J Neurosci* 2009;29(12):3760–3765.
 30. Sharot T, Shiner T, Brown AC, *et al.* Dopamine enhances expectation of pleasure in humans. *Curr Biol* 2009;19(24):2077–2080.
 31. Breiter HC, Aharon I, Kahneman D, *et al.* Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 2001;30(2):619–639.
 32. Hare TA, O'Doherty J, Camerer CF, *et al.* Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci* 2008;28(22):5623–5630.
 33. De Martino B, Kumaran D, Holt B, *et al.* The neurobiology of reference-dependent value computation. *J Neurosci* 2009;29(12): 3833–3842.
 34. Tom SM, Fox CR, Trepel C, *et al.* The neural basis of loss aversion in decision-making under risk. *Science* 2007;315(5811):515–518.
 35. Fliessbach K, Weber B, Trautner P, *et al.* Social comparison affects reward-related brain activity in the human ventral striatum. *Science* 2007;318(5854):1305–1308.
 36. Weber B, Rangel A, Wibral M, *et al.* The medial prefrontal cortex exhibits money illusion. *Proc Natl Acad Sci U S A* 2009;106(13): 5025–5028.
 37. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981;211(4481):453–458.
 38. Roszkowski MJ, Snelbecker GE. Effects of “Framing” on measures of risk tolerance: financial planners are not immune. *J Behav Econ* 1990;19(3):237–246.
 39. Loewenstein GF, Weber EU, Hsee CK, *et al.* Risk as feelings. *Psychol Bull* 2001;127(2): 267–286.
 40. Hsee CK, Rottenstreich Y. Music, pandas, and muggers: on the affective psychology of value. *J Exp Psychol Gen* 2004;133(1):23–30.
 41. Mukherjee K. A dual system model of preferences under risk. *Psychol Rev* 2010;117(1): 243–255.
 42. De Martino B, Kumaran D, Seymour B, *et al.* Frames, biases, and rational decision-making in the human brain. *Science* 2006;313(5787):684–687.
 43. Roiser JP, de Martino B, Tan GCY, *et al.* A genetically mediated bias in decision making driven by failure of amygdala control. *J Neurosci* 2009;29(18):5985–5991.
 44. Schultz W, Dickinson A. Neuronal coding of prediction errors. *Annu Rev Neurosci* 2000;23:473–500.
 45. Paulus MP, Frank LR. Anterior cingulate activity modulates nonlinear decision weight function of uncertain prospects. *Neuroimage* 2006;30(2):668–677.

46. Sokol-Hessner P, Hsu M, Curley NG, *et al.* Thinking like a trader selectively reduces individuals' loss aversion. *Proc Natl Acad Sci U S A* 2009;106(13):5035–5040.
47. De Martino B, Camerer CF, Adolphs R. Amygdala damage eliminates monetary loss aversion. *Proc Natl Acad Sci U S A* 2010; 107(8):3788–3792.

FURTHER READING

Glimcher PW, Camerer CF, Fehr E, *et al.*, editors. *Neuroeconomics: decision making and the brain*. London, San Diego, Burlington; Academic Press; 2009.

NEWSVENDOR MODELS

NICHOLAS C. PETRUZZI
Department of Business
Administration, University of
Illinois, Champaign, Illinois

MAQBOOL DADA
Carey Business School, The Johns
Hopkins University, Baltimore,
Maryland

A *newsvendor* is a firm that maximizes expected profits in the face of uncertainty. Applied in the context of inventory management, then, a newsvendor is a firm that makes available for sale a chosen quantity of its single product, given that the demand for that product is uncertain. Correspondingly, the core problem faced by the newsvendor is the question of how much quantity to supply.

Historically, the newsvendor problem (or “newsboy” problem as it originally was coined) derives its name from the metaphorical example of a street-corner newsstand stocking its inventory in preparation for the day’s random foot traffic. Because demand is uncertain at the time that the newsvendor commits to a supply for its product, the fundamental principle at the heart of this metaphor is the subtle distinction between quantity *supplied* and quantity *sold*. Indeed, a newsvendor can sell all of its supplied stock if and only if it turns out that total demand for its product meets or exceeds the total quantity supplied. If, in contrast, the quantity supplied exceeds the realized demand, then the quantity that the newsvendor sells will be constrained by the amount demanded. Thus, newsvendor models are defined in essence by the following identity:

$$\text{sales} = \min\{\text{demand}, \text{supply}\}. \quad (1)$$

MANAGING SUPPLY: THE ESSENTIAL NEWSVENDOR

Given Equation (1), let random variable D denote demand, decision variable Q denote

supply, and function $S(Q) = E[\min\{D, Q\}]$ denote expected sales. Moreover, let D be characterized by a continuous cumulative distribution function $F(\xi)$, with corresponding probability density function $f(\xi) = dF(\xi)$. Then, expected sales can be expressed as follows:

$$\begin{aligned} S(Q) &= \int_0^\infty \min\{\xi, Q\} \cdot dF(\xi) \\ &= \int_0^\infty [Q - \max\{Q - \xi, 0\}] \cdot dF(\xi) \\ &= Q - \int_0^Q (Q - \xi) \cdot dF(\xi) \\ &= \int_0^Q d\xi - \int_0^Q F(\xi) \cdot d\xi \\ &= \int_0^Q [1 - F(\xi)] \cdot d\xi, \end{aligned} \quad (2)$$

where the second from last equality in Equation (2) follows from integration by parts.

By definition, expected profit is expected revenue less expected cost. Accordingly, let c denote the newsvendor’s per unit cost of supplying its Q units, and let r denote the per unit revenue associated with selling its $S(Q)$ units. Then, the quintessential form of the newsvendor expected profit function, $\Pi(Q)$, is as follows:

$$\Pi(Q) = r \cdot S(Q) - c \cdot Q. \quad (3)$$

Given Equation (3), the newsvendor’s optimal quantity to supply, denoted Q^* , is the value of Q that maximizes $\Pi(Q)$. To that determination, note from Equation (2) that $S(Q)$ is increasing and concave in Q . As a result, $\Pi(Q)$ is concave, which implies that Q^* is the unique value of Q that satisfies the first-order optimality condition $d\Pi(Q)/dQ = 0$. That is,

$$MR(Q^*) = c, \quad (4)$$

where, from Equations (3) and (2),

$$MR(Q) \equiv d[r \cdot S(Q)]/dQ = r \cdot [1 - F(Q)]. \quad (5)$$

Note that, because $r \cdot S(Q)$ represents a newsvendor's total expected revenue function, $MR(Q)$ correspondingly represents the newsvendor's marginal expected revenue function. Likewise, because $c \cdot Q$ represents a newsvendor's total expected cost function, c correspondingly represents a newsvendor's marginal expected cost function. Therefore, Equation (4) is significant in that it extends a first-principle result of profit maximization to the analog case of uncertainty as follows: Basic microeconomic theory indicates that, in the absence of uncertainty, a firm's optimal supply of a given resource is the level at which the associated marginal revenue is equal to marginal cost. With Equation (4), newsvendor theory thus generalizes this principle by establishing that a firm should set its supply level such that marginal *expected* revenue is equal to marginal *expected* cost.

To hone this intuition, we rearrange Equations (4) and (5) to express the newsvendor's optimal supply level in its more familiar form:

$$1 - F(Q^*) = c/r. \quad (6)$$

Written this way, the ratio c/r is commonly referred to, in operations research literature, as the *critical fractile* (or, *critical ratio*). In this context, c is typically referred to as the *per unit overage cost* because it reflects the economic penalty associated with overstocking. In other words if, *ex post*, a newsvendor's chosen supply (Q) exceeds its random demand (D), then the newsvendor incurs a cost of c for each unit supplied in excess of demand ($Q > D$) that it otherwise would not have incurred. Similarly, $r - c$ typically is referred to as the *per unit underage cost*, because it reflects the net economic penalty associated with understocking. In other words, if *ex post* $Q < D$, then the newsvendor in effect forfeits a revenue of r for each unit demanded in excess of supply, which otherwise would have been sold, although that forfeiture is subsidized in the sense that the newsvendor "saves" a cost of c for each of the units not sold by not

supplying them in the first place. Given these interpretations, notice that Equation (6) can be written as $(r - c) \cdot [1 - F(Q^*)] = c \cdot F(Q^*)$. Thus, one intuitively appealing property of the newsvendor's optimal solution is that it balances the expected cost of overstocking with the expected (opportunity) cost of understocking. Correspondingly, the higher the overage cost, everything else remaining equal, the lower the optimal supply level; and the higher the underage cost, everything else remaining equal, the higher the optimal supply level.

Referring to the left-hand side of Equation (6), notice also that for any given Q , the expression $1 - F(Q)$ can be interpreted as the probability that a newsvendor will sell all of the units that it supplies. Thus Equation (6), in effect, provides an optimal stopping rule to answer the newsvendor's core question of how much quantity to supply. In particular, Equation (6) suggests thinking of a newsvendor's fundamental supply question by asking instead, at what inventory level should the newsvendor cease supplying its product? In other words, at its core, Equation (6) dictates that the question of whether or not to supply a unit boils down to the question of whether or not the unit will sell, if supplied. In a deterministic world, a firm's answer to this question is simple and obvious: If an additional unit is going to sell, then supply it; if it is not going to sell, then do not supply it. However, a newsvendor does not know *ex ante* whether or not a unit supplied will sell. Thus, to a newsvendor, the question is not will the unit sell? The question is, what is the probability that a unit will sell? Intuitively, a newsvendor should supply an additional unit if and only if the probability that the unit will sell is "high" enough. But how high is high enough to justify the supply of the unit? That is where the critical fractile comes in. Simply put, one economic interpretation of the critical fractile is that it represents a sales target, a minimum threshold to establish whether or not a marginal unit should be supplied. As long as the probability of selling a marginal unit (namely, $1 - F(Q)$) meets that threshold, the unit should be supplied. Otherwise, it should not be supplied.

Incorporating Explicit Leftovers and Shortages Effects

The quintessential form of the newsvendor's expected profit function depicted by Equation (3) provides a canonical formulation of the newsvendor problem, in that it effectively captures the fundamental *ex ante* trade-off associated with supplying a unit of its product. A general formulation, however, not only reflects the *ex ante* trade-off associated with the supply decision, but also incorporates potential economic implications of any *ex post* opportunities or repercussions associated with leftovers or shortages. Examples of such direct economic effects include salvage revenues or disposal costs associated with leftovers, and net backordering costs or loss goodwill penalties associated with shortages.

Basically, *leftovers* refers to the excess supply that results if it turns out that supply exceeds demand, and *shortages* refers to the shortfall in supply that results if it turns out that demand exceeds supply. Thus, given Equation (1),

$$\text{leftovers} = \text{supply} - \text{sales}, \quad (7)$$

$$\text{shortages} = \text{demand} - \text{sales}. \quad (8)$$

Accordingly, let $\text{LO}(Q)$ and $\text{SH}(Q)$ denote expected leftovers and expected shortages, respectively, associated with supplying Q units. Moreover, let v denote a per unit salvage value applied to $\text{LO}(Q)$ (where $v < 0$ represents a disposal cost), and let s denote a per unit penalty applied to $\text{SH}(Q)$. Incorporating these elements into Equation (3), then, expands the newsvendor expected profit function to its general form:

$$\begin{aligned} \Pi(Q) = & r \cdot S(Q) + v \cdot \text{LO}(Q) \\ & - c \cdot Q - s \cdot \text{SH}(Q). \end{aligned} \quad (9)$$

Intuitively, Equation (9) depicts the following explicit accounting of supply units and customers: Each of the Q units that a newsvendor supplies *ex ante* incurs a cost of c and then ultimately results in either a sale or a leftover. Those that result in sales ($S(Q)$), each generates a sales revenue of r ; whereas those that result in leftovers ($\text{LO}(Q)$), each

generates a salvage revenue (or disposal cost) of v . In contrast, each of the $\text{SH}(Q)$ units that the newsvendor does not supply *ex ante* effectively invokes a penalty of s as an economic proxy for customer dissatisfaction.

Although Equation (9) provides an explicit accounting of cash flows and penalties associated with a newsvendor's operations, it is convenient for analysis to rewrite Equation (9) by applying Equations (7) and (8) as follows:

$$\hat{\Pi}(Q) = [\hat{r} \cdot S(Q) - \hat{c} \cdot Q] - s \cdot E[D], \quad (10)$$

where $\hat{r} \equiv (r - v) + s$, $\hat{c} \equiv c - v$, and $E[D]$ denotes expected demand. One way to think of Equation (10) is that it provides a reinterpretation of Equation (9) based on the following metaphorical accounting of supply units and customer demands: on the supply side, the newsvendor pays an adjusted cost of $c - v$ for each of the Q units supplied. But, in turn, the newsvendor likewise receives an adjusted revenue of $r - v$ for each of the $S(Q)$ units it sells. On the demand side, the newsvendor pays a penalty s to each of the $E[D]$ customers that seeks its product. That way, if a given customer does not find a unit of the newsvendor's product in stock, that customer appropriately walks away with s as consolation. However, if that customer does find a unit of the newsvendor's product in stock, then no consolation is necessary, in which case the customer returns the s and pays the (adjusted) price $r - v$ in exchange for the unit sold. In other words, looked at from the perspective of the newsvendor's quintessential expected profit function, Equation (3), v effectively translates c into $c - v$ while correspondingly translating r into $r - v$, thus indicating that a newsvendor should define its per unit supply cost and sales revenue net of any salvage (or disposal) effects. And, s effectively translates $r - v$ into $r - v + s$ while correspondingly translating $\Pi(Q)$ into $\Pi(Q) - s \cdot E[D]$, thus indicating that a newsvendor should view the total potential cost of turning away its customer base ($s \cdot E[D]$) as its economic basis for measuring the net return on its inventory investment.

From an optimization perspective, a key implication of Equation (10) is that it is

analogous to Equation (3). In particular, because $E[D]$ is independent of Q , the Q that maximizes $\hat{\Pi}(Q)$ in Equation (10) is the same Q that maximizes $\hat{r} \cdot S(Q) - \hat{c} \cdot Q$. But, this is precisely the same function as $\Pi(Q)$ in Equation (3) except that $\hat{r}$ and $\hat{c}$ replace r and c , respectively. Correspondingly, the Q that maximizes Equation (10) is, analogous to Equation (6),

$$1 - F(Q^*) = \hat{c}/\hat{r}. \quad (11)$$

Thus, the end effect of incorporating explicit economic consequences of leftovers and shortages is a straightforward reinterpretation of the newsvendor's per unit sales revenue and supply cost. Specifically, given v and s , the newsvendor's effective revenue per unit sold is $\hat{r} \equiv (r - v) + s$ and effective cost per unit supplied is $\hat{c} \equiv c - v$. But, given $\hat{r}$ and $\hat{c}$ in place of r and c , respectively, the principle insights and interpretations following from Equation (6) apply directly.

Initial Inventory

Strictly speaking, a newsvendor problem is a single-period problem. Nonetheless, the newsvendor model is foundational to the development of periodic-review decision policies for dynamic inventory control. To help demonstrate basic principles behind this connection, suppose that a newsvendor has x usable units of its product available before establishing its total supply quantity, Q . In this context, $Q \geq x$ refers to the final supply quantity available for possible sale. Thus $S(Q)$, given by Equation (2), represents the expected sales that generate per unit revenue r , but $Q - x$ represents the incremental supply quantity that incurs per unit cost c . In other words, given an initial inventory of x usable units available to the newsvendor before its supply decision, the newsvendor's expected profit function, Equation (3), is adapted as follows:

$$\begin{aligned} \Pi(Q) &= r \cdot S(Q) - c \cdot (Q - x) \\ &= [r \cdot S(Q) - c \cdot Q] + cx. \end{aligned} \quad (12)$$

The corresponding decision problem is to maximize Equation (12) subject to the constraint $Q \geq x$.

This is a constrained optimization problem, but the unconstrained Q that maximizes $\Pi(Q)$ in Equation (12) is the same Q that maximizes $\Pi(Q)$ in Equation (3). Thus, the Q^* implicitly defined by Equation (6) can be interpreted, in this context, as the newsvendor's *desired* supply quantity. However, given the initial inventory constraint, the newsvendor's *actual* supply quantity is Q^* only if $Q^* \geq x$; otherwise, it is x . In other words, as a function of the initial inventory level, x , the newsvendor's optimal supply quantity, denoted by $Q^*(x)$, is as follows:

$$Q^*(x) = \max\{Q^*, x\}, \quad (13)$$

where Q^* is implicitly defined by Equation (6).

More important for the purpose of developing dynamic replenishment policies, however, is the newsvendor's correspondingly optimal expected profit. Accordingly, let $\pi^*(x)$ denote the newsvendor's expected profit given that the optimal supply decision, $Q^*(x)$, is implemented. Then, substituting Equation (13) into Equation (12),

$$\begin{aligned} \pi^*(x) &= \Pi(Q^*(x)) \\ &= \begin{cases} r \cdot S(Q^*) - c \cdot (Q^* - x), & \text{if } x \leq Q^*, \\ r \cdot S(x), & \text{if } x > Q^*. \end{cases} \end{aligned} \quad (14)$$

Note, $\pi^*(x)$ is continuous. Moreover, from Equation (2),

$$\frac{d\pi^*(x)}{dx} = \begin{cases} c, & \text{if } x \leq Q^*, \\ r \cdot [1 - F(x)], & \text{if } x > Q^*, \end{cases}$$

which is continuous and nonincreasing. This implies that, as a function of the initial inventory level x , the newsvendor's expected profit given that an optimal supply policy is followed, is nondecreasing and concave. This property, in particular, provides structure to the Markov Decision Processes that define periodic-review dynamic inventory models. In this respect, the newsvendor model is fundamental to the development of stochastic inventory theory.

Stochastic inventory theory is a vast and important literature linked to a pioneering

era of operations research thought. Indeed, its modern history can be traced back to seminal works and compilations from the 1950s, including, for example, Refs 1 and 2. For authoritative treatments on the current state of the theory, including its important historical developments as well as its scope and applications, we point to Refs 3 and 4. See also the sections titled “Markov Decision Processes,” “Dynamic Programming,” and “Inventory Management and Control” in this encyclopedia.

Decentralization

Strictly speaking, a newsvendor problem is a centralized-control problem. Nonetheless, the newsvendor model is foundational to the development of vertical-control models for decentralized supply chains. To help demonstrate basic principles behind this connection, suppose that the newsvendor is a firm that purchases its supply quantity, Q , from an upstream producer. In this context, Q represents the supply quantity that is first produced upstream at one firm and then transferred downstream to a different firm (i.e., the newsvendor) in exchange for consideration dictated by contractual arrangement between the two firms. Thus, r represents the per unit revenue associated with the *newsvendor's* expected sales of $S(Q)$ units, but c represents the per unit cost associated with the *producer's* production of Q units.

Given this context, first consider a *price-only* contract between the two firms [5]. A price-only contract, also referred to as a *wholesale price* contract, is defined by a single parameter; namely, a per unit wholesale price paid by the newsvendor to the producer for the exchange of the Q units. Accordingly, let w denote such a per unit wholesale price. Then, the newsvendor's expected profits and the producer's (deterministic) profits, denoted here by $\Pi_N(Q)$ and $\Pi_P(Q)$, respectively, are as follows:

$$\Pi_N(Q) = r \cdot S(Q) - w \cdot Q, \quad (15)$$

$$\Pi_P(Q) = w \cdot Q - c \cdot Q. \quad (16)$$

Note, if this were a centralized system in which the two firms were vertically

integrated, then the resulting single firm's expected profits, denoted here by $\Pi_C(Q)$, would be the combination of Equations (15) and (16). That is,

$$\Pi_C(Q) = r \cdot S(Q) - c \cdot Q, \quad (17)$$

which is precisely the expected profit function from Equation (3). Accordingly, the optimal supply quantity for the centralized system, denoted here by Q_C^* , is prescribed directly by Equation (6)

$$1 - F(Q_C^*) = c/r. \quad (18)$$

Nonetheless, in the comparable decentralized supply chain characterized by Equations (15) and (16), the newsvendor chooses its supply quantity not by maximizing Equation (17), but instead by maximizing Equation (15). Thus, the newsvendor's optimal supply quantity in the decentralized supply chain, denoted here by Q_D^* is, by analogy,

$$1 - F(Q_D^*) = w/r. \quad (19)$$

In equilibrium, $w > c$ because from Equation (16), the upstream producer otherwise would not participate in the contract due to its nonprofitability. Thus, a direct comparison of Equations (19) and (18) reveals that $Q_D^* < Q_C^*$. In other words, in the end, a decentralized supply chain in which a price-only contractual arrangement governs the exchange of units between a producer and a newsvendor undersupplies the market as compared to the corresponding centralized system. This phenomenon is called *double marginalization* [6], and its existence fuels the notion of coordination. In short, a decentralized supply chain is *coordinated*, if $Q_D^* = Q_C^*$. Correspondingly, a principle focus of supply chain coordination theory is the question of how to characterize contractual constructs that achieve or reflect coordination.

To that determination, we return to Equations (15) and (16) and replace the price-only contract ($w \cdot Q$) with an abstract contractual construct, denoted simply by $K(Q)$. Applying this abstraction, we thus adapt Equations (15) and (16) as follows:

$$\Pi_N(Q) = r \cdot S(Q) - K(Q), \quad (20)$$

$$\Pi_P(Q) = K(Q) - c \cdot Q. \quad (21)$$

With this recharacterization of the decentralized supply chain as the back-drop, note that one sufficient condition for coordination is for $K(Q)$ to be such that $\Pi_N(Q) = \alpha \cdot \Pi_C(Q) - \beta$, where α and β are constants [7]. The reason for this is as follows: If $\Pi_N(Q) = \alpha \cdot \Pi_C(Q) - \beta$, then $d\Pi_N(Q)/dQ = \alpha \cdot d\Pi_C(Q)/dQ$, which implies that $Q_D^* = Q_C^*$. Accordingly, substituting the desired outcome, $\Pi_N(Q) = \alpha \cdot \Pi_C(Q) - \beta$, into Equation (20), we thus find that $K(Q)$ is a coordinating contract if $K(Q) = r \cdot S(Q) - [\alpha \cdot \Pi_C(Q) - \beta]$. Therefore, given Equation (17), $K(Q)$ is a coordinating contract if it is of the following form:

$$K(Q) = (1 - \alpha) \cdot r \cdot S(Q) + \alpha \cdot c \cdot Q + \beta. \quad (22)$$

One way to interpret this generic form of coordinating contract $K(Q)$ is that it represents a franchise agreement. In particular, Equation (22) is reflective of an agreement in which $\alpha \cdot c \cdot Q$ denotes a transfer payment in exchange for the supply quantity provided, and $\beta + (1 - \alpha) \cdot r \cdot S(Q)$ denotes a fixed royalty payment (β) plus a variable royalty payment $((1 - \alpha) \cdot r \cdot S(Q))$ from franchisee to franchiser. However, various interpretations of Equation (22) are possible, depending on how the parameters α and β are defined. Some common variations on the theme are as follows:

- If $\alpha = 1$ and $\beta > 0$, then $K(Q) = \beta + c \cdot Q$. Specified this way, $K(Q)$ achieves coordination as a *two-part tariff* contract, and is interpreted as follows: In exchange for the transfer of Q units, the newsvendor pays its upstream producer a fixed (β) plus variable (c) price.
- If $\alpha < 1$ and $\beta = 0$, then $K(Q) = (1 - \alpha) \cdot r \cdot S(Q) + \alpha \cdot c \cdot Q$. Specified this way, $K(Q)$ achieves coordination as a *revenue-sharing* contract [7], and is interpreted as follows: In exchange for the transfer of Q units at a wholesale price that is a fraction (α) of the production cost, the newsvendor agrees

to share with the upstream producer, a complementary fraction $(1 - \alpha)$ of its sales revenue.

Alternatively, applying the identity $S(Q) = Q - \text{LO}(Q)$ from Equation (7), this variation can be expressed as $K(Q) = [c + (1 - \alpha) \cdot (r - c)] \cdot Q - (1 - \alpha) \cdot r \cdot \text{LO}(Q)$. Specified this way, $K(Q)$ achieves coordination as a *returns* contract [8]. A returns contract, also referred to as a *buyback* contract, is interpreted as follows: In exchange for the transfer of Q units at a wholesale price that includes a premium $((1 - \alpha) \cdot (r - c))$ over the per unit production cost, the upstream producer agrees to repurchase all unsold units from the newsvendor at a discounted market price $((1 - \alpha) \cdot r)$.

- If $\alpha > 1$, then $K(Q) = \alpha \cdot c \cdot Q - [(\alpha - 1) \cdot r \cdot S(Q) - \beta]$. Specified this way, $K(Q)$ achieves coordination as a *sales-rebate* contract [9,10], and is interpreted as follows: In exchange for the transfer of Q units at a wholesale price that includes a premium (α) over the per unit production cost, the upstream producer agrees to provide a rebate $((\alpha - 1) \cdot r)$ for each unit sold after a sales threshold (β) is met.

Thus, specific interpretations of Equation (22) depend not only on the specified values of α and β , but also on the specified allocation between *ex ante* transfer payments (i.e., monetary exchanges that occur when the supply quantity is transferred between firms) and *ex post* transfer payments (i.e., monetary exchanges that occur after market demand is realized). In that vein, note that, in its generic form, coordinating contract $K(Q)$ can also be interpreted as a *quantity-discount* contract in the following sense: If $\alpha < 1$ and 100% of payment transfer is *ex ante*, then the total value of that exchange, namely, $K(Q)$, increases in Q at a decreasing rate. In other words, from Equations (22) and (2), $d^2K(Q)/dQ^2 = -(1 - \alpha) \cdot r \cdot f(Q) < 0$.

The newsvendor model as a fundamental building block of the theory on supply chain contracting and coordination under uncertainty has received much attention

in operations research and management sciences literatures in recent times, particularly over the last one or two decades. For a comprehensive and insightful perspective on the current state of that theory and its practice, we point to Ref. 11. See also the section titled “Contracts in Supply Chains” in this encyclopedia.

Multiple Sourcing and Other Variants

Strictly speaking, a newsvendor problem is a single sourcing, single location, and single product problem. Nonetheless, the newsvendor model is foundational to myriad variants, extensions, and applications. Although a comprehensive taxonomy of the vast scope and applicability of newsvendor models is beyond the scope of this introductory article, we close this section on the essential newsvendor by outlining a few such variants involving multiple sourcing in particular. For further taxonomic description and analysis, we point to Ref. 12.

As one important variant, consider a network of N newsvendors, each characterized by identical economic parameters but each facing a distinct demand distribution for the same product. In this construct each newsvendor is, in effect, provided a second sourcing opportunity in the form of *ex post* transshipments drawn from leftover inventory remaining at other newsvendors in the network. However, because each newsvendor faces uncertain demand, the leftover inventory from which any given newsvendor draws its transshipments also is random. In such a case, each newsvendor’s optimal solution is dictated by the same critical fractile. Moreover, that critical fractile can be interpreted as the weighted average of the critical fractile that would be associated with a single newsvendor and a critical fractile that is representative of an aggregation of the remaining $N - 1$ newsvendors in the network. See Ref. 13 for details and Ref. 14 for a recent synthesizing analysis. See also the section titled “Warehousing and Cross-Docking” in this encyclopedia.

In a similar vein to transshipment models, the explicit incorporation of second chances to rebalance supply for a given demand population represents another prevalent

variant of the essential newsvendor model. In quick response constructs, in particular, a newsvendor is provided a second sourcing opportunity in the form of an *ex ante* adjustment to its supply quantity after a demand signal is observed but before demand is fully revealed. In such a case, the newsvendor’s optimal solution is again amenable to the critical fractile interpretation. See Refs 15 and 16 for details and Ref. 17 for a recent synthesizing analysis. See also the section titled “Supply Chain Collaboration” in this encyclopedia.

Finally, we note that another reason to incorporate multiple sourcing opportunities into the newsvendor model arises from the notion of supply-side uncertainty. In random yield and unreliable supplier models, in particular, a newsvendor’s actual supply quantity differs from its intended supply quantity. Yet, in such cases, critical fractile solutions continue to prevail to characterize the optimal solution. See Ref. 18 for an influential survey and Ref. 19 for a recent synthesizing analysis.

MANAGING DEMAND: THE PRICE-SETTING NEWSVENDOR

Although the essential newsvendor model characterizes a one-variable problem that emphasizes operational efficiency, it also is foundational to the development of joint-decision models for defining and characterizing the operations/marketing interface. To help demonstrate the basic principles behind this connection, suppose that the newsvendor not only manages its supply quantity by choosing Q , but also influences demand D by choosing its per unit selling price r . In this context, the newsvendor is a *price-setting* newsvendor. Accordingly, the newsvendor’s fundamental question of how much quantity to supply is appended to include the question of what per unit price to set.

Given this context, the newsvendor’s demand is characterized by two effects; namely, randomness (an exogenous effect) and price dependency (an endogenous effect). Therefore, we adapt the formulation of

the newsvendor's expected profit function (Eq. 3) as follows. First, let $D(r, \xi) = y(r) \cdot \xi$ denote demand, where $y(r)$ is a decreasing function defined to represent the price effect on demand, and ξ is a random variable characterized by cumulative distribution function $F(\xi)$ defined to represent the randomness effect on demand. Specified this way, $D(r, \xi)$ denotes a demand function that is multiplicatively separable in its two effects, and can be interpreted such that ξ represents the (uncertain) size of a given market and $y(r)$ represents the (deterministic) fraction of the market that is willing to pay as much as r for a unit of the newsvendor's product. (To provide a contrast, suppose that demand were specified instead, as $D(r, \xi) = y(r) + \xi$, which is a function that is additively separable in its two arguments. Then, one interpretation would be that the demand comprises two distinct markets, a price-sensitive market of (deterministic) size $y(r)$ and a price-insensitive market of (uncertain) size ξ .)

Second, let z denote a transformation of variables defined such that the newsvendor's supply quantity can be written as $Q = y(r) \cdot z$. In this transformation, z represents the newsvendor's stocking factor [20], and is introduced to replace Q as a decision variable. To give stocking factor z intuitive significance, recall from the discussion following Equation (6) that a fundamental consideration of the newsvendor is the probability that a unit, if supplied, will sell. The stocking factor, in essence, is a proxy for that probability because

$$\Pr\{D \geq Q\} = \Pr\{y(r) \cdot z \geq y(r) \cdot \xi\} = 1 - F(z).$$

Finally, let $S(Q, r) = E[\min\{D(r, \xi), Q\}] = y(r) \cdot E[\min\{\xi, z\}]$ denote expected sales, and note that analogous to Equation (2),

$$\begin{aligned} E[\min\{\xi, z\}] &= \int_0^\infty \min\{\xi, z\} dF(\xi) \\ &= \int_0^z [1 - F(\xi)] d\xi. \end{aligned} \quad (23)$$

Then, it readily follows that the price-setting newsvendor's expected profit function, denoted here as $\Pi(z, r)$, is adapted from Equation (3) as follows:

$$\Pi(z, r) = y(r) \cdot [r \cdot E[\min\{\xi, z\}] - c \cdot z]. \quad (24)$$

Given Equation (24), the price-setting newsvendor's optimal stocking factor, denoted z^* , and optimal selling price, denoted r^* , are the values of z and r , respectively, that jointly maximize $\Pi(z, r)$. Correspondingly, first-order optimality conditions require that z^* and r^* satisfy $\partial \Pi(z, r) / \partial z = 0 = \partial \Pi(z, r) / \partial r$, which from Equations (24) and (23) imply that

$$1 - F(z^*) = c/r^*, \quad (25)$$

$$\varepsilon_r(r^*) \cdot [1 - \varepsilon_z(z^*)] = 1, \quad (26)$$

where $\varepsilon_r(r)$ and $\varepsilon_z(z)$ are defined as follows:

$$\varepsilon_r(r) = \frac{-r \cdot dy(r)/dr}{y(r)}, \quad (27)$$

$$\varepsilon_z(z) = \frac{z \cdot [1 - F(z)]}{E[\min\{\xi, z\}]}. \quad (28)$$

Note that Equation (25) is analogous to Equation (6); thus it likewise can be interpreted analogously. To interpret Equation (26), note that Equations (27) and (28) represent economic elasticity measures of the price-setting newsvendor's expected sales. In particular, referring to Equation (23) and the definition of expected sales, notice that $\varepsilon_r(r)$ denotes the positive percentage change in expected sales relative to a percentage change in r , and that $\varepsilon_z(z)$ denotes the positive percentage change in expected sales relative to a percentage change in z . Thus, by definition, $\varepsilon_r(r)$ represents the price-elasticity of expected sales, and $\varepsilon_z(z)$ represents the stocking factor elasticity of expected sales. In that sense, Equation (26) prescribes an optimal balance for the price-setting newsvendor's sales elasticities.

Although Equations (25) and (26) provide necessary conditions for z^* and r^* , the sufficiency of those conditions is not necessarily guaranteed. Basically, sufficiency depends on the specifications of $y(r)$ and $F(z)$. Nonetheless, one robust test for sufficiency, which follows directly from Equations (25) and (26), is as follows [21]: If $y(r)$ is such that $\varepsilon_r(r)$ is nondecreasing in r , and $F(z)$ is such that $\varepsilon_z(z)$

is nonincreasing in z , then Equations (25) and (26) uniquely characterize z^* and r^* .

To help provide insight into this particular sufficiency test, we close by highlighting a useful parallel between the price-setting newsvendor model and a basic formulation of a monopolist's problem. In this context, we define a monopolist as a firm that, like the price-setting newsvendor, influences demand through one or more of its actions, but unlike the price-setting newsvendor, does not face uncertainty in how those actions affect demand.

Regarding the monopolist's problem, let $d(r, z)$ denote a (deterministic) demand function in which r represents selling price and z represents any other control variable that influences demand. (In economics literature, z typically is interpreted as representing "promotion" or "effort" or, simply, "services.") Moreover, let c denote the per unit cost of production, and let $k(z)$ denote the per unit cost of providing z . Then, the corresponding monopolist's problem is to maximize

$$M(r, s) = [r - c - k(z)] \cdot d(r, z). \quad (29)$$

Returning now to the price-setting newsvendor model, notice from Equation (24) that the newsvendor's problem is to maximize

$$\Pi(z, r) = [r - c - k(z)] \cdot y(r) \cdot E[\min\{\xi, z\}],$$

where

$$k(z) = c \cdot (z - E[\min\{\xi, z\}]) / E[\min\{\xi, z\}].$$

But expressed this way, the price-setting newsvendor problem is precisely the problem of the monopolist, as specified by Equation (29), except that it includes a specific functional form for $k(z)$ and it specifies the monopolist's demand function to be $y(r) \cdot E[\min\{\xi, z\}]$. This suggests, first, that the notion of stocking factor is to the newsvendor what the notion of services is to the monopolist, and, second, that the notion of expected sales is to the newsvendor what the notion of demand is to the monopolist. Thus, from an optimization perspective, what the price-setting newsvendor needs in terms of regularity conditions to assure

sufficiency, in essence, is for its expected sales function to inherit the regularity properties stipulated for the monopolist's demand function. For additional insight on this parallel, see Ref. 22. See also the section titled "Marketing/Operations Interface" in this encyclopedia.

REFERENCES

1. Arrow KJ, Harris T, Marschak J. Optimal inventory policy. *Econometrica* 1951;19(3):250–272.
2. Arrow KJ, Karlin S, Scarf H. *Studies in the mathematical theory of inventory and production*. Stanford (CA): Stanford University Press; 1958.
3. Porteus EL. *Foundations of stochastic inventory theory*. Stanford: Stanford Business Books; 2002.
4. Zipkin P. *Foundations of inventory management*. New York: McGraw-Hill/Irwin; 2000.
5. Lariviere MA, Porteus EL. Selling to the newsvendor: an analysis of price-only contracts. *Manuf Serv Oper Manag* 2001;3(4):293–305.
6. Spengler JJ. Vertical integration and anti-trust policy. *J Polit Econ* 1950;58(4):347–352.
7. Cachon GP, Lariviere MA. Supply chain coordination with revenue-sharing contracts: strengths and limitations. *Manag Sci* 2005;51(1):30–44.
8. Pasternack B. Optimal pricing and returns policies for perishable commodities. *Market Sci* 1985;4(2):166–176.
9. Krishnan H, Kapuscinski R, Butz DA. Coordinating contracts for decentralized supply chains with retailer promotional effort. *Manag Sci* 2004;50(1):48–63.
10. Taylor TA. Supply chain coordination under channel rebates with sales effort effects. *Manag Sci* 2002;48(8):992–1007.
11. Cachon GP. Supply chain coordination with contracts. In: de Kok AG, Graves SC, editors. *Handbooks in operations research and management science: supply chain management: design, coordination, and operation*, volume 11. Amsterdam: Elsevier Science; 2003. pp 227–339.
12. Khouja M. The single-period (news-vendor) problem: literature review and suggestions for future research. *Omega: Int J Manag Sci* 1999;27(5):537–553.

13. Krishnan KS, Rao VRK. Inventory control in N warehouses. *J Ind Eng* 1965;16(3):212–215.
14. Wee KE, Dada M. Optimal policies for transshipping inventory in a retail network. *Manag Sci* 2005;51(19):1519–1533.
15. Fisher M, Raman A. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper Res* 1996;44(1):87–99.
16. Eppen GD, Iyer AV. Backup agreements in fashion buying—the value of upstream flexibility. *Manag Sci* 1997;43(11):1469–1484.
17. Li J, Chand S, Dada M, *et al.* Managing inventory over a short season: Models with two procurement opportunities. *Manuf Serv Oper Manag* 2009;11(1):174–184.
18. Yano CA, Lee HL. Lot-sizing with random yields—a review. *Oper Res* 1995;43(2):311–334.
19. Dada M, Petruzzi NC, Schwarz LB. A newsvendor’s procurement problem when suppliers are unreliable. *Manuf Serv Oper Manag* 2007;9(1):9–32.
20. Petruzzi NC, Dada M. Pricing and the newsvendor problem: a review with extensions. *Oper Res* 1999;47(2):183–194.
21. Petruzzi NC, Wee K, Dada M. The newsvendor model with consumer search costs. *Prod Oper Manag* 2009;18(6):693–704.
22. Salinger MA. Mark-up pricing under demand uncertainty. Working paper. 2009 June. Boston, MA: Boston University; 2009.

NEWTON-TYPE METHODS

WALTER MURRAY

Department of Management
Science and Engineering,
Stanford University,
Stanford, California

INTRODUCTION

Newton's method is famous: type it into the search in Youtube and you will get 144 videos with one having had over 30,000 hits. Despite its historic origins it would not be so well known today were it not for its widespread utility. It lies at the heart not just of methods to find roots, but also many other methods such as those for solving partial differential equations (PDEs). Even Karmakar's geometric-based algorithm [1] for linear programming that received such acclaim when it was discovered, was subsequently shown in Gill *et al.* [2] to be equivalent to Newton's method applied to a sequence of barrier functions. Newton's method also serves as a model for other algorithms. It has a theoretical purpose enabling rates of convergence to be determined easily by showing that the algorithm of interest behaves asymptotically, similarly to Newton's method. Naturally, a lot has been written about the method and a classic book well worth reading is that by Ortega and Rheinboldt [3]. Other books that cover the material here and in greater detail are Gill *et al.* [4], Bertsekas [5], and Nocedal and Wright [6].

There would not be so much to read were it not for the fact that Newton's method is only locally convergent. Moreover, as in most cases of algorithms that are only locally convergent it is usually not known *a priori* whether an initial estimate is indeed close enough. Ideally, algorithms should be globally convergent (converge from an arbitrary initial point). However, it should be remembered that proofs of global convergence do

need assumptions about the character of the problem that may not be verifiable for specific problems.

There are many approaches to incorporating Newton's method into a more complex algorithm to ensure global convergence and that is the issue we focus on here. It is somewhat more difficult to deal with the case for Newton's method to minimize a function rather than solving a system of nonlinear equations, so it is that class of problems that we give our main attention. Note that solving $f(x) = 0$ can be transformed into the minimization problem $\min_x \frac{1}{2}f(x)^T f(x)$, but minimizing a function cannot be transformed into solving a system of equations. Simply finding a stationary point by equating the gradient to zero does not do the job. Consequently, Newton's method for optimization, in addition to the deficiencies faced when solving systems of equations, needs to be augmented to enable iterates to move off saddle points. This is the key augmentation that is needed for minimization problems.

Note that there may not be a solution to $f(x) = 0$. Any algorithm will fail in some sense on such problems. It is useful to know how it fails and, should it converge, what are the properties of the point of convergence. Identifying a solution of a minimization is more problematic. We may converge to a point for which we can neither confirm the point is or is not a minimizer. That is another instance of why minimization is a more difficult problem.

Much is made, and rightly so, of Newton's method having a rapid rate of convergence. However, while little can be proved, observations show that when the Newton method converges, it usually behaves well away from the solution. It is the fact that it is a good local model of the behavior of a function that its popularity and success can be attributed. It is also the reason why alternative methods endeavor to emulate Newton's and why it is worthwhile trying to preserve the elements of Newton's method when modifying it to be robust.

We start by addressing the problem in one variable. While it does not highlight all the issues, it does illustrate the critical issue of local convergence being the best that can be achieved. In some of the text we drop the indices that denote the iteration when doing so does not cause confusion.

FUNCTIONS OF ONE VARIABLE

Newton's method for finding the root of a function of one variable is very simple to appreciate. Given some point, say, x_k , we may estimate the root of a function, say $f(x)$, by constructing the tangent to the curve of $f(x)$ at x_k and noting where that linear function is zero. Clearly, for Newton's method to be defined we need $f(x)$ to be differentiable, otherwise the tangent may not exist. Figure 1 shows a single step of Newton's method. Figure 2 illustrates that Newton's method may not give an improved estimate.

Algebraically, the method is that of approximating the nonlinear function at the current iterate by a linear one and using the location of the zero of the linear approximation as the next iterate. Consider

$$f(x) \approx ax + b,$$

where a and b are scalars. Since the two function values and their gradients are equal at x_k we have $a = f'_k \equiv f'(x_k)$ and $b = f_k - f'_k x_k$ giving

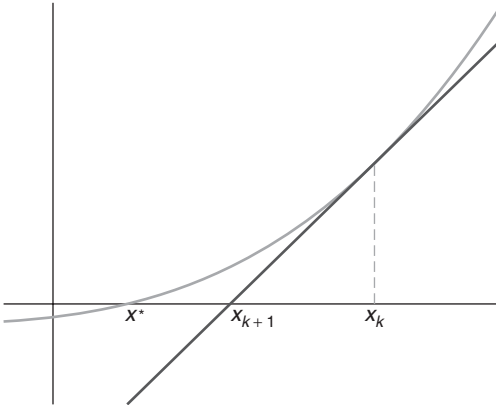


Figure 1. The tangent to $f(x)$ at x_k .

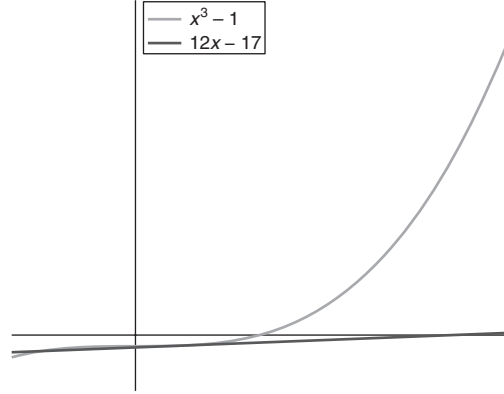


Figure 2. When f'_k is small x_{k+1} may be poor.

$$f(x) \approx f'_k x + f_k - f'_k x_k.$$

It follows that the zero of the linear function, say x_{k+1} , is given by

$$x_{k+1} = x_k - f_k / f'_k. \quad (1)$$

We can now repeat the process replacing x_k by x_{k+1} , which is Newton's method for finding the zero or root of a function of one variable. Note that the iteration given by Equation (1) is only defined if $f'_k \neq 0$ and it is this issue that is the flaw in Newton's method. It is not only when $f'_k = 0$ that is a cause of concern. If $|f'_k|$ is small compared to $|f_k|$, then the step may be so large as to make the new iterate a much worse approximation. It is not hard to construct a sequence of iterates that cycle.

Theorem 1 [Local convergence of Newton's method]. *Let $f(x)$ be a univariate continuously differentiable real-valued function on an open convex set $D \subseteq \mathbb{R}^1$. Assume that $f(x^*) = 0$ for some $x^* \in D$ and that $f'(x^*) \neq 0$. Then there exists an open interval $S \subset D$ containing x^* such that, for any x_0 in S , the Newton iterates (Eq. 1) are well defined, remain in S , and converge to x^* .*

Proof. Let α be a fixed constant in $(0, 1)$. Since f' is continuous at x^* and $f'(x^*)$ is nonzero, there is an open interval $S = (x^* - \epsilon, x^* + \epsilon)$ and a positive constant μ such that

$$\frac{1}{|f'(x)|} \leq \mu \quad \text{and} \quad |f'(y) - f'(x)| \leq \alpha/\mu, \quad (2)$$

for every x and y in S . Suppose that $x_k \in S$. Since $f(x^*) = 0$, it follows from Equation (1) that

$$x_k - x_{k+1} = (f_k - f(x^*) - f'_k(x_k - x^*)) / f'_k.$$

Hence, the first bound in Equation (2) implies

$$|x_{k+1} - x^*| \leq \mu |f_k - f(x^*) - f'_k(x_k - x^*)|.$$

From the integral mean-value theorem we have

$$\begin{aligned} f(x_k) - f(x^*) - f'_k(x_k - x^*) \\ = \int_0^1 (f'(x^* + \xi(x_k - x^*)) - f'_k(x_k - x^*)) d\xi. \end{aligned}$$

Hence,

$$\begin{aligned} |x_{k+1} - x^*| \leq \mu \\ \left\{ \max_{0 \leq \xi \leq 1} |f'(x^* + \xi(x_k - x^*)) - f'_k(x_k - x^*)| \right\} |x_k - x^*|. \quad (3) \end{aligned}$$

Thus, the second bound in Equation (2) gives

$$|x_{k+1} - x^*| \leq \alpha |x_k - x^*|$$

provided $x_k \in S$. Since $\alpha < 1$ and $x_0 \in S$, this last inequality implies that $x_k \in S$ for $k = 1, 2, \dots$, and that $\{x_k\}$ converges to x^* .

Theorem 2 [Rate of convergence of Newton's method]. *If the conditions of Theorem 1 hold, then the sequence $\{x_k\}$ generated by Newton's method converges q -superlinearly to x^* . Moreover, if $f'(x)$ is Lipschitz continuous in S , where S is defined by Theorem 1, with*

$$|f'(x) - f'(x^*)| \leq \gamma |x - x^*|, \quad x \in S, \quad (4)$$

for some constant $\gamma > 0$, then the sequence converges q -quadratically to x^* .

Proof. The linear convergence of the sequence $\{x_k\}$ to x^* is established in Theorem 1. Define

$$\beta_k = \mu \left\{ \max_{0 \leq \xi \leq 1} |f'(x^* + \xi(x_k - x^*)) - f'_k| \right\}, \quad (5)$$

and assume that $x_0 \in S$, with μ defined as in Theorem 1. The continuity of f' and the convergence of $\{x_k\}$ to x^* imply that β_k converges to zero. Since Equation (3) can be written as

$$|x_{k+1} - x^*| \leq \beta_k |x_k - x^*|,$$

the definition of q -superlinear convergence to x^* is satisfied by $\{x_k\}$. Let ξ^* denote the value of ξ for which the maximum in Equation (5) is achieved, and let y_k^* denote $x^* + \xi_k^*(x_k - x^*)$. Then

$$|f'(y_k^*) - f'_k| \leq |f'(y_k^*) - f'(x^*)| + |f'(x^*) - f'_k|. \quad (6)$$

Applying Equation (4) in Equations (6) and (5) gives

$$\beta_k \leq 2\mu\gamma |x_k - x^*|.$$

Hence, $\{x_k\}$ satisfies the definition that the rate of convergence to x^* is at least quadratic.

It should be stressed that Newton's method can converge even if $f'(x^*) = 0$. For example, suppose $f(x) = x^3$. Clearly $x^* = 0$. The Newton iteration is

$$x_{k+1} = x_k - x_k^3 / (3x_k^2) = 2x_k / 3,$$

which obviously converges to 0, but at a linear rate. We can see from Fig. 2 that simply shifting the function to be $x^3 - 1$ does not result in such a smooth set of iterates, although in that case when it converges, it does so at a quadratic rate. What this example also illustrates is that the interval of fast convergence may well be small when $f'(x^*)$ is small. In the example above, x^* is a point of inflection. It could well be that x^* is a local minimizer or a local maximizer of $f(x)$ in which case there exists a neighboring function that does not have a zero in the neighborhood of x^* . Typically, when problems are solved numerically, the best that can be achieved is to

solve exactly a neighboring problem. Even if the iterates converge numerically to a point for which $|f(x)|$ is small, there will remain some uncertainty as to whether a zero exists when $f'(x^*)$ is also very small. In the case where the function changes sign at x^* this may be checked. More importantly, knowing an interval for which the sign of the function at the end points are different provides a means of modifying Newton's method to enable it to find a zero. Specifically, when the Newton step lies outside the interval it needs to be replaced by a point in the interval. A more sophisticated approach is to replace the Newton step even when it lies in the interval. Since we are taking the Newton step from the better of the two endpoints we "expect" the next iterate to be close to that point.

A Hybrid Algorithm

If an interval is known, say $[a, b]$, at which $\text{sign}(f(a)) = -\text{sign}(f(b))$, then a zero may be found by the method of bisection. The function is evaluated at the midpoint of the interval and the end point of the same sign then discarded to give a new interval half the previous size. The bisection point is optimal in the sense it gives the best worst case performance. This approach may be combined with Newton's method by discarding the Newton iterate whenever it lies outside of the interval in which it is known a root lies. Some care needs to be exercised to ensure the algorithm converges and to ensure that the convergence is not very slow. For example, we need to reject the Newton iterate even if it lies within the interval if it is very close to an end point. When performing bisection alone, the optimal placement of the new iterate is the midpoint. When combining the method with Newton, we are assuming that eventually Newton's method will make bisection steps unnecessary. If the bisection step is such that the *best* iterate remains the same, then at the next iteration we will perform yet another bisection step. Consequently, we may wish to choose instead, a point that looks more likely to replace the best iterate by making it closer to that iterate than the bisection point.

Modifying Newton's Method to Ensure Global Convergence

For one variable, the hybrid algorithm described above is probably the best way to go. However, interval reduction does not generalize to n dimensions so it is worthwhile to discuss other means of making the method more robust. A necessary feature we need to ensure is the existence of the iterates. One way of achieving that is to replace the Newton iteration whenever $|f'_k| \leq \delta$, where $\delta > 0$. For example,

$$x_{k+1} = x_k - f_k/\beta, \quad (7)$$

where $\beta = \text{sign}(f'_k)\delta$ if $|f'_k| \leq \delta$, otherwise $\beta = f'_k$. This modification alone does not ensure convergence. What is needed is something that ensures the iterates are improving. One means of defining improvement is a reduction in $f(x)^2$. We could have chosen $|f(x)|$, but $f(x)^2$ has the advantage of being differentiable. The basic idea is to define the iterate as

$$x_{k+1} = x_k - \alpha_k f_k/\beta, \quad (8)$$

where α_k is chosen such that $f_{k+1}^2 < f_k^2$. We need something slightly stronger to prove convergence, but it is enough to choose $\alpha_k = \frac{1}{2}^j$, where j is the smallest index ≥ 1 such that $f(x_k + \frac{1}{2}^j f_k/\beta)^2 < f_k^2$. Determining α_k is a common procedure in n -dimensional problems and more elaborate and more efficient methods are known than the simple backtracking just described. However, they also require more elaborate termination conditions, which we discuss later.

Minimizing a Function of One Variable

Consider now the problem of

$$\min_x f(x).$$

As noted earlier, we could apply Newton's method to $g(x)$, where $g(x) = f'(x)$. The Newton iteration is then given by

$$x_{k+1} = x_k - g_k/g'_k. \quad (9)$$

Clearly, we can only expect this iteration to converge to x^* , where $g(x^*) = 0$. However, if

$g'(x^*) < 0$ we would know that x^* is a maximizer and not a minimizer. Another interpretation is that we are approximating $f(x)$ by a quadratic function, for example

$$f(x) \approx \frac{1}{2}ax^2 + bx + c.$$

The parameters a, b , and c may be determined by the three pieces of information f_k, g_k , and g'_k . The next iterate is defined to be the minimizer of this quadratic function. However, when $a = g'_k < 0$, the quadratic function does not have a minimizer. Even when $a = 0$, unless $b = 0$, this function does not have a minimizer. An essential first step then is to define the Newton iteration for minimizing a function as a modification to Equation (9), namely

$$x_{k+1} = x_k - g_k / |g'_k|. \quad (10)$$

In essence, we reverse the step when g'_k is negative and go in the direction away from the maximizer. To amend the algorithm when $|g'_k| = 0$ or is very small, we can make the same changes as in Equation (7) except $\beta = \delta$ if $|g'_k| \leq \delta$, otherwise $\beta = |g'_k|$. Also α_k is chosen to reduce $f(x)$ rather than $g(x)^T g(x)$. However, there is one more case to consider and that is if $g_k = 0$. In this case, if $g'_k = 0$, we may not be able to make progress. If $g'_k < 0$ then we are at a maximum and a tiny step in either direction from x_k will produce a new iterate x_{k+1} such that $f_{k+1} < f_k$.

As in the case of finding a zero, we could also combine Newton with an interval reduction algorithm such as bisection. We know a minimizer exists in an interval $[a, b]$ if $g(a) > 0$ and $g(b) < 0$. Consequently, we can always maintain an interval containing a minimizer by discarding the appropriate end point. There is a minor complication if $g(x) = 0$ at the new point. Also, if there is more than one minimizer in the interval there may be a choice.

FUNCTIONS WITH N VARIABLES

The extension to n variables is straightforward. The Newton iteration for $f(x) = 0$ becomes

$$x_{k+1} = x_k - J_k^{-1} f_k, \quad (11)$$

where $J_k \equiv J(x_k)$, and $J(x)$ is the Jacobian matrix of $f(x)$. Likewise, the iteration for $\min_x f(x)$ is

$$x_{k+1} = x_k - H_k^{-1} g_k, \quad (12)$$

where $H_k \equiv H(x_k)$, and $H(x)$ is the Hessian matrix of $f(x)$. Obviously, the iterations are only defined if the relevant inverses exist. The conditions for convergence and for the rate of convergence to be at least quadratic in the n -dimensional case is similar to that for the one-dimensional case. Rather than an interval we now require $J(x)$ (and $H(x)$) to be nonsingular in a ball about x^* . Despite being in n -dimensions, the Taylor expansions used in the proofs are still in one dimension since we are relating x_{k+1} to x_k along the vector connecting them.

Rather than define the iterations as in Equations (11) and (12), it is more typical to define them as

$$x_{k+1} = x_k + p_k,$$

where p_k is thought of as the “Newton step.” When modified as

$$x_{k+1} = x_k + \alpha_k p_k,$$

then the Newton step p_k is viewed as a search direction and the scalar $\alpha_k > 0$ is the step length. We are considering the Newton step as a trial to be rejected if it proves worse (in some sense) than the current approximation.

Rather than using inverses, we define p_k as the solution of a system of linear equations, which for solving $f(x) = 0$ becomes

$$J_k p_k = -f_k,$$

and for minimizing a function

$$H_k p_k = -g_k.$$

In the case of minimizing, the iteration only makes sense when H_k is positive definite. Unlike spotting whether g'_k is positive, determining whether a matrix is positive definite is nontrivial. Since H_k is symmetric, if it

was known to be positive definite we would usually determine p_k using Cholesky's algorithm. The algorithm can fail when H_k is indefinite and it is that failure that is the simplest way of determining whether or not H_k is positive definite.

LINESEARCH METHODS

Cast in the manner above, Newton's method may be viewed to be in the class of linesearch methods for minimizing a function [4]. From now on, we shall treat the problem of $f(x) = 0$ as $\min_x \frac{1}{2}f(x)^T f(x)$. A key requirement for linesearch methods is that p_k is a direction of *sufficient descent*, which is defined as

$$\frac{g_k^T p_k}{\|g_k\| \|p_k\|} \leq -\epsilon, \quad (13)$$

where $\epsilon > 0$, and $\|a^2\| \equiv a^T a$. This condition bounds the elements of the sequence $\{p_k\}$ from being arbitrarily close to orthogonality to the gradient. Typically methods are such that p_k is defined in a manner that satisfies Equation (13) even though an explicit value of ϵ is not known.

The other key requirement is that the step length α_k is chosen so as to not be too large or too small. This may be achieved by a suitable termination condition in the search for α_k . The termination criteria used in the optimization routine SNOPT [7] is

$$|g(x_k + \alpha p_k)^T p_k| \leq -\eta g_k^T p_k, \quad (14)$$

where $0 < \eta \leq 1$. Typically, η is chosen in the range $[.1, .9]$. This termination criterion ensures the step α_k is not too small. Usually, it will also ensure it is not too large. However, to be absolutely sure, we need an additional condition and

$$f(x_k + \alpha p_k) \leq f_k + \mu \alpha g(x_k + \alpha p_k)^T p_k, \quad (15)$$

where $0 < \mu < \frac{1}{2}$ will suffice. Typically, μ is chosen quite small and as a consequence, this condition is almost always satisfied once Equation (14) is satisfied. To find a step that satisfies Equation (14), one can search for a minimizer along p_k and terminate at the first

point that Equation (14) is satisfied. Since we are generating iterates that decrease $f(x)$, it follows that Equation (15) is satisfied with μ replaced by $\epsilon_k > 0$.

It can be shown that under modest assumptions, linesearch methods generate a sequence such that $\lim_{k \rightarrow \infty} g_k = 0$. For nonlinear equations, since $g_k = J_k^T f_k$, it implies $f_k \rightarrow 0$ provided J^* is nonsingular. In the minimization case we are only sure we are at a minimizer only if $\lim H_k = H^*$ and H^* is positive definite. When H^* is positive semidefinite with at least one zero eigenvalue and when higher derivatives of $f(x)$ are unknown, there is usually nothing further that can be done to confirm the iterates have converged to a minimizer. Indeed, it is not known just how many orders of higher derivatives will be needed to resolve the issue.

A more general issue is what do in any iteration when either J_k is singular or H_k is indefinite since under such circumstances, the Newton step is not necessarily a sufficient descent direction. If p_k is defined to be the solution of a system

$$B_k p_k = -g_k,$$

where $\{B_k\}$ is a sequence of bounded positive definite symmetric matrices whose condition number is also bounded (smallest eigenvalue bounded away from 0), then $\{p_k\}$ is a sequence of sufficient descent directions. It is easy to see why. Let A be a symmetric matrix with eigenvalues

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n.$$

Let p and g denote vectors in $\Re^n$ such that $Ap = -g$. From the definition of the smallest eigenvalue λ_1 , we have

$$u^T A u \geq \lambda_1 \|u\|^2.$$

From and the relationship $Ap = -g$ and the above inequality we have

$$\begin{aligned} -g^T p / (\|p\| \|g\|) &= p^T A p / (\|p\| \|g\|) \\ &\geq \lambda_1 \|p\| / \|g\|. \end{aligned}$$

Moreover, from norm inequalities we may assert that

$$\|g\| = \|Ap\| \leq \|A\| \|p\| = \lambda_n \|p\|.$$

Using the bound gives

$$-g^T p / (\|p\| \|g\|) \geq \lambda_1 / \lambda_n.$$

Applying this result to each matrix in the sequence $\{B_k\}$ gives

$$\cos \theta_k = -g_k^T p_k / (\|g_k\| \|p_k\|) \geq \lambda_1^{(k)} / \lambda_n^{(k)} \geq 1/M.$$

Consequently, $\{p_k\}$ is a sequence of sufficient descent directions as required. The main idea on how to modify the Newton system defining p_k is to ensure that it is the solution of a system that has the same properties as a B_k .

Consider first the nonlinear eqnarray case. When J_k is nonsingular, the search direction p_k is the solution of

$$J_k^T J_k p_k = -J_k^T f_k.$$

Obviously, if J_k is singular then so is $J_k^T J_k$. However, we may modify the symmetric system above by adding a multiple of the identity giving

$$(J_k^T J_k + \lambda_k I) p_k = -J_k^T f_k.$$

The above system will be nonsingular provided $\lambda_k > 0$. Hence, p_k is always well defined and, provided λ_k is chosen appropriately, is a sufficient descent direction for $\frac{1}{2}f(x)^T f(x)$. There are details that need explaining such as how best to solve this system and how to determine a good value for λ_k , but the issue of convergence is resolved. When $J(x^*)$ is nonsingular then there exists an index, say K , such that for $k \geq K$ we may set $\lambda_k = 0$. By doing so, the terminal search directions are pure Newton's steps.

We have left one detail out and that is the assumption that $\alpha_k = 1$ for $k \geq K$. Since we do not in general know K , when we compute the pure Newton direction in the neighborhood of the solution, we need the linesearch procedure to automatically take the unit step. It is easy to show this will happen provided the initial step taken is 1.

Setting $\alpha = 1$ in the LHS of Equation (14) and noting there exists a constant $M < \infty$ such that $\|g(x_k + p_k)\| \leq M \|g_k\|^2$ when $J(x^*)$ is nonsingular, we get

$$\begin{aligned} |g(x_k + p_k)^T p_k| &\leq \|g(x_k + p_k)\| \|p_k\| \\ &\leq M \|g_k\|^2 \|p_k\|. \end{aligned} \quad (16)$$

Since p_k is a sufficient descent direction, then the RHS of Equation (14) satisfies

$$-\eta g_k^T p_k \geq \eta \epsilon \|g_k\| \|p_k\|. \quad (17)$$

Since $\lim \|g_k\| \rightarrow 0$ it follows there exists K for which the unit step satisfies Equation (14) for $k \geq K$. Note that for the problem $f(x) = 0$ we have $g_k = J_k^T f_k$.

By a similar argument we can show that Equation (15) is also satisfied, which gives us the desired result.

Obtaining a Direction of Sufficient Descent from the Newton Equation

It may be thought that a simple way would be to solve the Newton equations and then reverse the direction if $g^T p > 0$. While this works in the one-variable case, it is flawed in the n -dimensional case. It is not hard to see why although doubtless this approach is still used. Suppose we have two variables and $f(x_1, x_2)$ is a separable function, that is $f(x) = f_1(x_1) + f_2(x_2)$. It follows that the Hessian is a diagonal matrix. Suppose $H_{1,1} > 0$ and $H_{2,2} < 0$ and the Newton step is p . In the x_1 variable, the step p_1 is correctly stepping toward the minimizer, while the step p_2 is stepping away from a minimizer. It could happen that $-p_1 g_1 < p_2 g_2$ in which case p is not a descent direction. What we would do if we knew this was a separable function is reverse the sign of p_2 only. More generally, what we wish to do is to preserve the Newton step in the subspace for which the H is positive definite and reverse it in the subspace it is not. We now explain how to do that. Note that reversing the sign of p does not give an assured descent direction since $p^T g$ could be zero even when g is not zero.

Consider the spectral decomposition: $H = V \Lambda V^T$, where V is the matrix of eigenvectors, $\Lambda = \text{Diag}(\lambda)$, and λ is the vector of

eigenvalues. Define

$$\bar{H} \equiv V\bar{\Lambda}V^T,$$

where $\bar{\lambda}_i = |\lambda_i|$, if $|\lambda_i| \geq \delta$, otherwise $\bar{\lambda}_i = \delta \geq 0$, and $\bar{\Lambda} = \text{Diag}(\bar{\lambda})$. A means of achieving our objective is to define p as

$$p \equiv -\bar{H}^{-1}g.$$

Provided the minimum eigenvalue of H^* is bigger than δ then Newton steps will be taken in the neighborhood of the solution. It may be thought that δ could be allowed to converge to zero as $g_k \rightarrow 0$, which would assure Newton's steps will always be taken in the neighborhood of the solution, provided H^* is positive definite. For example, we could define $\delta \equiv \min\{|g_k|, \epsilon\}$, where $\epsilon \geq 0$. However, in practice p is computed numerically and the error in the computed solution will overwhelm the Newton step when H^* is ill-conditioned. Indeed, the error analysis establishing that Cholesky is a stable algorithm assumes the matrix is sufficiently positive definite.

The drawback of the scheme above is that when n is large it is expensive both in computational effort and memory to compute the spectral decomposition. When H is sufficiently positive definite, it is also unnecessary. Since Cholesky is so cheap it is always worth trying first. An alternative is to compute a direction based on modifying the Cholesky factorization. Cholesky fails on indefinite matrices due to the need to compute a square root of a negative number when computing the diagonal elements of the factor. An obvious change is to define the diagonal elements in the same vein as the modification made to the eigenvalues. If no zero pivots occur, the factorization LDL^T exists, where L is a unit lower triangular matrix and D is diagonal. We can now define $\bar{D}$ in relationship to D in the identical manner to the relationship of $\bar{\Lambda}$ to Λ . It can be shown that $\bar{H} \equiv L\bar{D}L^T = H + E$, where E is also diagonal. In practice, the occurrence of a zero diagonal is of no consequence since should one arise, it is replaced by δ . Unfortunately, it can be shown that the bound on $\|E\|$ is $O(1/\delta)$ [4].

In order to obtain a Cholesky factor that is not close to being unbounded, it is necessary to perform the column form of the algorithm. The basic idea is not simply to alter pivots that are tiny or negative but to alter any pivot for which it is necessary to ensure that the magnitude of the elements of the Cholesky factor are *suitably* bounded. If on first computing an element it exceeds a prespecified bound, say β , then it is recomputed using an increase to the pivot that makes it match the bound. Care needs to be exercised in defining β to ensure that matrices that are sufficiently positive definite are not modified. Let $\bar{R}$ denote the Cholesky factor generated by this algorithm, then β is chosen such that $|\bar{R}_{i,j}| \leq \beta$, where

$$\beta^2 = \max\{\gamma, \xi/n, \delta\},$$

γ and ξ are the largest in modulus of the diagonal and off-diagonal elements of H respectively, and as before δ is some small positive scalar. It can be shown [4] that $\bar{H} \equiv H + E$, E is a diagonal matrix, $\|E\|$ is bounded and $\bar{H}$ has a bounded condition number. The algorithm requires almost the same computational effort as the regular Cholesky algorithm. Let $\bar{R}_k$ denote the Cholesky factor of $\bar{H}_k$. It follows a suitable sequence of sufficient descent directions, $\{p_k\}$, which are obtained by solving

$$\bar{R}_k^T \bar{R}_k p_k = -g_k.$$

When H^* is sufficiently positive definite, then in the neighborhood of the solution $H_k = \bar{H}_k$ and Newton steps are taken.

Generally using a modified direction as defined avoids saddle points. We are deliberately defining the direction not to be a step to such points. However, we could start at one or end up at one by chance. There are also problems [8] that have a form of symmetry that results in the gradient being in the space of the eigenvectors corresponding to the positive eigenvalues of H . It is a consequence of variables whose values, when they are switched, give the same objective. For such problems, the only way of breaking the tie is to use a direction of negative curvature (DNC).

Directions of Negative Curvature

If $g = 0$, then there is no direction of descent and if H is indefinite then x is a saddle point. The key to moving off a saddle point (or avoiding converging to one) is to use directions of negative curvature. A vector d is termed a DNC if $d^T H d < 0$. Obviously, the Hessian needs to be indefinite for such a direction to exist. It may be thought that it is relatively easy to find such a direction and in some cases it is. If H has half its eigenvalues that are negative and equal in value to the positive set then a random direction has a 50% chance of being a DNC. However, it can be shown [9] that the probability is extremely small when the number of positive eigenvalues dominate (approximately 10^{-14} when there are 99 eigenvalues of 1 and one of -1). Typically, in optimization problems there are only a few negative eigenvalues (the iterates are converging to a point the Hessian is positive semidefinite) and even these are small. A single large negative eigenvalue would signify that there is a DNC along which the objective will decrease rapidly and we are rarely that lucky.

Unlike directions of descent, there is a "best" DNC; namely, the eigenvector corresponding to the smallest eigenvalue. If the smallest eigenvalue is not unique, then any vector spanned by the set of corresponding eigenvalues is equally good. It is best in the sense that this direction has the smallest value of $d^T H d / d^T d$ and it is independent of a linear transformation of variables. A DNC predicts a potentially infinite reduction in the objective. Of course, in practice the step will be finite but there is no good initial estimate of what the finite step should be. This contrasts with a direction of descent which is typically derived from a positive definite quadratic model for which the unit step is the predicted to be optimal.

In order to show the iterates will converge to a point for which the Hessian is at least positive semidefinite, it is necessary for the DNC to be a direction of sufficient negative curvature. A sufficient condition for $\{d_k\}$ to be a sequence of directions of sufficient negative curvature is that

$$d_k^T H_k d_k \leq \theta \lambda_{\min_k} d_k^T d_k,$$

where $\lambda_{\min_k}$ is the smallest eigenvalue of H_k , and $\theta > 0$. Obviously $\theta \leq 1$. What the requirement does is to prevent d_k from becoming arbitrarily close to being a direction of zero curvature.

If d_k is a DNC then so is $-d_k$. The sign is chosen so that $d_k^T g_k \leq 0$. If $H(x^*)$ is positive definite, then there exists a index K such that for $k \geq K$ then $d_k = 0$.

A DNC may be computed from the modified Cholesky factorization. It can be shown an index, j , exists (it can be identified during the factorization) such that d , where

$$\bar{R}d = e_j, \quad (18)$$

is a DNC. However, it is not a direction of sufficiently negative curvature. Fortunately, this is of little consequence. Rather than applying the modified Cholesky algorithm to H it is better to apply it to $H + \delta I$, where $\delta > 0$ is suitably small. This modification serves a number of purposes. If this matrix is indefinite, then Equation (18) does give a sufficient DNC. When H^* is singular and positive semidefinite, it avoids tripping the modifications and from a getting a DNC that is not needed in the neighborhood of x^* . Given a DNC, it can be improved by using it as a starting point to minimize $d^T H d / d^T d$; for example, by just using a single pass through all the variables of a univariate search [9].

An issue is how best to use a DNC. An obvious choice is to define the search direction as

$$\bar{p} = p + d,$$

where p is a direction of sufficient descent. It is prudent to have scaled d . There are various schemes that can be used but the general principle is that if $d^T H d / d^T d$ is small then $\|d\|$ should be small.

It can be shown under suitable assumptions that this algorithm converges to a point satisfying the second-order conditions necessary for optimality.

TRUST-REGION METHODS

An alternative to modifying the Hessian is to minimize the quadratic approximation subject to a limit on the size of the step. This

is similar to using the interval to limit the step in problems with one variable. However, unlike the one-variable case, the solution does not usually lie within the limit. If at this new point the objective is not an improvement, then the limit is reduced. We wish to solve

$$\min_s \{\psi(s) \equiv \frac{1}{2}s^T H s + g^T s \mid s^T s \leq \Delta\}, \quad (19)$$

where $\Delta > 0$. Clearly a minimizer exists. It can be shown [6] that there exists a scalar $\lambda \geq 0$ such that the global minimizer, s^* , satisfies

$$(H + \lambda I)s^* = -g, \quad (20)$$

where $H + \lambda I$ is positive semidefinite and $\lambda(s^{*T}s - \Delta) = 0$. It is rare for $H + \lambda I$ to be singular. It is immediately apparent that the definition of s^* is similar to that of p , since the matrix is symmetric, at least positive semidefinite and is a diagonal modification of H . If H is positive definite and the solution of $Hs = -g$ is such that $s^T s \leq \Delta$ then s^* is the Newton step. When $f(x + s^*) \geq f(x)$ the step is rejected and the system (Eq. 20) is resolved with a smaller value of Δ , typically $\Delta \rightarrow \Delta/2$. It can be shown only a finite number of reductions need be taken. If a satisfactory step is obtained without the need to reduce Δ and $\psi(s^*)$ is a reasonable prediction of $f(x) - f(x + s^*)$ then Δ at the next iteration is increased, typically $\Delta \rightarrow 2\Delta$. It can be shown under typical assumptions that the sequence generated by the trust-region algorithm converges to a point satisfying second-order necessary conditions for optimality. It can also be shown that when H^* is positive definite that Δ does not converge to zero. Consequently, in the neighborhood of the solution Newton steps are taken.

In order to solve Equation (20), it is necessary to determine λ . To do so requires solving Equation (20) with estimates of λ . Also if Δ is unsatisfactory, the process has to be repeated, which requires more solutions of Equation (20). It is not necessary to determine λ accurately and often Δ is satisfactory. Nonetheless it is clear a trust-region algorithm which requires more solutions of similar linear systems than a linesearch method.

Note that when $g = 0$ and H is indefinite, then the solution of Equation (20) is the eigenvector corresponding to the minimum eigenvalue (assuming it is unique).

GRADIENT-FLOW METHODS

Unlike the other two classes of methods, these are not in widespread use. The methods were originally proposed by Courant to solve PDEs. Their simplest form is to compute the arc $x(t)$, where $x(t)$ satisfies

$$x'(t) = -g(x(t)) \text{ and } x(0) = x_0. \quad (21)$$

The minimization problem has been replaced by the problem of solving a system of nonlinear differential equations. This approach is grossly inefficient. The method is made reasonable when $g(x(t))$ is approximated at x_k by $g_k + H_k(x(t) - x_k)$. The idea now is compute x_{k+1} by searching along the arc $p(t)$, where $p(t)$ is the solution of a linear system of differential equations:

$$p'(t) = -g_k - H_k p(t) \text{ and } p(0) = 0. \quad (22)$$

The connection to Newton's method is now apparent. When H_k is positive definite, the arc is finite and the end point is the Newton step from x_k . If $H(x^*)$ is positive definite, then when the iterates are in a small neighborhood of the solution, the Newton step is taken. An interesting property of this approach is that the arcs and iterates generated are invariant under an affine transformation, a property that is typically not true for either linesearch or trust-region methods when H_k is not positive definite for all k .

ESTIMATING DERIVATIVES

For minimization problems, Newton-type methods assume the availability of second derivatives. When these derivatives are not available there are alternatives such as quasi-Newton methods. Usually these alternatives are to be preferred although there are problems where using the Newton-type methods described, except with the Hessian approximated by finite differences

of the gradients, are better. A key difference with quasi-Newton methods is that the Hessian approximation in those methods is usually positive definite, hence directions of negative curvature are not available. All the methods discussed may be applied to the case in which $H(x)$ is approximated by finite differences with little difficulty except for the need to compute the differences. The performance of such algorithms is almost indistinguishable, in most cases, from the case when an exact Hessian is used. Of course, the “exact” derivatives do themselves contain errors since they are computed in finite-precision arithmetic. There are not the same numerical issues in approximating second derivatives as there are to approximating first derivatives. It can be shown that when finite differences are used and the finite-difference interval does not converge to zero, which for numerical purposes it cannot, the quadratic rate of convergence is lost. However, this result is misleading. The iterates, except in pathological cases, mimic those of the exact case. Indeed the quadratic convergence of the exact cases also breaks down, but typically at the point the algorithm terminates.

When n is large, the number of differences could grow with n . When $H(x)$ (or $J(x)$ in the nonlinear eqnarray case) is a sparse matrix, there are clever algorithms that can choose the direction of the differences in a manner that significantly reduces the number of differences required [6]. For example, if $H(x)$ is tridiagonal, only two differences are required.

CONSTRAINED PROBLEMS

The general nonlinearly constrained problem may be stated as

$$\min_{x \in \mathbb{R}^n} f(x) \text{ such that } c(x) \geq 0, \quad (23)$$

where $c(x) \in \mathbb{R}^m$ is a vector of smooth functions. It is usual to distinguish linear from nonlinear constraints, and further, to distinguish simple bounds on the variables from linear constraints. Generally, equality constraints are easier to handle than inequality and for simplicity, are often

omitted. For problems with only linear constraints the iterates sequence generated is usually feasible and that has great benefits. Constraints complicate the adaptation of Newton’s method as does size and not surprisingly, large constrained problems are the most difficult of all. It is beyond the scope of this paper to consider all the issues in detail but we shall cover some of the salient issues.

Consider first a problem whose only constraints are simple equalities. By that we mean some subset of the variables are fixed at some value. Clearly, this creates no issues since the problem is unconstrained in the remaining variables. If there are simple inequalities then at each iteration an active-set method fixes some subset of the variables. Consequently, Newton-type methods using an active-set strategy are relatively easy to adapt to such problems. However, consider a trust-region method whose only constraints are $x \geq 0$. Adding the constraints $x_k + s_k \geq 0$ to the subproblem (20) makes the subproblem intractable. Typically the quadratic constraint is replaced by bounds on the step, which then gives a pure bounds constrained subproblem. However, finding the global minimizer of even this subproblem is intractable. It is still possible to find a step that reduces the objective. What is lost is that the modified algorithm is no longer assured of generating a sequence that converges to a point that satisfies the *second-order* necessary conditions.

As noted, for an active-set linesearch method, there are no complications for problems with simple bounds. There are also no complications with general linear constraints provided the problem is not too large. Issues do arise when the problem is large. These are not insurmountable but requires some ingenuity. When we have linear constraints that are active, we need to find a DNC that lies on the constraints or to put it another way, a DNC that is orthogonal to the constraint normals. We also need to modify the *reduced* Hessian rather than the Hessian. Suppose at the k th iteration we wish to remain on a set of constraints $A_k = b_k$. The subproblem we need to solve to obtain the Newton step is

$$\min_p \left\{ \frac{1}{2} p^T H_k p + g_k^T p \mid A_k p = 0 \right\}.$$

Let Z_k be a $n \times (n - m_k)$ full rank matrix such that $A_k Z_k = 0$, where m_k is the number of rows of A_k . We are assuming A_k is full rank. The solution to this problem is given by

$$Z_k^T H_k Z_k p = -Z_k^T g_k. \quad (24)$$

However, if $Z_k^T H_k Z_k$ is not positive definite, it needs to be modified in a manner similar to how H_k is modified in the unconstrained case. There is little difficulty in doing this when n is of modest size. There are not a lot of issues if $n - m_k$ is of modest size even if n is large, although forming the product may be expensive. Fortunately, H_k is typically sparse and an implicit form of Z_k may be found that is also sparse. In some cases, advantage can be made of structure both in A_k and H_k . However, there are cases where both n and $n - m_k$ are large and $Z_k^T H_k Z_k$ is not sparse. Under such circumstances, there are two approaches. One is to consider, instead of using the reduced system (24), solving the equivalent Karush-Kuhn-Tucker system

$$\begin{pmatrix} H_k & A_k \\ A_k^T & 0 \end{pmatrix} \begin{pmatrix} p_k \\ -\lambda_k \end{pmatrix} = \begin{pmatrix} -g_k \\ 0 \end{pmatrix}, \quad (25)$$

where λ_k is the Lagrange multiplier estimate. This matrix is indefinite regardless of the character of H_k . However, it may be deduced from the inertia of this matrix whether or not the reduced Hessian is positive definite. Factorizing this matrix in the form $PLBL^T P^T$, where P is a permutation matrix, L is a unit lower triangular matrix, and B is a 2×2 block-diagonal matrix reveals the inertia. Unfortunately, if the reduced Hessian is not positive definite, this factorization does not reveal how to obtain a suitable sufficient descent direction. How to obtain suitable descent directions by modifying this factorization is described in Forsgren and Murray [10]. They also suggest how to modify the original matrix prior to using an unmodified factorization.

An alternative to factorizations is to attempt to solve Equation (24) using the conjugate gradient (CG) method. Since the

CG algorithm requires only matrix-vector products, the reduced Hessian need not be formed explicitly. The CG method may fail on indefinite matrices, but it is precisely because it may fail that makes it useful. It can also be shown that when it fails, a direction of sufficient descent may be constructed from the partial solution. While it may not give a direction of sufficient negative curvature in general it will usually give a direction of at least zero curvature, which if need be, may be improved by using this as an initial vector in minimizing $d^T Z_k^T H_k Z_k d / d^T d$ either by a few iterations of the Lanczos algorithm or a few iterations of univariate search.

When there are nonlinear inequality constraints, then adapting Newton's method is considerably more problematic except if the nonlinear constrained problem is transformed into solving a sequence of linearly constrained problems. Assuming that is not the case, most methods generate infeasible iterates. As a consequence, the iterate is typically composed of two components: one is a Newton step to get feasible and the other to "reduce" (it may be increasing) the objective. If there are only equality constraints it is somewhat simpler, so we shall describe that first. We are now concerned not with the Hessian of the objective, but with the Hessian of the Lagrangian, where the Lagrangian $L(x, \lambda)$ is defined as $L(x, \lambda) \equiv f(x) - \lambda^T c(x)$. Assuming there are no issues with the reduced Hessian of the Lagrangian, the system of equations we need to solve are

$$g(x) - J(x)^T \lambda = 0 \text{ and } c(x) = 0, \quad (26)$$

which is a system of $n + m$ equations in $n + m$ unknowns. In other words, we are now applying Newton's method in both the primal and dual variables. The Newton step in these variables is given by

$$\begin{pmatrix} H_k^L & J_k \\ J_k^T & 0 \end{pmatrix} \begin{pmatrix} p_k \\ -\Delta \lambda_k \end{pmatrix} = \begin{pmatrix} -g_k + J_k^T \lambda_k \\ -c_k \end{pmatrix}, \quad (27)$$

where H_k^L is the Hessian of the Lagrangian at x_k, λ_k , and $\Delta \lambda_k$ is the Newton step in the

dual variables. The issues in addressing the case when the reduced Hessian is not positive definite in the space of the x -variables is similar to the linearly constrained case. We have no concern about the character of the Hessian in the space of the dual variables. However, since the search is in the space of both variables, there needs to be a search direction defined for those variables; the method is described in Murray and Prieto [11]. Note that we are leaving out the fact we are minimizing a merit function, so care needs to be taken to ensure directions of negative curvature for the Hessian of the Lagrangian are also directions of negative curvature for the merit function. Again, how this may be done is described in Murray and Prieto [11].

The issue with an active-set method for nonlinear constraints is that the most efficient approach is a sequential quadratic programming (SQP) method. Such methods construct a quadratic objective whose Hessian is an approximation to the Hessian of the Lagrangian. When second derivatives are known of both, the original objective and the of the nonlinear constraint functions, only the Lagrange multipliers need be estimated. If the quadratic programming (QP) is not convex, the solution of the QP is not necessarily a direction of sufficient descent. Indeed the QP may not have a bounded solution and even if it does, and the global minimizer is found, that may not yield a direction of descent. In solving the QP, the iterates move on to different sets of active constraints. It is possible for the reduced Hessian to be positive definite on every subspace and yet the solution is not a descent direction. Simply observing the character of the reduced Hessians may not reveal nonconvexity. Unlike the equality-constrained case, we do not know initially the null space of the active set of the constraints since we do not know which constraints are active at the solution of the QP. Moreover, altering the Hessian on that subspace will alter the set of constraints active at the solution. This conundrum may be resolved by monitoring the total step being taken in the QP from the initial feasible point. Assuming the initial reduced Hessian is positive definite (if it is not, the the Hessian may be

modified by applying the modified Cholesky factorization to the reduced Hessian), there are no problems provided only constraints are added. It is when a constraint is deleted that indefiniteness may arise. While it may not arise in the resulting reduced Hessian, the indefiniteness in the direction from the initial point must be checked. While the Hessian still has positive curvature along that direction, a descent direction is assured. When it does not, the situation is similar to that of applying the CG algorithm and it breaks down; the prior step is a descent direction and the new step yields a DNC.

Approximating second derivatives is sometimes cheaper for constrained problems, especially in the linearly constrained case. Since we only need to approximate $Z^T HZ$, then we can approximate HZ directly by differencing along the column vectors of Z . When $n - m$ is small, this is quite efficient. Also if the system (24) is solved using the CG algorithm we need only difference (this holds for the unconstrained case also) along the CG directions, which may be only a few before a sufficiently good search direction is obtained.

A Homotopy Approach

Rather than change the algorithm, another way to extend the range of convergence of Newton's method is to change the problem or rather replace the problem by a sequence of problems. For example, we can use a homotopy approach. The basic idea is to replace the original problem with a parameterized problem. At one end of the parameter range, the problem is easy to solve and at the other end, it is the original problem. Such an approach may be used for globally convergent algorithms also since it can be more efficient than simply solving the original problem directly. Here, our use is not because the problem is necessarily hard to solve but because Newton's method may not work at all. Note that when you do start close to the solution with Newton's method, the problem is usually easy to solve since the iterative sequence converges quickly.

Constructing a homotopy for the case of solving $f(x) = 0$ is particularly easy. Instead of solving $f(x) = 0$, we could solve

$f(x) = \gamma f(x_0)$, where x_0 is our initial estimate, and $0 \leq \gamma < 1$. Obviously if γ is very close to 1, then x_0 is very close to $x^*(\gamma)$, where $x^*(\gamma)$ is a zero of $f(x) = \gamma f(x_0)$. Since we can make γ as sufficiently close to 1 as we wish, we can be within the interval of convergence provided the gradient of $f'(x^*(\gamma)) \neq 0$. It is not necessary to find the solution to the intermediate problems precisely. The main difficulty is assessing how conservative to make the initial choice of γ and how quickly it can be reduced. In one variable, this approach makes the Newton step smaller and as such plays a similar role to α_k in Equation (8). However, note that it is not identical to a fixed scaling of the Newton step.

Another way of modifying functions is to make them more like the ideal case for Newton's method. When solving $f(x) = 0$, that can be done by adding a linear function, for example, solve $(1 - \gamma)f(x) + \gamma(x - x_0) = 0$. Here, we change the gradient of the function as well as move the zero closer to x_0 . A similar approach may be taken for the problem $\min_x f(x)$ by adding a quadratic function.

If at the initial estimate $f'_0 = 0$, then a different initial estimate is needed regardless of the approach being used to solve the problem.

It can be particularly useful to use a homotopy approach when using an SQP algorithm. As noted, when the Hessian is indefinite difficulties arise. Adding a term $\frac{1}{2}\rho(x - \bar{x})^T(x - \bar{x})$ to the objective, where $\bar{x}$ is the initial approximation, adds γI to the Hessian. Not only will this reduce the need to modify the Hessian, it also almost always reduces the number of iterations required to solve the QP subproblem. If γ was very large, then we would be taking a very small optimization step from the initial feasible point of the QP subproblem. It is advisable to change $\bar{x}$ periodically (perhaps even every iteration) and also to reduce γ to zero. When the iterates are close to the solution and the reduced Hessian of the Lagrangian is positive definite, then indefiniteness will not arise when solving the QP. Indeed, close to the solution, only one iteration of the QP is required since the correct active set has been identified. There is much more that could be written in particular on extensions of Newton's method to both,

preserve and improve upon the rate of convergence. It would also be remiss not to cite the classic work of Kantorovich [12].

Acknowledgments

Thanks are due to Nick Henderson for reading the manuscript and to the referees for their prompt reviews and insightful comments. Support for this work was provided by the Office of Naval Research (Grant: N00014-02-1-0076) and the army (Grant: W911NF-07-2-0027-1).

REFERENCES

1. Karmarkar N. A new polynomial time algorithm for linear programming. *Combinatorica* 1984;4(4):373–395.
2. Gill PE, Murray W, Saunders MA, *et al.* On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method. *Math Program* 1986;36(2):183–209.
3. Ortega JM, Rheinboldt WC. Iterative solutions of nonlinear equations in several variables. *SIAM Classics Appl Math* 2000;30.
4. Gill PE, Murray W, Wright MH. Practical optimization. London and New York: Academic Press; 1981.
5. Bertsekas D. Nonlinear programming. Belmont, Mass: Athena Scientific; 1999.
6. Nocedal J, Wright SJ. Numerical optimization. New York: Springer; 2006.
7. Gill PE, Murray W, Saunders MA. SNOPT: an SQP algorithm for large-scale constrained optimization. *SIAM Rev* 2005;47:99–131.
8. Ejov V, Filar JA, Murray W, *et al.* Determinants and longest cycles of graphs. *SIAM J Discrete Math* 2009;22(3):1215–1225.
9. Boman E. Infeasibility and negative curvature in optimization [PhD thesis]. Stanford University. 1999. <http://www.stanford.edu/group/SOL/dissertations.html>.
10. Forsgren A, Murray W. Newton methods for large-scale linear equality-constrained minimization. *SIAM J Matrix Anal Appl* 1993;14:560–587.
11. Murray W, Prieto F. A second-derivative method for nonlinearly constrained optimization. Report SOL 95–3. 1995.48 pp.
12. Kantorovich LV. Functional analysis and applied mathematics. *Usp Math Nauk* 1948;3:89–185.

NON-EXPECTED UTILITY THEORIES

ROBERT F. BORDLEY
General Motors Research and
Development Labs, Warren,
Michigan

Industrial and Operations
Engineering Department,
University of Michigan, Ann
Arbor, Michigan

PARADIGM-PRESERVING VERSUS PARADIGM-SHIFTING THEORIES OF RATIONAL CHOICE

Let S be a partitioning of the possible states of the world and let D be a set of possible decisions whose outcomes depends upon which state s in S obtains. Define a decision d in D as a function, $d[s]$, mapping each state into an outcome.

Savage's seminal book, the *Foundations of Statistics*, considers an individual rational if and only if that individual's preferences over decisions in D satisfy seven consistency postulates [1]. Savage showed that such an individual chooses that decision d with the highest expected utility (EU) where EU is defined by

$$u(d) = \sum_{s \in S} p(s)u(d[s])$$

with $p(s)$ being a probability distribution defined over states s in S and with $u(d[s])$ being a utility function defined over the outcome of making decision d in state s . While Savage's theory is called (subjective) EU theory, it does not presume that rational individuals use probability or utility functions in making their decisions.

Several problems with Savage's EU theory were identified in the second half of the twentieth century. In the *Foundations of Statistics* (from which all quotations are drawn unless otherwise indicated), Savage conceded that

it is undoubtedly true that the behavior of people does often flagrantly depart from the theory (Section 5.6, p. 100–101)

but nonetheless felt that his theory still provided a solid guide to rational behavior. In addition, he noted that

many apparent exceptions to this theory can be interpreted as not to be exceptions at all. For example, a flier may be observed doing a stunt that risks his life, apparently for nothing. That seems to be in contradiction to the theory, but if, in addition, the flier is known to have a real and practical need to convince himself and his colleagues of his courage, then he is simply paying for advertising with the risk of his life, which is not in itself in contradiction to the theory (Section 5.6, p. 101).

Savage considered this creative reinterpretation useful and noted that

The law of the conservation of energy and matter owes its success largely to being an expression of remarkable and reliable facts of nature but to some extent also to certain conventions by which new sorts of energy are so defined as to keep the law true (Section 5.6, p. 101).

For example, in response to Aumann's objection to his axioms, Savage wrote

The difficulties that you mention are all there; I have known about them in a confused way for a long time The theory of personal probability and utility is, as I see it, a sort of framework into which I hope to fit a large class of decision problems. In this process, a certain amount of pushing, pulling and departure from commonsense may be acceptable and even advisable. . . . I believe and examples have confirmed that decision situations can be usefully structured in terms of consequences, states and acts in such a way that the postulates of the Foundations of Statistics are satisfied. Just how to do that seems to be an art for which I can give no useful prescription and for which it is perhaps unreasonable to expect one—as we know for other postulate systems for application. [Letter from Leonard Savage to

R. Aumann, 1971 reprinted in Dreze [2], p. 78–81.)]

But even though Savage did allow for some *departure from commonsense*, Savage recognized that if extreme departures were constantly required to make the theory work, the theory would have to be questioned. As he notes:

In general, the reinterpretation needed to reconcile various sorts of behavior with the utility theory is sometimes quite acceptable and sometimes so strained as to lay whoever proposes it open to the charge of trying to save the theory by rendering it tautological. The same sort of thing arises in connection with many theories and I think there is general agreement that there is no hard-and-fast rule to be laid down as to when it becomes appropriate to make the necessary reinterpretation (Section 5.6, p. 101).

To interpret Savage’s comments, consider Lakatos’s [3] definition of a research program as an evolving body of research based on certain core principles (in this case, Savage’s seven postulates), which remain unchanged while various auxiliary hypotheses are added to the research program to successfully address ever broader sets of problems and applications. This paper refers to these core principles as the paradigm and the addition of auxiliary hypotheses as paradigm-preserving since they enable the research program to address real problems without changing the underlying paradigm.

A research program becomes degenerative if the added hypotheses are ad hoc attempts to defend the paradigm against being falsified by empirical data which do not significantly expand the explanatory power of the research program. A degenerative research program is eventually supplanted by a more successful research program based on a different core/paradigm. We refer to efforts to develop the core of this rival research program as *paradigm-shifting*.

In discussing the various challenges to Savage’s axioms, we will similarly distinguish between

1. Paradigm-preserving research which adds auxiliary hypotheses or reframes

the problem situation so that Savage’s axioms apply and

2. Paradigm-shifting research which addresses the empirical challenges by relaxing or replacing some or all of Savage’s postulates.

This paper focuses on the postulates relaxed or eliminated by the well known non-EU models. As a result, this paper will not discuss Savage’s fourth, fifth, and seventh postulate, even though the fourth postulate was challenged by Shafer [4] and the seventh postulate was challenged by Seidenfeld and Schervish [5].

THE TRANSITIVITY AXIOM: SAVAGE’S FIRST POSTULATE

The transitivity axiom states that if apples are preferred to bananas and if bananas are preferred to cantaloupes then apples must be preferred to cantaloupes. While this condition is generally considered essential even in decision making without uncertainty, violations do occur. For example, experimental subjects tend to choose gambles which have a higher probability of yielding some moderate payoff over riskier gambles with a higher payoff while assigning those riskier gambles a higher monetary value [6]. This “preference” reversal effect is sometimes interpreted as an artifact of the experimental design so that it does not imply intransitive behavior. But Loomes and Sugden [7] and Loomes *et al.* [8] view these (and other) experiments as evidence for intransitivity and have proposed regret theory as an intransitive generalization of Savage’s model.

While Savage’s theory indicates that an individual is only concerned with the consequences of his decision, regret theory indicates that an individual may also be concerned about whether his decision will still be considered the “right” decision after the uncertainty about state s is resolved. Typical formalizations of regret theory specify a regret function r and write the regret utility function as

$$u(d) = \sum_{s \in S} p(s)[u(d[s]) + r[u(d[s]) - u(s)]],$$

so that the utility of a decision given state s is measured not only by its utility in state s but by how that utility compares to the utility that would have been obtained, $u(s)$, had some other decision occurred. In many cases, $u(s)$ is the utility that would have obtained if we had known that state s would occur (and could have made the optimal decision given s occurred). Since $u(s)$ will be a function of the decisions in the choice set, D , an individual's preference between two decisions d^* and d^{**} , might change if we changed the choice set D to a different choice set D^* (where both D and D^* include decisions d and d^*). This often leads to intransitive behavior.

A more general approach [9] defines the relative utility of a consequence $d[s]$ in state s given that consequence $d^*[s^*]$ occurred in state s^* with the skew-symmetric bilinear (SSB) function $W(d[s], d[s^*])$. In their SSB utility theory, the individual chooses d over d^* if the expected value of this function is positive, that is, if

$$\sum_{s \in S} p(s)p(s^*)W(d[s], d[s^*]) > 0.$$

In the paradigm-preserving version of regret theory, the utility of decision d in state s is defined over both the direct consequences of the decision $d[s]$ as well as the indirect consequence of knowing what other consequences might have been achieved with a different decision. Thus, instead of defining regret as a function of utility, the utility function is defined as a function of regret. For example, let $V(d)$ be a random variable describing the possibly uncertain outcomes of decision d . Suppose an individual is working for stakeholders who will only reward the manager if the individual makes the “right choice,” that is, if, out of a possible set of decisions D , the manager chooses that decision which—once the state s is known—provides the most desirable outcome.

Define the random variable T by $T = \text{Max}_{d \in D} V(d)$.

Then the individual will choose that decision d maximizing the probability of $X(d)$ not being less than T . This behavior is fully rational given the individual's incentives and can be incorporated in Savage's

paradigm by redefining the consequence of a decision to include both direct and indirect consequences. But if the stakeholders thought the set of possible decisions was described by the choice set D^* , then the individual might make a different choice. By appropriately varying the choice set D , the individual might then exhibit a set of preferences which seems to violate intransitivity. However, this apparent intransitivity is not inconsistent with Savage's postulates because it arises from real changes in the situation the individual confronts. Thus, the paradigm-preserving interpretation of regret theory avoids violating the transitivity axiom by making an additional assumption about the environment the individual confronts.

THE SURE-THING PRINCIPLE: SAVAGE'S SECOND POSTULATE

Most of the early challenges to Savage's axioms focus on Savage's Sure-Thing Principle. Consider the situation in which two decisions lead to the same set of consequences given event E . The Sure-Thing Principle then concludes that an individual's preference between these two decisions should only be determined by the events where the decisions lead to different consequences. In other words, the individual's preference between the two decisions is independent of the consequences that both decisions incur when event E occurs. Unfortunately, there is ample evidence of a *common-consequence* effect where an individual's preferences between two gambles does change when the *common-consequence* (which both decisions incur when an event E occurs) changes.

Allais [10] presented the first *common-consequence* challenge to the Sure-Thing Principle with an experiment offering individuals two choices. The first choice was between a sure chance of a million dollars (Gamble 1) and a gamble offering a large chance of winning a million dollars, a smaller chance of winning five million and a tiny chance of winning nothing (Gamble 2). The second choice was between two gambles, both of which were likely to yield nothing with one gamble, Gamble

4, offering a chance of five million dollars (vs one million dollars) at a slightly increased risk of winning nothing. As Savage explained,

Many people prefer Gamble 1 to Gamble 2 because, speaking qualitatively, they do not find the chance of winning a very large fortune in place of receiving a large fortune outright adequate compensation for even a small risk of being left in the status quo. Many of the same people prefer Gamble 4 to Gamble 3 because, qualitatively speaking, the chance of winning is nearly the same in both gambles, so the one with the much larger prize seems preferable (Section 5.6, p. 102).

For the specific probabilities used in the Allais Paradox, these choices violated Savage's utility theory. Savage acknowledged that

Examples like the one cited do have a strong intuitive appeal; even if you do not personally feel a tendency to prefer Gamble 1 to Gamble 2 and simultaneously Gamble 4 to Gamble 3, I think that a few trials with other prizes and probabilities will provide you with an example appropriate to yourself... When the two situations were first presented, I immediately expressed preference for Gamble 1 as opposed to Gamble 2 and Gamble 4 as opposed to Gamble 3 and I still feel an intuitive attraction to these preferences (Section 5.6, p. 103).

To overcome this intuitive attraction (which violated his axioms), Savage redefined the four lotteries as defining differing payoffs depending upon which one of three events occurred. Savage felt that this new representation made the folly of violating his axioms more transparent. In this new representation, Savage declared that

I prefer Gamble 1 to Gamble 2 and (contrary to my initial reaction), Gamble 3 to Gamble 4. It seems to me that in reversing my preference between Gamble 3 and Gamble 4, I have corrected an error (Section 5.6, p. 103).

Thus, Savage argues that his own psychological tendency to violate the Sure-Thing Principle in the Allais Paradox could be overcome by appropriately defining a small world

for the Allais Paradox where all gambles had a common support.

A multitude of experiments followed the original experiments by Allais and eventually led to some paradigm-shifting efforts to generalize the Sure-Thing Principle. Machina [11] described the results of these experiments using a Marschak triangle defined over a three-outcome lottery (which gives payoffs with utility 1, y , and 0 with probabilities p, q , and $1 - p - q$.) For utility theory, this lottery has a utility of $U = p + qy$. The indifference curve—defined by those values of p and q which lead to a lottery with utility U —is linear and parallel since $p = U - qy$. However, the experimental results suggest that the indifference curves exhibit a 'fanning out' effect.

To be consistent with this "fanning out" phenomenon, Chew [12] modified Savage's second postulate. This led to a theory where, for some function v , individuals would maximize the following weighted EU function:

$$\sum_{s \in S} p(s)v(d[s])u(d[s]) / \sum_{s \in S} p(s)v(d[s])$$

But Bordley and Hazen [13] showed that this weighted EU function was consistent with Savage's EU theory by making the auxiliary paradigm-preserving assumption of suspicious decision makers. Specifically, suppose the decision maker's subjective probability for state s was initially $p(s)$. The individual is then offered a lottery promising unusually high payoffs (or unusually costly penalties) if state s^* occurs. Since the individual thinks that their wealth is unlikely to change dramatically, the individual could conclude that the chances of s^* occurring are lower than they originally believed. As a result, the individual uses Bayes Rule to lower the initial assessed probability for s^* from $p(s^*)$ to

$$p^*(s^*) = p(s^*) \left[\frac{p(\text{Lottery is offered} | s^*)}{p(\text{Lottery is offered})} \right].$$

If we denote the likelihood function by $v(d[s])$, that is, if we define

$$v(d[s]) = \frac{p(\text{Lottery is offered} | s)}{p(\text{Lottery is offered})},$$

then Chew's weighted EU function is equivalent to Savage's EU theory with $p(s)$ replaced by $p^*(s)$ for each s in S .

Thus, Chew's paradigm-shifting weighted EU theory might be interpretable in a paradigm-preserving way as Savage's model coupled with the auxiliary hypothesis that individuals update their prior beliefs given the event of being offered a lottery. Similar assumptions can also be used to derive the intransitive SSB utility model from EU theory [14].

Another more dramatic modification to Savage's axiom was Machina's [15] approach which simply eliminated the Sure-Thing Principle. This led to a nonlinear EU model with nonlinear indifferent curves and the fanning out property. Machina also observed that a linear approximation of his generalized utility function at a given probability was interpretable as a local EU function that satisfied all the rationality axioms and was applicable to that family of gambles in the neighborhood of this probability distribution.

Harless and Camerer [16] reviewed thousands of empirical tests of Savage's EU theory (EU) versus these (and other) proposed models of how individuals choose. Since increasing the number of estimable parameters in a model often increases its fit without improving its predictiveness, they evaluated each model based on empirical fit and on number of estimable parameters. Their article noted that

There are dramatic differences between theory accuracy when the gambles in a pair have different supports and when they have the same support. EU predicts poorly when the support is different and predicts well when the support is the same . . . When lotteries have different supports, there is never a price-of-precision which justifies using EU. [16], p. 1280.

Now Savage had shown that his own tendency to be misled by the Allais Paradox was eliminated when gambles were defined across common supports. This experience is consistent with Harless and Camerer's [16] finding that EU theory is reasonably descriptive when gambles are defined over a common support. When lotteries have different support, their results favored Kahneman and

Tversky's [17] prospect theory, a psychological theory of choice that, unlike generalized EU, was not explicitly derived from a relaxation of Savage's axioms and was not intended as a normative theory.

In fact, Hammond [18] challenged the normative status of non-EU theories by arguing that an individual violating the Sure-Thing Principle could make a decision at the start of a multistage decision and then "change that decision" in later stages—even in the absence of new information. In response to this argument, non EU theorists argued that standard decision analytic approaches to addressing multistage decision problems were not necessarily appropriate with non EU theory. (Note, of course, that theories which attributed apparent violations of the Sure-Thing Principle to factors like suspicion would be dynamically consistent once the suspicion is taken into account.)

Savage's comments were not intended to suggest that the Sure-Thing Principle is only normatively applicable to gambles defined over the same support. But it does suggest the possibility of creating a non-EU theory which restricts the class of decisions to which the Sure-Thing Principle applies. Building on work by Quiggin [19], Schmeidler [20] developed such a theory. Define a lottery where the individual receives payoff x_i if a state included in event E_i occurs by the vector $(E_1, x_1, \dots, E_n, x_n)$ where events are ordered so that $x_1 > x_2 > \dots > x_n$. Define the probability weighting function

$$q_j = W(E_1 \text{ or } \dots \text{ or } E_j) - W(E_1 \text{ or } \dots \text{ or } E_{j-1})$$

Schmeidler's axioms showed that individuals maximize the Choquet EU [21]:

$$\sum_j q_j u(d[E_j])$$

The preference ordering of decisions implied by Choquet EU theory is consistent with Savage's EU theory when decisions are co-monotonic [22], i.e., when the lotteries describing the uncertainty associated with different decisions order events in the same way [23]. Since Choquet EU theory satisfies

stochastic dominance and transitivity, its proposed transformation of probabilities was used to create an updated version of prospect theory. The resulting cumulative prospect theory [23] remains one of the most popular psychological theories of individual choice.

Diecidue and Wakker [24] interpreted q_j as the individual's "probability" for event E_j adjusted by the individual's perception of the desirability of event E_j occurring. Thus q_j is, in some limited ways, analogous to Bordley and Hazen's paradigm-preserving suspicion-adjusted probabilities.

SEPARATION OF BELIEFS AND VALUES: SAVAGE'S THIRD POSTULATE

Savage viewed decisions as functions that, for each possible state of the world, s in S , led to some consequence. Utility functions, which expressed the individual's values, were defined over consequences while probabilities, which expressed the individual's beliefs, were defined over states. Savage's third postulate argued that if one consequence was preferred to another (e.g., if more money was preferred to less), than a decision that led to the first consequence—if we knew the state of the world was s —would be preferred to a decision that led to the second consequence (given state s). But Savage (Section 2.7, p. 25) noted that the reasonableness of this postulate

depends largely on the interpretation we choose to make of our technical terms, as an example helps to bring out. Before going on a picnic with friends, a person decides to buy a bathing suit or a tennis racket, not having at the moment enough money for both. If we call possession of the tennis racket and possession of the bathing suit consequences, then we must say that the consequences of his decision will be independent of where the picnic is actually held. If the person prefers the bathing suit, this decision would presumably be reversed if he learned that the picnic were not going to be held near water . . . But under the interpretation of 'act' and 'consequence' I am trying to formulate, this is not the correct analysis of the situation. The possession of the tennis racket and the possession of the bathing suit are to be regarded as acts, not consequences . . . The

consequences relevant to the decision are such as these: a refreshing swim with friends, sitting on a shadeless beach twiddling a brand-new tennis racket while one's friends swim, etc.

Thus, the plausibility of the third postulate is contingent on defining consequences in terms of the ultimate objects over which utility is defined. From a practitioner's perspective, this complicates the application of EU theory. It becomes especially difficult in medical decision making where decisions can physically change the individual's ability to enjoy certain consequences. To avoid overly contrived definitions of consequence, Schervish *et al.* [25] and Karni [26] articulated a theory of state-dependent consequences where both probability and utility now depend on states. This can make independently eliciting probabilities and utilities more difficult than in Savage's original formalism.

Now in the history of physics, an alternate formally equivalent representation of Newtonian mechanics, known as *Hamiltonian mechanics*, emerged and was found to be easier to use than Newtonian mechanics in certain applications. Furthermore, quantum mechanics—a paradigm-shifting generalization of Newtonian mechanics—was more easily understood as a generalization of Hamiltonian mechanics than as a generalization of Newtonian mechanics. We now introduce an alternate (and formally equivalent) representation of EU known as *target-oriented utility theory* which likewise makes certain generalizations of EU easier to understand.

For convenience, suppose there is some performance metric v such that the utility of $d[s]$ can be written as some strictly monotonic function of $v(d[s])$. (Using this function v is purely a convenience which is not essential to target-oriented utility theory.) Since Fishburn [27] established that the utility function must be bounded, the utility of $d[s]$ can always be written as the probability that $v(d[s])$ exceeds T for some random variable T . Suppose $V(d)$ is a random variable with the probability of $V(d)$ exceeding v being the probability that decision d leads to an outcome $d[s]$ with $v(d[s]) > v$. Then, target-oriented

utility theory [28–30] writes the EU of decision d as the probability that $V(d)$ exceeds T where T must be uncorrelated with $V(d)$ for all d in D .

Relaxing the assumption of $V(d)$ and T being correlated leads to state-dependent EU theory. Savage had claimed that redefinition of consequences can eliminate state-dependence. To provide an example of where this claim is true, suppose there are random background influences W (e.g., changes in the value of the dollar) that affect both $V(d)$ and T . If there exist uncorrelated random variables $V^*(d)$ and T^* such that $V(d)$ and T are strictly monotonic functions of $V^*(d) + W$ and $T^* + W$, then $u(d)$ can easily be written as the probability that $V^*(d)$ exceeds T^* , that is, as a state-independent EU model. Unfortunately, this transformation only eliminates the simplest cases of state-dependence and hence does not show how state-dependence can be more generally eliminated.

To address violations of the Sure-Thing Principle and transitivity, Becker and Sarin [31] introduced their paradigm-shifting lottery-dependent utility theory. In this model, the EU of a lottery is

$$\sum_{s \in S} p(s)u(d[s], z) \text{ where } z = \sum_{s \in S} p(s)h(d[s])$$

The model states that the individual's EU is a multiattribute utility with two attributes—one determined by the direct consequences of the decision and the other determined by a statistic describing the gamble from which the individual received this consequence. But Castagnoli and LiCalzi [32] and Modesti [33] showed that this model can be reinterpreted as a paradigm-shifting generalization of target-oriented utility theory where T varies with the set of decisions the individual is evaluating.

Oden and Lopes [34] introduced S/A (security/aspiration level) theory where individuals have an uncertain security threshold on what is minimally acceptable and an uncertain threshold on the maximum they aspire to achieve. This is closely related to target-oriented utility theory with an

appropriately chosen random variable T . Oden and Lopes's theory generalizes target-oriented utility theory in a paradigm-shifting way by allowing these thresholds to vary depending upon how a problem is presented. In a series of experiments, Oden and Lopes argued that their theory was as predictively valid as cumulative prospect theory.

But target-oriented utility theory can be viewed as a generalization of Simon's [35] model of the satisfying individual who only tries to meet fixed targets. This generalization is directly applicable to goal-oriented decision makers

1. who are uncertain of what their goals are, or
2. who know their goals but are uncertain about what outcomes are required to achieve their goal.

In this second case, T represents the uncertain requirements for achieving the goal. Examples of such goals include the following:

1. the goal of exceeding the uncertain future performance of a competitor [36];
2. the goal of satisfying uncertain customer expectations [37];
3. a collective goal whose achievement depends on both the decision maker and the uncertain efforts of other team members [38];
4. the Hippocratic goal of making the patient healthier than the patient would be in the absence of medical treatment [39].

From the perspective of target-oriented utility theory, Savage's postulates imply that any rational individual acts like a goal-oriented individual who is uncertain about the requirements for achieving the goal. For certain purposes, this is a more transparent way to think about EU theory.

THE ARCHIMEDEAN AXIOM: SAVAGE'S SIXTH POSTULATE

Fishburn and LaValle [40] felt that Savage's Archimedean axiom was

the only axiom that can be violated in a prescriptively unassailable fashion.

For example, the axiom indicates that an individual should always be willing to take some possibly very small risk of the worst possible outcome for the sake of a modest gain. Hence, if a gamble had a small enough probability of annihilating the human race, an individual should be willing to accept the gamble for the sake of a modest increase in wealth. But Pascal [41] advocated rejecting any decision with even the slightest possibility of an ‘infinitely’ bad outcome.

One paradigm-preserving resolution of this problem presumes that the decision analyst defines the set of possible decisions D to exclude decisions with potential outcomes that are “infinitely worse” than the outcomes of other available decisions. In other words, the decision analyst will not even consider illegal or morally reprehensible decisions. A paradigm-shifting resolution of this problem is Fishburn and LaValle’s [40] lexicographic EU theory which replaces Savage’s scalar-valued EU theory with a theory of vector-valued utilities and matrix-valued probabilities.

But there is another problem with Savage’s sixth postulate—its assumption that any state could be partitioned indefinitely into some grand world in which all distinctions possible have been made in extreme detail. After explicitly noting that this notion of a grand world is *preposterous* and *ridiculous*, Savage writes

any claim to realism about this book . . . or indeed by almost any theory of individual decision making I know . . . is predicated on the idea that some of the individual decision situations into which actual people tend to subdivide the grand world do recapitulate in microcosm the mechanism of the idealized grand world decision . . . The general method of attack I propose to follow, for want of a better one, is to talk in terms of the grand situation . . . tongue in cheek—and in those terms to analyze and discuss isolated decision situations (Section 5.5, p. 83).

Savage will discuss these isolated decision situations by introducing the concept of a small world whose states

are, in fact, events in the grand world, that indeed they constitute a partition of the grand world [and] a restricted set of grand world acts, regarded as small world consequences . . . a small world is complete satisfactory . . . if and only if it satisfies the seven postulates (Section 5.5, p. 84–86).

The small world is called a microcosm if the probability (and utility) in the small world is consistent with the probability and utility in the grand world:

If the small world satisfies the postulates but does not necessarily have a probability or utility consistent with the grand world, call it a pseudo-microcosm (Section 5.5, p. 86).

Savage argues that if the grand world satisfies his axioms, then any small world

can without difficulty be expanded into a somewhat larger small world that is a pseudo-microcosm [which] constitute a Boolean subalgebra of the Boolean algebra of the grand world events (Section 5.5, p. 88).

But, whenever there is a correlation between the uncertainties in two different small worlds, a better than expected (or worse than expected) outcome in one small world could lead to an adjustment in the small world EU which is inconsistent with small world EU maximization [42]. Hence, expanding a small world into a pseudo-microcosm requires some special care. But even if we can successfully construct such pseudo-microcosms, Savage noted that

A pseudo-microcosm is, however, completely satisfactory only if it is actually a microcosm, i.e., leads to a probability measure and a utility well-articulated with those of the grand world (Section 5.5, p. 88).

But how do we know when a pseudo-microcosm is a microcosm? Savage concedes that

Unfortunately the condition . . . is not also automatic. The general condition that a pseudo-microcosm be a microcosm . . . seems incapable of verification without taking the grand world much too seriously (Section 5.5, p. 90).

This raises the possibility, as Shafer noted, that probabilities

calculated in a small world that did satisfy his postulates might change with refinement. If the probabilities are different for two different levels of refinement, then which level is right? (Shafer, p. 384).

But Savage hypothesized that this issue was not a problem in practice, i.e., that individuals could easily determine the right probability and level of refinement:

I feel, if I may be allowed to say so, that the possibility of being taken in by a pseudo-microcosm that is not a real microcosm is remote, but the difficulty I find in defining an operationally applicable criterion is, to say the least, grounds for caution. (Savage 5.5, p. 90).

Unfortunately, empirical data shows that the assessed probabilities for events do vary with the degree of refinement in those events [43–51], even when experienced decision analysts make the assessments [52]. Hence, distinguishing between a microcosm and a pseudo-microcosm is not trivial. As a result

A subjective EU analysis of a decision problem using one small world may fail to give the same results as an analysis using a more detailed small world (Shafer, p. 480).

It is also possible that this problem may not be solvable. Since a state in Savage's grand world could include

The exact and entire present, past and future history of the world understood in any sense, however wide (Section 2.3, p. 8).

Savage's notion of a grand world, taken literally, conflicts with the Heisenberg Uncertainty Principle which rules out the possibility of defining the exact position and momentum of a quantum mechanical particle. While Savage only proposed discussing the grand world *tongue-in-cheek*, Von Neumann's [53] seminal book, the *Mathematical Foundations of Quantum Mechanics*, rules out the meaningfulness of even a *tongue-in-cheek* discussion of the

exact position and momentum of a quantum mechanical particle.

This theoretical issue does have serious practical implications. Specifically, quantum mechanics—while allowing well-defined probabilities over the observable outcomes of a specific experiment (a small world)—defines a complex probability amplitude describing the uncertainty associated with the grand world. Since a complex probability amplitude corresponds to a square two-dimensional matrix (differing from Fishburn and LaValle's matrix-valued probabilities), the probabilities in the small world cannot agree with the matrix-valued probability in the grand world. Hence, microcosms may not exist.

Despite this inconsistency between Savage and von Neumann, much of Savage's work was based on von Neumann and Morgenstern's [54] treatment of decision making with known probabilities:

the treatment of utility as applied to gambles . . . is virtually copied from their book. Indeed their ideas on this subject are responsible for almost all my own. (Section 5.6, p. 97).

The inconsistency arises from Savage's intentionally replacing von Neumann and Morgenstern's known probabilities with DeFinetti's [55] subjective probabilities. DeFinetti long recognized the apparent contradiction between subjectivist theories of probability and von Neumann's theory of quantum probability [56]. But resolving this contradiction would still be of only theoretical interest if these quantum mechanical effects did not arise in the macroscopic work of decision making.

But according to Von Neumann's quantum logic [57], the incompatibility between different quantum mechanical representations of a problem suggests that there is a similar incompatibility between different representations of problems in the social sciences. This hypothesis is supported by empirical studies in the emerging field of quantum cognition [58–63]; and a more recent special issue of the *Journal of Mathematical Psychology* (2009).

SUMMARY

The focus of this review has been on the tension between

1. efforts to address the limitations of EU theory by preserving the paradigm with creative problem definition, target-oriented reformulations of utility theory, auxiliary hypothesis like suspicion etc., and
2. efforts to address these limitations by shifting the paradigm.

The resolution of this debate between paradigm-preserving and paradigm-shifting approaches will ultimately depend upon which approach provides the most insight and convenience in practical decision making [64, 65].

To better focus on this debate, this review has bypassed many important developments in nonEU [66–68] as well as related work showing how Savage's axioms could be simplified by postulating the existence of randomizing devices [69].

REFERENCES

1. Savage L. The foundations of statistics. New York: Wiley; 1954.
2. Dreze J. Essays on economic decisions under uncertainty. Cambridge: Cambridge University Press; 1987.
3. Lakatos I. Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave AE, editors. Criticism and the growth of knowledge. Cambridge: Cambridge University Press; 1970. pp. 91–196.
4. Shafer G. Savage revisited. *Stat Sci* 1986;1:463–501.
5. Seidenfeld T, Schervish M. Conflict between finite additivity and avoiding Dutch book. *Philos Sci* 1983;50(3):398–412.
6. Grether D, Plott C. The economic theory of choice and the preference reversal phenomenon. *Am Econ Rev* 1979;69:623–638.
7. Loomes G, Sugden B. Regret theory: an alternative theory of rational choice under uncertainty. *Econ J Nepal* 1982;92:805–842.
8. Loomes G, Starmer C, Sugden R. Observed violations of transitivity in experiment. *Econometrica* 1991;59(2):425–439.
9. LaValle I, Fishburn P. Decision analysis under states-additive SSB preferences. *Oper Res* 1987;35:722.
10. Allais M. Le comportement de l'homme rationnel devant le risque: critique des postulats et Axioms de l'Ecole Americaine. *Econometrica* 1953;21:503–546.
11. Machina M. Choice under uncertainty: problems solved and unsolved. *J Econ Perspect* 1987;1, (1):121–154.
12. Chew H. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica* 1983;51:1065–1092.
13. Bordley R, Hazen G. SSB and weighted linear utility as EU with suspicion. *Manage Sci* 1990;4(37):396–408.
14. Bordley F. An Intransitive Expectations-Based Bayesian Variant of Prospect Theory. *Risk and Uncertainty* 1992; 5(2):127–144.
15. Machina M. EU analysis without the independence axiom. *Econometrica* 1982;50:277–323.
16. Harless D, Camerer C. The predictive utility of generalized utility theories. *Econometrica* 1994;62:1251–1290.
17. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica* 1979;47:263–267.
18. Hammond PJ. Consequentialist foundations for EU. *Theory Decis* 1988;25:25–78.
19. Quiggin J. A theory of anticipated utility. *J Econ Behav Organ* 1982;3:323–343.
20. Schmeidler D. Subjective probability and EU without additivity. *Econometrica* 1989;57:571–587.
21. Choquet G. Theory of capacity. *Ann Inst Fourier* 1954;5:131–295.
22. Wakker P. The sure-thing principle and the comonotonic sure-thing principle: an axiomatic analysis. *J Math Econ* 1996;25:213–227.
23. Wakker P, Tversky A. An axiomatization of cumulative prospect theory. *J Risk Uncertain* 1993;7:147–176.
24. Diecidue E, Wakker P. On the intuition of rank-dependent utility theory. *J Risk Uncertain* 2001;23(3):281–298.
25. Schervish M, Seidenfeld T, Kadane J. State-dependent utilities. *J Am Stat Assoc* 1990;85:840–847.
26. Karni E. A definition of subjective probability with state-dependent preferences. *Econometrica* 1993;66:187–198.

27. Fishburn P. Utility theory for decision making. New York: John Wiley & Sons, Inc.; 1970.
28. Borch K. Economic objectives and decision problems. *IEEE Trans Syst Sci Cybernetics* 1968;3:266–270.
29. Castagnoli E, LiCalzi M. EU theory without utility. *Theory Decis* 1996;41(3):281–301.
30. Bordley R, Calzi ML. Decision analysis using targets instead of utility functions. *Decis Econ Finance* 2000;23:53–74.
31. Becker J, Sarin R. Lottery-dependent utility theory. *Manage Sci* 1987;33:1367–1382.
32. Castagnoli E, LiCalzi M. Non-EU theories and benchmarking under risk. *Sigma* 1999;29:199–221.
33. Modesti P. Lottery-dependent utility via stochastic benchmarking. *Theory Decis* 2003;55:45–57.
34. Oden G, Lopes L. The role of aspiration level in risky choice a comparison of cumulative prospect theory and SP/A theory. *J Math Psychol* 1999;43(2):286–313.
35. Simon H. Models of man: social and rational: mathematical essays on rational human behavior in social situations. New York: Wiley; 1957.
36. Bordley R, Kirkwood C. Multiattribute performance targets. *Oper Res* 2004;52(6):823–835.
37. Bordley R. Integrating gap analysis and utility theory in service research. *J Health Serv Res Policy* 2001;3(4):300–309.
38. Abbas A, Matheson J, Bordley R. Effective utility functions induced by organizational target-based incentives. *Manage Decis Econ* 2009;30(40):235–251.
39. Bordley R. The hippocratic oath, effect size and utility theory. *J Med Decis Making* 2009;29(3):1–2.
40. Fishburn P, LaValle I. Subjective expected lexicographic utility: axioms and assessments. *Ann Oper Res* 1980;80:183–206.
41. Pascal B. *Pensees*. Paris: Lefauima, Delmas; 1670.
42. Bordley R, Hazen G. Nonlinear utility models arising from unmodelled small world correlations. *Manage Sci* 1992;38(7):1010–1017.
43. Fox CR., Rottenstreich Y. Partitioning priming in judgements under uncertainty. *Psychological Science* 2003;14:195–200.
44. Hammond P, Seidl C. *Handbook of utility theory*. Dordrecht: Kluwer; 1999.
45. Fischhoff B, Slovic P, Lichtenstein S. Fault trees: sensitivity of assessed failure probabilities to problem representation. *Experiment Psychol: Human Perception and Performance* 1978;4:330–344.
46. Redelmeier D, Kohler D, Liberman V, *et al.* Probability judgments in medicine: discounting unspecified alternatives. *Med Decis Making* 1995;15:226–230.
47. Johnson EJ, Hershey J, Meszaros J, *et al.* Framing, probability distortions, and insurance decisions. *J Risk and Uncertainty* 1993;7:35–51.
48. Fox C, Tversky A. Ambiguity aversion and comparative ignorance. *Q J Econ* 1995;110:583–603.
49. Rottenstreich T, Tversky A. Unpacking, repacking and anchoring: advances in support theory. *Psychol Rev* 1997;2:406–415.
50. Tversky A, Koehler D. A non-extensional representation of subjective probability. *Psychol Rev* 1994;101:547–567.
51. See K, Fox C, Rottenstreich Y. Between ignorance and truth: partition-dependence and judgment in learning under uncertainty. *J Clin Exp Neuropsychol* 2006;32(6):1385–1402.
52. Fox C, Clemen R. Subjective probability assessment in decision analysis: partition-dependence and bias toward the ignorance prior. *Manage Sci* 2005;51:417–432.
53. von Neumann J. The mathematical foundations of quantum mechanics. Princeton (NJ): Princeton University Press; 1932.
54. Von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton (NJ): Princeton University Press; 1947.
55. De Finetti B. *Theory of probability*. New York: John Wiley & Sons, Inc.; 1970.
56. von Plato J. *Creating modern probability: its mathematics, physics and philosophy in historical perspective*. Cambridge: Cambridge University Press; 1994.
57. Foulis D, Randall C. Operational statistics i: basic concepts. *J Math Phys* 1972;13(11):1167–1175.
58. Bordley R. Experiment-dependent probabilities in quantum mechanics and psychology. *Phys Essays* 1997;10(3):401–406.
59. Bordley R. Econophysics and individual choice. *Phys A: Stat Mech Appl* 2005;35(4):479.
60. Bordley R, Kadane J. Experiment-dependent priors in psychology and physics. *Theory Decis* 1999;47(3):217–227.
61. Bordley R. Quantum Mechanical and Human Violations of Compound Probability

- Principles: Toward a Generalized Heisenberg Uncertainty Principle. *Oper Res* 1998;46: 923–926.
62. Pathos E, Busemeyer J. A quantum probability explanation for violations of ‘rational’ decision theory. *Proc Roy Soc B: Biol Sci* 2009;1–8.
 63. Busemeyer J, Wang Z, Townsend J. The quantum dynamics of human decision making. *J Math Psychol* 2006;50:220–224.
 64. Edwards W, Miles R, Winterfeldt D. *Advances in decision analysis*. New York: Cambridge University Press; 2007.
 65. Howard R, Matheson J. *Readings on the principles and applications of decision analysis*. Menlo Park (CA): Strategic Decisions Group; 1984.
 66. Gilboa I, Schmeidler D. Maximin EU with a non-unique prior. *J Math Econ* 1989;8:143–153.
 67. Epstein L, Snider M. Recursive multiple priors. *J Econ Theory* 2003;113(1):1–31.
 68. Ellsberg D. Risk, ambiguity and the savage axioms. *Q J Econ* 1961;75:64–669.
 69. Anscombe F, Aumann R. A definition of subjective probability. *Ann Math Stat* 1964;34:199–205.

NONLINEAR CONJUGATE GRADIENT METHODS

YU-HONG DAI
State Key Laboratory of
Scientific and Engineering
Computing, Institute of
Computational Mathematics
and Scientific/Engineering
Computing Academy of
Mathematics and Systems
Science, Chinese Academy of
Sciences, Beijing, P.R. China

INTRODUCTION

Conjugate gradient methods are a class of important methods for solving unconstrained optimization problems

$$\min f(x), \quad x \in R^n, \quad (1)$$

especially if the dimension n is large. They are of the form

$$x_{k+1} = x_k + \alpha_k d_k, \quad (2)$$

where α_k is a stepsize obtained by a line search, and d_k is the search direction defined by

$$d_k = \begin{cases} -g_k, & \text{for } k = 1; \\ -g_k + \beta_k d_{k-1}, & \text{for } k \geq 2, \end{cases} \quad (3)$$

where β_k is a parameter, and g_k denotes $\nabla f(x_k)$.

It is known from Equations (2) and (3) that only the stepsize α_k and the parameter β_k remain to be determined in the definition of conjugate gradient methods. In the case that f is a convex quadratic, the choice of β_k should be such that the methods (2) and (3) reduces to the *linear* conjugate gradient method if the line search is exact, namely,

$$\alpha_k = \arg \min \{f(x_k + \alpha d_k); \alpha > 0\}. \quad (4)$$

For nonlinear functions, however, different formulae for the parameter β_k result in different conjugate gradient methods and their properties can be significantly different. To differentiate the *linear* conjugate gradient method, sometimes we call the conjugate gradient method for unconstrained optimization as *nonlinear* conjugate gradient method. Meanwhile, the parameter β_k is called *conjugate gradient parameter*.

The linear conjugate gradient method can be dated back to a seminal paper by Hestenes and Stiefel [1] in 1952 for solving a symmetric positive definite linear system $Ax = b$, where $A \in R^{n \times n}$ and $b \in R^n$. An easy and geometrical interpretation of the linear conjugate gradient method can be founded in Shewchuk [2]. The equivalence of the linear system to the minimization problem of $\frac{1}{2}x^T Ax - b^T x$ motivated Fletcher and Reeves [3] to extend the linear conjugate gradient method for nonlinear optimization. This work of Fletcher and Reeves in 1964 not only opened the door to the nonlinear conjugate gradient field but greatly stimulated the study of nonlinear optimization. In general, the nonlinear conjugate gradient method without restarts is only linearly convergent [4], while n -step quadratic convergence rate can be established if the method is restarted along the negative gradient every n -step [5,6]. Some recent reviews on nonlinear conjugate gradient methods can be found in Hager and Zhang [7], Nazareth [8,9], and Nocedal [10,11]. This article aims to provide a perspective view on the methods from the angle of descent property and global convergence.

Since exact line search is usually expensive and impractical, the strong Wolfe line search is often considered in the implementation of nonlinear conjugate gradient methods. It aims to find a stepsize satisfying the strong Wolfe conditions

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \rho \alpha_k g_k^T d_k, \quad (5)$$

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma g_k^T d_k, \quad (6)$$

where $0 < \rho < \sigma < 1$. The strong Wolfe line search is often regarded as a suitable extension of the exact line search since it reduces to the latter if σ is equal to zero. In practical computations, a typical choice for σ that controls the inexactness of the line search is $\sigma = 0.1$.

On the other hand, for a general nonlinear function, one may be satisfied with a step-size satisfying the standard Wolfe conditions, namely, Equation (5) and

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \quad (7)$$

where again $0 < \rho < \sigma < 1$. As is well known, the standard Wolfe line search is normally used in the implementation of quasi-Newton methods, another important class of methods for unconstrained optimization. The work of Dai and Yuan [12,13] indicates that the use of standard Wolfe line searches is possible in the nonlinear conjugate gradient field. Besides this, there are quite a few Refs 14,15,16,17 that deal with Armijo-type line searches.

A requirement for an optimization method to use the above line searches is that, the search direction d_k must have the descent property, namely,

$$g_k^T d_k < 0. \quad (8)$$

For conjugate gradient methods, by multiplying Equation (3) with g_k^T , we have

$$g_k^T d_k = -\|g_k\|^2 + \beta_k g_k^T d_{k-1}. \quad (9)$$

Thus if the line search is exact, we have $g_k^T d_k = -\|g_k\|^2$ since $g_k^T d_{k-1} = 0$. Consequently, d_k is descent provided $g_k \neq 0$. However, this may not be true in case of inexact line searches for early conjugate gradient methods. A simple restart with $d_k = -g_k$ may remedy these bad situations, but will probably degrade the numerical performance since the second-derivative information along the previous direction d_{k-1} is discarded [18]. Assume that no restarts are used. In this article we say that, a conjugate gradient method is *descent* if Equation (8) holds for all k , and is *sufficient descent* if the sufficient descent condition

$$g_k^T d_k \leq -c \|g_k\|^2, \quad (10)$$

holds for all k and some constant $c > 0$. However, we have to point out that the borderlines between these conjugate gradient methods are not strict (see the discussion at the beginning of the section titled “Sufficient Descent Conjugate Gradient Methods”).

This survey is organized in the following way. In the next section, we will address two general convergence theorems for the methods of the form (2) and (3) assuming the descent property of each search direction. Afterwards, we divide conjugate gradient methods into three categories: early conjugate gradient methods, descent conjugate gradient methods, and sufficient descent conjugate gradient methods. They will be discussed in sections titled “Early Conjugate Gradient Methods”, “Descent Conjugate Gradient Methods”, and “Sufficient Descent Conjugate Gradient Methods”, respectively, with the emphases on the Fletcher–Reeves (FR) method, the Polak–Ribière–Polyak (PRP) method, the Hestenes–Stiefel (HS) method, the Dai–Yuan method, and the CG_DESCENT method by Hager and Zhang. Some research issues on conjugate gradient methods are mentioned in the last section.

GENERAL CONVERGENCE THEOREMS

In this section, we give two global convergence theorems for any methods of the form (2) and (3) assuming the descent condition (8) for all k . The first one deals with the strong Wolfe line search, while the second treats the standard Wolfe line search.

At first, we give the following basic assumptions on the objective function. Throughout this article, the symbol $\|\cdot\|$ denotes the two norm.

Assumption 1. (i) The level set $\mathcal{L} = \{x \in R^n : f(x) \leq f(x_1)\}$ is bounded, where x_1 is the starting point; (ii) in some neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable, and its gradient is Lipschitz continuous; namely, there exists a constant $L > 0$ such that

$$\|g(x) - g(y)\| \leq L \|x - y\|, \quad \text{for all } x, y \in \mathcal{N}. \quad (11)$$

Sometimes, the boundedness assumption for $\mathcal{L}$ in item (i) is unnecessary and we only require that f is bounded below in $\mathcal{L}$. However, we will just use Assumption 1 for the convergence results in this survey. Under Assumption 1 on f , we state a very useful result, which was obtained by Zoutendijk [19] and Wolfe [20,21]. The relation (12) is usually called as the *Zoutendijk condition*.

Lemma 1. *Suppose that Assumption 1 holds. Consider any iterative method of the form (2), where d_k satisfies $g_k^T d_k < 0$ and α_k is obtained by the standard Wolfe line search. Then we have that*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < +\infty. \quad (12)$$

To simplify the statements of the following results, we assume that $g_k \neq 0$, for all k , for otherwise a stationary point has been found. Assume also that $\beta_k \neq 0$, for all k . This is because if $\beta_k = 0$, the direction in Equation (3) reduces to the negative gradient direction. Thus, either the method converges globally if $\beta_k = 0$ for infinite number of k , or one can take some x_k as the new initial point. In addition, we say that a method is *globally convergent* if

$$\liminf_{k \rightarrow \infty} \|g_k\| = 0, \quad (13)$$

and is *strongly convergent* if

$$\lim_{k \rightarrow \infty} \|g_k\| = 0. \quad (14)$$

If the iterations $\{x_k\}$ stay in a bounded region, Equation (13) means that there exists at least one cluster point, which is a stationary point of f , while Equation (14) indicates that every cluster point of $\{x_k\}$ will be a stationary point of f .

To analyze the methods of the form (2) and (3), besides Equation (9), we derive another basic relation. By Equation (3), we have $d_k + g_k = \beta_k d_{k-1}$ for all $k \geq 2$. Squaring both sides of this relation yields

$$\|d_k\|^2 = -2g_k^T d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \quad (15)$$

The following theorem gives a general convergence result for any descent methods of the form (2) and (3) under the strong Wolfe line search. It indicates that, if $\|d_k\|^2$ is at most linearly increasing, namely, $\|d_k\|^2 \leq c_1 k + c_2$, for all k , a descent conjugate gradient method with strong Wolfe line search is globally convergent.

Theorem 1 [22]. *Suppose that Assumption 1 holds. Consider some methods of the form (2) and (3) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line search (5) and (6). Then the method is globally convergent if*

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} = +\infty. \quad (16)$$

Proof. It follows from Equations (9) and (6) that $|g_k^T d_k| + \sigma |\beta_k| |g_{k-1}^T d_{k-1}| \geq \|g_k\|^2$, which with the Cauchy–Schwarz inequality gives

$$(g_k^T d_k)^2 + \beta_k^2 (g_{k-1}^T d_{k-1})^2 \geq c_1 \|g_k\|^4, \quad (17)$$

where $c_1 = (1 + \sigma^2)^{-1}$ is constant. By Equation (15), $g_k^T d_k < 0$ and Equation (17), we have

$$\begin{aligned} & \frac{(g_k^T d_k)^2}{\|d_k\|^2} + \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \\ &= \frac{1}{\|d_k\|^2} \left[(g_k^T d_k)^2 + \frac{\|d_k\|^2}{\|d_{k-1}\|^2} (g_{k-1}^T d_{k-1})^2 \right] \\ &\geq \frac{1}{\|d_k\|^2} \left[(g_k^T d_k)^2 + \beta_k^2 (g_{k-1}^T d_{k-1})^2 \right. \\ &\quad \left. - \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \|g_k\|^2 \right] \\ &\geq \frac{1}{\|d_k\|^2} \left[c_1 \|g_k\|^4 - \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \|g_k\|^2 \right]. \end{aligned} \quad (18)$$

Assume that Equation (13) is false and there exists some constant $\gamma > 0$ such that

$$\|g_k\| \geq \gamma, \text{ for all } k \geq 1. \quad (19)$$

Notice that the Zoutendijk condition (12) implies that $g_k^T d_k / \|d_k\|$ tends to zero. By this,

Equations (18) and (19), we have for sufficiently large k ,

$$\frac{(g_k^T d_k)^2}{\|d_k\|^2} + \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \geq \frac{c_1}{2} \frac{\|g_k\|^2}{\|d_k\|^2}. \quad (20)$$

Thus, by the Zoutendijk condition and Equation (19), we must have that

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} \leq \frac{1}{\gamma^2} \sum_{k \geq 1} \frac{\|g_k\|^2}{\|d_k\|^2} < +\infty, \quad (21)$$

which is a contradiction to the assumption (16). Therefore, we must have that the convergence relation (13) holds.

We are now going to provide another general global convergence theorem for any descent methods (2) and (3) with the standard Wolfe line search. To this aim, we define

$$t_k = \frac{\|d_k\|^2}{\phi_k^2}, \quad \phi_k = \begin{cases} \|g_1\|^2, & \text{for } k=1; \\ \prod_{j=2}^k \beta_j^2, & \text{for } k \geq 2. \end{cases} \quad (22)$$

By dividing Equation (15) by ϕ_k^2 and noticing that $d_1 = -g_1$, we can obtain [23] that for all $k \geq 1$

$$t_k = -2 \sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \quad (23)$$

Theorem 2 [17]. *Suppose that Assumption 1 holds. Consider any methods of the form (2) and (3) with d_k satisfying $g_k^T d_k < 0$ and with the standard Wolfe line search (5) and (7). Then the method is globally convergent if the scalar β_k is such that*

$$\sum_{k \geq 1} \prod_{j=2}^k \beta_j^{-2} = +\infty. \quad (24)$$

Proof. Define ϕ_k as in Equation (22). The condition (24) is equivalent to

$$\sum_{k \geq 1} \frac{1}{\phi_k^2} = +\infty. \quad (25)$$

Noting that $-2g_i^T d_i - \|g_i\|^2 \leq (g_i^T d_i)^2 / \|g_i\|^2$, it follows from Equation (23) that

$$t_k \leq \sum_{i=1}^k \frac{(g_i^T d_i)^2}{\|g_i\|^2 \phi_i^2}. \quad (26)$$

Since $t_k \geq 0$, the relation (23) also gives

$$-2 \sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} \geq \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \quad (27)$$

Noting that $-4g_i^T d_i - \|g_i\|^2 \leq 4(g_i^T d_i)^2 / \|g_i\|^2$, we get by this and Equation (27) that

$$\begin{aligned} 4 \sum_{i=1}^k \frac{(g_i^T d_i)^2}{\|g_i\|^2 \phi_i^2} &\geq -4 \sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2} \\ &\geq \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \end{aligned} \quad (28)$$

Now we proceed by contradiction and assume that Equation (19) holds. Then by Equations (28), (25), and (19), we have that

$$\sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \phi_k^2} \geq \frac{\gamma^2}{4} \sum_{k \geq 1} \frac{1}{\phi_k^2} = +\infty, \quad (29)$$

which means that the sum series in the right-hand side of Equation (26) is divergent. By Lemma 6 in Pu and Yu [24], we then know that

$$\begin{aligned} +\infty &= \sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \phi_k^2} \frac{1}{t_k} = \sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \|d_k\|^2} \\ &\leq \frac{1}{\gamma^2} \sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2}, \end{aligned} \quad (30)$$

which contradicts the Zoutendijk condition (12). The contradiction shows the truth of Equation (13).

Theorem 2 provides a condition on β_k , which is sufficient for the global convergence of a conjugate gradient method using the standard Wolfe line search. Instead of the sufficient descent condition (10), only the descent condition $d_k^T g_k < 0$ is used here. An easy understanding between Theorems 1

and 2 is given in Dai [23] under the strong Wolfe line search, in which situation we have the estimate $d_k \approx \beta_k d_{k-1}$, if there is no convergence. Since different nonlinear conjugate gradient methods only vary with the scalar β_k , we believe that the condition (24) in Theorem 2 is very powerful in the convergence analysis of conjugate gradient methods. See Dai [23] for some further uses of Equation (24).

EARLY CONJUGATE GRADIENT METHODS

The Fletcher–Reeves Method

In 1964, Fletcher and Reeves [3] proposed the first nonlinear conjugate gradient method and used the following conjugate gradient parameter

$$\beta_k^{\text{FR}} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}. \quad (31)$$

The introduction of the FR method is a milestone in the field of large-scale nonlinear optimization.

Early analysis with the FR method is based on the exact line search. Zoutendijk [19] proved that the FR method with line search is globally convergent for nonlinear functions. Al-Baali [25] first analyzed the FR method with strong Wolfe inexact line searches (5) and (6). He showed that if $\sigma < 1/2$, the sufficient condition (10) holds and there is global convergence. Liu *et al.* [26] extended Al-Baali's result to the case that $\sigma = 1/2$. Dai and Yuan [27] presented a simpler proof to this result by showing that the sufficient condition (10) holds for at least one of any two neighboring iterations. Here it is worth noting that after the descent condition (8) has been verified, we can establish the global convergence easily by Theorem 2. More exactly, assuming that there is no convergence and Equation (19) holds, we can see that $\prod_{j=2}^k \beta_j^2$ is at most linearly increasing and hence Equation (24) holds. Consequently, there will be global convergence by Theorem 2, leading to a contradiction.

Further, if $\sigma > 1/2$, Dai and Yuan [27] proved that even for the one-dimensional

quadratic function

$$f(x) = \frac{1}{2}x^2, \quad x \in R,$$

the FR method may fail due to generating an uphill search direction. Interestingly enough, if we continue the FR method by searching its opposite direction once an uphill direction is generated and keeping $x_{k+1} = x_k$ if g_{k+1} is orthogonal to d_{k+1} , it is still possible to establish the global convergence of the method. Dai and Yuan [28] considered this idea and showed the global convergence of the FR method under a generalized Wolfe line search.

Powell [18] analyzed the global efficiency of the FR method with the exact line search. Denote by θ_k , the angle between d_k and $-g_k$. The exact line search implies that g_{k+1} is orthogonal to d_k for all k and hence

$$\|d_{k+1}\| = \sec \theta_k \|g_k\| \quad (32)$$

and

$$\beta_{k+1} \|d_k\| = \tan \theta_k \|g_{k+1}\|. \quad (33)$$

By using the above two relations and substituting the formula (31), we can obtain

$$\tan \theta_{k+1} = \sec \theta_k \|g_{k+1}\| / \|g_k\| > \tan \theta_k \|g_{k+1}\| / \|g_k\|. \quad (34)$$

Now, if θ_k is close to $\frac{1}{2}\pi$, the iteration may take a very small step, in which case both the step $s_k = x_{k+1} - x_k$ and the change $y_k = g_{k+1} - g_k$ are small. Thus the ratio $\|g_{k+1}\| / \|g_k\|$ is close to one. Consequently, by Equation (34), θ_{k+1} is close to $\frac{1}{2}\pi$, which indicates that slow progress may occur again on the next iteration. The drawback that the FR method may fall into some circle of tiny steps was extended by Gilbert and Nocedal [29] to the strong Wolfe line search, and was observed by many researchers in the community. It explains why the FR method is sometimes very slow in practical computations.

Suppose that after some iterations, the FR method enters a region in the space of the

variables where f is the quadratic function

$$f(x) = \frac{1}{2}x^T x, \quad x \in R^n. \quad (35)$$

In this case, the exact line search along d_k and $g_k = x_k$ implies that

$$\|g_{k+1}\| = \|g_k\| \sin \theta_k. \quad (36)$$

By this and the equality in (34), we obtain $\theta_{k+1} = \theta_k$. Thus the angle between the search direction and the steepest descent direction remains constant for all consecutive iterations, which makes the method very slow if θ_k is close to $\frac{1}{2}\pi$. This example was addressed by Powell [18] for the two-dimensional case and is actually valid for any dimension.

As will be seen in the section titled “The Polak–Ribière–Polyak Method”, unlike the FR method, the PRP method can generate a search direction close to the steepest descent direction once a small step occurs and hence can avoid cycles of tiny steps. On the other hand, the PRP method need not converge even with the exact line search. This motivated Touati-Ahmed and Storey [30] to extend Al-Baali’s convergence result on the FR method to the general methods (2) and (3) with

$$\beta_k \in [0, \beta_k^{\text{FR}}] \quad (37)$$

and suggested the formula

$$\beta_k^{\text{TS}} = \max \left\{ 0, \min \left\{ \beta_k^{\text{PRP}}, \beta_k^{\text{FR}} \right\} \right\}. \quad (38)$$

Gilbert and Nocedal [29] further extended Equation (37) to the interval

$$\beta_k \in [-\beta_k^{\text{FR}}, \beta_k^{\text{FR}}] \quad (39)$$

and proposed the formula

$$\beta_k^{\text{GN}} = \max \left\{ -\beta_k^{\text{FR}}, \min \left\{ \beta_k^{\text{PRP}}, \beta_k^{\text{FR}} \right\} \right\}. \quad (40)$$

However, the numerical results in Gilbert and Nocedal [29] show that the Gilbert and Nocedal (GN) method is not as good as the PRP method, although it indeed performs better than the FR method.

The Polak–Ribière–Polyak Method

In 1969, Polak and Ribière [31] and Polyak [32] proposed another conjugate gradient parameter, independently, that is

$$\beta_k^{\text{PRP}} = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}, \quad (41)$$

where $y_{k-1} = g_k - g_{k-1}$. In practical computations, the PRP method performs much better than the FR method for many optimization problems because it can automatically recover once a small step is generated. For this, we still consider the exact line search. It follows from Equation (41) that

$$|\beta_{k+1}^{\text{PRP}}| \leq \|g_{k+1}\| \|g_{k+1} - g_k\| / \|g_k\|^2. \quad (42)$$

By using the relations (32), (33), and (42), we can obtain

$$\tan \theta_{k+1} \leq \sec \theta_k \|g_{k+1} - g_k\| / \|g_k\|. \quad (43)$$

Assume that the angle θ_k between $-g_k$ and d_k is close to $\frac{1}{2}\pi$ and $\|s_k\| = \|x_{k+1} - x_k\| \approx 0$. Then we have that $\|g_{k+1} - g_k\| \ll \|g_k\|$ and hence

$$\tan \theta_{k+1} \ll \sec \theta_k. \quad (44)$$

Consequently, the next search direction d_{k+1} will tend to $-g_{k+1}$ and avoid the occurrence of continuous tiny steps. The PRP method was believed to be one of the efficient conjugate gradient methods in the last century.

Nevertheless, the global convergence of the PRP method only proves for strictly convex functions [33]; for general functions, Powell [34] showed that the PRP method can cycle infinitely without approaching a solution even if the stepsize α_k is chosen to the least positive minimizer of the line search function. To change this unbalanced state, Gilbert and Nocedal [29] considered Powell [35]’s suggestion of modifying the PRP method by setting

$$\beta_k^{\text{PRP}^+} = \max\{\beta_k^{\text{PRP}}, 0\}, \quad (45)$$

and showed that this modification of the PRP method, called PRP^+ , is globally convergent both for exact and inexact line searches. More

exactly, Gilbert and Nocedal established the following result.

Theorem 3. *Suppose that Assumption 1 holds. Consider the PRP⁺ method, namely, Equations (2) and (3), where β_k is given by Equation (45). If the line search satisfies the standard Wolfe conditions (5) and (7), and the sufficient descent condition (10) for some constant $c > 0$, the method is globally convergent.*

The technique of their proof is quite sophisticated. First, they define the so-called Property (*), which is a mathematical lifting of the property of avoiding cycles of tiny steps.

Property (*). Consider the methods (2) and (3) and assume that $0 < \gamma < \|g_k\| \leq \bar{\gamma}$. Then we say that the method has Property (*), if there exist constants $b > 1$ and $\zeta > 0$ such that for all k ,

$$|\beta_k| \leq b, \quad (46)$$

and

$$\|s_{k-1}\| \leq \zeta \implies |\beta_k| \leq \frac{1}{2b}. \quad (47)$$

It is not difficult to see that both PRP and PRP⁺ possess such a property. Secondly, defining $u_k = d_k / \|d_k\|$, Gilbert and Nocedal observed that if $\beta_k \geq 0$ and if $\|d_k\| \rightarrow \infty$, u_k and u_{k-1} will tend to be the same, namely, $\|u_k - u_{k-1}\| \rightarrow 0$. Their proof then proceeds by contradiction. If there is no convergence, then $\|d_k\|$ must tend to infinity. Consequently, by Property (*), the method has to take big steps for at least half of the iterations, otherwise $\|d_k\|$ becomes finite. However, since d_k tends to be the same direction for sufficiently large k , the iterations will lie out of the bounded level set $\mathcal{L}$ if there are many big steps. A contradiction is then obtained.

The convergence result of Gilbert and Nocedal requires the sufficient descent condition (10). If the strong Wolfe line search is used instead of the standard Wolfe line

search, the sufficient descent condition can be relaxed to the descent condition of the search direction [22]. However, the following one-dimensional quadratic example shows that the PRP method with the strong Wolfe line search may generate an uphill search direction [36]. Consider

$$f(x) = \frac{1}{2} \lambda x^2, \quad x \in \mathbb{R}^1, \quad (48)$$

where $\lambda = \min\{1 + \sigma, 2 - 2\delta\}$ and suppose that the initial point is $x_1 = 1$. Then for any constant δ and σ satisfying $0 < \delta < \sigma < 1/2$, direct calculations show that the unit step-size satisfies the strong Wolfe conditions (5) and (6). Consequently, $x_2 = 1 - \lambda$ and

$$g_2^T d_2 = \lambda^2 (\lambda - 1)^3 > 0, \quad (49)$$

which means that d_2 is uphill. Thus, for any small $\sigma \in (0, 1)$, the strong Wolfe line search cannot guarantee the descent property of the PRP method even for convex quadratic functions.

To ensure the sufficient descent condition, required by Theorem 3 for the PRP⁺ method, in practical computations, Gilbert and Nocedal [29] designed a dynamic inexact line search strategy. As a matter of fact, their strategy applies to any of the methods (2) and (3) with nonnegative β_k 's. Let us look at Equation (9). If $g_k^T d_{k-1} \leq 0$, we already have Equation (10) since $\beta_k \geq 0$. On the other hand, If $g_k^T d_k > 0$, it must be the case that $g_k^T d_{k-1} > 0$, which means that a one-dimensional minimizer has been bracketed. Then $g_k^T d_{k-1}$ can be reduced and Equation (10) holds by applying a line search algorithm, such as that given by Lemaréchal [37], Fletcher [38], or Moré and Thuente [39]. Comparing with the PRP method, however, no significant improvement is reported in Gilbert and Nocedal [29] for the PRP⁺ method.

Is there any clever inexact line search that can guarantee the global convergence of the original PRP method? Grippo and Lucidi [15] answered this question positively by generalizing an Armijo-type line search in De Leone *et al.* [40]. Given constants $\tau > 0$, $\sigma \in (0, 1)$, $\delta > 0$, and $0 < c_1 < 1 < c_2$, their line search

aims to find

$$\alpha_k = \max \left\{ \sigma^j \frac{\tau |g_k^T d_k|}{\|d_k\|^2}; j = 0, 1, \dots, \right\} \quad (50)$$

such that $x_{k+1} = x_k + \alpha_k d_k$ and $d_{k+1} = -g_{k+1} + \beta_{k+1}^{\text{PRP}} d_k$ satisfy

$$f(x_{k+1}) \leq f(x_k) - \delta \alpha_k^2 \|d_k\|^2 \quad (51)$$

and

$$-c_2 \|g_{k+1}\|^2 \leq g_{k+1}^T d_{k+1} \leq -c_1 \|g_{k+1}\|^2, \quad (52)$$

Such a stepsize must exist because of the following observations. If α_k or, equivalently, $\|s_k\| = \|x_{k+1} - x_k\| = \alpha_k \|d_k\|$ is small, β_{k+1}^{PRP} tends to zero and hence d_{k+1} gets close to $-g_{k+1}$. On the other hand, the difference in the objective function, $f(x_{k+1}) - f(x_k)$, is $O(-g_k^T s_k)$ or $O(\|s_k\|)$, whereas the expected decrease is only of the second order $O(\|s_k\|^2)$. Therefore, Equations (51) and (52) must hold provided that α_k is sufficiently small. Furthermore, since the total reduction of the objective function is finite, the line search condition (51) enforces $\lim_{k \rightarrow \infty} \|s_k\| = 0$. By this property, the strong global convergence relation (14) can be achieved for the PRP algorithm of Grippo and Lucidi. From the view point of computations, Grippo and Lucidi [15] refined their line search algorithm so that the first one-dimensional minimizer of the line search function can be accepted. Again, the numerical experience [41] does not suggest a significant improvement of their algorithm over the PRP method.

Along the line of Grippo and Lucidi [15], Dai and Yuan [36] builds the strong convergence of the PRP method with constant stepsizes

$$\alpha_k \equiv \eta, \text{ where } \eta \in (0, \frac{1}{4L}) \text{ is constant,} \quad (53)$$

where L is the Lipschitz constant in Assumption 1. This result was extended in Dai [42] for the case that $\alpha_k \equiv \frac{1}{4L}$. Chen and Sun [43] further studied the PRP method together with

other conjugate gradient methods using fixed stepsizes of the form

$$\alpha_k = \frac{-\delta g_k^T d_k}{d_k^T Q_k d_k}, \quad (54)$$

where $\delta > 0$ is constant and $\{Q_k\}$ is a sequence of positive definite matrices determined in some way.

The Hestenes–Stiefel Method

In this section, we briefly discuss the HS conjugate gradient methods, namely, (2) and (3) where β_k is calculated by

$$\beta_k^{\text{HS}} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}. \quad (55)$$

Such a formula is first used by Hestenes and Stiefel in the proposition of the linear conjugate gradient method in 1952.

A remarkable property of the HS method is that, no matter whether the line search is exact or not, by multiplying Equation (3) with y_{k-1} and using Equation (55), we always have that

$$d_k^T y_{k-1} = 0. \quad (56)$$

In the quadratic case, y_{k-1} is parallel to Ad_{k-1} , where A is the Hessian of the function. Then Equation (56) implies $d_k^T Ad_{k-1} = 0$, that is, d_k is conjugate to d_{k-1} . For this reason, the relation (56) is often called *conjugacy condition*. If the line search is exact, we have by Equation (9) that $g_k^T d_k = -\|g_k\|^2$ since $g_k^T d_{k-1} = 0$. It follows that $d_{k-1}^T y_{k-1} = \|g_{k-1}\|^2$ and $\beta_k^{\text{HS}} = \beta_k^{\text{PRP}}$. Therefore the HS method is identical to the PRP method in the case of exact line searches. As a result, Powell [34]’s counterexample for the PRP method also applies to the HS method, showing the nonconvergence of the HS method with the exact line search. Unlike the PRP method, whose convergence can be guaranteed by the line search of Grippo and Lucidi [15], it is still not known whether there exists a clever line search such that the (unmodified) HS method is well defined at each iteration and converges globally. The answer is perhaps negative. One major observation is that, when $\|s_{k-1}\|$

is small, both the nominator and denominator of β_k^{HS} become small so that β_k^{HS} might be unbounded. Another observation is that, for any one-dimensional function, we always have

$$\begin{aligned} d_2 &= -g_2 + \beta_2^{\text{HS}} d_1 \\ &= -g_2 + \frac{g_2 \cdot y_1}{d_1 \cdot y_1} d_1 \\ &= -g_2 + g_2 = 0 \end{aligned} \quad (57)$$

independent of the line search. Consequently, there is some special difficulty to ensure the descent property of the HS method with inexact line searches.

Similar to the PRP⁺ method, we can consider the HS⁺ method, where

$$\beta_k^{\text{HS}+} = \max \left\{ \beta_k^{\text{HS}}, 0 \right\}. \quad (58)$$

In case of the sufficient descent condition (10), it is easy to verify that both HS and HS⁺ have Property (*). Further, we can similarly modify the standard Wolfe line search to ensure the sufficient descent condition and global convergence for the HS⁺ method. If the sufficient descent condition (10) is relaxed to the descent condition, Qi *et al.* [44] established the global convergence of a modified HS method, where β_k takes the form

$$\beta_k^{\text{QHL}} = \max \left\{ 0, \min \left\{ \beta_k^{\text{HS}}, \frac{1}{\|g_k\|} \right\} \right\}. \quad (59)$$

Early in 1977, Perry [45] observed that the search direction in the HS method can be written as

$$d_k = -P_k g_k, \quad (60)$$

where

$$P_k = I - \frac{d_{k-1} y_{k-1}^T}{d_{k-1}^T y_{k-1}}. \quad (61)$$

Noting that $P_k^T y_{k-1} = 0$, P_k is an affine transformation that transforms R^n into the null space of y_{k-1} . To ensure the descent property of d_k , however, we may wish the matrix P_k is positive definite. It is obvious that there is no positive definite matrix P_k such that $P_k^T y_{k-1} = 0$. Instead, we look for a positive

definite matrix P_k such that the conjugacy condition (56) holds. In case of exact line searches, it is sufficient to require P_k to satisfy

$$P_k^T y_{k-1} = s_{k-1}, \quad (62)$$

which is exactly the quasi-Newton equation (see Yuan [33] or the article on **Quasi-Newton Methods**). Following this line, we can consider to generate P_k by using the BFGS update from $y_{k-1} I$, where y_{k-1} is some scaling factor. This yields

$$\begin{aligned} P_k(y_{k-1}) &= y_{k-1} \left(I - \frac{s_{k-1} y_{k-1}^T + y_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}} \right) \\ &\quad + \left(1 + \frac{y_{k-1} \|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \right) \frac{s_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}}. \end{aligned} \quad (63)$$

Shanno [46] explored this idea with $y_{k-1} = 1$ (namely, no scaling is considered in the BFGS update) and obtained the search direction

$$\begin{aligned} d_k &= -g_k + \left[\frac{g_k^T y_{k-1}}{s_{k-1}^T y_{k-1}} - \left(1 + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \right) \right. \\ &\quad \left. \frac{s_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] s_{k-1} + \frac{s_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}} y_{k-1}. \end{aligned} \quad (64)$$

The methods (2) and (64) are called *memoryless BFGS methods* by Buckley [47]. It is easy to see that the memoryless BFGS method reduces to the HS method if the line search is exact. Without much more calculations and storage at each iteration, the memoryless BFGS method performs much better than the HS method in practical computations.

In case of inexact line searches, Dai and Liao [48] derived the following relation directly from Equations (60) and (62),

$$\begin{aligned} d_k^T y_{k-1} &= -(P_k g_k)^T y_{k-1} = -g_k^T (P_k^T y_{k-1}) \\ &= -g_k^T s_{k-1}. \end{aligned} \quad (65)$$

By introducing a scaling factor t , Dai and Liao considered a generalized conjugacy condition,

$$d_k^T y_{k-1} = -t g_k^T s_{k-1}, \quad (66)$$

and proposed the following choice for β_k ,

$$\beta_k^{\text{DL}}(t) = \frac{g_k^T y_{k-1} - t g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \quad (67)$$

Clearly, if the line search is exact, namely, $g_k^T s_{k-1} = 0$, the DL direction is identical to the HS direction. If $g_k^T s_{k-1} \neq 0$, an analysis for quadratic functions is presented in Dai and Liao [48], showing that for small values of t , the DL direction can bring a bigger descent in the objective function than the HS direction if an exact line search is done at the k th iteration. The numerical experiments in Dai and Liao [48] showed that the DL method with $t = 0.1$ is a significant improvement of the HS method. In addition, similar to PRP⁺ and HS⁺, Dai and Liao [48] established the global convergence of a modified DL method, where

$$\beta_k^{\text{DL}^+}(t) = \max \left\{ \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, 0 \right\} - t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \quad (68)$$

that allows negative values.

Two further developments of the DL method are made by Yabe and Takano [49] and Li *et al.* [50]. Specifically, based on a modified secant condition given by Zhang *et al.* [51,52], Yabe and Takano [49] suggested the variants of Equations (67) and (68) with the vector y_{k-1} replaced with

$$z_{k-1} = y_{k-1} + \left(\frac{\rho \lambda_k}{s_{k-1}^T u_{k-1}} \right) u_{k-1}, \quad (69)$$

where $\lambda_k = 6(f_{k-1} - f_k) + 3(g_{k-1} + g_k)^T s_{k-1}$, $\rho \geq 0$ is a constant, and $u_{k-1} \in R^n$ satisfies $s_{k-1}^T u_{k-1} \neq 0$ (for example, $u_{k-1} = d_{k-1}$). Li *et al.* [50] considered the modified secant condition in Wei *et al.* [53] and suggested the following replacement of y_{k-1} in Equations (67) and (68):

$$y_{k-1}^* = y_{k-1} + \frac{v_{k-1}}{\|s_{k-1}\|^2} s_{k-1}, \quad (70)$$

where $v_{k-1} = 2(f_{k-1} - f_k) + (g_{k-1} + g_k)^T s_{k-1}$. Owing to the use of precise modified secant conditions, certain numerical improvements are expected for these variants over the DL and DL⁺ methods.

DESCENT CONJUGATE GRADIENT METHODS

From the previous section, we can see that none of the FR, PRP, and HS methods can ensure the descent property of the search direction even if the strong Wolfe conditions (5) and (6) with arbitrary $\sigma \in (0, 1)$. For the FR method, the descent condition can be guaranteed by restricting $\sigma \leq 1/2$. However, this is not true any more for $\sigma > 1/2$. For any constant value of $\sigma \in (0, 1)$, there is always some possibility for the PRP and HS methods not to generate a descent search direction.

If a descent search direction is not produced, a practical remedy is to restart the method along $-g_k$. However, this might degrade the efficiency of the method since the second-derivative information achieved along the previous search direction is discarded [18]. From the previous section, we see that many efforts have been made for early conjugate gradient methods to guarantee a descent direction and hence avoid the use of the remedy, including modifying the conjugate gradient parameter β_k or designing some special line search. In this section, we will first address the conjugate descent method and then emphasize the Dai–Yuan method, both of which can ensure the descent condition under strong Wolfe conditions and standard Wolfe conditions, respectively. A hybrid of the two methods is briefly mentioned at the end, which can ensure a descent direction at every iteration without line searches.

The Conjugate Descent Method

In his monograph [38], Fletcher proposed the conjugate descent (CD) methods, namely, (2) and (3) with β_k given by

$$\beta_k^{\text{CD}} = \frac{\|g_k\|^2}{-d_{k-1}^T g_{k-1}}. \quad (71)$$

Other than the FR, PRP, and HS methods, the CD method can ensure the descent property of each search condition, provided that the strong Wolfe conditions (5) and (6) are used. To see this, we first introduce the

following variants of the strong Wolfe conditions, namely, (5) and

$$\sigma_1 g_k^T d_k \leq g(x_k + \alpha_k d_k)^T d_k \leq -\sigma_2 g_k^T d_k, \quad (72)$$

where $0 < \delta < \sigma_1 < 1$ and $0 \leq \sigma_2 < 1$. If $\sigma_1 = \sigma_2 = \sigma$, the above conditions reduce to the strong Wolfe conditions (5) and (6). Now, by Equations (9) and (71), we have

$$-g_k^T d_k = \|g_k\|^2 \left[1 + g_k^T d_{k-1} / g_{k-1}^T d_{k-1} \right]. \quad (73)$$

The above relation and Equation (72) indicate that

$$1 - \sigma_2 \leq -g_k^T d_k / \|g_k\|^2 \leq 1 + \sigma_1. \quad (74)$$

Since $\sigma_2 < 1$, the left inequality in Equation (74) means that Equation (10) holds with $c = 1 - \sigma_2$ and hence the descent condition holds.

Global convergence analysis of the CD method is made in Dai and Yuan [54] using the generalized strong Wolfe conditions (5) and (72). Specifically, if $\sigma_1 < 1$ and $\sigma_2 = 0$, it follows from Equations (71), (74) and (31) that

$$0 \leq \beta_k^{\text{CD}} \leq \beta_k^{\text{FR}}. \quad (75)$$

Therefore, by the result of Touati-Ahmed and Storey [30] related to the relation (37), there is global convergence of the CD method. However, for any $\sigma_2 > 0$, it is possible that the square norm $\|d_k\|^2$ in the method increases to infinity at an exponential rate. Specifically, Dai and Yuan [54] considered the following two-dimensional function

$$f(x, y) = \xi x^2 - y, \quad \text{where } \xi \in (1, 9/8), \quad (76)$$

and showed that the CD method with the generalized strong Wolfe line search may not solve Equation (76). In real computations, the CD method is even inferior to the FR method.

The Dai–Yuan Method

To enforce a descent direction in case of the standard Wolfe line search, Dai and

Yuan [12] proposed a new conjugate gradient method, where

$$\beta_k^{\text{DY}} = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}}. \quad (77)$$

For the DY method, it follows by Equations (3), (77) and direct calculations that

$$g_k^T d_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} g_{k-1}^T d_{k-1}. \quad (78)$$

The fraction in Equation (78) is exactly the DY formula (77). With this observation, we can get an equivalent expression of β_k^{DY} from Equation (78),

$$\beta_k^{\text{DY}} = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \quad (79)$$

The following theorem establishes the descent property and global convergence of the DY method with the standard Wolfe line search.

Theorem 4. *Suppose that Assumption 1 holds. Consider the DY methods, namely, (2) and (3) where β_k is given by Equation (77). If the line search satisfies the standard Wolfe conditions (5) and (7), we have that $g_k^T d_k < 0$ for all $k \geq 1$. Further, the method converges in the sense that $\liminf_{k \rightarrow \infty} \|g_k\| = 0$.*

Proof. It is obvious that $d_1^T g_1 < 0$ since $d_1 = -g_1$. Assume that $g_{k-1}^T d_{k-1} < 0$. It follows by this and Equation (7) that $d_{k-1}^T y_{k-1} > 0$. Thus by Equation (78), we also have that $g_k^T d_k < 0$. Therefore by induction, $g_k^T d_k < 0$ for all $k \geq 1$.

Now, let us denote

$$q_k = \frac{\|d_k\|^2}{(g_k^T d_k)^2}, \quad r_k = -\frac{g_k^T d_k}{\|g_k\|^2}. \quad (80)$$

Dividing Equation (15) by $(g_k^T d_k)^2$ and using Equations (79) and (80), we obtain

$$q_k = q_{k-1} + \frac{1}{\|g_k\|^2} \frac{2}{r_k} - \frac{1}{\|g_k\|^2} \frac{1}{r_k^2}. \quad (81)$$

Noting that $\frac{2}{r_k} - \frac{1}{r_k^2} \leq 1$, an immediate corollary of Equation (81) is

$$q_k \leq q_{k-1} + \|g_k\|^{-2}. \quad (82)$$

Assume that $\liminf_{k \rightarrow \infty} \|g_k\| \neq 0$ and Equation (19) holds. By Equations (19), (82) and $d_1 = -g_1$, we have $q_k \leq k/\gamma^2$ and hence $\sum_{k \geq 1} q_k^{-1} = +\infty$, which contradicts the Zoutendijk condition (12). Therefore the statement is true.

By Equation (81), we can further exploit the self-adjusting property of the DY method [55]. To this aim, we first notice that the r_k defined in Equation (80) is a quantity that reflects the descent degree of the search direction d_k , since the descent condition (8) is equivalent to $r_k > 0$ and the sufficient condition (10) is the same as $r_k \geq c$. Now let us focus on the relation (81). The second term on the right side of Equation (81) increases the value of q_{k-1} , whereas the third term decreases the value of q_{k-1} . Considering the two terms together, we see that q_{k-1} increases if and only if $r_k \geq 1/2$. If r_k is close to zero, then q_{k-1} will be significantly reduced, since the order of $1/r_k$ in the second term is only one, but its order in the third term is two. This and the fact that $q_k \geq 0$ for all k imply that, in the case when q_{k-1} is very small, r_k must be relatively large. Further investigations along the observations can lead to the following result of the DY method independent of the line search.

Theorem 5. *Consider the DY methods (2),(3) and (77), where d_k is a descent direction. Assume that $0 < \gamma \leq \|g_k\| \leq \bar{\gamma}$ holds for all $k \geq 1$. There must exist positive constants δ_1, δ_2 and δ_3 such that the relations*

$$-g_k^T d_k \geq \frac{\delta_1}{\sqrt{k}}, \quad \|d_k\|^2 \geq \frac{\delta_2}{k}, \quad r_k \geq \frac{\delta_3}{\sqrt{k}} \quad (83)$$

hold for all $k \geq 1$. Further, for any $p \in (0, 1)$, there must positive constants $\delta_4, \delta_5, \delta_6$ such that, for any k , the relations

$$-g_i^T d_i \geq \delta_4, \quad \|d_i\|^2 \geq \delta_5, \quad r_i \geq \delta_6 \quad (84)$$

holds for at least $[pk]$ values of $i \in [1, k]$.

The above theorem enables us to establish global convergence for the DY method provided that the line search is such that

$$f_k - f_{k+1} \geq c \min \left\{ -g_k^T d_k, \|d_k\|^2, q_k^{-1} \right\}, \quad (85)$$

for all $k \geq 1$ and some $c > 0$. Consequently, we can analyze the convergence properties of the DY method using the standard Wolfe line search, the Armijo line search [56], and the line search proposed in De Leone *et al.* [40] and Grippo *et al.* [57] for no-derivative methods.

In general, once some optimization method fails to generate a descent direction, a usual remedy is to do a restart along $-g_k$. As shown in Dai [55], the DY direction can act the role of the negative gradient and meanwhile guarantee the global convergence. A numerical experiment with the memoryless BFGS method in Dai [55] demonstrated this finding.

Since the DY method has the same drawback as the FR method, namely, it cannot recover from cycles of tiny steps, it is natural to consider the hybrid of the DY and HS methods like those for the FR and PRP methods in Touati-Ahmed and Storey [30] and Gilbert and Nocedal [29]. Under the standard Wolfe line search, Dai and Yuan [13] extended Theorem 4 to any method (2),(3) with

$$\beta_k \in \left[-\frac{1-\sigma}{1+\sigma} \beta_k^{\text{DY}}, \beta_k^{\text{DY}} \right], \quad (86)$$

where σ is the parameter in the Wolfe condition (7). In spite of a large admissible interval, the numerical results of Dai and Yuan [13] indicated that the following hybrid is preferable in real computations

$$\beta_k^{\text{DYHS}} = \max \left\{ 0, \min \left\{ \beta_k^{\text{HS}}, \beta_k^{\text{DY}} \right\} \right\}. \quad (87)$$

Unlike the Touati-Ahmed and Storey (TS) and GN hybrid methods, the DYHS method using standard Wolfe line searches performs much better than the PRP method using

strong Wolfe line searches [13]. The latter was generally believed to be one of the most efficient conjugate gradient algorithms.

It is well known that some quasi-Newton methods can be expressed in a unified way and their properties can be analyzed uniformly [58,59]. On the contrary, nonlinear conjugate gradient methods were often analyzed individually. To change the situation, Dai and Yuan [60] proposed a family of conjugate gradient methods, in which

$$\begin{aligned} \beta_k(\lambda) &= \frac{\|g_k\|^2}{\lambda\|g_{k-1}\|^2 + (1-\lambda)d_{k-1}^T y_{k-1}}, \quad \lambda \in [0, 1]. \end{aligned} \quad (88)$$

This family can be regarded as some kind of convex combination of the FR and DY methods. Dai and Yuan [61] further extended the family to the case $\lambda \in (-\infty, +\infty)$ and presented some unified convergence results. Independently, Nazareth [8] regarded the FR, PRP, HS, and DY formulas as the four leading contenders for the conjugate gradient parameter and proposed a two-parameter family:

$$\begin{aligned} \beta_k(\lambda_k, \mu_k) &= \frac{\lambda_k\|g_k\|^2 + (1-\lambda_k)g_k^T y_{k-1}}{\mu_k\|g_{k-1}\|^2 + (1-\mu_k)d_{k-1}^T y_{k-1}}, \\ \lambda_k, \mu_k &\in [0, 1]. \end{aligned} \quad (89)$$

The methods that take the convex combination $\lambda_k\beta_k^{\text{HS}} + (1-\lambda_k)\beta_k^{\text{DY}}$, considered in Andrei [62], can be regarded as a subfamily of Equation (89) with $\mu_k = 0$. Several efficient choices for λ_k in this subfamily are also studied in Andrei [62] based on different secant conditions.

Later, based on FR, PRP, HS, DY, CD, and the formula

$$\rho_k^{\text{LS}} = \frac{g_k^T y_{k-1}}{-d_{k-1}^T g_{k-1}} \quad (90)$$

by Liu and Storey [63], Dai and Yuan [64] proposed a three-parameter family

$$\begin{aligned} \beta_k(\lambda_k, \mu_k, \omega_k) &= \frac{\|g_k\|^2 - \lambda_k g_k^T g_{k-1}}{\|g_{k-1}\|^2 + \mu_k g_k^T d_{k-1} - \omega_k \beta_{k-1} g_{k-1}^T d_{k-2}}, \end{aligned} \quad (91)$$

where $\lambda_k \in [0, 1]$, $\mu_k \in [0, 1]$, and $\omega_k \in [0, 1 - \mu_k]$ are parameters. One subfamily of the methods (91) with $\lambda_k = 1$, $\mu_k = 0$, and $\omega_k = u$ is studied in Shi and Guo [65] with an efficient nonmonotone line search. Further, Dai [66] studied a family of hybrid conjugate gradient methods, in which

$$\begin{aligned} \beta_k(\mu_k, \omega_k, \tau_k) &= \frac{\max\{0, \min\{g_k^T y_{k-1}, \tau_k \|g_k\|^2\}\}}{(\tau_k + \omega_k)g_k^T d_{k-1} + \mu_k \|g_{k-1}\|^2 + (1-\mu_k)(-d_{k-1}^T g_{k-1})}, \end{aligned} \quad (92)$$

where $\mu_k \in [0, 1]$, $\omega_k \in [0, 1 - \mu_k]$ and $\tau_k \in [1, +\infty)$ are parameters.

The DYCD Method

Suppose that M is some fixed positive integer, and λ and δ are constants in $(0, 1)$. Given an initial guess $\bar{\alpha}_k$ at the k th iteration, the nonmonotone line search by Grippo *et al.* [67] is to compute the least nonnegative integer m such that the steplength $\alpha_k = \bar{\alpha}_k \lambda^m$ satisfies the following relation:

$$f(x_k + \alpha_k d_k) \leq \max\{f_k, \dots, f_{k-M(k)}\} + \delta \alpha_k g_k^T d_k, \quad (93)$$

where $M(k) = \min(M, k-1)$. To enforce a descent search direction at every iteration in this situation, Dai [14] considered a hybrid of the DY and CD methods, namely, (2) and (3) with

$$\beta_k^{\text{DYCD}} = \frac{\|g_k\|^2}{\max\{d_{k-1}^T y_{k-1}, -d_{k-1}^T g_{k-1}\}}. \quad (94)$$

It is proved in Dai [14] that the DYCD method possesses the descent property without line searches. Further, there is the global convergence if the DYCD method is combined with

the above nonmonotone line search. Surprisingly, a variant of the DYCD method tested in Dai [14] was able to solve all the 18 test problems in Moré *et al* [68].

SUFFICIENT DESCENT CONJUGATE GRADIENT METHODS

In this section, we summarize several nonlinear conjugate gradient methods that can guarantee the sufficient descent condition (10), especially the CG_Descent method by Hager and Zhang [7,69].

Though the sufficient descent condition (10) is not scale invariant, there is, however, some difficulty to differentiate *descent* conjugate gradient methods and *sufficient descent* conjugate gradient methods. More exactly, if d_k satisfies $g_k^T d_k < 0$, we can define another method whose search direction is $\bar{d}_k = (-c \|g_k\|^2 / g_k^T d_k) d_k$, such that $g_k^T \bar{d}_k = -c \|g_k\|^2$.

Let us take the DY method as an illustrative example. A variant of the DY method is given in Dai [55], where d_k takes the form

$$d_k = -\frac{d_{k-1}^T y_{k-1}}{\|g_k\|^2} g_k + d_{k-1}. \quad (95)$$

Since $d_1 = -g_1$, we can get by the induction principle that

$$g_k^T d_k = -\|g_1\|^2, \quad \text{for all } k \geq 1. \quad (96)$$

Further, if a scaling factor $\|g_k\|^2 / \|g_{k-1}\|^2$, that is, the formula β_k^{FR} exactly, is introduced for each search direction d_k (except d_1), we obtain the scheme

$$d_k = -\frac{d_{k-1}^T y_{k-1}}{\|g_{k-1}\|^2} g_k + \beta_k^{\text{FR}} d_{k-1}. \quad (97)$$

In this case, we have that $-g_k^T d_k = \|g_k\|^2$ for all k , which implies that the sufficient descent condition (10) holds with $c = 1$. It is worth mentioning that the above scheme (97) is obtained by Zhang *et al.* [70] (see also the section titled “Several New Methods That Guarantee Sufficient Descent”) with the motivation of modifying the FR method. They

found that, the numerical performance of this scheme is very promising for a large collection of test problems in the CUTER library [71].

The CG_Descent Method

To ensure the sufficient descent condition (10), Hager and Zhang [7] proposed a family of conjugate gradient methods, where

$$\beta_k^{\text{HZ}}(\lambda_k) = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \lambda_k \left(\frac{\|y_{k-1}\|^2 g_k^T d_{k-1}}{(d_{k-1}^T y_{k-1})^2} \right), \quad (98)$$

where $\lambda_k \geq \bar{\lambda} > 1/4$ controls the relative weight placed on the conjugacy degree versus the descent degree of the search direction. This family is clearly related to the DL method (67) with

$$t = \lambda_k \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}}. \quad (99)$$

To verify the sufficient descent condition for the Hager and Zhang (HZ) method, we have by Equations (3) and (98) that

$$\begin{aligned} g_k^T d_k &= -\|g_k\|^2 + \left(\frac{g_k^T y_{k-1} (g_k^T d_{k-1})}{d_{k-1}^T y_{k-1}} \right) \\ &\quad - \lambda_k \left(\frac{\|y_{k-1}\|^2 (g_k^T d_{k-1})^2}{(d_{k-1}^T y_{k-1})^2} \right). \end{aligned} \quad (100)$$

Now, by applying

$$\begin{aligned} u_k &= \frac{1}{\sqrt{2\lambda_k}} (d_{k-1}^T y_{k-1}) g_k, \\ v_k &= \sqrt{2\lambda_k} (g_k^T d_{k-1}) y_{k-1} \end{aligned} \quad (101)$$

into the inequality

$$u_k^T v_k \leq \frac{1}{2} (\|u_k\|^2 + \|v_k\|^2), \quad (102)$$

we can obtain

$$\begin{aligned} \frac{g_k^T y_{k-1} (g_k^T d_{k-1})}{d_{k-1}^T y_{k-1}} &\leq \frac{1}{4\lambda_k} \|g_k\|^2 \\ &\quad + \lambda_k \left(\frac{\|y_{k-1}\|^2 g_k^T d_{k-1}}{(d_{k-1}^T y_{k-1})^2} \right). \end{aligned} \quad (103)$$

Therefore by Equations (100) and (103), we have that

$$g_k^T d_k \leq -\left(1 - \frac{1}{4\lambda_k}\right) \|g_k\|^2, \quad (104)$$

which with the restriction of λ_k means that the sufficient descent condition (10) holds with $c = 1 - (4\bar{\lambda})^{-1}$.

In order to obtain global convergence for general nonlinear functions, Hager and Zhang truncated their conjugate gradient parameter similar to the PRP⁺ method. More exactly, they suggested to choose

$$\begin{aligned} \beta_k^{\text{HZ}+}(\lambda_k) &= \max \left\{ \beta_k^{\text{HZ}}(\lambda_k), \eta_k \right\}, \\ \eta_k &= \frac{-1}{\|d_{k-1}\|^2 \min \{\eta, \|g_{k-1}\|\}}, \end{aligned} \quad (105)$$

where $\eta > 0$ is a constant. With this truncation, they established the global convergence of the modified method (105) with the standard Wolfe line search for general functions.

Hager and Zhang [69,72] tested the value of $\lambda_k = 2$ for the family with a precisely developed efficient line search. For a large collection of large-scale test problems in the CUTer library [71], the new method, called CG_DESCENT, performs better than both PRP⁺ of Gilbert and Nocedal and L-BFGS of Liu and Nocedal.

More efficient choices of λ_k , however, have been found in Kou and Dai [73] by projecting the scaled memoryless BFGS direction defined in Equations (60) and (63) into the one-dimensional manifold $\{-g_k + \beta d_k : \beta \in R\}$. By taking the scaling factors $\gamma_{k-1} = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2}$ and $\gamma_{k-1} = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$, they suggest the uses of $\lambda_k = 2 - \frac{d_{k-1}^T y_{k-1}}{\|d_{k-1}\|^2 \|y_{k-1}\|^2}$ and $\lambda_k = 1$, respectively. A simple and efficient nonmonotone line search criterion is also designed in Kou and Dai [73], that can guarantee the global convergence of the new methods.

Several New Methods That Guarantee Sufficient Descent

The remarkable property of the HZ method (98) that can guarantee the sufficient descent

condition (10) for general functions have attracted several further investigations.

A direct generalization of Equation (98) is given in Yu and Guan [74] (see also Yu *et al.* [75]). They found that, for any β_k of the form

$$\beta_k = \frac{g_k^T v_k}{\Delta_k}, \quad \text{for some } v_k \in R^n \text{ and } \Delta_k \in R, \quad (106)$$

there is a corresponding formula

$$\beta_k^{\text{YG}}(C) = \frac{g_k^T v_k}{\Delta_k} - \frac{C \|v_k\|^2}{\Delta_k^2} g_k^T d_{k-1}, \quad (107)$$

where $C > 1/4$, such that Equation (10) holds with $c = 1 - (4C)^{-1}$. Since almost all of the conjugate gradient parameters can be written into Equation (106), we can obtain various extensions that can guarantee sufficient descent. It is obvious that the HZ formula (98) is corresponding to Equation (107) with the HS formula (55), where $v_k = y_{k-1}$ and $\Delta_k = d_{k-1}^T y_{k-1}$. The extensions of β_k^{FR} , β_k^{PRP} , β_k^{DY} , β_k^{CD} , and β_k^{LS} are also provided in Yu and Guan [74]. A further generalization of this framework on the spectral conjugate gradient method (see Birgin and Martinez [76] or the section titled “Several Topics on Conjugate Gradient Methods”) is given in Yu *et al.* [75].

Another general way of producing sufficient descent conjugate gradient methods is provided in Cheng [77] and Cheng and Liu [78]. Its basic is as follows. For any search direction $-g_k + \beta_k d_{k-1}$, which need not be descent, an orthogonal projection to the null space of g_k leads to the vector

$$d_k^\perp = \left(I - \frac{g_k g_k^T}{\|g_k\|^2} \right) (-g_k + \beta_k d_{k-1}). \quad (108)$$

The search direction defined by

$$\begin{aligned} d_k &= -g_k + d_k^\perp = -\left(1 + \beta_k \frac{g_k^T d_{k-1}}{\|g_k\|^2} \right) \\ &\quad g_k + \beta_k d_{k-1} \end{aligned} \quad (109)$$

then always satisfies $g_k^T d_k = -\|g_k\|^2$. If the line search is exact, the second term in the paranthesis of Equation (109) is missing

since $g_k^T d_{k-1} = 0$. Hence the above scheme reduces to the linear conjugate gradient method in the ideal case. The above procedure with $\beta_k = \beta_k^{\text{PRP}}$ is studied in Cheng [77]. As shown in Cheng and Liu [78], setting $\beta_k = \beta_k^{\text{FR}}$ in Equation (109) leads to the scheme (97). Another variant corresponding to Yabe and Takano [49] (see the end of the section titled “The Hestenes–Stiefel Method”) is also investigated in Cheng and Liu [78].

Observing that the search direction (64) defined by the memoryless BFGS method is formed by the vectors $-g_k$, d_{k-1} , and y_{k-1} , Zhang *et al.* [17] proposed the following modification of the PRP method

$$d_k = -g_k + \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2} d_{k-1} - \frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} y_{k-1}. \quad (110)$$

By multiplying the above d_k with g_k^T , one can see that the corresponding values of the last two terms have opposite signs and hence $g_k^T d_k = -\|g_k\|^2$. The above scheme was implemented in Zhang *et al.* [17] with an Armijo-type line search in relation to De Leone *et al.* [40], yielding comparable numerical results with CG_DESCENT. Although Equation (110) reduces to Equation (3) in case of exact line searches, this scheme is not a standard conjugate gradient method of the form (3) any more.

SEVERAL TOPICS ON CONJUGATE GRADIENT METHODS

As shown in the previous sections, various formulas have been proposed for the nonlinear conjugate gradient parameter β_k , whereas there is not much to do with the choice of this parameter in the linear conjugate gradient method (it is a consensus to use the FR formula (31) there). While some of the existing conjugate gradient algorithms, like the DYHS method (87) and the CG_DESCENT method (98) among others, have proved more efficient than the PRP method, we feel there is still much more room to seek the best nonlinear conjugate gradient algorithms.

As a lot of attention has been paid to the choice of β_k , it is actually also important how to choose the stepsize α_k . Some joint consideration is given by Yuan and Stoer [79], which aims to find the best points of the function over the two-dimensional manifold

$$x_k + \text{Span} \left\{ -\alpha g_k + \beta d_{k-1} : \alpha, \beta \in \mathbb{R}^2 \right\} \quad (111)$$

as the next iterates. Motivated by the success of the Barzilai–Borwein stepsize in the steepest descent method [80,81], Birgin and Martinez [76] proposed the so-called spectral conjugate gradient method that takes the search direction

$$d_k = -\frac{1}{\delta_k} g_k + \frac{g_k^T (y_{k-1} - \delta_k s_{k-1})}{\delta_k d_{k-1}^T y_{k-1}} d_{k-1}. \quad (112)$$

The efficient combination of the Barzilai–Borwein method and the conjugate gradient method, however, is still not known to us. Specifically, the study of Dai and Liao [48] indicates that when the iterate gets close to the solution, the Barzilai–Borwein stepsize can always be accepted by the often-employed nonmonotone line search. We wonder whether there is a similar result for the spectral conjugate gradient method or some of its suitable alternatives.

In addition to the standard conjugate gradient methods of the form (2) and (3), another class of two-term conjugate gradient methods is called *method of shortest residuals* (SR), that was first presented by Hestenes [82] and studied by Pytlak and Tarnawski [83], Dai and Yuan [84], and the references therein. The SR method defines the search direction by

$$d_k = -Nr\{g_k, -\beta_k d_{k-1}\}, \quad (113)$$

where β_k is a scalar and $Nr\{a, b\}$ is defined as the point from a line segment spanned by the vectors a and b which has the smallest

norm, namely,

$$\|Nr\{a, b\}\| = \min \{ \|\lambda a + (1 - \lambda)b\| : 0 \leq \lambda \leq 1 \}. \quad (114)$$

If $\beta_k \equiv 1$, the corresponding variant of the SR method generates the same iterations as the FR method does in case of exact line searches. The formula of β_k corresponding to the PRP method is also given in Pytlak [85] and modified in Dai and Yuan [84]. If, further, the function is quadratic, these variants of the SR method are equivalent and the direction d_k proves to be the SR in the $(k - 1)$ -simplex whose vertices are $-g_1, \dots, -g_k$. For the SR method, the descent property of d_k is naturally implied by its definition (see Pytlak [85] and Dai and Yuan [84] for details). In contrast to the standard conjugate gradient method, where the size of d_k may become very large, the SR method has the trend of pushing $\|d_k\|$ very small. Therefore, we wonder whether there exists some family of methods that includes the standard conjugate gradient method and the SR method as its members. If this is the case, it might be possible to find more efficient methods that monitor the size of $\|d_k\|$ in a more efficient way.

If the storage of more vectors is admissible, one may consider to choose, for example, three-term conjugate gradient methods [17,18,86] and limited-memory quasi-Newton methods [87,88] for solving large-scale optimization problems, other than the two-term conjugate gradient methods. As an alternative, one may think of forming some preconditioner for conjugate gradient methods through the information already achieved in the previous fewer iterations [89–91]. Unlike the linear conjugate gradient method, where a constant preconditioner is usually satisfactory, a robust and efficient conjugate gradient method for highly nonlinear functions requires to be dynamically preconditioned. Therefore, it remains to study how to precondition the nonlinear conjugate gradient method in more effective ways.

Acknowledgments

The author thanks Professor Ya-xiang Yuan and the anonymous editor very much for their useful suggestions and comments. This work was partially supported by the Chinese NSF grants 10831106 and the CAS grant kjcx-yw-s7-03.

REFERENCES

1. Hestenes MR, Stiefel E. Method of conjugate gradient for solving linear system. *J Res Natl Bur Stand* 1952;49:409–436.
2. Shewchuk JR. An introduction to the conjugate gradient method without the agonizing pain. Technical Report CS-94-125. Pittsburgh (PA): Carnegie Mellon University; 1994.
3. Fletcher R, Reeves CM. Function minimization by conjugate gradients. *Comput J* 1964;7:149–154.
4. Crowder HP, Wolfe P. Linear convergence of the conjugate gradient method. *IBM J Res Dev* 1969;16:431–433.
5. Cohen A. Rate of convergence of several conjugate gradient method algorithms. *SIAM J Numer Anal* 1972;9:248–259.
6. McCormick P, Ritter K. Alternative proofs of the convergence properties of the conjugate-gradient method. *J Optim Theory Appl* 1975;13(5):497–518.
7. Hager WW, Zhang H. A survey of nonlinear conjugate gradient methods. *Pac J Optim* 2006;2(1):335–358.
8. Nazareth L. Conjugate-gradient methods. In: Floudas C, Pardalos P, editors. *Encyclopedia of optimization*. Boston (MA), Dordrecht, The Netherlands: Kluwer Academic Publishers; 2001. pp. 319–323.
9. Nazareth JL. Conjugate gradient method. *Comput Stat* 2009;1(3):348–353. *Wiley Interdisciplinary Reviews*.
10. Nocedal J. Theory of algorithms for unconstrained optimization. *Acta Numerica* 1991;1:199–242.
11. Nocedal J. Conjugate gradient methods and nonlinear optimization. In: Adams L, Nazareth JL, editors. *Linear and nonlinear conjugate gradient-related methods*. Philadelphia (PA): SIAM; 1996. pp. 9–23.
12. Dai YH, Yuan Y. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM J Optim* 1999;10(1):177–182.

13. Dai YH, Yuan Y. An efficient hybrid conjugate gradient method for unconstrained optimization. *Ann Oper Res* 2001;103:33–47.
14. Dai YH. A nonmonotone conjugate gradient algorithm for unconstrained optimization. *J Syst Sci Complex* 2002;15(2):139–145.
15. Grippo L, Lucidi S. A globally convergent version of the Polak-Ribière conjugate gradient method. *Math Program* 1997;78:375–391.
16. Wang CY, Zhang YZ. Global convergence properties of s -related conjugate gradient methods. *Chin Sci Bull* 1998;43(23):1959–1965.
17. Zhang L, Zhou W, Li D. A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence. *IMA J Numer Anal* 2006;26(4):629–640.
18. Powell MJD. Restart procedures of the conjugate gradient method. *Math Program* 1977;12:241–254.
19. Zoutendijk G. Nonlinear programming, computational methods. In: Abadie J, editor. *Integer and nonlinear programming*. Amsterdam: North-Holland Publishing Co.; 1970. pp. 37–86.
20. Wolfe P. Convergence conditions for ascent methods. *SIAM Rev* 1969;11:226–235.
21. Wolfe P. Convergence conditions for ascent methods II: some corrections. *SIAM Rev* 1971;13:185–188.
22. Dai YH, Han J, Liu G, et al. Convergence properties of nonlinear conjugate gradient methods. *SIAM J Optim* 1999;10(2):345–358.
23. Dai YH. Convergence analysis of nonlinear conjugate gradient methods. In: Wang Y, Yagola AG, Yang C, editor. *Optimization and regularization for computational inverse problems and applications*. Berlin Heidelberg: Springer; 2010. pp. 157–181.
24. Pu D, Yu W. On the convergence properties of the DFP algorithms. *Ann Oper Res* 1990;24:175–184.
25. Al-Baali M. Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch. *IMA J Numer Anal* 1985;5:121–124.
26. Liu GH, Han JY, Yin HX. Global convergence of the Fletcher-Reeves algorithm with an inexact line search. *Appl Math J Chin Univ Ser B* 1995;10:75–82.
27. Dai YH, Yuan Y. Convergence properties of the Fletcher-Reeves method. *IMA J Numer Anal* 1996;16:155–164.
28. Dai YH, Yuan Y. Convergence of the Fletcher-Reeves method under a generalized Wolfe search. *J Comput Math Chin Univ* 1996;2:142–148.
29. Gilbert JC, Nocedal J. Global convergence properties of conjugate gradient methods for optimization. *SIAM J Optim* 1992;2:21–42.
30. Touati-Ahmed D, Storey C. Global convergent hybrid conjugate gradient methods. *J Optim Theory Appl* 1990;64:379–397.
31. Polak E, Ribière G. Note sur la convergence de méthodes de directions conjuguées. *Rev Fr Inform Rech Oper* 1969;16:35–43.
32. Polyak BT. The conjugate gradient method in extremum problems. *USSR Comput Math Math Phys* 1969;9:94–112.
33. Yuan Y. *Numerical methods for nonlinear programming*. Shanghai: Shanghai Scientific & Technical Publishers; 1993 (in Chinese).
34. Powell MJD. Nonconvex minimization calculations and the conjugate gradient method. In: Griffiths DF, editor. *Numerical analysis, Lecture notes in mathematics 1066*. Berlin: Springer; 1984. pp. 122–141.
35. Powell MJD. Convergence properties of algorithms for nonlinear optimization. *SIAM Rev* 1986;28:487–500.
36. Dai YH, Yuan Y. *Nonlinear conjugate gradient methods*. Shanghai: Shanghai Scientific & Technical Publishers; 2000 (in Chinese).
37. Lemaréchal C. A view of line searches. In: Auslander A, Oetti W, Stoer J, editors. *Optimization and optimal control, Lecture notes in control and information 30*. Berlin: Springer; 1981. pp. 59–78.
38. Fletcher R. *Practical methods of optimization. Volume 1, Unconstrained Optimization*. New York: John Wiley & Sons, Inc.; 1987.
39. Moré J, Thuente DJ. On line search algorithms with guaranteed sufficient decrease. *ACM Trans Math Softw* 1994;20:286–307.
40. De Leone R, Gaudioso M, Grippo L. Stopping criteria for linesearch methods without derivatives. *Math Program* 1984;30:285–300.
41. Nocedal J. Large scale unconstrained optimization. In: Watson A, Duff I, editors. *The state of the art in numerical analysis*. Oxford: Oxford University Press; 1997. pp. 311–338.
42. Dai YH. Convergence of conjugate gradient methods with constant stepsizes. *Optim Meth Software* 2010; DOI: 10.1080/10556781003721042.

43. Chen XD, Sun J. Global convergence of two-parameter family of conjugate gradient methods without line search. *J Comput Appl Math* 2002;146:37–45.
44. Qi HD, Han JY, Liu GH. A modified Hestenes-Stiefel conjugate gradient algorithm. *Chin Ann Math Ser A* 1996;17(3):177–184.
45. Perry JM. A class of conjugate gradient algorithms with a two-step variable-metric memory. Discussion Paper 269. Evanston (IL): Center for Mathematical Studies in Economics and Management Sciences, Northwestern University; 1977.
46. Shanno DF. Conjugate gradient methods with inexact searches. *Math Oper Res* 1978;3:244–256.
47. Buckley A. Conjugate gradient methods. In: Powell MJD, editor. *Nonlinear optimization* 1981. London: Academic Press; 1982. pp. 17–22.
48. Dai YH, Liao LZ. New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl Math Optim* 2001;43(1):87–101.
49. Yabe H, Takano M. Global convergence properties of nonlinear conjugate gradient methods with modified secant condition. *Comput Optim Appl* 2004;28:203–225.
50. Li GY, Tang CM, Wei ZX. New conjugacy condition and related new conjugate gradient methods for unconstrained optimization. *J Comput Appl Math* 2007;202:523–539.
51. Zhang JZ, Deng NY, Chen LH. New quasi-Newton equation and related methods for unconstrained optimization. *J Optim Theory Appl* 1999;102:147–167.
52. Zhang JZ, Xu CX. Properties and numerical performance of quasi-Newton methods with modified quasi-Newton equations. *J Comput Appl Math* 2001;137:269–278.
53. Wei Z, Yu G, Yuan G, et al. The super-linear convergence of a modified BFGS-type method for unconstrained optimization. *Comput Optim Appl* 2004;29(3):315–332.
54. Dai YH, Yuan Y. Convergence properties of the conjugate descent method. *Adv Math* 1996;26(6):552–562.
55. Dai YH. New properties of a nonlinear conjugate gradient method. *Numer Math* 2001;89(1):83–98.
56. Armijo L. Minimization of functions having Lipschitz continuous partial derivatives. *Pac J Math* 1966;16:1–3.
57. Grippo L, Lampariello F, Lucidi S. Global convergence and stabilization of unconstrained minimization methods without derivatives. *J Optim Theory Appl* 1988;56:385–406.
58. Broyden CG. The convergence of a class of double-rank minimization algorithms 1. General considerations. *J Inst Math Appl* 1970;6:76–90.
59. Byrd R, Nocedal J, Yuan Y. Global convergence of a class of variable metric algorithms. *SIAM J Numer Anal* 1987;4:1171–1190.
60. Dai YH, Yuan Y. A class of globally convergent conjugate gradient methods. Research report ICM-98-030. Beijing: Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences; 1998.
61. Dai YH, Yuan Y. An extension class of nonlinear conjugate gradient methods. In: Li D, editors. *Proceedings of the 5th International Conference on Optimization: Techniques and Applications*. Hongkong: Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences; 2001. pp. 778–785.
62. Andrei N. Accelerated hybrid conjugate gradient algorithm with modified secant condition for unconstrained optimization. *Numer Algorithms* 2010;54(1):1017–1398.
63. Liu Y, Storey C. Efficient generalized conjugate gradient algorithms, part 1: theory. *J Optim Theory Appl* 1991;69:129–137.
64. Dai YH, Yuan Y. A three-parameter family of conjugate gradient methods. *Math Comput* 2001;70:1155–1167.
65. Shi ZJ, Guo J. A new family of conjugate gradient methods. *J Comput Appl Math* 2009;224:444–457.
66. Dai YH. A family of hybrid conjugate gradient methods for unconstrained optimization. *Math Comput* 2003;72:1317–1328.
67. Grippo L, Lamparillo F, Lucidi S. A nonmonotone line search technique for Newton's method. *SIAM J Numer Anal* 1986;23:707–716.
68. Moré JJ, Garbow BS, Hillstom KE. Testing unconstrained optimization software. *ACM Trans Math Softw* 1981;7:17–41.
69. Hager WW, Zhang H. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J Optim* 2005;16(1):170–192.
70. Zhang L, Zhou W, Li D. Global convergence of a modified Fletcher-Reeves conjugate gradient method with Armijo-type line search. *Numer Math* 2006;104(4):561–572.

71. Bongartz I, Conn AR, Gould NIM, et al. CUTE: constrained and unconstrained testing environments. *ACM Trans Math Softw* 1995;21:123–160.
72. Hager WW, Zhang H. Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent. *ACM Trans Math Softw* 2006;32(1):113–137.
73. Kou CX, Dai YH. A new conjugate gradient algorithm with nonmonotone line search, Research report. Beijing: LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences; 2010.
74. Yu GH, Guan LT. New descent nonlinear conjugate gradient methods for large-scale optimization. Technical Report. Department of Scientific Computation and Computer Applications, Sun Yat-Sen University, Guangzhou, P. R. China; 2005.
75. Yu GH, Guan LT, Chen WF. Spectral conjugate gradient methods with sufficient descent property for large-scale unconstrained optimization. *Optim Methods Softw* 2008;23(2):275–293.
76. Birgin EG, Martinez JM. A spectral conjugate gradient method for unconstrained optimization. *Appl Math Optim* 2001;43:117–128.
77. Cheng WY. A two-term PRP-based descent method. *Numer Funct Anal Optim* 2007;28:1217–1230.
78. Cheng WY, Liu QF. Sufficient descent nonlinear conjugate gradient methods with conjugacy conditions. *Numer Algorithms* 2010;53:113–131.
79. Yuan Y, Stoer J. A subspace study on conjugate gradient algorithms. *Z Angew Math Mech* 1995;75(1):69–77.
80. Barzilai J, Borwein JM. Two-point step size gradient methods. *IMA J Numer Anal* 1988;8:141–148.
81. Raydan M. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J Optim* 1997;7(1):26–33.
82. Hestenes MR. Conjugate direction methods in optimization. New York, Heidelberg, Berlin: Springer; 1980.
83. Pytlak R, Tarnawski T. On the method of shortest residuals for unconstrained optimization. *J Optim Theory Appl* 2007;133(1):99–110.
84. Dai YH, Yuan Y. Global convergence of the method of shortest residuals. *Numer Math* 1999;83:581–598.
85. Pytlak R. On the convergence of conjugate gradient algorithm. *IMA J Numer Anal* 1989;14:443–460.
86. Nazareth JL. A conjugate direction algorithm without line searches. *J Optim Theory Appl* 1977;23(3):373–387.
87. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program* 1989;45:503–528.
88. Sun LP. Updating the self-scaling symmetric rank one algorithm with limited memory for large-scale unconstrained optimization. *Comput Optim Appl* 2004;27:23–29.
89. Buckley A. A combined conjugate gradient quasi-Newton minimization algorithms. *Math Prog* 1978;15:200–210.
90. Morales JL, Nocedal J. Automatic preconditioning by limited memory quasi-newton updating. *SIAM J Optim* 2000;10(4):1079–1096.
91. Andrei N. Accelerated scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. *Eur J Oper Res* 2010;204:410–420.

NONLINEAR MULTIOBJECTIVE PROGRAMMING

ALEXANDER ENGAU

Department of Mathematical and
Statistical Sciences,
University of Colorado Denver,
Denver, Colorado

Multiojective programming is the field within mathematical programming that deals with the simultaneous optimization (maximization or minimization) of a finite number of two or more real-valued objective functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ($i = 1, 2, \dots, p$) subject to a finite number of real-valued equality or inequality constraints $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$ ($j = 1, 2, \dots, m$). The extension to more general vector spaces or an infinite number of objectives or constraints gives a more general vector optimization problem (VOP), but in this treatment we explicitly consider only the finite-dimensional Euclidean case. Hence, without loss of generality, we now assume that all objectives are to be minimized (otherwise replace $\max f_i(x)$ by $-\min -f_i(x)$) and that all constraints have the form of less-than-or-equal-to inequalities (otherwise replace $g_j(x) = 0$ by $g_j(x) \leq 0$ and $g_j(x) \geq 0$, and $g_j(x) \geq 0$ by $-g_j(x) \leq 0$). With these assumptions, we define the multiojective program (MOP)

MOP: Minimize

$$f(x) := (f_1(x), f_2(x), \dots, f_p(x))^T$$

subject to

$$g(x) := (g_1(x), g_2(x), \dots, g_m(x))^T \leq 0,$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are the vector functions comprised of objectives and constraints, respectively, and $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ is the vector of optimization or decision variables. Given MOP, the sets $S := \{x \in \mathbb{R}^n : g(x) \leq 0\}$ and $F := f[S] := \{y \in \mathbb{R}^p : y = f(x) \text{ for some } x \in S\}$ are called the feasible region or set of feasible points, and the set of attainable outcomes, respectively. As in the excluded special case $p = 1$, in

which the above formulation reduces to a single-objective program (SOP), MOP is said to be linear if all functions f_i and g_j are affine (linear plus a constant); combinatorial or (mixed-)integer if all (or some) components of the decision vector x have to be integers; and, respectively, nonlinear and continuous otherwise. Other classifications also distinguish deterministic MOPs from problems under uncertainty in which f and g involve random, stochastic, or fuzzy parameters, and single-level from multilevel problems in which several optimization problems are nested in some hierarchical fashion. Because the resulting additional challenges are often similar to those for SOPs, which are addressed in more detail elsewhere in this encyclopedia, this discussion focuses primarily on deterministic, continuous, nonlinear multiojective programming with the goal of highlighting some of the similarities and differences in the theory and methodology of solving MOPs in contrast to problems with only a single objective. Readers interested in a more comprehensive presentation are directed to the excellent monographs by Ehrgott [1] who offers a well-written introduction to nonlinear, linear, and combinatorial multicriteria optimization both with motivational examples and illustrations as well as numerous exercises; Jahn [2] who gives a rigorous exposition of the theory of more general vector optimization in the setting of partially ordered linear spaces together with many interesting applications in the mathematical and engineering sciences; and Miettinen [3] who provides a balanced treatment of the theory and numerous methods of nonlinear multiojective optimization with a particular focus on interactive procedures and decision making. In addition, in the last few years several bibliographic state-of-the-art surveys on multiojective programming have been written by Ehrgott and Wiecek [4] for an extensive volume on multiple criteria decision analysis [5], and more specifically on nonlinear MOPs [6], combinatorial MOPs [7],

fuzzy and multilevel MOPs [8–10], vector optimization [11], interactive procedures [12,13], and evolutionary algorithms [14–17] for an edited volume on multiple criteria optimization [18]. Finally, some case studies and practical applications in engineering and the sciences are collected in the books by Collette and Siarry [19], Stadler [20], and Wierzbicki *et al.* [21], and a lot of interesting and still relevant material can also be found in the classical but older texts by Chankong and Haimes [22], Cohon [23], Göpfert and Nehse [24], Hwang and Masud [25], Jahn [26], Luc [27], Sawaragi *et al.* [28], Steuer [29], White [30], Yu [31], Zeleny [32], and others. Much narrower in scope than many of these texts that treat the field of multiobjective programming together with the more general discipline of multicriteria decision making (MCDM), however, here we adopt the conceptual distinction between multiobjective programming as mathematical solution of MOPs, which forms the subject of this article, and MCDM as the preferential solution of multicriteria decision problems by one or multiple decision makers, which is discussed by numerous other authors in this encyclopedia. Consequently, this article restricts its focus to several solution concepts for MOPs in the first section, their characterization using optimality conditions and duality in the second section, and some representative generating methods in the third section, all of which are discussed independent of preferential input by a decision maker. Some concluding remarks are offered in the fourth section that is followed by a long list of supplemental references.

SOLUTION CONCEPTS

The fundamental concepts of Pareto optimality, cone efficiency, and nondominance that are reviewed in the section titled “Pareto Optimality, Efficiency, and Nondominance” can be found in essentially any of the above references on multiobjective programming or multicriteria optimization. The subsequent discussion in the section titled “Approximate, Epsilon and Proper Efficiency” describes,

first, approximate (epsilon-)efficiency as a further generalization of these classical concepts and, second, two concepts of proper efficiency with a particular focus on their relationships to trade-offs and trade-off rates.

Pareto Optimality, Efficiency, and Nondominance

Let us first consider the set $F = \{f(x) \in \mathbb{R}^p : g(x) \leq 0, x \in \mathbb{R}^n\}$, where f and g are the objectives and constraints of a deterministic, continuous, nonlinear or linear MOP. Different than for SOPs for which the corresponding sets are sets of real numbers with a well-defined minimum or infimum, F is a set of real vectors with no canonical order to determine an obvious meaning of minimum or minimization. In consequence, a variety of different (partial) orders can be used to compare two vectors or outcomes $f(x)$ and $f(x')$, among which the following three are of relevance for our general discussion in this section or some specific scalarization methods discussed later in the sections titled “Objective Aggregation and Prioritization” and “Reference Points and Norm Minimization”.

- Componentwise (Pareto) orders [33,34]

$$\begin{aligned} f(x) \leq f(x') &: \Leftrightarrow f_i(x) \leq f_i(x') \\ &\text{for all } i = 1, \dots, p, \\ f(x) \preceq f(x') &: \Leftrightarrow f(x) \leq f(x') \\ &\text{and } f(x) \neq f(x'), \\ f(x) < f(x') &: \Leftrightarrow f_i(x) < f_i(x') \\ &\text{for all } i = 1, \dots, p; \end{aligned}$$

- Lexicographic order [35–37]

$$\begin{aligned} f(x) \preceq_{\text{lex}} f(x') &: \Leftrightarrow \\ &f_k(x) < f_k(x') \text{ for some } k = 1, \dots, p, \\ &f_i(x) = f_i(x') \text{ for all } i = 1, \dots, k-1; \end{aligned}$$

- Max orders

$$\begin{aligned} f(x) \leq_{\max} f(x') &: \Leftrightarrow \max_i \{f_i(x)\} \leq \max_i \{f_i(x')\}, \\ f(x) <_{\max} f(x') &: \Leftrightarrow \max_i \{f_i(x)\} < \max_i \{f_i(x')\}. \end{aligned}$$

Pareto Optimality and Efficiency. While the lexicographic order $\leq_{\text{lex}}$ is a total order of $\mathbb{R}^p$ and therefore defines a unique lexicographic minimum in F , if one exists, the component-wise and max orders are only (strict) partial orders and accordingly may only identify a set of incomparable (weak) Pareto minima and min-max solutions, respectively. More precisely, a point $x^* \in S$ and its outcome $y^* = f(x^*) \in F$ are said to be Pareto optimal or weakly Pareto optimal for MOP if there is no other $x \in S \setminus \{x^*\}$ such that $f(x) \leq f(x^*)$ or $f(x) < f(x^*)$, respectively. More generally, if $\preceq$ and $<$ are any partial and associated strict partial orders of $\mathbb{R}^p$, then $x^* \in S$ is said to be efficient or weakly efficient for MOP with respect to $\preceq$ if there is no other $x \in S \setminus \{x^*\}$ such that $f(x) \preceq f(x^*)$ or $f(x) < f(x^*)$, respectively, and its associated outcome $y^* = f(x^*) \in F$ is said to be (weakly) nondominated. Equivalently, based on a definition by Yu [31,38,39], for every $y \in F$ we can define the two domination sets $D(y) := \{d \in \mathbb{R}^p : y \preceq y + d\}$ and $D_w(y) := \{d \in \mathbb{R}^p : y < y + d\}$ and then characterize the sets of efficient and weakly efficient points in S under f with respect to D by

$$E(S, f, D) := \{x^* \in S : \text{there is no}$$

$$y \in F \setminus \{f(x^*)\} \text{ such that } f(x^*) \in y + D(y)\}$$

and $E_w(S, f, D) := E(S, f, D_w)$, respectively, where $y + D := \{y\} + D := \{y + d \in \mathbb{R}^p : d \in D\}$ is defined as the set-algebraic Minkowski sum of the sets $\{y\}$ and D . The corresponding outcome set

$$N(F, D) := \{y^* \in F : \text{there is no } y \in F \setminus \{y^*\} \text{ such that } y^* \in y + D(y)\} = f[E(S, f, D)]$$

and $N_w(F, D) := N(F, D_w)$ of (weakly) nondominated points in F is also called the (weakly) efficient frontier for MOP. In the frequent case that all $D(y)$ are equal, the domination structure D is said to be constant and variable otherwise. In both theory and practice, the most common domination sets are constant pointed convex cones, that is, cones $C \subseteq \mathbb{R}^p$ that satisfy $C \cap (-C \setminus \{0\}) = \emptyset$ and $C + C \subseteq C$. In fact, for every pointed convex cone C with nonempty interior, $\text{int } C \neq \emptyset$, the

associated cone relations $y \preceq_C z : \Leftrightarrow z - y \in C$ and $y <_C z : \Leftrightarrow z - y \in \text{int } C$ determine (strict) partial orders, and conversely, $D(y)$ and $D_w(y)$ are respectively closed and open pointed convex cones if the underlying partial orders are additive and compatible with positive scalar multiplication. The two cones corresponding to the Pareto orders are the nonnegative and positive orthant $\mathbb{R}_{\geq}^p$ and $\mathbb{R}_{>}^p$, which in this context are also called the Pareto cones, and the sets $E(S, f, \mathbb{R}_{\geq}^p)$ and $N(F, \mathbb{R}_{\geq}^p)$ are also known as the Pareto set and alternatively as Pareto frontier, surface, or curve, respectively.

Properties of the Efficient and Nondominated Sets. Primarily focusing on closed convex domination cones, several properties of the efficient and nondominated sets have been investigated including their existence [40–65], structure (boundedness, connectedness, and compactness) [66–99], and stability [100–122,124,125]. An excellent account of these topics and seminal for much of their development with respect to general cones is the monograph by Sawaragi *et al.* [28]. Some specific results can be derived for polyhedral cones, that is, cones $C = \{d \in \mathbb{R}^p : Ad \geq 0\}$ with some matrix $A \in \mathbb{R}^{q \times p}$ that satisfy the well-known relationship $A[N(F, C)] \subseteq N(A[F], \mathbb{R}_{\geq}^q)$ between nondominated points with respect to C in F and Pareto points in $A[F] := \{z \in \mathbb{R}^q : z = Ay \text{ for some } y \in F\}$ with equality if C is pointed, or equivalently, if A has full column rank [28,31,126–130]. Because the underlying efficient sets are then identical, this result provides a useful technique to extend properties or generating methods for the Pareto set to efficient sets with respect to polyhedral cones, or in a generalized form to convex nonpolyhedral cones [131,132]. More general domination sets that are not necessarily convex cones have also been studied by several researchers for both MOP and VOP [31,38,131–144], and variable domination structures have been investigated in the context of vector-variational inequalities in equilibrium problems [145,146] and for modeling variable preferences in economics and decision analysis [147].

Approximate, Epsilon, and Proper Efficiency

Several other notions generalize or further refine the above concepts of Pareto optimality and cone efficiency to various types of approximate or epsilon efficiency [148–150] and proper efficiency.

Approximate and Epsilon Efficiency. Similar to the above definition, now let $D(y) = C$ be a pointed convex cone. For any vector $\varepsilon \in C$, the sets of (weakly) ε -efficient points can be defined by

$$E(S, f, C, \varepsilon) := \{x^* \in S : \text{there is no } y \in F \setminus \{f(x^*)\} \text{ such that } f(x^*) \in y + C + \varepsilon\}$$

and $E_w(S, f, C, \varepsilon) := E(S, f, \text{int } C, \varepsilon)$, respectively. The set $C_\varepsilon := C + \varepsilon$ with $\text{int } C_\varepsilon = \text{int } C + \varepsilon$ is a translated cone that satisfies $E(S, f, C_\varepsilon) = E(S, f, C, \varepsilon)$ and $E_w(S, f, C_\varepsilon) = E_w(S, f, C, \varepsilon)$, and similarly $N(F, C, \varepsilon) = N(F, C_\varepsilon)$ and $N_w(F, C, \varepsilon) = N_w(F, C_\varepsilon)$ for the corresponding (weakly) ε -nondominated sets. In particular, if C is a polyhedral cone with matrix $A \in \mathbb{R}^{q \times p}$, then similar to the result in the section titled “Pareto Optimality, Efficiency, and Nondominance” it follows that $A[N(F, C, \varepsilon)] \subseteq N(A[F], \mathbb{R}_{\geq b}^q)$, where $b = A\varepsilon$ and $\mathbb{R}_{\geq b}^q := \{z \in \mathbb{R}^q : z \geq b\}$ with equality if C is pointed, so that ε -nondominated points in F correspond to ε -Pareto points in the mapped space $A[F]$ [151,152]. Clearly, these results and the above definition reduce to the former if $\varepsilon = 0$, and to the Pareto case if $C = \mathbb{R}_{\geq}^p$ is the Pareto cone. Several other properties of approximate solutions and related concepts have been studied [153–185]. Not only of interest for various theoretical and computational reasons, the independent relevance of approximate solutions has also been highlighted in several practical applications for minimum-cost flows [186], finance [187], transportation [188], location [189], scheduling [190], engineering design [191,192], and decomposition techniques [193,194]. Finally, if $F \subseteq \mathbb{R}_{\geq}^p$, $C = \mathbb{R}_{\geq}^p$, and $0 \leq \varepsilon \leq 1$ is a real number, then an alternative multiplicative definition of approximate efficiency calls a point $x^* \in S$ ε -efficient if there is no other $x \in S \setminus \{x^*\}$ such that $f(x) \leq (1 - \varepsilon)f(x^*)$, which also has been used for the approximation of trade-offs [195].

Trade-Offs and Proper Efficiency. In general, a (point-to-point) trade-off between two feasible points x^1 and x^2 in S , with respect to two objectives f_i and f_j with $f_j(x^1) - f_j(x^2) \neq 0$, is defined by the ratio $(f_i(x^2) - f_i(x^1))/(f_j(x^1) - f_j(x^2))$ [22], and under suitable differentiability properties of f , the trade-off rate at a single point $x \in S$ with respect to f_i and f_j can be defined by $\partial f_i(x)/\partial f_j$. On the basis of these concepts, a Pareto optimal point $x^* \in S$ is said to be properly efficient in the sense of Geoffrion [196] or, if both f and g are continuously differentiable, in the sense of Kuhn and Tucker (KT) [197] if all trade-offs are bounded, that is, if there exists some positive number $M > 0$ such that

$$(f_i(x^*) - f_i(x))/(f_j(x) - f_j(x^*)) \leq M$$

for some j such that $f_j(x^*) < f_j(x)$,

whenever $f_i(x) < f_i(x^*)$ for some i and $x \in S$, or if there is no $h \in \mathbb{R}^n$ such that $\nabla f(x)h < 0$ and $\nabla g_j(x)h \leq 0$ whenever $g_j(x) = 0$, respectively. In particular, if all f_i and g_j are convex, then proper efficiency in the sense of KT implies proper efficiency in the sense of Geoffrion, and the reverse is true under an appropriate Kuhn–Tucker constraint qualification (KTCQ), for example, if for any $d \in \mathbb{R}^n$ such that $\nabla g_j(x)d \leq 0$ and $g_j(x) = 0$, there exists a differentiable function $h(t)$ and real numbers $(\alpha, T) > 0$ such that $h(0) = x$, $\dot{h}(0) = \alpha d$, and $g(h(t)) \leq 0$ for all $t \in [0, T]$ [3]. The above definitions of trade-offs have been extended to trade-offs along feasible directions [198–201] and other, more general trade-off directions based on the concepts of normal, polar, and tangent cones [202–206]. Many other properties as well as further generalizations to superefficiency [207,208] and other notions of proper efficiency with respect to cones have been investigated by Benson [209,210], Borwein [211], Corley [212], Hartley [76], Henig [213,214] and numerous other scholars [204,215–265]. Characterizations and connections between many of the different definitions of proper efficiency are collected by Sawaragi *et al.* [28]. In particular, if we denote the set of properly efficient points by $E_p(S, f, C)$ and choose $\varepsilon \in \text{int } C$ in the definition of epsilon

efficiency, then in each case one of the most general characterizations is the set inclusion

$$\begin{aligned} E_p(S, f, C) &\subseteq E(S, f, C) \subseteq E_w(S, f, C) \\ &\subseteq E(S, f, C, \varepsilon) \subseteq E_w(S, f, C, \varepsilon). \end{aligned}$$

OPTIMALITY CONDITIONS AND DUALITY

Classical nonlinear programs are typically solved using iterative methods applied to an associated system of nonlinear equations based on some set of optimality conditions or, alternatively, a dual or primal–dual penalty or barrier (interior-point) formulation. For MOP, similar optimality conditions and duality results have been derived for both Pareto optimality and cone efficiency, and are briefly reviewed in the sections titled “First- and Second-Order Conditions for Pareto Optimality” and “Duality and Sensitivity Analysis” based on presentations by Miettinen [3] and Tanino and Kuk [6].

First- and Second-Order Conditions for Pareto Optimality

The following conditions for MOP are very similar to those derived for SOP in terms of the Lagrangian function. To highlight this similarity, we restrict our discussion to the Pareto case; the extension to general cones is straightforward and merely requires an appropriate choice of multipliers from the respective dual cones as mentioned in some more detail at the end of the section.

First-Order Conditions. First, let f_i and g_j be continuously differentiable at a (weakly) Pareto optimal point $x^* \in S$. The multiobjective version of the Fritz–John (FJ) necessary conditions [266] state that then there exist two nonnegative vectors $\mu \in \mathbb{R}_{\geq}^p$ and $\lambda \in \mathbb{R}_{\geq}^m$ that satisfy $(\mu, \lambda) \succeq 0$ and

$$\begin{aligned} \sum_{i=1}^p \mu_i \nabla f_i(x^*) + \sum_{j=1}^m \lambda_j \nabla g_j(x^*) &= 0, \\ \lambda_j g_j(x^*) &= 0 \text{ for all } j = 1, \dots, m. \end{aligned}$$

More specifically, the multiobjective version of the KT conditions [197] states that the

same conditions hold with $\mu \succeq 0$ if x^* either satisfies the KT constraint qualification (KTCQ) from the section titled “Approximate, Epsilon and Proper Efficiency” or is regular, that is, the gradient vectors $\nabla g_j(x^*)$ of all active constraints $g_j(x^*) = 0$ are linear independent [267]. Furthermore, if all f_i and g_j are convex, or more generally, if all f_i are pseudoconvex and g_j quasiconvex, then these conditions are also sufficient for a (weakly) Pareto optimal point if μ can be chosen (strictly) positive and nonzero [3, 268–270]. Finally, if x^* is properly efficient (in the sense of KT), then μ can be chosen strictly positive and λ nonnegative, and these conditions are also sufficient if f_i and g_j are convex. While the FJ conditions were originally given for SOPs, the KT conditions were also prepared for multiple objectives in the seminal paper by Kuhn and Tucker [197] and similarly apply for proper efficiency in the sense of Geoffrion if x^* satisfies the KTCQ or one of several alternative regularity or constraint qualifications [271–276]. In particular, if f_i and g_j are locally Lipschitz but not necessarily continuously differentiable, then extensions of the above conditions remain true under appropriate constraint qualifications [3] if the gradients in the first condition are replaced by subdifferentials, so that 0 must only be contained in the corresponding set of subgradient vectors

$$0 \in \sum_{i=1}^p \mu_i \partial f_i(x^*) + \sum_{j=1}^m \lambda_j \partial g_j(x^*).$$

Second-Order Conditions. If f_i and g_j are twice continuously differentiable at a regular (weakly) Pareto optimal point $x^* \in S$, then the above conditions remain satisfied for $\mu \succeq 0$ and $\lambda \geq 0$, and

$$d^T \left(\sum_{i=1}^p \mu_i \nabla^2 f_i(x^*) + \sum_{j=1}^k \lambda_j \nabla^2 g_j(x^*) \right) d \geq 0,$$

for all $d \in \mathbb{R}^n$ such that $\nabla f_i(x^*)d \leq 0$ for all $i = 1, \dots, p$, and $\nabla g_j(x^*)d = 0$ whenever $g_j(x^*) = 0$. These conditions are also sufficient if the above inequality holds strictly for either all d such that $\nabla f_i(x^*)d \leq 0$ for

all $i = 1, \dots, p$ and $\nabla g_i(x^*)d \leq 0$ whenever $g_i(x^*) = 0$, or alternatively, for all d such that $\nabla g_i(x^*)d \leq 0$ whenever $g_i(x^*) = 0$ with equality if $\lambda_i > 0$ [3,277]. A large number of papers have extended the above results or proposed new first- and second-order conditions for Pareto optimality or efficiency with respect to general cones [277–355]. In particular, the results given above remain valid for weakly or properly efficient points with respect to a convex cone $C \subseteq \mathbb{R}^p$, if the choice of constraint multipliers μ is restricted to elements from one of the two respective dual cones $C_{\geq}^+ := \{\mu \in \mathbb{R}^p : \mu^T d \geq 0 \text{ for all } d \in C\}$ and $C_{>}^+ := \{\mu \in \mathbb{R}^p : \mu^T d > 0 \text{ for all } d \in C \setminus \{0\}\}$.

Duality and Sensitivity Analysis

Two other large parts of the literature deal with duality theory and different types of dual problems as well as the related topics of sensitivity analysis and stability already mentioned in the section titled “Pareto Optimality, Efficiency, and Nondominance”. Dependent on more advanced topics in set-valued optimization, therefore, we limit this only very brief introduction to the set-valued Lagrangian dual problem and the set-valued perturbation map.

Lagrangian Duality. Similar to the Lagrangian function in nonlinear programming, for MOP we can introduce a vector-valued Lagrangian function $L(x, \Lambda) := f(x) + \Lambda g(x)$ where $\Lambda \in \mathbb{R}^{p \times m}$ is a matrix that preserves nonnegativity, that is, $\Lambda d \geq 0$ whenever $d \geq 0$. In further extension of the Lagrangian dual problem, we define a set-valued dual map $\Phi(\Lambda) := N(L[S, \Lambda], \mathbb{R}_{\geq}^p)$ where $L[S, \Lambda] := \{L(x, \Lambda) \in \mathbb{R}^p : x \in S\}$ and consider the associated set-valued optimization problem

$$\begin{aligned} &\text{Maximize } \phi(\Lambda) \\ &\text{subject to } \Lambda \in \mathcal{L} := \{\Lambda \in \mathbb{R}^{m \times k} : \Lambda d \geq 0 \\ &\quad \text{for all } d \geq 0\} \end{aligned}$$

as the dual problem for MOP. Although the optimal solution sets for the original MOP and the above dual problem do not coincide, in general, several other properties between these two sets, and these two sets

of problems in general have been derived [6,28]. A more detailed discussion of this and other concepts of duality, including conjugate and axiomatic types, are given in the earlier surveys by Tanino and Kuk [6], Tammer and Göpfert [11], and a large number of other older and newer papers [258,356–496].

Sensitivity Analysis. Like in linear and nonlinear programming, sensitivity in MOP refers to the effects of changes in the problem data onto the set of optimal solutions, here, specifically the set of nondominated points. Hence, if we consider the parametrized MOP with parameter u (MOP(u))

$$\begin{aligned} &\text{MOP}(u): \text{Minimize} \\ &f(x, u) := (f_1(x, u), f_2(x, u), \dots, f_p(x, u))^T \\ &\text{subject to} \\ &g(x, u) := (g_1(x, u), g_2(x, u), \dots, g_m(x, u))^T \leq 0, \end{aligned}$$

then we can also define the associated sets of (weakly, properly) nondominated points with respect to u resulting in the set-valued maps $N(u) := N(F(u), D)$, $N_w(u) := N_w(F(u), D)$, and $N_p(u) := N_p(F(u), D)$ where $F(u) := \{f(x, u) \in \mathbb{R}^p : g(x, u) \leq 0\}$. The investigation of stability and sensitivity of MOP then pertains to the study of the continuity properties of the perturbation maps $N(u)$, $N_w(u)$, and $N_p(u)$ as well as their contingent derivatives, respectively. Precise definitions with more details and further references are provided in Gal and Wolf [110], and again in the annotated survey by Tanino and Kuk [6] as well as several other older and newer papers [123,497–517].

SCALARIZATION AND GENERATING METHODS

A great variety of solution methods exists for generating efficient solutions and nondominated points, which typically use or combine objective aggregations by a weighted sum, objective prioritization by constraining techniques and hierarchical orders, or some form of norm or distance minimization to some reference point [1–4]. The common characteristic of all these methods is that the original MOP is replaced by a parametrized

single-objective program (SOP) whose real-valued objective function $s : \mathbb{R}^p \rightarrow \mathbb{R}^1$ is (strictly) increasing with respect to the underlying order or ordering cone C , that is, $s(f(x)) \leq s(f(x'))$ or $s(f(x)) < s(f(x'))$ whenever $f(x') \in f(x) + C$ or $f(x') \in f(x) + \text{int } C$ ($f(x') \in f(x) + C \setminus \{0\}$), respectively. In this case, if x^* is an optimal solution for the parametrized problem (SOP(π))

$$\begin{aligned} \text{SOP}(\pi) : \text{Minimize } (s \circ f)(x) &:= s(f(x)) \\ \text{subject to } x &\in S, \end{aligned}$$

for some suitable scalarization parameter π , then x^* is a weakly efficient or efficient point for MOP if the scalarization function s is increasing or strictly increasing, respectively. Of particular interest are those functions s that can find any efficient solution (called completeness by Wierzbicki [518]) and produce only efficient solutions (“constructiveness”), or equivalently, that fully separate nondominated points from dominated points with variation of the scalarization parameters. General conditions for the existence and characterizations of such separating functions have been studied by Wierzbicki [518] and Gerth and Weidner [519].

Objective Aggregation and Prioritization

We begin this brief overview of solutions approaches with the two most popular scalarization methods.

Weighted-Sum Method. The arguably most common method to generate efficient solutions in practice in spite of its well-documented limitations [3, 29, 520–522] combines all objective functions in the form of a nonnegative linear combination or weighted sum [196, 523–527]

$$\begin{aligned} \text{WS}(w) : \text{Minimize } \sum_{i=1}^p w_i f_i(x) \\ \text{subject to } x &\in S, \end{aligned}$$

where the nonzero weighting vector w is chosen to be nonnegative or, more generally, from the dual cone $C_{\geq}^+ = \{w \in \mathbb{R}^p : w^T d \geq 0 \text{ for all } d \in C\}$ for the generation of Pareto

optimal points or efficient solutions with respect to C , respectively. In particular, an optimal solution for $\text{WS}(w)$ is weakly or properly efficient if w belongs to C^+ or $\text{int } C^+$, respectively, and any efficient point can be generated if MOP is convex, that is, if all objectives f_i and constraints g_j are convex functions so that both F and S are convex sets, or more generally, if F is C -convex, that is, if $F + C$ is convex. These restrictions (that unfortunately are often ignored in practice) follow from the geometric interpretation of w as the normal vector of a supporting hyperplane of F that exists at any boundary point $z \in F$ or $z \in N(F, C)$ if and only if F is convex or C -convex, respectively. Furthermore, if f is continuously differentiable at an optimal solution x^* of $\text{WS}(w)$ and provided that $w_i \neq 0$, then the weights satisfy

$$\partial f_i(x) / \partial f_j|_{x=x^*} = -w_j / w_i$$

and also give information about the trade-off rates between f_i and f_j at this solution [528, 529]. The weighted-sum method for Pareto points has been extended by additional constraints [530–532]

$$\begin{aligned} \text{WS}(w, b) : \text{Minimize } \sum_{i=1}^p w_i f_i(x) \\ \text{subject to } f(x) &\leq b \text{ and } x \in S, \end{aligned}$$

which can be generalized for arbitrary ordering cones by replacing the componentwise inequality constraint by the cone constraint $b - f(x) \in C$. While the results on generating (weakly) efficient solutions remain unchanged compared to $\text{WS}(w)$, now every efficient point $x^* \in S$ can be generated as the optimal solution for $b = f(x^*)$ without convexity assumption on f, g, S , or F . In particular, for $C = \mathbb{R}_{\geq}^p$ and $w = e := (1, \dots, 1)^T \in \mathbb{R}^p$, this modified method is equivalent to Benson’s method [42] that is often used to either generate solutions or to confirm Pareto optimality of some given point similar to one of several other direction-based approaches for general pointed convex cones [533–535]. Finally, another important special case

follows when $w = e_k$ is the k th-unit vector

$$\begin{aligned} \text{WS}(e_k, b) : \text{Minimize } & f_k(x) \\ \text{subject to } & f(x) \leq b \text{ and } x \in S, \end{aligned}$$

which is equivalent to the frequently used epsilon-constraint method (that we will therefore discuss separately) if we set $b_i = \varepsilon_i$ for all $i \neq k$ and $b_k \geq \sup\{f_k(x) : f_i(x) \leq b_i \text{ (} i \neq k \text{), } x \in S\}$ [22,536].

Epsilon-Constraint Method. Together with $\text{WS}(w)$ one of the two most commonly used approaches for generating Pareto points and essentially a special case of $\text{WS}(w, b)$, this method does not aggregate the objective vector but prioritizes a single objective f_k while reformulating all others as new constraints

$$\begin{aligned} \text{EC}_k(\varepsilon) : \text{Minimize } & f_k(x) \\ \text{subject to } & f_i(x) \leq \varepsilon_i \text{ (} i \neq k \text{) and } x \in S. \end{aligned}$$

In this case, a point $x^* \in S$ is weakly Pareto optimal or Pareto optimal for MOP if it solves $\text{EC}_k(\varepsilon)$ for some or all k , respectively, or uniquely for at least one k [22]. In particular, x^* is efficient if and only if x^* is optimal for $\text{CO}_k(f(x^*))$ for all $k = 1, \dots, p$, and several other characterizations [537,538] and modifications of this method have been proposed to also improve the detection of (properly) Pareto optimal points [539]. Furthermore, under the same differentiability assumption as for $\text{WS}(w)$ and now based on the sensitivity theorem from nonlinear programming [540,541], the associated trade-off rates between f_i and f_j at an optimal solution x^* of $\text{CO}_k(b)$ are given by

$$\begin{aligned} \partial f_k(x) / \partial f_i|_{x=x^*} &= -\lambda_{ki}^* \\ \text{for all } i &= 1, \dots, m, i \neq k, \end{aligned}$$

where λ_{ki}^* is the Lagrangian multiplier associated with the constraint $f_i(x^*) \leq b_i$ in $\text{CO}_k(b)$ [22]. Finally, a variation of this method is the multi-level lexicographic approach that solves the series of problems

$$\begin{aligned} \text{LM}_k : \text{Minimize } & f_k(x) \\ \text{subject to } & f_i(x) \leq y_i^* \text{ for all } i = 1, \\ & \dots, k-1 \text{ and } x \in S, \end{aligned}$$

where y_i^* are the optimal objective values from each previous subproblem $\text{LM}_i, i = 1, \dots, k-1$ [1]. Consequently, the final solution x^* of this hierarchical method will satisfy $y^* = f(x^*) \leq_{\text{lex}} f(x)$ for all $f(x) \in F \setminus \{f(x^*)\}$ and thus produce the (unique) lexicographic minimum of F , if it exists.

Reference Points and Norm Minimization

The following approaches are generally known as reference-point or direction-based methods [542] and typically formulated to find efficient points whose outcomes are in some sense closest to the ideal or an utopian point $u \in \mathbb{R}^p$, for which $u_i + \epsilon \leq \inf\{f_i(x) : x \in S\}$ for all $i = 1, \dots, p$ with (ideally) $\epsilon = 0$ or (utopian) $\epsilon > 0$. For numerical or comparability purposes, together with the nadir point $v \in \mathbb{R}^p$ for which $v_i = \sup\{y_i : y \in N(F, \mathbb{R}_{\leq}^p)\}$, ideal or (close) utopian points are also often used to scale or normalize each single objective f_i by $\tilde{f}_i(x) := (f_i(x) - u_i)/(v_i - u_i)$, provided that $u_i < v_i$.

Chebyshev Norm Method. Originating in the approximation of the ideal point in compromise programming [543] solvable using ℓ_p -programming techniques [544], the (weighted) Chebyshev norm method

$$\begin{aligned} \text{CN}(w, u) : \text{Minimize } & \max_{i=1, \dots, p} \{w_i(f_i(x) - u_i)\} \\ \text{subject to } & x \in S \end{aligned}$$

with nonnegative weights $w \geq 0$ is based on the (computationally easily tractable) ℓ_∞ -norm and has been proposed by several researchers [29,219,545] as a generating method for optimal solutions with respect to a weighted (by the weights w_i) and shifted (by the reference-point components u_i) version of the max-order introduced in the section titled “Pareto Optimality, Efficiency, and Nondominance”. In this case, a solution $x^* \in S$ is Pareto optimal for MOP if and only if there exists $w > 0$ such that x^* is optimal for $\text{CN}(w, u)$, and several variants and extensions to guarantee that every optimal solution for $\text{CN}(w, u)$ is also Pareto optimal for MOP have been proposed in the form of augmented Chebyshev norms [29,546] and other modifications [547]. In further

generalization, weighted- ℓ_p norms have been replaced by oblique norms [548–550], and consequences of choosing different locations for the reference point have been investigated [551]. Moreover, if $w = e$ and $u = 0$, then $\text{CN}(e, 0)$ reduces to the max-norm scalarization or max-ordering approach that has also been studied independently [1,552]. Finally, the associated trade-off rates

$$\partial f_i(x)/\partial f_j|_{x=x^*} = -\lambda_j^* w_j / \lambda_i^* w_i$$

can be obtained similar to the epsilon-constraint method by rewriting $\text{CN}(w, u)$ in the equivalent form

$$\begin{aligned} & \text{Minimize } \alpha \\ & \text{subject to } \alpha \geq w_i(f_i(x) - u_i) \text{ for all} \\ & \quad i = 1, \dots, p \text{ and } x \in S \end{aligned}$$

where λ_i^* is again the Lagrangian multiplier associated with each constraint $\alpha \geq w_i(f_i(x^*) - u_i)$, $i = 1, \dots, p$ [553].

Pascoletti–Serafini Method. Similar to the geometric interpretation of the weight vector w as normal vector of a separating hyperplane in the weighted-sum method, the weights w in the Chebyshev norm method can also be associated with a direction vector v using the more general scalarization method proposed by Pascoletti and Serafini [535]. Specifically, if we replace all weights $w_i > 0$ in $\text{CN}(w, u)$ by $v_i := w_i^{-1}$ (otherwise, if $w_i = 0$, we can simply drop objective f_i from the problem) and the componentwise inequality $\alpha v \geq f(x) - u$ by the corresponding cone constraint $u + \alpha v \in f(x) + C$ for some pointed convex cone C , then we obtain the Pascoletti–Serafini method

$$\begin{aligned} & \text{PS}(v, u) : \text{Minimize } \alpha \\ & \text{subject to } u + \alpha v \in f(x) + C \text{ and} \\ & \quad x \in S, \end{aligned}$$

which also extends several other direction-based approaches [533,534]. In particular, if $v = e_k$, $u = \varepsilon$, and $C = \mathbb{R}_{\geq}^p$ is the Pareto cone, then $\text{PS}(e_k, \varepsilon)$ reduces to the epsilon-constraint method $\text{EC}_k(\varepsilon)$ showing that

$\text{PS}(v, u)$ may serve as a generalization for both the epsilon-constraint and Chebyshev norm method for general ordering cones. Its main result states that if $(\alpha^*, x^*) \in \mathbb{R} \times S$ is optimal for $\text{PS}(v, u)$ with $u + \alpha^* v \in f(x^*) + C$, then x^* is weakly efficient for MOP and there exists an efficient solution $x^{**} \in E(S, f, C)$ such that $u + \alpha^* v \in f(x^{**}) + C$. This unifying method has the intuitive geometric interpretation to project the point u in direction of v onto the closest point of F after translation by some arbitrary element of the ordering cone C , is a direct consequence of more general separation results in nonlinear programming [519], and recently has also been used for a new adaptive scalarization method based on additional sensitivity considerations [554–556].

ADDITIONAL TOPICS AND CONCLUDING REMARK

In addition to the necessarily very limited list of concepts and methods selected for this short overview of nonlinear multiobjective programming, more recent research also focuses on other solution approaches that do not rely on scalarization but act directly on a certain set of optimality conditions in resemblance to classical nonlinear programming. Novel work in this interesting area is still ongoing with a particular focus on Newton or steepest descent methods [557–563], proximal methods [564–566], and barrier methods or other primal/dual strategies [567–569]. In addition, the continued theoretical investigation of duality and optimality conditions under particular assumptions clearly provides another unlimited plethora of creative research opportunities although not necessarily with a strong relevance for computations or practical purposes as the underlying assumptions and conditions are very often very difficult to actually verify. From that standpoint, the majority of current efforts in actually solving multicriteria optimization problems other than by using scalarization and standard single-objective programming seems to focus primarily on combinatorial problems and the development of new heuristics or metaheuristics. To conclude, however,

the author offers his humble apology should he have overlooked any other recent (or previous) contribution as well as any possible miscitation, and for inevitably missing many other related references of relevance. Specific suggestions in that regard would be greatly appreciated for future revisions of this article.

Acknowledgment

The author most gratefully thanks his doctoral advisor Margaret M. Wiecek for introducing him to the field of multiobjective programming, and to the editors and an anonymous referee for their insightful comments and helpful suggestions that have lead to the improved presentation of this contribution.

REFERENCES

1. Ehrgott M. Multicriteria optimization. 2nd ed. Berlin: Springer; 2005.
2. Jahn J. Vector optimization. Berlin: Springer; 2004.
3. Miettinen K. Nonlinear multiobjective optimization, International series in operations research & management science, 12. Boston (MA): Kluwer Academic Publishers; 1999.
4. Ehrgott M, Wiecek MM. Multiobjective programming. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys, International series in operations research & management science, 78. New York: Springer; 2005. pp. 667–722.
5. Figueira JR, Greco S, Ehrgott M, editors. Multiple criteria decision analysis - state of the art surveys, International series in operations research & management science, 78. New York: Springer; 2005.
6. Tanino T, Kuk H. Nonlinear multiobjective programming. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 71–128.
7. Ehrgott M, Gandibleux X. Multiobjective combinatorial optimization—theory, methodology, and applications. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 369–444.
8. Sakawa M. Fuzzy multiobjective and multi-level optimization. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 171–226.
9. Sakawa M. Volume 14, Genetic algorithms and fuzzy multiobjective optimization, Operations Research/Computer Science Interfaces Series. Boston (MA): Kluwer Academic Publishers; 2002.
10. Sakawa M, Nishizaki I. Volume 48, Cooperative and noncooperative multi-level programming, Operations Research/Computer Science Interfaces Series. New York: Springer; 2009.
11. Tammer C, Göpfert A. Theory of vector optimization. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 1–70.
12. Miettinen K. Interactive nonlinear multiobjective procedures. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 227–276.
13. Korhonen P. Interactive methods. Volume 78, Multiple criteria decision analysis - state of the art surveys, International series in operations research & management science. New York: Springer; 2005. pp. 641–665.
14. Coello Coello CA, Romero CEM. Evolutionary algorithms and multiple objective optimization. Volume 52, Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2002. pp. 277–331.
15. Gandibleux X, Sevaux M, Sörensen K, *et al.*, editors. Volume 535, Metaheuristics for multiobjective optimisation, Lecture Notes in Economics and Mathematical Systems. Berlin: Springer; 2004.
16. Knowles J, Corne D, Deb K, editors. Multiobjective problem solving from nature, Natural computing series. Berlin: Springer; 2008.

17. Branke J, Deb K, Miettinen K, *et al.* Multi-objective optimization: interactive and evolutionary approaches. Belin, Heidelberg: Springer; 2008.
18. Ehrgott M, Gandibleux X, editors. Multiple criteria optimization: state of the art annotated bibliographic surveys, International series in operations research & management science, 52. Boston (MA): Kluwer Academic Publishers; 2002.
19. Collette Y, Siarry P. Multiobjective optimization: principles and case studies (Decision Engineering). Berlin: Springer; 2003.
20. Stadler W, editor. Multicriteria optimization in engineering and in the sciences. Volume 37, Mathematical concepts and methods in science and engineering. New York: Plenum Press; 1988.
21. Wierzbicki A, Makowski M, Wessels J. Decision support methodology with environmental applications. Dordrecht: Kluwer; 2000.
22. Chankong V, Haimes YY. Volume 8, Multiobjective decision making, North-Holland series in system science and engineering. New York: North-Holland Publishing Co., 1983.
23. Cohon JL. Multiobjective programming and planning. Volume 140, Mathematics in science and engineering. New York: Academic Press Inc. [Harcourt Brace Jovanovich Publishers]; 1978.
24. Göpfert A, Nehse R. Volume 74, Vektoroptimierung, Mathematisch-Naturwissenschaftliche Bibliothek [Mathematical-Scientific Library]. Leipzig: BSB B. G. Teubner Verlagsgesellschaft; 1990.
25. Hwang CL, Masud ASM. Volume 164, Multiple objective decision making—methods and applications, Lecture notes in economics and mathematical systems. Berlin: Springer; 1979.
26. Jahn J. Volume 31, Mathematical vector optimization in partially ordered linear spaces, Methoden und Verfahren der Mathematischen Physik [Methods and Procedures in Mathematical Physics]. Frankfurt am Main: Verlag Peter D. Lang; 1986.
27. Luc DT. Volume 319, Theory of vector optimization, Lecture notes in economics and mathematical systems. Berlin: Springer; 1989.
28. Sawaragi Y, Nakayama H, Tanino T. Theory of multiobjective optimization. Volume 176, Mathematics in science and engineering. Orlando (FL): Academic Press Inc.; 1985.
29. Steuer RE. Multiple criteria optimization: theory, computation, and application, Wiley series in probability and mathematical statistics: applied probability and statistics. New York: John Wiley & Sons, Inc.; 1986.
30. White DJ. Optimality and efficiency. Chichester: John Wiley & Sons, Ltd.; 1982.
31. Yu PL. Multiple-criteria decision making. Volume 30, Mathematical concepts and methods in science and engineering. New York: Plenum Press; 1985.
32. Zeleny M. Volume 95, Linear multiobjective programming, Lecture notes in economics and mathematical systems. Berlin: Springer; 1974.
33. Edgeworth FY. Mathematical psychics: an essay on the application of mathematics to the moral sciences. London: Kegan Paul; 1881.
34. Pareto V. Cours d'Économie Politique. Lausanne: Rouge; 1896.
35. Debreu G. Representation of a preference ordering by a numerical function. In: Thrall R, Coombs C, Davies R, editors. Decision processes. New York: Wiley; 1954. pp. 159–175.
36. Georgescu-Roegen N. Choice, expectations and measurability. *Q J Econ* 1954;64:503–534.
37. Fishburn PC. Lexicographic orders, utilities and decision rules: a survey. *Manage Sci* 1974;20:1442–1471.
38. Yu P-L. Introduction to domination structures in multicriteria decision problems. In: Cochrane JL, Zeleny M, editors. Multiple criteria decision making. Columbia: University of South Carolina Press; 1973. pp. 249–261.
39. Yu P-L. Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives. *J Optim Theory Appl* 1974;14:319–377.
40. Bao TQ, Mordukhovich BS. Relative Pareto minimizers for multiobjective problems: existence and optimality conditions. *Math Program Ser A* 2010;122(2):301–347.
41. Batista Santos L, Rojas-Medar M, Ruiz-Garzón G. Existence of weakly efficient solutions in vector optimization. *Acta Math Appl Sin (Engl Ser)* 2008;24(4):599–606.
42. Benson HP. Existence of efficient solutions for vector maximization problems. *J Optim Theory Appl* 1978;26:569–580.
43. Chen GY, Craven BD. Existence and continuity of solutions for vector optimization. *J Optim Theory Appl* 1994;81(3):459–468.

44. Deng S. Characterizations of the nonemptiness and compactness of solution sets in convex vector optimization. *J Optim Theory Appl* 1998;96(1):123–131.
45. Deng S. On efficient solutions in vector optimization. *J Optim Theory Appl* 1998;96(1):201–209.
46. Deng S. Characterizations of the nonemptiness and boundedness of weakly efficient solution sets of convex vector optimization problems in real reflexive Banach spaces. *J Optim Theory Appl* 2009;140(1):1–7.
47. Huang XX, Yang XM. Some existence results of efficiency in vector optimization. *Math Methods Oper Res* 2001;53(3):391–401.
48. Isac G. On Pareto efficiency. A general constructive existence principle. Volume 14, Combinatorial and global optimization (Chania, 1998), Series Applied Mathematics. River Edge (NJ): World Scientific Publishing; 2002. pp. 133–144.
49. Jahn J. Existence theorems in vector optimization. *J Optim Theory Appl* 1986;50(3):397–406.
50. Kazmi KR. Existence of solutions for vector optimization. *Appl Math Lett* 1996;9(6):19–22.
51. Kim MH, Lee GM. Some existence results for vector optimization problems. Volume 3, Fixed point theory and applications. Huntington (NY): Nova Science Publishers; 2002. pp. 147–157.
52. Kim MH, Lee GM. Some existence results for vector optimization problem on Banach space. Fixed point theory and applications (Chinju/Masan, 2001). Hauppauge (NY): Nova Science Publishers; 2003. pp. 159–169.
53. Lee GM, Kim DS, Kuk H. Existence of solutions for vector optimization problems. *J Math Anal Appl* 1998;220(1):90–98.
54. Li L, Li J. Equivalence and existence of weak Pareto optima for multiobjective optimization problems with cone constraints. *Appl Math Lett* 2008;21(6):599–606.
55. Liu Z-B, Kim JK, Huang N-J. Existence of solutions for nonconvex and nonsmooth vector optimization problems. *J Inequal Appl* 2008;7:Art. ID 678014.
56. Luc DT. An existence theorem in vector optimization. *Math Oper Res* 1989;14(4):693–699.
57. Mastroeni G, Pappalardo M. On the existence of solutions to vector optimization problems. Volume 58, Equilibrium problems: nonsmooth optimization and variational inequality methods: Nonsmooth Optimization. Methods and Applications. Dordrecht: Kluwer Academic Publishers; 2001. 175–185.
58. Minh NB, Tan NX. The existence of solutions to generalized bilevel vector optimization problems. *Vietnam J Math* 2005;33(3):291–308.
59. Ng KF, Zheng XY. Existence of efficient points in vector optimization and generalized Bishop-Phelps theorem. *J Optim Theory Appl* 2002;115(1):29–47.
60. Oppezzi P, Rossi AM. Existence and convergence of Pareto minima. *J Optim Theory Appl* 2006;128(3):653–664.
61. Postolică V. New existence results of solutions for vector optimization programs with multifunctions. *Optimization* 2001;50(3-4):253–263.
62. Santos LB, Rojas-Medar MA, Ruiz-Garzón G, *et al.* Existence of weakly efficient solutions in nonsmooth vector optimization. *Appl Math Comput* 2008;200(2):547–556.
63. Sergienko IV, Lebedeva TT, Semenova NV. Existence of solutions in vector optimization problems. *Cybern Syst Anal* 2000;36(6):823–828.
64. Sonntag Y, Zalinescu C. Comparison of existence results for efficient points. *J Optim Theory Appl* 2000;105(1):161–188.
65. Zaslavski AJ. Existence of solutions of a vector optimization problem with a generic lower semicontinuous objective function. *J Optim Theory Appl* 2009;141(1):217–230.
66. Benoist J. Connectedness of the efficient set for strictly quasiconcave sets. *J Optim Theory Appl* 1998;96(3):627–654.
67. Bitran GR, Magnanti TL. The structure of admissible points with respect to cone dominance. *J Optim Theory Appl* 1979;29(4):573–614.
68. Cambini R, Marchi A. A note on the connectedness of the efficient frontier. *J Interdiscip Math* 2005;8(1):21–37.
69. Choo EU, Schaible S, Chew KP. Connectedness of the efficient set in three criteria: quasiconcave programming. *Cah Cent Etud Rech Oper* 1985;27(3-4):213–220.
70. Daniilidis A, Hadjisavvas N, Schaible S. Connectedness of the efficient set for three-objective quasiconcave maximization problems. *J Optim Theory Appl* 1997;93(3):517–524.

71. Fu WT, Zhou KP. Connectedness of the efficient solution sets for a strictly path quasiconvex programming problem. *Nonlinear Anal* 1993;21(12):903–910.
72. Gong XH. Connectedness of the efficient solution set of a convex vector optimization in normed spaces. *Nonlinear Anal* 1994;23(9):1105–1114.
73. Gong XH. Connectedness of super efficient solution sets for set-valued maps in Banach spaces. *Math Methods Oper Res* 1996;44(1):135–145.
74. Gong X. On the contractibility and connectedness of an efficient point set. *J Math Anal Appl* 2001;264(2):465–478.
75. Gong XH. Connectedness of the solution sets and scalarization for vector equilibrium problems. *J Optim Theory Appl* 2007;133(2):151–161.
76. Hartley R. On cone-efficiency, cone-convexity and cone-compactness. *SIAM J Appl Math* 1978;34:211–222.
77. Helbig S. On the connectedness of the set of weakly efficient points of a vector optimization problem in locally convex spaces. *J Optim Theory Appl* 1990;65(2):257–270.
78. Hirschberger M. Connectedness of efficient points in convex and convex transformable vector optimization. *Optimization* 2005;54(3):283–304.
79. Hu Y. An equivalent theorem on connectedness of the weakly cone-efficient sets. *J Math Res Exp* 1997;17(3):340–342.
80. Hu YD, Ling C. Connectedness of cone super-efficient point sets in locally convex topological vector spaces. *J Optim Theory Appl* 2000;107(2):433–446.
81. Hu YD, Sun EJ. Connectedness of the efficient set in strictly quasiconcave vector maximization. *J Optim Theory Appl* 1993;78(3):613–622.
82. Lim K-P. The set of all efficient solutions and their connectedness property in multi-objective optimization problems. *Bull Malays Math Soc* 1987;10(2):56–66.
83. Luc DT. Connectedness of the efficient point sets in quasiconcave vector maximization. *J Math Anal Appl* 1987;122(2):346–354.
84. Makarov EK, Rachkovski NN, Song W. Arcwise connectedness of closed efficient point sets. *J Math Anal Appl* 2000;247(2):377–383.
85. Makarov EK, Rachkovski NN. Efficient sets of convex compacta are arcwise connected. *J Optim Theory Appl* 2001;110(1):159–172.
86. Malivert C, Boissard N. Structure of efficient sets for strictly quasi-convex objectives. *J Convex Anal* 1994;1(2):143–150.
87. Malivert C, Popovici N. The structure of efficient sets in bicriteria quasilinear optimization. *J Nonlinear Convex Anal* 2001;2(3):291–304.
88. Molho E, Zaffaroni A. Quasiconcavity of sets and connectedness of the efficient frontier in ordered vector spaces. Volume 27, *Generalized convexity, generalized monotonicity: recent results* (Luminy, 1996), *Nonconvex Optimization Applications*. Dordrecht: Kluwer Academic Publishers; 1998. pp. 407–424.
89. Naccache PH. Connectedness of the set of nondominated outcomes in multicriteria optimization. *J Optim Theory Appl* 1978;25(3):459–467.
90. Nguyen QH. Arcwise connectedness of the solution sets of a semistrictly quasiconcave vector maximization problem. *Acta Math Vietnam* 2002;27(2):165–174.
91. Popovici N. Structure of efficient sets in lexicographic quasiconvex multicriteria optimization. *Oper Res Lett* 2006;34(2):142–148.
92. Song W, Wu B-Y, Zhang J-M. On connectedness of efficient solution sets for a star cone-quasiconvex vector optimization problem. *Acta Math Appl Sin (Engl Ser)* 2007;23(1):91–98.
93. Sun EJ. On the connectedness of the efficient set for strictly quasiconvex vector minimization problems. *J Optim Theory Appl* 1996;89(2):475–481.
94. Warburton AR. Quasiconcave vector maximization: connectedness of the sets of Pareto-optimal and weak Pareto-optimal alternatives. *J Optim Theory Appl* 1983;40(4):537–557.
95. Ward J. Structure of efficient sets for convex objectives. *Math Oper Res* 1989;14(2):249–257.
96. Wernsdorf R. On the connectedness of the set of efficient points in convex optimization problems with multiple or random objectives. *Math Operationsforsch Stat Ser Optim* 1984;15(3):379–387.
97. Zheng XY. Contractibility and connectedness of efficient point sets. *J Optim Theory Appl* 2000;104(3):717–737.
98. Zhong R-Y, Huang N-J, Wong M-M. Connectedness and path-connectedness of solution

- sets to symmetric vector equilibrium problems. *Taiwan J Math* 2009;13(2B):821–836.
99. Zhou X-W, Hu Y-D. Connectedness of cone-efficient solution set for cone-quasiconvex multiobjective programming in locally convex spaces. *Acta Math Appl Sin (Engl Ser)* 2004;20(2):309–316.
 100. Attouch H, Riahi H. Stability results for Ekeland's ϵ -variational principle and cone extremal solutions. *Sem Anal Convexe* 1990;20(5):43.
 101. Badra NM, Maaty MA. Stability of multicriteria quadratic programming. *J Inst Math Comput Sci Comput Sci Ser* 2000;11(2):133–140.
 102. de Melo W. Stability and optimization of several functions. *Topology* 1976;15(1):1–12.
 103. Dolecki S, Malivert C. Stability of efficient sets: continuity of mobile polarities. *Nonlinear Anal* 1988;12(12):1461–1486.
 104. Durea M. On the existence and stability of approximate solutions of perturbed vector equilibrium problems. *J Math Anal Appl* 2007;333(2):1165–1179.
 105. El-Banna A-ZH, Zarea SA. Stability of linear vector optimization problems corresponding to an efficient set. *Math Comput Simul* 2001;57(6):335–345.
 106. Emelichev VA, Gurevsky EE. On stability of some lexicographic multicriteria Boolean problem. *Control Cybern* 2007;36(2):333–346.
 107. Emelichev VA, Gurevsky EE. On stability of a Pareto-optimal solution under perturbations of the parameters for a multicriteria combinatorial partition problem. *Comput Sci J Moldova* 2008;16(2):286–297.
 108. Fiala P. Duality and stability in linear vector optimization. *Found Control Eng* 1984;8(3-4):155–163.
 109. Gal T. On efficient sets in vector maximum problems—a brief survey. *Eur J Oper Res* 1986;24(2):253–264.
 110. Gal T, Wolf K. Stability in vector maximization: a survey. *Eur J Oper Res* 1986;25(2):169–182.
 111. Gu G, Hu Y. Stability of majorly efficient points and solutions in multiobjective programming. *Appl Math J Chin Univ Ser B* 1995;10(3):313–324.
 112. Huang XX. Stability in vector-valued and set-valued optimization. *Math Methods Oper Res* 2000;52(2):185–193.
 113. Jurkiewicz E. Stability of compromise solution in multicriteria decision-making problems. *J Optim Theory Appl* 1983;40(1):77–83.
 114. Klose J. Volume 808, Stabilitäts- und Sensitivitätsanalyse in der Vektoroptimierung, Deutsche Hochschulschriften [German University Publications]. Egelsbach: Hänssel-Hohenhausen; 1993.
 115. Li S. Sensitivity and stability for contingent derivative in multiobjective optimization. *Math Appl (Wuhan)* 1998;11(2):49–53.
 116. Miglierina E. Stability of critical points for vector valued functions and Pareto efficiency. *J Inf Optim Sci* 2003;24(2):413–422.
 117. Miglierina E, Molho E. Scalarization and stability in vector optimization. *J Optim Theory Appl* 2002;114(3):657–670.
 118. Naccache PH. Stability in multicriteria optimization. *J Math Anal Appl* 1979;68(2):441–453.
 119. Papageorgiou NS. Pareto efficiency in locally convex spaces. II. Stability results. *Numer Funct Anal Optim* 1985;8(1-2):117–136.
 120. Tanino T. Stability and sensitivity analysis in multiobjective nonlinear programming. *Ann Oper Res* 1990;27(1-4):97–114.
 121. Tanino T, Sawaragi Y. Stability of the solution set in problems of finding cone extreme points. *Syst Control* 1978;22(11):693–700.
 122. Tanino T, Sawaragi Y. Stability of nondominated solutions in multicriteria decision-making. *J Optim Theory Appl* 1980;30(2):229–253.
 123. Thuan LV, Luc DT. On sensitivity in linear multiobjective programming. *J Optim Theory Appl* 2000;107(3):615–626.
 124. Vogel S. On stability in multiobjective programming—a stochastic approach. *Math Program Ser A* 1992;56(1):91–119.
 125. Xiang SW, Zhou YH. Continuity properties of solutions of vector optimization. *Nonlinear Anal* 2006;64(11):2496–2506.
 126. Weidner P. Complete efficiency and interdependencies between objective functions in vector optimization. *Z Oper Res (Math Methods Oper Res)* 1990;34(2):91–115.
 127. Cambini A, Luc DT, Martein L. Order-preserving transformations and applications. *J Optim Theory Appl* 2003;118(2):275–293.
 128. Noghin VD. Relative importance of criteria: a quantitative approach. *J Multi-Criteria Decis Anal* 1997;6:355–363.

129. Hunt BJ, Wiecek MM. Cones to aid decision making in multicriteria programming. In: Tanino T, Tanaka T, Inuiguchi M, editors. Multi-objective programming and goal programming. Berlin: Springer; 2003. pp. 153–158.
130. Hunt BJ. Multiobjective programming with convex cones: methodology and applications [PhD thesis]. Clemson University, Clemson, South Carolina; 2004. Margaret M. Wiecek, advisor.
131. Engau A. Domination and decomposition in multiobjective programming [PhD thesis]. Clemson University, Clemson, South Carolina; 2007. Margaret M. Wiecek, advisor.
132. Engau A, Wiecek MM. Introducing non-polyhedral cones into multiobjective programming. In: Barichard V, Ehrgott M, Gandibleux X, editors. Volume 618, Multiobjective programming and goal programming, Lecture notes in economics and mathematical systems. Berlin: Springer; 2009. pp. 35–45.
133. Yu P-L, Leitmann G. Compromise solutions, domination structures, and Salukvadze's solution. *J Optim Theory Appl* 1974;13: 362–378.
134. Yu P-L. Domination structures and non-dominated solutions. Multicriteria decision making, CISM International Centre for Mechanical Sciences CISM Courses and Lectures, 211. Vienna: Springer; 1975. pp. 227–280.
135. Bergstresser K, Charnes A, Yu PL. Generalization of domination structures and non-dominated solutions in multicriteria decision making. *J Optim Theory Appl* 1976;18(1): 3–13.
136. Bergstresser K, Yu PL. Domination structures and multicriteria problems in N -person games. *Theory Decis* 1977;8(1):5–48.
137. Chew KL. Domination structures in abstract spaces. Proceedings of the 1st Franco-Southeast Asian Mathematical Conference; 1979; Singapore, Vol. II, Special Issue b. 1979. pp. 190–204.
138. Weidner P. Domination sets and optimality conditions in vector optimization theory. *Wiss Z Tech Hochsch Ilmenau* 1985; 31(2):133–146. (in German).
139. Weidner P. On interdependencies between objective functions and dominance sets in vector optimization. Volume 90, 19. Jahrestagung "Mathematische Optimierung" (Sellin, 1987), Seminarberichte. Berlin: Humboldt University; 1987. pp. 136–145.
140. Weidner P. Tradeoff directions and dominance sets. Multi-objective programming and goal programming, Advances in soft computing. Berlin: Springer; 2003. pp. 275–280.
141. Hazen GB, Morin TL. Optimality conditions in nonconical multiple-objective programming. *J Optim Theory Appl* 1983; 40(1):25–60.
142. Hazen GB, Morin TL. Nonconical optimality conditions: some additional results. *J Optim Theory Appl* 1983;41(4):619–623.
143. Hazen GB, Morin TL. Steepest ascent algorithms for nonconical multiple objective programming. *J Math Anal Appl* 1984; 100(1):188–221.
144. Hazen GB. Differential characterizations of nonconical dominance in multiple objective decision making. *Math Oper Res* 1988;13(1): 174–189.
145. Chen GY, Yang XQ. Characterizations of variable domination structures via nonlinear scalarization. *J Optim Theory Appl* 2002;112(1):97–110.
146. Chen GY, Yang XQ, Yu H. A nonlinear scalarization function and generalized quasi-vector equilibrium problems. *J Glob Optim* 2005;32(4):451–466.
147. Engau A. Variable preference modeling with ideal-symmetric convex cones. *J Glob Optim* 2008;42(2):295–311.
148. Kutateladze SS. Convex ε -programming. *Sov Math Dokl* 1979;20(2):391–393.
149. Loridan P. ϵ -solutions in vector minimization problems. *J Optim Theory Appl* 1984;43(2):265–276.
150. White DJ. Epsilon efficiency. *J Optim Theory Appl* 1986;49(2):319–337.
151. Engau A. Exploring epsilon-efficiency in multiobjective programming [Master's thesis]. Clemson University, Clemson, South Carolina; 2004. Margaret M. Wiecek, advisor.
152. Engau A, Wiecek MM. Cone characterizations of approximate solutions in real vector optimization. *J Optim Theory Appl* 2007;134(3):499–513.
153. Bolintineanu S. Approximate efficiency and scalar stationarity in unbounded nonsmooth convex vector optimization problems. *J Optim Theory Appl* 2000;106(2):265–296.
154. Bolintineanu S. Vector variational principles; ϵ -efficiency and scalar stationarity. *J Convex Anal* 2001;8(1):71–85.

155. Deng S. On approximate solutions in convex vector optimization. *SIAM J Control Optim* 1997;35(6):2128–2136.
156. Dutta J, Vetrivel V. On approximate minima in vector optimization. *Numer Funct Anal Optim* 2001;22(7-8):845–859.
157. Engau A, Wiecek MM. Generating epsilon-efficient solutions in multiobjective programming. *Eur J Oper Res* 2007;177(3):1566–1579.
158. Engau A, Wiecek MM. Exact generation of epsilon-efficient solutions in multiple objective programming. *OR Spectrum* 2007;29(2):335–350.
159. Gupta P, Shiraishi S, Yokoyama K. ϵ -optimality without constraint qualification for multiobjective fractional program. *J Nonlinear Convex Anal* 2005;6(2):347–357.
160. Gutiérrez C, Jiménez B, Novo V. Multiplier rules and saddle-point theorems for Helbig's approximate solutions in convex Pareto problems. *J Glob Optim* 2005;32(3):367–383.
161. Gutiérrez C, Jiménez B, Novo V. Conditions and parametric representations of approximate minimal elements of a set through scalarization. Volume 83, *Large-scale nonlinear optimization, Nonconvex Optimization and its Applications*. New York: Springer; 2006. pp. 173–184.
162. Gutiérrez C, Jiménez B, Novo V. On approximate solutions in vector optimization problems via scalarization. *Comput Optim Appl* 2006;35(3):305–324.
163. Gutiérrez C, Jiménez B, Novo V. A unified approach and optimality conditions for approximate solutions of vector optimization problems. *SIAM J Optim* 2006;17(3):688–710.
164. Gutiérrez C, Jiménez B, Novo V. On approximate efficiency in multiobjective programming. *Math Methods Oper Res* 2006;64(1):165–185.
165. Gutiérrez C, Jiménez B, Novo V. Optimality conditions via scalarization for a new ϵ -efficiency concept in vector optimization problems. *Eur J Oper Res* 2010;201(1):11–22.
166. Helbig S, Pateva D. On several concepts for ϵ -efficiency. *OR Spektrum* 1994;16(3):179–186.
167. Kazmi KR. Existence of ϵ -minima for vector optimization problems. *J Optim Theory Appl* 2001;109(3):667–674.
168. Lemaire B. Approximation in multiobjective optimization. *J Glob Optim* 1992;2(2):117–132.
169. Liu JC. ϵ -Pareto optimality for non-differentiable multiobjective programming via penalty function. *J Math Anal Appl* 1996;198(1):248–261.
170. Liu J-C, Yokoyama K. ϵ -optimality and duality for multiobjective fractional programming. *Comput Math Appl* 1999;37(8):119–128.
171. Li Z, Wang S. ϵ -approximate solutions in multiobjective optimization. *Optimization* 1998;44(2):161–174.
172. Loridan P, Morgan J, Raucchi R. Convergence of minimal and approximate minimal elements of sets in partially ordered vector spaces. *J Math Anal Appl* 1999;239(2):427–439.
173. Miglierina E, Molho E, Patrone F, *et al.* Axiomatic approach to approximate solutions in multiobjective optimization. *Decis Econ Finance* 2008;31(2):95–115.
174. Németh AB. Between Pareto efficiency and Pareto ϵ -efficiency. *Optimization* 1989;20(5):615–637.
175. Panaitescu E, Dogaru L. On approximate solutions in multiobjective optimization. *An Univ Bucur sti Mat* 2008;57(2):285–291.
176. Rong W. Proper ϵ -efficiency in vector optimization problems with cone-subconvexlikeness. *Neimenggu Daxue Xuebao Ziran Kexue* 1997;28(5):609–613.
177. Rong WD, Wu YN. ϵ -weak minimal solutions of vector optimization problems with set-valued maps. *J Optim Theory Appl* 2000;106(3):569–579.
178. Taa A. ϵ -subdifferentials of set-valued maps and ϵ -weak Pareto optimality for multiobjective optimization. *Math Methods Oper Res* 2005;62(2):187–209.
179. Tammer C. Stability results for approximately efficient solutions. *OR Spektrum* 1994;16(1):47–52.
180. Tanaka T. Approximately efficient solutions in vector optimization. *J Multi-Criteria Decis Anal* 1996;5(4):271–278.
181. Vályi I. Approximate solutions of vector optimization problems. Volume 27, *Systems Analysis and Simulation 1985, I* (Berlin, 1985), *Mathematical Research*. Berlin: Akademie-Verlag; 1985. pp. 246–250.
182. Yokoyama K. ϵ -optimality criteria for convex programming problems via exact

- penalty functions. *Math Program Ser A* 1992;56(2):233–243.
183. Yokoyama K. ϵ -optimality criteria for vector minimization problems via exact penalty functions. *J Math Anal Appl* 1994;187(1):296–305.
184. Yokoyama K. Epsilon approximate solutions for multiobjective programming problems. *J Math Anal Appl* 1996;203(1):142–149.
185. Yokoyama K. Relationships between efficient set and ϵ -efficient set. *Nonlinear Analysis and Convex Analysis* (Niigata, 1998). River Edge (NJ): World Scientific Publishing; 1999. pp. 376–380.
186. Ruhe G, Fruhwirth B. ϵ -optimality for bicriteria programs and its application to minimum cost flows. *Computing* 1990;44(1):21–34.
187. White DJ. Epsilon-dominating solutions in mean-variance portfolio analysis. *Eur J Oper Res* 1998;105(3):457–466.
188. White DJ. Epsilon dominance and constraint partitioning in multiple objective problems. *J Glob Optim* 1998;12(4):435–448.
189. Blanquero R, Carrizosa E. A DC biobjective location model. *J Glob Optim* 2002;23(2):139–154.
190. Angel E, Bampis E, Kononov A. On the approximate tradeoff for bicriteria batching and parallel machine scheduling problems. *Theor Comput Sci* 2003;306(1-3):319–338.
191. Fadel G, Konda S, Blouin VY, *et al.* Epsilon-optimality in bi-criteria optimization. *Collect Tech Pap AIAA ASME ASCE AHS Struct Struct Dyn Mater* (43rd Structures, Structural, Dynamics and Materials Conference). Denver (CO): Volume 1, 2002. pp. 256–265.
192. Engau A, Wiecek MM. 2D decision-making for multicriteria design optimization. *Struct Multidiscip Optim* 2007;34(4):301–315.
193. Engau A, Wiecek MM. Interactive coordination of objective decompositions in multiobjective programming. *Manage Sci* 2008;54(7):1350–1363.
194. Engau A. Tradeoff-based decomposition and decision-making in multiobjective programming. *Eur J Oper Res* 2009;199(3):883–891.
195. Papadimitriou CH, Yannakakis M. On the approximability of trade-offs and optimal access of web sources (extended abstract). 41st Annual Symposium on Foundations of Computer Science (Redondo Beach (CA); 2000). Los Alamitos (CA): IEEE Comput. Soc. Press; 2000. pp. 86–92.
196. Geoffrion AM. Proper efficiency and the theory of vector maximization. *J Math Anal Appl* 1968;22:618–630.
197. Kuhn HW, Tucker AW. Nonlinear programming. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950. Berkeley, Los Angeles (CA): University of California Press; 1951. pp. 481–492.
198. Kaliszewski I. Quantitative Pareto analysis by cone separation technique. Boston (MA): Kluwer Academic Publishers; 1994.
199. Kaliszewski I, Michalowski W. Generation of outcomes with selectively bounded trade-offs. *Found Comput Decis Sci* 1995;20(2):113–122.
200. Kaliszewski I, Michalowski W. Efficient solutions and bounds on tradeoffs. *J Optim Theory Appl* 1997;94(2):381–394.
201. Kaliszewski I. Trade-offs—a lost dimension in multiple criteria decision making. *Multiple objective and goal programming* (Ustroń, 2000), *Advances in Soft Computing*. Heidelberg: Physica; 2002. pp. 115–126.
202. Henig MI, Buchanan JT. Tradeoff directions in multiobjective optimization problems. *Math Program Ser A* 1997;78(3):357–374.
203. Lee GM, Nakayama H. Generalized trade-off directions in multiobjective optimization problems. *Appl Math Lett* 1997;10(1):119–122.
204. Miettinen K, Mäkelä MM. On cone characterizations of weak, proper and Pareto optimality in multiobjective optimization. *Math Meth Oper Res* 2001;53(2):233–245.
205. Miettinen K, Mäkelä MM. On generalized trade-off directions in nonconvex multiobjective optimization. *Math Program Ser A* 2002;92(1):141–151.
206. Miettinen K, Mäkelä MM. Characterizing generalized trade-off directions. *Math Methods Oper Res* 2003;57(1):89–100.
207. Borwein JM, Zhuang DM. Super efficiency in convex vector optimization. *Z Oper Res* 1991;35(3):175–184.
208. Borwein JM, Zhuang D. Super efficiency in vector optimization. *Trans Am Math Soc* 1993;338(1):105–122.
209. Benson HP. An improved definition of proper efficiency for vector maximization with respect to cones. *J Math Anal Appl* 1979;71(1):232–241.
210. Benson HP. Efficiency and proper efficiency in vector maximization with respect

- to cones. *J Math Anal Appl* 1983;93(1): 273–289.
211. Borwein J. Proper efficient points for maximizations with respect to cones. *SIAM J Control Optim* 1977;15(1):57–63.
 212. Corley HW. On optimality conditions for maximizations with respect to cones. *J Optim Theory Appl* 1985;46(1):67–78.
 213. Henig MI. Proper efficiency with respect to cones. *J Optim Theory Appl* 1982;36(3): 387–407.
 214. Henig MI. Value functions, domination cones and proper efficiency in multicriteria optimization. *Math Program Ser A* 1990;46(2):205–217.
 215. Adán M, Novo V. Proper efficiency in vector optimization on real linear spaces. *J Optim Theory Appl* 2004;121(3):515–540.
 216. Bhattacharya DK. Proper efficiency and vector optimization on a Banach space. *Indian J Pure Appl Math* 1999;30(4):345–354.
 217. Chen GY, Rong WD. Characterizations of the Benson proper efficiency for nonconvex vector optimization. *J Optim Theory Appl* 1998;98(2):365–384.
 218. Chew KL, Choo EU. Pseudolinearity and efficiency. *Math Program* 1984;28(2):226–239.
 219. Choo EU, Atkins DR. Proper efficiency in nonconvex multicriteria programming. *Math Oper Res* 1983;8(3):467–470.
 220. Coladas Uría L. Nondominated solutions in multiobjective problems. *Trab. Estad. Invest. Oper.* 1981; 32(1):3–12.
 221. Crespi GP. Proper efficiency and vector variational inequalities. *J Inf Optim Sci* 2002;23(1):49–62.
 222. Das LN, Nanda S. Proper efficiency conditions and duality for multiobjective programming problems involving semilocally invex functions. *Optimization* 1995;34(1):43–51.
 223. Egudo RR. Proper efficiency and multiobjective duality in nonlinear programming. *J Inf Optim Sci* 1987;8(2):155–166.
 224. Fujita T. Characterization of the solutions of multiobjective linear programming with a general dominated cone. *Bull Inform Cybern* 1996;28(1):23–29.
 225. Gasimov RN. Characterization of the Benson proper efficiency and scalarization in nonconvex vector optimization. Volume 507, Multiple criteria decision making in the new millennium (Ankara, 2000), Lecture notes in Economics and Mathematical Systems systems. Berlin: Springer; 2001. pp. 189–198.
 226. Ginchev I, Guerraggio A, Rocca M. Isolated minimizers and proper efficiency for $C^{0,1}$ constrained vector optimization problems. *J Math Anal Appl* 2005;309(1):353–368.
 227. Giorgi G, Guerraggio A. Proper efficiency and generalized convexity in nonsmooth vector optimization problems. Volume 502, Generalized convexity and generalized monotonicity (Karlovasi, 1999), Lecture Notes in Economics and Mathematical Systems. Berlin: Springer; 2001. pp. 208–217.
 228. Guerraggio A, Molho E, Zaffaroni A. On the notion of proper efficiency in vector optimization. *J Optim Theory Appl* 1994;82(1):1–21.
 229. Gulati TR, Islam MA. Efficiency and proper efficiency in differentiable multiobjective programming. *Ganit* 1989;9(1-2):55–62.
 230. Gulati TR, Islam MA. Proper efficiency in linear vector maximum problems with nonlinear constraints. *J Aust Math Soc Ser A* 1989;46(2):229–235.
 231. Gulati TR, Islam MA. Efficiency and proper efficiency in nonlinear vector maximum problems. *Eur J Oper Res* 1990;44(3):373–382.
 232. Gulati TR, Talaat N. Proper efficiency in convex vector minimum problems. *J Inform Optim Sci* 1992;13(3):437–444.
 233. Huang XX, Yang XQ. On characterizations of proper efficiency for nonconvex multiobjective optimization. *J Glob Optim* 2002;23(3-4):213–231.
 234. Jahn J. A characterization of properly minimal elements of a set. *SIAM J Control Optim* 1985;23(5):649–656.
 235. Kaliszewski I. Norm scalarization and proper efficiency in vector optimization. *Found Control Eng* 1986;11(3):117–131.
 236. Kannappan P, Pandian P. Proper efficiency and generalized nonconvex duality for vector minimization problems. *Opsearch* 1997;34(2):105–115.
 237. Kim DS, Lee GM, Sach PH. Hartley proper efficiency in multifunction optimization. *J Optim Theory Appl* 2004;120(1):129–145.
 238. Lalitha CS, Arora R. Proximal proper efficiency for minimisation with respect to normal cones. *Bull Aust Math Soc* 2005;71(2): 215–224.
 239. Lee GM, Kim DS, Sach PH. Characterizations of Hartley proper efficiency in nonconvex vector optimization. *J Glob Optim* 2005;33(2):273–298.
 240. Li H-T, Li C-P. The generalized optimality conditions of vector optimization under

- Benson proper efficiency. *Pure Appl Math (Xi'an)* 2001;17(3):271–278.
241. Majumdar AAK. Counterexamples in efficiency and proper efficiency in differentiable multiobjective programming. *Ganit* 1998;18:75–89.
242. Makarov EK, Rachkovski NN. Density theorems for generalized Henig proper efficiency. *J Optim Theory Appl* 1996;91(2):419–437.
243. Makarov EK, Rachkovski NN. Unified representation of proper efficiency by means of dilating cones. *J Optim Theory Appl* 1999;101(1):141–165.
244. Mishra SK. Generalized proper efficiency and duality for a class of nondifferentiable multiobjective variational problems with V -invexity. *J Math Anal Appl* 1996;202(1):53–71.
245. Mishra SK, Mukherjee RN. Generalized convex composite multi-objective nonsmooth programming and conditional proper efficiency. *Optimization* 1995;34(1):53–66.
246. Preda V. On proper efficiency and duality for multiobjective programming problems. *Rev Roum Math Pures Appl* 1993;38(6):545–554.
247. Qiu J-H. Dual characterization and scalarization for Benson proper efficiency. *SIAM J Optim* 2008;19(1):144–162.
248. Rong W. Proper efficiency of nonconvex vector optimization problems. *Syst Sci Math Sci* 1998;11(1):18–25.
249. Sach PH. Hartley proper efficiency in multi-objective optimization problems with locally Lipschitz set-valued objectives and constraints. *J Glob Optim* 2006;35(1):1–25.
250. Sach PH, Kim DS, Lee GM. Strong duality for proper efficiency in vector optimization. *J Optim Theory Appl* 2006;130(1):139–151.
251. Sach PH, Tuan LA. Strong duality with proper efficiency in multiobjective optimization involving nonconvex set-valued maps. *Numer Funct Anal Optim* 2009;30(3-4):371–392.
252. Singh C, Hanson MA. Generalized proper efficiency in multiobjective programming. *J Inf Optim Sci* 1991;12(1):139–144.
253. Sterna-Karwat A. Approximating families of cones and proper efficiency in vector optimization. *Optimization* 1989;20(6):809–817.
254. Tamura K, Arai S. On proper and improper efficient solutions of optimal problems with multicriteria. *J Optim Theory Appl* 1982;38(2):191–205.
255. Truong XDH. Existence and density results for proper efficiency in cone compact sets. *J Optim Theory Appl* 2001;111(1):173–194.
256. Weidner P. The influence of proper efficiency on optimal solutions of scalarizing problems in multicriteria optimization. *OR Spektrum* 1994;16(4):255–260.
257. Weir T. Proper efficiency and duality for vector valued optimization problems. *J Aust Math Soc Ser A* 1987;43(1):21–34.
258. Weir T. On efficiency, proper efficiency and duality in multiobjective programming. *Asia Pac J Oper Res* 1990;7(1):46–54.
259. White DJ. Concepts of proper efficiency. *Eur J Oper Res* 1983;13(2):180–188.
260. Zălinescu C. On two notions of proper efficiency. Volume 34, *Optimization in mathematical physics (Oberwolfach, 1985)*, *Methoden Verfahren Math Phys*. Frankfurt am Main: Lang; 1987. pp. 77–86.
261. Zalmai GJ. Proper efficiency conditions and duality models for constrained multiobjective optimal control problems containing arbitrary norms. *Optimization* 1995;34(2):111–128.
262. Zhang YM. A note on equivalence between Hartley's and Benson's proper efficiency. *J Syst Sci Math Sci* 1987;7(4):380–381.
263. Zheng XY. Proper efficiency in locally convex topological vector spaces. *J Optim Theory Appl* 1997;94(2):469–486.
264. Zhuang DM. Bases of convex cones and Borwein's proper efficiency. *J Optim Theory Appl* 1991;71(3):613–620.
265. Zhuang D. Density results for proper efficiencies. *SIAM J Control Optim* 1994;32(1):51–58.
266. John F. Extremum problems with inequalities as subsidiary conditions. *Studies and Essays Presented to R. Courant on his 60th Birthday; 1948* 8 Jan. New York: Interscience Publishers, Inc.; 1948. pp. 187–204.
267. Da Cunha NO, Polak E. Constrained minimization under vector-valued criteria in finite dimensional spaces. *J Math Anal Appl* 1967;19:103–124.
268. Maruściac I. On Fritz John type optimality criterion in multi-objective optimization. *Anal Numér Théor Approx* 1982;11(1-2):109–114.
269. Simon CP. Scalar and vector maximization: calculus techniques with economics applications. Volume 25, *Studies in mathematical economics*, *MAA Studies in Mathematics*.

- Washington (DC): Mathematical Association of America; 1986. pp. 62–159.
270. Majumdar AAK. Optimality conditions in differentiable multiobjective programming. *J Optim Theory Appl* 1997;92(2):419–427.
 271. Giorgi G, Jiménez B, Novo V. On constraint qualifications in directionally differentiable multiobjective optimization problems. *RAIRO Oper Res* 2004;38(3):255–274.
 272. Klee V. A note on convex cones and constraint qualifications in infinite-dimensional vector spaces. *J Optim Theory Appl* 1982; 37(2):277–284.
 273. Li XF. Constraint qualifications in nonsmooth multiobjective optimization. *J Optim Theory Appl* 2000;106(2):373–398.
 274. Maeda T. Constraint qualifications in multiobjective optimization problems: differentiable case. *J Optim Theory Appl* 1994; 80(3):483–500.
 275. Preda V, Chişescu I. On constraint qualification in multiobjective optimization problems: semidifferentiable case. *J Optim Theory Appl* 1999;100(2):417–433.
 276. Zhou X, Sharifi Mokhtarian F, Zlobec S. A simple constraint qualification in convex programming. *Math Program Ser A* 1993;61(3):385–397.
 277. Wang SY. Second-order necessary and sufficient conditions in multiobjective programming. *Numer Funct Anal Optim* 1991;12(1-2):237–252.
 278. Aghezzaf B. Second-order necessary conditions of the Kuhn-Tucker type in multiobjective programming problems. *Control Cybern* 1999;28(2):213–224.
 279. Aghezzaf B, Hachimi M. Second-order optimality conditions in multiobjective optimization problems. *J Optim Theory Appl* 1999;102(1):37–50.
 280. Aghezzaf B, Hachimi M. Sufficient optimality conditions and duality in multiobjective optimization involving generalized convexity. *Numer Funct Anal Optim* 2001;22(7-8):775–788.
 281. Aghezzaf B, Hachimi M. Sufficient optimality conditions in multiobjective optimization problems. *Math Rech Appl* 2002;4:35–45.
 282. Ahmad I. Optimality conditions and mixed duality in nondifferentiable programming. *J Nonlinear Convex Anal* 2004;5(1):113–119.
 283. Amahroq T, Taa A. Sufficient conditions of optimality for multiobjective optimization problems with γ -paraconvex data. *Stud Math* 1997;124(3):239–247.
 284. Antczak T. Optimality conditions and duality for nondifferentiable multiobjective programming problems involving d - r -type I functions. *J Comput Appl Math* 2009;225(1):236–250.
 285. Arana-Jiménez M, Rufián-Lizana A, Osuna-Gómez R, *et al.* Pseudoinvexity, optimality conditions and efficiency in multiobjective problems; duality. *Nonlinear Anal* 2008;68(1):24–34.
 286. Bao TQ, Gupta P, Mordukhovich BS. Necessary conditions in multiobjective optimization with equilibrium constraints. *J Optim Theory Appl* 2007;135(2):179–203.
 287. Bao TQ, Mordukhovich BS. Necessary conditions for super minimizers in constrained multiobjective optimization. *J Glob Optim* 2009;43(4):533–552.
 288. Bigi G. On sufficient second order optimality conditions in multiobjective optimization. *Math Methods Oper Res* 2006;63(1): 77–85.
 289. Boţ RI, Hodrea IB, Wanka G. Optimality conditions for weak efficiency to vector optimization problems with composed convex functions. *Cent Eur J Math* 2008; 6(3):453–468.
 290. Cambini R. Second order optimality conditions in multiobjective programming. *Optimization* 1998;44(2):139–160.
 291. Cambini A, Martein L, Cambini R. Some optimality conditions in multiobjective programming. *Multicriteria analysis* (Coimbra, 1994). Berlin: Springer; 1997. pp. 168–178.
 292. Cambini A, Martein L. Some optimality conditions in vector optimization. *J Inf Optim Sci* 1989;10(1):141–151.
 293. Cambini A, Martein L, Cambini R. A new approach to second order optimality conditions in vector optimization. Volume 455, *Advances in multiple objective and goal programming* (Torremolinos, 1996), *Lecture notes in economics and mathematical systems*. Berlin: Springer; 1997. pp. 219–227.
 294. Cambini A, Martein L. First and second order optimality conditions in vector optimization. *J Stat Manage Syst* 2002;5(1-3):295–319.
 295. Censor Y. Pareto optimality in multiobjective problems. *Appl Math Optim* 1977/78; 4(1):41–59.
 296. Chen GY. Necessary conditions of non-dominated solutions in multicriteria decision making. *J Math Anal Appl* 1984; 104(1):38–46.

297. Chen X-H. Optimality conditions and duality in vector-valued optimization. *Math Appl (Wuhan)* 2003;16(2):112–117.
298. Crespi GP, La Torre D, Rocca M. Second order optimality conditions for nonsmooth multiobjective optimization problems. Volume 77, Generalized convexity, generalized monotonicity and applications, Non-convex Optimization Application. New York: Springer; 2005. pp. 213–228.
299. Ferrara M, Stefanescu MV. Optimality conditions and duality in multiobjective programming with (Φ, ρ) -invexity. *Yugosl J Oper Res* 2008;18(2):153–165.
300. Fulga C, Preda V. On optimality conditions for multiobjective optimization problems in topological vector space. *J Math Anal Appl* 2007;334(1):123–131.
301. Gadhi N. Sufficient second order optimality conditions for C^1 multiobjective optimization problems. *Serdica Math J* 2003;29(3):225–238.
302. Gadhi N. Necessary optimality conditions for Lipschitz multiobjective optimization problems. *Georgian Math J* 2005;12(1):65–74.
303. Giorgi G, Jiménez B, Novo V. Sufficient optimality conditions and duality in nonsmooth multiobjective optimization problems under generalized convexity. Volume 583, Generalized convexity and related topics, Lecture notes in economics and mathematical systems. Berlin: Springer; 2007. pp. 265–278.
304. Glover BM, Jeyakumar V, Rubinov AM. Dual conditions characterizing optimality for convex multi-objective programs. *Math Program Ser A* 1999;84(1):201–217.
305. Guerraggio A, Luc DT, Minh NB. Second-order optimality conditions for C^1 multiobjective programming problems. *Acta Math Vietnam* 2001;26(3):257–268.
306. Guerraggio A, Luc DT. Optimality conditions for $C^{1,1}$ constrained multiobjective problems. *J Optim Theory Appl* 2003;116(1):117–129.
307. Gutiérrez C, Jiménez B, Novo V. ϵ -Pareto optimality conditions for convex multiobjective programming via max function. *Numer Funct Anal Optim* 2006;27(1):57–70.
308. Gutiérrez C, Jiménez B, Novo V. Optimality conditions for metrically consistent approximate solutions in vector optimization. *J Optim Theory Appl* 2007;133(1):49–64.
309. Hachimi M, Aghezzaf B. New results on second-order optimality conditions in vector optimization problems. *J Optim Theory Appl* 2007;135(1):117–133.
310. Hu Y, Ling C. The generalized optimality conditions of multiobjective programming problem in topological vector space. *J Math Anal Appl* 2004;290(2):363–372.
311. Hu Y, Meng Z. Necessary conditions for major optimal solutions and major efficient solutions of multiobjective programming. *Syst Sci Math Sci* 1997;10(4):315–319.
312. Ishizuka Y. Optimality conditions for directionally differentiable multi-objective programming problems. *J Optim Theory Appl* 1992;72(1):91–111.
313. Jahn J. Some characterizations of the optimal solutions of a vector optimization problem. *OR Spektrum* 1985;7(1):7–17.
314. Jiménez B. Strict minimality conditions in nondifferentiable multiobjective programming. *J Optim Theory Appl* 2003;116(1):99–116.
315. Jiménez B, Novo V. First and second order sufficient conditions for strict minimality in multiobjective programming. *Numer Funct Anal Optim* 2002;23(3-4):303–322.
316. Jiménez B, Novo V, Sama M. Scalarization and optimality conditions for strict minimizers in multiobjective optimization via contingent epiderivatives. *J Math Anal Appl* 2009;352(2):788–798.
317. Jourani A. Necessary conditions for extremality and separation theorems with applications to multiobjective optimization. *Optimization* 1998;44(4):327–350.
318. Kannappan P. Necessary conditions for optimality of nondifferentiable convex multiobjective programming. *J Optim Theory Appl* 1983;40(2):167–174.
319. Khanh PQ, Tuan ND. Optimality conditions for nonsmooth multiobjective optimization using Hadamard directional derivatives. *J Optim Theory Appl* 2007;133(3):341–357.
320. Kim DS, Bae KD. Optimality conditions and duality for a class of nondifferentiable multiobjective programming problems. *Taiwan J Math* 2009;13(2B):789–804.
321. Kim DS, Lee GM, Lee BS, *et al.* Counterexample and optimality conditions in differentiable multiobjective programming. *J Optim Theory Appl* 2001;109(1):187–192.
322. Kouada I. Fritz John's type conditions and associated duality forms in convex nondifferentiable vector-optimization. *RAIRO Rech Oper* 1994;28(4):399–412.
323. Lai HC, Liu JC. A necessary and sufficient condition of convex multiobjective programming. *Tamkang J Math* 1989;20(1):7–17.

324. Lai HC, Liu JC. Optimality conditions for multiobjective programming with generalized (J, ρ, θ) -convex set functions. *J Math Anal Appl* 1997;215(2):443–460.
325. La Torre D. Optimality conditions for convex vector functions by mollified derivatives. Volume 583, Generalized convexity and related topics, Lecture notes in economics and mathematical systems. Berlin: Springer; 2007. pp. 327–335.
326. Lee GM. Optimality conditions in multiobjective optimization problems. *J Inf Optim Sci* 1992;13(1):107–111.
327. Li YX. Optimality conditions and duality theorems in nonconical multiobjective programming. *Syst Sci Math Sci* 1995;8(2):116–120.
328. Li Y. Nonconical nonsmooth multiobjective programming—optimality conditions and duality theorems. *Syst Sci Math Sci* 1998;11(1):61–68.
329. Liu L, Neittaanmäki P, Křížek M. Second-order optimality conditions for nondominated solutions of multiobjective programming with $C^{1,1}$ data. *Appl Math* 2000;45(5):381–397.
330. Liu Y, Mei J. Optimality conditions in bilevel multiobjective programming problems. *Southeast Asian Bull Math* 2009;33(1):79–87.
331. Li XF, Zhang JZ. Stronger Kuhn-Tucker type conditions in nonsmooth multiobjective optimization: locally Lipschitz case. *J Optim Theory Appl* 2005;127(2):367–388.
332. Lu J, Qin Z. Sufficient conditions for circular-cone efficient solutions of multiobjective programming. *Math Econ* 2003;20(4):86–90.
333. Maeda T. Second-order conditions for efficiency in nonsmooth multiobjective optimization problems. *J Optim Theory Appl* 2004;122(3):521–538.
334. Minami M. Weak Pareto optimality of multiobjective problem in a Banach space. *Bull Math Stat* 1980;19(3-4):19–23.
335. Minami M. Weak Pareto optimality of multiobjective problems in a locally convex linear topological space. *J Optim Theory Appl* 1981;34(4):469–484.
336. Niculescu C. Optimality conditions for multiobjective optimization involving generalized connected functions with respect to cones. *Rev Roum Math Pures Appl* 2005;50(3):293–299.
337. Popa M, Popa M. Sufficiency optimality conditions for multiobjective programming under generalized univexity type I with semidifferentiable functions. *An Univ Bucuresti Mat* 2003;52(2):235–246.
338. Preda V. On some sufficient optimality conditions in multiobjective differentiable programming. *Kybernetika (Prague)* 1992;28(4):263–270.
339. Preda V, Chițescu I, Bebu I. Kuhn-Tucker type necessary conditions for efficiency. Dini directional derivatives case. *Politehn Univ Bucharest Sci Bull Ser A Appl Math Phys* 2000;62(1):51–60.
340. Rizvi MM, Nasser M. New second-order optimality conditions in multiobjective optimization problems: differentiable case. *J Indian Inst Sci* 2006;86(3):279–286.
341. Staib T. On necessary and sufficient optimality conditions for multicriterial optimization problems. *Z Oper Res* 1991;35(3):231–248.
342. Suneja SK, Gupta S, Vani. Optimality conditions and duality in multiobjective nonlinear programming involving ρ -semilocally preinvex and related functions. *Opsearch* 2007;44(1):27–40.
343. Taa A. Necessary and sufficient conditions for multiobjective optimization problems. *Optimization* 1996;36(2):97–104.
344. Tamura K, Miura S. Necessary and sufficient conditions for local and global nondominated solutions in decision problems with multi-objectives. *J Optim Theory Appl* 1979;28(4):501–523.
345. Wan YH. On local Pareto Optima. *J Math Econom* 1975;2(1):35–42.
346. Wang SY. A note on optimality conditions in multiobjective programming. *Syst Sci Math Sci* 1988;1(2):184–190.
347. Wang LC, Dong JL, Liu QH. Optimality conditions in nonsmooth multiobjective programming. *Syst Sci Math Sci* 1994;7(3):250–255.
348. Wang C-L, Liu Q-H, Li Z-F. Optimality conditions and duality for nonsmooth multiobjective programs with generalized essential convexity. *Northeast Math J* 2008;24(5):377–385.
349. Wu H-C. The Karush-Kuhn-Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions. *Eur J Oper Res* 2009;196(1):49–60.
350. Xu Z. Necessary conditions for suboptimization over the weakly efficient set associated to generalized invex multiobjective programming. *J Math Anal Appl* 1996;201(2):502–515.

351. Xu Y, Liu S. Kuhn-Tucker necessary conditions for (h, ϕ) -multiobjective optimization problems. *J Syst Sci Complex* 2004;17(4):472–484.
352. Yang XQ, Jeyakumar V. First and second-order optimality conditions for convex composite multiobjective optimization. *J Optim Theory Appl* 1997;95(1):209–224.
353. Yuan D, Chinchuluun A, Liu X, *et al.* Optimality conditions and duality for multiobjective programming involving (C, α, ρ, d) type-I functions. Volume 583, Generalized convexity and related topics, Lecture notes in economics and mathematical systems. Berlin: Springer; 2007. pp. 73–87.
354. Zeng L-C. Necessary optimality conditions for super minimizers in structural problems of multiobjective optimization. *Asian-Eur J Math* 2009;2(2):321–358.
355. Zubiri J. A necessary condition for optimality in vector optimization problem. Volume 88, Mathematical analysis and systems theory, V, DM. Budapest: Karl Marx University of Economic; 1988. pp. 49–56.
356. Aghezzaf B. Second order mixed type duality in multiobjective programming problems. *J Math Anal Appl* 2003;285(1):97–106.
357. Aghezzaf B, Hachimi M. Generalized invexity and duality in multiobjective programming problems. *J Glob Optim* 2000;18(1):91–101.
358. Aghezzaf B, Hachimi M. Sufficiency and duality in multiobjective programming involving generalized (F, ρ) -convexity. *J Math Anal Appl* 2001;258(2):617–628.
359. Ahmad I. Multiobjective mixed symmetric duality with invexity. *N Z J Math* 2005;34(1):1–9.
360. Ahmad I. Second order symmetric duality in nondifferentiable multiobjective programming. *Inf Sci* 2005;173(1-3):23–34.
361. Ahmad I. Sufficiency and duality in multiobjective programming with generalized (F, ρ) -convexity. *J Appl Anal* 2005;11(1):19–33.
362. Ahmad I, Husain Z, Sharma S. Higher-order duality in nondifferentiable multiobjective programming. *Numer Funct Anal Optim* 2007;28(9-10):989–1002.
363. Ahmad I, Husain Z. Nondifferentiable second order symmetric duality in multiobjective programming. *Appl Math Lett* 2005;18(7):721–728.
364. Ahmad I, Husain Z. Second order (F, α, ρ, d) -convexity and duality in multiobjective programming. *Inf Sci* 2006;176(20):3094–3103.
365. Ahmad I, Husain Z. Minimax mixed integer symmetric duality for multiobjective variational problems. *Eur J Oper Res* 2007;177(1):71–82.
366. Ahmad I, Husain Z. Second order duality in multiobjective programming. *J Appl Anal* 2008;14(1):131–148.
367. Ahmad I, Sharma S. Optimality conditions and duality in nonsmooth multiobjective optimization. *J Nonlinear Convex Anal* 2007;8(3):417–430.
368. Ahmad I, Sharma S. Second-order duality for nondifferentiable multiobjective programming problems. *Numer Funct Anal Optim* 2007;28(9-10):975–988.
369. Antczak T. Saddle point criteria and duality in multiobjective programming via an η -approximation method. *ANZIAM J* 2005;47(2):155–172.
370. Antczak T. On G -invex multiobjective programming. II. Duality. *J Glob Optim* 2009;43(1):111–140.
371. Balbas A, Jimenez P, Heras A. Duality theory and slackness conditions in multiobjective linear programming. *Comput Math Appl* 1999;37(4-5):101–109.
372. Bector CR, Chandra S, Durga Prasad MV. Duality in pseudolinear multiobjective programming. *Asia Pac J Oper Res* 1988;5(2):150–159.
373. Bector CR, Chandra S, Singh C. A linearization approach to multiobjective programming duality. *J Math Anal Appl* 1993;175(1):268–279.
374. Bector CR, Bector MK, Gill A, *et al.* Duality for vector valued B -invex programming. Volume 405, Generalized convexity (Pécs, 1992), Lecture notes in economics and mathematical systems. Berlin: Springer; 1994. pp. 358–373.
375. Bector CR, Chandra S, Goyal A. On mixed symmetric duality in multiobjective programming. *Opsearch* 1999;36(4):399–407.
376. Bhatia D, Singh C, Rueda N. Generalized proper efficiency and duality in multiobjective fractional programming with Dini derivatives. Volume 8, Approximation and optimization in the Caribbean, II (Havana, 1993), Approximate Optimization. Frankfurt am Main: Lang; 1995. pp. 98–113.
377. Bhatia D, Mehra A. Optimality conditions and duality for multiobjective variational problems with generalized B -invexity. *J Math Anal Appl* 1999;234(2):341–360.

378. Bhatia D, Mehra A. Approximate saddle point and duality for multiobjective n -set optimization. *Numer Funct Anal Optim* 2003;24(5-6):475–488.
379. Bitran GR, Magnanti TL. Duality based characterizations of efficient facets. Volume 177, Multiple criteria decision making theory and application (Proceedings 3rd Conference, Hagen/Königswinter; 1979), Lecture notes in economics and mathematical systems. Berlin: Springer; 1980. pp. 12–25.
380. Bitran GR. Duality for nonlinear multiple-criteria optimization problems. *J Optim Theory Appl* 1981;35(3):367–401.
381. Boş RI, Vargyas E, Wanka G. Conjugate duality for multiobjective composed optimization problems. *Acta Math Hung* 2007;116(3):177–196.
382. Boş RI, Grad S-M, Wanka G. A general approach for studying duality in multiobjective optimization. *Math Methods Oper Res* 2007;65(3):417–444.
383. Boş RI, Wanka G. Duality for multiobjective optimization problems with convex objective functions and D.C. constraints. *J Math Anal Appl* 2006;315(2):526–543.
384. Bătaiorescu A. Sufficiency and duality in multiobjective programming involving (F, ρ) -convexity and support functions. *Math Rep (Bucur)* 2003;4(54):235–244.
385. Bătaiorescu A. Generalized duality involving V-type-I univexity and set-functions. *Rev Roum Math Pures Appl* 2004;49(5-6):433–445.
386. Bătaiorescu A, Preda V, Beldiman M. On higher-order duality for multiobjective programming involving generalized (F, ρ, γ, b) -convexity. *Math Rep (Bucur.)* 2007;9(59):161–174.
387. Cambini R, Carosi L. Duality in multiobjective optimization problems with set constraints. Volume 77, Generalized convexity, generalized monotonicity and applications, Nonconvex Optimization Application. New York: Springer; 2005. pp. 131–146.
388. Chandra S, Abha, Technical note on symmetric duality in multiobjective programming: some remarks on recent results. *Eur J Oper Res* 2000;124(3):651–654.
389. Chen X. Higher-order symmetric duality in nondifferentiable multiobjective programming problems. *J Math Anal Appl* 2004;290(2):423–435.
390. Chen X-H. Second-order symmetric duality for a multiobjective programming problem with cone constraints. *Math Appl (Wuhan)* 2006;19(1):127–133.
391. Corley HW. Duality theory for maximizations with respect to cones. *J Math Anal Appl* 1981;84(2):560–568.
392. Das L, Nanda S. Generalized invexity and duality in multiobjective nonlinear programming. *J Appl Math Comput* 2003;11(1-2):273–281.
393. Di Guglielmo F. Nonconvex duality in multiobjective optimization. *Math Oper Res* 1977;2(3):285–291.
394. Egudo RR, Weir T, Mond B. Duality without constraint qualification for multiobjective programming. *J Aust Math Soc Ser B* 1992;33(4):531–544.
395. Galperin E, Jimenez Guerra P. Duality of nonscalarized multiobjective linear programs: dual balance, level sets, and dual clusters of optimal vectors. *J Optim Theory Appl* 2001;108(1):109–137.
396. Göpfert A. Multicriterial duality, examples and advances. Volume 273, Large-scale modelling and interactive decision analysis (Eisenach, 1985), Lecture notes in economics and mathematical systems. Berlin: Springer; 1986. pp. 52–58.
397. Gulati TR, Agarwal D. Sufficiency and duality in nonsmooth multiobjective optimization involving generalized (F, α, ρ, d) -type I functions. *Comput Math Appl* 2006;52(1-2):81–94.
398. Gulati TR, Agarwal D. Second-order duality in multiobjective programming involving (F, α, ρ, d) -V-type I functions. *Numer Funct Anal Optim* 2007;28(11-12):1263–1277.
399. Gulati TR, Agarwal D. Optimality and duality in nondifferentiable multiobjective mathematical programming involving higher order (F, α, ρ, d) -type I functions. *J Appl Math Comput* 2008;27(1-2):345–364.
400. Gulati TR, Husain I, Ahmed A. Multiobjective symmetric duality with invexity. *Bull Aust Math Soc* 1997;56(1):25–36.
401. Gulati TR, Husain I, Ahmed A. Optimality conditions and duality for multiobjective control problems. *J Appl Anal* 2005;11(2):225–245.
402. Gulati TR, Ahmad I, Agarwal D. Sufficiency and duality in multiobjective programming under generalized type I functions. *J Optim Theory Appl* 2007;135(3):411–427.
403. Gulati TR, Islam MA. Sufficiency and duality in multiobjective programming involving

- generalized F -convex functions. *J Math Anal Appl* 1994;183(1):181–195.
404. Gulati TR, Talaat N. Duality in nonconvex multiobjective programming. *Asia Pac J Oper Res* 1991;8(1):62–69.
405. Hachimi M, Aghezzaf B. Second order duality in multiobjective programming involving generalized type I functions. *Numer Funct Anal Optim* 2004;25(7-8):725–736.
406. Hachimi M, Aghezzaf B. Sufficiency and duality in differentiable multiobjective programming involving generalized type I functions. *J Math Anal Appl* 2004;296(2):382–392.
407. Hachimi M, Aghezzaf B. Sufficiency and duality in nondifferentiable multiobjective programming involving generalized type I functions. *J Math Anal Appl* 2006;319(1):110–123.
408. Hsia W-S, Lee TY. Lagrangian function and duality theory in multiobjective programming with set functions. *J Optim Theory Appl* 1988;57(2):239–251.
409. Huang XX, Yang XQ. Duality and exact penalization for vector optimization via augmented Lagrangian. *J Optim Theory Appl* 2001;111(3):615–640.
410. Huang XX, Yang XQ. Nonlinear Lagrangian for multiobjective optimization and applications to duality and exact penalization. *SIAM J Optim* 2002;13(3):675–692.
411. Huang XX, Yang XQ. Duality for multiobjective optimization via nonlinear Lagrangian functions. *J Optim Theory Appl* 2004;120(1):111–127.
412. Isermann H. On some relations between a dual pair of multiple objective linear programs. *Z Oper Res* 1978;22(1):A33–A41.
413. Islam MA. Sufficiency and duality in nondifferentiable multiobjective programming. *Pure Appl Math Sci* 1994;39(1-2):31–39.
414. Jahn J. Duality in vector optimization. *Math Program* 1983;25(3):343–353.
415. Jo CL, Kim DS, Lee GM. Duality for multiobjective programming involving n -set functions. *Optimization* 1994;29(1):45–54.
416. Khan ZA. On invexity and duality in nondifferentiable multiobjective programming. *Aligarh J Stat* 1994;14:25–40.
417. Khurana S. Symmetric duality in multiobjective programming involving generalized cone-index functions. *Eur J Oper Res* 2005;165(3):592–597.
418. Kim DS. Generalized convexity and duality for multiobjective optimization problems. *J Inf Optim Sci* 1992;13(3):383–390.
419. Kim DS, Lee GM, Jo CL. Duality theorems for multiobjective fractional minimization problems involving set functions. *Southeast Asian Bull Math* 1996;20(2):65–72.
420. Kim DS, Yun YB, Kuk H. Second-order symmetric and self-duality in multiobjective programming. *Appl Math Lett* 1997;10(2):17–22.
421. Kim DS, Lee HJ, Lee YJ. Generalized second order symmetric duality in nondifferentiable multiobjective programming. *Taiwan J Math* 2007;11(3):745–764.
422. Kim DS, Kang HS, Lee YJ. Second order symmetric duality for nondifferentiable multiobjective programming involving cones. *Taiwan J Math* 2008;12(6):1347–1363.
423. Kim MH, Kim DS. Non-differentiable symmetric duality for multiobjective programming with cone constraints. *Eur J Oper Res* 2008;188(3):652–661.
424. Kim DS, Schaible S. Optimality and duality for invex nonsmooth multiobjective programming problems. *Optimization* 2004;53(2):165–176.
425. Kuk H, Tanino T. Optimality and duality in nonsmooth multiobjective optimization involving generalized type I functions. *Comput Math Appl* 2003;45(10-11):1497–1506.
426. Lai HC, Ho CP. Duality theorem of nondifferentiable convex multiobjective programming. *J Optim Theory Appl* 1986;50(3):407–420.
427. Lal SN, Nath B, Kumar A. Duality for some nondifferentiable static multiobjective programming problems. *J Math Anal Appl* 1994;186(3):862–867.
428. Lalitha CS, Suneja SK, Khurana S. Symmetric duality involving invexity in multiobjective fractional programming. *Asia Pac J Oper Res* 2003;20(1):57–72.
429. Lal SN, Kumar A. Some results on optimality and duality in multiobjective programming. *Bull Allahabad Math Soc* 1994;7:31–41.
430. Li ZF. ϵ -optimality and ϵ -duality of nonsmooth nonconvex multiobjective programming. *Neimenggu Daxue Xuebao Ziran Kexue* 1994;25(6):599–608.
431. Li Z-F, Liu Q-H, Yang R. Duality for Dini-invex nonsmooth multiobjective programming. *Northeast Math J* 2005;21(3):265–270.

432. Liu JC. ϵ -duality theorem of nondifferentiable nonconvex multiobjective programming. *J Optim Theory Appl* 1991;69(1): 153–167.
433. Liu JC. Optimality and duality for multiobjective programming involving subdifferentiable set functions. *Optimization* 1997;39(3):239–252.
434. Liu Q, Dong J, Wang L. Duality for nonsmooth pseudolinear multiobjective programming. *Math Appl (Wuhan)* 1996;9(3):395–398.
435. Liu S, Feng E. Fenchel duality theorem in multiobjective programming problems with set functions. *J Appl Math Comput* 2003;13 (1-2):139–152.
436. Luc DT. On duality theory in multiobjective programming. *J Optim Theory Appl* 1984;43(4):557–582.
437. Luc DT. About duality and alternative in multiobjective optimization. *J Optim Theory Appl* 1987;53(2):303–307.
438. Martínez-Legaz J-E. Lexicographical order and duality in multiobjective programming. *Eur J Oper Res* 1988;33(3):342–348.
439. Mishra SK, Wang SY, Lai KK. Higher-order duality for a class of nondifferentiable multiobjective programming problems. *Int J Pure Appl Math* 2004;11(2):221–232.
440. Mishra SK, Wang SY, Lai KK. Optimality and duality in nondifferentiable and multiobjective programming under generalized d -invexity. *J Glob Optim* 2004;29(4):425–438.
441. Mishra SK, Wang SY, Lai KK, *et al.* New generalized invexity for duality in multiobjective programming problems involving N -set functions. Volume 77, Generalized convexity, generalized monotonicity and applications, Nonconvex Optimization Application. New York: Springer; 2005. pp. 321–339.
442. Mishra SK, Wang SY, Lai KK. Generalized type I invexity and duality in nondifferentiable multiobjective variational problems. *Pac J Optim* 2007;3(2):309–322.
443. Mishra SK, Wang SY, Lai KK. Optimality and duality for V -invex non-smooth multiobjective programming problems. *Optimization* 2008;57(5):635–641.
444. Mishra SK, Wang SY, Lai KK. Optimality and duality for a nonsmooth multiobjective optimization involving generalized type I functions. *Math Methods Oper Res* 2008;67(3):493–504.
445. Mishra SK, Lai KK. Second order symmetric duality in multiobjective programming involving generalized cone-invex functions. *Eur J Oper Res* 2007;178(1):20–26.
446. Nahak C, Nanda S. Duality for multiobjective variational problems with invexity. *Optimization* 1996;36(3):235–248.
447. Nakayama H. Duality theory in vector optimization: an overview. Volume 242, Decision making with multiple objectives (Cleveland, Ohio, 1984), Lecture notes in economics and mathematical systems. Berlin: Springer; 1985. pp. 109–125.
448. Nakayama H. Duality in multi-objective optimization. Volume 21, Multicriteria decision making, International series of operations research and management science. Boston (MA): Kluwer Academic Publishers; 1999. pp. 3–1–329.
449. Niculescu C. On duality for a class of nondifferentiable multiobjective programming problems. *Rev Roum Math Pures Appl* 2003;47(4):445–449.
450. Niculescu C. Optimality and duality in multiobjective fractional programming involving ρ -semilocally type I-preinvex and related functions. *J Math Anal Appl* 2007;335(1):7–19.
451. Nobakhtian S. Duality without constraint qualification in nonsmooth optimization. *Int J Math Math Sci* 2006;11:Art. ID 30459.
452. Nobakhtian S. Optimality criteria and duality in multiobjective programming involving nonsmooth invex functions. *J Glob Optim* 2006;35(4):593–606.
453. Nobakhtian S. Generalized (F, ρ) -convexity and duality in nonsmooth problems of multiobjective optimization. *J Optim Theory Appl* 2008;136(1):61–68.
454. Ojha DB, Mukherjee RN. Some results on symmetric duality of multiobjective programmes with generalized (F, ρ) invexity. *Eur J Oper Res* 2006;168(2):333–339.
455. Osuna-Gómez R, Rufián-Lizana A, Ruiz-Canales P. Duality in nondifferentiable vector programming. *J Math Anal Appl* 2001;259(2):462–475.
456. Pandian P. Optimality and duality in generalized pseudolinear multiobjective programming. *Indian J Pure Appl Math* 2000;31(9):1151–1160.
457. Patel RB. Duality in multiobjective programming with generalized convex functions. *Indian J Math* 1998;40(3):331–345.
458. Patel R, Naik RK. Optimality and duality for multiobjective programming involving

- generalized b-invex functions. *Far East J Appl Math* 2008;30(3):385–407.
459. Peng J, Yang X. Duality for a class of non-smooth multiobjective programming problems. *J Syst Sci Complex* 2005;18(1):74–85.
 460. Preda V, Batătorescu A. On symmetric duality for multiobjective programming involving (η, ρ, θ) -invex functions. *An Univ Bucu sti Mat Inform* 2002;51(1):159–166.
 461. Preda V, Beldiman M. Optimality and duality for multiobjective programming involving n -set functions. *Rev Roum Math Pures Appl* 2005;50(3):301–314.
 462. Preda V, Stancu-Minasian IM, Koller E. On optimality and duality for multiobjective programming problems involving generalized d -type-I and related n -set functions. *J Math Anal Appl* 2003;283(1):114–128.
 463. Preda V, Koller E. General Mond-Weir duality for multiobjective programming with generalized (F, ρ, θ) convex set functions. *Rev Roum Math Pures Appl* 2001;45(6):1005–1018.
 464. Preda V, Stancu-Minasian IM. Mond-Weir duality for multiobjective mathematical programming of n -set functions. *Rev Roum Math Pures Appl* 2000;44(4):629–644.
 465. Preda V, Stancu-Minasian IM. Optimality and Wolfe duality for multiobjective programming problems involving n -set functions. Volume 502, *Generalized convexity and generalized monotonicity* (Karlovassi, 1999), *Lecture notes in economics and mathematical systems*. Berlin: Springer; 2001. pp. 349–361.
 466. Preda V, Stancu-Minasian IM. Mond-Weir duality for multiobjective programming problems involving d -type-I n -set functions. *Rev Roum Math Pures Appl* 2003; 47(4):499–508.
 467. Rueda NG, Mishra S. Second- and higher-order generalized invexity and duality in multiobjective programming problems. *An Univ Bucu sti Mat* 2006;54(2):257–264.
 468. Singh C. Continuous time multiobjective duality theory. *J Inf Optim Sci* 1989; 10(1):153–164.
 469. Singh C. Multiobjective programming strong vector duality. *J Inf Optim Sci* 1990; 11(2):319–326.
 470. Singh C, Bhatia D, Rueda N. Duality in nonlinear multiobjective programming using augmented Lagrangian functions. *J Optim Theory Appl* 1996;88(3):659–670.
 471. Srivastava MK, Bhatia M. Symmetric duality for multiobjective programming using second order (F, ρ) -convexity. *Opsearch* 2006;43(3):274–295.
 472. Srivastava MK, Govil MG. Second order duality for multiobjective programming involving (F, ρ, σ) -type I functions. *Opsearch* 2000;37(4):316–326.
 473. Suneja SK, Lalitha CS, Khurana S. Second order symmetric duality in multiobjective programming. *Eur J Oper Res* 2003;144(3):492–500.
 474. Suneja SK, Srivastava MK. Optimality and duality in nondifferentiable multiobjective optimization involving d -type I and related functions. *J Math Anal Appl* 1997;206(2):465–479.
 475. Tanino T. Conjugate maps and conjugate duality. Volume 289, *Mathematics of multiobjective optimization* (Udine, 1984), *CISM Courses and Lectures*. Vienna: Springer; 1985. pp. 129–155.
 476. Tanino T, Sawaragi Y. Duality theory in multiobjective programming. *J Optim Theory Appl* 1979;27(4):509–529.
 477. Tanino T, Sawaragi Y. Conjugate maps and duality in multiobjective optimization. *J Optim Theory Appl* 1980;31(4):473–499.
 478. Tarvainen K. Duality theory for preferences in multiobjective decision making. *J Optim Theory Appl* 1996;88(1):237–245.
 479. Wang S, Li Z. Scalarization and Lagrange duality in multiobjective optimization. *Optimization* 1992;26(3-4):315–324.
 480. Wanka G, Boř RI. Multiobjective duality for convex ratios. *J Math Anal Appl* 2002;275(1):354–368.
 481. Wanka G, Boř RI. A new duality approach for multiobjective convex optimization problems. *J Nonlinear Convex Anal* 2002;3(1):41–57.
 482. Wanka G, Boř RI, Grad SM. Multiobjective duality for convex semidefinite programming problems. *Z Anal Anwendungen* 2003;22(3):711–728.
 483. Wanka G, Boř R-I. Multiobjective duality for convex-linear problems. II. *Math Methods Oper Res* 2001;53(3):419–433.
 484. Weir T, Mond B. Multiple objective programming duality without a constraint qualification. *Util Math* 1991;39:41–55.
 485. Xu Z. Mixed type duality in multiobjective programming problems. *J Math Anal Appl* 1996;198(3):621–635.

486. Yan Z-X. Efficiency and duality for generalized convex multiobjective programming. *J Appl Anal* 2008;14(1):63–71.
487. Yang X. On second order symmetric duality in nondifferentiable multiobjective programming. *J Ind Manage Optim* 2009; 5(4):697–703.
488. Yang X, Wang S, Li D. Symmetric duality for a class of multiobjective programming. *Int J Math Math Sci* 2000;24(9):617–625.
489. Yang XM, Teo KL, Yang XQ. Duality for a class of nondifferentiable multiobjective programming problems. *J Math Anal Appl* 2000;252(2):999–1005.
490. Yang XM, Yang XQ, Teo KL, *et al.* Second order symmetric duality in nondifferentiable multiobjective programming with F -convexity. *Eur J Oper Res* 2005; 164(2):406–416.
491. Yang XM, Yang XQ, Teo KL. Mixed type converse duality in multiobjective programming problems. *J Math Anal Appl* 2005;304(1):394–398.
492. Yang X, Yang X, Teo KL. Higher-order symmetric duality in multiobjective programming with invexity. *J Ind Manage Optim* 2008;4(2):385–391.
493. Yang X-M, Hou S-H. Second-order symmetric duality in multiobjective programming. *Appl Math Lett* 2001;14(5):587–592.
494. Yan Z, Li S. Saddle point and generalized convex duality for multiobjective programming. *J Appl Math Comput* 2004; 15(1-2):227–235.
495. Zhang J. Higher order convexity and duality in multiobjective programming problems. Volume 30, *Progress in optimization, Application Optimization*. Dordrecht: Kluwer Academic Publishers; 1999. pp. 101–117.
496. Zhou H, Sun W. Duality and Lagrange multipliers for nonsmooth multiobjective programming. *Bull Aust Math Soc* 2006;74(3):369–383.
497. Balbás A, Galperin E, Jiménez Guerra P. Sensitivity of Pareto solutions in multiobjective optimization. *J Optim Theory Appl* 2005;126(2):247–264.
498. Balbás A, Jiménez Guerra P. Sensitivity analysis for convex multiobjective programming in abstract spaces. *J Math Anal Appl* 1996;202(2):645–658.
499. Bolintineanu S, Craven BD. Linear multicriteria sensitivity and shadow costs. *Optimization* 1992;26(1-2):115–127.
500. Craven BD. On sensitivity analysis for multicriteria optimization. *Optimization* 1988;19(4):513–523.
501. Craven BD, Luu DV. Perturbing convex multiobjective programs. *Optimization* 2000;48(4):391–407.
502. Dauer J, Liu Y-H. Multi-criteria and goal programming. Volume 6, *Advances in sensitivity analysis and parametric programming*, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 1997. pp. 11.1–11.31.
503. Ester J. About the stability of the solution of multiple objective decision making problems with generalized dominating sets. *Syst Anal Model Simul* 1984;1(4):287–291.
504. Jiménez Guerra P, Melguizo MA, Muñoz-Bouzo MJ. Sensitivity analysis in multiobjective differential programming. *Comput Math Appl* 2006;52(1-2):109–120.
505. Kuk H, Tanino T, Tanaka M. Sensitivity analysis in parametrized convex vector optimization. *J Math Anal Appl* 1996; 202(2):511–522.
506. Kuk H, Tanino T, Tanaka M. Sensitivity analysis in vector optimization. *J Optim Theory Appl* 1996;89(3):713–730.
507. Lee GM, Huy NQ. On sensitivity analysis in vector optimization. *Taiwan J Math* 2007;11(3):945–958.
508. Li D, Haimes YY. The uncertainty sensitivity index method (USIM) and its extension. *Nav Res Log* 1988;35(6):655–672.
509. Lucchetti R. Some existence results and stability in multiobjective optimization. Volume 289, *Mathematics of multiobjective optimization* (Udine, 1984), CISM courses and lectures. Vienna: Springer; 1985. pp. 179–187.
510. Rarig HM, Haimes YY. Risk/dispersion index method. *IEEE Trans Syst Man Cybern* 1983;13(3):317–328.
511. Robinson MS, Soland RM. The sensitivity analysis of “inexact” multicriteria decisions. In: Fandel G, Gal T, Hanne T, editors. Volume 448, *Multiple criteria decision making* (Hagen, 1995), *Lecture notes in economics and mathematical systems*. Berlin: Springer; 1997. pp. 185–191.
512. Shi DS. Sensitivity analysis in convex vector optimization. *J Optim Theory Appl* 1993;77(1):145–159.
513. Tanino T. Sensitivity analysis in multiobjective optimization. *J Optim Theory Appl* 1988;56(3):479–499.

514. Tanino T. Sensitivity analysis in MCDM. Volume 21, Multicriteria decision making, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 1999. pp. 7–1–729.
515. Tanino T, Kuk H, Tanaka M. Stability and sensitivity analysis in noncooperative games. In: Fandel G, Gal T, Hanne T, editors. Volume 448, Multiple criteria decision making (Hagen, 1995), Lecture notes in economics and mathematical systems. Berlin: Springer; 1997. pp. 93–102.
516. Tanino T. Stability and sensitivity analysis in convex vector optimization. *SIAM J Control Optim* 1988;26(3):521–536.
517. Venkat V, Jacobson SH, Stori JA. A post-optimality analysis algorithm for multi-objective optimization. *Comput Optim Appl* 2004;28(3):357–372.
518. Wierzbicki AP. On the completeness and constructiveness of parametric characterizations to vector optimization problems. *OR Spektrum* 1986;8(2):73–87.
519. Gerth C, Weidner P. Nonconvex separation theorems and some applications in vector optimization. *J Optim Theory Appl* 1990;67(2):297–320.
520. Das I, Dennis J. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Struct Optim* 1997;14(1):63–69.
521. Fliege J. Gap-free computation of Pareto-points by quadratic scalarizations. *Math Meth Oper Res* 2004;59(1):69–89.
522. Scott MJ, Antonsson EK. Compensation and weights for trade-offs in engineering design: Beyond the weighted sum. *J Mech Des Trans ASME* 2005;127(6):1045–1055.
523. Gass S, Saaty T. The computational algorithm for the parametric objective function. *Nav Res Log Q* 1955;2:39–45.
524. Zadeh LA. Optimality and nonscalar-valued performance criteria. *IEEE Trans Automat Control* 1963;8(1):59–60.
525. Gearhart WB. Characterization of properly efficient solutions by generalized scalarization methods. *J Optim Theory Appl* 1983;41(3):491–502.
526. Jahn J. Scalarization in vector optimization. *Math Program* 1984;29(2):203–218.
527. Graña Drummond LM, Maculan N, Svaiter BF. On the choice of parameters for the weighting method in vector optimization. *Math Program Ser B* 2008;111(1-2):201–216.
528. Sakawa M, Yano H. Trade-off rates in the hyperplane method for multiobjective optimization problems. *Eur J Oper Res* 1990;44(1):105–118.
529. Li D, Yang J-B, Biswal M. Quantitative parametric connections between methods for generating noninferior solutions in multiobjective optimization. *Eur J Oper Res* 1999;117(1):84–99.
530. Corley HW. A new scalar equivalence for Pareto optimization. *IEEE Trans Automat Control* 1980;25(4):829–830.
531. Wendell RE, Lee DN. Efficiency in multiple objective optimization problems. *Math Program* 1977;12(3):406–414.
532. Guddat J, Guerra Vasquez F, Tammer K, *et al.* Multiobjective and Stochastic Optimization Based on Parametric Optimization, Mathematical research, 26. Berlin: Akademie-Verlag; 1985.
533. Roy B. Problems and methods with multiple objective functions. *Math Program* 1971;1:239–266.
534. Gembicki FW, Haimes YY. Approach to performance and sensitivity multiobjective optimization: the goal attainment method. *IEEE Trans Automat Control* 1975;60:769–771.
535. Pascoletti A, Serafini P. Scalarizing vector optimization problems. *J Optim Theory Appl* 1984;42(4):499–524.
536. Haimes YY, Lasdon LS, Wismer DA. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Trans Syst Man Cybern* 1971;SMC-1:296–297.
537. Ruíz-Canales P, Rufián-Lizana A. A characterization of weakly efficient points. *Math Program Ser A* 1995;68(2):205–212.
538. Luc DT, Schaible S. Efficiency and generalized concavity. *J Optim Theory Appl* 1997;94(1):147–153.
539. Ehrgott M, Ruzika S. Improved ϵ -constraint method for multiobjective programming. *J Optim Theory Appl* 2008;138(3):375–396.
540. Luenberger DG. An approach to nonlinear programming. *J Optim Theory Appl* 1973;11:219–227.
541. Luenberger DG. Linear and Nonlinear Programming, 2nd ed. Reading (MA): Addison-Wesley Publishing Company; 1984.

542. Wierzbicki AP. Reference point approaches. Volume 21, Multicriteria decision making, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 1999. pp. 9–1–939.
543. Zeleny M. Compromise programming. In: Cochrane J, Zeleny M, editors. Multiple criteria decision making. Columbia: University of South Carolina Press; 1973. pp. 262–301.
544. Terlaky T. On l_p programming. *Eur J Oper Res* 1985;22(1):70–100.
545. Bowman VJ. On the relationship of the Tchebycheff norm and the efficient frontier of multiple-criteria objectives. In: Thirez H, Zionts S, editors. Multiple criteria decision making, Lecture notes in economics and mathematical systems, 130. Berlin: Springer; 1976. pp. 76–86.
546. Steuer RE, Choo EU. An interactive weighted Tchebycheff procedure for multiple objective programming. *Math Program* 1983;26(3):326–344.
547. Kaliszewski I. A modified weighted Tchebycheff metric for multiple objective programming. *Comput Oper Res* 1987;14(4):315–323.
548. Schandl B, Klamroth K, Wiecek MM. Norm-based approximation in bicriteria programming. *Comput Optim Appl* 2001;20(1):23–42.
549. Schandl B, Klamroth K, Wiecek MM. Introducing oblique norms into multiple criteria programming. *J Glob Optim* 2002;23(1):81–97.
550. Schandl B, Klamroth K, Wiecek MM. Norm-based approximation in multicriteria programming. *Comput Math Appl* 2002;44(7):925–942.
551. Lin JG. On min-norm and min-max methods of multi-objective optimization. *Math Program* 2005;103:1–33.
552. Kouvelis P, Yu G. Robust discrete optimization and its applications, Nonconvex optimization and its applications, 14. Dordrecht: Kluwer Academic Publishers; 1997.
553. Yano H, Sakawa M. Trade-off rates in the weighted Tchebycheff norm method. *Large Scale Syst* 1987;13(2):167–177.
554. Eichfelder G. Adaptive scalarization methods in multiobjective optimization, Vector Optimization. Berlin: Springer; 2008.
555. Eichfelder G. An adaptive scalarization method in multiobjective optimization. *SIAM J Optim* 2008;19(4):1694–1718.
556. Eichfelder G. Scalarizations for adaptively solving multi-objective optimization problems. *Comput Optim Appl* 2009;44(2):249–273.
557. Drummond LMG, Iusem AN. A projected gradient method for vector optimization problems. *Comput Optim Appl* 2004;28(1):5–29.
558. Graña Drummond LM, Svaiter BF. A steepest descent method for vector optimization. *J Comput Appl Math* 2005;175(2):395–414.
559. Fischer A, Shukla PK. A Levenberg-Marquardt algorithm for unconstrained multicriteria optimization. *Oper Res Lett* 2008;36(5):643–646.
560. Fliege J, Svaiter BF. Steepest descent methods for multicriteria optimization. *Math Methods Oper Res* 2000;51(3):479–494.
561. Fliege J. An efficient interior-point method for convex multicriteria optimization problems. *Math Oper Res* 2006;31(4):825–845.
562. Fliege J, Graña Drummond LM, Svaiter BF. Newton's method for multiobjective optimization. *SIAM J Optim* 2009;20(2):602–626.
563. Tlas M, Ghani BA. A logarithmic barrier function method for solving nonlinear multiobjective programming problems. *Control Cybern* 2005;34(2):487–504.
564. Bonnel H, Iusem AN, Svaiter BF. Proximal methods in vector optimization. *SIAM J Optim* 2005;15(4):953–970.
565. Ceng L-C, Yao J-C. Approximate proximal methods in vector optimization. *Eur J Oper Res* 2007;183(1):1–19.
566. Chen Z, Zhao K. A proximal-type method for convex vector optimization problem in Banach spaces. *Numer Funct Anal Optim* 2009;30(1-2):70–81.
567. Das I, Dennis JE. Normal-boundary intersection: a new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J Optim* 1998;8(3):631–657.
568. Sosa W, Raupp FMP. A new approach to a multicriteria optimization problem. *Numer Algorithms* 2004;35(2-4):233–247.
569. Recchioni MC. A path following method for box-constrained multiobjective optimization with applications to goal programming problems. *Math Methods Oper Res* 2003;58(1):69–85.

NONSTATIONARY INPUT PROCESSES

MICHAEL E. KUHL
Industrial & Systems Engineering
Department, Rochester Institute of
Technology, Rochester, New York

MODELING AND SIMULATING NONSTATIONARY INPUT PROCESSES

Nonstationary input processes are often encountered in the application of systems simulation. The arrival patterns of customers to a bank, incoming calls to a call center, arrival of patients to a hospital emergency room, demands for seasonal products such as lawn mowers, and the arrival of passengers to an airport are all examples of situations where the arrival rates may change over time. To successfully simulate the operation of these types of systems, it often requires stochastic input process that can accurately mimic the arrival pattern and changes in the arrival rates over time. This article focuses on two methods for representing nonstationary input processes, namely, nonhomogeneous Poisson processes and nonstationary non-Poisson processes.

Nonhomogeneous Poisson processes (NHPPs) have been used successfully to model complex time-dependent arrival processes in a broad range of applications including the arrival of events in a data base system [1], the arrival of storms to an off-shore drilling site [2], and the arrival patterns of organ transplant patients and organ donors in a simulation model used to analyze liver-transplant policies in the United States [3,4]. The behavior of NHPPs includes independence of the number of arrivals in nonoverlapping intervals of time, individual arrivals, and other characteristics that are defined in detail in the next section. Although there is a large class of nonstationary arrival processes that have the characteristics of

NHPPs, there also exists situations where other methods are needed. Research into these processes has been ongoing. In particular, Gerhardt and Nelson [5] have developed methods for modeling and simulating nonstationary non-Poisson arrival processes.

In this article we discuss various methods for modeling and simulating NHPPs as well as nonstationary non-Poisson process with emphasis on their advantages and limitations in practice. The content of this article includes material presented in a series of input modeling tutorials presented at the Winter Simulation Conference [6] as well as from Kuhl and Wilson [7].

NONHOMOGENEOUS POISSON PROCESSES

A nonhomogeneous Poisson process $\{N(t) : t \geq 0\}$ is a generalization of a Poisson process in which the instantaneous arrival rate $\lambda(t)$ at time t is a nonnegative integrable function of time. The mean-value function of the NHPP is defined by

$$\mu(t) \equiv E[N(t)] = \int_0^t \lambda(z) dz \text{ for all } t \geq 0.$$

An NHPP has the following properties [8]:

1. $\{N(t) : t \geq 0\}$ is a counting process;
2. $N(0) = 0$;
3. $\{N(t+s) - N(t) : s \geq 0\}$ is independent of $\{N(u) : 0 \leq u < t\}$ for all $t \geq 0$;
4. $\Pr\{N(t+h) - N(t) = 1\} = \lambda(t)h + o(h)$ for all $t \geq 0$ and $h > 0$; and
5. $\Pr\{N(t+h) - N(t) \geq 2\} = o(h)$ for all $t \geq 0$ and $h > 0$.

In the above properties, the function $f(\cdot)$ is said to be $o(h)$ if $\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0$. From these properties we observe that the rate or mean-value function of the NHPP $\{N(t) : t \geq 0\}$ completely characterizes the probabilistic behavior of the process. Thus the objective of modeling an NHPP is to define a rate function that closely represents the behavior

of the arrival process under study. In the next section we discuss nonparametric, parametric, and semiparametric methods for estimating an NHPP from observed arrival streams and for generating independent realizations of the fitted NHPP.

Nonparametric NHPP Methods

Leemis [9–11] develops a nonparametric method for estimating and simulating an NHPP. This method utilizes multiple observations of the arrival process over a given time interval $[0, S]$. The observed arrival times of these process realizations are superimposed, and the mean-value function over the observation interval is represented by a piecewise-linear estimator. The details of the method are as follows.

Given k independent realizations of the arrival process over the interval $[0, S]$. Let n_i be the number of arrivals observed on the i th realization for $i = 1, 2, \dots, k$. In addition, let $n = \sum_{i=1}^k n_i$ be the total number of arrivals accumulated over all realizations of the arrival process, and let $\{t_{(i)} : i = 1, \dots, n\}$ denote the overall set of arrival times for all arrivals expressed as an offset from the beginning of the observation interval $(0, S]$. Then sort the set of arrival times in increasing order represented by $t_{(0)} \equiv 0$ and $t_{(n+1)} \equiv S$ so that for $t_{(i)} < t \leq t_{(i+1)}$ and $i = 0, 1, \dots, n$. The resulting piecewise linear nonparametric estimator of $\mu(t)$ is

$$\hat{\mu}(t) = \frac{in}{(n+1)k} + \left\{ \frac{n[t - t_{(i)}]}{(n+1)k[t_{(i+1)} - t_{(i)}]} \right\}; \quad (1)$$

see Leemis [9]. Figure 1 depicts the structure of the estimator $\hat{\mu}(t)$.

Given the estimated mean-value function $\hat{\mu}(t)$ of the form (1), the inversion algorithm of Leemis [9] as displayed in Fig. 2 can be used to generate a stream of arrival times $\{A_i : i = 1, 2, \dots\}$ over the time interval $(0, S]$ with approximately the same general pattern of dependence on time as in Equation (1).

The main advantage of this approach to modeling and simulating NHPPs is that it does not require the assumption of any particular functional form for the way in which

the arrival rate $\lambda(t)$ depends on the time. Furthermore as the number of realizations k of the target arrival process becomes large, that is as $k \rightarrow \infty$, the estimated mean-value function $\hat{\mu}(t)$ of Equation (1) converges to the true mean-value function $\mu(t)$ for all $t \in (0, S]$ with probability 1. This means that the simulation algorithm given in Fig. 2 (which is based on inversion of $\hat{\mu}(t)$ so that $A_i = \hat{\mu}^{-1}(B_i)$ for $i = 1, \dots, N$) is also asymptotically valid as $k \rightarrow \infty$. For additional details on this approach to modeling and simulation of NHPPs, see Leemis [11].

Parametric NHPP Models

Parametric models for the rate function of an NHPP have been developed to represent arrival processes that exhibit systematic changes in the arrival rate over time such as long term trends or periodicities.

To represent arrival processes having a long term linear trend, Cox and Lewis [12] introduce an exponential linear rate function of the form

$$\lambda(t) = \exp\{\alpha_0 + \alpha_1 t\}. \quad (2)$$

This rate function has an exponential form to ensure that the arrival rate always remains positive. To represent more general long-term trends as well as cyclic behavior with a known frequency ω , Lewis [13] introduces a rate function of the form

$$\lambda(t) = \exp \left\{ \sum_{i=0}^2 \alpha_i t^i + \gamma \sin(\omega t + \phi) \right\}. \quad (3)$$

MacLean [14] proposes that any continuous rate function can be approximated arbitrarily closely using an exponential polynomial rate function of degree m

$$\lambda(t) = \exp \left\{ \sum_{i=0}^m \alpha_i t^i \right\}. \quad (4)$$

To model the occurrence of storms to an offshore drilling site, Lee *et al.* [2] use an NHPP with rate function of the form

$$\lambda(t) = \exp \left\{ \sum_{i=0}^m \alpha_i t^i + \gamma \sin(\omega t + \phi) \right\}. \quad (5)$$

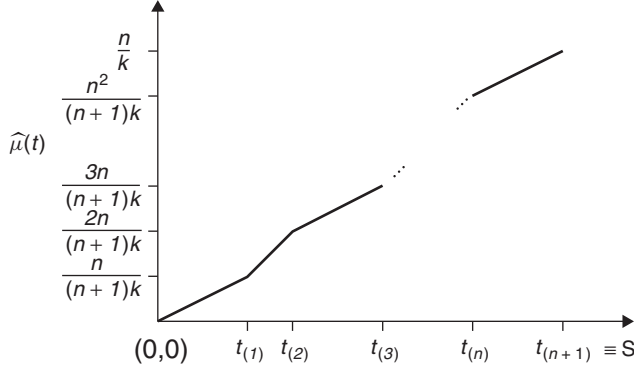


Figure 1. Nonparametric estimator of mean-value function.

```

[1] Set  $i \leftarrow 1$  and  $N \leftarrow 0$ .
[2] Generate  $U_i \sim \text{Uniform}(0,1)$ .
[3] Set  $B_i \leftarrow -\ln(1 - U_i)$ .
[4] While  $B_i < n/k$  do
  Begin
    Set  $m \leftarrow \left\lfloor \frac{(n+1)kB_i}{n} \right\rfloor$ ;
    Set  $A_i \leftarrow t_{(m)} + \{t_{(m+1)} - t_{(m)}\} \left\{ \frac{(n+1)kB_i}{n} - m \right\}$ ;
    Set  $N \leftarrow N + 1$ ; Set  $i \leftarrow i + 1$ ;
    Generate  $U_i \sim \text{Uniform}(0,1)$ ;
    Set  $B_i \leftarrow B_{i-1} - \ln(1 - U_i)$ .
  End

```

Figure 2. Algorithmic statement of the NHPP simulation procedure of Leemis [9].

For efficient simulation of an NHPP with the rate function (5), the authors develop maxline, an algorithm to construct a piecewise linear majorizing rate function $\tilde{\lambda}(t)$ that closely approximates $\lambda(t)$ on $[0, S]$; that is, $\tilde{\lambda}(t) \geq \lambda(t)$ for all $t \in [0, S]$. The advantage of this approach is that the corresponding mean-value function $\tilde{\mu}(t) = \int_0^t \tilde{\lambda}(u) du$ is piecewise quadratic and thus, is easily inverted so that a sequence of arrivals can be easily generated by the method of inversion. Then applying the thinning scheme of Lewis and Shedler [15], the authors develop an algorithm to generate arrival epochs $\{A_\ell\}$ from an NHPP with rate function $\lambda(t)$ as follows: the epoch $\tilde{A}_n$ is independently accepted for inclusion in $\{A_\ell\}$ with probability $\lambda(\tilde{A}_n) / \tilde{\lambda}(\tilde{A}_n)$ for $n = 1, 2, \dots$.

Kuhl *et al.* [16] and Kuhl and Wilson [17] extended the work of Lee *et al.* [2] to handle arrival processes with a general trend

or multiple periodicities. The model for the arrival process is an NHPP with rate function of the form

$$\lambda(t) = \exp\{h(t; m, p, \Theta)\}, \quad t \in (0, S], \quad (6)$$

with

$$h(t; m, p, \Theta) = \sum_{i=0}^m \alpha_i t^i + \sum_{k=1}^p \gamma_k \sin(\omega_k t + \phi_k),$$

where $\Theta = [\alpha_0, \alpha_1, \dots, \alpha_m, \gamma_1, \dots, \gamma_p, \phi_1, \dots, \phi_p, \omega_1, \dots, \omega_p]$ is the vector of continuous parameters. Kuhl *et al.* [16] use a form of inversion for simulating NHPPs with a rate function of the form (6). Rate functions of this type were originally used in the UNOS Liver Allocation Model [3]; and although the resulting fits were remarkably accurate, the times to generate realizations of the fitted NHPPs were too large in practice.

An extension of the maxline procedure of Lee *et al.* [2] to yield a piecewise linear rate function that majorizes Equation (6) could facilitate rapid simulation of such NHPPs based on thinning, but substantial effort may be required to develop such a procedure.

Semiparametric NHPP Methods

Semiparametric methods for representing NHPPs have been developed in an attempt to combine the flexibility of nonparametric methods with the smooth, continuous representations of parametric models. Kuhl and Wilson [18] develop a nonparametric method for modeling and simulating arrival processes that may exhibit a long-term trend or nested periodic effects (such as daily and weekly cycles), where the periodic effects might not necessarily be symmetric. The method is referred to as a *multiresolution* procedure.

The multiresolution procedure of Kuhl *et al.* [19] involves the following steps at each resolution level corresponding to a basic cycle: (i) transforming the cumulative relative frequency of arrivals within the cycle to obtain a statistical model with approximately normal, constant-variance responses; (ii) fitting a specially formulated polynomial to the transformed responses; (iii) performing a likelihood ratio test to determine the degree of the fitted polynomial; and (iv) fitting to the original (untransformed) responses a polynomial of the same form as in (ii) with the degree determined in (iii). Given the fitted functions at each resolution, the multiresolution model is constructed for the mean-value function.

Extending the semiparametric method of Kuhl *et al.* [19] to model more general nonstationary behavior, Kuhl *et al.* [20] introduce an alternative method to model the mean-value function of the NHPP. The final estimate is obtained for the mean-value function of the following form:

$$\mu(t) = \mu(S)R(t) \quad \text{for } t \in [0, S], \quad (7)$$

where $R(t)$ is a monotone increasing degree- r polynomial of the form

$$R(t) = \begin{cases} t/S, & \text{if } r = 1, \\ \sum_{k=1}^{r-1} \beta_k (t/S)^k + \left(1 - \sum_{k=1}^{r-1} \beta_k\right) (t/S)^r, & \\ & \text{if } r > 1. \end{cases}$$

This methodology can be used to fit an NHPP to one or more realizations of data from the process under study. The primary advantage of this method is that it fits a smooth (differentiable) mean-value function over the interval $[0, S]$.

The fitting and simulation methods of Kuhl *et al.* [19,20] have been implemented in Web-based software, which is available on-line via <http://www.rit.edu/simulation>.

NONSTATIONARY NON-POISSON ARRIVAL PROCESSES

Although NHPPs are used to model a large class of nonstationary arrival processes, there are applications where NHPPs do not adequately represent the arrival process of interest. For an NHPP, the random variable $N(t)$ has the Poisson distribution with mean $\mu(t)$ for all $t > 0$, we also have $\text{Var}[N(t)] = \mu(t)$ for all $t > 0$; and thus, an NHPP has $\text{Var}[N(t)]/\text{E}[N(t)] = 1$ for all $t > 0$. In many studies of manufacturing systems, telecommunications networks, and consumers' purchasing behavior, the relevant point process exhibits a variance-to-mean ratio that is either substantially more or less than that of an NHPP.

Gerhardt and Nelson [5] propose methods for modeling and simulating a point process that allow the user to achieve desired values of both the mean-value function and the ratio $\text{Var}[N(t)]/\text{E}[N(t)]$ of the process. Starting from a stationary renewal process $\{N^\circ(t)\}$ with rate 1 and the desired ratio $\lim_{t \rightarrow \infty} \text{Var}[N^\circ(t)]/\text{E}[N^\circ(t)] = C$, a corresponding sequence of renewal epochs $\{A^\circ n : n = 1, 2, \dots\}$ for the process $\{N^\circ(t)\}$ is generated. Finally a point process $\{N(t)\}$ is obtained with the desired mean-value function $\mu(t)$ and the corresponding arrival epochs $\{A_n : n = 1, 2, \dots\}$ by inversion: $A_n = \mu^{-1}(A^\circ n)$ for $n = 1, 2, \dots$. The inversion

procedure yields $E[N(t)] = \mu(t)$ for all $t \geq 0$ and $\lim_{t \rightarrow \infty} \text{Var}[N(t)]/E[N(t)] = C$.

Gerhardt and Nelson [5] also propose a method for achieving a point process $\{N(t)\}$ with the desired mean-value function $\mu(t)$ based on thinning. If the corresponding rate function $\lambda(t) = \frac{d}{dt}\mu(t)$ has finite upper bound λ^* , then the authors start from a stationary renewal process $\{N^*(t)\}$ with rate λ^* , arrival epochs $\{A_n^* : n = 1, 2, \dots\}$, and the desired variance-to-mean ratio C ; and the corresponding sequence of arrival epochs $\{A_\ell : \ell = 1, 2, \dots\}$ from a point process with the desired mean-value function $\mu(t)$ is obtained as follows: the epoch A_n^* is independently accepted for inclusion in $\{A_\ell\}$ with probability $\lambda(A_n^*)/\lambda^*$ for $n = 1, 2, \dots$. The authors show that the resulting thinned point process $\{N(t) : t \geq 0\}$ has mean-value function $E[N(t)] = \mu(t)$ for all $t \geq 0$.

CONCLUSIONS

In this article we have surveyed nonparametric, parametric, and semiparametric methods for modeling and simulating NHPPs; and we have also reviewed similar procedures for handling nonstationary non-Poisson arrival processes. Although these methods can be used to model a large class of nonstationary point process, research into more flexible and more efficient methods for both modeling and simulating these stochastic processes is ongoing.

REFERENCES

1. Lewis PAW, Shedler GS. Statistical analysis of non-stationary series of events in a data base system. *IBM J Res Dev* 1976;20: 465–482.
2. Lee S, Wilson JR, Crawford MM. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. *Commun Stat Simulat* 1991;20:777–809.
3. Pritsker AAB, Martin DL, Reust JS, Wagner MA, Daily OP, Harper AM, *et al.* Organ transplantation policy evaluation. In: Alexopoulos C, Kang K, Lilegdon WR, Goldsman D, editors. Proceedings of the 1995 winter simulation conference. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 1995. pp. 1314–1323.
4. Harper AM, Taranto SE, Edwards EB, Daily OP. An update on a successful simulation project: The UNOS liver allocation model. In: Joines JA, Barton RR, Kang K, Fishwick PA, editors. Proceedings of the 2000 winter simulation conference. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers; 2000. pp. 1955–1962.
5. Gerhardt I, Nelson BL. Transforming renewal processes for simulation of nonstationary arrival processes. *INFORMS J Comput* 2009;630–640.
6. Kuhl ME, Ivey J, Lada EK, Steiger NS, Wagner MA, Wilson JR. Introduction to modeling and generating probabilistic input processes for simulation. In: Rossetti MD, Hill RR, Johansson B, Dunkin A, Ingalls RG, editors. Proceedings of the 2009 winter simulation conference. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 2009. pp. 184–1202.
7. Kuhl ME, Wilson JR. Advances in modeling and simulation of nonstationary arrival processes. In: Lee LH, Kuhl ME, Fowler JW, Robinson S, editors. Proceedings of the 2009 INFORMS simulation society research workshop. INFORMS Simulation Society; 2009. pp. 1–5.
8. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
9. Leemis LM. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Manage Sci* 1991;37(7):886–900.
10. Leemis LM. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Manage Sci* 2000;46(7): 989–998.
11. Leemis LM. Nonparametric estimation and variate generation for a nonhomogeneous Poisson process from event count data. *IIE Trans* 2004;36(12):1155–1160.
12. Cox DR, Lewis PAW. The statistical analysis of series of events. London: Chapman and Hall; 1966.
13. Lewis PAW. Recent results in the statistical analysis of univariate point processes. In: Lewis PAW, editor. Stochastic point processes. New York: John Wiley & Sons; 1972. pp. 1–54.
14. MacLean CJ. Estimation and testing of an exponential polynomial rate function within the nonstationary poisson process. *Biometrika* 1974;61(8):1–85.

15. Lewis PAW, Shedler GS. Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res Logist Q* 1979;26:403–413.
16. Kuhl ME, Wilson JR, Johnson MA. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Trans* 1997;29(3):201–211.
17. Kuhl ME, Wilson JR. Least squares estimation of nonhomogeneous Poisson processes. *J Stat Comput Simulat* 2000;67:75–108.
18. Kuhl ME, Wilson JR. Modeling and simulating Poisson processes having trends or nontrigonometric cyclic effects. *Eur J Oper Res* 2001;133(3):566–582.
19. Kuhl ME, Sumant SG, Wilson JR. An automated multiresolution procedure for modeling complex arrival processes. *INFORMS J Comput* 2006;18(1):3–18.
20. Kuhl ME, Deo S, Wilson JR. Smooth flexible models of nonhomogeneous Poisson processes using one or more process realizations. In: Mason SJ, Hill RR, Mönch L, Rose O, Jefferson T, Fowler JW. *Proceedings of the 2008 winter simulation conference*. Piscataway (NJ): Institute of Electrical and Electronics Engineers; 2008. pp. 353–361.

NURSE SCHEDULING MODELS

JONATHAN F. BARD
Graduate Program in Operations
Research and Industrial
Engineering, The University of
Texas, Austin, Texas

INTRODUCTION

Nurse scheduling has become the prototypical example of personnel scheduling in the service industry. Unlike manufacturing, where standard shifts and days off are the rule, service organizations often operate 24 h a day, 7 days a week, and face widely fluctuating demand. In hospitals and other health-care facilities, improperly balanced nursing teams are too often the explanation for excessive medical errors, substandard patient care, job dissatisfaction, and high turnover rates [1]. What sets the nurse scheduling problem apart from other staffing applications is the need to take into account an increasing variety of skill requirements, labor restrictions, union contracts, individual preferences, and uncertain demand, all of which conspire to frustrate quantitative approaches.

In the most general sense, personnel planning and scheduling problems can be viewed temporally and decomposed along the time axis. This suggests the following hierarchical breakdown:

- *Long-Term Planning* fix the size and composition of the permanent workforce.
- *Midterm Scheduling* determine days off, shift assignments, and optimal skill mixes.
- *Short-Term Scheduling* assign tasks to workforce, manage vacations and leave, plan for the use of overtime, casuals, and part-time labor.
- *Real-Time Control* adjust workforce to deal with emergencies, absenteeism, and other last-minute disruptions.

In the long run, the goal is to structure the permanent workforce. This may include different levels of full-timers, part-timers, casuals, and per diem nurses, each with different skill sets. The demand data used for the analysis are typically average values and hence do not directly take into account seasonal or weekly variations.

The output of the long-term analysis is the number of nurses (by type) that are required for each unit or ward, along with a set of shift profiles. Hospital units use this information as input to the midterm scheduling problem, the next level in the hierarchy. The primary goal at this level is to generate nurse rosters, that is, patterns of days on and days off over a 2- to 6-week planning horizon. In the more traditional case, fixed patterns of days on and days off are established and the staff is rotated continuously through them; for example, see Ref. 2. At this level, *rostering* and *scheduling* are used interchangeably.

Alternatively, many hospital units have turned to methods that give individuals more control over their assignments and management more flexibility in the face of shortages. *Self-scheduling*, for example, uses a sign-up procedure [3]. Nurses are asked to sign up for those shifts that they wish to work over the planning horizon. When conflicts occur, the nurse manager tries to resolve them through consensus, often a difficult process.

A second approach, *preference scheduling*, applies a common set of rules and a cost measure that together are designed to achieve a balance between staff satisfaction and the use of external resources. The rules (constraints) are hospital-dependent but generally may be categorized as either hard or soft. When the rosters are maintained from one planning period to the next, we have what is sometimes called *cyclic scheduling*. In today's environment, it is more common

for rosters to be rebuilt from scratch, starting from a template or sign-up schedule.

In the short run, schedule adjustments must be made on a weekly and daily basis to account for planned leave and expected departures from average demand. To accomplish this, critical resources must be tracked and evaluated. The goal is to provide updated schedules that balance overtime and the use of part-time labor so that all requirements are met at minimum cost and without too much disruption to the midterm rosters. This can be thought of as a replanning problem.

Given the current schedules, further adjustments and decisions need to be made on a shift-by-shift basis to accommodate changing circumstances. When supply exceeds demand in a unit, a nurse may be asked to float to another unit or may be sent home. When a unit is short of staff, the hospital has several options such as requesting overtime, calling in casuals, asking a nurse to work on his or her day off, or requesting outside resources. The outside resources include a float pool and agency nurses. Each option has different cost consequences.

At the bottom level of the hierarchy, a real-time scheduling problem that falls under the heading of *control or disruption management* must be solved. The goal is to get back on track as soon as possible, at minimum cost and with minimum deviation from the original schedule.

The idea of decomposing planning problems along the time axis is an effective way of dealing with uncertainty. At each level, solutions are obtained using the best estimates of the uncertain parameters. As the time horizon shrinks and the uncertainty unfolds, adjustments are made; however, all higher level decisions remain in effect and become inputs at subsequent levels. An alternative to the hierarchical approach might be an integrated approach that accounts for the stochastic nature of demand in a single model. Although the field of stochastic optimization is growing, there have been few, if any, attempts to formulate and solve stochastic versions of the nurse scheduling problem due to its size and complexity [for a stochastic model developed for assigning nurses to patients, see Ref. 4].

Whether specialized to nurses or to other service sector workers, personnel scheduling problems are generally NP-hard in the strong sense [5]. In this article, we focus mainly on the midterm problem, which has received the lion's share of attention, but also mention some recent advances related to short-term adjustments. The long-term and real-time problems are not as well studied with respect to nurse staffing, so there is little to say [see Ref. 6 for one approach to estimating long-term nurse staffing needs, and Ref. 7 for a different application]. In the next section, the basic elements of the staffing problem are introduced, followed by an overview of models for midterm and short-term scheduling. The relevant literature is summarized in the section titled "Past and Current Research." We close with an assessment of current capabilities and suggest several topics for future research.

MODELING ISSUES

Hospitals in the United States typically employ full-time nurses to work either 8- or 12-h shifts, giving rise to five standard shift types: three 8-hour shifts called *Day* (7 a.m.–3 p.m.), *Evening* (3 p.m.–11 p.m.), *Night* (11:00 p.m.–7 a.m.) and two 12-hour shifts called *AM* (7 a.m.–7 p.m.) and *PM* (7 p.m.–7 a.m.). An AM shift starts at the same time as a Day shift and ends midway into the Evening shift. A similar interpretation exists for a PM shift. Part-time nurses may work a full shift or a fraction thereof. European hospitals have a parallel structure but may use up to 10 standard shift types.

Staffing requirements for a unit depend on the census (typically taken at midnight), the acuity or level of infirmity of the patients, and the time of day. Nurse rosters are usually generated 2 weeks prior to the start of the upcoming planning period which may extend anywhere from 2 to 6 weeks. Adjustments are made weekly and daily to account for call outs and other disruptions.

For modeling efficiency, it is convenient to divide the day into a set of contiguous (hourly) periods of uneven length such that each shift consists of one or more periods.

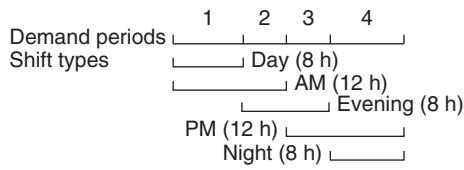


Figure 1. Shift types and demand periods.

By doing so, we can express the demand in terms of periods rather than shifts and hence reduce the size (number of rows) of the model. Figure 1 depicts the relationship between shifts and periods for our problem. As can be seen, only four periods are needed to cover 24 h. The first demand period coincides with the Day shift, the second period covers the first half of the Evening shift, the third covers the second half of the Evening shift, and the last coincides with the Night shift.

In developing a model, coverage constraints are needed to account for demand by skill category for each period, regardless of timescale. The nursing staff at most US hospitals includes registered nurses (RNs), licensed practical nurses (LPNs), technicians (TECHs), and nurse aides (AIDEs). Each category may be further divided by specialty and experience, the latter being a chief determinant of quality of care. RNs are the most versatile and are generally preferred because they can provide the widest range of care. LPNs are less flexible, while AIDEs have only limited training and skills. TECHs are commonly called upon to operate certain medical instruments specific to a unit, such as cardiology. For purposes of matching skills with requirements, some percentage of the average daily demand is assigned to each nurse category. The specification of this percentage is a management decision that represents a tradeoff between cost, availability, and quality of service. European health-care systems follow similar procedures and likewise group nurses by skill, but use more grades within each category.

MIDTERM SCHEDULING

When constructing midterm rosters, the process can be confined to a single nursing unit at

a time since there are few interdependencies among them. In addition, each skill category can generally be scheduled separately. When higher skilled personnel are permitted to fill in for lower skilled personnel, the process is called *downgrading* and leads to a more efficient use of the workforce. Bard and Purnomo [8] provide a series of algorithms for the corresponding problem.

An important component of nurse scheduling is the perception of fairness. When preferences are included in the process, soft constraints are used to embody the scheduling rules [9–12]. Some examples are given below:

- Number of one day-off or one day-on patterns occurring in a schedule. The higher the number, the worse the schedule. Such patterns are commonly referred to as on-off-on and off-on-off patterns, and written 1-0-1 and 0-1-0, respectively.
- Number of shift changes on consecutive days. Five-day schedules that have many shift changes, such as Day-Evening-Day-Evening-Day, are undesirable.
- Working three or more shifts in two consecutive weekends.
- No AM or Day shift after a requested day off or holiday. No PM or Night shift before a requested day off or holiday.

The most prevalent hard constraints are as follows:

- Nurses must be assigned to work as many hours as their contract specifies. Overtime is possible but it is generally considered in the daily adjustment stage.
- A nurse can work at most 12 h in 24 h. This means that at most one shift type can be assigned in a day. Also, consecutive shifts must be separated by at least 8 h.
- Rotational nurses must be assigned a minimum proportion (typically 40%) of each of the two shift types specified in their individual contracts. Nonrotational nurses can only be assigned to

work the shift type specified in their contracts.

- The number of consecutive working days (i.e., *workstretch*) should not exceed a specified number of days.
- Every nurse must work two weekend shifts every 2 weeks.

Two versions of the midterm scheduling problem are highlighted below and have their foundations in the pioneering work of [13,14]. Because of the possibility of staffing shortages, uncovered demand is identified as gaps in the schedule to be filled in on the actual day.

Preference Scheduling

In the development of the midterm scheduling model, we make use of the following notation.

Indices and Sets

- i index for nurses; $i \in N$
- j index for schedules; $j \in S_i$ nurse i
- p scheduling period (portion of a day); $p \in P$
- d day of the week; $d \in D$
- N set of nurses to be scheduled
- S_i set of schedules considered for nurse i
- P set of periods in a day
- D set of days in the planning horizon

Parameters and Data

- c_{ij} penalty “cost” of assigning schedule j to nurse i (this value is a function of the number and severity of preference and request violations in the schedule)
- a_{ijdp} 1 if schedule j for nurse i contains period p on day d , 0 otherwise
- LD_{dp} lower bound on demand for nurses on day d in period p
- UD_{dp} upper bound on demand for nurses on day d in period p
- M_p large number representing the cost of an outside nurse or under coverage in period p

Decision Variables

- x_{ij} (binary) 1 if nurse i is assigned to schedule j , 0 otherwise
- y_{dp} number of outside nurses used on day d in period p
- s_{dp} slack variable for the “left” constraint on day d in period p

A basic optimization model designed to minimize the cost of meeting the demand for a particular nurse category in each of the $m \equiv |P| \times |D|$ periods over the planning horizon, is given below in column-oriented format.

Minimize

$$\sum_{i \in N} \sum_{j \in S_i} c_{ij} x_{ij} + \sum_{d \in D} \sum_{p \in P} M_p y_{dp} \quad (1a)$$

subject to

$$-s_{dp} + \sum_{i \in N} \sum_{j \in S_i} a_{ijdp} x_{ij} + y_{dp} = LD_{dp},$$

$$\forall d \in D, p \in P \quad (1b)$$

$$\sum_{j \in S_i} x_{ij} = 1, \quad \forall i \in N \quad (1c)$$

$$x_{ij} = 0 \text{ or } 1, \quad \forall i \in N, j \in S_i \quad (1d)$$

$$0 \leq s_{dp} \leq UD_{dp} - LD_{dp}, y_{dp} \geq 0, \quad \forall d \in D, p \in P. \quad (1e)$$

The objective (1a) is to minimize the weighted sum of the costs associated with the schedules assigned to the regular nurses and the cost of using outside nurses to cover gaps. The choice of the parameter M_p implicitly defines the tradeoff between satisfying the collective preferences of the nursing staff and incurring additional costs by calling on outside nurses to meet unfilled demand. In general, we want $M_p \gg \max\{c_{ij} : i \in N, j \in S_i\}$ to ensure that costs are minimized before preferences, although other options are possible.

The first constraints (1b) ensure that the number of nurses scheduled in each period falls between some minimum and maximum value, and can be written as two separate constraints. In many situations, it might only be necessary to include the lower bound

component. Constraints (1c) along with (1d) guarantee that each nurse in the unit is assigned one and only one schedule. Note that for the sake of compactness, it is possible to combine d and p into a single index that references the time period over the planning horizon. Finally, observe that as long as LD_{dp} and UD_{dp} are integral, an optimal solution will always exist with s_{dp} and y_{dp} integral so it is not necessary to declare them as such.

Both exact and heuristic methods have been developed to find solutions to model Equations (1a–1e) and its variants. Column generation was used successfully by Bard and Purnomo [10] and Jaumard *et al.* [15] while Aickelin and White [9] implemented a tabu search approach. Further discussion of solution methodologies is given in the section titled “Past and Current Research.”

Cyclic Scheduling

In a cyclic scheduling environment, a portion of the staff has rotational shift profiles [16]. A profile specifies the shift types and the number of hours that a nurse is to be assigned over the planning horizon. Problem inputs include the number of nurses for every shift profile, demand per shift, and scheduling rules. When 8- and 12-h shift combinations are permitted, there are many possible shift profiles; that is, Day–Evening, Day–Night, Day–PM, Evening–Night, AM–Evening, AM–Night, and Day–PM.

The second input, demand, may be stated as a range in terms of the number of nurses needed for every shift. Specific values are easily converted to demand per period. The rules typically used in cyclic scheduling are associated with the days-on and days-off patterns listed above as soft constraints.

In preference scheduling, a column-oriented model is best because of the ease with which rule violations can be incorporated. Such models, however, become increasingly difficult to solve as the number of nurses in a unit grows. Because cyclic scheduling considers far fewer rules, a constraint-oriented formulation would seem to be a better choice [14,17]. The main

change with respect to the model expressed in Equations (1a–1e) involves the definition of the decision variables. For cyclic scheduling, the primary decision variables, denoted by x_{idt} , take on the value of 1 or 0 depending on whether or not nurse i is assigned to work on day d and shift t . As a result, the rules that determine the legality of a schedule (hard constraints) or measure preference violation (soft constraints) are required to be modeled explicitly as constraints. For days-on and days-off pattern violations, we have the following:

$$\begin{aligned} \sum_{t \in T_i} x_{idt} + \left(1 - \sum_{t \in T_i} x_{i,d+1,t}\right) \\ + \sum_{t \in T_i} x_{i,d+2,t} + p_{id} \geq 1, i \in N, \\ d = 1, \dots, |D| - 2 \end{aligned} \quad (2a)$$

$$\begin{aligned} \left(1 - \sum_{t \in T_i} x_{idt}\right) + \sum_{t \in T_i} x_{i,d+1,t} \\ + \left(1 - \sum_{t \in T_i} x_{i,d+2,t}\right) + q_{id} \geq 1, \\ i \in N, d = 1, \dots, |D| - 2. \end{aligned} \quad (2b)$$

where p_{id} is a binary variable equal to 1 when nurse i has off–on–off pattern that starts on day d and 0 otherwise, and q_{id} is a binary variable equal to 1 when nurse i has on–off–on pattern that starts on day d and 0 otherwise, and T_i is the set of shifts for which nurse i is hired to work. A complete model can be found in Ref. 17.

SHORT-TERM SCHEDULING

There are several levels of authority and responsibility within a hospital with respect to staff scheduling. At the unit level, the nurse manager, clinical manager, or charge nurse keeps track of the current situation and assesses whether the level of coverage is too low, appropriate, or too high for the number of patients and their acuties. When more nursing resources are available in a unit than are needed for a particular shift, the nurse

manager has several options. Beginning with the least senior staff member, the first is to request a reassignment to another unit. If the nurse in question is not willing to float and is not contractually obligated to do so, her shift is cancelled and her supervisor notified. Generally speaking, nurses would rather be cancelled than floated, although individual preferences may be overridden when the need is critical. If a nurse is cancelled, then one of the following designations is used for the time off: vacation, personal day, holiday, or unpaid leave. This information is reported to the nursing resources director.

Daily Adjustments

When a unit is short of staff, a number of corrective steps can be taken. Each has different cost consequences, although in most hospitals cost is not foremost on the minds of those directly responsible for a unit. Their primary concern is meeting demand, especially in critical situations. With this in mind, the order of action is typically as follows:

1. look for volunteers in the unit to work the next shift (or fraction thereof) as overtime;
2. call casuals and per diem nurses;
3. find floaters from other units that might be overstaffed or draw on the float pool;
4. cycle through the on-call list (used mostly in emergencies);
5. have the nursing resources director call in agency nurses;
6. try to reach unit staff who are not scheduled to work during the next 24 h;
7. invoke mandatory overtime by requesting that a nurse on the current shift stay for the next shift. This is done in reverse order of seniority on a rotating basis so the most junior nurse is not necessary singled out.

Several caveats and qualifications exist for each of these options. For example, when assigning either voluntary or mandatory overtime, there is a distinct possibility that the nurse will call in sick another day during

the pay period. Although he or she will not be paid overtime if his or her total hours do not exceed 80 in the 14-day period, there are no negative consequences. In effect, he or she works overtime in exchange for a day off. A similar situation arises in practice when a nurse is called in on the day off, to work a shift that is separated from the upcoming shift by only 8 h (e.g., Evening $\rightarrow$ Day). In this case, he or she is likely to call in sick for the day shift so little has been gained. This is called a *double back* and should be avoided.

The first two options fall within the purview of the unit manager and are not necessary to model. If contributions are to be made at this level, it is best to focus on the hospital-wide problem of optimally allocating floaters, on-call nurses, and agency nurses. This problem must be solved at least three times a day, and falls within the purview of the nursing resources director.

Modeling Approach

The majority of rules and constraints that govern the daily short-term scheduling problem are outlined above. The complete set may vary among hospitals but generally reflects institutional policies, union agreements, state or federal statutes, and financial considerations. The overall goal is to satisfy coverage requirements at minimum cost while taking into account nurse preferences, morale, the need for the perception of fairness, and the expected response of staff members whose work patterns are affected.

Bard and Purnomo [18] were the first to address this problem analytically by developing an integer linear program for a 24-h planning horizon of three shifts. They proposed that the solutions be implemented within a rolling horizon framework since staffing requirements change unpredictably over the day.

In formulating the model, it was assumed that all cancellation and overtime costs were known, that demand was given or could be estimated accurately for each of the three shifts in the planning horizon, and that the status of all unit nurses, pool nurses, casuals, and agency nurses was available. This is equivalent to saying that the nurse managers, supervisors, and nursing resources

director all have up-to-date information on call outs, shortages, surpluses, floaters, and pool nurses. Using a branch-and-price algorithm, the authors were able to find optimal solutions to instances with up to 200 nurses within a few minutes.

PAST AND CURRENT RESEARCH

The overwhelming majority of research on nurse scheduling has centered on the midterm problem [19,20], which, as mentioned, has been viewed at least three different ways. When rotational or cyclic scheduling is used, several sets of rosters are generated that collectively satisfy the demand requirements. Nurses are then rotated from one set to another in consecutive planning periods. In the simplest case, an exact solution can be found by using a pure set covering model in which cyclic work patterns are generated to minimize the size of the nursing staff needed to cover the demand. Burns [21] studied the case of 10 days on for a 14-day planning horizon with every second weekend off and workstretches limited to six consecutive days. Although easy to implement, cyclic schedules impose a rigidity that many nurses find objectionable.

In contrast, self-scheduling attempts to solve the problem in a more flexible way. Having determined the staffing demand for each period, nurses are asked to sign up for any shifts that they want to work during the planning horizon as long as they remain within their contractual bounds. To avoid an oversupply in any given period, an upper limit is placed on the number permitted to sign up in each block of time. When conflicts occur, further adjustments are made through consensus. Griesmer [3] reported a successful application involving 32 nurses. Although the approach is appealing, it is quite difficult to implement because of the impracticality of holding meetings to resolve conflicts, especially for large units. Moreover, the results may not necessarily be perceived as fair. Those who are savvy enough to game the system will always have an advantage over the procrastinators. Controlling the sign-up order and rotating it over the year is the common way of dealing with this concern.

The third approach—preference scheduling—is actually a combination of the first two. Preferences can be measured in terms of requests to work specific shifts or to be given specific days off, and by other rules such as the number of working hours, shift sequence patterns or even nurse-to-patient ratios. Like self-scheduling, nurses are asked to sign up for shifts prior to the beginning of the planning horizon. At that time, they usually submit a list of requests to the nurse manager who compiles them and decides which to approve immediately in light of the coverage requirements for the upcoming planning horizon. When generating the final rosters, he or she tries to satisfy as many preferences and pending requests as possible.

Exact methods, heuristics, and constraint satisfaction have all been applied to the preference scheduling problem over the last 25 years. Exact methods primarily involve the use of set covering-type models like Equations (1a–1e). In this approach, alternative rosters are generated by any number of methods and one is selected for each nurse. The typical objective is to minimize a weighted combination of costs and preference penalties. What has limited the usefulness of this approach is the size of the integer programs (IPs) that must be solved, and the determination of the penalty coefficients. The longer the planning horizon, the more alternatives that can be generated, and as a result, the harder the IP is to solve.

Warner [22] was the first to develop a methodology for solving the set covering model for the case in which all nurses work 8-h shifts. Jaumard *et al.* [15] extended his model to include both 8- and 12-h shifts, and different skill levels. This extension adopted the concept of periods rather than shifts as the basic time unit. The modified problem was then solved using Dantzig–Wolfe decomposition with the necessary adjustments for integrality restrictions. Columns were generated at each iteration by solving a resource-constrained shortest path subproblem, one for each nurse.

While the set covering methodology focuses on strategies for generating candidate rosters in the form of columns,

heuristics are designed to first obtain an initial solution and then improve upon it with neighborhood search. Miller *et al.* [13] used a simplified version of a rotation heuristic for a 4-week problem. The objective function was a weighted combination of personnel costs and preference penalties. A greedy heuristic with feasible neighborhood swaps was adopted to solve a real instance with up to 12 nurses, a difficult problem at the time.

Metaheuristics have also been used for preference scheduling. Dowsland [23] developed a tabu search approach with strategic oscillation. The algorithm starts with an initial roster obtained with a greedy heuristic that ignores the minimum coverage requirement and treats each nurse separately. In the next phase, three swap moves are used to try to satisfy the coverage constraints: shift swaps for a single nurse, two-nurse chain swaps, and rotation swaps. The quality of each schedule produced by the algorithm is measured by a weighted sum of the penalty coefficients associated with the preference violations.

Valoux and Houxos [14] used a combination of integer programming and heuristics to find solutions to a limited version of the preference scheduling problem. Instead of solving the classic set covering model, they solved a relaxation that minimized the total number of shifts needed to satisfy demand subject to several preference constraints. Arthur and Ravindran [24] were the first to apply goal programming to the preference scheduling problem. They considered the following four goals: contractual requirements, preferences, requests, and staffing requirements. In the first phase of their two-phase approach, a small IP is solved to decide the days on which each nurse is to work. Shift assignments are made in the second phase. In a similar vein, Berrada *et al.* [11] proposed a constraint satisfaction model that viewed demand coverage and the number of working days as the hard constraints and compliance to shift patterns, daily requirements for supervisory personnel, and the grouping of days off and weekends off as the soft constraints. Several solution approaches were discussed

including sequential optimization and tabu search.

For the same problem, Ferland *et al.* [12] developed a tabu search algorithm that included a diversification strategy and an adaptive memory structure. Similarly, Burke *et al.* [25] used tabu search to schedule nurses at several Belgium hospitals using a code implemented in the Plane software system. Subsequent refinements allowed for multiple objectives and more equitable weekend assignments.

Working independently, Cheng *et al.* [26] proposed about 20 heuristic rules to build rosters and recursively fill in the uncovered shifts. Following a different course, Kawanaka *et al.* [27] used a genetic algorithm to find solutions. Their method was a bit restrictive, though, because it was designed around six hard constraints specific to the hospital environment in which they were working. Other heuristics include the use of language theory, artificial intelligent techniques, and expert systems. The general idea is to represent problem constraints as clauses or strings that can be combined to form feasible schedules with the help of an inference engine like Prolog; for example, see Ref. 28.

Table 1 summarizes the work on exact and heuristic methods for midterm preference scheduling. The first column lists the researchers, the second column identifies the type of model used, and the last column highlights the solution methodology.

Somewhat surprisingly, only a handful of the approaches discussed in this section have found their way into commercial systems. The first was the work of Warner [22] who developed ANSOS in the late 1970s and who went on to develop ClairVia in the 1990s. More recently, the column generation approach of Bard and Purnomo [10] was realized in a system called *CareWare*. Over the years, many products have appeared, but few have taken advantage of the sophisticated techniques that have emerged from academic research, preferring instead to rely on simple rules and heuristics. This issue is further discussed by Kellogg and Walczak [29] who trace the transfer of nurse scheduling technology from academia to the marketplace.

Table 1. Summary of Midterm Scheduling Research

Researchers	Model Type	Solution Methodology
Aickelin and White [9]	Set covering	Genetic algorithm
Arthur and Ravindran [1]	Constraint-based	Goal programming
Bard and Purnomo [18]	Set covering	Column generation
Berrada <i>et al.</i> [11]	Constraint-based	Sequential optimization, tabu search
Cheng <i>et al.</i> [26]	Rule-based	Constraint programming
Dowland [23]	Set covering	Tabu search
Ferland <i>et al.</i> [16]	Constraint-based	Goal programming
Griesmer [3]	Implicit	Self-scheduling (sign-up)
Jaumard <i>et al.</i> [15]	Constraint-based	Dantzig–Wolfe decomposition
Kawanaka <i>et al.</i> [27]	Constraint-based	Genetic algorithm
Millar and Kiragu [2]	Constraint-based, network	Integer programming
Miller <i>et al.</i> [13]	Constraint-based	Greedy heuristic
Okada [28]	Logical clauses	Inference engine
Purnomo and Bard [17]	Constraint-based	Dantzig–Wolfe decomposition
Valouxis and Housos [14]	Constraint-based	Integer programming, heuristics
Warner [22]	Constraint-based	Integer programming

CONCLUSIONS

The continuing interest in nurse scheduling is driven by a combination of (i) the economic importance of the application, (ii) the enticing duality between the conciseness of the problem statement and the difficulty in deriving high quality solutions, (iii) the opportunity to advance the use of quantitative methods, and (iv) the rewards that come from industrial–academic collaboration. Although variations of the problem have been studied exhaustively for many years, a broad survey of the field suggests that many opportunities still exist for improving both, the prevailing models and our algorithmic capabilities, especially when it comes to incorporating uncertainty in the analysis. A more robust treatment of demand at all four planning levels and a better understanding of long-term staffing requirements, are two areas where good research is needed.

To narrow uncertainty, it will be necessary to develop better forecasting systems and better patient tracking methods. Each of these functions requires the collection and analysis of massive amounts of data drawn from the various and often incompatible databases that have come to characterize the state of information technology at most health-care facilities. To overcome this hurdle, an important step will be the establishment of

standard patient records and uniform interface protocols.

REFERENCES

1. Aiken LH, Clarke SP, Sloane DM, *et al.* Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *J Am Med Assoc* 2002;288(16):1987–1993.
2. Millar HH, Kiragu M. Cyclic and non-cyclic scheduling of 12-hour shift nurses by network programming. *Eur J Oper Res* 1998;104:582–592.
3. Griesmer H. Self-scheduling turned us into a winning team. *Manage Decis* 1993;56(12):21–23.
4. Punnaikitikashem P, Rosenberger JM, Buckley Behan D. Stochastic programming for nurse assignment. *Comput Optim Appl* 2008;40(3):321–349.
5. Lau HC. On the complexity of manpower shift scheduling. *Comput Oper Res* 1996;23(1):93–102.
6. Pickard B, Warner DM. Demand management: a methodology for outcomes-driven staffing and patient flow management. *Nurs Leadersh* 2007;5(2):30–34.
7. Fry MJ, Magazine MJ, Rao US. Firefighter staffing including temporary absences and wastage. *Oper Res* 2006;54(2):353–365.
8. Bard JF, Purnomo H. A column generation-based approach to solve the preference

- scheduling problem for nurses with downgrading. *Socio Econ Plann Sci* 2005;39(3): 193–213.
9. Aickelin U, White P. Building better nurse scheduling algorithms. *Ann Oper Res* 2004;128:159–177.
 10. Bard JF, Purnomo H. Preference scheduling for nurses using column generation. *Eur J Oper Res* 2005;164(2):510–534.
 11. Berrada I, Ferland JA, Michelon P. A multi-objective approach to nurse scheduling with both hard and soft constraints. *Socio Econ Plann Sci* 1996;30(3):183–193.
 12. Ferland JA, Berrada I, Nabli I, *et al.* Generalized assignment type goal programming problem: application to nurse scheduling. *J Heurist* 2001;7:391–413.
 13. Miller HE, Pierskalla WP, Rath GJ. Nurse scheduling using mathematical programming. *Oper Res* 1976;24(5):857–870.
 14. Valouxis C, Housos E. Hybrid optimization techniques for the workshift and rest assignment of nursing personnel. *Artif Intell Med* 2000;20:155–175.
 15. Jaumard B, Semet F, Vovor T. A generalized linear programming model for nurse scheduling. *Eur J Oper Res* 1998;107:1–18.
 16. Emmons H. Work-force scheduling with cyclic requirements and constraints on days off, weekends off, and work stretch. *IIE Trans* 1985;17(1):8–15.
 17. Purnomo H, Bard JF. Cyclic preference scheduling for nurses using branch and price. *Nav Res Log* 2007;54(2):200–220.
 18. Bard JF, Purnomo H. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Trans Oper Eng* 2005;37(7):589–608.
 19. Burke EK, Causmaecker PD, Berghe GV, *et al.* The state of the art of nurse rostering. *J Sched* 2004;7(6):441–499.
 20. Cheang B, Li H, Lim A, *et al.* Nurse rostering problems—a bibliographic survey. *Eur J Oper Res* 2003;151:447–460.
 21. Burns RN. Manpower scheduling with variable demands and alternate weekends off. *INFOR* 1978;16(2):101–111.
 22. Warner DM. Scheduling nursing personnel according to nursing preference: a mathematical programming approach. *Oper Res* 1976;24(5):842–856.
 23. Dowsland KA. Nurse scheduling with tabu search and strategic oscillation. *Eur J Oper Res* 1998;106(2–3):393–407.
 24. Arthur JL, Ravindran A. A multiple objective nurse scheduling model. *AIIE Trans* 1981;13(1):55–60.
 25. Burke EK, Causmaecker PD, Berghe GV. Novel metaheuristic approach to nurse rostering problems in Belgian hospitals. In: Leung J, editor. *Handbook of scheduling*. Boca Raton (FL): CRC Press; 2004, Chapter 44. pp. 1–18.
 26. Cheng BMW, Lee JML, Wu JCK. A nurse rostering system using constraint programming and redundant modeling. *IEEE Trans Inform Tech Biomed* 1997;1(1):44–54.
 27. Kawanaka H, Yamamoto K, Yoshikawa T, *et al.* Genetic algorithm with constraints for the nurse scheduling problem. Volume 2, *Proceedings of congress on evolutionary computation*. Seoul: IEEE Press; 2001. pp. 1123–1130.
 28. Okada M. An approach to the generalized nurse scheduling problem—generation of a declarative program to represent institution-specific knowledge. *Comput Biomed Res* 1992;28:417–434.
 29. Kellogg DL, Walczak W. Nurse scheduling: from academia to implementation or not? *Interfaces* 2007;37(4):355–369.

OLYMPICS

MARCOS ESTELLITA LINS
Department of Production Engineering,
Federal University of Rio de Janeiro,
Brazil

ANGELA MOREIRA DA SILVA
Faculty of Economic Sciences, State
University of Rio de Janeiro, Brazil

DISTINCTIVE FEATURES IN OLYMPIC GAMES

Applications of operations research (OR) and management science (MS) in sports comprise quantitative support tools for measuring and improving players' and teams' performances, analyzing league strategy, and assessing social and economic values of professional sports. One typical application is to help the sports teams to choose their best available players in a sports league [1]. A collection of articles that best represents the sports research using Statistics and Operational Research can be found in Refs 2 and 3, respectively. What distinguishes Olympic Games and characterizes OR/MS specific applicable techniques is to be seen in the following text.

The Olympic Games were named after the place where they were held, in a sanctuary site for the Greek deities in Olympia near the towns of Elis and Pisa, in Southern Greece. Since ancient Greece, though designed for individual contests the cities from where the winners originated would grant them numerous prerogatives, showing that the city felt that it had won as well. The modern Games, initiated in 1896 by Baron Coubertin, intended to preserve the initial spirit of individual competitions. Until 1908, athletes entered as individuals rather than as members of a national team. However, the scenario preceding Second World War contributed to reverse this view, when the Third Reich had tried to show the supremacy of the Arian race in the Games of 1936, although the results were quite different from those what Hitler had expected.

During the Cold War, the national character of the contest became even more noticeable and developed into a true battle between East and West. Not the individual performances but the national performances were in the focal point of attention.

Nowadays, the results at country levels are disclosed through the Organizing Committee of the Olympic Games (OCOG) reports and in the popular media, playing an important role on the self esteem of an entire nation as a symbolism for achievement.

Therefore, the first distinguishing requirement of applied quantitative methods in Olympic Games is to help in explaining the results and setting targets for contestant countries consistent with available resources.

The second important issue brought about by Olympic Games concerns the capacity of shaping and projecting images of the host country, both domestically and globally. The event fosters an enhanced sense of national pride and makes the venue and the host country highly attractive for political and economic elites [4]. OR applications also aim to help the host country providing successful integrated planning and execution of activities. This is in consonance with the broader perspective of OR formulated by Ackoff [5] and Rosenhead [6], among others, encompassing both structuring and solution of real-world problems. The schema in Fig. 1 is a tool borrowed from problem structuring methods [7] for representing elicited knowledge and synthesizes the rationale behind this twofold competitive-integrative approach to Olympic issues.

UNDERSTANDING OLYMPIC PERFORMANCE

Explaining and understanding awarding Olympic medals as a function of country resources is the formal counterpart for the country competition, which is an Olympic driving force. So this will be the initiation for applications of quantitative methods to Olympics.

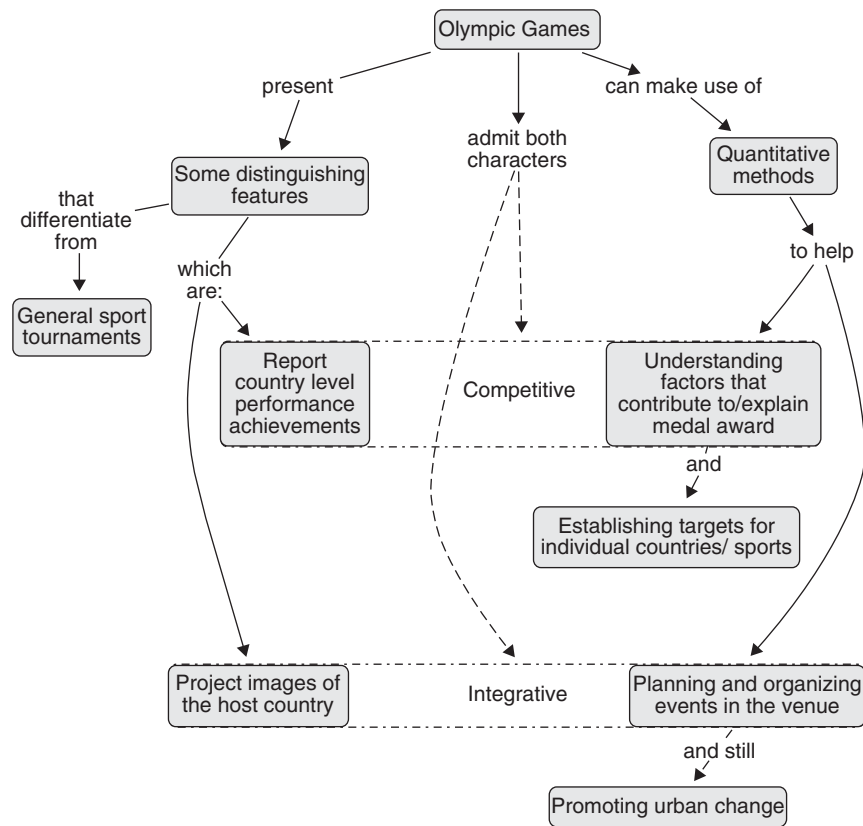


Figure 1. Rationale behind OR/MS applied to Olympics.

For the sake of motivating the theme, see in Ref. 8 the interesting debate about which explanatory variables play a causative role in awarding medals, illustrated by the proposal of dividing the total number of medals by the population of each country. It was first suggested in the “Globe and Mail,” analyzing the Atlanta games of 1996 and later in the “Los Angeles Times,” regarding the Beijing games of 2008. In the latter, the authors claimed to reflect real national accomplishment, since countries present huge differences in population scale. A forceful discussion follows, where opponents suggest several other relevant explaining variables, such as GDP funding (private or public) and delegation size. A large population offers a larger pool of potential athletes from which to withdraw successful contenders. High GDP allows to support higher fixed costs of infrastructure

and facilities for training and still to face the costs necessary to send athletes to the Games. Actually, understanding Olympic success requires additional explanatory variables, as demonstrated by the 1996 Olympic results, where Cuba won 25 medals, while India won 1 medal.

The second issue in Olympic performance regards the medals ranking, since results are traditionally published as a table in which countries are ranked in accordance with the number of gold medals their athletes have won. Silver medals are counted to break a tie and so are bronze medals in the sequence. This type of ranking is characterized as the lexicographic multicriteria method, where each criterion enjoys the exclusivity privilege by its turn. In this particular case, it implies a disadvantage to overvalue the gold medal, since countries that win the highest number

of silver and bronze medals but none of gold are ranked below countries that win a single gold medal.

Condon *et al.* [9] computed total scores (defined by the weighted sum of the top eight places in each event) for 195 countries that were represented at the 1996 Summer Games and gathered information on 17 independent variables. Acknowledging that usual explanatory variables would be connected to population density, economic system, and geographic distance, they included transportation indices in neural network and regression models, finding that “total length of railroad track” and “number of airports with usable runways” contributed to the explanation of variance in the Olympic scores. However, they conclude that “Although a country could use our models and new data to predict its total score in an upcoming Olympics, there is probably not much a country can do (in the short term) to make significant changes in the hope of increasing its total score.” In fact this applies to the left side of Fig. 2 where we display causal variables at the country level, but not to the right side, where we show the personal level explanatory variables.

Morton [10] performed statistical analysis for allometric approaches to medal scores, which included medals per capita, medals by GDP, and medals adjusted by several adjustment measures.

Hoffmann *et al.* [11] reported the results of regressions, testing hypotheses concerning the medal winning success of Olympic teams. They find that, while traditional economic and political factors are important, many inherent national characteristics such as geographical, demographic, and cultural factors have a significant and pronounced impact. The authors conclude that scope for public sports policy exists and discusses the degree to which the inherent characteristics limit it.

Bernard and Busse [12] examined the determinants of Olympic success at the country level and referred to fairness in achievements as a relevant issue. They asked, “Does the United States win its fair share of Olympic medals? Why does China win only 6% of the medals even though

it has one-fifth of the world’s population?” They considered the role of population and economic resources in determining medal totals from 1960 to 1996, suggesting that both a large population and a high per capita GDP are needed to generate high medal totals. They also provided out-of-sample predictions for the 2000 Olympics in Sydney.

According to Hawksworth [13], Head of Macroeconomics PricewaterhouseCoopers, “the number of medals won increases with the population and economic wealth of the country, but less than proportionately: David can sometimes beat Goliath in the Olympic arena, although superpowers such as the United States, China, and Russia continue to dominate at the top of the medal table.” This model includes data on medal performance from the Olympic Games since 1988, for the following statistically significant factors: population, average income levels, whether the country was previously a part of the former Soviet bloc (including Cuba in this case), whether the country is the host nation, and its medal share in the previous Olympic Games.

Courneya and Carron [14] identified three factors that can account for home advantage of the host nation, namely, crowd, learning, or familiarity and travel.

- Crowd may either influence players to play more recklessly or affect the match officials’ decisions to favor the home side. Crowd noise would appear to influence the observers to favor the home side and penalize the opponent side. In gymnastics, Ansorge and Scheer [15] found that judges at the 1984 Olympics not only scored their own gymnasts higher, but also that they scored immediate competitors lower. Balmer *et al.* [16] used simple statistical sign tests to identify bias demonstrating a clear observed imbalance. Corrective methods have been adopted to reduce nationalistic bias.
- Familiarity comprises the competition site and the cultural/environmental adaptation. Home advantage results relevant for events where there is wide variation in conditions (e.g., alpine

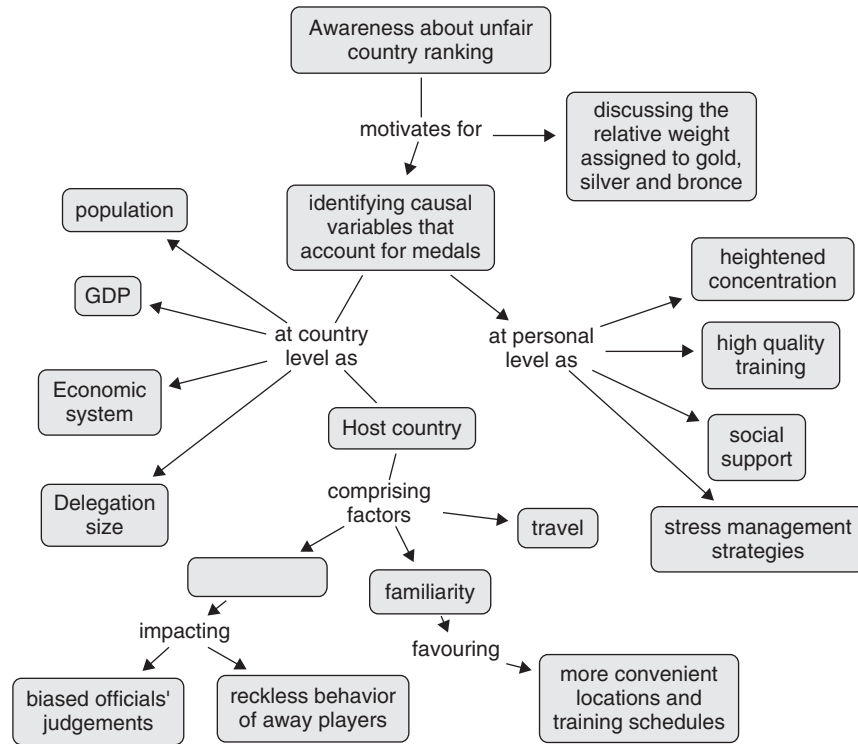


Figure 2. Explanatory variables for Olympic results at country and personal levels.

skiing, luge, and bobsled or where grass is used instead of artificial surfaces) as opposed to the far less variable rink size or stadia. Local athletes are more accustomed to facilities and they may be able to arrange more convenient locations and training schedules.

- Pace and Carron [17] proposed that “only a small portion of the variance in the home advantage/visitor disadvantage can be explained by travel related factors.” Jehue *et al.* [18] found direction of travel and time of day to be important, and suggested a possible “jet-lag” effect.

Balmer *et al.* [16] test for the prevalence of subjective officiating, distance traveled, and familiarity with local conditions, using linear, quadratic, and exponential terms in regression analysis. They conclude that subjective assessment by officials can explain a large proportion of the variation in observed home

advantage in Winter Olympic events. Familiarity with local conditions was shown to have an intermediate influence and no appreciable trends were noted for the travel factor, given that time would be available to overcome the latter adverse effects.

Kuper and Sterken [19] analyze Olympic participation and success in all the modern editions of the Summer Games, using a large data set including Olympic statistics, per capita income, and demographic information. They explain participation shares and medal success at the country level and show that, conditional on participation, it is possible to model success in the medal standings. They achieve the following conclusion:

- First, the economic condition of a country is still important for both participation and, to a lesser extent, for success. The impact of per capita income has diminished though. The same holds for

the impact of population size and distance to the Games.

- Second, the home effect is important, especially for participation. Here we have a regulated effect; the home team is allowed to send more athletes. This effect is still strong, but used to be more important at the older editions of the games. Probably, transport costs caused a bias to home representation. The home advantage in success is less clear. Before World War II the home advantage was strong via participation and success. At the recent games, the home advantage shifted from gold to bronze.
- Third, the legal tradition of a country, as a proxy of the sports culture is relevant for modeling participation and success. Especially, socialist countries send more athletes and earn more medals. French legal system countries are less impressive in their performance. Emancipation is found to be important. Media attention is also important in explaining participation. Finally, there seems to be an untested auto-regressive factor, since winning gold medals has important consequences for future success in the Olympics.

This is indeed, a fundamental task for quantitative models applied to Olympics: explaining success and failure in Olympic Games.

Lozano *et al.* [20] apply data envelopment analysis (DEA) to measure the performance of the nations participating in the last five Summer Olympic Games. The proposed approach considers two inputs (GNP and population) and three outputs (number of gold, silver, and bronze medals won). To increase the consistency of the results, weight restrictions are included, guaranteeing a higher valuation for gold medals than for silver medals and higher for the latter than for bronze medals. Variable returns to scale are assumed. In addition, plotting the performance of a specific country for the different games can help identify trends as well as objective successes and disappointments.

Lins *et al.* [21] proposed a zero sum gains data envelopment analysis (ZSG-DEA) model to assess Olympic performance of countries constrained to a bounded amount of medals. The weights assigned to each medal— w_G (gold), w_S (silver), and w_B (bronze)—can vary, allowing for different country preferences, though restricted by $w_G > w_S > w_B$. The ZSG-DEA assumes that the sum of outputs is constant. This is similar to a zero sum game in which whatever is won by a player is lost by one or more of the others. The overall changes are taken into account when measuring the efficiency of a given country.

Churilov and Flitman [22] focused on the issue of whether it is possible to design an objective impartial system of analysis of the Olympic results. The authors' approach aims at accounting for the multiple criteria and multiple attribute nature of the situation under consideration. Differentiation between results and performance reflects the concept of fairness, as a concern with "significant bias towards countries with large population, or rich countries—those whose GDPs are significantly higher than the average." The authors are quite persuasive putting their argument straightforwardly: "in order to financially afford getting 10 gold medals at Sydney Olympics, the United States of America would have had to spend 0.04% of its annual health budget, or 0.09% of its annual education budget, or 0.15% of its annual defense budget. For Australia, these numbers are 1.2%, 1.9%, and 4.9%, respectively, while for Laos they are 570%, 237%, and 146%, respectively (expressed in terms of 1998 USD)." The unsupervised data mining technique of self-organizing maps is used to group the participating countries into homogenous clusters. A DEA-based model is then used for producing a new ranking of participating teams that could be acceptable as "fair" by the majority of participants. An important innovation was the inclusion of disability adjusted life expectancy (DALE) index and index of equality of child survival (IECS).

Li *et al.* [23] formulate a context-dependent DEA model using different weight restrictions to analyze the achievements of nations during six Summer Olympic Games.

They report the efficiency scores showing that Australia, Norway, North Korea, and Azerbaijan, present very positive trends in terms of performance improvement.

Wu *et al.* [24] applied DEA to the last six Summer Olympic Games and obtained a unique ordering of the participant countries based on the average cross efficiency, also cluster analysis technique was used to select the more appropriate targets for poorly performing countries to use as benchmarks.

Wu *et al.* [25] proposed to apply a DEA game cross-efficiency model where each DMU is viewed as a competitor under a noncooperative game approach. The method is applied to the last six Summer Olympic Games.

The explanatory approach to Olympic results could be further extended to help guiding country strategies, when individual strategic features are included in assessment. Recent practices of the National Olympic Committees have moved away from funding mass participation and toward rewarding medal production [26]. Russia and China have long been funneling money to sports that offer multiple medal opportunities, such as diving, gymnastics, track and field, and swimming [27]. Until 2000, in the United States, the USOC was still dividing

up its funds equally among 45 different sports federations for use as they saw fit. In 2004, the USOC shifted to what it called a *venture capital* model requiring member National Governing Bodies (NGBs) to present specific plans detailing how they intend to use financial resources from the USOC to increase their chances of winning Olympic medals. Just as business executives, Olympic coaches in the United States now must present specific plans detailing how they intend to use the money to increase their chances of bringing home the gold. The rationale behind gold productivity from given resources can benefit from quantitative models used to explain Olympic performance. However, it requires the measurement and inclusion of social and psychological variables such as high quality training, heightened concentration, stress management strategies, social support, and others [28].

PLANNING AND ORGANIZING OLYMPIC EVENTS

OR/MS applications to Olympics planning aim at supporting the operation of venues, scheduling of events, information management, and selection of host city (Fig. 3).

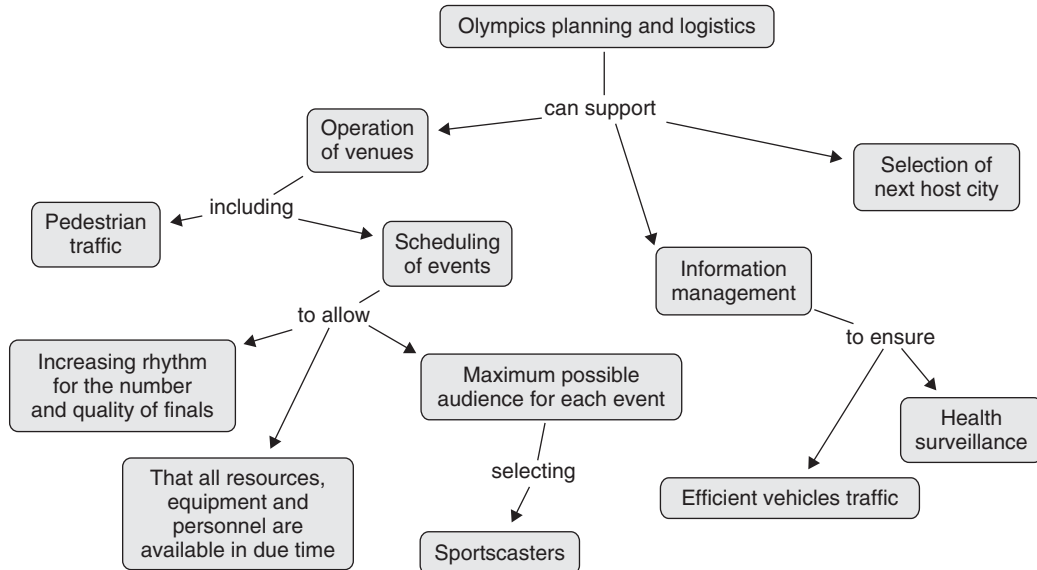


Figure 3. Problems in Olympics planning suited to OR/MS application.

OR/MS can help planning, designing, and implementing systems to support venue operations at the Olympic Games, helping the organizing committees to create designs that result in reliable, high-quality venue operations at reasonable cost. The Athens 2004 Olympic Games Organizing Committee (ATHOC) used innovative techniques from MS, systems engineering, and information technology to change the planning, design, and operation of venues [29]. The Process Logistics Advanced Technical Optimization (PLATO) project was developed as a systematic process for planning and designing venue operations by using knowledge modeling and resource management techniques and tools. They developed a rich library of models that is directly transferable to future Olympic organizing committees and other sports-oriented events. The direct financial benefit to ATHOC was the reduction of the costs of managing venue operations by over \$69.7 million. The success of the games raised Greece's international profile in terms of capabilities in managing large and complex projects, in the medium- to long term, with financial, political, and social benefits.

Andreu and Colominas [30] developed a Decision Support System (DSS) to support the scheduling of 2000 events in a 15-day period during the 1992 Olympic Games in Barcelona. The problem is one of multiple criteria; namely, ensure the maximum possible audience for each Olympic event (considering different times and interests across countries); while avoiding inconveniences to athletes, provide an increasing rhythm for the number and quality of finals; ensure that all resources, equipment and personnel are available in due time; and avoid traffic jams. The DSS implementation should satisfy some requirements such as allowing for checking fulfillment of feasibility constraints; adopting criteria categories that make good sense to the end user, instead of easy-to-manipulate from an algorithmic point of view; facilitating the definition, manipulation, examination, and evaluation of alternative schedules according to different criteria as naturally and convenient as possible. The system, called *SUCCESS92* was approved as a useful tool by the organizing committee.

Liao and Chang [31] applied a multiple criteria decision-making method, to help Taiwanese TV companies select optimal televised sportscasters for the 29th Olympic Games held in Beijing, China, from 8 to 24 August 2008. Expert televised sportscasters fully describe, analyze, and comment on the games. The televised sportscaster is a vital contributor to the audiences' appreciation of televised sports. Selection criteria were collected through interviews with 44 executives and televised sportscasters. Twelve criteria that were retained are the following:

Description: Describe the game as it is taking place

Explanation: The ability to analyze. For example: Why implement this tactic?

Expertise: Professional knowledge of sports

Comment: Comment the game, such as the performance of tactics

Order: The ability to finish orders

Cognition: The ability to resolve problem by oneself

Adaptation: Adapt to the external environment

Language: Familiar with foreign languages

Teamwork: Cooperate with others to finish work

Experience: Past experience about sports

Attitude: Conscientious toward work

Response: React appropriately to emergency

Unlike previous contributors who ignored the interdependence among criteria, a more feasible and accurate approach was applied in this article to handle such problems: the analytical Network Process (ANP), which captures the outcome of dependency among the criteria. The ANP extends the Analytical Hierarchic Process (AHP) to deal with dependence. Priorities are established in the same way they are in the AHP, using pair wise comparisons.

Vehicles and pedestrians traffic is a main concern in Olympic Games, given the increased burden imposed on transit

network. Theory of graphs and simulation are some of the OR methods usually applied to support decision making in transportation problems.

During the 1996 Atlanta Olympic Games, the authority deployed an Olympic shuttle bus and a 126 km of high-occupancy vehicle (HOV) lane, expanding rail transit network, to reduce regional commute trips and to alleviate traffic congestion. They also developed the most ambitious intelligent transportation systems (ITS) in the United States [32]. Some impressive results were obtained, concerning reduction of peak flows and increased public transportation volume.

To face the huge Olympic transport challenge, Sydney 2000 established an Olympic Road and Transport Authority (ORTA), a fully multimodal public agency in charge of planning and delivering of all Olympic transport services. According to ORTA's plan, all Olympic stadiums and venues were involved in the public transit system and a new high capacity rail loop with 5.3-km length was built to handle about 80% of the Olympic generated traffic. All measures combined resulted in a background traffic reduction of about 20% and rail traffic growth from 14 to 29.5 million passengers per day.

Transport departments in Beijing worked out "Traffic Assurance Plan during the 2008 Beijing Olympic and Paralympic Games." The general objective was to cut the total number of vehicles on the roads, ensure traffic safety and mobility during the Games, provide public transportation services for citizens to minimize the impact on people's work and daily lives, reduce vehicle emissions, and improve air quality. Liu *et al.* [32] analyzed traffic operation status of different periods, based on surveys and concluded that the successful implementation of traffic control measures during the 29th Beijing Olympics not only promoted the bus-oriented traffic flow composition but also effectively relieved road traffic load.

According to Zhu *et al.* [33], the pedestrian traffic issues aroused by large-scale of people's activities in special events are commonly considered more comprehensive and even cause more serious results than vehicle traffic. As the great demand of spectators

which may induce intermittent walking way, crowded space, and hidden safety problems, it is a crucial process of framing the pedestrian designing and management plan. The pedestrian simulation can be used to assist decision making in the planning phase. The advantages of program simulation are to test the pedestrian facilities and management planning beforehand in order to decrease the probability of potential risks. Usual simulation traffic parameters in train stations, metro stations, and airports are speed and density of pedestrians, which assume particular characters in Olympic events, given the huge centralization in time at a particular site. Zhu *et al.* [33] applied a simulation tool to assess the capacity of facilities, effect, and security of operational planning on different phases of designing and evacuating planning of the National Stadium in Beijing.

Gundlapalli *et al.* [34] report the development of a computer rule-based system that relies on the patient's electronic medical record, transmitted by clinical computing systems and encompassed a variety of types of patient-level data. These includes: laboratory test ordering and results, radiology ordering and reports, emergency room and outpatient clinic visits, and hospital admissions. Several types of rules were developed for this system. Single data source rules were designed to identify occurrences of single important events (e.g., test order for anthrax and test results for influenza). Other rules were focused on specific diagnoses, such as pneumonia by chest radiograph. Groups of rules based on multiple data sources with patient-level data were used as a proxy to identify infectious syndromes. For each rule, statistical process controls were applied to generate alerts when levels exceeded two standard deviations above the mean. The system was deployed at a large hospital in Salt Lake City, during the 2002 Winter Olympic Games, and it was accessed three times a day to perform surveillance. Daily reports were provided to local public health agencies after preliminary investigation of the alerts. Since this system uses objective, quantifiable events, and access to patient data, infection control practitioners could

play a front-line role in biosurveillance and facilitate bidirectional communication with public health agencies [34].

Following the Beijing Olympics, the most important task for the International Olympic Committee (IOC) will be choosing which city hosts the 2016 Summer Olympics. (2012 has already been awarded to London.) The scope of candidacy was narrowed to Madrid, Rio de Janeiro, Tokyo, and Chicago. The winner would be announced in October 2009. How does the IOC make their decision?

While their overall process is popularly described as a two-step process, it is clearly three as mentioned below [35]:

1. Potential cities (not countries) must submit a proposal demonstrating that they meet minimum hosting requirements. These focus on infrastructure, stability, and resources.
2. The 15-member executive board then uses an extensive questionnaire to reduce the list to about 5 to 10 cities.
3. Finally, the full board (currently 115 members) votes. A simple majority (>50%) is required to win. The city with the lowest vote count during each vote is eliminated from the next voting round until a majority is achieved.

According to Ref. 35, “the final phase contains two important flaws. First, the criteria for city selection are relatively loose. While the Olympic Charter provides guidelines and expectations, they lack true criteria, as well as any means to measure the criteria. Board members are left to determine their own criteria and priorities. Second, the voting process focuses on eliminating the least qualified cities rather than valuing the most prepared ones. Through rounds of voting, this process can become more an act of political will than a referendum on the best available option.”

The authors offer an on-line exercise to use a multicriteria model based on six criteria (objectives) and the four finalist cities, to help determine which city should host the Summer Olympics in 2016.

REFERENCES

1. Fry MJ, Lundberg AW, Ohlmann JW. A player selection heuristic for a sports league. *J Quant Anal Sports* 2007;3(2), Article 5.
2. Bennett J, Cochran JJ. In: Albert J, editor. *Anthology of statistics in sports ASA-SIAM series on statistics and applied probability*. Philadelphia, ASA, Alexandria (VA): SIAM; 2005.
3. Cochran JJ. *Operations Research and Sports: A Brief Overview*, StatOR, Vol 8, Num 2.
4. Manzenreiter W, Horne J. *Hosting Major International Events: Comparing Europe and Asia*; 2005. ASEF/Alliance workshop, Available at http://www.cba.uc.edu/or_sports. Assessed 2009 Nov 10.
5. Ackoff RL. *Creating the corporative future*. New York: Wiley; 1981.
6. Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world Revisited: Problem Structuring Methods for Complexity, Uncertainty, and conflict*. Chichester: Wiley, 2001.
7. Checkland P, Scholes J. *Soft systems methodology in action*. Chichester: John Wiley & Sons, Ltd.; 1990.
8. The true Olympic gauge: Medals Per Capita Table in Los Angeles Times – Sports. Available at http://latimesblogs.latimes.com/olympics_blog/2008/08/everybody-says.html. Accessed in 2010 June 23.
9. Condon EM, Golden BL, Wasil EA. Predicting the success of nations at the Summer Olympics using neural networks. *Comput Oper Res* 1999;26:1243–1265.
10. Morton RH. Who won the Sydney 2000 Olympics?: an allometric approach. *Statistician* 2002;51:147–155.
11. Hoffmann R, Ging LC, Ramasamy B. *Appl Econ Lett* 2002;9(8):545–548.
12. Bernard AB, Busse MR. Who wins the olympic games: economic resources and medal totals. *Rev Econ Stat* 2004;86(1):413–417.
13. Hawksworth J. *Economic Briefing Paper: Modeling Olympic Performance*. Accessed 2009 June 01. pwc.com/economics.
14. Courneya KS, Carron AV. The home advantage in sport competitions: a literature review. *J Sport Exerc Psychol* 1992;14:13–27.
15. Ansoorge CJ, Scheer JK. International bias detected in judging gymnastic competition at the 1984 Olympic games. *Res Q Exerc Sport* 1988;59:103–107.

16. Balmer NJ, Nevill AM, Williams AM. Home advantage in the Winter Olympics (1908–1998). *J Sports Sci* 2001;19:129–139.
17. Pace A, Carron AV. Travel and the home advantage. *Can J Sports Sci* 1992;17:60–64.
18. Jehue R, Street D, Huizenga R. Effect of time-zone and game time changes on team performance: National Football League. *Med Sci Sports Exerc* 1993;25:127–131.
19. Kuper G, Sterken E. Olympic participation and performance since 1896. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=274295. Assessed 2009 May 01.
20. Lozano S, Villa G, Guerrero F, *et al.* Measuring the performance of nations at the summer olympics using data envelopment analysis. *J Oper Res Soc* 2003;53(5):501–511.
21. Lins MPE, Gomes EG, Mello JCCBS, *et al.* Olympic ranking based on a zero sum gains DEA model. *Eur J Oper Res* 2003;148:312–322.
22. Churilov L, Flitman A. Towards fair ranking of Olympics achievements: the case of Sydney 2000. *Comput Oper Res* 2006;33:2057–2082.
23. Li Y, Liang L, Chen Y, *et al.* Models for measuring and benchmarking Olympics achievements. *Omega* 2008;36:933–940.
24. Wu J, Liang L, Chen Y. DEA game cross-efficiency approach to Olympic rankings. *Omega* 2009;37:909–918.
25. Wu J, Liang L, Yang F. Achievement and benchmarking of countries at the Summer Olympics using cross efficiency evaluation method. *Eur J Oper Res* 2009;197:722–730.
26. Dittmore S, Mahony D, Andrew DPS, *et al.* Examining fairness perceptions of financial resource allocations in U.S. Olympic Sport. *J Sport Manage* 2009;23:429–456.
27. Piore A. Sink or swim. *Newsweek*; 2004. Available at <http://www.msnbc.msn.com/id/5636114/site/newsweek>. Retrieved 2006 Feb 21.
28. Gold D, Guinan D, Greenleaf C, *et al.* Factors affecting olympic performance: perception of athletes and coaches from more and less successful teams. *Sports Psychol* 1999;13:371–294.
29. Beis DA, Loucopoulos P, Pyrgiotis Y, *et al.* PLATO helps athens win gold: olympic games knowledge modeling for organizational change and resource management. *Interfaces* 2006;36(1):26–42. DOI: 10.1287/inte.1060.0189.
30. Andrew R, Corominas A. SUCCCES92: a DSS for scheduling the Olympic Games. *Interfaces* 1989;19(5):1–12.
31. Liao SK, Chang KL. Select televised sportscasters for Olympic Games by analytic network process. *Manage Decis* 2009;47(1):14–23. DOI 10.1108/00251740910929678.
32. Liu M, Mao B, Huang Y, *et al.* Comparison of Pre- & Post-Olympic Traffic: a case study of several roads in Beijing. *J Trans Syst Eng & IT* 2008;8(6):67–72.
33. Zhu N, Wang J, Shi J. Application of pedestrian simulation in Olympic games. *J transport syst eng inform tech* 2008;8(6):85–90.
34. Gundlapalli AV, Olson J, Smith SP, *et al.* Hospital electronic medical record–based public health surveillance system deployed during the 2002 Winter Olympic Games. *Am J Infect Control* 2007;35(3):163–171.
35. Which city should host the 2016 Summer Olympics? by Expert Choice Inc. Available at <http://www.expertchoice.com/news-events/news-detail/expert-choice-going-for-gold-with-selection-process-of-the-2016-olympics/>. Accessed in 2010 June 23.

OMEGA RHO INTERNATIONAL HONOR SOCIETY FOR OPERATIONS RESEARCH AND MANAGEMENT SCIENCE

DAVID F. ROGERS*

Department of Quantitative Analysis
and Operations Management,
College of Business, University of
Cincinnati, Cincinnati, Ohio



OMEGA RHO

OMEGA RHO INTERNATIONAL HONOR SOCIETY
Operations Research and Management Science

Omega Rho International Honor Society is dedicated to serving students interested in Operations Research and Management Science (ORMS) and primarily does that by honoring students who have achieved sufficiently high grades during their undergraduate or graduate programs. Student inductees are provided with a membership certificate, a piece of Omega Rho jewelry, and a line on their resume that indicates achievement, recognition, involvement, and distinction. Chapters may also induct faculty who are actively involved in teaching or research in ORMS and the executive committee may induct honorary members—persons who have made conspicuous contributions to ORMS. Omega Rho also furthers the discipline of ORMS by sponsoring the *Omega Rho Distinguished Lecture Series*, since 1983 at joint national meetings of the Operations Research Society of America (ORSA) and The Institute of Management Sciences (TIMS), and after their merger in 1995 at the Institute for Operations Research and the Management Sciences (INFORMS) national

meetings. Select INFORMS International Meetings have hosted an *Omega Rho International Distinguished Lecture* since 2001. Omega Rho has consistently cosponsored the ORSA/TIMS/INFORMS Doctoral Colloquium since the 1980s and has supported numerous other ORMS-related student activities since its inception including, for example, travel stipends for students to present at INFORMS meetings, and sponsorship of regional student-centered conferences.

HISTORY AND TRADITIONS OF OMEGA RHO

In 1975, Clint Ancker, Chair of Industrial Engineering at the University of Southern California, lamented the fact that while there were many collegiate honor societies for intellectual achievement, there were none for operations research. With the help and encouragement of Joseph Engel, University of Illinois at Chicago Circle and a Past President (1968) of ORSA, Jack Borsting from the Naval Postgraduate School and President of ORSA during 1975, and others, Omega Rho was founded on April 1, 1976 at the Joint National Meeting of ORSA and TIMS in Philadelphia, Pennsylvania. These 15

*Historian, Website Manager, and Past President. Omega Rho International Honor Society of INFORMS.

universities were the inaugural set of schools to begin inducting members: Case Western Reserve, Clemson, Columbia, George Washington, Illinois-Chicago, Miami University (Ohio), Naval Postgraduate School, New York University, North Carolina State, Maryland, Oregon State, Pittsburgh, Southern California, Southern Methodist, and SUNY Buffalo.

Omega Rho was admitted to the Association of College Honor Societies (ACHS), an organization founded in 1925 to promote high and consistent standards for college and university honor societies, as an Associate Member in 1983 and obtained Full Member status in 1986. ACHS consists of 64 honor societies that are “individually and collaboratively exhibiting excellence in scholarship, service, programs, and governance” [1] and its main goal is to maintain high standards for the recognition and promotion of academic excellence in higher education.

“Ad Optimum per Omega Rho” is the formal public motto for Omega Rho as prescribed in the Constitution. When translated to English, the motto means “Toward the Optimum through Omega Rho.” The official seal is most appropriate for ORMS—a projection of a truncated blue saddle surface with a vertical red arrow centrally located at its base at the bottom of the circle and its head at the saddle point, with the Greek capital letters Ω and ρ in black, and the motto in red.

In October 27, 1998, at the Seattle INFORMS National Meeting, the Omega Rho Executive Committee modified its Constitution, and the INFORMS Board adopted a bylaw, to formally recognize Omega Rho as the Honor Society of INFORMS. Separate meetings for the general membership and the executive committee are held at the INFORMS Annual Meeting each Fall.

Overseeing the Society is the Omega Rho Executive Committee consisting of 15 officers. It is classically structured with a president, vice president/president-elect, secretary, treasurer, historian, two past presidents, and eight regional directors. There are four regional directors for North America (Northeastern, Southeastern, Central, and Western) and one each for Africa, Asia-Pacific, Central and South America, and Europe. Serving as one of these officers

requires two years of valuable service to the ORMS community. An INFORMS liaison is formally designated to work with the Society and the INFORMS Chapter Relations Coordinator currently fills that position. The Executive Committee meets annually at the INFORMS National Meeting, and a General Membership meeting, open to all Omega Rho members, is also held there.

OMEGA RHO MEMBERSHIP IS FOR RECOGNIZING EXCELLENCE

Student participation through one of the 38 established chapters is the life blood of Omega Rho. Chapters are located through universities and have individual constitutions and bylaws. Minimal requirements for membership at the society level are that undergraduate students be in the top 25% of their class and that graduate students have at least a 3.5/4.0 grade point average. However, some chapters have instituted more stringent requirements in their local bylaws. Members may be involved in a wide variety of curricular studies. While an Omega Rho Chapter may be based in a college of business (management science, decision science, quantitative analysis), engineering (industrial/systems engineering, operations research), or science (mathematics, statistics), any ORMS-based curricular study qualifies. The only requirement for student qualification, regardless of major, is success in a significant set of courses in ORMS as deemed satisfactory by the chapter faculty advisor.

If a university does not have a chapter, the Omega Rho Bylaws allow individuals to petition the President for membership. Joining Omega Rho means that a student becomes a member for the remainder of their university career and an alumni member after graduation for life. The one-time cost for joining, US \$20 in 2009, has held steady for at least two decades! A certificate and handsome jewelry depicting a saddle point as in the logo are provided to each member. Honor cords for wearing at graduation ceremonies are also available in blue and red, the official Omega Rho colors.

A faculty performing teaching and/or research in ORMS at an established university chapter qualifies for status as a faculty member. Visitors, such as someone from another university teaching a course or a visitor who is doing research with faculty or students while on a sabbatical leave, also qualify for induction by a local chapter. Upon leaving the university, faculty members also become alumni members of the local chapter. Approximately 6100 students and faculty have been inducted through 2009. The membership directory on the INFORMS website now lists Omega Rho members through their initiating chapter.

Honorary member status is only awarded to exceptional individuals in ORMS and must be approved by the executive committee. Primarily, this status is employed to honor the presenter of the *Omega Rho Distinguished Lecture* and *Omega Rho International Distinguished Lecture* but has also been occasionally used by local chapters to honor worthy visitors and speakers. Through 2009, 43 honorary members have been inducted into Omega Rho.

THE OMEGA RHO DISTINGUISHED LECTURE SERIES

“To recognize academic excellence”—a part of the Oath for Omega Rho—has been extended to significantly recognize the best in research and application through the *Omega Rho Distinguished Lecture Series*, arguably one of the most prominent series of lectures on ORMS in the world. Recognized as an “Area of Excellence” by ACHS [2], the distinguished lecture has been a plenary presentation at ORSA/TIMS/INFORMS meetings since Russell L. Ackoff delivered the first such lecture on November 8, 1983, at the ORSA/TIMS Orlando, Florida, meeting. Other notable distinguished lecturers have included George Dantzig, Nobelist Herbert A. Simon, Alan B. Pritsker, and a host of contemporary honorees, all in a series of presentations of unsurpassed quality in ORMS research and practice. *Omega Rho Distinguished Lectures* have addressed a wide variety of novel topics in such important areas as health-care delivery, airline

safety, public sector issues, e-commerce, service management, and manufacturing simulation, while other lectures were on the more traditional topics of queueing, linear programming, and artificial intelligence. Some of the more provocative talks in the series dealt with evaluating life and death decisions, resurrecting the economy, resizing the federal government, ethical and social responsibility issues in ORMS, human genome sequencing, and HIV prevention.

The *Omega Rho International Distinguished Lecture* has been delivered at four select INFORMS International Meetings. William R. Pulleyblank presented the first such lecture during the Society’s 25th anniversary at the INFORMS International Meeting in Hawaii in 2001. A total of 35 *Distinguished Lectures* and *International Distinguished Lectures* have been presented through 2009. Complete details for each lecture are available at the Omega Rho website, OmegaRho.INFORMS.org.

SERVICE FOR OMEGA RHO IS REWARDING

Service for an honor society is student-oriented, consistent with the educational goals of most university settings, and usually regarded as a worthy activity by department chairs, deans, and provosts. Working with Omega Rho is gratifying from many perspectives:

- Faculty work to honor and serve the best students and promote academic excellence with the ORMS community at large.
- Omega Rho promotes and sponsors activities to showcase the best of ORMS.
- Omega Rho provides service to the ORMS profession with support of the INFORMS doctoral colloquium and other student-based activities.
- Omega Rho helps spread the idea of the value of ORMS to the outside community and the value of “The Science of Better.”

Further credit comes to those who serve Omega Rho as president. The post

involves an eight-year commitment—two years as president, two prior years as vice president/president-elect, and a subsequent four-year term as past president and executive committee member. Often, the president has already served as secretary,

treasurer, historian, regional director, and/or a faculty advisor. Omega Rho has been fortunate to have had the following talented, hard-working senior faculty serve as President:

Clinton J. Ancker, Jr.	University of Southern California	1976–1978
Joseph H. Engel	University of Illinois at Chicago Circle	1978–1980
Burton V. Dean	Case Western Reserve University	1980–1982
E. Leonard Arnoff	University of Cincinnati	1982–1984
Dundar F. Kocaoglu	Portland State University	1984–1986
Saul I. Gass	University of Maryland	1986–1988
Paul Gray	Claremont Graduate University	1988–1990
A. Ravindran	University of Oklahoma	1990–1992
Mary W. Cooper	Southern Illinois University	1992–1994
U. Narayan Bhat	Southern Methodist University	1994–1996
Bruce W. Schmeiser	Purdue University	1996–1998
David F. Rogers	University of Cincinnati	1998–2000
Richard M. Soland	George Washington University	2000–2002
Yupo Chan	University of Arkansas at Little Rock	2002–2004
George G. Polak	Wright State University	2004–2005
Subhash C. Narula	Virginia Commonwealth University	2005–2008
Christopher M. Rump	Bowling Green State University	2008–2010
James K. Lowe	United States Air Force Academy	2010–2012

At the 75th anniversary celebration of the ACHS held on February 19, 2000, each honor society recognized “five persons who have contributed to the well-being of the Society.” The five recognized for Omega Rho were all former presidents:

- Clinton Ancker and Joe Engel, the first presidents and driving forces behind founding Omega Rho and making it operational.
- Richard Soland, who served as long-time executive director and treasurer and was key in maintaining and enhancing Omega Rho’s presence through the 1980s and 1990s.
- Saul Gass, involved in creating Omega Rho and in all the years since then recognized for consistently serving Omega Rho on many committees including the Bylaws Revision Committee.

- David Rogers, who expanded the presence of Omega Rho through new chapters, chaired the 1998 Bylaws Revision Committee, spearheaded Omega Rho becoming a part of INFORMS, and developed and maintained the extensive archival internet website for the society.

OMEGA RHO LOOKS TO THE FUTURE

Omega Rho has a ceaseless desire to first and foremost serve and recognize students, faculty, and professionals of ORMS. We are also relentless in our efforts to help spread awareness of the value of the profession. An ongoing activity is promoting growth by establishing new chapters and the Omega Rho website has information and documents to assist with this endeavor. The *Omega Rho Distinguished Lectures Series*, sponsorships

for the INFORMS Doctoral Colloquium, and other student-oriented activities will likely remain a constant of Omega Rho activities. As the world has grown smaller with virtual connections, as honor societies become more typical in universities where they have not previously existed, and as ORMS spreads to developing nations, Omega Rho will continue to increase its chapters worldwide to include recognition for deserving individuals around the globe.

More information about Omega Rho can be accessed from its website at OmegaRho.INFORMS.org or through the INFORMS website, www.INFORMS.org under "Communities." There one can find current and archival information including the constitution and bylaws, oath, current and past officers, chapter locations, meeting minutes, honorary members, the *Distinguished Lecture Series*, upcoming meetings, and assistance for starting a new chapter.

Acknowledgments

Past Omega Rho Presidents Paul Gray, Richard Soland, and Yupo Chan are gratefully recognized for their editorial inputs. Artwork in this article is the property of Omega Rho International Honor Society.

REFERENCES

1. Association of College Honor Societies. Mission. Available at www.achsntl.org. Accessed 2009 Sept 16.
2. Association of College Honor Societies, Omega Rho: Operations Research and Management Science—Distinguished Lecture Series, ACHS Areas of Excellence, 4990 Northwind Drive, Suite 400, East Lansing, Michigan, 2000.

“ON-THE-SPOT” MODELING AND ANALYSIS: THE FACILITATED MODELING APPROACH

LUIS A. FRANCO
Warwick Business School,
University of Warwick,
Coventry, UK

GILBERTO MONTIBELLER
Department of Management,
London School of Economics,
London, UK

APPROACHES TO MODEL-BASED CONSULTANCY

The most common approach to undertaking model-based consultancy in management science has been what is known as the *expert modeling* mode, where the management scientist uses powerful quantitative methods and models that enable an “objective” analysis of the client’s problem situation, together with the recommendation of optimal (or quasi-optimal) solutions to alleviate that problem situation. Most management science textbooks offer excellent advice on how to deploy the expert mode of model-based consultancy, which has been successfully used to solve a broad range of challenging management problems in areas such as logistics, operations, marketing, and finance. This is aptly illustrated by the EURO Medal and the Franz Edelman prizes, two of the most prestigious awards for the practice and proven impact of management science and operational research projects.

Broadly speaking, we can identify four tenets underlying the expert modeling mode of model-based consultancy. First, the expert modeling mode assumes that the problem situation to be addressed by the consultant exists as an external reality, and therefore its actual definition and scope should not depend on which stakeholder is describing

such reality. Indeed, the task of the consultant is to remove any “bias” in the description of the problem in order to solve the “real problem” [1]. This obviously has an impact on how problems are framed and formulated. In particular, it means that it is the task of the consultant alone to frame “correctly” the problem and formulate “precisely” the model, without the need of major client involvement in the modeling process.

Second, model-based analysis in the expert mode should be objective, and thus individual opinions and different views about the problem should be at least suspended and at best avoided [2]. This leads consultants to define, by themselves, the metrics that they will be using for evaluating solutions, based on prior categorizations of problems and how these are typically solved. It also means that analysis will be based mostly on “hard” quantitative data and conducted in the backroom, as modeling and analysis is likely to be technically challenging.

The expert modeling mode also assumes that clients hire consultants for their expertise in providing optimal (or quasi-optimal) solutions to problems [2]. This is usually communicated via a detailed report that contains the optimal recommendations as well as a summary of the main assumption(s) employed in the analysis. Finally, there is an expectation among expert consultants that the implementation of solutions derived from well-conducted model-based analysis will be relatively straightforward. As the solutions proposed are clearly optimal, there is little reason why they should not be implemented by the client who commissioned the analysis [3,4].

The expert modeling mode has proven to be a natural and sensible way to solving complex operational or recurrent problems in organizations. On the other hand, when dealing with problem situations at a more strategic or policy level, the expert mode of consultancy may not always be appropriate for several reasons. First, strategic or policy

issues typically generate competing hypotheses about the scope and depth of the problem situation to be tackled, which will make its definition a contested process for the organization's stakeholders. In addition, the distinct and, often conflictive, objectives, values, and interests of the stakeholders will have to be accommodated before and during the analysis, in order to reach a “working consensus” about the way forward. Finally, such situations pose a substantial challenge to both clients and consultants regarding decisions about the appropriate level of participation in the problem solving process, which will have a substantial impact on whether the solutions derived from the analysis are both desirable for the stakeholders, and politically feasible (i.e., implementable and supportable) for the organizational units they represent [5,6].

For problem situations such as those described above, an alternative mode of model-based consultancy, known as the *facilitated modeling* mode has been developed [7]. In this mode a team is drawn from the client organization (or sometimes more than one organization with a stake in the problem) and is typically placed as responsible for scoping, analyzing, and solving the problem situation of interest; the team is supported by the consultant, who acts both as an analyst and as a facilitator. Every step taken in the intervention—from defining what the problem really is, to creating and analyzing models, and providing the recommendations—is conducted interactively with the team. This produces noticeable differences in the way the consultancy engagement is undertaken, as we discuss next.

The facilitated modeling mode recognizes that, while some aspects of problems are tangible, their nature and salience will depend on how stakeholders subjectively construct them. Different stakeholders will perceive the same problem situation in quite diverse ways, influenced by their distinct interests and organizational roles and, therefore, will describe the situation according to those perceptions. Instead of trying to define the “real” problem, the consultant in this mode will support the team to come up with a joint problem definition, which encompasses the

main features of their individual perceptions about the situation [1,3,8].

In addition, the facilitated modeling mode acknowledges that problem situations always involve subjective elements; for example, different perceptions about the desired future of the organization, distinct opinions about what or which organizational objectives to pursue, and the presence of organizational objectives that are of a qualitative nature [4,5,9]. Instead of disregarding such aspects as “nonrational” or irrelevant, the consultant operating under the facilitated modeling paradigm helps the team to externalize such subjective issues and represent them in formal models.

The facilitated modeling mode also assumes that results from model-based analysis are considered by clients as a mere indication of what would happen if the solution were implemented. The view among facilitated modeling consultants is that clients are usually satisfied in finding a set of good options that can provide clear improvements to the organizational system, and which are deemed implementable at the same time [10,11].

Finally, a major tenet of the facilitated modeling mode is that the involvement of the key stakeholders in the process of modeling and analysis markedly increases the chance that recommendations of the analysis will actually be implemented [5,6,12–14]. The underlying rationale is that their involvement will make stakeholders more confident in the analysis performed, and more committed to the recommendations derived from it. Such confidence and commitment will be strengthened if: key stakeholders believe that their views, preferences, and objectives are taken into account; the model represents adequately the problem they want to tackle; the modeling assumptions are organizationally realistic; and the solutions derived from the analysis are sound and justifiable.

In summary, two alternative models of model-based consultancy are available to the management scientist: expert and facilitated modeling. Neither of these modes is necessarily superior. For complex operational problems, the expert modeling mode is usually quite appropriate: there is a clear and unarguable quantitative objective to be

optimized; the structure of the problem is well-known and renders itself to quantitative modeling; and solutions are clear-cut and easily implementable. On the other hand, strategic and policy issues and high-stake decisions are likely to require a facilitated modeling approach due to their complex quantitative and qualitative dimensions, and the need to engage a team of stakeholders in the problem solving process [5,11,15,16].

The rest of the article is structured as follows. In the next section we discuss four key dimensions of facilitated modeling: the characteristics of the modeling process; the nature of the models produced; facilitated modeling outcomes; and facilitation issues. We then briefly review several well-established facilitated modeling approaches developed within the operational research, decision sciences, and systems fields. In the latter part of the article we examine some of the key design choices open to the consultant interested in applying facilitated modeling in practice. We end the article with conclusions and suggest possible directions for future work in the field of facilitated modeling.

WHAT IS FACILITATED MODELING?

We use the term *facilitated modeling* to describe a process by which formal models are jointly developed with a group of stakeholders, with or without the assistance of computer support, and in which insights about the problem situation being modeled is the result from model-based analysis produced “on-the-spot” [7,16–18]. We consider a model as “formal” if it represents a problem situation in any of the following ways: (i) activity or process flows; (ii) cause and effect relationships; or (iii) relationships between decision choices and their (deterministic or uncertain) consequences. A formal model is amenable to analysis and manipulation, but not necessarily quantifiable. Models produced in the facilitated modeling mode are “transitional objects” [17,19] used by stakeholders to share and increase their individual understandings of the problem situation of interest, appreciate the potential impact of different options, and negotiate courses of action that are politically feasible.

In facilitated modeling, the consultant must use facilitation to support the modeling process. Model-based facilitation, however, is different from the type of group facilitation that is commonly used in the organizational development field [20,21]. Although both forms of facilitation are similar in the way the consultant provides process support to a group, it is the handling of the content of group members’ communicative exchanges where the main difference lies. In facilitated modeling, content is managed through the use of formal models, as opposed to other facilitative means. Modeling and model use is the defining characteristic of facilitated modeling, and what gives it its unambiguous management science identity.

Figure 1 captures the main aspects of facilitated modeling. The figure suggests that the consultant interacts simultaneously in two “spaces”: as an analyst in the *modeling space* and as a facilitator in the *group process space*. In the former, the consultant uses a particular management science methodology to build a model that is a representation of the problem situation as described by the group. Meanwhile, in the *group process space*, the consultant facilitates the group (informed by facilitation methods and tools), supports the group’s discussion and interacts with the group’s members. The group provides information that enables the consultant to model the problem situation of concern and, conversely, the model generates responses that inform the group’s discussions about the situation. Two types of outcomes are provided by facilitated modeling: in the modeling space, there are *model outcomes* (such as priorities for actions, system’s responses, etc.); in the group process space, there are *group outcomes* (such as commitment to action, learning, etc.). The two spaces cannot be divorced from each other [22], as there will be cross-impacts from one space to the other (for example, conceptual mistakes in the model may generate meaningless system’s responses, which then impact negatively on group outcomes, such as creating low commitment to action).

In the section below, we examine the notion of facilitated modeling in more detail.

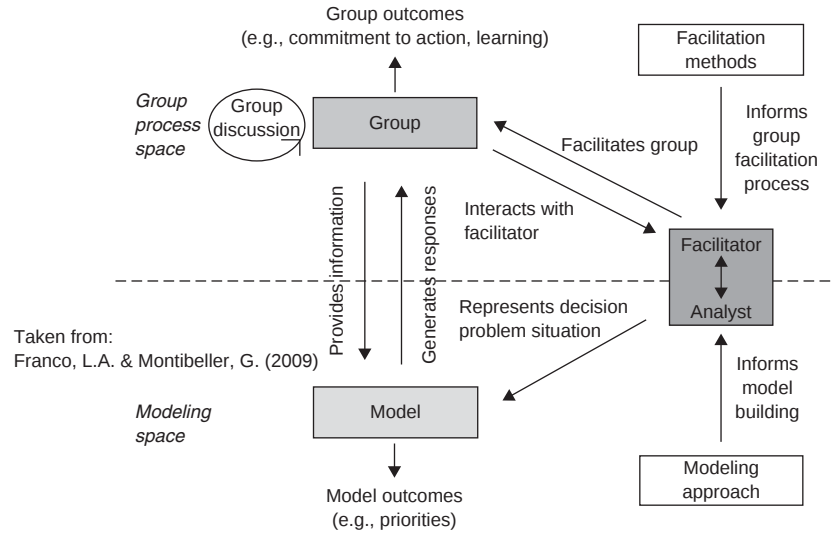


Figure 1. The facilitated modeling process (Source: Based on Franco and Montibeller [7]).

Our discussion will be organized around four key aspects of facilitated modeling: process, models, outcomes, and facilitation skills.

The Process of Facilitated Modeling

The primary orientation of facilitated modeling is to assist a group of stakeholders in agreeing with the nature of a problem situation of concern, and on a feasible action plan intended to tackle that situation so that progress can be made. This is because, when dealing with situations involving strategic problems, complex policy issues, or high-stake decisions, there is likely to be a plurality of stakeholders with different interests who will need to engage in a dialogue if the problems they face are to be resolved by means other than an overt exercise of power [5].

Consequently, facilitated modeling can be thought of as a conversational process [23] in which stakeholders take part to exchange their understandings and views about the situation that is being analyzed. This process is thus a participative one, in the sense that stakeholders are able to jointly construct the situation, make sense of it, arrive at a shared problem definition, and develop and evaluate a portfolio of options relevant to the problem so defined. This participatory process is

supported by the consultant acting both as a modeler and a facilitator [24,25].

Because interaction between the stakeholders, and of the stakeholders with the analysis, is needed to jointly build a model of the problem situation, facilitated modeling is also an interactive process. Stakeholders' interaction with the model reshapes the analysis, and the model analysis reshapes the group discussion. Such interactive processes continue until the problem situation is satisfactorily structured and analyzed, so that the group feels sufficiently confident in making commitments and implementing options.

Facilitated modeling is typically organized into group work stages that roughly correspond to: structuring the situation and agreeing on a problem focus; developing a model of organizational objectives or systems; creating, refining, and evaluating options for action; and, developing implementation plans. This “staged-ness” makes it possible for stakeholders to end the process without passing through all the stages, and still have a tangible product that can be of use to them. Furthermore, the stages of facilitated modeling do not have to be followed in a linear sequence; rather, it is possible for the participants to cycle between the stages.

A characteristic feature of facilitated modeling is its ability to enable stakeholders, during the modeling process, to distance themselves from previous bindings, effectively providing them with a certain degree of detachment regarding their own stakes [11,17]. This, it is argued, allows stakeholders to change their initial positions in response to what they have learned about the problem situation without destroying the social order in the group [17]. Changing positions imply stakeholders “changing their minds,” that is, changed beliefs, changed values, and changes in the salience of particular issues or values [26]. The consequence of these changes is that it becomes easier for participants to reconcile the position they eventually take both with principles and with past words and actions during the modeling process.

Thus far we have looked at the characteristics of a facilitated modeling process. As we have seen, facilitated modeling requires a consultant supporting a group model-building process that must be participatory, interactive, staged, and nonlinear, and one that enables stakeholders a certain degree of detachment from their stakes in order to make change possible. At the same time, the consultant must be responsive to the dynamics of group work and the particularities of the situation at hand.

In terms of technology, facilitated modeling can be a relatively unsophisticated activity, typically conducted in a workshop format, and one which does not necessarily require software to support it [27]. Noncomputer-supported environments typically involve [28–30]: a room spacious enough for participants to move around freely and with movable chairs laid out in a horse-shoe fashion; large sheets of paper attached around the walls of the room; a simple, nonpermanent means of sticking papers to these walls (e.g., blu-tack); and a good supply of marker pens with contrasting colors. Facilitated modeling can also be deployed with computer support. In this case, specialized software is used to support the processes [17,31,32]. This kind of software enables fast model building and real-time computing [17,27]. Some software, such as *Group Explorer* (www.banxia.com) and *VISA groupware*

(<http://www.simul8.com/products/visagroup.htm>), also allows participants to enter their views relating to a problem situation directly and anonymously into it. The system is then operated by the consultant who manipulates and analyzes the data according to the requirements of the group or the task. Once a model of the problem is built and stored in the system, several analyses can be performed “on-the-spot.” Figure 2 illustrates both computer and noncomputer-supported modeling environments.

The Nature of Models in Facilitated Modeling

As stated earlier, models produced in facilitated modeling interventions provide stakeholders with a facilitative device or transitional object [17,19], that is, a “play tool” that allows them to rehearse ideas and action possibilities about the situation of interest [18]. The availability of such a tool, it is argued, increases stakeholders’ multiple understandings of the problem situation, and supports them in negotiating courses of action that are desirable and feasible for the stakeholders.

The models resulting from facilitated modeling represent, among other things, relationships between concepts, activities, or stakeholders; relationships of similarity or influence; and relationships between options. The particular elements contained in any model will depend on the focus of the model (e.g., choices of options, design of systems) and on the group task they mainly support (i.e., problem structuring, decision making, or both).

It has been claimed that visual methods are of particular value in representing complexity to lay audiences who might otherwise find quantitative models opaque [5,11]. For a model to be considered a useful and legitimate facilitative device by stakeholders, it should be open (i.e., nothing is hidden), transparent (i.e., easy to understand), and accessible (i.e., simple to use). Thus, “facilitative models” are essentially visual displays (e.g., cognitive and causal maps, causal loop diagrams, decision graphs, value trees) containing the stakeholders’ own language used to describe the problem situation of concern,



Figure 2. Computer- and noncomputer-supported environments in facilitated modeling.

together with their own judgmental preferences implied by different decision options.

A consequence of considering a model as a facilitative mechanism is the challenge of assessing what constitutes a valid model and a good model solution. Probably the best answer to this issue is to employ the concept of “requisite-ness” [10]: a model is requisite if it contains sufficient knowledge and information to help the stakeholders find a way forward about the problem situation of concern.

The Outcomes of Facilitated Modeling

A number of outcomes have been claimed to be the result of the use of facilitated modeling approaches. Some of them will be tangible outputs of the modeling process itself, while others will be less visible but valuable in their own right.

The most visible facilitated modeling outcome is obviously “the model” object that is built during the process. The model contains the structure of the problem situation, and allows performing on-the-spot analyses and drawing conclusions from its responses. Furthermore, the model is thought to facilitate the achievement of a number of invisible outcomes, of which we will cite three considered central. First, it is argued that by allowing the mutual exploration of the problem structure as portrayed by the model, facilitated modeling enables the accommodation of multiple and differing positions among stakeholders [33]. This argument is based on the notion that strategic problems, policy issues, or high-stake decisions will commonly

require stakeholders to adjust their positions and/or expectations to take into consideration the possible objectives and interests of others [5]. Accommodations between stakeholders may also require coalition forming [26,34], which can produce a shift in power relations between the stakeholders participating in the modeling process [17].

Second, the analysis and manipulation of the relationships between model variables, together with the evaluation of decision options, is thought to give stakeholders an increased and shared understanding (rather than full consensus) of the problem situation, of the impact of potential courses of action, of others’ beliefs and values, and of organizational roles, processes and cultures [5]. Such gained understanding is taken to be conducive to developing a sense of common purpose that preserves individual differences of opinion among stakeholders [13,35], as well as learning [6,14,36,37].

Finally, it is argued that stakeholders’ active participation in modeling and analysis produces strong ownership of the problem formulation as well as a commitment to the way forward, in the sense of an acceptance of joint responsibility for the consequences of the actions taken [5,6,13].

Facilitation Issues

As already stated, facilitated modeling requires the consultant to act both as an analyst and as a facilitator during the modeling process. This means that the consultant should be not only proficient in the particular

modeling approach to be deployed with the group, but also adept in group facilitation to support the modeling work. A comprehensive discussion of the general skills and competences required to facilitate groups is beyond the scope of this article, and the interested reader is invited to consult the mainstream facilitation literature [20,38,39]. In the discussion below, we consider issues specific to facilitating groups within the context of modeling, drawing mainly upon the seminal work of Phillips and Phillips [24].

It is argued that when a group gets together to perform a task, the group acquires a character or a personality while working on the task. The group also has an emotional life that develops and changes over time, sometimes in abrupt and unpredictable ways, which can substantially affect the effectiveness with which the members of the group carry out their task [40].

While group life can affect the individuals in a group, so also each individual can affect the group. The individuals and the group mutually interact and influence each other and can be thought of as a “system.” Furthermore, tensions within this system can arise because of the interplay between individuals and the group. These tensions are the main source of conflict in the group, as well as the anxiety and even frustration commonly experienced by group members. If such conflict, anxiety or frustration are not acknowledged, tolerated, and held by group members, they may be dealt with it in non-effective ways. For example, group members may project it onto other group members or the facilitator; group members may form coalitions or become dependent on the facilitator [40]; or group members may avoid and table conflict altogether to maintain the cohesion of the group, a typical manifestation of “group thinking” [41] and “group centrism” [42]. These coping strategies reduce conflict, anxiety, or frustration by diverting the group from the task at hand.

Although the group life works its effect through the group members, it is often the case that they do not see it because their attention is focused on the task. The consultant must therefore be able to recognize and understand the effect of the group life

on the performance of the group task. This can be achieved in three main ways [24]: by observing group roles and relationships, by attending to symbolic content in the group discussion, and by monitoring feelings (the group’s as well as the consultant’s). Understanding the group life is, however, not sufficient. The consultant must also intervene when appropriate in order to help the group carry out their task more effectively. Without giving advice, providing evaluation or interpretation, or contributing to content, the consultant can intervene in at least four different ways [24]: by pacing the task against progress; by directing the group about work steps; by reflecting back content or handing it back in changed form to provide a new perspective; and by questioning and summarizing to increase understanding. The interested reader is referred to Phillips and Phillips [24] and Ackermann [25] for a more detailed discussion of observation and intervention strategies available to the facilitated modeling consultant.

Before we discuss choices related to the design of facilitated modeling interventions, we offer next a brief review of some of the more established facilitated modeling approaches.

BRIEF SUMMARY OF FACILITATED MODELING APPROACHES

We will attempt here to very briefly summarize a number of well-established facilitated modeling approaches within management science. We have selected a family of approaches that, in our view, sit comfortably within the definition and characteristics articulated above. We classify them in three groups according to the general modeling methodologies from which they are derived (in alphabetical order): decision analysis, problem structuring methods, and system dynamics.

Facilitated Decision Analysis

The combined use of decision analysis methods, group facilitation, and computer-supported modeling led to the development of a facilitated modeling approach known

as *decision conferencing* in the late 1970s [32,43,44]. During a decision conference, computer-based models are used as a focus for group discussion and the structured group process allows participants to debate issues constructively while forcing them to represent their collective judgments in a logically consistent and easily communicated fashion. Although supporting information is important, it is the substantive expertise of the participants that is key to the success of decision conferences. For further details, see Phillips [13].

Facilitated Problem Structuring

A set of methods originated mainly within the British operational research community in the late 1970s and early 1980s. Their main features are the assumption of subjectivism (different views about the world); the significance of problem formulation and definition in facilitating joint agreements and coordinated action; and the limited role of quantification in the analysis of problem situations (mostly qualitative modeling). Well-established problem structuring methods include strategic options development and analysis (SODA) [34,45], journey making [6], strategic choice approach [12], value focused thinking [9], and dialogue mapping [46]. For further details, see Rosenhead and Mingers [5] and Jackson [47].

Facilitated System Dynamics

Since the late 1980s, the system dynamics community has made considerable progress in developing tools and techniques to support a facilitated modeling process. The result has been an approach known as *group model building* [48–51], which employs a mixture of system dynamics modeling techniques, such as causal loop diagrams, stock and flows diagrams, and graphical functions, in combination with guidelines for facilitating group sessions to foster organizational learning and change [52–54]. The interested reader is directed to Vennix [14] for further details.

Facilitated modeling has been applied in a wide variety of areas including health [55–60]; transport [61,62]; natural

resource management [63–66]; manufacturing [67,68]; project management [69–71]; knowledge management [72,73]; strategy development [74–78]; information systems [79–81]; interorganizational collaboration [82,83]; community work [84–87]; portfolio analysis [88–90]; and project prioritization [91]. It is worth noting that, although we have presented the above families of facilitated modeling as separate entities, an increasing trend observed in practice combines or integrates different facilitated modeling approaches to create novel “multimethodology” designs [92] that aim to take benefit from their strengths—for recent multimethodology applications of facilitated modeling, see Bana e Costa [91], Andersen [93], Franco and Lord [94], and Montibeller [95].

DESIGN CHOICES IN FACILITATED MODELING

In this section we focus on design choices that the consultant may consider when conducting interventions based on facilitated modeling. Our emphasis will be on the design of the facilitated modeling process itself, rather than the entire intervention. Broader design issues related to facilitated modeling interventions, including the management of the consultant–client relationship, have already been treated extensively elsewhere [4,24,25,34,96], and the interested reader is referred to these works.

The focus of our discussion here will be on a single facilitated modeling session delivered in a one- to three-day workshop format, which is the typical format adopted by many facilitated modeling approaches. We have identified six dimensions as key when designing such a session: modeling focus, data collection strategy, data requirements, technology support, flexibility of modeling rules, and content facilitation requirements. Each of these are discussed next.

The first dimension is the *modeling focus*, as a facilitated modeling session can be designed mainly to support problem structuring or the evaluation of options. Problem structuring is typically supported via problem structuring methods such as SODA

or the strategic choice approach, whereas approaches such as group model building and decision conferencing have been used mostly to support the quantitative evaluation of options. Different modeling skills are required for each phase; for example, problem structuring requires the modeler to cope well with ambiguity, multiple perspectives, and a large amount of qualitative data; while the evaluation of options requires a modeler with synthesizing abilities and the ability to handle a large amount of quantitative data. Supporting both activities, within a single intervention, is also feasible, but it requires the consultant to be able to adopt a multi-methodology approach intervention [92,97].

The second dimension concerns the *data collection strategy*. Some facilitated modeling approaches use well-defined categories to elicit data up front, as in the case of, for example, the strategic choice approach (e.g., decision areas, uncertainty areas, comparison areas) [12]. Other approaches involve eliciting all relevant data before predefined categories are applied to it. This form of data elicitation is typical of approaches such as, for example, SODA, group model building, and dialogue mapping. Different data elicitation approaches pose different demands on how the consultant manages the content of the model: the first approach requires working inductively, whereas a deductive way is required for the second one. The choice of approach will thus depend on the particular preferences of the consultant working either inductively or deductively, as well as on the stakeholder group's adaptability in providing the data in the elicitation formats required by the modeling approach adopted.

A related issue, and the third dimension, is the type of *data requirements* for building the model. Some models are mainly diagrammatic representations of the problem situation and thus have reduced quantitative data requirements (e.g., a cognitive map or a causal loop diagram), closer to the way stakeholders think and communicate. Others, such as a decision analysis model require data in quantified form (even if only judgmental data is available) in order to build the model. The degree of quantification needed for some facilitated modeling approaches thus, requires

the consultant to be aware of the cognitive demands that this may impose on the group. For example, research has shown that quantitative judgmental data can be difficult to elicit [98] and influenced by cognitive biases [99–101]. On the other hand, quantitative models tend to produce outputs that are less ambiguous and more amenable to further analysis. Therefore, the consultant needs to make a choice between these two conflicting modeling objectives, given the type of problem situation that the stakeholder group wishes to address and the background of those participating in the modeling process.

The amount of data elicited (whether qualitative or quantitative) in facilitated modeling sessions is usually high, which can challenge participants' information processing capabilities [102,103]. This raises our fourth dimension, the *degree of technology support* needed for a facilitated modeling session (where the term *technology* is used to describe the level of computer support used in the session). Problem structuring methods such as the strategic choice approach, for example, have traditionally favored unsophisticated forms of technology support (i.e., manual rather than computer-based). On the other hand, approaches within which models are built to perform numerical calculations and/or quantitative sensitivity analyses, such as decision conferencing or group model building, tend to rely almost exclusively on computer support [32], which can be used in “single-user” mode (facilitator operates the modeling software) or multiple-user mode (participants are able to use the modeling software themselves guided by the facilitator) [104]. Other modeling approaches such as SODA tend to use a combination of both computer and noncomputer support.

A decision about which level of technology support is appropriate for a facilitated modeling session will depend on several factors. First, the consultant's personal preferences for the use of technology are an obvious influence, as well as the perceived complexity associated with the modeling task. As stated earlier, facilitated modeling requires “on-the-spot” modeling and analysis of a large amount of data. A noncomputer-supported environment will restrict the speed at which

the data is elicited and manipulated, making the task relatively easier for both the consultant and the participants. On the other hand, computer-supported environments may be more effective in terms of collecting large amounts of data in less time (particularly those that allow a multiuser mode of working), but they can pose considerable challenges to the modeler's ability to structure the data, as well as to the participants' understanding of what is going on, due to the speed at which data is elicited and structured. Consequently, when the different levels of computer technology are being considered, having more than one facilitator/modeler for the session may be needed [31].

The fifth dimension is the *degree of flexibility of the modeling rules* being employed. Some methodologies underlying particular facilitated modeling approaches, such as decision analysis (for decision conferencing) and system dynamics (for group model building), require stricter specification of variables and the fulfilment of a larger number of structural properties. On the other hand, facilitated modeling approaches such as SODA are usually more flexible in their modeling rules. Stricter rules usually permit a higher level of inference in the model, particularly important when appraising alternatives and policies. Conversely, the flexibility is beneficial for representing problem situations where there is high ambiguity of meanings and multiple perspectives, and common features of the modeling context in which problem structuring methods operate. Furthermore, some problems may require different levels of formal analysis and interaction with the group (for instance, from representing the problem with a qualitative model to evaluating quantitatively options). Such features then guide the choice about the degree of flexibility of its modeling rules required and, consequently, the choice of the method to be employed in the intervention.

The final dimension we suggest is the *degree of content facilitation* required in a modeling session. Some modeling approaches require strong content facilitation as in the case of SODA, decision conferencing, and group model building, where more than one facilitator/modeler for the session may be

advisable or even necessary [31]. On the other hand, approaches such as the strategic choice approach can allow a group to become self-facilitated after the modeling guidelines of the approach have been internalized by the participants. This means that a less central role may be required for the consultant, who will only have to ensure that the modeling guidelines are being followed throughout the process. In addition, allowing groups to self-facilitate themselves for certain tasks within the modeling session can help to keep up the momentum, and energy levels, within the group. On the other hand, it does require that the group is self-motivated and competent enough to perform the modeling tasks that the consultant asks them to carry out.

The six dimensions we discussed above are somehow idiosyncratic, as they are drawn from our particular experiences as academic practitioners of facilitated modeling. Nevertheless, our aim has been to provide some useful insights into the design choices open to the consultant interested in applying facilitated modeling. In Table 1 we present a summary of each dimension, with its description, range, and some implications for practice.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

This article reviewed a particular mode of model-based intervention in organizations: facilitated modeling—a process by which models are created jointly with a stakeholder group within a facilitation paradigm. Our discussion distinguished the facilitated modeling mode of intervention from the expert modeling approach typically adopted by consultants within the management science domain. We advanced a definition for facilitated modeling, and discussed several of its main aspects such as process, models, outcomes, and facilitation issues. We also proposed a classification of facilitated modeling approaches, based on the modeling methodologies from which they have been developed, and a set of choice dimensions that may be relevant when designing facilitated modeling interventions. Finally, we noted that facilitated modeling demands a particular set of skills

Table 1. Design Choices in Facilitated Modeling (based on Franco and Montibeller [7])

Dimension	Description		Range		Choice Influenced by:
Modeling focus	The extent to which the approach provides support to a particular phase of decision making	Emphasis on problem structuring	↔	Emphasis on the evaluation of options	- Facilitation skills of the consultant - Type of problem to be tackled
Data collection strategy	The way in which the data about the problem is gathered and employed to structure the model	Inductive (categories are created from data)	↔	Deductive (data is elicited for predefined categories)	- The consultant's preferences for working inductively or deductively with data - The adaptability of participants to provide the data required by the model
Data requirements	The type of data required by the model	Data of qualitative nature	↔	Data of quantitative nature	- Type of problem to be tackled - Balance between ambiguity and precision - Participants' background
11 Technology support	The degree of technology support required by the modeling approach	Manual	↔	Computer-supported (single-user or multiuser)	- Personal preferences of the consultant - Acceptable levels of risk for modeling "on-the-spot" at different speeds - Availability of more than one facilitator
Flexibility of modeling rules	How flexible are the modeling rules required by the methodology being employed	Flexible	↔	Strict	- Phase of the decision making being supported - Type of problem to be tackled - Type of analysis required
Content facilitation required	The degree of content facilitation required by the modeling process	Self-facilitation by the group	↔	Strong facilitation by the consultant	- Demands placed on the consultant - Need for keeping momentum and energy levels in the group - Degree of self-motivation and competency of participant

on the part of the consultant. Although these challenges have been discussed in separate management science subfields (e.g., decision analysis, problem structuring methods, system dynamics), we attempted in this article to provide a unifying framework that could allow us to increase our understanding of different forms of facilitated modeling.

Another aspect that is relevant to be highlighted is that, in our view, there is not a “best” mode of organizational intervention (facilitative vs expert mode): such choice has to be made based on the nature of the problem situation that the consultant is dealing with, as well as the consultant’s own preferences. This is in contrast to suggestions that one is better than the other, or that there is one “right” approach to intervention. Nevertheless, we should say that, from our practical experience, the facilitative modeling approach is particularly suitable for supporting organizational problem structuring and/or strategic decision making given its socio-technical complexities.

Several areas for further research can be envisaged, and we suggest some of them below.

- *Better understanding of facilitated modeling processes in practice.* Most of the evidence about the use of facilitative modeling in practice has tended to be purely anecdotal. There have been recent attempts to conducting research on this topic using more systematic observation and analysis [72,105,106], and much could be learnt with more extensive research programs in this area.
- *Generalization of best model-based facilitation practices:* Following from the previous item, a better understanding of facilitated modeling practices could (and should) be focused on trying to generalize best model-based facilitation. We feel that just importing the literature of standard facilitation is not always viable, as facilitated modeling is a complex endeavor requiring a variety of skills and behaviors. Although there is some evidence of work in this area

[25,107,108], there is still considerable scope for further work.

- *Systematic assessment of outcomes from facilitated modeling interventions.* Most of the literature on facilitated modeling makes bold claims on the outcomes that it provides, as reviewed in this article. However, the number of studies that tried to systematically assess such outcomes is quite limited [35,109,110,111]. There is thus wide scope for further studies in this area, both in the laboratory and the field.
- *Test the validity of design dimensions for facilitated interventions.* This article suggested a set of dimensions that, in our view, should be considered when designing a facilitated modeling intervention. It would be interesting to conduct empirical research to test the validity of such dimensions.
- *Increase the family of facilitated modeling approaches in management science.* Most facilitated modeling approaches were developed by adapting a mainstream management science methodology to a facilitation paradigm (problem structuring methods being an exception, as they were specifically developed within a facilitation paradigm in mind). We believe that the facilitation paradigm places constraints in the way that a management science methodology can be designed and deployed in practice (for example, on the type of data that it requires and how it displays information). There is some exciting work already taking place in this direction, for example, in the field of discrete event simulation [112–114]. Further research is required to explore novel ways for redesigning other management science methodologies as facilitated modeling approaches.

Concluding, we believe that the management science community has much to gain in better understanding and further adopting the facilitated modeling approach. The nature of modern organizations, less hierarchical, more networked, with increasing distributed knowledge and power, tends to favor

this type of intervention, particularly for problems characterized by high levels of complexity, uncertainty, and conflict. We hope this article stimulates further research into facilitated modeling, and encourages more management scientists to adopt this mode of organizational intervention.

Acknowledgments

Part of this article is based on research work published by Franco and Montibeller in 2010 [7]. We would like to thank Larry Phillips for encouraging the development of this article.

REFERENCES

- Landry M. A note on the concept of 'problem'. *Organ Stud* 1995;16(2):315–343.
- Ackoff R. The future of operational research is past. *J Oper Res Soc* 1979;30(2):93–104.
- Eden C. Problem construction and the influence of O.R. *Interfaces* 1982;12(2):50–60.
- Eden C, Sims D. On the nature of problems in consulting practice. *OMEGA Int J Manage Sci* 1979;7(2):119–127.
- Rosenhead J, Mingers J, editors. *Rational Analysis for a Problematic World Revisited: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley; 2001.
- Eden C, Ackermann F. *Strategy making: the journey of strategic management*. London: Sage; 1998.
- Franco LA, Montibeller G. Facilitated modelling in operational research (Invited Review). *Eur J Oper Res* 2010;205(3):489–500.
- Eden C, *et al.* The intersubjectivity of issues and issues of intersubjectivity. *J Manage Stud* 1981;18(1):37–47.
- Keeney RL. *Value-focused thinking: a path to creative decision-making*. Cambridge (MA): Harvard University Press; 1992.
- Phillips L. A theory of requisite decision models. *Acta Psychol (Amst)* 1984;56(1-3):29–48.
- Eden C, Ackermann F. Use of 'Soft OR' models by clients: what do they want from them? In: Pidd M, editor. *Systems modelling: theory and practice*. Chichester: Wiley; 2004. pp. 146–163.
- Friend J, Hickling A. *Planning under pressure: the strategic choice approach*. 3rd ed. Oxford: Elsevier; 2005.
- Phillips L. Decision conferencing. In: Edwards W, Miles R Jr, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. pp. 375–399.
- Vennix JAM. *Group model building: facilitating team learning using system dynamics*. Chichester: Wiley; 1996.
- Williams TM. *Management science in practice*. Chichester: Wiley; 2008.
- Eden C, Radford J. *Tackling strategic problems: the role of group decision support*. London: Sage; 1990.
- Eden C. A framework for thinking about group decision support systems. *Group Decis Negot* 1992;1:199–218.
- Eden C. From the playpen to the bomb site: the changing nature of management science. *OMEGA Int J Manage Sci* 1993;21(2):139–154.
- De Geus A. Planning as learning. *Harv Bus Rev* 1988;66(2):70–74.
- Schuman S, editor. *The IAF Handbook of Group Facilitation: best practices from the leading organization in facilitation*. San Francisco (CA): Jossey-Bass; 2005.
- Gallos JV, editor. *Organization development: a Jossey-Bass reader*. San Francisco (CA): Jossey-Bass; 2006.
- Eden C. The unfolding nature of group decision support: two dimensions of skill. In: Eden C, Radford J, editors. *Tackling strategic problems: the role of group decision support*. London: Sage; 1990. pp. 48–52.
- Franco LA. Forms of conversation and problem structuring methods: a conceptual development. *J Oper Res Soc* 2006;57(7):813–821.
- Phillips L, Phillips M. Facilitated work groups: theory and practice. *J Oper Res Soc* 1993;44(6):533–549.
- Ackermann F. Participant's perceptions on the role of facilitators using group decision support systems. *Group Decis Negot* 1996;5(1):93–112.
- Eden C. Problem solving or problem finishing. In: Jackson M, Keys P, editors. *New directions in management science*. Aldershot: Gower; 1986. pp. 97–107.
- Ackermann F, Eden C. Issues in computer and non-computer supported GDSSs. *Decis Support Syst* 1994;12(4,5):381–390.
- Eden C. Managing the environment as a means to managing complexity. In: Eden C,

- Radford J, editors. Tackling strategic problems: the role of group decision support. London: Sage; 1990. pp. 154–161.
29. Huxham C. On trivialities in process. In: Eden C, Radford J, editors. Tackling strategic problems: the role of group decision support. London: Sage; 1990. pp. 162–168.
 30. Hickling A. 'Decision Spaces': a scenario about designing appropriate rooms for group decision management. In: Eden C, Radford J, editors. Tackling strategic problems: the role of group decision support. London: Sage; 1990. pp. 169–177.
 31. Ackermann F. The role of computers in group decision support. In: Eden C, Radford J, editors. Tackling strategic problems: the role of group decision support. London: Sage; 1990. pp. 132–141.
 32. Phillips L. People-centred group decision support. In: Doukidis G, Land F, Miller G, editors. Knowledge-based management support systems. Chichester: Ellis-Horwood; 1989. pp. 208–224.
 33. Checkland P. Systems thinking, systems practice. Chichester: Wiley; 1981.
 34. Eden C, Ackermann F. SODA: the principles. In: Rosenhead J, Mingers J, editors. Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict. Chichester: Wiley; 2001. pp. 21–41.
 35. Schilling MS, Oeser N, Schaub C. How effective are decision analyses? Assessing decision process and group alignment effects. *Decis Anal* 2007;4(4):227–242.
 36. Checkland P. Soft systems methodology: a 30-year retrospective. Chichester: Wiley; 1999.
 37. Friend J, Hickling A. Planning under pressure: the strategic choice approach. 2nd ed. Oxford: Butterworth-Heinemann; 1997.
 38. Kaner S. Facilitator's guide to participatory decision making. San Francisco (CA): Jossey-Bass; 2007.
 39. Schwartz R. The skilled facilitator: a comprehensive resource for consultants, facilitators, managers, trainers, and coaches.. San Francisco (CA): Jossey-Bass; 2002.
 40. Bion W. Experiences in groups. London: Tavistock; 1961.
 41. Janis IL. Groupthink: psychological studies of policy decisions and fiascos. 2nd ed. Boston (MA): Houghton Mifflin Company; 1982.
 42. Kruglanski AW, *et al.* Groups as epistemic providers: need for closure and the unfolding of group-centrism. *Psychol Rev* 2006;113(1): 84–100.
 43. Quaddus MA, Atkinson DJ, Levy M. An application of decision conferencing to strategic planning for a voluntary organization. *Interfaces* 1992;22(6):61–71.
 44. Phillips L. Decision analysis for group decision support. In: Eden C, Radford J, editors. Tackling strategic problems: the role of group decision support. London: Sage; 1990. pp. 142–150.
 45. Ackermann F, Eden C. SODA and mapping in practice. In: Rosenhead J, Mingers J, editors. Rational analysis in a problematic world revisited. Chichester: Wiley; 2001. pp. 43–60.
 46. Conklin J. Dialog mapping: building shared understanding of wicked problems. Chichester: Wiley; 2006.
 47. Jackson M. Systems approaches to management. New York: Kluwer; 2000.
 48. Vennix JAM, Akkermans HA, Rouwette EAJA. Group model-building to facilitate organizational change: an exploratory study. *Syst Dyn Rev* 1996;12(1):39–58.
 49. Andersen D, Richardson G. Scripts for group model building. *Syst Dyn Rev* 1997;13(2): 107–130.
 50. Richardson G, Andersen D. Teamwork in group model building. *Syst Dyn Rev* 1995; 11(2):113–137.
 51. Eskinasi M, Rouwette EJA, Vennix JAM. Simulating urban transformation in Haaglanden, The Netherlands. *Syst Dyn Rev* 2009;25(3):182–206.
 52. Morecroft J, Sterman JD, editors. Modeling for learning organizations. Portland: Productivity Press; 1994.
 53. Senge P. The fifth discipline: the art and practice of the learning organization. New York: Doubleday; 1990.
 54. Lane D. Modelling as learning: a consultancy methodology for enhancing learning in management teams. *Eur J Oper Res* 1992;59(1):64–84.
 55. Hindle T, *et al.* Developing a methodology for multidisciplinary action research: a case study. *J Oper Res Soc* 1995;46(3):453–464.
 56. Lartindrake JM, Curran CR. All together now: the circular organization in a university hospital. *Syst Pract* 1996;9(5):391–401.
 57. Wells JSG. Discontent without focus: an analysis of nurse management and activity on a psychiatric inpatient facility using a soft

- systems approach. *J Adv Nurs* 1995;21(2): 214–221.
58. Belton V, Ackermann F, Shepherd I. Integrated support from problem structuring through to alternative evaluation using COPE and VISA. *J Multi-Criteria Decis Anal* 1997;6(3):115–130.
59. Royston G, *et al.* Using system dynamics to help develop and implement policies and programmes in health care in England. *Syst Dyn Rev* 1999;15(3):293–213.
60. Cavana RY, *et al.* Drivers of quality in health services: different worldviews of clinicians and policy managers revealed. *Syst Dyn Rev* 1999;15(3):331–340.
61. Khisty CJ. Soft systems methodology as learning and management tool. *J Urban Plann Dev* 1995;121(3):91–107.
62. Ulengin F, Topcu I. Cognitive map: KBDSS, integration in transportation planning. *J Oper Res Soc* 1997;48(11):1065–1075.
63. Brown JR, MacLeod ND. Integrating ecology into natural resource management policy. *Environ Manage* 1996;20(3):289–296.
64. Gough JD, Ward JC. Environmental decision making and lake management. *J Environ Manage* 1996;48(1):1–15.
65. Fielden D, Jacques JK. Systemic approach to energy rationalisation in Island communities. *Int J Energy Res* 1998;22(2):107–129.
66. Joubert A, Stewart TJ, Eberhard R. Evaluation of water supply augmentation and water demand management options for the City of Cape Town. *J Multi-Criteria Decis Anal* 2003;12(1):17–25.
67. Williams TM, Ackermann F, Eden C. Structuring a delay and disruption claim: an application of cause-mapping and system dynamics. *Eur J Oper Res* 2003;148(1):192.
68. Ackermann F. Modeling for litigation: mixing qualitative and quantitative approaches. *Interfaces* 1997;27(2):48–65.
69. Franco LA, Cushman M, Rosenhead J. Project review and learning in the UK construction industry: embedding a problem structuring method within a partnership context. *Eur J Oper Res* 2004;152(3): 586–601.
70. Barcus A, Montibeller G. Supporting the allocation of software development work in distributed teams with multi-criteria decision analysis. *OMEGA* 2008;36(3):464–475.
71. Winter M. Problem structuring in project management: an application of soft systems methodology (SSM). *J Oper Res Soc* 2006;57(7):802–812.
72. Shaw D, Ackermann F, Eden C. Approaches to sharing knowledge in group problem structuring. *J Oper Res Soc* 2003;54(9):936–948.
73. Montibeller G, Shaw D, Westcombe M. Using decision support systems to facilitate the social process of knowledge management. *Knowl Manage Res Pract* 2006;4(2): 125–137.
74. Eden C, Ackerman F. Mapping distinctive competencies: a systemic approach. *J Oper Res Soc* 2000;51(1):12–20.
75. O'Brien F, Meadows M. Developing a visioning methodology: visioning choices for the future of operational research. *J Oper Res Soc* 2007;58(5):557–575.
76. Salo A, Gustafsson T, Ramanathan R. Multicriteria methods for technology foresight. *J Forecast* 2003;22(2-3):235–255.
77. Howick S, Ackermann F, Andersen D. Linking event thinking with structural thinking: methods to improve client value in projects. *Syst Dyn Rev* 2006;22(2):113–140.
78. Montibeller G, Franco LA. Raising the Bar: Strategic Multi-Criteria Decision Analysis. *J Oper Res Soc*. In press.
79. Ormerod R. Putting soft OR methods to work: information systems strategy development at parabola. *OMEGA* 1998;26(1):75–98.
80. Ormerod R. Putting soft OR methods to work: information systems strategy development at Sainsbury's. *J Oper Res Soc* 1995;46(3):277–293.
81. Ormerod R. Putting soft OR methods to work: information systems development at Richards Bay. *J Oper Res Soc* 1996;47(9): 1083–1097.
82. Huxham C. Group decision support for collaboration. In: Huxham C, editor. *Creating collaborative advantage*. London: Sage; 1996. pp. 141–151.
83. Franco LA. Facilitating collaboration with problem structuring methods: a case of an inter-organisational construction partnership. *Group Decis Negot* 2008;17(4): 267–286.
84. Walsh M, Hostick T. Improving health care through community OR. *J Oper Res Soc* 2005;56(2):193–201.
85. White L. Development options for a rural community in Belize: alternative development and operational research. *Int Trans Oper Res* 1996;1(4):453–462.

86. White L, Taket A. Beyond appraisal: Participatory Appraisal of Needs and the Development of Action (PANDA). *OMEGA* 1997;25(5):523–534.
87. Hodgkin J, Belton V, Koulouri A. Supporting the intelligent MCDA user: a case study in multi-person multicriteria decision support. *Eur J Oper Res* 2005;160(1):172–189.
88. Matheson D, Matheson JE. The smart organization: creating value through strategic R & D. Boston (MA): Harvard Business School Press; 1998.
89. Montibeller G, *et al.* Structuring resource allocation decisions: a framework for building portfolio models with area-grouped projects evaluated by multiple criteria. *Eur J Oper Res* 2009;199(3):846–856.
90. Phillips L, Bana e Costa CA. Transparent prioritisation, budgeting, and resource allocation with multi-criteria decision analysis and decision conferencing. *Ann Oper Res* 2007;154(1):51–68.
91. Bana e Costa CA, *et al.* Decision support systems in action: integrated application in a multi-criteria aid process. *Eur J Oper Res* 1999;113(2):315–335.
92. Mingers J, Gill A, editors. *Multimethodology: the theory and practice of combining management science methodologies*. Chichester: Wiley; 1997.
93. Andersen D, *et al.* Using a group decision support system to add value to group model building. *Syst Dyn Rev*. In press.
94. Franco LA, Lord E. Understanding multimethodology: evaluating the perceived impact of mixing methods for group budgetary decisions. *OMEGA Int J Manage Sci*. doi:10.1016/j.omega.2010.06.008.
95. Montibeller G, *et al.* Structuring resource allocation decisions: a framework for building multi-criteria portfolio models with area-grouped projects. *Eur J Oper Res* 2009;199(3):846–856.
96. Schein EH. *Process consultation revisited: building the helping relationship*. Financial Times/Prentice Hall; 1998.
97. Mingers J, Brocklesby J. Multimethodology: towards a framework for mixing methodologies. *OMEGA* 1997;25(5):489–509.
98. Budescu DV, Wallsten TS. Consistency in interpretation of probabilistic phrases. *Organ Behav Hum Decis Process* 1985;36(3):391–405.
99. Kahneman D, Slovic P, Tversky A, editors. *Judgement under uncertainty: heuristics and biases*. Cambridge (MA): Cambridge University Press; 1982.
100. Poyhonen M, Vrolijk H, Hamalainen RP. Behavioral and procedural consequences of structural variation in value trees. *Eur J Oper Res* 2001;134:216–227.
101. Plous S. *The psychology of judgment and decision making*. New York: Mc Graw Hill; 1993.
102. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 1956;63:81–97.
103. Simon H. *Administrative behavior: a study of decision making processes in administrative organizations*. 2nd ed. New York: McMillan; 1957.
104. Ackermann F, Eden C. Contrasting single user and networked group decision support systems for strategy making. *Group Decis Negot* 2001;10(1):47–66.
105. Afordakos O, Franco LA, O'Brien F. Understanding facilitated modelling in-use: an exploratory study with non-experimental groups. In: Climaco J, Kersten G, Costa JP, editors. *Proceedings on the 2008 Group Decision and Negotiation Conference*. Coimbra: INESC Coimbra, Instituto de Engenharia de Sistemas e Computadores; 2008. pp. 95–96.
106. White L. Evaluating problem structuring methods: developing an approach to show the value and effectiveness of PSM interventions. *J Oper Res Soc* 2006;57(7):842–855.
107. Papamichail KN, *et al.* Facilitation practices in decision workshops. *J Oper Res Soc* 2007;58(5):614–632.
108. Akkermans H, Vennix J. Client's opinions on group-model building: an exploratory study. *Syst Dyn Rev* 1997;13(1):3–31.
109. Rouwette EAJA, Vennix JAM, van Mullekom T. Group model building effectiveness: a review of assessment studies. *Syst Dyn Rev* 2002;18(1):5–45.
110. Westcombe M, Franco LA, Shaw D. Where Next for PSMs - a grassroots revolution? *J Oper Res Soc* 2006;57(7):776–778.
111. Schilling MS, Oeser N, Schaub C. How Effective Are Decision Analyses? Assessing Decision Process and Group Alignment Effects. *Decis Anal* 2007;4(4):227–242.
112. Robinson S. Soft with a hard centre: discrete-event simulation in facilitation. *J Oper Res Soc* 2001;52(8):905–915.

113. Tako A, Kotiadis K, Vasilakis C. A conceptual modelling framework for stakeholder participation in simulation studies. Proceedings of the OR Society's Two-Day Simulation Workshop (SW10). Worcestershire; 2010.
114. Tako A, Kotiadis K, Vasilakis C. A participative modelling framework for developing conceptual models in healthcare simulation studies. Proceedings of the 2010 Winter Simulation Conference. Baltimore (MD); 2010.

OPERATION RESEARCH IN GOLF

WILLIAM J. HURLEY
Department of Business
Administration, Royal Military
College of Canada, Kingston,
Ontario, Canada

INTRODUCTION

According to the National Golf Foundation [1], the golf industry generated \$76 billion in direct expenditures on goods and services in the United States in 2005. This same report estimates the economic impact to be in excess of \$195 billion. Golf facility revenue (green fees, memberships, etc.) alone generated some \$28 billion in revenue. By comparison, the revenue generated from all spectator sports, including baseball, basketball, football, and hockey was \$24.4 billion. All this is to say that golf is big business. And given its growth in popularity over the last several decades, it is likely to continue. Given these expenditures and the interest, it is appropriate to ask whether there is a role for operations research (OR). Consequently, I describe how OR has been applied and opportunities for its further use.

HOGAN

Without a doubt, golf's first OR scientist was the American virtuoso, Ben Hogan. Hogan is on just about everybody's top-five all-time list, a list that would include Tiger Woods, Jack Nicklaus, and Bobby Jones. But Hogan was different. Whereas Woods, Nicklaus, and Jones were child prodigies, with blue-chip amateur records accumulated in their teens (Jones was an amateur his whole career), Hogan had little success as an amateur and only started to win consistently on the PGA Tour when he reached the age of 30. In fact, Hogan had to leave the Tour three times because he ran out of money. Yet, in time, he became one of the best players of all time.

One of the measures of his greatness is what other great players thought of him. When Jack Nicklaus was asked about his dream foursome, he responded that it would be Bobby Jones, Walter Hagen, and Hogan—Jones and Hagen because he had never played with them and Hogan because “I’d want him as my partner.” Tiger Woods’s foursome is Jones, Nicklaus, and Hogan. When Nicklaus was asked whether Woods was the best ball-striker he had seen, he responded “No, no—Ben Hogan easily.” (*Golf Digest*, 2004).

In my opinion, Hogan’s triumph was the result of a keen scientific intellect (even though his formal education in science was limited to whatever he got in high school) and an almost autistic preoccupation with trying to discover how to hit a golf ball properly. His practice sessions were legendary—4 hours in the morning and 4 hours in the afternoon hitting approximately three balls per minute. He discusses the role of practice in his famous instructional book, “Five Lessons: the Modern Fundamentals of Golf” [2]:

How then do you build a swing that you can depend on to repeat in all kinds of wind and weather, under all kinds of presses and pressure? Having devoted the bulk of my waking hours (and a few of my sleeping hours) for a quarter of a century to the pursuit of the answer, I now believe that what I have learned can be of tremendous assistance to all golfers. That is my reason for undertaking this series of lessons. I do not propose to deal in theory. What I have learned I have learned by laborious trial and error, watching a good player do something that looked right to me, stumbling across something that felt right to me, experimenting with that something to see if it helped or it hindered, adopting it if it helped, refining it sometimes, discarding it if it didn’t help, sometimes discarding it later if it proved undependable in competition, experimenting continually with new ideas and old ideas and all manner of variations until I arrived at a set of fundamentals that appeared to me to be right because they accomplished a very definite purpose, a set of fundamentals which proved to me they were

right because they stood up and produced under all kinds of pressure. To put it briefly, the information I will be presenting is a sifting of the knowledge I've tried to acquire since I first met up with golf at the age of 12 and knew, almost immediately, I wanted to make the game my lifework.

If Hogan's approach is not traditional OR, I am not sure what is. There was a continual experimentation with his swing until he was satisfied he had found a set of fundamentals that would stand the test of the PGA Tour.

There is no question that Hogan discovered a number of principles that modern golfers employ. Here is an example. I invite the reader to look at the following YouTube video clips of the golf swings of Bobby Jones and Tiger Woods, two swings from quite different eras:

Jones: <http://www.youtube.com/watch?v=d5XeIFcLyGw&feature=related>

Woods: <http://www.youtube.com/watch?v=cHcP6X7dEUo>

It is clear from Hogan's book and some of the stories of him instructing younger players that he felt that the lower body had to be as "quiet" as possible on the backswing, that part of the swing where you take the club back to the point where you are going to begin your downswing. So first look at Woods's swing. You will note that his right knee is absolutely fixed over his complete backswing. It only changes its position through the downswing as he shifts his weight from his right side to his left. His lower body just does not move much over the backswing. Now look at Jones's swing. Note that his lower body is much more active than Woods'. The two swings are like day and night.

I would ask the reader to do a further experiment. Try swinging a golf club keeping your lower body fixed and rotating only from the waist up. On the backswing, you will note that the muscles of your lower back and upper legs—the engine of the golf swing—really feel like they are coiling and ready to spring once the downswing begins. In my opinion, this quiet lower body does two things. It helps you to hit the ball straighter and generates more power to hit it further. This, then, is one

of the swing principles that Hogan brought to the game as a result of his experimentation.

In a 1955 *Life Magazine* article, published after his competitive career was essentially finished, Hogan supposedly revealed his "secret," the change he had made in his swing that allowed him to win more consistently. In that article, Hogan described these changes to be a weaker grip (his left hand rotated slightly to the left) and the position of his wrists at the top of his backswing (his wrists were pronated). With these adjustments, he argued, he was able to hit his signature power fade (a shot that moves slightly left to right for a right-handed golfer). But there remains some doubt that this was his real secret. In fact, those who knew him well suggested that his wisdom "I found it in the dirt," was closer to the truth. Of course, he was suggesting that a golfer needed to go to the practice tee and stay there until he or she found a swing that worked. In other words, Hogan felt that what we would call an OR approach—experimentation to see what works and what does not—was at the heart of playing better golf.

OPERATIONS RESEARCH AND THE PROBLEM OF GETTING BETTER

There has been significant research published on the nature of the golf swing and how to improve one's game. The vast majority of this has been in the field of biomechanics. The interested reader is referred to the collections in Refs 3–6. These volumes are essentially the proceedings of the four World Scientific Congress of Golf conferences. Representative titles include "Back Pain in Novice Golfers, A One-year Follow-up," "The Biomechanics of the Shoe-ground Interaction," and "Spine and Hip Motion Analysis During the Golf Swing." However, there are a few articles that are pure OR and I consider some of these in each of the following sections.

Putting

The renowned shortgame instructor, Dave Pelz [7], presents some convincing evidence that most golfers do not play enough break in their putts. They consistently hit putts

that end up below the hole on both left- and right-breaking putts. Pelz posits that this is a result of a poor read of the line and his advice is to assess the break of the putt from a position behind the ball and along the aim line (the direction in which the putt has to be hit in for it to just fall into the hole).

When assessing how hard to hit a putt, there is an old aphorism “Never up, never in.” This means that the putt has to be hit with enough speed so that it passes the hole if it misses, otherwise it has no chance of going in. But if you hit it too far past the hole, the next putt is more difficult and a three-putt is a possibility. Pelz suggests that, when a golfer is trying to maximize the probability of a one-putt, all putts should be hit so that they stop no more than 2 ft past the hole. Hoadley [8] develops a slightly different rule. On the basis of a Markov chain analysis for putting on a flat green, he suggests the heuristic rule that the distance past the hole should be 2 ft times the probability of making the putt. This advice makes sense. For very short putts, the rule really does not matter; for longer putts, it suggests that you try to get the ball just slightly past the hole. For instance, if you have a 30-ft putt, it is unlikely that you will make it. Hence, the aim should be to get it just by the hole and do no worse than a two-putt. As my colleague Mark Broadie has pointed out, one of the difficulties with his analysis is that his parameters are suspect. For instance, for PGA Tour professionals, he assumes that one standard deviation of distance error from 60 ft is about 13 ft. This is quite high. Nonetheless, the form of his rule makes sense.

The most sophisticated treatment of putting I have seen is the model developed by Bansal and Broadie [9]. Their model is based on a characterization of putter ability and a physics model based on putt trajectory and holeout considerations. The optimal strategy is the solution of a difficult dynamic stochastic programming problem. They use simulation to solve for optimal strategies for specific parameter sets. The resulting strategies are roughly equivalent in form to those produced by Hoadley.

The Long Game

If you have time to practise, how should you allocate it? The aphorism “Drive for show, putt for dough” would suggest that the majority should be spent on your short game (putting, chipping, and sand play) rather than the long game (woods and irons). What does the evidence suggest?

Berry [10] looked at PGA Tour performance statistics. He regressed a PGA Tour professional’s score on a number of factors, including driving ability (both length and accuracy), putting ability, sand save percentage, chipping ability, and iron play. He found that the factor that made the largest contribution to lowering scores was putting ability, followed closely by driving ability. Consequently, he rephrased the dictum to “You drive for dough and putt for more dough.” While this evidence is suggestive, we would need to understand the relative productivities of an hour spent on the driving range and the practice green before we could say for sure. Moreover, the data his analysis is based on are for Tour professionals, not the average amateur golfer.

There has been an interesting line of research beginning with Cochran and Stobbs [11] and continued by Landsberger [12], which attempts to assign a value to each shot in a round of golf. The most recent work is by Broadie [13]. He first defines the idea of *fractional par*, the average number of strokes that a scratch golfer needs to complete a hole. By way of example, a 140-yard par 3 hole might have a fractional par of 3.2. Suppose that you have 200 yards left into a par 4 hole. Then the fractional par might be 3.6. He defines *shot value*, v , to be the fractional par value before the shot, f_s , less the fractional value after the shot, f_e , less 1, or

$$v = f_s - f_e - 1. \quad (1)$$

As Broadie points out, this formula must be adjusted for penalty strokes. For instance suppose on a par 4 with a fractional par of 4.4, you hit your drive out of bounds. Then under the rules of golf you must hit another ball off the tee and incur a penalty so the

value of the first shot is

$$v = 4.4 - 4.4 - 2 = -2. \quad (2)$$

Here is another example. Suppose that the fractional par before you hit your second shot into a par 4 hole is 3.6 and you hit the ball to 10 feet and at that point your fractional par is 1.7. Then the value of the shot into the green is

$$v = 3.6 - 1.7 - 1 = 0.9. \quad (3)$$

We interpret this as 0.9 strokes better than what the average scratch golfer would do with the same shot.

Broadie argues that this shot value measure could be used in a variety of ways. One could use it to measure a golfer's *consistency* in the following way. We could classify shots into three types: *awful* shots, *typical* shots, and *great* shots. Awful shots have a value less than -0.8 ; typical shots are between -0.8 and $+0.8$; and great shots have a value exceeding 0.8 . A golfer could classify each shot over a number of rounds and in the process develop a picture of where he or she was falling down. For instance, it might be that the majority of awful shots are the result of poor putting. This would indicate that the golfer should devote relatively more practice time to putting.

I think Broadie's approach has considerable merit. He has developed a software program, *Golfmetrics*, which enables a golfer to record shots on six golf courses. Over time, a golfer could input individual shots from each round and from these data determine what he/she ought to practise to improve. By way of example, he analyzed shots from a group of players who normally shot between 84 and 97. His complete data set for this group and four others included 40,000 shots over 500 rounds of golf from over 130 golfers. Sample average shot values for four types of shots are shown in the following table:

<i>Putting</i>	-2.3
<i>Long game</i>	-9.3
<i>Short game</i>	-3.4
<i>Sand game</i>	-0.6
Total	-15.6

In total, these golfers are some 15.6 shots worse than a scratch golfer. Where they loose the most is in the long game (-9.3) and short game (-3.4). This would suggest that such golfers should be spending relatively more practice time on their long games but, again, it really depends on the marginal productivities of long game practice relative to those in the other areas.

OPERATIONS RESEARCH AND THE HANDICAP PROBLEM

One of the fundamental problems in golf competition, whether a friendly Nassau or an amateur tournament, is the setting of handicaps for individual golfers so that a competition is fair. For instance, if one golfer regularly shoots 75 and another shoots 85, how many strokes should the better player give to the other so that the competition is fair? This problem has received a lot of attention over the last 40 years. The interested reader is referred to papers by Bingham and Swartz [14], Pollock [15,16], Scheid [17-19], and Swartz [20].

In golf, there are two basic kinds of competitions: *match play* and *medal play*. Match play is a competition between two players where a particular hole is either tied, won, or lost. The golfer winning the most holes over 18 is then the winner of the match. With medal play, the winner of the competition is the competitor taking the fewest strokes. Most amateur tournaments tend to be medal play. To make these tournaments fair, a golfer's course-specific handicap is subtracted from his gross score to arrive at a net score and typically prizes are given on the basis of both net and gross scores.

The handicap systems in the United States

and Canada are essentially the same except for some minor differences in equitable stroke control—the term given to the maximum a golfer can count on any given hole for the purposes of computing a handicap. The systems of the Royal and Ancient and Australia are fundamentally different than those used in North America.

Under current North American handicapping systems, Bingham and Swartz [14] have shown that stronger golfers have the advantage in match play. Yet in handicapped tournaments with many players, weaker players have the advantage. So the question is how you alter the current systems to make them more equitable.

As Swartz [20] argues, golf scores are, by their nature, random and fluctuate around a golfer's true ability. Consequently, the problem of constructing equitable handicaps is, at its heart, a statistical problem. Here is roughly the way the system works in Canada. For each round, a differential is calculated according to

$$D = 113 \left(\frac{X - R}{S} \right), \quad (4)$$

where X is the golfer's score (adjusted for equitable stroke control), R is the course rating, and S is the course slope rating. Typically the rating, R , is what a scratch golfer would score on the course and S is an additional factor that helps measure course difficulty. Next a factor, F , is calculated by looking at the best 10 values of the golfer's last 20 differentials. Let the last 20 differentials be $D_1, D_2, \dots, D_{20}$ and let the 10 best (smallest) be represented by

$$D_{(1)}, D_{(2)}, \dots, D_{(10)}. \quad (5)$$

Then the factor is given by

$$F = 0.096 (D_{(1)} + D_{(2)} + \dots + D_{(10)}). \quad (6)$$

To get the course handicap, we calculate

$$FS/113 \quad (7)$$

and round it to the nearest integer.

This system has some obvious flaws from a statistical perspective. For instance, why

calculate the handicap on only the best 10 differentials? On statistical grounds, such a calculation ignores information and is not likely to be as efficient as one that uses more differentials. Swartz has suggested the following change. Rather than the factor, F , he calculates what he terms a mean, $\hat{\mu}$, given by

$$\hat{\mu} = \sum_{i=1}^{16} w_i D_{(i)}, \quad (8)$$

where $w_1, w_2, \dots, w_{16}$ are appropriately chosen weights.

To get at a good set of weights, suppose that a golfer's gross scores are normally distributed as suggested by Scheid [18] and Bingham and Swartz (1990). If this is the case, then

$$113 \left(\frac{Y - R}{S} \right) \sim \text{Normal}(\mu, \sigma^2), \quad (9)$$

where Y is the gross score, and μ and σ^2 are parameters that characterize a golfer's skill level and consistency, respectively. If the weights, $w_1, w_2, \dots, w_{16}$, are chosen properly, then not only is $\hat{\mu}$ unbiased ($E(\hat{\mu}) = \mu$), but it is also BLUE (best linear unbiased estimator), which is a very interesting result. These are nice properties and Swartz goes on to show that his new system substituting $\hat{\mu}$ for F results in a fairer system.

One of Swartz's difficulties in the design of a new system is that he is working with the Royal Canadian Golf Association and it is their wish that the new system not be too different than the old. Given the normal ups and downs in individual golf games over time, let us suppose that a handicap based on 20 scores is reasonable. Why then does he choose the best 16 rather than all 20 or some other form of truncation, say removing the three best and three worst scores? In a personal communication to me, Swartz explained that amateur golf scores are actually slightly skewed to the right. This is interesting by itself because it is certainly consistent with sandbagging (the claim that some golfers either do not register some of their good scores or inflate some of their actual scores). In this regard, it is interesting

to note that Grober [21] finds that PGA Tour scores are not skewed to the right. Swartz goes on to explain that “the censored data has better fit as a censored normal than the full sample has as a normal. I found that 16 gave roughly the best fit.”¹

As Swartz and others have pointed out, one of the difficulties with the handicapping system is that handicaps are designed to make medal play tournaments fair on the basis of net scores. So how do they work for match play? Swartz shows that his new system works fairly well for match play as well. But there are many other kinds of competitions where handicaps are used to make the game fair. One popular game at most courses is a friendly net bestball match within a regular foursome. At the course I play at, we normally match the best player with the worst in one team. With the exception of the best player, the other three players are given a number of strokes equal to 80% of the difference between their handicaps and the best player’s handicap. This arrangement has evolved over time and, in my judgment, works quite well. It would be interesting to see why it works so well.

OPERATIONS RESEARCH AND GOLF COURSE OPERATION

A number of researchers have looked at the problems associated with golf course operations. Shmanske [22] has written an interesting book that takes an economic approach to some of these problems. Representative OR work would include Refs [23–25]. These papers employ a process simulation to study the relationship among: the interval between tee-off times at the first hole; the duration of a round; and the throughput of a given course over a typical day.

The simulation model developed in Ref. 23 offers some interesting results. For the course they study, they show that there is only a slight decrease in the number of rounds played over a day if the tee time interval

at the first hole is raised from 6 to 10 min. But this increase in the tee time interval decreases the length of the average round by over an hour. This is a substantial decrease and would surely increase the quality of a round. This result also demonstrates that the tee time interval at the first hole is an important consideration for golf course managers.

In addition, Tiger *et al.* [23] look at where the bottlenecks are on the golf course they studied. The biggest bottlenecks occurred on two par 3 holes. They do not give the characteristics of these holes but I suspect they are difficult. At my course, we have two difficult par 3s and one is considered one of the toughest holes in Ontario. Invariably, we have to wait at both these holes. It seems to me that golf course designers need to look very carefully at where difficult par 3s are placed on a course and what the rules are for playing them. At some courses, players reaching the green on a difficult par 3 are asked to waive the players behind to hit their tee shots before the players on the green putt out. In my view, this is a good rule as long as the players at the green can move to a safe place. These simulation models can also be used to look at other rules such as compulsory cart rules to determine their effect on speed of play and throughput.

OPERATIONS RESEARCH AND STRATEGY FOR GOLF TEAM COMPETITIONS

Operations researchers have always had a keen interest in team strategies in sports competitions. Golf is no exception. Mulvey *et al.* [26] focus on College team selection in the United States. I have looked at some of the decisions that a Ryder Cup captain has to make [27–29]. There are some who suggest that a captain’s decisions are not that important. For instance, Jaime Diaz [30] argued that being Ryder Cup captain was “the most over-rated job in sports.” And some of the captains have felt this way as well. Jack Nicklaus, arguably the greatest player ever and Ryder Cup captain twice, said this [30]: “I don’t think the captain’s role really affects the result very much.” On another

¹Personal communication by e-mail dated March 19, 2009.

occasion he remarked that his job required making a few speeches and making sure his players had fresh towels, sunscreen, and tees [31]. Regardless of their importance, some of these decisions are quite interesting OR problems. So in this section I sketch a few of these and describe how OR can help.

The Ryder Cup is a very prestigious golf competition played between two teams representing the United States and Europe. Since it is such a great honor for a golfer to be selected to one of these teams, both teams comprise the best European and American golfers. The competition is held once every two years at sites that alternate between the United States and Europe. It is a three-day competition. On the first two days, the teams play eight “foursomes” and eight “four-ball” matches. These are two-man team match-play competitions. In four-ball matches (equivalently bestball matches), each player from a team plays his own ball and the score for a team on a hole is the lowest of their two scores. In foursomes matches the players from a team alternate shots. On the final day of competition, 12 singles matches are played. Each team captain submits an ordered list of golfers (a *slate*), and the matches are determined from these slates: USA1 versus Europe1, USA2 versus Europe2, and so on.

In my view, some of the interesting decisions are as follows:

1. *Which Golfers Should Be in the Team?* For the 2008 Ryder Cup, US captain Paul Azinger had four captain’s picks rather than the two that previous captains had. While he did not say too much about how these selections were made, it is clear that several were made just before the formal announcement.
2. *How Should a Captain Select the Slate for the Final Day of Competition?* In recent years, much has been made of front-end loading—putting the best players in the team out early in the Sunday singles matches. The question is whether this is the right thing to do.
3. *How Should a Captain Select Four-Ball and Foursomes Teams?* In the 2004 matches, US captain Hal Sutton paired

Tiger Woods and Phil Mickelson, arguably his two best players. Even if these two liked one another, the question is whether it was the right thing to do.

So how can we use OR to help answer these questions?

Ryder Cup Team Selection

Up to the 2008 Ryder Cup, US team selections were based on points accumulated for the two years prior to the competition. The top 10 players on this list were automatically in the team. The remaining two players—the wild cards—were selected by the captain. For the most recent Ryder Cup, US team captain Paul Azinger convinced the PGA that the automatic selections should be based on only 2008 PGA Tour results and that only eight of these selections would be automatic. This gave Azinger four wild-card selections. The strategy was clearly to select as many players as possible who were playing well at the time of the Ryder Cup. Azinger never discussed the precise details of how the wild-card selections were made. However, it is clear that the decision was not made until just before they had to be made.

There has been some interesting statistical research on characterizing the performance of PGA Tour players. This would include the work of Berry [32], Puterman and Wittman [33], McHale and Forrest [34], and Connolly and Rendleman [35]. I think the Connolly and Rendleman approach would be quite useful in the identification of Ryder Cup wild cards. They propose a method that breaks down an individual golfer’s scores into skill-based and unexpected components as a function of time using the smoothing spline model of Wang [36]. They find evidence of significant positive first-order autocorrelation in their error structure for 10% of the golfers in their sample. Hence, they provide statistical proof that there is method in Azinger’s madness of trying to get the players that are “hot” at the time the Ryder Cup is played. This result is also confirmed by the work of McHale and Forrest [34]. Moreover, Connolly and Rendleman’s estimates of the time-varying

skill for the players under consideration might be very useful to future Ryder Cup captains.

Choosing the Order of Play for the Singles Matches

Consider a simple variation of the actual problem. Two teams, the United States and Europe, have only two players: a very good golfer (V) and a good golfer (G). The probabilities for the US players winning individual matches are specified in the following table:

		Europe	
		Very Good (V)	Good (G)
US	Very Good (V)	$1/2$	p
	Good (G)	$1 - p$	$1/2$

We choose p to be a number bigger than $1/2$ since a very good golfer is assumed to have a better than even chance to beat a good golfer. Note that these probabilities are symmetric in the following sense. The US V golfer has a 50–50 chance against his European counterpart and a probability $p > 0.5$ of beating the European G golfer. This same property holds for the European V Golfer. By extension, both G golfers have a similar symmetric property.

Suppose that the Americans have not had a very good weekend so far, and to win the Ryder Cup, they must win both matches. We assume that no match can end in a tie. How should each captain choose his slate?

Each captain has two pure strategies: (V, G) and (G, V) The payoff matrix is shown below:

		Europe	
		(V,G)	(G,V)
US	(V,G)	$1/4$	$p(1 - p)$
	(G,V)	$p(1 - p)$	$1/4$

Each element of the table is the probability that the United States wins the Ryder Cup for the particular pair of pure strategies.

This type of game is known as a *coordination game*. Since $p(1 - p)$ is always less than $1/4$ for $p > 0.5$, the United States would like to pick (V,G) if Europe picks (V,G) and (G,V) if Europe picks (G,V). It is in the United State's best interest to have competitive matches; it is in Europe's best interest to have mismatches. Therefore, a side will have an advantage if it can somehow discover what strategy the other will use. Game theory suggests that it is in each side's best interest to keep its choice of slate secret.

Given this payoff matrix, the optimal mixed strategy for each side is to use a coin to pick a slate. That is, each captain would choose (V,G) with probability $1/2$ and (G,V) with probability $1/2$. Equivalently each could draw names out of a hat. In Hurley [29] I show that this strategy is optimal under more general conditions.² The critical assumption giving this result is that golfers' individual probabilities of winning a match are unaffected by the order of the matches. This assumption is discussed more fully in Ref. 28.

Constructing Bestball Teams

In Ref. 27, I look at the problem of how to construct four bestball teams from eight golfers of heterogeneous ability assuming the other side has the same eight golfers. I distinguish two strategies. The *Unbalanced* strategy puts together the pairings (1,2), (3,4), (5,6), and (7,8) where 1 is the best of the eight golfers and 8 is the worst; the *Balanced* strategy forms (1,8), (2,7), (3,6), and (4,5). The article suggests that the *Balanced* strategy is the best. But what is surprising is that the advantage is relatively small. So if a Ryder Cup captain felt that some of his *Balanced* pairings did not offer good "chemistry", and therefore moved away from the *Balanced* strategy, the cost to his side would be small at best.

²My colleague Hamdy Taha has pointed out an error with the proof of the general proposition given in the appendix of Ref. 29. It has been repaired and the revised proof is available from the author by request.

SOME INTERESTING PROBABILITIES

There are a number of interesting probability calculations that we could make:

1. *The Probability of the Hogan Grail.* James Dodson [37], in his biography of Hogan, tells this story. Jimmy Demaret, a fellow professional, was leaving the club house at Oak Hill after a tournament round in the mid-fifties and saw Hogan on the range, illuminated by car head-lights, hitting shots into the night. Demaret approached Hogan and remarked that he did not understand why he was practising when he had shot 64 that day with 10 birdies. Hogan responded: "Jimmy, if a man can make ten birdies in a round, there's no reason in this world why he can't birdie every hole." What is the chance, then, that a PGA Tour golfer could have a round with 18 birdies?
2. *The Probability of Shooting 59.* The Hogan Grail aside, there have been some remarkable rounds of golf. One standard is a round of 59. At the professional level, this was first accomplished by Al Geiberger on June 10, 1977, at the Danny Thomas Memphis Classic at the Colonial Country Club. Chip Beck and David Duval have also shot 59. What is the chance that a PGA Tour player would shoot 59?
3. *The Probability of Some of Tiger Woods's Winning Margins.* In 2000, Tiger Woods had a remarkable summer by any measure. He was the first golfer since Hogan in 1953 to hold three of the four major championships. He won the US Open by a remarkable 15 shots and then the British Open by 8 shots. In the modern era, these winning margins are not even close to the normal margin of 1 or 2 shots. Consequently, it would be interesting to examine what these margins tell us about Woods' abilities.
4. *The Probability That Byron Nelson's Mark of 11 Successive PGA Tour Wins Will Be broken.* In 1945, Byron Nelson won 11 consecutive PGA Tour

events. This is a remarkable feat by today's standards. The interested reader is referred to Grober [21] for the calculation.

5. *The Probability of the Miracle at Brookline.* After the first two days of the 1999 Ryder Cup matches, the US team was down 10 points to 6 to the European side. The United States won 8 1/2 of a possible 12 points over the Sunday singles matches to claim the Cup. This was the largest come-from-behind victory in Ryder Cup history. What, then, is the chance that this comeback happens assuming the teams were identical?

In unpublished work, I have documented the calculation of each of these probabilities with the exception of 4. By way of example, suppose that we calculate the probability of the Hogan Grail. Let us suppose that a particular course comprises four par 3s, 10 par 4s, and four par 5s. Let us consider Tiger Woods. On the basis of 2006 PGA Tour statistics, Woods' chances of birdies on each type of hole are estimated as follows:

Type	Probability of a Birdie
Par 3	0.1490
Par 4	0.2179
Par 5	0.5796

Woods' probability of a birdie on a par 5 hole is the sum of his frequency getting either a birdie (0.5341) or an eagle (0.0455). Consequently, his probability of at least birdieing every hole is

$$(0.1490)^4(0.2179)^{10}(0.5796)^4 = 10^{-11}. \quad (10)$$

If there are roughly 20,000 PGA Tour rounds played per year, my estimate is that the Hogan Grail should occur roughly once every 3.5 million years. In the pantheon of sports records, we are much more likely to see a 57-game hitting streak in baseball.

Acknowledgments

I would like to thank Mark Broadie and Tim Swartz for their very helpful comments on an earlier draft of this article.

REFERENCES

1. National Golf Foundation. Golf industry report. Volume 7, 2007.
2. Hogan B, Wind HW. Five lessons: the modern fundamentals of golf. New York: Golf Digest Classic Books; 1957.
3. Cochran AJ. Science and Golf: Proceedings of the First World Scientific Congress of Golf. London: E & FN Spon; 1990.
4. Cochran AJ, Farrally MR. Science and Golf II: Proceedings of the World Scientific Congress of Golf. London: E & FN Spon; 1994.
5. Farrally MR, Cochran AJ. Science and Golf III: Proceedings of the World Scientific Congress of Golf. London: Human Kinetics; 1998.
6. Thain E. Science and Golf IV: Proceedings of the World Scientific Congress of Golf. London: Routledge; 2002.
7. Pelz D. A study of golfers' abilities to read greens. In: Cochran AJ, Farrally MR, editors. Science and Golf II: Proceedings of the World Scientific Congress of Golf. London: E & FN Spon; 1994. pp. 180–184.
8. Hoadley B. A study of golfers' abilities to read greens. In: Cochran AJ, Farrally MR, editors. Science and Golf II: Proceedings of the World Scientific Congress of Golf. London: E & FN Spon; 1994. pp. 185–192.
9. Bansal M, Broadie M. A simulation model to analyze the impact of hole size on putting in golf. In: Mason SJ, Hill RR, Monch L, editors. Proceedings of the 2008 Winter Simulation Conference. Los Alamitos (CA): IEEE; 2008.
10. Berry SM. Drive for show and putt for dough. *Chance* 1999;12(4):50–55.
11. Cochran A, Stobbs J. The search for the perfect swing. Philadelphia (PA): J. B. Lippincott Company; 1968.
12. Landsberger LM. A unified golf stroke value scale for quantitative stroke-by-stroke assessment. In: Cochran AJ, Farrally MR, editors. Science and Golf II: Proceedings of the World Scientific Congress of Golf. London: E & FN Spon; 1994. pp. 216–221.
13. Broadie M. Assessing golfer performance using golfmetrics. In: Crews D, Lutz R, editors. Science and Golf: Proceedings of the 2008 World Scientific Congress of Golf. Mesa (AZ): Energy in Motion Inc.; 2008. pp. 253–262.
14. Bingham DR, Swartz TB. Equitable handicapping in golf. *Am Stat* 2000;54(3):170–177.
15. Pollock SM. A model for the evaluation of golf handicapping. *Oper Res* 1974;22(5):1040–1050.
16. Pollock SM. A model of the USGA handicap system and fairness of match and medal play. In: Ladany SP, Machol RE, editors. Optimal strategies in sports. Amsterdam: North Holland; 1977. pp. 141–150.
17. Scheid F. An evaluation of the handicap system of the United States golf association. In: Ladany SP, Machol RE, editors. Optimal strategies in sports. Amsterdam: North Holland; 1977. pp. 151–155.
18. Scheid F. Normality and independence of golf scores, with applications. In: Cochran AJ, editor. Science and Golf: Proceedings of the First World Scientific Congress of Golf. London: E & FN Spon; 1990. pp. 147–152.
19. Scheid F. A general principle in golf. In: Thain E, editor. Science and Golf IV: Proceedings of the World Scientific Congress of Golf. London: Routledge; 2002.
20. Swartz TB. A new handicapping system for golf. *J Quant Anal sports* 2009;5(2): Article 9. DOI: 10.2202/1559-0410.1168. Available at <http://www.bepress.com/jqas/vol5/iss2/9>.
21. Grober RD. PGA tour scores as a Gaussian random variable. Available at arXiv:0802.3890v1 [stat.AP]. Cornell University Library; 2008. Accessed 2009 Mar 19.
22. Shmanske S. Golfonomics. New Jersey: World Scientific Publishing; 2004.
23. Tiger AA, Speers J, Simpson C, *et al.* Using simulation modeling to develop a golf course flow DSS. *Issues Infor Syst* 2003;4:720–726.
24. Tiger AA, Salzer D. Daily play at a golf course: using spreadsheet simulation to identify system constraints. *INFORMS Trans Educ* 2004;4(2). Available at <http://ite.pubs.informs.org/Vol4No2/TigerSalzer>.
25. Kimes SE, Schruben L. Golf course revenue management: a study of tee time intervals. *J Revenue Pricing Manage* 2002;1(2):111–120.
26. Mulvey JM, Green W, Traub C. Team selection by portfolio optimization. In: Thain E, editor. Science and Golf IV: Proceedings of the World Scientific Congress of Golf. London: Routledge; 2002. pp. 305–315.

27. Hurley WJ. The Ryder cup: are balanced four-ball pairings optimal? *J Quant Anal Sports* 2008a;3(4), Article 6. Available at <http://www.bepress.com/jqas/vol3/iss4/6>.
28. Hurley WJ. The 2001 Ryder cup: was strange's strategy to put Tiger Woods in the anchor match a good one? *Decis Anal* 2008b;4(1):41–45.
29. Hurley WJ. How should team captains order golfers on the final day of the Ryder Cup matches? *Interfaces* 2002;32(2):74–77.
30. Diaz J. The most over-rated job in sports. *Golf Digest* 2004. Available at www.golfdigest.com. Accessed 2009.
31. Newport JP. Team USA's management victory. *Wall St J* 2008 Sep 27.
32. Berry SM. How ferocious is Tiger? *Chance* 2001;14(3):51–56.
33. Puterman ML, Wittman SM. Match play: using statistical methods to categorize PGA tour players' careers. *J Quant Anal Sports* 2009;5(1). Article 11. Available at <http://www.bepress.com/jqas/vol5/iss1/11>.
34. McHale I, Forrest DK. The importance of recent scores in a forecasting model for professional golf tournaments. *J Manage Math* 2005;16(2):131–140.
35. Connolly RA, Rendleman RJ Jr. Skill, luck, and streaky play on the PGA tour. *J Am Stat Assoc* 2008;103:74–88.
36. Wang Y. Smoothing spline models with correlated random errors. *J Am Stat Assoc* 1998;93:341–348.
37. Dodson J. Ben Hogan: an American life. New York: Broadway Books; 2004.

OPERATIONAL RESEARCH SOCIETY OF INDIA

BANI K. SINHA
Indian Institute of Management
Calcutta, Calcutta, India

HISTORY

The Operational Research Society of India (ORSI) is a national forum set up with the objective of promoting the education and applications of operational research (OR) in business, industry, and other organizations, and also to improve the quality of human lifestyle. It was established in 1957 with the late Prof. P. C. Mahalanobis, the then Deputy Chairman of Planning Commission, Government of India and founder of Indian Statistical Institute, as the founder President of the Society. The Society is one of the earliest members of the International Federation of Operational Research Societies (IFORS), having joined it in 1960, the second year of formation of IFORS. It is the first among the developing countries to become a member of IFORS. The affiliation continues. It is also an active member of the Asian Pacific Operational Research Societies (APORS).

Since inception, the Society has grown in both stature and strength. In the 1950s, the number of OR scientists in the country was limited. Chapters of the Society are being setup from time to time throughout the country with the growth of membership. The headquarters of the Society is located in Kolkata, with 11 chapters spread throughout the country. The number of members was 777 as of June 30, 2009.

NOTEWORTHY ACCOMPLISHMENTS

India lacked facilities for formal education in OR during the early 1960s. The Society introduced a course in 1973, the two-year

graduate program in OR, following the distance-learning mode. The course is recognized by the Government of India. Examinations are held in the month of May every year.

The Society holds annual conventions when practitioners and academicians in OR and allied fields from industry as well as educational institutes in the country and abroad take part, exchange knowledge and deliberate on the latest developments. The themes of the conventions are listed in the appendix. Besides, the Society organized the following international conferences:

- International Conference on Transportation in New Delhi;
- The IFORS, the sponsor of International Conference on Operational Research for Development (ICORD), has been interested in the role of OR in developing countries and for development in general. ICORD formally started in 1992. The aim of this conference has been to provide a forum for discussion on the role, relevance, and appropriate approach of OR for development and in developing countries. There are certain basic approaches to organizing the specialized conference ICORD. These, amongst others, are to ensure and encourage OR workers from developing countries to participate, allowing adequate time and environment for meaningful discussion, developing appropriate agenda and action plan for furthering the cause of ICORD, and the like;
- ICORD-I—The first ICORD was held in at the Indian Institute of Management in Ahmedabad (IIMA), India, from December 14–16, 1992. Though IFORS, ORS of Britain, and ORSI lent their names to the conference; no support in any tangible form came from them. Major financial support came from the Tata Iron & Steel Co. Ltd. The Council of Scientific and Industrial Research of

Government of India also supported the conference financially. IIMA, the venue of the conference, provided infrastructural support to the conference. The Commonwealth Secretariat, London, provided travel and other support to some of the participants toward attending the conference and also toward organizing a specialized workshop for these participants. A position paper titled "Ahmedabad Declaration" was produced on the basis of the deliberations at the conference. This formed the basis of various initiatives on the part of IFORS toward OR for development. In addition, after an editorial review process, a few selected articles from the conference were published in the form of a book titled "Operational Research for Development," edited by Jonathan Rosenhead and Arabinda Tripathy.

Approximately, 100 participants from nearly 20 countries attended the conference. The countries represented included Australia, Bangladesh, Brazil, Canada, Greece, Iran, Ireland, India, Kenya, Malaysia, Nigeria, Peru, Singapore, Sri Lanka, South Africa, United Kingdom, United States of America, and Venezuela.

- ICORD-V was also held in Jamshedpur, India, from December 19–21, 2005.
- ORSI is the founder member of the APORS under IFORS. The regional conference APORS 2003 was held in India during December 8–11, 2003. Dr. M.C. Puri from the Department of Mathematics, Indian Institute of Technology, Delhi, India was the Guest Editor for the feature issue on "OR and its Applications" from APORS 2003.
- The latest APORS conference was held in India, for the third time, from December 6–9, 2009. The theme of this conference was "Operations Research for Emerging Economies." The Interface Description Language (IDL) Lecture was delivered by Dr. Jonathan P. Caulkins on "Providing a Scientific Basis for Managing Illegal Drug Problems," and the plenary session was

delivered by Dr. Elise del Rosario, President, IFORS on "Operations Research: A Powerful and Versatile Discipline."

The other main highlight of the conference was "Teaching Effectiveness Workshop," which was conducted by (i) Dr. Tom Grossman on "Success in Teaching Operations Research to MBAs: Philosophy, Goals, Content, and Approach"; (ii) Dr. Jill Hardin on "Exploring Best Practices in OR/MS Education"; (iii) Dr. G. Raghuram on "Use of Cases for Teaching OR/OM/Modeling"; (iv) Dr. Jim Cochran on "Developing Cases for the Quantitative Classroom: Examples and Advice," and (v) Dr. Ashok K. Mittal on "Fostering Effective Operations Research Education through Active Learning."

Amongst many pioneering applications of OR in India, the major ones are as follows:

- Dr. Jagjit Singh, the second President of the ORSI was the pioneer in introducing OR in Indian railways.
- Dr. G. Raghuram, a past President of the ORSI has been a pioneer in OR applications in Indian rail transport and infrastructure management.
- Dr. G. Raghuram *et al.* developed "Strategies to Handle Pilgrim Population at Tirumala" in 1999.
- Dr. N. Ravichandran developed "A Process-Oriented Approach to Waiting Line Management in a Large Pilgrimage Center in India" in 2005.

Two of the past Presidents of ORSI, namely, the late Dr. Bani P. Banerjee and Prof. S.P. Mukherjee held the post of Vice President of IFORS. In particular, Prof. S. P. Mukherjee was the Vice President of IFORS during the period from 2004 to 2006.

PUBLICATIONS

The Society has been publishing its flagship journal, OPSEARCH, every quarter since 1964. OPSEARCH is the first journal on OR

published in a developing country. The post of Editor-in-Chief of the journal has been held by eminent OR scientists like the late Prof. N. K. Jaiswal, who was the first to receive a doctorate in OR in India from the Delhi University. Prof. Kanti Swarup, an eminent OR specialist followed the footsteps of late Prof. N. K. Jaiswal.

The journal publishes articles on applications and modern developments in OR and allied fields. Each and every article published in the journal is subjected to reviewing by experts in the respective fields. The journal is highly acclaimed by professionals and academicians in India as well as abroad. The abstract of the articles are also published in International Abstracts in Operations Research (IAOR) monitored by IFORS.

Currently, OPSEARCH is being published by Springer India Pvt. Ltd. The hard copies of the journal are distributed to the members and e-copies are available on-line for others.

MEETINGS

The Society is governed by a nationally elected Central Council. The election is held every two years. Three to four meetings of the council are held every year and the annual General Meeting of the Society is held during the annual convention in December every year.

The Society has initiated the Dr. M. C. Puri memorial "Lifetime Achievement" award to recognize a distinguished OR professional from amongst its members every year in its Annual General Body meeting. Dr. M. C. Puri, retired Professor of Mathematics, Indian Institute of Technology, Delhi, and an eminent OR professional, became a victim of terrorist attack in the Indian Institute of Science, Bangalore, on December 28, 2005 while he was attending the ORSI's annual convention there. The "Lifetime Achievement" award has been awarded to three professionals so far, as given below:

- Prof. R. N. Kaul from Delhi in 2007
- Prof. M. R. Rao from Tirupati in 2008
- Dr. Subir Chowdhury from Jaipur in 2009.

Of the three awardees, the most notable in terms of global recognition of research excellence is Dr. M. R. Rao, former Director of IIM Bangalore, current Emeritus Dean of ISB, Hyderabad, a world-class operations researcher, and a past President of ORSI.

PAST PRESIDENTS AND THEIR TERMS OF OFFICE

S. No.	Name of the President	Tenure
1.	Prof. P. C. Mahalanobis	(Expired)
2.	Dr. Jagjit Singh	
3.	Mr. A. Rahman	(Expired)
4.	Dr. V. Ranganathan	
5.	Brig. K. Pennathur	(Expired)
6.	Dr. B. P. Banerjee	1971 (Expired)
7.	Dr. Subir Chowdhury	1972
8.	Col. H. S. Subba Rao	1973
9.	Prof. N. K. Barat	1974
10.	Dr. M. Krishnamoorthy	1975
11.	Prof. S. K. Srinivasan	1976 (Expired)
12.	Dr. Rangalal Bandyopadhyay	1977–1978
13.	Mr. S. M. Sundara Raju	1979
14.	Prof. S. Subba Rao	1980
15.	Dr. Utpal K. Banerjee	1981
16.	Dr. Sushil K. Basu	1982
17.	Prof. A. H. Kalro	1983
18.	Prof. P. R. Ramakrishnan	1984
19.	Prof. K. C. Sahu	1985
20.	Prof. Kanti Swarup	1986
21.	Prof. Nitin R. Patel	1987
22.	Prof. Susy Philipose	1988 (Expired)
23.	Prof. A. Tripathy	1989
24.	Dr. S. R. K. Prasad	1990
25.	Dr. Amitava Mustafi	1991
26.	Prof. K. G. Ramamurty	1992
27.	Mr. M. N. Poddar	1993
28.	Prof. B. R. K. Kashyap	1994
29.	Prof. A. K. Rao	1995
30.	Prof. S. Krishnamoorthy	1996
31.	Dr. Gopal P. Sinha	1997
32.	Dr. I. C. Sharma	1998
33.	Prof. G. Raghuram	1999–2000
34.	Prof. S. P. Mukherjee	2001–2002

S. No.	Name of the President	Tenure
35.	Prof. M. R. Rao	2003–2004
36.	Prof. P. K. Kapur	2005–2006
37.	Prof. Ashok K. Mittal	2007–2008
38.	Prof. Bani K. Sinha	2009

Recently, ORSI has started to honor its past presidents for their outstanding contribution to the society by awarding the title of “Fellow” provided the number of such Fellows is not more than 25 at any time. The current list of Fellows is as given below:

S. No.	Name of the Fellow
1.	Dr. R. Bandyopadhyay
2.	Prof. N. K. Barat
3.	Prof. Sushil K. Basu
4.	Dr. Subir Chowdhury
5.	Mr. P. C. Datta
6.	Prof. A. H. Kalro

S. No.	Name of the Fellow
7.	Dr. Jagjit Singh
8.	Prof. S. Subba Rao
9.	Prof. M. N. Gopalan
10.	Prof. P. R. Ramakrishnan
11.	Mr. V. Ranganathan
12.	Prof. K. C. Sahu
13.	Dr. S. R. K. Prasad
14.	Dr. Amitava Mustafi
15.	Prof. B. R. K. Kashyap
16.	Prof. A. K. Rao
17.	Prof. S. Krishnamoorthy
18.	Dr. Gopal P. Sinha
19.	Dr. I. C. Sharma
20.	Prof. G. Raghuram
21.	Prof. S. P. Mukherjee
22.	Prof. M. R. Rao
23.	Prof. P. K. Kapur
24.	Prof. Ashok K. Mittal

OPERATIONAL RESEARCH SOCIETY OF NEPAL (ORSN): AN INTRODUCTION

SUNITY SHRESTHA HADA
ORSN, Kathmandu, Nepal

HISTORY OF ORSN

Background

The twenty-first century is the period of challenge for global development in multidimensional sectors, ranging from the manufacturing sector, service sector, government sector, nongovernment sector, economic and social sector, gender mainstreaming, natural resources, climatic changes, environmental factors, nuclear technology, and information technology. All these sectors and few other sectors need to make decisions at various levels, such as at operational level, middle level, higher level, or sometimes global level. The interaction of two or more sectors mentioned above make the scenario more complicated, and the decision making process need more rational and some theoretical ground and advanced scientific techniques and tools.

The traditional approach or the heuristic approach used to arrive at a decision or application of some simple or advanced tools come into the category of mathematics, statistics, or management science or the operations research (OR). There are no areas in the world at present, which can be imagined without the application of mathematics, statistics, or OR subject areas.

There are various organizations in the name of Society, Association, or others on the subject areas of mathematics, statistics, management science, and the OR around the world, country-wise, region-wise, and global-wise, which work for the benefit of the personnel working in these areas, giving a platform to interact and network with each other to

share their respective knowledge and their research findings so that everybody can be aware of what is going on around the world in their areas of interest.

In Nepal also, the personnel from statistics area and working in Faculty of Management as academicians, Prof. Dr. (Ms) Sunity Shrestha Hada and Mr. Govinda Tamang of Central Department of Management, Tribhuvan University (TU) realized the need of a society of people working in this area in various sectors of Nepal and came up with the idea of bringing all of them together in one platform. Both having being exposed of the existence of Operational Research Societies in India and other countries and the regional basis, they wanted to network with these societies, and as a result, they applied for the membership in Operations Research Society of India (ORSI) in 2005 and received the life membership.

It was easy to convince colleagues of mathematics, statistics, and engineering areas, as all of them were at the same boat, as they all observed that these areas are not getting the due acknowledgement they deserve in various sectors based on their application. One simple example of the importance not given to the area OR is observed that in the Central Department of Statistics, TU, the course on OR was absent because of the scarcity of faculties in this area, where this area of OR and/or operations management is a compulsory course in management faculty, engineering institutes, and mathematics departments.

Hence, a series of meetings with like-minded colleagues, Sunity and Govinda, was successful in making a team of nine persons, as patron or the founder members and with lots of effort and preparation, got the "Operational Research Society of Nepal (ORSN)" registered duly in the office of Nepal Government at February 1, 2007. Now, the ORSN has about 50 members consisting of both life members and ordinary members in it, with a common purpose of extracting the importance of OR in the society.

Among various objectives, some of them are to advocate the OR areas in academic fields, because without the people educated with OR knowledge, scientific decision making is beyond imagination in this century. It also aims in bringing the academicians, professionals, and practitioners of OR in a single platform of ORSN and share their problems, experience, and knowledge and have in win-win situation and get benefited.

The ORSN also realizes the importance of not only the applications of OR techniques and tools in different sectors but also the theoretical development in this area and desires to encourage the theoretical foundations in Nepal as well. ORSN became 53rd national member society of IFORS and is the member of APORS from 29th November 2011.

Mission

The mission of ORSN is to be a leader organization for the OR theoretical developments and OR applications in Nepal.

Vision

The vision of ORSN is to be an organization that works for the developments of OR by the following ways:

- establishing an “Institute of Operations Research in Nepal” with responsibility of providing academic and vocational degrees of OR and its related fields;
- networking with national and international organizations working in similar fields;
- publishing an international journal of OR.

Objectives

The ORSN has the following objectives as per its constitution at the time of registration:

- to make advocacy and provide orientation in different sectors of Nepal;
- to collect and publish the materials relating to operational research for making the operational research systematic and effective;
- to organize and operate the study and research, training, discussion, seminar, and workshop in national and international levels for promoting competency of persons involved in the area of operational research;
- to provide consultancy services of operational research to institutions working in the concerned area of different sectors such as education, management, industry, business, and others;
- to make significant contribution in the development of operational research in Nepal by maintaining coordination and network with different and international institutions;
- this organization shall be a nonprofit, public-welfare, and social-welfare institution.

STRUCTURE OF ORSN

The structure of ORSN constitutes an executive committee of nine members. The ad hoc committee that worked for initial processes and the registration of the ORSN, and the journal IJORN is the founder member and the patron of the OR Society. All the founder members are life members as referred in the constitution.

Membership Eligibility

The member of ORSN must hold a master's degree in any discipline from recognized university and involve in study-teaching, research, and consultancy in areas of OR. There is no discrimination to obtain the membership from the organization on the basis of race, caste, religion, sex, and political views.

Membership Types

The type of membership is classified as follows in ORSN:

- *Founder/Patron Member.* The office bearers signing the constitution during the registration process of ORSN are the founder members, or the patrons.

All the founder members are life members. There are nine founder members of ORSN.

- *Life Member.* The executive committee grants membership to the eligible persons according to the above criteria and paying NRs. 10,000 (ten thousand rupees in Nepali currency) to ORSN.
- *Ordinary Membership.* The eligible persons can get ordinary membership by paying an entrance fee of NRs. 500 (five hundred) and can be renewed annually with a fee of NRs. 500. The fiscal year for the membership starts from first date of Shrawan and ends at last date of Asadh (mid-july) each year.
- *Honorary Member.* The ORSN can provide this membership to honor personalities who made special contribution to the organization.
- *Institutional Membership.* Institutional membership is provided to any institution who is working in OR or its related areas. The ordinary membership can be provided with an annual fee of NRs. 10,000. Also, life membership to the institution can be provided with a payment of NRs. 100,000 at a time.

The Executive Committee

The structure of executive committee of ORSN constitutes of President, Vice-President, General Secretary, Secretary, Treasurer, and four members. Later, amendment was done and 2 members are added in the executive committee, and the new executive committee includes 11 members. Only life members are eligible for being in the executive committee Table 1.

Members Election

The general meeting is regarded as an apex body of the organization. In general, the general meeting is held once in a year. In special cases, the general meeting can be held more than once in a year. The proceedings of the general meeting cannot take place unless at least 51% of the members are present in the meeting. The executive members and the officials are elected from the general meeting. The tenure of each member is of two years. A person should be life member of this organization for being a candidate of the executive committee. A person worked continuously at least four years (two tenure periods) in the executive committee of this organization can be the candidate of the president of ORSN.

Table 1. Executive Committee

Position	2007–2009 (August)	2009–2011 (August)	2011–2013 (August)
President	Prof. Dr. Sunity Shrestha	Prof. Dr. Sunity Shrestha	Prof. Dr. Sunity Shrestha
Vice-President	Prof. Pushkar Kumar Sharma	Prof. Pushkar Kumar Sharma	Prof. Pushkar Kumar Sharma
General Secretary	Mr. Govinda Tamang	Mr. Govinda Tamang	Mr. Govinda Tamang
Secretary	Mr. Basant Dhakal	Mr. Basant Dhakal	Mr. Basant Dhakal
Treasurer	Mr. Krishna Nakarmi	Mr. Krishna Nakarmi	Mr. Krishna Nakarmi
Member	Dr. Kabita Bade Shrestha	Dr. Kabita Bade Shrestha	Dr. Kabita Bade Shrestha
Member	Ms. Sunil Amatya	Ms. Sunil Amatya	Ms. Sunil Amatya
Member	Mr. Raju Manandhar	Mr. Raju Manandhar	Mr. Raju Manandhar
Member	Mr. Binod Manandhar	Mr. Binod Manandhar	Prof. Hirendra Man Pradhan
Member	—	Prof. Hirendra Man Pradhan	Mr. Vijay Lal Shrestha
Member	—	Mr. Vijay Lal Shrestha	Mr. Dipendra Purush Dhakal

The founder president shall be the honorary president of the organization ORSN. The immediate past president (IPP) will be the executive member for the consequent tenure.

The executive committee may invite any national or foreign experts or any other distinguished personalities to the meeting.

Fiscal Management

The ORSN is a nonprofit organization. There is a separate fund for the organization, and the account has to be maintained as per the format of the prevailing law of Nepal.

The sources of the fund are membership entrance and renewal fee, overhead cost received from training, research, consultancy, and other services. The ORSN also considers the grants assistance or donations, foreign grants, and assistance with prior approval from the Nepal government as the sources of fund. An annual audit must be carried out by a licensed auditor from the Nepal government within one month from the date of expiry of every fiscal year. The audited report need to be approved by the Annual General Meeting (AGM).

THE PRESIDENT OF ORSN

It has been four years since the establishment of ORSN. The tenure of the president being two years, the third president has to take care of the ORSN office. Considering the progress of ORSN and the vision and performance of the current president, Prof. Dr. Sunity Shrestha has been the president for the third tenure.

Among a number of functions, some of the duties and powers of President are to monitor and control the functions and activities of ORSN, to represent the organization, and to make the agreements on behalf of ORSN and other institutions for operations of projects, and to provide directives and policy for operations of the projects.

ACTIVITIES OF ORSN

Website

The ORSN has its own website www.orsn.org.np. It covers the members of ORSN; past,

current, and upcoming events; and the constitution of ORSN. It has a photo gallery too to support the events.

Conferences

The ORSN has successfully organized ORSN Conference 2012 on February 1 and 2, in Kathmandu with the theme "Operations Research for Sustainable Development." The IPP of IFORS, Ms. Elise Del Rosario and other distinguished resource persons and participants from India, Brazil, New Zealand, Korea, and Nepalese graced the events.

However, the ORSN always encourage its members to participate in national and international seminars and conferences throughout the world, which helps in interaction with OR experts from around the world and networking with other organizations.

Mr. Govinda Tamang, General Secretary of ORSN attended the ORSI International Conference held at Kolkata Business School, Kolkata on January 6–8, 2012 with theme "International Conference on Operations Research for Sustainable Development in Globalized Environment" and presented a paper on "Total Quality Management of Commercial Banks in Nepal."

The ORSN president Dr. Sunity Shrestha Hada attended IFORS-2011 Conference held at Melbourne on July 10–15 and presented a paper on "Relative Efficiency of Nepalese Commercial Banks."

Mr. Govinda Tamang attended the SORMS (Society of Operations Research and Management Science) seminar in NITIE Mumbai, India held on 2010 and presented paper on "Relative Efficiency of Commercial Banks"

Also, president Dr. Sunity Shrestha attended the national seminar organized by the SORMS held at Indian Institute of Technology (IIT) Madras, Chennai, India, held on December 20–22, 2009 and presented a paper on "Productivity Management in Education Sector in Nepal."

The president Dr. Sunity Shrestha and general secretary Mr. Govinda Tamang of ORSN participated in the APORS conference in 2009 (December 6–9) at Jaipuria Institute of Management (JIM), Jaipur, India. They

Table 2. Workshop

Workshop Title	Resource Persons
Introduction to lean six sigma	Dr. Charles Aubrey, Chairman, Asia Pacific Quality Organization
Food safety and quality in food service establishments: A major tourism booster	Dr. Miflora Gatchalian, Chief Executive Officer, Quality Partners Co. Ltd, Philippines
Harmonizing workplace relations for quality through labor management councils (LMCs)	Dr. Jose C. Gatchalian, Board Vice-Chairman, Development Center of Asia/Africa/Pacific (DCAAP)

presented a paper on “Effectiveness of Higher Education System in Nepal.”

Seminar/Workshop

The ORSN has organized a number of workshops individually and in association with other organizations.

Training cum workshop on “Revenue Management and Dynamic Pricing” was held at Hotel Royal Singi, Kathmandu, Nepal on January 4–5, 2009. The resource person was Prof. Goutam Dutta from Indian Institute of Management, Ahmedabad (IIMA). Thirty-six participants from 24 different organizations, including government organizations, banks, hospitals, airlines, consultancy services, universities, campuses, and electricity authorities, attended the program.

The ORSN associating with Nepal Bureau of Standard and Meteorology (NBSM) conducted postconference workshops on September 21, 2010 in Kathmandu during the 16th International Conference on Quality

(ICQ) (September 18–20, 2010) coorganized by the Network for Quality, Productivity and Competitiveness-Nepal (NQPCN). The workshop was coordinated by Prof. Sunity Shrestha, president of ORSN Table 2.

The ORSN organized one-day workshop on February 16, 2011 at Kathmandu on Knapsack problems and Project Management. The resource person was invited from University of Graz, Austria.

The ORSN also actively participated in seminar cum workshop held on March 1–4, 2011 on “Emergency Planning” organized by the Central Department of Mathematics, Nepal German Alumni Association Society (NEGAAS) and DAAD. The main resource person was Prof. Dr. Horst Hamacher, who is the IPP of German Operations Research Society.

Talk Programs

ORSN has been organizing talk programs as a regular activity (Table 3).

Table 3. Talk Programs

Date	Talk Program Topics	Resource Person
August 27, 2010	Aggregation scheme: Revisit	Prof. Dr. Devendra Chhetry, Chairman, Nepal statistical Association
September 10, 2010	Survival analysis focusing on acute liver failure	Dr. Shankar Prasad Khanal, Associate professor, Central Department of Statistics Tribhuvan University
January 20, 2011	Carbon sequestration: Local solution to global problems	Dr. Umakanta Mishra, Post Doctoral Research Scholar, University of California, Berkely
May 4, 2011	Estimating health effect coefficients attributed to ambient particulate air pollution	Dr. Srijan Lal Shrestha, Associate Professor, Central Department of Statistics, Tribhuvan University
September 5, 2011	Relative efficiency measurement: DEA approach	Prof. Dr. Sunity Shrestha Hada, Central Department of Management, Tribhuvan University, Nepal

Training

The ORSN has priority in training programs and has planned for conducting trainings on the following areas, and the training cum workshop on *Applications of Computer Technology in Statistics: SPSS 18.0* was conducted at Kathmandu on November 28–December 28, 2011.

- SPSS
- Econometrics
- Research methodology
- Optimization
- Time series

- Forecasting
- Others

Research and Consultancy

The ORSN has the future programs of conducting research in various areas of OR. It also has future plan of performing consultancies with industries in Nepal.

Journal

The ORSN has been working on publishing an international journal annually. The journal has been legally registered under the Nepal Government on 2010. The name

Table 4. IJORN Committees

<i>Editorial Board</i>				
S. No.	Name	Organization	Country	Position in UORN
1	Prof. Dr. Devendra Bahadur Chhetry, devendrachhetry@yahoo.com	Consultant	Nepal	Chief Editor
2	Prof. Dr. Sunity Shrestha, sunity.shresthahada7@gmail.com	Tribhuvan University	Nepal	Member
3	Prof. Dr. Hans Kellerer, hans.kellerer@uni-graz.at	University of Graz	Austria	Member
4	Dr. Tank Dhamala, dhamala@yahoo.com	Tribhuvan University	Nepal	Member
<i>Advisory Board</i>				
1	Prof. Dr. Pushkar Man Bajracharya, pushkarbajracharya_r@hotmail.com	Tribhuvan University	Nepal	—
2	Prof. Dr. Elise del Rosario, elise@jgdelrosario.com	Executive Director, OR Society of Philippines	Philippines	—
3	Prof. Dr. Baikunth Nath, baikunth@unimelb.edu.au	University of Melbourne	Australia	—
4	Prof. Dr. Frank Werner, frank.werner@ovgu.de	University of Magdeburg	Germany	—
5	Prof. Dr. K. K. Lai, mskklai@cityu.edu.hk	City University of Hong Kong	Hong Kong	—
<i>Publication Committee</i>				
1	Prof. Dr. Sunity Shrestha, sunity.shresthahada7@gmail.com	Central Department of Management, TU	—	—
2	Prof. Pushkar Kumar Sharma, pushkar-s@hotmail.com	Nepal Commerce Campus, TU	—	—
3	Mr. Govinda Tamang, tamang_govinda@yahoo.com	Central Department of Management, TU	—	—
4	Mr. Krishna Nakarmi, nakarmikrishna@gmail.com	Agricultural Development Bank, Nepal	—	—

of the journal is International Journal of Operations research-Nepal (IJORN). Different committees have been formed to activate this. The first issue of IJORN has been published on February 1, 2012 at the occasion of the sixth anniversary of ORSN Table 4.

Institute of Operations Research

The ORSN has a highly optimistic vision of establishing an “Institute of Operations Research” in Nepal. The main purpose of the institute is to impart knowledge of theoretical development and applications of OR in various sectors.

DISCUSSION

The ORSN meets with its members about once in a month. All the members have their association with their own organizations that

ranges from universities, colleges, institutes, financial sectors, airline sectors, health sectors, government offices, and research consultancy firms. Talk program is a regular program of ORSN, where the members get opportunity to meet, interact, and discuss their academic and other activities and share their experience.

The ORSN has tie-up and collaboration relationship with other organizations in Nepal and in international institutions. It is working with “Institute of Organizations Career Development (IOCD)” in sharing the training programs and seminar/workshops. It is also working with “Nepal Evaluation Society (NES)” and “Innovative Solutions Pvt. Ltd” in research and consultancy.

The ORSN is corresponding to “University of Graz at Austria” with the initiative of Prof. Dr. Hans Kellerer for institutional interaction and exchange programs.

OPERATIONAL RESEARCH SOCIETY OF TURKEY

TANER BILGIÇ

REFİK GÜLLÜ

Department of Industrial Engineering,
Boğaziçi University, Bebek,
Istanbul, Turkey

Roots of the Operational Research Society of Turkey (ORST) go back to 1965, when the Scientific and Technological Research Council of Turkey (TÜBİTAK - responsible for the funding and coordination of government supported research) decided to establish an Operations Research (OR) unit within its organization. Halim Doğrusöz, who received his PhD in OR in the group Russel Ackoff brought together at Case Institute of Technology, established the OR unit within TÜBİTAK. He was its first chairperson until 1968.

Later, in 1975, Prof. Doğrusöz founded a graduate Operations Research and Statistics program at the Middle East Technical University, Ankara. ORST was founded March 10th of the same year by a group of fellow OR researchers including Halim Doğrusöz, Muhittin Oral, Kutay Arman, Ünver Çınar, Erkut Yücaoglu, and Ataç Sosyal. The last two names are from Boğaziçi University and Istanbul Technical University, respectively. These two universities in Istanbul had recently founded their Departments of Industrial Engineering with a strong emphasis on OR. The Operations Research and Statistics program at the Middle East Technical University later merged with the Department of Industrial Engineering in the same university. Over the years, OR education had been largely organized under Departments of Industrial Engineering in Turkey and enjoys a high reputation as a popular profession among university students. As of the end of 2009, there are 54 Industrial Engineering departments in 54

different universities with an intake of more than 4400 students every year.

One of the early mandates of ORST has been to make OR visible in Turkey. To that end, one of the first challenges of ORST had been the need to translate “OR” into Turkish. Military OR has been in existence in Turkey since 1954 within the “Scientific Investigation Council” of the Turkish Armed Forces. In 1956, the army instituted an OR unit. Although it seemed natural to use the word “Harekât” in Turkish as the translation of the term “Operations,” ORST noted the overwhelming military use of that term in Turkish and needed an alternative. They coined the term “Yöneylem” (which is an amalgamation of “direction” and “action” in Turkish) and translated “OR” as “Yöneylem Araştırması” (“Research on Directed Actions”) stripping it largely from its military meaning. When late Prof. Ackoff visited Turkey in the early 1970s, he was asked what he thought about the translation of OR into Turkish. He replied “Directed actions? So OR is actually ‘Research on Purposeful Actions.’ This is even better than the English name of the profession!” [1].

Past presidents of ORST are: Halim Doğrusöz (1975–1979), Merih Celasun (1979–1981), Ataç Soysal (1981–1989), Ömer Saatçioğlu (1989–1996), Çağlar Güven (1996–2001), Nesim Erkip (2001–2003), Haldun Süral (2003–2005), Refik Güllü (2005–2009), Taner Bilgiç (from 2009 to till date). The Society is governed by an executive committee comprising the president, secretary, treasurer, and two members. The executive committee is elected by the general assembly.

Today, ORST is a well established society with a total membership body of about 900, from both academics and industry. Major activities of ORST are the organization of a yearly Operational Research and Industrial Engineering (OR/IE) congress, and the publication of the journal *Transactions on Operational Research*. ORST is a full member of EURO (The Association of European

Operational Societies) and IFORS (International Federation of Operational Research Societies).

The first national OR/IE congress took place in Istanbul in 1975 and was organized by ORST, TÜBİTAK Marmara Research Center, and Boğaziçi University. OR/IE congress usually takes place in the summer and it is hosted by one of the universities in Turkey. In 2009, the congress was hosted by Bilkent University in Ankara. In 2010, the 30th Congress will be held in Istanbul at Sabancı University. Although the language of the congress has always been Turkish, over the years there are more sessions conducted in English and more international participation. Along with plenary speakers, tutorials, contributed and invited sessions, a student paper competition is also organized to honor the best student work of the year and prize for Excellence in Practice is awarded. Typical attendance to OR/IE congresses ranges between 250 and 350 with about 170–200 papers presented.

ORST also organizes a biannual Doctoral Colloquium for doctoral students with the aim of bringing together doctoral students from the nation to share their research and experiences. The doctoral programs in OR are relatively young but growing rapidly.

PhD degrees in OR were awarded for the first time in the late 1970s and early 1980s by the Middle East Technical University and Boğaziçi University. Owing to its popularity as a profession, Industrial Engineering departments have attracted and still attract very strong students. Many of these students have pursued graduate degrees abroad (mainly in North America) and there is a sizeable community of OR researchers and practitioners outside Turkey.

ORST began awarding a prize for *Excellence in Practice* in 2002. The objective of this award is to promote excellence in using OR while successfully solving problems of practical interest. The award also aims to coordinate the efforts of various isolated groups applying OR to practical problems.

ORST maintains a website (www.yad.org.tr) but it is mostly in Turkish. It also has access to its membership body via email lists.

REFERENCES

1. Doğrusöz H. Yöneylem Araştırması Serüvenim (My Adventure in Operational Research). Istanbul; Boğaziçi University Press, ORST Publications; 2009. (in Turkish).

OPERATIONAL RISK

RODNEY COLEMAN
Department of Mathematics,
Imperial College London,
London, UK

RISKS ARISING FROM BUSINESS ACTIVITY

You lose business opportunities while your computer systems are down. You settle a wrongful dismissal claim. You receive a penalty notice for the late filing of documents. Your employee embezzles from you. Your customer records have gone missing. A flood has breached your stock room damaging goods awaiting shipment. These are the day-to-day hazards of running a business beyond those from its money-making activities. This is operational risk, often shortened to *oprisk*. It excludes risks which result in losses from poor business decisions. Oprisk losses will generally stem from weak management, from outsourcing nonstrategic activities, or from external factors.

Operational risk events happen everywhere, not just in a business environment; for example, injuries sustained from a design weakness in playground or fairground equipment, or engineering flaws resulting in a mass recall of motor cars. Poor labeling on pharmaceutical packaging can lead to the wrong dose being administered. Swabs or instruments can be left in a patient after surgery. Attention to safety has a big role in mitigating these hazards.

Major construction projects—digging tunnels, mineral extraction, raising bridges, building skyscrapers—all need careful planning not just to support efficient operations but to minimize avoidable delays from operational risk events. Large companies involved with manufacturing, technology, telecommunications, and media services will all have operational risk management departments.

Business is entered into with the aim of generating wealth, in the pursuit of returns

on capital. In contrast, the supervision, control, and management of operational risk absorbs capital in the protection of wealth, without being directly profit generating. Indeed a major cost is often on business continuity planning, preparing for eventualities beyond the control of management that threaten the business environment. These include terrorism, civil unrest, systems failings including hacking, organized crime, and nature's worst behavior such as hurricanes, tsunamis, and volcanic ash clouds.

Measuring and modeling operational risk events is needed to answer questions of how high to build a flood barrier, how much to spend on railway safety, and how much to put into reserves to protect a bank from collapse.

Oprisk data for study are hard to come by; their publication might expose a firm's failings, and make it lose competitive advantage. The financial sector has led much of the study of operational risk, not only because losses are widely expressed in monetary terms but also since the industry is highly regulated. This is to ensure not only that our savings and pensions are safe but also that businesses and governments would be able to continue to borrow and lend following a worst case financial crisis. When the loss events are penalties imposed for contravening regulations, they enter the public domain.

Onetime categorized as the catch-all "other risks" after credit and market risk are excluded, operational risk was pretty universally considered to be unmeasurable. In 2001, however, Basel II, which set out a regulatory framework for operational risk in international banking, was to give a generally accepted definition for operational risk [1, para. 664, p. 140]. It was to be the risk of loss resulting from inadequate or failed internal processes, people, systems, or from external events. More importantly, Basel II was the driving force behind developments in operational risk management practices, the search for robust risk measurement, and the requirements of transparency through disclosure. This has made operational risk a

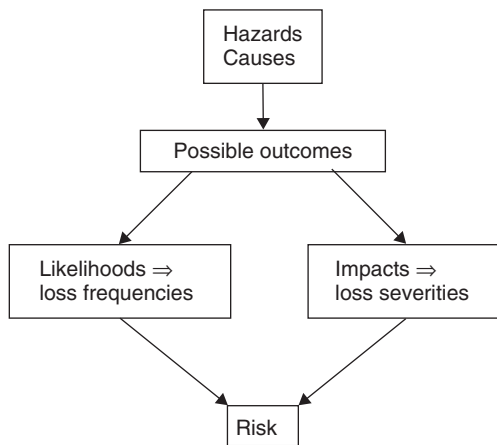


Figure 1. A basic risk structure.

significant topic in business and management studies. (For more on Basel II, see Box B.)

The basic operational risk structure is shown in Fig. 1. A risk management structure and an operational risk management structure follow later.

OPERATIONAL RISK EVENTS MAKE HEADLINES

Public attention gets drawn to operational risk when infrequent but high impact events occur. Fitch's OpVar database has nearly 500 events in the United States between 1978 and 2005 where the losses exceeded \$10 million. With 10 years of data, the 2004 loss data collection exercise (LDCE) has more than 100 events in the United States where the losses were \$100 million or more [2, p. 2].

Brief narratives follow on some extreme oprisk events.

- *Union Carbide.* During the night of December 3, 1984, 40 tons of poisonous gas leaked from the US owned Union Carbide pesticide factory in Bhopal, India, and settled over neighboring slums. More than half a million people were exposed to the deadly fumes. Within days, 3500 people were dead, and more than 15,000, possibly as many as 25,000, have died since. It stands as the world's worst ever industrial accident. Union Carbide

denied responsibility. The effects of the gas on survivors continues, with chronic illnesses, cancers, and children with birth defects. Environmentalists claim that the neighborhood has not been rid of toxins, said to be still polluting the local water supply.

- *Space Shuttle Challenger Disaster.* On January 28, 1986, shortly after take-off from the Kennedy Space Center in Florida, the space shuttle Challenger exploded killing all seven on board. A seal on a rocket booster motor had failed allowing hot pressurized gas to reach an external fuel tank, causing its structure to fail and the spacecraft to disintegrate.
- *Bank of Credit and Commerce International.* BCCI was closed down in 1991 after it was revealed that accumulated losses totaling more than a billion US dollars had been fraudulently hidden from auditors and regulators.
- *Sumitomo Corporation.* Between 1985 and 1999, a star copper trader at Sumitomo incurred massive losses through risky trading. It was sufficient to affect the entire copper market.
- *Barings Plc.* Barings Bank, a long established merchant bank in the city of London, became insolvent in 1995 when a young trader at its Singapore subsidiary lost £860 million from excessive and unauthorized speculation on the Tokyo stock exchange index. He hid this activity for two years. When brought to light, it was found that he had failed to hedge enormous outstanding positions, cover for which required payments that Barings could not meet. That its main assets were held offshore probably contributed to the Bank of England being prepared to let the bank fail.
- *Prudential Insurance Company of America.* The Prudential paid out \$2 billion in settlement of a class action lawsuit relating to its agents deliberately misleading purchasers throughout a period of 13 years up to 1995. More recently in the United Kingdom, there was similarly the costly endowment mis-selling scandal

of people being misled into exchanging defined company pension plans for riskier and less advantageous market-funded schemes. In December 2002 one of the insurers involved, Abbey Life, was fined a record £1 million and paid £160 million in compensation to 45,000 policy holders.

- *Hurricane Andrew.* In 1992, Hurricane Andrew swept through the south of Florida. The damage led to a \$16 billion insurance payout, a record insured loss, only exceeded with the terrorist attack that devastated the World Trade Centre in Manhattan on September 11, 2001.
- *The Global Financial Crisis of 2007–2009.* The worldwide financial breakdown began with the credit crunch, the seizing up of the markets on August 9, 2007 following problems created by subprime mortgage lending in the United States. It was the result of a bubble created by cheap borrowing, with offloaded debt insured against default. In their premium pricing, the insurers failed to take account of the risk that property values might fall with consequent defaults on mortgage repayment. The operational risk was in the lack of due diligence: the lenders in failing to check the means of their mortgage applicants, the purchasers of the debt similarly in respect of the mortgages supporting it, and the insurers in taking on the default risk. Poor business practices caused the crisis to propagate from insurers without the means to pay out on the mortgage defaults, to the banks who lent the money, and to the householders who lost their homes. This is illustrated by the fall of Northern Rock, which suffered the first run on a British retail bank in a century. (Box A.)
- *Deepwater Horizon.* On April 20, 2010, 50 miles off the Louisiana coast in the Gulf of Mexico, the Deepwater Horizon drilling rig exploded and sank killing 11 workers. Hundreds of thousands of gallons of oil began leaking daily from a ruptured pipe, causing untold environmental damage. The oil slick

washing up on the beaches threatened wildlife and businesses. The fishing industry was all but halted, and tourism was curtailed. The drilling platform was leased to the British oil giant BP from the Swiss headquartered rig operator Transocean, the world's largest offshore drilling company. The explosion happened after the US company Halliburton, the world's largest provider of services to the energy industry, laid cement casings round the well head, completing its construction. BP acknowledged that it had been unprepared for the disaster. Operational risk controls failed in this case.

- *Volcanic Ash.* On April 14, 2010, a volcano erupted through the Eyjafjallajökull glacier in Iceland and sent a cloud of ash into the upper atmosphere. For safety reasons, the airspace over parts of northern Europe was closed. With thousands of flights grounded, there was widespread disruption to travel and business activity. Travelers became stranded abroad. Airspace was cautiously reopened five days later. The previous volcanic eruption at the glacier had been in December 1821. The long passage of time since led to the 2010 eruption being called a “black swan” event, an event so unlikely as to be off everybody's radar.

RISKS AND RISK MANAGEMENT

Before considering operational risk and its management, we give a very brief introduction to risks in general and their management.

Classifying Risks

Risks are often sorted into categories, such as those given here with illustrative examples [3, p. 107].

- *People Risks.* Deliberate willful behavior such as fraud or malicious damage. Errors such as mistakes brought on by fatigue, incompetence, lack of management supervision, and inadequate staffing levels.

Box A. Northern Rock

This former Building Society (savings and loan company) had floated on the stock exchange as a bank in 1997. It moved into subprime lending in 2006, issuing mortgages in excess of 100% of property valuations, often on self-reported and unverified household incomes. It expanded quickly to become the United Kingdom's fifth largest bank, funding its rise by borrowing on money markets, securitizing its mortgage debt, and other financial instruments. When the US's \$8 trillion housing bubble burst, it became clear that Northern Rock's business model was severely defective.

The credit crunch of August 2007 caused the share price to fall fast. In September, there were depositors queueing round the block seeking to withdraw their generally modest savings. Withdrawals were also being made on-line. In December, the Bank of England stepped in by providing liquidity support, and sought a buyer, but without success. On February 22, 2008, it was taken into public ownership. Yet, it was considered to be well managed operationally.

No measures had been taken to provide for a fall in property prices. It has been reported that Lehman Brothers had underwritten £100 billion of Northern Rock's debt (collateralized debt obligations on mortgage-backed securities).

Lehman Brothers itself filed for bankruptcy on September 15, 2008. A week earlier, the giant US mortgage lenders Fannie Mae and Freddie Mac had received a government bailout, and the largest insurance company, AIG, and the largest savings and loan company, Washington Mutual, were in dire straits, lining up for support.

The response to the collapse of Northern Rock was to herald the international effort to stave off global collapse of the financial system.

- *Cumulative Interactive Risks.* Minor errors build up leading to a major loss. The effects accumulate as they propagate through a company. This is the so-called Swiss cheese model, where the holes/hazards line up, with all controls failing simultaneously. It is widely used in investigating catastrophic incidents [3, p. 113; 4, p. 234].
- *Systems and Process Risks.* Technology failure, programming errors, errors in data management, insufficient processing capacity. Lost or wrong paperwork, and inadequate segregation of duties.
- *External Factors.* The sinking of a cruise liner would severely damage the entire cruising industry, and have an immediate impact on shipping insurance premiums.
- *Randomness.* Low risk events having high impact can be due to pure chance. However unlikely, lottery jackpots are won.

We need also to understand that these categories are not necessarily mutually exclusive. Catastrophic industrial accidents at a plant can occur despite the plant having an exemplary safety record. In March 2005, there was an explosion at BP's Texas City refinery which left 16 people dead and more than 170 injured. A report on the incident found that BP had interpreted improvements in personal injury rates as indicating acceptable process safety performance. Elliott *et al.* [5] used this in the introduction to their study of the relationship between occupational illnesses and injuries (OII) and the major incidents reported to the US Environmental Protection Agency over the same period, 1996 to 2000. Their statistical analysis finds only weak evidence of a link.

A Simple Framework for Risk Management

This is given in Fig. 2.

Box B. Basel II

The Basel Committee for Banking Supervision (BCBS) was set up in 1974 as a committee of the Bank for International Settlement (BIS) to provide a regulatory framework for internationally active banks. In its Basel Accord of 1998, now known as Basel I, it settled the minimal level of capital to be held by banks as provision for credit risk and market risk. In 2001, it moved to do the same for operational risk in its New Basel Capital Accord, known as Basel II [1]. It was approved by the European Parliament in 2005, and came into effect across the entire European Union (EU) in 2008.

The accord sets out a risk sensitive way of calculating reserve capital to cover possible defaults. Institutions are required to categorize operational risk losses by event type, promoting identification of risk drivers. There is no mandated methodology.

Pillar 1 of Basel II gives three ways of calculating the operational risk capital charge, with increasing complexity, but benefiting from a reduced charge.

- The *basic indicator approach* (BIA) calculates the reserve capital simply as a proportion of gross revenue.
- The *standardized approach* (TSA) divides the activities of a bank into eight business lines (Table 2), with standard capital charges for each based on calculated risk indicators.
- The *advanced measurement approach* (AMA) requires that the banks model loss distributions of cells of a business line/loss event type grid from operational risk loss data that they themselves have collected, supplemented as required by external data.

Pillar 2 of the accord requires banks to demonstrate that their management and supervisory systems are satisfactory. Pillar 3 relates to transparency, requiring them to report on their operational risk management.

Solvency II, the EU's regulatory directive for insurers, has adopted the same three pillars. This directive will come into force throughout the EU in 2013.

In November 2007, the US banking agencies approved the US Final Rule for Basel II. Banks will be grouped into the large or internationally active banks that will be required to adopt AMA, those that voluntarily opt-in to AMA, and the rest who will adopt an extended version of the earlier Basel I. A Basel III is in preparation.

Management Tools

The following are approaches to risk management, and in particular to operational risk management (ORM).

- *Risk and Control Self-Assessment (RCSA)*. This is the identification, evaluation and checks on the effectiveness of all significant risks, and the recording and reporting of the results. This may involve the use of questionnaires to establish expert opinion, monitoring key performance indicators (KPIs), form filling, workshops, and independent assessments. External methodologies developed for RCSA, or by supervisory and regulatory

bodies for their own use, can support this.

- *Loss Event Monitoring*. This is the collection and tracking of loss events and near misses, with regular reporting. In some fields, regulators will have specified the events they require to be tracked. Good management will expand on these, particularly in respect of near-miss events which can be predictive of potential loss.
- *Key Risk Indicators (KRIs)*. Tracking and monitoring KRIs can provide early warnings of areas of concern. Doubts stand as to their predictive value.

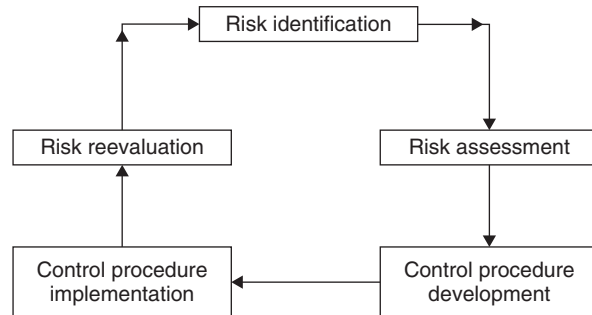


Figure 2. The risk management cycle.
(Source: Taken with permission from Risk Books [6, p. 228].)

- *Stress and Scenario Testing.* This is the contingency planning for possible adverse events, from payouts occasioned by a badly drawn up contract, to the disaster recovery and business continuity to follow from terrorism or natural catastrophe causing the loss of headquarters, paperwork, processing capacity, and so on. It involves testing impact tolerance and resilience.

CLASSIFYING OPERATIONAL EVENTS

When something is defined only by what it is not, there is always going to be a problem in giving it a taxonomy. A major hindrance prior to Basel II was not only its lack of a constructive definition but also the absence of data. Operational risk losses were generally treated as costs of doing business and allocated to the department where they occurred. As such they were not recorded specifically as operational losses. Even when they were identified as such, if the losses were small, they were not going to contribute significantly to business failure. Post Basel II, data can still be very sparse. An insurer recently had just two oprisk events to show financial supervisors, with another six possible (personal communication).

Operational Risk Loss Event Types in Banking and Insurance

In practice, we would need to identify the operational risk loss events particular to the business activity. A start at this classification can be made by using the seven designated categories of loss events given in Basel II, also adopted in Solvency II, the EU regulations.

- *Internal Fraud (IF).* Losses within the business from fraud, misappropriation of property, unauthorized activity, and circumventing regulations.
- *External Fraud (EF).* Fraudulent claims by an external party, forgery, and hacking damage to systems security.
- *Employment Practices and Workplace Safety (EP and WS).* Organized labor activity, violations of employee health and safety rules, discrimination in employment, and personal injury claims.
- *Clients, Products, and Business Practices (CP and BP).* Unintentional failure or negligence in meeting professional obligations to clients or customers (customer complaints, the suitability of advice, lack of disclosure, including breaches of trust). Flaws in the design or behavior of a product.
- *Damage to Physical Assets (DPA).* Losses from damage to property from natural catastrophes (hurricanes, floods) or man-made events (fires, explosions, terrorism, pollution).
- *Business Disruption and System Failures (BD and SF).* Losses due to hardware or software failure, system design failure, and other infrastructure issues.
- *Execution, Delivery, and Process Management (ED and PM).* Failed transaction processing or management, failed customer/client services (account errors, data entry errors, and incorrect payments), and inadequate monitoring and reporting.

Table 1. The Percentages of Losses of \$10,000 or More for Each Event Type (Taken from the 2005 LDCE in the United States)

Event type	IF	EF	EP and WS	CP and BP	DPA	BD and SF	ED and PM	Other	Total
Losses (%)	3.8	41.8	7.6	9.2	0.7	0.7	35.3	0.8	100

Source: [7].

Table 2. Business Units and Business Lines for International Banking Activities Under Basel II. Percentages of Losses are from the 2005 LDCE

Business unit	Business line	Frequency (%)
Investment banking	• Corporate finance	0.4
	• Trading and sales	7.9
Banking	• Retail banking	65.3
	• Commercial banking	5.5
	• Payment and settlement	4.8
	• Agency services	5.5
Others	• Asset management	2.7
	• Retail brokerage	7.9

Source: [7].

The percentages of losses of \$10,000 or more for each event type, derived from [7, slide 10], the 2005 LDCE in the United States, are shown in Table 1.

Business Lines

Managing operational risk requires a more granular breakdown of the business activities in which the losses occur. Basel II gives eight broad business lines within the banking sector, creating an 8×7 grid of 56 business line/event type cells. Table 2 shows the percentages of losses from each business line from the 2005 LDCE [7, slide 12].

The Operational Risk Consortium Ltd. (ORIC) database of operational risk events, established in 2005 by the Association of British Insurers (ABI), has nearly 2000 events showing losses exceeding £10,000

in the years 2000 to 2008. Insurance has a different set of business lines from banking. The most significant with respect to operational risk are given in Table 3. The event types are those mentioned in the section titled “Operational Risk Loss Event Types in Banking and Insurance.” The data, derived from Selvaggi [8, Fig. 4A, p. 14], give percentages of loss amounts for those business activity/event type cells having at least 4% of the total loss amount.

This information tells little about the actual events. For this we need level 2 and level 3 categories, the seven event types being level 1. To illustrate this, again from the ORIC database, Table 4, derived from Selvaggi [8, Fig. 4B, p. 15], shows the most significant level 2 and level 3 event types in terms of both severity and frequency (values over 4%). It excludes losses arising from the UK’s mis-selling of endowment policies scandal. We note that natural disasters do not feature as significant.

OPERATIONAL RISK MANAGEMENT

Is anything more than common sense needed for ORM? Or so a referee of an application for research funds reported to the awarding panel. We already manage our exposure to operational risk by locking filing cabinets and doors, by using fancy passwords, by installing antivirus software, and so on. It has been argued that it would be sufficient for every employee, from the shop floor through to the main board, to be educated about risk. Of course, one response might be that we would still need to address problems from outsourcing and other subcontracting, and from external events outside the control of the firm. We must remember also that personal attention to risk does not preclude process risk.

A widely accepted approach in risk management is to identify the major risks

Table 3. Business Activity/Event Type Grid Showing Percentages of Loss Amounts (Minimum 4%) in the ORIC Database (2000–2008)

Business activity	Event type				Total
	CP and BP	ED and PM	BD and SF	Others	
Sales and distribution	18.9	6.8		0.7	26.4
Customer service/policy		13.2		2.3	15.5
Accounting/finance		23.4		0.1	23.5
IT			6.0	6.6	12.6
Claims		4.0		1.4	5.4
Underwriting		6.3		0.3	6.6
Others	5.1	11.8	1.4		10.0
Total	24.0	65.5	7.4	11.4	100.0

Source: [8].

Table 4. Level 2 and Level 3 Event Categories from Insurance Losses in ORIC (2000–2008) that Show Loss Amounts and the Frequency of Losses of 4% or More

Level 2 events	Level 3 events	Size (%)	Frequency (%)
Advisory activities	Mis-selling (nonendowment)	13	9
Transaction capture, execution, maintenance	Accounting error	12	
	Inadequate process documentation	8	
	Transaction system error	8	6
	Management information error	7	
	Data entry errors	7	5
	Management failure	5	
	Customer service failure	4	16
Suitability, disclosure, fiduciary Systems	Customer complaints	6	4
	Software	6	
Customer/client account management	Incorrect payment to client/customer		9
	Payment to incorrect client/customer		4
Theft and fraud	Fraudulent claims		4
	Total	76	57

Source: [8].

faced by an organization, measured by their impact in terms of their frequency and severity. In many cases, from a list of 100 operational risks identified as occurring within an organization, no more than 5 or 10 will give rise to the most serious of the loss events and loss amounts. However, the ubiquitous nature of operational risk requires that day-to-day management must use bottom-up as well as top-down risk control. There needs to be an understanding of the risk in all activities. Aggregating losses over business lines and activities would tend to hide the low impact risks behind those having a more dramatic effect. The bottom-up approach is thus a necessary part of seeing the complete risk picture.

Aggregation at each higher level informs management throughout the business.

The perception of the risk needs to match the actual risk. Hazards that have not yet been encountered may not even be considered, and when considered have risks that are hard to calculate.

Operational Risk Management Structure

Figure 3, based closely on Álvarez [6, p. 231], shows the structure of an operational risk management program.

An internal operational loss event register would typically show high impact events at low frequency among events of high frequency but low impact. A financial institution might therefore sort its losses into “expected loss” (EL) to be absorbed

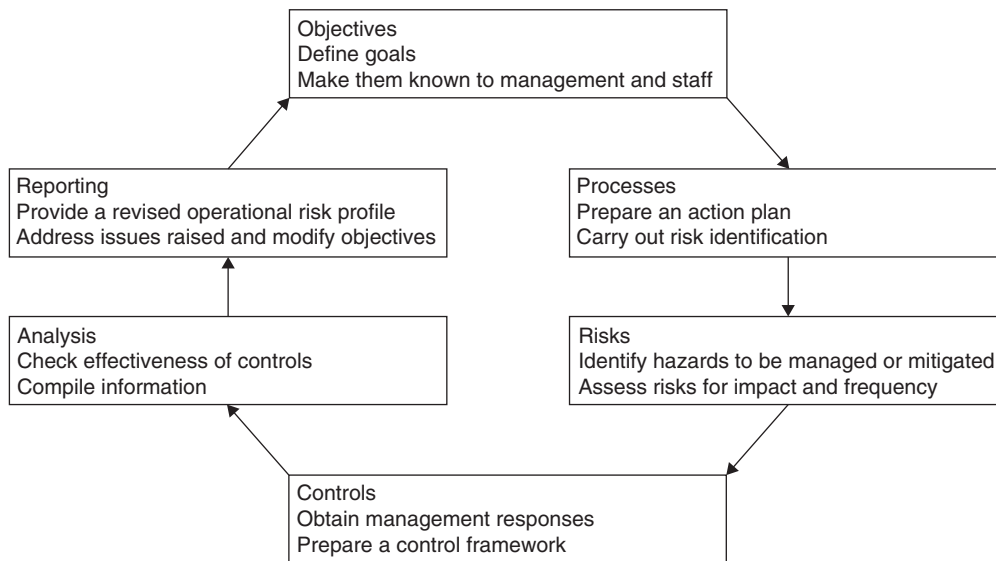


Figure 3. The operational risk management cycle. (Source: Taken with permission from Risk Books [6].)

by earnings, “unexpected loss” (UL) to be covered by risk reserves (so not totally unexpected), and “stress loss” (SL) requiring core capital or risk financing for cover. The “EL” per transaction can easily be embedded in the transaction pricing. It is the rare but extreme stress losses that the institution must be most concerned with. This structure is the basis of the loss data analysis approach to operational risk. Hard decisions need to be made in choosing the EL/UL and UL/SL boundaries. This latter would often be the maximum probable loss. Besides these, a threshold “petty cash” limit is needed to set a minimum loss for recording it as an operational loss. Loss events with recovery and other near-miss events also need to be in the internal database for the information they carry. The threshold and boundaries are set separately for each business activity/event type category.

External Data

In addition to the management tools mentioned in the section titled “Management Tools”, operational risk management will often need to supplement their internally collected data. The shortage of in-house oprisk data, particularly large loss events that

would have a major influence in estimating reserve funds to cover a major loss should it occur, is such that data publicly available or from commercial or consortia databases will need to be explored.

Insurance

Transferring operational risk through insurance is problematic as a risk management tool [3, p. 187].

For example:

- Blanket cover would not be available, leaving unforeseen events uncovered.
- Exclusions could deny payment. Similarly, delays in payment, possibly for years if legal proceedings take place, could put firms at risk.
- The absence of sufficient and appropriate data would make pricing the risk difficult.
- Risk transfer may lead to moral hazard, the abandonment of responsibility for risk management.

Most importantly perhaps is that, although insurance is a means of mitigating the consequence of operational risk losses, it does nothing to enhance control of the risk itself [9, pp. 11 and 259–260].

MODELING OPERATIONAL RISK

The Marked Point Process

The appropriate structure for modeling oprisk events is that of the *marked point process*. Along the time line (the x axis) we have points representing the times of occurrence of the events. We attach marks to each of these points. These are labels representing information about the events (see the section titled “Database Acquisition”). In practice, the first label is the loss size z , which can be represented by a vertical spike at its time point t , its height being proportional to the size (with the y axis showing the scale). The second label is a code i for the i th business activity, and a third with a code j for the j th event type. We are thus coding each point by a (t, z, i, j) .

The statistical modeling has to identify the probability structure for these four-dimensional descriptors. We can study each business line separately and compare its losses and their occurrence times with those for any other business line, or even the rest of the business lines; and similarly for the event types, and every business line/event type combination. This structure can be visualized as the superposition on the time line of processes for each business line/event type combination (i, j) separately. The interaction between the interval processes of gaps between events and their sizes can also be studied.

We have here a rich mixture to model. In practice, we would need substantial amounts of data to provide more than just basic statistical properties. In the section titled “Example: Fitting GEV and GPD,” we model data from a single activity and event type and fit only a loss severity probability distribution.

MEASURING OPERATIONAL RISK

Basel [1, para. 665] requires that the internal measurement system of banks adopting the advanced measurement approaches “must reasonably estimate unexpected losses based on the combined use of internal and

relevant external loss data, scenario analysis, and bank-specific business environment and internal control factors (BEICF).”

Using the prudent amount of risk capital allocation as a proxy for risk, and obtaining this measure by modeling collected loss data, may seem to satisfy the need for objectivity. It seems to avoid the subjective interpretation of risk indicators and expert opinion.

In practice, the quality of data and of its collection, the appropriateness of the model choice, the risk modeling of potential high impact outcomes that have not happened, the very shortage of data on those that have, and assessing and aggregating the loss from a major event, or the potential loss from a near miss, the dynamics of risk management practices making historic data out of date, the use of external data which cannot match the loss profile of the business, all of these point to loss data analysis being a highly subjective approach to quantitative measurement of operational risk.

Despite this, the act of establishing a price for risk highlights risky activities, as well as those held back because of exaggerated risk assessment. It can be used as the basis of cost–benefit analysis in capital allocation. Further, loss data modeling allows for predictions and forecasting, for identifying correlations between business lines and between event types, for scenario analysis, stress testing, benchmarking, validation, backtesting, and so on.

There is no true answer, and no methodology on its own can provide it. Multiple approaches should be used, both qualitative and quantitative. The objective must be to assist management in acquiring a sensitivity to data, and in its interpretation for use in decision making.

Database Acquisition

A loss event register is vital to understanding the risk environment. The data entered should show more than just the loss amounts. The records should show for each loss event

- a reference number for the event;
- the level 2 business line code;
- the level 3 business line code;

- the event type;
- where it occurred;
- when it occurred;
- when it was discovered;
- the gross loss amount;
- amount of any direct recovery;
- amount of any indirect recovery;
- if a near-miss event, an estimate of the potential loss amount;
- reference numbers of related events;
- a short narrative, to include, if needed, proposed corrective action. (*Source*: [3, p. 111; 10].)

External Data

Basel II requires internationally active banks to have a systematic process for determining as to when external data must be used and how to make use of the data (“the use test”), for example, scaling, qualitative adjustments, or improving scenario analysis. The accuracy and completeness of external data have to be validated.

Scenario analysis enables an organization to consider responses. However, merging extreme scenario data with actual loss data corrupts those data and distorts the loss distribution, and could lead to higher capital charges than might otherwise be the case.

External loss data are available from industry consortia and commercial sources. Reference has already been made in the sections titled “Operational Risk Events Make Headlines” and “Business Lines” to the OpVar and ORIC databases.

The Operational Riskdata eXchange (ORX) Association is a well-established database of operational risk events in banking. The not-for-profit consortium collects data quarterly from its 54 member banks from 15 countries, using a standard format. It has attempted to establish loss data standards for the secure exchange of high-quality anonymous oprisk loss data. The standard format of ORX informed the list proposed in the section titled “Database Acquisition” for internal database acquisition. In March 2010, ORX reported that it has more than

180,000 losses, each at least €20,000, and totaling more than €47 billion [10].

There are also national banking databases, and those using publicly available information from the press and trade publications. Insurance companies have their own claims registers, with accurate records of amounts paid in settlement.

The Small Sample Problem

An extreme loss in a small sample is over-representative of its 1-in-a-1000 year or 1-in-a-10,000 year chance, yet underrepresented if not observed. The largest losses are generally overly influential in any fitted model. So, we must conclude that fitting a small data set cannot truly represent the loss process whatever model is used.

Basic to statistical practice is reporting on the quality of any estimate. With only a small data sample of oprisk losses, with none very large, we have no way of reporting an accurate value of a loss that would be seen only once in a 1000 years. It is a well-understood principle that any statistical inferences about models in regions far outside the range of the available data should be treated with caution. In carrying out loss data analysis we are operating outside our statistical comfort zone.

Probability Modelling of Loss Data

The Gaussian (normal distribution) model of much of statistical inference is inadequate for loss data. Even a truncated normal distribution that excludes negative and small positive values gives too little chance to large and very large losses. The lognormal distribution has been used instead historically in econometrics theory, and the Weibull in reliability modeling.

Two models that can allow rare large observations come from extreme value theory. The generalized extreme value (GEV) distribution and the generalized Pareto distribution (GPD) are the limit distributions as sample sizes increase to infinity. Environmental studies (hydrology, pollution, sea defences, etc.) already use them. Each of them has parameters μ giving location, σ scale, and ξ shape, which we vary to obtain a good fit. The location μ and scale σ are

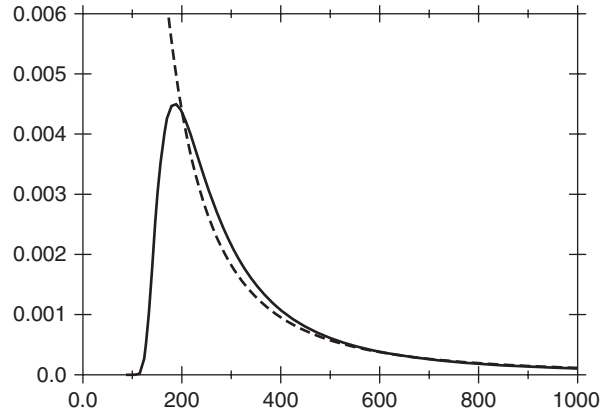


Figure 4. The probability densities of GEV (0.70, 230, 100) (line) and GPD (0.70, 150, 125) (dashes).

Table 5. Penalties Imposed by Financial Regulators (In Thousands of US Dollars)

3822	907	735	556	423	395	302	260	248	220	204	193	180	160	150
2568	845	660	550	417	360	297	255	239	220	202	191	176	157	147
1416	800	650	506	410	350	295	252	232	220	200	186	176	154	146
1299	750	630	484	406	350	275	251	230	215	200	185	165	151	143
917	743	600	426	400	332	270	250	229	211	194	182	165	151	143

Source: [12].

not to be identified with the population mean and standard deviation. For the GPD, μ is the lower bound of the range. Figure 4 shows the form of their respective probability density functions. Models with four and more parameters, such as Tukey's g-and-h class of distributions, are also gaining users: but they do require more data than is usually available. They have though been seen to capture the loss distribution of aggregated firm-wide losses. Readers are referred to Young and Coleman [3] for plots and properties of these and other models.

Example: Fitting GEV and GPD. This section follows the analysis in Young and Coleman [3, pp. 399–403], summarized in Coleman [11], for fitting the 75 losses given in Cruz [12, p. 83]. In Table 5, these data have been ordered and rounded to the nearest \$1000.

Figure 5 shows the sample cumulative distribution function (the observed proportion of values less than x) shown as steps, together with four fitted cumulative distribution functions (the height y is the probability of obtaining a future value less than x). The range of observation is (143, 3822). Figure 5 shows a good fit in each case. Table 6 shows

the values of four fitted models, and some large quantile values for each of them (the x values for given y values). For example, the 99th percentile $Q(0.99)$ is at a loss value of 3663 for the first model. The others are at 2794, 4457, and 2605. The 99.9th percentiles range from 7595 to 22,452. This $Q(0.999)$ is to be the basis of regulatory charging in banking, with $Q(0.995)$ for insurers. Estimation far outside a data set is always fraught. This can lead to significant errors in high quantile estimation. Quantiles give no information about how big a future loss larger than $Q(p)$ is likely to be. A measure used for this is the *mean excess*, also called *conditional value-at-risk* (*CVaR*). This computes the mean over the values greater than $Q(p)$ of the fitted model probability.

A simulation study of GPD (0.70, 150, 125) gave an approximate 95% confidence interval for $Q(0.999)$ of (5200, 9990).

A simulation of 4000 values from GEV(0.53, 230, 130) gave the estimates (0.50, 227, 126) for its parameters, emphasizing the need for large data sets.

The computations were made using Academic Xtremes, a computer package that accompanies Reiss and Thomas [13].

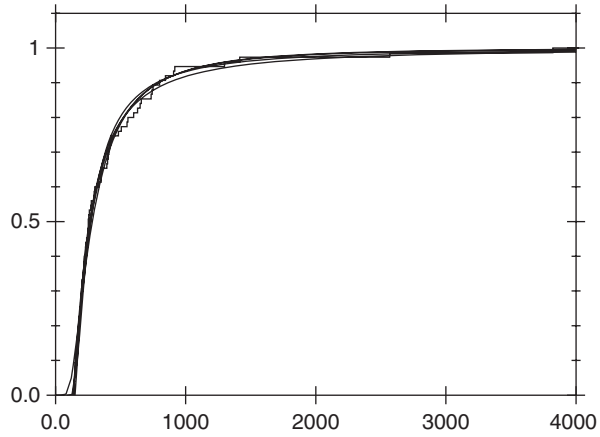


Figure 5. The sample cumulative distribution function (steps) with the four fitted models of Table 6. (Source: [3, p. 401].)

Table 6. The Parameters, Quantiles, and Fitted Values of the GEV and GPD Models when Fitted to Loss Data

Fitted model	GEV	GEV	GPD	GPD
<i>Parameter estimates</i>				
ξ	0.53	0.70	0.44	0.70
μ	230	230	135	150
σ	130	100	165	125
<i>Quantiles</i>				
$Q(0.9)$	777	793	866	793
$Q(0.95)$	1230	1169	1425	1161
$Q(0.99)$	3663	2794	4457	2605
$Q(0.995)$	5906	4046	7258	3619
$Q(0.999)$	18,066	9525	22,452	7595
<i>Data</i>				
	<i>Model values</i>			
1416	1614	1459	1902	1435
2568	2280	1925	2733	1857
3822	4842	3470	5929	3160

Source: [3, pp. 400–403].

Frequency Modeling

Standard statistical methods can be used to fit Poisson or negative binomial probability distributions to frequency counts. Experience shows that the daily, weekly, or monthly frequencies of loss events tend to occur in ways that cannot be fitted well by either of these models.

Combining Internal and External Data

When data from an external database are combined with an in-house data set, the former are relocated and scaled to match the latter. Then, we adopt the location and scale of the internal data. Finally, we check the

threshold used for the internal data against the new threshold for the relocated and rescaled external data, and set the larger of the two as the new threshold, eliminating all data points that fall below.

This is illustrated in Giacometti *et al.* [14, p. 7], where we see that it can lead to strange statistical results if the location and scale metrics are not chosen appropriately. A table shows a mean of pooled data having a mean smaller than the mean of both the internal and the external data. The same happens to the medians. Elsewhere another pooling shows skewness and a kurtosis of pooled data being larger than those of the separate

sets. We should perhaps be seeking a better standardization of operational risk loss data. In Reiss and Thomas [13], a logarithmic transformation of the data is taken prior to standardization. This corresponds to fitting a lognormal probability distribution to the data, but it did not entirely rectify the paradoxical result. A two-parameter extreme value theory model such as the standard Gumbel distribution should be a better choice for standardization. This choice has been applied in an environmental context relating different measures of air pollution [15].

CONCLUDING REMARKS

With risk being found in every enterprise, why does finance have such a prominent role in the study of it? Or does it? The theory of probability and statistics grew out of calculation of the odds in gambling. Knowing its true value gave an edge to any gambler. The mathematics was extended to set the price for life insurance, to determine the premium that should be charged for assuming the risk.

The financial services regulators make explicit in their rules the requirements of modeling, measuring, and using the results of the exercise. There are no prescriptive rules for the models that are to be used. The finance industry is still searching for the best practice. This is driving efforts at finding robust processes for operational risk, and evaluating methodologies for it. We must wait and see what emerges.

REFERENCES

1. Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards. Basel: Bank for International Settlements; 2006.
2. de Fontnouvelle P, De-Jesus V, Jordan J, *et al.* Capital and risk: new evidence on implications of large operational losses. Federal Reserve Bank of Boston. Working Paper No. 03-5. Available at www.bos.frb.org/bankinfo/oprisk/articles.htm.
3. Young B, Coleman R. Operational risk assessment. Chichester: John Wiley & Sons, Ltd.; 2009.
4. Reason J. Managing the risk of organisational accidents. Aldershot: Ashgate Publishing; 1997.
5. Elliott MR, Kleindorfer PR, DuBois JJ, *et al.* Linking OII and RMP data: does everyday safety prevent catastrophic loss? *Int J Risk Assess Manage* 2008;10:130–146.
6. Álvarez G. An operational risk management framework. In : Davis E, editor. Operational risk: practical approaches to implementation. London: Risk Books; 2005. pp. 227–236.
7. Cole RT. New challenges for operational risk measurement and management – a regulator's perspective. 2008. Keynote address. www.bos.frb.org/bankinfo/qau/conf/oprisk2008.
8. Selvaggi M. Analysing operational risk in insurance. ABI Research Paper 16. Association of British Insurers; 2009. www.abi.org.uk.
9. Williams C, Smith M, Young P. Risk management and insurance. 8th ed. New York: McGraw-Hill; 1998.
10. ORX. Available at www.orx.org/orx-data.
11. Coleman R. A VaR too far. *Capco Inst J Finan Trans* 2010;28:123–129.
12. Cruz MG. Modeling, measuring and hedging operational risk. Chichester: John Wiley & Sons, Ltd.; 2002.
13. Reiss R-D, Thomas M. Statistical analysis of extreme values. 3rd ed. Basel: Birkhäuser Verlag; 2007.
14. Giacometti R, Rachev S, Chernobai A, *et al.* Aggregation issues in operational risk. *J Oper Risk* 2008;3(3):3–23.
15. Heffernan JE, Tawn JA. A conditional approach for multivariate extreme values. *J Roy Stat Soc B* 2004;66:1–34.

OPERATIONS RESEARCH AND MANAGEMENT SCIENCE IN FIRE SAFETY

JOHN M. WATTS, JR
Fire Safety Institute,
Middlebury,
Vermont

Fire safety incorporates many different types of problems: sociological, technical, and managerial. In 1974, The Lanchester Prize, the most prestigious award offered by the Operations Research Society of America, went to the authors of a book describing how to reposition available fire companies in response to changing patterns of fire alarms [1]. This work was part of the comprehensive undertaking of the New York City/Rand (NYC/RAND) Institute to study the complexities, uncertainties, and multiple variables in the urban policy-making environment. Fire service deployment and emergency response remain a major area of application of operations research/management science (OR/MS) and are described in a separate article in this encyclopedia [2]. The broader concept of fire safety deals with encompassing aspects, such as combustibility of fuels, building construction, automatic detection and suppression systems, and occupant egress, for which the local fire department is only one component. The following are the types of OR/MS methods that have been employed among others:

- decision analysis,
- simulation,
- stochastic processes,
- networks,
- statistics, and
- economics.

WHAT IS FIRE SAFETY?

Fire has been described as one of the least understood common phenomena. Other animals have been found to use language

and tools, but use of fire distinguishes *Homo Sapiens* among that species that exist on the earth. Along with this use is its misuse, which constitutes a significant loss of life and property, mission failure, environmental destruction, and damage to our culture.

Aspects of fire safety include the chemi-cophysical nature of unwanted combustion, a weakly defined scientific process that can occur and be affected by a myriad of environmental conditions including ignition sources, types of combustibles, configurations of the fuel, and space in which it is burning, proximity to people and valuables, and the various response systems of fire detection and suppression.

Anyone looking to gain insight to the problem would do well to read the Scientific American book, *Fire*, by John Lyons [3], which provides an excellent description of the phenomenon and its implications to our society. The two-volume *Fire Protection Handbook*, published by the National Fire Protection Association [4], is an extensive compendium of basic approaches to fire safety in many aspects. The *Handbook of Fire Protection Engineering*, produced by the Society of Fire Protection Engineers, describes technical approaches to fire and fire safety that have evolved from our better understanding of the concept of fire safety science [5]. Section 5 of this Handbook covers the area of fire risk analysis, which employs some typical OR/MS approaches, including applications in health care, transportation, and consumer products.

While as early as 1965 Jarrett [6] identified aspects of fire safety that are amenable to OR/MS techniques, applications have been piecemeal. With the notable exception of the NYC/RAND project identified above, there has been little concerted effort to deal with fire safety in an operations research or management science approach. Areas of interest beyond municipal fire service deployment include wildland fires, process industries, and fire safety in buildings. This article introduces concepts of these applications and focuses on the mostchallenging issue: fire

safety in buildings.

WILDLAND FIRES

Among the more dramatic yet approachable fire safety problems are forest fires—or in the modern jargon, wildland fires—and in particular the interface of fire-prone wildland and encroachment of built environments, for example, the “urban–wildland interface.” Among the distinctions of this class of fires are a relatively homogenous “fuel load” (combustibles available to burn), a relatively limited (e.g., natural) variety of topography, and a major dependence on weather.

It is interesting to note that the third edition of Quade’s book, *Analysis for Public Decisions* [7], uses wildland fire [8] as an example of OR/MS modeling. The evolution of such models, within the US Forest Service and worldwide, has progressed measurably. An invited lecture was given by Frank Albini at the fifth triennial symposium of the International Association for Fire Safety Science [9]. This is an excellent introduction to the level of sophistication of current wildland fire models that integrate physicochemical aspects of fire spread into simulations of fire control. Applicable OR/MS methods also include allocation methods.

The significance of dealing with wildland fires has substantially increased as the extent of the wildland–urban interface has increased and greater incidence and severity seem correlated with global climate change. Significant research efforts are in progress in Australia and Southern Europe, parts of the world that have recently experienced devastating wildland fire events.

A prolific author of readable yet technical information on the subject is Pyne [10]. For a more complete description of the events and issues in the United States, the reader is referred to *The Wild Fire Reader: A Century of Failed Policies* [11]. A literary work worth reading is *Young Men and Fire* by Norman Maclean [12]. The first half of this book is a stirring tale of a wildland fire that occurred in 1947. The second half details the author’s 1990s investigation of the events that led to the deaths of 12 smokejumpers, which includes modern methods of wildland

fire modeling. It could be construed as an example of applied OR/MS.

INDUSTRIAL FIRE SAFETY

Under the guise of “process safety,” there are many OR/MS techniques that have been adapted to fire problems in the areas of aerospace industry, petrochemical facilities, and nuclear power generation plants. Some of these are described in the aforementioned *SFPE Handbook of Fire Protection Engineering* [5], and include

- probability models, such as stochastic fire growth;
- statistical models, such as extreme value;
- mathematical programming;
- cost–benefit analysis;
- networks such as fault trees; and
- reliability theory.

“System safety” techniques such as fault tree analysis evolved with the aerospace program where predictive models were needed in the absence of experience. See, for example, the *Systems Safety Handbook* [13]. The applicability of this approach was enhanced by the Rasmussen report on safety of nuclear facilities [14], also known as *WASH 1400*. This trend continued with standards such as MIL-STD 882C [15] and publications from the Center for Chemical Process Safety [16,17]. These techniques are not exclusively specific to fire but cover a multitude of potential safety hazards as well.

Many industry OR/MS safety system studies focus on reliability programs. The book *What Every Engineer Should Know about Reliability and Risk Analysis* uses a fire safety system as an example application [18, pp. 318–328].

Industry and insurance have motivated some of the better applications of OR/MS to fire safety economics. This can be traced back as least as far as Mandelbrot’s 1964 article in *Operations Research* [19]. The paper by Shpilberg and Neufville that appeared in *Fire Technology* in 1974 [20] is a classic example of a decision tree and utility theory approach to fire safety. It relates hardware alternatives for fire protection and insurance

deductibles in a decision tree model. This work and similar approaches are discussed in the book *Fire Protection Economics* by Ramachandran [21]. This is perhaps the only monograph that specifically addresses fire safety decision-making techniques.

Economic issues involving very tall buildings have taken on multifarious aspects that were formerly associated only with the more complex systems of industrial situations. New construction techniques, security concerns, and some deep pockets have created need for modeling the economics of alternative approaches to hazard mitigation. The National Institute for Standards and Technology (NIST) has recently produced and distributed a computer model to address these issues [22].

FIRE SAFETY IN BUILDINGS

The issues associated with fire safety in buildings fall within a class of problems that have many unexpected and little understood characteristics of a “complex system,” defined as a high order, multiple loop, nonlinear feedback structure. Feedback structure is common to all types of decisions, but feedback in complex systems has some special characteristics.

“Order” of a system refers to the number of state variables that define the system: for example, fire temperature, heat release, smoke concentration, spatial geometry, occupant density, and location. Complexity increases rapidly with the order of the system, and in fire safety we have a myriad of factors that affect the outcome of a fire scenario.

Complex systems have multiple feedback loops. Primary feedback loops in building fires are thermal energy of the fire, smoke, and toxic products of combustion, and the interaction of these thermodynamic conditions with people in the building. The evolution of combustion physics and computational fluid dynamics (CFD) modeling is moving us toward better understanding of mass (smoke and toxic gasses) transport in different types of buildings.

Most mathematics deals with linear processes. Fire and human behavior exhibit a great deal of nonlinearity. Multiple loop

realignment along nonlinear functions makes the complex system highly insensitive to most system parameters and thus resistant to efforts to change system behavior. In fire safety, we do not yet have a solid feel for the significance of the observed nonlinearity or how to deal with it.

Constraints on the fire safety system are often of a political nature in addition to fixed architectural or physical restraints. There are limitations imposed by construction practices and fire protection hardware technology. And marketing influences may come into play. But the most significant constraint is the incoherent agglomeration safety standards known as building codes. Building codes, along with other fire safety codes and regulations, are intended to be minimum standards for effective fire safety. However, as was demonstrated by Howard Emmons, the eminent Harvard professor of mechanical engineering, over 40 years ago [23], aspects of these requirements, when subject to scientific scrutiny, are found to be unnecessarily overbearing and hence may constrain fire safety rather than promote it.

One way to approach the analysis of complex systems is by decomposition into component parts. By understanding the behavior of simpler subsystems, we can gain insight into the expected behavior of the complex system. As this decomposition continues into smaller and smaller components, the diversity of knowledge brought to bear expands greatly. This is reflected in the wide range of details that have been considered in dealing with fire safety. But reality is not rigidly compartmentalized. We risk further loss of control if the complex system is not dealt with in its totality. This is substantially the current state of the art of building fire safety. The synthesis step is still missing.

Most readers of this article would have their scientific sensibilities offended by the fire safety codes and standards used to dictate building construction and occupancy. Inroads to more systemic fire safety evaluation have evolved in the last decade under the sobriquet of “performance” engineering, for example, Society of Fire Protection Engineers [24]. While still in a neonatal stage, the concept of building design based on fire performance evaluation is encouraging.

FIRE SAFETY IN BUILDINGS: AN APPLICATION

Fire risk indexing is an aggregation of system variables into a single index to reflect an ordinal evaluation of fire safety in a building. The approach uses many tools and techniques of OR/MS, including decision analysis, Delphi, multiattribute evaluation, analytic hierarchic process (AHP), scaling techniques, utility theory, hierarchical methods, decision tables, and careful thought. The following description and process is adapted from Chapter 13 in the book *Evaluation of Fire Safety* [25].

In general, a fire risk index assigns values to selected attributes based on professional judgment and experience. The selected variables represent both positive and negative fire safety features and the assigned values are then operated on by some combination of arithmetic functions to arrive at a single value or index. This value can be compared with other similar assessments or to a standard.

Fire risk indexing is not a substitute for detailed theoretic fire safety evaluation. It is a planning tool useful for screening, ranking, and setting priorities. Fire risk indexing systems come in a large variety of formats and with a broad spectrum of purposes. Reviews of example fire risk indexes may be found in the *SFPE Handbook of Fire Protection Engineering* [26].

As has been pointed out in this article, fire safety evaluation involves the analysis of many complex factors that are difficult to assess in a uniform and consistent way. Defensibility, both internally and externally, is one of the strongest assets of a scientifically constructed index system. Internal defensibility provides management with justification for fire safety policies and expenditures. It facilitates the allocation of limited resources among fire and other risks. External justification of priority-setting is important in litigation and in dealing with regulatory agencies.

Fire risk indexing is a heuristic model of fire safety. It is a process of modeling and scoring fire hazard and exposure factors to produce a rapid and simple estimate of

comparative evaluation. The process heuristically relates known fire safety attributes that have varying degrees of accuracy in their measurement. The distinct advantage is a complete model that does not depend solely on assumptions about unsubstantiated physical processes.

For many situations where a quantitative fire safety evaluation is desirable, an in-depth theoretic analysis may not be cost effective or appropriate. This could be the fundamental case where great sophistication is not required, where prioritization is the principal objective, or where it is necessary to institutionalize an approach to fire safety evaluation for a wide base of use. The objective of fire risk indexing is to sacrifice details and individual features for the sake of making the assessment easier. Information should be reduced to a single score even in the most complex applications. OR/MS techniques have been espoused to combine technical, economic, and sociopolitical factors.

Because of our limited phenomenological knowledge, the accuracy demanded for a fire safety evaluation is different from other engineering purposes. Often, establishing an order of magnitude will suffice. Time and resource expenditure increases as the depth of analysis is increased. In an age where resources are scarce, and efficiency is prized, maximizing the utility of a fire risk indexing approach to fire safety evaluation is clearly desirable for the many situations where the evaluation of fire safety is fundamental.

Perhaps the most common implicit justification of fire risk indexing is the need for a simplistic process of fire safety evaluation. In most applications, the target of fire risk indexing is a broad class of products or facilities for which a detailed fire risk analysis of each individual case is not feasible. Fire risk indexing has an appeal to administrators charged with risk management decision-making responsibilities but who may be unfamiliar with the details and mechanics of the fire risk assessment process. Widespread implementation of a generalized approach to fire risk is contingent on its appeal to a broad class of users including architects, building officials, and property managers.

Multiattribute Evaluation

By nature of the circumstances, fire safety decisions often have to be made under conditions where the data are sparse and uncertain. The technical parameters of fire safety evaluation are very complex and normally involve a network of interacting components, the interactions generally being nonlinear and multidimensional. However, complexity and sparseness of data do not preclude useful and valid approaches. Such circumstances are not unusual in decision making in business or other risk venues and, unless such problems are addressed, developments that could be useful to society may be inhibited. The space program illustrates how success can be achieved when there is little relevant data. One applicable approach to fire safety evaluation is multiattribute evaluation.

The importance of multiattribute evaluation to fire safety is that there are several well-used and accepted approaches to fire risk indexing that do not have any credible structure to their development. Hence, there is no transparency or technical validity, yet they are used to make life threatening decisions.

As implied above, fire safety decisions require more than one attribute to capture all relevant aspects of the consequences. If the attributes for a decision problem are $x_1, x_2, x_3, \dots, x_n$, then an evaluation function $E(x_1, x_2, x_3, \dots, x_n)$ needs to be determined over these measures in order to conduct a performance assessment. However, determining an appropriate evaluation function over multiple performance measures is a complex problem. Subjective attribute values must be elicited by asking questions, and it seems difficult to answer questions that will directly determine an n -dimensional function.

If trade-offs among the attributes do not depend on the levels of the remaining attributes, then a single measure of the overall outcome of a system is given by

$$E(x_1, \dots, x_i, \dots, x_n) = \sum_{i=1}^n w_i R_i(x_i),$$

where w_i are weighting constants greater than zero and $R_i(x_i)$ are normalizing functions of the attributes.

Evaluation of fire safety can be a difficult endeavor. Many, sometimes conflicting, attributes must be juggled simultaneously. The field of OR/MS has long dealt with this type of problem. There is a large body of literature on the subject of multiattribute evaluation, also known variously as *multiattribute decision analysis*, *multicriteria decision making*, and *multiattribute utility theory*. These methods apply to problems where a decision maker must evaluate, rank, or classify alternatives characterized by two or more relevant attributes. Summaries and descriptions of the principal methods are in other articles in this encyclopedia.

Multiattribute evaluation begins with the generation of attributes that provide a means of evaluating goal achievements. These attributes, referred to in some applications as parameters, elements, factors, variables, and so on, identify the ingredients of fire safety.

Fire safety attributes are components of fire risk that are quantitatively determinable by direct or indirect measurement or estimation. They are intended to represent factors that account for an acceptably large portion of the total fire risk. Usually they are not directly measurable. This is especially true for existing buildings where only limited information is readily available. Thus, attributes may be either quantitative or qualitative, and both types of attributes are very important. Selection of attributes should result in a set that is nonconflicting, coherent, and logical.

ATTRIBUTE GENERATION

Fire safety is a complex system with an inordinately large number of factors that may affect it. These can range from ignitability of personal clothing to availability of a heliport for evacuation. Computational and cognitive feasibility limit consideration to only a relatively small number of these variables.

It is intuitively appealing to postulate that safety from fire is a Paretian phenomenon in that a relatively small number of attributes account for most of the problem. This is supported by general fire loss figures, which suggest that a small number of factors are

associated with a large proportion of fire deaths. It is necessary, then, to identify as attributes some defensible combination of factors that account for an acceptable portion of the fire risk. A desirable list of attributes should be as follows:

- *Complete and exhaustive.* That is, all important attributes should be represented.
- *Mutually exclusive.* Independence of the attributes facilitates evaluation of trade-offs.
- *Restricted to highest degree of importance.* Lower level criteria may be part of attribute rating discussed later in this section.

A study of fire safety effectiveness statements conducted for the US Fire Administration focused on logical and reproducible means of identifying key life safety variables [27]. Following is the list of 19 key life safety variables or attributes identified as a result of this study. The attributes are placed into four groupings for convenience and clarity.

Fire development

- Amount of combustibles
- Rate of heat release
- Toxicity of combustion products
- Obscuration by smoke

Fire spread

- Fire resistance of structural members
- Fire/smoke barriers
- Fire resistance of vertical shafts
- Isolation of hazardous functions/materials

Fire control

- Automatic extinguishing system
- Automatic smoke control
- Systems maintenance
- Suppression by in-house staff
- Suppression by municipal fire department

Egress

- Exit way dimensions

- Remoteness and independence of exits
- Height of building
- Automatic fire detection and alarm system
- Physiological/psychological condition of occupants

This detailed process is not typical for most fire risk indexing systems. Selection of attributes is usually more arbitrary, with correspondingly disparate results. Where the subject of attribute generation is addressed, it often involves an adapted Delphi exercise. Operational considerations of Delphi applications to multiattribute evaluation in fire safety have been addressed in the literature. Donegan [28] discuss Delphi briefly among other approaches to subjective measurement in fire protection engineering. Development of attributes relies heavily on intuition and subjective judgment. In the final analysis, it is most important that the evaluation vector include only those attributes that significantly vary among buildings and for which the variation is considered meaningful.

Attribute Weighting

Not all fire safety attributes are equally important. The role of weight serves to express the importance of each attribute compared with the others. Hence the assignment of weights is a key component of multiattribute evaluation. Although assigning weights by an ordinal scale is usually easier, most multiattribute evaluation methods require cardinal weights. The attribute weights are generally normalized to sum to one, that is, if y_i is the raw weight of attribute I , then

$$w_i = \frac{y_i}{\sum_{i=1}^n y_i}$$

and

$$\sum_{i=1}^n w_i = 1.$$

This produces a vector of n weights given by

$$W = [w_1, \dots, w_i, \dots, w_n],$$

where w_i is the resultant weight assigned to the i th attribute.

This relative importance of attributes is defined to be constant across building evaluations. There are many weight assessment techniques used in multiattribute evaluation. Generally, hierarchical methods have been found effective in fire safety evaluation.

Attribute Ratings

Each attribute weight represents a specific relative importance that is universal for all facilities within the scope of the evaluation method. Individual buildings will vary in the degree to which attributes exist or occur in a space. Attribute ratings or grades are a measure of the intensity level or degree of danger or security afforded by the attribute in a particular application.

The selected attributes may be either quantitative or qualitative. Because both types of attributes are very important, we need a method that accounts for attributes of a general nature. Qualitative attributes may be impractical, impossible, or too costly to measure directly. Likert scaling is most often used to empirically capture the essential meaning of the attribute and develop a scale upon which a surrogate measure or rating can be based.

Quantitative attributes are readily measured or quantified but may require judgment to convert into a compensatory measure. Each quantitative attribute typically has a different unit of measurement. Since multiattribute evaluation scoring generally requires a homogenous data type, data transformation techniques become necessary. Quantitative attribute ratings must be normalized to a scale that is common to all attributes.

Likert Scaling

Scaling and scale construction are of central concern in the measurement of any phenomena. This includes objective conditions as well as subjective states. Scaling identifies each individual object so that valid and reliable differences among objects can be represented.

No scaling model has more intuitive appeal than the Likert scale. Likert scaling rates attributes as 1, 2, 3, 4, or 5, reading

from unfavorable to favorable. A 5-point Likert scale is used predominantly, but a more detailed scale such as a 7-point or 10-point scale can also be used if its application does not create undue cognitive stress.

There are three underlying assumptions of Likert scaling:

- Each item is monotonically related to the underlying latent dimension continuum, that is, there is no obvious discontinuity or reversal of slope.
- The sum of the item scores is monotonic (and approximately linear) with respect to the dimension measured.
- The items as a group measure only the dimension sought. In other words, all items to be linearly combined should be related only to a single common factor. The sum of these items is expected to contain all the important information contained in the individual items.

Only this last assumption tends to be somewhat problematic. It is difficult to decide conclusively that the items as a whole are measuring only a single phenomenon.

In an experiment on Delphi methodology, the equivalence of three simple scaling techniques was examined and it was concluded that for practical purposes the results could be accepted as an interval scale. The experiment showed that Likert scales used in Delphi exercises have the equal difference property and therefore combinatorial calculations are appropriate. The intervals between scores on a Likert scale are meaningful but ratios of scores are not interpretable.

Normalization of Data

Typically, each quantitative attribute has a different unit of measurement. In order to attain the compensatory trade-offs that are an essential characteristic of multiattribute evaluation, commensurable attribute units are necessary. Attribute ratings are therefore normalized to eliminate computational problems caused by differing measurement units in the evaluation vector. Thus, we need to construct or adopt a normalizing function $R_i(x_i)$ for each attribute i .

Fire safety attributes may be beneficial, detrimental, or nonmonotonic. Beneficial attributes offer monotonically increasing utility: the greater the attribute value, the more its preference, for example, fire resistance. Detrimental attributes are monotonically decreasing in utility: the greater the attribute value the less its preference, for example, rate of heat release. Nonmonotonic fire safety attributes are uncommon. One, perhaps unique, example is floor level, where, for life safety, grade level is preferred over stories above or below grade.

The most common form of normalization is linear. For beneficial attributes, the normalized rating of attribute i , r_i , is given by

$$r_i = \frac{x_i - x_i^\vee}{x_i^\wedge - x_i^\vee},$$

where $x_i^\wedge$ is the ceiling or maximum possible value of x_i , and $x_i^\vee$ is the floor or smallest possible value of x_i . Thus the expression $x_i^\wedge - x_i^\vee$ is the range of all possible values of attribute x_i . If, as often happens, $x_i^\vee = 0$, then the normalized rating is given by the ratio of the attribute value to the maximum value, $r_i = x_i/x_i^\wedge$. The resultant ratings have the characteristic that $0 \leq r_i \leq 1$ and the attribute is more favorable as r_i approaches 1.

The linear normalized rating of a detrimental attribute i is given by

$$r_i = \frac{\hat{x}_i - x_i}{\hat{x}_i - x_i^\vee}.$$

Again, the resultant ratings have the characteristic that $0 \leq r_i \leq 1$ and have been adjusted to be consistent with beneficial attributes so that the attribute is more favorable as r_i approaches 1.

There are statistical procedures for normalizing nonmonotonic attributes, but they are typically not necessary in fire safety evaluation. Where quantitative and qualitative attributes are mixed, the normalized ratings should be multiplied by the modulus of the Likert scale used for the qualitative attributes. For example, if a 5-point Likert scale is used for qualitative data, then the normalized ratings should be multiplied by

5. This is essential to maintain the compensatory capability of the scoring method.

Decision Tables

Fire safety attribute rating can be facilitated by partitioning the attributes into measurable constituent parts. Usually, these parts will be directly measurable survey items. Sometimes, there may also be intermediate subattributes. A survey item is a measurable feature of a building or building space that serves as a constituent part of one or more attributes or subattributes.

Decision tables are commonly used in decision analysis and documentation. Their purpose is to provide orderly representation of information flow in elementary decisions. While such decisions can appear simple by relative comparison, their underlying logic may often be complex. The tabular approach is used to express decision logic in a way that encourages reduction of a problem to its simplest form by arranging and presenting logical alternatives under various conditions. Decision tables have been shown to present a useful logic for using survey items to develop ratings for fire safety attributes [29]. Three properties of decision tables that contribute to their effectiveness in fire safety evaluation are the following:

- Disciplined way to rate fire safety attributes using data on survey items.
- Concise and standardized documentation of the detailed design of a fire risk index.
- Simple transition to computerized application.

Analytic Hierarchy Process (AHP)

Ratings for fire safety attributes can also be developed using pairwise comparisons and the AHP. AHP has been widely reviewed and applied in the literature and its use is supported by several commercially available user-friendly software packages. In this approach, the relative importance of each survey item or subattribute is determined by setting up a square matrix A and making pairwise comparisons. Relative importance of each item can then be calculated

from the matrix using any of several methods. The best known and most supported by commercial software is the eigenvalue prioritization method.

When using pairwise comparisons and AHP, the number of fire safety subattributes or survey items for each attribute should be limited to around 7. This number is congruous with the theory that 7 ± 2 represents the greatest amount of information that an observer can give us about an object from an absolute judgment [30]. If the practical range of the attribute ratings is not known, AHP can be subject to distortion and rank reversal, although this is seldom the case in fire safety evaluation.

Scoring

A multiattribute evaluation may be viewed as a vector of attributes. The transformation of a vector to an appropriate scalar value is the purpose of the fire risk index, that is, formulating an index to represent the effectiveness of the system. In fire safety, we do not yet have a thorough understanding of the functional relationships among components, so simple heuristic scoring techniques are used.

Additive weighting is the most widely used method in fire risk indexing. The score is determined by adding the contribution from each attribute. Since two items with different measurement units cannot be added, a common numerical scaling system such as normalization, discussed in the previous section, is required to permit addition among attribute values. The total score for each evaluation then can be computed by multiplying the comparable rating for each attribute by the importance weight assigned to the attribute and then summing these products over all the attributes. Formally, the evaluation score S in the additive weight method can be expressed as

$$S = \sum_{i=1}^n w_i r_i,$$

where w_i is the weight of attribute i and r_i is the normalized rating of attribute i . Thus, the numeric evaluation is calculated as the scalar product of the weighting vector and rating vector of the attributes.

The underlying assumption of additive weighting is that attributes are preferentially independent. Less formally, this means that the contribution of an individual attribute to the total (multiattribute) score is independent of other attribute values. Therefore, preferences regarding the value of one attribute are not influenced in any way by the values of the other attributes. Fortunately, studies show that additive weighting yields extremely close approximations to “real” value functions even when independence among attributes does not exactly hold.

Additive weighting permits very lenient assumptions about individual components of a scale. This suggests that, because each item may contain considerable measurement error or specificity, the importance of this additive model is that it does not take any particular item very seriously. Additive weighting also assumes that the characteristic weights are proportional to the relative value of a unit change in each attribute’s value function. This is what makes the fire risk index compensatory.

Application Observations

Fire risk indexing has gained some acceptance because of its high utility and relative ease of application. Fire safety evaluation involves a large number of multifarious factors that are hard to assess in a uniform and consistent way. Analysis of such a complex system is difficult but not impossible as evidenced by activities in the fields of nuclear safety and environmental protection. Detailed risk assessment can be an expensive and labor-intensive process and there is considerable scope for improving the presentation of results. Fire risk indexing can provide a cost-effective means of fire safety evaluation that is sufficient in both utility and validity.

For analytical and procedural simplicity, common practice neglects both uncertainties and imprecision inherent in the evaluation vector data and in the additional elicited information about the attributes and objects. Neglect of uncertainty occurs when uncertain values are represented by their expected values rather than by probability distributions. Neglect of imprecision occurs when

ratings such as “good” and “bad” are converted to scalar numbers rather than ranges. All applications to fire safety follow this practice. Chance or random error is involved in any type of measurement. However, as Nunnally [31, p. 67] observes, “this unreliability averages out when scores on numerous items are summed to obtain a total score, which then frequently is highly reliable.”

Fire risk indexing tries to obtain a meaningful index from multidimensional data to evaluate fire safety. They are relatively simple models that do not rely exclusively on demonstrated principles of fire dynamics and OR/MS. But their credibility is enhanced to the extent that they do employ such principles.

CONCLUSION

Over its relatively short life, OR/MS has become known for the use and development of quantitative modeling tools and techniques such as described elsewhere in this encyclopedia. All of these are applicable to issues of fire safety. But, while application and refinement of algorithmic methods is a visible and useful component of OR/MS, it is the tip of the problem-solving iceberg. The broader aspect of operational problems includes the amorphous complexity of the real world. This requires problem formulation, evolving goals and objectives, setting boundaries for the work, designing attractive alternatives for study, screening them, considering unquantifiable (and quantifiable) aspects, and evaluating the resulting outcomes. Perhaps, most importantly, it requires a focus on the problem that is unhampered by preconceived notions of what the solution should be.

Potential OR/MS applications include greater use of techniques such as network flows, for example, system dynamics; stochastic programming of combined fire and human dynamics; reliability and maintainability of fire safety systems; and simulation modeling and analysis of complex interactions of variables including rare events. Of most promise is the use of risk analysis to model the relative value of input from fire physics and statistical analyses of fire losses to modify the

experienced judgment upon which today’s fire safety codes and standards are mostly based.

Often, the issue of fire safety appears too big and too complex for us to comprehend in its entirety. Our solution is to attack bits and pieces of the problem through laboratory research and standards writing. We do not often try to look at the larger picture because that perspective is too difficult. Many tools of OR/MS are designed to deal with large complex problems. A broader perspective may reveal more efficient approaches to fire safety. The late ORSA President Hugh Miser was wont to quote Schön [32] in regard to seemingly intractable problems such as fire safety:

In the varied topography of professional practice, there is a high, hard ground overlooking a swamp. On the high ground, manageable problems lend themselves to solution through the application of research-based theory and technique. In the swampy lowland; messy, confusing problems defy technical solution. The irony of this situation is that the problems of the high ground tend to be relatively unimportant to individuals or society at large, while in the swamp lay the problems of greatest human concern.

To date, there has not been much concerted effort to apply OR/MS to fire safety. In particular, there has been noticeably little OR/MS endeavor on structural fires, where lives are lost, property is destroyed, missions are aborted, and great damage is done to our environment and our cultural heritage.

REFERENCES

1. Walker WE, Chaiken JM, Ignall EJ, editors. Fire department deployment analysis. New York: North Holland Publishing Company; 1979.
2. Hall JR. Fire department deployment and service analysis. Encyclopedia of operations research and management science. New York: John Wiley; 2010.
3. Lyons JW. Fire. New York: Scientific American Library; 1985.
4. Cote AE, editor. Fire protection handbook. Quincy (MA): National Fire Protection Association; 2007.
5. Society of Fire Protection Engineers. Handbook of fire protection engineering. 4th ed.

- Quincy (MA): National Fire Protection Association; 2008.
6. Jarrett H. The systems approach to fire protection problems: an overview. *Fire Technol* 1965;1(3):182–187.
 7. Quade ES. In: Carter GM, editor. *Analysis for public decisions*. 3rd ed. New York: Prentice Hall; 1989.
 8. Jewell WS. Forest fire problems: a progress report. *Oper Res* 1963;11:678–692.
 9. Albini FA. An overview of research on wildland fires. In: Hasemi Y, editor. *Fire Safety Science: Proceedings of the Fifth International Symposium*. Melbourne: International Association for Fire Safety Science; 1997.
 10. Pyne SJ. *Introduction to wildland fire*. New York: John Wiley; 1984.
 11. Wuerthner G, editor. *The wild fire reader: a century of failed policies*. Washington (DC): Island Press; 2006.
 12. Maclean N. *Young men and fire*. Chicago: University of Chicago Press; 1992.
 13. Hansen M. *System safety analysis handbook*. 2nd ed. Albuquerque (NM): Systems Safety Society; 1997.
 14. Rasmussen NC. *Reactor safety study: an assessment of accident risks in U.S. commercial nuclear power plants*. Report no WASH-1400 (NUREG-75/014). Washington (DC): US Nuclear Regulatory Commission; 1975.
 15. MIL-STD-882 C. *System safety program planning*. Washington (DC): Department of Defense; 1993.
 16. Center for Chemical Process Safety. *Tools for making acute risk decisions with chemical process safety applications*. New York: AIChE; 1995.
 17. Center for Chemical Process Safety. *Guidelines for chemical process quantitative risk analysis*. New York: AIChE; 1989.
 18. Modarres M. *What every engineer should know about reliability and risk analysis*. New York: Marcel Dekker; 1993.
 19. Mandelbrot B. Random walks, fire damage amount, and other Paretian risk phenomena. *Oper Res* 1964;12:582–585.
 20. Shpilberg DC, Neufville RD. Best choice of fire protection: an airport study. *Fire Technol* 1974;10:5–14.
 21. Ramachandran G. *The economics of fire protection*. London: Spon; 1998.
 22. Thomas DS, Chapman RE. *A guide to printed and electronic resources for developing a cost-effective risk management plan for new and existing constructed facilities*. NIST Special Publication 1082. Washington (DC): Building and Fire Research Laboratory, US Department of Commerce; 2008 Jul.
 23. Emmons HW. Fire research abroad. *Fire Technol* 1967;3(3):225–231.
 24. Society of Fire Protection Engineers. *SFPE engineering guide to performance-based fire protection*. 2nd ed. Quincy (MA): National Fire Protection Association; 2007.
 25. Rasbash DG, Ramachandran G, Kandola B, *et al*. *Evaluation of fire safety*. New York: John Wiley; 2004.
 26. Watts JM Jr. Fire risk indexing. In: DiNenno PJ *et al.*, editors. *SFPE handbook of fire protection engineering*. 4th ed. Quincy (MA): National Fire Protection Association; 2009. Section 5, Chapter 10.
 27. Watts J, Milke JA, Bryan JL, *et al*. *A study of fire safety effectiveness statements*. College Park (MD): Department of Fire Protection Engineering, University of Maryland; 1979. (NTIS: PB-299169).
 28. Donegan HA. Decision analysis. In: DiNenno PJ, *et al.* editors. *SFPE handbook of fire protection engineering*. 4th ed. Quincy (MA): National Fire Protection Association; 2008. Section 5, Chapter 2.
 29. Watts JM Jr, Budnick EK, Kushler BD. Using decision tables to quantify fire risk parameters. *Proceedings of the International Conference on Fire Research and Engineering*. Boston (MA): Society of Fire Protection Engineers; 1995.
 30. Miller GA. The magic number seven, plus or minus two. *Psychol Rev* 1956;63:81–97.
 31. Nunnally JC. *Psychometric theory*. New York: McGraw-Hill; 1978.
 32. Schön DA. *Educating the reflective practitioner*. San Francisco (CA): Jossey-Bass; 1987.

OPERATIONS RESEARCH APPLICATIONS IN TRUCKLOAD FREIGHT TRANSPORTATION NETWORKS

TED L. GIFFORD
Schneider National, Inc.,
Green Bay, Wisconsin

This article describes a set of decision support systems based on operations research (OR) methodologies that arise in the planning, management, and operations of a large truckload transportation company. These tools support policy development activities and business processes that range across the time horizon from longer term planning cycles to real-time driver dispatching and load management decisions. This discussion will highlight problems and solution models that appear to be either unique to truckload freight transportation or of particular interest in this area.

We draw examples from Schneider National Inc., one of the nation's largest transportation and logistics companies. Although the company's activities are much broader than *truckload* freight, encompassing intermodal operations, supply chain management and engineering, freight management and forwarding, and third-party brokerage, we restrict our examples to the full-truckload segment of the operations. In 2008, this group operated a fleet of 14,000 tractors and 48,000 trailers with a workforce of 15,500 drivers. This freight transportation network spans the contiguous United States and many provinces of Canada. A significant amount of freight also originates or terminates in Mexico and is transferred to/from Mexican carriers at border stations. The freight volume for the year approached two million loads and the total distance traveled was well over 1.5 billion miles. Fleet operations are supported by 15 regional operating centers which include maintenance, refueling, and driver rest facilities. Freight that is carried by these various groups

comes from thousands of distinct customers, although a significant portion originates with a relatively small group of high-volume, consistent, repeat-business customers. Most pricing is determined by long-term agreements, although a significant and growing percentage of pricing is negotiated as part of the tendering process in a daily spot market environment. Almost all freight is tendered within a few days of the required pickup time.

Profitable, efficient management of an enterprise of this size and complexity requires extensive planning processes and tools as well as an array of decision support systems that are based on various models drawn from OR and management science (MS). The review article, "Planning models for freight transportation" [1], provides a good overview of a broad range of models that are common across many transportation modes and describes a commonly-used framework for categorizing these processes and models by time frames. *Strategic planning* encompasses decisions made in a multiyear time frame. These include determining markets, customer demographics, and geographic regions to either increase or decrease focus on, allocating capital funds and resources among various operating groups, deciding whether to open, close, or relocate operating centers based on changes in freight flows, and making adjustments to target levels for driver hiring. *Tactical planning* focuses on those decisions that tend to be durable for one to several quarters. These include specific customers interactions such as contract bid prices and capacity commitments, targeted hiring of drivers from specific domicile regions, schedules for equipment purchase and resale, setting targets and policies for maintenance of trailer pools. *Operational planning* occurs in a time window ranging from a day to one or two weeks. Work in this area is centered on decisions that drive the efficiency and profitability of daily network operations. The most important of these fall into two areas; freight selection—acceptance, targeted solicitation, appointment schedul-

ing; and asset repositioning—empty moves of tractors and trailers to adjust resources to respond to anticipated supply–demand imbalances. We use the term *execution* to denote the daily activity of matching drivers/tractors and trailers to specific loads, monitoring freight in transit, adjusting to changes brought on by such things as weather conditions, accidents, driver emergencies, and customer requests, and managing additional requirements for equipment repositioning and appointment rescheduling.

CONTRAST WITH OTHER TRANSPORTATION MODES

There are some key differences between *truckload* transportation network operations as described here and most other freight transportation *modes*. These other modes include air, railroad, ocean, as well as less-than truckload (LTL) and parcel trucking. The notable distinction between truckload transportation and these other freight transport modes is that the *tactical planning* phase in these modes normally includes the *explicit* determination of routes and schedules for the specific *vehicles* (planes, trains, boats) which constitute the freight transportation network. These routes and schedules are set in advance of knowledge of the realized freight flows and remain relatively static across a predefined horizon. A brief description of air networks (both freight and passenger are similar) will serve to illustrate the general approach for these networks. Forecasts are generated to estimate expected demand over a horizon that typically ranges from three to six months, and also account for expected revenue, costs, and resulting profitability. Using these forecasts as problem parameters, the network design problem of determining routes, route capacities, and schedules is addressed directly, often as a network flow problem.

After setting the structure and capacity of the network, the *operational* problem that remains is to maximize utilization and profitability in response to actualized demand as it arrives. Schedules are published months in advance so that the number of available seats for each route and time slot is fixed. Airlines

then apply sophisticated *revenue management* optimization models to segment customers by buying behavior and adjust prices as a function of booking patterns (accumulating demand), remaining capacity, and time left to departure. These techniques are now notably visible to the general public as on-line airline ticket booking allows customers to see prices change in a matter of minutes. LTL and small package shipping companies also maintain static network routes and schedules, but respond to demand variability by adjusting the number of short trailers that are pulled by a single tractor. Similarly, railroads can vary the number of freight cars and locomotives on scheduled trains to account for actualized freight.

This *two-stage* methodology of *tactically* determining network routing and scheduling and then *operationally* filling out its capacity does not apply for a large *truckload freight* common-carrier. This is driven by the fact that freight is generated by a diverse set of customers, who typically provide short lead times for specific freight movements with service conditions that require shortest-path point-to-point transit. The tractor and sometimes the trailer are now available for redeployment at a time and location that could not be anticipated in advance of *operational planning* and *execution* activity.

MODELING A TRUCKLOAD TRANSPORTATION NETWORK

From a planning and execution management perspective, we should also make clear what we mean by *the network* and how, using OR/MS terminology, we define the *nodes* of our network. As there are well in excess of 250,000 actual *point-to-point* origin and destination locations visited by the tractor fleet in a year, any model that incorporated these directly would clearly be intractable. As such, we subdivide the country into approximately 500 nonoverlapping regions which then become the *operative* locations for modeling and analysis. Postal ZIP3 zones are commonly used as a basis to construct these. Corresponding distances and transit durations are generated by accessing commercial

software packages that provide distance calculations based on detailed information of road and highway networks.

As noted above, truckload freight networks generally exhibit an inherent randomness as the next destination for an asset may not be determined until a short time before the scheduled departure as several or perhaps many different new shipment opportunities may be available for an asset (including moving empty to a better location) when it becomes available at a new place and time. A fundamental concept for analysis of these dynamic, random networks is quantifying the opportunity value of a resource asset at a particular location in the network at the point in time when it becomes available there. As a quantity, this is usually viewed as an estimate of the profit, which can be generated over a specified finite time horizon by the asset looking forward. The quantity is commonly referred to as the *network value* of the asset and is specified by a network location and point (or interval) in time. It may, additionally, be parameterized by the number of similar resources that may arrive concurrently and by attribute values that characterize subclasses of assets. Several of the models and methods described below include as a critical component, the calculation of *network value*.

STRATEGIC PLANNING

Strategic planning problems faced by a large truckload transportation company are generally similar to those encountered in the other freight transportation mentioned above. These include the following:

1. locating and sizing Operating Centers, which provide facilities for equipment maintenance, refueling, and driver services;
2. developing equipment acquisition and disposal schedule and policies. This encompasses determining quantities and types of tractors, trailers, and containers. Decisions on when to retire assets must be based on analysis of maintenance costs, regulatory compliance, and prevailing resale or salvage values;
3. implementing driving hiring and training policies to include setting targets for workforce size, training programs, and determining the desired demographic mix of drivers;
4. negotiating long-term contracts for fuel prices and fuel hedging strategies;
5. entering into bid agreements with customers for longer term pricing and capacity commitments.

Of the above, only the third appears to present a unique challenge to long-haul trucking, which we describe in the following section. For the other areas, Crainic and Laporte [1] provide an introduction to strategic planning problems that apply across a range of transportation modes and a broad range of references.

TACTICAL PLANNING

Among the planning problems and operational policy decisions, there are several that appear to be unique to large-scale truckload transportation networks. In particular, the *truck driver* is a critical resource and subject to unique constraints and cost factors. The industry average for driver turnover is approximately 100% annually, that is to say, in order to maintain a workforce of 10,000 truck drivers, an organization will need to hire approximately 10,000 new drivers each year. The cost of recruiting and training replacement driver costs can cost up to \$7,000 and costs related to both efficiency and safety are considerably higher for drivers with less experience. Consequently, improved driver retention is both a key competitive differentiator and contributor to low-cost operations. The primary factor affecting retention is the ability to get drivers back to their domicile (home location) on a consistent and relatively frequent basis. Because of the random nature of large truckload networks, operational efficiency requires drivers to typically remain away from home for up to several weeks. Figure 1 depicts a sample one-month tour for a long-haul driver with two layovers (one occurring at home and the other on the road).

Another dimension of driver cost directly relates to the hours of service (HOS) rules

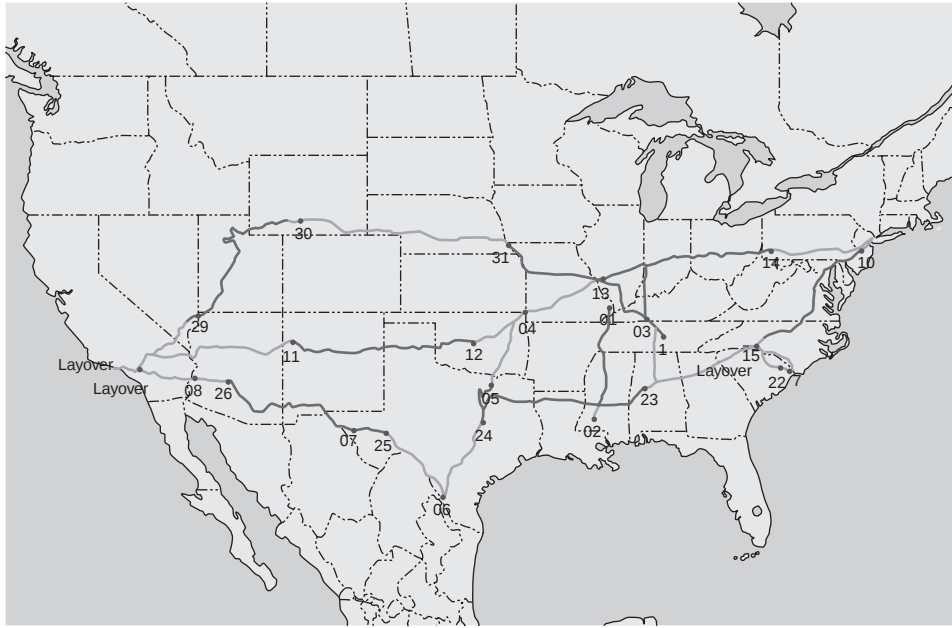


Figure 1. Typical month-on-the-road for a long-haul driver.

mandated by the Department of Transportation, which governs driver work schedules (see below for further discussion of these rules). The ability to model these complex rules in the context of large random transportation networks can improve operating decision policy choices as well as provide an effective tool to evaluate the impact of proposed changes in these rules.

A successful approach to these problems has been developed using the framework of approximate dynamic programming (ADP) [see, in particular, Powell 2], and close collaboration between industry subject-matter experts and academic researchers. This project as well as a thorough overview of the technology is reported in Ref. 3. Building on an extensive body of research conducted at CASTLE Lab, Princeton University, a tactical planning simulator (TPS) was developed to simulate real-time dispatch operations for a large truckload freight transportation fleet. The TPS has proved to be particularly well suited to investigate issues wherein the detail-level attributes of specific resources and their interaction with decision rules and

operating policies have a dominant impact on system performance.

The TPS provides a realistic simulation of 6000 drivers moving across a nationwide network transporting over a three-week time period during which approximately 50,000 shipments are simulated. Figure 2 below depicts the general framework for the system. The nodes in a vertical plane represent one of the planning regions for a 6-h time interval and a row of nodes represents successive time steps for a given region across a three-week simulation cycle. The figure is simplified representation, as the actual TPS simulator explicitly models each driver and load, with arrivals and departures captured in continuous-time. At each node, a set of shipments are available corresponding to loads outbound from the region with pickup appointment times that fall in the corresponding time interval. These shipments can either be randomly generated to simulate actual freight flows as derived from historical data or may represent forecasts of expected freight. Drivers initially become available either at the beginning time step

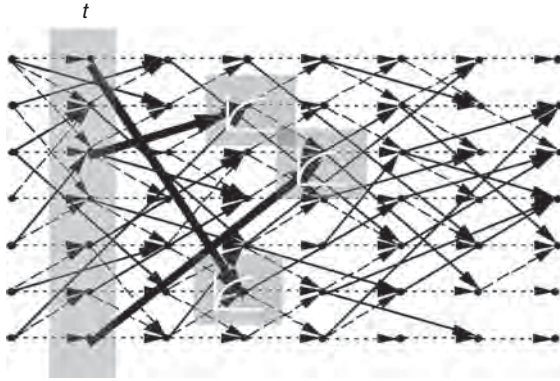


Figure 2. Planning simulator network flow structure with value functions (from Ref. 2, p. 410).

or at a point corresponding to a *simulated* entry into service. During an initial *forward pass*, a driver-to-load assignment problem is solved at each node *assigned* drivers next become available at the node corresponding to the destination and arrival time of their *assigned* load. The *objective function* for these sequential assignment problems is comprised of four types of costs or contributions:

1. direct *hard* costs associated with the particular driver-load match, most notably the *deadhead* travel distance and time;
2. *soft* costs intended to capture difficult-to-quantify ramifications of the particular assignment, such as the extent to which the load may move a driver back toward his/her domicile;
3. costs intended to steer the model toward decisions that are representative of the *real-world* solution accounting for qualitative trade-offs and adherence to policy guidelines not captured in a traditional mathematical model;
4. future value that represents the *network value* described above: it is an estimate of the marginal *future* value that the driver will contribute to the overall solution at the destination location (and time) if the assignment is selected.

Cost types (1) and (2) are described in more detail in the section titled “Execution” below as they are used in the driver-load matching algorithm used to support the real-time

dispatch process. Type (3) cost factors are derived by establishing a set of *business operating metrics*, which form *patterns* to which the solution will be *encouraged* to adhere. These *pattern metrics* specify the expected percentage of freight with certain attributes that should be transported by drivers with various characteristics, as well as the distribution of freight mix that should be moved in cases where extra freight is available. For further discussion of this technique and its application, see Refs 3 and 4. Cost type (4) lies at the core of the ADP approach and is computed empirically via dynamic programming recursion. Each iteration of the simulation consists of a *forward pass* during which the driver-load matching problems are solved by simulating dispatch decisions forward through time, followed by a backward pass in which the *realized* marginal values of drivers of each attribute class (e.g., domicile, type, remaining driver hours, and number of days to scheduled time-off) are updated with new information from the most recent forward pass. Figure 2 is a representation of the TPS network structure showing the value functions (cost type 4) that are updated with each successive pass.

To address the time-at-home (TAH) aspect of the driver retention, the TPS was configured to estimate the relative changes in operating profit as a function of driver populations for the various domicile regions. The home location of long-haul truck drivers has a significant impact on network operating efficiency, since the trip from home to the first load start and to home from

the last load end results in additional cost due to unbillable miles and unproductive time. As profitable freight flows are largely determined by geographic production, distribution, and consumption patterns, targeted driver hiring in specific regions is the most effective lever to address this issue. It is also the case that driver hiring costs and salaries exhibit geographic variation, often producing the highest wages in regions where Schneider originates the most loads. Using TPS simulations, it has been possible to quantify the marginal contribution differentials between regions, which may vary by up to \$1000 per month per driver. This analysis is used to determine more efficient advertising, incentive, and regional salary differential policies. Benefits have been conservatively estimated at \$5 M annually. A closely related analysis, which led to adjustments in proposed driver self-scheduling policies, is estimated to achieve a net annual benefit of \$25 M. For a comparison with other techniques, see Ref. 5. This work studies the driver domicile problem in the context of an LTL shipper and provides an interesting contrast to our truckload discussion as both the constraining factors and the solution approach are very different.

The TPS has also been effective in evaluating the economic impact of HOS rules changes and providing insight into strategies to mitigate productivity declines. Several years ago, the US Department of Transportation has introduced significant changes to driver work schedule constraints. The new rules increase the maximum driving time from 10 to 11 h but require all driving to occur within a single 14-h interval. Additionally, the new rules allow the 60 h/week limit to be refreshed with a continuous 34-h break. Using TPS simulations, it was possible to both quantify these impacts and evaluate the mitigating effect of adjustments to customer constraints such as pickup/delivery time-window and yard management policies. With this information, it was possible to appropriately negotiate adjustments in customer billing rates and freight tendering/handling procedures, resulting in a near complete offset of these potential productive impacts. The simulator was also used to evaluate the efficacy of the

34-h *restart rule* and to develop guidelines for practical implementation.

OPERATIONAL PLANNING

Operational planning activities normally occur within the week preceding freight movements that these activities will influence. The primary goal is to make decisions that maximize the efficiency and profitability of daily network operations. There are two key metrics that measure this success, *billed ratio* and *load profitability*. *Billed ratio* is the ratio of revenue-generating miles to total miles driven and is improved by reducing empty miles required to reposition equipment to balance supply and demand. *Load profitability* measures the profit attributable to a shipment and is a function of revenue, direct and allocated costs, and the *network impact* of the resource repositioning resulting from the shipment execution. To influence these metrics, OR-derived decision support tools for both *freight selection* and *network management* have been developed to help planners directly influence *profitability*, *demand*, and *supply* within limitations determined by market conditions, customer commitments, and fleet capacity. These tools focus on two business process areas, freight selection and asset repositioning.

With 9500 loads per day tendered by customers, a majority of which are communicated and processed electronically, automated systems are required. The *Freight selection* process is supported by automated tools for *load acceptance*, *spot pricing*, and *solicitation*. Although a majority of loads fall under long-term contract pricing agreements, a significant number may be *spot priced*, particularly during times of greater competitive market pressure. Additionally, for select large customers, the carrier may have a contract commitment to move a minimum number of loads from an outbound region on any given day.

The *load acceptance tool* is a real-time decision support system that evaluates customer loads as they are tendered to determine whether or not it should be accepted for movement and additionally evaluates feasible options for both, the capacity type and

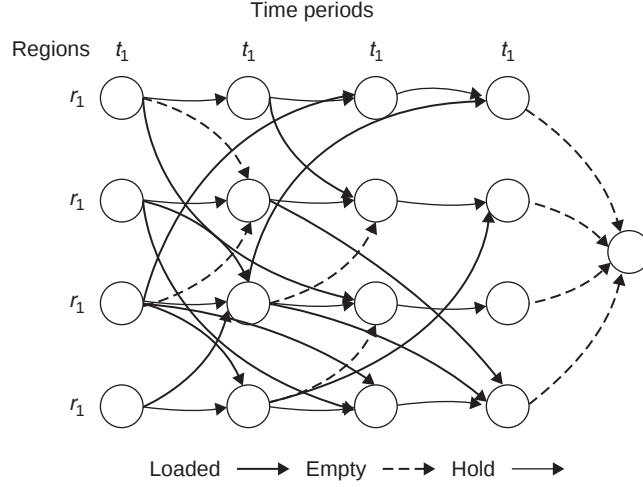


Figure 3. Time-space flow network for LP formulation.

possible pickup and delivery options. The system provides the estimated profit that will be generated by the load for each of the feasible parameter combinations, which may be available for execution of the load movement. Since the contemplated load movement will result in a redistribution of network capability, the impact of this asset reposition must be taken into account. The equation used to determine this value takes the form

$$P = R - C + V_{d,t} - V_{o,s}, \quad (1)$$

where P is the profit or value to be realized; R is the revenue associated with the shipment, C is the total direct and allocated cost associated with the shipment, and $V_{d,t}$ and $V_{o,s}$ represent the opportunity cost of the asset at the destination at time t and the origin at time s . These are specific examples of the *network value* described above. Their importance is reflected in the fact that freight transportation networks can often exhibit significant demand imbalances. Two oft-cited examples of these are (i) the flow of Asia-originating consumer products from West Coast ports to population centers in the Midwest and East that cannot be completely offset by the flow of freight westward and (ii) the lack of freight moving north out of Florida other than seasonal fruit products to provide a balance to consumer products moving to the large populations in south Florida. C is defined to include all costs

that can either be attributed directly to the shipment or can be allocated independently of how flows in the network are realized. Consequently, the calculation of values for $V_{o,s}$ and $V_{d,t}$ represents the main challenge. $V_{d,t}$ represents the anticipated profit opportunity for the asset (tractor/trailer) when it completes the current shipment at location d and becomes available for a subsequent shipment at time t . Similarly $V_{o,s}$ represents the value of alternative profit opportunities, which were otherwise available to the asset at location o and time s .

There are several methods available to estimate the values $V_{d,t}$. The classical approach is to leverage linear programming (LP) duality, using a network flow formulation with nodes corresponding to distinct geographic region/time interval combinations. Figure 3 depicts the directed graph for such a network flow formulation. A common approach is to implement a three-week planning horizon using short time intervals in the first week and longer ones in the second and third weeks; for example, 6-, 12-, and 24-h intervals for weeks 1, 2, and 3, respectively. This choice is based on the fact that approximately 90% of all freight is tendered within the week prior to pickup and at the end of a three-week look-ahead, it is reasonable to assume a common *salvage* value for assets. Consequently, a representative example for a large national freight network with up to

500 distinct planning regions and 50 time intervals would have 25,000 nodes. The average number of outbound arcs per node is about 50, giving a total of 1,250,000 arcs.

There are three type of arcs in the network; (i) *loaded arcs* represent freight movement, an origin/pickup time-window to a destination/delivery time-window corresponding to the estimate transit time to travel from the origin to the destination, (ii) *empty arcs* represent repositioning of trucks from one location to another with the appropriate time displacement, and (iii) *hold arcs* represent the decision to hold a truck idle until the next time interval in anticipation of a better opportunity. These arc flows are constrained by lower bounds representing shipments already accepted together with additional known commitments and by upper bounds representing forecasts for the number of additional loads that are expected to become available. For arcs in later time periods, less actual information is known and forecasts are more uncertain. There will often be multiple parallel *loaded arcs* representing shipment opportunities at different profit levels corresponding to varying customer contracts and spot market price estimates.

Additionally, node balance constraints insure that the number of trucks entering a node generally equals the number departing. Nodes in the first time period are source nodes corresponding to the starting (i.e., current) distribution of trucks and the *sink* node at the right receives inbound flows representing asset “salvage” values at the problem horizon. Intermediate nodes, particularly those in early time intervals, may have positive or negative balance corresponding to known plans for trucks entering or leaving service.

Although the *primal solution* to this network flow problem gives an *optimal* set of flows for the problem as described, it is of no direct interest, as the realization depends both on the freight selected and the *execution* level driver–load matching described below, the *dual values* associated with various model constraints can provide reasonable estimates of opportunity cost.

While the dual values associated with lower-bound (upper-bound) arc constraints represent the potential value of relaxing a commitment to move freight at a specified profit level [of the value of additional (i.e., above the forecast) freight], our interest is in the node balance constraint duals. If the value is positive, it provides an estimate of the value of obtaining an additional truck in the corresponding region and time slot, and, if negative, the savings associated with having on fewer. This approach has proved reasonably successful as a mechanism for generating the values $V_{d,t}$, used by the *load acceptance tool* described above. In current practice, the model is refreshed with updated information on known and forecasted freight and driver/truck available and rerun approximately every 15 min.

The *spot pricing tool* provides guidance for price negotiation to customer service representatives (CSRs) receiving freight tenders in the spot market. In this environment, customers are often shopping for the lowest price that meets their service requirements. This tool is comprised of two components. A binary logit choice model is used to derive the probability that the customer will accept a specified offered price. Some of the factors found to be significant for this model include customer incumbency, day of week, time of day, guideline or corresponding contract prices, and whether the offer is for multiple loads. The price associated with a target acceptance rate is then applied in Equation (1), to estimate the resultant profit. Depending on current capacity availability and general market activity, management can then adjust thresholds for profit and acceptance to balance fleet utilization with profit objectives.

Network management encompasses a planning process that attempts to minimize supply/demand imbalances across the broad network over a moving several-day time horizon. Using a hierarchical approach, high-level area planners have responsibility for freight flows in and out of large multi-state regions of the country; for example, the northeast, the upper Midwest, and so on. At the next level, region planners review current balance forecasts within

specific regions and time intervals. These are segmented to be small enough and narrow so that a driver destined to arrive in one of these region-time slots would be a reasonable choice for assignment to a shipment available for pickup in the same. Typically, such a region will have a diameter ranging from 25 to 100 miles. There are several hundred regions with sufficient volume to support active planning. Corresponding time slots normally range between 4 and 8 h depending on freight volume. Figure 4 illustrates this balance situation and depicts the contributing factors. On the supply side, trucks become available as (i) drivers complete delivery of inbound shipments and become available for a new assignment, (ii) drivers come off their *weekend* or TA, (iii) or transfer off of an in-transit or *relayed* load. On demand side, trucks are (i) assigned to outbound loads, (ii) go out of service as drivers start a TA interval, or (iii) assigned to take over *relayed* loads. During this stage, planners are not directly concerned with individual assignments, which we describe in the following section, but rather with trying to ensure that the number of available trucks will match the number of shipments that have been committed for movement.

In order to achieve this *node balance* within the network, there are several different *levers*, one can pull.

1. The *load acceptance tool* naturally encourages acceptance decisions that positively affect network balance through the influence of the V terms

in Equation (1), as shipments that contribute to balance at an origin node will have relatively smaller values for $V_{o,s}$ indicating less incentive to *conserve* capacity at O and those with relatively larger values for $V_{d,t}$ will encourage the movement of freight into a node with capacity at a premium.

2. Similarly, the *spot pricing tool* will tend to recommend softer pricing along lanes and time frames that will reduce imbalances. This occurs through the profit estimate provided through Equation (1) as well as explicit reduction of profit target.
3. A proactive *solicitation* program uses a decision support tool to match profitable lanes with customer tendering patterns and then directs CSRs to reach out to specific customers for additional freight.
4. In many cases, customer service constraints provide wide windows for either pickup or delivery times. In regions where imbalances appear short-lived rather than chronic, adjusting appointments earlier or later can help reduce imbalances.
5. The freight networks for different capacity types—team and solo drivers, intermodal, brokerage to third-party carriers—are managed somewhat independently. In some situations, different capacity types may exhibit *complementary* imbalances, allowing the conversion of compatible freight from one type to another to

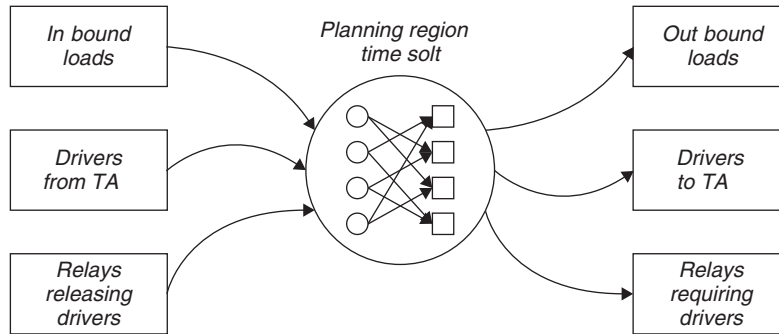


Figure 4. Network balance at a node with contributing flows.

positively influence balance in both networks.

6. The least desirable option is to move drivers from a region of excess capacity to one nearby with available freight. This is usually least desirable as near-full cost is incurred with no revenue generated. Repositioning schedules are established in this planning phase and individual driver moves are specified in the execution activities described below. When conditions require extensive multiregion repositioning, a straightforward min-cost network flow model is sometimes employed to determine the least cost region-to-region transfers to achieve the desired supply levels.

DISPATCH AND EXECUTION

Execution encompasses the decisions related to direct control of specific assets on a day-to-day basis. The related tasks are referred to as the *dispatch process* and include (i) assigning drivers/tractors to specific shipments, (ii) relaying in-transit loads from one driver to another to adjust to unexpected situations, and (iii) determining specific drivers to either reposition or hold.

As described in the previous section, one of the primary short-term planning goals is to assure that within a relatively small planning region and time slot, the number of tractors available for deployment is sufficient to cover the full trailer-loads of freight that have been accepted and scheduled for delivery. The *dispatcher's* task is to determine which specific drivers to assign to which loads. Depending on timing and other circumstances, the dispatcher may sometimes have either a single specific driver in mind, for whom a next load needs to be found or a specific load that needs to be covered. However, more general (and computationally interesting) is simultaneously finding a set of good assignments for a group of drivers who are converging on a *node* in the space-time network we have described above. The applicability of a classic

operations research model to this problem was first recognized and implemented in the 1970s. Powell [6] provides an excellent overview and perspective on this problem and its implementation at several truckload motor carriers. He also addresses important additional considerations including trailers and demand forecasting. A related article, by Powell *et al.* [7] provides an interesting account of human-system interaction in the application of this technology, addressing issues such as model trust and dispatcher compliance with system-generated recommendation.

This difficulty with effective implementation of this approach lies not with model or solution complexity, but in defining effective *filtering* and *costing* rules to generate the model arcs. The *filtering* step consists of determining whether a driver is a feasible choice for a given load. Some of these, notably whether a driver is qualified to haul a given load; for example, hazard materials or appropriate border crossing documentation, are straightforward to apply. For others, it may be unclear whether the best strategy is to implement a *hard* constraint or to allow the arc with an addition *penalty* cost component to allow a constraint to be violated when few or no other opportunities exist for the driver or load. The components that comprise the cost value for an arc corresponding to a specific driver-load match can include the following:

1. *NAL*. Distance from driver's next available location (NAL), usually either the prior destination or park location, if entering service, to the load pickup location multiplied by a *deadhead* travel cost factor;
2. *NAT*. Time delay from the driver's next available time (NAT) to the pickup time-window for the new load, multiplied by a *waste time* factor, if appropriate. (Typically, a filtering rule incorporating the driver's NAT and NAL is also applied to determine if the driver can meet the customer time window.)
3. *DOT Hours to Run*. Based on DOT driver time rules described earlier, an

available driver may either require a break before beginning the new load or will have a certain number of hours left to run. A *productivity fit* cost factor is calculated to estimate how well this remaining time meshes with the appointment time windows or service constraints of the load.

4. *Trailing Equipment Compatibility.* Upon completing a load, a driver may or may not have an empty trailer. Similarly, the next load may or may not require the driver to bring an empty trailer to the pickup location. If the driver and the new location *match* in this sense, then the driver may proceed directly to the pickup. If not, costs are applied to account for the time and distance required for the driver to drop, acquire, or exchange a trailer before proceeding to the pickup.
5. *Driver Time-At-Home.* In a previous section, we discussed ways to address driver time-off in the tactical time frame. In the dispatch process, we need to make specific assignments that will get the driver home as promised while minimizing costly empty drive time. To do this, a bonus (penalty) cost adjustment is made to the assignment arc to reflect how well (poorly) the load will *set a driver up* to meet his/her off time schedule. A common approach has been to implement a *time-to-distance* offset that conceptually defines concentric geographic rings around the driver domicile and rewards loads that move a driver to inner rings as his/her TAH date approaches. An alternative approach is currently being evaluated with the TPS which develops simulation-derived transition probabilities.

While the driver-load assignment process addresses normal “sunny day” operations, it is often the case that shipments must be handed off or relayed to another driver in transit. This may be necessary due to equipment breakdowns, unexpected driver time-off requirements, or time recovery for loads running late. Optimization tools have

been developed that extend the logic of the driver-load assignment model to scan network activity to, evaluate and rank *relay opportunities* considering hard and soft costs.

Although network planners work to setup a balance of driver and loads in each region on a daily basis, specific “stay or go” decisions often need to be made in real time when a driver is ready for a next assignment and none are available within the normal time or distance tolerance. Should the driver “stay” and remain idle at his/her current location until freight becomes available or should the driver run empty (incurring a significant deadhead cost) to another region where freight may be available? Using a decision support tool that relies on a very short-term demand forecast and accounts for costs associated with either option, the dispatcher is provided guidance on when and where to shift drivers.

CONCLUSION

This article has highlighted a subset of the planning and operations problem encountered in large-scale truckload networks and a selection of OR/MS based decision support tools that have been implemented to help solve them. While some of the models utilize advanced techniques and a level of sophistication that supports coverage in current literature, in every case, the initial and ongoing justification for investment has been borne out by verifiable economic benefit. In order to maintain this objective, improvements are ongoing. In particular, the TPS is being enhanced to address a wider range of issues, including the capability to evaluate and quantify specific customer requests related to service constraint changes and freight characteristics. Variations to *network value* calculations described in the section titled “Tactical Planning” are also under consideration. These include (i) the use of TPS value functions, (ii) techniques to lessen the sensitivity to forecast error by employing robust techniques along the lines of those described in Ref. 8 or unconstraining methods

as described in Ref. 9, and (iii) the application of average rather marginal costing for certain network value applications.

REFERENCES

1. Crainic TG, Laporte G. Planning models for freight transportation. *Eur J Oper Res* 1997;97:409–438.
2. Powell WB. Approximate Dynamic Programming: Solving the curses of dimensionality. Hoboken (NJ): John Wiley & Sons Inc.; 2007.
3. Simao HP, Day J, George A, *et al.* An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transport Sci* 2009;43(2):178–197.
4. Marar A, Powell WB. Capturing incomplete information in resource allocation problems through numerical patterns. *Eur J Oper Res* 2009;197(1):50–58.
5. Erera A, Hewitt M, Karacik B, *et al.* Locating drivers in a trucking terminal network. Working Paper, The Supply Chain and Logistics Institute, Georgia Institute of Technology, 2006.
6. Powell WB. Real-time dispatching for truckload motor carriers. In: Taylor GD, editor. *Logistics engineering handbook*. Boca Raton, FL: CRC Press; 2007a. pp. 15–30.
7. Powell WB, Marar A, Gelfand J, *et al.* Implementing operational planning models: a case application from the motor carrier industry. *Oper Res* 2002;50(4):571–581.
8. Erera A, Morales JC, Savelsbergh M. Robust optimization for empty repositioning problems. Working Paper, The Supply Chain and Logistics Institute, Georgia Institute of Technology, 2007.
9. Queenan C, Ferguson M, Higbie J, *et al.* A comparison of unconstraining methods to improve revenue management systems. *Prod Oper Manag* 2007;16(6):729–746.

OPERATIONS RESEARCH APPROACHES TO ASSET MANAGEMENT IN FREIGHT RAIL

MICHAEL F. GORMAN
STEVEN HARROD
Department of MIS, Operations
Management, and Decision
Sciences, University of Dayton,
Dayton, Ohio

INTRODUCTION

The objective of this article is to provide an overview of the theory and practice of operations research (OR) in asset management in the freight rail industry. In an accompanying article (see *Operations Research for Freight Train Routing and Scheduling*), we describe how freight rail service is designed using OR. That is, this article addresses, primarily, the provision of assets, and the companion article addresses the strategic and tactical deployment of those assets, through scheduling. Here, we describe the five primary assets required to provide freight rail service and how OR contributes to improved decision making surrounding each asset.

Trains require five different resources to run. The “physical plant” consists of the fixed infrastructure or network assets such as rail lines and rail yards or terminals (for intermodal). The “mobile resources” are crew, locomotives, and railcars. The second section describes the OR techniques that apply to mobile assets, the third section discusses OR modeling of lines and yards, and the fourth section concludes the article.

In each section, we describe OR modeling approaches as they are applied to strategic, tactical, and operational decision horizons. For example, locomotive fleet modeling might address strategic issues such as long-term fleet sizing over a multiple-year horizon, tactical modeling might focus on meeting train

needs on repeating weekly train scheduling, and operational modeling might focus on locomotive repositioning given a regional surplus or deficit. Similar decision horizons have been applied in the OR modeling literature surrounding the other rail topics.

MOBILE ASSETS: LOCOMOTIVE, CARS, AND CREW

The mobile assets that make up a train are cars, crews, and locomotives. Only the railcar follows a customer shipment from its origin to its destination. A train may carry a shipment for some portion of its route, which is not usually the same as the train’s origin and destination; a shipment often moves on many trains. A set of locomotives, called a *consist*, are typically assigned from the origination to the termination of a train. Crews are assigned to a small (often four to eight hour) portion of a train route called a *crew district*; trains have a crew change at each change in crew district. Because each asset is subject to entirely different management rules, different OR approaches govern each.

Railcars

Before service to a customer can be provided, an empty railcar is delivered to the customer origin. In the case of intermodal, an empty container or trailer must often be delivered. OR modeling is applied to deciding which empty should be assigned to which customer order. See Dejax and Crainic [1] for a survey on this subject.

Some work has been done on the strategic railcar fleet sizing problem, which establishes the appropriate number of cars to support forecasted traffic in steady state [2–6]. Considerations include on-time service and back-orders; stochastic supply; demand and travel times; and fleet utilization.

The tactical empty railcar assignment problem is typically expressed as a transportation problem or conservation of flow problem.

The first challenge in such problems is to sort through the plethora of car attributes and car orders to assure a quality solution.

Car attributes may be categorized as static or dynamic. Static attributes include car type (box car, gondola, flat, etc.), length, and capacity. Dynamic attributes include date and location of next availability. A set of customer preferences on car orders specifies desired and near-substitute car types and the desired delivery location, date, and allowable delivery window. Between subsets of cars and car orders over the horizon that spans a few weeks, there is a set of allowable assignments based on these attributes. The cost of any allowable assignment includes the time, distance, and yard handling cost of empty movement to the railroad, as well as qualitative factors such as car-to-order mismatch, early and late delivery, and unmet order cost. The total costs of assignments are minimized through optimal car-to-order assignments.

Because the OR modeling for this problem is relatively straightforward and quick to solve, standard LP solution methods have been applied. As a result, real-time tactical empty car distribution models have been among the most effective and widespread in the rail industry. Gorman *et al.* [7,8] describe successful applications and BNSF and CSX railroads, and Narisetty *et al.* [9] describe an implemented system at Union Pacific. Sherali and Suharko [10] describe an effort to distribute automotive railcars amongst the major North American Railroads, and Gorman [11] describes real-time container distribution in intermodal. Key challenges in implemented decision support systems include constantly evolving car supply and car order information coupled with sequential assignment practices, as well as stochastic delivery times.

Extensions to the commonly implemented approach include the inclusion of stochastic car supply, customer car orders, and delivery times for empties [12]. Each of these applications focuses on the near-term decisions of assigning empty railcars to current orders. The models could be critiqued as myopic; car assignments made today affect future supply conditions after a shipment is completed and the car becomes empty again

(often a week or two in the future). Powell and Topaloglu [13] and Topaloglu and Powell [14] propose an approximate dynamic programming approach to capture the effect of future network conditions when making car assignment decisions. Similarly, Powell and Carvalho [15] propose a logistics queueing network approach for tactical flat car and container management in intermodal. Along a different line, Joborn *et al.* [16] describe economies of scale in equipment distribution when empty cars do not travel in predetermined blocks. They propose a tabu search approach to this network mixed integer problem with side constraints.

Train Crews

Train crews work in teams to move a train across its route. The train goes through numerous crew districts or geographic corridors; in each crew district, the train is manned by a single crew (usually two people: a conductor and an engineer) that is qualified to operate a train in that district's unique topography. The crew district represents one job for a crew, and the job is usually between 4 and 8 h. At the transition point between two districts known as the *crew change point*, one crew exits and another enters the train, and the train continues on its way into the next district. Each crew has a "home," terminal, works a train to an "away" terminal, then subject to on duty and rest time restrictions, works another train in the opposite direction back to its home. In some cases, districts are "double-ended", meaning that there are two different pools of crews, each with a home at each end of the district.

The difficulty of managing train crews comes from managing the costs against the constraints of the problem. In most cases, crews must receive some minimal rest between trips; thus, an inventory of rested (and resting) crews must be maintained at both ends of the district because it is of paramount importance to avoid crew-related train delay. On the other hand, at the away terminal, if the time between trips exceeds some contractually agreed upon number of hours (usually 16), the crew receives detention pay. By labor agreement, particular to the North American freight rail

industry, crews must be called in *first in, first out* (FIFO) order. In many cases and over many time intervals, the number of trains in each direction may not be balanced; thus, either detention pay mounts, or the crew can be “deadheaded,” or moved back home (or away) at some cost via taxi or train without providing productive service.

Extensive work has taken place with airline and passenger European rail crew scheduling (see for examples, Barnhart *et al.* [17] for airline, and Caprara *et al.* [18], Kroon *et al.* [19], and Ernst *et al.* [20] for European passenger rail), but these methods do not apply directly to the freight rail crew management problem described above because of rules such as FIFO described above.

Only two articles [21,22] have been written specific to North American freight rail. Both models are tactical in nature, that is, they are solved with an initial crew starting position. The objective is to minimize the sum of crew wages, deadhead costs, and crew detention. In both cases, the solution is subject to the constraints: all scheduled trains are covered by exactly one crew with no delay and rest rules are obeyed. Gorman and Sarrafzadeh [21] use dynamic programming to solve the singled-ended crew district problem to optimality in seconds, but this method does not handle more complex two-ended districts.

Vaidyanathan *et al.* [22] describe the two-ended crew district problem decision process and model it as an integer multicommodity flow problem with side constraints on a time-space network. The approach of Vaidyanathan *et al.* requires explicit modeling of crew flow balance and handling of the FIFO constraint, which is inherent in the dynamic programming approach. Vaidyanathan *et al.* use LP relaxation and iterative branch and bound to solve the much larger and complex formulation. The authors report near optimum results in seconds or minutes, depending on the problem.

Of all the moving assets, crews are most subject to deviations in the train schedule because crew inventory and dwell time must be managed most closely. Train annulments, second sections, and delays in existing trains can directly affect the quality of the plan in a crew district, either increasing detention or

deadhead cost, or causing train-related crew delay. The methods developed to this point focus on the deterministic case, but the risk associated with stochastic train schedules is of great importance to decision makers and largely uncovered in the current research. See Walker *et al.* [23] for an example of crew recovery planning.

Locomotives

Locomotives are usually assigned to trains in “consists”—groups of locomotives with similar operating characteristics measured in horse power and number of axles. The minimum required consist for a train depends on the train size (in tons). The consist must have a sufficient number of required axles to provide sufficient tractive effort to move across the route’s topography, and the required horse power to move the train at speeds that meet its schedule, often described as the required HPT (horse power per ton) ratio. The inventory of locomotives must be managed at origination yards to assure that locomotive requirements and preferences are met without delay.

Similar to railcars at load destination and crews at crew change points, locomotive supply at train origin is driven by their previous use on a train. In this case, supply is usually created at the train termination. There is a handling cost to adding or subtracting locomotives from a consist; so where possible, consist “busting” is avoided by connecting the entire consist of a terminating train to an originating one.

Often a freight imbalance drives the need for locomotive repositioning to or from a yard. Locomotive repositioning occurs either in a scheduled train, known as *deadheading*, or independently, known as *light travel*.

As described in a survey article by Cordeau *et al.* [24], little published work has been conducted on strategic models on locomotive fleet sizing (one recent example is Godwin *et al.* [25]). The focus has been more on steady-state locomotive planning to support a weekly schedule, or tactical locomotive scheduling given a starting inventories at yards in the rail network. Early work focused on single-locomotive scheduling [26]. More recently, the more useful locomotive consist

problem has been considered [27,28], as well as joint distribution of locomotives and railcars [29].

The locomotive scheduling problem is formulated as a multicommodity flow mixed integer problem with side constraints. The decision at each yard is which locomotives to assign to which trains, recognizing that these assignments affect downstream locomotive supply (usually one to a few days in the future). The objective in the problem can vary, but often includes minimization of train overpowering (and underpowering, if allowed); locomotive idle time; deadheading and light travel; and consist busting and consist inconsistency costs. The constraints are to supply sufficient locomotives to all trains such that they can depart on time and run on schedule. Additionally, there are minimum and maximum locomotives on each train, as well as minimum times for locomotive connections between trains that must be obeyed.

Because of the locomotive scheduling problem size for realistic examples, heuristics have been used to provide near-optimal solutions. Early single-locomotive models used Benders decomposition and Lagrangian relaxation. Ziarati *et al.* [30] use Dantzig–Wolfe decomposition and shortest path algorithms applied to the multiple locomotive problem at the Canadian National (CN) railroad. Ahuja *et al.* [27] propose decomposing the problem into two stages and very large-scale neighborhood search to a problem at CSX railroad.

The problem is made more complicated when one considers that individual locomotives within a consist have mandated service requirements, and often must be routed to a shop, ideally while productively providing tonnage hauling in a train consist [31]. Similarly, locomotives have individual refueling histories and fuel efficiency, thus have independent fueling needs. These additional constraints restrict paths locomotives can take, and because of consist mix rules, can affect the flows of other locomotives as well. Locomotive schedules are further complicated by delayed or canceled trains, or yard congestion. The locomotive consist planning and scheduling problem remains

one of the more difficult OR problems in rail asset management.

NETWORK ASSETS: YARD AND LINE PLANNING AND OPERATIONS

The network on the ground, the mainlines, sidings, signals, and classification yards, determine the feasible region of railway operations. Poorly configured sidings or inefficient yards may severely handicap railway transport capabilities regardless of any efforts to mobilize labor or improve locomotives or rolling stock. Further, the network on the ground nearly always has a long functional life span, represents a large financial investment, and requires a lengthy investment return period. Consequently, changes to or extensions of the network are made cautiously and according to well-tested calculations. Pachl [32] provides a general purpose text on railway installations, signaling systems, and operation.

Signals and Communications

While signals and communications are significant factors in the capacity enhancement of dense passenger services, innovations in signals (beyond the established fixed block color light system) are less important to the capacity of heavy rail freight. The primary reasons for this are the acceleration and deceleration characteristics of heavy freight trains. In order to minimize wear and tear, heavy freight trains avoid frequent, heavy braking, and likewise avoid fast acceleration to save fuel. Consequently the signal response times and headways necessary for cost-effective operation are significantly less stringent than those required for safety or shared operation with passenger trains. Dinger *et al.* [33] make an extensive analysis of communications based or positive train control within a heavy freight network and enumerate scenarios where advanced signals and communications will improve capacity, and scenarios where they will reduce capacity.

Line Capacity and Planning

The decisions faced in specifying a railway line generally include the number of parallel

main lines and the size and location of additional short tracks to allow trains to wait and allow each other to pass, called *sidings*, as well as the location of connections between main lines, called *crossovers*. If more than one main line track is specified, there is a choice of whether to limit each track to a single direction of flow or to provide signals and communications for bidirectional flow on both tracks. Sidings may be specified at different lengths, either to support trains of different lengths or according to whether a train will stop and wait in the siding or attempt to continue moving.

The impact of siding placement and its effect on traffic is a frequent topic. An early analytical study by Frank [34] estimates maximum flow for a single track line operating at one speed and demonstrates that under some scenarios of dispatch double sidings supporting two waiting trains allow greater network flow. This observation received no further attention until Harrod [35] demonstrated that some scenarios with mixed speeds also benefit from double sidings. Petersen and Taylor [36] consider the design of a mixed traffic single track line for shared use by slow freight trains and high speed passenger trains, and conclude that a carefully planned and dispatched single track line could perform acceptably compared to a double track line. Higgins *et al.* [37] present a mixed integer programming model with the objective of determining the location of sidings on a single track line to support a given traffic pattern. A Benders decomposition solution scheme is used. Finally, Landex [38] presents single track parametric measures of capacity derived from the UIC (International Union of Railways) standard 406.

Simulation has played a prominent role in line planning. The origins of early train performance calculators and dispatch simulators are documented by Wilson and Hudson [39] and Wilson and Lach [40]. Prokopy and Rubin [41] simulate a wide variety of typical North American operating scenarios and then estimate statistical or parametric functions of capacity from the simulation results. Kraft [42] gives a focused comparison of analytical and simulation models for line planning. Welch and Gussow [43], finalists for the

1985 Franz Edelman Award, describe many innovations in their simulation model, again at Canadian National Railway. They describe improvements to lockup prevention (lockup is when trains are mistakenly dispatched into terminating, infeasible paths) and the representation of double main lines. They found, for their operating pattern, that signal blocks should be shorter and that double tracks should be evenly distributed along a route and at the entrances to yards. Public policy decisions in line planning are considered by Leilich [44]. Krueger [45] presents statistical estimates of capacity calibrated to simulated operations of Canadian National Railway. Lai and Barkan [46] extend this model with an application that automatically enumerates a wide range of line improvements and estimates the resulting network capacity.

Many railway infrastructure projects are now joint or publicly financed projects, introducing an additional element of financial and political risk to the outcome of the planning process. A number of authors give warnings on the interpretation of simulation and design analysis. Wood and Robertson [47] present cautionary lessons drawn from life experiences on the failure to calibrate line design to intended train timetables. Harris and Cordon [48] demonstrate that infrastructure investments may require some utility function of multiple measures to justify their approval. White [49] expresses concern that simulation output may be misinterpreted, particularly if only one performance measure is considered.

Yard Design and Operation

Yards are the heart of railway operations. They direct and measure the flow of traffic throughout the network, and are the most critical factor in accomplishing the transportation function. Indeed, Mierzejewski [50] has extensively documented that only after Allied forces shifted their bombing focus from main lines to yards was the German economy finally disrupted in World War II. Dramatic service failures ensued after the merger of Union Pacific and Southern Pacific due to the failure to properly manage classification in the Houston area in 1997 [51].

Yards occur in a range of sizes, but primarily two forms: the flat yard and the hump

yard. The flat yard is very labor intensive, but may be worked simultaneously by multiple switch crews. The hump yard contains a man-made hill at one end over which whole trains are pushed and released, one car at a time, so that they may roll by gravity into a programmed track. With automation a hump yard requires very little labor. However, a hump yard may only sort one inbound train of cars at time.

Differences in train technology can severely limit the application of hump yards. In particular, some traffic is prohibited in hump yards in North America due to technical limitations, and multicar blocks do not handle well on humps. Wong *et al.* [52] provide a comprehensive engineering and economic guide to the planning and installation of yards. Kraft [53,54] reviews the interplay between North American railway scheduling strategy and classification yard technology from a 100 years ago to the current date.

OR topics in classification have progressed from simulation, to queueing, and to analytical studies. To address the problem of excessive dwell time in yards, Crane *et al.* [55] demonstrate a very early simulation of yard operations. Applied simulation models have since played a vital role in strategic management. Indeed, Troup [56] refers to simulation as “the most advanced tool for studying and analyzing rail yard operations.” Troup provides a detailed and still relevant manual to yard operations.

A few studies derive queueing models of classification yard operations [57–60], but queueing methods have not generally been accepted in practice. Martland [61] expressed concern that prior queueing models required unrealistic presumptions of steady-state processes. Martland describes a method called PMAKE (“probability of making a connection”) and notes its extensive application record in North American operations. Martland calibrates a number of cumulative probability functions for the time to complete a connection as a function of yard dwell time, congestion, car status, and other variables.

Daganzo [62–64] investigates the specific quantity and sequence of switching moves required according to traffic volume and

yard dimensions. Daganzo considers fixed block assignments to yard tracks (“static” blocking), dynamic blocking, methods of handling strings of cars, and a variety of strategies for prioritizing the classification of cars. He *et al.* [65] apply computer-aided dispatching to classification yard scheduling. A large-scale mixed integer formulation is presented for the simultaneous matching of inbound cars to outbound trains, assignment of switch crews, and assignment of yard tracks along with solution algorithms. Yagar and Saccamanno [66] describe a dynamic programming application applied at a yard for Canadian National.

FURTHER READING AND OTHER TOPICS

A number of survey articles have been written describing research and application of OR methodologies in the rail industry (Cordeau *et al.* [24] is a recent example). For example, a recent tutorial of various methods applied to numerous rail problems can be found in Ahuja *et al.* [67]. Classic surveys in rail optimization can be found in Assad [68] and Haghani [69]. Recent summary discussion on optimization in rail have been published in Newman *et al.* [70], a broader discussion of long-haul freight transportation is in Crainic and Hall [71], and specific discussion of intermodal in Crainic and Kim [72].

Numerous other topics in OR and rail outside of service design and asset management have been pursued by researchers and applied in practice. For examples, Gorman [73,74] discusses pricing optimization and yield management in the rail context. Gorman and Kanet [75] and Podofilini *et al.* [76] explore line maintenance planning and scheduling. Finally, broader questions of modal choice and rail in the context of national transportation network design has been researched by Gorman [77] and Crainic [78].

New topics in rail research are emerging. Future topics of interest could center on topics such as mode sustainability and fuel conservation, security of rail and port operations, internationalization and containerization, public–private partnerships in rail

infrastructure investment, line capacity with satellite tracking and control, and the freight and passenger rail interface.

REFERENCES

1. Dejax P, Crainic T. A review of empty flows and fleet management models in freight transportation. *Transp Sci* 1987;21(4):227–248.
2. Beujon G, Turnquist M. A model for fleet sizing and vehicle allocation. *Transp Sci* 1991;25(1):19–45.
3. Sherali H, Maguire L. Determining rail fleet sizes for shipping automobiles. *Interfaces* 2000;30(6):80–90.
4. Papier F, Thonemann U. Queuing models for sizing and structuring rental fleets. *Transp Sci* 2008;42(3):302–317.
5. Sayarshada HR, Ghoseirib K. A simulated annealing approach for the multiperiodic rail-car fleet sizing problem. *Comp Oper Res* 2009;36(6):1789–1799.
6. Sayarshad H, Ghoseiri K. A simulated annealing approach for the multi-periodic rail-car fleet sizing problem. *Comput Oper Res* 2009;36(6):1789–1799.
7. Gorman M, Crook K, Sellers D. North American freight rail industry real-time optimized equipment distribution systems: state of the practice. *Transp Res: Part C* 2010. In press.
8. Gorman M, Acharya D, Sellers D. CSX railway uses OR to cash in on optimized equipment distribution. *Interfaces* 2010;40(1):1–12.
9. Narisetty A, Richard J, Ramcharan D, *et al.* An optimization model for empty freight car assignment at Union Pacific. *Interfaces* 2008;38(2):89–102.
10. Sherali H, Suharko A. A tactical decision support system for empty railcar management. *Transp Sci* 1998;32(4):306–329.
11. Gorman M. Hub group implements a suite of OR tools to improve operations. *Interfaces* 2010;40(5):1–17.
12. Crainic TG, Gendreau M, Dejax P. Dynamic and stochastic models for the allocation of empty containers. *Oper Res* 1993;41(1):102–126.
13. Powell WB, Topaloglu H. Fleet management. In: Wallace S, Ziemba W, editors. *Applications of stochastic programming*. Math programming society—SIAM series in optimization. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2005.
14. Topaloglu H, Powell WB. Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. *INFORMS J Comput* 2006;18(1):31–42.
15. Powell W, Carvalho T. Real-time optimization of containers and flatcars for intermodal operations. *Transp Sci* 1998;32(2):110–126.
16. Joborn M, Crainic T, Gendreau M, *et al.* Economies of scale in empty freight car distribution in scheduled railways. *Transp Sci* 2004;38(2):121–134.
17. Barnhart C, Johnson E, Nemhauser G, *et al.* Crew scheduling. In: Hall RW, editor. *Handbook of transportation science*. New York: Springer; 2003. pp. 493–521.
18. Caprara A, Fischetti M, Toth P, *et al.* Algorithms for railway crew management. *Math Program* 1997;79:124–141.
19. Kroon L, Huisman D, Abbink E, *et al.* The new Dutch timetable: the OR revolution. *Interfaces* 2009;39(1):6–17.
20. Ernst AT, Jiang H, Krishnamoorthy M, *et al.* An integrated optimization model for train crew management. *Ann Oper Res* 2001;108(1):211–224.
21. Gorman M, Sarrafzadeh M. An application of dynamic programming to crew balancing at Burlington Northern Santa Fe Railway. *Int J Ser Technol Manage* 2000;1(2/3):174–187.
22. Vaidyanathan B, Jha K, Ahuja R. Multicommodity network flow approach to the railroad crew-scheduling problem. *IBM J Res Dev* 2007;51(3/4):325–344.
23. Walker C, Snowdon J, Ryan D. Simultaneous disruption recovery of a train timetable and crew roster in real time. *Comput Oper Res* 2005;32(8):2077–2094.
24. Cordeau J, Toth P, Vigo D. A survey of optimization models for train routing and scheduling. *Transp Sci* 1998;32(4):380–404.
25. Godwin T, Gopalan R, Narendran T. Tactical locomotive fleet sizing for freight train operations. *Transp Res: Part E* 2008;44(3):440–454.
26. Forbes M, Holt J, Watts A. Exact solution of the locomotive scheduling problems. *J Oper Res Soc* 1991;42:121–138.
27. Ahuja R, Liu J, Orlin J, *et al.* Solving real-life locomotive-scheduling problems. *Transp Sci* 2005;39(4):503–517.
28. Vaidyanathan B, Ahuja R, Liu J, *et al.* Real-life locomotive planning: new formulations and computational results. *Transp Res: Part B* 2008;42(2):147–168.

29. Fügenschuh A, Homfeld H, Huck A, *et al.* Scheduling locomotives and car transfers in freight transport. *Transp Sci* 2008;42(4): 478–491.
30. Ziarati K, Soumis F, Desrosiers J, *et al.* Locomotive assignment with heterogeneous consists at CN North America. *Eur J Oper Res* 1997;97(2):281–292.
31. Vaidyanathan B, Ahuja R, Orlin J. The locomotive routing problem. *Transp Sci* 2008;42(4):492–507.
32. Pacht Joern. Railway operation and control. Mountlake Terrace, Washington (DC): VTD Rail Publishing; 2002.
33. Dingler M, Lai Y, Barkan CPL. A review of the effects of CBTC and ECP brakes on railroad capacity. *Transp Res Rec*. In press.
34. Frank Ove. Two-way traffic on a single line. *Oper Res* 1966;14(5):801–811.
35. Harrod S. Capacity factors of a mixed speed railway network. *Transp Res: Part E* 2009; 45:830–841.
36. Petersen E, Taylor A. Design of a single-track rail line for high-speed trains. *Transp Res: Part A* 1987;21A(1):47–57.
37. Higgins A, Kozan E, Ferreira L. Modelling the number and location of sidings on a single track railway. *Comput Oper Res* 1997;24(3): 209–220.
38. Landex A. Evaluation of railway networks with single track operation using the UIC 406 capacity method. *Netw Spatial Econ* 2009;9:7–23.
39. Wilson P, Hudson C. Simulation of over-the-road operations. *CNR Proceedings*. 1963. pp. 142–149.
40. Wilson P, Lach D. Computer simulation developments in Canadian national railways. *International Symposium on the Use of Cybernetics on the Railways*; Montreal. 1967. pp. 176–182.
41. Prokopy J, Rubin R. Parametric analysis of railway line capacity. Technical Report Document No. FRA-OPPD-75-1. U.S. Department of Transportation, Federal Railroad Administration; Peat, Marwick, Mitchell and Co; 1975.
42. Kraft E. Analytical models for rail line capacity analysis. *J Transp Res Forum* 1988;29(1): 153–162.
43. Welch N, Gussow J. Expansion of Canadian national railway's line capacity. *Interfaces* 1986;16(1):51–64.
44. Leilich R. Application of simulation models in capacity constrained rail corridors. In: Medeiros DJ, Watson EFR, Carson JS, *et al.*, editors. *Proceedings of the 1998 Winter Simulation Conference*. San Diego: Winter Simulation Conference Foundation, Society for Computer Simulation International; 1998. pp. 1125–1133. <http://www.informs-sim.org/wsc98papers153.pdf>.
45. Krueger H. Parametric modeling in rail capacity planning. In: Farrington PA, Nembhard HB, Sturrock DT, *et al.*, editors. *Proceedings of the 1999 Winter Simulation Conference*. San Diego: Society for Computer Simulation International; 1999. pp. 1194–1200. <http://www.informs-sim.org/wsc99papers173.pdf>.
46. Lai Y, Barkan CPL. Enhanced parametric railway capacity evaluation tool. *Railways 2009, transportation research record no. 2117*. 2009. pp. 33–40.
47. Wood D, Robertson S. Planning tomorrow's railway—role of technology in infrastructure and timetable options evaluation. In: Allan J, Hill RJ, Brebbia CA, *et al.*, editors. *Computers in railways VIII*. Southampton: Wessex Institute of Technology, WIT Press; 2002. pp. 721–730.
48. Harris N, Cordon M. Cost-effective timetable and infrastructure simulation. In: Allan J, Hill RJ, Brebbia CA, *et al.*, editors. *Computers in railways VIII*. Southampton: Wessex Institute of Technology, WIT Press; 2002. pp. 389–395.
49. White T. Alternatives for railroad traffic simulation analysis. *Transp Res Rec* 2005; 1916:34–41.
50. Mierzejewski A. The collapse of the German war economy, 1944–1945; allied air power and the German national railway. Chapel Hill: University of North Carolina Press; 1988.
51. Saunders R. Main lines, rebirth of the North American railroads, 1970–2002. DeKalb (IL): Northern Illinois University Press; 2003.
52. Wong P, Sakasita M, Stock W, *et al.* Railroad classification yard technology manual Volume I: yard design methods. Technical Report Document No. DOT-TSC-1337. Washington (DC): U.S. Department of Transportation, Federal Railroad Administration; 1981.
53. Kraft E. The yard: railroading's hidden half, part 1. *Trains* 2002;62(6):46–67.
54. Kraft E. The yard: railroading's hidden half, part 2. *Trains* 2002;62(7):36–47.
55. Crane R, Brown F, Blanchard R. An analysis of a railroad classification yard. *J Oper Res Soc Am* 1955;3(3):262–271.
56. Troup K. Railroad classification yard technology: an introductory analysis of functions

- and operations. Technical Report Document No. DOT-TSC-FRA-75-19. Washington (DC): U.S. Department of Transportation, Federal Railroad Administration; 1975.
57. Petersen E. Railyard modeling: part I. Prediction of put-through time. *Transp Sci* 1977; 11(1):37–49.
 58. Petersen E. Railyard modeling: part II. The effect of yard facilities on congestion. *Transp Sci* 1977;11(1):50–59.
 59. Turnquist M, Daskin M. Queuing models of classification and connection delay in rail-yards. *Transp Sci* 1982;16(2):207–230.
 60. Marinov M, Viegas J. A simulation modelling methodology for evaluating flat-shunted yard operations. *Simul Model Pract Theory* 2009;17:1106–1129.
 61. Martland C. PMAKE analysis: predicting rail yard time distributions using probabilistic train connection standards. *Transp Sci* 1982;16(4):476–506.
 62. Daganzo C. Static blocking at railyards: sorting implications and track requirements. *Transp Sci* 1986;20(3):189–199.
 63. Daganzo C. Dynamic blocking for railyards: part I. Homogeneous traffic. *Transp Res: Part B* 1987;21B(1):1–27.
 64. Daganzo C. Dynamic blocking for railyards: part II. Heterogeneous traffic. *Transp Res: Part B* 1987;21B(1):29–40.
 65. He S, Song R, Chaudhry SS. An integrated dispatching model for rail yards operations. *Comput Oper Res* 2003;30:939–966.
 66. Yagar S, Saccamanno F. An efficient sequencing model for humping in a rail yard. *Transp Res: Part A* 1983;17(4):251–262.
 67. Ahuja R, Cunha C, Sahin G. Network models in railroad planning and scheduling. *Tutorials Oper Res* 2005; 54–101.
 68. Assad A. Models for rail transportation. *Transp Res* 1980;14A:281–310.
 69. Haghani A. Rail freight transportation: a review of recent optimization models for train routing and empty car distribution. *J Adv Transp* 1987;21:31–38.
 70. Newman M, Nozick L, Yano C. Optimization in the rail industry. In: Pardalos P, Resende M, editors. *Handbook of applied optimization*. New York: Oxford University Press; 2002.
 71. Crainic T, Hall R. Long-haul freight transportation. *Handbook of transportation science*. New York: Kluwer Academic Publishers; 2003. pp. 451–516.
 72. Crainic T, Kim K. Intermodal transportation. In: Barnhart C, Laporte G, editors. *Transportation. Handbooks in operations research and management science*. Amsterdam: North-Holland Publishing Co.; 2007. pp. 467–537.
 73. Gorman M. Pricing and product mix optimization in freight transportation. *Transp Q* 2002;56(1):135–148.
 74. Gorman M. Intermodal pricing model creates a network pricing perspective at BNSF. *Interfaces* 2001;31(4):37–49.
 75. Gorman M, Kanet J. Formulation and solution approaches to the rail maintenance production gang scheduling problem. *J Transp Eng* 2010;136(8):701–708.
 76. Podofilini L, Zio E, Vatn J. Risk-informed optimisation of railway tracks inspection and maintenance procedures. *Reliab Eng Syst Saf* 2006;91(1):3495–3510.
 77. Gorman M. Evaluating the public investment mix in US freight transportation infrastructure. *Transp Res: Part A* 2008;42(1):1–14.
 78. Crainic T, Florian M, Léal J. A model for the strategic planning of national freight transportation by rail. *Transp Sci* 1990;24(1):1–24.

OPERATIONS RESEARCH FOR FREIGHT TRAIN ROUTING AND SCHEDULING

STEVEN HARROD
MICHAEL F. GORMAN
Department of MIS, Operations
Management, and Decision
Sciences, University of Dayton,
Dayton, Ohio

Although a very old technology with many traditions, today, freight railway managers make extensive use of operations research to address complex operations problems. This article surveys planning and control of freight train movements upon the rail network frequently referred to as the *operating plan*. We describe the typical practical decisions faced by railway management given a fixed infrastructure, and systematically describe relevant operations research theory and practice. The planning and control of assets such as locomotives and crews are typically subordinate decisions to the train movement plan and are not addressed here. We do address these topics in our companion article (see ***Operations Research Approaches to Asset Management in Freight Rail***) along with the long-term planning of lines and yards. Notable prior surveys of operations research applicable to freight rail are Cordeau *et al.* [1], Haghani [2], and Assad [3]. A higher abstraction of operational planning would be the determination of what services to offer and at what price to offer them. We do not address this topic directly, but instead recommend Crainic [4]. Although passenger train operations share many operational problems with freight rail, many of the procedures described here are irrelevant to passenger services and we refer the interested reader to Caprara *et al.* [5], Huisman *et al.* [6], and Kroon *et al.* [7] for literature on passenger services.

In addition to surveying recent literature, this article organizes the literature according to the alternative operational strategies

found in rail freight. In the next section we describe the significance of rail freight and describe significant technology differences between systems. In the section titled “Service Design,” we define service design and the activities that are typical of a block car routing strategy. A single car routing strategy is more common in high frequency hub and spoke services and is described in the section titled “Individual Car Routing.” We present scheduling algorithms and performance estimates for *ad hoc* dispatching without predefined train paths in the section titled “Timetable-Free Dispatching,” and a novel classification scheme for the decision variable structure of train pathing problems in the section titled “Timetable Generation.” In the section titled “Conclusion and Future Trends,” we conclude with comments on likely future operations research topics in freight rail.

SIGNIFICANCE AND DESCRIPTION OF RAIL FREIGHT OPERATIONS

Freight rail is an important and growing component in the world’s freight transportation infrastructure. In the United States, for example, ton-miles carried by rail have increased from 27% of all modes ton-miles in 1980 to 39% in 2007, and have doubled in absolute figures [8]. In Europe, rail freight captured only 17.3% of all modes ton-kilometers in 2006, but grew in absolute figures at an annual compounded growth rate of 2.2% from 2001 to 2006 [9].

Other technology and traffic conditions also motivate differences between typical North American and European applications of operations research to rail freight. The European network is typically limited to trains of 750 m length, with some exceptional trains of 1 km length, but the North American network regularly operates trains of 2.29 km (7500 ft) and some operations now support lengths of 3 km (10,000 ft). The mix of freight traffic in Europe contains a higher proportion of priority freight than the North

American market. According to Vassallo and Fagan [10], in 2000, coal accounted for 23% of North American railway and highway ton-kilometers, but in Europe it was only 1%. Finally, European rail freight must integrate with a dense and punctual passenger rail service. According to the International Union of Railways [11] the ratio of passenger kilometers to ton-kilometers in Europe in 2008 was 0.15, but in North America it was only 0.004.

Freight trains may be classified under three types: unit, loose car, and dedicated. Unit trains are effectively permanently coupled trains that make cyclical trips carrying commodities such as coal or grain. Dedicated trains are trains composed of one car type usually providing a carefully structured level of service, but they may consolidate traffic and require some switching. Examples of dedicated trains are solid trains of frozen concentrate orange juice between Florida and New Jersey, intermodal trains between Los Angeles and Chicago, and imported automobiles shipped from Pacific ports to inland distribution centers. Loose car or carload trains move a broad spectrum of mixed freight in a vast array of car types to and from every location on the network. Customers load and unload one or more cars at their private sidings. This requires frequent switching of cars between trains and adjustment of schedules to individual customer requests and levels of business activity. Loose car operations present very complicated combinatorial problems and consequently are frequently an operations research topic.

SERVICE DESIGN

The activities of train frequency setting and car routing are collectively categorized as *service design*, and are usually strategic, longer time horizon decisions. Medium time horizon or tactical decisions are composed of setting train departure and arrival times, and revising car routings to agree with train schedules. Short-term or operational decisions consist of assigning specific cars to specific trains, determining of specific train paths on the available tracks (and consequent further revision of specific timings),

and revising en route train paths in response to network delays. These plans are consolidated into the daily operating plan.

Typically in the literature a “train schedule” establishes the train routes (the set of lines and yards to traverse), train frequency and time at origin, destination and intermediate yards, and cars handled by a train (picked up or dropped off at intermediate yards). For example, Gorman [12,13] uses genetic search algorithms to determine the operating plan for intermodal services at BNSF predecessor Santa Fe. The train schedule does not establish an operational train path on the network as described in the section titled “Train Pathing”, but it does affect connection times for cars and, thus, service quality.

Cars may be individually routed to their destination on the next available train, or they may be held and consolidated to travel to common destinations together. The latter method is called anticipation classification or *blocking*, and dominates service design in North America not only because coupling and uncoupling is labor intensive, but also because the North American coupler design makes hump yard classification of some car types inadvisable. The challenge in service design is to create large enough blocks and trains to allow for economies of scale, but avoid excess handling and frequent train starts, while remaining feasible from both a yard and train capacity perspective and a customer service perspective.

The combined problem of assigning cars to blocks, assigning blocks to trains, establishing train frequencies, and setting departure times is vast in size and exceeds any current resolution capability. Instead, the problem is decomposed by activity and solved iteratively. Typically, the blocking plan focuses on minimizing handling costs and the distance traveled. The train plan brings service frequency and timing into account, thus including car delay costs and customer service considerations. Given a set of demands that must be serviced, the optimal train schedule and blocking plan minimizes cost of service subject to meeting customer service requirements. Ahuja *et al.* [14] offer a detailed explanation of these various levels of planning

in North American practice. Haghani [15] demonstrates in an early model formulation that significant performance improvements may be achieved by combining the car routing and train frequency problem.

Anticipation Classification or Blocking

In North America, rail freight's primary competitive advantage of rail is low cost over long distances, and blocking of cars supports this service strategy. Consequently, all of the literature in this section refers to North American applications. Anticipation classification or blocking reduces the frequency of hump yard classification, but incurs some additional delay while cars accumulate to form a block. Under this strategy, cars are consolidated and coupled by final destination very early, and a block is defined as a set of rail cars ready for departure from the same origin to the same destination.

Take, for example, a shipment from Bangor, Maine, to Louisville, Kentucky. It could be placed in a small block to Louisville, and not be reclassified, or could be placed in a large block to Cincinnati, where it would have to be switched to another block destined for Louisville. In the direct block, the car would experience a longer departure delay waiting for cars to form a block (because there is a limited amount of traffic between these two cities). In the separate blocks through Cincinnati, the car would enjoy frequent departures, but incur additional switching cost.

At the Canadian Pacific Railway, Ireland *et al.* [16] first define optimal blocks as a function of aggregate cargo flows, then define train schedules (departure times, not train paths) to support these blocks, simulate the schedules for reliability and yard inventories, and finally assign locomotives. Infeasibility or discovered alternatives at any stage may cause a revision of the prior activity solution. Crainic *et al.* [17] present an alternative sequence for Canadian National Railway, where train frequencies are defined first and then the blocking policy is defined upon the train schedule. The algorithm initializes at a high train frequency, successively reduces frequency, and generates a new blocking plan at each iteration, stopping at a predefined objective change. Newton *et al.* [18] focus

solely on the blocking activity without consideration of the train frequencies. Jha *et al.* [19] address the operational decision of assigning specific realized blocks to scheduled trains, where both the blocking policy and train schedule are predefined.

Blocking introduces a number of integer variables in formulations. For example, the assignment of blocks to trains is constrained by train capacity, potentially forcing a choice between dividing a block, dispatching a light train, and dispatching a heavy, underpowered train. The latter option affects over the road performance and introduces delay to other trains on the route. In addition, sometimes minimum block sizes are established, which imply some origin dwell time to accumulate rail cars. As blocking policy becomes more aggressive, car delays at the origin increase, larger yards are necessary to inventory cars, and car cycle times increase.

The blocking problem is a fertile ground for solution algorithms and strategy. Bodin *et al.* [20] present a nonlinear formulation that presumes a fixed train schedule, but is unable to solve practical problem sizes. Keaton [21] provides practical tips on the application of Lagrangian relaxation to the combined train frequency and car routing model. Martinelli and Teng [22] use artificial intelligence to generate fast solutions to an experimental example of the block to train assignment. Marín and Salmerón [23] compare simulated annealing, tabu search, and descending contour search heuristics for a generalized railway network problem, and Marín and Salmerón [24] develop statistical measures of the quality of the solutions generated.

Huntley *et al.* [25] present an actual implementation of simulated annealing to the weekly block scheduling problem at CSX. Ahuja *et al.* [26] describe very large-scale network search algorithms for blocking problems implemented at CSX, Norfolk Southern, and BNSF railways, under either unconstrained or predefined train frequencies. Kwon *et al.* [27] explicitly consider shipper delivery time expectations and train lengths, and apply a column generating solution algorithm. Barnhart *et al.* [28] apply both Lagrangian relaxation and column generation.

Individual Car Routing

Rail freight services in Europe typically cover much shorter distances and require a more punctual, responsive service. The infrastructure and dense passenger service schedule support and, in some cases, require operation of shorter, more frequent freight trains. There are also fewer technical limitations to the hump yard classification of cars in Europe. Consequently, blocking is less advantageous.

If blocking is dispensed with, a more even distribution of tonnage across trains may be achieved. Individual car routing allows smaller fractions of train capacity to be allocated. Ceselli *et al.* [29] describe the Swiss SBB Cargo Express service, which provides car load freight service within Switzerland on an overnight delivery cycle. They describe a hub and spoke network with two yards (hubs), and formulate a combined service design problem that assigns cars to trains, determines the train schedules (but not the dispatch paths), and respects capacity limits at the yards.

None of the models and service design methods yet described consider the return routing of empty cars. Campetella *et al.* [30] include the movement of empty cars in their solution to train frequency and car routing on the Italian FS network. The objective function is nonlinear, where costs decline as train frequency increases on a route because of a weighted improvement in the quality or reliability of service. The cost function is concave and the constraint set is linear and integer. Solutions from a genetic algorithm are described.

TRAIN PATHING

A train path is the complete authority for a train to use tracks and sidings from its origin to its destination. Track is usually divided into discrete controlled segments called *blocks*, and a train path is then a list of authorized blocks and times. The more highly utilized a network, and the more that punctual operation is desired, the more critical it is to configure and validate train paths in advance.

Train pathing is typically performed at either a strategic level or an operational level. If train paths have been configured and distributed in advance, then the trains are operating under a timetable. In the absence of a timetable, train paths are determined by dispatchers as the trains are en route. White and Krug [31] call this “timetable-free” or “improvised” operation.

The distinction between these two scenarios is increasingly blurred. Modern algorithms and computers allow computation of revised timetables (of varying quality) on demand. Many signal control systems provide automatic route setting functions that arrange groups of train paths within a limited area and time horizon. See Pahl [32] for a description of route setting and signaling technologies. Therefore, to provide a clear definition, we define timetable authority as any scenario in which a train possesses a defined and validated path before entering the network. A railway is operating under timetable if the majority of its trains operate under timetable authority.

The distinction between timetabling and “rescheduling” or “tactical” train path scheduling is largely unnecessary, as both activities require the same problem data, have the same constraints, and typically have similar objective functions. The primary difference is one of application and solution quality. Rescheduling typically must be performed quickly, and so the emphasis is on heuristics and quality suboptimal solutions returned in seconds. Consequently, we group these papers together.

Once again, environmental and market factors encourage distinct preferences in train pathing. European networks intermix freight trains with passenger trains and stringent punctuality is maintained. In addition, many of the national networks have transitioned to “open access”, where the management of the rail network and the operation of trains fall under separate business entities. Competing train operating companies share the same rail network. Operators negotiate and contract for train paths months in advance. Penalties are assessed for failure to provide a train path and failure

to conform to a train path (operator fault). Consequently, timetable operation is typical.

In contrast, vertical integration of infrastructure and operations is still the rule in North America. Traffic fluctuations lead to various train delays and cancellations. For example, blocking policy may lead to an inadequately powered train, which then has difficulty accelerating to scheduled speed. Passenger trains which share the freight rail network have considerable amounts of buffer time or “slack” included in their schedules to allow flexibility for these events. Consequently, although certain patterns may recur from day to day, the train paths are determined dynamically according to relative train priorities, over a shorter time horizon and smaller network segment, and not according to a predetermined timetable.

Timetable-Free Dispatching

Under a timetable-free operating plan, train paths and sequencing are determined by dispatchers as trains arrive within their region of supervision. Excepting passenger and select time sensitive freight trains, dispatchers have the authority to hold trains indefinitely at sidings according to their experience and judgment of what is feasible and best for the operating plan. It is very typical in these operations to find that locomotives, and train lengths are pressed to their limits, which further leads to variability in trip times. This variability in trip times makes adherence to a timetable very difficult, and thus argues for a timetable-free operation.

Under timetable-free operations, transit time over a route is not fixed and a variety of empirical and statistical methods are used to forecast trip times. Gorman [33] reviews a representative sample of North American lines and performs a linear regression of train delays as a function of dispatch conditions. Krueger [34] discusses a number of estimators of network capacity under an unscheduled, manual dispatch policy. Abbott [35] calculates the loss of train paths and incremental delay resulting from mixing high-speed and regular train traffic on a single direction route.

Hallowell [36] estimates delays to trains on either single or double track from simulations of a range of train frequencies and departure time variances under an advanced meet/pass planning algorithm. Carey and Kwieciński [37] devise probabilistic estimates of train delays on single direction lines as a function of dispatch headways and signal headways. Harker and Hong [38] devise probabilistic estimates of mean and variance of train delays on mixed single and double track lines where trains depart with some known variance around a scheduled time.

Dispatch decision making under timetable-free operation frequently requires the arbitration of conflicts between pairs of trains. Sauder and Westerman [39] describe a very early computer-aided dispatching system that estimated the delay cost of a set of dispatching alternatives. In many cases train values may be defined by linear functions, and thus the dispatching decision space is convex. Kraft [40] takes advantage of this property with a branch-and-bound decision algorithm that seeks to eliminate inferior dispatch options early in the decision process.

Jovanović and Harker [41] consider in detail what factors should be included in a dispatch objective. In particular, they express concern that the sum of weighted train delays is too narrow an objective for efficient dispatch. Şahin [42] models the dispatcher’s cognitive decision in the meet/pass planning problem and demonstrates that computer algorithms can significantly improve the quality of dispatching plans. Mills *et al.* [43] enhance the dispatch decision to consider fuel consumption. They propose both discrete-time network and nonlinear real-valued time models, and find that the simpler discrete-time model returns high quality solutions. Although intended for dispatching, this last model could also be classified as a timetabling tool.

Timetable Generation

The earliest timetabling models and solution methods were limited to long-term planning both because of the time required to obtain a solution, and because the solutions were frequently imperfect and required manual revision. Szpigel [44] is frequently cited as the

earliest example of global timetable optimization. That model minimizes the weighted sum of train transit times on a single track line, but does not constrain the capacity of sidings or stations. Direct manipulation of the integer variables is avoided by dynamically adding constraints at the branches of the solution tree, an application of the method of Greenberg [45].

Two model forms appear frequently. The first is the binary integer occupancy problem (BIOP), where a discrete-time network is defined and each binary decision variable represents the potential occupation of a track segment by one train during one discrete-time interval. The second is the mixed integer sequencing linear program (MISLP), where timing variables are real valued, and binary variables decide the precedence of pairs of trains at conflicted track segments. Both these model classes may be viewed as responses to the lack of siding or station capacity constraints in Szpigel. The BIOP formulation first appears in Mees [46], along with an approximate solution algorithm based upon Lagrangian relaxation.

The MISLP formulation first appears in Carey [47,48]. However, this model is too complex for optimal solution, and the authors describe algorithms that schedule trains sequentially and iteratively, with methods derived from observed practice of human dispatchers. Kraay and Harker [49] formulate a MISLP model with a nonlinear objective that considers the crew schedules and car blocking connections typical of North American operations. A prerequisite to the application of this model is the abandonment of timetable-free operation [50]. North American railways were unwilling to make this change in the 1990s; consequently, this model was not placed into service. Higgins *et al.* [51] formulate an MISLP model with a nonlinear objective of tardiness and fuel costs and decompose the problem into meet/pass planning and timing steps.

An open access network motivates Brannlund *et al.* [52] to describe sequential, iterative solution procedures for another BIOP formulation. In this article, Lagrangian dual variables serve to estimate the value or market price of a train path. Caprara

et al. [53] formulate a similar BIOP model, where train path enumeration is integrated within the primal problem. An application of this model specific to freight rail appears in Cacchiani *et al.* [54]. Borndörfer *et al.* [55] enhance the formulation of Caprara *et al.* to consider clique sets of mutually exclusive train path segments. Timetables are generated iteratively as part of a Parkes “i-bundle” auction of train paths.

Computing power and solution algorithms for multicommodity flow networks have improved significantly in the last decade, leading to more aggressive models and problem sizes. Ahuja *et al.* [14] present a BIOP model along with a greedy solution heuristic and a constraint tightening heuristic. Zhou and Zhong [56] apply an MISLP formulation to a double track high-speed railway in China, and develop a Pareto optimal frontier for two distinct objectives: minimization of the variance of interdeparture times and minimization of the total travel time. Harrod [57] visualizes the railway network as a hypergraph and formulates a binary decision variable model that constrains transitions between track segments over a time interval. Harrod [58] demonstrates that increasing the speed of priority trains relative to other trains increases the number of possible train paths on a single track line. Caimi *et al.* [59] demonstrate that the feasible solution set of a BIOP formulation may be reduced by deleting train paths that are dominated by other paths.

In recent years, the algorithms and available computing power have improved sufficiently that real-time, dynamic timetable solutions are available. These are usually presented as methods to respond to train delays and network disruption, but they also support impromptu additions of extra trains to existing timetable dispatched networks. Törnquist and Persson [60] address the rescheduling problem with an MISLP model, and find that the best solution strategy is to bound the problem by limiting the number of changes to relative train priorities. Later, Törnquist [61] found that local optimization over a rolling time horizon was competitive to optimization over the full time horizon.

Alternative graph theory has played a role in dynamic rescheduling. Mascis and Pacciarelli [62] present an algorithm called *Avoid Maximum Current Cmax* (AMCC), which they find suitable for the solution of train timetabling problems formulated as alternative graphs. Mazzarello and Ottaviani [63] apply this formulation and algorithm to dynamic rescheduling after delays on two Netherlands Railways examples. D'Ariano *et al.* [64] extend the AMCC algorithm further by compiling sets of "static implications," or sets of implied constraints, and demonstrate a 120 s solution time on problems drawn from the Netherlands Railways Schiphol district.

Lastly, the periodic event scheduling problem (PESP) should be mentioned. Primarily intended for passenger railways with cyclical schedules, it could also be applied to cyclical scheduled freight operations or networks, where freight paths must integrate with cyclical passenger schedules. The method originates with Serafini and Ukovich [65] and is first applied by Odijk [66]. Notable applications of PESP include the Netherlands Railways Design of Network Services (DONS) timetabling system [67], which was the 2008 Franz Edelman Award recipient [7], and the Berlin U-bahn network [68].

CONCLUSION AND FUTURE TRENDS

The service design problem, which combines both car classification and train scheduling, is at the heart of rail operations planning, and is one of the most difficult. The problem is nearly always expressed as a very large-scale dynamic network. The magnitude of these problems and their economic significance have motivated many fundamental achievements in operations research, such as network flow theory, column generation, and alternative graph theory. When the problems exceed reasonable solution capabilities, they are frequently decomposed into the blocking plan problem, the train frequency problem, and either the timetabling or dispatching problem. These problems are interrelated, and the unique strategic and environmental

concerns of each railway determine the order of their solution.

Significant technological and economic conditions differentiate the most common practices in North America and Europe. Long distances, tolerant clearances, long train lengths, and a preponderance of coal and grain motivate North American operators to emphasize consolidation of car loads into blocks. Economies of scale are more important than punctuality, so that stochastic analysis of transit times dominates over timetabling. We survey literature that addresses the blocking problem typical of North America and the dispatch problem of "meet/pass" planning, where trains proceed across the network without a predefined train path or timetable. The very large scale of blocking (network) problems motivates applications of Lagrangian relaxation, column generation, and genetic search.

European railways attract higher value shipments, operate over shorter distances, and must integrate freight trains with dense passenger services. Punctuality is emphasized and timetabled train paths are enforced. It is more common to route shipments as individual car loads through a hub and spoke system, without blocking. Timetabling, the advance determination of train paths, is common for freight traffic in Europe. This is frequently mandatory in order to share a railway line with scheduled passenger service, and also provides the benefit of higher service quality. The majority of timetabling models may be categorized as either managing the occupancy of discrete-timed track segments with binary variables or equating the resource priority of trains with binary variables and the timing with real-valued variables. On lines dominated by cyclical passenger traffic, the periodic event scheduling problem formulation may also determine freight train paths.

Research in this area has focused almost exclusively on planning, whereas only little research has been conducted on operating plan execution. Since operational flexibility can be valuable, and traffic demand is notoriously volatile and hard to predict, on-the-fly

schedule adjustments are common. For example, when disturbances occur to car routing plans or train frequencies, should an attempt be made to return to the prior plan, or should an alternate plan be formulated?

Commercial vendors have made significant advances in their train pathing and dispatch route setting systems, but little has been published on these systems. In at least one instance known to the authors, there was a desire to protect commercial intellectual property. Automated route setting systems increasingly can optimize groups of trains over longer time horizons according to defined global objectives. As these systems become more sophisticated, the distinction between meet/pass planning and network train pathing (timetable design) narrows.

Improvements in signaling and communications technology are underway in both North America and Europe that will expand the range of problems to be addressed by operations research and/or eliminate or revise standard problem constraints. In North America, Federal law asserts that positive train control, the direct monitoring of train positions by live communication links, shall be required within 10 years. On some networks this will change dispatch strategy. Long-term European plans are to standardize the signaling systems that currently limit circulation of locomotives. This will eliminate many constraints on train frequency and international services.

REFERENCES

1. Cordeau JF, Toth P, Vigo D. A survey of optimization models for train routing and scheduling. *Transp Sci* 1998;32(4): 380–404.
2. Haghani A. Rail freight transportation: a review of recent optimization models for train routing and empty car distribution. *J Adv Transp* 1987;21:31–38.
3. Assad AA. Models for rail transportation. *Transp Res Part A* 1980;14:205–220.
4. Crainic TG. Service network design in freight transportation. *Eur J Oper Res* 2000;122: 272–288.
5. Caprara A, Kroon L, Monaci M, *et al.* Passenger railway optimization. In: Barnhart C, Laporte G, editors. Volume 14, *Transportation; handbooks in operations research and management science*. Amsterdam: North-Holland Publishing Co.; 2007. pp. 129–187.
6. Huisman D, Kroon L, Lentink R, Vromans M. Operations research in passenger railway transportation. *Stat Neerl* 2005;59(4): 467–497.
7. Kroon L, Huisman D, Abbink E, *et al.* The new Dutch timetable: the OR revolution. *Interfaces* 2009;39(1):1–12.
8. Bureau of Transportation Statistics. *National transportation statistics*. Washington, DC: U.S. Department of Transportation; 2009.
9. CER. Rail freight quality progress report 2007/2008. Bruxelles: Community of European Railway and Infrastructure Companies; 2008.
10. Vassallo JM, Fagan M. Nature or nurture: why do railroads carry greater freight share in the United States than in Europe? *Transportation* 2007;34:177–193.
11. International Union of Railways. *International Railway Statistics, Full Year 2008*. Paris: International Union of Railways (UIC); 2008. Available at <http://www.uic.org/training/spip.php?article1348>. Retrieved 2010 Feb 4.
12. Gorman MF. An application of genetic and tabu searches to the freight railroad operating plan problem. *Ann Oper Res* 1998;78: 51–69.
13. Gorman MF. Santa Fe railway uses an operating-plan model to improve its service design. *Interfaces* 1998;28(4):1–12.
14. Ahuja RK, Cunha CB, Şahin G. Network models in railroad planning and scheduling. In: Greenberg HJ, editor. *Tutorials in operations research, emerging theory, methods, and applications*. Hanover (MD): Institute for Operations Research and the Management Sciences; 2005. pp. 54–101.
15. Haghani AE. Formulation and solution of a combined train routing and makeup, and empty car distribution model. *Transp Res Part B* 1989;23B(6):433–452.
16. Ireland P, Case R, Fallis J, *et al.* The Canadian pacific railway transforms operations by using models to develop its operating plans. *Interfaces* 2004;34(1):5–14.

17. Crainic T, Ferland JA, Rousseau JM. A tactical planning model for rail freight transportation. *Transp Sci* 1984;18(2):165–184.
18. Newton HN, Barnhart C, Vance PH. Constructing railroad blocking plans to minimize handling costs. *Transp Sci* 1998;32(4):330–345.
19. Jha KC, Ahuja RK, Şahin G. New approaches for solving the block-to-train assignment problem. *Networks* 2008;51(1):48–62.
20. Bodin LD, Golden BL, Schuster W, *et al.* Model for the blocking of trains. *Transp Res Part B* 1980;14B(1–2):115–120.
21. Keaton MH. Designing optimal railroad operating plans: Lagrangian relaxation and Heuristic approaches. *Transp Res Part B* 1989;23B(6):415–431.
22. Martinelli DR, Teng H. Optimization of railway operations using neural networks. *Transp Res Part C* 1996;4(1):33–49.
23. Marín A, Salmerón J. Tactical design of rail freight networks. Part I: exact and heuristic methods. *Eur J Oper Res* 1996;90(1):26–44.
24. Marín A, Salmerón J. Tactical design of rail freight networks. Part II: local search methods with statistical analysis. *Eur J Oper Res* 1996;94(1):43–53.
25. Huntley CL, Brown DE, Sappington DE, *et al.* Freight routing and scheduling at CSX transportation. *Interfaces* 1995;25(3):58–71.
26. Ahuja RK, Jha K, Liu J. Solving real life blocking problems. *Interfaces* 2007;37(5):404–419.
27. Kwon OK, Martland CD, Sussman JM. Routing and scheduling temporal and heterogeneous freight car traffic on rail networks. *Transp Res Part E* 1998;34(2):101–115.
28. Barnhart C, Jin H, Vance PH. Railroad blocking: a network design application. *Oper Res* 2000;48(4):603–614.
29. Ceselli A, Gatto M, Lübbecke ME, *et al.* Optimizing the cargo express service of Swiss Federal Railways. *Transp Sci* 2008;42(4):450–465.
30. Campetella M, Lulli G, Pietropaoli U, *et al.* Freight service design for an Italian Railways Company. *ATMOS 2006: 6th Workshop on Algorithmic Methods and Models for Optimization of Railways*. 2010. <http://drops.dagstuhl.de/opus/volltexte/2006/685> Accessed 2010 Jan 17.
31. White T, Krug A. Managing railroad transportation. Mountlake Terrace, Washington (DC): VTD Rail Publishing; 2005.
32. Pachl J. Railway operation and control. Mountlake Terrace, Washington (DC): VTD Rail Publishing; 2002.
33. Gorman MF. Statistical estimation of railroad congestion delay. *Transp Res Part E* 2009;45(3):446–456.
34. Krueger H. Parametric modeling in rail capacity planning. In: Farrington PA, Nemphard HB, Sturrock DT, *et al.*, editors. *Proceedings of the 1999 Winter Simulation Conference*. New York: ACM; 1999. pp. 1194–1200.
35. Abbott RK. Operation of high speed passenger trains in rail freight corridors. FRA-OR&D-76-07. Washington (DC): Federal Railroad Administration, U.S. Department of Transportation; 1975 Sep.
36. Hallowell SF, Harker PT. Predicting on-time performance in scheduled railroad operations: methodology and application to train scheduling. *Transp Res Part A* 1998;32(4):279–295.
37. Carey M, Kwieceński A. Stochastic approximation to the effects of headways on knock-on delays of trains. *Transp Res Part B* 1994;28B(4):251–267.
38. Harker PT, Hong S. Two moments estimation of the delay on a partially double-track rail line with scheduled traffic. *J Transp Res Forum* 1990;31(1):38–49.
39. Sauder RL, Westerman WM. Computer aided train dispatching: decision support through optimization. *Interfaces* 1983;13(6):24–37.
40. Kraft ER. A branch and bound procedure for optimal train dispatching. *J Transp Res Forum* 1987;28(1):263–276.
41. Jovanović D, Harker PT. A decision support system for train dispatching: an optimization-based methodology. *J Transp Res Forum* 1990;31(1):25–37.
42. Şahin I. Railway traffic control and train scheduling based on inter-train conflict management. *Transp Res Part B* 1999;33(7):511–534.
43. Mills RGJ, Perkins SE, Pudney PJ. Dynamic rescheduling of long-haul trains for improved timekeeping and energy conservation. *Asia Pac J Oper Res* 1991;8:146–165.
44. Szpigel B. Optimal train scheduling on a single track railway. *Operational Research '72. Proceedings of the Sixth IFORS International Conference on Operational Research*. London: North-Holland Publishing Co.; 1973. pp. 343–352.

45. Greenberg HH. A branch-bound solution to the general scheduling problem. *Oper Res* 1968;16(2):353–361.
46. Mees AI. Railway scheduling by network optimization. *Math Comput Model* 1991;15(1): 33–42.
47. Carey M. A model and strategy for train pathing with choice of lines, platforms, and routes. *Transp Res Part B* 1994;28B(5): 333–353.
48. Carey M. Extending a train pathing model from one-way to two-way track. *Transp Res Part B* 1994;28B(5):395–400.
49. Kraay DR, Harker PT. Real-time scheduling of freight railroads. *Transp Res Part B* 1995;29B(3):213–229.
50. Jovanović D, Harker PT. Tactical scheduling of rail operations: the SCAN I system. *Transp Sci* 1991;25(1):46–64.
51. Higgins A, Kozan E, Ferreira L. Optimal scheduling of trains on a single line track. *Transp Res Part B* 1996;30(2):147–161.
52. Brännlund U, Lindberg PO, Nöu A, *et al.* Railway timetabling using Lagrangian relaxation. *Transp Sci* 1998;32(4):358–369.
53. Caprara A, Monaci M, Toth P, *et al.* A Lagrangian heuristic algorithm for a real-world train timetabling problem. *Discrete Appl Math* 2006;154:738–753.
54. Cacchiani V, Caprara A, Toth P. Scheduling extra freight trains on railway networks. *Transp Res Part B* 2010;44:215–231.
55. Borndörfer R, Grötschel M, Lukac S, *et al.* An auctioning approach to railway slot allocation. *Compet Regul Netw Ind* 2006;1(2): 163–196.
56. Zhou X, Zhong M. Bicriteria train scheduling for high-speed passenger railroad planning applications. *Eur J Oper Res* 2005;167: 752–771.
57. Harrod S. Modeling network transition constraints with hypergraphs. *Transp Sci* 2010. In press.
58. Harrod S. Capacity factors of a mixed speed railway network. *Transp Res Part E* 2009;45:830–841.
59. Caimi G, Burkolter D, Herrmann T, *et al.* Design of a railway scheduling model for dense services. *Netw Spat Econ* 2009;9:25–46.
60. Törnquist J, Persson JA. N-tracked railway traffic re-scheduling during disturbances. *Transp Res Part B* 2007;41:342–362.
61. Törnquist J. Railway traffic disturbance management - an experimental analysis of disturbance complexity, management objectives and limitations in planning horizon. *Transp Res Part A* 2007;41:249–266.
62. Mascis A, Pacciarelli D. Job-shop scheduling with blocking and no-wait constraints. *Eur J Oper Res* 2002;143:498–517.
63. Mazzarello M, Ottaviani E. A traffic management system for real-time traffic optimisation in railways. *Transp Res Part B* 2007;41:246–274.
64. D'Ariano A, Pacciarelli D, Pranzo MA. Branch and bound algorithm for scheduling trains in a railway network. *Eur J Oper Res* 2007;183:643–657.
65. Serafini P, Ukovich W. A mathematical model for periodic scheduling problems. *SIAM J Discrete Math* 1989;2(4):550–581.
66. Odijk MA. A constraint generation algorithm for the construction of periodic railway timetables. *Transp Res Part B* 1996;30(6): 455–464.
67. Kroon LG, Peeters LWP. A variable trip time model for cyclic railway timetabling. *Transp Sci* 2003;37(2):198–212.
68. Liebchen C. The first optimized railway timetable in practice. *Transp Sci* 2008;42(4): 420–435.

OPERATIONS RESEARCH IN AUSTRALIA

ERHAN KOZAN
School of Mathematical Sciences,
Queensland University of
Technology, Brisbane,
Queensland, Australia

Operations research (OR) was initially introduced to Australia from Britain in about 1942 by the Australian Army. Early studies were restricted to weapons research. Later studies were in administrative and logistical areas, which were also valuable to the war effort. After the war, OR activities in the army were allowed to run down and it began to be used by civilians in the 1950s. According to Mellor [1], the first two civilian OR records were by Bowen (1948) on "Operational Research into Air Traffic Problem," and Muncey and Hutson (1953) on "The Effect of Floor on Foot Temperature."

In the 1960s, OR practitioners set up local OR groups in Melbourne and Sydney and this pattern was repeated in other major urban centers including the smaller cities of Wollongong and Newcastle, which had substantial industry. Informal links between the Melbourne and Sydney groups led to the establishment of the Australian Joint Council for Operational Research (AJCOR). As a result of a meeting held in November 1971, the council of the Statistical Society of Australia agreed to the initiative of its OR section to form an autonomous and national society. Then on January 1 1972, members of the OR section, also being members of AJCOR, agreed to have AJCOR superseded by a new national society which was to be named the Australian Society for Operations Research Incorporated (ASOR) [2].

Currently ASOR has about 250 members nationwide and like other national societies, it is affiliated to the International Federation of Operational Research Societies (IFORS). There are also a number of regional groups within IFORS. For instance, ASOR belongs

to Asian Pacific Operational Research Societies (APORS). ASOR serves the professional needs of OR analysts, managers, students, and educators by publishing a national bulletin and chapter newsletters. The society also serves as a focal point for operations researchers to communicate with each other and to reach out to other professional societies.

The Society's objectives are as follows:

- to foster the development of the science of OR;
- to promote the application of OR wherever appropriate;
- to facilitate the widest possible exchange of information and ideas on OR and related subjects;
- to define standards of proficiency in OR;
- to promote and facilitate the study of OR.

ASOR has had 18 presidents since 1971, their names and terms of office are given in Table 1.

The ASOR bulletin was founded in 1981 through the initiative of Ian Sadler, who was at that time the National President of ASOR. He launched a national newsletter to establish communication between chapters, and this eventually became the ASOR Bulletin. The first editor of the bulletin was Carlton Scott, who produced Volume 1 comprising three issues, in 1981. The ASOR Bulletin aims to provide news of ASOR activities, book reviews, Australian problem descriptions and solution approaches, and a forum on topics of interest to OR practitioners, researchers, academics, and students. The refereed articles are peer reviewed by at least two independent experts in the field. The ASOR Bulletin is published as four volumes per year in March, June, September, and December.

ASOR organizes its national conference every two years to keep up-to-date with OR issues in Australia and overseas. It is a good platform for discussions with practitioners

Table 1. Past Presidents of ASOR

President	Year	President	Year	President	Year
W.B Kennedy	1972–1973	J. Flanagan	1983–1985	E. Kozan	1998–1999
T.G. O'Meally	1973–1974	C. Pearce	1985–1987	C. Pearce	2000–2001
E.J. Buckman	1974–1976	P. Evans	1987–1989	S. Perkins	2002–2003
R.B. Potts	1976–1978	C.S. Newton	1989–1991	L. Caccetta	2004–2005
D. Atkinson	1978–1979	C. Malanos	1991–1993	B. Nath	2006–2007
J. Adams	1979–1980	P. Lochert	1993–1995	E. Kozan	2008–2009
I. Sadler	1980–1983	S. Weal	1996–1997		

and researchers. ASOR has organized 20 conferences since 1972 and rotated these conferences between the branches. Their distribution to the branches is as follows: South Australia, four times; Queensland, four times; Victoria, five times; New South Wales, three times, Australian Capital Territory, two times; and Western Australia, two times. Branches also hold one-day specialized workshops, industry workshops, and student conferences.

The Ren Potts award of the ASOR is intended to recognize individuals who have made outstanding contributions to theory or practice of OR in Australia. It is a national award restricted to Australian residents. Ren Potts was a leading researcher in traffic flow and later in other areas of OR, and was also very active in the formation and ongoing function of ASOR. He served as one of the first national presidents. The first worthy recipients of this medal were Dr. Bruce Craven, Prof. Charles Pierce, Prof. Pra Murthy, Prof. Jerzy Filar, and Prof. Santos Kumar. These awards are conferred at a special ceremony organized as part of ASOR's national conferences.

The ASOR National Council established an encouragement medal for new researchers, who have completed a research degree at the Master's or Ph. D. level in OR or in a related field with significant contribution to the advancement of OR from any Australian university. Normally those candidates will be considered who have completed all requirements for their degree between the two consecutive ASOR national conferences. The awards are conferred in a special ceremony held at the national conferences.

ASOR official web-site (<http://www.asor.org.au>) was set up in 2006 to present the latest information about the members, ASOR bulletin, constitution, by-laws, news, conferences, important events, and links.

The ASOR Inc. has fostered the profession of OR in Australia since 1972 and has served as a platform for exchange of ideas between practitioners. The initial enthusiasm evident in the 1970s has gradually declined and this has seen a movement away from the centralized OR group, although a number of large organizations such as BHP, Qantas, Queensland Rail, Commonwealth Bank, Shell, Queensland Transport, and AECOM still support small OR groups or individual OR practitioners. Developments in OR and the need for quantitative problem solving continue to flourish in the private and public sectors. OR continues to be applied under many guises with practitioners operating under related fields of practice, so that the application of OR is never diminishing.

Despite the loss of groups within the private and public enterprises specifically dedicated to OR, ASOR continued to grow in the 1980s and 1990s with an increasing academic component as time progressed. As the university sector expanded in Australia, OR became increasingly identified with academia. Practical issues in all areas of OR are addressed by our active researchers from the Commonwealth Scientific and Industrial Research Organization (CSIRO) and the Defense Science and Technology Organization (DSTO).

Many mathematics departments in Australia offer OR subjects at the undergraduate level and a doctoral program for graduate students in OR. Many OR subjects are also offered under different names within various departments of the universities. For graduate

students in OR, there are significant opportunities for part-time employment in many universities. In addition to providing income for the student, the experience gained by working on real problems can be invaluable to a future career in OR.

Till date, in the field of practice, many OR procedures have been standardized and practitioners commonly apply them. Practitioners within the private and public enterprises, who are affiliated with ASOR or similar groups within overseas parent organizations, continue to improve OR procedures thereby making strategic and tactical processes more efficient.

The importance of OR continues to grow since it is an indispensable tool for

solving difficult problems in the mining and resources sector, transport and communications, health and medicine, defense, and within the service sector. OR continues to play an important role in industry and will remain very influential in Australia's prosperity, particularly when considering the carbon-constrained future which lies ahead.

REFERENCES

1. Mellor DP. Australia in the war of 1939–1945. The role of science and industry. Adelaide, Australia: Griffin Press; 1958.
2. Hoffman D. 27 years ago. ASOR Bull 1998;17 (2):2–5.

OPERATIONS RESEARCH IN DATA MINING

SHOUYI WANG
WANPRACHA ART CHAOVALITWONGSE
Department of Industrial and Systems
Engineering, Rutgers University,
Piscataway, New Jersey

ONUR SEREF
Department of Business Information
Technology, Virginia Polytechnic
Institute and State University,
Blacksburg, Virginia

INTRODUCTION

Data analysis has become an essential part of today's world in the past decade. It is estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster [1]. With this explosion in data and information, data mining (DM) techniques have become ubiquitous with an upsurging interest from both academia and industry. The methods applied to extracting patterns from data have a long history for centuries, such as sorting data manually or using various hypothesis driven methods. However, it is extremely difficult to transform large volume of data into valuable knowledge by the traditional means of analysis in modern times. This motivates the development of modern DM methods, which are designed to discover meaningful representations or structures of data using pattern recognition, statistical, and mathematical techniques. Generally, DM techniques are first trained by sample data sets, and then scale the acquired knowledge up to large databases under the assumption that they have a structure similar to the training data. DM methods can be broadly categorized as either supervised or unsupervised learning. In supervised methods, the algorithm is provided with a set of training data whose class attributes are known. If the class information is not

available, unsupervised techniques can be employed, such that clustering algorithms are designed to discover underlying groupings of data instances, and rule association induction aims to explore associations and correlations among data instances.

Compared to DM, which is a relatively young discipline, the roots of operations research (OR) may be traced to early research of the optimization problems of transportation and mail system by the English mathematician Charles Babbage (1791–1871). Today, there are numerous well-developed OR techniques that are used routinely to solve problems in a wide range of application areas. OR and DM are two independent fields with distinct objectives. However, the intersection between OR and DM is also quite broad. Each of them could benefit from making use of the other. For example, OR techniques could enhance the efficiency of a DM process by embedding various optimization tools [2], and DM could provide a concise and meaningful data space that may greatly facilitate an OR process [3,4]. Recently, a growing interest in the integration of OR and DM can be observed in both OR and DM literature [5].

Primary tools of OR include optimization techniques, queueing theory, game theory, graph theory, probabilistic modeling, and simulation. The major focus of this article is the use of optimization techniques in DM. We review the state-of-the-art DM approaches and discuss how optimization techniques can be applied to solve these DM problems. The DM methods we discuss here include supervised methods, such as neural networks (NNs), support vector machine (SVM), K-nearest neighbor (KNN), decision tree, Bayesian learning, and unsupervised methods such as clustering and association rule induction. Although there are many other algorithms, these methods are among the most influential algorithms in DM research community. For each method, we introduce the basic concepts, review therecent advances on the algorithms, and

discuss the application of optimization techniques to these DM algorithms.

SUPERVISED LEARNING METHODS

Supervised learning approaches construct a predictive function from training data, which consists of sets of input–output pairs. They first find a global mapping between inputs and outputs, and then make predictions of new inputs that it has never seen by their generalization capability. Generally, a good generalization of supervised methods requires a training data set that is sufficiently large and representative of all cases so that a valid general mapping between outputs and inputs can be found. Supervised learning is the most frequently used paradigm in DM, and a large number of supervised learning algorithms have been developed in the last decades. The most popular ones are discussed in the following sections.

Neural Networks (NNs)

NNs, inspired by the structure of biological neurons, have been widely used to solve many classification problems where there exists sufficient amount of training data. NNs have gained this popularity owing to their powerful capacity to model extremely complex nonlinear functions and their relatively easy use with well-developed training algorithms. To train a NN, one first presents to the network a set of training data (a collection of input–output pairs), and then adjusts the weights of the NN in such a way that the prediction error between the desired and actual outputs of the training data is minimized. The mean squared error is the most commonly used cost function to represent prediction error, which is formulated by

$$E = \frac{1}{N} \sum_{i=1}^N (f(x_i) - y_i)^2, \quad (1)$$

where x_i, y_i are the input and output of training data, $f(x_i)$ is the output of NN with respect to the input x_i , and N is the number of training data.

To minimize prediction error, various update rules have been developed for training NNs, such as Perceptron learning rule and Widrow–Hoff rule [6]. Rumelhart *et al.* [7] proposed a pioneering work to apply a gradient descent–based optimization method to update the weights of a multilayer NN. This method was later on developed into the most well-known and widely used training algorithm of NNs called *backpropagation* (BP). BP employs error derivatives of a NN’s weights during training process and calculates how error changes as each weight is increased or decreased slightly. In general, the update rule of BP is given by

$$\omega_j = \omega_j - \alpha \frac{\partial E}{\partial \omega_j}, \quad (2)$$

where α is the learning rate, which controls the relative size of change in weights at every iteration, and $\frac{\partial E}{\partial \omega_j}$ is the partial derivative of error cost function E with respect to weight ω_j . BP has become quite popular in practice since it can often find a good set of weights in a reasonable amount of time. There have been many successful applications of BP in science, engineering, finance, and other disciplines [8].

Since BP training is a gradient descent process, it often gets trapped into local minima and thus is very inefficient in searching for the global optimum in a NN’s weight space. In recent years, genetic algorithms (GAS), as powerful global optimization tools, have been increasingly applied in optimizing a NN [9]. It is also very useful to solve the problems where gradient information is either not available or costly to obtain [10].

Simulated annealing techniques are also applied to gradient descent methods to help move out the local optima. However, these algorithms may experience cycling. An excellent alternative is to use tabu search introduced by Glover [11] in 1989. This method applies an iterative greedy search algorithm using the memory of past visited points. The tabu search imposes a “tabu” on some subset of the searching space so as to avoid making mistakes a second time. Battiti *et al.* [12] applied the concept of tabu search to train a NN, and their results demonstrated that

tabu search was effective in continuing the search after local minima and was also robust to random initial conditions.

Support Vector Machines (SVMs)

SVMs are an assembly of supervised machine learning algorithms with a rapidly growing spectrum of application areas, especially in DM. The widespread acceptance of SVMs are rooted in their robustness arising from the strong fundamentals of statistical learning theory (SLT) developed by Vladimir Vapnik [13,14]. SLT explains the learning process from a statistical point of view by empirical risk minimization and generalization. A major component of SLT is the Vapnik–Chervonenkis (VC) dimension of a statistical learning algorithm that reflects the algorithm’s capacity for learning complex data.

SVMs are most frequently used in classification problems [15,16]. The fundamental problem in SVM classification consists of two sets of instances as data points in $\mathfrak{N}^n$, which are referred to as *training* data with *positive* and the *negative* labels, and a decision boundary that separates them. Given positive and negative instances with nonoverlapping polytopes, the simplest form of a decision boundary is a hyperplane in $\mathfrak{N}^n$ that separates the two polytopes. One can find infinitely many such hyperplanes. However, STL shows that the hyperplane that maximizes its distance from the Closest point from these polytopes has the minimal empirical error and the best generalization capacity. This distance is referred to as the *margin*. The hyperplane is represented with a normal vector $\mathbf{w}$ and a bias term b , both of which can be multiplied by any scalar $\lambda \neq 0$ without changing the hyperplane. The distance of an instance $\mathbf{x} \in \mathfrak{N}^n$ from the hyperplane is $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$, where $\langle \cdot \rangle$ is the dot product. This distance is called the *functional* distance, whereas the (actual) *geometrical* distance is measured as $f(\mathbf{x})/\|\mathbf{w}\|$. The class of hyperplanes with a unit functional margin are called *canonical* hyperplanes and are used in modeling SVMs. For a canonical hyperplane, the geometric margin is equal to $1/\|\mathbf{w}\|$.

Given m pairs of training data $(\mathbf{x}_i, y_i)$, $i = 1, \dots, m$, where $y_i \in \{1, -1\}$ are the labels,

the hyperplane that maximizes the margin can be found by solving the following convex minimization problem:

$$\min_{\mathbf{w}, b} \left\{ \|\mathbf{w}\|^2/2 : y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \right. \\ \left. i = 1, \dots, m \right\}. \quad (3)$$

This problem, which is also known as the *maximal margin classifier*, has a feasible solution when the data is separable, which usually is not the case in real-life problems. In the case of linearly inseparable training data, a slack term is added to each constraint and penalized in the objective as shown in the following problem, which is known as *soft margin classifier*.

$$\min_{\mathbf{w}, b} \left\{ \|\mathbf{w}\|^2/2 + C \sum_{i=1}^n \xi_i : y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \right. \\ \left. \xi_i \geq 0, i = 1, \dots, m \right\}. \quad (4)$$

Following Lagrangian dual of this formulation actually paves the way for nonlinear classification and SVM implementations.

$$\max_{\alpha} \left\{ \sum_{i=1}^n \alpha_i - 1/2 \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle : \right. \\ \left. \sum_{i=1}^n y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, \dots, m \right\}. \quad (5)$$

In the solution to the dual problem, the instances with $0 < \alpha_i$ are the support vectors, and from the duality theory, we can determine the normal to the hyperplane as $\mathbf{w} = \sum_i y_i \alpha_i \mathbf{x}_i$. The instances that satisfy $0 < \alpha_i < C$ lie exactly at the margin with a target distance of $1/\|\mathbf{w}\|$. From complementary slackness, the constraints for these instances are binding without requiring any slack, from which the bias b can easily be calculated. The label for an instance whose class label is unknown can be determined with the

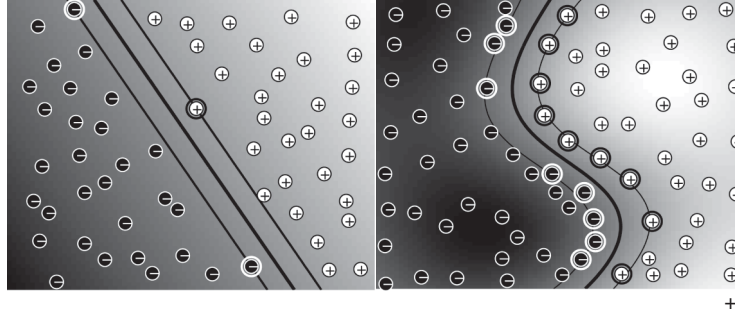


Figure 1. Linear and nonlinear SVM classification with highlighted support vectors on the class margins.

following function,

$$\text{class}(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^n y_i \alpha_i \langle \mathbf{x} \cdot \mathbf{x}_i \rangle + b \right). \quad (6)$$

The dual formulation and the distances between instances as dot products play the central role for expanding the capacity of an SVM classifier by implicitly mapping the data into a higher dimensional space. Such a mapping $\mathbf{x} \mapsto \phi(\mathbf{x})$ is done implicitly, since the dual formulation requires the dot product of the mapped points, $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$. Here, K is the *kernel* and can take many linear and nonlinear forms as long as it satisfies certain geometrical properties in the mapped space [15]. Given the new nonlinear distances through the kernel function, the rest is to find a linear separation in the mapped higher dimensional space. The calculations of these distances are done implicitly by the kernel function. Some of the most frequently used kernel functions are linear, $\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$, polynomial, $(\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1)^d$, Gaussian radial basis, $\exp(-\|\mathbf{x} - i - \mathbf{x}_j\|^2 / 2\sigma^2)$ and sigmoid, $\tanh(\kappa \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + c)$. In Fig. 1, examples of linear and nonlinear separation are given with highlighted support vectors on the class margins.

Although the underlying QP optimization problem for training SVM is convex with linear constraint, solving large problems become impractical with conventional quadratic optimization methods. Chunking and decomposition methods have been proposed to make SVM training practical [17,18]. The current

method used in most SVM implementations today is sequential minimal optimization (SMO). SMO is an extreme decomposition method that iteratively solves the QP problem two variables at a time, whose solution can be found analytically [19,20].

There are variations of SVM formulations based on penalization of the misclassification. The loss function in Equation (4) is known as *hinge-loss* or *linear* penalty, which can be replaced by a quadratic loss function. In this case, the penalty parameter is transferred to the objective in the dual, instead of appearing as an upper bound on the dual variables. Another possible variation is to maximize L_1 -norm of the margin, which provides a linear programming formulation with sparse solutions [21]. The sparsity of the solution is usually interpreted from a feature selection point of view. This linear formulation can be kernelized for nonlinear classification.

SVMs are extended to multiclass classification with one class against one class or one class against all classes training [22,23]. In *multiple instance learning*, the instances are replaced with bags of instances with asymmetrical classification rules, for which a number of SVM-based methods are developed [24–26]. These bags can also be classified symmetrically by a selective SVM method to obtain a large margin [27]. SVM concepts are also used for regression, in which case the class labels are replaced with real-valued predictions [28–30].

The applications of SVMs in real-life problems are overwhelmingly high in number

and wide in range. Some of these applications include handwritten digit recognition [31], object recognition [32], speaker identification [33], face detection in images [34], text categorization [35], pattern recognition [36], biomedical applications [37–40], and financial applications [41,42].

Decision Tree

Decision tree is a hierarchical tree structure that is used to classify data classes based on attributes of data instances. In a decision tree, nodes represent classification attributes and branches represent conjunctions of attributes that lead to those classifications. Given a set of training data, a decision tree can be induced in the form of a sequence of rules. Once a decision tree is formed, it can be easily used to classify unseen data instances based on their attribute values starting at the root node.

The key to build a decision tree is to choose attributes in order to branch. The objective is thus to reduce impurity or uncertainty in data as much as possible. A subset of data is pure if all instances belong to the same class. In general, decision tree algorithms are recursive. There are several popular algorithms that have been developed and performed well in decision tree regression, such as ID3, C4.5, and CART (classification and regression tree). The most well-known algorithm to generate decision trees is known as C4.5 [43]. It builds decision trees from a set of training data by using the concept of Shannon entropy [44], a measure of uncertainty associated with a random variable. On the basis of the fact that each attribute of data can be used to split the data set into smaller subsets, C4.5 examines the relative entropy for each attribute. The attribute with the highest normalized information gain is used to make a decision. Ruggieri [45] provided an efficient version of C4.5, called *EC4.5*, which was claimed to be able to achieve a performance gain up to five times to compute the same decision trees as C4.5. Yildiz and Dikmen [46] presented three parallel C4.5 algorithms which were designed to be applicable to large data sets. Baik and Bala [47] presented a distributed version of a decision tree by using an agent-based approach. The

designed agent generated partial trees and communicated the temporary results among them in a collaborative way. The distributed tree algorithm achieved a very good performance for several data sets collected from distributed hosts.

Decision trees provide an effective way for discovering meaningful patterns and classification rules for a data set. However, it might be impractical or impossible to construct a tree using all of the data instances for a very large data set. Even if an exhaustive tree may be possibly built, it might not be the best way to analyze the data. In this perspective, optimization techniques could play a valuable role in building accurate and concise decision trees from large data sets. One good example comes from the research of Quinlan [43] who proposed a method of selecting the best decision tree using the Minimum Description Length (MDL) principle [48]. A decision tree was first built using the standard DM algorithm C4.5, and then the optimal subtree structure was selected by the MDL, which selects a solution that minimizes the total number of bits needed to encode the tree and the description of the data given by the tree. The MDL-based optimization formulation offers a way to prune the designed tree structure and also avoids the problem of overfitting. Kennedy *et al.* [49] provided an example of applying GA to optimize a decision tree with respect to its classification accuracy. In particular, they encoded a decision tree as a linear array (chromosome), and prediction accuracy was the fitness function of GA. A standard crossover and mutation operations were performed to the chromosomes. The classification accuracy was monotonically increased after each generation of decision trees. Fu *et al.* [50] also applied GA to optimize a decision tree. The proposed GA-based method was shown to be robust to initial training samples, and outperformed the regular approach of building a decision tree using all the available data sets.

K-Nearest Neighbor (KNN)

The key concept of KNN algorithms is based on the assumption that data instances that are near to each other should have similar predicted values. In general, KNN methods

classify a data instance based on the majority category of its K closest neighbors. This means that an instance is assigned to the most common class amongst its K nearest neighbors. Any ties can be broken at random. If $K = 1$, an instance is just simply assigned to the class of its nearest neighbor.

The performance of KNN is highly affected by the choice of K . If K is too small, the prediction result might be very sensitive to noise points; if K is too large, the model tends to be inaccurate since too many points from other classes might be included. Another important issue of KNN is how to combine the class labels of nearest neighbors into prediction. As mentioned above, the simplest and basic method is to take a majority vote, however, problems may arise when the nearest neighbors are distributed widely or some classes dominate the prediction with more frequent samples. On the basis of the assumption that the closer neighbors are better candidates to indicate the class of a new instance, an enhanced approach is developed by assigning a weight to each neighbor according to its distance to the new instance. In particular, the weight of a neighbor is often defined as the reciprocal of its distance. The new instance is then classified to the class with the highest sum of weights. In addition, the distance measure is also essential to KNN approaches. Using a measure that is appropriate for the data at hand is important. A desirable measure should be the one that has a smaller distance value for the two instances with a greater possibility of having the same class. There are numerous similarity measures, which can be used to compute the distance between two instances, such as Euclidean distance, chi-square distance, and dynamic time warping.

Compared to other classifiers, KNN does not require a sophisticated modeling process and thus, is very easy to understand and implement. However, the cost of this is that KNN approaches are usually computational expensive since they have to compute distances to all training examples to classify a new instance.

Moreover, the classification accuracy of KNN can be severely degraded by the presence of irrelevant features. A good selection

of features could reduce the dimensionality of data, scale down the computation time, and, most importantly, improve the classification accuracy. Therefore, a key issue in much KNN research effort has been put into selecting appropriate sets of features. A good example of this field can be found in Chaovalitwongse *et al.* [51] who use a binary integer programming formulation to solve the feature selection problem of massive electroencephalogram data. This formulation actually represents a more general optimization framework of feature selection, which can be expressed as follows:

$$\begin{aligned} \min \quad & f(X) \\ \text{s.t.} \quad & L \leq \sum x_i \leq U, \\ & x_i = 1 \text{ if feature } i \text{ is selected,} \\ & 0 \text{ otherwise,} \\ & x_i \in X \text{ is the feature vector,} \end{aligned} \tag{7}$$

where U and L are the upper and lower bounds of the number of desired features, respectively.

Siedlecki and Sklansky [52] applied GA to performed feature selection for KNN classifiers. Each feature was encoded in a binary vector as a chromosome. If a bit is 1 then the corresponding feature was used in the KNN classifier; otherwise, the corresponding feature was unused. Each of the strings was penalized by the classification accuracy and the number of 1 bits (larger penalty for more 1s). The results indicated that GA-based optimization approach was effective to improve the performance of their KNN classifier for data sets with a large number of features (more than 20 in their studies). Punch *et al.* [53] expanded this 0–1 weighting approach into a 0–10 weighting structure for a better discrimination between the features. Their experiment results indicated that the enhanced KNN classifier worked better than the previous 0–1 approach.

Bayesian Learning

Bayesian decision theory is one of the most widely used statistical approaches to solve

problems in DM. The basic idea of Bayesian learning is built on the estimation of probability distributions [54], which can be described by the well-known naive Bayes probability model as follows:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \quad (8)$$

where $P(H)$ is called *prior probability* of hypothesis H , which is the probability that H is correct before data D was seen; $P(D|H)$ is called *likelihood probability*, which is the probability of observation data D given that the hypothesis H is hold. $P(D)$ is called *prior probability*, which is the probability of witnessing D under all possible hypotheses. The ratio $P(D|H)/P(D)$ is often referred to as *irrelevance index*. If the irrelevance index is 1, any knowledge about H is not relevant to D . In other words, the occurrence of D is independent with the hypothesis H . Any value of irrelevance index below 1 measures the relevancy between D and hypothesis H . $P(H|D)$ is the posterior probability. It is the probability that the hypothesis H is true given that the data D occurs. On the basis of the naive Bayes probability model, a naive Bayesian classifier can be built by introducing some decision rules. The most commonly used one is called *maximum a posteriori decision rule*, which simply picks the hypothesis that is most probable.

The naive Bayesian classifier starts with a prior belief of an event, and then the belief is revised after observing the event. It has become popular because this process models the human belief revision system quite closely [55]. More importantly, the naive Bayesian classifier can be used to construct various powerful Bayesian networks, which have been widely used to solve many complex real-world problems [34,56,57]. For example, a Bayesian network could be built to represent the probabilistic relationships between diseases and symptoms. Given a symptom, the Bayesian network can be used to determine which disease it is with the highest probability [58].

Formally, a Bayesian network constructs a set of random variables and their joint probability distributions as a directed acyclic graph. Each node corresponds to a variable,

and each arc represents the conditional dependency between two variables. For some simplest cases, a Bayesian network can be specified manually. However, this becomes impossible for human beings when the network is too complex, which is often the case in real-world problems. In practice, the construction of Bayesian networks is often formulated as an optimization problem, where the computational task is to find a structure that maximizes a statistically motivated score function. The score is defined to describe the fitness of each possible structure to the observed data. Several existing approaches in use today solve the structure optimization problem by various heuristic search techniques, such as greedy hill-climbing and simulated annealing. One excellent example is from Heckerman *et al.* [59], who applied a hill-climbing search method to find the best network structures. The hill-climbing method starts with an initial network, which can be roughly constructed from expert knowledge. For each search step, it examines all possible variations of the current network obtainable by adding, deleting, or reversing arcs. The best variation with the highest score becomes the new current network. The process repeats until no structural variation improves the current score.

One problem of such local search techniques is it is computationally expensive, especially for a data set with large number of instances or attributes. Friedman *et al.* [60] proposed a faster searching algorithm by restricting the search space only to a small number of candidates for each variable. At each step, the learned network selects the best candidates for the next iteration within the restricted space. The method was evaluated by a number of large real-life data sets, and their results demonstrated that the method was significantly faster than many standard heuristic search procedures without loss of quality in the learned structures.

UNSUPERVISED LEARNING

Unsupervised learning is inspired by brain's ability to extract statistical patterns and recognize complex visual scenes, sounds, and

odors from sensory data. It takes root in neuroscience/psychology and is founded on information theory and statistics. Unsupervised learning is used to reveal and capture unknown, but useful data groupings in a given data set autonomously. The goal of unsupervised learning is to build meaningful representations of data sets for decision making, prediction, and efficient communication. Unsupervised learning methods have wide applications in biology, medicine, market research, and robotics, among others [61]. The two most popular unsupervised learning approaches are discussed in the following sections.

Clustering

Clustering deals with data that have not been preclassified in any way, and does not need any type of supervision during the learning process. It is an unsupervised paradigm which automatically assigns the input data into meaningful clusters based on their similarity. Therefore, the similarity measures between two clusters are essential to most unsupervised learning algorithms. There are many approaches to define similarity between data instances, such as Euclidean distance, Manhattan distance, Hamming distance, Mahalanobis distance, and angular separation. [62]. Owing to the variety of similarity measures available, one must carefully choose the measures, since they might highly influence the shape of clusters. For example, some instances may be close to one cluster according to one measure and further away according to another.

The performance of clustering algorithms can be evaluated by relationships, such as intraconnectivity and interconnectivity. Intraconnectivity is a measure of the density of connections between the data instances within a single cluster. A higher intraconnectivity indicates a better clustering arrangement in a sense that the data instances in the same cluster are highly dependent on each other. Interconnectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is desirable since it indicates that the individual clusters are largely independent of each other.

The goal of clustering is to group a given data set based on a certain measure of similarity. Optimization techniques play an important role in various clustering methods, since the problem is fundamentally to find the optimal groupings or relations for a given data set. Rao [63] was one of the pioneers to apply linear and nonlinear programming formulations to solve clustering problems in an efficient way. A comprehensive review of clustering and optimization formulations can be found in Hansen and Jaumard [64]. Among various clustering algorithms, the most well-known one is called *k-means clustering*. The objective of *k-means clustering* is to find k cluster centers that minimize a squared-error criterion function [61]. Cluster centers are represented by the gravity centers of data instances: that is, the coordinates of a cluster center are the arithmetic mean for each dimension separately over all the data instances in the cluster. The *k-means clustering* assigns each instance to a cluster whose center is nearest to it. Since *k-means clustering* generates partitions such that each pattern belongs to one and only one cluster, the obtained clusters are disjoint. In Dunn [65], a popular clustering algorithm called *fuzzy c-means* (FCM) was developed to allow one data instance to belong to two or more clusters. Each data instance is associated with every cluster by a membership function, by which it has a degree of likelihood to every cluster, rather than just being assigned completely to one cluster. For example, the data instances on the edge of a cluster may be in the cluster to a lesser degree than the instances around the center of the cluster. FCM finds the most characteristic data instance in a cluster to be the “center” of the cluster, and then assigns the grade of membership to each data instance in the cluster.

Sherali and Desai [66] formulated the clustering problem as a nonlinear, discrete optimization problem, and developed a global optimization approach to derive a global optimal partition. They also developed a global optimization algorithm for solving fuzzy clustering problems. The proposed global optimization-based fuzzy clustering

algorithm was validated by several standard data sets from the literature and the synthetically generated data sets for large problems. Their results demonstrated that the proposed algorithm exhibited better performance over the popular FCM technique. Bradley *et al.* [67] provided another optimization structure which uses a concave minimization model to find optimal clusters. Given m points in a n -dimensional Euclidean space R^n , a fixed number of cluster k , the centers of the cluster c are determined such that the sum of the distances of each point to a nearest cluster center is minimized. The general nonconvex optimization formulation can be expressed as follows:

$$\begin{aligned} \min_{C,D} \quad & \sum_{i=1}^m \min_l e^T d_{il} \\ \text{s.t.} \quad & -d_{il} \leq x_i - c_l \leq d_{il} \\ & i = 1, \dots, m \\ & l = 1, \dots, k. \end{aligned} \quad (9)$$

where x_i is the i th point out of m , c_l is the center of l -th cluster out of k , $d_{il} \in R^n$ is the dummy variable used to bound the components of the difference between point x_i and the center c_l , e is the unity vector. This general nonconvex problem was further reformulated into an optimization structure which minimizes a bilinear function using a k -median algorithm.

There are also several other important forms of clustering algorithms that can be distinguished from the literature.

- *Biclustering.* Biclustering is a clustering technique which allows simultaneous clustering of the set of samples and the set of sample attributes (or features). The basic idea of biclustering is that each cluster may be induced by a certain subset of sample features. Therefore, we may find an associated cluster of features for each cluster. The pair of a data cluster and its associated cluster of features is called a *bicluster*. Biclustering aims at partitioning a data set into biclusters. A variety of optimization methods have been used in biclustering algorithms to find the optimal partition. A comprehensive review

of biclustering can be found in Busygin *et al.* [68], which discusses the most widely used and successful biclustering techniques with an emphasis on their mathematical concepts and formulations. It is noted that a great variety of biclustering algorithms have been applied in bioinformatics [69] and text mining [70].

- *Network Clustering.* In many cases, data sets can be conveniently represented by networks or graphs. For these problems, network clustering can be used to construct network clusters for such data sets. A recent survey on network clustering methods can be found in Balasundaram and Butenko [71]. Usually, each data instance is represented by a vertex. Two data instances are connected by a link if they are similar or related based on certain similarity measure. The similarity criterion of a network model considers all data instances in a cluster as a whole and not just the pairwise relations between data instances. Network clustering has become an effective approach in clustering biological networks, such as protein interaction networks [72] and gene expression networks [73].
- *Hierarchical Clustering.* The clustering algorithms mentioned above, all partition data instances directly in a single step. On the other hand, hierarchical clustering attempts to find successive clusters using previously established clusters. Hierarchical clustering takes a series of partitions, which may separate the previous clusters successively into finer clusters, or combine a series of current clusters into larger clusters. Some well developed hierarchical clustering algorithms are Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [74], Clustering Using Representatives (CURE) [75], a hierarchical clustering based on dynamic modeling called CHAMELEON [76], and an incremental hierarchical clustering algorithm called GRIN [77]. The major advantage of hierarchical clustering is that it does

not require the number of clusters to be known in advance. However, these methods suffer from their inability to perform adjustments once splitting or merging decisions are made.

- *Distributed Clustering.* Recently distributed clustering algorithms have attracted considerable attention to extract knowledge from large databases [78,79]. In many cases nowadays, the data are originally collected at different sites, and then brought together to extract information. These data sets are usually huge in size. Instead of being transmitted to a central site where we can analyze data by the standard clustering algorithms, data can be clustered independently on different local sites at first. In a subsequent step, the central site tries to establish a global clustering based on the local clustering results. This approach is computationally efficient, since local clusterings can be operated in a parallel way.

Association Rule Induction

Association rule induction is another popular unsupervised DM technique to discover interesting patterns extracted from a given data set. For each pattern, rule induction will statistically determine its accuracy and significance so that an end user knows how strong the pattern is and how likely it is to occur again. A typical association rule has a form as follows

$$X_1, X_2, \dots, X_n \longrightarrow Y_1, Y_2, \dots, Y_m, \quad (10)$$

where both of X , Y are combinations of binary attributes. To discover useful association rules from data efficiently, many optimization-based methods have been proposed in DM research community. “Apriori,” proposed by Agrawal and Srikant in [80], is the very popular algorithm to mine association rules. Apriori uses a breadth-first search strategy to count the support of itemsets and applies a candidate generation function to exploit the downward closure property of support. Since the first introduction of Apriori, various more efficient algorithms of

rule discovery have been developed, many of them share the same idea with Apriori in candidate generation, that is “if an itemset is not frequent, any of its superset is never frequent.” The Apriori algorithm is quite simple and easy to implement, thus Apriori-like algorithms are widely used in solving problems of association rule induction. Another interesting method was proposed by Rymon [81], who presented an optimization framework to discover minimal rules from binary data. The data set was encoded in a Boolean logic expression at first, and then logic minimization was used to simplify the Boolean logic expression to generate optimal rules.

CONCLUSIONS

In this article, we provided an overview of the most popular DM algorithms and paid particular attention on how optimization techniques can be applied to solve DM problems. Optimization techniques have been widely applied in DM, in both supervised and unsupervised methods. Optimization techniques can directly contribute to a DM model, and new DM approaches may be generated when they are combined with various optimization techniques. On the other hand, many DM methods are fundamentally optimization problems, such as SVM, clustering.

Optimization techniques can enhance a DM method by developing better solutions through various well-developed optimization tools. Many DM methods incorporate optimization as a part of the DM problem. A block diagram of a typical DM process and the application of optimization techniques in DM is shown in Fig. 2. In general, optimization techniques can significantly contribute to DM methods at three major stages:

- Optimal feature selection in data preprocessing step, which can significantly reduce the data dimensionality and search space for DM models.
- Construct optimization formulations directly in DM models, develop efficient optimization-based DM algorithms.
- Pick up the best patterns among a large number of candidates generated by DM models.

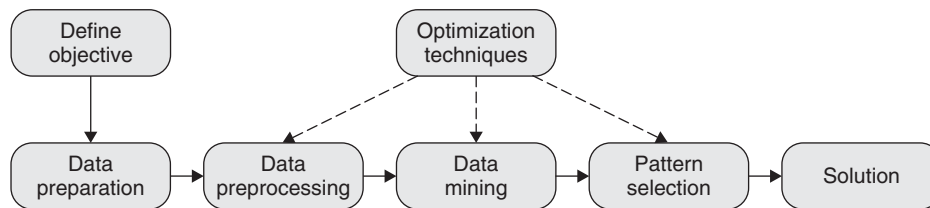


Figure 2. Block diagram of a typical data mining process, as well as the use of optimization techniques in DM at three major steps.

The use of optimization techniques can considerably improve the performance of DM in terms of efficiency, applicability, and accuracy. We provide the most distinct examples of optimization applications in DM and demonstrate how optimization frameworks can be formulated in DM structures. We believe that optimization techniques can play a critical role in providing better and faster solutions to DM problems. In conclusion, it can be stated that DM methods can also contribute to optimization problems in building efficient decision models by providing concise and meaningful representations for a given data set. It is noted that this area is relatively less studied compared to the numerous applications of optimization techniques in DM. The integration of optimization and DM in a closely complementary way constitutes a rich field of research for both OR and DM researchers.

REFERENCES

1. Piatetsky-Shapiro G, Frawley WJ, editors. Knowledge discovery in databases. Cambridge (MA): AAAI/MIT Press; 1991.
2. Basu A. Perspectives on operations research in data and knowledge management. *Eur J Oper Res* 1998;111(1):1–14.
3. Wu C. Applying frequent itemset mining to identify a small itemset that satisfies a large percentage of orders in a warehouse. *Comput Oper Res* 2006;33(11):3161–3170.
4. Delesie L, Croes L. Operations research and knowledge discovery: a data mining method applied to health care management. *Int Trans Oper Res* 2000;7(2):159–170.
5. Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: Formulations and challenges. *INFORMS J Comput* 1999;11(3):217–238.
6. Hagan MT, Demuth HB, Beale M. Neural network design. Boston (MA): PWS Publishing Co.; 1997.
7. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323(9):533–536.
8. Haykin S. Neural networks: a comprehensive foundation. 2nd ed. Upper Saddle River (NJ): Prentice Hall; 1998.
9. Yen GG, Lu H. Hierarchical genetic algorithm based neural network design. In: IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks. Piscataway (NJ): 2000. pp. 168–175.
10. Siddique MNH, Tokhi MO. Training neural networks: Backpropagation vs. genetic algorithms. In: IEEE International Joint Conference on Neural Networks, Volume 4. Washington (DC): 2001. pp. 2673–2678.
11. Glover E. Tabu search—Part I *ORSA J Comput* 1989;1(3):190–206.
12. Battiti R, Tecchiolli G. Training neural nets with the reactive tabu search. *IEEE Trans Neural Netw* 1995;6(5):1185–1200.
13. Vapnik VN. The nature of statistical learning theory. New York (NY): Springer New York, Inc.; 1995.
14. Vapnik VN. Statistical learning theory. New York: Wiley-Interscience; 1998.
15. Cristianini N, Taylor JS. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
16. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond (Adaptive Computation and Machine Learning). Cambridge (MA): MIT Press; 2001.
17. Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. In: Proceedings of IEEE Neural

- Networks for Signal Processing. Amelia Island (FL): 1997. pp. 276–285.
18. Joachims T. Making large-scale SVM learning practical. *Advances in kernel methods - support vector learning*. Cambridge (MA): MIT Press; 1999. pp. 169–184.
 19. Platt JC. Fast training of support vector machines using sequential minimal optimization. Cambridge (MA): MIT Press; 1999. pp. 185–208.
 20. Chang CC, Hsu CW, Lin CJ. The analysis of decomposition methods for support vector machines. *IEEE Trans Neural Netw* 1999;12(4):291–314.
 21. Zhu J, Rosset S, Hastie T, *et al.* 1-norm support vector machines. *Neural information processing systems*. Cambridge (MA): MIT Press; 2003. p. 16.
 22. Weston J, Watkins C. Multi-class support vector machines. Technical Report. Egham: Royal Holloway, University of London. 1998.
 23. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J Mach Learn Res* 2002;2(5):265–292.
 24. Andrews S, Hofmann T, Tsochantaris I. Multiple instance learning with generalized support vector machines. 18th national conference on Artificial intelligence. Menlo Park (CA): AAAI Press; 2002. pp. 943–944.
 25. Mangasarian OL, Wild EW. Multiple instance classification via successive linear programming. *J Optim Theor Appl* 2008;137(3):555–568.
 26. Kundakcioglu OE, Seref O, Pardalos PM. Multiple instance learning via margin maximization. *Appl Numer Math* 2009;60(4):358–369.
 27. Seref O, Kundakcioglu OE, Prokopyev OA, *et al.* Selective support vector machines. *J Comb Optim* 2009;17(1):3–20.
 28. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput* 2004;14(3):199–222.
 29. Gunn SR. Support vector machines for classification and regression. Technical report. Faculty of Engineering, Science and Mathematics School. Southampton: University of Southampton; May 1998.
 30. Drucker H, Burges CJC, Kaufman L, *et al.* Support vector regression machines. *Advances in neural information processing systems*. Cambridge (MA): MIT Press; 1997. pp. 155–161.
 31. Scholkopf B, Burges C, Vapnik V. Extracting support data for a given task. *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. Montreal: AAAI Press; 1995. pp. 252–257.
 32. Blanz V, Scholkopf B, Bulthoff H, *et al.* Comparison of view-based object recognition algorithms using realistic 3D models. Volume 1112, *Proceedings of the International Conference on Artificial Neural Networks*. Berlin: Springer; 1996. pp. 251–256.
 33. Schmidt M. Identifying speaker with support vector networks. In: *Proceedings of Interface*. Sydney: 1996.
 34. Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Puerto Rico: 1997. pp. 130–136.
 35. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: *Proceedings of 10th European Conference on Machine Learning*, Volume 1398. Chemnitz; 1998 Apr. pp. 137–142.
 36. Lee S, Verri A. Pattern recognition with support vector machines. *SVM 2002*. Niagara Falls: Springer; 2002.
 37. Brown M, Grundy W, Lin D, *et al.* Knowledge-base analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 2000;97(1):262–2267.
 38. Lal TN, Schroeder M, Hinterberger T, *et al.* Support vector channel selection in bci. *IEEE Trans Biomed Eng* 2004;51(6):1003–1010.
 39. Noble WS. Support vector machine applications in computational biology. *Computational molecular biology*. Cambridge (MA): MIT Press; 2004. Chapter 3.
 40. Seref O, Cifarelli C, Kundakcioglu OE, *et al.* Detecting categorical discrimination in a visuomotor task using selective support vector machines. In: Arabnia HR, Yang MQ, Yang J. Y., editors. Volume 2, *Proceedings of the 2007 International Conference on Bioinformatics & Computational Biology (BIOCOMP)*. Las Vegas: CSREA Press; 2007. pp. 580–587.
 41. Huang Z, Chen H, Hsu CJ, *et al.* Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decis Support Syst* 2004;37:543–558.
 42. Trafalis TB, Ince H. Support vector machine for regression and applications to financial forecasting. In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference*

- on Neural Networks, Volume 6. Como; 2002. pp. 348–353.
43. Quinlan JR. C4.5: programs for machine learning. San Mateo (CA): Morgan Kaufman; 1993.
 44. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;27: 379–423.
 45. Ruggieri S. Efficient C4.5. *IEEE Trans Knowl Data Eng* 2001;14(2):438–444.
 46. Yildiz OT, Dikmen O. Parallel univariate decision trees. *Pattern Recognit Lett* 2007;28(7):825–832.
 47. Baik S, Bala J. A decision tree algorithm for distributed data mining: towards network intrusion detection. In: *Proceedings of International Conference of Computational Science and Its Applications*, Volume 3046; Fukuoka, Japan. 2004. pp. 206–212.
 48. Rissanen J. Modeling by shortest data description. *Automatica* 1978;14:465–471.
 49. Kennedy H, Chinniah C, Bradbeer P, *et al.* The construction and evaluation of decision trees: a comparison of evolutionary and concept learning methods. *Evol Comput* 1997;1305:147–161.
 50. Fu Z, Golden B, Lele S, *et al.* A genetic algorithm-based approach for building accurate decision trees. *INFORMS J Comput* 2003;15(1):3–22.
 51. Chaovalitwongse W, Fan YJ, Sachdeo RC. Novel optimization models for abnormal brain activity classification. *Oper Res* 2008;56(6):1450–1460.
 52. Siedlecki W, Sklansky J. A note on genetic algorithms for large-scale feature selection. *Pattern Recognit Lett* 1989;10(5):335–347.
 53. Punch WF, Goodman ED, Pei M, *et al.* Further research on feature selection and classification using genetic algorithms. In: *Proceedings of International Conference on Genetic Algorithms*, Volume 93. Urbana-Champaign (IL): 1993. pp. 557–564.
 54. Duda RO, Hart PE, Stork DG. *Pattern classification*. 2nd ed. New York: John Wiley & Sons, Inc.; 2001.
 55. Schocken S, Kleindorfer PR. Artificial intelligence dialects of the Bayesian belief revision language. *IEEE Trans Syst Man Cybern* 1989;19(5):1106–1121.
 56. Myllymaki P, Silander T, Tirri H, and Uronen P. Bayesian data mining on the web with b-course. In: *Proceedings of the 1st IEEE International Conference on Data Mining*. page 626–629, 2001.
 57. Ting J, D'Souza A, Vijayakumar S, Schaal S. A bayesian approach to empirical local linearization for robotics. In: *International Conference on Robotics and Automation*. May 2008.
 58. Chen G, Yu H. Bayesian network and its application in maize diseases diagnosis. *Int Fed Inform Process* 2008;259:917–924.
 59. Heckerman D, Geiger D, Chickering DM. Learning bayesian networks: The combination of knowledge and statistical data. *Mach Learn* 1995;20(3):197–243.
 60. Friedman N, Nachman I, Pe'er D. Learning bayesian network structure from massive datasets: The 'sparse candidate' algorithm. In: *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden. pp. 206–215, 1999.
 61. Duda RO, Hart PE, Stork DG. *Unsupervised Learning and Clustering*. 2nd edition. New York: Wiley; 2001.
 62. Deza E, Deza MM. *Dictionary of Distances*. Elsevier; 2006.
 63. Rao M. Cluster analysis and mathematical programming. *J Am Stat Assoc* 1971;66: 622–626.
 64. Hansen P, Jaumard B. Cluster analysis and mathematical programming. *Math Program* 1997;71(1–3):191–215.
 65. Dunn JC. A fuzzy relative of the iso-data process and its use in detecting compact well-separated clusters. *J Cybernetics* 1973;3(3):32–57.
 66. Sherali HD, Desai J. An rlt-based global optimization algorithm for solving the hard clustering problem. *J Global Optim* 2005;32(2):281–306.
 67. Bradley PS, Mangasarian OL, Street WN. Clustering via concave minimization. *Adv Neural Inform Process Syst* 1997;9:368–374.
 68. Busygin S, Prokopyev OA, Pardalos PM. Biclustering in data mining. *Comput Oper Res* 2008;35(9):2964–2987.
 69. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinformatics* 2004;1(1):24–45.
 70. Sebastiani F. Machine learning in automated text categorization. *ACM Comput Surv* 2002;34(1):1–47.
 71. Balasundaram B, Butenko S. Network clustering. In: Junker BH, Schreiber F, editors. *Analysis of Biological Networks*, Wiley Series on Bioinformatics, Computational Techniques and Engineering. Hoboken (NJ): Wiley; 2008.

72. Gagneur J, Krause R, Bouwmeester T, Casari G. Modular decomposition of protein-protein interaction networks. *Genome Biol* 2004;5(8):R57–1–R57–12.
73. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. *IEEE Trans Knowl Data Eng* 2004;16(11):1370–1386.
74. Zhang T, Ramakrishnan R, Linvy M. Birch: An efficient data clustering method for very large data sets. *Data Min Knowl Discov* 1997;1(2):141–182.
75. Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large data sets. In *Proceedings of ACM SIGMOD International Conference on Management of Data*. Seattle, Washington, USA, Jun. 1998. pp. 73–84.
76. Karypis G, Han EH, Kumar V. Chameleon: A hierarchical clustering algorithm using dynamic modeling. *Computer* 1999;32(8): 68–75.
77. Chen CY, Hwang SC, Oyang YJ. An incremental hierarchical data clustering algorithm based on gravity theory gravity theory. In *Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, Volume 2336, May 2002*. pp. 237–250.
78. Januzaj E, Kriegel HP, Pfeifle M. Towards effective and efficient distributed clustering. In: *Workshop on Clustering Large Data Sets (ICDM)*. 2003. pp. 49–58.
79. Alevizos PD, Tasoulis DK, Vrahatis MN. Parallelizing the unsupervised k-windows clustering algorithm. In: *Parallel processing and applied mathematics*. Berlin, Heidelberg: Springer; 2004. pp. 225–232.
80. Agrawal R, Srikant R. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB conference 1994*. pp. 487–499.
81. Rymon R. On kernel rules and prime implicants. In *AAAI 94: Proceedings of the twelfth national conference on Artificial intelligence*. American Association for Artificial Intelligence, 1994. pp. 181–186.

OPERATIONS RESEARCH IN FORESTRY AND FOREST PRODUCTS INDUSTRY

SOPHIE D' AMOURS
FORAC Research
Consortium, Université
Laval, Québec, Canada

RAFAEL EPSTEIN
ANDRES WEINTRAUB
Department of Industrial
Engineering, University of
Chile, Santiago, Chile

MIKAEL RÖNNQVIST
The Norwegian School of
Economics and Business
Administration, Bergen,
Norway

TOWARD VALUE CHAIN MODELING

Operations research (OR) and management science (MS) have for many years played an important role in supporting forest products industry managers and public officials in their planning decisions. The industry covers many different applications, and the methodology used covers essentially all areas of OR/MS. Also, the industry is very important for several countries including Brazil, Canada, Chile, Finland, New Zealand, Russia, and Sweden. Depending on the country and company, the government, forest owners, and industries face different characteristics and slightly different models and methods are used. However, the basic properties and challenges are similar.

Several surveys exploring the different perspectives on the forest product supply chain can be found in the literature today. Some reviewed the contribution of OR to the forestry industry, focusing on issues related to forest management, harvesting, and transportation to wood-consuming industries [1,2]. Environmental and implementation

issues have also been included in the review [3]. More recently, these surveys were completed by including OR contributions to planning and distribution of forest products such as paper, lumber, engineered wood products, and biofuel [4]. In general, the quality of forest management and forest operations has an important and direct impact on the performance of the different wood fiber value chains. The *Handbook on Operations Research in Natural Resources* [5] presents articles showing multiple topics related to forest modeling, considering different hierarchical levels, spatial and environmental issues, and stochastic and multi-attribute approaches. A special issue of *INFOR* (Vol 3, 2010) also provides a number of applications of forest supply chains.

The literature in the forestry domain can be divided into two main categories. The first category focuses on the forest value chain, particularly forest management, harvesting, and transportation, and the second one focuses on forest products value chain planning for the different products/markets, such as pulp and paper, lumber, engineered wood products, and energy.

The outline of the article is as follows. In the section titled "Value Chains in the Forest Industry," we describe general aspects of value chains in the forest industry. The section titled "Planning Environment" describes the planning environment with strategic, tactical, and operational problems. In the sections titled "Forest Value Chain," "Pulp and Paper Value Chains," "Lumber, Panel, and Engineered Wood Value Chains," and "Energy Value Chains," each of the subchains is described. These are forest, pulp and paper, lumber, and energy chains. In the section titled "Discussion," emerging challenges are discussed.

VALUE CHAINS IN THE FOREST INDUSTRY

In order to transform raw materials from the forest into finished products, many

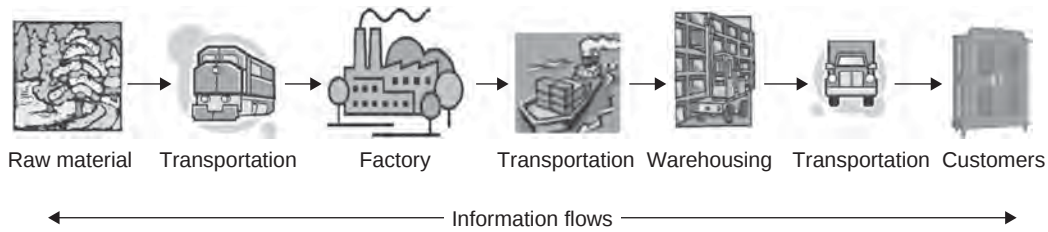


Figure 1. Illustration of the activities from forest to end customer.

operations have to be performed, and consequently, several enterprises and public organizations are involved. The ownership of the forest and the regulations to assess the wood fiber are fundamental characteristics of the different forest value chain structures that can be seen in the different countries around the world. The ownership of the forest will define how integrated the value chain is. When the forest is managed and owned by the public (e.g., under governmental management), the forest value chain is decoupled between the forest and the industry. When the forest is privately owned (e.g., industry or private owners), the forest value chain can be more or less decoupled. In the past years, whatever the structure of the forest value chain is, its activities have often been managed on the basis of local objectives and limited constraints. Today, because of greater knowledge about supply chain management, the involved organizations aim for higher coordination as well as increased value creation.

This complex set of entities that work together via relationships to create economic, environmental, and social values is known as a *value chain* or a *value creation network*. It encompasses all activities associated with the usage of the forest as well as its transformation into goods or services, from the raw materials stage to the end user stage, as well as the associated information flow. These activities are illustrated in Fig. 1.

The forest flows provide raw materials to the forest products industry. These wood fiber flows are mainly full-length trees, logs, or wood biomass. The fiber is transformed in different ways. From a modeling perspective, the transformative processes are either

defined as *one-to-one*, *one-to-many*, *many-to-one*, or *many-to-many*. Moreover, the processes in the forest value chain and the commodity business streams are not strictly defined; many recipes or cutting patterns can be used to provide the needed products. It is said that the processes are alternatives. However, it should be noted early that the value chain is divergent, that is, the number of products or stock keeping units (SKU) increases along the chain. Moreover, the value chain is mainly based on a push mode; that is, the production is supply driven rather than demand driven (i.e., pull based).

The value chains in the forest products industry include all forest owners, companies, and business units that are involved in the management of the forest, procurement, production, or transformation of a given forest product or service and its distribution to the market. They also extend to include recuperation value chains where products are either refurbished or converted into other products, for example, chemicals or energy. This can encompass those companies responsible for forestry operations, transportation, sawmilling, value-added production, pulp and paper, and so on.

The value chains extend in “time” and “space.” For example, the boreal forest may take up to a 100 years to regenerate and is normally widely spread. The value chains are composed of many units that can be associated to tightly connected chains; see Fig. 2. The forest value chain is responsible for managing the forest as well as for delivering the wood to the mills or the services to the population. The forest products value chains transforms the wood into products or services. It then sells and delivers the goods

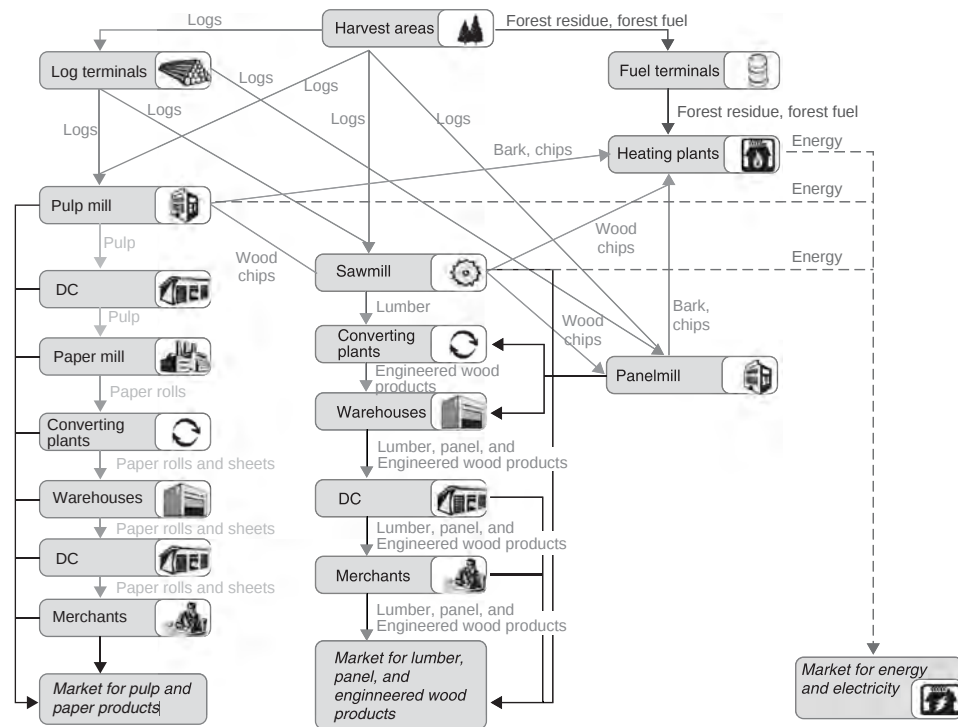


Figure 2. The forest products value subchains.

to the markets. Moreover, it recuperates the products at the end of their useful life and then either refurbishes, converts into new products, or disposes of the products.

The ownership of the chain's units can be shared between public and private organizations (firms). The links between the units vary depending on country legislations and enterprise business models. The main business streams are pulp and paper (e.g., newsprint, fine paper, tissue, and packaging), wood products (e.g., lumber, panels, and engineered wood products for structural or appearance application), and energy (e.g., ethanol, pellets, and biogas). An emerging business stream that is not shown in the figure is known as the *green chemicals*.

The different wood product firms typically own a set of business units that are involved in the transformation and distribution of forest products. When such a firm owns units covering all operations from harvesting activities to distribution activities, it

is said to be vertically integrated. Certain large international companies, in addition to encompassing a wide range of activities, are also active in all of the different markets shown in Fig. 1. For example, Stora Enso, an international corporation with its head office in Finland, has many interrelated supply chains. This is the same for other International companies such as International Paper (USA), Abitibi-Bowater (CAN), UPM (Finland), Arauco (Chile), Fibria (Brazil), and Fletcher Challenge (New Zealand).

As compared to discrete manufacturing-based value chains (e.g., computers, cars, and appliances), forest value chains are complex adaptive systems that are strongly driven by socio-economical, technological, and environmental forces. This is because they rely on a natural resource that contributes to society in many ways. For example, in Canada, more than 300 collectivities are looking upon the forest and the forest products industry to provide jobs and economic development

opportunities as well as to mitigate climate changes through certified forest management (e.g., carbon sequestration) as well as low carbon footprint products (e.g., substitution to fossil-fuel-based products). This is also true for other countries including Sweden, Finland, Chile, New Zealand, and Russia.

The markets are worldwide and particularly dynamic at this time when emerging countries are increasing their consumption of fiber per capita (e.g., paper usage per capita, wood-based product per capita, wood-based energy per capita) and developed countries are moving toward the usage of more Internet-based media and communication. This is illustrated in Fig. 3. For example, in the pulp and paper industry and more specifically the newsprint segment, the North American market is decreasing while the Asian markets are growing. The dynamics of the markets raise challenging questions when it comes to redesign of the value chains through various international investments. Again, these decisions can be taken without an integrated view of the wood fiber supply and location as well as the market location and needs.

PLANNING ENVIRONMENT

Planning the forest value chain is a challenging process because of intrinsic characteristics of the resource, the transformative processes, and the market structures and behaviors. In the next sections, the different planning decisions are reviewed in terms of their strategic, tactical, and operational impacts. Following this, specific descriptions of the decisions taken in the forest value chains are described in more details illustrating the need for well-designed OR tools.

The planning of the forest value chain involves many decisions. These decisions are interrelated, because they involve planning a subset of activities that need to be synchronized with other activities out of the scope of the planning process. For example, harvesting decisions are to be synchronized with machine crew scheduling decisions and transportation decisions, and depending on the role and responsibilities of the value chain entities these decisions may be driven by different individual or groups. The decisions are also linked to time, as long-term decisions such as investment (e.g.,

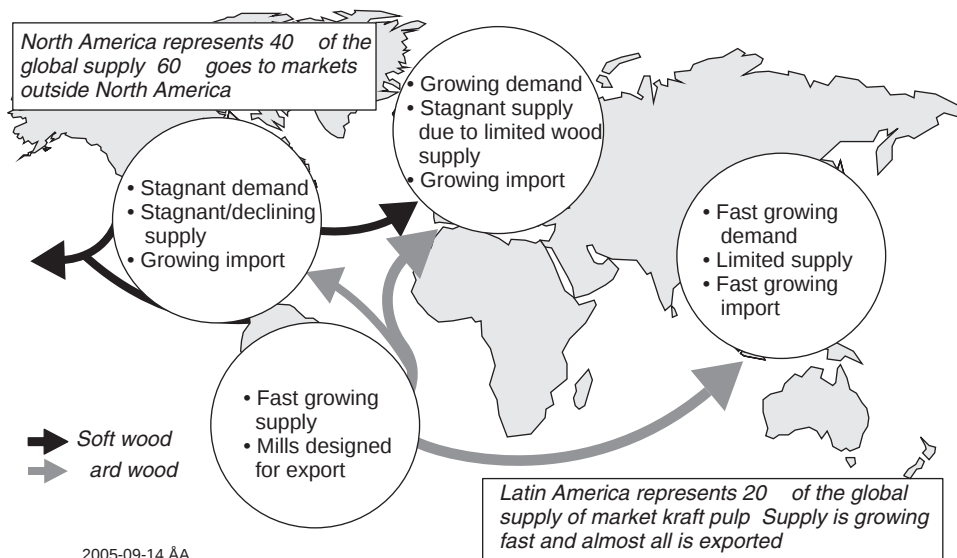


Figure 3. Anticipated development of the global flow of wood fibers. [Source: Axelsson, Södra Cell AB from [6].]

mill investment and forest road investment) constrain the possibilities at the operation level. Finally, space distribution imposes decision synchronization; for example, forest management decisions have a strong impact on industry decisions; therefore, the location of the forest, industry, and market is critical and often has to be taken into consideration to make coherent set of decisions even though these are taken by different business entities.

To better understand the challenges of decision making in the forest value chain, we first review the nature of the decisions to be taken in terms of their strategic, tactical, and operational impacts. To manage the complexity and the links between all the decisions, hierarchical planning is used in the value chains [7].

Strategic Planning

Strategic, or long-term, planning in the forest value chain does indeed involve very long periods of time. For example, the rotation of forest growth can take more than 80 years, and a new pulp or paper mill is normally intended to last more than 30 years. On the other hand, at forest plantation in South America the rotation may be as short as 10 years. Thus, strategic decision making includes making choices related to forest management strategies, silviculture treatments, conservation areas, road construction, the opening/closing of mills, the location/acquisition of new mills, process investments (e.g., machines, transportation equipment, and information technology), product and market development, financial and operational hedging, planning strategies (e.g., make-to-stock, make-to-order, and cut-to-order), and inventory location (e.g., location of decoupling points and warehouses).

The planning approach chosen has a major impact on all investment decisions. For example, the capacity needs and the type of equipment required to support a make-to-stock strategy would be different than those needed to support a make-to-order strategy. Therefore, the planning approach defines important parameters with respect to the necessary technology, capacity, inventory

levels, and maximum distances to customers. Such decisions naturally involve evaluating how the investment will fit into the whole value chain, including deciding which markets to target, how the distribution of the products should be carried out and at what cost, and how the production units should be supplied with the necessary wood fibers (i.e., wood or pulp). Other elements, such as logistics infrastructure as well as energy and water supplies, might also be crucial as it is in the case of pulp and paper industry.

The type of forest land tenure may also affect the way supply chain strategic decisions are made. Wood could come from public lands, private lands, or both, with each type requiring different procurement programs. The types of logs and fibers also differ between regions, and different characteristics are better or less suited for different end products. Other factors may also have to be considered. For example, governmental rules governing the amount of forest land to be set aside for biodiversity purposes, recreational use, and/or carbon sequestration must be taken into account in any decision. The strategic planning of forests needs to consider the industries it supplies downstream, long-term environmental issues, and sustainability as well as its own silvicultural goals [8].

Tactical Planning

Following strategic planning, the next level in the hierarchical planning structure is tactical or mid-term planning. Tactical planning is slightly different depending on whether a forest management problem or a production/distribution planning problem is being addressed. In forest management, hierarchical planning approaches are widely implemented as they permit the strategic planning problem to be initially addressed without taking spatial issues into account. Once this has been done, the problem is then tightly constrained spatially. While strategic forest management planning problems generally span 100 years, tactical planning problems are often reviewed annually over a five-year planning period. Spatial aspects are fundamental in tactical planning. At the forest level, road building accounts for over 30%

of total harvesting costs in native forests. It has been shown that a correct planning approach needs to simultaneously tackle harvesting and road building to provide access [9]. Also, environmental issues that have become significant in the last few decades need to consider spatial issues very carefully. For example, in North America, total harvesting is allowed in small continuous patches, usually not larger than 40 ha.

In planning problems dealing with schedules and resources (e.g., supply/production/distribution issues), tactical planning normally addresses the allocation rules that define the unit or the group of units that is responsible for executing the different supply chain activities, or the type of resources or the group of resources that will be used. It also sets the rules in terms of supply/production/distribution lead times, lot sizing, and inventory policies. Tactical planning allows these two types of rules to be defined through a global analysis of the supply chain. Tactical planning also serves as a bridge between the long-term comprehensive strategic planning and the short-term detailed operational planning that has a direct influence on the actual operations in the chain (e.g., truck routing and production schedules). Tactical planning should also ensure that the subsequent operational planning conforms to the directives established during the strategic planning stage, even though the planning horizon is much shorter. Other typical tactical decisions concern allocating customers to mills and defining the necessary distribution capacity. The advanced planning required for distribution depends on the transportation mode. For example, ship and rail transportation typically need to be planned earlier than truck transportation.

Another important reason for tactical planning in the forest value chain is tied to seasonality, which increases the need for advanced planning. Seasonality has a great influence on the procurement stage (i.e., the outbound flow of wood fiber from the forests). One reason for this seasonality is the shifting weather conditions throughout the year, which can make it impossible to transport logs/chips during certain periods because of

a lack of carrying capacity of forest roads caused by the spring thaw, for example. In many areas of the world, seasonality also plays a role in harvesting operations. In the Nordic countries, for example, a relatively small proportion of the annual harvesting is done during the summer period (July–August). During this period, operations are instead focused on silvicultural management, including regeneration and cleaning activities. A large proportion of the wood is harvested during the winter when the ground is frozen, thus reducing the risk of damage while forwarding (or moving) the logs out of the forest. Seasonality can also affect the production stage (e.g., in Nordic countries, hardwood drying times can vary over the season) or the demand process (e.g., again in Nordic countries, most construction projects are not conducted during the winter period). In Chile, the harvesting season is centered on summer, the rainy winter season being more complicated as less costly dirt roads get muddy.

Budget projection is another area in which tactical planning is useful. Most companies execute an important planning task when projecting the annual budget for the following year, deciding which products to offer to customers and in what quantities. In the process of elaborating these decisions, companies need to evaluate the implications of their decisions on their whole value chain (procurement, production, and distribution) with the aim of maximizing net profits.

Operational Planning

The third level of planning is operational or short-term planning, which is the planning that precedes and determines real-world operations. The operational planning is really the planning process that fixes how and when the value will be captured. For this reason, this planning process must adequately reflect the detailed reality in which the operations take place. The precise timing of operations is crucial. It is generally not enough to know the week or month that a certain action should take place; the time period must be defined in terms of days or hours. Operational planning is usually distributed among the different facilities, or units in the facilities, because

of the enormous quantity of data that has to be manipulated at the operational level [e.g., number of stock keeping unit (SKU) and other specific resources]. In the forest value chain, many operational decisions need to be synchronized (e.g., harvesting, transportation, production, and distribution decisions). Different approaches are used to permit this; among all, information sharing or price-based mechanisms are the most commonly used.

Within the production process, one type of operational planning problem deals with cutting and must be solved by many of the wood product mills (e.g., lumber, dimension parts, and pulp and paper mills). Scheduling the different products moving through the manufacturing lines is also a typical operational planning problem, as is process control involving real-time operational planning decisions. Process control is particularly critical in the pulp and paper industry as the characteristics of the output products depend greatly on the precision of the chemical-fiber mix. Another type of operational planning problem deals with the problems related to transportation, specifically the routing and dispatching done at several points in the supply chain. For instance, it is necessary to route the trucks used for hauling wood from the forest to the mills or for shipping finished products from mills to customers or distribution centers.

Table 1 presents an illustrative summary of the strategic, tactical, and operational planning decisions needed in the pulp and paper industry. This supply chain planning matrix was proposed by Carlsson *et al.* [6] on the basis of a series of case studies conducted with the Swedish pulp producer Södra Cell.

FOREST VALUE CHAIN

The forest supply chain covers harvesting, storage, and transportation to terminals and mills (Fig. 4). It also includes the first conversion processes, for example, bucking and forwarding. Different modes of transportation (e.g., trucks, trains, and ships) are used to transport the various products from the forest units to the industrial mills.

The forest value chain must provide trees suitable for different uses. This involves

dealing with a variety of issues, ranging from strategic forest management (e.g., land management and silviculture treatments) to operational tasks related to harvesting and transportation. Forest planning means dealing with a very long-term planning horizon, anticipating natural disruptions like fires, considering multiple societal needs, and meeting industrial demands. Depending on the nature of the forest (e.g., species, age, soil, and plantation management) and the type of land tenure (e.g., public or private), the planning problems may differ from country to country and from region to region. For example, in Canada the forest is mainly Crown land and is managed by public officers. In Sweden, the forest is mainly owned by thousands of independent land owners who decide, following public regulation, as to which prescription to apply and when, to their land.

The forest value chain is a strategic component of the three main value chains described in this article (i.e., pulp and paper; lumber, panel, and engineered wood; energy). The various business units in the chain are typically extremely widespread geographically, since they deal with different operations (e.g., planting, thinning, road construction, harvesting, storage, and transportation). In the forest value chain, it is first necessary to solve the problem of allocating wood for different uses. In terms of economics, the problem then becomes a question of distributing the wood to the different chains. The long-term strategic decisions aim to attain societal targets that are defined in order to meet sustainable socio-economic development. Some of the strategies decided upon span several forest rotations, which means that some countries have to establish plans covering more than 100 years. Over time, these plans have a great impact on the quality and volume of the available fiber.

The forest is hierarchically planned from long-term strategic planning to mid-term tactical planning and short-term operational planning. Natural forests are composed of trees of different species, age, quality, and volume. They are managed according to the mix of their attributes. Plantation forests are more uniform in age, quality, and volume and

Table 1. The Pulp and Paper Value Chain Planning Matrix [6]

	Procurement	Production	Distribution	Sales
∞ Strategic	• Wood procurement strategy (private vs public land) • Forest land acquisitions and harvesting contracts • Silvicultural regime and regeneration strategies • Harvesting and transportation technology and capacity investment • Transportation strategy and investment (e.g., roads construction, trucks, wagons, and terminal, vessels)	• Location decisions • Outsourcing decisions • Technology and capacity investments • Product family allocation to facilities • Order penetration point strategy • Investment in information technology and planning systems (e.g., advance planning and scheduling)	• Warehouse location • Allocation of markets/customers to warehouse • Investment in logistics resources (e.g., warehouse, handling technologies, and vessels) • Contracts with logistics providers • Investment in information technology and planning systems (e.g., warehouse execution)	• Selection of markets (e.g., location, segment)—customer segmentation • Product-solution portfolio • Pricing strategy • Services strategy • Contracts • Investment in information technology and planning systems (e.g., on-line tracking system, CRM)
Tactical	• Sourcing plan (log class planning) • Harvesting aggregate planning			

	• Route definition and transshipment yard location and planning • Allocation of harvesting and transportation equipment to cutting blocs • Allocation of product/bloc to mills • Yard layout design • Log yard management policies	• Campaigns duration • Product sequencing in campaign • Lot sizing • Outsourcing planning (e.g., external conversion) • Seasonal inventory target • Parent roll assortment optimization • Temporary mill shutdowns	• Warehouse management policies (e.g., dock management) • Seasonal inventory target at DCs • Routing (vessel, train, and truck) • 3PL contracts	• Aggregate demand planning per segment • Customer contracts • Demand forecasting, safety stocks • Available to promise aggregate need and planning • Available to promise allocation rules (including rationing rules and substitution rules) • Allocation of product and customer to mills and distribution centers
Operational	• Detailed log supply planning • Forest to mills daily carrier selection and routing	• Pulp mills/paper machines/ winders/sheeters daily production plans • Mills to converters/DCs/ customers daily carrier selection and routing • Roll cutting • Process control	• Warehouses/DCs inventory management • DCs to customer daily carrier selection and routing • Vehicle loads	• Available to promise consumption • Rationing • Online ordering • Customer inventory management and replenishment

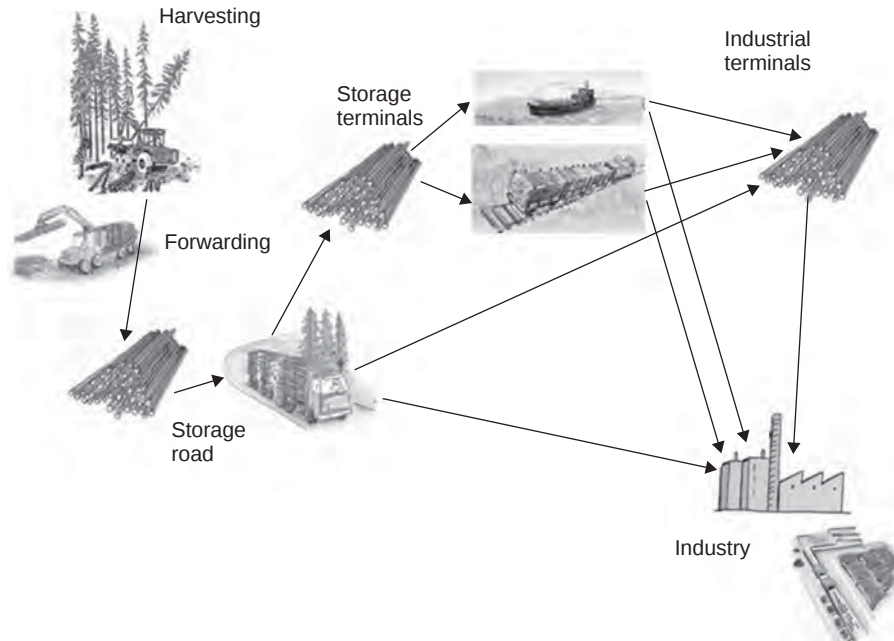


Figure 4. Illustration of harvesting, storage, and transportation activities.

typically provide one species. Silvicultural treatments may or may not involve flows of commercial fiber as opposed to harvesting, which always provides commercial flows. Estimating the basket of products to come out of the forest is very difficult and is often done through stochastic modeling of the forest growth and natural disturbances.

Forest Management

Strategic forest management places the emphasis on the relationship between decisions connected to forest use (e.g., harvesting areas, allocations, and silvicultural treatments) and their different socio-economic consequences (e.g., environmental problems, nondeclining yield, continued employment, forest access, and industrial competitiveness). Numerous models have been developed to aid forest managers and public forestry organizations in their decision making. Some of these models are based on operational research (e.g., harvest planning, road construction, and maintenance planning), while others are based on simulation

(e.g., simulations of growth or ecological impact). Economic models help to connect fiber availability to the value of the forest products. On public land, forest management regimes are established by the government. Moreover, on private land, several laws and regulations are also imposed by the government regarding forest management regimes. Since multiple criteria need to be considered during decision making, most governments and large companies use simulation to evaluate the impact of different forest management strategies.

Depending on the planning decisions to be taken, the forest is viewed from either landscape or regional standpoint. Forest planning includes deciding on silvicultural prescriptions to apply to one or many stands. The prescription includes a sequence of silvicultural treatments such as thinning and progressive or total cuts. These treatments have an impact on the growth of the forest and its sustainability; therefore, the set of possible prescriptions and treatments are normally defined by foresters. For example,

some thinning treatments result in higher quality trees.

During the 1970s and 1980s, linear programming (LP) models, such as the FORPLAN used by the USDA Forest Service, flourished [10]. LP models allow information about growth, biodiversity, spatial requirements, and requirements for protected area to be taken into account. In such models, forest management strategies are treated as constraints, though spatial constraints have not yet been considered. For example, these models normally include a set of nondeclining yield constraints.

With the advent of the geographical information systems (GIS) and the associated spatial data, integrated forest management and harvest planning practices have begun to show increasing concern for spatial relationships and environmental conditions. Particular issues of interest include promoting the richness and diversity of wildlife, creating favorable habitats for flora and fauna, ensuring the quality of soil and water, preserving scenic beauty, and guaranteeing sustainability. Tactical models seek to address these issues, implicitly or explicitly, by structuring the necessary constraining relationships and limiting spatial impact.

One of the primary ways by which spatial relationships and environmental conditions have been modeled at the tactical level is by using adjacency restrictions with green-up requirements. Specifically, a maximum local impact limit is established to restrict local activity for a given period of time. In the case of clear cutting, for example, this corresponds to a maximum open area, which is imposed on any management plan. Another important example for wildlife is the requirement that patches of mature habitats (i.e., contiguous areas of a certain age) must be maintained to allow animals to live and breed. To ensure this, potential areas must be grouped to form patches. A number of models incorporate the maximum open area and adjacency constraints. These models require binary variables to correctly describe the logical decisions and constraints. Both exact and heuristic methods have been proposed to solve these problems. The size ranges from a few

hundred to several hundred thousand management areas [11–13].

Fire is one example of a natural disruption that occurs in forests, thus affecting value chain planning. Fire management processes include both long-term integrated fire and forest management planning and the short-term dispatching of fire crews to stop fire from spreading. Fire management processes can vary from country to country, because of differences in climate, vegetation, and societal needs. The literature on the subject underscores the fact that fire can have a significant impact, both in terms of forest management and value chain planning. For this reason, the subject requires careful examination. Forest fire management can be viewed as getting the right amount of fire to the right place, at the right time, and at the right cost. This definition raises the question of finding the right balance between the beneficial and detrimental impacts of fire on people and forest ecosystems at a reasonable cost to society.

Fires are stochastic processes, which may be caused by humans or nature (e.g., lightning). Significant effort has been dedicated to building good forecasting models in order to anticipate the number of fires that will occur over certain time periods in a given space [14]. Fire prevention and fire detection are two important aspects of fire management. Some forest fire management agencies use fixed towers or lookout points to continuously scrutinize specific forest sectors; whereas, others use fire detection patrol aircraft. Designing such fire detection systems raises interesting OR challenges and questions. Also, once a fire has started, resources with air tankers and firefighters must be coordinated. Often, this must be coordinated for several simultaneous fires. Harvested areas can act as fire breaks that slow down the advance of fires. Hence, integration of fire management with harvest planning may be necessary to include fire aspects early in the planning process. Other examples of natural disruptions that may have a big impact are storms and pests.

Harvesting and Transportation

Harvesting includes the following main phases. The trees are cut and branches are

removed. The trees are then bucked (or cross-cut) into logs of specific dimensions and quality, which can be done in the harvest sectors or in lumber yards. The collection of logs and trees to roadside storage is called *forwarding*, and there are a number of different machines available. Trees, or logs, are then transported directly to mills or to terminals for intermediate storage. The forest residues are also used. They are converted into energy (e.g., wood chips) or chemicals. The harvesting is done by groups of harvest crews and the transportation is done by one or several transport companies.

Harvesting and transportation involve large costs and are often linked together. Figure 5 illustrates the main components of this part of the value chain. Several transportation modes including trucks (many different configurations) are combined with trains and ships. Tactical models are related to the selection and sequencing of stands, or cutting blocks, for harvesting in order to satisfy temporal demands for timber. In addition, road engineering is also a necessary component of tactical planning, since the industry is dependent on an efficient road network to provide access to harvest areas. Thus, the associated costs of road engineering and harvesting impact the viability of management plans. In the forest value chain, storage terminals (e.g., tree/log yards) are used for many reasons, including the benefit from the pulling effect and to protect against demand variation or changing weather conditions [15].

Tactical planning generally involves the use of mixed integer programming (MIP) to model decisions on the time and place of timber harvest, as well as the roads and terminals that should be built or maintained [16]. One important aspect is to get all data required to formulate models. For example, finding correct road investment cost involves many aspects, for example, distance to gravel supply, type of terrain, type of maintenance program, machines required, and so on. Another example is demand forecasting. The demand is a key driver in planning and the results are highly dependent on these values. Because these decision parameters typically vary, the use of robust and stochastic optimization is increasing.

Timber is typically defined by the length and diameter of the logs and by the quality of the wood. The lower part of the tree has a larger diameter, and thus a higher value, and is sent to high-end sawmills. The upper, thinnest part of the tree has a lower value, and is best suited for pulp and paper mills. It is not easy to match the available standing timber exactly to specific product orders. LP models have been particularly useful in this regard, significantly reducing the loss incurred when greater diameter logs are used for lower value purposes, such as pulp. Backhauling, which aims to combine allocations of timber in order to reduce unloaded distances, is often an important part of the models. Backhauling can also be modeled in LP models although the model size increases substantially.

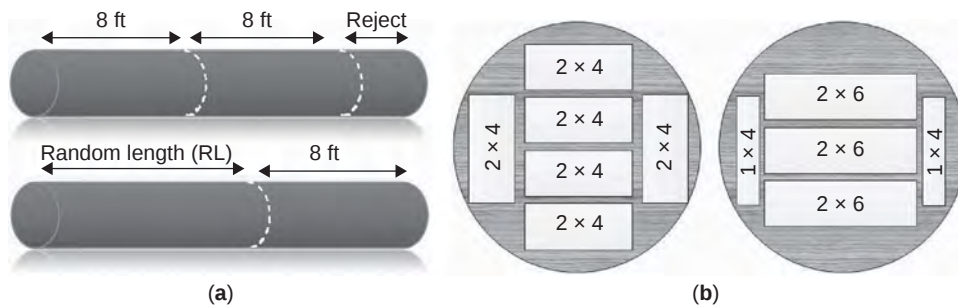


Figure 5. Example of alternative transformative divergent processes. (a) A conceptual representation of tree bucking into logs and (b) sawing patterns transforming logs into lumber or boards.

In many cases, bucking decisions are integrated into the decisions made about the stands that have to be harvested. The number of possible bucking patterns is very high, given the many combinations of lengths and diameters. Bucking can be carried out at sawmills, where each tree is scanned and analyzed individually, or in the forest, by implementing optimizers directly into mechanized harvesters. Dynamic Programming and metaheuristics are the main algorithms proposed for such processes, and commercial codes have been developed and are now used by many forest firms. Freshness, wood quality, and detailed distribution of diameter and length are the aspects that affect log prices and demand.

Operational routing of trucks in forestry has some special properties as compared to standard vehicle routing problems. The basic problem is a pickup and delivery problem with time windows. Trucks typically have different configuration and home bases. In addition, the demand is given for several days; so several daily routes must be found. The available supply is often larger than the demand due to inventory policies, and it is important to balance close supply points with more distant to balance truck capacity. Other aspects are queueing issues and coordinated planning with loaders. Depending on the country and company, some aspects can be simplified and used in the development of solution methods. There are a number of methods developed ranging from simple heuristic and metaheuristics to exact methods. Many models are based on set partitioning type models. GIS are required to provide computations of both distances and time estimates.

PULP AND PAPER VALUE CHAINS

As the logs or chips are transported through the pulp and paper supply chain, they are transformed first into pulp and then into paper, which is then formed into commercial rolls or sheets. In pulp production, the fibers are mixed with chemicals according to a specific recipe to produce the different pulp grades. Recycled paper is often introduced to the pulp, which is then used to produce

jumbo reels of paper of a specific grade, finish, base weight, and color. The jumbo reels are cut into rolls, which can be either sold directly on the market or sheeted for printing and writing paper products. From the mills, the paper is distributed either directly or through a network of wholesalers, distributors, and merchants. Customers vary with the type of paper; for example, printers may buy newsprint; retailers may buy fine paper; and food chains may buy packaging materials. A typical pulp and paper company owns many mills. This pulp and paper supply chain uses all modes of transportation (e.g., trucks, trains, and ships) to transport logs or chips from the forest to the pulp mills; though pulp products may be transported by truck, they mostly travel by train or ship; finished paper is usually moved by train or truck.

It is only recently that the issues of supply chain design in the pulp and paper industry have attracted the attention of practitioners as well as researchers. This can be partially explained by the fact that the industry has typically been driven by a push model in which the main decisions are related to when and where to cut the trees, followed by decisions about processing and selling the resulting products.

The pulp and paper value chain often covers large areas and it is important to design the network. This is a strategic planning problem as mills have to be used for 20–30 years. On a shorter term, it is also important to decide about terminal or distribution center (DC) structure, use of transportation modes (e.g., a combination of vessel, train, barge, and truck), allocation of products to mills, and production capacity. In addition, there is also a need to consider national taxation legislation, transfer price regulations, environmental restrictions, trade tariffs, and exchange rates. The OR models used are often large-scale MIP models. Such model can also be integrated with harvesting or sourcing decisions [17].

Given a distribution network, there may be several customers and one problem is to determine the orders and the level that should be accepted. For example, one order could be profitable only if there is another customer in the same region so that the

same terminal can be used. Pulp and paper mills are often specialized for particular products; so, it is important to combine products—mills and market zones. Also, if vessels are used for long transports, it is important to find good routes and integration between the self-owned fleet with short-term spot vessels. Such models are often MIP models, and they are used to support analysis of what-if scenarios.

The mills can schedule production around campaigns where products are produced in cycles, one at a time, introducing switching costs. These costs can be related to a lower production speed or addition of chemicals needed to satisfy specifications on properties (e.g., brightness). In the pulp and paper industry, switch costs are known to be sequence dependent. This means that the costs of switching between two products are affected by the sequence in which the products are produced, which has in turn led to various scheduling formulations with inventory planning.

Roll cutting is a well-known problem that arises with paper machines. Such problems can often be solved using column generation. However, in an industrial setting, there are many practical issues to consider, such as a limited number of knives in the cutting machine, the width of the cutting machine, the products that must (or must not) be cut in the same pattern, different product due dates, or limited inventory space or replenishment opportunities. Another practical issue is that, given a minimum number of rolls, the objective is to use as few cutting patterns as possible in order to limit setup costs and times. These concerns result in many new constraints, and heuristic approaches to these problems have been suggested [18].

Optimal control is used in many process control applications. When several closed loops have to be controlled simultaneously, nonlinear model can be used to propose the overall control.

LUMBER, PANEL, AND ENGINEERED WOOD VALUE CHAINS

The lumber, panel, and engineered wood industries transform the logs into boards (to

produce lumber and dimension parts) or into flakes (to produce panels). Boards and panels are used as components for engineered wood products, which are mainly used in construction or decoration.

Lumber is produced in stages. First, the logs are sawn into boards in sawmills. In modern sawmills, scanners read the geometry of the logs and optimize the cutting in order to produce maximum value. To prevent production bottlenecks, a mix of logs is used in the sawing lines. The boards are then dried in dry kilns, either by batch or through a continuous process, and are grouped according to specific configurations in order to maximize the quality of the output. Finally, the boards are planed on finishing lines, where setup constraints affect the finishing process. The setup times are mainly due to the process of emptying the buckets containing the finished products.

Panels are produced from wood parts, flakes, or fibers, which are dried, glued, and pressed together. The wood flakes are obtained from logs that have been stored in ponds to soften, while the fiber is obtained by mechanically refining sawdust, shavings, or residues. The wood is then mixed with resin and spread to form a mattress, which is then cut to length. The final board is produced by pressing this mattress under heat conditions. When the panels are used to make engineered wood products (e.g., prefabricated wood I-beams), they are again cut into smaller dimensions to meet specifications.

Engineered wood products are typically produced by assembling lumber and panels. They are used for structural parts in residential and nonresidential constructions (e.g., flooring or roofing systems). Many standards regulate lumber, panel, and engineered wood products. For example, in North America, softwood lumber must conform to the NLGA (National Lumber Grade Authority, www.nlga.org) certification that defines the grades, or quality, for softwood lumber of certain dimensions, while hardwood lumber must conform to the NHLA (National Hardwood Lumber Association, www.nhla.com) grading rules.

Customers exhibit different buying practices, all of which are greatly affected by the variation of the spot market prices. These practices can be grouped into three different categories: spot market, vendor-managed inventory, and contract based. The vendor-managed inventory and contract-based approaches usually offer a financial advantage to the producer in exchange for guaranteed deliveries.

At sawmills, wood is normally sorted into different log types with specific properties. Typical sorting is based on diameter, length, and quality. The sorting can also take place at the harvest areas. However, creating more assortments for the sorted wood is costly and generally a single party cannot independently make the decision to create more assortments. The challenge is to make wood sourcing decisions, including wood sorting strategies as well as technology investments, in order to maximize the profit of the supply chain taking into account fiber sourcing possibilities, production, and market opportunities [19]. LP models can be used to formulate tactical models.

The transformation activities involve generic many-to-many processes; these processes consume a set of input products that are combined in different ways or cut according to different cutting patterns (e.g., bucking or sawing patterns) in order to produce a set of output products. When multiple products are produced through these processes, the products are individually classified as either coproducts or by-products. Coproducts are demand-driven products; by-products are the secondary results generated by the process—usually low-value products (e.g., bark or saw dust)—and are sold on side markets. In many circumstances, several alternative processes (e.g., cutting patterns) are available to achieve the desired transformation; under these circumstances, the planning decisions must also select the processes to be used. Figure 5 illustrates different processes found at sawmills.

Production planning at sawmills includes many decisions. One example is the decision on the type of cutting pattern to be applied to a specific log type; since the process is divergent, it is important to balance production

against demand while taking into account the cost of inventory. A second example is the decision on the type of drying process to be used for each of the green lumber products produced, which is a campaign-scheduling problem. Often there are capacity restrictions on some processes and hence products may be transported between mills in order to better use the capacities. Some models also integrate bucking decisions and production planning decisions at the mills, in order to buck the trees so the mills get the best logs for an optimized efficient production of the ordered lumber products. In such cases, there is a need to use column generation techniques to generate different supply possibilities [20].

To analyze the performance of lumber supply chains, some work is done to build agent-based simulation platforms [21]. This permits the modeling of all business units within the chain and simulates the impact of the combined planning behaviors.

At the operational level, the cutting problem is often critical. Whether dealing with hardwood or softwood lumber, panels or engineered wood products, optimal cutting of incoming products is challenging in terms of material yield management as well as demand satisfaction. The general literature provides many models that deal with cutting problems. In the wood products industry, difficulties stemming from wood defects and wood grading must be considered, raising the need to tackle difficult two-dimensional or even three-dimensional problems. In addition, the high production speed requires fast solvers. Most cutting solvers are based on dynamic programming or heuristics [22].

ENERGY VALUE CHAINS

The energy industry uses forest product residues to produce energy. Forest biomass is directly supplied from the forest or from the mills. This biomass, which serves as biofuel for producing energy, can be transported bulk, bundled, or chipped. The raw materials are often branches, tops, stubs, and low-grade logs that are converted into wood chips. Chipping can be done in the forest, at the intermediate storage terminals

or at the heating plants. Depending on the location, different systems for chipping and transportation are used. Transportation constitutes a large proportion of the overall handling costs, and there is obviously a trade-off between chipping directly in the harvest areas or waiting to chip at the terminals. It is typically cheaper to chip at terminals, but transporting non-chipped forest residue is more expensive than transporting wood chips. The demand for wood chips is correlated to the weather and is as such uncertain. Often LP-based models with transportation, production, and inventory decisions are used [23]. Owing to uncertain demand, robust models are being developed and used [24].

Some chemical processes can be used to transform the residues into specific biofuels (such as ethanol). The energy supplied can be used to satisfy public needs (e.g., heating plants that supply residential and industrial sectors with hot water for heating) or industrial needs (e.g., drying kilns). As the cost of fuel rises, using wood residues for energy becomes an increasingly attractive alternative.

DISCUSSION

Integration and Methods

Although value chain planning has helped to improve the performance of many companies, the challenge of integrating the different planning problems from the forest to the market still remains [25,26]. This is the case when the procurement, production, distribution, and sales activities need to be synchronized throughout a set of independent business units (e.g., entrepreneurs, carriers, sawmills, and pulp and paper mills), their suppliers and their customers. Forest product value chains are generally composed of many interconnected business units that are constrained by their divergent processes. For example, one supply chain can include the producers who manage a mix of species in the forest, the various entrepreneurs who convert these trees into logs or chips, the sawmills that cut the logs into boards or dimension parts, and the pulp and paper

mills that use the wood chips to create reels of paper that are then cut into smaller product rolls or sheets. These varied many-to-many processes make the task of integrating the procurement, production, distribution, and sales activities very complex, given that these activities are always bounded by the trade-offs between yield, logistical costs, and service levels. The consequence of this is that although the number of products in the first stage is low, in the later stages it increases often to several thousands of end products.

In addition to integrating the varied activities, it is also crucial for value chain planning to integrate strategic, tactical, and operational decision making. Because of the size of the problems, due to the number of products, processes, suppliers, customers, and time periods, decomposition techniques and/or hierarchical planning approaches are typically needed. Strategic decisions impose constraints on the tactical planning process, and the ensuing tactical decisions impose constraints on the operational planning process. Many OR models and methods have been proposed to support planning problems in the forest products industry [1–5]. Because the planning horizon ranges from very short (parts of minutes) to very long (several hundred years) time periods, there are very different requirements of the solution time (Fig. 6). For this reason, heuristics, metaheuristics, and easy-to-solve network methods are generally used for operational problems, whereas MIP and stochastic programming methods are better for tactical and strategic planning problems. Many of the OR models are implemented in diverse industrial decision support systems (DSS), which are often integrated into application-specific databases holding all the information needed for the models and the GIS that are used to visualize the input data and results.

Collaboration

There are often many actors in the forest industry, which provides good opportunities for collaboration. This has been explored in the context of transporting logs to mills [27]. Often, many companies operate in different



Figure 6. Illustration of required solution times with respect to planning period. [Source: revised from [1].]

parts of the country, which provides opportunities for optimizing backhauling operations. This opportunity has been addressed in different parts of the world, using the specific wood allocation and trucking constraints found in each region. Models based on game theory have been used to propose a fair allocation of both revenues and costs. Another example is that many companies obtain their wood allocations from unevenly aged forests owned by the government; they often need to agree on a common in-forest harvesting plan. This problem can be addressed by proposing collaborative approaches to help the negotiation process converge on a profitable balanced solution. Studies on the relationship between paper mills and customers have also been carried out. Different approaches to integration can be simulated and optimized, starting with the traditional make-to-order, and then moving toward continuous replenishment, vendor-managed inventory (VMI), and collaborative planning forecasting and replenishment (CPFR) [28].

Future Challenges

The OR challenges in the forest industry are similar to many other areas. Integrating different parts of the supply chains still provides many challenges. The size of the problems becomes large and complex. Methods that are able to solve the problems in short times but with high-quality solutions must be developed. In many applications, there is uncertainty in the data. This is true for measured data, for example, supply information in forest areas as well as forecasts for future demand. Although better methods can increase the quality, stochastic and robust optimization will be more important in the future. To develop a useful DSS for planners, it is important to integrate GIS with the OR models and methods. Also, many planners have no OR background and it is critical to educate these persons regarding the proper use of OR tools.

Acknowledgments

The authors would like to acknowledge the support of the Natural Science and Engineering Council of Canada (NSERC) and the Norwegian School of Economics and Business Administration (NHH), as well as the industrial support of the FORAC Research Consortium, the Forestry Research Institute of Sweden (Skogforsk), and the Millennium Institute Complex Engineering Systems.

REFERENCES

1. Rönnqvist M. Optimization in forestry. *Math Program Ser B* 2003;97:267–284.
2. Martell DL, Gunn EA, Weintraub A. Forest management challenges for operational researchers. *Eur J Oper Res* 1998;104:1–17.
3. Weintraub A, Romero C. Operations research models and the management of agricultural and forestry resources: a review and comparison. *Interfaces* 2006;36(5):446–457.
4. D'Amours S, Rönnqvist M, Weintraub A. Using operational research for supply chain planning in the forest products industry. *INFOR* 2008;46(4):265–282.
5. Weintraub A, Romero C, Bjørndal T, *et al.*, editors. *Handbook on operations research in natural resources*. Volume 99, International series in operations research & management science. New York: Kluwer Academic Publishers; 2007.
6. Carlsson D, D'Amours S, Martel A, *et al.* Supply chain management in the pulp and paper industry. *INFOR* 2009;47(3):167–183.
7. Church R, Murray AT, Barber K. Designing a hierarchical planning model for USDA forest service planning. *Proceedings of the 1994 Symposium on Systems Analysis and Forest Resources*; Pacific Grove, CA. 1994. pp. 401–409.
8. Gunn EA, Rai AK. Modelling and decomposition for planning long-term forest harvesting in an integrated industry structure. *Can J Forest Res* 1987;17:1507–1518.
9. Andalaft N, Andalaft P, Guignard M, *et al.* A problem of forest harvesting and road building solved through model strengthening and Lagrangean relaxation. *Oper Res* 2003;51(4):613–628.
10. Johnson KN, Scheurman HL. Techniques prescribing optimal timber harvest and Investment under different objectives—discussion and synthesis. Volume 18, Forest science monograph. Washington (DC): Society of American Foresters; 1977.
11. Goycoolea M, Murray AT, Barahona F, *et al.* Harvest scheduling subject to maximum area restrictions: exploring exact approaches. *Oper Res* 2005;53(3):490–500.
12. McDill ME, Rabin SA, Braze J. Harvest scheduling with area-based adjacency constraints. *Forest Sci* 2002;48:631–642.
13. Murray A. Spatial restrictions in harvest scheduling. *Forest Sci* 1999;45:45–52.
14. Martell DL. A review of operational research studies in forest fire management. *Can J Forest Res* 1982;12(2):119–140.
15. Epstein R, Morales R, Seron J, *et al.* Use of OR systems in the Chilean forest industries. *Interface* 1999;29(1):7–29.
16. Beaudoin D, LeBel L, Frayret J-M. Tactical supply chain planning in the forest products industry through optimization and scenario-based analysis. *Can J Forest Res* 2007;37(1):128–140.
17. Bredström D, Lundgren JT, Rönnqvist M, *et al.* Supply chain optimization in the pulp mill industry—IP models, column generation and novel constraint branches. *Eur J Oper Res* 2004;156:2–22.
18. Bergman J, Flisberg P, Rönnqvist M. Roll cutting at paper mills. *Proceedings of the Control Systems*; 2002 Jun 3–5, Stockholm, Sweden: 2002. pp. 159–163.
19. Vila D, Martel A, Beauregard R. Designing logistics networks in divergent process industries: a methodology and its application to the lumber industry. *Int J Prod Econ* 2006;102:358–378.
20. Maness TC, Adams DM. The combined optimization of log bucking and sawing strategies. *Wood Fiber Sci* 1993;23:296–314.
21. Frayret J-M, D'Amours S, Rousseau A, *et al.* Agent-based supply chain planning in the forest products industry. *Int J Flex Manuf Syst* 2008;19(4):358–391.
22. Todoroki CL, Rönnqvist M. Combined primary and secondary log breakdown optimisation. *J Oper Res Soc* 1999;50(3):219–229.
23. Gunnarsson H, Lundgren JT, Rönnqvist M. Supply chain modeling of forest fuel. *Eur J Oper Res* 2004;158(1):101–123.
24. Bredström D, Rönnqvist M. A new method for robustness in rolling horizon planning. *SNF Report 25/08*. Bergen, Norway: 2008.

25. Ouhimmou M, D'Amours S, Ait-Kadi D, *et al.* Optimization helps shermag gain competitive edge. *Interface* 2009;39(4):329–345.
26. Feng Y, D'Amours S, Beauregard R. The value of sales and operations planning in the oriented strand board industry with make-to-order manufacturing system: cross functional integration under deterministic demand and spot market recourse. *Int J Prod Econ* 2008;115:189–209.
27. Frisk M, Göthe-Lundgren M, Jörnsten K, *et al.* Cost allocation in collaborative forest transportation. *Eur J Oper Res* 2010;205:448–458.
28. Lehoux N, D'Amours S, Frein Y, *et al.* Collaboration for a two-echelon supply chain in the pulp and paper industry: the use of incentives to increase profit. *J Oper Res Soc* 2010; In Press.

OPERATIONS RESEARCH IN THE VISUAL ARTS

CRAIG S. KAPLAN
University of Waterloo,
Cheriton School of
Computer Science, Ontario,
Canada

ROBERT BOSCH
Department of Mathematics,
Oberlin College, Oberlin,
Ohio

Optimization is a branch of operations research that deals with optimal performance in the face of constraints. The optimizer's goal is to find the best way to perform a difficult task. Usually, the task is difficult because of the constraints that accompany/define it. In this article, we show that visual art, a field that may seem—at least initially—to have no use for optimization whatsoever, is actually a very natural area of application, filled with opportunities for both optimization and optimizers.

The reason why is simple: Artists, like everyone else, face constraints. They must work within budgets. They must meet deadlines. If they enter competitions or juried shows, they must make sure that their pieces satisfy the rules of entry. If they take commissions, they must (or at least should) follow their clients' instructions. And no matter what media they choose to work with, they must deal with the particular constraints—imposed by the laws of physics—that govern how those media work. (Painting with watercolors is different from painting with oils, and painting on rice paper is different from painting on canvas.)

Given that artists are creative, one might think that if it were up to them, they would do away with constraints. After all, constraints *constrain*. They restrict. They limit one's choices. It would seem that constraints inhibit creativity.

But actually, there is much evidence to the contrary. Many artists embrace constraints.

Some need deadlines to be able to finish their work, and some believe that when their choices are limited, they are much more focused and creative:

When forced to work within a strict framework, the imagination is taxed to its utmost and will produce its richest ideas.

T.S. Eliot

Creativity arises out of the tension between spontaneity and limitations, the latter (like the river banks) forcing the spontaneity into the various forms which are essential to the work of art or poem.

Rollo May [1]

A large part of the beauty of a picture arises from the struggle which an artist wages with his limited medium.

Henri Matisse [2]

In fact, many artists go so far as to impose artificial constraints upon themselves. Consider George-Pierre Seurat. When viewing his painting *A Sunday on La Grande Jatte* (1884) from up close, one sees a mass of colorful dots. Upon backing away from it, one's eyes fuse the dots into an image of a group of Parisians relaxing on an island on the Seine. To create this masterpiece, Seurat set himself the task of producing the best possible depiction of what he saw on the riverbank, subject to two highly restrictive, manufactured constraints: he had to keep his colors separate, and he could only apply paint to the canvas with tiny, precise, dot-like brush strokes. Seurat's self-imposed constraints gave rise to a spectacular piece of art, the most widely reproduced example of what art historians call *pointillism*.

When the computer is considered as an artistic medium, the problem of satisfying constraints—formerly a purely cognitive process—becomes a practical opportunity for the application of optimization algorithms. If the constraints can be articulated in a form that the computer can understand, and a domain established within which a satisfactory artwork might be found, then an optimization algorithm can be made to search through the domain for a bestpossible solu-

tion. That solution can then be synthesized on the computer screen or used as a template for constructing an artwork by hand.

In the rest of this article, we explore applications of optimization algorithms to the computer-based synthesis of art. We highlight a few of the most indicative examples (rather than attempting to offer a comprehensive survey) and outline the strategies and algorithms used by those examples. We have chosen to break down our presentation broadly into three topics: tile-based methods, freeform arrangement of elements, and line art.

LAYING TILES

The predominant means of representing photographs or other pictorial information digitally is as a raster image. A raster image (or, more succinctly, an image) can be thought of as a function I that maps pixel locations (x,y) to some color space. Here, x and y are positive integers in the ranges $\{1, 2, \dots, W\}$ and $\{1, 2, \dots, H\}$, respectively, for width W and height H . The range of I might be monochromatic luminance values in the interval $[0, 1]$ (where 0 represents black and 1 represents white), colors represented by RGB triples in $[0, 1]^3$, or values in some other space. These pixels need not be restricted to individual color samples. For example, given an RGB source image I of size $W \times H$, we may wish to group pixels in I into $k \times k$ blocks. The result is a new image I' of size $\frac{W}{k} \times \frac{H}{k}$, in which each “pixel” is a tuple of size $3k^2$ holding all the RGB information from a $k \times k$ block of pixels in I .

Once a photograph is seen as divided into a grid of pixels (or blocks of pixels), it is natural to seek an artistic or playful rendition of that photograph in which each pixel is replaced by some novel object or picture. Every time an artist decides to construct a mosaic out of some new material, a new self-imposed constraint is born, giving rise to a new form of artistic expression. The pixels take on a life and identity of their own. The viewer’s eye is left in a dynamic tension between the image as a whole and the small elements that make it up. These elements

can complement or comment on the image of which they are a part.

Renowned photorealist painter Chuck Close has spent most of his career working in this medium. He typically divides a photographic portrait into a grid, and then approximates the average color of each grid cell in paint (using, for example, concentric rings of multicolored paint). Computer graphics pioneer Ken Knowlton has also created works that emphasize the interplay between an image and the elements used to represent it [3,4]. In many of his works, small objects are arranged in a regular grid, with the size or color of the element chosen to approximate the tone of an underlying image. Fragments of a broken teapot reassemble into an image of a teapot; portraits are formed from hundreds of tiny seashells. In one famous image (which we will soon revisit), a portrait of Martin Gardner is created from six sets of double-nine dominoes. More generally, mosaicists have used dice [3], dominoes [3–8], Lego[®], bricks [9], toy cars [3], spools of thread [4,10], baseball cards [4,11], and food [12].

As computers became capable of handling more (and larger) images, artists and computer scientists began to experiment with using images themselves as elements to compose larger images. In a “photographic mosaic,” blocks of pixels in a target image are approximated using tile images taken from a database [11,13,14]. The goal might simply be to minimize color differences between blocks and tiles, or to find tile images with structures that align with visual features in the blocks. Today, photographic mosaics are commonplace and many software tools exist that allow hobbyists to create their own. Photomosaics typically take advantage of an individual’s photograph collection or of the vast availability of digital photographs on-line; alternatively, frames of video can provide an abundant source of tiles [15].

When working in the realm of mosaics, some artists are compelled to go beyond the inherent constraints of their materials. The domino mosaics of Knowlton, Knuth, and Bosch are not only made out of dominoes but also made out of complete sets of dominoes. (Bosch’s 44-set mosaic of President

Obama contains exactly 44 dominoes of each type—exactly 44 double-blank dominoes, exactly 44 zero-one dominoes, etc.) Ken Knowlton’s portrait of Helen Keller is composed of the 64 characters of the Braille writing system, each appearing 16 times. Related constraints have also been applied in the construction of photographic mosaics. Chris Jordan’s *Denali/Denial* mosaic arranges 24,000 (digitally altered) logos from the GMC Yukon Denali sports utility vehicle (representing six weeks of sales in 2004) into an image of Denali (also known as Mount McKinley). A Robert Silvers photomosaic, commissioned by *Playboy* for its 45th anniversary, renders the centerfold of the magazine’s first issue as a photomosaic of the first $45 \cdot 12 = 540$ of its covers.

When approximating image pixels by objects taken from some inventory, the simplest case is to choose a best possible object independently for each pixel. (We might also imagine that our inventory consists of an unlimited supply of every possible object.) In this case, very little optimization is necessary; we can simply choose the best-fit object independently for every pixel. Consider the simple case of monominoes, black squares printed with zero to nine dots, arranged in the patterns found on double-nine dominoes. Each of those monominoes can be seen as approximating one of ten fixed luminance levels spread evenly between black and white. For each block of square pixels in a target image, we quantize its luminance to one of those ten values and place the corresponding monomino. This algorithm is analogous to ordered dithering techniques in computer graphics [16].

The need for optimization becomes more apparent when blocks must be approximated by tiles taken from a finite inventory. In a Knowlton-style mosaic, we might have a fixed supply of shells or teapot shards, and wish to find the most successful overall arrangement of those pieces that approximate the source image. Or, we might have a database of tile photographs to use in a photographic mosaic, but we wish each photograph to be used at most a small number of times. In such cases, a greedy choice in one part of the image will affect the availability of a good approximation

elsewhere; a successful image will achieve a globally optimal approximation.

Let I be a $W \times H$ image, and suppose we have an inventory of tiles $T = \{T_1, \dots, T_n\}$. (We may think of each tile as a small image of the same size and shape as the pixels in I .) In order to evaluate the quality of the approximation, we are given a distance function $\delta(C, T)$, where C is a color in the same space as the range of I and T is a tile. This function measures how effectively the given tile approximates a pixel in the source image; we assume that $\delta(C, T) \geq 0$ and that smaller values denote better approximations. The goal is to define a mapping $f(x, y)$ that chooses, for each pixel location (x, y) , a particular tile T_j in such a way that no tile is used more than once (if we wish to allow duplicates, we can preload them into the inventory). Of all such mappings, we seek the one that minimizes $\sum_{x,y} \delta(I(x, y), f(x, y))$, the total amount by which the tiles approximate the original image.

Problems of this type can be efficiently solved by constructing a complete, weighted bipartite graph, in which each pixel location (x, y) is connected to every tile T_j by an edge of weight $\delta(I(x, y), T_j)$, and computing the minimum-weight matching that uses all of the pixels. Analogously, they can be seen as instances of the assignment problem.

Our task becomes much more complex when tiles do not match up one to one with image pixels. Consider Knowlton’s domino portrait of Martin Gardner, mentioned above. Each tile is a domino that must serve as an effective approximation of two adjacent pixels. However, any given pairing of numbers occurs only once per set of dominoes, and so a desired pairing might be in short supply. Moreover, dominoes may be oriented horizontally or vertically, and we must contend with the superexponential number of ways to cover pixels in a rectangular image with dominoes.

In this case, a simple technique like bipartite matching is no longer rich enough to encode the space from which a successful solution will be found. However, the construction of domino art can be reduced in a natural way to an integer programming problem [5].

Let D_n represent the set of double- n dominoes. This set can be thought of as consisting of the $\binom{n+2}{2}$ distinct ordered pairs $d = (d_1, d_2)$ for integers $0 \leq d_1 \leq d_2 \leq n$. Now let us suppose that we are given an image I of size $W \times H$, and s sets of double- n dominoes to cover it with (whence $WH = 2s\binom{n+2}{2}$). We define a set P consisting of all pairs of adjacent pixel locations:

$$P = \{(x, y), (x+1, y)\} | 1 \leq x < W, 1 \leq y \leq H\} \cup \\ \{(x, y), (x, y+1)\} | 1 \leq x \leq W, 1 \leq y < H\}.$$

Any pair in P can potentially be covered by any domino. Furthermore, because the domino can be rotated by 180° when covering the pair, its two halves can be assigned to the pair's pixels in two possible ways. Let $p = \{(x_1, y_1), (x_2, y_2)\}$ be a pair, $d = (d_1, d_2)$ be a domino, and δ be a distance function that compares individual pixels to domino halves. We define a new distance function δ' that compares dominoes to pixel pairs, taking into account their two possible orientations:

$$\delta'(p, d) = \min(\delta(I(x_1, y_1), d_1) + \delta(I(x_2, y_2), d_2), \\ \delta(I(x_1, y_1), d_2) + \delta(I(x_2, y_2), d_1)).$$

We now formulate an integer programming problem as follows. We define a set of variables x_{dp} , where d ranges over all dominoes in D_n and p ranges over all pairs in P . The intention is that $x_{dp} = 1$ if and only if domino d is used to cover pair p ; otherwise $x_{dp} = 0$. With these variables in mind, we wish to minimize

$$\sum_{d \in D} \sum_{p \in P} x_{dp} \delta'(p, d),$$

subject to the following constraints:

$$x_{dp} = 0 \text{ or } 1 \text{ for all } d \in D \\ \text{and all } p \in P \quad (1)$$

$$\sum_{p \in P} x_{dp} = s \text{ for all } d \in D \quad (2)$$

$$\sum_{d \in D} \sum_{\substack{p \in P; \\ p \ni (x, y)}} x_{dp} = 1 \text{ for all } 1 \leq x \leq W \\ \text{and all } 1 \leq y \leq H. \quad (3)$$

The first set of equations above forces the variables x_{dp} to decide whether or not to place a domino atop a given pair. The equations in the second set require that each distinct domino be used exactly s times. The third set prevents dominoes from overlapping by asking that every pixel location is covered by precisely one domino.

The tangled interactions between variables introduced by the third set of equations complicate the domino portrait problem, to the point where it can no longer be seen as a simple instance of the assignment problem. Still, the problem instances that arise in the creation of a domino artwork have been found to be tractable in practice [6]. An example of a domino portrait is given in Fig. 1. (Knowlton's original domino portraits relied on a two-phase algorithm in which the image was first partitioned, once and for all, into pairs of adjacent squares, and then dominoes were placed into these domino locations. Knuth later showed that each of these phases could be modeled as an instance of the assignment problem [8].)

Having developed this integer programming framework for tile-based image approximation, it becomes possible to explore variations that depend on other constraints between tiles. For example, Bosch and Pike considered methods for approximating every image pixel with a gray square chosen from a small set of tones [17]. They impose the constraint that neighboring pixels receive contrasting tones, producing images they call *map-colored mosaics*. They achieve this goal with a similar objective function as above, with the addition of equations that prevent adjacent pixels from receiving the same tone. In other work, Bosch uses tiles that must seamlessly match across their boundaries with neighboring tiles [18]. Again, the compatibility of adjacent tiles can be expressed as constraints in an integer programming problem.

FREEFORM ARRANGEMENT OF PRIMITIVES

The previous section explored cases in which elements (colored squares, photographs, dominoes, or other tiles) could be placed in a discrete, finite set of configurations. Optimization has also successfully been used to create freeform artistic compositions, in which elements can be positioned arbitrarily in the image plane and possibly even deformed. Typically, such techniques use a continuous optimization process to find an optimal or near-optimal arrangement, though even the simple geometric optimality of Voronoi diagrams has been used for decades to create abstract compositions [19].

As discussed in the introduction, there is a natural analogy to be drawn with abstraction in many styles of art. Any artistic medium or technique imposes constraints on the artist's range of expression. The artist must then seek to create the best possible reproduction of a scene subject to those constraints. Even in a highly expressive medium such as oil paint and canvas, we might imagine that the artist would at least seek to minimize the amount of paint used.

For example, suppose we are given a color image $I(x,y)$, which we wish to approximate via a painting P . The painting consists of an ordered sequence of strokes, each of which is defined by a curved path in the image plane, a thickness, and a color. Any such painting can easily be rendered to yield a color image, and the resulting image can be compared to I as a measure of the quality of P .

Hertzmann uses this sort of stroke-based model to construct painted approximations of images via a relaxation process [20]. He articulates several ways that we may wish to measure the quality of a painting. Obviously, the similarity of the painting to the original image is one crucial measure; Hertzmann weights this measurement by a spatially varying scaling factor, allowing unimportant parts of the canvas to be painted more roughly. Other measures are the total surface area of all paint strokes, the number of strokes, and the surface area of the uncovered part of the canvas. A user of this algorithm may wish to vary the relative importance of these measurements. Hertzmann's overall objective function is a weighted sum of these individual terms; the user can adjust the

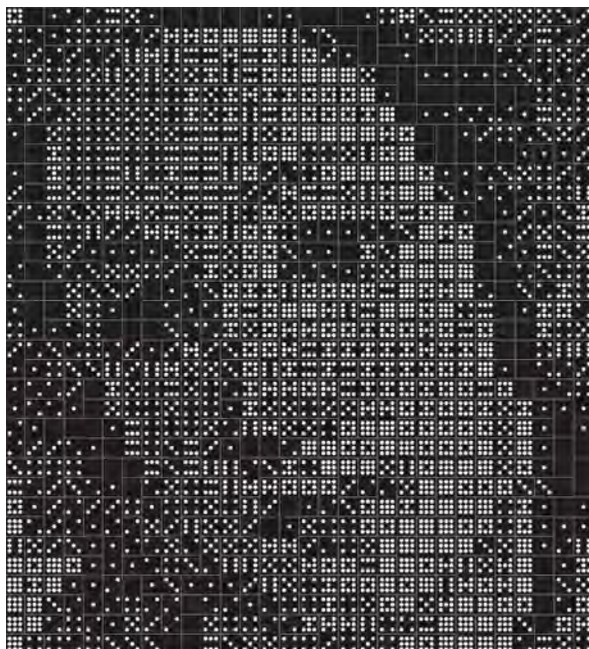


Figure 1. An example of a domino artwork: a portrait of Venus from Botticelli's "The Birth of Venus," constructed from precisely nine complete sets of double-nine dominoes.



Figure 2. Four sample paintings of a photograph of the Queen Mary, created using Hertzmann's paint by relaxation algorithm [20]. By varying the texture and transparency of strokes, together with the weights used in the objective function, different visual effects can be achieved. [Source: Images courtesy of Aaron Hertzmann, based on a photograph by Philip Greenspun (<http://philip.greenspun.com/>).]

painting style by varying the weights. The idea of constructing an objective function from a weighted sum of competing terms, and then putting the weights in the hands of the user, recurs frequently in the application of optimization algorithms to the creation of art in computer graphics.

Hertzmann's algorithm iteratively proposes perturbations to the current state of the painting, and accepts those perturbations if they reduce the value of the objective function. He defines a set of possible perturbations, including the addition, removal, deformation, and recoloring of strokes. Four sample paintings created using Hertzmann's algorithm are shown in Fig. 2.

Related approaches use genetic algorithms (GA) or genetic programming (GP) to construct an approximation of an image [21]. These approaches typically begin with a fixed set of strokes (for example, oriented line segments) and use an evolutionary computation to find optimal positions and orientations for those strokes. Here, the objective function is based entirely on the similarity of the drawing to the original image; Hertzmann's terms for number of strokes and total stroke surface area are implicit in the construction of the problem instance. A related approach, pioneered by Collomosse and Hall, distributes stroke locations based on a measure of perceptual salience [22]. Their algorithm performs a genetic crossover of two paintings

by preserving the strokes from each that contribute most toward the reproduction of the source image.

A hybrid between tile-based methods and freeform element placement can be found in the so-called *packing* algorithms. These techniques introduce additional constraints on spacing between elements; for example, we may wish to prevent elements from overlapping. One simple application is stippling, in which black discs are distributed across the image plane so that their varying radii and spacing approximate the darkness of a source image I . Most stippling algorithms are approximate forms of importance sampling, where samples are placed more densely in darker parts of an image. A practical stippling technique is weighted Voronoi stippling, a weighted version of Lloyd's method (an iterative algorithm for constructing a centroidal Voronoi diagram) that relaxes sample points into a stippled drawing of an image [23].

In many packing algorithms the goal is not to approximate continuous tone or color, but to fill a set of container regions as tightly as possible while controlling gaps and overlaps. For example, In Jigsaw image mosaics, each container region is filled using an iterative algorithm with backtracking that distributes elements inward from the container's outer

boundary [24]. The placement of tiles is evaluated based on a weighted sum of color reproduction, overlaps, gaps, and tile deformation.

In an alternative approach to stylized illustration, artistic thresholding begins with a subdivision of the image plane into disjoint regions derived from segmentation [25]. An optimization process then colors each of the regions black or white, resulting in a stylized monochromatic illustration with no depiction of tone. As with Hertzmann's algorithm, the quality of a given assignment is computed as a weighted sum of several terms. The two most important terms allow the user to trade off between the overall approximation of tones in the source image and the representation of high-contrast edges between adjacent segments.

LINE ART

Recently, some research in artistic optimization has explored the use of continuous line drawing to represent images and designs. The basic idea is simple: approximate the tone of an image via a simple closed path. The path will necessarily take a twisty, meandering route as it distributes its tone over the entire image plane. Contemporary visual artist and designer Mo Morales [26] has created similar artworks by hand.

TSP Art creates continuous line drawings using the traveling salesman problem, one of the cornerstones of optimization research [27]. In a first phase, cities (points) are distributed over the image plane so that their density approximates the tone of a given image. Weighted Voronoi stippling provides an excellent point distribution technique for this purpose, though it is also possible to use an approach such as point repulsion to achieve an even distribution [28]. Once cities are placed, they are joined into a simple closed path using a heuristic TSP solver such as the Lin–Kernighan approach provided by Concorde [29]. An example TSP portrait is shown in Fig. 3. We do not expect this solution to be optimal, but the heuristic will produce a single closed path with no self-intersections. The polygon corresponding to the tour can serve as the final drawing, or it can be further

modified for aesthetic purposes (for instance, its thickness can be modulated to improve tone reproduction or it can be smoothed into a curve). A set of city distributions used in the creation of TSP Art was recently offered as large-scale challenge problems for the TSP [30]. Because of the evenness of the spacing between cities, we can expect these problem instances to feature a broad range of near-optimal solutions, making the true optimum especially difficult to find.

Similar results have also been achieved using a geometric attraction-repulsion model inspired by reaction-diffusion systems [31]. In this approach, a physical simulation allows points in a closed path to exert forces on each other. The force model, based on the Lennard-Jones potential in chemistry, induces a strong repulsive force on points that are too close together and a mild attractive force on distant points. The “rest distance” varies spatially with image tone. Further steps in the simulation add and remove points as needed and filter the curve for smoothness. Beginning with a simple shape such as a circle, this technique will gradually evolve coral-like meandering curves that capture the tone of images.

The tour produced by the Concorde TSP solver will not intersect itself, but beyond that it is difficult to control what kind of tour is created. Recent research has considered ways that the tour can be “steered” to satisfy visual goals [28]. In particular, it would be valuable to be able to specify that certain regions in the image plane were definitely inside or outside of the tour path, or that two regions were on the same side. While TSP solvers like Concorde cannot encode such side constraints, they can easily be added as equations to the Dantzig–Fulkerson–Johnson integer programming formulation of the TSP.

CONCLUSION

This article has offered a glimpse of some of the ways that optimization has been used as a tool in the visual arts. We have encountered some of the more popular styles of art that have been explored via optimization

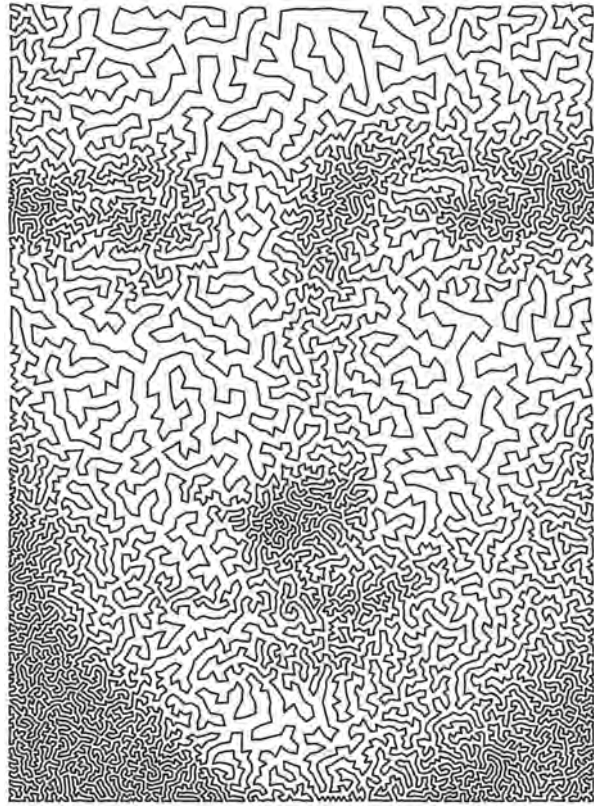


Figure 3. An example of TSP Art: the face of Leonardo's "Mona Lisa", rendered from a single, continuous closed curve.

(tile-based designs, paint strokes, and line art) and some of the optimization techniques that have figured into those explorations (the assignment problem, integer programming, continuous energy minimization, etc.).

When optimization is used to create visual art, it is alluring to envision an accurate model of human aesthetic response as the ultimate objective function. The constraints in which a human artist works could then be seen as delineating a search space, in which we seek an image that maximizes this idealized aesthetic response.

Of course, such an objective function is impossible, or at least well beyond our computational capabilities and our understanding of human perception. There is a great gulf between this ultimate view of artistic optimization and the simple mathematical objective functions we use in practice. In that gulf, we see endless opportunities for future applications of optimization in visual art.

Each new optimization offers another point of view from which to understand the way we see art, and another source of beautiful images.

FURTHER READING

Readers specifically interested in the use of GA and GP in the synthesis of art may wish to consult the book *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music* (Romero JJ and Machado P, editors. Springer, 2007).

REFERENCES

1. May R. The courage to create. New York (NY): W.W. Norton & Company; 1994.
2. Flam J. Matisse on art. Berkeley: University of California Press; 1995.
3. Knowlton KC. Representation of designs. US Patent 4,398,890. 1983 August 16.

4. Knowlton K. Knowlton mosaics. 2010. Available at www.knowltonmosaics.com. Accessed 2010 Feb 1.
5. Bosch RA. Constructing domino portraits. In: Cipra B *et al.*, editors. *Tribute to a mathematician*. Wellesley (MA): A.K. Peters; 2004. pp. 251–256.
6. Bosch RA. Opt art. *Math Horizons* 2006; February;13:6–9.
7. Bosch R. Domino artwork. 2010. Available at www.dominoartwork.com. Accessed 2010 Feb 1.
8. Knuth DE. *The Stanford GraphBase*. Reading (MA): Addison Wesley; 1993.
9. Harshbarger E. Eric Harshbarger's LEGO pages. 2010. Available at www.ericharshbarger.org/lego/. Accessed 2010 Feb 1.
10. Sperber D. Installation art, sculpture, and public commissions. 2010. Available at www.devorahsperber.com. Accessed 2010 Feb 1.
11. Silvers R. *Photomosaic portraits*. New York (NY): Viking Studio; 2000.
12. Mecier J. Jason Mecier. 2010. Available at jasonmecier.com/gallery.html. Accessed 2010 Feb 1.
13. Jordan C. Chris Jordan photography. 2010. Available at www.chrisjordan.com. Accessed 2010 Feb 1.
14. Silvers RS. *Photomosaics: putting pictures in their place* [MS Thesis]. The Media Lab, Massachusetts Institute of Technology; 1996.
15. Salavon J. *The grand unification theory*. 2010. Available at www.salavon.com/GUT/GUT.shtml. Accessed 2010 Feb 1.
16. Ulichney R. *Digital halftoning*. Massachusetts Institute of Technology. Cambridge (MA): The MIT Press; 1987.
17. Bosch R, Pike A. *Map-colored mosaics. Bridges Banff: mathematical connections in art, music, and science*. London (UK): Libra; 2009. pp. 139–146.
18. Bosch R. *Edge-constrained tile mosaics. Bridges donostia: mathematical connections in art, music, and science*. Leicester (UK): IEEE Computer Society, Los Alamitos, CA; 2007. pp. 351–360.
19. Haeberli P. Paint by numbers: abstract image representations. *SIGGRAPH '90: Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM Press; 1990. pp. 207–214.
20. Hertzmann A. Paint by relaxation. *CGI '01: Computer Graphics International 2001*: Hong Kong, China: IEEE Computer Society; 2001. pp. 47–54.
21. Barile P, Ciesielski V, Trist K. Non-photorealistic rendering using genetic programming. In: Li X *et al.*, editors. *Proceedings of SEAL 2008 LNCS 5361*. Springer; Heidelberg 2008. pp. 299–308.
22. Collomosse JP, Hall PM. Painterly rendering using image salience. *Proceedings of the 20th Eurographics UK Conference*. 2002. pp. 122–128.
23. Secord A. Weighted Voronoi stippling. *NPAR '02: Proceedings of the 2nd International Symposium on Non-photorealistic Animation and Rendering*. New York: ACM Press; 2002. pp. 37–43.
24. Kim J, Pellacini F. Jigsaw image mosaics. *SIGGRAPH '02: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*. New York: ACM Press; 2002. pp. 657–664.
25. Xu J, Kaplan C. Artistic thresholding. *NPAR '08: Proceedings of the 6th International Symposium on Non-photorealistic Animation and Rendering*. New York: ACM Press; 2008. pp. 39–47.
26. Morales M. *Virtual Mo*. 2010. Available at www.virtualmo.com/. Accessed 2010 Jun 1.
27. Kaplan C, Bosch R. TSP Art. *Bridges 2005: mathematical connections in art, music, and science*. 2005. pp. 301–308.
28. Bosch R. Simple-closed-curve sculptures of knots and links. *J Math Arts* 2010;4(2):57–71.
29. Applegate D, Bixby R, Chvátal V, *et al.* Concorde: a code for solving traveling salesman problems. 2010. Available at www.tsp.gatech.edu/concorde/. Accessed 2010 Jun 1.
30. Cook WJ. TSP test data: TSP art instances. 2010. Available at www.tsp.gatech.edu/data/art/. Accessed 2010 Jun 9.
31. Pedersen H, Singh K. Organic labyrinths and mazes. *NPAR '06: Proceedings of the 4th International Symposium on Non-photorealistic Animation and Rendering*. New York: ACM Press; 2006;79–86.

OPERATIONS RESEARCH MODELS FOR CANCER SCREENING

OGUZHAN ALAGOZ
TURGAY AYER
FATIH SAFA ERENAY
Department of Industrial and
Systems Engineering,
University of
Wisconsin-Madison, Madison,
Wisconsin

INTRODUCTION

In 2008, approximately 1,437,180 Americans were diagnosed with cancer and 565,650 died from it: more than 1500 people a day, making cancer the second leading cause of death in the United States [1]. American men have slightly less than a 1 in 2 lifetime risk of developing cancer; for American women, the risk is a little more than 1 in 3 [1]. Table 1 presents the estimated number of new cases and deaths by selected cancer types in 2008. Figure 1 shows how the incidence and mortality rates of cancer change over time. The economic impact of cancer is also huge. According to the National Institutes of Health, the total cost of cancer care in the United States in 2007 was \$219.2 billion, where direct medical costs accounted for \$89 billion [1].

Although cancer is a highly fatal disease, as a result of advanced treatment options, most cancers are now curable if detected early. For instance, Surveillance Epidemiology and End Results (SEER) program reports that five-year survival rates for localized (an early stage) and distant (an advanced stage) breast cancer patients are 98% and 23%, respectively [2]. Obviously, a successful screening program would improve the effectiveness of cancer treatment. The five-year relative survival rate for all cancers has improved from 50% in 1975–1977 to 66% in 1996–2003 [1], which can be attributed to the

progress in both diagnosing certain cancers at an earlier stage and improvements in treatment the progress [3]. Table 2 lists screening modalities for selected cancer types.

Although screening of a population of asymptomatic individuals can diagnose some early-stage cancers, the vast majority of these individuals receive no benefit since cancer deaths are prevented in only a small fraction of the screened population [5]. In fact, the screened population may be exposed to additional health risks and even death (e.g., perforation during a colonoscopy screening) as a result of screening, such as complications from the screening procedure, from false-positive test results that trigger unnecessary invasive follow-up procedures (e.g., breast biopsy after mammography), and from misdiagnosis and overtreatment of cancers that would have never caused death [5]. Furthermore, population screening is very expensive. Burnside *et al.* [6] estimate that the total costs of just mammography screening for breast cancer and associated work-up of positive findings were between \$3–5 billion in 2001. Screening asymptomatic individuals therefore requires a careful analysis of the benefits and risks of screening in both utilitarian and economic terms.

The complexity of selecting a cost-effective cancer screening program creates many challenges that can be addressed using a variety of operations research (OR) tools. There are several review articles that summarize statistical and OR applications in cancer screening. Heidenberger [7] reviews the studies using quantitative approaches that focus on determining which screening strategy should be used for a population or an individual. Stevenson [8] summarizes statistical models considering the problem of planning and evaluation of cancer screening, whereas Pierskalla and Brailer [9] summarize mathematical models for cancer screening including the earliest OR applications. Knudsen *et al.* [5] provide a comprehensive summary of articles that use simulation methodology, the most

Table 1. Estimated New Cancer Cases and Deaths in the US in 2008 for Selected Cancer Types [1]

Cancer Type	Estimated New Cases	Estimated Deaths
Lung cancer	215,020	161,840
Prostate cancer	186,320	28,660
Breast cancer	184,450	40,930
Colorectal cancer	148,810	49,960
Bladder cancer	68,810	14,100
Skin cancer	67,720	11,200
Leukemia	44,270	21,710
Pancreas cancer	37,680	34,290
Gastric cancer	21,500	10,880
Other cancers	462,600	192,080
Total	1,437,180	565,650

commonly used OR tool in cancer screening. Among all modeling (particularly simulation) studies, research by the Cancer Intervention and Surveillance Modeling Network (CISNET), a consortium of National Cancer Institute (NCI)-sponsored investigators, is noteworthy. CISNET investigators use simulation/statistical modeling and common inputs to assess the impact of interventions (e.g., screening and treatment) on current population trends in cancer incidence and mortality. More information about CISNET models is available elsewhere [3,10–12].

This article provides a review of the OR studies considering cancer screening that are either briefly described or not mentioned in previous review articles. We classify relevant literature by cancer type. We start with summarizing cancer screening studies that

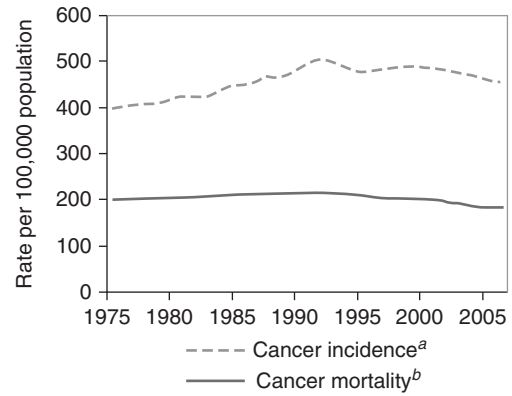


Figure 1. Cancer incidence and mortality rates in the United States between 1975–2006 [2]. Cancer incidence^a data come from SEER based on nine sites (San Francisco, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, and Atlanta) whereas cancer mortality^b data come from US Mortality Files, National Center for Health Statistics, Centers for Disease Control and Prevention. Both rates are per 100,000 and are age-adjusted to the 2000 US population.

are not applied to any specific cancer type. We then describe OR models on screening for breast cancer, colorectal cancer, cervical cancer, and other cancer types. We present a list of the articles described in this study in Table 3. Finally, we briefly summarize some issues for future research in cancer screening.

GENERAL MODELS FOR CANCER SCREENING

This section describes general (mostly theoretical) studies that are not applied to any

Table 2. Screening Modalities for Selected Cancer Types [4]

Cancer Type	Screening Modalities
Breast cancer	Clinical breast examination, X-ray mammography, magnetic resonance imaging (MRI), ultrasound computed tomography (CT)
Colorectal cancer	Colonoscopy, fecal occult blood test (FOBT), sigmoidoscopy, virtual colonoscopy, digital rectal examination
Prostate cancer	Digital rectal examination, prostate specific antigen (PSA) blood test
Lung cancer	Chest X-ray, computed tomography (CT) scan
Cervical cancer	Pap smear test, HPV test
Bladder cancer	Microscopic urinalysis, urine dipstick, urine cytology, bladder tumor antigen (BTA), nuclear matrix protein (NMP22) immunoassay
Ovarian cancer	Cancer antigen 125 (CA-125) test, pelvic examination, transvaginal ultrasound

Table 3. Summary of OR Models in Cancer Screening

References	Cancer Type	Objective	Methodology
Prorok [13,14]	General	Provide explicit formulation for the proportion of detected cancer	Stochastic stationary point process
Butler [15]		Maximize expected lifetime	Discrete-time Markov process
Lee and Pierskalla [16]		Minimize aggregate detection delay	General stochastic model
Milioni and Pliska [17]		Minimize screening cost and minimize death probability	Continuous-time semi-Markov process
Houshyar [18]		Minimize the total cost of screening program	General stochastic model
Hanin <i>et al.</i> [19]		Minimize the difference between the expected tumor size with and without screening	General stochastic model
Özekici and Pliska [20]		Minimize total costs	MDP
Parmigiani [21–23]		Minimize expected losses	Non-Markovian stochastic model
Zelen [24] Lee and Zelen [25–27], Lee <i>et al.</i> [28]		Maximize total utility	General stochastic model
Tsodikov <i>et al.</i> [30]		Minimize the expected delay time for the diagnosis of a recurrent cancer	General stochastic model
Baker [31]	Breast	Minimize the cost of screening and lost life-years	General stochastic model
Günes <i>et al.</i> [32]		Minimize breast cancer deaths	Queuing model
Maillart <i>et al.</i> [33]		Maximize survival probability	POMDP
Ayer <i>et al.</i> [34]		Maximize QALYs	POMDP
Ivy [35]		Minimize costs of screening and treatment	POMDP
Chhatwal <i>et al.</i> [36]		Maximize QALYs	MDP
Frazier <i>et al.</i> [37]		Find cost-effectiveness of alternative screening programs	Markov process and simulation
Preston and Smith [38]		Minimize the total cost of screening	Continuous-time Markov process and embedded Markov process
Clemen and Lacke [40]		Maximize the number of prevented cancers and minimize the total cost of screening	Simulation
Parmigiani <i>et al.</i> [41]		Minimize the probability of becoming infected before the screening examination	Continuous-time stochastic model
Leshno <i>et al.</i> [42]	CRC	Find cost-effectiveness of alternative screening policies	POMDP
Harper and Jones [43]		Maximize survival and the number of detected CRC	Semi-Markov process and simulation
Yaesoubi and Roberts [44]		Minimize a function of total cost and QALYs (negative term)	Game theoretical model
Erenay and Alagoz [45]		Maximize QALYs and minimize total cost of screening and treatment	POMDP
Myers <i>et al.</i> [46]		Find cost-effectiveness of alternative screening policies	Markov process

Table 3. (Continued)

References	Cancer Type	Objective	Methodology
Goldie <i>et al.</i> [47]		Find cost-effectiveness of alternative screening policies	Markov process
Kulasingam <i>et al.</i> [48]		Find cost-effectiveness of alternative screening policies	Markov process
Tsodikov and Yakovlev [29]	Lung	Minimize the delay time for cancer detection and minimize total cost	General stochastic model
Pinsky [49]		Provide explicit formulation for the expected number of detected cancers	Convolution model
Tsodikov <i>et al.</i> [50]	Prostate	Provide explicit formulation for prostate cancer lead-time and over diagnosis rate	Stochastic model
Kent <i>et al.</i> [51]	Bladder	Minimize detection delay	Nonlinear programming
Havrilesky <i>et al.</i> [52]	Ovarian	Find cost-effectiveness of alternative screening policies	Markov process
Davies <i>et al.</i> [53]	Gastric	Maximize reduction in mortality	Simulation

specific cancer type. In two related articles, Prorok [13,14] introduces a stochastic stationary point process that represents the natural history of a chronic disease such as cancer, under a given screening program. Prorok's stochastic model includes three states: disease-free, preclinical, and clinical states, where the sojourn time for each state is a random variable. The author defines the number of screenings as an exogenous variable to the model. The author also assumes that the time horizon is divided into equal-sized intervals and each screening is assumed to occur in one of these intervals. Prorok provides explicit formulations for the mean lead time, mean forward recurrence time, and proportion of detected cancers by screening for this model. The author performs a numerical analysis using hypothetical parameters.

Butler [15] provides a four-state discrete-time Markov process model to determine

the optimal inspection schedule for a system where the inspection itself may cause the failure of the system (e.g., X-ray screening to detect cancer). The author provides an explicit formulation for the inspection schedule that maximizes the expected lifetime of the system and shows in which cases it is optimal not to screen, always to screen, and to screen periodically. The author does not perform any computational analysis.

Lee and Pierskalla [16] develop a general stochastic model that can be used to make a cost-benefit analysis of population screening programs for contagious and noncontagious diseases. The objective of this model is to minimize the aggregate detection delay (or equivalently, minimize the average number of people who have the disease) subject to a budget constraint. The authors convert this stochastic model to a knapsack problem and solve it optimally. They consider both

perfect and imperfect tests, but assume constant screening test accuracy over time. They further assume that the number of susceptible people in the population is constant and the incidence rate is time-stationary. They account for noncompliance to recommended policies. They show that when tests are perfect (i.e., sensitivity and specificity of the tests are 100%), optimal screening policy consists of equally spaced screening intervals.

Milioni and Pliska [17] present a stochastic model to find the optimal inspection schedule for a deteriorating system progressing toward a failure, where the deterioration process is modeled as a continuous-time semi-Markov process. They consider two objectives: minimizing the inspection cost subject to the constraint that the probability of failure is less than a specified value and minimizing the failure probability while the total number of screenings is fixed. They note that their model can be applied to the cancer screening problem. They provide dynamic-programming-based methods to solve the optimal inspection scheduling problem for both objectives.

Houshyar [18] presents a general stochastic model formulation to derive the optimal screening policy for a given disease and population. The objective is to minimize the total cost of a screening program including the costs of screening tests, treatment, and death. The author classifies disease states as occult, clinically surfaced or detected by examination, cured, and death. The author assumes that the screened population is closed and death from cancer is independent of death from other causes. The author concludes that the optimal screening strategy is typically aperiodic and the screening interval is age-dependent.

Hanin *et al.* [19] provide a stochastic model that represents cancer recurrence under a specific screening schedule (or no screening) to evaluate the performance of cancer surveillance programs. They measure the performance of a given screening program by the difference between the expected tumor size with and without screening. They break the tumor progression into three stages: formation of initiated cells, transition from initiated cells to malignant clonogenic cells (cells

that have the ability to proliferate to regenerate the tumor), and transition from malignant clonogenic cells to a malignant tumor. They model time in each stage as a random variable and provide examples from different distributions. They present an explicit formulation for screening efficiency and derive the structural properties for an optimal screening schedule. They also present numerical results with hypothetical data. They find that a fixed number of screenings should be uniformly distributed over the decision horizon under exponential tumor growth.

MODELS FOR BREAST CANCER SCREENING

Breast cancer is probably the most commonly studied cancer type in the OR literature. There are several early studies on breast cancer screening such as Zelen and Feinleib [54], Kirch and Klein [55], Blumenson [56], Shwartz [57], Eddy [58], Voelker and Pierskalla [59], van Oortmarssen *et al.* [60], Houshyar and Al-Khayyal [61], which are described in the previous survey articles in detail.

Özekici and Pliska [20] develop an infinite-horizon, undiscounted, delayed Markov decision process (MDP) model that minimizes the total expected cost of inspection, false-positives, corrective actions (treatments), and failures such as deaths. They refer to this model as a delayed MDP because the sojourn time in healthy state is a general nonnegative random variable but once the process leaves healthy states, it behaves like a Markov process. They assume the disease aggression and test accuracies are stationary. They use breast cancer screening problem as an example and show that the optimal policy is not to inspect a person for whom the cost of failure is under \$55,000, where the cost of failure is defined as the sum of terminal medical care costs and a dollar value assigned to the possible loss of life.

Parmigiani [21] addresses the question of which age groups and what part of population to screen for breast cancer. The author utilizes a continuous-time non-Markovian stochastic model for disease progression and calculates the optimal policy that minimizes expected losses including financial costs,

side effects, wasted time, stress due to false-negative test results, mortality, morbidity, and cost of treatment. The author presents a set of sufficiency conditions which ensure that screening frequency increases with age. In another paper [22], Parmigiani finds an exact solution to this model when there are only two disease states and the test is imperfect. In a related paper, Parmigiani [23] studies the question of whether to screen or not, given the age and risk factors of the patient and whether to re-examine a recently screened patient.

Zelen [24] develops a continuous-time stochastic model to find the optimal screening policy such that the objective is to maximize a general weighted utility function. The author considers age-specific incidence while keeping all other parameters stationary. The author concludes that if the incidence is stationary and the sensitivity of the test is perfect, then the screening intervals should be equally spaced. As a result, the author suggests that women over 50 should have an annual mammography screening.

In a series of papers, Lee and Zelen [25–27] and Lee *et al.* [28] extend the model of Zelen [24] to address various problems in cancer screening such as the optimal time to initiate screening. Lee and Zelen [25] consider the problem of finding the best periodic screening schedule for the early detection of a given disease. They develop a three-state stochastic model based on the model of Zelen [24] and determine the best screening schedule using two approaches. The first approach imposes a limit to the probability of being in a preclinical state whereas the second approach maximizes the ratio of expected number of patients diagnosed by the screening test over the expected number of cancer patients. They assume patients have different disease development risks and the time in preclinical state is exponentially distributed. They then illustrate the performance of their proposed screening schedules on breast cancer. In a related paper, Lee and Zelen [27] develop a stochastic model for predicting the mortality for specific screening policies for breast cancer. They consider age-specific incidence and sensitivity. They use Breast Cancer

Surveillance Consortium (BCSC) data and conclude that initiating screening at age 40 instead of age 50 would lead to a small reduction in mortality.

Tsodikov *et al.* [30] consider breast cancer recurrence under a specific screening schedule. They build a stochastic model for the recurrence of a posttreatment cancer by assuming that the initial number of clonogenic cells follows Poisson distribution and the time until tumor recurrence from each clonogenic cell is identically distributed. They explicitly model cancer recurrence time and apply the optimization framework provided by an earlier article by Tsodikov and Yakovlev [29] to determine the optimal discrete-time cancer screening schedule that minimizes the expected delay time for the diagnosis of a recurrent cancer. They also perform a numerical analysis using BCSC data.

Baker [31] builds a stochastic model to evaluate the cost of screening and the cost of life-years lost due to breast cancer. The author measures costs such as screening costs in terms of months of life. Baker assumes one month of life is worth eight screenings. The author also assumes that ductal carcinoma *in situ* (DCIS) tumors, a noninvasive breast disease, are not harmful and thus can be ignored. The author evaluates a set of policies and concludes that the optimal screening policy should start at age 48, end at age 61.5, and should occur approximately every 2.5 years.

Günes *et al.* [32] develop a queueing model to investigate the impacts of various dynamic factors in mammography screening that are in interaction such as mammography reading volume and quality, access to screening, and delays in service. Their queueing model represents patients as the customers and radiologists as the servers, where the service capacity is limited. They also develop a simulation model. They minimize breast cancer deaths, while keeping system cost as a constraint. They find that improving the dissemination without an increase in the quality of the screening tests would result in excessive waste of resources.

Maillart *et al.* [33] develop a partially observable Markov chain model to investigate the relative value and frequency of

mammography screening for premenopausal women versus postmenopausal women. Their objective is to maximize the survival probability of a given patient. Using sample path enumeration, they evaluate 1223 policies, where each policy consists of two different screening intervals. They consider age-specific disease aggression and imperfect tests (both false-positive and false-negative test results). They show that screening should start at early ages and continue relatively late in life regardless of the screening intervals adopted.

Ayer *et al.* [34] develop a finite-horizon partially observable Markov decision process (POMDP) model to determine the optimal mammography screening strategy for an individual patient. The objective of this POMDP model is to find the optimal personalized screening policy that maximizes the total expected quality-adjusted life-years (QALYs) of a woman during her lifetime. They optimally solve this POMDP model using real data. They consider age-dependent disease progression and test accuracies as well as the possibility of self-detection. They show that the individualized personalized screening schedules outperform the existing guidelines with respect to the total expected QALYs, while significantly decreasing the number of mammograms. They further demonstrate that the mammography screening threshold risk increases with age. They also derive several structural properties of their POMDP model, including the sufficiency conditions that ensure the existence of a threshold cancer risk over which it is optimal to mammogram.

Ivy [35] builds a POMDP model that determines a cost-optimal screening and treatment policy from the payer's perspective while accounting for the patient's utility. The author assumes that both the likelihood of death from other causes and test accuracy are age-independent. The author discretizes the state space of the POMDP model, converts it into an MDP model, and solves this MDP using data from the literature. Ivy then presents the policy trade-off curves based on probabilities of *in situ* and invasive cancer while balancing payer cost and patient utility.

Chhatwal *et al.* [36,62] develop a finite-horizon MDP model to optimize breast-biopsy decisions based on mammography findings and demographic risk factors for women undergoing mammography. They compute the optimal biopsy decisions as a function of the probability of cancer while maximizing the total expected QALYs of a woman. They perform a structural analysis of the MDP model and prove several structured policies such as the optimal policy is a control-limit type for all ages (there exists a threshold probability of cancer such that it is always optimal to biopsy over that threshold) and the optimal control limit is nondecreasing with age. They use clinical data to solve the MDP model optimally and show that as women get older, the biopsy should be made less aggressively since women in older age groups benefit less from aggressive biopsying due to limited life expectancies.

Models for Colorectal Cancer Screening

The earlier studies on colorectal cancer (CRC) screening such as the studies by Eddy [63] and Eddy *et al.* [64] are described in previous review articles. Frazier *et al.* [37] compare the cost-effectiveness of several CRC screening policies suggested by several expert panels. They model the progression of CRC using a Markov process and develop a Markov-process-based simulation model to evaluate the performance of a given screening policy. They use data from literature and SEER database to simulate the US population by gender and age, and evaluate the performance of each screening policy. They perform extensive sensitivity analyses and report the best screening policies for each of these cases. They suggest the use of annual fecal occult blood test (FOBT) plus five-year sigmoidoscopy for CRC screening.

Preston [38] uses a continuous-time Markov process, an extension of the model described by Preston and Smith [39], to evaluate the performance of periodic screenings for a specific disease and a screening test with a particular sensitivity and specificity. The author models disease progression as an embedded Markov process with five system states. The author assumes that

the screening interval is constant (i.e., only periodic screenings are considered) and incorporates the total cost of cancer screening into the model. Preston also uses a parametric critical test value, a threshold that determines whether the test result is negative or positive. The author shows that the optimal combination of screening interval and critical test value minimizing the total expected cost of screening can be obtained by a nonlinear optimization model. The author presents numerical results for CRC screening when the only available screening modality is FOBT.

Clemen and Lacke [40] develop a simulation model by adding stochastic variables such as cancer dwelling time to an earlier deterministic model of Wagner *et al.* [65] and compare 30 different CRC screening policies. They use expert opinion and statistical methods to estimate the input parameters of the simulation model. They consider two criteria to compare various screening policies: number of prevented cancers and expected cost per saved life year. Their results show that undergoing colonoscopy every three years is the best screening policy.

Parmigiani *et al.* [41] develop a stochastic model to find the optimal timing of colonoscopy in a continuous-time setting in which individuals are examined only once in their lifetime. They derive a closed-form solution for the optimal timing of screening by maximizing a utility function which is equivalent to minimizing the probability of becoming infected before the screening examination. Therefore, they find the optimal time of screening by taking the derivative of this probability function. They assume that the sensitivity of the screening test is the same for all ages. They find that the optimal time for screening is at age 75 if the patient undergoes screening just once during his/her lifetime.

Leshno *et al.* [42] make a cost-effectiveness analysis for CRC screening for an individual with average risk of developing CRC. They build a POMDP model to represent the progression of CRC for a given screening strategy and use it to evaluate the performance of a set of common CRC screening strategies. They use data from the literature

to estimate the input parameters and perform an extensive numerical analysis, which suggests that the use of colonoscopy every ten years and the use of annual FOBT plus five-year sigmoidoscopy are the most cost-effective two strategies.

Harper and Jones [43] consider the problem of determining the best screening policy for an average patient using semi-Markov process and simulation models. They develop a five-state semi-Markov process model using general transition time distributions. They show how to calculate the probability of early polyp and cancer detection given that the time between two consecutive screenings is constant. The authors do not provide any numerical results for the semi-Markov process model; instead, they present a detailed simulation model where disease progression is modeled as a semi-Markov process. The state space of the simulation model consists of conventional cancer states (clean, polyp, cancer) and metastatic CRC state. They utilize clinical data from European health institutions to evaluate and compare six different screening policies with respect to several criteria such as mean survival time from CRC diagnosis, percentage of detected polyps, cancers, etc.

Yaesoubi and Roberts [44] develop a sequential game theory model with perfect information to analyze the decisions of an insurance company for covering CRC screenings. The insurer decides which screening alternatives to cover, how much the insurance premium would increase to cover CRC screenings, and how much the insurees would have to pay to cover recommended screenings. The insurer's objective is to minimize a weighted average of the total screening/treatment cost and total QALYs of all of the patients (a negative term). In return, the insurees choose the screening alternative that maximizes a utility function of screening cost and their total QALYs. The authors assume that the decisions are made every 10 years. They also assume that patients have perfect information about the progression of the CRC and are able to estimate their total expected QALYs. They use a simulation model and clinical data to estimate the progression of CRC under

certain screening alternatives. The details of this simulation model are explained in another related paper by Roberts *et al.* [72]. Yaesoubi and Roberts assume that the patients belong to the two risk groups (low and high risk) and they consider only two screening alternatives: FOBT and colonoscopy. Their numerical results show that covering CRC screening, especially for low risk patients, is not economically attractive for the insurer unless the insurer's valuation of QALY is very high.

Erenay and Alagoz [45] develop a finite-horizon POMDP model that can be used to obtain a patient-specific optimal lifetime screening policy. They use two optimization criteria: total expected QALYs of the patient and total cost of screening and treatment. They consider the effect of age, gender, and personal CRC history on CRC progression. They model patients with both low and high risk of developing polyps, the precursor of CRC, separately. They also account for the secondary occurrence of CRC, which enables the model to determine the optimal screening policy for CRC surveillance. They perform a numerical analysis by using data from SEER database, clinical databases, and literature.

Models for Cervical Cancer Screening

The earlier studies on cervical cancer screening such as Lincoln and Weiss [66], Eddy [67], Habbema *et al.* [68], and Boyd *et al.* [69] were described in previous survey articles in detail. Myers *et al.* [46] develop a Markov model that represents the natural history of the human papillomavirus (HPV), the primary cause of cervical cancer, and cervical cancer under a specific Pap smear screening. They use the Markov model to analyze the effects of different screening interval choices on cervical cancer incidence for a cohort population. They vary starting age for screening to evaluate the performance of screening policies using Pap smear test with 1-to-5-year intervals and compare them to no screening. They show that as the screening intervals decrease, the performance of the screening improves. However, the benefit of more frequent screening reduces for women over a threshold age. They also note that the performance difference is insignificant if the

screening starts at a very early age such as 15–25.

There are several other studies that analyze the effect of HPV screening and prevention techniques on cervical cancer prevention, by using Markov models. Goldie *et al.* [47] evaluate the cost-effectiveness of the HPV DNA test in addition to conventional cytology tests scheduled according to the national guidelines. Kulasingam *et al.* [48] consider whether adding HPV vaccination to the cervical cancer screening policy suggested by national guidelines as compared to screening alone is cost-effective or not. Their numerical results show that both DNA test with conventional screening and vaccination with conventional screening are, cost-effective when compared to conventional screening alone.

Models for Screening Other Cancers

Tsodikov and Yakovlev [29] consider the problem of optimizing cancer screening schedule for a given patient and apply their theoretical findings to lung cancer screening. They classify the health states as disease-free, infected, and heavily infected. They perform an analytical study by developing a stochastic model that optimally schedules cancer screenings over a given decision horizon for given data. They first assume that the number of total screenings is fixed and minimize the total expected delay time to diagnose the cancer. They then relax this assumption and minimize the total cost of cancer screening by assigning a unit cost for delay time to cancer detection. They further assume that there is no symptomatic diagnosis and cancer diagnosis probability after screening increases with time within the same health state. They derive several structural properties of the optimal screening schedule such as the existence of an optimal periodic screening schedule when the total number of screenings is fixed.

Pinsky [49] develops a natural history model for a specific disease and measures the performance of a screening policy for a target population. The author develops a five-state convolution model by relaxing the Markovian assumption in Chen *et al.*'s [70] five-state Markov process model. Pinsky demonstrates the computation of the expected number of

screen-detected preclinical and clinical disease states given a specific screening program and provides closed-form solutions for various components of the model such as sojourn time and over-diagnosis rate. The author uses data from literature and Mayo Lung Cancer Screening Trial (MLCST) [71] and performs a numerical analysis for lung cancer, which shows that the model's results fit well to the MLCST. The author also investigates how screening interval affects the model's outputs. For example, the author shows that lead times are not affected by changes in screening interval whereas the proportion of the cohort ever diagnosed with late-stage lung cancer is very sensitive to the screening intervals.

Tsodikov *et al.* [50] use a three-stage stochastic model to represent prostate cancer progression under a PSA screening program. They explicitly formulate certain characteristics of the prostate cancer progression including mean lead time, prostate cancer incidence under a given screening program, and over-diagnosis rate for each patient type and representative population. They then use SEER data to perform a numerical analysis and estimate these characteristics for prostate cancer.

Kent *et al.* [51] develop a nonlinear optimization model to determine the optimal time for the bladder cancer screening from the physician's perspective. They optimize the expected detection delay while minimizing the probability that the cancer spreads to other parts of the body. They assume that only a limited number of screenings is available and recurrence of tumors follows a Poisson process. They further assume that the screening tests are perfect. They conclude that their proposed screening policy would save three weeks of delay in detection when compared to clinical practice.

Havrilesky *et al.* [52] build a Markov-process-based simulation model to estimate the cost-effectiveness of various screening strategies for ovarian cancer. They use age-specific incidence and mortality rates from SEER data. Their outcome measures include mortality reduction, number of false-positive tests, positive predictive value, years of life saved, lifetime economic costs, and incremental cost-effectiveness

ratio. They conclude that annual screening is cost-effective and cost-effectiveness of screening is most sensitive to test frequency, specificity, and cost.

Davies *et al.* [53] build a discrete-time simulation model to measure the performance of a screening program against *Helicobacter pylori* infection, the primary cause for gastric cancer, in England and Wales, where individuals under age 50 would be screened only once. They use data from the literature and the simulation model to estimate the reduction in mortality due to the screening program and the cost of the program. They report sensitivity analyses for five parameters with high uncertainty.

FUTURE RESEARCH IN CANCER SCREENING

Cancer screening is very controversial with many open research questions such as when to start and end screening? How often should a target population undergo a particular screening modality? Is the use of dynamic screening interval/policy a better strategy than the current screening recommendations using static screening strategies? How can a more personalized screening policy based on individual risk factors be implemented? What would be the effects of requiring some patients to use more expensive but more sensitive screening modalities? Is it cost-effective to use a new screening technology over an existing one? What are the mortality and morbidity benefits of a particular screening program?

Because of its importance, there is a growing interest among operations researchers in studying the cancer screening problem. As Maillart *et al.* [33] note, there are two approaches taken in quantitative cancer screening models. On one extreme, which is taken by most studies in the literature, empirical, cost-effectiveness simulation models based on clinical trials and clinical data are used. A simulation model can be used to find the best strategy that maximizes an objective by testing all reasonable screening strategies and comparing their outcomes. This is feasible only when the number of reasonable screening policies is very small.

However, the number of possible screening policies (due to open questions such as when to start and end screening, what should be the optimal screening interval for different risk groups, which screening technology should be used for which type of patients) increases dramatically and hence the computation of all cost-effective strategies becomes intractable.

At the other extreme, analytical models are utilized, which have several modeling limitations. For instance, most of these studies do not perform extensive numerical analyses using reliable real clinical data. In particular, while several researchers solve their analytical models using data from the literature, some of the data sources include unreliable/inaccurate estimations, which may affect the policy recommendations made by these studies. The validation of the data used in these studies is typically missing whereas it is of crucial importance. Furthermore, the existing studies make several unrealistic assumptions by ignoring important factors affecting the screening decisions. For instance, many studies consider only fixed screening intervals (i.e., the screening interval does not depend on the results of the previous screening tests) whereas the screening policy should consider the results of the previous tests in making further screening recommendations.

Moreover, several studies do not explicitly model precursor states, which is very critical in modeling diseases (such as CRC) that have a precursor clinical state (such as polyps in CRC) which progresses to cancer with time. That is, certain cancer types may not be represented properly with these models. Because removing such precursors affects further cancer development, any realistic model should consider them explicitly. Most researchers assume constant screening test accuracy, that is, the accuracy of the screening tests does not depend on age. Similarly, several studies assume that the disease aggression does not change with age although there is strong evidence that the progression of the diseases is highly age-dependent.

In addition to these common limitations of the existing studies, some of the other limiting assumptions in these models include

the following: the number of the susceptible people in the population is constant and the incidence rate is time-stationary; the time in preclinical state is exponentially distributed; the risk of cancer is the same for all population; the number of available screening tests is equal to one; screening tests are perfect; tumor recurrence follows a Poisson process; self-diagnosis of diseases (such as breast cancer) is not possible; the disease mortality does not change with age.

Although the validation of OR models constitutes a barrier for their policy implementation, mathematical modeling is an essential tool for planning and improving cancer screening due to the complexity of the cancer screening problem. Future research in cancer screening may need to combine the two approaches described earlier. Furthermore, there is also a strong need to develop models without making any unrealistic assumptions, such as the ones listed earlier. Another important issue in cancer screening studies is to conduct numerical analyses using reliable data. Fortunately, there are many statistical modeling approaches in the literature that study disease progression and treatment decisions more rigorously. Combining the OR methods with statistical tools that analyze real clinical data more rigorously is another promising area of research. Considering huge economic cost of and lives lost due to cancer, more research is needed to design better screening programs, the most cost-effective means for controlling this disease.

Acknowledgments

This article was supported in part by National Science Foundation grant CMMI-0844423. The authors thank three anonymous referees for their suggestions and insights, which improved this manuscript.

REFERENCES

1. American Cancer Society. Cancer facts and figures. Available at <http://www.cancer.org/downloads/STT/2008CAFFfinalsecured.pdf>. Accessed 2008 Nov 24.
2. Surveillance Epidemiology and End Results. SEER cancer statistics review

- 1975–2006. Available at http://seer.cancer.gov/csr/1975_2006/results_single/sect_04_table.12.pdf. Accessed 2009 Jul 17.
3. Berry DA, Cronin KA, Plevritis SK, *et al.* Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med* 2005;353(17):1784–1792.
4. National Cancer Institute. Screening and testing to detect cancer. Available at <http://www.cancer.gov/cancertopics/screening>. Accessed 2009 Jul 16.
5. Knudsen AB, McMahon PM, Gazelle GS. Use of modeling to evaluate the cost-effectiveness of cancer screening programs. *J Clin Oncol* 2007;25(2):203–208.
6. Burnside E, Belkora J, Esserman L. The impact of alternative practices on the cost and quality of mammographic screening in the United States. *Clin Breast Cancer* 2001;2(2):145–152.
7. Heidenberger K. Strategic investment in preventive health care: quantitative modelling for programme selection and resource allocation. *OR Spectrum* 1996;18(1):1–14.
8. Stevenson C. Statistical models for cancer screening. *Stat Methods Med Res* 1995;4(1):18–32.
9. Pierskalla WP, Brailer DJ. Applications of operations research in health care delivery. In: Pollock SM, Rothkopf MH, Barnett A, editors. Volume 6, Handbooks in operations research and management science. Amsterdam: Elsevier Science; 1994. Chapter 13.
10. NCI Statistical Research & Applications Branch. Cancer intervention and surveillance modeling network. Available at <http://cisnet.cancer.gov/>. Accessed 2008 Nov 25.
11. Cancer Intervention and Surveillance Modeling Network (CISNET) Breast Cancer Collaborators. The impact of mammography and adjuvant therapy on US breast cancer mortality (1975–2000): collective results from the cancer intervention and surveillance modeling network. *J Natl Cancer Inst Monogr* 2006;36:1–126.
12. NCI Statistical Research & Applications Branch. CISNET Publications. Available at <http://cisnet.cancer.gov/publications/>. Accessed 2008 Nov 25.
13. Prorok P. The theory of periodic screening II. Doubly bounded recurrence times and mean lead time and detection probability estimation. *Adv Appl Probab* 1976;8(3):460–476.
14. Prorok P. The theory of periodic screening I: lead time and proportion detected. *Adv Appl Probab* 1976;8(1):127–143.
15. Butler D. A hazardous-inspection model. *Manage Sci* 1979;25(1):79–89.
16. Lee HL, Pierskalla WP. Mass screening models for contagious diseases with no latent period. *Oper Res* 1988;36(6):917–928.
17. Milioni A, Pliska S. Optimal inspection under semi-Markovian deterioration: the catastrophic case. *Nav Res Log* 1988;35(5):393–411.
18. Houshyar A. What can be done to screen progressive diseases optimally. *IIE Trans* 1991;23(1):40–50.
19. Hanin L, Tsodikov A, Yakovlev A. Optimal schedules of cancer surveillance and tumor size at detection. *Math Comput Model* 2001;33(12–13):1419–1430.
20. Özekici S, Pliska SR. Optimal scheduling of inspections: a delayed markov model with false positives and negatives. *Oper Res* 1991;39(2):261–273.
21. Parmigiani G. On optimal screening ages. *J Am Stat Assoc* 1993;88(422):622–628.
22. Parmigiani G. Optimal scheduling of fallible inspections. *Oper Res* 1996;44(2):360–367.
23. Parmigiani G. Timing medical examinations via intensity functions. *Biometrika* 1997;84(4):803–816.
24. Zelen M. Optimal scheduling of examinations for the early detection of disease. *Biometrika* 1993;80(2):279–293.
25. Lee S, Zelen M. Scheduling periodic examinations for the early detection of disease: applications to breast cancer. *J Am Stat Assoc* 1998;93(444):1271–1272.
26. Lee SJ, Zelen M. Modelling the early detection of breast cancer. *Ann Oncol* 2003;14(8):1199–1202.
27. Lee SJ, Zelen M. Mortality modeling of early detection programs. *Biometrics* 2008;64(2):386–395.
28. Lee S, Huang H, Zelen M. Early detection of disease and scheduling of screening examinations. *Stat Methods Med Res* 2004;13(6):443–456.
29. Tsodikov A, Yakovlev A. On the optimal policies of cancer screening. *Math Biosci* 1991;107(1):21–45.
30. Tsodikov A, Asselain B, Fourque A,

- et al.* Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems. *Biometrics* 1995;51(22): 437–447.
31. Baker RD. Use of a mathematical model to evaluate breast cancer screening policy. *Health Care Manag Sci* 1998;1(2): 103–113.
 32. Günes ED, Chick SE, Aksin OZ. Breast cancer screening services: trade-offs in quality, capacity, outreach, and centralization. *Health Care Manag Sci* 2004;7(4): 291–303.
 33. Maillart LM, Ivy JS, Diehl K, Ransom S. Assessing dynamic breast cancer screening policies. *Oper Res* 2008;56(6): 1411–1427.
 34. Ayer T, Alagoz O, Stout NK. A mathematical model to optimize breast cancer screening policy. Proceedings of the 31st Annual Meeting of the Society for Medical Decision Making Abstract; 2009.
 35. Ivy JS. A maintenance model for breast cancer treatment and detection. Working paper, also presented at INFORMS 2001 Annual Meeting; Miami (FL); 2001.
 36. Chhatwal JC, Alagoz O, Burnside EB. When to biopsy in breast cancer diagnosis? A quantitative model using Markov decision processes. Proceedings of the 29th Annual Meeting of the Society for Medical Decision Making Abstract; Pittsburgh (PA); 2007.
 37. Frazier A, Colditz G, Fuchs C, *et al.* Cost-effectiveness of screening for colorectal cancer in the general population. *J Am Med Assoc* 2000;284(15):1954–1961.
 38. Preston AJ, Smith W. Disease screening designs: sensitivity and screening frequency. Proceedings of the Annual Meeting of the American Statistical Association; August 5–9; Atlanta (GA); 2001.
 39. Preston AJ, Smith W. Stochastic model for optimal medical screening intervals: results for mixed disease models. Proceedings of Biopharmaceutical Section of American Statistical Association; Dallas (TX); 1998.
 40. Clemen R, Lacke C. Analysis of colorectal cancer screening regimens. *Health Care Manag Sci* 2001;4(4):257–267.
 41. Parmigiani G, Skates S, Zelen M. Modeling and optimization in early detection programs with a single exam. *Biometrics* 2002;58(1):30–36.
 42. Leshno M, Halpern Z, Arber N. Cost-effectiveness of colorectal cancer screening in the average risk population. *Health Care Manag Sci* 2003;6(3):165–174.
 43. Harper P, Jones S. Mathematical models for the early detection and treatment of colorectal cancer. *Health Care Manag Sci* 2005;8(2):101–109.
 44. Yaesoubi R, Roberts S. How much is a health insurer willing to pay for colorectal cancer screening tests? Winter Simulation Conference; 2008. pp. 1624–1631.
 45. Erenay FS, Alagoz O. Optimal policies for colorectal cancer screening. Working paper, also presented at INFORMS 2008 Annual Meeting; Washington (DC); 2008.
 46. Myers ER, McCrory DC, Nanda K, *et al.* Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. *Am J Epidemiol* 2000;151(12):1158–1171.
 47. Goldie S, Kim J, Wright T. Cost-effectiveness of human papillomavirus DNA testing for cervical cancer screening in women aged 30 years or more. *Obstet Gynecol* 2004;103(4):619–631.
 48. Kulasingam S, Benard S, Barnabas R, *et al.* Adding a quadrivalent human papillomavirus vaccine to the UK cervical cancer screening programme: a cost-effectiveness analysis. *Cost Eff Resour Alloc* 2008;6(4):1–11.
 49. Pinsky P. An early-and late-stage convolution model for disease natural history. *Biometrics* 2004;60(1):191–198.
 50. Tsodikov A, Szabo A, Wegelin J. A population model of prostate cancer incidence. *Stat Med* 2006;25(16):2846–2866.
 51. Kent DL, Shachter R, Sox HC. Efficient scheduling of cystoscopies in monitoring for recurrent bladder cancer. *Med Decis Making* 1989;9(1):26–37.
 52. Havrilesky LJ, Sanders GD, Kulasingam S, *et al.* Reducing ovarian cancer mortality through screening: is it possible, and can we afford it? *Gynecol Oncol* 2008;111(2):179–187.
 53. Davies R, Crabbe D, Roderick P, *et al.* A simulation to evaluate screening for helicobacter pylori infection in the prevention of peptic ulcers and gastric cancers. *Health Care Manag Sci* 2002;5(4):249–258.
 54. Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika* 1969;56(3):601–614.
 55. Kirch RLA, Klein M. Surveillance schedules for medical examinations. *Manage Sci* 1974;20(10):1403–1409.

56. Blumenson LE. Detection of disease with periodic screening: transient analysis and application to mammography examination. *Math Biosci* 1977;33:73–106.
57. Shwartz M. A mathematical model used to analyze breast cancer screening strategies. *Oper Res* 1978;26(6):937–955.
58. Eddy DM. Screening for cancer: theory, analysis, and design. Englewood cliffs (NJ): Prentice Hall; 1980.
59. Voelker J, Pierskalla W. Test selection for a mass screening program. *Nav Res Log Q* 1980;27:43–56.
60. van Oortmarssen GJ, Habbema JD, van der Maas PJ, *et al.* A model for breast cancer screening. *Cancer* 1990;66(7):1601–1612.
61. Houshyar A, Al-Khayyal F. A mathematical model for scheduling screening tests for progressive diseases. *Socioecon Plann Sci* 1990;24(3):187–197.
62. Chhatwal JC, Alagoz O, Burnside EB. Optimal policies for biopsy decision-making based on mammography findings and demographic factors. Working paper, also presented at INFORMS 2008 Annual Meeting; Washington (DC); 2008.
63. Eddy DM. A mathematical model for timing repeated medical tests. *Med Decis Making* 1983;3(1):45–62.
64. Eddy DM, Nugent FW, Eddy JF, *et al.* Screening for colorectal cancer in a high-risk population: results of a mathematical model. *Gastroenterology* 1987;92(3):682–692.
65. Wagner J, Tunis S, Brown M, *et al.* Prevention and early detection of colorectal cancer. The cost-effectiveness of colorectal cancer screening in average-risk adults. London: Saunders; 1996. pp. 321–356.
66. Lincoln T, Weiss GH. A statistical evaluation of recurrent medical examinations. *Oper Res* 1964;12(2):187–205.
67. Eddy DM. The frequency of cervical cancer screening. Comparison of a mathematical model with empirical data. *Cancer* 1987;60(5):1117–1122.
68. Habbema J, Lubbe J, van Oortmarssen G, *et al.* A simulation approach to cost-effectiveness and cost-benefit calculations of screening for the early detection of disease. *Eur J Oper Res* 1987;29(2):159–166.
69. Boyd A, Davies L, Bagust A. Modelling the implications for hospital services of cervical cytology screening: a case history. *J Oper Res Soc* 1989;40(6):529–537.
70. Chen HH, Kuo HS, Yen MF, *et al.* Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics* 2000;56(1):167–172.
71. Fontana R, Sanderson D, Woolner L, *et al.* Lung cancer screening: the Mayo program. *J Occup Environ Med* 1986;28(8):746–750.
72. Roberts S, Wang L, Klein R, *et al.* Development of a simulation model of colorectal cancer. *ACM Trans Model Comput Simul* 2007;18(1):1–30.

OPERATIONS RESEARCH SOCIETY OF CHINA*

YA-XIANG YUAN

State Key Laboratory of
Scientific/Engineering Computing,
Institute of Computational
Mathematics and Scientific/Engineering
Computing, Academy of Mathematics
and Systems Science, Chinese Academy
of Sciences, Beijing, China

XIANG-SUN ZHANG

DEGANG LIU
Academy of Mathematics and Systems
Sciences, Chinese Academy of Sciences,
Beijing, China

HISTORY OF THE OR SOCIETY IN CHINA

Modern OR activities in China were initiated in the late 1950s as the first OR group was founded at the Institute of Mechanics, Chinese Academy of Sciences (CAS) in 1956, promoted by Profs H.S. Tsien and K.C. Hsu. Tsien received his Master's degree from MIT and his Ph.D. from the California Institute of Technology, and went on to become Cal Tech's first Goddard Professor. Hsu earned his Ph.D. at the University of Kansas, and was a research associate in the Institute of Fluid Mechanics and Applied Mathematics at the University of Maryland. Tsien and Hsu returned to China in 1955.

By 1959, a second OR division was set up at the Institute of Mathematics, CAS. These two divisions merged into one research department in the Institute of Mathematics in 1960. The main research interests at that time were queueing theory, nonlinear programming, and graph theory along with transportation theory, dynamic programming, quality control, and economic analysis.

*Part of the historical review of ORSC is from an article by Xiang-Sun Zhang and Kan Cheng, "China: Ancient Country Makes OR History," *OR/MS Today*, October 1998.

In 1963, the division organized a series of specialized courses in OR at the Chinese University of Science and Technology, marking the first time in China that a university offered a systematic education in OR. Today, OR courses are a basic requirement in the business schools and engineering departments of almost every university in China.

Early OR application activities (late 1950s) in China focused on transportation problems. One such example was the allocation design of threshing grounds to save time and manpower. The classic Chinese postman problem model was presented by Gwan [1] in the same period.

One of the early highlights of OR application activities in China was conducted by the Chinese mathematician L.G. Hua, a member of CAS and president of the Chinese Society of Mathematics, during the so-called *Great Cultural Revolution*. Since theoretical research had been stopped at that time, Hua formed his own small group and turned to teaching basic optimization techniques at countryside factories in an effort to help managers and engineers apply OR in their daily operations. For a decade, beginning in 1965, Hua visited more than 20 provinces and innumerable towns and cities with his group and planted the seeds of OR principles and philosophies wherever he went [2]. His work promoted the development of OR in China then and continues to this day, long after his death.

The Operations Research Society of China (ORSC) was founded after the Great Cultural Revolution in 1980 when the First National Conference of OR was held in Shandong Province (on the east coast). Hua was elected the first president of ORSC. ORSC became a member of IFORS in 1982. Since the 1990s, ORSC has been active in the development of the Association of Asia-Pacific Operational Research Societies (APORS) within IFORS. As the president of APORS from 1991 to 1994, Professor Guanghui Hsu (ORSC president 1988–1996) organized the second conference of APORS in Beijing in the fall of 1991. As a representative of APORS, Hsu also served

as vice president of IFORS from 1992 to 1994.

As the most important events, ORSC convenes quadrennial national meetings, in which new president and its committee members, as well as officers, are elected for a term of four years. The most recent quadrennial meeting was held in 2008 in Nanjing and attended by over 450 participants. The national congress has traditionally been convened every four years, supplemented by a biannual academic conference. The Congress featured plenary talks, paper presentations, the ORSC election, prize awarding, and committee meetings. ORSC2008 awarded, for the first time, the Science and Technology Award of ORSC to Professor Minyi Yue, who was the president of ORSC from 1984 to 1988. It also marketed the first attendance of an IFORS president in the event. Other guests invited in the past include Sun Joo Park from Operations Research Society of Korea, Toshiharu Hasegawa and Masanori Fushimi from Operations Research Society of Japan (ORSJ) as well as Charles Larson and Mark Daskin from INFORMS.

It was ORSC's pleasure to host the IFORS 15th Triennial Conference in Beijing, since it was an opportunity for Chinese OR workers to learn about the new developments and trends of OR in other countries, especially in the developed countries. Since IFORS99 in Beijing, ORSC events have become more international and attracted more young operations researchers. The majority of its more than 1300 members come from universities and research institutes. ORSC plans to play an active part in getting OR applied in industry and be relevant to the rapidly developing economy of China. A step in this direction can be seen in the election into the ORSC council of more representatives from the consulting and financial arenas.

The ORSC now has 1350 qualified members and a dozen chapters located throughout the vast country. To obtain a membership, one just needs to send a request to the ORSC secretariat (orssc@amt.ac.cn), subject to annual membership fee RMB 40 Yuan (or US\$6).

Academically, there are 13 special committees (or subsocieties) under ORSC:

- Mathematical Programming

- Queueing Theory and Applications
- Reliability Theory
- Decision Making Theory and Applications
- Uncertainty Systems
- Intelligence Computing
- Scheduling Theory
- Graph Theories and Combinatorics
- Fuzzy Information and Engineering
- Game Theory and Applications
- Financial Engineering and financial risk management
- OR for Enterprise
- Computational Systems Biology

They organize symposiums every year or every other year during the interval of the two national conferences. Many of the committees and local branches present prizes recognizing outstanding work.

To promote cooperation with OR colleagues in the Asia-Pacific area and the rest of the world, the Asia-Pacific Operations Research Center (APORC) was set up in 1995 in Beijing. The center is affiliated with the CAS and APORS. APORC has already organized eight symposiums titled "International Symposium on Operations Research and Applications in Engineering, Technology, and Management (ISORA)." The upcoming ninth ISORA2010 is to be held in Jiuzhaigou, a world-famous natural reserve in China's south western Sichuan province. These symposium proceedings are published in the series of "Lecture Notes in Operations Research."

STATUS OF ACADEMIC RESEARCH AND APPLICATION OF OR IN CHINA

During the 1980s, many universities and colleges started to set up OR courses in their engineering schools, business schools or economic management schools. Linear programming, nonlinear programming, networks, queueing theory, reliability theory, inventory, and other major OR techniques are all taught at various levels. OR degrees (MS and PhD) are offered in some universities and institutes such as Tsinghua University, Chinese University of Science

and Technology, Fudan University, and some institutes in CAS such as Institute of Applied Mathematics and Institute of Systems Science.

ORSC has two official journals: *OR Transactions*, a quarterly journal that publishes original articles in OR and related fields (papers are published either in Chinese or English); and *OR and MS* (in Chinese only), a bimonthly aims at OR practitioners. Two journals sponsored by CAS—*Acta Mathematicae Applicatae Sinica* and *System Science and Mathematics*—also accept OR-related papers.

There are two major avenues for OR researchers in China to obtain grants. The first is through the National Natural Science Foundation of China (NSFC), which has given about one-third of its total funds for mathematicians to researchers involved in OR mathematics. Researchers can also apply for grants from other branches of the NSFC such as Management Science, Information Science and Engineering.

The second means to obtain grants is by providing consulting services to the public sector such as the central government, local government, industrial sectors, and so on. In these applications, traditional OR models and algorithms are built into software, and sometimes a complete management information system or a decision support system with OR-type philosophy and algorithms at its core are provided to the user.

One such example, a “Project Evaluation System in the State Economic Information System of China,” by Profs X.S. Zhang and J.C. Cui from the OR Division of Institute of Applied Mathematics, CAS, [3] won the IFORS Prize for OR in Developing Countries at the 14th Triennial Conference in Vancouver, Canada. The core part of the project is a reverse model of Data Envelopment Analysis (DEA).

Most OR activities in China have their foothold in industrial and agricultural applications, such as applications in petrochemical enterprises, manufacture engineering, and structure design of products. Some OR practitioners have focused their attention on supporting high-level

government decision makers in their strategic planning about the country’s continued advance in economy and ecology. The project, “Optimization of the Ecological Economic Development for the Upper Reaches of the Yangtse River,” by Prof. G.Z. Liu and his colleagues from Chengdu University of Science and Technology, was awarded the second place prize in the IFORS Prize for OR in Developing Countries competition. Another OR research project which had great social influence in China is “Grain Output Forecasting” conducted by Prof. X.K. Chen from the Institute of Systems Science, CAS [4]. With an input–occupancy–output technique at its core, the project has greatly increased the accuracy of the national grain output forecast. The group and its achievement received accolades from top government officials as well as CAS, which awarded the group its highest prize. Remarkably, Profs Yaxiang Yuan and Yuhong Dai’s recent research results [5–9] on nonlinear optimization theories and computational methods won the runner-up prize of the China National Natural Science Prize, which is one of the top scientific achievements recognized by the Government of China. Recently, Shouyang Wang and his research group [10] show that OR techniques play an important role in China’s reaction to the world financial crises.

2010 marks the 30th anniversary of the OR Society of China. The society is going to have a national meeting ORSC2010 in October in Beijing. Varied events have been planned to celebrate the rapid development of OR researches and applications in China over the past thirty years. ORSC is going to host APORS2011 in China’s Xi’an, a return of APORS triennial conference back to China after the second conference in Beijing in 1991. Chinese OR community is prepared to welcome colleagues from around the world as they are making ever greater efforts to contribute to the rapid growing economy.

ORSC PRESIDENTS AND THEIR TERMS OF OFFICE

Prof. Loo-Keng Hua, 1980–1984

Prof. Minyi Yue, 1984–1988

Prof. Guanghui Hsu, 1988–1996
 Prof. Xiang-Sun Zhang, 1996–2004
 Prof. Yaxiang Yuan, 2004–2012

REFERENCES

1. Gwan MK. Graphic programming using odd or even points. *Acta Math Sin* 1962;1:273–277.
2. Hua LK. Optimal selection techniques (in Chinese). Science Press 1981.
3. Zhang XS, Cui JC. A project evaluation system in the state economic information system of china—An operations research practice in public sectors. *Int Transactions Oper Res* 1999;6:441–452.
4. Chen XK, Guo J, Yang C. Yearly grain output prediction in China 1980-2004. *Econ Sys Res* 2008;20(2):139–150.
5. Yuan YX. On a subproblem of trust region algorithms for constrained optimization. *Math Program* 1990;47:53–63.
6. Yuan YX. On the convergence of a new trust region algorithm. *Numer Math* 1995;70:515–539.
7. Dai YH, Yuan YX. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM J Optim* 1999;10:177–182.
8. Dai YH, Yuan YX. *Nonlinear Conjugate Gradient Methods* (in Chinese). Shanghai: Shanghai Science and Technology Press; 2000.
9. Yuan YX. On the truncated conjugate gradient method. *Math Prog* 2000;87:561–571.
10. Wang SY. *et al.* Fighting China's Financial Crisis with OR/MS Today. 2010;27(2):Available online.

OPERATIONS RESEARCH SOCIETY OF NEW ZEALAND

ANDREW MASON
Operations Research Society of
New Zealand, Auckland,
New Zealand

The Operations Research Society of New Zealand (ORSNZ) was created on 24 May, 1965 when the Society was legally incorporated at the New Zealand Companies office. The Society was founded by staff at the Applied Maths Division of the Department of Scientific and Industrial Research (DSIR), a government department based primarily in Wellington, New Zealand. Known originally as the Operational Research Society of New Zealand, the “Operational Research” was changed to “Operations Research” in 2009 to reflect the now more popular term in New Zealand. The Society’s logo is shown in Fig. 1.

GOALS AND MEMBERSHIP

Formally registered as a charity in 2008, the ORSNZ is a nationwide not-for-profit organization with approximately 150 members including public servants, academics, students, consultants, and practitioners in industry. The Society’s membership also includes several companies active in operations research. The primary aim of the Society is to promote operations research and management science in New Zealand in both academic and industrial aspects. The ORSNZ is affiliated to the Association of Asia Pacific Operational Research Societies (APORS), the International Federation of Operational Research Societies (IFORS), and the Royal Society of New Zealand (RSNZ).

The Society’s activities include publishing a regular newsletter, organizing branch meetings in Auckland, Wellington, and Christchurch, hosting visitors to New Zealand, maintaining the web site (www.orsnz.org.nz), and organizing an

annual conference. The Society published the New Zealand Operational Research (NZOR) journal from September 1971 until July 1986 when NZOR merged with the Asia Pacific Journal of Operational Research (APJOR).

ANNUAL CONFERENCE AND VISITOR PROGRAM

The Society runs a two-day conference in late November or early December each year. The 40th such conference was held in December 2005 at the Society’s birthplace, Victoria University, Wellington. The conference location rotates between the University of Auckland, the University of Canterbury in Christchurch, and the University of Victoria in Wellington. Proceedings of the conference are published and posted to members, and also made available on-line on the Society’s web site.

A highlight of the conference is the Young Practitioner Prize (YPP) awarded for the best lecture given by a presenter under 30 years of age. Students and recent graduates are supported and encouraged to attend conferences by free registration, generous prize money for the YPP, and grants to cover travel expenses.

The Society encourages and financially supports international visitors to give plenary addresses at the conference. Recent plenary presenters include Dorit S. Hochbaum (USA) in 2009, Craig MacLeod (NZ) and Anita Schöbel (Germany) in 2008, and Michael Trick (USA) and Gerard P. Cachon (USA) in 2007. In 1985, the Society was honored to welcome George Dantzig as a guest at its 21st conference celebrations.

To further support international visitors, the ORSNZ invites nominations for ORSNZ honorary visiting lecturers to visit New Zealand. Each visiting lecturer is expected to give a lecture on some topic likely to be of general interest to ORSNZ members at three or more locations among Auckland, Hamilton, Wellington, and Christchurch. The ORSNZ contributes to the travel costs for



Figure 1. The ORSNZ logo.

this visit. Recent visitors include Ted Ralphs from Lehigh University, Michael Trick from Carnegie Mellon University, and Anita Schöbel from the University of Göttingen.

DAELLENBACH PRIZE

To honor the considerable contributions of Emeritus Professor Hans Daellenbach to operations research/management science in New Zealand, the ORSNZ established the ORSNZ Hans Daellenbach Prize in 2001. Professor Daellenbach's contributions reflect his belief that the best work in operations research combines strong innovative methodology with practical impact. The Daellenbach Prize is awarded for a body of work that has made a significant contribution and received international recognition. The inaugural prize was awarded to Professor David Ryan, University of Auckland, in 2001, for his work on scheduling problems and their application to crew scheduling and other problems at Air New Zealand. In 2003, the award went to Professor Les Foulds, University of Waikato, while in 2006 Professor Andy Philpott won the award for contributions ranging from yacht design with Team New Zealand, through telecommunications network design to the analysis of electricity markets and the optimization of electricity generation

Table 1. Recent Presidents of the ORSNZ.

1994–1996	Jonathon Lermitt, Transpower, Wellington
1996–2000	Andy Philpott, University of Auckland
2000–2003	Les Foulds, University of Waikato
2003–2007	David Ryan, University of Auckland
2007–	Andrew Mason, University of Auckland

and transmission. Perhaps as a reflection of the strong emphasis placed in New Zealand on using operations research to make a practical difference, Professors David Ryan and Andy Philpott have both been finalists in the prestigious international INFORMS Edelman prize, as has fellow ORSNZ member Professor Mikael Rönnqvist.

ORSNZ MANAGEMENT

The Society is run by an elected Council comprising a President, a Vice President, an Honorary Secretary, an Honorary Treasurer, and six other members. Recent presidents of the Society are listed in Table 1.

In accordance with legal requirements, the Society holds Annual General Meetings to present its accounts to members and discuss business and other matters. These meetings typically take place during the annual conference.

Further information on the ORSNZ can be found on the Society's web site, <http://www.orsnz.org.nz>. The author acknowledges this site as a source for much of the information included in this article.

OPERATIONS RESEARCH SOCIETY OF TAIWAN

HSING LUH
Department of Mathematical
Sciences, National Chengchi
University, Taipei City,
Taiwan

SUMMARY

Before starting to introduce the Operations Research Society of Taiwan, a short summary of Taiwan's economic and educational background is given for a better glance of its industrial and academic environment. Taiwan includes a population of 23,046,177 people as registered in 2009. The capital city is Taipei, which is located off the coast of mainland China. Real growth in GDP has averaged about 8% during the past three decades. Exports have provided the primary forward motion for industrialization. The trade surplus is substantial, and foreign reserves are the world's fifth largest as of December 31, 2007. Leading technologies of Taiwan include: bicycle manufacturing, biotechnology, semiconductor device fabrication, laptops, and smart phones; for example, Giant Bicycles, Merida, Acer, Asus, BenQ, HTC, and so on [1].

During the past 20 years, Taiwan has become a major foreign investor in mainland China, Thailand, Indonesia, the Philippines, Malaysia, and Vietnam. It is estimated that some 50,000 Taiwanese businesses and 1 million business people and their dependents are established in the People's Republic of China (PRC) [2]. Growth averaged more than 4% in the 2002–2006 period and the unemployment rate fell below 4%. Since the global financial crisis starting with United States in 2007, unemployment rate in Taiwan has risen to over 5.9% and economic growth has fallen to –2.9% [3]. As of December 2008, there were 1,337,455 students enrolled in

167 colleges (universities): 17% studying the humanities, 36% social science, and 47% science and technology. Of the more than 6000 academic departments or programs, almost 28% of students were studying technology and engineering; 2% mathematics- and statistics-related areas; and 13% general management [4].

For a long time, the methodologies of operations research (OR) have been adopted as analysis tools for production efficiency in business and industry in Taiwan. In a survey by Kuo and Liu [5], it was found that there were 59.30% of the methodologies were used in production, 57.56% in finance, 48.26% in marketing, 35.47% in human resource, and 53.49% in information. The specific OR subjects involved were linear programming (28.49%), network analysis (34.30%), other mathematical programming (18.02%), queueing models (10.47%), Markovian decision processes (4.65%), other probability models (40.12%), decision theory (27.91%), simulation (30.23%), statistical forecasting (64.53%), and in decision analysis with computers (70.93%).

THE SOCIETY

Because of its high demand in education and practice, setting up of an OR society was proposed in 1996 when many OR professionals, including professors, engineers, and managers called for an exclusive professional society. As per the law regarding professional societies in Taiwan, numerous meetings were initiated by the founding members; these took place in Taipei, Hsin-chu, Taichung, and Kaohsiung. Finally, two formal meetings were held at the Department of Mathematical Sciences in National Chengchi University, Taipei and Department of Industrial Engineering and Enterprise Information in Tunghai University, Taichung on April 26 and July 12, 2003, respectively. The testimonial events generated an enthusiastic exchange of ideas and culminated in the formation of a

working committee whose task was to define the organization that was to be the Operations Research Society of Taiwan (ORSTW), which is also responsible for the INFORMS Taiwan Chapter (INFORMS TWC) with the same committee members. On September 13, 2003, the committee presented the constitution and by-laws for approval by the first founding meeting of some 150 members, and the ORSTW formally came into being. Finally, ORSTW/INFORMS TWC has been officially at the forefront of promoting the advancement and practice of OR in Taiwan.

The officers of the society include the president, secretary, treasurer, and executive board. The terms of the president, secretary, treasurer, and board of executives is two years from the time of election. Up to now, the presidents and their terms of office are Dr. Ping Teng Chang, Professor at Department of Industrial Engineering and Enterprise Information, Tunghai University, 2003–2005; Dr. Hsing Paul Luh, Professor at Department of Mathematical Sciences, National Chengchi University, 2005–2007; Dr. Sy-Ming Guu at College of Management, Yuan Ze University, 2007–2009.

The logo of ORSTW depicted in Fig. 1 was presented and approved unanimously in the first annual meeting while its website was given at www.orstw.org.tw for advertisement and communication. Besides local and foreign experts who have presented papers featuring OR applications in various sectors, Ph.D. and Master's thesis competitions and working project competitions have been initiated for college students. This has become a regular event of the society in its drive to focus on the vast potential of OR in Taiwan. At about the same time, the call for papers for the first issue of the International Journal of Operations Research (IJOR) was brought out.

IJOR reports on developments in OR, information systems, and management sciences, including the latest research results and applications. Its aim is to meet the needs of accelerating technological change and changes in the global economy. It is available in abstracting/indexing services including Mathematical Reviews, EBSCOhost, Index of Information Systems Journals, CSA's

Technology Research Database, CSA's Engineering Research Database, CSA's Industrial and Manufacturing Engineering Database, Zentralblatt MATH, INSPEC. Twenty issues have been published since 2004 with both a regular print version, ISSN: 1813–713X, and an on-line version, ISSN: 1813–7148 [6].

Each year, the ORSTW's general membership meeting serves as a forum to discuss pertinent issues and new directions. Both academic and professional development programs are scheduled for its members through a paper presentation and project competition on issues such as traffic, agriculture, environment, energy, supply-chain management, and business logistics. Public service was also established to move closer to the long-term mission of helping in nation-building through the use of OR. Several projects have been incorporated with semiconductor manufacturing, transportation, and financial services. Workshops of OR methodologies have been organized with European and American universities to provide short-term or summer programs for Taiwanese college students as well as teachers to update their OR knowledge for future research and development. Since 2003, the membership has been growing by 15% due to the aggressive promotion and attractive activities.

The ORSTW's executive board was founded by 20 members and consists of five members in the board of supervisors; they are selected among all members biannually in the general membership meeting. The mission of the executive board includes the following:

- *Offer More Products and Services to the Practice Community.* ORSTW looks



Figure 1. The logo of the Operations Research Society of Taiwan.

across the whole set of interrelationships in the field while strengthening as well as balancing both sides between academia and industry. If ORSTW develops only strong academics without any link to the real-world, that would make no difference with pure theory analysis, such as mathematics. To understand the whole value chain, ORSTW needs to pay more attention to real-world problems. Therefore, in the annual ORSTW conference each year, more than half of the contributed sessions are reserved for practitioners. Besides, the committee encourages joint professional activities with Taiwanese enterprises by providing more management consultancy and services for them.

- *Boost Worldwide Recognition for the Profession.* It has tremendous opportunities to identify new practitioners and practice-related activities that have not necessarily been in the OR field in the past. ORSTW feels there are more possibilities that should be explored in those areas. For example, because of the emergent internet economy, finding more ways to put cutting-edge research into use is challenging and promising.
- *Identify New Initiatives to Take the Profession and the Organization Forward.* To appreciate the value chain created by OR professions, ORSTW needs to create a virtuous circle which not only brings new people and programs into ORSTW, but also makes the people and programs we already have even better. Thus, ORSTW is responsible for initiating any possible force and providing the momentum to take it forward.
- *Launch New Publications.* It is a very long and difficult process to start a new academic journal, but IJOR has been successfully published as a significant example. It needs to start thinking about new publications and journals and how ORSTW delivers them. To increase the presence of journals and services in China, India, and other emerging economies, ORSTW should

make it available for edited book volumes or on-line compilations.

- *Increase Membership.* ORSTW should have provided tremendous value to our academic members because of regularly published IJOR and annual conferences. But it is not enough. Many people doing things called “analytics” or some other name have not traditionally been allied with OR and management science. ORSTW should offer them the services, products, networking opportunities, and other things that they value and help them to be more effective professionally, so that we can really contribute to the growth of this organization. ORSTW should bring new membership for circulating the process to achieve better.
- *Improve Governance.* ORSTW keeps working to improve mechanisms to deal with building a strong organization. ORSTW receives enormous help from volunteers every year. Teachers, students, and practitioners have made great contributions toward planning, prioritizing and steering the organization. ORSTW should enhance strengths by identifying areas for growth and new opportunities.
- *International awareness.* ORSTW is working collaboratively with OR societies all over the world. ORSTW members have been actively involved in meetings among INFORMS, Asian Pacific Operational Research Societies (APORS), International Federation of Operational Research Societies (IFORS) and so on. Joint conferences with transportation, science and technology, service and military, and many other professional societies have taken place with mutual research and interest. Societies from Japan, Korea, Hong Kong, USA, and others have worked closely with ORSTW for general conferences. In December 2007, the ORSTW hosted its first group of Chinese OR scholars from mainland China at the fourth annual meeting.

THE FUTURE

Information technology is transforming the world. There are tremendous opportunities out there for OR in both education and industry. ORSTW and INFORMSTWC will work together to provide more services for OR professions, as well as academia and industry. With its rich past and strong support, ORSTW indeed has a lot to look forward to for its promising development.

REFERENCES

1. Available at <http://en.wikipedia.org/wiki/Taiwan>. Accessed 2009.
2. American Chamber of Commerce in Taipei. Convenience stores aim at differentiation. Taiwan Bus TOPICS 2009;34(11). Available at <http://www.amcham.com.tw/content/blogcategory/135/346/>.
3. Directorate-General of Budget, Accounting and statistics. Executive Yuan, R.O.C.(Taiwan). Available at <http://eng.dgbas.gov.tw/mp.asp?mp=2>. Accessed 2009.
4. Ministry of Education, R.O.C. (Taiwan). Available at <http://english.moe.gov.tw/ct.asp?xItem=552&CtNode=415&mp=1>. Accessed 2009.
5. Kao C, Liu S-T. Operation research application in Taiwan: a linguistic approach. Eur J Oper Res 1997;103:628–634.
6. Int J Oper Res. Available at <http://www.orstw.org.tw/IJOR/index.html>. Accessed 2009.

OPERATIONS RESEARCH TO IMPROVE DISASTER SUPPLY CHAIN MANAGEMENT

OZLEM ERGUN
GONCA KARAKUS
PINAR KESKINOCAK
JULIE SWANN
MONICA VILLARREAL
H. Milton Stewart School of
Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta,
Georgia

INTRODUCTION

From improving the performance of fire and police departments, optimizing the public transportation system, programming delivery of blood to hospitals, planning housing projects, and analyzing drug policies to improving delivery of meals to senior citizens, there are many examples of how operations research (OR) addresses community needs [1]. Public sector OR deals with solving public interest problems through the application of analytical tools. One of the public sector OR application areas is humanitarian OR, which particularly deals with the problem of delivering relief to people affected by a disaster. Hence, natural and man-made disasters are the most common subject of attention for humanitarian OR.

There were 6637 natural disasters between 1974 and 2003 worldwide, with more than 5.1 billion affected people, more than 182 million homeless, more than 2 million deaths, and with a reported damage of US\$1.38 trillion [2]. In 2005 alone, over 180,000 deaths and US\$200 billion economic losses have occurred due to disasters according to the Disaster Resource Network Humanitarian Relief Initiative [3]. The September 11 attacks (2001), tsunami in South Asia (2004), Hurricane Katrina (2005),

and earthquakes in Pakistan (2005) and Java (2006) are just some examples of the deadliest disasters witnessed by humankind in the past few years.

The consequences of these events are enormous, not only in the short term with injuries, loss of life, and damaged infrastructure, but also in the long term with changes in social and economic conditions. Even though the occurrence of these events could not have been avoided, the impact could have been reduced by different means including humanitarian OR. For example, adequate warning systems could have prevented injuries and fatalities during the 2004 tsunami, and help could not reach Pakistani earthquake victims because of logistics difficulties with limited infrastructure. Humanitarian OR differs from other OR applications because it deals with particularly unique and highly variable events, often in resource-poor and limited-infrastructure environments, with multiple organizations trying to work together in response activities simultaneously. These factors increase the complexity of responding to these events.

The focus of this article is on supply chain-related issues in humanitarian operations, with a greater focus on “disasters” rather than ongoing conditions. We discuss differences between regular supply chains and supply chains used for disaster planning and response, that is, “humanitarian supply chains”. First, we describe characteristics of supply and demand of disaster supply chains, followed by a discussion on the particularities of the execution and management of these supply chains. Next, an application of OR in humanitarian operations is examined in more detail, and finally main challenges and future research directions are presented.

OVERVIEW OF DISASTERS

Disasters can be grouped into two main categories: natural and man made.

Table 1. Disaster Categorization and Examples

	Man-made	Natural
Slow onset	Political crisis, refugee crisis	Famine, drought
Sudden onset	Terrorist attacks, chemical leaks	Hurricanes, floods, earthquakes, tsunamis

Natural disasters are the consequences of natural hazards that affect people, whereas man-made disasters are caused by human actions. A more detailed categorization of disasters is shown in Table 1 along with examples. Also, disasters could be categorized in predictable timing (or seasonal) such as floods or unpredictable timing like earthquakes; and predictable location such as hurricanes or unpredictable locations like tsunamis.

No matter what the type of disaster, the management of these events typically follows four sequential stages: mitigation, preparedness, response, and recovery. Mitigation is the application of actions to help prevent or reduce the hazards of the disaster. It differs from the other stages because it focuses on long-term measures for reducing or eliminating risk [4]. Preparedness activities help prepare for response once a disaster occurs. The response phase covers activities for mobilizing emergency responders and services for the affected region. Recovery is the stabilization phase during which restoration of the disaster area is conducted

in the long term. The disaster management timeline with the related operations can be seen in Fig. 1.

There has been some research on the use of OR techniques in disaster management. Altay and Green [5] survey the literature and summarize the research in this area. Most of the papers published deal with man-made disasters (47.6%) or general disaster operations management (40.5%). Among the papers reviewed, 61.1% are on the pre disaster phase (mitigation and preparedness) and mostly based on risk analysis, 23.9% are on response, and only 11% are about recovery. More than half of the research is based on model development, followed by theory and application development. Most of the application studies are made for the disaster phase, whereas the theory papers mostly focus on the pre-disaster phase.

SUPPLY

A primary role of a supplier in a supply chain is to source the required items downstream. The main sources in a humanitarian supply chain are vendors and donors. Vendors can be local to the region where the disaster occurred, or global. Donors are the sources of donations of any type (financial, products, services, etc.).

Supplies consist of relief items, personnel/volunteers, and transportation and construction resources, among others. Most of the supplies fall into the relief items

Predisaster	Disaster	Postdisaster
Mitigation and preparedness	Response	Recovery
Assessment Risk factors Vulnerability	Relief operations: First phase Medics, food, and shelter	Debris cleaning Infrastructure rebuilding Re-establishing communities
Planning Infrastructure Policy making Capacity building Prepositioning resources	Second phase Housing Food supply chain building	Measure the effects of: Infrastructure Planning Response short and long term
Training/education	Logistics stages: Mobilization and procurement Long haul The last mile	Lessons learned, feedback to planning and response

Figure 1. Disaster timeline and operations.

category. Figure 2 shows a categorization for relief items, as well as some examples. Consumable relief items require recurrent delivery to the affected community, and non-consumable relief items require a one-time delivery only. Nonconsumable operational relief items are required to set up an operation, while nonoperational items are required to meet the essential needs of the affected population.

There are specific challenges related to supplies that come from in-kind donations. First, since the quantity and mix of the supplies depend at least to some degree on the donor, there is a high uncertainty of what is going to be received. Moreover, the timing of these supplies might not be appropriate: for example, consumables that arrive too early cannot be stored for a long time, or nonconsumables that arrive after the operation was set up are wasted. There are many cases in the recent history where donated items were not needed and were not deployed to people affected by the disaster. Autier *et al.* [6] discuss the case of drug supplies after the 1988 Armenian earthquake, when at least 5000 tons of drugs and consumable medical supplies were sent by international relief operations, but only 30% were immediately

usable (sorted, relevant for the emergency situation, and easy to identify), and 20% of these supplies had to be destroyed by the end of 1989. Unsuitable donations caused bottlenecks in the supply chain, making storage and transportation processes more inefficient.

Donations place additional complications on the procurement process, since it is difficult to define what will come from donors and what will have to be sourced from vendors. But even if donations are not considered, the procurement process is by itself a challenging task. Supply availability is highly dependent on the location. Organizations often have a low visibility into existing inventory. Control of inventory is usually given to country offices, resulting in excess supplies in some locations and scarcity in others. The selection of suppliers during the procurement process includes not only total cost versus quality, response time, and reliability trade-offs, but also considers less measurable factors like activating local economies by choosing a local provider. Developing contracts with suppliers is difficult given the uncertainty of the type, quantity and timing of items required, and the available budget. Also, there is competition for supply sources when there are

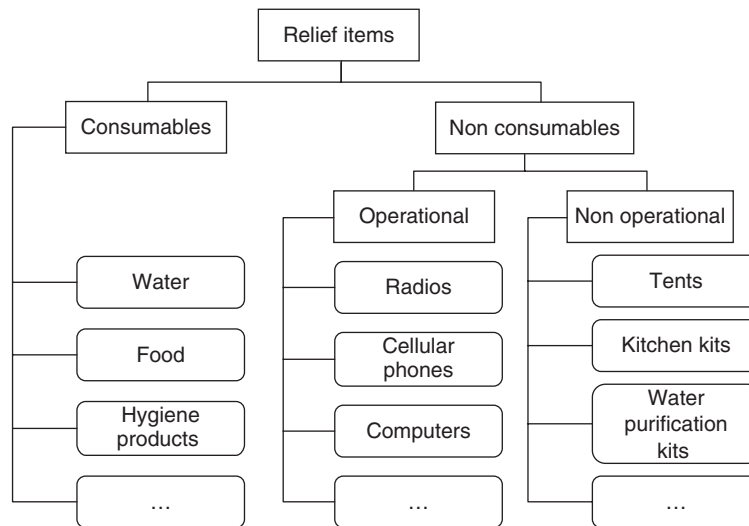


Figure 2. Relief supplies categorization and examples.

local or international nongovernmental (non-profit) organizations (NGOs) sharing similar relief objectives, and further there is lack of coordination among them.

Challenges faced while sourcing from donors or vendors after the event occurs may affect supply availability. The shortage of supplies may cause emergency response to be ineffective and result in increased human suffering [7]. Hence it is important to develop strategies to accelerate supply response or deal with unpredictability of demand. One strategic initiative that has been recently implemented by several humanitarian organizations is the prepositioning of inventory instead of procurement after the fact. Prepositioning allows not only faster response but also better procurement planning and an improvement on distribution costs; however, it requires an additional investment before the event occurs, and funds are more difficult to obtain.

In summary, the supply process in a humanitarian supply chain is different from regular supply chains. Supply in regular supply chains follows a standardized order fulfillment process, while in the case of disasters a portion of supplies comes from voluntary donations. Also, since humanitarian supply chains are often in resource-poor environments, supply may be especially variable. Greater certainty on the supply quantity, location, and time; longer relationships with suppliers; and information about the suppliers available at the location of interest before decisions are made allow better procurement contracts for normal supply chains. Finally, when common supply deals with routine events and not a one-time disaster event, developing a procurement “expertise” is easier.

DEMAND

The customers in a disaster supply chain include the population at the affected area, as well as intermediate customers at local or global storage facilities. Their needs change significantly according to disaster types and the phases in the disaster timeline. In the predisaster phase, protection-based items

like batteries and flashlights are highly demanded both by the people and local stores for preparation, while immediately after the disaster, the high demand changes to first-response items like drugs, medicine, food, water, and shelter. Long-term recovery items like infrastructure repair and construction equipment are among the items needed during the postdisaster period.

The demand patterns are also different in each phase. The predisaster stage consists of planning processes that are mostly based on forecasts, so the demand is not certain. Once a disaster hits, demand becomes complex: high and quickly changing, but more certain based on the reported needs that are sent eventually by the assessment team in the disaster area. During the postdisaster period, demand again stabilizes and becomes more predictable with the real data from the region. The overall demand structure of disasters is complicated and challenging because of the high unpredictability of its three main dimensions: time, location, and magnitude. Also, disaster demand has other drivers related to those dimensions such as population characteristics, economy, and political conditions; these factors are also complex to formulate.

Dependency of demand in disasters on these hard-to-measure factors and its high uncertainty are the main differences from the demand in regular supply chains. Unlike logisticians in the private sector, humanitarian workers are always faced with the unknown: when, where, what, how much, where from, and how many times; in short, the basic parameters needed for an efficient supply chain setup are highly uncertain [8]. Disaster demand forecasting is also difficult due to the lack of historical data. Even though there do exist some databases from the past experiences prepared by both NGOs and governments, such as the EMDAT (Emergency Events Database) by the Center of Research on the Epidemiology of Disasters, they are occasionally inadequate because of inconsistent and/or insufficient data collection and reporting problems [9]. Additionally, disasters are unique even if they occur in exactly the same location, since other factors such as population structure or economic conditions

could have changed since the previous occurrence. Hence, historical data is not always very useful for predicting future demand.

DISASTER SUPPLY CHAINS

Supply chains link the sources of “supply” (suppliers) to the owners of “demand” (end customers). In a typical humanitarian supply chain, governments and NGOs are the primary parties involved. Governments hold the main power with the control they have over political and economical conditions and directly affect supply chain processes with their decisions. After the 2004 tsunami, for instance, the Indian government did not invite international aid agencies to participate at all in the first 60 days of the relief effort, and functioned during that period with the local sources of supplies [10]. Donors, military, and the media are the other significant players in the humanitarian supply chains.

Disaster Supply Chain Management

Supply chain management is the process of planning and controlling the operations of the supply chain by coordinating the different parties involved in the process of fulfilling customer demand efficiently.

Coordination and management of disaster supply chains have challenging problems. The supply network is huge and complicated with numerous players (donors, NGOs, government, military, and suppliers), and it is hard to coordinate all of them along with all the items that need to be delivered. Despite the different cultural, political, geographical, and historical differences among them, collaboration and specialization of the tasks between NGOs, military, government, and private business are increasingly needed in the humanitarian supply chains [8]. Despite being experienced and aware of the key points in humanitarian supply chains, people in charge of logistics and supply chain management in most NGOs or other humanitarian organizations are not often specialized in this area, and therefore they are not experts in the tools for solving the problems that might occur during the operations. There could also

be domestic barriers such as the need of excessive paper work, specific policies of the region that may cause additional delays, as well as external complications due to foreign relations.

Use of technology is essential in managing supply chain operations and maintaining coordination, but the organizations usually do not use specific software or other technology tools for supply chain management. The case study written for El Salvador earthquake in 2001 [11] focuses on coordinating the logistics and supply chain operations using a humanitarian relief supplies management software called SUMA. SUMA’s objective is “to develop a standardized methodology and operational capacity at the national or regional level to manage relief supplies and equipment efficiently” [12]. The implementation of this software allowed the identification urgent needs, helped prevent unsolicited donations that could upset the system, and created reports with centralized information to inform the population about the development of the operations, building visibility and transparency through the supply chain.

Finally, goals and performance metrics of humanitarian and regular supply chains differ notably. Unlike the humanitarian supply chains, which do not have any profit targets and rely heavily on volunteers and donors, in regular supply chains, stakeholders are the “owners” of the chain. A company has to improve its profits and aims for making more money. It is easier to prove success using profit data in private supply chains, but how to prove success is not clear for humanitarian supply chains because cash flow data may not fully explain results. Nevertheless, the numerous models based on minimizing cost (or equivalently, maximizing profit) for building efficient supply chains can be applied to the humanitarian supply chains directly or with modifications. One example is an integrated, multiobjective supply chain model that uses fill rate, cost, and flexibility as measurement factors for simultaneous strategic and operational supply chain planning [13]. Lodree and Taskin [14] work on stochastic production/inventory

control models for recovery planning, specifically for hurricanes, to determine how long to postpone decision making to optimize the trade-off between logistics cost efficiency and hurricane forecast accuracy.

Disaster Supply Chain Execution

The execution or delivery process in a supply chain consists of bringing supply and demand together. Delivery could mean different things at each disaster stage, such as setting up temporary warehouses or shelters during the preevent stage or delivering relief to affected people during the response stage.

Distribution operations within a disaster framework are very challenging. Owing to the complexity of the sourcing process and the uncertainty of demand, there could be a big gap between supply and demand in terms of mix, quantity, timing, and location, and matching them becomes a hard task [15,16]. Handling expertise becomes more challenging due to the large span of relief items and the use of ad hoc warehouses. Transportation infrastructure is highly dependent on the location; it may be damaged or disrupted and suffer from dynamically changing conditions. Additionally, communication infrastructure may be disrupted. Because of the sudden onset of most disasters, problems associated with inadequate distribution planning are common, including high expediting costs, choice of wrong transportation mode and/or provider, bottlenecks in the port of entry, incomplete execution, and so on. Hence preparation during the preevent stage is vital, and strategies such as the use of staging areas for prepositioning and distribution of relief supplies [17] help overcome the difficulties in getting the right supplies to the right people, at the right time, at the right place.

Humanitarian supply chain delivery operations may have different targets than for-profit supply chains. Usually, there is a trade-off between cost and responsiveness of the delivery process. In the case of disaster supply chains, responsiveness concerns human welfare, and it has a higher priority compared to operational cost than regular supply chains. Since every human life is equally valuable, fairness is a particularly

important criterion for disaster supply chains, and their responsiveness should not be based to the social or economical condition of the affected people.

AN ILLUSTRATIVE EXAMPLE—DEBRIS MANAGEMENT OPERATIONS

Depending on the nature and severity of the disaster, and the characteristics of the affected area, there could be massive amounts of debris. The resources required to collect the debris might be limited, debris could be blocking roads and obstructing aid supply, and some types of waste can endanger community health and safety; therefore, resources have to be allocated adequately to speed up the debris collection and disposal process, while minimizing its impact. Debris collection becomes a complex logistics issue, and a debris management plan should be developed in advance addressing each disaster stage [18]. Figure 3 shows the main elements of such a plan, based on the guidelines of the Federal Emergency Management Agency of the United States [19]. While a management plan answers the question of what to do, OR answers the question of how to do it most efficiently and effectively.

During the preevent stage, forecasts are made to predict the quantities and location of debris. Debris forecasts are used to plan for the resources, to design debris management sites for temporary storage, and to develop adequate strategies for final debris disposal. An adequate debris forecast that considers the type and extent of the potential disaster and historical data is essential for effective preparedness. During the resource planning, debris collection authorities may realize that existing forces and equipment will not be sufficient and some services must be outsourced. Choosing both the right contract (unit price, lump sum, time and materials, etc.) and the right contractor is fundamental for controlling the costs, quality, and availability of the procured services. Finally, a facility location model (see *Integrated Supply Chain Design Models*) could be used to select among potential debris management

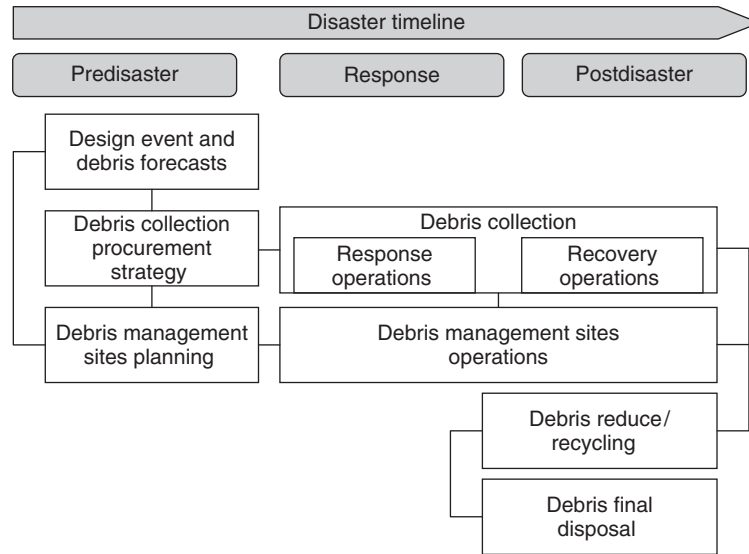


Figure 3. Elements of Debris Collection Management Plan.

sites such that the use of these locations expedites the collection process while minimizing additional costs.

During the response stage, having access to injured people and to critical facilities such as hospitals or police stations is crucial. Roads and streets that are essential for the emergency operations might be obstructed, and debris must be removed to facilitate rescue efforts. A network model can determine how resources should be allocated and which roads should be cleared first. Such a model takes into account the priority of the facilities to be connected in order to maximize the supplied relief as well as the available quantity of workforce and equipment.

The longest phase of the debris collection process occurs during the recovery stage of a disaster, once emergency and other major routes are cleared from debris. After the disaster, community residents begin to take debris to public rights-of-way, and this debris along with what was left after the response stage has to be collected. Debris such as white goods (containing refrigerants and other regulated machine fluids) and hazardous waste has to be handled with special care because of their environmental and health safety impact. Scheduling and

routing models for the debris collection resources can be developed to minimize the impact of the remaining debris while using the available resources efficiently to reduce costs. A model of this nature assigns each resource to loading tasks at the debris sites and unloading tasks at the debris management sites. After the debris is collected, it may go through some reduce or recycle processes, and then be taken to its final destination. Deciding the best policies to follow for each debris type is not a trivial problem when different constraints such as environmental regulations, available capacities, budget, and so on, are taken into account. A mathematical model can help the decision maker to find an optimal trade-off of cost and environmental impact.

CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The use of OR techniques for disaster preparedness and response can help improve the efficiency and effectiveness of processes extensively. The challenges defined in this article about the particular structure of disaster supply chains demonstrate their differences from regular supply chains. Since

disaster supply chain management is a newly emerging research area, there are many problems to be solved and new directions to be discovered for future research beyond the limited literature so far.

Most of the current research on disaster operations management focuses on the pre-event phase, which covers planning, mitigation, and preparedness processes. There is limited research in recovery planning phases. Furthermore, when a specific problem within a disaster context is modeled, the model usually focuses on one specific stage (preevent, response, postevent). However, there is an undeniable interaction between the decisions made at different stages; that is, what is done today affects what can be done tomorrow. Therefore, more comprehensive models integrating multiple disaster stages are needed.

One of the main characteristics of disaster supply chains is the presence of multiple stakeholders (NGOs, governments, local authorities, etc.). In this multifunctional environment, each of these players might have different objectives and priorities, leading to potential conflicts and inefficiencies in operations. Also, the best decision for each stakeholder alone is not the same when the entire supply chain is considered; that is, inefficiencies exist when individual incentives are not aligned. To handle the system as a whole, hierarchical planning, multiobjective models, centralized/decentralized system trade-offs, standard procedures, information sharing platforms, collaboration mechanisms, and incentive models need to be developed and analyzed.

When dealing with disasters, the human factor is always present. Even the best model that reflects the real conditions does not provide the perfect solution for disasters unless it takes human factors into account. Therefore, there is a need to incorporate behavioral aspects into the mathematical models. For example, in order to design an effective evacuation model, the model should be able to predict and incorporate the different reactions and attitudes of the people. There is a

limited amount of work connecting mathematical representation and behavioral dynamics [20], and this multidisciplinary approach has a high potential for more efficient solutions for disaster supply chains.

Finally, there is need for models that fully describe a disaster supply chain system within stochastic, dynamic, and adaptive settings. A stochastic model would be able to capture the high uncertainty in demand, supply, and delivery in these supply chains. Even though there is some work done modeling disaster- and emergency-related problems under uncertainty, many of the models in the literature are deterministic. Incorporating the dynamics of the choices made during each time period is crucial because of the strong interactions of these decisions. The proposed models should also be adaptive to incorporate new information when a change occurs. In a disaster framework, information is usually limited at the beginning, and as time passes, more and more accurate information becomes available. The model should capture and make use of the new pieces of information.

Disaster supply chain management is a rising area of OR, which has the potential to make a significant impact on the society. Defining the systems accurately and developing models that fit the structure of disasters are essential for applying OR techniques in disasters applications.

Acknowledgments

We gratefully acknowledge support of our work from the Harold R. and Mary Anne Nash Junior Faculty Endowment Fund and the Focused Research Program Grant from the College of Engineering at Georgia Tech.

REFERENCES

1. Kaplan E. Adventures in policy modeling! Operations research in the community and beyond. *Omega Int J Manage Sci* 2008;36:1–9.
2. Center of Research on the Epidemiology of Disasters (CRED). EMDAT: Emergency events database. Available at <http://www.emdat.be/>. Accessed 2009 Jan 15.

3. Disaster Resource Network Humanitarian Relief Initiative (HRI). Available at <http://www.weforum.org/en/initiatives/drn/index.htm>. Accessed 2009 Jan 15.
4. Haddow GD, Bullock JA, Coppola DP. Introduction to emergency management. Burlington (MA): Butterworth-Heinemann Homeland Security Series; 2008.
5. Altay N, Green WG. OR/MS research in disaster operations management. *Eur J Oper Res* 2006;175(1):475–493.
6. Autier P, Ferir MC, *et al.* Drug supply in the aftermath of the 1988 Armenian earthquake. *Lancet* 1990;335(8702):1388–1390.
7. Knott R. The logistics of bulk relief supplies. *Disasters* 1987;11(2):113–115.
8. Van Wassenhove LN. Blackett memorial lecture—Humanitarian aid logistics: supply chain management in high gear. *J Oper Res Soc* 2006;57(5):475–489.
9. Centre for Research on the Epidemiology of Disasters (CRED). Thirty years of natural disasters 1974–2003: the numbers. Available at http://www.emdat.be/Documents/Publications/publication_2004_emdat.pdf. Accessed 2009 Jan 15.
10. Thomas A, Fritz L. Disaster relief, Inc. *Harv Bus Rev* 2006;84(11):114–122.
11. Tomasini R, Van Wassenhove L. Coordinating disaster logistics after El Salvador's earthquakes using SUMA's humanitarian supply management system. INSEAD; 2003.
12. Degoyet CD. Post disaster relief: the supply-management challenge. *Disasters* 1993;17(2):169–176.
13. Sabri EH, Beamon BM. A multi-objective approach to simultaneous strategic and operational planning in supply chain design. *Omega Int J Manage Sci* 2000;28(5):581–598.
14. Lodree EJ, Taskin S. An insurance risk management framework for disaster relief and supply chain disruption inventory planning. *J Oper Res Soc* 2008;59(5):674–684.
15. Ozdamar L, Ekinici E, *et al.* Emergency logistics planning in natural disasters. *Ann Oper Res* 2004;129(1–4):217–245.
16. Sheu JB. An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transp Res Part E-Logist Transp Rev* 2007;43:687–709.
17. Kapucu N, *et al.* Logistics and staging areas in managing disasters and emergencies. *J Homeland Secur Emerg Manage* 2007;4(2).
18. Carbajal JA, Ergun O, Keskinocak P, *et al.* Debris management operations. Atlanta (GA): Center for Humanitarian Logistics, Georgia Institute of Technology; 2008.
19. Federal Emergency Management Agency (FEMA). Public assistance debris management guide. Available at <http://www.fema.gov/pdf/government/grant/pa/demagde.pdf>. Accessed 2009 Jan 15.
20. Dixit V, Pande A, Radwan E, *et al.* Understanding the impact of a recent hurricane on mobilization time during a subsequent hurricane. *Transp Res Rec: J Transp Res Board* 2008;2041:49–57.

OPERATIONS RESEARCH TOOLS FOR ADDRESSING CURRENT CHALLENGES IN EMERGENCY MEDICAL SERVICES

SHANE G. HENDERSON

School of Operations Research and
Information Engineering,
Cornell University, Ithaca,
New York

The job of an emergency medical service (EMS) provider is to respond to calls for assistance, render urgent medical care at the scene of a call and then, if necessary, transport the patient to an appropriate hospital. An EMS provider must coordinate the actions of many ambulances, the staff in which may have different levels of training, to address calls of many different types. Call volumes vary significantly according to cyclical patterns on a daily, weekly, and annual basis, and calls are certainly not evenly spread over geographic regions.

EMS problems are similar to those faced by other emergency services including fire and police, although there are important differences. For a superb overview of work in these various fields to the early 1990s, see Ref. 1; and for some recent commentary, see Ref. 2. The survey [3] is a valuable reference for both operations research (OR) specialists and EMS professionals, although it is written primarily for EMS professionals. EMS problems also have some similarities with the taxi and courier industries, although the latter industries can benefit from *a priori* scheduling of a nontrivial fraction of their calls.

The focus in this article is on *operational* issues rather than *clinical* issues. Clinical issues are those issues that are essentially medical in nature, such as the difference that prompt intubation can make to patient survival rates. The clinical side is extremely active. See, for example, the *Annals of Emergency Medicine*, the *Journal of Emergency Medical Services*, or *Academic Emergency*

Medicine. These journals occasionally publish articles that focus on operational aspects.

OR has been used to assist in decision making in EMS at least since the 1960s [4]. Typical applications include the siting of new bases and staff scheduling. But there are still a host of challenges faced by EMS providers, some of which are variants of these well-known problems, while others have arisen only recently. The goal of this article is to describe some of these challenges and suggest possible avenues of (operations) research that might help address them. We will particularly emphasize system-status management (SSM) as a tool for addressing several problems at once. The article is not meant to be a state-of-the-art survey, but rather a “call to arms.” To absorb the entire article, the reader should have a background in the tools of OR, say at the undergraduate or masters level, but much of the discussion should be accessible without this background.

This is one of two articles that deal with EMS. In the other article titled *Emergency Medical Service Systems that Improve Patient Survivability*, the goal is to provide an overview of next-generation EMS design that helps address the problems faced today. In contrast, we start with the problems of today, indicate potential solution approaches, and go into depth on one approach (SSM) that is already in use.

The good news for an operations researcher looking at EMS problems is that the area is data-rich. Computer-aided dispatch (CAD) systems have been in place in many locations for over a decade, and so in large centers it is invariably the case that data is available on every call. This is not necessarily the case in small centers, however. The amount of data varies, but there is usually at least three years’ data available as of 2008, and often much more—even decades in some places. Time stamps are logged for each call, including the time the call was received, the time the call was assigned to an ambulance, the time the ambulance was en route to the call, and so forth. Some of these time

stamps are logged manually by ambulance officers, while others are logged automatically by the CAD system, so the quality of the time stamps varies (people occasionally forget). A key component to planning systems is travel-time information, including time-dependent travel times or speeds on road networks. This sort of data is often available from town planning departments, but is usually intended for long-term planning questions, so often needs editing before it is suitable for EMS applications. A potential source of travel-time data is automatic vehicle location (AVL) data, which is becoming much more widely available as global positioning system (GPS) units on ambulances become standard. AVL data gives the location of ambulances as recorded by a GPS unit at regular intervals of time. GPS readings vary in quality depending on how accessible the GPS satellites are. For example, GPS readings often degrade when ambulances are in tunnels or in the urban canyons that are typical of the centers of large cities, so even this data source varies in quality.

The pressing need for efficiencies in EMS systems, together with the availability of large amounts of data, suggests that quantitative OR methods, together with appropriate statistical methodology, should be well positioned to strengthen an already well-established presence in EMS.

AN OVERVIEW OF EMERGENCY MEDICAL SERVICE CHALLENGES

Measuring Performance

EMS systems are designed to help reduce morbidity and mortality by quickly and safely responding to calls, providing emergency medical assistance at the scene, and transporting those patients who require further treatment to an appropriate hospital. As with many decision-making questions in health care, it is difficult to quantify the medical impact of changes. The traditional approach is to measure performance through response times. The response time for a call is the elapsed time from when the call is received to when an ambulance arrives at the scene. Performance is then measured as the

fraction of calls received that have response times of x minutes or less, where x is usually 8 or so; see Ref. 5 for historical notes and discussion from a practical perspective. Let us call this performance measure the response time threshold fraction (RTTF). RTTFs are often reported separately for different regions and different time intervals, so that a picture of performance in space and time can be obtained.

The RTTF is easily explained, easily measured, and unambiguous. It can also be computed using existing methodologies including simulation and hypercube methods. (See Ref. 6 for an overview of simulation in EMS and Refs 7–9 for how to do this using the hypercube method.) These are important advantages that help to explain why RTTF is so prevalent in the industry. RTTF has its origins in cardiac survival studies, where survival rates are linked to response times. But it is also somewhat arbitrary. The threshold of 8 min is chosen from the survival probability versus response time curve, and is often adjusted depending on whether one is in a rural setting (larger values are used) or a metropolitan area. It is also highly city dependent, with values 8, 9, 10, and 12 being common choices. Different thresholds are used for different types of calls, since some calls have lower priorities than others and so naturally receive slower response times.

For all of these reasons, there is great interest in finding better performance measures than RTTF. One approach is to use a utility function u so that a call with response time t receives a utility or “score” $u(t)$, and the goal is to maximize the average utility over all calls. RTTF is actually a special case where u only takes values 0 and 1. The utility approach is a central theme in both (also see *Emergency Medical Service Systems that Improve Patient Survivability*) [10], where the utility function adopted is again based on cardiac studies. Of course, there are other types of calls including, for example, trauma calls, so interpretation of the utility can be difficult, even when it is specialized to different types of calls. As another example of the difficulty in coming to agreement on utilities, many cardiac patients wait for some

time before calling EMS, so a few minutes difference in response time is somewhat immaterial when there has already been a delay of up to a few hours. Nevertheless, the idea of attaching utilities to calls and measuring average utility is important. For example, a key observation in Ref. 10 is that solutions to ambulance deployment problems obtained using RTTF can perform very poorly when measured in terms of average utility.

Perhaps, we should continue to err on the side of simplicity and simply measure performance in terms of *average* response times instead of RTTFs or more complex measures. See article titled *Emergency Medical Service Systems that Improve Patient Survivability* for some discussion of this question.

Some EMS providers additionally measure performance in other ways. For example, one approach to ambulance deployment used in Edmonton, Alberta, is to specify, for each number of available ambulances, a set of locations that they should occupy. When ambulances are appropriately positioned, the service is said to be “in compliance,” and performance is measured by the fraction of time the system is in compliance. One would expect that RTTFs are also tracked.

Delays at, and Diversion from, Emergency Departments

The time required to transfer a patient from an ambulance to an accepting emergency department varies greatly depending on the medical needs and priority of a patient and how busy the emergency department is. Typical desired transfer times are on the order of 15 min, but in many cases can be as large as 1 h or more. This places tremendous pressure on EMS organizations because it dramatically increases—even doubles—the time required to handle each call. Another serious problem is known as *ambulance diversion*, which occurs when an emergency department notifies dispatchers that it will temporarily not accept patients from ambulances. The resulting patient transport times are increased because ambulances are forced to go elsewhere, again impacting ambulance utilization.

Both transfer delays and diversion can have serious medical consequences. A statistical study [11] links ambulance diversion to increased death rates from heart attacks. The problem may be even more serious than this paper suggests, since transfer delays can be much larger than the additional driving time associated with diversion.

What can be done to reduce the impact of delays and diversion?

One approach that has been tried is to place an EMS paramedic at each receiving hospital. The paramedic receives patients from arriving ambulances and subsequently transfers those patients to the emergency department when possible. When not busy with this function, the paramedic can assist in the emergency department. One can think of the paramedic as a buffer (in the sense of manufacturing) between arriving ambulances and the emergency department. In one case, this was observed to work well for about one week, until the hospital realized it could use the EMS paramedic for its own ends, and patient transfer delays returned.

This points to a key problem in that hospitals and EMS systems are usually operated as separate organizations. The incentives that are in place for both organizations are not geared toward inducing cooperative behavior in the emergency department where these organizations interface.

There are at least three possible solutions to this problem.

The first, and most obvious, potential solution is to immediately conclude that hospitals need more resources to cope with the flows of patients entering their emergency departments. But as noted in several places [12, 13] any problem may lie downstream from the emergency department, in hospital wards that need to free capacity to allow emergency department patients to be transferred. It may also be the case that the systems in place in hospitals can be streamlined to make it easier for emergency department staff to keep pace with patient flows and reduce the need for ambulance delays and diversion. This is a key place where OR can help.

The second potential solution is to adopt some form of “regionalization,” as is currently under way in a number of provinces in

Canada to varying degrees. Here the hospital and pre-hospital organizations are placed under a single umbrella organization, with the goal that the pre-hospital and hospital organization are now a single organization and therefore will “pull in the same direction.” It is not yet clear whether this step will help though, because the size of the pre-hospital and hospital organizations makes it difficult to maintain a “systems view” of operations, and it is very likely that a hierarchical organization will result, with EMS on one branch and emergency departments on another, perhaps leading to the same situation we have now. This is not a foregone conclusion of course, but it is one possible outcome.

The third potential solution is to redesign the incentive schemes for both EMS and emergency departments, that is, change the performance measures that are used to report their performance, along with how those performance measures are written into contracts. Perhaps by aligning the incentive schemes using the game-theoretic field of mechanism design, one can design contracts that lead to outcomes that are simultaneously of high quality for patients, EMS providers, and emergency departments.

Increasing Traffic Congestion

The high-profile aspect of traffic congestion is large-scale traffic jams. But traffic congestion can also mean large traffic flows on a daily basis, and therefore slow travel on routes that were never intended to handle the volumes that they receive. Even when ambulances travel with lights and sirens, it can take time for drivers to react and clear a path for an ambulance. Furthermore, ambulances spend significantly more time traveling at regular traffic speeds than at lights and sirens speeds, in which case they are slowed by traffic congestion, thereby increasing the load on ambulances for the same call volume. This, in turn, translates into increased response times due to unavailability of the closest appropriate ambulance.

So what can EMS providers do about this problem?

Rather than maintain ambulance bases with three or more ambulances at a base,

many EMS providers now spread their available ambulances around the city. This is sometimes known as *parking on street corners*. Crews are asked to sit in their ambulances in parking areas, waiting for calls. The particular locations are chosen depending on time-dependent demand and travel speeds, and various methods are available for such planning [14]. This has the effect of reducing the driving distance to calls, and therefore can improve response times. It also creates the need for real-time ambulance location and relocation. To see why, picture a city with evenly spread ambulances. When a call comes in, one of those ambulances “disappears” from the picture, leaving a potentially large “gap” in coverage. It may then be advantageous to move one or more nearby ambulances to reduce the size of this gap. This process of using real-time information to relocate available ambulances is known as *system-status management*, and we will discuss it in depth later in this article. It is worth noting that under a system where multiple ambulances are stationed at a single base, only occasionally are all ambulances at the base busy with calls, so gaps in coverage do not open up as frequently. However, because of the clumping of ambulances at bases, the distance between available ambulances is typically large, and this translates into large response times.

The use of SSM methods makes having accurate travel-time or speed information more important than ever. As mentioned earlier, currently available speed information on road networks is usually not very accurate, so one would like to “tune” the speeds, probably using AVL data. But how should this be done, given that the data consists of noisy observations of ambulance locations at discrete points in time, so one is uncertain even about the exact route that an ambulance has taken? The field of map matching [15,16] is undoubtedly an important tool that can perhaps be combined with carefully designed statistical methods to update modeled travel speeds. See Budge *et al.* [17] for a recent statistical study of travel times that models travel times as random, and Henderson and Budge *et al.* [18,19] for the consequences of travel times being random.

Vehicle Mix Questions

Should one's ambulance fleet consist of advanced life support (ALS) vehicles only, or a mix of ALS and basic life support (BLS) vehicles? To understand the distinction, note that ambulance staff can have a variety of levels of training. *Ambulance officers* (also called *emergency medical technicians*) typically have on the order of several weeks of training and are highly adept first responders. However, they cannot deliver the higher level of care needed on a significant fraction of calls that a *paramedic* can provide (e.g., administer drugs) due to a much higher level of training. ALS vehicles have a paramedic on board, while BLS vehicles do not.

Most cities have a mix of ALS and BLS vehicles, although some authors strongly advocate an ALS-only system [20]. It is cheaper to run a BLS ambulance than an ALS ambulance, so for a given budget one can operate more vehicles in a mixed fleet than in an ALS-only fleet. With a mixed fleet, one sends BLS vehicles to calls where paramedic-level attention is not needed. However, this requires that dispatchers make a determination at the time of receiving a call about which type of vehicle to send, thereby inflating the time required before dispatching a vehicle, and creating the possibility of sending the wrong vehicle, leading to additional delays before a high level of care can be provided. The trade-offs between ALS-only fleets or a mixed fleet are therefore complex, and it is not clear which approach is better in a given situation. OR tools can help quantify the trade-off to assist in making a determination. For example, simulation was used in Ref. 6 to model the trade-offs between sending a vehicle to a call based on minimal information from a caller, as opposed to obtaining more information from the caller before dispatching (ProQA). Using ProQA ensures a better match of vehicle type to the call, but slows down dispatch times, thereby inflating response times.

The ALS/BLS mix is just one example of a number of vehicle mix questions. For example, some EMS organizations use vehicles staffed by a single paramedic, which can rapidly respond to calls but cannot transport patients. How many such vehicles should be

employed? Should scheduled patient transports be performed by emergency vehicles or by a fleet dedicated to the task? And to what extent should EMS organizations employ helicopters?

Increasing Call Volumes

Virtually all EMS providers are seeing their call volumes increase over time. This increase is probably a consequence of many factors, including population increase, demographic changes (which are strong predictors of call volumes [21,22]), and what the general public perceives as the threshold for calling an ambulance. This obviously poses challenges for an EMS provider whose budget is often constrained.

Using careful dispatch strategies can help. For example, it has been known for some time [23] that sending the closest available vehicle is not necessarily optimal, because it can lead to imbalances in workload and therefore inflated response times. Nevertheless, most dispatchers follow either exactly this policy or one very close to it, partly because of the danger of litigation in the event of an unfortunate outcome, but also partly because there is a lack of decision support systems that can help with this decision. What do optimal dispatching policies look like and how can they be deployed in a dispatcher-friendly manner?

Another question we must ask ourselves is whether all of the incoming calls actually *need* EMS response. During the September 11 attacks in New York City, some emergency departments noted a considerable *drop* in patient arrivals that was not due to a lack of ambulances. It appears that the general public felt that their problems were outweighed by the needs of potential casualties from the World Trade Center and did not call for help. In many cases, no doubt, they should still have called for help, but it does make one wonder how much EMS and emergency room demand could be mitigated using other strategies. Is there a way to shape demand, perhaps through public education schemes? Would such schemes lead to better outcomes in the sense of overall public health, or would such schemes reduce EMS call volumes at the expense of public health?

Yet another approach to dealing with increasing call volumes is to attempt to obtain better performance from the resources at one's disposal. One such method that can help immediately, and that is already in use, is system status management, and this is the subject of the remainder of this article.

EXISTING APPROACHES TO SYSTEM-STATUS MANAGEMENT

Many ambulance organizations are moving away from the model where ambulances are stationed in twos and threes at bases, instead preferring to spread the vehicles out over the city in an attempt to reduce the travel distance to calls. When a call is received and an ambulance is dispatched to the call, a "hole" is left in the coverage of the city. Should the nearby vehicles be rearranged? If so, how? Any method that gives a strategy for performing such rearrangements is known as *system-status management*, or by one of the synonyms move up, redeployment, dynamic relocation, or dynamic positioning.

In order to implement SSM, dispatchers need to be aware of the location and status of the ambulance fleet. This is not difficult for EMS organizations that have installed GPS units on all vehicles and where crews send updates on their status to the dispatch center. Indeed, virtually all large EMS organizations have adopted these practices.

Another key issue is ambulance crew receptiveness to SSM. Many implementations encounter some resistance from crews and perhaps with good reason. Ambulance crews have a stressful job that also requires them to work in shifts. On top of that, a poorly implemented SSM approach may lead to crews spending almost the entirety of their shift driving between locations to fill holes in coverage, or parked in some undesirable location that is far less comfortable than any ambulance base. The crews do not see the "birds-eye view" that leads to these redeployments, and so the trips can seem pointless to them. However, if implemented with care, SSM can obtain material improvements in response time with only modest disruption to crews.

We now describe specific methods for implementing SSM.

Manual Search and Selection

Suppose a dispatcher is considering redeploying a specific ambulance to one of several potential locations. Experience often guides the selection of a specific location. "What-if" tools can also help. For example, one approach is to assume that all other vehicles will not change status or location in the short term, and then consider the candidate locations, one at a time. As long as the number of candidate locations is small, it is typically a simple matter to recompute some measure of quality (usually "coverage," where coverage means the fraction of future demand that can be reached within some time threshold from the candidate locations) at each candidate location and select the one that is best. Graphical tools that display coverage, for example, by coloring locations in a map according to projected response times, can also help. This is the basic idea behind simple implementations of SSM. Such implementations are easy to understand by dispatchers, but they also have some important shortcomings.

1. It can be difficult to gain a picture of both call density by location and coverage in graphical displays. Therefore, dispatchers need experience to help in determining which vehicle relocations might prove effective.
2. It is difficult to consider all possibilities when simultaneously considering moving two or more vehicles because of the combinatorial increase in the number of candidate locations that need to be considered.
3. The status and location of ambulances is constantly in flux, so decisions based on such static models can miss the fact that the situation may look very different by the time any relocations have been completed.
4. In busy systems, it is quite possible that vehicles on their way to a new location may be dispatched to a call before they reach their intended destinations,

calling into question the basis for the relocation decision.

5. It is difficult to get a sense of which relocations are the most important ones in terms of reducing response times, which is important when one wishes to limit relocations to avoid frustrating crews.

Real-Time Optimization

Several of the disadvantages of the real-time manual approach discussed above motivate an automated approach. All such searches rate candidate solutions using some measure of quality, which is usually coverage, as defined above, although other measures have been used [24]. Exhaustive searches can work quite well when the number of potential relocations is small, but they can quickly become overwhelmed when multiple ambulance moves are simultaneously considered. It is then that optimization-based methods become the preferred approach.

The first optimization-based method was developed by Kolesar and Walker in the context of fire-fighting operations [25]. It involved solving, in real time, three integer programs using heuristics. Unfortunately, one cannot just “cut and paste” this approach into the EMS setting, because there are some important differences in the way fire services and EMS services operate. The key differences are that (i) fire companies have much lower utilization than EMS (to ensure that they can respond rapidly to major incidents), and (ii) fire companies can be engaged at a structure fire for 8 h or more, whereas it is unusual for an ambulance to be engaged in a single call for more than 2 h. These differences mean that it is reasonable to assume that the dominant system conditions will not change during a fire company deployment (call), while, as we argue below, in the EMS setting such an assumption is problematic.

A real-time optimization approach was developed in Ref. 26. The integer programming formulation ensures appropriate coverage of various locations, that vehicles are not relocated too often and that vehicles do not take too many “round trips.” A tabu-search

heuristic on a parallel computing platform was used to solve the model. Similar formulations have been adopted in some commercial systems.

Offline Optimization

An alternative to the methods above is to develop a lookup table which gives, for each number of available ambulances, the set of locations where available ambulances should be positioned. Dispatchers then attempt to direct the ambulances to occupy those locations with as little disruption to crews as possible. Lookup tables can be time-dependent to reflect time-dependent travel speeds and demand patterns, and are compact and easily understood. But how can they be constructed?

Here, one can exploit the large amount of available research on identifying optimal base locations in the *static* version of ambulance deployment where ambulances always return to their preassigned bases when free. Integer-programming methods (see Ref. 27 for a recent survey) are available for selecting ambulance locations for a given number of ambulances. One can solve one such optimization problem for each number of available ambulances. A difficulty here is that there is no attempt to ensure consistency between the set of locations chosen for n available ambulances, and those chosen for $n + 1$ available ambulances. This can lead to many redeployments, which should be avoidable. A partial solution is to require that the redeployment locations be *nested* in the sense that the set of locations chosen for n ambulances be a subset of those chosen for $n + 1$ ambulances, for each $n = 1, 2, \dots, N - 1$, where N is the total number of ambulances in the fleet [28]. These nesting constraints can be satisfied by solving a single, large integer program that simultaneously finds the optimal locations for all $n = 1, 2, \dots, N - 1$, and includes explicit constraints that enforce the nesting.

This method is a highly appealing and important option for SSM implementation, but it does possess some disadvantages. In particular, Issues 3, 4, and 5 identified as difficulties for manual selection are also important difficulties for this approach.

Stochastic Dynamic Programming

The most important issues with the methods we have discussed thus far are that (i) they are based on the assumption that vehicles are exactly at the desired locations at all times, and (ii) they cannot give a clear sense of the importance of particular redeployments. Both of these issues can be addressed using stochastic dynamic programming (DP) models.

DP models were first introduced for this problem in work by Berman and co-authors in Refs 29–33. DP is certainly a very natural framework to employ, since it is the study of problems where decisions must be made over time in the presence of uncertainty, which exactly describes our problem. To apply DP, one tracks the locations and status of all ambulances over time. At any decision point, one defines a set of actions (potential redeployment decisions) that might be taken. The actions are of the form “send ambulance x to location y .” The DP is designed to discover a *value function* V that gives the quality of a particular configuration of ambulances at a particular time. The function V maps the system states (ambulance locations and status, and time) to a measure of the quality of the configuration, which can be interpreted as the expected performance of the system (e.g., number of calls answered on time) starting from that particular state. Once one obtains the value function V^* associated with an optimal policy, it can be used to determine the optimal decision at any stage by implementing the associated greedy policy: At each decision point, one selects the action that maximizes the expected value of the immediate reward plus “downstream” rewards. See Puterman [34] for a comprehensive introduction to DP and full discussion of these concepts.

Recently, Zhang *et al.* [35] have revisited the use of DP for ambulance relocation, establishing a number of results that help build intuition for what optimal policies look like. Their results show that one should send vehicles to one of several distinguished locations, where the set of locations depends on the arrival rate of calls. Furthermore, they show numerically that the heavier the load of calls

on the system, the less effective SSM can be relative to a static policy.

In the papers mentioned above, the approach is to obtain the optimal value function V^* by storing a value for *every possible state* and applying standard DP techniques. This necessitates keeping the number of states small, which means that only very small examples can be treated. Therefore, the models addressed in this work are somewhat stylized. The results are important in that they help build our understanding of when SSM can be effective and what effective policies look like, but the approach cannot be applied directly to realistic-sized systems. To enable the leap to realistic-sized systems, a different approach to representing the value function is needed, which brings us to the subject of approximate dynamic programming.

APPROXIMATE DYNAMIC PROGRAMMING FOR SYSTEM-STATUS MANAGEMENT

Approximate dynamic programming (ADP) is a general-purpose suite of tools for solving problems where successive decisions must be made over time under uncertainty. Essentially, ADP involves an implementation of dynamic programming where one stores an *approximation* for the value function, instead of the value function itself. This allows one to tackle problems that would be impossible with exact DP owing to the huge (often infinite) state space involved. The catch is that one cannot, in general, expect ADP to find an *optimal* policy. The goal, instead, is to find a *high-quality* policy in a computationally tractable manner.

We will sketch the key ideas behind an ADP approach to SSM developed in Ref. 36 and further explored and improved in Ref. 37. We will refer to this approach as *our* approach because the present author was one of the authors of this work. The goal is to minimize the fraction of calls received that have a response time that exceeds a given threshold. We refer to such calls as *lost* or *missed*. The specific version of ADP we use is an adaptation of a generic approach for ADP that was initially described in Ref. 38 and involves a mix of simulation and regression.

Let V denote an approximation to the optimal value function. We will explain where the approximation comes from shortly, but for now assume that it is given. Associated with V is a policy, known as the *greedy policy* with respect to V , that works as follows. At each time point t at which a decision is required, let s denote the current state of the real system, which is essentially all of the information that dispatchers have at their disposal at time t . Therefore, s includes information on the location and status of all ambulances, including the time that each ambulance has been at its location at time t . It also includes information on any queued calls. It does not include information on future events such as might be obtained from an event list in a simulation, for example, the time remaining until an ambulance completes a hospital transfer in progress. One then considers each possible *action* in turn, where each action corresponds to a request for one or more ambulances to redeploy to specific locations. As a consequence of each action, the specified ambulances will begin to deploy to the new locations, but do not necessarily reach those locations before the situation changes. For each action a , we let $S(a)$ denote the random system state corresponding to the new situation at the time of the next system event, which could be an ambulance arriving at the scene of a call, a new call arriving, and so on. Also, let $C(a)$ be the immediate cost incurred by this action, which in our case is either 0 (no missed calls) or 1 (one missed call). We then compute or approximate $E[C(a) + V(S(a))]$, and choose the action that minimizes this quantity as our decision at time t . (The expectation is computed *conditional* on the system state at time t .) We then continue the simulation to the next decision point, at which time the process above is repeated.

We cannot compute the required conditional expectation analytically, and instead use a small number of simulations to obtain a Monte Carlo estimate of it for each action that is considered. We refer to this calculation as a *micro simulation*, because each replication of a micro simulation involves simulating from the current system state forward in time for at most a few minutes of simulated time, and

this can be accomplished with a very small amount of computation. In our implementation, the micro simulations for all potential actions take a fraction of a second. Furthermore, these micro simulations are only necessary when we make a relocation decision, and then only from the current state of the system, so they represent a minor computational overhead. The Monte Carlo error we incur means that we will occasionally take an action that does not minimize the true expectation. However, this will typically occur when the suboptimal action we choose has a value that is close to that of the optimal action, so we do not expect that this effect will cause practical problems, provided that the number of micro simulations is reasonable.

To summarize thus far, for a given function V there is an associated greedy policy that we can approximate using micro simulations, and this can be implemented in real time provided that one has a simulation model of the system along with the ability to observe the state of the system in real time.

So, where does the function V come from? We use a linear combination of prespecified basis functions $V_1, V_2, \dots, V_n$, that is, for each state s ,

$$V(s) = a_1 V_1(s) + a_2 V_2(s) + \dots + a_n V_n(s),$$

for some coefficients $a_1, a_2, \dots, a_n$. The basis functions are chosen heuristically to reflect characteristics or features of the state s that are thought to be important in determining a policy. In the present case, after experimentation with potentially useful functions we settled on $n = 6$ functions as follows:

- V_1 A constant.
- V_2 The number of queued calls that will definitely not be reached within the time threshold.
- V_3 The arrival rate of calls to areas that cannot be reached by any available ambulance.
- V_4 The arrival rate of calls to areas that can be reached by an available ambulance but will likely be missed because of congestion. Here we measure the loss rate of calls as the product of two terms, the first being

the arrival rate, and the second being the loss probability as measured by the Erlang loss function.

- V_5 A version of V_3 where the ambulances are assumed to have reached their destinations if en route somewhere, and where presently stationary ambulances remain where they are.
- V_6 A version of V_4 that is modified as in V_5 above.

Why did we choose these functions? The Erlang loss function was helpful in choosing static locations for ambulances [39] so it seems reasonable to include it, and SSM is designed to help in the short run—say over the next hour or so—so using versions of the basis functions associated with predicted ambulance locations on that timescale seems reasonable. But there is plenty of room for argument and art in this selection, and we do not claim that this set of functions is the best possible. Much more remains to be done in basis function selection.

Given a set of basis functions, the problem then reduces to determining the coefficients $a_i, i = 1, 2, \dots, n$. There are a variety of methods for doing this. We use regression as follows. The state s also includes the time t at which the state is observed. The function value $V(s)$ then represents the expected cost (number of missed calls) observed starting from state s at time t through to the end of the simulation horizon. (We use two weeks for the simulation horizon.) In implementing this algorithm, we have found it necessary to *discount* the costs in time. This is purely a computational device that helps ensure convergence; see Maxwell *et al.* [37] for more discussion on why this is needed and why it works. (A similar trick was employed in a different setting in Ref. 40.) So the function value $V(s)$ actually represents the expected *discounted* number of missed calls from the time t until the end of the simulation horizon. The simulation provides *observations* of the discounted number of calls till the end of the horizon. So we can use linear regression to try to align $V(s)$ with these observations.

The result of the linear regression is a set of coefficients $a_1, a_2, \dots, a_n$, which determine a new value function $\tilde{V}$. We can then iterate this process, with the next iteration simulating the greedy policy with respect to $\tilde{V}$ instead of V . After a modest number of iterations (usually less than 20 or so), we terminate the process, and select what appears to be the best policy (i.e., set of coefficients) seen thus far.

Figure 1 gives an example of the results of such a training session for data from Edmonton, Alberta. The data were modified prior to use to protect the confidentiality of Edmonton EMS performance. Here, the vertical axis plots the estimated *undiscounted* percentage of missed calls. The apparent-best policy arises after three iterations. There are two reasons why this is the *apparent* rather than necessarily the *actual* best policy. First, at each point we run a simulation to estimate the performance of the given policy. Therefore, there is simulation error in the estimates of performance. Second, even though each point gives an unbiased estimate of the performance of the corresponding policy, when we focus in on the point with the lowest function value in the plot, we are biased toward observing a point where the estimated cost is lower than the true cost, that is, the estimated cost associated with the best observed policy is biased low. If we perform an independent simulation of the estimated-best policy, we obtain an unbiased estimate of its performance, which in the case of the observed best policy in Fig. 1 gave an estimated performance of $25.4 \pm 0.1\%$. By way of comparison, the static policy (where ambulances return to preassigned bases after completing each call) misses $29.5 \pm 0.1\%$ of calls.

The results of Fig. 1 are based on policies that only consider an ambulance for redeployment when it is completing a call either at scene or at a hospital. This is appealing because it ensures that crews are not overly disrupted by excessive redeployments. However, it is possible to obtain even greater improvements in performance by considering additional redeployments. In Ref. 37, approximately an additional 3% of calls are not missed over and above the

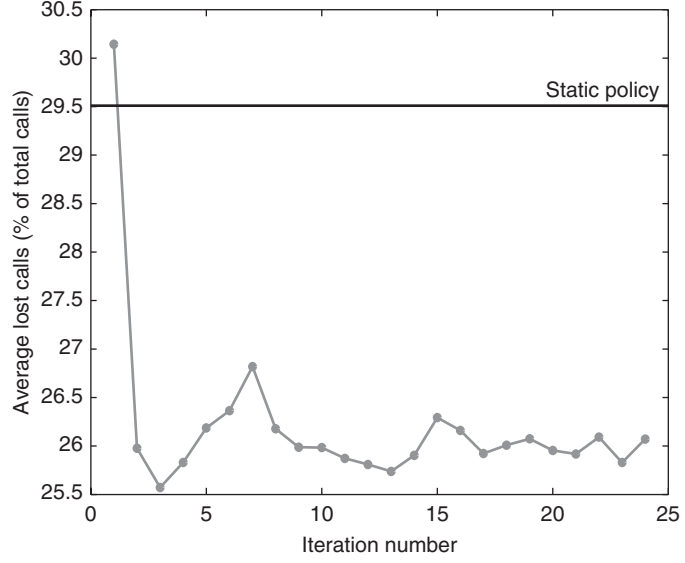


Figure 1. An example of the training process of ADP using representative data from Edmonton, Alberta.

completing-calls-only redeployment policy when additional redeployments are considered at regular intervals.

One might reasonably ask whether ADP identifies the optimal choice of coefficients $a_1, a_2, \dots, a_n$. As mentioned earlier, the algorithm discussed above is not designed to converge to any particular set of coefficients. This is because the computational burden in training is great, so we cannot simulate a

large number of policies, which is essential if we wish to design a converging algorithm. Instead we focus on exploring the coefficient space for a highly effective policy.

The natural question to ask then is how the policy identified by ADP ranks amongst the potential policies. In Fig. 2, we give a histogram of the performances of the top-performing policies amongst 1000 policies. These 1000 policies were selected uniformly

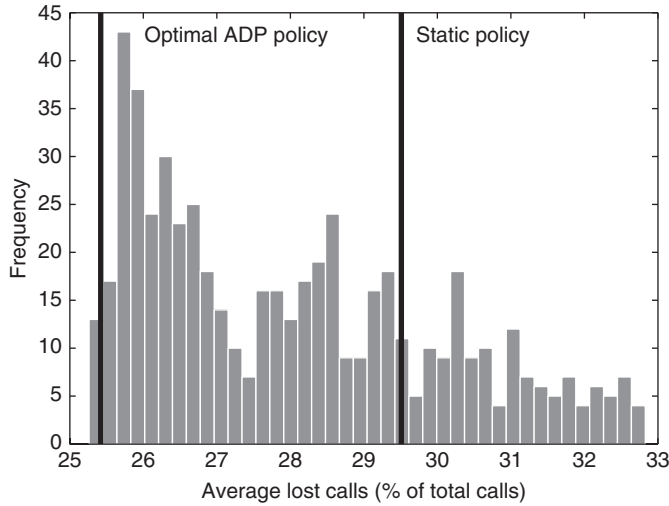


Figure 2. Histogram showing the estimated performance of a number of policies using representative data from Edmonton, Alberta.

at random from a cube surrounding the final recommended coefficient vector a with side lengths chosen to be commensurate with the variation in the coefficients seen during the course of the ADP iterations. We see that the performance of the optimal ADP policy is very close to the best policy seen. It may even be optimal, because we again obtain bias when we focus in on the policies appearing at the extreme left of the histogram. There are no theoretical guarantees that the ADP policy would do so well, and so it may be necessary to apply a separate optimization procedure after the ADP search to refine the coefficients. We are considering this step in current research.

CONCLUSION

EMS providers face significant challenges that, in the long run, could be addressed by a variety of remedies that involve OR methodologies. We have suggested some such remedies, but they are a long way from being developed, much less implemented. SSM can help essentially immediately with many of the difficulties faced by EMS practitioners, but it needs to be implemented carefully to fully realize the potential benefits in response times while not unduly inconveniencing crews. Existing implementations vary in quality, which has led to the technique receiving a bad name with crews.

Offline and on-line optimization implementations of SSM probably represent the current “best practice” in the area, with offline optimization perhaps being more transparent and accepted by dispatchers than the on-line approach. We have explained in broad terms how these methods work. We have also sketched the main ideas behind an approach to SSM based on ADP. This new approach is already competitive with the existing optimization-based methods, and has a number of advantages over the existing methods. It is not quite ready for commercial implementation, but is certainly close to that point.

All of the methods for SSM rely on quality estimates of call arrival rates broken down by time and location. Our ability to obtain

quality estimates of these quantities lags our ability to exploit those estimates in OR models. So, while there are many potential research directions described in this article, perhaps the most pressing one is the need to expand existing efforts [41], in statistical analysis of EMS data.

Acknowledgments

We thank the editors and referees for very helpful comments that improved this article. This article is based upon work supported by the National Science Foundation under the grants DMI 0400287 and CMMI 0758441. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Swersey AJ. The deployment of police, fire, and emergency medical units. In: Pollock SM, Rothkopf MH, Barnett A, editors. *Operations research and the public sector*. Amsterdam: North Holland; Publishing Co. 1994.
2. Green LV, Kolesar PJ. Improving emergency responsiveness with management science. *Manag Sci* 2004;50(8):1001–1014.
3. Goldberg JB. Operations research models for the deployment of emergency services vehicles. *EMS Manag J* 2004;1:20–39.
4. Savas ES. Simulation and cost-effectiveness analysis of New York’s emergency ambulance service. *Manag Sci* 1969;15:B608–B627.
5. Fitch J. Response times: myths, measurement & management. *J Emerg Med Serv* 2005;30(9):46–56.
6. Henderson SG, Mason AJ. Ambulance service planning: simulation and data visualization. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. *Operations research and health care: a handbook of methods and applications*. Boston (MA): Kluwer Academic; 2004. pp. 77–102.
7. Larson RC. A hypercube queueing model for facility location and redistricting in urban emergency services. *Comput Oper Res* 1974;1:67–95.
8. Larson RC. Approximating the performance of urban emergency service systems. *Oper Res* 1975;23:845–868.

9. Budge S, Ingolfsson A, Erkut E. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Oper Res* 2009;57(1):251–255.
10. Erkut E, Ingolfsson A, Erdoğan G. Ambulance deployment for maximum survival. *Nav Res Logist* 2008;55:42–58.
11. Yankovic N, Glied S, Grams M, *et al.* Ambulance diversion and myocardial infarction mortality. 2009. In press.
12. Green LV. Capacity planning and management in hospitals. In: Brandeau ML, Sainfort F, Pierskalla WP, editors. *Operations research and health care: a handbook of methods and applications*. Boston (MA): Kluwer; 2004. pp. 15–41.
13. Green LV. Using Operations Research to reduce delays for healthcare. In: Chen Z-L, Raghavan S, editors. *Tutorials in Operations Research*, P. Gray, Series Editor. Washington (DC): INFORMS; 2008.
14. Rajagopalan HK, Saydam C, Xiao J. A multi-period set covering location model for dynamic redeployment of ambulances. *Comput Oper Res* 2008;35:814–826.
15. Mason AJ, Henderson SG. Analysing ambulance travel: a dynamic program for map matching with sparse GPS data. Working Paper. 2007.
16. Krumm J, Letchner J, Horvitz E. Map matching with travel time constraints. Society of Automotive Engineers (SAE) 2007 World Congress. Detroit (MI): SAE International; 2007. Paper 2007-01-1102.
17. Budge S, Ingolfsson A, Zerom D. Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. *Manag Sci* 2010;56:716–723.
18. Henderson SG. Should we model dependence and nonstationarity, and if so how? In: Kuhl ME, Steiger NM, Armstrong FB, *et al.*, editors. *Proceedings of the 2005 Winter Simulation Conference*. Piscataway (NJ): IEEE; 2005. pp. 120–129.
19. Budge S, Ingolfsson A, Erkut E. Optimal ambulance location with random delays and travel times. *Health Care Manag Sci* 2008;11:262–274.
20. Balaker T, Summers AB. Emergency medical services privatization: frequently asked questions. 2003. Available at http://www.reason.org/policystudiesbysubject.shtml#priv_govreform.
21. Aldrich CA, Hisserich JC, Lave LB. An analysis of the demand for emergency ambulance service in an urban area. *Am J Public Health* 1971;61(6):1156–1169.
22. Setzler H, Saydam C, Park S. EMS call volume predictions: a comparative study. *Comput Oper Res* 2009;36(6):1843–1851.
23. Carter GM, Chaiken JM, Ignall E. Response areas for two emergency units. *Oper Res* 1972;20(3):571–594.
24. Andersson T, Värband P. Decision support tools for ambulance dispatch and relocation. *J Oper Res Soc* 2007;58:195–201.
25. Kolesar P, Walker WE. An algorithm for the dynamic relocation of fire companies. *Oper Res* 1974;22:249–274.
26. Gendreau M, Laporte G, Semet F. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Comput* 2001;27:1641–1653.
27. Brotcorne L, Laporte G, Semet F. Ambulance location and relocation models. *Eur J Oper Res* 2003;147:451–463.
28. Gendreau M, Laporte G, Semet F. The maximal expected coverage relocation problem for emergency vehicles. *J Oper Res Soc* 2006;57:22–28.
29. Berman O. Dynamic repositioning of indistinguishable service units on transportation networks. *Transport Sci* 1981;15:115–136.
30. Berman O. Repositioning of 2 distinguishable service vehicles on networks. *IEEE Trans Syst Man Cybern* 1981;11:187–193.
31. Berman O. Repositioning of distinguishable urban service units on networks. *Comput Oper Res* 1981;8:105–118.
32. Berman O, LeBlanc B. Location-relocation of mobile facilities on a stochastic network. *Transport Sci* 1984;18:315–330.
33. Berman O, Rahnama MR. Optimal location-relocation decisions on stochastic networks. *Transport Sci* 1985;19:203–221.
34. Puterman ML. *Markov decision processes: discrete stochastic dynamic programming*. New York: John Wiley & Sons, Inc.; 1994.
35. Zhang O, Mason AJ, Philpott AB. Simulation and optimisation for ambulance logistics and relocation. In: *Presentation at the INFORMS 2008 Conference*; Washington, (DC): 2008.
36. Restrepo M. *Computational methods for static allocation and real-time redeployment of ambulances* [PhD thesis]. Ithaca (NY): Cornell University; 2008.

37. Maxwell MS, Restrepo M, Henderson SG, *et al.* Approximate dynamic programming-based ambulance redeployment. *INFORMS J Comput* 2010;22:266–281.
38. Bertsekas DP, Tsitsiklis JN. *Neuro-dynamic programming*. Belmont (MA): Athena Scientific; 1997.
39. Restrepo M, Henderson SG, Topaloglu H. Erlang loss models for the static deployment of ambulances. *Health Care Manage Sci* 2009;12:67–79.
40. Henderson SG, Meyn SP, Tadić V. Performance evaluation and policy selection in multiclass networks. *Disc Event Dyn Syst* 2003;13:149–189.
41. Channouf N, L'Ecuyer P, Ingolfsson A, *et al.* The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Manage Sci* 2007;10:25–45.

OPTIMAL MONITORING STRATEGIES

IGOR KABASHKIN
Transport and Telecommunication
Institute, Riga, Latvia

TYPES OF MONITORING MISSIONS

To operate and support a system, different types of monitoring activities must be supported. Within the frame of reliability study these activities bring into correlation with diagnostic missions. The main tasks of these diagnostic missions are as follows [1]: verification of operational readiness, fault detection and characterization, fault isolation, diagnosis and repair, and other maintenance actions.

Verification of Operational Readiness

The *verification of operational readiness* is the process that verifies to the operators of a system the system's capability to meet the requirements of their operational mission. This process typically begins with system power-on and ends with the operator's assurance that the system is operable. Intermediate steps consist of a number of off-line diagnostics that run as expeditiously as possible with minimal operator involvement. These diagnostics may not be comprehensive in that they may not isolate failures, but they are complete in that all operational data and functional paths are verified.

Fault Detection and Characterization

The system commences operation after operational readiness has been achieved. *Fault detection and characterization* is an ongoing diagnostic activity to detect possible faults in the functioning system. The goal is to detect the presence of faults and to characterize their effects so that the operator can determine if the mission is affected. If the mission is affected, the system may

continue to operate in a degraded mode or it may be declared inoperable. If the system has failed, the operator executes a fault isolation and repair process to restore the system to operation.

Fault Isolation

Fault isolation is the next logical step in the fault detection and characterization process. The goal of fault isolation is to locate the failure within a single or small group of replaceable items, which if replaced, restores the system to operation. Diagnostic hardware and software may be aided by historical data and procedural activities to narrow the field of possible failure sources and focus on the diagnosis and repair activities. The correlation between the faults and the item or items that have failed can be very complex.

Diagnosis and Repair

After a failed item has been isolated, removed from the system, and replaced, it must be evaluated for repairability, repaired, or discarded if it is not repairable, and the repair verified for correct operation. The first and last of these steps are performed by the comprehensive tests available in the integrated diagnostics environment. In case these integrated diagnostics are insufficient to isolate the failure to a repairable unit (as opposed to a replaceable unit), an additional manual or automated test equipment (ATE) and custom diagnostics are employed. Manual, ATE, or custom diagnostic activities typically take place in either intermediate or depot maintenance support facilities.

Other Maintenance Actions

Systems will typically need *other maintenance actions* that must take place either periodically or on an event-driven basis. These activities are most likely associated with installation, configuration, or reconfiguration tasks, or are associated with ongoing calibration and alignment activities.

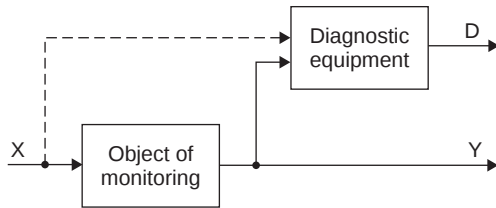


Figure 1. The structure of functional diagnostic system.

TECHNICAL REALIZATION OF DIAGNOSTIC STRATEGIES

There are two main types of diagnostic systems:

- functional diagnostics,
- test diagnostics.

The structure of a *functional diagnostic system* is shown on Fig. 1. An object of monitoring is operated in standard condition with real input signal X , real output signal Y , and in some cases input signal X is received by diagnostics equipment. The result of the diagnosis is formulated according to the analysis of X and Y signals.

The advantage of this type of diagnostic structure is the possibility to use a real operating signal without interruption of work for monitoring of an object.

The weakness of a functional diagnostic system is the impossibility of using special test signals with special parameters for more effective diagnosis.

The structure of a *test diagnostic system* is shown in Fig. 2. The object has two states switching through the commutator with two positions: “1”—normal operation, “2”—test operation. In the test status a special test signal from test generator comes to the monitoring object. Diagnostic equipment generates result of diagnosis based on object reaction.

The advantage and weakness of test diagnostic system are contrary to that of the functional diagnostic system, that is, the possibility to use a test signal with special parameters and the necessity to switch off the object from normal operation during the monitoring process.

OPTIMIZING MONITORING STRATEGIES

Monitoring strategy optimization requires the following questions to be asked:

- What are the goals for which test strategies are being optimized?
- How can the degree of optimization be assessed?
- What can be done to optimize a test strategy?
- How do diagnostic tools help perform optimization?

Monitoring strategy optimization is achieved by focusing on the overall “diagnosability” of a system or device during development. In the most limited sense, test strategy optimization refers to ensuring that tests are performed in an order that best serves development and maintenance goals. In a broader sense, optimization may encompass not only the ordering of tests, but also the enhancement of the hardware’s inherent capacity to facilitate and support effective diagnostics [2].

Optimization can serve a number of goals, including reducing cost of ownership, improved availability/readiness, reducing support environment requirements, and greater safety.

Optimization with multiple goals can be at cross-purposes. For example, improving availability may result in increased life cycle cost. Finding solutions that satisfy multiple goals can be achieved incrementally by iteratively improving diagnostic capability. The most out-of-range goals are used to provide limits to drive new cases for any iteration, thereby converging on a solution.

The degree to which case studies help to meet goals is often determined by examining familiar testability, reliability, and maintainability measures, such as fault detection/fault isolation levels, system mean time between failures (MTBF), mean cost/time to repair, likelihood of critical events, and inherent availability.

While diagnostic optimization may lead to more fundamental design changes (test point placement, partitioning, etc.), development costs can sometimes be reduced by first

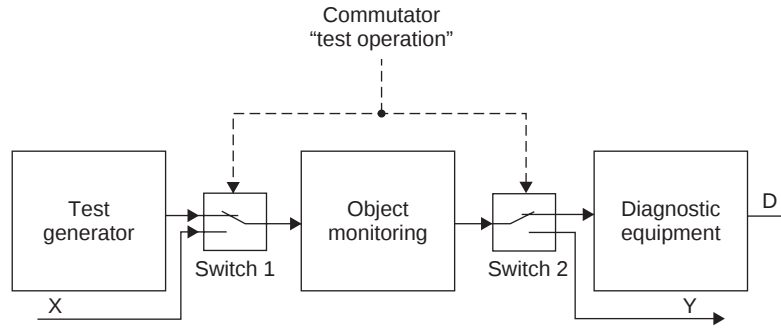


Figure 2. The structure of test diagnostic system.

focusing on changes to testing, such as refining test selection/ordering criteria, additional test procedures (automated or otherwise), and changing test implementation.

It is important to recognize when changes to testing result in diminishing returns—an indication that design changes should be considered, or that the proposed changes to testing are ineffective. More sophisticated testing can involve greater development costs, often overriding any benefits, which lead to its consideration. Similar changes made earlier in the design’s development, however, can achieve much of the same optimization at a substantially reduced cost. Optimization is about finding the balance between seemingly competing criteria.

A robust engineering environment requires a powerful diagnostics assessment and optimization tool. Considering properly, the

trade-offs between design changes and testing, further requires a tool that incorporates systems engineering methodology.

CASE STUDIES

Optimal Monitoring of the System with Continuous Operation for Maximum Availability

Every component, equipment, or system has an intrinsic design capability (Fig. 3). The demand or the expected value may be below this level.

Over time the curve of capability will decrease due to fouling, wear, fatigue, or chemical attack. When this happens some maintenance has to be done to bring the capability up to the acceptable design level.

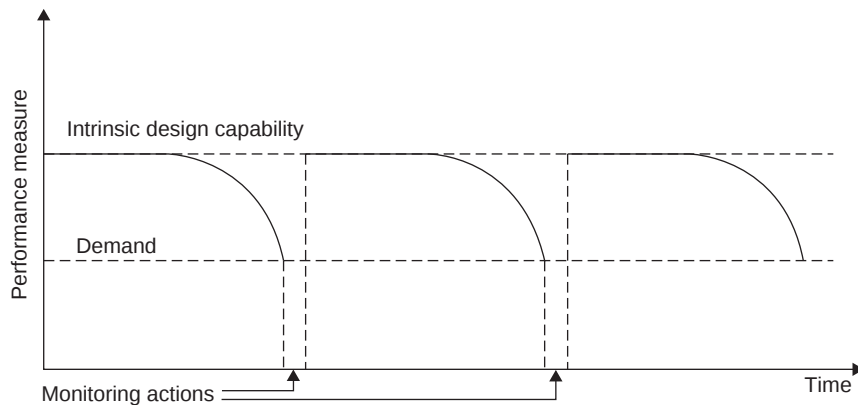


Figure 3. Effect of demand fluctuations on maintenance timing.

Unfortunately, not all components of the system have the built-in test equipment. In this case periodical test actions are required.

Markov's state transition diagram for such a system is shown in Fig. 4, where

- $\lambda = 1/T_0$ = failure rate of the system with MTBF T_0 ,
- $\mu = 1/T_R$ = repair rate with mean time to repeat T_R
- $T_m = 1/\lambda_m$ = mean of exponentially distributed monitoring times with parameter μ_m
- $\tau_m = 1/\mu_m$ = mean of exponentially distributed duration of monitoring with parameter μ_m ,
- $0 \leq a \leq 1$ = testing part of monitoring system (with the built-in test equipment).

The behavior of the examined system is described by the state transitions (Fig. 4): H_0 —state in which system is operating and available for use; H_1 —failure is in the part of the system with monitoring of the built-in test equipment; H_2 —failure is in the non-monitoring part of the system; H_3 —failure in the nontested part of the equipment is being repaired after manual monitoring; and H_4 —system is without failures and under manual monitoring.

On the base of the state transition diagram for a Markov's process shown in Fig. 4, we can write the system of Chapman–Kolmogorov's equations according to the general rule for a stationary process [3]—a set of equations has for each state H_i on the left-hand side the equilibrium, a probability of a given state P_i multiplied by the total rate of all departures

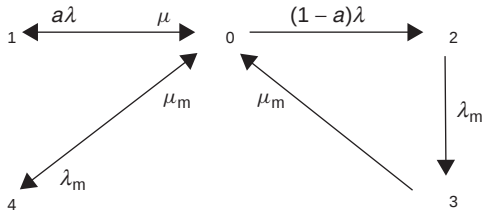


Figure 4. Markov's state transition diagram for system with continuous operation, built-in-test equipment and manual monitoring.

translating from this state, and on the right-hand side, the sum of rates of all arrivals translating into the i th state multiplied by the probabilities of the states; consequently,

$$\begin{aligned}\mu_1 P_1 &= a\lambda P_0 \\ \lambda_m P_2 &= (1-a)\lambda P_0 \\ \mu_m P_3 &= \lambda_m P_2 \\ \mu_m P_4 &= \lambda_m P_0.\end{aligned}$$

The normalizing condition is $P_0 + P_1 + P_2 + P_3 + P_4 = 1$.

After transformation of the above-mentioned set of equations we can obtain the value for $P_i (i = 1 \dots 4)$ via P_0 :

$$\begin{aligned}P_1 &= \frac{a\lambda}{\mu} P_0 \\ P_2 &= \frac{(1-a)\lambda}{\lambda_m} P_0 \\ P_3 &= \frac{\lambda_m}{\mu_m} P_2 = \frac{(1-a)\lambda}{\lambda_m} P_0 \\ P_4 &= \frac{\lambda_m}{\mu_m}.\end{aligned}$$

Value of P_0 can be obtained by replacement $P_i (i = 1 \dots 4)$ in the normalizing equation $\sum_{i=0}^4 P_i = 1$:

$$\begin{aligned}P_0 + \frac{a\lambda}{\mu} P_0 + \frac{(1-a)\lambda}{\lambda_m} P_0 \\ + \frac{(1-a)\lambda}{\mu_m} P_0 + \frac{\lambda_m}{\mu_m} P_0 &= 1 \\ P_0^{-1} &= 1 + \frac{a\lambda}{\mu} + \frac{(1-a)\lambda}{\lambda_m} \\ &+ \frac{(1-a)\lambda}{\mu_m} + \frac{\lambda_m}{\mu_m}.\end{aligned}$$

Let us define the unavailability U of the examined system.

Unavailability can be defined as the probability that an item will not operate correctly at a given time and under specified conditions [4]. Mathematically, unavailability is 1 minus the availability A .

For the examined system $A = P_0, U = 1 - A = P_1 + P_2 + P_3 + P_4$.

For a highly reliable system with $\lambda \ll \mu, \lambda_m \ll \mu_m$ approximately $P_0 \rightarrow 1$.

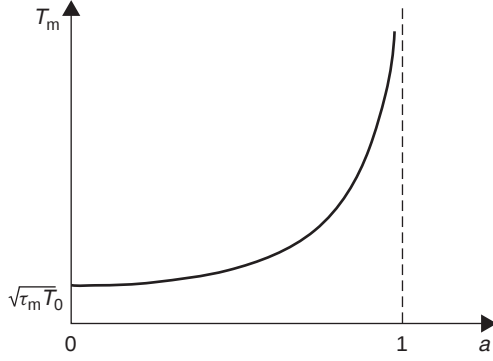


Figure 5. The graphic of function $T_m = f(a)$.

In this case the system unavailability U is defined as follows:

$$U = \sum_{i=1}^4 P_i \cong \frac{a\lambda}{\mu} + \frac{(1-a)\lambda}{\lambda_m} + \frac{(1-a)\lambda}{\mu_m} + \frac{\lambda_m}{\mu_m}. \quad (1)$$

By differentiating Equation (1) with respect to λ_m , we get

$$\frac{dU}{d\lambda_m} = -\frac{(1-a)\lambda}{\lambda_m^2} + \frac{1}{\mu_m}. \quad (2)$$

Setting Equation (2) equal to zero we get optimal value of λ_m

$$\frac{1}{\lambda_m^2} = \frac{1}{\mu_m(1-a)\lambda}. \quad (3)$$

From Equation (3) we can obtain an expression for optimal monitoring time $T_m = 1/\lambda_m$:

$$T_m^2 = \frac{\tau_m T_0}{1-a} \text{ or } T_m = \sqrt{\frac{\tau_m T_0}{1-a}}. \quad (4)$$

The graphic of function $T_m = f(a)$ is shown in Fig. 5.

For the special case of a full monitoring system $a \rightarrow 1$, the monitoring time $T_m \rightarrow \infty$. This means that if the system has a built-in test equipment for 100% monitoring, the manual monitoring of such a system is not needed.

The system has $T_0 = 2500$ h, $\tau_m = 2$ h. Half of the system is under monitoring by the built-in test equipment ($\alpha = 0.5$). In accordance with Equation 4 $T_m = \sqrt{2 \times 2 \times 2500} = 100$ h.

For optimal results (the highest availability), approximately one inspection should be performed every 4 days.

Optimal Monitoring of the System with Periodical Sessions of Operation

Periodical sessions of operation are typical for professional communication channels (CCs); for example, during conversations between the pilot and the controller in an Air Traffic Control (ATC) system [5].

The following symbols have been used to develop equations for the model:

- λ = failure rate
- μ = repair rate
- A = availability
- h_i = probability of being in state H_i
- T_m = periodicity of test operations with parameter of Poisson's flow
- $\omega = 1/T_m$
- t_m = time of test operations
- τ = parameter of exponential distribution of t_m
- t_{f1} = time of failure fixation by human operator
- $v_1 = 1/t_{f1}$ = parameter of exponential distribution of t_{f1}
- t_{f2} = time of failure fixation by automatic test system
- $v_2 = 1/t_{f2}$ = parameter of exponential distribution of t_{f2}
- T_c = periodicity of communication demands with parameter of Poisson's flow $\varphi = 1/T_c$
- t_c = time of communication session
- $\psi = 1/t_c$ = parameter of exponential distribution of t_c
- t_n = mean time of test interruption in the case of communication session begins
- $v_3 = 1/t_n$ = parameter of exponential distribution of t_n
- t_R = time of repeat demand on communication in the channel with failure

η = parameter of exponential distribution of t_R

Strategy 1: Communication Channel Does Not Have the Built-In Test Equipment. In the case when the test equipment is absent, the CC failure is detected by the operator during the process of operation.

In this case, it is possible to clarify the process of the CC failure detection by the time diagram (Fig. 6). The CC with an operational set of the equipment starts functioning at time t_0 . After some operation time T_0 , the failure of the main set of the communication equipment happens at the moment of time t_1 . However, due to the absence of an automatic test system, the fixation of the failure does not happen. After some time T_c , at the moment t_2 , the operator receives the requirement by a communication, for example, in the form of a request from the transport object. The operator does not know about the malfunction of the operation of the CC. The operator transmits the answer for the request to the transport object. In this case, the object does not receive the answer and after some time T_R repeats the request. The delivery of the repeated request in moment t_3 is the information about the malfunction of CC to the operator. In such a case, the operator goes over to the CC with the reserve on-ground communication equipment, fixing the failure of the main set of the equipment. A similar process iterates in the case of the refusal of the reserve radio station too.

The behavior of the examined system is described by the state transition diagram (Fig. 7), where H_1 —completely good state of the CC equipment; H_2 —communication session in CC; H_3 —failure of the main CC system; H_4 —communication session in the CC with failure of the main communication radio aid; H_5 —repeat demand on communication in CC with failure of equipment;



Figure 6. Time diagram of failure detection for test strategy 1.

and H_6 —detection of the failure state of the main communication radio aid, switch over to the reserve one.

Solving forward Kolmogorov's differential equation system by describing the operation of the examined channel it is possible to define A_1 availability of the channel with the first test strategy:

$$A_1 = (1 + a_1)/(1 + a_1 + a_2), \quad (5)$$

where

$$a_1 = \lambda[\gamma_1(1 + \beta) + 1/\varphi], \quad a_2 = \beta,$$

$$\gamma_1 = 1/\mu + 1/\eta + 1/\nu_1, \quad \beta = \varphi/(\lambda + \psi).$$

Comparing Equation (5) with the equation for the A_0 availability of the ideal system (immediate failure detection, immediate switchover, nonfailure switch) [6], it is possible to come to a conclusion that $A_1 = A_0$ with nonlimited increase of the intensity of communication φ and the reduction of the failure detection time ($\nu_1 \rightarrow \infty$).

It is possible to evaluate the deterioration of the reliability in the real CC in comparison with the ideal one with the help of the reliability deterioration coefficient

$$V_1 = (1 - A_1)/(1 - A_0).$$

The function of influence of average time between communication session T_c and average switch time t_{f1} for the CC are given (see in Fig. 8 line 1 for $t_{f1} = 2$ min and line 2 for $t_{f1} = 10$ s).

Strategy 2: Communication Channel Has the Built-In Test Equipment with Diagnosis Procedures during Communication Sessions.

At present many of the communication systems have the built-in test equipment with diagnosis procedures during communication

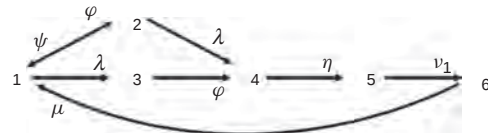


Figure 7. State transition diagram for test strategy 1.

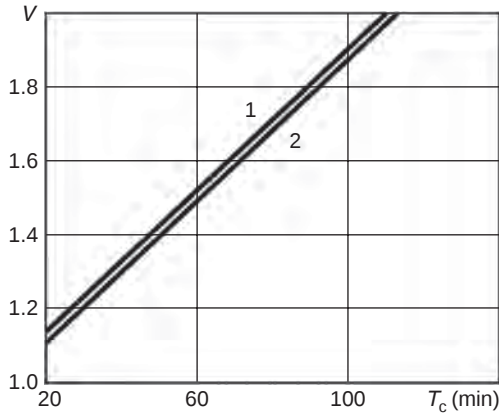


Figure 8. Deterioration of the reliability in the real communication channel with test strategy 1.

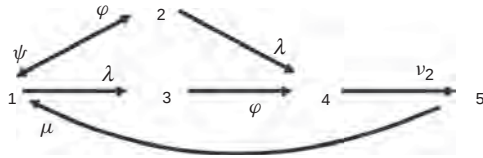


Figure 9. State transition diagram for test strategy 2.

sessions. The behavior of the examined system with the mentioned test strategy can be described by the graph of states submitted in Fig. 9, where the symbols of states entered earlier are saved.

The availability A_2 of the channel with the second test strategy may be defined by solution of the Kolmogorov's equation system, which describes the graph mentioned previously (Fig. 9):

$$A_2 = (1 + a_1)/(1 + a_1 + a_2),$$

where

$$a_1 = \lambda[\gamma_2(1 + \beta) + 1/\varphi], \quad a_2 = \beta, \\ \beta = \varphi/(\lambda + \psi), \quad \gamma_2 = 1/\mu + 1/\nu_2.$$

The analysis of the reliability deterioration coefficient $V_2 = (1 - A_2)/(1 - A_0)$ shows the dependency similar to the one submitted in Fig. 8.

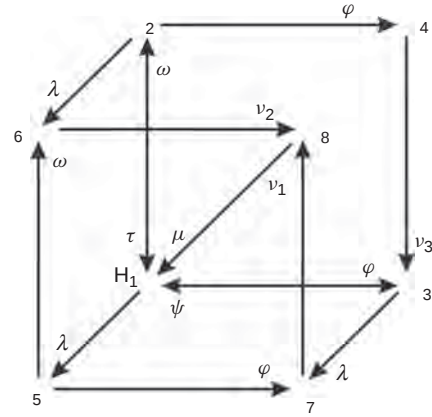


Figure 10. State transition diagram for test strategy 3.

Strategy 3: Communication Channel Has a Periodical Test in the Pauses between Communication Sessions. In this case the process of CC operation can be described by the following states: H_1 —good condition of CC equipment, absence of the communication session and test operations; H_2 —test in the channel without communication session; H_3 —communication session; H_4 —demand on communication in the test period; H_5 —failure of CC equipment when the communication is paused; H_6 —test of system with failure; H_7 —demand on communication in the channel with failure of equipment; H_8 —failure fact fixation by automatic test system or by human operator. There is a state transition diagram of the examined system showed in Fig. 10. The availability A_3 of the examined CC is defined by equation

$$A_3 = \sum_{i=1}^3 h_i.$$

By forming and solving the Kolmogorov's equation system describing the graph of the states, it is possible to define probabilities $h_i, i = 1, \dots, 8$, and the availability A_3 :

$$A_3 = (1 + a_1)/(1 + a_1 + a_2),$$

where

$$\begin{aligned}
 a_1 &= \omega/(\lambda + \varphi + \tau)[\lambda(1/\mu + 1/v_2) + \varphi/v_3 \\
 &\quad + \lambda\varphi/(\lambda + \psi)(1/\mu + 1/v_1)] \\
 &\quad + \lambda/(\omega + \varphi)[1 + \omega(1/\mu + 1/v_2) \\
 &\quad + \varphi(1/\mu + 1/v_1) + \varphi/(\lambda + \psi)(1/\lambda + v_1)] \\
 a_2 &= (\lambda + \varphi)^{-1}[\omega(\lambda + \varphi + \psi)/(\lambda + \varphi + \tau) + \varphi].
 \end{aligned}$$

The reliability deterioration of the real CC may be evaluated by the coefficient

$$V_3 = (1 - A_3)/(1 - A_0).$$

The analysis of the $V_3(\omega)$ shows that it is a unimodal function with the extremes in ω_{opt} point.

The expression for the definition of $T_{m,\text{opt}} = 1/\omega_{\text{opt}}$ is possible to find out from the condition $dA_3/d\omega = 0$. In a general case, the expression for the definition $T_{m,\text{opt}}$ appears quite cumbersome, but for the practical crucial case of the highly reliable systems ($\lambda \ll \mu$) with a short time for the failure fixation ($\mu \ll v_1, \mu \ll v_2$), the following approximate expression is justified for the definition of the optimal test periodicity:

$$\begin{aligned}
 T_{m,\text{opt}}^{-1} &= v_3\varphi^{-1}\{\lambda + [(\lambda - \varphi^2/v_3)^2 \\
 &\quad + \lambda\varphi\tau/v_3]^{1/2}\} - \varphi.
 \end{aligned}$$

Let us examine the CC availability of an ATC system in an airport area. The controller's channel radio stations are operating in the day/night operation regime. The sessions of communication are carried out at random moments of time. The failures of the radio stations could be fixed by the objective technical test aids executing the periodical test, or subjectively according to the operator's evaluation. There are $V(T_m)$ dependencies for various MTBF T_0 and typical average meanings of the characteristics of the ATC CC showed in Fig. 11. There are functions $T_{m,\text{opt}}(T_0)$ for the various T_c periodicity of communication session and time of test interruption t_n showed in Figs 12 and 13.

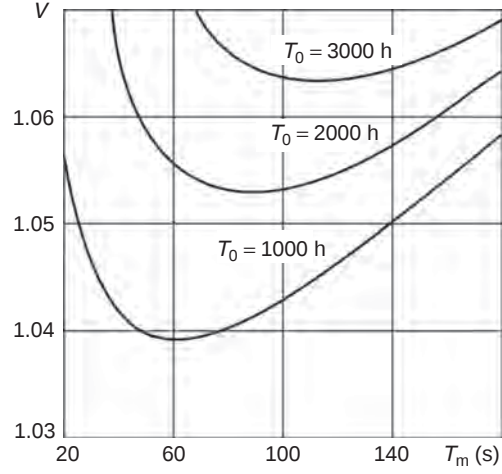


Figure 11. Functions $V(T_m)$ for test strategy 3.

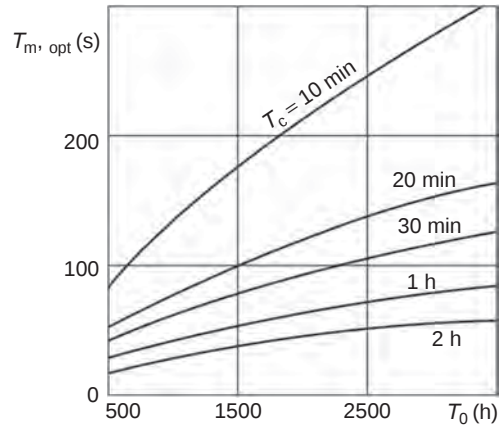


Figure 12. Functions $T_{m,\text{opt}}(T_0, T_c)$.

Optimal Inspection Frequency in order to Maximize Profit

The model is developed on the premise that the equipment under repair leads to zero output, thus less profit. Furthermore, if the equipment is inspected too often, there is a danger that there may be an increase in costs due to factors such as loss of production, cost of materials, and wages than losses due to breakdowns. The following assumptions are associated with this model [7]:

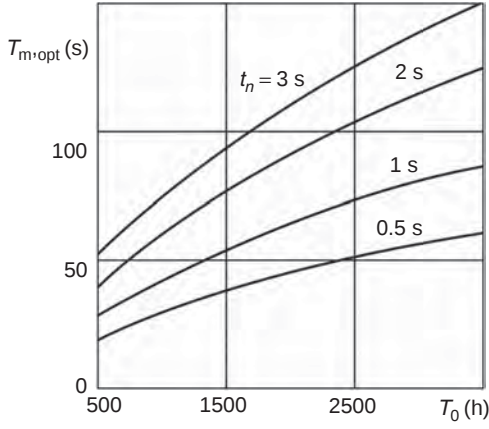


Figure 13. Functions $T_{m, \text{opt}}(T_0, t_n)$.

- The equipment failure rate is a function of inspections.
- Times of inspection are exponentially distributed.
- Equipment failure and repair rates are constant.

The following symbols are used to develop equations for the model:

- n = number of inspections performed per unit of time
- $1/\theta$ = mean of exponentially distributed inspection times
- p = profit at no downtime losses
- C_i = average inspection cost per uninterrupted unit of time
- C_r = average cost of repair per uninterrupted unit of time
- $\lambda(n)$ = equipment failure rate for maintenance strategy with n number of inspections
- μ = equipment repair rate

Profit per unit of time is expressed by Dhillon [7] as

$$\begin{aligned} \text{PR} &= p - \text{PL}_i - \text{PL}_r - \text{IC} - \text{RC} \\ &= p - \frac{pn}{\theta} - \frac{p\lambda(n)}{\mu} - \frac{nC_i}{\theta} - \frac{C_r\lambda(n)}{\mu}, \quad (6) \end{aligned}$$

where

PL_i = production output value loss per unit of time due to inspections

PL_r = production output value loss per unit of time due to repairs

IC = inspection cost per unit of time

RC = repair cost per unit of time.

By differentiating Equation (6) with respect to n and then equating it to zero yield

$$\frac{d\text{PR}}{dn} = -\frac{p}{\theta} - \frac{p}{\mu} \frac{d\lambda(n)}{dn} - \frac{C_i}{\theta} - \frac{C_r}{\mu} \frac{d\lambda(n)}{dn} = 0. \quad (7)$$

Rearranging Equation (7), we get

$$\begin{aligned} \frac{d\lambda(n)}{dn} &= -\left[\frac{1}{\theta}(p + C_i)\right] / \left(\frac{p}{\mu} + \frac{C_r}{\mu}\right) \\ &= -\frac{\mu(p + C_i)}{\theta(p + C_r)}. \end{aligned} \quad (8)$$

The value of n will be optimal when the left and right sides of Equation (8) are equal. At this point, the profit will be at its maximum value.

Assume the failure rate of a manufacturing system is defined as $\lambda(n) = \lambda_0 e^{-n}$, where λ_0 is the system failure rate at $n = 0$.

Let us develop an expression for the optimal value of n with the aid of Equation (8); using $\lambda(n) = \lambda_0 e^{-n}$ in Equation (8) yields

$$-\lambda_0 e^{-n} = -\frac{\mu(p + C_i)}{\theta(p + C_r)} \quad (9)$$

and by rearranging Equation (9), we get

$$n^* = \ln \left[\frac{\lambda_0 \theta (p + C_r)}{\mu (p + C_i)} \right], \quad (10)$$

where

n^* = optimal manufacturing system inspection frequency.

Suppose that we have the following data:

p = \$10,000 per month

$\lambda_0 = 2$ failures per month
 $\mu = 25$ repairs per month
 $\theta = 100$ per month
 $C_i = \$75$ per month
 $C_r = \$400$ per month.

Determine the optimal value of n by using Equation (10).

By inserting the specified data values into Equation (10), we obtain

$$n^* = \ln \left[\frac{2 \times 100 \times (10,000 + 400)}{25 \times (10,000 + 75)} \right] = 2.11.$$

For optimal performance, approximately two inspections per month should be performed [7].

CONCLUSIONS

Monitoring strategies are a key element of reliability-centered maintenance (RCM), which is the most proven proactive approach for developing maintenance programs from scratch. RCM includes preventive, predictive, and detective (failure finding) activities. Preventive maintenance includes activities that are scheduled and executed to replace components or restore them to their original condition from their apparent condition at the time. Predictive maintenance also known as *condition-based maintenance* or *on-condition maintenance* includes activities performed to look for signs of impending failures so that corrective maintenance can be done before equipment functionality is lost. RCM programs provide balance of cost versus reliability that is tailored specifically to each system in its operating environment. Maintenance optimization arises from the need to improve maintenance programs that are failing to meet the desired performance expectations. *Decision optimization* refers to the use of any one of several optimization tools. These tools are used to make decisions about component replacement, repair versus replace, inspection intervals, and repairs. Optimization tools enhance reliability and reduce costs. They add value to the output of RCM and may lead to modifications in

RCM output decisions. There are many mathematical approaches to solving typical maintenance problems and monitoring tasks. Many of them have come from the field of operations research. Some of the problems that can be solved include [8] optimizing of the following:

- the age at which equipment replacements are made using various applications and operating cost profiles to determine the most cost-effective age for replacement;
- the decision for the time swap duty and standby equipment combinations to achieve maximum backup availability and reduce the risk of multiple failures;
- the decision as to when to replace capital equipment to maximize discounted cash flow benefits or to minimize total cost;
- intervals between preventive replacements to achieve target availability;
- the time for group replacements of large numbers of identical assets instead of replacement as they fail in order to minimize overall cost;
- inspection frequencies to maximize profits, minimize downtime, and maximize availability of emergency equipment;
- decision making related to overhaul and repair versus replacement policies and cost limits.

Many maintenance problems can be solved using data analysis tools with special software [9–11]. There are many of these software tools available, most are stand-alone programs that can be interfaced or integrated with corporate databases. Several collect condition-monitoring signals that are input from field sensors and use that data in real time with history data and economic inputs to produce recommendations either to run or not run equipment or to shut it down for work. Another powerful modern technique is simulation modeling of the complex systems [12,13]. These models can show the effect of various monitoring strategies and reliability increase at

different points in the systems and help focus engineering improvement efforts more effectively.

Continuous improvement of maintenance management will be possible due to the utilization of emerging techniques and technologies in areas that are considered to be of higher impact as a result of the previous steps of the management process [14].

Regarding the application of new technologies to maintenance, the “e-maintenance” concept is put forward as a maintenance support, which includes the resources, services, and management necessary to enable proactive decision process execution. This support not only includes e-technologies but also e-maintenance activities (operations or processes) such as e-monitoring, e-diagnosis, and e-prognosis [15].

References 16–21 present a set of techniques that can be used to improve involvement of maintenance personnel in maintenance improvement programs, and a review of new technologies arising in maintenance, respectively.

REFERENCES

1. Marz TF. Integrated diagnostics: operational missions, diagnostic types, characteristics, and capability gaps. Technical Note CMU/SEI-2005-TN-035; 2005 Sep. pp. 3–5.
2. Optimizing Test Strategies by Testability.com. Published on 7/6/2004. Available at http://articles.testability.com/Articles/2/1/Optimizing_Test_Strategies.aspx
3. Wentzel ES. Operations research: a methodological approach: Mir Publisher; 1983. pp. 159–161.
4. Wikipedia. Available at <http://en.wikipedia.org/wiki/Unavailability>
5. Kabashkin I. Optimal monitoring strategy for communication channel of intelligent transport systems with periodical sessions of activity. In: Proceedings of the International Conference on Operation Research, Simulation and Optimisation in Business and Industry; 2006 May 17–20; Tallinn, Estonia. Kaunas: Technologia; 2006. pp. 201–206.
6. Barlow BE, Proshan F, Hunter LC. Mathematical theory of reliability. New York: John Wiley & Sons, Inc.; 1965.
7. Dhillon BS. Engineering maintenance: a modern approach. Boca Raton (FL): CRC Press; 2002. pp. 56–58.
8. Campbell JD, Reyes-Picknell JV. Uptime: strategies for excellence in maintenance management. New York: Productivity Press; 2006. p. 358.
9. Relex Reliability Software Tools. Available at <http://www.relex.com>
10. Simulation Software for Probabilistic Event and Risk Analysis. Available at <http://www.reliasoft.com>
11. Reliability Software and Safety Analysis Tools. Available at <http://www.item-software.com>
12. Knopov PS, Pardalos PM. Simulation and optimization methods in risk and reliability theory: Nova Science Publishers; 2008. p. 285.
13. Lawday G, Ireland D, Edlund GA. Signal integrity engineer's companion: real-time test and measurement and design simulation: Prentice Hall PTR; 2008. p. 496.
14. Blanchard BS, Verma D, Peterson EL. Maintainability: a key to effective serviceability and maintenance management. New York: John Wiley & Sons, Inc.; 1995. p. 537.
15. Marquez AC. The maintenance management framework: models and methods for complex systems maintenance. London: Springer; 2007. p. 333.
16. Wireman T. Developing performance indicators for managing maintenance. New York: Industrial Press; 2005. p. 250.
17. Bloom NB. Reliability centered maintenance. New York: McGraw-Hill Inc.; 2006. p. 292.
18. Levitt J. The handbook of maintenance management. New York: Industrial Press; 2009. p. 500.
19. Moubray J. Reliability-centered maintenance. New York: Industrial Press Inc.; 1997. p. 423.
20. Ebeling CE. An introduction to reliability and maintainability engineering. New York: The McGraw-Hill Inc.; 1997. p. 486.
21. Narayan V. Effective maintenance management: risk and reliability strategies for optimizing performance. New York: Industrial Press; 2004. p. 246.

OPTIMAL RELIABILITY ALLOCATION

YASHWANT K. MALAIYA
Computer Science Department,
Colorado State University,
Fort Collins, Colorado

INTRODUCTION

A large system is implemented by using a set of interconnected subsystems. While the architecture of the overall system is often fixed because of the functional requirements, implementation choices are generally available for individual subsystems. A designer needs to achieve the target reliability while minimizing the total cost, or alternatively maximize the reliability while using only the available budget. Intuitively, some of the lowest reliability components may need special attention to raise the overall reliability level. Such optimization problems arise while designing a complex software or a hardware system. Such problems also arise in mechanical or electrical systems. A number of studies since 1960 have examined some of the computational aspects of such problems [1].

All the subsystems are essential in a nonredundant system. However, an individual subsystem can often be made more reliable by using a more costly implementation. This additional cost may represent wider columns in a building or more thorough testing of software. In redundant implementations, higher reliability can sometimes be achieved by using several copies of a subsystem, such that the system incorporates a *parallel* or a *k-out-of-n* configuration.

The problem formulation is considered in the next section, followed by approaches used for setting up an optimization problem. As a detailed example, software reliability allocation is examined in detail with a numerical illustration. The last section considers reliability allocation in general complex systems.

ALLOCATION AS AN OPTIMIZATION PROBLEM

We consider a system that has been designed at a higher level as an assembly of appropriately connected subsystems. The functionality of each subsystem is often unique; however, there can be several choices for many of the subsystems that provide the same functionality but differently reliability levels.

Here we consider the problem formulation for a common and widely applicable case. Let there be n subsystems $SS_i, i = 1, \dots, n$, each with reliability R_i and cost C_i . Let the cost C_i be a function of the reliability given by $f_i(R_i)$, this is generally referred to as the *cost function*. Let C_s and R_s represent the total system cost and the overall reliability and let R_{ST} be the specified target reliability. If all the subsystems are essential to the system and if their failures are statistically independent, the system can be modeled as a *series system*. The cost minimization problem can be stated as

$$\begin{aligned} \text{Minimize} \quad & C_s = \sum_{i=1}^n C_i = \sum_{i=1}^n f_i(R_i) \end{aligned} \quad (1)$$

$$\begin{aligned} \text{Subject to} \quad & R_{ST} \leq R_s \\ \text{i.e., } & R_{ST} \leq \prod_{i=1}^n R_i \quad \text{since} \\ & R_s = \prod_{i=1}^n R_i. \end{aligned} \quad (2)$$

Note that Equation (1) assumes that the cost of interconnecting the subsystems is negligible. An alternative problem would be to maximize R_s while keeping C_s less than or equal to the allocated cost budget.

Depending on the type of the system, the i th subsystem SS_i may have several implementation choices with different reliability values:

1. The subsystem can be made more reliable by extending a continuous attribute (e.g., diameter of a column in building or time spent for software testing).
2. Different vendors may offer their own implementations of SS_i at different costs.
3. It may be possible to use multiple copies of SS_i (e.g., double wheels of a truck) to achieve higher reliability. Often the number of copies is constrained between a minimum (often one) and a maximum number because of implementation issues.

In the first case, both cost and the reliability can be varied continuously, whereas in the other two cases, the choices are discrete. In the first case, we can define a continuous cost function. In the second case too, the market forces may impose a cost function. In the third case, the subsystem may be modeled as a *parallel* or *k-out-of-n* system for reliability evaluation, provided the failures are statistically independent. A number of publications on reliability allocation consider only the third case, where the optimization problem becomes an integer optimization problem. It becomes a 0–1 optimization problem when choices are discrete and a component from a given list of candidates is either used or not used [2].

THE COST FUNCTION AND PROBLEM SETUP

It is reasonable to assume that the cost function f_i would satisfy these three conditions [3]:

1. f_i is a positive function;
2. f_i is nondecreasing, thus higher reliability will come at a higher cost;
3. f_i increases at a higher rate for higher values of R_i .

The third condition leads to the fact that it can be very expensive to achieve the reliability value of 1. In fact, for software, it has been shown that it is infeasible to achieve ultra-high reliability in software [4]. Note

that the cost function, in general, will also be a function of the relative complexity of the subsystem.

In some cases, the cost function can be derived from basic considerations, as we will do below for software reliability. In other cases, it may be derived empirically by compiling data for different choices. The cost function is often stated in terms of the reliability; for example, the cost function proposed by Mettas [3] is given in terms of the maximum and minimum achievable reliability values.

In some cases, there are choices that can be made that can enhance reliability with little or no additional cost, sometimes by using newer technology. For example, some disciplined software development processes have been found to yield more reliable software with the same effort. We assume that such choices have already been made, and optimization is sought by exploiting the available choices of attributes.

The cost function can also be given in terms of the failure rate as illustrated below for software reliability allocation.

A transformation of Equation (2) can be obtained by logarithms of both sides of the equation [5–7]

$$\ln(R_{ST}) \leq \sum_{i=1}^n \ln(R_i). \quad (3)$$

The transformation in Equation (3) can reduce the problem to a linear optimization problem. In some cases, the term $\ln(R_i)$ also has a well-defined physical significance and is termed the *failure rate*. When the failure rate of a subsystem SS_i is constant, its reliability is given by an exponential relationship $R_i(t) = \exp(-\lambda_i t)$, the system failure rate is given by the summation of the subsystem failure rates and hence Equation (3) can be restated as

$$\lambda_{ST} \geq \sum_{i=1}^n \lambda_i. \quad (4)$$

The failure rate is itself a frequently used reliability metric. In some cases such as in software reliability engineering, it is the failure rate that is often specified [8,9]. The cost

function of a subsystem can also be given in terms of its failure rate. If the cost C_i is given by the function $f_i(\lambda_i)$, Equation (1) can be restated as

$$\text{Minimize} \quad C_S = \sum_{i=1}^n C_i = \sum_{i=1}^n f_i(\lambda_i). \quad (5)$$

For example, let us consider software reliability. When a software is tested, defects in it are found and removed by debugging. A program tested more thoroughly will have fewer bugs and hence higher reliability. Several models relating software reliability growth with the time spent in testing, have been proposed and validated. These are termed software *reliability growth models* (SRGMs). For the popular *exponential software reliability growth model* [8–10], the failure rate as a function of testing time d is given by

$$\lambda_i(d) = \lambda_{0i} \exp(-\beta_i d),$$

where λ_{0i} and β_i are the model parameters. It should be noted that λ_{0i} depends on initial defect density and β_i depends inversely on program size [11]. If we assume that the variable cost is dominated by the testing time, the cost is given by the following function, which satisfies the three conditions mentioned above

$$d(\lambda_i) = \frac{1}{\beta_i} \ln \left(\frac{\lambda_{0i}}{\lambda_i} \right). \quad (6)$$

RELIABILITY ALLOCATION FOR BASIC SERIES AND PARALLEL SYSTEMS

The reliability allocation problem for two basic reliability structures such as *series* and *parallel* can be solved by linearizing the constraints [1,2,7]. In a *series* system, the constraint is given in Equation (2) above, which can be linearized by rewriting it as given in Equation (3) above, which may then be solved relatively easily. An example is given below for software reliability allocation.

In a *parallel system*, functionally identical subsystems are configured such that correct operation of at least one of them assures a correctly functioning system. It is assumed

that any overhead in implementing such a system is negligible. In real systems, the overhead involved will result in a lower level of reliability. In a number of studies, the problem assumes that the reliability of a subsystem can be increased by using functionally identical components in parallel [2,7]. For parallel systems, the constraint is given by

$$R_{ST} \leq 1 - \prod_{i=1}^n (1 - R_i). \quad (7)$$

The constraint can be linearized by using logarithms of the complements of reliability. Thus, Equation (7) can be rewritten as

$$\ln(1 - R_{ST}) \geq \sum_{i=1}^n \ln(1 - R_i). \quad (8)$$

Elegbede *et al.* have recently shown [7] that if the cost function satisfies the three properties given above, the cost is optimal if the all the parallel components have the same cost. For software, computer hardware, and mechanical systems, the number of discrete parallel component is likely to be very small.

RELIABILITY ALLOCATION IN SOFTWARE SYSTEMS

Let us examine the problem of software reliability allocation [6], which serves as a good illustration. Typically, a software consists of sequentially executed components, where only one of them is under execution at a time. Each component can be independently tested and debugged to reduce the failure rate below a target value. In some cases, the reliability of a component can be further increased by replication. For replication to be effective, each replicated version must be developed independently such that the failures are relatively independent. The impact of replication can be evaluated by assuming statistical independence. However, it has been shown that sometimes the statistical correlation can be significant, requiring analysis that is more complex.

In this example, we consider a nonredundant implementation of software, divided into n sequential components [6], which is a common case. Let us assume that a component i is under execution for a fraction x_i of the time where $\sum x_i = 1$ [5]. Software reliability is often described in terms of the failure rates. Then the reliability allocation problem can be written as

$$\text{Minimize} \quad C = \sum_{i=1}^n \frac{1}{\beta_i} \ln \left(\frac{\lambda_{0i}}{\lambda_i} \right) \quad (9)$$

$$\text{Subject to} \quad \lambda_{ST} \leq \sum_{i=1}^n x_i \lambda_i. \quad (10)$$

Solution: let us solve the problem posed by Equations (9) and (10) by using the Lagrange multiplier approach by finding the minimum of

$$F(\lambda_1, \dots, \lambda_n) = C + \theta(\lambda_1 + \dots + \lambda_n) - \lambda_{ST},$$

where θ is the Lagrange multiplier. The necessary conditions for the minimum to exist are (i) the partial derivatives of the function F are equal to 0, (ii) $\theta > 0$, and (iii) $x_1 \lambda_1 + x_2 \lambda_2 + \dots + x_n \lambda_n = \lambda_{ST}$ [6]. Equating the partial derivatives to 0 and using the third condition, the solutions for the optimal failure rates are found as following:

$$\begin{aligned} \lambda_1 &= \frac{\frac{\lambda_{ST}}{x_1}}{\sum_{i=1}^n \frac{\beta_1}{\beta_i}} \quad \lambda_2 = \frac{\beta_1 x_1}{\beta_2 x_2} \lambda_1 \quad \dots \\ \lambda_n &= \frac{\beta_1 x_1}{\beta_n x_n} \lambda_1. \end{aligned} \quad (11)$$

The optimal testing times of the individual modules, $d_i, i = 1$ to n , are given by Equation (12) below

$$\begin{aligned} d_1 &= \frac{1}{\beta_1} \ln \left(\frac{\lambda_{10} x_1 \sum_{i=1}^n \frac{\beta_1}{\beta_i}}{\lambda_{ST}} \right) \quad \text{and} \\ d_i &= \frac{1}{\beta_i} \ln \left(\frac{\lambda_{i0} \beta_i x_i}{\lambda_1 \beta_1 x_1} \right). \end{aligned} \quad (12)$$

Note that d_i is positive if $\lambda_i \leq \lambda_{i0}$. The testing time for a module must be nonnegative. If any of the testing times obtained using Equation (12) are negative, the optimization problem may have been solved iteratively [6] or using special approaches. Sometimes a reused subsystem has a significantly higher reliability because of past testing, which may result in $\lambda_i \geq \lambda_{i0}$. Since the reused subsystem is already reliable enough, it would make sense to spend any testing effort on the other subsystems.

In software reliability engineering, the assumptions involved in formulation of the exponential model imply that the parameter β_i is inversely proportional to the software size [8,11], measured in terms of the lines of code. In the examples below, we assume that value of x_i is proportional to the subsystem code size. The values of λ_i and λ_{i0} do not depend on size but on the initial defect densities [11] at the beginning of testing. Thus, if the exponential model indeed holds, the Equation (11) states that the optimal values of the post-test failure rates $\lambda_1, \dots, \lambda_n$ are equal. In addition, if the initial defect densities are also all equal for all the components, then the optimal test times for each module is proportional to its size.

The parameter values needed for Equations (12) and (13) may be determined by some initial testing or using empirical relationships.

Example 1. A software system uses five functional components B1–B5. This example has been constructed assuming sizes 1, 2, 3, 10, and 20 KLOC (thousand lines of code) respectively, and the initial defect densities of 10, 10, 10, 15, and 20 defects per KLOC, respectively. Let us assume that measured parameter values are given in the top three rows, which are the inputs to the optimization problem. The solution obtained using Equations (11) and (12) are given in the two bottom rows. The test cost has been minimized so that the overall failure rate is less than or equal to 0.04 per unit time. Here the time units can be person hours of testing time, or hours of CPU time used for testing.

Component	B ₁	B ₂	B ₃	B ₄	B ₅
β_i	4.59×10^{-3}	2.30×10^{-3}	1.53×10^{-3}	4.59×10^{-4}	2.30×10^{-4}
λ_{i0}	0.046	0.046	0.046	0.069	0.092
x_i	0.028	0.056	0.083	0.278	0.556
Optimal λ_i	0.04	0.04	0.04	0.04	0.04
Optimal d_i	30.1	60.1	90.2	1184	3620

Note that the optimal values of λ_i for the five components are equal, even though they started with different initial values. This implies that a substantial part of the test effort is allocated to largest components. The total cost in terms of testing is 4984 unit time.

If the total test time were equally distributed for all five components, it would have resulted in significantly higher failure rate of 0.055 per unit time.

RELIABILITY APPORTIONMENT RULES

The discussion in the section above suggests that some preliminary rules may be used for obtaining initial reliability apportionments before a more thorough optimization. Some apportionment rules have been suggested in the literature [5].

Equal Reliability Apportionment. One can test a set of software components, such that at the end they all individually have the failure rate equal to the target failure rate for the system. Newly developed modules need to be tested until their reliability equals that of the existing modules.

Complexity Based Apportionment. The software size itself is a complexity metric. Thus, the available test time can be apportioned in proportion to the software size.

Impact Based Apportionment. A component that is executed more frequently, or is more critical in terms of failures, should be assigned more resources. Note that in the analysis above, x_i can be chosen to reflect the product of frequency and criticality.

RELIABILITY ALLOCATION FOR COMPLEX SYSTEMS

Some systems can have complex reliability allocation choices and may require an iterative approach [1,2,7,12]. Such an approach is also needed if the objective function is multi-objective and includes both total cost and the system reliability [1]. An iterative approach involves the following steps [12]:

1. Design the system using functional subsystems.
2. Perform an initial apportionment of cost or reliability attributes based on suitable apportionment rules or preliminary computation.
3. Predict system reliability.
4. Determine if reallocation is feasible and will enhance the objective function. If so, perform reallocation.
5. Repeat until optimality is achieved.
6. See if this meets the objectives. If not, consider returning to step 1 and revising the design at a higher level.
7. Finalize the design with recommended reliability allocation and the cost projections.

Several optimization methods can be used for steps 2–5 above. These can be classified into three approaches [1] as follows:

1. *Exact Methods.* When the problem is not large, exact methods can be desirable. In general, the problem can be a non-linear optimization problem. In a few cases, the problem can be transformed into a linear problem, as shown in the example above.
2. *Heuristics Based Methods.* Several heuristics for reliability allocation have been developed. Many of them

are based on identifying the variable to which the solution is most sensitive, and varying its value.

3. *Metaheuristic Algorithms*. These algorithms are based on artificial reasoning. The best known of them are *genetic algorithms*, *simulated annealing*, and *tabu-search*. These algorithms can be useful when the search space is large and approximate results are sought.

Reliability allocation in systems with replicated subsystems can encounter correlated failures and thus would need a more careful modeling. In software, such correlation can be significant [13]. The factors that affect technology dependant reliability attributes such as defect density in software [14] have been studied by researchers and are still being researched.

Several software tools, both general purpose [15] and special purpose [6], have been developed that can simplify setting up and solving the optimal allocation problem. Reliability allocation problem can also be formulated to address other reliability attributes such as availability or maintainability.

REFERENCES

1. Kuo W, Prasad VR. An annotated overview of system-reliability optimization. *IEEE Trans Reliab* 2000;49:176–187.
2. Majety SRV, Dawande M, Rajgopal J. Optimal reliability allocation with discrete cost-reliability data for components. *Oper Res* 1999;47:899–906.
3. Mettas A. Reliability allocation and optimization for complex systems. *Proceedings of the Annual Reliability and Maintainability Symposium*; 2000 Jan; Los Angeles (CA). pp. 216–221.
4. Butler RW, Finelli GB. The infeasibility of quantifying the reliability of life-critical real-time software. *IEEE Trans Softw Eng* 1993;19:3–12.
5. Lakey PB, Neufelder AM. *System and software reliability assurance notebook*. Rome: Rome Laboratory; 1996. pp.6.1–6.24.
6. Lyu MR, Rangarajan S, van Moorsel APA. Optimal allocation of test resources for software reliability growth modeling in software development. *IEEE Trans Reliab* 2002;51: 183–192.
7. Elegbede AOC, Chengbin C, Adjallah KH, *et al.* Reliability allocation through cost minimization. *IEEE Trans Reliab* 2003;52: 106–111.
8. Musa JD, Iannino A, Okumoto K. *Software reliability, measurement, prediction, application*. New York: McGraw-Hill; 1897.
9. Lyu MR, editor. *Handbook of software reliability engineering*. Highstown (NJ): McGraw-Hill; 1995.
10. Goel AL, Okumoto K. Time-dependent error detection rate model for software and other performance measures. *IEEE Trans Reliab* 1979;28:206–211.
11. Malaiya YK, Denton J. What do the software reliability growth model parameters represent? *International Symposium on Software Reliability Engineering*; Albuquerque (NM); 1997. pp. 124–135.
12. NASA Document, *Organizational Instruction: Reliability Allocation*, QDR-008 Rev. E, Oct 1; 2004.
13. Dai YS, Xie M, Poh KL. Modeling and analysis of correlated software failures of multiple types. *IEEE Trans Reliab* 2005;54: 100–106.
14. Neufelder A. The facts about predicting software defects and reliability. A Neufelder; 2002 Apr 8; *The RAC Journal*. 2009. pp. 1–4. Available at <http://www.softrel.com/publicat.htm>.
15. ReliaSoft. *Reliability Importance and Optimized Reliability Allocation (Analytical)*. 2007. Available at http://www.weibull.com/SystemRelWeb/component_reliability_importance.htm. Accessed 2010 Aug 27.

OPTIMAL REPLACEMENT AND INSPECTION POLICIES

MAARTEN-JAN KALLEN
HKV Consultants, Lelystad,
The Netherlands

This article covers the selection of policies for the timing of inspections or replacements, which are optimal according to some predefined criterion. The primary application of such policies is in problems of maintenance optimization. Mechanical devices, electrical components, and civil infrastructures are examples of items which generally need some form of maintenance. Inspections and replacements in various forms are actions that may be performed as part of a maintenance strategy. Repairs may be performed before an item has failed, in which case it is a preventive maintenance action, or it may be performed after an item has failed in which case it is a corrective maintenance action. Depending on the risk involved with the failure of an item, the decision maker will have a preference for one of these two types of maintenance actions.

The times at which failures occur are usually uncertain. For systems that depend on the availability of an item, for example, a conveyor belt in a production line, the failure of such an item results in unplanned and costly disruptions of production. It will therefore pay to perform inspections and repair items preventively. An optimal policy describes how often these inspections and repairs should be performed, such that the cost of these maintenance actions is well balanced against the cost of simply letting an item fail. The decision to preventively replace an item may be based on its condition, in which case the policy may also include the quality criteria by which an item will be deemed necessary to be replaced or not.

Optimal maintenance policies are obtained using mathematical models, which

are almost always probabilistic in nature. Deterministic models are only applicable when the occurrence of failures is deterministic or when (probabilities of) the outcomes of an inspection are assumed to be known beforehand to the decision maker. Many models have been proposed for the purpose of optimizing maintenance actions of which Refs 1–10 all give an overview in some form or another. In the following section, many of the common assumptions that characterize these models are mentioned. The next section lists three cost-based criteria that can be used to compare different maintenance policies. This is followed by a section with the mathematical details of the two most common models: age-based and condition-based replacements.

MODEL ASSUMPTIONS AND NOTATIONS

Maintenance models are built upon a number of assumptions that help to keep the calculations tractable. They also ensure that the resulting policies are easy to implement in practice.

Inspections are generally performed to determine the condition of an item. However, in some cases they may be required to ascertain if an item is still functional or not. This means that a failure of the item is not noticed until an inspection is performed. Inspections are almost always assumed to be instantaneous. This is generally a good approximation if the duration of an inspection is negligible compared to the lifetime of the item. The same assumption is often made for repairs, although there are many applications in which the duration of a repair is relevant. This is especially true for cases in which the availability of an item is an important operating characteristic of the maintenance policy. Inspections are either periodic or not, where in the latter case the times between inspections may differ. An example of an aperiodic inspection policy is one in which the frequency of inspections increases

together with the age of an item.

The single most important assumption in maintenance policies is that a repair brings an item to an as-good-as-new state such that a repair is equivalent to a replacement. Implicit in this assumption is that the operating environment of an item after it has been replaced is the same as it was before the replacement. This means that the deterioration of the item is the same after each replacement or, in mathematical terms, the probability distribution of the time to failure is the same after each replacement. This assumption allows for the use of the results in renewal theory (see **Limit Theorems for Renewal Processes**). Although models for partial repair (also referred to as *imperfect maintenance*) are relevant in real-life situations, they are outside the scope of this article.

The following section uses the concept of a cycle. Each time an item is replaced, a new cycle starts. The cycle ends when an item is replaced, which may be before or after it fails. During a cycle several types of costs may be incurred: cost of a preventive or corrective replacement, cost of one or more inspections, and possibly the cost of unavailability. The latter cost should be included if the occurrence of a failure may only be determined using an inspection. In the absence of such a penalty, the optimal policy would be not to inspect at all.

COST-BASED CRITERIA FOR POLICY EVALUATION

This section presents three cost-based criteria, whose purpose is to reduce future streams of costs (and rewards) to a single value for policy evaluation. Let $K(t)$ denote the costs incurred during a time span $[0, t]$ with $t > 0$. A renewal process $\{N(t), t \geq 0\}$ is a nonnegative integer-valued stochastic process, which registers the successive renewals in the time interval $[0, t]$. Let $F(t)$ be the cumulative probability distribution function of renewal time T and let $c(t)$ be the cost associated with a renewal at time t .

An optimal maintenance policy over a finite time-span (or bounded horizon) may

be determined by calculating the expected value, denoted by $EK(t)$, for each policy and to select the policy with the lowest value. If the choice of a maintenance policy, or a particular design of the system, affects the length of the service life of an object, it is beneficial to compare these policies or designs based on (possibly fictive) annuity payments to be made during the service life. Annuities are specified payments made at fixed time intervals. Specifying the costs of maintenance policies in terms of these (annual) payments enables one to compare policies, which entail different service lives.

The first and most common criterion for policy evaluation over an unbounded horizon is the criterion of expected average cost per unit time:

$$\lim_{t \rightarrow \infty} \frac{EK(t)}{t} = \frac{\int_{t=0}^{\infty} c(t) dF(t)}{\int_{t=0}^{\infty} t dF(t)}. \quad (1)$$

Therefore, if a cycle is completed every time a renewal occurs, then Equation (1) states that the (expected) long-run average cost is simply the expected cost incurred during a cycle divided by the expected length of a cycle. This criterion is a direct result of renewal reward theory [11, Theorem 3.16].

According to Ref. 12, Chapter 11, there are two other cost-based criteria, which may be used for evaluating maintenance policies: (i) the expected discounted costs over an unbounded horizon and (ii) the expected equivalent average costs per unit time. Using the criterion in Equation (1), one does not have a preference for the times at which costs are incurred. In other words, costs incurred in the near future are weighted no differently than costs incurred a long time from now.

Future costs (and rewards) can be discounted to their present value on the basis of a discount rate. The notion of discounting assumes that the future utility of money decreases in time. In practice, this means that costs incurred at future dates are less bothersome to us than costs incurred today. In mathematical terms, the (present) discounted value of the costs c_n in unit time $n = 0, 1, 2, \dots$ is defined to be $\alpha^n c_n$ with

$\alpha = [1 + (r/100)]^{-1}$, the discount factor per unit time and $r\%$ ($r > 0$) the discount rate per unit time. This formulation assumes that the timescale is divided in discrete units of time. For continuous types of discounting, we may assume any continuous, nonnegative, and nonincreasing function $\alpha(t)$ with the convention $\alpha(0) = 1$. The simplest example of such a function is the example of exponential discounting with $\alpha(t) = \exp\{-rt\}$ proposed by Samuelson [13]. The expected discounted costs over an unbounded horizon are given by

$$\lim_{t \rightarrow \infty} EK(t, \alpha(t)) = \frac{\int_{t=0}^{\infty} \alpha(t)c(t) dF(t)}{1 - \int_{t=0}^{\infty} \alpha(t) dF(t)}. \quad (2)$$

See Ref. 14, for a proof of this result with $\alpha(t)$ equal to the exponential discounting function mentioned above. For more on continuous forms of discounting, like exponential and hyperbolic discounting, see Ref. 15. The latter article is also concerned with the derivation of the second moment of the discounted costs. This information is interesting because it gives an indication of the amount of uncertainty involved with the optimal policy. Experience shows that an optimal policy in terms of expected (discounted) costs may also have large uncertainty associated with it [16]. This suggests that a decision maker may prefer to select a policy with slightly higher expected costs, but with an outcome that is less uncertain.

The expected equivalent average costs per unit time are given by

$$\lim_{t \rightarrow \infty} \frac{EK(t, \alpha(t))}{\int_{x=0}^t \alpha(x) dx}. \quad (3)$$

This criterion is obtained by constructing an infinite stream of costs (or rewards) with the same present discounted value as the expected discounted costs over an unbounded horizon. It relates the notions of the expected average costs and the expected discounted costs. This criterion is particularly well suited for the inclusion of initial investment costs into the decision model. The concept of life cycle costing is an example of an analysis in

which the costs of construction and demolition are also included when determining an optimal maintenance policy. Design decisions may influence the future costs of maintaining the item, such that the costs of different designs may be compared over the full life-time of an item.

COMMON MAINTENANCE POLICIES

By far the most common policies applied today are the age-based and condition-based replacement policies. The basic theory of both policies are described next (see also **Age Replacement Policies** and the section titled “Condition-Based Maintenance and Reliability” in this encyclopedia).

Age-Based Replacements

In an age-based maintenance policy, an item is replaced at age $\tau > 0$ or at a failure, whichever occurs first. It is assumed that failures are immediately noticed. Let c_1 denote the cost of a corrective replacement after a failure and c_2 the cost of a preventive replacement before a failure. We always assume that a preventive replacement is less costly compared to a corrective replacement, that is $c_1 > c_2$. The total costs incurred during the interval $[0, t]$ are

$$EK(t) = c_1 EN_1(t) + c_2 EN_2(t), \quad (4)$$

where $EN_1(t)$ and $EN_2(t)$ are the expected number of corrective and preventive replacements respectively. Using Equation (1) with

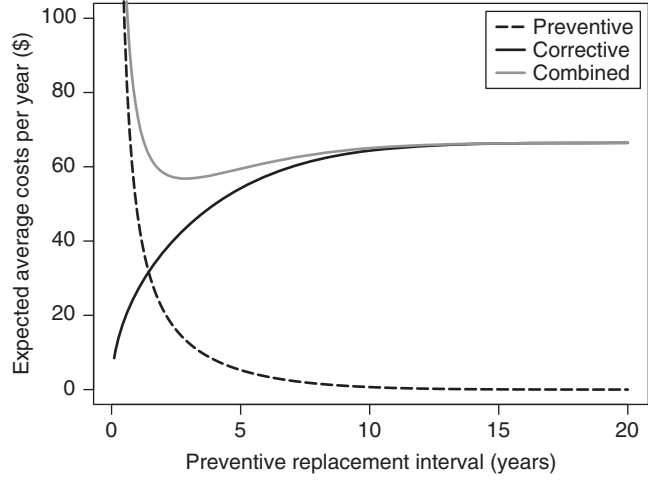
$$c(\tau) = \begin{cases} c_1, & T \leq \tau, \\ c_2, & T > \tau, \end{cases} \quad (5)$$

the expected average costs per unit time over an unbounded horizon are given by

$$\lim_{t \rightarrow \infty} \frac{EK(t)}{t} = \frac{c_1 F(\tau) + c_2 [1 - F(\tau)]}{\int_{t=0}^{\tau} [1 - F(t)] dt} = L(\tau). \quad (6)$$

Here $L(\tau)$ is the loss function for the replacement interval τ and $F(t) = \Pr\{T \leq t\}$ is the probability distribution function of the

Figure 1. Example of the loss function $L(\tau)$ for an age-based replacement policy giving the expected average costs as a function of the replacement interval τ .



renewal time T . An optimal policy τ^* is found by minimizing $L(\tau)$ with respect to τ . For a very large replacement interval Equation (6) states that only corrective replacements are performed: $L(\infty) = c_1/\mu$, with μ , the mean time to failure:

$$\mu = E[T] = \int_{x=0}^{\infty} x dF(x). \quad (7)$$

Let $r(x) = f(x)/(1 - F(x))$ denote the failure rate function of the distribution $F(x)$ (see **Hazard Rate Function**). If $r(x)$ is strictly increasing, then the minimum τ^* of the loss function $L(\tau)$ exists, but may be equal to infinity. Using heuristic arguments, Barlow and Proschan [17] show that it should always hold that $\tau^* \geq (c_2/c_1)\mu$. This is useful information, but in most cases the optimal replacement interval will be much larger than this lower bound. Note that the optimal replacement interval τ^* only depends on the ratio c_2/c_1 of the costs and not on the actual costs themselves, which can be seen by dividing the loss function in Equation (6) by c_1 . This changes the value at the minimum, but not its location and this also holds for an age-based replacement policy using discounted costs [18, p. 100].

Example. The following simple example helps to illustrate a typical result of an age-based replacement model. Let the lifetime of

an item be represented by a Weibull distribution with cumulative probability function

$$F(t) = 1 - \exp\left\{-\left[\frac{t}{b}\right]^a\right\},$$

with scale parameter $b = 5$ and shape parameter $a = 1.5$. The expected value of the lifetime T is then approximately 4.5 time units. Let $c_1 = \$300$ and $c_2 = \$50$, then the loss function $L(\tau)$ in Equation (6) may be plotted for a range of values of τ . This is shown in Fig. 1. Also shown in this figure are the individual contributions of preventive and corrective replacements to $L(\tau)$. It can be seen that the expected costs due to a preventive replacement decreases as the length of the replacement interval increases. This is due to the fact that the probability of a corrective replacement increases as replacements are performed less frequently. The minimum is attained at $\tau^* = 2.9$ time units with a value of $L(\tau^*) = \$56.81$. For large values of τ , the loss function approaches $c_1/\mu \approx \$300/4.5 \approx \66.67 .

Condition-Based Replacements

Consider now an item subjected to deterioration whose condition may be determined by inspection. Let the condition of this item be modeled by a stochastic process $\{X(t), t \geq 0\}$. This process may be continuous or discrete,

the latter being the case if the condition of the item may be classified in one of the finite number of states (see the section titled “Stochastic Processes” in this encyclopedia). Assume that an item is considered to be in a failed state when $X(t) > r_1$ and that a preventive replacement is performed when $r_1 \geq X(t) > r_2$, $r_1 > r_2 > 0$, which is often referred to as the item being in a marginal state. If $X(t) \leq r_2$, the item is in a functional state and no action is taken.

If the item is in a marginal state, there is an opportunity to replace the item preventively and at lower cost compared to a corrective replacement. Inspections are required to determine when the item is in a marginal state and possibly also to determine if an item has failed or not. This requirement differentiates this policy from an age-based replacement policy. Let $\tau > 0$ be the fixed time between inspections, such that $k\tau$, $k = 1, 2, \dots$, represents the time of the k th inspection. Also, each inspection entails a cost of $c_3 > 0$ monetary units. For inspections to be economically interesting, c_3 must be smaller than both c_1 and c_2 . Also, if failures are noticed immediately without the necessity of performing an inspection, the cost of a single inspection, followed by a preventive replacement, must be less than the cost of a corrective replacement. In this case, inspections are only useful if $c_3 < c_1 - c_2$. With these costs, the aim is to determine the optimal inspection interval τ^* , which minimizes the loss function $L(\tau)$. Such an optimal policy will present a balance between inspections, followed by a preventive replacement on the one side and a corrective replacement on the other side.

As in the age-based replacement model presented previously, each cycle in the condition-based replacement model ends with either a preventive or a corrective replacement. However, each cycle is now further partitioned at the times of one or more inspections. This becomes apparent in the cost function $c(k)$, which gives the costs incurred up to time $k\tau$:

$$c(k) = c_1 + (k - 1)c_3 \quad (8)$$

if $X((k - 1)\tau) \leq r_2$ and $X(k\tau) > r_1$ or

$$c(k) = c_2 + kc_3 \quad (9)$$

if $X((k - 1)\tau) \leq r_2$ and $r_1 \geq X(k\tau) > r_2$. Equation (8) is the sum of the cost of a corrective replacement and $k - 1$ inspections, which assumes that no inspection is required in the k th interval to determine that the item has failed. Equation (9) is the sum of a preventive replacement and k inspections. The influence of the inspections on this cost function is now obvious: a replacement (preventive or corrective) is only performed if no replacements were made during one of the $k - 1$ previous intervals. The following short notation is used for the probability of incurring the costs in Equations (8) and (9):

$$P_1(k, \tau) = \Pr\{X((k - 1)\tau) \leq r_2, X(k\tau) > r_1\}$$

and

$$P_2(k, \tau) = \Pr\{X((k - 1)\tau) \leq r_2, r_1 \geq X(k\tau) > r_2\}.$$

Using the expected average cost criterion in Equation (1) and the cost functions in Equations (8) and (9), the loss function is

$$\begin{aligned} L(\tau) = & \left\{ \sum_{k=1}^{\infty} [c_1 + (k - 1)c_3]P_1(k, \tau) \right. \\ & \left. + [c_2 + kc_3]P_2(k, \tau) \right\} \\ & \cdots \left\{ \sum_{k=1}^{\infty} E[T]P_1(k, \tau) + k\tau P_2(k, \tau) \right\}^{-1}, \end{aligned} \quad (10)$$

where the failure distribution is given by $F(t) = \Pr\{T \leq t\} = \Pr\{X(t) > r_1\}$. By dividing the loss function in Equation (10) by the cost of an inspection c_3 and rearranging the terms in the numerator, we obtain

$$\begin{aligned} & \sum_{k=1}^{\infty} \frac{c_1 - c_3}{c_3} P_1(k, \tau) + \frac{c_2}{c_3} P_2(k, \tau) \\ & + k[P_1(k, \tau) + P_2(k, \tau)] \end{aligned} \quad (11)$$

in the numerator. Like the age-based replacement model, this changes the value

at the minimum τ^* , but it does not change the location of the minimum. The result in Equation (11) shows that the location of the minimum depends on three things: the ratios $(c_1 - c_3)/c_3$ and c_2/c_3 and the expected number of inspections per cycle. The expected number of inspections per cycle depends on the threshold level for preventive maintenance,

$$\begin{aligned} &P_1(k, \tau) + P_2(k, \tau) \\ &= \Pr\{X((k-1)\tau) \leq r_2, X(k\tau) > r_2\}, \end{aligned}$$

therefore the preventive replacement threshold r_2 is also a parameter, which may be adjusted in order to obtain an optimal replacement policy.

As mentioned earlier, if a failure is only observed by inspection, the cost function $c(k)$ must include a penalty for each time unit that the item is left in a failed state. Otherwise the optimal policy would be to not inspect at all, that is, $\tau^* = \infty$, since this would make the expected cycle length infinitely large, whereas the expected costs per cycle remain finite. If c_4 is the cost of a nonfunctional item per time unit, then the cost function in the case of a preventive replacement in Equation (8) is given by

$$c(k) = c_1 + kc_3 + \int_{t=(k-1)\tau}^{k\tau} (k\tau - t)c_4 dF(t).$$

The denominator in Equation (10) simplifies to

$$\sum_{k=1}^{\infty} k\tau [P_1(k, \tau) + P_2(k, \tau)],$$

since the item is always replaced at the time of an inspection.

There are many variations on the basic model for condition-based maintenance. Obviously there is the formulation of the loss function in Equation (10) using discounted costs, but there are also many different types of stochastic processes that have been used to model the uncertain deterioration over time, represented here by $X(t)$. A minimum requirement for the model presented here is that $X(t)$ is monotonically increasing, such that $X(t_2) \geq X(t_1)$ for all $t_2 \geq t_1$.

For continuous types of deterioration, the gamma process is a good candidate. The gamma process is a continuous-time stochastic process with independent gamma-distributed increments. It also belongs to the family of Lévy processes. See Ref. 19 for a broad review of the application of gamma processes to maintenance optimization. Such a gamma process is applied by Jia and Christer [20] in a condition-based replacement model, which they extended to include the time of the first inspection, τ_1 as a separate decision variable. Such a model can be useful in situations in which the rate of deterioration proceeds slowly after the item is taken into service and then gradually increases. The optimal policy may then be the one in which the first inspection is followed by more frequent periodic inspections. A condition-based replacement policy may also be determined for discrete types of deterioration using a Markov chain for $X(t)$ [21] or in situations where inspections are not able to determine the exact state of deterioration $X(t)$ [22].

REFERENCES

1. Cho DI, Parlar M. A survey of maintenance models for multi-unit systems. *Eur J Oper Res* 1991;51(1):1–23.
2. Dekker R. Applications of maintenance optimization models: a review and analysis. *Reliab Eng Syst Saf* 1996;51(3):229–240.
3. Frangopol DM, Kallen MJ, van Noortwijk JM. Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. *Prog Struct Eng Mater* 2004;6(3):197–212. DOI: 10.1002/pse.180.
4. McCall JJ. Maintenance policies for stochastically failing equipment: a survey. *Manag Sci* 1965;11:493–524.
5. Nicolai RP, Dekker R. Optimal maintenance of multi-component systems: a review. In: Kobbacy KAH, Murthy DNP, editors. *Complex system maintenance handbook*. Berlin: Springer; 2007.
6. Pierskalla WP, Voelker JA. A survey of maintenance models: the control and surveillance of deteriorating systems. *Nav Res Log Q* 1976;23:353–388.

7. Sherif YS, Smith ML. Optimal maintenance models for systems subject to failure—a review. *Nav Res Log Q* 1981;28:47–74.
8. Thomas LC. A survey of maintenance and replacement models for maintainability and reliability of multi-item systems. *Reliab Eng* 1986;16(4):297–309.
9. Valdez-Flores C, Feldman RM. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Nav Res Log* 1989;36:419–446.
10. Wang H. A survey of maintenance policies of deteriorating systems. *Eur J Oper Res* 2002;139(3):469–489. DOI: 10.1016/S0377-2217(01)00197-7.
11. Ross SM. *Applied probability models with optimization applications*. San Francisco (CA): Holden-Day; 1970.
12. Wagner HM. *Principles of operations research*. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall; 1975.
13. Samuelson PA. A note on measurement of utility. *Rev Econ Stud* 1937;4(2):155–161.
14. Rackwitz R. Optimizing systematically renewed structures. *Reliab Eng Syst Saf* 2001;73(3):269–279. DOI: 10.1016/S0951-8320(01)00050-3.
15. van der Weide JAM, Suyono, van Noortwijk JM. Renewal theory with exponential and hyperbolic discounting. *Probab Eng Inform Sci* 2008; 22(1): 53–74. DOI: 10.1017/S0269964808000041.
16. van Noortwijk JM. Explicit formulas for the variance of discounted life-cycle cost. *Eng Reliab Syst Saf* 2003;80:185–195. DOI: 10.1016/S0951-8320(03)00023-1.
17. Barlow RE, Proschan F. *Mathematical theory of reliability*. New York (NY): John Wiley & Sons, Inc.; 1965.
18. Barlow RE. *Engineering reliability*. Philadelphia (PA): American Statistical Association and the Society for Industrial and Applied Mathematics (ASA-SIAM); 1998.
19. van Noortwijk JM. A survey of the application of gamma processes in maintenance. *Reliab Eng Syst Saf* 2009;94(1):2–21. DOI: 10.1016/j.res.2008.11.013.
20. Jia X, Christer AH. A prototype cost model of functional check decisions in reliability-centered maintenance. *J Oper Res Soc* 2002;53(12):1380–1384. DOI: 10.1057/palgrave.jors.2601453.
21. Kallen MJ, van Noortwijk JM. Optimal periodic inspection of a deterioration process with sequential condition states. *Int J Press Vessels Piping* 2006;83(4):249–255. DOI: 10.1016/j.ijpvp.2006.02.007.
22. Kallen MJ, van Noortwijk JM. Optimal maintenance decisions under imperfect inspection. *Reliab Eng Syst Saf* 2005;90:177–185. DOI: 10.1016/j.res.2004.10.004.

OPTIMAL RISK MITIGATION AND RISK-TAKING

CHRIS CHAPMAN

IAN HARWOOD

School of Management,
University of Southampton,
Southampton, UK

Risk mitigation and risk-taking is commonly approached in a different manner in different areas of application—operations management, project planning and strategy formation for example. To some extent these differences reflect different dominant background frameworks—decision analysis, portfolio selection, or scenario building for example. In any decision-making context, optimal risk mitigation and risk-taking demands an optimal set of working assumptions within an optimal set of framing assumptions. Working assumptions are context and application specific assumptions of convenience—robust simplifications to save effort and time which are appropriate to the circumstances. Framing assumptions include what is meant by overall optimality, what “risk” means, and an understanding of the purposes of risk management [1]. These assumptions are impacted by phenomena such as “problem framing,” “outcome history,” and “judgmental bias” [2].

The distinction between a local optimum and a global optimum is widely understood in contexts like mathematical programming—for a local optimization process to be consistent with a broader global optimum, conditions associated with global optimization have to be met. The overall optimization of interest here is a generalization of this notion, a form of universal optimization that addresses the nature, costs, and benefits of alternative decision-making processes as well as the nature, costs, and benefits of the decisions made.

To facilitate overall optimization in practical terms, optimal framing assumptions must be common for all application areas. Further, the basic tool kit must be portable, flexible,

and integrated—a modern approach to practical ORMS requires a multimethod approach with compatible components [3].

The framing assumptions and tools used in this article were initially developed for project risk management, where they have been used successfully in practice since the mid-1970s [4]. They have also been used successfully in many of the other areas covered in the section titled “Risk Analysis Applications” in this encyclopedia and could be used in the others [5]. They embrace the issues discussed in the rest of the section titled “Risk Analysis” plus some parts of the sections titled “Markov Decision Processes (MDPs)” and “Decision Analysis” in this encyclopedia.

The following section illustrates simple working assumptions within a general framework plus a simple version of a basic tool recommended for making choices. A more complex version of this basic tool is then illustrated together with its companion tool. Risk-taking at corporate levels and linked behavioral issues are considered next. Working assumptions when starting an analysis are then considered and linked to bottom-up strategy formation. This section also links alternative frameworks which may be compatible. Consideration of top-down strategy formation completes an introductory overview of the key issues.

A SIMPLE OPTION CHOICE EXAMPLE

A simple example illustrates a recommended basic tool for making choices in a form initially developed for IBM UK for bidding purposes [5, Chapter 3]. It also illustrates other simple working assumptions within a general framework, and how more complex working assumptions can be addressed as necessary.

The approach suggested uses a generalization of the “uncertainty” concept employed by modern decision analysis, a generalization of the “risk” concept employed by modern portfolio theory, and synthesizing these two usually disparate perspectives. It assumes “measured” uncertainty and risk

can be portrayed by cumulative probability distributions, but there is no direct metric or proxy for either, and there is no need to specify utility curves or functions. It assumes that more than one objective is usually relevant and some objectives may involve an attribute that is not measurable in a useful sense.

The single photocopier in a busy office failed terminally. The office manager had to replace it quickly, and justify his choice of replacement later. He could obtain the same machine from the same supplier on a comparable contract. The contract cost for a minimum number of years was a rental charge per month plus a maintenance charge per copy. The only source of uncertainty associated with contract cost over the minimum number of years was the number of copies needed. The option A line in Fig. 1 portrays the office manager's estimate of the contract cost per year of this choice, assuming a linear cumulative probability distribution, corresponding to a uniform probability density function. Figure 1 uses the absolute minimum and maximum points " a_1 " and " a_2 " for portrayal simplicity, and these values can be used for estimation purposes. However, using 10 and 90 percentile value estimates to optimize the elimination of bias is recommended. In either case it is important to recognize that the true curve will be nonlinear and asymmetric.

Alternative suppliers offering comparable machines on the same contractual basis were evaluated using the same 10 and 90 percentile estimates, portrayed as lines B and C in Fig. 1.

Assuming minimizing contract cost was the primary objective, and making this assumption clear, the office manager argued that A was the risk efficient choice (its expected cost was the lowest and it involved less risk), because the option A line was entirely to the left of the B and C lines. As the option A machine was marginally faster than B or C, option A dominated when considering this secondary objective too. Had this not been the case, employee time lost waiting for copies might have been estimated and converted to an opportunity cost, followed by aggregation to a single cost attribute. Alternatively, any important secondary objectives might have been considered using separate graphs in Fig. 1 form, or a simple judgment made about whether the gap between (Fig. 1) primary objective lines was large enough to more than counterbalance secondary considerations.

As the office manager knew that the color and design style of the current supplier's products was consistent with the corporate consensuses of an optimal house style, this third-order objective was not an issue either. Had this not been the case, the gap between the curves would have to be evaluated by the relevant person or group in terms of trade-offs between nonmeasurable objectives as well as measurable objectives. An optimal approach in the general optimum sense uses simple dominance tests for additional objectives unless more sophisticated approaches look useful [5].

Had the option A machine been unavailable in a feasible time frame, the B option

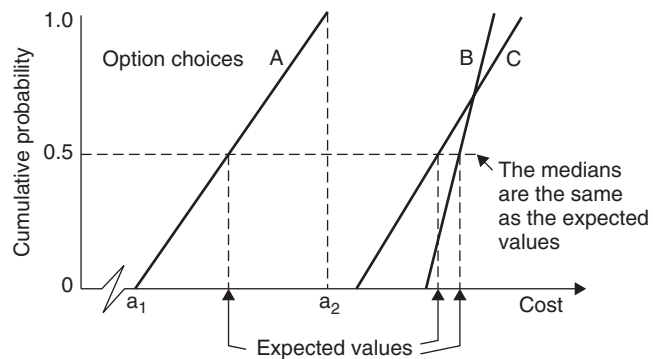


Figure 1. A simple example of an option choice graph.

would not dominate C in the risk efficient sense, indicted by the way the lines cross. However, the lower expected cost of B justifies the marginal increase in rental cost risk—at a corporate level this kind of additional variation should be regarded as just “noise.” This alternative starting position would also have to be tested in terms of additional objectives, measurable or nonmeasurable.

Had B or C seemed preferred choices because A was not available in a feasible time frame, reliability risk and risk associated with a new supplier who might not deal effectively with reliability or other issues would need care. Risk to the company and risk to the office manager might need separate consideration by the office manager—the office manager’s reputation might suffer from a highly visible mistake, perhaps costing the office manager their job if times were tough. Optimal mitigation of these risks might involve exploring more general working assumptions, like borrowing a copier from the existing supplier until A is available, a simple example of a reformulation of the choices available to deal with emerging concerns in a general framework.

More complex problem formulation and reformulation issues will be considered later, but this example illustrates a flexible general framework that can address multiple attributes (some not measurable), risk that is not limited to downside variability of measured attributes, and risk perceived differently by different parties.

This approach assumes one attribute per objective, but a multiple criteria approach to each attribute, involving expected outcome and risk. Risk is not measured by a single criterion like variance—it is depicted graphically by cumulative probability distributions. Comparing the B and C options graphically and choosing one implies an underlying decision function approximation to preference functions which addresses trade-offs between expected outcome and associated risk exists, but it does not require its specification or agreement between different relevant parties. Similarly, choices involving more than

one attribute imply the existence of appropriate decision function approximations to preference functions, which addresses trade-offs between attributes, but it does not require their specification or agreement between relevant parties.

“Risk efficiency” (a minimum level of risk for any given level of expected outcome for any attribute of interest including nonmeasurable attributes) in terms of all relevant attributes as a basis for optimal trade-offs between all relevant attributes is at the core of the underlying framework adopted here, the key framing assumption. A significant generalization of Markowitz’s notion [6] of risk efficiency is involved. One aspect of this generalization is recognizing that the measures of skew and other moments may be relevant, not just mean and variance, addressed by working directly with cumulative probability distributions. The other is considering all relevant attributes, using Fig. 1 equivalents for all measurable attributes. Nonmeasurable attributes like optimal house style and the office manager’s reputation may require considerations too, as may attributes not worth measuring, like the reliability of other machines and suppliers. Distinguishing between risk to the organization and risk to the office manager is important. Links like the effect on the office manager’s reputation of a commitment to a poor machine and supplier are important.

Working assumptions not addressed for robustness associated with this analysis include no need to look beyond the minimum rental contract period, or decompose average variability per annum within this period, or estimate Fig. 1 using more sophisticated probability distributions. Judgments about value, value trade-offs, and associated uncertainty, including process as well as option issues are core concerns explicitly addressed in this example, via a flexible approach to working assumptions suitable for the context. No concerns of potential importance are overlooked, even if they cannot be measured (like optimal house style or office manager reputation), or measurement is not practical (like the reliability of a new supplier).

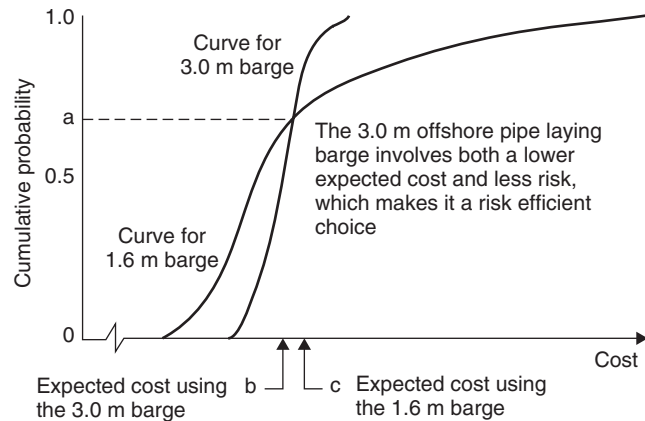


Figure 2. A North Sea project example of an option choice graph.

MORE COMPLEX OPTION CHOICE CURVES USED FOR RISK MITIGATION OPTIMIZATION

A more complex example illustrates a more sophisticated version of the same tool for making choices as it was used by BP International in the 1970s for the Magnus project, a £1000 million North Sea oil project. It also illustrates more sophisticated assumptions about the role of risk management.

Figure 2 was shown to the BP main board to illustrate why one risk mitigation measure developed via a new project risk management process [7] would more than pay for the application of that process. This was the basis of a case for using this new process to save money, perhaps £100 for every £1 invested in the process. This saving would be in addition to meeting the original objective of the process from a board perspective, providing the basis for time and cost estimates that were unbiased; on average BP could bring in projects on time and on cost using the new process, later demonstrated statistically.

The base plan approach to “hook-up”—linking the offshore pipeline to the production platform once both were in place—assumed the use of a barge with a 1.6-m nominal maximum wave height capability. The 1.6-m barge was shown to be sensible if this activity took place in August as planned, and August was shown to be a sensible target plan. However, the cumulative effect of delays to all earlier activities were shown to lead to likely delays to the start date, with about a 10%

chance of a failure to complete before winter set in, with £100 million consequences, portrayed by the long right-hand tail of the 1.6-m barge curve in Fig. 2. The 3-m barge was a mitigation measure to address this risk. It virtually eliminated the chance of failing to complete before winter, and its expected cost was shown to be about £5 million less, despite a much higher cost per day. Some reduction in expected cost was associated with fewer days being required, but eliminating the 10% chance of an extra £100 million was crucial. This simultaneous reduction in expected cost and risk—an improvement in risk efficiency—is portrayed in Fig. 2 by the more vertical 3-m curve with a lower expected cost. The 1.6-m barge had a 70% chance of being cheaper, indicated by point “a,” but “b” is to the left of “c.” On the basis of a discussion of Fig. 2, the board mandated the underlying process worldwide for all large or sensitive projects, because the board was convinced that the increase in project risk efficiency would more than pay for the process, the basic rationale for risk management based on the recommended framing assumptions.

LAYERED CONTRIBUTION GRAPHS TO PORTRAY HOW UNCERTAINTY ACCUMULATES

Before considering the cost implications of the completion of a late activity like “hook-up,” sources of uncertainty and associated

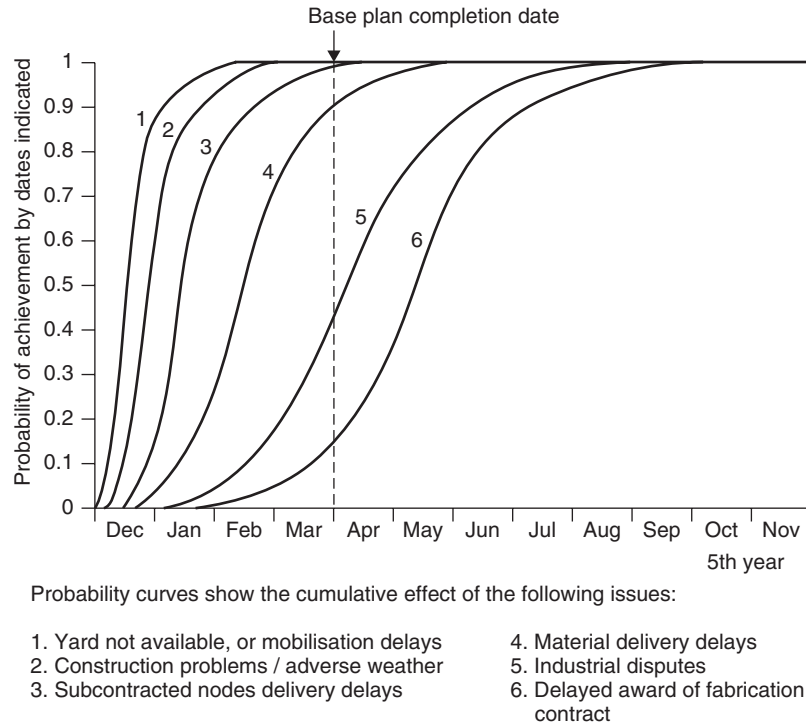


Figure 3. A North Sea project example of a layered contribution graph. Probability curves show the cumulative effect of the following issues: 1, yard not available, or mobilization delays; 2, construction problems/adverse weather; 3, subcontracted nodes' delivery delays; 4, material delivery delays; 5, industrial disputes; and 6, delayed award of fabrication contract.

mitigation had to be addressed for all earlier activities and this had to be linked in a nested structure, which reflected knock-on relationships and dependences. The choice curves in Fig. 2 measure composite sources of uncertainty, which needed decomposition in an optimal manner. Figure 3 illustrates an activity level portrayal of completion date sources of uncertainty and associated mitigation measures specific to that source within an activity, in this case the fabrication of the “jacket” (the steel structure forming the basis of the platform used to produce oil).

Curve 1 is the cumulative probability distribution associated with source of uncertainty 1 given any specific mitigation measures for that source. Curves 2–6 successively add comparable cumulative probability distributions for sources 2–6. Curve 6 provides the overall total, and curves 1–6 show the contribution of that component

via the successive gaps—5 is the biggest, 2 the smallest.

To compute these curves using a standard Monte Carlo simulation package assuming independence between sources of uncertainty, $X_1 - X_6$ could define the outcomes of samples from each of the six source of uncertainty distributions, with $X_7 = X_1 + X_2$, $X_8 = X_7 + X_3$, and so on to $X_{11} = X_{10} + X_9$. Then curve 1 is defined by X_1 , curve 2 is defined by X_7 , and so on with curve 6 defined by X_{11} . That is, n sources require $n - 1$ composites instead of just one overall composite to provide all n curves. Dependence specifications have to be defined in terms of preceding composite components for sources 3–6—source 2 dependence can be defined in relation to source 1. Independence is a very special case of the implied separability conditions [5, Chapter 10].

In this case, the ordering of sources 5 and 6 reflects mitigation discussions. Source 6

could be eliminated by senior management action fairly simply—the first mitigation measure discussed was early award of the fabrication contract. This still leaves an uncomfortable situation. Source 5 could be almost eliminated, by more complex management action, which might impact source 6. Source 5, industrial disputes, was initially sized subjectively, but its size led to data analysis, which suggested that industrial disputes were the result of a “feast or famine” environment for platform constructors. The only viable mitigation identified was collaboration with other oil companies to smooth the flow of work. This should lower contracted prices as well as the chance of industrial disputes—an example of an opportunity discovered while looking for risk efficient responses to overall uncertainty. If the uncertainty portrayed in Fig. 3 for both 5 and 6 were effectively eliminated by mitigation in this way, curve 4 becomes an acceptable picture within this activity.

Curve 1 is a composite of two sources of uncertainty, both assessed in relation to optimal mitigation measures prior to the estimation of curve 1. For example, “yard not available” might have arisen because another company’s jacket is not finished by the contracted time to start BP’s jacket, and the yard (a dry dock) cannot float it out until it is finished. The first choice response was “find an adjacent site and start stockpiling steel to cope with a short delay.” This involved a secondary source of uncertainty—“a long delay may be involved.” In this case, optimal mitigation shifts to “find another yard.” The secondary source of uncertainty here is “there may be none available.” “Accept a long delay” then becomes the only available choice. Figure 3 shows six sources of uncertainty at this activity level because more curves make these layered contribution graphs too complex to read easily. However, each curve can involve a lower level structure, and each of these can be further decomposed. The first source on an activity level graph is often “delays caused by earlier activities,” a composite built up by working through the activity network. The second source is often a composite like “production problems” in Fig. 3, a composite of sources

of uncertainty not deemed worthy of separate pre-emptive analysis. Judging how far to decompose uncertainty for a cost-effective understanding of optimal mitigation, as well as unbiased estimates, is the target of an optimal process which is discussed later.

Each activity needs to be addressed in this manner, managing mitigation specific to each source of uncertainty. Further, “general” mitigation measures, which can address accumulated uncertainty from many sources, including knock-on effects, require consideration at a number of levels. The 3-m barge option is one example of a general response. Another example is a second or third shift for the pipe coating operation—it would mitigate delays specific to pipe coating and earlier difficulties.

Decomposition of activity uncertainty can include the use of Markovian processes, to optimize mitigation measures on a time period by time period basis.

Building up an understanding of duration driven uncertainty and the impact of mitigation measures activity by activity in this way leads to the ability to portray related costs, including the implications of a mitigation measure like that portrayed by Fig. 2. Underlying both Fig. 2 curves is a nested set of Fig. 3 format curves. Further, any choice of mitigation options during the build-up can use the Fig. 2 format.

An optimal approach to risk mitigation for complex projects requiring detailed proactive risk mitigation demands this approach for projects being shaped at an implementation and delivery strategy shaping stage, as demonstrated in the 1970s, and many times subsequently. By the end of the 1970s, it had become clear that risk efficiency and unbiased estimates are not the only benefits, but processes working within the same framing assumptions can also meet a number of additional objectives, which are important [4, Chapter 2 and 5]. The very simple structure demonstrated by Fig. 1, perhaps supported by linear versions of Fig. 3, uses simpler working assumptions while drawing on the same general framing assumptions. All projects and all operations management choices need an optimal level of simplicity/complexity in

terms of these working assumptions, generally between that illustrated by Figs 1 and 2, drawing on the same framing assumptions. All projects and all operations management also have to adapt their risk management objectives to the context and application in an optimal manner.

A framing assumption illustrated by the role of Fig. 3 is the need to decompose measured overall uncertainty and associated risk as and when necessary into intermediate level composites as well as base level sources of uncertainty in an optimized structured manner, using the structure to deal with dependence and knock-on effects. Figure 3 portrays the way the measured uncertainty accumulates, providing built-in sensitivity analysis, an optimal approach to understanding how what has been decomposed fits together.

Another framing assumption is the use of Fig. 2 to assess alternative risk mitigation measures as well as basic choices like those addressed in Fig. 1.

Working assumptions associated with Figs 2 and 3 include the need to look at the possible delay to completing “hook-up,” and its cost implications in terms of component composites at a reasonably detailed level, considering the impact of mitigation measures in the process, and dependence. Some mitigation measures operate at the base level, some at the overall uncertainty level, and others at the intermediate levels. Some are reactive, some are proactive. Some are just fixes for threats, but others are

opportunities. Dependence may take a range of simple statistical and complex causal forms.

In the context of this example, the multiple attribute and non-measurable attribute and multiple party complications considered in the last section were not relevant—in this sense the consideration of risk efficient choices was relatively simple for BP.

In the context of Fig. 1, no Fig. 3 equivalent was optimal, but if both contract cost as measured and opportunity cost associated with employee time were relevant, a Fig. 3 curve equivalent could be very important. In some contexts, simple linear versions of Fig. 3 or spreadsheet equivalents may be optimal.

An overall optimum implies the need to always search for risk efficiency at an appropriate level of detail, understanding how the detail fits together with a flexible set of working assumptions suitable for the context.

RISK-TAKING AT CORPORATE LEVELS AND LINKED BEHAVIORAL IMPLICATIONS

During the 1990s, IBM UK undertook a culture change program involving a two-day “forum” run about 40 times to cover all senior staff. It used a discussion of Fig. 2 followed by a discussion of Fig. 4 as its starting position.

Having explained the implications of Fig. 2, it was pointed out that slightly different numbers might have shifted the 3-m curve to the right as shown in Fig. 4. Point “a” now suggests that the 1.6-m barge

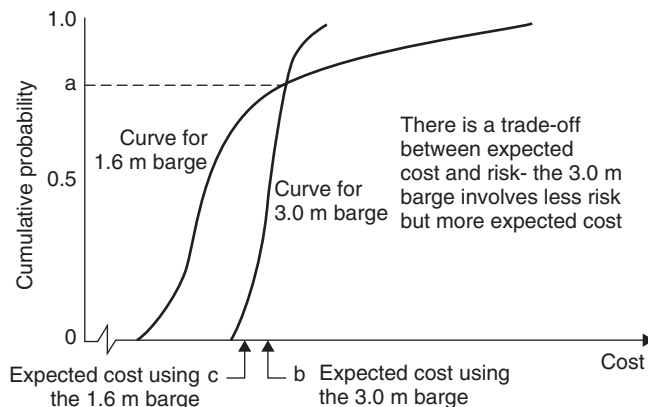


Figure 4. Shifting the 3-m curve on Fig. 2 to the right.

has an 80% chance of being cheaper. Of crucial importance, the reversal of points “b” and “c” suggests that the expected cost of the 3-m barge is now about £5 million more than the expected cost of the 1.6-m barge. However, the long tail for the 1.6-m barge implies much more risk, still comparable to a 10% chance of an extra £100 million. The key question is should this extra risk be taken?

A basic Markowitz approach implies that both options are risk efficient and the choice is a matter of decision-taker preference—the board needs to use an implied decision function based on the preferences of individual members to make a choice without necessarily agreeing about preferences. The position here, and the point of the IBM story, is that the choice should be a matter of corporate “risk appetite” [8], determined via risk efficient assessment of all projects and other corporate operations. Knowing some £1000 million projects could double in cost, BP used risk sharing partnerships with other oil companies to share big project risk. Having gone to this trouble and expense it was clear that it did not make sense to add £5 million to the expected cost to avoid a £100 million risk. BP would take an “enlightened gamble” in a case like this, to aggressively maximize profit within their risk appetite. IBM had to identify a suitable corporate risk appetite and persuade all decision-takers to aggressively take risk to remain competitive and avoid the risk of IBM going out of business because IBM was too expensive. Taking risk up to the limit of corporate risk appetite involves “enlightened gambles” illustrated by the Fig. 4 situation.

Once Figs 2 and 4 were understood in terms of the crucial role of enlightened gambles, “enlightened caution” was discussed. It was pointed out that the actual hook-up took place in late October, the weather was good, and it was evident after the fact that BP could have got away with the 1.6-m barge. Given the analysis associated with Fig. 2, it was clear that the project manager had done a good job, made the right barge choice, and been lucky with the weather. It was also clear that had no analysis been done and a 3-m barge chosen, the project manager would

have been accused of wasting money. Further, it was clear that if no analysis of this kind was available, any experienced project manager would see the 1.6-m barge as the best choice in terms of their career interests, even though it is a risk inefficient, “unenlightened gamble” in corporate terms. “Enlightened caution” involves avoiding risk inefficient gambles.

Optimal risk-taking has to address corporate risk appetite in terms of a risk efficient approach to overall corporate risk. It has to make the formal analysis approach discussed here available to support high profile enlightened gambles and caution. And it has to ensure that the concepts are used informally for all other choices at all levels of projects and operations decisions. In effect, it has to distinguish between good luck and good management, bad luck and bad management, for all decision making. This involves a culture change for many organizations—formal analysis processes and tools are only part of what is needed.

IBM recognized this and directly managed a culture change, introducing risk management tools and processes as part of a culture change. Moving from a risk-averse culture to a risk-taking culture was the key goal, with a view to significantly improving profitability. An appropriate corporate risk appetite and linked management of individual behavior is a central issue for most organizations.

PROJECT AND OPERATIONS CHOICES IMPACTING STRATEGY AND OTHER FRAMEWORKS

The basic Markowitz mean-variance approach to portfolio selection [6] as generalized earlier in this article needs to be linked to conventional decision analysis [9] and scenario building [10] frameworks. This discussion also needs linking to a generalisation of the problem formulation discussion initiated earlier and bottom-up strategy formation. Consider an example.

Stan, the sole owner of a UK firm, which manufactured injection-molded plastic point-of-sale displays, wanted to consider buying a new type of injection molding machine, which

Table 1. Payoff Matrix for a Simple Machine Choice Example

Action: Size of Machine	Event: Market Growth				Criteria: More?	
	Zero	Poor	Moderate	Good	Expected Payoff	Standard Deviation
No machine	0	0	0	0	0.0	0.0
Small	−1	3	2	1	1.6	1.5
Medium	−2	2	5	2	2.1	2.4
Large	−4	−2	2	8	0.3	3.9
Probabilities	0.20	0.35	0.30	0.15		

could produce a completely new type of display product. Three different production rate capacity “sizes” of machines were available. All were highly automated, sophisticated, and expensive. There were differences in associated technologies, but Stan did not see these as important. The new point-of-sale products might compete with current products, but they would probably complement them, and manufacturing them would probably strengthen Stan’s overall marketing position. Space for any of the new machines was available, staffing implications for all three raised no obvious difficulties, and the key issue from Stan’s perspective was matching the size of machine to the size of the market, because the size of the market was very uncertain.

Table 1 is a classic payoff matrix portrayal of Stan’s choices, summarized from an extended discussion in Ref. 5, Chapter 9.

Table 1 involves four mutually exclusive action choices—a large, medium, or small machine plus the no machine option. Table 1 also involves a scenario or event breakdown of the uncertain market size growth—zero, poor, moderate, and good, with probabilities for each, assumed to be common to all four actions. Further, payoffs for each are shown in £100,000 units, an estimate of the net present value (NPV) of additional future profits after costs. Two criteria are also shown, expected payoff and standard deviation. A very relevant question is posed—more criteria?

If maximizing expected payoff is the only criteria, the optimal choice is the medium machine. If a basic Markowitz approach is adopted, standard deviation is a simpler measure of risk than variance, and Table 1 indicates standard deviations. In these terms the small and medium choices plus the no plant

are risk efficient—only the large plant can be rejected as risk inefficient.

The basic mean-variance or mean-standard deviation approach does not yield a unique choice. This is a conventional decision analysis argument for ignoring a mean-variance approach. If risk may matter, this is sometimes the rationale for moving to an expected utility approach at this point, normally without actually considering the mean-variance alternative or other alternatives.

However, say we move directly from Table 1 to Fig. 5, adopting the option choice format of Figs 1 and 2 to portray the information in Table 1.

Figure 5 takes the discrete point value estimates of Table 1 literally. For example, the large machine has a minimum value of −4 associated with a 0.2 probability, and the next higher value is −2, so the cumulative probability curve rises from 0 to 0.2 at a payoff of −4, then moves horizontally to a payoff of −2. Figure 5 uses a dashed line for the large machine option because it is not risk efficient. Figure 5 also uses a bolder solid line for the medium choice because it is arguably the best choice unless Stan cannot afford to take the additional risk. This additional risk is portrayed by the dashed line relative to the bold solid line after they cross at a cumulative probability of 0.85—risk is not measured by a single metric, but the difference in risk associated with alternative choices is clearly portrayed in terms defined by whatever probability distribution working assumptions are appropriate.

It will be obvious that the true distributions approximated by the Fig. 5 model are smoother. It will also be obvious that Fig. 5 curves are not very easy to read, because

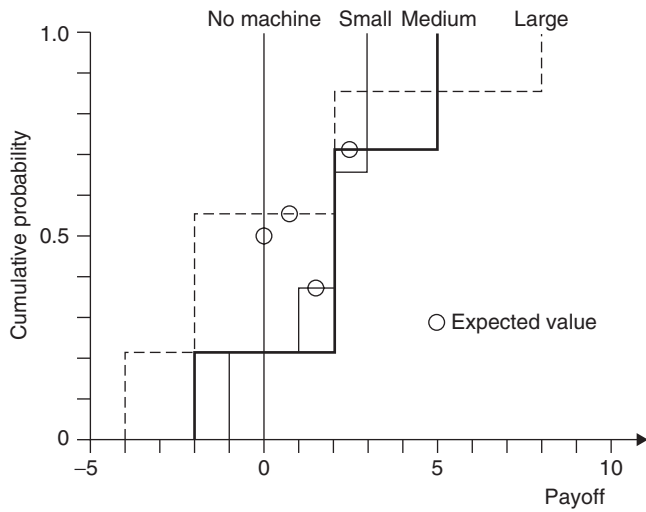


Figure 5. Plotting the cumulative distributions from Table 1.

of their angular nature. Indeed, the position taken here is the intermediate scenarios are not all that useful, and a better use of time and effort to this point is simple $p_{\min}$ and $p_{\max}$ scenarios for each option to define linear cumulative distributions. For present purposes, the Table 1 minimum and maximum values for each choice would serve as these $p_{\min}$ and $p_{\max}$ P10 and P90 approximations, yielding Fig. 6.

The P10 and P90 dotted lines were used to plot these simple linear cumulative probability distributions, for example, the large machine P10 was assumed to be -4 , with a

P90 of 8. Expected values can now be taken from the P50 dotted line, for example, the large machine P50 is 2 (the average of -4 and $+8$).

Apart from the relative ease of interpretation, an obvious difference is the large machine is now a risk efficient choice, with an increase in expected value from 0.3 to 2.0, while the medium and small machines both have lower expected values than on Table 1. All four options are now shown with the same weight solid line—none are risk inefficient and there is no obvious default choice. Advocates of conventional decision

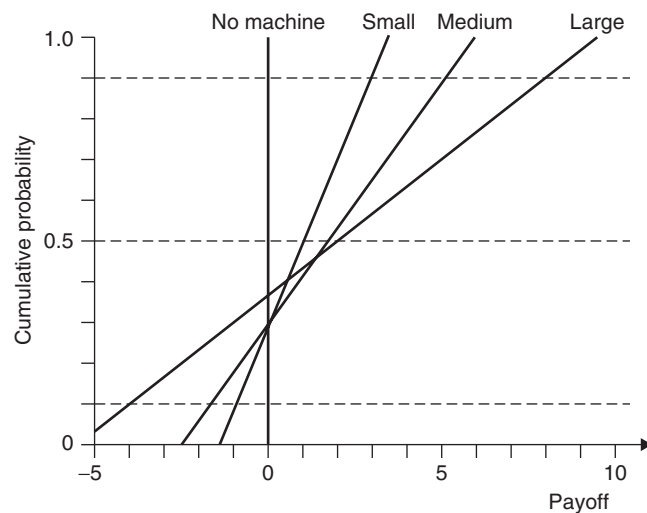


Figure 6. Simplifying the cumulative distributions from Table 1.

Table 2. Simple Example Alternative to the Table 1 Payoff Matrix

Size of Machine	Key Uncertainty Source: Market Growth		Expected NPV
	$p_{\min}$	$p_{\max}$	
No machine	0	0	0
Small	-1	3	1
Medium	-2	5	1.5
Large	-4	8	2

analysis might argue that information has been lost—Fig. 6 almost harps back to a classic “decision making under uncertainty” paradigm with the assumption that each scenario is equally likely—so the move from Table 1 and Figs 5 and 6 might be seen as counterproductive, moving in the wrong direction. The view taken here involves a series of counterarguments, but it might start by arguing that Table 2 is an effective simplification of Table 1, which is all that is needed to obtain Fig. 6. That is Table 2 is a better place to start, moving directly to Fig. 6. Table 1 involves additional information in an unduly constrained form, which makes Fig. 5 less helpful, so Table 1 is not as good a place to start.

One part of the reason Table 2 is a better place to start is the “lost information” associated with Table 1 is distorted by the assumption that all probabilities are common to all action choices—in practice a large machine will involve a lower unit cost, which will tend to create a larger market. This distortion is avoided if Table 1 is simplified to Table 2 as a basis for Fig. 6.

A second reason is empirical experience with scenario building approaches [10] suggests that two scenarios representing plausible extremes usually more effective than three or more, because more attention to what underlies two plausible extremes is time better spent than additional attention to central scenarios.

A third reason is empirical experience with risk management in various contexts [5] suggests that a first pass analysis which provides a very quick overview of the scale of all sources of uncertainty before detailed examination of what seems to matter on later

passes is the only efficient and effective way to proceed.

A fourth reason is the need to avoid closing down the options via an expected utility theory approach, which may be moving in the wrong direction. What may be needed is opening up the options via more clarity about what matters. Clues about what matters are often discovered by exploring important working assumptions. Working assumptions need identification, testing, and possible relaxation to yield these clues—what many decision sciences call “robustness analysis,” taking different forms in practice depending upon the school of thought. Consider illustrative examples of this approach, taken from Ref. 5, Chapter 9.

At a fairly fundamental level of analysis, we can start by recognizing that the real risk for Stan may not be variations in additional NPV in the terms measured thus far—it may be the threat to his current local market dominance if he cannot offer the new products at competitive prices. His current local market dominance may be the foundation of Stan’s business. Any threat to this dominance may be crucial. This is an issue, which the “no plant” payoff of zero for all market events clearly ignores, implicitly the case for the other options. The best defense for Stan may be attack, selling the new type of product to protect his overall market position with the most aggressive pricing feasible if necessary. Given this insight, it may be obvious that “no machine” as defined by Table 1 is not a viable choice, but neither are any of the other mismatches between machine size and market size. Indeed, it should become obvious that the four options of Tables 1 and 2 are not mutually exclusive, and it may be possible to buy capacity on someone else’s machine or sell capacity on a machine which is too big for the current point-of-sale display market. This insight, in turn, might suggest buying capacity until the potential market size becomes better defined, then moving to an appropriate machine or machines, with a view to using any spare capacity in some other way.

At an intermediate level of sophistication, uncertainty associated with competitors and feasible market prices might be explored,

using layered source of uncertainty format graphs like Figs 1 and 2 to break down revenue and cost sources of uncertainty, possibly in a nested structure.

At a more sophisticated level of analysis, this breakdown might be related to a variant of multiple attribute value theory (MAVT) approaches, considering the desirable and undesirable attributes of the different machines. This might reveal that the large machine involves leading edge advanced technology, with patents in place and excellent performance with guarantees from a high reputation manufacturer. The medium and small machines might be relatively old fashioned and unreliable. This might imply high resale values and low depreciation for the large machine. All these features in conjunction with the lowest variable cost per unit produced and the lowest purchase cost per unit of production capability may make competing against anyone with a large machine in any relevant market very difficult—the real threat Stan needs to avoid. Crucially, the large machine manufacturer may be non-EU, with limited EU penetration, and have no UK sales presence—Stan might be able to negotiate a UK franchise for sales of the large machine on the back of purchasing one which could be used as a demonstrator with trial capacity for sale. Best of all—Stan might really enjoy the opportunities and such a move might provide—most good entrepreneurs are motivated by the “fun” involved in this kind of opportunity.

The [5, Chapter 9] discussion leads to this conclusion considering a range of tools which might generate this kind of insight. Approaches discussed include variants of MAVT as advocated by Belton [11] and others. Multiple attribute utility theory (MAUT) as commonly linked to decision analysis by Keeney and Raiffa [12] and others is rejected because embedding risk in the metric used for each attribute via a utility function is not a suitable working assumption in many circumstances—obviously the case when non-measurable attributes are involved or more than one person is involved and agreement about preferences is not feasible. More generally, transparency is lost, which the

discussion of curves like those of Figs 1–4 and 6 enhances.

Some advocates of MAUT may see this and the rest of this section as contentious. A structured case to persuade them to alter their framing or favorite working assumptions is beyond the scope of this article, but it is important for all readers to understand that the kind of generalization of Markowitz’s risk efficiency ideas advocated earlier is not compatible with expected utility assumptions. As argued by Williams [3], practical ORMS requires compatible multimethodologies; so, choices need to be made about what features of conventional decision analysis we want to use in an integrated manner with appropriate features of conventional portfolio and scenario analysis. Very simple versions of basic Markowitz models and expected utility approaches are comparable, but not the generalizations discussed earlier.

A key linked point is the need to reject a conventional decision analysis payoff matrix, like Table 1, as a place to start when analyzing decisions in practice, if the associated working assumptions are not optimal for the choice being considered. The Table 1 framework reflects the conventional decision analysis starting point for framing assumptions—selecting a unique optimum from a predefined set of mutually exclusive options evaluated in terms of three or more discrete event scenarios common to all options using restrictive framing assumptions about risk and value trade-offs. A better place to start may be a simple table of options like that of Table 2, and graphs like those of Fig. 6, employing the option choice format of Figs 1 and 2. Each of the options in the Table 2 format can be considered a potential “project” or “operation” strategy, and the integration of two or more of these potential strategies can be viewed as a program or portfolio of projects and operations defining the development of an emerging strategy. All the analysis possibilities underlying layered source of uncertainty graphs like Fig. 3 as used by BP and others are part of the possibilities available from this starting position, plus a range of other exploratory approaches, like MAVT, in the very general framework discussed in relation to Fig. 1.

The focus of an approach providing overall optimality is seeking clarity about what really matters in an effective and efficient manner. Robustness analysis is a key part of this—understanding the role and importance of working assumptions. Table 1 and a conventional approach to decision analysis is a dead-end relative to Table 2 and an approach that opens up the options.

Another key point here is the compatibility of the Table 2 and Fig. 6 approach to decision analysis extended to decision trees as a basis for multiple stage decision choices and value of information concepts. Each decision node in any decision tree can be modeled in these terms. Multiple level decision trees cannot be reduced as simply—conditional choices may need to be addressed considering all branches of the tree—but in practice, this is a price worth paying for the clarity provided by avoiding restrictive assumptions. At any level in a decision tree choices are not mutually exclusive or uniquely defined, and their definition can be better shaped if it looks like it might pay to do so.

Other frameworks not mentioned here also need integration with respect to useful compatible features. Portfolio selection, decision analysis, and scenario building are simply the three key starting positions for addressing most risk mitigation and risk-taking in practice. Examples include influence diagramming to address dependencies and systems dynamics modeling to consider forensic risk management [3].

TOP-DOWN STRATEGY FORMATION

Some strategy formation can be driven bottom-up as illustrated in the last section, but top-down frameworks are also needed in many contexts. Consider another example.

In the 1990s, Ontario Hydro developed a 25-year strategic plan that included a significant development of their portfolio of nuclear power stations. Their proposed strategy might have been a good idea, but its presentation framework had a fundamental difficulty—an inappropriate view of risk and optimal mitigation of that risk. The plan was withdrawn when this difficulty became obvious.

An electric utility like Ontario Hydro needs to manage two kinds of risk of interest here in quite different ways, explored in detail in a report to the commission considering it and subsequent publications [5, Chapter 11], outlined here.

One is risk associated with the utility's inability to forecast which sources of electric power will be cheapest in 25 years time—renewable sources of various kinds, nuclear, oil, gas, and so on. Legislative issues and political pressures are involved as well as market forces. The only way to optimize in terms of this risk requires a generalized portfolio selection approach—most utilities cannot afford to “put all their eggs in one basket.” Optimal risk mitigation is virtually by definition a portfolio selection problem addressing efficient diversification—aiming for a position at the 25-year planning horizon which involves that mix of sources of electric power which minimizes the risk for any given expected cost, and choosing a level of risk that is appropriate for the organization.

A utility can change their view of the optimal mix to aim for as time evolves and current views of expected values and associated risk change. But it has to move toward that target mix as new units are added to meet rising demand and the need to retire existing units. Assessing the gap between the current target and the current position is the basis for deciding the sequence of new units to be added.

A second risk is introducing new units in the selected sequence too soon or too late—a timing issue best considered in the generalized decision analysis framework discussed earlier.

Most organizations have to blend top-down analysis of this kind with bottom-up analysis of the kind discussed in the previous section to achieve an overall optimum in their approach to strategy formation. Their diversification issues may be across product markets, input suppliers or areas of research—but diversification is a common risk mitigation issue.

Stress testing key assumptions may also involve top-down strategic consideration of scenario analysis, as developed by Shell [10]

to test corporate strategy against very high or low oil price scenarios. The relevant scenarios may include external circumstances which are not directly measurable, like a collapse of interbank lending because of a loss of trust if the organization is dependent upon this kind of finance, or a collapse of broader markets, or major political events.

Overall integration also has to address issues like the role of discounting in decision making [13 and 5, Chapter 8], safety [5, chapter 7] and internal and external contracts [14 and 5, Chapters 5 and 6].

Overall optimization requires overall integration of all this decision making, using compatible framing assumptions.

CONCLUSION

Optimal risk mitigation and risk-taking requires an optimal set of framing assumptions for all organizational contexts and optimal working assumptions specific to each context and application developed on an iterative basis as analysis progresses and insight is gained—ORMS processes in general are learning processes, which do not pursue a “right first time” policy. Simple tools like those illustrated by Fig. 1 may suffice, but the framework employed must be general, and an optimal approach usually requires a level of sophistication somewhere between that illustrated by Fig. 1 and that illustrated by Figs 2 and 3. Even the Fig. 1 analysis was within a sophisticated general framework, addressing subtle issues not relevant to the discussions on Figs 2 and 3. Understanding and exploiting the possibilities underlying Fig. 4 requires attention to behavioral issues as well as appropriate tools and processes. Problem formulation issues and the links between strategy formulation and more detailed operations and project management choices is complex, but they can be addressed within the same framing assumption set, as illustrated in the last two sections. Proper resolution of potential controversy associated with this topic in the context of Figs 5 and 6 is beyond the scope of this article, but appreciation of the need for an integrated set of compatible framing assumptions and tools

is essential for sound ORMS practice [3]. A somewhat different controversy, for the same basic reason, is addressed in Refs 1 and 13.

Acknowledgments

Figures 1–4 are reproduced from Ref. 4 with permission from John Wiley & Sons.

REFERENCES

1. Chapman CB. Key points of contention in framing assumptions for risk and uncertainty management. *Int J Proj Manage* 2006;24:303–313.
2. Sitkin SB, Weingart LR. Determinants of risky decision-making behaviour: a test of the mediating role of risk perceptions and propensity. *Acad Manage J* 1995;38(6):1573–1592.
3. Williams T. *Management science in practice*. Chichester: John Wiley & Sons, Ltd.; 2008.
4. Chapman CB, Ward SC. *Project risk management: processes, techniques and insights*. 2nd ed. Chichester: John Wiley & Sons, Ltd.; 2003.
5. Chapman CB, Ward SC. *Managing project risk and uncertainty: a constructively simple approach to decision making*. Chichester: John Wiley & Sons, Ltd.; 2002.
6. Markowitz H. *Portfolio selection: efficient diversification of investments*. New York: John Wiley & Sons, Inc.; 1959.
7. Chapman CB. Large engineering project risk analysis. *IEEE Trans Eng Manage* 1979;26: 78–86.
8. HM Treasury. *The Orange Book: Management of Risk - Principles and Concepts*. 2004. Available at www.who.int/management/general/risk/managementofrisk.pdf. Accessed 2008 Oct 16.
9. Raiffa H. *Decision analysis: introductory lectures on choices under uncertainty*. Reading (MA): Addison-Wesley; 1968.
10. Schoemaker PJH. Scenario planning: a tool for strategic thinking. *Sloan Manage Rev* 1995;36:25–40.
11. Belton V. Multiple criteria decision analysis—practically the only way to choose. In: Hendry LC, Eglise RW, editors. *Operational research tutorial papers*. Birmingham: Operational Research Society; 1990.
12. Keeney RL, Raiffa H. *Decisions with multiple objectives*. New York: John Wiley & Sons, Inc.; 1976.

13. Chapman CB, Ward SC, Klein JH. An optimised multiple test framework for project selection in the public sector, with a nuclear waste disposal case-based example. *Int J Proj Manage* 2006;24:373–384.
14. Chapman CB, Ward SC. Developing and implementing a balanced incentive and risk sharing contract. *Constr Manage Econ* 2008;26:659–669.

OPTIMIZATION AND DECISION SCIENCES (ITALY)

ROBERTO TADEI
Politecnico di Torino,
Corso Duca degli
Abruzzi, Torino, Italy

AIRO, Associazione Italiana di Ricerca Operativa—Italian Operations Research Society—Optimization and Decision Sciences, is a purposeful scientific organization comprising approximately 400 researchers as well as several University departments, companies, and other bodies that are interested in operations research (OR) in Italy. AIRO also has many foreign members. In this brief article, we want to outline the most important moments of the story of our association.

AIRO was founded on 20 April 1961 in Rome, at ISTAT (Istituto Centrale di Statistica—Italian Institute of Statistics) headquarters, owing to the work carried out by Prof. Benedetto Barberi, General Manager of ISTAT. A group formed by researchers, scientific research bodies, and big Italian enterprises accepted Prof. Barberi's invitation to participate in the formation of AIRO. These people include, but are not limited to, Cav. Lav. (Knight of Labour) Carlo Faina (President of Montedison), Giuseppe Pero (President of Olivetti), Prof. Giovanni Polvani (President of CNR—The National Research Council), Prof. Giuseppe Pompilj (Director of the Probability Calculation Institute in Rome University), Prof. Vittorio Valletta (President of FIAT), and the member of Parliament Prof. Enrico Mattei (President of ENI). Prof. Enrico Mattei (President of ENI), an important person in the Italian industrial history, is amongst the most ardent supporters of the importance of the operations research methods in economic and financial organizations.

In the same year, Prof. Barberi announced the first Work Days (Annual meeting of the

Association), which has become the primary meeting in Italy that deals with operations research from both the methodological and application perspectives. From 1961 to 1968, "Quaderni R.O." was published: the purpose of this seven-issue series was to collect the annual contributions of the Work Days carried out in those years; each of these issues focused on a different theme.

At the beginning of the second half of the twentieth century, industrialized nations faced major problems linked to economic planning, increases in per capita income, and reductions in regional and social economic gaps. This was brought to the attention of operations researchers and ORSA (Operations Research Society of America) in the early 1960s, and was dealt with again on the occasion of the second Conference of IFORS (International Federation of Operational Research Societies). Consequently, the identification of the operations research as an urgent economic planning instrument was introduced to the scientific and productive world through the statement submitted by a group of researchers inside the United Nations at the beginning of 1961. The interest of the United Nations dates back to a vote given by the Economic and Social Board of that organization for a systematic commitment to make the most modern and verified instruments available for the newly developing countries to accelerate their economic development. With regards to Italy, it was Prof. Barberi's desire that operations research, through AIRO, would play an important role in the construction and implementation of the Italian economic development program.

Since the beginning, AIRO has been organized in local chapters that have depended each year upon the assembly regarding the activities carried out that year. The first chapters were established in Turin, Genoa, Milan, Padua, Rome, and Naples.

AIRO publishes an information bulletin that provides information about scientific operations research activities both

at the national and international level. Furthermore, it publishes a series of volumes that explain OR concepts and methods.

In 1970, Prof. Barberi left his office as AIRO President, and, at the members' meeting, Eng. Armando Corso was elected as his successor.

In addition to the activities of the local chapters, new initiatives are carried out, including (but not limited to) the workgroups "Applications of Operations Research to the control of the environment," "Teaching Operations Research in Italy," and "Operations Research in banks."

The information bulletin is accompanied by the publication of the journal "Ricerca Operativa" (Operations Research), published through the support of the National Research Council, which features articles and reviews of scientific books.

In the beginning of the 1970s, there was a considerable increase in interest toward OR by Italian universities. In these years, the first university chairs of operations research were established, and many university courses were started in this discipline.

After two consecutive terms of office, Eng. Corso left the AIRO presidency, and, in 1975, Eng. Bruno Martinoli was elected in his place. AIRO began its cooperation with IFORS and EURO (The Association of European Operational Research Societies) at that time; seven European associations, including AIRO, cooperated in the management of EURO at that time.

At the same time, EURO began publishing the "European Journal of Operational Research," which AIRO has actively supported through qualified cooperation, special issues, and participations in workgroups.

Moreover, at the same time, AIRO began working cooperatively with other Italian scientific communities such as AICA, AISL, ANIPLA, SIRI, and ACM Charter. These activities included the unitary meeting held on March 1977 about "Informatics distributed in the industrial production." OR theoretical bases in those years ranged from matrix and linear algebra to probability, statistics, stochastic processes, calculating machines science, microeconomics, business administration, organization theories, and

behaviour sciences. In particular, it developed mathematics models and methods for linear programming, dynamical programming, queueing theory, network theory, game theory, and simulation.

During the years in which Eng. Martinoli was the President, AIRO cooperated with the organization as a member of EURO and IFORS in many international meetings. These meetings included EURO, Stockholm, 1976; IFORS, Toronto, 1978; EURO, Amsterdam, 1979; EURO, Cambridge, 1980; IFORS, Hamburg, 1981; and EURO, Lausanne, 1982.

With regards to cooperation with similar Italian scientific organizations in order to exchange information, share experiences, and promote common activities, AIRO established a committee composed of the presidents of the following associations: AICA, AIRO, AIS, ANIPLA, and SIRI.

At the beginning of the 1980s, there was a very low presence of corporate applications, as well as an absence of submissions to the annual AIRO conferences from industries where OR had its main development. Conversely, there were many submissions from university and research bodies. This dual-speed development situation of operations research in universities and the private industry continued throughout the 1980s. This brought considerable growth at the university level, but also led to a progressive decrease in OR corporate groups.

In September 1982, the members elected Prof. Massimo Merlino as AIRO President. Prof. Merlino gave the Presidency a more managerial orientation, and 20 important Italian companies joined AIRO. The president established a permanent advisory committee for corporate members, in which any company could participate through a representative. This committee was created to direct the AIRO policy toward the private industry. In particular, many Italian corporations reorientated themselves toward a higher quantitative culture, and they made a higher investment in the local AIRO chapters and the training of young graduates.

During this time period, there were many influential AIRO seminars on various managerial topics such as distribution problems, organizational issues linked to industrial

automation, interactive decision support methods, and combinatorial optimization.

In 1984, the National Operations Research Annual Prize for the best publishing in operations research was established and was given to Italy during the same year.

Corporate membership in AIRO increased by 30%. Research institutions, already strong, continued to grow. AIRO multiplied its central and local activities and improved their quality, enabling the organization to reacquire a prestigious position among Italy's large enterprises and institutions (similar to the position it held in the 1960s). In this period, the President appointed three illustrious honorary members: Senator Prof. Siro Lombardini, Prof. Antonio Ruberti, and Prof. Mario Volpato. The following year Prof. Giuseppe Brambilla was also appointed as an honorary member of AIRO.

Italian companies began looking for more operations research training courses; these organizations showed particular interest in introductory courses on decision techniques. The main methods of corporate interest were production planning, simulation of automation line, decision support systems, expert systems for planning and management of production lines, optimization techniques, and statistics and simulation applied to the factory.

The main element that now distinguished AIRO from other similar associations was the simultaneous presence of university researchers and professionals employed by the private industry and government.

Prof. Merlino eventually left the presidency after two consecutive trienniums recommending the creation of new openings and opportunities for exchange of ideas and information with other disciplines and cultures.

In October 1988, the members elected Prof. Paolo Toth as AIRO President. As a world-famous researcher, he was highly esteemed and appreciated by all those who dealt with OR both in Italy and abroad. In 1998, he was awarded the EURO Gold Medal for scientific merits. He held the office as EURO President from 1995 to 1996 and IFORS President from 2001 to 2003.

Pursuant to a request submitted by AIRO member companies for training courses that addressed institutional deficiencies in the most advanced operations research techniques, AIRO and the Institute for Systems Analysis and Computer Science (IASI) of the National Research Council began to carry out a programme to regulate Operations Research Courses. The courses were extremely popular and became a strength of the association. They were extremely successful due to the high educational level provided by qualified instructors and the presentation of optimization software packages.

AIRO simultaneously implemented a policy to increase exposure to operations research among young people (university and high school students) to create interest and participation in the subjects of our discipline. The actions that were carried out ranged from the identification of the topics that must be dealt with by OR to the choice of pilot schools for experimentation. The experimentation was supported by the Ministry of Public Education and had the training of teachers and subsequently students as its objective, providing them with the methodological tools that would enable them to deal with various kinds of decisional problems that young people have to face in their future work activities.

In order to promote the connection between scientific research and industrial applications, AIRO and Ferrovie dello Stato (Italian State Railway Institution) announced two International contests (F.A.R.O. and F.A.S.T.E.R.). These contests were intended specifically to promote the development of large-scale resource optimization methods applied to train management and personnel rostering.

In September 1995, Prof. Toth left the presidency and the members elected Prof. Giorgio Gallo as AIRO President. Training activities continued intensively through company-oriented courses. In this period, companies faced continuous change in the market, fast technological development, and the necessity of adjustment to the production processes. The significance of operations

research was increased through the improvement in computer science technologies that offered increasingly effective contributions to the solution of the abovementioned issues.

After recognizing the opportunity for a better integration with similar European associations, the President suggested the publication of a journal by AIRO and the Belgian and French associations. In 2001, the journal "Ricerca Operativa" ended its publishing with the issue of no.100, and the international journal "4OR" was published by Springer. This journal promptly acquired a high scientific level. Prof. Silvano Martello represented AIRO as one of the founding editors-in-chief of this new journal.

In those years, AIRO began publishing a news bulletin named "AIRONews" in both print and on the Internet. It was received very favourably by members and followers.

Pursuant to the proposal of the AIRO Steering Committee, the members' meeting appointed Prof. Lucio Bianco and Prof. Sergio De Julio as honorary members.

In September 2001, the members elected Prof. Roberto Tadei as the AIRO President.

He set the following goals for AIRO:

- analysis of corporate training needs;
- offering short courses, both inter-corporate and corporate, in the following fields: assessment of the marketing operations, logistics instruments, instruments for simulation and optimization, and methodologies and techniques for project financing and project management;
- publication, both print and on the Internet, of a monthly information brochure for the enterprises that illustrate successful applications of the operations research within the companies;
- setting up "Windows about Operations Research," a presentation by research groups in Italian universities of the products that they have developed for the private industry;
- realization of a service to facilitate the joint participation of companies and research centres in the European and National research projects.
- The reinforcement of the connection with young people through the extension of graduation awards (both for high schools and master's degree) as well as the establishment of a contest for high school students to increase the involvement of younger students and their teachers in our discipline.
- The creation of on-line databases that could collect employment applications and offers, making graduates aware of companies looking for individuals trained in operations research and of available training programs in operations research.
- The creation of a Thematic Chapter "AIRO Giovani" (AIRO Youth) to (i) support a network of relationships and competencies among young people and (ii) propose training for PhD courses and initiatives that aim at specialization and extension of research competencies in young people.

Furthermore, alongside the journal "4OR," the news bulletin AIRONews enforces the objective of spreading OR in Italy. There are two main branches, the first oriented to industry and the second oriented to education (universities and schools). The most qualifying proposals consist of submitting not only technical articles but also informative ones, publishing thoughts about daily life, taking care of the diffusion of the news bulletin so that specific people in companies and bodies are reached. This also enables cooperation with similar journals of other associations.

In addition, the internet publication ImpresAIRO is initiated; this is a readable journal intended for companies and is an instrument through which companies communicate with each other to show the relevant benefits of applying operations research methods to production processes.

Raising awareness and appreciation of OR among young people was one of the most important objectives of AIRO in those years. Indeed, initiatives were developed for the following:

In 2003, another new initiative was carried out. This initiative established an annual meeting, called *AIROWinter*, to be held during the winter in a mountainous location (the first event took place in Champoluc, Aosta Valley, Italy). The objective of these meetings was to gather Italian and foreign research groups working in the fields that were important to Prof. Mario Lucertini (a famous OR researcher who died prematurely in 2002 and to whom this meeting was dedicated). The meeting was organized in the following year to also remember Stefano Pallottino, another illustrious colleague who died prematurely in 2004.

On the occasion of the “Work Days” in 2005, as acknowledgment of the noteworthiness support given to the association when they were holding the office of president, the following past presidents were named as AIRO honorary members: Prof. Armando Corso, Eng. Bruno Martinoli, Prof. Massimo Merlini, Prof. Paolo Toth, and Prof. Giorgio Gallo.

On the occasion of the “Work Days” in 2006, pursuant to the proposal of a jury composed of the three editors-in-chief of the journal “4OR,” there was the appointment of the first AIRO fellow members: Prof. Giorgio Gallo, Prof. Pierre Hansen, and Prof. Paolo Toth.

Other Italian scientific associations (in particular AMASES and SIMAI) have been contacted in order to establish a common cooperation and to promote the realization of the Federazione Italiana di Matematica Applicata (FIMA), which is the Italian Federation of Applied Mathematics. FIMA has two purposes. First, on a national level, FIMA organizes common meetings, provides integrated training courses, and coordinates the activities of individual associations. Second, on an international level, FIMA

aims to engage with similar federations and associations of other European countries. Indeed, in addition to the research projects, in which it is hard to participate in the framework of scientific associations, there are many other projects aimed at propagating the results and training young people and companies. The structure and the organization of such associations can play an important role.

As its first initiative, in January 2006, FIMA organized the First International Conference FIMA with the theme “Models and Methods for Human Genomics,” in Champoluc (in the Aosta Valley of Italy) to encourage greater Italian and foreign participation in this area. On the basis of the works presented during the meeting, FIMA published a special issue of the journal *Computers and Mathematics with Applications*, entitled “Modelling and Computational Methods in Genomics.”

In September 2006, Prof. Renato De Leone was elected as AIRO President, and he still holds this office. This President has launched many interesting initiatives in Italy and abroad that promise to play very important roles in the future of the Italian Association of Operations Research.

Acknowledgments

The paper is based on “Storia dell’Associazione Italiana di Ricerca Operativa”, by Agnese Martinoli in *Scienza delle decisioni in Italia: applicazioni della Ricerca Operativa a problemi aziendali*, ECIG 2008, pp. 21–32. Mrs. Martinoli, now retired, has held her office as AIRO Secretary for almost 40 years, proving to be an indispensable reference for the Association, its presidents, and to all of the members. We would like to express our most sincere thanks to Mrs. Martinoli.

OPTIMIZATION FOR DISPATCH AND SHORT-TERM GENERATION IN WHOLESALE ELECTRICITY MARKETS

GOLBON ZAKERI

Electric Power Optimization Center,
Department of Engineering Science,
Faculty of Engineering, University of
Auckland, Auckland, New Zealand

INTRODUCTION

The first example of energy market concepts and privatization of electric power systems took place in Chile in the early 1980s. The idea behind the Chilean model was to bring rationality and transparency to the operations of the power system that would ultimately be reflected in power prices. Other rationales for the eventuation of electricity markets include better reliability (e.g., in the case of Argentinean electricity market) and signaling appropriate levels of investment in infrastructure in the energy sector. Today many countries and jurisdictions rely on electricity markets to meet their electricity needs. These include UK, Australia, New Zealand, the Nordpool market consisting of Scandinavian countries, Brazil and other South American countries, as well as many jurisdictions in the US.

Electricity markets typically consist of five main sectors: the system operator (sometimes there is also a separate transmission system operator), generators, consumers, distributors, and regulatory bodies overseeing the regulations and operations of the market. While operations research is utilized in each one of these sectors on a day to day basis, to keep this article to a reasonable length and be informative, we only discuss the first two sectors. In what follows, we describe the operation of an electricity market. We then discuss the generation sector and describe how operations research is utilized within that sector.

MARKET OPERATION AND THE SYSTEM OPERATOR

Electricity is not a storable commodity. It is injected into a transmission grid at certain nodes of that transmission grid often referred to as grid injection points (GIPs), and flows through the grid complying with physical constraints. Electricity is withdrawn at grid exit points (GXPs) and delivered to consumers. Owing to the physical constraints on the flow of electricity, in all electricity markets, the dispatch of electricity generation is left to an independent system operator (ISO). In most electricity markets, an additional function of the ISO is to determine the price of electricity at different nodes of the transmission network.

Economic Dispatch

Typically, in a wholesale electricity market, in each period of the day, each generator offers in generation quantities for each of its plants (possibly located at different GIPs), at certain prices. In its most general form, the generation offers are supply functions (also known as offer curves) denoted as $p = C(q)$, where $C(q)$ is the marginal price of producing quantity q . These supply offers are collected by the ISO. The ISO also estimates the demand over that period. The ISO then solves the so-called economic dispatch problem (EDP) which is also known as the optimal power flow problem. This is the optimization problem outlined in EDP below. For the readers familiar with optimization, EDP is a side-constrained network optimization problem where the objective is to minimize the total cost of production of electricity. The constraints of this optimization problem reflect that demand must be met at every node of the network, and that physical constraints such as transmission line capacities and Kirchhoff's laws govern the actual flows. Typically, system operators solve such an optimization problem on at least an hourly basis. Often reactive power modeling is left out of the

ISO's dispatch problem and the problem is in fact a direct current equivalent load flow model [1,2]. A general model for the ISO's EDP is formulated below.

$$\begin{aligned}
\text{EDP: } & \text{minimize } \sum_i \sum_{m \in \mathcal{O}(i)} \int_0^{q_m} C_m(x) dx \\
& \text{s.t. } g_i(y) + \sum_{m \in \mathcal{O}(i)} q_m \\
& \quad = D_i, \quad i \in \mathcal{N}, \quad [\pi_i] \\
& \quad q_m \in Q_m, \quad m \in \mathcal{O}(i), \\
& \quad i \in \mathcal{N}, \\
& \quad y \in Y.
\end{aligned}$$

We use i as the index for the nodes in the transmission grid. We use m as the index for the generators, and $\mathcal{O}(i)$ indicates the set of all generators located at node i . Generator m submits marginal cost curve C_m to the ISO and is called upon to supply quantity q_m . The demand at node i is denoted by D_i . Q_m indicates the capacity of generator m .

Here the components of vector x measure the dispatch of each generator and the components of the vector y measure the flow of power in each transmission line. We denote the flow in the directed line from i to k by y_{ik} , where by convention we assume that $i < k$. (A negative value of y_{ik} denotes flow in the direction from k to i .) It is required that the vector of line flows lies in the convex set Y , which means that each component satisfies the thermal limits on each line, and satisfies loop flow constraints that are implied by Kirchhoff's Law. The function $g_i(y)$ defines the amount of power arriving at node i for a given choice of y . This notation enables different loss functions to be modeled. For example, if there are no line losses, then we obtain

$$g_i(y) = \sum_{k < i} y_{ki} - \sum_{k > i} y_{ik}.$$

With quadratic losses (where r_{ik} denote the appropriate loss coefficient for the quadratic

loss), we obtain

$$\begin{aligned}
g_i(y) = & \sum_{k < i} y_{ki} - \sum_{k > i} y_{ik} - \sum_{k < i} \frac{1}{2} r_{ki} y_{ki}^2 \\
& - \sum_{k > i} \frac{1}{2} r_{ik} y_{ik}^2.
\end{aligned}$$

As indicated above, one of the functions of the ISO is to set the price. The price of electricity is determined as the shadow price π_i of the node balance constraints above, which indicates that demand must be met at all nodes. This price is the system cost of meeting one more unit of demand at node i . This method of determining the electricity price is sometimes referred to as locational marginal pricing (LMP). New Zealand and the PJM market in the US are examples of electricity markets with LMP.

It is worth noting that some wholesale electricity markets operate by assuming that demand and supply are located at the same node and trading takes place in that one node. This means that a single price of electricity is found. Nevertheless, in order to ensure that the demand is met at all nodes and that the flow complies with physical constraints, a balancing market would follow in real time where the residuals of the single node market are traded. The UK wholesale electricity market is an example of a single node market.

Unit Commitment

While the EDP described so far minimizes the total cost of generation based on the generators' offers, it is concerned with cost minimization over a single period. Besides the marginal cost of running generation units, other costs such as start up and shut down costs and constraints such as minimum up and down time for the generation units may need to be considered. In some markets, these concerns are left to the generators. They are to figure these costs and constraints and reflect them in the way they offer their generation into the market. Other markets, such as the PJM, have unit commitment constraints embedded, and the EDP runs over a time horizon spanning a day. For a comprehensive discussion on the unit commitment problem

and the problem of market design embedding unit commitment see Ref. 3.

GENERATORS

The generation sector in wholesale electricity markets usually consists of coal or gas fired thermal, nuclear power, hydro-electric, and/or other renewable (such as wind or solar powered) generation. As mentioned above, generators in a wholesale electricity market are required to submit an offer curve for each of their generation units. Therefore, the main question faced by any generator in a wholesale electricity market is how to offer into the market in such a way as to maximize its return. Before we explore this problem further we must make a distinction between two different types of generators, the so-called price makers and price takers. We should also note that since the future is uncertain and neither the competitor offers nor the demand is known, assuming risk neutrality, the generators will be maximizing their expected profits.

Price-Taker Generators

A generator is defined to be a price taker if the (nodal) price of power is not affected by the generation strategy of the generator. A true price taker is likely to have a small capacity for generation (relative to the rest of the market) with correspondingly little storage capacity. Therefore, the typical time horizon for the optimization problem is short to medium term. Here, the only significant uncertainty is in the electricity price process. Other factors such as inflows (in the case of a hydro generator) and the price of thermal fuel (in the case of coal or gas generators,) may be treated as deterministic as they are well forecasted over the short time horizon. If the electricity prices were independent from period to period, then the optimal policy would be to submit a generation offer so that dispatch is ensured if the electricity price is above expectation and no generation is dispatched (therefore, resources such as water are saved for the subsequent period) if the price is below expectation. However, this

independence assumption does not hold for electricity prices.

Pritchard and Zakeri [4] focus on a price-taker hydro generator operating a river chain. They have developed a time inhomogeneous Markov model for electricity prices (at a node); here the transition probabilities depend on the time of day and the time of week. They use this price process to set up and solve a stochastic dynamic program in order to determine the release policies for operating a river chain, so that the expected profit of a price-taker hydro-electric generator is maximized.

In a subsequent paper, Pritchard *et al.* [5] relax the assumption of limited storage. Here, they consider a model where weekly expected releases for a river chain are determined on the basis of a long-term model for electricity prices. This expected release, along with a standard deviation on release, is then passed to another optimization problem with a much shorter time horizon that focuses on a particular week. They then construct period to period offer stacks for the week in question using dynamic programming. This model integrates the long- and short-term plans and allows for target release levels (the expected release each week) as well as allowing deviations from this target should an opportunity present itself in the shorter time horizon model to take advantage of favorable prices.

Conejo *et al.* [6] develop a model that addresses similar questions pertaining to a thermal generator. The maximum output of a thermal unit is not available instantaneously. Thermal generators are bound by their ramp rate characteristics that dictate how the generator's available output rises toward its maximum output as a function of time. They must also comply with minimum up and down constraints that constitute them; once a unit is turned off it may not be turned on again until some time has elapsed and vice versa. Conejo *et al.* develop a mixed integer linear program that produces a bidding strategy for a price-taker thermal generator over the course of a day given a price forecast for electricity prices that day. This bidding strategy maximizes the generator's expected profit. It should be noted that

while most models in the literature assume that the generators are risk neutral and aim to maximize expected profits, there has been some research addressing production schedules in the presence of a risk attitude; see, for example, Ref. 7. For further reading on price-taker generation optimization see also Refs 8–10.

Price-Maker Generators

Owing to economies of scale and the spatial distribution of consumers, almost all electricity markets are oligopolies. Therefore, the most natural setting for a model is to assume that the offer strategies of a generator influence the (nodal) price of electricity. In such a setting, the generators' fundamental profit maximization question needs to be revisited. The price of electricity is no longer exogenous; hence, the profit optimization problem faced by the generator now must take into account how the actions of the generator influence the price of electricity. This problem was first addressed in a paper by Klemperer and Meyer [11], who were interested in modeling an oligopoly facing uncertain demand, where each firm bids a supply function as its strategy. This is in contrast to previous models in the economics literature where firms were restricted to strategize over their quantities only (Cournot models) or their prices only (Bertrand models) and allows a firm to adapt better to an uncertain environment. Green and Newbery address the same question but in the context of the British spot market [12].

To begin, let us assume that there are only two generators supplying the market (i.e., we are dealing with a duopoly) and suppose that the offer curve of the competitor is given by $q = S(p)$. Let us also assume that the demand curve is given by $q = D(p)$, that is, the market will absorb quantity q if the price is p . For their analysis, Klemperer and Meyer use the concept of the residual demand curve faced by the generator. Consider the curve given by $q = D(p) - S(p)$. This determines what quantity must be offered into the market if we desire the price to be p based on the demand curve and the competitor's offer strategy. The inverse of this curve describes how the price is influenced by the quantity we offer and is referred to as the residual demand curve. Therefore, for any potential offer quantity q , we can obtain the corresponding market clearing price p from the residual demand curve, and the product $q \cdot p$ will yield the revenue associated with that (potentially) offered quantity. With this information at hand, it is now easy to optimize our profits (see Fig. 1); we simply choose to offer in the quantity q that will maximize revenue less cost of production.

Recall that Klemperer and Meyer point out that supply functions allow a firm to adapt better to an uncertain environment. If there are multiple possible residual demand curves that a generator may face, the supply function response may allow selection of a point on each of these residual demand curves that would optimize the generator's profit given that residual demand curve

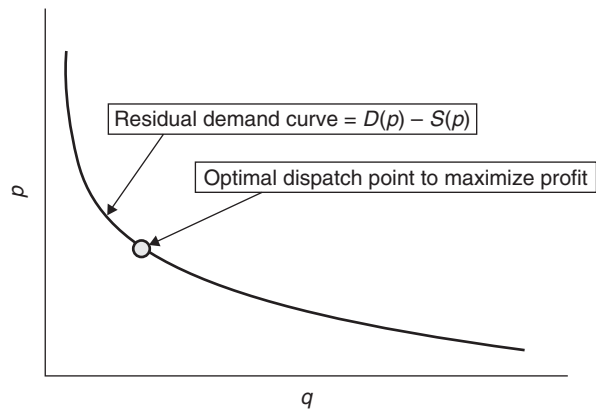


Figure 1. Optimal point for a generator to get dispatched along a residual demand curve.

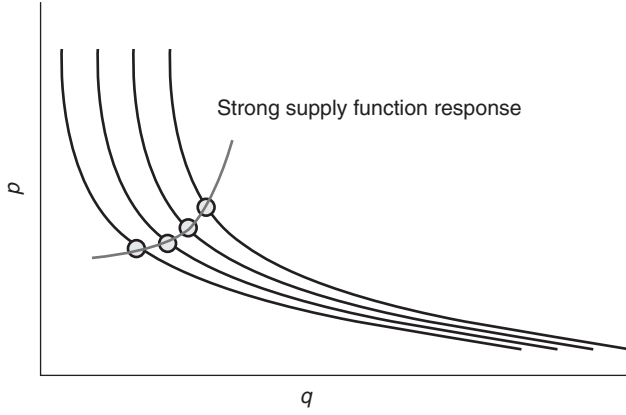


Figure 2. Building a strong supply function response from a distribution of residual demand curves.

has realized. This is referred to as a strong supply function response (see Fig. 2). A number of papers construct the residual demand curve by simulating the (single node) market and explicitly building the supply function response, see, for example, Refs 13 and 14. In Ref. 13, the residual demand curve takes on a step function form and the authors develop a nonlinear integer programming model of the generator's revenue optimization problem. They develop a combined coordinate search, branch and bound method, to solve this problem. Torre *et al.* exploit the nature of the previous problem to develop a more efficient solution method [14]. Baillo *et al.* further consider a robust price-maker decision problem that is an extension to the previous work [15].

In a sequence of papers Anderson and Philpott have also addressed the profit maximization problem of a price-maker generator under various assumptions. In Ref. 16, they assume that a price-maker generator knows its competitors' offer curves, but is faced with uncertain demand. They first establish the existence of a strong supply function response, for such a generator, that would be optimal for any realization of the uncertain demand. This strong supply function response is guaranteed to exist when the generation costs of the generator in question are increasing and convex, and the competitor offers are log concave. They discuss a procedure where the true aggregate offer stack of the competitors is approximated by a log concave function. Note that this (aggregate)

offer stack would be a step function in almost all real-world electricity markets. They construct a strong supply function response, S_g , for the generator in question. Subsequently, they approximate S_g in order to comply with market rules. Finally, they provide bounds on the performance of such an offer strategy.

In Ref. 17, Anderson and Philpott generalize their model by allowing uncertainty not only in the demand but also in the competitor offers. They introduce the concept of a market distribution function $\psi(q, p)$ pertaining to a specific generator at a specific transmission node. They define $\psi(q, p)$ to be the probability of not being fully dispatched if the generator submits a quantity q at price p . Let $R(q, p)$ denote the profit that the generator makes if it is dispatched quantity q at a clearing price of p . They demonstrate that if the generator submits the curve s and the pertinent market distribution function ψ is continuous, then the expected profit of the company is given by

$$V(s) = \int_s R(q, p) d\psi(q, p).$$

They proceed to provide conditions that guarantee (local) optimality of an offer stack s that would maximize $V(s)$. To address the question of estimating the market distribution function see Refs 18 and 19.

The work described thus far only deals with generators that are located at a single node of the market or alternatively assumes

that the wholesale market is a single node market. As noted in the section titled “Market Operation and the System Operator,” however, most wholesale electricity markets use locational marginal pricing where the price of electricity is different from node to node. To capture the effects of the transmission network, a generator must look at the variations in the prices from the dispatch problem EDP as a function of how it offers into the market. The revenue optimization problem is now posed as a bilevel program, or a mathematical program with equilibrium constraints (MPEC) and becomes a nonconvex optimization problem.

$$\begin{aligned}
 & \text{maximize } R(x, \pi) \\
 & \text{s.t. } (x, \pi) \in \arg \min \sum_i \sum_{m \in \mathcal{O}(i)} \int_0^{q_m} C_m(x) dx \\
 & \text{s.t. } g_i(y) + \sum_{m \in \mathcal{O}(i)} q_m = D_i, \quad i \in \mathcal{N}, \\
 & \quad q_m \in Q_m, \quad m \in \mathcal{O}(i), \quad i \in \mathcal{N}, \\
 & \quad y \in Y.
 \end{aligned}$$

Here x denotes the vector of quantities dispatched at each node if the generator offered at that node (or is 0 if the generator in question does not own generation at a particular node), and π is the vector of electricity prices. Note that the inner optimization problem, namely, the economic dispatch problem EDP can be replaced with its necessary and sufficient conditions for optimality as it is a convex problem (see, e.g., chapter 4 of [20]). In this case, the reformulation is referred to as an MPEC [21]. Several papers have addressed this problem and have developed techniques to produce local or global optimal solutions for it (see, e.g., Ref. 22). In Ref. 23, Fampa *et al.* develop a stochastic version of the above problem in which they consider various demand and market clearing scenarios. This results in a number of follow-on economic dispatch problems. They propose some heuristic methods for solving this problem. Pritchard also considers a stochastic version of the above problem [24]. In his model, the generator in question only owns generation assets at a single node of the transmission

network. He proposes a stochastic dynamic program to solve the generator optimization problem. It should be noted that the literature surveyed for price-maker generators is only concerned with short- to medium-term time horizons, where a reasonable distribution for competitor behavior and demand, or a reasonable estimate of the market distribution function, is available to a generator.

OTHER ELECTRICITY SECTORS

The focus of this short article has been on the generation side of wholesale electricity markets. Operations research is utilized in every sector of the electricity market, and in this section we provide a very brief overview of how it is utilized by the demand side and the regulators.

Consumers of electricity can be loosely classified as major or minor users. Major users of electricity are typically large industrial users who have the ability to observe electricity prices in real time. In many wholesale electricity markets, such users can also bid into the electricity market by specifying a demand curve where they indicate their willingness to consume at different prices. These consumers are similar to price-maker generators in that they have the ability to alter the price of electricity (through the amount that they consume). Frequently, they are not only able to withdraw electricity from a GXP but they may also be in a position to produce their own power through a gas or diesel generator say, or they may have flexibility to reduce their production of goods, which translates into reduction in consumption of electricity. Gomez-Villalva and Ramos [25] have developed a mixed integer linear program for such an industrial consumer in a wholesale electricity market.

For some major consumers, it may also be possible to shift the electricity usage from one period to another, albeit at a cost. Large industrial chillers may be able to cool their contents at earlier hours of the morning to lower degrees simply because the price of electricity is lower in those off-peak periods. The cooler may be able to retain the

low temperatures reasonably well and hence reduce its electricity consumption in the peak morning hours thereby reducing its total cost of consumption. This is referred to as load shifting. Middelberg *et al.* [26] study an optimal control model for load shifting with application to energy management of a colliery.

Regulators are interested in designing electricity markets so that they are as competitive as possible and an efficient and reliable supply is accomplished. They are also interested in diagnosing any ill functioning, such as abuse of power by generating firms, and introducing regulation that would prevent such behavior. They are therefore interested in models of steady-state behavior of the market. There is a vast amount of literature on such models ranging from models that derive Nash–Cournot equilibria and supply function equilibria to agent-based simulation models. For additional information, see Refs 11, 27–36.

REFERENCES

- Schweppe F, Caramanis M, Tabors R, *et al.* Market operations in electric power systems. Boston (MA): Kluwer; 1988.
- Oren SS, Spiller PT, Varaiya P, *et al.* Nodal prices and transmission rights: a critical appraisal. *Electric J* 1995;8(3):24–35.
- Hobbs BF, Rothkopf MH, O'Neill RP, *et al.* The next generation of electric power unit commitment models. Norwell (MA): Kluwer; 2001.
- Pritchard G, Zakeri G. Market offering strategies for hydroelectric generators. *Oper Res* 2003;51(4):602–612.
- Pritchard G, Philpott AB, Neame PJ. Hydroelectric reservoir optimization in a pool market. *Math Program* 2005;103(3):445–461.
- Conejo AJ, Nogales FJ, Arroyo JM. Price-taker bidding strategy under price uncertainty. *IEEE Trans Power Syst* 2002;17(4):1081–1088.
- Wang J. Risk constrained generation asset scheduling for price-takers in the electricity markets. Proceedings of 41st Hawaii International Conference on System Sciences; Hawaii: 2008. pp. 1–9.
- Li T, Shahidehpour M. Price-based unit commitment: a case of Lagrangian relaxation versus mixed integer programming. *IEEE Trans Power Syst* 2005;20:2015–2025.
- Shahidehpour M, Li T, Choi J. Optimal generation asset arbitrage in electricity markets. *KIEE (Korea) Int Trans Power Eng* 2005;5A(4):311–321.
- Shahidehpour M, Yamin HY, Li Z. Market operations in electric power systems. New York: Wiley; 2002.
- Klemperer P, Meyer M. Supply function equilibria in oligopoly under uncertainty. *Econometrica* 1989;57(6):1243–1277.
- Green RJ, Newbery DM. Competition in the British electricity spot market. *J Pol Econ* 1992;100(5):929–53.
- Conejo AJ, Contreras J, Arroyo JM, *et al.* Optimal response of an oligopolistic generating company to a competitive pool-based electric power market. *IEEE Trans Power Syst* 2002;17(2):424–430.
- de la Torre S, Arroyo JM, Conejo AJ, *et al.* Price maker self-scheduling in a pool-based electricity market: a mixed integer LP approach. *IEEE Trans Power Syst* 2002;17(4):1037–1042.
- Baillo A, Ventosa M, Rivier M, *et al.* Optimal offering strategies for generation companies operating in electricity spot markets. *IEEE Trans Power Syst* 2004;19(2):745–753.
- Anderson EJ, Philpott AB. Using supply functions for offering generation into an electricity market. *Oper Res* 2002;50(3):477–489.
- Anderson EJ, Philpott AB. Optimal offer construction in electricity markets. *Math Oper Res* 2002;27(1):82–100.
- Pritchard G, Zakeri G, Philpott AB. Non-parametric estimation of market distribution functions in electricity pool markets. *Math Oper Res* 2006;3(3):621–636.
- Anderson EJ, Philpott AB. Estimation of electricity market distribution functions. *Ann Oper Res* 2003;121:21–32.
- Bazaraa M, Sherali H, Shetty CM. Nonlinear programming theory and algorithms. New York: Wiley; 1993.
- Luo ZQ, Pang JS, Ralph D. Mathematical programs with equilibrium constraints. Cambridge, UK: Cambridge University Press; 1996.
- Hobbs BF, Metzler CB, Pang JS. Strategic gaming analysis for electric power systems: an MPEC approach. *IEEE Trans Power Syst* 2000;15(2):638–645.

23. Fampa M, Barroso LA, Candal D, *et al.* Bilevel optimization applied to strategic pricing in competitive electricity markets. *Comput Optim Appl* 2008;39:121–142.
24. Pritchard G. Optimal offering in electric power networks. *Pac J Optim* 2007;3(3):425–438.
25. Gomez-Villalve E, Ramos A. Optimal energy management of an industrial consumer in liberalized markets. *IEEE Trans Power Syst* 2003;18(2):716–723.
26. Middelberg A, Zhang J, Xia X. An optimal control model for load shifting with application in energy management of a colliery. *Appl Energy* 2009;86:1266–1273.
27. Stoft S. *Power system economics: designing markets for electricity*. New York: Wiley-IEEE Press; 2002.
28. Neuhoff K, Barquin J, Boots M, *et al.* Network-constrained Cournot models of liberalized electricity markets: the devil is in the details. *Energy Econ* 2005;27(3):495–525.
29. Rudkevich A. On the supply function equilibrium and its application in electricity markets. *Decis Support Syst* 2005;40(3):409–425.
30. Sioshansi R, Oren S. How good are supply function equilibrium models: an empirical analysis of the ERCOT market. *J Regul Econ* 2007;31(1):1–35.
31. Zhou Z, Chan WK, Chow JH. Agent-based simulation of electricity markets: a survey of tools. *Artif Intell Rev* 2007;8:305–342.
32. Fabra N, von der Fehr N, Harbord D. Designing electricity auctions. *RAND J Econ* 2007;37(1):23–46.
33. Berry CA, Hobbs BF, Meroney WA, *et al.* Understanding how market power can arise in network competition: a game theoretic approach. *Utilities Policy* 1999;8(3):139–158.
34. Newbery D. Predicting market power in wholesale electricity markets. EUI RSCAS Working Paper No. 2009/03. 2009. Available at <http://hdl.handle.net/1814/10620>.
35. Rivier M, Ventosa M, Ramos A, *et al.* A generation operation planning model in deregulated electricity markets based on the complementarity problem. In: Ferris MC, Mangasarian OL, Pang J-S, editors. *Complementarity: applications, algorithms and extensions*. Norwell (MA): Kluwer Academic Publishers; 2001.
36. Hurlbut D, Rogas K, S Oren. Protecting the market from “Hockey Stick” pricing: how the public utility commission of texas is dealing with potential price gouging. *Electric J* 2004; 26–33.

OPTIMIZATION MODELS FOR CANCER TREATMENT PLANNING

WENHUA CAO
GINO J. LIM
Department of Industrial
Engineering, University
of Houston, Houston,
Texas

INTRODUCTION

The National Cancer Institute estimates that 1,479,350 new cancer cases are expected to be diagnosed in the United States in 2009 [1]. They also estimate that about 562,340 Americans are expected to die of cancer this year. Cancer is the second leading cause of death in the United States, exceeded only by heart disease. Types of cancer treatment are determined by the cancer site and staging. Most common cancer treatments are surgery, radiation therapy, and chemotherapy. Other types include immunotherapy, targeted therapy, and photodynamic therapy. Physicians often use a combination of these treatments to obtain the best results.

The ultimate goal of various cancer treatments is to kill all cancerous cells without damaging patient's normal tissue and function. However, due to the proximity of healthy organs to the tumor volume, it is almost impossible to completely spare healthy tissues during the treatments. As the oldest form of cancer treatment, surgery is often used to remove localized cancerous cells; that is, it can be an option for those patients whose tumor has not spread out to other areas of the body. Surgeries still face risks and side effects. Radiation therapy uses different types of radiation particles such as photons, electrons, and protons to sterilize the tumor or stop cancerous cells from proliferation. There are two kinds of radiation therapy: the external (beam) radiation therapy or teletherapy uses external

radiation beams for treatment, and the internal radiation therapy or brachytherapy places radioactive sources inside or close to the tumor. Both types of the radiation therapy methods have been increasingly utilized for treating many cancers such as brain, lung, breast, bladder, and prostate, to name a few. Effective plans of radiation treatments require delivery of high dose of radiation to the tumor region while minimizing dose on the normal cells as much as possible. On the other hand, chemotherapy or a drug therapy uses medicines (drugs) to eliminate cancer. Patients can take medicines alone or combine with other treatments.

Designing a cancer treatment plan often involves a series of optimization problems with different kinds of objectives and constraints. Researchers in the optimization community have made significant contributions in improving the quality of various treatment plans for cancer patients: surgery [2], radiation therapy [3–6], and chemotherapy [7,8].

This article presents planning problems of major cancer treatments, and mathematical and computational optimization models that have been developed to solve them. The rest of the article is organized as follows. The second section introduces external radiation therapy planning problems including gamma knife, conventional three-dimensional conformal radiation therapy (3DCRT), intensity modulated radiation therapy (IMRT), and proton therapy, and a few optimization methods that are developed specifically for such treatment modalities. The third and fourth sections discuss cancer treatment models for brachytherapy and chemotherapy. And, the final section summarizes the article with future research directions.

EXTERNAL RADIATION THERAPY

Treatment Delivery and Optimization

In external radiation therapy, radiation is delivered from outside the body and

directed at the patient's tumor location using special radiation delivery machines. Different devices produce different types of radiation and they include Cobalt-60 machines for gamma knife radiosurgery, linear accelerators for IMRT, neutron beam machines, orthovoltage X-ray machines, and proton beam machines. With the exception of the gamma knife radiosurgery, most treatments are administered over several weeks. Patients receive small dose of radiation each day until the total prescribed dose is delivered. Such a regimen is used to protect normal tissues from being damaged by radiation, since they can have time to recover. If the treatment volume is sufficiently irradiated, the tumor may well be eliminated. But, a significant amount of normal tissues would also be damaged. This is not a desirable outcome of a treatment.

Before a radiation therapy treatment planning takes place, physicians need to obtain precise geometric information of the tumor from a patient. Several scanning procedures, such as computed tomography (CT), magnetic resonance imaging (MRI), or positron emission tomography (PET), are mainly used. These diagnostic images then help physicians determine the three-dimensional shape, location and size of the tumor, and surrounding normal tissues. Conventionally, there are three types of volumes to be considered: *planning target volume* (PTV), *organs-at-risk* (OARs) or *critical structures*, and normal tissues. PTV includes the suspected cancerous area; OARs are organs located close to the PTV where receiving some minor radiation dose is acceptable; and all remaining portions are normal tissues. The physician designs the *prescription* for radiation treatment that delivers a sufficient amount of radiation dose on the PTV to kill cancer cells while limiting radiation on the OARs and the normal tissues. These are two conflicting objectives in radiation treatment planning. One may easily expect a trade-off between these two objectives. Thereby, upper and lower bounds of radiation doses are employed to guarantee a successful treatment plan according to radiation treatment planners' experiences. Optimization methods are often used to find

a clinically acceptable solution that is a compromise between the two objectives [9].

The quality of a treatment plan can be measured by *uniformity*, *conformity*, and *avoidance* [5]. It is important for a treatment plan to have uniform dose distributions on the target, so that *cold* and *hot spots* can be minimized. A *cold spot* is a portion of an organ that receives less than its required radiation dose level. On the other hand, a *hot spot* is a portion of an organ that receives more than the desired dose level. The uniformity requirement ensures that radiation delivered to the tumor volume will have a minimum number of *hot spots* and *cold spots* on the target. This requirement can be enforced using lower and upper bounds on the dose, or approximated using penalization. The conformity requirement is used to achieve the target dose control while minimizing the damage to OARs or healthy normal tissue. This is generally expressed as a ratio of cumulative dose on the target over total dose prescribed for the entire treatment. This ratio can be used to control conformity in optimization models. As we mentioned earlier, a great difficulty of producing radiation treatment plans is the proximity of the target to the OARs. An avoidance requirement can be used to limit the dose delivered to OARs. Finally, simplicity requirements state that a treatment plan should be as simple as possible. Simple treatment plans typically reduce the treatment time as well as implementation error. These above-mentioned requirements provide guidelines for most treatment planning optimization models.

In the rest of this section, we discuss some optimization models and techniques that are practically useful for radiation treatment modalities: gamma knife radiosurgery, 3DCRT, IMRT, and proton therapy.

Gamma Knife Radiosurgery

Background. The gamma knife is an important device for treating tumors or vascular malformations located in patient's head [10]. It can deliver a single beam of high dose of radiation by 201 Cobalt-60 unit sources. All 201 beams simultaneously intersect at the same location to form an

approximately spherical region that is typically termed as *shot of radiation*. Multiple shots are often used in a treatment using a gamma knife due to the irregularity and size of tumor shapes and the fact that the focusing helmets are available only in four sizes (4, 8, 14, and 18 mm). The objective of treatment planning is to deliver a high dose of radiation to the intracranial target volume with minimum damage to the surrounding normal tissue. The uniformity, conformity, and avoidance are necessary requirements to a treatment plan. Treatment optimization models are constructed to incorporate these requirements, which may be case specific of different doctors and patients. In addition, minimization of the number of shots is also preferred.

Optimization Models. The optimization problem of the gamma knife radiosurgery treatment planning aims to determine the locations and duration times for different shots. Corresponding integer variables are usually included in the optimization model. A mixed integer nonlinear programming (MINLP) has been introduced by Ferris *et al.* [3].

Suppose that the number of radiation shots for the treatment is given *a priori*. Adding the goal of minimizing this number to the optimization model is typically straightforward. However, solving such models can be extremely difficult.

Decision Variables. Let T denote the subset of tumor voxels and N denote the subset of voxels that are not in the tumor. Let $D_{i,j,k}$ denote the amount of radiation dose that a voxel (i,j,k) receives. In general, there are three types of decision variables:

1. a set of shot center locations (x_s, y_s, z_s) for $s \in S$;
2. a discrete set of collimator sizes w ($w \in W$);
3. radiation exposure time $t_{s,w}$ for each radiation shot.

It is assumed that the dose contribution $d_w(x_s, y_s, z_s, i, j, k)$ to voxel (i, j, k) from a shot centered at (x_s, y_s, z_s) with width w is given

in the model. Thus, the total dose delivered to the voxel (i, j, k) can be calculated by

$$D_{i,j,k} = \sum_{(s,w) \in S \times W} t_{s,w} d_w(x_s, y_s, z_s, i, j, k). \quad (1)$$

Constraints

1. Uniformity—Isodose Line Coverage.

A treatment plan is normally considered acceptable if a $\theta\%$ isodose curve encompasses the tumor region. For example, 50% isodose curve is a curve that encompasses all voxels that receive at least 50% of the maximum radiation dose that is delivered to any voxels in the target volume.

2. Choosing Shot Sizes.

The location of a shot center can be chosen by a continuous optimization process, while the width at each shot location can be determined by a discrete optimization process, for example, binary integer programming. Thus, the number of shots used can be counted by tracking the exposure time $t_{s,w}$; that is, a shot is used when $t_{s,w}$ is positive. Let n denote the total number of shot and width combinations, then it is valid that $n = \sum_{(s,w) \in S \times W} H(t_{s,w})$, where H is the step function.

An optimization model of Ferris *et al.* [3] is described as follows:

$$\begin{aligned} \min \quad & \sum_{(i,j,k) \in N} D_{i,j,k} \\ \text{s.t.} \quad & D_{i,j,k} = \sum_{(s,w) \in S \times W} t_{s,w} d_w(x_s, y_s, z_s, i, j, k) \\ & \theta \leq D_{i,j,k} \leq 1, \quad \forall (i,j,k) \in T \\ & n = \sum_{(s,w) \in S \times W} H(t_{s,w}) \quad t_{s,w} \geq 0. \end{aligned} \quad (2)$$

The most critical problem for solving the optimization model (2) is the large number of voxels that are needed when dealing with large irregular tumors (both within and outside of the target). This makes the optimization problem computationally intractable. One approach to overcome this problem is by removing a large number of the nontarget voxels from the model. Although this improves the computation time, the conformity of the dose to the target is

typically weakened. Ferris *et al.* [3] propose a sequential solution approach to speed up the time while maintaining conformity. First, a *coarse grid* problem is solved as a nonlinear programming (NLP) model using a set of systematically reduced data points. Then the *finer grid* NLP problem with full data points is solved using the starting point that was obtained by the coarse grid model in the previous stage. Typically, the solution from this finer grid model is very close to a good local optimum for the MINLP. Using this solution, the full MINLP model is finally solved to determine the values of the three sets of decision variables for this problem.

Three-Dimensional Conformal Radiation Therapy

Background. 3DCRT is able to utilize 3D imaging data of tumor and surrounding anatomy to achieve high conformal dose of radiation delivered to the target volume, and therefore reduce the dose to normal tissues. This treatment can be administered for a variety of cancers. In 3DCRT, multiple external beams are generated by a linear accelerator and directed from different angles into the target volume (see Fig. 1a). Most importantly, each beam can be shaped by a computer automated multileaf collimator (MLC), which is mounted on a gantry

to confirm the *beam's-eye-view* of the tumor (see Figure 1b). Hence, a three-dimensional geometry can be confirmed by these beams using anatomical information.

Wedge Filters. A wedge (also called a *wedge filter*) is a tapered metallic block with a thick side (the *heel*) and a thin edge (the *toe*) (see fig. 2). This metallic wedge varies the intensity of the radiation in a linear fashion from one side of the radiation field to the other. When the wedge is placed in front of the aperture, less radiation is transmitted through the heel of the wedge than through the toe. 2b shows an external 45° wedge, so named because it produces isodose lines that are oriented at approximately 45° . The quality of the dose distribution can be improved by incorporating a wedge filter into one or more of the treatment beams. Wedge filters are particularly useful in compensating for a curved patient surface, which is common in breast cancer treatments.

Two different wedge systems are used in clinical practice. In the first system, four different wedges with angles 15° , 30° , 45° , and 60° are available, and the therapist is responsible for selecting one of these wedges and inserting it with the correct orientation. In the second system, a single 60° wedge (the *universal wedge*) is permanently located on a motorized mount located within the head

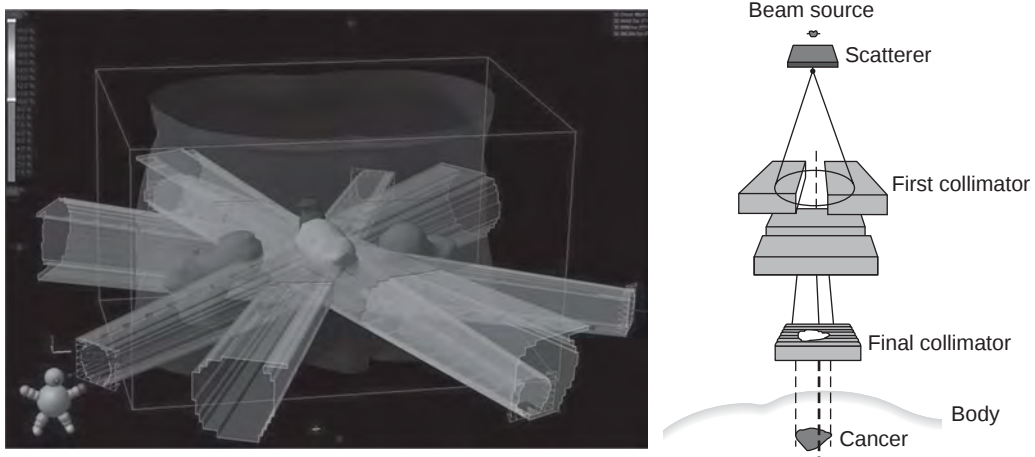


Figure 1. Multiple beams and beam's-eye-view: (a) Multiple external radiation beams. (b) A beam's-eye-view is a 2D shape of a tumor viewed by the beam source at a fixed angle.



Figure 2. Wedges: (a) a wedge filter and (b) an external wedge.

of the treatment unit. This wedge can be rotated to the desired orientation or removed altogether, as required by the treatment plan.

Optimization Models. Suppose that the data to the optimization models is given. Let $d_{(i,j,k),A}$ be the dose contribution to voxel (i,j,k) from a beam of weight 1 from angle A , S be a collection of voxels in the sensitive structure(s), and N be a collection of voxels in the normal tissues. When wedges are allowed in the optimization, the data will be provided as $d_{(i,j,k),A,F}$ that represents the dose contribution to voxel (i,j,k) from a beam of weight 1 from angle A , using wedge orientation F .

Beam Weight Optimization. The classical optimization problem in conformal radiation therapy is to choose the weights (or intensity levels) to be delivered from a given set of angles. Suppose, w_A represents the beam weight assigned to angle A , $D_{(i,j,k)}$ represents the total dose deposited to voxel (i,j,k) , and λ represents the relative weighting factors in the objective function. Given a set $\Omega = \{T \cup S \cup N\}$, a general optimization model that determines optimal radiation intensity is

$$\begin{aligned}
 \min_w \quad & \lambda_t f(D_T) + \lambda_s f(D_S) + \lambda_n f(D_N) \\
 \text{s.t.} \quad & D_{(i,j,k)} = \sum_{A \in \mathcal{A}} d_{(i,j,k),A} w_A, \\
 & \forall (i,j,k) \in \Omega, \\
 & l \leq D_T \leq u, \\
 & 0 \leq w_A, \forall A \in \mathcal{A}.
 \end{aligned} \tag{3}$$

Hard upper and lower bound constraints, that is, u and l , are imposed on the target

dose, so that, in the worst case, the resulting solution will satisfy the minimum requirement for a treatment plan. The penalty function $f(D)$ in the objective function can be defined based on the planner's preference, but a general function can be written as

$$f(D) = \|D(\cdot) - \theta\|_p, \quad p \in \{1, 2, \infty\}.$$

Note that, θ is the desired dose level for an organ of interest. These problems can be cast as a quadratic programming (QP) problem ($p = 2$), minimizing the Euclidean distance between the dose delivered to each voxel and the prescribed dose [11]. Furthermore, linear programming (LP) has also been extensively used to improve conventional treatment planning techniques [12]. The strength of LP is its ability to control *hot* and *cold* spots or integral dose on the organs using constraints, and the presence of many state-of-the-art LP solvers. The LP model replaces the Euclidean norm objective function of a QP with a convex one, for which standard reformulations [5] result in LP problems. While these techniques still suffer from large amounts of data in $d_{(i,j,k),A}$, they are typically solved in acceptable time frames. These models tend to find optimal solutions more quickly than the corresponding QP formulations.

Equivalent Uniform Dose (EUD). Some of the medical physics literature has been advocating the use of other forms of objective function in place of the ones outlined above. A popular alternative is that of generalized equivalent uniform dose (EUD) [13]. This is

defined on a per structure basis as

$$\text{EUD}_a(D, \Omega) := \left(\frac{1}{\text{card}(\Omega)} \sum_{(i,j,k) \in \Omega} D_{(i,j,k)}^a \right)^{1/a}.$$

Note that EUD is a scaled version of the a -norm of the dose to the particular structure, and hence is known to be a convex function for any $a \geq 1$ and concave for $a \leq 1$. Thus, the problem

$$\begin{aligned} \max_w \quad & \text{EUD}_a(D, T) \\ \text{s.t.} \quad & D_{(i,j,k)} = \sum_{A \in A} d_{(i,j,k),A} w_A, \quad \forall (i,j,k) \in \Omega, \\ & \text{EUD}_b(D, S) \leq \phi, \\ & \text{EUD}_c(D, N) \leq u, \\ & 0 \leq w_A, \quad \forall A \in A. \end{aligned}$$

is a convex optimization problem provided $a \leq 1$ and $b, c \geq 1$. As such, locally convergent NLP algorithms will find globally optimal solutions to these problems.

Beam Angle Selection and Wedge Orientation Optimization. Optimization also lends itself to solving the more complex problem of selecting which angles and wedge orientations to use, as well as their intensities. Mixed integer programming (MIP) is a straightforward technique for these types of problems. We describe an optimization model that simultaneously optimizes beam angles, wedge orientations, and beam intensities. Wedges are placed in front of the collimator to produce a gradient over the dose distribution and can be effective for reducing the radiation dose to organs-at-risk. This can be done by adding an extra dimension F to the variable w_A :

$$\begin{aligned} \min_{w, \psi} \quad & \lambda_t f(D_T) + \lambda_s f(D_S) + \lambda_n f(D_N) \\ \text{s.t.} \quad & D_\Omega = \sum_{A,F} d_{\Omega,A,F} w_{A,F}, \\ & w_{A,F} \leq M \cdot \psi_A, \\ & l \leq D_T \leq u, \\ & \sum_{A \in A} \psi_A \leq K, \\ & \psi_A \in \{0, 1\}, \quad \forall A \in A. \end{aligned} \tag{4}$$

The variable ψ_A is used to determine whether or not to use an angle A for delivery. The

choice of M plays a critical role in the speed of the optimization; further advice on its choice is given in Lim *et al.* [5].

Solution Techniques. Simulated annealing (SA) has been well accepted in the medical community [14]. But, the weakness of SA is its inability to verify the optimality. On the other hand, it is possible to find a global optimal solution for Equation (4). However, due to its slow convergence, using the MIP model has not been very useful for designing a treatment plan in the hospital. Recently, Lim *et al.* [5] proposed an iterative solution approach that solves the MIP problem fast (within 20 min for two clinical case examples presented). It is termed a *three-phase approach*.

Three-phase approach is a multiphase technique that “ramps up” to the solution of the full problem via a sequence of models. Essentially, the models are solved in increasing order of difficulty, with the solution of one model providing a good starting point for the next. The models differ from each other in the selection of voxels included in the formulation, and in the number of beam angles allowed.

If the most promising beam angles can be identified in advance, the full problem can be solved with a small number of discrete variables. One simple approach for eliminating unpromising beam angles is to remove those passing directly through any OAR [15]. A more elaborate approach [16] introduces a score function for each candidate angle, based on the ability of that angle to deliver a high dose to the PTV without exceeding the prescribed dose tolerance to OAR or to normal tissue located along its path. Only beam angles with the best scores are included in the model. Although there is no guarantee that this technique will produce the same solution as the original full-scale model (4), Lim *et al.* [5] have found that the quality of its approximate solution is close to optimal based on several numerical experiments.

Intensity Modulated Radiation Therapy

Background. *IMRT* is a high-precision radiotherapy that has been extensively used to treat a range of cancer types since the 1990s. As an advanced form of the



Figure 3. A multileaf collimator (MLC) for IMRT.

state-of-the-art 3DCRT and thanks to the continuous development of treatment techniques, IMRT is able to deliver more accurate radiation dose to precisely conform to the three-dimensional shape of the tumor while sparing the surrounding normal tissues. That is because the advanced MLC can move its leaves back and forth to block the portion of each beam's radiation dose (see Fig. 3). This capability allows the MLC to divide its *aperture* into hundreds of rectangular grids, and thus partitioning each beam into *sub-beams*, also called *beamlets* or *pencil beams*, from which the radiation is modulated. Therefore, a high degree of flexibility in delivering radiation dose distribution from each beam angle can be achieved by IMRT.

The IMRT treatment planning optimization framework typically consists of solving three sequential problems: beam angle optimization (BAO), fluence map optimization (FMO), and beam segmentation (BS). Assuming the geometry information of the treatment volume and prescribed dose to the target have been obtained. The (BAO) problem aims to find a subset A^* of a given set A of candidate beam angles ($A^* \subset A$) to deliver a satisfactory treatment. FMO aims to find an optimal beam shape intensity map, or known as a *fluence map* for each selected treatment angle. The fluence map specifies the intensity of radiation dose of each beamlet within a beam. Note that evaluating a BAO solution requires the knowledge of optimal fluence map for each selected angle;

hence, the BAO problem always combines FMO together. In BS, the interest is to decompose the fluence map into a minimum number of implementable MLC apertures and minimum corresponding intensities. These three problems have invited substantial research effort on optimization models to facilitate IMRT treatment planning in recent years.

Beam Angle Optimization and Fluence Map Optimization. With the patient lying on the treatment couch, the gantry of the linear accelerator can rotate 360° in a plane perpendicular to the couch. It stops at one angle to deliver certain amount of radiation. If we only allow the movement of the gantry during treatment and it is possible for the gantry to stop and deliver radiation at any angle, say every 1° , there are a set (A) of 360 beam angles available from which a BAO model aims to find an optimal subset ($A^* \subset A$) of angles that can yield best treatment. An FMO model determines the optimal fluence map for each selected angle.

As we discussed, the MLC discretizes a 2D beam into a matrix of grids. Let us define the beamlet intensity or weight as $W_{a,l,p}$, which is continuous and denotes the intensity of radiation delivered by the beam angle a and through the (l,p) th beamlet, $l = 1, 2, \dots, m$, $p = 1, 2, \dots, n$ (m is the number of MLC leaves, and n is the number of discrete units that a leaf can move). Also, the 3D treatment volume is discretized into thousands of voxels (i,j,k) . We term $d_{i,j,k,a,l,p}$ as the dose delivered by the beam angle a through the (l,p) th with a weight of 1 beamlet and received by voxels (i,j,k) . Thus, the total dose $D_{i,j,k}$ deposited in a voxel (i,j,k) is

$$D_{i,j,k} = \sum_{a,l,p} (\omega_{a,l,p} \cdot d_{i,j,k,a,l,p}). \quad (5)$$

Since the optimal set of beam angles are intimately related to the optimal fluence map for each angle, these two problems must be dealt with together in the problem formulation. Let a denote an angle, $a \in A$; then, formulating an optimization model for optimizing both beam angles and the fluence maps is a simple extension to Equation (4) that we discussed for the conventional 3DCRT, and the model can be obtained by

replacing the dose calculation constraint and big-M constraint in Equation (4) with constraints (6) and (7) as follows:

$$D_\Omega = \sum_{a,l,p} d_{\Omega,a,l,p} w_{a,l,p}, \quad \Omega = \{T \cup S \cup N\}, \quad (6)$$

$$w_{a,l,p} \leq M \cdot \psi_a. \quad (7)$$

See more details of this formulation and others in Lim *et al.* [17].

Obtaining global optimal solutions is difficult due to the computational complexity of the problem. Therefore, a good deal of heuristic methods has been discussed in the literature. Pugachev and Lei [18] introduced a scoring-based approach. Li *et al.* studied a particle swarm algorithm [19] and genetic algorithms [20]. Lim *et al.* [17] developed a *LP-based iterative beam angle elimination* (IBAELP) algorithm to obtain quality beam angles with optimized fluence maps. Aleman *et al.* [21] discussed a neighborhood search algorithm.

Optimal Beam Segmentation. An IMRT treatment plan can be completed by solving BAO and FMO problems. However, the deliver of the plan requires another optimization process, that is, solving the beam segmentation problem. In each beam angle, the MLC need to be configured to generate a series of apertures to deliver the given fluence map. Therefore, the BS problem aims to decompose the fluence map into a set of implementable MLC apertures and corresponding intensities. The number of delivering apertures associates with MLC setup times, while the intensity for each setup results in the length of beam-on time. The objective of the BS optimization consists of minimizing the two kinds of time simultaneously.

In a BS optimization model, the fluence map $\mathcal{W}$ is often represented from real matrix to integer matrix, that is,

$$\mathcal{W} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{pmatrix},$$

where $w_{l,p} \in \mathbb{Z}$, for $l = 1, \dots, m$, $p = 1, \dots, n$. Suppose that there are K binary matrices S^k and K integers μ_k . Suppose that K denotes the maximum number of different apertures needed to deliver the required fluence map, S^k represents the k th aperture profile, and μ_k represents the number of unit delivery time for the k th aperture. The primary requirement or constraint to the model is

$$\mathcal{W} = \sum_{k=1}^K \mu_k \cdot S^k, \quad (8)$$

where $S^k = [s_{l,p}^k]$, $s_{l,p}^k \in \{0, 1\}$, $l \in \{1, 2, \dots, m\}$, $p \in \{1, 2, \dots, n\}$, $\mu_k \in \mathbb{Z}$, $k \in \{1, 2, \dots, K\}$. Another important constraint required by the MLC movement is that the element in each row of S^k cannot change more than two times starting from one side to another side, also known as *consecutive one property*. For example,

$$0 \ 1 \ 1 \ 1 \ 0$$

is a feasible sequence. But,

$$0 \ 1 \ 0 \ 1 \ 0$$

is not allowed because the sequence of ones is not continuous.

There are two objectives of the BS model:

1. Minimize K , (setup time);
2. Minimize $\sum_{k=1}^K \mu_k$ (beam-on time).

The BS problem has been well studied so far. If the optimization model adopts the single objective of beam-on time, the problem is relatively easy and can be solved in polynomial time [22–24]. However, the problem is proven to be *strongly NP-hard* [23] when both objectives are included, that is, minimizing the total treatment time. Therefore, both researchers and planners use various heuristics. Engel [25] proposed an efficient algorithm that generates optimal beam-on time, but the optimality of setup time (K) is not guaranteed. Lim and Choi [26] proposed a two-stage IP approach that improves setup time with smaller values comparing Engel's approach, yet no proved optimality is present. Both methods claimed that solving problems

with a matrix size larger than 10×10 takes less than few seconds. Taskin *et al.* [27] developed a decomposition approach using MIP to obtain exact optimal solutions.

Proton Therapy

Introduction. Proton therapy is another form of radiation delivery method for cancer treatment. Thanks to continuous research and development of proton therapy for years; it is revealed that the radiation therapy with proton beams may perform a more precise dose delivery than widely used current practice of radiation treatment with photon beams in certain types of cancer. It is because the dose distribution from a monoenergetic beam of protons has an entrance region of slowly rising dose followed by a sharp increase, called *Bragg peak*, near the end of range. By superimposing beams of different energies, the Bragg peak can be spread to generate a dose distribution that provides moderate entrance dose, uniform high dose within the target tissue, and almost zero dose beyond the target (see Fig. 4). Highly conformal dose distributions with low integral dose can be achieved with a small number of proton treatment fields convergent on a target region.

IMPT Treatment Planning and Optimization. There are four modulation methods for radiation therapy formally defined by Lomax [28]:

two-dimensional (2D) modulation, distal edge tracking (DET), two-and-half-dimensional (2.5D) modulation, and three-dimensional (3D) modulation. The 2D (in the beam's-eye-view) modulation technique modulates the intensities or weights of spread-out Bragg peaks (SOBPs) with fixed extents. The DET technique modulates the intensity of single Bragg peaks with depths according to the distal edge of the target volume. The 2.5D modulation technique modulates the intensity of SOBPs with variable extents. The 3D modulation technique modulates the intensity of all Bragg peaks located within the target volume, which adds one additional dimension (depth) to the 2D modulation. The 3D modulation is performed by the pencil beam scanning (or called *spot scanning*) technique [29]. Their test results suggested that the 3D modulation may be superior in target coverage and sparing normal tissues than other methods. We consider 3D intensity modulated proton therapy (IMPT) treatment planning in the following discussion.

Unlike IMRT research, the study of IMPT treatment planning and associated optimization algorithms is in its beginning stage. There are various delivery techniques and identified system-specific optimization problems within IMPT treatment planning, or general proton therapy, are more clinically oriented. Therefore, most current research papers are primarily contributed

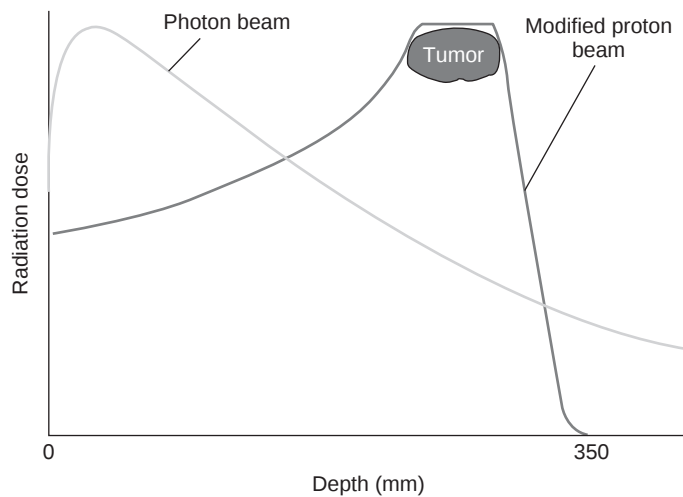


Figure 4. A comparison of the amount of radiation delivered with conventional photon beam radiation therapy versus proton therapy. Conventional therapy is distinguished by a relatively high entrance dose and an exit dose. By contrast, proton therapy has a much lower entrance dose and no exit dose.

by clinical researchers, while few are done by optimization experts. General IMPT treatment planning optimization problems are discussed in following paragraphs.

Beam Intensity Map Optimization. The beam (or spot) intensity map optimization problem is to determine the optimal intensity or weight of each beam spot within the three-dimensional target volume with the objective of achieving prescribed dose as close as possible. Two preliminary steps before performing the intensity map optimization includes (i) selecting the number and direction of treatment beam angles, and (ii) defining the spot position for each field. These two can be separate optimization problems.

The intensity map optimization problem for IMPT planning is essentially equivalent to conventional IMRT treatment planning, in which intensities of pencil beams need to be optimized for each treatment angle. Therefore, optimization methods used for IMRT intensity map can be adapted to IMPT use. The objective function (for minimization) can be the total deviation between prescribed dose and delivered dose, and dose constraints can be assigned to the model, for instance. The IMPT treatment planning system at Paul Scherrer Institute (PSI) uses an iterative least-square optimization approach [28]. Pflugfelder *et al.* [30] at German Cancer Research Center (DKFZ) conducted a study comparing with three optimization algorithms including steepest descent, conjugate gradient, and quasi-Newton methods.

Beam Direction Optimization. Currently, the selection of the number of beams and their directions, that is, the beam direction optimization problem, are manually performed by physicians based on their judgment, both for proton and photon beams. Unlike conventional IMRT with photons, the beam direction optimization algorithm has not been well documented for IMPT treatment planning. Some studies discussed the effect of the beam number and direction selection based in their case studies [28,31]. Since the beam direction optimization problem is typically a large-scale combinatorial problem, solution algorithms are often computationally expensive.

Beam Scanning Path Optimization. The beam scanning path optimization for IMPT treatment planning is to identify the optimal scanning path to deliver small proton pencil beams. Given the optimized spot intensity and predefined spot geometry in each treatment field, a scanning path that connects all scanning spots needs to be determined, so that the pencil beam can move continuously or discretely according to the path. The objective function to this problem can be formulated as the total length of the scanning path, which is meant to find minimal treatment time. Kang *et al.* [32] proposed a traveling salesman problem (TSP) formulation and a simulated annealing algorithm to solve this problem. Trofimov and Bortfeld [33] discussed the beam sequencing issues in a raster continuous scanning system.

Treatment Planning with Uncertainty. There are two main reasons to account for uncertainty in IMPT treatment planning. First, protons are physically very sensitive to inhomogeneous tissues when they pass through the patient's body. Range uncertainties may occur and result in misplacing Bragg peaks and deviating dose distribution from prescribed dose. Secondly, IMPT treatment plans, especially by the 3D modulation, are particularly sensitive to organ motions because the Bragg peaks require precision in millimeters. It is easy to result in delivering very high dose on normal tissues but very low dose on target. Therefore, treatment planning optimization with consideration of uncertainties is crucial for achieving better quality and robust IMPT treatment plans. A series of related studies have been reported and different uncertainties such as system errors, organ motion, patient positioning are discussed in Morávek *et al.* [34], Peterson *et al.* [35], and Pflugfelder *et al.* [36].

BRACHYTHERAPY

Background

Brachytherapy is an internal (invasive) radiation therapy modality that has been solely used or combined with external

radiation therapy to treat certain cancers such as prostate. In *brachytherapy*, radioactive substances are placed within or close to the tumor region in the form of wires, seeds, or rods. Types of brachytherapy are categorized depending on how the radioactive sources are placed inside the body, and include interstitial brachytherapy, intracavitary brachytherapy, intraluminal radiation therapy, and radioactively tagged molecules given intravenously. Interstitial brachytherapy is the focus of this section.

The problem of brachytherapy treatment planning (BTP) is primarily to determine the implantation location of radioactive sources. The goal of a brachytherapy plan consists of delivering enough radiation dose to the target tumor mass and sparing surround critical organs and normal tissues. Obtaining true optimal treatment plans for BTP is also extremely difficult in a clinical setting. Therefore, optimization models have been developed to find near optimal solutions within a clinically acceptable time [37,38].

Optimization Models

Modeling the BTP problem firstly requires the dose calculation scheme. Meyer *et al.* [38] described an approximation function for dose rate. Let us term the dose rate at a distance l from a single source as $D(l)$, referring to Meyer *et al.* [38] for calculations. Then, if j indexes the three-dimensional grid of prespecified possible locations of radioactive sources, where $j = 1, \dots, m$. For an arbitrary point i , the total dose at point i from all source locations is given by

$$D_i = \sum_{j=1}^m y_j D(l_{ij}), \quad (9)$$

where y_j is a binary vector indicating the selection or nonselection of location j , and l_{ij} is the Euclidean distance from point X_i to point X_j , that is, $l_{ij} = \|X_i - X_j\|$.

The constraints and objective of the BTP optimization model are homogeneously constructed as the RTP model. One can control the dose of radiation on PTV by hard

constraints:

$$D_i \leq T_U, \quad (10)$$

$$D_i \geq T_L, \quad (11)$$

where T_U and T_L are upper and lower bounds for dose deposition on PTV; or soft constraints:

$$D_i - R_U^T \leq T_U, \quad (12)$$

$$D_i + R_L^T \geq T_L, \quad (13)$$

where R_U^T and R_L^T are dose deviations from upper and lower bounds.

The specified maximum number of locations for seeds can be controlled by this constraint:

$$\sum_{j=1}^m y_j \leq \text{maximum number of locations.} \quad (14)$$

The objective functions may vary based on treatment planner's preferences. Lee *et al.* [39] described two kinds of objective functions. One minimizes the total dose deviation on PTV, and the other maximizes the total number of points that satisfying specified dose distribution requirements. Certainly, most objectives that are studied in external radiation therapy treatment optimization models can also be utilized to brachytherapy.

Solving the above BTP models in the clinical setting is still an ongoing research. Solution methods can be divided into two groups. One is an exact algorithm mainly of the branch-and-bound. Meyer *et al.* [38] tested various combination of branch-and-bound strategies of one commercial optimization solver. Cutting planes and problem preprocessing are also suggested to reduce the computational difficulty of BTP models. Another group of methods are heuristics such as SA and genetic algorithms [39].

In addition to the location problem, another optimization problem in BTP is to decide an appropriate amount of time (dwell time) for those sources staying in patient's

body [40]. In the dwell time optimization model, the dose calculation is defined as

$$D_i = \sum_{j=1}^m t_j D(l_{ij}), \quad (15)$$

where t_j is the dwell time. The constraints including (14)–(17) and objective functions for optimizing locations of radioactive sources can be used to optimize dwell time. Alterovitz *et al.* [40] introduced an LP formulation to the problem. Its solutions were compared with ones from the SA approach implemented in the clinical setting. Detailed discussions of their model can be found in Alterovitz *et al.* [40].

CHEMOTHERAPY

Chemotherapy or drug therapy uses drugs to control and even cure the cancer by minimizing tumor cells in certain treatment periods [41]. Although chemotherapy has been extensively used to treat cancer, it is well-known for its side effects; it brings not only efficacy to the tumor control, but also toxicity to normal tissues. Hence, the chemotherapy plan is often designed with two objectives: minimizing the tumor mass and minimizing the total drug dose [42], and to determine the schedule of drug treatment cycles and the amount of drug dose in specific cycles. Since there are very complex biological processes in patient's body during the drug treatment, it is difficult to achieve an optimal chemotherapy plan.

Many optimization approaches have been developed to improve the efficacy of chemotherapy [7, 41–43]. We take a typical chemotherapy planning optimization model given by Agur *et al.* [7] for illustration in three components as follows:

First, a simulation algorithm is used to predict body's response to chemotherapy. Two types of human cells are considered: *host cell* and *target cell* (tumor cell). The aim of the treatment is to reduce the target cells and maintain a certain level of host cells in patient's body. The growth of both cells can be interfered by chemotherapy. When the cells are at their sensitive life cycle and exposed to chemotherapy, they could

probably die. With the help of well-done biological studies, the cell behavior during chemotherapy is modeled by the simulation algorithm.

Secondly, the objective function named as fitness function is developed mainly based on the level of host and target cells that reflects the treatment plan performance. The primary variable of the fitness function is the chemotherapy treatment schedule over a series of n time intervals, which are represented by a vector of binary numbers. Let $\mathcal{F}(\mathbf{s})$ denote the fitness of a treatment schedule $\mathbf{s}$. It can be briefly constructed as follows:

$$\begin{aligned} \mathcal{F}(\mathbf{s}) = & c_1 x_h(\mathbf{T}) + c_2 x_a(\vec{T}) + c_3 I_{\text{alive}} + c_4 I_{\text{cure}} \\ & + c_4 t_{\text{death}} + c_5 t_{\text{cure}}, \end{aligned} \quad (16)$$

where $x_h(\mathbf{T})$ and $x_a(\mathbf{T})$ denote the levels of host and target (abnormal) cells at the end of time period $\mathbf{T}$, I_{alive} and I_{cure} denote the indicators of patient's condition being alive or cured at the end of treatment, t_{death} and t_{cure} denote the time of death or cure eventually, and c_i for $i = 1, \dots, 5$ denotes the weighting factor.

Suppose that one means the treatment is given and zero means the treatment is not given in a time interval. For example, the vector $s = (1, 0, 1, 1, 0, 0, 1, 1)$ indicates an eight-hour treatment schedule if the time unit is an hour, where the patient receives the treatment in the first hour, and stops taking it in the following hour, then continues the treatment for consecutive two hours, stops two hours, and continues another two-hour at the end. After a treatment schedule is fulfilled, the fitness function is evaluated to assess the condition of the patient, that is, cured, survival, or death.

Thirdly, the local search heuristics can be used to solve the problem. Overall, the aim is to find a good solution within computational capability.

CONCLUSIONS

Soon, cancer is projected to be the leading cause of death in the United States. Optimization techniques have been playing a key

role in facilitating cancer treatment planning and delivery. Well-designed optimization models can not only improve the treatment accuracy but also help save health-care cost by more efficient treatment delivery. We have described various cancer treatment optimization models and their solution methods in this article. The aim was to give a brief overview about recent advances in cancer treatment models. This is an area of research that requires multiple disciplinary collaborative efforts among engineers, mathematicians, computer scientists, physicians, physicists, and more. We truly hope that cancer becomes history in our generation.

REFERENCES

1. American Cancer Society. Cancer facts and figures 2009. Available at <http://www.cancer.org>. Accessed on 2009 Nov 1.
2. Malinen M, Huttunen T, Kaipio JP. Thermal dose optimization method for ultrasound surgery. *Phys Med Biol* 2003;48:745–762.
3. Ferris MC, Lim J-H, Shepard DM. Optimization approaches for treatment planning on a Gamma Knife. *SIAM J Optim* 2003;13:921–937.
4. Lee EK, Fox T, Crocker I. Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy. *Int J Radiat Oncol Biol Phys* 2006;64(1):301–320.
5. Lim GJ, Ferris MC, Wright SJ, *et al.* An optimization framework for conformal radiation treatment planning. *INFORMS J Comput* 2007;19(3):366–380.
6. Romeijn HE, Ahuja RK, Dempsey JF, *et al.* A new linear programming approach to radiation therapy treatment planning problems. *Oper Res* 2006;54(2):201–216.
7. Agur Z, Hassin R, Levy S. Optimizing chemotherapy scheduling using local search heuristics. *Oper Res* 2006;54(5):829–846.
8. De Pillis LG, Radanskaya A. A mathematical tumor model with immune resistance and drug therapy: an optimal control approach. *J Theor Med* 2000;3:79–100.
9. Holder A. Radiotherapy treatment design and linear programming. In: Brandeau ML, Saintfort F, Pierskalla WP, editors. *Operations research and health care: a handbook of methods and applications*. Boston (MA): Kluwer Academic Publishers; 2004. pp. 741–774.
10. Ganz JC. Gamma knife surgery. Wien: Springer; 1997.
11. Chen Y, Michalski D, Houser C, *et al.* A deterministic iterative least-squares algorithm for beam weight optimization in conformal radiotherapy. *Phys Med Biol* 2002;47:1647–1658.
12. Shepard DM, Ferris MC, Olivera G, *et al.* Optimizing the delivery of radiation to cancer patients. *SIAM Rev* 1999;41:721–744.
13. Choi B, Deasy JO. The generalized equivalent uniform dose function as a basis for intensity-modulated treatment planning. *Phys Med Biol* 2002;47:3579–3589.
14. Morrill SM, Lam KS, Lane RG, *et al.* Very fast simulated annealing in radiation therapy treatment plan optimization. *Int J Radiat Oncol Biol Phys* 1995;31:179–188.
15. Rowbottom CG, Khoo VS, Webb S. Simultaneous optimization of beam orientations and beam weights in conformal radiotherapy. *Med Phys* 2001;28(8):1696–1702.
16. Pugachev A, Lei X. Incorporating prior knowledge into beam orientation optimization in IMRT. *Int J Radiat Oncol Biol Phys* 2002;54:1565–1574.
17. Lim GJ, Choi J, Mohan R. Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning. *OR Spectr* 2008;30(2):289–309.
18. Pugachev A, Xing L. Pseudo beam's-eye-view as applied to beam orientation selection in intensity-modulated radiation therapy. *Int J Radiat Oncol Biol Phys* 2001;51(5):1361–1370.
19. Li Y, Yao J, Yao D, *et al.* A particle swarm optimization algorithm for beam angle selection in intensity-modulated radiotherapy planning. *Phys Med Biol* 2005;50:3491–3514.
20. Li Y, Yao J, Yao D. Automatic beam angle selection in IMRT planning genetic algorithm. *Phys Med Biol* 2004;49:1915–1932.
21. Aleman DM, Kumar A, Ahuja RK, *et al.* Neighborhood search approaches to beam orientation optimization in intensity modulated radiation therapy treatment planning. *J Glob Optim* 2008;42:587–607.
22. Ahuja RK, Hamacher HW. A network flow algorithm to minimize beam-on time for unconstrained multileaf collimator problems in cancer radiation therapy. *Networks* 2005;45:36–41.
23. Baatar D, Hamacher HW, Ehrgott M, *et al.* Decomposition of integer matrices and

- multileaf collimator sequencing. *Discrete Appl Math* 2005;152:6–34.
24. Boland N, Hamacher HW, Lenzen F. Minimizing beam-on time in cancer radiation treatment using multileaf collimators. *Networks* 2004;43(4):226–240.
 25. Engel K. A new algorithm for optimal multileaf collimator field segmentation. *Discrete Appl Math* 2005;152:35–51.
 26. Lim GJ, Choi J. A two-stage integer programming approach for optimizing leaf sequence in IMRT. IE Technical Report, IE0707-01. Houston (TX): University of Houston; 2007.
 27. Taşkın ZC, Smith JC, Romeijn HE, *et al.* Optimal multileaf collimator leaf sequencing in IMRT treatment planning. *Oper Res* 2010;58(3):674–690.
 28. Lomax A. Intensity modulation methods for proton radiotherapy. *Phys Med Biol* 1999;44:185–205.
 29. Smith AR. Present status and future developments in proton therapy. *Laser-Drive Relativistic Plasmas Applied for Science, Industry, and Medicine: 2nd International Symposium. AIP Conference Proceedings, Volume 1153.* Kyoto, Japan; 2009. pp. 426–437.
 30. Pflugfelder D, Wilkens JJ, Nill S, *et al.* A comparison of three optimization algorithms for intensity modulated radiation therapy. *J Med Phys* 2008;18(2):111–119.
 31. Li HS, Romeijn HE, Fox C, *et al.* A computational implementation and comparison of several intensity modulated proton therapy treatment planning algorithms. *Med Phys* 2008;35(3):1103–1112.
 32. Kang JH, Wilkens JJ, Oelfke U. Demonstration of scan path optimization in proton therapy. *Med Phys* 2007;34(9):3457–3464.
 33. Trofimov A, Bortfeld T. Beam delivery sequencing for intensity modulated proton therapy. *Phys Med Biol* 2003;48:1321–1331.
 34. Morávek Z, Rickhey M, Hartmann M, *et al.* Uncertainty reduction in intensity modulated proton therapy by inverse Monte Carlo treatment planning. *Phys Med Biol* 2009;54(15):4803–4819.
 35. Peterson S, Polf J, Ciangaru G, *et al.* Variations in proton scanned beam dose delivery due to uncertainties in magnetic beam steering. *Med Phys* 2009;36(8):3693–3702.
 36. Pflugfelder D, Wilkens JJ, Oelfke U. Worst case optimization: a method to account for uncertainties in the optimization of intensity modulated proton therapy. *Phys Med Biol* 2008;53(6):1689–1700.
 37. Lee EK, Zaider M. Mixed integer programming approaches to treatment planning for brachytherapy—application to permanent prostate implants. *Ann Oper Res* 2003;119:147–163.
 38. Meyer RR, D’Souza WD, Ferris MC, *et al.* Mip models and bb strategies in brachytherapy treatment optimization. *J Glob Optim* 2003;25(1):23–42.
 39. Lee EK, Gallagher RJ, Silvern D, *et al.* Treatment planning for brachytherapy: an integer programming model, two computational approaches and experiments with permanent prostate implant planning. *Phys Med Biol* 1999;44:145–165.
 40. Alterovitz R, Lessard E, Pouliot J, *et al.* Optimization of hdr brachytherapy dose distributions using linear programming with penalty costs. *Med Phys* 2006;33(11):4012–4019.
 41. Tse SM, Liang Y, Leung KS, *et al.* A memetic algorithm for multiple-drug cancer chemotherapy schedule optimization. *IEEE Trans Syst Man Cybern* 2007;37(1):84–91.
 42. Panetta JC, Fister KR. Optimal control applied to competing chemotherapeutic cell-kill strategies. *SIAM J Appl Math* 2003;63(6):1954–1971.
 43. De Pillis LG, Gu W, Fister KR, *et al.* Chemotherapy for tumors: an analysis of the dynamics and a study of quadratic and linear optimal controls. *Math Biosci* 2007;209(1):292–315.

OPTIMIZATION OF PUBLIC TRANSPORTATION SYSTEMS

JUAN CARLOS MUÑOZ
RICARDO GIESEN
Department of Transport
Engineering and Logistics,
Pontificia Universidad Católica
de Chile, Santiago, Chile

INTRODUCTION

Public transportation systems pose a highly complex set of challenges for the analyst that stem fundamentally from their sheer size. The task of scheduling a transit network involves determining the operations of thousands of vehicles and drivers making daily trips, often numbering in the millions. A prime example is the Interligado system in Sao Paulo, Brazil, which handles 9 million trips per day on approximately 14,000 vehicles. In the United States, more than 20 cities operate fleets of over 1000 buses [1]. The design, planning, and operation of a transit system must also address an array of other considerations including the structure of the network, its coexistence with other existing transportation systems, the congestion created by its use (which in turn effects service levels), and the autonomy on travel decisions of users, the system's lifeblood.

In light of these considerations, transit systems are typically analyzed from the opposing perspectives of demand and supply.

Demand

The demand for transit trips arises from the spatially distributed needs of individuals in accessing destinations at times that may be specific (e.g., travel to work) or flexible (e.g., shopping trips) and for which there is a certain maximum willingness to pay. This demand typically has a spatial and temporal structure that exhibits time-of-day variations that are clearly marked off into peak and off-peak periods. How demand translates into

actual trips depends on the levels of service supplied by the transit system (and alternative transportation modes, if any) and how they are perceived by users. If service levels are deficient, some potential trips shift to alternative modes or may not even be made at all. Any trip demand study thus requires an understanding of how demand is distributed across the service area, how it breaks down across different user socioeconomic strata, what transportation modes are available for each user (taking account of motorization rates), what are the purposes of the trips, and what attributes of the transit and alternative system services impact users' decisions (e.g., walking time, wait time, in-vehicle time, number of transfers, fare, comfort, reliability, safety, security, level of information, and seat availability).

Supply

Transit supply is expressed by the availability of *services* or *routes* to those willing to travel. Each service has a specific starting and ending place and follows a sequence of stops. Also, each service has a load limit (passenger capacity) and is therefore subject to congestion, implying that service levels begin to decline as usage increases. Congestion on public transit systems is manifested in terms of wait time increases due to fully loaded vehicles passing up waiting riders, longer boarding and alighting times at stops or stations, deterioration of comfort level aboard vehicles, and so on.

Most urban transit systems coexist with other transportation modes. The road infrastructure required by surface transit modes is often shared with flows of cars, motorcycles, bicycles, and trucks, complicating the design of transit networks and impacting its performance.

The sequence of decisions constituting the transit operations planning process for the supply of service is depicted by the flowchart in Fig. 1, adapted from Ceder and Wilson [2] and Ceder [3]. In practice, the boundaries between the various stages are somewhat

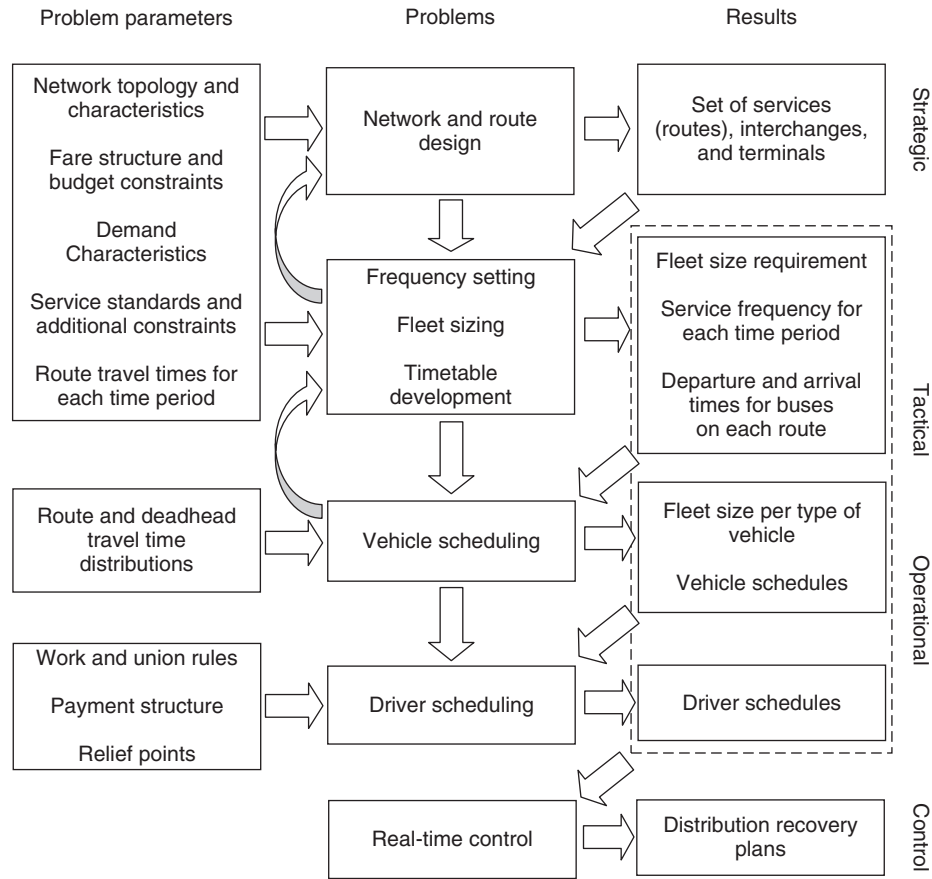


Figure 1. Transit operation planning process.

blurred and the process is enriched by feedback on previous decisions. Feedback is also employed in analyzing the impacts on subsequent stages of modifications to earlier ones.

According to this scheme, the design and analysis of a public transit system can be structured into four decision levels as follows.

Strategic decisions are those which are difficult to modify later, at least for relatively lengthy periods. Examples of such decisions are the transit mode (e.g., BRT versus LRT versus Metro); the type of vehicle to acquire (capacity); the structure of services to be offered; and, at a lower level, the location of stations and stops along the routes.

Tactical decisions are those that hold over the medium term and are revised anywhere from monthly to half-yearly. They include adjusting service frequencies at each period,

modifying fares, and offering operational variations on services such as deadheading, expressing, and short turning. In a two-way service, when demand for service is significantly imbalanced, a *deadhead* operation could prove beneficial. In such a case, transit vehicles would operate normally, loading passengers, on the high-demand direction and return empty (and as fast as possible, maybe using a different path) on the opposite direction. Such a scheme increases the supply where it is most needed while decreasing it in the low-demand direction, allowing for a balanced occupation rate of the vehicles in both directions. An *express* service is a service that skips some predefined stops, offering a fast trip to its passengers. A *short-turning* service associated with a certain route consists of providing a shorter service that operates only

in a section of the route with the purpose of increasing supply in its busiest section.

Operational decisions, such as drivers' shifts, and route and vehicle assignments, typically occur in the short term (daily basis).

Finally, control decisions define the system's recovery response to the inevitable disruptions that upset service schedules such as absent drivers, vehicle breakdowns, services running late, vehicle bunching, and so on.

In this article we discuss how operations research provides answers to the problems confronting each type of decision. The text is structured around the four decision-making levels: strategic or long term in the section titled "Strategic Decisions," tactical or medium term in the section titled "Tactical Decisions," operational or short term in the section titled "Operational Decisions," and disruption response or real-time control in the section titled "Real-Time Control." A deeper perspective of all these issues can be found in Vuchic [4].

STRATEGIC DECISIONS

This decision level deals with questions relating to the formal design of a new and complete transit system (a "greenfield" project) or the modification of an existing one. An example of a city designed and implemented from the ground up is Brasilia, created in the mid-twentieth century, but it was initially mostly designed for cars. Other cities have faced the challenge of drastically changing a transit system already in place [5]. In either situation, strategic decision making needs to address the following questions:

1. The market structure that will govern the system. How many operators will participate in the system? Will it be based on private operators? Do you expect operators to compete for the passengers? What role will the government have? Different cities have tried different approaches, from full deregulation and fully privately operated to a centralized system run by the government. In between, there are centrally operated systems with private operators such as Transmilenio in Bogota. In addition, these decisions are influenced by the institutional framework chosen by the government, for example, different agencies in charge of different modes (e.g., buses and metro) or a single agency in charge of the complete system, such as transport for London. The answers to these questions probably have significant impacts on the systems costs and the quality of service faced by the users.
2. The transportation modes that will configure the system (metro, light rail, BRT, and regular buses). For surface modes, it is important to define the level of interaction that the system will have with other modes, particularly private cars. The more exclusive is the transit right-of-way and the smoother is the boarding and alighting process, the higher will be the vehicle speeds, generating significant benefits in terms of level of service to users. Also essential is the definition of the type of energy to be used by the vehicles, since once decided it can be very costly to change. Note that per passenger transported, the transit system is less polluting than cars; so it is intrinsically associated with a more sustainable motorized urban mobility. Still, in many cities the transit system as a whole is a major actor in the generation of pollution (atmospheric and acoustic).
3. The type of route structure of the transit network. This decision is intimately linked to the level of fare and operational integration between the routes. If there is no fare integration, many of the routes will tend to be designed to cross the entire area served so that most riders can reach their destinations paying a single fare. Such routes tend to be inefficient given that along most of their length, vehicle load factors are relatively low. This oversupply of underutilized vehicle kilometers can be reduced by adopting a system design that ensures integration between routes. Typically, the design is a trunk-and-feeder system in which trunk services cover large distances

at high speeds and high vehicle load factors while feeder services shuttle to and from the trunk lines, picking up riders at their origin points and delivering them at their final destinations. This setup significantly reduces fleet and operating costs but forces users to transfer. It also requires a road infrastructure and right-of-way that enables high operating speeds along transit corridors with stops or stations that can efficiently handle transfers on a large scale. Varying levels of route integration are observed in transit systems around the world, including total competition among different vehicles on a single route, competition among different routes, competition among different modes, and total fare and operational integration.

4. The specific routes to be supplied. It is necessary to determine for each route its actual path or course, the type of vehicle, and the service frequency. At this decision level the analysis is typically confined to the highest demand periods of the day.

In responding to these challenges, planners attempt to identify a configuration that maximizes social welfare, this being defined as the total benefits of trips undertaken less the social cost of the system in terms of resources allocated by society as a whole to enable transit users to reach their chosen destinations. Of course, quite often this maximization procedure is restricted to satisfy some kind of budget constraint. Typically, this process involves building a model that represents overall generalized costs for a fixed and known level of transit trip demand specified by a trip matrix of travel between different zones of the area served. The following types of costs should be included:

1. *Costs That Reflect Impacts on Users.* Among the direct impacts of a trip are walking time (i.e., stop or station access time), waiting time at each stage of the trip (including added time when waiting users are passed up by fully

loaded vehicles), the disutility of transfers, and in-vehicle travel time. Vehicle load factors should also be considered for their effects on passenger comfort. In the case of high-frequency services that are not operating to a set schedule, waiting time is assumed to be inversely proportional to frequency. The fare is not included as a social cost since it is a mere transfer from one element of the system (the user) to another (the vehicle operator) rather than the consumption of a resource.

To attach cost figures to these impacts, the value of each of them—access time, waiting time, and penalties for transfers and overcrowding—must be determined. These values greatly vary from user to user if they are estimated on the basis of wage rates, but the use of a common value for all riders is recommended so that the model is not biased toward those with greater purchasing power (who highly value time). It is up to the transit authority to define the value of time, but it should be noted that the higher is this value, the denser will be the network design and the more frequent will be the services. It is therefore an extremely important decision that is typically grounded on how significant a role the transit system is expected to play in the city.

2. *Infrastructure Costs.* Different transit technologies and services require different levels of capital investment. Putting a value on these costs necessarily means defining the number of stops or stations, the level of segregation of the bus lanes, and the expenditures involved in building and equipping the corridor. The cost of installing any vehicle control equipment or traffic signal priority systems must also be included. Each of these decisions directly impacts on the operating speed of services.
3. *Operation Costs.* These typically include the cost of fleet acquisition and a unit operating cost per run for each route (for a given type of vehicle). By a *run*, we understand the trip made by

a single vehicle from the beginning to the end of a route.

4. *Costs Imposed by Externalities.* This includes impacts on travel times of other transportation modes (cars, bicycles, and walking), impacts on accident rates, atmospheric and noise pollution, and so on.

The formulation should model demand as a set of trips where each user chooses the best alternative available. In the absence of congestion or vehicle load (capacity) limits, such a set can be identified simply by minimizing the social costs of the system.

If congestion in the system is expected to be significant, however, with vehicle load levels close to maximum capacity, the model will have to accommodate this added complexity. A design that omits this factor and merely minimizes total costs would wrongly assume that riders spontaneously choose the alternative that is best for the system and not just for them. Consider, for example, the simple case illustrated in Fig. 2 of a transit corridor with two lines in which their frequencies need to be determined. The corridor has three nodes and two origin–destination pairs (A–B, A–C).

Let us assume that line 1 is very fast along its entire length but has high operating costs while line 2 is slow but with low costs. Both lines charge the same fare. Clearly, the frequency on line 1 cannot be zero since it is the only route supplying trips from A to C and, since it is offered, A–B passengers will also want to take it. But if a capacity constraint for line 1 were active, it could be cheaper for the system to have A–B passengers to travel on the slower line than to increase frequencies on the faster one, in which case the model would indicate that they wait for Line 2, something they would not do of their own accord.

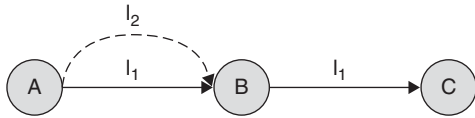


Figure 2. Corridor with two O–D pairs served by two lines.

Thus, if capacity constraints are included, the design must analyze how users decide their route when there is more than one alternative, each with its own combination of times, transfers, and fares. Models of user behavior need to take into account that riders normally place a higher value on waiting than walking and more value on walking than in-vehicle traveling [6]. They must also consider that these values greatly vary between users of different income levels.

Incorporating all of these different elements into a model is no easy task, and often some are left out of preliminary design versions. Nevertheless, it is important that the model accurately represents the transit system so that it can generate detailed scenarios for comparing alternatives and supporting decision making.

Certain other indicators are useful in providing general orientation for the various decisions confronting transit planners. For example, it can be shown that the optimal service frequency (f^*) of a route with no capacity or budget restraints is given by the following formula [7,8]:

$$f^* = \sqrt{\frac{k c_w q}{c_o}}, \quad (1)$$

where c_w is the value of wait time, q is the arrival rate of passengers to the system, c_o is the operating cost per run, and k is a headway constant defined so that $k = 1/2$ if headways are perfectly regular and grows as they become increasingly irregular. This formula can be used to deduce the effect of an abrupt decline in operating costs or a rise in demand on the optimal route frequency. Note, however, that in developing countries the service frequencies given by Equation (1) may be insufficient to meet demand. This is because the value of operating costs tends to be much higher than of wait time, implying $c_o \gg c_w$. In such cases, the frequencies will have to be adjusted to ensure that demand on the most heavily traveled segment of each route is satisfied.

The lower bound for the number of transit vehicles required during peak periods, n ,

can also be estimated from the service frequency if we recall that $n = \frac{t_c}{h}$, where t_c is the cycle time, h is the headway, and $h = \frac{1}{f}$. The *cycle time* on a two-way service is the amount of time needed by a vehicle to do both services sequentially, returning to its original position.

This type of analysis can further be utilized to determine the flows over a transit corridor that would justify the adoption of different transit solutions. Replacing a regular bus route with a BRT or a Metro line, for example, is an important question that many decision makers must face given the large investment involved (the per-kilometer cost of a metro line is at least US\$60 million and can reach as high as several hundred million) and many actors attempt to influence these decisions.

Often planners attempt to formulate the problem of deciding which technology should be considered for different demand levels. These simplified results must be applied with caution, however, as travel patterns of similar size but different structure (in average trip distance, for example) may or may not justify a decision such as that mentioned in the previous paragraph. In general, however, decisions should not be made on the basis of a subset of a network without considering its global structure, particularly in the presence of heavy congestion.

It should also be recalled that in reality trip demand is far from being fixed and known as we assumed above. Forecasting demand is a complex process that depends on many factors related to the characteristics of the various transportation modes and of the users themselves. Moreover, any changes to a transit system that impact perceived service levels may affect the number of trips taken not only by current users but by potential ones.

Finally, decisions at this level must determine the system's fare structure and method of payment. The fare structure influences users' route choices within the network. The higher is the charge for transfers, for example, the greater will be the proportion of users opting for direct routes even if other itineraries are faster.

TACTICAL DECISIONS

Once the routes to be served and the size of the system's fleet have been decided at the strategic level, the planning process turns to the tactical decisions. The issues addressed at this level are as follows:

1. The definition of regular operating periods. Travel pattern data are used to determine sequences of hours during which trip volumes and their space and time characteristics are relatively homogeneous.
2. As already noted, service frequencies for peak periods are normally defined at the strategic level and serve as the basis for identifying the size of the vehicle fleet required by the system. The frequencies for all other periods are defined at the tactical level. For this purpose, it is assumed that the demand for trips is known and that the frequencies can be obtained by Equation (1). More sophisticated methods for determining frequencies that incorporate variable and/or elastic demand assumptions, alternative concepts of waiting time, and explicit budget restraints are reviewed in Ceder [3]. The ongoing adjustment of frequencies at this decision level should keep them in step with observed demand levels in the system, and the expected demand pattern changes because of seasonal factors such as holiday periods or special events that attract large crowds to specific locations.
3. The definition of schedules based on the frequencies for each period. This involves determining the departure times from each terminal for each route. As a general rule, if demand is relatively uniform over a given period the departures are scheduled to maintain equal headways. If demand varies significantly, on the other hand, departure times can be adjusted so that passenger load levels are equalized over the most heavily used segment of a route. The problem is considerably more complex if it is desired

to synchronize transfers between different routes at transfer stations. Recent work on this issue is reviewed in Desaulniers and Hickman [9].

Given expected between-stop travel times and dwell times across the network for each period, the visit times of each run at each stop and arrival times at route terminals can be predicted. This result is central to vehicle scheduling as it determines the time when a vehicle becomes newly available for service after finishing a run.

For low-frequency routes (less than six vehicles per hour), schedules are normally published in various media so that users can plan their stop arrival times to minimize waiting. With high-frequency routes, schedules are rarely published, as the inevitable variations in actual service relative to the scheduled service intervals are relatively large, rendering detailed timetables of little practical use to riders. They may still be used by drivers as a general guide for maintaining regular intervals between successive runs.

4. Definition of special services for high-demand periods or routes with high spatial concentration. Implementation of such services, for example, short turning, deadheading, and express services may be justified by certain demand structures. A number of these strategies are discussed in Furth and Day [10].

One of the advantages of these services is that they constitute an alternative to putting on additional vehicles and thus reduce the required fleet size for serving a given demand, which in turn implies lower system costs. In operational terms, these strategies increase frequencies along high-load route segments to reduce wait times. Express services in particular are used to reduce the in-vehicle travel times of some trips.

5. Establishment of fares for each period. Fare structures can be defined to promote or discourage ridership at different times of the day. For example, since the number of required vehicles is

necessarily determined by the highest demand period, a means of reducing fleet size is to raise fares for peak hours, which tends to flatten the peak as some users shift their travel plans to lower fare off-peak periods. Fare variations can also be applied to different directions of the same route. Charging a higher rate on the high-demand direction will encourage greater use of services at points where there is more available capacity. Also, fares should depend on the distance traveled to represent in some sense the cost imposed on the system. If the exact distance cannot be measured, charging according to a zonal structure may provide the second-best alternative.

OPERATIONAL DECISIONS

Given the route schedules and available resources in terms of vehicles, drivers, garages, and possible driver relief points, the scheduling of vehicles and drivers should minimize operating costs subject to constraints imposed by garage capacity, employment standards and agreements, and other relevant factors peculiar to the area served by the system. Each run in the schedule requires a vehicle and a driver in the right place at the right time. Thus, the (often sequential) decisions to be made at this level are

1. *Vehicle Scheduling.* The arrival and departure times of each run at each terminal are drawn from the route schedules discussed in the previous section. These are used to define the number of vehicles of each type to be assigned to each terminal and the set of runs (called a *block*) to be assigned to each vehicle such that all runs can be carried out as scheduled. In formal terms, the fleet of vehicles $F_{v,k}$ of type v assigned to terminal k must satisfy the following relationship [11]:

$$F_{v,k} \geq \max_t \{D_{v,k}(t) - A_{v,k}(t)\}, \quad (2)$$

where $D_{v,k}(t)$ is the cumulative number of departures of type v vehicles from terminal k at instant t , and $A_{v,k}(t)$ is the cumulative number of arrivals of type v vehicles at terminal k at instant t . The sum of these assignments over all terminals and types of vehicles gives the minimum required fleet size for satisfying the entire system schedule. This value can be reduced through modifications to the route schedules such as marginally reducing frequencies, advancing or delaying arrivals, changing assigned vehicle types for a given run, and repositioning vehicles from terminals where more vehicles arrive than the number departing for terminals with a shortage of arrivals. This repositioning usually involves deadhead runs, but where demand exists it may also be traveled with passengers (creating new express services, for example).

The vehicle scheduling process aims at defining vehicle blocks consisting of feasible sequences of runs and repositionings (deadhead vehicle trips) for each vehicle between route terminals or terminals and the vehicle garage. The objective is to minimize a combination of required fleet size and non-revenue time subject to satisfying all scheduled runs. This problem of assigning vehicles to runs starting from a single depot is called the *single depot vehicle scheduling problem* (SDVSP). The SDVSP can be solved in polynomial time. For a detailed overview of the literature on the SDVSP, see Desrosiers [12].

2. *Driver Shift Scheduling.* Once a block has been assigned to each vehicle, they are initially divided into tasks that must be worked by a single driver. So *tasks* start and end at relief points of the network, where vehicles can exchange drivers. The problem of assigning drivers to tasks is called the *crew scheduling problem* (CSP). A CSP is solved for each day of the planning horizon. In the CSP, a minimum cost set of *duties* (sequence of tasks by a

single driver) must be determined so that all tasks are assigned to a duty and each duty can be performed by a driver. The possibility of drivers doing overtime at a higher salary is usually considered in a CSP. Driver constraints like maximum working time without a break, minimum break duration, and maximum working time are typically added. Although CSP looks similar to the SDVSP in structure, it is much more complicated. The CSP has been shown to be NP-hard. Wren and Rousseau [13] present an overview of different approaches for this problem.

Once the driver duties over the planning horizon have been determined, duties are assigned to specific drivers. At this stage, called the *rostering process*, other constraints like rest periods, holidays, training, and so on, are considered. A survey on different approaches for this problem can be found in Odoni *et al.* [14].

This sequential approach for transit systems is inherited from the airline industry, where vehicles are optimized before drivers (crews in this case) since vehicle-related costs have a bigger impact on the total budget. In the case of transit systems based on buses, driver-related costs may be the most significant. Therefore, transit systems may benefit from optimizing drivers first and should do it taking full advantage of driver's flexibility.

This sequential approach to design a transportation system has other important limitations. Performance improvements could be obtained if the problem was solved simultaneously instead, but the complexity of the problem makes it a challenge.

Another drawback of the sequential approach is that it assumes a deterministic system. However, uncertainties are present at all operation levels (e.g., driver absenteeism, passenger demand, roundtrip durations, and vehicle failure) and should be considered in these decisions (e.g., hiring extra drivers, called *extra-board*, may allow operation

in case of absenteeism). Also, an effective control reaction to these and other uncertainties should be designed, as is explained in the next section.

REAL-TIME CONTROL

In the absence of any control system, the vagaries of traffic and of passenger demand at the various stops along a route tend to result in the formation of convoys of road transit vehicles. This phenomenon, known as *bus bunching* [15], has negative effects on the functioning of the transit system. One of these is the impact on average system wait times and their variability. Formally, average wait time $E[W]$ assuming users arrive randomly at a stop is given by

$$E[W] = \frac{E[H]}{2} + \frac{\text{Var}[H]}{2E[H]}, \quad (3)$$

where $E[H]$ and $\text{Var}[H]$ are, respectively, the mean and variance of vehicle headways. Bunching is thus highly undesirable, because it generates a noticeable rise in the variance of the intervals separating successive vehicle runs, which translates into average wait times that are longer and more variable. It also leads to underutilization of vehicle capacity since most passengers will be concentrated on the first vehicle in the bunch. This means that the average user will not only wait longer but also be riding on a fuller vehicle. Moreover, in high-demand scenarios vehicle load capacity is often exceeded and waiting passengers are unable to board the first arriving vehicle. These bunching phenomena impact strongly on perceived service levels, especially since, as mentioned, waiting time has the highest subjective value of all travel time components.

To avoid bunching, various control strategies may be adopted for regularizing headways. These include (i) holding vehicles at stops to prevent them catching up to the vehicle ahead; (ii) skipping stops at which no passenger wants to alight in order to make up time; and (iii) installing a traffic signal priority (TSP) system that can speed up or hold back vehicles.

Other eventualities, such as vehicle breakdowns, closed streets, and so on, may arise. The appropriate actions to be taken in these and other cases should be set out in contingency plans defining how vehicles and personnel are to be reassigned in each type of situation.

CONCLUDING REMARKS

Although the division of decision making into four levels is useful in ordering and ranking the various transit scheduling problems, it must be kept in mind that the design decisions at the strategic level are dependent on the solutions found for the lower level decision problems because of their impact on cost indicators. To solve the strategic design questions, estimates are required of such factors as how many drivers are needed per vehicle, how frequent service on a given route should be, and how long will waiting times be at that frequency, and so on. One of the ongoing challenges for researchers in this area is to develop methodologies that enable these problems at different levels to be solved simultaneously.

For each of the decision levels, we have described a sequence of problems that must be addressed. Each of these problems is complex in and of itself and generally involves not just technical challenges but social, psychological, economic, political, and urban design aspects as well. The presence of these other issues indicates that although operations research, with its techniques based on various types of mathematical modeling (optimization, control, econometric, etc.), is a highly valuable tool, it must be complemented by other decision-making mechanisms.

For an advanced coverage on the mathematical models and solution methodologies of the different problems covered in this article, the reader is referred to Desaulniers and Hickman [9] and Ceder [3].

Acknowledgments

This research was supported by the Across Latitudes and Cultures—Bus Rapid Transit Centre of Excellence funded by the Volvo

Research and Educational Foundations (VREF). We also acknowledge funding received through the CONICYT project “Anillos Tecnológicos ACT-32.”

REFERENCES

1. American Public Transportation Association. Public transportation fact book. 56th ed. Washington (DC): American Public Transportation Association; 2005.
2. Ceder A, Wilson NHM. Bus Network Design. *Trans Res Part B* 1986;20B(4):331–344.
3. Ceder A. Public transit planning and operation: theory, modeling and practice. 1st ed. Amsterdam: Butterworth-Heinemann, Elsevier; 2007.
4. Vuchic VR. Urban transit: operations, planning and economics. 1st ed. Hoboken (NJ): John Wiley and Sons; 2005.
5. Muñoz JC, Gschwender A. Transantiago: a tale of two cities. *Res Trans Econ* 2008;22: 45–53.
6. Boardman A, Greenberg D, Vining A, *et al.* Cost-benefit analysis. New Jersey: Prentice Hall; 2001.
7. Welding PI. The instability of a close-interval service. *Oper Res Q* 1957;8(3):133–148.
8. Newell GF. Dispatching policies for a transportation route. *Trans Sci* 1971;5(1):91–105.
9. Desaulniers G, Hickman MD. Public transit. In: Barnhart C, Laporte G, editors. Volume 14, Handbook in OR and MS. Transportation. Amsterdam: Elsevier B.V; 2007.
10. Furth PG, Day FB. Transit routing and scheduling strategies for heavy-demand corridors. *Trans Res Rec* 1985;1011:23–26.
11. Salzbom FJM. Optimum bus scheduling. *Trans Sci* 1972;6(2):137–148.
12. Desrosiers J, Dumas Y, Solomon MM, *et al.* Time constrained routing and scheduling. In: Ball MO, Magnanti TL, Monma CL, *et al.*, editors. Volume 8, Handbooks in OR and MS. Network routing. Amsterdam: North-Holland Publishing Company; 1995. pp. 35–139.
13. Wren A, Rousseau JM. Bus driver scheduling—an overview. In: Daduna JR, Branco I, Paixao JMP, editors. Proceedings of computer-aided scheduling of public transport. Germany: Springer; 1995. pp. 173–187.
14. Odoni AR, Rousseau JM, Wilson NHM, *et al.* Volume 8, Models in urban and air transportation. Handbooks in OR & MS: operations research and the public sector. The Netherlands: North Holland Publishing Company; 1994. pp. 107–150.
15. Newell GF, Potts RB. Maintaining a bus schedule. Volume 2, Proceedings of the 2nd Australian Road Research Board, Vol 2; Melbourne; 1964. pp. 388–393.

OPTIMIZATION PROBLEMS IN PASSENGER RAILWAY SYSTEMS

ALBERTO CAPRARA
DEIS, Università di Bologna,
Bologna, Italy

LEO KROON
RSM, Erasmus University
Rotterdam, Rotterdam, The
Netherlands

Department of Logistics, NS
Reizigers, Utrecht, The
Netherlands

PAOLO TOTH
DEIS, University of Bologna,
Bologna, Italy

Planning and operational processes related to railway systems are generally highly complex and are therefore fields that are rich in interesting *operations research* (OR) problems. Railway transportation can be split into passenger and cargo transportation. This article mainly focuses on the planning processes of passenger railway transportation systems.

Currently, in many countries, new regulations specify that the *management* of the railway infrastructure should be separated from its *exploitation*. This introduces, on the other hand, the organization of the *Infrastructure Manager* (IM), who is responsible for expansion and maintenance of the railway infrastructure as well as for capacity allocation and real-time traffic control, and, on the other hand, the *Train Operators* (TOs). The TOs provide passenger and cargo rail services by developing their preferred timetable, and by operating their rolling stock and crews. They usually operate on a commercial basis. In cooperation with the TOs, the IM integrates the preferred timetables of the TOs into one published timetable.

Simulation has always been recognized by the railway industry to be a fundamental OR tool for the planning phase, to be applied

extensively before putting any new timetable into operation in order to hopefully highlight its main strengths and weaknesses. The main advantage of simulation is that it allows the user to consider very detailed models of reality, and that it is fairly easy to adapt the associated software to refine/update these models. On the other hand, simulation is aimed at *evaluating* only possible solutions to the planning problems to be faced, whereas the number of these solutions is huge due to their intrinsic combinatorial structure. For example, if there are 12 trains in a short time period on a certain route, then the number of possible cyclic orders of trains after another on this route equals already $11! = 39916800$. Furthermore, considering all routes in the railway network at the same time increases the number of possible solutions in an exponential way.

Until recently, the possible solutions that were tested through simulation and eventually put into operation were generally designed by human experts: although this approach may lead in the end to solutions that are acceptable in practice, the time to generate such solutions in this way is unavoidably long. This is undesirable in an environment that, due to competition, is increasingly aiming at service and efficiency. Moreover, it gives almost no room for systematic what-if analyses of different scenarios (e.g., what would change by increasing the minimum time between two consecutive train trips to be assigned to the same driver?). For these reasons, in the last years, we have witnessed an increasing interest toward the application of mathematical models and optimization techniques for solving railway planning problems.

This trend is facilitated by fast developments in computer hardware and algorithmic power. Furthermore, it also follows the example of the airline industry, where these techniques were introduced already some time ago in order to cope with privatization and competition. There are many similarities between railway and

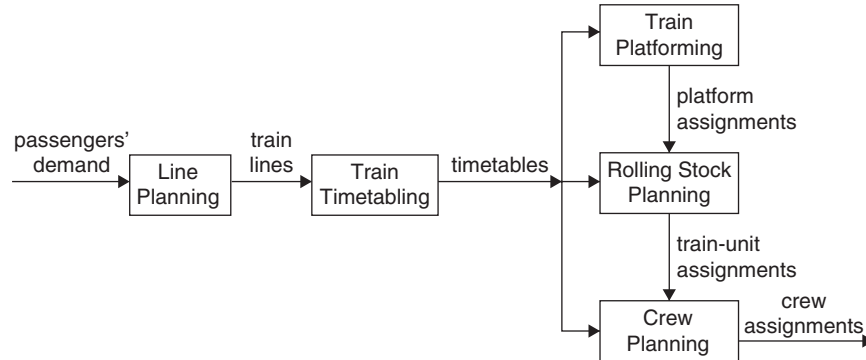


Figure 1. Illustration of the main problems solved in the planning of a passenger railway system.

airline systems. For example, both operate based on a timetable by allocating vehicles (airplanes vs trains) that are staffed by certain crews (cockpit and cabin crew versus train drivers and conductors). However, there are also differences. For example, an airline system usually has a much better insight into the numbers and destinations of the passengers than a railway system, since seat reservation is customary for airlines but often not for railways. Furthermore, in a railway system the infrastructure may be much more restrictive than in an airline system, and an airplane consists of a single vehicle whose type determines its capacity, whereas the capacity of a train can be adapted by coupling additional carriages or train units.

In this work, we focus attention on the various optimization problems that are faced in the planning and real-time operations of a passenger railway system, from the definition of the routes and frequencies of the trains in the railway network to the assignment of the workload to be carried out by the drivers and conductors, presenting them in the sequence in which they are faced in practice. We start with line planning, which amounts to deciding the routes for the trains as well as the types and frequencies of the trains on each route. Next, the actual timetable of each train has to be fixed by train timetabling and by assigning the trains to routes and platforms in the stations they visit. The assignment of rolling stock (locomotives and train carriages or train units) to the trains with a

known timetable is specified by the rolling stock planning. Finally, the assignment of the workload of drivers and conductors to operate the trains in a given timetable is the purpose of crew planning. This is illustrated in Fig. 1, which depicts the list of problems along with their input/output.

Surveys on optimization problems arising in passenger rail transportation systems can be found in Bussieck *et al.* [1], Caprara *et al.* [2], and Huisman *et al.* [3], to which the reader is referred for further details and references.

We conclude this introduction by noting that we will not discuss fundamental problems in strategic planning, related to deciding on the resources that are required to accomplish the passengers' demand, such as the capacities of the infrastructure, the rolling stock, and the crews. Although an enormous amount of money is involved with the management of these resources, the related strategic decisions are currently usually still taken without OR tools.

Additional passenger transportation problems are discussed in the articles titled *Air Traffic Management*, *Airline Resource Scheduling*, *Traffic Network Analysis and Design*, *Travel Demand Modeling*, and *Optimization of Public Transportation Systems* in this encyclopedia, whereas issues on rail cargo transportation and logistics are discussed in the article titled *Operations Research Applications in Truckload Freight Transportation Networks* in this encyclopedia.

Solution Methods

The simplest practical approach to solve the problems illustrated here is the use of various types of constructive and local search (meta-)heuristics, which often simulate what a human expert is doing to construct a solution manually. One clear drawback of these approaches is that they provide neither lower nor upper bounds to guarantee the quality of the solutions found. Moreover, this quality can often be improved by using mathematical programming tools.

For most problems considered here, as well as for most other optimization problems that are solved in practice (to the best of our knowledge), the mathematical programming methods that are used successfully are based on stating the problem as a *mixed-integer linear program* (MILP), often after having represented it on a suitable graph. These MILPs are solved as they stand (by using a general-purpose solver), or by applying constraint/column generation techniques combined with the solution of reduced *linear programming* (LP) relaxations (again by using a general-purpose solver), or also by applying methods to approximately solve (reduced) LP relaxations such as Lagrangian relaxation. The use of MILPs to model real-world problems and their practical solution is discussed in the articles titled **Formulating Good MILP Models**, **Branch-and-Bound Algorithms**, **Branch and Cut**, and **Branch-Price-and-Cut Algorithms** in this encyclopedia. In addition, constraint programming and its combination with MILP techniques appears to be of particular interest for railway timetabling [4].

THE MAIN PROBLEMS

In this section, for space reasons, we do not present the main MILP formulations of the problems explicitly, but we limit ourselves to mention their decision variables, along with the possible use of graphs to derive these formulations. Moreover, for each problem, we give some references where more details on mathematical models, exact and heuristic algorithms, and computational results on real-world instances can be found. Several

additional references are listed in the three surveys [1–3] mentioned above.

Line Planning

Most passenger TOs operate a timetable in which the scheduled trains can be partitioned into a number of so-called *lines*, containing trains with the same route and the same set of stop stations. The only difference between the trains of each line is the arrival and departure times. The *frequency* of a line denotes the number of trains that run in each direction per time unit (e.g., hour or day) on their common route, whereas its *capacity* is the overall number of seats provided by the trains on the line. In a *cyclic* line system, the difference of the departure times of the trains in the same direction is always about one (submultiple of the) cycle time.

The *line planning problem* (LPP) is the problem of designing a line system such that the *passengers' demand* is satisfied. There are two main conflicting objectives, namely, (i) maximizing the service toward the passengers and (ii) minimizing the operational costs of the railway system. According to (i), *direct passengers* are the passengers who can travel from their origin station to their destination station without having to change trains. Maximizing the number of direct passengers usually results in long lines: the longer a line, the more direct connections are provided. However, long lines may transfer delays more easily over a wider geographical area, and they may prohibit an efficient allocation of rolling stock.

LPP requires to define, for each possible line, whether to use it at all (i.e., to have trains running accordingly) and in this case with which frequency and capacity. The objective is to minimize the weighted sum of the connection costs for the passengers and the operating costs for the TO, subject to the constraint of satisfying the passengers' demand.

Note that if the line system is assumed to be symmetric (i.e., each line is operated in both directions with the same frequency), and all trains on a line have the same capacity, then it may be assumed without loss of generality that the origin/destination matrix is symmetric. However, if the

origin/destination matrix is strongly non-symmetric, for example, during rush hours, then it may be useful to allow the line system to be nonsymmetric as well.

A basic version of LPP assumes that the passengers have been already routed through the network before the optimization, for example, based on the assumption that they use shortest paths and intercity services as much as possible. On the basis of this system split assumption, LPP decomposes into two subproblems, concerning, respectively, intercity lines and regional lines. An advantage of this decomposition is that, within each subproblem, the stations may be considered as having the same type.

This basic version of LPP is naturally formulated by representing the railway network as an undirected graph $G = (V, E)$, with a node in V for each station and an edge in E for each track joining two stations. Moreover, assume that the passengers' demand is given by the matrix $d_{(i,j)}$, representing the number of passengers who want to travel from node i to node j . Then an initial MILP formulation is obtained by listing explicitly all possible lines with their frequencies and capacities, say l, f, c , and then by introducing binary variables $x_{l,f,c}$, which are equal to 1 if line l with frequency f and capacity c is activated, and "continuous" variables $y_{l,(i,j)}$ expressing the number of direct passengers between nodes i and j along line l . Details of this basic model can be found in Goossens *et al.* [5].

In extended versions of LPP, the passengers also have to be routed through the network. In this case, the intercity lines and the regional lines must be considered simultaneously. This approach clearly assumes that there is some "authority" that guides the passengers through the railway network from their origin to their destination. Furthermore, one may avoid considering the whole set of lines explicitly, using dynamic column generation to generate them. Details of these extended approaches can be found in Borndörfer *et al.* [6].

Train Timetabling

The general aim of the *train timetabling problem* (TTP) is to provide a timetable for the

trains of the TOs on a certain part of the railway network. For both cyclic and non-cyclic timetables, the problem for the general case of many TOs and a unique IM can be formalized schematically as follows. Each TO generates his/her own *ideal timetable* and submits this to the IM. Then the IM, in cooperation with the TOs, integrates the ideal timetables of the TOs into one single timetable. Thus, there are two stages of TTP. In the first stage, the ideal timetables of the individual TOs have to be generated. In the second stage, the ideal timetables of the TOs have to be integrated into a single timetable.

Relevant aspects to be taken into account in TTP are running times and dwell times of single trains, and headway and connection times between pairs of trains. Relevant objectives to be pursued in the first stage of TTP are to minimize the travel times of the passengers or to maximize the efficiency or the robustness of the timetable. In the second stage, the ideal timetables of the TOs are integrated into one single timetable. Here, the input consists of the ideal timetables of the TOs, along with a *fee* the TO is willing to pay to the IM for each of his/her trains and the *penalties* to be paid by the IM for possible deviations of the integrated timetable from the ideal timetables. Now the TTP is the problem of determining, for each train, whether to schedule it or not, and, in the former case, its *actual timetable*. The objective is to maximize the difference between the global fee of the trains scheduled and the global penalty for the deviations from the actual timetables with respect to the ideal ones. These deviations are required to satisfy the operational constraints that impose, for safety reasons, a minimum headway time between trains traveling on the same section of the infrastructure.

There are two main MILP formulations that are used both for the first and second stage of the TTP. The first formulation can be used to generate a cyclic or a noncyclic timetable. This MILP formulation discretizes the time horizon; for example, for a time horizon of one day, it assumes that all times are expressed by an integer number of time units (e.g., minutes). In the cyclic case, all departure and arrival times are

expressed modulo (a submultiple of) the cycle time. Then the problem is represented in a directed space–time graph whose nodes represent departure or arrival of trains at stations along specific tracks. The arcs of this graph represent either a train traveling between two adjacent stations (in case they join a departure node to an arrival node), or a train stopping at a station (in case they join an arrival node to a departure node). The timetable for a single train corresponds to a path in this graph. The natural MILP formulation has binary variables $x_{t,a}$ equal to 1 if the path for train t uses arc a . Details can be found in Caprara *et al.* [7].

The other MILP formulation is based on the periodic event scheduling problem (PESP), which was introduced by Serafini and Ukovich [8]. PESP focuses on the generation of a cyclic timetable with cycle time T . The MILP model has “continuous” variables $x_{t,e}$ expressing the time in the interval $[0, T - 1]$ at which the *event* e , corresponding to a departure/arrival at a station, occurs for train t . Furthermore, each constraint imposes that the difference between two variables $x_{t,e}$ and $x_{u,f}$ must be between a given lower bound and a given upper bound (modulo the cycle time T). All constraints can be represented in a so-called *constraint graph*. The modulo operation is linearized by introducing binary variables, whose “big M ” coefficient equals the cycle time. An alternative formulation of PESP does not use decision variables for the event times, but for the process times. Here one has to satisfy the additional constraint that the sum of the process times along any directed cycle in the constraint graph is a multiple of the cycle time T . Details can be found in Liebchen [9].

Train Platforming

The definition of the actual timetables within TTP generally does not take into account the actual routing of the trains within the stations. This is the purpose of the *train platforming problem* (TPP), which must be solved separately for each station, in which one has to find for each train a so-called *route* from the point where it enters the station to the point where it leaves the station, taking into account that the times in which these two

events occur have already been specified by TTP.

Assuming for simplicity that each train has to stop at a platform, a route is specified by the platform itself plus arrival/departure routes joining its entry/exit points to the platform. Several objectives are generally taken into account, such as the minimization of the number of platforms used and the penalties incurred if a train is not assigned to a platform within a given preference list. Also a maximal robustness of the platform assignment may be pursued, since the stations are often considered as the bottlenecks in the railway system. Operational constraints impose a minimum time interval between the departure of a train and the arrival of the next train at a platform, and forbid the simultaneous occupation of an incompatible pair of routes (generally physically crossing in the station layout) by two trains.

The commonly used MILP formulation of TPP considers explicitly the full list of routes for each train (which is generally acceptably long), and uses binary variables $x_{t,r}$ equal to 1 if train t is assigned to route r , plus possibly binary variables y_p equal to 1 if platform p is used by at least one train. Details can be found in Zwaneveld *et al.* [10] and Caprara *et al.* [11]. Recently Caimi [12] presented an approach focusing on the railway infrastructure: it guarantees that each section of the infrastructure is claimed by at most one train at the same time, which leads to a stronger LP relaxation.

Rolling Stock Planning

The rolling stock to be assigned to the trains whose timetable has been found by TTP can either be locomotives and train carriages, or aggregated modules, subsequently called *train units*. The latter are composed of a number of carriages in a fixed composition, and can move in both the directions without the need of an extra locomotive. A train can then be composed of several coupled train units. Train units may exist in different types, representing technical characteristics and capacity. For simplicity, we restrict our attention to the train unit case.

In order to obtain a better match between the available rolling stock and the passengers' seat demand, the compositions of the trains usually can be changed at several stations by adding or removing rolling stock from the trains. We let a *trip* to be a part of a train timetable that must be performed by the same rolling stock without changes. The *rolling stock planning problem* (RSPP) calls for assigning (combinations of) train units to trips.

The nature of the railway network plays a major role in the problem, in particular, in relation to the maintenance needs of the train units. Namely, in a sparse network with long distances, and hence long travel times and relatively low frequencies of trains, RSPP must specify in detail the circulations of the individual train units over the planning horizon, thereby taking into account the fact that each individual train unit should reach a maintenance center sufficiently often. In contrast, in a dense network with relatively short distances and high frequencies of trains, the train unit circulation is usually anonymous but contains sufficient time reservations for maintenance. Maintenance can be planned then just before the actual operations by exchanging duties of train units based on the assumption that all train units of the same type that do *not* need maintenance at that time are interchangeable.

In all cases, RSPP requires the minimization of a weighted sum of the costs of the total distance traveled by train units, of the composition change costs, and of the seat shortages with respect to the expected passengers' demand, subject to the train unit availability. Operational constraints impose a maximum length for each train (based on the shortest platform along which the train will dwell), besides requiring that composition changes can be done by respecting the timetable and that maintenance is carried out properly in the sparse network case.

A natural representation of RSPP is based on a directed multigraph $G = (V, A)$ with a node in V for each trip. The arc set A is partitioned into different sets A^t , one for each train unit type t . An arc $(i, j) \in A^t$ represents the fact that trip j can appear right after trip i in a feasible sequence for train unit type t . With

this representation, RSPP can be modeled as a multicommodity flow problem. Namely, it calls for a min-cost collection of *paths* of G , each associated with arcs of the same type, such that each node is visited by a sufficient number of paths, with several constraints on the path feasibility. This can either be modeled by using integer variables $x_{(i,j)}^t$ equal to the number of times arc $(i, j) \in A^t$ is in a path in the solution, or by considering the whole list $\mathcal{P}^t$ of paths for train unit type t and integer variables x_P equal to the number of paths $P \in \mathcal{P}^t$ in the solution. In this case, dynamic column generation is a useful approach due to the large number of potential paths. Details can be found in Cordeau *et al.* [13] and Cacchiani *et al.* [14]. Fioole *et al.* [15] use the concept of *transition graphs* to restrict the possible changes in the compositions of the trains in the stations.

Planning the shunting processes related to train units that are temporarily idle at a station, involving, for example, matching of arriving and departing train units, routing them through the station, and parking them in appropriate shunting tracks, is also an important part of RSPP. Several aspects of these problems have been modeled as MILPs [16].

Crew Planning

The *crew planning problem* (CPP) is the counterpart of RSPP involving crews (train drivers and conductors). It is concerned with covering each trip in the timetable with the required number of crews. That is, at least one driver and a sufficient number of conductors must be assigned to each trip. Here, a *trip* is a part of a train timetable that must be performed by the same crew without changes. Generally these trips are shorter than the trips for the rolling stock since a crew change requires less time. CPP represents a very complex and challenging problem, due to both the size of the instances to be solved and the type and number of operational constraints. It is usually decomposed into the following two phases.

In the *crew scheduling* phase, the short-term schedule of the crews is considered, and a convenient set of *duties* (also called *pairings*) covering all the trips is constructed.

Each duty represents a sequence of trips to be covered by a single crew member within a short time period overlapping at most one or two consecutive days. In the *crew rostering* phase, the duties selected in the crew scheduling phase are sequenced to obtain the final rosters, which describe for each crew member the sequence of duties to be carried out on consecutive days. Here, trips are no longer taken into account explicitly, but determine the *attributes* of the duties which are relevant for the roster feasibility and cost.

The main objective of CPP is the minimization of the global number of crews needed to perform all the daily occurrences of the trips in the given planning horizon. In some applications, for example when the considered trains cover a relatively small area and run mainly within the daytime, the crew rostering phase plays a minor role, since the corresponding constraints are rather weak and the number of crews is easily determined from the solution of the crew scheduling phase. In this case, the objective used in the crew scheduling phase mainly calls for the minimization of the number of working days corresponding to the duties. In applications involving several trains covering a wide area and/or running overnight, instead, considerable savings can be obtained through a clever sequencing of the duties obtained in the first phase. Therefore, the objective of the crew scheduling phase has to take into account the characteristics of the duties selected and their implication in the subsequent crew rostering phase.

Both crew scheduling and rostering phases in CPP, as well as RSPP, require finding min-cost *sequences* through a given set of *items*. Items correspond to trips for crew scheduling and for RSPP and to duties for crew rostering, whereas sequences correspond to duties for crew scheduling and to rosters for crew rostering. A natural representation is on the same graph considered for RSPP, and the MILP models are analogous, with the difference that often there is a unique crew type and it is sufficient to assign one crew to each item. Details can be found in Caprara *et al.* [17] and Kroon *et al.* [4]. For solving such problems of practical size, dynamic column generation

is usually necessary due to the huge number of potential sequences of items, and to the complex rules that must be satisfied by these sequences, in particular in the case of CPP.

ROBUSTNESS ISSUES

As mentioned in the introduction, simulation is often used to validate the solutions of the problems mentioned before, and it points out their strengths and weaknesses with respect to *disturbances*. These are external events in the real-time operations that were not taken into account in the optimization of the plans. An example of a disturbance is the extended duration of a dwell time due to an unexpectedly large number of alighting and boarding passengers.

Recently, considerable efforts have been made to take into account disturbances already in the optimization phase, although unavoidably with models of reality that are less accurate than in simulation. Roughly speaking, the general idea of robustness is to optimize a weighted combination of the original objective function and a penalty due to the operational costs in case of disturbances (taking into account an estimate of the probability of the occurrence of these disturbances). These methods are discussed in a general setting in the articles titled ***Two-Stage Stochastic Integer Programming: A Brief Introduction, Dynamic Models for Robust Optimization, Introduction to Robust Optimization*** in this encyclopedia.

The first, and most natural, approach to introduce robustness is *stochastic programming*, in which, besides the *nominal* scenario corresponding to the absence of disturbances, an explicit list of the other possible scenarios must be given along with the associated probability. The resulting mathematical programming models have a subproblem for each scenario, their size growing therefore linearly in the number of scenarios. This forces the number of scenarios to be small in practical applications. Later, other approaches to deal with robustness that are less cumbersome than stochastic programming (and allow one to implicitly consider infinitely many scenarios) have been proposed.

The approaches in Soyster [18], Ben-Tal and Nemirovski [19], and Bertsimas and Sim [20] keep the original (nominal) objective function and guarantee that the solution is feasible not only for the nominal scenario but also for all other scenarios. They implicitly list these scenarios as a mathematical program, through uncertainties in the data defining the problem constraints. (For example, in Bertsimas and Sim [20], each entry of the constraint matrix can vary within given intervals and at most a given number of entries of the same row can vary at the same time.) These approaches have the clear disadvantage that they focus on finding a solution that is feasible for all scenarios, which may not even exist at all, and in any case may be of very bad quality.

This is relaxed in the *light robustness* approach of Fischetti and Monaci [21], which, starting from the formulation in Bertsimas and Sim [20], includes a new objective function taking into account the amount of constraint violation for all scenarios and a new constraint stating that the value of the original objective function should not deviate too much from the optimal nominal value.

In any case, the main drawback of all the methods in the above paragraph is that, unlike stochastic optimization, they do not consider the possibility of changing the nominal solution within the operations to adapt it to the scenario that is occurring in practice. This inspired the notion of *recoverable robustness* [22], which essentially combines the notion of a recoverable algorithm that is used to adapt the nominal solution to the actual scenario and the implicit representation of the list of scenarios as a mathematical program, without any assumption on the probabilities associated with these scenarios.

REAL-TIME CONTROL

Real-time control aims at monitoring the real-time system and at restoring the system once one or more events occurred that made the original plan infeasible or that will make it infeasible soon. Such an event may be a relatively small *disturbance*, or a large *disruption*. Real-time control may be carried

out in a reactive way, when the plan has become already infeasible, or in a proactive way, when the plan is going to become infeasible. Therefore, forecasting future states of the system is an important issue in real-time control. Anyway, for real-time control short computation times are extremely important, possibly leading to suboptimal solutions.

In the case of a disturbance, the infeasibility can usually be restored by relatively small measures by traffic control, for example, giving speed advises to trains or rerouting trains inside stations. In these cases, it is often assumed that the changes to the timetable can be conducted without rescheduling rolling stock and crews. D'Ariano *et al.* [23] describe a branch-and-bound procedure that can be used to improve the quality of such timetabling decisions. Their model is based on the disjunctive arc model by Mascis and Pacciarelli [24].

A closely related problem is delay management. This problem is used to find optimal wait-depart decisions for connecting trains when the feeder train is delayed, so that the total passenger delay is minimized. Schöbel [25] describes MILP models for solving the delay management problem. Here, the wait-depart decisions are modeled by binary variables.

A large disruption may be caused, for example, by malfunctioning infrastructure or rolling stock, or by an accident. Such a large disruption may lead to less railway traffic on the involved infrastructure for a certain period of time, or to no railway traffic at all. As a consequence, the timetable, rolling stock circulation, and crew duties must be adapted accordingly. A problem here is that the duration of the disruption is usually not known in advance.

For adapting the structure of the timetable, often disruption scenarios are used, in particular, in the case of a cyclic timetable. However, recovering the rolling stock circulation and the crew duties may be more complicated. Nielsen *et al.* [26] describe an adapted version of the rolling stock planning model of Fioole *et al.* [15] for this purpose. In order to keep their running times low, the model is applied on a rolling horizon.

Rezanova and Ryan [27] and Potthoff *et al.* [28] describe an adapted version of a set covering model for rescheduling the crew duties. They keep the running times low by only considering a well-selected subset of the original duties: the duties that probably cannot be helpful for the rescheduling of the disrupted duties are not considered. For both rolling stock and crew rescheduling, it is important to keep intact the parts of the original plans that need not be modified.

FINAL REMARKS

An overview of optimization models for operational railway line planning, train timetabling and platforming, and rolling stock and crew planning has been given. These models are applied more and more in practical planning environments. The next challenge is to apply them also in the dynamic environment of real-time control and disruption management.

REFERENCES

1. Bussieck M, Winter T, Zimmermann U. Discrete optimization in public rail transport. *Math Program* 1997;79:415–444.
2. Caprara A, Kroon L, Monaci M *et al.* Passenger railway optimization. In: Barnhart C, Laporte G, editors. *Transportation. Handbooks in operations research and management science* 14. Amsterdam: Elsevier; 2007. pp. 129–187.
3. Huisman D, Kroon L, Lentink R *et al.* Operations research in passenger railway transportation. *Stat Neerl* 2005;59:467–497.
4. Kroon L, Huisman D, Abbink E *et al.* The new dutch timetable: the or revolution. *Interfaces* 2009;39:6–17.
5. Goossens J, van Hoesel C, Kroon L. A branch-and-cut approach for solving railway line-planning problems. *Transp Sci* 2004;38:379–393.
6. Borndörfer R, Grötschel M, Pfetsch M. A column generation approach to line planning in public transport. *Transp Sci* 2007;41:123–132.
7. Caprara A, Fischetti M, Toth P. Modeling and solving the train timetabling problem. *Oper Res* 2002;50:851–861.
8. Serafini P, Ukovich W. A mathematical model for periodic scheduling problems. *SIAM J Discrete Math* 1989;2(4):550–581.
9. Liebchen C. Periodic timetable optimization in public transport [PhD thesis]. TU Berlin; 2006.
10. Zwaneveld P, Kroon L, van Hoesel C. Routing trains through a railway station based on a node packing model. *Eur J Oper Res* 2001;128:14–33.
11. Caprara A, Galli L, Toth P. Solution of the train platforming problem. Technical Report. Bologna: DEIS, University of Bologna; 2008.
12. Caimi G. Algorithmic decision support for train scheduling in a large and highly utilised railway network [PhD thesis]. ETH Zürich; 2009.
13. Cordeau J-F, Soumis F, Desrosiers J. Simultaneous assignment of locomotives and cars to passenger trains. *Oper Res* 2001;49:531–548.
14. Cacchiani V, Caprara A, Toth P. Solving a real-world train unit assignment problem. *Math Program* 2010;124:207–231.
15. Fioole P, Kroon L, Maróti G *et al.* A rolling stock circulation model for combining and splitting of passenger trains. *Eur J Oper Res* 2006;174:1281–1297.
16. Kroon L, Lentink R, Schrijver A. Shunting of passenger train units: an integrated approach. *Transp Sci* 2008;42:436–449.
17. Caprara A, Fischetti M, Toth P *et al.* Algorithms for railway crew management. *Math Program* 1997;79:125–141.
18. Soyster A. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper Res* 1973;21:1154–1157.
19. Ben-Tal A, Nemirovski A. Robust optimization-methodology and application. *Math Program* 2002;92:453–480.
20. Bertsimas D, Sim M. The price of robustness. *Oper Res* 2004;52:35–53.
21. Fischetti M, Monaci M. Light robustness. In: Ahuja R, Möhring R, Zaroliagis C, editors. *Robust and online large-scale optimization. Lecture notes in computer science* 5868. Berlin: Springer; 2009. pp. 61–84.
22. Liebchen C, Lübbecke M, Möhring R *et al.* The concept of recoverable robustness, linear programming recovery, and railway applications. In: Ahuja R, Möhring R, Zaroliagis C, editors. *Robust and online large-scale*

- optimization. Lecture notes in computer science 5868. Berlin: Springer; 2009. pp. 1–27.
23. D'Ariano A, Corman F, Pacciarelli D *et al.* Reordering and local rerouting strategies to manage train traffic in real time. *Transp Sci* 2008;42:405–419.
 24. Mascis A, Pacciarelli D. Job-shop scheduling with blocking and no-wait constraints. *Eur J Oper Res* 2002;143:498–517.
 25. Schöbel A. Optimization in public transportation. Springer optimization and its applications 3. Berlin: Springer; 2006. ISBN: 978-0-387-32896-6.
 26. Nielsen L, Kroon L, Maróti G. Capacity oriented rolling stock rescheduling in passenger railways. Technical Report. Rotterdam: Erasmus University Rotterdam; 2009. (Submitted to Transportation Science.)
 27. Rezanova N, Ryan D. The train driver recovery problem - a set partitioning based model and solution method. *Comput Oper Res* 2010;37:845–856.
 28. Potthoff D, Huisman D, Desaulniers G. Column generation with dynamic duty selection for railway crew rescheduling. Technical Report. Rotterdam: Erasmus University Rotterdam; 2008. (To appear in Transportation Science.)

OPTIMIZING THE AVIATION CHECKPOINT PROCESS TO ENHANCE SECURITY AND EXPEDITE SCREENING

ADRIAN J. LEE*

Department of Mechanical Science and
Engineering, University of Illinois at
Urbana-Champaign, Champaign,
Illinois

SHELDON H. JACOBSON

Department of Computer Science,
University of Illinois at
Urbana-Champaign, Urbana

OPTIMIZING PASSENGER SCREENING FOR AVIATION SECURITY

Passenger screening at aviation security checkpoints is a critical component in protecting airports, airlines, and passengers from terrorist threats. Historically, terrorist activity has directly influenced the reform of security screening strategies in an effort to adapt to an ever-expanding list of improvised threats. However, the reactive nature of many of these changes, such as the 3-1-1 liquid and gel policy for carry-on baggage, for example, has caused increases in the cost associated with resolving alarms for non-threatening passengers and inconvenience to travelers due to longer screening times.

Passengers are screened prior to entering the airport terminal using a set of security detection devices and procedures, such as magnetometers, explosive detection systems, pat-down searches, and explosive trace detection devices. There are two possible ways to screen passengers for threat items: (i) uniform screening and (ii) risk-based screening. In uniform screening, all passengers are deemed to pose equal

threat to the aviation transportation system, by carrying a concealed threat item within their clothing or baggage. Because of budget constraints surrounding the purchase, maintenance, and operation of the detection devices, it is neither economically nor logistically feasible to screen all passengers at the same level of intensity, especially using specialized devices and procedures that are time-intensive. Furthermore, a random selection policy for assigning passengers to undergo higher screening intensity exhibits the vulnerability to under screening passengers who are indeed a threat.

A risk-based screening strategy addresses this vulnerability by quantifying a perceived risk of each passenger, and assigning this passenger to be screened by an appropriate set of devices. The computer-aided passenger prescreening system (CAPPS), developed by the Federal Aviation Administration, Northwest Airlines, and the United States Department of Justice, is an automated risk assessment system to assess passengers' risk, based on factors such as personal information, prior travel history, and other security-sensitive information [1]. CAPPS designates passengers as either *selectees* (i.e., those not cleared by CAPPS) or *nonselectees* (i.e., those cleared by CAPPS). When capacity permits, non-selectees are randomly chosen to undergo higher screening intensity in an effort to fully utilize the available security resources.

A risk-based screening strategy may also be used to quantify the perceived risk of passengers using a continuous spectrum of assessed threat values (as opposed to two distinct categories in CAPPS), where the passengers are directed to undergo certain levels of screening intensity based on a set of threshold values. This article develops a multiobjective optimization problem, where the goal is to obtain a threshold-based passenger assignment policy that balances the trade-off between maximizing security and minimizing the expected amount of time it takes to screen passengers and baggage through the security checkpoint. Stochastic models

*Affiliated with Vishwamitra Research Institute in 2009–2010.

are used to capture the queueing dynamics of the passengers as they proceed through the security checkpoint. Passengers arrive sequentially at the security checkpoint, and are directed to undergo the appropriate level of screening intensity based on uncertainties surrounding their corresponding perceived risk levels, the passenger arrival times, and the detection device screening times. An optimal policy is obtained for sequentially assigning passengers to a multilevel security system and to a security system comprising primary and secondary levels of screening.

PROBLEM FORMULATION

A multilevel security system within an airport terminal assigns each passenger to one of several available *security classes* according to the perceived risk for this passenger, where the security classes are composed of sets of security detection devices and procedures. Each device and procedure individually specializes in detecting a specific type of threat, such as guns, knives, and explosives, while operating collectively to screen a passenger or bag at a certain level of intensity. The problem of assigning passengers to the appropriate security class is explored dynamically, where the assignment decision must be made as each passenger arrives at the security checkpoint for screening, and without knowledge of future passenger risk.

Let M be the number of security classes available for screening passengers, and let L_m denote the security level of class $m = 1, 2, \dots, M$, with $0 \leq L_1 \leq L_2 \leq \dots \leq L_M \leq 1$. The value L_m is defined as the conditional probability that an alarm occurs for a passenger who is a threat, given that the passenger is assigned to security class m (i.e., the conditional true alarm rate). Because of the procedures and technologies associated with each screening device, the rates at which passengers are screened vary across the security classes. Assume that due to these sets of screening devices, the security class service rates are exponential, with rates $\mu_1 > \mu_2 > \dots > \mu_M > 0$, where passengers are screened faster (slower) through security classes of lower (higher) levels of intensity

using a set of routine (specialized) devices and procedures. Also, assume that passengers arrive independently and sequentially at the security checkpoint according to a Poisson process with rate $\lambda > 0$. For stability, assume $\lambda < \sum_{m=1,2,\dots,M} \mu_m$ (i.e., collectively, the multilevel security system is capable of screening passengers at a rate faster than the arrival rate).

Passengers sequentially arrive at the security checkpoint for screening, indexed as $i = 1, 2, \dots$ upon arrival at times $t_1, t_2, \dots$. Upon check-in, the perceived risk for each passenger is quantified by security personnel into an *assessed threat value*, scaled between zero and one, captured through the independently and identically distributed random variables $AT_1, AT_2, \dots$, with realizations $\alpha_1, \alpha_2, \dots$. The cumulative distribution function $F_{AT}(\alpha_i), i = 1, 2, \dots$ is defined over $[0, 1]$. The value of α_i quantifies a risk perception for use in a risk-based security system, where passengers may be screened at various degrees of intensity. Since security personnel cannot know if a passenger is indeed a threat, the assessed threat value reflects their perception of the probability of that passenger being a threat. The value α_i corresponds to the probability that passenger i is a threat, where a value of α_i close to one (zero) corresponds to a high-risk (low-risk) passenger. A passenger cannot be identified as a threat until a security detection device signals an alarm and the alarm is confirmed to be positive. The value α_i is assigned to passenger i upon check-in, either at the ticket counter or on-line, but is considered unknown to the security personnel located at the terminal checkpoint, since the order of passengers might change from check-in to screening. Also, this allows the assessed threat value to be altered based on other screening techniques, such as behavior detection officers, prior to the passenger entering the screening process. Thus, the assessed threat value becomes available for assigning the appropriate level of screening once the passenger first enters the screening process, reflected through the realization of the random variable AT_i .

The decision to assign passenger i to security class m in a multilevel security system is

a function of the passenger's assessed threat value in relation to a set of *security class threshold values*, $b_m(t), m = 1, 2, \dots, M, t \geq 0$. The value $b_m(t)$ determines the upper threshold value for security class m , where $0 < b_1(t) \leq b_2(t) \leq \dots \leq b_M(t) \equiv 1$ partition the interval $(0, 1]$ into M subintervals. The security class assignment for passenger i is represented by the Bernoulli random variable, $X_m(t_i) = 1$ (0) if passenger i is (not) assigned to class m , with observed value, $x_m(t_i)$. Define $p_m(t_i) = P(X_m(t_i) = 1), m = 1, 2, \dots, M, i = 1, 2, \dots$, as the probability that passenger i entering screening at time $t = t_i$ is assigned to security class m . The objective is to determine the values $b_m(t_i), m = 1, 2, \dots, M$ that assigns passenger i to the appropriate security class, subject to capacity constraints. With $p_m(t_i) = P(X_m(t_i) = 1) = P(b_{m-1}(t_i) < AT_i \leq b_m(t_i)), m = 1, 2, \dots, M, i = 1, 2, \dots$, the values $b_m(t_i)$ are related to $p_m(t_i)$ through $b_m(t_i) = F^{-1}_{AT} \left(\sum_{k=1,2,\dots,m} p_k(t_i) \right), m = 1, 2, \dots, M$. Therefore, obtaining the optimal security class assignment probabilities $p_m(t_i)$ for each passenger arriving sequentially at the security checkpoint is equivalent to obtaining the optimal security class threshold values $b_m(t_i)$, which assign each passenger to undergo the appropriate level of screening.

PASSENGER ASSIGNMENT POLICY

First, consider a passenger assignment policy in which each passenger is assigned to a security class independent of all prior passenger assignments. This implies that the assignment decision operates independent of the number of passengers currently in the security system, and sequentially assigns the passengers according to some predetermined set of constant security class threshold values. This type of policy can be applied, for example, to the random selection of nonselectee passengers for higher intensity screening, in an effort to completely utilize the set of devices used specifically for screening selectee passengers.

A *steady-state* queueing analysis is used to provide an expression for the steady-state expected amount of time a passenger spends

in a security system, $E[W]$. The passenger screening process is modeled as a multilevel, single-server exponential queueing system with infinite capacity (see also the section titled "Queueing Theory and Queueing Networks" in this encyclopedia). The objective of minimizing the steady-state amount of time each passenger spends in the security system can be solved through the following nonlinear program (NLP) for a set of constant values $p_m, m = 1, 2, \dots, M$ (see also the section titled "Nonlinear Programming and Global Optimization" in this encyclopedia)

$$\begin{aligned} \text{minimize } E[W] &= \sum_{m=1}^M p_m / (\mu_m - \lambda p_m) \\ \text{subject to } 0 &\leq p_m \leq 1, m = 1, 2, \dots, M, \\ p_m &< \mu_m / \lambda, m = 1, 2, \dots, M, \\ \sum_{m=1}^M p_m &= 1. \end{aligned} \quad (1)$$

The second (strict) inequality constraint, $p_m < \mu_m / \lambda$, corresponds to the stability condition, and, in practice, can be replaced with $p_m + \varepsilon_m \leq \mu_m / \lambda$, for some $\varepsilon_m > 0$. The minimum values of $p_m, m = 1, 2, \dots, M$ are obtained by taking the limit as $\varepsilon_m \rightarrow 0$. For a two-class system, the minimum solution to the NLP in Equation (1) can be obtained by substituting the relation $p_2 = 1 - p_1$ and differentiating to solve for the optimal value p_1^* . Note that, if the security class service rates are identical, $\mu_1 = \mu_2 = \dots = \mu_M$, then $p_1^* = p_2^* = \dots = p_M^* = 1/M$.

Alternatively, a *transient* queueing analysis allows the queue length to be used as feedback into a dynamic passenger assignment process to balance the expected amount of time a passenger spends in the security system with the expected number of true alarms. Let the discrete random variable TA_i be defined as the number of *true alarms* (i.e., correctly detecting a threat item) over the first i passengers to enter screening. Then,

$$E[TA_i] = \sum_{\tau=1}^i \sum_{m=1}^M L_m p_m(t_\tau) \alpha_\tau \quad (2)$$

is the expected number of true alarms in the security system over the first i passengers. Let the continuous random variable $W(t_i)$ be defined as the amount of time passenger i spends in the security system. The optimal assignment policy that determines the sequence of security class threshold values $\{b_m(t_i), i = 1, 2, \dots, m = 1, 2, \dots, M\}$ is the policy which minimizes $E[W(t_i)]$ and maximizes $E[TA_i]$ for each passenger assignment. Clearly, if minimizing $E[W(t_i)]$ is not an issue, then the policy that maximizes $E[TA_i]$ is simply to have all passengers undergo maximum screening (i.e., $p_M(t_i) = 1, i = 1, 2, \dots$). However, by defining a dual objective, a balance must be achieved between security and passenger throughput.

To analyze the passenger screening process, the airport security checkpoint is modeled as a multilevel, single-server exponential queueing system, where each security class has finite capacity $c_m, m = 1, 2, \dots, M$. The number of passengers in security class m at time t_i , denoted by the discrete random variable $S_m(t_i)$, is formulated as a discrete-time, nonhomogeneous Markov chain, with transition probabilities $P_m^{kj}(t_i), m = 1, 2, \dots, M, i = 1, 2, \dots$ (see also the section titled “Discrete-Time Markov Chains” in this encyclopedia). Let the continuous random variable $W_m(t_i)$ be defined as the amount of time passenger i spends in security class m . A recursion for $E[W_m(t_i)]$ is obtained as

$$\begin{aligned} E[W_m(t_{i+1})] \\ = E[W_m(t_i)] + p_m(t_i)(1 - P(N_m^s(t_i, t_{i+1}) = 0, \\ S_m(t_i) = c_m))/\mu_m - E[N_m^s(t_i, t_{i+1})]/\mu_m, \end{aligned} \quad (3)$$

$m = 1, 2, \dots, M, i = 1, 2, \dots$, with $E[W_m(t_1)] = 1/\mu_m$, and where $N_m^s(t_i, t_{i+1})$ is the discrete random variable corresponding to the number of passengers screened in class m during the time interval $(t_i, t_{i+1}]$. From Equation (3), the expected amount of time passenger i spends in the security system, $E[W(t_i)]$, can

be obtained from

$$\begin{aligned} E[W(t_i)] &= \sum_{m=1}^M p_m(t_i) E[W_m(t_i)], \\ m &= 1, 2, \dots, M, i = 1, 2, \dots \end{aligned} \quad (4)$$

The dual objectives of minimizing the expected amount of time each passenger spends in the security system and maximizing the expected number of true alarms can be combined into a single weighted objective function by normalizing each objective term, and introducing cost weight parameter η_1 . The mean square cost associated with the true alarm probability results from the difference between assigning passenger i to class m and assigning passenger i to class M (i.e., the highest level of screening intensity) is defined by

$$\begin{aligned} C^Z(t_i) &\equiv \left[1 - \sum_{m=1}^M p_m(t_i)(L_m - L_1)/(L_M - L_1) \right]^2, \\ i &= 1, 2, \dots, \end{aligned} \quad (5)$$

such that $0 \leq C^Z(t_i) \leq 1$. Next, the mean square cost associated with the expected amount of time passenger i spends in the security system is defined by

$$\begin{aligned} C^W(t_i) &\equiv \left[\left(\sum_{m=1}^M p_m(t_i) E[W_m(t_i)] - w_{\min} \right) \right. \\ &\quad \left. / (w_{\max} - w_{\min}) \right]^2, i = 1, 2, \dots, \end{aligned} \quad (6)$$

where $w_{\max} = \max_{m=1,2,\dots,M} \{E[W_m(t_i)]\}$ and $w_{\min} = \min_{m=1,2,\dots,M} \{E[W_m(t_i)]\}$ such that $0 \leq C^W(t_i) \leq 1$.

The policy that minimizes the cost-weighted dual objective can be obtained by solving the following NLP for $p_m(t_i), m = 1, 2, \dots, M$, for each successive passenger $i = 1, 2, \dots$ arriving at the security checkpoint:

$$\begin{aligned} &\text{minimize } (1 - \eta_1)C^Z(t_i) + \eta_1 C^W(t_i) \\ &\text{subject to } 0 \leq p_m(t_i) \leq 1, m = 1, 2, \dots, M, \\ &\quad \sum_{m=1}^M p_m(t_i) = 1. \end{aligned} \quad (7)$$

The scalar weight $0 \leq \eta_1 \leq 1$ is used to create a cost trade-off between security versus passenger throughput. The value $\eta_1 = 0$ places full emphasis on security, by assigning all passengers to undergo the highest level of screening intensity, while the value $\eta_1 = 1$ places full emphasis on expediting screening, by assigning each successive passenger into the security class with minimal expected amount of system time. For a two-class system, the minimum solution to the NLP in Equation (8) can be obtained by substituting the relation $p_2(t_i) = 1 - p_1(t_i)$ and differentiating to solve for the optimal value $p_1^*(t_i)$.

PRIMARY AND SECONDARY SCREENING

In a security system comprising primary and secondary levels of screening, passengers arrive at screening where all passengers and carry-on baggage undergo primary screening using a set of routine detection devices. Then, a selected number of passengers are directed to undergo secondary screening, consisting of specialized devices and procedures, while the remaining passengers exit without receiving additional screening. Passengers are directed toward secondary screening, with probability $q(t_i)$, based on the passenger's assessed threat value and the secondary screening queue length at the time the secondary screening decision is made. This type of security system structure is utilized when CAPPS designates passengers as either selectees or nonselectees, for example.

Let the number of nonselectee and selectee passengers arriving at the security checkpoint, $N_{NS}^a(t)$ and $N_S^a(t)$, follow a Poisson process with rates $\lambda(1 - p_s) > 0$ and $\lambda p_s > 0$, respectively, where $0 \leq p_s \leq 1$ is the fraction of passengers who are designated as selectees. The total number of passengers arriving at the security checkpoint, $N^a(t) = N_{NS}^a(t) + N_S^a(t)$, is Poisson with rate $\lambda > 0$. Assume that if passenger i is designated as a selectee (nonselectee), then the observed passenger's assessed threat value is $\alpha_i = 1$ ($0 < \alpha_i < 1$). Furthermore, assume that the primary and secondary screening service times are exponential with rates μ_1 and μ_2 , security levels L_1 and L_2 , and capacities c_1 and

c_2 , respectively. Define the Bernoulli random variable $X(t_i) = 1(0)$ if passenger i is (not) directed to secondary screening.

The queueing process is sampled when each passenger enters the decision point for assigning secondary screening. Following the primary screening stage, optimization of the secondary screening assignment decision is equivalent to that of a two-class security system, where passengers assigned to the lower security class exit without further screening (i.e., $L_1 = 0, E[W_1(t_i)] = 0$, and $w_{\min} = 0$), while passengers assigned to the higher security class undergo the secondary screening procedures. Substitution of $L_1 = 0, E[W_1(t_i)] = 0$, and $w_{\min} = 0$ into Equations (5)–(7) simplifies the NLP to

$$\begin{aligned} & \text{minimize } (1 - \eta_1)(1 - q(t_i))^2 + \eta_1 q^2(t_i) \\ & \text{subject to } 0 \leq q(t_i) \leq 1, \end{aligned} \quad (8)$$

with solution $q(t_i) = 1 - \eta_1$. Since the fraction of nonselectee passengers is $1 - p_s$, the optimal probability for assigning (nonselectee) passenger i exiting primary screening becomes

$$\begin{aligned} q^*(t_i) &= 1 - \frac{p_1^*(t_i)}{(1 - p_s)} \\ &= \begin{cases} \frac{(1 - \eta_1)}{(1 - p_s)} & \text{if } s_2(t_i) < c_2 \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots \end{aligned} \quad (9)$$

where $s_2(t_i)$ is the number of passengers in the queue for secondary screening at time t_i . Furthermore, $q^*(t_i) = 1$ if passenger i is designated as a selectee.

IMPACT AND LIMITATIONS OF PASSENGER SCREENING MODELS

Optimization of the passenger screening process is achieved by balancing the security and passenger throughput objectives via a cost weight parameter. This parameter can be qualitatively determined from current threat and environment factors to allow flexibility for either heightened security or expedited screening. Numerical solutions to the preceding NLP problems can be used to compare

security systems that partition the available detection devices into primary and secondary levels of screening versus those that partition the same set of devices into a two-class structure. This systematic approach to designing optimal screening strategies provides a unique perspective on the trade-off between security and passenger service, from which airport security operations personnel may use to fine-tune the security checkpoint process, expedite passenger screening, and further strengthen aviation security operation performance through an increase in the probability of threat detection. Incorporating a dual objective provides significant improvements in both air transportation safety and passenger satisfaction with the security screening process, and produces an automated system that is reliable, efficient, and cost effective. The results from this type of analysis play an important role in comparing different security screening structures, and provide a quantifiable means for presenting the benefits and limitations of each design to management officials within the aviation security sector.

Several mathematical assumptions place limitations on capturing the true behavior of the passenger screening process. In practice, service times do not follow an exponential distribution, since there exists a minimum amount of time to screen a passenger or bag. The screening process may be modeled using more complex service time distributions, which may also include either a time or queue length dependency on the service rate. Also, dependencies between the security classes can be modeled, where passengers who signal an alarm in one security class are directed to join the queue for a separate security class. This interaction would facilitate the use of highly specialized, time-consuming screening devices for high-risk passengers as well as for resolving alarms occurring from lower risk passengers. Lastly, optimization is performed over a security measure defined as the number of true alarms. In an effort to decrease cost and expedite screening, it is also desirable to minimize the number of *false alarms* (i.e., detecting a threat that does not exist) or minimizing the number of *false clears* (i.e., not detecting a threat).

However, there exists an inherent trade-off between minimizing the false alarm and false clear rates. These, as well as other alternative security measures, may be defined as objectives in the optimization problem.

ADDITIONAL OR TECHNIQUES IN PASSENGER SCREENING

In addition to queueing theory and non-linear programming, additional operations research techniques have been explored in related passenger and baggage screening research surrounding aviation security (see also the section titled “Anti-Terrorism and Homeland Security” in this encyclopedia). Discrete optimization models are used to develop policies that optimally screen passengers given a set of security screening devices [2]. Subsets of detection devices can be identified by solving an integer programming model, which defines multiple levels of security based on the device capacities within each security class. Each passenger is then assigned to a particular security class based on the perceived passenger’s risk level. Discrete optimization models have also been developed to sequentially assign passengers to a set of security classes, when the perceived passenger risk levels are known only upon check-in. Sequential passenger assignments are modeled as a Markov decision process, where the optimal policy for generating the set of assignments is obtained through dynamic programming [3], and also modeled as a discrete-time difference equation, where an optimal closed-loop passenger assignment algorithm is developed through both a probabilistic analysis and nonlinear control [4]. A two-stage model for a sequential stochastic security design problem analyzes the purchase of security devices and determines the screening assignments of sequentially arriving passengers, solved via a deterministic integer program [5]. Also, for sequential screening problems, a Bayesian analysis is used to incorporate prior knowledge of the device responses on a bag to develop metrics that assess the risk and cost associated with specific detection device configurations [6].

Linear programming models are used to investigate the benefit of dividing passengers into multiple levels of screening [7], with the objective of minimizing the number of false alarms, subject to a maximum false clear probability constraint. Extensions of this model accommodate various passenger risk levels and formulate the model as a mixed integer program [8], while also addressing the joint responses of the detection devices through a multivariate statistical analysis [9]. The deployment and utilization of new device technology can be addressed through the development of cost models to assess the risk and effectiveness among existing and proposed technologies [10–12]. Other applications of operations research techniques include the use of game theory to model the interaction between security system operations and terrorist-acquired knowledge of how the system behaves [13]. Further insight into the application of operations research techniques to aviation security issues can be found in the literature [14–16].

A key assumption in the trade-off between maximizing security and minimizing passenger screening times is that passengers become more dissatisfied with the security screening process the longer they wait to be screened. While many passengers may prefer to proceed through security with minimal hindrance, survey results indicate that many passengers also feel more secure when they experience a longer time in the screening process, given the thoroughness of passenger and baggage screening [17]. A statistical analysis studying the relationship between the waiting times of passengers at screening and their level of satisfaction with the screening procedures indicates that there exist factors other than waiting time that significantly affect the passengers' level of satisfaction [18]. The methodology employed throughout the research in aviation passenger screening provides the means to systematically address the problem of optimizing security given resource constraints. Resulting from various operations research techniques, priority can then be shifted readily between heightened security and expedited screening according to current threat and environmental factors.

REFERENCES

1. United States Government Accountability Office. Computer-assisted passenger pre-screening system faces significant implementation challenges. Publication GAO-04-385. 2004.
2. McLay LA, Jacobson SH, Kobza JE. Integer programming models and analysis for a multi-level passenger screening problem. *IIE Trans* 2007;39(1):73–81.
3. McLay LA, Jacobson SH, Nikolaev AG. A sequential stochastic passenger screening problem for aviation security. *IIE Trans* 2009;41(6):575–591.
4. Lee AJ, McLay LA, Jacobson SH. Designing aviation security passenger screening systems using nonlinear control. *SIAM J Control Optim* 2009;48(4):2085–2105.
5. Nikolaev A, Jacobson SH, McLay LA. A sequential stochastic security system design problem for aviation security. *Transport Sci* 2007;41(2):182–194.
6. Feng QM, Sahin H, Karson MJ. Bayesian analysis models for aviation baggage screening. *IIE Trans* 2009;41(11):995–1006.
7. Babu VLL, Batta R, Lin L. Passenger grouping under constant threat probability in an airport security system. *Eur J Oper Res* 2006;168:633–644.
8. Nie X, Batta R, Drury CG, *et al.* Passenger grouping with risk levels in an airport security system. *Eur J Oper Res* 2009;194(2):574–584.
9. Nie X, Batta R, Drury CG, *et al.* The impact of joint responses of devices in an airport security system. *Risk Anal* 2009;29(2):298–311.
10. Feng QM, Sahin H, Kapur KC. Designing airport checked-baggage-screening strategies considering system capability and reliability. *Reliab Eng Syst Saf* 2009;94(2):618–627.
11. Feng QM. On determining specifications and selections of alternative technologies for airport checked-baggage security screening. *Risk Anal* 2007;27(5):1299–1310.
12. Sahin H, Feng QM. A mathematical framework for sequential passenger and baggage screening to enhance aviation security. *Comput Ind Eng* 2009;57(1):148–155.
13. Pita J, Jain M, Ordenez F, *et al.* Using game theory for Los Angeles airport security. *AI Mag* 2009;30(1):43–57.
14. Lee AJ, Nikolaev AG, Jacobson SH. Protecting air transportation: a survey of operations

- research applications to aviation security. *J Transp Secur* 2008;1(3):160–184.
15. McLay LA, Jacobson SH, Kobza JE. Making skies safer. *OR/MS Today* 2005;32(5):24–31.
16. Wright PD, Liberatore MJ, Nydick RL. A survey of operations research models and applications in homeland security. *Interfaces* 2006;36(6):514–529.
17. Bureau of Transportation Statistics. Airline passenger opinions on security screening procedures. Omnibus household survey. 2004. Available at <http://www.bts.gov/publications>.
18. Gkritza K, Niemeier D, Mannering F. Airport security screening and changing passenger satisfaction: an exploratory assessment. *J Air Transport Manage* 2006;12(5):213–219.

FURTHER READING

Thomas AR, editor. Aviation security management: a three volume set. Westport (CT): Praeger Security International; 2008.

OPTION PRICING: THEORY AND NUMERICAL METHODS

VOLODYMYR BABICH
BARDIA KAMRAD
McDonough School of Business,
Georgetown University,
Washington, DC

A *derivative security* is a contract whose payoff depends on the stochastic price of another security, called an *underlying asset*. For example, consider a contract which gives you the right to sell a share of IBM stock in three months for \$100. This contract, called *European put option*, provides you with an insurance against IBM stock price dropping below \$100. In three months, if IBM stock price, $S_{\text{IBM}} < \$100$, you can buy IBM stock for S_{IBM} and *exercise* the option, selling the stock for \$100, and pocketing the difference $\$100 - S_{\text{IBM}}$. If IBM stock price is $S_{\text{IBM}} \geq \$100$, you allow the option to expire unexercised. What is the “fair” price for this put option today? In this article, we will describe the models, theory, and numerical methods for pricing this put option and derivative securities in general.

There is a cornucopia of derivative securities, varying by the underlying assets, by the form of the flexibility, by payoffs characteristics, and so on. One distinguishes between *financial derivative securities*, which are traded on financial markets (see websites for CME Group—www.cmegroup.com, and CBOE—www.cboe.com), and *real options*, which are just various forms of flexibility that decision makers have. The basic types of financial derivatives are forwards, futures, swaps, and financial options. In this article, we will use stock options to illustrate ideas underlying the theory and practice of option pricing. The same ideas apply to other financial derivative contracts and (to some extent) real options, but implementation details could differ. Recommendations on reading about those details are at the end of this article.

The two most common financial option contracts are *call* and *put* options (in common parlance, they are referred to as *plain vanilla* put and call options). The call (put) option gives the owner a right to buy (sell) the underlying asset at the specified time in the future at the specified price, called the *exercise price*. In the example above, the *time to expiration* is three months, the exercise price is \$100, and the underlying asset is the IBM stock. Call and put options can be *American* or *European*. These are simply monikers. They do not reflect option geography. Similarly Russian, Israeli, Bermudan, and Asian options are not traded exclusively in Russia, Israel, Bermuda, and Asia. American options are allowed to be exercised at any time prior to the expiration time. European options can be exercised only at the expiration time.

In this article we will use the following notation: $S(t)$ - stock price at time t , T - time to maturity (time to expiration) of the option, K - strike price, $\Pi(S, t)$ - payoff of the option when it is exercised at time t and the stock price is S (for a put option $\Pi(S, t) = (K - S)^+$, where $(x)^+ = \max(x, 0)$; for a call option $\Pi(S, t) = (S - K)^+$), r - risk-free rate, μ - expected stock return, δ - dividend yield on the stock (continuous), σ - volatility (standard deviation) of stock returns, and $C(S, t)$ - value of the derivative security at time t , when the stock price is S .

In the next section we describe commonly used models for stock prices. Then we present no-arbitrage derivatives pricing and illustrate how no-arbitrage arguments work in case of binomial and Black–Scholes and Merton models. After that, we discuss numerical methods for derivative pricing: binomial and trinomial lattice, simulation, finite difference, and other methods. We conclude with suggestions for further reading.

STOCHASTIC PROCESSES AND STOCK PRICE MODELS

Derivative securities derive their value from the underlying asset price. The common

approach is to assume that the underlying asset price follows a stochastic process, with the dual attributes of “relevance” and “tractability.” In particular, the choice of a specific stochastic process must be plausible, it should capture salient statistical properties of the stock price, and it should facilitate the development of option pricing algorithms. For example, for stock prices plausibility implies nonnegative prices. A salient statistical features of stock prices is the *random walk* property, described as the statistical independence between successive stock returns. A consequence of the random walk property is the statement that, given the current stock prices, the future stock prices do not depend on the past prices, which is known as the *weak form market efficiency* hypothesis in finance, and *Markov property* in operations research (OR) and math (see the section titled “Diffusion Processes and Random Walks” in this encyclopedia). The following are several examples of Markov processes used to model stock prices.

Random Walks as Binomial and Trinomial Trees

Binomial and trinomial trees are examples of discrete-time Markov processes. Consider an equi-width partition $\{0 = t_0 < t_1 < \dots < t_N = T\}$ of time interval $[0, T]$ into periods, so that the length of each period is $\Delta t = t_{i+1} - t_i$, where t_i is an observation time, $i = 0, 1, \dots, N - 1$. Denote by $S_i = S(t_i)$ the stock price at time t_i , $i = 0, \dots, N$.

According to the binomial tree model, each period the stock price can go up with probability p or down with probability $1 - p$, that is for $i = 0, \dots, N - 1$

$$S_{i+1} = S_i Z_i, \text{ where } Z_i = \begin{cases} u & \text{with probability } p, \\ d & \text{with probability } 1 - p. \end{cases} \quad (1)$$

Random variables $\{Z_i\}_{i=0}^{N-1}$ are independent and they represent shocks to stock prices every period (stock price $S_i = s$ either jumps up to $S_{i+1} = us$ or down to $S_{i+1} = ds$). We assume that $u > d$. An example of a two-step binomial tree is given in the Fig. 1a.

If the current stock price is S_0 , after N steps along a binomial tree, the distribution of the stock price, S_N , is binomial. Specifically, random variable S_N takes value $S_0 u^j d^{N-j}$ with probability $\binom{N}{j} p^j (1-p)^{N-j}$ for $j = 0, 1, \dots, N$. When implementing binomial models, parameters u , d , and p are chosen so that the mean and variance of Z_i are equal to those of the stock return, over time step Δt . Thus, one has to solve a system of two equations and three unknowns, which has an infinite number of solutions. Cox *et al.* [1] were the first to use binomial trees for option pricing by imposing an additional constraint: $d = 1/u$. With this constraint, for a stock with the standard deviation of returns $\sigma\sqrt{\Delta}$, dividend yield $\delta\Delta t$, and mean return $\mu\Delta t$ over time Δt , binomial tree model

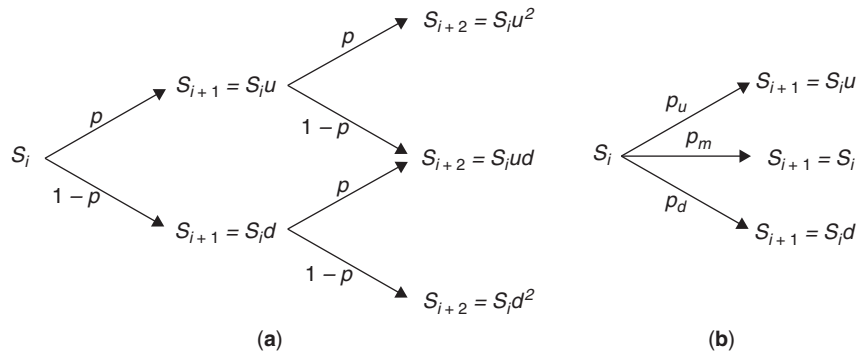


Figure 1. Binomial tree, two-step (left panel). Trinomial tree, one step (right panel).

parameters are

$$u = e^{\sigma\sqrt{\Delta t}}, d = e^{-\sigma\sqrt{\Delta t}}, \text{ and } p = \frac{e^{(\mu-\delta)\Delta t} - d}{u - d}. \quad (2)$$

The binomial tree nodes (price states) always recombine because the volatility is assumed homoscedastic (constant): that is, an up-state followed by a down-state will be in the same state (value) as a down-state followed by an up-state.

Kamrad and Ritchken [2] developed a trinomial lattice model accommodating a no-price-change as a possible state during a period of length Δt . That is for $i = 0, \dots, N - 1$,

$$S_{i+1} = S_i Z_i, \text{ where } Z_i = \begin{cases} u & \text{with probability } p_u, \\ 1 & \text{with probability } p_m, \\ d & \text{with probability } p_d. \end{cases} \quad (3)$$

Random variables $\{Z_i\}_{i=0}^N$ are independent and $p_u + p_m + p_d = 1$. Graphical representation of the trinomial model is given in the Fig. 1b. Similar to the binomial model, for a trinomial tree, parameters u , d , p_u , p_m , and p_d are chosen in order to match the first two moments of the actual stock returns. An extra degree of flexibility (compared to the binomial tree model, there is one extra variable in the trinomial tree model) affords more control over lattice implementation (i.e., jump sizes and probabilities, see the section titled “Numerical Methods in Derivatives Pricing.”) Kamrad and Ritchken [2] choose values for up and down jumps to be $u = 1/d = e^{\lambda\sigma\sqrt{\Delta t}}$ with the stretch parameter $\lambda \geq 1$ and the probabilities:

$$p_u = \frac{1}{2\lambda^2} + \frac{\left(\mu - \delta - \frac{\sigma^2}{2}\right)\sqrt{\Delta t}}{2\lambda\sigma}; \quad p_m = 1 - \frac{1}{\lambda^2}; \\ p_d = \frac{1}{2\lambda^2} - \frac{\left(\mu - \delta - \frac{\sigma^2}{2}\right)\sqrt{\Delta t}}{2\lambda\sigma}. \quad (4)$$

Geometric Brownian Motion

One of the most commonly used model for the evolution of the stock prices is the geometric Brownian motion (GBM) model, which is a continuous-time Markov process. The GBM model for stock prices is expressed as the following *stochastic differential equation*:

$$dS_t = (\mu - \delta)S_t dt + \sigma S_t dW_t, \quad (5)$$

where dW_t is an increment to the Wiener process, representing random shocks to stock returns.

The *Wiener process*, W_t , is a continuous-time stochastic process with satisfies the following properties: (i) $W_0 = 0$, (ii) independent increments, that is, for any $a < b \leq c < d$, $W(b) - W(a)$ and $W(d) - W(c)$ are independent, (iii) normality, that is, for any $b > a$, $W(b) - W(a)$ has a conditional (on $W(a)$) normal distribution with mean zero and standard deviation $\sqrt{b - a}$, (iv) path continuity, that is, any trajectory of W_t is continuous. Therefore, $dW(t)$ is a normal deviate with mean 0 and variance dt .

Informally, Equation (5) means that the percentage change in the stock price $\Delta S/S$ is due to the deterministic drift $(\mu - \delta)\Delta t$ and a random shock $\sigma\Delta W$. The rigorous definition and the solution of the stochastic differential equation (5) require tools of stochastic calculus [3–6]. Here, we simply present the solution (in the form similar to Equations (1) and (3)):

$$S_t = S_0 Z_t = S_0 e^{(\mu - \delta - \sigma^2/2)t + \sigma W_t}. \quad (6)$$

Therefore, for the GBM models of stock prices, for any t , random variable $\ln[S_t/S_0]$ is normally distributed with mean $(\mu - \delta - \sigma^2/2)t$ and standard deviation $\sigma\sqrt{t}$ or, in other words, S_t/S_0 follows the *log-normal distribution*. The binomial and the log-normal models of the stock prices are closely related. One can show that, as the number of subintervals N increases, the binomial model converges to the log-normal model [4, Chapter 3.2.7].

Other Models for Stock Prices

Although the GBM model is widely used in finance because of its mathematical properties, there is empirical evidence disproving its assumptions. Specifically, the GBM process (Eq. 5) is path continuous, but the movements of real stock prices are demonstratively discontinuous (see Cont and Tankov [7], Chapter 1). Next, the volatility parameter, σ , for real stock prices is not constant (e.g., French *et al.* [8]). Furthermore, the real stock returns may exhibit serial correlation, contrary to the independent increment property of the GBM process (see review by Richardson and Smith [9]). Finally, for real stock prices, $\ln[S_t/S_0]$ is not normally distributed because extreme movements of the stock prices happen much more frequently in practice than normal distribution allows (e.g., Jansen and de Vries [10]). To address the shortfalls of the GBM model, a number of alternative models have been considered, such as the constant elasticity of variance (CEV), stochastic volatility (SV), jump-diffusion, and Lévy processes. For discussion on alternative models for stock prices see Hull [11]. For theory and applications of Lévy processes Cont and Tankov [7]. As OR and management science academics and practitioners know, any model is just an approximation of reality and there is a constant tension between a model's precision and our ability to use the model for analysis. Therefore, it is likely that a variety of models, including the GBM model, will continue to be used to model stock prices.

NO ARBITRAGE DERIVATIVES PRICING

An *arbitrage opportunity* refers to a condition within financial markets such that a portfolio of traded securities exists, with the properties that (i) it can never generate negative cash flows to its holder, and (ii) it can generate positive cash flows with positive probability. In financial markets, arbitrage opportunities are short-lived in that supply and demand forces quickly dissolve price imbalances leading to such costless, riskless opportunities.

By disallowing arbitrage opportunities, we can limit the range of values that option prices can take. With additional

assumptions, we can derive a unique *no arbitrage* price. Typical assumptions made in no arbitrage analysis are the following: (i) Markets for stocks, bonds, and options are frictionless (no transaction costs; no bid-ask spreads; no restrictions on short sales; no taxes; shares of the stock are infinitely divisible); (ii) Markets are liquid (any position in financial securities is possible at any time); (iii) There are no arbitrage opportunities in the markets.

A typical derivation of no-arbitrage bounds on option prices proceeds as follows. One constructs a (replicating) portfolio comprising a short position in a risk-free asset such as T-bills and a long position in a number of shares of the stock. This portfolio is constructed in such way that all the cash inflow/outflow from this portfolio perfectly duplicates the option's cash flows during the maturity interval. Thus, the portfolio has the same exposure to the stock as does the option. Therefore, to avoid *arbitrage*, the option and the portfolio must command the same value.

No-arbitrage conditions guarantee the existence of a no-arbitrage option price. In addition, for uniqueness, we say that the market is *complete* when any derivative security's cash flows can be perfectly replicated through a combination of other existing traded assets at every point in time and for all possible states of nature. For complete markets, the no-arbitrage option price is unique and, in complete markets, options are redundant—it is always possible to construct replicating portfolios instead. In practice, however, options create tremendous economic value in terms of trading and financial market activity. We will demonstrate no-arbitrage pricing arguments on the binomial model and present the results of no-arbitrage analysis in the form of Black–Scholes and Merton option pricing formulas.

Binomial Model

Consider an economy with two assets: the stock and the bond. Suppose we would like to price a derivative security on the stock. Assume that stock price follows a one-step (for simplicity) binomial model (1) and bonds earn constant return r , continuously compounded (i.e., if the initial bond price is B_0 , at

time t this price is $B_t = B_0 e^{rt}$. Furthermore, for simplicity, the dividend yield on the stock is $\delta = 0$. One can prove that no-arbitrage condition in this economy is equivalent to $d < e^{rT} < u$.

To price the derivative security, first consider its time- T cash flows:

$$C(S(T), T) = \begin{cases} \Pi_u \stackrel{\text{def}}{=} \Pi(S_0 u, T) & \text{if } S(T) = S_0 u, \\ \Pi_d \stackrel{\text{def}}{=} \Pi(S_0 d, T) & \text{if } S(T) = S_0 d. \end{cases} \quad (7)$$

Create a portfolio by buying Δ shares of stock and investing b in bonds (note: Δ and b could be negative). This portfolio is replicating if the cash flows from this portfolio at time T (when portfolio is sold) match those from the derivative security:

$$\begin{cases} \Delta S_0 u + b e^{rT} = \Pi_u & \text{if } S(T) = S_0 u, \\ \Delta S_0 d + b e^{rT} = \Pi_d & \text{if } S(T) = S_0 d. \end{cases} \quad (8)$$

Because $u > d$, this system of equations has a unique solution (thus, our economy is complete):

$$\Delta = \frac{\Pi_u - \Pi_d}{(u - d)S_0}; \quad b = -e^{-rT} \frac{u\Pi_d - d\Pi_u}{u - d}. \quad (9)$$

Using these values, we compute the value of the portfolio to be

$$S_0 \Delta + b = \frac{\Pi_u - \Pi_d}{u - d} - e^{-rT} \frac{u\Pi_d - d\Pi_u}{u - d}. \quad (10)$$

To avoid arbitrage, this must be the price of the derivative security $C(S_0, 0)$ at time 0 (if this were not the case, e.g., the claim's price were lower, we would buy the claim and short the replicating portfolio, guaranteeing positive cash flow at time 0 and net zero cash flow at time T). Interestingly, after some algebra, Equation (10) can be written as

$$C(S_0, 0) = e^{-rT} \left[\frac{e^{rT} - d}{u - d} \Pi_u + \frac{u - e^{rT}}{u - d} \Pi_d \right]. \quad (11)$$

Define $q \stackrel{\text{def}}{=} \frac{e^{rT} - d}{u - d}$. Then, Equation (11) is the same as $C(S_0, 0) = e^{-rT} [q\Pi_u + (1 - q)\Pi_d]$. From no-arbitrage condition, $d < e^{rT} < u$, $q \in (0, 1)$. We will treat q as a probability, called the *risk-neutral probability*. Thus, $C(S_0, 0) = e^{-rT} E_q[\Pi(S(T), T)]$, where E_q is the expectation with respect to the risk-neutral probability.

The last observation is very important. It is an example of so-called *risk-neutral pricing*. Generally, no-arbitrage assumption implies the existence of a risk-neutral probability. Market completeness implies that this probability is unique. The “ $E = mc^2$ ” statement in financial asset pricing is that, in theory, any claim in the economy can be priced as $C(0) = E_q[e^{-rT} \Pi(T)]$, as long as one can find the appropriate risk-neutral probability measure. The challenge of asset pricing in practice is to find that probability measure.

For a binomial model with the continuous dividend yield, δ , the risk-neutral probability is

$$q = \frac{e^{(r-\delta)T} - d}{u - d}. \quad (12)$$

Note that the pricing formula (11) does not contain terms with the stock's expected return μ , the actual probability of stock going up, p , or any reference to the investor's risk preferences. It is as if all investors were risk-neutral and all assets were earning return of r . This explains the term *risk-neutral pricing*. The next section contains another illustration of risk-neutral pricing.

The Black–Scholes (1973) and Merton (1973) Option Pricing Model

Black and Scholes [12] and Merton [13], independently arrived at the same pricing formula that, in light of the assumptions made, governs the value of a European option. Our objective here is to provide an insight to and an intuitive understanding of the fundamental features of their model, absent mathematical rigor (a rigorous treatment can be found in many books [4,6]). We note that the B-S & M (1973) model can be derived in a number of ways including construction of a (self-financing, replicating) portfolio consisting of T-bills and shares of the stock, similar

to what was done for the binomial model above.

The set of assumptions required to derive the B-S & M (1973) model include the following: (i) Markets for stocks, bonds, and options are frictionless (no transaction costs; no restrictions on short sales; no taxes; shares of the stock are infinitely divisible). (ii) Risk-free rate of return, r is constant during the option's maturity $[0, T]$. (iii) Stock price process is given by Equation (5) with parameters μ , δ , and σ constant over the maturity interval $[0, T]$. We will demonstrate the model in the case of plain vanilla European options.

The analysis of the instantaneous change in the derivative security value ($C(S, t)$) over an instant of time results in a second-order partial differential equation (PDE), that is to be solved subject to the claim's terminal condition. This PDE and its solution is the B-S & M (1973) valuation model. The Black-Scholes PDE is (for $t \in [0, T]$ and $0 \leq S < \infty$)

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + (r - \delta)S \frac{\partial C}{\partial S} - rC = 0. \quad (13)$$

Any European derivative on the stock satisfies Equation (13). However, claims differ in their terminal conditions. For example, for a European call option, the terminal condition is $C(S, T) = (S - K)^+$. For a European put option, the terminal condition is $C(S, T) = (K - S)^+$.

For a European call option, the solution of the B-S & M (1973) equation is

$$C(S, 0) = C_0 = Se^{-\delta T}F(d_1) - Ke^{-rT}F(d_2), \quad \text{where} \quad (14)$$

$$d_1 = \frac{\ln\left(\frac{S}{K}\right) + (r - \delta + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}};$$

$$d_2 = \frac{\ln\left(\frac{S}{K}\right) + (r - \delta - \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}. \quad (15)$$

In the above equations, C_0 is the B-S & M (1973) call value at time zero, and $F(\cdot)$ is

the cumulative standard normal probability. We are now in a position to provide some intuition to the model

1. Equation (13) is commonly referred to as the *fundamental equation*. A striking aspect of this PDE is the second-order effect: reflecting the rate of rate of change in the call relative to the stock and is proportional to the volatility of the stock (the partials with respect to the stock price and time are to be expected. The second-order effect results from stochastic calculus.)
2. In describing Equations (13) and (14), it is important to note the following:
 - a. The average growth rate of the stock, μ , does not appear in the model: the option's value is independent of μ regardless of the investors' risk preference. (In the derivation of the PDE (13) the term with μ gets eliminated.) This risk "preference" independence feature of the B-S & M (1973) model is a consequence of the risk-neutral valuation.
 - b. To obtain results, the B-S & M (1973) model requires five input parameters as given. These are initial stock price, $S(0)$; exercise price, K ; maturity time, T ; risk-free rate, r ; and volatility (standard deviation of the stock price return), σ .
 - c. In Equation (14), the constant dividend yield is taken (in effect discounted) out of the stock price. This is so because any dividend paid out accrues to the owner of the stock and not to the owner of the option. This adjustment is necessary to avoid arbitrage opportunities.
 - d. $F(d_2)$ is the probability that at its maturity the call option is exercised. That is, $F(d_2) = \Pr[S(T) > K]$. More on this below.
 - e. $F(d_1)$ is also a probability, as we will discuss below. It is commonly referred to as the *perfect hedge ratio* or the *option delta*: $F(d_1) = \frac{\partial C}{\partial S} = \Delta$. Option delta indicates the call's sensitivity to the stock price. A \$1

increase in the stock price will lead to a $\$ \Delta$ increase in the call value.

3. Stocks pay dollar dividends and not dividend yield.

B-S & M (1973) formula (14) can also be derived by simply computing the expectation in $C(S, 0) = E_q[e^{-rT}\Pi(S(T), T)]$, where E_q is the expectation with respect to the risk-neutral probability. Under the risk-neutral probability, the evolution of the stock price is given by the GBM model (5) with $\mu = r$. It is instructive to go through a few steps of this approach.

$$\begin{aligned} C(S, 0) &= E_q[e^{-rT}\Pi(S(T), T)] \\ &= E_q[e^{-rT}(S(T) - K)^+] \\ &= E_q[e^{-rT}(S(T) - K)1_{\{S(T) > K\}}] \\ &= E_q[e^{-rT}S(T)1_{\{S(T) > K\}}] \\ &\quad - E_q[e^{-rT}K1_{\{S(T) > K\}}] \\ &= E_q[e^{-rT}S(T)1_{\{S(T) > K\}}] \\ &\quad - e^{-rT}K \Pr[S(T) > K]. \end{aligned} \quad (16)$$

The last equality is based on $E_q[1_{\{S(T) > K\}}] = \Pr_q[S(T) > K]$ where $\Pr_q[\cdot]$ is the probability with respect to the risk-neutral measure. It is easy to compute $\Pr_q[S(T) > K]$ and verify that $\Pr_q[S(T) > K] = F(d_2)$.

It would be convenient to convert the expectation, $E_q[e^{-rT}S(T)1_{\{S(T) > K\}}]$, into an expression with probabilities. This transformation is possible by the change of *numeraire* technique. This technical step is frequently very useful in proofs of pricing formulas for many types of derivative securities, particularly securities on multiple assets, foreign exchange rates, and interest rates. Mathematical details can be found in a number of textbooks (e.g., Bjork [4] and Shreve [6]) and in Geman *et al.* [14]. Applications to option pricing of the change of numeraire technique are discussed in Schroder [15]. At a high level, the idea of the change of numeraire is as follows. Let $B(t) = e^{rt}$ denote the value of the money market account at time t (with $B(0) = 1$). Then pricing formula $C(S, 0) = E_q[e^{-rT}\Pi(S(T), T)] =$

$B(0)E_q[\Pi(S(T), T)/B(T)]$ compares the value of the option payoff with the value of money market account, which thus serves as a numeraire asset. But one does not have to use the money market account as a numeraire. Any positively valued asset $\beta(t)$ will work, and a general pricing formula is $C(S, 0) = \beta(0)E_\beta[\Pi(S(T), T)/\beta(T)]$, where E_β is the expectation with respect to the probability measure corresponding to the numeraire asset β (one can compute the Radon–Nikodym derivative of the new measure relative to the risk-neutral measure, but typically one simply uses the fact that the process $S(t)/\beta(t)$ is a martingale under the new measure [4,6,14]). For Equation (16), using this general pricing formula, we can write $E_q[S(T)/B(T)1_{\{S(T) > K\}}] = S(0)/B(0)E_S[1_{\{S(T) > K\}}] = S(0)\Pr_S[S(T) > K]$, where E_S ($\Pr_S$) is the expectation (probability) with the process $S(t)/B(t)$ as the numeraire. Further computations result in $\Pr_S[S(T) > K] = e^{-\delta T}F(d_1)$.

For a European put option, the solution of the B-S & M (1973) equation is

$$P(S, 0) = P_0 = Ke^{-rT}F(-d_2) - Se^{-\delta T}F(-d_1), \quad (17)$$

where d_1 and d_2 are given by Equation (15). The delta of put option is $\Delta = \frac{\partial P}{\partial S} = -F(d_1) < 0$. That is, as the stock price increases, the value of the put option decreases. $F(-d_2)$ is the put exercise probability.

American Options

The plain vanilla stock options traded on exchanges are American options. Recall that an American option can be exercised at any time prior to the option's expiration. The extra flexibility makes American options more valuable than the corresponding European options, but complicates the valuation process. Typically, there are no closed-form expressions, similar to Equations (14) and (17), available for American options. An exception to this is an American call option written on a nondividend-paying stock. One can show that it is not optimal to exercise this option prior to expiration. Therefore, the

additional flexibility of the American feature is worthless; the value of an American call option is equal to the value of the corresponding European call option, and is given by formula (14). Note that it can be optimal to exercise an American put option on a nondividend paying stock early and, therefore, we cannot use formula (17) for American put options.

One approach to valuation of American options is to identify a region $R \subset \mathbf{R}_+ \times \mathbf{R}_+$ in the space of stock prices and times (S, t) , where the exercise of the option is optimal. Inside of the exercise region (i.e., for $(S, t) \in R$), the value of the option is equal to the option's payoff ($C(S, t) = \Pi(S, t)$). Outside of the exercise region, the value of the option satisfies the B-S & M (1973) PDE (13). Because, when solving the PDE (13), we do not know in advance where the boundary (∂R) of the exercise region (R) lies, to value an American option we need to solve a *free boundary problem*. An additional condition used to describe the free boundary problem for American options is the *smooth pasting condition*, that is $\frac{\partial C(S, t)}{\partial S} = \frac{\partial \Pi(S, t)}{\partial S}$ for all $(S, t) \in \partial R$. The smooth pasting condition is a no-arbitrage condition—it is necessary to avoid arbitrage opportunities in the model. When an American option does not have an expiration time, the free boundary problem is simplified because the B-S & M (1973) PDE (13) becomes an ordinary differential equation (ODE):

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + (r - \delta)S \frac{\partial C}{\partial S} - rC = 0, \quad (18)$$

yielding closed-form expressions for option prices (see Dixit and Pindyck [16]). However, when an option has a finite life T , the option value can be found only numerically (see discussion of numerical methods in the section titled “Numerical Methods in Derivatives Pricing”).

Another approach to valuation of American options is to combine the optimal stopping problem with the risk-neutral pricing. Mathematically, the exercise time of an American option is a *stopping time* (for definition and properties of stopping times, see Oksendal [3]). If $\Theta[t, T]$ is a set

of all stopping times $\theta \in [t, T]$, then the value of an American option at time t is found by solving an optimization problem $C(S(t), t) = \sup_{\theta \in \Theta[t, T]} E_q [e^{-r(\theta-t)} \Pi(S(\theta), \theta)]$, where, as before, E_q is the expectation with respect to the risk-neutral probability. The solution of this optimization problem requires tools beyond the scope of this introductory article (please, see Oksendal [3], Shreve [4], and the section titled “Martingales” in this encyclopedia for the discussion of the solution to the optimal stopping problem, *martingales*, and *supermartingales*). We would like to mention, in passing, that one can solve the optimal stopping problem (and, hence, value American options) using variational inequalities (see **Variational Inequalities** and Oksendal [3]). For further discussion of variational inequalities and other approaches, see a review article by Broadie and Detemple [17].

NUMERICAL METHODS IN DERIVATIVES PRICING

For some options, such as European put and call options, there exist closed-form pricing formulas (Black–Scholes formulas (14) and (17)). However, for the majority of derivative securities, both European and American types, valuation can be done only numerically. In this section we will discuss several numerical methods commonly used for option pricing: binomial and trinomial trees, simulation, and finite difference methods.

LATTICE MODELS

Binomial lattice (binomial tree) model of Cox *et al.* [1] is one of the most commonly used models. Its main strengths are simplicity and flexibility. It can be easily understood and programmed and it is difficult to make a mistake implementing it. Lattice models can be used to price practically any type of option and, compared to other models, they are particularly easy to use for pricing American options. The drawback of lattice models as numerical algorithms is that they tend to be slow, particularly for options which depend on several underlying assets. With lattice methods, the objective is to approximate the

stochastic evolution of the underlying asset over the maturity interval $[0, T]$. The lattice generates possible states of price realization in discrete-time steps with corresponding probabilities (see the section titled “Random Walks as Binomial and Trinomial Trees”). The state space and the probabilities are constructed in a manner to ensure convergence to the continuous-time stochastic process and eliminate arbitrage opportunities (see the section titled “Binomial Model”). With the lattice, the option’s terminal condition is superimposed on the final set of approximating stock price states and terminal option values are approximated. A backward stochastic *dynamic programming* recursion is then used to obtain the time-zero option value (for further discussion of dynamic programming, see the section titled “Markov Decision Processes” in this encyclopedia). If the option is American, then for every node on the tree, we simply compare the value computed by backward recursion and the exercise value of the option and select the larger of the two. The trinomial tree method is applied similarly.

Simulation

Monte Carlo simulation methods (see the section titled “Monte–Carlo Method” and “Markov Chain Monte–Carlo” in this encyclopedia) use the representation of option prices as the expected discounted option payoffs (with respect to the risk-neutral measure; see the section titled “No Arbitrage Derivatives Pricing”), that is, $C(S, 0) = E_q[e^{-rT}\Pi(S(T), T)]$. The main idea of simulation methods is to approximate the expectation by the sample mean of the sample generated from the random variable $e^{-rT}\Pi(S(T), T)$.

For example, to price the put option introduced at the beginning of this article, we (i) generate a sample of IBM stock prices at time $T = 3/12$: $\{S_1(T), S_2(T), \dots, S_N(T)\}$; (ii) compute discounted option payoff at each value of the IBM stock price: $\{\pi_1 = e^{-rT}\Pi(S_1(T), T), \pi_2 = e^{-rT}\Pi(S_2(T), T), \dots, \pi_N = e^{-rT}\Pi(S_N(T), T)\}$, where $\Pi(S, t) = (100 - S)^+$; and (iii) use sample mean as an approximation of the expectation: $p = \sum_{i=1}^N \pi_i / N$.

By the central limit theorem, this approximation converges to the true value of the expectation with the rate of $O(1/\sqrt{N})$. The convergence rate remains the same even if the option payoffs depend on the values of more than one underlying asset: $\Pi[S^1(T), S^2(T), \dots, S^M(T), T]$. As the dimension of the problem, M , increases, the computational effort for generating additional random variables increases more slowly than the computational effort for computing M -dimensional integrals using other numerical methods. Therefore, for high-dimensional problems, Monte Carlo simulation methods outperform other numerical methods.

To generate a sample $\{S_1(T), S_2(T), \dots, S_N(T)\}$ of the IBM stock prices, we invoke one of the stochastic models for stock prices (see the section titled “Stochastic Processes and Stock Price Models”). For example, for the GBM model, we generate a sample of independent standard normal random numbers $\{z_1, \dots, z_N\}$ and, using Equation (6), construct a sample of IBM stock prices as follows: $\{S_i(T) = S_0 e^{(r-\delta-\sigma^2/2)T + \sigma\sqrt{T}z_i}\}_{i=1}^N$. Note that, in Equation (6) the expected stock return μ was replaced by the risk-free rate, r , because the stock price process is considered under the risk-neutral measure.

Monte Carlo simulation is a very flexible method that can be used to price many types of options. In particular, Monte Carlo simulation works well for path-dependent options, such as Asian, barrier, and lookback options (see Boyle *et al.* [18], and Moel and Tufano [19]). Path-dependent options are examples of so called *exotic* options. Recall that the payoff for the European put option discussed at the beginning of this article is $[K - S(T)]^+$, where K is the strike price and $S(T)$ is the stock price at the expiration time. In contrast, the payoff of an Asian put option is $[K - A(T)]^+$, where $A(T)$ is the average price of the stock during time $[0, T]$. The payoff of a barrier option depends on the event of the stock price reaching a barrier during the life of the option. An *up-and-out put* option, for instance, has the same payoff as a regular European put option, $[K - S(T)]^+$, unless the stock price rises above the specified barrier, b , during time $[0, T]$, in which case the option becomes worthless. The payoff of a lookback

option depends on the highest or the lowest stock price during time $[0, T]$, and not just the price $S(T)$.

Monte Carlo simulation also works well for pricing derivatives that depend on several underlying assets, such as spread options, rainbow options, and outperformance options (e.g., Carmona and Durrleman [20]). For example, a rainbow option payoff is equal to $[S_1(T) + S_2(T) + \dots + S_L(T) - K]^+$, where $\{S_l\}_{l=1}^L$ are underlying assets.

The standard error of Monte Carlo simulation is $\sigma_p/\sqrt{N}$, where σ_p is the sample standard deviation of the sample of option payoffs, and N is the sample size. The standard error is reduced if the sample size, N , increases, but only at the rate of $\sqrt{N}$. Alternatively, one can reduce σ_p part of the standard error by applying *variance reduction techniques*, such as *control variates*, *moments matching*, *antithetic variables*, *stratified sampling*, and *importance sampling* (see the section titled “Variance Reduction Techniques” in this encyclopedia). The implementation of the variance reduction techniques is relatively simple and the computational improvements are significant. Therefore, in practice some form of the variance reduction technique is usually applied when using Monte Carlo simulation.

Monte Carlo simulation methods become computationally expensive when applied to pricing of American options. As discussed

earlier, American options can be exercised at any time during the life of the option. The interaction between the option exercise optimization problem (given the simulated asset prices) and the simulation of the option payoff (for the optimal option exercise) leads to considerable computational difficulties. Therefore, typically, other numerical methods are used for American option pricing (although simulation-based algorithms have been proposed, e.g., Longstaff and Schwartz [21]). For further discussion of the Monte Carlo simulation see Glasserman [22].

Finite Difference Methods

To solve Black–Scholes PDE (13) using a finite difference method, we begin by discretizing the time and stock price spaces. Specifically, consider time points $t = j\Delta t$, for $j = 0, 1, 2, \dots, J$ and stock values $S = i\Delta S$, for $i = 0, 1, 2, \dots, I$. These points make up a grid where Δt and ΔS are time and price steps and $J\Delta t = T$, see Fig. 2. The option value at any grid point $(S, t) = (i\Delta S, j\Delta t)$ is denoted by $C(i\Delta S, j\Delta t) = C_i^j$.

Next, we replace partial derivatives in PDE (13) by their discrete approximations (called *finite difference approximations*). There are a number of finite difference approximations. For example, $\frac{\partial C}{\partial t}(i\Delta S, j\Delta t) \approx \frac{C_i^j - C_i^{j-1}}{\Delta t}$, $\frac{\partial C}{\partial t}(i\Delta S, j\Delta t) \approx \frac{C_i^{j+1} - C_i^j}{\Delta t}$, and $\frac{\partial C}{\partial t}(i\Delta S,$

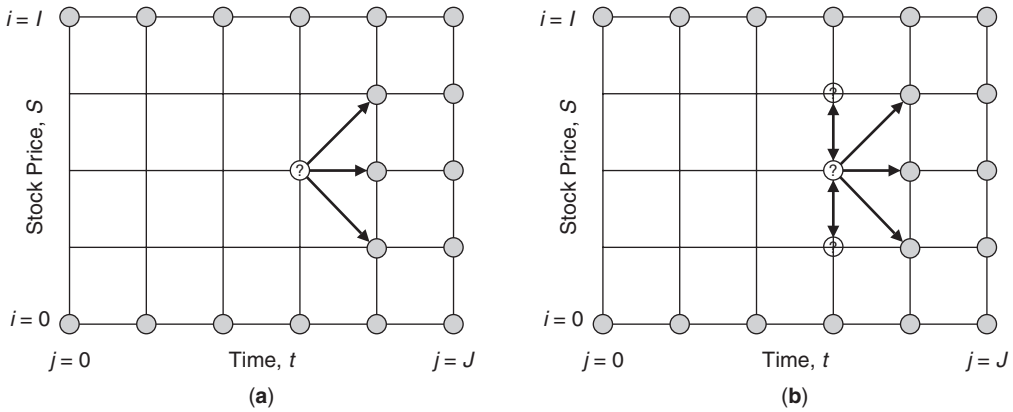


Figure 2. Graphical representation of an explicit (a) and an implicit (b) difference methods. Filled circles represent grid nodes where option values have been already computed. Circles with a question mark represent grid nodes where option values are being evaluated.

$j\Delta t) \approx \frac{C_i^{j+1} - C_i^{j-1}}{2\Delta t}$ are called *backward*, *forward*, and *central difference approximations* for $\frac{\partial C}{\partial t}$, respectively. The finite difference approximations of $\frac{\partial C}{\partial S}$ are defined similarly.

The second-order partial derivative is usually approximated as $\frac{\partial^2 C}{\partial S^2}(i\Delta S, j\Delta t) \approx \frac{C_{i+1}^j - 2C_i^j + C_{i-1}^j}{\Delta S^2}$. The forward, backward, and central differences possess different numerical properties (e.g., errors using forward and backward differences are of $O(\Delta S)$ and $O(\Delta t)$, whereas the error using central differences are of $O(\Delta S^2)$ and $O(\Delta t^2)$) and the properties of the finite difference method depend on which approximations are used.

We use option payoff, $\Pi(S, T)$, to specify option values at $j = J$. We also specify boundary conditions for $i = 0$ and $i = I$. Using backward finite difference in time and central difference in stock price and collecting terms C with same indices, PDE (13) is approximated by

$$C_i^{j-1} = a_i^j C_{i+1}^j + b_i^j C_i^j + c_i^j C_{i-1}^j, \quad (19)$$

where coefficients a_i^j , b_i^j , and c_i^j do not depend on the option prices, C . From conditions at expiration and boundary condition we know values C_i^J , for $i = 1, \dots, I$. Using Equation (19), we can compute values C_i^{J-1} for $i = 1, \dots, I$. Continuing this process recursively we derive C_i^0 for $i = 1, \dots, I$. Because option price C_i^{j-1} depends only on option prices which are already known: C_{i+1}^j , C_i^j , and C_{i-1}^j , the finite difference method corresponding to the difference equation (19) is called the *explicit difference method*. This explicit difference method is represented in Fig. 2a. Note that Equation (19) is equivalent to a relationship between option prices in a trinomial tree. Just as with the trinomial tree, if the option is American, then for every node, we simply choose the larger of (i) the option exercise value, and (ii) the value from Equation (19).

Using forward finite difference in time and central difference in stock price and collecting terms C with the same indices, PDE (13) is approximated by

$$d_i^j C_{i+1}^j + e_i^j C_i^j + f_i^j C_{i-1}^j = C_i^{j+1}. \quad (20)$$

This equation combines several unknown option prices (with superindex j) and is called an *implicit equation*. The corresponding implicit method is shown in Fig. 2b. By solving a system of linear equations of the form (20) together with boundary conditions and conditions at the expiration, one derives the option values at every point of the grid.

Implementation of implicit difference methods requires more programming sophistication than implementation of explicit methods. Fortunately, numerous software packages offer efficient routines for solving linear systems of equations. Implicit methods have one significant advantage over explicit methods—numerical stability (which roughly means the insusceptibility of an algorithm to accumulation of rounding errors). One can prove that an explicit method above is stable (and therefore generates an answer which is not dramatically affected by rounding errors) only if $\Delta t < A\Delta S^2$, where constant A can be computed. This constraint on the time step relative to the price step is a significant hindrance in practice, because small Δt translates into dramatic increase in the number of computations one needs to perform to find the option price. Implicit methods are numerically stable without any restrictions on the size of the time step. For additional discussion on the use of finite difference methods for option pricing, see Wilmott [23, Chapters 63–65].

Other Methods

Besides finite difference methods, there are many methods for solving PDEs [24]. Many of those approaches, specifically Laplace and Fourier transforms and expansions in Fourier and Taylor series, have been used for option pricing [25–28].

FURTHER READING

Quantitative finance, and derivative pricing in particular, is a vibrant research and applications area. New approaches are discovered and new models are introduced almost as quickly as the new types of derivative securities are invented. In this article we touched on only a small subset of important

ideas. Fortunately, there is a multitude of websites, articles, and books available to help those who are interested in derivatives. Comprehensive and accessible introductions to derivatives and derivative pricing can be found in Broadie and Detemple [17], Wilmott [23], Ritchken [29], Briys *et al.* [30], Wilmott *et al.* [31], and Hull [11]. Discrete-time arbitrage pricing is presented in Pliska [32] and Shreve [33]. Stochastic calculus and its applications in continuous-time finance are described in Oksendal [3], Shreve [4], Karatzas and Shreve [5], Björk [6], and Steele [34]. Commodity derivatives are discussed in Geman [35]. For models of interest rates and pricing of interest rate derivatives, see Jarrow [36]. Credit risk, credit derivatives, and credit derivatives pricing are presented in Schönbucher [37]. Real options are discussed in Dixit and Pindyck [16], Trigeorgis [38], Grenadier [39], Schwartz and Trigeorgis [40] and in the section titled “Value of Information and Option Values” in this encyclopedia. For a review of empirical research of stock and option price processes, see Bates [41].

REFERENCES

1. Cox JC, Ross SA, Rubenstein M. Option pricing: a simplified approach. *J Financ Econ* 1979;7:229–263.
2. Kamrad B, Ritchken P. Multinomial approximating models for options with k state variables. *Manage Sci* 1991;37(10):1640–1652.
3. Oksendal B. Stochastic differential equations: an introduction with applications. Berlin, New York: Springer; 2007.
4. Shreve SE. Stochastic calculus for finance II: continuous-time models. New York: Springer; 2008.
5. Karatzas I, Shreve SE. Brownian motion and stochastic calculus. New York: Springer; 2008.
6. Björk T. Arbitrage theory in continuous time. 2nd ed. Oxford, New York: Oxford University Press; 2004.
7. Cont R, Tankov P. Financial modeling with jump processes. Boca Raton (FL); Chapman and Hall/CRC; 2004.
8. French KR, Schwert GW, Stambaugh RF. Expected stock returns and volatility. *J Financ Econ* 1987;19:2–29.
9. Richardson M, Smith T. A unified approach to testing for serial correlation in stock returns. *J Bus* 1994;67(3):371–399.
10. Jansen DW, de Vries CG. On the frequency of large stock returns: putting booms and busts into perspective. *Rev Econ Stat* 1991;73(1): 18–24.
11. Hull J. Options, futures, and other derivatives. 9th ed. Upper Saddle River (NJ): Prentice Hall; 2009.
12. Black F, Scholes M. The pricing of options and corporate liabilities. *J Polit Econ* 1973;81: 637–654.
13. Merton RC. Theory of rational option pricing. *Bell J Econ Manage Sci* 1973;4(1):141–183.
14. Geman H, Karoui NE, Rochet J-C. Changes of numeraire, changes of probability measure and option pricing. *J Appl Probab* 1995;32(2): 443–458.
15. Schroder M. Changes of numeraire for pricing futures, forwards, and options. *Rev Financ Stud* 1999;12(5):1143–1163.
16. Dixit AK, Pindyck RS. Investment under uncertainty. Princeton (NJ): Princeton University Press; 1994.
17. Broadie M, Detemple JB. Option pricing: valuation models and applications. *Manage Sci* 2004;50(9):1145–1177.
18. Boyle P, Broadie M, Glasserman P. Monte Carlo methods for security pricing. *J Econ Dyn Control* 1997;21:1267–1321.
19. Moel A, Tufano P. Bidding for the Antamina mine: valuation and incentives in a real options context. In: Brennan MJ, Trigeorgis L, editors. Project flexibility, agency, and competition: new developments in the theory and application of real options. Oxford, New York: Oxford University Press; 2000. Chapter 8. pp. 128–150.
20. Carmona R, Durrleman V. Pricing and hedging spread options. *SIAM Rev* 2003;45(4): 627–685.
21. Longstaff FA, Schwartz ES. Valuing American options by simulation: a simple least-squares approach. *Rev Financ Stud* 2001;14(1): 113–147.
22. Glasserman P. Monte Carlo methods in financial engineering. New York: Springer; 2004.
23. Wilmott P. Paul Wilmott on quantitative finance. Chichester, West Sussex, England, New York: John Wiley & Sons, Inc.; 2000.
24. Zauderer E. Partial differential equations of applied mathematics. 3rd ed. New York: Wiley-Interscience; 2006.

25. Heston SL. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev Financ Stud* 1993;6(2):327–343.
26. Carr P, Madan D. Option pricing and the fast fourier transform. *J Comput Financ* 1999;2(4): 61–73.
27. Duffie D, Pan J, Singleton K. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* 2000;68(6):1343–1376.
28. Davydov D, Linetsky V. Pricing options on scalar diffusions: an eigenfunction expansion approach. *Oper Res* 2003;51(2):185–209.
29. Ritchken PH. Options: theory, strategy, and applications. Glenview (IL): Scott, Foresman and Company; 1987.
30. Briys E, Bellalah M, Mai HM, *et al.* Options, futures and exotic derivatives. Chichester, England, New York: John Wiley & Sons, Ltd; 1998.
31. Wilmott P, Howison S, Dewynne J. The mathematics of financial derivatives: a student introduction. Cambridge: Cambridge University Press; 1995.
32. Pliska SR. Introduction to mathematical finance: discrete time models. Oxford: Blackwell Publishers, Ltd.; 1997.
33. Shreve SE. Stochastic calculus for finance I: the binomial asset pricing model. New York: Springer; 2009.
34. Steele M. Stochastic calculus and financial applications. New York: Springer; 2003.
35. Geman H. Commodities and commodity derivatives: modelling and pricing for agriculturals, metals and energy. West Sussex, New York: Wiley; 2005.
36. Jarrow R. Modelling fixed income securities and interest rate options. Stanford (CA): Stanford Economics and Finance; 2002.
37. Schönbucher PJ. Credit derivatives pricing models: model, pricing and implementation. Chichester, England, Hoboken (NJ): Wiley; 2003.
38. Trigeorgis L. Real options: managerial flexibility and strategy in resource allocation. Cambridge (MA): MIT Press; 1996.
39. Grenadier SR. Game choices: the intersection of real options and game theory. London: Risk Books; 2000.
40. Schwartz ES, Trigeorgis L. Real options and investment under uncertainty: classical readings and recent contributions. Cambridge (MA): MIT Press; 2001.
41. Bates DS. Empirical option pricing: a retrospection. *J Econ* 2003;116:387–404.

OR MODELS IN FREIGHT RAILROAD INDUSTRY

ASHISH K. NEMANI

Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

RAVINDRA K. AHUJA

Innovative Scheduling, Inc.,
and Department of
Industrial and Systems
Engineering, University of
Florida, Gainesville,
Florida

Transportation services are the backbone of the economy and a vital parameter to help gauge the growth of a country. A streamlined flow of raw materials to production facilities and finished products to customers through different transportation modes (rail, road, air, and water) aids in market stability. Rail has always shared a major role in freight transport. After the deregulation act (Staggers Rail Act, 1980) and subsequent technological improvement, the US freight railroad's traffic volume has increased by almost 50%. On average, a typical Class I US railroad (BNSF, CN, CP, CSX, FFX, KCS, NS, TFM, and UP) receives about 4000–7000 new shipments (carloads) from its 5000 customers every day. These shipments are moved from some 1000 origins to about 2000 destinations. This steep rise in the demand coupled with higher customers' expectations has created the need for better resource planning in order to provide more efficient and cost-effective services.

The services can be improved by capacity expansion, which is capital extensive, and/or by optimizing the constituent operations, which we discuss in this article. A major railroad comprises more than 10,000 miles of tracks, with several thousand trains moving along them weekly. These trains include about 2000–3000 locomotives and approximately 80,000 railcars located in

about 200–300 rail yards, requiring more than 5000 crew members to operate. Because further expansion in the infrastructure would not prove cost effective, the second approach of implementing optimization-based rules of planning and scheduling is more practical. However, the railway system is by nature highly complex because of its interlinked subsystems (locomotives, railcars, crews, tracks, etc.). Each individual subsystem comprises very large scale optimization problems which render the standard solution techniques unusable. Most of the recent research papers on this topic deal with simplified models, or apply them to unrealistic or isolated instances. In addition, collecting and organizing the data required for implementing and installing software in the railroad industry is an arduous task. However, the situation is gradually improving with the development of effective and practical algorithms, and railroads have started implementing operations research (OR)-based decision tools because of the pressure of a competitive environment. Recently, Canadian Pacific Railways was chosen as the winner of the 2003 Franz Edelman Award for achievement in operations research and management sciences for their project titled, "Perfecting the Scheduled Railroad: Model-Driven Operating Plan Development" [1]. This project generated a cost savings of approximately US\$ 170 million between 1999 and 2000.

With the rapid development of other modes of transportation, the concept of "running a scheduled railroad" has gained increased importance. The traditional tonnage-based approach attempts to save costs by focusing on reducing the number of trains and creating longer trains, even if that causes delays or cancellations of scheduled trains. This differs from the modern "schedule-based approach," in which trains follow the operating plan even if they are not fully loaded. Although the current approach is more favorable for customers, several operating issues may arise without a dedicated

planning and scheduling optimization tool. Examples of such issues include: scheduling extra trains to meet the deadlines; more congested tracks and yards; increased idle time of expensive equipment (e.g., \$1 million locomotives, \$65,000 railcars); additional maintenance; and consequently a risk of higher operating costs.

In this article we discuss cutting-edge mathematical models and state-of-the-art algorithms for solving the enormous, immensely complex, and commonly interrelated problems arising from the above-mentioned issues. The decision-making issues arise at each level of planning: (i) strategic: design of track network; increasing line capacity; location of track sidings; building new yards, improving existing yards, and closing outdated yards; plus sizing and updating the fleet of locomotives and railcars; (ii) tactical: developing train and locomotive schedules, maintenance schedules, empty railcars and containers repositioning schedules; and (iii) operational: crew management; yard operations such as switching, consist-busting/formation, and so on. Also, we review the planning-related optimization problems (train planning, locomotive planning, and crew planning) which have been effectively solved using OR-tools, and implemented at various railroad companies that showed significant quantitative as well as qualitative benefits. We describe each problem in detail including the unique characteristics which make it so challenging.

A train consists of blocks of railcars with different origins and destinations, a set of locomotives to pull the railcars, and a set of crews to operate it. From the demand side, each major railroad may receive 4000–7000 new shipments every day, with each shipment consisting of a set of railcars with a common origin and destination [2]. These shipments are grouped together according to a blocking plan; each shipment, or block, is shipped as a single unit. Adhering to blocking plans helps to minimize the total transportation and handling costs. After the blocking plan is prepared, the next task is to develop train schedules (master schedule) in order to determine the train routes, their frequencies,

and the days of operation for each train. The master schedule is critical in minimizing the block carrying cost. Detailed timetables are then developed using several operational and safety constraints with a goal of limiting the total deviations and delays (train dispatching or meet-and-pass problems) from the master schedule. We study these three closely related problems in depth in the section titled “Train Planning.”

The section titled “Locomotive Planning” covers the locomotive planning problem and specifically deals with providing enough power to the scheduled trains by efficiently assigning and routing them over the network. A number of factors must be considered, such as fleet-size constraints on different locomotive types, fueling and maintenance restrictions, as well as matching restrictions between trains and locomotives. At the tactical level, besides deciding on the type of locomotives used, we also balance their availability throughout the network by deadheading (pulling locomotives from one place to another attached to scheduled trains such as railcars), or light travel (locomotives traveling on their own and not pulling any railcar). At the operational level, we assign specific units from the available set of locomotives to the scheduled trains.

After the train and locomotive planning is completed, crews are assigned to each train to execute these plans. The crew planning aspect is one of the most difficult challenges of railroads, similar to the problems experienced in the airline industry, owing to the presence of union rules in addition to other operational constraints (see the section titled “Crew Planning”). The objective is to assign crews to each train and maintain a balanced availability at each location in order to avoid crew-related delays so that crew costs are minimal.

Infrastructure modifications do not occur frequently in railroads, but the industrial revolution and subsequent changes have created the need to upgrade. A vast amount of literature exists on these related topics, such as finding rail-yard locations and track maintenance problems. Yard location is one of the foremost strategic-level concerns. The objective is to find the best possible locations

for the yards (venue for blocks classification and train switching) over the network, so that effective blocking plans can be developed. Another variation of this problem is to assess the effect of closing an existing yard on the railroad's overall performance, such as the impact on transportation cost. An optimally planned yard network has the potential of saving millions of dollars every year, as it reduces the number of handlings and helps in improving the blocking plan. We refer the readers to the papers by Ahuja *et al.* [3] for the details on this topic.

The freight trains carry tens of thousands of tons of weight and run at a very high speed. The tracks supporting these operations must have regular maintenance to ensure their safe and optimum performance. Nemani *et al.* [4] discuss the problem of preparing an annual timetable for the maintenance-crews to complete the track repair or replacement jobs (curfew planning). It is a highly constrained problem with a large number of restrictions, such as the number of regions being repaired simultaneously; work-continuity (no crews should be idle anytime except during vacations); and schedule development that minimizes the number of constraint violations. The paper discusses several exact and optimization-based heuristics to solve the problem.

In this article, we skip the infrastructure related problems and focus on the scheduling and planning of railcars, locomotive, and crews. In the following sections we discuss each of these problems, along with approaches for solving them.

TRAIN PLANNING

Train planning involves a myriad of decisions, from the arrival of customer orders (railcars) at the origin to delivery at their destination points, with intermediate attaching and detaching to existing trains along the route. In this section, we study three of the most complex and closely interrelated problems, which are solved sequentially to generate a train plan: railroad blocking, train scheduling, and train dispatching.

Railroad Blocking Problem

Railroad blocking is a classification problem which aims to reduce costs primarily through minimizing the intermediate (between origin and destination) handling of railcars. Each railroad receives thousands of shipments (set of railcars) everyday at its yards to ship between their origin–destination (O–D) pairs. Ideally, each shipment should be handled individually and transported via the shortest path between the O–D pair. However, the resources available at yards (crews, classification tracks, and switching locomotives) limit the classification capacity and render individual handling infeasible. Besides, scheduling a direct train between each O–D pair without any intermediate handling is impossible considering the huge railroad network. The small amount of freight for many O–D pairs and the limited rail network structure also encourage the consolidation of overall network. Therefore, several shipments are grouped into one block and transported between the block's O–D pair as a single unit. A block O–D pair may be different from the O–D pairs of individual shipments, so shipments may require reclassifications at intermediate yards. For example, suppose two shipments originating at the Jacksonville (JAX) yard have destinations of Los Angeles (LA) and San Diego (SD), respectively. Both can be grouped with other shipments headed for California, into a block with the O–D pair as JAX–LA. At LA, another block will be created with all shipments arriving at LA and headed to SD. In this example, the shipments with the O–D pair JAX–SD are transported by grouping them into two blocks with different O–D pairs.

It should also be noted that once a shipment is placed in a block, it will not be classified again until it reaches the destination of that block. The classification process often induces a one-day delay for the shipment, causing a potential bottleneck. Besides the indirect cost of delay, the (re)grouping costs are also substantial. For a typical railroad, the average cost of reclassification is \$40 per car. Assuming that a railroad ships 50,000 shipments per month, on average each

containing 10 cars and requiring 2.5 classifications, it would require six million cars per year costing approximately \$1 billion for the classification process. Even a minor reduction in these costs by employing OR-based tools will save tens of millions of dollars for a single US railroad company.

Mathematical Models. In the past, several researchers have attempted to solve the blocking problem by using mathematical models or some heuristics [5–9]. They use network-based models to formulate this problem, albeit with limited success from an implementation point of view. We briefly describe a *multicommodity flow* [10] based model, along with its limited applicability, followed by an implementable solution approach.

Let $G = (N, A)$ be the blocking network, where N is the set of all nodes denoting origin, destinations, or intermediate stations and A is the set of all potential arcs in $N \times N$, that is, $(i, j) \in A$ if a block can be built from node i to node j . The classification cost of a car is h_i and the cost of shipping a block through arc (i, j) is c_{ij} . We need to determine: (i) y_{ij} : which arcs to build, and (ii) x_{ij}^s : the flow of each shipment s of set S (collection of all shipments, containing n_s cars) on each arc (i, j) .

We develop a blocking model with the objective of minimizing all flow and handling costs (Eq. (1)), honoring several operational and modeling constraints. The problem requires that all n_s cars of a shipment s of set S , must be sent along a unique path (Eqs (2) and (7)). The resource availability restricts the maximum number of blocks, b_i , made at a node $i \in N$ (Eq. (4)), as well as the maximum number of cars, d_i , it can handle (Eq. (5)). The geometric configuration does not allow more than a specific number, u_{ij} , of cars to pass through the arc (i, j) of the network (Eq. (3)). Let $I(i)$ and $O(i)$ be the set of incoming and outgoing arcs of node i , respectively; $o(s)$, and $d(s)$ denote the origin and destination of s , respectively. All these notations will be followed throughout the article

unless otherwise stated.

$$\text{Minimize } \sum_{s \in S} \sum_{(i,j) \in A} c_{ij} x_{ij}^s + \sum_{i \in N} \sum_{s \in S} \sum_{(i,j) \in O(i)} h_i x_{ij}^s \quad (1)$$

$$\text{subject to } \sum_{(i,j) \in O(i)} x_{ij}^s - \sum_{(i,j) \in I(i)} x_{ij}^s = \begin{cases} n_s, & \text{if } i = o(s) \\ 0, & \text{if } i \neq o(s) \text{ or } d(s), \forall s \in S \\ -n_s, & \text{if } i = d(s) \end{cases} \quad (2)$$

$$\sum_{s \in S} x_{ij}^s \leq u_{ij} y_{ij}, \forall (i, j) \in A \quad (3)$$

$$\sum_{(i,j) \in O(i)} y_{ij} \leq b_i, \forall i \in N \quad (4)$$

$$\sum_{s \in S} \sum_{(i,j) \in I(i)} x_{ij}^s \leq d_i, \forall i \in N \quad (5)$$

$$y_{ij} \in \{0, 1\}, \forall (i, j) \in A \quad (6)$$

$$x_{ij}^s \in \{0, n_s\}, \forall (i, j) \in A, \forall s \in S. \quad (7)$$

A class I railroad on average receives 50,000 shipments (each containing 1–50 cars) at approximately 1000 origins, 300 yards, and 2000 destinations every day. The number of possible blocks for this network will be around one million (between each origin-yard pair, each yard-yard pair, and each yard-destination pair), pushing the total number of decision variables into the billions. The exponential size of the instances eliminates the use of any standard exact optimization model and encourages the use of heuristics to find near-optimal solutions in a reasonable time frame. We describe a very large scale neighborhood search algorithm for blocking problems, which has been successfully implemented at several railroads.

Very Large Scale Neighborhood (VLSN) Search Algorithm. The VLSN search algorithm is a variant of the local search algorithm that effectively explores the exponentially large neighborhood and converges to a good near-optimal solution. It is applicable to only those problems that can be expressed as an instance of partition problems [11]. Through the course of

developing algorithms for blocking problem, we observed that in most cases honoring the blocking capacity (b_i) at each yard is enough to ensure its car-handling limit (d_i) and maximum flow constraints on arcs (u_{ij}). Hence, these constraints can be ignored for all practical purposes and can be handled using the Lagrangian relaxation approach [10]. We also observe that these relaxations reduce the problem to finding only the arcs to be built, as all capacity constraints have been relaxed. Once the arcs are built, flows on each arc can be found by determining the shortest path for each shipment. The VLSN algorithm uses these concepts in its two-phase approach: (i) construction phase and (ii) improvement phase.

The construction phase aims to find an initial solution, which will form the basis for the improvement phase. The initial solution must be feasible. Therefore, it should ensure that the origin node of each shipment is connected to its destination node through a blocking path without violating the blocking capacity of any node. We obtain this solution in four simple steps: (i) construct a directed cycle that passes through all the yards exactly once and returns to the first yard; (ii) create a blocking arc from each origin node to the nearest yard; (iii) connect each destination node to the nearest yard that still has some blocking capacity; and (iv) route shipments along the shortest path over the network. The initial solution is then improved in the second phase by searching for better solutions in its neighborhood.

Creating the neighborhood structure is an art and the success of a neighborhood search algorithm is dependent on it. The neighborhood of a blocking problem is constructed by changing the blocks made at a single node and re-solving the shortest path problems on an updated arc-set. Clearly the neighborhood size is very large, as there are N nodes to choose from; for each node with p emanating arcs, there are pC_k ways in which a subset of k arcs can be selected. The problem of finding a better solution in this neighborhood is NP-complete and hence, an explicit search in a reasonable time frame is impossible. A heuristic algorithm has been developed by

Ahuja *et al.* [5], which uses a five-step subroutine at each node i : (i) delete all arcs emanating from node i ; (ii) reroute the flow on the updated arc-set; (iii) create one new arc at a time from node i and determine the potential savings, assuming no change at other nodes; (iv) add each new arc to the network that provides maximum savings; and (v) repeat steps (iii) and (iv) until the blocking capacity is met or all options are exhausted.

The proposed algorithm is very flexible in accommodating new constraints based on a variety of scenarios at different railroads. Some railroads may not want to deviate from their existing plans by more than a specific value, and hence fix several blocks in the input. They may also want to send a shipment through a specific route. Owing to demand patterns in a special direction and other external factors (empty car repositioning), they may want blocks to travel only in a specific direction between two nodes. Several such scenarios have been considered in the article by Ahuja *et al.* [5].

Conclusions. The proposed VLSN search algorithm has been tested and implemented at three large US railroads. Each railroad has realized significant reductions in their average car miles (0.1–4%), and in the number of reclassifications (31–42%) at the yards over their manual plans. The dynamic nature of the problem also restricts the computational time. The VLSN algorithms proved very effective on this front too, by solving different instances between 5 min and 2 h.

After a blocking plan is prepared, developing a train schedule is the next most important operational planning task. To realize the savings mentioned above, a proper match between the train schedules and blocking plan is required. In the next section, we discuss this problem and offer several solution approaches.

Train Scheduling Problem

A train schedule is a detailed description of the train routes, their frequencies, and their timetables on the railroad's network designed to minimize the cost of carrying cars

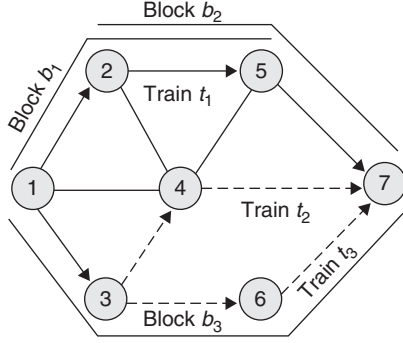


Figure 1. Structural components of a train scheduling problem.

from their origins to destinations. Figure 1 shows a simple example of a railroad network with nine nodes (stations), three trains and block paths. Train t_1 starts at node 1, follows the route 1–2–5–7, and terminates at node 7. Train t_2 follows the route 1–3–4–7, and train t_3 traverses the path 3–6–7. Three blocks of cars, b_1 , b_2 , and b_3 , travel on these trains. Block b_1 takes train t_1 over the train segments 1–2–5, while block b_2 takes the same train over the segments 2–5–7. As a train travels from its origin to destination, it picks up blocks at various nodes it visits and may also drop off blocks at those nodes. Typically, several blocks ride on a train at any segment. Likewise, a block may travel on several trains as it goes from its origin to its destination. For example, block b_3 starts at node 1 and travels on train t_2 on the segment 1–3. It is off-loaded by train t_2 at node 3, where it is picked up by train t_3 and then carried from node 3 to its final destination, node 7. The transfer of a block from one train to another is called a *block swap*. Among the blocks in our example, only b_3 performs a block swap.

Figure 1 shows the interplay between the railroad network, trains, and blocks for a simplified scenario. We have not included the time factor in the figure to keep it simple. In practice, from any station (origin/intermediate/destination) only a limited number of trains can pass through within a specified time window. To avoid congestion, railroads also limit the number of trains passing through any arc of the network. In

addition, the number of cars traveling on any segment must be within a prespecified range. We need to determine the total number of trains and their details (routes, arrival–departure time at each station on the route, frequency, and matching blocks) and honor all these constraints so that the operating cost is minimized. The cost structure may include several factors based on the specific scenario: total distance traveled by cars; total travel time of cars between their O–D pairs; total number of block swaps; total number of locomotives (for pulling the blocks) and crews required; total number of trains formed in a week; total number of train stops at stations; and the variation in block-to-train assignments over a week.

A typical railroad comprises about 400 trains that run on average five days a week, with an average travel time of one day between the O–D pairs. The daily operating cost of a train includes about \$1500 for the locomotive and \$1500 for the crew. Hence, the average annual cost of running one train is approximately \$0.8 million and the average cost of running all trains is around \$320 million. This data suggests that replacing the current manual approach with the OR-tools, in conjunction with improving the service level, will save tens of millions of dollars. We discuss such an optimization model in the next section.

Mathematical Models. The train scheduling problem is a very large scale network optimization problem with trillions of decision variables. Several attempts [12–17] have been made to solve this problem using a two-phase approach, which repetitively develops the train routes followed by the block routes. Some heuristics have also been designed to solve the integrated problem [18] but with limited success. We discuss a decomposition-based two-phase approach, which first determines the train routes and block matching and then decides other details of the train such as frequency and schedule times [2]. Before looking into details of the decomposition approach, we present a simplified mathematical model because the complete global problem is impossible to solve within a reasonable time frame using

current state-of-the-art computing power. We assume that each train runs every day of the week and ignore the train timing.

Let $G(N, A)$ be the scheduling network where N is the set of all origin/destinations/intermediate stations and A is the arc set. We add two dummy nodes s and t , representing the train's global origin and destination, respectively, and an arc (s, t) that denotes the flow of those trains that are not built. It is a variant of the multicommodity flow problem, where each train starting from source s is a unique commodity. We determine if a train $k \in K$ (set of trains) passes through arc $a \in A$ by variable $y_a^k \in \{0, 1\}$; and if it passes, whether block $b \in B$ belongs to the train by variable $x_{ka}^b \in \{0, 1\}$. We aim to minimize the cost of running trains, traveling blocks, and their intermediate swaps (Eq. (8)). As stated earlier, from each station i , the total number of trains originating (T_i^+) or terminating (T_i^-) should be within a limit (Eqs (10) and (11)). Each arc also has a limit (T_a) on the number of trains traveling on it (Eq. (12)). Finally, there are some modeling constraints: flow balance on trains (Eq. (9)), flow balance on blocks (Eq. (13)), and block–train relations (Eq. (14)). The number of cars carried by a train $k \in K$ through arc $a \in A$, should be within the given lower limit L_a^k and upper limit U_a^k . v^b represents the number of cars in a block (Eq. (14)). We assume that the total number of scheduled trains is enough to move all blocks.

$$\begin{aligned} \text{Minimize } & \sum_{k \in K} \sum_{a \in A} f_a^k y_a^k + \sum_{b \in B} \sum_{k \in K} \sum_{a \in A} c_b^k x_{ka}^b \\ & + \sum_{b \in B} \sum_{\substack{i \in N \\ i \neq o(b), i \neq d(b)}} s_i^b \left| \sum_{a \in O(i)} x_{ka}^b - \sum_{a \in I(i)} x_{ka}^b \right| \end{aligned} \quad (8)$$

$$\begin{aligned} \text{subject to } & \sum_{a \in O(i)} y_a^k - \sum_{a \in I(i)} y_a^k \\ & = \begin{cases} 1, & \text{if } i = s \\ 0, & \text{if } i \neq s, t, \forall i \in N, \forall k \in K, \\ -1, & \text{if } i = t \end{cases} \end{aligned} \quad (9)$$

$$\sum_{k \in K} \sum_{a=(s,i)} y_a^k \leq T_i^+, \forall i \in N \quad (10)$$

$$\sum_{k \in K} \sum_{a=(i,t)} y_a^k \leq T_i^-, \forall i \in N \quad (11)$$

$$\sum_{k \in K} y_a^k \leq T_a, \forall a \in A \quad (12)$$

$$\begin{aligned} & \sum_{k \in K} \sum_{a \in O(i)} x_{ka}^b - \sum_{k \in K} \sum_{a \in I(i)} x_{ka}^b \\ & = \begin{cases} 1, & \text{if } i = o(b) \\ 0, & \text{if } i \neq o(b) \text{ or } d(b), \\ & \forall i \in N, \forall b \in B \\ -1, & \text{if } i = d(b) \end{cases} \end{aligned} \quad (13)$$

$$L_a^k y_a^k \leq \sum_{b \in B} v^b x_{ka}^b \leq U_a^k y_a^k, \forall a \in A, \forall k \in K \quad (14)$$

$$y_a^k \in \{0, 1\}, \forall a \in A, \forall k \in K \quad (15)$$

$$x_{ka}^b \in \{0, 1\}, \forall a \in A, \forall k \in K, \forall b \in B. \quad (16)$$

A typical railroad comprises about 2000 stations, 12,000 arcs, 1300 blocks, and 400 trains. In the above simplified model, there are around seven billion X and five million Y binary variables. An instance of this size cannot be solved under current computational capabilities within a reasonable time frame. As we discussed earlier, such problems are solved using decomposition methods. Next, we describe a very effective algorithm to solve the train scheduling problem.

Decomposition Approach to Solve Train Scheduling Problem. The decomposition methods divide the single very large scale problem into several parts and then optimize each part independently using the output of the other parts. We decompose the train scheduling problem in two phases. In the first phase, we determine the routes of each train with its O–D pair and the blocks to be carried, along with optimizing the number of train starts, total train miles traveled, average number of cars per train mile, and the number of block swaps. The algorithm is described in Fig. 2. This second phase uses the solution of the first phase to help determine the other details such as train times and route frequency. It is comparatively more complex and needs more attention.

```

algorithm Train-Scheduling Phase 1;
begin
  Map blocks on the shortest paths;
  for each block destination do
    enumerate all possible trains under the assumption that the train can start only at block's origin;
  while any block  $b$  is not routed
    begin
      Empty the set  $T$ ; //  $T$  is set of trains in which block  $b$  can be attached.
      for each train  $t$  do
        if (O-D pair of the block  $b$  is on the route of any train  $t$  AND a train-path exists for  $b$  from its
          origin to the station to catch the train  $t$ ) then
          determine the goodness of train  $t$  and put it in set  $T$ ;
      create a set of link-disjoint best trains from  $T$ ;
    end;
  for each train  $t$  with per-mile car volume  $< n^{\text{threshold}}$  do // a pre-specified value
    re-initialize the set  $B_t$  with blocks on train  $t$ ;
    if (An alternate route exists for each unassigned block such that
      there are no more than  $k$  trains in the new train path of any of the block of  $B_t$  And
      the new train path should not be more than  $x\%$  of the shortest path) then // a pre-specified value
      delete train  $t$  and reroute its blocks;
  end;

```

Figure 2. Phase 1 of the decomposition algorithm for the train scheduling problem.

The second phase is based on an underlying seven-day train-space-time network, in which nodes represent the physical stops of the train and the arcs represent the train's path. At each of the train stops, additional d nodes are created in time dimension and sorted in the order of departure time to represent the trains running d days per week. *Each train running on a particular day of the seven-day planning horizon is denoted as a train leg.* Once the network is created, this phase is solved in three major steps as explained next. We assume that trains run only by the hour (i.e., at 00, 01, ..., 23 h) and once the departure time is fixed at the origin, then arrival and departure times at other stations can be automatically determined.

- Step 1a: Assume that each train runs every day of the week.
- Step 1b: Add the trains to the time-space network where times have been already specified, and assign the blocks that can be completely routed on these trains.
- Step 1c: Arrange the trains whose times have to be determined in descending order of the total car miles.

Step 1d: Find the departure time of each train in the order sorted by simple enumeration of 24 possible options, and then select the one that minimizes the total car flow time.

Step 2a: Consider the 10 *train legs* with the lowest car volume, and find the impact of the deletion of these legs one at a time.

Step 2b: Delete the train with the lowest impact. Repeat steps 2a and 2b until each train leg has sufficient volume (i.e., each leg contains more than a prespecified number of cars, say 25).

Step 3: Reoptimize the departure times of the remaining trains once again, as detailed in step 1d. This step is required, because while reducing the frequencies of trains in step 2 the block assignment changes. In the new scenario, the previously calculated times (in step 1d) may not remain optimal.

Conclusions. The proposed algorithms have been tested on the real-world data of NS

Corporation [2]. The results achieved were superior in each of the cost terms mentioned earlier in the article, to a varying degree (1–20%). Besides considering all practical constraints of the train scheduling problem, these algorithms are flexible enough to incorporate new business requirements arising at other railroads, such as giving preference to some trains over others and preassigning some blocks to particular trains. These results show the potential for railroads to save millions of dollars by optimizing the number of rolling stocks and the intermediate operations.

The train schedule creates an ideal plan (arrival and departure time at main stations) by routing trains on the given network and assigning blocks to them. However, because of the immense complexity of the problem, it is difficult to factor in all the possible conflicts that could occur as the trains travel and meet along the tracks. We resolve these issues by using the meet-pass planner (train dispatcher), which determines the detailed train movements and timetables.

Train Dispatching Problem

To implement a train schedule in real-life scenarios, it should provide the arrival and departure times of trains not only at the main terminals but also at each track section along their routes. The tracks of a railroad vary from one section to another in their capacities to carry trains. A very simplified track structure is shown in Fig. 3 to describe the relevant terminologies. It contains two major stations (*B* and *D*), two intermediate stations (*A* and *C*) where trains can wait for crossing or overtaking, and four track sections (tracks between two consecutive stations or sidings, e.g., 1, 2, 3, and 4). Sidings are used

for crossing or overtaking. Track section 3 is double-lined, which can allow at most two trains to pass through in any direction; the others are single-lined and can accommodate only one train at a time.

Most of the freight railroads comprise single-line track sections, which causes a delay and deviation from the original plan whenever a train is pulled over to cross or overtake another. The percentage deviation from the original plan is an indicator of deviation from the savings proposed in the section titled “Train Scheduling Problem.” Therefore, in this problem we aim to determine detailed timetables at each track section/sidings (where the trains will meet and pass), which minimizes the total deviation from the original plan honoring several operational and safety constraints. Some of these restrictions are listed below:

- Track capacities should never be violated. It restricts the stoppage of trains at small sidings or intermediate terminals.
- Trains that do not have good traction power to resume moving once stopped at sidings with a sharp slope toward the direction of movement cannot be stopped at those sidings.
- Some trains have extremely strict schedules and incur a very high penalty if they are delayed by more than a prespecified duration.
- Only those intermediate stations that have a well-developed infrastructure to support the needs of crews can be used as a waiting place for long periods.

All of these restrictions make the problem even more complex, thereby limiting the applicability of standard models. We describe

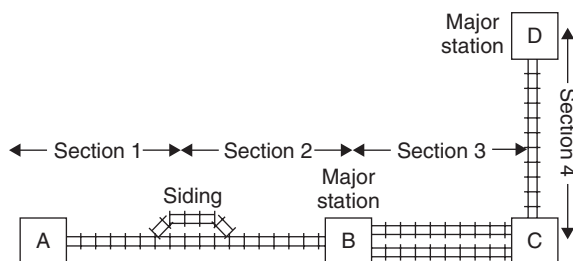


Figure 3. An illustration of track structures.

a mathematical model to show the enormity of the problem, which becomes the basis of IP-based heuristics.

Mathematical Models. The train dispatching problem is a very large scale problem and variants of it have been attempted in literature [19]. Several models have been proposed [20–22] with the assumption that all trains always operate at their constant maximum velocity. These approaches are based on branch and bound techniques [21], Lagrangian relaxation [20], tabu search [22], and other local search heuristics. A few papers [23–26] solve problems with variable trains’ velocity, but with other assumptions such as “only single-line tracks are available,” “traffic flows in only one direction,” or “crossing/overtaking takes place only at stations ignoring the roles of sidings.” All of these simplifications reduce the direct applicability of these approaches in real-world scenarios. In this section, we discuss the algorithm developed by Sahin *et al.* [27], which is based on an underlying space-time network, $G(N, A)$, where N and A denote the nodes and arcs, respectively (Fig. 4). N is the superset of three sets of nodes:

- *Departure Nodes* $D = (d^t)$. Set of artificial nodes each where each element d^t represents the origin of a train $t \in T$.
- *Arrival Nodes* $A = (a^t)$. Set of artificial nodes each where each element a^t represents the destination of a train $t \in T$.
- *Station-Time Nodes* $S = (s^q)$. Set of all pairs of stations/sidings, $R = \{0, 1, \dots, r\}$ and times, $Q = (1, 2, \dots, q)$, equally divided in the given time horizon. Each node (s^q) represents the state of a station/siding $s \in R$ any time $q \in Q$.

Modeling a highly constrained problem as a network flow problem has the intrinsic advantage of reducing the complexity by creating only feasible nodes and arcs. We demonstrate how most of the constraints discussed earlier can be handled while constructing the arc-set A . We create arcs for one train ($t \in T$) at a time following the sequence of origin–intermediate stations–destination,

through which train t passes and the consecutive sections between these stations. While creating these arcs, we assume that railroads comprise only single-line track sections. The network contains four sets of arcs each with unit capacity as follows:

- *Origin Arcs*: Arcs from the nodes of set D to those nodes of set S whose station attribute is the origin (o^t) of the train t ; and the time attribute is between its earliest and latest start times (departure time window of train $tdep^{tw}$). The cost of these arcs corresponds to the length of delays at the starting nodes.
- *Destination Arcs*: Arcs from those nodes of set S whose station attribute is the train’s destination (d^t); and the time attribute is between the corresponding scheduled and latest arrival time (arrival time window of train $tarr^{tw}$) to the nodes of set A . Costs for these arcs are set to 0.
- *Travel Arcs*: Arcs between each pair of nodes of set S whose station attribute represents the pair of consecutive stations on the train’s path; and the time attribute lies within the earliest and latest start time for the section between them. Costs for these arcs are set to 0.
- *Waiting Arcs*: Arcs between each such pair of nodes s^q and s^{q+1} of set S for which the train t is allowed to wait at station/siding $s \in R$ and there is an incoming arc to s^q representing the train’s arrival. If there is an additional constraint that the train t must wait at siding $s \in R$ from time q to $q + 1$, we ignore travel arc from s^q and create only the travel arc from node s^q to node s^{q+1} . Costs for these arcs is set to the length of time period, that is, $q + 1 - q$.

We give the IP formulation based on the space-time network, in which we need to determine the flow x_{ij}^t , of train $t \in T$, on arc $(i, j) \in A$ so that all costs are minimized (Eq. (17)). Constraints (18)–(20), along with the arc-capacity constraints (21), ensure a unit flow of train t from its departure node to arrival node. Constraints (20) maintain the

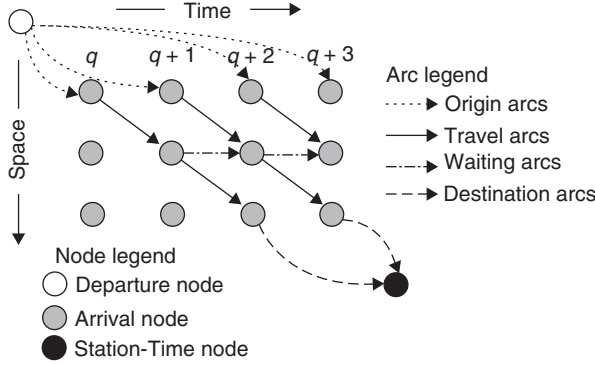


Figure 4. An illustration of the space-time network for the train dispatching problem.

balanced flow throughout the network. Constraints (22) ensure the flow of at most one train in any section at any time, by enforcing the sum of flows in both directions (T^{out} is outgoing trains from station $s \in R$).

$$\text{Minimize } \sum_{t \in T} \sum_{(i,j) \in A} c_{ij} x_{ij}^t \quad (17)$$

$$\text{subject to } \sum_{q \in \text{dep}^{\text{tw}}} x_{D^t o_q^t}^t = 1 \forall t \in T \quad (18)$$

$$\sum_{q \in \text{arr}^{\text{tw}}} x_{d_q^t A^t}^t = 1 \forall t \in T \quad (19)$$

$$\sum_{i \in S} x_{ij}^t - \sum_{i \in S} x_{ji}^t = 0 \forall t \in T, \forall j \in S \quad (20)$$

$$\sum_t x_{ij}^t \leq 1 \forall (i,j) \in A \quad (21)$$

$$\sum_{t \in T^{out}} \sum_{l \leq q-1} \sum_{m \geq q} x_{s_l(s+1)_m}^t + \sum_{t \in T^{in}} \sum_{l \leq q-1}$$

$$\sum_{m \geq q} x_{(s+1)_l s_m}^t \leq 1 \forall s \in R, \forall q \in Q \quad (22)$$

$$x_{ij}^t \in \{0, 1\} \forall (i,j) \in A, \forall t \in T. \quad (23)$$

A class I railroad generally comprises about 100 stations and sidings, and may require the time interval of 10 min included in the timetable [27]. In this case, a planning period of 48h will result in approximately 28,800 nodes; the flow of even 50 trains will push the number of binary variables up to 720,000. The vast size of the problem

makes the current IP-solvers inapplicable and encourages the development of a heuristic approach. Besides the quality, the solution-generation time should also be low (5–10 min) considering the dynamic behavior of the problem, with continuously changing scenarios depending on the occurrence of several events.

IP-Based Heuristics. The IP-based heuristics is based on two observations from the attempts to solve the IP-model using CPLEX: (i) CPLEX generated optimal solutions for the small-size instances within a few minutes; and (ii) CPLEX generated average quality feasible solution within a minute for large instances. These facts were used by Sahin *et al.* [27] in developing the IP-based heuristics. It is an iterative algorithm, with each iteration solving a smaller-sized problem for a prespecified time. Besides the number of stations, trains, and time windows, the size of the problem in the time-space network depends on how many delays are allowed for each train. A larger delay allowance permits trains to wait at different terminals and increases the number of potential train-paths in the network exponentially. We use only this characteristic to control the size of the problem to be solved in each iteration. The basic framework of the algorithm is given below and other implementation details can be found in Ref. 27. Let *SolutionTime* be the time limit to find the solution.

Step 1: Divide the scheduling horizon q into two periods: short ($q_{\text{short}} \leq q$) and long term ($q_{\text{long}} = q - q_{\text{short}}$).

- Step 2: Initialize allowed delays for each train $t \in T$ by $ad_t = \text{maximum allowed delays}$.
- Step 3: Find a feasible solution by solving the IP-model using CPLEX. Let δ^t be the delay of the train $t \in T$ in the current solution.
- Step 4: Reduce/update the allowed delays for each train $t \in T$. For short term, set $ad_t = \min(\delta^t + \delta^1, ad_t)$, and for long term, set $ad_t = \min(\delta^t + \delta^2, ad_t)$. δ^1 , and δ^2 are incremental changes in train delays.
- Step 5: Find the best feasible solution by solving the IP-model using CPLEX for a prespecified time *IterationTime*.
- Step 6: Repeat steps 4 and 5 until either *SolutionTime* is reached or no improvement is found in step 5.

Conclusions. The heuristics developed to solve the dispatching problems were tested on problem instances of a major Brazilian railroad that transports heavy freight. The resulting plan showed a very large improvement over the railroad's solution by decreasing the deviation from the optimal schedule from 71% to 0.06%. The high solution quality, combined with a fast response time of the algorithm, shows its potential of being developed into a real-life decision making tool.

LOCOMOTIVE PLANNING

After the train schedules have been finalized, the next important problem is to assign a set of locomotives (consist) to each scheduled train so as to provide sufficient power to pull the trains from their origins to destinations at a minimum cost. The schedules of trains differ from one another in several aspects, such as frequency per week, number of days taken to travel from origin to destination, pulling power required, and compatibility with the different types of locomotives. Locomotives also vary in their pulling power, costs, fleet size, fueling and servicing requirements, allowable combination to

form the consist, and so on. Each locomotive must be sent to maintenance shops at periodic intervals (every 3000 miles for regular, and every 10,000 miles for major maintenance) and, therefore, can be assigned only to those trains that do not take them too far from shop locations on their due date. Several other constraints are discussed in the section titled "Mathematical Models," which explains the mathematical model.

These vastly different characteristics of trains and locomotives cause an imbalance in the flow of locomotives on the network, resulting in a shortage at some terminals and a surplus at others. To balance this flow so that the availability of locomotives is ensured for each train, these are either *deadheaded* (pulled like railcars by attaching to a scheduled train) or *light traveled* (traveling on their own without attaching to other locomotives or railcars) from the surplus to shortage terminal. Both of these indicate poor utilization of the locomotives, with light travel also requiring extra crews. Deadheading is less costly than light travel, but it is dependent on the train's schedule and results in comparatively slower travel.

The difference in the power requirement of trains arriving at and departing from a station gives rise to *consist-busting*. Whenever a train arrives at a station, either its consist can be directly assigned to a departing train with same requirement or it can be sent to a pool. At the pool, consists are disassembled and regrouped into new sets of consists comprising different types and numbers of locomotives. Ideally, there should not be any consist-busting, as it is very costly (requires servicing crews and equipment) and is a potential source of delays. If a departing train takes locomotives from several incoming trains, the delay in any of these facets will have a rippling effect in the network (propagating delays all over the network).

A typical railroad has over 3000 locomotives, which translates into a capital investment of over \$5 billion, as well as over \$1 billion in yearly maintenance and operational costs. The locomotive utilization is around 50%. Accordingly, even a small improvement will free up many locomotives (300–400) and

result in a savings of over \$100 million per year.

Mathematical Models

The problem has been studied by several researchers [28–31] with a varying degree of success. Most of them deal with scenarios in which only one locomotive is required for each train, which is applicable only to European railroads. These approaches are not applicable to US railroads where up to 12 locomotives (including both the active and deadhead locomotives) can be assigned to a train. Issues such as the costs associated with deadheading and light travel, as well as the preference between trains and locomotives, are relaxed to keep the problem simpler. Ahuja *et al.* [32] solve the problem including all practical issues, which is further improved by Vaidyanathan *et al.* [33]. The second article is based on consist flow rather than locomotive flow, and thereby nearly eliminates the chances of consist-busting. We will discuss the second approach in this section.

Given a weekly train schedule, we construct a space-time network $G(N, A)$ for the problem and formulate it as a multicommodity flow instance with side constraints, in which each consist type defines a commodity (Fig. 5). The node-set N contains three set of nodes:

- *Arrival Nodes (ArrNodes)*: Each node has the arrival station as its place attribute and the arrival time as its time attributes.
- *Departure Nodes (DepNodes)*: Each node has the departure station as its place attribute and the departure time as its time attribute.
- *Ground Nodes (GrNodes)*: Each node represents a pool at a particular time and station. These nodes facilitate the transition of consists from inbound trains to outbound trains. They are of two subtypes: arrival ground nodes (one for each arrival node with the same place and time attributes) and departure ground nodes (one for each departure node with the same time and place attributes).

We create four set of arcs: $AllArcs = TrArcs \cup CoArcs \cup GrArcs \cup LiArcs$.

- *Train Arcs (TrArcs)*: For each train there is one arc from its *DepNode* to its *ArrNode*.
- *Connection Arcs (CoArcs)*: It contains two subsets of arcs: (i) from each *ArrNode* to its associated arrival ground node, and (ii) from each departure ground node to its associated *DepNode*.
- *Ground Arcs (GrArcs)*: Sort all the *GrNodes* at each station into the chronological order of their time attributes. *GrArcs* connect each ground node to the next ground node in this order. This represents the flow of consist in and out of the pool. These also allow consists to stay in the pool if the outgoing train is not immediately available. The last ground node of the week is connected back to the first node of the week in order to model the fact that the ending inventory of locomotives becomes the starting inventory for the next week.
- *Light Arcs (LiArcs)*: These arcs are from ground nodes of different stations and represent light travel of trains. A fixed cost of light travel is assigned to these arcs.

We now illustrate the mixed integer programming (MIP) formulation based on the space-time network. It should be noted that in this model we deal with the flow of consists rather than of locomotives. We have four sets of decision variables: (i) x_l^c is binary representing if consist type c flows on the *TrArc* l ; (ii) y_l^c denotes the number of nonactive (deadhead, light-traveling, or idling) consist of type c on *AllArc* l ; (iii) z_l is a binary variable to model the existence of any flow of any consist type on *LiArc* l ; and (iv) s^k is the number of unused locomotives of type k .

Railroads have several choices for consists, and they incur different cost (c_l^c based on the compatibility between a consist c and a train l). Most of the locomotive-compatibility-related constraints are resolved by not creating those consists. The objective function

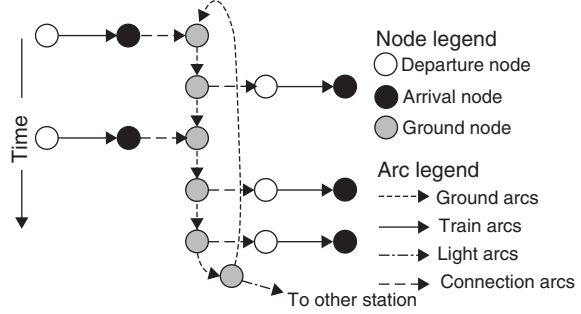


Figure 5. An illustration of the space-time network for locomotive planning.

has four cost terms: (i) total consists cost; (ii) deadhead/light travel/idle cost (d_l^c) of each consist c on any arc l ; (iii) cost of using light arc l (F_l); and (iv) savings (G^k) by not using a locomotive of type k . Constraints (25) ensure that each train is assigned a consist, while constraints (26) and (28) enforce the restriction that a train cannot contain more than 12 locomotives (α^{ck} is the predefined number of locomotives of type k in the consist c) on trains arcs and light arcs, respectively. Constraints (27) maintain a balanced flow through the network, while constraints (29) ensure that the number of locomotives of each fleet type used is no more than the fleet size (B^k). $day(l)$ denotes the number of times an arc l crosses the midnight timeline. For instance, suppose train l leaves at 10 a.m. on day 1 and arrives at 4 a.m. on day 2, then $day(l) = 1$. In contrast, suppose it leaves at 10 a.m. on day 1 and arrives at 4 a.m. on day 3, then $day(l) = 2$.

$$\begin{aligned} \text{Minimize } & \sum_{l \in TrArcs} \sum_{c \in C} c_l^c x_l^c + \sum_{l \in AllArcs} \sum_{c \in C} d_l^c y_l^c + \\ & \sum_{l \in LiArcs} F_l z_l - \sum_{k \in K} G^k s^k \end{aligned} \quad (24)$$

$$\text{subject to } \sum_{c \in C} x_l^c = 1 \forall l \in TrArcs \quad (25)$$

$$\sum_{c \in C} \sum_{k \in K} \alpha^{ck} (x_l^c + y_l^c) \leq 12 \forall l \in TrArcs \quad (26)$$

$$\begin{aligned} \sum_{l \in I(i)} (x_l^c + y_l^c) &= \sum_{l \in O(i)} (x_l^c + y_l^c) \\ \forall i \in AllNodes, c \in C \end{aligned} \quad (27)$$

$$\sum_{k \in K} \sum_{c \in C} \alpha^{ck} y_l^c \leq 12 z_l \forall l \in LiArcs \quad (28)$$

$$\sum_{l \in S} \sum_{c \in C} \alpha^{ck} (x_l^c + y_l^c) day(l) + s^k = B^k \forall k \in K \quad (29)$$

$$x_l^c \in \{0, 1\}, \forall l \in TrArcs, c \in C \quad (30)$$

$$y_l^c \geq 0 \text{ and Integer}, \forall l \in AllArcs, c \in C \quad (31)$$

$$z_l \in \{0, 1\}, \forall l \in LiArcs \quad (32)$$

$$s^k \geq 0, \forall k \in K. \quad (33)$$

The consist-based MIP formulation is smaller in comparison to other locomotive-based models, but the size of the problem still remains quite considerable. A large railroad contains about 600 trains, with each one having a different running frequency, running time, and consist requirement. For the trains that run multiple days, the railroad wants to assign the same consist on each day it runs for the convenience of its operators. Applying these aspects to the MIP will introduce thousands of fixed-charge variables, thereby significantly increasing the complexity and rendering the available software inapplicable. To include all of these issues in the solution model, a decomposition-based heuristics was proposed by Ahuja *et al.* [32]. We briefly discuss it here and refer the readers to Refs 32 and 33 for details. This approach solves the problem in two stages:

- *Daily Locomotive Planning:* We assume that all trains that are running more than a prespecified number of days (p), run every day of the week, and that all trains running less than p days

do not run at all. The space-time network is modified accordingly, and consequently some constraints such as the fleet-size constraints are adjusted. The network size can be decreased by reducing the number of fixed-charge variables. Finally, we use CPLEX to find the solution to the relaxed problem.

- *Weekly Locomotive Planning:* In this stage we adjust the solution of the first stage, which assigns extra consists to some trains (whose frequency increased), and leave selected trains unassigned. A modified space-time network is constructed, in which each train has actual frequency. This stage is solved by further decomposing the multicommodity flow problem into several single-commodity flow problems. We optimize the problem for each commodity (consist type) assuming no change in the assignment of the others. Once all single-commodity problems are solved, a local neighborhood search is applied to improve the solution.

Conclusions

The locomotive planning is a highly combinatorial problem with significant economical impact. As stated earlier, even a minor improvement in the locomotive plan has the potential of saving hundreds of millions of dollars. The proposed algorithm was tested on real-world data from a large North American railroad company, and it showed that the optimized plan schedules all trains with 400 less locomotives than the manual plan. There was also significant improvement in locomotive utilization. These results were obtained by a heuristic approach and they show the opportunities of developing other holistic approaches.

CREW PLANNING

A railroad requires a set of crews that comprise engineers to operate locomotives, conductors to coordinate movement, and other servicing personnel for running the trains. Crew planning involves assigning crews from the set of available crews (*crew*

pool) to trains. The objective is to operate the trains on schedule with a minimum of crew-related costs, while simultaneously satisfying a variety of Federal Railway Administration (FRA) regulations and trade union rules. Besides maintaining the train schedules, railroads also strive to ensure the crew's quality of life (e.g., regular work hours and favorable work locations). However, the FRA and union rules in conjunction with complex train schedules in an overwhelmingly dynamic environment often result in a very complex problem. Accordingly, railroads generally end up using "Assign-Available-Crews-to-Current-Task" rule. Railroads inform crews about their next task just 2h before the train's departure, which can have an origin and destination anywhere in the allowed crew districts. North American rail network is divided into crew districts, each comprising a subset of terminals and rail-links. Each crew is qualified to work only in a subset of crew districts. Crews can be assigned from their current terminal to an *away terminal* where they wait in a hotel until they are assigned back to some other train. *Each crew has a home terminal (generally the home city of the crew members), where railroads do not have to pay for lodging, and all other terminals are called away terminals.* Owing to the imbalanced flow of trains over the network, crews need to be moved by taxi or other trains as idle from the surplus terminal to the shortage terminal. Similar to the terminologies of locomotive planning, we call this idle movement *crew-deadheading*. The deadheading time is counted in normal work hours. Sometimes neither reassigning nor deadheading crews from an away terminal is possible, so they need to wait for a long duration (*crew detention*). If this wait is longer than a prespecified duration (generally 16 h), the crews are paid wages (detention cost) for the extra hours (*total detention hours—16*).

A crew must report for initial preparation before the train's departure, and it must remain to perform some tasks after the train's arrival. The length of time between the arrival (*on-duty time*) and departure (*tie-up time*) of the crew is called the duty period. According to federal law, a train crew must

be sent for sufficient rest (*dead/dog-lawed crew*) depending on the length of their duty period. The crew must wait at least 12 h (10 h of rest followed by a 2-h call period) if the duty period is greater than 10 h, and 10 h (8 h rest followed by 2-h call period) if the duty period is less than 10 h. The minimum 8-h rest period is required after the release from duty, except if no crew is available at the away terminal and either of the following is true: (i) if the recent travel time from the home terminal is followed by a rest of less than 4 h, plus the travel time of a new assignment back home is shorter than 12 h; or (ii) if the recent travel time from the home terminal plus the travel time of a new assignment back home is less than 12 h, and the rest time in between the assignments is more than 4 h. Crews cease doing all duties after consecutive 12 h of work, even if no extra crew is available and the train's departure has to be delayed. A delay also occurs if at any crew change location a crew is not available; this delay costs about \$1000 per hour. There have been some changes in crew's service restrictions as per rail safety improvement act of 2008 which became effective on July 16, 2009, such as (i) 276-h cap on the number of hours operating employee can work in a month; (ii) minimum off duty time has been increased from 8 h to 10 consecutive hours during the previous 24-h period; (iii) the last hour of travel time spent returning from the final trouble call will not be counted toward required off duty time. The complete list of these changes can be found in Ref. 34.

Crews are paid only for their time spent on duty, regardless of the length of their rest period; and they are guaranteed a minimum pay amount per month. The crews are deployed following the first-in-first-out (FIFO) rule at both the home terminal and the away terminals. This guideline requires that the crew with the maximum rest period of all the crews that are resting at the on duty team for the next train is assigned first.

These complex operational, regulatory, and contractual requirements, which also vary from state to state, make it a very challenging problem to solve using mathematical models. Although crew planning

related research has been performed in other industries (e.g., airlines and trucks), in railroads it remains largely unsolved and no commercial optimization product is available for use.

Mathematical Models

Several researchers [35–37] have attempted to solve the crew planning problem using a set-covering based approach. They decompose the problem in *crew pairing* and *crew rostering*. Crew pairing is a sequence of connected segments that not only start and end at the same crew base, but also satisfy all legal constraints. We first find the minimum cost set of the crew pairing so that each segment is covered, and then we use crew rostering to assign individual members to these pairings. A variety of articles solve the problems of different countries such as Hong Kong [38] and New Zealand [28] using heuristics, a localized optimization model, or other IP-based models. In this section, we describe the model proposed by Vaidyanathan *et al.* [39] for US railroads.

We discuss an integer multicommodity flow model with side constraints for crew planning based on an underlying space-time network, $G(N, A)$, as illustrated in Fig. 6. For the sake of simplicity, we show a single crew type of one district and assume that one of the O–D pairs is the home terminal and the rest of the stations are away terminals. It can be easily extended to a general setting. We create four subsets of nodes (N) as follows:

- *Supply Nodes (SupNode)*: One for each crew with time attribute corresponding to the time at which it becomes available for assignment, while the place attribute corresponds to the terminal from which it is released for duty. These nodes have a supply of one unit.
- *Sink Node (SinkNode)*: A common sink node for all crews. Time attribute: end of planning horizon; no space attribute. Its demand equals the total number of supplied crews.
- *Departure Nodes (DepNode)*: One node for each train passing through the crew district. The space attribute is the first

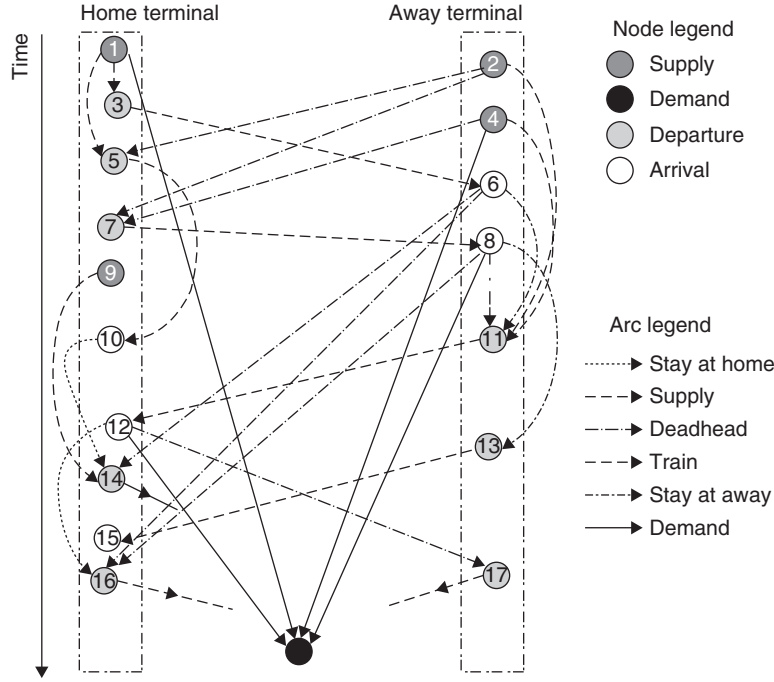


Figure 6. Simplified space-time network for crew planning.

departure station of the train in the crew district, and the time attribute is the on duty time of the train.

- **Arrival Nodes (ArrNode):** One node for each train passing through the crew district. The space attribute is the last arrival station in that crew district, and the time attribute is the tie-up time of the train.

The supply and sink nodes help to ensure the balanced flow of crews in and out of the network. We now create the arc-set, A , as follows:

- **Supply Arcs (SupArc):** Arcs from supply nodes to departure nodes representing the first assignment of a crew to a train.
- **Train Arcs (TrArc):** Each arc connects the *DepNode* and *ArrNode* of a train.
- **Deadhead Arcs (DhArc):** It connects an *ArrNode* or *SupNode* at one location to a *DepNode* at another location. It create only those arcs that satisfy all regulations.

- **Stay-at-Home Arcs (ShArc):** It connects an *ArrNode* or *SupNode* at the home location of a crew to a *DepNode* at the same location.
- **Stay-at-Away Arcs (SaArc):** It is same as *ShArc*, but for at all the away locations of a crew.
- **Demand Arcs (DeArc):** From all *ArrNode* and *SupNode* to the *SinkNode*.

Each arc has an equivalent cost based on its type, such as the crew wage, dead-head cost, or detention cost. All arcs are constructed to satisfy all contractual and regulatory requirements except for the FIFO constraints. The current network does not incorporate train delays caused by the unavailability of qualified crews; however, it can be handled by drawing additional arcs as explained in Vaidyanathan *et al.* [39]. It should be noted that regulations are often crew-pool specific, so on each arc only crews from a specific crew-pools can flow. Therefore, we formulate it as a

multicommodity flow problem in which each crew pool represents a commodity.

We need to find the flow, x_a , on each arc $a \in A$ with the objective (Eq. (34)) of minimizing all crew-related costs as mentioned earlier. Constraints (35) ensure that each train departs with a crew. Constraints (36–38) maintain a balanced flow throughout the network. n denotes the total number of available crew. Constraints (39) take care of FIFO constraints. A_a corresponds to the set of arcs that cannot have a positive flow if there is a flow on arc a . All other notations have been derived from earlier sections.

$$\text{Minimize } \sum_{a \in A} c_a x_a \quad (34)$$

$$\text{subject to } \sum_{a \in O(i)} x_a = 1 \forall i \in \text{DepNode} \quad (35)$$

$$\sum_{a \in O(i)} x_a = 1 \forall i \in \text{SupNode} \quad (36)$$

$$\sum_{a \in I(i)} x_a = n \forall i \in \text{SinkNode} \quad (37)$$

$$\sum_{a \in I(i)} x_a = \sum_{a \in O(i)} x_a \forall i \in \text{DepNode} \cup \text{ArrNode} \quad (38)$$

$$\sum_{a' \in A_a} x'_a \leq M(1 - x_a) \forall a \in A \quad (39)$$

$$x_a \in \{0, 1\}, \forall a \in A. \quad (40)$$

At a typical railroad most crew districts have two terminals, two to four crew types, and about 50 crews. A train schedule has approximately 500 trains running over the course of several weeks in a crew district. In this problem setting, the network has around 1050 ($50 + 2 \times 500$) nodes, 25,000 deadhead arcs, and 100,000 rest arcs. Each rest arc has one FIFO constraint, which makes the problem computationally intractable.

We observe that without FIFO constraints the model reduces to a simple minimum cost flow formulation, and the relaxed problem can be solved up to optimality within 5 s using CPLEX. On the basis of this observation, a two-phase approach has been proposed, whereby the first phase solves the relaxed problem and the second phase handles FIFO constraints in several ways. We suggest a

simple way to consider FIFO constraints and refer our readers to Vaidyanathan *et al.* [39] for other approaches.

Step 1: Solve the relaxed problem and find all FIFO violated constraints. If there are any violated constraints go to step 2, otherwise the current solution is optimal.

Step 2: Add violated FIFO constraints to the model and go back to step 1.

Conclusions

Crew planning is a highly complex problem because of the presence of numerous operational, contractual, and regulatory requirements. A typical railroad employs approximately 20,000 locomotive crews and the other supporting workers receiving a competitive salary. The proposed algorithm has been tested on real-world instances of a major US railroad, and our computational analysis shows a significant benefit in all parameters (crew-caused train delays, crew-deadheading, and crew detentions). The data shows the potential of a reduction in crew-related operating cost by improving the efficiency and effectiveness of the crews that use optimization tools.

SUMMARY

The railroad industry is a very rich source of optimization problems that are of significant economical impact. Numerous complex and conflicting operating rules, external regulations, and other requirements make it overwhelmingly difficult to design the models that can solve real-life instances in a reasonable time frame. We have grouped these intrinsically complex and interrelated problems into three major classes: train planning, locomotive planning, and crew planning. We have discussed the main challenges of each problem and provided some of the most effective solution techniques. The importance of these optimization tools has been shown by a computational study on real-world data, which indicates a total savings of billions of dollars each year.

In this article, we have emphasized the practical approaches to achieve near-optimal solutions, rather than focusing on exact approaches that generate optimal solutions in an unacceptable time frame. This is justified by two key facts: (i) in the highly dynamic real-world, planners generally need to make decisions within a few minutes; and (ii) an excessively large amount of data (input) collected for the optimization model includes some errors. Therefore, even an exact procedure will not generate ideal solutions. The approaches suggested are generic and can be applied to solve other large-scale problems after modifying the models accordingly. There are plenty of opportunities to develop more efficient and effective approaches to further improve the solution quality and response time.

REFERENCES

1. Ireland P, Case R, Fallis J, *et al.* The Canadian pacific railway transforms operations research by using models to develop its operating plans. *Interfaces* 2004;34(1):5–14.
2. Ahuja RK, Cunha CB, Sahin G. Network models in railroad planning and scheduling. *Tutorials Oper Res INFORMS* 2005;1:54–101.
3. Ahuja RK, Liu J, Sahin G. Railroads yard location problem. Working paper. Gainesville (FL): Innovative Scheduling Systems, Inc.; 2005.
4. Nemani AK, Bog S, Ahuja RK. Solving the curfew planning problem. *Trans Sci.* Available at <http://transci.journal.informs.org/cgi/content/abstract/trsc.1100.0323v1>.
5. Ahuja RK, Jha KC, Liu J. Solving real-life railroad blocking problems. *Interfaces* 2007;37:404–419.
6. Barnhart C, Jin H, Vance PH. Railroad blocking: a network design application. *Oper Res* 2000;48:603–614.
7. Bodin LD, Golden BL, Schuster AD, *et al.* A model for the blocking of trains. *Trans Res B* 1980;14:115–120.
8. Newton HN. Network design under budget constraints with application to the railroad blocking problem [PhD thesis]. Auburn (AL): Industrial and Systems Engineering Department, Auburn University. 1996.
9. Newton HN, Barnhart C, Vance PM. Constructing railroad blocking plans to minimize handling costs. *Trans Sci* 1998;32:330–345.
10. Ahuja RK, Magnanti TL. Network flows: theory, algorithms, and applications. New Jersey: Prentice Hall; 1993.
11. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123:75–102.
12. Assad A. Models for rail transportation. *Trans Res A* 1980;14:205–220.
13. Crainic TG, Rosseau JM. Multimode, multi-commodity freight transportation: a general modeling and algorithmic framework for the service network design problem. *Trans Res B* 1986;20(3):225–242.
14. Keaton MH. Designing optimal railroad operating plans - Lagrangian-relaxation and Heuristic approaches. *Trans Res B* 1989;23(6):415–431.
15. Keaton MH. Service-cost tradeoffs for carload freight traffic in the United States rail industry. *Trans Res A* 1991;25(6):363–374.
16. Keaton MH. Designing railroad operating plans - a dual adjustment method for implementing Lagrangian-relaxation. *Trans Sci* 1992;26(4):263–279.
17. Hallowell SF, Harker PT. Predicting on-time performance in scheduled railroad operations: methodology and application to train scheduling. *Trans Res Part A* 1998;32:279–295.
18. Gorman MF. An application of genetic and tabu searches to the freight railroad operating plan problem. *Ann Oper Res* 1998;78:51–69.
19. Sauder RL, Westerman WM. Computer aided train dispatching: decision support through optimization. *Interfaces* 1983;13:24–37.
20. Brännlund U, Lindberg PO, Nöu A, *et al.* Railway timetabling using Lagrangian relaxation. *Trans Sci* 1998;32(4):358–369.
21. Jovanovic D, Harjer PT. Tactical scheduling of rail operations: the SCAN I System. *Trans Sci* 1991;25:46–64.
22. Nou A, Desrosiers J, Soumis F. Weekly locomotive scheduling at Swedish state railways. Technical Report G-97-35. Quebec, Canada: GERARD, École des Hautes Études Commerciales de Montreal, Montreal. 1997.
23. Caprara A, Fischetti M, Toth P. Modeling and solving the train timetabling problem. *Oper Res* 2002;50(5):851–861.
24. Higgins A, Kozan E, Ferreira L. Optimal scheduling of trains on a single line track. *Trans Res B* 1996;30:147–161.
25. Higgins A, Kozan E, Ferreira L. Heuristic techniques for single line train scheduling. *J Heuristics* 1997;3:43–62.

26. Kraay D, Harker P. Real-time scheduling of freight networks. *Trans Res B* 1994;29(3):213–229.
27. Şahin G, Ahuja RK, Cunha CB. New approaches for the train dispatching problem. *Trans Res B* 2005. In press.
28. Cordeau JF, Toth P, Vigo D. A survey of optimization models for train routing and scheduling. *Trans Sci* 1998;32:988–1005.
29. Fischetti M, Toth P. A package for locomotive scheduling. Technical Report DEIS-OR-97-16. Bologna, Italy: University of Bologna; 1997.
30. Florian M, Bushell G, Ferland J, *et al.* The engine scheduling problem in a railway network. *INFOR* 1976;14:121–138.
31. Forbes MA, Holt JN, Watts AM. Exact solution of locomotive scheduling problems. *J Oper Res Soc* 1991;42:825–831.
32. Ahuja RK, Liu J, Mukherjee A, *et al.* Solving real-life locomotive scheduling problems. *Trans Sci* 2005;39:503–517.
33. Vaidyanathan B, Ahuja RK, Liu J, *et al.* Real-life locomotive planning: new formulations and computational results. *Trans Res Part B* 2008;42:147–168.
34. Transportation safety. Rail safety improvement act of 2008. 2008. <http://www.utu.org/safety/safetyact2008.htm>.
35. Caprara A, Fischetti M, Guida PL, *et al.* Solution of large-scale railway crew planning problems: the Italian experience. *Lect Notes Econ Math Syst* 1999;471:1–18.
36. Carrararesi P, Gallo G. Network models for vehicle and crew scheduling. *Eur J Oper Res* 1984;16:139–151.
37. Chu SCK, Chan ECH. Crew scheduling of light rail transit in Hong Kong: from modeling to implementation. *Comput Oper Res* 1998;25(11):887–894.
38. Cordeau JF, Soumis F, Desrosiers J. A Benders' decomposition approach for the locomotive and car assignment problem. Technical Report G-98-35. Quebec, Canada: GERARD, École des Hautes Études Commerciales de Montreal, Montreal; 1998.
39. Vaidyanathan B, Jha KC, Ahuja RK. Network flow-based approaches for the railroad crew scheduling problem. *IBM J Res Dev* 2007;51:325–344.

OR/MS APPLIED TO CRICKET

MIKE WRIGHT
Department of Management
Science, Lancaster
University, Lancaster, UK

There has been a steady trickle of academic papers over the past 20 years relating to the application of OR/MS to cricket. There are several reasons for this, including

- cricket's widespread popularity;
- the vast array of strategic and tactical decisions involved in playing cricket;
- cricket's discrete nature that lends itself easily to being modeled;
- the availability of enormous quantities of statistical data relating to professional cricket;
- the large amounts of money involved in professional cricket;
- the growing popularity of betting on cricket;
- the complexity of managing cricket competitions;
- the amenability of cricket timetabling and scheduling problems to OR/MS formulations and solutions.

Moreover, among professional practitioners, the game has recently become much more analytical. It is now commonplace to see not only administrators and managers but also players consulting their laptops in the pavilion during play. At the moment, this is probably mainly used for video analysis of their own play and that of their opponents, but it opens the way for the more widespread practical use of OR/MS implementations before, during, and after a match.

CRICKET'S RANGE AND POPULARITY

Cricket is played across the world, especially in English-speaking countries. Its popularity

is highest in the Indian subcontinent, where it is commonly referred to as a religion. In India, "the country comes to a stop when a cricket match is being played—the roads are deserted, parties and weddings are postponed, operations in hospitals are rescheduled, parliament goes in for early closing" [1]. In Sri Lanka, it was reported that the imminence of a new world record temporarily stopped a war [2].

Cricket is generally regarded as the world's second most popular sport. The number of players in the United Kingdom has been reliably estimated as 3.5 million [3]; with the population of the Indian subcontinent being about 30 times that of the United Kingdom, this suggests that the number of players worldwide must be well over 100 million. The number of professional players in England is about 450 [4]—the worldwide total is probably around 2000.

The number of cricket fans has been reported to be between 2 and 3 billion [5], and the largest TV audience, for the 2007 Twenty20 World Cup Final between India and Pakistan, has been reported variously as 400 million [6] and 1.4 billion [7]. Perhaps most surprisingly, the North American TV figures have been reported to be 37 million [7]—and North America is traditionally indifferent to cricket.

There is no doubt that cricket is of major importance to an enormous number of people and is thus an appropriate area for the serious application of the tools and approaches of OR/MS.

A BRIEF DESCRIPTION OF A CRICKET MATCH

Cricket is a sport played between two teams of 11 players on a large field. At any given time, one team is batting with two batsmen on the field, and the other is fielding with all 11 players on the field. One batsman is the "striker," who stands at one end of the "pitch" (a strip of well-rolled grass, matting, or other material 22 yards long) and receives a delivery (or "ball") bowled by the

“bowler” (one member of the fielding side) from the “nonstriker’s” end of the pitch; the other batsman stands at the nonstriker’s end. A “wicket keeper” stands behind the striker’s stumps and the other nine members of the fielding side are “fielders,” placed anywhere on the field (though there may be some restrictions) by the team captain.

The batsman wields a bat in order to hit the ball and thus score “runs.” When the ball is hit (or even if it is not) the batsmen may choose to attempt to score one or more “runs,” by running from one end of the pitch to the other, as a result of which they may change ends. “Boundaries” may also be scored by hitting the ball out of play—this counts for four runs if the ball hits the ground before going over the boundary, six otherwise.

The fielding side aims to get batsmen “out” along with trying to restrict the scoring of runs. There are 10 distinct ways of getting a batsman out, including “bowled” (where the ball hits one of three vertical pieces of wood known as “stumps”), caught (where the batsman hits the ball and a fielder catches it before it bounces), and “run out” (where the fielder hits the stumps with the ball before the batsman has completed his run). The dismissal of a batsman is known as a “wicket,” or the fall of a wicket. (Cricket terminology can be very confusing to the layman—for example, the pitch is often referred to as the wicket, and the three stumps are also often called wickets.)

Balls are bowled in batches of six (“overs”) by the same bowler, after which a different bowler must be used from the opposite end of the pitch. There may be limits to the number of overs any individual bowler may bowl.

Usually each team just has one “innings” in which they bat, but in some competitions each team has two innings. An innings ends when 10 wickets have fallen, but possibly also after a given number of overs (typically 20, 40, or 50), or after a time limit. In addition, it is allowable and sometimes wise for a team captain to declare his team’s innings “closed” early (see later). The team that scores the most runs wins, except that in some forms of the game a time limit may be overstepped before a result has been reached, in which case the game is a “draw.” If both sides

score the same number of runs, the match is a “tie”—this is quite unusual. A typical number of runs scored in an innings of a professional match may be as low as 100, but is frequently over 500 if there is no overs limit; the record low and high totals are 12 and 1107.

One-innings matches (i.e., where each team has one innings) may take anywhere between two and seven hours. Two-innings matches usually last between three and five days (for “Test” matches between national teams).

For the full laws of cricket, which cover every game of cricket anywhere in the world, see [8]. Every cricket competition also has its own “rules” in addition to the laws. For noncompetitive matches, the team captains agree the rules to be applied concerning, for example, the time or overs limit and any limits applying to individual batsmen or bowlers.

THE CONTRIBUTION OF OR/MS

OR/MS has been used to analyze decision making in cricket at many levels. It has had the most impact at the macro level, addressing issues of importance to the sport’s governing bodies and tournament organizers. However, there have also been several studies of strategic and tactical decision making and it is expected that such studies will have increasing practical influence in future.

The next three sections are therefore organized according to the level at which decisions are taken, starting with the areas where most impact has already been achieved.

DECISIONS FOR ADMINISTRATORS

How to Determine a Winner for a Shortened Match

Often play is deemed impractical or unsafe, usually because of heavy rain, a wet ground, or poor light. This situation is improving as more of the top grounds install improved covers and devices to remove surface water, and more install floodlights, but it is still

frequently impossible to play a match to its natural conclusion.

Often this is unproblematic: the match is declared a draw. However, in some circumstances, it is important that a winner be declared—this is especially the case in knock-out matches, where the winner proceeds to the next round. If there has been a reasonable amount of play, then it may be possible to determine a winner even though the normal winning criteria may not have been met.

Several different methods were originally tried for this, but all led to very unjust outcomes at times. However, two OR analysts devised a system [9] which is now used throughout the world and accepted as being fair by virtually all professionals, though probably few of them understand its intricacies. Given professional cricket's international following, it seems certain that the names of Duckworth and Lewis are far better known than those of any other OR/MS analyst in any field.

The precise details are quite complex, and the model requires calibration using real data, but the main feature of the Duckworth/Lewis method involves the consideration of wickets left and overs remaining as resources; thus, for example, when the number of overs remaining for the side batting second changes because of an interruption, the revised number of runs needed to win depends upon the overall resources available before and after the interruption.

A useful potential by-product of this system is that it can also determine putative winning margins [10] to be used as a tiebreakers for teams with identical win records in a league.

Other methods have been devised [11] which aim to preserve win probabilities. While it is arguable that these may be superior methods, the cricket authorities are very unlikely to consider discarding Duckworth/Lewis.

Fixture Scheduling

For some competitions, scheduling fixtures is easy, using fixed patterns. However, often a more analytical approach is needed using computer algorithms. It may sometimes

be appropriate to use theoretical models [12], but often the situation is too complex for anything other than a custom-built solution.

Sometimes, it must be decided which teams are to play against each other, where, and how often, which can involve issues of fairness [13]. However, it is usually preordained which matches are to be scheduled; just the dates must be decided.

The most extreme case concerns English professional cricket. Eighteen county teams take part in four different competitions—three one-day, one four-day—with different formats. The competitions overlap throughout the season; a team may frequently finish a match in one competition one day and start another match in a different competition the next day.

There are many constraints and objectives; also, the teams have strong preferences that are far from identical for all counties. Travel considerations are important, especially late evening travel, as is the overall pattern of home matches. Teams do not like long gaps between matches; there are matches against touring international teams to fit in; some pairs of teams do not like to have home matches simultaneously; many teams have traditional “festivals,” often at holiday resorts, for which the dates possible are usually very restrictive; often there are also dates to avoid if possible, such as when other sporting fixtures are involved, where the ground is required for a concert, etc.

This problem was successfully approached [14] using a variant that takes advantage of information derived from each separate objective as well as the overall objective.

Other successful implementations have been carried out for the World Cup [15] using a mixture of human decision making and integer programming (see the section titled “Combinatorial Optimization/Integer Programming” in this encyclopedia), for Australian professional cricket [16] using simulated annealing, for New Zealand professional cricket [17] using a mixture of simulated annealing and tree search, and for an amateur league in England [18] using a partly manual heuristic (see the section

titled “Heuristics and Metaheuristics” in this encyclopedia).

Umpire Scheduling

Every match requires two umpires out in the field. For televised matches, there is usually another umpire who analyzes replays of incidents to help the two in the field with difficult decisions. Moreover, for Test matches, there is a fourth umpire mainly acting as a reserve in case another umpire falls ill.

Scheduling these umpires can be difficult. A successful implementation for English professional cricket [19] considered a variety of constraints and objectives and solved the problem using local search, though more recently an improved simulated annealing procedure has been used. Objectives and constraints included journey time, balancing various types of matches, limiting consecutive days worked and spreading umpires out between the teams and each other.

This can also be difficult in amateur leagues; in one implementation [20], an important consideration was the value of one umpire giving a lift to the other, so as to save money on petrol.

STRATEGIC DECISIONS

Strategic decisions so far addressed by OR/MS analysts include team selection, batting order, whether or not to bat first, choosing a “night watchman” and whether and when to declare an innings closed. However, there are plenty of other important decisions that have not yet been studied.

Team Selection

This is an issue about which every fan of any sport has strong views, but very little headway has been made by OR/MS. There have been attempts [21] to devise measures of performance that measure a player’s actual or potential contribution more accurately than those prevalently used, but there is much more work to be done. For example, highly sophisticated techniques have been developed to select football teams [22] and these are widely used among top clubs in the

Netherlands. This is a target that OR/MS analysts could aim at for cricket.

Batting Order

Normally, the best batsmen go in early in a team’s innings and the less able batsmen come in later. This is partly because it is wasteful to have a good batsman stranded if all 10 wickets have fallen, and also because not all batsmen may be needed—this is especially important in games with a low over limit. One study showed how making assumptions for transition probabilities for each batsman separately, any potential order can be simulated [23]; permutations of possible orders were generated using simulated annealing in order to find an order close to optimal.

A batting order does not have to be decided upon at the start of an innings, and a Dynamic Programming (DP) study [24] has shown that captains should make dynamic decisions depending on the state of the game. Another study [25] has considered situations where a day’s play is nearly over when a wicket falls, for a team whose innings will continue the following day. In such circumstances, some captains like to send in a “night watchman,” who is good at defence but poor at scoring runs, to see out the day’s play. This may be useful because any batsman, even a good one, is especially vulnerable at the start of his innings or at the start of a day’s play, and such a batsman would then have to make two “starts.” Some captains do not like to risk this for their good batsmen, whereas other captains think this unimportant and send in their better batsmen ahead of their weaker batsmen in all circumstances. The study’s results came down between these positions, concluding that in some circumstances a night watchman should be used and in other circumstances he should not.

Whether to Bat or Field First

Before the start of every match the captains toss a coin; the winner of the toss decides whether to bat or field first. The captain makes his decision based on factors including the current and predicted future condition

of the weather and the pitch, plus perhaps factors relating to particular players. The toss is regarded as an important part of the game and pundits often put forward opposing views as to the right decision. Where it is not clear-cut, a toss can be regarded as a “good toss to lose,” on the grounds that the captain cannot later be blamed for making the wrong decision.

Regression models [26, 27] have been used to analyze captains’ choices. Results appear to indicate that for one-day matches it is generally better to field first unless the match is due to finish under floodlights, in which case it is preferable to bat first. However, in Test Matches, it appears to be usually advantageous to field first. These results are surprising as most toss winners choose to bat first in all forms of the game.

When to Declare

Where the length of an innings is not limited by a number of overs, but instead there is an overall time limit, a captain may be faced with the decision whether and when to terminate his team’s innings early, or “declare.” The timing of a declaration is often very difficult. Too early and he may be giving the opposition too good a chance of winning. Too late and a draw may be almost inevitable—there will be insufficient incentive for the opposition to take risks, and not enough time for his bowlers to dismiss them.

This has been analyzed using a multinomial logistic regression model [28], and a decision tool has been created with the potential to help second innings declaration decisions in Test matches. Declarations are also possible in a team’s first innings—this again is an important decision but rather more difficult to analyze. Further OR/MS work would be very helpful here.

TACTICAL DECISION MAKING

At all times during his opponent’s innings, a fielding captain must decide where to place the fielders, normally in consultation with the bowler. His decisions are influenced by the prowess of the batsman, the bowler, and the fielders, by the state of the pitch, the weather,

and the state of the game. Professional captains have become more inventive with their field placings recently, making use of video analysis portraying the opposing batsman’s strengths and weaknesses. In tense situations, a captain may make changes every ball, but often the positions remain the same throughout an over.

The bowler decides what kind of ball to attempt to bowl. Variables include ball speed, the location where the ball will bounce (if it does bounce), whether to apply swing, whether to induce lateral movement off the seam of the ball, whether to spin the ball and if so what type of spin—leg-spin, off-spin, googly, top-spin, “doosra,” and possibly others. The more ambitious the intention, the less likely even a top-class bowler is to achieve exactly the desired effect, and most bowlers specialize in particular types of deliveries (e.g., fast swing, slow spin).

The batsman has a split second to react to the ball that is hurtling toward him at up to 100 miles/hour. Not only does he have to decide whether to hit the ball or leave it, the types of shots he can play include the glance, glide, square cut, late cut, upper cut, pull, hook, cow-shot, slog, straight drive, on-drive, off-drive, cover drive, square drive, sweep, reverse sweep, slog sweep, paddle sweep, switch-hit, forward defensive, backward defensive, prod, nudge, nurdle, flick, scoop and possibly others—every now and then, an unexpected new shot is invented. He must also decide how hard to hit it and with what degree of elevation. His decision depends upon the pace, line (direction), length (where it bounces), height of bounce, spin, swing, and so on of the delivery, upon his own strengths and weaknesses, and upon the state of the game—whether he needs to play aggressively or defensively. When he has hit it, or even if he has not, he and his partner need collectively to decide whether to attempt one or more runs.

While there is no time for rational real-time analysis and batsmen react using a mix of reflexes, skill, and experience, there are plentiful opportunities for rational OR/MS interventions in terms of the approach adopted by batsmen. Two examples out of many are discussed below.

Balance between Attack and Defence

One of the most fundamental decisions for a batting team concerns the balance between attack and defence. An aggressive approach will score runs more quickly, but at a greater risk of dismissal. Conversely, with a defensive approach, a batsman may well stay in longer, but his runs will come more slowly.

The optimal balance depends upon the state of the match—the number of overs remaining, the number of batsmen out, and, for the team batting second, the number of runs still needed for victory. (This assumes that the match is of one innings and that there is a fixed limit to the number of overs available. The situation is a little more complex where there is no such limit, especially for two-innings matches, but the same general principles apply.)

This issue has been approached [29] by first creating a model of a limited-overs match, based on stochastic DP. A state of the system is defined by the number of runs required (for the team batting second), the number of wickets lost, and the number of balls remaining to be bowled. Using data from actual matches, probabilities are assumed for the following events: one wicket and no run; neither wicket nor run; one run; two runs; three runs; four runs; and six runs. (This does not quite cover all eventualities: for example, it is possible for runs to be scored and a wicket to fall at the same ball, usually by means of a run out; it is possible to score five or even seven runs; and some deliveries may be “wides” or “no balls,” resulting in an extra run but not counting toward the total number of balls to be bowled. However, it is adequate for the study in question.)

These probabilities vary depending upon the approach taken; under a very aggressive approach, the probability of a wicket is high, the probability of scoring runs is high, and the probability of neither run nor wicket (known in cricket parlance as a “dot ball”) is very low; conversely, with a very defensive approach, the probability of a wicket is low, of scoring runs is low, and of a dot ball is high.

The study calculated the optimal scoring rate to aim for, from any given state of the system, in order to maximize the number of

runs scored for the team batting first, or the probability of victory for the team batting second. The results showed that the prevalent strategy of starting slowly and then speeding up was suboptimal, and that a more aggressive attitude should be adopted right from the start. Subsequent matches have supported this conclusion, and teams now usually show some aggression right from the start.

This analysis has been extended [30] using larger data sets and with variable probabilities—reasonable enough because not all batsmen or bowlers are of equal ability, and probabilities can also vary significantly with pitch and weather conditions. This can be used as a simulator (see the section titled “Simulation Modeling and Analysis” in this encyclopedia) to answer all manner of strategic and tactical questions.

A Strong Batsman Protecting a Weaker One

An intriguing tactical battle often occurs when nine wickets have fallen, if a specialist batsman is accompanied by a much less proficient batsman (probably selected in the team for his bowling). It is up to the good batsman to score runs, as the other batsman is unlikely to score many, and if the weak batsman faces the bowling he has a high probability of being out, leaving the good batsman stranded; even though he is not out himself, the team will have lost 10 wickets and thus the innings ends. So he will aim to face most of the balls bowled.

As the bowling changes ends every over, his ideal is to score an even number of runs from each of the first five balls and an odd number on the sixth ball—generally a single run, as it is very hard to score exactly three runs on purpose. Thus, he would face every ball. However, he cannot guarantee that the runs will come exactly as he wishes; and if he fails to score a single on the final ball, the weak batsman may then have to face all six balls of the next over.

DP analysis [31] has concluded that the strong batsman should normally refuse to take an odd number of runs on the first four balls, but should accept a single run on the fifth or sixth ball, if nine wickets have fallen; otherwise, both batsmen should take every run on offer.

However, this is also a tactical issue for the fielding side. If the strong batsman is on strike at the start of an over, the captain may place the fielders toward the boundary edge, such that it is easy for the batsman to score a single, but very hard for him to run two safely or to pierce the field for a boundary four. He could try to hit the ball over the top of the fielders for a six, but this is often very risky—he could easily be caught out.

This creates a game of “cat and mouse,” where for the first four balls the batsman may hit the ball to a faraway fielder and refuse to run, and for the fifth and sixth balls the captain will bring all the fielders in much closer to minimize the probability of a single. Batsmen may react by accepting a single off the fourth ball, or by trying to hit the ball over the fielders on the fifth or sixth, perhaps happy to accept a boundary four or six as compensation for not being the facing batsman at the start of the next over. The situation could be approached using game theory analysis (see the section titled “Decision and Risk Analysis and Game Theory” in this encyclopedia).

FORECASTING AND BETTING

This is a growing area for study and it is probable that much OR/MS analysis has been undertaken confidentially by betting companies. However, an interesting study [32] has been published using not only regression modeling but also the Duckworth/Lewis analysis. This considered betting during the course of a match, concluding that punters tend to overreact to important events (such as wickets falling).

CONCLUSION

Cricket is a very fruitful area for OR/MS study, with a plethora of decisions to be analyzed, and a fast-growing data bank to help the analysis. Nowadays, for professional matches, a substantial amount of data is recorded for every ball, including field placements, the type of delivery, the type of shot played, and the outcome. Moreover, issues for administrators get ever more complex as new

types of tournament are created. We can look forward to many more papers on the subject.

REFERENCES

1. <http://news.bbc.co.uk/1/hi/programmes/3734-038.stm>. Accessed on August 20th 2009.
2. <http://www.cricinfo.com/columns/content/story/255212.html>. Accessed on August 20th 2009.
3. http://timesonline.typepad.com/line_and_length/2009/03/here-come-the-g.html. Accessed on August 20th 2009.
4. <http://www.independent.co.uk/sport/cricket/henry-blofeld-surfeit-of-counties-leaves-england-chasing-their-tail-579197.html>. Accessed on August 20th 2009.
5. <http://ezinearticles.com/?Most-Popular-Sports-Around-The-World&id=551180>. Accessed on August 20th 2009.
6. <http://blog.livecurrent.com/archives/102-T20,-IPL-among-biggest-sports-stories-of-2008.html>. Accessed on August 20th 2009.
7. <http://news.bbc.co.uk/sport1/hi/cricket/73403-31.stm>. Accessed on August 20th 2009.
8. <http://www.lords.org/laws-and-spirit/laws-of-cricket/>. Accessed on August 20th 2009.
9. Duckworth FC, Lewis AJ. A fair method for resetting the target in interrupted one-day cricket matches. *J Oper Res Soc* 1998;49(3): 220–227.
10. de Silva BM, Pond GR, Swartz TB. Estimation of the magnitude of victory in one-day cricket. *Austral New Zeal J Statist* 2002; 43(3):259–268.
11. Carter M, Guthrie G. Cricket interruptus: fairness and incentive in limited overs cricket matches. *J Oper Res Soc* 2004;55(8): 822–829.
12. Easton K, Nemhauser GL, Trick M. In: Leung JT, editor. *Handbook of scheduling: algorithms, models and performance analysis*. Boca Raton: CRC Press; 2004. pp. 52.1–52.19.
13. Wright MB. A fair allocation of county cricket opponents. *J Oper Res Soc* 1992;43(3): 195–201.
14. Wright MB. Timetabling county cricket fixtures using a form of tabu search. *J Oper Res Soc* 1994;45(7):758–770.
15. Armstrong J, Willis RJ. Scheduling the cricket World Cup—a case study. *J Oper Res Soc* 1993;44(11):1067–1072.

16. Willis RJ, Terrill BJ. Scheduling the Australian state cricket season using simulated annealing. *J Oper Res Soc* 1994;45(3): 276–280.
17. Wright MB. Scheduling fixtures for New Zealand Cricket. *IMA J Manag Math* 2005; 16(2):99–112.
18. Johns S. Complexity in amateur sports scheduling. Presented at the Operational Research Society Conference, Bath, 2001.
19. Wright MB. Scheduling English cricket umpires. *J Oper Res Soc* 1991;42(6):447–452.
20. Wright MB. Case study: problem formulation and solution for a real-world sports scheduling problem. *J Oper Res Soc* 1997;58(4):439–445.
21. Barr GDI, Kantor BS. A criterion for comparing and selecting batsmen in limited overs cricket. *J Oper Res Soc* 2004;55(12): 1266–1274.
22. Sierksma G. Computer support for coaching and scouting in football. *Sports Eng* 2006;9(4):229–249.
23. Swartz TB, Gill PS, Beaudoin D, deSilva BM. Optimal batting orders in one-day cricket. *Comput Oper Res* 2006;33(7):1939–1950.
24. Norman JM, Clarke SR. Optimal batting orders in cricket. *J Oper Res Soc*. 2010;61(6):980–986.
25. Clarke SR, Norman JM. Dynamic programming in cricket: choosing a night watchman. *J Oper Res Soc* 2003;54(8):838–845.
26. Allsopp PE, Clarke SR. Rating teams and analysing outcomes in one-day and Test cricket. *J R Statist Soc A* 2004;167(4): 657–667.
27. Dawson P, Morley B, Paton D, Thomas D. To bat or not to bat: an examination of match outcomes in day-night limited overs cricket. *J Oper Res Soc* online DOI: 10.1057/jors.2008.135.
28. Scarf P, Shi X. Modelling match outcomes and decision support for setting a final innings target in Test cricket. *IMA J Manag Math* 2005;16:161–178.
29. Clarke SR. Dynamic programming in one-day cricket—optimal scoring rates. *J Oper Res Soc* 1988;39(4):331–337.
30. Swartz TB, Gill PS, Muthukumarana S. Modelling and simulation for one-day cricket. *Can J Statist* 2009;37(2):143–160.
31. Clarke SR, Norman JM. To run or not to run? Some dynamic programming models in cricket. *J Oper Res Soc* 1999;50(5): 536–545.
32. Bailey M, Clarke SR. Predicting the match outcome in one day international cricket matches while the game is in progress. *J Sports Sci Med* 2006;5:480–487.

ORBEL, THE BELGIAN OPERATIONS RESEARCH SOCIETY (SOGESCI-BVWB)

PIERRE L. KUNSCH
Université Libre de Bruxelles,
SMG-CODE Department,
Brussels, Belgium

Vrije Universiteit Brussel, MOSI
Department, Brussels, Belgium

THE BELGIAN OR SOCIETY

SOGESCI-BVWB, the Belgian Operations Research Society, today most commonly called ORBEL, is the Belgian society for the promotion of scientific methods of management. Its goal is to contribute to the diffusion of new tools, methods, and knowledge in the fields of operational research, mathematical programming, statistics, pattern recognition, and related disciplines. It also aims at stimulating research in the above fields as well as developing the cooperation between researchers and with the industry and business, both at national and international levels.

The society today has about 120 members. It represents Belgium in a number of international associations and federations of national societies including

- EURO, the European Association of Operational Research Societies;
- IFORS, the International Federation of Operational Research Societies.

SOGESCI-BVWB also coedits the journal 4OR: a quarterly journal in operations research published by Springer, along with the French and Italian OR societies.

The SOGESCI-BVWB website is at www.orbel.be.

HISTORY

SOGESCI-BVWB started as a Belgian nonprofit organization in today's form in January 1962, as a merger between the "Belgian Society for the Industrial Application of Statistics" (ABSI) founded in Brussels in 1957 and the "Belgian Association for the Application of Scientific Methods of Management" (AGESCI) founded in Brussels in 1958. Starting with the mergers of two associations founded in 1957–1958, SOGESCI-BVWB may claim over 50 years of existence.

Belgium being a trilingual country with French and Dutch as the two main languages, the new society needed a composite name.

SOGESCI is the French acronym for "Société Belge pour l'application des Méthodes Scientifiques de Gestion ASBL," while BVWB is the Dutch acronym for "Belgische Vereniging Voor de Toepassing van Wetenschappelijke Methoden in het Bedrijfsbeheer." No official German acronym was provided.

To make things easier it has been—unofficially—decided for some years to use the name of the annual national conference name of ORBEL, standing for "Operations Research in Belgium," as a convenient acronym for common use between members.

The list of past presidents going back to 1980 from today is as follows:

F. Broeckx 1980–1984
J. Teghem 1984–1986
M. Despontin 1986–1988
Ph. Vincke 1988–1990
H. Pastijn 1990–1992
M. Roubens 1992–1994
F. Broeckx 1994–1996
M. Pirlot 1996–1998
D. Cattrysse 1998–2000
P. Kunsch 2000–2002
F. Plastria 2002–2004

E. Loute 2004–2006
 G. Janssens 2006–2008
 R. Bisdorff 2008–2010
 P. De Causmaecker 2010–2012

Several members of SOGESCI-BVWB are very active in EURO and IFORS and they have occupied important positions therein, such as the following:

Jean-Pierre Brans, former professor at Vrije Universiteit Brussel (VUB), and SOGESCI-BVWB president 1974–1976, has been a cofounder of EURO, organizing the first EURO I conference in Brussels in 1975. He was president of EURO in 1983–1984, and received the EURO Gold Medal in 1994. He was also IFORS vice president in 1977–1980 and 1989–1992.

Philippe Vincke, Professor at Université Libre de Bruxelles (ULB), its rector in 2006–2012, and SOGESCI-BVWB president 1988–1990, was EURO president in 2001–2002.

Martine Labbé, Professor at ULB, was EURO president 2007–2008.

Jacques Teghem, professor at Faculté Polytechnique de Mons (FPMs), and SOGESCI-BVWB president 1984–1986, was editor of EJOR, the European Journal of Operational Research of EURO till 2007.

Raymond Bisdorff professor at Luxembourg University, and SOGESCI-BVWB president 2008–2010, was EURO vice president in 1998–2000. Note that, for the first time in the history of ORBEL, a non-Belgian citizen was elected for the ORBEL presidency due to the very close links between OR researchers in Luxemburg and Belgium.

ORGANIZATION

The Board of Administration gathers all administrators of SOGESCI-BVWB. Each mandate is renewed every two years.

The Executive Committee emanating from the Executive Committee consists of the president, the three last past presidents acting

as vice presidents, the president-elect, the treasurer, the secretary, the delegate administrator, the webmaster, and administrators responsible for the newsletters, or for the organization of the ORBEL activities.

The meetings of the General Assembly usually take place at the occasion of the annual ORBEL conference.

Meetings of the executive committee and of the board of administration take place at the headquarters of SOGESCI-BVWB located at the Royal Military Academy (RMA) in Brussels. All meetings are held in English to simplify administrative matters. English is also the official language of the annual ORBEL conference.

The RMA has traditionally hosted in its premises, the annual national conference ORBEL, which was biannual till 1990; it is organized by an academic chairperson, the post being a rotating one. For many years now, the annual, typically two-day, ORBEL conference takes place alternately in a French-speaking or Dutch-speaking university in the country. The 50th anniversary of SOGESCI-BVWB was celebrated in January 2008 at RMA at the occasion of the 22nd ORBEL conference. Some speakers from the early days of the society were invited for delivering talks about the history of our society.

SERVICES TO MEMBERS

The Society offers members a number of services to its member including the following ones:

- free subscription to the journal 4OR;
- organization of annual conferences and workshops;
- awarding of the SOGESCI-BVWB master's thesis award;
- an email newsletter about OR-related calls for papers, jobs, books, seminars, and conferences.

There are privileged links to many other national societies which are members of EURO, such as

- Italian Association for Operations Research (AIRO), Italy;

- Australian Society for Operations Research (ASOR), Australia;
- German Operations Research (GOR) Society, Germany;
- Danish Operations Research Society (DORS), Denmark;
- French Society of Operations Research and Decision Support (ROADEF), France;
- Nederlands Genootschap van Besliskunde (NGB), The Netherlands;
- Institute for Operations Research and the Management Sciences (INFORMS), United States;
- Operations Research Society (ORS), United Kingdom.

Furthermore, there are links to the Belgian Statistical Society (BSS), the Belgian Engineers Societies FABI (French speaking), and KVIV (Dutch speaking).

THE ORBEL AWARD

The ORBEL award, organized by ORBEL, is bestowed on the author(s) of a scientifically oriented master's thesis.

The criterion for judgment is the appropriateness of the scientific approach applied to solve the problem under study. The author(s) should emphasize the application of existing methods in new areas or the application of new methods of dealing with models recognized as useful for solving real-life management problems. Therefore, actual collaboration with industry during the thesis elaboration is one of the selection criteria.

Examples of recently awarded master theses are as follows:

- Measures of congestion in container terminals.
- Matrix approximation and under approximation by nonnegative factorization: algorithms, complexity and applications.
- A discrete-time queueing model with batch service.
- Fire station location models and applications in Belgium.

CONFERENCES AND WORKSHOPS

Although the big majority of ORBEL members are academics, very many efforts have been made to attract more and more practitioners from the industry, especially for the annual conferences or topical workshops. These scientific events are understood by our society as being meeting places between academia and industry.

ORBEL conferences are organized every year, in January, and bring together most of the people concerned with OR and quantitative techniques for decision making in Belgium; they also attract a number of specialists from the neighboring countries.

Young researchers and PhD students are invited to attend the ORBEL conference at a reduced registration fee which also includes a one-year membership to SOGESCI-BVWB.

The ORBEL conferences consist of invited lectures delivered by prominent foreign specialists and practitioners on topics of general interest; more specialized contributed papers are presented in parallel sessions by young researchers.

The winner of the ORBEL award for the best Master's thesis in OR is also announced at the conference and the student is invited to present his/her work there.

The last 23rd ORBEL conference took place at Leuven University (KUL) in 2009 with more than 120 participants. ORBEL 24 will take place at Université de Liège (ULg) in 2010.

Workshops on a particular theme are also organized. These one-day meetings are devoted to a single topic; the invited lectures are selected in order to provide an overview of recent and important developments in some fields of interest.

In recent years, workshops have been devoted to discrete-event simulation, fuzzy logic and constraint programming, data mining and operations research, current research in location theory, and optimization in engineering.

THE 4OR JOURNAL

The journal "4OR, A Quarterly Journal of Operations Research" started in 2003. This

refereed joint journal of the three national OR societies of Belgium, France, and Italy is offered free to all their members. It is published by Springer, and available on-line. It will be listed by the ISI, starting from 2010. Professor F. Plastria of the Vrije Universiteit Brussel, a former ORBEL president, has been editor-in-chief representing the Belgian society from the beginning till the end of 2009.

The former journal JORBEL stopped after the last volume 41 (2001). This refereed quarterly journal published scientific contributions; special issues were devoted to

particular topics and contained a review or tutorial paper. The most recent special issues were centered around general heuristics, production management, and constraint programming; a collection of tutorial lectures presented at the Francophone Operations Research (FRANCORO) conference in 1996 also appeared.

The Borsalino project aimed at having the complete archives of JORBEL freely available on-line. It is now completed, and all articles from the journal are today available from the ORBEL website www.orbel.be.

ORSP: FORGING AHEAD IN ITS TWENTY-FIFTH YEAR

ELISE DEL ROSARIO

Formation of the national OR Society of the Philippines (ORSP) was an initiative of the private sector with the full support of the academe. The San Miguel Corporation's Operations Research Department (awarded the 1992 ORSA prize for consistent and sustained OR/MS support to decision making at all levels of the organization) hosted a dinner dubbed "A Toast to OR Professionals" in December 1986. This led to a frenzy of activities that culminated in the official recognition of the ORSP in February 1987.

The very first symposium of this organization that featured OR applications in various sectors metamorphosed into quarterly technical forums on issues such as traffic, agriculture, environment, energy, business logistics, health, sustainable development, and finance. Regular workshops on the use of spreadsheets and various simulation software were held. Its annual general membership meeting coincides with the annual national conference in which professionals from all over the country discuss directions for the society, present their technical papers, and get updated on local and international OR developments.

ORSP became a member of the International Federation of Operational Research Societies (IFORS) in 1990. In 2007, one of its founders, Elise del Rosario, assumed the presidency of the international body. ORSP organized its first international conference, the 1990 International Conference on OR/MS (ICORMS) in the same year in which the first issue of the *Philippine Journal of Operations Research* appeared (currently published online in its web site, www.orsp.org.ph; the *ORSP Newsletter* is also available in this web site). In 1997, ORSP hosted the second ICORMS jointly with the IFORS' annual International Conference on OR in Development (ICORD), followed in 2004 by the

IFORS Workshop for Teachers in Developing Countries. In 2006, ORSP hosted the Association of Asia Pacific OR Societies (APORS) seventh conference. In this meeting, Philippine President Gloria Arroyo acknowledged the contribution of OR in development and challenged ORSP to help the government in some of its programs.

At this time, the ORSP Committee on OR for Public Service (ORSP Corps) had been well established. With a mission to help in nation building through the use of OR, the committee has helped government agencies such as Philippine Ports Authority, Bureau of Customs, Commission on Elections, and National Power Corporation since 1999. Before this, the organization had links with the Department of Science and Technology, which sponsored its forums on public sector applications, and had two Philippine presidents grace its Annual General Membership meetings.

ORSP does not lose sight of the future OR professionals: the students. Early on, ORSP laid down the foundation for the formation of student chapters and inducted its first one in 1989. It has since accredited 19, who comprise the ORSP Federation of Student Chapters. The Federation actively sponsors get-togethers, and an annual student congress, lectures, interuniversity OR quiz, and paper writing contest.

It can indeed be concluded that since its founding in 1987, ORSP has remained at the forefront of promoting the advancement and practice of OR in the Philippines for 25 years.

Table 1 shows that ORSP has seen 14 presidents in the 25 years it has been in existence. The president is joined by 10 board directors including a vice president, secretary, and treasurer, who are elected for a term of two years with a maximum two-term limit.

Table 1. The ORSP Presidents

Year	ORSP President
1987–1988	William Torres
1988–1990	Elise del Rosario
1990–1992	Felipe Fondevilla
1992–1994	Manuel Agustin
1994–1996	Olegario Villoria Jr
1996	Elvira Zamora
1997–1998	Aura Matias
1998–1999	Emmanuel Macalalag
1999–2000	Wilson Wy Tiu
2000–2002	JC Mercado
2002–2004	Vicente Reventar
2004–2006	Elise del Rosario
2006–2008	Alleli Domingo
2008–2010	Alleli Domingo
2010–2012	Francis Miranda

OVERWEIGHTING OF SMALL PROBABILITIES

ZACH BURNS

ANDREW CHIU

GEORGE WU

University of Chicago Booth School of
Business, Center for Decision
Research, Chicago, Illinois

Consider the decisions of whether or not to undergo heart surgery, evacuate a city when a tropical storm is approaching, or increase counterterrorist measures. These decisions involve consideration of both the likelihood and the impact of surgical death, a Category 3 hurricane, or a September 11-magnitude terrorist attack, respectively. Assessing the likelihood of these events is difficult, with experts often disagreeing on the chance that the relevant event will occur [1]. Rare events are also important when they involve extreme consequences, whether good or bad. Indeed, much of risk and policy analysis is devoted to the study of low probability, high consequence events, such as a nuclear accident or a terrorist attack [2].

Research in judgment and decision making suggests that decision makers do not weight rare events according to their actuarial chances of occurring. Instead, small probability events tend to be overweighted for two reasons: (i) decision makers tend to overestimate the chance that rare events will occur; and (ii) small probabilities are overweighted in terms of their impact on decisions. Combined, these two pieces suggest that rare events are given greater psychological weight in our minds than is normatively appropriate.

Although both parts contribute to the overweighting of small probability events, it is nevertheless important to distinguish between these two separate processes. Events with a low probability of occurring may be overestimated, or perceived as more likely than they actually are. For

example, someone considering a move to San Francisco may overestimate the probability that the Bay Area will suffer from an earthquake in the next 5 years. However, small probability events can loom large in decisions, even when these events are not overestimated. For example, consumers find lottery tickets attractive, even though they may be very much aware of the low chances of winning [3,4]. In this article, we review empirical findings from judgment and decision making research that illustrate how these two stages contribute to the overweighting of small probability events.

A TWO-STAGE FRAMEWORK

In the classical theory of decision under uncertainty, subjective expected utility theory, the utility of each outcome is weighted by the decision maker's subjective probability that this outcome will obtain [5,6]. Consider a prospect, $A = (E_1, x_1; \dots; E_i, x_i; \dots; E_n, x_n)$, that offers outcome x_i if event E_i occurs, where events E_i are mutually exclusive and collectively exhaustive. Subjective expected utility evaluates the prospect A as follows,

$$U(A) = \sum_{i=1}^n P(E_i)u(x_i),$$

where the utility function, $u(x_i)$, measures the value the decision maker attaches to outcome x_i , and the subjective probabilities, $P(E_i)$, capture the decision maker's assessment of the likelihood of event E_i and satisfy the basic axioms of probability.

Although subjective expected utility is a standard building block for economic analysis [7], a large number of experimental studies have found that people are inconsistent with this model: subjective probability judgments violate the basic axioms of probability [8], and preferences are inconsistent with the axioms

underlying expected utility theory [9,10]. Of the numerous alternative models that have been proposed [11], prospect theory is consistent with the broadest set of empirical observations critiquing expected utility theory, such as the common-consequence effect (e.g., the Allais Paradox), the common-ratio effect, the fourfold pattern of risk preferences, and the simultaneous attraction of lottery tickets and insurance (see *Prospect Theory and Paradoxes and Violations of Normative Decision Theory*) [3,10,11]. Prospect theory uses a probability weighting function, $\pi(p)$, to capture the distortion of probabilities (see *Probability Weighting Functions*), and a value function, $v(x)$, to account for the evaluation of outcomes.

We extend prospect theory using a two-stage framework proposed by Tversky and Fox [12,13]. Consider a prospect that offers \$100 if a thumbtack is flipped point down three times in a row (and \$0 otherwise). The two-stage process suggests that the weight assigned to the outcome \$100 reflects: (i) a decision maker's *subjective probability* that a thumbtack will be flipped point down three consecutive times; and (ii) a distortion of the subjective probability that captures the impact of that probability on the attractiveness of the prospect. This framework allows for bias at each of the two stages. More formally, consider a prospect that pays x if event E occurs (and 0 otherwise), denoted (E, x) for simplicity. The weight assigned to outcome x is $W(E) = \pi(P(E))$, where $P(E)$ captures stage 1 (the judgment of the likelihood of E) and $\pi(P(E))$ captures stage 2 (the distortion of $P(E)$) as measured by $\pi(p)$, the probability weighting function. Overestimation of rare events occurs if $P(E)$, the judged likelihood of event E , is higher than the actuarial probability of E , whereas overweighting of probability occurs if $\pi(P(E)) > P(E)$.

In the section titled "Stage 1: Overestimation of Rare Events," we discuss overestimation in likelihood judgments, and identify some psychological factors that contribute to this phenomenon. We then discuss overweighting in decisions involving small probabilities in the section titled "Stage 2: Overweighting of Small Probabilities in Decision Making."

STAGE 1: OVERESTIMATION OF RARE EVENTS

Overestimation

Mortality risks have been commonly used to examine the perception of low probability events. During the last three decades, a number of studies have shown that participants consistently overestimate the risks of death due to low probability causes, while underestimating risks of death from high probability causes [14–16]. Consider a classic study conducted by Lichtenstein and colleagues, where participants judged the frequency of lethal events in the United States. Lichtenstein *et al.* compared the judged estimates to the actual death rates and found that low frequency events (such as smallpox, poisoning by vitamins, and botulism) were overestimated by a factor of 10, while high frequency events (such as stomach cancer, stroke, and heart disease) were underestimated [17].

In the next three sections, we sketch three psychological mechanisms that drive overestimation of rare events. Our account of mechanisms is by no means comprehensive; the literature contains a number of other often complementary psychological accounts [18–20]. In addition, the overestimation described in this section is not universal. For example, because rare events are not always experienced, they are actually underestimated in some cases as a result [21]. Of course, when rare events are actually experienced, they tend to be overestimated, independent of the psychological mechanisms described below [22].

Availability

Tversky and Kahneman proposed that people deal with the cognitively difficult task of assigning probabilities by using simplifying heuristics [23]. If you are asked to estimate the divorce rate in your state, you may mentally work through your list of friends, and tally how many of these friends have been divorced (or may likely to become so soon). This example illustrates the *availability heuristic*, whereby judged likelihood is based on the ease with which relevant instances

come to mind. Although this heuristic often produces quite sensible probability estimates, in some cases it may lead to systematic biases. For example, two people with the same number of divorced friends may nevertheless judge the divorce rate differently. The availability heuristic suggests that a person whose best friends are divorced will tend to overestimate this rate compared to the person whose best friends are married.

Tversky and Kahneman provided a clear empirical demonstration of this heuristic [24]. One group of participants estimated the proportion of words in a typical English text with the letter “n” in the penultimate position. Another group estimated the proportion of words ending in “-ing.” Although the former must be larger than the latter, participants overwhelmingly judged “-ing” words to be more common than “-n-” words. This is because it is much easier to recruit examples of “-ing” words than “-n-” words. In essence, “-ing” words . . . , “-ing” words are more psychologically available.

We suggest that the availability heuristic will generally give rise to the overestimation of small probability events. The most important low probability events involve extreme consequences, such as plane crashes or winning the lottery [2]. These events tend to be discussed disproportionately and are psychologically more salient, at least in part because high impact events receive media coverage disproportionate to the actual rate of occurrence. For example, Combs and Slovic found that newspapers tend to over-report deaths by homicides, accidents, and natural disasters, and under-report deaths from disease relative to their frequency [25]. In addition, accidents tend to be more dramatic and emotionally evocative, and thus, easier to recall [17,26].

Anchoring on the “ignorance prior”

The French mathematician Laplace proposed a principle that has become known as the *principle of insufficient reason* [27]. Suppose someone is asked to provide probabilities that each of the 16 baseball teams in the American League will advance to the World Series. A non-American who is completely unfamiliar with baseball may lack a compelling reason

to judge one team more likely to win than any other, and thus may conclude that each of the 16 teams has an equal chance (1/16) of winning the American League division. Of course, in many situations, individuals have some knowledge to use for making probabilistic judgments. Nevertheless, recent research has suggested that individuals often start by assigning equal probabilities to each of the possible events (the “ignorance prior”) and then make insufficient adjustment to each of these probabilities [28,29]. We suggest that this process will often result in overestimation of small probabilities.

We first present evidence that decision makers anchor on the ignorance prior. One implication of this process is partition-dependence, in which dramatically different probability estimates are produced, depending on how the event space is partitioned. Fox and Clemen asked Duke MBA students to estimate the likelihood the salary of a random MBA student would fall in each of several salary brackets [28]. In the low condition, there were four salary brackets below the median income of \$85,000, and one above. In contrast, the high condition had one bracket below the median and four above that value. Overall, participants in the low condition estimated that there was a 75% chance that a student would earn less than \$85,000, compared to 40% in the high condition. As predicted, these probabilities are close to the ignorance prior of 80% for the low condition and 20% for the high condition.

Anchoring and adjustment is a process in which decision makers start by “anchoring” on a particular quantity—even if this number is arbitrary or unrelated to the problem at hand—and then adjust insufficiently away from the anchor [23]. Of course, in many situations the anchoring step is well-justified. For example, this year’s sales is usually a good starting point for estimating next year’s sales. However, the anchoring and adjustment process often leads to systematic biases. In one demonstration, Kahneman and Tversky asked participants to estimate the percentage of African countries in the United Nations. In the first stage, a wheel was spun with equal chance of it landing on numbers 0 to 100. Participants were then asked

whether the actual percentage was higher or lower than the number on the wheel. In the second stage, participants were asked to estimate the actual probability. The correlation between the number spun and the eventual estimate was very high, even though it was clear to participants that the number was randomly generated, and had no informational value for the estimate at hand.

Anchoring and adjustment from the ignorance prior thus leads to overestimation of rare events when the number of events in the partition is not large. In most cases, the partition of events includes a small number of events, such as 2 (e.g., “bags will be lost in transport” or “bags will not be lost in transport”) or 3 (e.g., “the Pittsburgh Steelers win,” “the Pittsburgh Steelers lose,” or “the Pittsburgh Steelers tie”). Thus, rare events will be overestimated if a decision maker begins by assigning each event a $1/n$ chance, where n is the number of events in the partition, and then adjusts insufficiently toward the “true probability.”

Coarse Chance Categories

A classic study by Piaget and Inhelder found that the typical 4-year-old child divides chance events into three categories, “certainly will happen,” “certainly will not happen,” and “will possibly happen” [30]. Adults often act as if their categorization of the event space is similarly coarse. For example, Fischhoff and Bruine de Bruin found that participants provided a disproportionate number of probability judgments of 50%, even for mortality-related risks from smoking-induced lung cancer and breast cancer (which are relatively small probabilities in the overall population), indicating equal probabilities of “will occur” and “will not occur” [31]. At the other extreme, professional weather forecasts and expert bridge players are relatively well-calibrated (i.e., events which they judge to have a 20% chance of occurring actually occur 20% of the time), partly because these experts have a finer categorization of the probability interval [32,33].

By definition, low probability events are encountered infrequently, and thus, individuals do not have very much practice thinking about such events. Reyna and

Brainerd suggest that people understand probabilities in “gists,” meaning that they form a general impression rather than a precise probability of whether an event will occur [19]. In this way, it is likely that people form broad notions of probabilities in coarse categories instead of a specific number. Put differently, people seem to differentiate *across* chance categories, but not *within* [34]. Indeed, people prefer communicating probabilistic information verbally rather than numerically, even though verbal statements are less precise [35]. As a result, rare events are lumped into a single category of “unlikely” outcomes. To the extent that “unlikely” corresponds to probabilities between 0.15 and 0.20 [36], this coarse categorization process will lead decision makers to overestimate rare events.

STAGE 2: OVERWEIGHTING OF SMALL PROBABILITIES IN DECISION MAKING

Prospect Theory

We next turn to evidence that small probabilities are overweighted in decision making. Research leading up to the development of prospect theory showed that individuals do not weight probabilities linearly, as would be predicted by expected utility theory. Like expected utility theory, prospect theory assumes that individuals evaluate each alternative and then choose the alternative with the highest subjective valuation. Prospect theory also assumes that the decision maker values the possible outcomes, the probabilities of obtaining those outcomes, and then combines those valuations together. The major difference between prospect theory and the classical theory is that the valuation of outcomes is governed by the value function, $v(x)$, while the valuation of the probabilities is governed by the probability weighting function, $\pi(p)$. The value function is S-shaped, concave for gains and convex for loss, and steeper for losses than gains (see *Prospect Theory*). The second component, the probability weighting function, is the piece of prospect theory that captures the psychological phenomenon of overweighting of small probabilities.

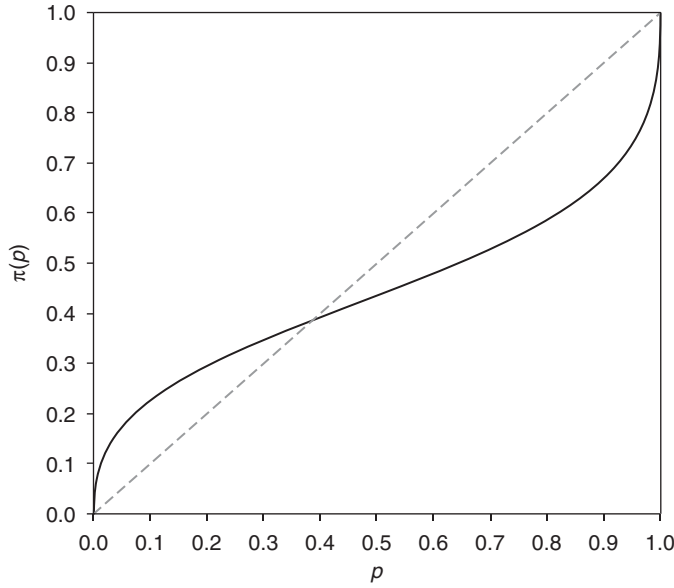


Figure 1. A typical prospect theory probability weighting function, $\pi(p)$, which is concave for small probabilities and convex for medium to large probabilities (and thus, consistent with the principle of diminishing sensitivity).

We should note that small probabilities are not always overweighted. On the contrary, they may be ignored if the probability is below some threshold [37].

Psychophysics of the Probability Weighting Function

Prospect theory assumes that individuals do not weigh outcomes by their probability, as in expected utility theory, but by some distortion of probabilities. This distortion of probability is captured by prospect theory's probability weighting function, $\pi(p)$. The main psychological principle governing the distortion of probability is *diminishing sensitivity* away from a reference point. For probability, there are two obvious reference points, impossibility and certainty, or 0% chance and 100% chance. The distortion of probability captured by the probability weighting function captures the diminishing sensitivity away from these two reference points. People are most sensitive to changes in probability when they are near 0% or 100% than when the change applies to intermediate probabilities. Thus, most people regard the change from 0% to 1% as significant, because it changes the chance of winning from impossible to possible, a phenomenon known as the *possibility effect*. In addition, improving a 99% chance at winning by 1% is also

substantial, because the chance of winning has shifted from almost certain to certain, a phenomenon known as the *certainty effect*. In contrast, a chance from 32% to 33% is seen as inconsequential. Algebraically, both $\pi(0.01) - \pi(0)$ and $\pi(1) - \pi(0.99)$ are larger than $\pi(0.33) - \pi(0.32)$. More generally, a typical probability weighting function is concave for low probabilities and convex for medium to high probabilities (see Fig. 1).

Tversky and Kahneman conducted a study that illustrates this nonlinearity [10]. Participants provided certainty equivalents for gambles that provided 5, 50, and 95% chances at winning \$100 [10]. They found that the certainty equivalents for these gambles were \$14, \$36, and \$78, respectively. The first 5% contributes \$14 of value (or \$2.80 for each percentage), while the next 45% is only worth \$22 (or \$0.48 for each percentage). The final 5% adds \$22 in value (or \$4.40 for each percentage), compared to the previous 45%, which contributes \$42 (or \$0.93 for each percentage).

Diminishing sensitivity implies that low probabilities are typically given more weight than they would receive using expected utility. Empirically, there is generally overweighting, $\pi(p) > p$, for $p < 0.35$ [38,39]. This overweighting, combined with an S-shaped value function, is consistent with

risk-seeking for low probability gains (such as lottery tickets) and risk-aversion for low probability losses (such as insurance). (Note, however, that overweighting of probabilities is distinct from risk-seeking. For gains, concavity of the probability weighting function for low probabilities “competes” again with concavity of the value function. Thus, risk-seeking is observed for probabilities around 0.10 or less [10].) Medium to high probabilities are typically given less weight than they would receive using expected value. Such underweighting is consistent with risk-aversion for medium to high probability gains, and risk-seeking for medium to high probability losses.

Affect and Risk

The shape of the probability weighting function also depends on the affective reactions associated with potential outcomes of a risky choice. Affect is the psychological concept that encompasses noncognitive reactions to stimuli, such as arousal, emotions, or moods. Some outcomes are relatively affect-rich, while others are relatively affect-poor. In a set of studies, the probability weighting function was more curved for affect-rich targets (such as kisses and electric shocks) than affect-poor targets (such as money) [40]. In particular, Rottenstreich and Hsee demonstrated a preference reversal in which participants preferred \$50 in cash to “the opportunity to meet and kiss your favorite movie star,” but chose a 1% chance at a kiss over a 1% chance at \$50. This study suggests that there is more overweighting of small probabilities for affect-rich outcomes than affect-poor outcomes.

The affective account is compelling because many of the risks that are overweighted in judgment are dramatic, fear-inducing and catastrophic, and hence affect-rich. While rare events that are affect-poor in nature may also be overweighted, events that are affect-rich are likely to be overweighted to a larger extent. In the aftermath of the attacks on 9/11, Vice President Dick Cheney set out what has become known as the *One Percent Doctrine*: “We have to deal with this new type of threat in a way we have not yet defined...With a low-probability,

high-impact event like this...if there is a 1% chance that Pakistani scientists are helping al Qaeda to build or develop a nuclear weapon, we have to treat it as a certainty in terms of our response” [41]. The One Percent Doctrine is an extreme example of overweighting of small probabilities, but less extreme risks are typically overweighted as well, although to a lesser extent.

SUMMARY

The overweighting of small probability events is the result of a two-stage process. In the first stage of overestimation of rare events, decision makers take information they receive from the world and then recruit from their memory and convert that information into a psychological representation of probability. As such, this stage is particularly susceptible to many of the basic cognitive limitations and shortcuts identified by psychological research, biasing the resulting judgments in a predictable manner. Though by no means an exhaustive list, we have identified several of the most important, including the availability heuristic, anchoring on the “ignorance prior,” and coarse chance categories. In the second stage of overweighting of small probability events, these (perhaps already inflated) probabilities are distorted upward due to the psychophysics of chance and affect.

Although eliminating these biases altogether is probably unrealistic, it is nevertheless possible to reduce these biases. Many of the most effective approaches will require identifying and then addressing the particular psychological mechanism that produced the biased probability judgment [42]. Consider, for example, the availability heuristic. To the extent that salient and vivid instances tend to come to mind and are thus overweighted in a decision, organizations might force decision makers to rely on historical actuarial evidence rather than what evidence they would naturally recall. Alternatively, organizations may develop “cognitive repairs” [43]. For example, a division at Motorola was overly focused on the problems of large customers, even though small customers constituted a large revenue

base for the division. Motorola developed a process in which problems were surveyed more systematically and weighted by customer volume. In contrast, anchoring on the ignorance prior requires a different intervention, such as a thoughtful consideration of how to partition the space of events [28].

In the second stage, small probability events are additional inflated, as captured by the probability weighting function. Fortunately, this bias is easily addressed when decisions are made using the standard analytical methods of decision analysis [44]. In decision analysis, outcomes (or the utility of these outcomes) are multiplied by their probability of occurring, with the alternative with the highest expected value (or expected utility) being selected.

REFERENCES

1. Clemen RT, Winkler RL. Combining probability distributions from experts in risk analysis. *Risk Anal* 1999;19(2):187–203.
2. Camerer CF, Kunreuther H. Decision processes for low probability events: policy implications. *J Policy Anal Manag* 1989;8(4):565–592.
3. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
4. Clotfelter CT, Cook PJ. *Selling hope: state lotteries in america*. Cambridge (MA): Harvard University Press; 1989.
5. Savage LJ. *The foundations of statistics*. 1st ed. New York: Wiley; 1954.
6. Von NJ, Morgenstern O. *Theory of games and economic behavior*. 1st ed. Princeton (NJ): Princeton University Press; 1944.
7. Machina MJ. Choice under uncertainty: problems solved and unsolved. *J Econ Perspect* 1987;1(1):121–154.
8. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185:1124–1131.
9. Wu G, Gonzalez R. Common consequence conditions in decision making under risk. *J Risk Uncertain* 1998;16(1):115–139.
10. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5(4):297–323.
11. Starmer C. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *J Econ Lit* 2000;38(2):332–382.
12. Fox CR, See KE. Belief and preference in decision under uncertainty. In: Hardman D, Macchi L, editors. *Thinking: psychological perspectives on reasoning, judgment and decision making*. New York: Wiley; 2003. pp. 273–314.
13. Tversky A, Fox CR. Weighing risk and uncertainty. *Psychol Rev* 1995;102:269–283.
14. Armantier O. Estimates of own lethal risks and anchoring effects. *J Risk Uncertain* 2006;32(1):37–56.
15. Hakes JK, Viscusi WK. Dead reckoning: demographic determinants of the accuracy of mortality risk perceptions. *Risk Anal* 2004;24(3):651–664.
16. Viscusi WK, Hakes JK, Carlin A. Measures of mortality risks. *J Risk Uncertain* 1997;14(3):213–233.
17. Lichtenstein S, Slovic P, Fischhoff B, *et al.* Judged frequency of lethal events. *J Exp Psychol [Hum Learn]* 1978;4(6):551–578.
18. Brenner L, Griffin D, Koehler DJ. Modeling patterns of probability calibration with random support theory: diagnosing case-based judgment. *Organ Behav Hum Decis Process* 2005;97(1):64–81.
19. Reyna VF, Brainerd CJ. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learn Individ Differ* 2008;18(1):89–107.
20. Budescu DV, Erev I, Au WT. On the importance of random error in the study of probability judgment: Part II: applying the stochastic judgment model to detect systematic trends. *J Behav Decis Making* 1997;10(3):173–188.
21. Hertwig R, Barron GM, Weber EU, *et al.* Decisions from experience and the effect of rare events in risky choice. *Psychol Sci* 2004;15(8):534–539.
22. Hadar L, Fox CR. Information asymmetry in decision from description versus decision from experience. *Judgm Decis Mak* 2009;4(4):317–325.
23. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185:1124–1131.
24. Tversky A, Kahneman D. Availability - heuristic for judging frequency and probability. *Cogn Psychol* 1973;5(2):207–232.
25. Combs B, Slovic P. Newspaper coverage of causes of death. *J Q* 1979;56(4):837–843.

26. Carroll JS. The effect of imagining an event on expectations for the event: an interpretation in terms of the availability heuristic. *J Exp Soc Psychol* 1978;14(1):88–96.
27. Hacking I. The emergence of probability. 1st ed. Cambridge: Cambridge University Press; 1975.
28. Fox CR, Clemen RT. Subjective probability assessment in decision analysis: partition dependence and bias toward the ignorance prior. *Manag Sci* 2005;51(9):1417–1432.
29. Fox CR, Rottenstreich Y. Partition priming in judgment under uncertainty. *Psychol Sci* 2003;14(3):195–200.
30. Piaget JP, Inhelder B. In: Leake PB, Fishbein HD, Leake L, translators. The origin of the idea of chance in children. 1st ed. New York: Norton; 1976.
31. Fischhoff B, Bruine de Bruin W. Fifty-Fifty=50%? *J Behav Decis Making* 1999; 12(2):149–163.
32. Murphy AH, Winkler RL. Reliability of subjective probability forecasts of precipitation and temperature. *J R Stat Soc Ser C Appl Stat* 1977;26(1):41–47.
33. Keren G. Facing uncertainty in the game of bridge: a calibration study. *Organ Behav Hum Decis Process* 1987;39(1):98–114.
34. Sun Y, Wang H, Zhang J, *et al.* Probabilistic judgment on a coarser scale. *Cogn Sys Res* 2008;9(3):161–172.
35. Wallsten TS, Budescu DV, Zwick R. Comparing the calibration and coherence of numerical and verbal probability judgments. *Manag Sci* 1993;39(2):176–190.
36. Lichtenstein S, Newman JR. Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychon Sci* 1967;9(10):563–564.
37. Slovic P, Fischhoff B, Lichtenstein S, *et al.* Preference for insuring against probable small losses: insurance implications. *J Risk Insur* 1977;44(2):237–258.
38. Wu G, Gonzalez R. Curvature of the probability weighting function. *Manag Sci* 1996;42(12):1676–1690.
39. Gonzalez R, Wu G. On the shape of the probability weighting function. *Cogn Psychol* 1999;38(1):129–166.
40. Rottenstreich Y, Hsee CK. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol Sci* 2001;12(3):185–190.
41. Suskind R. The one percent doctrine: deep inside America's pursuit of its enemies since 9/11. 1st ed. New York: Simon and Schuster; 2006.
42. Arkes HR. Costs and benefits of judgment errors: implications for debiasing. *Psychol Bull* 1991;110(3):486–498.
43. Heath C, Larrick RP, Klayman J. Cognitive repairs: how organizational practices and compensate for individual shortcomings. *Res Organ Behav* 1998;20:1–37.
44. Raiffa H. Decision analysis: introductory lectures on choices under uncertainty. New York: Addison-Wesley; 1968.

PARADOXES AND VIOLATIONS OF NORMATIVE DECISION THEORY

JAY SIMON
Defense Resources Management
Institute, Naval Postgraduate
School, Monterey, California

YITONG WANG
L. ROBIN KELLER
The Paul Merage School of
Business, University of
California Irvine, Irvine,
California

Traditional economic decision theory proposes that people behave in certain ways when faced with a well-formulated set of alternatives and information. It is a normative theory; it suggests that people should act according to certain decision rules, but not that they will necessarily do so in reality. The standard assumption is that a person has a utility function over all possible outcomes, which defines the desirability or usefulness of any given scenario. Traditional utility theory asserts that in a decision without uncertainty, a person should choose the alternative resulting in the highest level of utility, and that in a decision with uncertainty, a person should choose the alternative with the highest expected utility. Detailed explanations are given by von Neumann and Morgenstern and Savage [1,2].

These ideas are generally sound and can be valuable when used as a guide for decision makers. However, it was initially thought that they were also valuable as a descriptive theory. It has since been determined that people violate normative decision theory in many clear and predictable ways. In this article, we introduce and discuss several of these interesting behavioral phenomena.

One of the earliest researchers to demonstrate a deviation from expected utility theory was Allais [3]. He distributed two surveys. In the first survey, respondents were asked to choose between the following

two options (original amounts were in francs):

Choice 1.

A:	\$1 million for sure
B:	0.10 probability of \$2 million 0.89 probability of \$1 million 0.01 probability of \$0

In the second survey, respondents were asked to choose between the following two options:

Choice 2.

C:	0.11 probability of \$1 million 0.89 probability of \$0
D:	0.10 probability of \$2 million 0.90 probability of \$0

The majority of respondents preferred *A* to *B*, and preferred *D* to *C*.

Choosing *A* and *D* appears to be reasonable. However, this example has been called the *Allais Paradox*, because these two choices contradict the idea that people are maximizing expected utility. If we let U_0 represent the person's utility for \$0, U_1 represent the person's utility for \$1 million, and U_2 represent the person's utility for \$2 million, then preferring *A* to *B* would imply that

$$U_1 > 0.10U_2 + 0.89U_1 + 0.01U_0,$$

or equivalently,

$$U_1 > \frac{10}{11}U_2 + \frac{1}{11}U_0.$$

If we examine the second survey choice, we find that preferring *D* to *C* implies that

$$0.11U_1 + 0.89U_0 < 0.10U_2 + 0.90U_0.$$

This may not seem like a problem at first glance. However, with simple

algebraic manipulation, we see that this is equivalent to

$$U_1 < \frac{10}{11}U_2 + \frac{1}{11}U_0.$$

This is the exact opposite result of what we saw in the first survey! It means that regardless of the utilities of the three outcomes, choosing both A and D (or both B and C) contradicts the theory of expected utility maximization.

Table 1 reveals the structure of the transformation of the first choice between A and B into the second choice between C and D . The final column corresponds to the 0.89 chance of the \$1 million outcome received in either A or B in the first choice. This \$1 million was replaced with a \$0 in both C and D in the second choice. Changing the amount of this “sure thing” of getting the same outcome of \$1 million in the 0.89 probability outcome in choice 1 to a different sure thing of \$0 in choice 2 should not change the preference order between the options. Savage (2, Postulate 2) called this the *sure-thing* principle, which must be obeyed under expected utility maximization.

This was one of the first clear demonstrations of a systematic violation of the principles of expected utility theory. That in itself is a very influential finding. However, we may want to ask why it happened. What element of people’s underlying preferences is not being captured by expected utility theory?

It turns out there are multiple explanations for Allais’ results, but the main reason they occurred is a phenomenon often referred to as the *certainty effect*. Reducing the probability of a negative outcome from 0.01 to 0 is usually judged to be a greater improvement than reducing the probability of that same negative outcome from, say, 0.90 to 0.89. Expected utility maximization requires

that preferences over gambles are linear in probabilities, and in Allais’ example, they clearly are not.

Another major finding in this area was discussed by Ellsberg [4]. He conducted an experiment in which subjects were presented with a jar. The subjects were told that the jar contains 30 red balls, and 60 balls which are some unknown mix of yellow and black. They were asked to choose between two gambles:

A:	Receive \$100 if the ball is red
B:	Receive \$100 if the ball is black

The majority of participants chose gamble A . They were then asked to choose between the following two gambles:

C:	Receive \$100 if the ball is red or yellow
D:	Receive \$100 if the ball is black or yellow

Given this choice, the majority of participants selected gamble D . As in the Allais’ example, choosing both A and D seems reasonable. However, this has been called the *Ellsberg Paradox*, because it is also a violation of expected utility theory. It may not be immediately clear how expected utility is violated, so let us write a participant’s estimated fraction of black balls as B , and observe the following contradiction:

Choosing A implies that

$$\begin{aligned} \frac{1}{3}U(\$100) + \frac{2}{3}U(\$0) &> BU(\$100) \\ &+ (1 - B)U(\$0). \end{aligned}$$

Assuming that $U(\$100) > U(\$0)$, this simplifies to $B < 1/3$.

Choosing D implies that

$$\begin{aligned} (1 - B)U(\$100) + BU(\$0) &< \frac{2}{3}U(\$100) \\ &+ \frac{1}{3}U(\$0). \end{aligned}$$

Table 1. Allais Paradox

	0.10	0.01	0.89
A	\$1 million	\$1 million	\$1 million
B	\$2 million	\$0	\$1 million
C	\$1 million	\$1 million	\$0
D	\$2 million	\$0	\$0

Assuming that $U(\$100) > U(\$0)$, this simplifies to $B > 1/3$.

Thus, there is no consistent set of beliefs (about the fraction of black balls and the fraction of yellow balls) which allows us, using traditional subjective expected utility theory, to prefer both gamble A and gamble D . That is, no single probability of drawing a black ball would lead to this pair of responses. However, there is a simple reason that most people choose both gamble A and gamble D : these are the choices for which we can easily compute the exact probability of receiving \$100. Ellsberg showed that not only do people display aversion to risk, they also display aversion to “ambiguity.” People tend to prefer gambles for which they are confident that they know the exact probabilities involved.

Table 2 shows the structure of this paradox. The choice between A and B is transformed by changing the sure thing of \$0 when a yellow ball is chosen to a sure thing of \$100 when yellow is chosen. So, if A is preferred over B , then C should be preferred over D .

Possibly the most groundbreaking work to bring violations of normative utility theory to a wide audience is from Kahneman and Tversky [5]. They highlighted two major behavioral violations of expected utility maximization, and demonstrated them using simple survey results.

Their first major violation resulted from a very simple pair of questions:

1. Would you prefer to receive \$3000 for sure, or \$4000 with probability 0.8?
2. Would you prefer to lose \$3000 for sure, or lose \$4000 with probability 0.8?

Expected utility maximization does not tell us precisely what the answers to these questions should be. However, notice that in

the second question, the expected value of losing \$4000 with probability 0.8 is $-\$3200$. It has greater risk than the sure loss, and a lower expected value. Preferring the gamble would imply that the person is risk-seeking. Utility curves formed over total wealth are almost always concave, implying risk aversion.

The answers to these two questions were startling. People (in aggregate) overwhelmingly preferred receiving \$3000 for sure in question 1, and overwhelmingly preferred losing \$4000 with probability 0.8 in question 2. That is, they were clearly risk-averse in question 1, and clearly risk-seeking in question 2. Kahneman and Tversky also asked similar types of questions in different contexts, using different probabilities and outcomes. In all cases, they found that the majority’s preferred answer to the question dealing with losses was the opposite of that for the question dealing with gains. They referred to this phenomenon as the *reflection effect*.

Since these people did not all have the same level of wealth, it would be impossible to construct a utility function over wealth that could incorporate the reflection effect and accurately represent these subjects’ responses. Instead, Kahneman and Tversky proposed that people evaluate outcomes as gains and losses rather than using the resulting total wealth levels. Based on their survey data, they postulated that people tend to be risk-averse over gains and risk-seeking over losses. Numerous subsequent studies have demonstrated this phenomenon, yet it is still not fully accepted or incorporated into traditional economic theory.

Kahneman and Tversky also found nonlinearities in probability, and expressed them in some more detail than Allais. In addition to certainty effects, they found that people tend to place too much weight on outcomes with very small (nonzero) probabilities. That is, they do not fully understand or incorporate how truly unlikely these outcomes are. They postulated that people treat probabilities according to the weighting function shown in Fig. 1. The horizontal axis represents the actual given probability, and the vertical axis represents the probability that people implicitly interpret it to be

Table 2. Ellsberg Paradox

	30 Balls Red balls	60 Balls are Either Black or Yellow	
		Black balls	Yellow balls
A \$100		\$0	\$0
B \$0		\$100	\$0
C \$100		\$0	\$100
D \$0		\$100	\$100

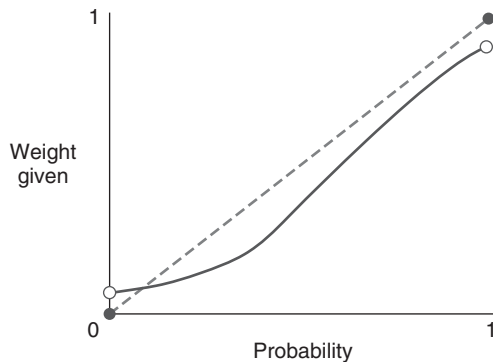


Figure 1. Nonlinear application of probabilities as observed by Kahneman and Tversky [5].

when making decisions. The “correct” interpretation of probability, as implied by traditional expected utility theory, is given by the 45 degree dotted line. This graph incorporates the certainty effect (at both zero and one), the overweighting of small probabilities, and the underweighting of middle and higher probabilities.

Another related behavioral phenomenon is what Tversky and Kahneman [6] referred to as *loss aversion*. Loss aversion states that a loss is deemed to be more significant than an equal-sized gain. The theoretical loss aversion concept explains two other more specific observed concepts. These two specific concepts are the *endowment effect* and the *status quo bias*.

The *endowment effect* is discussed by Thaler [7]. The premise is that the loss in utility from giving up an item is greater than the gain in utility from acquiring it. It was demonstrated experimentally by Kahneman *et al.* [8]. They asked subjects to place a monetary value on a cheap decorative mug. The catch was that some subjects had been given the mug beforehand, and others had not. Since the mugs were assigned randomly, there was no apparent reason that the underlying preferences should differ between those with mugs and those without. However, they found that the subjects who were given the mug placed *much* higher values on it than those who were not. In one experiment, the median value for the mug was \$7.12 among the subjects who had one, and \$3.12 among those who had not. In a second experiment,

these median values were \$7 and \$3.50, respectively. This was clear evidence that people tend to place much higher values on items they possess, or are “endowed” with.

The *status quo* bias was discussed by Samuelson and Zeckhauser [9]. *Status quo* bias is a preference for remaining in the current situation. The example that the authors used dealt with financial investments. They conducted an experiment in which several different funds were described to the participant. When participants were told that they currently had money invested in one of the funds, then they generally preferred that fund to the others. When they were given a “neutral” scenario in which they did not have money invested in any of the funds, then no such bias was observed.

Loss aversion explains both the endowment effect and the status quo bias. If a loss is more significant than an equal-sized gain (in terms of utility), then losing an item is clearly more significant than receiving that same item. It is not difficult to use loss aversion to explain the endowment effect. To understand the status quo bias, we can interpret a change as “losing” the current scenario and obtaining the new one. If the two scenarios are equally desirable when viewed from a neutral perspective, then loss aversion suggests that the decision maker will prefer the status quo.

Notice that in each of the two preceding examples, if the outcomes were viewed in terms of raw overall outcomes rather than gains and losses, the groups would be facing identical questions. In the endowment effect example, both groups would be expressing the difference in monetary value between a “mug” outcome and a “no mug” outcome. In the status quo bias example, each individual would be simply selecting the most desirable fund.

One implication of expected utility theory is that a person’s chosen course of action should not depend on the manner in which the alternatives are presented. If two decision situations involve the same outcomes with the same probabilities resulting from the corresponding alternatives, then the person’s decision should be identical for both. This property is called *invariance*. It is discussed by von Neumann and Morgenstern [1],

and studied further by Arrow [10]. Invariance is often violated in reality. These violations are referred to as *framing* effects.

Tversky and Kahneman [11] conducted a pair of surveys which demonstrated a clear framing effect. In both survey scenarios, a group of 600 people is about to be exposed to a deadly disease. In the first survey, the choice is whether to save 200 people for sure, or save all 600 with probability $1/3$ (and save nobody with probability $2/3$). The vast majority of participants chose the sure option. In the second survey, the choice is whether to allow 400 people to die for sure, or accept a $2/3$ probability of all 600 people dying (and a $1/3$ probability of nobody dying). The vast majority of participants chose the gamble.

Look at those two survey questions again. They are actually asking the same thing! One is framed in terms of number of lives saved, and the other is framed in terms of number of deaths. The outcomes and their associated probabilities, however, are identical. Simply due to the ways in which the situations are described, we find that many people's preferences are completely reversed. An observant reader may notice that this particular framing effect occurs because one question is described in terms of gains, and the other in terms of losses. Tversky and Kahneman [12] discussed this in greater detail. As prospect theory tells us, people tend to be risk-averse for gains and risk-seeking for losses.

One of the arguments in favor of using a risk-averse utility curve over wealth is the observation that most people will never accept a 50/50 gamble between gaining and losing $\$X$, regardless of the value of X . This implies a concave (risk-averse) utility function for wealth. Loss aversion is an alternative explanation for this observation; it asserts that " $-\$X$ " has a larger effect on utility than " $+\$X$ " does, without having to assess or incorporate wealth level.

Loss aversion, along with Kahneman and Tversky's earlier data, implies that rather than looking at potential overall wealth levels when making decisions, people tend to view outcomes as gains and losses relative to a particular reference level. Kahneman and Tversky's [5] prospect theory suggests that

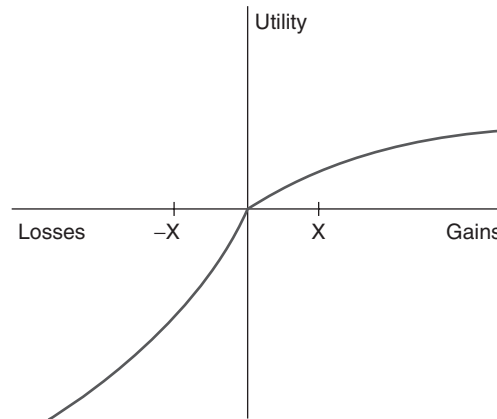


Figure 2. The utility function associated with prospect theory Kahneman and Tversky [5].

people act according to a utility function¹ similar to the one shown in Fig. 2, where the slope of the utility is steeper when a loss of $\$X$ is incurred than when a gain of $\$X$ is incurred. This idea alone can explain many real-world deviations from traditional expected utility theory. Baucells and Rata [13] demonstrated empirically that using a reference point in this manner is an important factor in risk-taking behavior; Camerer [14] provides many additional real-world examples.

Given that people do not always conform to expected utility, there are different ways to proceed. One is to investigate how to aid people to do so. For example, Keller [15,16] investigated how different visual representations of gambles (most promisingly drawings of color coded balls labeled with the monetary amount that would be received if a ball is drawn) could lower violations of expected utility for questions similar to the Allais Paradox. Another approach is to relax the assumptions required by expected utility and

¹Kahneman and Tversky refer to this as a value function rather than a utility function, and it is generally understood to be an appropriate model for preferences under certainty. However, they proposed it initially based on observed preferences over gambles, as we have discussed earlier. Thus, to be consistent with our usage of the terms "value" and "utility," we refer to the curve in Fig. 2 as a utility function.

construct generalizations of expected utility, such as models that relax the requirement of linearity in probabilities. Kahneman and Tversky's [5] prospect theory is the most well-known generalized utility theory, but there are many others.

Becker and Sarin [17] introduced a lottery-dependent utility model that allows the utility achieved in an outcome to vary according to the lottery in which it occurs, but requires the true probabilities to be used. Their model was rather general, but can be used to explain some violations of expected utility. Daniels and Keller [18] tested this model experimentally, and found that it was useful in predicting choice among gambles, but not noticeably better than expected utility when actually determining decision makers' underlying preferences. Some other generalized utility models include weighted utility [19], regret theory (20, 21), and skew-symmetric bilinear utility [22,23]. For more examples and detailed discussion the reader is directed to Refs 24, 25, or 26.

In general, expected utility theory is quite valuable as a normative theory for guiding good decision making. However, between the certainty effect, ambiguity aversion, the reflection effect, the endowment effect, status quo bias, loss aversion, and framing effects, it is painfully apparent that people violate the traditional theory in many clear and predictable ways.

A final question for thought: in the real world, how often do these sorts of paradoxes and violations occur? Certainly the effects and biases that we have discussed do exist, but how often do they lead to incorrect or suboptimal decisions? If they occur very infrequently, then they are interesting as academic issues only. However, if they do come up reasonably often in the real world, then they undoubtedly merit further study and awareness.

REFERENCES

1. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton University Press; 1947, second edition.
2. Savage L. The foundations of statistics. New York: John Wiley & Sons, Inc.; 1954.
3. Allais M. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'Ecole Americaine. *Econometrica* 1953;21(4):503–546.
4. Ellsberg D. Risk, ambiguity, and the Savage axioms. *Q J Econ* 1961;75(4):643–669.
5. Kahneman D, Tversky A. An analysis of decision under risk. *Econometrica* 1979; 47(2):263–292.
6. Tversky A, Kahneman D. Loss aversion in riskless choice: a reference dependent model. *Q J Econ* 1991;107(4):1039–1061.
7. Thaler R. Toward a positive theory of consumer choice. *J Econ Behav Organ* 1980; 1(1):39–60.
8. Kahneman D, Knetsch JL, Thaler RH. Experimental tests of the endowment effect and the Coase theorem. *J Polit Econ* 1990; 98(6):1325–1348.
9. Samuelson W, Zeckhauser R. Status quo bias in decision making. *J Risk Uncertain* 1988;1(1):7–59.
10. Arrow K. Risk perception in psychology and economics. *Econ Inq* 1982;20(1):1–9.
11. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981;211(4481):453–458.
12. Tversky A, Kahneman D. Rational choice and the framing of decisions. *J Bus* 1986;59(4):S251–S278.
13. Baucells M, Rata C. A survey study of factors influencing risk-taking behavior in real-world decisions under uncertainty. *Decis Anal* 2006;3:163–176.
14. Camerer C. Prospect theory in the wild: evidence from the field. In: Kahneman D, Tversky A, editors. Choices, values, and frames. Cambridge: Cambridge University Press; 2001. pp. 288–300.
15. Keller LR. The effects of problem representation on the sure-thing and substitution principles. *Manage Sci* 1985a;31(6):738–751.
16. Keller LR. Testing the “reduction of compound alternatives” principle. *Omega, Int J Manage Sci* 1985b;13(4):349–358.
17. Becker J, Sarin R. Lottery dependent utility. *Manage Sci* 1987;33:1367–1382.
18. Daniels RL, Keller LR. An experimental evaluation of the descriptive validity of lottery-dependent utility theory. *J Risk Uncertain* 1990;3(2):115–134.
19. Chew SH, Waller WS. Empirical tests of weighted utility theory. *J Math Psychol* 1986;30:55–72.

20. Bell D. Regret in decision making under uncertainty. *Oper Res* 1982;30:961–981.
21. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. *Econ J Nepal* 1982;92:805–824.
22. Fishburn PC. Transitive measurable utility. *J Econ Theory* 1983;31:293–317.
23. Fishburn PC. SSB utility theory: an economic perspective. *Math Soc Sci* 1984;8:63–94.
24. Edwards W, editor. *Utility theories: measurements and applications*. Boston (MA): Kluwer; 1992.
25. Fishburn PC, LaValle I, editors. *Annals of operations research 19: choice under uncertainty*. Basel: J. C. Baltzer; 1989.
26. Weber M, Camerer C. Recent developments in modelling preferences under risk. *OR Spektrum* 1987;9:129–151.

PARALLEL CONFIGURATIONS

MARCUS AGUSTIN
Department of Mathematics and
Statistics, Southern Illinois
University Edwardsville,
Edwardsville, Illinois

PARALLEL SYSTEM VERSUS SERIES SYSTEM

In the area of reliability, the primary goal is to ensure that a system will function adequately with a high probability. For instance, Clements [1] noted that due to safety reasons, command systems in aircrafts require high probability of functioning. Moreover, it may be desired to maintain a minimum acceptable level in which a system will function. The papers by Kvam and Peña [2] and Sinkovic, Lovrek, and Nemeth [3] modeled how a load can be shared by components in a system in such a way that failure of a component will require the remaining components to compensate for the failed component.

In certain cases, components are connected in a manner where system failure will only be observed as soon as all components have failed. Such a system is known as a *parallel system*. In contrast to a series system (see **Series Systems** in this encyclopedia) where a system fails as soon as a component has failed, a parallel system continues to function as long as at least one component is still functioning. A partial listing of papers that compare a series system and a parallel system include Refs 4 and 5. On the other hand, the papers [6–8], among others, considered the similarity between a parallel system and a system of redundant components. Finally Refs 9 and 10 are some of the recent papers that investigated the failure in the interaction of components in a parallel system.

The probability of observing a functioning system is known as the *system reliability*. The system reliability depends on the manner in which components are connected and the

individual component reliability is usually observed as a function of time. In this paper, we will formulate the system reliability for a parallel system and consider an example in which the individual component reliability will depend on some known probability distribution.

SYSTEM RELIABILITY

Structure Function of a Parallel System

We consider a system of n components such that the state of the system, that is, whether the system is functioning or has already failed, depends on how the components are connected and the current state of each component. Let binary variable X_i indicate whether component i is still functioning, denoted by $X_i = 1$, or has failed, denoted by $X_i = 0$. Each component is assumed to be independent the other resulting in independent random variables $X_1, \dots, X_n$. The vector $\mathbf{X} = (X_1, \dots, X_n)$ is the state vector for a system of n components. On the other hand, the binary function $\phi(\mathbf{X})$ represents the state of a system, that is, $\phi(\mathbf{X}) = 1$ if the system is still functioning and $\phi(\mathbf{X}) = 0$ if the system has failed. For a more detailed discussion of the properties of X_i and $\phi(\mathbf{X})$, the reader is referred to articles **Coherent Systems** and **Series Systems** in this encyclopedia.

A parallel system is a system that functions if and only if at least one component is functioning. From a different point of view, a parallel system fails as soon as all components have failed. Owing to the binary nature of X_i , it follows that $1 - X_i = 1$ whenever component i has failed. Clearly, the events $\{\phi(\mathbf{X}) = 0\}$ and $\{X_1 = 0, \dots, X_n = 0\}$ are equivalent, and $\{\phi(\mathbf{X}) = 1\} = \{\text{at least one } X_i = 1\}$, and thus the structure function of a parallel system is

$$\begin{aligned}\phi(\mathbf{X}) &= 1 - (1 - X_1) \cdots (1 - X_n) \\ &= 1 - \prod_{i=1}^n (1 - X_i).\end{aligned}\tag{1}$$

2 PARALLEL CONFIGURATIONS

An alternative way of expressing the structure function of a parallel system is

$$\phi(\mathbf{X}) = \max(X_1, \dots, X_n).$$

Figure 1 presents a diagram for a parallel system with two components.

Reliability of a Parallel System

Consider a parallel system that is observed at a specified time instant t . To indicate this time-dependence in our notation, let $X_i(t)$ indicate the state of component i at t . The corresponding state vector is $\mathbf{X}(t) = (X_1(t), \dots, X_n(t))$. Let $p_i(t)$ be the probability that component i is functioning at time t . In terms of $X_i(t)$, we have

$$p_i(t) = \mathbf{P}\{X_i(t) = 1\}, \quad i = 1, \dots, n.$$

If we let $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$, then the system reliability at time t is defined to be

$$h(\mathbf{p}(t)) = \mathbf{P}\{\phi(\mathbf{X}(t)) = 1\} = \mathbf{E}[\phi(\mathbf{X}(t))].$$

For a parallel system, it is more convenient to write the system reliability in the form

$$h(\mathbf{p}(t)) = 1 - \mathbf{P}\{\phi(\mathbf{X}(t)) = 0\}.$$

By the independence of the $X_i(t)$'s, we get

$$\begin{aligned} h(\mathbf{p}(t)) &= 1 - \mathbf{P}\{X_1(t) = 0, \dots, X_n(t) = 0\} \\ &= 1 - \mathbf{P}\{X_1(t) = 0\} \cdots \mathbf{P}\{X_n(t) = 0\}. \end{aligned}$$

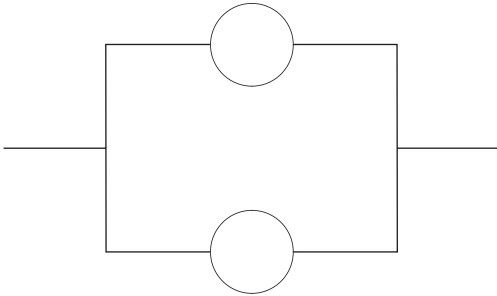


Figure 1. Diagram for a parallel system with two components.

Thus, the reliability of a parallel system at time t is

$$\begin{aligned} h(\mathbf{p}(t)) &= 1 - [1 - \mathbf{P}\{X_1(t) = 1\}] \cdots \\ &\quad \times [1 - \mathbf{P}\{X_n(t) = 1\}] \\ &= 1 - \prod_{i=1}^n [1 - p_i(t)]. \end{aligned}$$

TIME-DEPENDENT SYSTEM RELIABILITY

Survival Function for a Parallel System

We now consider a parallel system where each component's failure process is a function of the time the component has been in use. In other words, we consider a random variable T_i denoting the lifetime of component i with corresponding probability density function $f_i(t; \theta)$ for some parameter vector θ . Following Ref. 11, we define the survival function (or reliability function) and the hazard function of component i at time t to be

$$\begin{aligned} S_i(t; \theta) &= \mathbf{P}\{T_i > t\} \\ &= \int_t^\infty f_i(u; \theta) du \quad \text{and} \\ h_i(t; \theta) &= \frac{f_i(t; \theta)}{S_i(t; \theta)}, \end{aligned}$$

respectively. We will assume that $T_1, \dots, T_n$ are independent random variables. Since all the n components have random lifetimes, the system lifetime, denoted by Y , is also a random variable. Extending Equation (1), it is easily seen that

$$Y = \max(T_1, \dots, T_n). \quad (2)$$

Whenever a system is in use, the main interest is to obtain the survival function and the hazard function of Y . Note that the survival function of Y is the system reliability as we have previously defined. From Equation (2),

$$\begin{aligned} S(y; \theta) &= \mathbf{P}\{\max(T_1, \dots, T_n) > y\} \\ &= 1 - \mathbf{P}\{\max(T_1, \dots, T_n) \leq y\} \\ &= 1 - \mathbf{P}\{T_1 \leq y, \dots, T_n \leq y\}. \end{aligned}$$

By the independence of the component lifetimes, the resulting system reliability is

$$\begin{aligned} S(y; \theta) &= 1 - \prod_{i=1}^n P\{T_i \leq y\} \\ &= 1 - \prod_{i=1}^n [1 - S_i(y; \theta)], \end{aligned} \quad (3)$$

where $S_i(t; \theta)$ is the survival function of the i th component. The corresponding density function of the system lifetime for a parallel system is

$$\begin{aligned} f(y; \theta) &= - \sum_{i=1}^n S'_i(y; \theta) \prod_{\substack{j=1 \\ j \neq i}}^n [1 - S_j(y; \theta)] \\ &= \sum_{i=1}^n f_i(y; \theta) \prod_{\substack{j=1 \\ j \neq i}}^n [1 - S_j(y; \theta)]. \end{aligned} \quad (4)$$

The hazard function of Y is easily obtained by taking the quotient of the density function and reliability function obtained from Equations (4) and (3), respectively. In a more specific case when the component lifetimes are identically distributed with density function $f_*(t; \theta)$ and survival function $S_*(t; \theta)$, the system reliability function and density function of the system lifetime Y become $S(y; \theta) = 1 - [1 - S_*(y; \theta)]^n$ and $f(y; \theta) = n f_*(y; \theta) [1 - S_*(y; \theta)]^{n-1}$, respectively.

Exponential Lifetimes

We will now consider a parallel system consisting of components whose lifetimes are exponentially distributed. The component lifetimes are assumed to be independent of one another but not necessarily identically distributed. Hence, the density function of the i th component lifetime, T_i , is

$$f_i(t; \theta_i) = \theta_i e^{-\theta_i t}, \quad \theta_i > 0.$$

Some recent papers that considered exponential lifetimes for a parallel system include Refs 4 and 12–14. For a more detailed discussion of the exponential distribution, the reader is referred to articles **Common Failure Distributions** and **Series Systems** in this encyclopedia.

We let $Y = \max(T_1, \dots, T_n)$ denote the lifetime of the parallel system and the vector $\theta = (\theta_1, \dots, \theta_n)$ the scale parameter vector. From Equation (3), the system reliability of a parallel system is

$$S(y; \theta) = 1 - \prod_{i=1}^n (1 - e^{-\theta_i y}).$$

Since $S'_i(y; \theta) = -f_i(y; \theta)$, it follows from Equation (4) that the density function of Y is

$$f(y; \theta) = \sum_{i=1}^n \theta_i e^{-\theta_i y} \prod_{\substack{j=1 \\ j \neq i}}^n (1 - e^{-\theta_j y}).$$

Whenever the component lifetimes are also identically distributed, that is, $\theta_i = \theta$ for $i = 1, \dots, n$, the system reliability and density function become

$$\begin{aligned} S(y; \theta) &= 1 - (1 - e^{-\theta y})^n \quad \text{and} \\ f(y; \theta) &= n \theta e^{-\theta y} (1 - e^{-\theta y})^{n-1}, \end{aligned}$$

respectively. Consequently, in this special case the hazard function of a parallel system is

$$h(y; \theta) = \frac{f(y; \theta)}{S(y; \theta)} = \frac{n \theta e^{-\theta y} (1 - e^{-\theta y})^{n-1}}{1 - (1 - e^{-\theta y})^n}.$$

Let us examine the behavior of the hazard function $h(y; \theta)$ as $y \rightarrow \infty$. It has been shown in the literature that $\lim_{y \rightarrow \infty} h(y; \theta) = \min(\theta_1, \dots, \theta_n)$ and in particular, $\lim_{y \rightarrow \infty} h(y; \theta) = \theta$ in the identically distributed case. Thus, the limit of the hazard function of a parallel system is the same as the hazard function of the strongest exponential component of the system.

REFERENCES

1. Clements RR. Reliability of multichannel systems. SIAM Rev 1980;22:88–95.
2. Kvam PH, Pena EA. Estimating load-sharing properties in a dynamic reliability system. J Am Stat Assoc 2005;100:262–272.
3. Sinkovic V, Lovrek I, Nemeth G. Load balancing in distributed parallel systems for telecommunications. Computing 1999;63:201–218.

4 PARALLEL CONFIGURATIONS

4. Tan Z. Estimation of exponential component reliability from uncertain life data in series and parallel systems. *Reliab Eng Syst Saf* 2007;92:223–230.
5. Hausken K. Strategic defense and attack for series and parallel reliability systems. *Eur J Oper Res* 2008;186:856–881.
6. Altinel K, Ozekici S, Feyzioglu O. Component testing of repairable systems in multistage missions. *J Oper Res Soc* 2001;52:937–944.
7. Koide T, Sandoh H. Economic analysis of an n-unit parallel redundant system based on a Stackelberg game formulation. *Comput Ind Eng* 2009;56:388–398.
8. Papageorgiou E, Kokolakis G. Reliability analysis of a two-unit general parallel system with n-2 warm standbys. *Eur J Oper Res* 2010;201:821–827.
9. Zequeira RI, Berenguer C. On the inspection policy of a two-component parallel system with failure interaction. *Reliab Eng Syst Saf* 2005;88:99–107.
10. Asadi M. On the mean past lifetime of the components of a parallel system. *J Stat Plan Inference* 2006;136:1197–1206.
11. Leemis L. Reliability: probabilistic models and statistical methods. New Jersey: Prentice Hall; 1995.
12. Sarhan AM, El-Bassiouny AH. Estimation of components reliability in a parallel system using masked system life data. *Appl Math Comput* 2003;138:61–75.
13. Sarhan AM. Reliability equivalence factors of a parallel system. *Reliab Eng Syst Saf* 2005;87:405–411.
14. Vellaisamy P, Kumar M. Optimal component test plans for a parallel system based on type-II censoring. *Stat Meth* 2008;5:454–461.

PARALLEL DISCRETE-EVENT SIMULATION

JASON LIU
School of Computing and Information
Sciences, Florida International
University, Miami, Florida

SIMULATION AND PARALLELIZATION METHODS

Discrete-Event Simulation

Simulation, defined as a representation of the operation of real-world processes or systems over time [1], is one of the most important and widely used techniques in operations research. These processes or systems are abstracted as models in the form of mathematical or logical relationships. In simulation, one uses computers to evaluate the models numerically in a controlled environment, where data are gathered and used to estimate the behavior of the target systems. Simulation is effective for prototyping new system designs, as well as providing insight to the true characteristics of existing systems. Simulation is particularly indispensable for studying large-scale and complex systems, which can be otherwise intractable to closed-form mathematical or analytical solutions. Owing to the practical nature of the design process where simplified assumptions fly in the face of required complexities, simulation is an irrefutable choice for both system design and evaluation.

A simulation model specifies the state evolution over time. The target system can be viewed either as a *continuous* system, where the state changes continuously with respect to time; as a *discrete* system, where the state is modified only at specific points in time; or as a *hybrid* system, which consists of both continuous and discrete elements. In simulation, time progression is described by the so-called “time advancement function,” the treatment of which signifies two classes of

simulation methods: time-driven and event-driven. In a *time-driven* (or *time-stepping*) simulation, time is measured at small intervals giving the impression that the system evolves continuously over time. As such, it is naturally more suitable for simulating continuous systems.

In an *event-driven* (or *discrete-event*) simulation, time leaps through distinct points in time, which we call *events*. Consequently, event-driven simulation is more appropriate for simulating discrete systems. Note that one can combine both types of simulations, for example in a computer network simulation, using discrete events to represent detailed network transactions such as sending and receiving packets, and using continuous simulation to capture the fluid dynamics of overall network traffic [2].

This article mainly focuses on discrete-event simulations and efficient techniques to parallelize them. It is necessary to first understand what constitutes a discrete-event simulation before we proceed to review the parallelization methods. At the application level, modelers can describe the processing of events using three major kinds of simulation worldviews (also known as *simulation paradigms*): event-oriented, process-oriented, and activity scanning. The *event-oriented* worldview focuses on events and their effects on the state of the simulation. Event handlers are defined and used in the simulation program to process events of different types. Event-oriented worldview provides the lowest modeling perspective and yet, the most efficient mechanism among the three worldviews.

The *process-oriented* worldview describes a higher level of modeling abstraction than does the event-oriented worldview. A model is expressed as a set of interacting simulation processes, each being a thread of control that describes the state evolution over time. A simulation process usually has a life cycle longer than just a single event, and is provided with mechanisms to interact with other simulation processes: a process can wait for

a period of simulation time to elapse or some specific events to happen, such as the arrival of a message. It is important to note that the process-oriented worldview can be implemented on top of an event-oriented simulation engine with additional support for thread management.

The third kind of simulation worldview is called *activity scanning*. It is closely related to time-driven simulation. A collection of procedures are provided, each associated with a predicate. At each timestep, all predicates are evaluated and the corresponding procedures are invoked if the predicates turn out to be true. All simulation worldviews are merely user perspectives, which can be supported by the same discrete-event simulation engines. We choose to focus on the event-oriented worldview in this article as it is closest among the three simulation worldviews to the underlying discrete-event simulation mechanism, which we seek to parallelize.

A discrete-event simulator maintains a data structure called the *event list*, basically a priority queue that sorts events according to the time at which they are scheduled to happen in the simulated future. A clock variable T is used to denote the current time in simulation. At the heart of the program is a loop; the simulator repeatedly removes an event with the smallest time stamp from the event list, sets the clock variable T to the time stamp of this event, and processes the event. Processing an event typically changes the state of the model and may generate more future events to be inserted into the event list. The loop continues until the simulation termination condition is met; for example, when the event list becomes empty or when the simulation clock has reached a designated simulation completion time.

Parallelization Methods

Parallel discrete-event simulation (PDES), also known simply as *parallel simulation*, is an important research area cross-cutting modeling and simulation, and high-performance computing. PDES is concerned with executing a single discrete-event simulation program on parallel computers, which can be shared-memory multiprocessors, distributed-memory clusters of

interconnected computers, or a combination of both. By exploiting the potential parallelism in the model, PDES can overcome the limitations imposed by sequential simulation in both the execution time and the memory space. As such, PDES can bring substantial benefits to time-critical simulations and simulations of large-scale systems that demand a tremendous amount of computing resources.

The simplest form of parallel simulation is called *replicated trials* [3]. This method executes multiple instances of a sequential simulation program concurrently on parallel computers. These “replicated” trials may be replications belonging to the same experiment (e.g., started off with different random seeds), or separate simulation runs with different model parameters. This approach has the obvious advantage of simplicity. It can be an efficient variance reduction method or a method for expediting the exploration of a large parameter space. The disadvantage is that each replicated trial does not provide any speedup and cannot overcome the memory limitation imposed upon sequential execution. To address the latter problem, Hybinette and Fujimoto [4] introduced a cloning method as an efficient parallel computation technique to allow simultaneous exploration of different simulation branches resulted from alternative decisions made during simulation.

Another form of parallel simulation is to assign different functions of a simulation program, such as random number generation or event handling, to separate processors. This method is called *functional decomposition* [5]. The main problem of this approach is the lack of ample parallelism. Also, the tight coupling of the simulation functions creates an excessive demand for communication and synchronization among the parallel components, which can easily defeat the parallelization effort.

More generally, one can view simulation as a set of state variables that evolve over time. Chandy and Sherman [6] presented a space-time view of simulation, where each event can be characterized by a temporal coordinate, indicated by the time stamp of the event, and a spatial coordinate, indicated

by the location of the state variables affected by the event. Accordingly, the state space of a discrete-event simulation can be perceived as consisting of a continuous time axis and a discrete space axis; the objective of the simulation in general is therefore to compute the value at each point in the space–time continuum. This space–time view provides a high-level unifying concept for parallel simulation, where one can divide the space–time graph into regions of arbitrary shape and assign them to separate processors for parallel processing. For example, Bagrodia *et al.* [7] presented a distributed space-time algorithm based on fixed-point computations.

A special case of the space–time view is the *time-parallel* approach, which is based on temporal decomposition of the time–space continuum. Time-parallel simulation divides the space–time graph along the time axis into nonoverlapping time intervals and assigns them to different processors for parallel processing. Due to the obvious dependency issue, that is, the initial state of a time interval must match the final state of the preceding time interval, the efficiency of this approach relies heavily on the model’s ability to either rapidly compute the initial state or achieve fast convergence under relaxation. For this reason, only a limited number of cases using time-parallel simulation exist in the literature. Successful examples include trace-driven cache simulations [8,9], queueing network and Petri net simulations [10,11], and road traffic simulations [12].

Orthogonal to the time-parallel approach, *space-parallel* simulation is based on data decomposition, where the target system is divided into a collection of subsystems, each simulated by a *logical process (LP)*. Each LP maintains its own simulation clock and event list, and is capable of processing only those events that pertain to the subsystem to which it is assigned. These LPs can be assigned to different processors and executed concurrently. Communication between the LPs takes place exclusively by exchanging time-stamped events. In general, space-parallel simulation is considered more robust than the other parallelization approaches, mainly because data decomposition can be naturally

applicable to most models. For this reason, we focus on space-parallel simulation for the rest of this article.

SYNCHRONIZATION ALGORITHMS

In a discrete-event simulation, events need be processed in a nondecreasing time stamp order, because an event with a smaller time stamp has the potential to modify the state of the system and thereby affect events that happen later. This property is called the *causality constraint*. Provided that simultaneous events—events with the same time stamps—are sorted deterministically and consistently using certain tie-breaking rules, the causality constraint implies a total ordering of events. In parallel simulation, the global event list in sequential simulation is replaced by a set of event lists; each LP maintains its own simulation clock and a separate event list that contains events that can only affect the state of the corresponding LP. Since each LP processes events on its own event list in time stamp ordering—a property also known as the *local causality constraint*—the total ordering maintained by the original sequential discrete-event model is replaced by a partial ordering similar to Lamport’s “happens before” relationship [13]. The fundamental challenge is therefore associated with the difficulty of preserving the local causality constraint at each LP without the use of a global simulation clock as LPs advance their own simulation clocks more or less independently from one another.

We use a simple example to illustrate this problem (Fig. 1). Suppose there are two interconnected queues, A and B, with jobs in both queues waiting to be serviced. After a job is serviced at one queue, it joins the other queue after a certain delay. There are two types of simulation events: one representing the job arrival and the other representing the job departure. The two queues are simulated by two LPs running on two separate processors. Upon processing the “job departure” event at one LP, a “job arrival” event will be sent from the LP to the other LP. Figure 1 shows that a job enters Queue A at time 4 (represented by $E1$), departs at 10 ($E3$), and then

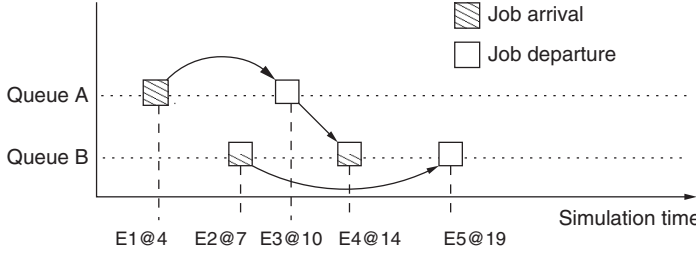


Figure 1. Simulation of two queues.

travels to Queue B at 14 ($E4$), while another job enters Queue B at time 7 ($E2$) and leaves at 19 ($E5$). Since the two queues are simulated on separate processors, it is possible that the LP handling Queue B gets to process $E5$ immediately after processing $E2$, while the other LP is still in the middle of processing $E1$ at Queue A. That is, it is probable that an event (in this case, $E4$) could arrive from the other LP later in real time carrying a time stamp smaller than the LP's current simulation clock.

To avoid potential causality errors, some synchronization mechanism needs to be in place. Specifically, an LP must be able to make the right decision about whether to process the smallest time-stamped event on its local event list.

The distinction on how local causality constraint is enforced underscores the difference between two major parallel simulation methods. *Conservative parallel simulation* eliminates the possibility of any causality errors; that is, an LP must be blocked from processing the next event in its event list until it is sure that processing such event will not cause out-of-order event execution due to future events to arrive from other LPs.

In contrast, *optimistic parallel simulation* allows events to be processed out of order. Once a causality error is detected, the offending LP must be rolled back to recover from such an error. In order for the simulation to roll back from possible erroneous computations, state saving and recovery mechanisms must be provided.

CONSERVATIVE SYNCHRONIZATION

The conservative approach strictly prohibits out-of-order event execution: an LP must be

blocked from processing the next event in its event queue until it is safe to do so. Synchronization needs to take place among the LPs to ensure that no event will arrive at an LP with a time stamp earlier than its local simulation clock.

CMB Algorithm

The first parallel simulation synchronization protocol is the *CMB algorithm* proposed independently by Chandy and Misra [14], and Bryant [15].

In CMB, LPs are connected via directional links, through which events are transferred from one LP to another in chronological order (with nondecreasing time stamps). Events are enqueued at the receiving LPs, there is one input queue for each incoming link at an LP. Also, each input queue is associated with a clock variable, set to be either the time stamp of the first event in the queue, or, if the queue is empty, the time stamp of the last processed event (or zero initially). Each LP maintains a loop: at each iteration, the LP selects an input queue with the smallest clock value and processes the event at the beginning of the input queue; if the input queue is empty, the LP blocks until an event arrives at the queue and then continues at the next iteration.

Since LPs block on empty input queues (with the smallest clock value), deadlock may happen once a waiting cycle is formed, in which case no progress can be made even if there are events in other input queues. The CMB algorithm uses *null messages* to avoid this pathological situation. A null message does not represent any real activities in the model; it carries only a time stamp, which can be treated as a guarantee from the sending LP that it will not send events in the

future with time stamps smaller than the time stamp of the null message. Upon receiving a null message, an LP can advance the clock associated with the input queue. The LP can further propagate the time advancement to its successor LPs, possibly by sending out more null messages. Consequently, no waiting cycle can be formed and deadlock is avoided.

It is important to note that the use of null messages is not the only way to prevent deadlocks. Alternatively, one can allow deadlock to happen, and subsequently detect and recover from deadlock situations [16]. Deadlock recovery is based on the observation that events with the smallest time stamp in the system can always be processed safely. In cases where deadlocks happen more frequently, however, this may result in a significant amount of sequential execution that can adversely affect the overall performance.

Exploiting Lookahead

The CMB algorithm introduced an important concept regarding event processing. Each LP must determine the lower bound on the time stamps (LBTS) of future events to arrive from other LPs. LBTS is actually the upper bound up to which the LP can safely process its local events and advance its simulation clock without introducing causality errors. CMB uses a small time increment above the LP's current simulation clock as the time stamp of the null messages it sends to the successor LPs. This is the essential idea behind the concept of *lookahead*, which is defined as the amount of simulation time that an LP can predict into the simulated future (and therefore allow other LPs to safely advance their simulation clocks). In other words, lookahead is the inherent asynchrony in the simulation model—with a positive lookahead, an LP can process events within a certain range independent of its neighbors. Extensive performance studies [17–19] confirmed the positive correlation between lookahead and the parallel performance of conservative algorithms.

Exploiting lookahead takes on two directions. One direction focuses on extracting lookahead from model characteristics. Nicol [20] provided a classification of lookahead

based on different levels of knowledge that can be extracted from the model. Several models exhibit good lookahead properties. A notable example is the simulation of first-come-first-served (FCFS) stochastic queueing networks [21]. By presampling the job service time and determining branch destination at the time when a job enters the queue, one can significantly improve the lookahead. Another example is the simulation of continuous-time Markov chains (CTMC) [22–24]. The method exploits the mathematical structure of CTMC models (more specifically, the uniformization property) that allows the preselection of synchronization points ahead of time. Consequently, the synchronization complexity can be reduced significantly due to the improved lookahead.

The other direction focuses on lookahead extrapolation for general applications. Chandy and Sherman's *conditional-event* approach [25] takes into account two types of events: *definite* events, which are bound to happen regardless of the earlier events, and *conditional* events, which can be canceled because of earlier events. This algorithm requires each LP to determine the LBTS of future events it will send to each of its neighbor LPs on the premise that its earliest local conditional event will not be canceled. This lower bound can be calculated as the time stamp of the earliest conditional event on the local event list plus the lookahead. Since the earliest conditional event in the system is actually a definite event, the algorithm uses a global min-reduction to find out the true lower bound and use it as the LBTS for all LPs to process their local events.

The topology of the model—how LPs are connected with each other—plays an important role in the lookahead computation. Lubachevsky's *bounded lag* algorithm [26] takes advantage of the minimum propagation delay between LPs and the so-called opaque period within which the state of an LP is not affected by other LPs due to the model's non-preemptive behavior. The algorithm introduces a time interval B , called the *lag*, through which the algorithm computes the “sphere of influence”, which

is the set of LPs that can possibly affect a given LP within B units of simulation time. These are the LPs that need to be considered to determine whether events on the given LP with time stamps between T (the current simulation time) and $T + B$ can be safely processed. Ayani's *distance-between-objects* algorithm [27] also exploits the distance between the LPs; it uses a shortest-path algorithm to determine the LP's LBTS. The *time-of-next-event* algorithm proposed by Groselj and Tropper [28] also explores LP topology; the algorithm considers situations where multiple LPs reside on a single processor and allows efficient computation of the greatest lower bound of all LPs on a processor. Nicol [29] established a scalability analysis based on a window-based synchronization algorithm. The algorithm is similar to the conditional-event approach in that it also uses a global min-reduction to determine the time of the next synchronization point.

In a certain way, LP-based conservative synchronization methods can be viewed as a scheduling problem where LPs are selected to run on parallel processors. Xiao *et al.* [30] introduced an asynchronous multilevel LP-scheduling algorithm, called the *critical channel traversing* (CCT) algorithm, to accommodate simulation models with small computational granularity. CCT schedules groups of LPs among the processors using a shared data structure to reduce unnecessary low-level synchronization overheads. Nicol and Liu [31] proposed a composite synchronization scheme combining both synchronous and asynchronous scheduling algorithms to avoid the potential performance pitfalls from either synchronous or asynchronous mechanisms used alone.

OPTIMISTIC SYNCHRONIZATION

Optimistic synchronization enforces the local causality constraint differently from its conservative counterpart: an LP is allowed to process events that arrive in the simulated past, as long as the simulation is able to detect such causality errors, rewind the simulation clock, and roll back the erroneous computations.

Time Warp

It is safe to say by and large that Jefferson's *Time Warp* algorithm [32] is the most far-reaching parallel synchronization protocol. The protocol was first conceived to be offering a general solution to all parallel simulation problems, although the overhead associated with many aspects of the protocol's execution eventually caught up. In retrospect, the protocol set forth an era of intensive research in the PDES field.

In Time Warp, each LP saves the events received from other LPs in the input queue, and those sent to other LPs in the output queue. Also, the state variables are saved in the state queue each time before an event is processed. When an event arrives with a time stamp smaller than the current simulation clock (called the *straggler* event), the LP must be rolled back to the saved state immediately before the time stamp of the straggler event. All actions that the LP might have affected on other LPs later than the time stamp of the straggler event must also be canceled. This can be achieved by sending *anti-messages* corresponding to the original messages that are stored at the output queue. Upon receiving an anti-message, the LP will annihilate the corresponding message in the input queue. If the anti-message carries a time stamp smaller than the LP's current simulation clock (and thus become a straggler event itself), the LP will also be rolled back accordingly.

Simple and elegant as it appears to be, the algorithm must address the important issue of memory consumption: as the simulation progresses, the size of the input, output, and state queues could increase without bound. Furthermore, since rollback does not apply to irrevocable operations, such as I/O, the algorithm must also determine when these irrevocable operations can be executed. The *global virtual time* (GVT) is defined as the minimum time stamp of all events and messages (including both positive messages and anti-messages) in the system at a given wall-clock time. One can show that the system will never roll back to a time earlier than GVT. As such, GVT can be viewed as a measure of progress, where messages or state vectors

earlier than GVT can be reclaimed (in a process called *fossil collection*). Similarly, irrevocable operations that happen before GVT can also be committed. Another important problem with Time Warp is containment. As indicated in a paper by Nicol and Liu [33], a correct simulation program that runs sequentially may fail catastrophically due to an unpredictable inconsistent state resulted from out-of-order event execution. Such failures must be captured or contained by the Time Warp simulation engine (not without significant undertaking to modify operating system services), so that a failed state can be rolled back when Time Warp corrects itself.

Time Warp also needs to overcome several practical problems pertaining to its performance: state saving inevitably introduces memory overhead; rollback entails more complexities, where state needs to be restored, and anti-messages need to be sent, which can possibly cause further rollbacks rippling through the entire system. These problems have prompted a lot of interesting research, some of which we describe next.

State Saving, GVT, and Rollback

State saving is necessary to undo modifications to the state variables, caused by erroneous computations. There are three types of state-saving techniques: *copy state saving* makes a copy of all state variables each time an event is processed; *incremental state saving* saves only those state variables modified as a result of processing an event; and *infrequent state saving* adjusts the interval between checkpoints and reduces the frequency of state saving. Incremental state saving can be made automatic, by overloading the assignment operators in certain object-oriented programming languages [34], by using specific hardware support [35], or by directly modifying the executable files [36]. To enable infrequent state saving, the rollback mechanism must be able to handle situations where an LP may be rolled back to a state not immediately before the straggler event. In this case, the LP must *coast forward*, reprocessing events until the desired state is reached. For this reason infrequent state saving must be applied judiciously. Lin *et al.* [37] studied the trade-off between

the state-saving interval length and the rollback cost, and suggested an analytically optimal solution.

GVT computation is a major source of overhead for Time Warp. The task of computing GVT amounts to capturing a consistent snapshot of the state of a distributed system. According to Samadi [38], there are two major issues: the *transient message* problem and the *simultaneous reporting* problem. A transient message is a message that has been sent but not yet received; either the sender or the receiver LP must account for the message in the GVT computation, or inconsistencies would occur. Simultaneous reporting is also related to the possible miscount of transient messages; it is, however, prompted by the difference in the ordering of events perceived by different LPs. Samadi's algorithm [38] employs a message acknowledgment scheme to solve the transient message problem and adopts a barrier synchronization method to solve the simultaneous reporting problem. To avoid the expensive barrier synchronization, Preiss's algorithm [39] arranges the LPs in a ring and uses a token traversal approach to compute GVT. Mattern's algorithm [40] uses two broadcasts to define separate "cuts" and uses messages with "colors" to distinguish those spanning across the cuts.

Optimizations can also be applied to reduce the overhead from rollbacks. Gafni [41] proposed *lazy cancellation* to avoid canceling messages that are more likely to be recreated after rollbacks. The algorithm does not immediately send anti-messages during a rollback; instead, they are sent only when the same positive messages are not recreated. West [42] proposed another method called *lazy reevaluation* to avoid reprocessing the events, which can be achieved by comparing the state vectors. Both methods require additional time for comparing messages or state vectors, which could cause erroneous computation to be spread further.

One interesting alternative to the state-saving approach is *reverse computation*, proposed by Carothers *et al.* [43]. Rather than saving the state to enable rollback, reverse computation performs the inverses of the individual operations that correspond

to the normal event processing (i.e., by executing the code backwards) to get the system back to an earlier state. This method alleviates the state-saving cost from the forwarding computation path and transfers the cost to the reverse computation path. It can also be viewed as trading the memory space (which would otherwise be consumed by state saving) with the execution time (for the more expensive rollback). Significant improvements over traditional state-saving methods have been reported both in run time and memory utilization for fine-grain models, including queueing network models. One potential problem for reverse computation is that not all operations are reversible, in which case the method is degenerated to the traditional state-saving approach. The increased rollback cost may also get erroneous computations to be propagated to more LPs, possibly causing unstable behavior.

Curbing Optimism

In practice, “pure” Time Warp usually does not provide an acceptable performance. Many alternative solutions have been proposed that effectively limit optimistic execution to circumvent various performance hazards.

Lin and Preiss [44] first introduced the concept of *storage optimality*. A storage optimal parallel simulator can complete the execution of any model using no more than the memory space required to complete a sequential execution multiplied by some constant factor. To achieve storage optimality, a necessary condition is that the parallel simulator should be able to reclaim the memory space beyond the sequential execution requirement, even if it implies restraint in the speculative execution. Several algorithms run on shared-memory multiprocessors qualify for the storage optimality, which includes Jefferson’s *cancel-back* algorithm [45], and Lin and Preiss’s *artificial rollback* algorithm [44]. Preiss and Loucks’s *prune-back* algorithm [46] is designed for distributed-memory architectures; however, it fails to meet the storage optimality requirement.

A number of optimistic parallel simulation algorithms are proposed to curb optimism.

For example, the *moving time window* approach [47] uses a window to control how far an LP may get ahead of other LPs. The effectiveness of this algorithm obviously depends on the size of the moving time window.

Reynolds [48] introduced a classification method for optimistic synchronization protocols based on two attributes: *aggressiveness* and *risk*. Aggressiveness specifies that events on an LP can be processed out of time stamp order, where risk suggests that messages generated by aggressive event processing are allowed to be sent to other LPs. For example, the SRADS protocol by Dickens and Reynolds [49] is aggressive, but allows no risk: it permits only local rollbacks; messages are not sent unless the LP is sure that the messages will not be rolled back. As an extension, the *breathing time bucket* approach, proposed by Steinman [50], introduced a quantity called an *event horizon*. An LP’s local event horizon is the smallest time stamp of any new messages that are generated as a result of processing local events. The global event horizon can be computed as the minimum of the local event horizons among all LPs. The algorithm allows only messages with a sending time prior to the global event horizon to be sent out. By doing that, it can be shown that these messages will never be rolled back.

Many have tried to compare conservative and optimistic simulations; yet it comes as no surprise that there is no conclusive answer to which is a better approach. This is mainly because parallel simulation performance largely depends on the characteristics of the simulation model. On the one hand, conservative synchronization, due to its low operational overhead and small memory footprint, can usually achieve good performance as long as good lookahead can be guaranteed from the model. On the other hand, the optimistic approach can exploit better parallelism in the model through speculative execution. However, optimistic execution requires state saving, GVT computation, fossil collection, and rollbacks, which could bring in significant design complexity and memory overhead.

HIGH-PERFORMANCE MODELING AND SIMULATION OF LARGE-SCALE NETWORKS

PDES has been applied in many areas, such as military applications (including war games and training exercises), on-line gaming, business operations, manufacturing, logistics and distribution, transportation, computer systems, and computer networks. In this section, we review the latest development in high-performance modeling and simulation of large-scale computer networks as an example of successful PDES applications.

Internet is described by Paxson and Floyd [51] as “an immense moving target”: its size, heterogeneity, and rapid change pose a daunting challenge for accurately simulating its complex behavior. Over the last ten years or so, researchers have developed several parallel simulators that can model networks at unprecedented scales. For example, Riley *et al.* [52] used a federated approach to parallelize the popular *ns-2* simulator [53]. The original sequential discrete-event simulation engine is “refurbished” with parallel functions to allow time-synchronized event distribution. There are also unadulterated efforts in developing parallel network simulations from scratch. One early effort is the Telecommunications Description Language (TeD) proposed by Perumalla *et al.* [54], which allows parallel execution of large network models with an optimistically synchronized parallel simulation engine.

The Scalable Simulation Framework (SSF) was later developed as a common Application Programming Interface (API) for parallel simulations of large-scale telecommunication systems [55]. SSF is different from TeD in that it does not require a separate programming language; SSF network models are written in common programming languages (Java or C++) using simple simulation constructs that support transparent conservative parallel execution on different platforms. A similar effort is the Georgia Tech Network Simulator (GTNetS) [56], which is a full-fledged simulation environment embellished with a rich set of network protocol implementations. The GTNetS effort has recently evolved into *ns-3* [57], which is a renewed effort for

open community development of network simulations. Both SSF and GTNetS are developed based on conservative synchronization protocols. ROSSNet [58], on the other hand, is a large-scale network simulator using an optimistic approach. It offers a compact and light-weight implementation framework that reduces the amount of state required to simulate large-scale network models. ROSSNet uses reverse computation to enable fast optimistic execution on commodity multiprocessor systems.

Two important application areas using high-performance network simulation have shown very promising results: on-line simulation and real-time simulation. *On-line network simulation* refers to the use of network simulation as an integrated service for real-time network management for parameter tuning, network planning, network monitoring, and network traffic engineering [59,60]. Parallel simulation allows detailed performance analysis faster than real time and thus can be used to predict, control, and fine-tune network parameters in real time. *Real-time network simulation* refers to simulation of large-scale networks in real time so that the virtual network can interact with real implementations of network protocols, network services, and distributed applications. Real-time simulation allows network immersion, where a simulated network is made indistinguishable from a physical network in terms of conducting network traffic between real applications. It provides the necessary realism, scalability, and flexibility for experimental networking research [61].

ADDITIONAL READING

This article is but scratching the surface of the exciting field of PDES. Owing to space limitations, we are far from being able to provide an in-depth review of all the problems and solutions. There are many excellent surveys of the parallel discrete-event simulation field [62–65]. Fujimoto’s book on PDES [66] provides a good review of the intriguing problems and solutions in this area. The Workshop on Principles of Advanced and Distributed Simulation (PADS) is the primary conference for parallel discrete-event

simulation. Many research papers on PDES applications also appear regularly in the Winter Simulation Conference (WSC).

REFERENCES

1. Banks J, Carson JS II, Nelson BL, *et al.* Discrete-event system simulation. 4th ed. Upper Saddle River (NJ): Prentice Hall; 2004.
2. Liu J. Packet-level integration of fluid TCP models in real-time network simulation. Proceedings of the 2006 Winter Simulation Conference (WSC'06); 2006 Dec. pp. 2162–2169.
3. Henderson SG. Input model uncertainty: why do we care and what should we do about it? Proceedings of the 35th Conference on Winter Simulation (WSC'03), Volume 1; 2003 Dec. pp. 90–100.
4. Hybinette M, Fujimoto R. Cloning parallel simulations. ACM Trans Model Comput Simul 2001;11(4):378–407.
5. Comfort JC. The simulation of a master-slave event set processor. Simulation 1984;42(3):117–124.
6. Chandy KM, Sherman R. Space-time and simulation. Proceedings of the 1989 SCS Multiconference on Distributed Simulation, Volume 21, No. 2; 1989 Mar. pp. 53–57.
7. Bagrodia R, Chandy KM, Liao WT. A unifying framework for distributed simulation. ACM Trans Model Comput Simul 1991;1(4):348–385.
8. Heidelberg P, Stone HS. Parallel trace-driven cache simulation by time partitioning. Proceedings of the 1990 Winter Simulation Conference (WSC'90); 1990 Dec. pp. 734–737.
9. Nicol DM, Greenberg AG, Lubachevsky BD. Massively parallel algorithms for trace-driven cache simulations. IEEE Trans Parallel Distrib Syst 1994;5(8):849–858.
10. Lin YB, Lazowska ED. A time-division algorithm for parallel simulation. ACM Trans Model Comput Simul 1991;1(1):73–83.
11. Ammar H, Deng S. Time Warp simulation using time scale decomposition. ACM Trans Model Comput Simul 1992;2(2):158–177.
12. Kiesling T, Luthi J. Towards time-parallel road traffic simulation. Proceedings of the 19th Workshop on Principles of Advanced and Distributed Simulation (PADS 2005); 2005 Jun. pp. 7–15.
13. Lamport L. Time, clocks, and the ordering of events in a distributed system. Commun ACM 1978;21(7):558–565.
14. Chandy KM, Misra J. Distributed simulation: a case study in design and verification of distributed programs. IEEE Trans Softw Eng 1979;SE-5(5):440–452.
15. Bryant RE. Simulation of packet communication architecture computer systems. Technical Report MIT-LCS-TR-188. MIT; 1977.
16. Chandy KM, Misra J. Asynchronous distributed simulation via a sequence of parallel computations. Commun ACM 1981;24(4):198–205.
17. Reed DA, Malony AD, McCredie BD. Parallel discrete event simulation using shared memory. IEEE Trans Softw Eng 1988;14(4):541–553.
18. Fujimoto RM. Lookahead in parallel discrete event simulation. Proceedings of the 1988 International Conference on Parallel Processing; 1988 Aug. pp. 34–41.
19. Fujimoto RM. Performance measurements of distributed simulation strategies. Trans Soc Comput Simul 1989;6(2):89–132.
20. Nicol DM. Principles of conservative parallel simulation. Proceedings of the 1996 Winter Simulation Conference (WSC'96); 1996 Dec. pp. 128–135.
21. Nicol DM. Parallel discrete-event simulation of FCFS stochastic queueing networks. ACM SIGPLAN Not 1988;23(9):124–137.
22. Heidelberg P, Nicol D. Conservative parallel simulation of continuous time Markov chains using uniformization. IEEE Trans Parallel Distrib Syst 1993;4(8):906–921.
23. Nicol D, Heidelberg P. Optimistic parallel simulation of continuous time Markov chains using uniformization. J Parallel Distrib Comput 1993;18(4):395–410.
24. Nicol D, Heidelberg P. A comparative study of parallel algorithms for simulating continuous time Markov chains. ACM Trans Model Comput Simul 1995;5(4):326–354.
25. Chandy KM, Sherman R. The conditional event approach to distributed simulation. In: Proceedings of the 1989 SCS Multiconference on Distributed Simulation, Volume 21, No. 2; 1989 Mar. pp. 93–99.
26. Lubachevsky BD. Bounded lag distributed discrete event simulation. Proceedings of the 1988 SCS Multiconference on Distributed Simulation, Volume 19, No. 3; 1988. pp. 183–191.
27. Ayani R. A parallel simulation scheme based on the distance between objects. Proceedings

- of the 1989 SCS Multiconference on Distributed Simulation, Volume 21, No. 2; 1989 Mar. pp. 113–118.
28. Groselj B, Tropper C. The time of next event algorithm. Proceedings of the 1988 SCS Multiconference on Distributed Simulation, Volume 19, No. 3; 1988. pp. 25–29.
 29. Nicol DM. The cost of conservative synchronization in parallel discrete event simulations. *J Assoc Comput Mech* 1993;40(2):304–333.
 30. Xiao Z, Unger B, Simmonds R, *et al.* Scheduling critical channels in conservative parallel discrete event simulation. Proceedings of the 13th Workshop on Parallel and Distributed Simulation (PADS'99); 1999 May. pp. 20–28.
 31. Nicol D, Liu J. Composite synchronization in parallel discrete-event simulation. *IEEE Trans Parallel Distrib Syst* 2002;13(5):433–446.
 32. Jefferson DR. Virtual time. *ACM Trans Program Lang Syst* 1985;7(3):404–425.
 33. Nicol D., Liu X. The dark side of risk (what your mother never told you about Time Warp). Proceedings of the 11th Workshop on Parallel and Distributed Simulation (PADS'97); 1997 May. pp. 188–195.
 34. Rönngren R, Liljenstam M, Montagnat J, *et al.* Transparent incremental state saving in Time Warp parallel discrete event simulation. Proceedings of the 10th Workshop on Parallel and Distributed Simulation (PADS'96); 1996 May. pp. 70–77.
 35. Fujimoto RM. The virtual time machine. Proceedings of the 1st Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA'89); 1989 Jun. pp. 199–208.
 36. West D, Panesar K. Automatic incremental state saving. Proceedings of the 10th Workshop on Parallel and Distributed Simulation (PADS'96); 1996 May. pp. 78–85.
 37. Lin YB, Preiss BR, Loucks WM, *et al.* Selecting the checkpoint interval in Time Warp simulation. Proceedings of the 7th Workshop on Parallel and Distributed Simulation (PADS'93); 1993 May. pp. 3–10.
 38. Samadi B. Distributed simulation, algorithms and performance analysis [PhD thesis]. Department of Computer Science, UCLA; 1985.
 39. Preiss BR. The Yaddes distributed discrete event simulation specification language and execution environments. Proceedings of the 1989 SCS Multiconference on Distributed Simulation, Volume 21(2); 1989 Mar. pp. 139–144.
 40. Mattern F. Efficient distributed snapshots and global virtual time approximation. *J Parallel Distrib Comput* 1993;18(4):423–434.
 41. Gafni A. Rollback mechanisms for optimistic distributed simulation systems. Proceedings of the 1988 SCS Multiconference on Distributed Simulation; Volume 19(3); 1988. pp. 61–67.
 42. West D. Optimizing Time Warp: Lazy rollback and lazy re-evaluation [Master's thesis]. Department of Computer Science, University of Calgary; 1988.
 43. Carothers CD, Perumalla KS, Fujimoto RM. Efficient optimistic parallel simulations using reverse computation. *ACM Trans Model Comput Simul* 1999;9(3):224–253.
 44. Lin YB, Preiss BR. Optimal memory management for Time Warp parallel simulation. *ACM Trans Model Comput Simul* 1991;1(4):283–307.
 45. Jefferson DR. Virtual time II: storage management in distributed simulation. Proceedings of the 9th Annual ACM Symposium on Principles of Distributed Computing; 1990 Aug. pp. 75–89.
 46. Preiss BR, Loucks WM. Memory management techniques for Time Warp on a distributed memory machine. Proceedings of the 9th Workshop on Parallel and Distributed Simulation (PADS'95); 1995 Jun. pp. 30–39.
 47. Sokol LM, Briscoe DP, Wieland AP. MTW: a strategy for scheduling discrete simulation events for concurrent execution. Proceedings of the 1988 SCS Multiconference on Distributed Simulation, Volume 19(3); 1988. pp. 34–42.
 48. Reynolds PF Jr. A spectrum of options for parallel simulation. Proceedings of the 1988 Winter Simulation Conference (WSC'88); 1988 Dec. pp. 325–332.
 49. Dickens PM, Reynolds PF Jr. SRADS with local rollback. Proceedings of the 1990 SCS Multiconference on Distributed Simulation, Volume 22(1); 1990 Jan. pp. 161–164.
 50. Steinman JS. Breathing Time Warp. Proceedings of the 7th Workshop on Parallel and Distributed Simulation (PADS'93); 1993 May. pp. 109–118.
 51. Paxson V, Floyd S. Why we don't know how to simulate the Internet. Proceedings of the 1997 Winter Simulation Conference (WSC'97); 1997 Dec. pp. 1037–1044.

52. Riley GF, Fujimoto RM, Ammar MH. A generic framework for parallelization of network simulations. Proceedings of the 7th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'99); 1999 Oct. pp. 128–135.
53. Breslau L, Estrin D, Fall K, *et al.* Advances in network simulation. IEEE Comput 2000;33(5):59–67.
54. Perumalla K, Ogielski A, Fujimoto R. TeD—a language for modeling telecommunication networks. ACM SIGMETRICS Perform Eval Rev 1998;25(4):4–11.
55. Cowie JH, Nicol DM, Ogielski AT. Modeling the global Internet. Comput Sci Eng 1999;1(1):42–50.
56. Riley GF. The Georgia Tech network simulator. Proceedings of the ACM SIGCOMM Workshop on Models, Methods and Tools for Reproducible Network Research (MoMeTools'03); 2003 Aug. pp. 5–12.
57. Henderson TR, Roy S, Floyd S, *et al.* ns-3 project goals. Proceeding of the 2006 Workshop on ns-2: The IP Network Simulator (WNS2); 2006 Oct.
58. Yaun GR, Bauer D, Bhutada HL, *et al.* Large-scale network simulation techniques: examples of TCP and OSPF models. ACM SIGCOMM Comput Commun Rev 2003;33(3):27–41.
59. Szymanski BK, Saifee A, Sastry A, *et al.* Genesis: a system for large-scale parallel network simulation. Proceedings of the 16th Workshop on Parallel and Distributed Simulation (PADS 2002); 2002 May. pp. 89–96.
60. Ye T, Kaur HT, Kalyanaraman S, *et al.* Large-scale network parameter configuration using an on-line simulation framework. IEEE/ACM Trans Network 2008;16(4):777–790.
61. Liu J. A primer for real-time simulation of large-scale networks. Proceedings of the 41st Annual Simulation Symposium (ANSS'08); 2008 Apr. pp. 307–323.
62. Fujimoto RM. Parallel discrete event simulation. Commun ACM 1990;33(10):30–53.
63. Nicol D, Fujimoto R. Parallel simulation today. Ann Oper Res 1994;53:249–286.
64. Lin YB, Fishwick P. Asynchronous parallel discrete event simulation. IEEE Trans Syst Man Cybern 1996;26(4):397–412.
65. Low YH, Lim CC, Cai W, *et al.* Survey of languages and runtime libraries for parallel discrete-event simulation. Simulation and Trans of SCS 1999;72(3):170–186.
66. Fujimoto RM. Parallel and distributed simulation systems. New York: John Wiley & Sons, Inc.; 2000.

PARALLEL SYSTEMS

MINJAE PARK

HOANG PHAM

Rutgers University, The State
University of New Jersey, New
Brunswick, New Jersey

INTRODUCTION

It is important to understand the concept of redundancy in parallel system as it contributes toward improving the reliability of the system. Redundancy is a state where more than one possible ways exists for the system to perform the required operations. When an engineer is designing a system, simply reducing the number of components is not the best strategy. Because in cases where high system reliabilities are needed the engineer must design the system to avoid any single failure and possibly duplicate components to improve the reliabilities, that is, introduce redundancy. And, the common structure of redundancy is the *k-out-of-n* systems where a special case of *k-out-of-n* systems, that is $k = 1$, becomes parallel systems [1,2].

Unlike the series system, in which a single failure of a component results in failure of the whole system, in the parallel system all the elements of the system need to fail to cause failure in operation. As can be seen in Fig. 1 where there are q components connected in parallel such that even if one component fails the link to the other component is still available resulting in successful functioning of the system. In other words, a system is said to be a parallel system if and only if the failure of all components within the system results in the failure of the entire system [3,4]. We study the parallel systems with respect to the warranty policy and also present numerical examples to illustrate the results.

Similar to the case in the series system, the reliability of the parallel system can be expressed as follows: let R_s be the reliability

of the system, $\bar{Z}_i$ be the event of component i that has failed, and $P(\bar{Z}_i)$ be the probability that component i has failed. The reliability of the system is then given by

$$\begin{aligned} R_s &= 1 - P(\bar{Z}_1 \cap \bar{Z}_2 \cap \dots \cap \bar{Z}_q) \\ &= 1 - P(\bar{Z}_1) \cdot P(\bar{Z}_2|\bar{Z}_1) \\ &\quad \times P(\bar{Z}_3|\bar{Z}_1 \cdot \bar{Z}_2) \dots \\ &\quad \times P(\bar{Z}_q|\bar{Z}_1 \cdot \bar{Z}_2 \dots \bar{Z}_{q-1}). \end{aligned} \quad (1)$$

Since the above equation is expressed as a conditional probability, this also includes the case where the previous failure impacts the following failure rates. However, if each component's failure rate is independent, then the equation can be written as

$$\begin{aligned} R_s &= 1 - P(\bar{Z}_1) \cdot P(\bar{Z}_2) \cdot P(\bar{Z}_3) \dots P(\bar{Z}_q) \\ &= 1 - \prod_{i=1}^q P(\bar{Z}_i). \end{aligned} \quad (2)$$

Additionally, for a parallel system, the reliability of the system can be calculated from the reliability of its individual components.

$$R_s = 1 - \prod_{i=1}^q (1 - R_i). \quad (3)$$

We are going to develop the warranty cost model for the parallel system. First, we briefly review the warranty policy. In general, warranty is an obligation attached to products that require the manufacturers to provide compensation for consumers according to the warranty terms when the warranted products fail to perform their intended functions [3]. The warranty is nowadays important not only to the manufacturers but also to the buyers of any commercial product. The consumer is provided a resource for dealing with items that fail due to the uncertainty of product performance. The manufacturer is provided protection because warranty terms explicitly

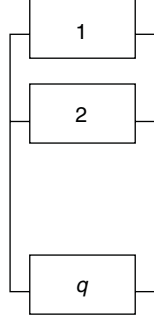


Figure 1. Parallel system with q components.

limit responsibility in terms of both time and type of product failure.

To develop the cost model for the series system, several assumptions have been made in this article as follows: (i) The repair cost and the replacement cost are constant. (ii) Repair service time and replacement service time are negligible. (iii) All warranty claims are executed and all claims are valid. (iv) Repairs are imperfect and the repair process can be modeled by a quasi-renewal process. (v) The time to perform an inspection in which to determine whether the failed component needs a repair or a replacement is negligible. (vi) The warranty period is renewable.

BACKGROUND

Several researchers [1,5–9] have studied the optimization of system reliability using various techniques and have investigated various characteristics and reliability of the parallel system. Yeh [7] calculates the rate of occurrence of failures of a continuous-time Markov chain and his result is applied to the parallel system. Nakagawa [6] considers the replacement policy for the parallel system in a random environment. Jonkers *et al.* [5] develop the set of machine model building blocks, which is a new algorithm for the prediction of multiple-class parallel section completion times. Asadi and Bayramoglu [1] propose the definition for the mean residual life function of a parallel system and obtain some of its properties. Bairamov *et al.* [8] define the survival and mean residual life

function of a system consisting of n identical and independent components having series or parallel structure. Shen and Xie [9] investigate the effect of such parallel redundancy upon system reliability when applied at various places and in various systems.

We consider that a replacement service would be possible during the warranty period by introducing two alternate quasi-renewal concepts based on the quasi-renewal process. The first is called an *altered quasi-renewal process*. In the quasi-renewal process model, upon each imperfect repair, the time to failure is reduced to a fraction α of the immediately previous one and is dependent on all previous ones. The altered quasi-renewal process, however, does not necessarily have the same patterns. The second is called a *mixed quasi-renewal process* by considering both repairs and replacements subject to imperfect strategies. Using these two alternate quasi-renewal processes, the cost analyses are conducted for a parallel system. For the warranty cost analysis, we obtain a distribution function of the number of failures for the warranty cost analysis, assuming that the warranty policy is renewable.

WARRANTY COST ANALYSIS

In this section, warranty cost analyses are conducted by obtaining the expected warranty cost as well as the variance of the warranty. We also refer to the article titled **Series Systems** in this encyclopedia for the distribution of N to derive several statistical properties of warranty cost function per cycle or per product sold.

Let $f_{ij}(\cdot)$, $F_{ij}(\cdot)$, and $R_{ij}(\cdot)$ be the probability density function (pdf), the cumulative density function (cdf), and reliability function of component j 's failure times after $(i - 1)$ th repair/replacement within a warranty period w , respectively. Additionally, let $f_{is}(\cdot)$, $F_{is}(\cdot)$, and $R_{is}(\cdot)$ be pdf, cdf, and reliability function of system failure times after $(i - 1)$ th repair/replacement within a warranty period w , respectively. β denotes a parameter for altered quasi-renewal process. Using Equation (3), the reliability function

of the j th component and i th failure inter-occurrence interval using the altered quasi-renewal process is as follows:

$$\begin{aligned} R_{is}(w) &= 1 - \prod_{j=1}^q (1 - R_{ij}(w)) \\ &= 1 - \prod_{j=1}^q \left(1 - \left(1 - \int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx \right) \right). \end{aligned} \quad (4)$$

The reliability function for $(n_s + 1)$ th inter-failure interval is

$$\begin{aligned} R_{(n_s+1)s}(w) &= 1 - \prod_{j=1}^q (1 - R_{(n_s+1)j}(w)) \\ &= 1 - \prod_{j=1}^q \left(\int \frac{1}{\beta_{(n_s+1)j}} f_{1j} \left(\frac{1}{\beta_{(n_s+1)j}} x \right) dx \right). \end{aligned} \quad (5)$$

For the cost analysis, we obtain the expected warranty cost function and the variance of the cost function. Let a random variable N_a be the number of repair within a warranty period. Using the pdf of number of failures from the article titled **Series Systems** in this encyclopedia, $f(n_a)$ of the parallel system failure time is given by

$$\begin{aligned} f(n_a) &= \left(\prod_{i=1}^{n_s} (F_{is}(w)) \right) (R_{(n_s+1)s}(w)) \\ &= \left[\left(\prod_{j=1}^q \left(\int f_{1j}(x) dx \right) \right) \right. \\ &\quad \times \prod_{i=2}^{n_s} \left(\prod_{j=1}^q \left(\int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx \right) \right) \Big] \\ &\quad \times \left[1 - \prod_{j=1}^q \left(\int \frac{1}{\beta_{(n_s+1)j}} \right. \right. \\ &\quad \left. \left. \times f_{1j} \left(\frac{1}{\beta_{(n_s+1)j}} x \right) dx \right) \right]. \end{aligned} \quad (6)$$

Using Equation (6), the expected number of warranty repair services is given by

$$E(N_a) = \int n_a \cdot f(n_a) \cdot dn_a. \quad (7)$$

Let r.v. N_b be the number of replacements within a warranty period, respectively. Similar to Equation (7), we obtain the expected number of replacement warranty services:

$$\begin{aligned} E(N_b) &= \int n_b \cdot \left(\prod_{i=1}^{n_b} \left(\int f_{is}(y) dy \right) \right) \\ &\quad \times \left(1 - \int f_{is}(y) dy \right) \cdot dn_b. \end{aligned} \quad (8)$$

Let c_a and c_b be warranty costs for repairs and replacements within a warranty period w respectively. And let C be the warranty system cost for a warranty period w . Eventually, the expected warranty cost for the parallel system is given by

$$E(C) = c_a E(N_a) + c_b E(N_b), \quad (9)$$

where $E(N_a)$ and $E(N_b)$ are given in Equations (7) and (8).

Additionally, using Equations (6) and (7), we can easily derive the second moment as $E(N_a^2) = \int n_a^2 \cdot f(n_a) \cdot dn_a$. Using the first moment and the second moment, the variance of the number of repair services is given by

$$\text{Var}(N_a) = E(N_a^2) - [E(N_a)]^2. \quad (10)$$

Also, the variance in the number of replacement services $\text{Var}(N_b)$ could be obtained by a similar way.

Note that

$$\begin{aligned} E[N_a N_b] &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b \\ &\quad \times P[N_a = n_a, N_b = n_b] \\ &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b P[N_s = n_s] \\ &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b \left(\prod_{i=1}^{n_a+n_b} (F_{is}(w)) \right) \\ &\quad \times (R_{(n_a+n_b+1)s}(w)). \end{aligned} \quad (11)$$

4 PARALLEL SYSTEMS

Using Equations (6) and (11), we also obtain

$$\begin{aligned}
 E[N_a N_b] &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b \prod_{j=1}^q \left(\int f_{1j}(x) dx \right) \\
 &\times \prod_{i=2}^{n_a+n_b} \left(\prod_{j=1}^q \left(\int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx \right) \right) \\
 &\times \left[1 - \prod_{j=1}^q \left(\int \frac{1}{\beta_{(n_a+n_b+1)j}} \right. \right. \\
 &\times \left. \left. f_{1j} \left(\frac{1}{\beta_{(n_a+n_b+1)j}} \right) dx \right) \right] \quad (12)
 \end{aligned}$$

Additionally, the covariance of the repair services and the replacement services is given as follows:

$$\text{Cov}(N_a, N_b) = E[N_a N_b] - E[N_a]E[N_b], \quad (13)$$

where $E[N_a], E[N_b]$ are given in Equations (7) and (8) and $E[N_a N_b]$ is given in Equation (12).

Combining Equations (10) and (13), the variance of warranty system cost is given by

$$\begin{aligned}
 \text{Var}(C) &= c_a^2 \cdot \text{Var}(N_a) + c_b^2 \cdot \text{Var}(N_b) \\
 &+ 2c_a c_b \cdot \text{Cov}(N_a, N_b). \quad (14)
 \end{aligned}$$

NUMERICAL EXAMPLES

In this section, a numerical example is presented to illustrate the warranty cost analysis for a two-component parallel system (see Fig. 2). We assume that the failure time of the system follows the exponential distribution with $\lambda = 1$.

Given $q = 2$ and $h = 1$. We assume that the warranty costs and β values are as follows:

$$\begin{aligned}
 c_a &= \$5,000, c_b = \$8,000 \quad \text{and} \\
 \beta_{i1} &= 0.99, \beta_{i2} = 0.98.
 \end{aligned}$$

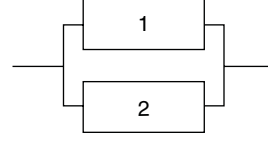


Figure 2. Parallel system with two components.

The results of the expected warranty cost, $E(C)$, and standard deviation, $SD(C)$, of warranty cost and coefficient of variation (CV) are listed in Table 1 for various values of the warranty period w .

In Fig. 3, the expected warranty cost increases as warranty time increases, which is similar to the plots of the series system. When they are dependent, it can be seen that both the expectation and the standard deviation of warranty cost functions increase monotonically over the warranty period. Furthermore, when they are independent, the CV decreases as the warranty time progresses.

Table 1. $E(C)$, $SD(C)$, and CV for the Parallel System

W	$E(C)$	SD^a	CV^a	SD^b	CV^b
0.1	39	579	15.02	125	3.24
0.2	126	616	4.90	308	2.46
0.3	231	897	3.88	497	2.15
0.4	349	1174	3.36	688	1.97
0.5	476	1436	3.02	878	1.85
0.6	609	1684	2.77	1067	1.75
0.7	747	1917	2.57	1254	1.68
0.8	889	2137	2.40	1440	1.62
0.9	1036	2344	2.26	1624	1.57
1	1185	2781	2.35	1806	1.52
1.1	1338	2724	2.04	1987	1.48
1.2	1494	2898	1.94	2166	1.45
1.3	1652	3061	1.85	2345	1.42
1.4	1813	3215	1.77	2522	1.39
1.5	1975	3360	1.70	2698	1.37
1.6	2140	3497	1.63	2873	1.34
1.7	2306	3625	1.57	3046	1.32
1.8	2475	3746	1.51	3219	1.30
1.9	2644	3859	1.46	3391	1.28
2	2816	3966	1.41	3563	1.27

^awhen repair and replacement are dependent.

^bwhen repair and replacement are independent.

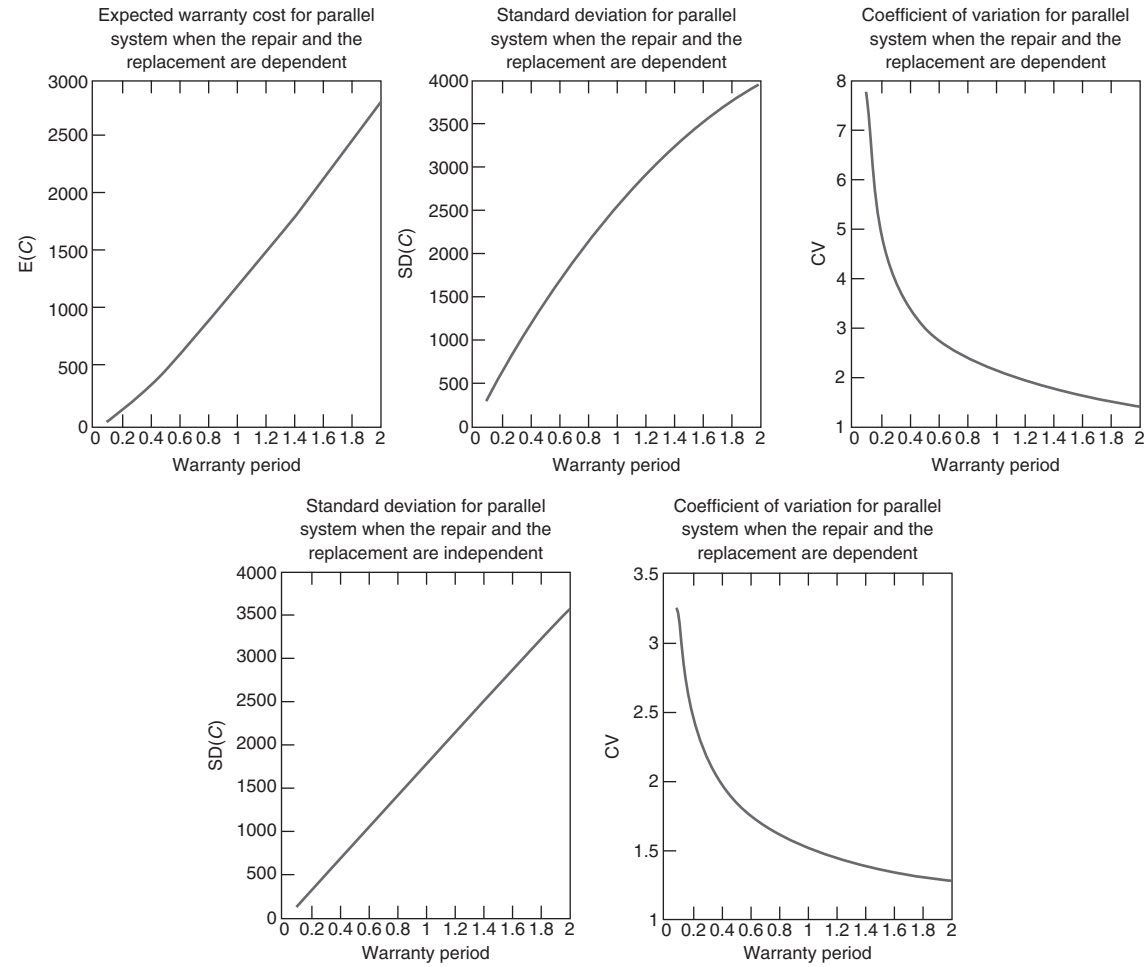


Figure 3. $E(C)$, $SD(C)$, and CV for the parallel system when the repair and the replacement are dependent or independent.

REFERENCES

1. Asadi M, Bairamov I. A note on the mean residual life function of a parallel system. *Commun Stat Theor Meth* 2005;34:475–484.
2. Pham H. *Handbook of reliability engineering*. Springer; 2003.
3. Pham H. *Handbook of engineering statistics*. Hoboken, NJ, USA: Springer; 2006.
4. http://www.weibull.com/SystemRelWeb/simple_parallel_systems.htm. 2009.
5. Jonkers H, Van Gemund A, Reijns G. A probabilistic approach to parallel system performance modelling. 1995. pp. 412–421.
6. Nakagawa T. Replacement problem of a parallel system in random environment. *J Appl Probab* 1979;203–205.
7. Yeh L. Calculating the rate of occurrence of failures for continuous-time Markov chains with application to a two-component parallel system. *J Oper Res Soc* 1995;528–536.
8. Bairamov I, Ahsanullah M, Akhundov I. A residual life function of a system having parallel or series structures. *J Stat Theor Appl* 2002;1:119–132.
9. Shen K, Xie M. On the increase of system reliability by parallel redundancy. *IEEE Trans Reliab* 1990;39:607–611.

PARALLEL-SERIES AND SERIES-PARALLEL SYSTEMS

MARCUS AGUSTIN
Department of Mathematics and
Statistics, Southern Illinois
University Edwardsville,
Edwardsville, Illinois

INTRODUCTION

In fields such as operations research, engineering, and biostatistics, it is of interest to consider a system where components are connected according to some structure. Two systems that are at the extreme ends of the spectrum are the series system and the parallel system. For a series system, it is necessary for all components to function in order for the system to function. On the other hand, for a parallel system, it only takes one functioning component in order for the system to function. For a more thorough discussion concerning the reliability of these systems, the reader is referred to the articles on *Systems in Series* and *Parallel Configurations*.

A series-parallel system and a parallel-series system are two different systems that result from combining the attributes of a series system and a parallel system. For a series-parallel system, subsystems are connected in series and each subsystem consists of components connected in parallel. The subsystem in a series-parallel system was described in Refs 1 and 2 as a redundant system since the components making up the subsystem can be viewed as similar or redundant components. A subsystem fails only if all of the redundant components have failed, resulting in failure of the main system. In contrast, a parallel-series system consists of subsystems connected in parallel, and each subsystem is made up of components connected in series. Hence, failure of a component in a subsystem results in failure

of the subsystem. However, all subsystems must fail in order for the main system to fail.

One of the earliest papers that considered series-parallel and parallel-series systems is Ref. 3, where a system consisting of m subsystems with each subsystem containing n components was considered. The type of component failure may result in two types of system failure. A component that exhibited a type 1 failure would cause the failure of the subsystem containing the component, while failure of the second type of all components in a subsystem resulted in subsystem failure. Whenever each subsystem has failed due to a type 1 failure, system failure is observed. This is synonymous to the failure of a parallel-series system. On the other hand, a type 2 failure of a single subsystem resulted in system failure. The event is similar to the failure of a series-parallel system. One of the aims in this paper was to find the optimal number of subsystems in order to maximize the probability that the system will function. On the other hand, a special form of a series-parallel system was considered by Boland *et al.* [4] to examine the allocation of a redundant or spare component for a system of n components that functions if $k \leq n$ components function. For such systems, it was shown that in order to optimize system lifetime it is stochastically preferable to place a spare component in parallel to the weakest component.

Majority of the work that dealt with series-parallel or parallel-series systems either investigated the problem of finding the number of components or subsystems required to optimize a given function (e.g., probability that system will function, operational cost associated with system function, etc.) or how to allocate the components in order to make use of the redundant nature of a parallel subsystem. A partial, but not exhaustive, list of papers on finding the number of components or subsystems for a series-parallel or parallel-series system in order to either maximize the probability

of system function or minimize system cost include Refs 5–8. The allocation of components in series-parallel or parallel-series systems via redundancy optimization was also considered, among others, by Elegbede *et al.* [9], Nourelfath and Ait-Kadi [10], Tavakkoli-Moghaddam [11], Wang and Watada [12], and Li *et al.* [13].

In this article, we will obtain the system reliability of a parallel-series system as well as the system reliability of a series-parallel system. Furthermore, we will consider the setting where the reliability of each component is a function of how long a component has been in use.

SYSTEM RELIABILITY

Structure Function

Consider a system of n components, where a component is either in a “functioning” state or in a “failed” state. We will consider the random variables $X_1, \dots, X_n$ assumed to be mutually independent and X_i is a binary random variable such that

$$X_i = \begin{cases} 1, & \text{if component } i \text{ is functioning,} \\ 0, & \text{otherwise.} \end{cases}$$

We will define $\mathbf{X} = (X_1, \dots, X_n)$ to be the state vector of the system. The state of the system is referred to as the *structure function* of the system and is denoted by $\phi(\mathbf{X})$. The function $\phi(\mathbf{X})$ is a binary random variable where

$$\phi(\mathbf{X}) = \begin{cases} 1, & \text{if the system is functioning,} \\ 0, & \text{otherwise.} \end{cases}$$

A primary interest in the study of systems in operations research and reliability is to examine systems, where replacing a failed component by a functioning component does not result in the system being “worse off.” In other words, interest is on systems whose structure functions are nondecreasing. Such a system is called a *coherent system*. For a more thorough discussion of the structure function of a coherent system as well as its properties, the reader is referred to **Coherent Systems**.

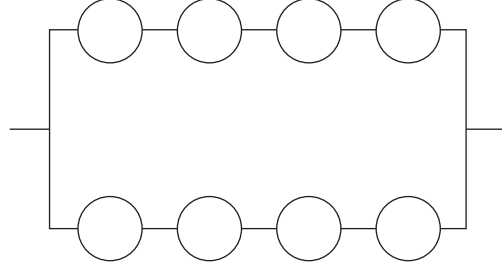


Figure 1. Diagram for a parallel-series system with $m = 2$ subsystems and $k = 4$ components per subsystem.

Parallel-Series System

A parallel-series system with mk components where k components are in each of the m subsystems is a system that functions if and only if at least one subsystem is functioning. Thus, the m subsystems form a parallel system. However, within each subsystem, the k components are connected in series so a subsystem will function only if all components are functioning. Figure 1 presents a diagram for a parallel-series system with $m = 2$ subsystems and $k = 4$ components per subsystem.

In order to find the structure function $\phi(\mathbf{X})$ of a parallel-series system, let us first represent the structure function of the j th subsystem by $\phi_j(\mathbf{X}_j)$ for $j = 1, \dots, m$. The vector $\mathbf{X}_j = (X_{j1}, \dots, X_{jk})$ is the state vector of the j th subsystem and X_{ji} is the 0-1 random variable representing the state of the i th component in the j th subsystem, where $X_{ji} = 1$ if component i in subsystem j is functioning. Since the components in a subsystem are connected in series, it follows from the results presented in the **Systems in Series** that the structure function of the j th subsystem is

$$\phi_j(\mathbf{X}_j) = \prod_{i=1}^k X_{ji} = \min(X_{j1}, \dots, X_{jk}). \quad (1)$$

Letting $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_m)$ denote the state vector of the system, the structure function of a parallel-series system, following results

presented in **Parallel Configurations**, is

$$\begin{aligned}\phi(\mathbf{X}) &= 1 - \prod_{j=1}^m (1 - \phi_j(\mathbf{X}_j)) \\ &= 1 - \prod_{j=1}^m \left(1 - \left[\prod_{i=1}^k X_{ji} \right] \right).\end{aligned}\quad (2)$$

An alternative representation for the structure function of a parallel-series system is

$$\phi(\mathbf{X}) = \max_{1 \leq j \leq m} \left[\min_{1 \leq i \leq k} X_{ji} \right].$$

As an illustration, consider a parallel-series system with $m = 2$ subsystems and $k = 4$ components per subsystem. If $X_{11} = 0$ and $X_{22} = 0$ while the remaining X_{ji} 's are all equal to 1, then $\phi_1(\mathbf{X}_1) = \min(0, 1, 1, 1) = 0$ and $\phi_2(\mathbf{X}_2) = \min(1, 0, 1, 1) = 0$, resulting in $\phi(\mathbf{X}) = 0$. On the other hand, if $X_{11} = 0$ and $X_{13} = 0$ while the remaining X_{ji} 's are all 1, then $\phi_1(\mathbf{X}_1) = \min(0, 1, 0, 1) = 0$ and $\phi_2(\mathbf{X}_2) = \min(1, 1, 1, 1) = 1$, giving us $\phi(\mathbf{X}) = 1$, a functioning system.

Series-Parallel System

We define a series-parallel system with mk components as a system comprised of m subsystems connected in series and each of the m subsystems consists of k components connected in parallel. From the point of view of failure, a subsystem will fail if all of the components fail, and as soon as one subsystem fails, the series-parallel system will fail. Hence, a subsystem functions if at least one of its k components functions. However, the whole system functions only if all m subsystems function. Using the same notation in the section titled "Parallel-Series System" and using the results in **Parallel Configurations**, the structure function of the j th subsystem is

$$\phi_j(\mathbf{X}_j) = 1 - \prod_{i=1}^k (1 - X_{ji}) = \max(X_{j1}, \dots, X_{jk}).$$

Invoking the results in **Systems in Series**, the structure function of the series-parallel

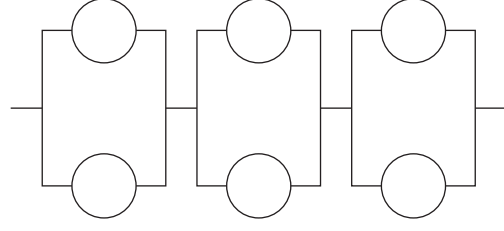


Figure 2. Diagram for a series-parallel system with $m = 3$ subsystems and $k = 2$ components per subsystem.

system is expressed as

$$\phi(\mathbf{X}) = \prod_{j=1}^m \phi_j(\mathbf{X}_j) = \prod_{j=1}^m \left[1 - \prod_{i=1}^k (1 - X_{ji}) \right].$$

Alternatively, the structure function of the series-parallel system can be written as

$$\phi(\mathbf{X}) = \min_{1 \leq j \leq m} \left[\max_{1 \leq i \leq k} X_{ji} \right].$$

For a concrete example of a series-parallel system, consider a series-parallel system with $m = 3$ subsystems, with each subsystem consisting of $k = 2$ components. An illustration of this system is given in Fig. 2. If $X_{11} = 0, X_{12} = 0$, and $X_{22} = 0$ while the remaining X_{ji} are all equal to 1, then $\phi_1(X_{11}, X_{12}) = \max(0, 0) = 0$, $\phi_2(X_{21}, X_{22}) = \max(1, 0) = 1$, and $\phi_3(X_{31}, X_{32}) = \max(1, 1) = 1$. Thus, one can see that the first subsystem has failed while the remaining subsystems are still functioning. Consequently, $\phi(\mathbf{X}) = \min(0, 1, 1) = 0$, which indicates that the whole system has failed. On the other hand, if $X_{11} = X_{21} = X_{31} = 0$, then $\phi_1(X_{11}, X_{12}) = \max(0, 1) = 1$, $\phi_2(X_{21}, X_{22}) = \max(1, 0) = 1$, and $\phi_3(X_{31}, X_{32}) = \max(0, 1) = 1$; that is, all three subsystems are functioning. Hence, $\phi(\mathbf{X}) = \min(1, 1, 1) = 1$ which shows that the series-parallel system is still functioning.

System Reliability

Consider either a series-parallel system or a parallel-series system that is observed at a specified time instant t . In order to indicate the time-dependence, let $X_{ij}(t)$

indicate the state of the i th component in the j th subsystem at time t . The corresponding state vector of the j th subsystem is $\mathbf{X}_j(t) = (X_{j1}(t), \dots, X_{jk}(t))$. Let $p_{ji}(t) = P\{X_{ji}(t) = 1\}$ be the probability that the i th component in the j th subsystem is functioning at time t , for $j = 1, \dots, m$ and $i = 1, \dots, k$. Let $\mathbf{p}_j(t) = (p_{j1}(t), \dots, p_{jk}(t))$ be the vector of probabilities for subsystem j and $\mathbf{p}(t) = (\mathbf{p}_1(t), \dots, \mathbf{p}_m(t))$ be the vector of probabilities for the m subsystems. The reliability of a system at time t is defined to be

$$h(\mathbf{p}(t)) = P\{\phi(\mathbf{X}(t)) = 1\} = E[\phi(\mathbf{X}(t))].$$

The random variables $X_{ji}(t)$ s are independent. To find the reliability of a parallel-series system, we first note that the probability that the j th subsystem is functioning at t is

$$\begin{aligned} P\{\phi_j(\mathbf{X}_j(t)) = 1\} &= P\{\min(X_{j1}(t), \dots, X_{jk}(t)) = 1\} \\ &= P\{X_{j1}(t) = 1, \dots, X_{jk}(t) = 1\} \\ &= \prod_{i=1}^k p_{ji}(t). \end{aligned} \quad (3)$$

On the other hand, the probability that the system is functioning at time t is equal to the probability that at least one subsystem is functioning at that time instant. Owing to the fact that $\phi(\mathbf{X}(t))$ is either 0 or 1, it follows that $h(\mathbf{p}(t)) = 1 - P\{\phi(\mathbf{X}(t)) = 0\}$. Since the subsystems will also be mutually independent, then

$$\begin{aligned} h(\mathbf{p}(t)) &= 1 - P\{\phi_1(\mathbf{X}_1(t)) = 0, \dots, \phi_m(\mathbf{X}_m(t)) = 0\} \\ &= 1 - \prod_{j=1}^m P\{\phi_j(\mathbf{X}_j(t)) = 0\}. \end{aligned} \quad (4)$$

Combining Equations (3) and (4), the reliability of a parallel-series system is

$$\begin{aligned} h(\mathbf{p}(t)) &= 1 - \prod_{j=1}^m [1 - P\{\phi_j(\mathbf{X}_j(t)) = 1\}] \\ &= 1 - \prod_{j=1}^m \left[1 - \prod_{i=1}^k p_{ji}(t) \right]. \end{aligned} \quad (5)$$

For a series-parallel system, we note that since the components in a subsystem are connected in parallel, then

$$\begin{aligned} P\{\phi_j(\mathbf{X}_j(t)) = 1\} &= 1 - P\{\phi_j(\mathbf{X}_j(t)) = 0\} \\ &= 1 - P\left\{ \bigcup_{i=1}^k (X_{ji}(t) = 0) \right\}. \end{aligned}$$

Since components are mutually independent, it follows that

$$\begin{aligned} P\{\phi_j(\mathbf{X}_j(t)) = 1\} &= 1 - P\{X_{j1}(t) = 0\} \cdots \\ &\quad \times P\{X_{jk}(t) = 0\} \\ &= 1 - \prod_{i=1}^k (1 - p_{ji}(t)). \end{aligned} \quad (6)$$

Furthermore, the subsystems are connected in series which results in

$$\begin{aligned} h(\mathbf{p}(t)) &= P\{\phi_1(\mathbf{X}_1(t)) = 1, \dots, \phi_m(\mathbf{X}_m(t)) = 1\} \\ &= \prod_{j=1}^m P\{\phi_j(\mathbf{X}_j(t)) = 1\}. \end{aligned} \quad (7)$$

Thus, using Equations (6) and (7), the reliability of a series-parallel system at time t is obtained to be

$$\begin{aligned} h(\mathbf{p}(t)) &= \left[1 - \prod_{i=1}^k (1 - p_{1i}(t)) \right] \cdots \\ &\quad \times \left[1 - \prod_{i=1}^k (1 - p_{mi}(t)) \right] \\ &= \prod_{j=1}^m \left[1 - \prod_{i=1}^k (1 - p_{ji}(t)) \right]. \end{aligned} \quad (8)$$

We can generalize the notion of the state of a component by considering the random variable T_{ji} to denote the lifetime of the component i in subsystem j , $j = 1, \dots, m$ and $i = 1, \dots, k$, where $f_{ji}(t; \theta)$ is the corresponding probability density function that is parameterized by some parameter vector θ . From Ref. 14, the probability that this component is functioning at time t is

$$p_{ji}(t; \theta) \equiv P\{T_{ji} > t\} = \int_t^\infty f_{ji}(u; \theta) du. \quad (9)$$

For a parallel-series system, we define the lifetime of subsystem j by Y_j and the lifetime of the system by Y . From Equation (1), we get $Y_j = \min(T_{j1}, \dots, T_{jk})$ and while using Equation (2), we obtain $Y = \max(Y_1, \dots, Y_m)$. Since the component lifetimes are independent and $P\{Y_j > t\} = \prod_{i=1}^k P\{T_{ji} > t\}$, the system reliability in Equation (5) simplifies to

$$\begin{aligned} h(\mathbf{p}(t; \theta)) &= 1 - \prod_{j=1}^m \left[1 - \prod_{i=1}^k P\{T_{ji} > t\} \right] \\ &= 1 - \prod_{j=1}^m \left[1 - \prod_{i=1}^k p_{ji}(t; \theta) \right]. \end{aligned} \quad (10)$$

For a series-parallel system, the lifetime of the j th subsystem is $Y_j = \max(T_{j1}, \dots, T_{jk})$ and the resulting system lifetime is $Y = \min(Y_1, \dots, Y_m)$. Since

$$\begin{aligned} P\{Y_j > t\} &= 1 - P\{\max(T_{j1}, \dots, T_{jk}) \leq t\} \\ &= 1 - \prod_{i=1}^k P\{T_{ji} \leq t\}, \end{aligned}$$

it follows from Equation (8) that the reliability of a series-parallel system at time t is

$$\begin{aligned} h(\mathbf{p}(t; \theta)) &= \prod_{j=1}^m \left[1 - \prod_{i=1}^k P\{T_{ji} \leq t\} \right] \\ &= \prod_{j=1}^m \left[1 - \prod_{i=1}^k [1 - p_{ji}(t; \theta)] \right]. \end{aligned} \quad (11)$$

NUMERICAL EXAMPLE

As an illustration, consider the case when each component lifetime follows an exponential distribution with mean $1/\lambda_{ji}$, that is, $T_{ji} \sim \text{EXP}(\lambda_{ji})$. Hence, the random variable T_{ji} has density function

$$f_{ji}(t; \lambda_{ji}) = \lambda_{ji} \exp(-\lambda_{ji} t), \quad \lambda_{ji} > 0, \quad t > 0.$$

A partial list of papers that considered the exponential distribution for series-parallel or parallel-series systems include Refs 8 and 15–18. For further discussion

regarding the exponential distribution, the reader is referred to **Systems in Series**.

Using Equation (9), the probability that component i in the j th subsystem is functioning at time t is

$$\begin{aligned} p_{ji}(t; \lambda_{ji}) &= \int_t^\infty \lambda_{ji} \exp(-\lambda_{ji} u) du \\ &= \exp(-\lambda_{ji} t). \end{aligned} \quad (12)$$

Let $\lambda = (\lambda_{11}, \dots, \lambda_{1k}, \lambda_{21}, \dots, \lambda_{2k}, \dots, \lambda_{m1}, \dots, \lambda_{mk})$ be the vector of parameters. For a parallel-series system, the system reliability given in Equation (10) simplifies to

$$\begin{aligned} h(\mathbf{p}(t; \lambda)) &= 1 - \prod_{j=1}^m \left[1 - \prod_{i=1}^k \exp(-\lambda_{ji} t) \right] \\ &= 1 - \prod_{j=1}^m \left[1 - \exp\left(-\sum_{i=1}^k \lambda_{ji} t\right) \right]. \end{aligned}$$

On the other hand, using Equations (11) and (12), the reliability of a series-parallel system is obtained to be

$$h(\mathbf{p}(t; \lambda)) = \prod_{j=1}^m \left[1 - \prod_{i=1}^k (1 - \exp(-\lambda_{ji} t)) \right].$$

As a special case, suppose $T_{ji} \sim \text{EXP}(\beta)$, that is, component lifetimes are independent and identically distributed with mean $1/\beta$. The system reliability of a parallel-series system simplifies to

$$h(\mathbf{p}(t; \beta)) = 1 - [1 - \exp(-k(\beta t))]^m,$$

while the system reliability of a series-parallel system simplifies to

$$h(\mathbf{p}(t; \beta)) = [1 - (1 - \exp(-\beta t))^k]^m.$$

REFERENCES

1. Jensen PA. Optimization of series-parallel-series networks. *Oper Res* 1970;18:471–482.
2. Monga A, Zuo MJ. Optimal design of series-parallel systems considering maintenance and salvage value. *Comput Ind Eng* 2001;40:323–337.

3. Barlow RE, Hunter LC, Proschan F. Optimum redundancy when components are subject to two kinds of failure. *J Soc Ind Appl Math* 1963;11:64–73.
4. Boland PJ, El-Newehi E, Proschan F. Stochastic order for redundancy allocation in series and parallel systems. *Adv Appl Probab* 1992;24:161–171.
5. El-Newehi E, Proschan F, Sethuraman J. Optimal allocation of components in parallel-series and series-parallel systems. *J Appl Probab* 1986;23:770–777.
6. Rajendra Prasad V, Nair KPK, Aneja YP. Optimal assignment of components to parallel-series and series-parallel systems. *Oper Res* 1991;39:407–414.
7. Ramirez-Marquez JE, Coit DW. Optimization of system reliability in the presence of common cause failures. *Reliab Eng Syst Saf* 2007;92:1421–1434.
8. Tian Z, Levitin G, Zuo MJ. A joint reliability-redundancy optimization approach for multi-state series-parallel systems. *Reliab Eng Syst Saf* 2009;94:1568–1576.
9. Elegbede AOC, Chu C, Adjallah KH, *et al.* Reliability allocation through cost minimization. *IEEE Trans Reliab* 2003;52:106–111.
10. Noureldath M, Ait-Kadi D. Optimization of series-parallel multi-state systems under maintenance policies. *Reliab Eng Syst Saf* 2007;92:1620–1626.
11. Tavakkoli-Moghaddam R, Safari J, Sassani F. Reliability optimization of series-parallel systems with a choice of redundancy strategies using genetic algorithm. *Reliab Eng Syst Saf* 2008;93:550–556.
12. Wang S, Watada J. Modelling redundancy allocation for a fuzzy random parallel-series system. *J Comput Appl Math* 2009; 232:539–557.
13. Li C, Chen X, Yi X, *et al.* Heterogeneous redundancy optimization for multi-state series-parallel systems subject to common cause failures. *Reliab Eng Syst Saf* 2010;95:202–207.
14. Leemis L. Reliability: probabilistic models and statistical methods. New Jersey: Prentice Hall; 1995.
15. Revyakov M. Component allocation for a distributed system: reliability maximization. *J Appl Probab* 1993;30:471–477.
16. Sarhan AM, Al-Ruzaiza AS, Alwasel IA, *et al.* Reliability equivalence of a series-parallel system. *Appl Math Comput* 2004;154:257–277.
17. Sarhan AM. Reliability equivalence factors of a general series-parallel system. *Reliab Eng Syst Saf* 2009;94:229–236.
18. Levitin G, Amari SV. Optimal load distribution in series-parallel systems. *Reliab Eng Syst Saf* 2009;94:254–260.

PARAMETRIC LP ANALYSIS

ALLEN HOLDER
Department of Mathematics,
Rose-Hulman Institute of
Technology, Terre Haute,
Indiana

The study of parametric linear programming dates back to the work of Gass and Saaty [1], Mills [2] and Saaty and Gass [3] in the middle 1950s (see also Refs 4–6). The analysis of how optimal properties depend on a problem's data is important since it allows a model to be used for its intended purpose of explaining the underlying phenomena. This is because models are often constructed with imperfect information, and the study of parametric and sensitivity analysis relates optimal properties to the problem's data description. The topic is a mainstay in introductory texts on operations research and linear programming. Here we advance the typical introduction and point to some modern applications. For the sake of brevity we omit proofs and instead cite publications in which proofs can be located.

We assume the standard form primal and dual throughout,

$$(LP) \min\{c^T x : Ax = b, x \geq 0\} \text{ and}$$

$$(LD) \max\{b^T y : A^T y + s = c, s \geq 0\},$$

where $A \in \mathbb{R}^{m \times n}$ has full row rank, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$. For any $B \subseteq \{1, 2, \dots, n\}$ we let A_B be the submatrix of A whose column indices are in B . We further let $N = \{1, 2, \dots, n\} \setminus B$ so that A_N contains the columns of A not in A_B . Similarly, c_B and c_N denote the subvectors of c whose components are in the set subscripts. The partition (B, N) is *optimal* if both

$$A_B x_B = b, x_B \geq 0 \text{ and}$$

$$A_B^T y = c_B, A_N^T y + s_N = c_N, s_N \geq 0 \quad (1)$$

are consistent. One special case is if A_B is invertible, and we refer to such a partition as

basic or as a *basis*. If a basis is optimal, the above systems can be expressed as

$$A_B^{-1}b \geq 0 \text{ and } c_N - A_N^T(A_B^T)^{-1}c_B \geq 0. \quad (2)$$

Another special case is if B is *maximal*, meaning that it is not contained in another optimal B . There is a unique maximal B for any A , b , and c , and we denote the corresponding partition as $(\hat{B}, \hat{N})$. This fact was first established in Ref. 7. We mention that the terminology used here differs from that in some of the supporting literature. The maximal partition is often called the *optimal partition*, but since it is generally only one of many optimal partitions, we distinguish it with the term *maximal* since this is the mathematical property it has in relation to the other optimal partitions. An optimal solution to (LP) and (LD) can be constructed for any optimal (B, N) by letting y , x_B , and s_N be any solution to Equation (1) and by letting $x_N = 0$ and $s_B = 0$.

The classical study in parametric analysis considers linear changes in one of b or c , and it is this case that we primarily study. Let $\mathcal{B} \in \mathbb{R}^m$ and consider the parametrized linear program

$$z^*(\theta) = \min\{c^T x : Ax = b + \theta\mathcal{B}, x \geq 0\}.$$

A typical question is, "Over what range of θ does an optimal basis remain optimal?" Since the dual inequality in Equation (2) is unaffected by a change in b , we have that a basic optimal solution remains optimal so long as

$$x_B(\theta) = A_B^{-1}(b + \theta\mathcal{B}) \geq 0.$$

Hence the basis remains optimal as long as θ is at least

$$\begin{aligned} & \max \left\{ -[A_B^{-1}b]_i / \mathcal{B}_i : \mathcal{B}_i > 0 \right\} \\ & = \min \{ \theta : A_B x_B = b + \theta\mathcal{B}, x \geq 0, x_N = 0 \} \end{aligned} \quad (3)$$

and at most

$$\begin{aligned} & \min \left\{ -[A_B^{-1}b]_i / \mathcal{B}_i : \mathcal{B}_i < 0 \right\} \\ & = \max \{ \theta : A_B x_B = b + \theta \mathcal{B}, x \geq 0, x_N = 0 \}. \end{aligned} \quad (4)$$

The left-hand side of these two equalities is a common ratio test, but it is insightful to realize that these equate to the stated linear programs. The objective value's response over this range is

$$z^*(\theta) = c_B^T x_B(\theta) = c_B^T A_B^{-1} b + \theta c_B^T \mathcal{B},$$

which shows that the objective function is linear as long as the basis remains optimal.

The above calculation is often used to support the result that $z^*(\theta)$ is continuous, piecewise linear, and convex. Although true, the above analysis of a basic optimal solution does not immediately lead to a characterization of z^* since bases need not remain optimal as long as z^* is linear. As an illustration, we consider the following example throughout,

$$A = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad b = \begin{pmatrix} 1 \\ 3 \end{pmatrix},$$

$$\mathcal{B} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \text{ and } c = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The unique optimal basis for $\theta = 0$ is $B = \{2, 3\}$, and this basis remains optimal as long as $0 \leq \theta \leq 2$. However, for $2 < \theta < 3$ the unique optimal basis is $B = \{1, 2\}$, and we have

$$x_B(\theta) = x_B(0) + \theta A_B^{-1} \mathcal{B} = \begin{pmatrix} -2 \\ 3 \end{pmatrix} + \theta \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

For $\theta > 3$ the unique optimal basis is $\{1, 4\}$, and

$$x_B(\theta) = x_B(0) + \theta A_B^{-1} \mathcal{B} = \begin{pmatrix} 1 \\ -3 \end{pmatrix} + \theta \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

For these three bases we have

$$B = \{2, 3\} \Rightarrow c_B^T A_B^{-1} \mathcal{B} = -1$$

$$B = \{1, 2\} \Rightarrow c_B^T A_B^{-1} \mathcal{B} = -1$$

$$B = \{1, 4\} \Rightarrow c_B^T A_B^{-1} \mathcal{B} = 0,$$

and the objective value is

$$z^*(\theta) = \begin{cases} 3 - \theta, & 0 \leq \theta \leq 3 \\ 0, & \theta > 3. \end{cases}$$

The *linearity intervals* of z^* are $[0, 3]$ and $[3, \infty)$, and over the interior of the first we have $dz^*/d\theta = -1$ and over the interior of the second we have $dz^*/d\theta = 0$. This example shows that a linearity interval is not necessarily identified from the initial collection of optimal bases, that is, no basis that is optimal for $\theta = 0$ remains optimal over the first linearity interval. The intersection of two adjoining linearity intervals is called a *break point*, and two linearity intervals sharing a break point are said to *adjoin*. Although the objective z^* is not differentiable at a break point, the left and right derivatives exist as long as feasibility is possible on both sides of the break point. Additional examples of this behavior are found in Ref. 8.

The observation that no basis is guaranteed to remain optimal over a linearity interval leads to the question of how to characterize the behavior of the objective value under parametrization. One solution to this question was motivated by the advent of interior-point methods in the 1980s and 1990s. Specifically, we consider the optimal solution that is the limit of the central path. The central path is the collection of unique solutions to

$$Ax = b, \quad x > 0, \quad A^T y + s = c, \quad s > 0,$$

$$x_i s_i = \mu > 0, \quad \forall i,$$

which we denote by the parametrization $(x(\mu), y(\mu), s(\mu))$. The path converges as $\mu \downarrow 0$ to a unique optimal solution (x^*, y^*, s^*) up to the representation of the problem defined by A , b , and c . This solution induces $(\hat{B}, \hat{N})$ as follows:

$$\hat{B} = \{i : x_i^* > 0\} \text{ and } \hat{N} = \{i : x_i^* = 0\}.$$

The solution x^* is continuous with respect to b [9], which immediately implies $z^*(\theta)$ is continuous in θ .

Returning to the example, we have $\hat{B} = \{1, 2, 3\}$ for $0 \leq \theta < 3$, since each variable indexed by this set can be positive

in an optimal solution. For example, $x^*(\theta) = (5, 3 - \theta, 7 - \theta, 0)$ is optimal over this range. The variable x_2 is forced to zero once θ reaches 3, and the maximal partition changes to $\hat{B} = \{1, 3\}$ and $\hat{N} = \{2, 4\}$. For $\theta > 3$, the maximal partition is $\hat{B} = \{1, 4\}$ and $\hat{N} = \{2, 3\}$. This suggests that the maximal partition remains constant over the interior of each linearity interval. This is indeed true, and an analysis based on the maximal partition characterizes $z^*(\theta)$ since it algebraically describes the optimal sets of (LP) and (LD), which are respectively

$$\mathcal{P}^* = \{x : Ax = b, x \geq 0, x_{\hat{N}} = 0\} \quad \text{and} \\ \mathcal{D}^* = \{(y, s) : A^T y + s = c, s_{\hat{B}} = 0\}.$$

The counterparts to the right-hand sides of Equations (3) and (4) are to solve

$$\min \{\theta : Ax = b + \theta \mathcal{B}, x \geq 0, x_{\hat{N}} = 0\} \quad \text{and} \\ \max \{\theta : Ax = b + \theta \mathcal{B}, x \geq 0, x_{\hat{N}} = 0\}.$$
 (5)

As long as θ is between these optimal values, the objective z^* is linear, and the derivative of z^* over the interior of this range is

$$D = \max\{\mathcal{B}^T y : A^T y + s = c, s \geq 0, s_{\hat{B}} = 0\},$$
 (6)

which is a solution to a linear program defined over the dual optimal set. For the example, the optimal value of this math program is -1 for $\hat{B} = \{1, 2, 3\}$, which is the maximal partition for $0 < \theta < 3$ and is 0 for $\hat{B} = \{1, 4\}$, which is the maximal partition for $3 < \theta$.

The linear programs in Equations (5) and (6) are stated in terms of $(\hat{B}, \hat{N})$, and although mathematically correct, these sets are non-trivial to calculate due to numerical round-off, presolving, and scaling. An alternative is to replace the linear programs in Equations (5) and (6) with

$$\min \{\theta : Ax = b + \theta \mathcal{B}, x \geq 0, c^T x = z^*(0) + \theta D\}, \\ \max \{\theta : Ax = b + \theta \mathcal{B}, x \geq 0, c^T x = z^*(0) + \theta D\}, \\ \text{and} \\ \max\{\mathcal{B}^T y : A^T y + s = c, s \geq 0, b^T y = z^*(0)\}.$$

The equalities guaranteeing optimality can further be relaxed to $c^T x \leq z^*(0) + \theta D + \varepsilon$ and $b^T y \geq z^*(0) - \varepsilon$ to improve numerical stability if needed, where the parameter ε approximates the numerical tolerance of zero.

Results under the parametrization $c + \lambda \mathcal{C}$ mirror those for the right-hand side parametrization; the main difference being that the roles of the primal and dual reverse. In this case, we indicate the dependence on λ by using $z^*(\lambda)$ instead of $z^*(\theta)$. The following theorem presents the results for changes in one of θ or λ . See Refs 10–15 for various proofs.

Theorem 1. *Let $\mathcal{B}$ and $\mathcal{C}$ be directions of perturbation for b and c , and let $(\hat{B}, \hat{N})$ be the maximal partition for the unperturbed linear programs. Then the following hold:*

1. $(\hat{B}, \hat{N})$ remains the unique maximal partition under the parametrization $b + \theta \mathcal{B}$ as long as

$$\theta^- = \min\{\theta' : Ax = b + \theta' \mathcal{B}, x \geq 0, \\ x_{\hat{N}} = 0\} \\ < \theta < \max\{\theta' : Ax = b + \theta' \mathcal{B}, \\ x \geq 0, x_{\hat{N}} = 0\} = \theta^+,$$

provided that one of θ^- or θ^+ is nonzero.

2. Either $\theta^- = \theta^+ = 0$ or $\theta^- < 0 < \theta^+$.
3. The largest interval containing zero over which $z^*(\theta)$ is linear is $[\theta^-, \theta^+]$.
4. The dual optimal set, $\{(y, s) : A^T y + s = c, s \geq 0, s_{\hat{B}} = 0\}$, is invariant for $\theta \in (\theta^-, \theta^+)$, provided that θ^+ and θ^- are nonzero.
5. Assume θ^- and θ^+ are nonzero. Let $(\hat{B}', \hat{N}')$ and $(\hat{B}'', \hat{N}'')$ be the respective maximal partitions for $\theta = \theta^-$ and $\theta = \theta^+$. Then, $\hat{B}' \subset \hat{B}$, $\hat{N}' \supset \hat{N}$, $\hat{B}'' \subset \hat{B}$, and $\hat{N}'' \supset \hat{N}$.
6. If θ^- and θ^+ are nonzero and $\theta \in (\theta^-, \theta^+)$, then

$$\frac{dz^*(\theta)}{d\theta} = \max\{\mathcal{B}^T y : A^T y + s = c, s \geq 0, \\ s_{\hat{B}} = 0\}.$$

7. $(\hat{B}, \hat{N})$ remains the unique maximal partition under the parametrization $c + \lambda \hat{c}$ as long as

$$\begin{aligned} \lambda^- &= \min\{\lambda' : A^T y + s = c + \lambda' \hat{c}, s \geq 0, \\ &\quad s_{\hat{B}} = 0\} \\ &< \lambda < \max\{\lambda' : A^T y + s = c + \lambda' \hat{c}, \\ &\quad s \geq 0, s_{\hat{B}} = 0\} = \lambda^+, \end{aligned}$$

provided that one of λ^- or λ^+ is nonzero.

8. Either $\lambda^- = \lambda^+ = 0$ or $\lambda^- < 0 < \lambda^+$.
9. The largest interval containing zero over which $z^*(\lambda)$ is linear is $[\lambda^-, \lambda^+]$.
10. The primal optimal set, $\{x : Ax = b, x \geq 0, x_{\hat{N}} = 0\}$, is invariant for $\lambda \in (\lambda^-, \lambda^+)$, provided that λ^- and λ^+ are nonzero.
11. Assume λ^- and λ^+ are nonzero. Let $(\hat{B}', \hat{N}')$ and $(\hat{B}'', \hat{N}'')$ be the respective maximal partitions for $\lambda = \lambda^-$ and $\lambda = \lambda^+$. Then, $\hat{B}' \supset \hat{B}$, $\hat{N}' \subset \hat{N}$, $\hat{B}'' \supset \hat{B}$, and $\hat{N}'' \subset \hat{N}$.
12. If λ^- and λ^+ are nonzero and $\lambda \in (\lambda^-, \lambda^+)$, then

$$\begin{aligned} \frac{dz^*(\lambda)}{d\lambda} &= \min\{\hat{c}^T x : Ax = b, x \geq 0, \\ &\quad x_{\hat{N}} = 0\}. \end{aligned}$$

Theorem 1 states that the maximal partition characterizes the linearity of z^* about $\theta = 0$. Compared to the basis approach in Equations (3) and (4), the calculation guaranteeing the identification of the entire linearity interval requires the solution of two linear programs. Theorem 1 also distinguishes the roles of the primal and dual. The fourth and tenth statements show that one of the primal or the dual optimal sets is invariant over the interior of a linearity interval. For example, under the parametrization of b the dual optimal set remains intact, and hence, any basis within the maximal partition corresponds to a vertex of the dual optimal set independent of the θ in (θ^-, θ^+) . Similarly, as c is parametrized over the interior of its linearity interval, the primal optimal set is unchanged, and any basis within the maximal partition corresponds to a vertex of the primal optimal set. To be precise, let

(B, N) be any optimal partition for $\theta = 0$, for which there is a dual optimal solution (y, s) so that

$$A^T y + s = c, s \geq 0, s_B = 0.$$

Assuming that θ^- and θ^+ are nonzero, we have from Theorem 1 that if $\theta \in (\theta^-, \theta^+)$, then there is a primal solution x such that

$$Ax = b + \theta \hat{b}, x \geq 0, x^T s = 0.$$

In the case that (B, N) is basic, we say that the basis is *dual optimal* for $\theta \in (\theta^-, \theta^+)$. Similarly, if c is parametrized, then for any x such that

$$A_B x_B = b, x \geq 0, x_N = 0,$$

we know that there are y and s so that

$$A^T y + s = c + \lambda \hat{c}, s \geq 0, x^T s = 0,$$

provided that $\lambda \in (\lambda^-, \lambda^+)$. Again, if (B, N) is basic, we say that the basis is *primal optimal* for $\lambda \in (\lambda^-, \lambda^+)$.

The interplay between the basic and the maximal partitions is of special interest at a break point. As an illustration, consider the example for $\theta = 3$. There are two optimal basic partitions, which we denote by $(B', N') = (\{1, 2\}, \{3, 4\})$ and $(B'', N'') = (\{1, 4\}, \{2, 3\})$. The maximal partition is $(\hat{B}, \hat{N}) = (\{1, 3\}, \{2, 4\})$. In this case we have

$$\begin{aligned} 0 &= \min\{\theta : Ax = b + (3 + \theta)\hat{b}, x \geq 0, x_{\hat{N}} = 0\} \\ &= \max\{\theta : Ax = b + (3 + \theta)\hat{b}, x \geq 0, x_{\hat{N}} = 0\}. \end{aligned}$$

This shows the linearity interval for the perturbed right-hand side of $b + 3\hat{b}$ is the singleton $[0, 0] = \{0\}$, and we say that the maximal partition is *incompatible* with parametrizations away from $b + 3\hat{b}$ along $\hat{b}$. However, the basic optimal partitions give

$$\begin{aligned} -1 &= \min\{\theta : Ax = b + (3 + \theta)\hat{b}, x \geq 0, x_{N'} = 0\} \\ 0 &= \max\{\theta : Ax = b + (3 + \theta)\hat{b}, x \geq 0, x_{N'} = 0\} \end{aligned} \tag{7}$$

and

$$0 = \min\{\theta : Ax = b + (3 + \theta)\mathcal{B}, x \geq 0, x_{N''} = 0\}$$

$$\infty = \max\{\theta : Ax = b + (3 + \theta)\mathcal{B}, x \geq 0, x_{N''} = 0\}.$$

So (B', N') is *compatible* with $\mathcal{B}$ as θ decreases and (B'', N'') is *compatible* with $\mathcal{B}$ as θ increases. In the case of θ decreasing we note that (B', N') does not identify the entire linearity interval. If θ increases, then the last linear program being unbounded shows that the linearity interval is unbounded. From the example we see that different partitions are compatible with different directions of perturbation. A theory of compatibility is developed in Refs 12 and 16.

Combining Theorem 1 with the study of what occurs at a break point gives a complete analysis of z^* as either θ or λ traverse through all possible values for which the primal and dual are feasible, in which case the parameters could move through multiple linearity intervals. The question that remains is how to move from one linearity interval, through a break point, and to the adjoining interval. From the above example we see that we can always move into a linearity interval with a basis, even from a break point. However, identifying a compatible basis in a degenerate problem is not simple. From Theorem 1 we see that the maximal partition provides a definitive technique that rests on solving a linear program. Suppose the maximal partition $(\hat{B}, \hat{N})$ is incompatible with $\mathcal{B}$, a fact that would be known upon calculating θ^+ to be zero in the first statement of Theorem 1. The adjoining linearity interval is of the form $[0, \theta']$, where $\theta' > 0$. We let $(\hat{B}', \hat{N}')$ be the unique maximal partition for $\theta \in (0, \theta')$. To find $(\hat{B}', \hat{N}')$ we solve

$$\max\{\mathcal{B}^T y : A^T y + s = c, s \geq 0, s_{\hat{B}} = 0\}.$$

From Ref. 10 we have that

$$\hat{B}' = \{i : s_i = 0 \text{ for all optimal } (y, s)\} \text{ and}$$

$$\hat{N}' = \{1, 2, \dots, n\} \setminus B''. \quad (8)$$

Subsequently, we have from the fifth and sixth statements of Theorem 1 that the solution to this problem is the right-sided derivative of z^* at $\theta = 0$.

The change in the maximal partition denoted in Equation (8) leads to the following algorithm to calculate z^* . As with Equations (5) and (6) the linear programs can be stated in terms of the objective function for stability reasons, and it is this presentation that we use.

1. Calculate $z^*(0) = \min\{c^T x : Ax = b, x \geq 0\}$ and initialize θ^- to be zero.
2. Calculate $D^+ = \max\{\mathcal{B}^T y : Ay + s = c, s \geq 0, b^T y = z^*(\theta^-)\}$. If the problem is unbounded, then stop.
3. Calculate $\max\{\theta : Ax = b + \theta\mathcal{B}, x \geq 0, c^T x = z^*(\theta^-) + \theta D^+\}$ and let θ^+ be the optimal value (possibly infinity). Note that we always have $\theta^- < \theta^+$ since we are moving through a linearity interval.
4. Let $z^*(\theta) = z^*(\theta^-) + \theta D^+$ for either $\theta \in (\theta^-, \theta^+]$, if $\theta^+ < \infty$, or $\theta \in (\theta^-, \theta^+)$, if $\theta^+ = \infty$.
5. If $\theta^+ < \infty$, then let $\theta^- = \theta^+$ and return to step 3. Otherwise, stop.

We only state the algorithm for parametrizations in b . The algorithm for changes in c is analogous.

The study of parametrizations in one of b or c naturally continues with questions of simultaneous changes in b and c . In fact, the general question is to study the essence of optimality as the data (A, b, c) is parametrized along $(\mathcal{A}, \mathcal{B}, \mathcal{C})$. An expanded study of this topic is beyond the scope of this article, and we point readers to the bibliography, which includes works outside those cited in support of a continued study. One result from Ref. 12 is particularly germane to our discussion of characterizing z^* . Consider what happens if A remains constant and b and c change simultaneously. In this situation we are interested in

$$z^*(\theta) = \min\{(c + \theta\mathcal{C})^T x : Ax = b + \theta\mathcal{B}, x \geq 0\}.$$

Let (x^*, y^*, s^*) be a primal-dual solution for $\theta = 0$ such that $x_B^* > 0$ and $s_N^* > 0$. As long as $\mathcal{B}$ and $\mathcal{C}$ are compatible with $(\hat{B}, \hat{N})$, we have

that

$$z^*(\theta) = z(0) + \theta \left(\bar{x}_B x_B^* + \bar{y}_N^T y_N \right) + \theta^2 \left(\bar{x}_B A_B^+ \bar{y}_N \right),$$

where A_B^+ is any generalized inverse of A_B . The first-order term is the sum of the primal and dual derivatives for the case in which they are parametrized individually. The second-order term shows that the curvature of z^* is $\bar{x}_B A_B^+ \bar{y}_N$, which is invariant with respect to the choice of the generalized inverse.

Parametric linear programming has been used outside of typical query questions of the type “What if ...?” Parametric analysis on matching problems has been used in Ref. 17 to identify undesirable job assignments for the US Navy. It was also used in Ref. 18 to design a regional network for efficient and equitable liver allocation, and in Ref. 19 to prune radiotherapy treatments by restricting the search space. Each of these applications is based on the fact that the collection of Pareto optimal solutions to a bi-objective linear programming problem can be expressed parametrically. For example, consider the bi-objective linear program

$$\min \left\{ \begin{pmatrix} c^T x \\ d^T x \end{pmatrix} : Ax = b, x \geq 0 \right\}.$$

Owing to the convexity of the problem we know that each Pareto optimal solution is also a solution to

$$\min\{(c + \lambda(d - c))x : Ax = b, x \geq 0\},$$

for some, not necessarily unique, $\lambda \in (0, 1)$ [20]. The analysis above shows that we can calculate the linearity intervals of this problem and the rate at which the objective value changes over each linearity interval by solving a series of linear programs. If we let (B^k, N^k) be the finite sequence of maximal partitions at the break points in $(0, 1)$, then any variable whose index is not in

$$\bigcup_k B^k$$

is zero in every Pareto solution. This conveniently allows us to mathematically identify

the variables that can be removed without altering the Pareto set. This works only for the bi-objective case, which somewhat limits its applicability. A host of related applications is found in Ref. 21.

REFERENCES

1. Gass S, Saaty T. Parametric objective function (part 2)—generalization. *Oper Res* 1955;3(4):395–401.
2. Mills H. Marginal values of matrix games and linear programs. In: Kuhn H, Tucker A, editors. *Linear inequalities and related systems*. Princeton (NJ): Princeton University Press; 1956.
3. Saaty T, Gass S. Parametric objective function (part 1). *Oper Res* 1954;2(3):316–319.
4. Gal T. Postoptimal analysis: parametric programming and related topics: degeneracy, multicriteria decision-making, redundancy. 2nd ed. Berlin, FRG: Walter de Gruyter; 1995.
5. Orchard-Hays W. History of mathematical programming systems. *Ann Hist Comput* 1984;6(3):296–312.
6. Williams A. Marginal values in linear programming. *J Soc Ind Appl Math* 1963;11:82–94.
7. Goldman A, Tucker A. Theory of linear programming. In: Kuhn H, Tucker A, editors. *Volume 38, Linear inequalities and related systems*. Princeton (NJ): Princeton University Press; 1956. pp. 53–97.
8. Jansen B, de Jong JJ, Roos C, *et al.* Sensitivity analysis in linear programming: just be careful!. *Eur J Oper Res* 1997;101:15–28.
9. Holder A. Simultaneous data perturbations and analytic center convergence. *SIAM J Optim* 2004;14(3):841–868.
10. Adler I, Monteiro R. A geometric view of parametric linear programming. *Algorithmica* 1992;8:161–176.
11. Greenberg H. The use of the optimal partition in a linear programming solution for postoptimal analysis. *Oper Res Lett* 1994;15(4):179–185.
12. Greenberg H. Simultaneous primal-dual right-hand-side sensitivity analysis from a strictly complementary solution of a linear program. *SIAM J Optim* 2000;10:427–442.
13. Jansen B. Interior point techniques in optimization [PhD thesis]. Delft, The Netherlands: Delft University of Technology, Delft; 1995.

14. Monteiro R, Mehrotra S. A general parametric analysis approach and its implication to sensitivity analysis in interior point methods. *Math Program* 1996;72:65–82.
15. Roos C, Terlaky T, Vial J-Ph. *Theory and algorithms for linear optimization: an interior point approach*. New York: John Wiley & Sons, Inc.; 1997.
16. Greenberg H, Holder A, Roos C, *et al.* On the dimension of the set of rim perturbations for optimal partition invariance. *SIAM J Optim* 1998;9(1):207–216.
17. Holder A. Navy personnel planning and the optimal partition. *Oper Res* 2005;53:77–89.
18. Demirci M, Schaefer A, Romeijn H, *et al.* An exact method for balancing efficiency and equity in the liver allocation hierarchy. Technical report. Department of Industrial Engineering, University of Pittsburgh; 2010.
19. Holder A. Partitioning multiple objective optimal solutions with applications in radiotherapy design. *Optim Eng* 2006;7:501–526.
20. Ehrgott M. Volume 491, *Multicriteria optimization*, Lecture Notes in Economics and Mathematical Systems. New York: Springer; 2000.
21. Gal T, Greenberg HJ. *Advances in sensitivity analysis and parametric programming*. Norwell (MA): Kluwer Academic Publishers; 1997.

FURTHER READING

- Caron R, Greenberg H, Holder A. Analytic centers and repelling inequalities. *Eur J Oper Res* 2002;143:268–290.
- Ghadigheh AG, Terlaky T. Generalized support set invariancy sensitivity analysis in linear optimization. *J Ind Manage Optim* 2006;2:1–18.
- Ghaffari-Hadigheh A, Ghaffari-Hadigheh H, Terlaky T. Bi-parametric optimal partition invariancy sensitivity analysis in linear optimization. *Cent Eur J Oper Res* 2008;16:215–238.
- Greenberg H. An analysis of degeneracy. *Nav Log Res Q* 1986;33:635–655.
- Greenberg H. Matrix sensitivity analysis from an interior solution of a linear program. *INFORMS J Comput* 1999;11(3):316–327.
- Greenberg H. *A computer-assisted analysis system for mathematical programming models and solutions: a user's guide for analyze*. Boston (MA): Kluwer Academic Publishers; 1993.
- Hadigheh AG, Mirnia K, Terlaky T. Active constraint set invariancy sensitivity analysis in linear optimization. *J Optim Theory Appl* 2007;133:303–315.
- Hadigheh AG, Terlaky T. Sensitivity analysis in linear optimization: invariant support set intervals. *Eur J Oper Res* 2006;169:1158–1175.

PARTIALLY OBSERVABLE MDPS (POMDPS): INTRODUCTION AND EXAMPLES

EMINE YAYLALI

JULIE S. IVY

Edward P. Fitts Department of Industrial
and Systems Engineering, North Carolina
State University, Raleigh, North Carolina

A Markov decision process (MDP) is an optimization model used for sequential decision making in a stochastic environment. The theory, applications, and solution techniques for MDPs have been investigated in several papers and books [1,2]. The objective of an MDP model is often to determine a set of decision rules called *policy* for selecting an appropriate *action* in every *decision epoch*. A key assumption of Markov models is that the transition probability from the current state of the system to the state of the system at next decision epoch depends only on the current state, and not on the prior states or actions of the system. Any stochastic dynamic system that satisfies this Markovian property becomes a controlled Markov process when the transition probabilities can be affected by the decision maker's actions. MDPs and partially observable Markov decision processes (POMDPs) are the examples of the controlled Markov processes. In the MDP framework, the state of the system is assumed to be completely observable by decision maker with the help of perfect observations (and hence known with certainty by the decision maker). Relaxation of this assumption yields a POMDP or a hidden Markov chain (HMC). In the case of HMCs, the underlying stochastic system is hidden and assumed to be Markov based on observations that signal the system's hidden states. Similarly, the underlying state of the POMDP is determined through a sequence of observations. However, in a POMDP, the decision maker can affect the transition of the system through the selection of actions. This is the key

difference between a POMDP and an HMC. Table 1, which was proposed by Littman, best illustrates these distinctions [3].

A POMDP is a generalization of MDPs in which there is incomplete information and uncertainty regarding the current state of the Markov process. Incomplete information regarding the state of process affects the optimal choice of actions. Physical constraints, noise-corrupted readings in the observation of state, or simply being too costly to observe state can result in uncertainty. Some examples for which the uncertainty of the state is a realistic key assumption include machine maintenance, learning theory, artificial intelligence, and health care. Shiryaev [4] initially formulated a detection problem with random process whose parameters were unknown and explored applying a Markov process as a special case of this problem. Drake, Astrom and Sondik later defined explicit formulations for a POMDP [5–7]. Since then, the theory of POMDPs including solution techniques and structural characteristics has been an active research topic.

The generalization of an MDP to a POMDP increases computational complexity by adding dimensionality due to the uncertainty of states (depending on the definition of the MDP state space this uncertainty can add several dimensions to the state space). In comparison to MDPs, POMDPs have an infinite state space. Furthermore, an optimal policy for a POMDP is defined over a continuous state space. These features of a POMDP increase the computational effort and memory required to solve the model optimally. Typical exact solution procedures for a POMDP rely on dynamic programming approaches and often are insufficient for solving real-world problems because of the computational burden (i.e., the curse of dimensionality). Research on efficient algorithms to solve POMDPs varies from exact algorithms [7–12] to approximation methods. Finite-memory, finite-grid approximations, and heuristics for developing feasible bounds are examples of approximate solution

Table 1. Relationship of Markov Models

Markov Models		Do We Have Control over State Transitions?	
		No	Yes
Are the states completely observable?	Yes	Markov chain	MDP
	No	Hidden Markov chain	POMDP

techniques for POMDPs [13,14]. For detailed information, see Lovejoy’s and White’s surveys on solution techniques [15,16].

The remainder of this article is organized as follows: the section titled “Partially Observable Markov Decision Processes” presents a general formulation for a POMDP and the section titled “Applications of POMDP” summarizes several applications of POMDPs.

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

A POMDP is completely characterized by $\langle S, A, Z, P, Q, R \rangle$: the core state space, the action space, the set of observations, the transition probability matrix, the information matrix, and the reward function. Each of the elements of a POMDP model is discussed in detail the following sections.

Core States

Let $\{s(t), t = 0, 1, \dots\}$ be the *core state* of the system at *decision epoch*, t . For simplicity in presentation, $s(t)$ is assumed to be a finite state discrete-time Markov chain with stationary transition probability matrix $P = [p_{ij}]$, $i, j \in S$ where S is called *core state space*. The transition probability matrix P completely defines the core process; however, decision maker cannot directly observe this core process at time t . Let T be the number of decision epochs, that is, planning horizon. If $T < \infty$, POMDP model is a finite horizon model, otherwise it is an infinite horizon POMDP.

Observations

To learn about the true value of state at time t , *observations* $\{z(t), t = 1, 2, \dots\}$ are collected.

The observation space, Z , is assumed to be finite just as the core state space S . The relationship between the core state and the observations is known to the decision maker and defined by the *information matrix* $Q = [q_{ik}]$, $i \in S, k \in Z$ where q_{ik} is the probability of observing $k \in Z$ given that the underlying state is $i \in S$.

Actions

The decision maker’s control over the system is provided by actions. At each decision epoch, t , the decision maker chooses an *action* $\{a(t), t = 0, 1, \dots\}$ from the set of available actions from the action space A where it is assumed to be finite. The state transition matrix P and the information matrix Q could be functions of the action $a(t)$. For the most general case, both matrices are assumed to be affected by the choice of action. Specifically, $P(a) = [p_{ij}(a)]$ is the probability that the system moves from core state $i \in S$ to core state $j \in S$ when action $a \in A$ is chosen. Similarly, $Q(a) = [q_{ik}(a)]$ is the probability of observing $k \in Z$ while the system’s actual state is $i \in S$ and action $a \in A$ is chosen.

Data Process and Belief States

The *data process* $\{d(t), t = 0, 1, \dots\}$ is the collection of all past and present observations and actions taken prior to time t including the initial distribution of core states, $\pi(0)$, that is, $d(t) = (\pi(0), z(t), \dots, z(1), a(t-1), \dots, a(0))$ for $t \geq 1$, where $\pi(0) = (\pi_i(0), i \in S)$ and $\pi_i(0) = \Pr\{s(0) = i\}$. The core states are only partially observable by the decision maker. Given the data process available to the decision maker, the belief state, $\pi_i(t)$, is defined as the probability of being in the core state, $i \in S$, that is, $\pi_i(t) \equiv \Pr\{s(t) =$

$i|d(t)\}$. The set of belief states, $\Pi(S)$ defined over core state space S is as follows: $\Pi(S) \equiv \left\{ \pi \in \mathbb{R}^N : \sum_{i=1}^N \pi_i = 1, \pi_i \geq 0, i = 1, \dots, N \right\}$, where N is the cardinality of core state space S .

Updating Belief States

The system transitions among belief states. An important element of this process is the timing of the state transitions. According to this timing, we present two models that differ with respect to the update function for the belief state. Although there is a slight difference in formulation, the models have been shown to be equivalent [17,18]. In the first model, MI, given the core state $i \in S$ and the chosen action $a \in A$, a transition occurs with probability $p_{ij}(a)$, and then given the transition from core state $i \in S$ to $j \in S, k \in Z$ is observed with probability q_{jk} . In the second model, MII, if the current state is $i \in S$ and action $a \in A$ is chosen, $k \in Z$ is observed with probability q_{ik} , then a transition to state $j \in S$ occurs with probability $p_{ij}(a)$. The sequence of the process in MI is action, then the state transition, and last, the observation, whereas in MII, the sequence is action first, then observation, and finally the transition of the state. Figures 1 and 2 show the schematics of the

decision process for MI and MII, respectively [7,17].

We can update our belief state for each model using Bayes' formula as follows:

$$\pi_j(t+1) = \frac{\sum_{i \in S} q_{jk}(a) p_{ij}(a) \pi_i(t)}{\sum_{i \in S} \sum_{l \in S} q_{lk}(a) p_{il}(a) \pi_i(t)} \text{ for MI} \quad (1)$$

and

$$\pi_j(t+1) = \frac{\sum_{i \in S} q_{ik}(a) p_{ij}(a) \pi_i(t)}{\sum_{i \in S} q_{ik}(a) \pi_i(t)} \text{ for MII.} \quad (2)$$

The belief state $\pi(t)$ is a sufficient statistic and contains all of the information required to update the system and for making a decision at time t [7,19].

Reward Function

To construct a POMDP model, we define an immediate reward function. The decision maker receives an immediate reward $r(i, a) \in R$ when core state is $i \in S$ and action $a \in A$ is taken. The POMDP model, $\langle S, A, Z, P, Q, R \rangle$, determines the decision rule, or policy, that yields the maximum

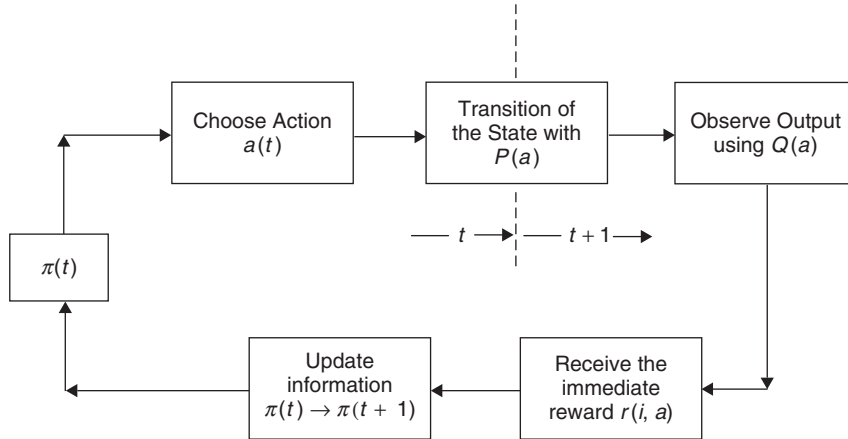


Figure 1. Partially observable Markov decision process for model 1, that is, assuming that the state transition occurs prior to the observation.

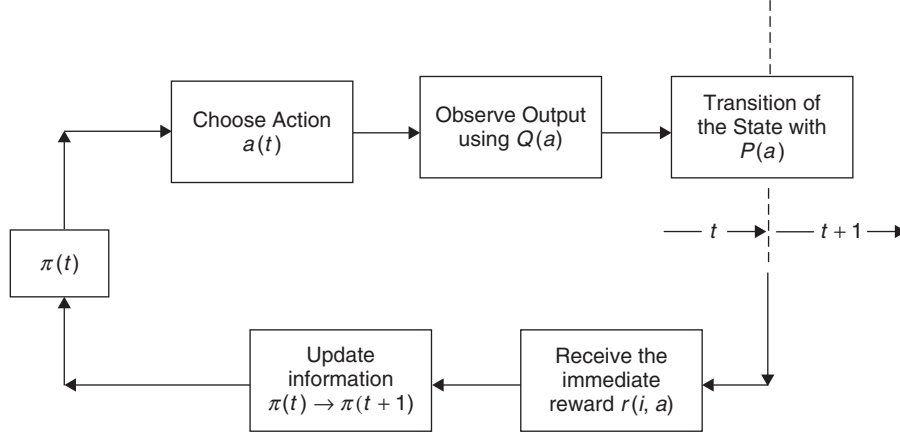


Figure 2. Partially observable Markov decision process for model 2, that is, assuming that the state transition occurs after the observation.

total expected reward. When determining the maximum total expected reward, a discount factor $\lambda \in [0, 1]$ can be introduced. Because the POMDP model is defined over a planning horizon, a discounted reward function is often a reasonable assumption for applications where the reward or cost accumulates over a long horizon.

Policy and Strategy

A *strategy* is defined as a sequence of policies, $\delta = \{\delta_1, \delta_2, \dots, \delta_t, \dots\}$, such that δ_t is a *policy* that determines which action in A to select at time t of the process given $d(t)$, the data available at time t . Because the process is Markovian, it is not necessary to use all of the information in the data process, $d(t)$, and the current state of the system is sufficient information for action selection. Thus, the policy can be redefined as a function of the state, which is $\delta_t : \Pi(S) \rightarrow A$. A strategy is stationary if the policy does not change from one decision epoch to another, given that the state of the process, $\pi(t)$, stays the same.

Optimal Value Function

Using the above notation, we define an optimal value function, $V_\lambda(\pi)$, to determine the policy that maximizes the total expected discounted reward. For infinite horizon problem, optimal policy can be determined by solving

the following recursion:

$$V_\lambda(\pi) = \max_{a \in A} \left\{ \sum_{i \in S} \pi_i r(i, a) + \lambda \sum_{k \in Z} V_\lambda(\pi(k, a)) \sigma(k|\pi, a) \right\} \text{ for } \pi \in \Pi(S), \quad (3)$$

where $r(i, a) \in R$ is the immediate reward function, $V_\lambda(\pi)$ is the optimal value function for belief state $\pi \in \Pi(S)$, and $\lambda \in [0, 1]$ is the discount factor [4, 12]. The probability of observing $k \in Z$, given the state $\pi(t)$ and the selected action $a \in A$, is $\sigma(k|\pi(t), a(t))$ (which is also the denominator of Equation (1)) and it is computed by

$$\begin{aligned} \Pr\{k|\pi(t), a(t)\} &\equiv \sigma(k|\pi(t), a(t)) \\ &= \sum_{j \in S} \sum_{i \in S} q_{jk} p_{ij} \pi_i(t). \end{aligned}$$

For the finite horizon problem, where $T < \infty$, the analogous recursive equation is defined as follows:

$$\begin{aligned} V_\lambda^{T+1}(\pi) &= \max_{a \in A} \left\{ \sum_{i \in S} \pi_i r^{T+1}(i, a) \right\} \\ &\text{for } \pi \in \Pi(S), \end{aligned}$$

$$V_{\lambda}^t(\pi) = \max_{a \in A} \left\{ \sum_{i \in S} \pi_i r(i, a) + \lambda \sum_{k \in Z} V_{\lambda}^{t+1}(\pi(t+1)) \sigma(k|\pi, a) \right\} \text{ for } \pi \in \Pi(S), t = 1, 2, \dots, T, \quad (4)$$

where $r^{T+1}(i, a)$ is the terminal reward received when the core state is $i \in S$ and $\lambda \in [0, 1]$ is the discount factor. For $t = 1, 2, \dots, T$, $V_{\lambda}^t(\pi)$ is the optimal value function if the system is in state $\pi \in \Pi(S)$ and there are $T - t$ periods remaining before the process ends.

Equations (3) and (4) determine the action $a^*(\pi) \in A$ that maximizes the total discounted reward in that decision epoch. The collection of these actions provides the optimal policy $\delta^* : \Pi(S) \rightarrow A$ for the process. The following simple example illustrates the application of a POMDP model to a classic machine maintenance problem.

Example

Machine maintenance is one of the earliest applications of POMDP. In this model, a machine or a system, which is observed periodically, deteriorates over time. Since deterioration involves internal components of the machine/system, the true condition of the machine/system is unknown to the decision maker. It may be possible to determine the true condition of the machine by disassembling the machine or through some other type of destructive testing, however, this procedure can be costly or unproductive. Consider a simple machine that is operating in one of the two conditions, “good” (0) or “bad” (1), $S \in \{0, 1\}$. The “bad” state corresponds to a machine that is operational but producing defective or unusable. This failure may only be identified by destructive testing; hence, the condition of the machine is partially observed. Further, observations may be based on a sample of products from a lot making the observations imperfect. There are two belief states: $\pi_t \equiv \text{Probability}\{S_t = 0\}$ and $1 - \pi_t \equiv \text{Probability}\{S_t = 1\}$. The machine starts in state 0 and enters state 1 after operating for time I , where I is a discrete random variable with a probability mass

function $f_I(i) = p_{00}(1 - p_{00})^{i-1}, i = 1, 2, \dots$. Once the machine enters 1 it stays there until exogenous action is taken. Monitoring gathers independent observations, Z_t , at time $t, t = 1, 2, \dots$, where the density of Z_t depends on the machine state with pdf, q_{0k} , if $S_t = 0$, or q_{1k} , if $S_t = 1$.

At any decision epoch, t , there are two possible actions: “do nothing” action $a_0(t)$, which allows the machine to continue to operate, and the “checking” action $a_1(t)$, which stops the machine and obtains perfect information about its condition. If the machine is found to be in state 0, it is a false alarm; if the machine is found to be in state 1, the machine is replaced and restarts in state 0. Assume that there is a fixed cost, $r(i, a)$, associated with performing action, a_1 , when the machine is in state i , where $r(1, a_1) < r(0, a_1)$. Assume that the cost per unit time for operating the machine in state i is $r(i, a_0)$, where $r(0, a_0) < r(1, a_0)$. The cost $r(1, a_0)$ serves as a penalty for false alarms while the cost $r(0, a_1)$ is the penalty per unit time associated with failing to detect.

When action $a \in \{a_0, a_1\}$ is taken, define the immediate expected cost accrued associated with each action as $r(1, a_0)(1 - \pi)$ and $r(0, a_1)\pi + r(1, a_1)(1 - \pi)$, respectively. We assume without loss of generality that $r(0, a_0) = 0$. Then, we define the minimum total expected cost with $T - t$ periods remaining, and probability π that the system is in state 0 as

$$V_{\lambda}^t(\pi) = \min_{i \in \{0, 1\}} \left\{ \begin{aligned} & r(1, a_0)(1 - \pi) + E_k[V_{\lambda}^{t+1}(\pi_{a_0}^k(t+1))] \\ & r(0, a_1)\pi + r(1, a_1)(1 - \pi) \\ & + E_k[V_{\lambda}^{t+1}(\pi_{a_1}^k(t+1))] \end{aligned} \right\},$$

where $E_k[V_{\lambda}^{t+1}(\pi(t+1)|k, a_i)] \equiv$ expectation of the minimum total cost with respect to the next observation k ; and $\pi_{a_i}^k(t+1) \equiv$ probability the state is 0 given that observation k and action a_i are selected, $i = 0, 1$.

Using Bayes’ rule, one gets

$$\pi_{a_0}^k(t+1) = \frac{(1 - p_{00})\pi q_{0k}(a_0)}{\pi q_{0k}(a_0) + (1 - \pi)q_{1k}(a_0)}$$

and because a_1 is the *renew* action, $\pi_{a_1}^k(t+1) = 1$. We assume that the value of the machine at the end of the horizon, when

Table 2. Applications of POMDP

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
9 Girshik and Rubin [23], Eckles [24], Ross [25], Smallwood and Sondik [8], Rosenfield [26], Pierskalla and Voelker [27], Wang [28,29], White [30–33], Hopp and Wu [34], Hopp and Kuo [35], Ivy and Pollock [36], Maillart [37], Kuo [38], Ghasemi <i>et al.</i> [39], Yeung <i>et al.</i> [40]	A machine or a system which is observed periodically deteriorates over time. Determination of a maintenance/replacement/inspection policy at every period that satisfies objectives of the model. Observations about machine can be used as a basis for maintenance or replacement decisions. Two types of observation can exist: outputs of the machine or outcomes of the system and inspection of the machine/system	Objective function of the model can vary. Some examples of different objective types that are applied in some of machine maintenance applications are to minimize total cost (operating cost), to maximize the production capacity, to maximize the expected length of time between replacements, and to maximize total reward	States of POMDP model are defined as the state of the machine. Maintenance models can be classified depending on the number of states. Two-state models assume the state of the machine is either “good” or “bad”. Multistate models generalize two-state models by allowing the machine to have N -level of deterioration	Since deterioration includes internal parts of the machine, the true condition of the machine is unknown to the decision maker. It is possible to disassemble the machine in order to determine the true state; however, this procedure can be costly or unproductive	Available actions for maintenance models can be any set of the following: do nothing, perform maintenance, inspect the machine, replace components, and replace the machine. (Different types of inspection can be included as well.)	Machine maintenance is one of the earliest and widely explored applications of POMDP. Not only are various types of maintenance models formulated as POMDP, for example, whether perfect information is available after inspection action, but also many researchers have developed structural properties, characterized the optimal policy, and produced conditions for control-limit type policies
Karush and Dear [41]	A subject is taught n items in the course of N trials. The model is developed for the determination of a strategy of trial-by-trial item selection	To maximize the expected terminal level of achievement of the subject	States are the knowledge of subject on any given item	It is not possible to observe knowledge fully. Responses of the subjects on the test of given items are observations for the underlying state	A reinforcement or teaching action related to item	Application of POMDP for model development in education and teaching strategies

	Kaplan [42]	Assessment of periodic financial reports of divisions in a large decentralized corporation. Reports are used as observations regarding whether or not the division is out of control	The minimum expected discounted cost given a policy is chosen.	Finite two-state system where the state is 1 if the division is operating in control or 2 if the reduction of incurred cost in the division is possible with the actions of management	Observing financial reports does not give complete information about the state of the system, that is, whether division is in control or out of control	Do nothing and investigate the division incurring a cost	Adapted a model that was developed for industrial quality control
	Pollock [43]	When a target moves between two regions in Markovian fashion, the model allocates the search effort to one region in order to locate the moving target	To minimize the expected number of observations or to maximize the probability of locating the target	Two-state model where state 1 means the target is in region 1 and state 2 means the target is in region 2. An extended model including state 3 which is the state the target moves to when it is detected	The decision maker does not exactly know the location of the moving target	To look in region 1, to look in region 2	Knowing the location of the moving target could be useful for military applications, such as locating an opponent's submarines and mobile missile platform. Technological advances like satellite images and sonar could be used for observations in the system that increases the chance of locating the target

(continued overleaf)

Table 2. (Continued)

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
Smallwood [44]	Optimization of the instructional decision process based on learning of students	To minimize the total expected cost of future instruction for each of the available alternative teaching strategies	State is defined as student's internal state of knowledge	Similar to Karush and Dear model, knowledge cannot be assessed entirely via student's response to the test	For an item, possible actions are to use present instructional alternative, to use "test" instructional alternative, and to terminate the instruction	Built a model to use for day-to-day operations of an educational system, developed an algorithm to solve the model, proposed an extension incorporating student's past history in the model
8 Smallwood <i>et al.</i> , [45]	Identifying patient state and physician state of information as two interrelated concepts, development of a decision-making methodology related to problems in health-care systems. Problems can be individual, such as medical diagnosis and treatment and individual facility design; or societal, such as health service programs and regional health system design	For an immunization program, the objective function can be to improve the health state of patients for a certain disease. Depending on the type of the problem, the objective function of the model differs	Patient state is defined as the internal physiological or psychological state of the patient, which is usually unobservable. Physician state of information corresponds to the physician belief regarding the patient state and depends on the physician, his prior experience, and his knowledge of the patient. These two types of states are paired for formulating the problems	The internal states of the patient are not directly observable. The physician will have data available (diagnostic tests etc.) which influence his state of information about the patient's state. Natural change in the patient's state or change in the application of the physician's action also affect the form of the state pairs.	The available actions vary in different models. In medical diagnosis and treatment problems, the actions are medical ones in response to a need. In health system design problem, the actions are policy decisions such as how much and where to spend funds	One of the earliest health-care-related paper that proposes the utilization of POMDP while defining patient health as unobserved state pairing with physician state of information. It explores a broad range of health-care system problems as promising areas to integrate twin states

	White [46]	Determination of the optimal sequence of questions, or a questionnaire in order to identify the state of the system while allowing for answers that are not necessarily truth	To minimize the expected terminal cost	Any system with an underlying process which can be identified with a set of diagnostic questions. If the system has finite number of states and it doesn't change state during questions, then the given model can be used to diagnose the state optimally.	As a prior assumption, the state of the system is unobservable and the questionnaire is used to diagnose the underlying state.	The sequence of questions to be asked.	A POMDP application on questionnaire design. Model is slightly different version of the POMDP where transition of the states during questionnaire is not allowed. Sufficient conditions for an optimal questionnaire to obtain a lower limit are also determined.
6	Segall [47]	While only one copy of a file is allowed to exist in a computer network, the goal is to determine the optimal policies for dynamic allocation of files in the computer network that works under time-varying operating conditions	To minimize total expected cost where the cost function includes storage cost of a file in a computer and transmission cost of a file from one computer to another in the network	Demand rate of files is assumed to be the state of the system	Two models are developed in the article. One assumes a known demand rate, whereas the second model is constructed with unknown demand rates since it is usually difficult to know demand rates perfectly as a function of time	While the file is stored in a computer i at time t , the available actions are to leave the file in computer i , to transfer the file to computer j which requests that file at time t	A POMDP application on network design

(continued overleaf)

Table 2. (Continued)

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
Hughes [48]	Construction of a general decision model appropriate for the determination of the optimal timing of internal audits. Audit is one of the actions that provides information about the internal control process of a firm	The minimum total expected discounted costs	States of the model are the effectiveness of internal control process of the firm	Auditors cannot observe the state of internal control process. However, they have a belief about the state	Do nothing, audit, and restore. It is also mentioned that as long as action space is finite, any control action related to internal control process can be added as action into the model	Model is similar (or adapted from) the Ross model of quality control under Markovian deterioration
Lane [49]	Determining dynamic decision policies for the intraseasonal decisions of individual fishing vessel operators. When and where to fish throughout the season with the highest return on seasonal income for independent fishermen will be decided according to the optimal policy.	To maximize net operating income	The state space is assumed to be fish abundance; in a particular model, it is salmon abundance.	The fish stock abundance is not directly observable by the decision maker. However, the average fish catch by individual fishermen provides an imperfect measure of stock abundance.	The action space is assumed to be interzonal movements. In other words, action in any period describes the movement from current zone to another in the fishery, including no fishing option.	A real-life application of POMDP that may be used as a decision tool in the regulation of fisheries.

Koenig and Simmons [50,51]	An autonomous robot requires having a reliable navigation system to complete its function. Authors developed a navigation architecture on the basis of POMDP that robustly tracks a robot's location indoors and directs its goal-oriented actions	To minimize the expected total cost while in every decision epoch an immediate cost is measured as the expected travel times needed to complete the actions	The pose of the robot is the state of the model. The orientation and the position of the robot are defined as the pose. The states are classified according to environmental parameters such as doors and walls	Uncertainty in actuator, sensor noise, incomplete knowledge of the environment, and uncertainty in the initial pose of the robot make the state unobservable. Thus, the robot never actually knows its exact location but has some belief regarding its location from sensor data	The actions are “the motion reports” for pose estimation and “the directives” for policy generation	POMDP application on robot navigation. It is proved that this architecture leads robust indoor navigation for an actual robot
Crites and Barto [52]	Elevator dispatching problem explores good policies for the control of elevator systems while improving performance. The article provides an application of reinforcement learning to the elevator dispatching problem which is known to be DP intractable	Possible objectives are minimizing the average wait time or the average system time, minimizing the percentage of passengers that are waiting longer than a dissatisfaction threshold time, and minimizing sum of squared wait times	Positions and directions of the elevator cars as well as the desired destinations of the passengers waiting in each floor are established as the state of the model	Because the system has varying passenger arrival rates and the call buttons on every floor, it does not provide enough information about the number of the passengers waiting and their desired destinations, and the state of the system is unobservable	Action space consists of some possible primitive actions for each elevator car: if it is stopped at a floor, it must either “move up” or “move down”, and if it is in motion between floors, it must either “stop at the next floor” or “continue past the next stop”	A real-life application of reinforcement learning (RL) for elevator dispatching. Since an optimal policy is unknown, the performance of RL-based algorithm is compared to some other elevator dispatching heuristics and it is shown to perform well in comparison heuristics. This demonstrates utility of RL in large-scale dynamic optimization problems

(continued overleaf)

Table 2. (Continued)

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
Darrell and Pentland [53]	An active gesture recognition (AGR) is designed based on POMDP in which an active camera is guided to foveate (focus on) for discriminating a certain gesture of the target.	To maximize the expected future reward where reward function consists of a unit positive reward when system detects the target and a unit negative reward when system fails to recognize target.	The states of the model are the states of the targeted person such as the targets' pose, facial expression, and hand configurations.	Since a foveated image sensor can only focus on a single body part not the whole body, the target's gesture cannot be recognized fully resulting with partial observability of the surroundings.	The set of actions are four foveation commands: look-body, look-head, look-right-hand, and look-left-hand. Each action lets the active camera to track a certain body part.	A POMDP application on machine vision. Q-learning algorithm adapted for gesture recognition of a foveated vision system.
Hu <i>et al.</i> [54]	Efficient dosage policies in drug therapy should maintain drug concentration at a target. Presentation of an infinite horizon POMDP model for a drug infusion problem which determines how to choose the infusion rate and how to administer the most effective dosage pattern	To minimize the total expected discounted cost where the cost function depends on the dosage and penalizes deviations from the target concentration	State space corresponds to the volume, clearance, and current drug concentration	Drug concentration is measured through observations that are prone to errors. Because exact drug concentration could not be observed, the model is partially observable	Action space contains nonnegative infusion rates	A POMDP model is developed to provide policies for adaptive control of drug concentration. However, the model is intractable and authors introduce approximation techniques to generate suboptimal candidate policies

Hauskrecht [55], Hauskrecht and Fraser [56,57]	Development of a POMDP framework for medical therapy planning for patients with ischemic heart disease (IHD)	To minimize the expected cumulative cost of the treatment, where the cost is defined in terms of dead–alive trade-off, quality of life, invasiveness of procedures and their economic cost	The state of the model is defined as the state of the patient or disease. If the patient’s status is alive, the state is described with a set of variables and their values. Some examples of these variables are coronary artery disease (mild, severe etc.), chest pain (no pain, mild etc.), stress test result (positive, negative), ischemia level (no ischemia, mild, severe), and so on	The diagnosis of a disease and its treatment are often practised without knowing the underlying patient state. The underlying disease state is observed indirectly using incomplete or imperfect observations	Actions are categorized as treatment and investigative actions. Treatment actions are “wait”, “medication”, “angioplasty,” and “coronary artery bypass graft surgery,” and in general, these actions result in an active change in the patient state to a more appropriate one. Investigative actions are “stress test” and “coronary angiogram,” and they may not only reveal something about the hidden disease state but also yield a change in the state	POMDP is suggested as an elegant framework for modeling medical therapy since it captures both uncertainty related to the control process (outcomes of treatment and diagnostic actions are not predictable) and partial observability of the disease process. The application of POMDP to a medical therapy planning problem is demonstrated with the IHD model
---	--	--	--	---	--	--

(continued overleaf)

Table 2. (Continued)

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
Murphy [58]	When an autonomous robot navigates in a dynamic environment, it must handle changes in the environment, such as closing doors and moving objects. A model for the map learning of a robot in such an environment is presented	A simple policy is developed for the model instead of solving the model for optimal policy. The policy is to visit the nearest uncertain cell. After successfully learning the map of an environment, the robot keeps running to update its map	The location of the robot corresponds to the state of the model	Similar to the Koenig and Simmons model, the robot's understanding of its location is incomplete and the sensors provide observations about its possible location	The actions are moving one step ahead in each of the four compass directions	Incorporation of dynamic environment into the POMDP models of robot navigation
Roy <i>et al.</i> [59]	A POMDP dialog manager is generated for recovering from noisy or ambiguous utterances in speech recognition systems	To maximize the reward obtained for fulfilling a user's request. Choosing the correct action leads to a positive reward,; on the other hand, incorrect actions are penalized by an equivalent negative amount	The states are the intentions of the user with respect to dialog task. A nursing home assistant robot is used to conduct experiments. In this example, the states of the model are determined as the tasks that are relevant for an assisted living in nursing home such as "bring medication" and "want TV information"	Human speech can be noisy and ambiguous. The dialog manager does not have direct knowledge regarding what the user wants, that is, the state of the system. It uses noisy speech utterances from the user to understand the task	Robot's responses to the dialog are the actions of the model. In the nursing home robot example, these actions are determined as follows: 10 actions correspond to the different abilities of the robot, such as going to kitchen, or providing the time, the remaining 10 actions are clarification and confirmation actions such as reconfirming the desired TV channel	A POMDP-based spoken dialog is developed and tested to handle noisy speeches. It is shown that under the same conditions, the POMDP dialog manager outperforms MDP dialog manager

Regan <i>et al.</i> [60]	A POMDP model (advisor-POMDP) that examines trust through reputation in electronic markets is proposed. This decision-theoretic framework describes the problem with the help of buying and selling agents. Buyers make purchases depending on advisors' information regarding seller's reputation	To maximize the expected reward, where the reward function has two components: a cost associated with asking an advisor about a selling agent and a reward which can be positive or negative depending on satisfaction from the purchase	A state of the advisor-POMDP consists of a vector of real values between 0 and 1 representing the reputation of each seller and a scalar value representing satisfaction resulting from a purchase	States of the model include the reputation of each selling agent. However, the buyer will have limited information about each seller's reputation since it only has some knowledge about a seller's past interactions	Actions that are designed for buying agents are asking an advisor for information about a selling agent or buying from a selling agent	The advisor-POMDP models the uncertainty inherent in social reputation systems for electronic markets. The framework allows buyers to make informed purchases using the information about a seller's reputation
Boger <i>et al.</i> [61]	Task assistance for people with dementia provides guidance when problems arise with completing activities of daily living (ADL) such as handwashing and dressing. MDP and POMDP models are developed for a specific ADL, that is, handwashing, and it is stated that the model is generalizable for almost any ADL assistance	To maximize the expected discounted sum of rewards. The reward function is designed to cover conflicting objectives such as maximizing the odds of task completion, minimizing caregiver intervention, and minimizing user frustration	Four classes of variables construct the state space of the model: those that capture the state of environment (hand location, water on/off), those that summarize ADL plan steps completed thus far (most recent step, maximum plan steps completed, current step that is repeated, etc.), those summarizing system behavior	Sensor noise causes environment variables and task steps to be estimated imperfectly	The model has 20 actions. Eighteen actions comprise prompts for six different plan tasks (water on/off, use soap, wet hands, rinse hands, dry hands) at three levels of specificity (general, moderate, and specific). The other two actions are "null" action and "call caregiver"	POMDP is offered as an ideal model for the decision process for ADL assistance since it captures stochasticity, noise, conflicting objectives, and the ability to customize behavior to specific individuals. The POMDP model for handwashing is solved approximately to develop optimal policies for ADL prompting

(continued overleaf)

Table 2. (Continued)

References	Short Summary of Problem	Objective Function	Definition of States	Why It Is Partially Observable	Action Space	Comments
		(number of prompts issued for the current plan step, type of last prompt, etc.), and those reflecting certain hidden aspects of the user's personality or mental state (e.g., general responsiveness, low or high)				
16 Williams and Young [62]	POMDP is proposed as a single framework (SDS-POMDP) which can unify and extend existing methods for improving the robustness of spoken dialog systems	The reward function in the framework is not specified explicitly, since it depends on design objectives of the target system. Some objective function examples given are maximizing the chance of successful closure and minimizing the number of turns required	A factored state-space representation is presented where the unobserved state of the SDS-POMDP can be factored into three distinct components: the user's goal, the user's action, and the dialog history	One reason why spoken dialog systems are challenging is that the estimate of a user's action is noisy, that is, the user's goal derived from its speech contains recognition errors. As a result, the state of the model is not directly observable and can never be known with certainty	The machine's actions are cast as the action space in the factored model	The SDS-POMDP framework is presented as a well-suited model in order to cope with the uncertainty present in spoken dialog systems and ambiguity of speech patterns. Practical advantages of utilizing the framework are also demonstrated

Maillart <i>et al.</i> [63]	Development of a partially observed Markov chain model of breast cancer progression in order to assess screening policies. The aim is to investigate the relative value and frequency of mammography screening for premenopausal women versus postmenopausal women	Although the model is not solved to optimality, the value of a screening policy is measured by lifetime mortality risk, that is, how likely a patient is to die from breast cancer over her lifetime under a certain screening policy	The states of the model are no breast cancer (state 0), early breast cancer (state 1), later/advanced breast cancer (state 2), breast-cancer-induced death (state 3) and non-breast-cancer-induced death (state 4)	Because transition between states 0, 1, and 2 is not visible without screening or diagnostic tests and mammograms may produce false positive and false negative results, the disease state is partially observable	Different screening policies are the actions of the model since they affect the underlying core state	Natural history of breast cancer is formulated as a partially observed Markov chain to generate a frontier of efficient screening policies by evaluating them according to lifetime mortality risk and expected mammogram count.
--------------------------------	--	---	--	--	---	--

there are no observations remaining and the process terminates, is zero regardless of the state, that is, $V_{\lambda}^0(\pi) = 0 \forall \pi$. $V_{\lambda}^T(\pi)$ is the optimal or minimal cost for a T -period finite horizon. Shiryaev proved that the fixed probability threshold rule for this type of two-state model with geometric failure times minimizes the average cost per unit time. This rule states that if π , the posterior probability of the system operating in the “good” condition, is less than or equal to some threshold value, then a “checking” action should be taken. Refer to Ivy and Nembhard [20] for sample solutions for the two-state machine-problem probability thresholds. Structural results such as this one can provide information about the form of the optimal policy and typically accelerate the solution procedure. However, owing to the uncertainty of the initial state and imperfect information regarding the underlying state, structural properties that are often exploited in MDPs cannot be determined for POMDPs [17]. In the following section, we present several applications of POMDPs.

APPLICATIONS OF POMDP

A POMDP is recognized as an appropriate technique for modeling problems with incomplete information and stochastic nature. While numerous articles provide theoretical contributions for the analysis and characterization of POMDPs, there is significant research that applies POMDPs to practical problems. Several problems in a wide variety of areas with these features can be formulated as POMDPs. Monahan and Cassandra’s surveys provide summaries of the literature for several of these applications [17,21]. In addition, the recent tutorial by Littman presents more examples and suggests that the POMDP may be an effective tool for behavioral science problems [22]. Table 2 presents a summary of several significant application domains for POMDPs.

SUMMARY AND CONCLUSION

In this article partially observable Markov decision processes are introduced and

reviewed and the relationship between POMDPs and MDPs is discussed. The general model formulation and the decision process of POMDP models are explained. Applications of POMDPs in the literature are presented chronologically. The states, objective function, and action sets of the POMDP models for these applications are discussed with a brief explanation of the model topic. The reasons for the partial observability of the states are explained for each application.

REFERENCES

1. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley & Sons; 1994.
2. White CC, White DJ. Markov decision processes. *Eur J Oper Res* 1989;39:1–16.
3. Cassandra AR. The POMDP page [Internet]. The Cassandra Organization; 2003. Available at <http://www.pomdp.org/pomdp/pomdp-faq.shtml>. Updated 2009 Jul 26; cited 2010 Jan 10.
4. Shiryaev AN. On optimum methods in quickest detection problems. *Theory Probab Appl* 1963;8:22–46.
5. Drake A. Observation of a Markov process through a noisy channel [ScD thesis]. Massachusetts: Massachusetts Institute of Technology; 1962.
6. Astrom K. Optimal control of Markov processes with incomplete state information. *J Math Anal Appl* 1965;10:174–205.
7. Sondik EJ. The optimal control of partially observable Markov processes [dissertation]. California: Stanford University; 1971.
8. Smallwood RD, Sondik EJ. The optimal control of partially observable processes over a finite horizon. *Oper Res* 1973;21:1071–1088.
9. Cassandra AR, Kaelbling LP, Littman ML. Acting optimally in partially observable stochastic domains. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Menlo Park (CA): AAAI Press/The MIT Press; 1994. pp. 1023–1028.
10. Cassandra AR, Littman ML, Zhang NL. Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes. *Proceedings Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*. San Francisco (CA): Morgan Kaufmann; 1997. pp. 54–61.

11. Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. *Artif Intell* 1998;101: 99–134.
12. Cheng HT. Algorithms for partially observable Markov processes [dissertation]. Vancouver: University of British Columbia; 1988.
13. Hauskrecht M. Value-function approximations for partially observable Markov decision processes. *J Artif Intell Res* 2000;13:33–94.
14. Lovejoy WS. Computationally feasible bounds for partially observed Markov decision processes. *Oper Res* 1991;39(1):192–175.
15. Lovejoy WS. A survey of algorithmic methods for partially observed Markov decision processes. *Ann Oper Res* 1991;28:47–66.
16. White CC III. Partially observed Markov decision processes: a survey. *Ann Oper Res* 1991;32:215–230.
17. Monahan G. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Manage Sci* 1982;28:1–16.
18. Albright S. Structural results for partially observable Markov decision processes. *Oper Res* 1979;27(5):1041–1053.
19. Striebel C. Sufficient statistics in the control of stochastic systems. *J Math Appl* 1965;12: 576–592.
20. Ivy J, Nembhard H. A modeling approach to maintenance decisions using statistical quality control and optimization. *Qual Reliab Eng Int* 2005;21(4):355–366.
21. Cassandra AR. A survey of POMDPs applications. Planning with Partially Observable Markov Decision Processes. AAAI Symposium; 1998 Oct 22–24; Orlando, FL. 1998.
22. Littman ML. A tutorial on partially observable Markov decision processes. *J Math Psychol* 2009;53:119–125.
23. Girshick M, Rubin H. A Bayes' approach to a quality control model. *Ann Math Stat* 1952; 23:114–125.
24. Eckles J. Optimum maintenance with incomplete information. *Oper Res* 1968;16:1058–1067.
25. Ross S. Quality control under Markovian deterioration. *Manage Sci* 1971;17:587–596.
26. Rosenfield D. Markovian deterioration with uncertain information. *Oper Res* 1976;24: 141–155.
27. Pierskalla W, Voelker J. A survey of maintenance models: the control and surveillance of deteriorating systems. *Naval Res Logist Q* 1976;23:353–388.
28. Wang R. Computing optimal quality control policies-two actions. *J Appl Probab* 1976;13: 826–832.
29. Wang R. Optimal replacement policy under unobservable states. *J Appl Probab* 1977;14: 340–348.
30. White C. A Markov quality control process subject to partial observation. *Manage Sci* 1977;23:843–852.
31. White C. Optimal inspection and repair of a production process subject to deterioration. *J Oper Res Soc* 1978;29:235–243.
32. White C. Bounds on optimal cost for a replacement problem with partial observation. *Naval Res Logist Q* 1979;26:415–422.
33. White C. Optimal control-limit strategies for a partially observed replacement problem. *Int J Syst Sci* 1979;10:321–331.
34. Hopp W, Wu S. Multiaction maintenance under Markovian deterioration and incomplete state information. *Naval Res Logist* 1988;35:447–462.
35. Hopp W, Kuo Y. An optimal structured policy for maintenance of partially observable aircraft engine components. *Naval Res Logist* 1998;54:335–352.
36. Ivy JS, Pollock SM. Marginally monotonic maintenance policies for a multi-state deteriorating machine with probabilistic monitoring and silent failures. *Reliab IEEE Trans* 2005;54:489–497.
37. Maillart L. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Trans* 2006;38:463–475.
38. Kuo Y. Optimal adaptive control policy for joint machine maintenance and product quality control. *Eur J Oper Res* 2006;171:586–597.
39. Ghasemi A, Yacout S, Ouali MS. Optimal condition based maintenance with imperfect information and the proportional hazards model. *Int J Prod Res* 2007;45:989–1012.
40. Yeung T, Cassady R, Schneider K. Simultaneous optimization of $\bar{X}$ control chart and age-based preventive maintenance policies under an economic objective. *IIE Trans* 2008;40:147–159.
41. Karush W, Dear EE. Optimal stimulus presentation strategy for a stimulus sampling model of learning. *J Math Psychol* 1966;3:15–47.
42. Kaplan R. Optimal investigation strategies with imperfect information. *J Account Res* 1969;7:32–43.

43. Pollock S. A simple model of search for a moving target. *Oper Res* 1970;18:883–903.
44. Smallwood RD. The analysis of economic teaching strategies for a simple learning model. *J Math Psychol* 1971;8(2):285–301.
45. Smallwood RD, Sondik E, Offensend F. Toward an integrated methodology for the analysis of health-care systems. *Oper Res* 1971;19:1300–1322.
46. White CC III. Optimal diagnostic questionnaires which allow less than truthful responses. *Inf Control* 1976;32:61–74.
47. Segall A. Dynamic file assignment in a computer network. *IEEE Trans Automat Control* 1976;AC-21:161–173.
48. Hughes J. Optimal internal auditing timing. *Account Rev* 1977;LII:56–58.
49. Lane DE. A partially observable model of decision making by fishermen. *Oper Res* 1989;37(2):240–254.
50. Simmons R, Koenig S. Probabilistic robot navigation in partially observable environments. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Montreal, Canada; 1995. pp. 1080–1087.
51. Koenig S, Simmons R. Xavier: a robot navigation architecture based on partially observable Markov decision process models. In: Kortenkamp D, Bonasso R, Murphy R, editors. *Artificial intelligence based mobile robotics: case studies of successful robot systems*. Cambridge (MA): MIT Press; 1998. pp. 91–122.
52. Crites RH, Barto AG. Improving elevator performance using reinforcement learning. *Advances in neural information processing systems* 8. Cambridge (MA): MIT Press; 1996.
53. Darrell T, Pentland A. Active gesture recognition using partially observable Markov decision processes. *Thirteenth IEEE International Conference on Pattern Recognition (IPCR'96)*. Vienna, Austria; 1996. (Also appears as Technical Report No. 367 from MIT Media Laboratory Perceptual Computing Section.)
54. Hu C, Lovejoy WS, Shafer SL. Comparison of some sub-optimal control policies in medical drug therapy. *Oper Res* 1996;44(5):696–709.
55. Hauskrecht M. Planning and control in stochastic domains with imperfect information [dissertation]. Massachusetts: Massachusetts Institute of Technology; 1997.
56. Hauskrecht M, Fraser H. Modeling treatment of ischemic heart disease with partially observable Markov decision processes. *Proceedings of American Medical Informatics Association Annual Symposium on Computer Applications in Health Care*. Orlando, Florida; 1998. pp. 538–542.
57. Hauskrecht M, Fraser H. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artif Intell Med* 2000;18:221–244.
58. Murphy K. Bayesian map learning in dynamic environments. *Advances in neural information processing systems (NIPS'99)*; Nov 29–Dec 4; Denver, CO. 1999.
59. Roy N, Pineau J, Thrun S. Spoken dialogue management using probabilistic reasoning. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*; Hong Kong. 2000.
60. Regan K, Cohen R, Poupart P. The Advisor-POMDP: a principled approach to trust through reputation in electronic markets. *Conference of Privacy Security and Trust*. St. Andrews, New Brunswick, Canada; 2005. pp. 121–130.
61. Boger J, Poupart P, Hoey J, *et al.* A decision theoretic approach to task assistance for persons with dementia. *Proceedings of the International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland; 2005.
62. Williams J, Young S. Partially observable Markov decision processes for spoken dialog systems. *Comput Speech Lang* 2007;21(2): 231–422.
63. Maillart L, Ivy J, Ransom S, *et al.* Assessing dynamic breast cancer screening policies. *Oper Res* 2008;56(6):1411–1427.

PASTA AND RELATED RESULTS

MUHAMMAD EL-TAHA
Department of Mathematics and
Statistics, University of Southern
Maine, Portland, Maine

Poisson arrivals see time averages (*PASTA*) is one of the fundamental properties of queueing and related stochastic systems. *PASTA* and its generalization arrivals see time averages (*ASTA*) play an important role in the analysis and understanding of stochastic systems. Its usefulness covers a wide range of applications.

Consider a queueing process with an associated arrival process satisfying a lack-of-anticipation assumption (*LAA*), which says that future arrivals occur independently of the present and past behavior of the queueing process. If the arrival process is Poisson, then *PASTA* says that an arriving customer sees the same as an independent outside observer (i.e., customer averages are the same as time averages). More generally, consider a random process X (the state of the system) that evolves in continuous time such that its states may be sampled (observed) at a sequence of discrete events (an imbedded point process N). For a given set of states, form the time-average and the event-average frequencies of the process. Then the question of whether or not *arrivals see time averages* is concerned with finding sufficient and/or necessary conditions for the equality of event and time averages.

This article is organized as follows: in the section titled “Setting and Problem Definition,” we provide the problem setting and definitions. We also give Wolff’s *PASTA* under *LAA* and give examples of where *PASTA* holds, and where it fails. In the section titled “Sample-Path Framework,” we discuss *PASTA/ASTA* in a sample-path framework. The results pertaining to conditional *ASTA*, and reversed *ASTA* have been included. Several simple examples are provided. In the section titled “Stochastic

Framework,” *ASTA* results are given in a stochastic framework. We use the expectation framework provided by Melamed and Whitt [1], where we give the lack-of-bias-assumption (*LBA*) and discuss its connection to a similar condition given in the sample-path framework. Finally, in the section titled “Remarks and References,” we provide a literature review emphasizing *PASTA* type results and applications that are not covered in this article.

SETTING AND PROBLEM DEFINITION

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the common probability space over which all stochastic processes will be defined. Let $X \equiv \{X(t), t \geq 0\}$ be a stochastic process taking values in a complete separable metric space, S , with the associated σ -algebra, $\mathcal{G}$, generated by the open sets. Let $N \equiv \{N(t), t \geq 0\}$ be a simple counting process, which is nonnull and nonexplosive, that is, $0 < E[N(t)] < \infty, t > 0$. The jump times of N form a stochastic sequence $T \equiv \{T_n, n \geq 0\}$, where $T_n = \inf\{t > 0, N(t) \geq n\}$ for $n \geq 1$, and $T_0 = 0$ by convention. The process X is referred to as the *state* process and represents the stochastic evolution of the system of interest. Let the sample paths of X be right-continuous with left-hand limits. The counting process N is referred to as the *arrival* process, and its jump times constitute the imbedded times of the state process, that is, the imbedded state process is a stochastic sequence $X^N \equiv \{X(T_n-), n \geq 0\}$. Define the following limits when they exist.

$$V(t) = t^{-1} \int_0^t f(X(s)) ds \quad \text{and} \quad (1)$$

$$W(n) = n^{-1} \sum_{k=1}^n f(X(T_k-)), \quad (2)$$

where f is any real-valued bounded function. Here, $V(t)$ is a time average while $W(n)$ is an imbedded event (arrival) average.

In many applications, the points of N will be a subset of the set of points where X

changes state. For example, X can be the queue length process in a $G/G/1$ queue and N , the points at which X makes a transition from state i to state $i + 1$ for some $i \geq 0$ (i.e., an arrival occurs). In this example, our definition of N ensures that an arrival does not see itself. In such applications, $\{N(s) : 0 \leq s \leq t\}$ is completely determined by $\{X(s) : 0 \leq s \leq t\}$. See Refs 2–5 for applications to Markov processes that exploit this property. Our more general framework is consistent with that of Refs 1, 6–10 in the sense that it is not necessary that all, or any of the points of N coincide with the changes of state of X .

As an extreme case, N could represent the time points of observations that occur independently of the evolution of X . As an application of this situation, suppose that we are interested in estimating the probability that a (strictly) stationary stochastic process X belongs to a set of states B . By the Ergodic Theorem, this probability is (a.s.) equal to the asymptotic time-average fraction of time that the process spends in B . The results in this article confirm that we can obtain an estimate of these probabilities by sampling the process over a sequence of discrete-time points (i.e., there is no need to continuously observe the process over time) provided that the sequence of sampling instants and the original process satisfy the *LAA* assumption. One way to guarantee that *LAA* is satisfied is to ensure that the time between sampling instants are *iid*, exponentially distributed, and independent of the state of the original process. For instance, consider a communication link between two subnets. It is often necessary to collect data to measure the performance of the link (utilization, job processing times, delay, and so on). Continuous monitoring of the link can degrade the performance due to unacceptable overhead. If the link is observed (sampled) at predetermined discrete-time points (so that *LAA* holds), we can ensure that the collected event-average statistics provide good approximations of the statistics of interest without seriously degrading system performance.

We shall need the following *LAA* introduced by Wolff [6].

Assumption [LAA]. For all $t \geq 0$, $\{N(t + h) - N(t); h \geq 0\}$ is independent of $\{X(\tau); 0 \leq \tau \leq t\}$.

The *LAA* assumption implies that the number of points from N after time t is independent of the past behavior of the process X . Melamed and Whitt [8] and Stidham and El-Taha [11] show that the whole past may be replaced by the present when joint stationarity of X and N is assumed. We state a slight variation of *PASTA* as given by Wolff [6].

Theorem 1. *Under the assumptions that N is a Poisson process and *LAA* holds, $V(t) \rightarrow V(\infty)$ w.p.1 if and only if $W(n) \rightarrow W(\infty)$ w.p.1 in which case*

$$W(\infty) = V(\infty) \text{ w.p.1.}$$

Special Cases

(i) Let B be any subset of the state space of X . Now let f be an indicator function such that $f(\cdot) = \mathbf{1}\{X(\cdot) \in B\}$. Then we obtain Wolff's *PASTA*, which says that the probability that an arrival sees X in set B is the same as the time-average probability that X is in set B ; that is, the probability that an arrival sees $X \in B$ is the same as the probability a random observer sees $X \in B$. More precisely,

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n \mathbf{1}\{X(T_k-) \in B\} \\ = \lim_{t \rightarrow \infty} t^{-1} \int_0^t \mathbf{1}\{X(s) \in B\} ds \text{ w.p.1. } (3) \end{aligned}$$

(ii) Consider an $M/GI/1$ queue where $X(t)$ is the number of customers in the system at time t , and service times are *iid* and independent of the arrival process. Let f be the identity function. Clearly, the conditions of *PASTA* hold. Then the *PASTA* property says that the long run time-average number of customers in the system $L = \lim_{t \rightarrow \infty} t^{-1} \int_0^t X(s) ds$ is equal to the long run average number of customers in the system as seen by arrivals $L_A = \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X(T_k-)$. Here, $\{T_k, k \geq 1\}$ represent arrival instants. Now if T_k represents departure instants, then *PASTA* says L is equal to the mean

number of customers in the system as left behind by departing customers.

(iii) Let $X(t)$ represent the work in a single server queueing system at time t , and f be the identity function. Then $W(\infty)$ is the average work in the system at arrival instants (average delay per customer) and $V(\infty)$ is the average work in the system as seen by a random observer (virtual work in the system). PASTA ensures the equality of the time and arrival averages, that is, $W(\infty) = V(\infty)$ w.p.1.

(iv) Here, we give an example where PASTA does not hold. Consider a $GI/GI/1$ queueing model where time between arrivals are iid and uniform (1.5,2.5), and service times are deterministic and equal to 1. Then $\rho = \lambda/\mu = 0.5$, so $p(0) = p(1) = 0.5$. However, the arrival probabilities are $\pi^-(0) = 1$ and $\pi^-(n) = 0, n \geq 1$.

SAMPLE-PATH FRAMEWORK

In this section, we present ASTA in a sample-path framework. Included are the results pertaining to conditional ASTA, and reversed ASTA. Several simple examples are provided. We shall begin with the following preliminary results.

Background on Point Processes; $Y = \lambda X$

Let $\{T_k, k \geq 1\}$ be a deterministic sequence of time points, with $0 \leq T_k \leq T_{k+1} < \infty, T_0 = 0$, and $k \geq 1$. Define $N(t) := \max\{k \geq 1 : T_k \leq t\}$, $t \geq 0$. We interpret T_k as the time point at which the k th of a sequence of events occurs, such as the k th arrival to a queue, and $N(t)$ as the number of events in $[0, t]$. Note that our definition allows more than one event to occur at the same time point (e.g., batch arrivals). That is, the point process $N = \{N(t), t \geq 0\}$ need not be simple. We assume that $T_k \rightarrow \infty$ as $k \rightarrow \infty$, so that there are only a finite number of events in any finite time interval ($N(t) < \infty$ for all $t \geq 0$). Note that $N(t) \rightarrow \infty$ as $t \rightarrow \infty$, since $T_k < \infty$ for all $k \geq 1$.

First, we state the following elementary lemma, which is a sample-path analog of the elementary renewal theorem. (For a proof, see Refs 12, 13.)

Lemma 1. *For any $0 \leq \lambda \leq \infty$, the following are equivalent:*

1. $N(t)/t \rightarrow \lambda$, as $t \rightarrow \infty$;
2. $T_n/n \rightarrow \lambda^{-1}$, as $n \rightarrow \infty$.

The following lemma is a sample-path analog of the renewal-reward theorem ($Y = \lambda X$).

Lemma 2. *Suppose $t^{-1}N(t) \rightarrow \lambda$, $0 \leq \lambda \leq \infty$, as $t \rightarrow \infty$. Let $\{X_k, k \geq 1\}$ be a deterministic sequence of nonnegative real numbers and define*

$$Y(t) := \sum_{k=1}^{N(t)} X_k, \quad t \geq 0. \quad (4)$$

Then

1. *if $n^{-1} \sum_{k=1}^n X_k \rightarrow X$, $0 \leq X \leq \infty$, as $n \rightarrow \infty$, then $t^{-1}Y(t) \rightarrow Y = \lambda X$ as $t \rightarrow \infty$, provided that λX is well defined;*
2. *if $t^{-1}Y(t) \rightarrow Y$, $0 \leq Y \leq \infty$, as $t \rightarrow \infty$, then $n^{-1} \sum_{k=1}^n X_k \rightarrow X = \lambda^{-1}Y$ as $n \rightarrow \infty$, provided that $\lambda^{-1}Y$ is well defined.*

Remark. Lemmas 1 and 2 pertain to deterministic processes, or, equivalently, to individual sample paths of stochastic processes. Thus, they do not require explicit stochastic assumptions, only the existence of certain limiting averages on the sample path in question. One can apply these results to stochastic models in which the stochastic assumptions imply (via an ergodic theorem or law of large numbers) that the sample-path averages in question exist almost surely. For example, suppose that $\{(T_n - T_{n-1}, X_n)\}$ is an ergodic, strictly stationary sequence and that $\{Y(t), t \geq 0\}$ is defined in terms of $\{X_k, k \geq 1\}$ by Equation (4). Then the ergodic theorem and Lemma 2 imply that $n^{-1}T_n \rightarrow E[T_1] = \lambda^{-1}$ and $n^{-1} \sum_{k=1}^n X_k \rightarrow E[X_1]$ a.s. as $n \rightarrow \infty$, and that $t^{-1}Y(t) \rightarrow Y = E[X_1]/E[T_1]$ a.s. as $t \rightarrow \infty$. Other variants of Lemma 2 are given by El-Taha and Stidham [7,14]. See the sections on “Point Processes and Renewal” in this encyclopedia and the article **Regenerative Processes** for further information on these topics.

Relations between Frequencies for a Process with an Imbedded Point Process

In this section, we use $Y = \lambda X$ to give simple derivations of relations between time-average and point-average frequency distributions for processes with imbedded point process. Specifically, we will give the covariance formula that leads to conditions for ASTA.

Let $X = \{X(t), t \geq 0\}$ be a deterministic continuous-time process with state space S . We assume that S is a Polish (complete separable metric) space, with the Borel σ -field $\mathcal{B}(S)$, and that X is right-continuous with left-hand limits. Let $\{T_n, n \geq 1\}$ be an associated deterministic point process with counting process $N = \{N(t), t \geq 0\}$ satisfying the conditions of the section titled “Background on Point Processes; $Y = \lambda X$.” (Recall that these conditions do not include a requirement that N be simple.)

The process N may (but does not necessarily) count transition instants of process X ; for example, level crossings or points where upward or downward jumps (e.g., arrivals or departures in a queueing system) occur. In general, transition instants can be points of continuity as well as points of discontinuity of process X .

For an arbitrary set of states $B \in \mathcal{B}(S)$, define the following limits, when they exist:

$$\lambda := \lim_{t \rightarrow \infty} \frac{N(t)}{t} = \lim_{n \rightarrow \infty} \frac{n}{T_n}, \quad (5)$$

$$p(B) := \lim_{t \rightarrow \infty} t^{-1} \int_0^t \mathbf{1}\{X(s) \in B\} ds, \quad (6)$$

$$q(B) := \lim_{t \rightarrow \infty} \frac{\int_0^t \mathbf{1}\{X(s-) \in B\} dN(s)}{\int_0^t \mathbf{1}\{X(s) \in B\} ds}, \quad (7)$$

$$\pi^-(B) := \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n \mathbf{1}\{X(T_k-) \in B\}. \quad (8)$$

That is, λ is the rate of the point process N , $\{p(B), B \in \mathcal{B}(S)\}$ is the time-average frequency distribution of the process X , $q(B)$ is the conditional rate of N while $X(t-)$ is in B , and $\{\pi^-(B), B \in \mathcal{B}(S)\}$ is the point-average frequency distribution of $\{X(t-), t \geq 0\}$. In the “classical” application to queueing systems,

N is the arrival process and $\pi^-(B)$ is the frequency with which arrivals “see” the system in the set of states B . (Recall that $X(t)$ is right-continuous with left limits, so that $X(T_k-)$ is the state of the system just before the k th “arrival.”)

In what follows, we shall make frequent use of the identity

$$\int_0^t \mathbf{1}\{X(s-) \in B\} dN(s) = \sum_{k=1}^{N(t)} \mathbf{1}\{X(T_k-) \in B\}. \quad (9)$$

The following theorem gives a relation between the time-average frequency of the set B and the frequency of B “as seen by an arrival.”

Theorem 2. *Let $B \in \mathcal{B}(S)$ and assume that the limits λ and $q(B)$ exist, $0 \leq \lambda < \infty$, $0 \leq q(B) < \infty$. Then $p(B)$ exists if and only if $\pi^-(B)$ exists, in which case*

$$p(B)q(B) = \lambda\pi^-(B). \quad (10)$$

Proof. Let $Y(t) = \int_0^t \mathbf{1}\{X(s-) \in B\} dN(s)$, so that $Y(t)$ is the number of points that occur in $[0, t]$ while $X(s-) \in B$, $t \geq 0$. Let $X_k = \mathbf{1}\{X(T_k-) \in B\}$, $k \geq 1$ and apply Lemma 2, using Equation (9) and the fact that $\lim_{t \rightarrow \infty} t^{-1}Y(t) = p(B)q(B)$, when the limits exist.

Remark. Theorem 2 plays a role similar to the Palm transformation formula in the stochastic theory of marked point processes and processes related to point processes (7, Appendix A), [15–18] (see also the section on “Point Processes” in this encyclopedia). Some authors call Theorem 2 the *covariance formula* [1, 14].

Reversed Processes. Now we give sample-path versions of these relations for reversed processes, again using the same framework as for ASTA. We begin by defining

the following limits when they exist, for $B \in \mathcal{B}(S)$:

$$q'(B) := \lim_{t \rightarrow \infty} \frac{\int_0^t \mathbf{1}\{X(s) \in B\} dN(s)}{\int_0^t \mathbf{1}\{X(s) \in B\} ds}, \quad (11)$$

$$\pi^+(B) := \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n \mathbf{1}\{X(T_k) \in B\}. \quad (12)$$

That is, $q'(B)$ is the *conditional rate* of N while the reversed version of process X is in B , and $\{\pi^+(B), B \in \mathcal{B}(S)\}$ is the *point-average frequency distribution* of $X = \{X(t+), t \geq 0\}$. (Recall that X is right-continuous.) Informally, we refer to $\{\pi^+(B), B \in \mathcal{B}(S)\}$ as the frequency distribution for X “just after” a point from N has occurred. Equivalently, it is the frequency distribution of the reversed version of X “just before” a point from N occurs. In applications to queueing systems, N is often the process that counts departures and hence $\{\pi^+(B), B \in \mathcal{B}(S)\}$ is the frequency distribution of the state of the system “left behind” by a departure.

By a proof exactly parallel to that used to prove Theorem 2 one can derive the following relation between $p(B)$ and $\pi^+(B)$, assuming the relevant limits are well defined:

$$p(B)q'(B) = \lambda\pi^+(B). \quad (13)$$

Equation (13) will be used to give necessary and sufficient conditions for *ASTA*-type results for reversed processes.

Characterization of *ASTA* and Related Properties

Theorem 2 yields necessary and sufficient conditions for the *ASTA* property, as shown in the following theorem. (To simplify the presentation of our results throughout this section, we shall assume that all relevant limits are well defined and that $\{p(B), B \in \mathcal{B}(S)\}$ and $\{\pi^-(B), B \in \mathcal{B}(S)\}$ are proper frequency distributions.)

Theorem 3 [Conditions for *ASTA*]. For all $B \in \mathcal{B}(S)$,

$$p(B) = \pi^-(B) \Leftrightarrow q(B) = \lambda.$$

Moreover, the following are equivalent:

1. $q(B)$ is independent of B , $B \in \mathcal{B}(S)$;
2. $q(B) = \lambda$, for all $B \in \mathcal{B}(S)$;
3. the frequency distributions $\{p(B), B \in \mathcal{B}(S)\}$ and $\{\pi^-(B), B \in \mathcal{B}(S)\}$ coincide.

Proof. The proof follows immediately from Theorem 2.

Remark. This theorem can be used as the starting point for a proof of *PASTA*. Most proofs of *PASTA* in the literature [6] contain a step in which it is shown that Poisson arrivals together with a *LAA* (roughly speaking, the future of N does not depend on the past or present of X) imply that $q(B) = \lambda$, for all $B \in \mathcal{B}(S)$, or the probabilistic equivalent of this statement. El-Taha and Stidham [19,20] use martingale and weak-convergence theory to prove stronger results under the assumptions of Wolff [6]. In particular, it is shown that the Poisson point process N , restricted to times t when $X(t) \in B$, is a probabilistic replica of the original point process; that is, it is Poisson with the same rate λ .

Conditional *ASTA*. Now we provide a sample-path proof of conditional *ASTA* in the same framework as used above for *ASTA*. Let $A \in \mathcal{B}(S)$, $B \in \mathcal{B}(S)$. Then by Theorem 2,

$$p(A \cap B)q(A \cap B) = \lambda\pi^-(A \cap B). \quad (14)$$

Define the following limits, assuming they exist:

$$p(A | B) := \lim_{t \rightarrow \infty} \frac{\int_0^t \mathbf{1}\{X(s) \in A \cap B\} ds}{\int_0^t \mathbf{1}\{X(s) \in B\} ds}, \quad (15)$$

$$\pi^-(A | B) := \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \mathbf{1}\{X(T_k-) \in A \cap B\}}{\sum_{k=1}^n \mathbf{1}\{X(T_k-) \in B\}}. \quad (16)$$

Then it follows from Equations (6) and (15) that

$$p(A | B) = \frac{p(A \cap B)}{p(B)}, \quad (17)$$

and from Equations (8) and (16) that

$$\pi^-(A | B) = \frac{\pi^-(A \cap B)}{\pi^-(B)}, \quad (18)$$

Clearly then $p(A | B)$ and $\pi^-(A | B)$ are sample-path versions of the conditional probabilities of A given B at an arbitrary time point and at a point of the imbedded point process, respectively. Now, multiplying both sides of Equation (17) by $q(A \cap B)$ and dividing both sides of Equation (17) by $q(B)$,

$$\begin{aligned} \frac{p(A | B)q(A \cap B)}{q(B)} &= \frac{p(A \cap B)q(A \cap B)}{p(B)q(B)} \\ &= \frac{\lambda \pi^-(A \cap B)}{\lambda \pi^-(B)} \\ &= \pi^-(A | B), \end{aligned}$$

from which we obtain the following sample-path version of the *covariance formula* for conditional probabilities:

$$p(A | B)q(A \cap B) = \pi^-(A | B)q(B). \quad (19)$$

As a corollary of this result we obtain the following conditions for conditional ASTA:

Theorem 4 [Conditional ASTA]. *Let $B \in \mathcal{B}(S)$ be given. For all $A \in \mathcal{B}(S)$,*

$$p(A | B) = \pi^-(A | B) \Leftrightarrow q(A \cap B) = q(B).$$

Moreover, the following are equivalent:

1. $q(A \cap B)$ is independent of A , $A \in \mathcal{B}(S)$;
2. $q(A \cap B) = q(B)$, for all $B \in \mathcal{B}(S)$;
3. the conditional frequency distributions $\{p(A | B), A \in \mathcal{B}(S)\}$, and $\{\pi^-(A | B), A \in \mathcal{B}(S)\}$ coincide.

Remark. Intuitively, these results are just Theorems 2 and 3 restricted to B -time, that is, to time points when the process $\{X(t), t \geq 0\}$ is in the fixed set B . One can use this intuition

to construct alternate proofs of Equation (19) and Theorem 4 based on a change-of-time argument (Section 3.5, of Ref. 19), where the concept of B -time is used for a different but related purpose).

Remark. In Refs 1, 21, and 22, conditional ASTA is studied in the special case of a two-dimensional process, $X = (X_1, X_2)$, with state space $S = S_1 \times S_2$ and σ -field $\mathcal{B}(S) = \mathcal{B}(S_1) \otimes \mathcal{B}(S_2)$. In this context, one is interested in the limiting time-average and point-average conditional frequencies with which $X_1(t) \in A_1$, given $X_2(t) \in A_2$, for $A_1 \in \mathcal{B}(S_1)$, $A_2 \in \mathcal{B}(S_2)$. By setting $A = A_1 \times S_2$ and $B = S_1 \times A_2$, in Equation (19) and Theorem 4, one obtains the results presented in Refs 1, 21, and 22.

Reversed Processes and DSTA. Now we give sample-path versions of ASTA-type results for reversed processes, using the same framework as for ASTA. Equation (13) immediately gives necessary and sufficient conditions for departures see time averages: DSTA.

Theorem 5 [Conditions for DSTA]. *For all $B \in \mathcal{B}(S)$,*

$$p(B) = \pi^+(B) \Leftrightarrow q'(B) = \lambda.$$

Moreover, the following are equivalent:

1. $q'(B)$ is independent of B , $B \in \mathcal{B}(S)$;
2. $q'(B) = \lambda$, for all $B \in \mathcal{B}(S)$;
3. the frequency distributions $\{p(B), B \in \mathcal{B}(S)\}$ and $\{\pi^+(B), B \in \mathcal{B}(S)\}$ coincide.

Examples

Example 1 [Finite Capacity Model]. Consider an $M/G/1/K$ model with Poisson arrivals having rate λ , general service times distribution function and finite capacity K . Arrivals that find the system full are lost. Here arrivals that enter the system are not Poisson and PASTA does not hold. But the sample-path version of the covariance formula (Eq. 3) still holds; that is,

$$\bar{\lambda} \pi_n^- = \lambda_n p_n$$

where $\bar{\lambda}$ is the effective arrival rate, and π_n^- ; $n = 0, 1, \dots, K-1$ is the conditional arrival-time probability of arrivals that enter the system (i.e., the long-term fraction of arrivals, among those that join, that see the system in state n). It is well known that $\bar{\lambda} = \lambda(1 - p_K)$, and $\lambda_n = \lambda$ for $n = 0, 1, \dots, K-1$; and 0 otherwise, so $\lambda(1 - p_K)\pi_n^- = \lambda p_n$. Thus,

$$\pi_n^- = \frac{p_n}{1 - p_K}, \quad n = 0, 1, \dots, K-1,$$

which gives a relationship between π_n^- and p_n .

Example 2 [Machine Interference Model]. Consider the machine interference model with a finite population of size N and c exponential identical servers, each work at rate μ . When active, a machine time-to-failure is exponential with rate λ . Here, we have a state-dependent arrival rate $\lambda_n = (N - n)\lambda$ where n is the number of down machines. Note that the time between arrivals at the repair center are independent and exponentially distributed with state dependent rate $\lambda_n = (N - n)\lambda$, that is, the arrival process is not Poisson. Also, for this model, the LAA/LBA does not hold since $\lambda_n \neq \lambda$, so ASTA does not hold. However, relation (10) holds and reduces to

$$\bar{\lambda}\pi_n^- = \lambda_n p_n,$$

where $\bar{\lambda} = \lambda(N - L)$ is the effective arrival rate, and L is the mean number of down machines. Therefore,

$$\pi_n^- = \frac{N - n}{N - L} p_n. \quad (20)$$

Equation (20) gives a relationship between customer-average, π_n^- , and time-average, p_n , probabilities. It can be shown that

$$\pi_n^-(N) = p_n(N - 1), \quad (21)$$

when $\pi_n^-(N)$ is the arrival point probability with N customers in the system (e.g., up and down machines), and $p_n(N - 1)$ is the random observer (time average) probability

of n machines in the repair center with $N - 1$ customers in the entire system. Equation (21) says that an arrival sees the system time-average probability with one customer removed, that is, arrival does not see itself. It is worth noting that Equation (21) is known as the *arrival theorem* in closed Markovian queueing networks. Refer to the section on “*Queueing Theory and Queueing Networks*” in this encyclopedia for more discussion on queueing models.

STOCHASTIC FRAMEWORK

Here, we present ASTA using a stochastic framework based on the approach given by Melamed and Whitt [1] and Melamed and Yao [23], among others. The time-average and customer-average quantities are defined as expected values using stochastic Riemann–Stieltjes integrals.

Consider (X, N) as defined in the section titled “Setting and Problem Definition.” Recall that $X \equiv \{X(t), t \geq 0\}$ is any stochastic process, $N \equiv \{N(t), t \geq 0\}$ is any point process, the sequence of points $\{T_n, n = 0, 1, 2, \dots\}$ form an imbedded point process of X , and the imbedded state process is a stochastic sequence $X^N \equiv \{X^N(T_n^-), n \geq 0\}$, where for notational convenience $X^N(T_n^-) = X(T_n^-)$. The ASTA problem is concerned with finding sufficient and/or necessary conditions for which, in the limit, X and X^N converge to the same distribution. More specifically let

$$X(t) \xRightarrow[t \rightarrow \infty]{} X(\infty) \text{ and } X^N(T_n^-) \xRightarrow[n \rightarrow \infty]{} X^N(\infty),$$

where $\Rightarrow$ denotes convergence in distribution (weak convergence). Then, we say ASTA holds for (X, N) if

$$X(\infty) \stackrel{d}{=} X^N(\infty),$$

where $\stackrel{d}{=}$ denotes equality in distribution.

It is convenient to work with the transformed state process $U \equiv \{U(t), t \geq 0\}$, where $U(t) = f(X(t-))$ and f is a bounded

real-valued measurable function on $\mathcal{G}$. The function f is designed to extract the relevant information from the state description of X . For example, f can be an indicator function, so that the corresponding $U(t)$ can extract detailed stochastic information from X , whose limiting time averages are the state probabilities of X . Let

$$\begin{aligned}\bar{V}(t) &= E \left[\int_0^t U(s) ds \right] / t, \quad t \geq 0 \\ \bar{W}(t) &= \frac{E \left[\sum_{k=1}^{N(t)} U(T_k -) \right]}{E[N(t)]} \\ &= \frac{E \left[\int_0^t U(s-) dN(s) \right]}{E[N(t)]}, \quad t \geq 0,\end{aligned}$$

where $\bar{V}(0) = 0$, and $\bar{W}(t) = 0$, whenever $E[N(t)] = 0$. Then ASTA holds for the pair (U, N) if

$$\lim_{t \rightarrow \infty} \bar{V}(t) = \bar{V}(\infty) = \bar{W}(\infty) = \lim_{t \rightarrow \infty} \bar{W}(t)$$

provided one of the limits exist.

Remark. This framework is connected to the sample-path framework in the following sense. When we have weak convergence, then as $t \rightarrow \infty$

$$V(t) \Rightarrow V(\infty) = \bar{V}(\infty) = \lim_{t \rightarrow \infty} \bar{V}(t), \quad (22)$$

$$W(t) \Rightarrow W(\infty) = \bar{W}(\infty) = \lim_{t \rightarrow \infty} \bar{W}(t). \quad (23)$$

Equations (22) and (23) hold in a regenerative framework or in a stationary ergodic framework. (See the sections on “Point Processes” and “Renewal and Regenerative Processes” in this encyclopedia.) These frameworks suffice for most applications.

ASTA under the LBA Condition. The LBA weakens the LAA by providing a general necessary and sufficient condition for ASTA. The LBA condition requires the existence of a conditional stochastic intensity process, $\Lambda_c = \{\Lambda_c(t), t > 0\}$ of N given U .

Definition. [Conditional Stochastic Intensity]. The conditional intensity, $\Lambda_c(t)$, of N at t given $U(t)$, is a random variable

$$\Lambda_c(t) = \lim_{h \rightarrow 0} \frac{E[N(t+h) - N(t) | U(t)]}{h}, \quad t \geq 0$$

provided the limit exists.

A motivation for calling $\Lambda_c(t)$ an intensity process is offered by the fact that, if the interchange of limit and expectation operations is allowed, then

$$\begin{aligned}E[\Lambda_c(t)] &= E \left[\lim_{h \rightarrow 0} \frac{E[N(t+h) - N(t) | U(t)]}{h} \right] \\ &= \lim_{h \rightarrow 0} E \left[\frac{E[N(t+h) - N(t) | U(t)]}{h} \right] \\ &= \lambda(t),\end{aligned} \quad (24)$$

where $\lambda(t) \triangleq \frac{d}{dt} E[N(t)]$ is the unconditional intensity (rate) of $N(t)$. It is useful to think of $\Lambda_c(t)$ as the conditional (random) rate of a new arrival (point of N) at time t , given the state of the process is $U(t)$.

Remark. Let f be the identity function so that $U = X$. Let the state of X be discrete (e.g., number of customers in a queueing system) and assume that X is strictly stationary with known stationary distribution. Now given $X(t)$ in any set of states B , then $\Lambda_c(t) = q(B)$ w.p.1, in particular, if $X(t) = n$, $\Lambda_c(t) = \lambda_n$ w.p.1, where λ_n is first given in the section titled “Sample-Path Framework.”

Definition [LBA]. The pair (U, N) is said to satisfy the LBA, if U and Λ_c are point-wise uncorrelated, that is, $\text{cov}[U(t), \Lambda_c(t)] = 0$, for all $t \geq 0$.

The LBA weakens LAA by substituting a point-wise condition for an interval-based condition. This fact, together with stochastic intensity definition, imply that LBA is a local condition in the sense that the interactions between U and N (using Λ_c) need to be known only in an infinitesimal forward interval, and only through covariances which are a measure of linear dependence.

Lemma 3 [The Bias Formula]. *Suppose that*

1. *U and N are jointly stationary in the sense that*

$$\begin{aligned} & [U(t), N(t+h) - N(t)] \\ & \stackrel{d}{=} (U(t+s), N(t+s+h) - N(t+s)), \\ & t, s, h > 0 \end{aligned}$$

2. *Λ_c is well-defined and obeys Equation (24), so that*

$$E[\Lambda_c(t)] = \lambda, \quad t \geq 0,$$

where $\lambda = \lim_{h \rightarrow 0} \frac{N(h)}{h}$. Then, both $\bar{V}(t)$ and $\bar{W}(t)$ are independent of t and

$$\lambda(\bar{W}(t) - \bar{V}(t)) = \text{cov}[U(0), \Lambda_c(0)], t \geq 0. \quad (25)$$

The Bias formula (Eq. 25), also called *covariance formula*, gives a relation between the event averages $\bar{W}(t)$, time averages $\bar{V}(t)$, and the $\text{cov}[U(0), \Lambda_c(0)]$. Let X be a strictly stationary ergodic stochastic process, and let $U(t) = 1\{X(t) \in B\}$ where B is an arbitrary set in S . In this case, we show that Equation (25) is equivalent to the sample-path covariance formula (Eq. 10). By the ergodic theorem, it follows that

$$\begin{aligned} \bar{W}(t) &= \pi^-(B), \quad \bar{V}(t) = p(B), \text{ and} \\ E[U(0)] &= p(B) \quad \text{w.p.1.} \end{aligned} \quad (26)$$

Moreover,

$$\begin{aligned} \text{Cov}[U(0), \Lambda_c(0)] &= E[U(0)\Lambda_c(0)] - E[U(0)]E[\Lambda_c(0)] \\ &= E[U(0)E[\Lambda_c(0) | U(0)]] - \lambda p(B) \\ &= q(B)E[U(0)] - \lambda p(B) \\ &= q(B)p(B) - \lambda p(B). \end{aligned} \quad (27)$$

Substitute Equations (26) and (27) in Equation (25) to obtain (w.p.1)

$$\lambda \pi^-(B) = q(B)p(B),$$

which is the sample-path covariance formula given by Equation (10).

The following theorem and its proof are given in Ref. 1.

Theorem 6. [ASTA under Stochastic LBA]. *Suppose that the conditions of Lemma 3 hold for the pair (U, N) . Then $\bar{W}(\infty) = \bar{V}(\infty)$ iff LBA holds for (U, N) .*

Proof. Assume f is continuous, then

$$\begin{aligned} & E \left[\int_0^t U(s-) dN(s) \right] \\ &= E \left[\lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} U \left(\frac{kt-}{n} \right) \right. \\ & \quad \left. \times \left\{ N \left(\frac{(k+1)t}{n} \right) - N \left(\frac{kt}{n} \right) \right\} \right] \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} E \left[U \left(\frac{kt-}{n} \right) \right. \\ & \quad \left. \times \left\{ N \left(\frac{(k+1)t}{n} \right) - N \left(\frac{kt}{n} \right) \right\} \right] \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} E[U(0)N(t/n)] \\ &= \lim_{n \rightarrow \infty} nE[U(0)E[N(t/n) | U(0)]] \\ &= E \left[U(0) \lim_{n \rightarrow \infty} nE[N(t/n) | U(0)] \right] \\ &= tE \left[U(0) \lim_{n \rightarrow \infty} \frac{E[N(t/n) | U(0)]}{t/n} \right] \\ &= tE[U(0)\Lambda_c(0)], \end{aligned}$$

where steps 1 and 2 are justified using Riemann–Stieltjes sum approach and Lebesgue-Dominated Convergence Theorem. Further details are given in Ref. 1. We also use the first condition in Lemma 3 in step 3, and calculus of conditioning in step 4. In the second-last step, convergence and uniform integrability remain unaltered on multiplication by the bounded random variable $U(0)$. Hence

$$\begin{aligned} \bar{W}(t) &= \frac{tE[U(t)\Lambda_c(t)]}{E[N(t)]} = \frac{tE[U(t)\Lambda_c(t)]}{E[\Lambda_c(t)]t} \\ &= E[U(t)], \end{aligned}$$

where we use *LBA* in the last step. But, by the definition of $\bar{V}(t)$ and the stationarity of U

$$\bar{V}(t) = E[U(t)].$$

So the result follows for continuous f . The extension to all measurable f , for which the expectations are well defined, is obtained by taking limits (24, pp. 7–9).

One can show that *ASTA* holds under the following *weak lack-of-anticipation assumption* (*WLAA*) given by Melamed and Whitt [1].

Definition. [*WLAA*]. The pair (U, N) is said to satisfy the *WLAA*, if for each $t \geq 0$ there exists $s_0 > 0$, such that $U(t)$ and $N(t+s) - N(t)$ are uncorrelated for all $0 \leq s \leq s_0$.

Note that *WLAA* is weaker than *LAA* and stronger than the *LBA*. Its primary advantage is that it may be easier than *LBA* to verify in certain applications. Moreover, the proof of *ASTA* under *WLAA* is straightforward.

REMARKS AND REFERENCES

Descloux [25], Strauch [26], Stidham [27], and Cooper [28] give elementary arguments for *PASTA*. In Ref. 6, Wolff proves *PASTA* in a general setting. In subsequent articles [1,10,11,14,22,29–33], the requirement that the point process be Poisson has been eliminated as a condition for events to see time averages, thus extending *PASTA*. In addition, the *LAA* assumption has been replaced by the weaker *LBA* [1,34,23,7]. Additional work on conditional *PASTA/ASTA* is found in Refs 21, 22 and 35.

In this article we present *PASTA* and related results, focusing on *PASTA* as given by Wolff [6], sample-path *ASTA* including conditional *ASTA* and reversed *ASTA* as given by Stidham and El-Taha [11,36], El-Taha and Stidham [7,14,19,20,37], and El-Taha [22]. We also cover *ASTA* in the expectation framework using stochastic Riemann–Stieltjes sum approach as given by Melamed and Whitt [1] and Melamed

and Yao [23]. We give several elementary examples to illustrate the issues discussed.

There are additional topics on *PASTA* and related problems that are not covered in this article. Below, we give a brief literature review of the martingale approach, the stationary framework approach, the filtered *ASTA* problem, *ASTA* in discrete-time, fluid *ASTA*, and applications to queueing networks among other topics.

Several authors use martingale theory to prove *PASTA* and related results. The martingale approach takes advantage of the martingale decomposition, stochastic integration, martingale *SLLN*, and stochastic intensities. It has the advantage of leading to general versions of *ASTA* without the stationarity assumptions. The martingale approach is used by Wolff [6] in proving *PASTA*. It is also used to extend *PASTA* to *ASTA* by eliminating the need for Poisson arrivals and replacing *LAA* by *LBA*. The list of authors include Melamed and Whitt [8], Melamed and Yao [23], Yao [38], Brémaud *et al.* [34], Bacelli and Brémaud [17], and Brémaud [9].

Rather than extending *ASTA* by weakening the conditions under which it holds, El-Taha and Stidham [19,20] show that the *ASTA* property can be strengthened under Wolff's original assumptions of Poisson arrivals and *LAA*. They prove that for any subset B , if *LAA* holds and the point process satisfies certain conditions (e.g., doubly stochastic Poisson), the B -process [i.e., the original point process restricted to the time spent by X in set B (B -time)] not only has the same intensity but is in fact a probabilistic replica of the original imbedded point process. In particular, the B -process is Poisson with intensity λ for every set B if (and only if) the original point process is Poisson with intensity λ . Stidham [27] gives an informal proof of this result for the special case of regenerative processes.

The stationary approach, which focuses on the underlying stochastic process being strictly stationary and views the *ASTA* problem as relating two expectations with underlying probability measures, $\mathbf{P}$, and its Palm-measure version, $\mathbf{P}^0$, via the Palm inversion formula, is used by König and Schmidt [32], Franken *et al.* [15], and again

by König and Schmidt [10]. Brémaud [9] connects the stationary framework and the martingale approach through Papangelou's lemma and Radon Nykodym derivative.

Given that *ASTA* holds, then anti-*PASTA*, a converse of *PASTA*, finds conditions under which the arrival process is Poisson. In a Markovian framework, the arrival process must be Poisson. For a discussion on the anti-*PASTA* problem, the reader may refer to Refs 3, 33, and 4. Using statistical sampling theory, Glynn *et al.* (39 and references therein) consider the issue of estimating customer averages and time averages.

Discrete time *ASTA* is given by Makowski *et al.* [40], using discrete-time martingales, and El-Taha and Stidham [7,19,20], using probabilistic arguments. Conditional and time-reversed *ASTA* results are given in Refs 19, and 21 among others. In Refs 7, and 20, weak convergence theory and limiting arguments are used to go from a discrete-time version to the continuous case. In the process, they extend the results of Ref. 19 to the case of a process with an imbedded cumulative input process, $A = \{A(t), t \geq 0\}$, which is assumed to be a Lévy process with nondecreasing sample paths. This framework allows for modeling fluid processes, as well as compound Poisson processes with noninteger increments.

Applications to queueing networks are considered in Refs 1, 7, 41 and references there in. In particular, in closed Markovian networks, it is shown that in equilibrium arrival-time probabilities are equal to time-average probabilities of the system state with one customer removed, that is, the customer in transit does not see itself. In Markovian networks, Melamed and Yao [23] show the equivalence of quasi-reversibility, *LBA*, and *ASTA*. El-Taha and Stidham [7] give a sample-path proof of the Arrival Theorem, and Serfozo [41] shows that moving units see time averages (*MUSTA*) holds for Markovian Jackson and Whittle networks.

REFERENCES

1. Melamed B, Whitt W. On arrivals that see time averages. *Oper Res* 1990;38:156–172.
2. Walrand J. An introduction to queueing networks. Englewood Cliffs (NJ): Prentice Hall; 1988.
3. Green L, Melamed B. An ANTIPASTA result for Markovian systems. *Oper Res* 1990;38:173–175.
4. Wolff R. A note on PASTA and ANTIPASTA for continuous-time Markov chains. *Oper Res* 1990;38:176–177.
5. Serfozo R. Poisson functionals of Markov processes and queueing networks. *Adv Appl Probab* 1989;21:595–611.
6. Wolff R. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
7. El-Taha M, Stidham S Jr. Sample-path analysis of queueing systems. Boston (MA): Kluwer Academic Publishing; 1999.
8. Melamed B, Whitt W. On arrivals that see time averages: a martingale approach. *J Appl Probab* 1990;27:376–384.
9. Brémaud P. Characteristics of queueing systems observed at events and the connection between stochastic intensity and Palm probability. *Queueing Syst* 1989;5:99–112.
10. König D, Schmidt V. EPSTA: the coincidence of time-stationary and customer-stationary distributions. *Queueing Syst Theory Appl* 1989;5:247–264.
11. Stidham S Jr, El-Taha M. Sample-path analysis of processes with imbedded point processes. *Queueing Syst* 1989;5:131–165.
12. Stidham S Jr. $L = \lambda W$: a discounted analogue and a new proof. *Oper Res* 1972;20:708–732.
13. Stidham S Jr. A last word on $L = \lambda W$. *Oper Res* 1974;22:417–421.
14. El-Taha M, Stidham S Jr. Sample-path analysis of stochastic discrete-event systems. *Discrete Event Dyn Syst* 1993;3:325–346.
15. Franken P, König D, Arndt U, *et al.* Queues and point processes. Chichester: Wiley; 1982.
16. Baccelli F, Brémaud P. Palm probabilities and stationary queues. Volume 41, Lecture notes in statistics. Berlin, New York: Springer; 1987.
17. Baccelli F, Brémaud P. Elements of queueing theory: palm martingale calculus and stochastic recurrences. Volume 26, Applications of Mathematics. New York: Springer; 1994.
18. Rolski T. Stationary random processes associated with point processes. Lecture notes in statistics. New York: Springer; 1981.
19. El-Taha M, Stidham S Jr. A filtered *ASTA* property. *Queueing Syst Theory Appl* 1992;11:211–222.

20. El-Taha M, Stidham S Jr. Filtration of ASTA: a weak convergence approach. *J Stat Plann Infer* 1997;100:171–183.
21. Van Doorn EA, Regterschot GJK. Conditional PASTA. *Oper Res Lett* 1988;7(5):229–232.
22. El-Taha M. On conditional ASTA: a sample-path approach. *Stoch Model* 1992;8:157–177.
23. Melamed B, Yao D. The ASTA property. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Raton (FL): CRC Press; 1995. pp. 195–224.
24. Billingsley P. *Weak convergence*. New York: John Wiley & Sons, Inc.; 1968.
25. Descoux A. On the validity of a particular subscriber's view. In: *Proceedings 5th International Traffic Conference*. New York; 1967. p. 309.
26. Strauch RE. When a queue looks the same to an arriving customer as to an observer. *Manage Sci* 1970;17:140–141.
27. Stidham S Jr. Regenerative processes in the theory of queues with applications to the alternating-priority queue. *Adv Appl Probab* 1972;4:542–577.
28. Cooper RB. *Introduction to queueing theory*. 2nd ed. New York: North Holland; 1981.
29. El-Taha M. Sample-path analysis of queueing systems: new results [PhD thesis]. School of Engineering, Graduate Program in O.R., NCSU, Raleigh; 1986.
30. König D, Miyazawa D, Schmidt V. On the identification of Poisson arrivals in queues with coinciding time-stationary and customer-stationary state distributions. *J Appl Probab* 1983;20:860–871.
31. König V, Schmidt D, Van Doorn EV. On the PASTA property and a further relationship between customer and time averages in stationary queueing systems. *Stoch Model* 1989;5:261–272.
32. König D, Schmidt V. Extended and conditional versions of the PASTA property. *Adv Appl Probab* 1990;22:510–512.
33. Miyazawa M, Wolff RW. Further results on ASTA for general stationary processes and related problems. *J Appl Probab* 1990;28:729–804.
34. Brémaud P, Kannurpatti R, Mazumdar R. Event and time averages: A review. *J Appl Probab* 1992;24:377–411.
35. Rosenkrantz W, Simha R. Some theorem on conditional PASTA: a stochastic integral approach. *Oper Res Lett* 1992;11:173–177.
36. Stidham S Jr, El-Taha M. Sample-path techniques in queueing theory. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Raton (FL): CRC Press; 1995. pp. 119–166.
37. El-Taha M, Stidham S Jr. Sample-path analysis of stochastic discrete-event systems. In: *Proceedings of the 30th IEEE CDC Meeting*; Brighton, UK; 1991. pp. 1145–1150.
38. Yao DD. On Wolff's PASTA martingale. *Oper Res* 1992;40:352–355.
39. Glynn P, Melamed B, Whitt W. Estimating customer and time averages. *Oper Res* 1993;41:400–408.
40. Makowski A, Melamed B, Whitt W. On averages seen by arrivals in discrete-time. In: *IEEE Conference on Decision and Control, Conference Proceedings, Volume 28*; 1989. Tampa (FL). pp. 1084–1086.
41. Serfozo R. *Introduction to stochastic networks*. New York: Springer; 1999.

PENALTY AND BARRIER METHODS

STEPHEN G. NASH

Systems Engineering and Operations
Research Department, George Mason
University, Fairfax, Virginia

Penalty and barrier methods solve a constrained optimization problem via a sequence of unconstrained problems. Under appropriate assumptions, the solutions of the unconstrained problems are guaranteed to converge to the solution of the original constrained problem. However, to obtain an accurate solution, the unconstrained problems will typically become ever more ill-conditioned, and special algorithmic techniques are needed to ensure good performance.

Penalty and barrier methods were important during the 1960s, but lost favor because of the difficulties associated with ill-conditioning (even though remedies were known at that time). However, interest in barrier methods revived in the 1980s in connection with the development of interior-point methods for linear programming [1]. Penalty and barrier methods are useful methods in their own right, but they also form the theoretical and practical underpinning to other important and widely used methods. In particular, penalty methods are the foundation for augmented Lagrangian methods, and barrier methods are the DNA of interior-point methods.

Penalty and barrier methods can be grouped together under the label of *penalization methods*. Penalty methods impose a penalty for violating a constraint, and barrier methods impose a penalty for approaching the boundary of an inequality constraint. Penalty methods usually approach the solution of the constrained problem via a sequence of infeasible points, and barrier methods approach the solution via a sequence of strictly feasible (interior) points. As a result, barrier techniques are typically only be applied to inequality constraints.

Penalty techniques are normally used to handle equality constraints, since standard approaches for applying penalty techniques to inequality constraints lead to unconstrained problems with discontinuous second derivatives. Penalty and barrier methods have much in common. Their convergence theories are similar, and the underlying structure of their unconstrained problems is similar.

Because of their influence on interior-point methods, barrier methods have received more attention than penalty methods. The discussion here also places more emphasis on barrier methods. Much of the theory for barrier methods can be replicated for penalty methods.

Here is an outline of the article. Sections titled “Barrier Methods,” “Penalty Methods,” and “Convergence” give an overview of barrier and penalty methods and their convergence properties. Sections titled “Ill-Conditioning” and “Stabilized Penalty and Barrier Methods” discuss the ill-conditioning associated with the methods. Sections titled “Extrapolation” and “Other Issues” describe algorithm enhancements to the methods. The section titled “Further Reading” gives suggestions for further reading.

BARRIER METHODS

Consider the nonlinear inequality-constrained problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_i(x) \geq 0, \quad i = 1, \dots, m, \quad (1) \end{aligned}$$

where the functions f and g_i are assumed to be twice continuously differentiable. We assume that the feasible set has a nonempty interior; that is, there exists some point x_0 such that $g_i(x_0) > 0$ for all i . We also assume that it is possible to reach any boundary point by approaching it from the interior.

The barrier method is defined in terms of a function $\phi(x)$ that is continuous on the interior of the feasible set and becomes

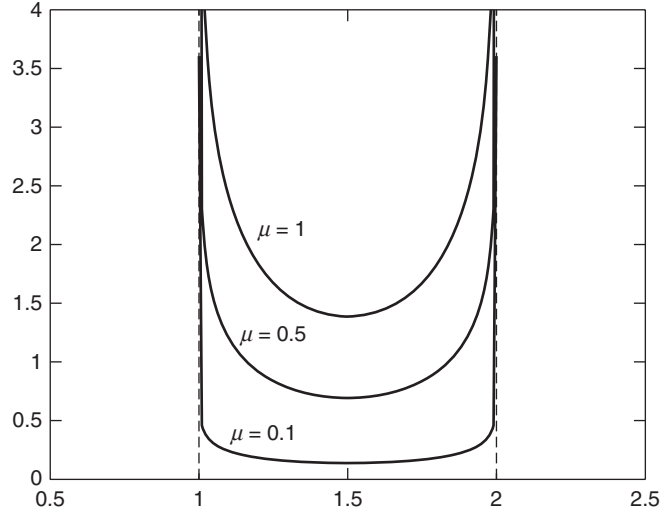


Figure 1. Effect of the barrier term.

unbounded as the boundary of the set is approached from its interior

$$\phi(x) \rightarrow \infty \quad \text{as} \quad g_i(x) \rightarrow 0_+.$$

The most common choices are the logarithmic function

$$\phi(x) = -\sum_{i=1}^m \log(g_i(x))$$

and the inverse function

$$\phi(x) = \sum_{i=1}^m \frac{1}{g_i(x)}.$$

The *barrier function* is defined as

$$\beta_\mu(x) = f(x) + \mu\phi(x),$$

where μ is a positive scalar called the *barrier parameter*. The hope is that, as $\mu \rightarrow 0$, the minimizer of the barrier function approaches the solution of the original constrained problem. Figure 1 plots the logarithmic function $\phi(x)$ for the constraints $x \geq 1$ and $x \leq 2$ for various values of μ .

In discussing barrier methods, there is an expectation that the solution of the constrained optimization problem will lie on the boundary of the feasible region. If the solution is in the strict interior of the feasible region,

then (at least in a small neighborhood of the solution) the constraints are redundant, and the optimization problem shares many of the properties of an unconstrained problem. In the discussion below, the focus is on the behavior of barrier methods for problems where at least one constraint is binding at the solution.

It is usually necessary to solve a sequence of unconstrained minimization subproblems, rather than just solving one subproblem using a small value of the barrier parameter μ . That is, we solve

$$\text{minimize } \beta_{\mu_k}(x)$$

for a sequence $\mu_k \rightarrow 0$ of positive barrier parameters. For small values of μ , the problems are difficult to solve since the barrier function will have a singularity at the boundary of the feasible region. Thus, the barrier method is started with a larger value of μ , and μ is gradually decreased. If the solution of one unconstrained subproblem is used as the starting point for the next problem, these unconstrained minimization subproblems tend to be much easier to solve. If extrapolation is used to specify the starting point for the next problem (see below), algorithmic performance is further enhanced.

Because the original constrained problem is replaced by a sequence of unconstrained

subproblems, penalty and barrier methods have been called *sequential unconstrained minimization techniques* (SUMT) [2]. A related software package was developed around 1970, which was also called SUMT [3].

Under appropriate conditions on the optimization problem (1) and the barrier term $\phi(x)$, it is possible to prove convergence for barrier methods (see the section titled “Convergence”). Further, the sequence of barrier minimizers defines a differentiable curve $x(\mu)$ called the *barrier trajectory* [2]. The trajectory exists for the logarithmic barrier function, as well as for barrier terms $\phi(x)$ that satisfy appropriate convexity and monotonicity assumptions, provided that x_* is a regular point of the constraints that satisfies the second-order sufficiency conditions as well as the strict complementarity conditions. Path-following algorithms are based on this trajectory (see also **Central Path and Barrier Algorithms for Linear Optimization**). A point is *regular* if the gradients of the constraints binding at that point are linearly independent.

Barrier methods produce Lagrange multiplier estimates. Suppose that the barrier function has the form $\phi(x) = \sum_i \theta(g_i(x))$, for either the logarithmic barrier function, $\theta(y) = -\log(y)$, or for the inverse barrier function $\theta(y) = 1/y$. Let $x = x(\mu)$ be a minimizer of the barrier function for a specific value of μ . Setting the gradient of the barrier function to zero, we obtain

$$\nabla f(x) + \sum_{i=1}^m \mu \theta'(g_i(x)) \nabla g_i(x) = 0.$$

This can be re-written as

$$\nabla f(x) - \sum_{i=1}^m \lambda_i \nabla g_i(x) = 0,$$

where $\lambda_i = \lambda_i(\mu)$ is defined by

$$\lambda_i = -\mu \theta'(g_i(x)).$$

Thus, we have a feasible point $x(\mu)$ and a vector $\lambda(\mu)$ that satisfy the following relations:

$$\begin{aligned} \nabla f(x(\mu)) - \sum_{i=1}^m \lambda_i(\mu) \nabla g_i(x(\mu)) &= 0 \\ \lambda_i(\mu) g_i(x(\mu)) &= \mu, \quad i = 1, \dots, m \\ \lambda_i(\mu) &\geq 0, \quad i = 1, \dots, m. \end{aligned}$$

This is a perturbation of the first-order necessary conditions for optimality of the constrained problem. The only difference is that the complementary slackness conditions are perturbed, so that $\lambda_i(\mu) g_i(x(\mu))$ is equal to μ rather than to zero.

If x_* is a regular point of the constraints, then as $x(\mu)$ converges to x_* , $\lambda(\mu)$ converges to λ_* . The above results show that the points on the barrier trajectory, together with their associated Lagrange multiplier estimates, are the solutions to a perturbation of the first-order optimality conditions.

Another attractive property of barrier methods is that, if we use a convex barrier term such as the logarithmic barrier function or the inverse barrier function, then the barrier function is convex if the constrained problem is a convex optimization problem defined in terms of a convex objective function and concave constraint functions.

However, barrier methods suffer from ill-conditioning as the barrier parameter decreases. In most situations, the condition number of the Hessian of the barrier function tends to infinity as $\mu \rightarrow 0$. This is discussed in the section titled “Ill-Conditioning”. The reasons for this are visible in Fig. 1. If the constrained problem has a solution on the boundary of the feasible region then the barrier function will have a minimizer near the boundary of the feasible region. As the barrier parameter goes to zero, the distance from this minimizer to the singularity will go to zero as well.

This ill-conditioning influences the way the unconstrained problems are solved. Newton-type methods should be used, since their performance is less sensitive to the condition number of the Hessian than methods such as quasi-Newton or conjugate-gradient methods (see also **Non-linear Conjugate Gradient Methods**

and *Quasi-Newton Methods*). Within the unconstrained algorithm, special-purpose techniques have been developed for use in the context of barrier methods (see the section titled “Other Issues”).

Concerns about ill-conditioning caused barrier methods to fall into disfavor in the 1970s. However, interest in these methods was renewed in the 1980s with the development of Karmarkar’s method for linear programming [4] and the subsequent discovery that this method is a form of barrier method [5]. (When a barrier method is applied to a linear program with a unique solution, ill-conditioning does not arise.) Since then, there has been a renewed interest in barrier methods for nonlinear optimization, leading to the development of numerically stable algorithms (see the section titled “Stabilized Penalty and Barrier Methods”).

If the feasible region is unbounded, the unconstrained problem may not have a minimizer. This situation can occur only if the problem is nonconvex. For the following example,

$$\begin{aligned} &\text{minimize} && \frac{-1}{x^2 + 1} \\ &\text{subject to} && x \geq 1, \end{aligned}$$

the logarithmic barrier function is unbounded below in the feasible region, although it has a local minimizer that approaches the solution $x_* = 1$ as $\mu \rightarrow 0$.

Barrier methods require that the initial guess of the solution be strictly feasible. In general problems a strictly feasible point may not be known. It is sometimes possible to find an initial point by solving an auxiliary optimization problem. This is analogous to the use of a two-phase method in linear programming.

PENALTY METHODS

Penalty methods solve a sequence of unconstrained optimization problems where a penalty is imposed for violation of the constraints. As the penalty is increased, the iterates are forced toward the feasible region.

Penalty methods do not require the iterates to be strictly feasible. Thus, they

are suitable for problems with equality constraints.

Consider first, the equality-constrained problem

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g(x) = 0, \end{aligned}$$

where $g(x)$ is an m -dimensional vector whose i th component is $g_i(x)$. We assume that all functions are twice continuously differentiable.

The penalty for constraint violation will be a continuous function ψ satisfying

$$\begin{aligned} \psi(x) &= 0, && \text{if } x \text{ is feasible,} \\ \psi(x) &> 0, && \text{otherwise.} \end{aligned} \tag{2}$$

The best-known such penalty is the *quadratic-loss function*

$$\psi(x) = \frac{1}{2} \sum_{i=1}^m g_i(x)^2 = \frac{1}{2} g(x)^T g(x),$$

based on the 2-norm. Penalties can also be based on other norms, for example, exact penalty functions. The weight of the penalty is controlled by a positive *penalty parameter* ρ .

The *penalty function* is defined as

$$\pi_\rho(x) = f(x) + \rho\psi(x).$$

The penalty method solves a sequence of unconstrained minimization subproblems of the form

$$\text{minimize } \pi_{\rho_k}(x)$$

for an increasing sequence $\{\rho_k\}$ of positive values tending to infinity.

Penalty methods have many properties in common with barrier methods. First, under mild conditions, it is possible to guarantee convergence. Also, under appropriate conditions, the sequence of penalty function minimizers defines a continuous trajectory. In the latter case, it is possible to get estimates of the Lagrange multipliers in the solution. For

example, consider the quadratic-loss penalty function

$$\pi_\rho(x) = f(x) + \frac{1}{2}\rho \sum_{i=1}^m g_i(x)^2.$$

Its minimizer $x(\rho)$ satisfies

$$\begin{aligned} \nabla_x \pi_\rho(x(\rho)) \\ = \nabla f(x(\rho)) + \rho \sum_{i=1}^m \nabla g_i(x(\rho)) g_i(x(\rho)) = 0. \end{aligned}$$

Defining $\lambda_i(\rho) = -\rho g_i(x(\rho))$, we obtain that

$$\nabla f(x(\rho)) - \sum_{i=1}^m \lambda_i(\rho) \nabla g_i(x(\rho)) = 0.$$

If $x(\rho)$ converges to a solution x_* , which is a regular point of the constraints, then $\lambda(\rho)$ converges to the Lagrange multiplier λ_* associated with x_* [2].

Penalty functions also suffer from ill-conditioning. As the penalty parameter increases, the condition number of the Hessian matrix of $\pi_\rho(x(\rho))$ increases, tending to ∞ as $\rho \rightarrow \infty$. Therefore, the unconstrained minimization problems can become increasingly difficult to solve.

Penalty methods can be applied to problems with inequality constraints. Consider the problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_i(x) \geq 0, \quad i = 1, \dots, m. \end{aligned}$$

In this case, it is possible to use the penalty

$$\psi(x) = \frac{1}{2} \sum_{i=1}^m [\min(g_i(x), 0)]^2.$$

This function has continuous first derivatives

$$\nabla \psi(x) = \sum_{i=1}^m [\min(g_i(x), 0)] \nabla g_i(x),$$

but its second derivatives can be discontinuous at points $\bar{x}$ where $g_i(\bar{x}) = 0$ for some i . Thus, one must exercise care when using Newton's method to minimize the function.

CONVERGENCE

We focus on the convergence of barrier methods when applied to the inequality-constrained problem. Convergence results for penalty methods can be developed in a similar manner [2].

Consider the problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in S, \end{aligned}$$

where

$$S = \{x : g_i(x) \geq 0, \quad i = 1, \dots, m\}.$$

Denote the interior of the feasible region by

$$S^0 = \{x : g_i(x) > 0, \quad i = 1, \dots, m\}.$$

The following theorem assumes that:

- (i) The functions f and $g_1, \dots, g_m$ are continuous on $\mathbb{R}^n$.
- (ii) The set $\{x : x \in S, f(x) \leq \alpha\}$ is bounded for any finite α .
- (iii) The set S^0 is nonempty.
- (iv) S is the closure of S^0 .

Assumptions (i) and (ii) imply that the function f has a minimum value f_* on the set S . Assumption (iii) is necessary to define the barrier subproblems. Assumption (iv) is necessary to avoid situations where the minimum point is isolated and does not have neighboring interior points. The barrier function is defined in terms of a continuous function ϕ that satisfies

$$\phi(x) \rightarrow \infty \quad \text{as } g_i(x) \rightarrow 0_+.$$

A proof of the theorem can be found in Ref. 6.

Theorem 1. *Suppose that a nonlinear inequality-constrained problem satisfies conditions (i)–(iv) above. Suppose that a sequence of unconstrained minimization problems*

$$\text{minimize } \beta_\mu(x) = f(x) + \mu\phi(x)$$

is solved for μ taking values $\mu_1 > \mu_2 > \dots > \mu_k > \dots$, where $\lim_{k \rightarrow \infty} \mu_k = 0$. Suppose also that the functions $\beta_{\mu_k}(x)$ have a minimum in S^0 for each k . Let x_k denote a global minimizer of $\beta_{\mu_k}(x)$. Then

- (i) $f(x_{k+1}) \leq f(x_k)$,
- (ii) $\phi(x_{k+1}) \geq \phi(x_k)$,
- (iii) the sequence x_k has a convergent subsequence,
- (iv) if $\{x_k : k \in \mathcal{K}\}$ is any convergent subsequence of unconstrained minimizers of β , then its limit point is a global solution of the constrained problem.

The theorem assumes that the barrier functions have a minimizer. This will be true if the set S is compact, or if the problem is convex [2].

It is also possible to prove convergence to a local minimum of the constrained problem if additional assumptions are made [2]. It is not true that every limit point of a sequence of local minimizers of the barrier function is a constrained minimizer [7].

ILL-CONDITIONING

In most cases, the barrier function becomes increasingly ill-conditioned as the barrier parameter μ decreases [8,9]. The analysis below is in terms of the logarithmic barrier function

$$\text{minimize } \beta_\mu(x) = f(x) - \mu \sum_{i=1}^m \log(g_i(x)),$$

but the result is true, in general, for other barrier and penalty functions. The results in the section titled “Stabilized Penalty and Barrier Methods” show how to overcome this ill-conditioning by deriving numerically stable formulas for the Newton direction in a barrier method.

The ill-conditioning can be observed by examining the formulas for the Hessian matrix of β :

$$\nabla_{xx}^2 \beta = \nabla^2 f - \mu \sum_{i=1}^m \frac{\nabla^2 g_i}{g_i} + \mu \sum_{i=1}^m \frac{\nabla g_i \nabla g_i^T}{g_i^2},$$

where we have written f in place of $f(x)$, and so on. Let $x(\mu)$ be an unconstrained minimizer of $\beta_\mu(x)$, and let $\lambda_i = \mu/g_i(x(\mu))$ be the associated Lagrange multiplier estimate. Then,

$$\nabla_{xx}^2 \beta = \nabla^2 f - \sum_{i=1}^m \lambda_i \nabla^2 g_i + \frac{1}{\mu} \sum_{i=1}^m \lambda_i^2 \nabla g_i \nabla g_i^T.$$

The term $\nabla^2 f - \sum_{i=1}^m \lambda_i \nabla^2 g_i$ is an approximation to the Hessian matrix of the Lagrangian of the problem. Its condition number reflects the conditioning of the Lagrangian Hessian for the original constrained problem.

The final term is the source of the ill-conditioning. If a constraint is binding at the solution of the constrained problem (and its corresponding Lagrange multiplier is not zero), then the ratio λ_i^2/μ tends to infinity as μ approaches zero. As a result, the Hessian matrix in general becomes increasingly ill-conditioned as the solution is approached. This structural ill-conditioning occurs even when the underlying constrained problem is well conditioned.

If a barrier method is applied to a linear programming problem, then the objective function and the constraint functions are all linear. Hence the Hessian of the barrier function simplifies to

$$\nabla_{xx}^2 \beta = \frac{1}{\mu} \sum_{i=1}^m \lambda_i^2 \nabla g_i \nabla g_i^T.$$

This matrix only involves the constraints in the linear program, and its conditioning corresponds to that of the underlying linear program. There is no ill-conditioning induced by the barrier method.

STABILIZED PENALTY AND BARRIER METHODS

In a number of settings, it is possible to implement a penalization method in such a way so as to avoid the effects of ill-conditioning. One approach is to use “stabilized” formulas for the search direction in a Newton-type method [9,10]. For example, consider using a

logarithmic barrier method to solve

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } g(x) \geq 0. \end{aligned}$$

(The results can be extended to other penalty and barrier methods.) The ill-conditioning is avoided by using an approximate formula for the Newton direction for the barrier problem. This formula becomes more accurate as the barrier parameter goes to zero, that is, as the ill-conditioning becomes more severe.

For simplicity, assume that all the constraints are binding at the solution, and that all of these constraints have positive Lagrange multipliers. In practice, the techniques here are applied only to the binding constraints. The approach separates the Newton direction into two components, one in the null space and one in the range space of the constraint gradients. The approximations make it possible to compute both these components in a stable manner.

Let $A = \nabla g(x)^T$ be the Jacobian matrix of the constraints, and assume that A has full rank. Let Z be a basis matrix for the null space of A , where Z is obtained from an orthogonal QR factorization, so that

$$(Z \ A^T) \begin{pmatrix} Z^T \\ A_r^T \end{pmatrix} = I.$$

In addition, define the Lagrange multiplier estimates $\lambda_i = \mu/g_i(x)$ and the diagonal matrix D , whose i th diagonal entry is λ_i^2 . Finally, let $B = \nabla_{xx}^2 f$ be the Hessian of the barrier function, and let H be the Hessian matrix of the Lagrangian:

$$H = \nabla^2 f - \sum_{i=1}^m \lambda_i \nabla^2 g_i.$$

Then for small values of μ , it is possible to derive an approximation to the Newton direction p :

$$p = B^{-1} \nabla_x \beta \approx p_1 + \mu p_2,$$

where

$$\begin{aligned} p_1 &= -Z(Z^T H Z)^{-1} Z^T \nabla_x \beta, \\ \lambda &= A_r^T (H p_1 + \nabla_x \beta), \\ p_2 &= (Z(Z^T H Z)^{-1} Z^T H - I) A_r D^{-1} \lambda. \end{aligned}$$

The approximation is based on the inverse of $Z^T H Z$. This is a projection of the Hessian of the Lagrangian, which has no structural ill-condition, that is, no ill-conditioning due to the presence of the barrier parameter.

These formulas are expressed in terms of the inverse of $Z^T H Z$. This inverse matrix should not be formed. Instead a factorization should be computed. This factorization need only be computed once; back-substitution can be used to compute p_1 and p_2 . As $\mu \rightarrow 0$, it can be shown that the error in the approximate search direction is $O(\mu)$ [10].

If $Z^T H Z$ is positive definite at the point x , then these formulas produce a descent direction if the barrier parameter μ is sufficiently small. If $Z^T H Z$ is not positive definite, it is still possible to compute a descent direction by modifying the Hessian matrix using techniques adapted from Refs 11 and 12.

When solving an optimization problem with a large number of variables, it is often advantageous to use an iterative method (such as the linear conjugate-gradient method) to solve the systems of equations associated with $(Z^T H Z)^{-1}$. Consequently, these formulas require two runs of the iterative method. A related set of formulas can be derived that requires only one use of the iterative method and that is more suitable in this case [10].

Related results have been developed for more general interior-point methods. For example, if an interior-point method is developed by applying logarithmic barrier terms to bound constraints on the slack variables and the Lagrange multipliers, and if direct matrix factorizations are used, then the computations can be performed in a numerically stable manner [13,14].

EXTRAPOLATION

Extrapolation can be an important component of a penalization method. The discussion

here is based on using the logarithmic barrier method, but related results are valid for other penalization methods.

Given minimizers $x(\mu_k)$, $x(\mu_{k-1})$, $\dots$, of the barrier function for successive values of the barrier parameter, extrapolation approximates these minimizers using polynomial approximations parameterized by μ . For example, given $x(\mu_k)$, $x(\mu_{k-1})$, and $x(\mu_{k-2})$, a quadratic approximation is used. Then the initial guess $x_0(\mu_{k+1})$ for the new value of the barrier parameter is obtained from the polynomial approximation. If we use a linear approximation based on two successive values of μ then

$$x_0(\mu_{k+1}) = x(\mu_k) + \frac{\mu_{k+1} - \mu_k}{\mu_k - \mu_{k-1}} (x(\mu_k) - x(\mu_{k-1})).$$

Formulas for quadratic and cubic extrapolation are only slightly more elaborate.

Extrapolation is an inexpensive enhancement to a penalization method, but it can produce major improvements in the performance of the method [15].

The results here are based on the following assumptions:

- the third derivatives of the objective and constraint functions are continuous;
- the solution x_* of the constrained problem is a regular point;
- the reduced Hessian at x_* is positive definite; and
- strict complementarity holds.

Under these assumptions, x_* is an isolated local minimizer, and the corresponding Lagrange multipliers are unique [2]. Also, under these conditions, there exists an isolated twice-differentiable trajectory $x(\mu)$ converging to x_* , and the multiplier estimates converge to the Lagrange multipliers at the solution.

If linear extrapolation is used then it can be shown that:

- the initial guess $x_0(\mu_{k+1})$ is likely to be close to the solution of the new subproblem;

- the initial guess is likely to be feasible for the new subproblem; and
- the Newton step at the initial guess is likely to be feasible.

In contrast, if extrapolation is not used it can be shown that the gradient at this initial guess is not likely to be close to zero, and hence the initial guess is not close to the solution of the new subproblem [15]. Further, the Newton step at the initial guess is likely to be infeasible [16].

Related results on the barrier trajectory and extrapolation can be obtained with fewer assumptions than are made here [17,18]. It is also possible to apply extrapolation to penalty methods [19].

In degenerate cases, it may be appropriate to use alternative extrapolation formulas [20].

OTHER ISSUES

Penalization methods are based on solving a sequence of unconstrained problems. Because of the ill-conditioning, it is advisable to use a Newton-type method. In the context of penalization methods, it has been traditional to use a line search to ensure convergence of the unconstrained method. Thus, at each iteration of the unconstrained method the new estimate of the solution has been obtained via

$$x_{k+1} = x_k + \alpha p,$$

where $\alpha > 0$ is a scalar chosen to ensure that x_{k+1} is an improved estimate of the solution, and p is a descent direction for the penalty or barrier function at the point x_k .

Standard line-search algorithms are based on polynomial approximations. These are not ideal in the context of a penalization method, since the penalty or barrier term introduces a singularity or near singularity. For this reason, specialized line-search algorithms have been developed for use in this setting [21]. For a barrier method, this involves using a log-quadratic approximation in the line search:

$$\beta_\mu(x_k + \alpha p) \approx t_1 + t_2 \alpha + t_3 \alpha^2 - \mu \log(t_4 - \alpha),$$

where the coefficients $\{t_i\}$ are chosen to interpolate β and $\nabla\beta$ at two trial points. In addition, safeguards must be used to ensure that the trial points are strictly interior to the feasible region.

The performance of a barrier method can be improved by proper choice of the sequence of barrier parameters. The initial barrier parameter should not be too small (making the first unconstrained problem too hard to solve) nor too large (resulting in a large number of unconstrained problems). Also, if the barrier parameter is reduced too quickly or too slowly, the overall method will be less efficient. Techniques for initializing and adjusting the barrier parameter are discussed in Refs 10 and 22.

It is possible to prove complexity results for logarithmic barrier methods applied to convex optimization problems [23–25]. These results provide a bound on the number of iterations required to solve a barrier subproblem to within a tolerance $\epsilon = 2^{-M}$ in a number of operations that is polynomial in M and the problem dimensions. If polynomial algorithms exist to evaluate the gradient and Hessian, the overall algorithm for the barrier subproblem is polynomial in the dimensions of the optimization problem.

In general, barrier methods and penalty methods compute a sequence of unconstrained optima that approximate, but never equal, the solution of the underlying constrained problem. *Exact penalty methods* make it possible to solve the constrained problem *exactly* for a finite positive penalty parameter (see **Methods for Large-Scale Unconstrained Optimization**).

The ill-conditioning of penalty methods can be ameliorated by including multipliers explicitly in the penalty function. This idea is the basis of *augmented Lagrangian methods*. *Modified barrier methods* are an analogous extension of barrier methods [26].

Nonlinear primal-dual methods are also related to penalization methods. As discussed earlier, penalization methods produce Lagrange multiplier estimates that satisfy a perturbation of the optimality conditions for the constrained problem. In particular, they correspond to a perturbed set of complementary slackness conditions.

Primal-dual methods work explicitly with both primal and dual variables, and compute new estimates of the solution based on the perturbed optimality conditions. These methods were developed as an out-growth of interior-point (i.e., barrier) methods.

FURTHER READING

The most important reference on penalty and barrier methods is the book by Fiacco and McCormick (1968, reprinted 1990) [2]. It includes an extensive survey on the methods. A more recent discussion can be found in the paper by Wright [7]. For a history of these methods, see the report by Fiacco [27]. A general discussion of penalty and barrier methods, including examples and related methods, is in the book by Griva *et al.* [6].

REFERENCES

1. Breitfeld MG, Shanno DF. Computational experience with penalty/barrier methods for nonlinear programming. *Ann Oper Res* 1996;62:439–464.
2. Fiacco AV, McCormick GP. *Nonlinear programming: sequential unconstrained minimization techniques*. New York: John Wiley & Sons, Inc.; 1968. Reprinted by SIAM Publications, 1987.
3. Mylander WC, Holmes RL, McCormick GP. *A guide to SUMT—version 4*. Report RAC-P-63. McLean (VA): Research Analysis Corporation; 1974.
4. Karmarkar N. A new polynomial-time algorithm for linear programming. *Combinatorica* 1984;4:373–395.
5. Gill PE, Murray W, Saunders MA, *et al.* On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method. *Math Program* 1986;36:183–209.
6. Griva I, Nash SG, Sofer A. *Linear and nonlinear optimization*, Philadelphia (PA): SIAM; 2008.
7. Wright MH. Interior methods for constrained optimization. *Acta Numer* 1992;1:341–407.
8. Lootsma FA. Hessian matrices of penalty functions for solving constrained optimization problems. *Philips Res Rep* 1969;24:322–331.
9. Murray W. Analytical expressions for the eigenvalues and eigenvectors of the Hessian

- matrices of barrier and penalty functions. *J Optim Theory Appl* 1971;7:189–196.
10. Nash SG, Sofer A. A barrier method for large-scale constrained optimization. *ORSA J Comput* 1993;5:40–53.
 11. Gill PE, Murray W. Newton-type methods for unconstrained and linearly constrained optimization. *Math Program* 1974;28:311–350.
 12. Schnabel RB, Eskow E. A new modified Cholesky factorization. *SIAM J Sci Stat Comput* 1990;11:1136–1158.
 13. Forsgren A, Gill PE, Shinnerl JR. Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization. *SIAM J Matrix Anal Appl* 1996;17:187–211.
 14. Wright SG. Effects of finite-precision arithmetic on interior-point methods for nonlinear programming. *SIAM J Optim* 2001;12:36–78.
 15. Nash SG, Sofer A. Why extrapolation helps barrier methods; 1998. Available at <http://iris.gmu.edu/snash/papers/sofer13.ps>.
 16. Wright MH. Why a pure primal Newton barrier step may be infeasible. *SIAM J Optim* 1995;5:1–12.
 17. El Afia A, Benchakroun A, Dussault J-P. Asymptotic analysis of the trajectories of the logarithmic barrier algorithm without differentiable objective function. *Int J Pure Appl Math* 2003;9(1):99–123.
 18. El Afia A, Benchakroun A, Dussault J-P, *et al.* Asymptotic analysis of the trajectories of the logarithmic barrier algorithm without constraint qualifications. *RAIRO Oper Res* 2008;42:157–198.
 19. Dussault J-P. High-order Newton-penalty algorithms. *J Comput Appl Math* 2005;182(1): 117–133.
 20. Jittorntrum K, Osborne MR. A modified barrier function method with improved rate of convergence for degenerate problems. *J Aust Math Soc, Ser B* 1980;21:305–329.
 21. Murray W, Wright MH. Line search procedures for the logarithmic barrier function. *SIAM J Optim* 1994;4:229–246.
 22. Dussault J-P, El Afia A. On the convergence rate of the logarithmic barrier algorithm. *Comput Optim Appl* 2001;19(1):31–54.
 23. den Hertog D. Interior point approach to linear, quadratic and convex programming. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1994.
 24. Nesterov Y, Nemirovskii AS. Interior-point polynomial algorithms in convex programming. Philadelphia (PA): SIAM; 1994.
 25. Nash SG, Sofer A. On the complexity of a practical interior-point method. *SIAM J Optim* 1998;8:833–849.
 26. Polyak R. Modified barrier functions (theory and methods). *Math Program* 1992;54: 177–222.
 27. Fiacco AV. Historical survey of sequential unconstrained methods for solving constrained minimization problems. Technical Paper RAC–TP–267. McLean (VA): Research Analysis Corporation; 1967.

PERCOLATION THEORY

JOHN WIERMAN

Department of Applied Mathematics and
Statistics, Whiting School of Engineering,
Johns Hopkins University, Baltimore,
Maryland

In physics, chemistry, biology, and materials science, percolation concerns the movement of fluid or clustering of particles in random media, when the nature of the medium has a substantial effect on the possible behaviors. The mathematical model of percolation, which involves randomly distributed connections in a graph, which represents the medium, is an alternative to diffusion models. Percolation theory has brought understanding to a wide range of scientific phenomena, and provides insight into more complicated models exhibiting phase transitions and critical phenomena. The study of percolation theory uses methods and techniques from probability, stochastic processes, and discrete mathematics. The reader may consult Grimmett [1] and Kesten [2] for the probabilistic perspective, Bollobás and Riordan [3] for the discrete mathematics perspective, and Sahimi [4], Hughes [5], Stauffer and Aharony [6], and Deutscher *et al.* [7] for scientists' perspectives and applications.

BOND AND SITE PERCOLATION MODELS

In a *bond percolation* model, each edge of an infinite graph G is declared to be open with probability p , $0 < p < 1$, independently, and closed otherwise. Denote the corresponding probability measure by P_p and the expectation with respect to P_p by E_p . The cluster C_v containing a vertex v of G is the set of vertices that can be reached from v through a path of open edges. The percolation probability function is $\theta_v(p) = P_p[|C_v| = \infty]$, where $|\cdot|$ denotes cardinality, represents the probability that v is in an infinite cluster. $\theta_v(p)$ is a nondecreasing and right-continuous function

for every v . If G is a connected graph, for any vertices v and w in G , $\theta_v(p) = 0$ if and only if $\theta_w(p) = 0$. The *critical probability* or *percolation threshold* of a connected graph G is defined by $p_c(G) = \sup\{p : \theta_v(p) = 0\}$, which is independent of the choice of v . Note that, by Kolmogorov's zero-one law, when $p > p_c$, there is an infinite open cluster in the lattice with probability one. The percolation threshold thus corresponds to a phase transition, such that the global behavior of the system is substantially different in the regions $p < p_c$ and $p > p_c$ of the parameter space. The model is said to be in the subcritical, critical, or supercritical phase if $p < p_c$, $p = p_c$, or $p > p_c$, respectively.

A *site percolation* model is defined similarly, with each vertex being open with probability p , with an edge declared to be open if both end points are open. Clusters, percolation probability function, and percolation threshold are defined analogously.

PERCOLATION THRESHOLD RELATIONSHIPS

Two simple relationships establish inequalities between percolation thresholds. If H is a subgraph of G , then $p_c(H) \geq p_c(G)$, for both site and bond models. If H is obtained by contracting a set of edges in G , then $p_c(H) \leq p_c(G)$ for bond models. Under broad conditions, these are strict inequalities, by a result of Aizenman and Grimmett [8] regarding the effect of enhancing one model by systematically adding connections based on local information.

The bond percolation model on graph G may be transformed into an equivalent site percolation model on the *line graph* of G , denoted by $L(G)$. Thus, $p_c^{\text{bond}}(G) = p_c^{\text{site}}(L(G))$. The vertices of $L(G)$ are the edges of G , with edges of $L(G)$ between two vertices if their corresponding edges have a common end-vertex. However, not every site model can be obtained in this manner, so the site percolation model is more general than the bond percolation model. Another relationship

between bond and site models is that, for any G , $p_c^{\text{bond}}(G) \leq p_c^{\text{site}}(G)$ [9].

There exist two important alternative definitions of the percolation threshold. A threshold $p_T(G)$ may be defined in terms of the expected cluster size: $p_T(G) = \sup\{p : E_p[|C|] < \infty\}$. Clearly, $p_T(G) \leq p_c(G)$ for any connected graph G . A third definition, p_S , in terms of probabilities of crossings of rectangular regions by open paths, was crucial to establishing exact percolation threshold values.

An infinite graph G is periodic if it can be embedded in $\mathbf{R}^d$ with $d \geq 2$, so that the graph is invariant under translations by d linearly independent vectors and the vertex set has no accumulation points. If so, we say that G has dimension d . Menshikov [10] and Aizenman and Barsky [11] independently proved the exponential decay of the probability distribution of $|C_v|$ when $p < p_c$, so that $p_T = p_c$ for periodic graphs in any dimension.

PERCOLATION THRESHOLD VALUES

The only exact percolation threshold solutions are for infinite trees and some two-dimensional periodic graphs. It is a major open problem to determine the exact percolation thresholds for additional graphs, and particularly for any lattice in three or higher dimensions.

Exact Values

The first exact solutions were for regular trees of degree d , obtaining the value $p_c = 1/(d-1)$ for both bond and site models by using branching process methods. Later, Lyons [12] characterized the percolation threshold of an arbitrary infinite tree.

Predictions of exact percolation thresholds were derived for the square, triangular, and hexagonal bond models and the triangular site model by an insightful heuristic approach by Sykes and Essam [13]. Their predictions were later confirmed, showing $p_c = 1/2$ for the square lattice bond model [64], $p_c = 2 \sin(\pi/18) \approx 0.347296$ for the triangular lattice bond model and $p_c = 1 - 2 \sin(\pi/18) \approx 0.652704$ for the

hexagonal lattice bond model [14], and $p_c = 1/2$ for the triangular lattice site model and the site model on any full-triangulated periodic lattice [2]. The bond thresholds of the bow tie lattice and its dual graph are the solutions $p_0 \in (0, 1)$ of the polynomial equation $1 - p - 6p^2 + 6p^3 - p^5 = 0$ and $1 - p_0$, respectively, where $p_0 \approx 0.404518$ [15]. Recent research has produced an approach for constructing an infinite class of graphs for which bond percolation thresholds may be exactly determined [16–18].

The solutions for bond models depend crucially on graph duality. Since dual graphs exist only for planar graphs, there are currently solutions only for two-dimensional lattices. Similarly, site model solutions rely on a matching property, which corresponds to duality for line graphs, and is valid only for two-dimensional lattices. Exact solution proofs also have relied heavily on symmetry properties of the graph. Evidently, symmetry conditions on the model are no longer necessary, due to results of Sheffield [19].

Since exact solutions do not exist for bond or site percolation on most lattice graphs of interest to physical scientists, considerable research has been focused on rigorous bounds, estimates, and approximations.

Bounds

Rigorous bounds for the threshold of a lattice graph are typically relatively far apart. For two of the most studied unsolved graphs, we have $0.556 < p_c < 0.679492$ for the square lattice site model [20,21] and $0.652703 < p_c < 0.743359$ for the hexagonal lattice site model [22], in which case the lower bound has not been improved since 1981. The substitution method [14,23] computes percolation threshold bounds for an unsolved lattice by comparing it to a solved lattice, comparing probability distributions of connections by stochastic ordering, obtaining $0.522197 < p_c < 0.526873$ for the kagomé lattice bond model and $0.739773 < p_c < 0.741125$ for the Archimedean $(3, 12^2)$ lattice [22].

Estimates

Over the years, many nonrigorous numerical estimations of percolation thresholds have been generated by simulating configurations on finite portions of the lattice, then extrapolating results from these to obtain the estimate. A variety of simulation approaches and algorithms for extrapolation are used, leading to some disagreement regarding error bounds. One established efficient method is due to Newman and Ziff [24,25], which produced an estimate of $0.59274621 \pm 0.00000013$ for the square lattice site percolation threshold. Riordan and Walters [26] developed a method for computing narrow confidence interval estimates with extremely high confidence levels, such as error probabilities of one per million, which does not use any extrapolation.

Approximation formulas [27,28], based on a small number of features of the graph, such as average degree and dimension, also provide estimates for percolation thresholds.

Critical Surfaces

In inhomogeneous percolation models, different directions or vertices in different locations have different probabilities of being open. For such models, there is a critical surface in the parameter space which corresponds to the percolation threshold in the homogeneous model. Only a few mathematically rigorous results have been established. For the square lattice, with each horizontal edge open with probability p_h and each vertical edge open with probability p_v , the critical surface is $p_h + p_v = 1$. The triangular lattice, with edges in the three directions open with probability parameters p_h , p_v , and p_d , the critical surface is $p_h + p_v + p_d - p_h p_v p_d = 1$. [1, Section 11.9].

CONTINUITY

The percolation probability function $\theta_v(p)$ can be successively approximated by the probability that v is connected to some vertex at distance n by an open path. Each such probability is a polynomial in p , so, being the

limit of nondecreasing continuous functions, $\theta_v(p)$ is right continuous. Thus, it is continuous at p_c if and only if the probability of an infinite open cluster is zero at p_c . With probability one, there are no infinite clusters at the percolation threshold for regular trees or for the periodic lattices in two dimensions, and this result also holds for periodic lattices in dimensions $d \geq 19$. However, for aperiodic graphs, there exist examples where there is an infinite cluster containing a vertex v at p_c with probability one.

UNIQUENESS

From reasoning based on Kolmogorov's zero-one law, for any value of p there may be zero, one, or infinitely many infinite open clusters in a graph. For $p > p_c$ there may be infinitely many infinite clusters, while for two-dimensional periodic graphs there is a unique infinite cluster. In fact, there exists a threshold $p_u \geq p_c$ such that for $p > p_u$ there is a unique infinite cluster [29,30]. For two-dimensional periodic lattices, $p_c = p_u$, but for trees $p_u = 1$, and for some other graphs $p_c < p_u < 1$. The concept of amenability is crucial for the study of uniqueness in percolation. The edge-isoperimetric constant of a graph G with vertex set V is $\kappa(G) = \inf_W \{|\partial W|/|W|\}$, where the infimum ranges over all finite connected subsets W of V and ∂W is the set of edges with one end point in W and one in $V \setminus W$. An infinite graph G is amenable if $\kappa(G) = 0$ and nonamenable otherwise. Partial results suggest the conjecture that if G is an infinite connected vertex-transitive graph, then $p_u > p_c$ if and only if G is amenable [31].

UNIVERSALITY AND CRITICAL EXPONENTS

It is generally believed that, although the value of the percolation threshold is related to the local structure of the graph, the nature and behavior of open clusters below, at, and above the threshold are invariant with respect to the local structure. The expected behaviors of various percolation functions are described in terms of power laws with the exponents referred to as *critical exponents*. The *universality principle* is that the critical

exponents depends only on the dimension of the lattice, and are independent of the detailed structure of the graph.

To describe the power law behavior, the notation $f(p) \approx |p - p_c|^\lambda$ means $\lim_{p \rightarrow p_c} \frac{\log f(p)}{\log |p - p_c|} = \lambda$ as $p \rightarrow p_c$. Similar notation may be used for functions of n as $n \rightarrow \infty$. The lattices are assumed to be periodic, with dimension d . The most common percolation functions, and the expected form of their power laws, are:

The percolation probability: $\theta(p) = P_p[|C| = \infty] \approx (p - p_c)^\beta$.

The mean finite cluster size: $\chi^f(p) = E_p[|C|, |C| < \infty] \approx |p - p_c|^{-\gamma}$.

The mean number of clusters per vertex: $\kappa(p) = E_p[|C|^{-1}]$, with power law $\kappa'''(p) \approx |p - p_c|^{-1-\alpha}$.

The cluster moments: $\chi_k^f(p) = E_p[|C|^k, |C| < \infty]$, with power laws

$$\frac{\chi_{k+1}^f(p)}{\chi_k^f(p)} \approx |p - p_c|^{-\Delta}.$$

The cluster volume: $P_{p_c}[|C| = n] \approx n^{-1-1/\delta}$.

The cluster radius: $P_{p_c}[\text{rad}(C) = n] \approx n^{-1-1/\rho}$.

The pair connectivity function:

$$P_{p_c}[0 \rightarrow x] \approx |x|^{2-d-\eta}.$$

The correlation length has several different definitions. Loosely, it represents the minimal distance beyond which the open clusters of two vertices are essentially uncorrelated. The notation and power law are $\xi(p) \approx |p - p_c|^{-\nu}$.

In almost all cases, it has not been proved that the limits defining critical exponents exist. For all these percolation functions, there are proofs of power bounds, instead of actual power laws, for many periodic graphs with $d = 2$ or sufficiently large d .

Based on heuristic derivations and simulation evidence, a number of *scaling relations* between the critical exponents are expected to be satisfied for graphs in any dimension: $2 - \alpha = \gamma + 2\beta = \beta(\delta + 1)$, $\Delta = \delta\beta$, and $\gamma = \nu(2 - \eta)$. In addition, in

dimensions $d = 2, \dots, 6$ there are anticipated *hyperscaling relations*: $d\rho = \delta + 1$ and $d\nu = 2 - \alpha$. Under the assumption that the limits defining the critical exponents exist, the scaling and hyperscaling relations (except that involving α) have been proved for a class of models in $d = 2$ [32] and the scaling relations for sufficiently high dimensions [33].

Physicists have predictions of the values of critical exponents when $d = 2$ or d is large [34–36]. Conjectured values for $d = 2$ are $\alpha = -\frac{2}{5}$, $\beta = \frac{5}{36}$, $\gamma = \frac{43}{18}$, $\delta = \frac{91}{5}$, and $\nu = \frac{4}{3}$. These values have been rigorously verified only for the triangular lattice site model [37]. It is widely believed that the critical exponents for lattices with dimension $d \geq 6$ have the corresponding values for a regular tree, which are $\alpha = -1$, $\beta = 1$, $\gamma = 1$, $\delta = 2$, $\rho = \frac{1}{2}$, $\eta = 0$, and $\nu = \frac{1}{2}$. The *lace expansion* has been extensively developed by Hara and Slade [33], and has been valuable for understanding percolation in higher dimensions. They proved many of the physicists' conjectures for bond percolation on the hypercubic lattices $\mathbb{Z}^d$ with $d \geq 19$. However, they can prove their results in any dimension $d > 6$ for *spread-out* models, which have edges between vertices, which are not the nearest neighbors in the lattice.

CONFORMAL INVARIANCE AND SCALING LIMIT

Percolation models on two-dimensional lattices are conjectured to have conformal invariance, in the sense that, if the lattice is rescaled so the spacing tends to zero, then certain limiting probabilities of crossing of regions are invariant under conformal mappings of the plane when the model is at criticality. Conformal invariance has been proved for the triangular site percolation model [38].

Stochastic Loewner Evolution [39–42], or Schramm–Loewner evolution, is a stochastic process of random planar curves generated by a differential equation with Brownian motion as the input. It is conjectured to be the scaling limit of various critical percolation models and other stochastic processes in the

plane, such as uniform spanning trees and planar loop-erased random walk. A parameter κ determines the distribution of the process, with $\kappa = 6$ believed to correspond to the scaling limit of the boundary of percolation clusters at criticality in two-dimensional models. This connection would yield proofs of the conjectured values for critical exponents in such cases, and has been proven to be correct for site percolation on the triangular lattice.

EXTENSIONS AND GENERALIZATIONS

A plethora of extensions of the classical percolation models have been developed, which allow a wide variety of phenomena to be modeled. We briefly introduce a few of the major alternative models to suggest the range of applications and behaviors.

Mixed Percolation

Mixed percolation models allow both vertices and edges, and sometimes faces also, to be open or closed at random. The most commonly treated case is “site-bond” percolation, in which both edges and vertices are randomized.

Directed Percolation

Directed percolation, sometimes called *oriented percolation*, randomly retains edges of a directed graph, usually to model the movement of a fluid in a random medium under the influence of a force such as gravity. A survey of methods and results is given in Ref. 43.

AB Percolation

The AB percolation model, also called *antipercolation*, was introduced as a model for gelation. Each vertex is labeled A with probability p and B otherwise, representing two possible chemicals or species that could occupy a site. An edge connects two vertices if they have opposite labels. Wierman and Appel [44] and Appel and Wierman [45] proved results concerning the existence and nonexistence of infinite AB percolation clusters on certain lattices.

Invasion Percolation

Invasion percolation is a dynamic process which models a fluid being pumped into a random medium under pressure, displacing other materials with least resistance. Each edge of the graph is independently assigned a random variable that is uniformly distributed on $(0, 1)$. A sequence of connected clusters is created by starting from one vertex and successively adding the edge on the boundary of the cluster which has the smallest value. For any $p > p_c$, the set of edges with a value less than p will contain an infinite component of the graph. Once the process reaches this component, it will never invade another edge with value greater than p . The principal result is that the empirical distribution function of the values of edges in the cluster converges to the uniform distribution on $(0, p_c(G))$ [46]. The version described above is called *nontrapping*, while a *trapping* model does not allow fluid to invade edges in regions that are already surrounded.

Continuum Percolation

It is important to model phenomena, which do not occur in a lattice structure, and a variety of models of continuum percolation have been introduced to do so. Common approaches are to place random objects, such as discs or balls in the space, or construct a random irregular graph from a set of random vertices located in the space. Bollobás and Riordan recently made a major advance, proving that the site percolation threshold is $1/2$ for the Delaunay triangulation constructed from the points of a homogeneous Poisson point process in the plane. Various models and results are described in the books by Meester and Roy [47], Bollobás and Riordan [3, Chapter 8], and Hall [48].

Fractal Percolation

Several models of fractal percolation have been proposed. A typical construction generates a random Cantor-like set by partitioning the unit square into N^2 subsquares, where $N \geq 2$, retaining each subsquare with probability p , then iterating this process for each retained subsquare. Although each point in the square is retained with probability zero,

depending on the parameters N and p it is possible to have an infinite set of points retained in the limit, and the limiting set may even contain a path which crosses the square with positive probability [49].

First-Passage Percolation

In the first-passage percolation model, each edge of the underlying graph G is assigned a nonnegative random variable which represents its travel time. Interest focuses on the shortest travel time between two vertices over the infinite set of paths available. Ergodic theory for subadditive processes is used to prove that a limiting velocity exists in each direction from the origin, and that a limiting shape exists for the set of vertices that are reached from the origin by time t , as $t \rightarrow \infty$ [50,51].

Bootstrap Percolation

The bootstrap percolation process is a cellular automata started from a random configuration. As an example, consider an $n \times n$ grid of the square lattice, with vertices occupied randomly with probability p , and, at each stage, an unoccupied vertex becomes occupied if it has at least two occupied neighbors. For p sufficiently large, every vertex eventually becomes occupied. A remarkably precise result of Holroyd [52] states shows that the asymptotic behavior of the threshold satisfies $\lim_{p \rightarrow 0} p \log n = \frac{1}{18} \pi^2$. A typical application is to advance the fluid flow front in porous media when the state of a location is dependent on the nearby pores and microstructure.

Dynamical Percolation

In dynamical percolation [53], time dynamics are added to the classical bond or site percolation model. In the dynamical bond model, for fixed p , edges of the underlying graph change between open and closed states according to independent two-state Markov processes. Changes to the open state occur at rate p , while changes to the closed state occur at the rate $1 - p$. The unique stationary distribution of the model is ordinary bond percolation with parameter p . Interest centers on the differences in behavior from the classical model.

Additional Models

The proliferation of variations has been extensive, including models named ρ -percolation, self-destructive percolation, clique percolation, last passage percolation, gradient percolation, and line-of-sight percolation. The different types of models may be used in various combinations, such as directed first-passage percolation.

CONNECTIONS TO OTHER MATHEMATICAL TOPICS

Random Graph Theory

The classical Erdős-Rényi random graph model $G_{n,p}$ is defined similarly to the bond percolation model, starting with the complete graph on n vertices and retaining each edge independently with probability p , $0 < p < 1$. Interest also focuses on the existence of thresholds at which the behavior of the graph changes dramatically. There has been a very substantial and fruitful exchange of concepts and methods between random graph theory and percolation theory. The theory is developed in [54–56].

Correlation Inequalities

Percolation theory stimulated the development of correlation inequalities, particularly the FKG inequality [57] and BK inequality [58], which are useful in reliability theory, random graph theory, and statistical mechanics.

Interacting Particle Systems

Interacting particle systems model a collection of objects on a lattice which may change the state according to the states of their neighbors [59,60]. A common application is the spatial epidemic processes. Percolation techniques have played an important role in the development of the theory.

Statistical Mechanics

Percolation is in the class of Potts models, corresponding to setting the parameter $q = 1$. The Ising model for ferromagnetism corresponds to the Potts model with $q = 2$.

Percolation research has provided insight and methods to advance understanding of these models and the more general the Fortuin–Kasteleyn random cluster models [61].

Subadditive Processes

The theory of subadditive stochastic processes originated in first-passage percolation, and the ergodic theorem for subadditive processes [62,63] plays a crucial role in the proof of the asymptotic shape theorem. Subadditive processes have found extensive applications in probability, stochastic processes, and statistics.

REFERENCES

1. Grimmett GR. Percolation. 2nd ed. Berlin: Springer; 1999.
2. Kesten H. Percolation theory for mathematicians. Boston (MA): Birkhäuser; 1982.
3. Bollobás B, Riordan O. Percolation. Cambridge: Cambridge University Press; 2006.
4. Sahimi M. Applications of percolation theory. London: Taylor & Francis; 1994.
5. Hughes B. Random walks in random environments II: random environments. Oxford: Clarendon Press; 1996.
6. Stauffer D, Aharony A. Introduction to percolation theory. 2nd ed. London: Taylor & Francis; 1991.
7. Deutscher G, Zallen R, Adler J, editors. Percolation structures and processes. Volume 5, Annals of the Israel physical society. Haifa: Inst of physics Pub Inc.; 1983.
8. Aizenman M, Grimmett GR. Strict monotonicity for critical points in percolation and ferromagnetic models. *J Stat Phys* 1991;63:817–835.
9. Hammersley JM. Comparison of atom and bond percolation. *J Math Phys* 1961;2: 728–733.
10. Menshikov MV. Coincidence of critical points in percolation problems. *Sov Math Dokl* 1986;33:856–859.
11. Aizenman M, Barsky D. Sharpness of the phase transition in percolation models. *Commun Math Phys* 1987;108:489–526.
12. (a) Lyons R. Random walks and percolation on trees. *Ann Probab* 1990;18:931–958; (b) Lyons R. Random walks, capacity and percolation on trees. *Ann Probab* 1992;20: 2043–2088.
13. Sykes MF, Essam JW. Exact critical percolation probabilities for site and bond problems in two dimensions. *J Math Phys* 1964;5:1117–1127.
14. Wierman JC. Bond percolation on the honeycomb and triangular lattices. *Adv Appl Probab* 1981;13:293–313.
15. Wierman JC. A bond percolation critical probability determination based on the star-triangle transformation. *J Phys A Math Gen* 1984;17:1525–1530.
16. Ziff RM. Generalized cell-dual-cell transformation and exact thresholds for percolation. *Phys Rev E* 2006;73:016134.
17. Ziff RM, Scullard CR. Exact bond percolation thresholds in two dimensions. *J Phys A Math Gen* 2006;39:15083–15090.
18. Wierman JC, Ziff RM. Self-dual planar hypergraphs and exact bond percolation thresholds arXiv:0903.3135.
19. Sheffield S. Random surfaces. *Astérisque* 2005;304:1–175.
20. van den Berg J, Ermakov A. A new lower bound for the critical probability of site percolation on the square lattice. *Random Struct Algorithm* 1996;8:199–212.
21. Wierman JC. Substitution method critical probability bounds for the square lattice site percolation model. *Comb Probab Comput* 1995;4:181–188.
22. May WD, Wierman JC. The application of non-crossing partitions to improving percolation threshold bounds. *Comb Probab Comput* 2007;16:285–307.
23. May WD, Wierman JC. Using symmetry to improve percolation threshold bounds. *Comb Probab Comput* 2005;14:549–566.
24. Newman MEJ, Ziff RM. Efficient Monte Carlo algorithm and high-precision results for percolation. *Phys Rev Lett* 2000;85:4104–4107.
25. Newman MEJ, Ziff RM. Fast Monte Carlo algorithm for site or bond percolation. *Phys Rev E* 2001;64:016706,16.
26. Riordan O, Walters M. Rigorous confidence intervals for critical probabilities. *Phys Rev E* 2006;76:010001.
27. Galam S, Mauger A. Universal formulas for percolation thresholds. *Phys Rev E* 1996;53:2177–2181.
28. Wierman JC, Naor DP, Smalley J. Incorporating variability into an approximation formula

- for bond percolation thresholds of planar periodic lattices. *Phys Rev E* 2007;75:011114.
29. Häggström O, Peres Y. Monotonicity of uniqueness for percolation on transitive graphs: all infinite clusters are born simultaneously. *Probab Theory Relat Fields* 1999;113:273–285.
 30. Schonmann RH. Stability of infinite clusters in supercritical percolation. *Probab Theory Relat Fields* 1999;113:287–300.
 31. Häggström O, Jonasson J. Uniqueness and non-uniqueness in percolation theory. *Probab Surv* 2006;3:289–344.
 32. Kesten H. Scaling relations for 2D-percolation. *Commun Math Phys* 1987;109:109–156.
 33. Hara T, Slade G. Mean-field behaviour and the lace expansion. *Commun Math Phys* 1990;128:333–391.
 34. den Nijs MPM. A relation between the temperature exponents of the eight-vertex and q -state Potts model. *J Phys A Math Gen* 1979;12:1857–1868.
 35. Nienhuis B, Riedel EK, Schick M. Magnetic exponents of the two-dimensional q -state Potts model. *J Phys A Math Gen* 1980;13:L189–L192.
 36. Pearson RP. Conjecture for the extended Potts model magnetic eigenvalue. *Phys Rev B* 1980;22:2579–2580.
 37. Smirnov S, Werner W. Critical exponents for two-dimensional percolation. *Math Res Lett* 2001;8:729–744.
 38. Smirnov S. Critical percolation in the plane: Conformal invariance, Cardy's formula, and scaling limits. *C R Acad Sci Paris* 2001;333:239–244.
 39. Schramm O. Scaling limits of loop-erased random walks and uniform spanning trees. *Isr J Math* 2000;118:221–288.
 40. Rohde S, Schramm O. Basic properties of SLE. *Ann Math* 2005;161:883–924.
 41. Werner W. Random planar curves and Schramm-Loewner evolutions. Volume 1840, *Lectures on probability theory and statistics: lecture notes in mathematics*. Berlin: Springer; 2003. pp. 107–195.
 42. Lawler G. Volume 114, *Conformally invariant processes in the plane: mathematical surveys and monographs*. Providence (RI): American Mathematical Society; 2005.
 43. Durrett R. Oriented percolation in two dimensions. *Ann Probab* 1984;12:999–1040.
 44. Wierman JC, Appel MJ, Infinite AB. percolation clusters exist on the triangular lattice. *J Phys A Math Gen* 1987;20:2533–2537.
 45. Appel MJ, Wierman JC. On the absence of AB percolation in bipartite graphs. *J Phys A Math Gen* 1987;20:2527–2531.
 46. Chayes JT, Chayes L, Newman CM. The stochastic geometry of invasion percolation. *Commun Math Phys* 1985;101:383–407.
 47. Meester R, Roy R. *Continuum percolation*. Cambridge: Cambridge University Press; 1996.
 48. Hall P. *An introduction to the theory of coverage processes*. New York: Wiley; 1988.
 49. Chayes JT, Chayes L, Durrett RT. Connectivity properties of Mandelbrot's percolation process. *Probab Theor Relat Field* 1988;77:307–324.
 50. Smythe RT, Wierman JC. Volume 671, *First-passage percolation on the square lattice: lecture notes in mathematics*. Berlin: Springer; 1978.
 51. Kesten H. Aspects of first-passage percolation. In: Hennequin PL, editor. Volume 1180, *Ecole d'Été de probabilités de Saint Fluor XIV: lecture notes in mathematics*. Berlin: Springer; 1986. pp. 125–264.
 52. Holroyd A. Sharp metastability threshold for two-dimensional bootstrap percolation. *Probab Theor Relat Field* 2003;125:195–224.
 53. Häggström O, Peres Y, Steif JE. Dynamical percolation. *Ann Inst Henri Poincaré, Probab et Stat* 1997;33:497–528.
 54. Bollobás B. Volume 73, *Random graphs: Cambridge studies in advanced mathematics*. 2nd ed. Cambridge: Cambridge University Press; 2001.
 55. Alon N, Spencer JH. *The probabilistic method*. 3rd ed. New York: Wiley; 2008.
 56. Janson S, Luczak T, Ruciński A. *Random graphs*. Hoboken (NJ): Wiley; 2000.
 57. Fortuin CM, Kasteleyn PW, Ginibre J. Correlation inequalities on some partially ordered sets. *Commun Math Phys* 1971;22:89–103.
 58. Berg J, Kesten H. Inequalities with applications to percolation and reliability. *J Appl Probab* 1985;22:556–569.
 59. Liggett T. *Interacting particle systems*. 2nd ed. Berlin: Springer; 1999.
 60. Durrett R. *Lecture notes on particle systems and percolation*. Pacific Grove (CA): Wadsworth and Brooks/Cole; 1988.

- 61. Grimmett GR. The random cluster model. Berlin: Springer; 2006.
- 62. Kingman JFC. Ergodic theory of subadditive stochastic processes. J Royal Stat Soc Ser B 1968;30:499–510.
- 63. Liggett T. An improved subadditive ergodic theorem. Ann Probab 1985;13:1279–1285.
- 64. Kesten H. The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. Commun Math Phys 1980;74:41–59.

PERFECT BAYESIAN EQUILIBRIUM AND SEQUENTIAL EQUILIBRIUM

AKIRA OKADA
Graduate School of Economics,
Hitotsubashi University,
Tokyo, Japan

INTRODUCTION

A subgame perfect equilibrium as defined by Selten [1] is a fundamental equilibrium concept for an extensive-form game. Specifically, it is effective for games with perfect information in which every player knows perfectly all past moves; in other words, every information set consists of a single node. The concept, however, is too weak for games with imperfect or incomplete information. A subgame perfect equilibrium may prescribe irrational player behavior after some unexpected deviations from the equilibrium strategies have occurred. To overcome this difficulty of a subgame perfect equilibrium, a concept in noncooperative game theory termed refinements of the Nash equilibrium has emerged and is the focus of active research.

The crux of the research is to define rational behavior for players when their moves are reached unexpectedly on the play of a game. The standard approach is to use the Bayesian decision theory developed by Savage [2], assuming that when a rational player makes a choice under uncertainty, he constructs a personal probability judgment about each uncertain event and maximizes the expected payoff with respect to these probabilities. Two important equilibrium concepts in this approach are those of a (weak) perfect Bayesian equilibrium defined by Fudenberg and Tirole [3] and of a sequential equilibrium defined by Kreps and Willson [4]. Roughly, these equilibrium concepts require that every player's strategy be optimal for his corresponding information set with respect to his probability judgment ("belief") about what

actions have been chosen on the previous play and players' future strategies. This condition, however, is not sufficient. In order for an equilibrium to be sensible, a strategy profile must be "consistent" with players' beliefs.

A (weak) perfect Bayesian equilibrium satisfies a weak notion of consistency that players' beliefs should be computed from the Bayes rule whenever possible, that is, for all information sets reached with positive probabilities by the equilibrium play. However, it allows any probability distribution for players' beliefs for information sets off the equilibrium play. Thus, the (weak) perfect Bayesian equilibrium may be too weak for a broad class of games and may not even be a subgame perfect equilibrium. A sequential equilibrium satisfies a stronger notion of consistency using the "trembling-hand" approach underlying a perfect equilibrium as introduced by Selten [1].

The rest of the article is organized as follows: The section titled "Preliminaries" gives some preliminaries on an extensive-form game; the section titled "Perfect Bayesian Equilibrium" reviews a (weak) perfect Bayesian equilibrium; the section titled "Sequential Equilibrium" reviews a sequential equilibrium; and the final section concludes.

PRELIMINARIES

We first present some preliminaries on extensive-form games. The definitions follow Selten [1]. For more detailed explanations, see the other sections. An *extensive-form game* is described as a set of five elements $\Gamma = (K, P, U, p, h)$. The *game tree* K is a finite directed tree with origin o . Let Z be the set of all terminal nodes in K , and X be the set of all other nodes. Each node $x \in X$ represents a decision point (move) whereby some player makes a choice. An edge connecting two nodes represents a choice. Let $S(x)$ be the set of all choices for x . A unique sequence of nodes and edges that connects o to a node x is called the *path* to x . The *player partition*

$P = (P_0, P_1, \dots, P_n)$ is a partition of X . P_0 is the set of *chance moves* whereby choices are made according to random mechanisms. $P_i (i = 1, \dots, n)$ is the set of moves by player i .

The *information partition* U is a collection $U = (U_1, \dots, U_n)$ where each U_i is a partition of P_i . An element $u \in U_i$ is called the *information set* of player i . Every information set u satisfies: (i) every path intersects u at least once, and (ii) all nodes in u have the same number of choices. When player i makes a choice for $x \in u$, he does not distinguish all other nodes in u from x . Owing to (ii), it can be assumed with no loss of generality that $S(x) = S(y)$ for every x and y in u given some suitable identification of choices. The set of choices for u , $S(u)$, is defined by $S(x)$ where $x \in u$.

The *probability assignment* p assigns to every $x \in P_0$ a completely mixed probability distribution over $S(x)$. The *payoff function* u assigns a payoff vector $h(z) = (h_1(z), \dots, h_n(z))$ to every terminal node $z \in Z$.

A *behavioral strategy* of player i is a function b_i that assigns to every $u \in U_i$ a probability distribution b_{iu} over the set $S(u)$ of choices at u . Let B_i be the set of behavioral strategies of player i . For a behavioral strategy profile $b = (b_1, \dots, b_n)$ of players, the probability of a terminal node z is defined and denoted by $\rho(z|b)$. The *expected payoff* $H_i(b)$ for player i for b is defined by $H_i(b) = \sum_{z \in Z} \rho(z|b) h_i(z)$.

A subtree K' of K is called a *subgame* of Γ if it is true for every information set u of Γ such that u never contains two nodes, one in K' and the other outside K' , simultaneously. Γ is a subgame of itself. A subgame of Γ is called *proper* if it is not equal to Γ . A behavioral strategy profile b for Γ naturally induces a behavioral strategy profile on its subgame. The expected payoff for a behavioral strategy profile in a subgame can be defined in the same way as for Γ .

We now define the two fundamental solution concepts for an extensive-form game.

A Nash equilibrium is defined by Nash [5] and a subgame perfect equilibrium is defined by Selten [1].

Definition 1. (i) A behavioral strategy profile $b^* = (b_1^*, \dots, b_n^*)$ for Γ is a *Nash equilibrium* of Γ if for every $i = 1, \dots, n$ and

for every $b_i \in B_i$, $H_i(b^*) \geq H_i(b_i, b_{-i}^*)$ where (b_i, b_{-i}^*) is the strategy profile obtained from b^* by replacing b_i^* with b_i . (ii) A behavioral strategy profile $b^* = (b_1^*, \dots, b_n^*)$ for Γ is a *subgame perfect equilibrium* of Γ if it induces a Nash equilibrium on every subgame of Γ .

PERFECT BAYESIAN EQUILIBRIUM

The notion of a subgame perfect equilibrium is very effective for eliminating unreasonable Nash equilibria in games with perfect information whereby every information set consists of only one node. Unfortunately, this is not the case in games with imperfect information.

To illustrate this point, consider a three-person extensive-form game Γ^1 with imperfect information in Fig. 1 in Ref. 1. In Γ^1 , all three players have two choices. A payoff vector assigned to each terminal node represents payoffs for players 1, 2, and 3 from top to bottom. In the game, player 1 first makes a choice. If he chooses R_1 , then player 2 makes a move knowing the choice made by player 1. If player 2 chooses R_2 , then the game ends and all players receive payoffs 1. If player 2 chooses L_2 , then the game proceeds to a move by player 3. Player 3 has another opportunity to make a choice if player 1 chooses L_1 at the

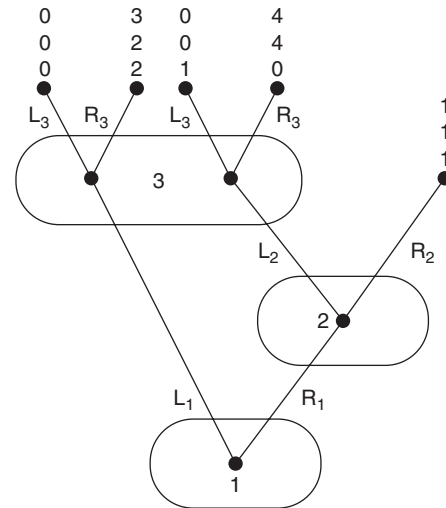


Figure 1. A game in which all Nash equilibria are subgame perfect.

start of the game. The information set of player 3, containing two nodes, reflects the fact that player 3 does not know which node is reached on his move.

Γ^1 has two Nash equilibria in pure strategies, (L_1, R_2, R_3) and (R_1, R_2, L_3) . To demonstrate why (L_1, R_2, R_3) is a Nash equilibrium, it suffices to know that the choice by every player is optimal for the strategies of all other players. Player 1's choice L_1 is optimal to player 2's choice R_2 and player 3's choice R_3 since he receives payoff 3 from L_1 larger than payoff 1 from R_1 . Note that player 2's information set is not reached when player 1 chooses L_1 . Thus, player 2 is indifferent between R_2 and L_2 . R_3 is player 3's optimal choice to L_1 . Readers can confirm for themselves that (R_1, R_2, L_3) is also a Nash equilibrium.

Since Γ^1 has no proper subgame, every Nash equilibrium of Γ^1 is a subgame perfect equilibrium. That is, (L_1, R_2, R_3) and (R_1, R_2, L_3) are subgame perfect equilibria. Are both equilibria reasonable? Consider (L_1, R_2, R_3) . When player 2 has a move unexpectedly, he receives payoff 1 from R_2 . Player 2, however, can receive a higher payoff 4 by choosing L_2 , expecting that player 3 will choose his equilibrium strategy R_3 (player 3 does not know that player 2 deviates from R_2). The Nash equilibrium (L_1, R_2, R_3) prescribes a nonoptimal strategy R_2 for player 2. If player 2's move is actually reached, it is optimal for him to choose L_2 . Realizing this fact, player 1 will choose R_1 and will receive payoff 4 higher than his equilibrium payoff 3. Clearly, this Nash equilibrium cannot be regarded as reasonable.

The game Γ^1 shows that the notion of a subgame perfect equilibrium is weak for an extensive-form game with imperfect information. A subgame perfect equilibrium may include disequilibrium (nonpayoff-maximizing) behavior for players on moves not reached by the equilibrium play with positive probability. The easiest way to overcome this difficulty of a subgame perfect equilibrium is to impose the following condition:

- (P) A subgame perfect equilibrium should prescribe player behavior that maximizes

his expected payoff after an information set containing a single node is obtained.

In Γ^1 , (R_1, R_2, L_3) satisfies this condition and (L_1, R_2, R_3) does not. Although condition (P) is effective for information sets with single nodes, it is not effective in general.

Consider a two-person game Γ^2 in Fig. 2 in Ref. 4. Γ^2 has a subgame perfect equilibrium (R_1, R_2) for pure strategies. This equilibrium satisfies (P) since player 1's choice R_1 is optimal for player 2's choice R_2 . However, it is not rational for player 2 to choose R_2 because his optimal choice is L_2 , regardless of which node is reached in the information set.

The game Γ^2 poses a general question: how should we define rational behavior for information sets with more than one node? According to Bayesian decision theory, when a rational player makes a choice under uncertainty, he constructs a personal probability judgment about each uncertain event and maximizes the expected payoff with respect to these probabilities. Such a subjective probability judgment is called a player's *belief*.

Formally, a system of beliefs is a function $\mu: X \rightarrow [0, 1]$ such that $\sum_{x \in u} \mu(x) = 1$ for each information set u . $\mu(x)$ is interpreted as the probability the player assigns to x when u is reached. An *assessment* is a pair

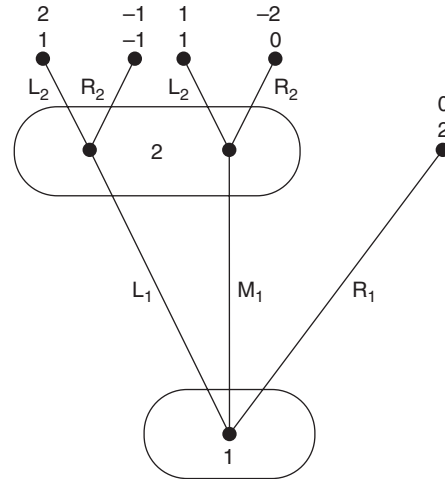


Figure 2. A subgame perfect equilibrium may prescribe irrational behavior for information sets with more than one node.

(μ, b) , where μ is a system of beliefs and b is a behavioral strategy profile. Given an assessment (μ, b) and an information set u of player i , the *conditional probability* $\rho(z|\mu, b)$ over Z_u , the set of all terminal nodes z reachable from some nodes in u , is well defined. Then the *conditional expected payoff* for player i for (μ, b) at u is defined as $H_{i,u}(b|\mu) = \sum_{z \in Z_u} \rho(z|\mu, b)h_i(z)$.

What are the types of conditions needed for an assessment (μ, b) to be a solution of an extensive-form game? Obviously, it is necessary that it should satisfy *sequential rationality*: the strategy of every player starting from his information set must be optimal, that is, maximize his conditional expected payoff after the information set, given his belief and all other players' strategies.

Sequential rationality is not sufficient for an assessment (μ, b) to be an equilibrium since a system μ of beliefs must be "consistent" with a behavioral strategy profile b . What belief should a rational player have? One reasonable condition is that a player's belief must be computed from the Bayes rule whenever possible. If an information set is reached with positive probability, then a player's belief about which node is reached in it should be equal to the conditional probability given by the Bayes rule. Then, what is his belief when an information set has zero probability ex ante? What is a reasonable belief when a player finds that his information set is reached unexpectedly? A possible answer is that we simply allow any probability distribution in such a case. This is a weak version of the perfect Bayesian equilibrium in Ref. 3.

Definition 2. An assessment (μ^*, b^*) is a (weak) *perfect Bayesian equilibrium* of an extensive-form game Γ if the following conditions hold for every $i = 1, \dots, n$.

1. *Sequential Rationality.* For every information set $u \in U_i$ and every $b_i \in B_i$, $H_{i,u}(b^*|\mu^*) \geq H_{i,u}(b_i, b_{-i}^*|\mu^*)$.
2. *The Bayes Rule.* For every $u \in U_i$ and every $x \in u$,

$$\mu(x) = \frac{\rho(x|b^*)}{\sum_{y \in u} \rho(y|b^*)},$$

if $\sum_{y \in u} \rho(y|b^*) > 0$ where $\rho(y|b^*)$ is the probability that a node $y \in u$ is reached under b^* . If $\sum_{y \in u} \rho(y|b^*) = 0$, then $\mu(x)$ can be any probability distribution over u .

The conditions of a perfect Bayesian equilibrium are simple and easy to check. For example, a Nash equilibrium (R_1, R_2, L_3) of Γ^1 is a perfect Bayesian equilibrium. Since the information sets of players 1 and 2 contain single nodes, their beliefs are trivial and they satisfy sequential rationality. Equilibrium strategy L_3 of player 3 satisfies sequential rationality if his belief assigns to the right node a probability greater than $2/3$. Since the information set of player 3 is not reached on the equilibrium play, any belief is allowed.

Owing its simplicity, the notion of a (weak) perfect Bayesian equilibrium is frequently applied to economic models, and specifically to games with two sequential moves. However, it may not work well if a game has more than two sequential moves. The game Γ^3 in Fig. 3 illustrates this point. In Γ^3 , player 1 has two information sets and player 2 has one information set. Consider a behavioral strategy profile $b = (R_1 r_1, R_2)$. An assessment (μ, b) is a weak perfect Bayesian equilibrium if the belief μ assigns probability

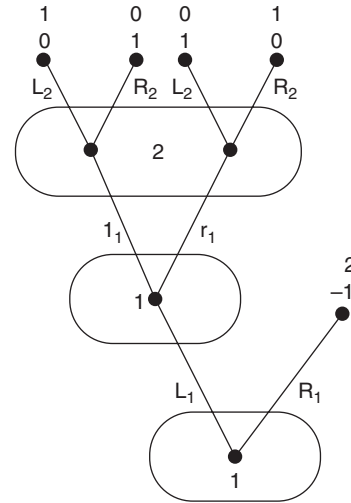


Figure 3. A weak perfect Bayesian equilibrium may not be subgame perfect.

1 to the left node in the information set of player 2. Since the latter information set is not reached on the play of b , such a belief is admissible. By contrast, b is not a subgame perfect equilibrium. Namely, the strategy profile (r_1, R_2) induced by b is not a Nash equilibrium of the subgame starting at the player 1's second information set.

The game Γ^3 shows that it is necessary to impose a consistency between players' strategies and their beliefs for information sets not reached by the equilibrium play. This task is achieved using the concept of a sequential equilibrium.

SEQUENTIAL EQUILIBRIUM

Kreps and Wilson [4] define the notion of consistency for an assessment in an extensive-form game using the trembling-hand approach introduced by Selten [1]. The basic idea is as follows: If an information set of a player is reached unexpectedly on a game play, then the player imagines that some previous players may have deviated from their equilibrium strategies by "mistake" before his move. Then, according to his imagination about possible mistakes on the past play, the player constructs his belief about which node has been reached in the information set using the Bayes rule. Formally, a consistent assessment is defined as follows:

Definition 3. An assessment (μ^*, b^*) of an extensive-form game Γ is *consistent* if there exists some sequence $\{(\mu_n, b_n)\}_{n=1}^\infty$ of assessments such that:

1. Each b_n is a completely mixed behavioral strategy profile, that is, it assigns a strictly positive probability to every choice in Γ , and μ_n is the system of beliefs generated from b_n by the Bayes rule; and
2. (μ_n, b_n) converges to (μ^*, b^*) as $n \rightarrow \infty$.

It should be noted that every information set is reached with a positive probability under a completely mixed behavioral strategy profile. Thus, the Bayes rule can be

applied for every information set. A sequential equilibrium is a weak perfect Bayesian equilibrium satisfying consistency in Definition 3.

Definition 4. An assessment (μ^*, b^*) of an extensive-form game Γ is a *sequential equilibrium* if it is both consistent and sequentially rational.

To understand the consistency criterion for a sequential equilibrium, consider again the game Γ^1 in Fig. 1. We show that a Nash equilibrium $b^* = (R_1, R_2, L_3)$ is a sequential equilibrium with the system μ^* of beliefs where μ^* assigns probability 1 to single nodes in the information sets of players 1 and 2, and probability 2/3 to the right node in the information set of player 3. It is easy to see that the assessment (μ^*, b^*) is sequentially rational. To demonstrate that (μ^*, b^*) is consistent, consider a completely mixed behavioral strategy profile $b_n = (b_{1,n}, b_{2,n}, b_{3,n})$ such that $b_{1,n} = (1 - \epsilon_n, \epsilon_n)$, $b_{2,n} = (1 - 2\epsilon_n, 2\epsilon_n)$, and $b_{3,n} = (\epsilon_n, 1 - \epsilon_n)$, where the first component in $b_{i,n}$ ($i = 1, 2, 3$) represents the probability assigned to choice R_i . Suppose that each $\epsilon_n > 0$ and ϵ_n converges to 0 as n goes to infinity. From $(b_{1,n}, b_{2,n})$, we can compute the belief of player 3 for his information set using the Bayes rule. This assigns probability

$$\mu_n(x) = \frac{2(1 - \epsilon_n)\epsilon_n}{2(1 - \epsilon_n)\epsilon_n + \epsilon_n} = \frac{2(1 - \epsilon_n)}{2(1 - \epsilon_n) + 1},$$

to the right node x in the information set of player 3. $\mu_n(x)$ converges to 2/3 as n goes to infinity. This demonstrates that (μ^*, b^*) is consistent. The construction of the completely mixed behavioral strategy profile b_n means that when his information set is reached unexpectedly, player 3 believes that players 1 and 2 may have chosen L_1 and L_2 mistakenly, and moreover that the error probability for player 2 is twice as great as that for player 1.

Kreps and Wilson [4] prove that a sequential equilibrium has the following properties.

Theorem 1. *For every extensive-form game with perfect recall by Kuhn [6], there exists at least one sequential equilibrium.*

Theorem 2. *If an assessment (μ^*, b^*) is a sequential equilibrium, then b^* is a subgame perfect equilibrium.*

The concept of a sequential equilibrium has prompted active research into noncooperative game theory to scrutinize “credible” beliefs for off-equilibrium play in Ref. 7–9. Although the definition of consistency by Kreps and Wilson [4] is mathematically simple, it may be too weak for some classes of extensive-form games.

Consider the game Γ^4 in Fig. 4 in Ref. 10. (R_1, R_2) is a sequential equilibrium with the belief of player 2 that assigns probability 1 to the right node in his information set. Note that player 2’s information set is not reached when player 1 chooses R_1 . It is evident that player 2’s belief is consistent with player 1’s strategy R_1 . If his information set is reached unexpectedly, player 2 believes that choice of M_1 is infinitely more likely than L_1 . In fact, any belief of player 2 can be consistent with player 1’s strategy R_1 (the proof is left to readers). However, it can be claimed that player 2’s belief above is not intuitively reasonable. The reason is as follows: M_1 is dominated by R_1 , that is, it gives player 1 a lower payoff than R_1 , regardless of which action player 2 chooses. If we take the view that a rational player should not choose any dominated action, then it is not sensible for player 2 to believe that player 1 has chosen the dominated action M_1 . When his information set is reached, player 2 should conclude that player 1 has chosen L_1 , not M_1 . Then his optimal choice will be L_2 . By this reasoning, we can say that (R_1, R_2) is not a sensible equilibrium and that the only sensible equilibrium is (L_1, L_2) .

In non-cooperative game theory, the question as to what a credible belief should be for an information set for off-equilibrium play has not been fully answered yet. The answer seems to depend on how a player should interpret players’ possible deviations ex post, whereas such events are impossible ex ante. Selten’s trembling-hand approach supposes players’ random mistakes by some unspecified psychological mechanisms. Other approaches discussed in the literature are to

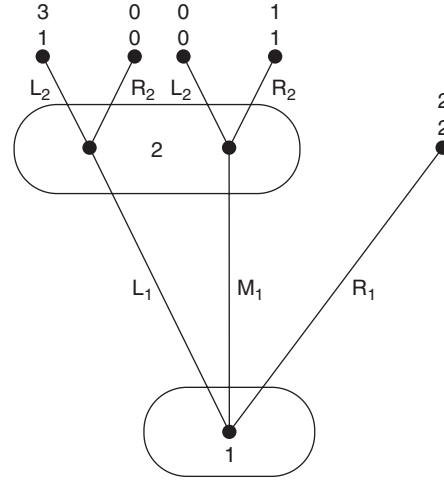


Figure 4. A sequential equilibrium may be too weak.

suppose that different players may have different theories about how to behave or that players deviate with a certain intention and send some signals by doing so. The latter approach is called *forward induction*.

Finally, there is also active research into weakening the concept of a sequential equilibrium. Fudenberg and Tirole [3] present the notion of perfect Bayesian equilibrium that satisfies sequential rationality and a weaker notion of consistency without the trembling-hand approach.

CONCLUSION

We have reviewed two important refinement concepts of a subgame perfect equilibrium for an extensive-form game with imperfect information, a (weak) perfect Bayesian equilibrium and a sequential equilibrium. These equilibrium concepts are defined in terms of a pair comprising a behavioral strategy profile and a system of beliefs so that they satisfy sequential rationality and consistency. A (weak) perfect Bayesian equilibrium requires a weak consistency that the Bayes rule should be applied whenever possible. A sequential equilibrium requires a stronger consistency that uses the trembling-hand approach of Selten [1].

REFERENCES

1. Selten R. Reexamination of the perfectness concept for equilibrium points in extensive games. *Int J Game Theory* 1975;4:25–55.
2. Savage LJ. *The foundations of statistics*. New York: John Wiley & Sons, Inc.; 1954.
3. Fudenberg D, Tirole J. Perfect Bayesian equilibrium and sequential equilibrium. *J Econ Theory* 1991;53:236–260.
4. Kreps D, Wilson R. Sequential equilibria. *Econometrica* 1982;50:863–894.
5. Nash JF. Non-cooperative games. *Ann Math* 1951;54:286–295.
6. Kuhn HW. Extensive games and the problem of information. In: Kuhn HW, Tucker AW, editors. *Volume II, Contributions to the theory of games*, *Annals of Mathematical Studies* 28. Princeton (NJ): Princeton University Press; 1953. pp. 193–216.
7. Myerson RB. Refinements of the Nash equilibrium concept. *Int J Game Theory* 1978;7: 73–80.
8. Cho IK, Kreps D. Signaling games and stable equilibria. *Q J Econ* 1987;102:179–221.
9. Kohlberg E, Mertens JF. On the strategic stability of equilibria. *Econometrica* 1986;54: 1003–1037.
10. Damme EV. *Stability and Perfection of Nash Equilibria*. Berlin: Springer; 1991.

PERFECT INFORMATION AND BACKWARD INDUCTION

MAREK M. KAMINSKI
Department of Political Science and
Institute for Mathematical Behavioral
Science, University of California,
Irvine, California

The origins of backward induction (BI) are unclear. Zermelo [1] was the first to analyze winning in chess, but his method of analysis was based on a different principle than BI [2]. Reasoning based on BI was implicit in Stackelberg's [3] construction of his alternative to the Cournot equilibrium. As a general procedure for solving two-person zero-sum games of perfect information, BI appeared in the von Neumann and Morgenstern's 1944 founding book [4], p. 117]. It was used to prove a precursor of Kuhn's Theorem for chess and similar games. The von Neumann's exceedingly complex formulation was later clarified and elevated to a high theoretical status by Kuhn's work [5], especially Corollary 1], Schelling's [6] ideas of incredible threats, and Selten's [7] introduction of subgame perfection. In the late 1970s, BI went under critical fire when the chain-store paradox, Centipede, and other games questioned its universal appeal [8,9]. The criticism, supported by experimental evidence, greatly diminished the appeal of backward induction in some games.

DESCRIPTION OF THE BACKWARD INDUCTION PROCEDURE

A bum approaches you in a dark street, displays a gun, and asks for a \$100 donation. If you give him the money (M), the game ends. If you refuse his offer (R), he promises to shoot *himself* (see the left panel in Fig. 1 BUM).

The solution to the game based on BI would proceed as follows: First, let us consider what happens when you refuse to give the money to the bum. He chooses between

shooting himself and receiving the payoff of 0 versus not shooting and receiving 1; thus, he selects N . In the first case, you receive 0 and in the second case you receive 2. You can anticipate his action N and assume that refusing automatically gives you 2 (since he will not shoot himself), while giving the money to the bum gives you 1. You choose not to give him the money.

In your reasoning, you used a small part of the original game, where the bum's subgame is substituted with the payoffs resulting from his best choice (Fig. 1, right panel). By analyzing the smaller game, you choose R , and the strategy profile chosen by BI is (R, N) . The game was solved by running backward with respect to the decision making order. First, the last decision in the game was analyzed and then the reasoning turned to the first decision.

The main idea behind the BI algorithm is to keep reducing games to smaller games, and to keep in mind the partial solution obtained in the process of reduction. Slightly more formally, the algorithm may be described for any finite extensive form game of perfect information in the following way.

Start at some of the final decision nodes, D (let us assume that the player who moves at D is j). Then select an action that offers the highest payoff to j at D (let us call such an action a). The existence of at least one payoff-maximizing action is guaranteed by the finite character of the game. Let us assume that there is just a single action that leads to the highest payoff. (If there are many actions with such a property, the reasoning is conducted separately for all of them and the more general algorithm applicable to such a case is slightly more complex.) Next, a new game is constructed by substituting D and the branches following D in the old game with the payoff vector assigned to a . The intuition behind the reduction is that all players in the game may predict that, once D is reached, j will correctly choose the action a that maximizes his payoff. Thus, once we reach D , the payoff vector assigned

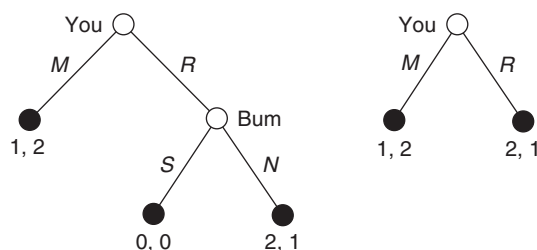


Figure 1. The game of Bum (left panel) and Bum reduced after applying the first step of backward induction (right panel).

to a will automatically follow. This means that there is no point in considering D and the following actions. The game may be now simplified. One has to remember the action a that j selected at D since this particular action becomes a part of the game's solution.

The reasoning described above, that is, the reduction of the original game to smaller games, is repeated until a final single vector of payoffs is obtained. What will result from the reduction is a strategy profile that consists of all best actions obtained at some stage of reduction procedure. Those actions (including the actions that will not be actually played) constitute a backward induction strategy profile (BISP). It can be proved that the final strategy profile is the same for all possible reductions of the original game.

When there are two or more actions available, one has to create separate partial strategy profiles for all such actions that include all previously obtained actions. Then the reasoning proceeds in a similar fashion for every partial strategy profile.

NASH EQUILIBRIUM, INCREDIBLE THREATS AND SUBGAME PERFECTION

In the game of Bum, the strategy profile selected by BI was (R, N) . It is easy to see that this particular strategy profile constitutes a Nash equilibrium of this game (see **Nash Equilibrium (Pure and Mixed)**). If Player 1 switches to M , then her payoff goes down from 2 to 1. If Player 2 switches to S , his payoff goes down from 1 to 0. This example illustrates the first important property of BI: it selects only Nash equilibria.

(R, N) is not the only Nash equilibrium. Another one is (M, S) in which you offer the bum his donation; if you did not, he would

shoot himself. This equilibrium has one flaw: it does not look reasonable. It is based on an *incredible threat* that the bum would shoot himself once you refuse the donation. However, if you refuse, the bum has no incentive to carry out the threat. He would be better off by abandoning his threat.

The concept of incredible threat was introduced by Schelling [6]. The interplay between credible versus incredible threats was at the center of the hugely popular movie "Dr. Strangelove." BI cleverly disregards Nash equilibria that are based on incredible threats. Thus, BI offers a *refinement* of the Nash equilibrium, that is, for every game, BISP are a subset of all Nash equilibria that excludes from consideration equilibria that are intuitively "unreasonable." This is the second important property of BI. However, while being very intuitive, the idea of "incredible threat" turned out to be difficult to formalize. Moreover, BI eliminates other unreasonable Nash equilibria that are not based on incredible threats.

Let us assume that our bum has an option of shooting himself also when he receives the \$100 donation. Thus, the game does not end with the donation M but two more actions are available to the bum once he receives the donation: s and n (see Fig. 2 Crazy Bum).

In Crazy Bum, there are three Nash equilibria. (M, nS) and (R, nN) intuitively correspond to (M, S) and (R, N) in the original Bum. The third equilibrium, (R, sN) , is the weirdest one. In this equilibrium, you refuse to make the donation, and the bum does not shoot himself. However, he would shoot himself if you gave him the money! The behavior associated with the strategy profile (R, sN) makes even less sense than the one associated with the equilibrium based on an

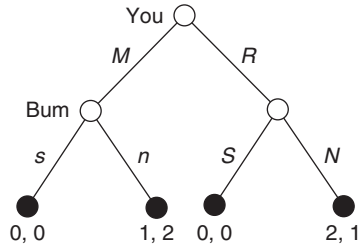


Figure 2. The game of Crazy Bum.

incredible threat. And yet, (R, sN) is a Nash equilibrium.

The ideas incorporating incredible threats and other unreasonable equilibria were formalized by Selten with his solution concept of subgame perfect equilibrium [7]. Briefly, a strategy profile is subgame perfect equilibrium (SPE) if it is a Nash equilibrium in all subgames of a game, including the game itself. The requirement imposed by the Nash equilibrium on players is strengthened. Since a subgame can be imagined as a small self-enclosed game, in an SPE, players are required to play Nash equilibrium strategies at every decision node that could demarcate such a new smaller game. In the game of Bum, SPE requires that the Nash equilibrium in this game must have an additional property: in the subgame, when the bum makes his decision, he must choose an action that is better for him. This is precisely the requirement of eliminating incredible threats. In the game of Crazy Bum, SPE requires additionally that the bum must choose the best action after receiving the money.

It is relatively easy to prove that, for finite games of perfect information, BI and subgame perfection represent the same solution concept, that is, the sets of BISP and SPE are always identical. It can also be proved that this coincidence is in fact true for all games (according to a very general definition of a game) and all pure strategies [10]. Thus, the final important fact about BI is that it selects precisely the same set of strategy profiles as subgame perfection. At the same time, for many games it offers an effective algorithm of finding all SPEs, which is simpler than looking for SPEs directly from its definition.

VARIANTS

In the version of BI described earlier, every smaller game was obtained by removing a single set of branches that originated at the same decision node. This version of BI can be generalized and multiple branches can be removed simultaneously.

Consider the following two-player game: you start with choosing a positive integer that is not greater than ten. Next, your opponent chooses another integer that is greater than your number, but the difference is again not greater than 10. Then you choose another integer by adding to the previous number no more than ten, and so on. The player who is the first to say a number at least equal to 100 is the winner.

Who can claim victory in this game?

The reader may be willing to pause reading at this point and to try solving the game. In fact, the problem is not easy to solve until we apply BI. Then it becomes trivial:

The game is finite, and there is a winner in each specific play. Let us assume that the winning player made the total of n moves. She could say 100 if she said 89 in her move $n - 1$ since her opponent must say at least 90 and no more than 99. Thus, saying 89 is a sufficient condition for guaranteeing herself the victory.

By the same logic, a player could say 89 in move $n - 1$, she could say 78 in her move $n - 2$, and so on. The strategy to say consecutively 1, 12, 23, 34, 45, 56, 67, 78, 89, 100 regardless of the other player's choices guarantees a player winning. Thus, Player 1—who can start by saying “one”—can claim the victory.

Our solution obtained above clearly uses some form of BI. However, a more careful examination of the reasoning reveals that, in every step that takes us backwards, we remove more than just one decision node with corresponding branches. For instance, in the last step, after Player 1 says 89, Player 2 could say any number from 90 to 99 before Player 1 says 100. All such numbers represent one or more subgames. Thus, in our solution, we remove several subgames and several decision nodes at the same time. More specifically, we remove an entire complex

subgame that starts with the decision node following Player 1 saying 89.

Kaminski [10] implies that the above-described procedure is a legitimate extension of BI. Moreover, we can substitute any disjoint subset of subgames with the SPE payoffs in those subgames and obtain all SPEs in the original game in the same way as it happens with the original BI. This property becomes critical for games with infinite actions or of infinite length.

In the Ultimatum game, Player 1 chooses a number $x \in [0, 1]$ that represents the amount Player 1 wants to keep. Next, Player 2 makes the choice. She may accept the proposed division and receive $1 - x$, while Player 1 receives x , or she may reject it, in which case both players receive zero. The BI reasoning stipulates that Player 2 accepts any positive amount offered to her and that she is indifferent between accepting and rejecting zero. Thus, in SPE, Player 1 can propose only one, that is, everything for himself. The unique SPE in the ultimatum game is when Player 1 proposes one and Player 2 accepts any proposed amount, including zero.

The ultimatum game provides a simple illustration of the usefulness of BI outside of the realm of finite games. However, as discussed later, its predictive power is questionable. A better example of how the strategic intuition motivating BI allows gaining insights into an important political process is provided by the Agenda-setter game [11].

Two players, the Agenda setter A and the Legislator L, have Euclidean preferences in the issue space $[0, 3]$ and the ideal points $a = 0$ and $l = 2$, respectively. The status quo is $q = 3$. First, A proposes a policy $x \in [0, 3]$. Next, L exercises his veto power and chooses the final law from $\{x, q\}$, that is, either rejects (R) or agrees that the new policy be implemented (P). (Euclidean preferences define the payoffs as negatives of distances between a player's ideal point and the point in the issue space resulting from the strategy profile.)

To solve the game by BI, let us consider L's problem. Since the distance between l and q is 1, L prefers any new policy that is closer to l than 1; is indifferent between 1 and q ; and prefers q to any policy that is farther from l than 1. Thus, the best action responding to

$x \in [0, 1]$ is R ; the best answer to $x \in (1, 3)$ is P ; and for 1 both R and P are best answers. Now, let us consider A's problem. Anticipating L's response, A does not try to impose on L his ideal point, zero. He offers a policy that is closer to L's ideal point. The policy that is closest to A's ideal point and acceptable for L is 1. Thus, the unique outcome selected by BI is 1.

The Agenda-setter game explains a vast variety of strategic interactions taking place between the two chambers of parliament, parliaments, and supreme courts or presidents endowed with veto power, and so on, all over the world. While the calibration of issue spaces is a perennial problem, the universally observed phenomenon is that policy proposals take into account the preferences of the relevant veto players, and incorporate them into the offer.

CRITICISM

It is sometimes argued that in the ultimatum game, zero offers to the second player happen infrequently, and that players sometimes reject even positive amounts offered to them. However, the phenomenon may be explained not by the failure of BI to predict the outcome correctly, but rather by an incorrect specification of player payoffs in the experimental setting. The actual player payoffs may not only represent the amount that the players receive but also their feelings of fairness. When monetary rewards are small, as is the case in experiments, such feelings may seriously disturb the payoffs.

There are more serious problems with BI than its alleged failure to predict the outcome of the ultimatum and similar games. One may ask a simple question: What would happen with games like chess if all potential players were using BI in a consistent fashion? The answer is now truly troublesome: nobody would be willing to play a game. Perfect "backward inductionists" would always reach the same outcome. We have not learned so far whether it would be the victory of whites or blacks, or possibly a tie. What we do know is that in the world of perfect BI this determinate outcome would be well known and there would be no thrill from playing

the game. Similarly, all board or computer games of perfect information would instantly lose their appeal.

Obviously, people play chess and other games and reach various outcomes, which implies that at least some of them are not perfect backward inductionists.

In fact, there are quite simple games in which real-world players practically always deviate from the principles of BI. The deviations can be attributed to the complexity of decision environment, the presence of a small amount of incomplete information, or making assumptions about other players' reasoning not being perfectly compatible with BI. In games such as the chain-store paradox [8,9], competing firms may engage in behavior that contradicts the predictions drawn from BI. Consider the following game of Centipede (see Fig. 3 Centipede)

Player 1 begins. She can continue or she can stop. When she stops, both players receive the payoff of one. Then Player 2 faces a similar choice, and so on. A player who continues increases the payoff of the other player by two and decreases his own payoff by one, assuming that the other player would immediately stop. The unique BI solution to the Centipede is that Player 1 stops immediately (and that both players would stop at any moment of the game). However, one can notice that, if both players continue for just one round, they will always get at least as much as in the unique BI equilibrium (Player 1's payoff may drop to 1 while Player 2's payoff will always remain strictly higher than 1). If they continue for at least two rounds, they always get strictly more. In fact, in experiments players typically recognize the rewards coming from

a longer play and practically always continue until nearly the end of the game. (When demonstrating Centipede to his students, the author of the present article started with ten-rounds-long Centipedes and payoffs in real dollars. He quickly learned that the three-round version is sufficient to illustrate the problem and, needless to say, much less expensive!)

The epistemic literature on BI puts emphasis on player expectations. If Player 2 gets the chance to make her first move, she understands that Player 1 deviated from the reasoning prescribed by BI and did not stop. Thus, she may be willing to assume a substantial probability of another future deviation, and continue herself. But Player 1 can predict this reasoning of Player 2 and continue in the first round precisely because of such expectations.

Another problem of BI is also closely connected to the insufficient power of subgame perfection to eliminate unreasonable Nash equilibria. In the game shown below, there are three Nash equilibria (see Fig. 4).

Player 2 is indifferent between the two available actions L and R . The version of BI that is suitable for such cases would admit all three Nash equilibria: When 2 selects R , 1 selects a (equilibrium (a, R)); When 2 selects L , 1 selects either a or b (equilibria (a, L) and (b, L)). However, the third equilibrium (b, L) does not seem reasonable since the strategy b of player 1 is weakly dominated by his strategy a . One can hardly justify the use of this particular strategy by Player 1—even if b performs against L as well as a , why would he take the risk that Player 2 will choose R and will bring him a lower payoff? Player 2 is

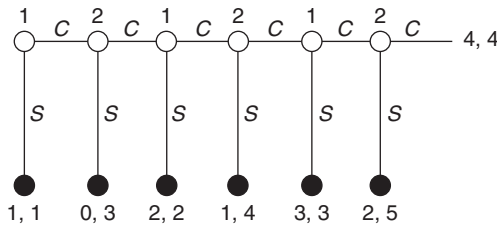


Figure 3. The game of Centipede with three rounds.

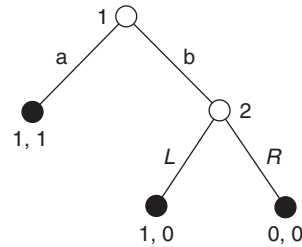


Figure 4. A game with BISP that includes a weakly dominated strategy.

indifferent between L and R and Player 1 has no reason to believe that one will be chosen over another.

The unreasonably looking strategy profile (b, L) can be eliminated by the solution concept such as perfect equilibrium that is more restrictive than SPE [12].

INCORRECT USES OF BI

A popular BI anecdote tells the following story. The prison warden announces to a game theorist on death row: "Today is Sunday. You will be executed during the next week. You will be very surprised when the day of your execution comes." The game theorist thinks for a while and responds with relief: "In such a case, I won't be executed at all!"

What was his reasoning?

It is easy to reconstruct the poor game theorist's way of thinking. He cannot be executed on next Sunday, the last day his execution is possible, since in such a case he would not be surprised at all. Thus, since next Sunday is impossible, the last possible day of execution is Saturday. But it is also impossible to hold the execution on Saturday since, once it comes, it would be the last possible day for the execution because Sunday had already been eliminated. But this means that our game theorist would certainly expect execution on Saturday and would not be surprised. By going backward, we can eliminate every single day in the next week. The execution is not possible.

How did the story end? The game theorist was executed on Wednesday and he was very surprised!

The flaw in the backward reasoning described above is that it does not refer to any specific game (the reader may try to construct a game that would correspond to the story). Instead, it uses some concepts that do not even have formal equivalents in extensive form games, such as "being surprised." Thus, while there is some superficial similarity to BI, the backward reasoning is not BI. There is some reduction applied and the inference goes backward, but this is

about where the similarity between the two types of reasoning ends.

REFERENCES

1. Zermelo E. Über eine anwendung der mengenlehre auf die theorie des schachspiels. Cambridge: Cambridge University Press; 1913. pp. 501–504.
2. Schwalbe U, Walker P. Zermelo and the early history of game theory. *Games Econ Behav* 2001;34:123–137.
3. Stackelberg HFV. Marktform Und Gleichgewicht (Market Structure and Equilibrium). Vienna; 1934.
4. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton: Princeton University Press; 1944.
5. Kuhn HW. Extensive games and the problem of information. In: Kuhn HW, Tucker AW, editors. Contributions to the theory of games 1. Princeton: Princeton University Press; 1953. pp. 193–216.
6. Schelling TC. The strategy of conflict. Cambridge, MA: Harvard University Press; 1960.
7. Selten R. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Z Gesamte Staatswiss.* 1965;121:301–324, 667–689.
8. Selten R. The chain store paradox. *Theory Decis* 1978;9:127–159.
9. Rosenthal RW. Games of perfect information, predatory pricing, and the chain-store paradox. *J Econ Theory* 1981;25:92–100.
10. Kaminski MM. Backward induction and subgame perfection. IMBS working paper, University of California, Irvine. 2009; pp. 09–01.
11. Romer T, Rosenthal H. Political resource allocation, controlled agendas, and the status quo. *Public Choice* 1978;33:27–44.
12. Selten R. Reexamination of the perfectness concept for equilibrium points in extensive games. *Int J Game Theory* 1975;4:25–55.

FURTHER READING

- Basu K. Strategic irrationality in extensive games. *Math Social Sci* 1988;15247–260.
- Basu K. On the non-existence of a rationality definition for extensive games. *Int J Game Theory* 1990;19:33–44.
- Binmore K. Modeling rational players, Part 1. *Econ Philos* 1987;3:179–214.

- Binmore K. Modeling rational players, Part 2. *Econ Philos* 1988;4:9–55.
- Bonanno G. The logic of rational play in games of perfect information. *Econ Philos* 1991;7: 37–65.
- Fudenberg D, Tirole J. *Game theory*. Cambridge, MA: The MIT Press; 1991.
- Myerson R. *Game theory. Analysis of conflict*. Cambridge, MA: Harvard University Press; 1991.
- Petit P, Sugden R. The backward induction paradox. *Econ Philos* 1989;14:95–125.
- Reny PJ. *Rationality, common knowledge and the theory of games*. Princeton: Princeton University; 1988.

PERFORMANCE BOUNDS IN QUEUEING NETWORKS

MICHAEL H. VEATCH
Department of Mathematics,
Gordon College, Wenham,
Massachusetts

Efficient scheduling of queueing networks is an important topic in applications including manufacturing systems and communication networks. However, because of the intractability of their optimal control formulations and the complexity of optimal policies, most analysis has focused on simple policies or stability. Given these limitations, one useful approach has been to compute bounds on the performance of any policy. Computing a lower bound on a cost measure allows analysts to also bound the suboptimality of a proposed heuristic policy, which can be very useful in the development of heuristics. Computing an upper bound on the performance of a specific policy or class of policies can be useful in establishing stability or establishing the robustness of a class of policies.

Most of the bounds we discuss are based on a Markov decision process (MDP) formulation of the network control problem with exponentially distributed interarrival and service times. We also restrict attention to sequencing control, that is, each station decides at each time what class of customer to serve, or possibly to idle. The network structure is an open multiclass network with probabilistic routing (see *Multiclass Queueing Network Models*). The objective is to minimize the long run average number of customers in the system, or average cost, where cost is a linear function of the number of customers of each class. Closed reentrant lines are also briefly considered.

The bounds considered are, to varying degrees, computable. Some are in closed form, while others require solving a linear program (LP) or computing average cost under a simple policy. Another body of research has developed theoretical bounds on the error in

some computable average cost estimate. For example, Refs 1–3 present bounds on the error using approximate LP techniques for discounted cost problems. They also bound how far the cost achieved by the approximate policy is from the optimal cost. Both bounds are instructive but involve quantities that cannot be computed, such as the stationary distribution under the optimal policy.

The bounds are presented in a roughly historical order. The sections titled “Achievable Region Method,” “Approximate Linear Programming,” and “LP Upper Bounds” cover LP methods and some nonlinear programming extensions of them. The approximate LP method involves approximating the dynamic programming value function. A dual relationship between these methods demonstrates that certain approximate LP bounds are tighter. The section titled “LP Upper Bounds” modifies the LPs in the first two sections. The next section is titled “Pathwise Bounds.” The section titled “Lyapunov Function Bounds” interprets the approximate LP inequalities as the condition for a Lyapunov function and describes bounds on expected cost and tail probabilities for general Lyapunov functions.

QUEUEING NETWORK SCHEDULING

An open multiclass network is defined as follows (see *Multiclass Queueing Network Models*). Customers arrive and move through different classes before exiting the network. There are n customer classes and m stations, each of which serves one or more classes. Let $X_i(t)$ be the number of class i customers at time t , including any that are being processed. Class i customers are served by station $\sigma(i)$. Each station can serve only one customer at a time. Unless otherwise noted, we assume that exogenous arrivals at each class are independent Poisson processes with rates α_i and processing times are independently exponentially distributed with mean $1/\mu_i$ in class i . Both the processing times and the routing of a customer depend on the class (not the station) in a multiclass network.

Routing may be probabilistic: $P = [p_{ij}]$ is the routing matrix with p_{ij} the probability that a customer finishing service at class i will be routed to class j , independent of all other history. We let p_{i0} represent the probability that a customer finishing service at class i leaves the system; it is equal to $1 - \sum_j p_{ij}$.

The network is open in that every customer eventually leaves with probability 1. For open networks, the effective arrival rate to class i , λ_i , is the unique solution to the traffic equation

$$\lambda = \alpha + P^T \lambda.$$

The utilization of server σ is $\rho_\sigma = \sum_{i:\sigma(i)=\sigma} \lambda_i / \mu_i$. We assume that $\rho < 1$.

We wish to bound the performance of any control policy. The performance measure is long-run average cost, J , where the cost rate is $c^T x$ in state $X(t) = x$. The minimum value of J for all policies is denoted by J^* . Often the objective is to minimize the expected number of customers in the system, in which case $c = \mathbf{1}$. For these system performance measures, the queue discipline within a class does not matter. Instead, each server must decide which job class to work on at each time, or possibly to idle. Preemption is allowed. Let $u_i(t) = 1$ if class i is served at time t and 0 otherwise. The infinite horizon objective and exponential distributions imply that there is an optimal policy that is stationary and Markov. Hence, we consider only such policies, writing $u(t) = u(X(t))$. In a slight abuse of notation, we also let u denote an action. The allowable actions in state x are

$$\mathcal{A}(x) = \left\{ u : \sum_{i:\sigma(i)=\sigma} u_i \leq 1 \quad \text{for all } \sigma, u \leq x, \right. \\ \left. u_i = 0 \text{ or } 1 \right\}.$$

The first constraint says that a server can only work on one class at a time; the second prevents serving an empty buffer. Also, let $\mathcal{A}$ be all the allowable actions. Because $\rho < 1$, there exists a stabilizing policy, so J^* is finite.

To uniformize the process, introduce an independent Poisson process with rate μ_i for

each class i , representing potential service completions. A discrete-time Markov chain is constructed by observing the system at each arrival and potential service completion. A potential service completion is an actual service completion if that class is being served. Assume that the potential event rate $\sum_{i=1}^n (\alpha_i + \mu_i) = 1$ so that these rates are also transition probabilities. In the sequel $X(t)$, $t = 1, 2, 3, \dots$ denotes the discrete-time Markov chain.

The network is called a *reentrant line* if all customers follow the same cycle through the network, say visiting classes in the order $1, 2, \dots, n$. In this case, $\alpha = (\alpha_1, 0, \dots, 0)$, $p_{i,i+1} = 1$ for $i < n$, $p_{n0} = 1$, and $\lambda_i = \alpha_1$ for all i . Customers may reenter a station, $\sigma(i) = \sigma(j)$ for some $i \neq j$. We will briefly consider closed reentrant lines. A network is called *closed* if no customers arrive or depart and a certain number of customers, N , circulate endlessly. In a closed reentrant line, all customers follow the same cycle through the network and are routed from class n to class 1. Each class has the same long run throughput, which will also be called λ .

ACHIEVABLE REGION METHOD

The idea of the achievable region method is to write flow balance equations, which hold in steady-state under any stable policy, for certain functions of $X(t)$. In particular, quadratic test functions lead to linear equations in the expected queue length and some auxiliary variables. These define a region in the space of expected queue lengths which contains the region that can be achieved under various policies, known as the *achievable region*. Minimizing average cost over the larger region bounds the average cost of any policy. This method was introduced in Ref. 4 for single-station and in Ref. 5 for multistation networks. One quadratic function, the square of workload, is used in Ref. 6 for stability and in Ref. 7 to bound performance. Many extensions are developed in Ref. 8, such as considering specific buffer priority policies, machine failure models, probabilistic routing, and general distributions. Extensions to higher-order test functions, other performance measures, and nonlinear constraints are made in Ref. 9.

Introduce the variables

$$z_{ij} = E[u_i(t)X_j(t)], \quad x_j = E[X_j(t)],$$

where the expectation is with respect to the stationary distribution for some policy u . The policy must have a unique stationary distribution and finite second moments, $E[X_i(t)^2] < \infty$, with respect to this distribution.

The logical constraint $\sum_{i:\sigma(i)=\sigma} u_i(t) \leq 1$ for each server σ leads to the constraint

$$\sum_{i:\sigma(i)=\sigma} z_{ij} - x_j \leq 0 \quad \text{for all } j, \sigma. \quad (1)$$

Let $\Delta X_i(t) = X_i(t+1) - X_i(t)$. Defining $\Delta(X_i X_j)$ similarly, under the stationary distribution, we must have

$$E[\Delta(X_i(t) X_j(t))] = 0 \quad \text{for } i \leq j.$$

This constraint can be thought of as stationarity with respect to a general quadratic test function [8, Section 2]. Now,

$$\begin{aligned} E[\Delta(X_i(t) X_j(t))] &= E(\Delta X_i(t) X_j(t)) \\ &\quad + E(X_i(t) \Delta X_j(t)) \\ &\quad + E(\Delta X_i(t) \Delta X_j(t)). \end{aligned}$$

Taking expectations with respect to the transition and then with respect to $X(t)$,

$$\begin{aligned} E(\Delta X_i(t) X_j(t)) &= \alpha_i x_j + \sum_{k=1}^n \mu_k p_{ki} z_{kj} - \mu_i z_{ij}, \\ E(\Delta X_i(t) \Delta X_j(t)) &= \begin{cases} 2\lambda_i(1 - p_{ii}) & \text{if } i = j \\ -(\lambda_i p_{ij} + \lambda_j p_{ji}) & \text{otherwise} \end{cases} \end{aligned}$$

with the conventions that $x_0 = z_{i0} = z_{0j} = 0$. The stationarity constraints, after dividing by 2 on the diagonal, are

$$\begin{aligned} \alpha_i x_i + \sum_{k=1}^n \mu_k p_{ki} z_{ki} - \mu_i z_{ii} &= -\lambda_i(1 - p_{ii}) \\ &\quad \text{for all } i, \end{aligned} \quad (2)$$

$$\begin{aligned} \alpha_i x_j + \sum_{k=1}^n \mu_k p_{ki} z_{kj} - \mu_i z_{ij} + \alpha_j x_i \\ + \sum_{k=1}^n \mu_k p_{kj} z_{ki} - \mu_j z_{ji} &= \lambda_i p_{ij} + \lambda_j p_{ji} \\ &\quad \text{for } i < j. \end{aligned} \quad (3)$$

Although our derivation assumed $E[X_i(t)^2] < \infty$, this is true for optimal policies, so the *performance* LP

$$(\text{Perf}) \quad \min_{x, z} \quad c^T x$$

Subject to: (1), (2), (3), and $z_{ij} \geq 0$

gives a lower bound on optimal average cost without a separate stability check [10, Sections 3 and 5]. A reformulation of the dual, called the *drift* LP, will allow us to compare this bound to others, so we state it here. See Ref. 11, Appendix A for a derivation.

$$(\text{Drift}) \quad \max_{q_{ij}} \sum_{i=1}^n \lambda_i \left(q_{ii} - \sum_{j=1}^n p_{ij} q_{ij} \right)$$

$$\begin{aligned} \text{Subject to: } \sum_{i=1}^n \left(\alpha_i q_{ij} + \sum_{k=0}^n u_i \mu_i p_{ik} (q_{kj} - q_{ij}) \right) \\ \geq -c_j \quad \text{for } u \in \mathcal{A} \text{ and all } j. \end{aligned} \quad (4)$$

Here, $\{q_{ij}\}$ are the dual variables for Equations (2) and (3), with the conventions that $q_{ij} = q_{ji}$ (the lower diagonal variables can be eliminated) and $q_{i0} = q_{0i} = 0$. The results in Refs 8 and 10 are stated for the special case of a reentrant line (defined in the section titled “Queueing Network Scheduling”). In this model the drift LP is

$$(\text{Reent-Drift}) \quad \max_{q_{ij}} \sum_{i=1}^n \alpha_1 (q_{ii} - q_{i,i+1})$$

$$\begin{aligned} \text{Subject to: } \alpha_1 q_{1j} + \sum_{i=1}^n u_i \mu_i (q_{i+1,j} - q_{ij}) \\ \geq -c_j \quad \text{for } u \in \mathcal{A} \text{ and all } j. \end{aligned}$$

Reentrant flow is captured through the allowable actions $\mathcal{A}$.

Many extensions to this method involve adding constraints to the performance LP. For example, if only nonidling policies are considered then $X_j(t) = \sum_{i:\sigma(i)=\sigma(j)} u_i(t)X_j(t)$ and Equation (1) is replaced by

$$x_j - \sum_{i:\sigma(i)=\sigma(j)} z_{ij} = 0 \quad \text{for all } j \quad (5)$$

$$\sum_{i:\sigma(i)=\sigma} z_{ij} - x_j \leq 0 \quad \text{for all } j, \sigma \neq \sigma(j) \quad (6)$$

To consider a specific buffer priority policy where class j is always given priority over class i by their shared station, set $z_{ij} = 0$ [8, Section 7]. Higher-order test functions are used in Ref. 9, Section 6.1 to generate additional constraints, potentially improving the bound. For example, cubic test functions $X_i(t)X_j(t)X_k(t)$ can be used to write additional constraints that include new, triply indexed variables $E[u_i(t)X_j(t)X_k(t)]$. Another advantage of adding these variables is that the performance space considered is not limited to expected queue lengths: it can include second moments, so that variance can be bounded. Once these second moment variables are included, another extension [9, Section 6.2] is to add constraints relating the first and second moments, of the form $E[Y^2] \geq E[Y]^2$, where Y is a linear combination of the $X_i(t)$. These are quadratic constraints; however, they are equivalent to the constraint that a certain matrix be positive semidefinite and can be handled efficiently by semidefinite programming.

An LP similar to (Perf) can be used to bound the throughput λ of closed reentrant lines with N customers; [8, Section 8] and [9, Section 5.2]. Noting that $\mu_i E[u_i(t)] = \lambda$ for all classes i , Equation (1) is replaced by the constraints

$$\begin{aligned} \sum_{j=1}^n \sum_{i:\sigma(i)=\sigma(j)} z_{ij} &= N, \\ \sum_{j=1}^n z_{ij} &= N \frac{\lambda}{\mu_i} \quad \text{for all } i. \end{aligned}$$

The objective is to maximize λ .

For certain queueing networks and nonidling policies, the performance LP gives

the exact achievable region: for any point (x, u) in the feasible region, x is achieved by some nonidling policy. An elegant proof is based on the connection with conservation laws (see **Conservation Laws and Related Applications**) made in Ref. 9. The quadratic test function is replaced by 2^n “potential” functions, one for each set S of classes, of the form $[f_S^T X(t)]^2$, where the vectors f_S contain the expected remaining processing time at classes in S for each class. The stationarity equations are

$$E[\Delta [f_S^T X(t)]^2] = 0. \quad (7)$$

Instead of introducing the supplementary variables z_{ij} , use the approximation that servers serve a class in S whenever they can and that classes not in S are only served when all classes in S are empty. Then Equation (7) has the form

$$\sum_{i \in S} \lambda_i f_S(i) x_i \geq b(S), \quad (8)$$

for some $b(S)$ and all S ; see Ref. 9, Section 4.1.

Now consider a single-server network, known as *Klimov’s problem* [4]. The approximations leading to Equation (8) are exact for policies that give priority to classes in S over those not in S . As a result,

$$\sum_{i \in S} \lambda_i f_S(i) x_i = b(S), \quad (9)$$

for such policies for all S . Furthermore, Equation (9) holds for $S = \{1, \dots, n\}$ for all nonidling policies. These conditions are known as *strong conservation laws*. One consequence is that Equation (8), with equality for $S = \{1, \dots, n\}$, defines the exact achievable region for nonidling policies, with vertices corresponding to buffer priority policies [12]. It is shown in Ref. 9, Section 4.2 that, for general networks, Equation (8) is a relaxation of (Perf). Thus, (Perf) gives stronger bounds than those based on Equation (8). However, for Klimov’s problem, we can conclude that (Perf) also gives the exact achievable region. The exact achievable region for Klimov’s problem was found by Tsoucas [13] using different methods.

Another implication of the conservation law approach is that the region defined by Equation (8), with equality for $S = \{1, \dots, n\}$, is an extended polymatroid that allows the LP to be easily solved using a one-pass algorithm [14, Section 2] and optimal policies to be more simply expressed as index policies [15]. Some other networks also satisfy strong conservation laws, for example, Jackson networks [16] and networks with switching times [17,18]. Certain systems nearly satisfy conservation laws, allowing bounds to be developed as in Ref. 19. In Ref. 9, the potential function method (Eq. 8) is applied to general networks using other parameters f_S ; however, as noted above, this approach gives weaker bounds than the quadratic test functions used in (Perf).

APPROXIMATE LINEAR PROGRAMMING

Approximate linear programming (ALP) was originally proposed by Schweitzer and Seidmann [20]. It starts with the LP form of the Bellman optimality equations, then approximates the differential cost by a linear combination of a small number of functions, giving a small number of variables. For quadratic approximating functions and our queueing network, the structure of the constraints allows them to be reduced to a smaller, though still exponentially large, set. This LP is only slightly larger than the achievable region LP of the section titled “Achievable Region Method” but gives a tighter bound. Let $p_u(x, y)$ be the transition probability from state x to state y under action u . Define the dynamic programming operator

$$(T_u h)(x) = c^T x + \sum_y p_u(x, y) h(y).$$

The average cost optimality equation is

$$J + h(x) = \min_{u \in \mathcal{A}(x)} (T_u h)(x) \quad \text{for all } x. \quad (10)$$

Because Equation (10) only determines h up to an additive constant, we can set $h(0) = 0$. Due to the infinite state space, additional conditions are needed for Equation (10) to have a unique solution J^*, h^* . For networks

with linear costs, the appropriate conditions are that h is bounded below by a constant and above by a quadratic; see Ref. 21, Theorem 2.1 and Section 7. Under these conditions, $h^*(x)$ is the differential cost, that is, the expected additional cost of starting in state x compared to state 0 under the optimal policy.

Any J, h , where h is bounded above and below by a quadratic, satisfying the average cost inequality

$$J + h(x) \leq (T_u h)(x), \quad \text{for all } u \in \mathcal{A}(x) \text{ and all } x \quad (11)$$

is a lower bound, $J \leq J^*$ (see Ref. 22, Appendix A). For a discussion of when this conclusion is valid for general MDPs, see Ref. 23, Section 3 and the references therein. Maximizing over J gives J^* . Now, consider the quadratic approximation $h(x) = \frac{1}{2} x^T Q x + p x$ where Q is symmetric. Substitution into Equation (11) gives an LP in the variables J, Q, p whose solution is a lower bound on J^* . There is one constraint for each state-action pair; however, because they are linear in the index x , the constraints for each u can be reduced to the constraint at the smallest x , that is, $x = u$, and a limiting constraint at each $x_i \rightarrow \infty$. The reduced ALP is

$$\begin{aligned} (\text{ALP}) \quad & \max_{J, Q, p} J \\ \text{Subject to:} \quad & J \leq d^u \quad \text{for all } u \in \mathcal{A} \\ & c_i^u \geq 0 \quad \text{for all } i \text{ for all } u \in \mathcal{A}, \end{aligned}$$

where

$$d^u = \sum_{k=1}^n \left[u_k \left(c_k^u + \mu_k \sum_{j=0}^n p_{kj} \left(\frac{1}{2} q_{kk} + \frac{1}{2} q_{jj} - q_{kj} + p_j - p_k \right) \right) + \alpha_k \left(\frac{1}{2} q_{kk} + p_k \right) \right], \quad (12)$$

$$c_i^u = c_i + \sum_{k=1}^n \left(\alpha_k q_{ik} + \sum_{j=0}^n u_k \mu_k p_{kj} (q_{ji} - q_{ki}) \right). \quad (13)$$

For a derivation, see Ref. 22, Section 4.1 or Ref. 24, Appendix A. For the special case of a reentrant line,

$$d^u = \sum_{i=1}^n \left[u_i \left(c_i^u + \mu_i \left(\frac{1}{2} q_{ii} + \frac{1}{2} q_{i+1,i+1} - q_{i,i+1} + p_{i+1} - p_i \right) \right) \right] + \alpha_1 \left(\frac{1}{2} q_{11} + p_1 \right),$$

$$c_i^u = c_i + \alpha_1 q_{1i} + \sum_{j=1}^n u_j \mu_j (q_{i,j+1} - q_{ij}).$$

Note that the second constraint in ALP is the same as the constraints (4) in (Drift). It is shown in Ref. 11 that ALP is a stronger bound than the achievable region LP; numerical tests show that the improvement can be substantial.

Stronger ALP bounds can be obtained by adding terms to the approximation architecture for h ; however, this makes Equation (11) nonlinear in x , so that it cannot be reduced to as small a set of constraints. Piecewise quadratic functions are used in Refs 24, 25, and 22, Section 4.3. Exponential and other functions are tested numerically in Ref. 22.

For a closed reentrant line, throughput λ can be bounded using a quadratic approximation [26]. The key change is to use the cost rate $\sum_i u_i(x) \mu_i$, which has expectation λ . It is also possible to obtain a throughput bound as a function of N . Choose the form of the bound to be $\lambda(N) = \alpha' N / (N + \nu)$ and also parameterize $h(x, N) = h(x) / (N + \nu)$. Letting $N \rightarrow \infty$ gives an LP that can be solved for α' . Substituting α' into the constraints and multiplying by $N + \nu$ eliminates N , yielding a second LP that can be solved for ν . Both of these bounds are tighter than the LP bounds in Ref. 8, Section 8 and Ref. 27.

LP UPPER BOUNDS

Simulating any policy gives an upper bound on J^* . However, upper bounds on classes of policies are of interest when studying stability or sensitivity of performance to the policy. A common approach is to consider all nonidling policies, for which the (worst) performance LP is

$$(\text{UB-Perf}) \max_{x, z} c^T x$$

Subject to: (2), (3), (5), (6) and $z_{ij} \geq 0$

and its dual is

$$(\text{UB-Drift}) \min_{q_{ij}} \sum_{i=1}^n \lambda_i \left(q_{ii} - \sum_{j=1}^n p_{ij} q_{ij} \right)$$

$$\text{Subject to: } \sum_{i=1}^n \left(\alpha_i q_{ij} + \sum_{k=0}^n u_i \mu_i p_{ik} (q_{kj} - q_{ij}) \right) \leq -c_j \quad \text{for all } j \text{ such that}$$

$$\sum_{i: \sigma(i)=\sigma(j)} u_i = 1, \quad u \in \mathcal{A}.$$

Note that the constraints are indexed by actions u that are nonidling at server $\sigma(j)$ and that the dual variables associated with Equations (2) and (3) are now $-q_{ij}$. Some numerical results for these upper bounds are reported in Ref. 8, Section 7. The corresponding upper bound ALP is

$$(\text{UB-ALP}) \min_{J, Q, p} J$$

Subject to: $J \geq d^u$ for all $u \in \mathcal{A}$,

$c_i^u \leq 0$ for all i such that

$$\sum_{j: \sigma(j)=\sigma(i)} u_j = 1 \text{ for all } u \in \mathcal{A},$$

using Equations (12) and (13) for d^u and c_i^u . Once again (UB-ALP) is a stronger bound than the achievable region LP [11]. Upper bound results for some ALPs are given in Ref. 24.

PATHWISE BOUNDS

Another way to obtain a lower bound is to define a simpler system $X^*(t)$ that has at least as many service completions as the original, either on a sample path or in expectation. This section describes two bounds from Ref. 28, Section 1. Sample path methods apply to general arrival and service processes. In this section, we will allow general arrival processes but still assume that service times are independently exponentially distributed.

The roughest approximation simply ignores the competition for servers by allowing all classes to be served simultaneously using a nonidling policy, which makes it a generalized Jackson network [see **Jackson Networks (Open and Closed)**]. To obtain a sample path bound, let $S_i(t)$ be a Poisson process with rate μ_i , independent for each class i , representing the number of potential class i service completions by time t . A class i service completion occurs at time t if there is a potential service completion and $u_i(t) = 1$. In the Jackson network, a service completion occurs if there is a potential service completion and $X_i^*(t) > 0$. Also, if there is probabilistic routing, a customer is assigned the same route in both systems. Customers in the Jackson network depart each class on their route at least as soon as in the original network; in particular, they depart the network at least as soon. Hence, the Jackson network gives a pathwise lower bound on the total number of customers: $\sum_i X_i^*(t) \leq \sum_i X_i(t)$ with probability 1.

Better bounds can be constructed using the workload in the system for each station, which can only be reduced at unit rate by any policy. A variant of workload is used in Ref. 28, Section 2. Assume that routing is deterministic and that service times depend only on the station, not the class, with a Poisson process $S_\sigma(t)$ with rate μ_σ representing potential service completions for station σ . Define

$$W(t) = MX(t),$$

where $M_{\sigma,i} = 1$ if class i customers require at least one more service from station σ and 0 otherwise, that is, $W_\sigma(t)$ is the number of customers in the system at t that require at least one more service from station σ before exiting. Consider a modified system in which arrivals split into one customer for each station σ that they will visit. These customers move through a series queue (no competition for servers) until they arrive at their first class served by station σ . This station operates in isolation as a Klimov model (see **Klimov's Model**): customers skip classes in their route that are not at this station. A last-buffer-first-served (LBFS) policy is used. Consider the departure process $D_\sigma^*(t)$

of customers departing the Klimov model for station σ . The LBFS policy is optimal in the sense that it gives a pathwise upper bound on $D_\sigma^*(t)$ for any service discipline. Note that the assumption that service rates are by station is needed for LBFS to be optimal.

Now compare the corresponding departure process $D_\sigma(t)$ for the original network, that is, the number of service completions by server σ by time t that are last visits by a customer to station σ . Customers in the Klimov model arrive at least as soon as customers arrive at their first class served by that station in the original network. Because it sees earlier arrivals and skips some classes, any departure process for the original process is feasible for the Klimov model, inserting idle times as needed. Hence, the departure process for the Klimov model using the LBFS policy is also a pathwise upper bound for departures from the original network: $D_\sigma^*(t) \geq D_\sigma(t)$, or equivalently, $W^*(t) \leq W(t)$ with probability 1.

Each value of $W^*(t)$ can be converted to a minimum number of customers using the LP

$$\begin{aligned} \min \quad & \sum_{i=1}^n X_i \\ \text{Subject to:} \quad & MX \geq W^*(t) \\ & X \geq 0. \end{aligned}$$

The optimal objective function value is a pathwise lower bound on $\sum_i X_i(t)$. To obtain a lower bound on the expected number of customers, one could simulate the modified system, solving the LP each time $W^*(t)$ changed, or solving the LP parametrically for all $W^*(t)$ before simulating.

LYAPUNOV FUNCTION BOUNDS

The section titled “Achievable Region Method” used equilibrium equations for quadratic test functions. A related idea is to find a Lyapunov function that satisfies a drift condition, namely, it decreases in expectation after one transition from all large states for some class of policies. Lyapunov functions are typically used to establish stability; in Ref. 29 they are used to obtain bounds. Their bounds are explicit; however, because of the

approximations involved, they are not likely to be as tight as the bounds presented in the previous sections. If the drift parameter is known, it can be used to construct an upper bound on the tail probabilities and expectation of the Lyapunov function under the stationary distribution. The first bound uses the workload of each station as a “lower” Lyapunov function, giving lower bounds on the expected workload.

For any stationary, countable-state discrete-time Markov chain with transition probabilities $p(x, y)$, Φ is a lower Lyapunov function if $\Phi(x) \geq 0$ and, for some drift parameter $\gamma > 0$,

$$\sum_y p(x, y) \Phi(y) \geq \Phi(x) - \gamma \text{ for all } x. \quad (14)$$

Note that the inequality (14) is reversed from that of a Lyapunov function. For networks,

$$\Phi(x) = W_\sigma(x) = M_\sigma x,$$

where $M_{\sigma,i}$ is the expected time station σ serves a class i customer before it leaves the system, is a lower Lyapunov function. Here $W_\sigma(x)$ is the workload for station σ , that is, the expected time for station σ to serve all customers in the system. It satisfies Equation (14) with $\gamma = 1 - \rho_\sigma$, that is, the station can only decrease its workload at rate $1 - \rho_\sigma$. Letting $\Delta\Phi(t) = \Phi(X(t+1)) - \Phi(X(t))$, the equilibrium equation $E[\Delta\Phi(t)] = 0$ and Equation (14) impose a limit on the increase in Φ . However,

$$E[\Delta\Phi(t) | \Delta\Phi(t) > 0] P(\Delta\Phi(t) > 0) \geq p_{\min} \nu_{\min},$$

where $p_{\min}$ is the smallest probability, over all states x and actions u , that $\Phi(x)$ will increase and $\nu_{\min} > 0$ is the smallest amount by which $\Phi(x)$ can increase. For reentrant lines, $p_{\min} = \alpha_1$ and $\nu_{\min} = M_{\sigma,1}$. These relationships imply a geometric lower bound on the tail probabilities and a lower bound on the expected workload at station σ

$$E[W_\sigma(X(t))] \geq \frac{\alpha_1 \rho_\sigma^2}{4(1 - \rho_\sigma)}.$$

Extensions to more general networks are given in Ref. 29 using the concept of a virtual station. Upper bounds on the expected

number of customers are also found using piecewise linear Lyapunov functions for networks with two stations.

It is sometimes possible to construct an (upper) Lyapunov function h and J satisfying Equation (11) without solving the LP of the section titled “Approximate Linear Programming.” Under the conditions on h discussed there, J gives a lower bound on J^* .

REFERENCES

1. de Farias DP, Van Roy B. The linear programming approach to approximate dynamic programming. *Oper Res* 2003;51(6):850–865.
2. de Farias DP, Weber T. State-relevance weights for the linear programming approach to dynamic programming. In: *Proceedings of the 47th IEEE Conference on Decision and Control*; 2008 Dec 9–11; Cancun, Mexico.
3. Desai VV, Farias VF, Moallemi CC. Approximate dynamic programming via a smoothed linear program. Working paper, MIT Sloan School of Management 2009.
4. Klimov GP. Time-sharing service systems. I. *Theor Probab Appl* 1975;19(3):532–551.
5. Rosberg Z. Process scheduling in a computer system. *IEEE Trans Comput* 1985;34(7):633–645.
6. Meyn SP, Down D. Stability of generalized Jackson networks. *Ann Appl Probab* 1994; 4(1):124–148.
7. Kumar PR. Re-entrant lines. *Queueing systems: special issue on Queueing Networks* 1993;13(1–3):87–110.
8. Kumar S, Kumar PR. Performance bounds for queueing networks and scheduling policies. *IEEE Trans Automat Control* 1994;39(8):1600–1611.
9. Bertsimas D, Paschaladis IC, Tsitsiklis JN. Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann Appl Probab* 1994;4(1):43–75.
10. Kumar PR, Meyn SP. Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies. *IEEE Trans Automat Control* 1996;41(4):4–17.
11. Russell MC, Fraser J, Rizzo S, *et al.* Comparing LP bounds for queueing networks. *IEEE Trans Automat Control* 2009; 54(11):2703–2707.

12. Bertsimas D, Niño-Mora J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math Oper Res* 1996; 21(2):257–306.
13. Tsoucas P. The region of achievable performance in a model of Klimov. Research Report RC16543. New York: IBM T.J. Watson Research Center, Yorktown Heights; 1991.
14. Bertsimas D. The achievable region method in the optimal control of queueing systems: formulations, bounds and policies. *Queueing Syst* 1995;21(3–4):337–389.
15. Glazebrook KD, Dacre MJ, Nino-Mora J. The achievable region approach to the optimal control of stochastic systems. *J R Stat Soc (Series B)* 1999;61(4):747–791.
16. Ross KW, Yao DD. Optimal dynamic scheduling in Jackson networks. *IEEE Trans Automat Control* 1989;34(11):47–53.
17. Bertsimas D, Nino-Mora J. Optimization of multiclass queueing networks with changeover times via the achievable region approach: part I, the single-station case. *Math Oper Res* 1999;24(2):306–330.
18. Bertsimas D, Nino-Mora J. Optimization of multiclass queueing networks with changeover times via the achievable region approach: part II, the multi-station case. *Math Oper Res* 1999;24(2):331–361.
19. Glazebrook KD, Nino-Mora J. Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance. *Oper Res* 2001; 49(4):609–623.
20. Schweitzer P, Seidmann A. Generalized polynomial approximations in Markovian decision processes. *J Math Anal Appl* 1985; 110(2):568–582.
21. Meyn SP. The policy iteration algorithm for Markov decision processes with general state space. *IEEE Trans Automat Control* 1997;42(12):191–197.
22. Veatch MH. Approximate dynamic programming for networks: fluid models and constraint reduction. Working paper. Gordon College, Department of Mathematics. 2009. Available at faculty.gordon.edu/ns/mc/Mike_Veatch.
23. Glynn PW, Zeevi A. Bounding stationary expectations of Markov processes. In: Ethier SN, Feng J, Stockbridge RH, editors. *Markov processes and related topics: a festschrift for Thomas G. Kurtz*. Beachwood (OH): Institute of Mathematical Statistics; 2008. pp. 195–214.
24. Morrison JR, Kumar PR. New linear program performance bounds for queueing networks. *J Optim Theor Appl* 1999;100(3):575–597.
25. Morrison JR, Kumar PR. Computational performance bounds for Markov chains with applications. *IEEE Trans Automat Control* 2008;53(5):1306–1311.
26. Morrison JR, Kumar PR. New linear program performance bounds for closed queueing networks. *Disc Event Dyn Syst* 2001; 11(4):291–317.
27. Jin H, Ou J, Kumar PR. The throughput of irreducible closed Markovian queueing networks: functional bounds, asymptotic loss, efficiency, and the Harrison-Wein conjectures. *Math Oper Res* 1997;22(4):886–920.
28. Ou J, Wein LM. Performance bounds for scheduling queueing networks. *Ann Appl Probab* 1992;2(2):460–480.
29. Bertsimas D, Gamarnik D, Tsitsiklis JN. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Ann Appl Probab* 2001; 11(4):1384–1428.

PHASE-TYPE (PH) DISTRIBUTIONS

RAFAEL PÉREZ-OCÓN
Department of Statistics and
Operational Research,
University of Granada,
Granada, Spain

The phase-type (PH) distributions, as they are known, were introduced by Neuts [1]. They play an important role in modeling random phenomena due to their interesting properties, and they also have a widespread use in applied probability. A PH distribution is the distribution of the absorption time in a Markov chain in continuous time with a finite-state space and one absorbent state. The class of PH distributions has important closure properties which make it very versatile in application. This class is weakly dense in the class of distribution functions (CDFs) defined on the positive real half-line, which allows us to approach general distributions by elements of this class. The use of PH distributions in stochastic modeling leads to results that can be presented in algorithmic form, and consequently, the results are susceptible to computational implementation. The PH distributions and the Markovian arrival processes are the basic elements in what is known as matrix-analytic methods. The definition and properties of PH distributions and their applications in different domains of application are presented in the following sections in a condensed form.

DEFINITION AND EXAMPLES

First, some definitions of operations related to PH distributions and frequently used in the article are presented.

Exponential of a Matrix. The matrix exponential of a finite matrix A of order $n \times n$, is defined by

$$\exp(A) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n.$$

Product and Sum of Kronecker. If A and B are rectangular matrices of orders $m \times n$ and $p \times q$ respectively, their Kronecker product, denoted by $A \otimes B$, is the matrix of order $mp \times nq$ defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

In compact form it is written as $A \otimes B = [a_{ij}B]$.

If A and B are square matrices of orders m and n , respectively, their Kronecker sum, is defined as $A \oplus B = A \otimes I_n + I_m \otimes B$, where I_j indicates the identity matrix of order j .

PH Distributions. Let a continuous-time Markov chain $\{X(t), t \geq 0\}$ be with state space $S = \{1, 2, \dots, m, m+1\}$. States $1, \dots, m$ are transient and state $m+1$ is absorbent. Ordering the states appropriately, the infinitesimal generator (or simply generator), can be written as

$$Q = \begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix}.$$

T is a matrix of order m with negative diagonal entries and nondiagonal entries that are nonnegative. Its nondiagonal entries are the transition rates among the transient states $T^0 = -Te$ (Vector e denotes a column vector with all entries equal to 1 and of order m). The eventual absorption implies that T is nonsingular.

Let α be the initial probability vector of the Markov process on the transient states, it is a row m -vector. The CDF of the time until the absorption occurs is

$$H(t) = 1 - \alpha \exp(Tt)e, t \geq 0.$$

It is said that $H(\cdot)$ is a PH distribution with representation (α, T) of order m and it is written as $\text{PH}(\alpha, T)_m$.

The quantity α_{m+1} denotes the probability that the process initiates on the absorbent state. When it is non-null, the initial vector takes the form (α, α_{m+1}) , and the distribution function $H(\cdot)$ has an atom in the origin with mass α_{m+1} . From now on, it will be considered that $\alpha_{m+1} = 0$, what is usual in applications.

The function $H(\cdot)$ is absolutely continuous on $(0, \infty)$. The probability density function $h(t)$ is

$$H'(t) = h(t) = \alpha \exp(Tt)T^0, \quad t \geq 0,$$

the hazard rate function is

$$r(t) = \frac{\alpha \exp(Tt)T^0}{\alpha \exp(Tt)e}, \quad t \geq 0.$$

The Laplace–Stieltjes transform (LST) $\phi(s)$ is a rational function, given by

$$\phi(s) = \alpha(sI - T)^{-1}T^0, \quad \text{Re}(s) \geq 0.$$

The moment of order r is

$$\mu_r = (-1)^r r! (\alpha T^{-r} e), \quad r \geq 0.$$

The representation of a PH distribution is not unique. It is denoted minimal representation the one with minor order, which is not unique either.

Examples. The generalized Erlang distribution is a PH distribution with representation

$$\alpha = (1, 0, \dots, 0),$$

$$T = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & -\lambda_{m-1} \\ 0 & 0 & 0 & \cdots & -\lambda_m \end{pmatrix}.$$

The hyperexponential distribution is a PH distribution with representation

$$\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m),$$

$$T = \begin{pmatrix} -\lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & 0 & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & -\lambda_m \end{pmatrix}.$$

PROPERTIES

The following closure properties are of interest in applications, particularly with reference to reliability. Some references related to closure properties are [2–4].

Convolution. Finite convolution of PH distributions is a PH distribution.

Suppose that $F(\cdot)$ and $G(\cdot)$ are PH distributions with representations $(\alpha, T)_m, (\beta, S)_n$ respectively. Then, the convolution of $F^*G(\cdot)$ is a PH distribution of order $m + n$ with representation (γ, L) , being $\gamma = (\alpha, 0)$ and

$$L = \begin{pmatrix} T & T^0\beta \\ 0 & S \end{pmatrix}.$$

Finite Mixture. A finite convex mixture of PH distributions is a PH distribution.

Suppose that $F_i(\cdot), 1 \leq i \leq n$, are $\text{PH}(\alpha(i), T(i))$ distributions of order m_i . Let $p_i, 1 \leq i \leq n$, be the parameters of the mixture. The mixture $p_1 F_1(\cdot) + \dots + p_n F_n(\cdot)$ is a PH distribution of order $m_1 + \dots + m_n$ with representation (δ, V) , being $\delta = (p_1 \alpha(1), \dots, p_n \alpha(n))$, and $V = \text{diag}(T(i))$.

Maximum and Minimum. The maximum and minimum of PH distributions are PH distributions.

We denote by $\otimes$ and $\oplus$ the Kronecker product and sum, respectively. Suppose two random variables X and Y with PH distributions $F(\cdot)$ and $G(\cdot)$ with representations $(\alpha, T)_m, (\beta, S)_n$ respectively. Then, $A = \min\{X, Y\}$ and $B = \max\{X, Y\}$ have the respective representations $(\gamma(1), L(1))_{mn}$ and $(\gamma(2), L(2))_{mn+m+n}$, given by

$$\gamma(1) = \alpha \otimes \beta, L(1) = T \otimes I + I \otimes S = T \oplus S,$$

$$\text{and } \gamma(2) = (\alpha \otimes \beta, 0, 0),$$

$$L(2) = \begin{pmatrix} T \oplus S & I \otimes S^0 & T^0 \otimes I \\ 0 & T & 0 \\ 0 & 0 & S \end{pmatrix}.$$

Coherent Systems. A coherent system with finite independent units following PH distributions has a lifetime that follows a PH distribution.

This result is proved in [3]. New closure properties have been established [see [5,6]].

Characterization Properties

These distributions have been characterized by means of the LST and for its closure properties.

Characterization in Terms of the Laplace–Stieltjes Transforms

A distribution is a PH distribution if and only if it has the following:

- A continuous probability density function on the positive real half-line, with a possible mass at 0.
- An LST rationale.
- LST has a unique pole of maximal real part.

This result is proved in [7], and it is a versatile modeling tool. The LST of a PH distribution is rational with this characterization property.

The class of distributions with LST rational has been characterized in [8]. They are named *matrix-exponential distributions* (ME). The analytic expression of these distributions is formally very similar to the distribution of the PH ones, but the initial vector is not a probability vector, and matrix T is not a generator. The interpretation of the phases is not valid, since there is no underlying Markovian structure. In Éltetô *et al.* [9] the coefficient of variation of the ME distributions is studied in order to be minimized.

Characterization in Terms of Closure Properties

The PH distributions are the smallest class of distributions defined on the positive real half-line which have the following properties:

- They contain the point mass at zero and all exponential distributions.
- They are closed under finite mixtures and convolutions.
- They are closed under the infinite mixture.

$$G(\cdot) = \sum_{n \geq 0} (1-p)p^n F^{*n}(\cdot), \quad \text{for } 0 \leq p \leq 1.$$

This result is proved in [4]. In Latouche and Ramaswami [10], chapters 1, 2 are dedicated to the PH distributions, and basic ideas and algorithms of the matrix-analytic theory in the discrete and continuous cases are presented

Approximation Properties

The following results related to approximating properties of PH distributions are proved in [11]: The family of PH distributions is weakly dense in the family of distributions defined on the positive real half-line. There are several families that play an important role in approximating general distributions.

- *FMER class.* Finite mixtures of Erlang distributions with the same intensity parameter.
- *(S/P) class.* Series-parallel distributions (exponential distributions in series and/or parallel).
- *ME class.* Matrix-exponential distributions.

The following relations are given among these classes: $\text{FMER} \subset \text{S/P} \subset \text{PH} \subset \text{ME}$. Consequently, all these classes are weakly dense in the family of distributions defined on the positive real half-line. A procedure for fitting PH distributions to general ones and to a data-set using the EM (expectation-maximization) algorithm is given in [12].

Results of combining fitting of moments and some other features, such as percentiles, by techniques of nonlinear optimization are in [13–15]. In Lang and Arthur [16], different techniques for approximating parameters for PH distributions are applied. In Schmickler [17] mixed Erlang distributions are used for fitting empirical distributions. A survey on

statistical estimation on PH distributions including different computational programs for fitting these distributions to a data-set and to other distribution is done in [18]. In Asmussen *et al.* [19] is presented a computational program based on the EM algorithm for fitting PH distributions. In Thümmeler *et al.* [20] an approach is presented for fitting mixtures of Erlang distributions to a data-set and constructing an algorithm of the EM type.

PH-RENEWAL PROCESSES

An interesting procedure used in applications is the construction of the renewal process associated with a PH distribution. This is a renewal process in which the interarrival times are PH distributed. Let $F(\cdot)$ be a $\text{PH}(\alpha, T)_m$ distribution with absorbent state $m+1$. Every time that the state $m+1$ is reached, the Markov process with generator Q associated with the PH distribution returns instantaneously to a transient state following vector α . Considering that every visit to the state $m+1$ is a renewal, we have a PH-renewal process with underlying distribution $F(\cdot)$. The renewals are governed by a Markov process $\{J(t), t \geq 0\}$ with generator $Q^* = T + T^0\alpha$, with m states.

The renewal density is $f(t) = \alpha \exp(Q * t)T^0, t \geq 0$.

The stationary probability vector $\pi = (\pi_1, \dots, \pi_m)$ associated with Q^* satisfies $\pi Q^* = 0, \pi e = 1$, and it is $\pi = (\alpha T^{-1}e)^{-1}(\alpha T^{-1})$.

The mean value of $F(\cdot)$ is μ , and it satisfies $\mu^{-1} = \pi T^0$.

The limiting distribution of the PH-renewal process, defined by

$$F * (x) = \frac{1}{\mu} \int_0^x [1 - F(u)] du \quad \text{for } x \geq 0,$$

is a $\text{PH}(\pi, T)$ distribution.

The expected number of renewals in $(0, t]$ is

$$\begin{aligned} EN(t) &= \frac{t}{\mu} + \frac{\sigma^2 + \mu^2}{2\mu^2} \\ &+ \frac{1}{\mu} \alpha [\Pi - \exp(Q * t)] T^{-1} e, \end{aligned}$$

where $\Pi = \text{diag}(\pi_1, \dots, \pi_m)$ and σ^2 is the variance of $F(\cdot)$.

The residual lifetime, that is, the time between a time t and the next renewal, follows a $\text{PH}(\gamma, T)$ distribution with $\gamma = \alpha \exp(Q * t)$.

Given that $\exp(Q * t) \rightarrow e\pi$ for $t \rightarrow \infty$, the density of renewal approximates to $\pi T^0 = \mu^{-1}$, the inverse of the mean of the distribution $F(\cdot)$.

In this renewal process, the number of renewals in $(0, t]$ is calculated as follows. We introduce the matrices $P(n, t) = (P_{ij}(n, t)), n \geq 0$, whose entries are

$$\begin{aligned} P_{ij}(n, t) &= P\{N(t) = j, J(t) \\ &= j | N(0) = 0, J(0) = i\}, \end{aligned}$$

for $t \geq 0, 1 \leq i \leq m, 1 \leq j \leq m$. These matrices satisfy the forward Kolmogorov equations:

$$\begin{aligned} P'(0, t) &= TP(0, t), \\ P'(n, t) &= TP(n, t) + P(n-1, t)T^0\alpha, n \geq 1, \\ P(n, 0) &= \delta_{0n}I \text{ (initial condition)}, \end{aligned}$$

with δ being the delta of Kronecker. The solution of this matrix equation system is, in general, a complex problem which implies algorithmic procedures. These previous results are established in [2], including an algorithm to calculate the matrices $P(n, t)$. This renewal process can be applied to general conditions using the approximation properties of the PH distributions.

DISCRETE PH DISTRIBUTIONS

The corresponding discrete class for these distributions is defined considering a discrete-time Markov chain $\{X_n, n = 0, 1, \dots\}$ with state space $S = \{1, 2, \dots, m, m+1\}$ and transition matrix P of the form

$$P = \begin{pmatrix} T & T^0 \\ 0 & 1 \end{pmatrix}.$$

T is a substochastic matrix of order m and $I - T$ is nonsingular. If α is the initial m -vector of the Markov chain, the probability density p_k is given by

$$p_0 = 0, \\ p_k = \alpha T^{k-1} T^0, \quad \text{for } k \geq 1.$$

As in the continuous case, it is assumed that the process is not initiated in the absorbent state.

The probability generating function is

$$P(z) = z\alpha(I - zT)^{-1}T^0$$

and the factorial moment of order k ($k = 1, 2, \dots$) is

$$d^k P(z)/dz^k|_{z=1} = k!\alpha T^{k-1}(I - T)^{-k}e.$$

The geometric distribution is the discrete analog to the exponential distribution, and the binomial negative distribution is the discrete distribution corresponding to the Erlang one.

Properties similar to the ones in the continuous case are valid for the discrete PH distributions. In [9], some closure properties are proved for the discrete case.

In the continuous case, any distribution defined on the real half-line can be approximated by a PH one, but there are distributions that are not of this class. For the discrete case, the following results establish that any discrete distribution with finite support is a PH distribution.

General Property. Any probability distribution with finite support is a discrete PH distribution [1]. The representation for any discrete distribution is constructed in [21]. We give two closure properties involving discrete and continuous PH distributions and that are of interest in applications.

Infinite Mixture. Consider the mixture

$$G(\cdot) = \sum_{k \geq 0} p_k F^{*k}(\cdot),$$

where $\{p_k\}$ is a discrete $\text{PH}(\beta, S)_n$ distribution, $F(\cdot)$ a continuous $\text{PH}(\alpha, T)_m$ distribution, and $F^{*k}(\cdot)$ denotes the k -fold convolution of $F(\cdot)$ with itself. Distribution $G(\cdot)$ is a $\text{PH}(\gamma, L)_{mn}$ distribution with

$$\gamma = \alpha \otimes \beta \text{ and } L = T \otimes I + T^0 \alpha \otimes S.$$

This result is used in queueing theory, for calculating the waiting time distribution in the model M/G/1 in stationary regime [2]. Infinite mixtures of PH-continuous distributions are not in general PH distributions.

Shock Model. Consider a device subject to shocks. Every shock can produce the failure of the device, $\bar{P}_k$ being the probability of surviving to the k th shock. If $N(t)$ is the random number of shocks in $(0, t)$, the survival function of the model is

$$\bar{H}(t) = \sum_{k=0}^{\infty} P\{N(t) = k\} \bar{P}_k, \quad t \geq 0.$$

The probability density $\{p_k\}$ of the number of shocks that can cause the device to fail is represented by

$$\bar{P}_k = \sum_{r=k+1}^{\infty} p_r.$$

This is a classic model studied in [22] under nonparametric methodology when $N(t)$ is the Poisson process. Many works were dedicated to the preservation of the reliability classes, that is, to show that $H(t)$ belonged to a continuous reliability class if $\{p_k\}$ belonged to the corresponding discrete class.

If the shocks occur according to a PH-renewal process with the underlying representation $\text{PH}(\alpha, T)_m$ distribution and it starts according to an initial probability vector η , and the probability density $\{p_k\}$ is a discrete $\text{PH}(\beta, S)_n$ distribution, then the distribution function $H(\cdot)$ is a $\text{PH}(\gamma, L)_{mn}$ with

$$\gamma = \eta \otimes \beta, L = T \otimes I + T^0 \alpha \otimes S.$$

This result is proved in [23], and it is the first time that an explicit form of the distribution function $H(\cdot)$ is determined. In a certain sense, it is a closure property for the PH distributions. Two shock models under inspections and perfect repair involving PH distributions are studied in [24]. A shock-and-wear model under policy N involving PH distributions is studied in [25].

APPLICATIONS OF PH DISTRIBUTIONS

The PH distributions are playing a central role in stochastic modeling. There are several practical reasons for using them: the Markovian structures when they are used, the closure properties, and the approximating properties.

The PH distribution are based in the exponential distribution by means of the phases, and so the no-memory property can be included in the study. But this implies that the number of states increases and the matrices are of high order. So, algorithmic procedures and computational treatment of the results are considered in applications. In Li and Cao [26] is presented a factorization method for reducing the order of the matrices.

The closure properties are very useful in modeling. In reliability systems with multiple units under repair, or in queueing systems, these properties allow operating with the same class of distributions, simplifying the calculations. The model that governs the systems is a Markov process though with vector state space, and matrices play a central role in the study. This shows that PH distributions take an essential part in the matrix-analytic methods. In Ref. 2, these methods are applied to the queue PH/PH/1, and in [27] to a reliability system with units in cold standby. In both cases, the process governing the system has the same generator, showing the versatility of the methods.

The inclusion of phases allows introduction of memory in the behavior of the systems, as in [28]. In general, the phases in the representation of a PH distribution do not have a physical significance. However, in reliability, the phases can be interpreted as different steps in a degradation [29] and/or recovery process, and also for indicating the design and control of the repair [30].

A survey with examples where these distributions are applied in different domains such as engineering, social sciences, and survival is in [31]. The PH distributions have been extensively applied in queueing theory, as in [2,8–10,32,33] and references therein. Formulas for calculating the stationary distribution in matrix-geometric form are calculated in [34]. The first paper considering

PH distributions in the study of reliability systems is [35], and from then, many papers have been published, as in [24,25,28] and references therein. They are being applied in other domains; in [36] PH distributions are applied to medical, pharmacological or societal data-sets.

The approximation of the transition functions in a semi-Markov process by PH distributions is studied in [37], but a systematic treatment of semi-Markov processes under PH distributions is not performed.

Further Properties and Extensions

PH distributions have been presented from the point of view of their applicability in constructing and solving models. Some theoretical elements completing the previous results, such as the irreducibility of the generator and a discussion about the minimal representation, are studied in [2]. In Aldous and Shepp [38], it is proved that the order of a PH distribution is at least as large as the inverse of its coefficient of variation -1 . An upper triangular representation for a particular subclass of PH distributions is constructed in [39]. Many other properties are discussed in other papers; see Refs 39–41, among others.

The extension to the multidimensional case of the univariate PH distributions has been considered in applications to reliability [40], and in systems with dependent components [41].

In Shi and Guo [42] the extension of the PH distributions to the case of infinite phases is studied; this new class is denoted by SPH, and certain approximation properties are established.

REFERENCES

1. Neuts MF. Probability distributions of phase type. Liber Amicorum. Prof. emeritus H. Florin University of Louvain. Belgium: Department of Mathematics; 1975.
2. Neuts MF. Matrix-geometric solutions to stochastic models: an algorithmic approach. Baltimore (MD): Johns Hopkins University Press; 1981.
3. Assaf D, Levikson B. Closure of phase-type distributions under operation arising in reliability theory. *Ann Probab* 1982;10:265–269.

4. Maier RS, O'Cinneide CA. A closure characterization of phase-type distributions. *J Appl Probab* 1992;29:92–103.
5. Bobbio A, Trivedi KS. Computation of the distribution of the completion time when the work requirement is a Ph random variable. *Stoch Model* 1990;6:133–150.
6. Neuts MF. Two further closure properties of PH-distributions. *Asia Pac J Oper Res* 1992;9:77–85.
7. O'Cinneide CA. Characterization of phase-type distributions. *Stoch Models* 1990;6:1–57.
8. Asmussen S, Bladt M. Renewal theory and queuing algorithms for matrix-exponential distributions. In: Chakravarty AR, Alfa A, editors. *Matrix-Analytic methods in stochastic models*. New York: Marcel Dekker; 1996. pp. 313–341.
9. Éltető T, Rácz S, Telek M. Matrix exponential distributions with minimal coefficient of variation M. *Madrid Conference on Queueing Theory*; 2006 July; Madrid. 2006
10. Latouche G, Ramaswami V. *Introduction to matrix analytic methods in stochastic modeling*. New York: ASA-SIAM; 1999.
11. Asmussen S. *Applied probability and queues*. New York: Springer; 2003.
12. Asmussen S, Nerman O, Olson M. Fitting phase-type distributions via the EM algorithm. *Scand J Stat* 1996;23:419–441.
13. Johnson MA, Taaffe MR. Matching moments to phase distributions: mixtures of Erlang distributions of common order. *Stoch Models* 1989;5:711–743.
14. Johnson MA, Taaffe MR. Matching moments with a class of phase distributions: Nonlinear programming approaches. *Stoch Models* 1990;6:259–281.
15. Johnson MA, Taaffe MR. Matching moments with a class of phase distributions: Density function shapes. *Stoch Models* 1990;6:283–306.
16. Lang A, Arthur JL. Parameter approximation for phase-type distributions. In: Chakravarty AR, Alfa A, editors. *Matrix-analytic methods in stochastic models*. New York: Marcel Dekker; 1996. pp. 151–206.
17. Schmickler L. Mixed Erlang distributions as phase type representations of empirical distribution functions. *Stochastic Models* 1992;8:131–156.
18. Asmussen S. Phase-type distributions and related point processes: fitting and recent advances. In: Chakravarty AR, Alfa A, editors. *Matrix-analytic methods in stochastic models*. New York: Marcel Dekker; 1996. p 137–149.
19. Häggström O, Asmussen S, Nerman O. EMPHT—A program for fitting phase-type distributions. Technical Report. Department of Mathematics. Göteborg, Sweden: Chalmers University of Technology; 1992.
20. Thümmel A, Buchholz P, Telek M. A novel approach for fitting probability distributions to trace data with the EM algorithm. *International Conference on Dependable Systems and Networks (DSN)*; Washington, DC: IEEE Computer Society; 2005. pp. 712–721.
21. Alfa AS, Castro IT. Discrete time analysis of a repairable machine. *J Appl Probab* 2002;3: 503–516.
22. Esary JD, Marshall AW, Proschan F. Shock models and wear processes. *Ann Probab* 1973; 1:627–649.
23. Neuts MF, Bhattacharjee MC. Shock models with phase type survival and shock resistance. *Nav Res Log* 1981;28(2):13–219.
24. Frostig E, Kenzin M. Availability of inspected systems subject to shocks: A matrix algorithmic approach. *Eur J Oper Res* 2009;193: 168–183.
25. Pérez-Ocón R, Montoro-Cazorla D, Segovia MC. Shock and wear models under policy N using phase-type distributions. *Appl Math Model* 2009;33:543–554.
26. Li Q-L, Cao J. Two types of RG-factorization of quasi-birth-and-death processes and their applications to stochastic integral functionals. *Stoch Models* 2004;20(3):299–340.
27. Pérez-Ocón R, Montoro-Cazorla D. A multiple system governed by a quasi-birth-and-death process. *Reliab Eng Syst Saf* 2004;84: 187–196.
28. Neuts MF, Pérez-Ocón R, Torres-Castro I. Repairable models with operating and repair times governed by phase type distributions. *Adv Appl Probab* 2000;34:468–479.
29. Pérez-Ocón R, Montoro-Cazorla D. Transient analysis of a repairable system, using phase-type distributions and geometric processes. *IEEE Trans Reliab* 2004;53(2):185–192.
30. Pérez-Ocón R, Ruiz-Castro JE. Two models for a repairable two-system with phase-type sojourn time distributions. *Reliab Eng Syst Saf* 2004;84:253–260.
31. Aalen OO. On phase type distributions in survival analysis. *Scand J Stat* 1995;22:27–463.
32. Artalejo JR, Economou A, Gómez-Corral A. Applications of maximum queue lengths to

- call center management. *Comput Oper Res* 2007;34(4):983–996.
33. Albores FX, Bocharov P. Two finites queues with relative priority in a single server system with phase-type distributions. *Automatica I Telemekhanika* 1993;4:96–107.
 34. Bocharov P, Naoumov V. Matrix-geometric stationary distribution for the PH/PH/1/r queue. *J Inf Proc Cybern* 1986;2:179–186.
 35. Neuts MF, Meier K. On the use of phase type distribution in reliability modeling of systems with two components. *OR Spektrum* 1981;2: 227–234.
 36. Faddy MJ. A structured compartmental model for drug kinetics. *Biometrics* 1993;49: 243–248.
 37. Limnios N, Oprisan G. *Semi-Markov processes and reliability*. Boston: Birkhauser; 2001.
 38. Aldous D, Shepp L. The least variable phase-type distribution is Erlang. *Stoch Models* 1987;3:467–473.
 39. Éltetô T, Vadera P. Finding upper-triangular representations for phase-type distributions with 3 distinct real poles. *Ann Oper Res* 2008; 160:139–172.
 40. Assaf D, Langberg NA, Savits TH, *et al*. Multivariate phase type distributions. *Oper Res* 1984;32:688–702.
 41. Cui L, Li H. Analytical method for reliability and MTTF assessment of coherent systems with dependent components. *Reliab Syst Saf Syst* 2007;92:300–307.
 42. Shi DH, Guo JL. SPH-distributions and the rectangle-iterative algorithm. In: Chakravarty AR, Alfa A, editors. *Matrix-analytic methods in stochastic models*: Marcel Dekker; 1996. pp. 207–224

POINT AND INTERVAL AVAILABILITY

TERJE AVEN

Department of Industrial Economics,
Risk Management, and Planning,
University of Stavanger,
Stavanger, Norway

INTRODUCTION

We study a system subject to random deterioration and failure. At failures, repairs or replacements are performed, and the system resumes operation. To characterize the performance of the system, different measures can be used. In this article, we address the following two measures:

- The probability that the system is functioning at a certain point in time (point availability).
- The probability that the system is functioning during a specified interval $[u,v]$ (interval availability).

Traditionally, focus has been placed on analyzing the point availability and its limit (the steady-state availability). For a single component the steady-state formula is given by $MTTF/(MTTF + MTTR)$, where MTTF and MTTR represent the mean time to failure and the mean time to repair (mean repair time), respectively. The steady-state probability of a system comprising several components can then be calculated using the theory of complex (monotone) systems.

Compared to the limiting point availability it is of course more complicated to compute the performance measures related to a time interval. Using simplifications and approximations, it is, however, possible to establish formulae that can be used in practice.

In this article, we will look closer at the above availability measures. We establish methods and formulae for computing these

measures. We also present some asymptotic results, for large time intervals and for systems having highly available components. We first address the one-component case, then more complex systems.

Our focus are situations where the system is being repaired or replaced immediately at failures. Safety systems where the possible failures are hidden until the system is activated or tested will not be covered. We refer to Rausand and Høyland [1] for a review of availability models applicable in such cases.

Our main reference is Aven and Jensen [2], which also covers analysis of other performance measures than the availability measures, including measures related to the number of system failures. Two key references to the classical availability theory are Barlow and Proschan [3,4].

BASIC SETUP

Let X_t be the state of the system, defined by $X_t = 1$ if the system is functioning at time t and 0 otherwise. Furthermore, let T_k , $k \in \mathbb{N}$ (here $\mathbb{N} = \{1, 2, \dots\}$) represent the length of the k th operation period, and let R_k , $k \in \mathbb{N}$, represent the length of the k th repair/replacement time for the system (Fig. 1). We assume that (T_k) , $k \in \mathbb{N}$, and (R_k) , $k \in \mathbb{N}$, are independent i.i.d. sequences of positive random variables. We denote the probability distributions of T_k and R_k by F and G , respectively, and assume that they have finite means μ_F and μ_G , respectively.

In reliability engineering, μ_F and μ_G are referred to as the *mean time to failure* (MTTF) and the *mean time to repair* (MTTR), respectively.

To simplify the presentation, we also assume that F is an absolutely continuous distribution, that is, F has a density function f and failure rate function λ . We do not make the same assumption for the distribution function G , since that would exclude discrete repair time distributions that are often used in practice.

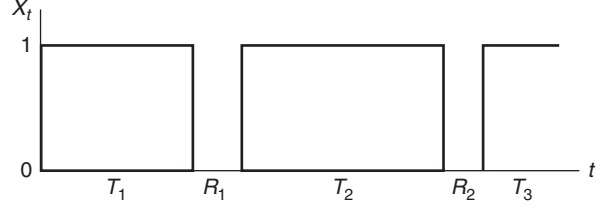


Figure 1. Time evolution of a failure and repair process for a one-component system starting at time $t = 0$ in the operating state.

In some cases, we also need the variances of F and G , denoted as σ_F^2 and σ_G^2 , respectively. In the following, when writing the variance of a random variable, it is tacitly assumed that these are finite.

The sequence $T_1, R_1, T_2, R_2, \dots$ forms an alternating renewal process. We introduce the following variables:

$$S_n = T_1 + \sum_{k=1}^{n-1} (R_k + T_{k+1}), \quad n \in \mathbb{N}$$

and

$$S_n^\circ = \sum_{k=1}^n (T_k + R_k), \quad n \in \mathbb{N}.$$

By convention, $S_0 = S_0^\circ = 0$ and sums over empty sets are zero. We see that S_n represents the n th failure time, and S_n° represents the completion time of the n th repair.

The S_n sequence generates a modified (delayed) renewal process N with renewal function M . The first interarrival time has distribution F . All other interarrival times have distribution $F * G$ (the convolution of F and G , defined by $F * G(t) = \int_0^t F(t-s) dG(s)$), with mean $\mu_F + \mu_G$. Let $H^{(n)}$ denote the distribution function of S_n . Then,

$$H^{(n)} = F * (F * G)^{(n-1)},$$

where B^{*n} denotes the n -fold convolution of a distribution B and as usual B^{*0} equals the distribution with mass of 1 at 0. Note that we have

$$M(t) = \sum_{n=1}^{\infty} H^{(n)}(t),$$

as $N_t = \sum_{n=1}^{\infty} I(S_n \leq t)$, where I is the indicator function. The S_n° sequence generates an ordinary renewal process N° with renewal function M° . The interarrival times, $T_k + R_k$,

have distribution $F * G$, with mean $\mu_F + \mu_G$. Let $H^{(n)}$ denote the distribution function of S_n . Then,

$$H^{(n)} = (F * G)^{*n}.$$

Let α_t denote the forward recurrence time at time t , that is, the time from t to the next event: $\alpha_t = S_{N_t+1} - t$ on $\{X_t = 1\}$ and $\alpha_t = S_{N_t^\circ+1}^\circ - t$ on $\{X_t = 0\}$. Hence, given that the system is up at time t , the forward recurrence time α_t equals the time to the next failure time. If the system is down at time t , the forward recurrence time equals the time to complete the repair. Let F_{α_t} and G_{α_t} denote the conditional distribution functions of α_t given that $X_t = 1$ and $X_t = 0$, respectively. Then, we have for $x \in \mathbb{R}$ (here $\mathbb{R} = (-\infty, \infty)$)

$$\begin{aligned} F_{\alpha_t}(x) &= P(\alpha_t \leq x | X_t = 1) \\ &= P(S_{N_t+1} - t \leq x | X_t = 1) \end{aligned}$$

and

$$\begin{aligned} G_{\alpha_t}(x) &= P(\alpha_t \leq x | X_t = 0) \\ &= P(S_{N_t^\circ+1}^\circ - t \leq x | X_t = 0). \end{aligned}$$

Similarly for the backward recurrence time, we define β_t, F_{β_t} , and G_{β_t} . The backward recurrence time β_t equals the age of the system if the system is up at time t and the duration of the repair if the system is down at time t , that is, $\beta_t = t - S_{N_t}^\circ$ on $\{X_t = 1\}$ and $\beta_t = t - S_{N_t}$ on $\{X_t = 0\}$.

POINT AVAILABILITY

The point availability $A(t)$, defined as $A(t) = P(X_t = 1)$, is given as

$$\begin{aligned} A(t) &= \bar{F}(t) + \int_0^t \bar{F}(t-x) dM^\circ(x) \\ &= \bar{F}(t) + \bar{F} * M^\circ(t), \end{aligned} \quad (1)$$

where $\bar{F} = 1 - F$. This follows by using a standard renewal argument. Conditioning on the duration of $T_1 + R_1$, it is not difficult to see that $A(t)$ satisfies the following equation:

$$A(t) = \bar{F}(t) + \int_0^t A(t-x) d(F * G)(x).$$

Hence, by the renewal equation (2, p. 245) Formula 1 follows. Alternatively, we may use a more direct approach, writing

$$X_t = I(T_1 > t) + \sum_{n=1}^{\infty} I(S_n^{\circ} \leq t, S_n^{\circ} + T_{n+1} > t),$$

which gives Equation (1) by taking expectations. The point unavailability $\bar{A}(t)$ is given by $\bar{A}(t) = 1 - A(t) = F(t) - \bar{F} * M^{\circ}(t)$.

Calculating the availability at a given point in time t , $A(t)$ is difficult except for a few special cases. If, for example, the lifetime and repair times are independent and they have exponential distributions, it can be shown using Markov theory [1] that the availability of the system at time t equals

$$\frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t},$$

where λ is the failure rate and μ is the repair rate of the unit. It is assumed that the unit is functioning at time zero.

By the Key Renewal Theorem (2, p. 247), it follows that

$$\lim_{t \rightarrow \infty} A(t) = \frac{\mu_F}{\mu_F + \mu_G}. \quad (2)$$

The right-hand side of Equation (2) is called the *limiting availability* (or steady-state availability) and is for short denoted A . The *limiting unavailability* is defined as $\bar{A} = 1 - A$. Usually μ_G is small compared to μ_F , so that

$$\bar{A} \approx \frac{\mu_G}{\mu_F}.$$

If the system has an exponential lifetime distribution with failure rate λ , it can be shown that

$$\bar{A}(t) \leq \lambda \int_0^t \bar{G}(s) ds \leq \lambda \mu_G,$$

Ref. 2, p. 107.

INTERVAL AVAILABILITY

The interval availability $A[u, v]$ is defined as $A[u, v] = P(X_t = 1, \forall t \in [u, v])$. Using that $A[u, v] = A(u)P(\alpha_u > v - u | X_u = 1)$, we see that

$$A[u, v] = \bar{F}_{\alpha_u}(v - u)A(u).$$

The interval availability for the interval $[t, t + w]$, that is, the probability that the system is up at time t and the forward recurrence time at time t is greater than w is given by

$$\begin{aligned} A[t, t + w] &= P(X_t = 1, \alpha_t > w) \\ &= \bar{F}(t + w) + \int_0^t \bar{F}(t - x + w) dM^{\circ}(x). \end{aligned}$$

This is shown by noting that

$$X_t I(\alpha_t > w) = \sum_{n=0}^{\infty} I(S_n^{\circ} \leq t, S_n^{\circ} + T_{n+1} > t + w).$$

By applying the Key Renewal Theorem (2, p. 247), it follows that

$$\lim_{t \rightarrow \infty} A[t, t + w] = \frac{\int_0^w \bar{F}(x) dx}{\mu_F + \mu_G}.$$

The limit $\int_0^w \bar{F}(x) dx / (\mu_F + \mu_G)$ is the asymptotic distribution of the forward recurrence time α_t and is denoted $F_{\infty}(w)$. Similarly we define the asymptotic distribution $G_{\infty}(w)$ of the backward recurrence time β_t .

STEADY-STATE DISTRIBUTION

The asymptotic results established above provide good approximations for the performance measures related to a given point in time or an interval. Based on the asymptotic values, we can define a stationary (steady-state) process having these asymptotic values as their distributions and means. To define such a process in our case, we generalize the model analyzed above by allowing X_0 to be 0 or 1.

Thus the time evolution of the process is, as shown in Fig. 1, beginning with an uptime, or similarly starting with a downtime.

The process is characterized by the parameters $A(0)$, $F^*(t)$, $F(t)$, $G^*(t)$, $G(t)$, where $F^*(t)$ denotes the distribution of the first uptime provided the system starts in state 1 at time 0 (i.e., $X_0 = 1$) and $G^*(t)$ denotes the distribution of the first downtime provided the system starts in state 0 at time 0 (i.e., $X_0 = 0$). Now assuming that $F^*(t)$ and $G^*(t)$ are equal to the asymptotic distributions of the recurrence times, that is, $F_\infty(t)$ and $G_\infty(t)$, respectively, and $A(0) = A$, it can be shown that the process (X_t, α_t) is stationary [5]. This means that we have, for example,

$$\begin{aligned} A(t) &= A, \quad \forall t \in \mathbb{R}_+ \\ A[u, u+w] &= \frac{\int_w^\infty \bar{F}(x) dx}{\mu_F + \mu_G}, \\ \forall u, w &\in \mathbb{R}_+ = [0, \infty). \end{aligned}$$

MONOTONE SYSTEMS

Consider now a binary monotone system comprising n independent components. For each component, we define a model as in the section titled “Basic Setup,” indexed by “ i .” The uptimes and downtimes of component i are thus denoted T_{ik} and R_{ik} with distributions F_i and G_i , respectively. The lifetime distribution F_i is absolutely continuous with a failure rate function $\lambda_i(t)$. Let $\Phi : \{0, 1\}^n \rightarrow \{0, 1\}$ be the structure function of the system, which is assumed to be monotone, that is, the structure function Φ is non-decreasing in each argument, and $\Phi(\mathbf{0}) = 0$ and $\Phi(\mathbf{1}) = 1$. The possible states of the system, “functioning” and “not functioning,” are indicated by “1” and “0,” respectively, as for the components. Then $\Phi_t = \Phi(\mathbf{X}_t)$ describes the state of the system at time t , where $\mathbf{X}_t = (X_t(1), \dots, X_t(n))$.

Let p_i denote the reliability of component i , that is, $p_i = P(X_i = 1)$. Then, if the components (i.e., the X_i s) are independent, the reliability of the system, $P(\Phi = 1)$, is a function of the p_i s. This function is referred to as the *reliability function* and is denoted $h(\mathbf{p})$, where $\mathbf{p} = (p_1, p_2, \dots, p_n)$.

Point Availability

The following results show that the point availability (limiting availability) of a monotone system is equal to the reliability function h with the component reliabilities replaced by the component availabilities $A_i(t)$ (A_i).

Theorem 1. *The system availability at time t , $A(t)$, and the limiting system availability, $\lim_{t \rightarrow \infty} A(t)$, are given by*

$$A(t) = h(A_1(t), A_2(t), \dots, A_n(t)) = h(\mathbf{A}(t)), \quad (3)$$

$$\lim_{t \rightarrow \infty} A(t) = h(A_1, A_2, \dots, A_n) = h(\mathbf{A}). \quad (4)$$

Formula 3 is simply an application of the reliability function formula with $A_i(t) = P(X_t(i) = 1)$. Since the reliability function $h(\mathbf{p})$ is a linear function in each p_i and therefore a continuous function, it follows that $A(t) \rightarrow h(A_1, A_2, \dots, A_n)$ as $t \rightarrow \infty$, which proves Equation (4).

The limiting system availability can also be interpreted as the expected proportion of time the system is operating in the long run, or as the long run average availability noting that

$$\begin{aligned} \lim_{t \rightarrow \infty} E \left[\frac{1}{t} \int_0^t \Phi_s ds \right] &= \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s) ds \\ &= \lim_{t \rightarrow \infty} A(t). \end{aligned}$$

Interval Availability

In general, it is difficult to calculate the interval availability for a monotone system. Only in some special cases it is possible to obtain practical computation formulae, and in the following we look closer into some of them.

If the repair times are small compared to the lifetimes and the lifetimes are exponentially distributed with parameter λ_i , then clearly the number of failures of component i in the time interval $(u, u+w]$, $N_{u+w}(i) - N_u(i)$, is approximately Poisson distributed with parameter $\lambda_i w$. Hence the interval availability $A[0, u]$, which is equal to $P(N_u = 0)$, is approximately exponentially distributed.

If the system is a series system, and we make the same assumptions as above, it is also clear that the number of system failures in the interval $(u, u + w]$ is approximately Poisson distributed with parameter $\sum_{i=1}^n \lambda_i w$. The number of system failures in $[0, t]$, N_t , is approximately a Poisson process with intensity $\sum_{i=1}^n \lambda_i$.

If the system is highly available and the components have constant failure rates, the Poisson distribution will in fact produce good approximations also for more general systems. To motivate this, we observe that the expected number of system failures per unit of time during the interval $(u, u + w]$, $EN_{(u, u+w]}/w$, is approximately equal to the asymptotic system failure rate λ_Φ , and $N_{(u, u+w]}$ is “nearly independent” of the history of N up to u , noting that the process $\mathbf{X}$ frequently restarts itself probabilistically [i.e., $\mathbf{X}$ reenters the state $(1, 1, \dots, 1)$]. The system failure rate λ_Φ is given by

$$\lambda_\Phi = \lim_{u \rightarrow \infty} \frac{EN_u}{u} = \sum_{i=1}^n \frac{h(1_i, \mathbf{A}) - h(0_i, \mathbf{A})}{\mu_{F_i} + \mu_{G_i}}, \quad (5)$$

where $h(x_i, \cdot)$ equals the reliability of the system given that the state of component i is x_i . For a proof of the second equality of Equation (5), see Aven and Jensen [2], p. 114.

The above asymptotic result is more formally stated and proved below.

We assume that the process $\mathbf{X}$ is a regenerative process [with regenerative state $(1, 1, \dots, 1)$]. The variable S denotes the length of the first renewal cycle of the process $\mathbf{X}$, that is, the time until the process returns to state $(1, 1, \dots, 1)$. Let T_Φ denote the time to the first system failure and q the probability that a system failure occurs in a renewal cycle, that is,

$$q = P(N_S \geq 1) = P(T_\Phi < S).$$

For $q \in (0, 1)$, let P_0 and P_1 denote the conditional probability given $N_S = 0$ and $N_S \geq 1$, that is, $P_0(\cdot) = P(\cdot | N_S = 0)$ and $P_1(\cdot) = P(\cdot | N_S \geq 1)$. The corresponding expectations are denoted as E_0 and E_1 . Furthermore, let $c_{0S}^2 = [E_0 S^2 / (E_0 S)^2] - 1$ denote the squared coefficient of variation of S under P_0 .

The notation $\xrightarrow{P}$ is used for convergence in probability and $\xrightarrow{D}$ for convergence in distribution. We write $\text{Exp}(t)$ for the exponential distribution with parameter t and $N(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2 .

For each component i ($i \in \{1, 2, \dots, n\}$) we assume that there is a sequence of uptime and downtime distributions (F_{ij}, G_{ij}) , $j = 1, 2, \dots$.

To simplify notation, we normally omit the index j . When assuming in the following that $\mathbf{X}$ is a regenerative process, it is tacitly understood for all $j \in \mathbb{N}$. We shall formulate conditions which guarantee that αT_Φ is asymptotically exponentially distributed with parameter 1, where α is a suitable normalizing “factor” (more precisely, a normalizing sequence depending on j). Below we focus on the factor λ_Φ . Other factors are studied in Aven and Jensen [2], see the proof of Theorem 2 below and the Bibliographic notes in Ref. 2.

The basic idea of the proof of the asymptotic exponentiality of αT_Φ is as follows: If we assume that $\mathbf{X}$ is a regenerative process and the probability that a system failure occurs in a renewal cycle, that is, q , is small (converges to zero), then the time to the first system failure will be approximately equal to the sum of a number of renewal cycles having no system failures; and this number of cycles is geometrically distributed with parameter q . Now if $q \rightarrow 0$ as $j \rightarrow \infty$, the desired result follows by using Laplace transformation. The result can be formulated in general terms as shown in Lemma 1.

A result similar to Lemma 1 was first proved by Keilson [6], see also Refs 7–9. Our version of this lemma is taken from Aven and Jensen [10].

Asymptotic results similar to those presented below have been reported in many papers and books.

A survey is given by Gertsbakh [11]. See also the summary given in the books by Gnedenko and Ushakov [8], Ushakov [12], and Kovalenko *et al.* [13, 14]. Some of the earliest results go back to the works done by Keilson [6] and Solov'yev [15].

See Bibliographic notes in Aven and Jensen [2], p. 166.

Lemma 1. Let $S, S_i, i = 1, 2, \dots$, be a sequence of nonnegative i.i.d. random variables with distribution function $F(t)$ having finite mean a , $a > 0$, and finite variance, and let v be a random variable independent of (S_i) , geometrically distributed with parameter q ($0 < q \leq 1$), that is, $P(v = k) = qp^{k-1}$, $k = 1, 2, \dots$, $p = 1 - q$. Furthermore, let

$$S^* = \sum_{i=1}^{v-1} S_i.$$

Consider now a sequence F_j, q_j ($j = 1, 2, \dots$) satisfying the above conditions for each j . Then if (as $j \rightarrow \infty$)

$$q \rightarrow 0 \quad (6)$$

and

$$qc_S^2 \rightarrow 0, \quad (7)$$

where c_S^2 denotes the squared coefficient of variation of S , we have (as $j \rightarrow \infty$)

$$\frac{qS^*}{a} \xrightarrow{D} \text{Exp}(1). \quad (8)$$

Proof. Let $\tilde{S}^* = qS^*/a$. By conditioning on the value of v , it is seen that the Laplace transform of S^* , $L_{S^*}(x) = Ee^{-xS^*}$, equals $q/(1 - pL(x))$ where $L(x)$ is the Laplace transform of S_i . Let $\psi(x) = (L(x) - 1 + ax)/x$. Then,

$$L_{S^*}(x) = \frac{q}{1 - p(1 - ax + x\psi(x))}.$$

We need to show that

$$L_{\tilde{S}^*}(x) = Ee^{-(qx/a)S^*} \rightarrow \frac{1}{1+x},$$

since the convergence theorem for Laplace transforms then give the desired result. Noting that

$$Ee^{-(qx/a)S^*} = \frac{1}{1 + px - (px/a)\psi(qx/a)}.$$

we must require that

$$(x/a)\psi(qx/a) \rightarrow 0,$$

that is,

$$[L(qx/a) - 1 + qx]/q \rightarrow 0.$$

Using that $ES = a$ and the inequalities $0 \leq e^{-t} - 1 + t \leq t^2/2$, we find that

$$\begin{aligned} 0 &\leq [L(qx/a) - 1 + qx]/q \\ &= E[e^{-(qx/a)S} - 1 + (qx/a)S]/q \\ &\leq E[(qx/a)S]^2/2q \\ &= \frac{x^2}{2} \frac{q}{a^2} ES^2 \\ &= \frac{x^2}{2} q(1 + c_S^2). \end{aligned}$$

The desired conclusion (Eq. 8) follows now since $q \rightarrow 0$ and $qc_S^2 \rightarrow 0$ (Assumptions 6 and 7).

Theorem 2. Assume that $\mathbf{X}$ is a regenerative process, and that F_{ij} and G_{ij} change in such a way that the following conditions hold (as $j \rightarrow \infty$):

$$q \rightarrow 0 \quad (9)$$

$$qc_{0S}^2 \rightarrow 0 \quad (10)$$

$$\frac{qE_1S}{ES} \rightarrow 0 \quad (11)$$

$$E_1(N_S - 1) \rightarrow 0. \quad (12)$$

Then,

$$A[0, u/\lambda_\Phi] \rightarrow e^{-u}, \text{ i.e., } \lambda_\Phi T_\Phi \xrightarrow{D} \text{Exp}(1). \quad (13)$$

Proof. Using Lemma 1, we first prove that under conditions 9–11, we have

$$\frac{T_\Phi q}{E_0 S} \xrightarrow{D} \text{Exp}(1). \quad (14)$$

Let v denote the renewal cycle index associated with the time of the first system failure, T_Φ . Then it is seen that T_Φ has the same distribution as

$$\sum_{k=1}^{v-1} S_{0k} + W_v,$$

where (S_{0k}) and (W_k) are independent sequences of i.i.d. random variables with

$$P(S_{0k} \leq s) = P_0(S \leq s)$$

and

$$P(W_k \leq w) = P_1(T_\Phi \leq w).$$

Both sequences are independent of ν , which has a geometrical distribution with parameter $q = P(N_S \geq 1)$. Hence Equation (14) follows from Lemma 1 provided

$$\frac{W_\nu q}{E_0 S} \xrightarrow{P} 0. \quad (15)$$

By a standard conditional probability argument it follows that

$$ES = (1 - q)E_0 S + qE_1 S,$$

and by noting that

$$\begin{aligned} \frac{qE_1 T_\phi}{E_0 S} &= \frac{qEW}{E_0 S} \leq \frac{qE_1 S}{E_0 S} = \frac{qE_1 S(1 - q)}{ES - qE_1 S} \\ &= \frac{\frac{qE_1 S}{ES}(1 - q)}{1 - \frac{qE_1 S}{ES}} \rightarrow 0, \end{aligned} \quad (16)$$

we see that Equation (15) holds.

Next, we will show that the ratio of λ_ϕ and $q/E_0 S$ converges to 1.

Observe that

$$\lambda_\phi = \frac{EN_S}{ES} \quad (17)$$

by the Renewal Reward Theorem (Aven and Jensen [2], p. 250).

Using Equation (17), we obtain

$$\begin{aligned} \frac{\lambda_\phi}{q/E_0 S} &= \frac{\lambda_\phi}{q/ES} \frac{E_0 S}{ES} \\ &= \frac{EN_S/ES}{q/ES} \frac{E_0 S}{ES} \\ &= \frac{EN_S}{q} \frac{E_0 S}{ES}. \end{aligned}$$

Now $EN_S/q = 1 + E_1(N_S - 1) \rightarrow 1$ in view of Equation (12), and

$$\frac{E_0 S}{ES} = \frac{1 - q \frac{E_1 S}{ES}}{1 - q} \rightarrow 1$$

by Equations (9) and (11). Hence the ratio of λ_ϕ and $q/E_0 S$ converges to 1. Combining this with Equation (14), the conclusion of the theorem follows.

It is intuitively clear that if the components have constant failure rates, and the component unavailabilities converge to zero, then the conditions of Theorem 2 would hold. Below this result will be formally established. We assume, for the sake of simplicity, that no single component is in series with rest of the system. If there are one or more components in series with the rest of the system, we know that the time to failure of these components has an exact exponential distribution, and by independence, it is straightforward to establish the limiting distribution of the total system.

Define

$$\lambda = \sum_{i=1}^n \lambda_i, \quad \tilde{\mu} = \sum_{i=1}^n \tilde{\mu}_i,$$

where

$$\tilde{\mu}_i = \sup_{0 \leq t < t^*} \{E[R_{i1} - t | R_{i1} > t]\},$$

and $t^* = \sup\{t \in \mathbb{R}_+ : \bar{G}_i(t) > 0\}$. We see that $\tilde{\mu}_i$ expresses the maximum expected residual repair time of the component i . We might have $\tilde{\mu}_i = \infty$, but we shall restrict attention to the finite case in the following. If G_i has the NBUE property, then

$$\tilde{\mu}_i \leq \mu_{G_i}.$$

If the repair times are bounded by a constant c , that is, $P(R_{ik} \leq c) = 1$, then $\tilde{\mu}_i \leq c$.

Theorem 3. *Assume that the system has no components in series with rest of the system. Furthermore, assume that component i has an exponential lifetime distribution with failure rate $\lambda_i > 0$, $i = 1, 2, \dots, n$. If $d' \rightarrow 0$, where $d' = \tilde{\lambda}\tilde{\mu}$, and there exist constants c_1 and c_2 such that $\lambda_i \leq c_1 < \infty$ and $ER_i^2 \leq c_2 < \infty$ for all i , then the conditions (Eqs 9–12) of Theorem 2 are all met, and consequently the limiting result (Eq. 13) holds, that is, $\lambda_\phi T_\phi \xrightarrow{D} \text{Exp}(1)$.*

The formal proof of this theorem is lengthy and detailed, and is omitted (2, pp. 123–127). We refer to Aven and Jensen [2] for further results providing sufficient conditions for Equations 9–12 to hold.

The above results show that the time to the first system failure is approximately exponentially distributed with parameter λ_Φ . For a system comprising highly available components, it is clear that $P(\mathbf{X}_t = \mathbf{1})$ would be close to one, hence the above approximations for the interval $[0, u]$ can also be used for an interval $(t, t + u]$.

Asymptotic Normality

Now we turn to a completely different way to approximate the distribution of the time to the first system failure. Above the up and downtime distributions are assumed to change such that the system availability increases and after a time rescaling the time to failure converges to an exponential variable. Now we leave the up and downtime distribution unchanged and establish a central limit theorem as t increases to infinity.

Theorem 4. *If $\mathbf{X}$ is a regenerative process with cycle length S , $\text{Var}[S] < \infty$ and $\text{Var}[N_S] < \infty$, then as $t \rightarrow \infty$,*

$$\sqrt{t} \left(\frac{N_{u+t} - N_u}{t} - \lambda_\Phi \right) \xrightarrow{D} N(0, \gamma_\Phi^2),$$

where

$$\gamma_\Phi^2 ES = \text{Var}(N_S - \lambda_\Phi S). \quad (18)$$

Noting that the system failure rate λ_Φ is given by

$$\lambda_\Phi = \frac{EN_S}{ES}, \quad (19)$$

the result follows from the Central limit theorem for renewal reward processes (2, p. 251). From Theorem 4 we obtain an approximate normal distribution for the interval availability.

If the system failure rate is small, then we have

$$\gamma_\Phi^2 \approx \lambda_\Phi.$$

This is seen by noting that

$$\begin{aligned} \gamma_\Phi^2 &= \frac{\text{Var}(N_S - \lambda_\Phi S)}{ES} = \frac{E(N_S - \lambda_\Phi S)^2}{ES} \\ &\approx \frac{EN_S^2}{ES} \approx \frac{EN_S}{ES} = \lambda_\Phi, \end{aligned}$$

where the last approximation follows by observing that if the system failure rate is small, then N_S with a probability close to one is equal to the indicator function $I(N_S \geq 1)$. More formally, it is possible to show that under certain conditions, $\gamma_\Phi^2/\lambda_\Phi$ converges to one (2, p. 130).

From the above theorem, we establish an approximation for the interval availability:

$$A[t, t + u] \approx P(N(0, 1) \leq -\lambda_\Phi \sqrt{t}/\gamma_\Phi).$$

A STANDBY SYSTEM

In this section, we study the performance of a standby system comprising two identical components of which one is normally operating and the other is in (cold) standby. The following additional assumptions are made:

- Failed components are repaired. Only one component can be repaired at a time (one repair facility/channel), the repair policy is “first come first served.”
- Switchover to the standby component is perfect, that is, instantaneous and failure free.
- A standby component that has completed its repair is functioning at demand, that is, the failure rate is zero in the standby state.
- All failure times and repair times are independent with probability distributions $F(t)$ and $G(t)$, respectively. F is absolutely continuous and has finite mean, and G has finite third-order moment. We assume

$$\int_0^\infty F(t) dG(t) > 0.$$

In the following, T refers to a failure time of a component and R refers to a repair time.

The squared coefficient of variation of the repair time distribution is denoted c_G^2 .

Let Φ_t denote the state of the system at time t , that is, the number of components functioning at time t ($\Phi_t \in \{2, 1, 0\}$). The process Φ is a regenerative process. It is seen that the time points when Φ jumps to state 1 are regenerative points, that is, the time points when (i) the operating component fails and the second component is not under repair (the process jumps from state 2 to 1) or (ii) both the components are failed and the repair of the component being repaired is completed (the process jumps from state 0 to 1).

A cycle refers to the length between two consecutive regenerative points. We define Y and S as the system downtime in a cycle and the length of a cycle, respectively. It follows from the Renewal Reward Theorem that

$$\bar{A} = \frac{EY}{ES}. \quad (20)$$

Here, system downtime corresponds to the time where two of the components are not functioning. Let us now look closer into the problem of computing $\bar{A}$, given in Equation (20).

Using Equation (20) and that the regenerative points for the process Φ are generated by the jumps from state 2 to 1, the limiting system unavailability $\bar{A}$ can be written as

$$\bar{A} = \frac{E[\max\{R - T, 0\}]}{ET + E[\max\{R - T, 0\}]} \quad (21)$$

$$= \frac{(\mu_G - w)}{\mu_F + (\mu_G - w)}, \quad (22)$$

where

$$w = E[\min\{R, T\}] = \int_0^\infty \bar{F}(t)\bar{G}(t) dt,$$

noting that $\max\{R - T, 0\} = R - \min\{R, T\}$ and the system downtime equals zero if the repair of the failed component is completed before the failure of the operating component, and equals the difference between the repair time of the failed component and the time to failure of the operating component

if this difference is positive. Thus, we have proved the following theorem.

Theorem 5. *The unavailability $\bar{A}$ is given by Equation (22).*

We now assume an exponential failure time distribution $F(t) = 1 - e^{-\lambda t}$. Then, we have

$$\bar{A} \approx \bar{A}', \quad (23)$$

where

$$\bar{A}' = \frac{\lambda^2}{2} ER^2 = \frac{\delta^2}{2} [1 + c_G^2]. \quad (24)$$

This gives a simple approximation formula for computing $\bar{A}$. The approximation (Eq. 23) is established formally by the following proposition.

Proposition 1. *If $F(t) = 1 - e^{-\lambda t}$, then*

$$0 \leq \bar{A}' - \bar{A} \leq (\bar{A}')^2 + \frac{\delta^3}{6} \frac{ER^3}{\mu_G^3}. \quad (25)$$

Proof. Using that $1 - e^{-\lambda t} \leq \lambda t$ and changing the order of integration, it follows that

$$\begin{aligned} \bar{A} &= \frac{\lambda(\mu_G - w)}{1 + \lambda(\mu_G - w)} \leq \lambda(\mu_G - w) \quad (26) \\ &= \lambda \int_0^\infty F(t)\bar{G}(t) dt \\ &\leq \lambda \int_0^\infty (\lambda t)\bar{G}(t) dt \\ &= \lambda^2 \frac{1}{2} ER^2 = \bar{A}'. \end{aligned} \quad (27)$$

It remains to show the right-hand inequality of Equation (25). Considering

$$\begin{aligned} \bar{A}(1 + \lambda(\mu_G - w)) &= \lambda \int_0^\infty F(t)\bar{G}(t) dt \\ &\geq \lambda \int_0^\infty \left(\lambda t - \frac{1}{2}(\lambda t)^2 \right) \bar{G}(t) dt \\ &= \bar{A}' - \frac{1}{6} \lambda^3 ER^3 \end{aligned}$$

and the inequalities $\bar{A} \leq \lambda(\mu_G - w) \leq \bar{A}'$ obtained above, it is not difficult to see that

$$\begin{aligned}
0 \leq \bar{A}' - \bar{A} &\leq \bar{A}\lambda(\mu_G - w) + \frac{1}{6}\lambda^3 ER^3 \\
&\leq (\bar{A}')^2 + \frac{1}{6}\lambda^3 ER^3,
\end{aligned}$$

which completes the proof.

Hence $\bar{A}'$ overestimates the unavailability and the error term will be negligible provided that $\delta = \mu_G/\mu_F$ is sufficiently small.

FINAL REMARKS

This article covers only a small number of availability models compared to the large number of models presented in the literature, see the survey paper by Smith *et al.* [16]. Our focus has been the basic models. A large category of models not included are models addressing preventive maintenance.

This category of models is often developed with the purpose of optimizing maintenance (replacement) policies. We refer to Aven and Jensen [2], Chapter 5 and its Bibliographic notes. See also Dijkhuizen and Heijden [17].

For some recent papers on availability models, see Zhang *et al.* [18] and Montoro-Cazorla and Pérez-Ocón [19], and the references therein.

Furthermore, we have not included models where some components remain in “suspended animation” while a component is being repaired/replaced. When repair of the failed component is completed, the remaining components resume operation. At any instant, they are not “as good as new,” but rather as good as they were when the system stopped operating. Then, it can be shown that the long run (expected) proportion of time the system is functioning equals (Barlow and Proschan [4] and Aven [20])

$$\frac{1}{1 + \sum_{i=1}^n \frac{\text{MTTR}_i}{\text{MTTF}_i}}. \quad (28)$$

This expression is to be compared to

$$h(\mathbf{A}) = \prod_{i=1}^n \frac{\text{MTTF}_i}{\text{MTTF}_i + \text{MTTR}_i}.$$

If the component availabilities are high, then

$$h(\mathbf{A}) \approx 1 - \sum_{i=1}^n \bar{A}_i,$$

which is seen to be approximately equal to Equation (28).

Markov models are often used for availability analysis [21]. Markov models are used to describe and analyze the movement of the system among various states, from the perfect functioning state to the complete failure state. Markov models are particularly suitable for analyzing smaller systems with complicated maintenance policies and possible dependencies between components. On the other hand, the computational problems can be significant and the assumption of exponential distributions may seem to restrict applicability. However, the so-called phase-type approach can open for other distributions. The phase-type approach makes use of the fact that a distribution function can be approximated by a mixture of Erlang distributions (with the same scale parameter), see, for example, Asmussen [22] and Tijms [23].

Markov models is a type of multistate systems, and many of the availability results obtained for the binary case have been extended to such systems (2, p. 151). See also Zio *et al.* [24] and the references therein. Finally, we would like to draw attention to semi-Markov models for availability analysis of multistate systems and we refer the readers to Csenki [25] and Limnios and Oprisan [26].

Acknowledgments

The author is grateful to an anonymous reviewer for useful comments and suggestions to an earlier version of the manuscript.

REFERENCES

1. Rausand M, Høyland A. System reliability theory. 2nd ed. New York: Wiley; 2004.
2. Aven T, Jensen U. Stochastic models of reliability. New York: Springer; 1999.
3. Barlow R, Proschan F. Mathematical theory of reliability. New York: Wiley; 1965.

4. Barlow R, Proschan F. Statistical theory of reliability and life testing. New York: Holt, Rinehart, and Winston; 1975.
5. Birolini A. Quality and reliability of technical systems. Berlin: Springer; 1994.
6. Keilson J. A limit theorem for passage times in ergodic regenerative processes. *Ann Math Stat* 1966;37:866–870.
7. Keilson J. Markov chain models - rarity and exponentiality. Berlin: Springer; 1979.
8. Gnedenko BV, Ushakov IA, Falk JA, editors. Probabilistic reliability engineering. Chichester: Wiley; 1995.
9. Kovalenko IN. Rare events in queuing systems—a survey. *Queuing Syst* 1994; 16:1–49.
10. Aven T, Jensen U. Asymptotic distribution of the downtime of a monotone system. *Math Meth Oper Res* 1997;45:355–375.
11. Gertsbakh IB. Asymptotic methods in reliability: a review. *Adv Appl Probab* 1984;16:147–175.
12. Ushakov IA, editor. Handbook of reliability engineering. Chichester: Wiley; 1994.
13. Kovalenko IN, Kuznetsov NY, Pegg PA. Mathematical theory of reliability of time dependent systems with practical applications. New York: Wiley; 1997.
14. Kovalenko IN, Kuznetsov NY, Shurenkov VM. Models of random processes. London: CRC Press; 1996.
15. Solov'yev AD. Asymptotic behavior of the time to the first occurrence of a rare event. *Eng Cybern* 1971;9(6):1038–1048.
16. Smith M, Aven T, Dekker R, *et al.* A survey on the interval availability of failure prone systems. Proceedings ESREL'97 Conference; 1997 Jun 17–20; Lisbon; 1997. pp. 1727–1737.
17. Dijkhuizen G, Heijden M. Preventive maintenance and the interval availability distribution of an unreliable production system. *Reliab Eng Syst Saf* 1999;66:13–27.
18. Zhang T, Xie M, Horigome M. Availability and reliability of k-out-of-(M+N): G warm standby systems. *Reliab Eng Syst Saf* 2006;91:381–387.
19. Montoro-Cazorla D, Pérez-Ocón M. A deteriorating two-system with two repair modes and sojourn times phased-type distributed. *Reliab Eng Syst Saf* 2006;91:1–9.
20. Aven T. Reliability and risk analysis. Oxford: Elsevier; 1992.
21. Misra KB. Reliability analysis and prediction. Oxford: Elsevier; 1992.
22. Asmussen S. Applied probability and queues. New York: Wiley; 1987.
23. Tijms HC. Stochastic modeling and analysis: a computational approach. New York: Wiley; 1994.
24. Zio E, Marella M, Pedofillini L. A Monte Carlo simulation approach to the availability assessment of multi-state systems with operational dependencies. *Reliab Eng Syst Saf* 2007;92:871–882.
25. Csenki A. An integral equation approach to the interval reliability of systems modeled by finite semi-Markov processes. *Reliab Eng Syst Saf* 1995;47:37–45.
26. Limnios N, Oprisan G. Semi-Markov processes and reliability. Boston (MA): Birkhauser; 2001.

POISSON PROCESS AND ITS GENERALIZATIONS

VAIDYANATHAN RAMASWAMI
AT&T Labs Research,
Florham Park, New Jersey

A point process is a random scattering of points over a subset of the n -dimensional Euclidean space R^n , $n \geq 1$ [1]. Mathematically, it is a function $N(\cdot)$, which provides for each Borel set B , the number of points $N(B)$ falling in it. When dealing with a point process on $[0, \infty)$, it is customary to denote $N(0, t]$ by $N(t)$ and to refer to $\{N(t) : t \geq 0\}$ as the point process. Point processes arise naturally in the study of random events over time, space, or space-time and find applications in numerous areas.

Among point processes, the Poisson process [1–3] is the simplest. Yet, it is the archetype of substantially large classes of point processes through generalization in various directions.

STATIONARY POISSON PROCESS ON $[0, \infty)$

The stationary Poisson process on $[0, \infty)$ was introduced independently by Erlang [4] in 1909 as a model for call arrivals to a telephone trunking system and by Bateman [5] in 1910 as a model for α -particle emissions in a nuclear experiment. The name Poisson process, however, may have come to pass since the number of counts in any interval of length t in such a process has a distribution discovered by Siméon-Denis Poisson [6], bearing his name, and characterized by the density $p_n(t) = e^{-\lambda t} (\lambda t)^n / n!$, $n = 0, 1, 2, \dots$ for some constant $\lambda > 0$.

Poisson distribution arises as a limit of the simple binomial distribution, the distribution of the count of an event with probability p in a set of n independent repetitions of the experiment. Indeed, the conditions ($n \rightarrow \infty, p \rightarrow 0, np \equiv \lambda$) of the limit theorem render the Poisson distribution a natural

candidate for counts of rare events, a fact confirmed by many data sets including the famous example provided by Von Bortkiewicz [7,8] of the number of fatalities due to horse kicks in the Prussian army; see Feller [9] for many others.

Definition 1. The stationary Poisson process on $[0, \infty)$ is a stochastic process $N(\cdot)$ such that for any set of intervals $(a_i, b_i]$, $i = 1, \dots, k$ which are nonoverlapping and for all integers $n_i \geq 0$, we have

$$P[N(a_i, b_i] = n_i, i = 1, \dots, k] \\ = \prod_{i=1}^k e^{-\lambda(b_i - a_i)} \frac{[\lambda(b_i - a_i)]^{n_i}}{n_i!}, \quad (1)$$

where $\lambda > 0$ is a constant called the *intensity* of the process.

This definition adapted from Ref. 1 and given in terms of the probability law of $N(\cdot)$, is one that is easy to generalize to higher dimensions. However, since the specification of the probability law does not by itself guarantee any regularity properties for sample paths often required in analysis, some authors, for example Çinlar [2], prefer to define the Poisson process directly in terms of its sample path properties.

Definition 2. A stochastic process $\{N(t) : t \geq 0\}$ of counts defined on a probability space $(\Omega, \mathcal{F}, P)$ is a stationary Poisson process if (a) for almost all ω , its paths $N_\omega(\cdot)$ are right continuous with finite left limits at all points t ; (b) for almost all ω , $N_\omega(t+) - N_\omega(t-) \leq 1$ (jumps are simple, i.e., of size 1); (c) for any $t, s \geq 0$, $[N(t+s) - N(t)]$ is independent of the history $\{N(u) : u \leq t\}$ (i.e., the process has independent increments); and (d) for any $s, t \geq 0$, the distribution of $N(t+s) - N(t)$ is independent of t (i.e., increments are stationary).

In the language of Palm [10], (c) and (d) above constitute an assumption of “no after-effects.”

Characterizations of the Stationary Poisson Process

It is clear from Equation (1) that the count $N(B)$ in any set B that can be expressed as the union of a finite number of finite intervals has Poisson distribution with parameter $\lambda l(B)$, where $l(\cdot)$ is the Lebesgue measure assigning to each interval its length as its measure. This is a characterizing property of the Poisson process [11].

Theorem 1. *A stationary point process with $N(I) < \infty$ a.s. for all intervals I of finite length is a Poisson process iff there exists a $\lambda > 0$ such that, for each set B that can be expressed as the union of a finite number of finite intervals, $N(B)$ obeys the Poisson distribution with parameter $\lambda l(B)$.*

The striking feature of this theorem is that independence of the increments comes as a consequence of its simple assumptions. We do hasten to note, however, that the theorem does not hold if one attempts to weaken the assumption concerning the range of the sets B to a smaller class like the set of intervals (relevant counterexamples and references in Ref. 1, p. 30).

Theorem 2. *For the stationary Poisson process $\{N(t) : t \geq 0\}$ with intensity λ , as $\Delta t \downarrow 0$,*

$$P[N(t + \Delta t) - N(t) = 1] = \lambda \Delta t + o(\Delta t), \quad (2)$$

$$P[N(t + \Delta t) - N(t) \geq 2] = o(\Delta t). \quad (3)$$

See Çinlar [2] for a proof. Equation (2) justifies calling λ the intensity of the process; it is the expected rate at which an increment occurs in a small interval of time. Equation (3) is a condition called *orderliness* by Khinchin [12] and is the analytic analogue of the unit jump condition for sample paths. The notion of orderliness helps to obtain the following result of Renyi [11] that weakens the above condition to the probabilities of a zero count.

Theorem 3. *$N(\cdot)$, an orderly point process on $[0, \infty)$, is a stationary Poisson process iff there exists a $\lambda > 0$ such that $P(N(B) = 0) =$*

$\exp(-\lambda l(B))$ for all sets B that can be expressed as a union of a finite number of finite intervals.

Indeed, under mild conditions, the avoidance function $P[N(B) = 0]$ characterizes the law of a point process in general spaces ([1], Section 7.3). A special case of that extends Theorem 3 to Poisson processes over R^n .

For the Poisson process, it is clear that

$$E[N(t + s) - N(t) | N(u), 0 \leq u \leq t] = \lambda s, \quad \text{for all } t, s \geq 0. \quad (4)$$

The following theorem of Watanabe [13] provides one of the simplest qualitative characterizations of the Poisson process in terms of this result.

Theorem 4. *The point process $\{N(t) : t \geq 0\}$ is a stationary Poisson process with intensity λ if and only if for almost all ω , each jump of $N_\omega(\cdot)$ is of unit magnitude and Equation (4) holds.*

Poisson Process as a Markov Chain

The memory-less property embodied in condition (c) of Definition 2, and Equations (2) and (3) lead to the following.

Theorem 5. *For the Poisson process with intensity λ , $\{N(t) : t \geq 0\}$ is Markov with infinitesimal generator Q with elements*

$$Q(i, j) = \begin{cases} -\lambda & \text{if } j = i, \\ \lambda & \text{if } j = i + 1, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

The Markov property of the Poisson process makes conditioning arguments easy in the models in which it is used; see Ref. 14, Example 2.1.2, for an example. Also, in such models one is often able to unearth embedded processes [15] that are Markovian even if the original model is not. This considerable ease of analysis accounts for the extensive use of the Poisson process in the literature on queues, inventories, dams, storage processes, insurance risk, etc. [16–18].

Inter-Event Times in a Poisson Process

Let $0 < \tau_1 < \tau_2 < \dots$ denote the successive epochs of jumps of the stationary Poisson process with rate λ , and let $X_i = \tau_i - \tau_{i-1}$, with $\tau_0 = 0$.

Theorem 6. *The inter-event times $X_1, X_2, \dots$ are independent and identically distributed with the exponential density $f(x) = \lambda e^{-\lambda x}$, $x > 0$.*

The exponential distribution has the memory-less property, namely,

$$P[X > t + s | X > s] = P[X > t].$$

For the Poisson process, this is explained by the fact that the elapsed inter-event time (which is a part of the history of the process) does not provide any information on the future waiting time for the event to occur.

An interesting set of intervals for the Poisson process comprises F_t the forward recurrence time (time to the next event after time t) and the elapsed time E_t (time since the last event before t). The Markov property of the Poisson process entails that just as any inter-event time, F_t is also exponentially distributed with parameter λ . Note, however, that $E_t \leq t$. By using a simple renewal argument, one can easily show that E_t has a point mass $e^{-\lambda t}$ at t and a density $\lambda e^{-\lambda u}$, $0 < u < t$. As $t \rightarrow \infty$, the elapsed time density also converges to the exponential distribution. Thus, the stationary forward recurrence time and stationary elapsed time for the Poisson process are both exponentially distributed with parameter λ . Indeed, among simple point processes, only the stationary Poisson process has this property.

The above facts demonstrate an interesting paradox called the *waiting time paradox* [9]. The inter-event interval containing a given epoch t , being the sum of the forward and elapsed times at t , has a mean larger than that of an inter-event time. This paradox underlying the oft-expressed despair, “Why does the bus I am waiting for always come late?” is resolved by noting that for a fixed t , a larger inter-event interval

has a higher chance of covering t compared to a shorter interval, implying that the choice of the interval covering t is inherently length-biased. This simple fact is indeed an example both of the novelty brought by probabilistic thinking as well as of how deceptive the simplicity of the Poisson process can be.

Conditional Distributions

It is straightforward [1,2] to show that for a stationary Poisson process, given that exactly n points occur in $[0, t]$ the location of those points $\tau_1 < \tau_2 < \dots < \tau_n$ in $[0, t]$ are distributed as an ordered independent sample of size n from the uniform distribution over $[0, t]$. The memory-less property and constant intensity imply that no epoch is favored over another.

The PASTA Property

In considering stationary distributions (i.e., distributions as time $t \rightarrow \infty$) of stochastic processes, one must differentiate between limits taken at an arbitrary time (i.e., as t goes to ∞ in the usual sense) or via selected subsets of epochs. As an example, in a queueing model, one may consider $X(t)$, the virtual waiting time (the waiting time as seen at an arbitrary t), or consider $X(\tau_k)$ where τ_k is the epoch of the k th arrival, and their stationary versions when they exist. These stationary distributions have ergodic interpretations respectively as the a.s. limits of the long run sample averages $\lim_{t \rightarrow \infty} t^{-1} \int_0^t X(t) dt$ taken over all time points and as long run sample averages $\lim_{N \rightarrow \infty} N^{-1} \sum_{k=1}^N X(\tau_k)$ taken over the selected set of time points $\{\tau_k\}$ only. PASTA (Poisson arrivals see time averages) [19] asserts, under general conditions, that these averages are equal when the selected epochs form a Poisson process. The conditions comprise of a nonanticipating attribute namely, that the next epoch of the observing Poisson process should be independent of the history of the observed process. The PASTA result when it holds is helpful in analysis, simulations, and statistical estimation from samples. For details, interested readers may refer to Ref. 19.

NONSTATIONARY POISSON PROCESS ON $[0, \infty)$

Generalization to the nonstationary simple Poisson process is obtained by relaxing the condition of constant intensity and assuming instead that $E[N(t)] = \Lambda(t)$, where $\Lambda(t) = \int_0^t \lambda(u) du$, where $\lambda(\cdot)$ is a nonnegative function called the *time-dependent intensity* of the process.

Definition 3. A simple nonstationary Poisson process on $[0, \infty)$ is a stochastic process $N(\cdot)$ such that for any set of intervals $(a_i, b_i]$, $i = 1, \dots, k$ which are nonoverlapping (i.e., no two of which have a common subinterval) and for any set of integers $n_i \geq 0$, we have

$$P[N(a_i, b_i] = n_i, i = 1, \dots, k] = \prod_{i=1}^k e^{-[\Lambda(b_i) - \Lambda(a_i)]} \frac{[\Lambda(b_i) - \Lambda(a_i)]^{n_i}}{n_i!}, \quad (6)$$

where $\Lambda(t) = \int_0^t \lambda(u) du$ and $\lambda(\cdot)$ is a nonnegative function on $[0, \infty)$.

One may also define a nonstationary Poisson process as in Definition 2 by replacing (d) there with the condition $E[N(t)] = \Lambda(t)$.

Assume $\lambda(\cdot)$ is continuous, and define a time change

$$u(t) = \inf\{s : \Lambda(s) > t\}, \quad t \geq 0.$$

Then $M(t) = N(u(t))$ is a stationary Poisson process with unit intensity. This artifice helps obtain results for the nonstationary Poisson process from those governing the stationary case; see Ref. 2, Chapter 4, for some examples.

Compound Poisson Process

Suppose that in Definition 2, we replace (b) by the condition that each jump can be of a random size $k \geq 1$ with probability p_k , where $\sum_1^\infty p_k = 1$, independently of the history of the process. Then, the resulting process is called a compound Poisson process. Such processes are useful in modeling batch arrivals of customers to a queue, of demands of random sizes to an inventory system, and so on. Many applied models with compound Poisson

processes lend themselves to an easy analysis with explicit formulae solutions as well; see examples and details in Ref. 2.

Martingales and Stochastic Integration

In the theory of stochastic integration [20] due to Itô—see Steele [21] for a nice summary—one considers processes defined through integrals of the type $I(t) = \int_0^t f(t) dB(t)$ where $B(\cdot)$ is a Brownian motion and f is a suitably restricted function such that $f(t)$ is measurable with respect to the history $\{B(u) : u \leq t\}$. A discussion of Itô integration is beyond our scope, but it suffices to note that this is not the same as the Stieltjes integral in classical analysis, and that it has become an important tool. A simple example of an Itô process is the value process of a portfolio wherein $f(t)$ denotes the size of the holding at time t of an asset and $B(\cdot)$ represents the price process. The standard Brownian motion $B(\cdot)$ is a martingale, and a famous result called *Girsanov's Theorem* asserts that a Brownian motion with drift can be turned into a martingale by a change of measure. These enable the construction and study of Itô integrals via the theory of martingales.

In many ways, the nonstationary Poisson process on the line is a discrete analogue of Brownian motion with drift. The Itô integral here is the sum $I(t) = \sum_{\tau_k \leq t} f(\tau_k)$ where τ_k are the event times of a Poisson process $N(\cdot)$, and f is a function whose value at t is determined by the history $\{N(u) : u \leq t\}$. A connection of the Poisson process to a closely related martingale and the development of a discrete version of the stochastic integral paralleling the continuous case can be made [22]. That development has placed the Poisson process in yet another favorable light as a tool in stochastic integration. In addition to unifying the continuous and discrete state space development of the stochastic integral, the accompanying techniques have provided powerful tools for analyzing processes with jumps and for proving many results in very general settings (details in Brémaud [22]).

Closure Properties

The set of Poisson processes is closed under independent superpositions, a fact useful in

many contexts like networks of queues where streams are merged. Thus, if N_1 and N_2 are independent Poisson processes, so is $N_1 + N_2$. The set is also closed under independent Bernoulli thinning, a result useful in networks with probabilistic routing. Thus, if each of the points of a Poisson process $N(\cdot)$ with intensity λ is independently kept or discarded with respective probabilities p and $1 - p$, the kept and discarded epochs form independent Poisson processes with respective intensities λp and $\lambda(1 - p)$. For substantial generalizations of this thinning result, see Ref. 23. Also, if the superposition of two renewal processes is Poisson, then each must be Poisson. For these and related characterizations, refer to Ref. 1.

SPATIAL POISSON PROCESS

Spatial point processes have become of increasing interest in operations research, particularly in the context of characterizing traffic and performance in wireless networks. The basic ideas parallel those of the line process. We assume as given a boundedly finite (i.e., finite on bounded sets) parameter measure $\Lambda(\cdot)$ which is a measure on the Borel sets of the underlying space $\mathcal{X}$ assumed to be a complete, separable, metric space (c.s.m.s.).

Definition 4. A stochastic process $N(\cdot)$ which assigns a (random) count of points to each Borel set B of $\mathcal{X}$ is a spatial Poisson process over $\mathcal{X}$ with parameter measure $\Lambda(\cdot)$ iff for all finite collections of bounded, disjoint Borel sets B_i and nonnegative integers n_i , we have

$$\begin{aligned} P[N(B_i) = n_i, i = 1, \dots, k] \\ = \prod_{i=1}^k e^{-\Lambda(B_i)} \frac{(\Lambda(B_i))^{n_i}}{n_i!}. \end{aligned} \quad (7)$$

If $\Lambda(B) = \int_B \lambda(x) dx$ for all B for a nonnegative, continuous function λ , the analysis is routine, but when that is not the case, technical complexities can arise; herein lies the non-triviality of the extension ([1], Section 2.4).

For $\epsilon > 0$, denote by $S_{\epsilon, x}$ the open sphere with center x and radius ϵ . The Poisson process on $\mathcal{X}$ is said to be orderly iff for all $x \in \mathcal{X}$,

$$P[N(S_{\epsilon, x}) > 1] = o(P[N(S_{\epsilon, x}) > 0]), \quad \epsilon \rightarrow 0.$$

One can show that the Poisson process is orderly iff its parameter measure $\Lambda(\cdot)$ has no atoms.

One can show in general that if a point process is such that for all bounded Borel sets, $N(B)$ has a Poisson distribution with parameter $\Lambda(B)$ for some boundedly finite measure $\Lambda(\cdot)$, then the process is Poisson. Also, an orderly process is Poisson iff we have $P[N(B) = 0] \equiv \exp[-\Lambda(B)]$ for some nonatomic boundedly finite measure Λ . These results are natural generalizations of similar results we stated for the Poisson process on the line. Closure properties under independent superposition and Bernoulli thinning also hold for spatial Poisson processes (details in Ref. 1).

Spatial Poisson processes are well studied from the point of view of characterization, computation, simulation, and statistical inference. For a detailed discussion, we refer the reader to Daley and Vere-Jones [1], Illian *et al.* [24], and Cressie [25].

A Useful Construction

We can construct a spatial Poisson process from a stationary Poisson process on the line using a time change and a random rotation.

Theorem 7. Let $\{T_n\}$ denote the successive epochs of a stationary Poisson process on $[0, \infty)$ with intensity λ . Then the set of points with radial coordinates $(\sqrt{T_n/\pi}, \Theta_n)$, where $\{\Theta_n\}$ is a sequence of i.i.d. $U(0, 2\pi)$ variables constitutes a stationary Poisson pattern on the plane with intensity λ .

Similar ideas have been used by Isham [26] on renewal processes to generate interesting planar point processes. Later, Latouche and Ramaswami [27] developed a new class called *spatial phase-type point patterns* by starting with a Markovian arrival process (MAP, see section titled, “Batch Markovian Arrival Process”) on $[0, \infty)$, general time changes, and general scattering rules. In addition to being algorithmically tractable, such processes can model a variety of spatial clusters and do not suffer from

the major restrictions of spatial homogeneity and independence of the stationary Poisson. For these reasons, they can be of use in areas like wireless and mobile communications where such assumptions are often untenable. A detailed study of the isometric case may be found in the dissertation [28] and subsequent publications of Remiche who has used the models to demonstrate the impact of spatial dependence of call arrivals on the performance of directional antennae in wireless communications.

A Limit Theorem for Superpositions

We noted that the Poisson distribution is a natural candidate for rare events. So is the spatial Poisson process in that under certain technical conditions, the superposition of a set of independent “sparse” point processes converges to a Poisson process. In many practical instances, this limit result forms an unstated justification for using a Poisson process. Note that the telephone calls to a large switch form the superposition of a large number of sparse processes of calls generated by individual users; this may give a mathematical explanation for Erlang’s success with the Poisson model for call arrivals. Due to its importance in applications, we shall state the superposition theorem carefully below. For proof, refer to Ref. 1, Chapter 9.

First, we provide a definition of sparsity.

Definition 5. A sequence of point processes $N_{ni}, n, i = 1, 2, \dots$ defined on a c.s.m.s. $\mathcal{X}$ will be called uniformly asymptotically negligible (u.a.n.) if for all bounded Borel sets $B \subset \mathcal{X}$, we have

$$\lim_{n \rightarrow \infty} \sup_i P[N_{ni}(B) > 0] = 0.$$

In addition to the above, call the set of processes an independent array if for each n , the processes $\{N_{ni} : i = 1, 2, \dots\}$ are mutually independent.

Theorem 8. Let N_{ni} denote a u.a.n. independent array of point processes on a c.s.m.s. $\mathcal{X}$, and let $N_n = \sum_{i=1}^{m_n} N_{ni}$ where m_n is a sequence of positive integers going to ∞ as

$n \rightarrow \infty$. As $n \rightarrow \infty$, N_n converges weakly to a Poisson process on $\mathcal{X}$ with parameter measure $\Lambda(\cdot)$ iff for all bounded Borel sets B with $\Lambda(\delta B) = 0$, as $n \rightarrow \infty$,

$$\sum_{i=1}^{m_n} P[N_{ni} \geq 2] \rightarrow 0 \text{ and} \\ \sum_{i=1}^{m_n} P[N_{ni}(B) \geq 1] \rightarrow \Lambda(B).$$

Weak convergence here is in the sense of stochastic process convergence; see Whitt [29] for applicable definitions.

DOUBLY STOCHASTIC POISSON PROCESS

Assume that there exists a positive random variable Λ , and that given $\Lambda = \lambda$, the random variable X has a Poisson distribution with parameter λ . The (unconditional) distribution of X is called a *doubly Poisson* or *mixed Poisson distribution*. The earliest instance of such a model is due to Greenwood and Yule [30] who considered the number of accidents to be Poisson with a parameter called *accident intensity* that varies across a worker pool and may therefore be considered random. The mixed Poisson distribution is widely used in Bayesian statistics and forms the basis of a commonly used technique called *Poisson Regression* [31] and many of its variants.

Just like the parameter of a single Poisson random variable, in a variety of applications, it may be appropriate to consider the intensity $\Lambda(\cdot)$ of a Poisson process to be random. What results is the doubly stochastic Poisson process. Thus, one assumes a stochastic process $\Lambda(\cdot)$ such that given a realization of $\lambda(\cdot)$ thereof, $N(\cdot)$ is a Poisson process with intensity function given by the sample path $\lambda(\cdot)$. As an example, we could consider the number N_t of interrupts in $(0, t]$ in a computer system to be a Poisson process for which the intensity at time t is $M_t \lambda$, where M_t , the number of jobs multitasking at time t , itself evolves as a stochastic process.

The doubly stochastic Poisson process was introduced by Cox [32] and also goes by the name Cox process. A rigorous treatment

is usually made in the context of random measures [1,33]. Cox processes find considerable use in time series analysis and Bayesian methods for them. For some interesting applications in credit modeling, refer to Lando [34] and Schönbucher [35]. Thus, the Cox process does find applications in many contexts. However, when used in their full generality in queueing models and the like, one finds that the models become quite hard to analyze; only some bounds for performance measures could be obtained and that too for certain special models [36]. However, for Cox processes on the line with an additional Markovian structure on the intensity process, much can be accomplished. That leads us to our next set of processes which can also be considered as special instances of a Cox process with batch events.

BATCH MARKOVIAN ARRIVAL PROCESS

The MAP and the Batch Markovian Arrival Process (BMAP) are generalizations obtained by assuming that the counting process combined with an auxiliary Markov process is a bivariate Markov process. The motivation for them is the fact that Markov property aids analysis of complex models by simplifying conditioning arguments [37,38].

Definition 6. A counting process $\{N_t\}$ giving the number of events in the interval $(0, t]$ is called a BMAP with parameter set $\{D_k : k = 0, 1, \dots\}$ where each D_k is an $m \times m$ matrix iff (a) there exists a continuous-time Markov chain $\{J_t\}$, called the Markov chain of phases, with infinitesimal generator $D = \sum_{k=0}^{\infty} D_k$; (b) the bivariate process $\{(N_t, J_t) : t \geq 0\}$ is Markov, and (c) after grouping the state space into sets called levels where level $L_k = \{(k, 1), \dots, (k, m)\}$, $k = 0, 1, \dots$, the infinitesimal generator matrix Q of the bivariate Markov chain has the block partitioned form

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & \cdots & \cdots \\ 0 & D_0 & D_1 & \cdots & \cdots \\ 0 & 0 & D_0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \quad (8)$$

When $D_k = 0$ for all $k \geq 2$, the process is simply called an MAP.

It is customary to assume that D is irreducible (no redundant phases) and that D_0 is nonsingular ($N(\infty) = \infty$) [39]. However, for those instances where we need a terminating MAP or BMAP, we could assume that one of the m phases, say m , is absorbing and that no events occur once m is reached; that is, $D_n(m, j) = 0$ for all n, j .

The BMAP was introduced by Neuts [40] and was called the *Neuts process* by Ramaswami [39] who provided a detailed steady state analysis of the single server queue with BMAP input along with a substantial matrix generalization of the Pollaczek–Khinchin formula [16] for the virtual waiting time distribution. The notation used by Neuts [40] and Ramaswami [39] is different from that given here which is based on Lucantoni [41] comprising results from Ref. 39 translated into the new notation. A subsequent work by Lucantoni *et al.* [42] contains a detailed time-dependent analysis of the BMAP/G/1 queue and serves as an excellent demonstration of the tractability of BMAP. The BMAP can also be used as a model for a service mechanism with batch services [43]. A simple introduction to the BMAP can be found in Ref. 14, Chapter 3.

The inter-event times in the MAP (and BMAP) are in general correlated phase-type random variables. Indeed, the extreme generality of the BMAP follows from Schassberger [44] and Asmussen and Koole [45] who show that the MAPs form a dense subset of the set of all simple point processes on $[0, \infty)$. Models involving the BMAP and various special cases continue to be an active area of research.

REFERENCES

1. Daley DJ, Vere-Jones D. An introduction to the theory of point processes. New York (NY): Springer; 1988.
2. Cinlar E. Introduction to stochastic processes. Englewood (NJ): Prentice-Hall, Inc; 1975.
3. Grandell J. Mixed poisson processes. Suffolk (UK): Chapman & Hall; 1997.

4. Erlang AK. The theory of probabilities and telephone conversations. *Nyt Tidsskr Mat* 1909;B20:33–41. [Also reprinted in Brockmeyer E, *et al.* editors. The life and works of A.K. Erlang. Copenhagen, Denmark: Copenhagen Telephone Company; pp. 131–137.]
5. Bateman H. Note on the probability distribution of α -particles. *Philos Mag* 1910;20(6): 704–707.
6. Poisson SD. Recherches sur la probabilité des jugements en matière criminelle et en matière civile. Précédées des règles générales du calcul des probabilités. Paris: Bachelier; 1837.
7. Von Bortkiewicz L. Das Gesetz der kleinen Zahlen. Leipzig: G. Teubner; 1898.
8. Quine MP, Seneta E. Bortkiewicz's data and the law of small numbers. *Int Stat Rev* 1987;55:173–181.
9. Feller W. Volume 1. 1950, An introduction to probability theory and its applications. 3rd ed. New York: Wiley; 1968.
10. Palm C. Intensitätsschwankungen in fernsprechverkehr. *Ericsson Tech* 1943;44:1–189.
11. Renyi A. Remark on the Poisson process. *Proc R Soc London [Ser A]* 1967;2:119–123.
12. Khinchin AY. Volume 49, Mathematical methods in the theory of queueing. *Trudy Mat Inst Steklov*; 1955. (in Russian) [Translated (1960) Griffin, London]
13. Watanabe S. On discontinuous additive functionals and Levy measures of a Markov process. *Jpn J Math* 1964;34:53–70.
14. Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. SIAM and ASA; 1999.
15. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain. *Ann Math Stat* 1953;24:338–354.
16. Takacs L. Introduction to the theory of queues. New York: Oxford University Press; 1962.
17. Prabhu NU. Stochastic storage processes, queues, insurance risk, and dams. New York (NY): Springer; 1980.
18. Asmussen S. Applied probability and queues. New York (NY): Springer; 2003.
19. Wolffe RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
20. Protter PE. Stochastic integration and differential equations. 2nd ed. New York (NY): Springer; 2005.
21. Steele M. Ito calculus. In: Teugels J, Sundt B, editors. The encyclopaedia of actuarial science. New York (NY): John Wiley & Sons, Ltd.; 2004. Available at <http://www-stat.wharton.upenn.edu/~steele/Publications/index.html>.
22. Bremaud P. Point processes and queues: martingale dynamics. New York (NY): Springer; 1981.
23. Kingman JFC. Poisson processes. Oxford (UK): Oxford University Press; 1995.
24. Illian J, *et al.* Statistical analysis and modelling of spatial point patterns (statistics in practice). New York (NY): John Wiley & Sons, Ltd.; 2008.
25. Cressie NAC. Statistics for spatial data. New York (NY): John Wiley & Sons, Inc.; 1993.
26. Isham L. Construction of planar point processes using concentric circles. *Stoch Proc Appl* 1977;5:131–141.
27. Latouche G, Ramaswami V. Spatial point patterns of phase type. In: Ramaswami V, Wirth PE, editors. Teletraffic contributions for the information age, Proceedings of the International teletraffic Congress ITC 15. Amsterdam: Elsevier Science B.V.; 1997. pp. 381–390.
28. Remiche MA. Isotropic phase planar point processes: analysis and application to cellular mobile telecommunication [PhD thesis]. Belgium: Université Libre de Bruxelles; 1999.
29. Whitt W. Stochastic process limits. New York (NY): Springer; 2002.
30. Greenwood M, Yule GU. An enquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *J Roy Stat Soc* 1920;83:255–279.
31. Gelman A, *et al.* Bayesian data analysis. New York (NY): Chapman & Hall/CRC; 1995.
32. Cox DR. Some statistical methods connected with series of events (with discussion). *J R Stat Soc [Ser B]* 1955;17:129–164.
33. Cox DR, Isham V. Point processes. London: Chapman and Hall; 1980.
34. Lando D. Credit risk modeling. Princeton (NJ): Princeton University Press; 2004.
35. Schonbucher PJ. Credit derivatives pricing models. Models, pricing and implementation. New York (NY): John Wiley & Sons, Inc.; 2007.

36. Rolski T. Upper bounds for single server queues with doubly stochastic Poisson arrivals. *Math Oper Res* 1986;11(3):442–450.
37. Neuts MF. *Matrix-geometric solutions in stochastic models*. Baltimore (MD): Johns Hopkins Press; 1980.
38. Neuts MF. *Structured stochastic matrices of M/G/1 type and their applications*. New York (NY): Marcel Dekker Inc.; 1989.
39. Ramaswami V. The N/G/1 queue and its detailed analysis. *Adv Appl Probab* 1980;12: 222–261.
40. Neuts MF. A versatile Markovian point process. *J Appl Probab* 1979;16:764–779.
41. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Stoch Mod* 1991;7(1):1–46.
42. Lucantoni DM, Choudhury GL, Whitt W. The transient BMAP/G/1 queue. *Stoch Mod* 1997;13(2):223–238.
43. Ramaswami V. From the matrix geometric to the matrix exponential. *Queueing Syst Theory Appl* 1990;6:229–260.
44. Schassberger R. *Warteschlangen*. Wien: Springer; 1973.
45. Asmussen S, Koole G. Marked point processes as limits of Markovian arrival streams. *J Appl Probab* 1993;30:365–372.

POLYNOMIAL TIME PRIMAL INTEGER PROGRAMMING VIA GRAVER BASES

SHMUEL ONN

Technion, Israel Institute of Technology,
Haifa, Israel

INTRODUCTION

Linear integer programming is the following optimization problem,

$$\min \{wx : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}, \quad (1)$$

where $w \in \mathbb{Z}^n$ is a cost, A is an integer $m \times n$ matrix, $b \in \mathbb{Z}^m$ is a right-hand side, and $l, u \in \mathbb{Z}_\infty^n$ are lower and upper bounds, with $\mathbb{Z}_\infty := \mathbb{Z} \cup \{\pm\infty\}$. It is one of the most fundamental discrete optimization problems and has a very broad modeling power and numerous applications in a variety of areas. A comprehensive treatment of the classical theory of linear integer programming is in the excellent book by Schrijver [1].

Integer programming is generally NP-hard and hence presumably intractable, but polynomial time solvable in two fundamental situations: first, when the underlying matrix is totally unimodular [2]; and second, when the number of variables is fixed [3]. The goal of this article is to describe the recent exciting developments in so-called *primal methods* for integer programming which, in particular, lead to the polynomial time solution of new broad classes of integer programming problems with linear, and more generally, various nonlinear, objective functions of the form

$$\min \{f(x) : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}. \quad (2)$$

Our emphasis is on primal methods that lead to polynomial time solution. For other primal strategies for integer programming, see Refs 4 and 5 and the references therein.

To simplify the presentation, and since the feasible set in most of the applications is finite or can be made finite by more careful modeling, whenever an algorithm detects that the feasible set is infinite, it simply stops. So, throughout our discussion, an algorithm is said to *solve a (nonlinear) integer programming problem* if it either finds an optimal solution x or concludes that the feasible set is infinite or empty.

The function f involved in a nonlinear integer program (2) is presented either by a mere comparison oracle that, queried on two vectors x, y , answers whether or not $f(x) \leq f(y)$, or by an evaluation oracle that, queried on vector x , returns $f(x)$.

In the following section, we define Graver bases and show that they can be used to design primal algorithms for solving (non)linear integer programs in polynomial time. In the section titled “ n -Fold Integer Programming,” we consider the universal class of n -fold integer programming problems which, using the results of the following section, can be solved in polynomial time. In the section titled “Some Applications,” we describe some applications to multi-index and multicommodity transportation.

GRAVER BASED PRIMAL INTEGER PROGRAMMING

The *Graver basis* is a fundamental object in the theory of integer programming which was introduced by Graver, already back in 1975 [6]. However, only recently, in the series of articles [7–9], it was established that the Graver basis can be used to solve linear, as well as various nonlinear, integer programming problems in polynomial time. We proceed to describe these important new developments.

The *lattice* of an integer $m \times n$ matrix A is the set $\mathcal{L}(A) := \{x \in \mathbb{Z}^n : Ax = 0\}$. Let $\mathcal{L}^*(A) := \{x \in \mathbb{Z}^n : Ax = 0, x \neq 0\}$. Define a partial order $\sqsubseteq$ on $\mathbb{R}^n$ as follows: for two vectors $x, y \in \mathbb{R}^n$ put $x \sqsubseteq y$ and say that x

is *conformal* to y if for $i = 1, \dots, n$, we have $x_i y_i \geq 0$ (so x and y lie in the same orthant of $\mathbb{R}^n$) and $|x_i| \leq |y_i|$. The classical lemma of Gordan [10] implies that every subset of $\mathbb{Z}^n$ has finitely many $\sqsubseteq$ -minimal elements. We have the following fundamental definition.

Definition 1 [6]. The *Graver basis* of an integer matrix A is defined to be the finite set $\mathcal{G}(A) \subset \mathbb{Z}^n$ of $\sqsubseteq$ -minimal elements in $\mathcal{L}^*(A) = \{x \in \mathbb{Z}^n : Ax = 0, x \neq 0\}$.

For instance, the Graver basis of the 1×3 matrix $A := [1 \ 2 \ 1]$ consists of 8 elements,

$$\mathcal{G}(A) = \pm \{(2, -1, 0), (0, -1, 2), (1, 0, -1), (1, -1, 1)\}.$$

We now show the fundamental role of the Graver basis of a matrix A in primal methods for linear integer programs of form (1) defined by A . A finite sum $u := \sum_i v_i$ of vectors in $\mathbb{R}^n$ is called *conformal* if all summands lie in the same orthant and hence $v_i \sqsubseteq u$ for all i . It is not hard to see that every $x \in \mathcal{L}^*(A)$ is a conformal sum $x = \sum_i g_i$ of Graver basis elements $g_i \in \mathcal{G}(A)$. This implies the following lemma.

Lemma 1. *A feasible point x in the linear integer program defined by matrix A ,*

$$\min \{wx : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\},$$

is optimal if and only if there is no $g \in \mathcal{G}(A)$ such that $x + g$ is feasible and better.

Proof. One direction is obvious. Suppose then x is not optimal. Let x^* be any better feasible point and put $h := x^* - x$. Then $Ah = b - b = 0$ so $h \in \mathcal{L}^*(A)$ and hence is a conformal sum $h = \sum_i g_i$ with $g_i \in \mathcal{G}(A)$ for all i . Now, $l \leq x, x + h = x^* \leq u$ and $g_i \sqsubseteq h$ imply that $l \leq x + g_i \leq u$ for all i . Also, $A(x + g_i) = Ax = b$ for all i . Therefore, $x + g_i$ is feasible for all i . Now $\sum_i wg_i = wh = wx^* - wx < 0$ implies that $wg_i < 0$ for some g_i in this sum. Therefore, $x + g_i$ is feasible and better.

A primal method for solving Equation (1) or (2) is one that starts from some

feasible point x^0 and generates a sequence $x^0, x^1, x^2, \dots$ of better and better feasible points, namely,

$$\begin{aligned} x^k &\in \mathbb{Z}^n, \quad Ax^k = b, \quad l \leq x^k \leq u, \\ f(x^{k+1}) &< f(x^k), \quad k = 0, 1, 2, \dots \end{aligned} \quad (3)$$

Lemma 1 implies that the Graver basis $\mathcal{G}(A)$ along with an initial feasible point x^0 enables to implement a primal method for problem (1) as follows: given a feasible point x^k in the sequence (3) do: if $wg < 0$ for some $g \in \mathcal{G}(A)$ with $l \leq x^k + g \leq u$ then append the new better feasible point $x^{k+1} := x^k + g$ to the sequence and repeat; otherwise, x^k is optimal. However, this simple implementation does not lead to a polynomial time solution, and a more sophisticated use of the Graver basis is needed.

We proceed to show that the Graver basis *does* enable primal methods for solving linear and various nonlinear integer programming problems in polynomial time.

Separable Convex Objective Functions

We begin with separable convex minimization. So we consider programs of the form

$$\min \{f(x) : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\},$$

where $f : \mathbb{Z}^n \rightarrow \mathbb{Z}$ is a *separable convex* function, that is, $f(x) = \sum_{j=1}^n f_j(x_j)$ with $f_j : \mathbb{Z} \rightarrow \mathbb{Z}$ a univariate convex function for all j . We denote by $\hat{f}$ the maximum value of $|f(x)|$ over the feasible set (which need not be part of the input).

We now show that the Graver basis enables efficient separable convex minimization given an initial feasible point. We later show how to find an initial point as well. Here, we only outline the proof. Complete details can be found in Refs 11 and 12.

Lemma 2. *There is an algorithm that, given an integer $m \times n$ matrix A , its Graver basis $\mathcal{G}(A)$, vectors $l, u \in \mathbb{Z}_\infty^n$ and $x \in \mathbb{Z}^n$ with $l \leq x \leq u$, and separable convex function $f : \mathbb{Z}^n \rightarrow \mathbb{Z}$ presented by a comparison oracle, solves the integer program*

$$\begin{aligned} \min \{f(z) : z \in \mathbb{Z}^n, Az = b, l \leq z \leq u\}, \\ b := Ax, \end{aligned} \quad (4)$$

in time polynomial in the binary-encoding length $\langle A, \mathcal{G}(A), l, u, x, \hat{f} \rangle$ of the data.

Proof. Using linear programming, in polynomial time, either detect that the feasible set is infinite and stop, or conclude that it is finite and continue. Next, produce a sequence of feasible points $x_0, x_1, \dots, x_s$ with $x_0 := x$ the given input point, as follows. Having obtained x_k , solve the minimization problem

$$\min \{ f(x_k + \lambda g) : \lambda \in \mathbb{Z}_+, g \in \mathcal{G}(A), l \leq x_k + \lambda g \leq u \} \quad (5)$$

by minimizing, for each $g \in \mathcal{G}(A)$, the convex univariate function in the integer variable λ using repeated bisections over the relevant interval. If the minimal value in Equation (5) satisfies $f(x_k + \lambda g) < f(x_k)$ then set $x_{k+1} := x_k + \lambda g$ and repeat, else stop and output the last point x_s in the sequence. Now, $Ax_{k+1} = A(x_k + \lambda g) = Ax_k = b$ by induction on k , so each x_k is feasible. Since the feasible set is finite and the x_k have decreasing objective values and hence distinct, the algorithm terminates.

We now show that the point x_s output by the algorithm is optimal. Let x^* be any optimal solution to Equation (4). Consider any point x_k in the sequence and suppose it is not optimal. We claim that a new point x_{k+1} will be produced and will satisfy

$$f(x_{k+1}) - f(x^*) \leq \frac{2n-3}{2n-2} (f(x_k) - f(x^*)) \quad (6)$$

By the integer Carathéodory's theorem [13,14], we can write the difference $x^* - x_k = \sum_{i=1}^t \lambda_i g_i$ as conformational sum involving $1 \leq t \leq 2n-2$ elements $g_i \in \mathcal{G}(A)$ with all $\lambda_i \in \mathbb{Z}_+$. By the supermodularity of separable convex functions [15,16],

$$\begin{aligned} f(x^*) - f(x_k) &= f\left(x_k + \sum_{i=1}^t \lambda_i g_i\right) - f(x_k) \\ &\geq \sum_{i=1}^t (f(x_k + \lambda_i g_i) - f(x_k)). \end{aligned}$$

Suitable manipulation now implies that for some $1 \leq i \leq t$, we have

$$\begin{aligned} f(x_k + \lambda_i g_i) - f(x^*) &\leq \frac{t-1}{t} (f(x_k) - f(x^*)) \\ &\leq \frac{2n-3}{2n-2} (f(x_k) - f(x^*)). \end{aligned}$$

So the point $x_k + \lambda g$ attaining minimum in Equation (5) satisfies

$$\begin{aligned} f(x_k + \lambda g) - f(x^*) &\leq f(x_k + \lambda_i g_i) - f(x^*) \\ &\leq \frac{2n-3}{2n-2} (f(x_k) - f(x^*)) \end{aligned}$$

and so indeed $x_{k+1} := x_k + \lambda g$ will be produced and will satisfy Equation (6). This shows that the last point x_s produced and output by the algorithm is indeed optimal.

We proceed to bound the number s of points. Consider any point $i < s$ and the intermediate nonoptimal point x_i in the sequence produced by the algorithm. Then $f(x_i) > f(x^*)$ with both values being integers, and so repeated use of Equation (6) gives

$$\begin{aligned} 1 &\leq f(x_i) - f(x^*) \\ &= \prod_{k=0}^{i-1} \frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} (f(x) - f(x^*)) \\ &\leq \left(\frac{2n-3}{2n-2} \right)^i (f(x) - f(x^*)) \end{aligned}$$

from which it follows that the number of points s produced by the algorithm satisfies

$$s = O(n \log(f(x) - f(x^*))).$$

Thus, the number of points produced and the total running time are polynomial.

Next, we show that Lemma 2 can also be used to find an initial feasible point for the given integer program or assert that none exists in polynomial time.

Lemma 3. *There is an algorithm that, given integer $m \times n$ matrix A , its Graver basis $\mathcal{G}(A)$,*

$l, u \in \mathbb{Z}_\infty^n$, and $b \in \mathbb{Z}^m$, in time which is polynomial in $\langle A, \mathcal{G}(A), l, u, b \rangle$, either finds a feasible point $x \in S$ or asserts that S is empty or infinite, where

$$S := \{z \in \mathbb{Z}^n : Az = b, l \leq z \leq u\}.$$

Proof. For simplicity, assume that $l, u \in \mathbb{Z}^n$ are finite. For the general case, see Refs 11 or 12. We also may assume $l \leq u$. Now, either detect that there is no integer solution to the system of equations $Ax = b$ (without the lower and upper bound constraints) and stop, or determine such solution $\hat{x} \in \mathbb{Z}^n$ and continue; it is well known that this can be done in polynomial time, say, using the Hermite normal form of A [1]. Next, define a separable convex function on $\mathbb{Z}^n$ by $f(x) := \sum_{j=1}^n f_j(x_j)$ with

$$f_j(x_j) := \begin{cases} l_j - x_j, & \text{if } x_j < l_j \\ 0, & \text{if } l_j \leq x_j \leq u_j \\ x_j - u_j, & \text{if } x_j > u_j \end{cases}, \quad j = 1, \dots, n$$

and extended lower and upper bounds

$$\hat{l}_j := \min\{l_j, \hat{x}_j\}, \quad \hat{u}_j := \max\{u_j, \hat{x}_j\}, \quad j = 1, \dots, n.$$

Consider the auxiliary separable convex integer program

$$\min\{f(z) : z \in \mathbb{Z}^n, Az = b, \hat{l} \leq z \leq \hat{u}\}. \quad (7)$$

Apply the algorithm of Lemma 2 to Equation (7) and obtain, in polynomial time, an optimal solution x . Now note that every point $z \in S$ is feasible in Equation (7), and every point z feasible in Equation (7), satisfies $f(z) \geq 0$ with equality if and only if $z \in S$. So, if $f(x) > 0$ then the original set S is empty, whereas if $f(x) = 0$ then $x \in S$ is a feasible point.

Lemmas 2 and 3 combined together give at once the following theorem showing that the Graver basis enables separable convex minimization in polynomial time.

Theorem 1 [9]. *There is an algorithm that, given integer $m \times n$ matrix A , its*

Graver basis $\mathcal{G}(A)$, $l, u \in \mathbb{Z}_\infty^n$, $b \in \mathbb{Z}^m$, and separable convex $f : \mathbb{Z}^n \rightarrow \mathbb{Z}$ presented by comparison oracle, solves in time polynomial in $\langle A, \mathcal{G}(A), l, u, b, \hat{f} \rangle$ the problem

$$\min\{f(x) : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}.$$

Ramifications

Linear Integer Programming. Any linear function $wx = \sum_{i=1}^n w_i x_i$ is separable convex. Moreover, an upper bound on $|wx|$ over the feasible set (when finite), which is polynomial in the binary-encoding length of the data, readily follows from Cramer's rule. Therefore we obtain, as an immediate special case of Theorem 1, the following important result, asserting that Graver bases enable the polynomial time solution of linear integer programming.

Theorem 2 [7]. *There is an algorithm that, given an integer $m \times n$ matrix A , its Graver basis $\mathcal{G}(A)$, $l, u \in \mathbb{Z}_\infty^n$, $b \in \mathbb{Z}^m$, and $w \in \mathbb{Z}^n$, solves in time which is polynomial in $\langle A, \mathcal{G}(A), l, u, b, w \rangle$, the following linear integer programming problem,*

$$\min\{wx : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}.$$

Distance Minimization. Another useful special case of Theorem 1, which is natural in various applications such as image processing, tomography, communication, and error correcting codes, is the following result, which asserts that the Graver basis enables to determine a feasible point which is l_p -closest to a given desired goal point in polynomial time.

Theorem 3 [9]. *There is an algorithm that, given integer $m \times n$ matrix A , its Graver basis $\mathcal{G}(A)$, positive integer p , vectors $l, u \in \mathbb{Z}_\infty^n$, $b \in \mathbb{Z}^m$, and $\hat{x} \in \mathbb{Z}^n$, solves in time polynomial in p and $\langle A, \mathcal{G}(A), l, u, b, \hat{x} \rangle$, the distance minimization problem*

$$\min\{\|x - \hat{x}\|_p : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}. \quad (8)$$

For $p = \infty$ the problem (8) can be solved in time polynomial in $\langle A, \mathcal{G}(A), l, u, b, \hat{x} \rangle$.

Convex Maximization. We proceed to discuss the *maximization* of a convex function of the composite form $f(Wx)$, with $f: \mathbb{Z}^d \rightarrow \mathbb{Z}$ any convex function and W any integer $d \times n$ matrix.

A *direction* of an edge (1-dimensional face) e of a polyhedron P is any nonzero scalar multiple of $u - v$, where u, v are any two distinct points in e . A result of Ref. 17 asserts, roughly, that if we can optimize in polynomial time linear functions over a set $S \subseteq \mathbb{Z}^n$ and we are given a set $E \subseteq \mathbb{Z}^n$ containing a direction of each edge of the polyhedron $\text{conv}(S)$, then we can also maximize over S convex functions of the above composite form $f(Wx)$ in polynomial time. Nicely enough, it turns out that the Graver basis of a matrix A contains a direction of each edge of $\text{conv}(S)$ with

$$S := \{z \in \mathbb{Z}^n : Az = b, l \leq z \leq u\}.$$

Thus, the result of Ref. 17 together with Theorem 2 imply the following theorem.

Theorem 4 [8]. *For every fixed d , there is an algorithm that, given integer $m \times n$ matrix A , its Graver basis $\mathcal{G}(A)$, $l, u \in \mathbb{Z}_\infty^n$, $b \in \mathbb{Z}^m$, integer $d \times n$ matrix W , and convex function $f: \mathbb{Z}^d \rightarrow \mathbb{R}$ presented by a comparison oracle, solves in time which is polynomial in $\langle A, W, \mathcal{G}(A), l, u, b \rangle$, the convex integer maximization problem*

$$\max \{f(Wx) : x \in \mathbb{Z}^n, Ax = b, l \leq x \leq u\}.$$

N-FOLD INTEGER PROGRAMMING

We now discuss a broad fundamental class of integer programming problems, for which the Graver basis itself can be computed in polynomial time, and hence the Graver based primal methods of the section titled “Graver Based Primal Integer Programming,” lead to polynomial time algorithms.

An $(r, s) \times t$ *bimatrix* is a matrix A consisting of two blocks A_1, A_2 , with A_1 its $r \times t$ submatrix consisting of the first r rows and A_2 its $s \times t$ submatrix consisting of the last s

rows. The n -fold product of A is the following $(r + ns) \times nt$ matrix,

$$A^{(n)} := \begin{pmatrix} A_1 & A_1 & \cdots & A_1 \\ A_2 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_2 \end{pmatrix}.$$

An n -fold integer programming problem is one of the form

$$\min \left\{ f(x) : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, l \leq x \leq u \right\}, \quad (9)$$

that is, with defining matrix which is the n -fold product $A^{(n)}$ of a fixed bimatrix A , and (possibly nonlinear) objective function f . Note that the dimension of an n -fold integer program is nt and is *variable*. Also, n -fold products $A^{(n)}$ are typically not totally unimodular: the n -fold product of the simple $(0, 1) \times 1$ bimatrix with A_1 empty and $A_2 := 2$ satisfies $A^{(n)} = 2I_n$ and has exponential determinant 2^n . So this class of programs cannot be solved by methods of fixed dimension or totally unimodular matrices. The class of n -fold integer programs turns out very natural and has numerous applications. In fact it is *universal*, a result of Ref. 18 implies that every integer program is an n -fold program over the $(3m, 3 + m) \times 3m$ bimatrix A with $A_1 = I_{3m}$ and A_2 the incidence matrix of the bipartite graph $K_{3,m}$ for some m .

It is useful to index each feasible point in Equation (9) as a tuple $x = (x^1, \dots, x^n)$ of n bricks $x^k \in \mathbb{Z}^t$. It turns out [19,20] that for every integer bimatrix A , there is a nonnegative integer $g(A)$, termed the *Graver complexity* of A , such that, for every n , every $x \in \mathcal{G}(A^{(n)})$ in the Graver basis of the n -fold product has at most $g(A)$ nonzero bricks regardless of n . This can be exploited to obtain the following result.

Theorem 5 [7]. *For every fixed integer bimatrix A , there is an algorithm that, given n , computes the Graver basis $\mathcal{G}(A^{(n)})$ in time which is polynomial in n .*

Theorem 5 and the general results on primal integer programming via Graver bases

from the section titled “Graver Based Primal Integer Programming,” along with some further extensions, lead to a sequence of results providing polynomial time solvability of n -fold integer programming with various classes of objective functions, collected in the following theorem. The functions f, g involved are presented by either comparison oracles or evaluation oracles. Also, $\hat{f}, \hat{g}$ denote the maximum values of $|f(Wx)|, |g(x)|$ over the feasible set.

Theorem 6. *For every fixed integer $(r, s) \times t$ bimatrix A , there are algorithms that, given $n, l, u \in \mathbb{Z}_{\infty}^{nt}$, and $b \in \mathbb{Z}^{r+ns}$, solve the following n -fold integer programs:*

1. *Linear Integer Programming [7]. In time polynomial in n and $\langle l, u, b, w \rangle$,*

$$\min\{wx : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, l \leq x \leq u\},$$

with $w \in \mathbb{Z}^{nt}$ any given cost vector.

2. *Convex Maximization [8]. In time polynomial in n and $\langle W, l, u, b \rangle$,*

$$\max\{f(Wx) : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, l \leq x \leq u\},$$

with d fixed, W any given integer $d \times nt$ matrix, and $f : \mathbb{Z}^d \rightarrow \mathbb{R}$ any given convex function presented by a comparison oracle.

3. *Separable Convex Minimization [9].*

$$\min\{f(x) : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, l \leq x \leq u\},$$

with $f : \mathbb{Z}^{nt} \rightarrow \mathbb{Z}$ separable convex function presented by a comparison oracle, in time polynomial in n and $\langle l, u, b, \hat{f} \rangle$.

4. *Distance Minimization [9]. In time polynomial in n, p , and $\langle l, u, b, \hat{x} \rangle$,*

$$\min\{\|x - \hat{x}\|_p : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, l \leq x \leq u\},$$

with p any given positive integer and $\hat{x} \in \mathbb{Z}^{nt}$ any given integer goal point. For $p = \infty$, the problem can be solved in time polynomial in n and $\langle l, u, b, \hat{x} \rangle$.

5. *Weighted Separable Convex Minimization [21].*

$$\min\{f(W^{(n)}x) + g(x) : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, \hat{l} \leq W^{(n)}x \leq \hat{u}, l \leq x \leq u\},$$

with W a fixed integer $(p, q) \times t$ bimatrix, $\hat{l}, \hat{u} \in \mathbb{Z}_{\infty}^{p+nq}$ any given additional bounds, and $f : \mathbb{Z}^{p+nq} \rightarrow \mathbb{Z}, g : \mathbb{Z}^{nt} \rightarrow \mathbb{Z}$ any given separable convex functions presented by evaluation oracles, in time polynomial in n and $\langle l, u, \hat{l}, \hat{u}, b, \hat{f}, \hat{g} \rangle$.

SOME APPLICATIONS

We now demonstrate the power of Graver bases and n -fold integer programming, and describe important applications for transportation and multicommodity flow.

The Multi-Index Transportation Problem.

Problems involving multiply-indexed variables have been studied extensively in mathematical programming in the context of multi-index dimensional transportation (see Refs 22 and 23 and references therein) and in statistics in the context of disclosure control and privacy in statistical databases (see Refs 24 and 25 and references therein). Theorem 6 leads to the first polynomial time algorithms for a variety of problems in these two contexts. Here, we discuss only applications to triple-index transportation problems. Discussion of further applications can be found in Refs 11 and 12.

The multi-index transportation problem was introduced by Motzkin in Ref. 26. For simplicity, we discuss here only the triple-index problem with specified line-sums,

$$\min\left\{wx : x \in \mathbb{Z}_+^{l \times m \times n}, \sum_i x_{i,j,k} = v_{*,j,k}, \sum_j x_{i,j,k} = v_{i,*,k}, \sum_k x_{i,j,k} = v_{i,j,*}\right\}, \quad (10)$$

where the specified line-sums are $mn + ln + lm$ given nonnegative integer numbers

$$v_{*,j,k}, v_{i,*,k}, v_{i,j,*}, \quad 1 \leq i \leq l, 1 \leq j \leq m, 1 \leq k \leq n.$$

If l, m, n are all fixed then the problem is solvable in polynomial time using integer programming in fixed dimension lmn [3]. If l, m, n are all variables then the problem is NP-hard [27]. The in-between cases are much more delicate and were resolved only recently. If two sides are variable and one is fixed then the problem is still NP-hard [28]. In fact, the recent universality theorem of Ref. 18 implies that *every* integer program is a $3 \times m \times n$ transportation problem for some m and n . We now show that if two sides are fixed and one is variable, then the problem can be solved in polynomial time by the Graver based primal algorithms. Note that even over $3 \times 3 \times n$ tables, the only solution of the problem available to-date is the one given below.

Theorem 7 [7]. *For every fixed l and m , there is an algorithm that, given n , integer $l \times m \times n$ cost w , and integer line-sums $v = (v_{*,j,k}, v_{i,*,k}, v_{i,j,*})$, solves Motzkin's triple-index transportation problem (10) in time which is polynomial in n and $\langle w, v \rangle$.*

Proof. Reindex arrays as $x = (x^1, \dots, x^n)$ with each $x^k = (x_{i,j,k})$ a suitably indexed lm vector representing the k th layer of x . Similarly reindex w . Let $t := r := lm$ and $s := (l + m)$. Let $b := (b^0, b^1, \dots, b^n) \in \mathbb{Z}^{r+ns}$, where $b^0 := (v_{i,j,*})$ and

$$b^k := ((v_{*,j,k}), (v_{i,*,k})), \quad k = 1, \dots, n.$$

Let A be the $(t, s) \times t$ bimatix with first block $A_1 := I_t$ the $t \times t$ identity matrix and second block A_2 a matrix defining the line-sum equations on $l \times m$ arrays. Then the equations $A_1(\sum_{k=1}^n x^k) = b^0$ are the line-sum equations $\sum_{k=1}^n x_{i,j,k} = v_{i,j,*}$ where summations over layers occur, whereas the equations $A_2 x^k = b^k$ for $k = 1, \dots, n$ represent all other line-sum equations, where summations are within a single layer at a time. So the transportation problem is encoded as the n -fold program

$$\min \{wx : x \in \mathbb{Z}^{nt}, A^{(n)}x = b, x \geq 0\}.$$

By Theorem 6 part 1, this n -fold integer program, and hence the given multi-index

transportation problem, can be solved in polynomial time.

This proof can be extended to multi-index transportation problems of higher dimension $m_1 \times \dots \times m_d \times n$, to any margin constraints, and to nonlinear objective functions of the forms appearing in the various parts of Theorem 6 [12].

The Multicommodity Transportation Problem. The *multicommodity transportation* problem seeks minimum cost routing of several discrete commodities from m suppliers to n consumers subject to supply, demand, and capacity constraints. The problem data is as follows. There are l types of commodities with $v_k \in \mathbb{Z}_+$ the volume per unit flow of commodity k . Each supplier i has a supply vector $s^i \in \mathbb{Z}_+^l$ with s_k^i its supply in commodity k , and each consumer j has a consumption vector $c^j \in \mathbb{Z}_+^l$ with c_k^j its consumption in commodity k . A multicommodity transportation is a vector $x = (x^1, \dots, x^n)$ with $x^j = (x_{1,1}^j, \dots, x_{1,l}^j, \dots, x_{m,1}^j, \dots, x_{m,l}^j)$, where $x_{i,k}^j$ is the flow of commodity k from supplier i to consumer j . The capacity constraint on channel (i, j) is $\sum_{k=1}^l v_k x_{i,k}^j \leq u_{i,j}$. The cost of transportation x is defined as follows. There are cost functions $f_{i,j} : \mathbb{Z} \rightarrow \mathbb{Z}$ for each channel and $g_{i,k}^j : \mathbb{Z} \rightarrow \mathbb{Z}$ for each channel and commodity. The transportation cost on channel (i, j) is $f_{i,j}(\sum_{k=1}^l v_k x_{i,k}^j) + \sum_{k=1}^l g_{i,k}^j(x_{i,k}^j)$ with the first term being the value of $f_{i,j}$ on the combined volume of flow of all commodities on channel (i, j) , and the second term being the sum of costs that depend on both the channel and the commodity. The total cost of transportation x is

$$\sum_{i,j} \left(f_{i,j} \left(\sum_{k=1}^l v_k x_{i,k}^j \right) + \sum_{k=1}^l g_{i,k}^j(x_{i,k}^j) \right).$$

Our results apply to cost functions which can be standard linear or convex such as $\alpha_{i,j} |\sum_{k=1}^l v_k x_{i,k}^j|^{\beta_{i,j}} + \sum_{k=1}^l \gamma_{i,k}^j |x_{i,k}^j|^{\delta_{i,k}^j}$ for some nonnegative integers $\alpha_{i,j}, \beta_{i,j}, \gamma_{i,k}^j, \delta_{i,k}^j$, which take into account the increase in cost due to channel congestion when subject to heavy traffic or communication load (the linear case occurs with $\beta_{i,j} = \delta_{i,k}^j = 1$).

We consider the problem with fixed (but arbitrary) number l of commodities, fixed number m of suppliers, and *variable* number n of consumers. This is very natural in operations research applications, where few facilities serve many customers. (If the number m of suppliers is also variable then the problem is NP-hard already for $l = 2$ commodities.) We point out that even with only $m = 3$ suppliers and $l = 2$ commodities, the only solution available to-date is the one given below.

We have the following theorem on (non-linear) multicommodity transportation. As before, $\hat{f}, \hat{g}$ denote the maximum absolute values of f, g over the feasible set.

Theorem 8 [21]. *For any fixed l commodities, m suppliers, and volumes v_k , there is an algorithm that, given n consumers, supplies and demands $s^i, c^j \in \mathbb{Z}_{+}^l$, capacities $u_{i,j} \in \mathbb{Z}_{+}$, and convex costs $f_{i,j}, g_{i,k}^j : \mathbb{Z} \rightarrow \mathbb{Z}$ presented by evaluation oracles, solves in time polynomial in n and $\langle s^i, c^j, u, \hat{f}, \hat{g} \rangle$, the multicommodity transportation problem,*

$$\begin{aligned} \min \quad & \sum_{i,j} \left(f_{i,j} \left(\sum_{k=1}^l v_k x_{i,k}^j \right) + \sum_{k=1}^l g_{i,k}^j (x_{i,k}^j) \right) \\ \text{s.t.} \quad & x_{i,k}^j \in \mathbb{Z}, \quad \sum_j x_{i,k}^j = s_k^i, \quad \sum_i x_{i,k}^j = c_k^j, \\ & \sum_{k=1}^l v_k x_{i,k}^j \leq u_{i,j}, \quad x_{i,k}^j \geq 0. \end{aligned}$$

Proof. Construct bimatrices A and W as follows: Let D be the $(l, 0) \times l$ bimatrix with first block $D_1 := I_l$ and second block D_2 empty. Let V be the $(0, 1) \times l$ bimatrix with first block V_1 empty and second block $V_2 := (v_1, \dots, v_l)$. Let A be the $(ml, l) \times ml$ bimatrix with first block $A_1 := I_{ml}$ and second block $A_2 := D^{(m)}$. Let W be the $(0, m) \times ml$ bimatrix with first block W_1 empty and second block $W_2 := V^{(m)}$. Let b be the $(ml + nl)$ -vector $b := (s^1, \dots, s^m, c^1, \dots, c^n)$.

Let $f : \mathbb{Z}^{nm} \rightarrow \mathbb{Z}$ and $g : \mathbb{Z}^{nml} \rightarrow \mathbb{Z}$ be the separable convex functions defined by $f(y) := \sum_{i,j} f_{i,j}(y_{i,j})$ with $y_{i,j} := \sum_{k=1}^l v_k x_{i,k}^j$ and $g(x) := \sum_{i,j} \sum_{k=1}^l g_{i,k}^j(x_{i,k}^j)$.

Now note that $A^{(n)}x$ is an $(ml + nl)$ -vector, whose first ml entries are the flows from each supplier of each commodity to all consumers, and whose last nl entries are the flows to each consumer of each commodity from all suppliers. Therefore the supply and consumption equations are encoded by $A^{(n)}x = b$. Next note that the nm -vector $y = (y_{1,1}, \dots, y_{m,1}, \dots, y_{1,n}, \dots, y_{m,n})$ satisfies $y = W^{(n)}x$. So the capacity constraints become $W^{(n)}x \leq u$ and the cost function becomes $f(W^{(n)}x) + g(x)$. Therefore, the problem is precisely the following n -fold integer program,

$$\min \left\{ f(W^{(n)}x) + g(x) : x \in \mathbb{Z}^{nml}, A^{(n)}x = b, W^{(n)}x \leq u, x \geq 0 \right\}.$$

By Theorem 6 part 5, this program can be solved in polynomial time as claimed.

REFERENCES

1. Schrijver A. Theory of linear and integer programming. New York: Wiley; 1998.
2. Hoffman AJ, Kruskal JB. Integral boundary points of convex polyhedra. Linear inequalities and Related Systems, Annals of Mathematics Studies 38:223–246. Princeton (NJ): Princeton University Press; 1956.
3. Lenstra HW Jr. Integer programming with a fixed number of variables. Math Oper Res 1983;8:538–548.
4. Haus U, Köppe M, Weismantel R. The integral basis method for integer programming. Math Meth Oper Res 2001;53:353–361.
5. Haus U, Köppe M, Weismantel R. A primal all-integer algorithm based on irreducible solutions. Math Program 2003;96: 205–246.
6. Graver JE. On the foundations of linear and linear integer programming I. Math Program 1975;9:207–226.
7. De Loera J, Hemmecke R, Onn S, *et al.* N-fold integer programming. Disc Optim 2008;5 (Volume in memory of George B. Dantzig): 231–241.
8. De Loera J, Hemmecke R, Onn S, *et al.* Convex integer maximization via Graver bases. J Pure Appl Algeb 2009;213:1569–1577.

9. Hemmecke R, Onn S, Weismantel R. A polynomial oracle-time algorithm for convex integer minimization. *Math Program* 2009. DOI: 10.1007/s10107-009-0276-7.
10. Gordan P. Über die Auflösung linearer Gleichungen mit reellen Coefficienten. *Math Ann* 1873;6:23–28.
11. Onn S. Theory and applications of n-fold integer programming. In: Lee J, Leyffer S, editors. *IMA volume on mixed integer nonlinear programming*, Frontier Series. Springer. pp. 1–35. In press.
12. Onn S. Nonlinear discrete optimization: an algorithmic theory. *Zürich Lectures in Advanced Mathematics*. European Mathematical Society. pp. 1–143. In press.
13. Cook W, Fonlupt J, Schrijver A. An integer analogue of Carathéodory's theorem. *J Combin Theor Ser B* 1986;40:63–70.
14. Sebö A. Hilbert bases, Carathéodory's theorem and combinatorial optimization. In: Kannan R, Pulleyblank WR, editors. *Proceedings IPCO 1 – 1st Conference on Integer Programming and Combinatorial Optimization*. Waterloo, Ontario: University of Waterloo Press; 1990. pp. 431–455.
15. Lee J, Onn S, Weismantel R. On test sets for nonlinear integer maximization. *Oper Res Lett* 2008;36:439–443.
16. Murota K, Saito H, Weismantel R. Optimality criterion for a class of nonlinear integer programs. *Oper Res Lett* 2004;32:468–472.
17. Onn S, Rothblum UG. Convex combinatorial optimization. *Disc Comput Geom* 2004;32: 549–566.
18. De Loera J, Onn S. All linear and integer programs are slim 3-way transportation programs. *SIAM J Optim* 2006;17:806–821.
19. Hoşten S, Sullivant S. Finiteness theorems for Markov bases of hierarchical models. *J Combin Theor Ser A* 2007;114:311–321.
20. Santos F, Sturmfels B. Higher Lawrence configurations. *J Combin Theor Ser A* 2003;103: 151–164.
21. Hemmecke R, Onn S, Weismantel R. N-fold integer programming and nonlinear multi-transshipment. Submitted for publication.
22. Vlach M. Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Disc Appl Math* 1986;13:61–78.
23. Yemelichev VA, Kovalev MM, Kravtsov MK. *Polytopes, graphs and optimisation*. Cambridge: Cambridge University Press; 1984.
24. Cox LH. On properties of multi-dimensional statistical tables. *J Stat Plan Infer* 2003;117: 251–273.
25. Fienberg SE, Rinaldo A. Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J Stat Plan Infer* 2007;137:3430–3445.
26. Motzkin TS. The multi-index transportation problem. *Bull Am Math Soc* 1952;58:494.
27. Irving R, Jerrum MR. Three-dimensional statistical data security problems. *SIAM J Comput* 1994;23:170–184.
28. De Loera J, Onn S. The complexity of three-way statistical tables. *SIAM J Comput* 2004;33:819–836.

PORTUGUESE OPERATIONAL RESEARCH SOCIETY—APDIO

DOMINGOS M. CARDOSO
Departamento de Matemática,
Universidade de Aveiro,
Aveiro, Portugal

PEDRO OLIVEIRA
Instituto de Ciências Biomédica
Abel Salazar, Universidade
do Porto, Porto, Portugal

MISSION AND ORGANIZATION

The Portuguese Operational Research (OR) Society (in Portuguese: Associação Portuguesa de Investigação Operacional—APDIO) is a scientific society that brings together the Portuguese community interested in OR. APDIO was founded in 1978 by 140 founding members, including university researchers, industrial professionals, and several Portuguese institutes and companies as institutional associates. The list of founding institutional associates included LISNAVE—Shipyard company, PETROGAL—Oil and Gas Company of Portugal, INESC—Institute of Systems Engineering and Computers, Portuguese General Directorate of Hydro-energy, LNETI—National Laboratory of Engineering and Industrial Technology, PROFABRIL—Portuguese Consulting Group, CIFAG—Government Investment Institute, Espírito Santo e Comercial Bank of Lisbon, CESUR—Center for Urban and Regional Systems, CTT—Post and Telecommunications of Portugal, CAUTL—Center of Automation of the Technical University of Lisbon, INA—National Institute for Public Administration, Direcção Geral de Energia—Portuguese General Directorate for Energy Policy, and PARTEX—Oil and Gas Corporation.

The main objective of the APDIO is to disseminate the most recent advances in OR

and its best practices and, furthermore, to strengthen the links between the OR community members on the basis of their research interests and challenges.

The APDIO defines its mission as contributing to the advance of OR through courses, seminars, workshops, conferences, and publications (such as scientific journals, bulletins, and books). Furthermore, the APDIO promotes the exchange of results, experiences, and OR applications by electronic means such as its web page ([//apdio.pt/home](http://apdio.pt/home)).

The APDIO is composed of a General Assembly (which is directed by a president and two secretaries and includes all its members), a Directive Committee (composed of a president, four vice presidents, a treasurer, and a secretary), and an Accounting Committee (composed of a president, a rapporteur, and a secretary). Each of these committees is elected for a period of two years. The General Assembly and the Accounting Committee meet regularly once a year, and the Directive Committee meets every two months. At present, the APDIO has about 300 associate members, and it is hosted (since its beginning) at CESUR—IST, Technical University of Lisbon (E-mail: apdio@civil.ist.utl.pt).

THE APDIO ACTIVITIES

The Portuguese OR Society is a member of the International Federation of Operational Research Societies (IFORS) and the International Federation of Automatic Control IFAC since its very beginning. Later, it became a member of the Association of European Operational Research Societies (EURO) and was involved in the foundation, in 1982, of the Association of Latin–Iberoamerican Operational Research Societies (ALIO). APDIO has a reciprocity agreement with the Brazilian Society of Operations Research (SOBRAPO) since 2001.

The APDIO has organized 15 national conferences in the following places and years: Lisboa, 1982; Porto, 1984; Coimbra, 1987;

Lisboa, 1989; Évora, 1992; Braga, 1994; Aveiro, 1996; Faro, 1998; Setúbal, 2000; Guimarães, 2002; Porto, 2004; Lisboa, 2006; Vila Real, 2008; Caparica, 2009; and Coimbra, 2011. The next National Conference will be held in Bragança in 2013.

The APDIO has organized and supported about 30 international conferences; among these conferences, we highlight the following:

- EURO VIII—European Conference on Operational Research, Lisbon, September, 1986
- IFORS 93—XIII World Conference on Operations Research, Lisbon, July, 1993
- The Optimization Conferences (Braga, 1995; Coimbra, 1998; Aveiro, 2001; Lisbon, 2004; Porto, 2007; and Lisbon, 2011)
- EURO XXIV—European Conference on Operational Research, Lisbon, July, 2010

The scientific journal of the APDIO, *Investigação Operacional*, was published between 1981 and 2008, totaling 28 volumes.

The editors in chief were Isabel Hall Themido, Technical University of Lisbon, 1981–1983; José Manuel Viegas, Technical University of Lisbon, 1984–1985; José Pinto Paixão, University of Lisbon, 1986–1987; Joaquim João Júdice, University of Coimbra, 1988–2004; and José Fernando Oliveira, University of Porto, 2004–2008.

The newsletter of the APDIO, *Boletim*, was published, twice a year, between 1984 and 2005, and its publication was resumed in 2010. The editors were José Manuel Viegas, Technical University of Lisbon, 1984–1986; António José Rodrigues, University of Lisbon, 1987–1990; Jorge Pinho de Sousa, University of Porto, 1991–1992; Carlos Henggeler Antunes, University of Coimbra, 1993–1999; João Paulo Costa, University of Coimbra, 2000–2002; João Luís Soares, University of Coimbra, 2003–2005; and Ana Camanho and Bernardo Almada Lobo, University of Porto, 2010–2011.

The current editors are Ana Luísa Custódio and Isabel Correia, New University of Lisbon.

Since OR's visibility comes from its applications in industry and services, in 2000, APDIO sponsored the publication of a book about case studies of real-world applied problems of OR in Portugal (in Portuguese), edited by Carlos Henggeler Antunes (University of Coimbra) and Luís Valadares Tavares (Technical University of Lisbon) [1]. Now, another book reporting successful applications of OR, edited by José Soeiro Ferreira (University of Porto) and Rui Carvalho Oliveira (Technical University of Lisbon), is in preparation.

Over the past 30 years, the activities of the Portuguese OR Society were held, mostly through its associates, researchers, and practitioners, who have dedicated much of their effort to promote the OR through the APDIO. We just mention those who had the utmost responsibility, the presidents of the Portuguese OR Society. They were Luís Valadares Tavares, Technical University of Lisbon, 1978–1985; José Dias Coelho, New University of Lisbon, 1986–1989; Rui Campos Guimarães, University of Porto, 1990–1994; Isabel Hall Themido, Technical University of Lisbon, 1995–1996; Luís Valadares Tavares, Technical University of Lisbon, 1997–1998; José Soeiro Ferreira, University of Porto, 1999–2001; José Valério de Carvalho, University of Minho, 2002–2004; Jorge Pinho de Sousa, University of Porto, 2005–2009; and Joaquim João Júdice, University of Coimbra, 2010–2011.

FUTURE CHALLENGES

Future challenges can be summarized in the strengthening of the presence of the OR community in the society at large. The present financial crisis, although with negative direct effects on financing research and higher education institutions, can become an opportunity for the OR community. The pressure to reduce costs at all levels and for a more efficient use of limited resources motivates the OR community to intervene, using its techniques and tools to propose solutions. In particular, the current Directive Committee has defined two major areas of public intervention, energy, and health systems (namely, the problems raised by their high costs and

its impact on the economic and social development of the country).

Most of the APDIO members belong to the Portuguese universities and polytechnic institutes. However, several private institutions and companies are also presented by their OR practitioners and some of them as institutional associates. One of the main goals of the current Directive Committee is to increase the participation of industrial and service sectors in APDIO initiatives and to improve the utilization of OR methodologies in such environments. The one-day meetings “OR in Companies” with the aim of analyzing particular problems by academics and practitioners and sharing their experience will be more frequent and extended to more institutions.

The Portuguese OR Society also considers that its presence in the education system has to be reinforced. In the first place, the APDIO must intervene in the definition of the mathematics curricula of secondary education, so that the curricula can be an introduction to the field of knowledge, which is largely unknown to the general public. With respect to higher education, the APDIO feels that the Bologna reform may have, in some graduations, reduced the presence of OR in the

curricula. Therefore, the teaching of OR has been largely shifted to postgraduate level [2]. In this particular aspect, the APDIO is challenging the different universities, for the development of a PhD program in OR. Research is also of a great concern, in particular, with the expected reduced funds. APDIO, jointly with the different research groups, is pressing the Portuguese National Science Foundation for the creation of a funding area dedicated solely to OR.

Finally, it is worth to mention that we believe in the continuity of the APDIO as an institutional research reference for the OR community in Portugal. The renovation of generations and the new social and economic challenging problems for the development of human societies ensures that OR will be even more visible and more effective in the near future.

REFERENCES

1. Antunes CH and Tavares LV. *Casos de Aplicação da Investigação Operacional em Portugal*. Portugal, Lisboa: McGraw-Hill; 2000.
2. Tavares LV. *Investigação Operacional em Portugal: os próximos anos*. Boletim da APDIO 2010; 42, 3.

POWER INDICES

JOSEP FREIXAS

Department of Applied Mathematics 3
and High Engineering School
(Campus of Manresa), Technical
University of Catalonia, Madrid,
Spain

INTRODUCTION

Roughly speaking, a power index is a numerical measure that estimates the a priori capacity or influence of each voter in a simple game. Of course the notion of power is complex and has been analyzed in depth by several authors. An interesting reference is the subtle book by Morriss [1], which analyzes power under a philosophical point of view.

Simple games form an interesting subclass of transferable utility cooperative games and are frequently used as a test bed for cooperative concepts. In a simple game, the members of a coalition can be seen as those who vote in favor of a certain law, while the others vote against it. The binary output $\{0, 1\}$ in simple games converts them in appropriate models for binary voting systems in which each coalition either becomes winning (output 1) and can force the passage of the bill at hand, or becomes losing (output 0) and cannot force the passage of the bill. It should be noted that the binary output allows the description of the game by only giving the set of winning (or, alternatively, losing) coalitions.

Several power indices found in the literature are, simply, the restriction of a cooperative value to this subclass: for example, the Shapley–Shubik index [2] is the restriction of the well-known Shapley value [3]. The power of a voter as discussed in Ref. 2 can be viewed as his/her *expected share in a fixed total prize*, that is, the power of a voter is meant to be voter's expected payoff. This idea was followed by several scholars (e.g. Deegan and Packel [4], Johnston [5]

and Holler [6], among others) who introduced alternative efficient power indices. All these authors followed the seminal ideas by Dubey [7] and Dubey and Shapley [8], and provided axiomatic characterizations of their indices. Hence, the acceptance of an index as a valid measure of power depends on how we are willing to accept the properties that uniquely characterize it.

Another alternative point of view, far from cooperative game theory, was followed by other scholars, for which a member's voting power is the degree to which that member's vote is able to *influence* the outcome of a division: whether the bill in question will pass or fail. A way to formally explicate this idea is in terms of probabilities. The three authors—Penrose [9], Banzhaf [10], and Coleman [11]—who, independent of each other, offered a mathematical explication of power as influence, adopted a probabilistic approach.

Felsenthal and Machover [12] in collaboration with Zwicker distinguished between these two previous sources of power. However, this distinction is not exclusive since; for example, the Shapley–Shubik index can be derived from the two approaches.

The main assumption when computing power indices is that all relevant information is encapsulated in the game. Hence, power indices can directly be computed from the characteristic function or, equivalently, from its set of winning coalitions. However, there are many other more sophisticated models that require exogenous information, specially concerning the formation of coalitions. Both sources have been extensively treated in the literature.

Simple games have served as theoretical models in many fields far from game theory, in spite of the fact that no rationality needs to be assumed for the “agents” in these fields. Nevertheless, the main application of simple games has always been the analysis of binary voting procedures and hence of political structures. Note, for instance, that the Barlow and Proshan index in reliability

corresponds to the Shapley–Shubik index and the Birnbaum index in reliability corresponds to the Banzhaf index. Here we regard only power indices in the politics context.

This short compendium on power indices is organized as follows. Section titled “Simple Games” recalls the notion of simple game and introduces some subclasses of them mainly based on the completeness of some preorderings. Section titled “Power Indices” provides the definitions of some of the most recognized power indices. Section titled “Properties of Power Indices” lists some reasonable properties for power indices, but disregards many others for lack of space, and explains which of them are satisfied for the power indices introduced. Section titled “Axiomatizations versus Probabilistic Approach” contains a brief probabilistic approach of some indices. Section titled “Discussion” summarizes the occurrence of some desirable properties for power indices and considers their ordinal equivalence. Finally, section titled “Miscellaneous” contains a compendium of significant issues related with power indices that have been developed in the literature.

SIMPLE GAMES

In the sequel, $N = \{1, 2, \dots, n\}$ denotes a fixed but otherwise arbitrary finite set of *players*. Any subset $S \subseteq N$ is a *coalition* and N itself is called the *grand coalition*. A cooperative game v (in N , omitted hereafter) is a *simple game* if (i) $v(S) = 0$ or 1 for all S , (ii) is monotonic, that is, $v(S) \leq v(T)$ whenever $S \subset T$, and (iii) $v(N) = 1$ (if condition (iii) is not required, then $v \equiv 0$ and hence $\mathcal{W} = \emptyset$ is also a simple game, which is convenient when restriction of simple games is considered, see first paragraph of section titled “Coalition Structures”). Either the family of *winning* coalitions $\mathcal{W} = \mathcal{W}(v) = \{S \subseteq N : v(S) = 1\}$ or the subfamily of *minimal* winning coalitions $\mathcal{W}^m = \mathcal{W}^m(v) = \{S \in \mathcal{W} : T \subset S \Rightarrow T \notin \mathcal{W}\}$ determines the game. We denote the set of simple games by S_N on N .

Von Neumann and Morgenstern [13] considered a very restrictive class of simple games. Some authors (Shapley in Ref. 3) extended their class. They consider the

definition above plus the requirement of *properness*: whenever $S \in \mathcal{W}$ and $T \in \mathcal{W}$ then $S \cap T \neq \emptyset$. Improper games rarely occur in politics, but they appear, for example, when the US Supreme Court gives a grant of *certiorari*, which requires approval by four out of the nine justices of court.

For any $\mathcal{W}, \mathcal{V} \in S_N$, we can define two new simple games by the ordinary set operations:

$$\begin{aligned}\mathcal{W} \cap \mathcal{V} &= \{S \in N : S \in \mathcal{W} \text{ and } S \in \mathcal{V}\}, \\ \mathcal{W} \cup \mathcal{V} &= \{S \in N : S \in \mathcal{W} \text{ or } S \in \mathcal{V}\}.\end{aligned}$$

Endowed with these set operations, S_N becomes a distributive lattice.

Power index. Any map $f : S_N \rightarrow \mathbb{R}^n$ with $|N| = n$ is a *power index*. The i th component is interpreted as a measure of the power that the simple game $\mathcal{W}$ confers to i . Any power index f induces a total preordering on the set of voters by $i \succsim_f j \Leftrightarrow f(i) \geq f(j)$. In this case $\succsim_f$ is called the *separability* relation for the power index f .

Crucial voters. This concept is in the ground of the definition of most power indices. We say that player i is *crucial* for a coalition S if $i \in S$ and $S \in \mathcal{W}$ but $S \setminus \{i\} \notin \mathcal{W}$. We will use notations:

$$\begin{aligned}C_i &= \{S \subseteq N : i \text{ is crucial in } S\}, \\ C_i(k) &= \{S \subseteq N : i \text{ is crucial in } S \\ &\quad \text{and } |S| = k\}.\end{aligned}$$

Some Significant Subclasses of Simple Games

There are two important subclasses of simple games that are worth taking into account: complete games (which include the well-known weighted games) and weakly complete games. Both subclasses are characterized by the completeness of a certain preordering on N .

There are at least two purposes of introducing these subclasses. First, practically all real-life binary voting systems are of some of these types and, second, the two preorderings we are going to introduce attempt to formalize the intuitive notion that underlies expressions such as the following: “ i and j have equal power,” “ i has at least as much power as j ,” “ i is preferable to j as a coalitional partner.”

Desirability relation. It was introduced by Isbell [14]. Let $\mathcal{W} \in \mathcal{S}_N$, $i, j \in N$. Define

$$i \succsim_D j \quad \text{iff} \quad S \cup \{j\} \in \mathcal{W} \Rightarrow S \cup \{i\} \in \mathcal{W} \\ \text{for all } S \subseteq N \setminus \{i, j\}.$$

The desirability relation $\succsim_D$ is a preordering on N (i.e., $\succsim_D$ is a reflexive and transitive relation). We write $i \succ_D j$ if $i \succsim_D j$ but $j \not\succsim_D i$, and $i \approx_D j$ stands for $i \succsim_D j$ and $j \succsim_D i$. A simple game $\mathcal{W}$ in N is *complete* whenever the desirability relation $\succsim_D$ is total.

A simple game $\mathcal{W}$ in N is a *weighted game* iff there exist nonnegative *weights* $w_1, w_2, \dots, w_n$ allocated to the players and a *quota* $q \in (0, \sum_{i \in N} w_i]$ such that $S \in \mathcal{W}$ iff $\sum_{i \in S} w_i \geq q$. We then write $\mathcal{W} \equiv [q; w_1, w_2, \dots, w_n]$. Any weighted game is complete because $w_i \geq w_j$ implies $i \succsim_D j$. Weighted games are the most frequent real-life voting procedures of simple games.

Complete games have been widely studied. Taylor and Zwicker [15,16] gave respective characterizations of weighted games and complete games in terms of trades among players and within coalitions. In Carreras and Freixas [17] an existence, uniqueness, and classification theorem for complete games was provided that enables us to enumerate all these games up to isomorphism. This theorem also supplies an alternative efficient way to determine, in terms of a set of inequalities, which complete games are weighted. The numbers in Table 1 were obtained using this result, with the exception of weighted games for $n = 9$, which was obtained from Tautenhahn [18].

The sources of real-life examples of complete games are the count and account games or games with consensus (see Peleg [19] and Carreras and Freixas [20] among others).

The weak desirability relation. It was considered by Carreras and Freixas [21]. Let

$\mathcal{W} \in \mathcal{S}_N$ and $i, j \in N$. Define

$$i \succsim_d j \quad \text{iff} \quad |C_i(k)| \geq |C_j(k)| \quad \text{for all} \\ k = 1, 2, \dots, n.$$

A simple game $\mathcal{W}$ in N is *weakly complete* whenever $\succsim_d$, the weak desirability relation on N , is total.

Any complete game is a weakly complete game because $i \succsim_D j$ implies $i \succsim_d j$.

Examples

The three examples below are real-life binary voting games. Example 1 models the first European Economic Community where France, Germany, and Italy were given four votes each, while Belgium and the Netherlands were given two votes and Luxembourg one. Passage required a total of at least 12 of 17 votes. It is an example of a weighted game. Example 2 is not weighted, but complete and is equivalent to an old system to amend the Canadian Constitution ([22] and [23]). Example 3 corresponds to the United States Federal System, which is not complete, but weakly complete. For compelling explanations on these real-life examples, we refer the interested reader to Taylor's book [24].

1. [12; 4, 4, 4, 2, 2, 1];
2. [9; 3, 3, 1, 1, 1, 1, 1, 1, 1] $\cap$ [7; 1, 1, 1, 1, 1, 1, 1, 1, 1];
3. $\left[67; \underbrace{0, \dots, 0}_{435}, \underbrace{1, \dots, 1}_{100}, 0.5, 16.5 \right] \cap$
 $\left[290; \underbrace{1, \dots, 1}_{435}, \underbrace{0, \dots, 0}_{100}, 0, 72 \right].$

Table 1. Number of Complete (CG) and Weighted Games (WG) with n Voters

n	1	2	3	4	5	6	7	8	9
CG	1	3	8	25	117	1171	44,313	16,175,188	284,432,730,174
WG	1	3	8	25	117	1111	29,373	27,30,164	989,913,344

POWER INDICES

We briefly introduce some of the most recognized power indices. The first two are the most distinguished power indices in the literature and are particular cases of semivalues, which are introduced immediately afterward; all of them are weighted sums of the numbers $|C_i(k)|$ when k ranges from 1 to n . For each player $i \in N$, we consider the following power indices.

Shapley–Shubik index. It was introduced by Shapley and Shubik [2] and is the restriction of the Shapley value [3] to simple games (set $|S| = s$ here).

$$\begin{aligned}\phi_i(\mathcal{W}) &= \sum_{S \in \mathcal{C}_i} \frac{(s-1)!(n-s)!}{n!} \\ &= \sum_{k=1}^n |C_i(k)| \frac{1}{\binom{n-1}{k-1} n}.\end{aligned}\quad (1)$$

Apart from the axiomatic treatment we discuss later, the Shapley–Shubik index, as defined in Equation (1), can be given the following heuristic explanation: suppose that the players agree to meet at a specified place and time. Naturally all arrive at different times; it is assumed, however, that all orders of arrival (permutations of the players) have the same probability $\frac{1}{n!}$. Suppose that if a player i arrives and finds the members of the coalition $S \setminus \{i\}$ (and no others) already there, he/she receives 1 as payoff if $S \in \mathcal{C}_i$ and 0 otherwise.

Banzhaf indices (absolute and relative respectively).

$$\beta_i(\mathcal{W}) = \frac{\eta_i(\mathcal{W})}{2^{n-1}}, \quad \beta'_i(\mathcal{W}) = \frac{\eta_i(\mathcal{W})}{\sum_{j=1}^n \eta_j(\mathcal{W})},$$

where the raw Banzhaf index $\eta_i(\mathcal{W})$ is the number of coalitions for which a voter is crucial, that is, $\eta_i(\mathcal{W}) = |\mathcal{C}_i| = \sum_{k=1}^n |C_i(k)|$.

Semivalues as power indices. Semivalues were first introduced in Weber [25] only for simple games and extended to all cooperative games in Dubey *et al.* [26]. There is an interesting characterization of semivalues in terms of *weighting coefficients*, provided by Dubey *et al.* [26] from a group of axioms

for cooperative games. For every weighting vector $p = (p_1, p_2, \dots, p_n)$ such that $p_k \geq 0$ for all $k = 1, \dots, n$ and $\sum_{k=1}^n p_k \binom{n-1}{k-1} = 1$, let ψ be defined by

$$\psi_i(\mathcal{W}) = \sum_{k=1}^n p_k |C_i(k)|$$

for all $i \in N$ and all $\mathcal{W} \in \mathcal{S}_N$. Then ψ is a semivalue. Conversely, any semivalue can be defined in this way, and the correspondence $p \mapsto \psi$ is one-to-one. If p_k is interpreted as the probability that a given player i joins a coalition of size k , provided that all the coalitions of a given size have the same probability of being joined, then, in the case of simple games, $\phi_i(\mathcal{W})$ gives the probability of player i to be crucial under this random scheme, so that it can be interpreted as a measure of this player's expected power.

For example, the Shapley–Shubik index ϕ is defined by weighting coefficients $p_k = 1/\binom{n-1}{k-1}$, and therefore $p_k \binom{n-1}{k-1}$ is constant for all k . The absolute Banzhaf index β is the only semivalue whose weighting coefficients are constant, that is, $p_k = 1/2^{n-1}$ for all k . A *semivalue* ψ is *regular* if $p_k > 0$ for all $k = 1, \dots, n$. In Carreras and Freixas [27] there are characterizations for regular semivalues and for semivalues with binomial coefficients, that is, a semivalue ψ^p is *p-binomial* whenever $p_k = \alpha^{k-1}(1-\alpha)^{n-k}$ for all $k = 1, \dots, n$ and $\alpha \in [0, 1]$. Semivalues have been studied as power indices in Carreras *et al.* [28] and have been applied to real-life situations Refs 27, 29–32.

Johnston index. It was introduced by Johnston [5]. For any winning coalition $S \in \mathcal{W}$, we define $\mathcal{X}(S)$ as the number of elements of S which are crucial for it:

$$\mathcal{X}(S) = |\{j \in S : S \in \mathcal{C}_j\}|$$

and, for each $i \in N$ and each $1 \leq h \leq n$, $\mathcal{C}_i^h$ is the collection of winning coalitions for which voter i and $h-1$ other elements are crucial, that is, $\mathcal{C}_i^h = \{S \in \mathcal{C}_i : \mathcal{X}(S) = h\}$. The raw Johnston is defined as

$$r_i(\mathcal{W}) = \sum_{S \in \mathcal{C}_i} \frac{1}{\mathcal{X}(S)} = \sum_{h=1}^n |\mathcal{C}_i^h| \frac{1}{h}$$

and the Johnston index is

$$\gamma_i(\mathcal{W}) = \frac{r_i(\mathcal{W})}{\sum_{j=1}^n r_j(\mathcal{W})}$$

Note that to measure power the Johnston index equally counts all coalitions with crucial voters and, within each coalition, equally divides power among crucial voters for it.

The indices below are essentially funded by probabilistic approaches.

Rae index. It was introduced in Rae [33] and is based on two assumptions: the votes are independent from one another, and each player votes “yes” with probability 1/2 and votes “no” with probability 1/2.

$$\mathcal{R}_i(\mathcal{W}) = \frac{|\{S : i \in S \in \mathcal{W}\}|}{2^n} + \frac{|\{S : i \notin S \notin \mathcal{W}\}|}{2^n}.$$

The Rae index captures the idea of “success”. It is understood that a voter is satisfied if he/she votes “yes” and the bill is passed or he/she votes “no” and the bill is rejected.

Coleman indices. Coleman [11,34] defines three different indices in terms of different ratios. The first is an index associated to the game and therefore is of different nature to the others. Hence, we do not consider it in the next two sections. The “power of a collectivity to act” measures the ease of decision making by means of a simple game $\mathcal{W}$, and is given by

$$\mathcal{A}(\mathcal{W}) = \frac{|\{S : S \in \mathcal{W}\}|}{|\{S : S \subseteq N\}|} = \frac{|\mathcal{W}|}{2^n}.$$

This notion, extensively treated in Carreras [35], can be used to measure the level of inertia/protectionism of any voting procedure

and to decide whether this level is or not a convenient feature.

Voter i ’s “Coleman index to prevent” action (C_i^P) is given by

$$C_i^P(\mathcal{W}) = \frac{|\{S : i \in S \in \mathcal{W}, S \setminus \{i\} \notin \mathcal{W}\}|}{|\mathcal{W}|} = \frac{\eta_i(\mathcal{W})}{|\mathcal{W}|},$$

while voter i ’s “Coleman index to initiate” action (C_i^I) is given by

$$C_i^I(\mathcal{W}) = \frac{|\{S : i \notin S \notin \mathcal{W}, S \cup \{i\} \in \mathcal{W}\}|}{|2^{N \setminus \mathcal{W}}|}.$$

Brams and Affuso [36] note that Banzhaf and Coleman indices are proportional:

$$\frac{\beta_i(\mathcal{W})}{\sum_{j \in N} \beta_j(\mathcal{W})} = \frac{\beta'_i(\mathcal{W})}{\sum_{j \in N} \beta'_j(\mathcal{W})} = \frac{C_i^P(\mathcal{W})}{\sum_{j \in N} C_j^P(\mathcal{W})} = \frac{C_i^I(\mathcal{W})}{\sum_{j \in N} C_j^I(\mathcal{W})},$$

but this coincidence due to normalization may provoke the loss of their probability interpretation.

König and Bräuninger’s inclusiveness index. It was introduced in König and Bräuninger [37], and it can be understood also as a measure of success.

$$KB_i(\mathcal{W}) = \frac{|\{S : i \in S \in \mathcal{W}\}|}{|\mathcal{W}|} = \frac{|\mathcal{W}_i|}{|\mathcal{W}|}.$$

Computing these power indices by hand for a limited number of voters is quite simple; we provide here the results for Example 1 and the main introduced indices.

$$\phi = \left(\frac{14}{60}, \frac{14}{60}, \frac{14}{60}, \frac{9}{60}, \frac{9}{60}, 0\right), \quad \beta = \left(\frac{5}{16}, \frac{5}{16}, \frac{5}{16}, \frac{3}{16}, \frac{3}{16}, 0\right), \quad \beta' = \left(\frac{5}{21}, \frac{5}{21}, \frac{5}{21}, \frac{3}{21}, \frac{3}{21}, 0\right),$$

$$\gamma = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, 0\right), \quad \mathcal{R} = \left(\frac{21}{32}, \frac{21}{32}, \frac{21}{32}, \frac{19}{32}, \frac{19}{32}, \frac{16}{32}\right), \quad \mathcal{A} = \frac{7}{32},$$

$$C^P = \left(\frac{5}{7}, \frac{5}{7}, \frac{5}{7}, \frac{3}{7}, \frac{3}{7}, 0\right), \quad C^I = \left(\frac{5}{25}, \frac{5}{25}, \frac{5}{25}, \frac{3}{25}, \frac{3}{25}, 0\right), \quad KB = \left(\frac{12}{14}, \frac{12}{14}, \frac{12}{14}, \frac{10}{14}, \frac{10}{14}, \frac{7}{14}\right).$$

PROPERTIES OF POWER INDICES

We start by considering some properties essentially following Freixas and Gambarelli [38] or Felsenthal and Machover [39]. For any voter $i \in N$, we consider a numerical measure $f_i(\mathcal{W})$ of the power either as voter's expected payoff or as influence.

Efficiency. If we consider power derived from cooperative game theory as expected share in a fixed total prize, then the measure should be *relative*, in the sense that $f_i(\mathcal{W})$ should measure what *share* of the total power of the grand coalition is held by i as compared to all the other voters. Algebraically, the relative character of a measure is expressed by the *efficiency* property:

$$\sum_{i=1}^n f_i(\mathcal{W}) = 1. \quad (2)$$

The indices ϕ, β', γ satisfy efficiency, while $\beta, \psi \neq \phi$ (i.e., any semivalue different from the Shapley–Shubik one) $\mathcal{R}, C^P, C^I, KB$ do not.

Positivity. The monotonicity property required for a simple game also suggests that each individual measure assigned to a voter i must be nonnegative, that is to say the *positivity* property:

$$f_i(\mathcal{W}) \geq 0. \quad (3)$$

All considered indices satisfy positivity. Note that the Rae index and König and Bräuninger's inclusiveness index also satisfy the more restrictive property: $f_i(\mathcal{W}) \geq 1/2$.

Invariance under isomorphism. In order to formulate it we need the following. Let $(N, \mathcal{W})$ and $(N', \mathcal{W}')$ be arbitrary simple games. An *isomorphism* from $(N, \mathcal{W})$ to $(N', \mathcal{W}')$ is a one-to-one mapping g of N onto N' such that, for any coalition $S \subseteq N$, $S \in \mathcal{W} \iff g(S) \in \mathcal{W}'$, where $g(S) = \{g(i) : i \in S\}$. If $(N, \mathcal{W})$ and $(N', \mathcal{W}')$ are isomorphic simple games—that is, if there is an isomorphism from one to the other—then they may be regarded as mere copies of each other, in the sense that they have an identical structure. If g is an isomorphism from $\mathcal{W}$ to $\mathcal{W}'$, then

$$f_i(\mathcal{W}) = f_{g(i)}(\mathcal{W}'). \quad (4)$$

All considered indices satisfy invariance under isomorphism.

Null voter. In a simple game, a null voter ($i \in N$ is *null* if $i \notin S$ for all $S \in \mathcal{W}^m$) has no power, that is to say,

$$f_i(\mathcal{W}) = 0 \quad \text{if } i \text{ is a null voter.} \quad (5)$$

All considered indices satisfy this property with the sole exceptions of $\mathcal{R}$ and KB , which satisfy the analogous property for measures of success rather than decisiveness: $f_i(\mathcal{W}) = 1/2$ for a null voter i . This is because $\mathcal{R}$ and KB are seen as measures of success, while the others are seen as measures of cruciality or decisiveness (see more details in section titled “Probabilistic Approach”).

Some Additional Properties

Let us consider some more properties of power indices that are mainly suggested by their performance on weighted games.

Dominance properties. Consider a weighted game $[q; w_1, \dots, w_n]$. It seems intuitively obvious that if $w_i \geq w_j$ then voter i has at least as much power as voter j , because any contribution that j can make to the passage of a resolution can be equaled or bettered by i . This suggests that any reasonable power index f should be required to satisfy the following condition

$$\begin{aligned} \text{If } \mathcal{W} &\equiv [q; w_1, \dots, w_n] \text{ and} \\ w_i &\geq w_j \text{ then } f_i(\mathcal{W}) \geq f_j(\mathcal{W}). \end{aligned} \quad (6)$$

Two gradually more restrictive versions are

$$\begin{aligned} i \approx_D j, \text{ then } f_i(\mathcal{W}) &= f_j(\mathcal{W}) \quad \text{and} \\ i >_D j, \text{ then } f_i(\mathcal{W}) &> f_j(\mathcal{W}), \end{aligned} \quad (7)$$

$$\begin{aligned} i \approx_d j, \text{ then } f_i(\mathcal{W}) &= f_j(\mathcal{W}) \quad \text{and} \\ i >_d j, \text{ then } f_i(\mathcal{W}) &> f_j(\mathcal{W}), \end{aligned} \quad (8)$$

which are, respectively, named the *D-dominance* and the *d-dominance* properties. It is worth noting that dominance property (6) fails for indices like the Deegan and

Packel [4] or Holler [6] indices, while is satisfied for the other indices considered in section titled “Power Indices”. Only regular semivalues satisfy D -dominance (7) and d -dominance (8). The Johnston index satisfies D -dominance but not d -dominance. The other indices satisfy all three forms of dominance.

Donation. Consider a pair of weighted games, $\mathcal{W}$ and $\mathcal{W}'$, on the same set N whose respective representations have the same quota q and the same total weight ($\sum_{i=1}^n w_i = \sum_{i=1}^n w'_i$):

$$\begin{aligned}\mathcal{W} &\equiv [q; w_1, \dots, w_n], \\ \mathcal{W}' &\equiv [q; w'_1, w'_2, \dots, w'_n].\end{aligned}\quad (9)$$

In this configuration (9), a voter i for which $w_i > w'_i$ is called a *donor* and a voter i for which $w_i < w'_i$ is called a *recipient*. Intuitively, the idea is that $\mathcal{W}$ represents the initial distribution of weights and $\mathcal{W}'$ arises from it by redistribution, the donor “donating” some weight to the recipient; so each donor loses weight and the recipient gains some, while the other weights and the total weight are unchanged. It seems quite unacceptable that the donor gains power; if f is any power index, we say that f satisfies the *donation* property if the following holds in the configuration (9) in which there is a sole donor i and a sole recipient

$$f_i(\mathcal{W}) \geq f'_i(\mathcal{W}') \quad (10)$$

All the indices considered here fulfill the donation (10) property with the exceptions of β' and γ . To see a failure of the donation property consider $\mathcal{W} \equiv [23; 9, 8, 7, 0, 1, 1, 1, 1, 1, 1]$ and $\mathcal{W}' \equiv [23; 8, 8, 7, 1, 1, 1, 1, 1, 1, 1]$. The player with weight 9 in $\mathcal{W}$ becomes the only donor and the player with weight 0 in $\mathcal{W}$ becomes the only recipient in the configuration formed by $\mathcal{W}$ and $\mathcal{W}'$. The relative Banzhaf index β' for the voter with weight 9 in $\mathcal{W}$ is 65/199, while it is 129/392 for the voter with weight 8 in $\mathcal{W}'$. Since $65/199 < 129/392$ the donation property fails. The Johnston index γ for the voter with weight 9 in $\mathcal{W}$ is 173/529, while it is 773/2322 for the voter with weight 8 in $\mathcal{W}'$. Since

$173/529 < 773/2322$ the donation property fails.

Block. Let $\mathcal{W} \in S_N$ and let i and j be distinct voters of N . Taking $i \& j$ to be a new entity, not belonging to N , define a new simple game $\mathcal{W}[i \& j]$ as the collection of all coalitions $S \in (N \setminus \{i, j\}) \cup \{i \& j\}$ satisfying one of the following two conditions: (i) $S \subseteq N \setminus \{i, j\}$ and $S \in \mathcal{W}$; (ii) $S = T \cup \{i \& j\}$ for some T such that $T \subseteq N \setminus \{i, j\}$ and $T \cup \{i, j\} \in \mathcal{W}$.

How should the power of the bloc $i \& j$ in the new simple game compare with the powers of the two components i and j in the original simple game?

At first sight it may perhaps seem reasonable to assume that if i and j are two voters in a simple game $\mathcal{W}$ and j is not null then

$$f_{i \& j}(\mathcal{W}[i \& j]) \geq f_i(\mathcal{W}). \quad (11)$$

This is the *block* property. The bloc $i \& j$ may be regarded as a result of a voluntary merger between voters i and j . A voluntary merger will take place only if, as a result of it, both parties are at least as well off as they were before. But the bloc can equally be regarded as a result of a takeover, in which i , having annexed j 's voting rights, now trades under the new name $i \& j$. Intuitively, it seems inconceivable to us that a voter should not find it beneficial to annex the voting rights of another voter who is not null. Property (11) is a formal statement of this intuition.

All the indices considered here fulfill the block property (11) with the exceptions of β' and γ . To see a failure of these two indices, it will be enough to force a block in the example we took of the donation property; so consider again $\mathcal{W} \equiv [23; 8, 8, 7, 1, 1, 1, 1, 1, 1, 1]$, which we considered for illustrating a failure of the donation property, for β' and γ . Suppose now that a voter with weight 8 and a voter with weight 1 in $\mathcal{W}$ form a block giving rise to the game $\mathcal{W}' \equiv [23; 9, 8, 7, 1, 1, 1, 1, 1, 1, 1]$. Then, the non-null values of both β' and γ for $\mathcal{W}$ (where $\mathcal{W}$ appears in the example of the donation property) coincide with those of $\mathcal{W}'$. Note that the only difference between $\mathcal{W}$ and $\mathcal{W}'$ is that the first is an additional null voter. The failure of the donation property for the two indices provides failure of these indices for the block property.

Transfer. Dubey [7] introduces the following condition. A power index $f: \mathcal{S}_N \rightarrow [0, 1]^n$ satisfies the *transfer* condition if

$$f(\mathcal{W}) + f(\mathcal{V}) = f(\mathcal{W} \cup \mathcal{V}) + f(\mathcal{W} \cap \mathcal{V}). \quad (12)$$

Note that $\mathcal{W} \cup \mathcal{V}$ and $\mathcal{W} \cap \mathcal{V}$ are counterparts of $\mathcal{W}$ and $\mathcal{V}$ and therefore ψ respects the balance of coalitions in both sides of the equality. The indices ϕ , β (and $\mathcal{R}$) and more generally semivalues as power indices satisfy Equation (12). The transfer condition is equivalent to saying that the power index is a *valuation* on the lattice $\mathcal{S}_N$.

Bicameral meet. Suppose that a bill requires the approval of two separate chambers to be passed. Let N_1 and N_2 denote the seats in the two chambers ($N_1 \cap N_2 = \emptyset$), and let $\mathcal{W}_{N_1}$ and $\mathcal{W}_{N_2}$ denote the simple games used by each chamber. Then the bicameral meet based on these two basic simple games is defined by the simple game $\mathcal{W}_N$, with $N = N_1 \cup N_2$, where

$$\mathcal{W}_N = \{S \subseteq N : S \cap N_1 \in \mathcal{W}_{N_1} \text{ and } S \cap N_2 \in \mathcal{W}_{N_2}\}.$$

The *bicameral meet* property requires of a power index f that if i and j are two non-null voters in N_1 then

$$f_i(\mathcal{W}_{N_1})/f_j(\mathcal{W}_{N_1}) = f_i(\mathcal{W}_N)/f_j(\mathcal{W}_N).$$

The two Banzhaf, the two Coleman, and the König and Bräuninger indices satisfy the bicameral meet property, while those of Shapley–Shubik and Johnston fail. Consider the weighted games: $\mathcal{W} \equiv [3; 2, 1, 1]$ and $\mathcal{W}' \equiv [2; 2]$, and $\mathcal{W}'' \equiv [5; 2, 1, 1, 2]$, which is the bicameral meet of the first two. For the Shapley–Shubik index, the yield is $\phi(\mathcal{W}) = (4/6, 1/6, 1/6)$ and $\phi(\mathcal{W}') = (10/24, 2/24, 2/24, 10/24)$. Thus, $\phi_1(\mathcal{W})/\phi_2(\mathcal{W}) = 4$ and $\phi_1(\mathcal{W}')/\phi_2(\mathcal{W}') = 5$. Hence, there is a failure of the bicameral meet property. Similarly, for the Johnston index: $\gamma_1(\mathcal{W})/\gamma_2(\mathcal{W}) = 4$, but $\gamma_1(\mathcal{W}')/\gamma_2(\mathcal{W}') = 3.5$. This example also illustrates a failure of the Rae index. Indeed, $\mathcal{R}_1(\mathcal{W})/\mathcal{R}_2(\mathcal{W}) = 7/5$, but $\mathcal{R}_1(\mathcal{W}')/\mathcal{R}_2(\mathcal{W}') = 11/9$.

AXIOMATIZATIONS VERSUS PROBABILISTIC APPROACH

Some power indices are founded on axiomatizations. An axiomatization for a power index consists of a list of *independent* properties that characterize it as the *unique* power index that satisfies the proposed properties. As more compelling is the list of properties characterizing a given power index, more valuable is the index as a measure of power.

Other power indices are founded on probabilistic approaches. This means that a power index can be interpreted as the probability of a certain event belonging to a sample space related to the game.

The purpose here is to mention some of the main axiomatizations for some of the power indices introduced in section “Power Indices” and to provide a brief probabilistic approach for other indices. A good review on the precise meaning and implications of the power notion in politics can be found in Laruelle and Valenciano [40], wherein many other interesting references on the topic by the same authors are given.

Some Axiomatizations of Power Indices

For lack of space we do not provide any axiomatization, but just mention some of them here, and refer the interested reader to the references. Of course, there exist many other axiomatic characterizations than those cited here.

The Shapley–Shubik index was the first to be characterized. Dubey [7] proves that it is the only map $f: \mathcal{S}_N \rightarrow \mathbb{R}^n$ that satisfies efficiency, anonymity, null player and transfer properties.

Dubey and Shapley [8] provide an adaptation of the characterization by Dubey for the Shapley–Shubik index in order to axiomatize the absolute Banzhaf index by replacing “efficiency” by the following somewhat ad hoc condition (known as *total power property*):

$$\sum_{i=1}^n f_i(\mathcal{W}) = \frac{|\mathcal{C}_i|}{2^{n-1}} \quad (13)$$

to obtain the following result. The absolute Banzhaf index is the only $f: \mathcal{S}_N \rightarrow \mathbb{R}^n$ that

satisfies condition (13), anonymity, null player, and transfer properties.

There is a rich literature on obtaining different axiomatizations of the Shapley–Shubik and the Banzhaf indices; among others, consider Owen [41], Lehrer [42], Roth [43], Straffin [44,45], Haller [46], Feltkamp [47], van der Brink and van der Laan [48], Nowak and Radzik [49], Albizuri [50], and Barua *et al.* [51].

The power indices ϕ and β share three axioms (anonymity, null player, and transfer). These three conditions are also satisfied by all semivalues. The axiomatic characterization given in Weber [52] for this class of games has been adapted to simple games in Carreras *et al.* [28]. Semivalues are basically the family of “values” that result by dropping “efficiency” in Shapley’s characterizing system of the Shapley value. Hence, semivalues for simple games can be seen as “generalized power indices”, and they are uniquely characterized by anonymity, positivity, dummy player, and transfer property.

The Johnston index [5] shares efficiency, null, and anonymity axioms with indices ϕ and β' . Lorenzo-Freire *et al.* [53] characterize it by using a property called *critical mergeability*.

Other relevant relative power indices considered in the literature are the Holler [6] and the Deegan and Packel [4] indices. However, these two indices do not fulfill the domination property in either of its versions: either Equation (6) or Equation (7) or Equation (8). It seems that the failure of domination is a handicap for them. However, they are interesting alternative efficient power indices to those considered here.

In summary, different axiomatizations for an index give different views of it, so that one may find one axiomatization more compelling than the other. In any case, several independent axiomatizations of an index add to its degrees of support as a good measure to evaluate power.

The property of efficiency is a natural and rational condition if the goal is to measure power as an expected payoff (i.e., we want to divide a unit of utility among voters). If we regard power as influence, whatever this might mean, then this condition is not

compelling. None of the measures that we consider in the next section are efficient. There we try to explain power as convenient probability of events rather than with an axiomatic approach.

Probabilistic Approach

Assume that a probability distribution over vote configurations p (exogenous information) enters as a second input besides the simple game $\mathcal{W}$. Of course, $0 \leq p(S) \leq 1$ for all $S \subseteq N$, and $\sum_{S \subseteq N} p(S) = 1$. The probability of i voting “yes” is denoted by $\delta_i(p)$, that is

$$\delta_i(p) = \text{Prob}(i \text{ votes “yes”}) = \sum_{S: i \in S} p(S).$$

In a voting situation thus described by the pair $(\mathcal{W}, p)$, the ease of passing proposals or probability of acceptance is given by

$$\alpha(\mathcal{W}, p) = \text{Prob}(\text{acceptance}) = \sum_{S: S \in \mathcal{W}} p(S).$$

If the proposal is accepted (rejected), those voters who have voted in favor (against) are satisfied with the result, while others are not. We also say that a successful voter has been *decisive* in a vote if his/her vote was crucial, that is, if he/she had changed his/her vote the outcome would have been different.

$$\begin{aligned} \Omega_i(\mathcal{W}, p) &= \text{Prob}(i \text{ is successful}) \\ &= \sum_{S: i \in S \in \mathcal{W}} p(S) + \sum_{S: i \notin S \notin \mathcal{W}} p(S). \end{aligned}$$

$$\begin{aligned} \Phi_i(\mathcal{W}, p) &= \text{Prob}(i \text{ is decisive}) = \sum_{\substack{S: i \in S \in \mathcal{W} \\ S \setminus \{i\} \notin \mathcal{W}}} p(S) \\ &\quad + \sum_{\substack{S: i \notin S \notin \mathcal{W} \\ S \cup \{i\} \in \mathcal{W}}} p(S). \end{aligned}$$

According to Laruelle and Valenciano [40], Barry’s [54] equation for “success” = “decisiveness” + “lucky” remains valid in a much more precise and general version:

$$\Omega_i(\mathcal{W}, p) = \Phi_i(\mathcal{W}, p) + \Lambda_i(\mathcal{W}, p),$$

where $\Lambda_i(\mathcal{W}, p)$ denotes voter i 's "luck" or probability of being "lucky", that is,

$$\Lambda_i(\mathcal{W}, p) = \sum_{\substack{S: i \in S \\ S \setminus \{i\} \in \mathcal{W}}} p(S) + \sum_{\substack{S: i \notin S \\ S \cup \{i\} \notin \mathcal{W}}} p(S)$$

Consider the following notation:

$$\begin{aligned} \Omega_i^+(\mathcal{W}, p) &= \text{Prob}(i \text{ is successful \& } i \text{ votes "yes"}) \\ &= \sum_{S: i \in S \in \mathcal{W}} p(S), \\ \Omega_i^-(\mathcal{W}, p) &= \text{Prob}(i \text{ is successful \& } i \text{ votes "no"}) \\ &= \sum_{S: i \notin S \in \mathcal{W}} p(S), \end{aligned}$$

so that voter i 's probability of success is:

$$\Omega_i(\mathcal{W}, p) = \Omega_i^+(\mathcal{W}, p) + \Omega_i^-(\mathcal{W}, p),$$

and voter i 's probability decisiveness is:

$$\Phi_i(\mathcal{W}, p) = \Phi_i^+(\mathcal{W}, p) + \Phi_i^-(\mathcal{W}, p).$$

Conditional probabilities of success and decisiveness under different conditions are also meaningful. The following questions arise naturally in this context:

Question 1: What is voter i 's conditional probability of success (decisiveness), given that voter i votes in favor of (against) the proposal?

Question 2: What is voter i 's conditional probability of success (decisiveness), given that the proposal is accepted (rejected)?

This makes for eight possible conditional probabilities. A little notation is necessary for variations of the measures: (Ω_i and Φ_i). The superindex " $i+$ " (" $i-$ ") expresses the condition given that i votes "yes" ("no"). So the answers to the first question are given by Ω_i^{i+} , Φ_i^{i+} , Ω_i^{i-} , and Φ_i^{i-} , respectively. The superindex "Acc"("Rej") expresses the condition "given that the proposal is accepted (rejected)". Thus, the answers to the second

question are given by Φ_i^{Acc} , Ω_i^{Acc} , Φ_i^{Rej} and Ω_i^{Rej} , respectively. As an illustration, we formulate two of them explicitly. Voter i 's conditional probability of being decisive given that voter i votes in favor of the proposal, is given by

$$\begin{aligned} \Phi_i^{i+}(\mathcal{W}, p) &= \text{Prob}(i \text{ is decisive} \mid i \text{ votes "yes"}) \\ &= \frac{1}{\delta_i(p)} \sum_{S \in \mathcal{C}_i} p(S). \end{aligned}$$

Voter i 's conditional probability of success, given that the proposal is accepted, is given by

$$\begin{aligned} \Omega_i^{\text{Acc}}(\mathcal{W}, p) &= \text{Prob}(i \text{ is successful} \mid \text{acceptance}) \\ &= \frac{1}{\alpha(\mathcal{W}, p)} \sum_{S: i \in S \in \mathcal{W}} p(S). \end{aligned}$$

The ten different unconditional and conditional probabilities of success and of decisiveness considered so far are summarized in Table 2.

Measurement is based on a probability distribution over vote configurations, but the relevant information for estimating this probability is usually not known. One simplification is to assume that all vote configurations are equally probable *a priori*: $p^*(S) = \frac{1}{2^n}$ for any configuration $S \subseteq N$. For this special unbiased distribution of probability some special relations hold:

$$\Omega_i(\mathcal{W}, p^*) = 0.5 + 0.5\Phi_i(\mathcal{W}, p^*). \quad (14)$$

Unconditional decisiveness and conditional decisiveness, given a positive vote or a negative vote, are indistinguishable:

$$\Phi_i(\mathcal{W}, p^*) = \Phi_i^{i+}(\mathcal{W}, p^*) = \Phi_i^{i-}(\mathcal{W}, p^*).$$

But it is worth remarking that these relationships do not hold in general for $p \neq p^*$. Some "power indices" in the literature can be seen as a particularization of some of the measures just introduced. Indeed,

$$\mathcal{R}_i(\mathcal{W}) = \Omega_i(\mathcal{W}, p^*), \quad \beta(\mathcal{W}) = \Phi_i(\mathcal{W}, p^*),$$

and from Equation 14 it follows that

Table 2. Ten Different Unconditional and Conditional Probabilities of Success and Decisiveness

Condition:	None	i votes “yes”	i votes “no”	Acceptance	Rejection
Success	Ω_i	Ω_i^{i+}	Ω_i^{i-}	Ω_i^{Acc}	Ω_i^{Rej}
Decisiveness	Φ_i	Φ_i^{i+}	Φ_i^{i-}	Φ_i^{Acc}	Φ_i^{Rej}

Table 3. Relationships between “Power Indices” and the Current Probabilistic Model

Condition	None	i votes “yes”	i votes “no”	Acceptance	Rejection
Success	$R_i(\mathcal{W})$	Ω_i^{i+}	Ω_i^{i-}	$KB_i(\mathcal{W})$	Ω_i^{Rej}
Decisiveness	$\beta_i(\mathcal{W})$	$\beta_i(\mathcal{W})$	$\beta_i(\mathcal{W})$	$C_i^P(\mathcal{W})$	$C_i^I(\mathcal{W})$

$$\mathcal{R}_i(\mathcal{W}) = 0.5 + 0.5\beta_i(\mathcal{W}).$$

However, as noticed in Laruelle and Valenciano [40], Equation (14) is not necessarily true if $p \neq p^*$. Therefore, success and decisiveness become unrelated concepts for particular distributions of probability different from p^* . Concerning the different Coleman indices and the König–Bräuninger’s index we have

$$\begin{aligned} \mathcal{A}(\mathcal{W}) &= \alpha(\mathcal{W}, p^*), \quad C_i^P(\mathcal{W}) = \Phi_i^{\text{Acc}}(\mathcal{W}, p^*), \\ C_i^I(\mathcal{W}) &= \Phi_i^{\text{Rej}}(\mathcal{W}, p^*), \quad KB_i(\mathcal{W}) = \Omega_i^{\text{Acc}}(\mathcal{W}, p^*). \end{aligned}$$

Table 3 summarizes these relationships: The Shapley–Shubik index and more generally semivalues are also measures of decisiveness if we consider particular choices of p .

DISCUSSION

In this section, we have two purposes. The first purpose is to summarize the degree of accomplishment of some properties considered in the section titled “Properties of Power

Indices” and to discuss the frequency of some failures of these properties. The second purpose is to analyze the ordinal equivalence of power indices, that is to say, to study whether they rank voters equally.

The occurrence of paradoxes and its frequency. Table 4 summarizes the failures of some properties for some power indices. We use the following notation: decisiveness = $\Phi(\mathcal{W}, p)$; success = $\Omega(\mathcal{W}, p)$; ND: means “not if $p(S)$ dependent on $|S|$ ”; NI: means “not if independence”; NI’: means “not if independence and $\delta_i = \delta_j$ ” for all $i, j \in N$.

Freixas and Molinero have done an experimental analysis that gives us some kind of evidence on whether some of these properties’ (dominance, donation, and bloc) failures are simply occasional facts and so can be regarded as occasional phenomena or, on the contrary, they appear frequently; hence, the term “paradoxical” for the failure of these properties seems to be rather inappropriate. We have found that

For β' and γ and for all games between five and eight voters, the percentage of times that the donation paradox arises is close

Table 4. Testing Several Indices

↓ Property/measure →	Shapley–Shubik (ϕ)	Banzhaf (β')	Johnston (γ)	Decisiveness	Success
D-dominance	Never	Never	Never	ND	ND
d-dominance	Never	Never	May occur	ND	ND
Donation	Never	May occur	May occur	Never	Never
Bloc	Never	May occur	May occur	Never	Never
Bicameral	May occur	Never	May occur	NI	NI

to 3%, whereas for the bloc paradox the percentage is close to 1%.

For β' and γ and for games with a large number of voters both the donation and the bloc paradoxes tend to appear very frequently.

In summary, we have checked that these failures are not isolated facts due to some particularities of the voting system at hand. Given an arbitrary game with a large number of voters, one should expect that these failures will be produced and therefore we should regard these failures as normal rather than as paradoxes. It seems clear that efficiency (Eq. (2)) is a very strong demand for a power index that may cause these failures, and, in fact, it is an unnecessary property when we regard power as an influence. Naturally, decisiveness (and success) behaves better than efficient power indices.

Power indices and their rankings. Of course, different power indices provide different measures and lead to different rankings for the power of voters; in this context a question naturally arises. How different are the separability relations induced by the most well-known power indices?

Let $\mathcal{W} \in \mathcal{S}_N$. Two power indices f and f' are *ordinally equivalent* if their separability relations coincide, i.e., for any $i, j \in N$ the following equivalences hold:

$$\begin{aligned} f_i > f_j &\iff f'_i > f'_j, \\ f_i = f_j &\iff f'_i = f'_j. \end{aligned}$$

Two ordinally equivalent power indices rank players in N in exactly the same way. We say that they induce the same *hierarchy*.

As it was proved in Carreras and Freixas [21], the rankings of voters given by any regular semivalue *coincide* within the class of weakly complete games, which includes almost *all* binary real-life voting systems. In particular, the Shapley–Shubik index and the Penrose–Banzhaf–Coleman indices coincide in weakly complete games. Hence, if we are interested in rankings (qualitative approach) rather than in exact measures (quantitative approach), we regard both the Shapley–Shubik and the Penrose–Banzhaf–Coleman indices as

being *qualitatively equivalent* within weakly complete games. Note, for instance, that the real-life simple games in Examples 1, 2, and 3 are, respectively, weighted, complete, and weakly complete. Hence, for these three classes of games, these indices are ordinally equivalent. Furthermore, the ordinal equivalence of the Shapley and Banzhaf values (or their restrictions for simple games) is still true for a larger class of games called *semicoherent* ([55]). In other words, we have not found any real-life simple game where the rankings for Shapley–Shubik and Banzhaf–Coleman indices do not coincide. So whatever “power” means we find that the most recognized power indices rank voters equally for simple games used in real life.

Another interesting question is to find out which rankings can be induced for regular semivalues. In Freixas and Pons [56], it is proved that all rankings are allowable for these indices inside weakly complete games with the sole exception of the four rankings: $> >$, $> > >$, $= > >$, $= > > >$, which are never available in a simple game.

We finally point out that the Johnston index ranks voters in the same manner as the Shapley–Shubik and Penrose–Banzhaf–Coleman power indices within complete games, and hence it is qualitatively equivalent to them within this class. The Johnston index has also good properties, not discussed here, concerning egalitarianism. We think that the Johnston index (or its raw version, not efficient) is an interesting power index under the shadow of the Shapley–Shubik and Penrose–Banzhaf–Coleman power indices.

MISCELLANEOUS

The section contains some explanations and references about the modification of power indices in the first five sections. In the subsequent sections other issues concerning power indices are treated.

When trying to apply power indices to actual situations, one frequently finds that the personality of the involved agents exerts a great influence in the outcomes. In other words, exogenous relationships must be considered if we do not want to keep the

analysis just in an “*in vitro*” stage but wish to confer to it an “*in vivo*” category, in order to better fit the reality. Curiously, this influence of the agents’ personality is clearly stronger in simple games describing political contexts than in cooperative games describing economic contexts: ideological constraints seem to be much more important in politics than in economics.

Different options have been proposed and used to incorporate additional information into the analysis of games, which is not stored in the characteristic function. Most of them are defined for cooperative games in general, but we emphasize here on its relevance for simple games only. And, of course, the effect of this additional information on the use of power indices is a main point to be considered. Several options are mentioned here: coalition structures, affinities, incompatibilities, cooperation indices, semivalues, partnerships, and persuasion and bribes among others. In some of these cases, the procedure is quite common: the exogenous information gives rise to a restricted game and the original power index is replaced with a modified power index, which is nothing but the original index applied to the restricted game.

Coalition Structures

The notion of coalition structure in a cooperative game was first introduced in Aumann and Drèze [57]. A coalition structure is a partition of the set of players, that is, a family of pairwise disjoint unions that cover the set. The authors argued that the effect of the coalition structure was that of confining the players to play a subgame in each one of the unions. They then defined the modified Shapley value of a game with a coalition structure by assigning to each player his Shapley value in the subgame he is playing within the union he belongs to. No application of this idea to simple games has been developed: the reason is that it very often leads to triviality because most of the unions are, usually, losing coalitions.

However, later on, a different approach that opened a prolific research line, especially on its application to simple games, was given in Owen [58] (also in Ref. 59). The coalition structure induces a quotient game

played by the unions. Then, each union gets the payoff given by the Shapley value in this game. Next, by considering pseudo-quotient games where each subset of a union replaces it in bargaining with the remaining unions and gets a theoretical payoff, we obtain a subgame within each union. By applying again the Shapley value, now to this subgame, the members of that union share its payoff among themselves. This gives rise to the now so-called Owen value. The method applies perfectly to simple games since the subclass they form is closed under the passing to quotient games and subgames. Then a “coalitional Shapley–Shubik index of power” is obtained.

Several applications of this index to real-world political setups have been published. In general, one is interested in discussing which coalition would form, and hence all possible coalition structures are tested using the Owen value. The selected ones, if any, are those that show stability in the very sense of the Nash equilibrium concept: those where no agent (party, nation. . .) possesses incentives to break the coalition structure unilaterally. The case of the Catalonia Parliament [60] illustrates this procedure.

Affinities, Incompatibilities, and Cooperation Indices

In Carreras [61], two main qualitative methods to modify simple games, and hence power indices, were suggested in order to get a better description of real-world situations. The selection of one method or the other will depend in each case on the view that the observer holds on the political relationships, ideological or strategic, among the agents. The former method is based on the affinity idea. A graph describing links between the agents is provided: it is interpreted that only agents connected by a path of links will be able to cooperate among themselves. This implies a modification of the game where only connected minimal winning coalitions are allowed, and one gets a restricted game in this way. Accordingly, the modified Shapley–Shubik power index is defined as the Shapley–Shubik index of the restricted game. This idea, closely related to a similar

work on general cooperative games (Myerson [62]), was applied in Carreras and Owen [60].

The latter method is based on the incompatibility idea and becomes more radical than the former in the sense that, in general, it gives rise to a more drastic modification of the original game. In this case, an irreflexive and symmetrical relation between the agents is assumed to be given, and only the minimal winning coalitions without incompatible members are considered admissible. Again, we get a restricted game and modify the Shapley–Shubik index as discussed earlier.

In Carreras [61], the possibility to combine both methods and simultaneously introduce an affinity graph and an incompatibility relation is also mentioned, without developing, however, the possibilities of this third option.

Nevertheless, a wide generalization, of a quantitative type and hence an increased ability to accurately describe the relationships among the agents, was provided in Amer and Carreras [63] in the general setup of cooperative games by introducing the notion of cooperation index, which encompasses the setup of Myerson [62]. A cooperation index in a game assigns a cooperation degree between 0 and 1 to each coalition. Again, this gives rise to a modified game, and the modified Shapley value is defined as the Shapley value of the modified game. It is worthy of mention that if this technique is applied to a simple game, then a non-simple game may arise as a modified game, so that the modified Shapley–Shubik index is, really, the Shapley value of the modified game.

In Amer and Carreras [64], these ideas are extended to games with a coalition structure and, therefore, yield a modification of the Owen value or power index by means of a cooperation index. Applications to real-world examples can be seen in Amer and Carreras [65,66]. In Amer and Carreras [67] the cooperation index theory is extended to weighted Shapley values (introduced in Kalai and Samet [68]).

Semivalues

Each semivalue is defined by a series of parameters that satisfy certain conditions, and hence the use of semivalues represents an alternative for introducing exogenous

information in the evaluation of games: one need not modify the game but only the viewpoint from which it is analyzed. For instance, the binomial semivalues constitute a special mono-parametric subfamily of semivalues.

Semivalues have been studied as power indices in Carreras *et al.* [28] and have been applied to real-world situations ([27,29–31]). They have been extended to games with a coalition structure in Albizuri and Zarzuelo [69] in the homogeneous case, where a unique semivalue acts in the quotient game and within unions, and in the heterogeneous case, where two arbitrary semivalues are combined.

Partnerships

The notion of partnership was first introduced in Kalai Samet [68] in the setup of weighted Shapley values. In Carreras [70], a detailed study of this concept was carried out, including the possibility to introduce one or more (disjoint) partnerships in a game and obtain a modified game to which any power index can be adapted using the standard method mentioned before. The procedure behaves nicely within the subclass of simple games.

Persuasion and Bribes

Suppose that certain committee is going to decide, using some fixed voting rules, either to accept or to reject a proposal that affects your interests. From your perception about each voter's position, you can make an *a priori* estimation of the probability of the proposal being accepted. Wishing to increase this probability of acceptance before the votes are cast, assume further that you are able to convince (at least) one voter to improve his/her perception in favor of the proposal. The question is: which voters should be persuaded/bribed in order to get the highest possible increase in the probability of acceptance? In other words, which are the optimal persuadable voters? To answer this question, a measure of “circumstantial power” is considered in Freixas and Pons [71]. Three preorderings in the set of voters, based on the voting rules, are defined and they are used

for finding optimal persuadable voters, even in the case that only a qualitative ranking of each voter's inclination for the proposal has been made.

The presidential Election Game and the Power in the European Union

The method of choosing a president for the United States presents a very interesting problem, which is treated in detail in Owen [59] (Chapter XII Section 4). The voters within each state elect "great electors," or members of the Electoral College, who in turn vote for the president. It is generally assumed that all of a given state's electors will vote for whichever candidate is preferred by a majority of the state's voters; thus, a very thin majority in a large state will more than annul large majorities in several small states. Owen analyzes the game from the point of view of both the Shapley–Shubik and Penrose–Banzhaf–Coleman power indices. He did the computations with generating functions and found that for the second index the calculi are considerably simpler since they are based on the square-roots of the census figures. He illustrates that with both indices a voter in California has approximately three times as much power as a voter in the District of Columbia. An observation which is, doubtless, of interest on terms of fairness of the voting system used.

Several binary voting systems have been used in the European Council between 1958 (treaty of Rome) and 2005, the number of members in the Council increasing from 7 to 25. The most complex of these systems are the two established in the treaty of Nice, both of which are complete simple games of dimension 3 ([72]). Several authors have analyzed power in these systems, among them being Bilbao *et al.* [73].

Complexity and Computing Power Indices

Complexity is an interesting issue for all lines of research, and particularly for simple games and power indices. Some easy problems for a small number of voters can become, for a large number of voters, very long processes even for a computer. Hence, it is interesting to determine the complexity of some of

Table 5. Complexity of Changing the Representation Form of a Simple Game

Input → Output ↓	$(N, \mathcal{W})$	$(N, \mathcal{L})$	$(N, \mathcal{W}^m)$	$(N, \mathcal{L}^M)$
$(N, \mathcal{W})$	—	EXP	EXP	EXP
$(N, \mathcal{L})$	EXP	—	EXP	EXP
$(N, \mathcal{W}^m)$	PT	PT	—	EXP
$(N, \mathcal{L}^M)$	PT	PT	EXP	—

these steps. As an example, suppose that we consider four natural ways of representing a simple game: winning, losing, minimal winning, and maximal losing coalitions. The computational cost of passing from one to the other is summarized in Table 5, which is obtained from Freixas *et al.* [74] and where other related problems concerning complexity of simple games are discussed. We use the following notations in the table: $\mathcal{L}$ for the set of losing coalitions of a simple game; PT for polynomial time; EXP for exponential time.

In general, the computation of power indices is NP-hard. And, therefore, a great number of operations are needed even for a computer. Generating functions, see Brams and Affuso [36], are a very useful tool to compute the Banzhaf score and the Shapley–Shubik index (Cantor in 1962 already computed the Shapley–Shubik index in a similar way). Several research groups have developed web pages with their own software to compute power indices; two examples are <http://www.warwick.ac.uk/~ecaae/> and <http://powerslave.val.utu.fi/>.

Dimension and Codimension of Simple Games

Weighted simple games are probably the most important subclass of simple games. It is well known that every simple game can be represented as an intersection of weighted games. It is, nevertheless, of interest to ask how efficiently this can be done for a given simple game. The question of efficiency leads to the definition of *dimension*, [15,16,24]. A simple game is said to be of *dimension* k if and only if it can be represented as the intersection of exactly k weighted games, but not as the intersection of $k - 1$ weighted games. In this sense, we can regard weighted simple games as a generating system of the class

of simple games with the *intersection* as the basic operation.

In Freixas and Marciniak [75], several extensions of the notion of dimension of a simple game are given. The existence of a minimum subclass of weighted games is proved with the property that every simple game can be expressed as its intersection. Some further generalizations lead to the new concept of codimension, which is obtained by considering the union instead of the intersection as the basic operation. Dimension and codimension are significant concepts since most of the real-life binary voting systems are of small dimension or codimension (e.g., the Council of Ministers of the European Union is a system of dimension 3, see details in Ref. 72).

Minimal Integer Representations of Weighted Games

In the section titled “Power Indices,” we have omitted many other power indices regarded as payoffs and which mainly come from cooperative game theory as, for example, the nucleolus, which shares the following properties with ϕ , β' and γ : Equations (2)–(5). There are also set solution concepts like the core or the kernel that may be empty for simple games, but a modification of them, for example, the least core, always gives a nonempty payoff for voters. Some computations have been done on these notions, specially for weighted games or their subclasses. Finding minimal integer representations of weighted games is a twofold interesting problem. On one side, they provide the simplest integer equivalent representation to any arbitrary weighted game, and on the other side these minimal representations are usually related with efficient power indices, see, for example, Peleg [76,77], who proves that the nucleolus of a constant-sum homogeneous weighted game coincides with the unique minimal integer representations of the game. The least core is, in many cases, related to the convex hull of the minimal integer representations of a weighted game.

Isbell in 1959 was the first to find a weighted game *without* a *minimum* integer representation, in which the affected players do not play a symmetric role (i.e., they are not equi-desirable) in the game. His example

has 12 players and is a *weighted decisive* game, that is, a weighted game for which a coalition wins *iff* its complement loses. Quite recently Freixas and Molinero proved in Ref. 78 that the minimum number of voters required to get an example of this type is 9. One of these examples with two minimal representations is formed by the representations: [54; 14, 11, 7, 5, 5, 5, 3, 2, 2] and [54; 14, 12, 6, 5, 5, 5, 3, 2, 2]. Note that the second and third players in such representations are not symmetric (not equi-desirable).

For $n = 8$ voters, there are 154 non-isomorphic weighted games without a unique minimal weight integer realization. For instance, the representations [12; 7, 6, 6, 4, 4, 4, 3, 2] and [12; 7, 6, 6, 4, 4, 4, 2, 3] are equivalent and componentwise minimal. These two equivalent realizations for the game are also equivalent to [14; 8, 7, 7, 5, 5, 5, 3, 3], which preserves types (equi-desirability classes) of players.

Games with Abstention and with Several Levels of Approval

A promising future line of research concerns voting systems with several levels of approval, including abstention or absence. Before dealing with these systems, it is required to provide the basic structure to analyze them. Extensions of simple games, weighted games, desirability, anonymity are important. We refer the reader to Freixas [79] for an appropriate framework dealing with the so-called (j, k) simple games (generalizations of both simple and cooperative games) and weighted (j, k) simple games, to Tchantcho *et al.* [80] for feasible extensions of the desirability relation, to Freixas and Zwicker [81] for anonymous (j, k) simple games, and to Freixas [79,82] for extensions of the Shapley and Banzhaf values.

Acknowledgments

The author wishes to thank Montserrat Pons and Nadezhda V. Smirnova for reading an earlier version of this work and pointing out useful comments that have allowed me to improve the presentation of this brief survey on power indices. I apologize for the omission of many interesting works due to the lack of

space or knowledge. Research was partially supported by Grant MTM 2009–08037 from the Spanish Science and Innovation Ministry, and SGR 2009–1029 of “Generalitat de Catalunya”.

REFERENCES

1. Morriss P. Power: a philosophical analysis. 2nd ed. New York: Manchester University Press; 2002;
2. Shapley LS, Shubik M. A method for evaluating the distribution of power in a committee system. *Am Pol Sci Rev* 1954;48:787–792.
3. Shapley LS. A value for n -person games. In: Tucker AW, Kuhn HW, editors. Contributions to the theory of games II. Princeton (NJ): Princeton University Press; 1953. pp. 307–317.
4. Deegan J, Packel EW. A new index of power for simple n -person games. *Int J Game Theory* 1978;7:113–123.
5. Johnston RJ. On the measurement of power: some reactions to Laver. *Environ Plann A* 1978;10:907–914.
6. Holler MJ. Forming coalitions and measuring voting power. *Pol Stud* 1982;30:262–271.
7. Dubey P. On the uniqueness of the Shapley value. *Am Pol Sci Rev* 1975;4:131–139.
8. Dubey P, Shapley LS. Mathematical properties of the Banzhaf power index. *Math Oper Res* 1979;4:99–131.
9. Penrose LS. The elementary statistics of majority voting. *J R Stat Soc* 1946;109:53–57.
10. Banzhaf JF. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Rev* 1965;19:317–343.
11. Coleman JS. Control of collectivities and the power of a collectivity to act. In: Lieberman B, editor. Social choice. New York: Gordon and Breach; 1971. pp. 269–300.
12. Felsenthal DS, Machover M. The measurement of voting power: theory and practice, problems and paradoxes. Cheltenham: Edward Elgar; 1998.
13. Von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton (NJ): Princeton University Press; 1944.
14. Isbell JR. A class of majority games. *Q J Math Oxford Ser* 1956;7:183–187.
15. Taylor AD, Zwicker WS. Weighted voting, multicameral representation, and power. *Games Econ Behav* 1993;5:170–181.
16. Taylor AD, Zwicker WS. Simple games: desirability relations, trading, and pseudoweightings. New Jersey: Princeton University Press; 1999.
17. Carreras F, Freixas J. Complete simple games. *Math Soc Sci* 1996;32:139–155.
18. Tautenhahn N. Enumeration of simple games with application in the distribution of voting weights [PhD dissertation] in German (also submitted in European Journal of Operational Research in 2008 with Sascha Kurz). Germany: Department of Mathematics, Physics, and Informatics, University of Bayreuth; 2008.
19. Peleg B. Voting by count and account. In: Selten R, editor. Rational interaction. Springer; 1992. pp. 45–52.
20. Carreras F, Freixas J. A power analysis of linear games with consensus. *Math Soc Sci* 2004;48:207–221.
21. Carreras F, Freixas J. On ordinal equivalence of power measures given by regular semivalues. *Math Soc Sci* 2008;55:221–234.
22. Kilgour DM. A formal analysis of the amending formula of Canada's Constitution Act. *Can J Polit Sci* 1983;16:771–777.
23. Kilgour DM, Levesque TJ. The Canadian constitutional amending formula: Bargaining in the past and in the future. *Public Choice* 1984;44:457–480.
24. Taylor AD. Mathematics and politics. New York: Springer; 1995.
25. Weber RJ. Subjectivity in the valuation of games. In: Moeschlin O, Pallaschke D, editors. Game theory and related topics. Amsterdam: North Holland Publishing Co.; 1979. pp. 129–136.
26. Dubey P, Neyman P, Weber RJ. Value theory without efficiency. *Math Oper Res* 1981;6:122–128.
27. Carreras F, Freixas J. Some theoretical reasons for using regular semivalues. In: De Swart H, editor. Logic, game theory and social choice (Proceedings of the International Conference). LGS Tilburg, the Netherlands: Tilburg University Press; 1999. pp. 140–154.
28. Carreras F, Freixas J, Puente MA. Semivalues as power indices. *Eur J Oper Res* 2003;149:676–687.
29. Carreras F, Freixas J. A note on regular semivalues. *Int Game Theory Rev* 2000;2:345–352.
30. Carreras F, Freixas J. Semivalue versatility and applications. *Ann Oper Res* 2002;109:341–356.

31. Carreras F, Freixas J. On power distribution in weighted voting. *Soc Choice Welfare* 2005;24:269–282.
32. Carreras F, Llongueras MD, Puente MA. Partnership formation and binomial semivalues. *Eur J Oper Res* 2009;192:487–499.
33. Rae D. Decision rules and individual values in Constitutional choice. *Am Pol Sci Rev* 1969;63:40–56.
34. Coleman JS. *Combinatorics: individual interests and collective action: selected essays*. Cambridge: Cambridge University Press; 1986.
35. Carreras F. A decisiveness index for simple games. *Eur J Oper Res* 2005;163:370–387.
36. Brams SJ, Affuso PJ. Power and size: a new paradox. *Theory Decision* 1976;7:29–56.
37. König T, Bräuninger T. The inclusiveness of European decision rules. *J Theor Polit* 1998;10:125–142.
38. Freixas J, Gambarelli G. Common internal properties among power indices. *Control Cybern* 1997;26:591–604.
39. Felsenthal DS, Machover M. Postulates and paradoxes of relative voting power – A critical reappraisal. *Theory Decis* 1995;38:195–229.
40. Laruelle A, Valenciano F. *Voting and collective decision-making*. Cambridge: Cambridge University Press; 2008.
41. Owen G. Characterization of the Banzhaf–Coleman index. *SIAM J Appl Math* 1978;35:315–327.
42. Lehrer E. An axiomatization of the Banzhaf value. *Int J Game Theory* 1988;17:89–99.
43. Roth AE. Introduction to the Shapley value. In: Roth AE, editor. *The Shapley value: Essays in Honor of L. S. Shapley*. Cambridge: Cambridge University Press; 1988. pp. 1–27.
44. Straffin PD. The Shapley–Shubik and Banzhaf power indices. In: Roth AE, editor. *The Shapley value: Essays in Honor of Lloyd S. Shapley*. Cambridge: Cambridge University Press; 1988. pp. 71–81.
45. Straffin PD. Power and stability in politics. In: Aumann RJ, Hart S, editors. *Volume 2, Handbook of game theory*. Amsterdam: Elsevier Sciences; 1994. pp. 1127–1151.
46. Haller H. Collusion properties of values. *Int J Game Theory* 1994;23:261–281.
47. Feltkamp V. Alternative axiomatic characterizations of the Shapley and Banzhaf values. *Int J Game Theory* 1995;24:179–186.
48. van den Brink R, van der Laan G. Axiomatizations of the normalized Banzhaf value and the Shapley value. *Soc Choice Welfare* 1998;15:567–582.
49. Nowak AS, Radzik T. An alternative characterization of the weighted Banzhaf value. *Int J Game Theory* 2000;29:127–132.
50. Albizuri MJ. An axiomatization of the modified Banzhaf–Coleman index. *Int J Game Theory* 2001;30:167–176.
51. Barua R, Chakravarty SR, Roy S, *et al.* A characterization and some properties of the Banzhaf–Coleman–Dubey–Shapley sensitivity index. *Soc Choice Welfare* 2004;15:567–582.
52. Weber JR. Probabilistic values for games. In: Roth AE, editor. *The Shapley value*. Cambridge: Cambridge University Press; 1988. pp. 101–119.
53. Lorenzo-Freire S, Alonso-Meijide JM, Casas-Méndez B, *et al.* Characterizations of the Deegan–Packel and Johnston power indices. *Eur J Oper Res* 2007;177:431–444.
54. Barry B. Is it better to be powerful or lucky?, part I and part II. *Pol Stud* 1980;28:183–194,338–352.
55. Freixas J. On ordinal equivalence of the Shapley and Banzhaf values. *Int J Game Theory* 2010. DOI: 10.1007/s00182-009-0179-0.
56. Freixas J, Pons M. Hierarchies achievable in simple games. *Theory Decis* 2010;68:393–404.
57. Aumann RJ, Drèze J. Cooperative games with coalition structures. *Int J Game Theory* 1974;3:217–237.
58. Owen G. Values of games with a priori unions. In: Henn R, Moeschlin O, editors. *Mathematical economics and game theory*. Berlin: Springer; 1977. pp. 78–88.
59. Owen G. *Game theory*. 3rd ed. San Diego (CA): Academic Press; 1995.
60. Carreras F, Owen G. An analysis of the Catalanian Parliament, 1980–1984. *Math Soc Sci* 1988;15:87–92.
61. Carreras F. Restriction of simple games. *Math Soc Sci* 1991;21:245–260.
62. Myerson RB. Graphs and cooperation in games. *Math Oper Res* 1977;2:225–229.
63. Amer R, Carreras F. Games and cooperation indices. *Int J Game Theory* 1995;24:239–258.
64. Amer R, Carreras F. Cooperation indices and coalitional value. *TOP* 1995;3:117–135.
65. Amer R, Carreras F. Power, cooperation indices and coalitional structures. *Homo Oeconomicus* 2000;17:11–29.

66. Amer R, Carreras F. Power, cooperation indices and coalition structures. In: Holler MJ, Owen G, editors. Power indices and coalition formation. Hamburg: Kluwer Academic Publishers; 2001. pp. 153–173.
67. Amer R, Carreras F. Cooperation indices and weighted Shapley values. *Math Oper Res* 1997;22:955–968.
68. Kalai E, Samet D. On weighted Shapley values. *Int J Game Theory* 1987;16:205–222.
69. Albizuri MJ, Zarzuelo JM. On coalitional semivalues. *Games Econ Behav* 2004;49:221–243.
70. Carreras F. On the existence and formation of partnerships in a game. *Games Econ Behav* 1996;12:54–67.
71. Freixas J, Pons M. Circumstantial power: optimal persuadable voters. *Eur J Oper Res* 2008;186:1114–1126.
72. Freixas J. The dimension for the European Union Council under the Nice rules. *Eur J Oper Res* 2004;156:415–419.
73. Algaba E, Bilbao JM, Fernández JR, *et al.* El índice de poder de Banzhaf en la Unión Europea ampliada. *Qüestió* 2001;25:71–90.
74. Freixas J, Molinero X, Olsen M, *et al.* On the complexity of problems on simple games. Technical report, archiv.org; 2008. Available at <http://arxiv.org/PScache/arxiv/pdf/0803/0803.0404v1.pdf>. Accessed in 2008.
75. Freixas J, Marciniak D. A minimum dimensional class of simple games. *TOP* 2009;17:407–414.
76. Peleg B. On weight of constant sum majority games. *SIAM J Appl Math* 1968;16:527–532.
77. Peleg B. On the kernel of constant-sum simple games with homogeneous weights. *Ill J Math* 1966;10:39–48.
78. Freixas J, Molinero X. Weighted games without a unique minimal representation in integers. *Optim Methods Softw* 2010;25: 203–215.
79. Freixas J. Banzhaf measures for games with several levels of approval in the input and output. *Ann Oper Res* 2005;137:45–66.
80. Tchantcho B, Diffio Lambo L, Pongou R, *et al.* Engoulou. Voters' power in voting games with abstention: influence relation and ordinal equivalence of power theories. *Games Econ Behav* 2008;64:335–350.
81. Freixas J, Zwicker W. Anonymous yes–no voting with abstention and multiple levels of approval. *Games Econ Behav* 2009;69:428–444.
82. Freixas J. The Shapley–Shubik power index for games with several levels of approval in the input and output. *Decis Support Syst* 2005;39:185–195.

PRESOLVING MIXED-INTEGER LINEAR PROGRAMS

ASHUTOSH MAHAJAN
Mathematics and Computer
Science Division, Argonne
National Laboratory,
Argonne, Illinois

INTRODUCTION

Presolving or preprocessing techniques constitute a broad class of methods used to transform a given instance of a mixed-integer linear program (MILP) or to collect certain useful information about it. The main aim of presolving techniques is to simplify the given instance for solving by the more sophisticated and often time-consuming algorithms such as cutting-plane or branch-and-bound algorithms. Almost all modern solvers, both free [1–4] and commercial [5–7], deploy these in one way or the other. Bixby *et al.* [8] claim that presolving alone speeds the solution times on benchmark instances by a factor of 2. Another important function of a presolver is to analyze the structure of the instance and collect information that will be useful when the problem is solved.

Presolving is also useful for the purpose of modeling, as it can be used to check the instance for obvious errors. It can, for instance, inform the user whether certain constraints or bounds on variables make the instance infeasible or whether certain variables will have arbitrarily high values (unbounded) in a solution. Such information can help the user improve the description of the instance and correct errors. It can also sometimes tell the user whether the values used in the model may cause numerical problems for the subsequent algorithms used by the solver. As such, some presolving is also performed by software such as AMPL [9] that is used for creating MILP instances.

In this article, we describe the various techniques used for basic and advanced

presolving. Throughout the discussion, we assume that the original MILP is of the form

$$\begin{aligned} \min \quad & \sum_{j=1}^n c_j x_j \\ \text{s.t.} \quad & \sum_{j=1}^n a_{ij} x_j \leq b_i, \quad i = 1, \dots, m, \\ & l_j \leq x_j \leq u_j, \quad j = 1, 2, \dots, n, \\ & x_j \in \mathbb{Z}, \quad j \in I, \end{aligned} \quad (\text{P})$$

where $m \in \mathbb{N}$, $n \in \mathbb{N}$, $c \in \mathbb{Q}^n$, $A \in \mathbb{Q}^{m \times n}$, $l \in \mathbb{Q}^n$, $u \in \mathbb{Q}^n$, $I \subseteq \{1, 2, \dots, n\}$ are given as inputs. We will explicitly clarify whenever a technique applies to any equality constraints ($\sum_{j=1}^n a_{ij} = b_i$) and not to the inequalities in Equation (P). Some techniques for presolving MILPs are the same as those used for presolving linear programs (LPs). We also describe these techniques here because they are a starting point for more advanced techniques for presolving MILPs. Most of the presolving techniques presented here are derived from the work of [10–13]. Details of implementing the techniques in an MILP solver can be found in the work of Suhl and Szymanski [14] and Achterberg [15].

We start by describing basic presolving techniques in the section titled “Basic Presolving” and then describe more advanced techniques in the section titled “Advanced Presolving”. We also discuss methods for detecting structure (in the section titled “Identifying Structure”) and for postprocessing (in the section titled “Postsolve”).

BASIC PRESOLVING

Basic presolving methods are simple procedures that are directly applied to simplify Equation (P). These methods may be called many times as subroutines in more advanced methods. Hence, implementing them efficiently is important.

Simple Rearrangements and Substitutions

Removing constraints and variables that are not necessary in the instance can reduce the amount of memory required by a computer to solve the instance and also speed up the calculations. Rearranging the order of variables and constraints in the instance can speed several routines in the subsequent preprocessing and solve methods.

Removing constraints. If we have $a_{ij} = 0 \forall j \in \{1, 2, \dots, n\}$ for some $i \in \{1, 2, \dots, m\}$ and if $b_i \geq 0$, then the i th constraint can be deleted. If for the same constraint $b_i < 0$, then the instance is infeasible because of this constraint. If we have a singleton row, namely, $a_{ik} \neq 0, a_{ij} = 0 \forall j \neq k, j \in \{1, 2, \dots, n\}$, for some $i \in \{1, 2, \dots, m\}$ and some $k \in \{1, 2, \dots, n\}$, then the i th constraint can be removed, and the bounds of variable x_k can be tightened if necessary.

Removing variables. If we have $a_{ij} = 0 \forall i \in \{1, 2, \dots, m\}$ for some $j \in \{1, 2, \dots, n\}$, then we can fix the variable x_j to l_j if $c_j \geq 0$ or to u_j if $c_j < 0$. After fixing the variable, we may have an additional constant term in the objective function. If we have a singleton column x_j , and if $j \notin I$, $a_{ij} > 0$, $l_j = -\infty$, $u_j = \infty$, $c_j \leq 0$ then we can drop the constraint i and substitute $x_j = \frac{b_i - \sum_{k: k \neq j} a_{ik} x_k}{a_{ij}}$ in the objective function. After the substitution, we have one less variable and one less constraint. The number of nonzeros in A is also reduced. The number of nonzeros in the objective function, however, may increase.

Rearrangements. The variables and constraints of the instance can be rearranged depending on how the methods of solving are implemented. For example, it may be beneficial to have all binary variables stored in the end of the array of variables so that deleting them (in case we fix them) from the sparse representation of the matrix A is faster. Similarly, the constraints can be rearranged so that “constraints of interest” appear together in the arrays in which they are stored. If it is known that an interior-point method will initially be used for solving the relaxation of Equation (P), then rows and columns may be

rearranged so as to reduce the “fill in.” Rothberg and Hendrickson [16] study the effects of various such permutation techniques on interior-point methods.

Granularity

Given an instance of form (P) and a vector $\mathbf{a} \in \mathbb{Q}^n$, we define granularity of $\mathbf{a}$ as

$$g(\mathbf{a}) = \min_{x^1, x^2 \in \mathcal{P}} \left\{ \left| \sum_{j=1}^n a_{ij} x_j^1 - \sum_{j=1}^n a_{ij} x_j^2 \right| : \sum_{j=1}^n a_{ij} x_j^1 \neq \sum_{j=1}^n a_{ij} x_j^2 \right\},$$

where $\mathcal{P}$ is the set of all points satisfying the constraints of Equation (P). Granularity refers to the minimum difference (greater than zero) between the values of activity that a constraint or objective function may have at two different feasible points. For a constraint that has only integer-constrained variables, then granularity is at least the greatest common divisor (GCD) of all the coefficients. If any of the coefficients in such a constraint are not integers, we can find a multiplier 10^K , where $K \in \mathbb{Z}^+$, to make all coefficients integers. The GCD can be divided by 10^K to get granularity. Hoffman and Padberg [17] call this procedure Euclidean reduction. The problem can be declared infeasible if there is an equality constraint whose right-hand side is not an integer multiple of the GCD. For an inequality constraint $i \in \{1, 2, \dots, m\}$, if b_i is not an integer multiple of the granularity $g(\mathbf{a}_i)$, then b_i can be reduced to the nearest multiple of g_i less than b_i . So the constraint

$$3x_1 + 6x_2 - 3x_3 \leq 5.3$$

can be tightened to

$$3x_1 + 6x_2 - 3x_3 \leq 3,$$

if $x_1, x_2, x_3 \in \mathbb{Z}$. The bound tightening in this case is a special case of a Chvátal–Gomory inequality introduced by Gomory [18]. Similarly, one can find the granularity of the objective vector $\mathbf{c}$ when only

integer-constrained variables have nonzero coefficients in the objective function. The difference between the upper bound obtained from a feasible solution and the optimal solution value will always be at least $g(\mathbf{c})$. If the difference between an upper bound and a lower bound falls below the granularity, we can stop the branch-and-bound or cutting-plane algorithm and claim that the optimal solution has been found.

Constraint and Variable Duplication

Sometimes, an instance may have duplicate constraints that are exact copies of each other. At other times, we may have constraints that are identical except for a scalar multiple. The obvious benefit of detecting such constraints is that we can reduce the memory needed for storing and solving the instance. It also helps improve the numerical stability of solution of the associated LP relaxations. Bixby and Wagner [19] and Tomlin and Welch [20] describe efficient ways to detect from a column-ordered format of A , whether two constraints are identical except for scalar multiples. When constraints with duplicate entries are found, we can, based on the values of the right-hand sides and the scalar multiples, declare the problem infeasible or delete one of the constraints or combine the two constraints into one.

Similarly, we can detect whether two columns are alike except for a scalar multiple. However, the possible integrality constraint on one or both of these variables can make any substitution complicated. If $A_j = \alpha A_i$ and if $c_j = \alpha c_i$ for some $i, j \in \{1, 2, \dots, n\} \setminus I$, $\alpha \in \mathbb{Q}$, then we can replace the two columns A_i, A_j by a single column A_k that is identical to A_i . The lower and upper bounds of the new variable x_k are $l_k = l_i + \alpha l_j$, $u_k = u_i + \alpha u_j$ if $\alpha > 0$ and $l_k = l_i + \alpha u_j$, $u_k = u_i + \alpha l_j$ otherwise. If one or both variables are integer-constrained, we can use the substitution $x_k = x_i + \alpha x_j$ if we can find feasible values of x_i, x_j for each $x_k \in [l_k, u_k]$. We can also add a constraint $x_k = x_i + \alpha x_j$ to avoid such complications.

Constraint Domination

One way to identify a redundant constraint is to check whether it can be termwise

dominated by another constraint. If $b_k \geq b_i$ for $i, k \in \{1, 2, \dots, m\}$, each variable in constraints i, k is either nonnegative or nonpositive, $a_{kj} \leq a_{ij}$ for j such that x_j is nonnegative and $a_{kj} \geq a_{ij}$ for j such that x_j is nonpositive, then constraint k can be deleted because it is redundant in the presence of constraint i and the bounds on variables.

Bound Improvement

Bound improvement is one of the most important steps of presolving. If we can improve the bounds on variables or reduce b_i , then we can tighten the LP relaxation of Equation (P). The tightest bounds on a constraint can be obtained [13] by solving the following problems:

$$\begin{aligned} \mathcal{L}_i &= \min \sum_{j=1}^n a_{ij}^T x \\ \text{s.t. } A^i x &\leq b^i, \\ l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I, \end{aligned} \quad (1)$$

and

$$\begin{aligned} \mathcal{U}_i &= \max \sum_{j=1}^n a_{ij}^T x \\ \text{s.t. } A^i x &\leq b^i, \\ l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I, \end{aligned} \quad (2)$$

where the set of constraints $A^i x \leq b^i$ is obtained by deleting the i th constraint $\sum_{j=1}^n a_{ij} x_j \leq b_i$ from Equation (P). However, solving the bound improvement problems 1 or 2 can be as difficult as solving Equation (P). Thus, there is a trade-off between the quality of a good bound and the time spent in finding it. The most common method of tightening the bound is to calculate bounds on $\mathcal{L}_i$ and $\mathcal{U}_i$:

$$\hat{\mathcal{L}}_i = \sum_{j: a_{ij} > 0} a_{ij} l_j + \sum_{j: a_{ij} < 0} a_{ij} u_j,$$

$$\hat{\mathcal{U}}_i = \sum_{j: a_{ij} > 0} a_{ij} u_j + \sum_{j: a_{ij} < 0} a_{ij} l_j. \quad (3)$$

Both $\hat{\mathcal{L}}_i, \hat{\mathcal{U}}_i$ can be calculated in $\mathcal{O}(nz)$ steps, by visiting each nonzero coefficient of the A matrix.

Clearly $\hat{\mathcal{L}}_i \leq \mathcal{L}_i \leq \mathcal{U}_i \leq \hat{\mathcal{U}}_i$. If $b_i < \mathcal{L}_i$, then the instance is infeasible. If $b_i = \mathcal{L}_i$, then all variables appearing in the constraint are forced to a bound: $x_j = l_j$ if $a_{ij} > 0$ and $x_j = u_j$ if $a_{ij} < 0$. This constraint can then be deleted. If $b_i \geq \mathcal{U}_i$, then the constraint can be deleted because it is redundant.

Similarly, we can improve the bounds on variables. The tightest lower and upper bounds for the variable x_k can be obtained by solving the following:

$$\begin{aligned} L_k &= \min x_k \\ \text{s.t. } Ax &\leq b, \\ l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I, \end{aligned} \quad (4)$$

and

$$\begin{aligned} U_k &= \max x_k \\ \text{s.t. } Ax &\leq b, \\ l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I. \end{aligned} \quad (5)$$

Again, solving the problems (4) and (5) can be as difficult as the original problem (P). A similar trade-off then occurs on the quality of the bound and the time spent in obtaining it. The bounds $\mathcal{L}_i$ and $\mathcal{U}_i$ on constraints can also be used to improve the bounds on variables. If $a_{ik} > 0$, then let

$$\hat{U}_{ik} = \frac{b_i}{a_{ik}} - \frac{\mathcal{L}_i - a_{ik} l_k}{a_{ik}}.$$

Clearly, $\hat{U}_{ik} (\geq U_k)$ is an upper bound on the feasible values of x_k . If $U_k < u_k$ (or $\hat{U}_{ik} < u_k$), then update the bound on variable x_k to U_k . If $U_k < l_k$, then the problem is infeasible. If $U_k = l_k$, then the variable x_k can be fixed to the value l_k . Similarly, if $a_{ik} < 0$, then let

$$\hat{L}_{ik} = \frac{b_i}{a_{ik}} - \frac{\mathcal{U}_i - a_{ik} u_k}{a_{ik}}.$$

Then $\hat{L}_{ik} (\leq L_k)$ is also a lower bound on feasible values of x_k . If $L_k < l_k$, then update the lower bound on variable x_k to L_k . If $L_k > u_k$, then the problem is infeasible. If $L_k = u_k$, then the variable x_k can be fixed to the value u_k . Once $\mathcal{L}_i$ and $\mathcal{U}_i$ (or their bounds $\hat{\mathcal{L}}_i, \hat{\mathcal{U}}_i$) are known, then the bounds $\hat{L}, \hat{U}$ can be calculated in $\mathcal{O}(nz)$ steps.

If the bounds are tightened using the complete models 1 and 2, then constraint domination is just a special case of bound tightening. This is no longer true, however, when the bounds are estimated from each row. In the latter case, we need only $\mathcal{O}(nz)$ calculations, where nz denotes the number of nonzeros in the A matrix.

If the bounds of a variable x_j are implied by a constraint, then the variable can be declared “free”, and we can try substituting the variable out as described in the section titled “Simple Rearrangements and Substitutions.”

Dual/Reduced-Cost Improvement

If, for a variable $x_j, j \in \{1, 2, \dots, n\}$, $a_{ij} \geq 0 \forall i \in \{1, 2, \dots, m\}$, and $c_j \geq 0$, then the variable x_j can be fixed to its lower bound l_j . Similarly, if $a_{ij} \leq 0 \forall i \in \{1, 2, \dots, m\}$ and $c_j \leq 0$, then the variable x_j can be fixed to its upper bound u_j .

Sometimes, an upper bound on the optimal value of solution of Equation (P) is known (it may be provided by the user or may be obtained from some heuristics) before presolving. In such a case, we have an additional inequality $\sum_{j=1}^n c_j x_j \leq z_u$, where z_u is the best-known upper bound. The basic presolving techniques applicable to other constraints can be applied to this constraint as well. Valid inequalities of the form $\sum_{j=1}^n \hat{c}_j x_j \geq z_l$, where $z_l \in \mathbb{Q}$, can also be obtained by linear combinations of the objective function with other constraints. Such inequalities are automatically obtained from the reduced costs calculated by the simplex method. When both z_l and z_u are available as described above, we can do bound tightening on variables to obtain tighter bounds. This is equivalent to the reduced-cost fixing method described by Balas and Martin [21].

ADVANCED PRESOLVING

Advanced presolving procedures try to modify the problem and then call basic preprocessing procedures repeatedly to derive more information about the effects of such changes. Such procedures are also called *probing*. Unlike basic preprocessing techniques, probing may be expensive because one can probe on many possible changes or on combinations of several changes. The simplest changes are the changes of bounds on binary variables. We can fix such a variable to zero or 1 and then probe on the reduced problem.

Fixing Variables

When probing on a binary variable, say x_i , we first fix x_i to zero by changing its upper bound. We then perform basic preprocessing on the modified problem. If the modified problem can be shown to be infeasible, then x_i can be fixed to one in the original problem. Similarly, if the problem becomes infeasible on setting x_i to 1, we can fix x_i to zero. If the problem is infeasible on setting x_i to 1 and zero both, we can declare that the problem is infeasible.

Improving Coefficients

If a constraint i , $\sum_{j=1}^n a_{ij}x_j \leq b_i$ is redundant when a binary variable x_k is set to zero, then we can improve the coefficient a_{ik} of x_k in this constraint using the approach of Crowder *et al.* [22]. We first observe that changing a_{ik} does not make any difference when $x_k = 0$. When x_k is set to 1, the original constraint is equivalent to $\sum_{j=1}^n a_{ij}x_j - \delta_0 x_k \leq b_i - \delta_0$, for any $\delta_0 \in \mathbb{R}$. We would like to have as high a value of δ_0 as possible, while still keeping the constraint redundant at $x_k = 0$. This value of δ_0 can be obtained by solving the following problem:

$$\delta_0 = b_i - \max_Q \sum_{j=1}^n a_{ij}x_j,$$

where Q is the set of values of x such that

$$\begin{aligned} A^i x &\leq b^i, \\ x_k &= 0, \end{aligned}$$

$$\begin{aligned} l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I. \end{aligned}$$

Again, the problem of finding the optimal value of δ_0 can be as hard as solving the original problem (P). A more easily computable bound,

$$\hat{\delta}_0 = b_i - \sum_{j:j \neq k, a_{ij} > 0} a_{ij}u_j - \sum_{j:j \neq k, a_{ij} < 0} a_{ij}l_j,$$

can be used instead.

Similarly, if the constraint i , $\sum_{j=1}^n a_{ij}x_j \leq b_i$ is redundant when a binary variable x_k is set to 1, we can replace the constraint by $\sum_{j=1}^n a_{ij}x_j + \delta_1 x_k \leq b_i$, with

$$\delta_1 = b_i - \max_Q \sum_{j=1}^n a_{ij}x_j,$$

where Q is the set of values of x such that

$$\begin{aligned} A^i x &\leq b^i, \\ x_k &= 1, \\ l_j &\leq x_j \leq u_j, \quad j = 1, 2, \dots, n \\ x_j &\in \mathbb{Z}, \quad j \in I. \end{aligned}$$

As above, we can use a bound that is computationally easier to calculate,

$$\hat{\delta}_1 = b_i - a_{ik} - \sum_{j:j \neq k, a_{ij} > 0} a_{ij}u_j - \sum_{j:j \neq k, a_{ij} < 0} a_{ij}l_j.$$

Deriving Implications

Implications are relations that the variable values of any optimal feasible solution must satisfy. One such implication can be found by fixing two binary variables, say, x_i, x_j , where $i, j \in I$ to zero. If the modified problem is shown to be infeasible, then we know that all solutions must satisfy the inequality $x_i + x_j \geq 1$. An equivalent way of looking at this implication is that, if we fix x_i to zero, then x_j is constrained to be 1. A more general case is when fixing x_i to zero implies $x_j = v_j, k \in \{1, 2, \dots, n\}$, where $v_j \in \mathbb{R}$. Then

the following two inequalities are valid for Equation (P):

$$\begin{aligned} x_j &\leq v_j + (u_j - v_j)x_i \\ x_j &\geq v_j - (v_j - l_j)x_i. \end{aligned}$$

Similarly, if by fixing x_i to 1 we can fix x_j to some $v_j \in \mathbb{R}$, then we can write

$$x_i = 1 \Rightarrow x_j = v_j \iff \begin{cases} x_j \leq u_j - (u_j - v_j)x_i, \\ x_j \geq l_j + (v_j - l_j)x_i. \end{cases}$$

When x_j is also binary, we can use the above procedure to obtain the following implications:

$$\begin{aligned} x_i = 0 \Rightarrow x_j = 0 &\iff x_i - x_j \geq 0, \\ x_i = 0 \Rightarrow x_j = 1 &\iff x_i + x_j \geq 1, \\ x_i = 1 \Rightarrow x_j = 0 &\iff x_i + x_j \leq 1, \\ x_i = 1 \Rightarrow x_j = 1 &\iff x_i - x_j \leq 0. \end{aligned}$$

Deriving such implications is useful not only for preprocessing but also for other techniques such as primal heuristics, branching, and generating valid inequalities, which are subsequently used in solving Equation (P). The number of such implications can, however, be very large for certain types of problems. For example, if we have a set partitioning constraint of the form

$$\sum_{i \in S} x_i = 1,$$

where x_i is binary for $i \in S$, then the number of valid implications with only two variables would be $\mathcal{O}(|S|^2)$. Thus it is important to implement methods to store and retrieve them quickly. Several implications can also be combined to fix variables, to delete constraints, and to derive new implications: e.g., if we have implications $x_1 = 0 \Rightarrow x_2 = 0$, and $x_1 = 1 \Rightarrow x_2 = 0$, then we can fix x_2 at 0. Similarly, if a constraint is redundant when $x_1 = 0$ and also when $x_1 = 1$, then we can delete that constraint. The algorithms for deriving such new information from an existing list of implications can be studied with the help of what is called a *conflict graph*.

A conflict graph is a graph that shows what values of binary variables are not

feasible (or optimal) for Equation (P). The graph has two sets $X, \bar{X}$ of k vertices each, where k is the number of binary variables. For each binary variable x_i in Equation (P), we have a vertex each in X (denoted i), and in $\bar{X}$ (denoted $\bar{i}$). A vertex $i \in X$ corresponding to variable x_i is associated with the implication “ $x_i = 1$ ”. Similarly, a vertex $i \in \bar{X}$ is associated with “ $x_i = 0$ ”. An edge between two vertices denotes a conflict, i.e., any optimal solution of Equation (P) cannot have values associated with the two vertices of any edge in the conflict graph. Figure 1 shows a conflict graph in three variables. Three implications that create the edges are $x_1 = 1 \Rightarrow x_2 = 1$, $x_3 = 1 \Rightarrow x_2 = 1$, and $x_3 = 1 \Rightarrow x_1 = 0$. The dark edges in the graph denote that either $x_i = 1$ or $x_i = 0$, and one of them necessarily holds. New implications can now be derived by studying this graph: e.g., if there is a vertex j with neighbors $i, \bar{i}$, then x_j can be fixed to 0. We refer the readers to the work of Savelsbergh and Atamtürk [13] and Savelsbergh [23] for algorithms that can be used to derive such implications. The latter also describe data structures for storing them efficiently. We now mention two more ways in which implications are useful for improving a formulation.

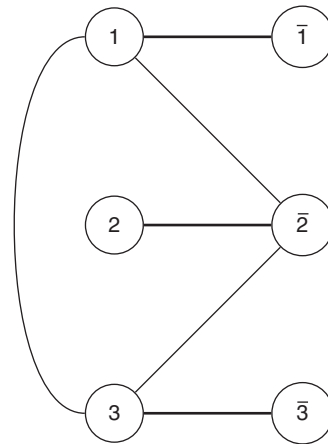


Figure 1. An example of a conflict graph.

Automatic disaggregation. Savelsbergh and Atamtürk [13] show that, if we add the inequalities derived from implications as above, then we automatically disaggregate each constraint of the form

$$\sum_{i \in S} x_i \leq Ux_k,$$

where $S \subseteq \{1, 2, \dots, n\}$, $k \in \{1, 2, \dots, n\}$, and $U \leq \sum_{i \in S} u_i$, into the inequalities

$$x_i \leq u_i x_k, i \in S.$$

Elimination of variables. If fixing a binary variable x_k to zero implies $x_i = r_i$ for some $i \in \{1, 2, \dots, n\}$ and fixing x_k to one implies $x_i = s_i$ for some $r_i, s_i \in \mathbb{R}$, then we can substitute

$$x_i = r_i + (s_i - r_i)x_k.$$

The variable x_i can now be eliminated from the problem without increasing the number of nonzeros in the A matrix.

Probing on Constraints

Consider a clique inequality of the form

$$\sum_{i \in S^+} x_i - \sum_{i \in S^-} x_i \leq 1 - |S^-|,$$

where $S^+, S^- \subseteq \{1, 2, \dots, n\}$, and all x_i are binary for $i \in S^+ \cup S^-$. The tuple $\{x_i\}, i \in S^+ \cup S^-$ can assume only $|S^+ \cup S^-| + 1$ values for any feasible point. Out of these, $|S^+ \cup S^-|$ values correspond to exactly one of $x_i, i \in S^+$ being 1 or exactly one of $x_i, i \in S^-$ being zero. All these cases are therefore considered automatically when probing on these variables. The only remaining case that is feasible for this constraint is when $x_i = 0, \forall i \in S^+$ and $x_i = 1, \forall i \in S^-$. If the problem is infeasible for this case, then the inequality can be tightened to an equality. If some other inequality is redundant, then the coefficients can be tightened following the procedure in the section titled “Improving Coefficients”.

IDENTIFYING STRUCTURE

Many techniques for generating valid inequalities, branching, and heuristics are applicable only for particular types of constraints and variables. Yet others may be greatly improved if some special structures are known to exist. Since such techniques are used repeatedly many times in the branch-and-bound or cutting-plane methods, it is important to detect up front any useful structures in the given MILP. Some of these structures are listed below.

- Special ordered sets of type 1 (SOS1) are of the form

$$\sum_{i \in S} x_i = 1,$$

$$x_k = \sum_{i \in S} a_i x_i,$$

$$x_i \in \{0, 1\}, i \in S \subseteq \{1, 2, \dots, n\}.$$

These can be used for special branching rules [24].

- Set partitioning ($\sum_{i \in S} x_i = \alpha, x_i \in \{0, 1\}, i \in S$), set covering ($\sum_{i \in S} x_i \geq \alpha$), and set packing inequalities ($\sum_{i \in S} x_i \leq \alpha$) can be used for primal heuristics [25] and for generating valid inequalities [26].
- Knapsack inequalities including 0-1 knapsack, mixed 0-1 knapsack, integer knapsack, and mixed-integer knapsack inequalities [27].
- Generalized upper bound constraints can be used to generate valid inequalities [28].
- Variable lower bound and upper bound constraints can also be used to generate valid inequalities [29].

POSTSOLVE

After an instance has been treated by a presolver and then solved, the user may ask for the solution and other problem characteristics. The presolver can change some instances substantially and the solution of the transformed problem may be quite different from that of the original

problem. Hence the presolve routines are designed to save the changes that were made to the original instance. Andersen and Andersen [10] suggest keeping a “stack” of all the transformations that were made by the presolver and reverting them in a last-in-first-out sequence. Sometimes, the solution obtained by reverting the changes may not remain feasible for the specified tolerance values even though the presolved solution was feasible. In such cases, the user must either decrease the tolerance limit or disable presolve.

CONCLUDING REMARKS

In this article, we have surveyed many techniques for presolving. Some of these techniques are simple, like removing empty constraints, and empty columns, identifying special structures and rearranging constraints and variables. Some others are more powerful, like detecting duplicate rows and columns. They require specialized algorithms to avoid spending excessive amount of time. Then there are some methods like bound tightening, detecting infeasibility, and identifying redundant constraints, which in theory can be as difficult as solving the original problem itself. Heuristic methods are used to trade-off the time spent in these problems against the benefits of solving them. All of these methods usually consider only bounds on variables and a single constraint. Advanced presolving methods can help simplifying the problems a lot but each probe requires repeating basic presolving methods. Further, one needs to probe over many candidate variables. In general, it is difficult to predict the optimal effort that should be put into each method of presolving. Most implementations of a presolver rely on several rules of thumb and the past experiences with these methods.

A presolver is an important component of modern MILP solvers. It can substantially reduce the time taken to solve MILPs. It is also useful for conveying to the user obvious problems with the formulation of the instance. However, the percentage of time spent by a solver in presolve routines is usually much smaller than the percentage of

lines of code required to write a presolver. Even though the techniques of presolving are simple, implementing them effectively for all types of formulations and inputs can be a challenging activity.

Acknowledgments

The author gratefully acknowledges the anonymous referee for suggestions and recommendations that helped improve this survey, and also Gail Pieper for proofreading an earlier draft. This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, US Department of Energy, under Contract DE-AC02-06CH11357. This work was also supported by the US Department of Energy through grant DE-FG02-05ER25694.

REFERENCES

1. Achterberg T. SCIP: solving constraint integer programs. *Math Program Comput* 2009;1(1):1–49.
2. Forrest J. COIN-OR branch and cut. 2010. Available at <https://projects.coin-or.org/Cbc>. Accessed 2010.
3. Nemhauser G, Savelsbergh M, Sigismondi G. MINTO, a mixed INTEGER optimizer. *Oper Res Lett* 1994;15(1):47–58.
4. Ralphs T, Guzelsoy M, Mahajan A. SYMPHONY 5.2.3 user’s manual. 2010. Available at <https://projects.coin-or.org/SYMPHONY>. Accessed 2010.
5. FICO. Xpress-Mosel user guide. 2009. Available at <http://optimization.fico.com>. Accessed 2009.
6. Gurobi. Gurobi optimizer reference manual. 2009. Available at <http://www.gurobi.com>. Accessed 2009.
7. IBM. User’s manual for CPLEX V12.1. 2009. Available at <http://ibm.com/software/integration/optimization/cplex/>. Accessed 2009.
8. Bixby RE, Fenelon M, Gu Z, *et al.* MIP: theory and practice - closing the gap. In: Powell MJD, Scholtes S, editors. *System modelling and optimization: methods, theory, and applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000. pp. 19–49.
9. Fourer R, Gay D, Kernighan B. AMPL: a modelling language for mathematical programming. Pacific Grove (CA): Brooks/Cole Publishing Company; 2003.

10. Andersen E, Andersen K. Presolving in linear programming. *Math Program* 1995;71:221–245.
11. Brearley A, Mitra G. Analysis of mathematical programming problems prior to applying the simplex algorithm. *Math Program* 1975;8:54–83.
12. Guignard M, Spielberg K. Logical reduction methods in zero-one programming: minimal preferred variables. *Oper Res* 1981;29(1):49–74.
13. Savelsbergh MWP. Preprocessing and probing techniques for mixed integer programming problems. *ORSA J Comput* 1994;6:445–454.
14. Suhl UH, Szymanski R. Supernode processing of mixed-integer models. *Comput Optim Appl* 1994;3:317–331.
15. Achterberg T. Constraint integer programming [PhD thesis]. Technical University of Berlin; 2007.
16. Rothberg E, Hendrickson B. Sparse matrix ordering methods for interior point linear programming. *INFORMS J Comput* 1998;10(1):107–113.
17. Hoffman K, Padberg M. Improving LP-representations of zero-one linear programs for branch-and-cut. *ORSA J Comput* 1991;3(2):121–134.
18. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64:275–278.
19. Bixby RE, Wagner D. A note on detecting simple redundancies in linear systems. *Oper Res Lett* 1987;6(1):15–17.
20. Tomlin J, Welch J. Finding duplicate rows in a linear programming model. *Oper Res Lett* 1986;5(1):7–11.
21. Balas E, Martin C. Pivot and complement - a heuristic for 0–1 programming. *Manage Sci* 1980;26(1):86–96.
22. Crowder H, Johnson EL, Padberg M. Solving large scale zero-one linear programming problems. *Oper Res* 1983;31(4):803–834.
23. Atamtürk A, Savelsbergh MWP. Conflict graphs in solving integer programming problems. *Eur J Oper Res* 2000;121:40–55.
24. Beale EML, Tomlin JA. Special facilities in general mathematical programming system for non-convex problems using ordered sets of variables. In: Lawrence J, editor. *Proceedings of the 5th International Conference on Operations Research*. London: Tavistock Publications; 1970. pp. 447–454.
25. Atamtürk A, Nemhauser GL, Savelsbergh MWP. A combined Lagrangian, linear programming and implication heuristic for large-scale set partitioning problems. *J Heuristics* 1995;1:247–259.
26. Balas E, Padberg MW. Set partitioning: a survey. *SIAM Rev* 1976;18(3):710–760.
27. Atamtürk A. Cover and pack inequalities for (mixed) integer programming. *Ann Oper Res* 2005;139(1):21–38.
28. Wolsey LA. Valid inequalities for 0–1 knapsacks and MIPs with generalized upper bound constraints. *Discrete Appl Math* 1990;29:251–261.
29. van Roy TJ, Wolsey LA. Solving mixed integer programming problems using automatic reformulation. *Oper Res* 1987;35(1):45–57.

PRICING AND LEAD-TIME DECISIONS

PELİN PEKGÜN
JDA Software Group,
Marietta, Georgia

In today's increasingly global and competitive marketplace, the demand for goods and services and in turn, the profits of the companies, are not only shaped by their inventory policies and capacity allocation decisions but also by their price and lead-time decisions. For companies that move from a make-to-stock to a make-to-order (MTO) model to satisfy their customers' unique needs, quoting the right price and the reliable lead-time is even more important. Promising an unrealistic due date to an upcoming order may result in monetary penalties as well as diminished prospects for future business. From a firm's point of view, although a shorter lead-time may attract customers and generate more demand, it puts pressure on the firm's production resources with a risk of demand exceeding capacity. On the other hand, customers might be willing to wait longer for their orders if they are offered lower prices. Therefore, depending on its current workload and market conditions, the firm may find it more profitable to offer customers shorter lead-times at the expense of higher prices or vice versa. In this article, we provide an introductory review on monopolistic and competitive models for price and lead-time decisions. The reader should note that papers reviewed in this article are by no means exhaustive. For an extensive review on due date management policies, including simulation-based papers and descriptions of industrial applications, see Keskinocak and Tayur [1]. We also refer the reader to Kaminsky and Hochbaum [2] for a detailed review of analytical models for due date quotations and to Cheng and Gupta [3] for an early survey in this area.

Quoting the right prices and reliable lead-times involves the decisions and actions of both, the marketing and operations/

production functions of a company. In many companies, production is evaluated based on cost and service while marketing is evaluated based on revenue and volume. However, dividing a firm into independent units for performance measurement based on accounting terms often leads to misaligned incentives and suboptimal system performance. In this article, we review studies that assume a centralized decision maker as well as those that assume a decentralized decision making structure and develop coordination mechanisms to align the incentives of marketing and production.

We discuss papers that study monopolistic models in the first section, and those that study the interaction among multiple firms under competition in the section titled "Competitive Setting." Note that within the scope of this article, we refer to the production lead-time and do not consider shipping lead-times. We provide our conclusions in the concluding section.

MONOPOLISTIC SETTING

This section reviews steady-state and dynamic models, where customer demand is a function of the quoted price and lead-time, for the case of a single firm or a single supply chain. We first review models with a centralized decision maker for price and lead-time. At the end of the section, we review decentralized models, where price and lead-time are chosen by different decision makers with different objectives.

Uniform Lead-time Guarantees and Price Decisions

In steady-state models, firm operations are typically modeled as an M/M/1 queue with mean production rate (μ) corresponding to the capacity of the firm and mean arrival rate (λ) corresponding to customer demand. It is shown in Ref. 4 that for high service levels, the exponential distribution provides a close approximation to the tail of the waiting time

distribution even for a G/G/s queue. Thus, the M/M/1 queue assumption makes the problem tractable while the derived results can be perceived as approximately valid for more general demand and service time characteristics. These models develop uniform lead-time guarantees and a single price decision independent of the current state of the system. Moreover, a fixed scheduling rule, such as first-come-first-serve (FCFS), is assumed. Customer demand is modeled as an additive or multiplicative function of price and lead-time.

The additive demand function decreases linearly in the quoted price and lead-time and is of the following form:

$$\lambda = D(p, L) = a - bp - cL. \quad (1)$$

In this equation, a is the total market potential, b is the price sensitivity of demand, c is the lead-time sensitivity of demand, and p and L are the quoted price and lead-time, respectively. As discussed in Ref. 5, the linear demand function has several desirable properties. One desirable property is that both the price elasticity and lead-time elasticity of demand increase in the quoted price and lead-time. This captures the practical observations that customers are more sensitive to the prices they are paying when they have to wait longer, or more sensitive to the waiting times if they are paying higher prices. Note that the Cobb–Douglas demand function, which we will present next, does not exhibit this property. Another desirable property of the linear demand function is the separability of price and lead-time, which reflects customers’ perception of time and money as substitutes, and thus, captures the trade-off between quoted price and lead-time.

Palaka *et al.* [5] study a firm, which serves customer demand that is linearly decreasing in the quoted price and lead-time. The objective of the firm is to maximize total profit. Three types of cost components are considered: variable production costs; congestion-related costs given by the mean number of jobs in the system multiplied by the unit holding cost; and lateness penalty costs incurred as a result of not meeting the quoted lead-times. In order to prevent

unrealistically short lead-times, the authors impose a service-level constraint, which specifies the minimum probability of meeting the quoted lead-time (s):

$$1 - e^{-(\mu - \lambda)L} \geq s. \quad (2)$$

For computational simplicity, a transformation of variables is employed and the service-level constraint is redefined as

$$(\mu - \lambda)L \geq k, \quad (3)$$

where $k = \log(1/(1 - s))$. It is shown that if the service level, s , is at least as large as a certain threshold, which is a function of the price and lead-time sensitivities and lateness penalty, then, the service-level constraint is tight at optimality. In this case, the problem can be transformed into a univariate one in λ for which there exists a unique optimal solution. The authors also consider capacity expansion with a linear expansion cost, and show that whenever the firm is faced with increased inventory holding or lateness penalty costs, higher lead-time sensitivity or service level, the optimal demand rate decreases despite the increase in the optimal capacity.

The following papers use the linear demand function but do not include congestion-related costs or lateness penalty costs in the objective function. They consider capacity expansion as a decision variable, and impose the same service-level constraint, which is shown to be tight at optimality. Ray and Jewkes [6] study a variant of the linear customer demand model in Ref. 5 by treating price as a function of lead-time. Their results show that the lead-time strategy for firms whose customers are more sensitive toward price than lead-time is different from firms whose customers want shorter lead-time and are ready to pay a price premium. The authors also extend this model by incorporating economies of scale, where the unit operating cost is a decreasing function of the demand rate. Boyaci and Ray [7] employ a linear demand function in the case of two substitutable products for which dedicated capacities are allocated:

$$\lambda_1 = a - bp_1 - cL_1 + \beta(p_2 - p_1) + \gamma(L_2 - L_1), \quad (4)$$

$$\lambda_2 = a - bp_2 - cL_2 + \beta(p_1 - p_2) + \gamma(L_1 - L_2). \quad (5)$$

In these equations, β is the sensitivity of switch-overs toward price difference, while γ is the sensitivity of switch-overs toward guaranteed lead-time difference. The lead-time of the standard (slower) product is assumed to be fixed. It is shown that whenever the cost differential is constant, higher marginal capacity costs motivate the firm to differentiate less between the two products, both in price and time, whereas higher cost differentials lead to less time differentiation but more price differentiation. Moreover, ignoring product substitutability can result in higher or lower prices, lead-time, and differentiation, depending on the capacity costs and the market parameters.

As compared to the linear demand approach of the above papers, So and Song [4] use the multiplicative Cobb–Douglas demand function:

$$D(p, L) = ap^{-b}L^{-c}. \quad (6)$$

The log transformation of this function leads to a linear functional form. The price elasticity and lead-time elasticity of demand are constant and directly given by parameters b and c , respectively. Among other results, it is shown that as the capacity of the firm increases, a shorter lead-time should be used, but the direction of the price change depends on the lead-time elasticity of demand. Moreover, optimal lead-time selection is more critical for firms with higher operating costs than those with lower operating costs.

Dynamic Price and Lead-time Quotation

Despite its popularity, there are shortcomings of uniform lead-time guarantees. When demand is high, these lead-times may be understated, leading to missed due dates and disappointed customers, or to higher costs due to expediting. When demand is low, they may be overstated and some customers

may choose to go elsewhere. There are a number of papers that study dynamic price and lead-time quotations taking into account shop-floor congestion for MTO firms, where sequencing decisions are also considered. We refer the reader to the article titled **Pricing and Scheduling Decisions** for an introductory review on pricing and scheduling decisions, and to Ref. 8 for a detailed review on queueing games.

Easton and Moodie [9] treat a tendered bid of price and lead-time decisions as contingent on the firm's future production capacity until the customer makes a decision to accept, reject or modify the firm's bid. The authors use a two-dimensional logit model, based on the contract price and lead-time, to estimate the probability of a successful bid assuming a tardiness penalty if the firm fails to complete the job by the promised date. In order to evaluate this penalty for various lead-times, a model is devised for the probability distribution of the amount of available shop time from the current period to any period in the future, based on the firm's current backlog level. Watanapa and Techanitisawad [10] extend this work and propose a bidding model with multiple customer segments enabling resequencing of orders to accommodate attractive schedules of time-sensitive jobs.

Plambeck [11] models a manufacturer's operations as an exponential single server queue with two classes of customers that differ in their price and lead-time sensitivities. The manufacturer chooses a service rate and a static price for each class of customer, and then dynamically quotes lead-times to potential customers as well as choosing their processing order. The arrival rate for each class decreases linearly with price and lead-time. A policy is proposed such that impatient customers pay a premium for immediate delivery and receive priority in scheduling, while patient customers are quoted a lead-time proportional to the current queue length. Queue length and lead-time are approximated by a reflected Ornstein–Uhlenbeck diffusion process. Celik and Maglaras [12] also use an approximated diffusion control problem to derive near-optimal dynamic pricing, lead-time quotation, sequencing, and expediting policies for the case of an MTO manufacturer

who offers multiple products to a market of price and lead-time sensitive users. Note that in this paper, the firm commits to a lead-time from a set of predetermined lead-times instead of using dynamic lead-time control, and focuses on dynamic pricing for these products.

There is also another stream of research in this area to understand the impact of the delay cost shape on pricing and lead-time decisions. Ata and Olsen [13] consider an MTO firm that dynamically quotes lead-times and prices. Customers are homogenous with the same delay cost function but have nonlinear disutility for the lead-times that they are quoted. As the firm is a monopolist, the problem reduces to one of lead-time quotation and order sequencing. The price for a quoted lead-time is given by the price for the maximum acceptable lead-time plus the customer's delay cost savings for the shorter lead-times. Lead-times are quoted via a shift-based approach and once a lead-time is quoted, it must be met. This approach helps to model the trade-off between committing future capacity beforehand or reserving it for potential higher revenue customers who will arrive later. Using a large-capacity asymptotic regime, the authors provide recommended policies for convex, concave, and convex-concave delay cost functions and prove that these policies are asymptotically optimal. FCFS sequencing is found to be optimal when costs are convex but not necessarily so in other cases. For nonconvex delay cost functions, the convex hull is shown to provide a lower bound for the costs. The general convex-concave, or "S-shaped," cost curve models the case where customers have a particular deadline in mind but once that deadline passes, there is increasingly little difference in the added lead-time. The authors find that in this case, most customers are given relatively short lead-times but a small portion of customers are given very long lead-times as a hedge against demand uncertainty. The numerical studies show that significant benefits can be gained by using the proposed dynamic policies instead of FCFS-based policies. Akan *et al.* [14] extend this work to the nonhomogeneous customer setting, where the firm dynamically

offers menus of prices and lead-times. There are two classes of customers with the same valuation for the product but a different level of patience. Readily implementable policies are proposed for convex-concave delay costs in the context of a large-capacity asymptotic regime. The authors first consider the full information case, where the system manager is aware of the customer types and extracts the entire surplus. On the other hand, under the incomplete information case, where the system manager is unaware of the customer type while offering incentive-compatible menus, achieving the optimal profit under full information depends on how dissimilar the two types of customers are.

Decentralized Decision Making for Price and Lead-Time

So far, we have discussed papers that employ a centralized decision making approach. There are only a few studies that measure the impact of decentralization of price and lead-time decisions. Liu *et al.* [15] consider price and lead-time decisions in a two-firm setting within a supply chain; a supplier and a retailer, where the supplier also needs to choose the transfer price. The authors focus on the inefficiencies that result from the double marginalization in the supply chain and find that decentralization leads to lower profits, higher prices, shorter lead-times, and lower demand. However, they do not consider coordinating mechanisms. A recent paper by Pekgun *et al.* [16] studies a single firm with decentralized marketing and production functions, where marketing is evaluated as a revenue center and production is evaluated as a cost center. Pricing decisions are made by marketing, whereas lead-time decisions are made by production. Since the decentralized settings in the two papers are different, in Ref. 16, decentralization leads to lower profits but lower prices, longer lead-times and higher demand as opposed to the findings in Ref. 15. The authors compare different decision making sequences within the firm and find that a decentralized setting with marketing as the leader dominates one in which production is the leader, particularly for firms operating under tight capacity. It is shown that coordination can be achieved

through a transfer price contract with bonus payments, where marketing pays production a transfer price per unit produced, and both departments receive a fraction of the overall revenue generated as a bonus payment. Note that both Refs 15 and 16 model customer demand as a linear function of price and lead-time in steady state as in Ref. 5.

We conclude the monopolistic setting with a brief discussion on the implications of firm parameters. Under both types of models, higher operating costs imply a higher price and a lower lead-time resulting in lower optimal demand and profit. Additionally, higher capacity implies a lower lead-time but not always a lower price. Although higher capacity implies a higher profit under a centralized structure, Pekgun *et al.* [16] show that it may result in a decrease in profit when marketing is the leader and production is the follower within the organization. Finally, as the service level increases, So and Song [4] show that the price should be reduced to keep the demand under a multiplicative demand model, while Pekgun *et al.* [16] show that the price can be increased up to a certain level, where capacity is sufficient to charge higher for better service, under an additive demand model.

COMPETITIVE SETTING

Most of the research on price and lead-time decisions in steady-state under competition uses queueing models and considers centralized firms. Several researchers employ a “waiting time standard,” which is determined by the time spent in a queue given the allocated capacity. Some researchers aggregate price and waiting time into a “full price” instead of modeling customers’ sensitivity to price and lead-time independently. Loch [17] and Armony and Haviv [18] study two competing service providers with an M/G/1 queue in the former and an M/M/1 queue in the latter and two customer classes, where each class has a distinct waiting cost rate and chooses a provider based on its full price. In Ref. 17, it is shown that the game has a unique Nash equilibrium when firms are symmetric, that is, they have the same

processing time distribution parameters. In Ref. 18, competition is modeled in two stages such that providers compete on the basis of service charges in the first stage, and customer classes compete with allocation decisions in the second stage, where they decide to join or give up the service based on full prices although they cannot observe queue sizes. It is shown that a pure price equilibrium may fail to exist.

Chen and Wan [19] study two M/M/1 service providers that compete for a single customer class on the basis of full price. Providers are differentiated by their service capacities, values of provided service and unit costs of waiting. It is found that the firm with the higher speed or value of service and lower waiting cost charges a price premium, and generates a larger market share when there is a unique equilibrium; however, such a firm may charge lower prices and generate smaller market share and/or lower profit when multiple equilibria exist. Lederer and Li [20] consider N M/G/1 service providers, who compete for N customer classes by choosing prices, production rates, and scheduling policies. They assume that providers are full price takers, and show that Nash equilibrium exists when expected waiting time for a class at a firm is a convex function of the firm’s demand rates for different customer classes. They find that the firm with the faster service, lower variability, and lower cost always generates a larger market share, higher capacity utilization, and profit. This result is accompanied with either higher prices and shorter lead-times, or lower prices and longer lead-times. When customers are differentiated by their demand function and delay sensitivity, competing firms that jointly serve several types of customers match prices and lead-times except for those customers that are the most and least patient (based on their delay costs). While capacity is treated constant in these studies, Cachon and Harker [21] consider the option of outsourcing to a supplier for two competing firms, which experience economies of scale as their unit costs are decreasing in the demand volume. Two types of competition are analyzed: an M/M/1 queueing game with price and

time-sensitive demand, which is modeled as a function of the full prices of both firms in two forms; linear and truncated logit, and an economic order quantity (EOQ) game with fixed ordering costs and price-sensitive demand.

In contrast to the full price approach of the above papers, Allon and Federgruen [22,23] treat price and waiting time independently, where demand decreases in own-price/time effects and increases in cross-price/time effects. Both papers model N M/M/1 firms, the former for a single customer class and the latter for N customer classes. In Ref. 22, the authors use service level, which is defined as the difference between an upper-bound benchmark for waiting time and the firm's actual waiting time standard, whereas in Ref. 23, waiting time is explicitly incorporated into the demand model. Along the same line of research, So [24] extends the work in Ref. 4 to a competitive setting of N M/M/1 firms using a multiplicative competitive interaction model, where the market size is fixed and shared among firms based on their "attraction" given their quoted prices and lead-times. It is shown that the equilibrium solution in the case of identical firms behave in a similar fashion as the optimal solution in the monopolistic firm setting in Ref. 4. On the other hand, if the firms in the market are nonidentical, then, these firms will exploit their distinctive firm characteristics to differentiate their services, where the high-capacity firms provide better lead-time guarantees, and firms with lower operating costs offer lower prices. The differentiation becomes more intense as demand becomes more lead-time sensitive.

A recent work by Pekgun *et al.* [25] extends the work in Ref. 16 and studies centralized and decentralized decision making comparatively for price and lead-time decisions under competition at steady state. The problem is modeled as a two-stage game. In the first stage, firms simultaneously choose their organizational structures to operate in centralized or decentralized mode. In the second stage, they simultaneously choose their price and lead-time decisions. Under decentralization, the price and lead-time strategy is established through a Stackelberg

game, where marketing first sets the price as the leader and then production sets the lead-time as the follower. The authors find that under intense price competition, where the intensity is characterized by the underlying parameters of market demand, firms may suffer from a decentralized structure. This finding has the most impact under high flexibility induced by high capacity, where revenue-based sales incentives motivate sales/marketing for more aggressive price cuts resulting in eroding margins. On the other hand, low-cost firms may benefit from a decentralized structure without hurting margins as long as the intensity of price competition is less than the intensity of lead-time competition. The authors also find that higher capacity does not always result in higher profits under competition even if it comes for free.

CONCLUSIONS

In this article, we have provided an introductory review on price and lead-time decisions, both in monopolistic and competitive settings. In the monopolistic setting, we first reviewed the papers that study uniform lead-time guarantees and price decisions at steady state, where firm operations are typically modeled as an M/M/1 queue. Extensions to general queueing models and multiple product/customer types (particularly under shared capacity) are still open for future research. Given the shortcomings of uniform lead-time guarantees since the steady-state results (e.g., the service-level constraint) may not always hold in the day-to-day operations of a firm, we discussed papers that study dynamic price and lead-time quotations taking into account shop-floor congestion and incorporating sequencing decisions. Extensions in this area for assemble-to-order or make-to-stock systems are potential areas for future research. We also reviewed recent work on the impact of the form of delay cost function on price and lead-time decisions, which is a fairly new area and can be further investigated with more detailed customer modeling and for the impact of competition. Next, we discussed papers that

study decentralized decision making for price and lead-time decisions and the inefficiencies that arise as compared to a centralized structure. We are not aware of any papers that study the impact of decentralization for multiplicative demand models or for dynamic price and lead-time quotations, which can be considered for future research. Finally, we reviewed papers on price and lead-time decisions under competition. While the existence of an equilibrium usually depends on the problem structure and certain conditions, it is possible to observe insights consistent with the monopolistic setting in most cases. However, there does not exist much work yet for dynamic models under the presence of competitors with a centralized or decentralized organizational structure, which can be attributed to the analytical complexity encountered in identifying an equilibrium solution and proving whether the solution is unique.

There is also an increasing body of literature within the marketing/operations interface that addresses the strategic role of production lead-time and its effects on a firm's pricing policies as well as decisions on the lag between when to produce and when to sell. Desai *et al.* [26] discuss that uncertainty about the length of the lead time or about the level of market demand can significantly affect a firm's pricing and production decisions. In a two-period model, they focus on demand uncertainty and production lead time such that the firm has to commit to a specific production quantity before demand is realized. Thus, the firm needs to decide on the production quantities for each period as well as how much to market in the first period, given that customers that purchase the product in the first period can resell it in the second period. This is also a potential area that can be pursued for future research.

REFERENCES

1. Keskinocak P, Tayur S. Due date management policies. Handbook of quantitative supply chain analysis: modeling in the e-business era. Boston (MA): Kluwer Academic Publishers; 2004. pp. 485–554.
2. Kaminsky P, Hochbaum D. Due date quotation models and algorithms. Handbook of scheduling: algorithms, models and performance analysis. Boca Raton (FL): Chapman & Hall/CRC; 2004.
3. Cheng TCE, Gupta MC. Survey of scheduling research involving due date determinations decisions. Euro J Oper Res 1989;38:156–166.
4. So KC, Song J. Price, delivery time guarantees and capacity selection. Euro J Oper Res 1998;111:28–49.
5. Palaka K, Erlebacher S, Kropp DH. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Trans 1998;30:151–163.
6. Ray S, Jewkes EM. Customer lead time management when both demand and price are lead time sensitive. Eur J Oper Res 2004;153(3): 769–781.
7. Boyaci T, Ray S. Product differentiation and capacity cost interaction in time and price sensitive markets. Manuf Serv Oper Manag 2003;5(1):18–36.
8. Hassin R, Haviv M. To queue or not to queue: equilibrium behavior in queuing systems. Boston (MA): Kluwer Academic Publishers; 2003.
9. Easton FF, Moodie DR. Pricing and lead-time decisions for make-to-order firms with contingent orders. Eur J Oper Res 1999;116: 305–318.
10. Watanapa B, Techanitisawad A. Simultaneous price and due date settings for multiple customer classes. Eur J Oper Res 2005;166: 351–368.
11. Plambeck EL. Optimal lead-time differentiation via diffusion approximations. Oper Res 2004;52(2):213–228.
12. Celik S, Maglaras C. Dynamic pricing and lead time quotation for a multi-class make-to-order queue. Manage Sci 2008;54(6):1132–1146.
13. Ata B, Olsen TL. Near-optimal dynamic lead time quotation and scheduling under convex–concave customer delay costs. Oper Res 2009;57(3):753–768.
14. Akan M, Ata B, Olsen TL. Congestion-based lead time quotation and pricing for revenue maximization with heterogeneous customers. Carnegie Mellon University, Pittsburgh, PA: Tepper School of Business; 2008. Working paper.
15. Liu L, Parlar M, Zhu SX. Pricing and lead time decisions in decentralized supply chains. Manag Sci 2007;53(5):713–725.

16. Pekkün P, Griffin PM, Keskinocak P. Coordination of marketing and production for price and lead time decisions. *IIE Trans* 2008;40(1): 12–30.
17. Loch C. Pricing in markets sensitive to delay. PhD thesis, Stanford, CA: Stanford University; 1991.
18. Armony M, Haviv M. Price and delay competition between two service providers. *Eur J Oper Res* 2003;147(1):32–50.
19. Chen H, Wan Y. Price competition of make-to-order firms. *IIE Trans* 2003;35:817–832.
20. Lederer PJ, Li L. Pricing, production, scheduling and delivery-time competition. *Oper Res* 1997;45(3):407–420.
21. Cachon PC, Harker PT. Competition and outsourcing with scale economies. *Manag Sci* 2002;48(10):1314–1333.
22. Allon G, Federgruen A. Competition in service industries. *Oper Res* 2007;55(1):37–55.
23. Allon G, Federgruen A. Competition in service industries with segmented markets. *Manage Sci* 2009;55(4):619–634.
24. So KC. Price and time competition for service delivery. *Manuf Serv Oper Manag* 2000;2(4): 392–409.
25. Pekkün P, Griffin PM, Keskinocak P. Centralized vs. decentralized competition for price and lead-time sensitive demand. Atlanta, GA: H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology; 2008. Working paper.
26. Desai PS, Koenigsberg O, Purohit D. The role of production lead time and demand uncertainty in marketing durable goods. *Manag Sci* 2007;53(1):150–158.

PRICING AND REPLENISHMENT DECISIONS

FERNANDO BERNSTEIN
Fuqua School of Business,
Duke University, Durham,
North Carolina

INTRODUCTION

An important problem in the interface between marketing and operations is the simultaneous determination of pricing and inventory replenishment decisions. Traditional inventory models focus on effective replenishment strategies and generally assume that demand for a product is independent of the product's price. In these models, the demand process and all revenue streams are exogenously determined. Marketing models, on the other hand, focus on pricing strategies and their impact on sales and revenues but, for the most part, overlook the firms' inventory decisions.

In recent years, a number of papers have focused on the integration of pricing decisions and inventory replenishment strategies. These decisions have traditionally been made by separate divisions within the organization. However, firms are increasingly pursuing more integrative approaches to manage their inventory and pricing strategies.

This article reviews some of the most significant results regarding the joint determination of pricing and inventory replenishment decisions. There is a vast literature addressing dynamic pricing issues in settings where selling must stop after a deadline and there is no inventory replenishment throughout the selling season. This includes papers focusing on clearance and markdown pricing. Many of these papers have applications to revenue management problems, particularly those pertinent to retail operations. We refer to the section titled "Pricing and Revenue Management" in this encyclopedia and

the book by Talluri and Van Ryzin [1] for extensive reviews of revenue management models. With the exception of single-period models, in which there is a one-time pricing and inventory decision, our focus is on situations in which inventory can be replenished frequently throughout the time horizon.

The papers discussed in this article analyze pricing and inventory decisions for both the case of a single firm and the case of multiple firms. In settings with multiple firms, the article reviews models in which firms operate in nonoverlapping markets (so each retail firm is a monopolist in its own market) and models in which firms compete for demand in the downstream market. The article is organized according to this distinction. The section titled "Single Firm" reviews work related to decisions made within an organization (i.e., the case of a single firm). The next section looks at broader supply-chain management settings in which multiple independent firms make decisions regarding their prices and replenishment strategies. The final section concludes the article. The review in this article is by no means exhaustive. Chan *et al.* [2] and Yano and Gilbert [3] provide excellent survey chapters reviewing models that address the coordination of pricing policies with inventory management and production decisions. In addition, the survey by Chan *et al.* provides a comprehensive taxonomy for work in this field. The article by Alon *Pricing and Scheduling Decisions*, [4] provides a review on work related to pricing and scheduling decisions.

SINGLE FIRM

This section reviews single-period, finite-horizon, and infinite-horizon inventory models in which demand is endogenously determined by a price p . In all these settings, a firm stocks a product in anticipation of uncertain demand. At the same time, the firm can affect demand by determining the

selling price. Joint pricing and replenishment models generally assume that the price-sensitive stochastic demand is of the multiplicative form or of the additive form, while demand randomness is independent of the selling price. In its more general form, demand is given by

$$D(p, \epsilon) = \gamma(p)\epsilon + \delta(p), \quad (1)$$

where ϵ is a random variable independent of price, and $\gamma(p)$ and $\delta(p)$ are nonincreasing functions of p . When $\delta(p) = 0$, demand is of the multiplicative form, and when $\gamma(p) = 1$, demand is of the additive form. In addition to the selling price, the firm decides on a stocking quantity q (or order-up-to level y , so that the order quantity is given by $y - x$ if x is the initial inventory at the beginning of the period in a multiperiod setting).

In a single-period setting, the firm makes a stocking decision for a product that becomes obsolete at the end of the selling period. Any demand in excess of the stocking quantity is lost. [The problem of determining an order quantity in a single-period setting with random demand is sometimes called the *newsvendor problem*. For a review on the newsvendor model, we refer to the article by Petruzzi *Newsvendor Models*, [5].] In addition, the firm decides a selling price for the product over this selling period. The one-period expected profit function is given by $\Pi(q, p) = pED(p, \epsilon) - cq - G(q, p)$, where c is the per unit purchasing or production cost, $pED(p, \epsilon)$ is the expected revenue assuming that all demand is satisfied, and $G(q, p) = h^+E[q - D(p, \epsilon)]^+ + (p + h^-)E[D(p, \epsilon) - q]^+$ is the expected inventory cost function that depends on the order quantity q and on the price p . (Note that $[z]^+ = \max\{z, 0\}$.) Here, h^+ represents the cost of disposing of any left-over inventory at the end of the period and h^- is the penalty cost associated with each unit of unsatisfied demand.

Whitin [6] was the first to consider the price-setting newsvendor problem. In Whitin's model, the probability distribution of demand depends on the selling price, and the price is itself a decision variable. Petruzzi and Dada [7] provide a comprehensive review of the single-period problem in

which a stocking quantity and selling price are set simultaneously. The paper examines the cases of additive demand and multiplicative demand. In both cases, the paper shows properties of the expected profit function that facilitate the optimization of the price and order-quantity variables. The cases of additive and multiplicative demand differ in how the price affects demand variability. Under additive demand, the variance of demand is independent of the selling price, while the coefficient of variation of demand (standard deviation/mean) increases in the price. On the other hand, in the case of multiplicative demand, the coefficient of variation is independent of the price, but the variance is a decreasing function of price. This distinction affects the way in which the selling price is used to mitigate the risk of demand uncertainty, that is, the risks of overstocking (ending the selling period with excess inventory) or understocking (causing lost sales). In the case of additive demand, it is possible to reduce the coefficient of variation of demand by selecting a lower price. This does not affect demand variance. When demand is of the multiplicative form, a firm can decrease demand variance, without affecting the coefficient of variation of demand, by selecting a higher price. Therefore, the paper concludes that a pricing strategy aimed at reducing supply and demand mismatches must depend on the way that price affects the demand function.

Initial work on the simultaneous choice of prices and replenishment strategies in multiperiod settings was confined to the case of price-dependent deterministic demand. In a multiperiod stochastic setting, Thowsen [8] studies the problem of integrating pricing and replenishment decisions under specific forms of the demand function. The paper proposes a heuristic model for a setting without fixed ordering costs. Thomas [9] considers a periodic review, finite-horizon model with fixed ordering costs and stochastic, price-dependent demand. The paper proposes a policy, referred to as (s, S, p) , under which inventory is replenished according to an (s, S) policy and the price depends on the initial inventory level at the beginning of the period.

An (s, S) policy stipulates that if the inventory position at the beginning of the period is below a reorder point s , then an order is placed to raise the inventory position to the order-up-to level S .

Federgruen and Heching [10] address simultaneous pricing and replenishment decisions in periodic review models (both finite and infinite horizons) with stochastic, price-dependent demand. The paper characterizes the structure of the optimal dynamic pricing and inventory strategy for a general stochastic demand function as in Equation (1). In their model, there are no fixed costs for placing orders and all unsatisfied demands are backlogged. One form of the one-period expected inventory and backlogging cost considered in the paper is given by $G(y, p) = h^+ E[y - D(p, \epsilon)]^+ + h^- E[D(p, \epsilon) - y]^+$, with y the order-up-to level, h^+ the per unit holding cost rate, and h^- the per unit backlogging penalty cost rate. The optimal dynamic pricing and inventory policy is a so-called base stock list price policy, characterized by a base stock level y_t^* and list price p_t^* for each period t . If the initial inventory level in period t , x_t , is below the base stock level, then the firm places an order for the difference $y_t^* - x_t$ and charges the list price. Otherwise, the firm places no order in that period and charges a discounted price to increase sales. This result is extended to the infinite-horizon setting with stationary parameters. The paper reports the results of a numerical study that suggests that there are significant benefits to using a dynamic pricing policy, as opposed to a fixed price policy. It is also noted that an increase in the coefficient of variation leads to higher optimal prices. This finding is consistent with the observations regarding the role of the pricing strategy in a single-period setting.

The work by Federgruen and Heching was extended to a setting in which there are both, a fixed cost for placing an order and a variable cost that is proportional to the amount ordered. Chen and Simchi-Levi [11] identify a dynamic inventory policy and pricing strategy that maximizes the expected profit over a finite horizon. As in Federgruen and Heching, this model assumes that unsatisfied demand

in a period is backlogged. The paper also considers demand functions as in Equation (1), with general forms of $\gamma(p)$ and $\delta(p)$. In the case of additive demand, the paper shows that the (s, S, p) policy identified in [9] is indeed optimal. The paper provides separate treatment for the case of multiplicative demand, for which an (s, S, p) policy may fail to be optimal. For the general demand model, the paper shows that the optimal policy is of the form (s, S, A, p) . In each period t , A_t is a (possibly empty) set of initial values of the inventory level. Under this policy, the firm places an order if the inventory level at the beginning of a period t is below s_t or if it is in the set A_t . In either case, an order is placed to raise the inventory level up to the target value S_t . In those cases, the firm charges a list price p_t^* . For all other initial inventory levels, no order is placed and the optimal price depends on the initial inventory level.

Huh and Janakiraman [13] use an alternative approach to prove the optimality of the (s, S, p) policy in a setting with additive demand and fixed ordering cost, both under the backordering and lost sales assumptions. This approach, based on a property related to the single-period expected profit function, can be used to show that the (s, S, p) policy is optimal under more general forms of the demand function. Song *et al.* [12] study the case of multiplicative demand functions, also in the context of lost sales and in settings in which the firm incurs a fixed cost for placing orders. The paper establishes the optimality of the (s, S, A, p) policy in this setting.

Deng and Yano [14] consider the problem of making pricing and production quantity decisions in a finite-horizon setting for a firm facing capacity constraints. The paper assumes that both variable and fixed setup costs are present in the model. Demand is a deterministic downward sloping function of price. The paper characterizes the optimal joint pricing and replenishment policy and shows that optimal prices may increase as the capacity increases.

A number of papers study the problem of jointly determining prices and inventory replenishment decisions for a firm selling multiple products. In a single-period newsvendor setting, Aydin and Porteus [15]

investigate the optimal prices and inventory levels for multiple products. The products' prices affect the overall size of the market and the way in which demand is allocated among products. Under certain classes of demand functions, the paper shows that there is a unique vector of prices and inventory levels that satisfy the first-order optimality conditions. Zhu and Thonemann [16] consider a firm selling two products over a finite horizon. The demand for each product depends on both products' prices. Demands are of the additive form with linear cross-price effects. The paper shows that a base stock list price policy is no longer optimal in this setting. Instead, they show how prices and order quantities can be computed optimally. The paper shows that both the products' prices and the replenishment decisions depend on the initial inventory levels of the two products.

MULTIPLE FIRMS

This section describes research addressing joint pricing and replenishment decisions in settings with multiple firms. These models generally consider two-echelon supply chains. In those settings, a manufacturer or supplier sells a product to one or multiple retail firms that, in turn, offer this product to the market. The retail firms may be geographically dispersed, in which case they do not directly compete in the downstream market, or they may engage in price or quantity competition. In either case, a firm's demand function decreases in its own price. In a competitive setting, a firm's demand increases with increases in the prices of its competitors.

There are two primary decision making paradigms that are analyzed and compared in the multiple-firm setting. Under centralized control, all pricing and replenishment decisions are made by a central planner who has control over the decisions made by all the firms in the supply chain. In contrast, in a decentralized system, each firm makes the pricing and inventory decisions that optimize its own performance, irrespective of the impact that these decisions may have on the

other firms in the supply chain. Competition among firms is inherent to decentralized supply-chain settings. This competitive interaction is modeled and analyzed using tools of Game Theory. We refer to the section titled "Foundations and Principles of Game Theory" in this encyclopedia and the book by Fudenberg and Tirole [17] for comprehensive treatments of this area of study.

In addition to the distinction between centralized and decentralized control, the papers in this section differ in terms of the type of demand functions they consider (e.g., deterministic or stochastic) and in terms of the relevant operational costs present in the system. The various assumptions about the decision making paradigm, the demand functions, and the cost structure, lead to very different dynamics regarding the firms' pricing and inventory replenishment decisions. In particular, in settings with multiple competing retailers, the analysis of equilibrium pricing and replenishment strategies depends on the specific forms of the underlying demand and cost functions. Most papers investigating decentralized supply chains also explore the design of coordinating contracts. These contractual agreements induce firms in the decentralized supply chain to adopt pricing and inventory replenishment strategies equal to those that would be adopted under centralized control, thus leading to higher system-wide profit. We refer to the article by Zhang *Supply Chain Coordination*, [18] for a review of issues related to supply-chain coordination.

Single or Multiple Geographically Dispersed Retailers

The paper by Zhao and Wang [19] studies joint pricing and production decisions in a supply chain consisting of a manufacturer distributing its product through a single retailer or distributor. This firm faces a price-dependent deterministic demand. The paper considers a finite-horizon setting and in each period, the distributor chooses its price and order quantity, while the manufacturer determines a wholesale price and its own replenishment quantity. The paper studies

the setting under decentralized decision making and proposes a price schedule that induces the distributor to adopt pricing and production strategies equal to those that would arise under centralized control. The paper by Granot and Yin [20] also studies a manufacturer selling a product to a retailer in a newsvendor setting with multiplicative, price-dependent demand. The manufacturer offers a contract to the retailer, and this firm determines the selling price and order quantity. The paper examines how the timing of the pricing and order-quantity decisions, with respect to the realization of random demand, affects the firms' profit performances. The paper by Ray *et al.* [21] studies a two-echelon system consisting of a distributor selling products through a retailer that faces stochastic price-sensitive demand. The retailer follows a continuous review base stock policy and faces a variable delivery lead time. The paper studies and compares optimal pricing and inventory decisions in a centralized setting and in a decentralized setting under a wholesale price contract.

Chen *et al.* [22] consider a two-echelon supply chain consisting of a supplier distributing a product through multiple noncompeting retail firms. Demand for each retailer occurs continuously at a constant rate, which decreases in the retailer's price. The retailers incur holding costs for the inventory they carry and fixed costs for the orders placed with the supplier. The supplier also incurs fixed and variable costs. The paper develops an efficient algorithm to determine the optimal pricing and replenishment strategy for the centralized system. The paper also considers a decentralized system in which the supplier and the retailers are independent profit-maximizing firms. In this setting, the supplier acts as the Stackelberg leader, offering the retailers a wholesale price that maximizes its profit, in anticipation of the subsequent joint pricing and replenishment strategy selected by the retailers. Chen *et al.* [23] identify a wholesale pricing mechanism that ensures that the retailers make their decisions in the decentralized system equal to those that would be made under centralized control.

Multiple Competing Retailers

Bernstein and Federgruen [24] consider a competing newsvendor setting in which multiple retail firms make pricing and order-quantity decisions. The distribution of a retailer's random demand depends on its own price as well as on those of the other retailers according to a general stochastic demand function. The paper analyzes the decentralized setting and provides conditions that ensure the existence of a unique price and order-quantity equilibrium. The paper then proceeds to the design of a coordinating contract arrangement that, if adopted, induces retailers to select the centralized optimal prices and order quantities. The paper by Zhao and Atkins [25] studies a setting with competitive newsvendors, in which firms compete simultaneously in terms of their prices and inventory levels. In their setting, a retailer's demand is a function of all firms' retail prices and stocking levels. Specifically, if demand at a particular retailer exceeds its available inventory, then a portion of the unsatisfied demand spills over to the retailer's competitors. The paper shows properties of each retailer's profit function that ensure the existence of a unique equilibrium.

In an infinite-horizon setting, Bernstein and Federgruen [26] study a two-echelon supply chain consisting of a supplier distributing a product through a network of competing retailers. In this setting, all firms incur inventory-carrying costs, while all supplier orders and transfers to the retailers incur fixed and variable costs. Like in Ref. 22, each retailer's demand occurs at a constant (deterministic) rate, over an infinite horizon. However, in this setting, a retailer's demand rate depends on the price charged by this retailer as well as those charged by its competitors. Alternatively, the price each retailer charges may depend on the sales volumes targeted by all the retailers, leading to quantity competition. The paper first characterizes the solution to the centralized system in which the joint pricing and replenishment decisions are determined to maximize chain-wide profit. This is followed by the analysis of the decentralized system, in which the

supplier acts as the Stackelberg leader offering a wholesale pricing scheme, followed by the retailers choosing their policy variables. The paper presents a condition that relates each retailer's price elasticity with the ratio of its annual sales to its combined inventory and fixed setup costs, that guarantees the existence of a unique Nash equilibrium. Finally, the paper proposes a wholesale pricing scheme that induces coordination.

In contrast to the deterministic demand setting explored in [26], Bernstein and Federgruen [27] explore a periodic review, infinite-horizon model in which in every period, each retailer faces a random demand volume. The distribution of a retailer's demand depends on all firms' retail prices. The paper considers the case of multiplicative demands and the more general case of nonmultiplicative demand structures, where the coefficient of variation of demand of each retailer may depend on its mean demand and, consequently, on all the prices in the market. The paper characterizes structural properties that guarantee the existence of an equilibrium of infinite-horizon joint pricing and replenishment strategies. It is shown that under retailer competition, and for the case of nonmultiplicative demand structures, a retailer's expected profit may decrease when reducing the uncertainty in its demand process.

Work on joint pricing and order quantity or production decisions also extends to assembly systems. In those settings, a number of firms (or suppliers) sell complementary products to a buyer (or assembler), who in turn, sells these items in a bundle (i.e., an assembled product). Wang [28] considers pricing and order-quantity decisions for a set of manufacturers that produce complementary products. Demand for the products (which are always purchased and consumed jointly) is random and occurs during a single selling season. The paper assumes a multiplicative demand model. The pricing and production decisions are made either simultaneously or sequentially. The paper characterizes and compares the firms' pricing and production equilibrium decisions and their resulting performance, under simultaneous and sequential decision making.

CONCLUSIONS

This article provides an overview of work addressing the joint determination of prices and inventory replenishment policies. The article reviews papers that focus on settings with a single firm and papers that consider more general settings with multiple firms. In settings with multiple firms, the article provides separate treatment for those papers that consider competition in the downstream market and those that assume that firms are geographically dispersed. Some of the papers reviewed in this article consider dynamic inventory and pricing decisions, while others restrict attention to static decisions. The case of dynamic joint pricing and replenishment decisions in settings with multiple competing firms represents a challenging but interesting direction for future research.

REFERENCES

1. Talluri KT, van Ryzin GJ. The theory and practice of revenue management. Norwell (MA): Kluwer Academic Publishers; 2004.
2. Chan L, Shen Z, Simchi-Levi D, *et al.* Coordination of pricing and inventory decisions: a survey and classification. In: Simchi-Levi D, Wu S, Shen Z, editors. Handbook of quantitative supply chain analysis: modeling in the E-business era. Dordrecht: Kluwer; 2004. pp. 335–392.
3. Yano CA, Gilbert SM. Coordinated pricing and production/procurement decisions: a review. In: Chakravarty A, Eliashberg J, editors. Volume 16, Managing business interfaces. Boston (MA): Kluwer; 2004. pp. 65–103.
4. Alon G. Pricing and scheduling decisions. Wiley Encyclopedia of Operations Research and Management Science; Hoboken, NJ: In press.
5. Petruzzi NC. Newsvendor models. Wiley Encyclopedia of Operations Research and Management Science; Hoboken, NJ: In press.
6. Whitin TM. Inventory control and price theory. *Manage Sci* 1955;2(1):61–68.
7. Petruzzi NC, Dada M. Pricing and the newsvendor problem: a review with extensions. *Oper Res* 1999;47(2):183–194.
8. Thowsen GT. A dynamic, nonstationary inventory problem for a price/quantity setting firm. *Nav Res Logist* 1975;22:461–476.

9. Thomas LJ. Price and production decisions with random demand. *Oper Res* 1974;22:513–518.
10. Federgruen A, Heching A. Combined pricing and inventory control under uncertainty. *Oper Res* 1999;47(3):454–475.
11. Chen X, Simchi-Levi D. Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: the finite horizon case. *Oper Res* 2004;52(6):887–896.
12. Song Y, Ray S, Boyaci T. Optimal dynamic joint pricing and inventory control for multiplicative demand with fixed order costs and lost sales. *Oper Res* 2008. In press.
13. Huh WT, Janakiraman G. optimality in joint inventory-pricing control: an alternate approach. *Oper Res* 2008;56(3):783–790.
14. Deng S, Yano CA. Joint production and pricing decisions with setup costs and capacity constraints. *Manage Sci* 2006;52(5):741–756.
15. Aydin G, Porteus EL. Joint inventory and pricing decisions for an assortment. *Oper Res* 2008;56(5):1247–1255.
16. Zhu K, Thonemann UW. Coordination of pricing and inventory across products. 2002. Working Paper. Stanford University.
17. Fudenberg D, Tirole J. *Game theory*. Boston, MA: The MIT Press; 1991.
18. Zhang F. *Supply chain coordination*. Wiley Encyclopedia of Operations Research and Management Science, Hoboken, NJ: In press.
19. Zhao W, Wang Y. Coordination of joint pricing-production decisions in a supply chain. *IIE Trans* 2002;34:701–715.
20. Granot D, Yin S. Price and order postponement in a decentralized newsvendor model with multiplicative and price-dependent demand. *Oper Res* 2007;56(1):121–139.
21. Ray S, Li S, Song Y. Tailored supply chain decision making under price-sensitive stochastic demand and delivery uncertainty. *Manage Sci* 2005;51(12):1873–1891.
22. Chen F, Federgruen A, Zheng YS. Near-optimal pricing and replenishment strategies for a retail/distribution system. *Oper Res* 2001;49(6):839–853.
23. Chen F, Federgruen A, Zheng YS. Coordination mechanisms for a distribution system with one supplier and multiple retailers. *Manage Sci* 2001;47(5):693–708.
24. Bernstein F, Federgruen A. Decentralized supply chains with competing retailers under uncertainty. *Manage Sci* 2005;51(1):18–29.
25. Zhao X, Atkins DR. Newsvendors under simultaneous price and inventory competition. *Manuf Serv Oper Manage* 2008;10(3):539–546.
26. Bernstein F, Federgruen A. Pricing and replenishment strategies in a distribution system with competing retailers. *Oper Res* 2003;51(3):409–426.
27. Bernstein F, Federgruen A. Dynamic inventory and pricing models for competing retailers. *Nav Res Logist* 2004;51(2):258–274.
28. Wang Y. Joint pricing-production decisions in supply chains of complementary products with uncertain demand. *Oper Res* 2006;54(6):1110–1127.

PRICING AND SCHEDULING DECISIONS

GAD ALLON
Kellogg School of Management,
Northwestern University,
Evanston, Illinois

Service providers as well as make-to-order manufacturers use priority schemes in conjunction with pricing as allocation mechanisms. By carefully selecting the appropriate scheduling scheme and the optimal prices, service providers can best utilize their resources to improve their customers' social welfare or improve their overall value (through improved service quality or cost reduction). This article studies the interplay between pricing and scheduling decisions focusing on the interaction between the service provider and the delay-sensitive customers, while dealing with different types of providers (a social planner, a monopolist, or oligopolies) in different types of service settings (either when the queues are observable or when they are not), and different types of customer populations (homogeneous or heterogeneous, segmented or unsegmented).

In exploring the relationship between pricing and scheduling decisions, we give separate treatment to different settings according to the role of service providers in the economy:

- *Social Planner.* In this case, the capacity of the service process is viewed as a scarce resource, and thus granting priorities is treated as such. The role of the pricing of the service is usually viewed in this context as a transfer mechanism to guarantee that only those who need high priority (due to high cost of delay or short service time) are granted this priority. In particular, pricing is used to ensure that the pricing-priority mechanism is incentive-compatible in the sense that customers have no incentive to misrepresent their true identity

when they hold such information privately (that is to claim that their delay cost is higher or service time is shorter). The pricing may also play a role in regulating usage in the system as a whole; this is a common practice in managing congested systems.

- *Monopolistic View.* In such a setting, the firm would like to use priorities and price them so as to improve its profits by virtue of price differentiation. When facing different types of customers, with different sensitivities to delay and different values obtained from the service, serving customers according to a priority rule while charging a priority-dependent price can outperform charging a fixed price, and serving according to a first-come-first-served manner. In settings in which there is asymmetric information regarding the identity of the customers, the prices and the priority schemes should be selected so that the overall pricing-priority mechanism is incentive-compatible.
- *Competitive View.* When competing in a market in which customers are sensitive to delay, firms can choose their prices, waiting time standards, capacity levels, and priority schemes. The role of the priority scheme is to allow the firm to accommodate all demand streams (driven by both the prices and the service level of the different segments), while meeting the promised service levels. As we will show, the price charged to each segment by each firm depends on the characteristics of all firms and also on the level of priority that the firm allocates to this specific segment.

We will initially review the use of priorities in models, where a social planner sets the prices and the scheduling rules. The main focus of these models will be on the incentive compatibility of the joint pricing-scheduling decisions. We will then survey settings where

profit-maximizing firms use the combination of pricing and scheduling, and which we will use to highlight the key differences between a social planner and a monopolist. We will initially discuss a setting in which the firm is not informed about the customer's waiting time cost and value, giving rise to incentive-compatible pricing and scheduling. We will also discuss the setting in which the queues are observable, giving rise to the phenomenon of "following the crowd" (which we will describe in detail). We will complete the article with a discussion of the role of priorities and pricing in the context of a competitive market. In this context, we will also discuss how pricing and scheduling decisions affect the benefits of each segment depending on the level of priority given to it. Note that the article deals with the interplay between pricing and scheduling in settings, which are modeled using queueing systems; that is, service settings and make-to-order manufacturers, and does not address make-to-stock settings.

These decisions have been considered (jointly) also in a make-to-stock environment. For examples of such models, see Charnsirisakskul *et al.* [1]. Also, for a more complete discussion of possible scheduling rules in queueing, see the article titled **Power Indices**. For a comprehensive discussion of strategic customer behavior in queueing, see the article titled **Strategic Customer Behavior in a Single Server Queue**. Lastly, for a discussion of the interplay between pricing and lead-time decisions, see the article titled **Pricing and Lead-Time Decisions**.

SOCIAL WELFARE MAXIMIZATION: INCENTIVE COMPATIBILITY

In many service systems, different customers typically have rather disparate sensitivities to the delay encountered, and obtain different value from the service. In order to improve the social welfare, a social planner can exploit these differences by optimally prioritizing among the different customers while charging them the optimal prices. The social planner, who maximizes the social welfare

of the players, views its limited available capacity as a scarce resource that needs to be allocated. Using the same logic, high priority is a scarce resource that needs to be properly allocated to those who need it. However, the design of optimal price and scheduling mechanisms may require knowledge of characteristics of the customers. These characteristics are seldom known to the queue manager and must be estimated or obtained from the customers' own statements, a situation that may create an incentive for customers to declare untruthful values.

The role of pricing, in many service systems, is to modulate demand. This is the case in particular for service providers striving to maximize social welfare, where customers are asked to pay for their use in order to make sure that the optimal usage level is achieved. Naor [2], for example, showed that in the context of homogeneous customers, tolls have to be levied in order to make sure that customers do not join the system when it is too congested. When customers are heterogeneous in terms of their waiting time cost and service times, the role of pricing is not only to regulate the usage in the different priority levels, (where, within each customer class, the optimal level of usage is achieved) but also to guarantee that customers either disclose their true costs or use the priority level that was designed to their needs and do not exploit the information asymmetry to obtain a higher priority than the one they deserve.

The question that this section tries to answer is how to construct a pricing mechanism that induces all customers to make decisions, which are not only optimal from their myopic, self-interested point of view, but also optimize the overall objective function of the organization.

Pricing Based on Externalities

Consider a model in which customers are identical except for having different costs of waiting, and priority levels are assigned to customers according to their declared costs. The information that is available to an arriving customer includes the queue length upon his arrival, the declared time values of the customers in the queue, and

the residual service time of the customer in service. Using this information, the customer declares a value to his time and obtains priority accordingly. The goal is to induce customers to declare their true time value, so that the resulting order of service optimizes social welfare. Dolan [3] suggested the use of Clarke prices. The idea is that each customer pays for the costs that he imposes on others when joining the queue. These costs are calculated given that customers reveal their true time value, which means that a customer pays an amount that is equal to the externalities that he imposes on other customers. It follows that under this pricing rule, the strategy that prescribes declaring the true time value is an equilibrium one. If a customer overstates his time cost, he must compensate the customers, who will have to wait longer as a result of this deviation by exactly the cost he saved (due to the work-conservation principle). Since these customers have higher waiting costs than his, compensating these customers for their increased waiting time costs the customer more than what he saves.

The model of Dolan [3] focuses on the ability to use a bidding rule, which is based on externalities, geared to induce customers to reveal their true delay cost. However, the model remains silent regarding the ability to use the pricing and the priority schemes to regulate usage and to control the arrival rate to the system, either to maximize profits or social welfare. Mendelson and Whang [4] address settings in which the arrival rate to the system is a function of the prices and the delay that the customers experience in the system.

Incentive Compatibility of the $c\mu$ Priority Rule

We now turn to the more general case, in which the service provider has to use pricing to regulate usage while being incentive-compatible. Consider a system that is modeled as an M/M/1 queueing system with multiple classes. Each class is characterized by its delay cost per unit of time and the expected service requirement. The goal is to find a pricing mechanism that is also optimal; that is, that the arrival rates and the scheduling scheme jointly maximize the

expected value of the system while allowing each customer to make individual decisions on whether or not to purchase the service and at what priority level. In making these decisions, while the customers are assumed to maximize their own utility functions under the optimal incentive-compatible pricing and scheduling schemes, they also maximize the objective of the organization as a whole.

When minimizing the expected organization-wide delay cost per unit of time, it is well known that given an exogenously specified arrival rate, and given that the delay cost and the expected service requirements are known, the optimal scheduling scheme is the so-called $c\mu$ rule, which gives absolute priority to customers according to the ratio of their delay cost to the expected service time required to complete their request. It is clear from the definition of the priority scheme that knowledge of these parameters is crucial for successful execution.

Mendelson and Whang [4] address the problem of resource pricing, which determines the demand rate of customers from each class, and priority pricing, which determines which customers are served at each point in time. For the case in which all customers have the same mean service times, regardless of their class, the authors show, in an analogous manner to Dolan [3], that charging each customer the externalities that he imposes on other customer classes under the optimal demand level, while using the $c\mu$ rule to prioritize among customers, is both optimal and incentive-compatible. In particular, the authors show that the optimal joint pricing-priority scheme automatically achieves incentive compatibility.

In the case of identical mean service times, the authors show that a price that depends only on the priority level is optimal and incentive-compatible. However, for the more general case, where customers differ in their delay cost as well as in their mean service time, such a pricing scheme may not be incentive-compatible anymore. In that case, the authors suggest a more general pricing scheme, which can be decomposed into two parts: a basic charge and a priority surcharge. The basic charge corresponds to the price of the lowest priority class, is equal

for all customers (regardless of their class or the priority they choose), and is quadratic in the service time. The priority charge is proportional to the service time with a coefficient that increases as the priority level goes up. One may view this pricing scheme as “externality pricing” in the following sense: when a customer buys a priority level, he pays the conditional expected externality of his decision on the other customers.

The main finding in this article is that the incentive-compatible pricing scheme induces the customers to behave in a way that is consistent with this rule, while aligning their interest with that of the overall organization. For a social planner, the prices are treated as transfer payments, whose objective is to merely be an instrument to regulate usage (both in the system as a whole and in each priority level). We next turn to ask how the pricing scheme and the scheduling decisions will change in a monopolistic setting.

PROFIT-MAXIMIZING PRICING AND SCHEDULING

Afeche [5] answers the question of how a capacity-constrained firm should design an incentive-compatible price-scheduling mechanism to maximize revenues from a heterogeneous pool of time-sensitive customers with private information on their willingness to pay, time-sensitivity, and processing requirements.

The article provides the following insights: First, the author shows that the familiar $c\mu$ rule, which is known to minimize the expected delay cost and to be incentive-compatible under social welfare optimization, need not be optimal in this setting. This specific fact suggests a more general guideline: in designing incentive-compatible and revenue-maximizing scheduling policies, delay cost-minimization, which plays a prominent role in controlling and pricing queueing systems, should not be a dominant criterion. Second, the author identifies optimal scheduling policies with novel features. One such policy prioritizes the more time-sensitive customers but voluntarily delays the completed orders of low

priority customers. This insertion of *strategic delays* deters time-sensitive customers from purchasing the lower-priority class. In other situations, the author shows that it is optimal to appropriately randomize priority assignment, in one extreme case, serving customers in reverse $c\mu$ order, which maximizes the system delay cost among all work-conserving policies. The author also shows that the optimal level of delay differentiation systematically emerges from a trade-off between operational constraints and customers’ incentives.

Numerical Example

In order to demonstrate the role of intentional delays, we will use a simple numerical example. Consider a firm that serves two types of customers:

1. H-type customers obtain \$100 from the service and incur \$20 per unit of time spent in the system.
2. L-type customers obtain \$30 from the service and incur \$4 per unit of time spent in the system.

We will assume that the arrival rates from each class is 0.2 H-type customers per unit of time, and 0.3 L-type customers per unit of time, independent of the price charged. We also assume that $\mu = 1$.

If the firm employs a $c\mu$ rule, the sojourn time of an H-type customer is $1/(1 - 0.2) = 1.25$. Using a similar logic, an L-type customer’s waiting time under the $c\mu$ rule is 2.5 time units.

If the firm tries to extract the entire consumer surplus, it would charge $\$75 = 100 - (1.25 \times 20)$ from each H-type customer, and $\$20 = 30 - (2.5 \times 4)$ from each L-type customer. This can presumably result in a revenue rate of $(0.2 \times 75) + (0.3 \times 20) = \21 per unit of time.

However, this solution is not incentive-compatible: Under this price and under this priority scheme, an H-type customer would opt to purchase the service grade designed for an L-type customer; this will leave him with utility of $100 - (2.5 \times 20) - 20 = 30$, which is better than what he obtains by purchasing

the service grade “designed” for him, which is $100 - (1.25 \times 20) = 75$.

Under the $c\mu$ rule, the highest incentive-compatible price that the firm can charge an H-type customer is \$45, which results approximately in a revenue rate of \$15 per unit of time. However, if the firm chooses to inject one unit of time of intentional delays for all L-type customers, the firm can charge only \$16 from each L-type customer, but the incentive-compatible price for the H-type customers is then \$61, which results in an increased revenue rate of $(3.2 \times 0.2) - (1.2 \times 0.3) > 0$. This means that the firm can improve its revenues by inserting artificial unnecessary delays. By degrading the quality of the low priority service, the service provider can charge more for high priority service. Note that in this case, the idleness is increasing the overall delay cost, yet the provider shrinks the pie in order to be able to “eat a larger piece.”

Bidding for Priorities in Observable Systems

We now turn to a discussion of the role of priority in systems in which the queue is observable to an arriving customer. We will first discuss several models that focus on the consumer behavior in such systems (fixing the priorities and the prices) and then discuss the behavior of a monopolist in such a market.

Consider an observable M/M/1 queue in which arriving customers choose from a discrete set of possible payments. In this model, customers are assigned priority levels according to their payments. All customers arriving to the system must obtain service (i.e., balking is not allowed), and customers are not informed on the payments made by others. Balachandran [6] shows that the equilibrium strategy of all customers is to choose the lowest price that will guarantee joining the head of the queue, which induces a “last-come-first-served” scheduling policy, when implemented by all customers. The rationale behind this equilibrium strategy is that a customer purchases high priority level not only because he can be scheduled ahead of other customers, but also so that others are not scheduled ahead of him. In many service systems, this leads to what Hassin and Haviv [7] refer to as “follow-the-crowd” behavior in which the

more customers purchase priority, the more inclined an individual is to do so himself. Haviv and Hassin [8] study a model with multiple priority classes in which agents can decide on joining strategies. While in a single class model, customers should follow a threshold policy; that is, join as long as the queue is below a certain level and balk otherwise. They showed that this is not true in a multiclass case.

The above models fix the pricing scheme and focus on the equilibrium behavior among customers in making the decision when to pay for priority, or how much to pay given the set of possible prices and priorities and in that, they are more descriptive of customer behavior in the presence of priorities.

The question that arises is how a monopolist utilizes such a consumer behavior when deciding what priority levels to offer the customers to choose from.

Alpersteins [9] considers a similar model in which the decisions on the available priorities and the associated prices are made by a profit-maximizing firm. Alperstein showed that the profit increases with the number of priority levels. The main conclusion from this model is that the profit-maximizing pricing scheme is one that induces a threshold of one for each priority type except for the highest one. That is, an arriving customer would always choose the lowest possible priority level that is above any existing customer.

COMPETITIVE MODELS

We will next discuss models in which service providers use prices in conjunction with priorities, to compete in the market. Examples of such industries are numerous. Banks and credit card companies segment their customers into regular and VIP or Gold and Platinum customers. Computer software and hardware firms often segment their customers, for example, into home and home office users, small businesses, large businesses and the government, education, and health-care sectors, using an integrated pool of technical support personnel to serve the different customer segments according to a specific priority discipline; each customer

segment is associated with a specific price and waiting time expectation. Finally, overnight delivery services use their planes and trucks to deliver letters, boxes, and cargo, each with different prices and delivery time standards. In many service industries, waiting time standards are used as a primary advertised competitive instrument. For example, most major electronic brokerage firms, (e.g., Ameritrade, Fidelity, and E-trade) all prominently feature the average or median execution speed per transaction, which is monitored by independent firms.

References Loch [10], Lederer and Li [11], and Armony and Haviv [12] as well as Allon and Federgruen [13] are the only published articles we are aware of that have directly addressed competition models in which waiting-time-sensitive customers are segmented into multiple classes. We will briefly survey the first three and then discuss in greater detail the model presented in Allon and Federgruen [13].

Loch [10] considers an industry with M/M/1 service providers and two customer classes, each with a given waiting cost rate and average service time. All customers within a class select the firm, which offers the lowest full price; that is, the sum of the direct price and the waiting cost, where the total demand volume in the class is given by a known function of this full price value. Under quantity competition, the author establishes the existence of a Nash equilibrium under which the customers are prioritized according to the $c\mu$ rule. Lederer and Li [11] generalize this model to allow for an arbitrary number of nonidentical M/G/1 firms and an arbitrary number of customer classes. Assuming the firms engage in price competition, the authors establish the existence of a Nash equilibrium, under which each firm, once again, prioritizes customers according to the $c\mu$ rule. The existence result is based on the assumption that each class' expected waiting time at a given firm is a convex function of all of the firm's demand rates for the different customer classes. Note also that while in Afeche's [5] monopoly model, the incentive-compatible optimal policy cannot, in general, be based on the $c\mu$ priority rule; the $c\mu$ priority rules are part of

the Nash equilibrium in Lederer and Li's [11] perfect price competition model (provided the above convexity assumption is satisfied).

Armony and Haviv [12] analyze a two-stage competition model with two M/M/1 service providers and two customer classes, each, again, acting as a single entity in deciding what fraction of its business to assign to each of the providers. In the first stage, the two providers compete with each other by announcing their service charges. In the second stage, the customer classes compete with their allocation decisions. They show that pure price equilibrium may fail to exist in this two-stage game.

All of the previous models assume that the service organization (or the make-to-order manufacturer) has a fixed capacity and thus the role of the priority rule is to allocate the capacity among the different customers according to their cost of waiting as well as their expected service requirements.

We next present a more general model in which the firm can invest in capacity in order to accommodate the different customers, in conjunction with a priority rule, in a way that will allow the firm to position itself in the market. In such markets, firms cater to multiple customer classes or market segments with the help of shared service facilities or processes, so as to exploit pooling benefits. As in all of the settings studied in the previous section, different customer classes typically have rather disparate sensitivities to the price of service as well as the delays encountered. Thus, from the firm's perspective, it is vital to offer differentiated service charges and levels of service to different customer classes so as to maximize (long run) profits.

Allon and Federgruen [13] propose and analyze a model in which firms select all or part of the following: (i) the prices charged to all customer classes, (ii) the waiting time standards promised to all classes measured in terms of the average waiting time, (iii) the capacity level, and (iv) a priority discipline enabling the firm to meet the promised waiting time standards under the chosen capacity level. While making these decisions in a competitive setting can be quite complex, one can observe that while the price and the waiting time decisions impact all firms in the

industry, the capacity and the priority (given the demand and the waiting time standard) impact only the cost of the firm itself. The authors, thus, first present the capacity problem, in conjunction with the priority setting. The authors then embed them in the competitive setting in which firms compete in terms of their prices.

Capacity Choice and Associated Priority Rules

Consider a service provider, modeled as an M/M/1 queueing facility. Each customer class generates an independent Poisson stream of customers to this service provider at the rate determined by the above-mentioned demand functions. Its service times are i.i.d. with a firm- and class-dependent service rate that is proportional to the firm's capacity level. Each firm incurs a given class-dependent cost per customer, as well as a cost per unit of time, proportional to the adopted capacity level. Each firm attempts to maximize its expected profits.

Allon and Federgruen [13] derive the analytical expression of the capacity level that each firm needs to adopt to accommodate a given vector of demand volumes and waiting time standards under an optimal associated dynamic priority rule. They show that this capacity level is the maximum of a number of closed-form capacity bounds, one for each subset of the customer classes. Interestingly, for arbitrarily specified waiting time standards, the maximum may be achieved for a strict subset of the collection of all classes, the so-called *bottleneck set*, in which case strategic idleness times, that is artificial after-service delays, may be adopted for the so-called *residual* classes outside the bottleneck set. The optimal capacity level is to be complemented with a randomized absolute priority rule.

The authors study three types of competition: price competition (where waiting times are exogenously determined), waiting time competition (when the prices are exogenously determined), and simultaneous price-waiting time competition. They show that in the waiting time and the simultaneous competition models, the bottleneck set is always the full set. However, this is not necessarily true in the price competition model, where

waiting times are not part of the strategic choices. Thus the need for intentional delays may arise only under price competition.

It is important to note that we again see the use of intentional delays, when selecting capacity levels and priority schemes as observed already, in a different setting and for a different purpose in Afèche [14].

Pooled versus Dedicated Capacity

The authors then compare the equilibria that arise when the firms use pooled capacity with those achieved when the firms service each market segment with a dedicated service process; that is, without pooling service resources. In the price competition model, for example, the equilibrium is obtained, both under service pooling and dedicated service facilities, when for each class the relative markup vis-a-vis the marginal cost equals the reciprocal of the demand elasticity. The marginal cost per customer per unit of time always consists of the variable service cost plus the marginal capacity cost (per unit of time). When service is provided with dedicated facilities, the marginal increase in the required capacity equals the expected amount of work per customer of the considered class. Under service pooling, it is zero for residual classes and less (or more) than this benchmark value, depending upon whether the customer class receives worse (or better) than average service.

Allon and Federgruen [13] then show that, while all firms reduce their total cost by switching from dedicated to pooled service, these cost savings do not necessarily result in price reductions for all customer classes. Indeed, if a customer class gets better than average service (and belongs to the bottleneck set) at all firms, it is charged a higher price under service pooling than under dedicated service. The following provides some intuition behind this result: assume all classes initially get identical normalized waiting time standards; if class 1, say, subsequently bargains for a lower waiting time standard, the cost for the other customer classes increases,

for which externalities class 1 is made to pay.

CONCLUSIONS

Many service providers use pricing and scheduling schemes in order to improve their value proposition. In summarizing this article, it is important to highlight several important concepts that arise in the intersection between pricing and scheduling.

1. *Externalities-Based Pricing.* The concept of charging customers the externalities that they impose on other customers arises in settings in which the customers have private information regarding their delay cost as well as in settings in which firms compete while investing in a pooled capacity to accommodate different segments. The former is used to achieve incentive compatibility while the latter to make sure that different segments, those with high service levels as well as those with low service levels, are charged a competitive price.
2. *Intentional Delays.* While in many cases service organizations set priorities to minimize the total waiting time cost, via the $c\mu$ rule, this is not always the case. Both, when a monopolist needs to set priorities that are incentive-compatible and profit-maximizing, as well as in the case in which the firm needs to choose its capacity level in conjunction with priority rules in order to accommodate different service levels, the firm might be introducing intentional delays. The idea is to delay the customers of one of the segments upon completion to achieve different goals that cannot be achieved when restricted to waiting-time-minimizing rules.
3. *Follow-the-Crowd Behaviour.* The consumer behavior in the simplest service system, the M/M/1 queue, can be characterized as “avoid-the-crowd” behaviour: customers prefer joining the queue when others do not do so.

However, in the presence of priorities, the consumer behavior can be characterized as “follow-the-crowd” behaviour: the more customers purchase priorities, the more likely other customers purchase these as well. As discussed above, a monopolist can exploit such behavior when introducing pricing and scheduling scheme to customers.

REFERENCES

1. Charnsirisakskul K, Griffin P, Keskinocak P. Pricing and scheduling decisions with lead-time flexibility. *Eur J Oper Res* 2006;171(1): 153–169.
2. Naor P. The regulation of queue sizes by levying tolls. *Econometrica* 1969;37(1): 15–24.
3. Dolan RJ. Incentive mechanisms for priority queuing problems. *Bell J Econ* 1978;9:421–436.
4. Mendelson H, Whang S. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper Res* 1990;38:870–883.
5. Afeche P. Incentive-compatible revenue management in queuing systems: Optimal strategic idleness and other delaying tactics. Working paper, Kellogg School of Management, Northwestern University; 2004.
6. Balachandran KR. Purchasing priorities in queues. *Manag Sci* 1972;18:319–326.
7. Hassin R, Haviv M. To queue or not to queue: equilibrium behavior in queuing systems. Boston (MA): Kluwer Academic Publishers; 2003.
8. Hassin R, Haviv M. Equilibrium threshold strategies: the case of queues with priorities. *Oper Res* 1997;45:966–973.
9. Alperstein H. Optimal pricing policy for the service facility offering a set of priority prices. *Manag Sci* 1988;34:666–671.
10. Loch C. Pricing in markets sensitive to delay [Ph.D. Dissertation]. Stanford (CA): Stanford University; 1991.
11. Lederer PJ, Li L. Pricing, production, scheduling, and delivery-time competition. *Oper Res* 1997;45(3):407–420.
12. Armony M, Haviv M. Price and delay competition between two service providers. *Eur J Oper Res* 2003;147(1):32–50.

13. Allon G, Federgruen A. Competition in service industries with segmented markets. *Manag Sci* 2009;55(4):619–634.
14. Afèche P. Incentive-compatible revenue management in queuing systems: optimal strategic delay and other delay tactics. Toronto: Rotman School of Management; 2006. Working paper.

PRIMAL–DUAL METHODS FOR NONLINEAR CONSTRAINED OPTIMIZATION

IGOR GRIVA

Department of Mathematical Sciences,
George Mason University, Fairfax,
Virginia

ROMAN A. POLYAK

Department of SEOR and Mathematical
Sciences, George Mason University,
Fairfax, Virginia

HISTORICAL NOTE

In 1797 in his book “Theorie des fonctions analytiques” Joseph Luis Lagrange (1736–1813) introduced the Lagrangian and Lagrange multipliers rule for solving optimization problems with equality constraints.

... If a function of several variables should be a maximum or minimum and there are between these variables one or several equations, then it will be suffice to add to the proposed function the functions that should be zero, each multiplied by an undetermined quantity, and then to look for the maximum and the minimum as if the variables were independent; the equation that one will find combined with the given equations, will serve to determine all the unknowns.

J.-L. Lagrange

The Lagrange multipliers rule is only a necessary but not sufficient condition for an equality constrained optimum. The primal–dual (PD) vector is neither a maximum nor a minimum of the Lagrangian: it is a saddle point. Nevertheless, for more than 200 years the Lagrangian has remained an invaluable tool for optimization. Moreover, with time, the great value of the Lagrangian has become increasingly evident. The Lagrangian is the main instrument for establishing optimality conditions, the basic

ingredient in the Lagrange duality, and one of the most important tools in numerical constrained optimization. Any PD method for solving constrained optimization problems in one way or another uses the Lagrangian and the Lagrange multipliers.

OPTIMIZATION WITH EQUALITY CONSTRAINTS: LAGRANGE SYSTEM OF EQUATIONS

Consider $q + 1$ smooth enough functions $f, g_j : \mathbb{R}^n \rightarrow \mathbb{R}$, $j = 1, \dots, q$ and the feasible set

$$\Omega = \{x : g_j(x) = 0, j = 1, \dots, q\}.$$

The equality constrained optimization (ECO) problem consists of finding

$$(\mathcal{ECO}) \quad f(x^*) = \min\{f(x) | x \in \Omega\}.$$

The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^q \rightarrow \mathbb{R}^1$ for (ECO) is given by the formula

$$L(x, v) = f(x) - \sum_{j=1}^q v_j g_j(x).$$

Let us consider the vector function $g^T(x) = (g_1(x), \dots, g_q(x))$, the Jacobian $J(g(x)) = \nabla g(x) = (\nabla g_1(x), \dots, \nabla g_q(x))^T$ and assume

$$\text{rank } \nabla g(x^*) = q < n. \quad (1)$$

Then for x^* to be a (ECO) solution, it is necessary the existence of $v^* \in \mathbb{R}^q$ that the pair (x^*, v^*) is a solution to the following Lagrange system of equations:

$$\nabla_x L(x, v) = \nabla f(x) - \nabla g^T(x)v = 0, \quad (2)$$

$$g_i(x) = 0, \quad i = 1, \dots, q. \quad (3)$$

We consider the Hessian

$$\nabla_{xx}^2 L(x, v) = \nabla^2 f(x) - \sum_{i=1}^q v_i \nabla^2 g_i(x)$$

of the Lagrangian $L(x, v)$. The regularity condition (1), together with sufficient condition for the minimizer x^* to be isolated that is

$$\begin{aligned} \langle \nabla_{xx}^2 L(x^*, v^*) \xi, \xi \rangle &\geq m \langle \xi, \xi \rangle, \\ \forall \xi : \nabla g(x^*) \\ \xi &= 0, m > 0, \end{aligned} \quad (4)$$

comprise the standard second-order optimality conditions for the $(\mathcal{ECO})$ problem.

Application of Newton's method to the nonlinear PD system of Equations (2), (3) leads to one of the first PD methods for constrained optimization [1].

By linearizing the system (2), (3) and ignoring terms of the second and higher order, one obtains the following linear PD system for finding the Newton direction $(\Delta x, \Delta v)$:

$$\begin{bmatrix} \nabla_{xx}^2 L(\cdot) & -\nabla g^T(\cdot) \\ \nabla g(\cdot) & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ -g(\cdot) \end{bmatrix}. \quad (5)$$

PDECOM-Primal-Dual Equality Constrained Optimization Method

Step 0. Let $y \in S(y^*, \delta)$.

Step 1. If $v(y) \leq \epsilon$, Output y as the solution.

Step 2. Find Δx and Δv from (5).

Step 3. Update the primal-dual pair by the following formulas

$$x := x + \Delta x, \quad v := v + \Delta v.$$

Step 4. Goto Step 1.

If the standard second-order optimality conditions for $(\mathcal{ECO})$ are satisfied, the Lipschitz conditions for the Hessians $\nabla^2 f(x)$, $\nabla^2 g_i$, $i = 1, \dots, q$ hold, $y = (x, v) \in S(y^*, \delta)$ and $\delta > 0$ is small enough, then the PDECOM generates the PD sequence that converges to the PD solution $y^* = (x^*, v^*)$ with quadratic rate; that is, the following bound holds:

$$\|\hat{y} - y^*\| \leq C \|y - y^*\|^2, \quad (9)$$

where $C > 0$ is independent from $y \in S(y^*, \delta)$ and depends only on the problem data [1].

Let $S(y^*, \delta) = \{y = (x, v) : \|y - y^*\| \leq \delta\}$, where $\|\cdot\|$ is the Euclidean norm, δ small enough, and $0 < \epsilon \ll \delta$ is the desired accuracy. The following merit function

$$\begin{aligned} v(y) &= v(x, v) \\ &= \max \left\{ \|\nabla_x L(x, v)\|, \max_{1 \leq i \leq q} |g_i(x)| \right\}, \end{aligned} \quad (6)$$

measures the distance between the PD approximation (x, v) and the solution (x^*, v^*) . If f and g_i are smooth enough and the second-order optimality conditions for $(\mathcal{ECO})$ are satisfied, then for $y \in S(y^*, \delta)$ we have

$$v(y) = 0 \Leftrightarrow y = y^*, \quad (7)$$

and there are $0 < m_0 < M_0$ that

$$m_0 \|y - y^*\| \leq v(y) \leq M_0 \|y - y^*\|. \quad (8)$$

The PD method for $(\mathcal{ECO})$ consists of the following operations:

AUGMENTED LAGRANGIAN AND PRIMAL-DUAL AUGMENTED LAGRANGIAN METHOD

For a given $k > 0$, the augmented Lagrangian (AL) $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}_+ \rightarrow \mathbb{R}^1$ is defined by the following formula [2,3]:

$$\mathcal{L}(x, v, k) = f(x) - \sum_{j=1}^q v_j g_j(x) + \frac{k}{2} \sum_{j=1}^q g_j^2(x). \quad (10)$$

The AL method alternates the unconstrained minimization of the AL $\mathcal{L}(x, v, k)$ in the primal space with a Lagrange multipliers update.

Let $\delta > 0$ be sufficiently small, $0 < \epsilon \ll \delta$ be the desired accuracy, $k > 0$ be sufficiently large, and $y = (x, v) \in S(y^*, \delta)$. One step of the classical AL method consists of finding the primal minimizer

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} \mathcal{L}(x, v, k), \quad (11)$$

followed by the Lagrange multipliers update.

In other words, for a given scaling parameter $k > 0$ and a starting point $y = (x, v) \in S(y^*, \delta)$, one step of the AL method is equivalent to solving the following nonlinear PD system for $\hat{x}$ and $\hat{v}$:

$$\nabla_x L(\hat{x}, \hat{v}) = \nabla f(\hat{x}) - \sum_{j=1}^q \hat{v}_j \nabla g_j(\hat{x}) = 0, \quad (12)$$

$$\hat{v} - v + kg(\hat{x}) = 0. \quad (13)$$

PDALM - Primal-Dual Augmented Lagrangian Method

Step 0. Let $y \in S(y^*, \delta)$ and $k > 0$ is large enough.

Step 1. If $v(y) \leq \epsilon$, Output y as a solution.

Step 2. Set $k = v(x, v)^{-1}$.

Step 3. Find Δx and Δv from (14).

Step 4. Update the primal-dual pair by the formulas

$$x := x + \Delta x, \quad v := v + \Delta v.$$

Step 5. Goto Step 1.

If the standard second-order optimality conditions for $(\mathcal{ECO})$ are satisfied, the Lipschitz conditions for the Hessians $\nabla^2 f(x)$, $\nabla^2 g_i$, $i = 1, \dots, q$ hold, $y = (x, v) \in S(y^*, \delta)$, and $\delta > 0$ is sufficiently small, then the PDALM generates a PD sequence that converges to the PD solution $y^* = (x^*, v^*)$ with a quadratic rate, that is Equation (9) holds [4].

OPTIMIZATION WITH INEQUALITY CONSTRAINTS

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$ be convex and all $c_i: \mathbb{R}^n \rightarrow \mathbb{R}^1$, $i = 1, \dots, p$ be concave and

Application of Newton's method for solving the PD system (12)–(13) for $(\hat{x}, \hat{v})$ together with a proper update of the penalty parameter $k > 0$ leads to the primal-dual augmented Lagrangian (PDAL) method for solving $(\mathcal{ECO})$ problems [4].

By linearizing the system (12)–(13) and ignoring terms of the second and higher order, we obtain the following linear PD system for finding the Newton direction $(\Delta x, \Delta v)$:

$$\begin{bmatrix} \nabla_{xx}^2 L(\cdot) & -\nabla g^T(\cdot) \\ \nabla g(\cdot) & \frac{1}{k} I_q \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ -g(\cdot) \end{bmatrix}, \quad (14)$$

where I_q is the identity matrices in $\mathbb{R}^q$. Note that if $k \rightarrow \infty$, then the system (14) gets close to the system (5). Therefore, by changing $k > 0$ properly it is possible to achieve a quadratic rate of convergence for the PDAL method.

smooth functions. Consider a convex set $\Omega = \{x \in \mathbb{R}^n : c_i(x) \geq 0, i = 1, \dots, p\}$ and the following convex inequality constrained optimization (ICO) problem:

$$(\mathcal{ICO}) \quad f(x^*) = \min\{f(x) | x \in \Omega\}.$$

Let us assume that

- (a) The primal optimal set X^* is not empty and bounded.
- (b) The Slater's condition holds; that is, there exists $\hat{x} \in \mathbb{R}^n : c_i(\hat{x}) > 0, i = 1, \dots, p$.

Due to the assumption **B**, the Karush–Kuhn–Tucker’s (KKT’s) conditions hold true; that is, there exists a vector $u^* = (u_1^*, \dots, u_p^*) \in \mathbb{R}_+^p$ such that

$$\nabla_x L(x^*, u^*) = \nabla f(x^*) - \sum_{i=1}^p u_i^* \nabla c_i(x^*) = 0, \quad (15)$$

$$c_i(x^*) \geq 0, \quad u_i^* \geq 0, \quad i = 1, \dots, p, \quad (16)$$

and the complementary slackness conditions

$$u_i^* c_i(x^*) = 0, \quad i = 1, \dots, p \quad (17)$$

hold true.

Let $I = \{i : c_i(x^*) = 0\} = \{1, \dots, r\}$ be the set of indices of the active at x^* constraints, $c_{(r)}^T(x) = (c_1(x), \dots, c_r(x))$ be the vector function of the active at x^* constraints and $\nabla c_{(r)}(x) = J(c_{(r)}(x))$ be the corresponding Jacobian. The standard second-order optimality conditions for (ICO) consists of existence $u^* \in \mathbb{R}_+^p$ and $m > 0$ such that

$$\text{rank}(\nabla c_{(r)}(x^*)) = r, \quad u_i^* > 0, \quad i = 1, \dots, r \quad (18)$$

and

$$\langle \nabla_{xx}^2 L(x^*, u^*) \xi, \xi \rangle \geq m \langle \xi, \xi \rangle, \quad \forall \xi : \nabla c_{(r)}(x^*) \xi = 0. \quad (19)$$

The dual to (ICO) problem consists of finding

$$(\mathcal{D}) \quad d(u^*) = \max \{d(u) | u \in \mathbb{R}_+^p\},$$

where $d(u) = \inf_{x \in \mathbb{R}^n} L(x, u)$ is the dual function.

For convex (ICO) that satisfies the Slater condition we have

$$f(x^*) = d(u^*).$$

The PD methods generate PD sequences $\{x^s, u^s\}_{s=1}^\infty$ that

$$f(x^*) = \lim_{s \rightarrow \infty} f(x^s) = \lim_{s \rightarrow \infty} d(u^s) = d(u^*).$$

LOG-BARRIER FUNCTION AND INTERIOR-POINT METHODS

The (ECO) problem is not combinatorial by nature because all constraints are active at the solution (by definition, active constraints are those that are satisfied as equalities). At the same time, the methods for (ECO) require an initial approximation from the neighborhood $S(y^*, \delta)$. The (ICO) in general and convex (ICO) in particular, are combinatorial by nature because the set of active at the solution constraints is unknown *a priori*. On the other hand, for convex (ICO) there is no need to have an initial approximation $y \in S(y^*, \delta)$.

Consider PD interior-point methods for convex optimization that are based on the classical log-barrier function $\beta : \text{int } \Omega \times \mathbb{R}_+ \rightarrow \mathbb{R}$ defined by the formula

$$\beta(x, \mu) = f(x) - \mu \sum_{i=1}^p \ln c_i(x).$$

Let $\ln t = -\infty$ for $t \leq 0$; then for any given $\mu > 0$ the Frisch log-barrier function $F : \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}$ is defined as follows:

$$F(x, \mu) = \begin{cases} \beta(x, \mu), & x \in \text{int } \Omega; \\ \infty, & x \notin \text{int } \Omega. \end{cases}$$

The classical sequential unconstrained minimization technique (SUMT) [5] finds the PD trajectory $\{y(\mu) = (x(\mu), u(\mu))\}$ by the following formulas:

$$\begin{aligned} \nabla_x F(x(\mu), \mu) &= \nabla f(x(\mu)) \\ - \sum_{i=1}^p \mu (c_i(x(\mu)))^{-1} \nabla c_i(x(\mu)) &= 0 \end{aligned} \quad (20)$$

and

$$\begin{aligned} u(\mu) &= (u_i(\mu)) = \mu (c_i(x(\mu)))^{-1}, \\ i &= 1, \dots, p. \end{aligned} \quad (21)$$

Equation (20) is the optimality condition for the unconstrained problem

$$x(\mu) = \arg \min \{F(x, \mu) | x \in \mathbb{R}^n\}.$$

Thus, for a given $\mu > 0$, finding the PD approximation from Equations (20), (21) is equivalent to solving the following system for $(\hat{x}, \hat{u})$:

$$\begin{aligned} \nabla_x L(\hat{x}, \hat{u}) &= \nabla f(\hat{x}) - \sum_{i=1}^p \hat{u}_i \nabla c_i(\hat{x}) = 0, \\ \hat{u}_i c_i(\hat{x}) &= \mu, \quad i = 1, \dots, p. \end{aligned} \quad (22)$$

It follows from Equation (20) that $x(\mu) \in \text{int } \Omega$. It follows from Equation (21) that $u(\mu) \in \mathbb{R}_{++}^p$, that is $x(\mu)$ and $u(\mu)$ are primal and dual interior points. It follows from Equations (20), (21) that

$$\nabla_x F(x(\mu), \mu) = \nabla_x L(x(\mu), u(\mu)) = 0,$$

or

$$d(u(\mu)) = \min_{x \in \mathbb{R}^n} L(x, u(\mu)) = L(x(\mu), u(\mu))$$

and the PD gap (i.e., the difference between the values of the primal and dual objective functions) is

$$\begin{aligned} \Delta(\mu) &= f(x(\mu)) - d(u(\mu)) \\ &= \sum_{i=1}^p u_i(\mu) c_i(x(\mu)) = p\mu, \end{aligned}$$

that is

$$\lim_{\mu \rightarrow 0} x(\mu) = x^* \quad \text{and} \quad \lim_{\mu \rightarrow 0} u(\mu) = u^*. \quad (23)$$

Finding $(\hat{x}, \hat{u}) = (x(\mu), u(\mu))$ or its approximation from Equation (22) and reducing $\mu > 0$ is the main idea of the SUMT [5].

Solving the nonlinear PD system (22) is an infinite procedure in general. The main idea of the PD interior-point methods is to replace the nonlinear PD system (22) by one Newton step toward solution of the system (22) and follow that by the barrier parameter update. For a given approximation $y = (x, u)$ and the barrier parameter $\mu > 0$, the application of Newton's method to the nonlinear PD system

(22) leads to the following linear PD system:

$$\begin{bmatrix} \nabla_{xx}^2 L(x, u) & -\nabla c(x)^T \\ U \nabla c(x) & C(x) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta u \end{bmatrix} = \begin{bmatrix} -\nabla L(x, u) \\ -U c(x) + \mu e \end{bmatrix}, \quad (24)$$

where $C(x) = \text{diag } (c_i(x))_{i=1}^p$, $U = \text{diag } (u_i)_{i=1}^p$, and $e = (1, \dots, 1)^T \in \mathbb{R}^p$.

The system (24) finds the Newton direction $\Delta y = (\Delta x, \Delta u)$, which is used to update the current approximation $y = (x, u)$:

$$\bar{x} = x + \alpha \Delta x; \quad \bar{u} = u + \alpha \Delta u. \quad (25)$$

One must determine the step length $\alpha > 0$ in such a way that a new approximation $\bar{y} = (\bar{x}, \bar{u})$ not only remains primal and dual interior, but also remains in the area for which Newton's method for PD system (22) is well defined for the updated barrier parameter [6]. For some classes of constrained optimization problems, including linear programming problems (LPs) and quadratic programming problems (QPs), it is possible to take $\alpha = 1$ in Equation (25) and update the barrier parameter μ by the formula $\bar{\mu} = \mu(1 - \rho/\sqrt{n})$, where $0 < \rho < 1$ is independent on n . The new approximation belongs to the neighborhood of the solution of the system (22) with μ replaced by $\bar{\mu}$. Moreover, each step reduces the PD gap by the same factor $(1 - \rho/\sqrt{n})$. This leads to polynomial complexity of the PD interior-point methods for LP and QP [6]. LP, QP, and QP with quadratic constraints problems are well structured, which means that constraints and epigraph of the objective function can be equipped with a self-concordant (SC) barrier [7,8].

If (ICQ) problem is not well structured, establishing polynomial complexity of the path-following methods becomes problematic, if not impossible. Nevertheless, the PD interior-point approach remains productive and leads to globally convergent and numerically efficient PD methods [9–13].

There are two main classes of PD methods for (ICQ): interior-point PD method based on the path-following idea (which goes back to SUMT) and exterior-point PD methods (which are based on nonlinear rescaling (NR) theory [14–16]).

PRIMAL-DUAL INTERIOR-POINT METHODS

corresponds to (ECO) (27).

Consider the following problem

$$\begin{aligned} \min & f(x), \\ \text{s.t. } & c(x) - w = 0, \\ & w \geq 0, \end{aligned} \quad (26)$$

which is equivalent to (ICO) problem.

The log-barrier function $F(x, w, \mu) = f(x) - \mu \sum_{i=1}^p \ln w_i$ is used to handle the nonnegativity of the slack vector $w \in \mathbb{R}_+^p$, which guarantees the primal feasibility.

One step of the path-following method consists of solving the following (ECO) problem:

$$\begin{aligned} \min & F(x, w, \mu), \\ \text{s.t. } & c(x) - w = 0, \end{aligned} \quad (27)$$

followed by the barrier parameter $\mu > 0$ update.

The Lagrangian for the problem (27) is defined by formula

$$\begin{aligned} L(x, w, u) \\ = f(x) - \mu \sum_{i=1}^p \log w_i - \sum_{i=1}^p u_i (c_i(x) - w_i). \end{aligned}$$

Let $W = \text{diag}(w_i)_{i=1}^p$ and $U = \text{diag}(u_i)_{i=1}^p$. The following Lagrange system of equations

$$\begin{aligned} \nabla f(x) - \nabla c(x)^T u &= 0 \\ -\mu e + W U e &= 0 \\ c(x) - w &= 0 \end{aligned} \quad (28)$$

Application of Newton's method to nonlinear PD system (28) leads to the following linear PD system for finding the Newton directions:

$$\begin{bmatrix} \nabla_{xx}^2 L(x, u) & 0 & -\nabla c(x)^T \\ 0 & U & W \\ \nabla c(x) & -I_p & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta w \\ \Delta u \end{bmatrix} = \begin{bmatrix} -\nabla f(x) + \nabla c(x)^T u \\ \mu e - W U e \\ -c(x) + w \end{bmatrix}, \quad (29)$$

where $\nabla_{xx}^2 L(x, u) = \nabla^2 f(x) - \sum_{i=1}^p u_i \nabla^2 c_i(x)$ is the Hessian in x of the Lagrangian for (ICO) problem.

For convex (ICO) the merit function

$$\begin{aligned} v(y) \equiv v(x, w, u) &= \max \{ \|\nabla_x L(x, w, u)\|, \\ &\quad \|c(x) - w\|, \|W U e\|, \} \end{aligned}$$

is used. Under the standard second-order optimality conditions, the merit function $v(y)$ satisfies Equations (7) and (8).

Take $\delta > 0$ to be sufficiently small and $0 < \epsilon \ll \delta$ to be the defined accuracy. Consider the following PD interior-point method:

PDIPM - Primal-Dual Interior-Point Method

Step 0. Let $y \in S(y^*, \delta) = \{y : \|y - y^*\| \leq \delta\}$ be the initial approximation.

Step 1. If $v(x, w, u) \leq \epsilon$, Output $y = (x, u)$ as a solution.

Step 2. Calculate the barrier parameter $\mu = \min \{\theta \mu, v^2(x, w, u)\}$, $0 < \theta < 1$.

Step 3. Find Δx , Δw , and Δu from (29).

Step 4. Calculate the parameter κ and the step lengths α_P and α_D by formulas

$$\begin{aligned} \kappa &= \max \{ \bar{\kappa}, 1 - v(x, w, u) \}, \quad 0 < \bar{\kappa} < 1, \\ \alpha_P &= \min_{1 \leq i \leq m} \left\{ 1; -\kappa \frac{(w^s)_i}{(\Delta w^s)_i} : (\Delta w^s)_i < 0 \right\}, \\ \alpha_D &= \min_{1 \leq i \leq m} \left\{ 1; -\kappa \frac{(u^s)_i}{(\Delta u^s)_i} : (\Delta u^s)_i < 0 \right\}. \end{aligned}$$

Step 5. Update the primal-dual pair by the formulas

$$\hat{x} := x + \alpha_P \Delta x, \quad \hat{w} := w + \alpha_P \Delta w, \quad \hat{u} := u + \alpha_D \Delta u.$$

Step 6. Goto Step 1.

If the standard second-order optimality conditions are satisfied for $(\mathcal{ICO})$, the Hessians $\nabla^2 f(x)$ and $\nabla^2 c_i(x)$, $i = 1, \dots, p$ are Lipschitz continuous, $\delta > 0$ is sufficiently small, and a starting point $y \in S(y^*, \delta)$, then the PD sequence generated by the PDIPM converges to the KKT's point with a quadratic rate [10,17].

PRIMAL-DUAL NONLINEAR RESCALING METHOD

Consider a class Ψ of strictly concave and twice continuously differentiable functions $\psi : (t_0, t_1) \rightarrow \mathbb{R}$, $-\infty < t_0 < 0 < t_1 < \infty$ that satisfy the following properties:

- 1⁰. $\psi(0) = 0$.
- 2⁰. $\psi'(t) > 0$.
- 3⁰. $\psi'(0) = 1$.
- 4⁰. $\psi''(t) < 0$.
- 5⁰. there is $a > 0$ that $\psi(t) \leq -at^2$, $t \leq 0$.
- 6⁰. $a)\psi'(t) \leq b_1 t^{-1}$, $b) -\psi''(t) \leq b_2 t^{-2}$, $t > 0$,
 $b_1 > 0, b_2 > 0$.

The following transformations $\psi \in \Psi$ satisfy the above properties:

1. Exponential transformation [18]

$$\psi_1(t) = 1 - e^{-t}.$$

2. Logarithmic MBF [14]

$$\psi_2(t) = \ln(t + 1).$$

3. Hyperbolic MBF [14]

$$\psi_3(t) = \frac{t}{1+t}.$$

Each of the above transformations can be modified in the following way. For a

given $-1 < \tau < 0$ we define quadratic extrapolation of the transformations 1 – 3 by the formulas

4.

$$\psi_{q_i}(t) = \begin{cases} \psi_i(t), & t \geq \tau, \\ q_i(t) = a_i t^2 + b_i t + c_i, & t \leq \tau, \end{cases}$$

where a_i, b_i, c_i one finds from the following equations: $\psi_i(\tau) = q_i(\tau)$, $\psi'_i(\tau) = q'_i(\tau)$, $\psi''_i(\tau) = q''_i(\tau)$.

Modification 4 leads to transformations that are defined on $(-\infty, \infty)$ and, along with penalty function properties, have some additional important features [15,16,19].

Due to 1⁰ – 3⁰ for any $\psi \in \Psi$ and any $k > 0$, the following problem

$$f(x^*) = \min \left\{ f(x) | k^{-1} \psi(k c_i(x)) \geq 0, \right. \\ \left. i = 1, \dots, p \right\} \quad (30)$$

is equivalent to the original $(\mathcal{ICO})$ problem. The classical Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}_+ \rightarrow \mathbb{R}^1$ for the equivalent problem (30) is defined by the formula

$$\mathcal{L}(x, u, k) = f(x) - k^{-1} \sum_{i=1}^p u_i \psi(k c_i(x)).$$

The NR principle consists of transforming the original $(\mathcal{ICO})$ problem into an equivalent one and then using the classical Lagrangian for the equivalent problem for theoretical analysis and numerical methods. In contrast to SUMT, the NR principle leads to exterior-point methods. Convergence of the NR methods is due to both, the Lagrange multipliers and the scaling parameter updates.

We use the following merit function:

$$v \equiv v(x, u) = \max \left\{ \|\nabla_x L(x, u)\|, -\min_{1 \leq i \leq p} c_i(x), \sum_{i=1}^p |u_i| |c_i(x)|, -\min_{1 \leq i \leq p} u_i \right\}. \quad (31)$$

For convex (ICCO) under the standard second-order optimality condition, the merit function (31) satisfies Equations (7) and (8).

For a given $u \in \mathbb{R}_{++}^p$ and $k > 0$, one step of the NR method with scaling parameter update consists in finding

$$\hat{x} = \arg \min \{ \mathcal{L}(x, u, k) \mid x \in \mathbb{R}^n \}, \quad (32)$$

or

$$\begin{aligned} \hat{x} : \nabla_x \mathcal{L}(\hat{x}, u, k) &= \nabla f(\hat{x}) - \sum_{i=1}^p \psi'(kc_i(\hat{x})) u_i \nabla c_i(\hat{x}) \\ u_i \nabla c_i(\hat{x}) &= 0, \end{aligned}$$

followed by the Lagrange multipliers update by formula

$$\hat{u}_i = \psi'(kc_i(\hat{x})) u_i, i = 1, \dots, p,$$

or

$$\hat{u} = \Psi'(kc(\hat{x})) u, \quad (33)$$

where $\Psi'(kc(x)) = \text{diag}(\psi'(kc_i(x)))_{i=1}^p$, and scaling parameter update by formula

$$\hat{k} := v(\hat{x}, \hat{u})^{-0.5}. \quad (34)$$

In other words, for a given Lagrange multipliers vector $u \in \mathbb{R}_{++}^p$ and the scaling

parameter $k > 0$, one step of the NR method is equivalent to solving the following nonlinear PD system:

$$\begin{aligned} \nabla_x \mathcal{L}(\hat{x}, u, k) &= \nabla f(\hat{x}) - \sum_{i=1}^p \psi'(kc_i(\hat{x})) u_i \nabla c_i(\hat{x}) \\ &= \nabla_x L(\hat{x}, \hat{u}) = 0, \end{aligned} \quad (35)$$

$$\hat{u} = \Psi'(kc(\hat{x})) u, \quad (36)$$

for $\hat{x}$ and $\hat{u}$, followed by the scaling parameter $k > 0$ update by Equation (34).

Application of Newton's method for solving the nonlinear PD system (35)–(36) for $\hat{x}$ and $\hat{u}$ leads to the primal-dual nonlinear rescaling (PDNR) method for solving (ICCO) problem [20].

By linearizing Equations (35)–(36) and assuming that $\bar{u} = \Psi'(kc(x)) u$, we obtain the following linear PD system for finding the Newton direction $(\Delta x, \Delta u)$:

$$\begin{bmatrix} \nabla_{xx}^2 L(\cdot) & -\nabla c^T(\cdot) \\ -U \Psi''(\cdot) \nabla c(\cdot) & \frac{1}{k} I_p \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta u \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ \frac{1}{k} (\bar{u} - u) \end{bmatrix}, \quad (37)$$

where $\nabla c(\cdot) = \nabla c(x)$, $\Psi'(\cdot) = \Psi'(kc(x)) = \text{diag}(\psi''(kc_i(x)))_{i=1}^p$, $U = \text{diag}(u_i)_{i=1}^p$ and I_p is an identity matrix in $\mathbb{R}^p$.

Let $\delta > 0$ be small enough and $0 < \epsilon \ll \delta$ be the desired accuracy. Then for a given $x \in \mathbb{R}^n$, Lagrange multipliers vector $u \in \mathbb{R}_{++}^p$, and scaling parameter $k > 0$, one step of the PDNR method consists of the following operations:

PDNRM - Primal-Dual Nonlinear Rescaling Method

Step 0. Let $y = (x, u) \in S(y^*, \delta)$.

Step 1. If $v(y) \leq \epsilon$, then output y as a solution.

Step 2. Set $k = v(x, u)^{-0.5}$.

Step 3. Find Δx and Δu from (37).

Step 4. Update the primal-dual pair by the formulas

$$\hat{x} := x + \Delta x, \quad \hat{u} := u + \Delta u.$$

Step 5. Goto Step 1.

If the standard second-order optimality conditions are satisfied, the Lipschitz conditions for the Hessians $\nabla^2 f(x)$, $\nabla^2 c_i$, $i = 1, \dots, p$ hold, and $y = (x, u) \in S(y^*, \delta)$, then the PDNR algorithm generates a PD sequence that converges to the PD solution with 1.5-Q-superlinear rate. A small modification of the PDNR algorithm generates a PD sequence that converges from any starting point $y = (x, u) \in \mathbb{R}^n \times \mathbb{R}_{++}^p$ with asymptotic 1.5-Q-superlinear rate [20].

PRIMAL-DUAL EXTERIOR-POINT METHOD FOR INEQUALITY CONSTRAINED OPTIMIZATION

Due to properties 1⁰ and 2⁰ for any given vector $k = (k_1, \dots, k_p) \in \mathbb{R}_{++}^p$, we have

$$c_i(x) \geq 0 \iff k_i \psi(k_i c_i(x)) \geq 0, \quad i = 1, \dots, p.$$

Consider the Lagrangian $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}_{++}^p \rightarrow \mathbb{R}$

$$\mathcal{L}(x, u, \mathbf{k}) = f(x) - \sum_{i=1}^p k_i^{-1} u_i \psi(k_i c_i(x))$$

for the equivalent problem. Let $k > 0$, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^p$ and $\mathbf{k} = (k_i = k(u_i)^{-1}, i = 1, \dots, p)$.

One step of the NR method with “dynamic” scaling parameters update maps the given triple $(x, u, \mathbf{k}) \in \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}_{++}^p$ into the triple $(\hat{x}, \hat{u}, \hat{\mathbf{k}}) \in \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}_{++}^p$ defined by formulas

$$\begin{aligned} \hat{x} &: \nabla_x \mathcal{L}(\hat{x}, u, \mathbf{k}) \\ &= \nabla f(\hat{x}) - \sum_{i=1}^p \psi'(k_i c_i(\hat{x})) u_i \nabla c_i(\hat{x}) \\ &= \nabla f(\hat{x}) - \sum_{i=1}^p \hat{u}_i \nabla c_i(\hat{x}), \end{aligned} \quad (38)$$

$$\hat{u} : \hat{u}_i = u_i \psi'(k_i c_i(\hat{x})), \quad i = 1, \dots, p, \quad (39)$$

$$\hat{\mathbf{k}} : \hat{k}_i = k \hat{u}_i^{-1}, \quad i = 1, \dots, p. \quad (40)$$

Such a method was considered in Ref. 18 for exponential transformation ψ_1 . By removing the formula for the scaling vector update (40)

from the system (38), (39), (40) we obtain the following nonlinear PD system:

$$\nabla_x L(\hat{x}, \hat{u}) = \nabla f(\hat{x}) - \sum_{i=1}^p \hat{u}_i \nabla c_i(\hat{x}), \quad (41)$$

$$\hat{u} = \Psi'(k c(\hat{x})) u, \quad (42)$$

where $\Psi'(k c(\hat{x})) = \text{diag}(\psi'(k_i c_i(\hat{x})))_{i=1}^p$. To measure the distance between the current approximation $y = (x, u)$ and the solution y^* , the merit function (31) is used.

Due to Equation (40), the Lagrange multipliers that correspond to the passive constraints (LMPC) converge to zero with at least a quadratic rate. Therefore, for $0 < \epsilon \ll 1$, it requires at most $s = \mathcal{O}(\ln \ln \epsilon^{-1})$ Lagrange multipliers updates for the LMPC to become of the order of $\mathcal{O}(\epsilon^2)$. So the PD system (41), (42) is reduced to the following nonlinear PD system:

$$\nabla_x L(\hat{x}, \hat{u}) = \nabla f(\hat{x}) - \sum_{i=1}^r \hat{u}_i \nabla c_i(\hat{x}), \quad (43)$$

$$\hat{u}_i = u_i \psi'(k_i c_i(\hat{x})), \quad i = 1, \dots, r, \quad (44)$$

where $I^* = \{i : c_i(x^*) = 0\} = \{1, \dots, r\}$.

Application of Newton’s method for solving the PD system (43), (44) leads to a local PDEP method for (ICCO) problems. Start with linearization of the system (44); due to the property 2⁰ of the transformation $\psi \in \Psi$, the inverse ψ'^{-1} exists. Therefore using the identity $\psi'^{-1} = \psi'^*$, where $\psi^*(s) = \inf_t \{st - \psi(t) | t \in \mathbb{R}\}$ is the conjugate of ψ , and denoting $\varphi = -\psi^*$, we can rewrite Equation (44) as follows:

$$\begin{aligned} c_i(\hat{x}) &= k^{-1} u_i \psi'^{-1} \left(\frac{\hat{u}_i}{u_i} \right) = k^{-1} u_i \psi'^* \left(\frac{\hat{u}_i}{u_i} \right) \\ &= -k^{-1} u_i \varphi' \left(\frac{\hat{u}_i}{u_i} \right). \end{aligned}$$

It follows from property 3⁰ of transformation ψ that $\varphi'(1) = 0$. Assuming $\hat{x} = x + \Delta x$ and $\hat{u} = u + \Delta u$, and ignoring terms of the second and higher order, we obtain

$$c_i(\hat{x}) = c_i(x) + \nabla c_i(x) \Delta x$$

$$\begin{aligned}
&= -k^{-1}u_i\varphi'\left(\frac{u_i + \Delta u_i}{u_i}\right) \\
&= -k^{-1}u_i\varphi'\left(1 + \frac{\Delta u_i}{u_i}\right) = -k^{-1}\varphi''(1)\Delta u_i, \\
&i = 1, \dots, r,
\end{aligned}$$

or

$$\begin{aligned}
&c_i(x) + \nabla c_i(x)\Delta x + k^{-1}\varphi''(1)\Delta u_i = 0, \\
&i = 1, \dots, r.
\end{aligned}$$

By linearizing the system (43) at $y = (x, u)$, we obtain the following linear PD system for

finding the PD Newton directions:

$$\begin{bmatrix} \nabla_{xx}L(\cdot) & -\nabla c^T(\cdot) \\ \nabla c(\cdot) & k^{-1}\varphi''(1)I_r \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta u \end{bmatrix} = \begin{bmatrix} \nabla_x L(\cdot) \\ -c(\cdot) \end{bmatrix} \quad (45)$$

where I_r is the identity matrix in $\mathbb{R}^r$ and $\nabla c(x) = J(c(x))$ is the Jacobian of the vector function $c(x)$.

Let $\delta > 0$ be small enough and $0 < \epsilon \ll \delta$ be the desired accuracy. Then the PDEP method for (ICCO) problems consists of the following operations.

PDEPICOM - Primal-Dual Exterior-Point Method for ICCO

- Step 0. Let $y = (x, u) \in S(y^*, \delta)$.
Step 1. If $v(y) \leq \epsilon$, then output y as a solution.
Step 2. Set $k = v(x, u)^{-1}$.
Step 3. Find Δx and Δu from (45).
Step 4. Update the primal-dual pair by the formulas

$$\hat{x} := x + \Delta x, \quad \hat{u} := u + \Delta u.$$

Step 5. Goto Step 1.

Under the standard second-order optimality conditions and Lipschitz conditions for the Hessians $\nabla^2 f(x)$, $\nabla^2 c_i$, $i = 1, \dots, r$, the PDEP algorithm generates a PD sequence that converges to the PD solution $y^* = (x^*, u^*)$ with quadratic rate. The globally convergent PDEP method with an asymptotic quadratic rate is given in [21].

PRIMAL-DUAL EXTERIOR-POINT METHOD FOR OPTIMIZATION PROBLEMS WITH BOTH EQUALITY AND INEQUALITY CONSTRAINTS

Consider $p + q + 1$ twice continuously differential functions $f, c_i, g_j: \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, p, j = 1, \dots, q$ and the feasible set

$$\begin{aligned}
\Omega = \{x : c_i(x) \geq 0, i = 1, \dots, p; \\
g_j(x) = 0, j = 1, \dots, q\}.
\end{aligned}$$

The problem with both inequality and equality constraints consists of finding

$$(\mathcal{IECO}) \quad f(x^*) = \min\{f(x) | x \in \Omega\}.$$

We use $\psi \in \Psi$ to transform the inequality constraints $c_i(x) \geq 0$, $i = 1, \dots, p$ into an equivalent set of constraints. For any fixed $k > 0$, the following problem is equivalent to the original (IECO) problem due to the properties of $\psi \in \Psi$; that is,

$$\begin{aligned}
f(x^*) &= \min\{f(x) | k^{-1}\psi(kc_i(x)) \geq 0, \\
&i = 1, \dots, p; g_j(x) = 0, j = 1, \dots, q\}.
\end{aligned}$$

For a given $k > 0$, the AL $\mathcal{L}_k: \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}^q \rightarrow \mathbb{R}^1$ for the equivalent problem is defined by the formula

$$\begin{aligned}
\mathcal{L}_k(x, u, v) &= f(x) - k^{-1} \sum_{i=1}^p u_i \psi(kc_i(x)) \\
&- \sum_{j=1}^q v_j g_j(x) + \frac{k}{2} \sum_{j=1}^q g_j^2(x). \quad (46)
\end{aligned}$$

The first two terms define the classical Lagrangian for the equivalent problem in the absence of equality constraints. The last two

terms coincide with the AL terms associated with equality constraints.

For a given $k > 0$ and $y = (x, u, v) \in S(y^*, \delta)$, a step of nonlinear rescaling-AL method maps the triple $y = (x, u, v)$ into the triple $\hat{y} = (\hat{x}, \hat{u}, \hat{v})$, where

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} \mathcal{L}_k(x, u, v), \quad (47)$$

or

$$\begin{aligned} \hat{x} : \nabla f(\hat{x}) - \sum_{i=1}^p u_i \psi'(kc_i(\hat{x})) \nabla c_i(\hat{x}) \\ - \sum_{j=1}^q (v_j - kg_j(\hat{x})) \nabla g_j(\hat{x}) = 0, \end{aligned} \quad (48)$$

$$\hat{u}_i = u_i \psi'(kc_i(\hat{x})), \quad i = 1, \dots, p; \quad (49)$$

$$\hat{v}_j = v_j - kg_j(\hat{x}), \quad j = 1, \dots, q. \quad (50)$$

The system (48)–(50) can be replaced by the following nonlinear PD system:

$$\begin{aligned} \nabla_x L(\hat{x}, \hat{u}, \hat{v}) = \nabla f(\hat{x}) - \sum_{i=1}^p \hat{u}_i \nabla c_i(\hat{x}) \\ - \sum_{j=1}^q \hat{v}_j \nabla g_j(\hat{x}) = 0, \end{aligned} \quad (51)$$

$$\hat{u} - \Psi'(kc(\hat{x})) u = 0, \quad (52)$$

$$\hat{v} - v + kg(\hat{x}) = 0, \quad (53)$$

where $\Psi'(kc(\hat{x})) = \text{diag}(\psi'(kc_i(\hat{x})))_{i=1}^p$.

We use the following merit function:

PDEPM - Primal-Dual Exterior-Point Method

Step 0. Let $y \in S(y^*, \delta)$.

Step 1. If $\mu(y) \leq \epsilon$, then output y as a solution.

Step 2. Set $k = v(x, u, v)^{-0.5}$.

Step 3. Find Δx , Δu , and Δv from Equation (55).

Step 4. Update the primal-dual pair by the formulas

$$\hat{x} := x + \Delta x, \quad \hat{u} := u + \Delta u, \quad \hat{v} := v + \Delta v.$$

Step 5. Goto Step 1.

$$v(y) = v(x, u, v)$$

$$= \max\{\|\nabla_x L(x, u, v)\|, \quad - \min_{1 \leq i \leq p} c_i(x),$$

$$\max_{1 \leq i \leq q} |g_i(x)|, \quad \sum_{i=1}^p |u_i| |c_i(x)|, \quad - \min_{1 \leq i \leq p} u_i\}. \quad (54)$$

Under the standard second-order optimality conditions, $v(y)$ possesses properties (7), (8).

Application of Newton's method to the system (51), (52), (53) for $\hat{x}$, $\hat{u}$, and $\hat{v}$ from the starting point $y = (x, u, v) \in S(y^*, \delta)$ leads to the PD exterior-point method (PDEPM) [22]. By linearizing the system (51), (52), (53) and ignoring terms of the second and higher order, we obtain the following linear PD system for finding the Newton direction $\Delta y = (\Delta x, \Delta u, \Delta v)$:

$$\begin{bmatrix} \nabla_{xx}^2 L(\cdot) & -\nabla c^T(\cdot) & -\nabla g^T(\cdot) \\ -U \Psi''(\cdot) \nabla c(\cdot) & \frac{1}{k} I_p & 0 \\ \nabla g(\cdot) & 0 & \frac{1}{k} I_q \end{bmatrix} \times \begin{bmatrix} \Delta x \\ \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} -\nabla_x L(\cdot) \\ \frac{1}{k} (\bar{u} - u) \\ -g(\cdot) \end{bmatrix}, \quad (55)$$

where $\nabla c(\cdot) = \nabla c(x)$, $\nabla g(\cdot) = \nabla g(x)$, $\Psi'(\cdot) = \Psi'(kc(x)) = \text{diag}(\psi'(kc_i(x)))_{i=1}^p$, $\bar{u} = \Psi'(kc(x))u$, $U = \text{diag}(u_i)_{i=1}^p$, and I_p , I_q are the identity matrices in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively.

The PD exterior-point method for constrained optimization problems with both inequality and equality constraints consists of the following operations.

Under the standard second-order optimality conditions and Lipschitz conditions for the Hessians $\nabla^2 f(x)$, $\nabla^2 c_i$, $i = 1, \dots, m$, $\nabla^2 g_i$, $i = 1, \dots, q$, for any $y \in S(y^*, \delta)$, the PDEPM algorithm generates the PD sequence that converges to the PD solution with 1.5-Q-superlinear rate [22].

CONCLUDING REMARKS

As we have seen, any PD method for constrained optimization is associated with particular PD nonlinear system of equations. Application of Newton's method to the system is equivalent to one step of either an interior- or an exterior-point method. Therefore, the computational process is defined by a particular PD nonlinear system and excellent convergence properties of Newton's method. As a result, under the standard second-order optimality conditions, the PD methods generate sequences that converge to the PD solution with an asymptotic quadratic rate.

As a practical matter it is worth mentioning that in the neighborhood of the solution the PD interior-point method generates a sequence similar to that generated by Newton's method for solving KKT's system, while the PD exterior-point method generates a sequence similar to that generated by Newton's method for solving the Lagrange system of equations that correspond to the active constraints. Unlike the former nonlinear system, the latter system does not contain complementarity equations. Therefore, the practical PD exterior-point method often is capable of finding the PD solution with a very high of accuracy.

REFERENCES

1. Polyak BT. Introduction to optimization. New York: Software Inc.; 1987.
2. Hestenes MR. Multipliers and gradient methods. *J Optim Theory Appl* 1969;4(5):303–320.
3. Powell MJD. A method for nonlinear constraints in minimization problems. In: Fletcher R, editor. *Optimization*. London: Academic Press; 1969. pp. 283–298.
4. Polyak RA. On the local quadratic convergence of the primal–dual augmented Lagrangian method. *Optim Method Soft* 2009;24(3):369–379.
5. Fiacco AV, McCormick GP. *Nonlinear programming: sequential unconstrained minimization techniques*. New York: John Wiley & Sons, Inc.; 1968.
6. Wright SJ. *Primal–dual interior-point methods*. Philadelphia (PA): SIAM; 1997.
7. Nesterov Y, Nemirovsky A. *Interior-point polynomial algorithm in convex programming*. Philadelphia (PA): SIAM; 1994.
8. Nesterov Y. *Introductory lectures on convex optimization*. Norwell (MA): Kluwer Academic Publishers; 2004.
9. Armand P, Benoist J, Orban D. Dynamic updates of the barrier parameter in primal–dual methods for nonlinear programming. *Comput Optim Appl* 2008;41(1):1–25.
10. Griva I, Shanno DF, Vanderbei RJ, *et al.* Global convergence of a primal–dual interior-point method for nonlinear programming. *Algorithm Oper Res* 2008;3(1):12–29.
11. Forsgren A, Gill P. Primal–dual interior methods for nonconvex nonlinear programming. *SIAM J Optim* 1998;8(4):1132–1152.
12. Wachter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program* 2006;106(1):25–57.
13. Waltz RA, Morales JL, Nocedal J, *et al.* An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Math Program* 2006;107:391–408.
14. Polyak RA. Modified barrier functions theory and methods. *Math Program* 1992;54:177–222.
15. Polyak R, Teboulle M. Nonlinear rescaling and proximal-like methods in convex optimization. *Math Program* 1997;76:265–284.
16. Polyak RA. Nonlinear rescaling vs. smoothing technique in convex optimization. *Math Program* 2002;92:197–235.
17. Yamashita H, Yabe H. Superlinear and quadratic convergence of some primal–dual interior point methods for constrained optimization. *Math Program* 1996;75:377–397.
18. Tseng P, Bertsekas D. On the convergence of exponential multipliers method for convex programming. *Math Program* 1993;60:1–19.
19. Ben-Tal A, Zibulevsky M. Penalty-barrier methods for convex optimization. *SIAM J Control Optim* 1997;7:347–366.

20. Griva I, Polyak RA. Primal-dual nonlinear rescaling method with dynamic scaling parameter update. *Math Program* 2006;106(2):237–259.
21. Polyak RA. Primal-dual exterior-point method for convex optimization. *Optim Method Softw* 2008;23(1):141–160.
22. Griva I, Polyak RA. 1.5-Q-superlinear convergence of an exterior-point method for constrained optimization. *J Glob Optim* 2008;40(4):679–695.

PROBABILISTIC DISTANCE CLUSTERING

CEM IYIGUN

Department of Industrial Engineering,
Middle East Technical University,
Ankara, Turkey

INTRODUCTION

A *cluster* is a set of data points (patterns, observations) that are similar in some sense and *clustering* is a process of organizing a collection of data items (data set) into disjoint clusters. The items within a cluster are more *similar* to each other than to the items in other clusters. The *similarity* term can be expressed in very different ways, based on the context and the purpose of the study [38].

Clustering is a basic tool in *statistics* [2] and *machine learning* [3], and has been applied in *pattern recognition*, *medical diagnostics*, *operations research* [4], *data mining* [5], biology, finance, and also in applied areas including genetics, taxonomy, medicine, marketing, and e-commerce (see Refs 6 and 7 for applications of clustering).

It is useful to begin by reviewing various issues in clustering analysis and then we introduce our notation and terminology that is used throughout the paper.

Generally, clustering is called *unsupervised learning* because unlike classification (known as *supervised learning*), it is performed when no prior information is available concerning the membership of data items for categorizing them to predefined classes. Besides, the traditional definition of clustering, in some recent studies called *semisupervised clustering*, a small amount of information is used to guide or adjust the clustering structure [8] and [9].

Clustering aims at exploring relationships among a collection of data points and organizing them into disjoint clusters. Two basic

measures, *similarity* in a cluster and *dissimilarity* between clusters, are considered while creating clusters. Similarity in a cluster, called *intraconnectivity*, measures the density of connections between data points of a single cluster. In a good clustering arrangement, since the data points within the same cluster are highly similar to each other, the value of the intraconnectivity will be high, whereas dissimilarity between clusters, called *interconnectivity*, measures the connectivity between distinct clusters. A low degree of interconnectivity is aimed at because it indicates that individual clusters are largely dissimilar to each other.

The objects of clustering are *data points* (also *observations*, or *instances*). Each data point is an ordered list of *attributes* (or *features*), such as height, age, and ID number. In a given collection of data points, named a *data set*, every data point is represented by using the same set of *attributes* (features). The attributes can be in the form of *continuous*, *categorical* or *binary* data [10].

The interpretation of clusters is a difficult task. The *number of clusters* is given, however in many applications, the “right” number of clusters (to fit best the data) needs to be determined. Even if there were no cluster structure in a given data set the algorithm will always assign the data points to clusters and form a clustering structure. Therefore, if the goal is to draw conclusions about the cluster structure in a data set, it is essential to analyze whether the data set has a clustering tendency. Another problem that one can face in clustering is the quality of the data. In many applications, there may be missing values or errors in the data set due to inaccurate measurement. So a strategy needs to be chosen for handling missing values, see Ref. 11. A preprocessing stage may also be needed to organize the data before the analysis. *Feature selection*, *aggregation*, and *dimensionality reduction* are some of the common preprocessing methods used.

Clustering is a difficult problem due to

many factors. How to process and prepare the data, which similarity measures are effective, which algorithm is appropriate, and how sensitive the algorithm output is to initial conditions are the questions that a clustering method needs to deal with [12]. These factors come into play in devising a well-tuned clustering technique for a given clustering problem. Moreover, there is no technique that is applicable to all sorts of cluster structures.

Clustering algorithms can be categorized into *partitioning (center-based)*, *hierarchical*, *density-based*, *grid-based*, and *model-based methods*. An excellent survey of clustering techniques can be found in Ref. 12. Some more recent works in clustering techniques can also be found in Ref. 13. In each of the clustering categories listed above, there are many papers in the literature that describe different algorithms and relevant work. For comprehensive explanations and further details, the reader can see Refs 14–16 for partitioning methods; Refs 2, 17, and 18 for hierarchical methods; Refs 19, 20 for density-based methods, and Refs 21, 22 for model-based methods.

Notation and Terminology

We take data points to be vectors $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, and interpret “similar” as “close,” in terms of a *distance* function $d(\mathbf{x}, \mathbf{y})$ in $\mathbb{R}^n$, such as

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad (1)$$

where the norm $\|\cdot\|$ is *elliptic*, defined for $\mathbf{u} = (u_i)$ by

$$\|\mathbf{u}\| = \langle \mathbf{u}, Q\mathbf{u} \rangle^{1/2}, \quad (2)$$

with $\langle \cdot, \cdot \rangle$ the standard inner product, and Q a positive definite matrix. In particular, $Q = I$ gives the *Euclidean norm*,

$$\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}, \quad (3)$$

and the *Mahalanobis distance* corresponds to $Q = \Sigma^{-1}$, where Σ is the covariance matrix of the data involved.

Example 1. A data set in $\mathbb{R}^2$ with $N = 200$ data points is shown in Figure 1. The data

was simulated, from normal distributions $N(\mu_i, \Sigma_i)$, with

$$\mu_1 = (0, 0), \Sigma_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}, (100 \text{ points}),$$

$$\mu_2 = (3, 0), \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, (100 \text{ points}).$$

This data will serve to illustrate Examples 2–5 below.

The *clustering problem* is, given a *data set* $\mathcal{D}$ consisting of N data points

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^n,$$

and an integer K , $1 < K < N$, to partition $\mathcal{D}$ into K clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$.

Data points are assigned to clusters using a *clustering criterion*. In *distance clustering*, abbreviated *d-clustering*, the clustering criterion is metric: with each cluster $\mathcal{C}_k$ we associate a *center* $\mathbf{c}_k$, for example its centroid, and each data point is assigned to the cluster to whose center it is the nearest. After each such assignment, the cluster centers may change, resulting in reassignments. Such an algorithm will therefore iterate between updating the centers and reassignments.

A commonly used clustering criterion is the sum-of-squares of Euclidean distances,

$$\sum_{k=1}^K \sum_{\mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad (4)$$

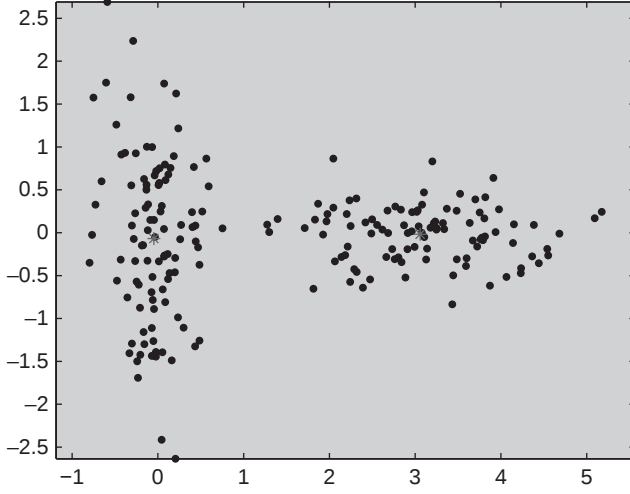
to be minimized by the sought clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$. The well known *k-means clustering algorithm* [14] uses this criterion.

In *probabilistic clustering* the assignment of points to clusters is “soft,” in the sense that the membership of a data point $\mathbf{x}$ in a cluster $\mathcal{C}_k$ is given as a *probability*, denoted by $p_k(\mathbf{x})$. These are subjective probabilities, indicating strength of belief in the event in question.

Let a *distance function*

$$d_k(\cdot, \cdot) \quad (5)$$

be defined for each cluster $\mathcal{C}_k$. These distance functions are, in general, different from one

Figure 1. A data set in $\mathbb{R}^2$.

cluster to another. For each data point $\mathbf{x} \in \mathcal{D}$, we then compute:

- the *distance* $d_k(\mathbf{x}, \mathbf{c}_k)$, also denoted by $d_k(\mathbf{x})$ (since d_k is used only for distances from $\mathbf{c}_k$), or just d_k if $\mathbf{x}$ is understood, and
- a *probability* that $\mathbf{x}$ is a member of C_k , denoted by $p_k(\mathbf{x})$, or just p_k .

Various relations between probabilities and distances can be assumed, resulting in different ways of clustering the data. In our experience, the following assumption has proved useful: for any point $\mathbf{x}$, and all $k = 1, \dots, K$

$$p_k(\mathbf{x}) d_k(\mathbf{x}) = \text{constant, depending on } \mathbf{x}.$$

This model is our working *principle* in what follows, and the basis of the *probabilistic d-clustering* [1], approach of the following section.

The above principle owes its versatility to the different ways of choosing the distances $d_k(\cdot)$. It is also natural to consider increasing functions of such distances, and one useful choice is

$$p_k(\mathbf{x}) e^{d_k(\mathbf{x})} = \text{constant, depending on } \mathbf{x},$$

giving the *probabilistic exponential d-clustering*.

The probabilistic d-clustering algorithm is presented in the following sections. It is a generalization, to several centers, of the Weiszfeld method for solving the Fermat–Weber location problem, and convergence follows as in Ref. 23. The updates of the centers use an extremal principle. The progress of the algorithm is monitored by the *joint distance function* (JDF), a distance function that captures the data in its low contours. The centers updated by the algorithm are stationary points of the joint distance function.

In the final part of the paper, a new function, *classification uncertainty function*, is introduced whose values indicate the uncertainty of classifying a data point to one of the clusters. The paper concludes with the problem of determining the “right” number of clusters in a given data set, using the classification of uncertainty function.

For other approaches to probabilistic clustering, see the surveys in Refs 11 and 24, and the seminal article [25] unifying clustering methods in the framework of modern optimization theory.

PROBABILISTIC D-CLUSTERING

There are several ways to model the relationship between distances and probabilities. The

simplest model, and our working *principle* (or axiom), is the following:

Principle 1. For each $\mathbf{x} \in \mathcal{D}$, and each cluster C_k ,

$$p_k(\mathbf{x})d_k(\mathbf{x}) = \text{constant, depending on } \mathbf{x}. \quad (6)$$

Cluster membership is thus more probable the closer the data point is to the cluster center. Note that the constant in Equation (6) is independent of the cluster k .

Probabilities

From Principle 1, and the fact that probabilities add to one, we get the following.

Theorem 1. Let the cluster centers $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K\}$ be given, let $\mathbf{x}$ be a data point, and let $\{d_k(\mathbf{x}) : k = 1, \dots, K\}$ be its distances from the given centers. Then the membership probabilities of $\mathbf{x}$ are

$$p_k(\mathbf{x}) = \frac{\prod_{j \neq k} d_j(\mathbf{x})}{\sum_{t=1}^K \prod_{j \neq t} d_j(\mathbf{x})}, \quad k = 1, \dots, K. \quad (7)$$

Proof. Using Equation (6) we write for t, k

$$p_t(\mathbf{x}) = \left(\frac{p_k(\mathbf{x})d_k(\mathbf{x})}{d_t(\mathbf{x})} \right).$$

Since $\sum_{t=1}^K p_t(\mathbf{x}) = 1$,

$$p_k(\mathbf{x}) \sum_{t=1}^K \left(\frac{d_k(\mathbf{x})}{d_t(\mathbf{x})} \right) = 1.$$

$$\therefore p_k(\mathbf{x}) = \frac{1}{\sum_{t=1}^K \left(\frac{d_k(\mathbf{x})}{d_t(\mathbf{x})} \right)} = \frac{\prod_{j \neq k} d_j(\mathbf{x})}{\sum_{t=1}^K \prod_{j \neq t} d_j(\mathbf{x})}.$$

In particular, for $K = 2$,

$$\begin{aligned} p_1(\mathbf{x}) &= \frac{d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}, \\ p_2(\mathbf{x}) &= \frac{d_1(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}, \end{aligned} \quad (8)$$

and for $K = 3$,

$$\begin{aligned} p_1(\mathbf{x}) &= \frac{d_2(\mathbf{x})d_3(\mathbf{x})}{d_1(\mathbf{x})d_2(\mathbf{x}) + d_1(\mathbf{x})d_3(\mathbf{x}) + d_2(\mathbf{x})d_3(\mathbf{x})}, \text{ etc.} \\ & \quad (9) \end{aligned}$$

Note: See Ref. 26 for related work in a different context. In particular, our Equation (8) is closely related to Ref. 26, Equation (5).

The Joint Distance Function

We denote the constant in Equation (6) by $D(\mathbf{x})$, a function of $\mathbf{x}$. Then

$$p_k(\mathbf{x}) = \frac{D(\mathbf{x})}{d_k(\mathbf{x})}, \quad k = 1, \dots, K.$$

Since the probabilities add to one we get,

$$D(\mathbf{x}) = \frac{\prod_{k=1}^K d_k(\mathbf{x}, \mathbf{c}_k)}{\sum_{t=1}^K \prod_{j \neq t} d_j(\mathbf{x}, \mathbf{c}_j)}. \quad (10)$$

The function $D(\mathbf{x})$, called the *JDF* of $\mathbf{x}$, has the dimension of distance, and measures the distance of $\mathbf{x}$ from all cluster centers. Here are special cases of Equation (10), for $K = 2$,

$$D(\mathbf{x}) = \frac{d_1(\mathbf{x})d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}, \quad (11)$$

and for $K = 3$,

$$D(\mathbf{x}) = \frac{d_1(\mathbf{x})d_2(\mathbf{x})d_3(\mathbf{x})}{d_1(\mathbf{x})d_2(\mathbf{x}) + d_1(\mathbf{x})d_3(\mathbf{x}) + d_2(\mathbf{x})d_3(\mathbf{x})}. \quad (12)$$

The JDF of the whole data set $\mathcal{D}$ is the sum of Equation (10) over all points, and is a

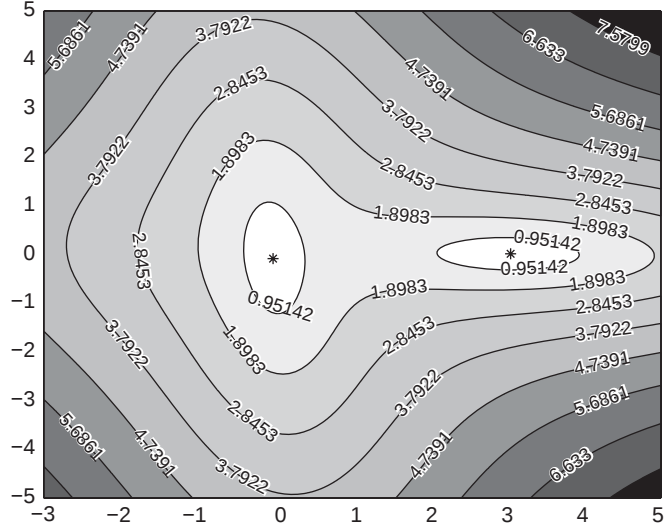


Figure 2. Level sets of a joint distance function.

function of the K cluster centers, say,

$$F(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K) = \sum_{i=1}^N \frac{\prod_{k=1}^K d_k(\mathbf{x}_i, \mathbf{c}_k)}{\sum_{t=1}^N \prod_{j \neq t} d_j(\mathbf{x}_i, \mathbf{c}_j)}. \quad (13)$$

Example 2. Figure 2 shows level sets of the JDF (Eq. 11), with Mahalanobis distances

$$d_k(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)}, \quad (14)$$

$\mathbf{c}_1 = \mu_1, \mathbf{c}_2 = \mu_2$, and Σ_1, Σ_2 as in Example 1.

Notes:

- (a) The JDF $D(\mathbf{x})$ of Equation (10) is a measure of the classifiability of the point $\mathbf{x}$ in question. It is zero if and only if $\mathbf{x}$ coincides with one of the cluster centers, in which case $\mathbf{x}$ belongs to that cluster with probability 1. If all the distances $d_k(\mathbf{x}, \mathbf{c}_k)$ are equal, say equal to d , then $D(\mathbf{x}) = d/k$ and all $p_k(\mathbf{x}) = 1/K$, showing indifference between the clusters. As the distances $d_k(\mathbf{x})$ increase, so does $D(\mathbf{x})$, indicating greater uncertainty about the cluster where $\mathbf{x}$ belongs.

- (b) The JDF (Eq. 10) is, up to a constant, the *harmonic mean* of the distances involved, see Ref. 27 for an elucidation of the role of the harmonic mean in contour approximation of data. A related concept in ecology is the *home range*, shown in Ref. 28 to be the harmonic mean of the area moments in question.

An Extremal Principle

For simplicity, consider the case of two clusters (the results are easily extended to the general case).

Let $\mathbf{x}$ be a given data point with distances $d_1(\mathbf{x}), d_2(\mathbf{x})$ to the cluster centers. Then the probabilities in Equation (8) are the optimal solutions p_1, p_2 of the extremal problem

$$\begin{aligned} &\text{Minimize} && d_1(\mathbf{x})p_1^2 + d_2(\mathbf{x})p_2^2 \\ &\text{subject to} && p_1 + p_2 = 1 \\ &&& p_1, p_2 \geq 0. \end{aligned} \quad (15)$$

Indeed, the *Lagrangian* of this problem is

$$\begin{aligned} L(p_1, p_2, \lambda) &= d_1(\mathbf{x})p_1^2 + d_2(\mathbf{x}) \\ &\quad \times p_2^2 - \lambda(p_1 + p_2 - 1), \end{aligned} \quad (16)$$

and setting the partial derivatives (with respect to p_1, p_2) equal to zero gives the

principle (Eq. 6)

$$p_1 d_1(\mathbf{x}) = p_2 d_2(\mathbf{x}).$$

Substituting the probabilities (Eq. 8) in the Lagrangian (Eq. 16) we get the optimal value of Equation (15)

$$L^*(p_1(\mathbf{x}), p_2(\mathbf{x}), \lambda) = \frac{d_1(\mathbf{x})d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}, \quad (17)$$

which is the JDF (Eq. 11) again.

The extremal problem for a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^n$ is, accordingly,

$$\begin{aligned} & \text{Minimize } \sum_{i=1}^N \left(d_1(\mathbf{x}_i) p_1(\mathbf{x}_i)^2 + d_2(\mathbf{x}_i) p_2(\mathbf{x}_i)^2 \right) \\ & \text{subject to } p_1(\mathbf{x}_i) + p_2(\mathbf{x}_i) = 1 \\ & \quad p_1(\mathbf{x}_i), p_2(\mathbf{x}_i) \geq 0, \quad i = 1, \dots, N. \end{aligned} \quad (18)$$

This problem separates into N problems like Equation (15), and its optimal value is

$$\sum_{i=1}^N \frac{d_1(\mathbf{x}_i) d_2(\mathbf{x}_i)}{d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)}, \quad (19)$$

the JDF (Eq. 13) of the data set, with $K = 2$. *Note:* The explanation for the strange appearance of “probabilities squared” above, is that Equation (15) is a smoothed version of the “real” clustering problem, namely,

$$\min \{d_1, d_2\},$$

which is nonsmooth; see Ref. 25 for a unified development of smoothed clustering methods.

Centers

We write Equation (18) as a function of the cluster centers $\mathbf{c}_1, \mathbf{c}_2$,

$$\begin{aligned} f(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^N & \left(d_1(\mathbf{x}_i, \mathbf{c}_1) p_1(\mathbf{x}_i)^2 \right. \\ & \left. + d_2(\mathbf{x}_i, \mathbf{c}_2) p_2(\mathbf{x}_i)^2 \right). \end{aligned} \quad (20)$$

If a point $\mathbf{x}_i$ coincides with a center, say $\mathbf{x}_i = \mathbf{c}_1$, then $d_1(\mathbf{x}_i) = 0$, $p_1(\mathbf{x}_i) = 1$, and $p_2(\mathbf{x}_i) = 0$. This point contributes zero to the summation.

For the special case of Euclidean distances, the minimizers of Equation (20) assume a simple form as convex combinations of the data points.

Theorem 2. *Let the distance functions d_1, d_2 in Equation (20) be Euclidean,*

$$d_k(\mathbf{x}, \mathbf{c}_k) = \|\mathbf{x} - \mathbf{c}_k\|, \quad k = 1, 2, \quad (21)$$

so that

$$\begin{aligned} f(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1, \dots, N} & \left(\|\mathbf{x}_i - \mathbf{c}_1\| p_1(\mathbf{x}_i)^2 \right. \\ & \left. + \|\mathbf{x}_i - \mathbf{c}_2\| p_2(\mathbf{x}_i)^2 \right), \end{aligned} \quad (22)$$

and let the probabilities $p_1(\mathbf{x}_i), p_2(\mathbf{x}_i)$ be given for $i = 1, \dots, N$. We make the following assumption about the minimizers $\mathbf{c}_1, \mathbf{c}_2$ of Equation (22):

$$\begin{aligned} & \mathbf{c}_1, \mathbf{c}_2 \text{ do not coincide with any of the points} \\ & \mathbf{x}_i, i = 1, \dots, N. \end{aligned} \quad (23)$$

Then the minimizers $\mathbf{c}_1, \mathbf{c}_2$ are given by

$$\mathbf{c}_k = \sum_{i=1, \dots, N} \left(\frac{u_k(\mathbf{x}_i)}{\sum_{j=1, \dots, N} u_k(\mathbf{x}_j)} \right) \mathbf{x}_i, \quad (24)$$

where

$$u_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)}, \quad (25)$$

for $k = 1, 2$, or equivalently, using Equation (8)

$$\begin{aligned} u_1(\mathbf{x}_i) &= \frac{d_2(\mathbf{x}_i, \mathbf{c}_2)^2}{d_1(\mathbf{x}_i, \mathbf{c}_1) (d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2))^2}, \\ u_2(\mathbf{x}_i) &= \frac{d_1(\mathbf{x}_i, \mathbf{c}_1)^2}{d_2(\mathbf{x}_i, \mathbf{c}_2) (d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2))^2}. \end{aligned} \quad (26)$$

Proof. The gradient of $d(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - \mathbf{c}\|$ with respect to $\mathbf{c}$ is, for $\mathbf{x} \neq \mathbf{c}$,

$$\nabla_{\mathbf{c}} \|\mathbf{x} - \mathbf{c}\| = -\frac{\mathbf{x} - \mathbf{c}}{\|\mathbf{x} - \mathbf{c}\|} = -\frac{\mathbf{x} - \mathbf{c}}{d(\mathbf{x}, \mathbf{c})}. \quad (27)$$

By Assumption (23), the gradient of Equation (22) with respect to $\mathbf{c}_k$ is

$$\begin{aligned} \nabla_{\mathbf{c}_k} f(\mathbf{c}_1, \mathbf{c}_2) &= - \sum_{i=1, \dots, N} \frac{\mathbf{x}_i - \mathbf{c}_k}{\|\mathbf{x}_i - \mathbf{c}_k\|} p_k(\mathbf{x}_i)^2 \\ &= - \sum_{i=1, \dots, N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d_k(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2, \\ &\quad k = 1, 2. \end{aligned} \quad (28)$$

Setting the gradient equal to zero, and summing like terms, we get

$$\begin{aligned} &\sum_{i=1, \dots, N} \left(\frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} \right) \\ &= \mathbf{x}_i \left(\sum_{i=1, \dots, N} \frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i, \mathbf{c}_k)} \right) \mathbf{c}_k, \end{aligned} \quad (29)$$

proving Equations (24)–(26).

The same formulas for the centers $\mathbf{c}_1, \mathbf{c}_2$ hold if the norm used in Equation (22) is elliptic.

Corollary 1. *Let the distance functions d_1, d_2 in Equation (20) be elliptic*

$$d_k(\mathbf{x}, \mathbf{c}_k) = \langle (\mathbf{x} - \mathbf{c}_k), \mathbf{Q}_k(\mathbf{x} - \mathbf{c}_k) \rangle^{1/2}, \quad (30)$$

with positive-definite matrices $\mathbf{Q}_k$. Then the minimizers $\mathbf{c}_1, \mathbf{c}_2$ of Equation (20) are given by Equations (24)–(26).

Proof. The gradient of $d(\mathbf{x}, \mathbf{c}) = \langle (\mathbf{x} - \mathbf{c}), \mathbf{Q}(\mathbf{x} - \mathbf{c}) \rangle^{1/2}$ with respect to $\mathbf{c}$ is, for $\mathbf{x} \neq \mathbf{c}$,

$$\nabla_{\mathbf{c}} d(\mathbf{x}, \mathbf{c}) = -\frac{\mathbf{Q}(\mathbf{x} - \mathbf{c})}{d(\mathbf{x}, \mathbf{c})}.$$

Therefore, the analog of Equation (28) is

$$\nabla_{\mathbf{c}_k} f(\mathbf{c}_1, \mathbf{c}_2) = -\mathbf{Q}_k \sum_{i=1, \dots, N} \frac{\mathbf{x}_i - \mathbf{c}_k}{d_k(\mathbf{x}_i, \mathbf{c}_k)} p_k(\mathbf{x}_i)^2, \quad (31)$$

and since $\mathbf{Q}_k$ is nonsingular, it can be “cancelled” when we set the gradient equal to zero. The rest of the proof is as in Theorem 2.

Corollary 1 applies, in particular, to the Mahalanobis distance (Eq. 14)

$$d_k(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)},$$

where Σ_k is the covariance matrix of the cluster $\mathcal{C}_k$.

The formulas (24) and (25) are also valid in the general case of K clusters, where the analog of Equation (20) is

$$f(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K) = \sum_{i=1, \dots, N} \sum_{k=1}^K d_k(\mathbf{x}_i, \mathbf{c}_k) p_k(\mathbf{x}_i)^2. \quad (32)$$

Corollary 2. *Let the distance functions d_k in Equation (32) be elliptic, as in Equation (30), and let the probabilities $p_k(\mathbf{x}_i)$ be given. Then, the minimizers $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K$ of Equation (32) are given by Equations (24) and (25) for $k = 1, 2, \dots, K$.*

Proof. The proof of Corollary 1 holds in the general case, since the minimizers are calculated separately.

The Weiszfeld Method

In the case of one cluster (where the probabilities are all 1 and therefore of no interest), the center formulas (24) and (25) reduce to

$$\mathbf{c} = \sum_{i=1, \dots, N} \left(\frac{1/d(\mathbf{x}_i, \mathbf{c})}{\sum_{j=1, \dots, N} 1/d(\mathbf{x}_j, \mathbf{c})} \right) \mathbf{x}_i, \quad (33)$$

giving the minimizer of $f(\mathbf{c}) = \sum_{i=1}^N d(\mathbf{x}_i, \mathbf{c})$. Formula (33) can be used iteratively to update the center $\mathbf{c}$ (on the left) as a convex combination of the points $\mathbf{x}_i$ with weights depending on the current center. This iteration is the *Weiszfeld method* [29] for solving the Fermat–Weber location problem, see Refs 29 and 30. Convergence

of Weiszfeld's method was established in Kuhn [23] by modifying the gradient $\nabla f(\mathbf{c})$ so that it is always defined, see Ref. 31 for further details. However, the modification is not carried out in practice since, as shown by Kuhn, the set of initial points $\mathbf{c}$ for which it ever becomes necessary is denumerable.

In what follows, we use the formulas (24) and (25) iteratively to update the centers. Convergence can be proved by adapting the arguments of Kuhn [23], but as before, it requires no special steps in practice.

The Centers and The Joint Distance Function

The centers given by Equations (24) and (25) are related to the JDF (Eq. 13) of the data set. Consider first the case of $K = 2$ clusters, where (Eq. 13) reduces to

$$F(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^N \frac{d_1(\mathbf{x}_i, \mathbf{c}_1) d_2(\mathbf{x}_i, \mathbf{c}_2)}{d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2)}. \quad (34)$$

The points $\mathbf{c}_k$ where $\nabla_{\mathbf{c}_k} F(\mathbf{c}_1, \mathbf{c}_2) = \mathbf{0}$, $k = 1, 2$, are called *stationary points* of Equation (34).

Theorem 3. *Let the distances d_1, d_2 in Equation (34) be elliptic, as in Equation (30). Then the stationary points of $F(\mathbf{c}_1, \mathbf{c}_2)$ are given by Equations (24)–(26).*

Proof. Let the distances d_k be Euclidean, $d_k(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_k\|$. It is enough to prove the theorem for one center, say $\mathbf{c}_1$. Using Equation (27) we derive

$$\begin{aligned} \nabla_{\mathbf{c}_1} F(\mathbf{c}_1, \mathbf{c}_2) &= \sum_{i=1}^N \frac{(d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)) d_2(\mathbf{x}_i) \left(-\frac{\mathbf{x}_i - \mathbf{c}_1}{d_1(\mathbf{x}_i)} \right) + d_1(\mathbf{x}_i) d_2(\mathbf{x}_i) \left(\frac{\mathbf{x}_i - \mathbf{c}_1}{d_1(\mathbf{x}_i)} \right)}{(d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2} \\ &= \sum_{i=1}^N \frac{-d_2(\mathbf{x}_i)^2 (\mathbf{x}_i - \mathbf{c}_1)}{d_1(\mathbf{x}_i) (d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2}. \end{aligned} \quad (35)$$

Setting Equation (35) equal to zero, and summing like terms, we get

$$\begin{aligned} &\left(\sum_{j=1}^N \frac{d_2(\mathbf{x}_j)^2}{d_1(\mathbf{x}_j) (d_1(\mathbf{x}_j) + d_2(\mathbf{x}_j))^2} \right) \mathbf{c}_1 \\ &= \sum_{i=1}^N \left(\frac{d_2(\mathbf{x}_i)^2}{d_1(\mathbf{x}_i) (d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2} \right) \mathbf{x}_i, \end{aligned}$$

duplicating Equations (24)–(26). If the distances are elliptic, as in Equation (30), then the analog of Equation (35) is

$$\nabla_{\mathbf{c}_1} F(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^N \frac{-d_2(\mathbf{x}_i)^2 Q_1(\mathbf{x}_i - \mathbf{c}_1)}{d_1(\mathbf{x}_i) (d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i))^2},$$

and since Q_1 is nonsingular, it can be “cancelled” when the gradient is set equal to zero.

In the above proof, the stationary points $\mathbf{c}_1, \mathbf{c}_2$ are calculated separately, and the calculation does not depend on there being 2 clusters. We thus have the following corollary.

Corollary 3. *Consider a data set with K clusters, and elliptic distances d_k . Then the stationary points of the JDF (Eq. 13) are the centers $\mathbf{c}_k$ given by Equations (24) and (25).*

Note: The JDF (10) is zero exactly at the K centers $\{\mathbf{c}_k\}$, and is positive elsewhere. These centers are therefore the global minimizers of Equation (10). However, the function (Eq. 10) is not convex, not even quasi-convex, and may have other stationary points that are necessarily saddle points.

Why d and Not d^2 ?

The extremal principle (Eq. 18), which is the basis of our work, is linear in the distances d_k ,

$$\text{Minimize} \quad \sum_k d_k p_k^2.$$

We refer to this as the *d-model*.

In clustering, and statistics in general, it is customary to use the distances squared in the objective function,

$$\text{Minimize} \quad \sum_k d_k^2.$$

We call this the d^2 -model.

The d^2 -model has a long tradition, dating back to Gauss, and is endowed with a rich statistical theory. There are geometrical advantages (Pythagoras Theorem), as well as analytical (linear derivatives).

The d -model is suggested by the analogy between clustering and location problems, where sums of distances (not distances squared) are minimized. Our center formulas (Eqs 24 and 25) are thus generalizations of the Weiszfeld method to several facilities.

An advantage of the d -model is its robustness. Indeed formula (25), which does not follow from the d^2 -model, guarantees that outliers will not affect the center locations.

Other Principles

There are alternative ways of modeling the relations between distances and probabilities. For example:

Principle 2. For each $\mathbf{x} \in \mathcal{D}$, and each cluster C_k , the probability $p_k = p_k(\mathbf{x})$ and distance $d_k = d_k(\mathbf{x})$ are related by

$$p_k^\alpha d_k^\beta = \text{constant, depending on } \mathbf{x}, \quad (36)$$

where the exponents α, β are positive.

For the case of two clusters we get, by analogy with Equations (8) and (18) respectively, the probabilities

$$\begin{aligned} p_1(\mathbf{x}) &= \frac{d_2(\mathbf{x})^{\beta/\alpha}}{d_1(\mathbf{x})^{\beta/\alpha} + d_2(\mathbf{x})^{\beta/\alpha}}, \\ p_2(\mathbf{x}) &= \frac{d_1(\mathbf{x})^{\beta/\alpha}}{d_1(\mathbf{x})^{\beta/\alpha} + d_2(\mathbf{x})^{\beta/\alpha}}, \end{aligned} \quad (37)$$

and an extremal principle,

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^N \left(d_1(\mathbf{x}_i)^\beta p_1(i)^{\alpha+1} + d_2(\mathbf{x}_i)^\beta p_2(i)^{\alpha+1} \right) \\ \text{subject to} \quad & p_1(i) + p_2(i) = 1 \\ & p_1(i), p_2(i) \geq 0, \end{aligned} \quad (38)$$

where $p_1(i), p_2(i)$ are the cluster probabilities at $\mathbf{x}_i$.

The *Fuzzy Clustering Method* [16,32], which is an extension of k -means method, uses $\beta = 2$ and allows different choices of α . For $\alpha = 2$, it gives the same probabilities as Equation (7), however the center updates are different from Equations (24) and (25).

PROBABILISTIC EXPONENTIAL D-CLUSTERING

Any increasing function of the distance can be used in Principle 1. The following model, with probabilities decaying exponentially as distances increase, has proved useful in our experience.

Principle 3. For each $\mathbf{x} \in \mathcal{D}$, and each cluster C_k , the probability $p_k(\mathbf{x})$ and distance $d_k(\mathbf{x})$ are related by

$$p_k(\mathbf{x}) e^{d_k(\mathbf{x})} = E(\mathbf{x}), \text{ a constant depending on } \mathbf{x}. \quad (39)$$

Most results of the section titled Probabilistic D-Clustering also hold for Principle 3, with the distance $d_k(\mathbf{x})$ replaced by $e^{d_k(\mathbf{x})}$. Thus the analog of the probabilities (Eq. 8) is

$$\begin{aligned} p_1(\mathbf{x}) &= \frac{e^{d_2(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}, \\ p_2(\mathbf{x}) &= \frac{e^{d_1(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}, \end{aligned} \quad (40)$$

or equivalently

$$\begin{aligned} p_1(\mathbf{x}) &= \frac{e^{-d_1(\mathbf{x})}}{e^{-d_1(\mathbf{x})} + e^{-d_2(\mathbf{x})}}, \\ p_2(\mathbf{x}) &= \frac{e^{-d_2(\mathbf{x})}}{e^{-d_1(\mathbf{x})} + e^{-d_2(\mathbf{x})}}. \end{aligned} \quad (41)$$

Similarly, since the probabilities add to 1, the constant in Equation (39) is

$$E(\mathbf{x}) = \frac{e^{d_1(\mathbf{x})+d_2(\mathbf{x})}}{e^{d_1(\mathbf{x})} + e^{d_2(\mathbf{x})}}, \quad (42)$$

also called the *exponential JDF*.

An Extremal Principle

The probabilities (Eq. 40) are the optimal solutions of the problem

$$\min_{p_1, p_2} \left\{ e^{d_1} p_1^2 + e^{d_2} p_2^2 : p_1 + p_2 = 1, \right. \\ \left. p_1, p_2 \geq 0 \right\}, \quad (43)$$

whose optimal value, obtained by substituting the probabilities (Eq. 40), is again the exponential JDF of Equation (42).

The extremal problem for a data set $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \subset \mathbb{R}^n$, partitioned into two clusters, is the following analog of Equation (18):

$$\text{Minimize } \sum_{i=1}^N \left(e^{d_1(\mathbf{x}_i)} p_1(i)^2 + e^{d_2(\mathbf{x}_i)} p_2(i)^2 \right) \\ \text{subject to } p_1(i) + p_2(i) = 1 \\ p_1(i), p_2(i) \geq 0, \quad (44)$$

where $p_1(i), p_2(i)$ are the cluster probabilities at $\mathbf{x}_i$. The problem separates into N problems like Equation (43), and its optimal value is

$$\sum_{i=1}^N \frac{e^{d_1(\mathbf{x}_i) + d_2(\mathbf{x}_i)}}{e^{d_1(\mathbf{x}_i)} + e^{d_2(\mathbf{x}_i)}}, \quad (45)$$

the exponential JDF of the whole data set.

Alternatively, Equation (39) follows from the “smoothed” extremal principle

$$\min_{p_1, p_2} \left\{ \sum_{k=1}^2 p_k d_k + \sum_{k=1}^2 p_k \log p_k \right. \\ \left. : p_1 + p_2 = 1, p_1, p_2 \geq 0 \right\}, \quad (46)$$

obtained by adding an entropy term to $\sum p_k d_k$. Indeed the Lagrangian of Equation (46) is

$$L(p_1, p_2, \lambda) = \sum_{k=1}^2 p_k d_k + \sum_{k=1}^2 p_k \log p_k \\ - \lambda (p_1 + p_2 - 1).$$

Differentiation with respect to p_k , and equating to 0, gives

$$d_k + 1 + \log p_k - \lambda = 0,$$

which is Equation (39).

Centers

We write Equation (44) as a function of the cluster centers $\mathbf{c}_1, \mathbf{c}_2$,

$$f(\mathbf{c}_1, \mathbf{c}_2) \\ = \sum_{i=1}^N \left(e^{d_1(\mathbf{x}_i, \mathbf{c}_1)} p_1(\mathbf{x}_i)^2 + e^{d_2(\mathbf{x}_i, \mathbf{c}_2)} p_2(\mathbf{x}_i)^2 \right), \quad (47)$$

and for elliptic distances we can verify, as in Theorem 2, that the minimizers of Equation (47) are given by

$$\mathbf{c}_k = \sum_{i=1}^N \left(\frac{u_k(\mathbf{x}_i)}{\sum_{j=1}^N u_k(\mathbf{x}_j)} \right) \mathbf{x}_i, \quad (48)$$

where [compare with Equation (25)],

$$u_k(\mathbf{x}_i) = \frac{p_k(\mathbf{x}_i)^2 e^{d_k(\mathbf{x}_i)}}{d_k(\mathbf{x}_i)}, \quad (49)$$

or equivalently,

$$u_1(\mathbf{x}_i) = \frac{e^{-d_1(\mathbf{x}_i)}/d_1(\mathbf{x}_i)}{(e^{-d_1(\mathbf{x}_i)} + e^{-d_2(\mathbf{x}_i)})^2}, \\ u_2(\mathbf{x}_i) = \frac{e^{-d_2(\mathbf{x}_i)}/d_2(\mathbf{x}_i)}{(e^{-d_1(\mathbf{x}_i)} + e^{-d_2(\mathbf{x}_i)})^2}. \quad (50)$$

As in Theorem 3, these minimizers are the stationary points of the JDF, given here as

$$F(\mathbf{c}_1, \mathbf{c}_2) = \sum_{i=1}^N \frac{e^{d_1(\mathbf{x}_i, \mathbf{c}_1) + d_2(\mathbf{x}_i, \mathbf{c}_2)}}{e^{d_1(\mathbf{x}_i, \mathbf{c}_1)} + e^{d_2(\mathbf{x}_i, \mathbf{c}_2)}}. \quad (51)$$

Finally we can verify, as in Corollary 2, that the results hold in the general case of K clusters.

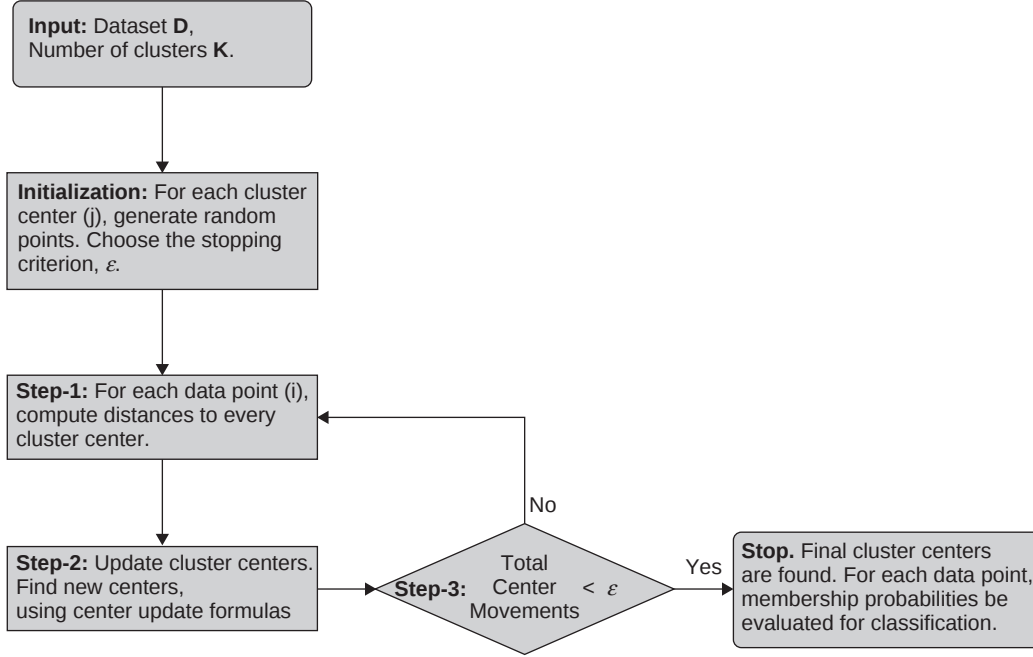


Figure 3. Flowchart of the probabilistic distance clustering algorithm.

A PROBABILISTIC D-CLUSTERING ALGORITHM

The ideas in the sections titled “*Probabilistic D-Clustering*” and “*Probabilistic Exponential D-Clustering*” are implemented in the

following algorithm for unsupervised clustering of data (Fig. 3). Steps of the algorithm are explained in detailed below and Algorithm 1 presents those steps—for simplicity—for the case of two clusters.

Algorithm 1. Steps of a probabilistic d-clustering algorithm

Initialization: given data $\mathcal{D}$, any two points $\mathbf{c}_1, \mathbf{c}_2$, and $\epsilon > 0$

Iteration:

Step 1 **compute** distances $d_1(\mathbf{x}), d_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}$
 Step 2 **update** the centers $\mathbf{c}_1^+, \mathbf{c}_2^+$
 Step 3 **if** $\|\mathbf{c}_1^+ - \mathbf{c}_1\| + \|\mathbf{c}_2^+ - \mathbf{c}_2\| < \epsilon$ **stop**
 return to step 1

The algorithm starts with random initial centers (number of clusters are given by the user). In *Step-1*, the distances of the data points to each center are computed. In *Step-2*, the centers are updated with the distances computed in Step-1 using Equations (24) or (48). The algorithm iterates between the cluster centers and the distances of the data points to these centers. The computations stop (in *Step 3*) when the centers stop moving.

The distance used in *Step 1* can be Euclidean or elliptic (the formulas 24–26, and 48–50, are valid in both cases). In *Step 2*, the centers are updated by Equations (24)–(26) if Principle 1 is used, and by Equations (48)–(50) for Principle 3. In particular, if the Mahalanobis distance (Eq. 14)

$$d(\mathbf{x}, \mathbf{c}_k) = \sqrt{(\mathbf{x} - \mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{c}_k)}$$

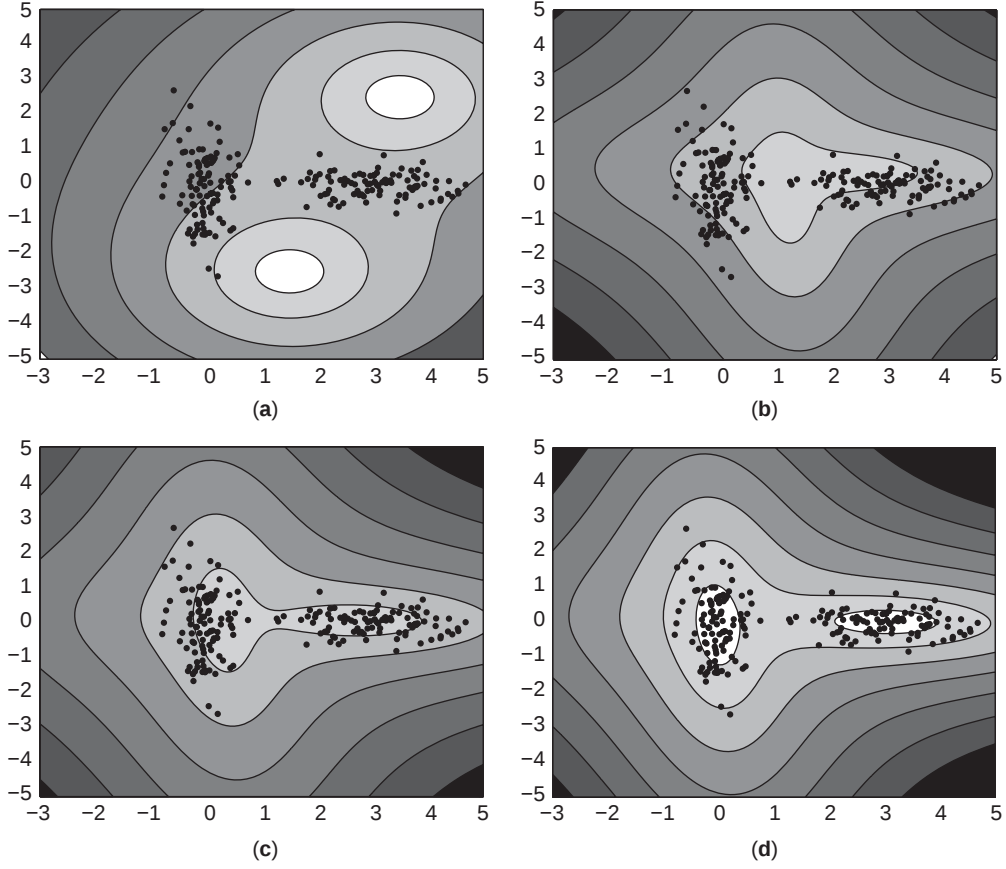


Figure 4. The level sets of the evolving *joint distance function* at (a) iteration 0, (b) iteration 2, (c) iteration 5, and (d) iteration 14.

is used, the covariance matrix Σ_k of the k th-cluster, can be estimated at each iteration by

$$\Sigma_k = \frac{\sum_{i=1}^N u_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T}{\sum_{i=1}^N u_k(\mathbf{x}_i)}, \quad (52)$$

with $u_k(\mathbf{x}_i)$ given by Equation (26) or Equation (50).

The cluster membership *probabilities* of the data points are not used explicitly in the algorithm. When the computations stop in (Step-3), the cluster membership probabilities can be computed by Equation (8) or Equation (40). These probabilities are not needed in the algorithm but they can be used

for classifying the data points after computing the cluster centers.

Using the arguments of Ref. 23, it can be shown that the objective function (Eq. 32) decreases at each iteration, and the algorithm converges. The cluster centers and distance functions change at each iteration, and so does the *JDF function* (Eq. 13) itself, which decreases at each iteration. The JDF may have stationary points that are not minimizers, however such points are necessarily saddle points and will be missed by the algorithm with probability 1.

Example 3. We apply *Algorithm 1*, using Mahalanobis distance, to the data of Example 1. Figure 4 shows the evolution of the JDF, represented by its level sets.

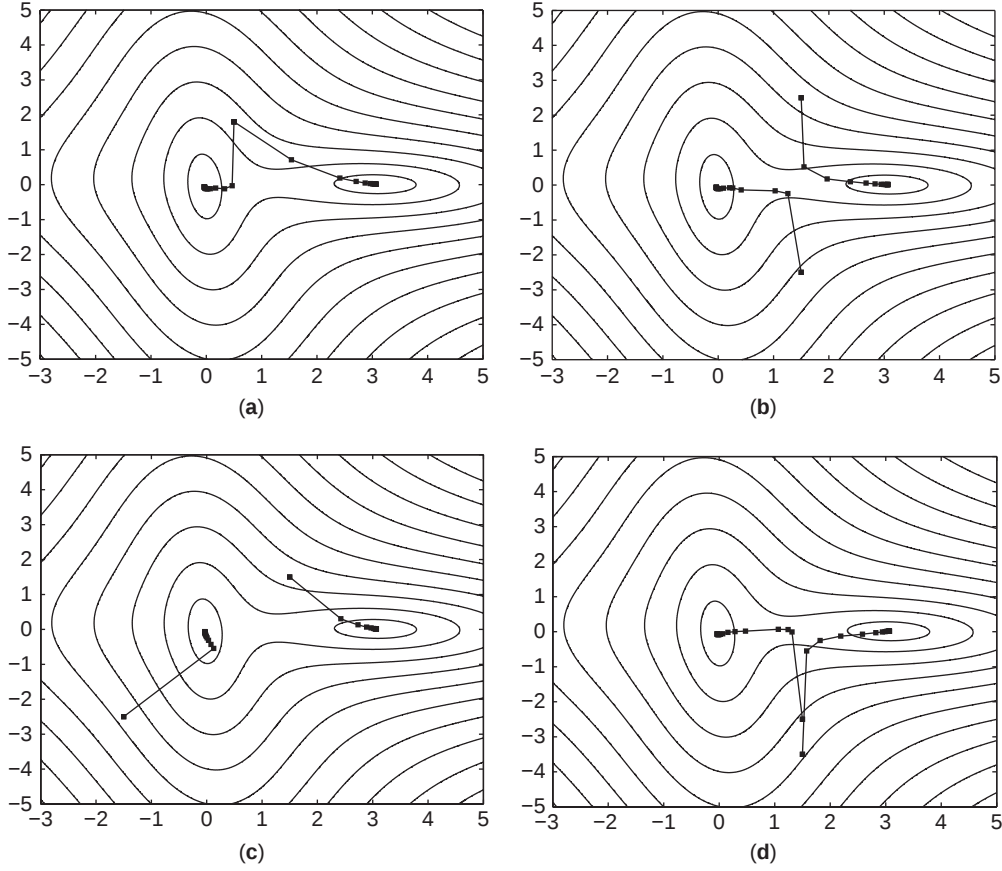


Figure 5. Movements of the cluster centers for different starts. (a) shows the centers corresponding to Fig. 4. (b) shows very close initial centers.

The initial function, shown in Fig. 5a, corresponds to the (arbitrarily chosen) initial centers and initial covariances $\Sigma_1 = \Sigma_2 = I$. The covariances are updated at each iteration using Equation (52), and by iteration 8 the function is already very close to its final form, as shown in Fig. 5d. For a tolerance of $\epsilon = 0.01$, the algorithm terminated in 12 iterations.

Example 4. In Fig. 5 we illustrate the movement of the cluster centers for different initial centers. The centers at each run are shown with the final level sets of the JDF found in Example 3.

The algorithm gives the correct cluster centers, for all initial starts. In particular,

the two initial centers may be arbitrarily close, as shown in Fig. 5b.

Example 5. The class membership probabilities (Eq. 8) were then computed using the centers determined by the algorithm. The level sets of the probability $p_1(\mathbf{x})$ are shown in Fig. 6. The curve $p_1(\mathbf{x}) = 0.5$, the thick curve shown in Fig. 6a, may serve as the clustering rule. Alternatively, the two clusters can be defined as

$$C_1 = \{\mathbf{x} : p_1(\mathbf{x}) \geq 0.6\}, C_2 = \{\mathbf{x} : p_1(\mathbf{x}) \leq 0.4\},$$

with points $\{\mathbf{x} : 0.4 < p_1(\mathbf{x}) < 0.6\}$ left unclassified, as in Fig. 6b.

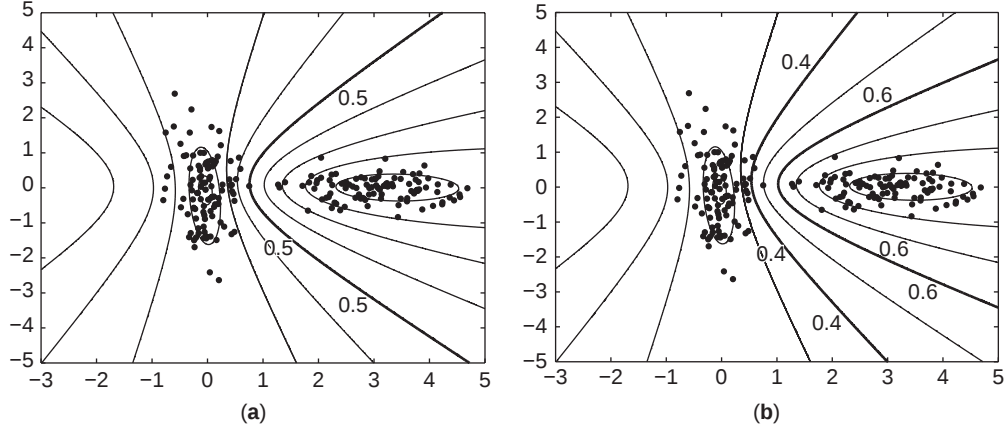


Figure 6. The level sets of the probabilities $p_1(\mathbf{x})$ and two clustering rules.

THE CLASSIFICATION UNCERTAINTY FUNCTION

The JDF has the dimension of distance. Normalizing it, we get the dimensionless function

$$U(\mathbf{x}) = K D(\mathbf{x}) / \left(\prod_{j=1}^K d_j(\mathbf{x}) \right)^{1/K}, \quad (53)$$

with $0/0$ interpreted as zero. Using Equations (7) and (10), $U(\mathbf{x})$ can be written as the geometric mean of the probabilities (up

to a constant)

$$U(\mathbf{x}) = K \left(\prod_{j=1}^K p_j(\mathbf{x}) \right)^{1/K}. \quad (54)$$

In particular, for $K = 2$,

$$U(\mathbf{x}) = 2 \frac{\sqrt{d_1(\mathbf{x})d_2(\mathbf{x})}}{d_1(\mathbf{x}) + d_2(\mathbf{x})} = 2\sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}. \quad (55)$$

The function $U(\mathbf{x})$ in Equation (53) is the harmonic mean of the distances $\{d_j(\mathbf{x})\}$ divided by their geometric mean. It follows

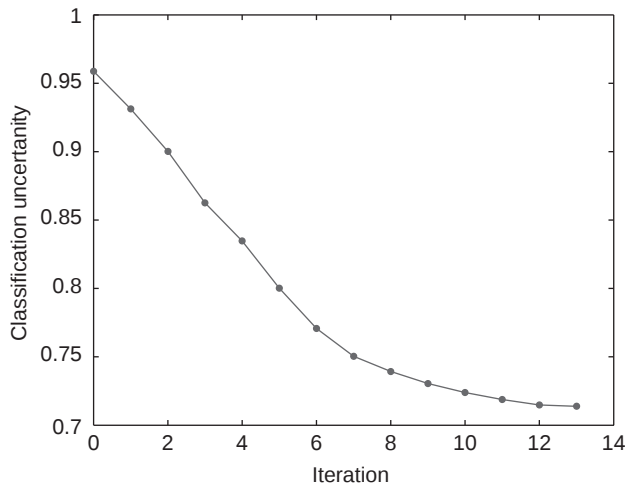


Figure 7. Illustration of Example 1: the decrease of classification uncertainty.

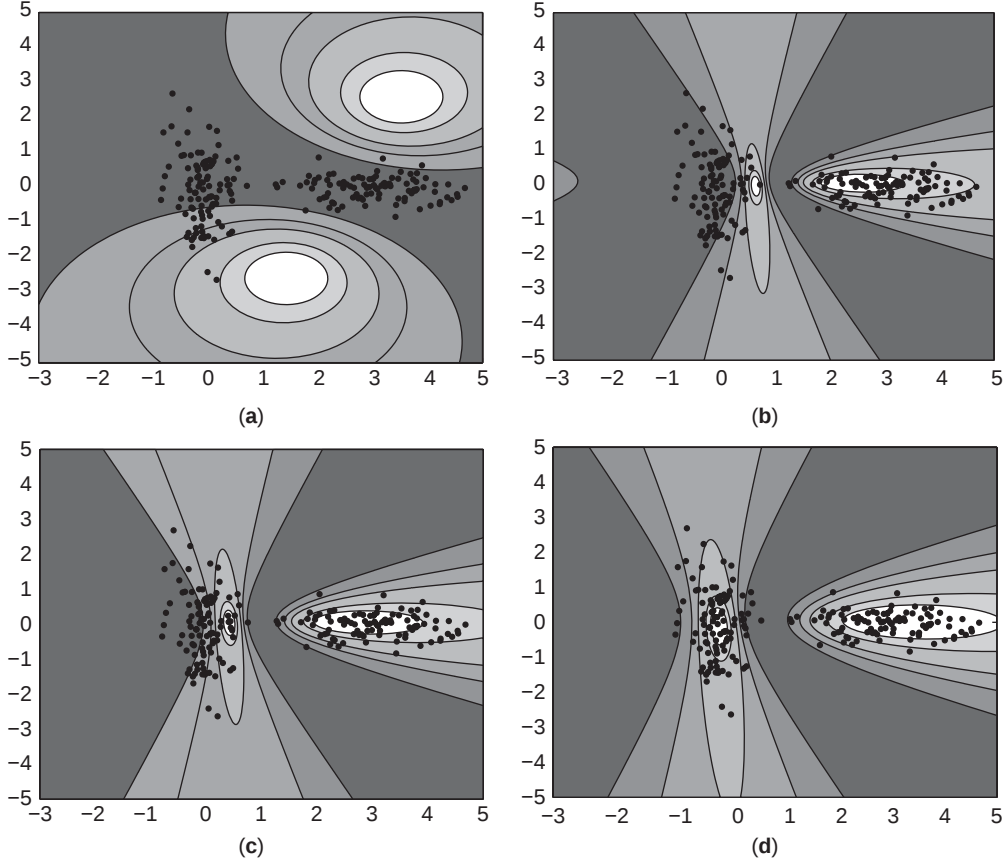


Figure 8. The level sets of the evolving classification uncertainty function at (a) iteration 0, (b) iteration 2, (c) iteration 3, and (d) iteration 14.

that $0 \leq U(\mathbf{x}) \leq 1$, with $U(\mathbf{x}) = 0$ if any $d_j(\mathbf{x}) = 0$, that is if $\mathbf{x}$ is a cluster center, and $U(\mathbf{x}) = 1$ if and only if the probabilities $p_j(\mathbf{x})$ are all equal. The values of $U(\mathbf{x})$ thus indicate the uncertainty of classifying the point $\mathbf{x}$ to one of the clusters. We call $U(\mathbf{x})$ the *classification uncertainty function*, abbreviated *CUF*, at $\mathbf{x}$. See Ref. 9 for a discussion of the CUF, and a justification for interpreting it as an entropic measure of uncertainty.

The CUF of the data set $\mathcal{D} = \{\mathbf{x}_i : i \in \overline{1, N}\} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \dots \cup \mathcal{C}_K$ with K clusters is

defined as

$$U_K(\mathcal{D}) := \frac{1}{N} \sum_{i=1}^N U(\mathbf{x}_i). \quad (56)$$

By definition, $U_1(\mathcal{D}) = 1$ since all probabilities are trivially 1 if there is just one cluster. At the other extreme, if every point is a separate cluster, it then has a zero probability of belonging to the other clusters, and therefore $U_N(\mathcal{D}) = 0$.

The classification uncertainty of the data set, measured by the CUF of the data set in Example 1, decreases at each iteration, converging to a residual uncertainty of about 0.7 (Fig. 7). This decrease is illustrated in Fig. 8, showing the level sets of the CUF

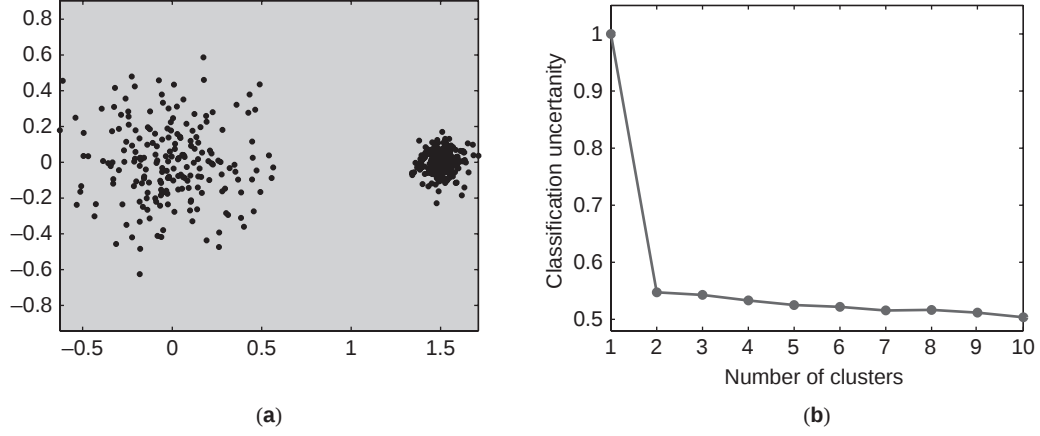


Figure 9. Results for Example 6.(a) A data set with $K = 2$ clusters and (b)the CUF as a function of K .

in Equation (54), with data points gradually “moving” into the level sets of lower uncertainty.

CLUSTERING VALIDATION

An important issue in clustering is to determine the “right” number K of clusters that fits a given data set with N points. In dichotomous situations the answer $K = 2$ is obvious, but in General, the answer lies between the two extremes of $K = 1$ (one cluster fits all), and $K = N$ (each point is a cluster). The CUF

of the data set, $U_K(\mathcal{D})$ given in Equation (56), helps resolve this issue [39].

Indeed, the value of $U_K(\mathcal{D})$ decreases monotonically with K , from $U_1(\mathcal{D}) = 1$ down to $U_N(\mathcal{D}) = 0$. The decrease of the uncertainty $U_K(\mathcal{D})$ is precipitous until the “right” K is reached, and after that the graph is almost flat. This suggests clustering the data set in question into $K = 1, 2, 3, \dots$ clusters, stopping at a value of K where the rate of decrease of $U_K(\mathcal{D})$ drops below a certain tolerance.

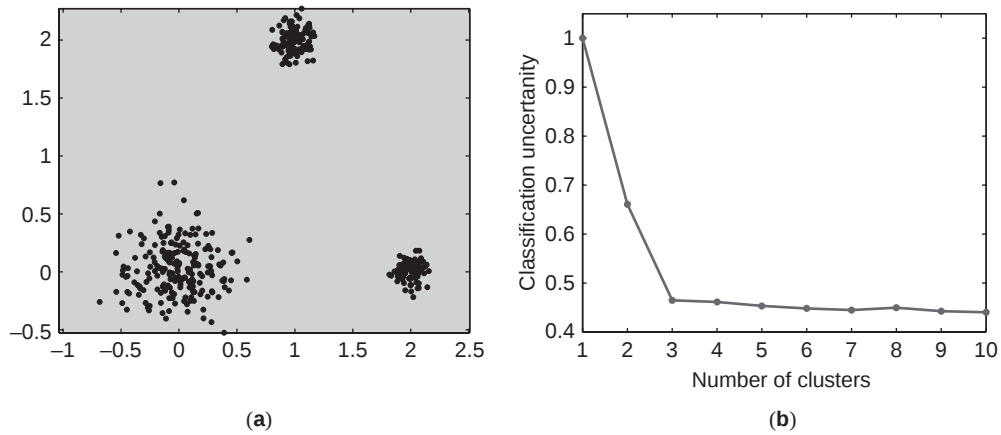


Figure 10. Note the change of slope of the CUF at $K = 3$. (a) A data set with $K = 3$ clusters and (b) the CUF as a function of K .

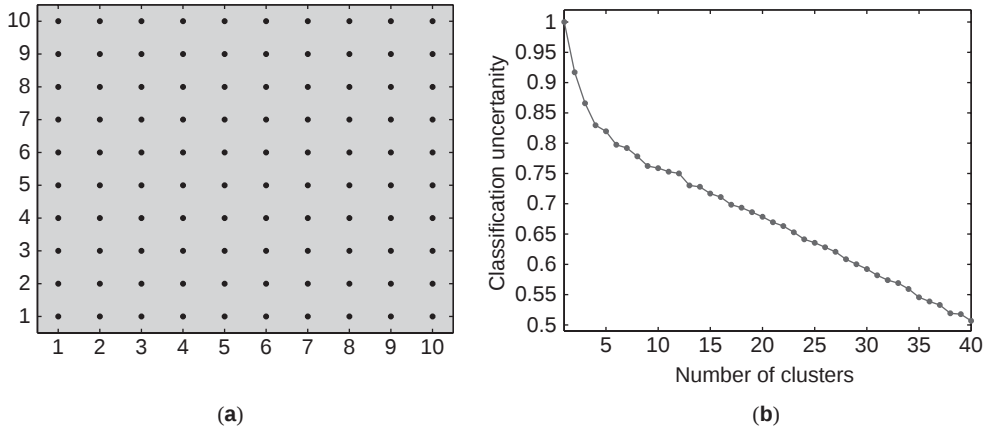


Figure 11. The slope of the CUF changes continuously. (a) A data set with no structure and (b) the CUF as a function of K .

This approach is practical because the d -clustering algorithm is fast, and it generates the information needed for computing the criterion (the value of the CUF), unlike other approaches that require external criteria. See also the survey in Ref. 33.

Example 6. Figure 9a shows a data set with two clusters. The d -clustering algorithm was applied to this data set, and the values of the $U_K(\mathcal{D})$ were computed for K from 1 to 10, the results are plotted in Fig. 9b. Note the drop from $U_1(\mathcal{D}) = 1$ to $U_2(\mathcal{D}) = 0.55$, and the change of slope at $K = 2$.

Figure 10a similarly shows a data set with $K = 3$ clusters. The CUF is plotted in Fig. 10b, revealing the correct number of clusters. Indeed, $U_K(\mathcal{D})$ drops from $U_2(\mathcal{D}) = 0.66$ to $U_3(\mathcal{D}) = 0.47$, and the rate of decrease is much smaller for $K > 3$.

If there is no clear answer for the number of clusters, as in Fig. 11a, then the CUF does not give an answer (Fig. 11b).

RELATED WORK

The data is given as a similarity matrix in many applications. For the analysis of such data using probabilistic distance clustering, see Ref. 34.

There are applications where the *cluster sizes* (ignored here) need to be estimated.

An important example is parameter estimation in mixtures of distributions. The above method, adjusted for cluster sizes, is applicable, and in particular presents a viable alternative to the Expectation-Maximization (EM) method, see Refs 35 and 36.

Our method allows an extension of the classical Weiszfeld method to several facilities. This is the subject of Ref. 37, giving the solution of *multifacility location problems*, including the capacitated case (which corresponds to given cluster sizes).

Semisupervised clustering is a framework for reconciling *supervised learning*, using any prior information (“labels”) on the data, with *unsupervised clustering*, based on the intrinsic properties and geometry of the data set. A new method for semisupervised clustering, combining probabilistic distance clustering for the unlabeled data points and a least squares criterion for the labeled ones, is given in Ref. 9.

REFERENCES

1. Ben-Israel A, Iyigun C. Probabilistic D-clustering. *J Classif* 2008;25:5–26.
2. Kaufman L, Rousseeuw P. Finding groups in data: an introduction to cluster analysis. New York: John Wiley & Sons, Inc.; 1990.

3. Fisher D. Knowledge acquisition via incremental conceptual clustering. *Mach Learn* 1987;2:139–172.
4. Hansen P, Jaumard B. Cluster analysis and mathematical programming. *Math Program* 1997;79:191–215.
5. Fayyad UM, Piatetsky-Shapiro G, Smyth P, *et al.* Advances in knowledge discovery and data mining. Cambridge (MA): MIT Press; 1996.
6. Bertsimas D, Mersearau AJ, Patel NR. Dynamic classification of online customers. In: Proceedings of SIAM International Conference on Data Mining; San Francisco (CA). 2003.
7. Gavrilov M, Anguelov D, Indyk P, *et al.* Mining the stock market: which measure is best? In: 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Boston (MA). 2000. pp. 487–496.
8. Wagstaff K, Cardie C, Rogers S, *et al.* Constrained k -means clustering with background knowledge. In: Proceedings of the 18th International Conference on Machine Learning; Williamstown (MA). 2001. pp. 577–584.
9. Iyigun C, Ben-Israel A. Semi-supervised probabilistic distance clustering and the uncertainty of classification. *Studies in classification, data analysis and knowledge organization*. Berlin, Heidelberg: Springer; 2010.
10. Everitt B, Landau S, Leese M. Cluster analysis. New York (NY): Oxford University Press Inc.; 2001.
11. Tan P, Steinbach M, Kumar V. Introduction to data mining. Boston (MA): Addison Wesley; 2006.
12. Jain AK, Murty MN, Flynn PJ. Data clustering: A review. *ACM Comput Surv* 1999;31(3):264–323.
13. Kotsiantis SB, Pintelas PE. Recent advances in clustering: a brief survey. Greece: Department of Mathematics, University of Patras; 2003.
14. Hartigan J. Clustering algorithms. New York (NY): John Wiley & Sons, Inc.; 1975.
15. Huang Z. Extensions to the k -means algorithm for clustering large data sets with categorical values. *Data Min Knowl Discov* 1998;2:283–304.
16. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Plenum; 1981.
17. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, New Jersey (NJ): Prentice Hall; 1988.
18. Zhang T, Ramakrishnan R, Linvy M. Birch. An efficient data clustering method for very large data sets. *Data Min Knowl Discov* 1997;1(2):141–182.
19. Ester M, Kriegel HP, Sander J, *et al.* A density-based algorithm for discovering clusters in large spatial data sets with noise. In: Proceedings 2nd International Conference on Knowledge Discovery and Data Mining. Portland (OR); 1996. pp. 226–231.
20. Xu X, Jager J, Kriegel HP. A fast parallel clustering algorithm for large spatial data sets. *Data Min Knowl Discov* 1999;3:263–290.
21. Kohonen T. Volume 30, Self-organizing maps, Springer Series in Information Sciences. 2nd extended ed. Berlin, Heidelberg, New York: Springer; 1997.
22. Jensen F. An introduction to Bayesian networks. New York (NY): Springer; 1996.
23. Kuhn HW. A note on Fermat's problem. *Math Program* 1973;4:98–107.
24. Höppner F, Klawonn F, Kruse R, *et al.* Fuzzy cluster analysis. Chichester, England: John Wiley & Sons, Inc.; 1999.
25. Teboulle M. A unified continuous optimization framework for center-based clustering methods. *J Mach Learn* 2007;8:65–102.
26. Heiser WJ. Geometric representation of association between categories. *Psychometrika* 2004;69:513–545.
27. Arav M. Contour approximation of data and the harmonic mean. *J Math Inequal* 2008;2(2):261–267.
28. Dixon KR, Chapman JA. Harmonic mean measure of animal activity areas. *Ecology* 1980;61:1040–1044.
29. Weiszfeld E. Sur le point par lequel la somme des distances de n points donnés est minimum. *Tohoku Math J* 1937;43:355–386.
30. Love R, Morris J, Wesolowsky G. Facilities location: models and methods. NY: North-Holland; 1988.
31. Ostresh LM Jr. On the convergence of a class of iterative methods for solving the Weber location problem. *Oper Res* 1978;26:597–609.
32. Bezdek JC. Fuzzy mathematics in pattern classification [PhD thesis]. (Applied Mathematics). Ithaca (NY): Cornell University; 1973.
33. Halkidi M, Batistakis Y, Vazirgiannis M. Cluster validity methods. *SIGMOD Rec* 2002;31:40–45.
34. Iyigun C. Probabilistic distance clustering.

- [PhD thesis] (Operations Research). Piscataway (NJ): Rutgers University; 2007.
- 35. Iyigun C, Ben-Israel A. Probabilistic distance clustering adjusted for cluster size. *Probab Eng Inf Sci* 2008;22:1–19.
- 36. Iyigun C, Ben-Israel A. Probabilistic distance clustering, theory and applications. In: Chaovalitwongse W, Pardalos PM, editors. *Clustering challenges in biological networks*. Singapore: World Scientific; 2008.
- 37. Iyigun C, Ben-Israel A. A generalized Weiszfeld method for the multifacility location problem. *Oper Res Lett*. 2010;38(3):207–214.
- 38. Han J, Kamber M. *Data mining: concepts and techniques*. San Francisco (CA): Morgan Kaufmann Publishers; 2006.
- 39. Iyigun C, Ben-Israel A. A new criterion for clustering validity via contour approximation of data. In press.

PROBABILITY WEIGHTING FUNCTIONS

ALI AL-NOWAIHI
SANJIT DHAMI
Department of Economics,
University of Leicester,
Leicester, UK

Under expected utility (EU) theory, decision makers weight probabilities linearly. But the evidence strongly suggests nonlinear weighting of probabilities. Consider the following example from Kahneman and Tversky [[1], p. 283]. Suppose that one is compelled to play Russian roulette. One would be willing to pay much more to reduce the number of bullets from one to zero than from four to three. However, in each case, the reduction in the probability of a bullet firing is 1/6, and so, under EU, the decision maker should be willing to pay the same amount. This suggests nonlinear weighting of probabilities, which is also supported by the emerging neuroeconomic evidence [2].

The main alternatives to EU—*rank-dependent utility* (RDU), *prospect theory* (PT), and *cumulative prospect* (CP) *theory*—incorporate nonlinear weighting of probabilities using the device of a *probability weighting function* (PWF). Sections titled “Foundations and Principles of Decision Analysis,” “Psychological Basis of Decision-making under Uncertainty and Risk,” “Normative Structuring of Decision Problems,” “Modeling a Decision Maker’s Preferences for Outcomes,” and “Modeling a Decision Maker’s Risk Attitude” of this encyclopedia and, in particular, the five articles in the section titled “Foundations and Principles of Decision Analysis” discuss the axiomatic foundations of EU, RDU, PT, and CP. They discuss the nonparametric restrictions that must be imposed on the PWF. The main restrictions are that the PWF is a continuous, increasing function. This, for instance, is the case in the axiomatizations of RDU by Abdellaoui [3] and Wakker [4].

By contrast, the aim of this article is to explore the parametric PWFs. Our approach, therefore, complements the material elsewhere in the Encyclopedia, gives concrete guidance to the use of PWFs in practical applications, and imposes parametric restrictions on the PWF, which allows for sharper tests of the predictions.

SOME BASICS

Let $X = \{x_1, x_2, \dots, x_n\}$ be a fixed finite set of real numbers, which represents the possible monetary *outcomes/wealth levels*. Assume that $x_1 < x_2 < \dots < x_n$. The decision maker has a set of feasible actions, A . Any action in A induces a probability distribution $(p_1, p_2, \dots, p_n)$, $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$, over the outcomes. We then define a *first-order lottery*, L , as

$$L = (x_1, p_1; x_2, p_2; \dots; x_n, p_n), \quad (1)$$

with the interpretation that outcome x_i occurs with probability p_i . Higher order lotteries (lotteries of lotteries) can then be defined by recursion. Let $\mathcal{L}$ be the set of all lotteries. How should the decision maker choose an action in A ? Expected utility (EU) theory postulates that a decision maker has a complete transitive preference ordering, $\preceq$, on $\mathcal{L}$. Under well-known assumptions [5] each lottery is equivalent (under $\preceq$) to a first-order lottery. Furthermore, $\preceq$ can be represented by an expected utility functional, $EU : \mathcal{L} \rightarrow \mathbb{R}$, which takes the form

$$EU(L) = \sum_{i=1}^n p_i u(x_i), \quad (2)$$

where $u(x_i)$ is the utility of the outcome x_i and $(x_1, p_1; x_2, p_2; \dots; x_n, p_n)$ is the first-order lottery equivalent to L . Thus, $L_1 \preceq L_2 \Leftrightarrow EU(L_1) \leq EU(L_2)$. A key axiom used in the derivation of Equation (2) is the *independence axiom*.

Definition 1. [Independence axiom].

Suppose that $\preceq$ is a preference relation defined over the set of lotteries. Then for all lotteries L_1, L_2, L , and all $p \in [0, 1]$, $L_1 \preceq L_2 \Leftrightarrow (L_1, p; L, 1-p) \preceq (L_2, p; L, 1-p)$.

The independence axiom is often violated by the evidence. Alternatives to EU mainly relax this axiom. Two main features of EU stand out. (i) There is additive separability across outcomes. (ii) The objective function is linear in probabilities. Most alternatives to EU relax the second feature, that is, linearity in probabilities.

Suppose that the decision maker weights outcome x_i with the weight π_i . Suppose that in place of Equation (2) we have

$$V(L) = \sum_{i=1}^n \pi_i u(x_i). \quad (3)$$

Clearly, Equation (2) is the special case of Equation (3) with $\pi_i = p_i$. Early generalizations of EU took π_i to be a function of p_i . However, it is well known that such a point transformation of objective probabilities can lead the decision maker to choose stochastically dominated options (violation of monotonicity) even when such dominance is obvious [6].

STYLIZED FACTS ON NONLINEAR WEIGHTING OF PROBABILITIES

On the basis of brief discussion so far, we summarize two stylized facts, S1 and S2.

- S1. Evidence suggests that decision makers weight probabilities in a nonlinear manner. In particular, the evidence suggests that decision makers overweight low probabilities and underweight high probabilities (inverse S-shaped probability weighting).
- S2. Nonlinear point transformation of probabilities, $\pi_i = \pi_i(p_i)$, can violate monotonicity even when such violation is obvious.

A third stylized fact of at least equal significance but that has received relatively less attention in theoretical work is the following.

- S3. For events close to the boundary of the probability interval, $[0, 1]$, extensive empirical evidence suggests that decision makers (i) ignore events of extremely low probability and (ii) treat extremely high probability events as certain. In contrast to S1, S3 applies only to very low and very high probabilities.

We emphasize that we are not making a case for ditching S1. Indeed S1 is a robust empirical finding that allows one to resolve many important problems and puzzles in the literature. However, S1, in the absence of S3, leads to severe problems for theory in explaining facts concerning events with probabilities close to the end points of the probability interval $[0, 1]$. Indeed al-Nowaihi and Dhami [7–9] show that the set of such problems is large and economically important. Hence, we make the case for a class of probability weighting functions that simultaneously incorporates both S1 and S3. Furthermore, our own view is that one cannot argue that either S1 or S3 is more important than the other.

S1 and S2 are well documented; see Starmer [6] and Kahneman and Tversky [10], and Appendix A in Wakker [11]. However, S3 is less well documented, and theoretical work that incorporates it explicitly and formally has only just begun to appear. We now briefly review the evidence for S3.

EVIDENCE FOR S3

In their Nobel prize winning work on *prospect theory* (PT), Kahneman and Tversky [1] were acutely aware of S3 and they explicitly chose to emphasize it's salience. In PT, there is a psychologically rich *editing* phase followed by an *evaluation/decision* phase; the latter uses a rule of the form (3). From our perspective, the most critical aspect of the editing phase arises when decision makers decide which extreme probability events to ignore. This is reflected in the manner that $\pi(p)$ is drawn by Kahneman and Tversky [1] (Fig. 1).

Kahneman and Tversky [1] wrote the following (pp. 282–283) to cogently summarize the evidence on the end points of the probability interval $[0, 1]$: *The sharp drops or*

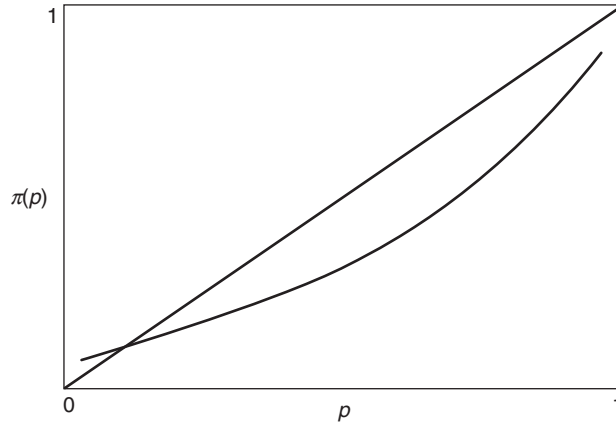


Figure 1. Ignorance at the end points.
[Source: Kahneman and Tversky [1, p. 283]].

apparent discontinuities of $\pi(p)$ at the end points are consistent with the notion that there is a limit to how small a decision weight can be attached to an event, if it is given any weight at all. A similar quantum of doubt could impose an upper limit on any decision weight that is less than unity... On the other hand, the simplification of prospects can lead the individual to discard events of extremely low probability and to treat events of extremely high probability as if they were certain. Because people are limited in their ability to comprehend and evaluate extreme probabilities, highly unlikely events are either ignored or overweighted, and the difference between high probability and certainty is either neglected or exaggerated. Consequently $\pi(p)$ is not well-behaved near the end points.

Dharm and al-Nowaihi [7–9]) review evidence strongly supportive of the claim of Kahneman and Tversky [1], from many disparate contexts. We now draw on these papers to briefly, and selectively, summarize the evidence.

Remark 1. Just so as to be absolutely clear we are not arguing that stylized fact S3 applies to all individuals. However, depending on the context, we argue that it could apply to a sizable number of people. Our model is capable of incorporating both sets of individuals: those to whom S3 applies and those to whom it does not. Indeed there is a well-developed notion of *bimodal perception of risks* in psychology that we outline in the

section titled “A Possible Formalization of the Bimodal Perception of Risk”. In this notion, some individuals are concerned mainly with the size of a loss while others, to whom stylized fact S3 applies, are mainly concerned about the probability. For the latter set of individuals, attention to losses declines dramatically as the probability falls below a minimum threshold. Theory needs to be able to accommodate both sets of individuals; we show that this is true of our theory in the section titled “A Possible Formalization of the Bimodal Perception of Risk”

Insurance for Low Probability Events

The insurance industry is of tremendous economic importance. Yet, despite impressive progress, existing theoretical models are unable to explain the stylized facts on the take-up of insurance for low probability events. The seminal study by Kunreuther *et al.* [12] provides striking evidence of individuals buying inadequate, nonmandatory, insurance against low probability events (e.g., earthquake, flood, and hurricane damage in areas prone to these hazards).

EU predicts that a risk-averse decision maker facing an actuarially fair premium will, in the absence of transaction costs, buy *full* insurance for all probabilities, however small. To test this prediction, Kunreuther *et al.* [[12], Chapter 7] presented subjects with varying potential losses with various probabilities, keeping the expected value of

the loss constant. Subjects faced actuarially fair, unfair, or subsidized premiums. In each case, they found that there is a point below which the take-up of insurance drops dramatically, as the probability of the loss decreases (and as the magnitude of the loss increases, keeping the expected loss constant). These results were shown to be robust to a very large number of perturbations and factors that include generous government incentives to reduce premiums to below market rates, controlling for moral hazard, and so on; see al-Nowaihi and Dhami [8] for details.

Arrow's own reading of the evidence in Kunreuther *et al.* [12], summarized in the foreword to Kunreuther *et al.* [12], is that: *Obviously in some sense it is right that he or she [the insuree] be less aware of low probability events, other things being equal; but it does appear from the data that the sensitivity goes down too rapidly as the probability decreases.*

Kunreuther *et al.* [12] write: *Based on these results, we hypothesize that most homeowners in hazard-prone areas have not even considered how they would recover should they suffer flood or earthquake damage. Rather they treat such events as having a probability of occurrence sufficiently low to permit them to ignore the consequences* These results are borne out from further testing by Camerer and Kunreuther [13] who conclude that: *Individuals seem to buy insurance only when the probability of risk is above a threshold . . .* This behavior is in close conformity to the observations of Kahneman and Tversky [1] outlined above.

The Becker Paradox

A celebrated result of Becker [14], which is known as the Becker proposition states that the most efficient way to deter a crime is to impose the *severest possible penalty with the lowest possible probability*. By reducing the probability of detection and conviction, society can economize on the costs of enforcement such as policing and trial costs. But by increasing the severity of the punishment (e.g., fines) the deterrence effect of the punishment is maintained. Indeed, under EU,

risk neutrality or risk aversion, and allowing for infinitely severe punishments, the Becker proposition implies that crime would be deterred completely, however small the probability of detection and conviction. Risk seeking behavior is not sufficient to overturn the Becker proposition either [9]. The Becker proposition is sometimes phrased as *it is efficient to hang offenders with probability zero*. Empirical evidence, summarized in Dhami and al-Nowaihi [9], however, is not supportive of the Becker proposition. This, we call as the Becker paradox.

Remark 2. It is shown in Dhami and al-Nowaihi [9], in detail, that nine explanations that have been seriously proposed in the literature cannot explain the Becker paradox. These explanations include the following: risk seeking behavior on the part of offenders, the ability to avoid severe fines by declaring bankruptcy, the need for differential punishments, type-I and type-II errors in conviction, rent seeking behavior in the presence of severe punishments, abhorrence of severe punishments, objectives other than deterrence, non-availability of severe punishments, and the psychological traits of offenders.

The behavior of decision makers for low probability events is critical to understanding the Becker paradox. Dhami and al-Nowaihi [9] show that if the probability weighting function respects stylized fact S1 but not S3, then the Becker paradox survives even under RDU and CP. However, once one allows for a probability weighting function that respects both stylized facts S1 and S3, then in conjunction with the reference point feature of CP, Dhami and al-Nowaihi [9] show that the Becker paradox is easily resolved. Indeed, S3 turns out to be a necessary condition for the resolution of the Becker paradox (and related paradoxes, see Remark 3).

Remark 3. The Becker paradox is a general paradox that applies to all situations where a decision maker is *not* deterred from some act, a , that would result in *almost infinite punishment with small probability*. In each

of these cases, and under plausible assumptions on risk attitudes, EU predicts that the decision maker will be *dissuaded* from the act a . Under RDU and CP the standard probability weighting functions respect S1 but not S3. This acts to dissuade the decision maker even more powerfully from the act a , as compared to EU. Hence, as al-Nowaihi and Dhami [7–9] show formally, standard theories with nonlinear weights make the Becker paradox even worse. Furthermore, they show that the simultaneous incorporation of S1 and S3 into the same probability weighting function is a necessary condition for solving the class of Becker paradoxes. In the remaining part of this section we consider more examples of the Becker paradox. These include evidence from jumping red traffic lights, driving and talking on mobile phones, seat belt usage, and the take-up of breast cancer examinations.

Evidence from Jumping Red Traffic Lights

In running red lights there is a *small probability* of an accident. However, the consequences are self-inflicted and potentially have *infinite costs*. This act is, therefore, like the act a in Remark 3. Rephrased, running red traffic lights is similar to *hanging oneself with a very small probability*, which is similar to the Becker proposition.

It is estimated in Bar Ilan and Sacerdote [15] that there are approximately 260,000 accidents per year in the United States caused by red-light running with implied costs of car repair alone of the order of \$520 million per year. Using Israeli data, the authors [16] calculated that the expected gain from jumping one red traffic is, at most, 1 min (the length of a typical light cycle). Given the known probabilities, they find that: *If a slight injury causes a loss greater or equal to 0.9 days, a risk neutral person will be deterred by that risk alone. However, the corresponding numbers for the additional risks of serious and fatal injuries are 13.9 days and 69.4 days respectively.* To this must be added other costs arising from an accident. It stretches plausibility to assume that these are simply mistakes.

Bar Ilan and Sacerdote [15–17] provide near-decisive evidence that the alternative

explanations listed in Remark 2 for the Becker paradox, cannot explain the decision to run red traffic lights. It is quite clear that several of those explanations do not work because the punishment is self-inflicted; see the report by The Royal Society for the Prevention of Accidents [9] for a detailed discussion. A far more natural explanation is that the probability of an accident is low enough for stylized fact S3 to apply for many people. Thus, red traffic light running is simply caused by some individuals ignoring (or seriously underweighting) the very low probability of an accident. The *bimodal perception of risk* (see the section titled “A Possible Formalization of the Bimodal Perception of Risk”) then ensures that a fraction of the drivers do not run traffic lights. If the PWF incorporates S1 but not S3, then we should not observe any red traffic light running at all (Remark 3), which is contrary to the evidence.

Driving and Talking on Car Mobile Phones

Consider the usage of mobile phones by drivers in moving vehicles. Such an individual faces *potentially infinite punishment* (e.g., loss of one’s and/or the family’s life) with *low probability*, in the event of an accident. The Becker proposition applied to this situation would suggest that drivers of vehicles will not use mobile phones while driving or perhaps use hands-free phones, and so, avoid the self-inflicted punishment. Yet, evidence suggests that drivers use mobile phones in moving vehicles [18,19]; hence, this problem belongs to the general class of problems in the Becker paradox (see Remark 3 above). Evidence also suggests that hands-free phone equipment does not offer significantly more safety. A natural explanation of this class of problems is that individuals simply ignore or substantially underweight the low probability of an accident caused by driving and talking on mobile phones, as in stylized fact S3.

Other Examples

People were reluctant to employ seat belts prior to their mandatory use despite publicly available evidence that seat belts can make

the difference between surviving and dying in a car crash, or simply that they provide better safety. Prior to 1985, in the United States, only 10–20% of motorists wore seat belts voluntarily, hence, denying themselves *self-insurance* [20]. *Potentially fatal* car accidents may be perceived by individuals as *low probability events*, particularly if they are overconfident of their driving abilities. Overconfidence is supported from a wide range of contexts [7]. Reluctance to wear seat belts can, thus, be seen to be a special case of the general Becker paradox outlined in Remark 3, which requires the use of stylized fact S3 as a necessary condition for an explanation.

Even as evidence accumulated about the dangers of breast cancer (which has a *low unconditional probability but fatal consequences*) women took up the offer of breast cancer examination, only sparingly. In the United States, this only changed after the greatly publicized events of the mastectomies of Betty Ford and Happy Rockefeller [12, p. xiii, pp. 13–14]. It is immediate that this too is a special case of the Becker paradox that requires an incorporation of S3 for its resolution.

Conclusion from These Disparate Contexts

al-Nowaihi and Dhami [7–9] draw two main conclusion from the evidence.

1. Human behavior for very low probability events cannot easily be explained by the existing mainstream theoretical models of risk. EU and the associated auxiliary assumptions are unable to explain the stylized facts. Furthermore, the leading non-expected utility alternatives such as (*RDU*) and (*CP*) make the problem even worse for those events where stylized fact S3 has bite. (*PT*) on the other hand is able to account for S1 and S3; however, its treatment of S3 is at an informal/ heuristic level only; see particularly the criticisms in Birnbaum [21].
2. A natural explanation for these phenomena seems to be that individuals simply ignore or seriously underweight very low probability events (stylized fact S3).

PROBABILITY WEIGHTING FUNCTIONS (PWFS)

Stylized facts S1, S2, and S3 would seem, at least to us, to be the minimum requirements that a theory of risk should address. Most alternatives to EU that use nonlinear weighting of probabilities, such as *RDU*, *PT*, *CP* invoke the concept of a *PWF* to incorporate stylized facts S1 and S2. None of these theories can incorporate all three stylized facts S1, S2, and S3. We examine below emerging work of al-Nowaihi and Dhami [7–9], who propose *composite cumulative prospect (CCP) theory* that successfully addresses all three stylized facts S1, S2, and S3.

Remark 4. A PWF, by itself, is not a theory of risk. It needs to be embedded within the framework of a decision theory, say, *RDU*, *PT*, and *CP*, for it to have significant predictive content in concrete economic situations.

Definition 2. [PWF]. By a PWF we mean a strictly increasing function $w(p) : [0, 1] \xrightarrow{\text{onto}} [0, 1]$.

A simple proof, that we omit, can be used to demonstrate the following properties of a PWF; see al-Nowaihi and Dhami [7] for the proofs.

Proposition 1. A PWF has the following properties:

(a) $w(0) = 0$, $w(1) = 1$. (b) w has a unique inverse, w^{-1} , and w^{-1} is also a strictly increasing function from $[0, 1]$ onto $[0, 1]$. (c) w and w^{-1} are continuous.

Definition 3. The function, $w(p)$, (a) *infinitely overweights infinitesimal probabilities*, if $\lim_{p \rightarrow 0} \frac{w(p)}{p} = \infty$, and (b) *infinitely underweights near-one probabilities*, if $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} = \infty$.

Definition 4. The function, $w(p)$, (a) *zero-underweights infinitesimal probabilities*, if

$\lim_{p \rightarrow 0} \frac{w(p)}{p} = 0$, and (b) *zero-overweights near-one probabilities*, if $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} = 0$.

Definition 5. (a) $w(p)$ *finitely overweights infinitesimal probabilities*, if $\lim_{p \rightarrow 0} \frac{w(p)}{p} \in (1, \infty)$, and (b) $w(p)$ *finitely underweights near-one probabilities*, if $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} \in (1, \infty)$.

Definition 6. (a) $w(p)$ *positively underweights infinitesimal probabilities*, if $\lim_{p \rightarrow 0} \frac{w(p)}{p} \in (0, 1)$, and (b) $w(p)$ *positively overweights near-one probabilities*, if $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} \in (0, 1)$.

The following are some of the references where single-parameter PWFs are mentioned [22–27]. The two-parameter PWFs are mentioned in the following papers: [28–33]. The Prelec (1998) function appears to be the one with the strongest empirical support. To quote from Stott [[34], p. 102]: “the most predictive version of [cumulative prospect theory] has a power value curve, a single-parameter risky weighting function due to Prelec [31] and a Logit stochastic process.” It was also the first axiomatically derived PWF.

Definition 7. [31]. By the *Prelec function* we mean the PWF $w(p) : [0, 1] \rightarrow [0, 1]$ given by

$$w(0) = 0, w(1) = 1, \quad (4)$$

$$w(p) = e^{-\beta(-\ln p)^\alpha}, \quad 0 < p \leq 1, \alpha > 0, \beta > 0. \quad (5)$$

Definition 8. By the *standard Prelec function* we mean the Prelec function (Definition 7) with $\alpha < 1$.

The following Proposition can be easily checked.

Proposition 2. *The Prelec function (Definition 7) is a probability weighting function in the sense of Definition 2.*

We plot the Prelec PWF in Fig. 2; this is a plot of $w(p) = e^{-(\ln p)^{0.5}}$, that is, $\beta = 1$ and $\alpha = 0.5$. Since $\alpha < 1$, the Prelec function plotted in Fig. 2 is, in fact, a standard Prelec function (Definition 8).

The parameter α controls the convexity/concavity of the Prelec function. Between the region of strict convexity ($w'' > 0$) and the region of strict concavity ($w'' < 0$), there is a point of inflexion ($w'' = 0$). The parameter β in the Prelec function controls the location of the inflexion point relative to the 45° line. Sometimes, the respective roles of α and β are also referred to as the *curvature* and *elevation* properties of a PWF [29,35]. Not all PWFs allow for a clear separation between curvature and elevation. This is particularly the case for PWFs that involve one parameter rather than two parameters.

The full set of possibilities for the Prelec PWF, for different combinations of α, β is established in al-Nowaihi and Dhami [7]. A simple proof leads to the following result.

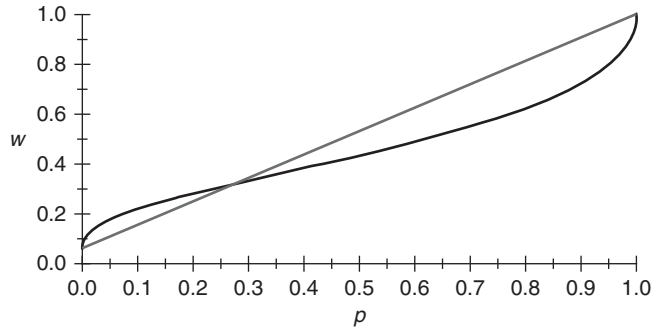


Figure 2. A plot of the Prelec (1988) function, $w(p) = e^{-(\ln p)^{1/2}}$.

Proposition 3. [7]

- (a) For $\alpha < 1$ the Prelec function, that is, the standard Prelec function (Definitions 7 and 8) (i) is inverse-S-shaped, (ii) infinitely overweights infinitesimal probabilities, that is, $\lim_{p \rightarrow 0} \frac{w(p)}{p} = \infty$, and (iii) infinitely underweights near-one probabilities, that is, $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} = \infty$.
- (b) For $\alpha > 1$ the Prelec function (i) is S-shaped, (ii) zero-underweights infinitesimal probabilities, that is, $\lim_{p \rightarrow 0} \frac{w(p)}{p} = 0$, and (iii) zero-overweights near-one probabilities, that is, $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} = 0$.

According to Prelec [[31], p. 505], the infinite limits in Proposition 3 (a) capture the qualitative change as we move from certainty to probability and from impossibility to improbability. On the other hand, they contradict stylized fact S3, that is, the observed behavior that people ignore events of very low probability and treat very high probability events as certain. These specific problems are avoided for $\alpha > 1$. However, for $\alpha > 1$, the Prelec function is S-shaped. This, however, is in conflict with stylized fact S1.

Remark 5. [Standard PWFs]. A large number of PWFs have been proposed, for example, those by Gonzalez and Wu [29], Lattimore Baker [30], and Prelec [31] (with $\alpha < 1$, that is, the standard Prelec function) [32,33]. They all overweight infinitesimal probabilities. We shall call these the standard PWFs. All these functions violate stylized fact S3.

Axiomatic Derivations of Prelec's PWF

Here, we overview three derivations of Prelec's function: Prelec [31] based on *compound invariance*, Luce [36] based on *reduction invariance*, and al-Nowaihi and Dhimi [37] based on *power invariance*.

We assume here that $0 \in X$, the set of outcomes, and we shall restrict ourselves to the special class of lotteries defined as follows.

Definition 9. Let $\mathbf{S} \subset \mathcal{L}$ be the subset of all lotteries of the forms (x) , (x, p_1) , $((x, p_1), p_2)$ and $((x, p_1), p_2), p_3)$, where $x \in X$ and $p_1, p_2, p_3 \in [0, 1]$.

To simplify notation, we shall refer to $((x, p_1), p_2)$ and $((x, p_1), p_2), p_3)$ by $(x; p_1, p_2)$ and $(x; p_1, p_2, p_3)$, respectively. Thus, (x) is the lottery whose outcome is x for sure, $(x; p_1)$ is the lottery whose outcomes are x with probability p_1 and 0 with probability $1 - p_1$, $(x; p_1, p_2)$ is the lottery whose outcomes are $(x; p_1)$ with probability p_2 and 0 with probability $1 - p_2$, and $(x; p_1, p_2, p_3)$ is the lottery whose outcomes are $(x; p_1, p_2)$ with probability p_3 and 0 with probability $1 - p_3$.

Given a strictly increasing function, $u : X \rightarrow \mathbb{R}$, and a PWF, w , we can extend u to a function, $U : \mathbf{S} \rightarrow \mathbb{R}$, by the following definition.

Definition 10. $U(x) = u(x)$, $U(x; p_1) = w(p_1)U(x)$, $U(x; p_1, p_2) = w(p_2)U(x; p_1)$, and $U(x; p_1, p_2, p_3) = w(p_3)U(x; p_1, p_2)$.

Definition 11. Let $\leq$ be the order on $\mathbf{S}$ induced by U , that is, for all $L_1, L_2 \in \mathbf{S}$, $L_1 \leq L_2 \Leftrightarrow U(L_1) \leq U(L_2)$.

Definition 12. [31]. The preference relation, $\leq$, satisfies *compound invariance* if, for all outcomes $x, y, x', y' \in X$, probabilities $p, q, r, s \in [0, 1]$, and integers $n \geq 1$, the following holds. If $(x, p) \sim (y, q)$ and $(x, r) \sim (y, s)$, then $(x', p^n) \sim (y', q^n)$ implies $(x', r^n) \sim (y', s^n)$.

Definition 13. [36]. The preference relation, $\leq$, satisfies *reduction invariance* if, for all outcomes $x \in X$, probabilities $p_1, p_2, q \in [0, 1]$ and $\lambda \in \{2, 3\}$: $(x; p_1, p_2) \sim (x; q) \Rightarrow (x; p_1^\lambda, p_2^\lambda) \sim (x; q^\lambda)$.

Definition 14. [37]. The preference relation, $\leq$, satisfies *power invariance* if, for all outcomes $x \in X$, probabilities $p, q \in [0, 1]$ and $\lambda \in \{2, 3\}$: $(x; p, p) \sim (x; q) \Rightarrow (x; p^\lambda, p^\lambda) \sim (x; q^\lambda)$ and $(x; p, p, p) \sim (x; q) \Rightarrow (x; p^\lambda, p^\lambda, p^\lambda) \sim (x; q^\lambda)$.

It is easy to check that for $w(p) = p$ (i.e., the EU case) *compound invariance*, *reduction invariance*, and *power invariance* are all satisfied. Hence, each of these constitutes a weakening (or generalization) of EU.

Proposition 4. [37]. *A PWF, w , is the Prelec PWF if, and only if, the induced preference relation, $\preceq$, satisfies either compound invariance, reduction invariance, or power invariance.*

From Proposition 4, all three axioms, *compound invariance*, *reduction invariance*, and *power invariance* are equivalent.

ADDRESSING STYLIZED FACT S1

Addressing stylized fact S1 requires a PWF that is inverse-S-shaped, that is, one that overweights low probabilities and underweights high probabilities. All the standard PWFs have this feature. In particular, this is true for the Prelec PWF if $\alpha < 1$, that is, what we called the *standard Prelec function* (Proposition 3(a)).

ADDRESSING STYLIZED FACT S2

There are two main ways of addressing S2. Either one uses *RDU* or *CP theory*. Quiggin's main insight in Quiggin [38,39] is that it is not individual probabilities that should be transformed (which gave rise to the violation of monotonicity reported in S2) but, rather, *cumulative probabilities*. When EU is applied to the transformed cumulative probabilities, we get what is now known as *RDU*.

In *RDU*, the decision maker uses Equation (3) with the decision weights generated as follows.

Definition 15. Consider the lottery $(x_1, p_1; x_2, p_2; \dots; x_n, p_n)$, where $x_1 < x_2 < \dots < x_n$. Let w be the PWF. For *RDU*, the *decision*

weights, π_i , are defined as follows.

$$\begin{aligned} \pi_n &= w(p_n), \\ \pi_{n-1} &= w(p_{n-1} + p_n) - w(p_n), \\ &\vdots \\ \pi_i &= w\left(\sum_{j=i}^n p_j\right) - w\left(\sum_{j=i+1}^n p_j\right), \\ &\vdots \\ \pi_1 &= w\left(\sum_{j=1}^n p_j\right) - w\left(\sum_{j=2}^n p_j\right) = w(1) \\ &\quad - w\left(\sum_{j=2}^n p_j\right) = 1 - w\left(\sum_{j=2}^n p_j\right). \end{aligned}$$

From Definition 15, we get that

$$\pi_j \geq 0 \text{ and } \sum_{j=1}^n \pi_j = 1. \quad (6)$$

Proposition 5. [3,38]. *A decision maker who uses RDU, never chooses stochastically dominated options (i.e., does not violate monotonicity). In other words, for an RDU decision maker there is no problem with explaining S2.*

A result analogous to that in Proposition 5 also holds for the case of *CP theory* [33,40,41]. In *CP*, the domain of outcomes is split into the domain of *gains* and the domain of *losses* by expressing each outcome relative to some *reference point*. We note here that the decision weights across the domain of gains and losses under *CP* do not necessarily add up to 1. This contrasts with the case of *RDU*, in which there is no conception of different domains of gains and losses; so the decision weights add up to 1. Since *CP* uses the cumulative weighting machinery from *RDU*, the following proposition holds.

Proposition 6. *A decision maker who uses CP does not choose stochastically dominated options. Hence, CP can address stylized fact S2.*

The PWF plays an important role in determining a rich set of attitudes toward risk

under CP. Under EU, attitudes to risk are determined purely by the curvature of the utility function. Under CP, however, the utility function is concave in the domain of gains and convex in the domain of losses [1]. Hence, it might be tempting to conclude that under CP the decision maker is risk averse in the domain of gains and risk loving in the domain of losses. This is not true because of the role played by the interaction of the PWF and the curvature of the utility function under CP, in determining attitudes to risk. The following fourfold pattern of risk preferences can be shown under CP [10]. The decision maker is risk loving for small probabilities in the domain of gains and non-small probabilities in the domain of losses. He/she is also risk averse for non-small probabilities in the domain of gains and small probabilities in the domain of losses.

ADDRESSING STYLIZED FACT S3

While RDU and CP in conjunction with, say, the standard Prelec function (Definitions 7 and 8), are able to explain S1 and S2, they are unable to address stylized fact S3.

al-Nowaihi and Dhami [7] make the ambitious proposal of combining the psychological richness of PT (in accounting for S3 in the editing phase) with the more satisfactory cumulative transformation of probabilities in CP. They combine PT and CP into a single theory, which they call *CCP theory*, which essentially combines the editing and decision phases of PT into a single phase, while retaining cumulative transformations of probability, as in CP. CCP successfully accounts for all three stylized facts S1, S2, and S3. It can explain everything that RDU and CP can, but the reverse is false.

An immediate implication of Proposition 3(a) is that the standard Prelec function ($\alpha < 1$) can explain S1 but not S3. From Proposition 3(b), we know that for $\alpha > 1$, the Prelec function is S-shaped, which explains S3 but contradicts S1. In order to explain S1, S2, and S3 using CCP, al-Nowaihi and Dhami [7] introduce a modification to the Prelec PWF in a manner that is consistent with the empirical evidence. They eliminate the discontinuities at the end points in Fig. 1 with

empirically founded as well as axiomatically founded behavior. They call their suggested modification as *composite Prelec weighting function* (CPF). Figure 3 depicts the general shape of CPF.

In Fig. 3, decision makers heavily underweight very low probabilities in the range $[0, p_1]$. Decision makers who use the weighting function in Fig. 3 would ignore very low probability events by assigning low subjective weights to them. Hence, in conformity with the evidence [7–9], they are unlikely to be dissuaded from “low-probability high-punishment” crimes, reluctant to buy insurance for very low probability events, reluctant to wear seat belts, reluctant to participate in voluntary breast screening programs, willing to run red traffic lights, and so on. For an even fuller description of behavior that accounts for the observed pattern of heterogeneity, one would need to incorporate further considerations such as those discussed in [7].

In the probability range $[p_3, 1]$, events are overweighted as suggested by Kahneman and Tversky [1, pp. 282–283]. In the middle segment, $p \in [p_1, p_3]$, the PWF is similar to the Prelec function with $\alpha < 1$, that is, a standard Prelec function successfully addresses S1.

For examples of the CPF, fitted to actual data, see al-Nowaihi and Dhami [7,8]. Owing to space limitations, we restrict ourselves to a brief, formal, description of the CPF. This implements the general shape of the CPF in Fig. 2. Define

$$\underline{p} = e^{-\left(\frac{\beta}{\beta_0}\right)^{\frac{1}{\alpha_0 - \alpha}}}, \quad \bar{p} = e^{-\left(\frac{\beta}{\beta_1}\right)^{\frac{1}{\alpha_1 - \alpha}}}. \quad (7)$$

Essentially, the CPF in Fig. 3 is described by segments from three different Prelec PWFs. The first is defined over the range $0 \leq p \leq \underline{p}$, the second is defined over the range $\underline{p} < p \leq \bar{p}$, and the third is defined over the range $\bar{p} < p \leq 1$. The choice of $\underline{p}$ and $\bar{p}$ will guarantee that the three segments are joined continuously.

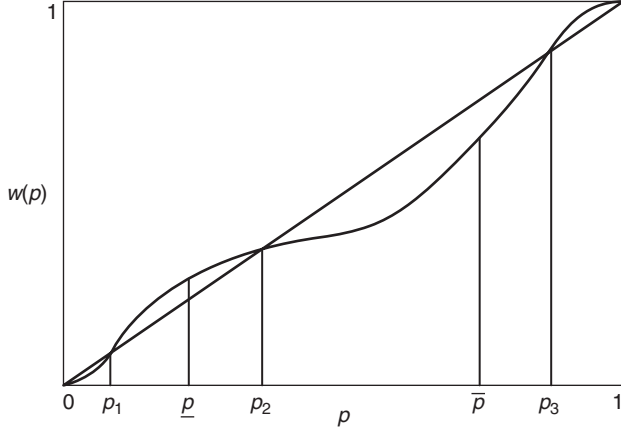


Figure 3. The composite Prelec weighting function (CPF).

Definition 16. [7]. By the *CPF* we mean the PWF $w : [0, 1] \rightarrow [0, 1]$ given by

$$w(p) = \begin{cases} 0 & \text{if } p = 0 \\ e^{-\beta_0(-\ln p)^{\alpha_0}} & \text{if } 0 < p \leq \underline{p} \\ e^{-\beta(-\ln p)^\alpha} & \text{if } \underline{p} < p \leq \bar{p} \\ e^{-\beta_1(-\ln p)^{\alpha_1}} & \text{if } \bar{p} < p \leq 1, \end{cases} \quad (8)$$

where $\underline{p}$ and $\bar{p}$ are given by Equation (8) and

$$0 < \alpha < 1, \beta > 0; \alpha_0 > 1, \beta_0 > 0; \alpha_1 > 1, \beta_1 > 0, \\ \beta_0 < 1/\beta^{\frac{\alpha_0-1}{1-\alpha}}, \beta_1 > 1/\beta^{\frac{\alpha_1-1}{1-\alpha}}. \quad (9)$$

Note the following. The first segment ($0 \leq p \leq \underline{p}$) of the CPF is identical to the Prelec function $e^{-\beta_0(-\ln p)^{\alpha_0}}$, which is S-shaped because $\alpha_0 > 1$. The second segment ($\underline{p} < p \leq \bar{p}$) is identical to the standard Prelec function $e^{-\beta(-\ln p)^\alpha}$, which is inverse-S shaped because $\alpha < 1$. The third, and final, segment of the CPF ($\bar{p} < p \leq 1$) is identical to the Prelec function $e^{-\beta_1(-\ln p)^{\alpha_1}}$, which is S-shaped because $\alpha_1 > 1$.

Proposition 7. [7]. *The CPF is a PWF.*

Define p_1, p_2, p_3 that correspond to the notation used for the general shape of a CPF in Fig. 3.

$$p_1 = e^{-\left(\frac{1}{\beta_0}\right)^{\frac{1}{\alpha_0-1}}}, p_2 = e^{-\left(\frac{1}{\beta}\right)^{\frac{1}{\alpha-1}}}, \\ p_3 = e^{-\left(\frac{1}{\beta_1}\right)^{\frac{1}{\alpha_1-1}}}. \quad (10)$$

Proposition 8. [7]. (a) $p_1 < \underline{p} < p_2 < \bar{p} < p_3$. (b) $p \in (0, p_1) \Rightarrow w(p) < p$. (c) $p \in (p_1, p_2) \Rightarrow w(p) > p$. (d) $p \in (p_2, p_3) \Rightarrow w(p) < p$. (e) $p \in (p_3, 1) \Rightarrow w(p) > p$.

By Proposition 7, the CPF in Equations (9) and (10) is a PWF in the sense of Definition 2. By Proposition 8, a CPF overweights low probabilities, that is, those in the range (p_1, p_2) , and underweights high probabilities, that is, those in the range (p_2, p_3) . Thus, it accounts for stylized fact S1. But, in addition, and unlike all the standard PWFs, it underweights probabilities near 0, that is, those in the range $(0, p_1)$, and overweights probabilities close to 1, that, those in the range $(p_3, 1)$ as required in S3.

Proposition 9. [7]. *The CPF (9):*

- (a) *zero-underweights infinitesimal probabilities, that is, $\lim_{p \rightarrow 0} \frac{w(p)}{p} = 0$ (Definition 4a),*
- (b) *zero-overweights near-one probabilities, that is, $\lim_{p \rightarrow 1} \frac{1-w(p)}{1-p} = 0$ (Definition 4b).*

Remark 6. [Axiomatic foundations of CPF-I]. The axiomatic foundations for the CPF in al-Nowaihi and Dhimi [7] rely on the axiom of *local power invariance*. This axiom, in turn, is a modification of the axiom of *power*

invariance used in al-Nowaihi and Dhami [37] to axiomatically derive the Prelec PWF.

Remark 7. [Axiomatic foundations of CPF-II]. Several axiomatizations of RDU allow for additive separability between terms that multiply utility with probability weights [3,4,11]. These axiomatizations are discussed elsewhere in the encyclopedia (see the section titled “Foundations and Principles of Decision Analysis”) and lie outside the scope of this entry. However, Abdellaoui [3] and Wakker [4] assume increasing PWFs; thus, CPF is consistent with them. In order to ensure concavity or convexity of the weighting function, these contributions introduce additional global auxiliary conditions such as probabilistic risk aversion. See, in particular, Wakker [11], who allows for a preference axiomatization of convex capacities. This allows for a weighting function that is either concave or convex throughout; hence, these cannot, simultaneously address stylized facts S1 and S3. Diecidue *et al.* [42] introduce local conditions that allow for probabilistic risk aversion to vary over the probability interval $[0, 1]$. For one set of parameter values, they can address S1 (in which case $\lim_{p \rightarrow 0} \frac{w(p)}{p} = \infty$) while for the complementary set of parameter values, they can address S3 (in which case $\lim_{p \rightarrow 0} \frac{w(p)}{p} = 0$). However, their proposal does not simultaneously address S1 and S2.

Remark 8. [Axiomatic foundations of CPF-III]. An advantage of the axiomatization of CPF by al-Nowaihi and Dhami [7] is that from the class of all increasing PWFs, which is an assumption made in all axiomatizations, they choose a particular parametric member, the CPF. Thus, the CPF (and the associated axiomatization that relies on local power invariance) allows for a sharper prediction, which turns out to be consistent with the evidence. Therefore, relative to the other axiomatizations, the work by al-Nowaihi and Dhami [7] addresses the further question “what extra axioms are needed to get more specific (and, hence, more predictive) PWFs?”

Definition 17. [7]. Otherwise standard CP, when combined with a CPF, is called *CCP theory*. Analogously, otherwise standard RDU, when combined with a CPF, is referred to as *composite rank-dependent utility* (CRDU).

al-Nowaihi and Dhami [7] prove the following proposition, whose intuition would by now be largely clear to the reader from our discussion of the CPF. Since probabilities in the middle ranges are weighted as in Prelec [31], stylized fact S1 is explained. Since cumulative transformations of probability are undertaken in CCP and CRDU, S2 is explained. And because of the property of the CPF in Proposition 9, S3 is explained.

Proposition 10. [7]. *CCP and CRDU can explain S1, S2, and S3.*

Blavatskyy [43] shows that the St. Petersburg paradox re-emerges under CP. al-Nowaihi and Dhami [7] show that the St. Petersburg paradox can be resolved under CCP mainly through the role played by the CPF. Rieger and Wang [32] also propose a PWF that resolves the St. Petersburg paradox, but it cannot explain stylized fact S3.

In comparison to CRDU, CCP, in addition, incorporates reference dependence, loss aversion, and richer attitudes toward risk. Hence, it can explain everything that CRDU can, but the converse is false. Furthermore, since CCP explains S1, S2, and S3, while CP (and RDU) can only explain S1 and S2, CCP can explain everything that CP (and RDU) can, but the converse is false. In light of these observations, it is interesting to note the observation in Machina [44] that “RDU is currently the most popular decision theory under risk.”

THE ALLAIS PARADOX UNDER CP AND CCP

Consider the Allais paradox as a *common consequence effect*. We omit a discussion of the *common ratio form*, as similar comments apply in that case. Consider the following two pairs of lotteries where outcomes are

presented relative to the status quo (so the reference point is 0):

$$a_1 = (1, 1); a_2 = (0, 0.01; 1, 0.89; 5, 0.1), \quad (11)$$

$$a_3 = (0, 0.89; 1, 0.11); a_4 = (0, 0.9; 5, 0.10). \quad (12)$$

Allais (1953) conjectured that decision makers would prefer a_1 to a_2 , while at the same time preferring a_4 to a_3 (this conjecture has been confirmed by a large number of experiments). As is well known, this pattern of preferences is violated under EU (on account of a violation of the independence axiom). Notice that all outcomes in the lotteries in Equations (12) and (13) are in the domain of gains. Let the utility function be of the usual power form, $u(x) = x^\gamma$, with $\gamma = 0.88$ as suggested by the experimental data [33]. Then, under CP (and also CCP), the stated pattern of preferences implies the following:

$$a_1 \succ a_2 \Leftrightarrow 1^{0.88} > [w(0.89 + 0.1) - w(0.1)]1^{0.88} + w(0.01)5^{0.88}, \quad (13)$$

$$a_4 \succ a_3 \Leftrightarrow w(0.1)5^{0.88} > w(0.11)1^{0.88}. \quad (14)$$

Suppose that under CP, the weighting function is the Prelec function with parameters $\alpha = 0.5$ and $\beta = 1$; so, $w(p) = e^{-(\ln p)^{0.5}}$. Thus, we can rewrite Equations (14) and (15) as

$$a_1 \succ a_2 \Leftrightarrow 1^{0.88} > \left[e^{-(\ln 0.99)^{0.5}} - e^{-(\ln 0.1)^{0.5}} \right] 1^{0.88} + w(0.01)5^{0.88}. \quad (15)$$

which is true.

$$\begin{aligned} a_4 \succ a_3 &\Leftrightarrow e^{-(\ln 0.1)^{0.5}} 5^{0.88} \\ &> e^{-(\ln 0.11)^{0.5}} 1^{0.88} \end{aligned} \quad (16)$$

which is also true. Hence, under CP, nonlinear weighting of probabilities allows for an explanation of the Allais paradox (in the common consequence effect form) that could not be explained under EU.

Remark 9. Can the Allais paradox (in the common consequence and common ratio effects forms) also be explained under CCP? From Definition 16 and Equation (9), we know that over the probability range $[\underline{p}, \bar{p}]$, CPF and CCP are identical. Hence, for lotteries that only have probabilities in the interval $[\underline{p}, \bar{p}]$ both theories make identical predictions. The intuition is that these probabilities are sufficiently nonextreme that only stylized fact S1 has any bite. Stylized fact S3 has bite only if probabilities lie outside the range $[\underline{p}, \bar{p}]$, say, in the two extreme probability intervals $[0, p_1]$ and $[p_3, 1]$ (see Fig. 3, Definition 16, and Equation (9)). In these two intervals, different segments of the Prelec function apply (because the CPF is made up of three different segments of a Prelec function) and one simply needs to alter the relevant values of α, β in inequalities such as Equations (16) and (17) above and recheck them.

In the light of Remark 9, it becomes important to compute $\underline{p}, \bar{p}$. In an ideal world one would obtain data from the relevant context and problem in order to elicit the values $\underline{p}, \bar{p}$. Here, we rely on the computation made by al-Nowaihi and Dhami [7] of these values based on data in Kunreuther *et al.* [12]. From two separate sets of data, the farm and the urn experiments, respectively, al-Nowaihi and Dhami [7] derive the following two sets of values for the range $(\underline{p}, \bar{p})$: $(0.05, 0.95]$ and $(0.25, 0.75]$.

The probabilities involved in the two comparisons in Equations (16) and (17) are 0.33, 0.34, and 0.99. The last of these probabilities lies outside the upper limit of both ranges $(0.05, 0.95]$ and $(0.25, 0.75]$. This does not affect the comparison in Equation (17) but does affect the comparison in Equation (16). For the urn and the farm experiments, al-Nowaihi and Dhami [7] show that the respective values of (α, β) for the upper segment of the Prelec function in the CPF are $(2, 6.4808)$ and $(2, 86.081)$. We now apply these values to the comparison in Equation (16), sequentially, for the urn and farm

experiments.

$$\text{Urn: } a_1 \succ a_2 \Leftrightarrow 1^{0.88} > \left[e^{-6.4808(-\ln 0.99)^2} - e^{-(\ln 0.1)^{0.5}} \right] 1^{0.88} + w(0.01)5^{0.88},$$

which is true.

$$\text{Farm: } a_1 \succ a_2 \Leftrightarrow 1^{0.88} > \left[e^{-86.081(-\ln 0.99)^2} - e^{-(\ln 0.1)^{0.5}} \right] 1^{0.88} + w(0.01)5^{0.88},$$

which is also true. Hence, the Allais paradox in common consequence effect form also holds true for CCP when we use a CPF that is fitted to the data in Kunreuther *et al.* [12]. We emphasize that we have not provided a proof that the Allais paradox will necessarily hold under CCP for all parameter values. Ideally one should have experimental data for the particular situation and then proceed along the lines of Remark 9 to check the resolution of the paradox.

A COMPARISON OF CCP WITH CONFIGURAL WEIGHTS MODELS

We now compare CCP with the class of *configural weight models*, which are reviewed elsewhere in the Encyclopedia (see the section titled “Foundations and Principles of Decision Analysis”). So we shall be brief and keep references to a minimum. The interested reader can directly consult Birnbaum [21] for an exhaustive set of references and a recent survey of these models.

Configural weight models, like several other alternatives to EU, nonlinearly weight probabilities of outcomes. In these models, one first writes down the decision tree corresponding to the problem. Consider some lottery, L , that has n possible consequences that arise from the different *branches* of the tree, ordered as, $x_n \leq x_{n-1} \leq \dots \leq x_2 \leq x_1$ with associated probabilities $p_n, p_{n-1}, \dots, p_2, p_1$. Outcome x_j is then said to have *rank* j , $j = 1, 2, \dots, n$ (one could use the convention $x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq x_n$ but in this case the rank of the j th number in the list (counting from the left) will be $n - j$).

THE RANK-AFFECTED MULTIPLICATIVE WEIGHTS MODEL (RAM)

Using Equation (8) in Birnbaum [21], in the *rank-affected multiplicative weights model* (RAM), the utility of lottery, L , is given by

$$\text{RAM}(L) = \frac{\sum_{i=1}^n a(i, n, s) t(p_i) u(x_i)}{\sum_{j=1}^n a(j, n, s) t(p_j)}. \quad (17)$$

Indices i, j denote the *rank* of the outcome (defined above). s is the *augmented sign* of the branch's consequence (positive, zero, or negative). $t(p)$ is an increasing function of p , which typically takes the form

$$t(p) = p^\gamma; 0 < \gamma < 1. \quad (18)$$

The utility function $u(x)$ typically takes the power form

$$u(x) = x^\beta; 0 < \beta < 1. \quad (19)$$

$a(i, n, s)$ is the *rank and sign augmented branch weight* corresponding to the outcome that has rank i . For two-branch gambles, for example, the lottery $L = (0, 0.5; 100, 0.5)$, the branch corresponding to the lower outcome ($\$0$) is associated with $a = 2$ and the higher outcome is associated with $a = 1$. For m branch gambles, such that $m \geq 3$, the weights are simply the ranks of the outcomes for that branch. For example, for the set of outcomes, $x_m \leq x_{m-1} \leq \dots \leq x_1$, the weights (the a 's) are $m, m-1, m-2, \dots, 1$. Thus, lower outcomes are given higher weights, which is one of the main features of the model. From Equation (18) the weight associated with outcome x_i is

$$w(p_i) = \frac{a(i, n, s) t(p_i)}{\sum_{j=1}^n a(j, n, s) t(p_j)}.$$

Using Equation (19), this can be written as

$$w(p_i) = \frac{1}{1 + \sum_{j \neq i} \frac{a(j, n, s)}{a(i, n, s)} \left(\frac{p_j}{p_i} \right)^\beta}. \quad (20)$$

From Equation (21) $w(0) = 1$ and $w(1) < 1$. So this does not satisfy the assumptions of a PWF. In particular, from Equation (21)

$$\frac{w(p_i)}{p_i^\gamma} = \frac{1}{p_i^\gamma + \sum_{j \neq i} \frac{a(j, n, s)}{a(i, n, s)} p_j^\gamma}.$$

Since $a(i, n, s)$ is nonzero and $\frac{a(j, n, s)}{a(i, n, s)}$ is finite,

$$\lim_{p_i \rightarrow 0} \frac{w(p_i)}{p_i^\gamma} = \frac{1}{\sum_{j \neq i} \frac{a(j, n, s)}{a(i, n, s)} p_j^\gamma},$$

which is finite. Hence, extremely low probabilities are not ignored, which contradicts stylized fact S3.

TRANSFER OF ATTENTION IN EXCHANGE (TAX) MODELS

In the *transfer of attention in exchange (TAX) models*, the decision maker is assumed to possess a fixed amount of limited attention. Owing to this, the decision maker transfers weight from higher ranked to lower ranked outcomes. Consider a *three-branch* lottery, L , given by $L = (x_1, p_1; x_2, p_2; x_3, p_3)$. Then, from Equation (10) in Birnbaum [21] we get that the utility of this lottery to the decision maker under TAX is given by

$$\text{TAX}(L) = \frac{w_1 u(x_1) + w_2 u(x_2) + w_3 u(x_3)}{w_1 + w_2 + w_3}, \quad (21)$$

where

$$w_1 = t(p_1) - \frac{2}{4} \delta t(p_1); w_2 = t(p_2) - \frac{1}{4} \delta t(p_2) + \frac{1}{4} \delta t(p_1); w_3 = t(p_3) + \frac{1}{4} \delta t(p_1) + \frac{1}{4} \delta t(p_2). \quad (22)$$

$t(p)$ continues to be given in Equation (19) and $\delta > 0$ is some constant. In this version of the TAX model, also known as the *special TAX model*, one transfers a fixed amount of probability weight from a higher outcome, proportionately, to lower outcomes. Thus, $w_1 + w_2 + w_3 = t(p_1) + t(p_2) + t(p_3)$. From w_1 in Equation (23), it follows that the weight on the highest outcome, x_1 , is reduced by $\frac{2}{4} \delta t(p_1)$

and each of the lower two outcomes gains one-half of this weight. Similarly, the middle ranked outcome, x_2 , loses weight $\frac{1}{4} \delta t(p_2)$ which is gained by the worse outcome, x_3 .

From Equations (19) and (23), we can compute the ratio of weights to probabilities as

$$\begin{aligned} \frac{w_1}{p_1^\gamma} &= 1 - \frac{1}{2} \delta; \frac{w_2}{p_2^\gamma} = 1 - \frac{1}{4} \delta + \left(\frac{p_1}{p_2}\right)^\gamma; \\ \frac{w_3}{p_3^\gamma} &= 1 + \frac{1}{4} \delta \left(\frac{p_1}{p_3}\right)^\gamma + \frac{1}{4} \delta \left(\frac{p_2}{p_3}\right)^\gamma. \end{aligned}$$

We assume that $0 < \delta \leq 2$ so that all weights are positive. There are three possibilities: (i) p_1 is the lowest probability. (ii) p_2 is the lowest probability. (iii) p_3 is the lowest probability. Since $t(p)$ is given in Equation (19), it follows that in these three cases, respectively,

$$\begin{aligned} \lim_{p_1 \rightarrow 0} \frac{w_1}{p_1^\gamma} &= 1 - \frac{1}{2} \delta \in (0, 1]; \lim_{p_2 \rightarrow 0} \frac{w_2}{p_2^\gamma} = \infty; \\ \lim_{p_3 \rightarrow 0} \frac{w_3}{p_3^\gamma} &= \infty. \end{aligned} \quad (23)$$

From Equation (24), we see again that stylized fact S3 is violated for this class of theories.

THE GAINS DECOMPOSITION UTILITY (GDU) MODEL

This model is set out as Equations (11)–(13) in Birnbaum [21]. The idea is to reduce multi-branch lotteries to simple, two-branch, lotteries. For instance, consider the binary lottery $L_2 = (x_1, p; x_2, 1 - p)$ such that $x_1 \leq x_2$. Then, under the *gains decomposition utility (GDU) model*, the utility of this binary lottery is

$$\text{GDU}(x_1, p, x_2) = w(p)u(x_1) + (1 - w(p))u(x_2). \quad (24)$$

Notice that the decision weights are formed exactly as under RDU. Three-branch lotteries such as $L_3 = (x_1, p_1; x_2, p_2; x_3, p_3)$ are then evaluated as follows. First, a choice is made between the lowest outcome x_3 and the binary lottery that comprises the remaining two outcomes. Probabilities are renormalized

so that they add up to 1 in the binary lottery. In other words, L_3 is treated as

$$\begin{aligned} L_3 &= \left(\left(x_1, \frac{p_1}{p_1 + p_2}; x_2, 1 - \frac{p_1}{p_1 + p_2} \right); \right. \\ &\quad \left. x_3, 1 - p_1 - p_2 \right) \Leftrightarrow L_3 \\ &= (L_2, p_1 + p_2; x_3, 1 - p_1 - p_2). \end{aligned} \quad (25)$$

Thus, L_3 is transformed into a binary lottery, which can be evaluated using Equation (25) as follows:

$$\begin{aligned} \text{GDU}(L_3) &= w(p_1 + p_2) \text{GDU} \left(x_1, \frac{p_1}{p_1 + p_2}, x_2 \right) \\ &\quad + [1 - w(p_1 + p_2)] u(x_3). \end{aligned}$$

In terms of the weighting of outcomes, this is exactly like RDU. The weighting functions used in this theory are the standard weighting function (e.g., the standard Prelec function is used by Birnbaum [21]). Thus, $\lim_{p \rightarrow 0} \frac{w(p)}{p} = \infty$ and, so, stylized fact S3 cannot be addressed by this theory either.

Remark 10. A great deal of the differences in prediction between CP (and by implication, CCP) and the configural weight models arise in the case of *event splitting* [21]. This occurs when there are two identical outcomes in a lottery. However, for the examples listed in the section titled “Evidence for S3”, event splitting does not occur [7–9].

Remark 11. For lotteries of the form $(x, p; y, 1 - p)$, where $x < 0 < y$, Birnbaum [21] reports that CP and TAX give close results. So, for TAX to be different, we need lotteries with, at least, three outcomes. More generally, it is difficult to distinguish between CP and TAX within the probability triangle. These observations are important. First, they show that where CP is successful (the experiments reported by Kahneman and Tversky [1,33] and, more generally, events that can be represented in the probability triangle), TAX is also successful. But for the “new” paradoxes, Birnbaum [21] argues that TAX succeeds where CP fails. In many

applications that involve use of stylized fact S3, configural weights models are likely to fail as do EU, RDU, and CP, while CCP is likely to succeed in these cases. In these cases, the configural weight models can benefit by incorporating stylized fact S3 by using a composite Prelec PWF.

A POSSIBLE FORMALIZATION OF THE BIMODAL PERCEPTION OF RISK

There is evidence of a *bimodal perception of risks* [13]. For the evidence, see Schade *et al.* [45]. Type-I individuals focus more on the probability of a loss while Type-II individuals focus more on the size of the loss.

Type-I individuals focus on the probability and do not pay attention to losses whose probability falls below a certain threshold. Hence, S3 applies to these individuals. Their behavior is explained by CCP and the CPF probability weighting functions. These individuals are the ones that jump red traffic lights, do not take out adequate insurance against low probability hazards, do not wear seat belts unless mandatory, are not deterred from crime by low probability capital punishment, and so on.

For Type-II individuals, the size of the loss is relatively more salient. They appear to fear large losses, even for extremely small probabilities. This could be accounted for by the standard Prelec function, since for that probability weighting function, $\lim_{p \rightarrow 0} \frac{w(p)}{p} = \infty$.

Whatever the underlying cause of their behavior, it appears to be best described by CP and the standard Prelec function. Indeed, since the underlying causes can be complex, assigning such individuals a Prelec function is a parsimonious and useful representation of that complexity.

In line with the evidence on the bimodal perception of risk, one may conjecture two subpopulations of humans. One whose behavior under risk is described by the standard Prelec probability weighting function and standard CP (Type-II individuals)

and the other by the composite Prelec probability weighting function and CCP (Type-I individuals). Although the focus of this section has been on the latter, our work should be regarded as complementing the (mainstream) Becker approach, thus extending it into areas it hitherto has been unsuccessful.

Example 1. Recall the discussion in the section titled “Evidence from Jumping Red Traffic Lights.” The reader may, rightly, wonder that if individuals ignore very low probabilities, then why does not everyone engage in, say, red traffic light running? There are many potential explanations. For instance, some individuals are simply “law abiding” under any circumstances. Others might be daunted by the possibility of losing their lives. Individuals who do not run traffic lights (for whatever reasons) behave “as if” they use the Prelec weighting function who infinitely overweight infinitesimal probabilities so the potential loss even with a low probability is very salient. Individuals who, on the other hand, run red traffic lights behave “as if” they satisfy the *CPF*. Such individuals ignore the very low probability of an accident and, so, the potentially infinite loss is not salient. This comment applies in spirit to all sorts of Becker paradoxes mentioned in this article.

Acknowledgments

Dr. Sanjit Dhami would like to acknowledge a one semester study leave provided by the University of Leicester. He also wishes to thank the International School of Economics (ISET) at Tbilisi State University, Georgia, for their excellent hospitality and academic environment during the time that he was a visiting Professor in the spring/summer of 2010. We are grateful to the editor and the referee for useful comments and suggestions.

REFERENCES

1. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47:263–291.
2. Berns GS, Capra CM, Chappelow J, *et al.* Non-linear neurobiological probability weighting functions for aversive outcomes. *Neuroimage* 2008;39:2047–2057.
3. Abdellaoui M. A genuine rank-dependent generalization of the von neumann-morgenstern expected utility theorem. *Econometrica* 2002;70:717–736.
4. Wakker P. Separating marginal utility and probabilistic risk aversion. *Theory Decis* 1994;36:1–44.
5. Fishburn PC. The foundations of expected utility. Dordrecht, Holland/Boston (MA), London: England: D. Reidel Publishing Company; 1982.
6. Starmer C. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *J Econ Lit* 2000;38:332–382.
7. al-Nowaihi A, Dhami S. Composite prospect theory: a proposal to combine prospect theory and cumulative prospect theory. University of Leicester.; 2010. Discussion paper.
8. al-Nowaihi A, Dhami S. Insurance behavior for low probability events. University of Leicester.; 2010. Discussion paper.
9. Dhami S, al-Nowaihi A. The Becker paradox reconsidered through the lens of behavioral economics. University of Leicester.; 2010. Discussion paper.
10. Kahneman D, Tversky A. Choices, values and frames. New York: Cambridge University Press; 2000.
11. Wakker P. Testing and characterizing properties of nonadditive measures through violations of the sure-thing principle. *Econometrica* 2001;69:1039–1059.
12. Kunreuther H, Ginsberg R, Miller L, *et al.* Disaster insurance protection: public policy lessons. New York: Wiley; 1978.
13. Camerer C, Kunreuther H. Experimental markets for insurance. *J Risk Uncertain* 1989;2:265–300.
14. Becker G. Crime and punishment: an economic approach. *J Polit Econ* 1968;76:169–217.
15. Bar Ilan A, Sacerdote B. The response of criminals and noncriminals to fines. *J Law Econ* 2004;47:1–17.
16. Bar Ilan A. The response to large and small penalties in a natural experiment. Haifa,

- Israel: Department of Economics. University of Haifa No. 31905; 2000.
17. Bar Ilan A, Sacerdote B. The response to fines and probability of detection in a series of experiments. National Bureau of Economic Research Working Paper. No. 8638; 2001.
 18. Pöystia L, Rajalina S, Summala H. Factors influencing the use of cellular (mobile) phone during driving and hazards while using it. *Accid Anal Prev* 2005;37:47–51.
 19. The Royal Society for the Prevention of Accidents. The risk of using a mobile phone while driving. 2005. Full text of the report can be found at www.rosipa.com.
 20. Williams A, Lund A. Seat belt use laws and occupant crash protection in the United States. *Am J Public Health* 1986;76:1438–1442.
 21. Birnbaum MH. New paradoxes of risky decision making. *Psychol Rev* 2008;115:463–501.
 22. Currim IS, Rakesh KS. Prospect versus utility. *Manage Sci* 1989;35:22–41.
 23. Hey JD, Orme C. Investigating generalizations of expected utility theory using experimental data. *Econometrica* 1994;62:1291–1326.
 24. Karmarkar US. Subjectively weighted utility and the allais paradox. *Organ Behav Hum Perform* 1979;24:67–72.
 25. Luce RD, Mellers BA, Chang S-J. Is choice the correct primitive? On using certainty equivalents and reference levels to predict choices among gambles. *J Risk Uncertain* 1993;6:115–143.
 26. Röell A. Risk aversion in Quiggin and Yaari's rank-order model of choice under uncertainty. *Econ J* 1987;97:143–160.
 27. Safra Z, Segal U. Constant risk aversion. *J Econ Theory* 1998;83:19–42.
 28. Goldstein WM, Einhorn HJ. Expression theory and the preference reversal phenomena. *Psychol Rev* 1987;94:236–254.
 29. Gonzalez R, Wu G. On the shape of the probability weighting function. *Cogn Psychol* 1999;38:129–166.
 30. Lattimore JR, Baker JK, Witte AD. The influence of probability on risky choice: a parametric investigation. *J Econ Behav Organ* 1992;17:377–400.
 31. Prelec D. The probability weighting function. *Econometrica* 1998;60:497–528.
 32. Rieger MO, Wang M. Cumulative prospect theory and the St. Petersburg paradox. *Econ Theory* 2006;28:665–679.
 33. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5:297–323.
 34. Stott HP. Choosing from cumulative prospect theory's functional menagerie. *J Risk Uncertain* 2006;32:101–130.
 35. Kilka M, Weber M. What determines the shape of the probability weighting function under uncertainty? *Manage Sci* 2001;47:1712–1726.
 36. Luce RD. Reduction invariance and Prelec's weighting functions. *J Math Psychol* 2001;45:167–179.
 37. al-Nowaihi A, Dhami S. A simple derivation of Prelec's probability weighting function. *J Math Psychol* 2006;50:521–524.
 38. Quiggin J. A theory of anticipated utility. *J Econ Behav Organ* 1982;3:323–343.
 39. Quiggin J. Generalized expected utility theory. Boston (MA): Kluwer Academic Publishers; 1993.
 40. Luce RD, Fishburn PC. Rank- and sign-dependent linear utility models for finite first-order gambles. *J Risk Uncertain* 1991;4:29–59.
 41. Starmer C, Sugden R. Probability and juxtaposition effects: an experimental investigation of the common ratio effect. *J Risk Uncertain* 1989;2:159–178.
 42. Diecidue E, Ulrich S, Horst Z. Parametric weighting functions. *J Econ Theory* 2009;144:1102–1118.
 43. Blavatsky PR. Back to the St. Petersburg paradox? *Manage Sci* 2005;51:677–678.
 44. Machina MJ. Non-expected utility theory. In: Durlaf SN, Blume LE, editors. *The new palgrave dictionary of economics*. 2nd ed. Basingstoke, New York: Macmillan; 2008.
 45. Schade C, Kunreuther H, Kaas KP. Low-Probability Insurance: Are Decisions Consistent with Normative Predictions? mimeo. University of Humboldt; 2001.

PROBLEM STRUCTURING FOR MULTICRITERIA DECISION ANALYSIS INTERVENTIONS

LUIS A. FRANCO
Warwick Business School,
University of Warwick,
Coventry, UK

GILBERTO MONTIBELLER
Department of Management,
London School of Economics,
London, UK

INTRODUCTION

Multicriteria decision analysis (MCDA), a methodology for supporting decision making when multiple objectives have to be pursued [1–3], has been extensively used to support a wide variety of complex decision problems [4,5]. While the literature on axiomatic aspects of multicriteria decision analysis models is extensive, much less attention has been devoted to the process of structuring these models, with few exceptions [6–8].

The task of structuring MCDA models in real-world interventions is far from trivial. This is mainly due to the intrinsic complexity of the models, where several objectives have to be articulated, defined, and measured by attributes. Furthermore, the definition of a set of alternatives to be evaluated is not always straightforward, as decision makers may struggle to think creatively about the problem and consider innovative alternatives.

At a broader level, much of the MCDA literature neglects the role of problem structuring as a prelude to the structuring of an MCDA model, a phase of the intervention whose proper management is absolutely crucial if both the decision analysts and the decision analysis are to have some effect on the organization.

In this article, we discuss problem structuring for MCDA interventions. There is a

limited literature on how to structure MCDA models, and this will be reviewed here. Furthermore, while there is a large body of literature on problem structuring and on problem-structuring methods, most of it is disconnected from the mainstream MCDA literature. We will build on this body of work to give a coherent perspective on problem structuring for MCDA, which we hope will be useful for both MCDA researchers and practitioners.

The chapter is structured as follows: we start by discussing two key problem-structuring tasks concerning the earlier stages of an MCDA intervention. The subsequent section reviews general guidelines for structuring MCDA evaluation models. In the final section, we build on the preceding sections to propose a general framework for conducting MCDA interventions, in which problem structuring plays a significant role. The chapter ends with concluding remarks and some directions for further research in the field.

STRUCTURING THE PROBLEM SITUATION

There are two main problem-structuring tasks faced by decision analysts when conducting MCDA interventions: *defining the problem*, and *scoping participation*. Below we discuss each of these tasks and comment on the challenges the decision analyst may encounter when carrying them out, together with a set of tools/techniques that could be used to facilitate their achievement. Although, in the discussion that follows, we present each problem-structuring task separately, it is worth noting that in practice these tasks are not necessarily undertaken in a linear fashion. Rather, there are two “modes” of problem structuring between which the decision analyst is continually “cycling” during the early stages of an MCDA intervention.

Defining the Problem

Given the significance of problem formulation

in organizational decision making [9–12], it is surprising that the literature on MCDA has devoted relatively minor attention to the processes of articulating and defining a multicriteria problem. It seems that the underlying assumption is that arriving at a well-structured multicriteria decision problem is somehow a relatively trivial task. There is also a widespread belief among many practitioners that structuring a decision problem is more “art than science” and that it can best be learned through experience. This view suggests that experienced analysts are able to recognize familiar patterns or structures of problems, and use them as templates to build their decision models [13]. Our experience as researchers and consultants, however, suggests that the use of decision analytic structures are well suited to problem situations that are clearly defined, but less so when they are ill-structured or ‘messy’ [14]. In such situations, attempts to impose a structure too early in the intervention can lead to focusing on and solving the ‘wrong problem’ and thus incurring in what is known as the *Type III* error [15].

Indeed empirical research has shown that the definition of problems, particularly those of the ill-structured type, is not given but continually negotiated among members of the organization before and during an intervention [16]. This process of negotiation can be conceptualized as follows. First, managers are constantly striving to make sense of their internal and external environments in order to manage and control their organizations [17]. This sense-making process is aided with the help of a unique mental framework that is developed through experience, and which includes systems of beliefs and values. A ‘problem’ emerges when the use of such a mental framework, to make sense of a particular situation, leaves the manager uneasy or dissatisfied because she/he does not know how to deal with that situation. Because different managers will experience different problems by applying their own unique mental frameworks to what might be thought of as the same situation, the decision analyst will not be able to think and talk about the ‘problem’ without ascribing an owner or owners to it.

Second, the problem which will eventually be presented to the analyst is the result of a process of ‘problem framing’ within the organization, most typically within a team of managers. As Eden and Sims aptly illustrate, a manager who wishes to get others in the team to take on a problem she/he has identified as being the team’s, “...will present the problem in such a way as to make it apparent that there are gains to be had or losses to be averted for other members of the team by solving this problem. He (sic) may seek to show some member of the team that a solution to his (the initial problem’s definer) problem would also solve some different problem, which he believes this member to be experiencing. ...he may define his problem to be in line with other problems which seem to be being experienced at that time. ... (or) express concern and commitment about some problem being stated by another member in the hope of getting some concern and commitment about his problem in return” (16, p. 121).

Thus, we might expect that when the analyst starts an MCDA intervention with a given problem situation presented by the client, the reality is that other versions of the same situation are likely to exist. These other versions will become apparent as the analyst listens to others in the organization. One challenge for the analyst at this stage is then not so much to model what will become the actual multicriteria decision problem to be solved, but to identify and model the different perceptions of the *problem situation* held by different managers. Several problem-structuring tools are available to support this task. These include, for example, cognitive mapping [18,19]; soft systems methodology [20,21]; dialog mapping [22]; strategic choice approach [23], and group model building [24,25]. (For an overview of these tools see Ref. 10.)

Most of these tools have been developed to capture multiple aspects of a problem situation, including objective and subjective ones. This is important because when managers define a problem situation, it will be defined in their own language and based on their own interpretations of the situation, their own experience or expertise and their own value

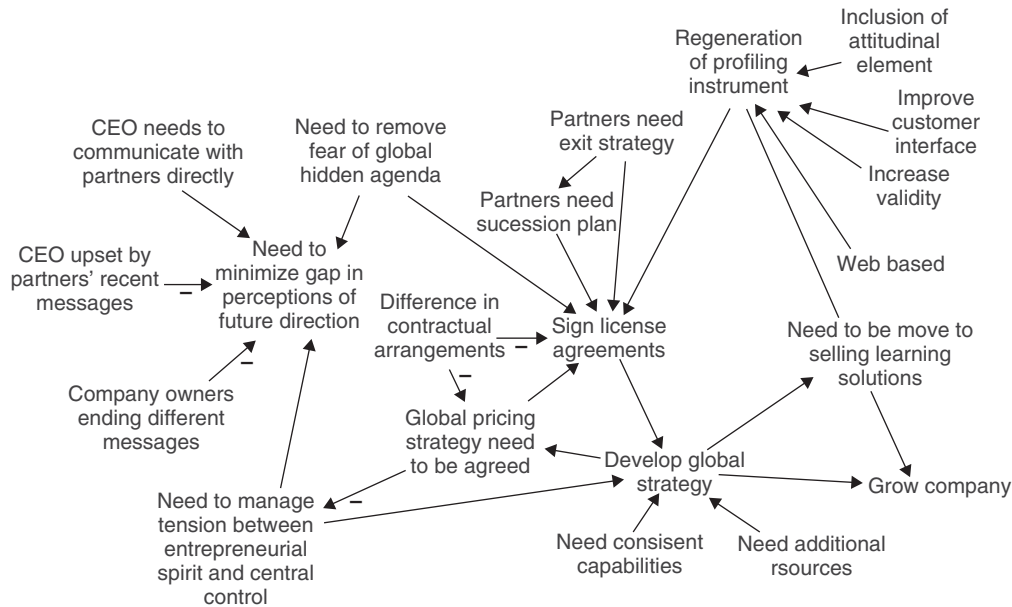


Figure 1. An example of a cognitive map, representing strategies for growth of an organization.

systems. A problem situation defined in this way will thus include factors that may not be typically regarded as legitimate variables in a standard MCDA modeling project, but that are nevertheless important if the analyst wishes to understand the needs and concerns of any particular client or client group. The challenge for the decision analyst is, therefore to be able to formally map aspects of the problem situation in terms of the concepts used by the client. For if there is a doubt in the client's mind about whether correct concepts have been taken into account, she/he is unlikely to believe in the solution to the problem, let alone act upon it [16].

For example, Fig. 1 illustrates a way to capture a client's understanding of a problem using the client's own concepts. The figure shows the beginning of a cognitive map that contains different aspects of a problem faced by an organization operating in the learning and professional development sector. Here the client is concerned about the growth of the organization, which eventually led to a multicriteria evaluation of strategic priorities at a later stage in the intervention. Nodes in the map contain statements describing different aspects of the problem. The links between

the statements denote means–end chains of arguments. For example, the “regeneration of profiling instrument” (top right in the map) is seen by this client as a way to get partners to “sign license agreements” (center right in the map).

The recognition that problem definition in organizations involves negotiation between managers with multiple world-views [20] about the problem has some practical implications for the analyst. First, if the MCDA intervention is intended to have some effect on the organization, the decision analyst may need to discuss a redefinition of the problem with the client before trying to help. The structuring tools cited above can all assist in this process [10]. Secondly, when working with members of a client group that have different views or interpretations of the problem situation of interest, the analyst must choose whose interpretation to pay attention to. The choice does not necessarily imply favoring one particular interpretation over another. Rather, it is about focusing on some combination of interpretations which for reasons that are explained later in the chapter, will often be a reflection of the analyst's understanding of the key stakeholders of the organization.

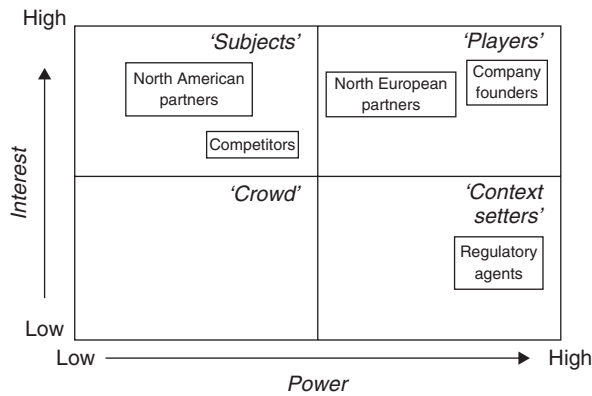


Figure 2. An example of a power–interest grid of stakeholders when considering strategies for growth of an organization.

Once the problem situation has been defined and agreed with the client or client group, the decision analyst should be in a good position to identify a particular decisional element of the situation upon which a relevant multicriteria evaluation model can be built. A quite useful tool at this stage is Keeney's concept of decision framing [6], which connects the strategic objectives of the organization with the fundamental objectives for the particular decision and the alternatives to be considered [as illustrated by Barcus and Montibeller [26]]. However, before proceeding, the decision analyst must scope the required levels of participation needed for the subsequent stages of the intervention. This aspect is discussed next.

Scoping Participation

Nutt conducted a careful analysis of 400 decisions in a variety of organizations and found that almost half of them 'failed' in terms of implementation (e.g., not implemented or only partially implemented) or the achieved results (e.g., poor results rather than good results) [27]. He discovered that the overriding reason for these failures was due, in large part, to the failure of decision makers to attend to the interests and information held by the *key stakeholders* of the organization. Although several definitions of stakeholders are possible [28], we conceptualize them here as those individuals, or groups, who have the power to affect the decision under consideration; or those groups that are affected, or perceived to be affected,

by the decision. This broad definition thus considers the internal as well as the external stakeholders of the organization.

Within the context of an MCDA intervention, attention to stakeholders is needed to assess and enhance political feasibility of decision implementation. Attention to stakeholders is also important to satisfy those involved in, or affected by, the decision that the intervention has followed rational, fair, and legitimate procedures. This does not imply that all possible stakeholders should be satisfied by or involved in the intervention; only that the key stakeholders must be. As in the case of defining the problem, the choice of which stakeholders are "key" should be the result of a discussion between the client and the analyst.

In the literature, there are several tools available for stakeholder analysis [28,29]. The most widely used techniques include the power–interest grid, star diagram, and stakeholder influence map [17]; and stakeholder–issue interrelation diagram and problem-frame stakeholder maps [28]. For example, Fig. 2 shows a power–interest grid for the problem situation discussed earlier. The grid arrays stakeholders on a two-by-two matrix, where the dimensions are the stakeholder's interest or stake in the decision at hand (i.e., they care about the decision or are affected by it), and the stakeholder's power to affect its implementation or impact. Four broad categories of stakeholders are shown in Fig. 2: 'players' who have both, an interest and significant power (e.g., Northern

European partners); ‘subjects,’ who have an interest but little power (e.g., North American partners); ‘context setters,’ who have power but little direct interest (e.g., regulatory agencies); and the ‘crowd,’ which consists of stakeholders with little interest or power. The grid allows the decision analyst to determine which players’ interests and power bases must be taken into account in order to address the decision at hand.

Whichever stakeholder identification technique is used, the actual process of choosing which stakeholders to involve in the intervention is often the result of several iterations along the following generic stages [28]:

- The analyst and client initiate the process by doing a preliminary stakeholder analysis, using any of the analysis techniques cited above. This step is useful in helping the client to think strategically about how to create the conditions needed for the intervention to reach a successful outcome.
- After reviewing the results of this analysis, a larger group of stakeholders can be assembled if judged appropriate. The assembled group should be asked to brainstorm the list of stakeholders, who might need to be involved in the intervention. Again, many of the techniques cited above might be used as a starting point. After this analysis has been completed, the analyst should encourage the group to think carefully about who is not at the meeting but that should be at subsequent meetings during the intervention. The analyst should ask the group to carefully think through the positive and negative consequences of involving—or not—other stakeholders or their representatives, and in what ways to do so.
- Last, both analyst and client finalize the various groups, who will have some role to play in the intervention. These will typically include the sponsors and champions, a coordinating group, a core decision analysis team, and various advisory or support groups [23].

The above process should be designed by the decision analyst to gain needed information, build political acceptance, and address some important questions about legitimacy, representation, and credibility [28]. However, the analyst should encourage the client to include stakeholders only when there are good and prudent reasons to do so. They should not be included when their involvement is not needed, impractical, or inappropriate.

Once the required participation is scoped, the next stage in the intervention process is to structure the MCDA evaluation model, which we present next.

STRUCTURING MCDA EVALUATION MODELS

There are three main tasks in structuring MCDA evaluation models: the *representation of objectives* in a value tree, the *definition of attributes* to measure the achievement of objectives, and the *identification of decision alternatives*. Below, we discuss each of these tasks and discuss the challenges that an analyst may encounter when undertaking them, as well as the tools/techniques that may be used to support their accomplishment.

Structuring Value Trees

The first step in building an MCDA evaluation model is always to represent the objectives that decision makers want to achieve (e.g., increase profitability, increase flexibility, reduce damage to the environment, and so on). In many multicriteria models, but particularly so in multi-attribute utility/value models [2], these objectives are organized as a value tree [1,30]. A value tree decomposes the overall objective of an evaluation into operational objectives, which can be more easily employed to assess the performances of decision alternatives. For example, Fig. 3 presents a value tree for evaluating different sites for building an industrial plant in Brazil. The client was concerned with the logistic costs associated with each site but also wanted to take into consideration the potential benefits from each site, such as accessibility to logistic

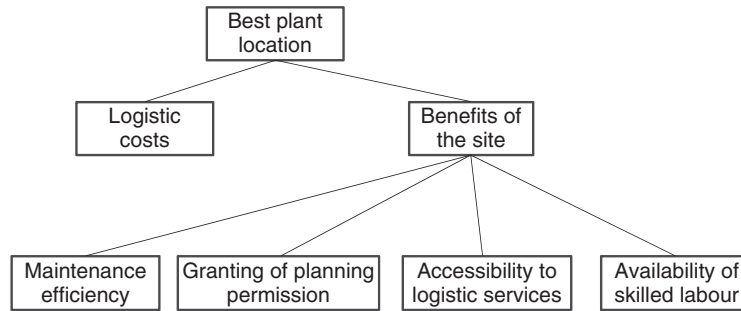


Figure 3. A value tree for selecting an industrial plant location.

systems (e.g., warehouses) and availability of skilled labor required for operating the plant.

Two approaches have classically been suggested for structuring a value tree [31,32]: top-down and bottom-up. The top-down approach is driven by the overall objective, which is then decomposed into objectives and the latter ones into sub-objectives, and so on. For example, if an analyst is structuring a value tree for the plant location problem described above, using a top-down approach, he/she would start with the overall objective (best location for the plant) and decompose it into logistic costs and benefits of the site. Each of these objectives could be decomposed even further, if required. The bottom-up approach is driven by the alternatives. In this case, the analyst would try to identify which attributes distinguish the alternatives and they would be included in the value tree. These attributes would then be grouped by their nature (e.g., in the plant location problem, all the attributes related to the potential benefits from a given site) and these groups could be further grouped upwards, composing the value tree.

There are compelling arguments that MCDA should employ a value-focused thinking approach for supporting decision making [6], as alternatives should be seen as mere means for organizations to achieve their fundamental and strategic objectives. This calls for a more top-down approach for structuring value trees. On the other hand, behavioral decision research has shown that individuals may struggle to think about their fundamental objectives [33], and may need

prompts from the analyst to reflect about the objectives prior to their explicit articulation. Behavioral research has also discovered that these two approaches (top-down and bottom-up) may generate value trees with different shapes [34], as values are “constructed” instead of merely extracted from decision makers’ minds [35]. Therefore, the choice of approach is clearly an important modeling decision that the analyst has to make.

Other possible tools for structuring a value tree involve the use of probes and grouping of ideas, such as Belton and Stewart’s CAUSE probes [1] and Parnell’s affinity diagrams [36]. Another set of tools for such purpose involves qualitative models that represent causality/influence between variables. Along these lines, Keeney [6] suggests the use of networks of means–ends objectives, where arrows represent the influence between a means and an end objective. Cognitive maps (illustrated in Fig. 1), a network of ideas connected by perceived influence and having a means–ends structure, have also been employed for structuring value trees [37–39] as discussed in Montibeller and Belton [40]. In a similar way, Merkhofer [41] suggests the use of qualitative influence diagrams to help the structuring of value trees. The main advantage of using these causality/influence tools is that they permit laddering-up toward the decision makers’ values, and laddering-down toward the attributes and decision alternatives, in a systematic and integrated way.

Objectives in a value tree must follow a set of properties that need to be checked when structuring it [1,2,6]. These properties are the following:

- *Essential*. They should consider all the essential organizational objectives involved in the decision.
- *Understandable*. They should have a clear meaning for all the members of the group involved in making the decision.
- *Operational*. It should be possible to measure the performance of decision alternatives against each of the fundamental objectives.
- *Nonredundant*. They should not measure the same concern twice.
- *Concise*. It should be the smallest number of objectives required for the analysis.
- *Preferentially independent*. If it is possible to measure the performance of decision alternatives on one objective disregarding their performance on all other objectives, then a simpler aggregation function can be used to aggregate partial performances.

Checking that these properties are observed in practice will usually, impact on the structure of a value tree. For example, a new objective may be included if the initial set does not cover all the essential issues in the evaluation. An objective may be removed, if it is not operational (e.g., if the information is considered as important but is unobtainable) or if it is redundant. Concerns about conciseness also can reduce the size of a value tree. Finally, if there are objectives that are preferentially dependent, the analyst may choose to restructure them to avoid using a complex aggregation function (for a detailed discussion on how to deal with preferential dependences, see Ref. 6).

Defining Attributes

For each objective placed at the bottom level of the value tree, an associated attribute or criterion should be specified. This attribute is a performance indicator employed to measure the impact of adopting each decision alternative on the organizational objective being pursued. There are two dimensions for classifying attributes in terms of: 1) the way it is measured; and 2) its alignment

with the objective being pursued [6,36,42]. We describe these two dimensions below.

The way the objective is measured: Direct or Indirect

- A *direct attribute* measures directly the degree of attaining the objective. For example, in Fig. 3, logistic costs have a direct attribute: the total logistic cost in US dollars.
- A *proxy attribute* measures indirectly the concern expressed by the objective, by assessing the degree of achievement of its associated objective. For instance, in the value tree shown in Fig. 3, the concern about having the planning permission granted is assessed by the number of months required for the processing of such permission.

The type of attribute: Natural or Constructed

- *Natural attributes* measure directly the concern expressed by the objective, are of general use and have a common interpretation. An example, in the value tree shown in Fig. 3, is to measure the logistic costs in US dollars.
- *Constructed attributes* measure directly, using indicators created specifically by the analyst, the concern expressed by the objective. In the plant location example, the availability of skilled labor (Fig. 3) is measured by a set of labels ranging from the best level ("wide availability of skilled labor from similar production plants in the region") to the worst one ("the plant will need to provide training to all its new employees").

Attributes can then be classified using these two dimensions; for example, a direct-natural attribute or a direct-constructed one. In terms of the way the objective is measured, whenever possible, it is usually better to use a direct attribute instead of a proxy one. If a direct attribute is not available, many times it is feasible to decompose an objective into subobjectives—with these subobjectives being assessed via direct

attributes—but avoiding excessive decomposition. In the same way, regarding the type of the attribute, a natural attribute is typically better than a constructed one, if the former is available and provides a clear way for decision makers to assess the alternatives. (See Refs 36, 42, and 43 for a comprehensive discussion on defining attributes and guidelines on how to develop suitable ones.)

Independently of its type, each attribute should possess five properties [42] if it were to be employed in a MCDA evaluation model:

- *Unambiguous.* The attribute should present a clear relationship between the impact of adopting a decision alternative and the description of such impact.
- *Comprehensive.* The attribute should cover the full range of possible consequences, if the decision alternatives were implemented.
- *Direct.* The attribute levels should describe as directly as possible the consequences of implementing a decision alternative.
- *Operational.* The information required by the attribute can be obtained in practice and it is possible to make value trade-offs between objectives [1,2].
- *Understandable.* Consequences and value trade-offs using the attribute can be clearly understood by the decision making group and communicated to other stakeholders.

Quantitative attributes tend to be less ambiguous than qualitative ones. A key point about comprehensiveness is that the upper and lower limits of the attribute are well-specified (maximum feasible and minimum acceptable, respectively) otherwise it would distort value trade-offs. Finally, it is critical that attributes are understandable, particularly if the analysis involves a group of decision makers and the modeling is conducted in a facilitated mode [44], such as in a decision conference [45].

Identifying Decision Alternatives

The other major task in structuring an MCDA evaluation model is the definition of

which decision alternatives will be assessed by the model. Traditionally, MCDA has taken an alternative-focused thinking perspective, where the set of options was assumed as given and stable [3]. However, the identification and creation of new alternatives is certainly one of the most important aspects of any MCDA intervention. No matter how careful and sophisticated the evaluation model is; if the decision alternatives under consideration are weak, it will lead to a poor choice [46].

An important aspect in structuring an MCDA model is that the decision alternatives should have the same nature (in the plant location example, for instance, all the alternatives are potential sites). If the analyst is careless about this aspect, it may be difficult to create a coherent value tree. There are several tools that may be employed in the creation/definition of decision alternatives, such as brainstorming techniques [47], cognitive mapping [18], dialog maps [22] among others.

Particularly useful tools are the ones, where decision alternatives are created from considering the decision makers' objectives [6] or stakeholders' values [48]. For example, the analyst can ask the decision makers to imagine options that could perform really well on a single objective. This process can be repeated for each of the fundamental objectives present in the value tree. Once the list of objectives is exhausted, the same procedure can be done for two objectives at once. Another way of creating a new option is by combining the existing alternatives, trying to maintain the best features of each alternative. Recently we have used a value-focused brainstorming using a cognitive map—which allowed eliciting, organizing, and displaying a large set of ideas from a client group—these ideas were then grouped as decision alternatives [39]. [For an extensive review of tools for creating alternatives see Keeney [6], Keller and Ho [49] and Parnell *et al.* [29].]

Although there is a natural tendency by decision makers to discard decision alternatives or options that may appear to generate some negative outcomes, any attempt at option evaluation should be contained at this stage. The assessment of alternatives should

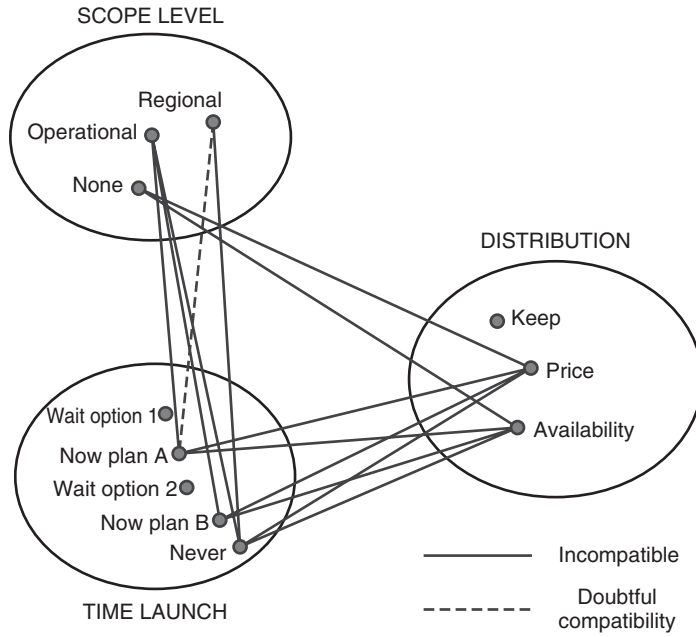


Figure 4. An example of areas of interconnected options.

be left for the evaluation phase of the process and not intermingled with their generation.

Another aspect concerning the identification of decision alternatives is that there are instances where the alternatives are comprised by a large set of sub-options. There are some methods that can be used to structure complex decision alternatives. The strategy generation table proposed by Howard [50] is a simple way of creating decision strategies from the combination of options under several dimensions. Another tool is the analysis of interconnected decision areas (AIDA) technique that is a part of the strategic choice approach [23]. In this technique, the links between several “decision areas” are represented, each one with several options, with their compatibility explored, in order to generate a list of possible option portfolios. For example, in an intervention with a major international hotel company, we used AIDA to initially shape a strategic decision concerning how to tackle “cost of sale,” and produced a list of candidate interconnected strategic options, grouped in three areas (*distribution, timing launch, and scope level*). This is shown in Fig. 4, where the links between

specific options represent incompatible combinations.

AN INTERVENTION FRAMEWORK

The techniques for multicriteria evaluations are already well-established in the literature. However, there has been much less investment in the development of techniques to support the structuring stages of MCDA interventions. We have reviewed both the mainstream problem structuring and MCDA literatures, and identified a number of modeling tools, which can be used to support problem structuring in MCDA interventions. Perhaps, more importantly, our foregoing discussion should have made clear to the reader of the important role that problem structuring plays in MCDA interventions.

In Fig. 5, we suggest a framework for conducting MCDA interventions, in which the role of problem structuring is made explicit. In Phase 1, the analyst structures the problem situation, helping the client to arrive at an agreed problem definition, and designs a decision process with the appropriate level of participation. Once this phase is finished,

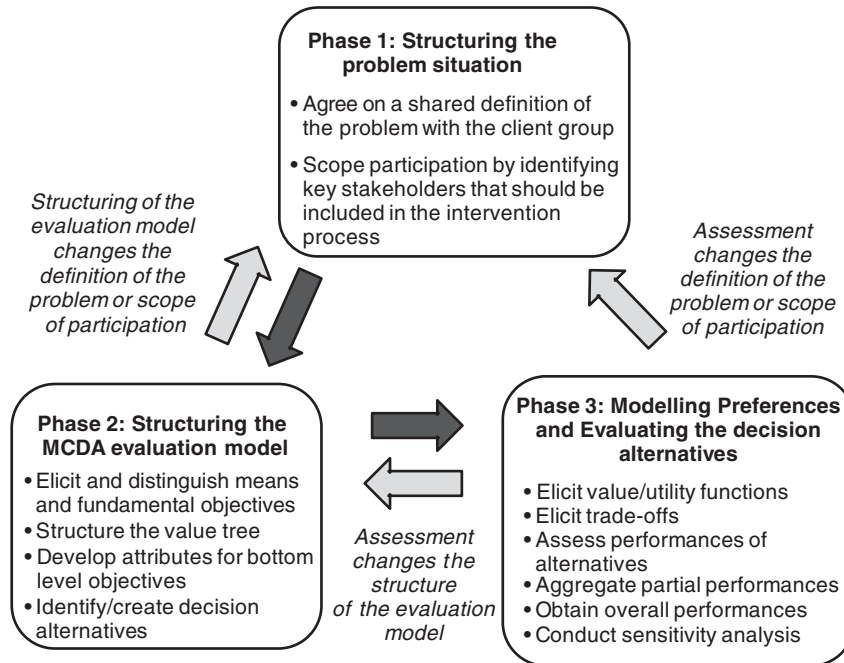


Figure 5. A framework for structuring MCDA models.

the analyst then can start Phase 2, the structuring of an MCDA model, which consists of eliciting and distinguishing means and fundamental objectives, structuring a value tree, developing attributes and identifying decision alternatives. With this second phase completed, the analyst can finally proceed to undertake Phase 3, modelling preferences and the evaluation of decision alternatives. The natural flow of phases is indicated with black arrows in Fig. 5, but notice that the process is not linear and can cycle back to earlier phases (grey arrows): back from Phase 2 to 1, if the structuring of the MCDA model changes the definition of the problem or the scope of stakeholders' participation; back from Phase 3 to 2, if modelling of preferences or the assessment of alternatives changes the structure of the MCDA model; and back from Phase 3 to 1, if modelling of preferences or the assessment of alternatives changes either the definition of the problem or the participation required. Table 1 contains a list of useful tools for supporting the different activities within each of the structuring phases of an MCDA intervention.

CONCLUSIONS AND DIRECTIONS FOR RESEARCH

While decision analysts have recognized for a long time the importance of problem structuring for successful MCDA interventions, most of them have relied on ad hoc practices for structuring the decision problem. The main aim of this chapter has been to provide a review of tools that can help this pre-MCDA evaluation phase of problem structuring. Furthermore, we have also reviewed the main tasks involved in building an MCDA evaluation model per se, while attempting to provide a more integrated view by relating these tasks with the problem-structuring literature.

As discussed in this chapter, there are a number of problem-structuring tools available to help decision analysts deploy effective MCDA interventions. However, our discussion should have also made clear that when the client comprises a group of managers, mastering the tools will not be sufficient. The analyst will also need skills for facilitating the group processes associated with

Table 1. Tasks and Tools for Structuring MCDA Models

Phase 1: Problem Structuring		
Activity	Task	Supporting Tools and Useful References
Defining the Problem	Capture the different understandings about the multicriteria problem and facilitate a definition of the problem that is shared by the client (or client group).	• Cognitive mapping [18,19] • Dialog mapping [22] • Soft systems methodology [20,21] • Strategic choice approach [23] • Group model building [24] • Decision framing [6]
Scoping Participation	Determine the type and level of participation of different stakeholders required for the intervention.	• Power-interest grid; star diagrams and stakeholder influence diagrams [17] • Stakeholder-issue interrelation diagram and problem-frame stakeholder maps [28]
Phase 2: Structuring the MCDA Evaluation Model		
Activity	Task	Supporting Tools and Useful References
Structuring Value Trees	Organize the objectives to be considered in the evaluation as a hierarchy.	• Top-down or bottom-up approaches [31] • Checklist and grouping of ideas [1,36] • Means-ends objective networks [6] • Cognitive maps [37–39] • Qualitative influence diagrams [41] • Checklist of properties for a value tree [1,6]
Defining Attributes	Specify, for each bottom level objective in the value tree, an associated attribute.	• Keeney's and Gregory's [42] decision model for selecting attributes and Parnell's [36] preference ranking for selecting attributes • Kirkwood's [43] classification of attributes and guidelines for their development • Checklist of properties for an attribute [6]
Identifying Decision Alternatives	Define/identify/create decision alternatives to be assessed by the MCDA model.	• Brainstorming [47] • Laddering-down in a cognitive map [18,19] • Dialog maps [22] • Focus on the objectives to be achieved [6,49] • Ideation techniques [29] • Strategy tables [50] • Analysis of interconnected decision areas [23]

defining the problem, which are significantly influenced by the power and interests of the managers in that group [16,44].

It worth noting that the chapter has focused on modeling decision making with multiple objectives. Frequently, however, key uncertainties are present and should also be represented. Useful tools for modeling

decision making under uncertainty are influence diagrams [51] and decision trees [52]. A good introduction to this type of modeling is provided by Clemen and Reilly [53] and Kirkwood [54].

We believe that problem structuring for MCDA is a rich field of research, where the focus can be not about developing and testing

suitable problem structuring tools, as well as about the study of facilitated modeling in this intervention context. We thus suggest some directions for further research:

- *Development of Problem-Structuring Methods.* While the field of problem-structuring methods (PSMs) is already well-established in management science, more research could be conducted on tools that could be tailored specifically for MCDA interventions.
- *Integrated Use of PSMs.* The use of standard PSMs with MCDA requires transitions from a problem-structuring model to a multicriteria decision analysis model, which may prove challenging [40]. Consequently, a direction of research is the development of methods that could provide a seamless transition. The reasoning maps method, suggested by Montibeller *et al.* [55], and the use of means objectives to assess the performance of decision alternatives on fundamental objectives, suggested by Butler *et al.* [56] are examples of research in this direction.
- *Tools for Supporting Structuring MCDA Tasks.* The paper reviewed some tools that could be employed for structuring value trees, defining attributes and identifying decision alternatives. The development of new tools is, however, still a potentially area of research—particularly if it were based more on psychological aspects [e.g., how to spark off creativity when creating alternatives; how to identify/display complex options to a group of decision makers, such as the approach proposed by Montibeller *et al.* [39]].

To summarize, this chapter provided an overview of the phases and tasks involved in structuring an MCDA evaluation model within an intervention, from defining the problem and identifying key stakeholders to building the MCDA model itself. Problem structuring is a fundamental and challenging task for any MCDA intervention; thus, we hope this chapter may be of help to decision analysts involved in such interventions

and may also serve as a useful resource for researchers interested in this field.

Acknowledgment

We are thankful for the insightful comments provided by an anonymous referee; this helped us to improve the draft.

REFERENCES

1. Belton V, Stewart TJ. Multiple criteria decision analysis: an integrated approach. Dordrecht: Kluwer; 2002.
2. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. 2nd ed. Cambridge (MA): Cambridge University Press; 1993.
3. Roy B. Multi-criteria methodology for decision aiding. Dordrecht: Kluwer; 1996.
4. Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys. New York: Springer; 2005.
5. Keefer DL, Kirkwood CW, Corner JL. Perspective on decision analysis applications, 1990–2001. *Decis Anal* 2004;1(1):5–24.
6. Keeney RL. Value-focused thinking: a path to creative decision-making. Cambridge (MA): Harvard University Press; 1992.
7. von Winterfeldt D, Fasolo B. Structuring decision problems: a case study and reflections for practitioners. *Eur J Oper Res* 2009;199(3):857–866.
8. Belton V, Stewart TJ. Problem structuring and MCDA. In: Ehrgott M, Figueira J, Greco S, editors. Trends in multiple criteria decision analysis. Boston (MA): Springer; In press.
9. Nutt PC. Formulation tactics and the success of organizational decision making. *Decis Sci* 1992;23(3):519–540.
10. Rosenhead J, Mingers J, editors. Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict. Chichester: Wiley; 2001.
11. Smith GF. Classifying managerial problems: an empirical study of definitional content. *J Manage Stud* 1994;32:679–706.
12. Lyles MA, Mitroff II. Organizational problem formulation: an empirical study. *Adm Sci Q* 1980;25:109–119.
13. von Winterfeldt D, Edwards W. Defining a decision analytical structure. In: Edwards W, Miles RF, von Winterfeldt D,

- editors. *Advances in decision analysis: from foundations to applications*. Cambridge: Cambridge University Press; 2007. pp. 81–103.
14. Ackoff R. *Redesigning the future: a systems approach to societal problems*. New York: Wiley; 1974.
 15. Mitroff IL, Ernshoff JR. On systemic problem solving and the error of the third kind. *Behav Sci* 1974;19:383–393.
 16. Eden C, Sims D. On the nature of problems in consulting practice. *Omega Int J Manage Sci* 1979;7(2):119–127.
 17. Eden C, Ackermann F. *Strategy Making: the journey of strategic management*. London: Sage; 1998.
 18. Eden C. Cognitive mapping: a review. *Eur J Oper Res* 1988;36(1):1–13.
 19. Eden C. Analyzing cognitive maps to help structure issues or problems. *Eur J Oper Res* 2004;159(3):673–686.
 20. Checkland P. *Systems thinking, systems practice*. Chichester: Wiley; 1981.
 21. Checkland P, Scholes J. *Soft systems methodology in action*. Chichester: Wiley; 1990.
 22. Conklin J. *Dialog mapping: building shared understanding of wicked problems*. Chichester: Wiley; 2006.
 23. Friend J, Hickling A. *Planning under pressure: the strategic choice approach*. 3rd ed. Oxford: Elsevier; 2005.
 24. Vennix J. *Group model building: facilitating team learning using system dynamics*. Chichester: Wiley; 1996.
 25. Richardson G, Andersen D. Teamwork in group model building. *Syst Dyn Rev* 1995;11(2):113–137.
 26. Barcus A, Montibeller G. Supporting the allocation of software development work in distributed teams with multi-criteria decision analysis. *Omega (Westport)* 2008;36(3):464–475.
 27. Nutt PC. *Why decisions fail? Avoiding the blunders and traps that lead to debacles*. San Francisco (CA): Berrett-Koehler Publishers; 2002.
 28. Bryson JM. *What to do when stakeholders matter: stakeholder identification and analysis techniques*. *Public Manage Rev* 2004;6(1):21–53.
 29. Parnell GS, Driscoll PJ, Henderson DL, editors. *Decision making for systems engineering and management*. Wiley Series in Systems Engineering, Sage AP, editor. Hoboken (NJ): John Wiley & Sons, Inc.; 2008.
 30. Goodwin P, Wright G. *Decision analysis for management judgment*. 4th ed. Chichester: Wiley; 2009.
 31. Buede DM. Structuring value attributes. *Interfaces* 1986;16(2):52–62.
 32. von Winterfeldt D, Edwards W. *Decision analysis and behavioral research*. Cambridge (MA): Cambridge University Press; 1986.
 33. Bond SD, Carlson KA, Keeney RL. Generating objectives: can decision makers articulate what they want? *Manage Sci* 2008;54(1):56–70.
 34. Morton A, Fasolo B. Behavioural decision theory for multi-criteria decision analysis: a guided tour. *J Oper Res Soc* 2009;60(2):268–275.
 35. Slovic P. The construction of preference. *Am Psychol* 1995;50:364–371.
 36. Parnell GS. Value-focused thinking. In: Rainey L, Loerch A, editors. *Methods for conducting military operational analysis*. Alexandria (VA): Military Operations Research Society; 2007. pp. 619–656.
 37. Bana e Costa CA, Ensslin L, Correa EC, *et al.* Decision support systems in action: integrated application in a multi-criteria aid process. *Eur J Oper Res* 1999;113:315–335.
 38. Belton V, Ackermann F, Shepherd I. Integrated support from problem structuring through to alternative evaluation using COPE and VISA. *J Multi Criteria Decis Anal* 1997;6:115–130.
 39. Montibeller G, Franco LA, Lord E, *et al.* Structuring resource allocation decisions: a framework for building multi-criteria portfolio models with area-grouped projects. *Eur J Oper Res* 2009;199(3):846–856.
 40. Montibeller G, Belton V. Causal maps and the evaluation of decision options: a review. *J Oper Res Soc* 2006;57(7):779–791.
 41. Merkhofer MW. Using influence diagrams in multi-attribute utility analysis: improving effectiveness through improving communication. In: Olivier RM, Smith JQ, editors. *Influence diagrams, belief nets and decision analysis*. Chichester: Wiley; 1990. pp. 297–317.
 42. Keeney RL, Gregory RS. Selecting attributes to measure the achievement of objectives. *Oper Res* 2005;53(1):1–11.
 43. Kirkwood CW. *Strategic decision making: multiobjective decision analysis with spreadsheets*. Belmont (CA): Duxbury Press; 1997.

44. Franco LA, Montibeller G. Facilitated modelling in operational research (invited review). *Eur J Oper Res*. 2010;205(3):489–500.
45. Phillips L. Decision conferencing. In: Edwards W, Miles R Jr, von Winterfeldt D, editors. *Advances in decision analysis: from foundations to applications*. New York: Cambridge University Press; 2007. pp. 375–399.
46. Brown R. Rational choice and judgment: decision analysis for the decider. New York: Wiley; 2005.
47. Osborn AF. *Applied imagination: principles and procedures of creative problem-solving*. New York: Charles Scribner's Sons; 1957.
48. Gregory R, Keeney RL. Creating policy alternatives using stakeholder values. *Manage Sci* 1994;40(8):1035–1048.
49. Keller LR, Ho J. Decision problem structuring: generating options. *IEEE Trans Syst Man Cybern* 1988;18(5):715–728.
50. Howard RA. Decision analysis: practice and promise. *Manage Sci* 1988;34(6):679–695.
51. Howard RA, Matheson JE. Influence diagrams. In: Howard RA, Matheson JE, editors. *The principles and applications of decision analysis*, Vol 2. Menlo Park (CA): Strategic Decisions Group; 1981.
52. Raiffa H. *Decision analysis: introductory lectures on choices under uncertainty*. Oxford: Addison-Wesley; 1968.
53. Clemen R, Reilly T. *Making hard decision with decision tools*. Pacific Grove (CA): Duxbury; 2001.
54. Kirkwood CW. An overview of methods for applied decision analysis. *Interfaces* 1992;22(6):28–39.
55. Montibeller G, Belton V, Ackermann F, *et al.* Reasoning maps for decision aid: an integrated approach for problem-structuring and multi-criteria evaluation. *J Oper Res Soc* 2008;59(5):575–589.
56. Butler JC, Dyer JS, Jia JU. Using attributes to predict objectives in preference models. *Decis Anal* 2006;3(2):100–116.

PROCUREMENT CONTRACTS

YEHUDA BASSOK
Department of Information and
Operations Management,
University of Southern
California, Los Angeles,
California

INTRODUCTION

The problem of securing a supply of raw materials, components, subassemblies, and finished goods at a stable and known price is not a new one. It is one of the main difficulties of every industrial (and not only industrial) organization. This problem is becoming even more acute due to the common practice of outsourcing and decentralization. In a decentralized environment, independent entities, each with its own plans and objectives, are purchasing raw materials, components, subassemblies, and finished products from each other. Clearly, each entity is interested in ensuring that it obtains the right quantity of inputs, of the right quality, at the right time, while minimizing cost. Coordinating the supply chain is complex and difficult. Some of the main reasons are as follows: players in the supply chain are facing uncertainties regarding the supply and demand; production lead times are long, and thus procurement managers are forced to place orders long before they observe the actual demand; their decisions are based on forecasts that are usually not accurate; economies of scale often force decision makers to place large orders and thus prevent them from observing the actual demand and continuously adjusting to the changing market conditions. Finally, one of the most important difficulties is that each entity in the supply chain is interested in maximizing its own utility but not necessarily the utility of the entire supply chain.

The issue of designing supply contracts is not new, but today it has become a crucial element of the procurement function. This is

due to the decentralization of companies and the fact that outsourcing of services and products is by now the common mode of operation. This is in contrast to the not-so-distant past, in which industrial companies were vertically integrated and coordinating the supply channel was much easier. In such environments, a single decision maker made the decisions for all of the involved entities of the organization, thereby optimizing the performance of the entire organization and not just the performance of each of the entities as a “stand-alone” unit.

Usually, supply contracts are part of the process of procuring goods in manufacturing and retail industries. More recently, it has been observed that such contracts are also an important part of health-care delivery, especially the global health-care delivery system. The reason is that many of the patients live in low-income countries and are too poor to pay the full price of expensive medications. In such environments, where markets do not exist, manufacturers are reluctant to invest in research and development of new drugs and in production capacity. We believe, as we describe in what follows, that supply contracts are an important tool to ensure availability and affordability of drugs.

In this article, we concentrate on contracts that deal with risk and uncertainty: in most cases, we deal with the uncertainty in the demand for goods and services. But, contracts are also used in other cases, especially in cases in which there is no uncertainty. A good example is quantity discount contracts. Quantity discount contracts can be applied in a deterministic environment. The main idea is to induce buyers and sellers to produce and purchase quantities that generate operational efficiencies. Such contracts are common when economies of scale are presented. Transportation systems are a good example. Sellers offer price discount to induce buyers to purchase larger quantities than those that they would have purchased otherwise. By doing so, sellers are able to ship full truck loads and reduce the per unit ship-

ping cost. The interested reader is referred to the seminal paper by Rosenblatt and Lee [1]. A description and classification of quantity discount contracts is in Munson and Rosenblatt [2].

For obvious reasons, we are unable to discuss all types of procurement contracts here. We describe some of the most important contracts in short. A more inclusive and detailed review can be found in Cachon [3]. The section titled “Contracts in Supply Chains” in this encyclopedia provides a very good discussion on the many different aspects of contracts.

FLEXIBILITY AND STABILITY

As mentioned above, the main objective of supply contracts is to ensure the availability of the right quantity and quality of goods and services at the right time and at the lowest possible cost. In most of the contracts that we discuss here, one of the main drivers of cost is risk. More specifically, all supply contracts are concerned with the allocation of risk between the different entities in the supply chain. (Here referred to as *suppliers and buyers*). The question that all supply contracts are addressing is the question of who should bear the risk of producing ahead of time and keeping raw materials, components, and finished goods. Suppliers like a stable environment in which all orders are

firm, and once a buyer places an order the buyer must take and pay for the goods. On the other hand, buyers like the flexibility to place orders and then change these orders as they learn more about the actual demand. We argue that in many cases the right balance between stability and flexibility maximizes the utility of the supply chain and can ensure a “win-win” situation.

A good example that illustrates this phenomenon was provided long ago by Blois [4], as shown in Table 1.

One can see how the monthly order changes as we approach the delivery time. For example, in March, the scheduled delivery for December was 22,700 units, but in June the scheduled quantity was reduced to 14,500. It is not difficult to imagine the difficulties that the supplier is facing by such big changes in the orders and the resulting additional cost. Thus, the role of a supply contract is not only to ensure demand (for the supplier), supply (for the buyer), and to improve the production planning process but also to create a flexible environment in which buyers can adjust to the changing conditions of the market. At the same time, it is necessary to keep the suppliers’ environment stable without imposing too much risk and costs on the suppliers.

To further illustrate the trade-off between stability and flexibility, we use the graph (Fig. 1) provided by Rietveld Hans [5].

Table 1. Company B’s Schedules of its Requirements from Company A

Schedule Issued on:		12/1/60	29/3/60	29/6/60	18/7/60	23/11/60
Delivery Required in:						
July ^a	1960	10,000				
August ^b		25,000		16,800		
September		20,000		13,500		
October ^b			28,400	18,100		
November			22,700	14,500		
December			22,700	14,500		5400
January ^b	1961				12,900	5700
February					10,300	5400
March					10,300	5400
April						5700
May ^b						5400
June						6200

^aHoliday month—only two weeks production.

^bFive week month.

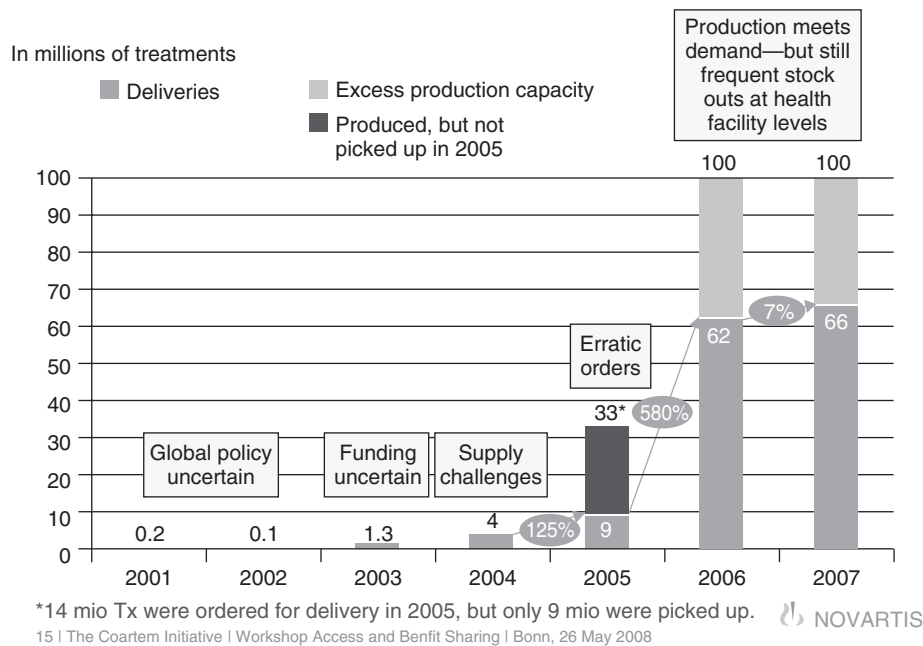


Figure 1. Production capacity confirmed orders and actual deliveries.

This graph illustrates several key issues crucial to the design of supply contracts. Despite the high demand for Malaria drugs (300–500 M instances per year) and despite the available capacity (100 M units), the manufacturer does not use all of its capacity. While there are several reasons for this, one of the main reasons mentioned by the manufacturer (Novartis) is the uncertainty in the delivery quantity. In 2005, a large quantity was ordered but only a relatively small quantity was picked and delivered. The reason is that orders are placed long in advance because of the long production lead time. When delivery time arrives, the market conditions may be different, forcing the buyers to change their previously made orders, resulting in a large loss to the manufacturer. This graph illustrates the general issue of balancing flexibility and stability. Buyers would like the flexibility to change previously made orders to better adjust to changing market conditions. Suppliers, on the other hand, would like to operate in a more stable environment in which the flexibility to change previously made orders is

limited. If the “right” balance between flexibility and stability is not achieved, sellers (buyers) may produce (purchase) less than the desired quantity and reduce their exposure to risk. How to balance flexibility and stability and allocate the risk between buyers and manufacturers is exactly the role of the procurement contracts that are discussed in this article.

MINIMUM COMMITMENTS

The total-minimum-quantity contract addresses two different issues as follows; (i) It gives the buyer some flexibility to change orders and to adjust to changing market conditions, but at the same time it limits the buyer’s flexibility to change order sizes drastically, which helps in avoiding increases in the production cost for the supplier. (ii) It provides incentives to suppliers to build capacity even when the demand is uncertain, and it shares the risk of investing in capacity between the supplier and the buyers.

Bassok and Anupindi [6] use the framework of a multiperiod news-vendor model,

which assumes that the periodical (month, week) demand is unknown but that a forecaster is able to provide a probability distribution of the demand. Also, in this model, it is assumed that the demand is stationary and independent—that is, the demand in each period follows the same probability distribution and demands across periods are not correlated. The authors consider a finite horizon problem that is a problem with a limited number of periods (for example, one year). The optimal solution for this problem is well known: every period has a critical number S_t (the “base stock”), so that in period t it is optimal to raise the available inventory (on-hand inventory) and order up to the base stock. The reader will notice that in such a model the buyer firm has all the flexibility to order any quantity it wishes, which means that the supplier environment is far from stable. This “order-up-to” policy obviously minimizes the buyer’s cost, but just as obviously it ignores its effects on the supplier’s cost.

To mitigate the negative effects of demand uncertainty on the supplier’s performance, the buyer is required to announce its minimum purchasing quantity for the entire horizon. For making the minimum purchasing commitment, the supplier offers a discounted purchasing price per unit. By making the “minimum commitment,” the buyer limits its own flexibility. If, for example, the total demand during the horizon turns out to be lower than the total commitment, then the buyer must still take delivery of the committed quantity and pay for it. If the total demand turns out to be higher than the total commitment, then the buyer will have to purchase the difference between the total demand and the total commitment at a higher price. On the other hand, the buyer still has the flexibility to place periodical orders of any size, as long as the sum of the periodical orders is not greater than the “minimum commitment.”

The supplier also benefits from such a contract because it knows with certainty that during the horizon it will have a buyer for the committed minimum quantity, which means the firm can plan its production in an efficient way.

The authors concentrate on the effects of the minimum commitment contract on the buyer’s performance. They are mainly interested in determining the optimal purchasing policy of the buyer assuming that it makes a minimum commitment to purchase K units during the horizon. They prove that the optimal purchasing policy of the buyer has a very simple structure and that it is very easy to calculate the optimal periodical purchasing quantity. The crux of the optimal policy is as follows. At the beginning of every period, if the sum of the buyer’s earlier purchases is still smaller than the minimum commitment, then it is optimal for the buyer to raise its inventory to the base stock S^M , where S^M is the optimal base stock for a single-period news-vendor problem assuming that the purchasing cost is equal to zero. If, at the beginning of a period, the sum of earlier purchases is larger than the minimum commitment, then the buyer needs to solve a standard multiperiod problem without any commitments.

By following such a simple policy, the buyer minimizes its cost. Clearly, the supplier is interested in the buyer making a large commitment and, to encourage the buyer to do so, the supplier is willing to provide the buyer with a per-unit discount. The effect of the commitment on the buyer’s cost is illustrated by the following graph (Fig. 2) [6].

The graph shows us that as the total commitment is increasing the total saving is decreasing. This is expected because the larger the commitment, the larger is the buyer’s risk. The graph is useful in determining the optimal commitments. For example, it is better for the buyer to commit for 750 units with a 5% discount than for 900 units with 10% discount.

The minimum-commitment contract can be used to encourage companies to build production capacity. This is especially true in environments in which there are no markets for the produced goods. A good example is the sub-Saharan countries, in which most of the population is unable to purchase life-saving drugs. Private and public organizations are willing to subsidize the drugs and to ensure that they are available and affordable to the needy, but usually the size and the duration

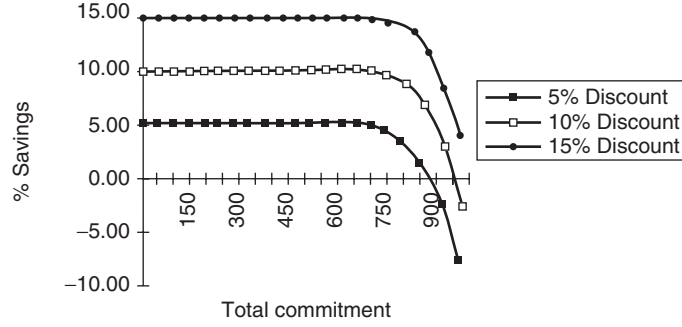


Figure 2. Effect of price discount.

of subsidy is not well defined and is uncertain. As a result, drug manufacturers are unwilling to take the risk and build capacity to produce drugs for low-income countries.

To solve this problem, several researchers have suggested that the private and public organizations make clear and well-defined commitments to subsidize certain drugs (for example, malaria drugs). The total amount and the duration of such a commitment should be advertised ahead of time and ensured by a third party. Finally, a neutral organization (e.g., the World Health Organization) will have to provide a list of drugs that should be subsidized by the private and public organizations. By providing a total minimum-quantity commitment, the private and public donors are able to mimic market conditions and provide a more stable environment for the manufacturers, who will now be certain that they will sell at least the minimum-commitment quantity. Clearly, this is an incentive for pharmaceutical companies to enter the market and invest in production capacity.

PERIODICAL COMMITMENTS AND FLEXIBILITY

The total-minimum-quantity commitment solves several problems, especially ensuring that suppliers will sell at least a minimum quantity during the horizon. But it does not deal with the changing periodical orders. Buyers still have the flexibility to change their previously made orders to better meet the actual demand. It is clear that buyers would like the flexibility to change

previously made orders at any time, but it is also evident that suppliers would like to limit this flexibility, ensure stability, and force the buyer to commit and make firm periodical orders. Bassok *et al.* [7], Anupindi and Bassok [8], and Tsay and Lovejoy [9] all address the issue of balancing flexibility and stability.

Common to these studies are several concepts. The buyer, before the horizon starts, makes periodical commitments $q_{0,1}, q_{0,2}, \dots, q_{0,N}$, where $q_{i,j}$ is the commitment made at period i for period j . At the beginning of period 1, the buyer may change the order quantity to be delivered during this period $q_{1,1}$ to be: $(1 - l_0)q_{0,1} \leq q_{1,1} \leq (1 + u_0)q_{0,1}$. Thus, the buyer has the downward flexibility of l_0 and the upward flexibility of u_0 . Thus, the buyer does not have the infinite flexibility to purchase any quantity it wishes but rather must stay within the limits determined by its original commitment and the downward and upward flexibility limits. In addition, the buyer may change its future commitment in a restricted way

$$\begin{aligned} (1 - l_1)q_{0,2} &\leq q_{1,2} \leq (1 + u_1)q_{0,2} \\ (1 - l_2)q_{0,3} &\leq q_{1,3} \leq (1 + u_2)q_{0,3} \\ (1 - l_{N-1})q_{0,N} &\leq q_{1,N} \leq (1 + u_{N-1})q_{0,N}, \end{aligned}$$

where $l_1 \leq l_2 \leq \dots \leq l_{N-1}$ and $u_1 \leq u_2 \leq \dots \leq u_{N-1}$.

Anupindi and Bassok [8] graphically describe the dynamics of this contract in the following way (Fig. 3):

The current period is represented by a diamond and the future periods by a rectangle. Suppose the commitment for period 3

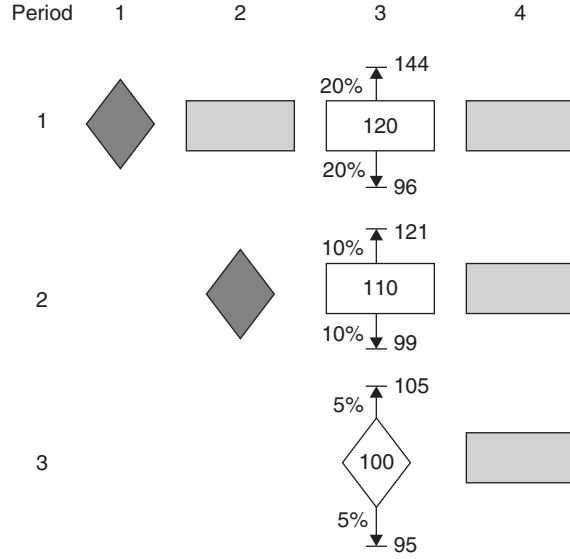


Figure 3. The dynamics of the periodical commitment and flexibility contracts.

made in period 0 was 120 units. In period 1, this commitment can be adjusted upward or downward by 20%. Thus, in period 1, the buyer can adjust the commitment in period 3 to any quantity between [96,144]. Suppose the buyer decides, at the beginning of period 1, to commit to purchase 110 units. In period 2, the commitment made for period 3 can be adjusted upward or downward only by 10%—that is, the adjusted commitment could be between [99,121]. Now let us say that the buyer chooses 100 units. In period 3, the buyer has 5% upward and downward flexibility and can purchase any quantity between [95,100].

This process repeats itself every period. The idea is to enable buyers to change their previously made commitments in a restricted way. When the time between making a new commitment (period i) and the actual delivery time (period j) is short, the flexibility is very limited. But when the time between making the commitment and the actual delivery is long, the flexibility is large. This is achieved by l_{j-i} and u_{j-i} increasing with $j - i$, and the number of updates that the buyer is able to make. The reason behind the structure of flexibility contract downward and upward constraints is that changes in the order quantity, with a very short notice, are costly for the supplier, while changes with a long notice

are much cheaper, and in this case changes in the previously made commitment can be relatively large.

All three studies aim to identify the optimal initial periodical commitments and the optimal periodical updates. Determining the optimal policy is very difficult, and all studies resort to approximation and heuristic. Basok *et al.* [7] develop a lower and upper bound on the total costs. The lower bound is achieved by assuming that the flexibility is very large—that is, assuming that changing previously made commitments is not limited. In this case, the structure of the optimal policy is well known: there are critical numbers (base stocks) $S_1, S_2, \dots, S_N$ so that at the beginning of each period it is optimal to raise the inventory up to the base stock. The idea behind the upper bound on the total cost is somewhat more involved. The upper bound policy is a feasible policy (changes in the commitment quantities are limited and satisfy the flexibility constraints). The changes are made so that the probability of reaching the base stocks $S_1, S_2, \dots, S_N$ is maximized. The authors show that the gap between the upper and lower bounds is not too large. For example, when the coefficient of variation of the demand is smaller than 0.5, the maximum gap is only 7% (assuming upper and lower flexibility of only 5%). Since the gap is

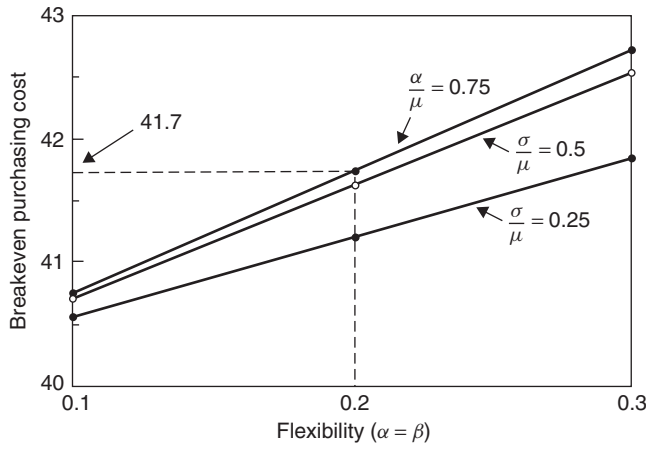


Figure 4. Breakeven purchasing cost versus flexibility under stationary demands ($c = 40, h = 1, p = 100, T = 12, \mu = 1000$), where c is the purchasing cost per unit, h is the holding cost per unit, p is the shortage cost per unit, μ is the mean demand, and σ is the standard deviation of the demand. [Source: Bassok *et al.* [7].]

small, the authors use the upper bound to calculate the optimal initial commitments and the updates of the periodical commitments.

The ability to determine close-to-optimal commitments and then to update them provides a very important tool in the negotiation process between buyers and suppliers. The suppliers may provide different levels of upward and downward flexibility, but for a higher flexibility the supplier demands a higher per-unit cost. This relationship is illustrated in the following graph (Fig. 4).

The reader can see that when the coefficient of variation is 0.75, the buyer should pay no more than 41.7 per unit for 0.2 upward and downward flexibility.

FIXED-DELIVERY CONTRACT

A contract that has similar features to the periodical orders and flexibility is presented by Moynzadeh and Nahmias [10]. These authors consider a fixed-delivery contract, and they assume a finite-horizon model with backorders and independent and stationary demand. The buyer is committed to purchasing the same quantity Q at a cost per unit of c in every period. The buyers have upward flexibility to order more than Q , but they have to pay a premium for the additional quantity, and the cost per unit is c_H (where $c_H > c$). In addition, every time the buyer deviates from the committed quantity Q , it must pay a fixed cost K . The fixed-delivery

contract captures the fact that in certain industries it is very costly to change the production quantity. Thus, suppliers prefer to produce the same quantity in every period. On the other hand, the buyer prefers to have flexibility to adapt to the uncertain demand and update the fixed-delivery quantity. The fixed-delivery contract offers limited flexibility because the buyer may purchase more than the fixed-delivery quantity but not less.

At first glance, it seems that the fixed-delivery contract is mostly advantageous to the supplier, who is able to reduce the variability in the orders. However, this observation is misleading. The benefit to the buyer depends upon the price discount that the supplier offers, c , and the premium price c_H . There is always a large-enough discount so that the buyer also benefits from the fixed discount contract.

BACKUP CONTRACTS

Eppen and Iyer [11] present a backup contract that has some of the flavors of the commitment-and-flexibility contract presented above. They consider a model that is very common in the apparel industry, especially between the manufacturer and a catalog (retailer). They suggest a two-period model. Before the first period, the buyer must make a purchasing decision, q , for the two periods. Yet, they take delivery of

only $Q(1 - \beta)$ units. Once they observe the demand, at the first period, they take a delivery of up to $Q\beta$ units, but not more than that. While the cost per unit is c , the cost per unit that the buyer does not take (at the most $Q\beta$ unit) is b . This contract is similar to the periodical-commitment-with-flexibility contract.

Notice that the buyer (catalog) makes a commitment to purchase Q units. The upward flexibility is zero and the downward flexibility is β . The main differences between the two contracts are (i) the periodical-commitment-with-flexibility contract is a multiperiod contract, while the backup contract is limited to only two periods; (ii) in the periodical-commitment-with-flexibility contract, the demands in the different periods are assumed to be independent, while the demands in the backup contracts are assumed to be correlated; and (iii) in the backup contract there is an explicit penalty per unit, b , for not purchasing units that were actually ordered. In the periodical-commitment-with-flexibility contract, the cost of flexibility is captured by the cost per unit of the product—the greater the flexibility, the higher is the cost per unit. But there is no explicit cost for units that were ordered but not purchased.

As the value of b (the penalty per unit for not purchasing the committed quantity) increases, the value of the backup contract decreases. This is quite intuitive because when b is very large, the buyer will tend to purchase all of the committed quantity and thus the backup option has zero value. Similarly, as l increases and the buyer has the opportunity to purchase a smaller fraction of the commitment, the total cost for the buyer increases.

Perhaps more interesting is the fact that the backup contracts provide the buyer with an incentive to commit to a larger quantity than the quantity that it would have purchased without the backup contract. This is a very important observation. The backup contract provides the buyer with flexibility and thus reduces the buyer's risk and induces the buyer to commit for more units. The supplier, who assumes a larger share of the risk benefits by selling a larger quantity.

The magnitude in this difference can be large, and the authors provide an example in which the committed quantity in the backup contract is larger by 63% than the purchasing quantity without a buyback agreement.

It is interesting to observe that for certain parameters both the supplier and the buyer benefit from the backup contract. Thus, one can ask what the parameters of the backup contracts that minimize the sum of the cost of both the buyer and the supplier are, and when is it possible to coordinate the channel. We will address this issue shortly.

OPTIONS CONTRACTS

The main idea of the above contracts is that they provide the buyer with flexibility to adjust to changing market conditions and, at the same time, they provide the supplier with information that helps the supplier to create a stable environment and plan its production and reduce cost better. All of the above contracts can be viewed as special cases of the following option contract. In a two-period option contract, the buyer makes periodical orders Q_1 and Q_2 , at a price c per unit. These orders are firm orders and cannot be changed. In addition, the buyer purchases M options, at the option price c_o per unit. These options may be exercised at the second period at the exercise price c_e .

As is the case in all of the previous contracts, the buyer would like to have the flexibility to change its previously made orders, while the supplier would like to limit the flexibility and encourage the buyer to limit the changes to the previously made orders. Clearly, if $c_o + c_e$ is large enough, then the buyer will tend not to purchase any options and will have zero flexibility. On the other hand, if $c_o + c_e = c$, the buyer will not make any commitment for the second period and will only purchase options. Choosing the right levels of c_o and c_e makes it possible to strike a balance between the interests of the buyer and the supplier and perhaps even to reduce (increase) the cost (profits) of both. The reader may notice that the option contract is a generalization of the two-period flexibility contract and the backup contracts.

Table 2. Special Cases

General model	Q_1	Q_2	M	c_1	c_2	c_o	c_e
Backup agreement	$Q(1 - \beta)$	0	$Q\beta$	c		b	$C - b$
QF contract	$\tilde{Q}_1$	$(1 - \alpha_d)\tilde{Q}_2$	$(\alpha_d + \alpha_u)\tilde{Q}_2$	c	c	0	c

[Source: Barnes-Schuster *et al.* [12].]

The table below (Table 2) demonstrates that these contracts are special cases of the option contract.

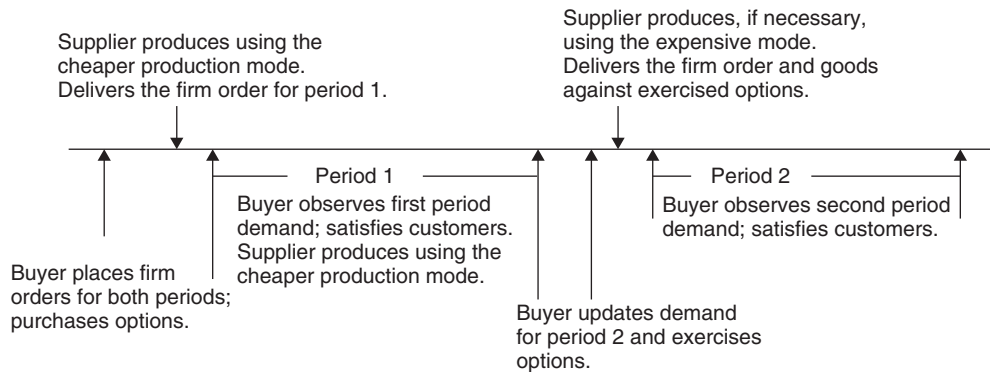
It is clear, for example, that the backup contract is a special case of the option contract. In both contracts, the buyer commits to purchase a certain quantity. The actions in the second period are different. In the backup contract, the buyer does not commit to any quantity, while in the options contract the buyer may commit to any quantity, Q_2 . Second, the backup contract limits the number of options that the buyer may purchase; it is a function of the buyer's first-period commitment. On the other hand, the option contract provides the buyer with the flexibility to purchase any number of options.

Barnes-Schuster *et al.* [12] provide a detailed two-period option model. In addition to studying the behavior of the buyer, they also study in detail the behavior of the supplier and the entire channel. The sequence of action assumed in this study is provided below (Fig. 5):

This model assumes that the supplier may produce ahead of time up to $Q_1 + Q_2 + M$ units. By producing ahead of time with a long production lead time, the buyer utilizes a less costly technology, but the supplier incurs

holding cost, and perhaps more importantly it assumes the risk of having, at the end of the last period, obsolete inventory. Alternatively, the supplier can produce a smaller quantity, wait until the buyer decides how many options to exercise, and only then produce the goods. In this case, the production cost per unit is high due to the short production lead time. Clearly, such a policy seems to be costly due to the high production cost. Yet, it reduces the inventory holding cost and the risk of obsolete goods that were produced but cannot be sold.

Such contracts are common practice in the apparel industry. The difficulty in this industry is the long lead time in producing the raw material. The knitting and weaving, dying and printing operations have a long production lead time. The sewing itself can be done relatively quickly. Suppliers (manufacturers), in order to minimize risk, tend to keep low inventories of raw materials—just enough to produce the ordered quantity. This results in one production run that is performed long before the season start, which creates a situation in which it is retailers who are unable to place a second order during the season when they have better information

**Figure 5.** A timeline of buyer-supplier decisions.

about the actual demand. The option contract reduces the supplier's risk of keeping raw materials because the buyer shares this risk by paying the option price. By keeping raw materials in stock, the suppliers are able to produce quickly and, if necessary, to produce and deliver during the season. Thus, it provides the buyer with the needed flexibility.

This detailed model not only determines the optimal commitments, the number of options, and the options and exercise prices, but it is also instrumental in identifying the optimal production policy for the supplier—in other words, how much should the supplier produce ahead of time at a low production cost, and which fraction of the total should be produced at a later time only and at a higher production cost. The fact that it is possible to consider that the supplier production policy leads the authors to ask whether it is possible to coordinate the channel. We remind the reader that we say that the channel is coordinated when the sum of the costs (profits) of the supplier and the buyer are the same as the costs (profits) in a centralized system with a single decision maker who optimizes the performance of the entire supply chain. In addition, if it is possible to coordinate the channel, it will be interesting to find out whether it is possible to allocate the cost (profit) between the buyer and the supplier in any arbitrary way. If the answer to both questions is positive, then it is possible to find options and exercise prices so that both players improve their costs (profits) as compared with other contracts. A detailed discussion of channel coordination is found in the article titled *Supply Chain Coordination* in this encyclopedia.

It turns out that in general it is impossible to coordinate the channel with linear prices. There are some instances (depending upon the cost parameters) that enable the coordination of the channel. But in these cases the prices are not individually rational for the supplier. Even return policies that are known to coordinate the channel in many environments fail to coordinate the channel in this case. One of the main difficulties in coordinating the channel is the fact that production cost is not linear. To see this point,

consider a buyer that purchases M options. The supplier may produce ahead of time, at a low cost, some fraction of the M options, then wait until a later time to see the actual number of options that are exercised and, finally, if necessary, produce an additional quantity at a higher cost. Thus, the actual production cost is not linear. The option contract assumes that the supplier charges the buyer a linear price. As a result, the number of options bought and exercised, assuming an option contract, is not the same as in the centralized system. Thus, in general, options contracts do not coordinate the channel. More on option contracts can be found in Refs 13 and 14.

CONCLUSIONS

The main issue in designing supply contracts is to ensure supplies and goods at the right quantity, quality, time, and cost. The tension is always between the wishes of buyers to have as much flexibility as possible and suppliers, who benefit from a stable environment and early firm commitments. Balancing this with flexibility can be done in different ways, but in all the cases stability and flexibility have real costs that can be evaluated, quantified, and used to calculate acceptable solutions for suppliers as well as buyers.

REFERENCES

1. Rosenblatt MJ, Lee HL. Improving profitability with quantity discount with fixed demand. *IIE Trans* 1985;17:388–395.
2. Munson C, Rosenblatt MJ. Theories and realities of quantity discounts: an exploratory study. *Prod Oper Manag* 1998;7(4):352–369.
3. Cachon G. Supply chain coordination with contracts. In: de Kok AG, Graves SC, editors. *Handbooks in operations research and management science*, 11: supply chain management: design, coordination and operations. North-Holland: 2003.
4. Blois KJ. Supply contracts in the Galbraithian planning system. *J Ind Econ* 1975;24(1):29–39.
5. Rietveld H. A new class of malaria drugs: the coartem breakthrough from Novartis and its Chinese partners. Workshop on access and benefit sharing. Bonn; 2008.

6. Bassok Y, Anupindi R. Analysis of supply contracts with total minimum commitment. *IEE Trans* 1997;29(5):373–381.
7. Bassok Y, Bixby A, Srinivasan R, *et al.* Design of component-supply contract with commitment-revision flexibility. *IBM J Res Dev* 1999;41(6):693–704.
8. Anupindi R, Bassok Y. Analysis of supply contracts with commitments and flexibility. *Nav Res Logist* 2008;55(2):459–477.
9. Tsay A, Lovejoy WS. Quantity flexibility contracts and supply chain performance. *Manuf Serv Oper Manag* 1999;1(2):89–111.
10. Moinzadeh K, Nahmias S. Adjustment strategies for fixed delivery contracts. *Manag Sci* 2000;48(3):408–423.
11. Eppen DG, Iyer AV. Backup agreement in fashion buying-the value of upstream flexibility. *Manag Sci* 1997;43(11):1469–1484.
12. Barnes-Schuster D, Bassok Y, Anupindi R. Coordination and flexibility in supply contracts with options. *Manuf Serv Oper Manag* 2002;4:171–207.
13. Donohue K. Efficient supply contracts for fashion goods with forecast updating and two production modes. *Manag Sci* 2000;46(11):1397–1411.
14. Spinler S. Capacity reservation for capital-intensive technologies: an options approach. Volume 525, *Lecture notes in economics and mathematical systems*. New York: Springer; 2003.

PRODUCT/SERVICE DESIGN COLLABORATION: MANAGING THE PRODUCT LIFE CYCLE¹

M. ERIC JOHNSON
Tuck School of Business,
Dartmouth College,
Hanover, New Hampshire

INTRODUCTION

Revenue growth in any industry requires a steady stream of innovative products and services. Without new market offerings, firms quickly find themselves left behind as customers defect to innovative competitors. This is particularly true for short-life products, like apparel or smart phones, where product life cycles have shrunk, driving the need for even more new products. Services are a little different—without innovation hotels become unattractive commodities and theme parks grow stale. Yet developing and bringing new offerings to market is becoming increasingly complex. Products with many components form challenging coupled problems, requiring substantial design iteration [4]. The driving forces of outsourcing and globalization have fragmented supply chains, with design activities now spread across firms and locations [1]. Product designers, marketers, and manufacturers are no longer in the same building or organization. More likely, they are spread over several continents in organizations with different cultures, languages, and business objectives. Once immune from such pressures, services are no longer local businesses and are also becoming more fragmented and global. The result is even greater challenges in managing the design process.

For example, brands like Hewlett-Packard, Dell, or Levi's used to do it all—operating their own US production plants along with their core design and marketing activities. Now, these firms have closed the production plants that once dotted the United States and outsourced much of that production and even design. Likewise, service help desks have moved to India and product recovery to China. While such outsourcing has had positive effects on product cost structures and asset management, it has dramatically increased supply chain complexity. Supply chains now span the globe with many hands touching a garment or electronic product before it reaches the consumer. Without solid processes and good communication, coordination failures can quickly snowball. When this happens, services fail to meet customer expectations or products arrive too late for the intended season, missing market share opportunities or requiring deep markdowns to liquidate the inventory before the next season arrives.

To be successful with such fragmented supply chains, firms must find ways to leverage their global partnerships through better design collaboration. Product design collaboration is a key element of product life cycle management [5,6]. Many think of product design as something that happens once, early in the development of a product, and then is over. However, design decisions [7] and the need to collaborate around the design continue throughout the life of a product. And the benefits of strong collaborative processes are not limited to short-life products. For long-life products, such as construction equipment or aircraft, products require ongoing maintenance and service enhancements to provide high lifetime value to end customers. Again, collaboration is critical for improving products, introducing new services, driving down cycle times, reducing both supply chain costs and the long-term cost of ownership for customers.

This article examines how firms in different businesses use collaborative approaches

¹This article substantially relies on the work of Johnson [1,2], and Kopczak and Johnson [3].

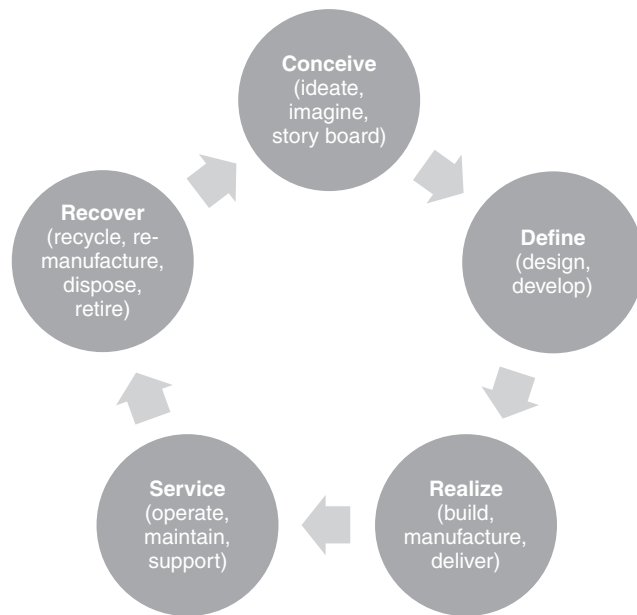


Figure 1. Stages of the product life cycle.

throughout the product life cycle. The life cycle of a product follows multiple stages (Fig. 1), including everything from the initial idea to the final recovery and recycling.

Collaboration around the definition of the product or service can happen at every stage of life. It is important to note that collaboration is not negotiation, nor is it simply learning from a partner. True product collaboration involves reliance on the unique expertise and experience of suppliers, partners, and customers, and delegation of a share of meaningful decision making. As companies implement collaborative software tools with their partners, the concurrent changes in technology, structure, work process, and organization depend on and impact trust within the supply chain. Thus, success in collaboration is as much about processes and technology as it is about organizational design and trust.

Often specific information technology tools are employed to enable collaboration at each stage. Sometimes those software products are focused on very specific problems facing particular industry issues (for example, storyboarding in apparel or engineering changes in computer hardware). Recently, many large software vendors (for example, Oracle and

SAP) have begun consolidating many of these point solutions into a suite of integrated tools that addresses the whole product life cycle. The focus of this article is the underlying challenges of life cycle management and the potential value of collaboration. To illustrate collaborative approaches, this article will present the key concepts of collaboration within the context of three case studies from different industries.

MANAGING PRODUCT CONTENT—THE ELECTRONICS INDUSTRY

Product collaboration in the electronics industry grew out of the need to better manage and synchronize product content information across the supply chain. Product content information is all of the data needed to manufacture a product to the correct specifications and at the most recent release. This includes such details as the bill of materials (BOM), drawings, lists of approved manufacturers for each component, and the process information needed by manufacturing to build and test a quality product. Typically, the BOM is a list of parts and subassemblies that make up the product, arranged in a

hierarchical format. For example, at the highest level is the final product. The next level of the BOM consists of major subassemblies followed by the smaller assemblies and components that make up each subassembly. Accuracy of this information is critical for procurement to buy the correct parts and manufacturing to assemble the right product.

As supply chains fragment into a string of outsourced services coordinated by virtual manufacturers, the need for content synchronization grows. In firms where marketing, R&D, procurement, manufacturing, and distribution are co-located on single site, synchronization (and collaboration) can occur without good information management. In the early idea stages of a product, designers can propose ideas over coffee with colleagues in marketing and manufacturing. During a new product introduction, manufacturing engineers can stroll down to R&D any time questions arise. When changes occur in the design or sourcing of material, development and procurement specialists can run downstairs to the shop floor to be sure the changes are understood and will not create production conflicts.

However, few electronic companies have such control over their products. More likely, designers are in Boston, assembly is outsourced to Southeast Asia, components are procured from an array of global firms, and distribution is handled by third-party providers who not only fulfill orders but also customize the products for unique customer needs (Fig. 2). For original equipment manufacturers (OEM), product licensees each have their own unique design needs, marketing

issues, and distribution requirements. Each partner in the chain must, at a minimum, have accurate, current product content information. Traditionally, critical product content information was often transferred in a hodgepodge of drawings, CAD files, spreadsheets, and text documents. Poorly managed product content information is a key source of supply chain headaches—particularly during new product introduction. One slip in the information relay race and products are assembled with the wrong parts or orders pile up waiting for components.

For example, consider a typical blunder made by an OEM who designed and marketed an electronic product. In a move to improve the functionality of the product, designers at the OEM made a routine change in one key specialized component that was supplied from their component distributor. The procurement managers at both the OEM and the manufacturing partner made the change and started ordering small quantities of the new component from the distributor. However, because the product content data was not integrated with the procurement systems, no one updated the forecast for the new component to reflect the ramping volumes that were planned. Since the distributor and manufacturing service provider did not realize the oversight, no plans were made with the component manufacturer to ensure the needed quantities would be in place. The component distributor had a small number of the components in inventory and quickly delivered initial orders, giving the manufacturing service provider the feeling that the part was readily available. Finally when a large order

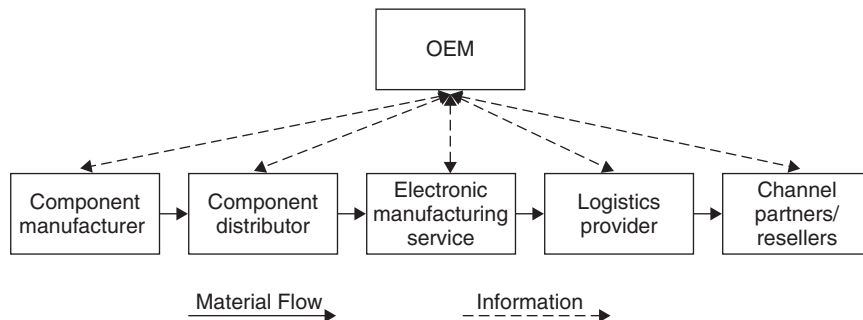


Figure 2. A typical virtual supply chain in the computer and electronics industry.

for the component was placed, the distributor realized the mistake, but it was too late. The component manufacturer could not deliver the component in time for the pending ramp up. For the OEM, this meant having to go to the channel partner and explain why the new product would have to be delayed for several weeks. The channel partner had already planned a major product rollout supported by advertising and had made many promises to its retail customers. After tense negotiations, the component manufacturer offered to switch production at one of its plants to expedite the chip, but at a cost of \$3 million. In the end, rather than delay the product introduction the three partners, the OEM, electronic manufacturing service (EMS) provider, and component distributor split the cost and expedited the chip.

While not all coordination mistakes are as costly as this one, the cumulative cost of many smaller mistakes is significant. Even managing simple tasks, such as the approval process for engineering changes, is a nightmare in many organizations. Proposed changes are often routed by phone, e-mails, text messages, and meetings to capture the necessary approvals to be released to manufacturing. Once released, changes often come as a surprise to manufacturing and sourcing partners, adding cost and time. On a routine basis, mistakes that create short delays in component deliveries to the EMS mean that reserved capacity at the EMS will go idle waiting for the part. In cases where the delay is clearly related to stumbles in the OEM product change process, the EMS may ask the OEM to help pay for the cost of the disruptions, typically \$5000–\$10,000 per day. Of course, these costs are small when compared to the substantial effect on total supply chain cost and potential revenue loss from delayed products.

To solve these problems, many firms—from Hewlett-Packard and General Motors—created their own homegrown design tools to help better coordinate designers and manufacturers. Many third-party companies, including CAD and workflow software vendors, also jumped into product data management market. Yet maintaining up-to-date product content is not enough.

The information must be shared with supply chain partners. A myriad of proprietary systems, while improving data quality, often hinders coordination between supply chain partners.

Today, a strong set of product collaboration software vendors such as Oracle and PTC offer web-based services that dramatically improve the ability of the supply chain members to communicate and collaborate with one another about new or changing product content. By bringing together all product content including drawings, BOM, approved vendors, process instructions, and a complete product history into a single portal, web-centric systems automate many painful tasks. For example, using such a system, a designer making an engineering change can work through a web browser to create the proposal. For each change, comments and product history is included so anyone examining the product can see why changes were made. When the proposal is complete, the software simply generates an e-mail containing a hyperlink to the proposed change to all of the people on the product team. By clicking on the hyperlink (with appropriate security), those who need to approve the proposal can see the changes, including drawings and product history. If the approvers agree with the change, they simply click an acceptance button and their response and comments are recorded with the proposal. After that, anyone who looks at the proposal could see who has accepted it thus far and who has not yet registered an acceptance or rejection. When all the required team members have approved the change, the system will automatically notify everyone related to the product that the change has been approved and release it to manufacturing and procurement.

As we have seen, synchronizing product data management translates into many cost savings. However, a far more interesting impact of a web-centric approach is how it changes the process of content management. Traditionally, distribution of product content follows a push process with the design engineers making product design decisions and sending those decisions to others in the supply chain. In disconnected push

systems, changes are expensive and thus are discouraged. Without instant access to content information, others within the supply chain operate with old information. More importantly, communication about product content is slow, hindering feedback from supply chain partners.

For example, if a designer specified a new component or product change that made the manufacturing process more difficult or affected the end quality of the product, it may take days or weeks for the supply chain partners to discover the issue and provide appropriate feedback. Automating the change process reduces the time and cost of engineering changes, making it possible to turn out frequent design improvements. Allowing supply chain partners to see the most recent information and suggest changes enables true collaboration and concurrent engineering where all supply chain partners participate in the design process. For example, a firm selling products worldwide could get feedback from regional partners, ensuring that the product would conform to local tastes and regulations well before the design is finalized. This leads to better products with fewer defects, improved functionality, and lower cost. The result is increased revenue.

ENSURING FRESH PRODUCTS—THE APPAREL INDUSTRY

The apparel supply chain is equally complex. Even simple garments like t-shirts are often touched by hands in several countries before ending up in the target markets of Europe or the United States. A more complex product, like a winter parka, often contains components from all over the world: snaps from Germany, zippers from Japan, insulation from China and Thailand, and the outer shell from Taiwan. To manage this complexity, many brand owners work with facilitating agents like Li & Fung who add value by coordinating the far-flung supply chains.

Brands and agents that bring their products to market orchestrate a long supply chain starting with fibers (wool, cotton,

synthetic) that are spun, woven, knit, and dyed in large mills, then on to cut-and-sew assembly operations, and finally laundry and finishing facilities before joining a global distribution system. With components like zippers and snaps added along the way, the final garment is the joint effort of many companies like DuPont (fiber), Burlington (mill), Kellwood (cut and sew), Liz Claiborne (brand), Dillard's (retailer), along with numerous transportation providers, freight forwarders, export agents, and warehouse providers. This whole process typically takes 6–12 months. The actual cycle time through the system is far less. Including transportation times, the value added cycle time without problems or changes is more like four to eight weeks. But along the way, the complexity of coordinating the product definition and managing the associated miscommunications across multiple companies slows the process to a crawl. So brands routinely begin working on their product lines well over a year before the selling season.

One of the biggest challenges in the industry is simply defining the product. Each garment has many technical specifications. First, there is the fiber and fabric itself. Working with fiber producers and mills, brand designers first must define many attributes of the fabric: its make-up, texture, weight, associated strength and elastic qualities, color, and finish. Even for what seems like a simple product, like jeanswear, this process can take months. After developing the concept with the brand, the mill runs a pilot production run of the fabric and submits it for testing. After review with the brand, changes are made and another pilot is run. Because of the scale of a high volume mill, each pilot run means producing several thousand yards of material that ultimately gets scrapped or dumped as off-goods. Getting the fabric right can mean multiple trials—each requiring weeks (Fig. 3).

But developing a technical specification for the fabric is only the beginning. Artwork for prints that are transferred to the fabric surface come next. Then there is the specification of the garment itself—the patterns, cutting and sewing instructions,

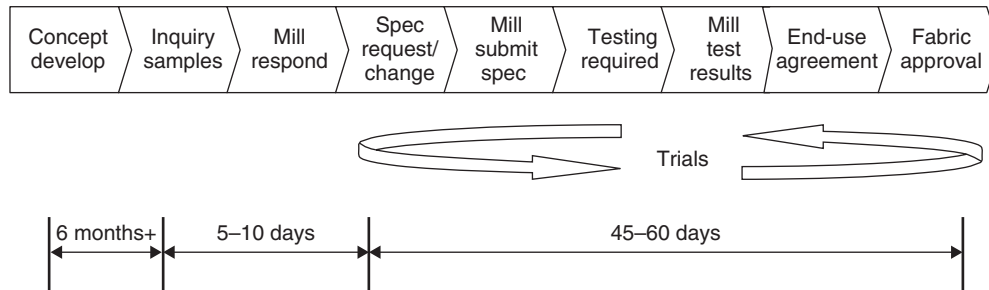


Figure 3. The fabric development process.

sizing specifications, and preferred vendors for buttons, snaps, zippers, and other components. Together with the BOM, associated drawings, process instructions, testing specifications, and technical color definitions, the total information content defining each product is massive. Even final garment finishing can be complex. For jeanswear products, the science behind laundering the garments to achieve the desired look is a key area of competition among brands and a high value-adding step to the garment.

As with many old industries, the information management supporting all the different facets of a garment specification has been slow to keep up with the increased complexity of the supply chain. Ensuring that everyone in the supply chain has an accurate and up-to-date description of the product is one of the biggest challenges. Like all products, once designed, garments are subject to many design changes in the preproduction phases. Typical products see more than 50 changes or enhancements before production is complete. And often changes made by brand designers are slow to reach the production floor of a contract manufacturer. In many companies, the change process is conducted through phone, e-mails, and test messages—all poor means of managing a distributed, complex supply chain. An executive at Burlington lamented that the change process looks like a child's whispering game. Starting at one end of the chain, a designer may call a manufacturing coordinator requesting a change to the left rear pocket of the pants. But, by the time the request makes its way to the manufacturing floor in Mexico, the request becomes a change to the right front pocket.

Getting the right information to the right people at the right time is the biggest challenge. Equally important is visibility to the entire product and sourcing team with a documented history of product changes. All too often, a change made by one member of the design team would be unseen by others creating confusion and finger pointing. Off-spec products arriving at a brand distribution center would be turned back by inspectors only to find out later that a single manager in the chain verbally approved the changes.

At Liz Claiborne, the first step toward information integration was bringing designers on-line using a consistent set of tools. Designers, long focused on hand sketching, were hesitant to move to computer-aided design. The organizational change of bringing hundreds of users on-line with digital design tools was painful, but after a five-year effort they have cut design time by 50%. Yet firms like Liz Claiborne found out that automating the design process, while improving internal efficiencies, did not help solve the supply chain problems. Vendors, who were less technically sophisticated or used different proprietary design systems, could not profit from the digital product designs. Often artwork created in a CAD system by a brand would be printed and shipped by overnight mail to a manufacturing partner where the document would be scanned (redigitized) to drive the manufacturing system that actually created the material. Even for those who could use the digital artwork, moving very large files over the web required more than simply attaching the CAD file to an e-mail message. It required a content management system that provided centralized

product information, where changes could be tracked and everyone was sure to have the most recent information.

At Dillard's, information transfer between in-house private-label designers and their 50 subcontracting partners once depended on massive faxing and overnight mail shipments. At a subcontractor's in India or China, the collection of documents would arrive over a period of days and weeks—completing the specification of the products. However, if there was a change in the design and assembly process, the product would be side-lined until the issues could be resolved and the required people in each organization had signed-off on the changes. When crunch time came, changes often happened over the phone or via e-mail with others in the design team left out of the loop, causing confusion and mistakes.

Tormented by the rising cost of complexity, many apparel designers began adopting digital design systems over 10 years ago, but like Liz, found that transferring the information within the rapidly disintegrating supply chain was tedious. Large firms like Gap and Limited invested in their own systems only to be paralyzed by the integration issues. With the opportunities of moving design processes onto the web, third-party apparel product management software companies like Lawson and Gerber Technology began developing web-based solutions. Levi's was one of the first large apparel companies to begin implementing a collaborative product management system from Lawson (then Freeborders) to facilitate fabric development, linking textile mills to Levi's product development, sourcing, pattern-making, and quality. For Dillard's, even the simplest features of these tools dramatically changed their product content management. Using a publishing and content management tool, Dillard's eliminated the faxing and e-mailing of product documents to their vendors. Each style is now managed with a virtual folder where designers and manufacturing partners can access documents over the web. With important product content stored in a single place, everyone is assured to be working from the most recent version. More importantly, changes cannot occur without being visible to

all parties. E-mail is limited to reminders to check changes in the folder and no changes can be made outside of the system. Besides saving weeks in communication time, the system helped eliminate costly mistakes and confusion.

Liz Claiborne implemented a similar system, cutting weeks out of the cycle time. At Liz, the time saved using a web-based system ensured more on-time deliveries and helped avoid costly markdowns due to late shipments. In some cases, with extra time in the design cycle, products could be further improved or shipments could leverage less-cost ocean transport rather than air freight. The advantages of even a simple web-based system are enormous and include the following:

- reduced cycle time;
- faster time to volume production;
- reduced scrap costs and mistakes;
- fewer markdowns due to late arrivals;
- secure environment with less paper and e-mailed documents;
- more reuse of standard design elements saving time in the design process;
- archived product history and change information leading to better accountability.

For contractors and mills, getting the product right the first time reduces scrap and costly delays that drive down manufacturing utilization. But escaping the supply chain tensions and the endless blame game may be one of the biggest benefits.

Collaborative systems can also be used by upstream suppliers, allowing vendors to share everything from product assembly specifications to more complex specifications of the fabric itself. For example, with its extensive experience with fibers, DuPont is able to work with textile mills to help them better use Lycra (a DuPont fiber) in manufacturing. As with many other industries, trust is important to successful collaboration. Decades of cost-focused procurement resulted in an environment of distrust that sometimes hinders firms from working together.

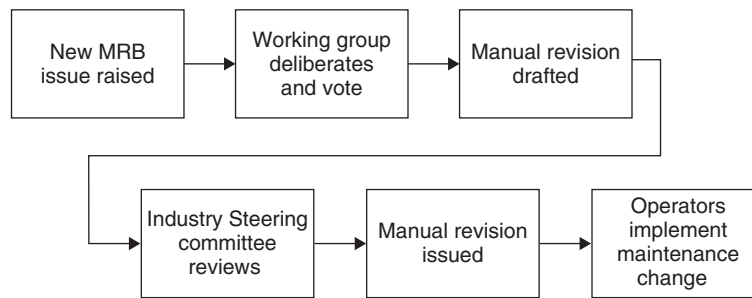


Figure 4. Maintenance review board (MRB) process.

COLLABORATION WITH CUSTOMERS—THE AEROSPACE INDUSTRY

In industries with long-life products, like construction equipment or aerospace, product collaboration has moved far beyond the initial product design to ongoing service innovation. Consider the case of BAE Systems Regional Aircraft, a UK-based company that provides postsales customer support for regional aircraft operators such as SWISS, Emerald Airlines, and Air Wisconsin. BAE Systems recently implemented collaborative software in its maintenance review board (MRB) process, the process by which BAE Systems and its customers (the operators), under the oversight of regulators, review and revise the maintenance review board report (MRBR) for each model of aircraft. The competitiveness of both BAE Systems and its customers has increased through this collaboration. And trust issues are being addressed along the way.

BAE Systems is responsible for customer support, engineering, and asset management activities covering a worldwide base of 160 customers operating over 1100 aircraft, in addition to managing a trading-and-leasing portfolio of some 350 aircraft.² BAE Systems currently supports the BAE Systems 146/Avro RJ, ATP, JS41, and JS 31/32 commercial aircraft models.

As part of its customer support activities, BAE Systems maintains the regulator-mandated MRBR for each model of plane. This manual is used by operators to create

their maintenance programs. While there has always been an opportunity to update maintenance procedures based on the experience of operators, relatively few changes have been made to these manuals. Updating has been limited by the labor-intensiveness and structure of the MRB process, as well as by the lack of collaboration with customers in this area. For example, as of 2004, while the BAe146/RJ model was 23 years old, the MRBR was on Revision 11. The fact that it was still a typewritten document also impeded updates.

The regulator-approved MRB process progresses as follows (Fig. 4): A proposed change (item) is introduced by an operator. The item is then discussed, analyzed, and voted on by the working group, a group including representatives from BAE Systems, major operators, and the regulating agencies (US Federal Aviation Administration (FAA), UK Civil Aviation Authority (CAA), and Transport Canada). Each operator in the working group gets one vote. If the item is approved by the majority of those casting votes, the revision to the MRBR is drafted and submitted to the Industry Steering Committee (ISC), a group including higher-level representatives from BAE Systems, the operators, and the regulators. If the manual revision is approved by the majority of ISC-member operators, then it is incorporated into the next revision of the MRBR. Finally, operators revise their maintenance programs to reflect the revised MRBR.

Traditionally, the intent was to issue a new revision of the MRBR once a year. BAE Systems accumulated and analyzed items submitted by the operators, and

²BAE Systems press release, January 2006.

then submitted them as approximately two 250-page packets for discussion at an annual three-day meeting of the Working Group. Based on that meeting, approved items were submitted as a group to the annual three-day meeting of the ISC. Thus, the number and depth of items that could be addressed was limited by the deliberation time per item and the meeting length. During the 1990s, an average of 20 changes per year were made to the BAe146/RJ, and an average of 10 items were carried over to the next meeting due to lack of time. (The 2006 MRBR for the BAe146/RJ included 1722 maintenance tasks).

Things started to change in mid-2004, when BAE Systems saw the opportunity to improve the competitiveness of its planes and reduce the operating costs of its customers by lengthening the maintenance intervals (time between scheduled maintenance) for its planes. This could only be done through close collaboration with the operators, as they had both the reliability data and the experience-knowledge that would be needed for making these changes without adversely impacting safety or availability of aircraft. Furthermore, it would be the operators that would cast the votes determining whether or not the changes would be adopted.

In mid-2004, BAE Systems established the maintenance optimization program (MOP), an 18-month effort to review "A" Check and "C" Check items for its BAe 146/Avro RJ and ATP models. The program involved analyzing all 1722 maintenance tasks for the BAe 146/RJ model and 555 maintenance tasks for the ATP model.

BAE Systems proposed to the operators and the regulating authorities that they implement a collaborative software tool called ForumPass developed by Exostar LLC. ForumPass is a secure on-demand solution for business collaboration built on top of Parametric Technology Corporation's Windchill collaboration engine ProjectLink. The collaborative software tool allows the working group to discuss proposed maintenance changes on-line as a stream of individual items, rather than in an annual face-to-face meeting as a large group of

items. After doing a pilot project, the group approved the implementation.

Implementation of the ForumPass software was accompanied by changes to the working group structure, work process, and organization. Rather than being structured around an annual face-to-face meeting of a year's accumulation of items, the structure of the discussion interaction is now an ongoing on-line threaded discussion of a small group of rotating items. Work is done at the operators' (and the BAE Systems representatives') home sites, rather than at the BAE Systems site. The size of the working group expanded, from 8 voting members to 13 voting members, and from 3 additional nonvoting members to 6 additional nonvoting members. (Travel time and cost was no longer a factor limiting participation.) Several of these new members are operators with fewer, more recently purchased BAE Systems aircraft, so have less experience with those aircraft. A quorum is set once 70% of the members have responded on-line, and an item is approved if more than 50% of those responding vote "Accept."

Under the MOP program, 13 operators and BAE Systems discussed the 1722 maintenance tasks for the BAe 146/Avro RJ models over a period of six months. Lengthening of the intervals was approved for 70% of the tasks—roughly 1200 tasks. The AMM was reprinted six months after the commencement of the program, using ForumPass.

The MOP program has benefited both the operators and BAE Systems. Maintenance intervals have been lengthened by 25% or more on the BAe 146/Avro RJ and ATP models. The operators have lowered the cost of ownership for their planes, as well as the cost of collaboration with BAE Systems. Operators of the BAe 146/RJ fleet saw a reduction in direct maintenance cost ranging from \$9–31 per flying hour, depending on the model.

For many years, BAE Systems and other aircraft makers have engaged in customer collaboration, relying on their customers to initiate, deliberate (with BAE Systems), and vote to approve proposed changes to maintenance procedures, procedures that in the long run will determine the safety and reliability record of the aircraft. BAE Systems has taken customer collaboration a

step further by implementing collaborative software for its working group process. BAE Systems has benefited by having more competitive planes, which has boosted demand. In addition, BAE Systems can incorporate the knowledge it has gained through the MOP into the maintenance programs for those planes. Finally, BAE Systems has increased customer loyalty toward its service and spare parts business.

CONCLUSION

The focus of this article has been collaboration in product and service design. As we have seen, collaboration can occur anywhere in the supply chain, both with suppliers or customers. Of course there are many objectives of such collaboration, such as cost reduction (as in the electronics case), speed to market (as in the apparel case), or service (as in the aerospace case). Alternatively, the goal of collaboration may be to facilitate manufacture [8]; ensure robust, product platform architectures [9]; reduce the environmental impact of the product by enhancing recyclability [10] or its ability to be remanufactured [11]. Recovery, remanufacturing, recycling, and carbon reduction pose many opportunities and challenges [12] for designers and supply chain managers—for a discussion of closed-loop supply chains [13] see the article titled *Take-Back Legislation and Its Impact on Closed-Loop Supply Chains*. Likewise, besides collaboration on the product/service design, there are many opportunities for firms to collaborate on the management of the supply chain—for a discussion see the article titled *Supply Chain Coordination*. Given both the range of objectives of collaboration and the number of possible participants, significant research opportunities exist in better harmonizing the collaboration goals (e.g., carbon versus cost, speed versus service) as well as enabling and motivating partners to participate.

REFERENCES

1. Johnson ME. Harnessing the power of partnerships. *Financ Times* 2004;8:4–5. (Also printed under the title “Adding another link to the supply chain”.)
2. Johnson ME. Supply chain synchronizing through web-centric product content management. Volume 2, Achieving supply chain excellence through technology. Redwood City (CA): Montgomery Research, Inc.; 2000. <http://www.ascet.com>.
3. Kopczak LR, Johnson ME. Digital customer collaboration at BAE systems: trust in the supply chain. Center for Digital Strategies Technical Report. Dartmouth: Tuck School of Business; 2006.
4. Smith RP, Smith SD. Identifying controlling features of engineering design iteration *Manage Sci* 1997;43(1):276–293.
5. Stark J. Product lifecycle management: 21st century paradigm for product realisation (Hardcover). New York (NY): Springer; 2004. ISBN 1-85233-810-5.
6. Saaksvuori A. Product lifecycle management. 3rd ed. New York (NY): Springer; 2008. ISBN 3540781730.
7. Krishnan V, Ulrich KT. Product development decisions: a review of the literature *Manage Sci* 2001;47(1):1–21.
8. Boothroyd G, Dewhurst P, Knight WA. Product design for manufacture and assembly. New York: Marcel Dekker, Inc.; 2002.
9. Martin MV, Kosuke I. Design for variety: developing standardized and modularized product platform architectures. *Res Eng Des* 2002;13(4):213–235.
10. Ferrão P, Amaral J. Design for recycling in the automobile industry: new approaches and new tools. *J Eng Des* 2006;17(5):447–462.
11. Charter M, Gray C. Remanufacturing and product design. *Int J Prod Dev* 2008; 6(3–4):375–392.
12. Thierry M, Salomon M, Van Nunen J. Strategic issues in product recovery management. *Calif Manage Rev* 1995;32(2):114–135.
13. Savaskan RC, Bhattacharya S, van Wassenhove LN. Closed-loop supply chain models with product remanufacturing. *Manage Sci* 2004;50(2):239–252.

PROGRESSIVE ADAPTIVE USER SELECTION ENVIRONMENT (PAUSE) AUCTION PROCEDURE

RICHARD STEINBERG
London School of Economics,
London, UK

COMBINATORIAL AUCTIONS

In an auction involving multiple items, a bidder might have a higher valuation for a specified set of items than the sum of his valuations for those items individually. In such a case, the bidder is said to have a *synergy* for those items and those items are said to be *complementary* for that bidder. A *combinatorial auction* is an auction in which bidders can place bids on combinations of items, called *packages*, rather than just on individual items. Without the allowance for package bidding, a bidder would be exposed to a possible loss if his bids included synergistic gains that might not be achieved; this is called the *exposure problem*.

The package bidding of combinatorial auctions eliminates the exposure problem for the bidders, but gives rise to the *winner determination problem (WDP)* for the auctioneer. This is the problem of determining the winning bids by labeling bids as either winning or losing so as to maximize the sum of accepted bids, under the constraint that each item can be allocated to at most one bidder. In practice, this is a computationally intractable problem for any auction of significant size, since the WDP is equivalent to the weighted set packing problem and thus is NP-hard [1].

One might try to achieve *computational tractability*—that is, solvability in polynomial time—for the WDP by restricting the package bids in some way. However, when combinatorial auctions were first considered for use in the US Federal Communications Commission (FCC) spectrum auctions, McAfee [2] and other economists

argued that the only choice was either completely disallowing, or permitting all possible, combinatorial bids. This principle thus asserts that a combinatorial auction should allow all possible combinatorial bids, that is, it should be *fully combinatorial*.

There is also a mathematical reason for not restricting package bids. Although perhaps not their intent, Rothkopf *et al.* [3] showed that essentially any restriction that would reduce the WDP to a computationally tractable problem will, of necessity, result in a setting that is too simplistic for most real-world auctions. For example, they examine cardinality-based structures, and consider the restriction of allowing package bids to include only up to k items, for some k . How small must k be in order that such combinatorial auction structures will be computationally tractable? Answer: $k = 2$. In other words, restricting package bids to contain only two items will indeed yield a WDP that can be solved in polynomial time, but allowing packages of up to three items will already result in a WDP that is NP-hard.

Another approach would be to allow the auctioneer to approximate the solution to the WDP. However, there is the doubtful matter of whether bidders would find an approximate solution to be acceptable. Even if an approximation were to be acceptable, there is the question of whether a good approximation to the WDP is even possible. Hastad [4] showed that, unless any problem in NP can be solved in probabilistic polynomial time, then for any $\varepsilon > 0$, there is no polynomial-time algorithm that guarantees an approximate solution to the WDP within a factor of $n^{1-\varepsilon}$ from optimal, where n is the number of submitted bids.

Two additional issues arise in combinatorial auctions. The first is the desideratum that a losing bidder has the ability to see why he lost; a property known as *transparency*. In single item auctions, transparency is not an issue, but in combinatorial auctions, where the set of winning bids is by no means obvious, transparency may be difficult to achieve.

The second additional issue is known as the *threshold problem*. Due to the fact that bidders on smaller packages may not have the resources individually to overtake a large package bid, or might not have the ability to coordinate with each other in order to do so, the allowance of package bidding can favor bidders seeking larger packages. It would be desirable to have a combinatorial auction procedure that alleviates this bias.

There is in fact an auction that addresses all these issues. The PAUSE (progressive adaptive user selection environment) procedure introduced by Kelly and Steinberg [5] is a combinatorial auction that is transparent and allows all possible package bids, that is, is full-combinatorial, and where the auctioneer does not face the WDP but instead need only solve a polynomial-time problem. Further, PAUSE mitigates the threshold problem. (See the section titled “Additional Properties of PAUSE.”) In addition, the allocation obtained by PAUSE is *envy-free*, that is, at the conclusion of the auction, no bidder will desire to exchange his allocation with that of another bidder. In order to explain how PAUSE works, we will first discuss in the next section an earlier, noncombinatorial, auction procedure.

THE SIMULTANEOUS ASCENDING AUCTION

The *simultaneous ascending auction* (SAA) [6] for auctioning multiple items proceeds by discrete rounds, where multiple items are auctioned simultaneously, each with its own price. Each bidder is free to bid in a round on any number of items, subject to: (i) a *minimum bid* increment; (ii) an *activity rule* that restricts the pace of the bidding; and (iii) an allowance for *bid withdrawal*, whereby players can withdraw their bids subject to a payment equal to the difference between the withdrawn bid and the bid that replaces it. The auction terminates when a round completes in which no new bids have been submitted on any item. The winner on each item is the high bidder on that item, who is required to pay his bid.

The minimum bid increment assures that the auction concludes in a reasonable amount of time.

The activity rule requires bidders to maintain a minimum level of bidding activity during the auction where, as the auction progresses, the level of activity increases. The activity rule forces a bidder desiring a large quantity of items to bid for a relatively large quantity earlier in the auction, thus preventing against the “snake in the grass” strategy, in which a bidder deliberately maintains a low level of activity early in the auction and then greatly expands his demand late in the auction. The activity rule also promotes *price discovery*, the process in which information about the bidding is revealed to the bidders, who have the opportunity to adjust their subsequent bids accordingly.

The bid withdrawal rule facilitates, to some extent, the realization of bidder synergies, but it can be gamed. In Australia, in 1993, two licenses for satellite-television services were offered by an auction that allowed for bid withdrawal. The licenses were won by two bidders by offering “startling high bids” [7]. However, it soon became clear that these two bidders had no intention of paying their winning bids, having put in a series of lower bids, and then proceeded to default on their highest bids, and then default on their next highest bids, and so forth until they finally purchased the licenses at considerably less than their original high bids.

PAUSE

The key idea of Kelly and Steinberg [5] is that the inherent computational complexity of combinatorial auctions cannot be eliminated, and thus the computational burden of evaluating a package bid is transferred to the bidder submitting the bid, leaving the auctioneer to no longer face the WDP. Although in theory a bidder might face a computationally intractable problem, in practice the bidder may often have a relatively easy problem. The Kelly–Steinberg design is called *PAUSE*.

PAUSE proceeds in stages comprised of discrete rounds. In stage 1, a SAA is held for all the items, thus facilitating price discovery. In stages 2, 3, . . . bidders can realize their synergies via package bidding. However, the bids on packages cannot be submitted in isolation: each bidder is required to submit them

as part of a *composite bid*, which is a set of nonoverlapping package bids (including individual bids) that cover *all the items in the auction*. Of course, a bidder will generally be interested in bidding only on a subset of the items in the auction—and in any given round, perhaps only a subset of these. A *composite bid consists of the bidder's own bids, together with previously submitted bids by any of the bidders*.

In general, there are only two restrictions on composite bids. First, for a composite bid placed in stage k , the package bids that comprise it are limited in size, viz., each package bid cannot contain more than k items each. Second, composite bids are restricted in how they use bids from other bidders, viz., when previously submitted bids from a bidder Y are included in the composite bid, the bids from Y must have all been submitted by Y in a single round of the auction.

The following example illustrates composite bidding (see Fig. 1). There are six items in the auction: $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$ (Fig. 1a). Stage 1 ended with a bid of 5 on each item,

respectively, by bidders A, B, C, D, E, F, and consequently a revenue to the auctioneer from these six bids totaling 30 (Fig. 1b).

Stage 2 ended with a standing composite bid by bidder B consisting of the following component bids: 11 by bidder A on the package $\alpha\alpha'$, 20 by bidder B on package $\beta\beta'$, and 14 by bidder C on package $\gamma\gamma'$, with revenue to the auctioneer totaling 45 (Figure 1c).

At the beginning of stage 3, two composite bids are submitted from Bidders A and E. Bidder A's composite bid consists of a bid from himself of 35 on the package $\alpha\beta\gamma$, together with the earlier bids of 5 each on α', β' , and γ' , respectively, from bidders D, E, and F, with revenue to the auctioneer of 50 (Fig. 1d). Bidder E's composite bid (Fig. 1e) consists of a bid from himself of 30 on the package $\alpha\alpha'\beta'$, together with the earlier package bid of 14 on $\gamma\gamma'$ from bidder C, and the earlier bid of 5 on β from B, with revenue to the auctioneer of 49. Since both A and E have submitted valid composite bids, the auctioneer chooses bidder A's composite bid, as it yields higher revenue. The auction progresses from there,

α	β	γ
α'	β'	γ'

(a) The six items in the auction.

5 _A	5 _B	5 _C
5 _D	5 _E	5 _F

(b) End of Stage 1. Auctioneer revenue: 30.

11 _A	20 _B	14 _C
-----------------	-----------------	-----------------

(c) Bidder B's stage 2 composite bid. Auctioneer revenue: 45.

35 _A		
5 _D	5 _E	5 _F

(d) Bidder A's stage 3 composite bid. Auctioneer revenue: 50.

	5 _B	
30 _E		14 _C

(e) Bidder E's stage 3 composite bid. Auctioneer revenue: 49.

Figure 1. Illustration of composite bidding.

with bidders submitting composite bids of increasingly higher revenue until the auction terminates.

Composite bidding has three important consequences: (i) the auctioneer is relieved of the combinatorial burden of the WDP (since he needs only choose the highest valid composite bid); (ii) each losing bidder can compare his bids with the winning composite bid to see why he lost; and (iii) at the conclusion of the auction, no bidder would prefer to exchange his allocation with that of another bidder. These first two features are, of course, computational tractability and transparency. The third feature is *envy-freeness*.

Note that a package bid can be described as a special type of “contingent” bid in the sense that the bidder is offering to serve all the properties in the package or none. PAUSE handles more general forms of contingency as well through the allowance of *Boolean bids*, that is, strings of package bids together with Boolean “AND,” “OR,” and bracket connectors (bracket being left bracket or right bracket), together with a value or values.

THE STANDARD PROCEDURE VERSUS THE PAUSE PROCEDURE

The example in Figs 2, 3, and 4 compares the workings of the PAUSE procedure with that of the “standard” combinatorial auction procedure in which bidders submit bids to the

auctioneer who then faces the WDP. Consider an auction on four items as shown in Fig. 2. (You may think of this diagram as representing four adjacent parcels of land being offered for sale in a real estate auction.) There are three bidders such as ●, ○, and ●, who have interests as shown. Thus, bidder ● does not value any one of the items individually, but has positive valuations for the two top and the two bottom items; bidder ○ has positive valuations for each of the items individually, but has no synergies for any set of items; and, bidder ● has a synergy for three of the four items, as well as having a positive valuation for one of those three items on its own.

Figure 3 shows how the “standard” procedure operates. Bidders submit all their package bids, including bids on individual items, leaving the auctioneer to solve the WDP, an NP-hard problem. This may go through a number of rounds until a stopping rule terminates the auction and reveals the allocation to bidders. This solution is not transparent, as the step in which the auctioneer solves the problem is a “black box.”

Figure 4 shows how the PAUSE procedure operates. In stage 1 of the auction, the bidders submit their individual bids only. Thus, bidder ● submits no bids, ○ submits all four of his bids, and ● submits his bid for the single item only. The auctioneer has a simple problem of determining the winning set of individual bids via a SAA. In later stages

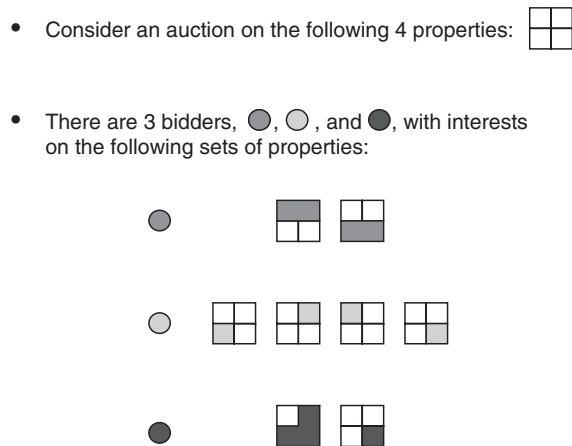


Figure 2. Four items and three bidders.

of the auction, each bidder can submit a single composite bid in each round and, as in stage 1, this process may go through a number of rounds until a stopping rule terminates the process. As it happens, bidder ● submits a composite bid including the top two items as a package from him, and fills out the bottom two items with previous bids from bidders ○ and ●. Bidder ○ has no synergies for any items, so acts passively by refraining from

submitting any composite bid. Bidder ● submits a composite bid consisting of the 3-item package bid from him, together with a bid from bidder ○ on the remaining item. The auctioneer needs only compare the composite bids submitted bidders ● and ●; specifically, by checking the validity of these composite bids and choosing the one with the highest revenue, an easy task. The solution is of course the same as obtained under the

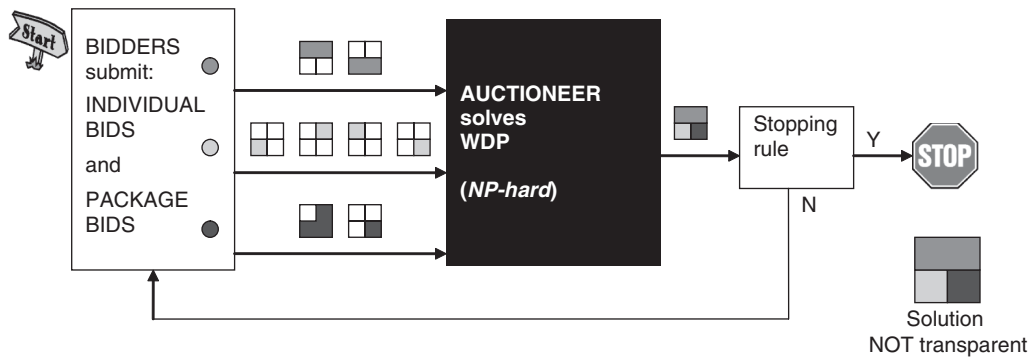


Figure 3. Standard procedure.

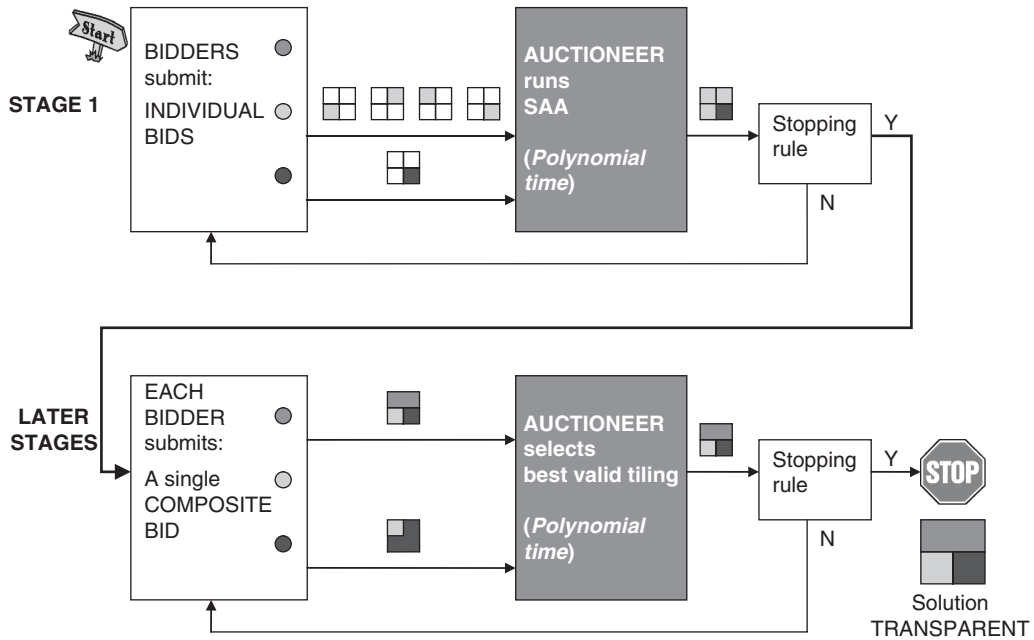


Figure 4. PAUSE procedure.

standard procedure, but here the procedure is transparent. Every bidder can see how the solution was obtained.

FORMAL DESCRIPTION OF PAUSE

Composite Bids

Let $\Omega = \{\alpha, \beta, \gamma, \dots\}$ denote the set of *items*, where $|\Omega| = m$, the number of items in the set Ω . Label packages $k \in \mathbf{K}(\Omega, A)$, where $\mathbf{K}(\Omega, A)$ is a set of subsets of Ω defined by a set of *attributes* A that are computationally tractable for the auctioneer to verify for each member of $\mathbf{K}(\Omega, A)$. Let

$$\mathbf{K}_q = \{k \in \mathbf{K}(\Omega, A) : 1 \leq |k| \leq q\},$$

where $|k|$ denotes the number of items in package k , that is, the *size* of package k . Thus $\mathbf{K}_q$ denotes the set of packages up to size q possessing attributes A . The set A would normally include the attribute, call it A_0 , that previously submitted bids from a bidder Y that are included in the composite bid must have all been submitted by Y in a single round. (Another attribute might be, for example, in the case where the items are parcels of land, that the items in each package bid must be geographically contiguous.) When we do not need to make the set A explicit, we can write $\mathbf{K}(\Omega, A)$ somewhat more concisely as $\mathbf{K}(\Omega)$.

A partition $\mathbf{P} = \{b_1, b_2, \dots, b_r\}$ is a collection of packages called *blocks* $b_1, b_2, \dots, b_r \in \mathbf{K}(\Omega)$ such that $\bigcup_{i=1}^r b_i = \Omega$, and $b_i \cap b_j = \emptyset$ for $i \neq j$. If a bidder X places a bid at price $p(b)$ on a block b of size k , then the triple $(X, b, p(b))$ is called a *block bid*.

A *composite bid* by bidder X is a collection of disjoint block bids that cover all the items of the auction. Note that a composite bid will, in general, contain block bids of other bidders in addition to one or more bids from bidder X . More formally, a composite bid comprises a partition

$$\mathbf{P} = \{b_1, b_2, \dots, b_r\},$$

together with an *evaluation*

$$(E(\mathbf{P}); p(b_1), p(b_2), \dots, p(b_r)),$$

where

$$E(\mathbf{P}) = \sum_{i=1}^r p(b_i).$$

Thus, a composite bid consists of $3r + 1$ pieces of information (in addition to the identification of the bidder placing the composite bid) capable of registration in a database. The first piece of information is the total value of the composite bid, $E(\mathbf{P})$. The remaining $3r$ pieces of information are, for i ($i = 1, 2, \dots, r$):

1. The specification of the block b_i
2. The bid $p(b_i)$ on the block b_i
3. The identity of the bidder for block b_i .

All $3r + 1$ pieces of information are available from the database to all bidders.

Specification of the Auction Mechanism

Stage 1: Bidding on Individual Items.

The Bidders. Each bidder submits bids on one or more items. In each round, there is an *improvement margin requirement*.

The new bid must improve on the previous best bid on that item by at least ε and less than 2ε , for a specified ε .

The Auctioneer. In each round, for each item the auctioneer confirms that a bid on that item is *valid* by verifying *increment validity*.

The bid satisfies the bounds of the improvement margin requirement.

In each round, the highest valid bid on each item is accepted by the auctioneer. The stage ends when bidding ends on all items. At the conclusion of stage 1, the auctioneer registers in a database the leading (i.e., high) bid on each item to their respective owners, as well as the stage in which their owner submitted the bid, and the stage and round in which the bid was submitted. This database is maintained by the auctioneer.

(Note: Stage 1 can be divided into sub-stages along the lines of the simultaneous ascending auction structure used by the US FCC, where progressively more severe activity rules are imposed over stages. See Ref. 8 for more details.)

Stages $k \geq 2$: Package Bidding.

The Bidders. Each bidder submits a single composite bid, $\mathbf{T}$ (which, by definition, covers all the items in the auction). In each round, there is an *improvement margin requirement*:

Let $w(\mathbf{T})$ denote the number of *new* bids in the composite bid. The new evaluation must improve on the previous best evaluation by *at least* $w(\mathbf{T})\varepsilon$ and *strictly less than* $2w(\mathbf{T})\varepsilon$. (Thus, the requirement is an average improvement per block of at least ε but less than 2ε .)

In stage k , each component b_i of a bidder's partition $\mathbf{P} = \{b_1, b_2, \dots, b_r\}$ is restricted to $b_i \in \mathbf{K}_k$, where $p(b_i)$ is either a new bid for block b_i , or a registered bid.

The Auctioneer. In each round, the auctioneer checks *composite bid validity* by checking:

1. *Bid Validity.* Each bid contained in the composite bid that is asserted to be registered in the data base is indeed so registered, and that new bids in the composite bid identify correctly the bidder for block b_i , and satisfy $b_i \in \mathbf{K}_k$.
2. *Evaluation Validity.* The value of the composite bid, $E(\mathbf{P})$, is indeed equal to $\sum_{i=1}^r p(b_i)$, the sum of the bids on each of its blocks.
3. *Increment Validity.* The value of the composite bid, $E(\mathbf{P})$, satisfies the bounds of the improvement margin requirement.

At the end of each round in stages $k \geq 2$, the new collection of valid bids on all the blocks from all the bidders is registered in the database to their respective owners, as well as the stage and round in which they submitted the bid. (Note that a composite bid might not be valid, but it may nevertheless, contain valid new component bids by the bidder submitting the composite bid. These valid new bids are to be registered in the database.) Further, the highest valid composite bid is accepted as the standing composite bid.

Thus, in each round, the auctioneer accepts *one* composite bid from among all the composite bids submitted by the bidders, but

updates the database with *all* valid bids on blocks from among *all* the bidders.

A stage ends when bidding ceases.

For suggested activity rules for PAUSE, see Ref. 5.

APPLICATION TO UNIVERSAL SERVICE SUPPORT AND MULTIPLE WINNERS

Kelly and Steinberg introduced the PAUSE mechanism for a specific application arising from the US *Telecommunications Act of 1996*, one of whose requirements was that regulators consider ways to reform the method of providing *universal service* subsidies for high cost areas. This refers to the situation that had been in place in the United States for many years, in which telephone companies were granted a monopoly to provide telephone service within their operating region, but had a concomitant responsibility to provide basic telephone service to all telephone subscribers within their region, no matter how costly. This was alleviated by a provision in which the telephone companies would receive subsidies for designated "high cost areas." Around the time of the passage of the Act, several parties advocated that "competitive bidding," that is, auctions, be used to determine universal service subsidies.

Kelly and Steinberg's paper was written in response to this suggestion. Since a firm might find it less costly to serve an area if it were to serve it together with other areas, a combinatorial auction was required. Kelly and Steinberg's universal service support auction was designed as a reverse auction using the PAUSE mechanism, where firms would bid for subsidies on specified areas and the winning firm would be the one that bid the lowest subsidy. Their design also allowed for competition within areas, that is, "multiple winners," a strong interest of the FCC at the time. The FCC's idea was that it could be useful to allow competition *in the market*, as well as competition *for the market*.

With the allowance for multiple winners, the auction design would need to minimize a form of collusion known as *price-level signaling*. This is the situation in which two or more players become aware that they are

actively interested in the same property on which they have the potential to share as multiple winners. Consequently, each firm bids noncompetitively so as to maximize its subsidy as a joint winner on the property. A discussion of how PAUSE minimizes price-level signaling appears in the section titled “Additional Properties of PAUSE.”

In order to accommodate multiple winners, the modifications to the basic PAUSE procedure are as follows. For the remainder of this section the reverse auction formulation will be employed, and the term for the areas will be “properties” rather than items. At the conclusion of stage 1, the auctioneer announces the number of multiple winners that will be accepted and necessary for each item χ , as determined by the auction’s *outcome rule*. One possible outcome rule (due to Milgrom [9]) has the following three-part formulation, where M_1 and M_2 ($0 < M_1 < M_2 < 100$) are specified in advance: (i) If at least one competing bid on χ is within M_1 percent of the lowest bid, then all firms that bid within M_1 percent of the lowest bid are designated as winners; (ii) if no competing bid on χ is within M_1 percent but one is within M_2 percent, then the two lowest bidders are winners; and (iii) if no bid on χ is within M_2 percent, then there is a single winner, viz., the lowest bidder. Denote the number of winners determined for property χ by $m(\chi)$. Before the start of stage 2, each item χ is replaced by $m(\chi)$ copies, $\chi_1, \chi_2, \chi_3, \dots, \chi_{m(\chi)}$, where each is allocated a nominal number of subscribers equal to $sub(\chi)/m(\chi)$, where $sub(\chi)$ denotes the number of subscribers in property χ .

The bidding rules from stage 2 onward are modified as follows. In order for a composite bid to be valid, there is the additional requirement that for each “multiple property” χ , the bid must not allocate χ_i and χ_j to the same bidder, for $i \neq j$. As before, the validity of a composite bid can be checked by the player constructing the composite bid, and this is to be checked by the auctioneer upon receiving the composite bid.

At the conclusion of the bidding, the $m(\chi)$ winning firms on property χ are each designated a $1/m(\chi)$ share of the responsibility on property χ . Specifically, the contractual

obligation carried by each winning firm is as follows:

1. The firm will receive its bid subsidy per subscriber on up to $1/m(\chi)$ of the total number of subscribers in that property;
2. The regulatory authority may require any firm with responsibility in that property who is not serving the full amount of its contractual share of subscribers to serve any unsubscribed subscriber in that property.

The particular subscribers that constitute a contractual share are not specified; the winning firm will compete with the other winning firms on that property for subscribers. Thus, there is an incentive for winners to actively seek to serve their share of subscribers, lest they be required to serve subscribers not of their choice.

A winning firm’s bid on property χ will in general be part of a composite bid. Thus, the limitation on the fraction of customers for which a firm will be subsidized prevents the firm from cross-subsidizing property χ from the other properties that comprise its bid. Each firm is free to compete for any or all the customer in property χ , although the firm will not receive subsidy for any subscribers beyond the share he has won responsibility for in the auction.

Kelly and Steinberg [5] point out that the method by which the numbers of multiple winners are determined is consistent with Milgrom [10] who showed that when there are no synergies, an optimal auction design entails endogenous market structure. In an appendix to their paper, Kelly and Steinberg describe a modification to PAUSE where the first stage is conducted as a sealed-bid auction.

More on the application to universal service can be found in Kelly and Steinberg [5], which had been submitted 3 years earlier *ex parte* to the FCC [11], the FCC’s related “Further Notice of Proposed Rule Making” [12, N.B. pp. 47–48 and 73–75], as well as Kelly and Steinberg’s comments in response to this document [13].

ADDITIONAL PROPERTIES OF PAUSE

Prevention Against Jump Bidding and Mitigation of the Threshold Problem

McAfee and McMillan [8] report a potential problem with SAAs, called *jump bidding*. This is where a bidder places a bid that significantly exceeds the required minimum required increment in order to signal an aggressive strategy to the other bidders not to compete with the jump bidder on the same item. This message discourages competition because it suggests that the jump bidder values the item more than anyone else. In Kelly and Steinberg [5], this is referred to more specifically as *price-jump bidding*. This is to distinguish it from another tactic they call *block-jump bidding*, in which a bid by a powerful bidder for a block of several items could be effective at preventing small bidders from piecing together a comparable composite bid.

In PAUSE, the improvement margin requirement reduces the possibility of price-jump bidding. The progression of allowable block size $k = 2, 3, \dots$, prevents block-jump bidding and mitigates the threshold problem. (Cramton [6] provides an example of how the absence of the upper bound on the improvement margin requirement can favor bidders seeking large aggregations.)

Minimization of Bidder Collusion

Klemperer [14] points out that a major concern for practical auction design involves the risk that participants may explicitly or tacitly collude to avoid bidding up prices. He specifically singles out the SAA where, early in the auction when prices are still low, bidders might be able to signal to each other who should win which items, and then tacitly agree to stop pushing up prices. Klemperer provides convincing evidence that this is what occurred in 1999 when Germany held a SAA for 10 blocks of spectrum. The auction closed after only two rounds, with Mannesmann and T-Mobile each ending up with half of the blocks at the same low price.

Under PAUSE, this problem of bidder collusion is less likely to occur, since the auction does not end at stage 1. In later stages of the auction, it is possible that another bidder

would place a package bid covering some of the items provisionally won by one (or more) of the colluding bidders. This could force one or more of the colluding bidders to raise their bids in order to stay in the auction.

In the version of PAUSE designed for universal service subsidy, which allows for multiple winners, a related type of bidder collusion is possible, called *price-level signaling* (see the section titled “Application to Universal Service Support and Multiple Winners”). However, in PAUSE price-level signaling is minimized by having the number of multiple winners determined prior to the combinatorial bidding stages that follow stage 1. Thus, most of the bidder surplus for the individual properties will have already been extracted. Any additional surplus mostly likely would be due to synergies, for which price-level signaling is very difficult to take advantage of, due to the fact that combinatorial bids overlap in ways that usually cannot be easily disaggregated.

FURTHER READING

Land *et al.* [15] focus on bidding behavior in PAUSE by considering in detail a general case of a two-item combinatorial auction. Mendoza and Vidal [16] develop two bidding algorithms for use in a PAUSE auction and analyze them via numerical methods.

Acknowledgments

The diagrams comparing PAUSE with the standard combinatorial auction procedure were devised together with Richard Gibbens. The author would like to thank Richard Gibbens, Frank Kelly, and Thomas Kittsteiner for helpful comments on earlier drafts of this article.

REFERENCES

1. Lehmann D, Müller R, Sandholm T. The winner determination problem. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006. pp. 297–317.
2. McAfee RP. Auction design for personal communications services: reply comments. PacTel

- Exhibit. Federal Communications Commission, PP Docket No. 93-253, Washington (DC). 1993.
3. Rothkopf MH, Pekec A, Harstad RM. Computationally manageable combinatorial auctions. *Manage Sci* 1998;44(8):1131–1147.
 4. Hastad J. Clique is hard to Approximate within $n^{1-\epsilon}$. *Acta Math* 1999;182(1):105–142.
 5. Kelly F, Steinberg R. A combinatorial auction with multiple winners for universal service. *Manage Sci* 2000;46(4):586–596.
 6. Cramton P. Simultaneous ascending auctions. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006. pp. 99–114.
 7. McMillan J. Selling spectrum rights. *J Econ Perspect* 1994;8(3):142–162.
 8. McAfee RP, McMillan J. Analyzing the airwaves auction. *J Econ Perspect* 1996;10(1):159–175.
 9. Milgrom PR. Statement attached to GTE's comments in response to questions. Federal Communications Commission. Notice of Proposed Rulemaking, CC Docket No. 96-45—Federal-State Joint Board on Universal Service. Washington (DC). 1996.
 10. Milgrom PR. Procuring universal service: putting auction theory to work. *Le Prix Nobel: The Nobel Prizes*. Stockholm: Nobel Foundation; 1996. pp. 382–392.
 11. Kelly F, Steinberg R. A combinatorial auction with multiple winners for COLR. University of Cambridge. 1997. Submitted ex parte 18 March 1997 by Citizens for a Sound Economy Foundation before the U.S. Federal Communications Commission re CC Docket No. 96-45—Federal-State Joint Board on Universal Service, Washington (DC).
 12. Federal Communications Commission. Further notice of proposed rule making, CC Docket No. 96-45, FCC 99-204. Released 1999. [See pp. 47–48, 73–75].
 13. Kelly F, Steinberg R. Comments on public notice DA 00-1075, In the matter of auction of licenses in the 747-762 and 777-792 MHz bands scheduled for September 6, 2000. Submitted before the U.S. Federal Communications Commission, 7 June 2000.
 14. Klemperer P. What really matters in auction design. *J Econ Perspect* 2002;16(1):169–189.
 15. Land A, Powell S, Steinberg R. PAUSE: a computationally tractable combinatorial auction. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Cambridge (MA): MIT Press; 2006. pp. 139–157.
 16. Mendoza B, Vidal JM. Bidding algorithms for a distributed combinatorial auction. In: *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*. Honolulu, Hawaii; 2007.

PROJECT-BASED ORMS EDUCATION

SUSAN E. MARTONOSI
Harvey Mudd College,
Claremont, California

There is much more to operations research and management science (ORMS) than rote application of formulas and algorithms. To effectively apply ORMS tools in the private and public sectors, students must also learn how to define an ill-posed problem; identify opportunities for cross-disciplinary methodology; choose an appropriate model; obtain, clean and analyze data; and decide when a model is useful. In addition, they must also develop the so-called “soft skills” such as professional and effective communication with stakeholders throughout the modeling and implementation process, project management, and effective team dynamics [1,2].

Ben-Zvi and Carton [3] and the references therein indicate that a traditional lecture-based course is insufficient for preparing students for managing ambiguities and decision making under uncertainty. Business and engineering school curricula focus on skill-development rather than the integration of multiple skills to solve a problem [3,4]. Grossman [5] and Reisman [6] encourage problem-centered, rather than method-centered, teaching with opportunities for hands-on work.

Projects, of all sorts and scopes, have been used in undergraduate and graduate ORMS curricula for decades to bridge the gap between the classroom, where tidy problems and clean data prevail, and the real world, where problems are complex and poorly defined, and data range from nonexistent to abundant but messy. In addition, project-based learning is prominent in other disciplines. The Accreditation Board for Engineering and Technology (ABET) requires accredited engineering programs to culminate in a major design project [7]. The Harvey Mudd College Clinic program, in which a team of undergraduates in their

final year completes a project posed by an external sponsor, spans their engineering, physics, computer science, and mathematics departments [8].

In this article, the author focuses on projects used in ORMS education. Common types of project-based educational activities that are described in the ORMS literature is outlined first: management games, case studies, class projects without external clients, curricular internships, and the so-called “field” projects with external clients. The author then focuses specifically on field projects with external clients and summarizes some of the best practices for instituting and maintaining such programs.

TYPES OF PROJECT-BASED EDUCATION

Management Games

In management games, teams of students are faced with fictitious but realistic business scenarios (and accompanying data) and must make a decision in real time. Often, management games are played repeatedly over a term, offering teams the opportunity to develop and refine their strategy, thus offering a project-like experience. The benefits of such games include

- learning the interrelation of functional specialties and applying cross-disciplinary knowledge;
- practicing decision making in a more dynamic environment than that afforded by a conventional case study;
- practicing decision making under uncertainty;
- low stakes environment;
- immediate feedback, typically via simulated financial outcomes, of the results of the team's decisions;
- illustrating organizational problems of team decision making;

- exploring game theoretic dynamics when multiple teams are acting as competing businesses within a same game; and
- improving one's decisions via repeated games, for instance, over the course of a semester [3,9,10].

Cohen and Rhenman [9] give an overview of the early use of simulation games in executive education. The first widely known business game was developed in 1956, based on military war games: American Management Association Top Management Business Decision Simulation. This game gave business executives the opportunity to gain experience in real-time decision making in a structured, fictitious environment. Players input their business decisions in the form of parameters, and the simulation output the results of those decisions in the form of business performance metrics. This game led to other business decision simulation games that were used in undergraduate, Masters, and PhD programs around the world.

More recently, Ben-Zvi and Carton [3] describe a business game course wherein students get practice in real-time decision making over the course of a semester through the use of the International Operations Simulation Mark/2000 game. Teams of students represent several companies that are operating in a simulated market environment. Each episode, the team makes a set of production and other operational and management decisions. All teams input their decisions into a database. The simulation runs, and the output is a set of financial reports for the companies. In addition to students observing the direct results of their decisions, they also begin to understand market dynamics as they witness the impact of their classmates' decisions on their company.

One downside of management games is the capital cost of any software [9]. However, free games are available on the internet, including the well-known MIT Beer Game [11]. This game consists of a retailer, a wholesaler, a distributor, and a factory who supply each other in a chain. In each round of the game, each party must place an order in the hopes of meeting its customers' demand. The

purpose of the game is to illustrate the value of information sharing across levels of a supply chain. Like other management games, it is dynamic and students work in groups to discuss their decisions [10].

Other recent examples include Refs 12–17, which are included in the Special Issue on the Use of Classroom Games in Management Science and Operations Research in *INFORMS Transactions on Education*.

Case Studies

A case study is a written scenario about a business situation in which a protagonist, typically a business executive, is faced with a managerial decision. Some cases are fictional, often anonymized versions of a true business situation, while many concern actual companies at critical points in their history. Cases are often authored by business school faculty who served as consultants to the company in question, advised a student consulting project, or maintain connections with their alumni in industry [18]. Cases are typically used in Masters of Business Administration (MBA) programs, and several business schools (e.g., Harvard Business School, the Ivey School of Business at the University of Western Ontario, and the University of Virginia Darden School of Business) are known for teaching almost exclusively using case studies.

Case studies bridge the gap between mastery of tools and techniques and their effective application to real-world problems [18]. Rarely do business problems conveniently pigeon-hole themselves within specific disciplines. Cases provide students with opportunities to apply interdisciplinary knowledge from across their education. Moreover, they demonstrate the role that quantitative methods can play within management [18]. In addition, there is rarely one correct decision to be made by the protagonist in the case, which allows students to use their creativity to consider many different options and interpretations. A good case has several possible directions and provides extensive information and data about the organization [19].

There are several models for using cases in the classroom. The simplest is an in-class

discussion of the issues of the case, either closely directed by the instructor to guide the students toward key points, or using “discussion leadership,” where students are completely responsible for directing the discussion toward key points [19]. In another model, students complete the quantitative analysis before the class and a student is selected to present the case to the class. In still another model, students can read and critique a completed case analysis [18,20]. In these examples, only one or two class sessions are spent discussing a single case. Some programs will look at a new case every class session [18], while others will intersperse cases among lectures and other classroom activities [21]. Although case studies are typically used in smaller classes, where the grading burden is less onerous and classroom discussion is more easily managed, Simpson [22] describes using cases in large classes with computer-aided evaluation of submitted assignments.

In a model that is closer to the “projects” theme of this article, students use a case study as the basis for a semester-long team project. In Ref. 23, the author describes a course at the Swinburne University in Australia in which students worked in teams of four to five students over 12 weeks to solve a case study. Each team appointed a project leader. At the end of the semester, teams gave oral presentations of their results.

In addition to providing the opportunity to synthesize multidisciplinary knowledge to make a business decision, cases also help students to develop soft skills such as communication (through in-class discussions and oral presentations), team work (in preparing the case solutions), and organization (in filtering through the provided data in the case to determine which pieces are the most critical) [20].

As stated in Ref. 20, cases can grow stale to instructors teaching the same course every year. The solution he discovered was to use articles from the journal *Interfaces* as a source of cases that student teams would try to improve on by developing their own models and solution approaches. In addition, there are several sources for published cases. Cases and accompanying

teaching materials are published in the journal *INFORMS (Institute for Operations Research and Management Science) Transactions on Education*. INFORMS also hosts an annual case competition, where authors of case studies can compete; participants are then encouraged to submit their cases to *INFORMS Transactions on Education*. Several business schools publish and sell cases authored by their faculty [24–27]. There also exist compilations of cases in book form, e.g. Refs 28,29.

Class Projects without External Clients

In addition to management games and case studies, a variety of projects can be incorporated into a course on ORMS to allow students to put theory to practice. The section titled “Student Field Projects” describes projects with external clients. In this section, projects that do not involve external clients are summarized.

These projects are typically one component (weighted less than one-third of the course grade) of a traditional undergraduate or graduate course. Some instructors use a collection of short projects, such as those described in Ref. 30. In these short projects, students observe and manage queues; follow news about service operations; analyze publicly available data, such as airline flight delay data; or map a service process and improve it, for example. Weal [23] describes a project in which groups of four or five students solve a real-world problem chosen by a member of academic staff.

In other courses, students are responsible for selecting their project topic, which can allow for great creativity. Belton *et al.* [31] describe a two-semester capstone course at the University of Strathclyde in the United Kingdom in which ORMS students in their final year design, as a class, a sequence of projects that they will complete. These projects range in structure from research papers, to game shows, to planning a start-up company. Anthony [32] and Martonosi [33] describe introductory courses at liberal arts colleges. Southwestern University and

Harvey Mudd College in which students identify their own projects of applying operations research (OR) methodology to a real-world problem. Projects have included course scheduling, locating electric vehicle charging stations, food pantry operations, and staffing of sports teams.

There are ample ways to incorporate the soft skills of OR into such projects. For example, Refs 32 and 33 report assigning intermediate deliverables such as a project proposal, a final report draft, and an executive summary, in addition to the final report; these deliverables undergo peer review by other students in the course to permit learning from each other. Students also evaluate each other's contributions both on the project teams and in the peer review groups to foster good teamwork. In Ref. 33, the quality of the written report counts for nearly half of the report grade, which encourages students to thoroughly proofread and edit their reports for clarity.

Grossman [5] explains that these course projects serve as a useful diagnostic tool for the instructor to recognize topics that have confused the majority of students. For instance, emphasis of models over modeling can result in students who do not know how to frame new problems. Instructors can use this feedback immediately to improve how they teach the material.

Curricular Internships

Some schools require students to complete internships in order to graduate to ensure that their graduates have had the opportunity to apply ORMS tools to practical problems. In an internship, the students work full time at an organization and have the opportunity to see firsthand how the methodology they have been learning in the classroom can be applied in a business setting. By working at the organization's site rather than completing a sponsored project on their own campus, they have greater access to the advice of professionals with domain expertise. The immersion experience also gives ample opportunity to develop professional communication, effective teamwork, and project management skills.

Although it is common for students to pursue internships during the vacation months of the academic year, this section addresses specifically the internships that are a required part of a university's degree curriculum. The extent to which these internships are arranged by the school for the students and monitored by faculty varies considerably in the literature.

In the shortest duration examples found by the author, students of the Cornell University's Semester in Manufacturing course spent three days participating in kaizen (rapid process improvement) activities at a company [34]. Bennet and MacFarlane [35] describe a program at the University of Strathclyde in the United Kingdom in which students spent three weeks working in a company's OR group as part of their graduation requirements.

Other universities require a full year internship for graduation. Weal [23] describes an OR program at the Swinburne Institute of Technology in Australia in which students were placed in jobs by a program coordinator who is also responsible for identifying participating companies. Moreover, representatives from the department visited the students three to four times during the year to monitor their progress and offer guidance. At the end, the students were required to write a report and deliver a poster session summarizing their experiences.

At the Eindhoven University of Technology in the Netherlands, a highly coordinated program was instituted involving six supply chains of three companies each. Students were assigned to work full time for one year at these companies. During that time, students implemented supply chain methodology from their courses to streamline interactions across each supply chain. The endeavor involved senior executives from each company and seven faculty advisors. The companies paid a large fee to the university to cover the cost of student stipends, faculty time, and travel [36].

Student Field Projects

In student field projects (also known as student consultancies), clients such as businesses, municipal offices, and nonprofit

organizations propose operational and managerial problems to teams of students. In contrast to an internship, field projects operate like a traditional consultancy: the student teams spend a period of time (typically one or two semesters) on the university campus solving the problem and present the final results to the client in the form of written and oral reports. Eaves [20] reports that the advantage of field projects over the examination of case studies is that field projects allow students to have a dialogue with the subject company that guides the project.

At many universities, field projects are used as the basis for “capstone” experiences, the culmination of a degree program. A capstone project permits students to synthesize material learned in earlier courses of their degree program, including courses fulfilling breadth requirements in other disciplines. At the undergraduate level, these include capstone programs in OR [37], operations management (OM) [38], and mathematics [39,40]. In other cases, the student field projects are conducted as one part of a more traditional course [41].

A wide range of project topics is reported in the literature, and the provided references can offer more detail on project scope. Abstracts of over 40 years of mathematics projects from the Harvey Mudd College Clinic Program, many related to ORMS, are found on the program’s web site [42]. Gorman [38] mentions projects involving machine scheduling, inventory management, process improvement, human resource planning and management, and statistical forecasting. Cooke and Williams [43] describe two software development projects. Armacost and Lowe [37] also list a wide array of topics including facility location; vehicle routing; production simulations; forecasting models for natural resource usage, catastrophic events and pension funds; and computer programming for database applications. They also summarize three successful projects. The first involved lawn maintenance crew routing at the Colorado Spring Cemeteries. Students used both a set-covering integer program and traveling salesperson heuristics to solve the problem as a multiple vehicle

routing problem, developed a prioritization scheme so that high priority regions of the cemetery lawn could be mowed once per week and lower priority regions less frequently, and identified resources necessary to achieve the cemetery’s original goal of mowing the full grounds every week. The second project pertained to pension portfolio management in which students solved quadratic programs to identify the efficient frontier of portfolio management strategies. They also developed a user interface sitting atop a simulation engine so that clients can select different strategies and assess their performance. In the third project, students created a custom database application to help the Colorado Air National Guard assess pilots’ accuracy on bombing runs. Although the authors report that this third project did not involve traditional OR methodology, the process of communicating with the client, clarifying the project’s objectives, and developing the tool to the client’s satisfaction gave the students valuable experience in the OR problem-solving process.

The educational benefits of student field projects include

- learning to frame ill-defined problems [37,43];
- learning to design the model with the end user in mind [5];
- improved communication skills [37,40,43];
- making connections between many disciplines within the field, or learning new tools to complete the projects [38,40];
- improved performance on course exams or in subsequent courses [5,44];
- developing project management skills [40];
- developing effective teamwork and leadership skills [40];
- understanding the ethical implications of professional work [39,40]; and
- learning basic research skills [40].

References in Ref. 43 point out that field projects emphasize frontline decision making over mere discussions of management in the hypothetical context of case studies.

Alumni also later report benefits from field project opportunities, such as improved job prospects, advancement within their organization, and the satisfaction of seeing their work implemented [38,45,46].

Student field projects have been formally assessed in several institutions. Gorman [47] reports on extensive assessment of the field project course at the University of Dayton, which included surveys of alumni zero to seven years postgraduation. Approximately 90% of respondents reported that the field project course was more valuable than other courses they had taken. In another study, 92% of students reported that the field project was an effective learning experience, and students who had completed a field project showed a statistically significant improvement in examination scores over students who had not [44]. At the University of Technology in Sydney, Quantitative Management Science students were asked to compare their capstone field project to other mathematics courses. Significant differences favoring field projects were found on the questions, “The subject was relevant to me,” “The subject was delivered in a way which was consistent with its stated objectives,” “My learning experiences were interesting and thought provoking,” and “Overall I am satisfied with the quality of this subject” [40].

In addition, institutional benefits in the form of revenue, research collaborations, publications, case studies, and community service are often reported in the literature [5,30,37,38,43,45,46]. Armacost and Lowe [37] summarizes two short teaching cases that resulted from student field projects at the US Air Force Academy. Grossman [5] provides references to several research papers and case studies that resulted from field projects.

Global Field Projects. While most student field projects work with companies that are local to the university, programs are emerging where students work with an international client, often in partnership with a team of students from an international university.

Kopczak and Fransoo [48] describe a “Global Project Coordination Course” between the Stanford University in the

United States and the Eindhoven University in the Netherlands. Three Stanford students partner with three Eindhoven students to work on a project having a corporate sponsor. For Eindhoven students, the course is a graduation requirement, whereas for Stanford students, it is an elective, and students submit résumés and cover letters in order to be selected. Through conference calls and periodic site visits to each other’s campuses and to the client’s site, the two teams of students work together over the course of one semester to solve the client’s real-world problem. In addition to offering the usual advantages of student field projects, these global field projects also teach the students important skills for effective cross-cultural collaboration. Stanford has partnered with several international universities and has run as many as 14 projects per year.

In a similar fashion, the Harvey Mudd College *Global Clinic* is a component of its college-wide Clinic capstone program. While the departmental Clinics have teams of three to five Harvey Mudd students working locally on a project sponsored by a client, in Global Clinic, a team of three Harvey Mudd students partners with a team of three students at an overseas university to work on a project sponsored by a multinational corporation or overseas nonprofit [49]. Harvey Mudd College has partnered with universities in Cambodia, Iceland, India, Japan, Puerto Rico, and Singapore; it runs one to three global field projects per year.

An important characteristic for the success of global field projects is that student abilities at the two partnering universities should be equally matched [48].

BEST PRACTICES IN STUDENT FIELD PROJECTS

Having established the many educational and institutional benefits of student field projects, it is now appropriate to consider the many details surrounding the implementation of such projects:

- Recruitment of projects;
- Fee structure;

- Confidentiality;
- Faculty role;
- Duration of projects;
- Team structure;
- Interactions between teams and clients;
- Deliverables;
- Grading;
- Institutionalizing these experiences;
- Challenges;

The literature presents many approaches to these details that have worked well in a wide variety of educational settings. This section attempts to summarize the approaches. In addition, Gorman [38] has a table describing the structure of several student field project courses.

Recruitment of Projects

Identifying good clients is repeatedly noted as critical to the success of a field project [38,39,48,50]. In addition to providing a project that is sufficiently challenging and interesting to the students, a good client is one who is invested in the outcome of a project and willing to communicate regularly with a team of students. Students can sense when the client has thrown together a problem statement that is of little importance, and they can lose motivation when this is the case [38]. In addition, the client must be willing to meet with students periodically, either face-to-face or via teleconference to ensure that the students are pursuing a direction that will be useful to the organization. Another desirable trait for some schools is the opportunity for future research collaborations and publications that could result from the project [46,48].

To find good clients, most field project programs rely heavily on their alumni base, local community organizations and businesses, industry affiliates of the university, personal and faculty contacts, companies that recruit through the school's career services office, and repeat customers because of the good reputation of their program [5,20,23,37–39,41,46,50]. Some programs, such as the Harvey Mudd College Clinic, recruit clients from across the nation and

rely heavily on teleconferences, e-mail, and one or two site visits for communication between the client and the student team [39]. The undergraduate capstone project at the University of Dayton uses an interview process to select clients, ultimately choosing ~60% of those organizations that have expressed an interest in participating. They prefer to maintain a mix of repeat clients and new clients, as well as a mix of industries and organization types [38].

Hand-in-hand with finding clients comes the problem of defining and scoping the project appropriately, which Bradley and Willett [34], Giaque and Sawaya [51], and others feel is the most critical determinant of project success. The first concern is ensuring that the techniques required to solve the problem are of the right level of sophistication for the students: too easy and the educational objectives of synthesizing material from several courses will not be met; too hard and the students might be unable to make progress. Armacost and Lowe [37] report needing to reject projects that are either too narrow (e.g. data gathering) or too broad. An additional concern is making sure the project scope will fit into the appropriate timeline [38,50]. Heriot *et al.* [52] recommend projects that have a tactical, rather than strategic, focus to reduce project scope. They also point out if the project scope is not defined in a way that is satisfactory to the client, then the client is more likely to back out. Borrelli [39] describes an iterative process in which the client presents several project ideas and works with the director of the field project program to select one idea and refine it suitably. However, despite one's best efforts, the project scope often changes as students start to understand the project better [38]. Importantly, the appropriate scope is highly dependent on the institution and its students; a project that can be completed in a semester at one institution might require a year or longer at another.

For field projects for large classes or that are requirements for graduation, the faculty member or program director in charge of recruiting clients has a duty to recruit a sufficient number of projects, which can be time demanding and stressful [52]. Eaves

[20] describes a personal rule of making five solicitations per day to prospective clients. Therefore, some field project programs require students to identify their own clients [5,30,43–45,52,53]. Business students, particularly part-time students, can often easily identify projects solving operational issues within their own employers; for these students, it is important that the selected project be of some benefit to them professionally. For undergraduates or full-time students, sourcing projects entails leveraging existing networks of contacts, cold-calling plant managers, and developing other useful entrepreneurial skills [45]. Once a client has been identified, students must also define the project in collaboration with their chosen client [45].

Student field projects offer several advantages to the client company:

- They are typically substantially cheaper than the cost of hiring a professional consultant (see the section titled “Fee Structure” and Refs 50 and 51), while providing access to talented students and *the faculty*.
- They are a good opportunity for a company to make progress on project ideas that have been sitting on a back burner without adequate staffing to pursue [20,43].
- They facilitate the client’s recruiting of students for jobs [5,20,39,41,50,51].
- Many programs have documented substantial savings for their clients: the University of Alabama Productivity Center estimates their 25-year contribution to be ~\$600 million in benefit to its clients [46]; the University of Dayton reports \$4 million in savings over eight years [38].
- They establish connections with academia [20].

Fee Structure

The ease of recruiting clients is somewhat dependent on the fee structure of the program. Many student field projects are free

to the client [37,38,45], and it is no surprise that these are the same programs that reported needing to turn away prospective clients. Gorman [38], Fraiman [41], Miller [46] report that these industry collaborations benefit both parties, and Armacost and Lowe [37] report that publishing papers and case studies based on the projects is an additional benefit. Haines [54] states that although the initial plan of the University of Toronto’s entrepreneurship program was to charge no fee to the clients (the program being funded by the Ontario government), the students in charge of managing the consultancy noticed that clients were less likely to be committed to a project for which they had paid nothing, suggesting that a nominal fee might be beneficial.

In some programs, clients pay direct expenses, such as software and travel [37,50,51]. In others, clients pay a fixed fee [39,48]. At Harvey Mudd College, the fixed fee goes to project expenses, administrative support, and institutional overhead [39]. Some programs are even more ambitious, securing state and federal funds in addition to client fees to support graduate students in addition to the items listed earlier [46]. Cooke and Williams [43] describe for-fee student field projects as an opportunity for businesses to outsource work at a low cost and for universities to supplement their income. They describe two programs, one at the University of North Texas where the clients pay a fee to the university, and a consultancy program at Clemson University where students are paid directly for their work, and a director bids on projects based on his or her contacts.

Confidentiality

Because any real-world project requires data, potential clients might be hesitant to participate in a student field project without requiring a confidentiality agreement. It is not uncommon for students, and occasionally faculty advisors, to sign such agreements [30,39]. In addition, care needs to be made in any public presentations that information is not being shared that could, for instance, preclude the client from filing a patent. This can usually be achieved by requesting the client

to vet any presentation materials in advance. The programs described in Refs 30 and 45 chose to have students conduct individual projects with their own employers in order to avoid the need for a confidentiality agreement. In Eindhoven's yearlong supply chain project described earlier, some company data was shared between companies within a single supply chain but never across different supply chains; clients restricted even some of the information communicated within the supply chain [36].

Faculty Role

There is wide variety documented in the available literature in the role of faculty members on student field projects, in terms of both technical contribution and time commitment.

Far to one end are student consultancies that are entirely student-run, such as that described in Ref. 54. Each term, the students filed as a legal partnership and eventually became an official nonprofit organization. Students identified clients and completed the work, occasionally, *hiring* faculty to act as consultants. Likewise, when students are responsible for selecting their own projects, the faculty advisor typically plays a backseat role [52]. On the other end of the spectrum, the faculty advisor or clinical professor acts as the project manager, and the students (undergraduates or graduates) are research assistants [43,46].

Most programs in the available literature describe a faculty role that is between these two extremes. The faculty advisor is primarily a coach who meets regularly with the team to monitor their progress and point them in the right direction but tries to allow them to proceed as autonomously as possible [5,20,23,37–39,48,53]. The faculty advisor is also responsible for quality control, by attending meetings with the client, reviewing reports, and helping the teams prepare for oral presentations [38,39,45,50,51]. Other important responsibilities include managing client expectations throughout the project [45], teaching supplemental topics that students need for the project [38,45], and tying up loose ends after project completion [51].

In addition, the number of teams a faculty member must advise can often dictate the amount of time that can be devoted to each team. The US Air Force Academy assigns a faculty instructor to the capstone course; this person must recruit 8–10 projects and oversee the program as part of that teaching responsibility. Faculty mentors then work more closely with a single team but do not receive teaching credit for doing so; they are compensated by the prospect of publishing the results of the project [37]. At the University of Dayton, the capstone course is taught by one or two faculty, each of whom manages roughly four teams and is responsible for recruiting clients. This involves ~20 h per team per faculty each year [38]. At Harvey Mudd College, a faculty director oversees the program and is responsible for recruiting projects, in exchange for teaching credit. In addition, each project's faculty advisor receives teaching credit for supervising a project. In the Stanford-Eindhoven program, one project entails about 250 h of faculty time because the faculty recruit the projects, manage the international logistics, as well as mentor the teams [48].

Duration of Projects

The majority of student field projects documented in the literature take place within a single semester, usually the final semester of the students' degree program [20,23,37,40,43,44,50]. The capstone project at the University of Dayton involves one credit hour in the fall semester for problem scoping, and five credit hours (or nearly two full courses) in the spring semester for completing the project. In between the fall and spring semesters, clients have ample time to gather needed data and prepare confidentiality agreements [38]. At Harvey Mudd College, the senior capstone spans both the semesters of the senior year and counts as a full course in each semester [39]. At the University of Alabama Productivity Center, projects range from short-term technical assistance projects to multiyear projects suitable for a PhD dissertation [46].

Team Structure

Most teams described in the literature have three to five student members [37–41,45,53], with [45] mentioning individual projects for MBA students and [5] describing teams as large as six to accommodate the large number of students in the bachelors degree program. Overall program sizes vary considerably. The Math Clinic program at Harvey Mudd College has 3–4 projects running annually [39], while the US Air Force Academy runs 8–10 projects each year [37], and the Indiana University South Bend had 70 OM field projects running during the 1998–1999 academic year, between a junior-level operations management course and an MBA-level operations management course [45]. The Alabama Productivity Center has funding for 25–30 students at a time and has participated in over 1300 projects over 25 years [46]. Most teams have a faculty mentor involved, and some also have professional engineering staff [46]. One of the team members (student, faculty, or staff) typically acts as a project manager. Teams can consist of students from a single major or can be interdisciplinary across several majors, depending on the needs of the project.

In some universities, students are assigned to teams by faculty, who sometimes take student preferences into consideration [39]. In other universities where field projects are not graduation requirements, students might have to interview in order to participate [48,50]. At the US Air Force Academy, clients compete for student interest by giving presentations on the first day of the course; students then vie for a client's attention to be the chosen team for the project. Teams are assigned to projects based on client and team preferences. When students choose their own project, there is a greater chance that the student will remain interested and engaged throughout [37]. On the other hand, assigning students to teams offers a greater growth opportunity for learning effective team dynamics, mimicking reality where one cannot always choose one's coworkers.

Another variable in the structure of these student field projects is the time commitment made by the students to the project during

the semester. Some programs require students to keep track of their billable hours [37,54]. At the Air Force Academy, each student spends ~4 h per week on the project [37]. At the University of Sydney, each student is expected to spend 7.5 h per week, but half of that time is spent in class [40]. Time commitments at the University of Dayton and Harvey Mudd College are larger, on the order of 8–10 h per week per student, amounting to ~500 person-hours per one-semester project involving teams of three at Dayton and 1200 person-hours per two-semester project involving teams of four at Harvey Mudd College [38,39].

Interactions between Teams and Client

Without regular interactions between the team and client, it is very easy for the students to pursue their own model without knowing whether or not it meets the client's objectives [39,46,50]. Therefore, a director of a student field projects program needs an assurance up front that the client will be committed to the project so that the students receive the guidance they need to solve the client's problem, not their own problem [50].

This guidance typically takes several forms. At a minimum, projects usually start with a kickoff event in which clients not just describe their desired outcomes of the project but also explain to the students the industry-specific context in which the project will be conducted and provide students the opportunity to brainstorm and ask questions [23,39,48]. Of the student field projects discussed in the literature, regular meetings with clients are also a common component, face-to-face if the client is local and via regular teleconference or occasional site visit if not [37–39,50]. These regular meetings serve to keep the client apprised of the progress being made by the team and to ensure that the project is not proceeding down an unfruitful path. In addition, most projects also require the client to respond promptly to questions and requests for data and information that can arise ad hoc. This communication typically takes place via e-mail. If the client is not committed to communicating regularly and responding to questions and requests in a timely fashion,

the project will be delayed [38]. Indeed, this can make or break a project, but clients do not always understand how critical this can be. The most successful projects are those with active and dedicated clients.

There is a delicate balance, however, between encouraging students to communicate regularly with the client and avoiding excessive demands on the client's time. One option is to appoint a team leader to manage all communications with the client. In addition, there are often library resources and the expertise of other faculty at the university that can be leveraged to provide the team with necessary background information without burdening the client.

Deliverables

Nearly all student field projects described in the literature require a written report at the end of the project that is given to the client. Several programs also require oral presentations before an audience of classmates, faculty, clients, and occasionally, outside reviewers [5,23,30,37,39,50,52]. In these oral and written reports, the students must communicate appropriately for the audience. Often, the clients are nontechnical or experts in different technical domains than ORMS [37,39]. Borrelli [39] describes Projects Day at Harvey Mudd College, where the entire campus and representatives from the ~40 clients watch a series of presentations and poster sessions given by the student teams. Other deliverables can include software and collected data that can help the client implement the project.

Gene Woolsey, a famous practitioner of OR, wrote "It is easier to do the math correctly (which requires no politics) than to get the method implemented (which always requires politics)," [55]. Implementing the results of an OR project can require several months or years of shepherding, as well as the full commitment of the sponsor to the outcomes of the project. Therefore, despite the potential educational benefits of involving students in this process, most of the student field project programs described in the literature omit the implementation phase. However, the MBA course described in Ref. 53 expects students to have a successful

implementation of the methodology within one year of the course. Thus, a deliverable is not just a report of recommendations; the students are also expected to help the client shepherd the project through implementation (this is feasible in an environment where the part-time students are still working full time in the organization in which the project is conducted). They report that many projects are implemented immediately, and only a few have little chance of ever being implemented.

Grading

In courses where a field project is only one component of the course grade, in addition to assignments and examinations, the literature suggests the project should count for 10–25% for undergraduate students and 15–30% for graduate students [30,44,45,52], with the difference attributed to the greater maturity and independence of graduate students.

Regardless of the weight of the project in the course grade, projects can be graded based on technical content, business content, the team's communication skills, project management, business outcomes, peer evaluations, as well as client evaluations [43,47]. Such evaluations can be used not only to assess the team's performance at the end of the project but also to provide formative feedback that can help the team during the project.

At Harvey Mudd College, each student submits a confidential self-evaluation and peer evaluation of each team member. The criteria used in this evaluation include numerical scores on cooperativeness, initiative, quality of technical work, quality of documentation, oral communication, reliability and follow-through on commitments, creation of useful results, creativity, promptness, and preparation. In addition, each student writes a narrative about the salient contributions, strengths, and weaknesses of each team member. These narratives typically convey useful behind-the-scenes information about the relative contributions of each student that might not already be apparent to the faculty advisor. This permits the faculty advisor to give a different grade

to each student and discourages students from “free-riding.”

Team presentations can also be evaluated by student peers, other faculty, and the clients to assess both the technical merits of the project and the team’s communication skills. Presentation rubric items can include clarity, effective use of technology, and appropriate technical content for intended audience. Additional free comments can be solicited to suggest improvements to both the presentation style and the technical approach to the project.

Because many student field project programs operate as consultancies (with students receiving academic credit, or being paid, or having the option to choose, as in Ref. 54), some programs have created policies for terminating unproductive members of a team. At the University of North Texas, where students receive academic credit for their work, a team has recourse to “fire” an unproductive student, provided they prepare sufficient documentation demonstrating the grounds for termination as would be required in a business environment [43]. In consultancies where students are paid rather than receiving academic credit, the precedent for terminating employment is more straightforward [43,54].

Institutionalizing These Experiences

The long-term success and continuity of these student field project programs can depend on individual personalities who spearhead the programs as well as institutional infrastructure [41], raising the question of what will happen when that person goes on sabbatical or leaves the university. Clearly, the program needs to be fully embraced by the department and the university administration. In a traditional publish-or-perish environment, Miller [46] indicates that a consultancy might not be taken seriously unless it is led by a respected faculty member with research credibility, rather than an administrator. At undergraduate colleges, such as the US Air Force Academy and Harvey Mudd College, there is greater hope for the longevity of a student-centered program. Indeed at Harvey Mudd College, nearly three-quarters of the student body participates in the field project

capstone program, across several majors [39]; its broad institutional support has no doubt played a major role in its success for over fifty years.

Miller [46] makes an important point that there is a risk when a program becomes institutionalized to the point of being required by its university to be self-funded. At that point, the program faces an urgent need to attract billable work, even if the pedagogical benefits of that work are insignificant. The program might need to turn down a challenging and interesting project from a nonprofit organization in favor of technical drudgery that is well funded by the client. Miller argues that if the school at least partially subsidizes the program then the program can focus on finding meaningful projects.

Challenges

One of the greatest obstacles to running a field project program is limited faculty time. Recruiting clients can take considerable effort. If the field projects are instituted as a supplement to an existing curriculum then there is less pressure to recruit projects than if the project is a graduation requirement [36]. The recruiting barrier becomes greater if the program needs to charge fees to cover costs, unless the program has a track record of success. However, as Wouters and van Donselaar [36] point out, with fees come the expectation of satisfying the client, to the possible detriment of the educational objectives of the program. In addition, considerable faculty time is needed to advise the teams in a meaningful way. Gorman [38] reports that feedback on the field experience from student interviews focused on improved selection and scoping of projects and more structured feedback and formal coaching for the teams, the very aspects that are the most time consuming for faculty. As described earlier, some schools count these programs toward a faculty member’s teaching load, but others do not. Failure to compensate faculty members for their time can result in a program structure that is convenient but perhaps not optimal for student learning. Lastly, advising such projects can bring faculty members out of their comfort zone, in that they must

relinquish some control over the outcome of a course or project [43].

In addition to those barriers, occasionally, projects are a disappointment to the student teams and the clients. Gorman [38] gives several reasons why a project might fail: the client's representative could lose her or his job or change positions; the students might procrastinate and try to complete the project at the last minute; the client might delay in giving the team necessary data; the project may turn out to be a "toy problem" that is not of any value to the client; the project might turn out to be significantly harder than anticipated; or the team might have difficulty in framing the problem.

Lastly, at many research universities, a tension can exist between conducting "applied research," such as typically arises from field projects, and more theoretical "basic research," that yields publications and grant funding. In one of the earlier discussions of ORMS education, Ackoff wrote of field projects: "In my opinion such involvement of students in research under faculty guidance is the most important instrument of education in the Management Sciences. Such activities do not preclude basic research; to the contrary, they stimulate it. No real problem worth solving can be solved without some basic research. Therefore the engagement of faculty and students on real problems yields basic research problems whose solutions are of practical significance," (Ref. 56, p. B4).

These potential challenges should not be treated as excuses for not pursuing field projects. As discussed earlier, there are myriad educational, academic, and institutional benefits to such projects.

CONCLUSIONS

Smith cautions: "It would be a mistake to sacrifice theoretical skills, which a university best teaches, for practical skills, which practice best teaches," (Ref. 57, p. 925). The project-based pedagogy described above should not be seen as a substitute for technical training in the mathematical methodology of ORMS. However, if practice

best teaches practical skills, as Smith suggests, then the above should provide inspiration for how to incorporate actual practice into ORMS education.

As described earlier, several institutions have begun assessing their project-based educational programs and comparing student outcomes to those achieved in more traditional courses. The early results are encouraging, suggesting that students who have completed a project-based course achieve better exam scores, better job placement and deeper satisfaction with the course than those who have not. However, more studies are needed to better assess the ability of project-based education to achieve specific learning objectives and long-term retention of ORMS skills.

FOR FURTHER READING

INFORMS Transactions on Education anticipates publishing in 2012 a special issue on Student Projects with Industry.

REFERENCES

1. Murphy FH. ASP, the art and science of practice: Elements of a theory of the practice of operations research: A framework. *Interfaces* 2005;35(2):154–163.
2. Murphy FH. ASP, the art and science of practice: Elements of a theory of the practice of operations research: Expertise in practice. *Interfaces* 2005;35(4):313–322.
3. Ben-Zvi T, Carton TC. From rhetoric to reality: Business games as educational tools. *INFORMS Trans Educ* 2007;8(1):10–18.
4. Noble JS. An approach for engineering curriculum integration in capstone design courses. *Int J Eng Educ* 1998;14(3):197–203.
5. Grossman TA. Student consulting projects benefit faculty and industry. *Interfaces* 2002;32(2):42–48.
6. Reisman A. OR/MS education: A need for course correction. *Int J Oper Quant Manage* 1995;1(1):67–76.
7. Accreditation Board for Engineering and Technology. Criteria for accrediting engineering programs, 2012-2013; 2012. [Online; Available at <http://www.abet.org/DisplayTemplates/DocsHandbook.aspx?id=1807>. Accessed 2012 Aug 19].

8. Harvey Mudd College, Clinic Program: An Education/Industry Collaboration 2012. [Online; Available at <http://www.hmc.edu/clinic>. Accessed 2012 Aug 19].
9. Cohen KJ, Rhenman E. The role of management games in education and research. *Manage Sci* 1961;7(2):131–166.
10. Griffin P. The use of classroom games in management science and operations research. *INFORMS Trans Educ* 2007;8(1):1–2.
11. Li M, Simchi-Levy D. The web based beer game: Demonstrating the value of integrated supply chain management; 2011. [Online; <http://beergame.mit.edu/guide.htm>. Accessed 2011 Dec 19].
12. Drake MJ, Mawhinney JR. Inventory control at Spiegel Grove. *INFORMS Trans Educ* 2007;8(1):19–24.
13. Hans EW, Nieberg T. Operating room manager game. *INFORMS Trans Educ* 2007;8(1):25–36.
14. Olsen T. Deming's quality experiments revisited. *INFORMS Trans Educ* 2007;8(1):37–40.
15. Villalobos JR. The stock portfolio game. *INFORMS Trans Educ* 2007;8(1):41–48.
16. Wolf JR, Portegys TE. Technology adoption in the presence of network externalities: A web-based classroom game. *INFORMS Trans Educ* 2007;8(1):49–54.
17. Wood SC. Online games to teach operations. *INFORMS Trans Educ* 2007;8(1):3–9.
18. Bodily SE. Teacher's forum: Teaching MBA quantitative business analysis using cases. *Interfaces* 1996;26(6):132–138.
19. Corner J, Corner PD. Teaching OR/MS using discussion leadership. *Interfaces* 2003;33(3):60–69.
20. Eaves BC. Learning the practice of operations research. *Interfaces* 1997;27(5):104–115.
21. Robinson S, Meadows M, Mingers J, O'Brien FA, Shale EA, Stray S. Teaching OR/MS to MBAs at Warwick Business School: A turnaround story. *Interfaces* 2003;33(2):67–76.
22. Simpson N. Educational automation: Creating case study challenges for large enrollment classes. *Decis Line* 2005;36(3):4–7.
23. Weal SE. Practical OR for undergraduates in the Swinburne course. *J Oper Res Soc* 1991;42(12):1047–1059.
24. Darden Business Publishing, Business Case Studies and Business Publications 2011. [Online; <https://store.darden.virginia.edu/>. Accessed 2011 Dec 19].
25. Harvard Business Publishing, Business Course Materials 2011. [Online; <http://hbsp.harvard.edu/>. Accessed 2011 Dec 19].
26. Richard Ivey School of Business, Ivey Publishing 2011. [Online; <https://www.iveycases.com/>. Accessed 2011 Dec 19].
27. Stanford Graduate School of Business, Cases 2011. [Online; <http://gsbapps.stanford.edu/cases/>. Accessed 2011 Dec 19].
28. Drucker PF. Management Cases, Revised Edition. Harper Business; 2008.
29. Carraway RL, Sherwood J, Frey C, Pfeifer PE, Bodily SE. Quantitative Business Analysis: Text and Cases. Richard D Irwin; 1997.
30. Behara RS, Davis MM. Active learning projects in service operations management. *INFORMS Trans Educ* 2010;11(1):20–28.
31. Belton V, Gould HT, Scott JL. Developing the reflective practitioner - Designing an undergraduate class. *Interfaces* 2006;36(2):150–164.
32. Anthony B. Operations research: Broadening computer science in a liberal arts college. Working paper; 2011.
33. Martonosi SE. The first (and often only) O.R. course: Something for everybody. INFORMS Annual Meeting, Austin, Texas, USA. 2010.
34. Bradley JR, Willett J. Cornell students participate in Lord Corporation's kaizen projects. *Interfaces* 2004;34(6):451–459.
35. Bennet P, MacFarlane J. Sampling the OR world: The Strathclyde 'apprenticeship' scheme. *J Oper Res Soc* 1992;43(10):933–944.
36. Wouters MJF, van Donselaar KH. Design of operations management internships across organizations - Learning OM by doing OM. *Interfaces* 2000;30(4):81–93.
37. Armacost AP, Lowe JK. Operations research capstone course: A project-based process of discovery and application. *INFORMS Trans Educ* 2003;3(2):1–25.
38. Gorman MF. The University of Dayton operations management capstone course: Undergraduate student field consulting applies theory to practice. *Interfaces* 2010;40(6):432–443.
39. Borrelli R. The doctor is in. *Not Am Math Soc* 2010;57(9):1127–1129.
40. Groen L. Meeting expectations - A focus on professional practice in a final year undergraduate mathematics course. Proceedings of the Assessment in Scientific

- Teaching and Learning Symposium; University of Sydney; 2007 28-29 Sep. [Online; <http://science.uniserve.edu.au/pubs/procs/2007/10.pdf>. Accessed 2011 Dec 16.
41. Fraiman NM. Building relationships between universities and businesses: The case at Columbia Business School'. *Interfaces* 2002;32(2):52–55.
 42. Harvey Mudd College, Mathematics Clinic: Projects 2012. [Online; Available at <http://www.math.hmc.edu/clinic/projects/>. Accessed 2012 Aug 19].
 43. Cooke L, Williams S. Two approaches to using client projects in the college classroom. *Bus Commun Q* 2004;67(2):139–152.
 44. Fish LA. Graduate student project: Employer operations management analysis. *J Educ Bus* 2008;84(1):18–30.
 45. Ahire SL. Linking operations management students directly to the real world. *Interfaces* 2001;31(5):104–120.
 46. Miller DM. A quarter of a century of academia-industry interfacing: The Alabama Productivity Center. *Interfaces* 2010;40(6):423–431.
 47. Gorman MF. Student reactions to the field consulting capstone course in operations management at the University of Dayton. *Interfaces* 2011;41(6):564–577.
 48. Kopczak LR, Fransoo JC. Teaching supply chain management through global projects with global project teams. *Prod Oper Manage* 2000;9(1):91–104.
 49. de Pillis LG. Harvey Mudd College Global Clinic; 2011. [Online; http://www.math.hmc.edu/depillis/GLOBAL_CLINIC/Global_Clinic_Welcome_Page.html. Accessed 2011 Dec 19.
 50. Giaque WC. Taking the classroom into reality: A field consulting experience for MBA's. *Interfaces* 1980;10(4):1–10.
 51. Giaque WC, Sawaya WJ. Taking the classroom into reality: Even better the second time around. *Interfaces* 1982;12(1):27–32.
 52. Heriot KC, Cook R, Jones RC, Simpson L. The use of student consulting projects as an active learning pedagogy: A case study in a production/operations management course. *Decis Sci J Innovat Educ* 2008;6(2):463–481.
 53. Liberatore MJ, Nydick RL. Breaking the mold: A new approach for teaching the first MBA course in management science. *Interfaces* 1999;29(4):99–116.
 54. Haines GH. The ombudsman: Teaching entrepreneurship. *Interfaces* 1988;18(5):23–30.
 55. Woolsey RED. Where were we, where are we, where are we going, and who cares? *Interfaces* 1993;23(5):40–46.
 56. Ackoff RL. Some ideas on education in the management sciences. *Manage Sci* 1970;17(2):B2–B4.
 57. Smith VL. The content of MSc OR courses. *J Oper Res Soc* 1991;42:924–926.

PROSPECT THEORY

ANDREW CHIU
GEORGE WU
University of Chicago Booth
School of Business,
Center for Decision
Research, Chicago, Illinois

Prospect theory is a descriptive theory of how people make choices that involve risk. The theory was developed by psychologists Kahneman and Tversky in 1979 [1]. In 2002, Kahneman was awarded the Nobel Prize in Economics, partially due to prospect theory's enormous influence on economics and other social sciences, and partially due to Kahneman and Tversky's research on heuristics and biases [2]. (The Nobel Prize cannot be awarded posthumously, and Tversky died in 1996.) Prospect theory challenged the traditional view in economics that decision makers are rational and, in particular, expected utility maximizers. However, prospect theory's reach has not been limited to economics—the theory has been influential in understanding a wide variety of real-world phenomena, in fields ranging from business, law, and medicine to political science and public policy.

Most real-world decisions must be made without full knowledge of what will transpire in the future. Consider, for example, an oil and gas company debating whether to drill for oil, an investor deciding whether to purchase Microsoft stock, or a family choosing whether to vacation in Chicago in May. These decisions would be straightforward if the decision makers could predict the future with certainty. But people are not clairvoyant. There may or may not be oil in the ground. Investing in Microsoft stock may increase your wealth, but you may also end up losing a sizable portion of your initial investment. Weather in Chicago in May could be glorious or unbearable.

The objective of prospect theory is to describe how people make decisions when there is uncertainty about the consequences of their choices. Decision theorists distinguish between *decision under risk*, situations in which the likelihood of events are known or objective (such as a spin of a roulette wheel), and *decision under uncertainty*, situations in which the decision maker must assess the probability of the uncertain events and hence the likelihood of events are subjective (such as the outcome of a sports game) [3]. Although prospect theory applies to both risk and uncertainty, we will focus here on risk for simplicity.

EXPECTED UTILITY THEORY

Before prospect theory, the dominant theory of how people choose among risky alternatives was expected utility theory. This theory was originally proposed by Bernoulli in 1738 [4] and was formalized by von Neumann and Morgenstern in 1947 [5]. von Neumann and Morgenstern proposed a set of axioms, or basic conditions, that are necessary and sufficient for expected utility. Expected utility theory is a standard building block of modern economic theory [6,7]. It is regarded to be rational to be an expected utility maximizer, as this theory is based on compelling axioms about how people *should* behave [8].

Expected utility theory posits that decision makers choose the prospect that maximizes their expected (or average) utility. More formally, consider a risky prospect, $P = (p_1, x_1; \dots; p_i, x_i; \dots; p_n, x_n)$, that offers outcome x_i with probability p_i , where $\sum p_i = 1$.

According to expected utility, the prospect P is evaluated as $U(P) = \sum_{i=1}^n p_i u(x_i)$, with decision makers choosing the prospect with the highest expected utility.

Under expected utility, risk preferences are completely captured by the shape of the utility function. Decision makers are risk averse if $U(x)$ is concave, and risk seeking

if $U(x)$ is convex, with most classical economic theory based on the tenet that decisions makers are risk averse. One standard interpretation for risk aversion is that the usefulness of an additional dollar decreases as a person gets wealthier, a principle known as *diminishing marginal utility* [9]. A utility function that exhibits diminishing marginal utility is concave, and hence the decision maker is risk averse.

PROSPECT THEORY

Kahneman and Tversky showed that decision makers are not rational as required by expected utility theory. The theory has been successful in making this point, largely because the theory consists of three parts that are individually attractive and work together well. First, Kahneman and Tversky produced a number of intuitive and elegant empirical demonstrations that were at odds with expected utility theory. Second, they proposed a mathematical model that could explain and organize these violations. Finally, they posited some psychological principles that suggest how people evaluate the alternatives they face. We begin by reviewing the primary empirical demonstrations that challenged the standard theory that decision makers are rational. We then describe the pieces of the mathematical model used to organize these violations. Throughout, we discuss how the psychological principles provide insight into these empirical demonstrations and review these principles at the end of this section.

Empirical Demonstrations

Fourfold Pattern of Risk Attitudes. Standard economic theory assumes that individuals are *risk averse*. Consider a lottery that offers a 50% chance to win \$1000 and a 50% chance to win \$0. The expected value of this lottery is $\$500 (0.5 \times \$1000 + 0.5 \times \$0)$. A risk-averse individual prefers winning the expected value of this lottery, \$500, for certain, to the lottery itself. Indeed, surveys show that the vast majority of people are risk averse for this choice.

But individuals are not universally risk averse. They dislike risk in some situations, while liking risk in others. For example, lotteries and casinos are billion-dollar industries in the United States. Many individuals purchase lottery tickets, even though the expected value of a state lottery ticket is often as little as half the price of the lottery ticket [10].

Kahneman and Tversky demonstrated that the purchase of lottery tickets is not an isolated example of risk-seeking behavior. They documented a pattern that has become known as the *fourfold pattern of risk attitudes* [11]. Individuals are risk averse for most gains, but risk seeking for most losses. As previously mentioned, most people would choose \$500 for sure over a lottery that offered them an equal chance at \$1000 or \$0. On the other hand, the same person typically prefers an even chance at losing \$0 or losing \$1000 to losing \$500 for sure. This pattern, however, switches when probabilities are small. In this case, decision makers are risk seeking for gains, and risk averse for losses. Consider a choice between winning \$1 for sure and a prospect that offered a 1% chance at winning \$100. Gambling in this case seems pretty reasonable. After all, \$1 is not much money, and the lottery offers some chance at winning a sizable amount of cash. The purchase of a lottery ticket is indeed an example of risk seeking for small probability gains. However, preferences reverse when we change “winning” to “losing” in the above example. Now, the typical preference is for losing \$1 for sure over taking a prospect in which there is a 1% chance of losing \$100. The attractiveness of insurance provides an example of risk aversion for small probability losses. In sum, the fourfold pattern of risk preferences is risk aversion for medium to large probability gains, risk seeking for small probability gains, risk seeking for medium to large probability losses, and risk aversion for small probability losses. This pattern is often called the *reflection effect*, because preferences flip or “reflect” when outcomes are changed from gains to losses.

Framing. Expected utility theory assumes that preferences between prospects do not depend on the manner in which they are described. For example, \$20 should be viewed the same whether you are given a single \$20 bill, two \$10 bills, or 20 singles. This compelling principle is known as the *invariance assumption*. However, prospect theory demonstrates that the same options can be framed in different ways to produce dramatically different preferences. In other words, our choices do not always obey the invariance assumption.

To illustrate, consider the following two-choice situations:

- In situation 1, you are first given \$1000. You must now choose between two options. If you choose *A*, you will receive an additional \$500 for sure. If you choose *B*, there is an equal chance that you will receive either an additional \$1000 or nothing.
- Now consider situation 2; this time you are first given \$2000. Again, there are two options. Perhaps you would like *C*, a sure loss of \$500? Or, maybe you would prefer *D*, a 50% chance at losing \$1000 and a 50% chance at losing \$0?

When confronted with situation 1, most people prefer *A*, the sure \$500, to the gamble *B*. On the other hand, when posed with a choice between *C* and *D*, most choose *D*, the uncertain loss, over the sure loss *C*. While both choices seem reasonable in isolation, situations 1 and 2 are identical in terms of final consequences, reducing to a choice between \$1500 for sure (*A* and *C*) and a lottery that offers an even chance at \$1000 or \$2000 (*B* and *D*).

The standard preferences described above violate the invariance assumption invoked by expected utility. However, they are consistent with the risk attitudes revealed by the fourfold pattern of risk attitudes. When the problem is framed as a gain (situation 1), people are usually risk averse. But when the problem is framed as a loss (situation 2), people tend to be risk seeking.

Aversion to Losses. Expected utility theory assumes that choices only reflect final outcomes. For example, if one were the beneficiary of a \$100 check, but also received a \$100 speeding ticket, these two events would offset one another in monetary terms. However, Kahneman and Tversky demonstrated that individuals tend to dislike losses much more than they like gains. Thus, the two events do not offset one another in psychological terms. The \$100 check makes you happy, to be sure. But, your pleasurable feelings for the check pale in comparison with the pain you feel when you are reminded about your \$100 speeding ticket. This dislike of losses is known as *loss aversion*. Put simply, losses loom larger than gains.

To illustrate loss aversion, consider a gamble in which you may lose \$1000 or gain \$1100 with equal chance. Although the expected value of this gamble is positive, most people find this prospect unappealing. The potential loss of \$1000 cannot be offset by the potential gain of \$1100. In order to make this gamble attractive, the gain has to be increased considerably, to about \$2000. In other words, roughly speaking, the pain of a loss is about twice as much as the pleasure from a comparably sized gain [12].

Common-Consequence Effect. Suppose that you are faced with the following choice. If you choose *A*, you receive \$200 with 10% chance. If you choose *B*, you receive \$100 with 20% chance. Next, imagine that we improve both options by adding to each a 10% chance of receiving \$1000. Now if you choose *A*, you receive \$200 with 10% chance, \$1000 with 10% chance, and \$0 with 80% chance. If you choose *B*, you receive \$100 with 20% chance, \$1000 with 10% chance, and \$0 with 70% chance. It seems reasonable that your choices in these two situations be identical, that is, unaffected by the additional chance at receiving \$1000. Since both of the second options have a 10% chance of receiving \$1000, this common consequence is irrelevant for this choice. Expected utility theory assumes this principle—adding a common consequence to two prospects should not change which alternative the decision maker

prefers. This principle is known as the *independence axiom* [13,14].

Kahneman and Tversky documented a violation of this principle. Consider the following two choices:

Situation 1

- option *A* offers a 33% chance at \$2500 and a 67% chance at \$0;
- option *B* offers a 34% chance at \$2400 and a 66% chance at \$0.

Situation 2

- option *C* offers a 33% chance at \$2500, a 66% chance at \$2400, and a 1% chance at \$0;
- option *D* offers \$2400 for sure.

Most people prefer *A* to *B*, arguing that 33% and 34% seem about the same, but \$2500 is clearly better than \$2400. On the other hand, people like *D* over *C*, reasoning that it is not worth taking a chance with *C* since *D* has a sure chance of winning a good amount of money. To see that this pattern of choices violates the independence axiom and hence is inconsistent with expected utility, note that situation 2 is created from situation 1 by adding a common consequence to both *A* and *B*: a 66% chance at \$2400.

This problem illustrates the strong preference for certainty over uncertainty. Indeed, this example has been termed the *certainty effect*. Adding a 66% chance at \$2400 to *B*, a 34% chance at \$2400, is particularly attractive, because it changes the prospect from possible to certain. On the other hand, for option *A*, the additional 66% chance at \$2400 increases the chance of winning from 33% to 99%, or from probable to more probable.

This demonstration is also called the *common-consequence effect*, because adding a common consequence to two options changes preferences, contrary to expected utility theory [15]. Indeed, a variant of this problem posited by the French economist Maurice Allais was one of the earliest challenges to the descriptive validity of expected utility [16]. Allais' challenge, which has become known as the *Allais Paradox*, consisted of two pairs of gambles. The first involved a choice between either (*A*) \$1 million for

sure or (*B*) a 0.10 chance at \$5 million, a 0.89 chance at \$1 million, and 0.01 chance at \$0, whereas the second was a choice between (*C*) a 0.11 chance at \$1 million (and a 0.89 chance at \$0) and (*D*) a 0.10 chance at \$5 million (and a 0.90 chance at \$0). Most people prefer *A* and *D*, a pattern that contradicts expected utility. Under expected utility theory, a choice of *A* over *B* implies that $0.11u(1) > 0.10u(5) + 0.01u(0)$, whereas a preference for *D* over *C* implies the opposite inequality, $0.10u(5) + 0.01u(0) > 0.11u(1)$.

The Prospect Theory Model

Prospect theory offered a rich set of empirical challenges to expected utility theory, including the fourfold pattern of risk attitudes, framing effects, aversion to losses, and the common-consequence effect. These demonstrations are robust, attractive, and intuitive. But do these demonstrations inform us about how individuals make decisions that involve risk in general? Kahneman and Tversky proposed a model for organizing the empirical demonstrations described above and making more accurate predictions about how individuals would behave in other choice situations.

Like expected utility theory, prospect theory assumes that individuals evaluate each alternative and then choose the alternative with the highest subjective valuation. However, prospect theory differs from expected utility theory in two substantive ways. Suppose a decision maker might choose an alternative that offers a p chance at outcome x . Expected utility theory assumes that the utility of this alternative is $pu(x)$. Prospect theory, like expected utility, assumes that the decision maker evaluates the x that he/she might get and the p chance at winning, and then combines these two evaluations together. However, the valuation of outcomes is governed by the *value function*, $v(x)$, and the valuation of the probabilities is governed by the *probability weighting function*, $\pi(p)$, and thus the value of this alternative is given by $\pi(p)v(x)$. [The valuation of more complicated alternatives with more than one nonzero outcome follows a revision of prospect theory, *cumulative prospect theory* [11].] These two pieces are described below.

Value Function. The prospect theory value function, $v(x)$, captures an individual's subjective perception of a particular outcome (Fig. 1). Prospect theory assumes that outcomes are coded as gains and losses, or outcomes above and below a reference point, and the value function captures how much better one gain is than another gain, how much worse one loss is than another loss, or alternatively how a particular loss compares with a gain of the same magnitude.

Three psychological principles constrain the shape of the value function. The first principle, *reference dependence*, suggests that outcomes are viewed relative to a reference point and hence coded as gains or losses. A \$10,000 bonus is seen differently if an employee is expecting a \$5000 bonus or a \$15,000 bonus. The bonus is seen as a \$5000 gain in the first case, but a \$5000 loss in the second case.

The second principle, *diminishing sensitivity*, borrows from research on the psychophysics of perception. Research has shown that an individual is highly likely to discriminate between a 2 and 3 kg weight, but not very likely to notice the difference between 22 and 23 kg weights [17]. Analogous findings appear in tasks ranging from the perception of line lengths to the perception of temperature. Prospect theory thus suggests that changes in value have a greater impact near the reference point than away from the reference point. Thus, there is a big difference between a \$100 gain and a \$200 gain, but a much smaller difference between gains of \$1100 and \$1200. Similarly, a loss of \$100 seems quite distinct from a loss of \$200, but losses of \$1100 and \$1200 seem pretty similar.

The principle of diminishing sensitivity away from the reference point is consistent with the reflection effect described above. Most people would choose \$500 for sure over an even chance at winning \$1000 or winning \$0. Diminishing sensitivity suggests that the first \$500 is more pleasurable than the second \$500. When this choice is changed to losses, however, preferences switch. The majority of individuals prefers an even chance at losing \$1000 or losing \$0 to losing \$500 for sure. This

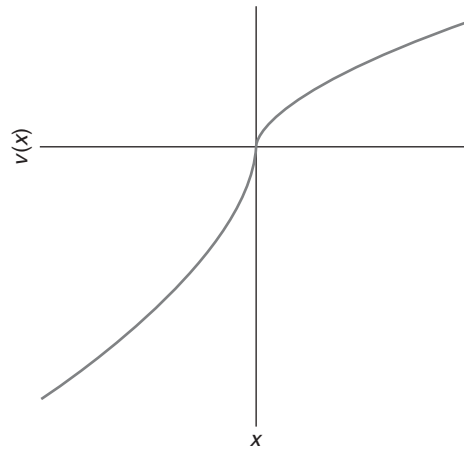


Figure 1. A typical prospect theory value function, $v(x)$, which is concave for gains and convex for losses (and thus consistent with the principle of diminishing sensitivity) and is steeper for losses than gains (and thus exhibits loss aversion).

preference is also consistent with diminishing sensitivity, because the first \$500 loss is much more painful than the second \$500 loss. Thus, the value function in Fig. 1 is concave for gains and convex for losses.

The third psychological principle underlying the value function is *loss aversion*. Recall that losses loom larger than gains. A loss of \$100 seems much more painful than a gain of \$100 seems pleasurable. Most people dislike a prospect that gives an equal chance at winning \$1000 or losing \$1000. These two examples are captured by a value function that is steeper for losses than gains ($-v(-x) > v(x)$), as is depicted in the value function in Fig. 1.

Probability Weighting Function. Prospect theory assumes that individuals do not weigh outcomes by their probability, as in expected utility theory, but by some distortion of probabilities. This distortion of probability is captured by prospect theory's probability weighting function, $\pi(p)$ (Fig. 2). Interestingly, the psychology governing the distortion of probability involves two of the three principles used to understand the value function, reference dependence and diminishing sensitivity away from a reference point. For probability, there are two obvious

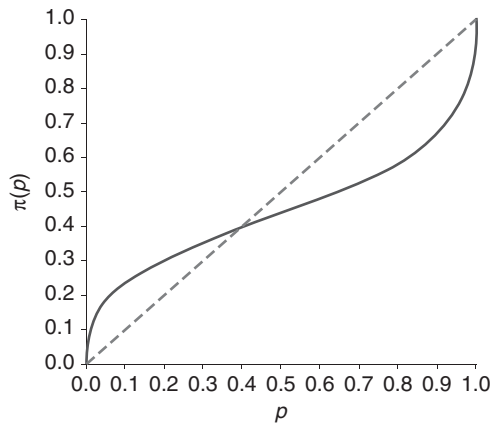


Figure 2. A typical prospect theory probability weighting function, $\pi(p)$, which is concave for small probabilities and convex for medium to large probabilities (and thus consistent with the principle of diminishing sensitivity).

reference points, certainty and impossibility, or 100% chance and 0% chance. The distortion of probability shown in the probability weighting function captures the diminishing sensitivity away from these two reference points. People are most sensitive to changes in probability when they are near 0% or 100% than when the change applies to intermediate probabilities. Consider adding a chance to win a large prize, such as a new car. Which of the following appears to be the least significant increase: increasing your chance of winning from 0% to 1%, from 33% to 34%, or from 99% to 100%? Most people regard the change from 0% to 1% as significant, because it changes the chance of winning from impossible to possible [18]. The change from 99% to 100%, too, seems noteworthy, because it shifts the odds from likely to certain. In contrast, the change from 33% to 34% seems negligible. Put simply, there is a categorical difference between impossible and possible or between likely and certain, but the difference between possible and slightly more possible is a matter of degree. The probability weighting function in Fig. 2 exhibits diminishing sensitivity: $\pi(0.01) - \pi(0) > \pi(0.34) - \pi(0.33)$ and $\pi(1) - \pi(0.99) > \pi(0.34) - \pi(0.33)$. The function is thus concave for small probabilities and convex for medium and large probabilities.

Diminishing sensitivity implies that low probabilities are typically given more weight than they would receive using expected utility. This overweighing is consistent with risk seeking for low probability gains (such as lottery tickets) and risk aversion for low probability losses (such as insurance). Medium to high probabilities are typically given less weight than they would receive using expected value. Such underweighing is consistent with risk aversion for medium to high probability gains, and risk seeking for medium to high probability losses. The probability weighting function also helps us understand the common-consequence effect described above. In the first choice, the probabilities of winning a prize are 33% and 34%, whereas the probabilities of winning with the second choice are 99% and 100%. The probability weighting function suggests that the first difference is trivial (thus decision makers look at outcomes to make this choice), whereas the second difference is consequential (and thus sufficient to drive the decision).

Psychological Principles

Prospect theory invokes three psychological principles: reference dependence, diminishing sensitivity, and loss aversion. The first two principles produce the characteristic form of the value function and the probability weighting function shown in Figs 1 and 2. For the value function, outcomes are viewed relative to a reference point, with outcomes above the reference point coded as gains and outcomes below the reference point coded as losses. The valuation of both gains and losses exhibits diminishing sensitivity away from the reference point. For the probability weighting function, probabilities are viewed relative to the reference points of impossibility (0) and certainty [1]. Changes away from these two reference points also exhibit diminishing sensitivity.

APPLICATIONS

Prospect theory has been useful for understanding real-world phenomena in a variety of domains. A few applications are discussed

below to provide a flavor of the broad reach the theory has had. For a description of other applications, see Ref. 19.

Mental Accounting

Although prospect theory was developed to explain how people make choices involving risk, the principles of prospect theory have been used to understand how people choose among certain alternatives as well. One influential application of prospect theory is known as *mental accounting* [20,21]. Mental accounting is a theory of how individuals and households make consumer and financial decisions. Most of the implications of mental accounting follow from the shape of prospect theory's value function.

Mental accounting suggests that people prefer to segregate gains and aggregate losses. Consider, for example, two pleasurable outcomes, a \$100 check from your aunt and \$50 won in a school raffle. Most people find it more enjoyable to view these outcomes separately, \$100 *plus* \$50, than together, \$150. The attractiveness of segregating gains follows from the shape of the value function and the psychological principle of diminishing sensitivity away from the reference point of zero. Next consider two painful outcomes, a \$100 speeding ticket and a \$150 tax liability. It seems clear that aggregating the losses is a good idea psychologically. In other words, a loss of \$250 seems less painful than two separate losses of \$100 and \$150. Mental accounting suggests that aggregating losses is more attractive than segregating losses. This simple principle helps us understand a number of real-world phenomena, including the attractiveness of flat-rate pricing plans. Most people find it less painful to pay \$70 for a monthly health club membership than \$10 each time they work out [22].

Mental accounting also suggests that different ways of framing the same activity can be viewed very differently. Suppose that you have decided to purchase a new car for \$22,000 and are contemplating purchasing a stereo upgrade. A car salesperson can frame the stereo as a \$300 purchase or an increase in the total price of the new car from \$22,000

to \$22,300. It is clear that the salesperson will be more effective at convincing you to purchase the upgrade if he/she frames the upgrade as an increase in price. Again, the principle at work is diminishing sensitivity of the value function.

Consider one final example of mental accounting. Imagine that you can walk 10 minutes to save \$5. Would you do so? People give different answers if the purchase is a \$25 calculator or a \$125 jacket. A \$5 discount seems more significant when it is applied to the inexpensive calculator rather than the expensive jacket. Of course, this behavior is not rational, since the decision to walk 10 minutes boils down to a choice of whether walking 10 minutes is worth \$5.

Status Quo Bias

In many situations, the decision maker must decide whether to choose a particular alternative or stick with the status quo. However, the status quo is frequently chosen much more often than it should be if the evaluation of the alternative were made simply based on the positive and negative features of that alternative. One study found a strong status quo bias for the choice of automobile insurance plans. Prior to deregulation in the early 1990s, Pennsylvania had an expensive insurance plan that allowed a driver to sue another driver in many situations. New Jersey had a cheaper plan that permitted its driver to sue only in limited situations. As a result of deregulation, citizens of Pennsylvania were permitted to trade down to the cheaper plan, and people in New Jersey were allowed to trade up to the more expensive plan. Nevertheless, 75–80% of drivers stuck with their original plan [23].

How does prospect theory help us understand why people tend to prefer the status quo? When giving up the status quo for another alternative there are, typically, trade-offs to be made. An option will have advantages and disadvantages relative to the status quo. The advantages are likely to be perceived as gains relative to the status quo, with the disadvantages perceived as losses. Because of loss aversion, losses

appear more significant than gains of comparable magnitude. The principle that losses loom larger than gains explains why people have a tendency to remain at the status quo—a phenomenon known as the *status quo bias*.

A related demonstration, called the *endowment effect*, shows once again the asymmetry between losing and gaining [24]. In one study, half of the participants are given a lottery ticket and half of the participants are given \$2 [25]. In each case, the participants can keep what they were given or trade it for what the others were given. However, very few participants chose to switch. Those who were given lottery tickets largely stuck with the lottery tickets, but those who were given the cash mostly elected to keep the cash. Once again, prospect theory's principle of loss aversion can explain this phenomenon. When people are endowed with an item, it becomes the reference point from which other options are evaluated. Since losses loom larger than gains, giving up an item that one possesses is perceived as more painful than acquiring an item of comparable magnitude that one does not already possess.

Other Applications

Prospect theory provides insights into numerous other real-world decisions. We sketch a few applications below to give a flavor of the power of prospect theory. First, in horse track betting, favorites are usually underbet, and longshots are typically overbet [26]. This behavior can be explained by the probability weighting function. Second, consider an investor who bought two stocks at \$50. One has increased in value to \$75, and the other has fallen in price to \$25. If the investor has to sell one stock, which would he/she sell? Financial economists have shown that investors overwhelmingly sell the winner rather than the loser [27]. The *disposition effect* suggests that selling the loser is unattractive, because the investor must come to terms with his/her losses.

Finally, we consider choices of retirement plans. Employers often provide a default

plan, or a plan that will be chosen if the employee does not make a choice. This plan is often a sensible plan, but it is often chosen by the vast majority of employees, even when another plan might make more sense [28]. To see why prospect theory can explain this choice, we assume that the default plan is the reference point for evaluating another retirement plan. A second plan will be better on some dimensions and worse on other dimensions. Because of loss aversion, the dimensions on which the second plan is inferior loom large compared to the dimensions on which it is superior, providing a substantial advantage to the default plan.

REFERENCES

1. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
2. Rabin M. The Nobel Memorial Prize for Daniel Kahneman. *Scandinavian Journal of Economics* 2003;105(2):157–180.
3. Knight FH. Risk, uncertainty and profit. New York: Houghton-Mifflin; 1921.
4. Bernoulli D. Specimen theoriae novae de mensura sortis. *Comment Acad Sci Imp Petrop* 1738;5:175–192.
5. von Neumann J, Morgenstern O. Theory of games and economic behavior. 2nd ed. Princeton (NJ): Princeton University Press; 1947.
6. Varian HR. Microeconomic analysis. New York: W.W. Norton; 1992.
7. Pratt JW. Risk aversion in the small and in the large. *Econometrica* 1964;32(1/2):122–136.
8. Raiffa H. Decision analysis: introductory lectures on choice under uncertainty. New York: Addison-Wesley; 1968.
9. Stigler GJ. The development of utility theory II. *J Polit Econ* 1950;58(5):373–396.
10. Clotfelter CT, Cook PJ. Selling hope: state lotteries in America. Cambridge (MA): Harvard University Press; 1989.
11. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5(4):297–323.
12. Abdellaoui M, Bleichrodt H, Paraschiv C. Loss aversion under prospect theory: a parameter-free measurement. *Manage Sci* 2007;53(10):1659–1674.

13. Machina MJ. Choice under uncertainty: problems solved and unsolved. *J Econ Perspect* 1987;1(1):121–154.
14. Samuelson P. Probability, utility, and the independence axiom. *Econometrica* 1952;20:670–678.
15. Wu G, Gonzalez R. Curvature of the probability weighting function. *Manage Sci* 1996;42(12):1676–1690.
16. Allais M. Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école Américaine. *Econometrica* 1953;21(4):503–546.
17. Stevens SS. On the psychophysical law. *Psychol Rev* 1957;64:153–181.
18. Gonzalez R, Wu G. On the shape of the probability weighting function. *Cognit Psychol* 1999;38(1):129–166.
19. Camerer CF. Prospect theory in the wild: evidence from the field In: Kahneman Daniel, Tversky Amos, editors. *Choices, values and frames*. New York: Cambridge University Press; 2000. pp. 288–300.
20. Thaler RH. Mental accounting and consumer choice. *Mark Sci* 1985;4(3):199–214.
21. Thaler RH. Mental accounting matters. *J Behav Decis Mak* 1999;12(3):183–206.
22. DellaVigna S, Malmendier U. Paying not to go to the gym. *Am Econ Rev* 2006;96(3):694–719.
23. Johnson E, Hershey J, Meszaros J, Kunreuther H. Framing, probability distortions, and insurance decisions. *J Risk Uncertain* 1993;7:35–51.
24. Kahneman D, Knetsch JL, Thaler RH. Experimental tests of the endowment effect and the Coase Theorem. *J Polit Econ* 1990;98:1325–1348.
25. Knetsch JL. The endowment effect and evidence of nonreversible indifference curves. *Am Econ Rev* 1989;79(5):1277–1284.
26. Thaler RH, Ziemba W. Anomalies: parimutuel betting markets: racetracks and lotteries. *J Econ Perspect* 1988;2(2):161–174.
27. Odean T. Are investors reluctant to realize their losses? *J Finance* 1998;53(5):1775–1798.
28. Madrian BC, Shea DF. The power of suggestion: inertia in 401(k) participation and savings behavior. *Q J Econ* 2001;116(4):1149–1187.

FURTHER READING

- Trepel C, Fox CR, Poldrack RA. Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cogn Brain Res* 2005;23(1):34–50.
- Wakker PP. *Decision under risk Prospect theory for risk and ambiguity*. Cambridge: Cambridge University Press; 2009.
- Wu G, Zhang J, Gonzalez R. In: Koehler DJ, Harvey N, editors. *Blackwell handbook of judgment and decision making*. Oxford: Blackwell; 2004. pp. 399–423.

PSYCHOLOGY OF RISK PERCEPTION

KATHERINE V. KORTENKAMP
Department of Psychology, University
of Wisconsin-La Crosse,
La Crosse, Wisconsin

COLLEEN F. MOORE
Department of Psychology, University
of Wisconsin-Madison, Madison,
Wisconsin

Risk perceptions refer to subjective judgments of risk that may involve both cognitions and emotions. Risk perceptions can be assessed in a variety of ways such as asking for estimates of the subjective probability and severity of an event, estimates of the risk of death or other negative outcomes, or ratings of the overall risk of an activity for a specific person, group of people, or the environment. Risk perceptions can be used as one predictor of public acceptance or opposition to new and existing technologies. Research in psychology has examined factors that influence or correlate with risk perceptions.

DEVIATIONS FROM NUMERICAL RISK ESTIMATES: PROBABILITY WEIGHTING, HEURISTICS, AND BIASES

It is well established that risk perceptions deviate from numerical risk estimates. One explanation for this is that subjective probability is nonlinearly related to numerical probability. Research has shown that people tend to overestimate small probabilities and underestimate large ones [1,2]. Kahneman and Tversky's [3] well-tested theory of decision making under risk, prospect theory, also reveals a tendency for people to over-weight small probabilities and under-weight larger ones. More recent research has found that whether people over- or under-weight small probabilities depends on the type of risky decision task employed. When presented with only descriptions of probabilities, people tend to over-weight small

probabilities, but when participants learn about probabilities through experience and feedback, they tend to under-weight small probabilities [4–6].

Apart from laboratory experiences with risk, experiences in the real world can also lead risk perceptions to deviate from numerical estimates through the availability heuristic. The availability heuristic says that people perceive events as more likely to occur if they are easier to bring to mind or imagine. According to the availability heuristic, risks that people have personally experienced, that are more vivid, and that receive more media coverage are judged as more prevalent relative to equally probable risks that are less available [7,8]. In addition, this heuristic can lead to what is called the *optimism bias*: people tend to judge themselves as less vulnerable to risks than others [9,10]. For example, if people hear news reports of others getting into car accidents, but they themselves have never gotten in one, they will perceive their personal risk as lower than others' risk [7].

BEYOND NUMERICAL RISK ESTIMATES: A PSYCHOMETRIC MODEL OF RISK PERCEPTION

Risk perception is not just determined by statistics and probabilities, but also by a number of qualitative factors related to the risks. Fischhoff and colleagues [11] first tested a psychometric model of risk perception and risk acceptability that examined several factors hypothesized to influence risk responses. Participants evaluated a wide range of technological and nontechnological risks in terms of their perceived risk, benefit, and acceptability. They also included variables to measure the degree to which the risks were voluntary, controllable, known, catastrophic, dreaded, and so on. A factor analysis of these variables revealed a two-dimensional model characterizing risk judgments that has been replicated

across many different groups of people [12]. One dimension is called *dread risk* and includes such factors as uncontrollable, dreaded, catastrophic, involuntary, inequitable, fatal, new, global, and not easily reduced. Risks found to be high on the dread dimension include nuclear weapons and reactors, radioactive waste, and DNA technology. Risks found to be low on dread include highly familiar technologies such as caffeine, aspirin, and power mowers. The second dimension is the known-unknown spectrum, characterized at one extreme by risks that are not understood by science or that are unobservable, new, and delayed in their effects. Risks found to be high in this dimension include DNA technology, electric fields, and nitrogen fertilizers; risks found to be low on this dimension include handguns, dynamite, and auto accidents. Nonexperts' risk perceptions are correlated with these two dimensions, especially the dread dimension. Technologies and activities that are rated more highly on the dread and unknown dimensions are perceived as riskier and less acceptable than those rated lower on these dimensions.

This model has been replicated in many different cultures [13–16]. In addition, the dread and knowledge dimensions have been identified in case studies of individuals encountering real risk events [17].

KNOWLEDGE: EXPERTS' VERSUS NONEXPERTS' RISK PERCEPTIONS

As knowledge about risks plays an important role in the psychometric model of risk, it makes sense that differences in knowledge would also influence risk estimates. Slovic [12] reported that experts' risk perceptions were not influenced by their ratings on the dread and unknown dimensions of risk, but instead their risk perceptions correlated with estimates of annual fatalities. Indeed, nonexperts can produce estimates of fatalities with rank-order similar to experts' technical estimates [12,18], nevertheless, their risk perceptions seem to be not entirely based on fatality estimates. Sjöberg [19] categorized comparisons of expert and nonexpert

risk perceptions into three types. First, there are similarities between experts and nonexperts when estimating fatalities caused by common, well-known risks. Second, nonexperts' risk perceptions are lower than experts' for some hazards. This seems to occur when individuals believe that they personally have control over a risk, and so, whereas risks to others might be high, risks to themselves are perceived to be low. Examples of these risks are smoking, drinking, driving, and adventure sports. Third, sometimes experts estimate risks to be lower than nonexperts do. This occurs most commonly for complex technological systems. The most extreme example is with nuclear power, where nonexperts perceive risk to be much higher than experts [12,20,21].

The difference in risk perceptions for complex technologies has been the most commonly studied. Several factors from the psychometric model could partially account for these differences, including differences between experts and nonexperts in familiarity, knowledge, and perceived controllability of technological risks. One study found that experts' risk judgments for biotechnology in food and medicine were higher than nonexperts', and that whether the applications of biotechnology were viewed as harmful, beneficial, unknown to science, and new partially accounted for the differences between experts' and nonexperts' risk perceptions [22].

A review of studies comparing risk perceptions made by experts and nonexperts revealed that differences in risk perceptions may also be due to socio-demographic differences between the groups, in terms of gender, education, age, and income [23]. Data on socio-demographic differences between expert and nonexpert samples is scarce in past studies and few included these characteristics in predictive models. But when reported, it seems that expert samples are more likely to be male, older, more highly educated, and wealthier. Some of these characteristics are predictive of lower risk perceptions, as discussed later.

TRUST AND CREDIBILITY

Having less knowledge about a technology or scientific process implies that nonexperts have to rely more on getting risk information from institutions and organizations that manage risks and communicate risks to the public. There is evidence that the public's perceptions of risk are related to whether they view these sources of risk information and policy as credible and trustworthy [24–28]. One study showed that greater trust in companies and universities that are involved in gene technology predicted lower perceptions of risks, higher perceptions of benefits, and greater acceptance of the technology [25]. In another study, participants who felt that authorities shared their values in public health and information policy were more trusting of the authorities, and this trust predicted lower risk perceptions of a cancer cluster [26].

The effect of trust and credibility on risk perceptions can also vary depending on the agency, authority, or institution involved [22,29]. Trust tends to be highest for research institutes, environmental groups, and consumer groups. Government agencies and political organizations are trusted less, and industry is the least trusted source of information about risks. However, these differences can vary somewhat depending on the nationality of the participants [29]. For example, in Italy and the United Kingdom, the government is perceived to be almost as equally trustworthy as consumer groups, whereas in Germany, the government is perceived as much closer in trustworthiness to industry.

A survey of New York residents about local environmental risks employed a measure of credibility developed by Meyer [30] that included five factors: trust, accuracy, fairness, completeness, and bias [24]. Lower ratings of credibility of the state government and the offending industry predicted higher risk perceptions; however, perceived credibility of the local newspaper did not predict risk perceptions. In another study, lower risk perceptions (pertaining to possible cancer clusters) were predicted by higher ratings of credibility for government and industry

sources of risk information but lower credibility ratings for citizen groups [28]. Also, perceiving government and industry sources as having higher credibility and citizen groups as having lower credibility predicted a reliance on heuristic (i.e., quicker and less effortful) processing of risk information, which, in turn, predicted lower risk perceptions. However, other research showed that attitudes toward risky technologies (in this case genetically modified food) can account for much greater variance in trust than the source of the risk information (industry, government, or consumer group) [29]. This suggests that risk perceptions may in fact influence trust. The direction of this relationship is impossible to untangle with correlational data.

A review of the literature concluded that there are three dimensions of trust that can be important for predicting risk perceptions: competence, care, and consensual values [31]. Competence refers to the fact that the credentials, knowledge, and experience of the agents assessing and managing the risk will predict how much they are trusted. According to the care dimension, people will trust a risk decision-making process more if they feel they have some control over or involvement in the process and ultimate decisions made, thus increasing procedural fairness, and if they are treated with respect and compassion. The values dimension stresses that people will trust risk assessors and managers more if they feel they share their values (e.g., cultural, political, and social values [32]).

Trust in institutions has also been shown to vary by gender, with women reporting lower levels of trust than men [25,26]. Differences in trust levels may partly explain gender and other individual differences (e.g., race, socioeconomic status) in risk perceptions [33].

INDIVIDUAL DIFFERENCES: RACE, GENDER, SOCIOECONOMIC STATUS

Many studies have shown gender differences in risk perceptions, with women showing higher risk perceptions than men [33–35]. Studies have also reported racial differences

with whites showing lower risk perceptions than nonwhites [33,36]. Interestingly, studies have shown what has been called a *white male* effect; that is, white males often have lower risk perceptions than white females, nonwhite males, and nonwhite females across a variety of different hazards, from smoking, to climate change, to air travel [33,37,38]. In one study, this effect was found to be driven by about 30% of the white male sample that had extremely low risk perceptions [33]. When compared to the rest of the sample, this group was found to have higher levels of education and income, more conservative political beliefs, and more trust in institutions and authorities. Another study showed that white men were more likely than women and nonwhite men to feel that they have control over health risks, that small risks can be imposed on people without their consent, and that the government should not be trusted to manage technological risks [37]. Differences in age, income, and education did not fully account for these gender, race, and “white male” effects.

Although differences in knowledge have been argued to be a reason for different risk perceptions between men and women, even in a sample of scientists, female scientists were shown to have higher perceptions of nuclear risks than male scientists [34]. There were also gender differences in attitudes about technology, the environment, and individual rights in this sample. For example, male scientists were more likely than female scientists to agree that existing technology can reduce risks from nuclear waste to acceptable levels, that the environment can be kept clean without requiring changes in lifestyle, and that risks can be imposed on people without their consent. Thus, there is more support for differences in attitudes between genders and races as explanations for differences in risk perceptions than there is support for differences in knowledge [35]. Specific perceptions of technological risk may reflect more general attitudes about technology, and trust in institutions as well as be influenced by perceived personal control over or vulnerability to risks.

Less research has specifically explored other demographic variables such as age, income, and education as predictors of risk perceptions. Effects of these variables on risk perceptions and attitudes have been found, but the small number of results does not seem to be entirely consistent [39,40].

PERCEPTIONS OF UNCERTAINTY AND RISK

Beliefs about the credibility of sources of risk information can also be influenced by the type of information that gets communicated. Uncertainty factors are almost always incorporated into formal risk assessments to account for what is unknown or for extrapolations [41], and as a result, a range of risk estimates might be reported including a best point estimate and an interval around that estimate. Research has shown that when uncertainty is communicated via interval estimates it can lead many people to perceive that the source of the risk information is more honest but less competent [42,43].

How does uncertainty about risk estimates affect risk perceptions? Research has shown that uncertainty can increase perceptions of risk. For example, when a range of risk estimates was presented, more people believed the true risk was at the higher end of the interval rather than the lower end [42–44]. Other research has found only small increases or no increase in risk perceptions due to uncertainty [45,46]. Uncertainty might increase risk perceptions because people tend to imagine the worst-case scenario [43–45] or because uncertainty alters perceptions of the scientists or institutions estimating the risks [42,43]. However, uncertainty could also potentially lead to lower risk perceptions for people who are motivated to perceive lower risks [47]. For example, smokers may be motivated to perceive lower risks from smoking in the face of uncertainty.

FAIRNESS, VALUES, ETHICAL CONSIDERATIONS

In perceptions of environmental and technological risks, often what matters more to

people than probabilities and statistics are issues of ethics such as justice and fairness in the distribution of risk burdens, informed consent to exposure, participation in the risk decision process, and duties to future generations [41,48–52]. For example, an analysis of letters of protest on the siting of nuclear waste facilities revealed that many of the arguments were about fairness, in terms of consent and unequal distribution of risk [52].

Several studies with survey data have found correlations between relevant moral issues and risk perceptions. The “dread” dimension of risk, which predicts risk perception and acceptability, includes the inequitable and involuntary components of risk that are related to the ethical issues of unequal distributions of risks and lack of informed consent to risk exposure [12]. The moral valuation of risks is a strong predictor (25% variance) of both the acceptability and perception of those risks [51,53,54]. Judgments of environmental injustice also predict higher perceived risk [55]. Finally, research showing a relationship between different value systems (e.g., cultural worldviews) and risk perceptions suggests that moral issues have the potential to influence risk perceptions by upsetting certain core values [56,57].

Two recent experimental studies seem to corroborate these correlational findings about the impact of ethical issues on risk perceptions. Providing information about the moral implications (for animals, the environment, and humans) of eating meat increased perceptions and reduced acceptability of the risks of meat consumption [58]. Support for a risk decision increased and risk perceptions decreased when the risk decision process included public involvement [59]. This effect may have occurred because people thought the decision-making process was more fair and more reflective of public values. In addition, public involvement may have led to a reframing of the risk from an involuntary risk imposed on the public to a voluntary risk agreed to by the public [59].

EMOTIONAL REACTIONS TO RISKS: RECENT THEORETICAL PERSPECTIVES

Feelings of trust, fairness, and dread can be seen as having emotional qualities, in addition to cognitive content. Research has begun to examine how emotional reactions to hazards influence risk perceptions. Some work on emotions and risk has examined the effect of anticipated emotions, that is emotions such as regret or pleasure that you would anticipate feeling after experiencing the outcome of a risky decision [60]. This research showed that 44–74% of the variance in people’s risk decisions can be explained by a preference for maximizing the subjective likelihood of experiencing positive anticipated emotions. Researchers have also made a distinction between anticipated emotions and emotions experienced during the risk decision-making process. Recent theoretical models of risk have emphasized the role of experienced emotions, and research has examined the effect of direct emotional reactions to risk information as inputs in risk judgments [61,62]. The focus of these models is the effect of emotional reactions on risk perceptions, but it is also understood that risk perceptions can influence emotional reactions as well.

It has been proposed that people rely on an “affect heuristic,” an easily accessible impression of how good or bad a hazard is, as an efficient cue for determining their risk perceptions [62,63]. Considering emotion as an important factor in risk perceptions can help explain some of the findings in the risk perception literature. For example, people are less sensitive to probabilities of risks when the risks are associated with emotionally laden images that trigger an affective response [64]. This is one explanation for why potentially catastrophic hazards with small probabilities (e.g., terrorist attacks, nuclear reactors), nonetheless, can be perceived as very risky. Some evidence for the role of emotion in risk perception also comes from neuroscience research, which has shown that damage to an area of the brain involved in emotion generation (ventromedial prefrontal cortex) hinders performance on risky decision tasks [65]. Current psychological theories of risk perception emphasize the importance

of both affective and cognitive evaluations of risk, and the interplay between these two processes [62,66].

CONCLUSION

Risk perception differs from numerical risk estimation and formal risk assessments because it is influenced by a wide array of psychological considerations such as trust, moral and social values, personal experience, and emotional reactions. Some variables that influence risk perceptions, especially ethical issues, trust, and future unknown outcomes, are very difficult to include in numerical risk assessments. For this reason, risk perceptions of members of the public sometimes add important information that is excluded from formal numerical risk assessments. In addition, risk communicators and managers must acknowledge and understand the role of psychological factors in risk perception in order to effectively gauge public responses to risks and design informative and effective risk messages and warnings.

REFERENCES

1. Attneave F. Psychological probability as a function of experienced frequency. *J Exp Psychol* 1953;46:81–86.
2. Lichtenstein S, Slovic P, Fischhoff B, *et al.* Judged frequency of lethal events. *J Exp Psychol Hum Learn Mem* 1978;4:551–578.
3. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica* 1979;47:263–291.
4. Barron G, Erev I. Small feedback-based decisions and their limited correspondence to description-based decisions. *J Behav Decis Mak* 2003;16:215–233.
5. Jessup RK, Bishara AJ, Busemeyer JR. Feedback produces divergences from prospect theory in descriptive choice. *Psychol Sci* 2008;19:1015–1022.
6. Hertwig R, Barron G, Weber EU, *et al.* Decisions from experience and the effect of rare events in risky choices. *Psychol Sci* 2004;15:534–539.
7. Slovic P, Fischhoff B, Lichtenstein S. Facts versus fears: understanding perceived risk. In: Kahneman D, Slovic P, Tversky A, editors. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press; 1982. pp. 462–492.
8. Weinstein ND. Effects of personal experience on self-protective behavior. *Psychol Bull* 1989;105:31–50.
9. Weinstein ND. Unrealistic optimism about future life events. *J Pers Soc Psychol* 1980;39:806–820.
10. Weinstein ND. Optimistic biases about personal risks. *Science* 1989;246:1232–1233.
11. Fischhoff B, Slovic P, Lichtenstein S, *et al.* How safe is safe enough? A psychometric study of attitudes toward technological risks and benefits. *Policy Sci* 1978;9:127–152.
12. Slovic P. Perception of risk. *Science* 1987;236:280–285.
13. Englander T, Farago K, Slovic P, *et al.* A comparative analysis of risk perception in Hungary and the United States. *Soc Behav* 1986;1:55–66.
14. Kleinhesselink RR, Rosa EA. Cognitive representation of risk perceptions: a comparison of Japan and the United States. *J Cross Cult Psychol* 1991;22:11–28.
15. Teigen KH, Brun W, Slovic P. Societal risk as seen by a Norwegian public. *J Behav Decis Mak* 1988;1:111–130.
16. Keown CF. Risk perception of Hong Kongese vs. Americans. *Risk Anal* 1989;9:401–405.
17. Trumbo CW. Examining psychometrics and polarization in a single-risk case study. *Risk Anal* 1996;16(3):423–432.
18. Wright G, Bolger F, Rowe G. An empirical test of the relative validity of expert and lay judgments of risk. *Risk Anal* 2002;22:1107–1122.
19. Sjöberg L. Risk perception: experts and the public. *Eur Psychol* 1998;3:1–12.
20. Barke RP, Jenkins-Smith HC. Politics and scientific expertise: scientists, risk perception, and nuclear waste policy. *Risk Anal* 1993;13:425–439.
21. Flynn J, Slovic P, Mertz CK. Decidedly different: expert and public views of risks from a radioactive waste repository. *Risk Anal* 1993;13:643–648.
22. Savadori L, Savio S, Nicotra E, *et al.* Expert and public perceptions of risk from biotechnology. *Risk Anal* 2004;24(5):1289–1299.
23. Rowe G, Wright G. Differences in expert and lay judgments of risk: myth or reality? *Risk Anal* 2001;21:341–356.
24. McComas KA, Trumbo CW. Source credibility in environmental health-risk controversies:

- application of Meyer's credibility index. *Risk Anal* 2001;21:467–480.
25. Siegrist M. The influence of trust and perceptions of risks and benefits on the acceptance of gene technology. *Risk Anal* 2000;20:195–204.
26. Siegrist M, Cvetkovich GT, Gutscher H. Shared values, social trust, and the perception of geographic cancer clusters. *Risk Anal* 2001;21:1047–1053.
27. Slovic P, Flynn JH, Layman M. Perceived risk, trust, and the politics of nuclear waste. *Science* 1991;254:1603–1607.
28. Trumbo CW, McComas KA. The function of credibility in information processing for risk perception. *Risk Anal* 2003;23:343–353.
29. Frewer LJ, Scholderer J, Bredahl L. Communicating about the risks and benefits of genetically modified foods: the mediating role of trust. *Risk Anal* 2003;23:1117–1133.
30. Meyer P. Defining and measuring credibility of newspapers: developing an index. *Journalism Q* 1988;65:567–574.
31. Johnson BB. Exploring dimensionality in the origins of hazard-related trust. *J Risk Res* 1999;2:325–354.
32. Earle TC, Cvetkovich GT. *Social trust: toward a cosmopolitan society*. Westport (CT): Greenwood Publishing Group; 1995.
33. Flynn J, Slovic P, Mertz CK. Gender, race, and perception of environmental health risks. *Risk Anal* 1994;14:1101–1108.
34. Barke R, Jenkins-Smith H, Slovic P. Risk perceptions of men and women scientists. *Soc Sci Q* 1997;78:167–176.
35. Davidson DJ, Freudenburg WR. Gender and environmental risk concerns a review and analysis of available research. *Environ Behav* 1996;28:302–339.
36. Johnson BB. Gender and race in beliefs about outdoor air pollution. *Risk Anal* 2002;22:725–738.
37. Finucane ML, Slovic P, Mertz CK, *et al.* Gender, race, and perceived risk: the “white male” effect. *Health Risk Soc* 2000;2:159–172.
38. Palmer CGS. Risk perception: another look at the “white male” effect. *Health Risk Soc* 2003;5:71–83.
39. Kraus N, Malmfors T, Slovic P. Intuitive toxicology: expert and lay judgments of chemical risks. *Risk Anal* 1992;12:215–232.
40. Lazo JK, Kinnell JC, Fisher A. Expert and layperson perceptions of ecosystem risk. *Risk Anal* 2000;20:179–193.
41. National Research Council (NRC). *Understanding risk: Informing decisions in a democratic society*. Washington (DC): National Academy Press; 1996.
42. Johnson BB. Further notes on public response to uncertainty in risks and science. *Risk Anal* 2003;23:781–789.
43. Johnson BB, Slovic P. Lay views on uncertainty in environmental health risk assessment. *J Risk Res* 1998;1:261–279.
44. Viscusi WK, Magat WA, Huber J. Communication of ambiguous risk information. *Theor Decis* 1991;31:159–173.
45. Johnson BB, Slovic P. Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. *Risk Anal* 1995;15:485–494.
46. Kuhn KM. Message format and audience values: interactive effects of uncertainty information and environmental attitudes on perceived risk. *J Environ Psychol* 2000;20:41–51.
47. Kunda Z. The case for motivated reasoning. *Psychol Bull* 1990;108:480–498.
48. Moore CF. *Children, pollution, and why scientists disagree*. 2nd ed. Oxford: Oxford University Press; 2009.
49. Rayner S, Cantor R. How fair is safe enough? The cultural approach to societal technology choice. *Risk Anal* 1987;7:3–9.
50. Shrader-Frechette K. Duties to future generations, proxy consent, intra- and intergenerational equity: the case of nuclear waste. *Risk Anal* 2000;20:771–778.
51. Sjöberg L, Drottz-Sjöberg B. Fairness, risk and risk tolerance in the siting of a nuclear waste repository. *J Risk Res* 2001;4:75–101.
52. Vari A. Public perceptions about equity and fairness: siting low-level radioactive waste disposal facilities in the U.S. and Hungary. *Risk Health Saf Environ* 1996;7:181–196.
53. Sjöberg L, Torell G. The development of risk acceptance and moral valuation. *Scand J Psychol* 1993;34:223–236.
54. Sjöberg L, Winroth E. Risk, moral value of actions, and mood. *Scand J Psychol* 1986;27:191–208.
55. Satterfield TA, Mertz CK, Slovic P. Discrimination, vulnerability, and justice in the face of risk. *Risk Anal* 2004;24:115–129.
56. Lima ML, Castro P. Cultural theory meets the community: worldviews and local issues. *J Environ Psychol* 2005;25:23–35.

57. Wildavsky A, Dake K. Theories of risk perception: who fears what and why? *Daedalus* 1990;119:41–60.
58. Berndsen M, van der Pligt J. Risks of meat: the relative impact of cognitive, affective and moral concerns. *Appetite* 2005;44:195–205.
59. Arvai JL. Using risk communication to disclose the outcome of a participatory decision-making process: effects on the perceived acceptability of risk-policy decisions. *Risk Anal* 2004;23:281–289.
60. Mellers B, Schwartz A, Ritov I. Emotion-based choice. *J Exp Psychol Gen* 1999;128:332–345.
61. Loewenstein GF, Weber EU, Hsee CK, *et al.* Risk as feelings. *Psychol Bull* 2001;127:267–286.
62. Slovic P, Finucane ML, Peters E, *et al.* Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal* 2004;24:311–321.
63. Finucane ML, Alhakami A, Slovic P, *et al.* The affect heuristic in judgments of risks and benefits. *J Behav Decis Mak* 2000;13:1–17.
64. Rottenstreich Y, Hsee CK. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol Sci* 2001;12:185–190.
65. Bechara A, Damasio H, Damasio AR. Emotion, decision making, and the orbitofrontal cortex. *Cereb Cortex* 2000;10:295–307.
66. Finucane ML, Holup JL. Risk as value: combining affect and analysis in risk judgments. *J Risk Res* 2006;9:141–164.

PUBLIC HEALTH, EMERGENCY RESPONSE, AND MEDICAL PREPAREDNESS I: MEDICAL SURGE

EVA K. LEE

ANNA YANG YANG

Center for Operations Research in
Medicine and Healthcare, School
of Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta, Georgia

NSF I/UCRC Center for Health
Organization Transformation,
Atlanta, Georgia

School of Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta, Georgia

FERDINAND PIETZ

Strategic National Stockpile, Office
for Public Health Preparedness
Response, Centers for Disease
Control and Prevention,
Atlanta, Georgia

BERNARD BENECKE

Global Disease Detection and
Emergency Response, Center for
Global Health, Centers for
Disease Control and Prevention,
Atlanta, Georgia

INTRODUCTION

In the aftermath of the September 11 terrorist attacks and the dissemination of anthrax in 2001, the ability of the US health care system to provide an effective and coordinated response to mass casualty or complex incidents came under intense scrutiny. More recently, the devastation caused by Hurricane Katrina and the mass disruption of public health and medical services along the Gulf Coast spotlighted the need for cohesive strategies that focus on management systems for major public health and medical response.

As stated by the US Department of Health and Human Services (HHS), the concept of medical surge forms the cornerstone of preparedness planning efforts for major medical incidents. Medical surge refers to the ability to provide adequate medical evaluation and care during events that exceed the limits of the normal medical infrastructure of an affected community. It encompasses the ability of health care organizations to survive a hazard impact and maintain or rapidly recover operations that were compromised (a concept known as *medical system resiliency*).

HHS makes a distinction between medical surge capacity and medical surge capability. Capacity relates to the ability to handle an overall increase in volume of patients, whereas capability refers to the ability to handle patients who are presenting with problems (e.g., a highly contagious disease) that are not normally handled at the location, as well as patient populations who are not normally handled (e.g., pediatric patients at a nonpediatric facility).

During a disaster, three categories of “players” must be ready to act: public health, individual facilities, and the community at large. Therefore it is important to consider surge capacity as it relates to each of these players. Table 1 summarizes the definitions of surge capacities at each of these levels.

Surge capacity plans should be scalable and flexible to cope with the many types and varied timelines of disasters. Improving surge capacity involves a multidisciplinary approach that includes all relevant partners and functions, from each of the community’s health care facilities to local, state, and federal agencies. Facility-based or “surge in place” solutions maximize health care facility capacity for patients during a disaster. When these resources are exceeded, community-based solutions, including the establishment of off-site hospital facilities, may be implemented [1].

This article focuses primarily on approaches for hospital surge capacity planning. The section titled “Surge Patterns” discusses

Table 1. Definition

Term	Definition
Surge capacity	Ability to manage a sudden, unexpected increase in patient volume that would otherwise severely challenge or exceed the current capacity of the health care system
Surge capability	Ability to manage patients who require specialized evaluation or interventions that are not normally provided at the facility/location (e.g., contaminated, highly contagious, or burn patients)
Public health surge capacity	Ability of the public health system to increase capacity not only for patient care but also for epidemiologic investigation, risk communication, mass prophylaxis or vaccination, mass fatality management, mental health support, laboratory services, and other activities
Facility-based surge capacity	Actions taken at the health care facility level that augment services within the response structure of the health care facility; may include responses that are external to the actual structure of the facility but are proximate to it (e.g., medical care provided in tenting on the hospital grounds). These responses are under the control of the facility's incident management system and primarily depend on the facility's emergency operations plans.
Community-based surge capacity	Actions taken at a community level to supplement health care facility responses. These may provide for triage and initial treatment, nonambulatory care overflow, or isolation (e.g., off-site "hospital" facility). These responses are under the control of the jurisdictional response (e.g., public health, emergency management) and represent a public effort to support and augment the health care system.

related issues on hospital surge patterns. The section titled "Components of Surge Capacity" reviews the four component of surge capacity: care personnel and staff, facilities, equipment and supplies, and incident management system. This is followed by a discussion of balancing daily operational efficiency and maintaining appropriate surge capacity in the section titled "Balancing Daily Operational Efficiency and Maintaining Appropriate Surge Capacity." The section titled "International Collaboration" highlights the importance of international collaboration. The section "Operations Research Methodologies" provides a brief overview of OR methodologies, including simulation, optimization, stochastic processes and systems dynamics models used in medical surge analysis. The article concludes with the section "challenges."

MEDICAL SURGE

Surge Patterns

In order to better plan for hospital surge capacity during emergency events, researchers have studied data from various disaster events in the past decade to obtain a glimpse of the impact on the health care facilities from these events.

Okumura *et al.* [2] investigated the case of 640 victims of the 1995 Tokyo gas attack, focusing on their symptoms and treatment. Case studies were given to describe the symptoms of patients upon arrival and the associated treatment provided. Emergency department (ED) treatment consisted mostly of administration of atropine sulfate and dosage with 2-PAM within a few hours of the initial exposure to sarin gas for severe and moderate cases. A few clinical management

issues were raised. These included poor ventilation in the ED area, inadequate decontamination due to privacy provision and the large number of affected population, required patient followup, and monitoring for long-term health effects.

Hogan *et al.* [3] studied the pattern in epidemiologic injury data on patients who suffered acute injuries after the 1999 Oklahoma City bombing and concluded that more seriously injured patients tend to arrive in a second wave at EDs since they are less mobile and often require on-site treatment. It is understandable that the closest hospitals received the greatest number of victims by all transportation methods initially. There are lessons learnt for EDs in terms of preparing for such phenomenon and to prepare and allocate/mobilize resources accordingly. Disaster response planning should avoid overloading hospitals in the vicinity of disasters. Rapid arrival of patients with nonlife threatening injuries may overwhelm the EDs. In this case, EDs must equip themselves with clinical personnel who can perform rapid assessment and stabilization, instead of staffing with specialty doctors who generally lack such skills.

Roccaforte [4] reviewed various aspects of response by Bellevue Hospital Center to the 2001 World Trade Center attack, including challenges faced with communication, organization, staffing, and logistics. The hospital is 2.5 miles from the site of the attack. When the news just arrived, communication channels were overloaded. Poor communication hindered adequate response planning and therefore wasted resources. Fortunately, the response was well supported by full staffing on the day of the event. Obviously, full staffing would likely not be in place if an emergency event occurred at night or on a weekend or holiday. In any event, the on-hand workforce should be allocated to control and preserve important resources for likely more serious cases in the second wave. Upon arrival of patients, they were triaged based on injury conditions. Out of the 90 patients that arrived during the first 5 h, 56 had minor injuries and most of them had respiratory complaints. Five of the 90 patients suffered mixture of fractures and burns. Three patients were seriously injured.

Schull *et al.* [5] investigated the surge capacity associated with restrictions on nonurgent hospital utilization and expected admissions during the Toronto Severe Acute Respiratory Syndrome (SARS) outbreak. The study compared expected influenza-related hospitalization with the actual reduction in hospital admission for eight weeks for three different pandemic severity levels: mild, moderate, and severe. It was found that influenza-related admissions generally exceed the reduction in admissions of nonurgent cases. Thus, simply reducing hospital admissions may not be sufficient for handling an influenza-related medical surge. Other strategies should be explored.

Gavagan *et al.* [6] examined medical response in the aftermath of Hurricane Katrina at the Houston Astrodome/Reliant Center Complex. During the two weeks of operation, the center provided shelter to 27,000 evacuees, over 11,000 of whom sought some form of medical care. Common health problems with the patients included uncontrolled hypertension, respiratory infection, and acute gastroenteritis. At the medical center, five levels of care characterized treatment, from immediate triage upon arrival to the use of community resources. Most of the adult visits to the Katrina clinic were related to chronic disease or medication refills. Another major health condition found in adults related to skin problems as a result of exposure to tainted water. Over 3500 children and infants were seen at the medical center during a two-week span. People 65 or older accounted for 56% of the total patients seen, and 62% of the 37 total deaths. Other problems that were addressed include mental health as well as obstetrics and gynecology. In terms of personnel management, instances were found where noncredentialed people setup medical clinics within the center, raising the critical issue of careful verification of staff qualification.

Stratton and Tyler [7] studied characteristics of medical surge capacity demand data based on existing databases and literature and concluded that the earliest outside assistance for a community subject to a sudden-impact disaster arrived in roughly 24 h, with a range that reached to 96 h. After

sudden-impact disasters, 84–96% of health care demand was managed by ambulatory care. EDs were the access point for care, with peak demand time occurring within 24 h. The authors also suggested that lack of standards for reporting disaster data and limited depth of the current disaster literature are limitations for surge capacity related research.

Components of Surge Capacity

Surge capacity encompasses multidimensional aspects such as potential patient beds; available space in which patients may be triaged, managed, vaccinated, decontaminated, or simply located; available care personnel of all types; necessary medications, supplies, and equipment; and even the legal capacity to deliver health care under situations that exceed authorized capacity. Further, close collaboration between public health and health care facilities is very important to effective disaster response.

Surge capacity is tight in the nation's hospitals, as hospitals have undergone years of budget tightening and consequent efforts to control costs, resulting in diminished patient capacity. Another factor that affects rapid response to surges includes an overall shortage of nurses. In a 2006 nationwide blue-ribbon panel of health care experts, led by Dr. Kelen, head of emergency medicine at The Johns Hopkins Hospital and Director of the Johns Hopkins Office of Critical Event Preparedness and Response, it was recommended that hospital plans for a surge of disaster victims should begin with a strategy to empty their beds of relatively healthy patients. Preliminary data suggest that such a strategy could safely empty 70% of a hospital's inpatient population within 72 h.

There is consensus among health care officials that, whether dealing with a natural disaster like Hurricane Katrina, a possible terrorist attack like September 11, or epidemics like SARS or avian flu, affected hospitals have few means of making room for large numbers of incoming casualties.

In one common disaster response, medical centers would setup additional beds wherever they can (in hallways, cafeterias, etc.). But, there is serious concern that staffing

levels could not expand to care for so many new patients, and "disposition classification" is being recommended as a must. Such an approach is important as it ensures that both the disaster victims as well as those patients who are already hospitalized, all would receive adequate treatment.

Disposition classification puts patients in one of five categories, based on their considered risk of a life-threatening or life-impairing medical problem within 72 h of hospital discharge. Patients classified as "minimum" risk could go home upon being discharged. Those in the "low-risk" group could also be transferred home depending on the severity and scope of the disaster. Those in the "moderate" category could not go home but could be transferred to a facility offering basic medical resources. "High-risk" patients could only be transferred to an acute-care facility and "very high-risk" patients could only be served in a critical care facility.

The decision making process is complex and the logistics behind transferring such a large number of patients to other facilities is hardly trivial. However, it offers an ethical and nonemotional approach, and could be implemented outside of a disaster situation as a tool to manage even routine, everyday overloads of hospital resources [8].

The four critical areas comprising hospital surge capacity identified by various investigators are personnel, facility, equipment and supply, and management [9]. Generally, hospitals have a concept on the amount of resources required for each category above a given number of patients. Therefore, improving surge capacity centers around the problem of determining how much a hospital could expand their capacity, how quickly the expansion can be implemented, and how long it can be sustained. For example, an arbitrary amount of 20% on top of normal capacity has been set for Israeli hospitals. In the United States, the Health Resource and Service Administration provides a surge guideline of 500 beds per one million population for patients with symptoms of acute infectious disease and 50 beds per one million population for noninfectious disease and injury [10]. It is also noted that a hospital will be unable to further increase surge

capacity once its saturation point has been reached. Strategies for determining when hospitals will reach this position currently do not exist [9].

We will now discuss each of these four areas in more detail.

Care Personnel and Staff. When disaster happens, it is natural for hospitals to increase the number of staff by maximizing the use of their own personnel. This requires established protocol for revision of staff work hours (e.g., a 12-h shift is standard in disaster events), call-back of off-duty personnel, use of nonclinical staff (e.g., nurse administration) in clinical roles as appropriate, and reallocation of outpatient staff resources [1].

One possibility to further support hospital staffing requirements is by providing emergency credentialing to volunteer health care professionals. However, many issues relating to this practice need to be addressed: for example, an institution's willingness to permit such individuals to work in their facility, credibility and confidentiality issues, and retention of such workers if they are needed for a time period that goes beyond expectation. Schultz and Stratton [11] suggested a solution to credential issues by sharing a database of credentialed health care workers among mutually acceptable organizations which will then assist in staff planning during disasters. In addition, family members could also assist in taking care of patients.

As discussed, the least serious casualties usually arrive at the hospital first during a disaster. Health care personnel may not be aware that more serious patients are yet to arrive. Careful planning must take this factor into consideration to avoid filling existing beds with minor injuries [12]. The "disposition classification" scheme can be used to help with proper resource allocation [8].

Research was conducted on health care workers' ability and willingness to report to duty during catastrophic disasters [13]. It was reported that one-fourth to one-third of the workforce may be deliberately absent for some period when the use of weapons of mass exposure was involved. Therefore, staffing during disaster will be a serious problem. In our own work, we focus on

optimal (minimizing) resource allocation and staffing assignment for mass population protection, and improving system throughput via maximizing operations efficiency, optimal clinic layout design, and process workflow [14–17] (see **Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts**).

Facility. Expanding bed capacity during an emergency can be accomplished by multiple strategies. These include the following:

- *Reduce Current Bed Occupancy:* such as canceling elective procedures, early discharge of inpatients [18], and clearing patients with minor injuries;
- *Expand Capacity Within Hospital:* including converting single rooms to double occupancy, using flat space as patient care wards, and using alternative sites in the parking lot;
- *Collaboration with Other Hospitals and Community Facilities:* including transferring patients to nearby hospitals or health care facilities and collaborating with other community organizations such as schools and churches for patient treatment and storage of supplies.

Some of the issues associated with facility expansion include the following:

- *Nature of Disaster and Effected Areas.* For example, in times of natural disasters such as hurricanes or earthquakes, large areas may be affected at the same time, thus making the transfer between hospitals or nearby community centers impossible;
- *Security Issues with Temporary Health Care Facilities.* For example, setting up temporary hospitals in tents is a common practice after earthquake events. These facilities are often open to the public, thus making it difficult to control access and prevent theft of equipment and supplies;

- *Local Practicality.* The practicality of the site is influenced by the climate and infrastructure support (e.g., water, electricity, and other utilities).

Equipment and Supplies. Building surge capacity must also take into consideration the requirement on equipment, supplies and pharmaceuticals on top of availability of wards, space, and medical staff. In the United States, the Strategic National Stockpile (SNS) is a program developed by the federal Centers for Disease Control and Prevention (CDC) Bioterrorism Preparedness and Response Initiative that assists states and communities in responding to extreme public health emergencies from terrorist attacks or natural disasters. SNS has large quantities of medicine and medical supplies to protect the public in a health emergency. The SNS is operated as a collaborative effort between local, state, and federal officials. The SNS contains medicines, vaccines, medical and surgical supplies, life-support medications, and breathing supplies. It is not a first response tool. Rather, it is designed to supplement and resupply a local and/or state response to an emergency within the United States or its territories. The two major components of the stockpile are 12-h Push Packages and the Managed Inventory. The Push Packages contain 50 tons of pharmaceuticals, antidotes, and medical supplies that will treat a variety of illnesses. Managed inventory can either replenish the Push Package or be tailored with specific medical supplies to respond to a defined agent or other specific public health threat. The 12-h Push Packages arrive within 12 h once deployed at the federal level. Supplies are distributed to local communities as quickly as possible. Managed inventory is shipped to arrive within 24–36 h after deployment (see ***Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts***).

Major issues with medical supplies and equipment held are as follows:

- *Speed of Delivery.* Hanfling [19] argued that it is necessary for each local

community to create a local cache for these items. He further explored the funding requirements for this approach.

- *Amount of Supplies to Stockpile.* It is possible that stockpiled supplies that are very expensive are no longer functional when a disaster outbreak happens (e.g., ventilators used for pandemic influenza). Whether we should stockpile for worst-case scenario or most likely scenario is another issue to be discussed.
- *Strategies for Effective Dispensing.* For example, in the event of an infectious disease outbreak or a biological attack, operational and strategic plans must be designed and put in place to dispense the medical supplies (e.g., vaccines, antibiotics) to the affected population effectively and efficiently [14–17] (see ***Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts***).

Incident Management System. Incident management systems and cooperative planning processes will facilitate maximal use of available resources. However, resource limitations may require implementation of triage strategies.

In building surge capacity, Burkle [20] investigated a population-based bioevent triage management with the goal of identifying the population requiring immediate care, preventing secondary transmission and building surge capacities that are population specific. In this management method, the population self-identify into one of five categories: susceptible but not exposed, exposed but not yet infectious, infectious, removed by death or recovery, and vaccination or prophylactic medication.

One of the most important issues of surge capacity is the ability to implement an incident management system. In the United States, one of the most conflicted issues with incident management is that the health care system includes many private entities. Private hospitals have no jurisdictional boundaries and are not under

any governmental or municipal operational authority or control. Successful implementation of an incident management system for effective disaster response requires careful planning and close collaboration between public health and (private) health care facilities, and organizations from the local level to the federal government level.

Balancing Daily Operational Efficiency and Maintaining Appropriate Surge Capacity

With the need to balance budgets, hospital administrators often look toward closures and downsizing to reduce costs, and toward process optimization to improve efficiency. These efforts to decrease costs and increase efficiency often run counter to simultaneous efforts to enhancing or maintaining surge capacity, as staffing and usage of resources are maintained at very high utilization rates, leaving very little room for medical surge. There is also a serious gap between daily surge capacity versus rapid large-scale surge requirements for catastrophic events.

Among the many issues that the US Congress needs to address for public health and medical preparedness and response, medical surge capacity has been identified as inadequate. The Government Accountability Office (GAO) issued an updated report in January 2010, “State Efforts to Plan for Medical Surge Could Benefit from Shared Guidance for Allocating Scarce Medical Resources.” The report states that many hospitals in the United States are unprepared for treating the overwhelming “surge” of victims from a large-scale mass casualty event. GAO said that “based on a review of state emergency preparedness documents and interviews with 20 state emergency preparedness officials, many states had made efforts related to three of the four key components of medical surge that GAO had identified—increasing hospital capacity, identifying alternate care sites, and registering medical volunteers, but fewer had implemented the fourth: planning for altering established standards of care. More than half of the 50 states had met or were close to meeting the criteria for the five medical surge-related sentinel indicators for hospital capacity reported in

the Hospital Preparedness Program’s 2006 midyear progress reports. In a 20-state review, GAO found the following. All 20 were developing bed reporting systems and most were coordinating with military and veterans hospitals to expand hospital capacity. Eighteen were selecting various facilities for alternate care sites. Fifteen had begun electronic registering of medical volunteers, and fewer of the states—7 of the 20—were planning for altered standards of medical care to be used in response to a mass casualty event.”

International Collaboration

As illustrated vividly by the recent tsunamis and the 2010 Haiti earthquake, surge capacity remains critical when catastrophic events elsewhere in the world demand it. Research has been done internationally on improving health care surge capacity in the face of a high-consequence event. Peleg and Kellermann [21] summarized 14 principles developed by Israel’s Ministry of Health on enhancing system surge capacity. Shih and Koenig [22] also discussed Taiwan’s experience in managing surge needs through sending doctors to the scene of an event, immediately recalling off-duty hospital personnel, managing volunteers, designating specialty hospitals, and use of incident management systems.

In the recent Haiti situation, the destruction of the local health care infrastructure required reaching out to hospitals outside the country, as well as immediate establishment of a network of on-site health-care facilities. In this case, the United States naval ship USS Comfort served as an acute treatment site, whereas temporary make-shift hospital tents were setup to provide medical care for the mass injured. This illustrates the intimate and critical interplay between hospital surge capacity and the urgent need for setting up (temporary) facility-based and community-based solutions to manage medical surges.

OPERATIONS RESEARCH METHODOLOGIES

Among operations research methodologies, simulation, optimization, stochastic

processes, and system dynamics models are often employed in the field of medical surge analysis.

A system dynamics model was developed by Manley *et al.* to assess hospitals ability to handle surges of demand during disasters [23]. The model represents all major flows of patients through the hospitals and helps the hospital to identify potential bottlenecks in the system when facing various situations.

Thompson *et al.* [24] attempted to solve the optimal reallocation problem for proactively transferring patients between floors in anticipation of demand surge by modeling it as a finite-horizon Markov decision process. A decision-support system based on the optimization model was implemented at Windham Hospital in Willimantic, Connecticut. It was estimated that this implementation projects a financial gain of 1% of the hospitals total revenue and a 50% reduction on the average wait time for an admitted patient to be transferred to a floor.

Gerchak *et al.* [25] developed a stochastic dynamic programming model for scheduling of elective surgery when the operating rooms capacity utilization is uncertain. The model addresses the trade-off between timeliness of admission and reserving capacity for emergency cases.

Einav *et al.* [26] suggested guidelines for hospital resource utilization during terror-related multiple casualty incidents based on data gathered from six level I trauma centers. It was found that there is a high demand for ED, OR, and ICU. Anesthesiologists, general, thoracic, and vascular surgeons are in immediate demand. ICU admissions occur simultaneously with ongoing patient arrival to the ED. Most patients operated on within the first 2 h require multidisciplinary surgical teams. Demand for orthopedic and plastic surgery and anesthesiology services continues for more than 24 h.

Patrick and Puterman [27] demonstrated how queueing theory provides managers with insights into the causes for excessive wait times and the relationship between wait times and capacity. A case study involving Markov decision process, linear programming, and simulation was included to optimize the scheduling of patients with

multiple priorities. It was shown that a wait time target can be achieved with judicious use of surge capacity in the form of overtime when applying the above methods.

Paul *et al.* [28] attempted to accurately model the behavior of the system by a transient modeling approach using simulation and exponential models. The parameters of the exponential model were regressed using outputs from designed simulation experiments. The model allows real-time estimation of capacity for hospitals of various sizes. Priority-based routing of patients within the hospital is also allowed in the modeling, where patients are guided based on the severity of injuries.

Roccaforte and Cushman [29] researched resource organization for optimally matching the demand and capacity of the treating facility.

Hoard *et al.* [30] introduced a conceptual framework for a study that applies a system dynamics approach to rural disaster preparedness and planning. This approach can be used for building computer simulation models and applies to hospital surge capacity to answer various “what-if” questions such as placing supplementary nursing staff for emergency outbreaks.

Green [31] addressed some problems faced in hospital capacity planning, such as government impact, competition, and influence of managed care. Oftentimes, the decision maker makes these complex decisions based on routine or experience. Green provided examples of how OR models can be used to provide insights into operational strategies and practices.

In our own work, the decision-support system RealOpt combines OR modeling techniques, novel and large-scale computational engines, sophisticated graph-drawing tools, 3D geographical spatial information with federal census data, and demographic and socio-economic data for operational and strategic planning, and policy analysis (see **Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts**). Within the medical surge area, RealOpt allows public health emergency coordinators to (re)design

customized, efficient floor plans for each facility via an automatic graph-drawing tool; identify effective triage strategies, and determine (in real-time) optimal resource allocation through advanced computational techniques in simulation and optimization [14–17]. This allows hospital managers the ability to perform scenario-based analysis and determine the best course of action in response to the medical surge requirement.

CHALLENGES

Large-scale public health emergencies may involve thousands of sick or injured people who will require various levels of medical care, ranging from patient evacuation (see **Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation**), hospital care, and sustainable and potentially long-term health-recovery procedures. Thus, such emergencies present a daunting set of challenges, including the surge capacity and capability, and flexibility of our existing medical systems, federal and state emergency capacity for rapid medical dispensing for population protection (see **Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts**), and the resolve and resilience of health-care workers and emergency responders to perform under critical timelines and exceedingly stressful conditions.

Surge capacity encompasses multidimensional aspects such as potential patient beds; available space in which patients may be triaged, managed, vaccinated, decontaminated, or simply located; available care personnel of all types; necessary medications, supplies, and equipment; and even the legal capacity to deliver health care under situations that exceed authorized capacity. Further, close collaboration between public health and health care facilities are very important to effective disaster response. The 2006 Academic Emergency Medicine Consensus Conference discussed the key concepts within the field of surge capacity. Specifically, five top topics were listed as research priorities [32]:

- defining criteria and methods for decision making regarding allocation of scarce resources;
- determining effective triage protocols,
- determining key decision makers for surge-capacity planning and means to evaluate response efficacy (e.g., incident command);
- developing effective communication and information-sharing strategies (situational awareness) for public-health decision support, and
- developing methods and evaluations for meeting workforce needs.

Modeling and optimizing public health infrastructure involve elements of resource allocation under risk, uncertainty, and time pressure; large-scale supply-chain management; transportation and operational logistics; and medical treatment (*medical surge*), population protection (see **Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts**), and effective evacuation (see **Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation**). The operations must be supported by an effective communication infrastructure (see **Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure**). There is a necessity for vertical and horizontal integration and communication, where federal, state, local, tribal, territorial, private and business stakeholders work toward a common goal of a resilient public health system. The infrastructure must be *flexible, scalable, sustainable, and elastic* to support an effective and timely response, and to mount rapid recovery and mitigation operations.

Acknowledgment

Part of the material and results reported herein are based on our work in this area and interaction with public health agencies, and discussion with many state and federal public health and emergency response experts. We also acknowledge the funding from the CDC.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

- Hick JL, Hanfling D, Burstein JL, *et al.* Health care facility and community strategies for patient care surge capacity. *Ann Emerg Med* 2004;44:253–261.
- Okumura T, Takasu N, Ishimatsu S, *et al.* Report on 640 victims of the Tokyo Sarin attack subway. *Ann Emerg Med* 1996;28:129–135.
- Hogan DE, Waeckerle JF, Dire DJ, *et al.* Emergency department impact of the Oklahoma city terrorist bombing. *Ann Emerg Med* Mosby 1999;34:160–167.
- Roccaforte JD. The world trade center attack observations from New York's Bellevue hospital. *Critical Care* 2001;5:307–309.
- Schull MJ, Stukel TA, Vermeulen MJ, *et al.* Surge capacity associated with restrictions on nonurgent hospital utilization and expected admissions during an influenza pandemic: lessons from the Toronto severe acute respiratory syndrome outbreak. *Acad Emerg Med* 2006;13:1228–1231.
- Gavagan TF, Smart K, Palacio H, *et al.* Hurricane Katrina: medical response at the Houston astrodome/reliant center complex. *South Med J* 2006;99:933–939.
- Stratton SJ, Tyler RD. Characteristics of medical surge capacity demand for sudden-impact disasters. *Acad Emerg Med* 2006;13:1193–1197.
- Kelen GD, Kraus CK, McCarthy ML, *et al.* Inpatient disposition classification for the creation of hospital surge capacity: a multiphase study. *Lancet* 2006;368:1984–1990.
- Schultz CH, Koenig KL. State of research in high-consequence hospital surge capacity. *Acad Emerg Med* 2006;13:1153–1156.
- Barbisch DF, Koenig KL. Understanding surge capacity: essential elements. *Acad Emerg Med* 2006;13:1098–1102.
- Schultz CH, Stratton SJ. Improving hospital surge capacity: a new concept for emergency credentialing of volunteers. *Ann Emerg Med* 2007;49:602–609.
- Kaji AH, Koenig KL, Lewis RJ. Current hospital disaster preparedness. *JAMA* 2007;298:2188–2190.
- Qureshi Ketal. Health care workers' ability and willingness to report to duty during catastrophic disasters. *J Urban Health: Bull N Y Acad Med* 2005;82:378–388.
- Lee EK, Maheshwary S, Mason J, *et al.* Large-scale dispensing for emergency response to bioterrorism and infectious disease outbreak. *Interfaces* 2006;36(6):591–607.
- Lee EK, Maheshwary S, Mason J, *et al.* Decision support system for mass dispensing of medications for infectious disease outbreaks and bioterrorist attacks. *Ann Oper Res Comput Optim Med Life Sci* 2006;148:25–53.
- Lee EK, Chen CH, Pietz F, *et al.* Modeling and optimizing the public health infrastructure for emergency response. *Interfaces* 2009;39(5):476–490.
- Lee EK, Smalley HK, Zhang Y, *et al.* Facility location and multi-modality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. *Int J Risk Assess Manage* 2009;12(2/3/4):311–351.
- Kelen GD, McCarthy ML, Kraus CK, *et al.* Creation of surge capacity by early discharge of hospitalized patients at low risk for untoward events. *Disaster Med Public Health Prep* 2009;3:S10–16.
- Hanfling D. Equipment, supplies, and pharmaceuticals: how much might it cost to achieve basic surge capacity? *Acad Emerg Med* 2006;13:1232–1237.
- Burkle JFM. Population-based triage management in response to surge-capacity requirements during a large-scale bioevent disaster. *Acad Emerg Med* 2006;13:1118–1129.
- Peleg K, Kellermann AL. Enhancing hospital surge capacity for mass casualty events. *JAMA* 2009;302:565–567.
- Shih FY, Koenig KL. Improving surge capacity for biotreats: experience from Taiwan. *Acad Emerg Med* 2006;13:1114–1117.
- Manley W, *et al.* A dynamic model to support surge capacity planning in a rural hospital. Conference Proceedings, the 23rd International Conference of the System Dynamics Society Boston, Massachusetts: 2005; p. 22.
- Thompson S, Nunez M, Garfinkel R, *et al.* Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Oper Res* 2009;57(2):261–273.

25. Gerchak Y, Gupta D, Henig M. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manage Sci* 1996;42(3):321–334.
26. Einav S, *et al.* In-hospital resource utilization during multiple casualty incidents. *Ann Surg* 2006;243(4):533–540.
27. Patrick J, Puterman M. Reducing wait time through Operations Research: Optimizing the Use of Surge Capacity. *Healthcare Policy* 2008; 3(3):75–88.
28. Paul JA, George SK, Yi P, Lin L. Transient modeling in simulation of hospital operations for emergency response. *Prehosp Disast Med* 2006; 21(4):223–236.
29. Roccaforte J, Cushman J. Disaster preparedness, triage, and surge capacity for hospital definitive care areas: Optimizing outcomes when demands exceed resources. *Anesthesiology Clinic* 2007; 25(1):161–177.
30. Hoard M, Homer J, Manley W, Furbee P, Hague A, Helmkamp J. Systems modeling in support of evidence-based disaster planning for rural areas. *International Journal of Hygiene and Environmental Health*. 2005; 208(1-2):117–125.
31. Green L. Capacity planning and management in hospitals. *Operations Research and Health Care Chapter 2*. Springer New York: 2005; pp.15–41.
32. Rothman R, Hsu E, Kahn C, Kelen G. Research Priorities for Surge Capacity. *Academic Emergency Medicine* 2006; 13(11):1160–1168.

PUBLIC HEALTH, EMERGENCY RESPONSE, AND MEDICAL PREPAREDNESS II: MEDICAL COUNTERMEASURES DISPENSING AND LARGE-SCALE DISASTER RELIEF EFFORTS

EVA K. LEE

Center for Operations Research in
Medicine and Healthcare, School
of Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta, Georgia

NSF I/UCRC Center for Health
Organization Transformation,
Atlanta, Georgia

School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

FERDINAND PIETZ

Strategic National Stockpile, Office
for Public Health Preparedness
Response, Centers for Disease
Control and Prevention,
Atlanta, Georgia

BERNARD BENECKE

Global Disease Detection and
Emergency Response, Center for
Global Health, Centers for
Disease Control and Prevention,
Atlanta, Georgia

INTRODUCTION

A catastrophic health event, such as a terrorist attack with a biological agent, a naturally occurring pandemic, or a calamitous meteorological or geological event, could cause tens or hundreds of thousands of casualties or more, weaken the economy, damage public morale and confidence, create panic and civil unrest, and threaten national security. It is therefore critical to establish a strategic vision that will enable a level of public health and medical

preparedness sufficient to address a range of possible disasters. Although present public health and medical preparedness plans incorporate the concept of “surging” existing medical and public health capabilities in response to an event that threatens a large number of lives, the assumption that conventional public health and medical systems can function effectively in catastrophic health events has proven to be incorrect in real-world situations. Therefore, it is necessary to transform the approach to health care in the context of a catastrophic health event in order to enable public health and medical systems to respond effectively to a broad range of incidents.

According to the US Department of Homeland Security Presidential Directive 21 2007, public health and medical preparedness refers to “the existence of plans, procedures, policies, training, and equipment necessary to maximize the ability to prevent, respond to, and recover from major events, including efforts that result in the capability to render an appropriate public health and medical response that will mitigate the effects of illness and injury, limit morbidity and mortality to the maximum extent possible, and sustain societal, economic, and political infrastructure.”

Planning for a catastrophe involving a disease outbreak or mass casualties is an ongoing challenge for first responders and emergency managers. They must make critical decisions on treatment distribution points, staffing levels, impacted populations, and potential impact in a compressed window of time when seconds could mean life or death. Although extensive resources have been devoted to planning for a worse-case scenario on the local, regional and national scale, the US Government Accountability Office (GAO) found that gaps still exist. While many states have made progress in planning for mass casualty events, many noted continued concerns related to maintaining adequate staffing levels and accessing other resources necessary for an effective response.

This article highlights our own experience on projects with the Centers for Disease Control and Prevention (CDC) and various public health jurisdictions in emergency response and medical preparedness for mass dispensing for disease prevention and treatment, and large-scale disaster relief efforts. Specifically, the section titled “Population Protection: Medical Countermeasures Dispensing and Large-scale Disaster Relief Efforts” provides a brief introduction and motivation for medical countermeasure dispensing. The section titled “A Systems View of Mass Dispensing Operations: Integrating All the Elements” offers a systems view of mass dispensing operations.

In the section titled “Mode of Dispensing and POD Placement”, we describe various modes of dispensing and POD (point of dispensing) placement. This is followed by resource allocation and POD layout design in the next section. The section titled “Disease Propagation Analysis: Mitigation Strategies and Choice of Dispensing Modalities” highlights the importance of disease propagation analysis and design of mitigation strategies, while the section titled “Supply Chain Management: Demand, Supply, Fulfillment, and Partnership” describes supply chain management. In particular, we briefly explain the mission and responsibility of CDC’s Strategic National Stockpile (SNS). This is followed by the section titled “Communication and Public Information.” This entails the communication infrastructure (hardware tools and software programs) that supports emergency operations, as well as public information and risk communication. The section titled “Lessons Learnt and Continued Challenges” offers insights on lessons learned from flu mass vaccination and anthrax drill exercises, and shares our knowledge on continued challenges and the need for multilayer protection.

The section titled “Large-scale Disaster Relief Efforts” summarizes large-scale disaster relief efforts, including the establishment of a network of service constructs for food, medical needs, and shelters; staffing and resource constraints; susceptibility of displaced population to infectious

disease outbreaks; and supply chain management. Coordination among different stakeholders; risk and uncertainty that are faced by on-the-ground rescue and relief workers; and media exposure are also discussed.

In the section titled “Summary,” we summarize some methodologies that are commonly employed in emergency responses. These techniques include optimization, stochastic processes, information technology, and an integrated framework of decision systems. The article concludes with some challenges for future research.

POPULATION PROTECTION: MEDICAL COUNTERMEASURES DISPENSING AND LARGE-SCALE DISASTER RELIEF EFFORTS

Public-health emergencies such as bioterrorist attacks or pandemics demand fast, efficient, large-scale dispensing of critical medical countermeasures (i.e., vaccines, drugs, and therapeutics). Such dispensing is complex and requires careful planning and coordination from multiple federal, state, and local agencies, and the potential involvement of the private sector. Dispensing medications quickly (within 48 h for anthrax prophylactic) to large population centers (with tens of thousands or even millions of people) is urgent; moreover, the multifaceted nature of dispensing (e.g., sending federal stockpiles to local PODs, coordination at the local level to manage the transportation of citizens to PODs, and the POD operations) makes the process highly unpredictable. Thus, emergency managers and public health administrators must be able to quickly investigate alternative response strategies as an emergency unfolds.

The focus of this article is on *mass dispensing* of medical countermeasures for protection of the general population. Other issues pertinent to large-scale disaster relief efforts will also be highlighted. Much of the writing below are excerpts from our recent work with the Centers for Disease Control and Prevention on modeling and optimizing public health emergency response infrastructure and the development of a

large-scale simulation-optimization decision support system, RealOpt [1–6].

Large-scale public health emergencies may involve thousands of sick or injured people who will require various levels of medical care, ranging from patient evacuation (see ***Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation***), hospital care, and sustainable and potentially long-term health-recovery procedures. Thus, such emergencies present a daunting set of challenges, including the surge capacity and flexibility of our existing medical systems (see ***Public Health, Emergency Response, and Medical Preparedness I: Medical Surge***), federal and state emergency capacity for rapid medical dispatching, and the resolve and resilience of health-care workers and emergency responders to perform under critical time lines and exceedingly stressful conditions.

In the wake of the 2001 anthrax attacks, the Department of Health and Human Services (HHS) increased its order for smallpox vaccine, accelerated production, and began working to develop a detailed plan for the public health response to an outbreak of smallpox. By January 2003, the United States had sufficient quantities of the vaccine for every person in the country in an emergency situation [7]. Subsequently, HHS required each state to submit a mass vaccination plan for administering the smallpox vaccine. Further, states are charged with developing city-readiness programs that deal with establishing regional treatment and dispensing centers, and developing procedures, policies, and a planning framework for efficient allocation of staff and resources in response to these events.

The importance of such population protection has been carefully studied for human, social, and economic benefits. Kaplan *et al.* [8] argued that immediate mass vaccination after a smallpox bioterrorist attack would result in fewer deaths and faster eradication of the potential epidemic; Wein *et al.* [9] concluded that immediate and aggressive dispersion of oral antibiotics and the full use of available resources (local nonemergency care workers, federal and military resources,

and nationwide medical volunteers) are extremely important.

A SYSTEMS VIEW OF MASS DISPENSING OPERATIONS: INTEGRATING ALL THE ELEMENTS

Public health and medical preparedness involves three phases: (i) preparedness and prevention, (ii) detection and response, and (iii) recovery and mitigation.

Modeling and optimizing public health infrastructure involve elements of resource allocation under risk, uncertainty, and time pressure; large-scale supply chain management; transportation and operational logistics; and medical treatment and population protection. The operations must be supported by an effective communication infrastructure (see ***Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure***). There is a necessity for vertical and horizontal integration and communication, where federal, state, local, tribal, territorial, private, and business stakeholders work toward a common goal of a resilient public health system. The infrastructure must be *flexible, scalable, sustainable, and elastic* to support an effective and timely response, and to mount rapid recovery and mitigation operations [5,6,10].

The global integration plan for population protection in the event of medical and emergency response includes multiple interconnected components: strategic planning; management command and control with key leaders working cohesively together as a team; requesting supplies and equipment from Strategic National Stockpile (SNS) tactical communication and information technology; public information and risk communication; security; regional and local distribution sites (for receiving, staging, and storing of supplies, as well as transportation and routing); inventory control and management; distribution supply and resupply; dispensing; treatment centers; and planning, training, and evaluation [10]. The ultimate goal is to *dispense* the medical countermeasures to the affected regional population in a timely manner.

Mode of Dispensing and POD Placement

Mass dispensing requires the rapid establishment of a network of dispensing sites and health facilities that are *flexible*, *scalable*, and *sustainable* for medical prophylaxis and treatment of the general population. Moreover, each POD must be capable of serving the affected local population within a specified short time frame (Fig. 1). Clearly, for very large-scale dispensing, the sophisticated logistical expertise needed to deal with the complexities of selecting an adequate number of strategically well-placed POD locations, and of designing and staffing each POD, is beyond the capability of any human planner or public health administrator. The limited availability of trained critical staff, such as public health professionals, further compounds the inherent complexities.

The CDC and public health administrators work closely with one another to prepare for and document the steps required to administer medication in the event that mass dispensing is needed. The goal and objectives of a dispensing facility, *POD*, are to deliver appropriate emergency services (e.g., vaccine, medical service, and education/training) to high-risk populations in an orderly, expeditious, and safe manner. Within the POD facility, the potential tasks and objectives may include the following:

1. Assess health status of clients.
2. Assess eligibility of clients to receive service.
3. Assess implications of each case and refer case for further investigation if necessary.
4. Counsel clients regarding service and associated risks.
5. Administer service.
6. Educate regarding adverse events.
7. Document services.
8. Monitor vaccine take rates.
9. Monitor adverse reactions.
10. Monitor development of disease.

Mode of Dispensing. The *key* to mass dispensing is to protect the general population efficiently and effectively under time pressure. For example, in an anthrax attack, the goal is for citizens to receive antibiotic prophylaxis within 48 h of the determination that an attack has occurred, as the mortality rate for persons demonstrating symptoms of inhalation anthrax is extremely high. (Lawler JV, Mecher CE. Homeland Security Council. 2007. Private Communication.) Thus, it is recognized that multiple dispensing modalities often must be employed in order to *serve (cover)* the entire regional population. For example, special dispensing services will be utilized to serve homebound, disabled, and special-need populations. In

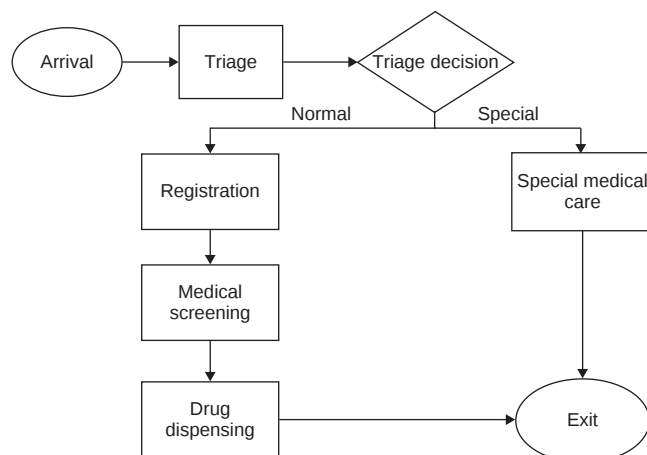


Figure 1. The flowchart shows a POD that was set up in a national drill exercise to dispense anthrax antibiotics.

some instances, it is unreasonable to expect residents to travel to a designated POD facility. For example, nursing homes, assisted living facilities, homeless shelters, hospitals, and prisons house many residents for whom it would be inconvenient or inadvisable to travel to a public dispensing facility. Moreover, in many of these instances, there are already medical personnel on site who can assist in the dispensing process. In such cases, it would be more efficient to set up a *closed* POD inside these locations for dispensing, or to have medication dispensed by a mobile POD facility near the site. In this case, the POD is *closed* as it provides services only to the residents on-site, and is not open to walk-ins from the general public. Corporate offices that staff a large number of employees could be served in a similar manner. Once these sites receive prophylactic supplies, they could set up a closed POD within their building, with their own health-care staff and volunteers, or with public health staff supplemented by the state. Several factors suggest that such closed PODs will have fewer security concerns and will be easier to manage than public PODs. These factors include familiarity with the environment and people (e.g., fellow residents/employees), existing security measures including established checkpoints and previously authenticated identification badges with photo and/or biometric markers, and less stress than having to commute to a public POD.

Airports and hotels, where a large number of nonresident travelers can be found, are also candidates for setting up PODs to service specific vulnerable populations. Universities can use their own health facilities (and if necessary, additional mobile on-campus PODs provided by the state) to provide prophylaxis to on-campus students, staff, and faculty. Clearly, if large employers and medical facilities provide prophylaxis to their own employees, families, and patients, it will eliminate a high percentage of the population (may be as high as 40% in some large cities) from visiting public PODs, thus reducing the load on those facilities.

Public PODs are *open* facilities that are setup to serve the general public.

In our work with CDC [2,3,5,6], public PODs have generally been described as being setup inside existing facilities, or in outdoor tents, with areas set aside for various activities in the dispensing process, including assembly/intake, triage, orientation, registration, screening, service, education, and discharge. Public PODs can be mobile or stationary, and in the latter case, they can be setup as *facility-based* or *drive-through*.

A facility-based POD operates within physical locations, such as buildings, warehouses, open fields, or large parking lots. Citizens are asked to arrive at the POD location and then *walk through* the POD to receive their medication or other treatment. Facility-based PODs may be scaled to operate within a setting as large as a professional sports stadium or as small as a volunteer fire house within a rural community.

These walk-through PODs are suitable for relatively large capacity facilities, where the possibility of traffic jams preclude the use of the drive-throughs. Because parking is typically limited, individuals may be directed to arrive at designated points, and they are taken by bus to the POD. Examples of pickup points include bus stations, subway stations, and parking lots of large shopping malls, where sufficient parking is available.

High schools are often selected as potential POD locations. The fact that they are government-owned makes logistics easier. Other considerations include existence of offices, computer communications, cafeterias, and storage. Shopping malls, churches, and stadiums are also suitable and in some cases, PODs are set up outdoors using tents and temporary constructs.

Drive-through PODs are suitable to serve a spread-out population. Ideally, government-owned properties are preferred over privately owned ones for logistic reasons. Locations for setting up drive-through facilities should have enough space for multiple dispensing lanes; surrounding access roads; and room for command tent, employee rest area, and medication storage.

POD (Facility) Placement. Facility location problems are classical optimization problems, and have been a critical element in

strategic planning for a wide range of private and public organizations. The earliest facility location problems incorporating emergency response relate to location of emergency facilities [11–13]. Chaiken and Larson [14] provided a survey on urban emergency unit allocation. Some researchers [15–17] explicitly take the need for backup facilities (in case the main facility is overloaded) into account. More recently, Jia *et al.* [18] provided a modeling framework for location of medical services for large-scale emergencies. Berman and Gavious [19] took a game theoretic approach toward location of terror response facilities. Church *et al.* [20,21] studied the problem of identifying and protecting critical infrastructure. A comprehensive review of facility location research can be found in Brandeau and Chiu [22], where the authors presented a survey of over 50 representative problems in location research. Most of the problems reviewed have been formulated as optimization problems. Owen and Daskin [23] provided another review of the strategic facility location problem; they considered a wide range of model formulations across numerous industries, including stochastic formulations, and discussed solution approaches.

While many of the facility location problems involve permanent construction and establishment of service facilities, facility location problem for mass dispensing concerns the placement of PODs (mobile or stationary) in existing facilities (e.g., high schools, stadiums, shopping malls, open parking lots) in a region to provide the necessary services to the population within a designated period of time.

Various objectives can be incorporated within the models. In the event of catastrophic incidents, it is critical that PODs are strategically located so as to allow easy access by the affected public. Hence, minimizing transportation time can be one critical objective. Further, the setup and operating costs of PODs cannot be neglected. A POD must be accessible by service workers, it should include a good communication infrastructure, it must be easily protected by law-enforcement personnel, and the facility must be capable of handling a large flow of

people. Physical constraints on the facility must be modeled properly, for example, capacity of a facility cannot be violated (POD parking capacity is limited and fire codes limit the number of individuals who can be inside a facility simultaneously).

For operational purposes, there is also a desire to ensure that the number of PODs in each jurisdiction is at least two. This is due to the concern that if a catastrophic event at one site necessitates shutting down of a POD, emergency dispensing can still be carried out in the remaining location. In this case, the response manager can reroute populations to the remaining site, while reestablishing a new POD, if deemed necessary.

The problem of modeling POD placement involves the traditional facility location issues, as well as the incorporation of spatial, geographical, and demographic information. In our work with CDC, the problem is modeled using a two-stage integer programming approach. For a large metropolitan area with a population over 5 million, such a POD placement problem can result in optimization (integer programming) instances involving millions of variables and constraints for each jurisdiction that consists of hundreds of thousands of people in the region. The challenge is to determine the trade-offs between the quality of solution, the practicality of planning, and real-time optimization. Exact algorithms and heuristics must be developed and advanced to address such computational challenges [5,6].

Resource Allocation and POD Layout Design

Resource Allocation. Mason and Washington [24] at CDC investigated optimal staffing arrangements for dispensing sites in the face of limited resources via a simulation/optimization system “Maxi-Vac” that they developed. Their study offered insights on the practicality of such a system as a planning tool for emergency situations but revealed critical bottlenecks between the commercial simulation software and the optimization software: over 10 h were needed to obtain a usable feasible solution in each scenario with about 25–30 staff. This initiated our collaboration with CDC and the development of RealOpt, a large-scale

real-time simulation-optimization decision support system [1–3].

Given a staff assignment (obtained from an initial optimization step), and input of service distributions at each station, we can model and simulate the movement of individuals inside a POD. The simulation output is a set of parameters (including statistics of average flow time, queue length, wait time, and utilization rate) that enable evaluation of the objective function being optimized (e.g., the resulting throughput).

The optimization of labor resources involves placement of staff at various stations in the POD to maximize throughput or minimize the staffing needs to satisfy a preset throughput population. The cost at each station depends on the type and number of workers who are assigned to that station and have the required skills, and on the average wait time, queue length, and utilization rate of the station. The total system cost depends on the cost at each station and on system parameters, such as cycle time and throughput. These cost functions are not necessarily expressible in closed form.

Constraints in the model include maximum limits on average wait time and queue length, range of utilization desired at each station, and upper and lower bounds on the number of workers with the required skills who are needed to perform various tasks at the POD. Constraining the average cycle time to be less than a prespecified upper bound is critical for emergency response, because individuals must move through the system as quickly as possible to facilitate crowd control, reduce sources of human frustration and potential disorderly outbursts, and reduce the potential spread of disease or contamination. The resulting nonlinear mixed-integer program poses unique challenges for existing optimization engines.

POD Layout Design. Designing the appropriate POD for various medical dispensing is critical. Further, POD layout will affect the overall staffing and efficiency of the dispensing operations.

Figure 2 contrasts two POD designs that were employed in a hepatitis A booster shot

event for 10,000 citizens [25]. The left shows the drive-through POD design used in the morning, and the right shows the redesign in the afternoon after real-time reconfiguration (based on service times collected on site) was performed. The redesign offers 10% improved throughput, 18% improved utilization, and a range of 10–85% reduction in wait time and queue length at various stations. *This illustrates the paramount importance of POD design to any emergency operation where resources are scarce, time is precious, and there is a large affected population to serve.*

Disease Propagation Analysis: Mitigation Strategies and Choice of Dispensing Modalities

Large-scale dispensing clinics could facilitate the spread of disease because of their high-volume population flow. The field of dynamical systems (mostly differential equation systems) provides the principle methods of modeling in classical mathematical epidemiology [26,27]. Despite their simplicity when compared to recent complex simulation studies [28–31], these methods have helped generate functional insights, such as the transmission threshold for the start of an epidemic and the vaccination threshold for containment of an outbreak. As modelers attempt to incorporate more realistic dynamics into their models (such as stochasticity, nonexponential waiting times, sample-path dependent events, and demographical and geographical data), more flexible tools such as individual-based stochastic simulations are preferable. Although simulation is a powerful approach, it is less mathematically tractable (i.e., it requires intensive computing time) than the classical methods.

The rapid and large-scale simulator in RealOpt opens up an opportunity to explore disease propagation studies in which stochasticity of systems can be incorporated readily. It includes a disease propagation module that aids users in understanding facility design and flow strategies that mitigate the spread of disease. The module incorporates the standard four-stage SEIR (susceptible, exposed, infectious, and recovered) model [32], and a novel six-stage SEPAIR model to capture the disease development (i.e., asymptomatic or

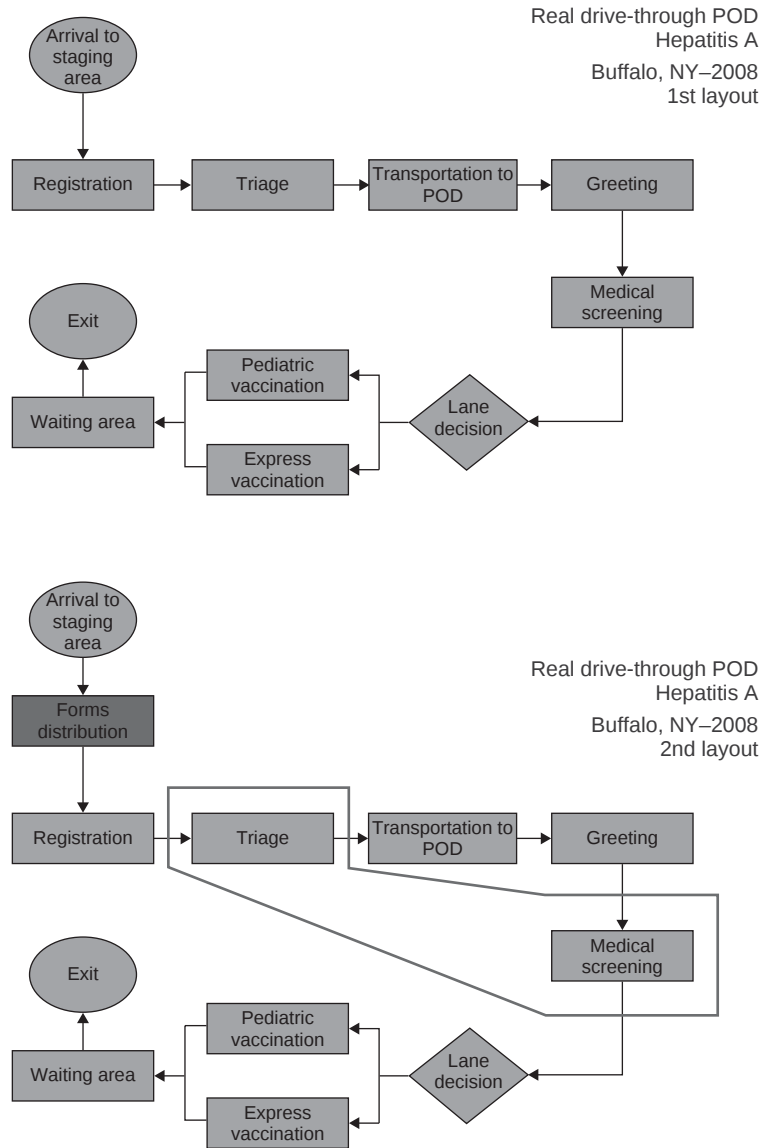


Figure 2. Two POD layouts used during an actual 2008 hepatitis vaccination event.

symptomatic). By distinguishing the symptomatic stage from the asymptomatic stage, this model allows one to examine the effect of triage accuracy in POD facility design.

Lee *et al.* [6,33] gave a detailed theoretical and computational analysis of disease propagation and strategies for mitigation during biological or pandemic outbreaks and mass dispensing. In addition to the incorporation

of stochasticity of client arrival and service distribution into the model, it also accommodates the following factors:

- The clinic model can be represented as an n -server system with queueing; transmission can occur between clients or between clients and staff. (In a real emergency, staff members will be given

medical countermeasures to protect them from the disease prior to their assignment to POD services. However, a medical countermeasure does not provide 100% protection; each staff member still has a small probability of being infected by clients.)

- The intraclinic infectivity between clients and staff can vary.
- If symptomatic individuals are not triaged out properly during the initial screening, they could infect other people inside the POD. The system allows users to observe the effect of triage and screening errors, determine improved strategies for triage and screening, and establish guidelines for mitigating the spread of disease because of such errors.
- Inhomogeneous mixing within the community is possible.
- The infectious, asymptomatic, and symptomatic individuals can infect at various rates.

Figure 3 contrasts the triage accuracy with respect to the symptomatic proportion, when simple mass-action incidence infection is considered [6,33]. This analysis assesses errors in triage and their infection consequences. It provides estimates for POD planners and epidemiologists to help determine the level and expertise of triage that should be in place with respect to the transmission coefficient. It also allows for scenario-based comparison of effective POD design.

Such analyses may influence the selection of dispensing modalities. Specifically, over the past few years, we have observed more use of drive-through PODs for infectious disease prophylactic dispensing (e.g., seasonal flu vaccination for communities and the H1N1 mass vaccination in 2009 and early 2010). In January 2008, a hepatitis A confirmation of a grocery worker triggered the prophylactic vaccination of 10,000 residents in Erie County in New York who were potentially exposed to the disease, costing the county's public health agency at least \$500,000. The health department dispensed the first vaccination in February when it set

up a stationary clinic (walk-through POD). Because of the medical logistics and infectious nature of the disease, some people had to wait for hours in frigid temperatures. In September 2008, the health department used a hepatitis A follow-up drive-through POD to provide the required second shot of vaccination. The POD was also the first test of the county's "drive-through" plan. The drive-through process is quick, efficient, and convenient, and minimizes the potential of intrainfectivity [25].

Supply Chain Management: Demand, Supply, Fulfillment, and Partnership

Although supply chain management during a disaster response mirrors that of business supply chain management, damaged or destroyed infrastructure force the use of *ad hoc* solutions that limit the effectiveness and efficiency of the operation.

Demand management can be extremely difficult due to the fluidity of the population and the potential collapse of the supporting infrastructure. Problems vary depending on the nature of the disaster. In the event of an infectious disease outbreak, as in the recent H1N1 event, when there is not a sufficient supply of medical countermeasures for the affected populations, predicting and allocating the proper distribution across the nation becomes critical, but is highly stochastic and uncertain. Decisions can have a major impact on overall infectivity and mortality rates. In the same manner, demand within an earthquake zone fluctuates rapidly due to the movement of people fleeing from one site to another. And the time it takes to implement an effective response can be critical to the survivors of the affected population.

Within the United States, an act of terrorism (or a large-scale natural disaster) targeting the US civilian population will require rapid access to large quantities of pharmaceuticals and medical supplies. Such quantities may not be readily available unless special stockpiles are created. No one can anticipate exactly where a terrorist will strike and few state or local governments have the resources to create sufficient stockpiles on their own. Therefore, a national

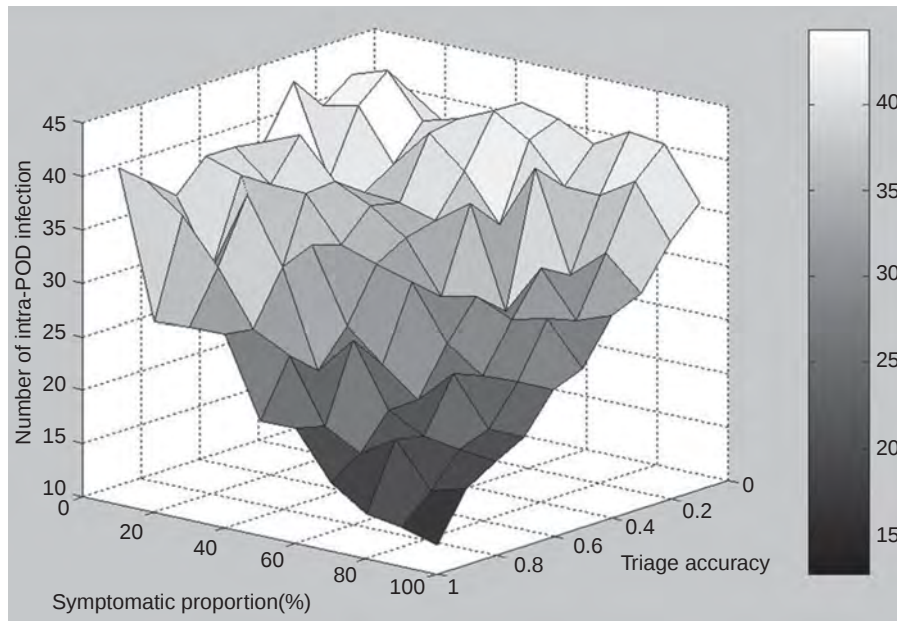


Figure 3. The graph shows the triage accuracy versus symptomatic proportion and the importance of using the SEPAIR six-stage propagation model, because it allows us to examine the effect of implementing triage accuracy. The graph shows the number of intra-POD infections (vertical axis) under different triage accuracy and symptomatic proportions (horizontal axes). The throughput is 36,000 over a period of 36 h. The contact number is 193 (for outer-POD disease propagation), and the transmission coefficient is $0.18E^{-5}/\text{min}$. The incoming percentage for susceptible is 95%, and for infectious is 5%. The mean dwell time is one day for both exposed and infectious and three days for asymptomatic and symptomatic.

stockpile has been created as a resource for all.

The CDC's SNS is a national repository of antibiotics, chemical antidotes, antitoxins, life-support medications, IV administration, airway maintenance supplies, and medical/surgical items to protect the American public if there is a public health emergency (terrorist attack, flu outbreak, earthquake) severe enough to cause local supplies to run out. Once Federal and local authorities agree that the SNS is needed, the SNS will supplement and resupply state and local public health agencies anywhere and at anytime within the United States or its territories.

The SNS is organized for flexible response. The first line of support lies within the immediate response 12-h Push Packages. These are caches of pharmaceuticals, antidotes, and medical supplies designed to provide rapid delivery of a broad spectrum of assets for

an ill-defined threat in the early hours of an event. These Push Packages are positioned in strategically located, secure warehouses ready for immediate deployment to a designated site within 12 h of the federal decision to deploy SNS assets. These 12-h Push Packages have been configured to be immediately loaded onto either trucks or commercial cargo aircraft for the most rapid transportation. Concurrent to SNS transport, the SNS Program will deploy its Stockpile Service Advance Group (SSAG). The SSAG staff will coordinate with state and local officials so that the SNS assets can be efficiently received and distributed upon arrival at the site.

If the incident requires additional pharmaceuticals and/or medical supplies, follow-on vendor managed inventory (VMI) supplies will be shipped to arrive within 24 to 36 h. If the agent is well defined, VMI can be tailored to provide pharmaceuticals,

supplies and/or products specific to the suspected or confirmed agent(s). In this case, the VMI could act as the first option for immediate response from the SNS Program.

The SNS Program works with governmental and nongovernmental partners to upgrade the nation's public health capacity to respond to a national emergency. Critical to the success of this initiative is ensuring capacity is developed at federal, state, and local levels to receive, stage, and dispense SNS assets.

In our work on modeling and optimization of public health infrastructure and mass dispensing strategies [5,6], we describe the importance of collaboration among federal, state, local, and private sectors for successful response to large-scale emergency scenarios, and report public/private solutions for meeting the SNS dispensing requirements as prescribed within the Cities Readiness Initiative program.

Communication and Public Information

Communication and Information Technology. Communication programs and infrastructure for emergency response are discussed in detail in the article titled *Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure*. Communication efforts should be coordinated, planned, and tested. Regular training on usage should be performed.

Briefly, multifaceted communication infrastructures are needed for effective medical and emergency response. Two-way communication lines must be kept open between emergency/medical responders and commanders for coordination and facilitation of the response effort; broadcast communication lines must be available to inform the general public about medical precautions and available services; and phone service and call answering stations must be maintained to allow the public to report critical emergency situations. Further, communication infrastructure of hospitals and related medical care entities are critical for any emergency response.

Developing and maintaining a robust communication infrastructure among hospitals, emergency medical services, and other health care facilities, as well as private medical transport companies and medical supplies vendors, is vital to providing relief operations and services during emergency situations. Such infrastructure provides the core foundation for basic knowledge sharing such as patient volume and severity, emergency room capacity, hospital bed availability; special wards availability; medical personnel specialties; transport vehicle availability; and blood, medicine, and medical supplies inventories. During an emergency situation, such critical information can be communicated to a regional coordinating control center which can assess available resources, identify surpluses and shortages, and coordinate distribution efforts. In fact, building capacity for an interoperable communication system for emergency response is identified as one of the areas that must be prioritized according to The Hospital Preparedness Program established by the US HHS.

Previous instances of emergency situations such as the September 11 terrorist incident and the Rhode Island night club fire incident have brought forth the challenges and necessity in establishing and sustaining a redundant emergency communication network. Kapucu [34] has examined the importance and challenges in establishing a coordinated interagency communication network and the role information technology serves to enhance such communication and decision making as in the context of the 9/11 terrorist incident. Challenges include interoperability among the various communication tools, reliability of the tools during emergency situations, technical limitations, barriers in interagency communication, and related aspects.

Public Information and Risk Communication. Risk communication plays a crucial role to any successful emergency response. The social challenges that arise as a result of human behavioral patterns cannot be overemphasized. During the high concern and high stress situation, the messages to the public must be concise, unified, and

coordinated. Getting correct and credible information out quickly assures the public that they can trust the authority in protecting themselves and their families. The public needs to know in a timely manner that there is a problem and the nature of the problem. Appropriate measures must be performed to curtail rumors that may cause unnecessary panic and fear, and potential unrest. Multimedia hot lines should be established to share factual information and timely updates. Educational information pertinent to the medical countermeasures and dispensing sites should be clearly stated. Finally, care must be taken to ensure that vulnerable and special-need populations are served appropriately.

Lessons Learnt and Continued Challenges

Our Experience through Anthrax Drills and Actual Vaccination Events. Our experience illustrates that by combining mathematical modeling, large-scale simulation, and powerful optimization engines and coupling these with automatic graph-drawing tools and a user-friendly interface, we can design and implement a fast and practical emergency response decision support tool that can run on a wide range of computing platforms, including personal digital assistants (PDAs) for real-time usage. The system RealOpt, developed in collaboration with CDC SNS investigators, offers public health emergency coordinators the following capabilities:

1. Determine strategically most effective locations for POD facilities to best serve the affected population.
2. Design customized and efficient POD floor plans via an automatic graph-drawing tool. Users can design and compare various floor plans to determine the trade-offs in personnel usage as well as operations efficiency.
3. Determine optimal labor resources required and provide the most-efficient placement of staff at individual stations within the POD. The resulting staffing plans maximize the number of individuals who can be treated, minimize the average time patients

spend in the clinic, and equalize utilization across clinic stations.

4. Perform disease propagation analysis, understand and monitor the intra-POD disease dilemma, and help to derive dynamic response strategies to mitigate casualties. The what-if analysis and worst-case scenarios can also equip epidemiologists with knowledge that may assist emergency planners in testing alternative POD facility layouts, assess batch size for patient orientation, and analyze the trade-offs between POD throughput operations and degree of infectiousness.
5. Assess current resources and determine minimum needs to prepare for readiness in emergency situations for their regional population.
6. Carry out large-scale virtual drills and performance analyses, and investigate alternative strategies.
7. Train personnel, and design emergency exercises with a variety of dispensing scenarios. Such training exercises could be used to quickly get new emergency preparedness planners up to speed and to keep existing planners sharp.

The computational advances also provide flexibility to quickly analyze design strategies and decisions, and can generate a feasible regional dispensing plan (with a network of cost-effective PODs, each operating at various throughput rates and utilizing various dispensing modalities) based on the best estimates and analyses available, and then allow for reconfiguration of various PODs as the event unfolds.

Some of the regional planning and operational analysis reveal that

1. sharing labor resources across counties and districts within the same jurisdiction is important;
2. the most cost-effective dispensing plan across a region consists of a combination of drive-through, walk-through, and closed PODs, each operating at a throughput rate that depends on the surrounding population density, facility type, and labor availability;

3. the optimal combination of POD modalities changes according to various facility capacity restrictions and the availability of critical public health personnel;
4. an increase in the number of PODs in operation does not necessarily increase the total number of core public health personnel needed;
5. optimal staffing is nonlinear with respect to throughput, thus we cannot estimate the optimal staffing and throughput by simply using an average estimate;
6. depending on the population, an “optimal” capacity that provides the most effective staffing exists for each POD location. If a POD is operating above its optimal capacity, reduction in capacity (and thus hourly throughput) eases the crowd-control tasks of law-enforcement personnel and helps to minimize potential operational problems inside the POD.

RealOpt has been used successfully by over 2000 public health and emergency directors and coordinators in planning for biodefense drills (e.g., anthrax, smallpox) and pandemic response events in various locations in the United States since 2005. It has also been used for dispensing clinic design and staff allocation for the recent H1N1 mass vaccination campaign. Users have tested various POD layouts, including drive-through, walk-through, and closed PODs. Because of the system’s rapid speed, it facilitates analysis of what-if scenarios, and serves not only as a decision tool for strategic and operational planning, actual drill preparation, and personnel training, but also allows dynamic reconfigurations as an emergency event unfolds. In addition, it supports performing “virtual field exercises,” offering insight into operation flows and bottlenecks when mass dispensing is required.

Continued Challenges and the Need for Multilayered Protection. The ability to analyze planning strategies, compare the various options, and determine the most cost-effective combination dispensing strategy

are critical to the ultimate success of mass dispensing.

The federal government continues to seek advances in this challenging area. In the recent Executive Order, signed on December 30, 2009 by the President of the United States, a policy was setup to plan and prepare for the timely provision of medical countermeasures to the American people in the event of a biological attack in the United States through a rapid federal response in coordination with state, local, territorial, and tribal governments. The policy seeks to (i) mitigate illness and prevent death; (ii) sustain critical infrastructure; and (iii) complement and supplement state, local, territorial, and tribal government medical countermeasure distribution capacity.

Specifically, the executive order stipulates that the federal government shall pursue the establishment of a national US Postal Service medical countermeasures dispensing model to respond to a large-scale biological attack with anthrax as the primary threat consideration. Further the federal government must develop the capacity to anticipate and immediately supplement the capabilities of affected jurisdictions to rapidly distribute medical countermeasures following a biological attack by establishment of a federal rapid response capability. The executive order also asks for Continuity of Operations where the federal government must establish mechanisms for the provision of medical countermeasures to personnel performing mission essential functions to ensure that mission essential functions of federal agencies continue to be performed following a biological attack.

Postal delivery of medical countermeasures has been discussed in detail in Wein [35]. Because postal workers deliver mail service to all households on a regular basis, the possibility of their usage for the first 12-h initial dispensing is appealing. This will allow time for the establishment of PODs for continued dispensing service. The availability of personnel during crisis can be volatile, as articulated in medical surge article (see *Public Health, Emergency Response, and Medical Preparedness I: Medical Surge*). However, such a multilayered dispensing

plan allows flexibility in response as well as leveraging and integration of heterogeneous resources for maximum coverage and outcome. Postal and POD distribution strategies require different levels of security; their trade-offs and complimentary characteristics should be carefully analyzed.

LARGE-SCALE DISASTER RELIEF EFFORTS

Large-scale disaster (humanitarian) relief efforts (e.g., in response to earthquakes, hurricanes, forest fires) where homes are destroyed, critical infrastructures are damaged, and tens of thousands or millions of people's lives are affected, require rapid establishment of "*service constructs*." These service constructs serve as home shelters for the population being displaced; as distribution nodes for receiving supplies for on-the-ground responders; as dispensing sites for handing out food and water to the affected population; and as hospital tents for medical care of the sick. In the aftermath of such an event, the medical surge requirement is acute (see ***Public Health, Emergency Response, and Medical Preparedness I: Medical Surge***), evacuation orders are highly probable (see ***Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation***), and the need for effective communication and coordination for humanitarian relief effort is crucial (see ***Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure***).

We highlight below various issues that are prominent within disaster relief efforts.

Network of Service Constructs, Staffing and Resource Constraints, Opportunistic Disease Spread, and Supply Chain Management

The scope and complexity of establishing service constructs (for food and water, shelters, and medical care etc.) are very similar to those for large-scale mass dispensing efforts (as described in previous sections in this article). Past disaster relief efforts such as in the aftermath of the 2010 earthquake in Haiti highlight the daunting challenges as

the regional response must rapidly establish (in an *ad hoc* manner) dispensing and distribution networks. Many elements discussed in the section titled "A Systems View of Mass Dispensing Operations: Integrating All the Elements" on mass dispensing and population protection come into play. Further, the lack of water, electricity, shelters, poor sanitation, and the existence of at most a barebones critical infrastructure present enormous challenges.

Staffing and resources are often under severe shortage where *ad hoc* skills and just-in-time training are provided. Skilled workers such as those with medical background, logistics and operational, and security experiences may often be deployed from other countries and nonprofit organizations to help with rapid response, relief, and rebuild mission.

Crowded, unsanitary conditions in improvised refugee shelters could spread illnesses such as typhoid and measles. Thus, mass prophylactic treatments are often needed to prevent disease spread, even though illness and infections can still be a threat to survivors. Puddles of filthy water accumulated near service constructs may become a breeding ground for mosquitoes. That, in turn, may lead to the spread of deadly diseases and epidemics such as dengue, malaria, and cholera. Disease mitigation and facility layout strategies are absolute critical tasks to the livelihood of these survivors.

In disasters, suppliers and stakeholders can be very diverse, and there is usually no unifying business and operating theme to manage them. Further, unsolicited and unwanted donations can burden the management team, causing overload of arrivals at sea or airports, or congesting the warehouses [36–38].

Routing of available and necessary goods from entry ports (sea or air) to affected sites can pose daunting challenges. Efficient usage of potentially limited sea/air space and landing sites, available roads, vehicles, fuels, drivers, and materiel handling equipment at the receiving ends are all uncertain and unreliable. Food safety, sanitation hygiene, and opportunistic looting further complicate the process.

Effective coordination of multiple humanitarian agencies, in addition to military, government, and private entities, is of great importance in response to disasters. This is both a challenge and an opportunity given the differing missions, histories, and expertise of these institutions.

Coordination among Different Stakeholders

Disaster and humanitarian operations often have a large number and variety of stakeholders and multiple organizations operating in the same place simultaneously but without (formal) coordination. A loosely coupled coordination of different aid agencies, suppliers, and local and regional actors, each with their own way of operating and own organizational structures, can work charmingly yet it can also pose discord [39]. The different political agendas, ideologies, and religious beliefs and the need for appeal to public for donations and media attention complicate the work. Perhaps, the greatest challenge here lies in aligning these organizations properly without compromising their mandates and beliefs. Further, people from different cultural background may have different traditions that may hinder the communication and coordination between organizations [40]. As seen in the recent Haiti event, operational and organizational differences can also create frictions and misunderstanding among various responding nations, thus masking the effectiveness of the operations.

Sometimes, affected areas may not be reachable due to political reasons. For example, after the 2005 South Asian Tsunami, the Indonesian government felt compelled to allow free entry in a region that had been very restricted for a long time. This caused a huge influx of personnel from humanitarian organizations, *ad hoc* organizations and volunteers on sites, overcrowding the sites.

Risk and Uncertainty

Personnel working within a disaster response environment are often exposed to destabilized infrastructure [37,38]. Not only do they need to work in facilities or areas that are physically damaged, the

social effect taking place after a disaster could often become overwhelming. Many disasters such as tsunamis and earthquakes have aftereffects that could further cause panic or disruption to response operations. Uncertainty in risk, in when the suppliers will arrive and where, in the amount of supplies that are available, in the overwhelming demand needed by the affected population, add layers of anxiety and stress to these on-the-ground workers. “Disaster relief supply chain” shows the extremes of a trend toward more uncertainty and risk prevalent in today’s global business supply chain.

Media Exposure

Disaster events often have high exposure to the media. However their relationship is described by some as a love-hate one [40]. On the positive side, high exposure to the media means more public attention on the affected areas. This often translates into more donations and general support. However, mass media is short-sighted and tends to be more interested in catastrophic events while putting less focus on long-term humanitarian and relief effort. There is a need to educate and collaborate with news media personnel on disseminating long-term challenges and efforts to the general public. In recent years, some improvement has been seen as news media and public have been made aware of response failures in the emergency responses related to Hurricane Katrina and other natural disasters, and the high-profile public scrutiny of the federal and local response effort. In general, appropriate media usage can educate the public and help spread word of the situation so as to garner financial donations that are much needed in any disaster response scenario.

SUMMARY

Optimization, Simulation and Dynamical Systems Methodologies

Operations research, with its roots in defense and military operations, has a natural place in emergency response planning and execution. Optimization, stochastic simulation,

systems modeling, and decision analysis approaches are routinely used to plan for and/or aid in analyzing a broad spectrum of emergency responses as a result of natural or man-made disaster [2,6,12,41,42]. Specifically, Green and Kolesar [42] traced the history of operations research and management science applications in emergency response, with a particular focus on the work done in New York City between 1969 and 1989. Many projects were undertaken by a group of researchers as part of the New York City-RAND Institute (NYCRI) initiative. These included applications to ambulance, fire, and police car location and deployment.

Resource allocation, scheduling, facility location, vehicle routing, inventory control, and transportation logistics in emergency response have all been formulated into optimization models. Among the resource allocation models, Fiedrich *et al.* [43] used dynamic optimization model for the initial search-and-rescue period after a strong earthquake. Branas *et al.* [44] used a trauma resource allocation model for ambulances and hospitals. Tzeng *et al.* [45] presented a multiobjective optimal planning model for designing relief delivery systems. The three goals are minimizing the total cost, minimizing the total travel time, and maximizing the minimal satisfaction during the planning period. Yan and Shih [46] used a time-space network model (based on an integer network flow problem with side constraints) to minimize the length of time needed for emergency repair, with related operating constraints for emergency repair work team scheduling.

Lee *et al.* [2,5] investigated resource allocation, staffing, facility location, and multimodality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. Some of the integer programming instances have on the order of 10 million variables and the authors provided rapid solution engines to arrive within 5% to optimality under 3 min. Zhang and Yang [47] presented an optimization model and algorithm of a facility location problem in a perishable commodities emergency system. Doerner *et al.* [48] presented a model for multiobjective decision analysis with respect to

the location of public facilities such as schools in areas near coasts, taking risk of inundation by tsunamis into account. Coskun and Erol [49] used an integer optimization model to decide locations and types of service stations, and regions covered by these stations under service constraints in order to minimize the total cost of the overall system. The model can produce optimal solutions within a reasonable time for large cities having up to 130 districts or regions.

Many authors present routing and transportation studies: Harewood [50] used a multiobjective version of the maximum availability location problem to determine emergency ambulance deployment. Wang *et al.* [51] analyzed and proposed the concept of postearthquake road safety, dividing emergency vehicle routing choice and optimization problem into two decision-making stages: pretrip and en route. Sheu [52] used a hybrid fuzzy clustering-optimization approach to study the operation of emergency logistics codistribution when responding to urgent relief demands in the crucial rescue period. Yi and Kumar [53] presented a metaheuristic of ant colony optimization for solving a logistics problem arising in disaster relief activities. Yuan and Wang [54] presented two mathematical models for path selection in emergency logistics management. The models include actual factors in time of disaster. Liu and Zhao [55] modeled an emergency materials distribution problem in an antibioterrorism system as a multiple traveling salesman problem.

Simulation has been used in numerous public health and medical preparedness topics, including evacuation, resource allocation and patient flow, routing in emergency medical services and surge planning, and disease propagation analysis.

Simulation has also been used in resource allocation and patient flow [2,3,24,56–60], and in the area of pandemic response strategies and mitigation [2,3,5,6,33,56,58,61–66]. It has found many applications in routing in emergency medical services and surge planning [62,67–69].

As modelers attempt to incorporate more realistic dynamics into epidemiology and disease propagation models (such as

stochasticity, nonexponential waiting times, sample-path dependent events, and demographical and geographical data), more flexible tools, such as individual-based stochastic simulations, are preferable. Although simulation is a powerful approach, the resulting models are often mathematically intractable, and require advances in computational strategies [6,28–31,33,70–72].

The discipline of dynamical systems, mostly differential equation systems, provides the principle methods of modeling in classical mathematical epidemiology [26,27]. It has also been used to analyze strategies and policies. Wein *et al.* [9] used a system of differential equations to model the effects of different policies in response to an anthrax bioterror attack. The system includes an atmospheric dispersion model for the spread of the bacterium causing anthrax, an age-dependent dose-response model for the impact of treatment on an individual, a disease progression model to capture the stages through which an infected individual goes, and a set of two-stage queueing systems for antibiotic distribution and hospital care. Kaplan [8] imbedded the evaluation of vaccination logistics policy within a disease propagation model and compared strategies of traced vaccination, mass vaccination, and the mixed response advocated by the Centers for Disease Control. Eichner [73] developed a deterministic model for evaluating impact of different intervention strategies during pandemic influenza. The model is based on over 1000 differential equations which extend the classic SEIR model by clinical and demographic parameters relevant for pandemic preparedness planning. The model aims to operate with an optimal combination of precision, realism, and generality. Wu *et al.* [74] developed models to demonstrate that a prepandemic vaccine allocation policy that allocates vaccine to each state in proportion to the population size is not the most efficient. In fact, an inequitable strategy that allows no allocation to some regions while the sufficient vaccines being allocated to other regions demonstrates larger benefit. However, if considering other strategy selection criteria such as simplicity, robustness, and equity, the current prorata policy is a good compromise.

Integrated Approaches, and Information and Decision Support Systems

The burden of responding to a public health or medical disaster is multifaceted and a genuine test to the sustainability of critical infrastructure. Negotiating emergency operations is especially difficult due to interagency goal conflicts, differences in organizational culture and bureaucratic constraints, discrepancies in situation assessment, scarcity of resources, and organizational complexity. Nevertheless, interplay among many agencies is critical, and consequently, integrated approaches, and information and decision support systems prove to be very beneficial as part of the solution strategies [75–80].

Iakovou and Douligeris [81] presented the development of IMASH, an Information Management System for Hurricane disasters. IMASH is an intelligent integrated dynamic information management tool, capable of providing comprehensive data pertaining to emergency planning and response for hurricane disasters. Popp *et al.* [82] developed information analysis tools for an effective multiagency information-sharing effort.

Zografos *et al.* [83] described an integrated framework consisting of a data management module, a vehicle monitoring and communication module, and a modeling module, for managing emergency response of the electric utility companies. The framework integrates a GIS system with a decision-making modeling module to deliver solutions in real-time that optimize deployment of the available emergency resources. El-Anwar *et al.* [84] presented the development of an automated system to support decision makers in optimizing postdisaster temporary housing arrangements. The system has been integrated in MAEviz and provides the capability of optimizing a number of important objectives, including minimizing negative socioeconomic impacts, maximizing housing safety, minimizing negative environmental impacts, and minimizing public expenditures.

Bui [85] proposed a framework for developing a global information network (GIN). The application would incorporate four factors that affect the design of a GIN: nature of

disaster relief operations, negotiation styles of participants, social/cultural/organizational characteristics of participants, and resource availability. Such a framework could provide a set of basic metrics/factors to characterize any disaster situation. GIN would use high-speed internet as backbone, it includes a command center where the disaster management team would be, and telecommunication channels that connect to expert advice groups from around the world. The GIN would also be linked to an array of data and knowledge base warehouses.

NYU's PLAN C is an innovative tool for emergency managers, urban planners, and public health officials to prepare and evaluate Pareto-optimal plans to respond to urban catastrophic situations. Doheny and Fraser [86] described a software tool for modeling the decisions that people make in emergency situations in offshore environments. It can be used to predict the likely behaviors of a population in hazardous situations and help evaluate the effectiveness of emergency procedures and training.

Mondschein [87] reviewed the use of spatial data by environmental managers and emergency responders who are charged with the responsibility to perform hazard assessments, identify the location of toxic and hazardous materials, deploy emergency resources, and review demographic data to ensure the safety of the public and the surrounding communities.

In our own work, the decision support system RealOpt combines OR modeling techniques, novel and large-scale computational engines, sophisticated graph-drawing tools, 3D geographical spatial information with federal census data, and demographic and socioeconomic data for operational and strategic planning, and policy analysis. RealOpt allows public health emergency coordinators to (i) determine locations for service facilities setup; (ii) design customized, efficient floor plans for each facility via an automatic graph-drawing tool; (iii) determine (in real-time) optimal resource allocation through advanced computational techniques in simulation and optimization; (iv) monitor intrafacility disease propagation through a novel disease propagation model,

and derive dynamic response strategies to reduce the spread of disease and mitigate the risk of casualties; (v) assess resources and determine minimum needs to prepare for treating regional populations; (vi) carry out large-scale virtual drills and performance analyses, and investigate alternative dispensing strategies; and (vii) design a variety of dispensing scenarios and emergency-event exercises to train personnel [2,3,5,6,10,33].

Challenges

Modeling and optimizing public health infrastructure involve elements of resource allocation under risk, uncertainty, and time pressure; large-scale supply chain management; transportation and operational logistics; and medical treatment and population protection. The operations must be supported by an effective communication infrastructure. There is a necessity for vertical and horizontal integration and communication, where federal, state, local, tribal, territorial, private, and business stakeholders work toward a common goal of a resilient public health system. The infrastructure must be *flexible, scalable, sustainable, and elastic* to support an effective and timely response, and to mount rapid recovery and mitigation operations.

The 2007 Homeland Security Presidential Directive-21 (HSPD-21) [88] established a National Strategy for Public Health and Medical Preparedness, which builds upon a four pillar framework—threat awareness, prevention and protection, surveillance and detection, and response and recovery. It aims to transform our national approach to protecting the health of the citizens against all disasters. Although the four pillars were developed initially to guide the efforts to defend against a bioterrorist attack, they are applicable to a broad array of natural and man-made public health and medical challenges, and are appropriate to serve as the core functions of the strategy for public health and medical preparedness.

Public health and medical preparedness continue to shower challenges to the scientific community. Some critical issues include (i) realistic systems modeling; (ii) intractability of large-scale instances;

(iii) interdependencies among multiple critical components/agencies; and (iv) the importance and necessity for end-to-end systems modeling and design. Technological advances are needed to allow for complex realistic modeling while providing users with affordable computational power that result in decision systems that are practical for actual scenario-based analysis. The effective integration and alignment of care personnel, facilities, and equipment and supply for optimal outcome remains essential. Capability to solve large-scale resource allocation and location problems is a must. Tracking of disease and designing and implementing dynamic mitigation strategies will have a tremendous impact on population protection. Supply chain management needs to be dynamic, and multiagency partnership models should be developed. Information sharing and management, and risk and communication strategies continue to evolve. Multimodality integration of technologies and reliable dissemination are critical. Policy and coordination among different stakeholders across country borders need to be studied and potentially streamlined. While many issues relate to operational and strategic planning, many others involve policies, risk management, security, public communication, and cultural and human behavior.

The key to success is flexibility and adaptability—in staffing, operations strategies, coordinating and communications strategies, and the willingness (for multiple agencies) to collaborate. Initial plans must be put in place and executed rapidly, and yet allow reconfiguration on the fly as an event unfolds.

Acknowledgments

Part of the material and results reported herein are based on our work in this area and interaction with public health agencies, and discussion with many state and federal public health and emergency response experts. While there are many people to thank in this multiagency and multidisciplinary collaboration, the authors would like to specially thank Dr. Jacquelyn Mason of the CDC, and Tom Tubesing, formerly of the CDC, Dr. Lawler

and Dr. Mecher at the Homeland Security Council in the White House, William Glisson at ESI, Bernard Hicks at DeKalb Emergency Preparedness Department, and the many public health and emergency managers throughout the nation. We also acknowledge the funding from the CDC. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

1. Lee EK, Maheshwary S, Mason J. Real-time staff allocation for emergency treatment response of biologic threats and infectious disease outbreak. INFORMS William Pierskalla Best Paper Award on Healthcare and Management Science. 2005.
2. Lee EK, Maheshwary S, Mason J, *et al.* Large-scale dispensing for emergency response to bioterrorism and infectious disease outbreak. *Interfaces* 2006;36(6):591–607.
3. Lee EK, Maheshwary S, Mason J, *et al.* Decision support system for mass dispensing of medications for infectious disease outbreaks and bioterrorist attacks. *Ann Oper Res Comput Optim Med Life Sci* 2006;148:25–53.
4. Lee EK. Doing good with good O.R. O.R.s do-gooders. *National biodefense—in case of emergency. OR/MS Today* 2008;35(1):28–34.
5. Lee EK, Chen CH, Pietz F, Benecke B. Modeling and optimizing the public health infrastructure for emergency response. *Interfaces* 2009;39(5):476–490. (The Daniel H. Wagner Prize for Excellence in Operations Research Practice.)
6. Lee EK, Smalley HK, Zhang Y, *et al.* Facility location and multi-modality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. *Int J Risk Assess Manage* 2001;12(2/3/4):311–351. (Special issue on biosecurity assurance in a threatening world: challenges, explorations, and breakthroughs.)
7. Gerberding JL. CDC efforts in implementing a smallpox vaccination program. *Testimony*. Washington, DC: United States Department of Health and Human Service; 2003 Jan 30.
8. Kaplan EH, Craft DL, Wein LM. Emergency response to a smallpox attack: the case for mass vaccination. *Proc Natl Acad Sci* 2002;99:10935–10940.

9. Wein LM, Craft DL, Kaplan EH. Emergency response to an anthrax attack. *Proc Natl Acad Sci* 2003;100:4346–4351.
10. Lee EK, Institute of Medicine. A systems view of POD operations: integrating all the elements. Institute of Medicine Workshop “Medical Countermeasures Dispensing: Emergency Use Authorization”. Washington (DC): 2009 Nov.
11. Swain R, ReVelle C, Toregas C, *et al.* The location of emergency service facilities. *Oper Res* 1971;19(6):1363–1373.
12. Larson RC. Approximating the performance of urban emergency service systems. *Oper Res* 1975;23(5):845–868.
13. Aly AA, White JA. Probabilistic formulation of the emergency service location problem. *J Oper Res Soc* 1978;29(12):1167–1179.
14. Chaiken JM, Larson RC. Methods for allocation urban emergency units: a survey. *Manage Sci* 1972;19(4):P110–P130.
15. Hogan K, ReVelle C. Concepts and applications of backup coverage. *Manage Sci* 1986;32(11):1434–1444.
16. Pirkul H, Schilling DA. The siting of emergency service facilities with workload capacities and backup service. *Manage Sci* 1988;34(7):896–908.
17. Narasimhan S, Pirkul H, Schilling DA. Capacitated emergency facility siting with multiple levels of backup. *Ann Oper Res* 1992;40(1):323–337.
18. Jia H, Ordóñez F, Dessouky M. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Trans* 2002;39(1):41–55.
19. Berman O, Gavious A. Location of terror response facilities: a game between state and terrorist. *Eur J Oper Res* 2007;177(2):1113–1133.
20. Church RL, Scaparra MP, Middleton RS. Identifying critical infrastructure: the median and covering facility interdiction problems. *Ann Assoc Am Geogr* 2004;94(3):491–502.
21. Church RL, Scaparra MP. Protecting critical assets: The R-interdiction median problem with fortification. *Geogr Anal* 2007;39(2):129–146.
22. Brandeau ML, Chiu SS. An overview of representative problems in location research. *Manage Sci* 1989;35(6):645–674.
23. Owen SH, Daskin MS. Strategic facility location: a review. *Eur J Oper Res* 1998;111(3):423–447.
24. Mason J, Washington M. Optimizing staff allocation in large-scale dispensing centers. CDC Report. 2003.
25. The Buffalo News. Drive-through vaccination effort a success in Amherst: 1,385 people receive hepatitis A booster. 2008 Sep 22.
26. Anderson RM, May RM, Anderson E. Infectious disease of human: dynamics and control, Oxford, UK: Oxford University Press; 1992.
27. Diekmann O, Heesterbeek J. Mathematical epidemiology of infectious diseases: model building. New York: Wiley; 2000.
28. Ferguson NM, Cummings DA, Cauchemez S, *et al.* Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 2005;437:209–214.
29. Ferguson NM, Cummings DAT, Fraser C, *et al.* Strategies for mitigating an influenza pandemic. *Nature* 2006;442:448–452.
30. Longini IM Jr, Nizam A, Xu S, *et al.* Containing pandemic influenza at the source. *Science* 2005;309:1083–1087.
31. Germann TC, Kadau K, Longini IM, *et al.* Mitigation strategies for pandemic influenza in the United States. *Proc Natl Acad Sci* 2006;103:5935–5940.
32. Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics—III. Further studies of the problem of endemicity. *Bull Math Biol* 1991;53:89–118.
33. Lee EK, Chen CH, Pietz F, *et al.* Disease propagation analysis and mitigation strategies for effective mass dispensing. Proceedings of the American Medical Informatics Association 2010 conference. Washington (DC): 2010. In press.
34. Kapucu N. Interagency communication networks during emergencies: boundary spanners in multiagency coordination. *Am Rev Public Adm* 2006;36:207–225.
35. Wein LM. 2008. Neither snow, nor rain, nor anthrax. *The New York Times*; 2008 Oct 13. http://www.nytimes.com/2008/10/13/opinion/13wein.html?_r=1. Accessed 2010 Feb 22.
36. Chomilier B, Samii R, Wassenhove LNV. The central role of supply chain management at IFRC. *Forced Migr Rev* 2003;18:iii.
37. Cassidy W. A logistics lifeline. *Traffic World*. 2003.
38. Murray S. How to deliver on the promises: supply chain logistics: humanitarian agencies are learning lessons from business in bringing essential supplies to regions hit by the tsunami. *Financial Times* 2005 Jan 6.

39. Long DC, Wood DF. The logistics of famine relief. *J Bus Logist* 1995;16:213–229.
40. Van Wassenhove L. Blackett memorial lecture humanitarian aid logistics: supply chain management in high gear. *J Oper Res Soc* 2006;57:475–489.
41. Larson RC, Metzger MD, Cahn MF. Responding to emergencies: lessons learned and the need for analysis. *Interfaces* 2006;36(6):486–501.
42. Green LV, Kolesar PJ. Anniversary article: improving emergency responsiveness with management science. *Manage sci* 2004;50:1001–1014.
43. Fiedrich F, Gehbauer F, Rickers U. Optimized resource allocation for emergency response after earthquake disasters. *Saf Sci* 2000;35:41–57.
44. Branas CC, MacKenzie EJ, ReVelle CS. A trauma resource allocation model for ambulances and hospitals (managerial and policy impact). *Health Serv Res* 2000;35(2):489–507.
45. Tzeng G-H, Cheng H-J, Huang TD. Multi-objective optimal planning for designing relief delivery systems. *Transp Res Part E: Logist Transp Rev* 2007;43:673–686.
46. Yan S, Shih Y-L. A time-space network model for work team scheduling after a major disaster. *J Chin Inst Eng* 2007;30:63–75.
47. Zhang M, Yang J. Optimization modeling and algorithm of facility location problem in perishable commodities emergency system ICNC '07: Proceedings of the Third International Conference on Natural Computation. IEEE Computer Society; 2007;246–250.
48. Doerner KF, Gutjahr WJ, Nolz PC. Multi-criteria location planning for public facilities in tsunami-prone coastal areas. *OR Spectr* 2009;31:651–678.
49. Coskun N, Erol R. An optimization model for locating and sizing emergency medical service stations. *J Med Syst* 2010;34:43–49.
50. Harewood S. Emergency ambulance deployment in Barbados: a multi-objective approach. *J Oper Res Soc* 2002;53:185–192.
51. Wang J, Hu X, Xie B. Emergency vehicle routing problem in post-earthquake city road network. International Conference on Transportation Engineering. Chengdu, China: 2009.
52. Sheu J-B. An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transp Res Part E: Logist Transp Rev* 2007;43:687–709.
53. Yi W, Kumar A. Ant colony optimization for disaster relief operations. *Transp Res Part E: Logist Transp Rev* 2007;43:660–672.
54. Yuan Y, Wang D. Path selection model and algorithm for emergency logistics management. *Comput Ind Eng* 2009;56:1081–1094.
55. Liu M, Zhao L. Optimization of the emergency materials distribution network with time windows in anti-bioterrorism system. *IJICIC* 2009;5:3615–3624.
56. Hupert N, Mushlin AJ, Callahan MA. Modeling the public health response to bioterrorism: using discrete event simulation to design antibiotic distribution centers. *Med Decis Making* 2002;22(5 Suppl): S17–S25.
57. Hupert N, Bearman GML, Mushlin AI *et al.* Accuracy of screening for inhalational anthrax after a bioterrorist attack. *Ann Int Med* 2003;139(5):337–345.
58. Wang J, Yu H, Luo J, *et al.* Medical treatment capability analysis using queuing theory in a biochemical terrorist attack. The 7th International Symposium on Operations Research and Its Applications (ISORA'08). 2008. pp. 2415–2424.
59. Rossetti MD, Trzcinski GF, Syverud SAPA, *et al.*, editors. Emergency department simulation and determination of optimal attending physician staffing schedules. Proceedings of the 1999 Winter Simulation Conference. New York: ACM; 1999.
60. Saleh MS, Othman ZZA. ASRTS: an agent-based simulator for real-time schedulers. Second Asia International Conference on Modeling and Simulation AICMS 08; 2008 May 13–15; Kuala Lumpur, Malaysia. 2008.
61. Aaby K, Herrmann JW, Jordan CS, *et al.* Montgomery county's public health service uses operations research to plan emergency mass dispensing and vaccination clinics. *Interfaces* 2006;36(6):569–579.
62. Barnes AJ, Jacobson JO, Solomon MD, *et al.* Los Angeles county pandemic flu hospital surge planning model. National Health Foundation; 2009.
63. Das TK, Savachkin AA, Zhu Y. A large scale simulation model of pandemic influenza outbreaks for development of dynamic mitigation strategies. *IIE Trans* 2007;40(9):893–905.
64. Lee EK, Yang AY, Chinnappan SG, *et al.* Survey and analysis of emergency communication infrastructure among Georgia state hospitals.

- Altalanta (GA): Georgia Division of Public Health; 2009.
65. Meltzer M, Damon I, LeDunc JW, *et al.* Modeling potential responses to smallpox as a bioterrorist weapon. *Emerg Infect Dis* 2001;7(6):959–969.
 66. Wu JT, Leung GM, Lipsitch M, *et al.* Hedging against antiviral resistance during the next influenza pandemic using small stockpiles of an alternative chemotherapy. *PLoS Med* 2009;6:1–11.
 67. Goldberg J, Paz L. Locating emergency vehicle bases when service time depends on call location. *Transp Sci* 1991;25(4):264–280.
 68. Haghani A, Tian Q, Hu H. Simulation model for real-time emergency vehicle dispatching and routing. *Transp Res Rec: J Transp Res Board* 2004;1882:176–183.
 69. Su S, Shih C-L. Modeling an emergency medical services system using computer simulation. *Int J Med Inform* 2003;72:57–72.
 70. Eubank S. Scalable, efficient epidemiological simulation. SAC '02: Proceedings of the 2002 ACM symposium on applied computing. New York: ACM Press; 2002. pp. 139–145.
 71. Gani R, Leach S. Transmission potential of smallpox in contemporary populations. *Nature* 2001;414(6865):748–751.
 72. Epstein J, Cummings DAT, Chakravarty S, *et al.* Toward a containment strategy for smallpox bioterror: an individual-based computational approach. Washington, DC: The Brookings Institution Press; 2004.
 73. Eichner M, Schwehm M, Duerr H-P, *et al.* The influenza pandemic preparedness planning tool Influsim. *BMC Infect Dis* 2007; 7.
 74. Wu JT, Riley S, Leung GM. Spatial considerations for the allocation of pre-pandemic influenza vaccination in the United States. *Proc R Soc B: Biol Sci* 2007;274:2811–2817.
 75. Kananen I, Korhonen P, Wallenius J, *et al.* Multiple objective analysis of input-output models for emergency management. *Oper Res* 1990;38:193–201.
 76. Kwan M-P, Lee J. Emergency response after 9/11: the potential of real-time 3D GIS for quick emergency response in micro-spatial environments. *Comput Environ Urban Syst* 2005;29:93–113.
 77. Nguyen S, Rosen J, Koop C. Emerging technologies for bioweapons defense. *Stud Health Technol Inform* 2005;111:356–361.
 78. Raghu TS, Ramesh R, Whinston AB. Addressing the homeland security problem: A collaborative decision-making framework. *J Am Soc Inform Sci Technol* 2005;56(3):310–324.
 79. Rotz LD, Hughes JM. Advances in detecting and responding to threats from bioterrorism and emerging infectious disease. *Natl Med* 2004;10:S130–S136.
 80. Subramaniam C, Kerpedjiev S. Dissemination of weather information to emergency managers: a decision support tool. *Eng Manage IEEE Trans* 1998;45:106–114.
 81. Iakovou E, Douligeris C. An information management system for the emergency management of hurricane disasters. *Int J Risk Assess Manage* 2001;2:243–262.
 82. Popp R, Armour T, Senator T, *et al.* Countering terrorism through information technology. *Commun ACM* 2004;47:36–43.
 83. Zografos KG, Douligeris C, Tsoumpas P. An integrated framework for managing emergency-response logistics: the case of the electric utility companies. *IEEE Trans Eng Manage* 1998;45:115–126.
 84. El-Anwar O, El-Rayes K, Elnashai A. An automated system for optimizing post-disaster temporary housing allocation. *Automat Constr* 2009;18:983–993.
 85. Bui T, Cho S, Sankaran S, *et al.* A framework for designing a global information network for multinational humanitarian assistance/disaster relief. *Inf Syst Front* 2000;1:427–442.
 86. Doheny JG, Fraser JL. MOBEDIC—A decision modelling tool for emergency situations. *Expert Syst Appl* 1996;10:17–27.
 87. Mondschein LG. The role of spatial information systems in environmental emergency management. *J Am Soc Inf Sci* 1994;45: 678–685.
 88. The White House. Homeland Security Presidential Directive 21 [HSPD-21]: Public Health and Medical Preparedness. 2007. Available at <http://www.fas.org/irp/offdocs/nspd/hspd-21.htm>. Accessed 2010 Feb 27.

PUBLIC HEALTH, EMERGENCY RESPONSE, AND MEDICAL PREPAREDNESS III: COMMUNICATION INFRASTRUCTURE

EVA K. LEE

ANNA YANG YANG

Center for Operations Research in
Medicine and Healthcare, School
of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

NSF I/UCRC Center for Health
Organization Transformation,
Atlanta, Georgia

School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

TYRUS GUILFORD

Division of Emergency Preparedness
and Response, Georgia
Department of Community
Health, Atlanta, Georgia

FERDINAND PIETZ

Strategic National Stockpile, Office
for Public Health Preparedness
Response, Centers for Disease
Control and Prevention, Atlanta,
Georgia

BERNARD BENECKE

Global Disease Detection and
Emergency Response, Center for
Global Health, Centers for
Disease Control and Prevention,
Atlanta, Georgia

INTRODUCTION

Medical and public health systems must prepare for major emergencies or disasters involving human casualties. Such events will severely challenge the ability of health-care systems to adequately care for large numbers of patients (surge capacity) and/or victims with unusual or highly specialized medical

needs (surge capability) (see *Public Health, Emergency Response, and Medical Preparedness I: Medical Surge*). In addition, medical and public health systems can expect incidents that significantly impact their usual operations, as occurred with Hurricane Katrina. These so-called *mass effect* events can have devastating consequences for medically fragile segments of society and those living with chronic health conditions. Limited or no access to routine health-care services can cause these populations to rapidly deteriorate, producing a downstream surge of demand for acute care that can overwhelm local capabilities.

Within the United States since 2003, health-care entities as well as public health and other response partners have been directed to build redundant communications systems with multiple communication technologies that would ensure connectivity and operability in the event of a public health emergency. A significant amount of funding has been provided to the states on communication devices and the building of a redundant communication system.

By FY 2007, states health departments were required to show an operational redundant communication system that is capable of communicating both horizontally between health-care providers and vertically with the jurisdiction incident command structure as described in the tiered response system outlined in the Medical Surge Capacity and Capability Handbook and in the Hospital Preparedness Program (HPP) FY 2006 guidance. The system must link all health-related organizations that participate in the program and deemed necessary to the State and local jurisdiction health and medical response operations plan, including law enforcement, public works, and others. This system should have the ability to exchange voice and/or data with all partners on demand, in real time, when needed and as authorized in the operational plans developed by the State and local jurisdictions.

All systems must meet SAFECOM requirements for communications interoperability, as established in April 2004 by the Department of Homeland Security and introduced in the FY 2006 HPP guidance. SAFECOM defines the future requirements for crucial voice and data communications in day-to-day, task force, and mutual aid operations, and establishes parameters around owned or leased radio equipment use.

This article presents an overview of communication infrastructure needs, and findings from our recent study and evaluation of statewide emergency communication systems in Georgia. Specifically, the section titled “Emergency Communication System” describes the importance and necessity of the emergency communication system. Physical infrastructure including various phone and radio technology are described and contrasted in the section titled “Physical Infrastructure for Emergency Communication,” and communication infrastructure is highlighted. This is followed by discussion of communication-facilitated programs including Telecommunications Service Priority Program, Government Emergency Telecommunications Services, and Wireless Priority Service Program in the section titled “Communication Facilitation Programs.” The section titled “Organizational Factors in Emergency Communication” states the organizational factors that are critical to cultivate and develop a strong communication infrastructure and personnel capability. The section titled “Sociological Factors and Linkage to Human Behavior” focuses on sociological factors and human behavior. Complementing the communication network for emergency responders, the section titled “Public Information and Risk Communication” touches on the need to maintain multimedia broadcast communication technology to inform the public of safety precautions, where to seek help, and any other important public service announcements.

EMERGENCY COMMUNICATION SYSTEM

Multifaceted communication infrastructures are needed for effective medical and

emergency response. Two-way communication lines must be kept open between emergency/medical responders and commanders for coordination and facilitation of the response effort; broadcast communication lines must be available to inform the general public about medical precautions and available services; and phone service and call answering stations must be maintained to allow the public to report critical emergency situations. In this section, we focus on the communication infrastructure of hospitals and related medical care entities.

Developing and maintaining a robust communication infrastructure among hospitals, emergency medical services (EMS), and other health-care facilities, as well as private medical transport companies and medical supplies vendors, is vital to providing relief operations and services during emergency situations (see *Public Health, Emergency Response, and Medical Preparedness I: Medical Surge*; *Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts*; and *Public Health, Emergency Response, and Medical Preparedness IV: Emergency Evacuation*). Such infrastructure provides the core foundation for basic knowledge-sharing such as patient volume and severity, emergency room capacity, hospital bed availability; special wards availability; medical personnel specialties; transport vehicle availability; and blood, medicine, and medical supplies inventories. During an emergency situation, such critical information can be communicated to a regional coordinating control center, which can assess available resources, identify surpluses and shortages, and coordinate distribution efforts. In fact, building capacity for an interoperable communication system for emergency response is identified as one of the areas that must be prioritized according to The Hospital Preparedness Program established by the US Department of Health and Human Services.

Previous instances of emergency situations such as the September 11 terrorist incident and the Rhode Island night club fire incident have brought forth the challenges in

and necessity for establishing and sustaining a redundant emergency communication network. Kapucu [1] examined the importance and challenges in establishing a coordinated interagency communication network and the role information technology serves to enhance such communication and decision making as in the context of the September 11 terrorist incident. Challenges include interoperability among the various communication tools, reliability of the tools during emergency situations, technical limitations, barriers in interagency communication, and related aspects.

Physical Infrastructure for Emergency Communication

Studies have been conducted in analyzing the requirement for effective hospital communication systems. Becker [2] discussed the effectiveness of focus groups and tabletop exercises involving key stakeholders in the public health community for initial structuring of the emergency communication infrastructure. Such exercises serve as a platform for discussing the functional and performance requirements for redundant communication and the type of information to be delivered. Manoj and Baker [3] identified three categories of communication challenges from literature review: technological, sociological, and organizational. Recently, we have completed an evaluation of emergency communication infrastructure among a network of 148 hospitals in Georgia.

Below, we summarize some of our findings, and discuss the status quo and challenges faced by the health systems [4].

Communication Tools. A wide range of communication tools are used in the health-care facilities. They can be classified into three main categories: phones (such as landline, mobile phones, satellite, dedicated, and VoIP), radios (such as 700 MHz, 800 MHz, UHF, VHF, microwave, and HAM) and computer-based tools (such as email and web-based collaborating platforms). When comparing the effectiveness of the tools, some of the metrics to take into consideration are medium of communication, security, emergency reliability factors, emergency

response services, cost of installation and usage, general application in public and private domain, and technological limitation. Table 1 compares telephones used in health-care facilities.

Landlines and cellular phones are commonly used in hospitals in both daily operations and emergencies as primary communications tools due to their low learning curve and relatively low cost. However, in a disaster event, many factors could contribute to the vulnerability of these tools, such as network congestion, network structural damage as observed in natural disaster, or a carefully crafted terrorist attack with the communication system as a target. On the other hand, tools like satellite phone and dedicated phone lines require higher initial investment and higher maintenance, but demonstrate high reliability. VoIP phones have high reliance on availability and stability of broadband internet. Current statistics on usage of VoIP phones is likely to change drastically in the near future given the emerging trend of dependency toward the internet for communication.

Table 2 compares the functions and characteristics of various types of radios. Among all the hospitals in the State of Georgia, 800-HMz radio, VHF, UHF, and HAM radio are more commonly seen than other radio technologies [4]. The percentage of hospitals possessing these tools ranges from 36 to 50%. Microwave radios, in particular, offer a wide range of features for emergency communication but are currently not used for hospital emergency communications. Adoption of new technology can be slow due to the lack of financial resource as well as potential training cost in equipment usage.

Researchers continue to examine the functionality and effectiveness of some novel tools with hybrid technologies. Ammenwerth *et al.* [5] examined the usability of a multifunctional mobile information and communication system through a simulation study conducted in a model hospital environment. Holzman [6] proposed the design of a computer-based communication device that can be used for onsite patient information retrieval during traumatic situations. The purpose of the user interface is to facilitate a

Table 1. Telephony Networks Comparison [4]

Features	Landline	Cellular	Satellite	Dedicated	VoIP
Medium of communication	Metal wire Optical fiber Airwaves	Airwaves	Airwaves	Metal wire Optical fiber	Metal wire Optical fiber Airwaves
Security	Secure	Insecure	Secure	Secure	Insecure
Emergency reliability factors	Network congestion	Network congestion Infrastructure reliability	Satellite congestion	Infrastructure reliability	Network congestion Infrastructure reliability Transmission system crash
Emergency applications	Text messaging	Push to talk Text messaging Email	Text messaging Calling services	Seamless connectivity between terminals	E911 services Text messaging Calling services
Cost of installation and usage	Installation: \$50-\$75 Usage: around \$40/ month	Installation: \$25-\$50 Usage: Ranges from \$30-\$100/ month	Installation: \$200-\$2600 (Handset cost) Usage: prepaid cards ranging from \$200-\$5000	Requires specialized cable installation Applicable for high priority communication services, e.g., military applications	Free Installation Usage depends on subscription and internet connection
Strength	Ease of usage and subscription	Ease of usage and subscription Mobility	Seamless connectivity High mobility	Dedicated connection between terminals No network congestion	Wide range of applications Ease of usage and subscription Mobile connectivity

Limitations	System getting outdated owing to better features of cellular/VoIP providers Higher maintenance and service requirements	Network congestion at peak/emergency hours Service provider infrastructure and network capabilities	High cost of handset as well as usage Satellite congestion lead to service disruption	Very high cost of installation System depends on cable infrastructure Higher maintenance and service requirements	Available bandwidth Network latency Echo, reliability Quality of service Lack of technical assistance
Caller tracking	Based on caller information Tracking based on requirement	User-assisted GPS to track caller	Uses GPS recognition to track callers	Standard tracking based on requirement	Depends solely on subscriber responsibility No tracking of caller possible
Additional features	Caller ID Conferencing Voice-based services	Broadband Internet Voice-based services Text alert services Flexible calling tariff	Broadband Internet Voice-based services Text messaging E-mail	Secure call transmission	Conferencing Secured subscriber preference Location independence Advanced telephony features

Table 2. Radio Networks Comparison [4]

Features	700 MHz	800 MHz	VHF	UHF	Microwave	HAM
Frequency	700–800 MHz	800–900 MHz	30–300 MHz	300–3 000 MHz	7–38 GHz	1.6–27 MHz
Security	Digital encryption Local IP address	Digital encryption Privacy plus Private conversation	Analog direct Unintercepted communication	Digital encryption	Highest link reliability IP/data transmission protocols	Digital encryption Dual/round table transmission
Emergency networks range	Block D: 758–763 and 788–793 MHz	Public safety 21–824 MHz 866–869 MHz	156.8 MHz, 164.4 MHz	450–470, 800 MHz subband, 2500 MHz subband	7–38 GHz	148.8 MHz
Usage features	VoIP services High-speed broadband access Video for surveillance High speed in built access for MTU/MDU	Trunked radio service Computer-controlled network assignment Call monitoring Interference reduction Frequency allocation	Voice operations in repeaters and voice links Link with computer-controlled equipment Reliable telemetry link in place of expensive phone lines	Trunked frequency radio communication Computerized call/data monitoring	Cellular networks for high speed links between stations Wireless local loop systems Broadband internet Voice/data application and emergency restoration	Automatic dialer Node terminal operation Multiple scan Computer interface
Cost	Handset: \$99–\$1 000	Handset: \$80–\$1 000	Handset: \$75–\$500	Handset: \$100–\$3 000	System: \$2000 and above	Handset: \$50–\$2500

✓	Pros	Direct operation with other 700–800 MHz users No infrastructure needed for field communications Nonrepeated system Simple reliable system No audio delays	Wide area coverage Digital signaling features Interoperability Talk group creation	No special infrastructure requirements for field communications Non-repeated systems Simple reliable systems No audio delays	Physical barrier penetration. Less interference from other bands High data transmission efficiencies Widely used	More signal transmission than wired systems Immediate spread spectrum Microwave link setup for disaster situations Military use Wide ranging frequencies for transmission	Networking for amateur usage Used in emergency communication during/after disasters
	Cons	Frequencies not available in all areas Units keying up simultaneously Audio quality least preferred No warning of fading	Complex infrastructure Noticeable audio delays Loss of system coverage required in buildings Inconsistent interior communication	System must meet FCC narrow banding requirements by 2013 Limited interoperability with other agencies on 700–800 MHz Systems not seamless Analog equipment may not be available due to digital transition	Commercial licensing underway Digital transmission lessens bands for emergency amateur usage Requires advances in equipment Battery requirements for sustained communication	Free space loss Transmitter sensitivity Infrastructure requirements Multipath interference Signal-to-noise ratio considerations	Amateur licensing requirements

Table 2. (Continued)

Features	700 MHz	800 MHz	VHF	UHF	Microwave	HAM
Application	Public safety Government entities Logistics Property management Hotel Port	Public safety Amateur radio communica- tion Closed network communica- tion	Airband (aircraft) radio Amateur radio FM radio broadcasts Marine VHF radio	Emergency response in public networks Private closed circle data transmission Amateur radio	Low power radio for wireless networking within a building Long distance telephone signal transmission	Amateur radio Emergency response systems
Special features	Communication range alert Programmable frequencies CTCSS/CDCSS work group authorization	Interoperability Interference control through computerized network traffic monitoring	IP switch frequency allocation Noise synthesizer Commercial grade TCXO for tight frequency accuracy	Programmable frequencies Interference reduction techniques	Digital data transmission Superior technology for 100% data transmission capabilities Interfaced with computer	High frequency stability with built in TCXO grade modules

common platform for patient data collection and transmission so that the medical facilities can be prepared for the proper medical procedure immediately upon the patient's arrival rather than spending time on information retrieval and diagnosis. Among the various information technology applications for out-of-hospital response phase, Arnold and Levine [7] analyzed the usage of P2P wireless communication application in a hospital setting and concluded that it provides superior performance than related applications by integrating reliability, flexibility, and information security.

Communication Infrastructure. Building a reliable communication infrastructure requires not only the use of the right tool but also using the tool right. Jennex [8] proposed a model for emergency response system and discussed the imperative need for a dynamic integrated collaborative method to communicate between users and data sources and the necessity for establishing protocols to facilitate communication and to improve decision making. Related work by Jan and Hwang [9] included optimizing the topological layers of links within the design of a communication network that simultaneously minimizes cost while satisfying reliability constraints over the links.

Communication during a disaster involves the connection between multiple agents, such as a regional coordinating hospital, community emergency operations center, local EMS services, county health department, and the local police and fire department. A reliable system also requires multiple layers of communication systems. According to our analysis of hospitals in the state of Georgia [4], most hospitals have two levels of communication infrastructure built with tools that are commonly seen but are subject to damage/congestion during emergency (such as cell phones and landlines). Manoj and Baker [3] identified the multiorganizational radio interoperability issue as an important obstacle to overcome.

Communication Facilitation Programs

The Federal Communications Commission and the National Communication System

have established various emergency communication restoration systems and protocols that facilitate public as well as hospital communications during emergency situations. Below we compare three existing programs, namely, Telecommunications Service Priority (TSP) Program, Government Emergency Telecommunications Services (GETS), and Wireless Priority Service (WPS) Program.

The TSP program, a Federal Communication commission and National Communications Systems initiative, is aimed at maintaining priority over the telecommunication links (wireline and wireless) during emergency so that essential services are not disrupted in operation. This system is of vital importance to services such as disaster communication centers (911 call centers), police, fire, and so on. The enrolling criterion includes a nonrefundable fee of \$100, monthly subscription maintenance charge of \$3 per line, and enrollment in the NS/EP program. The limitations of the system include monthly subscription, renewal of the license every three years, and the lack of interoperability in case of landline disconnection or wireline provider shift.

The GETS program, a National Communication Commission initiative, is aimed at providing wireline subscriber priority and communication restoration through usage of GETS card. The enrolling requirements include subscription in the NS/EP program, and the cost of usage of the GETS card is 10c per minute within the United States. GETS is an essential program for unrestricted emergency operations with around 90% call completion rates even when the congestion factor is 8. It offers seamless connectivity at times of emergency. Some of the unique features of the system include a toll-free access number, access control through PIN number usage, failsafe access, enhanced carrier routing/alternative carrier routing, priority treatment with trunk queueing, subgrouping and reservation, and interoperability with other networks.

The WPS, a joint initiative by the Federal Communication Commission and private wireless carriers, is aimed at Wireless priority and communication restoration for authorized personnel at times of emergency. The

enrollment in the program is subjected to a federal scrutiny and is restricted to key members in Public health/safety/law enforcement and the likes. The benefits of the system include high call completion probability by NS/EP call marking, originating radio channel priority, and terminating radio channel priority by call queueing. The limitations of the system include monthly payment for service restoration and program enrollment, service being carrier dependant, and lack of system support for prepaid cell phones and TDMA technology. Since the cellular prioritizing is computerized, the call connectivity depends on the priority ranking. Although the system is currently available throughout the country, the service is carrier dependent.

In Table 3, we first analyze the operational attributes of each of these facilitation systems and then compare their effectiveness. Although the three existing facilitating programs that are sanctioned and run by the federal government and dedicated for emergency communications restoration offer some valuable features for emergency response, they are not fully utilized as our findings show low enrollment among the hospital networks [4]. Such communication facilitation programs can be invaluable during an emergency situation and should be encouraged in a national level.

Organizational Factors in Emergency Communication

Organizational factors in emergency communication involve managerial attitude toward building healthy communication systems as well as the flexibility and adaptability of an organization in terms of hierarchical reporting structure in the event of a disaster.

Adapting to novel technologies could involve a somewhat steep learning curve. It is therefore important that training and testing of emergency communication tools are emphasized from the managerial level. It is also not practical to assume that operations will function properly in a crisis. A regular testing of the communication system will improve the organization's responsiveness to crisis in the following ways:

- testing the effectiveness of training;
- familiarizing staff and workers with emergency procedures; and
- identifying further problems existing within the system, as well as organization policies.

Another important factor in emergency communication is the reporting hierarchy and decision making. Response to emergencies may require decision making to take place as they might in a flatter, more dynamic, *ad hoc* organization. While hierarchical organizations can lead to wider information gaps across the organizations, flat organizations are not easily scalable. (A hierarchical organization refers to those with more structured pyramid-like organization where one person is in charge of a functional area with one of more subordinates handling the subfunctions. Flat organization refers to an *organization structure* with few or no levels of intervening management between staff and managers.) Therefore, a hybrid organizational model needs to be developed to best utilize the two organizational approaches [3,10–12].

Sociological Factors and Linkage to Human Behavior

The social challenges that arise with communication as a result of human behavioral pattern must be considered when designing a communication system.

Confidentiality and information security are always legitimate concerns in the dissemination of medical information. These concerns are escalated in emergency situations since the setting is unfamiliar and emergency response usually involves multiple agencies and stakeholders, some of whom may use volunteers who may not understand the sensitivity of information and the need to protect patient privacy.

Emergency situations can also cause volatile emotions, which may hinder communication. Fear, stress, and other emotions can be aggravated by the lack of information. Therefore, periodic information updates are important. Hegde *et al.* [13] presented a technological solution that provides differentiated services for an agitated caller by

Table 3. Comparison Between the TSP, GETS, and WPS [4]

Features	TSP	GETS	WPS
Objective	Prioritize communication of wireline and wireless subscriber	Prioritize communication of wireline subscriber Communication restoration through usage of GETS card	Wireless priority and communication restoration for authorized personnel's
Controlling authority	Federal Communication Commission and National Communications Systems	National Communication Systems	Federal communication Commission and private wireless carriers
Enrollment requirements	Enrollment in one of the National Security/Emergency preparedness program Pay the required dues and submit requisition form After scrutiny, notify authorities for emergency restoration protocols	Enrollment in National Security/Emergency preparedness program Public health, safety and maintenance of law & order	Key federal, state, local and tribal government and critical infrastructure Personnel requirements: – Executive leaderships – Disaster response/military command – Public health/safety/law enforcement command – Public services/utilities
Cost of enrollment	Enrollment fee: \$100 Monthly charge per line:\$3	No enrollment charge Call charge: 10c/min for calls within United States	Activation charge: \$10 Monthly charge: \$4.50
Benefits	Restoration of services in essential disaster communication centers such as 911 call centers, police, fire, and so on	Essential program for unrestricted emergency operations Around 90% call completion rates when congestion factor is 8 No additional software requirements	High call completion probability Originating radio channel priority Terminating radio channel priority by call queueing

Table 3. (Continued)

Features	TSP	GETS	WPS
Limitations	Monthly payment for service restoration Revalidation of TSP authorization every three years Entire restoration structure pertains to subscribed telephone company and needs to be modified in case of phone change	Usage charge and tariff based on wireline provider Requires calls to be placed in landline for priority treatment Federal thresholds based on tariff, usage Private entities require further scrutiny	Monthly payment for service restoration and program enrollment Cannot be provisioned on prepaid cellphones Service not available on TDMA technology
Special features	Prioritizing serves as a source of National Security/Emergency preparedness	Toll-free access number PIN number access control Failsafe access Enhanced/alternative carrier routing Interoperability with other networks International calling/number translation	Computerized cellular prioritizing and priority ranking Service availability is carrier dependent High availability throughout the country Software enhancements are provided for priority restoration of services

detecting the emotional content in speech packets over a wireless network.

The lack of common vocabulary between response agencies and between organizations and victim populations makes communication even more difficult. While concerted effort has been made to improve the communication between organizations through common vocabulary, efficiency is still lacking. Social science research has been conducted to investigate common languages and principles such as icon languages for use between response organizations and the affected population. Social networks such as Twitter and Facebook are common meeting places nowadays. Such communication infrastructure can be very effective in message dissemination. However, its full effect relies on users' capability to differentiate facts from rumors, and to perform measures

according to instruction upon receiving the information.

PUBLIC INFORMATION AND RISK COMMUNICATION

While the emergency communication systems across various health systems, EMS, and emergency departments provide responders and commanders the capability to maximize their rescue and response ability and manage their available resources so as to best serve the affected population during crisis, parallel communication needs to exist to inform the public of safety precautions, where to seek help, and any other important public service announcements. Risk communication plays a crucial role to any successful emergency response. The social challenges

that arise as a result of human behavioral patterns cannot be overemphasized. During the high concern and high stress situation, the messages to the public must be concise, unified, and coordinated. Getting correct and credible information out quickly assures the public that they can trust the authority in protecting themselves and their families. The public needs to know in a timely manner that there is a problem and the nature of the problem. Appropriate measures must be performed to curtail rumors that may cause unnecessary panic and fear, and potential unrest. Multimedia hotlines should be established to share factual information and timely updates. Educational information pertinent to the medical countermeasures and dispensing sites should be clearly stated. Finally, care must be taken to ensure that vulnerable and special need populations are served appropriately.

SUMMARY

Developing and maintaining a robust communication infrastructure among hospitals, EMS, and other health-care facilities, as well as private medical transport companies and medical supplies vendors, is vital to providing relief operations and services during emergency situations. Such infrastructure provides the core foundation for basic knowledge-sharing such as patient volume and severity, emergency room capacity, hospital bed availability; special wards availability; medical personnel specialties; transport vehicle availability; and blood, medicine, and medical supplies inventories. During an emergency situation, such critical information can be communicated to a regional coordinating control center, which can assess available resources, identify surpluses and shortages, and coordinate distribution efforts. In fact, building capacity for an interoperable communication system for emergency response is identified as one of the areas that must be prioritized according to The Hospital Preparedness Program established by the US Department of Health and Human Services. The September 11 terrorist incident showcases the importance of redundancy of communication infrastructure.

Multifaceted communication infrastructures are needed for effective medical and emergency response. Two-way communication lines must be kept open between emergency/medical responders and commanders for coordination and facilitation of the response effort; broadcast communication lines must be available to inform the general public about medical precautions and available services; and phone service and call answering stations must be maintained to allow the public to report critical emergency situations. Further, each state must strive to establish an operational redundant communication system that is capable of communicating both horizontally between health-care providers and vertically with the jurisdiction incident command.

Communication tools stockpiled must be tested and emergency personnel must learn to use them properly. Communication-facilitating programs are invaluable tools, and public health and emergency health organizations must enroll in them and learn to use their functionalities for practical emergency response execution. Organization must cultivate the atmosphere of proper emergency communication training. Sociological and human factors can play a critical role in the success of such communication infrastructure.

Communication to the public is an integral feature for mounting successful response. Critical components of the public information and risk communication plan are to inform, educate, and reassure those most impacted by a public health emergency in a timely manner. During a large-scale emergency, public fear and anxiety may create a challenge to dispensing preventive medication (prophylaxis) and following guidelines to those individuals in need. An effective risk communications plan informing and reassuring the public will reduce fear and anxiety, and will be crucial for earning public confidence and cooperation for distributing and dispensing medication.

Acknowledgment

Part of the material and results reported herein are based on our work in this area and interaction with public health agencies,

and discussion with many state and federal public health and emergency response experts. We also acknowledge the funding from the CDC and the State of Georgia Department of Human Resources.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

1. Kapucu N. Interagency communication networks during emergencies: boundary spanners in multiagency coordination. *Am Rev Public Adm* 2006;36:207–225.
2. Becker SM. Emergency communication and information issues in terrorist events involving radioactive materials. *Biosecur Bioterror* 2004;2:195–207.
3. Manoj B, Baker AH. Communication challenges in emergency response. *Commun ACM* 2007;50:51–53.
4. Lee EK, Yang AY, Chinnappan SG, *et al.* Survey & analysis of emergency communication infrastructure among Georgia State Hospitals. Atlanta, Georgia: Georgia Division of Public Health; 2009.
5. Ammenwerth E, Buchauer A, Bludau B, *et al.* Mobile information and communication tools in the hospital. *Int J Med Inf* 2000;57:21–40.
6. Holzman TG. Computer-human interface solutions for emergency medical care. *Interact ACM* 1999;6:13–24.
7. Arnold JL, Levine BNMR. Information-sharing in out-of-hospital disaster response: the future role of information technology. *Prehosp Disaster Med* 2004;19:201–207.
8. Jennex ME. Modeling emergency response systems. HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences. Big Island, Hawaii: IEEE Computer Society; 2007. p. 22.
9. Jan R-H, Hwang F-JCS-T. Topological optimization of a communication network subject to a reliability constraint. *Reliab IEEE Trans* 1993;42:63–70.
10. Lee-IOM. A systems view of POD operations: integrating all the elements. Institute of Medicine Workshop “Medical Countermeasures Dispensing: Emergency Use Authorization”; 2009 Nov; Washington (DC): 2009;
11. Lee EK, Chen CH, Pietz F, *et al.* Modeling and optimizing the public health infrastructure for emergency response. *Interfaces* 2009;39(5): 476–490.
12. Lee EK, Smalley HK, Zhang Y, *et al.* Facility location and multi-modality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. *Int J Risk Assess Manage* 2009;12(2/3/4): 311–351.
13. Hegde R, Manoj BS, Rao BD, *et al.* Emotion detection from speech signals and its applications in supporting enhanced QoS in emergency response. Proceedings of the 3rd International ISCRAM Conference; 2006 May; Newark (NJ): 2006;

PUBLIC HEALTH, EMERGENCY RESPONSE, AND MEDICAL PREPAREDNESS IV: EMERGENCY EVACUATION

EVA K. LEE

Center for Operations Research in
Medicine and Healthcare, School of
Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

NSF I/UCRC Center for Health
Organization Transformation,
Atlanta, Georgia

School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

ANNA YANG YANG

Center for Operations Research in
Medicine and HealthCare

NSFI/UCRC Center for Health
Organization Transformation

School of Industrial and Systems
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

FERDINAND PIETZ

Strategic National Stockpile, Office for
Public Health Preparedness
Response, Centers for Disease
Control and Prevention, Atlanta,
Georgia

BERNARD BENECKE

Global Disease Detection and
Emergency Response, Center for
Global Health, Centers for Disease
Control and Prevention, Atlanta,
Georgia

INTRODUCTION

In the aftermath of the September 11 terrorist attacks and the dissemination of anthrax in 2001, the ability of the U.S. healthcare system to provide an effective and coordinated response and evacuation capability to

mitigate mass casualty or complex incidents came under intense scrutiny. More recently, the devastation caused by Hurricane Katrina and the mass disruption of public health and medical services along the Gulf Coast spotlighted the need for cohesive strategies that focus on mounting effective evacuation and response for catastrophic events. Determining the most effective evacuation plan for a large public facility requires in-depth analysis of multiple interdependent and competing factors. Determining the best routes, foreseeing potential problems, addressing the chaos/panic factor, and orchestrating the evacuation are all critical aspects that should be evaluated in a well developed disaster management plan. Encouragingly, an evacuation conducted after a bomb threat at the stadium Santiago Bernabeu in Madrid, Spain, where 80,000 people were able to complete evacuation in 7 min (in comparison to prior drill results estimated 15 min) shed some light on the potential of saving many lives by careful planning and proper evacuation responses [1].

In this article, we seek to understand the elements that influence evacuation planning, issues faced by decision makers, as well as methodologies from the field of operations research that are leading to more effective planning and execution.

We will examine evacuation at both a health-care facility level and a community level. The discussion will be limited to evacuation caused by emergency scenarios, and exclude war-time evacuation.

The section titled "Emergency Evacuation" first provides the formal definition of emergency evacuation. The section titled "Factors that Affect Emergency Evacuation Planning" briefly describes the factors that affect emergency evacuation planning, including the nature of the threat, the potential extent of damage, and the risk to patients, if it relates to a health-care facility. The section titled "Issues in Hospital Evacuation" describes issues related to hospital evacuation, including facility, people, and

supporting service. The evacuation behavior is summarized in the section titled “Evacuation Behavior.” The section titled “Operations Research Methodologies” provides an overview of some of the operations research methodologies employed in emergency evacuation. Challenges in this area are highlighted in the section titled “Challenges.”

EMERGENCY EVACUATION

The early work by Quarantelli [2] provides a comprehensive report on evacuation behavior based on existing literature and evacuation data available at the Disaster Research Center at the Ohio State University. Evacuation is defined as

“...the mass physical movement of people, of a temporary nature, that collectively emerges in coping with community threats, damages or disruptions.”

It involves movement of people and equipment from unsafe areas to relatively safe places.

The scope of an evacuation refers to the general number of people affected. The scope could grow over time depending on the nature of the event. The following are some examples of escalating scope of evacuations [3]:

- defense in place, where minor adjustments are made to accommodate the event, but essentially no one is moved;
- single department/floor/unit in a venue;
- section, that is, multiple floors/units within a single building;
- entire building to another location on campus;
- entire campus evacuation;
- citywide evacuation.

Factors that Affect Emergency Evacuation Planning

The first factor affecting evacuation planning is the nature of the threat itself. Depending on the geographic location and/or time of year, communities may have a different probability for facing different natural or

man-made threats. While cities located close to the coast might be mostly concerned with seasonal hurricanes, those located on or near tectonic faults must be prepared at all times for a possible earthquake. And in high density population areas, major concerns may involve fire, hazardous material spills, or even terrorist attacks.

A second factor concerns the potential extent of damage. In the case of a fire or chemical spill, the affected area is generally smaller than for natural disasters that hit an entire geographical region. In the latter case, various facilities in the region may all be affected, and thus a large-scale evacuation is triggered (e.g., city-wide evacuation). Thus, this factor has a direct impact on the scope of evacuation.

In the event of a health-care facility evacuation, a third factor to consider is the risk to patients. Risk to patients varies depending on patients' conditions, acuity, and nature of the event. For example, in an emergency involving loss of electric power, patients in need of ventilators are more vulnerable compared to patients with mobility issues; whereas an event involving structural damage to the building may require allocating more resources to transferring patients with less mobility.

Overall, the resources demanded for evacuation is a critical issue. Indeed, evacuation requires resources of types and levels very different from routine operations. Further, structural damage (e.g., to airports and roads) or severe conditions may prohibit the transportation of much-needed supplies to the affected population or in a timely manner.

Issues in Hospital Evacuation

Traditional hospital evacuation plans involve horizontal (moving to a safer location on the same floor) and vertical (moving to another floor that is unaffected by the event) evacuations. However, full hospital evacuation requires moving patients and staff and others to a safe haven in preparation for a move to another facility.

According to Continuum Health Partners [3], a comprehensive evacuation plan addresses three basic elements: *facility*, *people*, and *support services*. Facility involves

rapidly identifying areas of the hospital that require a high priority for evacuation, areas of vulnerability, and areas that have potential risk. People involves assessing, triaging, tracking, and reconciling patients, staff, visitors, and others as they move throughout the evacuation, which is the single most important aspect of a hospital evacuation plan. Support service addresses areas such as systematic shutdown of medical gases, support of utilities and generators, telecommunication systems (see ***Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure***), and operations logistics (see ***Public Health, Emergency Response, and Medical Preparedness II: Medical Countermeasures Dispensing and Large-Scale Disaster Relief Efforts***).

Facility. Facility issues involve rapidly identifying areas of the hospital that require a high priority for evacuation, areas of vulnerability, and areas that have potential risk.

Collaboration between nearby hospitals and health-care facilities are important in identifying facility problems and solutions. In the United States, many counties have their own emergency operation center to help facilitate communication, coordination, and transportation between facilities. However, it was reported in one study that hospitals are capable of communicating among themselves during emergency events and can successfully transfer a large number of patients into safe facilities [4].

People. Accounting for people, including patients, staff, visitors, and other personnel, is a core component of hospital evacuation planning.

Emergency situations present many difficulties that are not faced during regular operations. For example, in the event of electric power outage, basic tracking and updating of patient records may be difficult without availability of the computer system. Cases have been reported as early as the mid-1980s where patient records were not updated until days after the event, resulting in confusion among caretakers [5]. This problem can

become very acute when electronic medical records are in use throughout the facility. It is recommended that a paper record system be in place to record patient information, their attending physician information, accepting facility and accepting physician as well as discharges, diagnostic tests, procedures, equipment needed, mode of transport, time of departure and arrival for transferring [3,6].

It is also necessary to identify and prepare patients for evacuation off the in-patient unit by preparing necessary medical information, medical records, and medications for accepting facility to carry on continuous care when patients are transferred [3].

On the basis of patient record, a triage system must be in place to determine the urgency of each patient case and prioritize transferring of inpatients. It is often required to facilitate the evacuation process. To achieve the maximum number of evacuated patients in the shortest time, it is common practice that hospitals evacuate the most mobile patients first. However, when faced with limited resources, hospitals tend to transfer the most care-intensive patients first.

In terms of staff, research has suggested that each evacuation area should designate a Unit Evacuation Leader who should coordinate the evacuation of the entire area. The role of volunteers in emergency evacuation has been mentioned.

Support Service. Support service addresses processes such as systematic shutdown of medical gases, assessing and maintaining utilities, emergency generators, telecommunication systems, and emergency logistics.

It is possible that an emergency event arises from nonstructural damage scenarios such as a water or power outage. In this case, hospitals should have backup power system that can operate until patients can be transferred safely. (All business-critical computer systems, including financial and clinical systems, should of course have offsite backup.)

Normal communications could be disrupted with structural damage of the landline or cellular towers. Emergency communication infrastructure should be put in place to plan for unforeseeable

events (see *Public Health, Emergency Response, and Medical Preparedness III: Communication Infrastructure*).

A variety of transportation modes can be deployed. These include helicopters, buses, trains, ambulances, or family vehicles. The choice of different transportation modes largely depends on the condition of the patients. Decisions have to be made based on an assessment of the potential benefit of transport versus the potential risks.

Evacuation on this scale is expensive, as it entails extra staffing, EMS utilization, and overnight housing. In the decision making process, a cost–benefit analysis is often used in evaluating different evacuation plans. However, it is difficult to put a monetary value on patients' lives. Just as with many events within an emergency response, the key lies in optimizing the operations under limited resources.

Evacuation Behavior

For general community evacuation, one critical element concerns evacuation behavior, including factors that affect decision making and traffic pattern of evacuees.

There has been mounting interest in gaining a better knowledge of how individuals in emergency situations behave when evacuating from a venue. One study suggests that a person's total evacuation process can be divided into four phases: assessment of the situation, movement, path-finding, and evacuation [7]. The variations with respect to time of individuals' progression through different phases of evacuation were observed in a study of the 1993 bombing of the World Trade Center [8]. People in a panic-causing emergency typically show stubborn and sub-optimal behavior, such as over-crowding and congestion [9]. Their research shows that panic may cause people to become oblivious to alternative (preferable) evacuation routes as their instincts are to follow the crowd and get to the head of the pack. The effect of structural parameters on individuals' ability to successfully egress has also been studied. Research shows that egression time increases as corridor complexity increases, and decreases if clear signage is provided [7]. Although emergency signs are

believed to play an important role in ensuring public safety in facilities during emergencies, in practice, specific and clear standards for placing emergency signs have not yet been established.

Much research has been done on evaluating evacuation decision making. Fischer *et al.* [10] investigated the factors and circumstances in which citizens would actually evacuate when receiving a warning from the authority. On the basis of survey data collected after a fire incident at Hamilton Distributing Company in Ephrata, Pennsylvania on May 1, 1990, it was found that residents are most likely to evacuate if they are ordered to do so, if they are contacted frequently by authority, if past warnings were accurate, and if dependent children were at home.

Lindell *et al.* [11] collected survey data from 2002 Hurricane Lili evacuation from some counties in Louisiana and Texas to investigate factors that affect evacuation decision for households. Some of the major findings are as follows:

- local news media is the most extensive source of information for hurricane development;
- the environmental cue is the most important factor in making an evacuation decision, followed by personal experience and social cues;
- evacuation impediments include job requirement and traffic congestion.

Drabek [12] looked at the difference between public response versus response made by business executives of tourist firms when making a disaster evacuation decision. It was found that most business executives have evacuation plans although they may vary greatly in term of preparedness. Managerial decision making is parallel to public response. Five key differences between public response and business response are influence and planning, firm versus family priorities, shelter selection, looting concerns, and media contacts.

Cohn *et al.* [13] interviewed both residents and public officials from western United States on the issues and problems they are facing when dealing with wildfire evacuation.

A contrast of interests was found between the two parties when making evacuation decisions through stages. For example, during anticipation stage, the public officials' greatest interest lies in ensuring safety of the entire community. Residents tend to gauge the seriousness of the threat, preparing appropriately and at the same time maintaining a normal life. This contrast in interest pertains throughout all stages of evacuation: anticipation, warning, displacement, notification, and return and recovery. Keeping these differences in mind may better assist authorities for designing effective evacuation planning.

Aguirre [14] surveyed the population in Cancun, Mexico during Hurricane Gilbert in 1988 and identified factors that affected their evacuation behavior. A post disaster survey was conducted one week after the impact and collected responses from 431 individuals, among whom 25% of them were evacuated during the impact. Family and friends tended to be the first stop shelter evacuees sought. The survey result also demonstrated that environmental context and physical characteristics of residence greatly impacted the perception of risk, which subsequently influenced the evacuation behavior. Other factors, such as socioeconomic factors have also been studied.

Other researchers have been concerned with the logistics aspects of evacuation, such as cost and transportation. Dow and Cutter [15] examined the demand of transportation infrastructure during the 1999 Hurricane Floyd evacuation. It was found that about 65% of South Carolina residents decided to evacuate. The study revealed three factors that contributed to the traffic congestion on interstate highways: (i) 25% of families took two or more cars for evacuation; (ii) only half of those with a map made an effort to find alternative routes instead of staying on the interstate; and (iii) majority of evacuees made extra trips that were unnecessary in order to get away from the hurricane.

Evacuation without guidance on timeline and traffic routes often proves to be deficient. Even in well-planned scenarios, as in the September 2005 evacuation of Texas citizens from Hurricane Rita, major problems can

ensue. Although the governor had announced a staggered schedule of evacuation for various districts and population zones within a 24-h period, well in advance of the storm's possible landfall later in the week, it was still not early enough to ensure that all residents could evacuate safely in advance of the storm. This reinforces the key elements of emergency response—that careful planning with procedural and policy guidelines must be in place, and that improvised and dynamic reconfiguration must be executed as the event unfolds.

OPERATIONS RESEARCH METHODOLOGIES

Much research has been done on emergency evacuation simulations. Usually, the focus is on one of the following four aspects of emergency evacuation: simulation methodology, human factors, model scale, and model speed.

Zheng *et al.* [16] studied seven methodological approaches: cellular automata models, lattice gas models, social force models, fluid-dynamic models, agent-based models, game theoretic models, and approaches based on experiments with animals for modeling crowd evacuation from a building. The paper compared and contrasted these approaches and discussed their associated advantages and disadvantages. The authors concluded that a variety of approaches should be combined to more accurately model crowd evacuation. Chen *et al.* [17] performed a study using agent-based simulation to model and analyze hurricane evacuation procedures in the Florida Keys. Taaffe *et al.* [18] described a simulation model for assessing hospital evacuation under emergency situations, such as a hurricane. They designed experiments based on resource utilization, such as number of staff and vehicles used in the model. Different scenarios were tested to identify the bottleneck of the described system.

It is important that all occupants should be able to evacuate to safety from large public buildings under fire conditions. In the paper by Shi *et al.* [19,20], a system simulation model was presented, in which a physical model and a mathematical model were both

included. On the basis of agent technology, a computer program was developed to simulate and analyze the egress progress in large public buildings through combining rule reasoning with numerical calculation, and some crowd pedestrian flow phenomenon. By coupling with the fire scenario simulated by CFD technology, the computer simulation program may represent the overall and dynamic process of occupants' evacuation under fire expansion, and the mutual relationship between occupants' safety and fire hazard. An indoor stadium which was used as a competition venue for the 2008 Beijing Olympic Games was studied.

Since the strategies and factors influencing evacuation differ greatly based on the size of the venue, much research has been conducted for both, small physical parameters and large-scale parameters such as geographical layouts. Simulating large area evacuation usually requires more parameters and data. The physical characteristics of a structure affect the success of an evacuation. It has been shown in different studies that building environment, involving exit position, exit width, obstacle distribution, slope of stairs, and so on, significantly affect the emergency egress time [19–21]. Also, according to Wong and Fong [22], the number of occupants, the ratio of floor space to the number of occupants, and the specific flow rate at the exit are significant factors of the risk to people during an evacuation.

Saadatseresht *et al.* [23] considered the traveling distance between a safe area and the point of evacuation to be an important parameter in building a multiobjective optimization model for constructing a city-wide evacuation plan. Pidd *et al.* [24] developed a decision support system that combines a geographical information system (GIS) with microsimulator tools. It can model the mitigation of the public in a certain area under emergency threats such as chemical leak. However, the system is computationally intensive and is not suitable for real-time simulation or scenario-based decision analysis.

There has also been much discussion focused on model speed, because the computational time of an emergency evacuation

planning model is a key factor of its functionality. Many of the models developed for planning emergency evacuations are of high computational complexity, which prohibits them from providing real-time decision support capability. Two examples of these systems are the multiobjective optimization model developed by Saadatseresht *et al.* [23], and the spatial decision support systems developed by Pidd [24] and de Silva and Eglese [25]. Microsimulators, which capture the detailed movement of individual entities, are usually computationally expensive, but these models allow integration of real-life factors. On the other hand, macrosimulators are inadequate for tracking detailed behavior of individual agents, but they are less computationally demanding than microsimulators [24]. Few models have been developed for real-time simulation. Papamichail and French [26] developed a real-time on-line decision support system (RODOS) that assists in the evaluation and ranking of alternative strategies in order to specify countermeasures. A coarse expert system (CES) was developed at the University of Manchester to cut down the large number of strategies that may arise in nuclear emergencies.

Bakuli and Smith [27] used state-dependent queueing network models incorporating the mean value analysis algorithm within Powell's derivative-free unconstrained optimization algorithm to design an emergency evacuation network. This method helps to specify design parameters for circulation space that would accommodate the requirement for both normal circulation and an emergency.

Chen *et al.* [28] developed a heuristic algorithm based on the Lagrangian relaxation for optimizing emergency sign locations with consideration of light-occlusion effects. On the basis of a cellular automata model, the evacuation efficiency of the optimized locations of emergency signs derived from the algorithm was verified. Through simulation to compare evacuation efficiency of the same supermarket with and without existing emergency signs, it showed that by enlarging the coverage of emergency signs, evacuation efficiency can be improved.

The article by Hobeika and Kim [29] described a software tool, MASSVAC 4.0, developed to model the evacuation of an area near a nuclear power plant. Although radiation exposure is a very rare event, the Nuclear Regulatory Commission requires detailed evacuation plans for all plants. The article described several generations of evacuation software designed to analyze, evaluate, and develop different evacuation strategies. The software combines a traffic-assignment module, a simulation module, and a user-interface module. The most recent addition to the software is a traffic-assignment method based on user equilibrium.

A crowd simulation system, MACES, was developed at the University of Pennsylvania that integrates a psychological model for emergency behaviors [30]. This crowd simulation system was capable of a high level way-finding that explores unknown environment to build knowledge of the surrounding areas for navigation. The model was capable of dealing with low level motion within each room, based on social forces. Communication and roles were also included to simulate information spread during emergencies. To expand the range of realistic human behaviors, a system called PMFserv were also included. The two systems were integrated to add validated agent behavior to crowd simulation.

Samuelson *et al.* [31] described real-time models that simulate mass egress of a stadium and a subway station. The models were built based on NetLogo, a simulation model platform. An identified problem with subway simulation was that adding too many choices of crowd behavior could push the model beyond what NetLogo can handle with acceptable execution times.

Luo *et al.* [32] developed a model that adopts an agent-based approach and employs a layered framework to reflect the natural pattern of a human-like decision making process, which generally involves a person's awareness of the situation and consequent changes on the internal attributes.

Chen and Zhan [33] used an agent-based technique to model traffic flows at the level of individual vehicles and investigated the collective behaviors of evacuating vehicles.

The problem with network flow models is that they cannot capture the behavior of individual vehicles. They compared the performance of simultaneous and staged evacuation strategies, but found that performance of the strategies depend on both road network structure and population density.

Shendarkar *et al.* [34] presented a novel VR (Virtual Reality)-trained BDI (belief, desire, intention) software agent used to construct crowd simulations for emergency response. The BDI framework allows modeling of human behavior with a high degree of fidelity. The proposed simulation has been developed using AnyLogic software to mimic crowd evacuation from an area under a terrorist bomb attack. The attributes that govern the BDI characteristics of the agent were studied by conducting human in-the-loop experiments in VR using the Cave Automatic Virtual Environment. To enhance the generality and interoperability of the crowd simulation modeling scheme, input data models have been developed to define environment attributes. Experiments were also conducted to demonstrate the effect of various parameters on key performance indicators such as crowd evacuation rate and densities. However, this model was not real-time.

Agent routing in E-SIM is computed using a two-level hierarchical path-finding approach based on the A* algorithm, a tree search algorithm that finds a path from a given initial node to a given goal node. It employs a heuristic estimate which ranks each node by an estimate of the best route that goes through that node. Each agent also has the ability to remember problems or obstructions that they encountered while attempting to exit the facility [36].

CHALLENGES

Evacuating large crowds of people under any circumstance is a challenge. Evacuation of large facilities during a disaster or an emergency further complicates the complex task because of the added elements of chaos, panic, the high density of the population, and the potential dependencies of supporting services. As terrorists move more toward

preying on soft targets such as airports, train stations, sport stadiums, and public theater districts, the potential for large government, commercial, residential, and sports facilities to become targets become a reality for which we must prepare. Further, such preparedness is necessary as we have seen the devastation resulting from inadequate planning and training for mass evacuations. The 2005 Hurricane Katerina killed at least 1836 people in the actual hurricane and in the subsequent floods, making it the deadliest US hurricane since the 1928 Okeechobee hurricane, with total property damage estimated at \$81 billion (2005 USD) [35], nearly triple the damage wrought by Hurricane Andrew in 1992.

Determining the most effective evacuation plan for a large public facility requires in-depth analysis of multiple interdependent and competing factors. Determining the best routes, foreseeing potential problems, addressing the chaos/panic factor, and orchestrating the evacuation are all critical aspects that should be evaluated in a well-developed disaster management plan. The success of the evacuation of 80,000 at the stadium Santiago Bernabeu in Madrid, Spain within seven min sheds some light on the potential to save many lives through careful planning and proper evacuation responses. Emergency responders and coordinators must be equipped with such knowledge so as to protect the regional population during critical incidents [1].

In the academic and research frontier, challenges remain in the trade-offs between the realism of the models that can accommodate the multifaceted complexity of the evacuation process versus their computational intractability. Researchers have to continue to push the frontier of computational strategies that can solve large-scale evacuation instances where they allow for scenario-based analysis and development of useful and practical evacuation guidance for maximizing evacuation success.

Acknowledgement

We acknowledge the funding from the CDC. The findings and conclusions in this report are those of the authors and do not

necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

1. U.S. Department of Homeland Security, Mass Evacuations. Planning for sports venues. Available at <http://www.cops.usdoj.gov/files/ric/CDROMs/PlanningSecurity/modules/10/module%2010%20handout%20%20lessons.pdf>. Accessed 2010 Jul 14.
2. Quarantelli EL. Evacuation behavior and problems: findings and implications from the research literature. Columbus (OH): Disaster Research Center, Ohio State University; 1980.
3. Continuum Health Partners. Evacuation planning for hospitals (Draft Document). Center for Bioterrorism Preparedness and Planning; 2006.
4. Schultz CH, Koenig KL, Lewis RJ. Implications of hospital evacuation after the Northridge, California, earthquake. *N Engl J Med* 2003;348:1349–1355.
5. Blaser MJ, Ellison RTI. Rapid nighttime evacuation of a veterans hospital. *J Emerg Med* 1985;3:387–394.
6. Cocanour CS, Allen SJ, Mazabob J, *et al.* Lessons learned from the evacuation of an urban teaching hospital. *Arch Surg* 2002; 137:1141–1145.
7. Notake H, Ebihara M, Yashirob Y. Assessment of legibility of egress route in a building from the viewpoint of evacuation behavior. *Saf Sci* 2001;38:127–138.
8. Fahy RF, Proulx G. Human behavior in the world trade center evacuation. *Fire Saf Sci* 1997;5:713–724.
9. Helbing D, Farkas I, Vicsek T. Simulating dynamical features of escape panic. *Nature* 2000;407:487–490.
10. Fischer HW, Stine GF, Stoker BL, *et al.* Evacuation behaviour: why do some evacuate, while others do not? A case study of the Ephrata, Pennsylvania (USA) evacuation. *Disaster Prev Manage* 1995;4(4): 30–36.
11. Lindell MK, Lu J-C, Prater CS. Household decision making and evacuation in response to Hurricane Lili. *Nat Hazards Rev ASCE* 2005;6(4):171–179.
12. Drabek TE. Variations in disaster evacuation behavior: public responses versus private

- sector executive decision-making processes. *Disasters* 1992;16(2):104–118.
13. Cohn PJ, Carroll MS, Kumagai Y. Evacuation behavior during wildfires: results of three case studies. *West J Appl Forest* 2006;21(10):39–48.
 14. Aguirre BE. Evacuation in Cancun during Hurricane Gilbert. *Int J Mass Emerg Disaster* 1991;9:31–45.
 15. Dow K, Cutter SL. Emerging hurricane evacuation issues: Hurricane Floyd and South Carolina. *Nat Hazards Rev ASCE* 2002;3:12–18.
 16. Zheng X, Zhong T, Liu M. Modeling crowd evacuation of a building based on seven methodological approaches. *Build Environ* 2009;44:437–445.
 17. Chen X, Meaker JW, Zhan FB. Agent-based modeling and analysis of hurricane evacuation procedures for the Florida keys. *Nat Hazards* 2006;38:321–338.
 18. Taaffe K, Johnson M, Steinmann D. Improving Hospital Evacuation Planning Using Simulation. *Proceedings of the 38th Conference on Winter Simulation*; 2006. pp. 509–515.
 19. Shi J, Ren AZ, Chen C. Agent-based evacuation model of large public buildings under fire conditions. *Automat Constr* 2009;18:338–347.
 20. Shi L, Xie Q, Cheng X. Developing a database for emergency evacuation model. *Build Environ* 2009;44:1724–1729.
 21. Graat E, Midden C, Bockholts P. Complex evacuation; effects of motivation level and slope of stairs on emergency egress time in a sports stadium. *Saf Sci* 1999;31:127–141.
 22. Wong L, Fong N. Risk analysis of escape time from buildings. *Facilities* 2005;23:487–495.
 23. Saadatseresht M, Mansourian A, Taleai M. Evacuation planning using multiobjective evolutionary optimization approach. *Eur J Oper Res* 2009;198:305–314.
 24. Pidd M, de Silva FN, Eglese RW. A simulation model for emergency evacuation. *Eur J Oper Res* 1996;90:413–419.
 25. de Silva FN, Eglese RW. Integrating simulation modeling and GIS: spatial decision support systems for evacuation planning. *J Oper Res Soc* 2000;51:423–430.
 26. Papamichail KN, French S. Generating feasible strategies in nuclear emergencies—a constraint satisfaction problem. *J Oper Res Soc* 1999;50:617–626.
 27. Bakuli DL, Smith JM. Resource allocation in state-dependent emergency evacuation networks. *Eur J Oper Res* 1996;89:543–555.
 28. Chen CX, Li Q, Kaneko S, *et al.* Location optimization algorithm for emergency signs in public facilities and its application to a single-floor supermarket. *Fire Saf J* 2009;44:113–120.
 29. Hobeika AG, Kim C. Comparison of traffic assignments in evacuation modeling. *IEEE Trans Eng Manage* 1998;45:192–198.
 30. Pelechano N, O'Brien K, Silverman B, *et al.* Crowd simulation incorporating agent psychological models, roles and communication. Philadelphia (PA): Center for Human Modeling and Simulation, University of Pennsylvania; 2005.
 31. Samuelson DA, Zimmerman A, Thorp J, *et al.* Panel: Agent-based Modeling of Mass Egress and Evacuations. *Proceedings of the 2007 Winter Simulation Conference*; 2007.
 32. Luo L, Zhou S, Cai W, *et al.* Agent-based human behavior modeling for crowd simulation. *Comput Animat Virtual Worlds* 2008;19:271–281.
 33. Chen X, Zhan FB. Agent-based modeling and simulation of urban evacuation: relative effectiveness of simultaneous and staged evacuation strategies. *J Oper Res Soc* 2008;59:25–33.
 34. Shendarkar A, Vasudevan K, Lee S, *et al.* Crowd simulation for emergency response using BDI agent based on virtual reality. WSC '06: *Proceedings of the 38th Conference on Winter simulation, Winter Simulation Conference*; 2006. pp. 545–553.
 35. Knabb RD, Rhone JR, Brown DP. Tropical cyclone report: Hurricane Katrina: 23–30 August 2005 (PDF). National Hurricane Center; (Retrieved 2006-05-30. December 20, 2005; updated August 10, 2006). http://www.nhc.noaa.gov/pdf/TCR-AL122005_Katrina.pdf.
 36. Smith JL. Agent-based simulation of human movements during emergency evacuations of facilities. Available at http://www.wbdg.org/pdfs/agent_based_sim_paper.pdf. Accessed 2010 Jul 14.

PURE CUTTING-PLANE ALGORITHMS AND THEIR CONVERGENCE

DINAKAR GADE
Department of Industrial and
Manufacturing Systems
Engineering, Iowa State
University, Ames, Iowa

SIMGE KÜÇÜKYAVUZ
Department of Integrated
Systems Engineering, The
Ohio State University,
Columbus, Ohio

INTRODUCTION

This survey article addresses the solution of the mixed-integer program (MIP)

$$z = \min_{x \in X} \left\{ c^\top x : X = \left\{ Ax \geq b, x \in \mathbb{Z}_+^p \times \mathbb{R}_+^{n-p} \right\} \right\}, \quad (1)$$

by means of adding valid linear inequalities (also referred to as *cutting planes* or *cuts*) to its *linear programming* (LP) relaxation. The LP relaxation of Equation 1 is given by,

$$z_\ell = \min_{x \in X_\ell} \left\{ c^\top x : X_\ell = \left\{ Ax \geq b, x \in \mathbb{R}_+^n \right\} \right\}. \quad (2)$$

Here $c \in \mathbb{Q}^n$, $A \in \mathbb{Q}^{m \times n}$, and $b \in \mathbb{Q}^m$. A *valid inequality* for MIP Equation 1 is an inequality $\pi^\top x \geq \pi_0$, where $(\pi, \pi_0) \in \mathbb{R}^{n+1}$ such that $X \subseteq \{x : \pi^\top x \geq \pi_0\}$. We refer the reader to Marchand *et al.* [1], Cornuéjols [2], and other articles in the encyclopedia for an introduction to valid inequalities for structured and unstructured MIPs.

A fundamental result in MIP (see, e.g., Theorems 6.2 and 6.3 Chapter I.4 in Nemhauser and Wolsey [3]) states that solving problem 1 is equivalent to solving the LP:

$$z = \min \left\{ c^\top x : x \in \text{clconv}(X) \right\},$$

where $\text{clconv}(X)$ denotes the closure of the convex hull of X . Moreover, it can be shown that $\text{clconv}(X)$ is a polyhedron. However, constructing a linear description of $\text{clconv}(X)$ *a priori* is a difficult task and instead, the cutting-plane approach iteratively constructs tighter linear approximations of $\text{clconv}(X)$. Given a family of valid inequalities $\Pi := \{(\pi, \pi_0) \in \mathbb{R}^{n+1} : X \subseteq \{x : \pi^\top x \geq \pi_0\}\}$, a pure *cutting-plane* method for solving MIPs attempts to find a valid inequality (π, π_0) for X that cuts off the fractional solution $\bar{x}$ of the LP relaxation, that is, $\pi^\top \bar{x} < \pi_0$. The problem of finding a violated inequality is referred to as *separation*. A generic cutting-plane algorithm for a given family of valid inequalities Π for X is given in Algorithm 1.

The algorithm begins with using the linear relaxation X_ℓ of X as an initial approximation of $\text{clconv}(X)$. In an iteration k of the algorithm, an LP corresponding to the k th approximation, X_ℓ^k is solved and the solution vector x^k is obtained. If x^k satisfies the mixed-integer requirement, then, because x^k is an optimal solution to a relaxation of Equation 1, and feasible in X , it is also optimal. If this is not the case, then the algorithm proceeds by choosing a valid inequality $\pi^\top x \geq \pi_0$ that is violated by the current solution with $\pi^\top x^k < \pi_0$ (assuming that it exists in the given family), and the subsequent relaxation X_ℓ^{k+1} is formed by adding the valid inequality to X_ℓ^k . These steps are repeated until a mixed-integer feasible solution is found or until no valid inequalities to separate x^k can be found.

Algorithm 1 A Generic Cutting-Plane Algorithm for MIP

- 1: *Initialization.* $k \leftarrow 0, X_\ell^k \leftarrow X_\ell$.
 - 2: **while** $x^k := \arg \min_{x \in X_\ell^k} c^\top x \notin \mathbb{Z}_+^p \times \mathbb{R}_+^{n-p}$
 - 3: *Separation.* Find an inequality $(\pi, \pi_0) \in \Pi$ such that $\pi^\top x^k < \pi_0$. If none exists, goto 6
 - 4: $X_\ell^{k+1} \leftarrow X_\ell^k \cap \{x : \pi^\top x \geq \pi_0\}$.
 - 5: $k \leftarrow k + 1$.
 - 6: **end while**
-

A natural question that arises regarding a cutting-plane algorithm is whether it converges to an optimal solution of MIP (if it exists) in finitely many steps. This question is important not only in the context of MIP but also for designing algorithms for stochastic MIPs. In this article, we survey cutting-plane algorithms for subclasses of MIPs and discuss the key results needed to show their convergence for different classes of MIPs. We also give examples to show nonconvergence when applied to other classes of problems. We begin with a discussion of the Gomory cutting-plane method and its convergence for pure integer programs.

ALGORITHMS USING LEXICOGRAPHY AND ROUNDING

First, we discuss the Gomory cutting-plane algorithm [4], which is one of the first cutting-plane algorithms developed to solve pure integer programming problems ($p = n$ in Eq. 1). Gomory cuts and extensions are discussed elsewhere in this encyclopedia (please refer to the encyclopedia article by R. Fukasawa).

We make the following assumptions:

- The data of the MIP are integral, that is, $c \in \mathbb{Z}^n, A \in \mathbb{Z}^{m \times n}, b \in \mathbb{Z}^m$. This assumption is without the loss of generality because the original rational coefficients can be appropriately scaled to obtain an equivalent representation of X using integer coefficients.
- The problem is stated equivalently by minimizing the variable x_0 and adding the constraint $x_0 = c^\top x$ to the constraint set.

For the sake of completeness, we derive the Gomory fractional cut in this article. Let $\bar{x}$ be the optimal solution to the LP relaxation (Eq. 2), let $\mathcal{B}$ and $\mathcal{N}$ denote the index set of basic and nonbasic variables associated with this optimal solution, respectively, and let B and N denote the columns of basic and nonbasic variables, respectively. If $\bar{x} \in \mathbb{Z}_+^n$, then we have found an optimal solution to the MIP. Otherwise, we choose a row of the

simplex tableau corresponding to a variable $i \in \mathcal{B}$, such that $\bar{x}_{\mathcal{B}(i)} \notin \mathbb{Z}$:

$$x_{\mathcal{B}(i)} + \sum_{j \in \mathcal{N}} \bar{a}_{ij} x_j = \bar{b}_i, \quad (3)$$

where $\bar{a}_{ij}$ is a component of the matrix $B^{-1}N$, $\mathcal{B}(i)$ is the i th basic variable and $\bar{b}_i := \bar{x}_{\mathcal{B}(i)}$. Because $x_j \geq 0$ for all j , we can write

$$x_{\mathcal{B}(i)} + \sum_{j \in \mathcal{N}} \lceil \bar{a}_{ij} \rceil x_j \geq \bar{b}_i.$$

Furthermore, because $x_j \in \mathbb{Z}$ for all j , we have,

$$x_{\mathcal{B}(i)} + \sum_{j \in \mathcal{N}} \lceil \bar{a}_{ij} \rceil x_j \geq \lceil \bar{b}_i \rceil. \quad (4)$$

Subtracting Equation 3 from Equation 4, we get the Gomory fractional cut as

$$\sum_{j \in \mathcal{N}} \phi(\bar{a}_{ij}) x_j \geq \phi(\bar{b}_i), \quad (5)$$

where $\phi(\beta) := \lceil \beta \rceil - \beta$. Note that the current basis and solution is cut off by Equation 5. An iteration of Gomory's cutting-plane algorithm involves solving an LP associated with the current iteration. If the solution is fractional, then a cut in the form (5) is generated, which is guaranteed to cut off this fractional solution.

The proof of convergence of Gomory's fractional cutting-plane algorithm relies on the lexicographic dual simplex method [3, 7, 8] (see the article by M. Banciu in the encyclopedia for a description of the lexicographic dual simplex method). Given two vectors $v^1, v^2 \in \mathbb{R}^n$, v^1 is said to be *lexicographically larger* than v^2 , denoted as $v^1 \succ v^2$, if there exists a k such that $v_k^1 > v_k^2$ and $v_j^1 = v_j^2$ for all $j = 1, \dots, k-1$. We say that v^1 is lexicographically larger than or equal to v^2 , denoted by $v^1 \succeq v^2$, if either $v^1 \succ v^2$ or $v^1 = v^2$. In the lexicographic dual simplex method, one begins with and maintains simplex tableaus that are lexicographically dual feasible, that is, tableaus that are dual feasible and have all the nonbasic columns lexicographically smaller than the zero vector. This is accomplished using a

lexicographic pivoting rule. This approach ensures that the simplex method does not cycle, and furthermore, guarantees that the lexicographically smallest optimal solution to the LP, if it exists, is found in *finitely many steps*.

In line 3 of Algorithm 1, when one uses the Gomory fractional cutting-plane method for pure integer programs, one solves the k th relaxation using the lexicographical dual simplex method. Because $X_\ell^{k+1} \subset X_\ell^k$ for each iteration k with a fractional solution x^k , we obtain a sequence of lexicographically increasing solutions $\{x^t\}_{t=0}^k$ during the execution of the cutting-plane procedure. What remains to be seen is whether this sequence converges to an integer solution in finitely many steps.

It is a well-known fact that if $\hat{x}$ is an extreme point of $\text{conv}(X)$, then there exists $M < \infty$ such that $\hat{x}_j \leq M$ with $M \in \mathbb{Z}$ for $j = 1, \dots, n$ (see Nemhauser and Wolsey [3], Theorem 4.1 in Chapter I.5). That is, the extreme point components of the convex hull of an integer set within a polyhedron are bounded. Moreover, the bound depends only on the problem data. Let $\beta^- := \min(\beta, 0)$ and $\beta^+ := \max(0, \beta)$. From the preceding fact, it follows that if $x \in X$, then

$$\begin{aligned} \left(\sum_{j=1}^n c_j^+ M, M, \dots, M \right)^\top &\succeq x \\ &\succeq \left(\sum_{j=1}^n c_j^- M, 0, \dots, 0 \right)^\top. \end{aligned} \quad (6)$$

The following result establishes the fact that Gomory cuts, when used with the lexicographic dual simplex method and when cuts are chosen from the smallest index variable that is fractional in the solution, have a certain “rounding” property.

Proposition 1 [8]. *Let x^t denote the lexicographically smallest optimal solution to $\min_{x \in X_\ell^t} c^\top x$, $t = k-1, k$, where the updated approximation X_ℓ^k is formed by adding to X_ℓ^{k-1} a Gomory cut (Eq. 5) corresponding to the smallest index source row of the fractional solution x^{k-1} . Let i_k denote the index of the*

variable used to form the Gomory cut and define

$$\alpha^k := \left(x_0^{k-1}, x_1^{k-1}, \dots, x_{i_k-1}^{k-1}, \lceil x_{i_k}^{k-1} \rceil, 0, \dots, 0 \right)^\top. \quad (7)$$

Then,

$$x^k \succeq \alpha^k \succeq x^{k-1}.$$

It is important to note that the fractional variable with the smallest index is chosen for cut generation. Finite convergence of the Gomory cutting-plane method follows because from Proposition 1, we obtain a sequence of lexicographically increasing vectors at each iteration, and there are only finitely many integer vectors that satisfy Equation 6.

Theorem 1 [3, 8]. *Suppose that at every iteration of the Gomory cutting-plane method, the cut is formed by choosing the smallest index source row of the fractional solution, and the lexicographic dual simplex method is used to optimize the updated problems. Then, the cutting-plane method finds an optimal solution or concludes that the problem is infeasible in finitely many steps.*

Note that as long as the Gomory cut from the smallest index source row is present, lexicography allows other valid inequalities to be added to the subsequent approximation X_ℓ^{k+1} of the Gomory cutting-plane method without losing finite convergence. For example, one can add as many Gomory cuts as there are fractional variables in the solution to the current approximation. This approach of adding cuts corresponding to all fractional variables in the linear relaxation is referred to as adding a *round of cuts*.

For the Gomory cutting-plane algorithm, the key elements needed to show convergence are lexicography and the choice of the source row for cut generation. The nature of convergence can be interpreted as pruning nodes in a lexicographic enumeration tree (cf. [3]). In this tree, the nodes at level j correspond to an index of a variable. For example, the root node corresponds to the objective function

variable x_0 and there are as many branches as there are possible values of the variable. All possible integer vectors satisfying Equation 6 are arranged in a lexicographic order. In an iteration k , any vector in the tree that is strictly lexicographically smaller than α^k can be thought of as pruned. Thus, by means of this tree, lexicography implicitly provides a “memory” of vectors pruned. We shall see that the concept of “memory” is a recurrent theme in the convergence proofs of various cutting-plane algorithms throughout this article.

Lexicography has also been used in other finitely convergent cutting-plane algorithms. Neto [9] proposes an iterative lexicographic optimization method for the solution of MIPs. The author shows that under the assumptions that the optimal objective function is integral and bounded, and the integer variables are bounded, the algorithm converges in finitely many iterations to the lexicographically smallest optimal solution. Unlike Gomory’s cutting-plane method, the cutting planes in this algorithm are not derived from the simplex tableau of the LP relaxation. They are not necessarily valid for the convex hull of solutions to the MIP, but they cut off a fractional LP solution, while ensuring that any lexicographically smaller integer feasible solution is not cut off. This algorithm is a generalization of the algorithm in [10], which gives a finite cutting-plane algorithm for mixed-binary problems with integral optimal objective function values. Other finitely convergent algorithms for pure integer programs are proposed in [11, 12, 13].

Gomory also proposed the Gomory mixed integer (GMI) cuts [14] for general MIP problems. When the objective function value is known to be integer, the proof for the pure integer case can be adapted to show finite convergence for the mixed-integer case. However, if the objective function value is no longer required to be integer, a cutting-plane algorithm based on GMI cuts may not converge. The following example [15] (see also Padberg [16]) is an instance on which GMI

cuts do not converge.

$$\begin{aligned} \min & -x_3 \\ & -x_1 - x_2 - x_3 \geq -2 \\ & x_1 - x_3 \geq 0 \\ & x_2 - x_3 \geq 0 \\ & x_1, x_2 \in \mathbb{Z}_+, x_3 \in \mathbb{R}_+. \end{aligned} \tag{8}$$

One can easily observe that $x_3 = 0$ in an optimal solution to this example. When one does not consider the objective function row corresponding to x_3 for cut generation, Padberg [16] shows that the optimal solution at the k th iteration of the GMI cutting-plane method (when a round of cuts is added in every iteration) is $x_1^k = x_2^k = (2k + 2)/(2k + 3)$, and $x_3^k = 2/(2k + 3)$. Thus, for any iteration k one has $x_3 > 0$ and only obtains an optimal solution as $k \rightarrow \infty$.

Although Gomory fractional and mixed-integer cuts were developed in the late 1950s and early 1960s, they were not widely used until the late 1990s. Balas *et al.* [17] demonstrated the computational effectiveness of Gomory cuts when used within a branch-and-cut scheme. More recently, Zanette *et al.* [18] showed that the performance of a pure Gomory fractional cutting-plane method is far superior when the lexicographic dual simplex method is used instead of the regular dual simplex. In their implementation, instead of using the lexicographic pivoting rule in the dual simplex, they use a post-processing scheme to obtain the lexicographically smallest solution. Other computational enhancements are reported in [19].

ALGORITHMS USING SEQUENTIAL CONVEXIFICATION

Lift-and-Project Cutting-Plane Algorithm

In this section, we discuss the lift-and-project cutting-plane algorithm [5] for mixed binary programs, that is, the mixed-integer restrictions in Equation 1 are $x \in \{0, 1\}^p \times \mathbb{R}_+^{n-p}$. The lift-and-project algorithm has its roots in disjunctive programming. Disjunctive programming was introduced by

Balas [20] and refers to optimization over union of polyhedra. Disjunctions arise naturally in integer programming. For example, a binary restriction on a variable $x_j \in \{0, 1\}$ can be represented as a disjunction $(x_j = 0) \vee (x_j = 1)$, where $\vee$ denotes the logical “or” operator. Similarly, an integer variable $x_j \in [0, U_j] := \{0, 1, \dots, U_j\}$ can be represented using the disjunction $(x_j = 0) \vee (x_j = 1) \vee \dots \vee (x_j = U_j)$. A variety of disjunctions can be used to generate cuts for an MIP. For example, when the LP relaxation of an MIP yields a solution $\bar{x}$ with $\bar{x}_j$ fractional for $j \in \{1, \dots, p\}$, a disjunction of the type $(X_\ell \cap \{x_j \leq \lfloor \bar{x}_j \rfloor\}) \vee (X_\ell \cap \{x_j \geq \lceil \bar{x}_j \rceil\})$ can be written to produce a cut that separates this point from $\text{clconv}(X)$. When the disjunction used for cut generation contains only two terms of the type mentioned earlier, we refer to such cuts as elementary disjunctive cuts. More general two-term disjunctions (e.g. *split disjunctions* [21]), disjunctions with more than two terms (multiterm disjunctions), and so forth can be considered to generate violated inequalities. We refer the reader to the articles in this encyclopedia by L. Liberti, Q. Louveaux and K. Andersen, and the references therein for further details.

The lift-and-project algorithm relies on a sequential convexification property of mixed binary programs. Mixed binary programs inherit this property from a more general class of problems called *facial disjunctive programs* (cf. [22]). Given a general MIP (Eq. 1), the convexification with respect to a variable $x_j, j \in \{1, \dots, p\}$ and a polyhedron Y is defined as:

$$P_j(Y) := \text{clconv}(Y \cap \{x : x_j \in \mathbb{Z}\}).$$

Let $i_1, \dots, i_t$, where $i_j \in \{1, \dots, p\}, j = 1, \dots, t$ be variable indices. For $t \geq 2$ define

$$P_{i_t, \dots, i_1}(X_\ell) := P_{i_t}(P_{i_{t-1}}(\dots(P_{i_1}(X_\ell))\dots)),$$

that is, first, the mixed-integer set defined by imposing integrality of x_{i_1} to X_ℓ is convexified. This results in the set $P_{i_1}(X_\ell)$. Then, the set defined by imposing integrality of x_{i_2} to $P_{i_1}(X_\ell)$ is convexified, and so forth. Imposing

integrality variable-by-variable and taking the closure of the convex hull at each step is referred to as *sequential convexification* in the sequence defined by the permutation $i_1, \dots, i_p$.

For mixed-binary programs, when the sets are convexified with respect to the binary variables sequentially in the manner described earlier, Balas [20] showed that after p steps, one obtains the closure of the convex hull of X , that is, $P_{i_p, \dots, i_1}(X_\ell) = \text{clconv}(X)$. Balas [20] also gave an example showing that for general MIPs (with $x \in \mathbb{Z}_+^p \times \mathbb{R}_+^{n-p}$), sequential convexification may fail to yield $\text{clconv}(X)$. Note that general MIPs are not a subclass of facial disjunctive programs.

Theorem 2 [5, 20]. *For mixed binary programs*

$$P_{i_p, \dots, i_1}(X_\ell) = \text{clconv}(X).$$

Sherali and Adams [23, 24] give an alternative finite characterization of the convex hull of the binary and mixed-binary integer programs based on the reformulation-linearization technique (RLT). We refer the reader to the encyclopedia article by H. D. Sherali for further details. See also [25] for another finite construction of the convex hull.

Assume that the constraints $Ax \geq b$ in the mixed-binary program include the constraints $x \leq 1$. In the lift-and-project cutting-plane method, cut generation is a two-step process. The first step is a lifting step, which involves representing the convex hull of the disjunctive program as an LP in a higher dimensional space (of dimension larger than n). We shall consider the constructions in [5]. If the solution to the linear relaxation is fractional, with x_j fractional, then one first constructs the nonlinear system

$$(1 - x_j)(Ax - b) \geq 0$$

$$x_j(Ax - b) \geq 0.$$

This system is linearized by setting $x_i x_j =: z_i, i \neq j$ and $x_i^2 = x_i$ (the latter identity holds because x_j is binary). The new linear system

has additional variables; let its feasible set be denoted by $M_j(X_\ell)$. Alternatively, one can write a disjunction

$$D_j := (X_\ell \cap \{x_j = 0\}) \vee (X_\ell \cap \{x_j = 1\}). \quad (9)$$

Consider the LP,

$$\begin{aligned} & \min c^\top x \\ & Ay^1 - by_0^1 \geq 0 \\ & Ay^2 - by_0^2 \geq 0 \\ & -y_j^1 = 0 \\ & y_j^2 - y_0^2 = 0 \\ & y_0^1 + y_0^2 = 1 \\ & x = y^1 = y^2. \\ & x, y^1, y^2 \geq 0, y_0^1, y_0^2 \geq 0, \end{aligned}$$

and let $K_j(X_\ell)$ denote the feasible region of this LP.

The second step of the lift-and-project algorithm is to project the higher dimensional LP onto the space of the original variables. Balas *et al.* [5] show that the projection of $M_j(X_\ell)$ (equivalently, the projection of $K_j(X_\ell)$) onto the space of x variables yields $P_j(X_\ell)$. Balas *et al.* [5] show that valid inequalities for $P_j(X_\ell)$ correspond to feasible points of projection cone of $M_j(X_\ell)$. If X_ℓ is full dimensional, the facets of $P_j(X_\ell)$ correspond to the extreme rays of the projection cone of $M_j(X_\ell)$. These results are specialized versions of the more general result on the compact representation of the closure of the convex hull of a disjunctive program in a higher dimensional space [22]. In order to produce violated lift-and-project cuts for a fractional solution to the linear relaxation, one writes the so-called cut generation linear program (CGLP), which is formulated by writing the projection cone of $M_j(X_\ell)$ (or $K_j(X_\ell)$) and adding a normalization constraint that truncates the cone (see also [26]). In addition, one adds an objective function, for example, an objective function that maximizes the violation of the cut.

In a cutting-plane algorithm that employs lift-and-project cuts for mixed-binary programs, one can always generate a violated

inequality for a fractional solution by forming a disjunction. However, as shown in an example in Sen and Serali [27], a naive implementation of the lift-and-project cutting-plane algorithm may not be finitely convergent. They show that appropriate memory needs to be maintained to guarantee finite convergence. The main idea behind the convergence proof of the lift-and-project algorithm is to carefully select the fractional variables and polyhedra for generating cuts that lead to sequential convexification of the linear relaxation with respect to the binary variables. To this end, let a j -cut be a valid inequality for $P_j(\cdot)$ obtained by solving a CGLP with an appropriate normalization. Also let S_i^k be the polyhedron defined by the inequalities $Ax \geq b$ and all j -cuts for $j = 1, \dots, i$ at iteration k , and let S_0^k denote the linear relaxation X_ℓ . In line 3 of Algorithm 1, to generate a cut, one chooses the *largest index* $j \in \{1, \dots, p\}$ such that x_j is fractional, and generates an inequality for $P_j(S_{j-1}^k)$ using the CGLP with an appropriate normalization constraint. One can show that the cut generated for $P_j(S_{j-1}^k)$ cuts off the fractional solution x^k . To show finite convergence, observe that because the projection cone of $M_1(S_0^k) = M_1(X_\ell)$ has finitely many extreme points, this set is convexified with respect to x_1 after the addition of finitely many 1-cuts. By induction, one can show that the number of j -cuts generated is finite for all $j = 1, \dots, p$. Thus, in the worst case, we will have constructed $\text{clconv}(X)$ in finitely many steps, at which point an optimal solution (if it exists) is found. Let m_j be the iteration index when the last i -cut was generated for $i = 1, \dots, j$. Note that $S_j^k = S_j^{m_j}$ in all iterations $k \geq m_j$. Therefore, the iteration index m_j provides the memory of inequalities defining S_j^k for each j , and it plays an important role in the convergence proof. We refer the reader to Sen and Serali [27] for a related proof of convergence and [28, 29] for an alternative finitely convergent implementation of a pure cutting-plane algorithm for facial disjunctive programs.

Theorem 3 [5]. *Suppose at each iteration k , the lift-and-project cutting-plane algorithm is implemented by selecting*

the largest index j that is fractional and generating a cut corresponding to an extreme point of CGLP (with appropriate normalization) formed using the polyhedron S_{j-1}^k . Then, the cutting-plane method finds an optimal solution or concludes that the problem is infeasible in finitely many steps.

The first computational results incorporating lift-and-project cuts within a branch-and-cut algorithm are reported in [30]. We refer the reader to the encyclopedia article by Q. Louveaux for a survey of recent computational developments.

Cutting-Plane-Tree Algorithm

As discussed earlier, a variable-by-variable sequential convexification may not yield the convex hull of feasible solutions for general MIPs. In fact, Owen and Mehrotra [31] give the following example, showing that when the lift-and-project cutting-plane algorithm is applied to the mixed-integer case, one may not converge to an optimal solution even in the limit.

$$\begin{aligned} \min & -x_1 - x_2 \\ \text{s.t.} & 8x_1 + 12x_2 \leq 27 \\ & 8x_1 + 3x_2 \leq 18 \\ & 0 \leq x_1, x_2 \leq 3 \\ & x \in \mathbb{Z}_+^2. \end{aligned}$$

It can be easily verified that the convex hull of the feasible set of MIP given above is $\{x \in \mathbb{R}_+^2 : x_1 + x_2 \leq 2\}$. However, an application of the lift-and-project algorithm using elementary (two-term) disjunctions of the type $(X_\ell \cap \{x_j \leq \lfloor \bar{x}_j \rfloor\}) \vee (X_\ell \cap \{x_j \geq \lceil \bar{x}_j \rceil\})$ yields in the limit the set $\{x \in \mathbb{R}_+^2 : x_1 + x_2 \leq 2.25\}$.

In a recent article, Chen *et al.* [6] propose the cutting-plane-tree (CPT) algorithm, which uses cuts derived from multiterm disjunctions. The multiterm disjunctions partition the feasible region into a finite collection of intervals for each variable, the union of which gives contains the feasible domain. Iteratively finer partitions are obtained over the course of the algorithm.

We first review a convexification result for general MIP proposed by Chen *et al.* [6]. Assume that each integer variable is bounded and can take integer values in the interval $[0, U_j]$. Let each interval be further divided into subintervals

$$[L_{1j} := 0, U_{1j}], [L_{2j}, U_{2j}], \dots, [L_{t_{jj}}, U_{t_{jj}} := U_j],$$

where $L_{\kappa_{jj}}, U_{\kappa_{jj}} \in \mathbb{Z}_+$, $L_{\kappa_{jj}} \leq U_{\kappa_{jj}}$ and $\kappa_j \in \{1, \dots, t_j\}$, $j = 1, \dots, p$ with $L_{\kappa_{j+1}j} - U_{\kappa_{jj}} \leq 1$ for $\kappa_j \in 1, \dots, t_j - 1$ so that the subintervals span all integers in $[0, U_j]$. Given such partition $\mathcal{P}$, the collection of all p -tuples of the form $\kappa = (\kappa_1, \kappa_2, \dots, \kappa_p)$ is denoted by $K(\mathcal{P})$. A unit partition $\mathcal{P}^*$ is a partition where $U_{\kappa_{jj}} - L_{\kappa_{jj}} \leq 1$ for all $\kappa_j \in \{1, \dots, t_j\}$, $j = 1, \dots, p$. For $\kappa \in K(\mathcal{P}^*)$, $j \in \{1, \dots, p\}$, and a polyhedron $\bar{X}$, we define the following sets:

$$\begin{aligned} \mathcal{P}^-(\kappa, j, \bar{X}) &:= \left\{ x \in \bar{X} : L_{\kappa_i i} \leq x_i \leq U_{\kappa_i i}, \right. \\ &\quad \left. i = 1, \dots, p; x_j \leq L_{\kappa_{jj}} \right\} \end{aligned}$$

and,

$$\begin{aligned} \mathcal{P}^+(\kappa, j, \bar{X}) &:= \left\{ x \in \bar{X} : L_{\kappa_i i} \leq x_i \leq U_{\kappa_i i}, \right. \\ &\quad \left. i = 1, \dots, p; x_j \geq U_{\kappa_{jj}} \right\}. \end{aligned}$$

In addition, we define $\mathcal{H}_j^\kappa(\bar{X}) := \text{clconv}(\mathcal{P}^-(\kappa, j, \bar{X}) \cup \mathcal{P}^+(\kappa, j, \bar{X}))$. The following theorem gives a sequential convexification result for a general MIP with respect to a unit partition $\mathcal{P}^*$.

Theorem 4 [6]. *Assume that the set X as defined in (1) is nonempty and has bounded integer variables. Then for any unit partition $\mathcal{P}^*$,*

$$\begin{aligned} \text{clconv}(X) &= \text{clconv}\left\{ \bigcup_{\kappa \in K(\mathcal{P}^*)} \right. \\ &\quad \left. [\mathcal{H}_p^\kappa(\mathcal{H}_{p-1}^\kappa(\dots(\mathcal{H}_1^\kappa(X_\ell))\dots)) \setminus \emptyset] \right\}. \end{aligned}$$

Adams and Sherali [32] propose an alternative finite convex hull characterization for general MIP based on Lagrange interpolation polynomials. Similar to the case with mixed-binary programs described in the section

titled “Lift-and-Project Cutting-Plane Algorithm”, the existence of a finite convex hull representation suggests the existence of a finite pure cutting-plane algorithm. However, care must be taken in the design of the algorithm to guarantee convergence. Next, we describe such an algorithm, referred to as the *CPT algorithm* [6].

In the CPT $\mathcal{T}$, there is a single root node o . For each node $\sigma \in \mathcal{T}$, an integer $v_\sigma \in \{1, 2, \dots, n_1\}$ stores the index of the integer variable that is split, an integer q_σ stores the (lower) level of the splitting. Let l_σ , r_σ , and p_σ denote the left child, right child, and parent nodes of node σ , respectively. Let $S(\sigma)$ be all nodes on the subtree rooted at node σ (not including node σ and the leaf nodes). Let $\mathcal{N}(\sigma)$ be the collection of the nodes on the path from the root node to node σ (not including the root node), $\mathcal{N}^-(\sigma)$ be the collection of nodes in $\mathcal{N}(\sigma)$ that were formed as the left child node of its parent, and $\mathcal{N}^+(\sigma)$ be the collection of nodes in $\mathcal{N}(\sigma)$ that were formed as the right child node of its parent. Given $\sigma \in \mathcal{T}$ define

$$\begin{aligned} \mathcal{C}_\sigma &= \{x | x_j \in [0, U_j], x_{v_{ps}} \leq q_{ps}, \forall s \in \mathcal{N}^-(\sigma), \\ x_{v_{ps}} &\geq q_{ps} + 1, \forall s \in \mathcal{N}^+(\sigma)\} \end{aligned}$$

We let m_σ store an iteration index, which keeps track of the cutting planes that will be used to generate a disjunctive cut when this node is revisited. In other words, m_σ determines the set X^{m_σ} to be used in CGLP [20, 26]. The set X^{m_σ} corresponds to X_ℓ together with the first $m_\sigma - 1$ cuts added to it. If $X^{m_\sigma} \cap \mathcal{C}_{l_\sigma} = \emptyset$ ($X^{m_\sigma} \cap \mathcal{C}_{r_\sigma} = \emptyset$), we say that the left (right) child node of σ is “fathomed”, that is, $l_\sigma = \text{null}$ ($r_\sigma = \text{null}$).

At iteration k of the CPT algorithm, if the current extreme point solution to $\min_{x \in X_\ell^k} c^T x$, given by x^k is integral, then we have found an optimal solution to the MIP. Otherwise, we search the CPT to find the last node σ on the path from the root node such that $x^k \in \mathcal{C}_\sigma$. There are two cases. Case (1) σ is a leaf node ($\sigma \in \mathcal{L}_k$, where $\mathcal{L}_k$ is the set of leaf nodes of the CPT at iteration k) and Case (2) σ is not a leaf node ($\sigma \notin \mathcal{L}_k$, $x^k \in \mathcal{C}_\sigma$, $x^k \notin \mathcal{C}_{l_\sigma}$, and $x^k \notin \mathcal{C}_{r_\sigma}$). In Case (1), we choose a fractional variable

$x_j, j = 1, \dots, n_1$ with the smallest index, and let the split variable be $v_\sigma = j$. We create two new nodes: left (l_σ) and right (r_σ) children of σ at the split level $q_\sigma = \lfloor x_j^k \rfloor$. We let $\mathcal{C}_{l_\sigma} = \{x \in \mathcal{C}_\sigma | x_j \leq \lfloor x_j^k \rfloor\}$ and $\mathcal{C}_{r_\sigma} = \{x \in \mathcal{C}_\sigma | x_j \geq \lceil x_j^k \rceil\}$. In this case, we also let $m_\sigma = k$, as this is the first time the tree search for a fractional solution stops at node σ . Let $\mathcal{L}_{k+1} \leftarrow (\mathcal{L}_k \cup \{l_\sigma, r_\sigma\}) \setminus \{\sigma\}$. In Case (2), the CPT and m_σ are unchanged, that is, $\mathcal{L}_{k+1} \leftarrow \mathcal{L}_k$. However, in this case, we update $m_t = k$ for all successors of σ , $t \in S(\sigma)$. We generate a valid inequality for the set $\text{clconv}\{\cup_{t \in \mathcal{L}_{k+1}} (X^{m_\sigma} \cap \mathcal{C}_t)\}$ that cuts off x^k (from an extreme point of the associated CGLP). The new inequality is included along with those defining X^k , and the resulting set is denoted X^{k+1} . This process continues until an optimal solution to the MIP is obtained.

Theorem 5 [6]. *Assume that the set X is nonempty and has bounded integer variables. Then, the CPT algorithm converges to an optimal solution of the MIP in finitely many iterations.*

The convergence proof relies on the tree data structure and the iteration index m_σ that serves as memory for storing the subset of cuts generated until that iteration of the algorithm. A careful reader will recognize the similarity between the memory index m_j in the section titled “Lift-and-Project Cutting-Plane Algorithm” with the memory index m_σ introduced in this section.

Table 1 shows the first three iterations of the execution of the CPT algorithm on the example of Owen and Mehrotra [31] that was presented at the beginning of this section. In this table, k , x^k , σ , and m_σ denote the iteration number, solution of the relaxation, the node, and the memory parameter, respectively. The sets $\mathcal{Q}_i$ in an iteration represent a subset of $[0, 3] \times [0, 3]$ in which the current solution can fall. We begin the algorithm by solving the linear relaxation corresponding to the root node $\sigma = 1$ (setting $m_1 = 1$) and we obtain a fractional solution (15/8, 1). We create two child nodes for 1, namely 2 and 3 representing $x_1 \leq 1$ and $x_1 \geq 2$, respectively. We then generate a disjunctive cut using the CGLP corresponding to the disjunction

Table 1. First Three Iterations of the CPT Algorithm on Owen and Mehrotra's [31] example

k	x^k	σ	m_σ	$\mathcal{Q}_{k+1}$			Cut
1	(15/8, 1)	1	1	$Q_1 = [0, 1] \times [0, 3]$	$Q_2 = [2, 3] \times [0, 3]$		$11/12x_1 + x_2 \leq 5/2$
2	(2, 2/3)	3	2	$Q_1 = [0, 1] \times [0, 3]$	$Q_2 = [2, 3] \times [0, 0]$		$x_1 + 15/19x_2 \leq 9/4$
3	(1, 19/12)	2	3	$Q_1 = [2, 3] \times [0, 0]$	$Q_2 = [0, 1] \times [0, 1]$	$Q_3 = [0, 1] \times [2, 3]$	$x_1 + 15/16x_2 \leq 9/4$

defined by Q_1 and Q_2 on X_ℓ , add it to the linear relaxation and re-solve the problem to get $x^2 = (2, 2/3)$. As x^2 satisfies the bounds specified by node 3, and node 3 is a leaf node, we set $m_3 = 2$ and create two child nodes for node 3. As the set $X_\ell \cap \{x : 2 \leq x_1 \leq 3, 1 \leq x_2 \leq 3\}$ corresponding to the right child of 3 is infeasible, we remove it from further consideration. The CGLP in iteration 2 is constructed using X_ℓ , cut 1 (as $m_3 = 2$) and the disjunctions are defined by $Q_i, i = 1, 2$ and a disjunctive cut is generated. In the third iteration, on solving the updated relaxation, we get another fractional solution (1, 19/12), which falls at node 2, which is a leaf node. Therefore, we set $m_2 = 3$ and create two child nodes for node 2. It turns out that the right node of node 2 is infeasible; therefore, we discard it from further consideration and we are left with the sets $Q_i, i = 1, \dots, 3$. The algorithm proceeds in this manner and generates the cut $x_1 + x_2 \leq 2$ in the sixth iteration (see [6]). In the subsequent solution of the linear relaxation, we find an optimal integer solution and the algorithm terminates. Note that while the convex hull was generated at termination in this example, this generally need not be the case for the algorithm to terminate.

Computational results with a branch-and-cut algorithm based on cuts from the CPT are presented in [33]. Jörg [34] proposes another finitely convergent pure cutting-plane algorithm for general MIP using multiterm disjunctions in conjunction with Gomory cuts. An iteration of this approach involves the identification of all alternative optimal solutions of a related LP and the enumeration of all extreme rays of a polyhedral projection cone.

CONCLUSIONS

In this article, we survey different pure cutting-plane algorithms and the key results needed to show their convergence. Throughout, we emphasize the important role that memory plays in the convergence proofs. Various studies indicate that pure cutting-plane algorithms are not practical for deterministic MIPs, but cutting planes are crucial for the solution of difficult MIPs in a branch-and-cut algorithm (see the article by J. Mitchell in the encyclopedia for an introduction to branch-and-cut algorithms). However, pure cutting-plane algorithms have important applications in the solution of two-stage stochastic MIPs. We refer the reader to the articles by Sen and Hingle [35], for a decomposition algorithm based on lift-and-project cuts for two-stage stochastic mixed-binary programs, Gade *et al.* [36], for decomposition algorithms using Gomory cuts for two-stage stochastic pure integer programs, and Ntaimo [37], for a decomposition algorithm for two-stage stochastic mixed-binary programs based on the so-called Fenchel cuts. (A pure cutting-plane algorithm based on Fenchel cuts [38] is shown to be finitely convergent for deterministic MIPs in [39]. The separation of Fenchel cuts involves the solution of MIP.)

An area closely related to the convergence of pure cutting-plane algorithms is the study of the convergence of *closures* of valid inequalities in MIPs. Given a polyhedron and a family of valid inequalities, the closure of the family of valid inequalities with respect to the polyhedron is obtained by adding all valid inequalities in the family to the polyhedron. Note that the number of inequalities in a given family can be infinite. For example, the family of Chvátal inequalities (which are equivalent to Gomory fractional cuts, not

necessarily constructed from a basis of the LP relaxation) are of the type $\lceil u^\top A \rceil \geq \lceil u^\top b \rceil$ for $u \in \mathbb{R}_+^m$ and the vector ceiling is performed component-wise. Chvátal [40] shows that for pure integer programs, the Chvátal closure is a polyhedron, that is, only finitely many Chvátal inequalities are needed to describe the closure. Moreover, starting with the linear relaxation of a pure integer program, taking Chvátal closures recursively yields the convex hull in finitely many steps. For MIPs, Cook *et al.* [21] show that the split closure is a polyhedron. However, using an example similar to the one given in this article in the section titled “Algorithms using Lexicography and Rounding”, they show that taking the split closure recursively does not yield the convex hull in finitely many steps. Nemhauser and Wolsey [41] show the equivalence between GMI cuts, mixed integer rounding cuts, and split cuts; and as a result, a cutting-plane algorithm based on any of these classes is not finitely convergent for general MIPs. Recently, Del Pia and Weismantel [42] show that for MIPs, the split closure converges to the convex hull in the limit. They also show that the closure of a certain family of cuts called *lattice-free cuts* is a polyhedron and that taking this closure recursively yields the convex hull in finitely many steps for general MIPs. Connections between lattice-free cuts and general multiterm disjunctive cuts are studied in [43], who also give a procedure to construct the closure of the convex hull in finitely many steps. The relationships among the closures of various families of valid inequalities are studied in [44].

ACKNOWLEDGMENTS

The authors thank Suvrajeet Sen for many discussions on convergence of cutting planes and their applications in decomposition algorithms for stochastic integer programs. This work was supported, in part, by NSF-CMMI Grant 1100383.

REFERENCES

1. Marchand H, Martin A, Weismantel R, *et al.* Cutting planes in integer and mixed integer programming. *Discrete Appl Math* 2002;123(1):397–446.
2. Cornuéjols G. Valid inequalities for mixed integer linear programs. *Math Program* 2008;112(1):3–44.
3. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. New York: John Wiley & Sons; 1988.
4. Gomory RE. Outline of an algorithm for integer solutions to linear programs. *Bull Am Math Soc* 1958;64(5):275–278.
5. Balas E, Ceria S, Cornuéjols G. A lift-and-project cutting plane algorithm for mixed 0–1 programs. *Math Program* 1993;58(1):295–324.
6. Chen B, Küçükyavuz S, Sen S. Finite disjunctive programming characterizations for general mixed-integer linear programs. *Oper Res* 2011;59(1):202–210.
7. Gomory RE. An algorithm for integer solutions to linear programming. In: Graves RL, Wolfe P, editors. *Recent Advances in Mathematical Programming*. New York: McGraw-Hill; 1963. p 269–302.
8. Nourie FJ, Venta ER. An upper bound on the number of cuts needed in Gomory’s method of integer forms. *Oper Res Lett* 1982;1(4):129–133.
9. Neto J. A simple finite cutting plane algorithm for integer programs. *Oper Res Lett* 2012;40(6):578–580.
10. Orlin JB. A finitely converging cutting plane technique. *Oper Res Lett* 1985;4(1):1–3.
11. Armstrong R, Charnes A, Phillips F. Page cuts for integer interval linear programming. *Discrete Appl Math* 1979;1:1–14.
12. Bell DE. A cutting plane algorithm for integer programs with an easy proof of convergence. Technical Report WP-73-15. Laxenburg: International Institute for Applied Systems Analysis; 1973.
13. Bowman VJ, Nemhauser GL. A finiteness proof for modified Dantzig cuts in integer programming. *Nav Res Logist Q* 1970;17(3):309–313.
14. Gomory RE. An algorithm for the mixed integer problem. Technical Report RM-2597. Santa Monica, CA: RAND Corporation; 1960.
15. White W. On Gomory’s mixed integer algorithm [Senior Thesis] Princeton University; 1961.
16. Padberg M. Classical cuts for mixed-integer programming and branch-and-cut. *Math Methods Oper Res* 2001;53(2):173–203.

17. Balas E, Ceria S, Cornuéjols G, *et al.* Gomory cuts revisited. *Oper Res Lett* 1996;19(1):1–9.
18. Zanette A, Fischetti M, Balas E. Lexicography and degeneracy: can a pure cutting plane algorithm work? *Math Program* 2010;130(1):1–24.
19. Balas E, Fischetti M, Zanette A. On the enumerative nature of Gomory’s dual cutting plane method. *Math Program* 2010;125(2):325–351.
20. Balas E. Disjunctive programming. In: Hammer PL, Johnson EL, Korte BH, editors. *Discrete Optimization II*. Volume 5, *Annals of discrete mathematics*. North-Holland: Elsevier; 1979. p 3–51.
21. Cook W, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program* 1990;47(1):155–174.
22. Balas E. Disjunctive programming: properties of the convex hull of feasible points. *Discrete Appl Math* 1998;89(1):3–44.
23. Sherali HD, Adams WP. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J Discrete Math* 1990;3(3):411–430.
24. Sherali HD, Adams WP. A hierarchy of relaxations and convex hull representations for mixed-integer zero-one programming problems. *Discrete Appl Math* 1994;52(1):83–106.
25. Lovász L, Schrijver A. Cones of matrices and set-functions and 0–1 optimization. *SIAM J Optim* 1991;1(2):166–190.
26. Sherali HD, Shetty CM. *Optimization with disjunctive constraints*. Berlin: Springer-Verlag; 1980.
27. Sen S, Sherali HD. On the convergence of cutting plane algorithms for a class of nonconvex mathematical programs. *Math Program* 1985;31(1):42–56.
28. Blair CE. Facial disjunctive programs and sequence of cutting-planes. *Discrete Appl Math* 1980;2(3):173–179.
29. Jeroslow RG. A cutting-plane game for facial disjunctive programs. *SIAM J Control Optim* 1980;18(3):264–281.
30. Balas E, Ceria S, Cornuéjols G. Mixed 0–1 programming by lift-and-project in a branch-and-cut framework. *Manage Sci* 1996;42:1229–1246.
31. Owen JH, Mehrotra S. A disjunctive cutting plane procedure for general mixed-integer linear programs. *Math Program* 2001;89(3):437–448.
32. Adams WP, Sherali HD. A hierarchy of relaxations leading to the convex hull representation for general discrete optimization problems. *Ann Oper Res* 2005;140(1):21–47.
33. Chen B, Küçükyavuz S, Sen S. A computational study of the cutting plane tree algorithm for general mixed-integer linear programs. *Oper Res Lett* 2012;40(1):15–19.
34. Jörg M. 2007. *k*-disjunctive cuts and a finite cutting plane algorithm for general mixed integer linear programs. Available at <http://arxiv.org/PScache/arxiv/pdf/0707/0707.3945v1.pdf>. Accessed 2013 July 18.
35. Sen S, Hige JL. The C^3 theorem and a D^2 algorithm for large scale stochastic mixed-integer programming: set convexification. *Math Program* 2005;104(1):1–20.
36. Gade D, Küçükyavuz S, Sen S. Decomposition algorithms with parametric Gomory cuts for two-stage stochastic integer programs. *Math Program* 2012, Article in Advance. DOI: 10.1007/s10107-012-0615-y.
37. Ntamo L. Fenchel decomposition for stochastic mixed-integer programming. *J Glob Optim* 2013;55(1):141–163.
38. Boyd EA. Fenchel cutting planes for integer programs. *Oper Res* 1994;42(1):53–64.
39. Boyd EA. On the convergence of Fenchel cutting planes in mixed-integer programming. *SIAM J Optim* 1995;5(2):421–435.
40. Chvátal V. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete Math* 1973;4(4):305–337.
41. Nemhauser GL, Wolsey LA. A recursive procedure to generate all cuts for 0–1 mixed integer programs. *Math Program* 1990;46(1):379–390.
42. Del Pia A, Weismantel R. On convergence in mixed integer programming. *Math Program* 2012;135(1):397–412.
43. Dash S, Dobbs NB, Günlük O, *et al.* 2012. Lattice-free sets, branching disjunctions, and mixed-integer programming. IBM Research Report. Available at <http://researcher.watson.ibm.com/researcher/files/us-sanjeebd/purecuts-oo.pdf>. Accessed 2013 July 18.
44. Cornuéjols G, Li Y. Elementary closures for integer programs. *Oper Res Lett* 2001;28(1):1–8.

PUSH AND PULL PRODUCTION SYSTEMS

W.C. BENTON, JR
Department of Management
Sciences, Fisher College of
Business, The Ohio State
University, Columbus, Ohio

INTRODUCTION

The terms Push and Pull are used to describe alternative systems for executing the production process in manufacturing organizations. A Push system is based on forecasted demand that is completed and sent to the next work station or in the case of the final work station, is pushed to finished goods inventory. Specifically, in a push system, work output moves to the next station when it is completed without knowing whether the next station is ready to receive the additional work output. If the succeeding workstation is not ready to receive the new work, excess inventories sometimes result in cost and quality problems. On the other hand, in a pull system, the movement of work is based on the requirements of the following work station. Each succeeding workstation pulls (demands) output from the previous workstation as needed. The next work station determines when and how much output is requested. The output from the final workstation is pulled by customer demand or the master production schedule. In a pull system, work moves to the next station only in response to demand from the next station.

A Hybrid production system combines both Push and Pull production systems. Careful analysis and planning is required prior to implementing this.

In a push production system, the emphasis is on material planning and procurement. In a pull production system, the emphasis is on shop floor scheduling. The hybrid production system thus makes best use of both material planning and shop floor scheduling efficiencies.

There is a common agreement among operations management researchers that JIT functions as a pull system and MRP as a push system. See *Just-in-Time/Lean Production Systems* and *Materials Requirements Planning*. Developed along with JIT, a Kanban system is a production control approach that uses containers, cards, or visual cues to control the production movement of goods through a manufacturing system. Yet, this proposition does not always hold true. For example, Rice and Yoshikawa [1] suggest that there are many practical similarities between the Kanban system and MRP and both are “pull-through” approaches. Also, Karmarker [2] suggests that JIT can be either a pull system or a push system because it is an integral philosophy rather than timing-based manufacturing technique. Summarizing the existing viewpoints on push/pull principles, Pyke and Cohen [3] conclude that an entire manufacturing system cannot be labeled as a push or pull system. They also suggest that the push/pull principles are an attribute of all underlying manufacturing and distribution control systems.

There are *three* ways to define or distinguish the nature of push and pull systems in general. The most common way is to characterize the differences in terms of the order release [2,4,5]. According to this viewpoint, in a pull system, removing an end item (or a fixed lot of end items) triggers the order release, by which the flow of materials or components is initiated. In contrast, push systems allow for the production or material flow in anticipation of future demand.

The second way is to examine the structure of the information flow [6–8]. In a pull system, the physical flow of materials is triggered by the local demand from the next server (individual station). The local demand in this system is signified by the local information (i.e., empty kanbans). In this context, the pure pull system is a decentralized control strategy in which the ultimate goal of meeting demand (orders) is disregarded in the local servers (individual

stations). However, a push system uses global and centralized information. Global information of customer orders and demand forecasts is released and processed to control all the levels of production in the push system.

On the basis of this notion, manufacturing planning and control (MPC) systems may consist of both push/pull elements simultaneously. According to Pyke and Cohen [3], centralized planning systems always have push elements. Therefore, even in the pull-manufacturing environment, outcomes of the central planning decisions may limit certain operating variables of the production control systems. For example, the JIT production system has a push element in setting capacity limit of a batch, the container (bin) sizes.

Finally, the third way to interpret push/pull systems is a practical approach associated with work-in-process (WIP) level on the shop floor. The WIP are orders or materials that have been started in the production process, but are not yet complete. WIP includes orders or materials waiting to be started, orders being setup, orders being worked on, and orders or materials are waiting to be moved to the next work station. A variant of a pull system is the *CONstant Work in Process (CONWIP)* system [9] which is known for its ease of implementation. Stockpiling WIP is a major drawback of MRP systems. In a simulation study comparing push, pull, and CONWIP systems, Spearman and Zazanis [10] found that the merits of the pull system are generated by the bounded WIP, not by the characteristics of pull logic. They show that a hybrid approach, CONWIP, functions in a very similar manner to the pull approach due to its bounded WIP level. Simply, a push system is considered to have infinite production capacity, and a pull system is considered to have finite production capacity. Therefore, the MRP controlled shop floor system without a WIP blocking mechanism would create more WIP inventory than a pull system, unless the production strictly follows the original planned schedule. One of the critical implications of this result is that comparison based on WIP level between an MRP and Kanban system at the shop floor

level may not be fair since WIP level of the Kanban system is inhibited by the size and number of containers.

Based on the three viewpoints suggested above, it can be concluded that if materials flow is initiated by the central planning system without controlling WIP level, the resulting system is close to the pure push system. In a pure push system, the parts or component will precede based on the predetermined schedule regardless of whether the next server is busy or idle. In a pure pull system, the subsequent process will withdraw (i.e., pull) the parts from the preceding process using local information and controlling WIP inventory level.

At the shop floor level, a Kanban system functions as a pull system, and MRP works as a push system. At other levels, the assumption that MRP means push and JIT means pull are not always true. In practice, for example, JIT production systems also use global information for the long-term production planning. The push/pull and MRP/JIT studies are related in a way of deliberating issues associated with material flow. But, MRP or JIT production system plays a role as a structure of a MPC system, whereas push or pull logic is a principle only for material flow in the MPC system. Thus, it is possible for a MPC system to operate under both the push and pull principles simultaneously depending on the stage of production. Notice that the combination of push/pull principles has been adopted and tested in the push/pull integration literature (see Table 1) and Benton and Shin [11]. In this push/pull hybrid-manufacturing environment, the distinction between push and pull is more ambiguous.

KANBAN PRODUCTION CONTROL SYSTEM

The Kanban production system was developed by Toyota. In the most basic Kanban production system, a card is attached to each container of items that have been produced. The container is filled with a percentage of the daily requirements for an item. When the items in the container are used, the card is removed from the container. The

Table 1. MRP/JIT Analytical Integration Literature

Category			
References	Title	Discussion	Methodology
Chaudhury and Whinston [12]	Towards an adaptive Kanban system	MPC shop floor control	Mathematical analysis and simulation
Ding and Yuen [5]	A modified MRP for a production system with the coexistence of MRP and Kanban	MPC	Simulation
Betz [13]	Common sense manufacturing, a method of production control	MPC	Field Study
Hodgson and Wang [7]	Optimal Hybrid push/pull control strategies for a parallel multistage system: Part 1	Shop floor control	Optimization
Hodgson and Wang [7]	Optimal Hybrid push/pull control strategies for a parallel multistage system: Part 2	Shop floor control	Optimization
Hirakawa <i>et al.</i> [14]	A hybrid push/pull control system for multistage manufacturing processes	MPC shop floor control	Mathematical analysis
Deleersnyder <i>et al.</i> [15]	Integrating Kanban type pull systems and MRP type push systems: Insights from a Markovian model	Shop floor control	Mathematical analysis
Hirakawa [16]	Performance of a multistage hybrid push/pull production control system	MPC shop floor control	Mathematical analysis

card is the indicator or signal that another container of items is needed. The empty containers are then moved to the storage area. When the container is refilled, the card is attached to the container. The container is then returned to the storage area. The cycle is repeated when the user obtains the refilled container.

In a pull system, work output is based on the demand of the next workstation. A kanban card is used to communicate (signal) the needed demand at the next workstation. Kanban is the Japanese word for card. The kanban card contains a part number, the part description, type of container, and various work station information. A Kanban production control system uses simple, visual signals to control the movement of materials between work centers as well as the production of new materials to replenish those

sent downstream to the next work center. The kanban card is attached to a storage and transport container. The kanban card is used to provide an easily understood, visual signal that a specific activity is required. Kanbans are similar to fixed order inventory systems where an order for Q is placed when the inventory level falls below the reorder point. The reorder point is determined based on the demand during the lead time. The only inventory required is the inventory required during the replenishment lead time.

In a dual-card Kanban system, there are two main types of kanban.

1. *Production Kanban* signals the need to produce more parts. Each kanban is physically attached to a container.
2. *Withdrawal Kanban* signals the need to withdraw parts from one work

center and deliver them to the next work center.

A Kanban controlled manufacturing system functions as a pull system. As discussed earlier, in a pull system, removing an end item (or a fixed lot of end items) triggers the order release, by which the flow of materials or components is initiated. In contrast, push systems allow for the production or material flow in anticipation of future demand.

In some pull systems, other signaling approaches are used in place of kanban cards. For example, an empty container alone (with appropriate identification on the container) could serve as a signal for replenishment. Similarly, a labeled, pallet-sized square painted on the shop floor, if uncovered and visible, could indicate the need to go get another pallet of materials from its point of production and move it on top of the empty square at its point of use.

A Kanban system is referred to as a pull system, because the kanban is used to pull parts to the next production stage only when they are needed. In contrast, an MRP system (or any schedule-based system) is a push system, in which a detailed production schedule for each part is used to push parts to the next production stage when scheduled. Thus, in a pull system, material movement occurs only when the work station needing more material asks for it to be sent, while in a push system the station producing the material initiates its movement to the receiving station, assuming that it is needed because it was scheduled for production. The weakness of a push system (MRP) is that customer demand must be forecast and production lead times must be estimated. Bad guesses (forecasts or estimates) result in excess inventory and the longer the lead time, and more room for error. The weakness of a pull system (i.e. Kanban) is that following the JIT production philosophy is essential, especially concerning the elements of short setup times and small lot sizes, because each station in the process must be able to respond quickly to requests for more materials. The dual-card kanban rules follow:

1. No parts are made unless there is a production kanban to authorize production. If no production kanban are in the “inbox” at a work center, the process remains idle, and workers perform other assigned activities. This rule enforces the “pull” nature of the process control.
2. There is exactly one kanban per container.
3. Containers for each specific part are standardized, and they are always filled with the same (ideally, small) quantity. (Think of an egg carton, always filled with exactly one dozen eggs.)

Decisions regarding the number of kanban (and containers) at each stage of the process are carefully considered, because this number sets an upper bound on the WIP inventory at that stage. For example, if 10 containers holding 12 units each are used to move materials between two work centers, the maximum inventory possible is 120 units, occurring only when all 10 containers are full. At this point, all kanban will be attached to full containers, so no additional units will be produced (because there are no unattached production kanban to authorize production). This feature of a dual-card Kanban system enables systematic productivity improvement to take place. By deliberately removing one or more kanban (and containers) from the system, a manager will also reduce the maximum level of WIP (buffer) inventory. This reduction can be done until a shortage of materials occurs. This shortage is an indication of problems (accidents, machine breakdowns, production delays, and defective products) that were previously hidden by excessive inventory. Once the problem is observed and a solution is identified, corrective action is taken so that the system can function at the lower level of buffer inventory. This simple, systematic method of inventory reduction is a key benefit of a dual-card kanban system. See the kanban determination example given in the next section.

Determining the Number of Kanbans (Containers)

The formula for determining the number of kanbans needed to control the production of a particular product/component part is shown below:

$$\text{Number of kanbans} = \frac{\text{average demand during lead time} + \text{safety stock}}{\text{container size}}$$

$$N = \frac{dL + S}{C}$$

where

N = number of kanbans or containers

d = average demand during lead time

L = lead time; the time it takes to replenish an order

S = safety stock: usually given as a percentage of demand during lead time or the level of expected variance

C = container size

To promote continuous improvement, the container size is usually set at a smaller level than the demand during lead time, say, and 10% of daily demand. This approach will result in reducing the number of kanbans. The reduced number of kanbans causes work methods and process problems to become visible.

Example. Marcus works as an operator at Health First, a supplement manufacturer. He is asked to process an average of 90 bottles an hour through his work station. If one kanban is attached to each container, a container holds 6 bottles, and it takes 20 min to receive new bottles from the upstream work station. The safety stock level is set at 10%. How many kanbans are required for the bottling process?

Solution.

$$N = \frac{(90 \times 0.33) + (90 \times 0.33 \times 0.10)}{6} = \frac{29.7 + 2.97}{6} = 5.44$$

Note: We can either round up or down. If we round down we would be promoting continuous improvement. If we round up we will be allowing more inventory into the process.

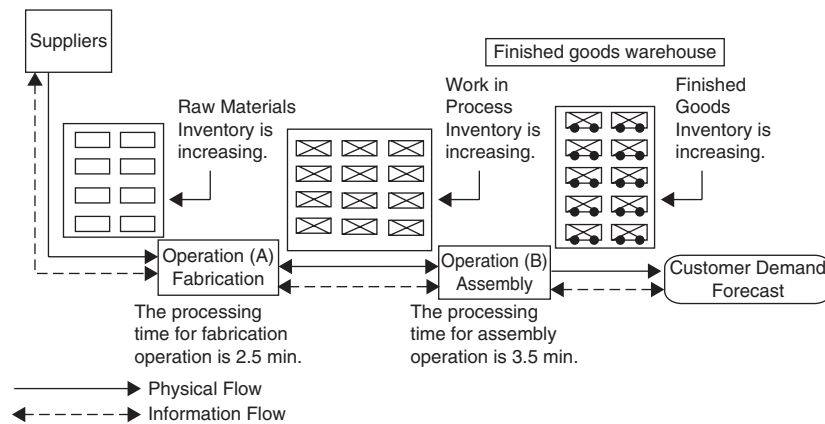
Comparison of a Push and Pull Production System

An illustration of the push and pull manufacturing concepts is given in the example above. Traditional MRP manufacturing systems use “push” systems, where the goal is to ensure all workstations and equipments are fully utilized. A result of this is that serial operations (i.e., fabrication, assembly, and finished goods) are independently optimized. Thus, if earlier operations are faster than a later operation, increasing levels of inventory will build at the succeeding work as in Fig. 1a.

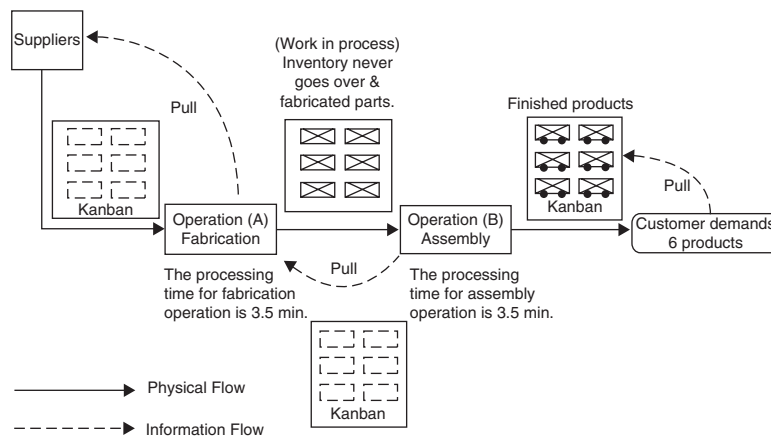
As shown in Fig. 1a, overproduction is an element of waste. There are significant

levels of excess raw materials, WIP, and finished goods inventories compared to the pull system shown in Fig. 1b. The resulting excess inventory uses up working capital and is liable to damage and devaluation. The ultimate inventory pile is at the end of the assembly process, where finished goods are stacked up in warehouses, waiting for customers to buy them.

The principle of “pull” is that control is transferred from the beginning of the line to the end. Thus, in the earlier example, Operation B needs to control what is transferred to assembly (Operation B). The secret of this is the “Kanban card.” A kanban card is a control device which effectively says to the recipient



(a)



(b)

Figure 1. (a) Traditional production planning system (push system). (b) JIT kanban production phasing system (pull system).

“Give me N items and N items only. When you have done that, stop! Wait until you get the next kanban card.” The diagram below now shows the changed conversation between the suppliers, Operation A and Operation B. See Fig. 1b.

Operation A now may stand idle for a while, which may seem like a cost, but it is not as great as the cost of idle inventory.

The objective of the pull systems is to see how small number of items you can pull using a kanban. A certain number is needed

to manage natural variation in the process, but no more than minimum needed for the natural variation. Reducing the number of items pulled, even by one, is an effective way of highlighting problems and bottlenecks in the system. For example, reducing the kanban number in the example to five may lead to Operation B being idle sometimes, but if some process improvement is done on it, it may even be able to handle as low as four kanban items. See the kanban determination example on pages.

CONCLUSION

Production control systems are classified as pull and push systems. In a push system, the production order is scheduled and the material is pushed into the production line. In a pull system, the start of each product assembly process is triggered by the completion of another at the end of production line.

There is a common agreement among operations management researchers that the Kanban controlled manufacturing system functions as a pull system and MRP as a push system. In this article, push and pull production systems are discussed.

MRP or JIT production system play a role as a structure of a MPC system, whereas push or pull logic is a principle only for material flow in the MPC system. Thus, it is possible for an MPC system to operate under both the push and pull principles simultaneously, depending on the stage of production. The entry ends with a push and pull MPC illustrations comparison.

REFERENCES

1. Rice JW, Yoshikawa T. A comparison of Kanban and MRP concepts for the control of repetitive manufacturing systems. *Prod Inventory Manage J* 1982;First Quarter:1–13.
2. Karmarker U. Getting control of just-in-time. *Harv Bus Rev* 1989;67(5):122–131.
3. Pyke DF, Cohen MA. Push and pull in manufacturing and distribution systems. *J Oper Manage* 1990;9(1):24–43.
4. De Toni A, Caputo M, Vinelli A. Production management techniques: push-pull classification and application conditions. *Int J Oper Prod Manage* 1988;8(2):35–51.
5. Ding F-Y, Yuen MN. A modified MRP for a production system with the coexistence of MRP and Kanbans. *J Oper Manage* 1991;10(2):267–277.
6. Olhager J, Ostlund B. An integrated push-pull manufacturing strategy. *Eur J Oper Res* 1990;45:135–142.
7. Hodgson TJ, Wang D. Optimal hybrid push/pull control strategies for a parallel multistage system: part I. *Prod Inventory Manage J* 1991;Second Quarter:74–83.
8. Hodgson TJ, Wang D. Optimal hybrid push/pull control strategies for a parallel multistage system: part II. *Prod Inventory Manage* 1991;29(7).
9. Spearman ML, Woodruff DL, and Hopp WJ. Push and pull production systems: issues and comparisons, *Oper res* 1990;40(14).
10. Spearman ML, Zazanis MA. Push and pull production systems: issues and comparisons. *Oper Res* 1992;40(3):521–553.
11. Benton WC, Shin H. Manufacturing planning and control: the evolution of MRP and JIT integration. *Eur J Oper Res* 1998;110:411–440.
12. Chaudhury A, Whinston AB. Towards an adaptive Kanban system. *Int J Prod Res* 1990;28(3):437–458.
13. Betz HJ Jr. Common sense manufacturing: a method of production control. *Prod Inventory Manage J* 1996;First Quarter:77–81.
14. Hirakawa Y, Hoshino K, Katayama H. A hybrid push/pull production control system for multistage manufacturing processes. *Int J Oper Prod Manage* 1992;12(4):69–81.
15. Deleersnyder J-L, Hodgson TJ, King RE, *et al.* Integrating Kanban type pull systems and MRP type push systems: insights from a Markovian model. *IIE Trans* 1992;24(3):43–56.
16. Hirakawa Y. Performance of a multistage hybrid push/pull production control system. *Int J Prod Econ* 1996;44:129–135.

QUALITY AND PRICING DECISIONS

CHESTER CHAMBERS
Carey Business School,
Johns Hopkins
University, Baltimore,
Maryland

When firms decide to offer a product or service, they must make integrated decisions regarding the characteristics of the product to be offered and the price that will be charged. This is particularly significant when the characteristics chosen for a product alter or determine a consumer's utility associated with its purchase. An additional layer of complexity exists when quality choices determine the product's production or delivery cost. A major area of scholarly research deals with the simultaneous selection of product quality and price. This research spans the economics, industrial organization, and operations management literatures. Intuitively, it is useful to think of quality in this context as the level of some attribute or some scalar metric representing a vector of attributes (e.g., variety, functionality, and reliability) such that all potential buyers agree that more is preferable to less. With this understanding in mind, it follows that multiple products can be presented to a potential buyer and if these products have different quality levels, it is natural to say that each offering is differentiated from the others. This notion of quality-based or "vertical" differentiation is related to, but distinct from, settings with "horizontal" differentiation. This concept relates to settings in which each buyer type can identify a product that is ideal for him but its purchase by any other type of buyer would provide less utility to that buyer than he could get from buying his ideal product. The relationship between these two notions is discussed in Refs 1 and 2.

In many instances, it is easy to see why a higher quality product would involve greater production or delivery cost. Assuming that the firm is not interested in producing a

product at a loss, it immediately follows that quality choices directly impact pricing decisions. This pair of decisions is made more complex when competing firms offer substitutable products to the same customer pool. In such cases, firms compete in how they "position" their product, where the term position refers to the product's price–quality pair. As an introduction to this topic we will discuss a collection of scenarios in some detail. First, we will deal with a monopolist choosing the profit maximizing quality level and price for a single offering. We initially consider the setting in which quality is free—meaning that production and/or delivery costs are independent of the quality of the offering. We then turn to the setting in which the cost of production is an increasing convex function of product quality. In some cases a monopolist considers multiple product offerings, either to increase profits or to make market entry by a potential competitor less attractive. We offer a few comments on this topic next. Following that we turn to the case of duopolists sharing a single market in which each firm offers a single product to the potential buyers. We then turn to more recent developments in this stream of research and end by highlighting several open questions in the field.

PRELIMINARIES

Several earlier sections of the Wiley Encyclopedia of Operations Research and Management Science provide excellent primers and background material that supplement the discussion presented here. Also, to make the nature of the pricing and quality decisions discussed here precise, we establish the defining assumptions that underlay this discussion below.

Other EORMS Articles of Interest

Speaking most generally, we can say that the quality and pricing decisions under consideration here can be viewed as economic games

involving one or more players. The concept of an economic game is introduced in the section titled “Foundations and Principles of Game Theory.” The objective of the producer of a good in such games is typically quite straight forward, and involves maximizing expected contribution or profit. The objective of the consumer of the good can be much more complex. The section titled “Modeling a Decision Maker’s Preference for Outcomes” discusses how modeling a decision maker’s preference for outcomes has been handled in prior works. Typically, the consumer is not trying to maximize cash flow or the NPV of the outcome of his actions, but some notion of utility that can be derived from the purchase or use of a product being considered. Article on deals with this topic as part of a discussion of SAUT, and, a parallel discussion of multiattribute utility theory is also discussed. An appreciation of these concepts of utility theory is needed to think through the games involved here. To understand how the games will be played out or be resolved, the reader needs an appreciation of the concept of nash equilibria and background on computing nash equilibria (see *Software for Solving Noncooperative Strategic Form Games*). Several additional subtleties of game theory, which are particularly relevant.

Mathematical Background

The marketing and industrial organization literatures often discuss the forces driving product positioning for quality differentiated products via games of vertical differentiation. The economic definition of quality for these settings is an attribute or a scalar, which represents a collection of attributes such that all buyers agree that more is preferable to less. (See, Ref. 1 for a review, and Ref. 3 for a collection of the most frequently cited research.) The notion that quality can be expressed as a scalar value is not as restrictive as it may appear. It can be argued that products are differentiated by a collection of attribute levels, which can be represented by a vector $\mathbf{w}$. Let us assume that this is true. Let us also restrict the discussion to cases in which all feasible products lie in some compact, convex set. Let us define a “generic” good with a quality level of 1. We label this good

y . If we assume that customer preferences are separable between this generic good and characteristics of the product that make it a “nongeneric” good, we can write any individual’s utility function as $u(y, \mathbf{w}) = v(y, z(\mathbf{w}))$. If we are also willing to assume that these $z(\bullet)$ functions are identical across individuals, then we can talk about quality as if it were a scalar. This approach assumes identity at the subutility level but allows individuals to differ in their marginal rates of substitution between the “generic” good and the higher quality good. While this argument is not entirely compelling, it does suggest that treating quality as a scalar value is reasonable even when goods differ regarding multiple attributes. This approach is developed in Ref. 4 and elsewhere. This highlights the fact that quality need not be a physical characteristic at all, but is a concept that relates to the utility that a potential buyer gets from use of the item in question. Consequently, vertical differentiation models have been used to explore a wide variety of differences that increase the demand for a good such as advertising, reputation, or reliability.

To make our discussion easier to follow, we will focus on a set of assumptions that dominate the literature. More specifically, we generally restrict the discussion to settings in which the following conditions hold:

1. We let q refer to the underlying hedonic attributes that characterize a particular product. Thus q is referred to as product quality. As discussed above, q is restricted to one dimension and higher values refer to products of higher quality. When participating firms consider only a single offering at a time, there is a one to one correspondence between product offerings and players in the game. In such settings, no confusion is introduced by referring to “product q ” or even “firm q ” where q refers to either the product or firm offering it.
2. Each customer buys at most one unit of the good or service under study, and has only one reservation price for any particular level of quality. Several researchers [5] have shown that this

assumption is restrictive when considering settings with multiple players or multiple product offerings if the firm's quantity decisions are also included in the analysis.

3. If costs are ignored, there are a fixed number of sellers (typically one or two). Without this assumption, entry could take place repeatedly resulting in a unique product at every quality level with each seller offering a product at marginal cost.
4. Given any finite set of products, all customers possess the same preference ranking for products within that set if they are identically priced. In some settings this is quite natural. If all other things are equal, the automobile offering higher gas mileage is simply better, one razor blade may provide more close shaves than another, one digital camera may record more pixels than another, and so on.
5. Each customer's willingness to pay for quality (θ) or their "taste" for quality is uniformly distributed between some known lower and upper bounds. We label these bounds as $\underline{\theta}$ and $\bar{\theta}$. On some level this is a rather odd assumption. Empirical investigations include Ref. 6, which uses surveys in supermarkets to measure customers' willingness to pay for more tender rib eye steaks. Virtually all customers agree that the more tender steak is of higher quality, but the willingness to pay extra for this difference is clearly not uniformly distributed. Other works [7] argue that the taste for quality is driven by income distributions, and this distribution is also far from uniformity. However, this assumption dominates the literature and offers great advantages at an expositional level. We also note that when considering any single customer, there is a one to one correspondence between his θ value and his utility specification. Therefore, no confusion should ensue when we refer to a customer or its type simply as θ .
6. For each quality level offered, the producer chooses exactly one price. Note

that when comparing two potential customers, say customers A and B , if $\theta_B > \theta_A$ then customer B would be willing to pay more than customer A for the same item. We assume throughout that this type of price discrimination is infeasible for whatever reason. For example, in retail settings, it is simply impractical to attempt to elicit the maximum amount that each customer is willing to pay for each item. Thus, the retailer sets a price for each item and customers either "take it or leave it" based on their internal valuation and the price offered.

7. If costs are invariant in product quality, a maximum achievable quality level $q_{\max}$ exists and is known to all parties in the game. We also make use of the assumption that a minimum allowable quality level $q_{\min} > 0$ exists and is known by all parties. These assumptions guarantee that an equilibrium can exist in such settings. The rationale for the existence of a value of $q_{\max}$ may stem simply from the "state of the art" in the production of the item under consideration. The rationale for the existence of a strictly positive value of $q_{\min}$ may stem from regulatory factors or liability concerns.
8. Market size is assumed to be fixed over time and is normalized to 1. This makes it natural to focus on a representative shopper seeking a single unit of a single good. It also allows us to discuss market shares or demand levels quite easily as functions of q , p , and θ .

Assumptions 5 and 6 imply that the buyer's surplus utility resulting from the purchase of a unit with quality q at price p can be written as $U = \theta q - p$. This utility function has dominated the literature in this field for several decades. However, we discuss an alternate formulation of the problem when considering emerging works.

THE QUALITY AND PRICE SETTING MONOPOLIST

We first consider the setting in which quality is free—meaning that production and/or delivery costs are independent of the quality of the offering. We then turn to the setting in which the cost of production is an increasing, convex function of product quality. In both instances we focus on the profit maximizing selections of price–quality pairs by a monopolist.

Optimal Behavior when Quality Is Free

It is useful here to follow the example developed in Gabszewicz and Wauthy [5]. Consider a monopolist who will choose a price and quality pair for a single offering when production cost is independent of product quality. This assumption is difficult to accept in many manufacturing settings, but does appear to be reasonable in some settings including that for information goods as described in Bhargava and Choudhary [8]. We note that the assumption that production costs are literally 0 need not hold for the following analysis to be relevant; only that the differences in production costs between goods of differing quality levels are negligible. Since any potential buyer's utility level is strictly increasing in product quality, if quality is not costly, the monopolist will offer a product with the best possible quality ($q_{\max}$). It is important to note that when the monopolist offers this quality level, he/she is also choosing the quality level that maximizes social welfare. Thus the only loss in efficiency that arises in this setting is from price selection. Also, consistent with Bhargava and Choudhary [8], we assume that consumers' tastes for quality are uniformly distributed over the interval $[0, 1]$. Now, if a consumer of type θ buys a good with quality $q_{\max}$ and price p , his utility is given by $U = \theta q_{\max} - p$. On the other hand, if he abstains from buying, his utility is simply 0. Therefore, market demand is derived by identifying the consumer θ_1 who is indifferent between the alternatives of buying one unit at price p or not buying at all. Solving $U = \theta_1 q_{\max} - p = 0$, we get $\theta_1 = p/q_{\max}$. Consequently, market demand is simply $D(p, q_{\max}) = 1 - p/q_{\max}$. Since the

monopolist sells his/her output at a per unit price p , we can write his/her profit level as $\Pi = (1 - p/q_{\max})p$. This value is maximized when $p^* = q_{\max}/2$ and the corresponding profit level is $\Pi^* = q_{\max}/4$. Salant [9] and Acharyya [10] have shown that it is sub-optimal for the monopolist to offer multiple products in this setting unless he/she anticipates possible entry by another player. Thus, the highest achievable quality level and a corresponding price level of $q_{\max}/4$ completely describes the outcome in this scenario.

Optimal Behavior when Quality Is Costly

In this section, we build on the approach developed in the seminal work by Mussa and Rosen [11]. One key problem for the monopolist in this setting arises from assumption 5. Given two customer types, say A and B with $\theta_B > \theta_A$, the monopolist knows that type B would be willing to pay more for any product offering than type A . Thus, ideally (from the monopolist's perspective), he/she would charge different prices to the two customer types. However, if this is not practical, then the problem of selecting which products to offer and at what prices becomes more involved. The unit production or delivery cost of any particular quality, $C(q)$, is assumed to be constant for any level of q . In other words, we ignore economies of scale in this development. The unit production or delivery cost is assumed to be an increasing and convex function of product quality. In other words, $C'(q) > 0$ and $C''(q) > 0$. Each customer's surplus utility derived from the purchase of a product is written as $U = \theta q - p$ as explained earlier. Note that if no barriers to entry are present (such as a fixed cost of entry), competition would result in an equilibrium outcome with a product offered at every possible quality level, and this product would be priced at marginal cost. Hence, we would have $P(q) = C(q)$, or market price would equal the production cost for any and all quality levels. Given such a price schedule, each customer type would choose the quality level that maximized his utility. This results in a first-order condition for each buyer of $P'(q) = U'(q) = \theta$. In other words, each buyer purchases a unit such that the marginal cost of increasing the

quality of that unit has the same value as the customer type. Furthermore, each customer type obtains the highest quality good that he/she is willing to buy at his/her marginal cost. The efficiency of such an outcome is quite appealing, but the monopolist finds such an outcome wholly unsatisfactory in that he/she earns no profit. Consequently, it is entirely clear that the monopolist will not work to bring about such an outcome.

To understand how the monopolist will behave, we follow a simple example. Given any continuous, differentiable cost function, it is natural to discuss a curve $C(q)$ that relates cost to quality. For this discussion we let q range from 0 to 1 and assume that $C(q) = q^2$. Now, assume that the market consists of only 2 types of customers, A and B . Let type A have $\theta_A = 0.2$ and let $\theta_B = 1$. The efficient solution includes a product targeting type A with $P(q_A) = C(q_A)$ and $P'(q_A) = \theta_A = C'(q_A) = 2q_A$. Consequently, we end up with $q_A = \theta_A/2 = 0.1$ and $P_A = \theta_A^2/4 = 0.01$. A parallel analysis shows that the efficient solution includes $q_B = 0.5$ and $P_B = 0.25$. Note that type A is left with a surplus of $U = \theta_A(q_A) - p_A = \theta_A(\theta_A/2) - \theta_A^2/4 = \theta_A^2/4$. Similarly type B ends up with a surplus of $U = \theta_B^2/4$.

One approach to understand how the monopolist will behave is to ask how he/she would alter this menu of offerings to extract this surplus from the two buyer types. This directly increases the producer's profits. It is clear that if the monopolist had the power to offer different prices to the different customer types and could restrict each customer type to have access to only one product, he/she would offer quality levels $q_A = \theta_A/2 = 0.1$ and $q_B = \theta_B/2 = 0.5$ and charge respective prices to types A and B of $\theta_A^2/4$ and $\theta_B^2/4$. In this case, the producer maximizes profits and no buyer gets a positive surplus from the purchase of either item. However, this outcome is not feasible if both products are simultaneously offered at these prices and customers are free to choose which product they buy. To understand why, consider the consumer's surplus if type A buys the product which targets type B and vice versa. Type A gets a surplus of $U_A = \theta_A(\theta_B/2) - \theta_A^2/2$. Since $\theta_A < \theta_B$ this value is strictly negative. On the

other hand, $U_B = \theta_B(\theta_A/2) - \theta_A^2/2$ and this value is strictly positive. Thus, if two buyer types exist and the monopolist offers the "efficient" quality levels to the two types and prices them in a way intended to extract all surplus, the high type buyer would prefer the lower quality product. The low type buyer receives no surplus, but the high type buyer receives a positive surplus and reduces the monopolist's profits. Consequently, offering these two product positions is not optimal to the monopolist.

Note that if these two product positions are considered, the monopolist can increase his/her profits by reducing the price of product q_B just enough to entice the higher type buyer to acquire it. This can be accomplished by pricing product B at $\theta_A q_A + \theta_B(q_B - q_A)$. At these price points, the monopolist extracts all surplus from type A and some of the surplus from type B . This is better for the monopolist but still not the best that he/she can do in this setting for the following reason. If the quality of the lower quality good is reduced by a small amount, say Δq and its price is reduced by $\theta_A \Delta q$, consumers of type A are content to buy this new lower quality item and the buyer of type B still prefers the higher quality good. However, the high type buyer is much more sensitive to this drop in quality when compared to the lower buyer type. By reducing the quality targeting type A by an amount Δq the monopolist can actually increase the price charged to type B by an amount $(\theta_B - \theta_A)\Delta q$ without inducing buyers to shift which product type they prefer. Therefore, it pays for the monopolist to reduce the quality offered to "low" types to increase the revenue earned from the "high" types.

Though this example is greatly simplified, its logic applies in general and leads to several important results. First, if $\theta \sim \text{Uniform}[\underline{\theta}, \bar{\theta}]$, then the buyer of type $\bar{\theta}$ is offered the same quality product whether the market is perfectly competitive or served by a monopolist. For every other consumer that the monopolist chooses to serve, the monopolist sells a lower quality good than the buyer would obtain under perfect competition. We reiterate the key point; offering products that lower type buyers will buy makes it impossible for the monopolist to extract all surplus

from the higher type buyers. Consequently, it sometimes becomes optimal for the monopolist to not serve lower type buyers at all. The range of buyer types that the monopolist serves is never greater than that served under perfect competition. Also, for the set of consumers that the monopolist chooses to serve, the difference between maximum and minimum quality offerings is always greater than that would be the case under competition. This is true because the upper end of the quality levels offered is unchanged and the lower bound is always driven downward by the effort to discriminate among buyer types. For any quality level sold both by the monopolist and under competition, the monopoly price is higher than the competitive price and the price differential increases in q . The monopolist uses his/her power to extract as much surplus from consumers as possible. The higher type buyers have more surplus to extract, and the higher the type, the more the monopolist is willing to sacrifice when dealing with low types to get it. Consequently, the higher the type, the greater is the gap between prices charged in perfect competition and that charged under monopoly.

COMPETING DUOPOLISTS

The discussion of the monopolist in the section titled “The Quality and Price Setting Monopolist” raises the natural question as to why the market should contain only one entrant. This question is particularly important if production costs are assumed to be 0. To gain some understanding about how competing firms will behave in such a setting, we consider three scenarios that parallel our earlier discussion. We begin with a treatment of cases in which production costs are ignored. We then turn to cases in which a firm’s fixed cost is increasing and convex in quality but variable costs are ignored. This can be contrasted with settings in which fixed costs are ignored, but variable costs are assumed to be convex and increasing in product quality.

Optimal Behavior of Duopolists when Quality Is Free

The analysis of this problem proceeds slightly differently depending on whether the market

is covered or not, where covered means that each potential buyer actually purchases a product. The setting in which we assume that the market is to be covered dates to the seminal works of Gabszewicz and Thisse [12] and Shaked and Sutton [13]. The notation used is most similar to that described by Tirole [1]. We again assume that the customers’ preferences are described by $U = \theta q - p$ and that $\theta \sim \text{Uniform}[\underline{\theta}, \bar{\theta} = \underline{\theta} + 1]$. We now consider two firms (1 and 2), each of which offers a single product of quality q_i with $q_2 > q_1$. Tirole [1] makes two assumptions in addition to those mentioned earlier.

1. $\bar{\theta} \geq 2\underline{\theta}$. This guarantees that the consumer heterogeneity is sufficient for what follows.
2. $c + \frac{\bar{\theta} - 2\underline{\theta}}{3}(q_2 - q_1) \leq \underline{\theta}q_1$. In this equation the variable production cost is fixed at c . Without loss of generality, we normalize this value to $c = 0$ in the discussion that follows. This guarantees that in the price equilibrium the market is covered. This means that each customer buys one of the two brands offered. It has been argued that this assumption is justified when considering goods or services deemed to be essential.

We note that many subsequent works in this literature normalize the problem such that $\theta \sim \text{Uniform}[0, 1]$. This approach satisfies assumption 1, but violates assumption 2, and guarantees that the market will not be covered. This is clear because for a customer with $\theta = 0$ we have $U = -p$, which guarantees that for some segment of the market no product will be purchased at any price. It is sometimes argued that this is reasonable for discretionary goods since it is clear that not all potential buyers will buy at least one of every type of items offered.

It is natural to model the game as playing out over two stages. In the first stage, firms simultaneously select quality levels. In the second stage they simultaneously select prices. This is consistent with the notion that it is far easier to set product prices than to alter product quality levels. Given quality levels $q_2 > q_1$, it is clear that if all customer

types purchase one item we are interested in cases in which high- θ customers buy the high quality good and low- θ customers buy the low quality good. A consumer of type θ is indifferent between the two products if $\theta q_1 - p_1 = \theta q_2 - p_2$. If we normalize the market size to 1 then, given these assumptions, each firm's market share, or each product's demand, can be equivalently written as $D_1(p_1, p_2) = \frac{p_2 - p_1}{q_2 - q_1} - \underline{\theta}$ and $D_2(p_1, p_2) = \bar{\theta} - \frac{p_2 - p_1}{q_2 - q_1}$. In equilibrium, each firm i maximizes $\pi_i = (p_i - c)D_i(p_i, p_j)$ with respect to p_i . The reaction functions are $p_2 = R_2(p_1) = (p_1 + c + \bar{\Delta})/2$ and $p_1 = R_1(p_2) = (p_2 + c - \bar{\Delta})/2$ where $\bar{\Delta} \equiv \bar{\theta}(q_2 - q_1)$ and $\underline{\Delta} \equiv \underline{\theta}(q_2 - q_1)$. We can say that $\underline{\Delta}$ and $\bar{\Delta}$ are the monetary values of the quality difference to the lowest and the highest type consumers. The Nash equilibrium prices are directly stated as $p_1^* = c + \frac{\bar{\theta} - 2\underline{\theta}}{3}(q_2 - q_1)$ and $p_2^* = c + \frac{2\bar{\theta} - \underline{\theta}}{3}(q_2 - q_1)$. We note that it immediately follows that $p_2^* > p_1^*$. The resulting demand levels are $D_1^* = (\bar{\theta} - 2\underline{\theta})/3$ and $D_2^* = (2\bar{\theta} - \underline{\theta})/3$. In addition, we see that equilibrium profit levels are $\Pi_1^*(q_1, q_2) = (\bar{\theta} - 2\underline{\theta})^2(q_2 - q_1)/9$ and $\Pi_2^*(q_1, q_2) = (2\bar{\theta} - \underline{\theta})^2(q_2 - q_1)/9$.

Now consider the two-stage game in which firms choose quality levels first. The decisions in this stage must anticipate the price equilibrium described above. Note that each firm's equilibrium profit level is monotonic in its selection of quality. Firm 2 wishes to make q_2 as large as possible because Π_2^* is strictly increasing in q_2 . This is also most attractive for Firm 1 since Π_1^* is also increasing in q_2 . Simultaneously, we see that Π_1^* is strictly decreasing in q_1 , as is Π_2^* . Thus, both firms seek "maximum differentiation" meaning that they each benefit from having as large a difference in selected quality levels as possible. Thus, we see that quality-based competition differs dramatically from price-based competition in that the equilibrium results in positive profit levels for both players. Each firm acts as a "near-monopolist" in its own segment of the market. The more dissimilar the two offerings, the less the customers in each segment will value the product offered by the firm serving the other segment.

A study of the second condition shown above shows that cases will exist with a positive lower bound on taste but the second assumption does not hold. In such cases, equilibrium behavior would leave the market uncovered. Choi and Shin [14] offer a comment on the work by Tirole [1] that extends the analysis to this setting. The key difference is that in such cases demand for product 1 can be written as $D_1(p_1, p_2) = \frac{p_2 - p_1}{q_2 - q_1} - \frac{p_1}{q_1}$. The main result is that in this setting the equilibrium will include $q_2 = q_{\max}$ and $q_1 = \frac{4q_{\max}}{7}$. The firms choose a high level of differentiation, but not as great as is possible. The key insight added is that cases exist in which some buyers' taste for quality is so low that the monopolist is better off not serving them even when it is costless to produce a product they would be willing to buy. We explore the rationale behind such behavior on the part of the firm below.

Optimal Behavior of Duopolists when the Fixed Cost of Quality is Increasing and Convex

The assumption that producing a unit of higher quality implies no additional costs is problematic for many product categories. However, in some settings the assumption that variable production costs are not sensitive to product quality while fixed costs, such as product design and marketing expenditures, are does seem reasonable. For example, Bhargava and Choudhary [8] argue that these are quite reasonable assumptions for information goods. To develop insights into this class of problems, we present a discussion motivated by Lehmann-Grubbe [15].

This model considers duopolists competing via a two-stage game in which the fixed costs associated with product quality are represented by a continuous, convex function. In the first stage, the firms choose quality levels. With no loss of generality, we can label the players such that $q_1 \leq q_2$. As assumed earlier, each consumer's utility is expressed as $U = \theta q - p$ and $\theta \sim \text{Uniform}[0, 1]$. What differs here is that the fixed cost of production $C(q)$ is increasing and convex with $C(0) = 0$, $C'(0) = 0$, $C'(q > 0) > 0$, $C''(q) > 0$, and $\lim_{q \rightarrow \infty} C'(q) = \infty$. Given this structure, it is

easy to show that the equilibrium revenue levels can ultimately be written as simple functions of the quality levels of the two product offerings.

$$R_1^* = \frac{q_1 q_2 (q_2 - q_1)}{(4q_2 - q_1)^2},$$

$$R_2^* = \frac{4q_2^2 (q_2 - q_1)}{(4q_2 - q_1)^2}.$$

Note that since $q_2 > q_1$, it immediately follows that the firm offering the higher quality product receives greater revenue in equilibrium. However, this does not prove that this firm will have higher profits because costs rise with quality levels as well. Let us assume for the moment that the firms choose quality levels simultaneously in stage 1 of the game and that a Nash equilibrium exists. In order for this to be the case, we must have marginal revenue equal to marginal costs at the selected quality levels. In other words, we will see

$$\frac{\partial R_1^*}{\partial q_1} = C'(q_1^*) = q_2^{*2} \frac{4q_2^* - 7q_1^*}{(4q_2^* - q_1^*)^3}$$

and

$$\frac{\partial R_2^*}{\partial q_2} = C'(q_2^*) = 4q_2^* \frac{4q_2^{*2} - 3q_1^* q_2^* + 2q_1^{*2}}{(4q_2^* - q_1^*)^3}.$$

To show the main result more clearly let us define $0 < a < 1$ and $0 < z$ such that $az \equiv q_1^* < q_2^* \equiv z$. Thus, the conditions defining an equilibrium can be rewritten as

$$\frac{\partial R_1^*}{\partial q_1} - C'(q_1^*) = \frac{4 - 7a}{(4 - 7a)^3} - C'(az) = 0$$

and

$$\frac{\partial R_2^*}{\partial q_2} - C'(q_2^*) = \frac{4(4 - 3a + 2a^2)}{(4 - a)^3} - C'(z) = 0.$$

The essence of the proof of the main result is that the authors are able to show that the convexity of the cost function along with the equations shown above are sufficient to prove the following:

$$4z \frac{1 - a}{(4 - a)^2} - C(z) > az \frac{1 - a}{(4 - a)^2} - F(az).$$

This, in turn, is sufficient to show that

$$R_2^* - C(q_2^*) > R_1^* - C(q_1^*).$$

In other words, the difference between revenue and production cost for the producer of the higher quality product is greater than the net revenue for the producer of the lower quality product.

Shaked and Sutton [13] argue that this type of setting is a “natural oligopoly” if there is any fixed cost associated with market entry. The firms need to have some distance between their offering and alternative offerings to generate profits. If we envision customer types spread uniformly over some line segment from $\underline{\theta}$ to $\bar{\theta}$, the revenue that a firm needs to cover its fixed costs must be proportional to the length of the line segment that the firm serves. Since the range of consumer tastes is finite, the presence of fixed costs means that the market can sustain only a finite number of competitors. If fixed costs are high enough, the market becomes a natural monopoly and this firm behaves as described earlier.

Additional application and consideration of this model includes the work of Donnenfeld and Weber [16], in which the problem of market entry when the market is large enough to support more than one firm is discussed. In this case if one firm is a first mover, it may be motivated to offer more than one product specifically to cover enough of the market to make market entry unprofitable for a second mover. It is important to note that if firms enter a new market sequentially, each player chooses at most one position, and firms focus only on the competitors present when it makes its positioning decision—the second firm that enters typically earns lower profits than the third. This happens because the first two entrants maximize the quality difference between them. Thus, one firm goes to the top of the quality scale ($q_{\max}$) while the other goes to the bottom ($q_{\min}$). If a third entrant arrives, it takes a position between the two and the resulting equilibrium profit levels have the characteristic that profit levels are increasing in quality. By understanding how the profit levels will evolve if additional players enter a market, an early mover is able

to choose positions that deter others from joining the game.

Optimal Behavior of Duopolists when the Variable Cost of Quality Is Increasing and Convex

Many settings exist in which the fixed cost associated with a higher quality product are not appreciably higher than the fixed cost associated with a lower quality product. In other settings, the fixed cost is sunk by the time the firm considers later decisions, but the variable costs of production vary dramatically with product quality. Gabszewicz and Thisse [12] motivate one of the seminal works in this literature using the purchase of a grand piano as a central example. Many observers including Barron [17] have focused on the production process for these items and found that the product designs for these objects do not change very often but that the variable costs of higher quality offerings do differ dramatically from those of lower quality substitutes because of the use of more expensive materials, higher skilled labor, and different production processes. Consequently, a need exists to consider settings in which the fixed costs of production are sunk but variable costs increase in a convex way with product quality.

The bulk of the research in this literature stream uses a formulation initially presented by Moorthy [18]. This work considers competing duopolists in which each firm's marginal cost of supplying a product of quality q is αq^2 with $\alpha > 0$. Again, it is assumed that customer types are uniformly distributed over some interval $[\underline{\theta}, \bar{\theta}]$, $0 < \underline{\theta} < \bar{\theta}$. The key element here is that the marginal cost of increasing quality grows no slower than linearly. Consequently, a quality level exists at which no producer would consider production because the production cost would exceed the amount that the most quality-conscious consumer would be willing to pay.

As before, we consider a two-stage game. In stage 1, the two firms simultaneously choose quality levels, and in stage 2 they choose equilibrium prices. Consider a setting with three products with three quality levels, $0 < q_1 < q_2$. Selecting a product with a quality level of 0 is synonymous with buying

nothing. If we consider these quality levels to be fixed for a moment, we can identify a customer type θ_1 who is indifferent between buying product q_1 and buying nothing. For this customer, the utility of purchasing product 1 is simply $U = \theta_1 q_1 - p_1 = 0$. Thus, we can see that $\theta_1 = \frac{p_1}{q_1}$. Simultaneously, there exists a customer who is indifferent between purchasing products q_1 and q_2 . For this customer, we will have $U = \theta_2 q_1 - p_1 = \theta_2 q_2 - p_2$. We can solve directly to see that $\theta_2 = \frac{p_1 - p_2}{q_1 - q_2}$. Thus, all customer types with $\theta \leq \frac{p_1}{q_1}$ will purchase nothing. Customers with $\frac{p_1}{q_1} < \theta \leq \frac{p_1 - p_2}{q_1 - q_2}$ will purchase product q_1 . Finally, all customers with $\theta > \frac{p_1 - p_2}{q_1 - q_2}$ will purchase the product with quality q_2 . The first-order conditions for a price equilibrium with $\underline{\theta} \leq \theta_1 \leq \theta_2 \leq \bar{\theta}$ are given by

$$\begin{aligned} p_1^* - \alpha q_1^2 &= \left(\frac{p_2 - p_1^*}{q_2 - q_1} - \frac{p_1^*}{q_1} \right) \left(\frac{q_1}{q_2} \right) (q_2 - q_1), \\ p_2^* - \alpha q_2^2 &= \left(\bar{\theta} - \frac{p_2^* - p_1}{q_2 - q_1} \right) (q_2 - q_1). \end{aligned}$$

Note that these two statements are simply the trade-offs between profit margins and market shares that each firm must serve to maximize profits. Rearranging terms shows that each firm's best price is simply the average of its marginal cost and the price that gives all of the market to its opponent, assuming that this price is greater than marginal cost. Consideration of these two functions simultaneously suggests a price equilibrium at which

$$\begin{aligned} p_1^* &= \alpha q_1^2 + \left[\frac{q_1(q_2 - q_1)}{(4q_2 - q_1)} \right] [\bar{\theta} + \alpha(q_2 - q_1)], \\ p_2^* &= \alpha q_2^2 + \left[\frac{2q_2(q_2 - q_1)}{(4q_2 - q_1)} \right] [\bar{\theta} - \alpha(q_2 + q_1/2)]. \end{aligned}$$

Finally, each firm's equilibrium profits are

$$\begin{aligned} \Pi_1^* &= q_1 q_2 (q_2 - q_1) \left[\frac{\bar{\theta} + \alpha(q_2 - q_1)}{4q_2 - q_1} \right]^2, \\ \Pi_2^* &= (q_2 - q_1) \left[\frac{2\bar{\theta} 2q_2 - \alpha q_2^2 - q_2(q_1 + q_2)}{(4q_2 - q_1)} \right]^2. \end{aligned}$$

With these profit functions in place, it immediately follows that firm 1's optimal

product satisfies the following first-order condition:

$$(q_2 - 2q_1)(4q_2 - q_1)(\bar{\theta} + \alpha(q_2 - q_1)) + 2q_1(q_2 - q_1)(\bar{\theta} - 3\alpha q_2) = 0.$$

The optimal product for firm 2 satisfies

$$q_2(4q_2 - q_1)(\bar{\theta} - \alpha(q_2 + q_1/2)) + (q_2 - q_1) \times (\alpha q_1^2 - 2q_1(\bar{\theta} - 2\alpha q_2) - 8\alpha q_2^2) = 0.$$

The simultaneous solution of these two first-order conditions yields the key result. As long as $\bar{\theta} \geq 2.27505\underline{\theta}$, an equilibrium will exist with $q_1^* = 0.2474\bar{\theta}/\alpha$ and $q_2^* = 0.5288\bar{\theta}/\alpha$. The equilibrium prices turn out to be $p_1^* = 0.109\bar{\theta}^2/\alpha$ and $p_2^* = 0.335\bar{\theta}^2/\alpha$. The resulting profit levels are calculated directly as $\Pi_1^* = 0.0173\bar{\theta}^3/\alpha$, and $\Pi_2^* = 0.0109\bar{\theta}^3/\alpha$. One important observation here is that the producer of the lower quality product earns greater profits in equilibrium. This stands in contrast with the results found when the quality related costs are fixed costs. This is understandable that when we note that the market share for firm 2 increases linearly in q_2 , but its cost increase is convex in q_2 . We also note that as $\bar{\theta}$ increases, the quality equilibrium includes an increase in quality for both products, but the gap between them also increases (if $\underline{\theta}$ is held constant). On the other hand, as α increases, both products decrease in quality, the gap between them decreases, as does profit for both players.

Several important extensions of these ideas have been developed more recently. These include work of Noh and Moshini [19], which considers market entry by a second firm when this entry requires the payment of a fixed fee that is independent of product quality, and both firms' variable production costs are quadratic functions of q . Rhee [20] considers settings in which products are differentiated with respect to quality but are also differentiated along another unobservable dimension. The presence of this unobserved attribute may explain why product positions in practice do not seem to differ as much as economic models

would suggest. Schmidt and Porteus [21] consider the interrelated problems of making investments that probabilistically lead to the ability to offer a product of higher quality and facing higher production cost if the higher quality product is introduced. One process describes the outcome of the investments in technology, while the model developed by Moorthy [18] is used to explain the quality–pricing game that follows.

RECENT DEVELOPMENTS

The work of Moorthy [18] was highly influential in this research stream because it was the first to include production costs in a meaningful way. The formulation in that work assumed that $c(q) = \alpha q^2$. This leads to very specific statements about equilibrium prices, profit levels, and market structures. Several subsequent works build on this foundation to look at related problems. For examples, see Rhee [20], Desai and Srinivasan [22], and Villas-Boas [23]. However, it remains an open question exactly how robust the essence of these results are when a problem setting does not perfectly match the assumptions of Moorthy [18].

Recent work that explores how generally these earlier results hold include that of Jing [24]. This work addresses the question of whether it is more profitable to offer the higher quality product depending on the variable cost function in place. Consider a market with K firms and each firm produces a good of quality q_k at a constant unit cost c_k , and there are no fixed costs. The firms are indexed such that $0 < q_1 < q_2 < \dots < q_K$. Thus each firm occupies a point (c_k, q_k) . The author then develops a way to rank the firms' relative cost efficiency. This is done by comparing each firm's cost–quality pair with those of its closest neighbors. One of the main findings is that the equilibrium profit levels of the firms will match the ranking of the relative cost efficiencies. The basic message is that if a firm's cost does not rise “too fast” with an increase in quality, the higher quality product will be more profitable. Of course, the phrase “too fast” must be understood relative to its closest competitors.

Recent work by Chambers *et al.* [25] explores the impact of altering the distribution of consumer taste among the population, to consider $U = \theta q - p$ when $f(\theta) = 1/\theta^2$ and $\theta \in [1, \infty)$. By dividing both sides of the utility function by θ and rescaling the problem, their work approximates the utility function by using $U = q - \theta p$ with $\theta \sim \text{Uniform}[0, 1]$. This simple transformation yields at least three unexpected and interesting results. First, allowing θ to be uniform on $[0, 1]$ accommodates problems with either covered or uncovered markets. Second, on the basis of the assumptions made in Chambers *et al.*'s work, the equilibrium price for the lower quality good equals the production cost of the higher quality good. This is a critical point, because it shows how the management of cost for one player affects the ability of its rival to manage the margin on its product offering. Third, their work finds that the behavior of competing duopolists leads to an equilibrium in which the profit level for the producer of the higher quality good is $\Pi_2 = q_2 - q_1$. This is true because the producer of the high quality good maintains enough price flexibility to drive the equilibrium price for the low quality good to the cost level of the high quality product. This fact allows the higher quality producer to compete "purely" on product quality. Thus, the quality gap defines his/her profit level.

The question of whether the market is covered or not is a bit more complex. It is easy to produce cost functions that produce either outcome. However, the authors are able to prove another useful result. If $c(0) = 0$, $c(q)$ is strictly increasing, twice differentiable, and convex on $[0, q_{\max}]$, and the average variable cost function $c(q)/q$ is convex and log-concave on $[0, q_{\max}]$, then the producer of the lower quality good earns higher profits when $q_L \leq c_H$. In addition, the authors show that it is never optimal for the lower quality producer to offer a product with a quality level above the minimum level that would cover the market for this class of cost functions. This result was developed assuming that both firms have identical cost functions, but does not assume any particular functional form. However, there is really no convincing argument that the cost functions

must always be equivalent. An extension of this work by Chambers and Ramachandran [26] has shown that this result holds even when the firms have different cost functions of the form $\alpha_i q_i^{r_i}$ where α_i , q_i , and r_i are each firm specific.

Underdeveloped Research Areas

Several important topics remain underdeveloped in this research stream. First, it is important to relax the assumption that all players in an industry and those considering market entry have identical cost functions. Many explanations are available which explain why firms have differing cost functions including learning effects, exchange rates, investment levels, and other elements of firm strategy. It is also important to consider both fixed and variable costs simultaneously. Many automated production approaches have higher fixed cost but lower variable cost when compared to older, "craftsman"-based systems they may be replacing. On the other hand, many emerging economies leverage their lower labor cost to use more labor-intensive approaches to compete with incumbent firms. Consequently, price-quality decisions are actually intertwined with technology management and/or outsourcing decisions.

Another important issue is that there is no good reason to believe that the distribution of taste parameters among consumers is stagnant. Consumers' expectations regarding product quality evolve over time and experience may drive the average taste upward for some items and downward for others. This could drive down the appeal of the lower quality good over time unless the quality of that offering keeps pace with changing consumer sentiment. Just as likely, it could force the producer of the higher quality good to lower his price over time if $q_{\max}$ cannot consistently be raised. Another problem is that deteriorating economic conditions may alter income distributions. This may serve to drive consumer tastes downward, increasing the market share of a lower quality offering. This may explain why sales at Wal-Mart and McDonald's rise while sales at Neiman Marcus and many exclusive restaurants decline during recessionary periods.

While this article has focused on the modeling of quality and pricing decisions, we would be remiss if we do not mention the need for empirical validation of the underlying assumptions embedded in these models. Some models do assume a general distribution of taste parameters but some nuances of the problem of product positioning must be sensitive to the functional form of that distribution. On the other hand, it is not at all clear that firm behavior is consistent with the economic conclusions. Stronger connections between economic models and firm practice is always enlightening.

REFERENCES

1. Tirole J. The theory of industrial organization. Cambridge (MA): MIT Press; 1988.
2. Cremer H, Thisse J. Location models of horizontal differentiation: a special case of vertical differentiation models. *J Ind Econ* 1991;39:383–390.
3. Thisse J-F, Norman G. Volumes 1 and 2, The economics of product differentiation. Brookfield (IL); Aldershot Publishing in the UK; 1994.
4. Beath J, Katsoulacos Y. The economic theory of product differentiation. New York: Cambridge University Press; 1991.
5. Gabszewicz JJ, Wauthy XY. Quality underprovision by a monopolist when quality is not costly. *Econ Lett* 2002;77:65–72.
6. Lusk JL, Fox JA, Schroeder TC, *et al.* In-store valuation of steak tenderness. *Am J Agr Econ* 2001;83:539–550.
7. Philips L, Thisse JF. Spatial competition and the theory of differentiated markets: an introduction. *J Ind Econ* 1982;31:1–9.
8. Bhargava H, Choudhary V. Information goods and vertical differentiation. *Res Pricing Econ Theory* 2001;18:89–106.
9. Salant SW. When is inducing self-selection suboptimal for a monopolist? *Q J Econ* 1989;104:391–397.
10. Acharyya R. Monopoly and product quality: separating or pooling menu? *Econ Lett* 1998;61:187–194.
11. Mussa M, Rosen S. Monopoly and product quality. *J Econ Theory* 1978;18:301–317.
12. Gabszewicz J, Thisse J. Price competition, quality and income disparities. *J Econ Theory* 1979;20:340–359.
13. Shaked A, Sutton J. Natural oligopolies. *Econometrica* 1983;51:1469–1484.
14. Choi JC, Shin HS. A comment on a model of vertical product differentiation. *J Ind Econ* 1992;60:229–231.
15. Lehmann-Grubbe U. Strategic choice of quality when quality is costly: the persistence of the high-quality advantage. *Rand J Econ* 1997;28:372–384.
16. Donnenfeld S, Weber S. Vertical product differentiation with entry. *Int J Ind Org* 1992; 10:449–472.
17. Barron J. Pianos: the making of a Steinway concert grand. Boston (MA): Time Press; 2006.
18. Moorthy K. Product and price competition in a duopoly. *Mark Sci* 1988;7:141–169.
19. Noh Y, Moshini G. Vertical product differentiation, entry-deterrence strategies, and entry qualities. *Rev Ind Org* 2006;29:227–252.
20. Rhee B. Consumer heterogeneity and strategic quality decisions. *Manage Sci* 1996;42: 157–172.
21. Schmidt G, Porteus E. Sustaining technology leadership can require both cost competence and innovative competence. *Manuf Serv Oper Manage* 2000;2:1–18.
22. Desai P, Kekre S, Radhakrishnan S. Product differentiation and commonality in design: balancing revenue and cost drivers. *Manage Sci* 2001;47:37–51.
23. Villas-Boas J. Product line design for a distribution channel. *Mark Sci* 1998;17:156–169.
24. Jing B. On the profitability of firms in a differentiated industry. *Mark Sci* 2006;25: 248–259.
25. Chambers C, Kouvelis P, Semple J. Quality-based competition, profitability, and variable costs. *Manage Sci* 2006;52(12):1884–1895.
26. Chambers C, Ramachandran K. Firm performance improvement strategies in vertically differentiated markets: *Cox Working Paper*; 2008. pp. 1–22.

QUALITY DESIGN, CONTROL, AND IMPROVEMENT

AMITAVA MITRA
College of Business,
Auburn University,
Auburn, Alabama

INTRODUCTION

Quality is fundamental to the success of all manufacturing and service processes. Management plays a critical role in ensuring quality in all phases, be it in the design of products/processes, the manufacturing of products or delivery of services, or in the operational phase of the product when used by the consumer.

This article links the concepts of quality design, quality control, and quality improvement (QI) and provides a perspective on the role of each of these functions. These functions are inter-related to each other and knowledge from subsequent phases is used as a feedback to ascertain action plans in the individual phases. Hence, creation of a quality design incorporates information from the processes that will be used to produce the product. Further, if product performance when actually used by the customer is not acceptable, it may influence design quality as well. In the following sections, the article describes the concepts of quality design, control of quality, and QI, respectively. An overview that links these functions is first presented.

QUALITY OF DESIGN, CONTROL, AND IMPROVEMENT FUNCTIONS

It is well understood that quality plays an important role in the production and performance of a product or the delivery of a service that meets the needs of the customer. The functions of designing a quality product or a process, maintaining stability in the process, and creating modifications to the product or

process design to adapt to the dynamic needs of the customer are integrally related to each other. The functions of quality design, quality control, and QI work in unison, each influencing and being influenced by the others.

Figure 1 provides a view of the conceptual linkage between the quality design, control, and improvement functions. As with any product or service, the first and foremost goal is meeting or exceeding the needs of the customer. Three prioritized categories of needs exist. Firstly, basic needs are those that are absolutely expected by the customer. While meeting these needs will not necessarily improve customer satisfaction, not meeting them will create a dissatisfied customer. Secondly, performance needs, on the other hand, typically influence customer satisfaction in a linear manner. Thirdly, excitement needs influence customer satisfaction exponentially. Such needs may not even be known or expected by the customer. Hence, their satisfaction thrills the customer.

All product/process designs are constrained by the availability of a limited budget. Further, it must be feasible to manufacture the product incorporating the developed design. Thus, product and process design go hand-in-hand. Once a feasible product and process design is established, the manufacturing phase must take place. Here also, there are constraints imposed by demand for various products, availability of raw material and parts, lead time for vendors, and availability of resources imposed by schedule constraints. The function of quality control deals with ensuring that the process is in a state of statistical control, that is, only common causes inherent to the system exist. Control chart methodology is used in this context. Even if a process is in a state of statistical control, it is still possible for the product to not meet the specifications as desired by the customer. If this is, in fact, the case, we conclude that the existing process is not capable and seek QI strategies

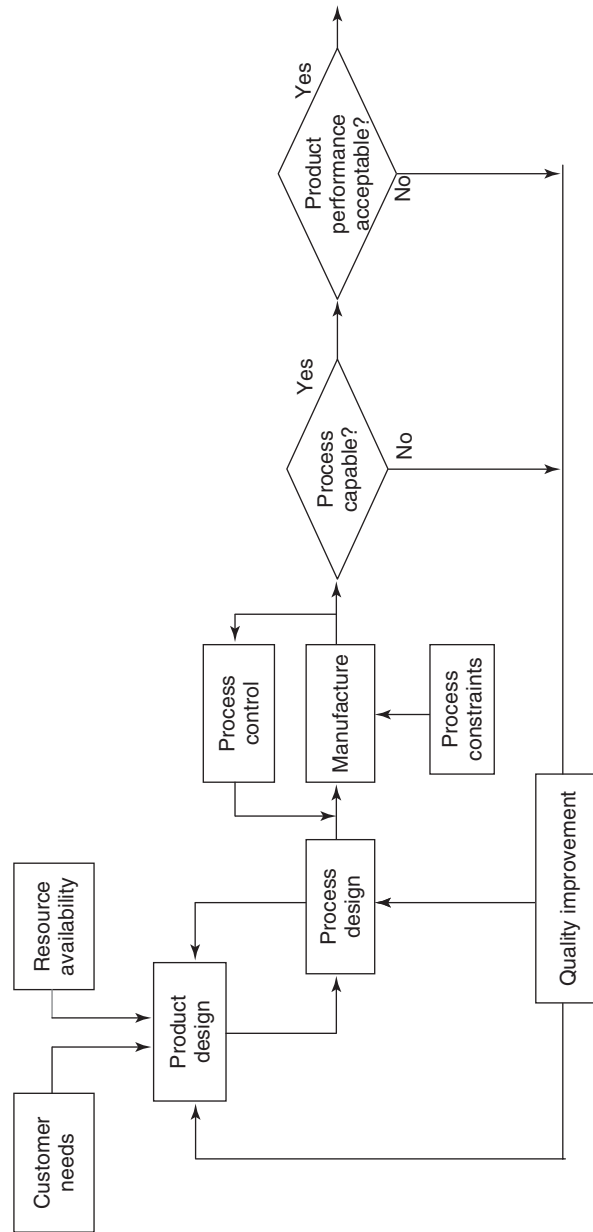


Figure 1. Linking quality design, control, and improvement functions.

to further improve product and/or process design. The eventual test of quality is based on how well the product functions when used by the customer. Issues such as failure within the warranty period merit consideration. Therefore, a feedback loop also exists for development of QI strategies based on field performance of the product.

QUALITY OF DESIGN

Designed quality in a product may be accomplished in a variety of ways. Resource availability, feasible manufacturing processes, and vendor quality are the factors that influence product quality. We first discuss the process of incorporating customer needs into the design of the product using the technique of quality function deployment (QFD). Thereupon, some concepts are illustrated on improving the system reliability of a product as most products are basically comprised of subsystems/components.

Quality Function Deployment

The central focus is on designing products that meet customer needs, and the product development cycle accomplishes this through a series of steps known as *QFD* [1]. QFD is essentially a planning tool that starts with product design to meet customer needs, continues to analysis of parts and components to create the desired product, progresses to process design, and eventually ends with production requirements. In each of these phases, a procedure utilizing the *house of quality* (HOQ) [2,3] is utilized. We briefly discuss this procedure in the context of product design. A similar approach could be used in the other phases as well.

In product design, a core concept is to prioritize customer requirements, usually through some ratings scale, such as a Likert scale with ordinal ratings of 1–5, with “5” representing “most important.” Thereafter, a relative assessment is performed, with respect to chosen competitors, on customers’ perceptions on each of the selected requirements. Such ratings are also on an ordinal scale, say 1–5. Review of the ratings indicates

the benchmark organization, for each of the customer requirements, and the relative standing of the company with respect to its competitors in satisfying each customer requirement. An outcome of such analyses could be to identify a desirable approach to pursue, to meet each of the requirements.

The next step in the procedure is the identification of technical product descriptors that will address the customer requirements. Attention should be given to satisfy critical customer needs first, as depicted by the priority ratings, and arriving at a small and cost-effective number of product features. It is possible for a given product feature, known as a *technical descriptor*, to satisfy more than one customer need. Following this is the assignment of rating factors of each technical descriptor to each customer need, using an ordinal scale of, say, 1–5, sometimes known as the *relationship matrix*. A zero may be assigned if the technical descriptor has no impact on meeting the particular customer requirement.

Using the relationship matrix, for each technical descriptor, the product of the customer importance rating of the requirement to the assigned rating score for the technical descriptor to meet the requirement, is found, and added up to determine the effectiveness of the particular product feature in meeting all customer requirements. A comparison of the sum of the weighted rating scores for each product technical descriptor provides us with an indication of their relative effectiveness in satisfying customer needs and also with management decisions to select the appropriate product descriptors, if all of them are not cost-feasible.

Additionally, a technical competitive assessment of the product descriptors may also be conducted. In this step, using an ordinal rating scale, say, of 1–5, the relative assessment with respect to the competitors may be performed for each descriptor. Rather than using customer information, this step addresses technical competence and the corresponding benchmark (or leader), for each product descriptor. It demonstrates the relative strengths and weaknesses of the company and areas, in terms of product features, where improvement efforts should

be focused. When technical information on competitors is publicly available, information from this step may be used along with the sum of the weighted rating scores for each technical descriptor, to make decisions regarding choice of product features.

System Reliability

Product designers are obviously interested in creating a design that not only meets but also exceeds customer expectations. With this goal in mind, a couple of constraints limit the designer. Resource availability and the availability of existing manufacturing processes are two major constraints. One basically has to address the question: Can we make what we designed within our allocated budget and available technology? The answer to this question is what creates a feedback loop between product design and process design. If we cannot make what we designed, the design may have to be altered considering the technology that is available, which is precisely the concept of *design for manufacturability* [4].

Most products consist of components or subassemblies that are assembled in a desired configuration for proper functioning of the product. To satisfy the customer, it is the reliability of the end product that is of interest. The reliability of a product or component represents the probability of it functioning effectively for a certain period of time under certain environmental conditions. The complexity of the configuration of components that create the end product will impact the reliability of the end product or the system of components. Let us examine the various factors that affect system reliability.

First, raw material quality, resource availability, and available technology will influence the component reliability. Manufacturing capability, as discussed earlier, will impact component reliability. As variations exist in all manufacturing processes, component reliability for a selected time period is always less than one. Second, component configuration greatly impacts reliability. For example, if components are in series, implying that all components must function for the system to function, the system or end product reliability is the product of the individual

component reliabilities, assuming that the component failures are independent of each other. Hence, if there are 20 components in series, each with a reliability of 0.990, the system reliability is only 0.818.

Several options exist for improving system reliability through design. One is the concept of adding components in parallel, creating redundancy in the system [5]. In this situation, the system or end product functions so long as at least one of the components, which are in parallel, functions. Suppose that there are p components in parallel, including the original component. Denoting the reliability of each component by R_i and assuming that the components operate independently of each other, the system reliability is given by

$$R_s = 1 - \prod_{i=1}^p (1 - R_i). \quad (1)$$

Consider an example where there are five components in parallel, each with a reliability of 0.8. The system reliability is greatly improved to 0.99968. As resource availability and manufacturing capability are usually bound on the reliability of a component, the concept of using a redundant design greatly improves system reliability. Of course, the product unit cost increases because of the added components.

A second concept, similar to the redundant design, is that of incorporating standby components in parallel. In this design, it is assumed that only one component in the parallel configuration is operating at any given time. In a parallel system, all components are assumed to operate simultaneously. Hence, system reliability for standby systems is higher than that of comparable systems with components in parallel. Suppose that the time to failure of components is an exponential distribution with failure rate λ , which amounts to the number of failures in time t following a Poisson distribution with parameter λt . Therefore, the probability of x failures in time t is given by Mitra [5]:

$$P(x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}. \quad (2)$$

Now, if we have a standby system with the basic component and n standby components, the system reliability is given by

$$R_s = e^{-\lambda t} \left[1 + \lambda t + \frac{(\lambda t)^2}{2!} + \cdots + \frac{(\lambda t)^n}{n!} \right]. \quad (3)$$

QUALITY CONTROL

There are two types of cause systems that influence manufacturing and service systems—*common causes* and *special causes*. Common causes are those that are inherent to the system. While they cannot be completely eliminated, their impact may be reduced through the design of better processes, equipment, procedures, raw material, and components. For example, variation of a certain characteristic, say part length, will always exist. The degree of variation may be reduced through *QI* efforts. *QI* thus deals with making systemic changes to improve products and processes.

Special causes, sometimes called *assignable causes*, are not inherent to the system. They occur due to some changes in the process that, hopefully, can be identified so as to take appropriate remedial action to eliminate them. In the manufacture of the part length, a new operator could be using a feed rate different from the specified norm, which may increase the variability and/or change the average value of the surface finish. The detection of such special causes and subsequently the identification of remedial actions may be facilitated through the process of *quality control*. In particular, *statistical process control* (SPC) deals with the monitoring of processes through statistical methods known as *control charts*. A process is said to be in statistical control when only common causes prevail in the system.

Control charts may be broadly classified into two categories—those for variables and those for attributes. Quality characteristics for variables are those that are measurable on a numerical scale. The length, thickness, surface finish, and tensile strength of a part are examples. For variables, the type of information collected from the process typically

dictates the type of control chart. Further, the formation of rational subgroups to collect data from the process is an important consideration. The idea is to choose subgroups such that the variation within subgroups is minimized, while the variation between subgroups should be used as a guideline to maximize the chance of detection of an out-of-control process.

Control charts typically have three reference lines: a center line that is usually an estimate of the mean of the quality characteristic and two bounds indicated by a lower control limit (LCL) and an upper control limit (UCL). These bounds are selected based on a chosen level of Type I error or false alarm in making decisions from the control chart. For quality characteristics that are variables, if each subgroup contains more than one observation, the guidelines are to construct a control chart for the mean ($\bar{X}$) and range (R), when the subgroup size is small (≤ 10). For large subgroup sizes, control charts for the mean ($\bar{X}$) and standard deviation (s) may be monitored. When each subgroup has only one observation, control charts for individuals (I) and moving range (MR) are used.

The second category of control charts is for attributes. Here, one class of control charts deals with the situation where the product or item is classified as conforming or nonconforming. The proportion nonconforming (p) chart is used when the subgroup size may remain constant or vary from one subgroup to another. The number nonconforming (np) chart is used when the subgroup size remains constant. A second class of control charts deals with the monitoring of nonconformities or defects. In this context, if the subgroup size remains constant, the c -chart for monitoring the number of nonconformities is used. On the other hand, if the subgroup size varies, the appropriate control chart is the u -chart that monitors the number of nonconformities per unit. An important consideration in attribute charts is the choice of an adequate subgroup size. The subgroup size must be large enough for nonconforming items or nonconformities to occur. A rule of thumb is that the expected number of such nonconforming items or nonconformities should be at least five. Details of SPC or statistical

quality control (SQC) are found in Statistical Process Control Section 4.1.5.3 of the encyclopedia.

A major objective of the control charting scheme is to determine the presence of special causes and take remedial action, as soon as possible, to bring the process back to a state of control. A process in control, therefore, is impacted by only common causes. Does this mean that if we have a process in control the customer satisfaction level is high? Not necessarily. There is a conceptual difference between a *process in control* and a *process that is capable*. *Process capability* addresses the degree to which the product, with the corresponding attributes, meets the requirements of the customer. A process can be in control but not be capable. Measures of process capability are found in Statistical Process Control Section 4.1.5.3 of the encyclopedia. Process control could be conducted on-line or off-line. In order to minimize the impact on outgoing product quality, for a process that is out-of-statistical-control, remedial action should be taken as soon as possible. For this reason, SPC through control charts, where feasible, could be implemented on an on-line basis.

QUALITY IMPROVEMENT

While quality control involves the identification of special causes and implementation of remedial actions to bring the process to a state of control, QI addresses the issue of the product or service meeting the needs of the customer. A process could be in statistical control yet the product may not satisfy customer requirements. This occurs when the inherent variation in the process, as defined by its natural tolerance limits, exceeds the variation as allowed by the specification limits. In such cases, the product and/or process may have to go through the design phase again so that the redesigned product manufactured by a redesigned process will meet customer specifications. QI is incorporated through such a feedback loop (as previously demonstrated in Fig. 1). Moreover, actual field results from the performance of the product when used by the customer may

also demonstrate a need for QI. For instance, field results may demonstrate a higher proportion of product/component failures within the warranty period. On identification of the causes of product failure, a product may have to be redesigned to avoid the repetition of early failures.

In the design phase of products and associated processes, one may identify the settings of process parameters that will lead to the accomplishment of desirable levels of the product parameters. Such analyses may be conducted off-line, much before production runs are implemented. It may be conducted through prototype runs of the process where it is feasible to change process parameter settings. Product output results could be obtained for such trials and analyzed to identify process settings that accomplish the product characteristic to be close to the target value and possess variability as small as possible. The topic of *design of experiments* (DOE) using statistical analysis is a viable approach for this off-line approach to QI.

Quality and Competitive Position

In the twenty-first century, importance of continual QI in order to stay competitive is paramount. Here, we discuss some specific ways in which improved quality affects the competitive position of the organization and improves its relative standing. In particular, we consider price/cost ratio, capacity, and market share.

With an improvement in the level of quality, there are fewer rework and scrap items and fewer returns by the customer. The unit cost of producing a conforming product decreases. Even if the unit selling price of the product does not change, the unit margin on the price/cost ratio improves, which leads to increased profitability of the organization.

An improvement in the quality level implicitly creates additional capacity. As there is a reduction in the amount of rework and scrap items, additional capacity (resources) will not be necessary to make up for items lost due to rework and scrap, assuming the demand to remain the same. Hence, the company will be able to produce more products to meet customer demand,

without necessarily going into overtime or subcontracting. In a sense, the company has been able to “increase capacity.” This leads to a more efficient utilization of the available resources.

A direct effect of the impact of an improvement in quality is increased customer satisfaction. As always, increased customer satisfaction may improve the firm’s ability to not only retain existing customers but also to draw customers from competitors. A positive impact is therefore a growth in market share. Another indirect impact is the ability to manage the unit price/cost ratio. With improved quality, the customer may be willing to pay a higher price. Thus, the profit margin is impacted in three ways. A higher unit price could be charged, a lower unit cost is achieved due to QI, and the market share is increased due to increased customer satisfaction. The contribution to total profit, in a simplistic sense, may be expressed as:

$$\begin{aligned} \text{Total profit} &= (\text{Unit price} - \text{Unit cost}) \\ &\quad \times \text{Market share,} \end{aligned} \quad (4)$$

where all three components on the right-hand side of the equation may be impacted through QI. Manufacturing a product usually requires many operations in sequence. The proportion nonconforming of the finished product is, thus, a function of the quality level of each operation and the number of operations. Let q_i be the quality level (proportion conforming) in operation i . Then, the proportion nonconforming at the finished product level, when there are n operations, is

$$p = 1 - (q_i)^n. \quad (5)$$

Hence, for a process with 30 operations, even if each operation is 99% conforming, the finished product is about 26% nonconforming, assuming that the operations are independent of each other. Sometimes, this assumption may not hold. In other words, a product that is nonconforming at stage j , may impact the rate of nonconformance in stage $(j + 1)$ adversely. This leads to an important consideration. What should the quality level be at a given operational stage?

Six-Sigma Quality

The notion of six-sigma quality [6] attempts to address this issue. Six-sigma may be viewed as a philosophy, methodology for continuous improvement, or as a metric of performance.

Let us consider its interpretation as a metric. Assuming that the distribution of the product characteristic is normal and that the product specification limits are three standard deviations from the mean, the proportion of nonconforming product is about 0.27% or 2700 (parts per million). As explained previously, the compounding effect of the number of operations comes into play. Thus, for a product with a 1000 parts or one that has 1000 operations, an average of 2.7 defects per product unit is expected. The rolled throughput yield, a measure of proportion conforming, is quite small.

If the specification limits are each six standard deviations away from the mean, the proportion nonconforming is only about 0.001 ppm on each tail, using the normal distribution. This quality level seems quite good compared to the 2700 ppm for a three-sigma capability level. In real-world situations, a process may drift (i.e., a shift in the location to the right or the left). Allowing for a shift in the process mean of 1.5 standard deviations from the center, the nearest specification limit is 4.5 standard deviations from the process mean. Using normal distribution theory, the proportion nonconforming is about 3.4 ppm, accounting for this drift. Hence, when six-sigma is viewed as a metric, with a nonconformance rate of 3.4 ppm, it should be noted that the nearest specification limit, for this situation, is 4.5 standard deviations away from the mean and not six standard deviations.

As a philosophy, six-sigma may be considered as a strategic initiative for organizations. It prescribes a road map to pursue the goal of continuous QI. The pursuit of quality is a never-ending journey. As with other approaches, the customer will be the key factor, with analysis of products and processes to improve customer satisfaction as the goal.

When six-sigma is considered as a methodology, it defines a sequence of logical steps to adopt to accomplish QI. In this context,

the *define, measure, analyze, improve, and control* (DMAIC) approach [7] becomes relevant. In the define phase, customer needs are translated to specific attributes that are critical to quality, cost, or delivery. Prioritization of needs as dictated by the customer will create a relevant focus. The measure phase creates metrics, which are measurable, for tracking process performance. Existing baseline levels may be used to determine if significant improvements occur based on the recommended action items. The analysis phase involves statistical analyses of collected data from the process to gauge the level of improvement. Sometimes, these may be simple graphical tools such as scatterplots, histograms, or multivariable charts. Subsequent analysis could be on hypothesis testing of process parameters, regression analysis of a selected response variable, or analysis of variance to investigate the significance of one or more factors on a chosen response variable. In the improve phase, identification of significant factors and their optimum levels is of interest. Finally, in the control phase, methods are identified to sustain the gains accomplished in the preceding phase. Statistical control charts, as found in Statistical Process Control Section 4.1.5.3, could be used in this context.

Quality Improvement Tools

Some simple graphical tools aid management in the QI process. One is the use of *Pareto diagrams* (based on the principle that 80% of the problems are created by 20% of the causes), which prioritize problems by arranging them in decreasing order of importance. A valuable tool in investigating process efficiency is the *flow chart*, which represents the sequence of events in a process. Analysis of a flow chart may lead to process simplification, identification of bottlenecks, and elimination of nonvalue added activities. A refined version of the flow chart is the *process map*, which may reflect the inputs and outputs for each event, the critical, controllable, and noise parameters for that operation, and the person who has ownership of that process.

An effective tool in QI is the *cause-and-effect* diagram, often known as a *fishbone diagram*. In this diagram, the problem of

interest, or the effect, is brainstormed to identify major causes. Major cause categories, for example, could be material, methods, people, machines, measuring equipment, and environmental factors. For each major cause, depicted graphically such as a fishbone, subcauses are listed. A QI team considers such causes and subcauses to identify ones that should be investigated.

Bivariate data in the form of a *scatterplot* depicts relationships between two variables. These may be used as a follow-up to a cause-and-effect analysis to obtain the impact of a controllable factor on an output characteristic. As realistic products and processes often involve more than two variables of interest, which includes dependent and independent variables, *multivariable* charts could be insightful. A *matrix plot*, for example, could depict how each independent variable impacts a given dependent variable, based on which a decision could be made on the independent variables to control.

Taguchi Loss Function

A traditional definition of quality is conformance to specifications. As long as the quality characteristic is within specifications, no loss is incurred. Therefore, for a characteristic with two-sided specifications, an upper specification limit (USL) and a lower specification limit (LSL), the loss is zero for the region between LSL and USL. For regions outside the specification limits, a constant loss (L_0) is encountered (Fig. 2).

This notion of a step-function loss is difficult to justify because the loss associated with a product whose characteristic is just above the LSL is zero, while for that which is just below the LSL it is L_0 . Moreover, under this definition of loss function, there is no motivation to improve quality as long as the characteristic is between LSL and USL.

The loss function advocated by Taguchi is more realistic [8,10]. A loss is incurred anytime if there is a deviation of the quality characteristic from the target value. This loss function is indicated by L_1 in Fig. 2. Losses may arise owing to customer complaints, warranty costs, lost opportunities, and other tangible and intangible costs. The

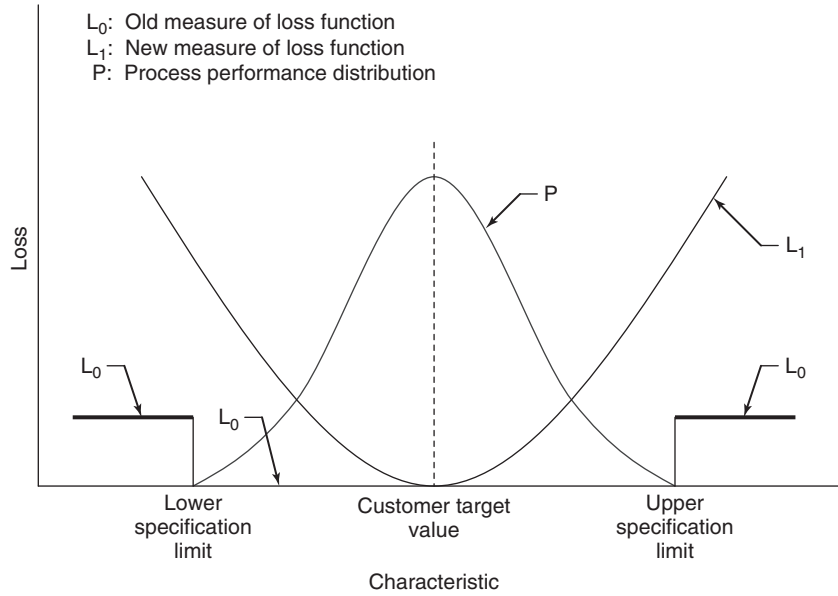


Figure 2. Comparison of old and new measures of the loss function. [Source: Reprinted with permission from Ref. 5. Copyright 2008 Wiley & Sons].

equation of such a loss function is given by

$$L_1(y) = k(y - T)^2, \quad (6)$$

where y represents the value of the characteristic, T is the target value, and k is a proportionality constant. Knowing the cost (or loss) to replace a product whose dimension is just below the LSL to be A , we obtain

$$k = A/\Delta^2 \quad (7)$$

where $\Delta = T - \text{LSL} = \text{USL} - T$ (assuming the target is centered within the specifications).

A feature of this loss function is that it promotes the philosophy of continuous improvement to reduce loss as the loss function increases with deviation of the process mean from the target value and with process variability.

Experimental Design

In classical experimental design [9] the primary objective is to create a model of

the response variable, that is, the quality characteristic in which we are interested, as a function of several factors, which are usually process variables. The predictors could be process variables and/or some functions of them as well as possible interactions between them. The objective is to develop a model for the mean response and thereby determine the significant predictors and their levels to accomplish a desired value for the mean response.

On the other hand, the Taguchi method of experimental design has an objective of not only achieving a desired mean response but also reducing the variability in the quality characteristic. Furthermore, this approach incorporates not only the process parameters but also the uncontrollable or noise factors.

As all processes have variation, the focus of the Taguchi concept is on robust designs, which are less sensitive to the uncontrollable or noise factors. While noise factors cannot be completely eliminated, Taguchi's approach to experimental design is to determine the settings of the noise factors and the controllable factors such that variability is minimized and

the quality characteristic is close to the target value. The performance measure used in Taguchi's experimental design is usually in the form of a signal/noise (S/N) ratio, that is maximized, and appropriate factor level settings are determined [10]. Thus, for the squared loss function, the signal is proportional to the squared difference between the process mean and the target (for a two-sided specification), while the noise is proportional to the variability of the characteristic.

Even though noise factors are uncontrollable, for the experimental design that investigates the settings of the controllable and uncontrollable factors and their impact on the response variable, Taguchi advocates selections based on joint considerations of these variables. The matrix settings of the controllable or design parameters are chosen in an inner array. The noise factor settings in the designed matrix are in an outer array, and these serve as replications for each setting of the design factors.

There are four basic phases in designing an experiment. In the planning phase, the factors or variables to be considered in the study along with the response variable, the levels of the factors, and the number of replications deserve consideration. Ideally, the runs should be selected in a randomized order to average out the effect of all other factors not considered in the analysis. Practical limitations on the process, for example, changing the temperature setting in an oven from a low to a high setting (skipping the medium setting) may cause restrictions on randomization. While more replications, for a given factor combination setting, are desirable, cost considerations may influence this choice.

One major advantage of a designed experiment is the ability to estimate *interaction* effects between factors. This represents the joint impact of factors on the response variable. For an experiment with two factors, for example, Factor A and Factor B, each at say two-level (low and high), interaction represents the degree of change in the response variable, as Factor A changes from a low to a high level, for a chosen level of Factor B. If the degree of change is different, for different levels of Factor B, an interaction is said to exist

between Factors A and B. In addition, if the interaction effect is significant, decisions on the optimum factor levels are based on such analyses rather than from the main effects.

The second phase in a designed experiment is that of process characterization or screening. This involves identification of significant factors from a multitude. Here, the choice of factor levels is kept minimal, say two or three, with usually the extreme values of the range of the factor level being selected. Taguchi designs could be used because they utilize a few factor levels and the total number of runs is quite small relative to the total number of factors chosen. For example, a 15-factor experiment with each factor at two levels may require only 16 runs using an orthogonal array. Experimental designs used in the Taguchi approach make use of orthogonal arrays. Such arrays have the property that every factor setting occurs the same number of times for every test setting of all other factors. In the design matrix setting, each row represents the levels or settings of the chosen factors (a run), while each column represents a specific factor. Any two columns of an orthogonal array form a two-factor full factorial design. Use of orthogonal arrays minimizes the number of runs while retaining the pairwise balancing property. An example of an orthogonal array for three factors, each having two levels (1 and 2), is shown in Table 1 with four experimental runs.

The third phase is that of optimization that involves determination of the optimum levels of significant factors to achieve desired characteristics associated with the response variable. As discussed previously, the S/N ratio is a performance measure considered in

Table 1. Orthogonal Array for a Three-Factor Experiment

Experiment	Factor 1	Factor 2	Factor 3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

Source: Reprinted with permission from Ref. 10.
Copyright 1987 American Supplier Institute.

the Taguchi approach. Lastly, the verification stage follows where the selected optimal settings of the factor levels are chosen to observe values of the response variable or performance measure to confirm actual process performance under these settings.

REFERENCES

1. Cohen L. Quality function deployment: how to make QFD work for you. Reading (MA): Addison-Wesley; 1995.
2. Akao Y. QFD: past, present, and future. Proceedings of the 3rd Annual International QFD Conference. Linkoping; 1997.
3. Mizuno S, Akao Y. In: translated by Mazur GH. QFD: the customer-driven approach to quality planning and deployment. Tokyo: Asian Productivity Organization; 1994.
4. Anderson DM. Design for manufacturability & concurrent engineering. Cambria (CA): CIM Press; 2010.
5. Mitra A. Fundamentals of quality control and improvement. 3rd ed. Hoboken (NJ): John Wiley & Sons; 2008.
6. Harry MJ. The vision of 6-sigma: a roadmap for breakthrough. Phoenix (AZ): Sigma Publishing Company; 1994.
7. Harry MJ, Schroeder R. Six sigma: the breakthrough management strategy revolutionizing the World's top corporations. New York: Doubleday; 2000.
8. Phadke MS. Quality engineering using robust design. Englewood Cliffs (NJ): Prentice Hall; 1989.
9. Box GEP, Hunter JS, Hunter WG. Statistics for experimenters. 2nd ed. Hoboken (NJ): Wiley-Interscience; 2005.
10. Taguchi G. System of experimental design. Volume 2. Bingham Farms (MI): American Supplier Institute; 1987.

QUALITY MANAGEMENT

AMITAVA MITRA
College of Business, Auburn
University, Auburn, AL

INTRODUCTION

What exactly is quality? A traditional definition, given by Crosby [1], relates to conformance to requirements or specifications. Such a definition focuses on some pre-designated standards for the product. Juran [2] proposed a more general definition incorporating fitness for use. Garvin [3], on the other hand, considered the definition of quality in five categories: product-based, user-based, manufacturing-based, value-based, and transcendental. At the core of product/service quality is the satisfaction of the customer. Hence, we advocate a definition of quality that incorporates meeting or exceeding the needs of the customer [4].

Management of the quality function is critical at all levels of the organization. Senior management sets the vision, mission, and associated strategic plans to accomplish the mission. They must also address aggregate costs, market share, and unit profitability so as to provide a reasonable rate of return to investors/shareholders. In this context, the role of quality in reducing unit total costs is discussed. Middle management usually deals with tactical issues. Whether it is determination of unit price, product mix, or product servicing after it is sold to the customer, product quality impacts such decisions. Line management deals with operational issues. Specific processes and their parameters, as well as desirable technical traits of employees involved in making the product, are issues related to quality considered at this operational level. Vendor quality of materials, components, and sub-assemblies also influence product quality.

Responsibility for quality, therefore, lies with everyone involved in the various phases

of product/process design, manufacturing, process control, product servicing on sale of the product, and meeting of customer needs. Identification of the customer is a first step that is assisted by knowledge of the market and competition. It should be noted that as both market and competition are dynamic in nature, customer identification is an iterative process that incorporates the changing nature of these entities. Consider a manufacturer of high end automobiles. It is incumbent on the manufacturer to identify the target customer group for such automobiles and possible factors that describe this group. These could be a base income level, an age classification level, or status of home ownership. Market research plays a key role in the identification of product features and characteristics that the customer prefers. The product design team faces the challenge of creating a feasible product that meets the requirements of the customer. Thereupon, process design ensures the feasibility of making the designed product using the available processes and meeting the limitations imposed by budget constraints. Several feedback cycles between product design and process design usually take place before finalization of a product design that meets the needs of the customer and can also be manufactured in a feasible manner. A final test of product quality is its performance during usage by the customer. A robust product design encompasses a variety of environmental conditions with the objective of achieving a product performance that does not change drastically with a change in the operating conditions.

Quality management is a broad area. It is not feasible to deal with the general foundations and philosophies as well as operational details of implementation of the tools and techniques in one article. Therefore, an approach that shows the links between various components of a quality management system has been chosen. We provide a roadmap that starts with customer identification followed by prioritization of the needs of the

customer to create an appropriate product or provide a desired service. As meeting or exceeding customer needs is fundamental, the Kano model that incorporates the prioritized needs of the customer is presented. Basic quality management tenets associated with Juran, Crosby, and Deming are discussed. Further, the role of total quality management, a philosophy that holds the three components of the customer, process, and people together in a cohesive manner to create an organization that meets the desired quality objectives, is also presented.

At the strategic level, the impact of the various phases of adoption of a quality system on the organizational objectives is discussed. The components of quality costs are presented next, where the importance of quality in reducing total costs is demonstrated. We include this section as all levels of management understand the associated costs at their level and, further, they need to grasp the impact of quality improvement on total costs. With the twenty-first century global economy, a brief exposure to the processes of certification in quality management using global standards is important to multinational organizations or vendors that conduct business with organizations abroad. Hence, the role of the International Organization for Standardization (ISO) is briefly presented, followed by criteria for a national quality award in the United States, the Malcolm Baldrige Award.

HIERARCHICAL NEEDS OF THE CUSTOMER AND THEIR INFLUENCE ON QUALITY

A psychological theory of human needs proposed by Maslow [5] depicts a pyramidal structure of the various categories of needs. At the base lie physiological needs. Moving up the pyramid, we find the needs associated with safety, belonging, esteem, and self-actualization, respectively. Needs at the lower level have to be satisfied before those at the upper level are addressed. Similar to this concept, in the realm of quality, customer satisfaction is also influenced by certain types of needs [6]. The Kano model hypothesizes three prioritized categories of customer needs: basic needs, performance

needs, and excitement needs. Basic needs are those that are taken for granted by the customer—the customer expects such attributes to be present in a product. If they are missing, it will cause customer dissatisfaction and possible loss of market share. On the contrary, their presence may not necessarily increase market share. For example, a bearing must be able to withstand a certain minimal load. Meeting this requirement does not necessarily increase the satisfaction level of the customer; however, not meeting it will certainly cause dissatisfaction and possible loss of a customer.

The next category of performance needs consists of those that do influence the level of customer satisfaction, which normally changes linearly as a function of the level of the performance need. For instance, it might be desirable for the bearing to operate within a temperature range. Hence, if the bearing performs well at very high temperatures, customer satisfaction may increase linearly with the level of temperature under which it is able to adequately operate. Designing and manufacturing bearings that can function well at temperature extremes may well please the customer and improve market share. One strategic approach to retain and attract new customers could be through the satisfaction of performance needs, relative to that of the competitors' level of offering.

The third and most influential category is that of excitement needs. These are known as *delighters*. The customer sometimes may not even be aware of these before the purchase of the product. This category addresses those needs, if satisfied, will thrill the customer. For instance, the product has certain features that exceed the expectations of the customer. Customer satisfaction increases exponentially with the level of excitement needs. Identification of such needs and their incorporation into the design and manufacture of products is certainly an avenue to increase market share and leapfrog over competitors. For example, a certain ceramic material may increase the life of the bearing 10-fold over those produced by competitors. Further, it could be lighter in weight and may require minimal maintenance. The ability to manufacture such a bearing at a competitive cost

will have a tremendous impact on market share.

QUALITY MANAGEMENT PHILOSOPHIES

The principles of quality management apply to manufacturing and service industries. In a manufacturing industry, it is imperative for the product to perform its intended function in a way that meets or exceeds customer expectations. This is accomplished by the product having desirable features and operating characteristics that surpass the needs of the customer. Similarly, there are some quality characteristics that are unique to the service industry, distinct from those in manufacturing that must be met in order to achieve high customer satisfaction.

The delivery of service involves an interaction between the provider and the recipient. Thus, behavioral characteristics and human factors come into play. The courtesy and caring attitude of a nurse impacts patient satisfaction. Timeliness in providing the service is often of importance to the customer. Waiting time in a bank before a transaction or the waiting time to receive bags at an airport has an impact on customer satisfaction. A third type of quality characteristic involves service nonconformity. For example, the number of billing errors per 1000 accounts by a utility company is a measure of service quality. Finally, a fourth category consists of facility-related characteristics. Lack of a swimming pool in a hotel or insufficient legroom in an aircraft is an example.

All of the quality management philosophies strive for continuous quality improvement—a never-ending process. Top management commitment is also a common thread in these philosophies. Without adoption by senior management and support through resource allocation, any of the proposed approaches will fail after a short time span. Finally, quality must be a measurable entity. What cannot be measured is hard to monitor.

Philip B. Crosby's [1] approach utilizes a quality management grid that consists of five stages of maturity of the firm and six

measurement categories that aid in the evaluation of the degree of maturity in the utilization of the quality management philosophy. The stages of maturity are uncertainty, awakening, enlightenment, wisdom, and certainty. They represent the degree to which the principles of quality management are accepted in the organization. For example, in the uncertainty stage, there is no understanding of why problems occur, nor is there a comprehensive approach to prevent problems from happening. On the contrary, in the certainty stage, quality management is an integral part of the organizational culture and the focus is on prevention of problems. Similarly, the measurement categories pertain to evaluation of a firm's maturity in quality management and may range from "management understanding and attitude" to "quality organization status" to "problem handling" to "cost of quality as a percentage of sales" to "quality improvement actions" to "summary of company quality posture." For an organization that is in a stage of "uncertainty," an understanding of the use of quality as a management tool does not exist in the measurement category of "management understanding and attitude." Crosby also proposes a 14-step plan for quality improvement.

Joseph M. Juran's [2] approach consists of a quality trilogy involving quality planning, quality control, and quality improvement. At the upper management level, quality planning consists of establishing quality goals, whereas at the middle management level, it may involve setting departmental goals that are consistent with the strategic goals set by top management. Finally, at the operational level, planning may include workforce assignment and individual goal setting. The quality control phase deals with process control to determine sporadic causes of nonconformities. Remedial actions are identified in this phase. In the quality improvement phase, steps are taken to accomplish continuous improvement of the product and the process. Such actions may deal with chronic nonconformities and deficiencies that are part of the existing product or process.

W. Edwards Deming's [7] philosophy stresses the importance of the role of management. His belief was that over 85%

of the problems are due to *common causes* of variation, those that are inherent to the system and, therefore, can be addressed by management only. Alternatively, only 15% or less of the problems are due to *special causes*, those that are external to the system and can be dealt with by engineering and operating personnel.

Deming's approach is unique in several ways. It is one of the few that emphasizes a data-driven action plan. While experience and intuition are good attributes in decision-making, Deming stressed the importance of a *plan-do-check-act* (PDCA) cycle. Hence, data collected from the process based on a proposed improvement plan should be analyzed to validate its suggested effectiveness. In the event that results from data analysis do not confirm predicted expectations, one may have to revert to the planning stage in the PDCA cycle. Deming was one of the first pioneers in quality management to focus on a statistical approach. His methodology relies heavily on the use of control charts using a statistical process control (SPC) approach. A major objective of SPC is the identification of special causes in a timely manner. This enables the adoption of appropriate remedial actions to bring a process to a state of statistical control. Deming also believed in utilizing data to create knowledge. Data by itself is not sufficient to draw meaningful conclusions. On the other hand, when the available data is analyzed to create predictions based on selected hypotheses, it becomes a powerful tool. Deming was keenly aware of the importance of psychology as managing people necessitate an understanding of the behavior and interactions of people with others. He realized the importance of education and training for all. Further, he recognized the presence of barriers in the organization that prevent the flow of information. These barriers exist within departments, between departments, and between the organization and vendors and investors. Deming is credited with the concept of the "*extended process*," which recognizes that organizations need to consider vendors, customers, investors, employees, and the community in their decision-making process. Thus, in product design, a dynamic

link needs to exist between the organization and its vendors that supply raw material, parts, and components so as to come up with a feasible design. Deming's concept of *optimization of the system* utilizes the above idea with the objective to create a win-win situation for all.

The total quality management philosophy accomplishes the organizational goals through the integration of three entities—customer, process, and people (Fig. 1).

The achievement of quality in products/services is fundamental, as the degree of customer satisfaction is a function of the difference between quality expectations and that experienced. Customer expectations may be influenced by the external environment (or competitors) and are not solely impacted by the organization. However, the satisfaction level experienced by the customer is directly influenced by the level of quality of the products/services, which is under the control of the organization.

The second entity, critical to achievement of quality, is the process itself. In the manufacturing environment, this consists of, for example, the machinery and equipment, procedures and operations, line and staff people and their degree of expertise, suppliers and vendors, and the degree of product servicing. The philosophy of "continuous improvement" is the key to sustaining and maintaining a competitive edge, as customer needs and competition encountered are dynamic. In order to ensure an effective and efficient process, the concept of an "extended process," as advocated by Deming [7] is relevant. By including the customer in the extended process, it provides the company better opportunities to determine needs of the customer through consumer research and thereby develop appropriate strategies to satisfy and surpass customer expectations. Similarly, keeping the vendors as part of the team that recommends requisite product/process changes helps in creating a more effective process that produces a finished product with the desired characteristics.

The third entity in the total quality management model involves the people. Production of a quality product requires a motivated workforce. Empowering people so that

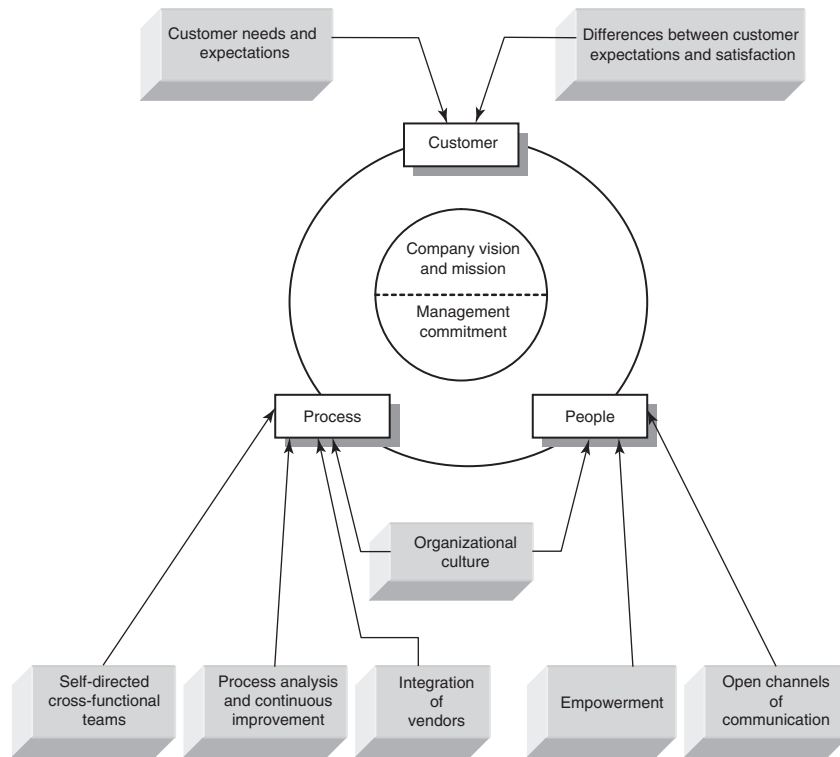


Figure 1. Features of a TQM model. [Source: Reprinted with permission from Ref. 4. Copyright 2008 Wiley & Sons.]

they may be involved in decision-making leads to creating a group that takes pride in its work skill. Through the promotion of open channels of communication, where management demonstrates its receptiveness to improvement ideas, a unified and encompassing approach to quality improvement is accomplished. All of this leads to the creation of a “quality culture” that permeates the organization.

IMPACT OF QUALITY ON ORGANIZATIONAL OBJECTIVES

The creation, delivery, and servicing of a quality product to the customer involves three phases, with feedback and interactions taking place among them. These are quality of design, quality of conformance, and quality of performance. Under quality of design, it must be ensured that the product meets the needs of the customer and is cost-effective

as well. Typically, the product is designed to at least exceed the customer needs using a safety factor. In the service sector, deciding on the means and procedures through which service will be provided is considered in the design phase. Quality of conformance addresses the issue of manufacturability or the providing of the service as planned. It must be feasible to produce what was designed at a cost-competitive rate. Alternatively, in the context of the service sector, it must be feasible to implement the procedures for rendering services, adopted in the design phase. Hence, feedback may take place between the phases of design and conformance. Quality of performance deals with the degree of satisfaction of the customer, when the product is used or when associated servicing on the product is conducted. In the service sector, it measures the degree to which the customer is satisfied on receiving the service. Thus, there may be feedback

from the performance to the design phase in the event that the product does not function accordingly, or if the service provided does not meet customer expectations.

QUALITY COSTS

Quality costs, in general, may be grouped into the four categories of prevention, appraisal, internal failure, and external failure [8]. In the category of prevention, are the costs associated with planning, implementing, and maintaining a quality system that includes product and process design. During the initial stage of implementation of a quality system, while prevention costs may increase, with the degree of quality improvement, the reductions in internal and external failure costs may more than off-set such increases leading to a reduction in total quality costs. Measuring and evaluating products, components, and raw materials to determine their degree of conformance to some specified standards falls in the category of appraisal costs. When products and components fail to meet established requirements, before being sold to the customer, the incurred costs are in the internal failure category. Costs of rework and scrap are examples. Lastly, when products do not function satisfactorily once sold to the customer, the costs incurred to handle complaints, warranty obligations, and product liability fall under external failure.

In general, as the level of quality increases, an increase may take place in the combined prevention and appraisal costs (Fig. 2). Associated with this increase in the level of quality, an exponential decline is usually encountered in the combined internal and external failure costs. Using a dynamic concept of the impact of quality improvement, beyond a certain level of quality, the shape of the prevention and appraisal function may change from a concave to a convex function. Hence, with further improvement in the quality level, and the continuing exponential decline in the internal and external cost functions, the total cost function may diminish. Thus, with the objective of minimizing the total cost function, the target becomes the achievement of a 100% level of conformance.

INTERNATIONAL STANDARDS ISO 9000

With manufacturing and service organizations becoming global in nature, a need exists for ensuring standardization in quality management systems across countries. The *International Organization for Standardization (ISO) 9000* [9] developed a system of quality management standards. These standards were first published in 1987 and were based on *BS 5750*, a series of standards developed by the British Standards Institute (BSI). The 1987 version had three models: ISO 9001: Model for quality assurance in design, development, production, installation, and servicing, which applied to organizations creating new products; ISO 9002: Model for quality assurance in production, installation, and servicing, which was similar to ISO 9001 but without the creation of new products; and ISO 9003: Model for quality assurance in final inspection and test, which dealt with inspection of the finished product only.

A revision of the ISO 9000 standards occurred in 1994. They emphasized quality assurance via prevention as opposed to just final inspection. The next revision took place in 2000 and integrated some key concepts in the standards. First, process management was at the heart of the standards, which included involvement of top management and system optimization, two concepts previously discussed. Second, the importance of process performance metrics was delineated, as continual process improvement could be monitored through such metrics. Third, customer satisfaction as the key objective was made explicit. The standards were revised again in 2008. This version only introduced clarifications to the 2000 version and added no new requirements.

Studies have investigated the impact of ISO 9001 certification on the business performance, financial performance, and process effectiveness of organizations [10–13]. While it is difficult to identify exact association between certification and profitability, there are some general advantages that may occur because of the process of certification. An effective and efficient process may be created through the satisfaction of the various requirements specified in the ISO standards.

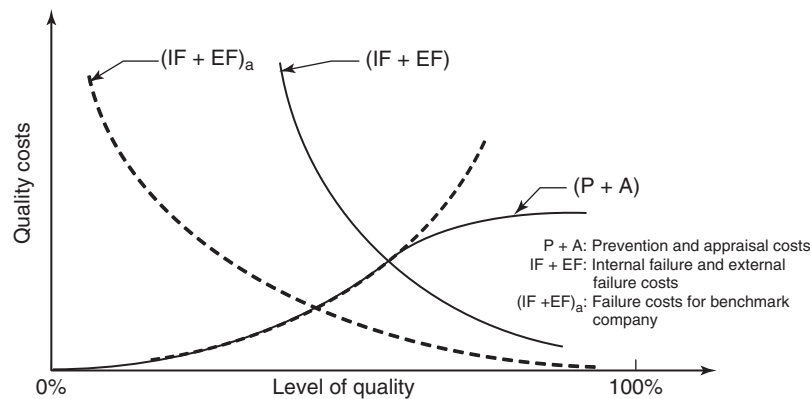


Figure 2. Target level of 100% conformance. [Source: Reprinted with permission from Ref. 4. Copyright 2008 Wiley & Sons.]

This, in turn, may lead to increased customer satisfaction, improved market share, and increased profitability. Certification of vendors may lead to improved quality of material and components that are procured. In general, meeting the ISO standards may assist in making the organization competitive. Certification takes place through third party audits and has to be renewed according to a defined interval, currently once every three years.

There have been some criticisms of the ISO 9000 standards, one being the exorbitant amount of time and paperwork required in preparation for registration. Second, if the quality culture is not engrained in the organization, quality manuals and documentation created in the registration process may merely serve to meet the requirements of the standard without actually being used to improve quality in the operations. Third, the registration process may promote specification control without actually being used for process understanding and improvement.

ISO 9001 is a generic standard applicable to manufacturing and service industries. There are various other industry standards, as well as other ISO standards, that have been developed for specific industries. For example, in the United States, the *QS 9000* standards have been adopted by the automotive industry. An international automotive quality standard accepted in the United States and European automakers

is *ISO/TS 16949—Quality Systems: Automotive Suppliers*. The aerospace industry follows the standard *AS 9100: Quality Systems: Aerospace-Model for Quality Assurance in Design, Development, Production, Installation, and Servicing*. In the telecommunications service industry, the *TL 9000* standard is used. A few other industry specific ISO standards are shown in Table 1.

MALCOLM BALDRIGE NATIONAL QUALITY AWARD

In the United States, the strategic importance of quality management is formally recognized through the *Malcolm Baldrige National Quality Award* [14]. The award is monitored through the National Institute of Standards and Technology (NIST), an agency of the US Department of Commerce. The American Society for Quality (ASQ) administers the award program under contract to NIST.

A fundamental reason for the creation of the award is to motivate US companies to improve quality and productivity and become effective and efficient in their operations. The intention is to recognize industry and non-profit organizations that may serve as role models for others in their sector. Awards are given annually, and award winners are expected to disseminate their best practices to others in their field, with the objective that this will lead to adoption of

Table 1. Industry Specific ISO Standards

Standard Number	Application Area
ISO 9069	Information processing—standard generalized markup language
ISO/IEC 9126	Software engineering—product quality
ISO 9362	Banking—business identifier code
ISO 9564	Financial services—personal identification number management
ISO/IEC 9579	Information technology—remote database access
ISO 9992	Financial transaction cards—messages between the integrated circuit card and the card accepting device
ISO 10012	Measurement management systems
ISO 10218	Robots and robotic devices
ISO 10993	Biological evaluation of medical devices
ISO 13485	Medical devices
ISO 13628	Petroleum and natural gas industries
ISO 14001	Environmental management systems
ISO/TS 16949	Automotive production and relevant service part organizations
ISO 22000	Food safety management systems
ISO 26000	Guidance on social responsibility
ISO 31000	Risk management

Table 2. Malcolm Baldrige Award Criteria Categories for 2013–2014

Categories	Point Values
1. Leadership	120
1.1 Senior leadership	70
1.2 Governance and social responsibilities	50
2. Strategic planning	85
2.1 Strategy development	45
2.2 Strategy implementation	40
3. Customer focus	85
3.1 Voice of the customer	40
3.2 Customer engagement	45
4. Measurement, analysis, and knowledge management	90
4.1 Measurement, analysis, and improvement of organizational performance	45
4.2 Knowledge management, information, and information technology	45
5. Workforce focus	85
5.1 Workforce environment	40
5.2 Workforce engagement	45
6. Operations focus	85
6.1 Work processes	45
6.2 Operational effectiveness	40
7. Results	450
7.1 Product and process results	120
7.2 Customer-focused results	85
7.3 Workforce-focused results	85
7.4 Leadership and governance results	80
7.5 Financial and market results	80
Total points	1000

Source: NIST, “2013–2014 Criteria for Performance Excellence—Malcolm Baldrige National Quality Award,” U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, www.nist.gov.

improved quality management practices and, therefore, a rise in the general health of the industry. Awards are given in the following categories: *manufacturing*—companies or units that sell manufactured products or manufacturing processes, and producers of agricultural, mining, or construction products; *service*—companies or subunits that sell services; *small business*—companies or subunits that are engaged in manufacturing and/or provision of services that have 500 or fewer employees; *nonprofit or government organizations*; *education*—organizations that include elementary and secondary schools, colleges, universities, schools or colleges within universities, professional schools, community colleges, and technical schools; and *health care*—organizations include hospitals, health maintenance organizations (HMOs), health care facilities, and home health agencies.

The award criteria are built on the core values [14] of visionary leadership, customer-driven excellence, organizational and personal learning, valuing employees and partners, agility, focus on the future, managing for innovation, operational effectiveness, social media and social responsibility, focus on results and creating value, and a systems perspective. A detailed list of award criteria categories, subcategories, and point values for the *2013–2014 Criteria for Performance Excellence* in the manufacturing, service, small business, nonprofit, and government categories is shown in Table 2.

In the United States, various States have their own awards for excellence in quality management. The majority of these awards are based on a format and criteria similar to the *Malcolm Baldrige National Quality Award*.

CONCLUSIONS

As tools and techniques of operational methods for achievement of quality have been omitted, we now mention some of the available methods and cite appropriate references. It should be noted that there are additional articles on quality control or SPC in Section 4.1.5 of the *Wiley Encyclopedia of Operations Research and Management Science* (2013).

An article on “Quality Design, Control, and Improvement” is also found in the *Wiley Encyclopedia* (2013).

In product design, identification of the prioritized needs of the customer is quite important. In this context, procedural steps outlined in a planning tool known as *quality function deployment* (QFD) is a possible approach [6, 15–17]. Designing a product that meets the needs of the customer may be accomplished through QFD. Following product design, parts design, and process design may also be accomplished through QFD. Subsequent applications of QFD could be in production planning, prototype testing, finished product testing, and after-sales troubleshooting.

To determine the prioritized customer needs, QFD utilizes a set of weights, typically on a 1–5 ordinal scale. As this is subjective in nature, a better approach could use the analytic hierarchy process (AHP) that determines relative weights based on pairwise comparisons [18]. AHP was initially used in a multicriteria decision-making environment where the decision maker is interested in satisfying multiple objectives and is able to come up with a prioritized ranking of the objectives. Consistency in the relative ranking of the objectives is achieved through a pairwise comparison approach.

The subject of SPC deals with the monitoring of processes through statistical methods using *control charts*. Two cause systems that abound in processes are *common* and *special* causes. Common causes are those that are inherent to the system, whereas special causes exist because of some changes in the process. A process is said to be in statistical control when only common causes prevail in the system. Two general categories of control charts exist—variables and attributes. A variable is a quality characteristic that is measurable on a numerical scale, for example, thickness or tensile strength of a part. In the category of attributes, the product or item is classified as conforming or nonconforming. Alternatively, the monitoring of the number of nonconformities or defects is another application of attribute charts. Details on SPC may be found in Section 4.1.5.3 of the *Encyclopedia*. Under

variables control charts, typical ones are for the mean and range ($\bar{X}$ and R), mean and standard deviation ($\bar{X}$ and s), individuals and moving range (I and MR), moving average (MA), and the exponentially weighted moving average ($EWMA$) charts [4, 19]. Similarly, for attributes, typical control charts are for the proportion nonconforming (p), the number nonconforming (np), number of nonconformities (c), or the number of nonconformities per unit (u) chart [4, 19].

A procedure that monitors the quality of outgoing product, when 100% inspection is not feasible, is that of acceptance sampling. A sample of items is chosen from a batch or lot, and a decision is made to accept or reject the lot on the basis of the information contained in the sample. The parameters of the sampling plan are the sample size (n) and the acceptance number (c), which represents the maximum number of nonconforming items that can be observed in the sample in order to accept the batch. These parameters are determined on the basis of certain stipulated conditions, such as the risk of rejecting a “good” lot or the risk of accepting a “bad” lot [4, 20].

Quality in a product is often achieved through product design. Most products consist of multiple components and/or subassemblies. Adequate operation of the product is influenced by the functioning of the components and subassemblies and by the configuration through which they constitute the product. The reliability of a product or system is the probability of the system functioning effectively for a certain period of time under stated environmental conditions. The interested reader may refer to the article [16] in the Encyclopedia for an overview. Additional details on reliability computations for complex systems, procedures for analyzing the criticality of various failure modes, reliability testing sampling plans, accelerated testing, and design of experiments (DOEs) to assess reliability may be found in the following references [4, 21–23]. While the above references are focused on reliability of physical systems, an approach for effective software development makes use of a capability maturity model (CMM) [24].

Often, it is of interest to determine optimum settings of process or product parameters that accomplish a desired objective. For example, one may wish to identify the optimum level of process parameters such as speed, feed, and depth of cut in a machining operation, as well as a product parameter indicating the type of alloy to use, so as to minimize surface roughness. As data needs to be collected under various parameter settings, the topic of DOEs indicates a framework under which experimental runs are to be conducted. Further, statistical techniques exist to analyze the data to determine the influential factors and their settings [25]. Typically, methods of analysis of variance are used.

Some operational methods for quality improvement that describe procedures and related tools and techniques within the procedure are also found in the literature. Among these are the Six-Sigma approach and the Lean Six-Sigma approach developed in the industry as practical guides. Six Sigma may be viewed as a philosophy, methodology for continuous improvement, or as a metric of performance [4, 26, 27]. As an operational procedure, the Six-Sigma approach utilizes the *Define, Measure, Analyze, Improve, and Control* (DMAIC) methodology, details of which may be found in the references.

To summarize, we have provided a strategic overview of quality management. This includes customer identification along with their prioritized needs, some philosophies of quality management, and the impact of quality on an organization’s objectives that include cost minimization. A brief exposure to certification in ISO 9001 standards is contained along with guidelines for the Malcolm Baldrige National Quality Award in the United States.

REFERENCES

1. Crosby PB. Quality is free. New York: McGraw-Hill; 1979.
2. Juran JM. Quality control handbook. 3rd ed. New York: McGraw-Hill; 1974.
3. Garvin DA. What does product quality really mean? Sloan Manage Rev 1984;26(1):25–43.

4. Mitra A. Fundamentals of quality control and improvement. 3rd ed. Hoboken, NJ: Wiley & Sons; 2008.
5. Maslow AH. Theory of human motivation. *Psychol Rev* 1943;50(4):370–396.
6. Cohen L. Quality function deployment: how to make QFD work for you. Reading, MA: Addison-Wesley; 1995.
7. Deming WE. Out of the crisis. Cambridge, MA: Center for Advanced Engineering Study, MIT; 1986.
8. Campanella J. Principles of quality costs: principles, implementation, and use. 3rd ed. Milwaukee, WI: ASQ; 1999.
9. International Organization for Standardization (ISO). *Quality management systems requirements, ISO 9001*; 2008; Geneva, Switzerland.
10. Corbett CJ, Montes-Sancho MJ, Kirsch DA. The financial impact of ISO 9000 certification in the United States: an empirical analysis. *Manage Sci* 2005;51(7):1607–1616.
11. Heras I, Dick GPM, Casadesus M. ISO 9000 registration's impact on sales and profitability: a longitudinal analysis of performance before and after accreditation. *Int J Qual & Reliab Manag* 2002;19(6):774–791.
12. Chow-Chua C, Goh M, Wan TB. Does ISO certification improve business performance? *Int J Qual & Reliab Manag* 2003;20(8):936–953.
13. Naveh E, Marcus A. When does the ISO 9000 quality assurance standard lead to performance improvement? Assimilation and going beyond. *IEEE Trans Eng Manag* 2004;51(3):352–363.
14. National Institute of Standards and Technology (NIST). 2013-2014 Criteria for performance excellence – Malcolm Baldrige national quality award, Gaithersburg, MD: U.S. Department of Commerce National Institute of Standards and Technology. Available at www.nist.gov. Accessed 2013 Oct 11.
15. Akao, Y. QFD: past, present, and future: Proceedings of the 3rd Annual International QFD Conference. Linköping; 1997.
16. Mitra A. Quality design, control, and improvement: *Wiley encyclopedia of operations research and management science*. Hoboken, N.J: Wiley; 2013.
17. Chan L-K, Wu M-L. Quality function deployment: a literature review. *Eur J Oper Res* 2002;143:463–497.
18. Saaty TL. How to make a decision: the analytic hierarchy process. *European Journal of Operational Research*. 1990;48:9–26.
19. Montgomery DC. Statistical quality control. 7th ed. Hoboken, NJ: Wiley & Sons; 2012.
20. Schilling EG, Neubauer DV. Acceptance sampling in quality control. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, Taylor & Francis Group; 2010.
21. Kapur KC, Lamberson LR. Reliability in engineering design. New York: Wiley & Sons; 1977.
22. Mahadevan S. Probability, reliability, and statistical methods in engineering design. New York: Wiley; 2000.
23. Condra LW. Reliability improvement with design of experiments. New York: Marcel Dekker; 2001.
24. Weber CV, Curtis B, Chrissis MB. The capability maturity model: guidelines for improving the software process. Reading, MA: Addison-Wesley; 1995.
25. Box GEP, Hunter JS, Hunter WG. Statistics for experimenters. 2nd ed. Hoboken, NJ: Wiley-Interscience; 2005.
26. Harry MJ. The vision of 6-sigma: a roadmap for breakthrough. Phoenix, AZ: Sigma Publishing Company; 1994.
27. Harry MJ, Schroeder R. *Six sigma: the breakthrough management strategy revolutionizing the world's top corporations*. New York: Doubleday; 2000.

QUANTUM COMMAND AND CONTROL THEORY

STEVE FORSYTHE
Applied Physics Laboratory-Precision
Engagement, Johns Hopkins
University, Dayton, Ohio

INTRODUCTION

Many military activities have an underlying theoretic body of knowledge. The search theory, for example, provides a basis for algorithms and situations involving searching, such as locating an enemy sub or trying to find a downed pilot. A theoretical foundation serves many useful purposes by enabling analysts to discuss and analyze the problem in a mathematical way. Competing methods can be evaluated on the basis of the theory to identify the relative merits of each. Command and Control (C2) is a critical military activity that has to date not been adequately described by an underlying computational theory. One of the more successful models for C2 is the Find, Fix, Track, Target, Engage, and Assess (F2T2EA) model. The limitation of this model is that it is not general enough: it assumes that there is a target to be engaged. This article presents a more generalized C2 model that focuses on the selection of a course of action (COA) by the commander.

When a military organization has a mission, it must decide how to carry out that mission. The commander's staff generates several different COAs (i.e., possible plans) and evaluates them based on some methodology. The results of these comparisons are presented to the commander so that the best COA can be selected. A common approach defined in the joint doctrine is the weighted numerical comparison method. This method for COA selection involves constructing a matrix with criteria to evaluate each COA against. Figure 1 shows a typical COA

decision matrix. COAs for this example are conventional ballistic missile (CBM), B-2 (bomber aircraft), and tomahawk land attack missile (TLAM). The criteria for the evaluation are listed on the left. Each COA is scored against each criterion and then the score is multiplied by a weight to get the weighted results. The weighted results are then summed and the highest sum is presumed to be the best COA. This approach has several limitations. What are the best criteria? How are appropriate weights chosen? Is the expected value really the best statistic to use? Other, more complex methods exist that overcome some of these limitations. Parnell [1] describes methods for dealing with uncertainty about the alternative scores as well as uncertainty about the weights. The Quantum C2 approach is being evaluated at the Johns Hopkins University Applied Physics Laboratory [2]. This simpler approach has the potential of being more easily adapted to military command and control problems. It is called Quantum C2 because, in Physics, the quantum theory takes objects that were formerly considered point masses and describes them using a probability distribution. Likewise, a key aspect of Quantum C2 is to describe a COA, not as a point estimate, but rather, as a probability distribution in the possible future outcome space.

Note that the assessments of the COAs in Fig. 1 are point estimates. These are the expected values but there is no insight into the extremes. In addition, Fig. 1 assumes that the results are all independent. In reality, if the attack misses the target, the risk of collateral damage goes way up and the probability of kill goes way down. A revised method is needed that takes these issues into account.

BACKGROUND

Military models of C2 take on many forms, each with particular strengths and weaknesses. With the revolution in commu-

COA Selection Decision Matrix							
Criteria	Weight	CBM	B-2	TLAM	Weighted Results		
					CBM	B-2	TLAM
Pd	20	0.6	0.9	0.8	20	60	40
Execution time	20	36	128	300	60	40	20
Collateral damage	20	Low	High	Medium	60	20	40
Fratricide risk	20	Low	Medium	Medium	60	40	40
Economy of force (risk to our forces)	20	Low	Medium	Low	60	40	60
Anti-access (ability to defeat)	20	High	High	Medium	60	60	40
Threats (against this option)	20	Low	Medium	Medium	60	40	40
Total					380	300	280
Recalculate							

Figure 1. Example of a COA selection matrix.

nications fueled by the Internet, new C2 paradigms have been proposed (e.g., net-centric), which propose a significantly different perspective on the organization and function of facilities. Current military missions typically require the participation of allies and nongovernment organizations (NGOs). This complex environment makes it challenging to choose the best COA. A model of C2 should be helpful in developing a methodology for COA selection.

Numerous conceptual models of C2 exist, such as John Boyd's OODA (observe, orient, decide, act) loop. This model was developed to explain fighter aircraft engagements but has been applied to a wide range of situations. Similar models include MAPE (monitor, access, plan, execute), SDA (sense, decide, act), and MAAPPER (monitor, access, analyze, predict, plan, execute, report) [3]. There is also the Lawson C2 model as well as the enterprise theory of C2 [4].

A representation of the OODA loop is shown in Fig. 2.

Other approaches such as Builder's Command Concepts focus on the art of command, suggesting that the essential communications up and down the chain of command can (and should) be limited to disseminating,

verifying, or modifying command concepts [5]. The ideal command concept is one that is so prescient, sound, and fully conveyed to subordinates that it would allow the commander to leave the battlefield before the battle commences, with no adverse effect upon the outcome [4].

THE C2 PROBLEM

The Department of Defense defines C2 as "The exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission. C2 functions are performed through an arrangement of personnel, equipment, communications, facilities, and procedures employed by a commander in planning, directing, coordinating, and controlling forces and operations in the accomplishment of the mission" [6]. For the purposes of this article, we adopt a slightly broader definition of C2, which encompasses both military and civil organizations. We shall define C2 as an allocation of resources by a leader over a set of opportunities to achieve a set of objectives while managing risks. The solution to the C2 problem is a

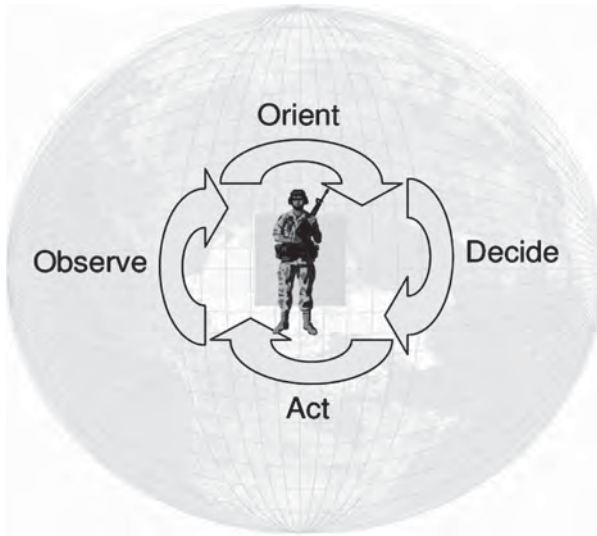


Figure 2. Representation of the OODA loop.

COA and the commander's challenge is to choose the best COA.

The C2 problem, then, addresses

1. how to define the set of opportunities,
2. how to best allocate resources among the opportunities, and
3. how best to carry out those decisions.

In order to analyze how a leader allocates resources over a set of opportunities, we must consider a plethora of different factors. These factors relate to the following:

1. *Desired outcomes*: the objectives we are trying to achieve.
2. *Unwanted outcomes*: those effects that we are trying to avoid or minimize.

Usually, there is a trade-off to be made between achieving the desired outcomes and avoiding unwanted outcomes. The classic example is to neutralize a military facility but not to harm the hospital nearby. With the currently available precision weapons, airstrikes have an excellent chance of destroying a target with minimal collateral damage, unlike in WWII, when bombing campaigns caused massive collateral damage because there were no other options. Numerous issues relate to the commander's

willingness to accept a higher risk of negative effects in order to achieve a greater likelihood of mission success. This trade-off between mission success and risk is the key to C2 and the art of command. The science of C2 is to use modern capabilities to inform the commander of the best set of options that span the mission success versus risk tradespace, so that the commander can apply his training and experience in selecting the best option from a set of good alternatives [7,8].

Figure 3 illustrates the key functions of the C2 process. In this figure, we define the boundary of the C2 process as the information interface with the rest of the world. Information flows in and the C2 system evaluates the information, selects opportunities to pursue, assigns and schedules resources against those opportunities, and then generates orders and other communications to execute the plan.

COA SELECTION

One of the reasons that military decision making is so challenging is that there is always a trade-off between possible unwanted outcomes (casualties and collateral damage) and achieving the desired outcomes of the mission objectives. It is clearly the commander's prerogative to make

Figure 3. The C2 problem. The cloud represents the things outside the C2 system; Φ , the set of information received by the commander (and staff); M, “Machine” that takes Φ as an input and produces Ω ; Ω , the set of opportunities that resources may be assigned against; S, “Machine” that takes Ω as input and produces orders; Orders, the set of communication that indicates the course of action (COA) selected.

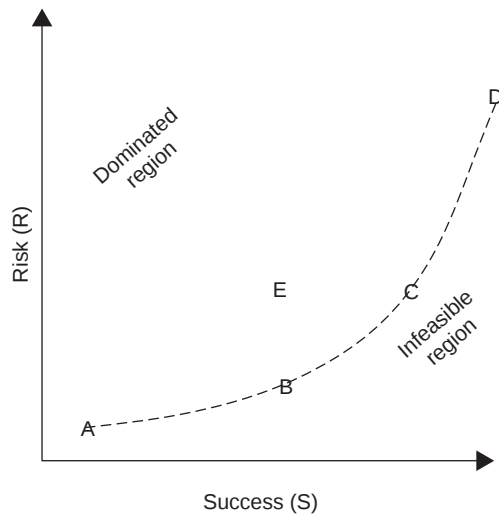
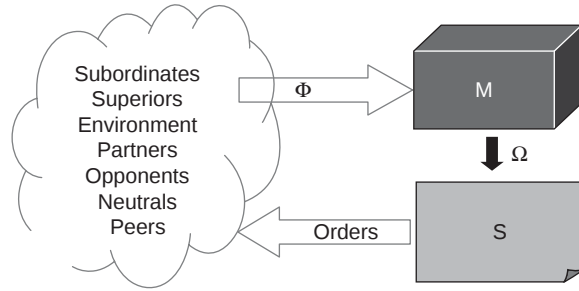


Figure 4. The efficient frontier. A–D represent solutions on the efficient frontier. Each represents the maximum chance of success for a given level of risk. Solution E is dominated by solutions B and C. For the same level of success, B offers lower risk than E, while C offers greater rewards for the same level of risk as E.

this risk versus reward trade-off in choosing the best COA (Fig. 4). Assuming that the commander does not quantify the explicit trade-off (very unlikely) ahead of time, the C2 problem becomes a multiple objective optimization problem.

In situations where there are multiple criteria to optimize, one optimal solution does not exist; there is an efficient frontier of solutions. These solutions maximize the chance of success for a given amount of risk. Other solutions, which do not lie on the efficient frontier, are said to be dominated solutions

as they are strictly inferior to one or more other solutions.

How does one measure risk and reward for military situations? For each of these measures, a utility (or value) function is created, which assigns each solution a value based on the probability of negative outcomes or positive outcomes happening. Casualties and loss of assets are clearly negative outcomes that a commander would like to minimize, while positive outcomes are the accomplishment of various mission objectives.

FACTORS FOR DESIRED OUTCOMES AND UNWANTED OUTCOMES

Figure 5 presents an example of the factors that could be used to calculate a COA's rating for desired outcomes and unwanted outcomes [9,10]. The mission objectives define the desired outcomes and are usually stated in the commander's intent. Also, rules of engagement and constraints on operations are defined by the commander and these define the unwanted outcomes that the commander wants to minimize while maximizing the desired outcomes.

Figure 6 illustrates the two-dimensional COA outcome space. Ideally, a commander would like to accomplish the desired outcomes with few losses, low collateral damage, and so on. Therefore, the lower right quadrant is preferred since most or all of the desired outcomes are achieved with few, if any, unwanted outcomes. The upper left quadrant represents the worst possible outcomes: failing to accomplish the mission yet suffering significant unwanted outcomes. The upper right represents a victory but at

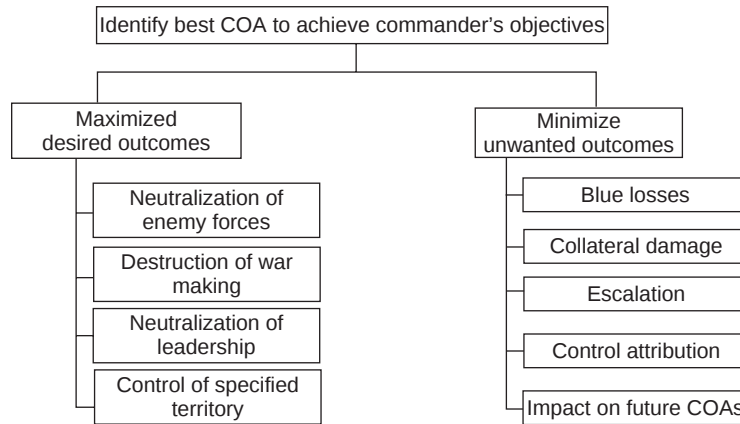


Figure 5. Subobjectives for desired outcomes and unwanted outcomes.

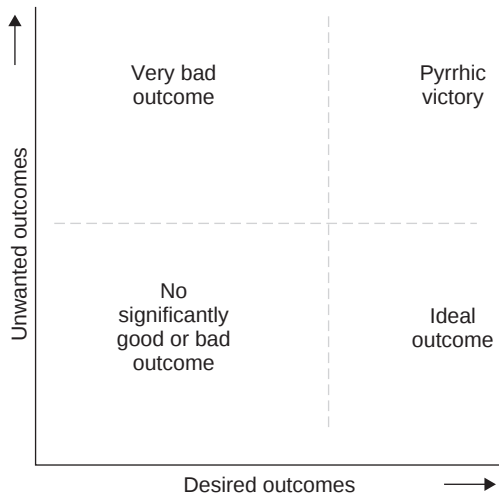


Figure 6. Two-dimensional COA outcome space.

a high cost. This is the very definition of a Pyrrhic victory. An example of a COA that is likely to fall into this quadrant would be the urban bombing campaigns of WWII. With the weapons of the day, there was no way to destroy the war-making capabilities of the enemy without significant collateral damage. The lower left quadrant represents reduced effectiveness but also low risk. An example might be the air campaign prior to ground combat in Desert Storm. Bombing was generally carried out from high altitude so as to minimize attrition, even though

this reduced the effectiveness of any given sortie. The mission requirements were such that moderate effectiveness at low risk was preferred to higher effectiveness at medium risk.

Any given COA has some probability of being in any of these quadrants. Figure 7 shows a probability distribution mapped to the two-dimensional COA outcome space.

Note that the distribution need not be unimodal. Figure 7 shows a bimodal distribution. The most likely result is to achieve most of the desired results with

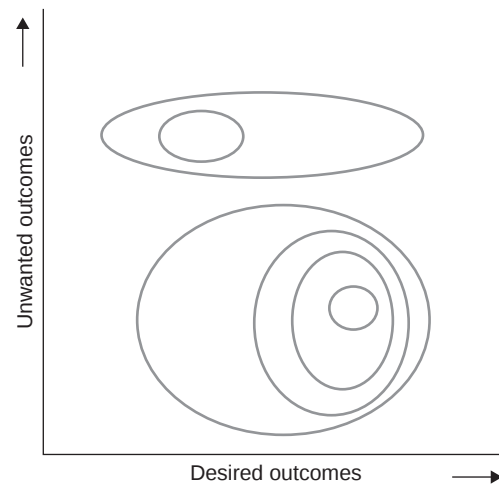


Figure 7. Probability distribution in the two-dimensional COA outcome space.

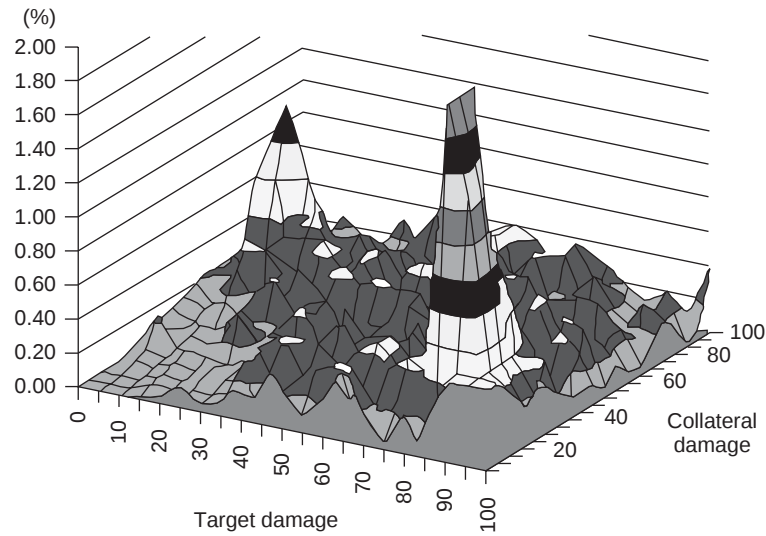


Figure 8. Example probability distribution in the two-dimensional COA outcome space.

some unwanted outcomes. The results in the upper left indicate that there is also some likelihood of achieving only a few of the desired outcomes while suffering significant unwanted outcomes.

In the case of attacking a single target, where the desired outcome is the percentage of the target and the unwanted outcome is collateral damage, the two-dimensional probability outcome space might look like Fig. 8.

Combat simulations often have bimodal distributions where some key early event causes the results to shift from one combatant to the other. It is this type of unusual distribution that makes treating the COA as a probability distribution so important. Just estimating the mean of these values does not convey enough information.

HOW TO EVALUATE COAs

There are several ways that the probability distribution can be used in COA selection. Depending on the commander's comfort with quantitative methods, different methods are available ranging from simple (but still very useful) to more complex. This approach allows a commander to choose the most appropriate approach for a given mission,

thus maximizing the use of the science of C2 in support of the commander's art of C2.

A commander specifies information such as the maximum risk rules of engagement and minimum conditions for success in the commander's intent. This information can be combined with a probability distribution such as the one shown in Fig. 9 to define

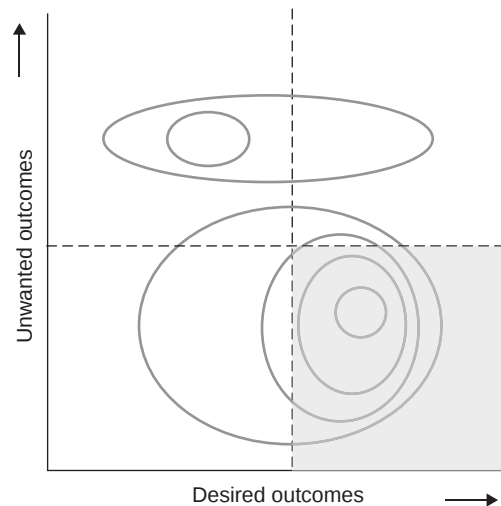


Figure 9. Method 1 for defining the acceptable outcome space.

the acceptable region. One measure of the desirability of a COA is the probability that the COA will have a satisfactory outcome. In many ways, this is the key question that drives a COA selection.

The shaded region in Fig. 9 is the acceptable outcome space. The probability that this COA will result in an acceptable set of outcomes is equal to the percentage of the distribution that lies within the shaded region. Other COAs would have different distributions and so each COA could be evaluated on the basis of the probability of achieving a set of acceptable outcomes.

One limitation to Method 1 is that a commander might be willing to have higher unwanted outcomes if the desired outcomes are achieved. Method 2 begins by asking the commander to refine his statements on risk. What are the maximum unwanted outcomes that are acceptable given minimum acceptable desired outcomes? What are the maximum unwanted outcomes that are acceptable given total success in all desired outcomes?

If a commander has numerous subobjectives for desired outcomes and unwanted outcomes and does not want to use a utility function to combine them into the two-dimensional decision space, an alternative method would be to define an n -dimensional acceptable region where there are n subobjectives. Each COA would have a probability of achieving a result that satisfied all n of the subobjectives. The advantage of this approach is that it avoids the subjective weights of the utility functions. The disadvantage of this approach is that it treats all outcomes in a binary fashion: all acceptable results are equally good and all unacceptable results are equally bad. This n -dimensional method also does not provide the commander with a risk versus mission success trade-off.

Figure 10 shows how the shape of the desired outcome space is modified on the basis of this information. It also clearly shows that a trade-off exists between the desired outcomes and the unwanted outcomes.

Military decision makers often use green/yellow/red connotations to quickly convey information. Method 3 defines the

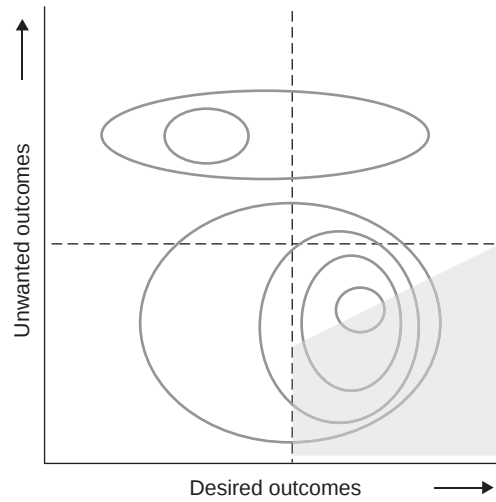


Figure 10. Method 2 for defining the acceptable outcome space.

acceptable outcome space as “green” and a portion of the outcome space could also be defined as “red.” This red space could be thought of as denoting disastrous results or other such descriptions based on the commander’s intent. A “Hail Mary” play in football, for example, has a high probability of getting a great result but it also has a high probability of being intercepted. It tends to be an all-or-nothing choice. Other strategies might have a lower chance of getting an acceptable outcome (getting a first down) but they avoid the disastrous results of a turnover. Using a green/red space such as that displayed in Fig. 11 would allow a commander to compare the likelihood of achieving acceptable results versus the likelihood of the results being extremely bad.

An n -dimensional approach can be taken with this method as well. The n subobjectives would have a range of “green” values that are acceptable, and each objective could have a “red” range of values that the commander wants to avoid at all costs. In this method, each COA would be evaluated to calculate the probability of it achieving an outcome where all the subobjectives are green as well as the probability that any of the subobjectives are red. The COAs could then be graphed on an x - y plot with the x values being the

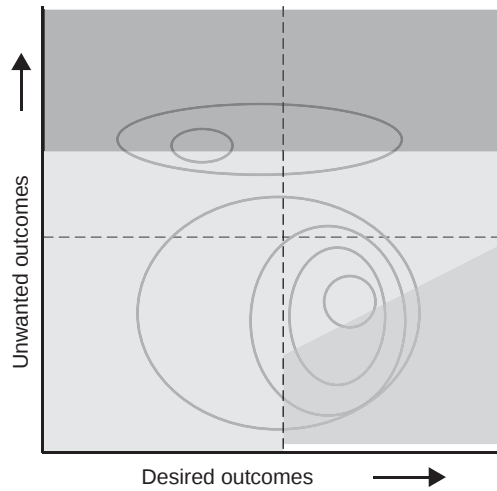


Figure 11. Method 3 for defining the acceptable outcome space.

probability of green outcome and the y values being the probability of a red outcome. Once again, the commander could then make a risk versus mission success trade-off.

Figure 12 shows how the results of using the third method could be displayed for a commander. The COAs are plotted on an xy plot with the x axis being the probability that the outcome will have a green result for all criteria. The y axis is the probability that the outcome will have a result where one or more criteria are red. Once again, the commander is given a trade-off between increasing the likelihood of mission success (all green) versus risk (any red results).

A fourth and final method can be used if the commander or his staff can define a set of relative values for each part of the outcome space. Such a set of relative values (called *utility values*) represents how much the commander values achieving the specified outcomes. Figure 13 shows a notional set of values with the best possible outcome being valued 1000 and the worst possible valued at zero. A simple polynomial surface can be fit to a set of values to allow extrapolation between the points. The probability distribution of each COA is then scored against the utility values of the commander.

Figure 14 shows how the expected utility value for each COA is calculated. By

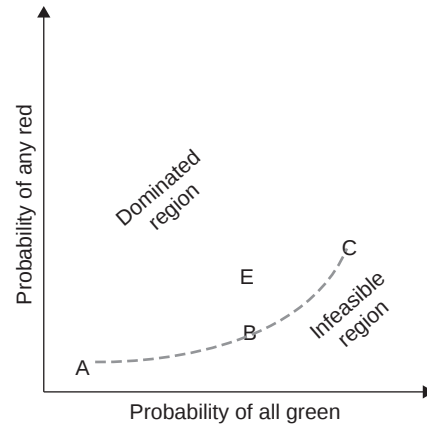


Figure 12. Methodology number three: mission success versus risk trade-off.

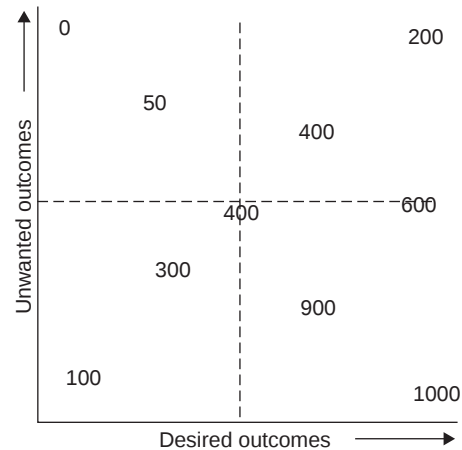


Figure 13. Commander's utility values.

applying values to each possible outcome, this method uniquely values each possible outcome. This utility theory method captures the commander's concern for unwanted outcomes as well as the importance of achieving the mission. This method has the most stringent requirements for data: utility values that the commander has confidence in. By using the utility data, the method allows for the calculation of the highest expected utility. This method therefore takes into account the

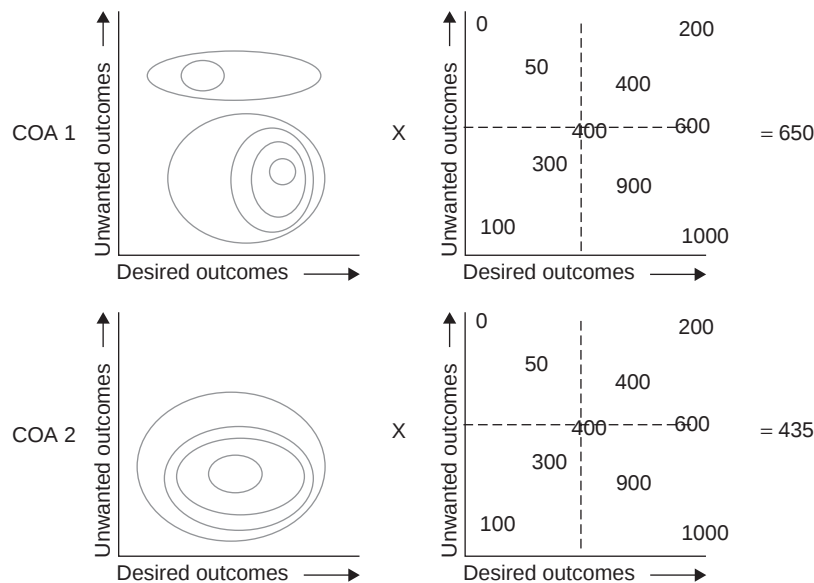


Figure 14. Calculating expected utility value for two COAs.

entire probability distribution and the commander's utility for outcomes associated with all parts of the outcome space.

CONCLUSION

This article provides a new approach to C2 COA selection, which has the potential to be an improvement on the current COA selection methodologies. The Quantum C2 theory treats the COA as a probability distribution that needs to be evaluated to determine the best COA based on the commander's guidance. Four methodologies are discussed, each of which offers some potential advantages over the current COA evaluation methodologies. The simplest method requires only knowledge of the minimum acceptable desired outcomes and unwanted outcomes. The most sophisticated approach uses utility theory to calculate the COA with the maximum expected utility. Thus, the Quantum C2 approach can accommodate different commanders with different approaches to COA evaluation.

REFERENCES

1. Parnell GS. Value-focused thinking. Methods for conducting military operational analysis. Washington (DC): Military Operations Research Society; 2007. Chapter 19.
2. Forsythe SL. Alternative COA selection methodologies: the quantum command and control theory. International Command and Control Research and Technology Symposium (ICCRTS). Santa Monica (CA); 2010.
3. North PD. Experimentation via the use of an executable workflow model to evaluate C2 decision quality. Command and Control Research and Technology Symposium. Seattle (WA); 2008.
4. North PD, Forsythe SL. A process decomposition approach for evaluating command and control (C2) functional performance. Command and Control Research and Technology Symposium. San Diego (CA); 2006.
5. Builder CH, Bankes SC, Nordin R. Command concepts, a theory derived from the practice of command and control. Santa Monica (CA): RAND; 1999.
6. Defense Technical Information Center. DoD Dictionary of Military and Associated Terms. Joint Publication 1-02. Available at <http://www.dtic.mil/doctrine/jel/doddict/data/c/01085.html> Accessed 2007 Jun.

7. Alberts DS, Hayes RE. Planning: complex endeavors. Washington (DC): CCRP Publication; 2007.
8. Alberts DS, Hayes RE. Understanding command and control. Washington (DC): CCRP Publication; 2006.
9. Loerch AG, Rainey LB. Methods for conducting military operations analysis. Washington (DC): MORS Publication; 2007.
10. Keeney RL. Value focused thinking. Cambridge (MA): Harvard University Press; 1992.

QUANTUM GAME THEORY

STEVEN E. LANDSBURG
Department of Economics,
University of Rochester,
Rochester, New York

Quantum game theory is the study of strategic behavior by agents with access to quantum technology. Broadly speaking, this technology can be employed in either of two ways: as part of a randomization device or as part of a communications protocol.

When it is used for randomization, quantum technology allows players to coordinate their strategies in certain ways. The equilibria that result are all correlated equilibria in the sense of Aumann [1], but they form a particularly interesting subclass of correlated equilibria, namely, those that are both achievable and deviation-proof when players have access to certain naturally defined technologies. Not all correlated equilibria can be implemented via quantum strategies, and of those that can, not all are quantum-deviation-proof.

When players have access to private information, the theories of correlated and quantum equilibrium diverge still further, with the classical equivalence between mixed and behavioral strategies breaking down, and the appearance of new equilibria that have no classical counterparts.

The second game theoretic application of quantum technology, other than randomization, is to communication. This leads to a new set of equilibria that seem to have no natural interpretation in terms of correlated equilibria or any other classical concepts.

In the first section, we review the elements of game theory, with special emphasis on those aspects that are generalized or modified by quantum phenomena. In the second section, we survey games with quantum randomization and in the third section, we survey games with quantum communication.

GAME THEORY

Games

A *two-person game* consists of two sets S_1 and S_2 and two maps

$$P_1 : S_1 \times S_2 \rightarrow \mathbf{R} \text{ and } P_2 : S_1 \times S_2 \rightarrow \mathbf{R},$$

where $\mathbf{R}$ denotes the real numbers. The sets S_i are called *strategy sets* and the functions P_i are called *payoff functions*. Games are intended to model strategic interactions between agents who are usually called *players* (though the players are not part of this formal definition). The value $P_i(x, y)$ is called the *payoff* to Player i when Player 1 chooses strategy x and Player 2 chooses strategy y .

For simplicity, we will usually assume the sets S_i are finite.

A *solution concept* is a function that associates to each game a subset of $S_1 \times S_2$; the idea is to pick out those pairs of strategies that we believe players might select in real-world situations modeled by the game. The appropriate solution concept depends on the intended application. The most studied solution concept is the *Nash equilibrium*. A pair (x, y) is called a Nash equilibrium if x maximizes the function $P_1(-, y)$ and y maximizes the function $P_2(x, -)$.

Mixed Strategies

To provide an accurate model of real-world strategic situations, we must allow for the possibility that players might bend the rules. For example, instead of choosing a single strategy, one or both players might randomize. Starting with a game $\mathbf{G}$, we model this possibility by constructing the *associated mixed strategy game* $\mathbf{G}^{\text{mixed}}$ in which the strategy space S_i is replaced with the set $\Omega(S_i)$ of probability distributions on S_i , and the payoff function P_i is replaced with the function

$$P_i^{\text{mixed}} : (\mu, \nu) \mapsto \int P_i(x, y) d\mu(x) d\nu(y)$$

Although $\mathbf{G}^{\text{mixed}}$ is not the same game as $\mathbf{G}$, it is traditional to refer to a Nash equilibrium in $\mathbf{G}^{\text{mixed}}$ as a *mixed strategy equilibrium* in the game $\mathbf{G}$.

An alternative but equivalent model would allow Player i to choose, not a probability distribution on S_i but an S_i -valued random variable from some allowable set. This is the approach we will generalize in what follows.

Correlated Strategies

In the play of $\mathbf{G}^{\text{mixed}}$, we can imagine that Player i first selects a probability distribution, then selects a random variable (with values in S_i) that realizes that distribution, then observes a realization of that random variable, and then plays accordingly. Implicit in this description is that the random variables available to Player 1 are statistically independent of those available to Player 2.

In the real world, however, this is not always true. In the extreme case, both agents might be able to observe the *same* random variable—such as the price of wheat as reported in the *Wall Street Journal*. We can model this extreme case by replacing $\mathbf{G}$ with a set of games $\{\mathbf{G}_\alpha\}$, one for each value α of the jointly observed random variable. We then analyze each game $\mathbf{G}_\alpha$ separately.

But in less extreme cases, we need a new concept. An *environment* for $\mathbf{G}$ is a pair $(\mathcal{X}_1, \mathcal{X}_2)$, where $\mathcal{X}_i$ is a set of S_i -valued random variables. (We do not assume that random variables in $\mathcal{X}_1$ are necessarily independent of those in $\mathcal{X}_2$.) It is natural to assume that for any $X \in \mathcal{X}_i$ and any map $\sigma : S_i \rightarrow S_i$, the random variable $\sigma \circ X$ is also in $\mathcal{X}_i$. (This models the notion that players should be able to map realizations to strategies any way they want to.)

Given such an environment, we define a new game $\mathbf{G}(E) = \mathbf{G}(\mathcal{X}_1, \mathcal{X}_2)$ as follows:

Player i 's strategy set is $\mathcal{X}_i$. The payoff functions are defined in the obvious way, namely

$$P_i(X, Y) = \int_{S_1 \times S_2} P_i(x, y) d\mu_{X, Y}(x, y),$$

where $\mu_{X, Y}$ is the joint probability distribution on $S_1 \times S_2$ induced by (X, Y) .

Definition 1. The pair of random variables (X, Y) is called a *correlated equilibrium* in $\mathbf{G}$ if it is a Nash equilibrium in the game $\mathbf{G}(\{X\}, \{Y\})$. Two correlated equilibria are *equivalent* if they induce the same probability distribution on $S_1 \times S_2$. We will frequently abuse language by treating equivalent correlated equilibria as if they were identical.

It is easy to prove the following.

Proposition 1. *Let E be an environment and suppose that (X, Y) is a Nash equilibrium in the game $\mathbf{G}(E)$. Then (X, Y) is a correlated equilibrium in $\mathbf{G}$.*

However, the converse to Proposition 1 does not hold.

Example 1. Let $S_1 = S_2 = \{\mathbf{C}, \mathbf{D}\}$. Let X , Y , and W be random variables such that

$$\begin{aligned} \text{Prob}(X = W = \mathbf{C}) &= \text{Prob}(X = W = \mathbf{D}) = 1/8, \\ \text{Prob}(X \neq W = \mathbf{C}) &= \text{Prob}(X \neq W = \mathbf{D}) = 3/8, \\ \text{Prob}(Y = W = \mathbf{C}) &= \text{Prob}(Y = W = \mathbf{D}) = 1/12, \\ \text{Prob}(Y \neq W = \mathbf{C}) &= \text{Prob}(Y \neq W = \mathbf{D}) = 5/12. \end{aligned}$$

Let $E = (\{X, Y\}, \{W\})$. Let $\mathbf{G}$ be the game with the following payoffs:

		Player 2	
		C	D
Player 1	C	(0,0)	(2,1)
	D	(1,2)	(0,0)

It is easy to check that both (X, W) and (Y, W) yield correlated equilibria in $\mathbf{G}$. But (X, W) is not an equilibrium in the game $\mathbf{G}(E)$, though (Y, W) is.

Games with Private Information

In real-world strategic interactions, either player might know something the other does not. We model this situation as a *game of private information*, consisting of two strategy spaces S_i , two sets (called *information sets*) $\mathcal{A}_i$, a probability distribution on $\mathcal{A}_1 \times \mathcal{A}_2$, and two payoff functions

$$P_i : \mathcal{A}_1 \times \mathcal{A}_2 \times S_1 \times S_2 \rightarrow \mathbf{R}.$$

Given a game of private information, the *associated game* $\mathbf{G}^\#$ has strategy sets $S_i^\# = \text{Hom}(\mathcal{A}_i, S_i)$ and payoff functions

$$P_i^\#(F_1, F_2) = \int_{\mathcal{A}_1 \times \mathcal{A}_2} P_i(A_1, A_2, F_1(A_1), F_2(A_2)).$$

(Here $\text{Hom}(A, S)$ denotes the set of all functions from A to S .)

If $\mathbf{G}$ is a game of private information, a *Nash equilibrium* in $\mathbf{G}$ is (by definition) a Nash equilibrium in the ordinary game $\mathbf{G}^\#$.

Now we want to enrich the model so players can randomize. To this end, let E be an environment in the sense of the section titled “Correlated Strategies”; that is, $E = (\mathcal{X}_1, \mathcal{X}_2)$, where $\mathcal{X}_i$ is a set of S_i -valued random variables. We define the *associated environment* $E^\# = (\mathcal{X}_1, \mathcal{X}_2)$ by setting $\mathcal{X}_i^\# = \text{Hom}(\mathcal{A}_i, \mathcal{X}_i)$ and identifying the latter set with a set of $S_i^\#$ -valued random variables.

Thus, if $\mathbf{G}$ is a game of private information with environment E , we can first “eliminate the private information” by passing to the associated game $\mathbf{G}^\#$ and environment $E^\#$, and then “eliminate the random variables” by passing to the associated game $\mathbf{G}^\#(E^\#)$ as in the discussion preceding Definition 1. Now we are studying an ordinary game, where we have the ordinary notion of Nash equilibrium.

Unfortunately, *this construction does not generalize to the quantum context*. To get a construction that generalizes, we need to proceed in the opposite order, by first eliminating the random variables and then eliminating the private information:

Construction 1. Given a game of private information $\mathbf{G}$ with an environment

$E = (\mathcal{X}_1, \mathcal{X}_2)$, define a new game of private information $\mathbf{G}(E)$ as follows:

The information sets $\mathcal{A}_i$ and the probability distribution on $\mathcal{A}_1 \times \mathcal{A}_2$ are as in $\mathbf{G}$. The strategy sets are the $\mathcal{X}_i$. The payoff functions are

$$\begin{aligned} P_i(A_1, A_2, X, Y) \\ = \int_{S_1 \times S_2} P_i(A_1, A_2, s_1, s_2) d\mu_{X,Y(s_1, s_2)}. \end{aligned}$$

Now applying the $\#$ construction to $\mathbf{G}(E)$ gives an ordinary game $\mathbf{G}(E)^\#$.

Theorem 1. $\mathbf{G}(E)^\# = \mathbf{G}^\#(E^\#)$.

Remark. In spirit, and in the language of Kuhn [2], $\mathbf{G}^\#(E^\#)$ is like the associated game with *mixed strategies*, where Player i chooses a probability distribution over maps $\mathcal{A}_i \rightarrow S_i$, while $\mathbf{G}(E)^\#$ is like the associated game with *behavioral strategies*, where Player i chooses, for each element of $\mathcal{A}_i$, a probability distribution over S_i . In Kuhn [2], these games are equivalent in an appropriate sense; here they are actually the same game. The difference is that in Kuhn [2], players choose probability distributions, whereas here they choose random variables. But the two approaches are fundamentally equivalent.

Quantum Game Theory

The notions of mixed strategy and correlated equilibria are meant to model the real-world behavior of strategic agents who have access to randomizing technologies (such as weighted coins). *Quantum game theory* is the analogous attempt to model the behavior of strategic agents who have access to quantum technologies (such as entangled particles).

Broadly speaking, there players might use these quantum technologies in either of two ways: as randomizing devices or as communication devices. We will consider each in turn.

QUANTUM RANDOMIZATION

Quantum Strategies

Just as the theory of mixed strategies models the behavior of players with access to

classical randomizing devices, the theory of quantum strategies models the behavior of players with access to quantum randomizing devices. These devices provide players with access to families of observable quantities that cannot be modeled as classical random variables.

For example, let X, Y, Z and W be binary random variables. Classically we have the near triviality:

$$\begin{aligned} \text{Prob}(X \neq W) &\leq \text{Prob}(X \neq Y) \\ &+ \text{Prob}(Y \neq Z) + \text{Prob}(Z \neq W). \end{aligned}$$

Proof. Imagine X, Y, Z, W lined up in a row; in order for X to differ from W , at least one of X, Y, Z, W must differ from its neighbor.

But if X, Y, Z , and W are quantum mechanical measurements, this inequality need not hold. (The most obvious paradoxes are avoided by the fact that neither X and Z , nor Y and W , are simultaneously observable.)

To model the behavior of agents who can make such measurements, we can mimic the construction preceding Definition 1, replacing the sets $\mathcal{X}_i$ of random variables with sets $\mathcal{X}_i$ of quantum mechanical observables. We require that any $X \in \mathcal{X}_1$ and any $Y \in \mathcal{X}_2$ be simultaneously observable. We call such a pair $E = (\mathcal{X}_1, \mathcal{X}_2)$ a *quantum environment*. (Quantum environments will be defined more precisely in the section titled “The Quantum Environment” below.) Given such a quantum environment and given a game $\mathbf{G}$, we construct a new game $\mathbf{G}(E)$ just as in the discussion preceding Definition 1.

If (X, Y) is a Nash equilibrium in $\mathbf{G}(E)$, we sometimes speak loosely enough to say that (X, Y) is a *quantum equilibrium* in $\mathbf{G}$, though the property of being a quantum equilibrium depends not just on $\mathbf{G}$ but on the environment E . Two quantum equilibria are called *equivalent* if they induce the same probability distribution on $S_1 \times S_2$, and, as with correlated equilibria, we will sometimes speak of equivalent quantum equilibria as if they were identical.

If (X, Y) is a quantum equilibrium then (by the definition of quantum environment),

X and Y are simultaneously observable and hence can be treated as classical random variables. With this identification, it is easy to show that any quantum equilibrium is a correlated equilibrium. (This generalizes Proposition 1.) However, just as in Example 1, the converse need not hold. A pair (X, Y) that is an equilibrium in one quantum environment need not be an equilibrium in another.

The Quantum Environment

Consider a game in which the strategy sets are $S_1 = S_2 = \{\mathbf{C}, \mathbf{D}\}$. Players can implement (ordinary classical) mixed strategies by flipping (weighted) pennies, mapping the outcome “heads” to the strategy $\mathbf{C}$, and the outcome “tails” to the strategy $\mathbf{D}$.

While a classical penny occupies either the state $\mathbf{H}$ (heads up) or $\mathbf{T}$ (tails up), a quantum penny can occupy any state of the form $\psi = \alpha\mathbf{H} + \beta\mathbf{T}$, where α and β are complex scalars, not both zero. A heads/tails measurement of such a penny yields the outcome either $\mathbf{H}$ or $\mathbf{T}$ with probabilities proportional to $|\alpha|^2$ and $|\beta|^2$. Physical actions such as rotating the penny induce unitary transformations of the state space, so that the state ψ is replaced by $U\psi$ where U is some unitary operator on the complex vector space $\mathbf{C}^2$.

(Of course literal macroscopic pennies do not behave this way, but spin-1/2 particles such as electrons do, with “heads” and “tails” replaced by “spin up” and “spin down”.)

A single quantum penny is no more or less useful than a classical randomizing device. If you want to play heads with probability p , you can first apply a unitary transformation that converts the state to some $\gamma\mathbf{H} + \delta\mathbf{T}$ with $|\gamma|^2/(|\gamma|^2 + |\delta|^2) = p$, then measure the heads/tails state of the penny and play accordingly.

However, two players equipped with quantum pennies have something more than a classical randomizing device. A pair of quantum pennies occupies a state of the form

$$\alpha\mathbf{H} \otimes \mathbf{H} + \beta\mathbf{H} \otimes \mathbf{T} + \gamma\mathbf{T} \otimes \mathbf{H} + \delta\mathbf{T} \otimes \mathbf{T},$$

where $\alpha, \beta, \gamma, \delta$ are complex scalars, not all zero. A physical manipulation of the first penny transforms the first factor unitarily and a physical manipulation of the second

penny transforms the second factor unitarily. Subsequent measurements yield the outcomes (heads, heads), (heads, tails), and so forth with probabilities proportional to $|\alpha|^2$, $|\beta|^2$, and so forth. This allows the players to achieve joint probability distributions that cannot be achieved via the observations of independent random variables.

Example 2. Suppose that two pennies begin in the *maximally entangled state* $\mathbf{H} \otimes \mathbf{H} + \mathbf{T} \otimes \mathbf{T}$. Players 1 and 2 apply the transformations U and V to the first and second pennies where

$$U = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

$$V = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}$$

This converts the state from $\mathbf{H} \otimes \mathbf{H} + \mathbf{T} \otimes \mathbf{T}$ to

$$\begin{aligned} U\mathbf{H} \otimes V\mathbf{H} + U\mathbf{T} \otimes V\mathbf{T} &= \cos(\theta - \phi)\mathbf{H} \otimes \mathbf{H} \\ &+ \sin(\theta - \phi)\mathbf{H} \otimes \mathbf{T} - \sin(\theta - \phi)\mathbf{T} \otimes \mathbf{H} \\ &+ \cos(\theta - \phi)\mathbf{T} \otimes \mathbf{T}. \end{aligned}$$

If players map the outcomes $\mathbf{H}$ and $\mathbf{T}$ to the strategies $\mathbf{C}$ and $\mathbf{D}$, then the resulting probability distribution over strategy pairs is

$$\begin{aligned} \text{Prob}(\mathbf{C}, \mathbf{C}) &= \text{Prob}(\mathbf{D}, \mathbf{D}) = \cos^2(\theta - \phi)/2 \\ \text{Prob}(\mathbf{C}, \mathbf{D}) &= \text{Prob}(\mathbf{D}, \mathbf{C}) = \sin^2(\theta - \phi)/2. \end{aligned}$$

We can now make precise the notion of *quantum environment*; a quantum environment is a triple $(\xi, \mathcal{X}_1, \mathcal{X}_2)$ where

1. ξ is a nonzero vector in a complex vector space $\mathbf{C}^{n_1} \otimes \mathbf{C}^{n_2}$ (with n_1 and n_2 assumed finite here, though this could all be generalized)
2. $\mathcal{X}_i$ is a set of unitary operators on $\mathbf{C}^{n_i}$.

The unitary operators fill the same role as the sets of random variables in the section titled “Game Theory”; they are the things that players can observe, and on whose realizations they can condition their strategies.

Quantum Equilibrium

Consider again the game $\mathbf{G}$ from Example 1.

		Player 2	
		C	D
Player 1	C	(0,0)	(2,1)
	D	(1,2)	(0,0)

Let E be the following quantum environment: $\xi = \mathbf{H} \otimes \mathbf{H} + \mathbf{T} \otimes \mathbf{T}$, where $\{\mathbf{H}, \mathbf{T}\}$ is a basis for the complex vector space $\mathbf{C}^2$; $\mathcal{X}_1 = \mathcal{X}_2$ is the set of all operators that take the form

$$M(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \quad (1)$$

when expressed in terms of the basis $\{\mathbf{H}, \mathbf{T}\}$.

Physically, this means that each player has access to one member of a pair of maximally entangled pennies, and can rotate that penny through any angle before measuring its heads/tails orientation.

Taking Player 2’s angle ϕ as given, Player 1 clearly optimizes by choosing θ so that $\sin(\theta - \phi) = 1$, and symmetrically with the players reversed. Thus in equilibrium, the outcomes (2, 1) and (1, 2) are each realized with probability 1/2.

As we have noted earlier, this is of course a correlated equilibrium, but it is more than that. For example, the correlated equilibria (X, W) and (Y, W) of Example 1 are not sustainable as quantum equilibria in this environment.

Quantum Games of Private Information

Let $\mathbf{G}$ be a game of private information as defined in the section titled “Games with Private Information”, and let E be a quantum environment.

We would like to model the play of $\mathbf{G}$ in the environment E as the play of an ordinary game. A natural attempt is to replace $\mathbf{G}$ with the game $\mathbf{G}^\#$ defined in the section titled “Games with Private Information”; recall that a strategy in $\mathbf{G}^\#$ is a map from the information set $\mathcal{A}_i$ to the strategy set S_i .

However, in the quantum case there is no natural way to define an environment $E^\#$ for this game. More precisely, it is shown in Dahl and Landsburg [3] that the natural definition of $E^\#$ makes sense when and only when the measurements in $\mathcal{X}_1 \cup \mathcal{X}_2$ have a single joint probability distribution, so that they can be thought of as classical random variables. In other words, the existence of an environment $E^\#$ is equivalent to the absence of any specifically quantum phenomena.

Therefore, we must generalize not the construction $\mathbf{G}^\#(E^\#)$ from the section titled “Games with Private Information”, but the classically equivalent construction $\mathbf{G}(E)^\#$. In the language of Kuhn [2] (and of the remark at the end of the section titled “Games with Private Information”), quantum game theory allows players to choose behavioral strategies without equivalent mixed strategies.

Example 3 [This example is adapted from Cleve *et al.* [4]]. Let $\mathbf{G}$ be the following game of private information.

The information sets are $\mathcal{A}_1 = \mathcal{A}_2 = \{\text{red}, \text{green}\}$. The probability distribution on $\mathcal{A}_1 \times \mathcal{A}_2$ is uniform. The payoff functions are specified as follows:

If both players observe red

		Player Two	
		C	D
Player One	C	(1,1)	(0,0)
	D	(0,0)	(1,1)

If either player observes green

		Player Two	
		C	D
Player One	C	(0,0)	(1,1)
	D	(1,1)	(0,0)

The environment E is as in the paragraph preceding Equation (1).

In the game $\mathbf{G}(E)^\#$, players choose maps $\mathcal{A}_i \rightarrow \mathcal{X}_i$.

To find equilibria in this game, suppose that Player 1 has chosen to map “red” and “green” to $M(\theta_{\text{red}})$ and $M(\theta_{\text{green}})$, while Player 2 has chosen $M(\phi_{\text{red}})$ and $M(\phi_{\text{green}})$ (where the M matrices are as defined in Equation (1)). Then we can compute Player 1’s expected payoffs as functions on $\mathcal{A}_1 \times \mathcal{A}_2$:

$$EP_1(\text{red}, \text{red}) = \cos^2(\theta_{\text{red}} - \phi_{\text{red}}) \quad (2)$$

$$EP_1(\text{red}, \text{green}) = \sin^2(\theta_{\text{red}} - \phi_{\text{green}}) \quad (3)$$

$$EP_1(\text{green}, \text{red}) = \sin^2(\theta_{\text{green}} - \phi_{\text{red}}) \quad (4)$$

$$EP_1(\text{green}, \text{green}) = \sin^2(\theta_{\text{green}} - \phi_{\text{green}}). \quad (5)$$

Because the probability distribution on $\mathcal{A}_1 \times \mathcal{A}_2$ is uniform, Player 1 seeks to maximize the sum of these four expressions, taking ϕ_{red} and ϕ_{green} as given. At the same time (due to the symmetry of the problem) Player 2 seeks to maximize the identical sum, taking θ_{red} and θ_{green} as given.

Given this, we can compute that there are two types of equilibria:

1. Equilibria in which exactly three of the four expressions (2)–(5) are equal to 1 and the fourth is equal to 0. All four possibilities occur with $\phi_r, \phi_g, \theta_r, \theta_g \in \{0, \frac{\pi}{2}, \pi\}$. In these equilibria each player receives a payoff of $3/4 = 0.75$.
2. Equilibria equivalent to $\{\phi_{\text{red}} = 0, \phi_{\text{green}} = 3\pi/4, \theta_{\text{red}} = \pi/8, \theta_{\text{green}} = 3\pi/8\}$. In these equilibria each player receives a payoff of $(1/2 + \sqrt{2}/4) \approx 0.85$.

The best equilibrium that can be reached in any classical environment is equivalent to an equilibrium of type (1). This is so even when the classical environment includes correlated random variables.

For ordinary games, we observed that every quantum equilibrium is also a correlated equilibrium. For games of private information, this example demonstrates that no analogous statement is true.

QUANTUM COMMUNICATION

Communication

In any real-world implementation of a game, players must somehow communicate their strategies before they can receive payoffs. This follows from the fact that the payoff functions take both players' strategies as arguments; therefore, information about both strategies must somehow be present at the same place and time.

Ordinarily, the communication process is left unmodeled. But in this section, we need an explicit model so that we can explore the effects of allowing quantum communication technology. To that end we postulate a referee who communicates with the players by handing them markers (say pennies), which the players can transform from one state to another (say by flipping them over or leaving them unflipped) to indicated their strategy choices; the markers are eventually returned to the referee who examines them and computes payoffs.

Of course not all real-world games have literal referees; sometimes the players are firms, their strategies are prices, and the prices are communicated not to a referee but to a marketplace, where payoffs are "computed" via market processes. In the models to follow, the referee can be understood as a metaphor for such processes.

A Quantum Move

Consider the game with the following payoff matrix (for now, view the labels **NN**, **NF**, etc., as arbitrary labels for strategies):

		Player 2	
		N	F
Player 1	NN	(1,0)	(0,1)
	NF	(0,1)	(1,0)
	FN	(0,1)	(1,0)
	FF	(1,0)	(0,1)

As noted above, classical game theory does not ask how players communicate with the referee. If we want to embellish our model

with an explicit communication protocol, there are several essentially equivalent ways to do it.

The Simplest Protocol. The referee passes two pennies to Player 1, who flips both to indicate a play of **FF**, flips only the first to indicate a play of **FN**, and so forth, and one penny to Player 2, who flips (**F**) or does not (**N**). The pennies are returned to the referee, who examines their states and makes payoffs accordingly.

An Alternative Protocol. A single penny in state **H** is passed to Player 1, who either flips or does not; the penny is then passed to Player 2, who either flips or does not; the penny is then returned to the Player 1, who either flips or does not; the penny is then returned to the referee who makes payoffs of (1, 0) if the final state is **H** or (0, 1) if the final state is **T**. (The players are blindfolded and cannot observe each others' plays.)

We can set things up so that flipping and not flipping correspond to the applications of the unitary matrices

$$\mathbf{F} = \begin{pmatrix} 0 & 1 \\ -i & 0 \end{pmatrix}, \mathbf{N} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (6)$$

Now suppose that Player 1 manages to cheat by employing arbitrary unitary operations. Under the simplest protocol, this is equivalent to playing a mixed strategy and gives Player 1 no advantage. But under the alternative protocol, Player 1 can guarantee himself a win by choosing the unitary matrix

$$U = \frac{1}{2} \begin{pmatrix} 1+i & \sqrt{2} \\ -\sqrt{2} & 1-i \end{pmatrix}$$

on his first turn and U^{-1} on his second. This guarantees him a win because $U^{-1} \circ \mathbf{F} \circ U$ and $U^{-1} \circ \mathbf{N} \circ U$ are both diagonal matrices, so that both fix the state **H**. (Recall that states are unchanged by scalar multiplication.) In other words, the final state is now **H** regardless of Player 2's strategy.

Note that it might be quite impossible to prohibit Player 1 from employing the matrix U , for exactly the same reason that it is usually quite impossible to prohibit players for

adopting mixed strategies: When the referee makes his final measurement, the only information revealed is the final state of the penny, not the process by which it achieved that state.

This example, due to Meyer [5], illustrates two points: First, quantum communication can matter. Second, whether or not quantum communication matters depends not just on the game $\mathbf{G}$; it depends on the specified communications protocol.

The Eisert–Wilkens–Lewenstein Protocol

The Eisert–Wilkens–Lewenstein protocol ([6,7]) captures the potential effects of quantum communication in a quite general context and is therefore the most studied protocol in games of quantum communication.

We start with a game $\mathbf{G}$ in which the strategy sets are $S_1 = S_2 = \{\mathbf{C}, \mathbf{D}\}$. (Everything can be generalized to larger strategy sets, but we will stick to this simplest case.) The referee prepares a pair of pennies in the state $\mathbf{H} \otimes \mathbf{H} + \mathbf{T} \otimes \mathbf{T}$, and passes one penny to each player, with Player 1 receiving the “left-hand” penny.

Players are instructed to play/operate with one of the unitary matrices $\mathbf{F}$ and $\mathbf{N}$ of Equation (6), with $\mathbf{F}$ indicating a desire to play $\mathbf{D}$ and $\mathbf{N}$ a desire to play $\mathbf{C}$. Players actually operate with the unitary matrices of their choices and then return the pennies to the referee, who makes a measurement that distinguishes among the four states that could result if the players followed instructions.

We can of course model this situation by saying that the original game $\mathbf{G}$ has been replaced by the *associated quantum game* $\mathbf{G}^Q$ (not to be confused with the associated quantum games that arise in the very different context of the section titled “Quantum Randomization”), with strategy sets equal to the unitary group U_2 (or, equivalently—because states are unchanged by scalar multiplication—the special unitary group SU_2) and payoff functions having the obvious definition. It is observed in Landsburg [8] that $\mathbf{G}^Q$ is equivalent to the following game:

- The strategy sets S_i^Q are both equal to the group of unit quaternions (which is

isomorphic to the special unitary group SU_2).

- The payoff functions are defined as follows:

$$P_i^Q(\mathbf{p}, \mathbf{q}) = A^2 P_i(\mathbf{C}, \mathbf{C}) + B^2 P_i(\mathbf{C}, \mathbf{D}) \\ + C^2 P_i(\mathbf{D}, \mathbf{C}) + D^2 P_i(\mathbf{D}, \mathbf{D}),$$

where P_i is the payoff function in $\mathbf{G}$ and where

$$\mathbf{pq} = A + Bi + Cj + Dk.$$

The group structure on the strategy sets guarantees that (except in the uninteresting case where both payoff functions are maximized at the same arguments) there can be no pure-strategy equilibria in this game, because Player 1, taking Player 2’s strategy $\mathbf{q}$ as given, can always choose $\mathbf{p}$ to maximize his own payoff, in which case Player 2 cannot be maximizing. So the game $\mathbf{G}^Q$ is quite uninteresting unless we allow mixed strategies.

A mixed strategy in $\mathbf{G}^Q$ is a probability measure on the strategy space $S_1^Q = SU_2 = S^3$, where S^3 is the three sphere. Thus the strategy spaces in $(\mathbf{G}^Q)^{\text{mixed}}$ are quite large. However, it is shown in Landsburg [8] that in equilibrium, both players can be assumed to adopt strategies supported on at most four points, and that each set of four points must lie in one of a small number of highly restrictive geometric configurations. This considerably eases the problem of searching for equilibria.

Example 4 [The Prisoner’s Dilemma]. Consider the “Prisoner’s Dilemma” game

		Player 2	
		C	D
Player 1	C	(3,3)	(0,5)
	D	(5,0)	(1,1)

There is only one Nash equilibrium, only one (classical) mixed strategy equilibrium, and only one (classical) correlated equilibrium, namely, $(\mathbf{D}, \mathbf{D})$ in every case. But it is an easy exercise to check that there are multiple mixed-strategy equilibria in the Eisert–Wilkens–Lewenstein game $\mathbf{G}^Q$.

First example: Each player chooses any of the four orthogonal quaternions and plays each of the four with equal probability. This induces the probability distribution in which each of the four payoffs is equiprobable and each player earns an expected payoff of $9/4$.

Second example: Player 1 plays the quaternions 1 and i with equal probability and Player 2 plays the quaternions j and k with equal probability. This induces the probability distribution where the payoffs $(0, 5)$ and $(5, 0)$ each occur with probability $1/2$ so that each player earns an expected payoff of $5/2$.

The techniques of Landsburg [8] reveal that, up to a suitable notion of equivalence, these are the only mixed strategy equilibria in $\mathbf{G}^Q$.

REFERENCES

1. Aumann R. Subjectivity and correlation in randomized strategies. *J Math Econ* 1974;1:67–96.
2. Kuhn H. Extensive games and the problem of information. *Contributions to the theory of games II*. Annals of Math Studies 28. Princeton (NJ): Princeton University Press; 1953.
3. Dahl G, Landsburg S. Quantum strategies. Preprint. Available at <http://www.landsburg.com/dahcurrent.pdf>.
4. Cleve R, Hoyer P, Toner B *et al*. Consequences and limits of nonlocal strategies. *Proceedings of the 19th Annual Conference on Computational Complexity*. Amherst (MA): IEEE Computer Society; 2004. pp. 236–249.
5. Meyer D. Quantum strategies. *Phys Rev Lett* 1999;82:1052–1055.
6. Eisert J, Wilkens M. Quantum games. *J Mod Opt* 2000;47:2543–2556.
7. Eisert J, Wilkens M, Lewenstein M. Quantum games and quantum strategies. *Phys Rev Lett* 1999;83:3077.
8. Landsburg S. Nash equilibria in quantum games. To appear in *Proceedings of the American Mathematical Society*. Preprint Available at <http://www.landsburg.com/qgt.pdf>. Accessed in 2010.

QUASI-NEWTON METHODS

YA-XIANG YUAN

State Key Laboratory of
Scientific/Engineering Computing,
Institute of Computational Mathematics
and Scientific/Engineering Computing,
Academy of Mathematics and Systems
Science, Chinese Academy of Sciences,
Beijing, China

INTRODUCTION

Quasi-Newton methods are an important class of methods for solving the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x). \quad (1)$$

Newton step for Equation (1) is obtained by minimizing the second-order Taylor expansion of $f(x)$ at x_k :

$$d_k^N = -(\nabla^2 f(x_k))^{-1} g_k, \quad (2)$$

where $g_k = \nabla f(x_k)$. Newton's method has desirable properties, including fast local convergence. Let x^* be a solution of Equation (1) at which $\nabla^2 f(x^*)$ is positive definite, Newton's method converges at q-quadratically, namely,

$$\|x_k + d_k^N - x^*\| = O(\|x_k - x^*\|^2). \quad (3)$$

However, calculating $\nabla^2 f(x_k)$ is time-consuming and sometimes impossible, particularly for large-scale problems. Quasi-Newton methods are similar to Newton's method, but they do not use the second-derivative matrix $\nabla^2 f(x_k)$. Quasi-Newton methods:

$$x_{k+1} = x_k + d_k = x_k - (B_k)^{-1} g_k \quad (4)$$

can be viewed as a modification of Newton's method. B_k is an $n \times n$ matrix, which satisfies the quasi-Newton condition:

$$B_{k+1}(x_{k+1} - x_k) = g_{k+1} - g_k. \quad (5)$$

The reason for imposing Equation (5) is that we have

$$\nabla^2 f(x_{k+1})(x_{k+1} - x_k) \approx g_{k+1} - g_k. \quad (6)$$

A quasi-Newton method is also called a *variable metric method* if B_k is symmetric and positive definite, because the search direction $d_k = -B_k^{-1} g_k$ is the steepest descent direction under the metric $\|d\|_{B_k} = \sqrt{d^T B_k d}$. In a practical implementation of a quasi-Newton algorithm, either line search or trust region technique should be used to ensure global convergence. In a line search algorithm, we need to compute a steplength $\alpha_k > 0$ satisfying certain line search conditions and let

$$x_{k+1} = x_k + \alpha_k d_k = x_k - \alpha_k (B_k)^{-1} g_k. \quad (7)$$

For exact line search, we require

$$f(x_k + \alpha_k d_k) = \min_{\alpha > 0} f(x_k + \alpha d_k). \quad (8)$$

For Wolfe inexact line search, the stepsize α_k satisfies

$$f(x_k + \alpha_k d_k) < f(x_k) + c_1 \alpha_k d_k^T g_k, \quad (9)$$

$$d_k^T \nabla f(x_k + \alpha_k d_k) \geq c_2 d_k^T g_k, \quad (10)$$

where $c_2 \geq c_1$ are two constants in $(0, 1)$. In a trust region algorithm, we compute the trial step d_k by solving the trust region subproblem:

$$\min_{d \in \mathbb{R}^n} g_k^T d + \frac{1}{2} d^T B_k d \quad (11)$$

$$\text{s. t. } \|d\|_2 \leq \Delta_k, \quad (12)$$

where $\Delta_k > 0$ is the trust region bound updated from iteration to iteration.

QUASI-NEWTON UPDATES

The first quasi-Newton method was invented by Davidon [1] in 1959. Davidon's method was later reformulated by Fletcher and Powell [2];

hence it is called the Davidon, Fletcher, and Powell (DFP) method. It updates the quasi-Newton matrix by the following formula

$$B_{k+1} = B_k - \frac{B_k s_k y_k^T + y_k s_k^T B_k}{s_k^T y_k} + \left(1 + \frac{s_k^T B_k s_k}{s_k^T y_k}\right) \frac{y_k y_k^T}{s_k^T y_k}, \quad (13)$$

where $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. It is easy to verify that B_{k+1} defined by Equation (13) satisfies quasi-Newton condition (5). One good property of the DFP update is that B_{k+1} remains positive definite if B_k is positive definite and if $s_k^T y_k > 0$. Please notice that the condition $s_k^T y_k > 0$ is always true for a strictly convex function. Even for nonconvex functions, $s_k^T y_k > 0$ also holds if α_k satisfies either Equation (8) or (10). Let the inverse of B_k be denoted by H_k , then update formula (13) implies that

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{s_k^T y_k}. \quad (14)$$

Another famous quasi-Newton update was given by Broyden [3], Fletcher [4], Goldfarb [5], and Shanno [6] independently:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k}, \quad (15)$$

$$H_{k+1} = H_k - \frac{H_k y_k s_k^T + s_k y_k^T H_k}{y_k^T s_k} + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k}. \quad (16)$$

The Broyden, Fletcher, Goldfarb, and Shanno (BFGS) update can be obtained by interchanging H and B and interchanging s and y in the DFP update. Hence, *BFGS* is called as the *dual update of DFP*.

For any two matrices $\bar{B}$ and $\hat{B}$ satisfying the quasi-Newton condition, the line combination $\bar{B} + \theta(\hat{B} - \bar{B})$ also satisfies the quasi-Newton condition. Thus, Broyden [3] gives a

family of quasi-Newton updates:

$$B_{k+1}(\theta) = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} + \theta w_k w_k^T, \quad (17)$$

where

$$w_k = \sqrt{s_k^T B_k s_k} \left(\frac{y_k}{s_k^T y_k} - \frac{B_k s_k}{s_k^T B_k s_k} \right),$$

$\theta \in \mathbb{R}^1$ being a parameter. As long as $\theta \neq 1/(1 - \beta_k \gamma_k)$, where $\beta_k = y_k^T H_k y_k / s_k^T y_k$ and $\gamma_k = s_k^T B_k s_k / s_k^T y_k$, we have

$$H_{k+1}(\phi) = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{s_k^T y_k} + \phi v_k v_k^T, \quad (18)$$

where

$$v_k = \sqrt{y_k^T H_k y_k} \left(\frac{s_k}{s_k^T y_k} - \frac{H_k y_k}{y_k^T H_k y_k} \right),$$

and $\phi = \phi(\theta) = (1 - \theta)/(1 + \theta(\beta_k \gamma_k - 1))$. If $\theta > 1/(1 - \beta_k \gamma_k)$, we call the updates defined by Equation (17) as a Broyden-positive family because $B_{k+1}(\theta)$ is positive definite as long as B_k is positive definite and $s_k^T y_k > 0$. If we restrict $\theta \in [0, 1]$, updates defined by Equation (17) are called a *Broyden convex family*.

Different values of the parameter θ lead to different quasi-Newton updates. If we let $\theta = 1/(1 - \gamma_k)$, formula (17) becomes the symmetric rank-1 (SR1) update

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}, \quad (19)$$

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k}, \quad (20)$$

which is given by Broyden [7], Davidon [8], Fiacco and McCormick [9], and Murtagh and Sargent [10] independently. It is easy to see that the dual update of SR1 is still SR1. Therefore, we call SR1 a *self-dual update*. By choosing $\theta = 1/(1 + \gamma_k)$, $1/(1 + \sqrt{\beta_k \gamma_k})$ and

$(\beta_k - 1)/(\beta_k \gamma_k - 1)$ respectively, we obtain the Hoshino [11] update

$$B_{k+1} = B_k - \frac{(B_k s_k + y_k)(B_k s_k + y_k)^T}{s_k^T y_k + s_k^T B_k s_k} + 2 \frac{y_k y_k^T}{s_k^T y_k}, \quad (21)$$

the Xie [12] update

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} + \frac{1}{1 + \sqrt{\beta_k \gamma_k}} w_k w_k^T, \quad (22)$$

and the Yuan [13] update

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} + \frac{\beta_k - 1}{\beta_k \gamma_k - 1} w_k w_k^T. \quad (23)$$

It is easy to see that all the update formulae in Equations (21)–(23) are self-dual updates.

A family of update formulae with three parameters were given by Huang [14]:

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \rho_k \frac{s_k s_k^T}{s_k^T y_k} + (\tau_k s_k + \xi_k H_k y_k) v_k^T, \quad (24)$$

where ρ_k , τ_k , ξ_k are three parameters, and v_k is defined above. The inverse form of the above update formula is

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{1}{\rho_k} \frac{y_k y_k^T}{s_k^T y_k} + (\tau'_k y_k + \xi'_k B_k s_k) w_k^T. \quad (25)$$

If we require B_{k+1} symmetric, we obtain the Huang symmetric family

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \rho_k \frac{s_k s_k^T}{s_k^T y_k} + \phi_k v_k v_k^T. \quad (26)$$

A special update of the Huang family is

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{1}{\rho_k} \frac{y_k y_k^T}{s_k^T y_k}, \quad (27)$$

which can be viewed as a modified BFGS method. Specific choices of the value $1/\rho_k$ include $[2(f(x_k) - f(x_{k+1}) + s_k^T g_{k+1})/s_k^T y_k]$ [15] and $[s_k^T(4g_{k+1} + 2g_k) + 6(f(x_k) - f(x_{k+1}))]/s_k^T y_k]$ [16,17].

It is easy to see that Broyden family can be obtained from Huang family by requiring the quasi-Newton condition (5) and symmetric condition

$$B_{k+1} = B_{k+1}^T. \quad (28)$$

The update formulae discussed so far are all rank-2 updates, in the sense that the rank of the increment matrix $B_{k+1} - B_k$ is at most two.

It is easy to show that any rank-1 quasi-Newton update must be in the following form:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) c_k^T}{c_k^T s_k}, \quad (29)$$

where c_k satisfies $c_k^T s_k \neq 0$. A similar form for the inverse update was first given by McCormick (see Ref. 18). If we let $c_k = s_k$, $B_k^T s_k$, and y_k respectively in the above formula, we obtain Broyden unsymmetric Rank-1 [19] update

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) s_k^T}{s_k^T s_k}, \quad (30)$$

the McCormick update [20]

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) s_k^T B_k}{s_k^T B_k s_k}, \quad (31)$$

and the Pearson update [18]

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) y_k^T}{s_k^T y_k}. \quad (32)$$

It is easy to see that the Pearson update is the dual update of the McCormick update. If $u_k = B_k^T y_k$, Equation (29) is

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) y_k^T B_k}{y_k^T B_k s_k}, \quad (33)$$

which is obtained by Broyden [19], and is called *Broyden's other method* or *Broyden's poor method* [21].

An important rank-2 update that does not belong to Broyden family is the symmetrization of Broyden's unsymmetric update (30):

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) s_k^T + s_k (y_k - B_k s_k)^T}{s_k^T s_k} - \frac{s_k^T (y_k - B_k s_k) s_k s_k^T}{(s_k^T s_k)^2}, \quad (34)$$

which was first obtained by Powell [22], and is called *Powell's symmetric Broyden* (PSB) update. Powell's technique can be used to symmetrize the general rank-1 update (29) to obtain the Dennis–Moré [21,23] update

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k) c_k^T + c_k (y_k - B_k s_k)^T}{c_k^T s_k} - \frac{s_k^T (y_k - B_k s_k) c_k c_k^T}{(c_k^T s_k)^2}. \quad (35)$$

It is easy to see that DFP is a special case of Equation (35) when $c_k = y_k$. Moreover, if $c_k = y_k + t w_k$, (35) is Broyden family (17) with $\theta = (1 + t \sqrt{s_k^T B_k s_k / s_k^T y_k})^2 - t^2 / s_k^T y_k$. Hence all updates in Broden convex family ($\theta \in [0, 1]$) are special cases of Equation (35), for example, BFGS is a special case of Equation (35) when $c_k = y_k \sqrt{s_k^T y_k} + B_k s_k \sqrt{s_k^T B_k s_k}$. Moreover the SR1 is also a special case of Equation (35) when $c_k = y_k - B_k s_k$.

Davidon [24] generalized the Broyden family to the following family

$$H_{k+1} = H_k - \frac{H_k \bar{y}_k \bar{y}_k^T H_k}{\bar{y}_k^T H_k \bar{y}_k} + \frac{\bar{s}_k \bar{s}_k^T}{\bar{s}_k^T \bar{y}_k} + \phi \bar{v}_k \bar{v}_k^T, \quad (36)$$

where ϕ is a parameter and $\bar{v}_k = \sqrt{\bar{y}_k^T H_k \bar{y}_k}$ $\left(\frac{\bar{s}_k}{\bar{s}_k^T \bar{y}_k} - \frac{H_k \bar{y}_k}{\bar{y}_k^T H_k \bar{y}_k} \right)$ with $\bar{s}_k$ and $\bar{y}_k \in \mathbb{R}^n$ satisfying $\bar{s}_k - H_k \bar{y}_k = s_k - H_k y_k$ and $\bar{y}_k^T \bar{s}_k = \bar{y}_k^T s_k = y_k^T \bar{s}_k$. When $\bar{s}_k = s_k$ and $\bar{y}_k = y_k$, Davidon reduces to the Broyden family.

Another class of update formulae has the following form:

$$B_{k+1} = \rho \left(B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \theta w_k w_k^T \right) + \frac{y_k y_k^T}{s_k^T y_k}. \quad (37)$$

They are called *self-scaling update formulae*, which was first given by Oren [25]. One such specific formula is the self-scaling BFGS update [26,27]:

$$B_{k+1} = \frac{s_k^T y_k}{s_k^T B_k s_k} \left(B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \right) + \frac{y_k y_k^T}{s_k^T y_k}. \quad (38)$$

As given above, there are lots of different quasi-Newton updates. It is natural to question which quasi-Newton update is the best one. Numerical results indicate that the BFGS method is one of the most efficient quasi-Newton methods. Powell [28] studied the global behavior of BFGS and DFP by considering special quadratic objective functions. However, there are no firm theoretical results showing that the BFGS is better than other quasi-Newton methods. Though efforts for searching for better quasi-Newton updates have never been stopped, it seems very unlikely to find a quasi-Newton method that is much better the BFGS method.

INVARIANCE AND QUADRATIC TERMINATION

One important property of quasi-Newton method is invariance, namely, that it is not affected by linear transformations. Let A be a nonsingular matrix, $a \in \mathbb{R}^n$; consider the linear transformation

$$\tilde{x} = Ax + a, \quad (39)$$

the objective function is now

$$\tilde{f}(\tilde{x}) = f(A^{-1}(\tilde{x} - a)) = f(x). \quad (40)$$

It is easy to see that

$$\nabla_{\tilde{x}} \tilde{f}(\tilde{x}) = A^{-T} \nabla f(x). \quad (41)$$

The following result is obvious (for example, see Ref. 29).

Lemma 1. *Assume that $\tilde{x}_k = Ax_k + a$, and $\tilde{H}_k = AH_kA^T$, then $\tilde{d}_k = -\tilde{H}_k\tilde{g}_k = Ad_k$ and $\tilde{d}_k^T\tilde{g}_k = d_k^Tg_k$. Furthermore, if $\tilde{\alpha}_k = \alpha_k$, we have $\tilde{x}_{k+1} = Ax_{k+1} + a$, and $\tilde{y}_k = A^{-T}y_k$.*

The above lemma indicates that a quasi-Newton method is invariant as long as $\tilde{\alpha}_k = \alpha_k$ and $\tilde{H}_k = AH_kA^T$ holds for all k . It is obvious that $\tilde{\alpha}_k = \alpha_k$ if exact line searches are used. If inexact line search conditions are based on $f(x_k)$, $d_k^T\nabla f(x_k)$, $f(x_k + \alpha d_k)$, $d_k^T\nabla f(x_k + \alpha d_k)$, we can see that the same value of $\tilde{\alpha}_k$ and α_k will be accepted. Assume $\tilde{\alpha}_k = \alpha_k$; because $s_k^Ty_k$, $s_k^TB_k s_k$ and $y_k^TH_k y_k$ are all invariant scalars, we can easily show that $\tilde{H}_{k+1} = A\tilde{H}_{k+1}A^T$ if $\tilde{H}_{k+1}$ and H_{k+1} are generated by the same update formula in the Huang family. Thus, by induction we can establish the following theorem.

Theorem 1. *Any method in Huang family is invariant if α_k is determined by tests on $f(x_k)$, $d_k^T\nabla f(x_k)$, or other invariant scalars.*

It is due to this invariance property that we can always assume $H_1 = I$ for studying any quasi-Newton method. The following result is obtained under this assumption.

Theorem 2. *Assume that*

$$f(x) = g^Tx + \frac{1}{2}x^TGx, \quad (42)$$

where $g \in \mathbb{R}^n$ and G is a symmetric positive definite matrix. Let $H_1 = I$; α_k is obtained by exact line searches and $H_k (k > 1)$ are generated by an update formula from the Huang symmetric family (26). Then, for all k we have

$$\begin{aligned} s_i^T\nabla f(x_{k+1}) &= 0, & i &= 1, \dots, k, \\ H_{k+1}y_i &= \rho_i s_i, & i &= 1, \dots, k, \\ s_k^T G s_j &= 0, & i &\neq j; i, j = 1, \dots, k+1. \end{aligned}$$

The above theorem shows that the steps s_i ($i = 1, \dots, k$) are mutual G-conjugate. Thus,

we have the following the quadratic termination result.

Theorem 3. *Under the assumptions of Theorem 2, if $g_k^TH_k g_k \neq 0$ as long as $g_k \neq 0$, there exists $m \leq n$ such that $\|g_{m+1}\| = 0$ and the iterates $\{x_k\}$ are the same as generated by the conjugate gradient method.*

A direct corollary of the above theorem is that all methods in Broyden-positive family has quadratic termination. Quadratic termination was first noticed by Myers [30] for the DPF method.

Because the invariance property of quasi-Newton methods, Theorem 2 is also true if we replace the condition $H_1 = I$ by H_1 being any symmetric positive definite matrix.

From the proof of Theorem 2 (for example, see Ref. 29), the quadratic termination of quasi-Newton methods depends on exact line searches. However, for the SR1 method, this condition is not needed.

Theorem 4. *Assume that H_1 is any symmetric matrix, and that H_{k+1} is given by SR1 update (20). If for $k = 1, 2, \dots, n$, $(s_k - H_k y_k)^Ty_k \neq 0$, and if $s_1, s_2, \dots, s_n$ are linearly independent, we have that $H_{n+1} = H^{-1}$.*

LEAST CHANGE AND ERROR BOUND

The main idea of quasi-Newton methods is to update the quasi-Newton matrix B_k so that (hopefully) it approaches to the Hessian of the objective function $\nabla^2 f(x)$. Because each quasi-Newton matrix B_k satisfies the quasi-Newton condition in the previous search direction, it is reasonable to require B_{k+1} to be not too far away from B_k . Actually many quasi-Newton updates do have certain least-change properties, namely, B_{k+1} minimizes certain norms of $B_{k+1} - B_k$, subject to the quasi-Newton condition (5).

Following the notation of Dennis Jr. and Schnabel [31], define the set

$$Q(y, s) = \{B \in \mathbb{R}^{n \times n} | Bs = y\}. \quad (43)$$

Theorem 5. *Broyden unsymmetric update (30) is the solution of the problem*

$$\min \|B - B_k\|_2, \quad (44)$$

$$\text{s. t. } B \in Q(y_k, s_k). \quad (45)$$

For any $n \times n$ nonsingular symmetric matrix M , we define the matrix norm $\|\cdot\|_{M,F}$ by $\|A\|_{M,F} = \|MAM\|_F$. We have the following theorem [23]:

Theorem 6. *Let $B \in \mathbb{R}^{n \times n}$ symmetric, $s, y \in \mathbb{R}^n$, $M \in \mathbb{R}^{n \times n}$ symmetric and nonsingular. The unique solution for problem*

$$\min \|A - B\|_{M,F} \quad (46)$$

$$\text{s. t. } A \in Q(y, s), \quad A = A^T \quad (47)$$

is

$$B_+ = B + \frac{(y - Bs)s^T M^{-2} + M^{-2}s(y - Bs)^T}{s^T M^{-2}s} - \frac{(y - Bs)^T s}{(s^T M^{-2}s)^2} M^{-2} s s^T M^{-2}. \quad (48)$$

Please notice that, if we let $c = M^{-2}s$, the above general least change update (48) is Dennis–Moré update (35), which is essentially the Broyden family (17).

From the above theorem, we can get the following results.

Corollary 1. *Assume that B_k is symmetric, B_{k+1} defined by PSB update (34) is the solution of*

$$\min \|B - B_k\|_F, \quad (49)$$

$$\text{s. t. } B \in Q(y_k, s_k), \quad B = B^T. \quad (50)$$

Corollary 2. *Assume that B_k is symmetric, $s_k^T y_k > 0$. Let M be any symmetric nonsingular matrix M that satisfies $M^{-2}s_k = y_k$, the solution of problem*

$$\min \|B - B_k\|_{M,F} \quad (51)$$

$$\text{s. t. } B \in Q(y_k, s_k), \quad B = B^T \quad (52)$$

is the DFP update (13).

Corollary 3. *Assume that H_k is symmetric, $s_k^T y_k > 0$. Let M be any symmetric nonsingular matrix M that satisfies $M^{-2}s_k = y_k$, the solution of problem*

$$\min \|H - H_k\|_{M^{-1},F} \quad (53)$$

$$\text{s. t. } H \in Q(s_k, y_k), \quad H = H^T \quad (54)$$

is the BFGS update (16).

Another least change from B_k to B_{k+1} is to require that $B_k^{-1/2} B_{k+1} B_k^{-1/2}$ is as close as possible to the identity matrix I . Byrd and Nocedal [32] suggest the matrix function

$$\psi(A) = \text{tr}(A) - \ln(\det(A)), \quad (55)$$

which is well defined for all symmetric positive definite matrices and which achieves its minimum at $A = I$. Fletcher [33] established the following result.

Theorem 7. *Assume that B_k is symmetric and positive definite, and $s_k^T y_k > 0$. Then, problem*

$$\min \psi(B_k^{-1/2} B B_k^{-1/2}) \quad (56)$$

$$\text{s. t. } B \in Q(y_k, s_k), \quad B = B^T \quad (57)$$

is the matrix B_{k+1} defined by BFGS update (15).

In addition to least change property, quasi-Newton updates have another desirable property that the new quasi-Newton matrix is closer to the Hessian of the objective function than the previous one in certain sense. First, let us consider the special case when the objective function is strictly convex and quadratic. When we need to show B_{k+1} is closer to the Hessian matrix G of the objective function than B_k . Fletcher [4] discovered that Broyden convex family has the following desirable property.

Theorem 8. *Let $f(x)$ be given by Equation (42), H_1 symmetric and positive definite,*

$H_k (k > 1)$ be generated by Broyden convex family. Assume that $\sigma_i^{(k)} = \sigma_i(G^{\frac{1}{2}} H_k G^{\frac{1}{2}})$ is the i th largest eigenvalue of $G^{\frac{1}{2}} H_k G^{\frac{1}{2}}$, then

$$\min \left\{ \sigma_i^{(k)}, 1 \right\} \leq \sigma_i^{(k+1)} \leq \max \left\{ \sigma_i^{(k)}, 1 \right\}. \quad (58)$$

All updates in Broyden convex family can be expressed in the form of the general least change update (48), which has the following desirable error bound property.

Theorem 9. Let B_+ be given by update formula (48), and let G be any symmetric matrix satisfying $y = Gs$, then

$$\begin{aligned} \|B_+ - G\|_{M,F}^2 &\leq \|B - G\|_{M,F}^2 - \frac{\|M(B - G)s\|_2^2}{\|M^{-1}s\|_2^2} \\ &= \|B - G\|_{M,F}^2 - \frac{\|M(Bs - y)\|_2^2}{\|M^{-1}s\|_2^2}. \end{aligned} \quad (59)$$

The above theorem indicates that B_+ , comparing to B_k , is closer to G . Inequality (59) is very useful in the study of convergence of quasi-Newton methods.

CONVERGENCE RESULTS

The convergence analysis for quasi-Newton methods have been studied in the past few decades. There are already plenty results on this interesting topic. However, there are still open problems remaining to be solved, for example, the global convergence of the DFP method.

First we consider the case where exact line searches are used.

For general convex functions, Powell [34] established the convergence of DFP method.

Theorem 10. Assume that $f(x)$ is uniformly convex and twice continuously differentiable. Then, $\{x_k\}$ generated by the DFP method with exact line search converges to the unique minimizer x^* of $f(x)$. Moreover, the convergence rate is Q -superlinear, namely,

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0. \quad (60)$$

Dixon [35,36] proved that all methods in Huang family generate the identical sequence $\{x_k\}$ if exact line searches are used. Therefore, Theorem 10 and Dixon's results ensure that exact line search ensures that all quasi-Newton methods in the Huang family will converge superlinearly for uniformly convex functions. For nonconvex objective functions, if $n = 2$, Powell [37] proved that the DFP method ensures the weak convergence result:

$$\liminf_{k \rightarrow \infty} \|g_k\| = 0, \quad (61)$$

if a special exact line search is used, namely, if α_k is the largest number that $f(x_k + \alpha d_k)$ decreases monotonically for $0 \leq \alpha \leq \alpha_k$. If the sequence $\{x_k\}$ generated by the DFP method with exact line searches converges to a point, it can be shown [38] that

$$\lim_{k \rightarrow \infty} \|g_k\|_2 = 0. \quad (62)$$

It is shown by Burmeister [39] and by Schuller and Stoer [40] that the DFP and Broyden family converge to an n -step quadratic, namely, there exists a constant c such that

$$\|x_{k+n} - x^*\|_2 \leq c \|x_k - x^*\|_2^2 \quad (63)$$

holds for all k .

Ritter [41] improves Equation (63) to

$$\|x_{k+n} - x^*\|_2 = o\left(\|x_k - x^*\|_2^2\right). \quad (64)$$

Suppose we define the matrix U_k by

$$U_k = [d_{k+1}/\|d_{k+1}\|_2, \dots, d_{k+n}/\|d_{k+n}\|_2]. \quad (65)$$

Schuller [42] proved that

$$\|x_{k+n} - x^*\| = O\left(\|x_{k+n-1} - x^*\| \|x_{k-1} - x^*\|\right). \quad (66)$$

under the assumption that

$$\limsup_{k \rightarrow \infty} \|U_k^{-1}\|_2 < +\infty \quad (67)$$

The above bound was improved to

$$\|x_{k+n} - x^*\| = O(\|x_{k+n-1} - x^*\| \|x_k - x^*\|). \quad (68)$$

if Equation (67) holds [41]. However, Equation (67) is a very strong condition as it is easy to construct examples that all U_k are singular matrices.

Without assuming Equation (67), the best local Q-suplinearly convergence result is given by Powell [43]:

$$\|x_{k+n} - x^*\| = O(\|x_{k+1} - x^*\| \|x_k - x^*\|). \quad (69)$$

When $n = 2$, Equations (69) and (68) are the same, which is the optimal convergence rate. When $n > 2$, Equation (69) is weaker than Equation (68). Numerical tests indicate that Equation (68) is always true. Moreover, if U_k is singular, and the iterate points are all in a lower dimension subspace, we would have a faster convergence rate. But, to prove Equation (68) without assuming Equation (67) is still an open problem.

From Ortega and Rheinboldt [44], Newton's method with exact line searches converges Q-quadratically. By giving a three-time continuously differentiable example, Powell [43] shows that the Q-order of the quasi-Newton method can be less than 2. It is shown by Yuan [45] that the least Q-order of quasi-Newton method is only 1, for all twice continuously differentiable functions.

Now we consider a special case that $\alpha_k \equiv 1$. This special case is of interest. First, most line search techniques try the unit steplength first. Secondly, when x_k converges to a solution, most quasi-Newton methods (such as DFP, BFGS) with unit steplength satisfy the inexact line search conditions.

Many important results are given by Dennis and Moré [23]. One of them is given as follows.

Theorem 11. *Assume that $f(x)$ is twice continuously differentiable and $\nabla^2 f(x)$ is Lipschitz continuous. Let x^* be a minimizer of $f(x)$ and $\nabla^2 f(x^*)$ positive definite. Then the DFP method and the BFGS method with $\alpha_k = 1$*

converges locally Q-superlinearly if x_1 is sufficiently close to x^ and B_1 is sufficiently close to $\nabla^2 f(x^*)$.*

The key technique for proving the above theorem and other local convergence results is by first showing Q-linear convergence [23]. The local Q-linear convergence is proved by using the following error bound [46]

$$\|B_{k+1} - \nabla^2 f(x^*)\|_{M,F} \leq (1 + \eta_1 \sigma(x_k, x_{k+1})) \|B_k - \nabla^2 f(x^*)\|_{M,F} + \eta_2 \sigma(x_k, x_{k+1}), \quad (70)$$

for some fixed matrix M and positive constants η_1, η_2 , where $\sigma(x, \bar{x}) = \max[\|x - x^*\|, \|\bar{x} - x^*\|]$. Once the local Q-linear convergence is proved, inequality (59) implies that

$$\lim_{k \rightarrow \infty} \frac{\|(B_k - \nabla^2 f(x^*))s_k\|}{\|s_k\|} = 0. \quad (71)$$

The above relation ensures that $\|x_k + d_k - x^*\|/\|x_k - x^*\| \rightarrow 0$, which shows that $\alpha_k = 1$ implies Q-superlinearly convergence [47]. It is easy to see that the above theorem is also true if we replace $\alpha_k = 1$ by $\alpha_k \rightarrow 1$. Limit (71) also implies that $\alpha_k^* \rightarrow 1$, where α_k^* is the stepsize obtained by exact line search. Thus, it can be shown that $\alpha_k = 1$ is acceptable for all sufficiently large k if Wolfe inexact line search is used. Therefore, we know that both BFGS and DFP methods with the special Wolfe inexact line search (always trying $\alpha_k = 1$ first) are locally superlinearly convergent. For local convergence results of other quasi-Newton methods, please see Refs 23, 48.

Now, we consider global convergence of quasi-Newton methods with inexact line searches. The following pioneering result was obtained by Powell [49].

Theorem 12. *Assume that $f(x)$ is convex, and twice continuously differentiable in $\{x; f(x) \leq f(x_1)\}$. For any positive definite B_1 , the BFGS method with Wolfe line search x_k either terminates finitely or Equation (62) holds.*

The above result is extended from BFGS method to all methods in the Broyden convex family except DFP [50].

Theorem 13. *Under the assumptions of the above theorem, let B_{k+1} be given by Broyden convex family, namely, $B_{k+1} = B_{k+1}(\theta_k)$, if there exists $\delta > 0$, such that $1 - \theta_k \geq \delta$, then either $\{x_k\}$ terminates finitely or Equation (62) holds.*

The proof of the above theorem cannot be extended to the DFP method, which lacks the “self-correction” $-B_k s_k s_k^T B_k / s_k^T B_k s_k$. Therefore, the convergence of the DFP method with inexact line search remains as an open problem [51]:

Open Question 1. *Assume that the objective function $f(x)$ is uniformly convex and twice continuously differentiable. Does the DFP method with Wolfe inexact line search converge to the unique minimizer of $f(x)$ for any starting point x_1 and any positive definite initial matrix B_1 ?*

It would be interesting and important to solve the above open question, as the DFP method is the very first quasi-Newton method invented. However, up to now, we are not able to prove the convergence nor able to construct counterexamples. Under additional assumptions, we can give a positive answer to the above open question [52].

Theorem 14. *If either $\|g_k\|_2$ is monotone or $\sum_{k=1}^{\infty} \|s_k\|_2 < \infty$, DFP method with inexact line searches guarantees Equation (62).*

For nonconvex functions, there are no global convergence results for quasi-Newton methods with Wolfe inexact line search. Nocedal [51] gives another open question.

Open Question 2. *Assume that the objective function is twice continuously differentiable and bounded below. Does the BFGS method with Wolfe inexact line search ensure Equation (61) for any starting point x_1 and any positive definite initial matrix B_1 ?*

It is known [53] that Equation (61) may fail for the PRP (Polyak, Polak, and Ribière) method and the BFGS method if α_k is allowed to be any local minimizer of $f(x_k + \alpha d_k)$. A counterexample given in Powell [53] that the BFGS method cycles nearly 8 points. Owing

to Dixon [36], Powell’s counterexample is also valid for other methods in the Broyden family. Moreover, Dai [54] observed that Powell’s counterexample satisfies the Wolfe inexact line search condition if $c_1 < 1/84$. Hence, we have the negative answer for Open Question 2.

Theorem 15. *Consider the BFGS method with Wolfe line search. If $c_1 \leq 1/84$, then for any $n > 2$ there exists a starting point x_1 and a C^∞ function $f(x)$ in $\mathbb{R}^n$ such that Equation (61) fails.*

Dai [54] also gives a two-dimensional example that the BFGS method with Wolfe line search cycles nearly 6 points and Equation (61) does not hold.

LIMITED MEMORY AND SUBSPACE TECHNIQUE

For medium- and small-scale problems, quasi-Newton methods are very efficient, and the BFGS method is one of most widely used algorithm for solving medium- and small-scale problems. However, for large-scale problems (when n is very very large), a quasi-Newton method needs huge storage and the linear algebra computation cost per iteration is very high.

Limited memory quasi-Newton methods try to reduce the memory requirement, while maintain certain quasi-Newton property. Limited memory quasi-Newton methods were first proposed as an extension of the conjugate gradient methods by Perry [55] and Shanno [56], and studied by many others, including Gill and Murray [57], Buckley [58], Buckley and LeNir [59], and Nocedal [60].

Let us write the BFGS update (16) in the following form

$$H_{k+1} = \left(I - \frac{s_k y_k^T}{s_k^T y_k} \right) H_k \left(I - \frac{y_k s_k^T}{s_k^T y_k} \right) + \frac{s_k s_k^T}{s_k^T y_k}. \quad (72)$$

Let $\rho_k = 1/s_k^T y_k$ and $V_k = (I - \rho_k y_k s_k^T)$; applying Equation (72), repeatedly, gives the following formula

$$\begin{aligned} H_{k+1} &= (V_k^T \cdots V_{k-i}^T) H_{k-i} (V_{k-i} \cdots V_k) \\ &\quad + \sum_{j=0}^i \rho_{k-i+j} \left(\prod_{l=0}^{i-j-1} V_{k-l}^T \right) s_{k-i+j} s_{k-i+j}^T \\ &\quad \times \left(\prod_{l=0}^{i-j-1} V_{k-l} \right)^T. \end{aligned} \quad (73)$$

Thus, the $m+1$ step limited memory BFGS method uses the following update

$$\begin{aligned} H_{k+1} &= V_k^T \cdots V_{k-m}^T H_k^{(0)} V_{k-m} \cdots V_k \\ &\quad + \sum_{j=0}^m \rho_{k-m+j} \left(\prod_{l=0}^{m-j-1} V_{k-l}^T \right) s_{k-m+j} s_{k-m+j}^T \\ &\quad \times \left(\prod_{l=0}^{m-j-1} V_{k-l} \right)^T, \end{aligned} \quad (74)$$

where $H_k^{(0)}$ can be a simple scale matrix σI . Assume we know $H_k^{(0)}$, we only need to store $s_i, y_i (i = k-m, \dots, k)$. One particular $H_k^{(0)}$ is $H_k^{(0)} = \frac{s_k^T y_k}{\|y_k\|_2^2} I$. Other choices of $H^{(0)}$ can be found in Liu and Nocedal [61].

Another approach for reducing storage and computing cost is the subspace technique. Subspace quasi-Newton methods are based on the following result (for example, see Ref. 62).

Theorem 16. *Consider the BFGS method applied to a general nonlinear function. If $H_1 = \sigma I (\sigma > 0)$, then the search direction $d_k \in \mathcal{G}_k = \text{SPAN}\{g_1, g_2, \dots, g_k\}$ for all k . Moreover, if $z \in \mathcal{G}_k$ and $w \in \mathcal{G}_k^\perp$, then $H_k z = \mathcal{G}_k$ and $H_k w = \sigma w$.*

Actually, it is easy to see that the above theorem is also true if the BFGS method is replaced by any method in the Broyden family. Owing to this subspace property, in iteration k , we can obtain the search direction d_k by solving a subproblem defined in the lower dimensional subspace $\mathcal{G}_k$ instead of working

in the whole space $\mathfrak{R}_n$. Let $r_k = \dim(\mathcal{G}_k)$ and $z_1, z_2, \dots, z_{r_k}$ be an orthonormal base of $\mathcal{G}_k$. Let $Z_k = [z_1, z_2, \dots, z_{r_k}]$ and define $\bar{g}_k = Z_k^T g_k$ and $\bar{B}_k = Z_k^T B_k Z_k$. Then, the search direction is

$$d_k = -Z_k (\bar{B}_k)^{-1} \bar{g}_k. \quad (75)$$

Owing to Theorem 16, if B_{k+1} is updated by the BFGS formula, $\bar{B}_{k+1}$ can be obtained by the BFGS update formula in the subspace directly. If $r_{k+1} = r_k$, we have

$$\bar{B}_{k+1} = \bar{B}_k - \frac{\bar{B}_k \bar{s}_k \bar{s}_k^T \bar{B}_k}{\bar{s}_k^T \bar{B}_k \bar{s}_k} + \frac{\bar{y}_k \bar{y}_k^T}{\bar{s}_k^T \bar{y}_k}, \quad (76)$$

where $\bar{s}_k = Z_k^T s_k$ and $\bar{y}_k = Z_k^T y_k$. If $r_{k+1} = r_k + 1$, we have

$$\bar{B}_{k+1} = \hat{B}_k - \frac{\hat{B}_k \bar{s}_k \bar{s}_k^T \hat{B}_k}{\bar{s}_k^T \hat{B}_k \bar{s}_k} + \frac{\bar{y}_k \bar{y}_k^T}{\bar{s}_k^T \bar{y}_k} \quad (77)$$

where

$$\hat{B}_k = \begin{pmatrix} \bar{B}_k & 0 \\ 0 & \sigma \end{pmatrix}$$

and

$$\bar{s}_k = Z_{k+1} s_k, \quad \bar{y}_k = Z_{k+1}^T g_{k+1} - \begin{pmatrix} Z_k^T g_k \\ 0 \end{pmatrix}.$$

Thus, we can have a subspace BFGS method, which defines the search direction by Equation (75) with $\bar{B}_1 = \sigma$, and $\bar{B}_k (k \geq 2)$ being updated by Equations (76) and (77). It is proved [62] that the subspace BFGS method and conventional BFGS method in the full space generate the identical iterate sequence $\{x_k\}$. A similar result is also true for quasi-Newton methods with trust regions [63].

Another type of special quasi-Newton methods is that the quasi-Newton matrices are sparse. Schubert [64] is the first to study sparse quasi-Newton updates where a sparse unsymmetric rank-1 update is proposed for solving nonlinear equations. For unconstrained optimization, earliest work on sparse quasi-Newton method was given by Powell and Toint [65] and Toint [66].

For example, if we know that the Hessian of the objective has sparse structure:

$$\left(\nabla^2 f(x)\right)_{ij} = 0, \quad \text{for } (i, j) \in \mathcal{I}, \quad (78)$$

where $\mathcal{I}$ is a set of ordered pairs of integers in the range 1 to n , it is reasonable to require $(B_k)_{ij} = 0$ for $(i, j) \in \mathcal{I}$ and all k . Given B_k satisfying $(B_k)_{ij} = 0$ for $(i, j) \in \mathcal{I}$, Toint [66] suggests to define B_{k+1} by

$$\min \|B - B_k\|_{M, F} \quad (79)$$

$$\text{s. t. } B \in \mathcal{Q}(y_k, s_k), \quad B = B^T \quad (80)$$

$$(B)_{ij} = 0, \quad \text{for } (i, j) \in \mathcal{I}. \quad (81)$$

The solution for problems (79)–(81) can be obtained by solving its corresponding KKT (Karush, Kuhn, and Tucker) system, for details please see Refs 66, 67.

It is quite often that large-scale problems have separable structure, namely, the objective functions are in the following form:

$$f(x) = \sum_{i=1}^m f_i(x), \quad (82)$$

where each of the functions $f_i(x)$ depends only on a few variables. Thus, the partitioned quasi-Newton method defines the search direction by

$$d_k = -\left(\sum_{i=1}^m B_k^{(i)}\right)^{-1} g_k, \quad (83)$$

where each $B_k^{(i)}$ ($i = 1, \dots, m$) satisfies the corresponding quasi-Newton condition

$$B_k^{(i)}(x_k - x_{k-1}) = \nabla f_i(x_k) - \nabla f_i(x_{k-1}). \quad (84)$$

For more details about the partitioned quasi-Newton methods, please see Refs 68–72.

PRACTICAL IMPLEMENTATIONS

In an iteration of a line search quasi-Newton algorithm, the main computation cost is the calculation of the search direction

$d_k = -B_k^{-1}g_k$. If we generate B_k by a quasi-Newton update, the line system

$$B_k d = -g_k \quad (85)$$

needs to be solved. Solving a system of n linear equations requires $O(n^3)$ operations. If we generate H_k directly, $d_k = -H_k g_k$ can be obtained by $O(n^2)$ operations. However, updating H_k is not as stable as updating B_k . One simple explanation is that updating B_k needs $y_k y_k^T / s_k^T y_k$, while updating H_k requires $s_k s_k^T / s_k^T y_k$. Assuming the errors for computing s_k and y_k are both in the magnitude of ε , then the error bounds for computing $y_k y_k^T / s_k^T y_k$ and $s_k s_k^T / s_k^T y_k$ are $O(\varepsilon \|y_k\|_2^2 / s_k^T y_k)$ and $O(\varepsilon \|s_k\|_2^2 / s_k^T y_k)$ respectively. It is easy to see for functions that $\nabla^2 f(x)$ satisfy Lipschitz condition, there exists constant M such that $\|y_k\|_2^2 < M s_k^T y_k$. But on the other hand, it is possible that $\|s_k\|_2^2 / s_k^T y_k \rightarrow \infty$.

In order to update B_k directly while avoiding solving equation (85), Gill and Murray [73] suggest to decompose B_k by LDL^T , where L is a unit lower triangle and D is a diagonal matrix. Then we update both L and D . Because any rank-2 update formulae can be written as

$$B_{k+1} = B_k + \tau_k u_k u_k^T + \xi_k z_k z_k^T, \quad (86)$$

where $\tau_k, \xi_k \in \mathbb{R}$, $u_k, z_k \in \mathbb{R}^n$. Therefore we only need to consider the LDL^T factorization of

$$LDL^T + \tau z z^T. \quad (87)$$

It is easy to see that the above relation can be written as

$$L(D + \tau \bar{z} \bar{z}^T)L^T, \quad (88)$$

where $\bar{z}$ is obtained by solving $L\bar{z} = z$. Hence, the problem now is how to obtain the LDL^T factorization of the sum of a diagonal matrix and a rank-1 matrix. The computation cost for the LDL^T factorization of Equation (87) is about $\frac{3}{2}n^2$ operations. Thus, if we have $B_k = L_k D_k L_k^T$, obtaining the LDL^T of B_{k+1} given by Equation (86) needs about $3n^2$ operations. Once we have

the LDL^T factorization, the search direction d_k can be easily obtained by solving $L_k D_k L_k^T d = -g_k$, which requires solving two triangle linear systems. Thus, the operations for updating the quasi-Newton matrix and computing the search direction for each iteration are in the magnitude of $O(n^2)$. Another advantage of updating B by its LDL^T factorization is that the positive definiteness of the quasi-Newton matrices can be easily preserved, namely, B_k is positive definite as long as all the elements of D_k are positive. Of course, the price that we have to pay is that this approach requires a more sophisticated and complicated programming.

Powell [74] proposed an update based on the symmetric decomposition of H_k . Assume that $H_k = Z_k Z_k^T$, where Z_k is a nonsingular matrix. If we have Z_k , the search direction $d_k = -H_k g_k$ can be obtained within $O(n^2)$ operations. Powell [74] applies BFGS formula to compute the factorization $Z_{k+1} Z_{k+1}^T$ of H_{k+1} . It is not difficult to find that the BFGS formula (16) can be written as

$$H_{k+1} = (I - s_k p_k^T) H_k (I - p_k s_k^T), \quad (89)$$

where

$$p_k = \frac{B_k s_k}{\sqrt{s_k^T y_k s_k^T B_k s_k}} + \frac{y_k}{s_k^T y_k}.$$

Thus, we can let

$$Z_{k+1} = (I - s_k p_k^T) Z_k Q_k, \quad (90)$$

with Q_k being any orthogonal matrix. Because $Z_k^T B_k Z_k = I$, the column vectors $z_k^{(i)}$ of Z_k are B_k -conjugate. Thus, Powell [74] selects orthogonal matrix Q_k such that the first column of $Z_k Q_k$ and s_k are the same direction. Let Q_k be an lower-Hessenberg matrix, then

$$(z_{k+1}^{(i)})^T G z_{k+1}^{(j)} = 0, \quad 1 \leq i \leq k < j \leq n, \quad (91)$$

for convex quadratic functions with exact line searches, where G is the Hessian matrix of the objective function. Owing to the conjugate property (91), we can see the quadratic termination of BFGS method.

Powell [74] also suggests a scaling technique for $z_{k+1}^{(i)}$. Assume that $Z_{k+1} = (z_{k+1}^{(1)}, \dots, z_{k+1}^{(n)})$ is already computed; we scale columns 2 to n in the following way

$$z_{k+1}^{(i)} := \max \left\{ 1, \frac{\sigma_k}{\|z_{k+1}^{(i)}\|_2} \right\} z_{k+1}^{(i)}, \quad (92)$$

where $\sigma_1 = \|z_2^{(1)}\|_2$ and $\sigma_k = \min\{\sigma_{k-1}, \|z_{k+1}^{(1)}\|_2\}$. The scaling operation does not alter the directions of $z_k^{(i)}$, thus the conjugate property (91) holds, which preserves the quadratic termination. However, the scaled matrix does not satisfy BFGS update formula, therefore for nonquadratic functions, it is not a BFGS method. And the convergence of this method is not analyzed for general nonlinear functions.

On the basis of Powell [74], Lalee and Nocedal [75] proposed the automatic column scaling BFGS method, which is proved to be global convergent and locally Q-superlinearly convergent for uniformly convex functions. Another extension of Powell's technique is given by Siegel [76].

NONLINEAR EQUATIONS AND NONLINEAR LEAST SQUARES

Quasi-Newton methods can also be applied to solve nonlinear equations and nonlinear least squares. Nonlinear equations are in the form:

$$F(x) = 0, \quad (93)$$

where $F(x) = (F_1(x), F_2(x), \dots, F_n(x))^T \in \mathbb{R}^n$. The quasi-Newton step d_k for Equation (93) can be obtained by solving

$$B_k d + F(x_k) = 0. \quad (94)$$

$B_k \in \mathbb{R}^{n \times n}$ is the quasi-Newton matrix, which approximates the Jacobian of $F(x)$. We require the quasi-Newton condition $B_{k+1} s_k = y_k$ where $s_k = x_{k+1} - x_k$ and $y_k = F(x_{k+1}) - F(x_k)$. Detailed discussions can be found in Refs 23, 29.

Nonlinear least square problems can be written in the form

$$\min \|F(x)\|_2^2, \quad (95)$$

where $F(x) = (F_1(x), F_2(x), \dots, F_m(x))^T \in \mathbb{R}^m$. Please notice here that m is not necessarily equal n . Let the Jacobian of $F(x)$ be $J(x)$. The Hessian matrix of $\frac{1}{2}\|F(x)\|_2^2$ is the sum of two terms, namely $J(x)^T J(x) + \sum_{i=1}^m F_i(x) \nabla^2 F_i(x)$. Hence we only need to apply the quasi-Newton technique to the second term as the first term is already known. A typical quasi-Newton method for nonlinear least squares, Equation (95) computes the search direction d_k by solving

$$(J(x_k)^T J(x_k) + B_k)d = -J(x_k)^T F(x_k), \quad (96)$$

where B_k is a quasi-Newton matrix. There are different quasi-Newton conditions. For example,

$$B_{k+1}s_k = (J(x_{k+1}) - J(x_k))^T F(x_{k+1}), \quad (97)$$

$$B_{k+1}s_k = J(x_{k+1})^T F(x_{k+1}) - J(x_k)^T F(x_k), \quad (98)$$

and

$$\begin{aligned} B_{k+1}s_k &= J(x_{k+1})^T F(x_{k+1}) - J(x_k)^T F(x_k) \\ &\quad - J(x_{k+1})^T J(x_{k+1})s_k \end{aligned} \quad (99)$$

are three possibilities. For more detailed discussions on quasi-Newton methods for nonlinear least squares problems, please see Refs 67, 77–79.

Acknowledgments

This work is partially supported by Chinese NSF grant 10831006 and by CAS grant kjcxyw-s7.

REFERENCES

- Davidon WC. Variable metric method for minimization. AEC Res Dev Report ANL-5900. Argonne (IL): Argonne National Laboratory; 1959.
- Fletcher R, Powell MJD. A rapid convergent descent method for minimization. *Comput J* 1963;6:163–168.
- Broyden GC. The convergence of a class of double rank minimization algorithms: 2. The new algorithm. *J Inst Math Appl* 1970;6:222–231.
- Fletcher R. A new approach to variable metric algorithms. *Comput J* 1970;13:317–322.
- Goldfarb D. A family of variable metric method derived by variation mean. *Math Comput* 1970;23:23–26.
- Shanno DF. Conditioning of quasi-Newton methods for function minimization. *Math Comput* 1970;24:647–656.
- Broyden GC. Quasi-Newton methods and their application to function minimization. *Math Comput* 1967;21:368–381.
- Davidon WC. Variance algorithms for minimization. *Comput J* 1968;10:406–410.
- Fiacco AV, McCormick GP. Nonlinear programming: sequential unconstrained minimization techniques. New York: John Wiley & Sons, Inc.; 1968.
- Murtagh BA, Sargent RWH. A constrained optimization method with quadratic convergence. In: Fletcher R, editors. *Optimization*. London: Academic Press; 1969. pp. 215–246.
- Hoshino S. A formulation of variable metric methods. *J Inst Math Appl* 1972;10:394–403.
- Xie YF. Best variable metric formula under variations: a new variable metric formula (in Chinese). *Acta Math* 1989;32:721–726.
- Yuan Y. On self-dual update formulae in the Broyden family. *Optim Methods Softw* 1992;1:117–127.
- Huang HY. Unified approach to quadratically convergent algorithms for function minimization. *J Optim Theory Appl* 1970;5:405–423.
- Yuan Y. A modified BFGS algorithm for unconstrained optimization. *IMA J Numer Anal* 1991;11:325–332.
- Biggs MC. Minimization algorithms making use of non-quadratic properties of the objective functions. *J Inst Math Appl* 1971;8:315–327.
- Yuan Y, Byrd R. Non-quasi-Newton updates for unconstrained optimization. *J Comput Math* 1995;13:95–107.
- Pearson JD. Variable metric methods of minimization. *Comput J* 1969;12:171–178.
- Broyden GC. A class of methods for solving nonlinear simultaneous equations. *Math Comput* 1965;19:577–593.
- McCormick GP, Pearson JD. Variable metric methods and unconstrained optimization. In: Fletcher R, editor. *Optimization*. London: Academic Press; 1969.
- Dennis JE. On some methods based on Broyden's secant approximation of the Hessian. In: Lootsma FA, editor. *Numerical methods for nonlinear optimization*. London: Academic Press; 1972. pp. 19–33.

22. Powell MJD. A new algorithm for unconstrained optimization. In: Rosen JB, Mangasarian OL, Ritter K, editors. *Nonlinear programming*. New York: Academic Press; 1970. pp. 31–66.
23. Dennis JE, Moré JJ. Quasi-Newton methods, motivation and theory. *SIAM Rev* 1977;19: 46–89.
24. Davidon WC. Optimally conditioned optimization algorithms without line searches. *Math Program* 1975;9:1–30.
25. Oren SS. Self-scaling variable metric algorithms for unconstrained minimization [PhD thesis]. Department of Engineering Economics Systems. Stanford University; 1972.
26. Shanno DF, Phua KH. Matrix conditioning and nonlinear optimization. *Math Program* 1978;14:149–160.
27. Nocedal J, Yuan Y. Analysis of a self-scaling quasi-Newton method. *Math Program* 1993; 61:19–37.
28. Powell MJD. How bad are the BFGS and DFP methods when the objective function is quadratic? *Math Program* 1986;34:34–47.
29. Fletcher R. Volume 1, *Practical methods of optimization, Unconstrained optimization*. Chichester: John Wiley & Sons, Ltd.; 1980.
30. Myers GE. Properties of conjugate gradient and Davidon methods. *J Optim Theory Appl* 1968;2:209–219.
31. Dennis JE Jr, Schnabel RB. *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs (NJ): Prentice-Hall; 1983.
32. Byrd R, Nocedal J. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM J Numer Anal* 1989;26:727–732.
33. Fletcher R. A new variational result for quasi-Newton formulae. *SIAM J Optim* 1991;1:18–21.
34. Powell MJD. On the convergence of the variable metric algorithm. *J Inst Math Appl* 1971; 7:21–36.
35. Dixon LCW. Quasi-Newton algorithms generate identical points. *Math Prog* 1972;2: 383–387.
36. Dixon LCW. Variable metric algorithms: necessary and sufficient conditions for identical behavior of nonquadratic function. *J Optim Theory Appl* 1972;10:34–40.
37. Powell MJD. On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables. *Math Program Ser B* 2000;87:281–301.
38. PU DG, YU WC. On the convergence property of the DFP algorithm. *J Qüfu Norm Univ* 1988;14:63–69.
39. Burmeister W. Die konvergenzordnung des Fletcher-Powell algorithms. *Z Angew Math Mech* 1973;53:693–699.
40. Schuller G, Stoer J. Volume 23, *Über die Konvergenzordnung gewisser Rang-2-Verfahren zur Minimierung von Funktionen*, International series of numerical mathematics. Basel: Birkhäuser; 1974. pp. 125–147.
41. Ritter K. On the rate of superlinear convergence of a class of variable metric methods. *Numer Math* 1980;35:293–313.
42. Schuller G. On the order of convergence of certain quasi-Newton methods. *Numer Math* 1974;23:181–192.
43. Powell MJD. On the rate of convergence of variable metric algorithms for unconstrained optimization. In: Ciesielki Z, Olech C, editors. *Proceeding of the International Congress of Mathematicians*. New York: Elsevier; 1984. pp. 1525–1539.
44. Ortega JR, Rheinboldt WC. *Iterative solution of nonlinear equations in several variables*. London: Academic Press; 1970.
45. Yuan Y. On the least Q-order of convergence of variable metric algorithms. *IMAJ Numer Anal* 1984;4:233–239.
46. Broyden GC, Dennis JE, Moré JJ. On the local and superlinear convergence of quasi-Newton methods. *J Inst Math Appl* 1973;12: 223–245.
47. Dennis JE, Moré JJ. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math Comput* 1974;28:549–560.
48. Martinez JM. Local convergence theory of inexact Newton methods based on structured least change updates. *Math Comput* 1990;55:143–167.
49. Powell MJD. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In: Cottle RW, Lemke CE, editors. *Volume 9, Nonlinear programming, SIAM-AMS proceedings*. Philadelphia (PA): SIAM publications; 1976. pp. 53–72.
50. Byrd R, Nocedal J, Yuan Y. Global convergence of a class of variable metric algorithms. *SIAM J Numer Anal* 1987;4:1171–1190.

51. Nocedal J. Theory of algorithms for unconstrained optimization. *Acta Numer* 1992;1: 199–242.
52. Yuan Y. Convergence of DFP algorithm. *Sci China Series A* 1995;38:1281–1294.
53. Powell MJD. Nonconvex minimization calculations and the conjugate gradient method. In: Griffiths DF, editor. Volume 1066, Numerical analysis, Lecture notes in mathematics. Berlin: Springer; 1984. pp. 122–141.
54. Dai Y. Convergence properties of the BFGS algorithm. *SIAM J Optim* 2002;13:693–701.
55. Perry JM. A class of conjugate gradient algorithms with a two-step variable-metric memory. Discussion Paper 269. Illinois: Center for Mathematical Studies in Economics and Management Science. Northwestern University Evanston; 1977.
56. Shanno DF. Conjugate gradient methods with inexact searches. *Math Oper Res* 1978;3: 244–256.
57. Gill PE, Murray W. Conjugate gradient methods for large-scale nonlinear optimization. Technical Report SOL 79–15. Stanford (CA): Department of Operations Research, Stanford University; 1979.
58. Buckley A. A combined conjugate gradient quasi-Newton minimization algorithm. *Math Program* 1978;15:200–210.
59. Buckley A, LeNir A. QN-like variable storage conjugate gradients. *Math Program* 1983;27: 155–175.
60. Nocedal J. Updating quasi-Newton matrices with limited storage. *Math Comput* 1980;35: 773–782.
61. Liu DC, Nocedal J. On the limited memory BFGS method for large-scale optimization. *Math Program* 1989;45:503–528.
62. Gill PE, Leonard NW. Reduced-Hessian quasi-Newton methods for unconstrained optimization. *SIAM J Optim* 2001;12:209–237.
63. Wang ZH, Yuan Y. A Subspace implementation of quasi-Newton trust region methods for unconstrained optimization. *Numer Math* 2006;104:241–269.
64. Schubert LK. Modification of a quasi-Newton method for nonlinear equations with sparse Jacobian. *Math Comput* 1970;24:27–30.
65. Powell MJD, Toint PhL. On the estimation of sparse Hessian matrices. *SIAM J Numer Anal* 1979;16:1060–1074.
66. Toint PhL. A sparse quasi-Newton update derived variationally with a nondiagonally weighted Frobenius norm. *Math Comput* 1981;37:425–433.
67. Sun WY, Yuan Y. Volume 1, Optimization theory and methods: nonlinear programming, Springer optimization and its application series. New York: Springer; 2006.
68. Griewank A, Toint PhL. On the unconstrained optimization of partially separable objective functions. In: Powell MJD, editor. Nonlinear optimization 1981. London: Academic Press; 1982. pp. 301–312.
69. Griewank A, Toint PhL. Partitioned variable metric updates for large structured optimization problems. *Numer Math* 1982;39: 119–137.
70. Griewank A, Toint PhL. Local convergence analysis of partitioned quasi-Newton updates. *Numer Math* 1982;39:429–448.
71. Toint PhL. Global convergence of the partitioned BFGS algorithm for convex partially separable optimization. *Math Program* 1986; 36:290–306.
72. Griewank A. The global convergence of partitioned BFGS on problems with convex decompositions and Lipschitzian gradients. *Math Program* 1991;50:141–175.
73. Gill PE, Murray W. Quasi-Newton methods for unconstrained optimization. *J Inst Math Appl* 1972;9:91–108.
74. Powell MJD. Updating conjugate directions by the BFGS formula. *Math Program* 1987;38:29–46.
75. Lalee M, Nocedal J. Automatic column scaling strategies for quasi-Newton methods. *SIAM J Optim* 1993;3:637–653.
76. Siegel D. Modifying the BFGS update by a new column scaling techniques. *Math Program* 1993;66:45–78.
77. Dennis JE, Gay DM, Welsch RE. An adaptive nonlinear least squares algorithm. *ACM Trans Math Softw* 1981;7:348–368.
78. Fletcher R, Xu CX. Hybrid methods for nonlinear least squares. *IMA J Numer Anal* 1987; 7:371–389.
79. Yabe H, Takahashi T. Factorized quasi-Newton methods for nonlinear least squares problems. *Math Program* 1991;51: 75–100.

QUEUEING DISCIPLINES

MOR HARCHOL-BALTER
Computer Science Department,
Carnegie Mellon University,
Pittsburgh, Pennsylvania

WHY SCHEDULING MATTERS

Try to remember the last time you asked yourself a question like this: “Why is this computer system so slow?” or “Why is it taking so long for the phone company to pick up my call?” At the time, you probably came to the realization that delay was being caused by *other users* who were sharing the database, or who were simultaneously using the call center.

Given that in many situations we do not have resources all to ourselves, one can ask, “How can we reduce user waiting time?” One obvious answer is to purchase more resources. For example, buying more RAM or a faster CPU could help in a computer system. Likewise, adding (human) servers could reduce waiting time at a call center. Such solutions, of course, cost money.

The goal of *scheduling* is to reduce mean delay *for free*! By simply serving jobs in the “right order,” we can often reduce mean delay by an order of magnitude, without requiring the purchase of additional resources. How can this be? What is the “right order” in which jobs should be scheduled? The answer depends on parameters of the system being scheduled, like load and the job size distribution. These parameters and others will be explained in the section titled “Queueing Model.” Typical scheduling policies will then be defined in the next section and analyzed in the section titled “Performance of Scheduling Policies in M/G/1.” Finally, in the section titled “Scheduling in Practice” we will see some recent real-world success stories where scheduling has been used to

greatly decrease average delays. The versatility of scheduling has made it the topic of thousands of research papers. While the majority of this article introduces scheduling under the simplest queueing model (an M/G/1 queue) and considers only the simplest performance metric (mean response time), in the last section we briefly discuss more complex queueing models, performance metrics, and scheduling policies, and give pointers to research papers covering those topics.

QUEUEING MODEL

Most of this article will deal with results of scheduling policies for a single queue, the M/G/1 queue, shown in Fig. 1. In the M/G/1 queue, it is assumed that jobs arrive according to a Poisson process with some average rate λ jobs per second. A “job” may refer to a computing job or may be a customer placing a call to a call center. We depict a job as a person in the figure. As shown in the figure, jobs are not all the same: some jobs have large size (require a large amount of service) and others have small size. For the M/G/1, it is assumed that all job sizes are independent instances from a common general job size distribution (‘G’ = general). We will use the random variable X to denote a job’s size. The expected (mean) job size is $\mathbf{E}[X]$, the second moment of job size is $\mathbf{E}[X^2]$, and the squared coefficient of variation of job size, C_X^2 , is the variance normalized by the square of the mean, namely,

$$C_X^2 = \frac{\text{var}(X)}{\mathbf{E}[X]^2}.$$

The *load*, ρ , of the system is the time-average fraction of the time that the server is busy, and is defined by $\rho = \lambda \mathbf{E}[X]$, where we assume $0 \leq \rho < 1$. In the analysis of scheduling policies, we will assume that the job size distribution is continuous with finite variance. We use $f(\cdot)$ to denote its probability density function (pdf) and $F(\cdot)$ to denote its cumulative distribution function.

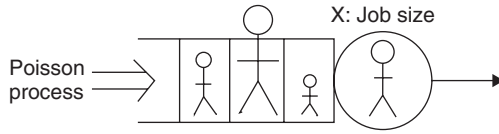


Figure 1. Illustration of an M/G/1 queue. Jobs (or customers) have different sizes (service demands).

In many applications, particularly computer science applications, it is frequently the case that empirical job size distributions exhibit *very high variability*, and are best modeled by a Pareto, a lognormal, or a Weibull distribution, with high C_X^2 . Some examples include Unix process CPU requirements ([1,2], where C^2 values of 25 to 49 have been measured), sizes of files transferred through the Web [3,4], sizes of files stored in Unix file systems [5], durations of FTP transfers in the Internet [6], and CPU requirements for supercomputing jobs [7]. In this article we will therefore focus a lot of attention on the effect of job size variability. *Decreasing failure rate (DFR)* (hereafter referred to as DFR) is also prevalent in computer science applications. Roughly speaking, under DFR, the more service a job has received so far, the greater its remaining service requirement is likely to be. For example, the more CPU a UNIX job has used so far, the more CPU it is expected to use in the future [1], and the more bytes a flow has transmitted so far, the more bytes it is likely to transmit in the future [8].

DEFINITION OF COMMON SCHEDULING POLICIES

In this section, we define the most common scheduling policies for a single-server system. These can be subdivided along two axes (Table 1). *Blind* policies do not make use of a job's size when determining which job to run next, while *size-based* policies do. *Nonpreemptive* policies can not preempt a job once it has started running, while *preemptive* policies allow jobs to be stopped and restarted at a later point, without losing intermediary work. All policies which we consider are *work-conserving*, meaning that the server is always busy working on a job whenever there

Table 1. Taxonomy of some Common Scheduling Policies

	Non-preemptive	Preemptive
Blind to Size	FCFS RANDOM LCFS	PS PLCFS FB
Uses Size	SJF	PSJF SRPT

are jobs to be served, and no work that is done is lost.

We start by describing policies which are nonpreemptive and do not make use of a job's size:

FCFS or FIFO (First-Come-First-Served or First-In-First-Out). Jobs are served in the order that they arrive, with each job being run to completion before the next job receives service. This policy is typically used in manufacturing systems, call centers, and supercomputing centers. In such settings it is often difficult to preempt a running job, and it seems “fair” to serve customers in the order that they arrive.

RANDOM (Random-Order-Service). Whenever the server is free, it chooses a random job in the queue to run next and runs that job to completion. This policy is mostly of theoretical interest.

LCFS (Last-Come-First-Served; nonpreemptive). Whenever the server is free, it chooses the last job to arrive and runs that job to completion. This policy is used for applications where jobs are piled onto a “stack” structure, with each new job being “pushed” on top of the stack. When the server becomes free, it is convenient to “pop” the job off the top of the stack.

The following policies are preemptive and do not make use of a job's size:

PS (Processor-Sharing). Whenever there are n jobs in the system, each job is allocated $1/n$ th fraction of the server. PS has its roots in round-robin scheduling of computing jobs by the operating

system CPU, whereby the CPU rotates among the jobs in the queue in a cyclic fashion, giving each job only a small fixed quantum of service before moving on to the next job [9]. If one imagines the limit as the quantum size goes to zero, one gets PS. Besides its use in modeling CPU processing, PS scheduling is also a good model of equal bandwidth sharing of a network link.

PLCFS (Preemptive-Last-Come-First-Served). At every moment, the server is always running the job that arrived last. When a new jobs arrives, it always preempts the job in service. This policy makes sense, for example, in a market information processing situation, where the most recent job carries the most relevant information (current stock price) and older jobs are less relevant.

FB or LAS or SET (Foreground–Background also known as Least-Attained-Service or Shortest-Elapsed-Time). This policy always serves the job which has obtained the least service so far. We say that a job's *age* is the total service it has received so far. FB (LAS) always serves that job which has the lowest age, resulting in PS if all jobs have the same age. This policy is often used in situations where the job sizes are not known, but it is known that the job size distribution has DFR, for example in CPU scheduling or flow scheduling. Given unknown job size and DFR, serving the job with lowest age is equivalent to serving that job that has the lowest expected *remaining* service requirement. FB has been proven to minimize the mean response time and queue length distribution among blind policies [10].

We now move on to policies which make use of a job's size, favoring shorter jobs in some way. Given that typically the great majority of jobs are "short," favoring short jobs reduces overall mean response time (it is preferable to have long jobs wait behind short jobs, rather than the reverse).

SRPT (Shortest-Remaining-Processing-Time). This preemptive policy always runs that job which has the currently smallest remaining time. If a job arrives with shorter processing requirement than the job in service, then it preempts the job in service. SRPT is known to minimize the mean response time and the queue length [11], however its use is limited to situations where job size and remaining size are known.

PSJF (Preemptive-Shortest-Job-First). This preemptive policy is very similar to SRPT, but favors that job with smallest *original* size rather than smallest *remaining* size. This can be a useful substitute for SRPT when remaining size is not known.

SJF (Shortest-Job-First). Under this non-preemptive variant of PSJF, whenever the server becomes free, it always grabs the shortest job to work on (shortest original size). SJF is used in nonpreemptive settings, where one wants to get the benefit of favoring short jobs, but jobs can not be preempted.

PERFORMANCE OF SCHEDULING POLICIES IN M/G/1

The *response time* (a.k.a., flow time or sojourn time) of a job is the time from when a job arrives until it completes service, and is denoted by T . Since some scheduling policies might favor short jobs, it is also instructive to consider $T(x)$, the response time for a job conditioned on that job being of size x . In this section we state results on mean response time, $\mathbf{E}[T]$, and on $\mathbf{E}[T(x)]$ for an M/G/1 queue under the above scheduling policies. While we do not have room to prove these results, we will provide high-level intuition behind each formula. We refer the reader to Refs 12–14, where complete proofs of these results can be found.

While the formulas for the response time that we present below are all well-known, the comparative ranking between the different scheduling policies is not obvious

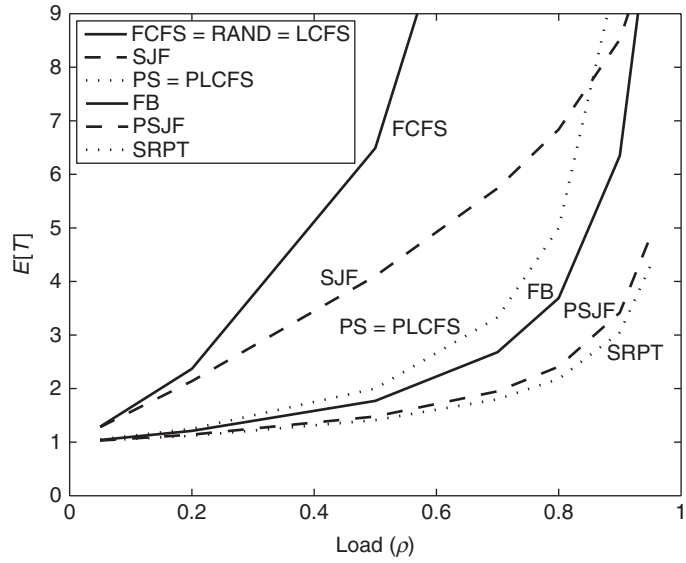


Figure 2. Mean response time as a function of load for the M/G/1 with various scheduling policies. The job size distribution is a Weibull with mean 1 and $C^2 = 10$.

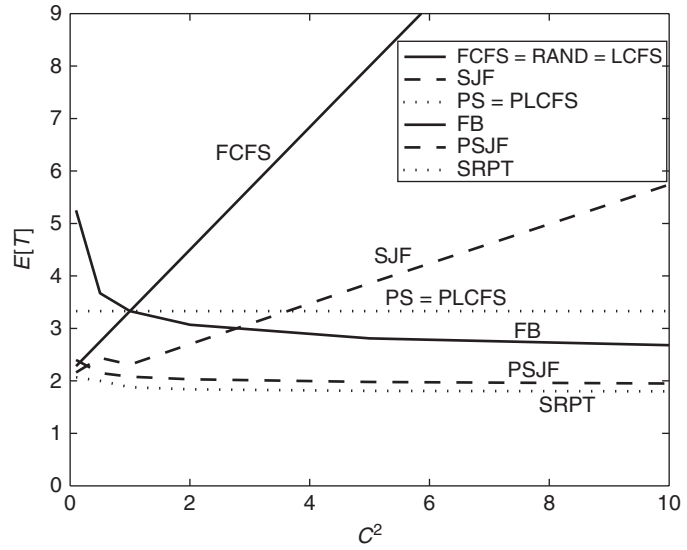


Figure 3. Mean response time as a function of variability (C^2) for the M/G/1 with various scheduling policies. Load is fixed at $\rho = 0.7$. The job size distribution is a Weibull with fixed mean 1 and changing C^2 .

from the formulas, many of which include triple-nested integrals. We have therefore evaluated these formulas using MathematicaTM to compare the mean response time under the different policies as a function of load (Fig. 2) and as a function of C^2 (Fig. 3). In both figures, we assume a Weibull job size distribution. The Weibull, defined by

$\bar{F}(x) = e^{-\left(\frac{x}{\lambda}\right)^\alpha}$ with parameters λ and α , is convenient because when $\alpha < 1$, it has DFR and C^2 can be made as high as desired.

In the analysis below, we will require a few terms that we have not yet introduced. First, we will need to talk about the load made up of just those jobs of size $< x$. We will

denote this by

$$\rho(x) = \lambda \int_0^x tf(t) dt \quad (1)$$

Second, we will often talk about a *busy period*, which is defined to be the time from when a job arrives to an empty system until the system is once again empty. We use B to denote the length of a busy period in an M/G/1 queue (for any work-conserving policy), and $B(x)$ to denote the length of a busy period conditioned on the fact that the starting job is of size x . Then, it is well known [15] that the expected length of $B(x)$ and B increase with load ρ in the following way:

$$\mathbf{E}[B(x)] = \frac{x}{1-\rho}, \quad (2)$$

$$\mathbf{E}[B] = \frac{\mathbf{E}[X]}{1-\rho}. \quad (3)$$

First-Come-First-Served (FCFS)

A job's response time, T , under FCFS can be viewed as its *waiting time*, W (the time from when the job arrives until it first gets service) plus its service time, X . The mean waiting time is given by the well-known Pollaczek–Khinchin (P–K) formula:

$$\mathbf{E}[W]^{\text{FCFS}} = \frac{\lambda \mathbf{E}[X^2]}{2(1-\rho)}. \quad (4)$$

Of key importance in this formula is the variability of the job size distribution (the second moment term $\mathbf{E}[X^2]$). Specifically, even if the load, ρ , is low (say $\rho = \frac{1}{2}$), the mean waiting time can still be quite high if the second moment of the job size distribution is high.

The second moment term shows that mean delay (waiting time) is proportional to the variability of the job size distribution. This makes sense because when there are short jobs mixed in with long jobs, some short jobs will inevitably get stuck behind long jobs, causing large mean delay. Another way to think about Equation (4) is to consider the excess (remaining) service time on the job which is currently in service when a random arrival occurs. One might think that the expected excess for the job in service is $\frac{\mathbf{E}[X]}{2}$, however this is only true when jobs

have deterministic job sizes. When job sizes are variable, a random arrival is more likely to see a larger-than-average excess, and, in fact, by the Inspection Paradox [16], the expected excess turns out to be $\mathbf{E}[X^2]/2\mathbf{E}[X]$, which comes up in the derivation of formula (4) above.

From the mean waiting time, it is easy to derive the mean response time as follows:

$$\mathbf{E}[T(x)]^{\text{FCFS}} = x + \mathbf{E}[W]^{\text{FCFS}} = x + \frac{\lambda \mathbf{E}[X^2]}{2(1-\rho)}, \quad (5)$$

$$\begin{aligned} \mathbf{E}[T]^{\text{FCFS}} &= \mathbf{E}[X] + \mathbf{E}[W]^{\text{FCFS}} \\ &= \mathbf{E}[X] + \frac{\lambda \mathbf{E}[X^2]}{2(1-\rho)}. \end{aligned} \quad (6)$$

Looking at Figs 2 and 3, we see that mean response time under FCFS is worse than any of the other policies when $C^2 > 2$. This is because of the huge role that job size variability plays in response time.

RANDOM and Last-Come-First-Served (LCFS)

While RANDOM and LCFS seem very different from FCFS, it turns out that their mean response time is the same of that of FCFS, and thus they lie on top of FCFS in Figs 2 and 3. In fact, an even stronger claim can be made [14]:

Lemma 1. *For an M/G/1, any nonpreemptive scheduling policy that does not make use of job size will yield the same distribution on the number of jobs in the system.*

The above lemma is easy to understand: Although the different nonpreemptive policies may choose their next job differently at each completion point, the job is *not* chosen based on its size. Therefore the time until the next completion is equal in distribution under all such policies. Thus the number of arrivals between each completion is also stochastically the same across policies, resulting in the same stochastic process for the number of jobs in the system. It then follows by Little's Law [17] that the mean response time is

also the same for all such policies. Hence,

$$\begin{aligned}\mathbf{E}[T]^{\text{RANDOM}} &= \mathbf{E}[T]^{\text{LCFS}} = \mathbf{E}[T]^{\text{FCFS}} \\ &= \mathbf{E}[X] + \frac{\lambda \mathbf{E}[X^2]}{2(1-\rho)}.\end{aligned}\quad (7)$$

Note that although the distribution of the number of jobs in the system is the same for all blind nonpreemptive policies, the distribution of response time is not the same. In particular, it can be shown that [14]

$$\text{var}(T)^{\text{FCFS}} < \text{var}(T)^{\text{RANDOM}} < \text{var}(T)^{\text{LCFS}}.$$

Processor-Sharing (PS)

The PS policy has many pleasing properties. First, the distribution of the number of jobs in the M/G/1/PS turns out to be the same as in an M/M/1/FCFS queue [16]. Letting $\mathbf{E}[N]$ denote the mean number of jobs, we have that

$$\mathbf{E}[N]^{\text{PS}} = \frac{\rho}{1-\rho}.\quad (8)$$

The conditional mean response time for a job of size x is similarly attractive:

$$\mathbf{E}[T(x)]^{\text{PS}} = x/(1-\rho).\quad (9)$$

Another way of stating this is that, under PS, every job, regardless of its size, is slowed down by the same constant factor:

$$\begin{aligned}\text{Constant slowdown factor} &= \frac{\mathbf{E}[T(x)]^{\text{PS}}}{x} \\ &= \frac{1}{1-\rho}.\end{aligned}$$

Observe that $1/(1-\rho) = \mathbf{E}[N]^{\text{PS}} + 1$ (Eq. 8). Intuitively, it makes sense that an arrival is slowed down by a factor equal to the number of jobs it sees plus itself (roughly speaking).

Finally, from Equation (9), we see that

$$\mathbf{E}[T] = \frac{\mathbf{E}[X]}{1-\rho}.\quad (10)$$

This is fascinating because of its complete independence of the job size variability. This property is best illustrated by the perfectly horizontal line in Figure 3, showing the

insensitivity of mean response time to C^2 . The survey paper [18] presents more results on PS.

Preemptive-Last-Come-First-Served (PLCFS)

The PLCFS policy has surprisingly little to do with the LCFS policy, although they are both based on serving the last job to arrive. While the LCFS mean response time is highly affected by job size variability, we will see that the PLCFS mean response time is independent of job size variability.

Consider a new arrival of size x to an M/G/1/PLCFS queue. We will refer to this job as “job x .” All arrivals prior to job x will not get worked on while job x is in the system. However, all arrivals after job x , which occur before job x completes, will have priority over job x . Thus, we can view the response time of job x under PLCFS as the length of a busy period started by a job of size x , namely $B(x)$. From Equation (3), it follows that

$$\mathbf{E}[T(x)]^{\text{PLCFS}} = \frac{x}{1-\rho},\quad (11)$$

$$\mathbf{E}[T]^{\text{PLCFS}} = \frac{\mathbf{E}[X]}{1-\rho}.\quad (12)$$

Thus, the mean response time of the M/G/1/PLCFS is identical to that of the M/G/1/PS queue, in agreement with Figs 2 and 3. However PLCFS has an advantage over PS: typically many fewer preemptions. In fact, under PLCFS, each job creates at most two preemptions: one when the job arrives and one when it departs. Since preemptions are not free, this can be a real advantage.

Foreground–Background (FB) a.k.a. Least-Attained-Service (LAS)

Again consider an arrival of size x , which we refer to as “job x .” Under FB (a.k.a., LAS), the arriving job x immediately receives service (since it is the youngest job in the system). However, it may later have to share both with jobs already in the system and with new arrivals. Jobs in the system only affect job x *until they hit age x* , after which point they are invisible to job x . New arrivals while job x is

in the system will all initially have priority over job x , and will each get to complete *up to* x units of work (actually the minimum of x and their size) before job x leaves.

From the above observations, we can see that if $f(y)$ denotes the pdf for job size, then we need to define a new pdf $f_x(y)$ where jobs of size greater than x have been replaced by jobs of size x . We define

$$f_x(y) = \begin{cases} f(y), & y < x; \\ 1 - F(x), & y = x. \end{cases}$$

We define X_x to be a random variable with density function $f_x(y)$, and define $\rho_x = \lambda \mathbf{E}[X_x]$. Then, the expected response time for a job of size x is given by

$$\begin{aligned} \mathbf{E}[T(x)]^{\text{FB}} &= \frac{x + \frac{\frac{1}{2}\lambda \mathbf{E}[X_x^2]}{1 - \rho_x}}{1 - \rho_x} \\ &= \frac{x(1 - \rho_x) + \frac{1}{2}\lambda \mathbf{E}[X_x^2]}{(1 - \rho_x)^2}, \end{aligned} \quad (13)$$

$$\mathbf{E}[T]^{\text{FB}} = \int_0^\infty \mathbf{E}[T(x)] f(x) dx. \quad (14)$$

Observe that in appearance Equation (13) does not seem all that different from Equation (5), the corresponding equation for FCFS. There is still a dependence on the second moment of the job size distribution, however the job size is now X_x , with lots of extra mass on x and no mass above x . Also the result is scaled by $(1 - \rho_x)$, due to delay incurred by waiting on later-arriving jobs.

In practice, however, these small changes can have a large effect, particularly when the job size distribution has DFR. Figure 2 shows that the mean response time for FB can be quite close to that of SRPT, which is optimal, and quite far from FCFS. Furthermore, Fig. 3 shows that the mean response time for FB can actually decrease with increased job size variability, particularly when increased variability implies even stronger DFR behavior, as in the case of the Weibull job size distribution or the Pareto distribution. The survey paper [19] presents more results on FB.

As a final point, we note that the fact that the curves for FCFS, PS, and FB all

cross in Fig. 3 at the point where $C^2 = 1$ is no accident. For exponentially distributed job sizes ($C^2 = 1$), it is easy to argue that the continuous-time Markov chain tracking the number of jobs in the system is identical under all these policies.

Shortest-Job-First (SJF)

The waiting time for a job of size x under SJF is given below:

$$\mathbf{E}[W(x)]^{\text{SJF}} = \frac{\lambda \mathbf{E}[X^2]}{2(1 - \rho(x))^2}, \quad (15)$$

where $\rho(x)$, defined in Equation (1), denotes the load made up by jobs of size less than x only.

The sharp resemblance of the above formula to FCFS, (Eq. 4), may seem surprising but should not be. While SJF favors short jobs, even a short job still has to wait for the server to first free up. Thus, the waiting time for a job of size x , $\mathbf{E}[W(x)]^{\text{SJF}}$ still includes the excess term, involving $\mathbf{E}[X^2]$, that we saw in FCFS. The only difference is that now we have $(1 - \rho(x))^2$ in the denominator rather than $(1 - \rho)$. The intuition is as follows: When a job of size x arrives, it first has to wait for the server to finish serving its current job (the excess). Then job x has to wait behind those jobs in the queue whose size are $< x$ and these contribute a $(1 - \rho(x))$ factor in the denominator. Finally job x has to wait behind all future arrivals of size $< x$ during its waiting time—these contribute another $(1 - \rho(x))$ factor. Note that in FCFS, job x does not have to wait behind future arrivals.

The mean response time for SJF is given as follows:

$$\mathbf{E}[T(x)]^{\text{SJF}} = x + \mathbf{E}[W(x)]^{\text{SJF}}, \quad (16)$$

$$\mathbf{E}[T]^{\text{SJF}} = \mathbf{E}[X] + \int_0^\infty \mathbf{E}[W(x)]^{\text{SJF}} f(x) dx. \quad (17)$$

From Figs 2 and 3, we see that the performance of SJF is far worse than that of the other policies that favor short jobs (PSJF and SRPT), and in fact most closely resembles that of FCFS. This is because the $\mathbf{E}[X^2]$ term dominates mean response time in SJF, as

it does in FCFS. The second thing to notice is that SJF appears to “get better” as load gets high. This is because the $(1 - \rho(x))^2$ in the denominator depends on $\rho(x)$, which can be far lower than ρ , for high ρ , causing the denominator in SJF (even with its squared effect) to be larger than that in FCFS.

Preemptive-Shortest-Job-First (PSJF)

PSJF is the preemptive version of SJF. There are two components to an arrival’s response time: the time the arriving job waits until it *first* receives service (we call this *initial waiting time*), and the time from when the arrival first receives service until it completes (we call this *residence time*). Since a job in service can be preempted by smaller arrivals, a job’s residence time typically consists of way more than its service time. The mean response time for a job of size x is shown below:

$$\mathbf{E}[T(x)]^{\text{PSJF}} = \frac{\frac{\lambda}{2} \int_0^x t^2 f(t) dt}{(1 - \rho(x))^2} + \frac{x}{(1 - \rho(x))}. \quad (18)$$

The first term in Equation (18) indicates the initial waiting time for an arrival of size x . This is very similar in form to the waiting time for SJF. The key difference is that the $\mathbf{E}[X^2]$ term has been replaced by $\int_0^x t^2 f(t) dt$, which represents the contribution to the second moment made up only of jobs of size $< x$. To understand this change, observe that an arrival of size x no longer has to consider any jobs whose original size is $> x$, since the arrival has immediate priority over those jobs; this is in contrast to SJF, where a small arrival is still affected by the excess of the job in service, no matter what its size.

The second term above is the residence time for a job of size x . Observe that this has the form of the duration of an M/G/1 busy period, started by a job of size x , where the system load is $\rho(x)$ (Eq. 2). This should make perfect sense, since a job of size x starts its residence time when there are no jobs of size $< x$ in the system, however this job is then interrupted by every arrival of size $< x$, and cannot depart until there are again, no jobs of size $< x$ in the system.

The mean response time for PSJF is given as follows:

$$\mathbf{E}[T]^{\text{PSJF}} = \int_0^\infty \mathbf{E}[T(x)]^{\text{PSJF}} f(x) dx \quad (19)$$

As can be seen by Fig. 2, the mean response time under PSJF is far lower than that under SJF, simply because the $\mathbf{E}[X^2]$ factor is gone. As expected, Fig. 3 shows that PSJF is far less sensitive to variability in the job size distribution than is SJF.

Shortest-Remaining-Processing-Time (SRPT)

The mean response time for a job of size x under SRPT is very similar to that under PSJF, as shown below:

$$\mathbf{E}[T(x)]^{\text{SRPT}} = \frac{\frac{\lambda}{2} \int_0^x t^2 f(t) dt + \frac{\lambda}{2} x^2 (1 - F(x))}{(1 - \rho(x))^2} + \int_0^x \frac{dt}{1 - \rho(t)} \quad (20)$$

The first term represents the initial waiting time for a job of size x . The only difference from PSJF is the additional term involving $x^2(1 - F(x))$. To get a rough feel for where this term comes from, observe that under SRPT, jobs of size $> x$ can still contribute to the waiting time of an arrival of size x , if those jobs of size $> x$ have remaining time $\leq x$ at the time job x arrives. Thus the $(1 - F(x))$ fraction of jobs of original size $> x$ can contribute at most x^2 to the second moment portion of the initial waiting time of job x .

The second term represents the residence time for a job of size x . This again resembles the residence time for PSJF, however it is actually a lot smaller. The point is that, as the remaining size of job x drops, the job has to compete with fewer and fewer interruptions (less load), because only those jobs with lower remaining size can bother it. Thus, the residence time for a job of size x can actually be viewed as a sum of busy periods, each started by a job of size dt , where each successive busy period fights against less and less load.

As expected from Ref. 11, Fig. 2 shows that the mean response time under SRPT is superior to all other policies. Also as seen in

Fig. 3, the response time actually decreases slightly with increased variability. Importantly, PSJF (with its lower initial waiting time and higher residence time) is a very close approximation to SRPT with respect to overall mean response time.

SCHEDULING IN PRACTICE

It is clear from Figs 2 and 3 that the choice of scheduling policy can dramatically affect mean response time. In situations where the job size variability is high, preemptive policies can be orders of magnitude more effective than nonpreemptive policies. When load and variability are high (e.g., $\rho = 0.9$ and $C^2 = 10$, as shown in Fig. 2), there is also significant differentiation among preemptive policies; for instance, SRPT offers more than a three-fold improvement over PS.

These dramatic improvements, at the cost of no additional resources, have not been lost on the computer systems community, which has incorporated smart scheduling in a wide variety of applications including web servers [20], routers [21], wireless networks [22], operating systems [23], and many more. Below we elaborate on just two such examples:

In a series of papers [20,24,25], Harchol-Balter *et al.* look at the problem of speeding up the mean response time in retrieving files at a web server. They find that the bottleneck resource for static “Get FILE” requests is the limited bandwidth in the web server’s access link. Traditionally, all requests share this bandwidth fairly in a manner resembling PS scheduling. The authors argue that scheduling requests in SRPT order would decrease mean response time, since file sizes at web sites are known to exhibit highly variable Pareto distributions. However, there are two potential stumbling blocks to implementing SRPT scheduling at web servers: first, one needs to know the “size” of the jobs. Second, there is a fear that by favoring short jobs, one might “starve” long jobs, or certainly treat them very unfairly. The authors deal with the first stumbling block by showing that the “size” of a job, in this case the time to retrieve a file, is well-approximated by the size of the

file, which is known. The authors deal with the starvation issue in Refs 26 and 27, where they prove a counterintuitive theorem, showing that, when job sizes follow a heavy-tailed distribution (like the Pareto or Bounded-Pareto), the expected response time of *every* job (including the largest-size job) is actually smaller under SRPT than under PS, provided load is not too high. Motivated by this “all-can-win” fairness result, the authors in Ref. 20 proceed to implement SRPT scheduling of HTTP requests by modifying the Linux kernel of an Apache web server to change the order that the server’s socket buffers are drained into the server’s access link, in accordance with SRPT scheduling. They demonstrate that this change improves mean response time by a factor of 5 for higher load, without penalizing requests for large files.

Similarly, Biersack *et al.* (see Ref. 8 and the references therein) consider flow scheduling at a bottleneck router, again with the goal of minimizing mean response time. However here the problem is complicated by the fact that the flow durations are not known *a priori*, so it is not possible to bias toward “short” flows (flows with few packets). Biersack *et al.* reason that since flow durations have a Pareto distribution (with DFR), favoring “young” flows is a good approximation to favoring “short” flows. They implement an FB-like policy which favors those flows which have transmitted the fewest packets so far, demonstrating an order of magnitude improvement over the traditional PS scheduling of flows. However, there are fairness issues with FB that are not present in SRPT, which they resolve by further tailoring of their FB policy.

OTHER TOPICS IN SCHEDULING AND REFERENCES

This document has offered a very brief introduction to scheduling in the M/G/1 queue. However, what is presented here is just the tip of the iceberg in terms of what is known about scheduling.

First of all, this document only considers mean response time. In fact, the full Laplace

transform of response time (and thus higher moments of response time) is known for all the policies discussed herein and is usually almost as easy to derive [13,14], although the high-level intuition is less easy to see from the transform. The tail behavior of response time has also been studied for many policies (see the survey in Ref. 28). For example, it is known that for heavy-tailed job size distributions, the response time tails for SRPT, PS, FB, and PLCFs all mimic the tail of the job size distribution, whereas the response time tails for FCFS and SJF mimic the tail of the excess of the job size distribution. Other performance metrics, like “fairness,” have also recently become very important in computer systems (see Refs 27 and 29 for an overview).

There are many scheduling policies which we do not have room to mention; in particular, there are many variants of PS that come up in networking (see the survey in Ref. 30). We also have not had room to delve into multiclass queues, where each class i of jobs has its own arrival rate λ_i and its own job size distribution, F_i , and jobs in one class might have preemptive, or nonpreemptive, priority over another class [15]. Also, while we have studied individual policies in this paper, there is a growing body of work that deals with broad classes of scheduling policies. In particular, the SMART class was introduced in Ref. 31 and defined so as to encompass a wide range of policies which favor short jobs.

Scheduling theory also exists for architectures far more general than the M/G/1 queue. Many computer systems follow a closed-loop architecture, where there is a fixed number of jobs, N , and a new job creation is only triggered by a job completion. The impact of scheduling in closed-loop systems has been shown to be far smaller than that in the open systems we have considered thus far [32]. By contrast, scheduling can have a huge impact in server farm architectures (multiserver systems); see Ref. 33, where scheduling decisions need to be made both at the central router level and at the individual host level. Lastly, fluctuations in load, for example alternations between very high and low load, can also greatly intensify the effect of scheduling, as seen in Refs 25 and 34.

Finally, while everything we have mentioned so far involves a stochastic setting, there is an entire field of on-line scheduling where the metric is worst-case response time and policies are ranked according to their “competitive ratio” when compared with the optimal policy (see the survey paper [35]).

REFERENCES

1. Harchol-Balter M, Downey A. Exploiting process lifetime distributions for dynamic load balancing. In: *Proceedings of ACM Sigmetrics*. Philadelphia (PA); 1996 May. pp. 13–24.
2. Leland WE, Ott TJ. Load-balancing heuristics and process behavior. In: *Proceedings of Performance and ACM Sigmetrics*. Raleigh (NC); 1986. pp. 54–69.
3. Crovella ME, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Trans Netw* 1997; 5(6):835–846.
4. Crovella ME, Taqqu MS, Bestavros A. Heavy-tailed probability distributions in the World Wide Web. *A Practical Guide To Heavy Tails*. New York: Chapman & Hall; 1998. Chapter 1. pp. 1–23.
5. Irlam G. Unix file size survey-1993. 1993 Sep. Available at <http://www.faqs.org/faqs/os-research/part2/section-12.html>.
6. Paxson V, Floyd S. Wide-area traffic: the failure of Poisson modeling. *Trans Netw* 1995;226–244.
7. Schroeder B, Harchol-Balter M. Evaluation of task assignment policies for supercomputing servers: the case for load unbalancing and fairness. *Cluster Comput J Netw Softw Tools Appl* 2004;7(2):151–161.
8. Biersack EW, Schroeder B, Urvoy-Keller G. Scheduling in practice. *Perform Eval Rev Spec Issue “New Perspectives in Scheduling”* 2007;34(4):21–28.
9. Kleinrock L. Queueing systems. Volume II, Computer applications. John Wiley & Sons; 1976.
10. Righter R, Shanthikumar J. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probab Eng Inf Sci* 1989;3:967–978.
11. Schrage LE. A proof of the optimality of the shortest remaining processing time discipline. *Oper Res* 1968;16:678–690.

12. Harchol-Balter M. Online lecture notes for performance modeling class at CMU. Available at <http://www.cs.cmu.edu/~harchol/Perfclass/class05.html/>. Accessed in 2005.
13. Wierman A. Scheduling for today's computer systems: bridging theory and practice [PhD thesis]. Carnegie Mellon University; 2007.
14. Conway RW, Maxwell WL, Miller LW. Theory of scheduling. Reading (MA): Addison-Wesley Publishing Company; 1967.
15. Takagi H. Queueing analysis—a foundation of performance evaluation. Volume 1, Vacation and priority systems, Part 1. North-Holland; 1991.
16. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1983.
17. Little JDC. A proof of the queueing formula $L = \lambda W$. Oper Res 1961;9:383–387.
18. Yashkov SF. Processor-sharing queues: some progress in analysis. Queueing Syst Theory Appl 1987;2:1–17.
19. Nuyens M, Wierman A. The foreground–background queue: a survey. Perform Eval 2008;65(3–4):286–307.
20. Harchol-Balter M, Schroeder B, Bansal N, *et al.* Size-based scheduling to improve web performance. ACM Trans Comput Syst 2003;21(2):207–233.
21. Rai IA, Urvoy-Keller G, Vernon M, *et al.* Performance modeling of las based scheduling in packet switched networks. In: ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems. New York; 2004.
22. Mangharam R, Demirhan M, Rajkumar R, *et al.* Size matters: Size-based scheduling for mpeg-4 over wireless channels. In: SPIE and ACM Proceedings in Multimedia Computing and Networking. San Jose (CA); 2004. pp. 110–122.
23. Feng H, Misra V, Rubenstein D. Pbs: a unified priority-based cpu scheduler. In: ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems. San Jose (CA); 2007.
24. Crovella M, Frangioso B, Harchol-Balter M. Connection scheduling in web servers. In: USENIX Symposium on Internet Technologies and Systems. Boulder (CO); 1999 Oct. pp. 243–254.
25. Schroeder B, Harchol-Balter M. Web servers under overload: How scheduling can help. ACM Trans Internet Technol 2006;6(1):20–52.
26. Bansal N, Harchol-Balter M. Analysis of SRPT scheduling: investigating unfairness. In: Proceedings of ACM Sigmetrics. Cambridge (MA); 2001.
27. Wierman A, Harchol-Balter M. Classifying scheduling policies with respect to unfairness in an M/GI/1. In: Proceedings of ACM Sigmetrics. San Diego (CA); 2003 June.
28. Boxma O, Zwart B. Tails in scheduling. Perform Eval Rev Spec Issue “New Perspectives in Scheduling” 2007;34(4):13–20.
29. Wierman A. Fairness and classifications. Perform Eval Rev Spec Issue “New Perspectives in Scheduling” 2007;34(4):4–12.
30. Aalto S, Ayesta U, Borst S, *et al.* Beyond processor sharing. Perform Eval Rev Spec Issue “New Perspectives in Scheduling” 2007;34(4):36–43.
31. Wierman A, Harchol-Balter M. Nearly insensitive bounds on SMART scheduling. In: ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems. Banff (Alberta); 2005.
32. Schroeder B, Wierman A, Harchol-Balter M. Closed versus open system models: a cautionary tale. In: Proceedings of Networked Systems Design and Implementation (NSDI). San Jose (CA); 2006.
33. Harchol-Balter M, Crovella M, Murta C. On choosing a task assignment policy for a distributed server system. IEEE J Parallel Distrib Comput 1999;59:204–228.
34. Gupta V, Harchol-Balter M, Scheller-Wolf A, *et al.* Fundamental characteristics of queues with fluctuating load. In: ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems. Saint Malo; 2006.
35. Pruhs K. Competitive online scheduling for server systems. Perform Eval Rev Spec Issue “New Perspectives in Scheduling” 2007;34(4):52–58.

QUEUEING NOTATION

ROBERT B. COOPER
Department of Computer
Science and Engineering,
Florida Atlantic University,
Boca Raton, Florida

Queueing theory is concerned with the mathematical analysis of systems that provide service to random demands. This theory has many application areas (e.g., industrial engineering, electrical engineering, computer science, telecommunications, and operations management). Consequently, the terminology and notation, while often intuitively sensible, are not always consistent. The mathematical theory is focused on simple (but meaningful) mathematical models that can be described in the precise terminology required for mathematical analysis. In this article, we give an overview of the notation and terminology used in describing the standard basic queueing models.

In the basic queueing model, “customers” arrive (to request service), wait (if necessary) for a “server” to become available to provide the required service, and then leave. Thus the model consists of three components: (i) the (stochastic) arrival process, (ii) the (stochastic) service requirement, and (iii) the physical configuration of the servers and their operating rules. The objective of the theory is to understand the relationship between these components and the behavior (performance measures) of the system.

In a now-classic 1953 paper, D.G. Kendall [1] proposed the following notation to describe a queueing model: $a/b/c$, where a describes the input (arrival) process, b describes the service process, and c is the number of servers. Implicit in this description is the assumption that each arriving customer is assigned to a server immediately if one is available and holds that server for the length of time required (the “service time”); if a server is not immediately available, the customer waits

in an infinite-capacity queue (or waiting room). Furthermore, the servers are all identical, and the waiting customers are served according to the FIFO (first in, first out) “queue discipline” (although the queue discipline might have no effect on the values of any particular performance measure, and therefore its omission from the explicit notation might be irrelevant).

Kendall’s notation is now widely used to describe queueing models. But, clearly, there are an infinite number of possible physical configurations, operating protocols, queue disciplines, and so on, with the consequence that sometimes authors extend this notation to include those variations; that is, they treat Kendall’s scheme as an algorithm that can be decoded to describe a particular model, rather than as a descriptive name of that model. For example, $a/b/c/n$ might be used to indicate that the model has a capacity of n customers, or that there are n different sources that generate the arrivals (see *Finite Population Models—Single Station Queues*), and so on. It has become commonplace to extend Kendall’s original three-symbol description to longer strings, such as $a/b/c/d/e$, where d is some measure of system capacity and e is the queue discipline. (However, sometimes even longer strings are used; clearly, this can get out of hand.) Besides FIFO, other queue disciplines include LIFO (last in, first out), LIFO-PR (last in, first out, preemptive-resume), SIRO (service in random order), SPTF (shortest processing time first), SRPT (shortest remaining processing time first), RR (round robin), PS (processor sharing), VAC (server vacations), and others. If, in any particular case, one wants to introduce Kendall-like notation to describe a particular model, then the shorthand notation should be described upfront, rather than assuming that the reader can decode its meaning; similarly, the reader should be careful in interpreting the meaning of any such notation without first ascertaining exactly what the model is. With this caveat, we describe Kendall’s notation.

Here are the conventions:

G	general (no particular assumption)
GI	general independent (the random variables in question are mutually independent and identically distributed)
M	Markov or memoryless (exponential random variables)
D	deterministic (the same constant value for each realization of the random variable)
E_k	k -phase Erlangian (sum of k independent, identical, exponentially distributed “phases”)
PH	phase-type (sum of a (possibly random) number of independent, exponentially distributed phases)
MAP	Markovian arrival process
$BMAP$	batch Markovian arrival process

Then, for example, a typical (and most important) model is $M/G/1$ (see **The $M/G/1$ Queue**). This describes a model with Poisson input (M denotes independent, identically distributed, exponential interarrival times), General service times, and 1 server. It is usually assumed implicitly that the service times are independent of each other (hence, $M/GI/1$ would be more appropriate, but is rarely used, although renewal-process input is often denoted explicitly, as in $GI/M/1$), and independent of the arrival process. It is also implicitly assumed here that there is infinite queueing capacity (so that no customer is ever “cleared” from the system because of lack of waiting space) and that the customers are served in FIFO order (although, depending on the performance measure of interest, the queue discipline might have no effect). The fact that the service times here are General means that the formulas that describe the behavior of $M/G/1$ are to be evaluated using the distribution function of the service times as if it were a parameter. (Thus, $M/M/1$ and $M/D/1$ are important special cases.) For $M/G/1$, the celebrated Pollaczek–Khintchine formula relates the mean waiting time to the utilization of the server (typically denoted by ρ) and to the

mean and variance of the distribution of service times.

Another important example is $M/G/s/s$ (see **The $M/G/s/s$ Queue**). Here, the second s specifies that the capacity of the system is s , which is the same as the number of servers; that is, in this model any customer who finds all servers busy does not wait, but is immediately cleared from the system. In a sense, this is not a queueing model, because there is no queue (unless one considers the customers in service to be part of the queue, which is a common convention). Some representations for the case of Poisson input and $n > 0$ waiting positions might be $M/G/s/n$, $M/G/s/s + n$, or $M/G/s + n$, all indicating s servers and n waiting positions, for a total system capacity of $s + n$, and, in the infinite-capacity case, the default notation $M/G/s$. Another example is $M^x/G/1$, say, in which the customers arrive in batches of size X (which may be a random variable), and the arrival times of the batches follow a Poisson process (see **Batch Arrivals and Service—Single Station Queues**). The point is that Kendall’s notation provides a convenient shorthand for describing queueing models, but the reader should be alert to avoid incorrect interpretations. With this background, most of the basic queueing models are now completely specified.

The notations PH , MAP , and $BMAP$ refer to generalizations of the method of phases, where the key idea is to model random time intervals as composed of a (possibly random) number of exponentially distributed phases and then to exploit the simplifications arising from the resulting Markovian structure (see **Matrix Analytic Method: Overview and History**).

The important model $M/G/s/s$ (s servers and 0 waiting positions) is often called the *Erlang B model* or the *Erlang loss model* (because customers who find all servers busy are cleared from the system and thus are lost). The formula that describes the probability of loss (the fraction of arriving customers who are cleared from the system) is typically called the *Erlang B formula* or *Erlang’s first formula*. Likewise, the important model $M/M/s$ (or, $M/M/s/\infty$, s servers and infinite waiting capacity) is typically called the *Erlang C model* or the *Erlang delay model*,

and the formula that describes the probability that an arriving customer will be forced to wait until a server becomes available is often called the *Erlang C formula* or *Erlang's second formula* (see **The M/M/s Queue**). The Erlang B and Erlang C formulas were first published in 1917 by A.K. Erlang. The context was telephone traffic (teletraffic) theory and engineering, and the terminology reflects this. The "B" in Erlang B refers to Blocking; a call is blocked if it finds all servers busy (but beware, there is not a universally accepted definition of "blocked"), in which case "lost calls are cleared." In the early teletraffic literature, the $M/G/\infty$ model is called *lost calls held*; this can be interpreted to say that the calls leave the system at the same rate whether they are in service or are waiting in the queue. (Aficionados should see Brockmeyer *et al.* [2] for reprints of Erlang's papers and interesting historical and biographical material. Syski [3] gives an invaluable authoritative and comprehensive treatment of teletraffic theory prior to 1960.) This model is now often called *Erlang A* (for Abandonment), generalizations of which can be used to describe systems (like the ubiquitous call center) in which customers renege or depart from the queue before entering service. In this context, the Kendall notation $a/b/s + M$, for example, would indicate that there are s servers and an infinite-capacity waiting room from which customers abandon after waiting an exponentially (M) distributed length of time (but beware, this could be interpreted to describe an s -server system with a waiting room of capacity M).

Two of the most important theorems in queueing theory are Little's Law [4] (see **Little's Law and Related Results**) and PASTA [5] (see **PASTA and Related Results**). Little's Law expresses the equality $L = \lambda W$, where L stands for expected number of customers in the "system," W stands for expected time spent by a customer in the system, and λ is the rate at which the customers enter the system. Originally, L denoted queue Length, W denoted Waiting time in queue, and λ denoted the customer arrival rate, but these

definitions are very flexible as long as one is consistent in the definition of "system." Little's Law is notable for its great generality; all that is required of L , λ , and W is, essentially, their existence.

PASTA is an acronym for Poisson Arrivals See Time Averages. In essence, it expresses the fact that if the customers arrive according to a Poisson process, then the states of the system that they observe at (just prior to) their arrival epochs (in the queueing theory vernacular, "epoch" means "instant") has the same distribution as the states observed at random epochs or averaged over time. PASTA is notable for its utility in solving queueing models (one can shift one's viewpoint as is convenient) and for its counterintuitive subtlety.

In summary, queueing theory is vast and, because of its many various application areas, its notation, while usually intuitive, is not always consistent across application areas. Finally, it is interesting and amusing to observe that almost all books and most technical journals use the spelling "queueing," which makes it the only (ordinary) word in the English language with five successive vowels, whereas virtually all spell-checkers and nontechnical editors (in the United States) omit the second E.

REFERENCES

1. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann Math Stat* 1953;24(3):338–354.
2. Brockmeyer E, Halstrøm HL, Jensen A. The life and works of A.K. Erlang. *Transactions 2. Copenhagen: Danish Academy of Technical Sciences; 1948. pp. 1–277.*
3. Syski R. Introduction to congestion theory in telephone systems. 2nd ed. Amsterdam: Elsevier Science Publishers B.V. 1986 (1st ed. Edinburgh: Oliver & Boyd; 1960.)
4. Little JDC. A proof for the queueing formula: $L = \lambda W$. *Oper Res* 1961;9(3):383–387.
5. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30(2):223–231.

R&D RISK MANAGEMENT

JAMES C. FELLI
Drug Disposition, Eli Lilly & Company,
Indianapolis, Indiana

JOHN S. ANDERSEN
Decision Sciences, Eli Lilly & Company,
Indianapolis, Indiana

RISK

Risk is a major concern in functional R&D organizations. It can arise from any number of sources, such as from accidents to misunderstandings to natural disasters. Yet, despite its importance and pervasiveness, it is no small feat to find a common understanding of “risk” within or across a firm’s R&D functions, let alone across industries.

The general concepts of risk and peoples’ attitudes toward risk have long held prominent places in game theory, decision analysis, and economics, and have been scrutinized using a wide range of methods from puzzles and games to prospect and utility theory [1–5]. It should come as no surprise that these concepts have attracted particular attention in the financial arena. In this setting, Holton provides an interesting discussion of the foundations of risk by exploring the relationship between risk, uncertainty, and exposure [6]. Of particular interest in business settings—and especially in R&D—are risks derived from perceptual and motivational biases, most notably optimism and personal desire [7–10]. Also noteworthy are project-related risks assumed either explicitly or implicitly as a consequence of forecast inaccuracies arising from technical, psychological, and political–economic interpretations [11].

Despite a rich history of exploration and scrutiny, the word “risk” has managed to maintain a slippery meaning. The International Organization for Standardization (ISO) has published standards on terminology and guidelines for risk management,

defining “risk” to be “the effect of uncertainty on objectives.” Despite this attempt at standardization, the term is persistently used by different people at different times to mean different things [12,13]. To a pharmaceutical discovery scientist, a “risky project” may be one with low technical feasibility. To an aerospace project manager, it may be a project that is likely to accrue large cost overruns on its way to certain completion. A “risky project” to a toy company might be one that is readily designed, developed, and manufactured but faces questionable potential for commercial success. The scientist, the manager and the executive are all articulate concern about uncertainty consistent with the high-level ISO definition, but focus on different types of uncertainty with different, predominantly negative consequences.

To an R&D organization, risk is about uncertainty and unwantedness. It is more than simply a measure of the likelihood of the obtainment of some unwanted event; it also carries with it some consideration of the attendant consequences of that event. For early stage R&D projects, tasks are undertaken to demonstrate feasibility. The obvious unwanted outcome is failure to do so, and the attendant consequence is that the money spent on the project is no longer available to be spent on other projects. As a project advances through the R&D process, from concept to product, the factors contributing to risk become more varied as the consequences become more numerous. A product may pass a critical technical success hurdle only to run the “risk of missing a launch deadline.” Or a product may clear its launch hurdle, but run the “risk of failing in the marketplace.”

Given such diversity and dynamics, it is easy to see why the cogent characterization and discussion of risk presents such great difficulty to R&D organizations. As a consequence, these organizations typically respond by implementing programs designed to highlight and dynamically manage a portfolio of risks throughout the R&D process stream.

RISK MANAGEMENT

Risk management is the general term for the collection of formal processes employed by R&D organizations to understand and mitigate risks during the development process. The fundamental steps involved include characterization of risk factors, monitoring of risk factors, and mitigation of risk factors. The risk analysis community generally considers the process of risk analysis to include both risk assessment and management.

A risk factor is anything that might proximately affect or directly cause the occurrence of an unwanted event, such as a delay in a development timeline. In terms of the ISO definition, a risk factor may be considered to be any uncertain event that might directly contribute to a negative “effect on objectives.” [13] The distinction between risks and risk factors is one of scope. For example, given a focus on minimizing the risk of a budget overrun, a project manager would consider all relevant factors that could reasonably affect the likelihood and/or magnitude of a budget overrun. Each of these factors, in turn, may be a fundamental risk from another perspective, with its own set of constituent factors. The distinction between risks and risk factors helps functional managers set operational objectives against those risks for which they will be held accountable and appropriately allocate resources to mitigate the underlying factors within their sphere of influence or control. Once identified, relevant risk factors can be prioritized to inform additional resource allocation decisions for their monitoring and mitigation. Given a well-defined risk factor with sufficient lead time to respond, most R&D organizations are able to enact robust sets of activities to assuage the consequences of its realization. In addition, an organization may engage in separate activities to influence either the likelihood or the consequences of specific risk factors.

Conventional risk management approaches tend to separate risk into schedule, cost, and performance risk [14], typically defining risk as a measure of the likelihood of not meeting a critical timeline, exceeding a fixed budget, or falling short of a performance benchmark. Several operational research

methods and models are commonly applied to characterize the consequences of specific sets of risk factors. These include, but are not limited to traditional and stochastic PERT/CPM [15] and critical chain [16] to quantify the likelihood and extent of project cost and timeline overruns; queueing models to investigate risk factors derived from the flow of information and materials [17]; inventory models to understand supply chain risks consequent to product and material stockouts [18]; mathematical and dynamic programming to optimize designs and processes [19]; and simulation modeling to investigate system and market behavior due to intrinsic and extrinsic uncertainties [20]. While a plethora of stand-alone or add-in software packages exist to help managers unleash the power of these models as part of their risk management arsenal, R&D organizations must also appropriately invest in personnel development and training to properly interpret and leverage their use.

Despite the sophisticated analytical tools available, many R&D organizations struggle with risk assessment and management. Their fundamental challenge is their inability to appropriately identify, characterize, and monitor relevant risk factors. This challenge is compounded when an organization suspects there are risk factors that it is unable to articulate, either due to lack of proximity, understanding, or experience. For example, an R&D organization may deal with a new partner to complete some work and may not have sufficient knowledge of the partner to appropriately gauge how the partner’s risk factors might translate into risks for the organization. Of particular consequence are risk factors that derive from cultural differences.

The characterization of a risk factor is typically accomplished by consideration of the various events that could cause it to manifest. The R&D organization must carefully define the risk factor to appropriately estimate its likelihood of manifestation and the consequences faced by the organization given its manifestation. Specifically, what exactly is the unwanted event, and how unwanted is it? Framing effect and motivation biases often interfere with these

tasks, as do the evaluation heuristics of availability and anchoring-and-adjustment [21]. These limitations to rationality may cause some risk factors to be ignored and others to be inappropriately characterized in terms of likelihood of occurrence or magnitude of effect. The combined effects of availability and anchoring-and-adjusting can be particularly troublesome in cases where one not only inadvertently bases one's assessment of an event's likelihood on memory rather than complete data (availability), but also inadequately captures the range of the event's potential consequences due to insufficient adjustment from its baseline outcome (anchoring-and-adjustment).

Some risk factors manifest slowly over time and may be preceded by clear signals (e.g., changes in the political environment) whereas others may be sudden and unexpected (e.g., a destructive fire at a supplier's manufacturing facility). Monitoring the set of identified risk factors costs the R&D organization time and money that cannot be invested in developing products. What should the organization monitor? How closely? If data are collected, how is the collection accomplished? With what frequency are the data refreshed? Who holds the data within the organization? These are all questions with potentially profound resource implications.

The management of risks in an R&D organization is typically accomplished through the creation and conditional execution of formal risk mitigation plans. Risk mitigation plans are typically triggered by a predefined event (e.g., a report of a risk factor obtaining). These plans define the organization's response to specific events and detail resource expenditures and actions of employees to lessen or counteract the consequences of the realized risk factor. Some industrial practices and/or policies require firms to maintain risk mitigation plans. Some firms have learned the hard way that risk mitigation plans are very valuable to mitigate well-understood risks in risky enterprises.

Risk mitigation plans require an organization to explicitly consider four things: risk factor characterization, risk factor likelihood,

risk factor consequence, and consequence alleviation or resolution.

RISK CHARACTERIZATION

While it seems obvious that risk factors should be understood in terms of both likelihood and consequence, it is less obvious but critically important to understand the context in their characterization. When one R&D manager describes a project as "risky," what exactly does the manager mean? Many problems in an R&D organization arise from people agreeing that a project is "risky" and then walking away with different action plans due to different perceptions of what "risky" actually means to them.

The general idea of risk can be decomposed into general contextual categories. An example is presented in Fig. 1, where the categories are technical risk, commercial risk, and execution risk. Execution risk is further decomposed into timeline risk and cost risk.

Technical Risk. This encapsulates the traditional idea of "probability of technical success" in terms of meeting a specific set of goals (e.g., manufacturing feasibility, efficacy and safety expectations, reliability commitments). The fundamental question is "can we do this project and accomplish these ends?" This is frequently presented as a binary outcome event to decision makers: yes or no, success or failure.

Commercial Risk. This presumes that the technical risk has been resolved and that

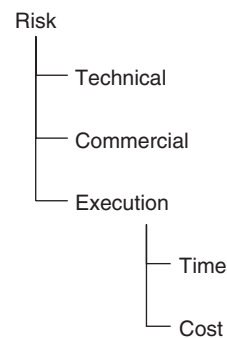


Figure 1. A composite model of risk for R&D management.

development can be accomplished. The focus here is on the subsequent commercial impact of the project or product. Will the product meet the market needs? How well will it do upon commercialization? This is most often addressed via scenario planning using financial metrics such as net present value (NPV).

Execution Risk. Execution risk focuses on the consequences of uncertainty in the successful completion of a project in accordance with its formal project plan. Given a project plan, the two principal metrics of interest to project managers and R&D executives are compliance with budget and adherence to timeline. Both of these are critical to R&D organizations pursuing multiple, simultaneous projects. Money spent on one project cannot be spent on another, while timelines drive allocations of resources, people and material. Where technical risk asks whether a project can be successfully completed, execution risk asks whether success can be secured on time and on budget.

This definition of risk as a composite measure of more well-defined, context-specific risks facilitates understanding and communication across R&D functions. Clear understanding flows from precise definitions of specific risk factors within the context of each category. This decomposition of risk facilitates clearer understanding of and metrics for the monitoring and characterization of risk factors. It additionally facilitates more focused and productive discussion among typically nontechnical managers and senior decision makers.

The risk decomposition illustrated in Fig. 1 is appropriate for a commercial R&D organization. Different categories may be appropriate for noncommercial organizations. For example, a military R&D organization may find it beneficial to decompose risk into technical, interoperability, and execution risks. A governmental agency may consider political risk to be an appropriate category.

Even within these risk components, we must carefully define what we mean by “risk.” It is here that we consider, in each context, what underlying elements are both uncertain and unwanted. Note that neither attribute

alone is sufficient. Pure association of risk with probability ignores consequence; pure association of risk with consequence ignores likelihood. To truly appreciate risk, we must consider both: how likely is an unwanted event to obtain, and what are the consequences if it does?

PROBABILITY AND CONSEQUENCE

For discussion purposes, Fig. 2 is a useful construct for the joint consideration of probability and consequence. Decision makers typically consider situations to present “low risk” if they are characterized by a low likelihood of occurrence and small potential consequence. Similarly, situations bearing a high chance of occurrence and great potential consequence are typically perceived as “high risk.” “Low-risk” situations typically receive low levels mitigation planning, whereas “high-risk” situations generally enjoy rich, detailed mitigation planning. In practice, likelihood-consequence tables akin to that in Fig. 2, but with more refinement in each dimension, are widely used to screen risks.

In a resource constrained environment, loosely regulated firms may opt to categorize situations of high probability and low consequence as “low risk” and treat them accordingly. Regulated firms, however, do not generally enjoy such discretion. For regulated firms, events with a high likelihood of obtainment and/or the potential for severe consequences (e.g., drug development, nuclear

Potential consequence	High	?	High risk
	Low	Low risk	?
		Low	High
		Probability of obtainment	

Figure 2. Example of a likelihood-consequence table to define risk.

reactor construction) typically require formal mitigation plans to assure regulatory compliance at prescribed points throughout a product's development and life cycle. In most cases, events characterized by low likelihood and great consequence require, at a minimum, thoughtful analysis. In such cases, it is important to understand the distribution of the potential consequence.

A PROJECT-LEVEL EXAMPLE: COST RISK

With today's readily available desktop simulation tools, R&D managers have the ability to simulate various project plans to generate plausible distributions for time and cost. While this effort requires time and cost distributions on component activities, it also affords the manager the opportunity to bring together different stakeholders for formal and structured discussion. This activity alone can often reveal important insights critical for appropriate modeling of uncertainty.

Consider the case of a manager who must set a budget for a project with the total cost density given in Fig. 3, where $C_{\min}$, $E[C]$, and $C_{\max}$ represent the project's minimum, expected, and maximum cost. The seemingly obvious solution is to set the budget at the project's expected cost, $E[C]$. In doing so, however, the manager must accept a significant probability of a budget overrun in the range of 0 to $C_{\max} - E[C]$. If the manager elects to set the budget at a higher level, C^* , for instance, both the likelihood and range of a budget overrun decrease, but the marginal budget increase $C^* - E[C]$ now represents funding no longer available to other projects.

To put some concrete numbers on this example, suppose that $f(C) \sim \text{Beta}(2,6)$ over $C_{\min} = \$100\text{M}$ and $C_{\max} = \$300\text{M}$. Given this density, $E[C] = \$250\text{M}$. If the manager sets the project budget at $E[C]$, there is a 56% chance that the project will go over budget. In fact, the expected budget overrun is \$11.75 M. If the manager sets the budget at $C^* = \$258\text{M}$, the chance of a budget overrun falls to 45% and the expected budget overrun drops to \$7.65 M. Clearly, setting the project budget equal to its expected cost presents a greater "risk" of a budget overrun, but is it worth an additional \$8 M increase to reduce the expected budget overrun by \$4.1 M? What else could have been done with \$8 M? These are the critical questions for management.

RISK AND OPPORTUNITY

Risk and opportunity are matters of perspective. Consider an investment with an expected return of 10%. The possibility of a return in excess of 10% is typically perceived as an opportunity for additional gain, whereas the possibility of a return less than 10% is typically perceived as a risk of insufficient gain. The intrinsic uncertainty in the investment's return is what it is; however, the potential investor's perspective, rooted in reference points (e.g., 10%) and objectives (e.g., maximize return), will ultimately determine whether it represents a risk to be avoided or an opportunity to be embraced. G. K. Chesterton summed this duality up nicely:

'An adventure is only an inconvenience rightly considered. An inconvenience is only an adventure wrongly considered.'

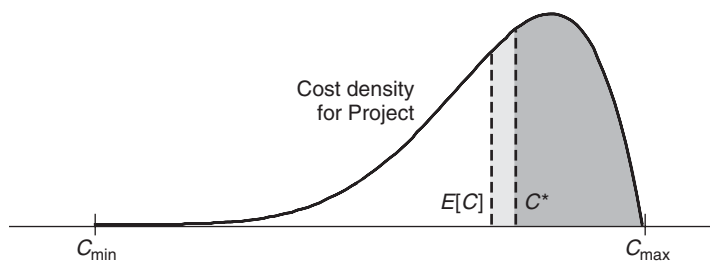


Figure 3. Density of total project cost for a hypothetical R&D project.

Given an appropriate scoring system, any R&D endeavor can be scored against a scale partitioned into two regions: “unwanted” and “wanted.” In monetary terms, we typically equate “losses” with “unwanted” and “gains” with “wanted.” Let the reference point r designate the boundary between “unwanted” and “wanted” values on this scale, and define the functions g and f to represent the cumulative and inverse cumulative distributions of the uncertain payoff a for a hypothetical project A. That is, for some benchmark α , $g(\alpha) = P[a < \alpha]$ and $f(\alpha) = P[a > \alpha]$. We can now calculate the expected loss and expected opportunity of project A. Define the quantities $a_0 = E[a|“unwanted”]$ and $a^0 = E[a|“wanted”]$. The “expected unwantedness” of project A is $a_0 g(r)$ and the “expected wantedness” of project A is $a^0 f(r)$. This “expected unwantedness” takes into account both likelihood and consequence and makes for a natural measure of “risk” in terms of the underlying scale. Similarly, “expected wantedness” can be thought of as an expression of “opportunity.”

INTEGRATED PROJECT RISK

From a project perspective, an integrated view of the risks noted in Fig. 1 can be of great value to a decision maker. One such

view for a hypothetical R&D project is provided in Fig. 4.

The number in the upper right corner of Fig. 4 quantifies the technical uncertainty surrounding the project. At the time of figure creation, there was a 65% chance that the project would be successfully developed and deployed. Successful development presumes that the project will not only be completed, but will also meet a specific set of performance targets upon completion (e.g., safety, reliability). The heavy line characterizes the project timeline uncertainty given successful completion. The choice of using the cumulative versus inverse cumulative distribution over launch date is matter of taste. We have found it more useful to frame timeline risk discussions with managers in terms of failing to launch by a given date as opposed to launching some time prior to a given date, and have therefore used an inverse cumulative distribution in Fig. 4. The small distributions over various parts of the inverse distribution in Fig. 4 are the distributions of the NPV at specific launch dates. The launch date and expected NPV are noted to the right of each distribution; the NPV range is noted below the distribution. The expected NPV is also indicated by a point on the axis. For example, given technical success and a product launch in March 2014, the expected NPV

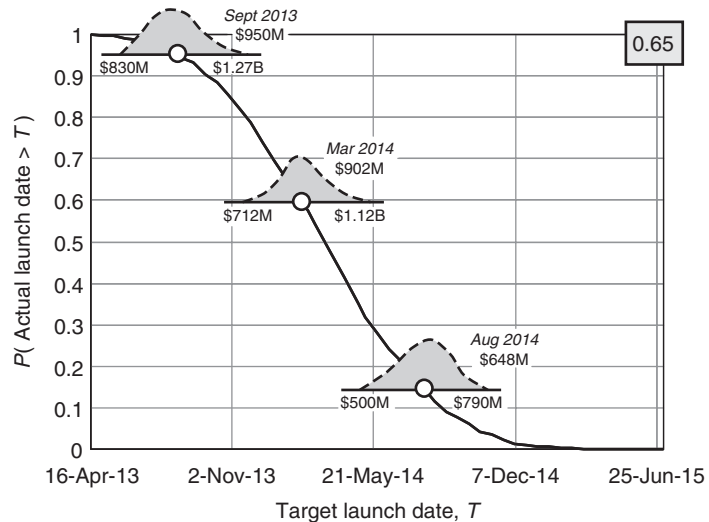


Figure 4. An example of an integrated risk view for a hypothetical R&D project.

of the project is \$902 M with a range of \$712 M to \$1.12B. Note the large drop in NPV for an August 2014 launch date. This may represent a situation in which a competitor has launched a rival product between March 2014 and August 2014.

This integrated risk picture provides information on three key aspects of project risk:

1. the likelihood that the project will be successfully completed,
2. the likelihood that a specific launch date will be met, and
3. the net monetary consequences associated with specific launch dates.

Taken together, these insights help R&D managers consider changes in plans to affect costs, timelines, and probabilities of technical success, either through or in addition to formal risk mitigation plans.

One may be tempted at this point to combine these aspects of project risk into a single distribution over project NPV. Although of great potential value to project and portfolio managers alike, such a construct would require a staggering amount of effort to produce. Most organizations partition their workforce into specialized functions (e.g., marketing, finance, project management); consequently, the information required to generate a single NPV distribution is typically fragmented across several functions. Collecting and integrating that information into a meaningful distribution requires functional cooperation, political alignment, and reconciling both data formatting diversity and modeling discrepancies. In addition, the cost and revenue streams associated with a new product launch typically exhibit sensitivity to time frames: the NPV of a product launch delayed by two years is rarely equal to the NPV calculated by simply sliding the cost and revenue streams out two years.

PORTFOLIO-LEVEL RISK

Careful consideration of project-level risk across all projects in an R&D pipeline can provide insight into portfolio-level risks not

typically appreciated on a project level. This is a “bottom-up” approach to portfolio risk. Each project’s risk is characterized by the process outlined above and common risks across several projects are identified. For example, suppose that several project teams have identified timeline uncertainty due to availability of material from a common, key, and overburdened manufacturing facility. If this facility runs into production difficulties, several projects will be affected, resulting in consequences across the portfolio. By taking a thorough and thoughtful approach to risk characterization at the project level, the portfolio manager will be able to identify mitigation plans whose implementation has the broadest impact across the portfolio.

A “top-down” approach to portfolio risk can provide a complimentary tool to the “bottom-up” approach. In this process, the portfolio is examined from a strategic perspective to identify patterns or trends in its composition. Consider the example of a pharmaceutical R&D organization. What fraction of the portfolio, and what fraction of the R&D budget, is devoted to projects that are targeting a particular disease state? Suppose that an important regulatory agency issues new guidelines for clinical drug development in this disease state. These new regulations may stipulate that clinical trials in this area must now provide patient exposure data for longer periods of time to better understand long-term safety of drug candidates. As a consequence, both expenses and timelines for projects in this area will be increased. An R&D portfolio that is heavily invested in this disease state is at higher risk for this type of change.

Note that the “top-down” approach has the ability to pick up risk elements that may have been overlooked by some of the project teams. Together, these two approaches both enable and allow for a comprehensive view of portfolio-level risk. While delving deeper into the mechanics and theory of portfolio management is beyond the scope of this work, we recommend Matheson [22] and Cooper *et al.* [23] as excellent departure points for further exploration into this rich area of study.

CONCLUSIONS

Risk management constitutes an important set of activities for managers at all levels in an R&D organization. To make these activities worthwhile, managers must first come to grips with the ambiguity of the term *risk* and recognize that it can apply to one of several uncertainties related to a project's development and commercialization. Thoughtful and careful identification and characterization of these risks can lead to greater insight in the appropriate development of risk mitigation plans at both the project and portfolio level. Well-articulated risk management plans can then effectively inform an organization's risk communication plan to guide internal and external messages to the project participants, portfolio managers, and outside stakeholders.

REFERENCES

1. von Neumann J, Morgenstern O. Theory of games and economic behavior. 1953 ed. Princeton (NJ): Princeton University Press; 1944.
2. Friedman M, Savage LP. The utility analysis of choices involving risk. *J Polit Econ* 1948;56:279–304.
3. Kreps DM. Notes on the theory of choice. Boulder (CO): Westview Press; 1988.
4. Clemen RT, Reilly T. Making hard decisions with decision tools. Pacific Grove (CA): Duxbury Press; 2001.
5. Goodwin P, Wright G. Decision analysis for management judgment. 3rd ed. England: John Wiley & Sons, Ltd.; 2004.
6. Holton GA. Defining risk. *Financ Anal J* 2004;60(6):19–25.
7. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica* 1979;47:313–327.
8. Lovall D, Kahneman D. Delusions of success: how optimism undermines executives' decisions. *Harv Bus Rev* 2003;81:56–63.
9. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47:263–292.
10. Gilovich T, Griffin D, Kahneman D, editors. Heuristics and biases: the psychology of intuitive judgment. New York: Cambridge University Press; 2002.
11. Flyvbjerg B. From nobel prize to project management: getting risks right. *Proj Manage J* 2006;37(3):5–15.
12. ISO/IEC Guide 73:2002. Risk management—Vocabulary—Guidelines for use in standards. International Organization for Standardization; 2002.
13. ISO/DIS 31000. Risk management—Principles and guidelines on implementation. International Organization for Standardization; 2009.
14. Galway L. RAND Corporation Report WR-112-RC; 2004.
15. Project Management Institute. A guide to the project management body of knowledge. 3rd ed. Drexel Hill: Project Management Institute; 2003. ISBN 1-930699-45-8.
16. Steyn H. An investigation into the fundamentals of critical chain project scheduling. *Int J Proj Manage* 2000;19:363–369.
17. Gross D, Shortle JF, Thompson JM, *et al.* Fundamentals of queueing theory. 4th ed. New York: Wiley; 2008. ISBN 978-0-471-79127-0.
18. Cheung KL. A risk-return framework for inventory management. *Supply Chain Manage Rev* 2002;6:50–55.
19. Beale EML, Mackley L. Introduction to optimization. New York: John Wiley & Sons, Inc.; 1988.
20. Evans JR, David LO. Introduction to simulation and risk analysis. Upper Saddle River (NJ): Prentice-Hall; 1998.
21. Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: heuristics and biases. New York: Cambridge University Press; 1982.
22. Matheson JE. Managing the corporate business portfolio. In: Howard RA, Matheson JE, editors. The principles and applications of decision analysis. Palo Alto (CA): SDG; 1983.
23. Cooper RG, Edgett SJ, Kleinschmidt EJ. Portfolio management for new products. Reading (MA): Addison Wesley; 1998.

RANDOM SEARCH ALGORITHMS

ZELDA B. ZABINSKY

Department of Industrial and Systems
Engineering, University of Washington,
Seattle, Washington

A random search algorithm refers to an algorithm that incorporates some kind of randomness or probability (typically in the form of a pseudorandom number generator) in its methodology. The random element may be introduced through sampling specifications of the algorithm, or through noise in the function observation. The random sampling method in the algorithm may include a random search direction and a random step size, as is common for continuous problems, or a random permutation of variables, such as cities in the traveling salesperson problem (TSP). Randomness may also be included when the objective function is only available through noisy measurements. In the literature, a random search algorithm may be called a *Monte Carlo method* or a *stochastic algorithm*. The term *metaheuristic* is also commonly associated with random search algorithms. Simulated annealing, tabu search, genetic algorithms (GAs), evolutionary programming, particle swarm optimization (PSO), ant colony optimization, cross-entropy, stochastic approximation, multistart, clustering algorithms, and other random search methods are being widely applied to continuous and discrete global optimization problems [1–8]. Random search algorithms are useful for ill-structured global optimization problems, where the objective function may be nonconvex, multimodal, nondifferentiable, and possibly discontinuous over a continuous, discrete, or mixed continuous–discrete domain. Random search algorithms have been applied to many ill-structured optimization problems that arise in complex systems including engineering design problems, scheduling and sequencing problems, and other applications in biological and economic systems.

The problem of designing algorithms that obtain global optimal solutions is very difficult when there is no overriding structure that indicates whether a local solution is indeed a global solution. In contrast to deterministic methods such as branch and bound, interval analysis, and tunneling methods [4,9,10], which typically guarantee asymptotic convergence to the optimum, random search algorithms ensure convergence in probability. The trade-off is in terms of computational effort. Random search algorithms are popular because they can provide a relatively good solution quickly and easily.

Random search methods have been shown to have a potential to solve large-scale problems efficiently in a way that is not possible for deterministic algorithms. Whereas it is known that a deterministic method for global optimization is NP-hard [11], there is evidence that a stochastic algorithm can be executed in polynomial time, on the average. For instance, Dyer and Frieze [12,13] showed that estimating the volume of a convex body takes an exponential number of function evaluations for *any* deterministic algorithm, but if one is willing to accept a weaker claim of being correct with a high probability, then a stochastic algorithm can provide such an estimate in polynomial time.

Another advantage of random search methods is that they are relatively easy to implement on complex problems with black-box function evaluations. Because the methods typically only rely on function evaluations, rather than gradient and Hessian information, they can be coded quickly, and applied to a broad class of global optimization problems. A disadvantage of these methods is that they are currently customized to each specific problem largely through trial and error. A common experience is that random search algorithms perform well and are “robust” in the sense that they give useful information quickly for ill-structured global optimization problems.

The term *stochastic optimization*, as used in Ref. 6, includes problems with noisy

measurements in the objective functions in addition to algorithms that incorporate a random element. Here we focus on algorithms that incorporate some form of randomness in their definition, even though they are applied to deterministic or noisy functions. Whether the function evaluations are deterministic or noisy, the function values are typically evaluated through a complicated computer program and/or Monte Carlo simulation and have no known structure. In contrast, well-structured noisy problems such as stochastic linear programs with recourse have specialized methods.

Different types of categorizations have been suggested for random search algorithms. Schoen [14] provides a classification of a two-phase method, where one phase is a global search phase and the other is a local search phase. The global phase is also referred to as an *exploration phase* aimed at exploring the entire feasible region, while the local phase is referred to as an *exploitation phase* aimed at exploiting local information. One example of a two-phase method is multistart, where the global phase is typically a uniform random search and the local phase is a deterministic gradient search. Another example is GAs, where the global or exploration phase is associated with the mutation operation and the local or exploitation phase is the crossover operation.

Zlochin [15] categorizes random search algorithms as instance-based or model-based, where instance-based methods generate new candidate points based on the current point or population of points, and model-based methods rely on an explicit sampling distribution and update parameters of the probability distribution. Examples of instance-based algorithms include simulated annealing, GAs, and tabu search, whereas examples of model-based algorithms include ant colony optimization, stochastic gradient search, and the cross-entropy method.

To understand the differences between random search algorithms, we concentrate on the procedures of generating new candidate points, and updating the collection of points maintained. The next section describes generating and updating procedures for several random search algorithms. Even though

instance-based methods may use heuristics to generate and update points, they do implicitly induce a sampling distribution. Abstracting the random search algorithms by a sequence of sampling distributions (whether implicit or explicit) provides a means to understand the performance of the algorithms. In the section titled, “Performance of Pure Random Search, Pure Adaptive Search, and Annealing Adaptive Search,” we analyze the performance of pure random search (PRS), pure adaptive search (PAS) and annealing adaptive search (AAS) by investigating their sampling distributions. We discuss model-based methods in the context of their explicit sampling distribution and a metacontrol framework in the section titled, “Cross-entropy and other Model-based Methods,” where algorithm parameters affecting the sampling distributions are adapted based on observed function values. We hope this overview serves to assist researchers in using random search algorithms and developing better methods in the future.

GENERIC RANDOM SEARCH AND SPECIFIC ALGORITHMS

The general global optimization problem (P) used here is defined as,

$$\min_{x \in S} f(x),$$

where x is a vector of n decision variables, S is an n -dimensional feasible region and assumed to be nonempty, and f is a real-valued function defined over S . The goal is to find a value for x contained in S that minimizes f . The objective function may be an oracle, or black-box function, in the sense that an explicit mathematical expression is not required but a numerical value for f must be returned for an input value of x . The feasible set S may be represented by a system of nonlinear equations, such as $g_j(x) \leq 0$ for $j = 1, \dots, m$, or by a membership function that simply returns whether or not a given x is in S . Notice that the feasible region may include both continuous and discrete variables. Denote the global optimal solution to

(P) by (x_*, y_*) where

$$x_* = \arg \min_{x \in S} f(x) \text{ and } y_* = f(x_*) = \min_{x \in S} f(x).$$

Random search algorithms are also frequently applied to stochastic optimization problems (also known as *simulation-optimization*) where the objective function and constraints involve randomness in the form of noisy measurements. In this case, the objective function is often the expectation of a function of a random variable, such as,

$$f(x) = E[F(x, Y_x)],$$

where F is a real-valued function of x and a random variable Y_x that may depend on $x \in S$. We assume the expectation exists and is finite, however, in practice the probability distribution is not known, and only realizations of F for specific values of x and y_x can be obtained.

In order to ensure a global optimum exists, we need to assume some regularity conditions. If (P) is a continuous problem and the feasible set S is nonempty and compact, then the Weierstrass theorem from classical analysis guarantees the existence of a global solution [16]. If (P) is a discrete problem and the feasible set S is nonempty and finite, a global solution exists. The existence of a global optimum can be guaranteed under slightly more general conditions. Note that the existence of a *unique* minimum at x_* is not required. If there are multiple optimal minima, let x_* be an arbitrary fixed global minimum.

In order to describe random search algorithms on continuous and/or discrete domains, it is necessary to assume some concept of neighborhood. In Euclidean space (real or integer valued variables), the notion of neighborhood is typically defined by a ball of radius $\delta, \delta > 0$, of points that are within a (Euclidean) distance δ of a point $x \in S$. However, the distance between points could be associated with the algorithm, rather than the problem domain. For example, the TSP in combinatorial optimization has several possible neighborhood structures (e.g., one-city swap, two-city swap) that are integrally associated with the search algorithm. Even for problems

with a continuous domain, the neighborhood associated with an algorithm with a short step size would differ from the neighborhood of an algorithm with a large step size. Thus the term *local search* must be tempered with the concept of the algorithm's neighborhood.

We describe a generic random search algorithm by a sequence of iterates X_k on iteration $k = 0, 1, \dots$ which may depend on previous points and algorithmic parameters. The current iterate X_k may represent a single point, or a collection of points, to include population-based algorithms. The iterates are also capitalized to denote that they are random variables, reflecting the probabilistic nature of the random search algorithm.

Generic Random Search Algorithm

Step 0. Initialize algorithm parameters Θ_0 , initial points $X_0 \subset S$ and iteration index $k = 0$.

Step 1. Generate a collection of candidate points $V_{k+1} \subset S$ according to a specific *generator* and associated *sampling distribution*.

Step 2. Update X_{k+1} based on the candidate points V_{k+1} , previous iterates and algorithmic parameters. Also update algorithm parameters Θ_{k+1} .

Step 3. If a stopping criterion is met, stop. Otherwise increment k and return to Step 1.

This generic random search algorithm depends on two basic procedures, the generator in Step 1 that produces candidate points, and the update procedure in Step 2. We first discuss some examples for the generator procedure.

Single-Point Generators. Many random search algorithms maintain and generate a single point at each iteration. For these single-point generators, the candidate point V_{k+1} is generated based on a combination of the current point and previous points. A common method to generate a candidate point is to take a step size in a vector direction, called a *direction-step paradigm* in Ref. 5, or also known as *step size algorithms*.

Step 1 of a step size algorithm can be expressed by

$$V_{k+1} = X_k + S_k D_k,$$

where the candidate point is generated by taking a step from the current point X_k of length S_k in a specified direction D_k on iteration k . In continuous problems, the direction of movement D_k may be motivated by gradient information, and if the objective function is noisy, then the probability distribution of D_k is related to the estimate of the gradient. The step length may be the result of a line search or an independent random step length. Quasi-Newton methods take advantage of an approximation of the Hessian to provide a search direction. In the basic method of stochastic approximation, also referred to as the *Robbins–Monro algorithm*, the direction of movement is a stochastic approximation to the gradient, and the step length is referred to as the *gain* [6]. If the direction is closely related to the gradient, as in stochastic gradient search, then local convergence may be proven under certain conditions, however, other variations must be introduced to escape local minima and find the global minimum.

As an alternative, a direction D_k that does not use any local information and is generated according to a uniform distribution on a hypersphere has been investigated. A collection of step size algorithms, including Refs 17–23, obtain a direction vector by sampling from a uniform distribution on a unit hypersphere. The method of choosing the step size may also be randomly generated, and may shrink or expand depending on the success of previously sampled points. It is interesting that several of these step size random search algorithms report experiencing computation that is linear in dimension.

The algorithm improving hit-and-run (IHR) also generates a direction uniformly on a hypersphere, and generates a step size according to a uniform distribution on the feasible portion of the line segment in direction D_k [23]. This simple generator is motivated by convergence properties of a Markov chain Monte Carlo sampler called *hit-and-run* to a target distribution [24–27].

The hit-and-run generator coupled with an update procedure has been used in several simulated annealing algorithms [28–30]. The update procedure in IHR is simply to only accept improving points. Even though this is a straightforward random search algorithm, its expected performance on a class of spherical programs is polynomial $O(n^{5/2})$ in dimension [23].

Examples of generators on discrete domains include nearest neighbor or local search, ball walk, k -city swap for TSP, and more recently, very large-scale neighborhood search (VLSN) [31]. Classical Markov chains such as the nearest neighbor random walk or the coordinate direction random walk can get trapped in isolated regions of the domain. A variation of hit-and-run has been proven to converge to an arbitrary target distribution over an arbitrary subset S of an integer hyperrectangle [32]. An extension of hit-and-run to mixed continuous/integer domains using a pattern generator has recently been embedded in IHR for global optimization [33]. The generator in tabu search is known for its restriction on neighborhoods—it uses a local neighborhood but prevents points from being generated that have been recently visited [3,34]. In simulated annealing, the specific method of generating a candidate point based on the neighborhood structure can be described as a Markov chain. As such, we view it as inducing an implicit sampling distribution.

Multiple-Point Generators. Population-based random search algorithms use a collection of current points to generate another collection of candidate points. Many of these algorithms are motivated by biological processes, and include GAs, evolutionary programming, PSO, and ant colony optimization. The concept behind GAs and evolutionary programming [35–37] is to produce “off-spring” of the current population with parameters related to reproduction (e.g., genetic crossover). PSO [38,39] is based on the social behavior of birds and schools of fish, communicating to find food. Once an individual in the group finds a food source (associated with an objective function value),

it informs the others. While initially the individuals of the group behave independently, they eventually communicate and move collectively toward one of the found sources of food. The individuals, or particles in the swarm, are typically modeled by position and velocity, which are updated based on the success of the other particles in the group. Ant colony optimization [40] is also considered a swarm intelligence, where the population of ants are moving individually, but communicate through a pheromone that provides feedback as to successful locations.

Another multiple-point generator that uses a population of points is differential evolution [41] which uses a weighted difference between two population vectors to generate a third vector. It is similar to the particle swarm generator because no explicit probability distribution is used, but an implicit empirical distribution is created through the interaction of the points. The generator procedure is just telling part of the story. The update procedure also influences the resulting probability distribution underlying the random search method.

Update Procedure. After a candidate point is generated, Step 2 of the generic random search algorithm specifies a procedure to update the current point and algorithm parameters. Algorithms that are strictly improving have a simple procedure, update the current point only if the candidate point is improving,

$$X_{k+1} = \begin{cases} V_{k+1}, & \text{if } f(V_{k+1}) < f(X_k), \\ X_k, & \text{otherwise.} \end{cases}$$

This type of improving algorithm may get trapped in a local optimum if the neighborhood, or procedure for generating candidate points is too restricted. One remedy is to sample a large neighborhood, as in VLSN or to draw from the entire feasible set. Although IHR has a simple acceptance procedure, it converges in probability to the global optimum because its generator has a positive probability of sampling anywhere in the bounded feasible region. Another approach is to accept nonimproving points, as in simulated annealing.

In simulated annealing [42–44], the update procedure is the Metropolis criterion, and the candidate point is accepted with a probability that reflects a Boltzmann distribution,

$$X_{k+1} = \begin{cases} V_{k+1}, & \text{with probability} \\ & \min \left\{ 1, \exp \left(\frac{f(X_k) - f(V_{k+1})}{T_k} \right) \right\}, \\ X_k, & \text{otherwise.} \end{cases}$$

In contrast to algorithms with strict improvement, simulated annealing allows nonimproving points to be accepted, which provides a means to escape a local optimum. With the Metropolis acceptance/rejection criterion, improving points are accepted with probability one, and the probability of accepting a worse point decreases as the temperature parameter cools, that is, decreases to zero. The update of the temperature parameter, called the *cooling schedule*, has been studied in the literature [28,45–49]. Typically the temperature parameter decreases monotonically, however, recent research on dynamically heating and cooling the temperature in an interacting particle algorithm [50] is discussed later in this article.

The update procedure for GAs and evolutionary programming is called the *selection criterion*. A straightforward scheme is to rank the population of points based on objective function value (also referred to as the *fitness function*), and select the top percentile of points to use in reproduction and recombination. However experience has shown that some diversity should be maintained to prevent premature convergence to a local optimum. The elitist strategy subdivides the population into three categories: the elite (best) solutions, the immigrant solutions (added to maintain diversity), and the crossover solutions [5]. This ensures that the best solutions are always carried to the next generation of points. The most effective proportion of elite, immigrant, and crossover solutions requires computational experimentation. An effective population size for a particular problem also requires experimentation. Researchers have used Markov chain analysis to analyze the behavior, such as the expected waiting time, of genetic algorithms [51,52].

Multistart and Clustering Algorithms. Algorithms may modify their generating and updating procedures per iteration. Consider multistart; on one iteration a candidate point is generated according to a global sampling distribution (e.g., a uniform distribution) on S , but on the next iteration a local search sampling distribution (e.g., a stochastic gradient search) is used. The update procedure is viewed as setting an indicating parameter to determine whether a global or local sample is generated for the next candidate point. These methods are often called *two-phase methods* [14]. Different combinations of sampling distributions can be created, such as generating uniformly distributed points in the global phase that initiate simulated annealing in the local phase, or use simulated annealing in the global search phase to initiate stochastic gradient search or nearest neighbor local searches.

An inefficiency in multistart is that the same local optimum may be repeatedly discovered from many different starting points. This has motivated clustering methods where clusters are formed (often grown around a seed point) to predict whether a starting point is likely to lead to a new local optimum or one that has already been discovered. Then local searches are only initiated at promising candidate points [7,14]. The update procedure may involve a sophisticated statistical analysis to decide whether to initiate a local search, as well as when to terminate a local search and generate more global candidate points. A related idea has led to linkage methods, which “link” points in the sample and essentially view clusters as trees, instead of spherical or ellipsoidal clusters. The well-known linkage method called *multilevel single linkage* [53,54] and several variants including random linkage [55] are summarized in Ref. 14.

More recently, a multistart heuristic for OptQuest/Nonlinear Programming, has been designed to operate on mixed integer nonlinear problems where the functions are differentiable with respect to the continuous variables [56]. They use scatter search (related to evolutionary programming) in the global search phase, and use filtering instead of clustering to decide when to initiate a

local search. The nested partitioning method [57] is based on a partitioning of the feasible region that can be viewed as the global phase, coupled with local random search within the subregions.

Convergence in Probability. Solis and Wets [22] provide a convergence proof, in probability, to the global minimum for general step size algorithms with conditions on the method of generating the step length and direction. Essentially, as long as the generator does not consistently ignore any region, then the algorithm will converge with probability one. Another convergence proof due to Bélisle [24] says that, even if the generator of an algorithm cannot reach any point in the domain in one iteration, if there is a means such as an acceptance probability to allow the algorithm to reach any point in a finite number of iterations, then the algorithm still converges with probability one to the global optimum. Stephens and Baritompä [58] go on to prove that an algorithm must either sample the entire region or use global information regarding the structure of the problem to provide convergence results. Examples of global information are bounds on the function, a Lipschitz constant, information on the level sets, number of local optima, functional form, and the global optimum itself. They conclude that using “global optimization heuristics is often far more practical than running general algorithms until the mathematically proven stopping criteria are satisfied” (58, p. 587).

The generating and updating procedures for random search algorithms can be viewed as Markov chain Monte Carlo samplers, and then the algorithms can be interpreted as implicitly or explicitly sampling from a sequence of distributions. Markov chain theory has been applied to the analysis of several versions of these algorithms, including simulated annealing and GAs, to prove convergence.

Several convergence results are also available for random search algorithms applied to noisy objective functions. Convergence results for simulated annealing applied to discrete problems with noisy function evaluations are reviewed in Ref. 59, and the authors prove that simulated annealing

with constant temperature converges almost surely to the set of global optimizers under mild assumptions on the Markov chain associated with the candidate point generator and with strongly consistent estimates of the objective function. Spall (6, Section 8.6) summarizes convergence results on continuous problems with noisy objective functions relating simulated annealing to stochastic approximation methods. GAs have also been applied to noisy functions but Spall remarks that “there appears to be relatively little formal analysis of GAs in the presence of noise” (6, p. 248). The nested partitioning method, introduced in Ref. 57 for objective functions that can be evaluated deterministically, was adapted for objective functions with noisy evaluations in Ref. 60 and shown to converge with probability one on discrete domains under mild assumptions. Cross-entropy and model reference adaptive search (MRAS) have also been applied to both deterministic and noisy objective functions, and are discussed in the section titled “Cross-Entropy and other Model-Based Methods”.

We next analyze the performance of three algorithms with explicit sampling distributions: PRS, PAS, and AAS, to gain insight into the distributions we could approximate with generating and updating procedures.

PERFORMANCE OF PURE RANDOM SEARCH, PURE ADAPTIVE SEARCH, AND ANNEALING ADAPTIVE SEARCH

The simplest and most obvious random search method is a “blind search” as discussed in Ref. 6, or called *pure random search* in Ref. 8. PRS was first defined in 1958 by Brooks [61] for problems with continuous variables, and discussed later in Refs 62 and 63. PRS is also easily defined for problems with discrete variables, as in Refs 6 and 8. PRS generates samples repeatedly from the feasible region S , typically according to a uniform sampling distribution. In the context of the generic random search algorithm, each candidate point is generated independently from a uniform distribution (continuous or discrete, as appropriate) on

S , and X_{k+1} is updated only if the candidate point is improving. It can be shown that PRS converges to within an ϵ distance of the global optimum with probability one [8,22]. PRS is often the global phase in multistart and clustering algorithms.

Even though PRS converges almost surely [6], it will take a long time. In order to describe the performance of PRS, we use $E[N(y_* + \epsilon)]$, the expected number of iterations until a point within ϵ distance of the global minimum is first sampled, as a measure of computational complexity. For PRS, an iteration generates one point uniformly (continuous or discrete, as appropriate) on S and performs exactly one function evaluation, so this captures the majority of the computational effort. The random variable $N(y)$, the number of iterations until a point is first sampled with an objective function value of y or less, has a geometric distribution [64], where the probability of a “success” is $p(y)$, the probability of generating a point in the level set $S(y)$, where $S(y) = \{x \in S : f(x) \leq y\}$. Then, the expected number of iterations until a point is first sampled within ϵ distance of the optimum is

$$E[N(y_* + \epsilon)] = 1/p(y_* + \epsilon).$$

This relationship supports the natural intuition that, as the target region close to the global optimum gets small (i.e., $p(y_* + \epsilon)$ gets small), the expected number of iterations gets large (inversely proportional). For a discrete problem, ϵ may equal zero, whereas for a continuous problem, ϵ is taken as a small positive value. The probability of failure to achieve $y_* + \epsilon$ after k iterations is $(1 - p(y_* + \epsilon))^k$. The variance of the number of iterations until a point first lands in $S(y)$ is

$$\text{Var}(N(y)) = \frac{1 - p(y)}{p(y)^2}.$$

This supports numerical observations that PRS (and other random search methods) experiences large variation; as the probability $p(y)$ decreases, the variance of $N(y)$ increases.

To gain insight into the performance of PRS, consider the TSP with N cities and

subsequently $(N - 1)!$ possible points in the domain. If there is a unique minimum, then $p(y_*) = 1/(N - 1)!$ and the expected number of PRS iterations to first sample the minimum is $(N - 1)!$, which explodes in N .

To illustrate the performance of PRS on a continuous domain problem, consider a global optimization problem where the domain S is an n -dimensional ball of radius 1, and the area within ϵ distance of the global optimum is a ball of radius ϵ , for $0 < \epsilon \leq 1$. Using a uniform sampling distribution on this problem, $p(y_* + \epsilon) = \epsilon^n$ for $0 < \epsilon \leq 1$ and the expected number of iterations until a sample point falls within the ϵ -ball is $(1/\epsilon)^n$, an exponential function of dimension.

To expand this example, it is shown in Ref. 8 that the expected number of PRS iterations on a global optimization problem where the objective function satisfies a Lipschitz condition with constant K is also exponential in the dimension of the problem,

$$E[N(y_* + \epsilon)] = (K/\epsilon)^n.$$

All random search algorithms are meant to improve upon PRS. However, the “No Free Lunch” theorems in Wolpert and Macready [65] show that, in essence, no single algorithm can improve on any other when averaged across all possible problems. As a consequence of these theorems, in order to improve upon PRS, we must either restrict the class of problems, have some prior information, or adapt the algorithm as properties of a specific problem are observed.

In contrast to PRS, where each sample is independent and identically distributed, we next consider PAS, where each sample depends on the one immediately before it. PAS was introduced in Ref. 66 for convex programs and later analyzed for global optimization problems with Lipschitz continuous functions in Ref. 67 and for finite domains in Ref. 68. It was generalized as hesitant adaptive search in Refs 69 and 70. PAS, by definition, generates a candidate point uniformly in the subset of the domain that is strictly improving, $V_{k+1} \in S(f(X_k)) = \{x : x \in S \text{ and } f(x) < f(X_k)\}$. It is an idealized algorithm to show potential performance for a random search

algorithm. Whereas, sampling from a uniform distribution in PRS over S is typically very easy (e.g., if S is a hyperrectangle), sampling from a uniform distribution in PAS over $S(f(X_k))$ is very difficult in general. However, the analysis shows the value of being able to find points in the improving level set; the number of iterations of PAS is an exponential improvement over the number of iterations in PRS (8, Theorem 2.2). Additionally, for Lipschitz continuous objective functions, the expected number of iterations of PAS is *linear* in dimension (8, Theorem 2.9), with an analogous result for finite domains (8, Corollary 2.9). The linearity result for PAS implies that adapting the search to sample improving points is very powerful.

The idealistic linear property of PAS has inspired the design of several algorithms with the goal of approximating PAS, including IHR [23] which uses hit-and-run as a Markov chain Monte Carlo sampler embedded in an optimization context, and Grover Adaptive Search [71] which uses quantum computing for optimization.

We now describe one more idealistic algorithm, AAS, which is a way to abstract simulated annealing. AAS differs from PAS in that the candidate points are generated from the entire feasible set S according to a Boltzmann distribution parameterized by temperature, whereas the sampling distribution in PAS only has positive support on nested level sets of S . The name AAS [72] is used because the algorithm assumes that points can be generated from exact Boltzmann distributions, which is a model of simulated annealing. The purpose of the AAS analysis is to gain understanding of simulated annealing and other stochastic algorithms.

The record values (best values observed) of AAS, first analyzed in Ref. 29, stochastically dominate those of PAS and hence inherit the ideal linear complexity of PAS. The number of sample points (including nonimproving points) of AAS is also shown to be linear on average under certain conditions on the cooling schedule for the temperature parameter. The analysis motivated a cooling schedule for simulated annealing which maintains a linear complexity in expected

number of sample points on a class of (Lipschitz) objective functions over both continuous and discrete domains [49].

CROSS-ENTROPY AND OTHER MODEL-BASED METHODS

In keeping with our generic random search algorithm, we now relate model-based methods to the generating and updating procedures. We view instance-based methods as emphasizing generating new points and subsequently updating them using algorithm parameters Θ_k , but now model-based methods change the emphasis to updating the algorithm parameters Θ_k and subsequently using them to generate new points. In this section, we discuss three methods that adapt the algorithm parameters using a model of the sampling distribution: cross-entropy, MRAS, and metacontrol.

To contrast instance-based and model-based methods, we compare them to an idealized algorithm, such as PAS or AAS, where we want to sample points according to a sequence of probability density functions $h_k(x)$. In PAS, the sampling distribution $h_k(x)$ is uniform on the level set associated with the best objective function value on the k th iteration. In AAS, the sampling distribution $h_k(x)$ is Boltzmann with a temperature parameter associated with the k th iteration. In both PAS and AAS, if we could sample directly from $h_k(x)$, the algorithm would have excellent performance. However, sampling directly from $h_k(x)$ is difficult, so we have to use an approximation. Instance-based methods can be interpreted as a Markov chain sampling approach to approximate $h_k(x)$. For example, IHR can be interpreted as a Markov chain sampling approach to approximate a series of PAS uniform distributions on level sets, and simulated annealing can be interpreted as a Markov chain sampling approach to approximate a series of AAS Boltzmann distributions. Model-based methods approximate a sequence of distributions $h_k(x)$ with a family of distributions that can be sampled from directly. The parameters of the distribution are adapted based on observed values.

Model-based methods, as discussed in Ref. 15, include ant colony optimization, stochastic gradient search, cross-entropy, and estimation of distribution methods. The MRAS algorithm introduced in Ref. 75 is also considered a model-based method.

Cross-entropy [73,74] and MRAS [75] use the idea of importance sampling, or sequential importance sampling, to approximate a sequence of target distributions $h_k(x)$. The idea is to use a simple parametrized probability density function $g(x, \Theta)$ that is easy to sample from, and update the parameter Θ so that the new samplings are biased toward the global optimum. Choices for $g(x, \Theta)$ typically come from the natural exponential family, which include Gaussian, Poisson, binomial, geometric, and other multivariate forms [73,74,75]. For example, when the Gaussian probability density function is used, then Θ consists of the mean vector and the covariance matrix. We now seek a parameter Θ_{k+1} that provides a good approximation to $h_k(x)$. This leads to the problem of minimizing the Kullback–Leibler divergence between $g(x, \Theta)$ and $h_k(x)$, that is

$$\Theta_{k+1} = \arg \inf_{\Theta} \int_S \ln \left(\frac{h_k(x)}{g(x, \Theta)} \right) h_k(x) dx.$$

The name of the cross-entropy method comes from this idea because the Kullback–Leibler divergence is also referred to as the *cross-entropy* between two probability density functions.

Now a new probability density function $h_{k+1}(x)$ is created by “tilting” $h_k(x)$ to bias it toward the improving regions in S using a performance function $G_y(x)$, as follows

$$h_{k+1}(x) = \frac{G_{y_{k+1}}(x)h_k(x)}{\int_S G_{y_{k+1}}(x)h_k(x) dx}.$$

The performance function $G_y(x)$ is assumed to be positive and relates to the objective function $f(x)$. One alternative for a performance function is $G_{y_{k+1}}(x) = \exp(f(x)) I_{S(y_{k+1})}(x)$ where

$$I_{S(y_{k+1})}(x) = \begin{cases} 1, & \text{if } x \in S(y_{k+1}), \\ 0, & \text{otherwise,} \end{cases}$$

is the indicator of the level set $S(y_{k+1})$. A sample of candidate points drawn from the sampling distribution $g(x, \Theta_{k+1})$ is used to determine a level set threshold y_{k+1} . In Ref. 75, y_{k+1} is estimated as a specified quantile of $f(x)$ with respect to $h_k(x)$.

Global convergence properties of these model-based methods appear in Refs 15, 73, 75 for both continuous and discrete domains. Cross-entropy and MRAS have also been applied successfully to noisy objective functions [73,76]. Convergence results for cross-entropy applied to the noisy buffer allocation problem and noisy TSP are summarized in Ref. 73, Chapter 6. Convergence properties for a version of MRAS for noisy objective functions on continuous and discrete domains are given in Ref. 76.

To complete this article, we mention recent work on a metacontrol methodology for adapting algorithm parameters based on observed values. Metacontrol of a deterministic optimization algorithm was introduced in Ref. 77, and extended to a random search algorithm in Ref. 50, where the temperature parameter is dynamically controlled in the interacting particle algorithm.

The interacting particle algorithm in Refs 50 and 78 is based on Feynman–Kac systems in statistical physics [79,80] and combines the ideas of simulated annealing with population-based algorithms. As such, it has a temperature parameter that influences the sampling distribution through acceptance/rejection. Instead of choosing a cooling schedule a priori, the metacontrol methodology dynamically determines the temperature parameter by adapting it based on observed sample points. An optimal control problem that controls the evolution of the probability density function of the particles with the temperature parameter is defined to make the algorithm’s search process satisfy specified performance criteria. The criteria in Ref. 50 include the expected objective function value of the particles, the spread of the particles, and the algorithm running time. Numerical results indicate improved performance by allowing heating and cooling of the temperature parameter through the metacontrol methodology [50]. The metacontrol methodology shares a

common goal with cross-entropy and MRAS: to provide a means to dynamically adapt algorithm parameters based on observed values in a closed-loop with feedback.

In this article, a generic random search algorithm is used to provide a common language to describe a variety of random search algorithms. Theoretical properties of convergence and performance are discussed to provide insight into the behavior of random search algorithms. Several model-based methods are reviewed with different approaches to adapt algorithm parameters. This serves as a starting point for researchers and practitioners interested in developing and applying random search algorithms for solving ill-structured global optimization problems.

REFERENCES

1. Blum C, Roli A. Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Comput Surv* 2003;35(3):268–308.
2. Boender CGE, Romeijn HE. In: Horst R, Pardalos PM, editors. *Handbook of global optimization*. Dordrecht: Kluwer Academic Publishers; 1995. pp. 829–869.
3. Glover F, Kochenberger GA. Volume 57, *Handbook of metaheuristics: international series in operations research & management science*. Boston (MA): Kluwer Academic Publishers; 2003.
4. Pardalos PM, Romeijn HE. Volume 2, *Handbook of global optimization*. Dordrecht: Kluwer Academic Publishers; 2002.
5. Rardin RL. *Optimization in operation research*. Upper Saddle River (NJ): Prentice Hall; 1998.
6. Spall JC. *Introduction to stochastic search and optimization: estimation, simulation and control*. Hoboken (NJ): Wiley; 2003.
7. Törn A, Žilinskas A. *Global optimization*. Berlin: Springer; 1989.
8. Zabinsky ZB. *Stochastic adaptive search for global optimization*. Boston (MA): Kluwer Academic Publishers; 2003.
9. Horst R, Hoang T. *Global optimization: deterministic approaches*. 3rd ed. Berlin: Springer; 1996.
10. Horst R, Pardalos PM. *Handbook of global optimization*. Dordrecht: Kluwer Academic Publishers; 1995.

11. Vavasis SA. Complexity issues in a global optimization: a survey In: Horst R, Pardalos PM, editors. Handbook of global optimization. Dordrecht: Kluwer Academic Publishers; 1995. pp. 27–41.
12. Dyer ME, Frieze AM. Computing the volume of convex bodies: a case where randomness provably helps. *Proc Symp Appl Math* 1991;44:123–169.
13. Dyer M, Frieze A, Kannan R. A random polynomial time algorithm for approximating the volume of convex bodies. *J ACM* 1991;38:1–17.
14. Schoen F. Two-phase methods for global optimization In: Pardalos PM, Romeijn HE, editors. Volume 2, Handbook of global optimization. Dordrecht: Kluwer Academic Publishers; 2002. pp. 151–177.
15. Zlochin M, Birattari M, Meuleau N, *et al.* Model-based search for combinatorial optimization: a critical survey. *Ann Oper Res* 2004;131:373–395.
16. Luenberger DG. Linear and nonlinear programming. 2nd ed. Reading (MA): Addison-Wesley; 1984.
17. Mutseniyeks VA, Rastrigin L. Extremal control of continuous multi-parameter systems by the method of random search. *Eng Cybern* 1964;1:82–90.
18. Rastrigin LA. Extremal control by the method of random scanning. *Autom Remote Control* 1960;21:891–896.
19. Rastrigin LA. The convergence of the random method in the extremal control of a many-parameter system. *Autom Remote Control* 1963;24:1337–1342.
20. Schrack G, Borowski N. An experimental comparison of three random searches In: Lootsma F, editor. Numerical methods for nonlinear optimization. London: Academic Press; 1972. pp. 137–147.
21. Schrack G, Choit M. Optimized relative step size random searches. *Math Program* 1976;10:270–276.
22. Solis FJ, Wets RJ-B. Minimization by random search techniques. *Math Oper Res* 1981;6:19–30.
23. Zabinsky ZB, Smith RL, McDonald JF, *et al.* Improving hit and run for global optimization. *J Glob Optimiz* 1993;3:171–192.
24. Bélisle CJP. Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *J Appl Probab* 1992;29:885–895.
25. Bélisle CJP, Romeijn HE, Smith RL. Hit-and-run algorithms for generating multivariate distributions. *Math Oper Res* 1993;18:255–266.
26. Lovász L. Hit-and-run mixes fast. *Math Program* 1999;86:443–461.
27. Smith RL. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper Res* 1984;32:1296–1308.
28. Romeijn HE, Smith RL. Simulated annealing for constrained global optimization. *J Glob Optimiz* 1994a;5:101–126.
29. Romeijn HE, Smith RL. Simulated annealing and adaptive search in global optimization. *Probab Eng Inform Sci* 1994b;8:571–590.
30. Romeijn HE, Zabinsky ZB, Graesser DL, *et al.* New reflection generator for simulated annealing in mixed-integer/continuous global optimization. *J Optimiz Theory Appl* 1999;101(2):403–427.
31. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123:75–102.
32. Baumert S, Ghate A, Kiatsupaibul S, *et al.* Discrete hit-and-run for generating multivariate distributions over arbitrary finite subsets of a lattice. *Oper Res* 2009;57(3):727–739.
33. Mete HO, Shen Y, Zabinsky ZB, *et al.* Pattern discrete and mixed hit-and-run for global optimization to appear in *J Glob Optim*. 2010.
34. Glover F, Laguna M. Tabu search In: Reeves CR, editor. Modern heuristic techniques for combinatorial problems. New York: Halsted Press; 1993. pp. 70–150.
35. Bäck T, Fogel D, Michalewicz Z. Handbook of evolutionary computation. New York: IOP Publishing and Oxford University Press; 1997.
36. Davis L. Handbook of genetic algorithms. New York: Van Nostrand Reinhold; 1991.
37. Koehler GJ. New directions in genetic algorithm theory. *Ann Oper Res* 1997;75:49–68.
38. Kennedy J, Eberhart RC. Particle swarm optimization. *Proc IEEE Int Conf Neural Netw* 1995;4:1942–1948.
39. Kennedy J, Eberhart RC, Shi Y. Swarm intelligence. San Francisco (CA): Morgan Kaufmann Publishers; 2001.
40. Dorigo M, Stützle T. Ant colony optimization. Cambridge (MA): MIT Press; 2004.
41. Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 1997;11(4):341–359.

42. Aarts E, Korst J. Simulated annealing and boltzmann machines - a stochastic approach to combinatorial optimization and neural computers. New York: Wiley; 1989.
43. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. *Science* 1983;20:671–680.
44. Metropolis N, Rosenbluth A, Rosenbluth M, *et al.* Equation of state calculations by fast computing machines. *J Chem Phys* 1953;21:1087–1090.
45. Cohn H, Fielding M. Simulated annealing: searching for an optimal temperature schedule. *SIAM J Optim* 1999;9(3):779–802.
46. Fielding M. Simulated annealing with an optimal fixed temperature. *SIAM J Optimiz* 2000;11(2):289–307.
47. Hajek B. Cooling schedules for optimal annealing. *Math Oper Res* 1988;13:311–329.
48. Locatelli M. Convergence of a simulated annealing algorithm for continuous global optimization. *J Glob Optim* 2000;18:219–234.
49. Shen Y, Kiatsupaibul S, Zabinsky ZB, *et al.* An analytically derived cooling schedule for simulated annealing. *J Glob Optim* 2007;38:333–365.
50. Molvalioglu O, Zabinsky ZB, Kohn W. The interacting-particle algorithm with dynamic heating and cooling. *J Glob Optim* 2009;43:329–356.
51. De Jong KA, Spears WM, Gordon DF. Using Markov chains to analyze GAFOs In: Whitley D, Vose M, editors. *Foundations of genetic algorithms 3*. San Francisco (CA): Morgan Kaufmann; 1995. pp. 115–137.
52. Koehler GJ. Computing simple GA expected waiting times. *Proc Genet Evol Comput Conf* 1999;795:1999.
53. Rinnooy Kan AHG, Timmer GT. Stochastic global optimization methods; part I: clustering methods. *Math Program* 1987;37:27–56.
54. Rinnooy Kan AHG, Timmer GT. Stochastic global optimization methods; part II: multi level methods. *Math Program* 1987;37: 57–78.
55. Locatelli M, Schoen F. Random linkage: a family of acceptance/rejection algorithms for global optimization. *Math Program* 1999;85(2):379–396.
56. Ugray Z, Lasdon L, Plummer J, *et al.* Scatter search and solvers: a multistart framework for global optimization. *Inf J Comput* 2007;19(3):328–340.
57. Shi L, Ólafsson S. Nested partitions method for global optimization. *Oper Res* 2000;48(3):390–407.
58. Stephens CP, Baritomba WP. Global optimization requires global information. *J Optim Theor Appl* 1998;96(3):575–588.
59. Alrafaei MH, Andradóttir S. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Manage Sci* 1999;45(5):748–764.
60. Shi L, Ólafsson S. Nested partitions method for stochastic optimization. *Methodol Comput Appl Probab* 2000;2(3):271–291.
61. Brooks SH. A discussion of random methods for seeking maxima. *Oper Res* 1958;6:244–251.
62. Dixon LCW, Szegő GP. *Towards global optimization*. Amsterdam: North-Holland; 1975.
63. Dixon LCW, Szegő GP. *Towards global optimization 2*. Amsterdam: North-Holland; 1978.
64. Devore JL. *Probability and statistics for engineering and the sciences*. 4th ed. Belmont (CA): Wadsworth, Inc.; 1995.
65. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997;1:67–82.
66. Patel NR, Smith RL, Zabinsky ZB. Pure adaptive search in Monte Carlo optimization. *Math Program* 1988;4:317–328.
67. Zabinsky ZB, Smith RL. Pure Adaptive Search in Global Optimization. *Math Program* 1992;53:323–338.
68. Zabinsky ZB, Wood GR, Steel MA, *et al.* Pure adaptive search for finite global optimization. *Math Program* 1995;69:443–448.
69. Bulger DW, Wood GR. Hesitant adaptive search for global optimization. *Math Program* 1998;81:89–102.
70. Wood GR, Zabinsky ZB, Kristinsdottir BP. Hesitant adaptive search: the distribution of the number of iterations to convergence. *Math Program* 2001;89(3):479–486.
71. Bulger D, Baritomba WP, Wood GR. Implementing pure adaptive search with grover's quantum algorithm. *J Optim Theor Appl* 2003;116:517–529.
72. Wood GR, Zabinsky ZB. Stochastic adaptive search In: Pardalos PM, Romeijn HE, editors. *Volume 2, Handbook of global optimization*. Dordrecht: Kluwer Academic Publishers; 2002. pp. 231–249.
73. Rubinstein RY, Kroese DP. The cross-entropy method: a unified approach to combinatorial optimization, monte carlo simulation, and

- machine learning. New York: Springer; 2004.
74. Rubinstein RY, Kroese DP. Simulation and the Monte Carlo method. 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008.
 75. Hu J, Fu MC, Marcus SI. A model reference adaptive search method for global optimization. *Oper Res* 2007;55:549–568.
 76. Hu J, Fu MC, Marcus SI. A model reference adaptive search method for stochastic global optimization. *Comm Inform Syst* 2008;8(3):245–276.
 77. Kohn W, Zabinsky ZB, Brayman V. Meta-control of an optimization algorithm. *J Glob Optimiz* 2006;34(2):293–316.
 78. Molvalioglu O, Zabinsky ZB, Kohn W. Multi-particle simulated annealing In: Törn A, Zilinskas J, Zilinskas A, editors. Models and algorithms for global optimization. New York: Springer; 2007. pp. 215–222.
 79. Del Moral P. Feynman-kac formulae: genealogical and interacting particle systems with applications. New York: Springer; 2004.
 80. Del Moral P, Miclo L. On the convergence and applications of generalized simulated annealing. *SIAM J Control Optim* 1999;37(4):1222–1250.

RANDOM VARIATE GENERATION

LAWRENCE M. LEEMIS
Department of Mathematics,
The College of William &
Mary, Williamsburg
Virginia

This article concerns algorithms for generating random variates that are typically used in Monte Carlo simulation (where the passage of time plays no role) or discrete-event simulation (where the passage of time plays a significant role). There is a subtle but important distinction between a random variable and a random variate. A *random variable* is a rule that assigns a real number to an outcome of an experiment. A *random variate* is a realization, or instantiation of a random variable, which is typically generated by a computer. Devroye [1] and Hörmann *et al.* [2] provide both an introductory and an encyclopedic treatment of variate generation.

The four functions—the cumulative distribution function, the probability density function, the hazard function, and the cumulative hazard function (chf)—describe the distribution of a random variable X .

The cumulative distribution function (cdf) of a random variable X ,

$$F(x) = P(X \leq x),$$

is a nondecreasing function of x satisfying $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$. When the cdf is differentiable,

$$f(x) = F'(x)$$

is the associated probability density function (pdf). For any interval (a, b) , where $a < b$,

$$P(a < X < b) = \int_a^b f(x) dx.$$

For discrete distributions, $F(x)$ is a step function and the mass values of $f(x)$ associated with the support values are equal to the heights of the steps in $F(x)$.

The hazard function, also known as the *rate function*, *failure rate*, and *force of mortality*, can be defined by

$$h(x) = \frac{f(x)}{1 - F(x)}.$$

The hazard function is popular in reliability and survival analysis because it has the intuitive interpretation as the amount of *risk* associated with an item that has survived to time x . The hazard function is mathematically equivalent to the intensity function for a nonhomogeneous Poisson process, and the failure time corresponds to the first event time in the process. Competing risk models are naturally formulated in terms of $h(x)$, as shown subsequently. The cumulative hazard function can be defined by

$$H(x) = \int_0^x h(y) dy.$$

We assume that there is a reliable source of pseudorandom numbers available, and we use U , with or without subscripts, to denote one instance of such a $U(0, 1)$ random variable. Any of the standard discrete-event simulation textbooks [3,4] contain a discussion of random number generation techniques. The purpose of this article is to present an overview of several random variate generation techniques that convert random numbers to random variates. The algorithms are broken into density-based algorithms and hazard-based algorithms, and there are analogies between the two sets of algorithms. A group of algorithms known as *special properties* (e.g., the sum of independent and identically distributed exponential random variables has the Erlang distribution) is neither density based nor hazard based.

There are a number of properties of variate generation algorithms that are important in the evaluation of the algorithms presented here. These include synchronization (one random number produces one random variate), monotonicity (a monotone relationship

between the random numbers and the associated random variates), robustness with respect to all values of parameters, number of lines of code, expected marginal time to generate a variate, setup time, susceptibility to computer round-off, memory requirements, portability, and so on. The choice between the various algorithms presented here must be made based on these criteria. Many of these properties are necessary for implementing “variance reduction techniques” [3].

DENSITY-BASED VARIATE GENERATION METHODS

There are three density-based algorithms for generating random variates: (i) the inverse-cdf technique; (ii) composition, which assumes that the pdf can be written as a convex combination; and (iii) acceptance–rejection, a majorization technique. These three algorithms are introduced in the sections that follow.

Inverse-cdf Technique

The inverse-cdf technique, often called *inversion*, is the algorithm of choice for generating random variate X . It is typically the fastest of the algorithms in terms of marginal execution time, produces a random variate from a single random number, and is monotone. The inverse-cdf technique is based on the probability integral transformation, which states that $F_X(X) \sim U(0, 1)$, where $F_X(x)$ is the cdf of the random variable X . This results in the following algorithm for generating a random variable X .

```
generate  $U \sim U(0, 1)$ 
 $X \leftarrow F^{-1}(U)$ 
return  $X$ 
```

The inverse-cdf technique works well on distributions that have a closed-form expression for $F^{-1}(u)$, for example, the exponential, Weibull, and log logistic distributions. Distributions that lack a closed-form expression for $F^{-1}(u)$, for example, the gamma and beta distributions, can still use the technique by

numerically integrating the pdf, although this tends to be slow. This difficulty leads to the development of other techniques for generating random variates.

Example. For an exponential random variable X with rate $\lambda > 0$, the cdf is given by

$$F(x) = 1 - e^{-\lambda x} \quad x > 0.$$

Inverting this expression yields the random variate generation expression

$$X \leftarrow -\frac{1}{\lambda} \log(1 - U).$$

Although the sense of the monotonicity is reversed, $1 - U$ can be replaced by U to speed up the algorithm by saving a subtraction.

In the previous example, the inverse cdf $F^{-1}(u)$ could be expressed in closed form. For discrete random variables and mixed random variables, however, a true inverse cdf $F^{-1}(u)$ does not exist in the mathematical sense. Devroye [1] gives a careful definition of $F^{-1}(u)$ that allows the inverse-cdf technique to be applied to discrete- and mixed-random variables, as illustrated next.

Example. For a Bernoulli random variable X with parameter $0 < p < 1$, the pdf is given by

$$f(x) = p^x(1 - p)^{1-x} \quad x = 0, 1.$$

A variate generation algorithm for generating a random Bernoulli variate is

```
generate  $U \sim U(0, 1)$ 
if  $X < 1 - p$ 
    return 0
else
    return 1
```

Here, and throughout this article, indentation is used in the algorithms to indicate nesting.

Composition

When the cdf of the random variable X can be written as the convex combination of n cdfs, that is,

$$F_X(x) = \sum_{i=1}^n p_i F_i(x),$$

where $\sum_{i=1}^n p_i = 1$ and $p_i > 0$ for $i = 1, 2, \dots, n$, then the composition algorithm can be used to generate a random variate X . Distributions that can be written in this form are also known as *finite mixture* distributions and there is a significant literature in this area [5,6]. The algorithm generates a random component distribution index I , then generates a variate from the chosen distribution using the inverse-cdf (or any other) technique.

```
generate  $I$  with probability  $p_i$ 
generate  $U \sim U(0, 1)$ 
 $X \leftarrow F_I^{-1}(U)$ 
return  $X$ 
```

This algorithm is not synchronized because two random numbers (one to generate the index I and another to generate X) are required. This difficulty can be overcome with a small alteration of the algorithm that recycles a portion of the random number used to generate the index I . In either case, however, the algorithm is not monotone, which creates difficulties in applying various variance reduction techniques to a simulation.

Acceptance–Rejection

The third and final algorithm is the acceptance–rejection technique, which requires finding a *majorizing function* $f^*(x)$ that satisfies

$$f^*(x) \geq f(x),$$

where $f(x)$ is the pdf of the random variable X . To improve execution time, it is best to find a majorizing function with minimum area. A scaled majorizing function $g(x)$ is

$$g(x) = \frac{f^*(x)}{\int_{-\infty}^{\infty} f^*(y) dy},$$

which is a legitimate pdf with associated cdf $G(x)$. The acceptance–rejection algorithm proceeds as described below.

```
repeat
  generate  $U \sim U(0, 1)$ 
   $X \leftarrow G^{-1}(U)$ 
  generate  $S \sim U(0, f^*(X))$ 
until  $(S \leq f(X))$ 
return  $X$ 
```

The acceptance–rejection algorithm uses a geometrically distributed number of $U(0,1)$'s to generate the random variate X . For this reason, the acceptance–rejection algorithm is not synchronized. The acceptance–rejection technique can be modified to cut its execution time by reducing the number of times that $f(X)$ needs to be evaluated. One of these techniques, the squeeze method, is illustrated in Schmeiser and Lal [7].

HAZARD-BASED VARIATE GENERATION METHODS

There are three hazard-based algorithms for generating random variates: (i) the inverse cumulative hazard function technique, an inversion technique that parallels the inverse-cdf technique; (ii) competing risks, a linear combination technique that parallels composition; and (iii) thinning, a majorization technique that parallels acceptance–rejection. These three algorithms are introduced in the sections that follow. Random numbers are again denoted by U and the associated random variates are denoted by X , consistent with the notation in the previous section. (Since hazard functions are commonly used in reliability and survival analysis, T , for *time*, is often used instead of X .)

Inverse Cumulative Hazard Function Technique

If X is a random variable with cumulative hazard function H , then $H(X)$ is an exponential random variable with a mean of 1. This

result, which is an extension of the probability integral transformation, is the basis for the inverse-chf technique. Therefore,

```
generate  $U \sim U(0, 1)$ 
 $X \leftarrow H^{-1}(-\log(1 - U))$ 
return  $X$ 
```

generates a single random lifetime X . This algorithm is the easiest to implement when H can be inverted in closed form. This algorithm is monotone and synchronized. Although the sense of the monotonicity is reversed, $1 - U$ can be replaced with U in order to save a subtraction. For identical values of U , the inverse-cdf technique and the inverse-chf technique generate the same random variate X .

Competing Risks

Competing risks [8,9] is a linear combination technique that is analogous to the density-based composition method. The competing risks technique applies when the hazard function can be written as the sum of hazard functions, each corresponding to an independent “cause” of failure

$$h(x) = \sum_{j=1}^k h_j(x),$$

where $h_j(x)$ is the hazard function associated with cause j of failure acting in a population. The minimum of the lifetimes from each of these risks corresponds to the system lifetime. Competing risks is most commonly used to analyze a series system of k independent components, but can also be used in actuarial applications with k independent causes of failure. The algorithm to generate a random variate X is

```
for  $j$  from 1 to  $k$ 
  generate  $X_j \sim h_j(x)$ 
 $X \leftarrow \min\{X_1, X_2, \dots, X_k\}$ 
return  $X$ 
```

The $X_1, X_2, \dots, X_k$ values can be generated by any of the standard random variate generation algorithms.

Thinning

The *thinning* algorithm, which was originally suggested by Lewis and Shedler [10] for generating the event times in a nonhomogeneous Poisson process, can be adapted to produce a single lifetime by returning only the first event time generated. A majorizing hazard function $h^*(x)$ must be found that satisfies $h^*(x) \geq h(x)$ for all values of x in the support of X . The algorithm is

```
 $X \leftarrow 0$ 
repeat
  generate  $Y$  from  $h^*(x)$  given  $Y > X$ 
   $X \leftarrow X + Y$ 
  generate  $S \sim U(0, h^*(X))$ 
until  $S \leq h(X)$ 
return  $X$ 
```

Generating Y in the repeat–until loop can be performed by inversion or any other method. The name *thinning* comes from the fact that X can make several steps, each of length Y , that are thinned out before the repeat–until loop condition is satisfied.

GENERATING STOCHASTIC PROCESSES

Most discrete-event simulation models have stochastic elements that mimic the probabilistic nature of the system under consideration. The focus in this section is on the generation of a sample realization of a small group of stochastic point processes. These models are used by probabilists to model the arrival times of customers to a queue, the arrival times of demands in an inventory system, the failure times of a repairable system, the times of the births of babies, and so on. These stochastic models generalize to model events that occur over time or space. The general question considered here is how to generate a sequence of event times when underlying stochastic process is known. We begin by introducing probabilistic models for sequences of events that occur at points in time. A special case of a point process is a “counting process,” where event occurrences increment a counter.

Counting Processes

A continuous-time, discrete-state stochastic process is often characterized by the *counting function* $\{N(t), t \geq 0\}$ which represents the total number of events that occur by time t [11]. A counting process satisfies the following properties:

1. $N(t) \geq 0$,
2. $N(t)$ is integer-valued,
3. if $s < t$, then $N(s) \leq N(t)$, and
4. for $s < t$, $N(t) - N(s)$ is the number of events in $(s, t]$.

Two important properties associated with some counting processes are *independent increments* and *stationarity*. A counting process has independent increments if the number of events that occur in mutually exclusive time intervals is independent. A counting process is stationary if the distribution of the number of events that occur in any time interval depends only on the length of the interval. Thus, the stationarity property should only apply to counting processes with a constant rate of occurrence of events.

Counting processes can be used in modeling events as diverse as earthquake occurrences [12], storm occurrences in the Arctic Sea [13], customer arrival times to an electronics store [14], and failure times of a repairable system [15].

We establish some additional notation at this point which are used in some results and process generation algorithms that follow. Let $X_1, X_2, \dots$ represent the times between events in a counting process. Let $T_n = X_1 + X_2 + \dots + X_n$ be the time of the n th event. With these basic definitions in place, we now define the Poisson process, which is the most fundamental of the counting processes.

Poisson Processes

A Poisson process is a special type of counting process that is a fundamental base case for defining many other types of counting processes.

Definition [11]. The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process* with rate λ , $\lambda > 0$, if

1. $N(0) = 0$,
2. the process has independent increments, and
3. the number of events in any interval of length t is Poisson distributed with mean λt .

The single parameter λ controls the rate at which events occur over time. Since λ is a constant, a Poisson process is often referred to as a *homogeneous* Poisson process. The third condition is equivalent to

$$P(N(t+s) - N(s) = n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!},$$

$$n = 0, 1, 2, \dots,$$

and the stationarity property follows from it.

Although there are many results that follow from the definition of a Poisson process, three are detailed in this paragraph that have applications in discrete-event simulation. Proofs are given in any introductory stochastic processes' textbook. First, given that n events occur in a given time interval $(s, t]$, the event times have the same distribution as the order statistics associated with n independent observations drawn from a uniform distribution on $(s, t]$; see, for example, Rigdon and Basu [16]. Second, the times between events in a Poisson process are independent and identically distributed exponential random variables with pdf $f(x) = \lambda e^{-\lambda x}$, for $x > 0$. Since the mode of the exponential distribution is 0, a realization of a Poisson process typically exhibits significant clustering of events. Since the sum of n independent and identically distributed exponential(λ) random variables is Erlang(λ, n), T_n has a cdf that can be expressed as a summation:

$$F_{T_n}(t) = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad t > 0.$$

Third, analogous to the central limit theorem, which shows that the sum of arbitrarily distributed random variables is asymptotically normal, the superposition of

renewal processes converges asymptotically to a Poisson process [17, p. 318].

The mathematical tractability associated with the Poisson process makes it a popular model. It is the base case for queueing theory (e.g., the M/M/1 queue as defined in Gross and Harris [18]) and reliability theory (e.g., the models for repairable systems described in Meeker and Escobar [19]). Its rather restrictive assumptions, however, limit its applicability. For this reason, we consider the following variants of the Poisson process that can be useful for modeling more complex failure time processes: the renewal process and the nonhomogeneous Poisson process. These variants are typically formulated by generalizing an assumption or a property of the Poisson process. Details associated with these models can be found, for example, in Resnick [20] or Nelson [21].

Renewal Processes

A renewal process is a generalization of a Poisson process. Recall that in a Poisson process, the inter-event times $X_1, X_2, \dots$ are independent and identically distributed exponential(λ) random variables. In a renewal process, the inter-event times are independent and identically distributed random variables from any distribution with positive support. One useful classification of renewal processes [22] concerns the coefficient of variation σ/μ of the distribution of the times between failures. This classification divides renewal processes into underdispersed and overdispersed processes. A renewal process is *underdispersed* (*overdispersed*) if the coefficient of variation of the distribution of the times between failures is less than (greater than) 1. An extreme case of an underdispersed process is when the coefficient of variation is 0 (i.e., deterministic inter-event times), which yields a deterministic renewal process. The underdispersed process is much more regular in its event times.

Nonhomogeneous Poisson Processes

A nonhomogeneous Poisson process is another generalization of a Poisson process that allows for a failure rate $\lambda(t)$ (known as

the *intensity* function) that can vary with time.

Definition [11]. The counting process $\{N(t), t \geq 0\}$ is said to be a *nonhomogeneous Poisson process* (NHPP) with intensity function $\lambda(t)$, $t \geq 0$, if

1. $N(0) = 0$,
2. the process has independent increments,
3. $P(N(t+h) - N(t) \geq 2) = o(h)$, and
4. $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$,

where a function $f(\cdot)$ is said to be $o(h)$ if $\lim_{h \rightarrow 0} f(h)/h = 0$.

An NHPP is often appropriate for modeling a series of events that occur over time in a nonstationary fashion. Two common application areas are the modeling of arrivals to a waiting line (queueing theory) and the failure times of a repairable system (reliability theory) with negligible repair times. The cumulative intensity function,

$$\Lambda(t) = \int_0^t \lambda(\tau) d\tau \quad t > 0,$$

gives the expected number of events by time t , that is, $\Lambda(t) = E[N(t)]$. The probability of exactly n events occurring in the interval $(a, b]$ is given by

$$\frac{\left[\int_a^b \lambda(t) dt \right]^n e^{-\int_a^b \lambda(t) dt}}{n!}$$

for $n = 0, 1, \dots$ [23].

Random Process Generation

The algorithms presented in this section generate a sequence of random event times on the time interval $(0, S]$, where S is a real, fixed constant. If the next-event approach is taken for placing events onto the calendar in a discrete-event simulation model, then these algorithms should be modified so that they take the current event time as an argument and return the next event time. All processes are assumed to begin at time 0. The random

event times that are generated by the counting process are denoted by $T_1, T_2, \dots$, and random numbers (i.e., $U(0, 1)$ random variables) are denoted by U or $U_1, U_2, \dots$. These algorithms allow for the possibility that no events were observed on $(0, S]$.

Poisson Processes. Since the times between events in a Poisson process are independent and identically distributed exponential(λ) random variables, the following algorithm generates the event times of a Poisson process on $(0, S]$.

```

 $T_0 \leftarrow 0$ 
 $i \leftarrow 0$ 
while  $T_i \leq S$ 
   $i \leftarrow i + 1$ 
  generate  $U_i \sim U(0, 1)$ 
   $T_i \leftarrow T_{i-1} - \log(1 - U_i) / \lambda$ 
return  $T_1, T_2, \dots, T_{i-1}$ 

```

Renewal Processes. Event times in a renewal process are generated in a similar fashion to a Poisson process. Let $F_X(x)$ denote the cdf of the inter-event times $X_1, X_2, \dots$ in a renewal process. The following algorithm generates the event times on $(0, S]$.

```

 $T_0 \leftarrow 0$ 
 $i \leftarrow 0$ 
while  $T_i \leq S$ 
   $i \leftarrow i + 1$ 
  generate  $U_i \sim U(0, 1)$ 
   $T_i \leftarrow T_{i-1} + F_X^{-1}(U_i)$ 
return  $T_1, T_2, \dots, T_{i-1}$ 

```

Nonhomogeneous Poisson Processes. Event times can be generated for use in discrete-event simulation as $\Lambda^{-1}(E_1), \Lambda^{-1}(E_2), \dots$, where $E_1, E_2, \dots$ are the event times in a unit Poisson process [23]. This technique is often referred to as *inversion*, and is implemented below.

```

 $T_0 \leftarrow 0$ 
 $E_0 \leftarrow 0$ 
 $i \leftarrow 0$ 
while  $T_i \leq S$ 
   $i \leftarrow i + 1$ 
  generate  $U_i \sim U(0, 1)$ 
   $E_i \leftarrow E_{i-1} - \log(1 - U_i)$ 
   $T_i \leftarrow \Lambda^{-1}(E_i)$ 

```

```

return  $T_1, T_2, \dots, T_{i-1}$ 

```

The inversion algorithm is ideal when $\Lambda(t)$ can be inverted analytically, although it also applies when $\Lambda(t)$ needs to be inverted numerically. There may be occasions when the numerical inversion of $\Lambda(t)$ is so onerous that the *thinning* algorithm devised by Lewis and Shedler [10] might be preferable. This algorithm assumes that the modeler has determined a *majorizing* value λ^* that satisfies $\lambda^* \geq \lambda(t)$ for all $t > 0$.

```

 $T_0 \leftarrow 0$ 
 $i \leftarrow 0$ 
while  $T_i \leq S$ 
   $t \leftarrow T_i$ 
  repeat
    generate  $U \sim U(0, 1)$ 
     $t \leftarrow t - \log(1 - U) / \lambda^*$ 
    generate  $U \sim U(0, \lambda^*)$ 
  until  $U \leq \lambda(t)$ 
   $i \leftarrow i + 1$ 
   $T_i \leftarrow t$ 
return  $T_1, T_2, \dots, T_{i-1}$ 

```

The majorizing value λ^* can be generalized to a majorizing function $\lambda^*(t)$ to decrease the CPU time by minimizing the probability of “rejection” in the **repeat–until** loop.

All the stochastic process models given in this article are elementary. More complex models are considered in Nelson *et al.* [24].

CONCLUSIONS AND FURTHER READING

The discussion here has been limited to the generation of univariate random variates and stochastic processes. This only scratches the surface of the extensive literature in random variate generation. Other topics that might be of interest to the reader include algorithms for generating from specific distributions [1,3], such as the polar method for generating normal random variables; generating multivariate random variables [25]; generating autocorrelated arrival processes and correlated random vectors using the ARTA, NORTA, and VARTA methods [26]; generating lifetimes under the accelerated life and Cox proportional

hazards models [27,28]; generating combinatorial objects [1,29,30]; generating random matrices [31–36]; generating random polynomials [37]; shuffling playing cards [38]; generating random spawning trees [39, p. 379]; generating random sequences [40]; generating Markov chains [11,41–44]. An overview of the simulation methodology can be found by reading an introductory text, for example, Law [3] or Banks *et al.* [4], or the comprehensive overview in Henderson and Nelson [45].

REFERENCES

- Devroye L. Non-uniform random variate generation. New York: Springer; 1986.
- Hörmann W, Leydold J, Deflinger G. Automatic nonuniform random variate generation. Berlin: Springer; 2004.
- Law AM. Simulation modeling and analysis. 4th ed. New York: McGraw–Hill; 2007.
- Banks J, Carson JS II, Nelson BL, *et al.* Discrete-event system simulation. 4th ed. Englewood Cliffs (NJ): Prentice–Hall; 2005.
- Everitt BS, Hand DJ. Finite mixture distributions. New York: Chapman and Hall; 1981.
- McLachlan G, Peel D. Finite mixture models. New York: John Wiley & Sons, Inc.; 2000.
- Schmeiser BW, Lal R. Squeeze methods for generating gamma random variates. *J Am Stat Assoc* 1980;75(371):679–682.
- Crowder MJ. Classical competing risks. Boca Raton (FL): Chapman and Hall/CRC Press; 2001.
- David HA, Moeschberger ML. The theory of competing risks. New York: Macmillan; 1978.
- Lewis PAW, Shedler GS. Simulation of non-homogeneous poisson processes by thinning. *Nav Res Log Q* 1979;26(3):403–413.
- Ross SL. Introduction to probability models. 8th ed. San Diego: Academic Press; 2003.
- Schoenberg FP. Multidimensional residual analysis of point process models for earthquake occurrences. *J Am Stat Assoc* 2003;98(464):789–795.
- Lee S, Wilson JR, Crawford MM. Modeling and simulation of a nonhomogeneous poisson process having cyclic behavior. *Commun Stat Simul Comput* 1991;20(2& 3):777–809.
- White KP. Simulating a nonstationary poisson process using bivariate thinning: the case of “Typical Weekday” arrivals at a consumer electronics store. In: Farrington P, Nembhard PA, Sturrock HB, *et al.*, editors. Proceedings of the 1999 Winter Simulation Conference. Piscataway (NJ): IEEE; 1999. pp. 458–461.
- Nelson WB. Recurrent events data analysis for product repairs, disease recurrences, and other applications. Philadelphia (PA): ASA/SIAM; 2003.
- Rigdon SE, Basu AP. Statistical methods for the reliability of repairable systems. New York: John Wiley & Sons, Inc.; 2000.
- Whitt W. Stochastic-process limits: an introduction to stochastic-process limits and their application to queues. New York: Springer; 2002.
- Gross D, Harris CM. Fundamentals of queueing theory. 2nd ed. New York: John Wiley & Sons, Inc.; 1985.
- Meeker WQ, Escobar LA. Statistical methods for reliability data. New York: John Wiley & Sons, Inc.; 1998.
- Resnick SI. Adventures in stochastic processes. Boston (MA): Birkhäuser; 1992.
- Nelson BL. Stochastic modeling: analysis and simulation. Mineola (NY): Dover Publications; 2002.
- Cox DR, Isham V. Point processes. New York: Chapman and Hall; 1980.
- Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice–Hall; 1975.
- Nelson BL, Ware P, Cario MC, *et al.* Input modeling when simple models fail. In: Alexopoulos C, Kang K, Lilegdon WR, *et al.*, editors. Proceedings of the 1995 Winter Simulation Conference. Piscataway (NJ): IEEE; 1995. pp. 93–100.
- Johnson ME. Multivariate statistical simulation. New York: John Wiley & Sons, Inc.; 1987.
- Biller B, Ghosh S. Multivariate input processes. In: Henderson SG, Nelson BL, editors. Simulation. Amsterdam North–Holland Publishing Co.; 2006. pp. 123–154.
- Cox DR, Oakes D. Analysis of survival data. London: Chapman and Hall; 1984.
- Leemis LM. Variate generation for the accelerated life and proportional hazards models. *Oper Res* 1987;35(6):892–894.
- Nijenhuis A, Wilf HS. Combinatorial algorithms for computers and calculators. 2nd ed. New York: Academic Press; 1978.
- Wilf HS. Combinatorial algorithms: an update. Philadelphia (PA): SIAM; 1989.

31. Carmeli M. Statistical theory and random matrices. New York: Marcel Dekker; 1983.
32. Deift P. Orthogonal polynomials and random matrices: a Riemann–Hilbert approach. New York: American Mathematical Society; 2000.
33. Edelman A, Kostlan E, Shub M. How many eigenvalues of a random matrix are real? *J Am Math Soc* 1994;7:247–267.
34. Mehta ML. Random matrices. 3rd ed. Amsterdam: Elsevier; 2004.
35. Marsaglia G, Olkin I. Generating correlation matrices. *SIAM J Sci Stat Comput* 1984;5(2):470–475.
36. Ghosh S, Henderson SG. Behavior of the NORTA method for correlated random vector generation as the dimension increases. *ACM Trans Model Comput Simul* 2003;13:276–294.
37. Edelman A, Kostlan E. How many zeros of a random polynomial are real? *Bull Am Math Soc* 1995;32(1):1–37.
38. Morris SB. Magic tricks, card shuffling and dynamic computer memories. Washington (DC): The Mathematical Association of America; 1998.
39. Fishman GS. Monte Carlo: concepts, algorithms, and applications. New York: Springer; 1996.
40. Knuth DE. The art of computer programming. Volume 2, Seminumerical algorithms. 3rd ed. Reading (MA): Addison–Wesley; 1998.
41. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. Boca Raton (FL): Chapman and Hall/CRC; 1996.
42. Devroye L. Nonuniform random variate generation. In: Henderson SG, Nelson BL, editors. *Simulation*. Amsterdam: North–Holland Publishing Co.; 2006. pp. 83–121.
43. Metropolis N, Rosenbluth AW, Rosenbluth MN, *et al.* Equations of state calculations by fast computing machine. *J Chem Phys* 1953;21:1087–1091.
44. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;57:97–109.
45. Henderson SG, Nelson BL, editors. *Simulation*. Volume 13, Handbooks in operations research and management science. Amsterdam: North–Holland Publishing Co.; 2006.

RANDOMIZED SIMPLEX ALGORITHMS

BERND GÄRTNER
Institute of Theoretical Computer
Science ETH, Zurich, Switzerland

INTRODUCTION

Here is the simplest and probably the most natural randomized simplex algorithm: in each pivot step, choose the entering variable uniformly at random among all candidates. More precisely, if there are k candidates for the entering variable, throw a fair k -sided dice to select one among them. The pivot rule underlying this algorithm has been termed *Random-Edge*.

In general, we can think of a randomized simplex algorithm as having access to a stream of random bits that can be tapped at discretion. While *Random-Edge* taps the stream in every pivot step, there are also other meaningful ways. For example, the randomized least-index simplex algorithm randomly reshuffles variable indices in the beginning, and after that it always enters the candidate of smallest index in the reshuffled order.

The Geometric View

In order for the algorithm's behavior to be determined by the rule for choosing the entering variable, we have to assume that the leaving variable is always unique. Symbolic perturbation of the right-hand side, for example, takes care of this. In analyzing the complexity of randomized simplex algorithms, one often makes further simplifying assumptions that do not incur a loss of generality for theoretical purposes. The first such assumption is that the feasible region is bounded and nondegenerate so that it has the form of a *simple polytope*. Moreover, the objective function is assumed

to be *generic*, which geometrically means that no two polytope vertices have the same value. Finally, one assumes that the algorithm has access to some initial vertex of the polytope. (see ***Degeneracy and Variable Entering/Exiting Rules***).

In this geometric setting, the simplex method can be interpreted as a path following method on *polytope digraphs* (vertex-edge graphs of simple polytopes, with edge orientations induced by linear functions): While the current vertex is not the (unique) sink, go along some outgoing edge to a neighboring vertex. Starting from a standard-form linear program (LP) in n nonnegative variables and m equality constraints, we obtain a polytope with n facets in dimension $d := n - m$.

The term *Random-Edge* has been coined with the polytope digraph model in mind. Indeed, *Random-Edge* chooses among the outgoing edges uniformly at random in each step. There is a second well-studied randomized simplex algorithm called *Random-Facet* that is also easy to describe in the geometric setting: choose a facet F incident to the current vertex uniformly at random and recursively optimize over F first. If the sink v in F is not yet the global sink, traverse the unique outgoing edge at v and repeat the process from the vertex w thus reached.

Why Randomized Simplex Algorithms?

The theory of randomized simplex algorithms is concerned with the analysis of the *worst-case* expected behavior; that is, we want bounds for the expected number of steps that hold for *all* LPs in a given class. In contrast, average case analysis [1] and smoothed analysis [2] of the simplex method are about performance bounds for deterministic simplex algorithms that hold for *random* LPs (or more strongly, for all LPs after a small random perturbation in case of smoothed analysis). The message of smoothed analysis is that the “hard” linear programs are rare. With randomized simplex algorithms, we are also fighting against the hard ones.

The main motivation to study randomized simplex algorithms is the unsolved theoretical question whether there exists a simplex algorithm whose worst-case number of steps is polynomial in n and d . All known results pertaining to deterministic simplex algorithms are discouraging. Starting with the seminal work of Klee and Minty [3], almost all known deterministic simplex algorithms have been shown to require an exponential number of steps in the worst case, on polytopes as simple as deformed cubes [4] (see *The Simplex Method and Its Complexity*). Here, randomization makes a difference: there is a randomized simplex algorithm whose expected number of steps is always *subexponential* [5,6] and therefore smaller than any bound that can currently be proved in the deterministic setting. Derandomization techniques exist [7] but are currently unable to reproduce the subexponential bounds. An easier (but also unsolved) question is whether there is any linear programming algorithm at all that runs in *strongly* polynomial time, meaning that there is a polynomial runtime bound in terms of only n and d , but not the bitsizes of the input. Such a result would considerably strengthen the classic polynomial-time results for linear programming, due to Khachiyan and Karmarkar [8,9].

In this context, a second and related open question has attracted a lot of attention. Given a polytope P with n facets in dimension d and two vertices v and w of P , is there always a short path between v and w in the vertex-edge graph of P ? By short we mean polynomial in n and d . In graph-theoretic terms, we are talking about the diameter of the polytope graph. The famous Hirsch conjecture even stipulates that the diameter is always at most $n - d$, which is the best possible bound [10]. While the Hirsch conjecture has been proved for small values of d and $n - d$, and for many special classes of polytopes, the general conjecture has recently been falsified by Santos [11]. In order for any simplex algorithm to exhibit (expected) polynomial-time worst-case behavior, there must be a polynomial bound on the diameter. Interestingly, the currently best known *quasipolynomial*

diameter bounds are obtained from an (algorithmically intractable) derandomization of the subexponential randomized simplex algorithm [12,13]. Both the subexponential runtime bound and the quasipolynomial diameter bound are obtained in the polytope digraph model, without reference to the geometry of the underlying LP.

The aforementioned results have generated quite some excitement concerning randomized simplex algorithms back in the early 1990s; however, no substantial progress has been made since then. On the contrary, there are results [14,15] showing that it may be very difficult to beat the subexponential runtime bounds without exploiting more of the underlying geometry; and, how to do this is completely unclear. For the diameter problem, the situation is very similar [16].

Today, the randomized simplex algorithms are still considered a relevant theoretical tool in the context of the two unsolved questions above, but there is no indication that the desired strong polynomiality results are anywhere nearby. Research has consequently shifted focus away from linear programming. A major virtue of most randomized simplex algorithms is that they can easily be translated to related and/or more abstract settings where much less is known and the potential for new applications is higher.

In the subsequent three sections, we will review the major results that are known for randomized simplex algorithms on LPs and other optimization problems. With very few exceptions, all of these results pertain to either *Random-Edge* or *Random-Facet*.

THE RANDOM-EDGE SIMPLEX ALGORITHM

It is fair to say that the research on the *Random-Edge* simplex algorithm started with an erroneous exercise. In their book *Combinatorial Optimization: Algorithms and Complexity* [17, Chapter 8], Papdimitriou and Steiglitz discuss the famous *Klee–Minty cube*, a d -dimensional deformed cube on which the simplex algorithm with Dantzig’s original pivot rule may take $2^d - 1$ steps.

Problem 10 of Chapter 8 (marked as “relatively difficult”) asks the reader to prove that $O(d)$ steps suffice to solve the Klee–Minty cube with *Random-Edge*.

An upper bound of $O(d^2)$ is easy to get and was first shown by Kelly [18], who also conjectured a better bound of $O(d \log^2 d)$. But this conjecture (and therefore also the statement of Problem 10 of Ref. 17, Chapter 8 was later falsified by lower bounds of $\Omega(d^2/\log d)$ [19] and $\Omega(d^2)$ [20], thus showing that the easy upper bound of $O(d^2)$ on the performance of *Random-Edge* on the Klee–Minty cube is in fact tight.

Already the analysis of *Random-Edge* on the specific class of d -dimensional Klee–Minty cubes is nontrivial; coming up with matching upper and lower bounds on general polytopes seems out of reach with known techniques. Even for the class of combinatorial cubes, currently, the best upper bound is only slightly better than the trivial bound [21].

Specific results for specific classes of polytopes are available, though. Broder *et al.* have constructed a family of three-dimensional polytopes for which (the number of facets and) the expected number of steps is exponential in the height (length of the longest shortest-directed path from any vertex to the unique sink in the polytope digraph) [22]. This means that *Random-Edge* may take very long even though there is a short-directed path to the sink from every vertex. In the meantime, the performance of *Random-Edge* on three-dimensional polytopes is reasonably well understood [23]. For the special class of dual-to-cyclic four-dimensional polytopes, Gillman provides an analysis [24].

On d -polytopes with $d + 1$ facets (simplices), *Random-Edge* is easily seen to require an expected number of $\Theta(\log d)$ steps. For d -polytopes with $d + 2$ facets, the bound is $\Theta(\log^2 d)$, but this is already considerably harder to obtain [25]. No bounds are known for the case of $d + 3$ facets.

THE RANDOM-FACET SIMPLEX ALGORITHM

The name of this algorithm is connected to a remarkable historical coincidence. Independently, Matoušek *et al.* [6] as well as

Kalai [5] managed to prove that this algorithm has *subexponential* expected performance on all LPs. Even the proofs are very similar. A function in n and d is called *subexponential* if its logarithm is sublinear in both parameters. As a consequence, *Random-Facet* is the first (and still the only) simplex algorithm with better than exponential (expected) worst-case complexity. The main idea behind the subexponential analysis is simple and reveals itself most easily on combinatorial cubes [26,27].

The subexponential bound had another remarkable consequence: A clairvoyant version of *Random-Facet* (that knows the optimal random choices) can be used to prove a *quasipolynomial* bound on the diameter of polytope graphs [12,13]. This bound is still the best known today.

GENERALIZATIONS AND OTHER RANDOMIZED PIVOT RULES

Both *Random-Edge* and *Random-Facet* can be described and analyzed in terms of polytope digraphs. Actual coordinates of the polytope (that come from the underlying LP) are not needed. This is in contrast to most other pivot rules for the simplex method that base their decisions on reduced costs or other numerical data extracted from the LP.

As a consequence, *Random-Edge* and *Random-Facet* can be (and have been) used not only for linear programming, but also for related and more general optimization problems. Even prior to the subexponential bounds [5,6], Sharir and Welzl have introduced the abstract class of *LP-type problems* [28]. This class contains all LPs (after making the above-mentioned assumptions underlying the geometric view), but a number of other relevant optimization problems as well. Examples are the problem of computing the smallest enclosing ball of a set of points in d dimensions, or the problem of finding optimal strategies in simple stochastic games [26].

The *Random-Facet* algorithm and its subexponential bound extend to all LP-type problems [6,29]. In form of the very similar randomized least-index algorithm, *Random-Facet* even works for cyclic polytope digraphs that arise for example from

P-matrix linear complementarity problems and more generally, from the abstract class of *unique sink orientations*, see Ref. 30 and the references therein. It is open whether the subexponential bound holds in the cyclic setting. Abstract problem classes are also relevant in the context of *lower* bounds for the expected worst-case performance of randomized simplex algorithms. So far, all attempts at finding “bad” LPs for either *Random-Edge* or *Random-Facet* have been unsuccessful. No available evidence contradicts the possibility that both algorithms have polynomial expected worst-case runtime. But we at least know that in order to prove such a result, a simplistic digraph model is insufficient.

Both for *Random-Facet* [14] and for *Random-Edge* [15], there are hypercube digraphs (with all properties [31,32] that are needed to derive the known upper bounds on the expected performance), such that the respective algorithm needs a superpolynomial expected number of steps. The major limitation for the relevance of these results is that the “bad” digraphs that are being constructed are known not to come from any LPs, since they violate the necessary *Holt–Klee condition* [33]. An intriguing open question is whether (and how) the simplistic digraph model, enhanced with just the Holt–Klee condition, still allows for superpolynomial lower bounds. If we admit cycles in the model, *Random-Edge* may be exponentially slow even under the Holt–Klee condition [34].

As far as other randomized simplex algorithms are concerned, the literature is sparse. The *Random-Facet* simplex algorithm has been generalized by Kalai to the *Random-Face* algorithm, and Kalai actually conjectures polynomial runtime for appropriate parameters [13]. Joswig and Kaibel have proposed two new randomized simplex algorithms and analyzed their behavior on Klee–Minty cubes [35]. A randomized simplex-type algorithm has been shown by Kelner and Spielman to require expected polynomial time [36], but with a polynomial that also involves the bitsizes of the input.

The question whether there is a strong polynomial-time algorithm for linear programming is still open.

REFERENCES

1. Borgwardt KH. The simplex method: a probabilistic analysis. Volume 1, Algorithms and combinatorics. Berlin Heidelberg: Springer; 1987.
2. Spielman DA, Teng SH. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *J ACM (JACM)* 2004;51(3):385–463.
3. Klee V, Minty GJ. How good is the simplex algorithm? In: Shisha O, editor. *Inequalities III*. New York: Academic Press; 1972. pp. 159–175.
4. Amenta N. Deformed products and maximal shadows of polytopes. Volume 999, *Advances in Discrete and Computational Geometry: Proceedings of the 1996 AMS-IMS-SIAM Joint Summer Research Conference, Discrete and Computational Geometry—Ten Years Later, 1996 July 14–18*. Mount Holyoke College: American Mathematical Society; 1998. pp. 57–99.
5. Kalai G. A subexponential randomized simplex algorithm. In: *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*. New York: Victoria, BC, Canada; 1992. pp. 475–482.
6. Matousek J, Sharir M, Welzl E. A subexponential bound for linear programming. *Algorithmica* 1996;16(4):498–516.
7. Chazelle B, Matousek J. On linear-time deterministic algorithms for optimization problems in fixed dimension. *J Algorithms* 1996;21(3): 579–597.
8. Khachiyan LG. Polynomial algorithms in linear programming. *USSR Comput Math Math Phys* 1980;20:53–72.
9. Karmarkar N. A new polynomial-time algorithm for linear programming. *Combinatorica* 1984;4(4):373–395.
10. Ziegler GM. *Lectures on polytopes*. Volume 152, *Graduate Texts in Mathematics*. New York: Springer; 1994.
11. Santos A. A counter-example to the Hirsch Conjecture. May 1994. Preprint
12. Kalai G, Kleitman DJ. A quasi-polynomial bound for the diameter of graphs of polyhedra. *Bull Am Math Soc* 1992;26:315–316.

13. Kalai G. Upper bounds for the diameter and height of graphs of convex polyhedra. *Disc Comput Geom* 1992;8(1):363–372.
14. Matoušek J. Lower bounds for a subexponential optimization algorithm. *Random Struct Algorithms* 1994;5(4):591–607.
15. Matoušek J, Szabó T. RANDOM EDGE can be exponential on abstract cubes. *Adv Math* 2006;204(1):262–277.
16. Eisenbrand F, Rothvoß T. Diameter of polyhedra: limits of abstraction. *Proceedings of the 25th Annual Symposium on Computational Geometry*. New York: ACM New; 2009. pp. 386–392.
17. Papadimitriou CH, Steiglitz K. *Combinatorial optimization: algorithms and complexity*. Englewood Cliffs (NJ): Prentice Hall; 1982.
18. Kelly DG. Some results on random linear programs. *Meth Oper Res* 1981;40:351–355.
19. Gärtner B, Henk M, Ziegler GM. Randomized simplex algorithms on Klee-Minty cubes. *Combinatorica* 1998;18(3):349–372.
20. Balogh J, Pemantle R. The Klee-Minty random edge chain moves with linear speed. *Random Struct Algorithms* 2006;30:464–483.
21. Gärtner B, Kaibel V. Two new bounds on the random-edge simplex algorithm. *SIAM J Disc Math* 2007;21:178–190.
22. Broder AZ, Dyer ME, Frieze AM, *et al.* The worst-case running time of the random simplex algorithm is exponential in the height. *Inf Process Lett* 1995;56(2):79–81.
23. Kaibel V, Mechtel R, Sharir M, *et al.* The simplex algorithm in dimension three. *SIAM J Comput* 2005;34(2):475–497.
24. Gillmann R. The random edge simplex algorithm on dual cyclic 4-polytopes. *Arxiv preprint math/0605117*. 2006.
25. Gärtner B, Solymosi J, Tschirschnitz F, *et al.* One line and n points. *Random Struct Algorithms* 2003;23(4):453–471.
26. Ludwig W. A subexponential randomized algorithm for the simple stochastic game problem. *Inf Comput* 1995;117(1):151–155.
27. Gärtner B. The Random-Facet simplex algorithm on combinatorial cubes. *Random Struct Algorithms* 2002;20(3):353–381.
28. Sharir M, Welzl E. A combinatorial bound for linear programming and related problems. Volume 577, *Proceedings 9th Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science*. Berlin: Springer; 1992. pp. 569–579.
29. Gärtner B. A subexponential algorithm for abstract optimization problems. *SIAM J Comput* 1995;24(5):1018–1035.
30. Foniok J, Fukuda K, Gärtner B, *et al.* Pivoting in linear complementarity: two polynomial-time cases. *Disc Comput Geom* 2009;42:187–205.
31. Kathy Williamson Hoke. Completely unimodal numberings of a simple polytope. *Disc Appl Math* 1988;20(1):69–81.
32. Kalai G. Linear programming, the simplex algorithm and simple polytopes. *Math Program* 1997;79(1):217–233.
33. Holt F, Klee V. A proof of the strict monotone 4-step conjecture. Volume 233, *Advances in Discrete and Computational Geometry: Proceedings of the 1996 AMS-IMS-SIAM Joint Summer Research Conference, Discrete and Computational Geometry—Ten Years Later; 1996 July 14–18*. Mount Holyoke College: American Mathematical Society; 1998. pp. 201.
34. Morris WD Jr. Randomized pivot algorithms for P-matrix linear complementarity problems. *Math Program* 2002;92(2):285–296.
35. Joswig M, Kaibel V. Randomized simplex algorithms and random cubes.
36. Kelner JA, Spielman DA. A randomized polynomial-time simplex algorithm for linear programming. *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*. New York: ACM; 2006. p. 60.

RECYCLING

ELIF AKÇALI

Department of Industrial and Systems
Engineering, University of Florida,
Gainesville, Florida

Increasing solid-waste generation rate worldwide is a growing concern for governments, environmental groups, and businesses. According to the Organization for Economic Cooperation and Development (OECD), increased prosperity, associated economic growth, and changes in consumption patterns contribute to the generation of higher rates of waste per capita. For instance, in the United States, the waste generation rate has increased from 2.68 to 4.62 pounds per day per capita from 1960 to 2007 [1]. During the same period, the recycling rate increased from about 6.4 to 24.9% in the country [1]. As industrial products are beginning to account for an increased fraction (from 11.3 to 17.9% by weight from 1960 to 2007 [1]) of the solid waste generated, the scope of the recycling industry in the United States, which was primarily limited to the recycling of glass, steel, aluminum, and paper products in the past, has been expanding to include the processing of a number of industrial durable products (e.g., appliances, cellular phones, and computers).

With the expanding scope and size of the recycling industry, operations research (OR) tools and techniques are being increasingly utilized for the optimization of recycling operations and the underlying reverse supply chain (RSC) systems. This article provides a general introduction to recycling and RSC systems and illustrates how OR methodologies can be used for the optimal design and operation of RSC systems for recycling. To this end, first, some of the issues encountered in recycling practice are highlighted, and a discussion of how these practical issues create some unique modeling challenges that necessitate the need to develop new analytical models for the design and operation of

RSC systems is presented. Second, an illustration of the use of OR techniques for the solution of an important pricing problem encountered in the context of RSC systems is provided. In this article, the term *recycling* is used in the broadest sense to include the recovery of materials and components for reuse wherever possible to avoid disposal, unlike remanufacturing (see the article titled **Remanufacturing** in this encyclopedia) that is primarily concerned with the recovery of products and components with the explicit intention to rebuild these products or reuse their components in their original design.

The remainder of this article is organized as follows. The section titled “Practical Issues in Recycling” provides a broader perspective on some of the practical issues encountered in the recycling industry. The section titled “Related Work” reviews the existing literature that address important planning and control problems that are encountered in the design and operation of RSC systems. The section titled “End-of-Life Vehicle Recycling” takes a closer look at the end-of-life vehicle recycling industry that motivates the work presented in this article. The section titled “Optimizing Price Decisions” illustrates how deterministic OR models can be utilized to address some important pricing decisions encountered in RSC systems. The section titled “Impact of Recovery Yield on Price Decisions” demonstrates how stochastic OR models can be used to investigate the influence of recovery yield uncertainty on the pricing decisions encountered in RSC systems. Finally, the last section provides a summary of the article and highlights future research directions.

PRACTICAL ISSUES IN RECYCLING

Upon a careful examination of the operations of a number of different recycling companies, several fundamental factors can be identified that influence the complexity of the

underlying RSC operations. In particular, the characteristics of the supply in the disposal market, the bill-of-material (BOM) structure of end-of-life products (ELPs), processing cost structures, and the demand in the reuse market pose different operational challenges.

Supply in the Disposal Market

The input to recycling is the supply of ELPs in the disposal market. Ideally, the collection process should control the *quantity*, *timing*, *composition*, and *quality (condition)* of the ELPs entering the RSC [2]. However, in practice, the ELPs can be received from either individual users one at a time (e.g., an individual may bring his/her used computer to a recycler) or a group of users in a batch (e.g., a large firm upgrading its entire computer network sends all used computers to the recycler). Similarly, a consumer may replace his/her cellular phone upon each airtime contract renewal, whereas another one may keep the same phone through a number of contract renewal cycles (Leff M. GRC Wireless Recycling, Inc., 2006. Personal communication). Hence, both the quantity and the timing of the ELPs entering the RSC *vary*. In addition, if a large firm upgrades its entire computer network and sells used computers to a recycler, this batch of ELPs may show little or no variation as to their make, model, age, and quality, that is, the composition of ELP supply is *homogeneous*. In contrast, a batch of products received from a third-party consolidator (e.g., a charity organization or an independent broker) may exhibit a lot of variation. For instance, Goodwill accepts all television set donations without making a distinction of the make, model, age, or quality of the set, and sells the inventory that cannot be liquidated through retail sales to recyclers. In this case, the composition of ELP supply is *heterogeneous*.

In the context of recycling, the quality of an ELP is associated with the economic value that can be recovered from it. The recoverable economic value depends on both the conditions under which the product was operated and whether any of the components of the product were removed or replaced during the course of its lifetime. For instance, an automotive engine that is driven in a cold climate

is subject to more wear and tear (due to extreme temperature changes), and hence, some of its components may not be recoverable. Therefore, the quality of the ELPs entering the RSC may exhibit a lot of *variation* and is *unknown* to the recycler until the products are disassembled.

For some product categories (e.g., cellular phones), recyclers are developing the capability to discriminate prices and selectively purchase ELPs, based on their make, model, age, and quality through product acquisition management [3,4]. As a result, *price-sensitive* markets are emerging, where the recyclers can exercise some control on the quantity, timing, composition, and quality of ELP supply.

BOM Structure of the ELP

Recycling entails the disassembly of an ELP to identify recoverable assets, and the BOM structure of the ELP has an influence on the disassembly operations used. First and foremost, BOM structure dictates the *nature* of disassembly, that is, *conserving* or *destructive*. Conserving disassembly preserves all of the components of the product, whereas destructive disassembly focuses on the recovery of certain components and may demolish some or all of the other components of the product. For instance, because of some irreversible operations (e.g., welding), disassembly may have to demolish some components of a product to recover a particular component. Similarly, BOM structure defines the *type* of the disassembly operation, that is, *dedicated* or *shared*. A dedicated operation retrieves a particular component, whereas a shared one releases multiple components simultaneously. Finally, disassembly operations may have precedence relationships that are dictated by the BOM structure of the ELP [5].

Processing Cost Structures

The nature of recycling operations as well as the characteristics of supply and demand impact the recycling cost structure. For instance, bulk material (e.g., ferrous metals) recovery processes exhibit significant differences in cost characteristics when compared to manual (e.g., cellular phone) recovery

processes. As bulk material recovery requires large amounts of input material to ensure overall equipment efficiency (Young S. CMC Recycling, 2007, personal communication), associated processing costs are largely volume dependent. Moreover, in bulk material recovery, the recycler may choose to collect a wider array of ELPs (e.g., scrap vehicles, household appliances) to stabilize the intermittent supply [6]. This, in turn, impacts the efficiency of the recovery process, that is, the same batch of material may have to be processed multiple times consecutively to improve the purity of the recovered material. Therefore, recycling costs may depend on the composition of the input stream, processing volume, and/or desired composition of the output stream.

Demand in the Reuse Market

The outputs of recycling are recovered materials and components that can be reused. The economic value associated with these outputs depends on the *quantity*, *quality*, and *timing* of recovery. That is, a recycler must ensure that the quantity of materials and components recovered must be sufficient to satisfy the demand in the reuse market. For instance, a steel manufacturer may refuse a transaction with a bulk metal recycler if the transaction involves an insufficient amount of material. Similarly, the quality of the recovered materials and components is important. For instance, a bulk metal recycler that processes plastic-coated copper wires has to take into account the trade-off between the cost of purifying the copper (i.e., stripping as much of the plastic as possible) and the revenue that he/she can collect. In addition, the timing of recovery is equally important, as it influences the revenue stream of the recycler.

The practical challenges explained above represent the distinguishing characteristics of the decision-making problems encountered in the context of RSC management for recycling. Hence, they lead to original and computationally challenging optimization problems that are not addressed by the traditional supply chain management literature. The section titled “Optimizing Price Decisions” demonstrates how deterministic OR models

can be used to determine the optimal ELP acquisition and reusable component selling prices. Later, the section titled “Impact of Recovery Yield on Price Decisions” illustrates how stochastic OR models can be utilized to investigate the impact of recovery yield rate on price decisions.

RELATED WORK

As we noted earlier, OR has made noteworthy contributions for the optimization of important design and control decisions encountered in RSC systems. Here, a brief overview of these contributions is presented. The analytical literature related to RSC systems for recycling primarily concentrate on network design, production planning, disassembly planning, and pricing problems. A detailed review of the existing work that focuses on network design, production planning, and disassembly planning is beyond the scope of this article. However, there are several papers and books that provide excellent reviews of the existing work in these areas [5,7–9].

With the increasing popularity of recycling and remanufacturing practices, ELPs are being sold and purchased as commodities and recyclers and remanufacturers are increasingly gaining market power to determine the purchase prices and exercise control on the supply quantity of used products [4]. Therefore, a recent stream of research investigates the pricing of recyclable ELPs and/or reusable recovered components. Considering settings where the supply of ELPs and/or the demand for reusable materials and components are price sensitive, several models that focus on determining the optimal prices have been proposed motivated by applications in reusable cellular phones [4], end-of-life vehicle recycling [10–12], and electronic product recycling [13]. With the exception of one study [10], where uncertain recovery yield is taken into account, the existing work focuses on the deterministic modeling of demand and/or supply [4,11–13]. In addition, while commercially available optimization software is used to obtain solutions for the proposed models [13], the majority of the work focuses on analyzing the structural properties of the optimal

solutions of the proposed models and develop efficient solution algorithms [4,10–12]. In the remainder of this article, some of these models studied by Karakayalı *et al.* [11,12], as well as Bakal and Akcali [10] are summarized to illustrate how deterministic and stochastic OR modeling and analysis techniques have been used to analyze these pricing decisions.

END-OF-LIFE VEHICLE RECYCLING

One of the complex industrial products that is processed by the recycling industry is automobiles. Each year around 15 million of vehicles reach the end of their useful service life in the United States. As more than 75% of an automobile's weight can be recycled in an economically viable manner by reusing its durable parts (e.g., engine, transmission, and catalytic converter) and recovering its reusable materials (e.g., ferrous and nonferrous metals), more than 95% of the retired vehicles are processed by automotive recyclers (i.e., parts dismantlers and scrap metal processors) [14].

Recyclers acquire end-of-life vehicles by purchasing them either at auctions or from final users who bring their vehicles for recycling [10,15]. Hence, the purchase price offered by a recycler influences the quantity of end-of-life vehicles acquired by the recycler and the recycler can increase the supply of end-of-life vehicles available by offering higher acquisition prices, that is, the supply is price sensitive. Moreover, recyclers may be willing to pay more for a two-year-old vehicle that is wrecked in a rear-end collision, since the front end of the car may still be intact from which several high-value parts can be recovered. Similarly, the recycler may be willing to pay less for a vehicle with aluminum engine and transmission, since the lighter weight castings are less durable (i.e., more susceptible to cracks), making them less likely to be remanufacturable than their traditional ferrous metal counterparts. Therefore, it is possible to classify end-of-life vehicles into multiple quality groups according to their age as well as their condition and offer different acquisition prices. Recyclers

typically perform several operations to ensure the environmentally safe disposal of end-of-life vehicles (see ***Closed-Loop Supply Chains: Environmental Impact*** for more information on the importance of environmental impact of supply chains) by removing all potentially hazardous liquids (e.g., gas, air conditioning gases, oils, coolants, and transmission and suspension fluids) and parts with hazardous materials (e.g., tires, airbags, and battery).

Then, the recycler may dismantle some parts from end-of-life vehicles for reuse or material recovery (e.g., transmission, engine, and catalytic converter). Moreover, the demand for the parts that can be remanufactured for reuse (e.g., transmission and engine) and parts that can be processed for material recovery (e.g., catalytic converter frame and facia components) can also be influenced through the selling prices. Finally, what is left of the end-of-life vehicles (i.e., the body along with parts that are not dismantled which is also known as the *hulk*) is shredded to recover reusable materials (e.g., ferrous and nonferrous metals). In this context, end-of-life vehicle recyclers may have the power to exercise control on either or both of these pricing decisions and have to characterize the acquisition prices of the end-of-life vehicles and the selling prices of the reusable automotive parts to align the supply in the disposal market with the demand in the reuse market.

Although the work reviewed in this article is primarily motivated by end-of-life vehicle recycling, the models and results summarized here can provide insights to a number of RSC systems for other complex industrial products. For instance, Aurora is engaged in disassembling processors from computers and selling them after a series of operations [16]. Similarly, GRC Wireless Recycling collects cellular phones and pagers from final users and sells them for reuse after minor modifications. (Leff M. GRC Wireless Recycling, Inc., 2006. Personal communication.) In the remainder of this article, the following generic problem is considered: given the parameters of supply functions for multiple types of quality groups of ELPs and demand

functions for multiple types of reusable components, our objective is to identify the optimal acquisition prices for the quality groups and optimal recovery quantities as well as the selling prices for the reusable components (or materials) to align the supply with the demand maximizing the net profit of the recycler.

OPTIMIZING PRICE DECISIONS

This section illustrates how deterministic OR models can be used to characterize the optimal acquisition and selling pricing decisions in RSC systems with price-sensitive ELP supply and reusable component demand markets. The following notation is used in the remainder of this section.

Sets and Indices

- $\mathcal{I}$ set of reusable components types indexed by i ; and
- $\mathcal{J}$ set of ELP quality groups indexed by j .

Parameters and Functions

- $D_i(p_i)$ demand of reusable component type i is a deterministic linear function of its selling price, that is, $D(p_i) = a_i - b_i p_i$ where $a_i, b_i \geq 0$;
- $S_j(f_j)$ supply of ELPs in quality group j is a deterministic linear function of its acquisition price, that is, $S(f_j) = \alpha_j + \beta_j f_j$ where $\alpha_j, \beta_j \geq 0$;
- c_j^d unit de-pollution cost per ELP in quality group j ;
- h_j unit salvage revenue per ELP (from which all reusable components are removed) in quality group j ;
- s_{ij} unit salvage revenue per reusable component type i recovered from an ELP in quality group j ;
- c_{ij}^p unit processing cost per reusable component type i recovered from an ELP in quality group j ;
- c_{ij}^r unit effective recovery cost associated with reusable component type i recovered from an ELP in quality group j , where $c_{ij}^r = s_{ij} + c_{ij}^p$; and

- κ_j unit effective salvage revenue associated with an ELP in quality group j where $\kappa_j = h_j + \sum_{i=1}^n s_{ij} - c_j^d$.

Decision Variables

- p_i unit selling price of the reusable component type i ;
- f_j unit acquisition price of the ELPs in quality group j ; and
- y_{ij} quantity of reusable components of type i recovered from ELPs in quality group j .

The profit maximization problem of the recycler can be modeled as follows:

$$\max \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} (p_i - c_{ij}^r) y_{ij} + \sum_{j \in \mathcal{J}} (\kappa_j - f_j)(\alpha_j + \beta_j f_j) \quad (1)$$

$$\text{subject to } y_{ij} \leq \alpha_j + \beta_j f_j \quad i \in \mathcal{I}, j \in \mathcal{J} \quad (2)$$

$$\sum_{j \in \mathcal{J}} y_{ij} = a_i - b_i p_i \quad i \in \mathcal{I} \quad (3)$$

$$- \alpha_j / \beta_j \leq f_j; \quad p_i \geq 0; \quad y_{ij} \geq 0 \quad i \in \mathcal{I}, j \in \mathcal{J}. \quad (4)$$

Note that objective function (1) maximizes the net profit of the recycler, which is given by the sum of the selling revenue obtained by selling reusable components, the salvaging revenue obtained by salvaging reusable components, and the salvaging revenue obtained by salvaging ELPs, that is, $\sum_{i \in \mathcal{I}} (p_i - c_{ij}^d, c_{ij}^p) (a_i - b_i p_i) + \sum_{j \in \mathcal{J}} (h_j - c_j^d - f_j) (\alpha_j - \beta_j f_j) + \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} s_{ij} y_{ij} - (a_i - b_i p_i)$. Constraint set (2) ensures that the supply of ELPs collected in a particular quality group is larger than the quantity of reusable components recovered from that group. Constraint set (3) ensures that the demand for a reusable component type is equal to the total quantity of components recovered from all quality groups. Constraint set (4) defines the feasible values for the decision variables.

To consider the simplest problem environment with homogeneous supply (i.e., a single ELP quality group) and a single type

of reusable component, indices i and j can be dropped. In this case, it can be shown that Equation (1) is concave in p and f , and a closed-form solution can easily be characterized using its first-order optimality conditions as follows:

$$(f^*, p^*) = \begin{cases} \left(\frac{\kappa}{2} - \frac{\alpha}{2\beta}, \frac{c^r}{2} + \frac{a}{2b} \right) & \text{if } \frac{a-bc^r}{2} < \frac{\alpha+\beta\kappa}{2} \\ \left(\frac{a-\alpha+b(\kappa-c^r)}{2(\beta+b)} - \frac{\alpha}{2\beta}, \frac{a-\alpha-\beta(\kappa-c^r)}{2(\beta+b)} + \frac{a}{2b} \right) & \text{otherwise.} \end{cases}$$

Another problem setting of interest is concerned with homogenous supply and multiple types of reusable components. In this case, again it can be shown that Equation (1) is concave in f and p_i and the optimal prices can be determined in polynomial time. In addition, for the problem setting with heterogeneous supply (i.e., multiple ELP quality groups) and one type of reusable component, a polynomial time algorithm can be developed to determine the optimal prices. Finally, the problem setting with heterogeneous supply and multiple types of reusable components yields a concave quadratic maximization problem. An algorithm that exploits the structural properties of an optimal solution to the problem that converges quickly in practice can be developed [11,12]. In addition, see the section titled “Constrained Optimization” in this encyclopedia for more information on how to develop solution algorithms for constrained optimization problems.

It should also be noted that these models can easily be extended to decentralized settings where the collection and recovery activities can be performed by different agents. These extended models can be used to investigate when and why an OEM would outsource the collection or the recycling activity, that is, prefer a collector- or a recycler-driven channel, in the presence of environmental regulations that specify target ELP collection rates for the OEMs [11].

IMPACT OF RECOVERY YIELD ON PRICE DECISIONS

An important issue that further complicates the acquisition and selling pricing decisions

is the recovery yield. In practice, the fraction of reusable components that can be recovered for reuse from ELPs is typically random and may be dependent on the acquisition price in some cases. This section illustrates how stochastic OR tools can be used to investigate the impact of random recovery yield rate on acquisition and selling pricing decisions in RSC systems with price-sensitive supply and demand markets. See the article titled **Modeling Uncertainty in Optimization Problems** in this encyclopedia for more information on how to incorporate uncertainty into optimization problems.

To this end, two cases can be considered both of which are motivated from real-life applications end-of-life vehicle recycling where the yield rate is random. Some salvagers purchase end-of-life vehicles at auctions. Once recoverable components are harvested, and yield is realized from the supply of end-of-life vehicles, the recycler sets the price and sells the reusable automotive parts. In this case, the recycler has the opportunity to adjust the selling price for the reusable automotive part to match the supply with the demand. In some cases, however, a part buyer contacts the recycler requesting a particular part and the recycler and the buyer agree on the sales price. Then, the recycler purchases end-of-life vehicles to satisfy the demand. If the yield is not enough to cover the demand, then the recycler is still bound by the price decision made *a priori*, and may have to incur a penalty for the unsatisfied demand in addition. In this case, the recycler does not have the opportunity to adjust the sales price after realizing the yield.

Let R be a random variable that represents the maximum attainable recovery yield rate and r a realization of R . Also, let $g(r)$ and $G(r)$ denote the probability distribution function and cumulative distribution function of random variable R , respectively. Moreover, as the recycler increases the acquisition price, ELPs in better condition becomes available to the recycler, which in turn can be interpreted as a higher probability of recovering a reusable component from the ELP. Let this probability of recovering a reusable component be represented by $t(f)$, which is a concave, nondecreasing function of f and

converges to 1 as f increases. Then, for the two cases discussed above, following models can be formulated:

Postponed Pricing Model (PPM)

In this setting, the recycler can postpone the decision to characterize the selling price for the reusable component until after the realization of the recovery rate of the ELPs. More specifically, the pricing decisions are made in two stages (see the section titled “Two-stage Stochastic Programming” in this encyclopedia). First, the acquisition price of the ELPs is identified and the corresponding supply is collected. After the realization of the recovery rate, selling price for the reusable component is determined, and demand is observed in the second stage. Note that the recycler never faces lost sales for the reusable component; demand for the reusable component can be aligned with the available supply by adjusting the selling price accordingly.

First, the second-stage decision is analyzed. Given the acquisition price, f , and recovery yield rate realization, r , the second-stage problem for the recycler reduces down to determining the selling price for the reusable component to maximize the total net profit. Then, the profit maximization problem can be formulated as follows:

$$\max (p - c^r)(a - bp) + (\kappa - f)(\alpha + \beta f) \quad (5)$$

$$\text{subject to } a - bp \leq rt(f)(\alpha + \beta f). \quad (6)$$

Given the acquisition price and recovery yield rate realization, the optimal selling price for the reusable component can be characterized as follows:

$$p^* = \begin{cases} \frac{a}{2b} + \frac{c^r}{2} & r \geq \frac{\ell}{t(f)(\alpha + \beta f)} \\ \frac{a - rt(f)(\alpha + \beta f)}{b} & r < \frac{\ell}{t(f)(\alpha + \beta f)} \end{cases}, \quad (7)$$

where $\ell = (a - bc^r)/2$.

Now, the first-stage decision can be analyzed. Given the optimal selling price for the reusable component obtained from the analysis of the second stage and the realization of

the recovery yield rate, the profit maximization problem faced by the recycler in the first stage can be formulated as follows:

$$\begin{aligned} \Pi(f|R=r) &= \begin{cases} \frac{\ell^2}{b} + (\alpha + \beta f)(\kappa - f), \\ r \geq \frac{\ell}{t(f)(\alpha + \beta f)}; \\ t(f)r(\alpha + \beta f)\frac{2\ell - t(f)r(\alpha + \beta f)}{b} \\ + (\alpha + \beta f)(\kappa - f), & r < \frac{\ell}{t(f)(\alpha + \beta f)}. \end{cases} \end{aligned}$$

Taking expectation with respect to R yields

$$\begin{aligned} E[\Pi(f)] &= \int_0^{\ell/(t(f)(\alpha + \beta f))} t(f)r(\alpha + \beta f) \\ &\times \frac{2\ell - t(f)r(\alpha + \beta f)}{b} g(r) dr \\ &+ \int_{\ell/(t(f)(\alpha + \beta f))}^1 \frac{\ell^2}{b} g(r) dr + (\alpha + \beta f)(\kappa - f). \end{aligned} \quad (8)$$

The objective function (8) is concave in the acquisition price for the ELPs, and its optimal value can be determined numerically. That is, the recycler determines the optimal acquisition price for the ELPs optimizing the expected profit function (8) and collects the resulting supply of ELPs. After the realization of the random recovery yield rate, selling price of the reusable components can be characterized using Equation (7).

Simultaneous Pricing Model (SPM)

In this case, the recycler has to determine the acquisition and selling prices simultaneously. That is, both acquisition and selling prices must be specified before the realization of the random recovery yield rate. Hence, the recycler may not be able to cover the demand for the reusable components entirely. Therefore, the profit function can be modified to include a term to account for the unsatisfied demand. Let v denote the unit penalty cost for unsatisfied demand (or the unit cost of supply from an outside source to satisfy the demand). Given any realization of the random recovery yield, the profit function of the

recycler is given by

$$\Pi(f, p|r) = \begin{cases} (p - c^r)(\alpha - bp) + (\kappa - f)(\alpha + \beta f), & r \geq \frac{a-bp}{t(f)(\alpha+\beta f)}; \\ t(f)r(\alpha + \beta f)(p - c^r) - v(a - bp - t(f)r(\alpha + \beta f)), & r < \frac{a-bp}{t(f)(\alpha+\beta f)}. \\ + (\alpha + \beta f)(\kappa - f), \end{cases}$$

Taking expectation with respect to R and rearranging the terms yields

$$\begin{aligned} E[\Pi(f, p)] &= \int_0^{\frac{a-bp}{t(f)(\alpha+\beta f)}} [rt(f)(\alpha + \beta f)(p - c^r + v) \\ &\quad - v(a - bp)] g(r) dr \\ &\quad + \int_{\frac{a-bp}{t(f)(\alpha+\beta f)}}^1 (a - bp)(p - c^r) g(r) dr \\ &\quad + (\alpha + \beta f)(\kappa - f). \end{aligned} \quad (9)$$

It can be shown that Equation (9) is pseudoconcave in f and p separately. However, it is analytically challenging to show that the Hessian of Equation (9) is strictly positive. Upon extensive numerical analysis with a variety of distribution functions, it can be conjectured that Equation (9) is pseudoconcave. This allows for to use first-order conditions of Equation (9) to obtain a local maximum and treat it as the global optimum in the remainder of our analysis.

Using these analytical results for PPM and SPM, it is possible to investigate the influence of recovery yield rate on the pricing decisions of a recycler. Results from a comprehensive computational analysis reveal that this influence critically depends on the profit margin associated with the reusable component. If the hulk and component salvage value for the component is high, which can be referred to as a *high-margin component*, then the recycler would not make any effort into increasing the recovery yield rate, if it is beyond a certain threshold. However, if the hulk and component salvage values of the component are low, which can be referred to as a *low-margin component*, then the recycler's profit would increase by increasing the recovery yield rate. Moreover, it is possible

to observe that the quantities of ELPs acquired and reusable components sold are higher for high-margin components under both pricing models. A comparison of the pricing models further reveals that as PPM always yields higher profits than SPM, as it provides the recycler with the opportunity to adjust the selling price after the realization of the random recovery yield, that is, suffer less from the effects of random recovery yield. Furthermore, there are no significant differences in expected profit under PPM and SPM for high-margin components, and the difference diminishes as the mean yield rate increases. In addition, profits generated in both models decrease as the standard deviation of the yield rate increases, and the benefits of PPM over SPM become more pronounced as the standard deviation increases. Finally, for low-margin components, the expected profit under PPM is higher than that under SPM, and the difference does not diminish as the mean yield rate increases. Hence, postponing the pricing decision is always more beneficial for the recycler, and the expected benefit is higher for lower yield rates, higher yield variability, and lower margin components. Using SPM, it is also possible to investigate the value of partial or perfect recovery yield rate information [10].

CONCLUDING REMARKS

This article illustrated how OR models can be utilized for the solution of important planning and control problems encountered in RSC systems for recycling. To this end, a brief introduction of some of the issues encountered in recycling practice that influence the complexity of the planning and control problems is provided. Specifically, a discussion of the characteristics of the supply in the disposal markets, the bill-of-materials structure of the ELPs, the processing cost structures, and the demand in the reuse markets is presented. Then, motivated by end-of-life vehicle recycling, a relatively simple setting is considered, where the supply in the disposal market can be modeled as a deterministic linear function of the acquisition price of the

ELP, the ELP has a simple BOM structure, the processing costs can be modeled as linear function of the quantity of ELPs processed, and the demand in the reuse market is also a deterministic linear function of the selling price of the reusable component. First, some deterministic OR models are developed to determine the optimal ELP acquisition and reusable component selling prices in RSC systems. Then, some simple stochastic OR models are developed to investigate the yield uncertainty on the pricing decision encountered in RSC systems. Future research in this area that takes the practical issues discussed in the section titled "Related Work" into account explicitly may have significant impact on some of the many challenging problems that arise in the context of RSC management.

REFERENCES

1. United States Environmental Protection Agency Office of Solid Waste. Municipal solid waste in the United States: 2007 facts and figures. www.epa.gov/epawaste/nonhaz/municipal/pubs/msw07-rpt.pdf, Accessed 2008.
2. Guide VDR. Production planning and control for remanufacturing: Industry practice and research needs. *J Oper Manage* 2000;18: 467–483.
3. Guide VDR, Van Wassenhove LN. Managing product returns for remanufacturing. *Prod Oper Manage* 2001;10(2):142–155.
4. Guide VDR, Teunter RH, Van Wassenhove LN. Matching supply and demand to maximize profits from remanufacturing. *Manuf Serv Oper Manage* 2003;5(4):303–316.
5. Lambert AJD, Gupta SM. Disassembly modeling for assembly, maintenance, reuse and recycling. Boca Raton: CRC Press; 2005.
6. White CD, Masanet E, Rosen CM, Beckman SL. Product recovery with some byte: An overview of management challenges and environmental consequences in reverse manufacturing for the computer industry. *J Clean Prod* 2003;11:445–458.
7. Akçalı E, Çetinkaya S, Üster H. Network design for reverse and closed-loop supply chains: an annotated bibliography of models and solution approaches. *Networks* 2009; 53(3):231–248.
8. Kim HJ, Lee DH, Xirouchakis P. Disassembly scheduling: literature review and future research directions. *Int J Prod Res* 2007; 45(18-19):4465–4484.
9. Lambert AJD. Disassembly sequencing: a survey. *Int J Prod Res* 2003;41(16):3721–3759.
10. Bakal IS, Akçalı E. Effects of random yield in remanufacturing with price-sensitive supply and demand. *Prod Oper Manage* 2006;15(3): 407–420.
11. Karakayalı I, Emir-Farinas H, Akçalı E. Analysis of decentralized collection and processing of end-of-life product. *J Oper Manage* 2007;25(6):1161–1183.
12. Karakayalı I, Emir-Farinas H, Akçalı E. Pricing and recovery planning for demanufacturing operations with multiple used products and multiple reusable components. *Comp Ind Eng* 2010;59(1):55–63.
13. Vadde S, Kamarth SV, Gupta SM. Optimal pricing of reusable and recyclable components under alternative product acquisition mechanisms. *Int J Prod Res* 2007;45(18-19): 4621–4652.
14. Pomykala JA, Jody BJ, Daniels EJ, Spangenberger JS. Automotive recycling in the united states: Energy conservation and environmental benefits. *J Miner Met Mater Soc* 2007;59(11):41–45.
15. Qu X, Williams JAS. An analytical model for reverse automotive production planning and pricing. *Eur J Oper Res* 2008;190:756–767.
16. Thierry M, Salomon M, van Nunen J, Wassenhove LNV. Strategic issues in product recovery management. *Calif Manage Rev* 1995;37(2):114–135.

REDUCTION OF A POMDP TO AN MDP

BURHANEDDIN SANDIKÇI
Booth School of Business,
University of Chicago,
Chicago, Illinois

A Markov decision process (MDP) is a standard mathematical tool for modeling, analyzing, and ultimately solving many sequential decision-making problems under uncertainty (see “*Introduction to MDPs*” for an introduction to MDPs). Many excellent texts [1–4] discuss MDP theory and its applications in a wide variety of settings in great detail. One of the key assumptions in an MDP model is that the *state* of the underlying process is *perfectly* known to the decision maker at all decision epochs. However, in many real-life situations, the nature of the situation being modeled may not allow the decision maker to completely observe the state of the underlying process or acquiring perfect state information may be too costly, in which case, a completely observable MDP model would no longer apply. Relaxing the perfect state information assumption in an MDP model yields to what is known as a “*partially observable Markov decision process (POMDP)*” model (also see ***Partially Observable MDPs (POMDPs): Introduction and Examples***”).

Examples of application domains where POMDPs are utilized include machine maintenance and replacement [5–10], inventory control [11], inspection of structural units such as paved roads, bridges, and buildings [12], network troubleshooting [13], search for a moving object [14,15], medical diagnosis and treatment [16–19], health care systems analysis [20], optimal design of questionnaires when the answers are not truthful [21], optimal teaching strategies [22], and fisheries [23]. We refer the reader to [24–26] for detailed reviews of POMDP applications.

The seemingly innocent relaxation of the assumption of perfect state information to

noise-corrupted state information comes at a significant cost in complexity. The so-called *curse of dimensionality* [1], which can be summarized as the exponential increase in the amount of resources (data, time, computing memory, and so on) needed to incorporate an extra dimension to the model, is a major computational burden in using the dynamic programming based approaches (see “*Dynamic Programming*” for deterministic dynamic programs) in many settings. This curse coupled with the stochastic elements in an MDP model further restricts the solvability of such models in practice.

Analyzing the structural properties of an MDP model, which oftentimes includes proving the existence of structured optimal policies (and value functions), is crucial to accelerate the solution speed and therefore to partially overcome this computational burden (also see “*Structured MDP Policies*”). Structural analyses of POMDP models, however, are usually much harder than their MDP counterparts, exactly because of the limited observability of the state of the process. As a result, in general, there is little hope to solve POMDP models through exploiting special structures. (Of course, if found, some special structures may help solve certain problems, but the main difficulty is identifying them.)

The exact algorithms [26–33] for POMDPs have not yet proved extremely useful for solving practical problems. The current literature reports solutions of POMDP models with only a few states (typically <10) using one of the exact algorithms. Given the scarcity of efficient exact methods to solve POMDPs, several approximation methods have been developed. Surveys of exact and approximate algorithms for POMDPs are available in Refs 26, 34, and 35.

Grid-based approximations are the most natural and widely used techniques to find approximate solutions to POMDPs. These techniques sample a finite number of points, called the *grid*, from the entire state space, compute the (approximate) value of the

points in the grid using only the points in the grid and approximate the values of the non-grid points (i.e., those points that are in the state space but not selected to the grid) from the computed grid points' values via some form of interpolation. Several variations exist in the literature [19, 36–39]. Grid-based approximations usually do not have any performance guarantees, therefore they should be viewed as heuristics. In this article, we summarize main results associated with approximating POMDP problems via (variants of) completely observable MDPs.

The rest of this article is organized as follows. We begin with a brief review of completely observable MDPs and continue with an analogous overview of POMDPs. Next, we discuss the main difficulties in solving POMDP models and compare how the completely observable MDP approximates the optimal value function of a POMDP model. Furthermore, we provide a concise description of grid-based approximation methods for POMDPs and compare the grid-based values to the exact values and show that grid-based values provide better estimates than the values obtained by completely ignoring partial observability. Finally, we also show how linear programming may be used to obtain tightest possible grid-based approximations.

MARKOV DECISION PROCESSES

We consider a sequential decision-making problem under *uncertainty*, where a decision maker can *control* the flow of the system at *discrete time* points $1, \dots, T$, called *decision epochs*. Note that T need not be finite. At any given decision epoch t , the system occupies a *core state* i that takes values from the *core state space* S . We assume that the actual state of the system, i , is completely known to the decision maker when a decision is to be made. We relax this assumption in the next section. The decision maker has control over the system by modifying its trajectory through choosing an *action* a from the *action space* $\mathcal{A}$ at each decision epoch t . We assume that S and $\mathcal{A}$ are both discrete finite sets. (We could generalize this setup to allow time-dependent state and action spaces, namely,

S_t and $\mathcal{A}_t$. However, this would not change the basic results presented in this article, hence we choose to drop the time index t from these sets for notational simplicity.) Uncertainty in the system is governed by its probabilistic evolution; that is, when the system is in state $i \in S$ and the decision maker chooses action $a \in \mathcal{A}$ at decision epoch t , the state of the system in the next decision epoch $t + 1$ becomes $j \in S$ with probability $p_{ij}^{a,t} \equiv P(j|i, a)$. The fact that the next state of the system is only dependent on the current state and action is called the *Markovian property* of the process. Furthermore, as a result of choosing action $a \in \mathcal{A}$ while the system is in state $i \in S$ at time t , the decision maker receives an *immediate reward* $r_t(i, a)$. The collection of objects $(T, S, \mathcal{A}, P\{\cdot|\cdot, \cdot\}, r(\cdot, \cdot))$ is referred to as a *completely observable Markov decision process*.

The objective of the decision maker is to identify a decision-making *policy* that maximizes the *expected total discounted reward* over the course of the entire process

$$\mathbb{E} \left[\sum_{t=0}^{T-1} \lambda^t r_t(i, a) \right], \quad (1)$$

where $\lambda \in [0, 1)$ is the discount rate. Other optimality criteria such as maximizing the expected total reward (see “*Total Expected Reward Criterion*”) or maximizing the average reward (see “*Average Reward Criterion*”) can also be handled.

For an infinite-horizon problem ($T = \infty$), a stationary model (i.e., a model in which the parameters do not change over time) is a standard assumption in the literature. Following this assumption, we drop the time index t for the infinite-horizon models presented in this article. An optimal policy that maximizes the expectation in Equation (1) can be found by solving the following set of recursive equations [3]:

$$u^*(i) = \max_{a \in \mathcal{A}} \left\{ r(i, a) + \lambda \sum_{j \in S} p_{ij}^a u^*(j) \right\} \quad \text{for } i \in S, \quad (2)$$

where $u^*(i)$ is the optimal MDP value function in state $i \in S$. The optimal action

$a^*(i)$ for state $i \in S$ is chosen as the argument $a \in \mathcal{A}$ that maximizes the right-hand side of Equation (2). *Value iteration* (see **Total Expected Discounted Reward MDPs: Value Iteration Algorithm**), *policy iteration* and its variants (see **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**) are typical exact algorithms used to solve the optimality equations (Eq. 2).

For a finite-horizon problem ($T < \infty$), the analog of Equation (2) can be written as

$$u_t^*(i) = \max_{a \in \mathcal{A}} \left\{ r_t(i, a) + \lambda \sum_{j \in S} p_{ij}^{a,t} u_{t+1}^*(j) \right\} \quad \text{for } i \in S; \quad t = 1, \dots, T-1, \quad (3)$$

with boundary conditions

$$u_T^*(i) = r_T(i) \quad \text{for } i \in S, \quad (4)$$

where $u_t^*(i)$ is the optimal MDP value function in state $i \in S$ when there are $T-t$ decision epochs left until the end of the horizon and $r_T(i)$ represents the terminal reward received when the process is in state i at the end of the planning horizon. The optimal action $a_t^*(i)$ at each decision epoch $t = 1, \dots, T$ for state $i \in S$ is chosen as the argument $a \in \mathcal{A}$ that maximizes the right-hand side of Equation (3). The *backward induction* algorithm provides an efficient and exact method for solving the optimality equations (Eq. 3).

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

We consider the same problem described in the previous section. However, we now relax the completely observable system state assumption. That is, we allow partial observability that yields imperfect state information. As a result, the system does not directly reveal its actual state $j \in S$, rather the decision maker observes a signal k , which takes values from the *observation set* $\mathcal{Z}$ and relates to the core state $j \in S$ through the *observation probability* $q_{jk}^{a,t} \equiv Q\{k|j, a\}$. More precisely, the observation probability $q_{jk}^{a,t}$ denotes the probability that the decision maker receives signal k at time

$t+1$ while the system actually occupies state j at time $t+1$ after action a was taken at time t . We assume that $\mathcal{Z}$ is a discrete finite set. The collection of objects $\langle T, S, \mathcal{A}, P\{\cdot|\cdot, \cdot\}, r(\cdot, \cdot), \mathcal{Z}, Q\{\cdot|\cdot, \cdot\} \rangle$ is referred to as a *partially observable* Markov decision process. Note that all but the last two objects in this collection are identically shared with completely observable MDPs.

It would no longer be useful to write the optimality equations for this new system using the core states, because they are not observable. However, we can write the optimality equations using a new concept called *belief state* and still possess the Markovian property, because a belief state is known to be a *sufficient statistic* for the entire history of the process [40,41]. This result implies that the probability of being in a particular belief state at time t is only dependent on which belief state was occupied at time $t-1$ and not on any of the previous belief states, hence the Markovian property holds. A belief state π is defined to be a probability distribution over the set of core states S , where the i th component of π , denoted by π_i , represents the probability of being in core state $i \in S$. Let $|S| = m$ and $\Pi(S)$ be the set of all probability distributions over S (also called the $(m-1)$ -dimensional probability simplex or sometimes referred to as the *belief simplex*), that is,

$$\Pi(S) = \left\{ \pi \in \mathbb{R}^m : \sum_{i=1}^m \pi_i = 1, \quad \pi_i \geq 0 \quad \forall i \right\}.$$

Note that there are infinitely many probability distributions that could be defined over S when $m \geq 2$. Therefore, a POMDP model defined over belief states has infinitely many states, which causes serious computational challenges.

Also note that a belief state POMDP model is equivalent to a completely observable MDP model with a continuous state space. We could, in theory, use the techniques available for continuous state MDPs to solve belief state POMDPs. However, it is usually more efficient to take advantage of the special

structure of the state space, which is not an arbitrary continuous space but rather it is a space composed only of discrete probability distributions.

As we will write the optimality equations using belief states, we need to know how to make transitions among belief states. For this purpose, let π be the current belief state and assume that action $a \in \mathcal{A}$ is taken while in this belief. As a result of this action, during the next decision epoch, the system occupies a (potentially new) *core* state, which is not known to the decision maker with certainty because it is not observable. Instead, the decision maker observes a signal $k \in \mathcal{Z}$ for this new decision epoch and needs to update her prior belief π using the information that just became available. The j th component of the updated belief state π' , denoted by π'_j , can be computed by [24]

$$\pi'_j = \frac{\sum_{i \in \mathcal{S}} \pi_i p_{ij}^a q_{jk}^a}{\sum_{i \in \mathcal{S}} \sum_{\ell \in \mathcal{S}} \pi_i p_{i\ell}^a q_{\ell k}^a}, \quad (5)$$

which is derived using Bayes' formula and simple probability arguments (see, for example, Ref. 26 for a detailed derivation of this formula.) Note that although π' depends on π , k , and a , we suppress this dependency on the left-hand side of Equation (5) for notational clarity. Also note that we drop the time index t from Equation (5) for notational clarity.

For an infinite-horizon problem ($T = \infty$), an optimal policy that maximizes the expectation in Equation (1) can be found by solving the following set of recursive equations [24,26]:

$$\begin{aligned} v^*(\pi) &= \max_{a \in \mathcal{A}} \left\{ \rho(\pi, a) + \lambda \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{Z}} \pi_i p_{ij}^a q_{jk}^a v^*(\pi') \right\} \\ &\text{for } \pi \in \Pi(\mathcal{S}), \end{aligned} \quad (6)$$

where $v^*(\pi)$ is the optimal POMDP value function for belief state $\pi \in \Pi(\mathcal{S})$ and $\rho(\pi, a) = \sum_{i \in \mathcal{S}} \pi_i r(i, a)$. The optimal action

$a^*(\pi)$ for state $\pi \in \Pi(\mathcal{S})$ is chosen as the argument $a \in \mathcal{A}$ that maximizes the right-hand side of Equation (6).

For a finite-horizon problem ($T < \infty$), the analog of Equation (6) can be written as [26,42]

$$\begin{aligned} v_t^*(\pi) &= \max_{a \in \mathcal{A}} \left\{ \rho_t(\pi, a) \right. \\ &\quad \left. + \lambda \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \sum_{k \in \mathcal{Z}} \pi_i p_{ij}^{a,t} q_{jk}^{a,t} v_{t+1}^*(\pi') \right\} \\ &\text{for } \pi \in \Pi(\mathcal{S}); \quad t = 1, \dots, T-1, \end{aligned} \quad (7)$$

with boundary conditions

$$v_T^*(\pi) = \sum_{i \in \mathcal{S}} \pi_i r_T(i) \quad \text{for } \pi \in \Pi(\mathcal{S}), \quad (8)$$

where $v_t^*(\pi)$ is the optimal value function in state $\pi \in \Pi(\mathcal{S})$ when there are $T - t$ decision epochs left until the end of the horizon. The optimal action $a_t^*(\pi)$ at each stage $t = 1, \dots, T$ for state $\pi \in \Pi(\mathcal{S})$ is chosen as the argument $a \in \mathcal{A}$ that maximizes the right-hand side of Equation (7).

APPROXIMATING A POMDP BY AN MDP

As we have seen in the previous section, a POMDP is a generalization of a completely observable MDP with an enlarged state space called the *belief space*, that is, the space of all probability distributions over the set of core states. Although this belief space allowed us to nicely recast the POMDP as a completely observable MDP, it is the source of the main difficulty when solving a POMDP model, because the dynamic programming recursions needed to solve a belief state POMDP is equivalent to having infinitely many equations with infinite number of variables. Evaluating the optimality equations over the entire belief space, therefore, suffers from computational tractability. Indeed, the current state of the art in the POMDPs can exactly evaluate these optimality equations only for problems with very few core states (typically considered toy problems) [31,43].

Approximation methods have been developed to partially overcome the computational challenges for POMDPs in exchange for forsaken optimal solutions. Grid-based approximations are the most frequently encountered methods in the literature. Grid-based approaches, as we will see below, have no performance guarantees in general, therefore they should be viewed as *heuristic* approximations at best and one-sided bounds obtained from such approaches should be treated with caution. Throughout the rest of this article, we focus on infinite-horizon problems for notational simplicity, but similar results can be easily written for finite-horizon problems as well.

Approximating with a Completely Observable MDP

We start with the simplest approximation method that can be viewed as a special type of grid-based technique. Ignoring the partial observability inherent in the system and solving the POMDP problem as if its states are completely observed is the simplest of all approximations and not an infrequent one in many real problems. In this approach, one recursively solves Equations (2) to obtain the optimal MDP value functions $u^*(i)$ for each core state $i \in \mathcal{S}$. For a given belief state $\pi \in \Pi(\mathcal{S})$, one can estimate the optimal POMDP value function $v^*(\pi)$ by

$$\hat{u}(\pi) = \sum_{i \in \mathcal{S}} \pi_i \cdot u^*(i). \quad (9)$$

When the belief state π is chosen to be one of the extreme points of the belief simplex, we denote it by the i th unit vector $\mathbf{e}^i$, which assigns 1 to its i th position and 0 elsewhere. Such a belief state is essentially equivalent to occupying core state i in a completely observable system. For any $i \in \mathcal{S}$, using $u^*(i)$ as an approximation for $v^*(\mathbf{e}^i)$ overestimates the true value [19,44], that is $v^*(\mathbf{e}^i) \leq u^*(i)$. Therefore, the approximation given in Equation (9) would overestimate the true value for arbitrary belief states, that is

$$v^*(\pi) \leq \hat{u}(\pi) \quad \text{for } \pi \in \Pi(\mathcal{S}). \quad (10)$$

Intuitively, inequality (10) implies that we cannot do any worse if we had access to complete state information and therefore the difference $\hat{u}(\pi) - v^*(\pi)$ can be interpreted as the value of accessing full information from the current partial information.

Approximating with a General Grid

In this section, we address the partial information availability by forming a grid composed of a finite number of sampled belief states from the entire belief simplex. Given a list of grid points, we recursively solve equations similar to Equation (6) adjusted for the finite number of states (or, alternatively, similar to Equation (2) adjusted for partial observability). We still need to resolve an additional difficulty: because of the finite sample, during the solution process, a point in the grid may be updated to a point not included in the grid by the Bayesian updating formula given by Equation (5). In such a case, the value associated with the nongrid point is estimated by some form of interpolation of the values of the grid points.

As a general rule, the limited (computational) resources should be wisely devoted to potentially useful belief points. In other words, one should refrain from including points that are highly unlikely to be reached during the solution process, because the contribution of such points would be negligibly small when taking the expectation in the recursive equations. Furthermore, assume that all the corner points of the probability simplex are already included in the grid. Note that there would be m corner points in an $(m - 1)$ -dimensional probability simplex. Let $\mathcal{G}$ denote the entire grid and $|\mathcal{G}| = n \geq m$.

Beyond these assumptions, there are several ways of forming and managing the grid. We briefly mention the most common of these approaches and note that our discussion below is independent of the grid management approach (except the fact that it must include all the corner points of the simplex.) A *fixed-resolution regular grid* samples points from the belief simplex in regular intervals (i.e., equidistant points in each dimension of the simplex) and does not vary these points throughout the rest of the solution process [36]. The advantage of this

method is its efficient interpolation, whereas the obvious disadvantage is the exponential growth in the size of the grid as the number of core states or the resolution of the grid increase [19,39,44]. A *variable-resolution irregular grid* samples points from the belief simplex in irregular intervals and adds/removes to/from the grid over iterations as perceived necessary by heuristic rules [37,38]. The advantage of this method is its more economic use of grid points in approximating the value function, whereas its disadvantage is inefficient interpolation times. A *variable-resolution regular grid* [39] and a *fixed-resolution irregular grid* [19] are additional methods that try to take advantage of the strong aspects of the previous two methods. Further details on grid-based methods can be found in Refs 34, 39, and 44 and the references therein.

The Approximation Function. We now derive the recursive equations to be solved with this approach. Because all the corner points of the belief simplex are included in grid $\mathcal{G}$, any belief point π from the simplex can be written as a convex combination of the points in $\mathcal{G}$, that is,

$$\pi = \sum_{\ell=1}^n \beta_\ell \mathbf{b}^\ell, \quad (11)$$

where $\mathbf{b}^\ell \in \mathcal{G}$ and $\beta_\ell \geq 0$ for $\ell = 1, \dots, n$, and $\sum_{\ell=1}^n \beta_\ell = 1$.

We compute the grid-based approximate value function $\hat{v}(\cdot)$ associated with a belief state $\mathbf{b} \in \mathcal{G}$ using a modified version of the POMDP optimality equations (Eq. 6). When we write the optimality Equation (6) for a given belief state $\mathbf{b} \in \mathcal{G}$, the problem will arise on the right-hand side as we will need the value associated with the updated belief vector, denoted by $\mathbf{b}'$, which is not known if $\mathbf{b}' \notin \mathcal{G}$. However, we can estimate the value associated with $\mathbf{b}'$ using a convex combination of the known values of the grid points $\mathbf{b} \in \mathcal{G}$. That is, replacing $v^*(\mathbf{b}')$ on the right-hand

side of Equation (6) with a convex interpolation estimate $\sum_{\ell=1}^n \beta_\ell \hat{v}(\mathbf{b}^\ell)$ yields the grid-based approximation

$$\begin{aligned} \hat{v}(\mathbf{b}) = \max_{a \in \mathcal{A}} & \left\{ \rho(\mathbf{b}, a) \right. \\ & \left. + \lambda \sum_{i \in S} \sum_{j \in S} \sum_{k \in \mathcal{Z}} b_i p_{ij}^a q_{jk}^a \sum_{\ell=1}^n \beta_\ell \hat{v}(\mathbf{b}^\ell) \right\} \\ & \text{for } \mathbf{b} \in \mathcal{G}. \end{aligned} \quad (12)$$

It should be noted here that some researchers use heuristic rules such as rounding or using the K nearest neighboring grid points (for some arbitrary positive integer K) to find the value of a nongrid point. Unfortunately, such heuristics do not work as intended in general when $m \geq 3$. The main reason for the failure of such heuristics is the fact that the nongrid point under consideration is not guaranteed to be spanned by the grid points found by such rules.

It can be shown that [19] the grid-based approximation given in Equation (12) yields identical results to the solution of the completely observable MDP for the corner points of the belief simplex (i.e., $\hat{v}(\mathbf{e}^i) = u^*(i)$ for $i \in S$) and that [19,44] the grid-based value also overestimates the true value, that is

$$v^*(\mathbf{b}) \leq \hat{v}(\mathbf{b}) \quad \text{for } \mathbf{b} \in \mathcal{G}. \quad (13)$$

More importantly, the grid-based value provides a better upper bound for the true value than the completely observable MDP value [19], that is

$$\hat{v}(\mathbf{b}) \leq \hat{u}(\mathbf{b}) \quad \text{for } \mathbf{b} \in \mathcal{G}. \quad (14)$$

It can further be shown that inequalities (Eqs 13 and 14) hold for $\pi \in \Pi(S) \setminus \mathcal{G}$. The difference $\hat{u}(\mathbf{b}) - \hat{v}(\mathbf{b})$ can be interpreted as the value of addressing partial observability, whereas the difference $\hat{v}(\mathbf{b}) - v^*(\mathbf{b})$ can be interpreted as the loss associated with not fully addressing the partial observability.

Bayesian Updating for the Grid. We have noted earlier that for a finite number of grid points, the Bayesian updating of a point does not guarantee the updated point to be in the grid. We have represented the updated belief ($\mathbf{b}'$) as a convex combination of the points in the grid, but we did not specify what the β_ℓ scalars would be.

We need at most m linearly independent vectors to represent any vector in an m -dimensional space. Since our grid, by construction, includes all the extreme points of the belief simplex, which span the entire space, we can always use these points to represent any updated belief $\mathbf{b}' \notin \mathcal{G}$. However, there may be other points in our grid that span the updated belief $\mathbf{b}'$ while, at the same time, yielding better approximate values (better in the sense that is closer to $v^*(\mathbf{b})$). The following linear program would actually find the set of grid points that provide the *best* approximate value while guaranteeing the representation of $\mathbf{b}'$ [19,39]:

$$\begin{aligned} & \text{minimize} && \sum_{\ell=1}^n \hat{v}(\mathbf{b}^\ell) \cdot \beta_\ell \\ & \text{subject to} && \sum_{\ell=1}^n \mathbf{b}^\ell \cdot \beta_\ell = \mathbf{b}', \\ & && \sum_{\ell=1}^n \beta_\ell = 1, \\ & && \beta_\ell \geq 0 \quad \text{for } \ell = 1, \dots, n. \end{aligned}$$

Note that the constraint set guarantees that the updated belief point $\mathbf{b}'$ is spanned by the selected set of grid points. Also note that minimizing the objective function guarantees the best approximate values because of inequality (13). Finally, observe that in order to identify the correct β -weights, this linear programming problem must be solved at every iteration of the solution algorithm (because the values $\hat{v}(\mathbf{b})$ may change from iteration to iteration) for each grid point for each observation and for each action (because each (grid point, observation, action) triplet may yield a different updated belief point). As a result, although each linear programming problem might be solved extremely fast, the number of programs to solve may be enormous.

SUMMARY AND CONCLUSIONS

This article reviews the completely and partially observable Markov decision process models and compares how the optimal value functions of both models relate to each other. Main computational challenges for POMDPs are discussed. The article also derives the grid-based approximation methods for POMDPs, shows how linear programming may be used to derive best possible approximations from these methods, and shows that, at best, grid-based MDPs overestimate true optimal values for POMDPs.

REFERENCES

1. Bellman RE. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
2. Howard R. Dynamic programming and Markov processes. Cambridge (MA): MIT Press; 1960.
3. Puterman ML. Markov decision processes: Discrete stochastic dynamic programming. New York: John Wiley & Sons, Inc.; 1994.
4. Bertsekas DP. Dynamic programming and optimal control. Volumes 1 and 2, 3rd ed. Belmont (MA): Athena Scientific; 2007.
5. Eckles JE. Optimum maintenance with incomplete information. Oper Res 1968;16(5):1058–1067.
6. Ross SM. Quality control under Markovian deterioration. Manage Sci 1971;17(9):587–596.
7. White CC. A Markov quality control process subject to partial observation. Manage Sci 1977;23(8):843–852.
8. White CC. Optimal inspection and repair of a production process subject to deterioration. J Oper Res Soc 1978;29(3):235–243.
9. White CC. Optimal control-limit strategies for a partially observed replacement problem. Int J Syst Sci 1979;20:321–331.
10. Maillart LM. Maintenance policies for systems with condition monitoring and obvious failures. IIE Trans 2006;38:463–475.
11. Treharne JT, Sox CR. Adaptive inventory control for nonstationary demand and partial information. Manage Sci 2002;48(5):607–624.
12. Ellis H, Jiang M, Corotis RB. Inspection, maintenance, and repair with partial observability. J Infrastruct Syst 1995;1(2):92–99.

13. Thiebaux S, Cordier M-O, Jehl O, *et al.* Supply restoration in power distribution systems: A case study in integrating model-based diagnosis and repair planning. Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-1996); Portland (OR). 1996. pp. 525–532.
14. Pollock SM. A simple model of search for a moving target. *Oper Res* 1970;18:883–903.
15. Eagle JN. The optimal search for a moving target when the search path is constrained. *Oper Res* 1984;32(5):1107–1115.
16. Hu C, Lovejoy WS, Shafer SL. Comparison of some suboptimal control policies in medical drug therapy. *Oper Res* 1993;44(5):696–709.
17. Hauskrecht M, Fraser H. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artif Intell Med* 2000;18:221–244.
18. Maillart LM, Ivy J Simmons, Ransom S, Diehl K. Assessing dynamic breast cancer screening policies. *Oper Res* 2008;56(6):1411–1427.
19. Sandıkçı B. Estimating the price of privacy in liver transplantation [PhD thesis]. University of Pittsburgh; 2008.
20. Smallwood RD, Sondik EJ, Offensend FL. Toward an integrated methodology for the analysis of health-care systems. *Oper Res* 1971;19(6):1300–1322.
21. White CC. Optimal diagnostic questionnaires which allow less than truthful responses. *Infor Control* 1976;32:61–74.
22. Smallwood RD. The analysis of economic teaching strategies for a simple learning model. *J Math Psychol* 1971;8:285–301.
23. Lane DE. A partially observable model of decision making by fishermen. *Oper Res* 1989;37(2):240–254.
24. Monahan GE. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Manage Sci* 1982;28(1):1–16.
25. Cassandra AR. Acting optimally in partially observable stochastic domains. Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97); Providence (RI). 1997. pp. 472–480.
26. Cassandra AR. Exact and approximate algorithms for partially observable Markov decision processes [PhD thesis]. Brown University; 1998.
27. Sondik EJ. The optimal control of partially observable Markov processes [PhD thesis]. Stanford University; 1971.
28. Smallwood RD, Sondik EJ. The optimal control of partially observable Markov processes over a finite horizon. *Oper Res* 1973;21(5):1071–1088.
29. Sondik EJ. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Oper Res* 1978;26(2):282–304.
30. Cheng H-T. Algorithms for partially observable Markov decision processes [PhD thesis]. University of British Columbia; 1988.
31. Hansen EA. An improved policy iteration algorithm for partially observable MDPs. 10th Neural Information Processing Systems Conference; Denver (CO); 1997.
32. Cassandra AR, Kaelbling LP, Littman ML. Acting optimally in partially observable stochastic domains. Proceedings of the 12th National Conference on Artificial Intelligence; Seattle (WA). 1994. pp. 1023–1028.
33. Cassandra AR, Littman ML, Zhang NL. Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes. Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97); Providence (RI). 1997.
34. Lovejoy WS. A survey of algorithmic methods for partially observed Markov decision processes. *Ann Oper Res* 1991;28:47–66.
35. Zhang W. Algorithms for partially observable Markov decision processes [PhD thesis]. The Hong Kong University of Science and Technology; 2001.
36. Lovejoy WS. Computationally feasible bounds for partially observed Markov decision processes. *Oper Res* 1991;39(1):162–175.
37. Brafman R. A heuristic variable grid solution method for POMDPs. Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97); Providence (RI). 1997.
38. Hauskrecht M. Incremental methods for computing bounds in partially observable Markov decision processes. Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97); Providence (RI). 1997.
39. Zhou R, Hansen EA. An improved grid-based approximation algorithm for POMDPs. Proceedings of the 17th International Conference on Artificial Intelligence; Seattle (WA). 2001.
40. Åström KJ. Optimal control of Markov processes with incomplete state information. *J Math Anal Appl* 1965;10:174–205.
41. Striebel CT. Sufficient statistics in the optimal control of stochastic systems. *J Math Anal Appl* 1965;12:576–592.

- 42. Lovejoy WS. Some monotonicity results for partially observed Markov decision processes. *Oper Res* 1987;35(5):736–743.
- 43. Lovejoy WS. Suboptimal policies, with bounds, for parameter adaptive decision processes. *Oper Res* 1993;41(3):583–599.
- 44. Hauskrecht M. Value-function approximations for partially observable Markov decision processes. *J Artif Intell Res* 2000;13:33–94.

REFLECTED BROWNIAN MOTION

A. B. DIEKER

Georgia Institute of Technology,
Atlanta, Georgia

This article is concerned with a class of d -dimensional Markov processes called *reflected Brownian motions* (RBMs), which are also known as *reflecting Brownian motions* or *regulated Brownian motions*. Attention is restricted from the outset to the case of primary interest in operations research/management science (OR/MS), where the state space of the RBM is the non-negative orthant $S = \mathbf{R}_+^d = \{x \in \mathbf{R}^d : x_1 \geq 0, \dots, x_d \geq 0\}$. The data of the RBM are a d -vector μ , a $d \times d$ covariance matrix Σ , and a $d \times d$ reflection matrix R that satisfies certain restrictions to be specified later.

The behavior of the process can be informally described as follows. In the interior of S , the process behaves like a Brownian motion with covariance matrix Σ and drift μ . Whenever the process touches the boundary ∂S of S and is about to escape from S , it receives an instantaneous push toward the interior of S with a force that is just right in order to keep the process in S . When recording the total push (displacement) received by the process during any time interval through a measure, this measure is singular with respect to Lebesgue measure due to the irregularity of Brownian paths.

The pushing directions on the boundary ∂S of the orthant are determined by the matrix R . Along the relative interior of the $(d-1)$ -dimensional face of points with a vanishing i th coordinate, the pushing direction is constant and is given by the i th column of R . At the intersection of several $(d-1)$ -dimensional faces, any convex combination of pushing directions of the intersecting faces may be used to constrain the process to $\mathbf{R}_+^d$. (This requirement is sometimes relaxed, but it is in force for all the RBMs in this paper.)

Without loss of generality, we follow the convention that the diagonal of R only contains ones.

All vectors in this article are column vectors. For $v \in \mathbf{R}^d$, we write $v \geq 0$ and $v > 0$ if all elements of v are nonnegative and strictly positive, respectively. We also write M' for the transpose of a matrix M . We let $C^d[0, \infty)$ be the set of continuous functions on $[0, \infty)$ with values in $\mathbf{R}^d$, and $C_S^d[0, \infty)$ the subset consisting of $x \in C^d[0, \infty)$ with $x(0) \in S = \mathbf{R}_+^d$.

An account of the developments on the theory of RBM until 1995 is given in the survey by Williams [1]. Since this article aims to discuss the foundations of RBM, some overlap with the material discussed by Williams is inevitable. Even though limited by space restrictions, we have strived to highlight the developments on the theory of RBM since 1995.

CONSTRUCTION OF RBM

This section discusses the mathematical construction of RBM. To successfully construct RBM, one needs conditions on the pushing mechanism to prevent the process from being absorbed on a lower-dimensional face such as the origin. Several approaches can be taken to construct RBM, and we primarily focus on a pathwise construction since it is the closest to the above informal description of RBM. Since we shall see that it is unclear if RBM can always be constructed pathwise, we also discuss methods different from the pathwise construction.

Pathwise Construction

Let a d -dimensional continuous path $x = \{x(t) : t \geq 0\}$ with $x(0) \in S$ be given, from which we aim to construct a path $z = \{z(t) : t \geq 0\}$, which is constrained to $S = \mathbf{R}_+^d$. We interpret x as the “free” path that determines the movements of z while it is in the interior of the state space S .

Definition 1. Let R and $x \in \mathcal{C}_S^d[0, \infty)$ be given. Then $(z, y) \in \mathcal{C}_S^d[0, \infty) \times \mathcal{C}^d[0, \infty)$ is said to be a solution of the *Skorokhod problem* for x if $z(0) = x(0)$ and if we have

- (i) $z(t) = x(t) + Ry(t)$ for all $t \geq 0$,
- (ii) $z(t) \geq 0$ for all $t \geq 0$,
- (iii) for $i = 1, \dots, d$, y_i is nondecreasing and $y(0) = 0$,
- (iv) $\sum_{i=1}^d \int_0^\infty \mathbf{1}_{\{z_i(t) > 0\}} dy_i(t) = 0$.

If (z, y) solves the Skorokhod problem for x , then we write $z \in \Gamma(x)$, and we refer to Γ as the *Skorokhod map*. This map may be multivalued in general.

The interpretation of $Ry(t)$ is the total push that is needed during the time interval $[0, t]$ to keep the path x in $S = \mathbf{R}_+^d$. In conjunction with (i), requirement (iii) ensures that the process can only be pushed in the directions given by the columns of the reflection matrix R . Equation (iv) states that z can only be pushed in an appropriate direction when it is on the boundary of the orthant, so that z moves freely (i.e., without pushing) in the interior of S . Equation (iv) also ensures that the pushing mechanism is “minimal” in the sense that the path cannot be pushed into the interior when it hits the boundary. Only in a few cases can one give “explicit” expressions for y (or equivalently z), a notable example being $d = 1$ and $S = \mathbf{R}_+$ with $R > 0$, in which case $y(t) = -(0 \wedge \inf_{0 \leq s \leq t} x(s)/R)$.

If the reflection matrix R is an M-matrix [2], that is, if it is of the form $I - Q$, where Q is nonnegative with largest absolute eigenvalue less than 1, then there is a unique solution to the Skorokhod problem for any $x \in \mathcal{C}_S^d[0, \infty)$ as shown by Harrison and Reiman [3]. Letting $\{X(t) : t \geq 0\}$ be a d -dimensional Brownian motion starting from $x \in S$ with constant drift $EX(1) - x = \mu$ and constant covariance matrix $\text{Cov}X(1) = \Sigma$, under this assumption on R , the first component of $\Gamma(X)$ becomes a process $Z = \{Z(t) : t \geq 0\}$ starting from x that is confined to the orthant. The resulting process is RBM in S starting from x with reflection matrix R , covariance matrix Σ , and drift μ . However, since the condition on R is not always satisfied in applications, one cannot always construct RBM in this way. The

Harrison–Reiman condition on R has been put in a wider context by Dupuis and Ishii [4] and Dupuis and Ramanan [5].

Two major obstacles for a pathwise construction of RBM are the possible nonexistence and nonuniqueness of the solution to a Skorokhod problem. Clearly, one can only have existence of the solution to a Skorokhod problem with $x(0)$ on the intersection of several $(d - 1)$ -dimensional faces if the process can be pushed onto a higher-dimensional face from there. This condition is known as the *completely-S* condition on the reflection matrix R [6], and immediately translates into the following matrix formulation. The matrix R is *completely-S* if each principal submatrix $\tilde{R}$ of R is an *S*-matrix, where the latter means that there exists some $\tilde{x} \geq 0$ such that $\tilde{R}\tilde{x} > 0$. Bernard and El Kharroubi [7] show that there exists a solution to the Skorokhod problem for any $x \in \mathcal{C}_S^d[0, \infty)$ if and only if R is *completely-S*. According to Williams [1], this result has independently been established by Mandelbaum and Van der Heyden [8]. However, even under the *completely-S* condition, the solution may not be unique since the Skorokhod map is multivalued in general [7,9]. Thus, this existence result cannot be used to construct RBM in the generality that one wishes. A formulation based on the Extended Skorokhod Map, introduced by Ramanan [10], allows one to use a pathwise construction for a class of RBMs that includes some non-SRBMs (semimartingale reflected Brownian motion), but the details fall outside of the current discussion. Lipschitz continuity of the (extended) Skorokhod map has emerged as a critical property for a successful pathwise construction of RBM, see Dupuis and Ramanan [5].

SRBM and Beyond

Unless RBM can be synthesized in a strong sense (pathwise), its construction relies on tools from stochastic calculus such as martingales and their variants, see the section titled “Martingales” in this encyclopedia for background on some of these tools. Semimartingales constitute a particularly convenient class of processes from this point of view, and the semimartingale framework can

be used to construct a process known as the *SRBM* in a “weak” sense.

The key difference between the strong (pathwise) and the weak constructions is that the weak approach constructs a probability space and a stochastic process constrained to the nonnegative orthant, while the strong approach synthesizes sample paths on the given probability space on which a Brownian motion is defined.

A triple $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ is called a *filtered space* if Ω is a set, $\mathcal{F}$ is a σ -field of subsets of Ω , and $\{\mathcal{F}_t\}$ is an increasing family of sub- σ -fields of $\mathcal{F}$, that is, a filtration. The following definition is taken from Ref. 11.

Definition 2. An SRBM in the orthant associated with the data (μ, Σ, R) is a continuous, $\{\mathcal{F}_t\}$ -adapted d -dimensional process Z , taking values in $S = \mathbf{R}_+^d$ together with a family of probability measures $\{\mathbb{P}_x : x \in \mathbf{R}_+^d\}$ defined on some filtered space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\})$ for each $x \in \mathbf{R}_+^d$, under $\mathbb{P}_x$, $Z(t) = X(t) + RL(t) \geq 0$ for all $t \geq 0$, where

- (i) X is a d -dimensional Brownian motion with drift vector μ and covariance matrix Σ , $\{X(t) - \mu t : t \geq 0\}$ is a martingale with respect to $\{\mathcal{F}_t\}$, and $X(0) = 0$ $\mathbb{P}_x$ -almost surely,
- (ii) L is an $\{\mathcal{F}_t\}$ -adapted, d -dimensional process such that $L(0) = 0$, L_i is continuous and nondecreasing for $i = 1, \dots, d$, and $\sum_{i=1}^d \int_0^\infty \mathbf{1}_{\{Z_i(s) > 0\}} dL_i(s) = 0$, all $\mathbb{P}_x$ -almost surely.

Relying in part on the results of Reiman and Williams [6], Taylor and Williams [11] show that an SRBM in the nonnegative orthant exists if and only if R is completely- S , and that it is then unique in law and defines a Feller continuous strong Markov process. Note that every SRBM is indeed a semimartingale, since RL is of bounded variation. We refer to Kang and Williams [12] for the construction of SRBM in more general piecewise smooth domains.

One can sometimes construct RBM through a so-called *submartingale problem* [13], but the resulting process may not be a semimartingale. As a consequence, one may need

special tools such as advanced stochastic calculus to study them, see Kang and Ramanan [14]. The submartingale approach itself requires machinery that is outside the scope of this article.

POSITIVE RECURRENCE AND STABILITY

This section discusses positive recurrence of SRBM as defined in Definition 2. Some authors use positive recurrence and stability interchangeably in view of the relation between RBM and stochastic networks. Following Dupuis and Williams [15], we say that an SRBM Z is a *positive recurrent* if for each closed set A in S having positive Lebesgue measure, we have $\mathbf{E}_x[\tau_A] < \infty$, for all $x \in S$, where $\tau_A = \inf\{t \geq 0 : Z(t) \in A\}$ and $\mathbf{E}_x$ denotes expectation under $\mathbb{P}_x$. The next theorem has served as an inspiration for introducing the paradigm of fluid models for studying stochastic networks, see the article titled **Fluid Models of Queueing Networks** in this encyclopedia. Under special conditions on the reflection matrix, this theorem has been refined further by Budhiraja and Dupuis [16].

Theorem 1 [15]. Let Z be a d -dimensional SRBM in $\mathbf{R}_+^d$ associated with the data (μ, Σ, R) for some matrix R with the completely S property.

Assume that the z -component of all solutions of the Skorokhod problem for paths of the form $x(t) = x(0) + \mu t$, for any $x(0) \in \mathbf{R}_+^d$, is attracted to the origin, implying that, for all $\epsilon > 0$, there exists some $T < \infty$ such that $|z(t)| < \epsilon$ for $t \geq T$. Then Z is a positive recurrent and it has a unique stationary distribution.

It has proven challenging to find necessary and sufficient conditions on R and μ for positive recurrence. The conditions

$$R \text{ is nonsingular, } R^{-1}\mu < 0$$

are necessary for positive recurrence [17,18]. They are also sufficient if R satisfies the Harrison–Reiman condition [19], but need not be sufficient otherwise [18]. Hobson

and Rogers [20] formulated necessary and sufficient conditions in the two-dimensional case, see also Ref. 21. The same has recently been achieved for the intrinsically more delicate three-dimensional case, see El Kharroubi *et al.* [22] and Bramson *et al.* [23]. Necessary and sufficient conditions for higher-dimensional SRBMs are currently not known. A by-product of the results in Ref. 23 is that, in three dimensions, the RBM is a positive recurrent if and only if every fluid path is attracted to the origin in the sense of Theorem 1.

THE STATIONARY DISTRIBUTION

This section discusses several aspects of the stationary distribution of SRBM, provided it exists. We focus on analytic formulas for the stationary density, numerical techniques, and large deviations (rough tail asymptotics).

Whenever it exists, the defining property of an invariant measure π is that, for any bounded Borel function ϕ on $\mathbf{R}_+^d$,

$$\int_{x \in S} \mathbf{E}_x[\phi(Z(t))]\pi(\mathrm{d}x) = \int_{x \in S} \phi(x)\pi(\mathrm{d}x).$$

Throughout this section, we assume that we work with a positive recurrent SRBM with a unique stationary distribution and a completely- S reflection matrix. Arguments from Harrison and Williams [19], applied to this more general setting, then show that the SRBM admits a stationary density p_0 with respect to the Lebesgue measure.

The basic adjoint relationship (BAR) [19] is a key tool for studying the stationary distribution for general (S)RBM. It informally gives rise to an elliptic partial differential equation (PDE) with mixed (or oblique) boundary conditions. It remains a challenge to understand the exact role of the nonnegativity constraint in the context of BAR or the associated PDE [17].

Product Forms

Some of the most well-studied stochastic networks admit a so-called product form stationary distribution, see the section titled “Product-Form Queueing Networks” in

this encyclopedia, for examples. Thus, it is natural to investigate when an SRBM in $\mathbf{R}_+^d$ has a stationary density of the form

$$p_0(x) = \prod_{j=1}^d p_j(x_j),$$

where each of the p_j is a density relative to the Lebesgue measure on $[0, \infty)$. Densities of this form are called *product form densities*. The following result has first been established in Ref. 19 under a slightly more restrictive assumption, which can readily be relaxed as noted in Ref. 17.

Theorem 2. *Assume that R is completely- S . A positive recurrent SRBM in $\mathbf{R}_+^d$ associated with data (μ, Σ, R) has a product form stationary density if and only if $-R^{-1}\mu > 0$ and $2\Sigma = R\Lambda + \Lambda R'$, where Λ is a diagonal matrix with the same diagonal entries as Σ . In this case, $p_0(x) = Ce^{-(\eta, x)}$ for $x \in \mathbf{R}_+^d$, where $\eta = -2\Lambda^{-1}R^{-1}\mu$ and $C = \prod_{i=1}^d \eta_i$.*

When the covariance matrix Σ is the identity matrix, the product form condition from Theorem 2 becomes the so-called *skew-symmetry* condition which has been shown to characterize exponential stationary distributions for SRBMs in general polyhedral domains by Harrison and Williams [24,25]. It is an interesting fact that Brownian product-form networks may arise as the heavy-traffic limit of nonproduct form networks, see Harrison and Williams [26]. In that work, the authors show that a product form for SRBM can be interpreted in terms of quasireversibility when R corresponds to a Brownian feedforward (multiclass) network; see the article titled in this encyclopedia for background.

Beyond Product Forms

Little is known regarding the stationary density of SRBM if it fails to be of product form. In some cases, however, the corresponding PDE with mixed boundary conditions can be solved. As a result, the stationary density has been found in seemingly isolated cases, which are currently mostly limited to a two-dimensional setting. The first and, arguably,

“most isolated” case is a specific SRBM given by Harrison [27]. The stationary density in Harrison’s example is found by reformulating the corresponding PDE in polar coordinates and then separating variables.

A relatively wide class of nonproduct-form RBMs is investigated by Knessl and Morrison [28]. They work in the two-dimensional orthant and assume that the covariance matrix is the identity matrix. Much in the spirit of classical results on PDEs with mixed boundary conditions, the resulting stationary densities are expressed as certain integrals and they involve special functions. Similar results have been obtained by Foddy [29] and Harrison *et al.* [30].

Sometimes, the stationary density of a two-dimensional SRBM can be written as a sum of several exponential terms. Foschini [31] and, independently, Karpelevich and Kreinin [32] provide two classes of SRBMs with this property. Through connections with a multidimensional reflection principle for (unconstrained) Brownian motion, Dieker and Moriarty [33] recently unified these results and characterized the class of SRBMs with sum-of-exponential stationary densities in terms of a condition reminiscent of the (product-form) skew-symmetry condition. (Although the results in Ref. 33 are formulated for a wedge, they readily translate into results for the two-dimensional nonnegative orthant through a linear transformation.)

A final isolated case is included in the results of Debicki *et al.* [34], who find the Laplace transform of the stationary distribution of a Brownian motion with a rank-one covariance matrix, which is confined to the orthant. Strictly speaking, owing to the singular covariance matrix, this result falls outside of the RBM framework.

Numerical Techniques

In the absence of analytical expressions, one needs to resort to numerical techniques in order to use RBM for performance analysis. Dai and Harrison [17] introduce a projection method for numerically finding solutions to the BAR, while Chen and Shen [35] investigate a “finite element” method based on earlier work for SRBM in a hypercube [36].

Performance metrics of solutions to the BAR, such as moments of the marginals, can also be found numerically using linear programming; see Schwerer [37] for details.

Large Deviations

The large deviation theory allows one to make powerful statements on the exponential rate at which probabilities converge to zero; see the article titled in this encyclopedia for background. It has been used to study the tail behavior of positive recurrent SRBMs. These tail probabilities indeed decay exponentially under the conditions of Theorem 1, as can be immediately deduced from the fact that SRBM in the nonnegative orthant has a finite moment-generating function in a neighborhood of the origin, that is, there exists some $\varepsilon > 0$ such that for $\theta \in \mathbf{R}^d$ with $\|\theta\| < \varepsilon$,

$$\int_{x \in S} e^{(\theta, x)} \pi(dx) < \infty.$$

This result is due to Budhiraja and Lee [38].

We say that the function $I : \mathbf{R}_+^d \rightarrow [0, \infty]$ is a *rate function* if its level sets $\{x \in \mathbf{R}_+^d : I(x) \leq \alpha\}$ are compact.

Definition 3 [LDP for SRBM]. Suppose an SRBM in $\mathbf{R}_+^d$ has a unique stationary distribution π . We say that it satisfies an LDP with rate function I if for any $A \subset \mathbf{R}_+^d$,

$$\begin{aligned} -\inf_{x \in A^\circ} I(x) &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \pi(nA) \\ &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \pi(nA) \leq -\inf_{x \in \bar{A}} I(x), \end{aligned}$$

where A° and $\bar{A}$ denote the interior and closure of A respectively.

Majewski [39] proves that an LDP holds for a positive recurrent SRBM with either a Harrison–Reiman or a triangular reflection matrix R . Dupuis and Ramanan [40] widen this class and give an alternative “time-reversed” representation of the rate function.

Evaluating the rate function amounts to solving a variational problem over a certain space of absolutely continuous paths. The solution to this problem is currently only

well understood for the two-dimensional non-negative orthant, see Avram *et al.* [41] and Harrison and Hasenbein [21].

Acknowledgments

The author gratefully acknowledges NSF grant EEC-0926308. He is also thankful to Ruth Williams and Kavita Ramanan for valuable suggestions.

REFERENCES

- Williams RJ. Semimartingale reflecting Brownian motions in the orthant. Volume 71, Stochastic Networks, the IMA Volumes in Mathematics and its Applications. New York: Springer; 1995. pp. 125–137.
- Berman A, Plemmons RJ. Nonnegative matrices in the mathematical sciences. New York: Academic Press; 1979.
- Harrison JM, Reiman MI. Reflected Brownian motion on an orthant. *Ann Probab* 1981;9:302–308. DOI: 10.1214/aop/1176994471.
- Dupuis P, Ishii H. SDEs with oblique reflection on nonsmooth domains. *Ann Probab* 1993;21(1):554–580. DOI: 10.1214/aop/1176989415.
- Dupuis P, Ramanan K. Convex duality and the Skorokhod problem. I, II. *Probab Theory Relat Fields* 1999;115:153–195, 197–236. DOI: 10.1007/s004400050269.
- Reiman MI, Williams RJ. A boundary property of semimartingale reflecting Brownian motions. *Probab Theory Relat Fields* 1988;77:87–97. DOI: 10.1007/BF01848132.
- Bernard A, El Kharroubi A. Régulations déterministes et stochastiques dans le premier “orthant” de $\mathbf{R}^n$. *Stoch Stoch Rep* 1991;34:149–167. DOI: 10.1080/17442509108833680.
- Mandelbaum A, van der Heyden L. Complementarity and reflection. 1987. Unpublished manuscript.
- Mandelbaum A. The dynamic complementarity problem. 1989. Unpublished manuscript.
- Ramanan K. Reflected diffusions defined via the extended Skorokhod map. *Electron J Probab* 2006;11:934–992.
- Taylor LM, Williams RJ. Existence and uniqueness of semimartingale reflecting Brownian motions in an orthant. *Probab Theory Relat Fields* 1993;96:283–317. DOI: 10.1007/BF01292674.
- Kang W, Williams RJ. An invariance principle for semimartingale reflecting Brownian motions in domains with piecewise smooth boundaries. *Ann Appl Probab* 2007;17:741–779. DOI: 10.1214/10505160600000899.
- Varadhan SRS, Williams RJ. Brownian motion in a wedge with oblique reflection. *Commun Pure Appl Math* 1985;38:405–443. DOI: 10.1002/cpa.3160380405.
- Kang W, Ramanan K. A Dirichlet process characterization of a class of multidimensional reflected diffusions. *Ann Probab* 2010;38:1062–1105. DOI: 10.1214/09-AOP487.
- Dupuis P, Williams RJ. Lyapunov functions for semimartingale reflecting Brownian motions. *Ann Probab* 1994;22:680–702. DOI: 10.1214/aop/1176988725.
- Budhiraja A, Dupuis P. Simple necessary and sufficient conditions for the stability of constrained processes. *SIAM J Appl Math* 1999;59:1686–1700. DOI: 10.1137/S0036139997330222.
- Dai JG, Harrison JM. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Ann Appl Probab* 1992;2:65–86. DOI: 10.1214/aoap/1177005771.
- El Kharroubi A, Ben Tahar A, Yaacoubi A. Sur la récurrence positive du mouvement brownien réfléchi dans l’orthant positif de $\mathbf{R}^n$. *Stoch Stoch Rep* 2000;68:229–253. DOI: 10.1080/17442500008834224.
- Harrison JM, Williams RJ. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics* 1987;22:77–115. DOI: 10.1080/17442508708833469.
- Hobson DG, Rogers LCG. Recurrence and transience of reflecting Brownian motion in the quadrant. *Math Proc Camb Philos Soc* 1993;113:387–399. DOI: 10.1017/S0305004100076040.
- Harrison JM, Hasenbein JJ. Reflected Brownian motion in the quadrant: tail behavior of the stationary distribution. *Queueing Syst* 2009;61:113–138. DOI: 10.1007/s11134-008-9102-9.
- El Kharroubi A, Ben Tahar A, Yaacoubi A. On the stability of the linear Skorokhod problem in an orthant. *Math Methods Oper Res* 2002;56:243–258. DOI: 10.1007/s001860200210.

23. Bramson M, Dai J, Harrison M. Positive recurrence of reflecting Brownian motion in three dimensions. *Ann Appl Probab* 2010;20:753–783. DOI: 10.1214/09-AAP631.
24. Harrison JM, Williams RJ. Multidimensional reflected Brownian motions having exponential stationary distributions. *Ann Probab* 1987;15:115–137. DOI: 10.1214/aop/1176992259.
25. Williams RJ. Reflected Brownian motion with skew symmetric data in a polyhedral domain. *Probab Theor Relat Fields* 1987;75:459–485. DOI: 10.1007/BF00320328.
26. Harrison JM, Williams RJ. Brownian models of feedforward queueing networks: quasireversibility and product form solutions. *Ann Appl Probab* 1992;2:263–293. DOI: 10.1214/aop/1177005704.
27. Harrison JM. The diffusion approximation for tandem queues in heavy traffic. *Adv Appl Probab* 1978;10:886–905. DOI: 10.2307/1426665.
28. Knessl C, Morrison JA. Heavy traffic analysis of two coupled processors. *Queueing Syst* 2003;43:173–220. DOI: 10.1023/A:1022816410813.
29. Foddy M. Analysis of Brownian motion with drift, confined to a quadrant by oblique reflection [PhD thesis]. Stanford University; 1983.
30. Harrison JM, Landau HJ, Shepp LA. The stationary distribution of reflected Brownian motion in a planar region. *Ann Probab* 1985;13:744–757. DOI: 10.1214/aop/1176992906.
31. Foschini GJ. Equilibria for diffusion models of pairs of communicating computers —symmetric case. *IEEE Trans Inf Theory* 1982;28:273–284. DOI: 10.1109/TIT.1982.1056473.
32. Karpelevich FL, Kreinin AY. Two-phase queueing system $GI/G/1 \rightarrow G/1/\infty$ under heavy traffic conditions. *Theor Probab Appl* 1981;26:293–313. DOI: 10.1137/1126030.
33. Dieker AB, Moriarty J. Reflected Brownian motion in a wedge: sum-of-exponential stationary densities. *Electron Commun Probab* 2009;14:1–16.
34. Dębicki K, Dieker AB, Rolski T. Quasi-product forms for Lévy-driven fluid networks. *Math Oper Res* 2007;32:629–647. DOI: 10.1287/moor.1070.0259.
35. Chen H, Shen X. Computing the stationary distribution of an SRBM in an orthant with applications to queueing networks. *Queueing Syst* 2003;45:27–45. DOI: 10.1023/A:1025691717137.
36. Shen X, Chen H, Dai JG, *et al.* The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks. *Queueing Syst* 2002;42(1):33–62. DOI: 10.1023/A:1019942711261.
37. Schwerer E. A linear programming approach to the steady-state analysis of reflected Brownian motion. *Stoch Models* 2001;17:341–368. DOI: 10.1081/STM-100002277.
38. Budhiraja A, Lee C. Long time asymptotics for constrained diffusions in polyhedral domains. *Stoch Process Appl* 2007;117:1014–1036. DOI: 10.1016/j.spa.2006.11.007.
39. Majewski K. Large deviations of the steady-state distribution of reflected processes with applications to queueing systems. *Queueing Syst Theor Appl* 1998;29:351–381. DOI: 10.1023/A:1019148501057.
40. Dupuis P, Ramanan K. A time-reversed representation for the tail probabilities of stationary reflected Brownian motion. *Stoch Process Appl* 2002;98:253–287. DOI: 10.1016/S0304-4149(01)00151-X.
41. Avram F, Dai JG, Hasenbein JJ. Explicit solutions for variational problems in the quadrant. *Queueing Syst* 2001;37:259–289. DOI: 10.1023/A:1011004620420.

REFORMULATION-LINEARIZATION TECHNIQUE FOR MIPs

HANIF D. SHERALI
Grado Department of Industrial
and Systems Engineering,
Virginia Polytechnic Institute
and State University,
Blacksburg, Virginia

Mixed-integer programming (MIP) problems involving binary (0–1) or general discrete decision variables in addition to continuous variables arise in a host of practical applications within the context of production planning and control, network design, location and distribution, and various operational problems in the service sector. Challenging instances of such problems are successfully tackled through a careful initial model construction, followed by a polyhedral analysis to tighten the model representation with respect to its linear programming (LP) relaxation through the addition of suitable valid inequalities, and then utilizing an effective branch-and-cut algorithm. The *reformulation-linearization technique* (RLT) is an automatic model reformulation process that can be used to generate a *hierarchy* of model representations whose continuous relaxations span the spectrum from the ordinary LP relaxation to the convex hull of feasible solutions. The RLT procedure can also be used to derive (facet-defining) valid inequalities for tightening or *lifting* the model representation. Such model augmentation lifting approaches can greatly enhance problem solvability.

To describe the essence of the RLT technique, consider a linear mixed-integer zero–one program, whose feasible region X is defined in terms of n binary variables $x_1, \dots, x_n$ and m (bounded) continuous variables $y_1, \dots, y_m$, respectively comprising the vectors x and y . The RLT process generates a hierarchy of relaxations X_d for the given region X , corresponding to levels $d = 0, 1, \dots, n$, where X_0 is the ordinary continuous or LP relaxation, and where the set

X_d for any $d \in N \equiv \{1, \dots, n\}$ is generated via the following two-step process. The first step is a *reformulation* phase, which utilizes the *bound factors* $x_j \geq 0$ and $(1 - x_j) \geq 0$ for $j \in N$ to construct *bound-factor products of order d* , defined as polynomial terms obtained by selecting some d distinct x -variables, choosing the bound-factor x_j or $(1 - x_j)$ for each selected variable, and then multiplying these bound factors together. Hence, this yields $\binom{n}{d} 2^d$ such bound-factor products.

The defining constraints of X (including the variable bounding restrictions) are multiplied by each of the bound-factor products of order d , where the identity $x_j^2 \equiv x_j$ (i.e., $x_j(1 - x_j) = 0$), $\forall j \in N$, is invoked and used in this multiplication process. This yields a polynomial mixed-integer 0–1 program of degree $\min\{d + 1, n\}$. The second step in the RLT process is a *linearization phase* in which the foregoing polynomial program is linearized by a simple substitution process that replaces each distinct term $\prod_{j \in J} x_j$ with a variable w_J , $\forall J \subseteq N$, and likewise replaces each distinct term $y_k \prod_{j \in J} x_j$ with a variable v_{Jk} , $\forall J \subseteq N, k = 1, \dots, m$. This produces a LP region or polyhedral set X_d in a lifted dimension involving the variable vector (x, y, w, v) . Denoting the projection of X_d onto the original variable space (x, y) as X_{Pd} , Sherali and Adams [1,2] demonstrate that the projections $X_{P0}, X_{P1}, \dots, X_{Pn}$ yield a progression of tighter relaxations leading to the precise convex hull representation, that is

$$\begin{aligned} X_0 &= X_{P0} \supseteq X_{P1} \supseteq X_{P2} \supseteq \dots \supseteq X_{Pn} \\ &= \text{conv}(X), \end{aligned} \quad (1)$$

where $\text{conv}(X)$ denotes the convex hull of X . This result is also shown to hold true when X is defined by polynomial constraints that are linear in y when x is fixed.

References 1 and 2 additionally provide specialized reduced operations that can be used in the reformulation phase in the presence of equality constraints defining X , as well as discuss the use of *constraint*

factors of the type $(\alpha x - \beta)$ based on structural constraints $\alpha x \geq \beta$, in concert with the bound factors, for generating the product terms within the reformulation phase. Also, Refs 3–5 present earlier specialized applications at level $d = 1$ of this novel concept of lifting a model representation through a bound/constraint - factor *multiplication process* followed by a *variable substitution step* to linearize the resulting polynomial program. Boros *et al.* [6] and Lovasz and Shrijver [7] have independently presented similar hierarchies of relaxations, where the latter further incorporates constraints from semidefinite programming. In addition, based on a special case of the RLT process that employs a convexification with respect to a single variable at a time, which is also motivated by disjunctive programming ideas [8,9], Balas *et al.* [10] have developed a *lift-and-project* cutting plane algorithm that is demonstrated to produce encouraging results [11,12].

EXPLOITING SPECIAL STRUCTURES

Sherali *et al.* [13] generalize the RLT process to present a framework that is particularly well suited to exploit frequently occurring special structures such as generalized or variable upper bounds: covering, partitioning, packing constraints, as well as sparsity. Here, suppose the given MIP constrained region X includes the constraints $x \in S \equiv \{x : g_i x \geq g_{0i}, \forall i = 1, \dots, p\}$, where the constraints defining S imply the bounds $0 \leq x_j \leq 1, \forall j \in N$, on the binary variables. Accordingly, define *S-factors of order d* , for any level $d \in \{1, \dots, n\}$, to comprise the collection of products of some d constraint factors $[g_i x - g_{0i}]$, including possible repetitions. Hence, there are $\binom{p+d-1}{d}$ distinct factors of this type at level d . At any given level d , using *S-factors of order d* that are nonnull and nonimplied based on the structure S [13], the generalized RLT process constructs products of these factors with the defining constraints of X while applying the identity $x_j^2 = x_j, \forall j \in N$, at the reformulation phase, and then adopts the variable substitution technique as before in the subsequent

linearization step to generate a hierarchy of relaxations leading to the convex hull representation. Again, specialized reductions are possible in the presence of equality constraints and equality factors [13].

Reference 13 also introduces a novel concept of *conditional logic* whereby the RLT constraints generated at any level can be tightened to improve the relaxation, which would otherwise have been possible only at some higher level in the hierarchy. Here, suppose an RLT constraint is generated by taking the product of some nonnegative polynomial factor $F(x)$ with a defining constraint $\gamma x \geq \gamma_0$ to yield the restriction $\{F(x)[\gamma x - \gamma_0]\}_L \geq 0$, where $\{\cdot\}_L$ denotes the linearization of the polynomial expression $\{\cdot\}$ using the RLT variable substitution process. Now, suppose we can tighten the constraint $\gamma x \geq \gamma_0$ to a stronger restriction $\gamma' x \geq \gamma'_0$ under the condition $F(x) > 0, x \in X$. Then, we can replace the foregoing RLT constraint with the lifted inequality $\{F(x)[\gamma' x - \gamma'_0]\}_L \geq 0$, which is valid whether $F(x) = 0$ or $F(x) > 0$, when $x \in X$. This strategy captures the concepts of strong branching and logical preprocessing, and can be judiciously applied to significantly tighten MIP model representations. Sherali and Driscoll [14] illustrate an application of this concept in the context of lifting the Miller–Tucker–Zemlin formulation of the asymmetric traveling salesman problem.

The hierarchy of higher-dimensional representations produced in this manner markedly strengthens the usual continuous LP relaxation as evidenced in many computational studies on different classes of problems, where even the first-level representation helps design algorithms that significantly dominate existing procedures in the literature [3,5,13,15–18]. The theoretical implications of this hierarchy are noteworthy: the algebraic representation available at level n promotes new methods for identifying and characterizing facets and strong valid linear inequalities in the original variable space, as well as for providing information that directly bridges the gap between discrete and continuous sets. Indeed, as the level n formulation characterizes the convex hull, all valid inequalities in the original variable space

must be obtainable via a suitable projection; thus, such a projection operation serves as an all-encompassing tool for generating valid inequalities. References 19 and 20 provide discussions on generating facets and tight valid inequalities for the Generalized Upper Bound (GUB)-constrained knapsack polytope and the Boolean quadric polytope and Ref. 21 discusses persistency issues for certain constrained and unconstrained pseudo-Boolean programming problems whereby variables that take on 0–1 values at an optimum to an RLT relaxation would persist to take on these same values at an optimum to the original problem. References 22–25 discuss the use of RLT to generate improved model representations for the set-partitioning problem, the quadratic assignment problem, and for 0–1 MIPs subject to various assignment constraints.

GENERAL MIXED-DISCRETE MIPs AND SEMI-INFINITE PROGRAMS

The RLT process has also been extended to address general mixed-discrete problems where the x -variables defining the MIP region X are restricted to take on general discrete values within $S_j \equiv \{\theta_{jk}, k = 1, \dots, k_j\}, \forall j \in N$, as opposed to simply binary values [26,27]. In concept, we can represent such a problem as a 0–1 MIP by adopting the following transformation for each $j = 1, \dots, n$:

$$x_j = \sum_{k=1}^{k_j} \theta_{jk} \lambda_{jk}, \quad \text{where } \sum_{k=1}^{k_j} \lambda_{jk} = 1, \\ \text{and } \lambda_{jk} \in \{0, 1\}, \forall k = 1, \dots, k_j, \quad (2)$$

and then applying the special-structured RLT process [13] discussed in the first section based on the GUB constraints defining the transformation (Eq. 2). In particular, the constraint multiplication process in the RLT reformulation phase recognizes the following identities that are prompted by Equation (2):

$$\{\lambda_{jk_1} \lambda_{jk_2} = 0, \forall k_1 \neq k_2, \forall j\}, \{\lambda_{jk}^2 = \lambda_{jk}, \forall j, k\}, \\ \text{and } \{x_j \lambda_{jk} = \theta_{jk} \lambda_{jk}, \forall j, k\}, \quad (3)$$

and the linearization phase adopts the usual variable substitution process to generate a

hierarchy of relaxations leading to the convex hull representation.

Reference 26 (Chapter 4) exhibits that, under a suitable transformation, the foregoing RLT process can be equivalently conducted directly in the original (x, y) -variable space, without employing the intermediate binary transformation (Eq. 2). The key observation is that the inverse relationship to Equation (2) is given by

$$\lambda_{jk} = \frac{\prod_{p \neq k} (x_j - \theta_{jp})}{\prod_{p \neq k} (\theta_{jk} - \theta_{jp})} \equiv L_{jk}, \text{ say,} \\ \forall k = 1, \dots, k_j, \forall j \in N. \quad (4)$$

Note that $\lambda_{jk} = 1$ if $x_j = \theta_{jk}$, and $\lambda_{jk} = 0$ if $x_j = \theta_{jp}$ for any $p \in \{1, \dots, k_j\} \setminus \{k\}, \forall j \in N$. The polynomial expression (Eq. 4) in x_j , denoted by L_{jk} , is called a *Lagrange interpolating polynomial* (LIP; [28]), and shares the same properties (Eq. 3) as $\lambda_{jk}, \forall j, k$. The reformulation phase of the resulting RLT procedure in the (x, y) -space adopts factors composed of the LIPs L_{jk} to multiply the problem defining constraints; then, applies the following identity akin to Equation (3):

$$x_j L_{jk} = \theta_{jk} L_{jk}, \forall j, k, \quad (5)$$

and finally utilizes a variable substitution process in the linearization phase to replace each distinct product of the x -variables and each distinct product of y_i with the x -variables, $\forall i$, with a new RLT variable, in order to recover the identical hierarchy of relaxations. Reference 27 provides an alternative direct proof of Equation (1) for this latter RLT procedure that uses the LIP factors in the x -space without relying on the intermediate transformation (Eq. 2) by recognizing that the collection of LIPs can be expressed as Kronecker products of matrices given by the inverse of the transpose of Vandermonde matrices. Noting that the Kronecker product of the inverses of a given set of matrices is the inverse of their Kronecker products, a proof of Equation (1) is derived, which thereby directly generalizes the application of RLT for 0–1 MIPs.

Indeed, note that if $S_j = \{0, 1\}$, then L_{jk} reduces to $(1 - x_j)$ and x_j for $k = 1$ and 2 respectively, and Equation (5) produces the familiar identities $x_j(1 - x_j) = 0$ and $x_j^2 = x_j$, respectively, which coincide in this binary case.

Sherali and Adams [29] have shown that Equation (1) holds true even for general mixed-discrete, linear semi-infinite programs. As such, this enables the extension of RLT to bounded 0–1 mixed-integer and mixed-discrete convex programs, where such problems can be equivalently written (implicitly) as semi-infinite programs by using suitable supporting hyperplanes for the region defined by each convex constraint based on the first-order characterization of convexity [30]. Reference 29 presents an inverse transformation mechanism whereby the resulting lifted semi-infinite relaxation at any level can be transformed back into an equivalent finitely constrained convex relaxation defined in terms of the original constraint functions for which the hierarchy (Eq. 1) holds true. In particular, for 0–1 mixed-integer convex programs, this analysis recovers, at the highest level, the convex hull representation derived by Stubbs and Mehrotra [31]. Sherali and Adams [29] also present useful first-level lifted relaxations that can be judiciously applied to solve practical mixed-integer convex programs arising in a wide range of contexts such as robotics, optimal control, finance, and various chemical process and engineering design problems [30,31]. Reference 31 presents encouraging computational results using only a partial first-level RLT relaxation of this type.

SEMIDEFINITE CUTS

Motivated by the semidefinite relaxations of Lovasz and Schrijver [7], Sherali and Fraticelli [32] derive various classes of so-called *semidefinite cuts* that can be used to further tighten RLT relaxations while maintaining linearity, thus facilitating the use of robust, commercial LP software as routines for solving such relaxations within the framework of branch-and-cut algorithms. To

illustrate, consider the first-level RLT relaxation for 0–1 MIPs and let $x_{(1)} \equiv \begin{bmatrix} 1 \\ x \end{bmatrix}$, represent an $(n + 1)$ -dimensional extension of the binary vector x , and consider the matrix $M_1 \equiv [x_{(1)}x_{(1)}^T]_L$, where, as before, $[\cdot]_L$ denotes the linearization of the expression $[\cdot]$ under the RLT variable substitution process, including the application of the identity $x_j^2 = x_j$, $\forall j \in N$ (in the present binary case). Noting that $x_{(1)}x_{(1)}^T$ is symmetric and positive semidefinite (denoted $\succeq 0$), we could require that $M_1 \succeq 0$, that is,

$$\alpha^T M_1 \alpha = [(\alpha^T x_{(1)})^2]_L \geq 0, \forall \alpha \in \mathbb{R}^{n+1}, \|\alpha\| = 1.$$

As such, letting $\bar{M}_1$ represent the evaluation of M_1 at a given solution to an underlying RLT relaxation in the lifted space of the original and new RLT variables, Sherali and Fraticelli [32] present a polynomial-time procedure, having a worst-case complexity $O(n^3)$, which either verifies that $\bar{M}_1 \succeq 0$, or generates a suitable $\bar{\alpha} \in \mathbb{R}^{n+1}$ for which $\bar{\alpha}^T \bar{M}_1 \bar{\alpha} < 0$, thus yielding the semidefinite cut $\bar{\alpha}^T M_1 \bar{\alpha} = [(\bar{\alpha}^T x_{(1)})^2]_L \geq 0$. Furthermore, motivated in part by the moment matrices utilized by Lasserre [33,34] in the context of continuous polynomial programs, Sherali and Desai [35] and Sherali *et al.* [36] extend this concept by replacing $x_{(1)}$ with a vector v that comprises the components of $x_{(1)}$ plus other suitable higher-order monomials, and then invoking $[vv^T]_L \succeq 0$ to generate so-called *v -semidefinite cuts*. Different techniques are described in [35,36] for generating various classes of such v -semidefinite cuts, along with encouraging computational results.

EXTENSION TO CONTINUOUS GLOBAL OPTIMIZATION

The RLT methodology provides a unifying framework for solving not only mixed-discrete polynomial programming problems but also continuous, nonconvex polynomial [37] and factorable [38] optimization problems to global optimality. Here, the *bound factors* $(x_j - \ell_j)$ and $(u_j - x_j)$ are defined with respect to the general bounding restrictions $\ell_j \leq x_j \leq u_j$, for all the variables x_j defining the given nonconvex program, and are used

in concert with the aforementioned constraint factors to generate tight relaxations, which are then embedded within a branch-and-bound algorithm. These relaxations can be further augmented by semidefinite cuts [32,35,36], as well as with other valid inequalities based on grid factors, LIP, and Chebyshev polynomials [38–40]. By utilizing a suitable partitioning process [37,38], such an enumeration process can be induced to converge to a global optimal solution.

References 26 and 41–50 provide additional reading on survey articles and the application of RLT to several special problems (including bilinear, indefinite quadratic, location allocation, and linear complementarity problems) in the domain of both discrete and continuous nonconvex optimization.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. CMMI-0969169.

REFERENCES

1. Sherali HD, Adams WP. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J Discrete Math* 1990;3(3):411–430.
2. Sherali HD, Adams WP. A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems. *Discrete Appl Math* 1994;52:83–106.
3. Adams WP, Sherali HD. A tight linearization and an algorithm for zero-one quadratic programming problems. *Manage Sci* 1986;32(10):1274–1290.
4. Adams WP, Sherali HD. Linearization strategies for a class of zero-one mixed integer programming problems. *Oper Res* 1990;38(2):217–226.
5. Adams WP, Sherali HD. Mixed-integer bilinear programming problems. *Math Program* 1993;59(3):279–306.
6. Boros E, Crama Y, Hammer PL. Upper bounds for quadratic 0–1 maximization problems. *Oper Res Lett* 1990;9:73–79.
7. Lovasz L, Schrijver A. Cones of matrices and set functions, and 0–1 optimization. *SIAM J Optim* 1990;1:166–190.
8. Balas E. Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM J Algebr Discrete Methods* 1985;6:466–485.
9. Sherali HD, Shetty CM. Optimization with disjunctive constraints. Berlin, Germany: Springer; 1980.
10. Balas E, Ceria S, Cornuejols G. A lift-and-project cutting plane algorithm for mixed 0–1 programs. *Math Program* 1993;58(3):295–324.
11. Balas E, Perregaard M. Lift-and-project for mixed 0–1 programming: recent progress. *Discrete Appl Math* 2002;123:129–154.
12. Balas E, Perregaard MA. Precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0–1 programming. *Math Program B* 2003;94:221–245.
13. Sherali HD, Adams WP, Driscoll P. Exploiting special structures in constructing a hierarchy of relaxations for 0–1 mixed integer problems. *Oper Res* 1998;46(3):396–405.
14. Sherali HD, Driscoll PJ. On tightening the relaxation of Miller–Tucker–Zemlin formulations for asymmetric traveling salesman problems. *Oper Res* 2002;50(4):656–669.
15. Sherali HD, Adams WP. A decomposition algorithm for a discrete location-allocation problem. *Oper Res* 1984;32(4):878–900.
16. Sherali HD, Brown EL. A quadratic partial assignment and packing model and algorithm for the airline gate assignment problem. In: Pardalos PM, Wolkowicz H, editors. Volume 16, DIMACS series in discrete mathematics and theoretical computer science, quadratic assignment and related problems. Providence (RI): American Mathematical Society; 1994. pp. 343–364.
17. Sherali HD, Ramachandran S, Kim S. A localization and reformulation discrete programming approach for the rectilinear distance location-allocation problem. *Discrete Appl Math* 1994;49(1–3):357–378.
18. Sherali HD, Sarin SC, Tsai PF. A class of lifted path and flow-based formulations for the asymmetric traveling salesman problem with and without precedence constraints. *Discrete Optim* 2006;3:20–32.
19. Sherali HD, Lee Y. Sequential and simultaneous liftings of minimal cover inequalities for GUB constrained knapsack polytopes. *SIAM J Discrete Math* 1995;8(1):133–153.
20. Sherali HD, Lee Y, Adams WP. A simultaneous lifting strategy for identifying new

- classes of facets for the Boolean quadric polytope. *Oper Res Lett* 1995;17(1):19–26.
21. Adams WP, Lasserre JB, Sherali HD. Persistence in 0–1 optimization. *Math Oper Res* 1998;23(2):359–389.
 22. Sherali HD, Lee Y. Tighter representations for set partitioning problems via a reformulation-linearization approach. *Discrete Appl Math* 1996;68:153–167.
 23. Adams WP, Guignard M, Hahn PM, Hightower WL. A level-2 reformulation-linearization technique bound for the quadratic assignment problem. *Eur J Oper Res*. 2007;180(3):983–996.
 24. Adams WP, Johnson TA. Improved linear programming-based lower bounds for the quadratic assignment problem. In: Pardalos PM, Wolkowicz H, editors. DIMACS series in discrete mathematics and theoretical computer science, quadratic assignment and related problems. Volume 16, 1994. pp. 43–75.
 25. Liberti L. Compact linearization of binary quadratic problems. 4OR 2007. DOI: 10.1007/s10288-006-0015-3. (published online 2007).
 26. Sherali HD, Adams WP. A reformulation-linearization technique for solving discrete and continuous nonconvex problems. Dordrecht/Boston/London: Kluwer Academic Publishers; 1999.
 27. Adams WP, Sherali HD. A hierarchy of relaxations leading to the convex hull representation for general discrete optimization problems. *Ann Oper Res* 2005;140(1):21–47.
 28. Volkov EA. Numerical methods. New York: Hemisphere Publishing; 1990.
 29. Sherali HD, Adams WP. A reformulation-linearization technique (RLT) for semi-infinite and convex programs under mixed 0–1 and general discrete restrictions. *Discrete Appl Math* 2009;157(6):1319–1333.
 30. Bazaraa MS, Sherali HD, Shetty CM. Nonlinear programming: theory and algorithms, 3rd ed. New York: John Wiley & Sons; 2006.
 31. Stubbs RA, Mehrotra S. A branch-and-cut method for 0–1 mixed convex programming. *Math Program* 1999;86:515–532.
 32. Sherali HD, Fraticelli BMP. Enhancing RLT relaxations via a new class of semidefinite cuts. *J Global Optim* 2002;22(1-4):233–261.
 33. Lasserre JB. Global optimization with polynomials and the problem of moments. *SIAM J Optim* 2001;11:796–817.
 34. Lasserre JB. Semidefinite programming vs. LP relaxations for polynomial programming. *Math Oper Res* 2002;27(2):347–360.
 35. Sherali HD, Desai J. On solving polynomial, factorable, and black-box optimization problems using the RLT methodology. In: Audet C, Hansen P, Savard G, editors. Essays and surveys on global optimization 2005. New York: Springer; pp. 131–163.
 36. Sherali HD, Dalkiran E, Desai J. A class of v -semidefinite cuts. Manuscript. Blacksburg, Virginia: Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University; 2009.
 37. Sherali HD, Tuncbilek CH. A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *J Global Optim* 1992; 2:101–112.
 38. Sherali HD, Wang H. Global optimization of nonconvex factorable programming problems. *Math Program* 2001;89(3):459–478.
 39. Sherali HD, Tuncbilek CH. A reformulation-convexification approach for solving nonconvex quadratic programming problems. *J Global Optim* 1995;7:1–31.
 40. Sherali HD, Tuncbilek CH. New reformulation-linearization technique based relaxations for univariate and multivariate polynomial programming problems. *Oper Res Lett* 1997;21(1):1–10.
 41. Liberti L. Reformulation and convex relaxation techniques for global optimization. 4OR 2004;2:255–258.
 42. Liberti L, Pantelides CC. An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms. *J Global Optim* 2006;36:161–189.
 43. Sherali HD. Global optimization of nonconvex polynomial programming problems having rational exponents. *J Global Optim* 1998;12(3):267–283.
 44. Sherali HD. Tight relaxations for nonconvex optimization problems using the reformulation-linearization/convexification technique (RLT). In: Pardalos PM, Romeijn HE, editors. Handbook of global optimization. Volume 2, Heuristic Approaches. Dordrecht/Boston/London: Kluwer Academic Publishers; 2002. pp. 1–63.
 45. Sherali HD. RLT: a unified approach for discrete and continuous nonconvex optimization. *Ann Oper Res* 2007;149:184–193.

46. Sherali HD, Liberti L. Reformulation-linearization for global optimization. Encyclopedia of optimization, 2nd ed. 2008. pp. 3263–3268.
47. Sherali HD, Alameddine A. A new reformulation-linearization technique for solving bilinear programming problems. J Global Optim 1992;2:379–410.
48. Sherali HD, Al-Loughani I, Subramanian S. Global optimization procedures for the capacitated Euclidean and LP distance multifacility location-allocation problems. Oper Res 2002;50(3):433–448.
49. Sherali HD, Ganesan V. A pseudo-global optimization approach with application to the design of containerships. J Global Optim 2003;26(4):335–360.
50. Sherali HD, Krishnamurty R, Al-Khayyal FA. Enumeration approach for linear complementarity problems based on a reformulation-linearization technique. J Optim Theory Appl 1998;99(2):481–507.

REGENERATIVE PROCESSES

MARIA VLASIOU
Department of Mathematics and
Computer Science, Eindhoven
University of Technology,
Eindhoven, The Netherlands

CLASSICAL REGENERATIVE PROCESSES

A stochastic process $\{X(t), t \geq 0\}$ is intuitively a regenerative process if it can be split into i.i.d. cycles. That is, we assume that a collection of time points exists, so that between any two consecutive time points in this sequence, (i.e., during a cycle), the process $\{X(t), t \geq 0\}$ has the same probabilistic behavior. There are several proposed definitions that attempt to make this description precise (see Asmussen [1], Çinlar [2], or Thorisson [3]). The classical definition of regenerative processes, which is given below, also demands that the process is independent of its history up to the previous regeneration point, and of the choice of the point itself. This assumption has been relaxed significantly and the notion of regenerative processes that we understand nowadays is wider; we will sketch these ideas in the following sections. This presentation is influenced by many books, which can be found in the references, and draws heavily from Sigman and Wolff [4].

The power of the concept of regenerative processes lies in the existence of a limiting distribution under conditions that are mild and usually easy to verify. For example, in continuous time it is only required that the mean cycle length of the embedded renewal process is finite, that the cycle length distribution is nonlattice, and that the sample paths of the regenerative process satisfy some conditions (e.g., right-continuity with left-hand limits) that are automatic in most examples [Chapter VI] [1].

Classic examples of regenerative processes are ergodic Markov chains and processes. For example, take a Markov chain

or a Markov process with a countable state space E and fix a state i . Assume that the process started at state i . Then every time at which state i is entered is a time of regeneration for the Markov process. As a second example, consider a single-server queueing system with Poisson arrivals and i.i.d. service times. Suppose that we start observing the system at a departure instant, which left behind i customers. Then every time a customer departs leaving behind him i customers is a regeneration epoch and the process describing the queue length after such a time has exactly the same probability law as the one it had at time zero.

We start with the classical definition of a regenerative process.

Definition 1 [Classical Definition].

A regenerative process $\{X(t), t \geq 0\}$ with state space E (endowed with a σ -field $\mathcal{E}$) is a stochastic process on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the following properties: there exists a random variable $R_1 > 0$ such that

- (i) $\{X(t + R_1), t \geq 0\}$ is independent of $\{X(t), t \leq R_1\}$ and R_1 ,
- (ii) $\{X(t + R_1), t \geq 0\}$ is stochastically equivalent to $\{X(t), t \geq 0\}$.

Stochastically equivalent means that the two stochastic processes in (ii) in Definition 1 have the same joint distributions. Because of property (i) in Definition 1, R_1 is a stopping time for $\{X(t), t \geq 0\}$. For a rigorous definition of property (i) in Definition 1, see also Sigman and Wolff [4]. A standard assumption that is often made is that the paths of $X(t)$ are right-continuous with left-hand limits almost surely. This ensures that the sample paths are continuous except possibly on a set of Lebesgue measure 0. We will comment on the necessity of this assumption in the main theorem later on.

The random variable R_1 is called a *regeneration epoch*, and its existence implies the

existence of a sequence of regeneration epochs $\{R_n\}$. This result is carefully proven in Sigman *et al.* [5] and Svertchkov [6]. We say that the process *regenerates* at such epochs. In other words, $\{X(t), t \geq 0\}$ and $\{X(t + R_n), t \geq 0\}$ are stochastically identical, and due to property (i) in Definition 1, $\{X(t + R_n), t \geq 0\}$ is independent of $\{X(t), 0 \leq t \leq R_n\}$. The regeneration epochs generate a renewal process $S_n = R_1 + \dots + R_n$ (see **Definition and Examples of Renewal Processes**), which is called the *embedded renewal process* for the regenerative process $\{X(t), t \geq 0\}$. Moreover, the interval $[S_{n-1}, S_n]$ is called the *nth regenerative cycle*, where we have assumed that $S_0 = 0$. A regenerative process with finite mean cycle length is called *positive recurrent*.

It will sometimes be convenient to allow the process during the first cycle to be different; this simply means that the process during the first cycle may have a distribution that is different from that of subsequent cycles. We call such a process *delayed regenerative*, if the emphasis on the assumptions of the first cycle is needed. From a time-average point of view, the first cycle is not important, provided that it eventually ends. That is, we assume that R_1 is a proper random variable. There are no new tools needed; in handling a delayed regenerative process, we first condition the event or expectation in question on the time R_1 of the first regeneration, and then we use the fact that at R_1 there starts an ordinary regenerative process. Thus, we use the term *regenerative* to cover both cases of delayed or not regenerative processes.

Remark 1. If the distribution of the process during the first special cycle and the length of that cycle are constructed in a particular way, then we can construct a stationary regenerative process, which implies that the distribution of $\{X(t), t \geq 0\}$ is independent of t for all t ; see Wolff [7] for a construction.

An important fact that follows directly from the classical definition is that functions of regenerative processes are also themselves regenerative.

Proposition 1 [Inheritance of Regenerations]. *If $\{X(t), t \geq 0\}$ with state space E is a regenerative process with regeneration epochs R_n , then $\{f(X(t)), t \geq 0\}$ is also a regenerative process (with some state space E') with the same regeneration epochs for any $f : E \rightarrow E'$.*

For instance, take $\{X(t), t \geq 0\}$ to be a regenerative process and consider the function $\hat{X}(t) = \mathbf{1}(X(t) \in B)$ for some set B . Since $X(t)$ is regenerative, then so is $\hat{X}(t)$, which can help one compute the distribution of one regenerative process as the expectation of another, that is, $\mathbb{P}(X(t) \in B) = \mathbb{E}[\hat{X}(t)]$. This proposition illustrates a more general observation that the embedded renewal times of a regenerative process need not be stopping times with respect to the evolution of $X(t)$. Observe that in contrast a function of a Markov process is usually not a Markov process itself. Nevertheless, it remains regenerative if the Markov process is so.

Other than computing the limiting distribution of one regenerative process via another, one could use the inheritance of regenerations to add a cost or reward function to the process. Namely, if we assume that the system earns rewards (that may be negative) at a rate $c(x)$ whenever it is in state x , then the total reward up to time t is given by $\int_0^t c(X(y)) dy$, where $\{c(X(y)), y \geq 0\}$ is a regenerative process. For finite mean cycle lengths, the time average of $c(x)$ defined to be equal to $\lim_{t \rightarrow \infty} \int_0^t c(X(y)) dy / t$ can be found by the renewal-reward theorem (see Theorem 1 in **Renewal Processes with Costs and Rewards**). In particular, it can be shown with standard methods that

$$\lim_{t \rightarrow \infty} \int_0^t c(X(y)) dy / t = \int_{-\infty}^{\infty} c(y) dF(y),$$

where $F(x)$ is the limiting distribution of the regenerative process $\{X(t), t \geq 0\}$, which is given below in Theorem 1 by choosing $(-\infty, x]$ for the set B appearing in Equation (1).

The primary use of the key renewal theorem (see **Limit Theorems for Renewal Processes**) is in characterizing the limiting behavior of regenerative processes. Renewal

theory is the main tool for studying regenerative processes in the absence of further properties. Conversely, it is this applicability to regenerative processes, which makes renewal theory one of the most important tools in elementary probability theory. The time average properties of regenerative processes are straightforward consequences of the renewal-reward theorem (which is based on the strong law of large numbers). This follows from the fact that the cycles, and the process during each cycle, are i.i.d. We proceed with the main theorem in this section, which gives the limiting distribution of a regenerative process.

Theorem 1 [Limiting Behavior of Regenerative Processes]. *Suppose that the process $\{X(t), t \geq 0\}$ is regenerative with regeneration epochs with finite mean, that is, $\mathbb{E}[R_1] < \infty$. We then have*

- (a) *The limiting distribution of the regenerative process in the time average sense exists; that is, the limit $\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbf{1}(X(s) \in B) ds$ exists for all measurable sets B , $B \in \mathcal{E}$, where $\mathbf{1}(\cdot)$ is the indicator function. For $B = (-\infty, x]$ we denote it by $F(x)$. Moreover, if we denote by X_∞ a generic random variable with this limiting distribution, we have that*

$$\begin{aligned} \mathbb{P}(X_\infty \in B) &= \frac{1}{\mathbb{E}[R_1]} \\ &\times \mathbb{E} \left[\int_0^{R_1} \mathbf{1}(X(s) \in B) ds \right], \quad B \in \mathcal{E}. \end{aligned} \quad (1)$$

- (b) *If the state space E is a complete separable metric space and the process $\{X(t), t \geq 0\}$ has right-continuous paths, then in case the cycle length distribution is nonlattice, we have that $\mathbb{P}(X(t) \in B)$ converges weakly to $\mathbb{P}(X_\infty \in B)$. Moreover, for a measurable function $f : E \rightarrow \mathbb{R}$ we have that*

$$\begin{aligned} &\lim_{t \rightarrow \infty} \mathbb{E}[f(X(t))] \\ &= \frac{1}{\mathbb{E}[R_1]} \mathbb{E} \left[\int_0^{R_1} f(X(s)) ds \right], \end{aligned} \quad (2)$$

which for the special case where f is the indicator function, yields that $\lim_{t \rightarrow \infty} \mathbb{P}(X(t) \in B) = \mathbb{P}(X_\infty \in B)$ is equal to the right-hand side of Equation (1).

- (c) *If the cycle length distribution is spread out, then $\mathbb{P}(X(t) \in B)$ converges to $\mathbb{P}(X_\infty \in B)$ in total variation (i.e., the convergence is uniform over all Borel sets B).*

For a regenerative process in discrete time, the above theorem holds in case the cycle length is aperiodic. If it is periodic with period d , the above limit holds if $t = nd$ and $n \rightarrow \infty$ [1,8]. Serfozo [9] derives Equation (2) without the assumptions under (b) in Theorem 1, but with the assumptions that the cycle length distribution is nonlattice and that for $M = \sup\{|f(X(t))| : t \leq R_1\}$, M and MR_1 have finite means.

From this theorem we see that the limiting distribution for a regenerative process, $\mathbb{P}(X_\infty \in B)$, is equal to the expected time that the process spent in B over the expected inter-renewal time. Evidently, if the embedded renewal process is recurrent but with $\mathbb{E}[R_1] = \infty$, then we have that $\mathbb{P}(X_\infty \in B) = 0$. One could use the above theorem to derive the Laplace–Stieltjes transform of the limiting distribution of $\{X(t), t \geq 0\}$ by the appropriate choice of the function f appearing in (b) in Theorem 1.

This theorem is also true, with slight modifications, for delayed regenerative processes. We, namely, have to integrate over a proper regenerative cycle, say $[S_1, S_2)$, rather than the first special cycle.

Example 1 [Age Process]. Evidently, the process Z_t giving the time from t until the next renewal in a renewal process is a regenerative process.

Example 2 [GI/GI/1 Queue]. Let $\{X(t), t \geq 0\}$ be the number of customers in a GI/GI/1 queue and let S_n be the n th time when a customer enters an empty system. Provided that the occupation rate of the system is less than 1 (i.e., that the mean service time is less than the mean interarrival time), the mean cycle length is finite. If the

process starts with a customer entering an empty system at time 0, then $\{X(t), t \geq 0\}$ is a regenerative process; otherwise it is a delayed regenerative process. The state space of this regenerative process is given by $\{0, 1, 2, \dots\}$. Then, under the assumptions of the theorem above

$$\begin{aligned} \lim_{t \rightarrow \infty} \mathbb{P}(X(t) = j) &= \frac{\mathbb{E}[\text{time in state } j \text{ during a cycle}]}{\mathbb{E}[\text{time of a cycle}]} \\ &= \lim_{t \rightarrow \infty} \frac{\mathbb{E}[\text{time in } j \text{ during } (0, t)]}{t}. \end{aligned}$$

That is, $\lim_{t \rightarrow \infty} \mathbb{P}(X(t) = j)$ is not only the probability that the regenerative process is in state j in steady state (i.e., the probability that there are j customers in the system in steady state), but also the long-run proportion of time that the process was in state j . To prove this result, one simply needs to assume that a reward is earned at rate 1 each time the process is in state j , which generated a renewal-reward process and the equality is then a consequence of the renewal-reward theorem (or of Theorem 1 above, for a particular choice of f).

Example 3 [I.i.d. Random Variables].

Any independent and identically distributed sequence of random variables is a regenerative process. The embedded renewal process is simply the counting process of the sequence $(S_n = n)$. If these random variables are continuous, we observe that the process returns to a fixed state with probability zero, and thus, the regeneration epochs cannot be defined as returns to any particular fixed state; contrast this with the fact that in the GI/GI/1 queue example above, returns to the fixed state “the number of customers in the queue is zero” were chosen to be the regenerative epochs.

As mentioned earlier, one can find several examples of regenerative processes in Markov chains and Markov processes.

Remark 2. There is a version of the central limit theorem for regenerative processes that can be used to construct confidence intervals

when simulating regenerative processes; see Sigman and Wolff [4] and Wolff [7] for details.

ONE-DEPENDENT REGENERATIVE PROCESSES

The definition of regenerative processes has been extended to allow for one-dependent cycles, which connects regenerative processes to Harris-recurrent Markov chains. Many-server queues formed the main motivation to extend the classical notion of regeneration to this definition [4]. Previously, for the GI/GI/1 queue we had taken the times when the system empties as the regeneration epochs. Suppose now you have a GI/GI/2 queue, which is stable, that is, the offered load to the system per server is less than 1. Take a completely deterministic system where service takes 1.5 time units and customers arrive every 1 time unit. Let the first arrival occur at time 0. Then we observe that this system never empties. Thus, we are not able to define the same regeneration epochs (i.e., an empty system) as for the single-server queue. The question is now if one should infer that the queue length process is not regenerative.

The definition of regenerative processes we have used so far demands that $\{X(t + R_1), t \geq 0\}$ is independent both of R_1 and of $\{X(t), t \leq R_1\}$ (see Definition 1). A more general definition is given in Asmussen [1] where the requirement now is that $\{X(t + R_1), t \geq 0\}$ is independent only of R_1 but not necessarily of $\{X(t), t \leq R_1\}$. Under both definitions, $\{X(t + R_1), t \geq 0\}$ is stochastically equivalent to $\{X(t), t \geq 0\}$. As a result, the embedded renewal process has i.i.d. cycles, but the regenerative process might be dependent. Asmussen [1] shows that under this definition, the process is *one-dependent*, namely, the process during adjacent cycles is dependent, but during nonadjacent cycles it is independent (see also [10]). If a process is regenerative under Definition 1, then it is naturally also so under this definition, while evidently, the converse is not true.

Definition 2 [One-Dependent Definition]. A regenerative process $\{X(t), t \geq 0\}$ with state space E (endowed with a σ -field $\mathcal{E}$) is a stochastic process on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the following properties: there exists a random variable $R_1 > 0$ such that

- (i) $\{X(t + R_1), t \geq 0\}$ is independent of R_1
- (ii) $\{X(t + R_1), t \geq 0\}$ is stochastically equivalent to $\{X(t), t \geq 0\}$.

Since the cycle lengths R_n are still i.i.d., they generate an embedded renewal process. Thus, Equation (2) still holds; the regenerative process converges in distribution to its limit (and also in total variation). This definition can be used to characterize Harris-recurrent Markov chains, as a Markov chain is positive Harris recurrent if and only if it is regenerative under the one-dependent definition and ergodic for every initial state x , with the distribution of the process during a cycle (possibly excluding the first one for the delayed case) not depending on the initial state x [4,11]. We usually think of queues as continuous-time processes, but we can embed discrete-time processes in them, as for example, instances when a customer arrives. Embedded processes for several queueing models turn out to be regenerative under the one-dependent definition, and with finite mean cycle lengths [12,13]. Returning to the GI/GI/2 example mentioned above, it has been shown [14] that the vector giving the workload for each server is a Harris ergodic Markov chain and thus it has a regenerative structure [7,13].

FURTHER READING

The interested reader should start with the comprehensive and in-depth review of regenerative processes by Sigman and Wolff [4]; several important references on the topic can be found there—we present only a few of them. Smith [15,16] constructs a regenerative process under the classical definition using random tours; see also [17]. As this

topic is an integral part of stochastic processes, most books on stochastic processes have some treatment of (classical) regenerative processes [2,8,9,17–19]. References on how to simulate a stationary version of a regenerative process have been omitted; few of them can be found in the sources cited here.

The notion of regeneration has been extended to even more general definitions. Rather than demanding that the regenerative epochs R_n form an i.i.d. sequence (cf. property (i) in Definition 2), one can extend the notion of the dependence of $\{X(t)\}$ on the process in adjacent cycles, to one-dependence of the cycle lengths themselves. In Sigman [11] it is shown that these processes are related to *continuous-time* Harris-recurrent Markov processes; (compare this to what has been mentioned in the previous section that Definition 2 is related to Harris ergodic chains, i.e., in discrete time). Further examples of this extended notion of one-dependence of regenerative processes include the superposition of two independent renewal processes where at least one of them has a spread-out cycle length distribution.

Another way of describing the regenerative structure for a recursive sequence as this was given under Definition 2 is by the *renewing events* or *renovation method* [20,21]. This amounts to first *algebraically* describing the independence of the past property rather than doing so in a probabilistic fashion; see Sigman and Wolff [4] for an example.

A further extension is the notion of *synchronous processes*; a stochastic process $\{X(t), t \geq 0\}$ is called *synchronous* if there exists a random variable $R_1 > 0$ such that $\{X(t + R_1), t \geq 0\}$ is stochastically equivalent to $\{X(t), t \geq 0\}$, possibly by enlarging the probability space, if necessary. That is, we no longer require property (i) in Definition 2 [22–24].

Other notions of regenerative processes worth mentioning are those of Kingman [25,26] in which the kind of regenerative structure found in discrete-state Markov processes is generalized by allowing any of the continuum of times during the visit to a recurrent state to serve as a regeneration epoch and that of Thorisson [3], who introduces the notion of a time-inhomogeneous

regenerative process, which can apply, for example, to an inhomogeneous Markov chain with a recurrent state.

The list of references here is by no means exhaustive, but should serve only as a starting point.

REFERENCES

1. Asmussen S. Applied probability and queues. New York: Springer; 2003.
2. Çinlar E. Introduction to stochastic processes. Englewood Cliffs, NJ: Prentice-Hall Inc.; 1975.
3. Thorisson H. The coupling of regenerative processes. *Adv Appl Probab* 1983;15:531–561.
4. Sigman K, Wolff RW. A review of regenerative processes. *SIAM Rev. A Publ Soc Ind Appl Math* 1993;35:269–288.
5. Sigman K, Thorisson H, Wolff RW. A note on the existence of regeneration times. *J Appl Probab* 1994;31:1116–1122.
6. Svertchkov MY. On wide-sense regeneration. In: Kalashnikov, Vladimir and Zolotarev, Vladimir. Stability problems for stochastic models (Suzdal, 1991). Volume 1546, Lecture notes in mathematics. Berlin, Heidelberg; Springer; 1993. pp. 167–169.
7. Wolff RW. Stochastic modeling and the theory of queues. Prentice Hall international series in industrial and systems engineering. Englewood Cliffs, NJ: Prentice Hall Inc.; 1989.
8. Kulkarni VG. Modeling and analysis of stochastic systems. Texts in statistical science series. London: Chapman and Hall Ltd.; 1995.
9. Serfozo R. Basics of applied stochastic processes. Probability and its applications (New York). Berlin: Springer; 2009.
10. Nummelin E. On distributionally regenerative processes. *J Theor Probab* 1994;7:739–756.
11. Sigman K. The stability of open queueing networks. *Stoch Process Appl* 1990;35:11–25.
12. Sigman K. Queues as Harris recurrent Markov chains. *Queueing Syst Theory Appl* 1988;3:179–198.
13. Sigman K. Regeneration in tandem queues with multiserver stations. *J Appl Probab* 1988; 25:391–403.
14. Charlot F, Ghidouche M, Hamami M. Irréductibilité et récurrence au sens de Harris des “temps d’attente” des files $GI/G/q$. *Z Wahr Ver Geb* 1978;43:187–203.
15. Smith WL. Regenerative stochastic processes. *Proc R Soc. London. Ser A. Math, Phys Eng Sci* 1955;232:6–31.
16. Smith WL. Renewal theory and its ramifications. *J R Stat Soc. Ser B. Method* 1958;20: 243–302.
17. Resnick S. Adventures in stochastic processes. Boston, MA: Birkhäuser Boston Inc.; 1992.
18. Cohen JW. On regenerative processes in queueing theory. Berlin. Volume 121, Lecture notes in economics and mathematical systems. Springer; 1976.
19. Ross SM. Stochastic processes. 2nd ed. Wiley series in probability and statistics: probability and statistics. New York: John Wiley & Sons Inc.; 1996.
20. Borovkov AA. Asymptotic Methods in queueing theory. Chichester: John Wiley & Sons; 1984.
21. Foss SG, Kalashnikov VV. Regeneration and renovation in queues. *Queueing Syst. Theory Appl* 1991;8:211–223.
22. Franken P, König D, Arndt U, *et al.* Queues and point processes. Wiley series in probability and mathematical statistics: applied probability and statistics. Chichester: John Wiley & Sons Ltd.; 1982.
23. Glynn P, Sigman K. Uniform Cesàro limit theorems for synchronous processes with applications to queues. *Stochastic Process Appl* 1992;40:29–43.
24. Rolski T. Stationary random processes associated with point processes. Volume 5, Lecture notes in statistics. New York: Springer; 1981.
25. Kingman JFC. Regenerative phenomena. Wiley series in probability and mathematical statistics. London-New York-Sydney: John Wiley & Sons Ltd.; 1972.
26. Kingman JFC. Progress and problems in the theory of regenerative phenomena. *Bull London Mathe Soc* 2006;38:881–896.

REINFORCEMENT LEARNING ALGORITHMS FOR MDPs

CSABA SZEPESVÁRI
Department of Computing Science,
University of Alberta, Edmonton,
Alberta, Canada

OVERVIEW

Reinforcement learning (RL) refers to both a learning problem and a subfield of machine learning. As a learning problem it refers to learning to control a system so as to maximize some numerical value which represents a long-term objective. A typical setting where RL operates is shown in Fig. 1: A controller receives the controlled system's state and a reward associated with the last state transition. It then calculates an action which is sent back to the system. In response, the system makes a transition to a new state and the cycle is repeated. The problem is to learn a way of controlling the system so as to maximize the total reward. The learning problems differ in the details of how the data is collected and how performance is measured.

In this article we assume that the system that we wish to control is stochastic. Further, we assume that the measurements available on the system's state are detailed enough so that the controller can avoid reasoning about how to collect information about the state. Problems with these characteristics are best described in the framework of Markovian decision processes (MDPs). The standard approach to "solve" MDPs is to use dynamic programming, which transforms the problem of finding a good controller into the problem of finding a good value function. However, apart from the simplest case when the MDP has very few states and actions, dynamic programming is infeasible. The RL algorithms discussed here can be thought of as a way of turning the infeasible dynamic programming methods into practical algorithms so that they can be applied to large-scale problems.

There are two key ideas that allow RL algorithms to achieve this goal. The first is to use samples to compactly represent the dynamics of the control problem. This is important for two reasons: First, it allows one to deal with learning scenarios when the dynamics is unknown. Second, even if the dynamics is available, exact reasoning might be intractable on its own. The second key idea behind RL algorithms is to use powerful function approximation methods to compactly represent value functions. The significance of this is that it makes dealing with large, high-dimensional state- and/or action-spaces possible. What is more, the two ideas fit nicely together: Samples may be focused on a small subset of the spaces they belong to, which clever function approximation techniques might exploit. It is the understanding of the interplay between dynamic programming, samples and function approximation that is at the heart of designing, analyzing and applying RL algorithms.

The interested reader is invited to learn more about RL by consulting the rich literature. The first survey that appeared on RL is due to Kaelbling *et al.* [1], which was followed by the publication of the books by Bertsekas and Tsitsiklis [2] and Sutton and Barto [3]. The book by Bertsekas and Tsitsiklis [2] discusses the theoretical foundations, while the book by Sutton and Barto [3] focuses on presenting the core ideas in an accessible manner. More recent books that are devoted in part or in whole to the subject include those by Gosavi [4], Cao [5], Bertsekas [6,7], Powell [8], Chang *et al.* [9], Busoniu *et al.* [10] and Szepesvári [11]. On-line, periodically updated summaries are provided by Bertsekas [12] and Szepesvári [13].

The rest of this article is organized according to Fig. 2. After introducing our notation (see the section titled "Preliminaries"), we consider value prediction problems, describe TD(λ) and its recent improved versions, along with least-squares methods. In the section titled "Control", we discuss a few algorithms for control, most notably Q-learning and

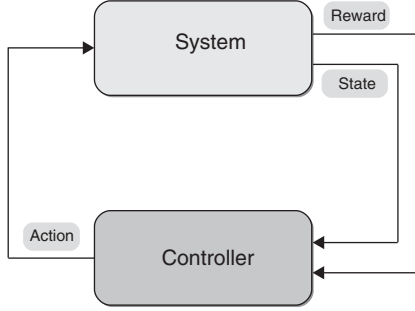


Figure 1. The basic reinforcement learning scenario.

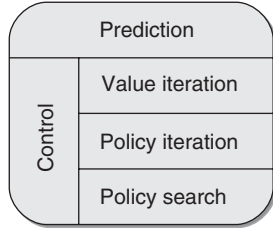


Figure 2. Types of reinforcement problems and approaches. The article gives examples of algorithms for all approaches except for (direct) policy search.

fitted Q-iteration, which can be thought of as implementing value iteration. This is followed by a description of an actor-critic algorithm, which can be thought of as implementing policy iteration. Owing to the lack of space, the discussion of (direct) policy search methods is not included.

PRELIMINARIES

We assume that the reader is familiar with MDPs. Accordingly, the sole purpose of this section is to introduce the notation used in the rest of the article.

We consider controlled stochastic process with Markovian dynamics:

$$\begin{aligned} X_{t+1} &= f_1(X_t, A_t, D_{t+1}), \\ R_{t+1} &= f_2(X_t, A_t, D_{t+1}), \quad t = 0, 1, \dots \end{aligned} \quad (1)$$

Here, $X_t \in \mathcal{X}$ is the state of the system at time t , $A_t \in \mathcal{A}$ is the action taken by the controller after X_t was observed, D_{t+1} is a disturbance

term (taking values in some Euclidean space), and $R_{t+1} \in \mathbb{R}$ is the reward received.¹ The disturbance sequence $(D_t; t \geq 1)$ is a sequence of independent, identically distributed (i.i.d.) random variables. The goal of the controller is to maximize the expected total discounted reward, or *expected return*, irrespective of the initial state

$$\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid X_0 = x \right] \rightarrow \max!, \quad x \in \mathcal{X}.$$

Here, $0 < \gamma < 1$ is the so-called *discount factor*. The process $((X_t, A_t, R_{t+1}); t \geq 0)$ is called a *Markov Decision Process*.² In the applications, f is usually a smooth function of its arguments.

Equation (1) specifies an MDP in the so-called state-space form. An equivalent definition can be given in terms of a transition probability kernel $\mathcal{P}_0$ which assigns to each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ a probability measure over $\mathcal{X} \times \mathbb{R}$. Denoting the latter by $\mathcal{P}_0(\cdot \mid x, a)$, the connection between the two formalisms is given by

$$\mathcal{P}_0(U \mid x, a) = \mathbb{P}(f(x, a, D) \in U),$$

where $f(x, a, d) = (f_1(x, a, d), f_2(x, a, d))$ and D is a random variable whose distribution is the same as that of D_t .

For simplicity, consider the case when the action set $\mathcal{A}$ is countable. A general policy determines the actions chosen as a function of history. The *value* of state x under a policy π is the expected return given that the policy is started in state x , while the *action value* of a state-action pair (x, a) is the expected

¹Here, we assumed that all actions are admissible in all states. Sometimes it is necessary to restrict the set of actions admissible to a subset of the actions in a state-dependent manner. The results and algorithms presented here extend to this case essentially without any change.

²The algorithms described below have been extended to other than the expected total discounted reward, such as the expected total reward (i.e., no discounting) or the average reward criteria. The discounted criterion is selected because it allows a simpler presentation of the main ideas.

return given that the process is started from state x , the first action is a after which the policy π is followed. The value of state x under policy π is denoted by $V^\pi(x)$, while the action value of the pair (x, a) under policy π is denoted by $Q^\pi(x, a)$.

The *optimal value* of a state is the value of the best possible expected return that can be obtained from that state. It is denoted by $V^*(x)$: $V^*(x) = \sup_\pi V^\pi(x)$. Similarly, the optimal value of a state-action pair is $Q^*(x, a) = \sup_\pi Q^\pi(x, a)$. A policy is called *optimal* if $V^\pi(x) = V^*(x)$ holds for all states $x \in \mathcal{X}$. A policy is called *stationary* if the actions selected at any step depend only on the last state. A stationary policy can thus be specified by specifying a distribution $\pi(\cdot|x)$ over actions for each state of the state space. Under mild regularity assumptions (e.g., assuming that $\sup_{x,a} \mathbb{E}[f_2(x, a, D)] < +\infty$ and that $\mathcal{A}$ is finite), an optimal stationary policy exists and in fact any stationary policy π that satisfies $\sum_{a \in \mathcal{A}} \pi(a|x) Q^*(x, a) = \max_{a \in \mathcal{A}} Q^*(x, a)$ for every state $x \in \mathcal{X}$ is optimal. Given some function $Q : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$, a maximizer of $Q(x, \cdot)$ is called a *greedy action* with respect to (w.r.t.) Q . A policy π is greedy w.r.t. Q if it is greedy w.r.t. Q in every state $x \in \mathcal{X}$.

Dynamic Programming Algorithms

Value- and *policy iteration* are the standard algorithms of dynamic programming to find optimal (or near-optimal) policies in MDPs.

Value iteration generates a sequence of value functions $V_{k+1} = T_1^* V_k$, $k \geq 0$, where V_0 is arbitrary. Thanks to Banach's fixed-point theorem, $(V_k; k \geq 0)$ converges to V^* at a geometric rate. Here, the operator $T_1^* : B(\mathcal{X}) \rightarrow B(\mathcal{X})$ is defined by $T_1^* V(x) = \sup_{a \in \mathcal{A}} \mathbb{E} [f_2(x, a, D) + \gamma V(f_1(x, a, D))]$, where $B(\mathcal{X})$ is the set of bounded functions over $\mathcal{X}$. The operator T_1^* is shown to be a γ -contraction w.r.t. the supremum norm.

Value iteration can also be used in conjunction with action-value functions, in which case it takes the form $Q_{k+1} = T_2^* Q_k$, $k \geq 0$, which again converges to Q^* at a geometric rate. Here, the operator $T_2^* : B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$ is defined by $T_2^* Q(x, a) = \mathbb{E} [f_2(x, a, D) + \gamma \sup_{a' \in \mathcal{A}} Q(f_1(x, a, D), a')]$ and is again a contraction. Both T_1^* and T_2^* are called *Bellman-optimality operators*.

It can be shown that once V_k (or Q_k) is close to V^* (respectively, Q^*), a policy that is greedy with respect to V_k (respectively, Q_k) will be close to optimal. In particular, the following bound is known to hold [[14], Corollary 2]:

$$V^\pi(x) \geq V^*(x) - \frac{2}{1-\gamma} \|Q - Q^*\|_\infty, \quad x \in \mathcal{X}. \quad (2)$$

Here, π is any policy that is greedy policy w.r.t. Q , and Q is an arbitrary real-valued function over the state-action space.

Let us now consider *policy iteration*. Fix an arbitrary initial policy π_0 . At iteration $k > 0$, compute the action-value function underlying π_k (this is called the *policy evaluation step*). Next, given Q^{π_k} , define π_{k+1} as a policy that is greedy with respect to Q^{π_k} (this is called the *policy improvement step*). The steps when π_{k+1} is computed from π_k define an update of *policy iteration*.

After k iterations, policy iteration gives a policy not worse than the policy that is greedy w.r.t. to the value function computed using k iterations of value iteration provided that the two procedures were started with the same initial value function. However, the computational cost of a single update in policy iteration is much higher (because of the policy evaluation step) than that of one update in value iteration.

For more information on MDPs consult Bertsekas and Shreve [15] or Puterman [16].

VALUE PREDICTION PROBLEMS

Value prediction is the problem of predicting the value of states (or state-action pairs) under a stationary policy π . The data is given in the form of a series of transitions, $\mathcal{D} = ((X_t, R_{t+1}, Y_{t+1}); t = 0, 1, \dots)$, where $(Y_{t+1}, R_{t+1}) = f(X_t, A_t, D_{t+1})$ and $A_t \sim \pi(\cdot|X_t)$. The goal is to learn a predictor for $V^\pi(\cdot)$. Here, the process generating X_t is either left unspecified, or $Y_{t+1} = X_{t+1}$; that is, the data is given in the form of a single trajectory. When X_t has a limiting distribution that corresponds to the stationary distribution of π (which is assumed to exist), we talk

about an *on-policy setting*, otherwise we talk about an *off-policy setting*. The off-policy setting comes up, for example, when one uses a simulator and time-to-time the states X_t are reset to specific states of interest. The on-policy setting comes up when such “resetting” of the state is not available, as in the case when learning happens while interacting with a real system.

TD(λ) with Linear Function Approximation

When the state space is large, it is infeasible to keep an estimate of the values of all states (the array storing the values may not fit into the memory). One way to get around this problem is to use a function approximation method, a special case of which is when a *linear function approximation method* is used. Such a method approximates the value at a state x by first calculating some *features* $\varphi(x) \in \mathbb{R}^d$ and then linearly combining these with some weights θ :

$$V^\pi(x) \approx \theta^\top \varphi(x).$$

The features are usually designed (picked) by hand so that they capture the properties of the state that the user thinks are important for predicting the values. The design of the features usually needs prior knowledge, though one can always resort to some general constructions, like tile coding or radial basis functions. Once the features are selected we are left with finding the weights so that V^π is well approximated. This is the job of the learning algorithms. Although it might seem a lot to ask the user to select the features, that the weights are tuned automatically is a major achievement whose significance should not be underestimated.

Algorithm 1. The function implementing the TD(λ) algorithm with linear function approximation. This function must be called after each transition.

function TDLambdaLinFApp(X, R, Y, θ, z)

Input: X is the last state, Y is the next state, R is the immediate reward associated with this transition, $\theta \in \mathbb{R}^d$ is the parameter vector of the linear function approximation, $z \in \mathbb{R}^d$ is the vector of eligibility traces

1: $\delta \leftarrow R + \gamma \cdot \theta^\top \varphi[Y] - \theta^\top \varphi[X]$

2: $z \leftarrow \varphi[X] + \gamma \cdot \lambda \cdot z$

3: $\theta \leftarrow \theta + \alpha \cdot \delta \cdot z$

4: **return**(θ, z)

The TD(λ) algorithm of Sutton [17] (see also Sutton [18]) is one widely used and successful method that allows one to tune the weights in an incremental manner. The pseudocode of the update rule for TD(λ) is shown as Algorithm 1. The routine shown must be called for each sampled transition. The name of the algorithm is derived from the abbreviation of temporal differences and it comes from line 1, where δ , which can be thought of as a temporal difference error is computed. In addition to the data underlying a transition, the method expects the current weights θ , and the current *eligibility traces* z as inputs. Here, both θ and z are d -dimensional vectors. The routine then returns the updated values of these vectors.

The tuning parameters of TD(λ) are $\lambda \in [0, 1]$ and the step-size $\alpha \geq 0$. In practice, one often uses d step-sizes, one for each component of the weight vector. The step-sizes should be inversely proportional to the expected magnitudes of the respective feature components. The λ parameter determines the amount of “bootstrapping.” When the dynamics is well preserved by the feature-extraction method (e.g., $\varphi(Y_{t+1})$ and the reward can be well predicted given $\varphi(X_t)$ and the action), a small value of λ is appropriate (and the algorithm can be thought of as implementing an incremental form of value iteration). In the opposite case, $\lambda \approx 1$ is more appropriate and the algorithm can be thought of as implementing a Monte Carlo algorithm. In general, increasing λ increases the variance, but decreases the bias. The best value of λ is problem dependent and is usually found by trial-and-error in small-scale experiments before the algorithm is run on a larger chunk of data. When $\lambda = 0$, the algorithm is also called TD(0), while it is also called TD(1) when $\lambda = 1$.

Almost sure convergence of the weights is guaranteed provided that (i) the stochastic process $(X_t; t \geq 0)$ is an ergodic Markov process whose stationary distribution μ is the same as the stationary distribution of the policy π ; and (ii) the step-size sequence satisfies the so-called Robbins–Monro (RM) conditions [[2], p. 222, Section 5.3.7; [19]].³ When the first condition is not met, that is, under off-policy sampling, convergence is no longer guaranteed, but in fact the parameters may diverge [[2], Example 6.7, p. 307]. This is true for linear function approximation when the distributions of $(X_t; t \geq 0)$ do not match the stationary distribution of π . The TD(λ) method has a version that is suitable for tuning the parameters of nonlinear function approximation methods, such as neural networks. However, this standard version might diverge [[2], Example 6.6, p. 292]. For further examples of instability, see Baird [20] and Boyan and Moore [21].

Gradient Temporal Difference Learning.

In this section, we present two algorithms, which overcome the instability of TD(λ) while retaining its incremental nature and the computational efficiency of the updates.

The key is to introduce an objective function and derive a gradient method from it. For simplicity, we consider the case when $\lambda = 0$. Define the objective function

$$J(\theta) = \mathbb{E} [\delta_{t+1}(\theta) \varphi_t]^\top \mathbb{E} [\varphi_t \varphi_t^\top]^{-1} \mathbb{E} [\delta_{t+1}(\theta) \varphi_t], \quad (3)$$

where

$$\begin{aligned} \delta_{t+1}(\theta) &= R_{t+1} + \gamma V_\theta(Y_{t+1}) - V_\theta(X_t) \\ &= R_{t+1} + \gamma \theta^\top \varphi'_{t+1} - \theta^\top \varphi_t, \\ \varphi_t &= \varphi(X_t), \\ \varphi'_{t+1} &= \varphi(Y_{t+1}). \end{aligned} \quad (4)$$

³The RM conditions state the following: if $(\alpha_t; t \geq 0)$ are the step-sizes used in time step t , then $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. They are satisfied by, for example, $\alpha_t = 1/t$. In practice, people often use small constant step-sizes or adaptive step-size selection methods.

Here, for simplicity we assumed that $(X_t; t \geq 0)$ is a stationary process, so the expectation in Equation (3) is well defined.

When TD(0) converges, it converges to a unique vector θ^* that satisfies

$$\mathbb{E} [\delta_{t+1}(\theta^*) \varphi_t] = 0. \quad (5)$$

In this case, the minimizer of J will coincide with θ^* . Therefore, it suffices to design an algorithm to minimize J . Accordingly, from now on let us redefine θ^* as the minimizer of J (for simplicity, assume that the minimizer is unique).

Taking the gradient of J , we get

$$\nabla_\theta J(\theta) = -2 \mathbb{E} [(\varphi_t - \gamma \varphi'_{t+1}) \varphi_t^\top] w(\theta), \quad (6)$$

where

$$w(\theta) = \mathbb{E} [\varphi_t \varphi_t^\top]^{-1} \mathbb{E} [\delta_{t+1}(\theta) \varphi_t].$$

Let us introduce two sets of weights: θ_t to approximate θ_* and w_t to approximate $w(\theta_*)$. In GTD2 (“gradient temporal difference learning, version 2”), the update of θ_t is chosen to follow the negative stochastic gradient of J based on Equation (6) assuming that $w_t \approx w(\theta_t)$, while the update of w_t is chosen so that for any fixed θ , w_t would converge almost surely to $w(\theta)$. The pseudocode of the resulting update, which Sutton *et al.* [22] called the GTD2 update, is shown as Algorithm 2. This algorithm needs two sets of step-sizes (α, β) .

Sutton *et al.* [22] have shown that (under the standard RM conditions on the step-sizes and some mild regularity conditions) the weight-sequence (θ_t) updated with GTD2 converges to the minimizer of $J(\theta)$ almost surely. However, unlike in the case of TD(0), convergence is guaranteed independently of what the distribution of $(X_t; t \geq 0)$ is. At the same time, the update of GTD2 costs only twice as much as that of TD(0).

One can arrive at another gradient method if the gradient of J is written as

$$\begin{aligned} \nabla_\theta J(\theta) &= -2 \left(\mathbb{E} [\delta_{t+1}(\theta) \varphi_t] - \right. \\ &\quad \left. \gamma \mathbb{E} [\varphi'_{t+1} \varphi_t^\top] w(\theta) \right). \end{aligned}$$

This suggests replacing the update in line 5 of GTD2 by

$$\theta \leftarrow \theta + \alpha \cdot (\delta \cdot f - \gamma \cdot a \cdot f'),$$

giving rise to the TDC (“temporal difference learning with corrections”) update of Sutton *et al.* [22]. Here the update of w_t must

use larger step-sizes than the update of θ_t : $\alpha_t = o(\beta_t)$. This makes TDC a member of the family of the so-called *two-timescale stochastic approximation algorithms* [23,24]. If, in addition to this, the standard RM conditions are also satisfied by both step-size sequences, $\theta_t \rightarrow \theta_*$ holds again almost surely [22].

Algorithm 2. The function implementing the GTD2 algorithm. This function must be called after each transition.

function GTD2(X, R, Y, θ, w)

Input: X is the last state, Y is the next state, R is the immediate reward associated with this transition, $\theta \in \mathbb{R}^d$ is the parameter vector of the linear function approximation, $w \in \mathbb{R}^d$ is the auxiliary weight

```

1:  $f \leftarrow \varphi[X]$ 
2:  $f' \leftarrow \varphi[Y]$ 
3:  $\delta \leftarrow R + \gamma \cdot \theta^\top f' - \theta^\top f$ 
4:  $a \leftarrow f^\top w$ 
5:  $\theta \leftarrow \theta + \alpha \cdot (f - \gamma \cdot f') \cdot a$ 
6:  $w \leftarrow w + \beta \cdot (\delta - a) \cdot f$ 
7: return  $(\theta, w)$ 
```

More recently, these algorithms have been extended to nonlinear function approximation [25] and to the $\lambda > 0$ case [26].

Note that although these algorithms are derived from the gradient of an objective function, they are not true stochastic gradient methods in the sense that the expected weight update direction can be different from the direction of the negative gradient of the objective function. In fact, these methods belong to the larger class of pseudogradient methods. The two methods differ in how they approximate the gradients. It remains to be seen whether one of them is better than the other.

LSTD: Least-Squares Temporal Difference Learning. As said earlier, in the limit, TD(0) converges to the solution of Equation (5). Given a finite sample

$$\mathcal{D}_n = ((X_0, R_1, Y_1), (X_1, R_2, Y_2), \dots, (X_{n-1}, R_n, Y_n)),$$

one can approximate Equation (5) by

$$\frac{1}{n} \sum_{t=0}^{n-1} \varphi_t \delta_{t+1}(\theta) = 0. \quad (7)$$

Plugging in $\delta_{t+1}(\theta) = R_{t+1} - (\varphi_t - \gamma \varphi'_{t+1})^\top \theta$, we see that this equation is linear in θ . In particular, if the matrix $\hat{A}_n = \frac{1}{n} \sum_{t=0}^{n-1} \varphi_t (\varphi_t - \gamma \varphi'_{t+1})^\top$ is invertible, the solution is simply

$$\theta_n = \hat{A}_n^{-1} \hat{\delta}_n, \quad (8)$$

where $\hat{\delta}_n = \frac{1}{n} \sum_{t=0}^{n-1} R_{t+1} \varphi_t$.

If inverting the matrix can be afforded (e.g., the dimensionality of the features is not too large and the method is not called too many times), this method can give a better approximation to θ^* than TD(0) or some other incremental first-order method, since the latter are negatively impacted by the eigenvalue spread of the matrix $A = \mathbb{E}[\hat{A}_n]$.

The idea of directly computing the solution of Equation (7) is due to Bradtke and Barto [27], who call the resulting algorithm *least-squares temporal difference* (LSTD) learning. Using the terminology of stochastic programming, LSTD can be seen to use *sample average approximation* [28]. In the terminology of statistics, it belongs to the so-called *Z-estimation* family of procedures [[29], Section 2.2.5].

Algorithm 3. The function implementing the RLSTD algorithm. This function must be called after each transition. Initially, C should be set to a diagonal matrix with small positive diagonal elements: $C = \beta I$, with $\beta > 0$.

function RLSTD(X, R, Y, C, θ)

Input: X is the last state, Y is the next state, R is the immediate reward associated with this transition, $C \in \mathbb{R}^{d \times d}$, and $\theta \in \mathbb{R}^d$ is the parameter vector of the linear function approximation

```

1:  $f \leftarrow \varphi[X]$ 
2:  $f' \leftarrow \varphi[Y]$ 
3:  $g \leftarrow (f - \gamma f')^\top C$   $g$  is a  $1 \times d$  row vector
4:  $a \leftarrow 1 + g f$ 
5:  $v \leftarrow C f$ 
6:  $\delta \leftarrow R + \gamma \cdot \theta^\top f' - \theta^\top f$ 
7:  $\theta \leftarrow \theta + \delta / a \cdot v$ 
8:  $C \leftarrow C - v g / a$ 
9: return ( $C, \theta$ )
```

Using the Sherman–Morrison formula, one can derive an incremental version of LSTD, analogously to how the recursive least-squares (RLS) method is derived in adaptive filtering [30]. The resulting algorithm is called *recursive least-squares temporal difference (RLSTD)* [27] and is shown as Algorithm 3. The computational complexity of one update is $O(d^2)$. When the number of samples n is large, it can be faster to use the direct form of LSTD that uses matrix inversion.

Boyan [31] extended LSTD to incorporate the λ parameter of TD(λ) and called the resulting algorithm LSTD(λ). (Note that for $\lambda > 0$ to make sense one needs $X_{t+1} = Y_{t+1}$, otherwise the TD errors do not telescope.) The recursive form of LSTD(λ), RLSTD(λ), has been studied by Xu *et al.* [32] and (independently) by Nedić and Bertsekas [33].

An alternative to LSTD (and LSTD(λ)) is λ *least-squares policy evaluation* (λ -LSPE for short) due to Bertsekas and Ioffe [34], which implements a form of fitted-value iteration.

Comparing Least-Squares and TD-Like Methods. To compare the two approaches, fix some time T available for computation and assume that samples are cheap to obtain (e.g., an efficient simulator is available). In time T , the least-squares methods are limited to process a sample of size $n \approx T/d^2$, while the lightweight methods can process a sample of size $n' \approx nd$. Let us now look at the precision of the resulting parameters. Assume

that the limit of the parameters is θ_* . Denote by θ_t the parameter obtained by (say) LSTD after processing t observations and denote by θ'_t the parameter obtained by a TD method. Then, one expects that $\|\theta_t - \theta_*\| \approx C_1 t^{-\frac{1}{2}}$ and $\|\theta'_t - \theta_*\| \approx C_2 t^{-\frac{1}{2}}$. Thus,

$$\frac{\|\theta'_{n'} - \theta_*\|}{\|\theta_n - \theta_*\|} \approx \frac{C_2}{C_1} d^{-\frac{1}{2}}. \quad (9)$$

Hence, if $C_2/C_1 < d^{1/2}$ then the lightweight TD-like method will achieve a better accuracy, while in the opposite case the least-squares procedures will perform better. As usual, it is difficult to decide this *a priori*. As a rule of thumb, based on Equation (8), we expect that when d is relatively small, least-squares methods might be converging faster, while if d is large then the lightweight, incremental methods will give better results. Notice that this analysis is not specific to RL methods, but applies in all cases when an incremental lightweight procedure is compared to a least-squares-like procedure and samples are cheap (for a similar analysis in a supervised learning problem, see, e.g., Bottou and Bousquet [35]). One attempt to marry the advantages of the computational efficiency of incremental methods and the statistical efficiency of least-squares methods is described by Geramifard *et al.* [36].

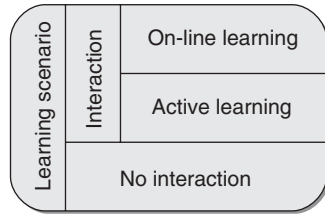


Figure 3. Types of reinforcement problems.

CONTROL

Figure 3 shows the basic types of control learning problems. The first criterion that the space of problems is split upon is whether the learner can actively influence the observations. In case she can, we talk about *interactive learning*, otherwise one is facing a *noninteractive learning* problem. Interactive learning is potentially easier since the learner has the additional option to influence the distribution of the sample. However, the goal of learning, is usually different in the two cases, making these problems incomparable in general.

In the case of noninteractive learning, the natural goal is to find a good policy given the observations. A common situation is when the sample is fixed. For example, the sample can be the result of some experimentation with some physical system that happened before learning started. In machine learning terms, this corresponds to *batch learning*. (Batch learning problems are not to be confused with batch learning methods, which are the opposite of incremental, also known as *recursive* or *iterative* methods.) Since the observations are uncontrolled, the learner working with a fixed sample has to deal with an off-policy learning situation. In other cases, the learner can ask for more data (i.e., when a simulator is used to generate new data). Here, the goal might be to learn a good policy as quickly as possible.

Consider now interactive learning. One possibility is that learning happens while interacting with a real system in a closed-loop manner. A reasonable goal then is to optimize *on-line performance*, making the learning problem an instance of *on-line learning*. In on-line learning, the key is to explore in a clever manner.

On-line performance can be measured in different ways. A natural measure is to use the sum of rewards incurred during learning. If this is offset by the expected sum of rewards under the optimal policy, one gets the notion of regret. Auer *et al.* [37] describe a state-of-the-art algorithm, UCRL2, that works in finite MDPs and give an overview of previous results. An alternative cost measure is the number of times the learner's future expected return falls short of the optimal return, that is, the number of times the learner commits a "mistake" [38]. Szita and Szepesvári [39] review the literature on this and describe a state-of-the-art algorithm, MORMAX, that commits a polynomial number of "mistakes" in finite MDPs. Another possible goal is to produce a well-performing policy as soon as possible (or find a good policy given a finite number of samples), just like in noninteractive learning. As opposed to the noninteractive situation, however, here the learner has the option to control the samples so as to maximize the chance of finding such a good policy. This learning problem is an instance of *active learning*. The E^3 -algorithm of Kearns and Singh [40] explores an unknown MDP and stops when it knows a good policy for the state *just visited*. E^3 needs a polynomial number of interactions and uses poly-resources in the relevant parameters of the problem before it stops. In a follow-up work, Brafman and Tennenholtz [41] introduced the R-max algorithm which refines the E^3 algorithm and proved similar results.

There exist only a few experimental studies in on-line learning. Jong and Stone [42] proposed a method that can be interpreted as a practical implementation of the ideas in Kakade *et al.* [38], while Nouri and Littman [43] experimented with multiresolution regression trees and the so-called fitted Q-iteration algorithm (reviewed in the next section). The main message of these works is that explicit exploration control can indeed be beneficial.

When a simulator is available, the learning algorithms can be used to solve *planning problems*. In planning, on-line learning becomes irrelevant and the algorithms'

running time and memory requirements become the primary concern. Kearns *et al.* [44], Szepesvári [45], Kocsis and Szepesvári [46], and Chang *et al.* [9] discuss some ideas for planning.

When the state or action space is large, one needs to resort to function approximation. Therefore, in what follows, we focus on the core algorithmic problem of learning good controllers in large spaces when a function approximation method is used and neglect the exploration issue. With this we do not want to imply that clever exploration methods are unimportant, but we merely make the point that the clever methods will need to work with efficient methods that use function approximation.

Algorithm 4. The function implementing the Q -learning algorithm with linear function approximation. This function must be called after each transition.

function QLEARNINGLINFAPP(X, A, R, Y, θ)

Input: X is the last state, Y is the next state, R is the immediate reward associated with this transition, $\theta \in \mathbb{R}^d$ parameter vector
 1: $\delta \leftarrow R + \gamma \cdot \max_{a' \in \mathcal{A}} \theta^\top \varphi[Y, a'] - \theta^\top \varphi[X, A]$
 2: $\theta \leftarrow \theta + \alpha \cdot \delta \cdot \varphi[X, A]$
 3: **return** θ

Although Q -learning is widely used in practice, little can be said about its convergence properties. In fact, since $\text{TD}(\lambda)$ is a special case of this algorithm (when there is only one action for every state), just like $\text{TD}(\lambda)$, this update rule will also fail to converge when off-policy sampling or nonlinear function approximation is used (cf. the section titled “ $\text{TD}(\lambda)$ with Linear Function Approximation”).

Algorithm 5. The function implementing one iteration of the fitted Q -iteration algorithm. The function must be called until some criterion of convergence is met. The methods `PREDICT` and `REGRESS` are specific to the regression method chosen. The method `PREDICT( $z, \theta$ )` should return the predicted value at the input z given the regression parameters θ , while `REGRESS( $S$ )`, given a list of input–output pairs S , should implement a regression algorithm that solves the regression problem given by S and returns new parameters that can be used in `PREDICT`.

function FITTEDQ(D, θ)

Input: $D = ((X_i, A_i, R_{i+1}, Y_{i+1}); i = 1, \dots, n)$ is a list of transitions, θ are the regressor parameters
 1: $S \leftarrow []$ ▷ Create empty list
 2: **for** $i = 1 \rightarrow n$ **do**
 3: $T \leftarrow R_{i+1} + \max_{a' \in \mathcal{A}} \text{predict}((Y_{i+1}, a'), \theta)$ ▷ Target at (X_i, A_i)
 4: $S \leftarrow \text{APPEND}(S, (X_i, A_i, T))$
 5: **end for**
 6: $\theta \leftarrow \text{REGRESS}(S)$
 7: **return** θ

Direct Methods

Direct methods aim at approximating the optimal action-value function directly. One class of methods, two members of which are reviewed here, can be thought of as implementing a form of value iteration with action-value functions.

Q -Learning with Function Approximation.

Assume that we are given features $\varphi(x, a) \in \mathbb{R}^d$ of state-action pairs. The problem is to find weights $\theta \in \mathbb{R}^d$ such that $Q^*(x, a) \approx \theta^\top \varphi(x, a)$ holds for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. The Q -learning algorithm by Watkins [47], whose update is shown as Algorithm 4, can be thought of as the extension of $\text{TD}(\lambda)$ to this case.

Fitted Q -Iteration. Fitted Q -iteration is an instance of the generic fitted-value iteration recipe. Given the previous iterate, Q_t , the idea is to form a Monte Carlo approximation to $T^*Q_t(x, a)$ at select state-action pairs and then learn a regressor based on the so-created data set. Algorithm 5 shows the corresponding pseudocode. This method must be called repeatedly in a loop until convergence.

It is known that fitted Q -iteration might diverge unless a special regressor is used [20,21,48].

Ormoneit and Sen [49] suggest to use kernel averaging, while Ernst *et al.* [50] suggest using tree-based regressors. These are guaranteed to converge (say, if the same data is fed to the algorithm in each iteration) as they implement local averaging and therefore the results of Gordon [51] and Tsitsiklis and Van Roy [48] are applicable to them. Riedmiller [52] reports good empirical results with neural networks, at least when new observations obtained by following a policy greedy with respect to the latest iterate are incrementally added to the set of samples used in the updates. That the sample is changed is essential if no good initial policy is available, that is, when in the initial sample the states that are frequently visited by “good” policies are underrepresented (a theoretical argument for why this is important is given by Van Roy [53] in the context of state aggregation).

Antos *et al.* [54] and Munos and Szepesvári [55] prove finite-sample performance bounds that apply to a large class of regression methods that use empirical risk minimization over a fixed space $\mathcal{F}$ of candidate action-value functions. Their bounds depend on the *worst-case Bellman error* of $\mathcal{F}$:

$$e_1^*(\mathcal{F}) = \sup_{Q \in \mathcal{F}} \inf_{Q' \in \mathcal{F}} \|Q' - T^*Q\|_\mu,$$

where μ is the distribution of state-action pairs in the training sample. That is, $e_1^*(\mathcal{F})$ measures how close $\mathcal{F}$ is to $T^*\mathcal{F} \stackrel{\text{def}}{=} \{T^*Q \mid Q \in \mathcal{F}\}$. The bounds derived have the form of the finite-sample bounds that hold in supervised learning, except that the approximation error is measured by $e_1^*(\mathcal{F})$. Note that in the earlier-mentioned counterexamples to the convergence of fitted-value iteration, $e_1^*(\mathcal{F}) = \infty$, suggesting that it is the lack of flexibility of the function approximation method that causes divergence.

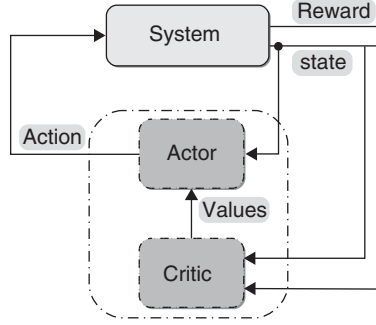


Figure 4. The actor-critic architecture.

Actor-Critic Methods

Actor-critic methods implement generalized policy iteration (GPI). Remember that policy iteration works by alternating between a complete policy evaluation and a complete policy improvement step. When using sample-based methods or function approximation, exact evaluation of the policies may require infinitely many samples or might be impossible due to the restrictions of the function approximation technique. Hence, RL algorithms simulating policy iteration must change the policy based on incomplete knowledge of the value function.

Algorithms that update the policy before it is completely evaluated are said to implement GPI. In GPI, there are two closely interacting processes of an actor and a critic: the actor aims at improving the current policy, while the critic evaluates the current policy, thus helping the actor. The interaction of the actor and the critic is illustrated in Figure 4 in a closed-loop learning situation.

Implementing a Critic. The job of the critic is to estimate the value of the current target policy of the actor. This is a value prediction problem. Therefore, the critic can use any value prediction method. Since the actor needs action values, the algorithms are typically modified so that they estimate action values directly. When $TD(\lambda)$ is appropriately extended, the algorithm known as SARSA(λ) is obtained due to Rummery and Niranjan [56]. The algorithm is named from its use of the current State, current Action, next Reward, next State, and next Action. The

pseudocode of the corresponding update rule is shown as Algorithm 6.

Being a TD algorithm, the resulting algorithm is subject to the same limitations as TD(λ) (cf. the section titled “TD(λ) with Linear Function Approximation”), that is, it might diverge in off-policy situations. It is, however, possible to extend GTD2 and TDC to work with action values (and use $\lambda > 0$) so that the resulting algorithms would become free of these limitations. For details consult Maei and Sutton [26].

Algorithm 6. The function implementing the SARSA(λ) algorithm with linear function approximation. This function must be called after each transition.

function SALSALAMBDA LINFAPP($X, A, R, Y, A', \theta, z$)

Input: X is the last state, A is the last action chosen, R is the immediate reward received when transitioning to Y , where action A' is chosen. $\theta \in \mathbb{R}^d$ is the parameter vector of the linear function approximation, $z \in \mathbb{R}^d$ is the vector of eligibility traces

- 1: $\delta \leftarrow R + \gamma \cdot \theta^\top \varphi[Y, A'] - \theta^\top \varphi[X, A]$
- 2: $z \leftarrow \varphi[X, A] + \gamma \cdot \lambda \cdot z$
- 3: $\theta \leftarrow \theta + \alpha \cdot \delta \cdot z$
- 4: **return** (θ, z)

Consider a smoothly parameterized policy class $\Pi = (\pi_\omega; \omega \in \mathbb{R}^{d_\omega})$ of stochastic stationary policies. When the action space is finite, a popular choice for Π is to use the so-called *Gibbs policies*:

$$\pi_\omega(a|x) = \frac{\exp(\omega^\top \xi(x, a))}{\sum_{a' \in \mathcal{A}} \exp(\omega^\top \xi(x, a'))},$$

$x \in \mathcal{X}, a \in \mathcal{A}.$

Here, $\xi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^{d_\omega}$ is an appropriate feature-extraction function. On the other hand, if the action space is a subset of a $d_{\mathcal{A}}$ -dimension Euclidean space, a popular choice is to use the Gaussian policies, in which case, given some parametric mean $g_\omega(x, a)$ and covariance $\Sigma_\omega(x, a)$ functions, the density specifying the action-selection distribution under ω is defined by

$$\begin{aligned} \pi_\omega(a|x) = & \frac{1}{\sqrt{(2\pi)^{d_{\mathcal{A}}} \det(\Sigma_\omega(x, a))}} \\ & \times \exp(-(a - g_\omega(x, a))^\top \\ & \times \Sigma_\omega^{-1}(x, a)(a - g_\omega(x, a))). \end{aligned}$$

Implementing an Actor. Policy improvement can be implemented in two ways: One idea is moving the current policy toward the greedy policy underlying the approximate action-value function obtained from the critic. Another idea is to perform gradient ascent directly on the performance surface underlying a chosen parametric policy class. Here, we describe a method based on the latter idea.

Care must be taken to ensure that Σ_ω is positive definite. Often, for simplicity, Σ_ω is taken to be $\Sigma_\omega = \beta I$ with some $\beta > 0$.

Given Π , formally, the problem is to find the value of ω corresponding to the best performing policy:

$$\operatorname{argmax}_{\omega} \rho_\omega = ?$$

Here, the performance, ρ_ω , is usually the expected return of policy π_ω with respect to the initial distribution over the states.

One possible actor-critic algorithm is shown in Algorithm 7. A special assumption in this method is that the features ψ_ω used by SARSA(λ), called in line 6, are *compatible* with π_ω in the sense that $\psi_\omega(x, a) = \frac{\partial}{\partial \omega} \log \pi_\omega(x, a)$. Note that the features depend on the value of ω , which, in the pseudocode, is signified by the presence of ω as a subindex to SALSALAMBDA LINFAPP. Konda and Tsitsiklis [57] proved that (under some regularity conditions) $\liminf_{t \rightarrow \infty} \nabla_\omega \rho_{\omega_t} = 0$ holds almost surely if the average-cost version of SARSA(1) is used to update θ_t . They have also shown that if SARSA(λ) is used and $m_\lambda = \liminf_{t \rightarrow \infty} \nabla_\omega \rho_{\omega_t}$, then $\lim_{\lambda \rightarrow 1} m_\lambda = 0$.

Algorithm 7. An actor-critic algorithm that uses compatible function approximation and SARSA(1).

```

function SARSAACTORCRITIC
1:  $\omega, \theta, z \leftarrow 0$ 
2:  $A \leftarrow a'_1$ 
3: repeat
4:    $(R, Y) \leftarrow \text{EXECUTEINWORLD}(A)$ 
5:    $A' \leftarrow \text{DRAW}(\pi_\omega(Y, \cdot))$ 
6:    $(\theta, z) \leftarrow \text{SARSALAMBDALINFAPP}_\omega(X, A, R, Y, A', \theta, z)$   $\triangleright$  Use  $\lambda = 1$  and  $\alpha \gg \beta$ 
7:    $\psi \leftarrow \frac{\partial}{\partial \omega} \log \pi_\omega(X, A)$ 
8:    $v \leftarrow \text{SUM} \pi_\omega(Y, \cdot) \cdot \theta^\top \varphi[X, \cdot]$ 
9:    $\omega \leftarrow \omega + \beta \cdot (\theta^\top \varphi[X, A] - v) \cdot \psi$ 
10:   $X \leftarrow Y$ 
11:   $A \leftarrow A'$ 
12: until True

```

Another interesting algorithm is obtained if one replaces the update in line 9 with

$$\omega \leftarrow \omega + \beta \cdot \theta,$$

which is called the *natural actor-critic* (NAC) update. Under some mild conditions, this algorithm can be seen to implement a stochastic pseudogradient algorithm, and thus the previous results extend to it.

For a fixed value of ω , the limit of the SARSA update, $\theta_*(\omega)$, can be shown to be a so-called *natural gradient* [58] of ρ_ω . This was first noted by Kakade [59]. It is believed that following a natural gradient generally improves the behavior of gradient ascent methods. This is nicely demonstrated by Kakade [59] on a simple two-state MDP, where the “normal” gradient is very small in a large part of the parameter space, while the natural gradient behaves in a reasonable manner. Other positive examples were given by Bagnell and Schneider [60], Peters *et al.* [61], and Peters and Schaal [62].

FOR FURTHER EXPLORATION

Inevitably, due to space constraints, in this review a large portion of the RL literature has not been mentioned.

One topic of particular interest not discussed is efficient sampling-based planning [9, 44–46]. The main lesson here is that off-line planning in the worst case can scale exponentially with the dimensionality of the

state space [63], while on-line planning (i.e., planning for the “current state”) can break the curse of dimensionality by amortizing the planning effort over multiple time steps [45, 64].

Other topics of interest include the linear programming-based approaches [65–67], dual dynamic programming [68], techniques based on sample average approximation [28] such as PEGASUS [69], on-line learning in MDPs with arbitrary reward processes [70–72], or learning with (almost) no restrictions in a competitive framework [73].

Other important topics include learning and acting in partially observed MDPs (for recent developments, see, e.g., Littman *et al.* [74], Toussaint *et al.* [75], Ross *et al.* [76]), learning and acting in games or under some other optimization criteria [77–80], or the development of hierarchical and multi-time-scale methods [81, 82].

Applications

The numerous successful applications of RL include (in no particular order) games, for example, Backgammon [83] and Go [84]; applications in networking, for example, packet routing [85] and channel allocation [86]; various operations research problems, for example, targeted marketing [87], maintenance problems [88], job-shop scheduling [89], elevator control [90], pricing [91], vehicle routing [92], inventory control [93], and fleet management [94]; learning in robotics, for example, controlling quadrupedals [95],

humanoid robots [61], or helicopters [96]; and finance, for example, option pricing [97–100]. For further applications, see, for example, the lists at the URLs

- <http://www.cs.ualberta.ca/szepesva/RESEARCH/RLApplications.html>
- <http://umichrl.pbworks.com/Successes-of-Reinforcement-Learning>.

Software

There are numerous software packages that support the development and testing of RL algorithms. Perhaps the most notable of these are the RL-GLUE and RL-LIBRARY packages. The RL-GLUE package available at <http://glue.rl-community.org> is intended for helping to standardize RL experiments. It is a free, language-neutral software package that implements a standardized RL interface [101]. The RL-LIBRARY (<http://library.rl-community.org>) builds on the top of RL-GLUE. Its purpose is to provide trusted implementations of various RL test beds and algorithms. The most notable other RL software packages are CLSquare,⁴ PIQLE,⁵ RL Toolbox,⁶ JRLF⁷, and LibPG.⁸ These offer the implementation of a large number of algorithms, test beds, intuitive visualizations, programming tools, and so on. Many of these packages support RL-GLUE.

REFERENCES

1. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res* 1996;4:237–285.
2. Bertsekas DP, Tsitsiklis JN. *Neuro-dynamic programming*. Belmont (MA): Athena Scientific; 1996.
3. Sutton RS, Barto AG. *Reinforcement learning: an introduction*. Cambridge (MA): MIT Press; 1998.
4. <http://www.ni.uos.de/index.php?id=70>.
5. <http://piqle.sourceforge.net/>.
6. <http://www.igi.tugraz.at/ril-toolbox/>.
7. http://mykel.kochenderfer.com/?page_id=19.
8. <http://code.google.com/p/libpg/>.
4. Gosavi A. *Simulation-based optimization: parametric optimization techniques and reinforcement learning*. Netherlands: Springer; 2003.
5. Cao XR. *Stochastic learning and optimization: a sensitivity-based approach*. New York: Springer; 2007.
6. Bertsekas DP. Volume 1, *Dynamic programming and optimal control*. 3rd ed. Belmont (MA): Athena Scientific; 2007.
7. Bertsekas DP. Volume 2, *Dynamic programming and optimal control*. 3rd ed. Belmont (MA): Athena Scientific; 2007.
8. Powell WB. *Approximate dynamic programming: solving the curses of dimensionality*. New York: John Wiley & Sons; 2007.
9. Chang HS, Fu MC, Hu J, *et al.* *Simulation-based algorithms for Markov decision processes*. London: Springer; 2008.
10. Busoniu L, Babuska R, De Schutter B, *et al.* *Reinforcement learning and dynamic programming using function approximators*. Automation and Control Engineering Series. Boca Raton (FL): CRC Press; 2010.
11. Szepesvári Cs. *Reinforcement learning. Synthesis lectures on artificial intelligence and machine learning*. Morgan & Claypool Publishers; 2010.
12. Bertsekas DP. *Approximate dynamic programming (online chapter)*. Volume 2, *Dynamic programming and optimal control*. 3rd ed. Belmont (MA): Athena Scientific; 2010. Chapter 6. Available at <http://web.mit.edu/dimitrib/www/dpchapter.pdf>.
13. Szepesvári Cs. *Reinforcement learning algorithms for MDPs – a survey*. Technical Report TR09-13. Department of Computing Science, University of Alberta; 2009.
14. Singh SP, Yee RC. An upper bound on the loss from approximate optimal-value functions. *Mach Learn* 1994;16(3):227–233.
15. Bertsekas DP, Shreve SE. *Stochastic optimal control (the discrete time case)*. New York: Academic Press; 1978.
16. Puterman ML. *Markov decision processes—discrete stochastic dynamic programming*. New York (NY): John Wiley & Sons, Inc.; 1994.
17. Sutton RS. *Temporal credit assignment in reinforcement learning [PhD thesis]*. Amherst (MA): University of Massachusetts; 1984.
18. Sutton RS. Learning to predict by the method of temporal differences. *Mach Learn* 1988;3(1):9–44.

19. Tsitsiklis JN, Van Roy B. An analysis of temporal difference learning with function approximation. *IEEE Trans Autom Control* 1997;42:674–690.
20. Baird LC. Residual algorithms: reinforcement learning with function approximation. In: Russell SJ, Prieditis A, editors. *Proceedings of the 12th International Conference on Machine Learning (ICML 1995)*. San Francisco (CA): Morgan Kaufmann; 1995. pp. 30–37. ISBN 1-55860-377-8.
21. Boyan JA, Moore AW. Generalization in reinforcement learning: safely approximating the value function. In: Tesauro G, Touretzky D, Leen T, editors. *NIPS-7: Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, Cambridge (MA): MIT Press; 1995. pp. 369–376.
22. Sutton RS, Maei HR, Precup D, *et al.* Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: Danyluk AP, Bottou L, Littman ML, editors. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*. Volume 382, ACM International Conference Proceeding Series. New York (NY): ACM; 2009. pp. 993–1000. ISBN 978-1-60558-516-1.
23. Borkar VS. Stochastic approximation with two time scales. *Syst Control Lett* 1997; 29(5):291–294.
24. Borkar VS. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge: Cambridge University Press; 2008.
25. Maei HR, Szepesvári Cs, Bhatnagar S, *et al.* Convergent temporal-difference learning with arbitrary smooth function approximation. In: Lafferty J and Williams C, editors. *Advances in Neural Information Processing Systems 22*. Cambridge (MA): MIT Press; 2010. pp. 1204–1212.
26. Maei HR, Sutton RS. $GQ(\lambda)$: a general gradient algorithm for temporal-difference prediction learning with eligibility traces. In: Baum E, Hutter M, Kitzelmann E, editors. *Proceedings of the Third Conference on Artificial General Intelligence*. Paris, France: Atlantis Press; 2010. pp. 91–96.
27. Bradtke SJ, Barto AG. Linear least-squares algorithms for temporal difference learning. *Mach Learn* 1996;22:33–57.
28. Shapiro A. *Monte Carlo sampling methods*. Volume 10, *Stochastic programming, handbooks in OR & MS*. Amsterdam: North-Holland Publishing Company; 2003.
29. Kosorok MR. *Introduction to empirical processes and semiparametric inference*. New York (NY): Springer; 2008.
30. Widrow B, Stearns SD. *Adaptive signal processing*. Englewood Cliffs (NJ): Prentice Hall; 1985.
31. Boyan JA. Technical update: least-squares temporal difference learning. *Mach Learn* 2002;49:233–246.
32. Xu X, He H, Hu D. Efficient reinforcement learning using recursive least-squares methods. *J Artif Intell Res* 2002;16:259–292.
33. Nedić A, Bertsekas DP. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dyn Syst* 2003;13(1):79–110.
34. Bertsekas DP, Ioffe S. Temporal differences-based policy iteration and applications in neuro-dynamic programming. LIDS-P-2349. MIT; 1996.
35. Bottou L, Bousquet O. The tradeoffs of large scale learning. In: Platt JC, Koller D, Singer Y, *et al.* editors. *Advances in Neural Information Processing Systems 20*. Cambridge (MA): MIT Press; 2008. pp. 161–168.
36. Geramifard A, Bowling MH, Zinkevich M, *et al.* iLSTD: eligibility traces and convergence analysis. In: Schölkopf B, Platt JC, Hoffman T, editors. *Advances in Neural Information Processing Systems 19*. Cambridge (MA): MIT Press; 2007. pp. 441–448. ISBN 0-262-19568-2.
37. Auer P, Jaksch T, Ortner R. Near-optimal regret bounds for reinforcement learning. *J Mach Learn Res* 2010;11:1563–1600.
38. Kakade S, Kearns MJ, Langford J. Exploration in metric state spaces. In: Fawcett T, Mishra N, editors. *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*. Washington (DC): AAAI Press; 2003. pp. 306–312. ISBN 1-57735-189-4.
39. Szita I, Szepesvári Cs. Model-based reinforcement learning with nearly tight exploration complexity bounds. In: Wrobel S, Fürnkranz J, Joachims T, editors. *Proceedings of the 27th Annual International Conference on Machine Learning (ICML 2010)*. ACM International Conference Proceeding Series. New York (NY): ACM; 2010.
40. Kearns MJ, Singh SP. Near-optimal reinforcement learning in polynomial time. *Mach Learn* 2002;49(2–3):209–232.

41. Brafman RI, Tennenholtz M. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *J Mach Learn Res* 2002;3:213–231.
42. Jong NK, Stone P. Model-based exploration in continuous state spaces. In: Miguel I, Ruml W, editors. 7th International Symposium on Abstraction, Reformulation, and Approximation (SARA 2007). Volume 4612, Lecture Notes in Computer Science, Whistler, Canada: Springer; 2007. pp. 258–272. ISBN 978-3-540-73579-3.
43. Nouri A, Littman ML. Multi-resolution exploration in continuous spaces. In: Koller D, Schuurmans D, Bengio Y, *et al.*, editors. Advances in neural information processing systems 21. Cambridge (MA): MIT Press; 2009. pp. 1209–1216.
44. Kearns MJ, Mansour Y, Ng AY. Approximate planning in large POMDPs via reusable trajectories. In: Solla SA, Leen TK, Müller KR, editors. Advances in neural information processing systems 12. Cambridge (MA): MIT Press; 1999. pp. 1001–1007.
45. Szepesvári Cs. Efficient approximate planning in continuous space Markovian decision problems. *AI Commun* 2001;13:163–176.
46. Kocsis L, Szepesvári Cs. Bandit based Monte-Carlo planning. In: Fürnkranz J, Scheffer T, Spiliopoulou M, editors. Proceedings of the 17th European Conference on Machine Learning (ECML-2006). Berlin, Heidelberg: Springer; 2006. pp. 282–293.
47. Watkins CJCH. Learning from delayed rewards [PhD thesis]. Cambridge: King's College; 1989.
48. Tsitsiklis JN, Van Roy B. Feature-based methods for large scale dynamic programming. *Mach Learn* 1996;22:59–94.
49. Ormoneit D, Sen S. Kernel-based reinforcement learning. *Mach Learn* 2002;49: 161–178.
50. Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. *J Mach Learn Res* 2005;6:503–556.
51. Gordon GJ. Stable function approximation in dynamic programming. In: Russell SJ, Prieditis A, editors. Proceedings of the 17th European Conference on Machine Learning (ECML-2006). Springer; 2006. pp. 261–268. ISBN 1-55860-377-8.
52. Riedmiller M. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In: Gama J, Camacho R, Brazdil P, *et al.*, editors. Proceedings of the 16th European Conference on Machine Learning (ECML-05). Volume 3720, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer; 2005. pp. 317–328. ISBN 3-540-29243-8.
53. Van Roy B. Performance loss bounds for approximate value iteration with state aggregation. *Math Oper Res* 2006;31(2): 234–244.
54. Antos A, Munos R, Szepesvári Cs. Fitted Q-iteration in continuous action-space MDPs. In: Platt JC, Koller D, Singer Y, *et al.* editors. Advances in neural information processing systems 20. Cambridge (MA): MIT Press; 2008. pp. 9–16.
55. Munos R, Szepesvári Cs. Finite-time bounds for fitted value iteration. *J Mach Learn Res* 2008;9:815–857.
56. Rummery GA, Niranjan M. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166. Cambridge: Engineering Department, Cambridge University; 1994.
57. Konda VR, Tsitsiklis JN. On actor-critic algorithms. *SIAM J Control Optim* 2003;42(4):1143–1166. DOI: 10.1137/S0363 012901385691. URL <http://link.aip.org/link/?SJC/42/1143/1>.
58. Amari S. Natural gradient works efficiently in learning. *Neural Comput* 1998; 10(2):251–276.
59. Kakade S. A natural policy gradient. In: Dietterich TG, Becker S, Ghahramani Z, editors. Advances in neural information processing systems 14. Cambridge (MA): MIT Press; 2001. pp. 1531–1538.
60. Bagnell JA, Schneider JG. Covariant policy search. In: Gottlob G, Walsh T, editors. Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03). San Francisco (CA): Morgan Kaufmann; 2003. pp. 1019–1024.
61. Peters J, Vijayakumar S, Schaal S. Reinforcement learning for humanoid robotics. Humanoids2003, Third IEEE-RAS International Conference on Humanoid Robots. Karlsruhe, Germany: 2003. pp. 225–230.
62. Peters J, Schaal S. Natural actor-critic. *Neurocomputing* 2008;71(7–9):1180–1190.
63. Chow CS, Tsitsiklis JN. The complexity of dynamic programming. *J Complex* 1989;5:466–488.
64. Rust J. Using randomization to break the curse of dimensionality. *Econometrica* 1996;65:487–516.

65. de Farias DP, Van Roy B. The linear programming approach to approximate dynamic programming. *Oper Res* 2003;51(6): 850–865.
66. de Farias DP, Van Roy B. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math Oper Res* 2004;29(3):462–478.
67. de Farias DP, Van Roy B. A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Math Oper Res* 2006; 31(3):597–620.
68. Wang T, Lizotte DJ, Bowling MH, *et al.* Stable dual dynamic programming. In: Platt JC, Koller D, Singer Y, *et al.* editors. *Advances in Neural Information Processing Systems 20*. Cambridge (MA): MIT Press; 2008.
69. Ng AY, Jordan M. PEGASUS: a policy search method for large MDPs and POMDPs. In: Boutillier C, Goldszmidt M, editors. *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI'00)*. San Francisco (CA): Morgan Kaufmann; 2000. pp. 406–415. ISBN 1-55860-709-9.
70. Even-Dar E, Kakade SM, Mansour Y. Experts in a Markov decision process. In: Saul LK, Weiss Y, Bottou L, editors. *Advances in neural information processing systems 17*. Cambridge (MA): MIT Press; 2005. pp. 401–408.
71. Yu JY, Mannor S, Shimkin N. Markov decision processes with arbitrary reward processes. *Math Oper Res* 2009;34(3):737–757.
72. Neu G, György A, Szepesvári Cs. The online loop-free stochastic shortest-path problem. *COLT-10*; 2010 Jun 27–29; Haifa, Israel. 2010.
73. Hutter M. *Universal artificial intelligence: sequential decisions based on algorithmic probability*. Berlin: Springer; 2004. 300. pp. Available at <http://www.hutter1.net/ai/uaibook.htm>; <http://www.idsia.ch/~marcus/ai/uaibook.htm>.
74. Littman ML, Sutton RS, Singh SP. Predictive representations of state. In: Dietterich TG, Becker S, Ghahramani Z, editors. *Advances in neural information processing systems 14*. Cambridge (MA): MIT Press; 2001. pp. 1555–1561.
75. Toussaint M, Charlin L, Poupart P. Hierarchical POMDP controller optimization by likelihood maximization. In: McAllester DA, Myllymäki P, editors. *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*. AUAI Press; 2008. pp. 562–570. ISBN 0-9749039-4-9.
76. Ross S, Pineau J, Paquet S, *et al.* Online planning algorithms for POMDPs. *J Artif Intell Res* 2008;32:663–704.
77. Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Cohen WW, Hirsh H, editors. *Proceedings of the 11th International Conference on Machine Learning (ICML 1994)*. San Francisco (CA): Morgan Kaufmann; 1994. pp. 157–163. ISBN 1-55860-335-2.
78. Heger M. Consideration of risk in reinforcement learning. In: Cohen WW, Hirsh H, editors. *Proceedings of the 11th International Conference on Machine Learning (ICML 1994)*. San Francisco (CA): Morgan Kaufmann; 1994. pp. 105–111. ISBN 1-55860-335-2.
79. Szepesvári Cs, Littman ML. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural Comput* 1999;11:2017–2059.
80. Borkar VS, Meyn SP. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Math Oper Res* 2002;27(1):192–209, URL <http://morjournal.informs.org/cgi/content/abstract/27/1/192>.
81. Dietterich T. The MAXQ method for hierarchical reinforcement learning. In: Shavlik JW, editor. *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*. San Francisco (CA): Morgan Kauffmann; 1998. pp. 118–126. ISBN 1-55860-556-8.
82. Sutton RS, Precup D, Singh SP. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif Intell* 1999;112:181–211.
83. Tesauro GJ. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput* 1994; 6(2):215–219.
84. Silver D, Sutton RS, Müller M. Reinforcement learning of local shape in the game of Go. In: Veloso MM, editor. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*; San Francisco, CA: Morgan Kauffman; 2007. pp. 1053–1058.
85. Boyan JA, Littman ML. Packet routing in dynamically changing networks: a reinforcement learning approach. In: Cowan JD, Tesauro G, Alspector J, editors. *NIPS-6*:

- Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference. San Francisco, CA: Morgan Kaufman; 1994. pp. 671–678.
86. Singh SP, Bertsekas DP. Reinforcement learning for dynamic channel allocation in cellular telephone systems. In: Mozer MC, Jordan MI, Petsche T, editors. NIPS-9: Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference. Cambridge (MA): MIT Press; 1997. pp. 974–980.
 87. Abe N, Verma NK, Apté C, *et al.* Cross channel optimized marketing by reinforcement learning. In: Kim W, Kohavi R, Gehrke J, *et al.* editors. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York (NY): ACM; 2004. pp. 767–772. ISBN 1-58113-888-1.
 88. Gosavi A. Reinforcement learning for long-run average cost. *Eur J Oper Res* 2004;155(3):654–674. DOI: 10.1016/S0377-2217(02)00874-3.
 89. Zhang W, Dietterich TG. A reinforcement learning approach to job-shop scheduling. In: Perrault CR, Mellish CS, editors. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 95). San Francisco (CA): Morgan Kaufmann; 1995. pp. 1114–1120.
 90. Crites RH, Barto AG. Improving elevator performance using reinforcement learning. In: Touretzky D, Mozer MC, Hasselmo ME, editors. NIPS-8: Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference. Cambridge (MA): MIT Press; 1996. pp. 1017–1023.
 91. Rusmevichientong P, Salisburry JA, Truss LT, *et al.* Opportunities and challenges in using online preference data for vehicle pricing: a case study at General Motors. *J Revenue Pricing Manage* 2006;5(1):45–61.
 92. Proper S, Tadepalli P. Scaling model-based average-reward reinforcement learning for product delivery. In: Fürnkranz J, Scheffer T, Spiliopoulou M, editors. Proceedings of the 17th European Conference on Machine Learning (ECML-2006). Springer; 2006. pp. 735–742.
 93. Chang HS, Fu MC, Hu J, *et al.* An asymptotically efficient simulation-based algorithm for finite horizon stochastic dynamic programming. *IEEE Trans Autom Control* 2007;52(1):89–94.
 94. Simão HP, Day J, George AP, *et al.* An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transport Sci* 2009;43(2):178–197.
 95. Kohl N, Stone P. Policy gradient reinforcement learning for fast quadrupedal locomotion. Proceedings of the 2004 IEEE International Conference on Robotics and Automation; Sendai, Japan: IEEE; 2004. pp. 2619–2624.
 96. Abbeel P, Coates A, Quigley M, *et al.* An application of reinforcement learning to aerobatic helicopter flight. In: Schölkopf B, Platt JC, Hoffman T, editors. Advances in Neural Information Processing Systems 19. Cambridge (MA): MIT Press; 2007. pp. 1–8. ISBN 0-262-19568-2.
 97. Tsitsiklis JN, Van Roy B. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives. *IEEE Trans Autom Control* 1999;44:1840–1851.
 98. Tsitsiklis JN, Van Roy B. Regression methods for pricing complex American-style options. *IEEE Trans Neural Netw* 2001;12:694–703.
 99. Yu H, Bertsekas DP. Q-learning algorithms for optimal stopping based on least squares. Proceedings of the European Control Conference; Kos, Greece; 2007.
 100. Li Y, Szepesvári Cs, Schuurmans D. Learning exercise policies for american options. *J Mach Learn Res* 2009;5:352–359.
 101. Tanner B, White A. RL-Glue: language-independent software for reinforcement-learning experiments. *J Mach Learn Res* 2009;10:2133–2136.

RELATIONSHIP AMONG BENDERS, DANTZIG–WOLFE, AND LAGRANGIAN OPTIMIZATION

CHURLZU LIM

University of North Carolina at
Charlotte, Charlotte, North
Carolina

When solving a linear programming problem, Benders decomposition, Dantzig–Wolfe decomposition, and Lagrangian optimization can be interpreted as equivalent techniques applied to different representations of the problem such as primal linear, and Lagrangian dual formulations. In particular, applying Dantzig–Wolfe decomposition to a linear program is equivalent to employing Benders decomposition to solve its dual linear program, and implementing a cutting plane method to solve its Lagrangian dual problem (Fig. 1). To illustrate their equivalence, consider the following linear program:

$$\mathbf{P}: \quad \text{Minimize} \quad \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{y} \quad (1a)$$

$$\text{subject to} \quad \mathbf{Ax} + \mathbf{Dy} = \mathbf{b}, \quad (1b)$$

$$\mathbf{Fx} = \mathbf{g}, \quad (1c)$$

$$\mathbf{x} \geq \mathbf{0}, \quad \mathbf{y} \geq \mathbf{0}, \quad (1d)$$

where $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{d}, \mathbf{y} \in \mathbb{R}^l$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{D} \in \mathbb{R}^{m \times l}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{F} \in \mathbb{R}^{k \times n}$, $\mathbf{g} \in \mathbb{R}^k$, and $\mathbf{0}$ s are vectors of zeros having appropriate dimensions. All vectors are defined as column vectors, and T denotes a transpose operator. Suppose that constraints in Equation (1b) are *complicating* (or *coupling*) constraints where both $\mathbf{x}$ and $\mathbf{y}$ appear in the constraints, while those in Equation (1c) are relatively simple. Using this specially structured linear program, we first present the equivalence between Dantzig–Wolfe decomposition and Benders decomposition; this is followed by demonstration of the equivalence between Benders decomposition and Lagrangian optimization. We also verify the equivalence when they are applied to block diagonal structures.

EQUIVALENCE OF DANTZIG–WOLFE DECOMPOSITION AND BENDERS DECOMPOSITION

Assume that the optimal objective value of $\mathbf{P}$ is finite. Furthermore, in order to avoid cumbersome presentation, let us assume that the polyhedral set $\mathbf{X} \equiv \{\mathbf{x} : \mathbf{Fx} = \mathbf{g}, \mathbf{x} \geq \mathbf{0}\}$ is nonempty and bounded. (We will briefly discuss the case of unbounded polyhedral sets later.) Then, each $\mathbf{x} \in \mathbf{X}$ can be written as a convex combination of a finite number of extreme points. Letting $E_{\mathbf{X}}$ denote the index set of extreme points in $\mathbf{X}$, we have

$$\mathbf{X} = \left\{ \sum_{j \in E_{\mathbf{X}}} \lambda_j \mathbf{x}_j : \sum_{j \in E_{\mathbf{X}}} \lambda_j = 1, \lambda_j \geq 0 \forall j \in E_{\mathbf{X}} \right\}, \quad (2)$$

Replacing $\mathbf{X}$ by Equation (2), we can rewrite $\mathbf{P}$ as

DW-MP:

$$\text{Minimize} \quad \sum_{j \in E_{\mathbf{X}}} (\mathbf{c}^T \mathbf{x}_j) \lambda_j + \mathbf{d}^T \mathbf{y} \quad (3a)$$

$$\text{subject to} \quad \sum_{j \in E_{\mathbf{X}}} (\mathbf{Ax}_j) \lambda_j + \mathbf{Dy} = \mathbf{b}, \quad (3b)$$

$$\sum_{j \in E_{\mathbf{X}}} \lambda_j = 1, \quad (3c)$$

$$\lambda_j \geq 0 \quad \text{for } j \in E_{\mathbf{X}}, \quad \mathbf{y} \geq \mathbf{0}. \quad (3d)$$

Note that **DW-MP** is the *Dantzig–Wolfe master program* of Dantzig–Wolfe decomposition (see **Dantzig–Wolfe Decomposition**). The decomposition algorithm commences with a master program having a subset of extreme points, which is called a *restricted master program*. $\mathbf{x}_j$'s and λ_j 's are iteratively added to this restricted master program when its solution is not optimal to **DW-MP**. Optimality is checked by solving the following subproblem:

DW-SP:

$$\text{Minimize} \quad \mathbf{c}^T \mathbf{x} - \bar{\mathbf{u}}^T \mathbf{Ax} - \bar{z} \quad (4a)$$

$$\text{subject to} \quad \mathbf{x} \in \mathbf{X}, \quad (4b)$$

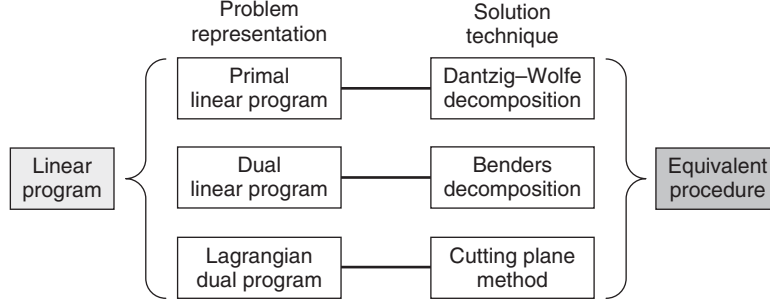


Figure 1. Schematic relationship of three decomposition methods.

where $\bar{\mathbf{u}}$ and $\bar{z}$ are the dual solutions of the current restricted master program. Note that the objective function of **DW-SP** (the Dantzig–Wolfe subproblem) represents the *reduced cost* of **DW-MP** (see *The Simplex Method and Its Complexity* and *Dual Simplex*). If the optimal objective value of **DW-SP** is nonnegative, then no extreme point in $\mathbf{X}$ yields negative reduced costs for **DW-MP** at the current basic feasible solution (see *The Simplex Method and Its Complexity*), and hence, the solution to the current restricted master program is also optimal to **DW-MP**. Otherwise, an entering variable λ_{j+1} that corresponds to the solution of **DW-SP**, $\mathbf{x}_{j+1}$, is added to the restricted master program to generate a new column (see Ref. 1; **Dantzig–Wolfe Decomposition** and **Column Generation**).

Since $\mathbf{P}$ is assumed to have a finite optimal objective value, solving its dual linear program will yield the optimal solution to $\mathbf{P}$ (see *LP Duality and KKT Conditions for LP*). Let $\mathbf{u}$ and $\mathbf{v}$ denote dual variable vectors associated with the constraints in Equations (1b) and (1c), respectively. Then, the dual linear program of $\mathbf{P}$ is as follows:

$$\begin{aligned} \mathbf{D}: \quad & \text{Maximize } \mathbf{b}^T \mathbf{u} + \mathbf{g}^T \mathbf{v} \\ & \text{subject to } \mathbf{A}^T \mathbf{u} + \mathbf{F}^T \mathbf{v} \leq \mathbf{c}, \\ & \mathbf{D}^T \mathbf{u} \leq \mathbf{d}. \end{aligned}$$

Let us first derive the master program of Benders decomposition for problem $\mathbf{D}$. Delaying the decision on $\mathbf{v}$, the dual linear

program can be written as

$$\text{Maximize}_{\mathbf{u} \in \mathbf{U}} \mathbf{b}^T \mathbf{u} + \left[\max_{\mathbf{v}} \mathbf{g}^T \mathbf{v} \right. \\ \left. \text{s.t. } \mathbf{F}^T \mathbf{v} \leq \mathbf{c} - \mathbf{A}^T \mathbf{u} \right], \quad (5)$$

where $\mathbf{U} = \{\mathbf{u} : \mathbf{D}^T \mathbf{u} \leq \mathbf{d}\}$.

Given $\bar{\mathbf{u}} \in \mathbf{U}$, let $\mathbf{x}$ be the dual variable vector of the inner maximization problem. Then, the dual linear program of the inner maximization problem becomes

$$\text{Minimize } (\mathbf{c} - \mathbf{A}^T \bar{\mathbf{u}})^T \mathbf{x} \quad (6a)$$

$$\text{subject to } \mathbf{x} \in \mathbf{X}. \quad (6b)$$

Since $\mathbf{X}$ is nonempty and bounded, the inner maximization problem in Equation (5) and the problem in Equation (6) will yield the same optimal objective value, which can be alternately obtained via

$$\min_{j \in E_{\mathbf{X}}} \{(\mathbf{c} - \mathbf{A}^T \bar{\mathbf{u}})^T \mathbf{x}_j\}, \quad (7)$$

where again $E_{\mathbf{X}}$ is the index set of extreme points of $\mathbf{X}$. Finally, substituting the minimum value in Equation (7) with a variable z , the problem in Equation (5) can be written as the following Benders master program (B-MP):

B-MP:

$$\text{Maximize } \mathbf{b}^T \mathbf{u} + z \quad (8a)$$

$$\text{subject to } \mathbf{D}^T \mathbf{u} \leq \mathbf{d}, \quad (8b)$$

$$z + (\mathbf{A} \mathbf{x}_j)^T \mathbf{u} \leq \mathbf{c}^T \mathbf{x}_j \\ \text{for } j \in E_{\mathbf{X}}. \quad (8c)$$

Note from Equation (8c) that z represents a lower bound of $\{(\mathbf{c} - \mathbf{A}^T \mathbf{u})^T \mathbf{x}_j : j \in E_{\mathbf{X}}\}$. Furthermore, maximizing the objective value of **B-MP** guarantees that the optimal value of z is indeed the greatest lower bound (i.e., $\min_{j \in E_{\mathbf{X}}} \{(\mathbf{c} - \mathbf{A}^T \mathbf{u})^T \mathbf{x}_j\}$).

Benders decomposition algorithm begins with a subset of constraints of **B-MP** (hence, it is called the *relaxed* master program), and then iteratively adds constraints along with new $\mathbf{x}_j$'s when the solution of the current relaxed master program is infeasible to **B-MP**. Given the current solution $(\bar{\mathbf{u}}, \bar{z})$, infeasibility is checked by solving the following subproblem:

$$\begin{aligned} \text{B-SP:} \quad & \text{Minimize} \quad \mathbf{c}^T \mathbf{x} - \bar{\mathbf{u}}^T \mathbf{A} \mathbf{x} \\ & \text{subject to} \quad \mathbf{x} \in \mathbf{X}. \end{aligned}$$

If its optimal objective value is not less than $\bar{z}$, it implies that the current solution satisfies all the constraints in **B-MP**. Otherwise, a new constraint is added to the relaxed master program (see Ref. 2, **Benders Decomposition**).

Let $\mathbf{y}$ and λ_j for $j \in E_{\mathbf{X}}$ denote the dual variable vector and dual variables associated with constraints given in Equations (8b) and (8c), respectively. Then, the dual linear program of **B-MP** becomes

$$\begin{aligned} & \text{Minimize} \quad \mathbf{d}^T \mathbf{y} + \sum_{j \in E_{\mathbf{X}}} (\mathbf{c}^T \mathbf{x}_j) \lambda_j \\ & \text{subject to} \quad \sum_{j \in E_{\mathbf{X}}} \lambda_j = 1, \\ & \quad \mathbf{D} \mathbf{y} + \sum_{j \in E_{\mathbf{X}}} (\mathbf{A} \mathbf{x}_j) \lambda_j = \mathbf{b}, \\ & \quad \lambda_j \geq 0 \forall j \in E_{\mathbf{X}}, \mathbf{y} \geq \mathbf{0}. \end{aligned}$$

This problem is identical to the master program of Dantzig–Wolfe decomposition in Equation (3). Furthermore, **DW-SP** is also equivalent to **B-SP** (Benders subproblem). Hence, we can conclude that the Dantzig–Wolfe decomposition method for solving the primal linear program **P** is equivalent to the Benders decomposition method for solving the dual linear program **D**.

EQUIVALENCE OF BENDERS DECOMPOSITION AND LAGRANGIAN OPTIMIZATION

Next, we present how the Lagrangian optimization for linear programs can be related to Benders decomposition. Let $\mathbf{u}$ denote the *Lagrangian multiplier* vector associated with the complicating constraints (Eq. 1b). Lifting the complicating constraints, we have the following *Lagrangian relaxation* (see Ref. 3, **Lagrangian Optimization for LP**)

$$\begin{aligned} \theta(\mathbf{u}) = \text{Minimize} \quad & \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{y} + \mathbf{u}^T \\ & \times (\mathbf{b} - \mathbf{A} \mathbf{x} - \mathbf{D} \mathbf{y}) \\ & \text{subject to} \quad \mathbf{F} \mathbf{x} = \mathbf{g} \\ & \quad \mathbf{x} \geq \mathbf{0}, \mathbf{y} \geq \mathbf{0}, \end{aligned}$$

where $\theta(\mathbf{u})$ is called the *Lagrangian function*. Then, the Lagrangian dual problem is to find the maximum value of the Lagrangian function:

LD:

$$\begin{aligned} \text{Maximize} \quad \theta(\mathbf{u}) = & \mathbf{u}^T \mathbf{b} + \min_{\mathbf{x} \in \mathbf{X}} (\mathbf{c}^T - \mathbf{u}^T \mathbf{A}) \mathbf{x} \\ & + \min_{\mathbf{y} \geq \mathbf{0}} (\mathbf{d}^T - \mathbf{u}^T \mathbf{D}) \mathbf{y}. \end{aligned}$$

Recall that **P** has a finite optimal objective value. Since **P** is a linear program having a feasible solution, **P** and **LD** yield the same optimal objective value.

The second minimization in $\theta(\mathbf{u})$ has a trivial solution; $\mathbf{y} = \mathbf{0}$ for $\mathbf{u}$ such that $\mathbf{u}^T \mathbf{D} \leq \mathbf{d}$, while it has an unbounded objective value if the inequality does not hold. Note that there exists $\mathbf{u}$ satisfying $\mathbf{u}^T \mathbf{D} \leq \mathbf{d}$ (otherwise, linear program **P**, and in turn **D**, do not have a bounded optimal objective value). Therefore, we can rewrite **LD** as follows:

LD:

$$\begin{aligned} \text{Maximize}_{\mathbf{u} \in \mathbf{U}} \quad \theta(\mathbf{u}) = & \mathbf{b}^T \mathbf{u} + \min_{\mathbf{x} \in \mathbf{X}} (\mathbf{c}^T - \mathbf{u}^T \mathbf{A}) \mathbf{x}. \end{aligned} \tag{9}$$

Note that this **LD** is identical to the problem in Equation (5) where the inner problem is replaced by Equation (6). Using Equation (7) and substituting the minimum value with a

single variable z , the master problem of the cutting plane algorithm for solving **LD** is

L-MP:

$$\begin{aligned} & \text{Maximize } \mathbf{b}^T \mathbf{u} + z, \\ & \mathbf{D}^T \mathbf{u} \leq \mathbf{d}, \\ & z + (\mathbf{Ax}_j)^T \mathbf{u} \leq \mathbf{c}^T \mathbf{x}_j \quad \forall j \in E_{\mathbf{X}}, \end{aligned}$$

which is identical to **B-MP**. Given a solution $(\bar{\mathbf{u}}, \bar{z})$ to the *relaxed* master program, the optimality is checked by solving the following Lagrangian subproblem:

$$\begin{aligned} \text{L-SP:} \quad & \text{Minimize } \mathbf{c}^T \mathbf{x} - \bar{\mathbf{u}}^T \mathbf{Ax} \\ & \text{subject to } \mathbf{x} \in \mathbf{X}, \end{aligned}$$

which is again identical to **B-SP**. Hence, when solving linear programs, the Lagrangian optimization via the cutting plane method is equivalent to Benders decomposition.

UNBOUNDED FEASIBLE REGION OF SUBPROBLEM

In the previous problem, we assumed that $\mathbf{X}$ is nonempty and bounded. When $\mathbf{X}$ is unbounded, Equation (2) can be modified as follows:

$$\mathbf{X} = \left\{ \sum_{j \in E_{\mathbf{X}}} \lambda_j \mathbf{x}_j + \sum_{j \in D_{\mathbf{X}}} \mu_j \mathbf{d}_j : \sum_{j \in E_{\mathbf{X}}} \lambda_j = 1, \right. \\ \left. \lambda_j \geq 0 \quad \forall j \in E_{\mathbf{X}}, \mu_j \geq 0 \quad \forall j \in D_{\mathbf{X}} \right\},$$

where $D_{\mathbf{X}}$ is the index set of extreme directions of $\mathbf{X}$, and $\mathbf{d}^j$, $\forall j \in D_{\mathbf{X}}$, are extreme directions. Then, the master program of Dantzig–Wolfe decomposition becomes

$$\begin{aligned} & \text{Minimize} \\ & \sum_{j \in E_{\mathbf{X}}} (\mathbf{c}^T \mathbf{x}_j) \lambda_j + \sum_{j \in D_{\mathbf{X}}} (\mathbf{c}^T \mathbf{d}_j) \mu_j + \mathbf{d}^T \mathbf{y} \quad (10a) \\ & \text{subject to} \end{aligned}$$

$$\sum_{j \in E_{\mathbf{X}}} (\mathbf{Ax}_j) \lambda_j + \sum_{j \in D_{\mathbf{X}}} (\mathbf{Ad}_j) \mu_j + \mathbf{Dy} = \mathbf{b}, \quad (10b)$$

$$\sum_{j \in E_{\mathbf{X}}} \lambda_j = 1, \quad (10c)$$

$$\lambda_j \geq 0 \quad \forall j \in E_{\mathbf{X}}, \mu_j \geq 0 \quad \forall j \in D_{\mathbf{X}}, \mathbf{y} \geq \mathbf{0}. \quad (10d)$$

Note that the subproblem remains the same as in Equation (4). When the subproblem yields an unbounded solution, an extreme direction is generated and added to the restricted master program.

From the Ref. 2 (see also **Benders Decomposition**), the Benders master program is

$$\text{Maximize } \mathbf{b}^T \mathbf{u} + z \quad (11a)$$

$$\text{subject to } \mathbf{D}^T \mathbf{u} \leq \mathbf{d}, \quad (11b)$$

$$z + (\mathbf{Ax}_j)^T \mathbf{u} \leq \mathbf{c}^T \mathbf{x}_j, \quad \forall j \in E_{\mathbf{X}}, \quad (11c)$$

$$(\mathbf{Ad}_j)^T \mathbf{u} \leq \mathbf{c}^T \mathbf{d}_j, \quad \forall j \in D_{\mathbf{X}}. \quad (11d)$$

One can verify that this master program is indeed the dual linear program of Equation (10).

Finally, note that the Lagrangian dual problem is same as in Equation (9). However, since $\mathbf{X}$ is now unbounded, the inner minimization problem may have an extreme direction along which the objective function value decreases. Such $\mathbf{u}$'s yielding this unboundedness are not of interest in the Lagrangian dual problem that maximizes $\theta(\mathbf{u})$, and hence, we can further restrict the selection of $\mathbf{u}$ such that

$$(\mathbf{c}^T - \mathbf{u}^T \mathbf{A}) \mathbf{d}_j \geq 0, \quad \forall j \in D_{\mathbf{X}}.$$

Equivalently, this can be written as

$$(\mathbf{Ad}_j)^T \mathbf{u} \leq \mathbf{c}^T \mathbf{d}_j, \quad \forall j \in D_{\mathbf{X}}.$$

Given that $\mathbf{u}$ satisfies these constraints, the minimum of the inner problem can be identified as Equation (7) and the master program of cutting plane method becomes identical to the master program of Benders decomposition in Equation (11).

$$\text{Minimize} \quad (\mathbf{c}^1)^T \mathbf{x}^1 + (\mathbf{c}^2)^T \mathbf{x}^2 + \dots + (\mathbf{c}^K)^T \mathbf{x}^K \quad (12a)$$

$$\text{subject to} \quad \mathbf{A}^1 \mathbf{x}^1 + \mathbf{A}^2 \mathbf{x}^2 + \dots + \mathbf{A}^K \mathbf{x}^K = \mathbf{b} \quad (12b)$$

$$\mathbf{D}^1 \mathbf{x}^1 = \mathbf{g}^1 \quad (12c)$$

$$\mathbf{D}^2 \mathbf{x}^2 = \mathbf{g}^2 \quad (12d)$$

$$\ddots = \vdots \quad (12e)$$

$$\mathbf{D}^K \mathbf{x}^K = \mathbf{g}^K \quad (12f)$$

$$\mathbf{x}^1 \quad \mathbf{x}^2 \quad \dots \quad \mathbf{x}^K \geq \mathbf{0}, \quad (12g)$$

where the constraint set can be partitioned into $K + 1$ subsets such that one is the set of complicating constraints (Eq. 12b) and others contain exclusive variable vectors (Eqs 12c–f). (Note that without the complicating constraints, the problem would have been separable to K small linear programs.) Assuming boundedness, each constraint set can be written as convex combinations of extreme points of respective polytopes. (A similar argument can be made when constraints sets are unbounded as before.) Then, the master program of Dantzig–Wolfe decomposition becomes

$$\text{Minimize} \quad \sum_{i=1}^K \sum_{j \in E_{\mathbf{X}^i}} (\mathbf{c}^T \mathbf{x}_j^i) \lambda_j^i \quad (13a)$$

$$\text{subject to} \quad \sum_{i=1}^K \sum_{j \in E_{\mathbf{X}^i}} (\mathbf{A}^i \mathbf{x}_j^i) \lambda_j^i = \mathbf{b}, \quad (13b)$$

$$\sum_{j \in E_{\mathbf{X}^i}} \lambda_j^i = 1, \quad i = 1, \dots, K \quad (13c)$$

$$\lambda_j^i \geq 0, \quad \forall i = 1, \dots, K, \forall j \in E_{\mathbf{X}^i}, \quad (13d)$$

where $E_{\mathbf{X}^i}$ is the index set of extreme points of $\mathbf{X}^i = \{\mathbf{x}^i : \mathbf{D}^i \mathbf{x} = \mathbf{g}^i, \mathbf{x}^i \geq \mathbf{0}\}$. Accordingly, the following K subproblems, for $i = 1, \dots, K$, need to be solved to check optimality of the current solution.

$$\begin{aligned} &\text{Minimize} \quad (\mathbf{c}^i)^T \mathbf{x}^i - (\bar{\mathbf{u}})^T \mathbf{A}^i \mathbf{x}^i - \bar{z}^i \\ &\text{subject to} \quad \mathbf{x}^i \in \mathbf{X}^i, \end{aligned}$$

where $\bar{\mathbf{u}}$ and $\bar{z}^i$ are dual solutions corresponding to Equations (13b) and (13c), respectively. On the other hand, the Benders master program of Equation (12) is

$$\text{Minimize} \quad \mathbf{b}^T \mathbf{u} + \sum_{i=1}^K z^i \quad (14a)$$

$$\begin{aligned} &\text{subject to} \quad z^i + (\mathbf{A}^i \mathbf{x}_j^i)^T \mathbf{u} \leq (\mathbf{c}^i)^T \mathbf{x}_j^i, \\ &\quad \forall i = 1, \dots, K, \forall j \in E_{\mathbf{X}^i}. \end{aligned} \quad (14b)$$

Again, one can see that this master program is the dual linear program of Equation (13).

Dualizing complicating constraints (Eq. 12b), the Lagrangian dual problem of Equation (12) can be written as follows:

$$\begin{aligned} &\text{Maximize} \\ &\theta(\mathbf{u}) = \mathbf{b}^T \mathbf{u} + \sum_{i=1}^K \min_{\mathbf{x}^i \in \mathbf{X}^i} [(\mathbf{c}^T)^i - \mathbf{u}^T \mathbf{A}^i] \mathbf{x}^i. \end{aligned}$$

Then, letting $z^i = \min_{\mathbf{x}^i \in \mathbf{X}^i} \{[(\mathbf{c}^T)^i - \mathbf{u}^T \mathbf{A}^i] \mathbf{x}^i\}$,

the master program of the cutting plane method is identical to Equation (14).

For details about various decomposition methods applied to linear and nonlinear programs and stochastic programming, readers are referred to Refs 3–5, and sections titled “Large-Scale Optimization,” “Scenario Generation and Sampling,” and “Stochastic Integer Programming” in this encyclopedia.

REFERENCES

1. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960; 8(1):101–111.
2. Benders JF. Partitioning procedures for solving mixed-variables programming problems. *Numer Math* 1962;4(1):238–252.
3. Bazaraa MS, Sherali HD, Shetty CM. *Nonlinear programming: theory and algorithms*. 3rd ed. New York: Wiley-Interscience; 2006.
4. Bazaraa MS, Jarvis JJ, Sherali HD. *Linear programming and network flows*. 3rd ed. New York: Wiley-Interscience; 2005.
5. Bertsimas D, Tsitsiklis JN. *Introduction to linear optimization*. Belmont (MA): Athena Scientific; 1997.

RELIABILITY INDICES

FLAVIO FRATTINI
ANTONIO BOVENZI
Dipartimento di Informatica e
Sistemistica, Università degli
Studi di Napoli Federico II,
Naples, Italy

JAVIER ALONSO
KISHOR TRIVEDI
Department of Electrical and
Computer Engineering, Duke
University, Durham, North
Carolina

INTRODUCTION

Computer and other technical systems are expected to be reliable during operation. However, their growing complexity, which is related to the intricate interdependencies among many heterogeneous components, makes the development of fault-free systems an unaffordable task in terms of time and cost. On the other hand, quality standards impose strict requirements on the reliability attributes and measures [1,2]. Fault-tolerance techniques are used to allow systems to continue their operations even in the presence of component failures, although in a possibly degraded mode. Hence, the quantification of the system reliability in combination with system performance is necessary.

Reliability can be evaluated using several approaches, generally classified into two categories: measurements-based and model-based [3]. Generally, the former is an attractive way to estimate the reliability of a system as it is based on real operational data. But it is costly and time consuming as it is based on real operational data collected from the system or a prototype. On the contrary, model-based approaches offer an abstraction of the real system, avoiding unnecessary details and system implementation. These models can be used to compute the reliability with

an acceptable approximation within affordable cost and time.

State-space models are often used because of their capacity of handling different failure/repair behaviors, such as imperfect coverage, correlated failures, and repair dependencies [3]. Continuous-time Markov chains (CTMCs) are state-space models commonly used for performance and reliability analysis. Homogeneous continuous-time Markov chains (HCTMCs) assume that the rates associated with events are time-independent and holding times exponentially distributed. Although HCTMCs are able to cover many cases, time-dependent rates and nonexponential distributions are also present in many real world situations [4,5]. In such cases, nonhomogeneous continuous-time Markov chains (NHCTMCs), semi-Markov processes (SMPs), Markov regenerative processes (MRGPs), or phase-type approximation can be used.

In order to show how to compute the reliability of a system based on a probabilistic model, consider the example of a *two-processor-parallel-redundant system with imperfect coverage*, adapted from Ref. 6. The system has two processors with the same time-independent failure rate and single repair facility with a certain repair rate. *Imperfect coverage* means that not all individual processor failures are recovered from. If a *covered failure* occurs in one of the two processors, the other one continues working and the system is up, although in a degraded mode. If a *not-covered failure* occurs, the system fails. Such a situation can be easily modeled by means of a HCTMC (see also the articles *Definition and Examples of DTMCs* and *Definition and Examples of CTMCs*). However, if we consider a time-dependent failure rate for the processors, the model is no longer homogeneous and the computation of the reliability becomes harder.

In the following, the reliability and related indices are formally defined; subsequently, we show how to compute them for the

mentioned example. In particular, three different techniques are presented: *piecewise constant approximation (PCA)*, *phase-type expansion*, and *discrete event simulation*. Finally, results obtained with these techniques are compared.

BASIC DEFINITIONS

Let X be a random variable representing the *time to failure* (TTF) (or *lifetime*) of a system. Such a random variable can be characterized by the cumulative distribution function (CDF), the probability density function (pdf), or the hazard rate function $h(t)$.

The CDF of the (non-negative) random variable X is simply defined as:

$$F_X(t) = P(X \leq t), \quad 0 < t < \infty \quad (1)$$

The pdf of X is defined by

$$f_X(t) = \frac{dF_X(t)}{dt} \quad (2)$$

The *hazard rate* (or *instantaneous failure rate*) is defined by

$$h(t) = \frac{f(t)}{1 - F_X(t)} \quad (3)$$

The *reliability* is defined by the Recommendation E.800 of the International Telecommunications Union (ITU-T) as the “ability of an item to perform a required function under given conditions for a given time interval.” Hence, for a time interval $(t_0, t_0 + t]$, the reliability $R(t|t_0)$ defines the probability that a system survives in this interval given that it was properly working at time t_0 .

If $t_0 = 0$, $R(t|0)$ defines the probability that a system is up until time t ,

$$R(t|0) = R(t) = P(X > t) = 1 - F_X(t) \quad (4)$$

The *conditional reliability* defines the probability that a system survives in the interval $(t_0, t_0 + t]$, given that the system has survived until time t_0 :

$$R_{t_0}(t) = \frac{R(t_0 + t)}{R(t_0)} \quad (5)$$

The expected life of a system is defined by the *mean time to failure* (MTTF) as:

$$E[X] = \int_0^\infty tf(t)dt = \int_0^\infty R(t)dt \quad (6)$$

The reliability represents the probability that a system is failure-free during a time interval. It is also useful to evaluate the probability of a system being failure-free at a precise instant of time. The *availability* is defined by ITU-T Recommendation E.800 as the “ability of an item to be in a state to perform a required function at a given instant of time or at any instant of time within a given time interval, assuming that the external resources, if required, are provided.”

Introducing the indicator random variable $I(t)$, which is equal to 1 when the system is up and 0 otherwise, we have the following definitions.

The *instantaneous availability* (or *point availability*) is the probability that a system is up at time t ,

$$A(t) = P(I(t) = 1) \quad (7)$$

Note that the instantaneous availability

$A(t)$ is always greater than or equal to the reliability

$R(t)$ and if the system has no repair or replacement strategies

$$A(t) = R(t).$$

The *steady-state availability* (or *limiting availability*) (A) is the value of $A(t)$ as t approaches infinity,

$$A = \lim_{t \rightarrow \infty} A(t) \quad (8)$$

Under very general conditions, the steady-state availability can be computed as:

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}} \quad (9)$$

where the *mean time to repair* (MTTR) includes both the time to detect a failure and the time to repair it.

The *mean time between failures* (MTBF), instead, is the sum of MTTF and MTTR:

$$\text{MTBF} = \text{MTTF} + \text{MTTR} \quad (10)$$

The *interval availability* (or *average availability*) is the expected fraction of time the system is up in a given interval of time $(0, t]$. It is defined as

$$A_I(t) = \frac{1}{t} \int_0^t A(x) dx \quad (11)$$

It is easy to demonstrate that, when both limits exist:

$$A = \lim_{t \rightarrow \infty} A_I(t) = \lim_{t \rightarrow \infty} A(t) \quad (12)$$

For more details about the availability indices, modeling, and evaluation, see the article *Point and Interval Availability*.

Other indices have been defined for combining the performance and reliability/availability measure of a system, which is called *performability* [7]. As a matter of fact, fault-tolerant systems are able to continue providing service even in the presence of component failures, although in a degraded mode. Assuming that at the initial state the system is operating at its maximum performance, when a failure occurs, system performance may degrade. This kind of system offers several levels of performance.

Let S denote the set of all possible configurations in which the system can perform its activity, and let $\{X(t), t \geq 0\}$ on S define a continuous-time stochastic process describing the structure of the system at time t . Let $\pi_i(t)$ be the probability that the system is in state $i \in S$ at time t and π_i be the probability that the system is in state $i \in S$ as t approaches infinity. We can associate a *reward rate* to every state indicating the performance level offered by the system in that state. r_i represents the reward obtained per unit time spent in state $i \in S$.

The *reward rate at time t* is indicated as $Z(t) = r_{X(t)}$. It can be shown that [8] $L_i(t) = \int_0^t \pi_i(\tau) d\tau$ is the expected total amount of time spent by the system in state i during the interval $(0, t]$.

The *expected instantaneous reward rate* at time t is given by

$$E[Z(t)] = \sum_{i \in S} r_i \pi_i(t) \quad (13)$$

The *expected steady-state reward rate* of the system is

$$E[Z] = \lim_{t \rightarrow \infty} E[Z(t)] = \sum_{i \in S} r_i \pi_i \quad (14)$$

The *accumulated reward in $(0, t]$* represents the total amount of work performed by system during the interval $(0, t]$:

$$Y(t) = \int_0^t Z(t) dt \quad (15)$$

Then, the *expected accumulated reward in $(0, t]$* is the amount of work done by the system during the interval of time $(0, t]$:

$$\begin{aligned} E[Y(t)] &= E\left[\int_0^t Z(t) dt\right] = \int_0^t E[Z(t)] dt \\ &= \sum_{i \in S} r_i \int_0^t \pi_i(t) dt = \sum_{i \in S} r_i L_i(t) \end{aligned} \quad (16)$$

Figure 1, adapted from Ref. 7, presents an example of a Markov reward model (MRM) with three states. The reward rate is 2 for *State 1*, 1 for *State 2*, and 0 for *State 3*. $X(t)$ represents the state of the Markov chain at time t , and $Z(t)$ is the corresponding reward rate of the system. $Y(t)$ is the accumulated reward over the time interval $(0, t]$.

For a CTMC with one or more absorbing states, we can also compute the *expected accumulated reward till absorption*

$$\begin{aligned} E[Y(\infty)] &= \sum_{i \in S} r_i \int_0^\infty \pi_i(x) dx \\ &= \sum_{i \in S} r_i L_i(\infty) = \sum_{i \in S} r_i \tau_i \end{aligned} \quad (17)$$

where $\tau_i = L_i(\infty) = \lim_{t \rightarrow \infty} L_i(t)$ is the expected total time spent in state i before absorption [8].

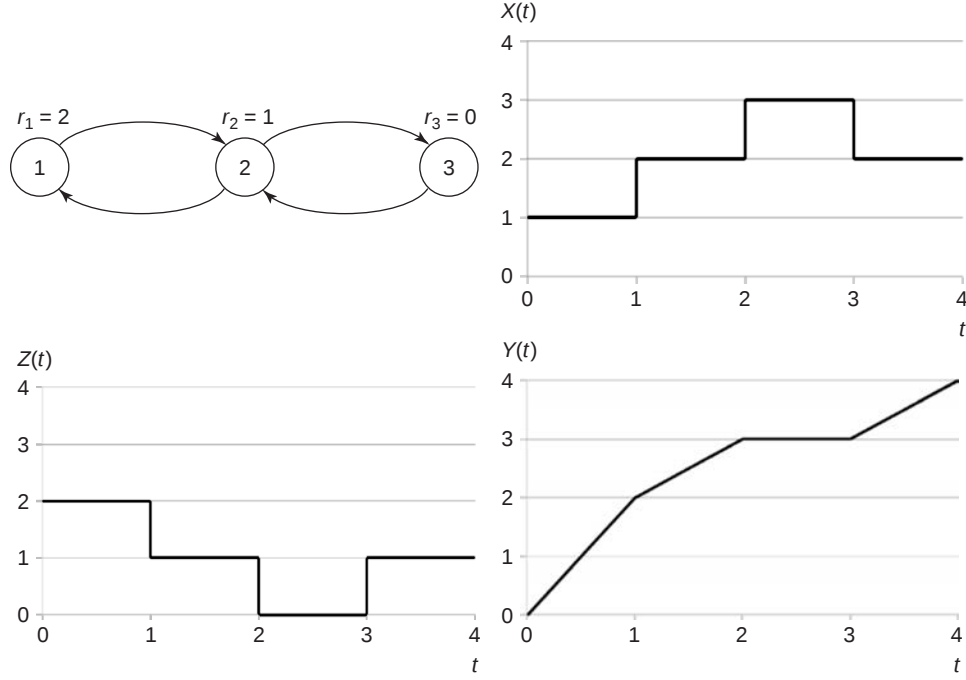


Figure 1. Markov reward model, evolution of the system over time (X), reward rate (Z), accumulated reward rate (Y).

The *distribution of the accumulated reward till absorption* $P(Y(\infty) \leq y)$ can also be computed [9,10].

A SIMPLE EXAMPLE

In this section, we present in detail the above-mentioned system to show how to compute the reliability and performability indices described in the previous section.

Consider the *two-processor-parallel-redundant system with imperfect coverage* introduced earlier. Let P_1 and P_2 be the two processors; both of them have the same time-dependent failure rate $\lambda(t)$, which is not dependent on the state or the task of the processor. The coverage factor, denoted by c , represents the proportion of individual processor failures from which the system can automatically be recovered. Once P_1 (or P_2) fails, if the failure is covered by the recovery strategy, the system can properly work with only one processor, although with degraded performance, otherwise it fails.

The processor repair facility is characterized by a constant repair rate μ .

The system can be modeled by a Markov chain with three states (Fig. 2):

- *State 2.* both processors are properly working; the failure rate of each processor is $\lambda(t)$. Hence the equivalent failure rate of this state is $2\lambda(t)$. If a failure covered by the recovery strategy occurs (rate $2\lambda(t)c$), the system goes to *State 1*, otherwise to *State 0* (rate $2\lambda(t)(1-c)$);
- *State 1.* one processor failed because of a covered failure; the system can continue working in degraded mode with one processor. In this case, two possible scenarios are possible: the failed processor is repaired (with constant repair rate μ) before the working one fails, then the system goes back to *State 2*; or the working processor fails (rate $\lambda(t)$), and the system fails (*State 0*);

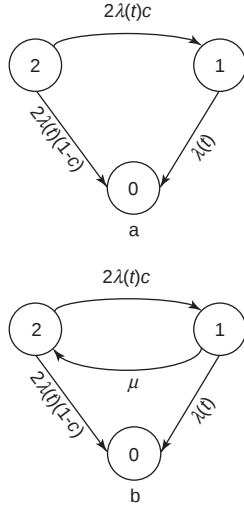


Figure 2. Two-processor-parallel-redundant system with imperfect coverage without (a) and with (b) repair.

- *State 0.* the system is down because either both processors failed or a non-recoverable failure occurred. This is an absorbing state (the system cannot be repaired, if this state is reached).

Note that in case there is no repair (Fig. 2a), the model is an NHCTMC. Since all rates are dependent on the same time-dependent function $\lambda(t)$, the time-dependent infinitesimal generator matrix can be factored into a time-independent matrix and a scalar function of time. Hence we can get the solution to the NHCTMC by solving a related HCTMC [8].

When repair from *State 1* back to *State 2* is introduced (Fig. 2b), we need to make two assumptions for the model to remain an NHCTMC:

1. Repair is minimal, that is, the repaired processor is in a state (age) equal to the one before its failure.
2. Repair time is negligible compared to time to failure.

The vector $\pi(t)$, whose i th element $\pi_i(t)$ is the probability that the system is in state i at time t , is equals to $[0 \ 0 \ 1]$ at $t = t_0$,

and the infinitesimal generator matrix is (see also the article **Definition and Examples of CTMCs**):

$$Q(t) = \begin{bmatrix} 0 & 0 & 0 \\ \lambda(t) & -\mu - \lambda(t) & \mu \\ 2\lambda(t)(1-c) & 2\lambda(t)c & -2\lambda(t) \end{bmatrix} \quad (18)$$

The failure rate $\lambda(t)$ is assumed to be the hazard rate of a two-parameter Weibull distribution:

$$\lambda(t) = \frac{\alpha}{\beta} \left(\frac{t}{\beta} \right)^{\alpha-1} \quad (19)$$

Table 1 shows the values of the parameters of the system. α and β are chosen to obtain a Weibull distribution with an increasing failure rate [11]; the rest of the parameters are typical values used to simplify computations.

We focus on the computation of the aforementioned *reliability* and *performability* indices for the described Markov model.

The transient behavior of a CTMC is defined by the Kolmogorov ordinary differential equations (ODE) system [12]:

$$\begin{aligned} \dot{\pi}(t) &= \pi(t)Q(t); \\ \sum_{i=1}^m \pi_i(t) &= 1 \end{aligned} \quad (20)$$

If $Q(t)$ is integrable, one solution exists and it is of the form $\pi(t) = \pi(t_0)\Phi(t, t_0) \cdot \Phi(t, t_0)$ can be evaluated using the Peano–Baker series [13] as:

$$\begin{aligned} \Phi(t, t_0) &= I + \int_{t_0}^t Q(\tau_1) d\tau_1 + \int_{t_0}^t Q(\tau_1) \int_{\tau_0}^{\tau_1} Q(\tau_2) d\tau_2 d\tau_1 \\ &\quad + \int_{t_0}^t Q(\tau_1) \int_{t_0}^{\tau_1} Q(\tau_2) \dots \int_{t_0}^{\tau_{h-1}} \\ &\quad Q(\tau_h) d\tau_h d\tau_{h-1} \dots d\tau_2 d\tau_1 + \dots \end{aligned} \quad (21)$$

However, the computation of $\Phi(t, t_0)$ for time-variant systems such as NHCTMCs is rather difficult. As a consequence, several methods have been proposed for the transient analysis of NHCTMCs as an alternative to

Table 1. Values of the parameters of the model.

Parameter	Value
α	2.1
β	1.02
μ	3.33×10^{-1} days
c	0.9
t_0	0

the ODE approach (see also the article *Transient Behavior of CTMCs*), such as PCA [14] and phase-type expansion (PH) [15].

The repair strategy of the system determines the method to use. Two different strategies have been considered [16]:

- *minimal repair* (or *as bad as before*);
- *maximal repair* (or *as good as new*), that is, the repaired processor behaves as a new one (with its age reset to zero).

In the first case, assuming that $\frac{1}{\mu} \ll \beta \Gamma(\frac{1}{\alpha} + 1)$ (MTTF of the Weibull distribution), that is, the recovery strategy is much faster than the inverse of the failure frequency, the model is an NHCTMC. Such a model requires only a *global clock* to describe all time-dependent transition rates as in every state, each component is as old as the system. To compute the reliability/performability of this model, PCA can be used.

In the case of maximal repair, instead, the system is non-Markovian as the transition rates from *State 2* to *State 1* and from *State 1* to *State 0* depend on how long the system has been in *State 2* or *State 1*, respectively. Furthermore, failure transition rate from *State 1* depends on the time spent by the nonfailed component in *States 2* and *1*. We note that for an SMP, all transition rates can depend on local time. In the current model with maximal repair, neither pure global clock nor pure local clock suffices. The model is neither an SMP nor an NHCTMC. However, using PH approximation, we can solve the problem.

For *performability* analysis, rewards are simply attached to the states of the Markov model by counting the number of active processors. In *State 2*, two processors are working, and hence, the reward rate is equal to

2; in *State 1*, where only one processor is working, the reward rate is 1. A reward rate equals to 0 is associated to the absorbing *State 0*. Hence, the reward rates are $r_2 = 2$, $r_1 = 1$, and $r_0 = 0$.

PIECEWISE CONSTANT APPROXIMATION

PCA is based on considering a time-variant function $f(t)$ as constant in certain intervals. Given the time interval $[0, t_0]$, it is divided into $n + 1$ shorter intervals of length δ : $t \in [0, t] \rightarrow t \in [i\delta, (i+1)\delta]$, $i = 0, 1, \dots, n$, where the function assumes the constant value $f(i\delta)$.

In the case of the failure rate $\lambda(t)$ of the considered Markov model:

$$\lambda(t) = \begin{cases} \lambda(\frac{\delta}{2}) & 0 \leq t < \delta \\ \lambda(\frac{\delta}{2} + \delta) & \delta \leq t < 2\delta \\ \vdots & \vdots \\ \lambda(\frac{\delta}{2} + i\delta) & i\delta \leq t < (i+1)\delta \\ \vdots & \vdots \end{cases}; i=0, 1, \dots, n \quad (22)$$

Figure 3 shows both the Weibull failure rate of the considered problem and its PCA with $\delta = 0.5$. With selected parameters of the Weibull function, the system has a monotonically increasing failure rate.

As a consequence of the failure rate approximation, also the infinitesimal generator matrix $Q(t)$ makes discrete changes:

$$Q(t) = \begin{cases} Q(\frac{\delta}{2}) & 0 \leq t < \delta \\ Q(\frac{\delta}{2} + \delta) & \delta \leq t < 2\delta \\ \vdots & \vdots \\ Q(\frac{\delta}{2} + i\delta) & i\delta \leq t < (i+1)\delta \\ \vdots & \vdots \end{cases}; i=0, 1, \dots, n \quad (23)$$

and the system is an HCTMC in each interval $[i\delta, (i+1)\delta]$, $i = 0, 1, \dots, n$. Hence, the transient behavior of the state probabilities can be computed as:

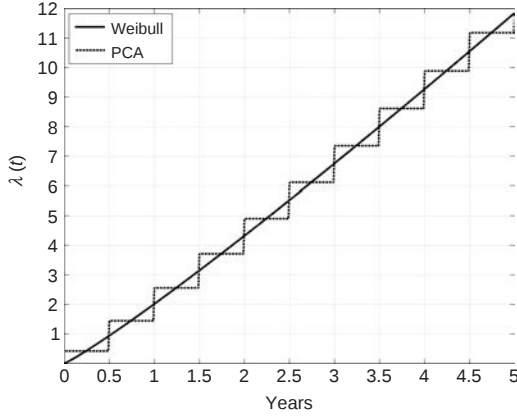


Figure 3. Comparison of the Weibull failure rate and its piecewise constant approximation.

$$\pi(t) = \begin{cases} \pi(0)e^{Q(\frac{\delta}{2})(t-0)} & 0 \leq t < \delta \\ \pi(0)e^{Q(\frac{\delta}{2}+\delta)(t-\delta)} & \delta \leq t < 2\delta \\ \vdots & \vdots \\ \pi(i\delta)e^{Q(\frac{\delta}{2}+i\delta)(t-i\delta)} & i\delta \leq t < (i+1)\delta \\ \vdots & \vdots \\ \vdots & \vdots \end{cases}; \quad (24)$$

$i = 0, 1, \dots, n$

where

$$\begin{aligned} \pi(\delta) &= \pi(0)e^{Q(\frac{\delta}{2})(\delta-0)} \\ \pi(2\delta) &= \pi(\delta)e^{Q(\frac{\delta}{2}+\delta)(2\delta-\delta)} \\ &\vdots \\ \pi(i\delta) &= \pi((i-1)\delta)e^{Q(\frac{\delta}{2}+(i-1)\delta)\delta} \\ &\vdots \end{aligned} \quad (25)$$

Reliability and performability indices for the *two-processor system* can be computed using SHARPE (Symbolic Hierarchical Automated Reliability and Performance Evaluator) [17], as shown in Fig. 4.

Since the model presents an absorbing state (*State 0*), that is, there is no repair strategy for the failure of the whole system, availability is not considered. Indeed, in this case, $A(t) = R(t)$.

Reliability

Since the CTMC has an absorbing state that represents the failed system (*State 0*), $\pi_0(t)$ is the probability that the system has failed at

or before time t ; the reliability can be simply computed as [8]:

$$R(t) = 1 - \pi_0(t) \quad (26)$$

Figure 5 shows the reliability of the considered system for different values of δ . We note that the two approximations are very similar.

Distribution Function

The CDF of the time to failure of the system is $F(t) = \pi_0(t)$ [8]. Figure 6 shows the $F(t)$ function of the considered system.

Probability Density Function

The *pdf* of the time to failure of the system is presented in Fig. 7. It is computed as $f(t) = \frac{d\pi_0(t)}{dt} = \lambda(t)\pi_1(t) + 2\lambda(t)(1-c)\pi_2(t)$ (see line 63 in Fig. 4 for the code).

Hazard Rate

The $h(t)$ is presented in Fig. 8. It is computed considering the state probabilities, as described in Ref. 18:

$$h(t) = \frac{\lambda(t)\pi_1(t) + 2\lambda(t)(1-c)\pi_2(t)}{\pi_1(t) + \pi_2(t)} \quad (27)$$

(see line 64 in Fig. 4). It has an increasing trend, but it is much less than the failure rate of each processor. Hence, the redundancy of processors and the recovery strategy are able to increase the reliability of the whole system.

```

1: format 15
2: epsilon basic 10e-25
3:
4: func lfr(alfa, beta, t)
(alfa/beta)*((t/beta)^(alfa-1))
5:
6: bind
7:   mu 120
8:   c 0.9
9:   alfa 2.1
10:  beta 1.02
11: end
12:
13: bind d 0.01
14: bind tmax 5.0
15:
16: *****
17:
18: markov nhctmc(t)
19: 1 2 mu
20: 1 0 lfr(alfa, beta, t)
21: 2 1 2*c*lfr(alfa, beta, t)
22: 2 0 2*(1-c)*lfr(alfa, beta, t)
23:
24: reward
25: 0 rew_nhctmc_0
26: 1 rew_nhctmc_1
27: 2 rew_nhctmc_2
28: end
29:
30: *probabilities
31: 0 init_nhctmc_0
32: 1 init_nhctmc_1
33: 2 init_nhctmc_2
34: end
35:
36: bind
37:   rew_nhctmc_0 0
38:   rew_nhctmc_1 1
39:   rew_nhctmc_2 2
40: end
41:
42: bind
43:   init_nhctmc_0 0
44:   init_nhctmc_1 0
45:   init_nhctmc_2 1
46: end
47:
48: *****
49:
50: bind crc 0
51: bind mttfa 0
52:
53: func sp0(t) tvalue(d; nhctmc, 0;
t)
54: func sp1(t) tvalue(d; nhctmc, 1;
t)
55: func sp2(t) tvalue(d; nhctmc, 2;
t)
56: func R(t) 1-sp0(t)
57: func f(t) lfr(alfa, beta,
t)*sp1(t) + 2*(1-c)*lfr(alfa, beta,
t)*sp2(t)
58: func h(t) ( lfr(alfa, beta,
t)*sp1(t) + 2*(1-c)*lfr(alfa, beta,
t)*sp2(t) ) / (100*( sp1(t) + sp2(t)
))
59: func rr(t) exrt(d; nhctmc; t)
60: func cr(t) cexrt(d; nhctmc; t)
61: loop t, 0, tmax, d
62:
63:   expr sp0(t)
64:   expr R(t)
65:   expr h(t)
66:   expr f(t)
67:   expr rr(t)
68:   bind crc crc+cr(t)
69:   bind mttfa mttfa+(R(t)*d)
70:
71:   bind
72:     init_nhctmc_0 1-sp1(t)-sp2(t)
73:     init_nhctmc_1 sp1(t)
74:     init_nhctmc_2 sp2(t)
75:   end
76:
77:   expr init_nhctmc_0
78:   expr init_nhctmc_1
79:   expr init_nhctmc_2
80:
81: end
82:
83: expr crc
84: expr mttfa
85:
86: end

```

Figure 4. Piecewise constant approximation—SHARPE code.

Mean Time to Failure

The MTTF of the considered system is 1.81 years. It is computed considering the approximate numerical solution of the integral $\int_0^\infty R(t)dt$ as the sum of the values of $R(t)$ in each time interval of length δ times δ itself (see lines 57, 80, and 96 in Fig. 4).

If we want to increase the average lifetime of the considered system without changing the number of processors of the system, we should

- use the processors with a smaller failure rate;
- improve the detection strategy to increase the coverage;
- improve the recovery strategy to increase the repair rate.

Expected Instantaneous Reward Rate

Figure 9 shows the expected instantaneous reward rate along time. It is computed in

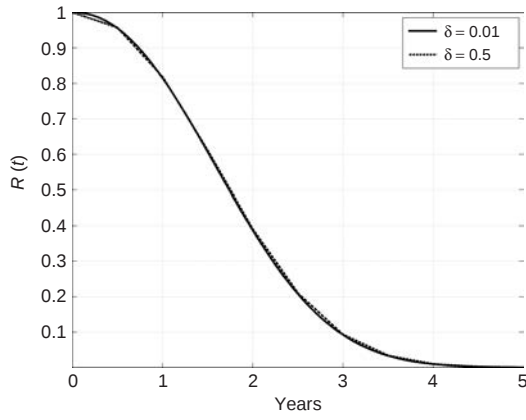


Figure 5. Piecewise constant approximation—reliability.

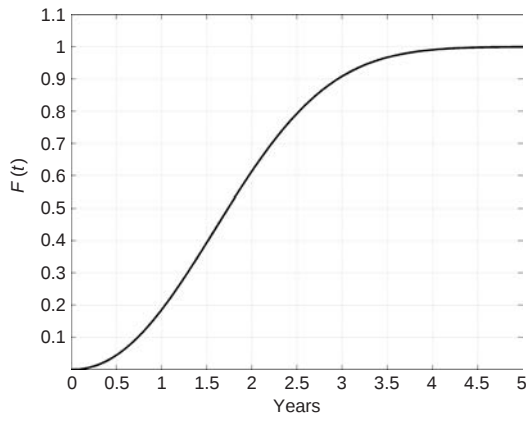


Figure 6. Piecewise constant approximation—distribution function.

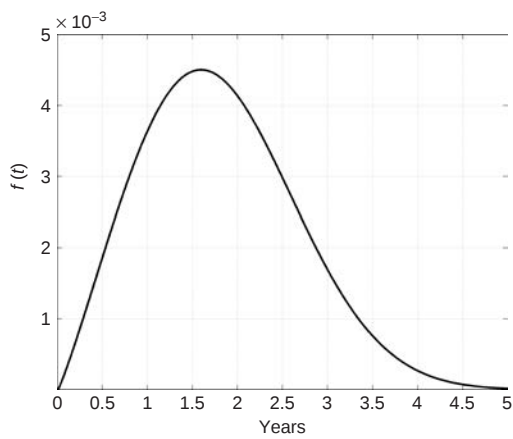


Figure 7. Piecewise constant approximation—probability density function.

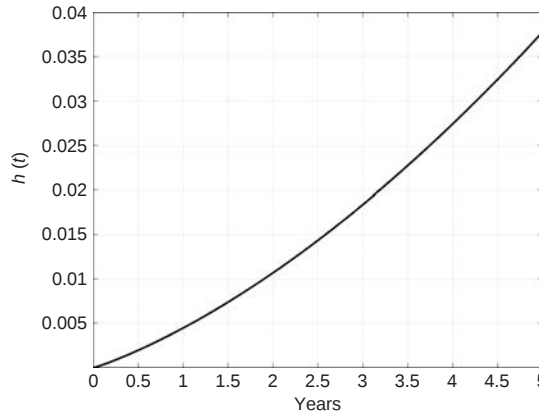


Figure 8. Piecewise constant approximation—hazard rate.

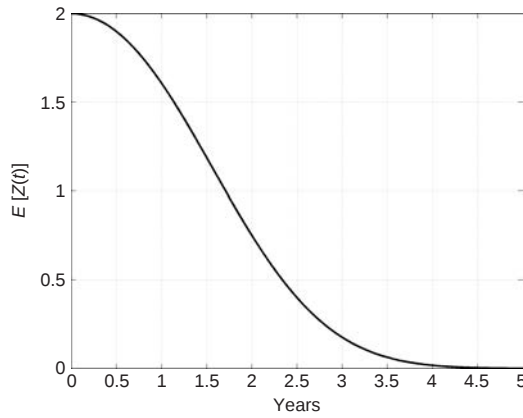


Figure 9. Piecewise constant approximation—expected instantaneous reward rate.

SHARPE with the *exrt()* function (see line 65 in Fig. 4). The trend is very similar to the one of the reliability of the system, apart from the amplitude of the curve. Such a measure is very important in order to quantify the performance of the system in the presence of failure. For instance, suppose the reward rate being the Microprocessor without Interlocked Pipeline Stages (MIPS) of the processors and that it is required for the system to perform at least 0.5 MIPS. The expected instantaneous reward rate shows that after about 2.3 years, the system is no longer able to provide the desired service level. Hence, a renewal strategy is necessary.

Expected Accumulated Reward

The behavior of such an index is shown in Fig. 10 and it is computed in SHARPE by

summing the results of the *cexrt()* function in all the time intervals of length δ (see lines 56 and 66 in Fig. 4). The maximum value is 3.55, reached after about 4 years. Since the model has an absorbing state, the line $y = 3.55$ is an asymptote and this value represents the *expected accumulated reward till absorption*, that is, the expected reward accumulated by the system during its lifetime. Hence, if we consider again the reward rate being the MIPS, the maximum number of instructions the system is able to perform during the observation period $(0, 5]$ years is 1.12×10^{14} .

PHASE-TYPE EXPANSION

In order to present how the repair policy affects the reliability and performability of

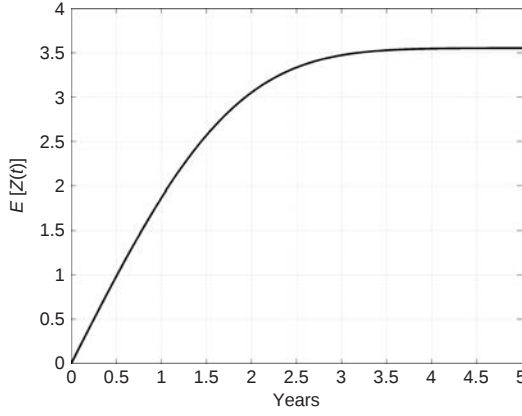


Figure 10. Piecewise constant approximation—expected accumulated reward.

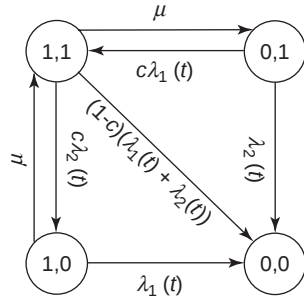


Figure 11. An alternative way for the representation of the system.

the system, now we consider the maximal repair policy. The failure times modeled by the Weibull distribution is approximated using the phase-type expansion technique [15]. Figure 11 shows an alternative way for the representation of the system that makes the description of the phase-type expansion technique simpler and emphasizes the different clocks of each processor. In this model, the failed component and the respective component failure rate are distinguished. For instance, if the system is in *State* (1, 1), both processors are up; in the *States* (0, 1) and (1, 0), one unique processor failed, P_1 and P_2 , respectively. Finally the *State* (0,0) is the absorbing state of the model, which represents that both processors failed, and thus the system failed as well.

The approximation of the Weibull distribution can be obtained using phase-type distributions. PH distributions are introduced in

Ref. 15 and have been widely used in stochastic modeling as they have different practical advantages: the Markovian properties, the closure properties, and the approximating properties. Many studies propose algorithms and numerical techniques for parameter fitting [11,19] and/or for comparing several approximations of nonexponential distributions, such as lognormal and Weibull. In Ref. 15 computation of reliability indices by means of PH distributions is discussed. In Ref. 20, the applicability of PH distributions for modeling the queuing networks is described. In Ref. 21, the authors compare several PH distributions, both continuous and discrete, and investigate which one best approximates a stochastic model.

A PH distribution is defined as the distribution of time to absorption of a CTMC with n transient states and one absorbing state (*State* $n+1$). The infinitesimal generator matrix Q of the CTMC can be partitioned as follows:

$$\begin{pmatrix} Q' & Q^0 \\ 0 & 0 \end{pmatrix} \quad (28)$$

where Q' is a $(n \times n)$ matrix that describes the transition rates between transient states of the CTMC, and Q^0 is the column vector of the transition rates to the absorbing state 0. The tuple $(\pi^{(0)}, Q')$ completely represents the PH distribution of order n , where $\pi^{(0)}$ is the $(n+1)$ initial probability vector. For the formal definition of PH distributions and

their properties, see the article *Phase-Type (PH) Distribution*.

The steps involved in PH expansion are: (i) the choice of stage combinations, for example, stages in series or parallel, and (ii) the derivation of the PH distribution parameters based on parameters of the original distribution. Depending on the approximated stochastic model, several stage combinations can be selected. In Ref. 22, some examples are provided. We use the n -stage-Erlang distribution. The stages are sequentially traversed, and the nonnegative continuous random variable represents the sum of the n independent exponentially distributed random variables with rate γ . Hence, the *pdf* of the resulting random variable X , which will be the approximation of the Weibull distribution, is the *pdf* of the n -stage-Erlang distribution:

$$f(x) = \frac{\gamma^n x^{n-1}}{(n-1)!} e^{-\gamma x} \quad (29)$$

Depending on the distribution to be approximated, different approaches can be used for evaluating the parameters of PH distribution (i.e., γ and n), such as moment matching [22], function fitting, and hybrid methods [11]. We will use the moment matching approach as it allows computing a closed-form expression for the failure rate γ and for the number of stages n as well:

$$\begin{aligned} m_1 &= \frac{n}{\gamma} \\ m_2 &= \frac{n(n+1)}{\gamma^2} \Rightarrow \gamma = \frac{m_1}{m_2 - m_1^2}, \quad n = \frac{m_1^2}{m_2 - m_1^2} \end{aligned} \quad (30)$$

Equalizing m_1 and m_2 to the first two moments of Weibull distribution,

$$m_1 = \beta \Gamma \left(1 + \frac{1}{\alpha} \right), \quad m_2 = \beta^2 \Gamma \left(1 + \frac{2}{\alpha} \right) \quad (31)$$

By substituting $\alpha = 2.1$ and $\beta = 1.02$ (Table 1) in the above formulas, we obtain $n = 4$ and $\gamma = 4.427$. If the computed value n is non-integer, it has to be rounded to the nearest integer. In this case, γ is to be recalculated accordingly.

Figure 12 shows the comparison of the Weibull *pdf* and the approximation obtained with PH expansion; the maximum error is 0.14. It is worth noting that, while selecting a proper stage combination, both the *pdf* and the *hazard rate* have to be checked. Figure 13 shows the comparison of the hazard rates. Since the hazard rates look quite different, it would be preferable to adopt a different stage combination in order to have a better approximation. However, the identification of the best PH approximation for the considered Weibull distribution is out of the scope of this article.

The system model resulting from using PH expansion is presented in Fig. 14. System states are represented specifying the number of phases for each processor. Hence, if the system is in *State (4, 4)*, both the components are up and they are at the stage 4. The failure of component P_1 is represented by *States (0, j)*, with $j = 1, \dots, 4$. The failure of component P_2 is represented by *States (i, 0)*, with $i = 1, \dots, 4$. *State (0, 0)* represents the system failure. A component failure, that is, the transitions from *States (1, j)* to *(0, j)*, or

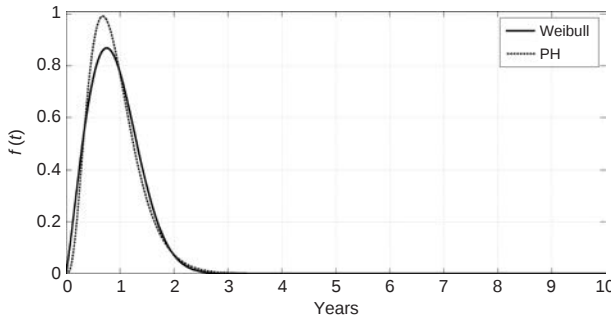


Figure 12. Weibull pdf and its phase type approximation.

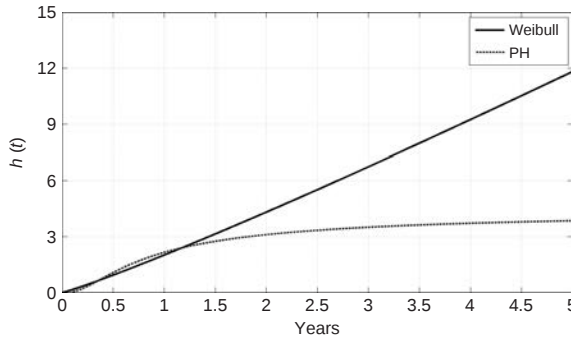


Figure 13. Weibull failure rate and its phase type approximation.

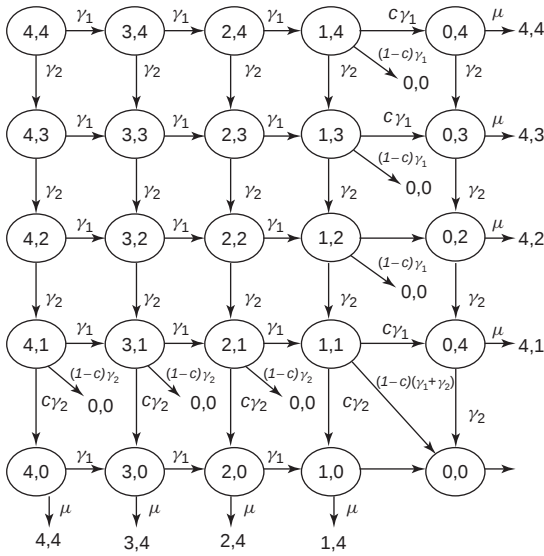


Figure 14. Phase type expansion of the system.

the transitions from *States* $(i, 1)$ to $(i, 0)$, can be covered with probability c . When a failure is not covered, the system fails [transitions from *States* $(i, 1)$ and $(1, j)$ to *State* $(0, 0)$]. When a component fails, it can be repaired or restored and becomes as good as new. This is represented by the transitions from *States* $(0, j)$ to *States* $(4, j)$ and by the transitions from *States* $(i, 0)$ to *States* $(i, 4)$.

An example of the SHARPE code for the specific example under analysis is shown in Fig. 15.

Reliability

The system reliability is computed as $R(t) = 1 - \pi_0(t)$ (Fig. 16). The reliability function is

computed using SHARPE (see line 165 in Fig. 15 for the code).

Cumulative Distribution Function

The *CDF* function is presented in Fig. 17. It is computed by means of SHARPE *tvalue()* function (see line 166 in Fig. 15).

Probability Density Function

Figure 18 presents the *pdf* of the time to failure computed with SHARPE (see line 167 in Fig. 15).

Hazard Rate

The failure rate is shown in Fig. 19. It is computed using the formula defined in the basic definitions section (see line 170 in Fig. 15).

<pre> 1: format 8 2: factor on 5: markov PHModel(lambdax, c, mu, lambday) 6: 1 0 lambdax 7: 1 21 mu 8: 2 1 lambdax 9: 2 22 mu ... 59: 24 23 lambdax 60: 24 19 lambday 62: *Reward definition 63: reward 64: 24 r2 65: 23 r2 ... 86: 2 r1 87: 1 r1 89: end 92: * Initial Probabilities defined 93: 0 init_PHModel_0 94: 1 init_PHModel_1 ... 115: 22 init_PHModel_22 116: 23 init_PHModel_23 117: 24 init_PHModel_24 118: end 120: * Initial Probabilities assigned: 121: bind 122: init_PHModel_0 0 123: init_PHModel_1 0 124: init_PHModel_2 0 ... 145: init_PHModel_23 0 146: init_PHModel_24 1 147: *Rewards 148: r2 2 149: r1 1 151: end 154: echo 155: echo *** Outputs *** 158: bind lambdax 4.427 159: bind c 0.9 160: bind mu 120 161: bind lambday 4.427 162: bind step 0.01 164: *Compute Reliability Indices* 165: func Reliability(t) 1- tvalue(t;PHModel; lambdax, c, mu, lambday) </pre>	<pre> 166: func CDF(t) tvalue(t;PHModel; lambdax, c, mu, lambday) 167: func pdf(t) (tvalue(t;PHModel,21; lambdax, c, mu, lambday)*lambdax+tvalue(t;PHModel,16; lambdax, c, mu, lambday)*lambdax+tvalue(t;PHModel,11; lambdax, c, mu, lambday)*lambdax+tvalue(t;PHModel,6; lambdax, c, mu, lambday)*(lambdax+lambday)+tvalue(t;PHM odel,1; lambdax, c, mu, lambday)*lambdax+tvalue(t;PHModel,7; lambdax, c, mu, lambday)*lambday+tvalue(t;PHModel,8; lambdax, c, mu, lambday)*lambday+tvalue(t;PHModel,9; lambdax, c, mu, lambday)*lambday+tvalue(t;PHModel,5; lambdax, c, mu, lambday)*lambday) 170: func h(t) (pdf(t)/(1- tvalue(t;PHModel; lambdax, c, mu, lambday))) 171: var MTTF mean(PHModel, 0; lambdax, c, mu, lambday) 172: expr MTTF 173: loop t,0, 20, step 174: expr Reliability(t) 175: expr CDF(t) 176: expr pdf(t) 177: expr h(t) 178: end 179: *Compute Performability indices* 180: echo expected reward rate: E[X(t)] 181: echo expected cumulated reward: E[Y(t)] 182: echo distribution of cumulated reward: P[Y(inf)<=r] 183: loop t,0,20,0.01 184: expr exrt(t;PHModel;lambdax, c, mu, lambday) 185: expr cexrt(t;PHModel;lambdax, c, mu, lambday) 186: end 187: loop r,0,50,0.01 188: expr rvalue(r;PHModel;lambdax, c, mu, lambday) 189: end 200: end </pre>
--	---

Figure 15. Phase type expansion—SHARPE code.

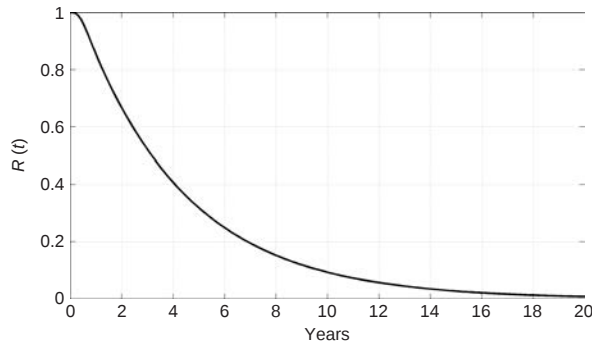


Figure 16. Phase type expansion—reliability.

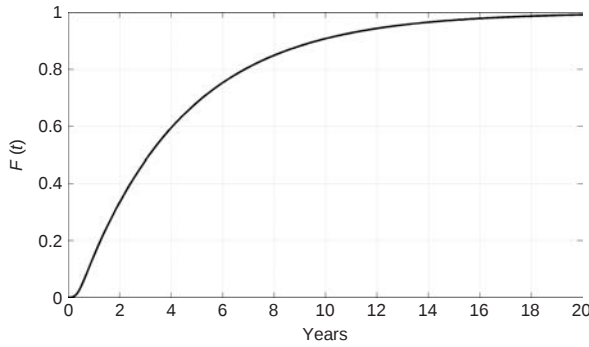


Figure 17. Phase type expansion—distribution Function.

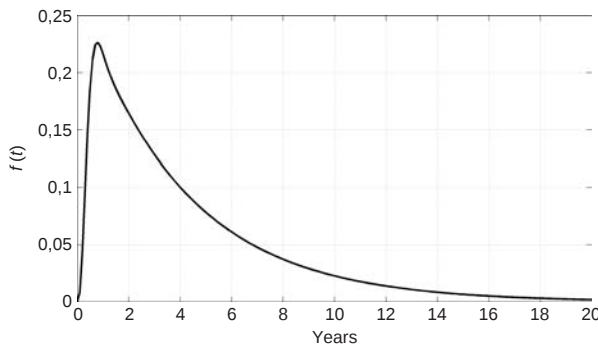


Figure 18. Phase type expansion—probability density function.

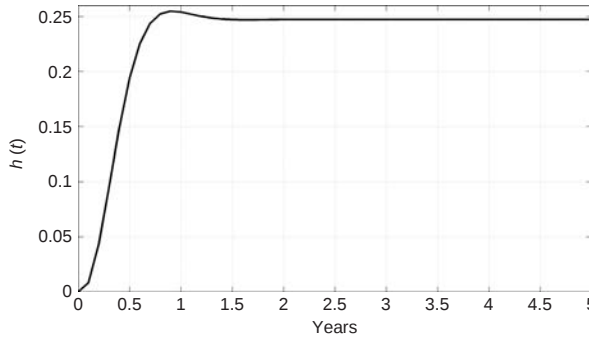


Figure 19. Phase type expansion—hazard rate.

Mean Time to Failure

The MTTF, or mean time to absorption, is using the *mean* function (see line 171 in Fig. 15): $\text{MTTF} = 4.39$ years.

Expected Instantaneous Reward Rate

This rate is computed by means of the *exrt()* function (see line 184 in Fig. 15). Figure 20 shows the trend of the expected reward rate at time t .

Expected Accumulated Reward

It is evaluated using the function *cexrt()* (see line 185 in Fig. 15) and plotted in Fig. 21.

The *expected accumulated reward till absorption* is: $E[Y(\infty)] = 8.66$.

Distribution of the Accumulated Reward till Absorption

The behavior of the distribution is shown in Fig. 22. The distribution of the accumulated

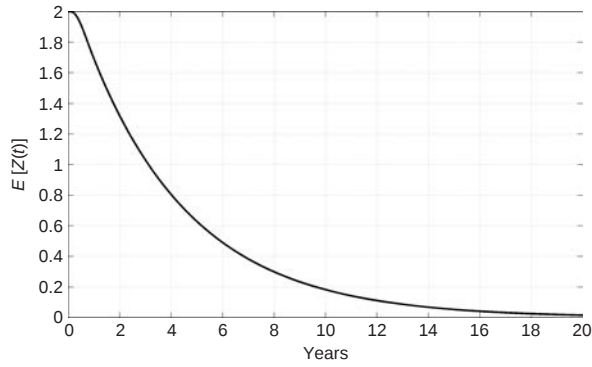


Figure 20. Phase type expansion—expected instantaneous reward rate.

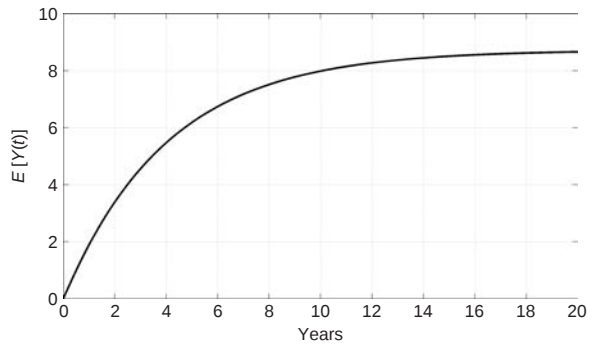


Figure 21. Phase type expansion—expected accumulated reward.

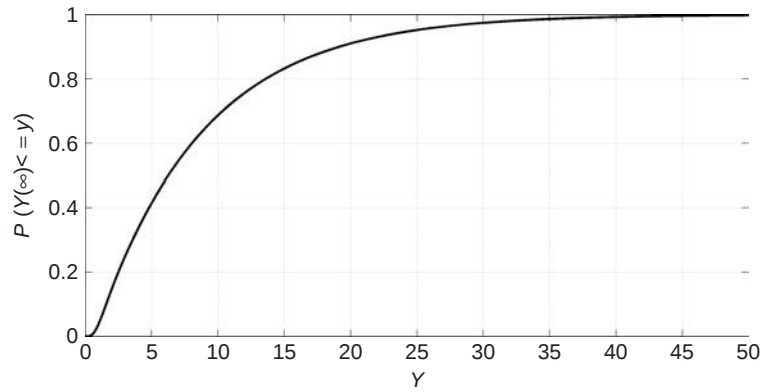


Figure 22. Phase type expansion—distribution of the accumulated reward till absorption.

reward until absorption can be computed numerically using the SHARPE built-in function *rvalue()* (see line 188 in Fig. 15).

SIMULATION

The simulation is an alternative to analytic-numeric solution of models,

especially for large and complex models [23]. Although, it is possible to simulate the Markov model itself, we propose to compute reliability and performability indices of the considered system using stochastic reward nets (SRNs) [24] and to perform simulation by means of SPNP (Stochastic Petri Net Package) [25]. To model the *two-processor*

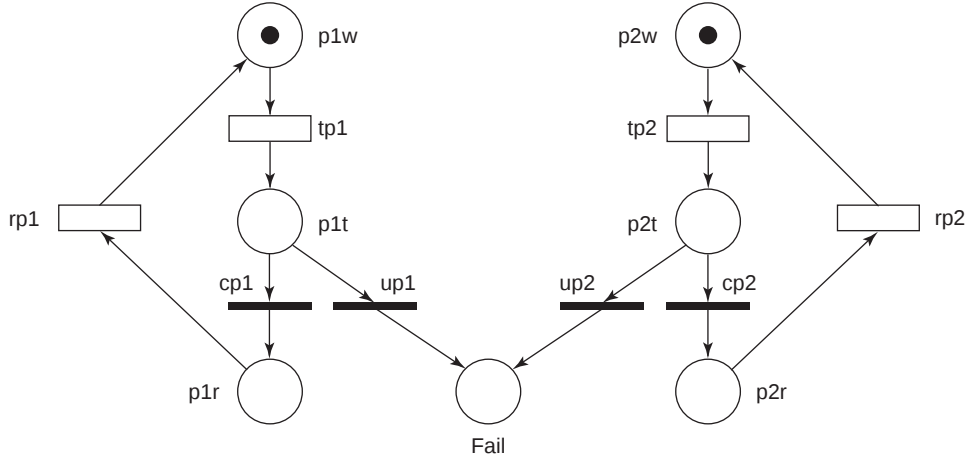


Figure 23. SRN representing the two-processor-parallel-redundant system with repair.

system with an SRN, the behavior of each processor is considered separately. Furthermore, the maximal repair policy (*as good as new*) is adopted. Figure 23 presents the SRN proposed to model the system described above. Let us consider the case of processor P_1 . At the beginning, it is properly working (place $p1w$). A failure can occur (transition $tp1$) leading the processor in a vanishing state (place $p1t$). If the failure is covered (transition $cp1$), the processor is to be repaired (place $p1r$). After repair (transition $rp1$), it works as good as when it was new (place $p1w$). If the failure is not covered (transition $up1$), the processor fails (place $fail$). The behavior of processor P_2 is the same.

Table 2 shows the rates and weights of the transitions.

Figure 24 presents the code to compute reliability and expected instantaneous reward rate in the SPNP software. The maximum number of simulation runs is 100,000, and the required confidence interval is 95% (see lines 22 and 24 of the code,

respectively). Reward functions $rel()$ and $rr()$ (lines 111 and 118 in Fig. 24, respectively) consider the system up if at least one processor is properly working (places $p1w$ and $p2w$) and no processor is failed (place $fail$). The reliability, the expected instantaneous reward rate, and the expected accumulated reward are presented in Figs 25, 26, and 27, respectively.

Mean Time to Failure

The MTTF, evaluated as mean time to absorption, is 4.62 years.

The *expected accumulated reward till absorption* is: $E[Y(\infty)] = 9.15$.

COMPARISON

The importance of the recovery strategy is evident considering the reliability computed with PCA and PH. As shown in Fig. 28, the reliability of the system with the maximal repair strategy, computed with PH expansion, is much greater than the one with minimal repair strategy computed with the PCA.

With minimal repair, the reliability becomes 0 after about 4.5 years, while after the same time reliability obtained with maximal repair is greater than 0.35. As a matter of fact, also the MTTF is much different. Maximal repair provides an average lifetime

Table 2. Rates and weights of the transitions.

Transition	Rate/weight
tp1, tp2	Weibull (α, β)
rp1, rp2	μ
cp1, cp2	c
up1, up2	$1 - c$

```

1: #include <stdio.h>
2: #include "user.h"
3:
4: double mu = 120;
5: double c = 0.9;
6:
7: double a = 2.1;
8: double b = 1/1.02;
9:
10: double rel();
11: double rr();
12:
13: /* OPTIONS */
14:
15: void options() {
16:
17:     iopt(IOP_SIMULATION, VAL_YES);
18:     iopt(IOP_SIM_RUNMETHOD, VAL_REPL);
19:     iopt(IOP_SIM_CUMULATIVE, VAL_NO);
20:     iopt(IOP_SIM_STD_REPORT, VAL_YES);
21:     iopt(IOP_SIM_SEED, 345983453);
22:     iopt(IOP_SIM_RUNS, 100000);
23:     fopt(FOP_SIM_ERROR, 0.0001);
24:     fopt(FOP_SIM_CONFIDENCE, .95);
25:     fopt(FOP_SIM_LENGTH, 20.0);
26:
27: }
28:
29: /* DEFINITION OF THE NET */
30:
31: void net() {
32:
33:     /* PLACES */
34:
35:     place("plw");
36:     place("plt");
37:     place("plr");
38:
39:     place("p2w");
40:     place("p2t");
41:     place("p2r");
42:
43:     place("fail");
44:
45:     init("plw", 1);
46:     init("p2w", 1);
47:
48:     /* TRANSITIONS */
49:
50:     weibval("tp1", b, a);
51:     imm("cp1");
52:     imm("up1");
53:     probval("cp1", c);
54:     probval("up1", (1-c));
55:     rateval("rp1", mu);
56:
57:     weibval("tp2", b, a);
58:     imm("cp2");
59:     imm("up2");
60:     probval("cp2", c);
61:     probval("up2", (1-c));
62:     rateval("rp2", mu);
63:
64:     /* ARCS */
65:
66:     /* Input Arcs */
67:     iarc("tp1", "plw");
68:     iarc("cp1", "plt");
69:     iarc("up1", "plt");
70:     iarc("rp1", "plr");
71:
72:     iarc("tp2", "p2w");
73:     iarc("cp2", "p2t");
74:     iarc("up2", "p2t");
75:     iarc("rp2", "p2r");
76:
77:     /* Output Arcs */
78:     oarc("tp1", "plt");
79:     oarc("cp1", "plr");
80:     oarc("up1", "fail");
81:     oarc("rp1", "plw");
82:
83:     oarc("tp2", "p2t");
84:     oarc("cp2", "p2r");
85:     oarc("up2", "fail");
86:     oarc("rp2", "p2w");
87:
88: }
89:
90:
91: /* FUNCTIONS */
92:
93: int assert() {}
94:
95:
96: void ac_init() {
97:     /* Information on the net structure */
98:     pr_net_info();
99: }
100:
101:
102: void ac_reach() {}
103:
104:
105: double rel() {
106:     if( (mark("plw")==1 ||
107: mark("p2w")==1) && (mark("fail")==0) )
108:         return(1.0);
109:     else
110:         return(0.0);
111: }
112:
113: double rr() {
114:     if( (mark("plw")==1 ||
115: mark("p2w")==1) && (mark("fail")==0) )
116:         return(mark("plw")+mark("p2w"));
117:     else
118:         return(0.0);
119: }
120:
121: /* OUTPUT */
122:
123: void ac_final() {
124:     pr_expected("Reliability", rel);
125:     pr_expected("Reward", rr);
126:     pr_cum_expected("Accumulated", rr);
127: }

```

Figure 24. Simulation—SPNP code.

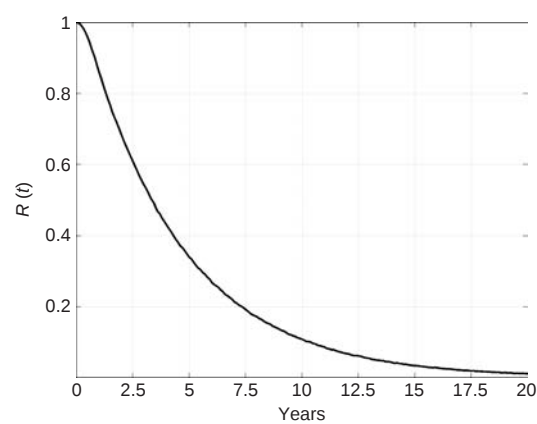


Figure 25. Simulation—reliability.

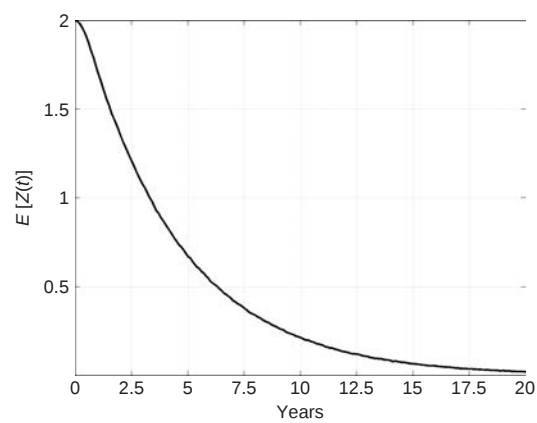


Figure 26. Simulation—expected instantaneous reward rate.

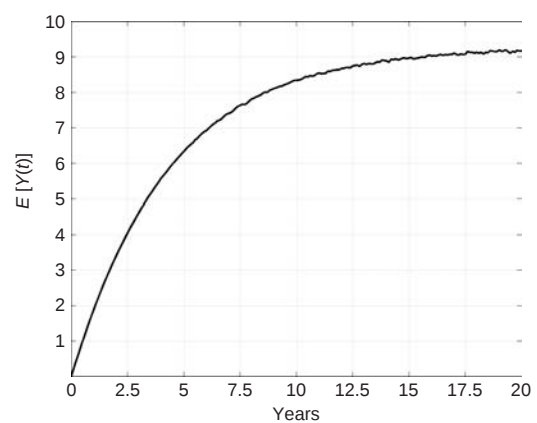


Figure 27. Simulation—expected accumulated reward.

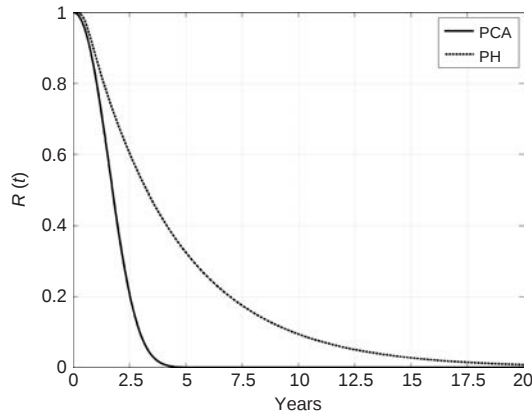


Figure 28. Comparison of the reliability computed with piecewise approximation and phase type expansion.

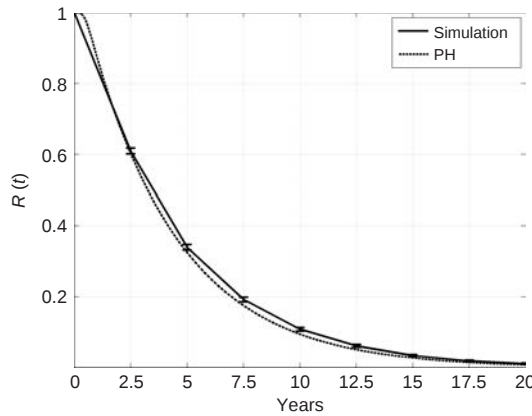


Figure 29. Comparison of the reliability computed with simulation and phase type expansion.

more than two times the one achieved when using an *as bad as before* repair strategy.

In the cases of PH expansion and simulation, the same recovery strategy is adopted. Figure 29 compares the reliability computed with the two techniques. Results obtained with PH expansion fall into the corresponding confidence intervals estimated with simulation when $t \leq 2.5$. Nevertheless, the trends of the reliability are very similar. Such a result is also confirmed by the mean times to failure.

REFERENCES

1. International Electrotechnical Commission. IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems. 2005.
2. RTCA SC-167, EUROCAE WG-12. DO-178B. Software Considerations in Airborne Systems and Equipment Certification. 1992.
3. Trivedi KS, Andrade EC, Machida F. Combining performance and availability analysis in practice. *Adv Comput* 2012;84:1–38.
4. Heath T, Martin RP, Nguyen TD. Improving cluster availability using workstation validation. *SIGMETRICS Perform Eval* 2002;30(1):217–227.
5. Sahoo RK, Sivasubramaniam A, Squillante MS, Zhang Y. Failure data analysis of a large-scale heterogeneous server environment. *Proceedings of International Conference on Dependable Systems and Networks*, Florence, Italy, 2004. p 772–781.
6. Arnold TF. The concept of coverage and its effect on the reliability model of a repairable system. *IEEE Trans Comput* 1973;C-22(3):251–254.

7. Sahner RA, Trivedi KS, Puliafito A. Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package. Boston, USA: Kluwer Academic Publishers; 1996.
8. Trivedi KS. Probability and statistics with reliability, queuing, and computer science applications. New York: John Wiley and Sons; 2001.
9. Beaudry MD. Performance-related reliability measures for computing systems. *IEEE Trans Comput* 1978;C-27(6):540–547.
10. Ciardo G, Marie RA, Sericola B, Trivedi KS. Performability analysis using semi-Markov reward processes. *IEEE Trans Comput* 1990;39(10):1251–1264.
11. Malhorta M, Reibman A. Selecting and implementing phase approximations for semi-Markov models. *Stoch Models* 1993;9(4):473–506.
12. Rindos A, Woollet S, Viniotis I, Trivedi KS. Exact methods for the transient analysis of nonhomogeneous continuous-time Markov chains. In: Stewart WJ, editor. 2nd International Workshop on the Numerical Solution of Markov Chains. Boston, USA: Kluwer Academic Publishers; 1995.
13. Fortmann TE, Hitz KL. An Introduction to Linear Control Systems, in Control and Systems Theory, Volume 5, A Series of Monographs and Textbooks. New York: Marcel Dekker; 1977. p 589–599.
14. Takacs L. Introduction to the theory of Queues. New York: Oxford University Press; 1962.
15. Neuts MF, Meier KS. On the use of phase type distributions in reliability modeling of systems with two components. *Operat Res Spectrum* 1980;2(4):227–234.
16. Ascher HE. Evaluation of repairable system reliability using the “Bad-As-Old” concept. *IEEE Trans Reliab* 1968;R-17(2):103–110.
17. Trivedi KS, Sahner R. SHARPE at the age of Twenty Two. *ACM SIGMETRICS Perform Eval Rev* 2009;36(4):52–57.
18. Bao Y, Sun X, Trivedi KS. A workload-based analysis of software aging, and rejuvenation. *IEEE Trans Reliab* 2005;54(3):541–548.
19. Johnson M, Taaffe M. Matching moments to phase distributions: nonlinear programming approaches. *Stoch Models* 1990;6:259–281.
20. Neuts MF. Matrix geometric solutions in stochastic models. Baltimore (MD): Johns Hopkins University Press; 1981.
21. Bobbio A, Horvath A, Telek M. The scale factor: a new degree of freedom in phase-type approximation. *Perform Eval* 2004;56(1–4):121–144.
22. Singh C, Billinton R, Lee SY. The method of stages for non-Markov models. *IEEE Trans Reliab* 1977;R-26(2):135–137.
23. Tuffin B, Choudhary PK, Hirel C, Trivedi KS. Simulation versus analytic-numeric methods: illustrative examples. Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools; Nantes, France, 2007. p 63.1–63.10.
24. Ciardo G, Blakemore A, Chimento PF, Muppala JK, Trivedi KS. Automated generation and analysis of Markov reward models using stochastic reward nets. In: Meyer C, Plemmons RJ, editors. Linear Algebra, Markov Chains, and Queueing Models—IMA volumes in mathematics and its applications. Heidelberg, Germany: Springer-Verlag; 1992.
25. Ciardo G, Muppala J, Trivedi KS. SPNP: Stochastic Petri Net Package. International workshop on petri nets and performance models. Proceedings of the 3rd International Workshop on Petri Nets and Performance Models (PNPM89), Volume 96; Kyoto, Japan, 1989. p 142–151.

REMANUFACTURING

GILVAN C. SOUZA
Kelley School of Business,
Indiana University,
Bloomington, Indiana

A supply chain is a network of facilities involved in the manufacturing and distribution of a product, involving procurement of raw materials, manufacturing of components, final assembly, storage in warehouses of distributors, and sales in retail stores or other distribution channels. The supply chain for beer, for example, includes procurement of beer ingredients such as yeast, barley, hops, and water; beer preparation through mixing and fermentation of ingredients; bottling, which could be done in a separate facility; shipment to national beer distributors, then to regional distributors, and finally to retail stores, where beer is sold. In supply chains, there are thus physical flows of raw materials, components, semifinished, and finished products; there are also financial flows, and information flows in the form of demand forecasts, inventory positions at different storage points, and other information. For more information on supply chains. The physical flows in a traditional supply chain are mostly unidirectional, from raw materials producers through manufacturers, distributors, and then to consumers. These are called *forward* flows.

In reverse supply chains, the flows of products are in the opposite direction, from consumers to producers. As an example, the reverse supply chain for beer cans involves collection of used beer cans, consolidation in intermediate storage points, and shipment to aluminum producers and/or recyclers. The term *closed-loop supply chain* (CLSC) indicates a supply chain where there is a combination of forward and reverse flows, such that these two types of flows may impact each other, and may thus require some level of coordination.

An example of CLSC is shown in Fig. 1. Pitney Bowes is an Original Equipment

Manufacturer (OEM) based in Stamford, Connecticut, which manufactures mailing equipment that matches customized documents to envelopes, weighs the parcel, prints the postage, and sorts mail by zip code. Pitney Bowes leases about 90% of its new product manufacturing, and sells the remainder. A typical leasing contract is for 4 years, and a typical life cycle for a Pitney Bowes product is 6 years. At the end of a leasing contract, customers may upgrade to newer generation equipment if available. If that is the case, the customer returns the used equipment to Pitney Bowes, which tests and evaluates the condition of the used machine, and makes a *disposition* decision: scrap for materials recovery (recycling), which is done for the worst quality returns; dismantle for spare parts harvesting, which is done for medium quality returns; or remanufacturing, which is done for the best quality returns so as to reduce remanufacturing cost. Remanufacturing consists of bringing the used product to a common operating and aesthetic standard, often with upgrades in some of the product's functions. More details on the remanufacturing planning process at Pitney Bowes can be found in Ref. 1. Remanufactured products are then sold as a cheaper alternative to new products. Note how the forward and reverse flows impact each other: customers may opt for either remanufactured or new products, depending on respective price levels, and perceived quality of the remanufactured product by the customer. Further, lease or sales of new products impact the flow of available used products for remanufacturing in the future. The CLSC for Pitney Bowes thus requires coordination between the forward and reverse flows.

It is important to distinguish the types of product returns in CLSCs, and the available disposition decisions for product returns. There are three types of product returns in CLSCs:

- *Consumer Returns.* These are lightly used products, most of which are not

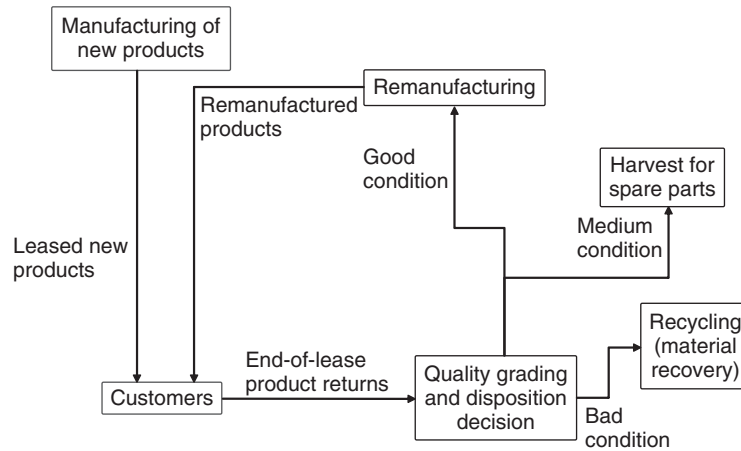


Figure 1. An example of a closed-loop supply chain with remanufacturing.

defective, which are returned to retailers (and then to manufacturers) by consumers as a result of liberal return policies. Typical reasons for returns include: buyer remorse, a better price elsewhere, a product that does not meet performance expectations, or technical complexity. Consumer returns account for 5–6% of personal computer sales, and up to 35% for fashion products.

- *End-of-Use Returns.* These are illustrated in the Pitney Bowes case above. Contrary to consumer returns, end-of-use returns have been extensively used by customers, their quality level is more variable depending on usage patterns, and thus remanufacturing here is a more extensive operation.
- *End-of-Life Returns.* These returns have reached the end of their useful life; their quality condition is poor and materials and/or part recovery are common disposition options.

In addition to landfilling—an option that is illegal for some products in some jurisdictions (e.g., electronic equipment cannot be landfilled in some US states such as California, Maine, Massachusetts, and Minnesota)—disposition decisions for product returns include:

- *Incineration.* Incineration reduces the amount of solid waste going to landfills—a significant issue for countries with limited landfill space—and it may be an attractive option for products or materials where recycling is difficult or uneconomical. In addition, incineration is commonly used for energy recovery. For example, Denmark incinerates 58% of its municipal solid waste toward energy recovery [2]. On the other hand, incineration increases the amount of toxic emissions such as mercury, and is therefore regulated.
- *Recycling.* Recycling means material recovery, and it is an attractive option for products where returns have little economic value due to technological obsolescence (e.g., old computers), or are in poor quality condition (e.g., product returns heavily damaged during transportation). Recycling can be mandated by regulation: in the Netherlands, 85% of the weight of each car at the end of its life needs to be recycled (as opposed to landfilled or incinerated); for electronic equipment, the current mandated recycling target is about 65% in the European Union. See introductory articles *Recycling* and *Take-Back Legislation and*

Its Impact on Closed-Loop Supply Chains.

- *Parts Harvesting.* Here, the firm recovers selected parts from a product return for use in warranty and service contracts. When the part is subject to wear and tear, as is common in mechanical components, this option is viable if the product has been lightly used such as consumer returns. Otherwise, for electronic components, this is a common disposition option. Fleischmann *et al.* [3] estimate that IBM can save as much as 80% per part by dismantling product returns relative to procuring the part from a supplier.
- *Remanufacturing.* This is a value-added operation as illustrated in the Pitney Bowes case above, and it can be the most profitable disposition decision. The economic viability of remanufacturing depends on several factors: the existence of a product core that retains significant value after usage (e.g., engine block), the existence of a market for remanufactured products, which is related to customers' acceptance of remanufactured products, total remanufacturing, and reverse logistics costs (collection, transportation, testing, labor and parts), among other factors.
- *Resale (As-Is).* This can happen if a secondary market for the used product exists, such as is the case with used cars, and some standard IT equipment.

Remanufacturing has the potential for higher profitability among the disposition decisions; it is thus the focus of this article. In new item manufacturing, materials and components are completely homogeneous. In used item remanufacturing, on the other hand, the key inputs—product returns, also called *cores*—are uncertain in quality, timing, and quantity. Product returns thus need to be graded according to their quality; returns with worse quality consume more capacity (more labor and equipment hours) and cost more to remanufacture than better quality returns. The uncertainty with respect

to timing and quantity can be minimized through proactive acquisition strategies, such as controlling the price paid for product returns [4]. We show subsequently, how operations research can be used for production planning with remanufacturing: the quantities of each type of product return that should be remanufactured, disposed of, or harvested for parts in each period.

Remanufacturing also offers another challenge on *remarketing*. It is often the case that remanufactured products are perceived as inferior to new products; as a result they are priced lower than their equivalent new counterparts. Hauser and Lund [5] report 35–55% average price discounts for remanufactured products relative to equivalent new products in industries such as automotive parts, compressors, machinery, office furniture, toner cartridges and the like. Similarly, Subramanian and Subramanyam [6] report price discounts for remanufactured products relative to new products, with the same warranty, ranging from 15% for consumer electronics to 40% for video game consoles at eBay. The key decisions in remarketing are pricing and distribution (channels of distribution) for remanufactured products relative to new products so as to maximize the firm's profitability, considering potential cannibalization of sales between new and remanufactured products. We present a model in the section titled "Pricing of Remanufactured versus New Products" that illustrates how this decision can be made.

Besides remarketing and production planning, there are many other decisions that need to be made in CLSCs. Like in regular forward chains, remanufacturing decisions in CLSCs can be classified into strategic, tactical, and operational decisions. We illustrate some of these decisions in Table 1, and offer further references in that table for readers interested in those particular topics.

PRICING OF REMANUFACTURED VERSUS NEW PRODUCTS

In this section, we show how operations research and economic models can be used to study the strategic interaction between

Table 1. Strategic, Tactical and Operational Decisions in Remanufacturing with Some References

Type of Decision	Time Frame	Examples
Strategic	Several years	• How should the reverse supply chain be designed? Specifically, where are the locations of collection sites, consolidation sites, remanufacturing, and recycling sites? [7–13] • Should an Original Equipment Manufacturer (OEM) remanufacture? If so, what should be the prices of remanufactured products? [14–16]) • Should the firm lease or sell its products?
Tactical	Several months	• Product acquisition for remanufacturing: how many returns should be procured, at which quality level, and at which price? [4,17–19] • Production planning for remanufacturing: how many returns of each type (quality grade) should be remanufactured and disposed of in each period? [1,20,21]
Operational	Hours/days	• How should the different remanufacturing operations be scheduled at the shop? [22] • Disassembly planning: what is the sequence and depth of disassembly? [23]

remanufactured and new products; the models provide recommendations regarding pricing of remanufactured and new products to maximize a firm’s profitability. We consider a specific business setting; however, the basic modeling setup can be extended to study other settings. Specifically, we consider a monopolist firm that sells (new) products, which have a useful life (to the consumer) of one period. For example, for photocopiers, a period would be, say, 5 years, after which it can no longer be used productively by the consumer.

To keep the exposition simple, we consider a time horizon of two periods, and use subscript i to indicate period $i, i = 1, 2$. In period 1, the firm only sells new products. Consumers who buy a new product in period 1 have a used product at the beginning of period 2. If a consumer has a used product at the beginning of period 2, and decides to buy another new product, she *can* return her used product to the firm “for free” (e.g., consumers can return old computers to Dell for free upon purchase of a new Dell computer). The firm *can* use these used products recovered from the consumers, for remanufacturing. Like new products, remanufactured products also have a useful life of one period; one can thus think of remanufacturing as extending the product’s useful life to another

period. For example, for a copier, remanufacturing would consist of replacing worn out parts, performing cosmetic repairs, such as removing scratches (which may involve replacing parts), and upgrading some modules to the latest technology. Assume that the firm collects, in period 2, a fraction $\rho < 1$ of all used products in the market (i.e., those sold in period 1), which it can remanufacture. The remaining fraction $(1 - \rho)$ is destroyed or kept by consumers (say, in their garages) and does not interfere in period 2’s market.

Thus, at the beginning of period 2, any consumer (regardless of whether she has a used product or not) has three choices: (i) she buys another new product, (ii) she buys a remanufactured product, or (iii) she does not buy anything. Remanufactured products are labeled as such, and sold at a discount relative to the new product; the firm would like to know which price would maximize its profitability. The basic argument for engaging in remanufacturing in this case is to reach customers who believe that the new product’s price is too high and thus do not buy it; remanufacturing thus allows the firm to expand its customer base and increase revenues. Although consumers prefer new to remanufactured products, a low enough price of the remanufactured product may provide enough of an incentive for them

to buy remanufactured products instead of new, causing *cannibalization*; too much cannibalization can be detrimental to profitability because the new product commands higher margins. We now provide a simple modeling framework to study this problem.

A key characteristic of the model is modeling consumer behavior. For a given product, different consumers have different willingness-to-pay (the maximum amount they are willing to pay) for the new product, which we denote by θ . For example, some consumers have little use for a photocopier, and therefore would not be willing to pay much for it. In this case, their willingness-to-pay for the copier is near zero. On the other hand, some consumers (e.g., law firms) make extensive use of photocopiers, and therefore are willing to pay a significant amount for a copier that meets their needs, say \$10,000. Between these two extremes, are most consumers: one can model the consumer base by assuming a willingness-to-pay for the new product for a random consumer to be between 0 and 10,000. To eliminate parameters in the model, the minimum willingness-to-pay (in this case, near zero), and the maximum willingness-to-pay (in this case 10,000) are normalized to 0 and 1, respectively. A common assumption on the way that consumers distribute between the two extremes is to assume a uniform distribution of consumer types: every willingness-to-pay θ for the new product between 0 and 1 is equally as likely. We denote this by $\theta \sim U[0, 1]$. A uniform distribution for willingness-to-pay implies a downward sloping, linear demand curve. To see this, denote the price of the new product in the first period by p_{n1} , $0 < p_{n1} < 1$. Then, all consumers with willingness-to-pay $\theta > p_{n1}$ derive a positive utility (defined as willingness-to-pay minus price) for the new product; given that the potential market size is normalized to 1, the quantity of new products sold is given by $q_{n1} = 1 - \Pr\{\theta > p_{n1}\} = 1 - p_{n1}$. A key assumption made here is that consumers are not *strategic*, that is, they are not able to anticipate a future presence of a remanufactured product in the second period (or a potentially lower price for the new product) when making a first period purchasing decision.

In the second period, the monopolist offers both new and remanufactured products. Again, given that any product purchased in the first period reached its useful life, all consumers in the market are again looking to buy a new or remanufactured product in the second period. As discussed in the introduction, consumers typically view remanufactured products as inferior relative to new products; this is modeled by considering that a consumer that is willing to pay θ for the new product is also willing to pay $\delta\theta$ for the remanufactured product, where $0 \leq \delta \leq 1$. If $\delta = 0$, the consumer does not consider the remanufactured price as a substitute, regardless of price. This could be the case, for example, for certain medical supplies that have contact with blood; consumers may perceive them as unsafe, even after sterilization. If $\delta = 1$, remanufactured and new products are perfect substitutes; a common example is toner cartridges—most consumers see no difference between remanufactured and new cartridges. Most markets fall in between these two extremes, with $0 < \delta < 1$: consumers prefer the new over the remanufactured product if they have the same price. To derive the quantities sold for new and remanufactured products, we consider that a consumer chooses the product that gives him/her the maximum utility. So, consider a remanufactured product priced at $p_{r2} < p_{n2}$. A customer of type $\theta < p_{n2}$ has negative utility for the new product; his/her utility for the remanufactured product is $\delta\theta - p_{r2}$, which is positive if $\theta > p_{r2}/\delta$. A customer of type $\theta > p_{n2}$, however, has positive utilities for both products but for $\theta > (p_{n2} - p_{r2})/(1 - \delta)$, the utility of the new product is higher than for the remanufactured product (this threshold can be found by solving $\theta - p_{n2} > \delta\theta - p_{r2}$ for θ). As a result, the quantity of new products sold in the second period is $q_{n2} = 1 - (p_{n2} - p_{r2})/(1 - \delta)$ (this comprises all customers who have a higher utility for the new product), whereas the quantity of remanufactured products sold is $q_{r2} = (p_{n2} - p_{r2})/(1 - \delta) - p_{r2}/\delta$ [this comprises all customers who have a higher (positive) utility for the remanufactured product]. With this basic interaction modeled, the firm can solve

an optimization problem to find the optimal prices of remanufactured and new products.

Assume that the variable production cost for the new product in the first period, new product in the second period, and remanufactured product are c_{n1} , c_{n2} , and $c_r < \min\{c_{n1}, c_{n2}\}$, respectively. The firm solves the following optimization problem:

$$\begin{aligned} \max_{p_{n1}, p_{n2}, p_r} \quad & \Pi = q_{n1}(p_{n1} - c_{n1}) + q_{n2}(p_{n2} - c_{n2}) \\ & + q_{r2}(p_{r2} - c_r), \end{aligned} \quad (1)$$

$$\text{s.t. } q_{r2} \leq \rho q_{n1}, \quad (2)$$

$$q_{r2}, q_{n1}, q_{n2} \geq 0, \quad (3)$$

where as shown above, $q_{n1} = 1 - p_{n1}$, $q_{n2} = 1 - (p_{n2} - p_{r2})/(1 - \delta)$, and $q_{r2} = (p_{n2} - p_{r2})/(1 - \delta) - p_{r2}/\delta$.

Equation (1) is the total profit across period; the terms represent, respectively, profit from the first period, new and remanufactured products from the second period. Constraint (2) indicates that the number of remanufactured products cannot exceed the number of available used products sold in the first period (i.e., this is a used products availability constraint). Finally, constraints (3) are logical constraints that ensure that the prices chosen do not result in nonnegative quantities. The nonlinear optimization problem (Eqs 1–3) can be solved analytically or numerically (see **Foundations of Constrained Optimization**); the analytical solution is not as straightforward as it results in multiple “regions,” such that the optimal prices are dependent on different combinations of the cost and δ parameters. In general, the firm remanufactures if $c_r < \delta c_{n2}$. For more detailed analyses the reader is referred to Refs 1 and 15 and references therein.

Finally, we note that the model above is deterministic: demand is a deterministic function of prices, and the quantity of returns is a deterministic function of previous sales. This is standard in higher-level strategic modeling of CLSCs, given the period length under consideration (the product’s useful life). Uncertainty clearly exists at a more tactical level (e.g., monthly demand for remanufactured products or number of

returns received or procured in a given week). Strategic models of CLSCs, however, consider longer periods of time (one period comprises several years in this case), and as a result there are several reasons why uncertainty is not included in models:

(i) there is less uncertainty in the data when it is aggregated over longer periods of time (e.g., for a stable product, yearly demand is less uncertain than monthly demand, and so forth); (ii) there is less to be gained in modeling the system at that level of detail, considering the decision variables of interest here (prices); and finally (iii) these models are not meant to be decision support models, but rather provide broader insights (e.g., under which conditions should a firm remanufacture.)

PRODUCTION PLANNING FOR REMANUFACTURED PRODUCTS

We now describe how production planning for remanufacturing is different from production planning for new products, and how operations research (specifically linear programming; see articles in the section titled “Linear Programming (LP)” in this encyclopedia) can help a planner decide on the optimal production plan. As discussed in the introduction, the basic difference is that the basic material input for remanufacturing—cores or returns—are not homogeneous; there are differences in their quality and availability during the planning horizon. A good quality return demands less processing capacity from the facility, and costs less to remanufacture than a bad quality return. If the firm has excess returns in a given period, it can salvage them (for example, by selling to a recycler, or dismantling for spare parts). We present a rather simple model for illustration; for more information on the environments faced by remanufacturing firms. Refs 1, 20, and 21 provide a more comprehensive coverage in the use of operations research models for production planning in remanufacturing.

Our simple model is meant to represent an *aggregate* production plan, where the planner decides upon the overall remanufacturing quantities for a given product family

(e.g., Motorola cell phones), as opposed to individual product models (e.g., the models V750, Razr2, and E8 by Motorola). Our model assumes that there are demand forecasts D_t for the family of remanufactured products under consideration for each period t in the planning horizon, and forecasts B_{it} for returns of each quality i for each period t in the planning horizon. There are mechanisms for forecasting the total quantity of returns received in each period (denoted by $B_{\bullet t} = \sum_i B_{it}$); this is reasonably straightforward if the firm leases its new item production (i.e., $B_{\bullet t}$ would be the number of expiring leases in period t), otherwise time series methods such as triple exponential smoothing can be used (see **Holt–Winters Exponential Smoothing**). Given $B_{\bullet t}$, an estimate of B_{it} would be $B_{it} = q_i B_{\bullet t}$, where q_i is the historic average proportion of returns of quality i . We introduce our complete notation below.

INDEXES

- i quality, $i = 1$ (best), $\dots, I$ (worse)
 t time period, $t = 1, \dots, T$

PARAMETERS

- D_t Demand forecast for remanufactured products at period t
 B_{it} Quantity of returns of quality i at period t
 s_i Unit salvage value for returns of quality i that are not remanufactured
 h_i Unit holding cost for return of quality i
 h_r Unit holding cost for remanufactured product
 b Unit backlogging cost
 c_i Unit remanufacturing cost for return of quality i
 a_i Unit capacity usage by return of quality i
 K_t Total capacity available at time t

DECISION VARIABLES

- z_{it} Quantity of quality i returns remanufactured in period t

- v_{it} Quantity of quality i returns salvaged in period t
 y_t^+ Inventory of remanufactured products at the end of period t
 y_t^- Backlog of remanufactured products at the end of period t
 u_{it} Inventory of quality i returns at the end of period t

The remanufacturing problem can be formulated as a linear program (LP) as follows:

$$\min TC = \sum_{t=1}^T \sum_{i=1}^I \{c_i z_{it} - s_i v_{it} + h_i u_{it}\} + h_r y_t^+ + b y_t^-, \quad (4)$$

$$\text{s.t. } (y_{t-1}^+ - y_{t-1}^-) + \sum_i z_{it} - (y_t^+ - y_t^-) = D_t, \quad t = 1, \dots, T; \quad (5)$$

$$z_{it} + u_{it} - u_{i,t-1} + v_{it} = B_{it}, \quad i = 1, \dots, I; \quad t = 1, \dots, T; \quad (6)$$

$$\sum_i a_i z_{it} \leq K_t \quad t = 1, \dots, T; \quad (7)$$

$$z_{it}, u_{it}, v_{it}, y_t^+, y_t^- \geq 0, \quad i = 1, \dots, I; \quad t = 1, \dots, T. \quad (8)$$

The objective function in Equation (4) minimizes total costs, comprising variable production cost, salvage cost (note the negative sign for the salvage revenue), holding cost for returns, holding cost for remanufactured products, and backlogging cost. The set of constraints (5) represents the inventory balance equations for remanufactured products: for each period t , beginning inventory net of backlogs ($y_{t-1}^+ - y_{t-1}^-$) plus remanufacturing production ($\sum_i z_{it}$) minus demand (D_t) is equal to ending inventory net of backlogs $y_t^+ - y_t^-$. The set of constraints (6) represents the inventory balance equations for returns according to quality level: for each quality i and time period t , beginning inventory ($u_{i,t-1}$) plus returns received (B_{it}) minus quantity remanufactured (z_{it}) minus quantity salvaged (v_{it}) equals ending inventory u_{it} . The set (7) represents the capacity constraints for each period. Finally, the set (8) represents the nonnegativity constraints.

Our model does not incorporate fixed costs for remanufacturing in a period (also called as *setup costs*) because remanufacturing is for the most part a labor intensive operation, not requiring elaborate machine setups; our model is also targeted at the family level, where setup costs are less of a concern.

The problems (4–8) can be solved numerically to find the optimal production plan. Ferguson *et al.* [1] show that if there are no capacity constraints (i.e., K_t is large enough for all t), then the firm always remanufactures the exact quantity demanded in each period if $h_r > h_i$. That is, the ending inventory of remanufactured products y_t^+ is always zero in each period. In essence, the firm carries inventory of remanufactured products only in situations where there are large demand spikes (i.e., D_t is large for some t) coupled with the inability of meeting the exact demand in each period due to capacity constraints.

One other point to make is that the formulation above assumes away any uncertainty in demand or returns; in practice both are likely to occur. There are two ways to deal with this problem. First, the firm may carry some safety stock of remanufactured products to protect against uncertainty. For example, one can multiply all values of D_t by $(1 + z \cdot \sigma)$, where z is a safety factor, and σ is the standard deviation of the ratio actual demand-to-forecast (A_t/D_t). This is clearly a heuristic, but one that is easy to implement and use in practice. Second, the firm solves the problem (Eqs 4–8) in a *rolling horizon* basis: start at the beginning of the planning horizon (period 1), solve the problem, and implement the first-period solution (z_{i1}, v_{i1}). At the end of period 1, after *actual* demand A_1 and return quantities are realized, the firm updates the inventory positions and demand forecasts, and re-solves the problem for the planning horizon 2, 3, ..., $T + 1$; this process continues each period. By using a rolling horizon approach coupled with some level of safety stock, the firm protects itself against forecast error. This is a *heuristic* solution to the problem with demand uncertainty; an optimal solution would require the formulation of Equations (4–8) as a stochastic program, where demand and return scenarios, with their associated probabilities, are

considered in the problem (see *Modeling Uncertainty in Optimization Problems*).

CONCLUSION

This article provides an introduction to applications of operations research to remanufacturing. We first introduce basic terminology, such as the different types of returns, different types of disposition decisions available in CLSCs, and the basic differences between remanufacturing and regular manufacturing. In essence, the basic input to remanufacturing operations—product returns or cores—are not homogeneous, as they arrive to the facility in different quality conditions, and at a timing and quantity that is more difficult to control than raw materials and parts in new item manufacturing. Second, remanufactured products are typically perceived as inferior to new products, even when sold with the same warranty and are thus priced at a lower price point; remanufactured products can thus be used to reach new customer segments, such as customers who wouldn't be able to afford the new product, however, remanufactured products can also cannibalize on the sales of new products (which typically have higher profit margins). We provide an example of how economic modeling and nonlinear programming can be used to answer the basic question: given a certain cost structure, and level of market acceptance for remanufactured products, should a firm offer a remanufactured product in addition to its new product? If so, what are the respective prices that maximize profitability? Further references on this issue include Refs 14–16.

An important strategic decision for a firm that engages or is considering engaging in remanufacturing is the overall design of a CLSC. Issues such as location of facilities (collection points, aggregation points, remanufacturing facilities, etc.) and optimization of flows in this network have been addressed through large-scale mathematical programming techniques, typically mixed-integer programming models [7–13].

At a more tactical level, the firm must implement a remanufacturing plan to meet demand for remanufactured products. First,

the firm must carefully plan the acquisition of product returns; this includes procuring an amount that takes into account the variable quality (and as a result, variable cost and yield) of returns. Research [17,18] has addressed this problem using variants of the classical newsvendor model (see **News vendor Models**), or economic modeling of supply and demand curves [4,17]. Given a forecast of returns (with their qualities) and demand for remanufactured products, the firm can use mathematical programming techniques (linear programming, stochastic programming) to provide the optimal production plan that minimizes the cost of meeting demand; we have provided a simple example of such problems but there are other approaches (see survey in Ref. 21).

Finally, at a more operational level, the firm must decide on the sequence and depth of disassembly of returns; these problems have been studied using mixed-integer linear programming models and graph theory (see survey in Ref. 23). It must also sequence the remanufacturing operations in the shop floor. Given the complexity of different shop configurations and bills of materials, these problems are more specific to each firm, and have been mostly addressed through simulation [22].

Given the introductory nature of this article, we have been able to only offer a glimpse of operations research applied to the study of CLSCs. We have provided several references for further study; many of these are reference books. For an overview of operations research and economic tools applied to CLSCs at a more advanced level, but still contained within 25 printed pages, the reader is referred to Ref. 24.

REFERENCES

1. Ferguson ME, Guide VDR Jr, Koca E, *et al.* The value of quality grading in remanufacturing. *Prod Oper Manage* 2009; 18(3):300–314.
2. Knox A. An overview of incineration and EFW technology as applied to the management of municipal solid waste (MSW). University of Western Ontario. Available at <http://www.oneia.ca/files/EFW%20-%20Knox.pdf>. 2005.
3. Fleischmann M, van Nunen J, Grave B. Integrating closed-loop supply chains and spare parts management at IBM. ERIM Report Series Reference No. ERS-2002-107-LIS. Available at <http://ssrn.com/abstract=371054>. 2002.
4. Guide VDR, Teunter R, van Wassenhove LN Jr. Matching demand and supply to maximize profits from remanufacturing. *Manuf Serv Oper Manage* 2003;5(4):303–316.
5. Hauser W, Lund R. The remanufacturing industry: anatomy of a giant. Boston (MA): Boston University, Department of Manufacturing Engineering Report; 2003.
6. Subramanian R, Subramanyam R. An empirical analysis of the market for remanufactured products: evidence from eBay. Working paper, College of Management, Georgia Institute of Technology.
7. Ammons JC, Realff MJ, Newton DJ. Decision models for reverse production system design. In: Madu CN, editor. *Handbook of environmentally conscious manufacturing*. Boston (MA): Kluwer Academic Publishers; 2001.
8. Fleischmann M, Beullens P, Bloemhof-Ruwaard J, *et al.* The impact of product recovery on logistics network design. *Prod Oper Manage* 2001;10(2):156–173.
9. Wojanowski R, Verter V, Boyaci T. Retail collection network design under deposit refund. *Comput Oper Res* 2007;34(2):324–345.
10. Sahyouni K, Savaskan C, Daskin M. A facility location model for bidirectional flows. *Transp Sci* 2007;41(4):484–499.
11. Bostel N, Dejax P, Lu Z. The design, planning and optimization of reverse logistics networks. In: Langevin A, Riopel D, editors. *Logistics systems: design and optimization*. Berlin: Springer; 2005.
12. Akcali E, Cetinkaya S. and Uster H., Network Design for Reverse and Closed-Loop Supply Chains: An Annotated Bibliography of Models and Solution Approaches. *Networks* 2009;53(3):231–248.
13. Pochampally KK, Nukala S, Gupta S. Strategic planning models for reverse and closed-loop supply chains. Boca Raton (FL): CRC Press; 2008.
14. Ferguson ME, Toktay B. The effect of competition on recovery strategies. *Prod Oper Manage* 2006;15:351–368.
15. Ferrer G, Swaminathan J. Managing new and remanufactured products. *Manage Sci* 2006;52(1):15–26.

16. Atasu A, Sarvary M, Wassenhove LV. Remanufacturing as a marketing strategy. *Manage Sci* 2008;54(10):1731–1746.
17. Galbreth M, Blackburn JD. Optimal acquisition and sorting policies for remanufacturing. *Prod Oper Manage* 2006;15:384–392.
18. Zikopoulos C, Tagaras G. Impact of uncertainty in the quality of returns on the profitability of a single-period refurbishing operation. *Eur J Oper Res* 2007;182:205–225.
19. Guide VDR Jr, van Wassenhove LN. Managing product returns for remanufacturing. *Prod Oper Manage* 2001;10(2):142–155.
20. Guide VDR Jr. Production planning and control for remanufacturing: Industry practice and research needs. *J Oper Manage* 2000;18:467–483.
21. Dekker R, Fleischmann M, Inderfurth K, *et al.*, editors. *Reverse logistics: quantitative models for closed-loop supply chains*. Berlin: Springer; 2004.
22. Guide VDR Jr, Srivastava R, Kraus M. Scheduling policies for remanufacturing. *Int J Prod Econ* 1997;48(2):187–204.
23. Lambert AJD, Gupta SM. *Disassembly modeling for assembly, maintenance, reuse, and recycling*. Boca Raton (FL): CRC Press; 2005.
24. Souza G. Closed-loop supply chains with remanufacturing. In: Chen ZL, Raghavan R, editors. *Tutorials in operations research*. Hanover (MD): Informs; 2008. pp. 130–153. Available at www.informs.org.

RENDEZVOUS SEARCH GAMES

STEVE ALPERN
Department of Mathematics,
London School of
Economics, London, UK

INTRODUCTION

Rendezvous search games arise in a very natural way from the general theory of search [1,2], if one simply assumes that the “hidden object” that the searcher wishes to find is another searcher who wants to be found. That is, rendezvous search is a common-interest game where both players want to minimize the expected time required for them to meet. These games may also be seen as search extensions of the one-shot coordination problems of Schelling [3], who asked where two parachutists should agree to meet after they are randomly scattered by the winds. The difference is that Schelling’s problem simply ends in failure if they do not meet at once, whereas in rendezvous search (as in all search problems) they continue to try to find each other until they meet. Schelling emphasized the importance of focal points in his problem, whereas in rendezvous search focal points are (usually) eliminated by symmetry. Rendezvous problems are faced quite commonly when individuals become separated, or when their specification of a meeting point is insufficient. They have been faced by convoys of ships at sea, scattered by a storm. They are faced by migratory birds dispersed during migration, and by parent penguins who go fishing and then have to find their offspring in large colonies. Computers solving problems in a parallel fashion have to use rendezvous protocols to regularly check the progress of the other processor. Rendezvous protocols when on dangerous patrols are discussed by McNab [4], and popular accounts have been given in the press [5].

The *rendezvous search problem* was first posed by Alpern [6] at the end of a 1976 talk on zero-sum (hide and seek) games, as

a cooperative contrast to those games (which are discussed in Shmuel Gal’s article *Search Games* in this volume). In the basic form of rendezvous search, two players are placed according to a known joint distribution (usually independently) in a known search region Q . They know their own location, but only up to some symmetry of Q . So, for example, if Q is a circle, they know nothing of use even about their own location (unless Q is labeled or directed). If Q is a disk, they would know their distance from the center. [These notions of spatial symmetry are discussed at length in Ref. 7 (and in Ref. 8) in terms of the corresponding subgroup of the symmetries of Q , but we shall treat symmetries informally here.] Each player moves with unit speed until the first time T when their distance is less than a prescribed “search radius,” which is taken to be zero when Q is a network. The minimum value of the expected meeting time $\hat{T}$ that the players can achieve is called the *rendezvous value* of Q , denoted $R = R(Q)$, and the corresponding strategies are called *optimal*. Note that $R(Q)$ depends crucially on which of certain symmetry assumptions are made, as discussed below.

There are two main forms of the problem determined by whether or not we can distinguish between the two players, a different kind of symmetry. In the *asymmetric* (or player-asymmetric) form of the game, the players can use distinct pure strategies. So, in asymmetric rendezvous one particular strategy pair that is always available is what we term wait for mommy (WFM), where Player I (Mommy) searches Q exhaustively to minimize the time taken to find a randomly placed stationary object, and Player II (Baby) simply waits. In fact, this is usually a good strategy (pair), but rarely an optimal one. If we cannot distinguish between the players and are (for example) writing a note in a hiker’s manual telling them what to do if they get separated, then we are restricted to suggesting a single strategy for both players to adopt. This is called the *symmetric* (or player-symmetric) form of the game. In this

form, optimal strategies are usually mixed (randomized). To see why pure strategies may not be useful, consider the case where Q is an infinite line and the players start a known fixed distance apart, and they are each initially faced in a random direction. If they happen to face the same direction (which has probability one-half) and they follow the same pure strategy, they will never meet. Hence, the expected meeting time of any doubly adopted pure strategy is infinite. The rendezvous values for the player-asymmetric and symmetric version are denoted, respectively, by $R^a(Q)$ and $R^s(Q)$.

The original formulation of the continuous rendezvous search problem, including a rigorous definition of the rendezvous value, is given in Alpern [8]. A survey of the field, covering most developments up to 2002, is given in Alpern [9]. A full treatment of rendezvous theory, considered as part of the field of search games, is given in the monograph of Alpern and Gal [10]. Here we will refer to the original publications.

This article is organized as follows. The section titled “Two Original Examples” describes the two original rendezvous problems posed in the 1976 talk, and subsequent results toward their solution. The section titled “Rendezvous on the Line” discusses rendezvous on the infinite line, when the players have information about the distribution of the distance, but not the direction, to their partner. The section titled “Uncertain Initial Conditions” deals with rendezvous on the line, for the case where the distribution of initial distance is not known. It seeks rendezvous strategies for which the expected meeting time is small with respect to the mean of the distance distribution. The section titled “Rendezvous in Higher Dimensions” covers rendezvous in higher dimensions. The section titled “Miscellaneous Rendezvous Problems” deals with other versions of rendezvous, including a brief description of how rendezvous problems are studied in the theoretical computer science literature.

TWO ORIGINAL EXAMPLES

To illustrate the flavor of specific rendezvous problems, we describe briefly the two

problems given in the original presentation of rendezvous search [6]. These were posed as perhaps a good starting point for the theory, as they were simple and presumably easy to solve. Surprisingly, neither is fully solved to this day.

In this section, we present the two problems and then give a summary of the work that has been done on them. Later sections will present the bulk of the literature on rendezvous theory, mostly in areas not foreseen in the original talk (some of these areas were suggested in Ref. 8).

Highly Symmetrical Regions

The first problem seeks to avoid the focal points of Schelling [3] by choosing the search region Q with maximal symmetry (a transitive symmetry group, so *no* distinguished points).

Problem 1 [Astronaut Problem]. *Two astronauts land independently and uniformly on large, smooth spherical body Q . The body does not have a fixed orientation in space, nor does it have an axis of rotation, so that no common notion of position or direction is available to the astronauts for coordination. Given unit walking speeds for both astronauts, how should they move about so as to minimize the expected meeting time $\hat{T}$ (before they come within the detection radius)?*

This two-dimensional (sphere) problem is still unsolved. The one-dimensional analog, the circle Q , has been extensively studied (see Table 1). There are six versions, depending on the assumptions we make regarding player symmetry and spatial symmetry. The three types of spatial symmetry are (i) a labeled circle, say a clockface; (ii) a directed circle; (iii) an undirected circle. The four strategies suggested for these in the initial paper of Alpern [8] where such symmetry notions were defined formally in terms of group theory are as follows: FOCAL (go to a specified point like 3 o'clock); OP-DIR (the two players go in opposite directions); COHATO (coin half tour: repeatedly go half way around, choosing direction equiprobably); WFM (the wait-for-mommy strategy discussed above). Of course,

Table 1. Conjectured Best Strategies for Circle (Circumference 1)

Players	Labeled Circle	Directed Circle	Undirected Circle
Symmetric	FOCAL, $\hat{T} = \frac{1}{3}$	COHATO, $\hat{T} = \frac{3}{4}$	COHATO, $\hat{T} = \frac{3}{4}$
Asymmetric	OP-DIR, $\hat{T} = \frac{1}{4}$	OP-DIR, $\hat{T} = \frac{1}{4}$	WFM, $\hat{T} = \frac{1}{2}$

not all of these are feasible for all versions. The conjectured optimal strategies for the six versions are given below, followed by their corresponding expected meeting times $\hat{T}$. It is perhaps an easy but useful exercise for the reader to check these times.

We note that, although it is an open question as to whether COHATO is indeed optimal in the two versions for which it is suggested, it has been shown that in the other four cases the suggested strategy is indeed optimal. For a class of initial distributions which includes the uniform, it was shown by Alpern [11] that OP-DIR is optimal for asymmetric rendezvous on the directed circle, extending the similar result of Howard [12] for the uniform distribution. Howard obtained this result by first establishing that OP-DIR is optimal for the labeled circle and then used the following argument. First note that rendezvous values (minimum of $\hat{T}$) can only increase (weakly) as we go to the right in the table, as the players have less information. Since OP-DIR is optimal for the labeled circle and feasible for the directed circle, it must also be optimal for the directed circle. Howard [12] also established that WFM is optimal for asymmetric rendezvous on the undirected circle. The optimality of the FOCAL strategy for symmetric rendezvous on the labeled circle was established by Alpern [13]. As a matter of curiosity, we note that COHATO is optimal for both players in the zero-sum search game on the circle (spatial symmetry is irrelevant in this context), for the case where one of the players wants to *maximize* T . So some might find this conjecture strange.

Rendezvous on Discrete Locations

Rendezvous games are mainly about continuous movement of players within a region but, as with similar games, discrete approximations are often very interesting. The second

problem posed in the original talk (sometimes known as the *telephone coordination problem*, or *rendezvous on the complete graph K_n*), can be stated as follows.

Problem 2 [Mozart Café Problem]. *Two friends agree to meet for lunch at the Mozart Café in Vienna on the 1st of January 2000. However, on arriving at Vienna airport, they are told there are three (or n) cafés with that name, no two close enough to visit in the same day. So, every day each can go to one of them, hoping to find his friend there. This is a player-symmetric problem, so what is the best randomized strategy for both to adopt?*

For general n , this problem is still open. When $n = 2$ it is not too difficult to show that the players cannot do better than choose a random café each day [14,15]. The Crawford–Haller result [15] is about strategy coordination with revealed information about previously chosen strategies, but for $n = 2$ the other player’s strategy is revealed as being the other café (unless you meet), so the two formulations are equivalent. Anderson and Weber [14] showed that in the player-asymmetric version, the optimal strategy pair is WFM, with say friend I searching all n cafés in random order, and friend II picking a random one and going there every day. Anderson and Weber introduced an important class of strategies (which we call AW_n) for the symmetric problem: Assuming you do not meet on the first day, do the following in every group of $n - 1$ days thereafter. With probability p_n , visit the remaining $n - 1$ cafés in random order; with probability $1 - p_n$, stay where you are (they calculated the optimal values of p_n). It has been conjectured that AW_n is indeed optimal. After some small progress [16,17] toward showing that AW_n has certain approximate optimality properties, in 2006 (after 30 years!) Richard Weber [18] used

semidefinite programming techniques to make the breakthrough of establishing that AW_3 is indeed optimal for $n = 3$, one of the outstanding results in rendezvous theory. Subsequently his PhD student J. Fan [19] has partially extended his techniques to related problems, including cases where the cafés are so large that the friends might not meet even if they go to the same one (i.e., there is a positive *overlooking probability*).

RENDEZVOUS ON THE LINE

The general formulation of rendezvous games given in Ref. 8 assumes that the search region Q is compact, but in fact the region that has been most studied is the infinite line. The two players are initially placed a distance d apart, and faced in random directions. The distance d may be fixed or chosen randomly from a given cumulative distribution function F . The player-asymmetric version has been solved for the cases where F is a point distribution or is convex on a bounded domain $[0, D]$. On the other hand, the player-symmetric version is still unsolved, and the progress that has been made in bounding the rendezvous value R^s in the case where d is known rests largely on numerical techniques.

Player-Asymmetric Version

In this version, a strategy for say Player I is a function $f(t)$, which represents his net forward motion at time t , where forward is the direction he is initially facing. Thus, f must satisfy the conditions $f(0) = 0$ and $|f(t) - f(t')| \leq |t - t'|$ (unit speed condition). It was shown by Alpern and Gal [20] that the optimal strategy pair for known d is a variation on WFM, which we call modified WFM. If player I plays Mommy, her optimal strategy is simply to go at unit speed a distance d in one direction (say forward) and then $2d$ in the other, finding Baby (player II) equiprobably at times d or $3d$, with mean meeting time $\hat{T} = 2d$. Suppose Baby (player II) does not just wait, but instead plays his optimal response. Clearly, this is to move in some direction for time $d/2$, hoping to meet Mommy at time $d/2$ (with probability $1/4$). If this fails, the best he can do is return to

his starting point at time d (again hoping Mommy will arrive there at time d , probability $1/4$). Failing this, he knows that Mommy went in the direction away from him and is now at distance $2d$ from him. So he moves in some direction (say forward) for the next time period of length d (meeting at time $2d$ with probability $1/4$). Finally, he returns to his start at time $3d$ where he will meet Mommy at last (unconditional probability $1/4$). This is the modified WFM strategy (pair), which has $\hat{T} = (1/4)(d/2 + d + 2d + 3d) = \frac{13}{8}d = R^a$.

It was shown by Alpern and Howard [21] that this problem is equivalent to the following *alternating search* problem (not a rendezvous problem): There are two searchers, each at the center of an interval of length d . There is an object hidden equiprobably at the four ends of the two intervals. The two searchers can move *one at a time* (hence the term *alternating*). What strategy for the searchers minimizes the expected time to find the object? Clearly, the searchers first alternate in going distance $d/2$ to one of the ends on their line, and then alternate in going distance d to the other end. Thus the same set of equiprobable times occurs, $\{d/2, d, 2d, 3d\}$. An easy way to solve more complicated problems of this type, without using dynamic programming, is given in Ref. 21. This alternating search theory was subsequently adapted by Alpern and Beck [22] to prove that the modified WFM strategy (taking steps which are multiples of $D/2$, where D is the maximum of d) is optimal when the distribution of d is convex on $[0, D]$, with $R^a = (4\mu + 9D)/8$. In particular, for the uniform distribution of d on $[0, D]$, $R^a = 11/8$.

A natural question to ask is whether waiting (staying still on a nontrivial time interval) is ever optimal. This was answered in the negative by Gal [23], who showed that there is no distribution F for which WFM is an optimal strategy. Alpern and Beck [22] strengthened this to show that in any optimal strategy (for any F) there is a nontrivial time interval on which both players are moving with (the maximum) speed 1. Note, however, that these negative results for waiting hold only for the line, as Howard [12] has shown that waiting may be optimal on the circle (as mentioned above). These

negative results for the line also require the assumption that both players have the same maximum speed. If Player I has maximum speed 1 but Player II has lower maximum speed $v < 1$, then it was shown by Alpern and Gal [24] That, for fixed $d = 1$, if v is less than the golden mean $\frac{\sqrt{5}-1}{2} \doteq .61803$, then Player II should stay still for time 1, and then move at maximum speed in a certain way, and the rendezvous value is $(v^2 + 4v + 2) / (v + 1)^2$. If v exceeds the golden mean, then a modification of the optimal strategy for $v = 1$ given above is optimal, and the rendezvous value is $(4v^2 + 7v + 2) / (v + 1)^3$.

Player-Symmetric Version

As noted above, the symmetric version of rendezvous on the line requires mixed (or behavioral) strategies. This problem was originally proposed in Ref. 8, where the following strategy 1F2B (one forwards, two backwards) was suggested (where μ and D denote the mean and maximum of the initial distance d): If the maximum distance is D , do the following in every time interval of length $3D/2$: (i) pick a random direction to call forward; (ii) go forward for time $D/2$; (iii) go backward for time D . Note that, if the players pick the same forwards direction, they will be at their initial distance apart again at time $3D/2$; otherwise, they will have met. The expected meeting time $\hat{T}$ for 1F2B satisfies the equation

$$\hat{T} = \frac{1}{4}(\mu/2) + \frac{1}{4}(D + \mu/2) + \frac{1}{2}(3D/2 + \hat{T}),$$

or

$$\hat{T} = \frac{1}{2}\mu + 2D.$$

In particular, for an initial distance of $d = 1$ (where $D = \mu = 1$), this gives the bound $R^s \leq 2.5$. This estimate has been improved a number of times, by Anderson and Essegaiier [25], Baston [26], Uthaisombut [27], Alpern [28], and Han *et al.* [29]. The last group have the best published estimates of R^s (normalizing to $d = 1$ as usual) for the line, of $2.076 \leq R^s \leq 2.1287$, which they obtained using absorbing Markov chain theory, fractional quadratic programming, and semidefinite programming.

It is an open question whether giving the players a common sense of direction on the

infinite line helps them at all in the player-symmetric version of rendezvous. It is conjectured that it does not help. (It certainly does help in the player-asymmetric version, as it reduces the problem to the celebrated linear search problem; see Ref. 20.)

The Labeled Interval

Rendezvous on the line has also been studied in the version where the points are labeled, for the case of the unit interval $Q = [0, 1]$. This project was initiated by Howard [12]. Players are initially placed independently according to a common distribution F . Howard showed that, if F has a nondecreasing density, then an optimal strategy for the asymmetric (and hence for the symmetric version) is for the players to go right until they meet the *right sweeper* (an imaginary agent whose location at time t is given by $1 - t$) and then to stick with him. In the case of the uniform distribution, another optimal strategy is simply to go to the middle of the interval and wait. Howard's results for the interval have been extended by Chester and Tutuncu [30] and Alpern [13]. Note that, in this setting, a strategy for a player is of the form $f_x(t)$: that is, it tells him where to be at every time t if his initial position is x . Howard's strategy given above has the property that, if $f_x(t) = f_{x'}(t)$ for some t , then $f_x(s) = f_{x'}(s)$ for all $s \geq t$. We call such strategies *sticky*. Alpern [13] studied analogous rendezvous problems on labeled networks and showed that for the symmetric version there is always an optimal strategy which is sticky, and also gave an example of a network for which there is a unique optimal strategy pair for the asymmetric problem, and it is not sticky. It is worth observing that, for symmetric rendezvous problems on *labeled* networks (including the labeled interval), pure strategies are sufficient for optimality.

Other Rendezvous Problems on the Line

There are several other versions of the rendezvous problem that have mainly been attacked when the search region is the line. We describe them briefly here.

Bounded Resources. Alpern and Beck [31,32] proposed the problem where the players have a limited amount of fuel (or stamina), so that Player I can travel a total distance (formally, *total variation*) of a and Player II a total distance of b , with $a \geq b$. The asymmetry of their resources means that we consider the player-asymmetric version and, as usual, they are placed a fixed distance d apart and faced in random directions. Of course, if $b = 0$, this is just the well known *linear search problem* of Bellman [33] and Beck [34]. The condition under which they can ensure a meeting is that $5a + 3b \geq 15d$ and in this case [31] when $d = 1$, the rendezvous value (least expected meeting time) $R_{a,b}$ is given by

$$R_{a,b} = \begin{cases} (31 - 5a - 3b)/8 & \text{if } 2a + b \leq 6 \\ (19 - a - b)/8 & \text{if } 2a + b \geq 6 \text{ and } a \leq 3 \\ (16 - b)/8 & \text{if } b \leq 3 \leq a \\ 13/8 & \text{if } 3 \leq b. \end{cases}$$

On the other hand, if $5a + 3b \geq 15d$, they cannot guarantee a meeting, so the best they can do is to maximize the probability of a meeting. For the case where $d = 1$ is known, this maximum probability $P_{a,b}$ is given by Alpern and Beck [32]

$$P_{a,b} = \begin{cases} 0 & \text{if } a + b < 1 \\ 1/4 & \text{if } a + b/3 < 1 \leq a + b \\ 1/2 & \text{if } a/3 + b/3 < 1 \leq a + b/3 \\ 3/4 & \text{if } a/3 + b/5 < 1 \leq a/3 + b/3 \\ 1 & \text{if } 1 \leq a/3 + b/5. \end{cases}$$

Similar results are given when the initial distance distribution is uniform, triangular, and negative exponential.

Multirendezvous. Thus far we have been considering the rendezvous of *two* players on the line. We now consider the problem where n players are placed randomly at consecutive integer locations on the line, faced in random directions, and wish to minimize the time $T_{n,m}$ taken for m of them to be at the same location. We call these problems $\Gamma_{n,m}^a$ and $\Gamma_{n,m}^s$, with respective rendezvous values $R_{n,m}^a$ and $R_{n,m}^s$ for the player-asymmetric and player-symmetric versions. Our previous

results concern the special case $n = m = 2$. The minimax time for all n of them to meet will be denoted M_n , considered in the player-asymmetric version. Note that, when $m > 2$, we must analyze information transfer between two players who meet.

The first of these problems to be studied was $\Gamma_{3,2}^a$. Optimal strategies move in time intervals of $1/2$ between adjacent half integers ($k/2$), and such strategies can be specified by indicating the integers k for which a player changes direction at time $k/2$. We give this list in vector form, so $x_i = 1$ if the player changes direction at time $i/2$ and $x_i = 0$ otherwise. So, for example, the vector $[1, 1, 1, \dots]$ indicates a strategy which always changes direction. Using a branch and bound method, the game $\Gamma_{3,2}^a$ was solved by Lim *et al.* [35], who showed that $R_{3,2}^a = 47/48$. Up to permutations of the labels of the players, there are six optimal strategy profiles, which we give in the $[]$ notation below:

$$\begin{aligned} & \{[0, 1, 0, 0], [1, 0, 0, 1], [1, 0, 1, 0]\}, \\ & \{[0, 1, 0, 0], [1, 0, 0, 0], [1, 0, 1, 0]\}, \\ & \{[0, 1, 0, 0], [1, 0, 0, 0], [1, 1, 0, 1]\}, \\ & \{[1, 1, 0, 1], [1, 0, 0, 0], [1, 1, 0, 1]\}, \\ & \{[0, 1, 0, 0], [1, 0, 0, 0], [1, 1, 1, 1]\}, \\ & \{[0, 1, 0, 0], [1, 0, 0, 1], [1, 1, 1, 1]\}. \end{aligned}$$

For large values of $m = n$, both rendezvous values are asymptotic to $n/2$, as shown by the following estimates of [35]:

$$\frac{n}{2} \leq R_{n,n}^a \leq R_{n,n}^s \leq \frac{n}{2} + 5.$$

Since $m = n > 2$, players will sometimes meet with another before the end of the game. We call the part of the game before a player's first meeting *Stage I* and the part after this meeting *Stage II*. The strategy that gives the upper bound is defined as follows (the word *start* refers to a player's initial position):

- *Rule 1 (overriding everything else).* If someone you meet says "follow me," then follow him (adopt same path).
- *Stage I.* Pick a random direction, independently and equiprobably each time.

Go a distance $1/2$ in that direction and then return to start. Then go a distance $1/2$ in the opposite direction and return to start. This takes total time 2, and you are back at start at integer times. Repeat until you are back at start and have met another player. (This return to start can occur at time 1 or at time 2.) Then proceed to Stage II.

- *Stage II.* Go at speed 1 a maximum distance 1 in the direction away from that in which you met another player earlier (in Stage I). If you meet another player in this time (after time $1/2$ or time 1 from beginning of the stage), then return immediately to your start and wait. If after time 1 you have not met another player, then proceed to Stage III.
- *Stage III.* Proceed at unit speed toward and then past your start, continuing until the game ends, instructing everyone you meet to follow you. (If two players in Stage III meet, then their groups together contain all the players, so the game is over.)

A related result of Alpern [9] for the multirendezvous of n players initially placed at unit distances along a circle of circumference n shows that the expected meeting time $R_{n,n}^s(C)$ has the upper bound $R_{n,n}^s(C) \leq 2^{n-2}/(2^{n-1}-1) + n/2$. To break symmetry, the associated strategy involves each player picking a random number in $[0, 1]$ at the outset, and using it to determine which player is the leader when two meet.

The minimax rendezvous problem, that of evaluating minimax time M_n for all the players to meet, was introduced by Lim and Alpern [36] (in the player-asymmetric form—in the symmetric form no maximum time can be guaranteed). The case $n = 2$ is the usual rendezvous problem in minimax form, and the expected time minimizing strategy (modified WFM), with maximum time 3, is also minimax. Simple WFM is also minimax. Thus $M_2 = 3$. The minimax strategy found by Baston [26] for $n = 3$ is more complicated. It establishes (together with an estimate from Ref. 36) that $M_3 = 3.5$. Baston's strategy was shown to be the unique optimal strategy by Alpern and Lim

[37]. Lim and Alpern [36] showed that, for large numbers of players, the minimax value M_n is asymptotic to $n/2$, which is the full information meeting time. Gal [23] established the upper bound

$$M_n \leq n/2 + 2 \log_2 n + 3.5.$$

An interesting aspect of Baston's strategy is that, sometimes when two players meet, they adopt different future paths (split up). If players who meet must stay together (sticky rendezvous), we denote the minimax time by $\tilde{M}_n$. It turns out that this is indeed a severe restriction, as it is shown in Ref. 36 that $\tilde{M}_3 = 5$. This contrasts with Alpern's result [13] that, for symmetric rendezvous on labeled networks, there is always an optimal (expected time minimizing) strategy which is sticky.

UNCERTAIN INITIAL CONDITIONS

In this section we deal with the question of how the rendezvousers on the line should proceed when there is uncertainty about the distribution function F of their initial distance. This question was vaguely suggested as one of the seven *questions for further work* in Ref. 8 as "Adversary-rendezvous games: Suppose the two searchers' initial position in Q is chosen by another player who wishes to maximize their meeting time." The question was first attacked by Baston and Gal [38], who presented a number of strategies which work well against arbitrary initial distributions F , in the sense that the expected meeting time against F is small with respect to the mean of F . One of their estimates was improved by Alpern and Beck [39].

Asymmetric Rendezvous

The first result in Baston and Gal [38] concerns the asymmetric rendezvous problem on the line. They showed that there exists an asymmetric pair of pure strategies such that the expected meeting time against any distribution F is at most 5.73μ , where μ is the mean of F .

Subsequently, Alpern and Beck [39] considered the following *alternating* strategy

(the players alternate reversing directions): In each time period n both players move in a fixed direction at maximum speed 1. At the end of each period, exactly one of the rendezvousers reverses direction, and they reverse in alternating periods. There is no first period; periods go $\dots, -3, -2, -1, 0, 1, 2, \dots$. Each period has length exactly $\hat{a}$ times the length of the previous period, where

$$\hat{a} = \frac{1}{3} + \frac{2^{2/3}}{3(8 + \sqrt{62})^{1/3}} + \frac{2^{1/3}(8 + \sqrt{62})^{1/3}}{3} \approx 1.5994.$$

If this strategy pair is used against any distribution F , the expected meeting time is at most $V\mu$, where

$$V = \frac{1}{2} + (1/8) \left(\frac{8}{\hat{a} - 1} + 7 + 5\hat{a} + 3\hat{a}^2 + \hat{a}^3 \right) \approx 5.514,$$

where μ is the mean of F . This result is sharp in that no other alternating strategy does better.

Note that, against a known distribution F , there is no need for the players in an asymmetric problem to randomize—pure strategies are sufficient. However, against an unknown F , the players are essentially playing a zero-sum game against Nature (who chooses F). In this case, it is well known that randomization may help. In the asymmetric rendezvous context, they may pick their *pair* of strategies randomly, which means their individual strategies will be correlated. The best estimate for this version of the problem is given in Ref. 38, Theorem 4.1, where the authors exhibit a probability distribution over pairs of distinct rendezvous strategies, so that if this “correlated” strategy is used against any distribution F , the expected meeting time is at most 4.42μ .

Symmetric Rendezvous

We now consider the scenario where the two players are constrained to use the same mixed strategy. The first result for this case is [38, Theorem 5.1], that there

exists a mixed search strategy $\bar{v}$ which, if adopted with independent randomization by the two players against any distribution F , gives an expected meeting time at most $(7 + 2\sqrt{10})\mu \approx 13.33\mu$.

The rendezvousers can improve this bound by using joint randomization. Here, S, S^* , and $(S^*)^*$, respectively, denote pure strategies, mixed strategies, and probability distributions over mixed strategies. Baston and Gal showed for this case that there exists a probability distribution $\bar{v}$ over mixed search strategies which has the following property: If a mixed strategy is chosen according to the distribution $\bar{v}$ and then adopted by both players with independent randomization, then the expected meeting time against any initial distance distribution $F \in \mathcal{F}$ is at most 11.4μ , where μ is the mean of F .

RENDEZVOUS IN HIGHER DIMENSIONS

Although one of the original problems in rendezvous theory (the “Astronaut Problem,” Problem 1 of the section titled “Highly Symmetrical Regions”) involves a two-dimensional search region, most of the theory has been developed for one-dimensional regions. Rendezvous on the plane was first attacked by Thomas and Hulme [40], who simulated the problem of a loud, fast helicopter seeking a slow, quiet walker. Their simulations indicate that (p. 48)

... there may be an advantage in role reversal. The helicopter should try to drag the walker toward it. Thus it should search less and make its presence audible more and eventually stay within a small area waiting to the walker to rendezvous with it

Planar Rendezvous, Diagonal Start

The first formal model for player-asymmetric planar rendezvous was developed by Anderson and Fekete [41]. They first considered the assumption that the players had a common sense of direction (each has a compass). In this case, as shown earlier [20] for the line, the problem can be reduced to a single-player search problem: minimize the expected time to find a stationary target

hidden uniformly on a circle of known radius. However, they then considered a true rendezvous problem by assuming no common sense of direction (no compasses). They take the search region as the integer grid on the plane ($\mathbb{Z} \times \mathbb{Z}$), with their initial vector difference taken equiprobably from the set $\{(1, 1), (1, -1), (-1, -1), (-1, 1)\}$: that is, they are on opposite vertices of a unit square. (We call this diagonal start.) Each player is faced in a random direction, which he calls N (North). They prove that the optimal (expected time minimizing) strategy pair is the *A-F strategy* $[(N, W, S, S, E, E, N, N), (N, S, N, S, N, S, N, S)]$. Note that A-F can be seen as a modified WFM strategy, with Player I doing a time minimizing exhaustive search (with meeting times 2, 4, 6, 8) and Player II at his start at these times but moving away at odd times in the hope of decreasing the meeting time by 1. Alpern and Baston [42] showed that, if there is a common sense of clockwise, there is another distinct optimal strategy pair, namely the *A-B strategy* $[(N, E, S, W, S, W, N, E), (N, S, E, S, W, N, E, N)]$. They also showed that, if there is no common sense of clockwise, then all optimal strategies are *generalized A-F strategies* (as defined below) and these strategies are also uniformly optimal (maximize probability of meeting by time t for all t).

Definition 1 [Generalized A-F Strategies]. A strategy pair is called *generalized A-F* if one player goes around the square $(\pm 1, \pm 1)$ (with respect to his start) in a cyclic fashion, while the other is back at his start at even times and away at odd times.

Parallel Start, $n \geq 2$

The Anderson-Fekete model [41] can be extended to n dimensions by initially placing the players on even nodes (those with even sum of coordinates) of the integer lattice $\mathbb{Z}^n$. In particular, Alpern and Baston [43] considered the case where the difference between the players is of length 2 and parallel to one of the coordinate axes. Let R_n denote the rendezvous value, assuming no common

sense of direction, and even *no common sense of clockwise*. They show that $R_2 = 197/32$ (extending the earlier result of Alpern and Gal [20] that $R_1 = 13/4$). However, unlike the diagonal start problem of Anderson and Fekete, none of the optimal strategies is uniformly optimal. Furthermore, for general n they show that

$$R_n \leq \frac{32n^3 + 12n^2 - 2n - 3}{12n^2} \rightarrow \frac{8}{3}.$$

In a related article, Alpern and Baston [44] showed that when the players *do have a common sense of clockwise*, it helps them, reducing the rendezvous time from $197/32 \approx 6.1563$ to $194/32 \approx 6.0625$. Also, unlike the lack of uniform optimality in the no-common-clockwise version [44], there exist uniformly optimal strategies.

Multiagent versions of planar rendezvous have been analysed by Lin *et al.* [45, 46]. They give strategies (which may or may not depend on a common clock) that ensure that the players will eventually all meet.

A related paper of Alpern [47] considered the problem faced by players who want to rendezvous in Manhattan and can see each other when they are on a common street or avenue. More work is needed on this “line of sight” version of rendezvous for other regions.

MISCELLANEOUS RENDEZVOUS PROBLEMS

In this section we discuss some other variations on the basic rendezvous problem.

Rendezvous-Evasion Games

We now consider a zero-sum win-lose game proposed by the author, in which the common aim of two rendezvouchers R_1 and R_2 is to meet strictly before either of them encounters an enemy searcher S (who has the opposite aim). This game models the situation, for example, where a rescuer R_2 is sent to find a downed mobile pilot R_1 before the enemy on the ground S finds him. The search region is the complete graph K_n , and the three agents start at distinct nodes at time 1. The value v_n of the game is therefore the probability of the searcher winning, with best play on both sides.

Asymmetric Rendezvous-Evasion on K_n . Lim [48] showed that $v_3 = 47/76$. The optimal strategies are described as follows. Each agent P labels his initial position as $P(1)$, his next *new* positions as $P(2)$, $P(3)$, and so on. When going to a new position, he can only move randomly (equiprobably) among the locations not already visited. The optimal mixed strategy for the rendezvous team is as follows: At Step 1, R_1 moves to location $R_1(2)$, while R_2 stays at location $R_2(1)$. Before they separate, they jointly choose $i = 1, 2, 3$ with probabilities $(7/19, 6/19, 6/19)$ and then: if $i = 1$, R_1 visits location $R_1(3)$ (uniquely determined), R_2 stays at $R_2(1)$; if $i = 2$, R_1 returns to $R_1(1)$; R_2 visits $R_2(2)$; if $i = 3$, R_1 stays at $R_1(2)$; R_2 visits $R_2(2)$. At this point, they have a common labeling of the locations and hence meet equiprobably at one of them at the next step. The searcher begins by moving to a new location $S(2)$. Then, he picks $j = 1, 2, 3$ with probabilities $(5/19, 5/19, 9/19)$ and goes to location $S(j)$. For steps $k > 3$, he visits each location with probability $1/3$.

Lim also gives a strategy for the searcher which guarantees him a win on K_4 with probability $31/54$. She further shows that, when all the agents have a common circular ordering of the locations (but may move between any two), this knowledge helps the rendezvouchers and the value is $\left((1 - 2/n)^{n-1} + 1\right)/2$.

Symmetric Rendezvous-Evasion on K_3 . In the symmetric version of rendezvous-evasion studied by Lim and Alpern [36], the two rendezvouchers must be given the same mixed strategy to follow. Unlike the asymmetric version in the previous section, the rendezvouchers cannot be certain to have met by a given time, so we require the rendezvous team to meet within a given time bound, namely by step 3.

We assume that all agents have a common notion of clockwise direction around the three nodes, which we view as vertices of a triangle. This assumption allows the agents to label their initial location as 0, and the remaining two as 1 and 2 in a clockwise direction from 0. In this notation, there are

nine pure strategies:

$$\{(0, 0), (1, 2), (2, 1), (0, 1), (1, 0), (2, 2), (0, 2), (1, 1), (2, 0)\}.$$

We will be concerned with four distributions over these strategies, which we call behavioral strategies.

$$b^1 = \frac{1}{3} (1, 1, 1, 0, 0, 0, 0, 0, 0),$$

$$b^2 = \frac{1}{3} (0, 0, 0, 1, 1, 1, 0, 0, 0),$$

$$b^3 = \frac{1}{3} (0, 0, 0, 0, 0, 0, 1, 1, 1),$$

$$b^4 = \frac{1}{9} (1, 1, 1, 1, 1, 1, 1, 1, 1).$$

Alpern and Lim [49] showed that the value of the common-direction, two-step, symmetric rendezvous-evasion game on K_3 is $57/81$. The optimal strategy for the rendezvouchers is for each of them to use (the same) behavioral strategies b^i , $i = 1, 2, 3$, equiprobably. The optimal strategy for the searcher is to use the random strategy b^4 .

The interpretation of the rendezvous strategy is as follows: The controller broadcasts a number $1, 2, 3$, equiprobably, which only the rendezvouchers receive. When number i is broadcast (and received), both rendezvouchers adopt the behavioral strategy b^i . The need for mixtures over behavioral strategies is unusual in game theory. Such strategies were introduced in Ref. 50.

Alpern and Lim [49] also found that, in the version with no common notion of direction, the value lies in the interval $[58/81, 59/81]$.

Uncertain Motives of “hider”

Sometimes when a person is lost, it is uncertain to the searcher whether or not the lost person wants to be found. For example, a missing teenager might be a “runaway”; a fictional Russian submarine in *The Hunt for Red October* by Clancy [51] might be lost or might be defecting. If p denotes the probability that the lost agent wants to be found, we obtain a class of games that connects a zero sum search game (in the sense of Gal’s article **Allocation Games** in this volume) to a rendezvous problem. Alpern and Gal [24] solved problems of this type when the search

region is a tree or the infinite line, and Gal and Howard [52] considered discrete locations. An opportunity for further study is the situation where *both* agents have unknown motives.

Information Transfer, Search Costs

In the basic form of rendezvous search, the players receive no additional information (other than that they have not yet met) until they meet. This assumption has been relaxed by Baston and Gal [53], who allowed each player to leave a mark at his starting point which the other player could see if he arrived there. They considered the complete graph and the line as search spaces. Baston and Kikuta (unpublished) have also worked in this area. Alpern [28] supposed that rendezvousers on the line get some information at certain stages in the game.

Another assumption of rendezvous is that search, for example, at a discrete location, is costless. This is unusual in search theory. The effect of including such costs has been studied by Kikuta and Ruckle [54].

Theoretical Computer Science

This volume is concerned with Operations Research, but we feel we should at least give a brief mention to a very vibrant area of rendezvous research in the theoretical computer science literature. See, for example, Refs 55–59.

REFERENCES

1. Stone LD. Theory of Optimal Search. 2nd ed. Arlington (VA): Operations Research Society of America; 1989.
2. Benkoski S, Monticino M, Weisinger J. A survey of the search theory literature. *Nav Res Logist* 1991;38:469–494.
3. Schelling T. The strategy of conflict. Cambridge: Harvard University Press; 1960.
4. McNab A. Bravo two zero. London: Corgi Books; 1994.
5. Matthews R. How to find your children when they are lost. *Sunday Telegraph*. Science 1995 Mar 26.
6. Alpern S. Hide and seek games. Vienna: Seminar, Institut für Höhere Studien; 1976.
7. Alpern S. The rendezvous search problem. LSE CDAM Research Report No. 53. London: London School of Economics; 1993.
8. Alpern S. The rendezvous search problem. *SIAM J Control Optim* 1995;33:673–683.
9. Alpern S. Rendezvous search: a personal perspective. *Oper Res* 2002;50(5):772–795.
10. Alpern S, Gal S. The theory of search games and rendezvous. International series on operations research and management science. Boston: Kluwer; 2003.
11. Alpern S. Asymmetric rendezvous search on the circle. *Dyn Control* 2000;10:33–45.
12. Howard JV. Rendezvous search on the interval and circle. *Oper Res* 1999;47(4):550–558.
13. Alpern S. Rendezvous search on labelled networks. *Nav Res Logist* 2002;49:256–274.
14. Anderson EJ, Weber R. The rendezvous problem on discrete locations. *J Appl Probab* 1990;27:839–851.
15. Crawford VP, Haller H. Learning how to cooperate: optimal play in repeated coordination games. *Econometrica* 1990;58(3):571–596.
16. Alpern S, Baston B, Essegai S. Rendezvous search on a graph. *J Appl Probab* 1999;36(1):223–231.
17. Alpern S, Pikounis M. The telephone coordination game. *Game Theor Appl* 2000;5:1–10.
18. Weber R. The optimal strategy for symmetric rendezvous search on K_3 . Cambridge University; 2006. In press. arXiv:0906.5447v1 [math.OC], 2009.
19. Fan J. Symmetric rendezvous problem with overlooking [PhD dissertation]. Cambridge University; 2009.
20. Alpern S, Gal S. Rendezvous search on the line with distinguishable players. *SIAM J Control Optim* 1995;33:1270–1276.
21. Alpern S, Howard JV. Alternating search at two locations. *Dyn Control* 2000;10:319–339.
22. Alpern S, Beck A. Asymmetric rendezvous on the line is a double linear search problem. *Math Oper Res* 1999;24(3):604–618.
23. Gal S. Rendezvous search on the line. *Oper Res* 1999;47:974–976.
24. Alpern S, Gal S. Search for an agent who may or may not want to be found. *Oper Res* 2002;50(2):311–323.
25. Anderson EJ, Essegai S. Rendezvous search on the line with indistinguishable players. *SIAM J Control Optim* 1995;33:1637–1642.
26. Baston VJ. Two rendezvous search problems on the line. *Nav Res Logist* 1999;46:335–340.

27. Uthaisombut P. Symmetric rendezvous search on the line using moving patterns with different lengths. University of Pittsburgh: Department of Computer Science; 2006.
28. Alpern S. Rendezvous search with revealed information: applications to the line. *J Appl Probab* 2007;44(1):1–15.
29. Han Q, Du D, Vera J, *et al.* Improved bounds for the symmetric rendezvous value on the line. *Oper Res* 2008;56(3):772–782.
30. Chester EJ, Tutuncu RH. Rendezvous search on the labelled line. *Oper Res* 2004;52(2):330–334.
31. Alpern S, Beck A. Rendezvous search on the line with bounded resources: expected time minimization. *Eur J Oper Res* 1997;101:588–597.
32. Alpern S, Beck A. Rendezvous search on the line with bounded resources: maximizing the probability of meeting. *Oper Res* 1999;47(6):849–861.
33. Bellman R. An optimal search problem. *SIAM Rev* 1963;5:274.
34. Beck A. On the linear search problem. *Isr J Math* 1964;2:221–228.
35. Lim WS, Alpern S, Beck A. Rendezvous search on the line with more than two players. *Oper Res* 1997;45(3):357–364.
36. Lim WS, Alpern S. Minimax rendezvous search on the line. *SIAM J Control Optim* 1996;34:1650–1665.
37. Alpern S, Lim WS. Rendezvous of three agents on the line. *Nav Res Logist* 2002;49:244–255.
38. Baston VJ, Gal S. Rendezvous on the line when the players' initial distance is given by an unknown probability distribution. *SIAM J Control Optim* 1998;36(6):1880–1889.
39. Alpern S, Beck A. Pure strategy asymmetric rendezvous on the line with an unknown initial distance. *Oper Res* 2000;48(3):1–4.
40. Thomas LC, Hulme PB. Searching for targets who want to be found. *J Oper Res Soc* 1997;48(1):44–50.
41. Anderson EJ, Fekete S. Two-dimensional rendezvous search. *Oper Res* 2001;49:107–188.
42. Alpern S, Baston V. Rendezvous on a planar lattice. *Oper Res* 2005;53(6):996–1006.
43. Alpern S, Baston V. Rendezvous in higher dimensions. *SIAM J Control Optim* 2006;44(6):2233–2252.
44. Alpern S, Baston V. A common notion of clockwise can help in planar rendezvous. *Eur J Oper Res* 2006;175(2):688–706.
45. Lin J, Morse AS, Anderson BDO. The multi-agent rendezvous problem. Part 1: the synchronous case. *SIAM J Control Optim* 2007;46(6):2096–2119.
46. Lin J, Morse AS, Anderson BDO. The multi-agent rendezvous problem. Part 2: the asynchronous case. *SIAM J Control Optim* 2007;46(6):2120–2147.
47. Alpern S. Line-of-sight rendezvous. *Eur J Oper Res* 2008;188(3):865–883.
48. Lim WS. A rendezvous-evasion game on discrete locations with joint randomization. *Adv Appl Probab* 1997;29:1004–1017.
49. Alpern S, Lim WS. The symmetric rendezvous-evasion game. *SIAM J Control Opt* 1998;36(3):948–959.
50. Alpern S. Games with repeated decisions. *SIAM J Control Opt* 1988;26(2):468–477.
51. Clancy T. *The hunt for Red October*. New York: Harper Collins; 1995.
52. Gal S, Howard JV. Rendezvous-evasion search in two boxes. *Oper Res* 2005;53(4):689–697.
53. Baston VJ, Gal S. Rendezvous search when marks are left at the starting point. *Nav Res Logist* 2001;48:722–731.
54. Kikuta K, Ruckle W. Rendezvous search on a star graph with examination costs. *Eur J Oper Res* 2007;181(1):298–304.
55. Dessmark A, Fraigniaud P, Kowalski D, *et al.* Deterministic rendezvous in graphs. *Algorithmica* 2006;46:69–96.
56. Kowalski D, Malinoski A. How to meet in anonymous network. *Theor Comput Sci* 2008;399:141–156.
57. Dobrev S, Flocchini P, Prencipe G, *et al.* Multiple agents rendezvous in a ring in spite of a black hole. Volume 3144/2004, *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2004 pp. 34–46.
58. De Marco G, Gargano L, Kranakis D, *et al.* Asynchronous deterministic rendezvous in graphs. *Theor Comput Sci* 2006;355:315–326.
59. Kranakis E, Krizanc D, Rajsbaum S. Mobile agent rendezvous: a survey. Volume 4056/2006, *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2006.

RENEWAL FUNCTION AND RENEWAL-TYPE EQUATIONS

LIQIANG LIU
EURANDOM, Eindhoven University of
Technology, Eindhoven, Netherlands

THE RENEWAL FUNCTION

Consider a renewal process $\{N(t), t \geq 0\}$ where $N(t)$ counts the number of occurrences of a certain event up to time t . The points in time distinguished by the event, called *renewal epochs*, are denoted, chronologically, by random variables $S_1, S_2, \dots$. The difference between any two successive members of this sequence is called *interrenewal time*. By the definition of renewal process, the interrenewal times are mutually independent and have a common distribution F . Throughout this article we assume the renewal process is pure, that is, the distribution of S_1 is independent and also follows F .

Definition 1 [Renewal Function]. The renewal function $M(\cdot)$ is defined as

$$M(t) = \mathbb{E}(N(t)), \quad t \geq 0. \quad (1)$$

The renewal function $M(\cdot)$ can be directly computed from its definition as follows:

$$M(t) = \sum_{n=1}^{\infty} n \mathbb{P}(N(t) = n) = \sum_{n=1}^{\infty} \mathbb{P}(N(t) \geq n).$$

Notice the following equivalence of events

$$N(t) \geq n, \text{ if and only if } S_n \leq t, \quad n = 1, 2, \dots$$

Now define $F^{*n}(t) = \mathbb{P}(S_n \leq t)$, $n = 1, 2, \dots$. Thus, F^{*n} is the distribution of the sum of n independent and identically distributed

random variables (the first n interrenewal times). Then immediately we get

$$M(t) = \sum_{n=1}^{\infty} F^{*n}(t). \quad (2)$$

For this consideration, Equation (2) is an equivalent definition of the renewal function. The convergence of the infinite sum that defines the renewal function is guaranteed [1, Theorem 8.8].

Theorem 1. *The renewal function $M(t)$ is finite for all $t \geq 0$.*

The renewal function plays a central role in renewal theory in the sense that a pure renewal process is completely determined by its renewal function. To see this, first take the Laplace–Stieljes transform (LST) on both sides of Equation (2). Since the renewal function is finite, we get

$$\tilde{M}(s) = \sum_{n=1}^{\infty} \tilde{F}^n(s) = \frac{\tilde{F}(s)}{1 - \tilde{F}(s)}, \quad (3)$$

or

$$\tilde{F}(s) = \frac{\tilde{M}(s)}{1 + \tilde{M}(s)},$$

where $\tilde{M}$ and $\tilde{F}$ are LSTs of M and F respectively. Owing to the uniqueness of the LST, F is completely determined, and so is the renewal process.

Example 1 [Poisson Process]. For a Poisson process with mean arrival rate λ , the interrenewal time distribution is exponential, that is, $F(t) = 1 - e^{-\lambda t}$ and $\tilde{F}(s) = \lambda/(s + \lambda)$. By Equation (3), $\tilde{M}(s) = \lambda/s$. Inverting this we get

$$M(t) = \lambda t, \quad t \geq 0.$$

In other words, a renewal process with a linear renewal function must be a Poisson process.

RENEWAL ARGUMENT AND RENEWAL EQUATION

The renewal argument is one of the most useful tools in renewal theory. The method involves two steps: conditioning on the epoch of the first renewal and then exploiting the regenerative feature of recurrent event. The renewal argument is used to derive an integral equation for certain probabilistic quantities in a renewal process. This section illustrates the renewal argument by deriving the *renewal equation*, which can also be used to compute the renewal function.

Let us compute $M(t)$ by conditioning on the first renewal epoch S_1 , namely,

$$M(t) = \int_0^\infty \mathbb{E}(N(t) | S_1 = x) dF(x).$$

If $x > t$, then the first renewal occurs after t and hence $N(t)$ is 0. If $x \leq t$, then $N(t)$ equals 1 (counting the one that occurs at time x) plus the number of renewals afterwards, up to time t . This is identical in distribution to $N(t - x)$, because we get a replica of the original renewal process after remarking the renewal epoch x as zero. So the integral above equals $\int_0^t \mathbb{E}(1 + N(t - x)) dF(x)$. Therefore,

$$M(t) = F(t) + \int_0^t M(t - x) dF(x). \quad (4)$$

Equation (4) is called a renewal equation. Notice that taking the LST on both sides yields $\tilde{M}(s) = \tilde{F}(s) + \tilde{M}(s)\tilde{F}(s)$, which gives Equation (3).

RENEWAL-TYPE EQUATIONS

This section introduces a class of integral equations that arise from the use of the renewal argument. *Renewal-type equations* are equations of the following form

$$H(t) = D(t) + \int_0^t H(t - u) dF(u),$$

where F is a distribution function of a nonnegative random variable that is almost surely bounded, D is given, and H is the unknown. The renewal equation such as

(4) is a special case of the renewal-type equations with $D = F$. The following theorem [1, Theorem 8.10] gives the conditions for the existence and uniqueness of the solution to a renewal-type equation.

Theorem 2. *Suppose*

$$|D(t)| < \infty \text{ for all } t \geq 0.$$

Then there exists a unique solution to the renewal-type equation

$$H(t) = D(t) + \int_0^t H(t - u) dF(u),$$

such that

$$|H(t)| < \infty \text{ for all } t \geq 0,$$

given by

$$H(t) = D(t) + \int_0^t D(t - u) dM(u),$$

where M is the renewal function as in Equation (2).

This theorem reveals the importance of the renewal function as a factor in the solution of the renewal-type equation. The importance of the renewal-type equation derives largely from a significant result in renewal theory, the *key renewal theorem* (see **Limit Theorems for Renewal Processes**).

In practice, if D has an LST $\tilde{D}$, then the transform version of the renewal-type equation,

$$\tilde{H}(s) = \tilde{D}(s) + \tilde{H}(s)\tilde{F}(s),$$

can be more useful.

In certain applications, one may be interested in an unknown F , where the distribution of interrenewal time, however, is difficult

to compute directly. In such situations, it can be helpful to find a renewal argument that establishes a renewal-type equation involving known H and D functions.

NOTES

The reader is encouraged to continue to article *Limit Theorems for Renewal Processes* on limit theorems, so that the notions of renewal function and renewal equation anchor to the grand picture of renewal theory. A classical and comprehensive reference for this topic is Feller [2, Chapter 11]. Other references include Karlin and Taylor [3], Resnick [4], and Ross [5].

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman & Hall/CRC; 1995.
2. Feller W. An introduction to probability theory and its applications. 3rd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley; 1966.
3. Karlin S, Taylor HM. A first course in stochastic processes. New York: Academic Press; 1975.
4. Resnick SI. Adventures in stochastic processes. Basel, Switzerland: Birkhauser Verlag; 1992.
5. Ross SM. Stochastic Processes. New York: John Wiley & Sons, Inc.; 1996.

RENEWAL PROCESSES WITH COSTS AND REWARDS

MARIA VLASIOU
Department of Mathematics
and Computer Science,
Eindhoven University of
Technology, Eindhoven,
The Netherlands

BASIC NOTIONS AND RESULTS

Many applications of renewal theory involve rewards or costs (which can be simply seen as a negative reward). For example, consider the classical example of a renewal process: a machine component gets replaced upon a failure or upon having operated for T time units. Then the time the n th component is in service is given by $Y_n = \min\{X_n, T\}$, where X_n is the life of the component. In this example, one might be interested in the rate of the number of replacements in the long run. An extension of this basic setup is as follows. A component that has failed will be replaced at a cost c_f , while a component that is replaced while still being operational (and thus at time T) costs only $c < c_f$. In this case, one might be interested in choosing the optimal time T that minimizes the long-run operational costs. The solution to this problem involves the analysis of *renewal processes with costs and rewards*.

Motivated by the above, let $\{N(t), t \geq 0\}$ be a renewal process with interarrival times X_n , $n \geq 1$, and denote the time of the n th renewal by $S_n = X_1 + \dots + X_n$. Now suppose that at the time of each renewal a reward is received; we denote by R_n the reward received at the end of the n th cycle. We further assume that (R_n, X_n) is a sequence of i.i.d. random variables, which allows for R_n to depend on X_n . For example, $N(t)$ might count the number of rides a taxi gets up to time t . In this case, X_n is the length of each trip, and one reasonably expects the fare R_n to depend on X_n . As usual, we denote by

(R, X) the generic bivariate random variable that is distributed identically to the sequence of rewards and interarrival times (R_n, X_n) . In the analysis, together with the standard assumption in renewal processes that the interarrival times have a finite expectation $\mathbb{E}[X] = \tau$, we further assume that $\mathbb{E}[|R|] < \infty$. The cumulative reward up to time t is given by $R(t) = \sum_{n=1}^{N(t)} R_n$, where the sum is taken to be equal to zero in the event that $N(t) = 0$. Depending on whether the reward is collected at the beginning or at the end of the renewal period, one might wish to adapt this definition by taking $R(t) = \sum_{n=1}^{N(t)+1} R_n$.

The importance of renewal processes with costs and rewards is evidenced by their application to Markovian processes and models. Most queueing processes in which customers arrive according to a renewal process are regenerative processes (see **Regenerative Processes**) with cycles beginning, e.g., each time an arrival finds the system empty. Moreover, for every regenerative process $Y(t)$ we may define a reward structure as follows: $R_n = \int_{S_{n-1}}^{S_n} Y(t) dt$, where all R_n , $n \geq 1$, are i.i.d., except possibly for R_1 that might follow a different distribution. The basic tools that are used are the computation of the reward per unit of time and the rate of the expected value of the reward. These two results are summarized in the following theorem.

Theorem 1. *Let $N(t)$ be a renewal reward process generated by (R, X) , $n \geq 1$. Assume that $\mathbb{E}[|R|] < \infty$ and let $r = \mathbb{E}[R]$, $\tau = \mathbb{E}[X] < \infty$. Then*

$$\lim_{t \rightarrow \infty} \frac{R(t)}{t} = \frac{r}{\tau}, \text{ with probability 1, and} \quad (1)$$

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}[R(t)]}{t} = \frac{r}{\tau}. \quad (2)$$

These results do not depend on whether rewards are collected at the beginning of the renewal cycle or at its end, or if they are collected in a more complicated manner (e.g., continuously or in a nonmonotone way

during the cycle), as long as (R_n, X_n) , $n \geq 1$ is a sequence of i.i.d. bivariate random variables and minor regularity conditions hold; see the following section. Moreover, the results remain valid also for delayed renewal reward processes, that is, when the first cycle might follow a different distribution.

Note that Equation (2) does not follow from Equation (1), since almost sure convergence does not imply the convergence of the expected values. Think of the distribution

$$\mathbb{P}[Y_n = x] = \begin{cases} n/(n+1), & x = 0, \\ 1/(n+1), & x = n+1. \end{cases}$$

Since for all n , $\mathbb{E}[Y_n] = 1$, we have $\lim_{n \rightarrow \infty} \mathbb{E}[Y_n] = 1$. Moreover, Y_n goes almost surely to zero, that is, $\mathbb{P}[\lim_{n \rightarrow \infty} Y_n = 0] = 1$, which means that the value of the random variable and its expectation differ in the limit.

Proof. To prove Equation (1), write

$$\frac{R(t)}{t} = \frac{\sum_{n=1}^{N(t)} R_n}{N(t)} \frac{N(t)}{t}.$$

By the strong law of large numbers we have that

$$\lim_{t \rightarrow \infty} \frac{\sum_{n=1}^{N(t)} R_n}{N(t)} = \mathbb{E}[R] = r,$$

since $N(t)$ goes to infinity almost surely, and by the strong law for renewal processes we know that

$$\lim_{t \rightarrow \infty} \frac{N(t)}{t} = \frac{1}{\mathbb{E}[X]} = \frac{1}{\tau}.$$

The proof of Equation (2) is a bit more involved. The classical proof applies Wald's equation [1] and requires the proof that $\lim_{t \rightarrow \infty} \frac{\mathbb{E}[R_{N(t)+1}]}{t} = 0$, by constructing a renewal-type equation (see **Renewal Function and Renewal-Type Equations**). Here, we present an alternative proof that follows the same steps of the classical proof for the Elementary Renewal Theorem (See e.g.[2]) but seems to be new in the setting of renewal reward processes.

We first construct a lower bound. From Fatou's lemma, we have that

$$\begin{aligned} \liminf_{t \rightarrow \infty} \frac{\mathbb{E}[R(t)]}{t} &\geq \mathbb{E} \left[\liminf_{t \rightarrow \infty} \frac{R(t)}{t} \right] \\ &= \mathbb{E} \left[\lim_{t \rightarrow \infty} \frac{R(t)}{t} \right] = \frac{r}{\tau} \end{aligned} \quad (3)$$

from Equation (1).

For the upper bound, we use truncation. Take $M < \infty$ and set $R_n^M = \max\{R_n, -M\}$. Note that

$$R(t) = \sum_{n=1}^{N(t)} R_n \leq \sum_{n=1}^{N(t)} R_n^M \stackrel{\text{def}}{=} R^M(t).$$

We then have that

$$\begin{aligned} \frac{\mathbb{E}[R(t)]}{t} &\leq \frac{\mathbb{E}[R^M(t)]}{t} \\ &= \frac{\mathbb{E} \left[\sum_{n=1}^{N(t)+1} R_n^M - R_{N(t)+1}^M \right]}{t} \\ &= \frac{\mathbb{E}[N(t)+1] \mathbb{E}[R_1^M]}{t} - \frac{\mathbb{E}[R_{N(t)+1}^M]}{t} \\ &\leq \frac{\mathbb{E}[N(t)+1]}{t} \mathbb{E}[R_1^M] + \frac{M}{t}. \end{aligned}$$

The last equality follows from Wald's equation since $N(t)+1$ is a stopping time, and for the last inequality we simply observe that we can bound $-\mathbb{E}[R_n^M]$ from above with M (by construction). Thus,

$$\begin{aligned} \limsup_{t \rightarrow \infty} \frac{\mathbb{E}[R(t)]}{t} &\leq \limsup_{t \rightarrow \infty} \left(\frac{\mathbb{E}[N(t)+1]}{t} \mathbb{E}[R_1^M] + \frac{M}{t} \right) \\ &= \frac{\mathbb{E}[R_1^M]}{\tau} \end{aligned}$$

by the Elementary Renewal Theorem. Since we can choose any $M < \infty$ in our construction, by monotone convergence we have that $\mathbb{E}[R_1^M] \rightarrow \mathbb{E}[R_1] = r$ as $M \rightarrow \infty$.

The advantage of this proof is that it avoids the computation of $\lim_{t \rightarrow \infty} \frac{\mathbb{E}[R_{N(t)+1}]}{t}$, and that it is, in essentials, identical to the classical proof for the corresponding theorem in renewal processes without rewards.

PARTIAL REWARDS

In some applications, rewards might not be given at the beginning or at the end of a cycle, but might be earned in a continuous (maybe nonmonotone) fashion during a cycle. The simplest case is when the reward during a cycle is given at a constant rate during the cycle, or only a portion thereof. The portion of the reward up to time t associated with a cycle starting at $X_{N(t)}$ is called a *partial reward*. Recall that we have assumed that $E[X] < \infty$ and $E[|R|] < \infty$. Under further assumptions, Theorem 1 still holds. In the case of partial rewards, we need to assume some form of good behavior as nothing so far prevents the rewards from escaping to plus or minus infinity while still having the average reward per cycle being finite. As an example, to avoid such erratic behavior, consider the restrictive assumption that the absolute value of rewards is uniformly bounded by some number; then Theorem 1 holds. This is also the case when rewards are nonnegative or if they accumulate in a monotone manner over any interval. Wolff [3] assumes that the partial reward during a cycle is the difference of two different monotone nonnegative processes, which could be interpreted as income and costs over the renewal cycle.

It seems to be a standard misconception that only having $E[|R|] < \infty$ is sufficient for Theorem 1 to hold also for partial rewards. In lack of necessary and sufficient conditions in the literature for this result, or at least of a counterexample where $E[|R|] < \infty$, but where Theorem 1 does not hold for partial rewards, we present here a counterexample that illustrates the needs for additional assumptions. In this case, the average reward over a cycle is finite (and we construct it to be equal to zero), but the average reward for every finite t is not defined, and thus also not its limit over t ; see Equation (2).

Take a renewal reward process that is defined as follows: $X_i = 1$ and $R_i = 0$ for all i . During a cycle, rewards build up and are depleted in the following manner. For cycle j , choose a Cauchy random variable C_j and take the auxiliary function $f(x) = x$ for $x \in [0, 1/2]$ and $f(x) = 1 - x$ for $x \in [1/2, 1]$. Define the reward for any time t during cycle j as

$C(t) = C_j \cdot f(t - [t])$. Then we see that the cumulative reward up to time t , for all $t > 0$, is given by $R(t) = C(t)$. Thus, since a Cauchy random variable does not have a well-defined first moment, we see that the average cumulative reward in this example cannot be defined, and thus Theorem 1 cannot be extended for this case, despite the fact that at the end of a cycle (and thus at all integer times) the cumulative reward is equal to zero.

In the following, we show two examples where partial rewards are assumed.

EXAMPLES

We proceed with a few applications of renewal processes with costs and rewards that exhibit the usefulness of these processes. The first two show how (cyclic) renewal processes can be seen as a special case of renewal processes with costs and rewards, while the last two make use of rewards earned continuously over time. Naturally, an abundance of examples of renewal (reward) processes might be found in Markov processes and chains.

Example 1 [Standard renewal processes]. Observe that the standard renewal process is simply a special case of a renewal reward process, where we assume that the reward at each cycle is equal to 1. In this case, we see, for example, that Equation (2) is simply the Elementary Renewal Theorem.

Example 2 [Alternating renewal processes]. For an alternating renewal process (see *Alternating Renewal Processes*), that is, a renewal process that can be in one of two phases (say ON and OFF), suppose that we earn at the end of a cycle an amount equal to the time the system was ON during that cycle. Alternatively, one may assume that we earn at rate one per unit of time when the system is ON, but one need not consider continuous rewards at this point. Then, the total reward earned in the interval $[0, t]$ is equal to the total ON time in that interval, and thus

by Equation (1) we have, as $t \rightarrow \infty$,

$$\frac{\text{ON time in } [0, t]}{t} \rightarrow \frac{\mathbb{E}[Y_{\text{ON}}]}{\mathbb{E}[Y_{\text{ON}}] + \mathbb{E}[Y_{\text{OFF}}]},$$

where Y_{ON} is a generic ON time in a cycle and equivalently for the OFF times. Thus, for nonlattice distributions, the limiting probability of the system being ON is equal to the long-run proportion of time it is ON. Naturally, these results extend to general cyclic renewal processes, that is, a process with more than two phases [4].

Example 3 [Average time of age and excess]. Let $A(t)$ denote the age at time t of a renewal process generated by $X_n \sim X$, and suppose that we are interested in computing the average value of the age. With a slight abuse of terminology, define this to be equal to

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t A(s) ds.$$

Suppose now that we are obtaining a reward continuously at a rate equal to the age of the renewal process at that time. Thus, $\int_0^t A(s) ds$ represents the total earnings by time t . Moreover, since the age of a renewal process at time s since the last renewal is simply equal to s , we have that the reward during a renewal cycle of length X is equal to $\int_0^X s ds = X^2/2$.

Then by Equation (1) we have that with probability 1,

$$\lim_{t \rightarrow \infty} \frac{\int_0^t A(s) ds}{t} = \frac{\mathbb{E}[X^2/2]}{\mathbb{E}[X]}.$$

We can apply the same logic to the residual life at time t , $B(t)$, and we will end up at the same result. Since the renewal epoch covering time t is equal to $X_{N(t)+1} = A(t) + B(t)$, we have that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t X_{N(t)+1} ds = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} \geq \mathbb{E}[X],$$

where we have equality only when $\text{Var}(X) = 0$. Thus, we retrieve the inspection paradox.

Example 4 [Little's law]. Consider a GI/GI/1 stable queue with the interarrival times X_i generating a nonlattice renewal process. Suppose that we start observing the system upon the arrival of a customer. Denote a generic interarrival time by X and let $\mathbb{E}[X] = 1/\lambda$. Also denote by $n(t)$ the number of customers in the system at time t . Consider the average number of customers in the system in the long run, which we denote by L . Then, with a slight abuse of notation we have that

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t n(y) dy.$$

Define a renewal reward process with cycles C starting each time an arrival finds the system empty, and suppose that we earn a reward at time y with a rate $n(y)$. Then from Equation (2) we have that

$$\begin{aligned} L &= \frac{\mathbb{E}[\text{reward during a cycle}]}{\mathbb{E}[\text{cycle time}]} \\ &= \frac{\mathbb{E}\left[\int_0^C n(y) dy\right]}{\mathbb{E}[C]}. \end{aligned} \quad (4)$$

Now let N be the number of customers served during a cycle, and define the long-run average sojourn time of customers as $T = \lim_{n \rightarrow \infty} (T_1 + \dots + T_n)/n$, where T_i is the time customer i spent in the system. Take now a cycle of length C where N customers were served. We can see then that the reward we have received in that cycle, which was defined to be equal to a rate of $n(y)$ at each time y , can also be seen as having each customer pay at rate 1 for each time unit he/she is in the system. In other words, in a cycle with N customers, the reward is equal to $T_1 + \dots + T_N$, that is, the total time all customers in that cycle spent in the system. Then T is the average reward per unit time (observe that since the second definition of the reward depends on the number of customers in a cycle, the duration of the cycle is also defined in terms of customers), and

again by Equation (2) we have

$$\begin{aligned} T &= \frac{\mathbb{E}[\text{reward during a cycle}]}{\mathbb{E}[\text{cycle time}]} \\ &= \frac{\mathbb{E}\left[\sum_1^N T_i\right]}{\mathbb{E}[N]}. \end{aligned} \quad (5)$$

We can now prove Little's law, one of the fundamental relations in queueing theory, which states that $L = \lambda T$. To see this, observe that

$$C = \sum_1^N X_i,$$

and since N is a stopping time for the sequence $\{X_i\}$, we have from Wald's equation

$$\mathbb{E}[C] = \mathbb{E}[N]\mathbb{E}[X] = \mathbb{E}[N]/\lambda.$$

Thus, by Equations (4) and (5), we see that

$$L = \lambda T \frac{\mathbb{E}\left[\int_0^C n(y) dy\right]}{\mathbb{E}\left[\sum_1^N T_i\right]}.$$

However, as we have seen, the fraction is equal to 1, since both the numerator and the denominator describe the reward earned during a cycle. Thus, we retrieve Little's law.

FURTHER READING

Almost all books on stochastic processes and introductory probability have a section on renewal theory. Here we only refer to the ones that have a separate mention of renewal processes with costs and rewards. Cox [5] is one of the few manuscripts exclusively devoted to renewal processes, where renewal processes with costs and rewards are found under the term *cumulative processes*. Another term that has been used to describe these processes is *compound renewal processes*. The general concept of a cumulative process and the asymptotic results related to them are based on Smith [6,7]. Mercer and Smith [8] investigate cumulative processes modulated by a Poisson process, in connection with a study of the wear of conveyor

belting. Brief mentions of renewal reward processes can also be found in Taylor and Karlin [9], as well as in Heyman and Sobel [10] in a contribution by Serfozo. In Moder and Elmaghraby [11], one may find a short treatment of cumulative processes, which also includes asymptotic results not only for the expectation of $R(t)$ as given by Equation (2) but also for its variance. As noted in their text, there are multivariate versions of these results as well. The following books offer a thorough treatment of this topic and a multitude of examples; moreover, they are approachable to a wide audience with only undergraduate knowledge of probability theory [3,12–16].

REFERENCES

1. Wald A. On cumulative sums of random variables. *Ann Math Stat* 1944;15:283–296.
2. Asmussen S. *Applied probability and queues*. New York: Springer; 2003.
3. Wolff RW. *Stochastic modeling and the theory of queues*. Prentice hall international series in industrial and systems engineering. Englewood Cliffs (NJ): Prentice Hall Inc.; 1989.
4. Serfozo R. *Basics of applied stochastic processes*. Probability and its applications. Berlin: Springer; 2009.
5. Cox DR. *Renewal theory*. London: Methuen & Co. Ltd.; 1962.
6. Smith WL. Regenerative stochastic processes. *Proc Roy Soc Lon Ser A Math Phys Eng Sci* 1955;232:6–31.
7. Smith WL. Renewal theory and its ramifications. *J Roy Stat Soc Ser B Met* 1958;20:243–302.
8. Mercer A, Smith CS. A random walk in which the steps occur randomly in time. *Biometrika* 1959;46:30–35.
9. Taylor HM, Karlin S. *An introduction to stochastic modeling*. 3rd ed. San Diego (CA): Academic Press Inc.; 1998.
10. Heyman DP, Sobel MJ, editors. *Stochastic models*. Volume 2, *Handbooks in operations research and management science*. Amsterdam: North-Holland Publishing Company; 1990.
11. Moder JJ, Elmaghraby SE, editors. *Handbook of operations research: foundations and fundamentals*. New York: Van Nostrand Reinhold Company; 1978.

12. Kulkarni VG. Modeling and analysis of stochastic systems. Texts in statistical science series. London: Chapman and Hall Ltd.; 1995.
13. Medhi J. Stochastic processes. 2nd ed. New York: John Wiley & Sons Inc.; 1994.
14. Resnick S. Adventures in stochastic processes. Boston (MA): Birkhäuser Boston Inc.; 1992.
15. Ross SM. Stochastic processes. 2nd ed. Wiley series in probability and statistics: probability and statistics. New York: John Wiley & Sons Inc.; 1996.
16. Tijms HC. Stochastic modelling and analysis. A computational approach. Wiley series in probability and mathematical statistics: applied probability and statistics. Chichester: John Wiley & Sons Ltd.; 1986.

RENT AND RENT LOSS IN THE ICELANDIC COD FISHERY

RAGNAR ARNASON
Department of Economics,
University of Iceland,
Reykjavik, Iceland

In a complete and smoothly operating market system, the net economic benefits of any production activity are measured by its profits [1,2]. In a natural resource-based industry such as fisheries, a part of the profits stem from the natural productivity of the underlying resource. This part may be referred to as natural resource rent or, in the case of the fishery, fisheries rent [3,4].

There is an important difference between fisheries rent and fisheries profit. Fisheries rent constitute that part of variable profits (profits without subtracting fixed costs), which is shared by all fishers. In this sense, the fisheries rents may be seen as a contribution of nature rather than individual ability. Above and beyond the rents, however, the more efficient fishers receive additional variable profits, sometimes referred to as intramarginal profits [5].

A simple example serves to illustrate these concepts. Imagine a fishery of three fishers, each of whom catches one unit of fish. Let the variable profits of the first be US\$10, those of the second US\$9, and those of the third US\$8. Then total profits in this fishery are US\$27 ($10 + 9 + 8$). The rents, however, are US\$24 ($8 \cdot 3$). The intramarginal profits are US\$3 ($2 + 1$).

In well-developed fisheries, intramarginal profits are usually quite small, so profits and rents are similar. The reason is that the knowledge, capital, and techniques generating superior efficiency tend, due to the forces of competition and market trade, to be quickly assimilated by all fishers. Therefore, in spite of technical progress, efficiency tends to be quite uniform across the fleet.

At each point of time, natural resource rents are bounded from above by the

productivity of the resource, the production technology, and input–output prices. These constraints define the maximum rents technically achievable. Obviously, the difference between the maximum rents attainable and the rents actually realized, the so-called rent loss, provides a measure of the economic efficiency of the resource use activity [4].

Ocean fisheries, since they are based on free-ranging ocean fish stocks, are extremely susceptible to the common property problem, that is, economic waste stemming from lack of private property rights in resources [6]. In the fishery, this problem leads to devastated fish stocks and a complete dissipation of all net economic benefits attainable from the fish resource [7]. As a result, the equilibrium state of common property fisheries is characterized by no profits and absence of any economic rents. This is confirmed by empirical observations. Most ocean fisheries in the world do not seem to be generating much net economic benefit or rent [4].

This article obtains a measure of the rent loss in the Icelandic cod fishery by estimating the actual resource rent generated and comparing it to the maximum attainable one. The fisheries rent (hereafter often referred simply to as *rent*) are calculated on the theoretical basis set out in Refs 4 and 5. The cod fishery is modeled in a simple, aggregate way. This is to keep the amount of data and modeling work within reasonable bounds. The construction of a more complete disaggregated model (with respect to substocks, cohorts, vessel classes, and fishing gear) of the Icelandic cod fishery is a major undertaking. Previous research employing both the disaggregate and aggregate approaches [8,9] indicates that the inaccuracies of employing a simple aggregative model for the Icelandic cod fishery rather than a more complete, disaggregated one are insignificant.

The Icelandic cod fishery is a particularly interesting case for a study of this nature. The reason is that this fishery was subjected to individual transferable quotas (ITQs) in stages beginning in 1984 [10]. In 1991, the

coverage of the ITQ system in the Icelandic cod fishery was substantially strengthened and expanded to cover all fishing vessels above 6 GRT (gross registered tons). In 2006, the system was further expanded to cover all commercial fishing vessels. Thus, from 1990 onwards, the Icelandic cod fishery has been managed for the most part on the basis of ITQs.

ITQs represent a certain fisheries management system under which the total allowable catch (TAC) each year is subdivided and allocated to fishers as their individual catch rights (quotas). Subsequently, the fishers may harvest according to these rights or transfer them to others. It is theoretically well-established [11,12] that ITQs lead to efficient fishing operations. First, the individual catch right greatly reduces the common property problem in the fishery. Second, the tradability of the rights allows the less efficient part of the fleet to leave the fishery with a monetary compensation and the TAC to be taken by fewer, more efficient vessels. Third, provided the TACs are set optimally over time so that the fish stocks can revert to their optimal levels—which for biological and ecosystem reasons may take a long time—the fishery will gradually become fully efficient. Obviously, this system promotes an overall gain in economic efficiency of the fishery and the generation of fisheries rent [11,12]. Thus, it is interesting to see to what extent the Icelandic cod fishery actually exhibits these theoretical gains.

In addition to its empirical relevance, this article provides an example of how the method of operations research can be used to derive useful results about a subject of considerable social importance. In this case, knowledge from economics and biology are combined with mathematical techniques to construct a reasonably tractable and theoretically consistent model of a fisheries situation. Numerical techniques are subsequently used to derive estimates of the fishery rent loss. Finally, statistical techniques are employed to assess the uncertainty about the true values and obtain estimates of confidence intervals. Needless to say, the methods outlined in this article can be used to perform similar calculations for other fisheries.

THE ICELANDIC COD FISHERY: BACKGROUND

The Icelandic cod fishery constitutes a sizable segment of the global ocean fishery. Annual catches have traditionally been in excess of 300 thousand metric tons (mt) annually. The unit value of landings is high (prices in 2007 were around US\$3 per kg.). Thus, the landed value of the fishery is potentially almost US\$1 billion. This may be compared to the estimated value of global marine capture fisheries landings of some US\$80 billion [13]. In the past, however, this fishery has been greatly overexploited. As a result, current stocks are quite depressed: estimates suggest that the fishable stock in 2007 was only about 30% of the virgin stock [14].

Cod fishing is pursued throughout the year, with a peak during the spawning season when large fish become highly catchable. It is conducted both, in inshore and off-shore waters by several types of vessels ranging in size from about 5 to 3000 GRT and employing different types of fishing gear. In terms of landed quantity, however, the most important part of the fleet are deep-sea bottom trawlers, typically 500–1500 GRT, many of which are freezer/factory trawlers.

The cod fishery is by far the most valuable component of the Icelandic demersal fisheries. These fisheries involve some 20 species of significance with cod haddock, saithe, redfish, and Greenland halibut (turbot) being the most important. In any given fishing trip, one of these species tends to be targeted. However, it is not common to haul in a clean, one-species harvest. In some fishing trips, especially the longer ones, a mix of two or more species is targeted.

As already mentioned, most vessels pursuing cod and other demersal fisheries were subjected to a form of ITQs in 1984. The ITQ system was substantially strengthened and expanded in 1991 although the smallest vessels (under 6 GRT) were still largely exempted from the ITQ restrictions. These small vessels were brought into the ITQ system in 2004.

Under the ITQ system, especially from 1991 onwards, various steps have been taken to restore the cod stock. Most important of

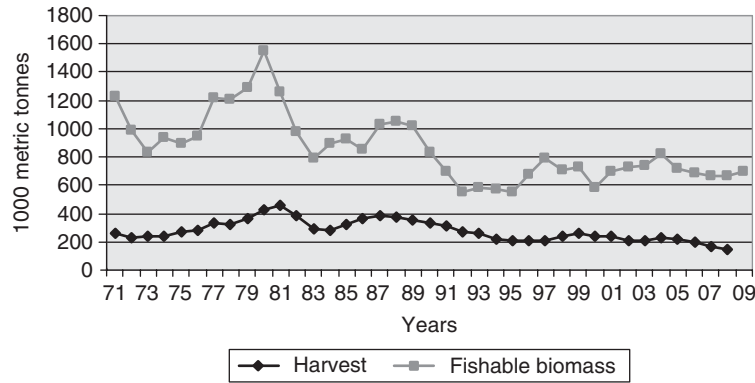


Figure 1. The cod fishery: fishable biomass and landings (1000 mt).

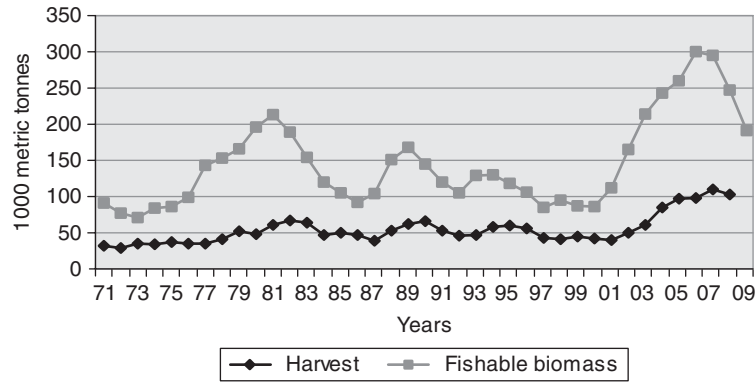


Figure 2. The haddock fishery: fishable biomass and landings (1000 mt).

these measures have been severe cut-backs in TACs. Other important measures have been extensive time and area closures. As can be gauged from Fig. 1, these measures have halted the decline in the stock, but have not managed to significantly increase it. Interestingly, similar measures for other demersal stocks, most notably the haddock stock, have produced much better results (Fig. 2).

Under the ITQ system, the size of the demersal fishing fleet as well as the demersal fishing effort (including the one on cod) has substantially contracted [10]. This as well as several other measures undertaken by the fishing firms, has greatly increased the economic performance of the fishery. Thus, productivity in the fishing industry has increased at a rapid rate. Similarly, the

profitability of the fishing firms has greatly improved. These developments have been reflected in a huge increase in the market value of the ITQs. These and further data regarding the economic and biological impact of the ITQs are recounted in Refs 10 and 15.

THE COD FISHERIES MODEL

To describe the cod fishery, the following basic fisheries model is adopted:

$$\dot{x} = G(x) - y \text{ (biomass growth function)} \quad (1)$$

$$y = Y(e, x) \quad \text{(harvesting function)} \quad (2)$$

$$\pi = p \cdot Y(e, x) - C(e) \text{ (profit function)} \quad (3)$$

$$p = P(y) \text{ (landings demand function).} \quad (4)$$

The five variables of this model, that is x, y, π, p , and e represent biomass, harvest, profits, landings price, and fishing effort, respectively. The first four are endogenous—determined within the model. The fifth, fishing effort, is exogenous in the sense of being a control variable for the fisheries operators. All variables depend on time, that is, they can change over time. The derivative, $\dot{x} \equiv \partial x / \partial t$ measures the change in biomass at a point of time. Note that parameters taken to be constants, such as various prices and technological coefficients, are not explicitly mentioned in the above presentation of the model.

In implementing the model, the following functional specifications, suggested by econometric investigations of the data [9,16], are employed.

$$x_{t+1} - x_t = G(x_t) = \alpha \cdot x_t - \beta \cdot x_t^2,$$

$$Y(e_t, x_t) = q \cdot e_t^a \cdot x_t^b,$$

$$C(e_t) = c \cdot e_t + fk,$$

$$P(y_t) = d \cdot y_t^f,$$

where $\alpha, \beta, q, a, b, c, fk, d$, and f are constants and t refers to time. Note that time is now measured in discrete years in accordance with the available data. The constant α is the intrinsic growth rate of the biomass, the ratio α/β , is the carrying capacity of the biomass, fk represents fixed costs, q catchability, and f is the elasticity of landings price with respect to the volume of landings.

Given these functional specifications, the complete model can be written in a condensed form as

$$x_{t+1} - x_t = \alpha \cdot x_t - \beta \cdot x_t^2 - y_t$$

(biomass growth function) (5)

$$\pi_t = p_t \cdot y_t - k \cdot \frac{y_t^\delta}{x_t^\varepsilon} - fk$$

(profit function) (6)

$$p_t = d \cdot y_t^f \quad (\text{price function}) \quad (7)$$

In this formulation fishing effort, e , has been substituted out. The endogenous variables

Table 1. Estimated Coefficients

Coefficients	Estimates	t -statistics
α	0.669853	8.6
β	$3.353 \cdot 10^{-4}$	-2.9
k	57604.95	6.5
δ	1.1	NA (nonlinear estimate)
ε	1.0	NA (restricted)
fk	10 billion	NA (from cost accounts)
d	220.0	14.8
f	0	NA (restricted)

Note: Economic values are in ISK. Exchange rate: 61 ISK per US\$.

are now x, π , and p . The exogenous (or control) variable is y . It is easy to check that the new constants k, δ , and ε are the following transformation of the original constants: $k = \frac{c}{q^{1/a}}, \delta = 1/a$ and $\varepsilon = b/a$.

Econometric and other estimation of the constants in the cod model defined by Equations (5), (6) and (7) led to the following results (Table 1) [16,17].

For the biomass function it is easy to verify that the maximum sustainable yield MSY and the stock carrying capacity XMAX, say, are given by the expressions:

$$\text{MSY} = \frac{\alpha^2}{4 \cdot \beta},$$

$$\text{XMAX} = \frac{\alpha}{\beta}.$$

It then follows from the estimates in Table 1 that the estimated MSY = 335 and the estimated XMAX = 1998 thousand mt.

Figure 3 provides a summary description of this fishery. The figure is drawn in the space of (fishable) biomass and landings (harvest) and applies at each point of time and, therefore, also in equilibrium. The parabolic graph represents the biomass growth function. As can be seen, it covers biomass from zero to the carrying capacity of almost 2 million mt and has an MSY of about 335 thousand mt. If, for any biomass level, landings lie on this curve, a biological equilibrium prevails. The other curves in this diagram are variable iso-profit curves, that is, loci of biomass and harvests which represent constant variable profits (measured in cod volume units). These functions are obtained by selecting a fixed level of

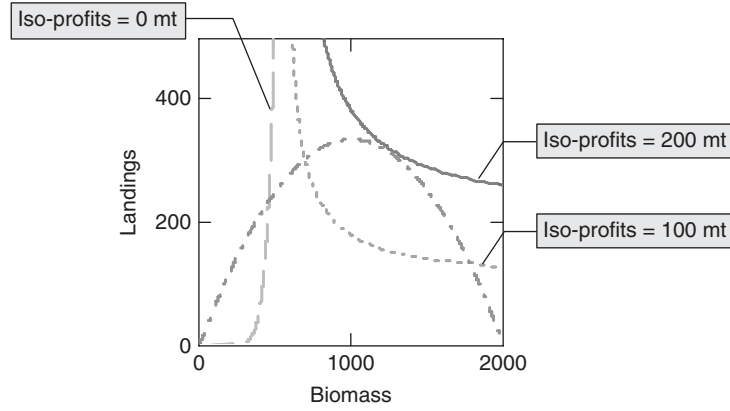


Figure 3. The estimated fisheries model (biomass growth and iso-profits in (biomass, harvest)-space).

Table 2. Rents and Rent Drain: Main Results

	Current (2005)	Maximum	Difference
Biomass (1000 mt)	715	1241	+526
Harvest (1000 mt)	215	315	+100
Effort (index)	1.0	0.86	-0.14
Total profits (million US\$)	126	546	+420
Variable profits (million US\$)	290	710	+420
Rents (million US\$)	241	667	+426

Note: Financial amounts in million US\$ per year.

profits in Equation (6) and tracing out the corresponding biomass, harvest coordinates. Any point where these iso-profit curves intersect the biomass growth curve represents a sustainable fishery with the corresponding variable profits. Thus, the traditional, zero-profit bioeconomic equilibrium is found where the zero-isoprofit curve intersects the biomass equilibrium curve. This happens, as indicated in the diagram, roughly where biomass is less than a quarter of the virgin stock equilibrium. The highest sustainable profits are obtained where an iso-profit curve is a tangent to the biomass growth function. As the diagram suggests, this occurs at a biomass of some 1300 thousand mt and harvest of over 300 thousand mt. At this point, annual variable profits from the fishery (approximately rents) amount to approximately 200 thousand mt of cod.

ESTIMATE OF RENT AND RENT LOSS

According to Refs 4 and 5, rents are defined as

$$\text{Rents}_t = \left(\frac{d\pi_t}{dy_t} \right) \cdot y_t,$$

where π_t represents profits and y_t harvest. For the profit function specified in the section titled “The Cod Fisheries Model,” this equation is

$$\text{Rents}_t = p_t \cdot y_t - k \cdot \delta \cdot \frac{y_t^\delta}{x_t^\varepsilon}. \quad (8)$$

Comparison with the profit function (6) shows that in this case, rents are unambiguously less than variable profits. This is because the profit function is strictly concave in harvest ($\delta > 1$). Note that because of the fixed costs, the rents may be greater or less than total

profits, depending on the level of harvest and biomass.

Given (8) and the fisheries model specified above, it is now an easy matter to estimate actual rent and rent loss. The latter, as explained in the introduction, is merely the difference between maximum attainable rent and current rent. In these calculations, the actual rent is calculated relative to conditions in 2005, the last year for which estimates of parameters and cod stock sizes are available. The estimated base year rent is referred to as the *current rent* below. The maximum rent is the maximum sustainable rent (Table 2).

So, it appears that in spite of its biologically depressed state, the Icelandic cod fishery is currently generating substantial economic rent as well as profit. However, compared to the maximum attainable rent, the fishery suffers from rent loss of some US\$426 million per year. In terms of current rent and the size of the fishery, this rent loss is very substantial. It should be noted, however, that to realize the maximum potential rent from this fishery, extensive rebuilding of the cod stock is required—a goal that has not been accomplished in spite of quite dramatic reduction in TAC levels since the early 1990s, and to a lesser extent, reduction in long-term fishing capital and effort. The former leads to substantial increase in sustainable catches that will be responsible for most of the economic gain.

The finding that the Icelandic cod fishery is apparently generating significant economic rent constitutes empirical evidence in support of the economically beneficial impact of the ITQ fisheries management system. That this rent appears to constitute less than 40% of the maximum attainable ones draws attention to the practical limitations of long overdue fisheries management. Before the introduction of the ITQ system, the Icelandic cod stock was subject to heavy overexploitation for decades. This will inevitably have altered ecosystem conditions, possibly to the point of transition to a different equilibrium attraction domain, and it may have adversely impact the genetic composition of the stock. As a result, it has proven much more difficult to restore the stock to economically optimal levels than predicted by the biological models.

UNCERTAINTY AND CONFIDENCE INTERVALS

The above estimates are subject to uncertainty and this uncertainty has many sources. There is uncertainty about the current state of the fishery, which is needed to estimate current rents. Similarly, there is uncertainty about the values of the parameters of the model used to calculate maximum future rents. Finally, there is uncertainty about the model itself. The first three types of uncertainty may be expressed as uncertainty about the numerical specifications of the current fishery and model parameters and the current state of the fishery. The importance and extent this uncertainty may be assessed by stochastic or Monte Carlo simulations [18].

Sensitivity analysis suggests that the estimated rent loss is most sensitive to the following: (i) initial stock size, (ii) MSY, (iii) carrying capacity of the biomass (i.e., XMAX) and (iv) price of landed catch, p . Hence, it makes sense to explore the uncertainty regarding these empirical estimates by stochastic simulations.

The stochastic simulations employed in this article are quite simplistic. Stochastic distributions for five model variables/coefficients, namely (i) landings price, (ii) initial biomass, (iii) cost coefficient, k , (iv) the estimated MSY, and (v) the estimated carrying capacity of the stock, XMAX, were specified. In all cases, the stochastic specifications took the form

$$z_i = \bar{z}_i \cdot (1 + u_i),$$

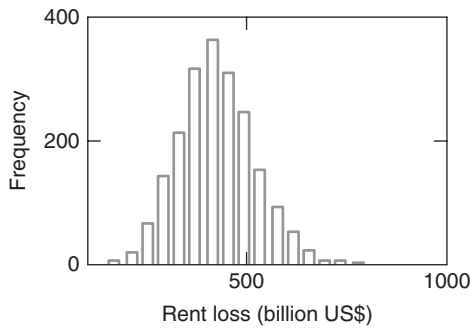
where z_i denotes the stochastic value of variable i , $\bar{z}_i$ the value specified in the section titled “The Cod Fisheries Model” above (especially Table 1 and below) and u_i is a normally distributed random variable with zero mean and some variance to be specified. It follows that each stochastic quantity z_i is normally distributed with expected value $\bar{z}_i$ and standard deviation $\bar{z}_i \cdot \sigma_i$ where σ_i is the standard deviation of variable i .

The stochastic specifications employed in the simulations are summarized in Table 3.

Two thousand random draws were drawn from the distributions specified in Table 3. For each set of draws, rent in the base

Table 3. Stochastic Specifications

Variable	Stochastic Version	Distribution of the Random Term	Expected Value	Standard Deviation
Landings price, p	$\bar{p} \cdot (1 + u_p)$	$u_p \sim N(0, 0.03)$	220	11
Initial biomass	$\bar{x} \cdot (1 + u_x)$	$u_x \sim N(0, 0.05)$	715	100
Cost parameter, k	$\bar{k} \cdot (1 + u_k)$	$u_k \sim N(0, 0.1)$	57605	5761
MSY	$\overline{\text{MSY}} \cdot (1 + u_{\text{MSY}})$	$u_{\text{MSY}} \sim N(0, 0.05)$	335	17
XMAX	$\overline{\text{XMAX}} \cdot (1 + u_{\text{XMAX}})$	$u_{\text{XMAX}} \sim N(0, 0.07)$	1998	140

**Figure 4.** Stochastically simulated rent loss: a histogram.**Figure 5.** Rent loss: an estimated distribution function.

year and maximum rent, and the difference between the two were calculated. Each difference represents one possible rent loss. The distribution of all these two thousand outcomes represents an estimate of the probability distribution of the rent loss. The corresponding histogram of rent loss is illustrated in Fig. 4.

As may be inferred from Fig. 4, the distribution of rent loss is skewed slightly to the right (a longer tail to the right). The coefficient of skewness was 0.314 (compared to zero for a normal distribution). Thus, the distribution is not quite normal. In accordance with this, the estimated mean rent loss was US\$428 million and the estimated median loss was US\$422 million. Compare this to the non-stochastic estimate of random loss derived in the section titled “Estimate of Rent and Rent Loss” (especially Table 2) of US\$426 million.

An estimated distribution function for the outcomes is drawn in Fig. 5. On the basis of this distribution function, confidence intervals for the rent loss are easy to calculate. A 95% confidence interval for the rent loss is US\$233 to 655 million. A 90% confidence

interval for the rent loss is US\$257 to 610 million.

SUMMARY AND CONCLUSION

According to the above, current (year 2005) total rent in the Icelandic cod fishery amounts to about US\$240 million annually. Maximum attainable annual rent, however, is estimated to be US\$667 million. Thus, the difference between current rent and maximum attainable rent, that is the rent loss, is estimated to be US\$426 million.

The main reason for the rent loss or submaximal generation of fisheries rent in the Icelandic cod fishery is the depressed state of the cod stock. Over the past several years, determined attempts have been made to restore this stock to its long-run optimal levels by imposing very restrictive TACs as compared to historical standards. Recent TACs for cod have been only about 40% of the estimated MSY and an even smaller fraction of the average harvest levels between 1970 and 1990. There are now signs that these

attempts may be bearing fruit and that the stock is slowly recovering. Nevertheless, it will probably take an additional 5 to 10 years to restore the stock to the level where it can generate the full economic rent of US\$667 million on a sustainable basis.

These numerical results concerning rents and rents loss are based on a fairly simplistic representation of the fishery, estimates of empirical quantities which cannot be accurate and several numerical approximations. Consequently the results cannot be very precise. Sensitivity analysis suggests that the calculated rents loss are most sensitive to the empirical assumptions about the initial stock size, the MSY, the carrying capacity of the biomass and the price of landed catch. Stochastic simulations over these and other empirical specifications suggest a 95% confidence interval for the rent loss of US\$233–655 million.

In spite of this comparatively high degree of uncertainty about the true numbers, the main message seems reasonably robust. The Icelandic cod fishery, in spite of being biologically quite depressed, is currently generating substantial economic rent. The rent, however, is less than half, perhaps as low as one-third of the maximum attainable one in sustainable equilibrium. To reach that point, the cod biomass must be increased greatly (almost by three-fourths) and fishing effort reduced by about one-sixth. Needless to say, this kind of change cannot be effected quickly—most likely, it will take several years even if all fishing for cod was immediately halted. It follows that in terms of present value, the potential rent gain is substantially less than that suggested above.

REFERENCES

1. Arrow KJ, Hahn FH. General competitive analysis. San Francisco (CA): Holden-Day; 1971.
2. Varian HR. Microeconomic analysis. 3rd ed. New York: Norton; 1992.
3. Alchian AA. Rent. In: Durlauf SN, Blume LE, editors. The New Palgrave Dictionary of Economics, Palgrave Macmillan, 09 Feb 2010, The New Palgrave Dictionary of Economics Online. Palgrave Macmillan; 2008. DOI: 10.1057/9780230226203.1420.
4. World Bank and FAO (Food and Agriculture Organization of the United Nations). The Sunken Billions: The Economic Justification for Fisheries Reform; Washington (DC); 2009.
5. Arnason R. Global fisheries rents loss: new results. A Paper Presented at the XVIIIth Annual EAFE Conference; 2007; Reykjavik, Iceland.
6. Hardin G. The tragedy of the commons. Science 1968;162:1243–1247.
7. Gordon HS. Economic theory of a common property resource: the fishery. J Polit Econ 1954;62(2):124–142.
8. Arnason R. Efficient harvesting of fish stocks: the case of the Icelandic demersal fisheries. PhD dissertation. Vancouver: University of British Columbia; 1984.
9. Arnason R, Sandal L, Steinshamn S, *et al.* Optimal feedback controls: comparative evaluations of the cod fisheries in Denmark, Iceland and Norway. Am J Agric Econ 2004;86(2): 531–542.
10. Arnason R. Property rights in fisheries: Iceland's experience with ITQs. Rev Fish Biol Fish 2006;15:243–264.
11. Arnason R. Minimum information management in fisheries. Can J Econ 1990;23(3): 630–653.
12. Anderson LG. The economics of fisheries management. Revised and enlarged Edition. New Jersey: Blackburn Press; 2004.
13. FAO. The State of the World Fisheries and Aquaculture 2006. Rome: Food and Agriculture Organization of the United Nations; 2007.
14. Hafrannsóknarstofnun. State of marine Stocks in Icelandic waters 2006–2007. Prospects for the quota year 2007/8. Fjölrit 129. Hafrannsóknarstofnunin Reykjavik; 2007.
15. Arnason R. Iceland's ITQ system creates new wealth. Electron J Sustain Dev 2008;1(2):35–41.
16. Agnarsson S, Arnason R, Johannsdottir K, *et al.* Comparative evaluation of the cod and herring fisheries in Denmark, Iceland and Norway: multispecies and stochastic issues. Report to the Nordic Council. Copenhagen: Nordic Council; 2007.
17. Hagfræðistofnun. Þjóðhagsleg áhrif aflareglu. Report No. C07:09. Reykjavik: Hagfræðistofnun. University of Iceland; 2007.
18. Hammersley JM, Handscomb DC. Monte Carlo methods. London: Methuen; 1975.

REPAIRABLE SYSTEMS: RENEWAL AND NONRENEWAL

STEVEN E. RIGDON

Department of Mathematics and Statistics,
Southern Illinois University, Edwardsville,
Illinois

OVERVIEW

When a system can be repaired upon failure, thereby bringing it to an operating condition, the system is said to be repairable. This repair could, but usually does not, mean complete replacement of the system. In contrast, a nonrepairable system is one that is usually not repaired upon failure. Items that were once considered repairable, such as a computer monitor or a handheld calculator, are now often discarded after failure because the replacement cost is lower than the repair cost. Automobiles, airplanes, and copy machines are examples of repairable systems, because the replacement cost is normally much higher than the repair cost.

For a repairable system, a model must account for the occurrence of events in time. We can describe such a model by using the actual failure times

$$0 < T_1 < T_2 < \cdots < T_n < \cdots,$$

measured from the initial start-up of the system. Time measured from system start-up is called *global time*. Alternatively, a model may describe the times between events or failures,

$$X_1, X_2, \dots, X_n, \dots,$$

Interevent times are measured in *local time*. Of course, these variables are related by

$$\begin{array}{ll} T_1 = X_1, & X_1 = T_1, \\ T_2 = X_1 + X_2, & X_2 = T_2 - T_1, \\ T_3 = X_1 + X_2 + X_3, & X_3 = T_3 - T_2, \\ \vdots & \vdots \end{array}$$

We say that a repairable system is *deteriorating* if the times between failure tend to get shorter, and we say that it is *improving* if the times between failures tend to get longer. Ebrahimi [1] gives a more precise and thorough description of deterioration and improvement.

EXAMPLES

In this section, we give a number of examples of data from repairable systems. Then in the section titled “Examples Revisited,” we consider the analysis of these data.

Example 1. Consider the failure times for the USS Halfbeak main propulsion engine [2]. Some of the failure times, measured in global time, are shown in Table 1. The plot of cumulative operating time against cumulative failure number is shown in Fig. 1. For this system, the failures seem to be getting closer together, indicating deterioration. Table 2 shows the number of failures in each 2000-h interval, indicating clearly that the rate of occurrence of failures (ROCOF) increases with the age of the system.

Table 1. Failure Times of USS Halfbeak

Failure number	Failure time	Time between failure
1	1,382	1,382
2	2,990	1,608
3	4,124	1,134
4	6,827	2,703
5	7,472	645
6	7,567	95
7	8,845	1,278
⋮	⋮	⋮
⋮	⋮	⋮
68	21,461	1,030
69	21,930	469
70	22,846	916
71	24,006	1,160

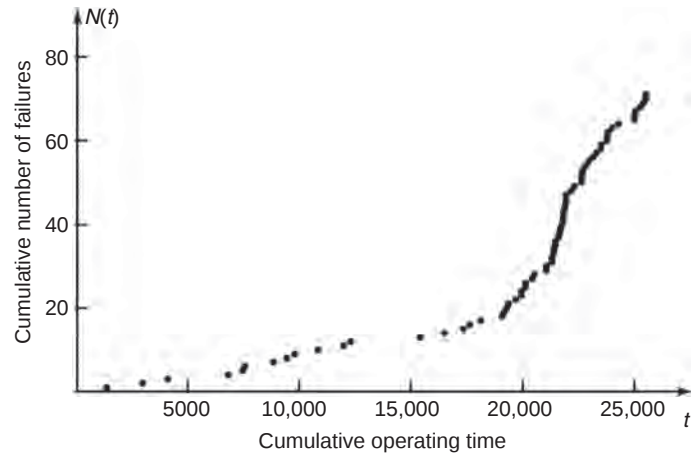


Figure 1. Failure times of USS Halfbeak.

Table 2. Failures in Intervals of 2000 h

Time interval	Number of failures
[0, 2000]	1
(2000, 4000]	1
(4000, 6000]	1
(6000, 8000]	3
(8000, 10000]	3
(10000, 12000]	2
(12000, 14000]	1
(14000, 16000]	1
(16000, 18000]	3
(18000, 20000]	8
(20000, 22000]	23
(22000, 24000]	15
(24000, 25518]	9

Example 2. Table 3, from Ref. 3, shows the failure times of a photocopy machine. Here, time is measured not in hours or days, but in the number of copies made. Similarly, the age of an automobile is often measured in miles, not years. The data seem to show neither improvement nor deterioration.

Example 3. Balaban [4] gave an example of the failures of a mechanical system. The failures during each month are shown in Table 4. This example differs from the previous examples, because the exact failure times are not known. All that is known is that a certain number of failures occurred in each month with a certain number of operating hours. This system seems to indicate improvement, because one-third of all failures occurred in the first two months,

Table 3. Failure Times, Measured in Number of Copies, for Copy Machine

Failure number	Copy count	Failure number	Copy count
1	2,827	10	131,832
2	13,128	11	142,555
3	25,732	12	153,277
4	46,237	13	158,395
5	58,014	14	159,703
6	71,165	15	159,720
7	79,749	16	177,205
8	90,556	17	191,661
9	112,319	18	196,877

Table 4. Failure Data for Balaban's Mechanical System, ARN-118

Month of operation	Number of operating hours	Cumulative operating hours	Number of failures
1	541	541	3
2	1,171	1,712	5
3	1,939	3,651	4
4	2,403	6,054	1
5	1,718	7,772	2
6	2,206	9,978	2
7	1,366	11,344	3
8	1,529	12,873	0
9	1,449	14,322	2
10	1,451	15,773	2

which also had the fewest number of operating hours.

Example 4. Jelinski and Moranda [5] gave the failure times for a software system. The failure times and times between failure are

Table 5. Failure Data for 41 Valves

System	Failure times	Truncation time
1		761
2		759
3	98	667
4	326 653 653	667
5		665
6	84	667
7	87	663
8	646	653
9	92	653
10		651
11	258 328 377 621	650
12	61 539	648
13	254 276 298 640	644
14	76 538	642
15	635	641
16	349 404 561	649
17		631
18		596
19	120 479	614
20	323 449	582
21	139 139	589
22		593
23	573	589
24	165 408 604	606
25	249	594
26	344 497	613
27	265 586	595
28	166 206 348	389
29		601
30	410 581	601
31		611
32		608
33		587
34	367	603
35	202 563 570	585
36		587
37		578
38		578
39		586
40		585
41		582

shown in Table 5. Reliability improvement is clear since the times between failure are much larger near the end than at the beginning and middle of the testing period.

Example 5. Nelson [6] gave data on the failure times of 41 valves; the data are presented in Table 6. Each valve is a repairable system, but many of the valves had no failures. The truncation

Table 6. Software Failure Data from Ref. 5

Failure number	Failure time	Time between failure
1	9	9
2	21	12
3	32	11
4	36	4
5	43	7
6	45	2
7	50	5
8	58	8
9	63	5
10	70	7
11	71	1
12	77	6
13	78	1
14	87	9
15	91	4
16	92	1
17	95	3
18	98	3
19	104	6
20	105	1
21	116	11
22	149	33
23	156	7
24	247	91
25	249	2
26	250	1
27	337	87
28	384	47
29	396	12
30	405	9
31	540	135
32	798	258
33	814	16
34	849	35

time, that is, the time that data collection stopped, is included in Table 6. A system that experiences no failures still contains information about the failure process. For example, system 1 contained no failures in 761 time units of operation, giving us some (but not much) information about the rate of occurrence of failures for that system.

NONREPAIRABLE SYSTEMS

Many concepts regarding repairable systems have direct analogs to concepts for nonrepairable systems, so we present a

brief review. Typically, the lifetime for a nonrepairable system is a random variable T . Let the PDF, CDF, and survivor function for T be denoted by $f(t)$, $F(t)$, and $S(t)$, respectively. The hazard function is the limit of a conditional probability

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

An increasing hazard indicates that the probability of an imminent failure increases with the age of the system, an indication of the system wearing out. It can be shown [3] that

$$h(t) = \frac{f(t)}{S(t)}.$$

In some books, this is taken as the definition, although the limit definition is often preferred because it gives insight into what the hazard measures. Leemis [7] gives a table showing the relationships between the PDF, CDF, survivor function, and hazard function.

Three commonly used lifetime distributions are the exponential, the Weibull, and the gamma. The exponential distribution with mean θ , denoted $\text{EXP}(\theta)$, has PDF

$$f(t) = \frac{1}{\theta} e^{-t/\theta}, \quad t > 0.$$

The hazard function for the exponential distribution is the constant $h(t) = \frac{1}{\theta}$. In addition, the exponential distribution has the memoryless property,

$$P(T > t + x | T > t) = P(T > x);$$

that is, given survival to time t , the probability of survival of an additional x time units is equal to the probability that a new unit survives x units. Thus, age does not affect the probability of survival.

The Weibull distribution, denoted $\text{WEI}(\beta, \theta)$ has PDF

$$f(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta}\right)^{\beta-1} \exp\left(-\left(\frac{t}{\theta}\right)^\beta\right), \quad t > 0.$$

The mean for the Weibull distribution is $\mu = \theta \Gamma(1 + 1/\beta)$ and the hazard function is

$$h(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta}\right)^{\beta-1}.$$

The Weibull distribution, unlike the exponential, can model systems that have an increasing or a decreasing hazard function, but it cannot model lifetimes that have a hazard function that is sometimes increasing and sometimes decreasing.

The gamma distribution, denoted $\text{GAM}(\kappa, \theta)$ has PDF

$$f(t) = \frac{t^{\kappa-1}}{\theta^\kappa \Gamma(\kappa)} e^{-t/\theta}, \quad t > 0.$$

The mean of this distribution is $\mu = \kappa\theta$. The hazard function is

$$h(t) = \frac{\frac{t^{\kappa-1}}{\theta^\kappa \Gamma(\kappa)} e^{-t/\theta}}{\int_t^\infty \frac{u^{\kappa-1}}{\theta^\kappa \Gamma(\kappa)} e^{-u/\theta} du} = \frac{t^{\kappa-1} e^{-t/\theta}}{\int_t^\infty u^{\kappa-1} e^{-u/\theta} du},$$

which converges to the value $1/\theta$ as $t \rightarrow \infty$.

For systems that have a high hazard when very young, followed by a fairly constant one, followed by a rapidly increasing hazard, a bathtub hazard is appropriate. Figure 2 illustrates this.

REPAIRABLE SYSTEMS: RENEWAL

When a system is repaired to a like-new condition, or if repair means complete system replacement, then the renewal process is used as a process model. The renewal process assumes that the times between failures $X_1, X_2, \dots$ are independent and identically distributed.

For any counting process, that is, the process that describes the occurrence of events in time, the complete intensity function is defined to be

$$\lambda(t | \mathcal{H}_t) = \lim_{\Delta t \rightarrow 0} \frac{P(\text{a failure occurs in } [t, t + \Delta t] | \mathcal{H}_t)}{\Delta t},$$

where $\mathcal{H}_t$ is the entire history of the process up to time t . If the distribution of the times between failures has an increasing hazard that begins with $h(0) = 0$, as for example the Weibull distribution does when $\beta > 1$, then

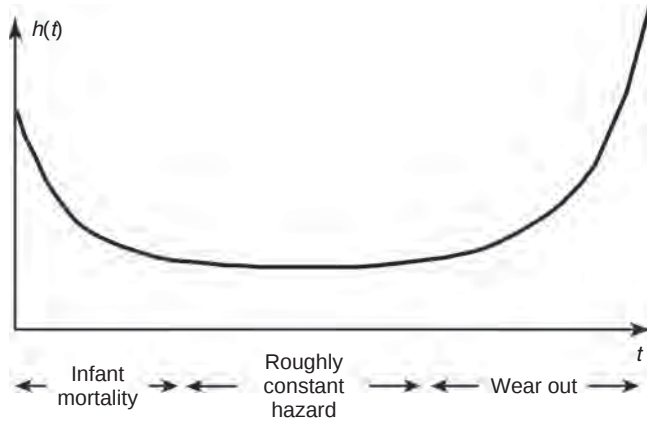


Figure 2. For a bathtub-shaped hazard function, there is a large hazard for very new components, followed by a period of roughly constant hazard and finally, wear out.

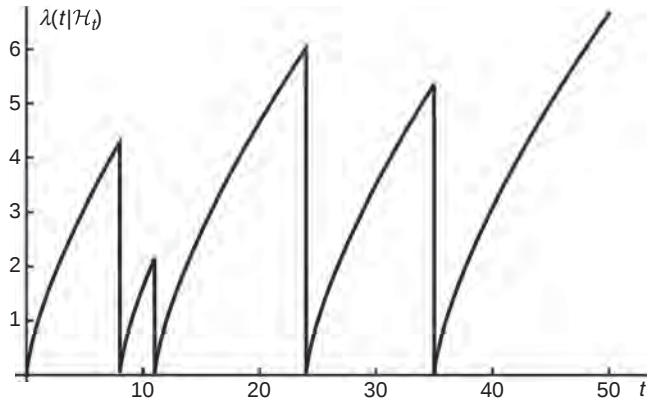


Figure 3. Complete intensity function for a renewal process with increasing hazard function, conditioned on failures at times 8, 11, 24, and 35.

the complete intensity function will increase from zero until a failure occurs; then it will reset to zero and increase until the next failure, and continue this process as long as the system operates. This is illustrated in Fig. 3. Care must be taken, however, to interpret this plot; it is conditioned on the failures at times 8, 11, 24, and 35.

Let

$$\Lambda(t) = E[\text{number of failures in } [0, t]].$$

The derivative $\mu(t) = \Lambda'(t)$ is called the *ROCOF*, or sometimes the *renewal density*. It turns out that the ROCOF is equal to limit of unconditional probabilities

$$\mu(t) = \lim_{\Delta t \rightarrow 0} \frac{P(\text{a failure occurs in } [t, t + \Delta t])}{\Delta t},$$

provided that simultaneous failures have probability zero [2]. It is not difficult to show that the ROCOF is equal to

$$\mu(t) = \sum_{k=1}^{\infty} f^{(k)}(t),$$

where $f^{(k)}(t)$ is equal to the PDF of the time of the k th failure, measured in global time, that is, the time since the initial start-up of the system [3].

Figure 4 shows the PDFs for the failure times (measured in global time) for a gamma renewal process, where the times between failure are IID $\text{GAM}(20, \frac{1}{10})$. The mean of this distribution is $20 \times \frac{1}{10} = 2$ and the variability is fairly low. The ROCOF, shown in Fig. 5, oscillates, but converges to 0.5, which is the reciprocal of the mean. This is to be

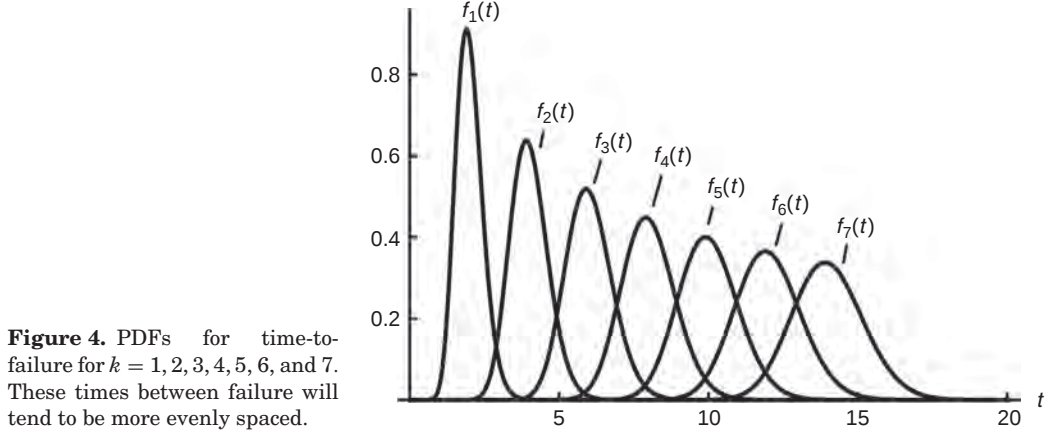


Figure 4. PDFs for time-to-failure for $k = 1, 2, 3, 4, 5, 6$, and 7 . These times between failure will tend to be more evenly spaced.

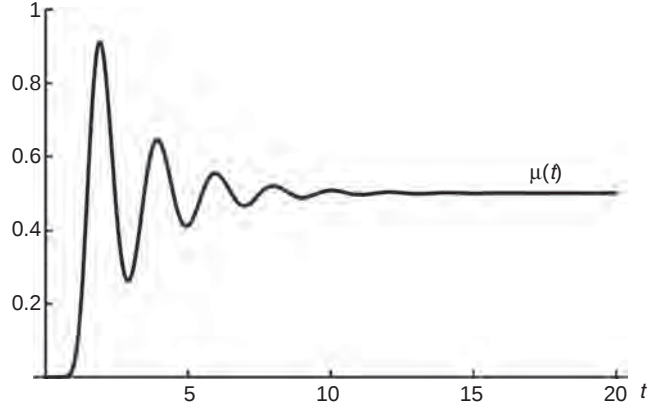


Figure 5. ROCOF for renewal process for which the PDFs for failure times are in Fig 4.

expected; if failures occur on average every 2 h, then in the long run, failures will occur at the rate of one per 2 h, or 0.5 failures per hour.

The results about the asymptotic behavior of the ROCOF are guaranteed by the next theorem.

Theorem 1. *For a renewal process with IID times between failure $X_1, X_2, \dots$, and $\eta = E(X_i)$,*

1. $\lim_{t \rightarrow \infty} \frac{\Lambda(t)}{t} = \frac{1}{\eta},$
2. $\lim_{t \rightarrow \infty} \mu(t) = \frac{1}{\eta}.$

See Ref. 3 for a proof.

In a renewal process situation, we typically know the distribution of the times between failure, that is, we know the PDF $f(x)$ and the CDF $F(x)$. The functions $\Lambda(t)$ and $\mu(t)$ are usually more difficult to obtain. There are, however, integral equations that express the relationships among these functions. Rigdon and Basu [3, pp. 72–73] show that $\Lambda(t)$ and $\mu(t)$ satisfy the following integral equations:

$$\Lambda(t) = F(t) + \int_0^t \Lambda(t-y)f(y)dy$$

and

$$\mu(t) = f(t) + \int_0^t \mu(t-y)f(y)dy.$$

If the repair time is taken into account, we can ask for the probability that the system is available at time t . The system availability, as a function of time t , is

$$A(t) = S(t) + \int_0^t S(t-y) \mu(y) dy.$$

This can be explained intuitively by saying that the system is available at time t if and only if one of the two conditions hold:

1. The system has operated with no failures up to time t . This occurs with probability $S(t)$.
2. The system has operated with no failures since the last repair. Conditioning on the time of the last repair and integrating gives us a probability of

$$\int_0^t S(t-y) \mu(y) dy.$$

The availability is thus, the sum of these two probabilities.

The average availability over the interval $[0, T]$ is defined to be

$$A_{\text{avg}}(t) = \frac{1}{T} \int_0^T A(t) dt.$$

The steady-state availability is the limit of the availability function as $t \rightarrow \infty$

$$A_{\text{ss}} = \lim_{t \rightarrow \infty} A(t) = \frac{E(X)}{E(X) + E(R)},$$

where X is the time-to-failure and R is the repair time. The steady-state availability can then be written as

$$A_{\text{ss}} = \frac{MTBF}{MTBF + MTTR}$$

where MTBF is the mean time-between-failures and MTTR is the mean time-to-repair. This steady-state formula makes intuitive sense; if a system fails on average every 9 h and it takes 1 h to repair it on average, then we would expect it to be working 9 times out of 10, for an availability of 0.9.

REPAIRABLE SYSTEMS: POISSON PROCESSES

The simplest model for repairable systems is the homogeneous Poisson process (HPP). Suppose we have a counting process where $N(t)$ is the (random) number of events in the interval $[0, t]$, and $N(a, b]$ is the number of events in the interval $(a, b]$. We say that a counting process $N(t)$ is a HPP if the following conditions hold.

1. $N(0) = 0$.
2. For every $a < b \leq c < d$, the random variables $N(a, b]$ and $N(c, d]$ are independent. This is called *independent increments*.
3. $\lim_{\Delta t \rightarrow 0} P(N(t, t + \Delta t] \geq 1) / \Delta t = \lambda$. The constant λ is called the *intensity* of the process.
4. $\lim_{\Delta t \rightarrow 0} P(N(t, t + \Delta t] \geq 2) / \Delta t = 0$. This assumption precludes simultaneous failures.

Under these conditions, it follows that $N(a, b]$ has a Poisson distribution with mean $\lambda(b - a)$, and the times between failures have independent exponential distributions with mean $1/\lambda$. The Poisson process is not a particularly reasonable model for repairable systems because it does not allow for either reliability improvement or deterioration.

If we replace assumption (3) above with the assumption

$$\lim_{\Delta t \rightarrow 0} \frac{P(N(t, t + \Delta t] \geq 1)}{\Delta t} = \lambda(t),$$

we are in effect allowing the probability of failure in a small interval to depend on the age of the system. The process obtained in this way is called the nonhomogeneous Poisson process (NHPP) and the function $\lambda(t)$ is called the *intensity function*. It can be shown that $N(a, b]$ has a Poisson distribution whose mean is $\int_a^b \lambda(t) dt$.

Various assumptions can be made about the functional form of $\lambda(t)$. Cox and Lewis[8] studied the NHPP with intensity

$$\lambda(t) = \exp(a + bt).$$

A more common form is

$$\lambda(t) = \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1}.$$

The NHPP obtained with this intensity function is called the *power law process* (PLP). Notice that the functional form of the intensity is identical to that of the hazard function of the Weibull distribution. The interpretations of the intensity and the hazard are completely different, however. The intensity is the instantaneous probability of a failure (any failure) of the system. The hazard is the instantaneous probability of the one and only failure of the system conditioned on survival up to that point. Because the PLP has an intensity function that is identical to the Weibull's hazard, the power law process was for a time called the *Weibull process*. Since there is a risk of confusing the Weibull distribution with the Weibull process, the term "power law process" is preferred over "Weibull process."

If $\beta > 1$, then the PLP has an increasing intensity function, indicating deterioration, since the failures tend to get closer together. If $\beta < 1$, then the PLP has a decreasing intensity, indicating reliability improvement. If $\beta = 1$, then the PLP reduces to the HPP with rate $1/\theta$.

The mean number of failures in $[0, t]$ is then

$$\Lambda(t) = E(N(t)) = \int_0^t \frac{\beta}{\theta} \left(\frac{y}{\theta} \right)^{\beta-1} dy = \left(\frac{t}{\theta} \right)^{\beta}.$$

Dividing by t and taking the log yields

$$\begin{aligned} \log \frac{\Lambda(t)}{t} &= (\beta - 1) \log t - \beta \log \theta \\ &= \alpha_0 + \alpha_1 \log t, \end{aligned}$$

where $\alpha_0 = -\beta \log \theta$ and $\alpha_1 = \beta - 1$. Duane [9] suggested plotting $\log [N(t_i)/t_i]$ (on the vertical axis) and $\log t_i$ (on the horizontal axis) in what became known as a *Duane plot*. It might seem that the PLP implies a linear Duane plot and vice versa. Neither implication, however, is true. Rigdon [10] showed that the PLP does not imply a linear Duane plot, and a linear Duane plot does not imply a PLP.

Suppose now that we have observed data from a failure process and we would like to estimate the parameters θ and β in the PLP. We consider two ways of collecting data from a repairable system. Data are said to be *failure truncated* if the observation of the repairable system was terminated after a fixed number, n , of failures. Data are said to be *time truncated* if the observation of the repairable system was terminated at a fixed time T . Point estimation, interval estimation, and hypothesis tests are somewhat different for these two kinds of truncation. (Note that we use the term truncation when referring to data collection for a repairable system, and censoring when we have a nonrepairable system.) When data are failure truncated, the number of failures, n , is fixed, but the time that testing stops, T_n , is a random variable. Conversely, when the data are time truncated, the number of failures, N , is random and the time that testing stops, T , is fixed.

When the data are failure truncated, the maximum likelihood estimators (MLEs) of the parameters are

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n-1} \log \left(\frac{t_n}{t_i} \right)}$$

and

$$\hat{\theta} = \frac{t_n}{n^{1/\hat{\beta}}}.$$

The random variable $2n\beta/\hat{\beta}$ is a pivotal quantity, having a chi-square distribution with $2(n-1)$ degrees of freedom, so a confidence interval for β is

$$\left(\frac{\chi_{1-\alpha/2}^2 (2(n-1)) \hat{\beta}}{2n}, \frac{\chi_{\alpha/2}^2 (2(n-1)) \hat{\beta}}{2n} \right).$$

Hypothesis tests for β can also be developed from this pivotal quantity. If we test $H_0: \beta = \beta_0$ against a two-sided alternative, we would reject H_0 for

$$\hat{\beta} < \frac{2n\beta_0}{\chi_{\alpha/2}^2 (2(n-1))} \text{ or } \hat{\beta} > \frac{2n\beta_0}{\chi_{1-\alpha/2}^2 (2(n-1))}.$$

Since $\beta = 1$ reduces the PLP to the HPP, we are often interested in testing $H_0: \beta = 1$.

Finkelstein [11] showed that the statistic $W = (\hat{\theta}/\theta)^\beta$ is also a pivotal quantity. Tables from Finkelstein can then be used to get a confidence interval for θ . Exact goodness-of-fit tests can be constructed from the result that, conditioned on $T_n = t_n$, the random variables

$$U_i = \log \left(\frac{t_n}{T_{n-i}} \right), i = 1, 2, \dots, n-1$$

are distributed as $n-1$ order statistics from an exponential distribution with mean $1/\beta$. Thus, any goodness-of-fit test for the exponential distribution with unknown mean can be used to test fit for the PLP. See Ref. 12 for a study of various goodness-of-fit tests, including estimates of the power.

When the data are time truncated with truncation time T , the MLEs are

$$\hat{\beta} = \frac{N}{\sum_{i=1}^N \log \left(\frac{T}{t_i} \right)}$$

and

$$\hat{\theta} = \frac{T}{N^{1/\hat{\beta}}}.$$

The random variable $2n\beta/\hat{\beta}$ is a pivotal quantity, that has a chi-square distribution with $2n$ degrees of freedom. A confidence interval for β is thus

$$\left(\frac{\chi_{1-\alpha/2}^2(2n)\hat{\beta}}{2n}, \frac{\chi_{\alpha/2}^2(2n)\hat{\beta}}{2n} \right).$$

Hypothesis tests for β can also be developed. If we test $H_0: \beta = \beta_0$ against a two-sided alternative, we would reject H_0 for

$$\hat{\beta} < \frac{2n\beta_0}{\chi_{\alpha/2}^2(2n)} \text{ or } \hat{\beta} > \frac{2n\beta_0}{\chi_{1-\alpha/2}^2(2n)}.$$

Goodness-of-fit tests can also be developed for the time-truncated case. See Refs 3 and 12 for details.

Bayesian inference procedures for the PLP were developed by Calabria *et al.* [13] and Bar-Lev *et al.* [14].

When we have data from multiple repairable systems, as in Example 5 from a

previous section, inference for the parameters depends on what assumptions we are willing to make about the similarity of the systems. If we have k systems, the possibilities include the following:

1. The k systems are identical. The only two parameters to estimate are θ and β .
2. The k systems are entirely different. We would then estimate θ_j and β_j for each system, $j = 1, 2, \dots, k$.
3. The k systems have the same β parameter, but different θ_j s.
4. The k systems have the same β parameter, but different θ_j s, where the θ_j s come from some prior distribution. This leads to an empirical Bayes model.
5. The k systems have different parameters, but the β_j s and the θ_j s each come from some prior distribution.

Details of inference procedures can be found in Ref. 3, Chapter 5.

REPAIRABLE SYSTEMS: OTHER MODELS

There are a number of other models that can be used for repairable systems.

Sen and Bhattacharyya [15] and Sen [16] proposed the piecewise exponential (PEXP) model. This model assumes that the times between failure are independent exponential random variables having means

$$E(X_i) = \frac{\delta}{\mu} i^{\delta-1}.$$

The times between failure are therefore, not identically distributed unless $\delta = 1$, in which case the PEXP reduces to the HPP. In the PEXP model with $\delta > 1$ the expected times between failures are increasing, indicating reliability improvement; however in the PEXP model, this improvement occurs at the times of failure, whereas with the PLP, or any other NHPP, the improvement is continuous. In the context of test-analyze-and-fix (TAAF), when system improvements are made when an item fails, the PEXP model may be more realistic.

Lakey and Rigdon [17] and Black and Rigdon [18] studied a model suggested by Berman [19] for repairable systems. This model is called the *modulated power law process* (MPLP). Suppose that “shocks” occur to a system according to a PLP, but a failure occurs, not at every shock but at every κ th shock, where κ is a positive integer. Under an NHPP, a repair brings the condition of the system to where it was just before the failure; for this reason, the NHPP is known as a bad-as-old, or same-as-old, model. By contrast, if the renewal process governs the failure times, then a failed/repared unit is in like-new condition. The renewal process is a good-as-new, or same-as-new, model. The MPLP is thus a compromise between the good-as-new and bad-as-old models. If $\kappa > 1$, then a failure/repair brings the system back to a condition better than it was just before the failure (because the shock counter gets reset), but not necessarily as good as a new system. The likelihood function turns out to be a valid likelihood if $\kappa > 0$, even if κ is not a positive integer, although in this case the model is harder to interpret. Inference procedures, including point and interval estimates of the three parameters, κ , β , and θ , are given in Ref. 18.

Doyen and Gaudoin [20] developed two models that are also considered a compromise between the good-as-new (renewal process) and the bad-as-old (NHPP) models. One model assumes that at the time of failure/repair, the intensity is reduced by a constant multiple, but not as far as it would be if it were renewed. Thus, a failed unit is better than just before the failure, but not as good as a new one. This model is called the *arithmetic reduction in intensity* (ARI) model. The second model reduces the “virtual age” of the system at the time of the failure by a fixed factor. This model is called the arithmetic reduction of age (ARA) model. Inference for a single system is given in Ref. 20 and a hierarchical Bayes method for multiple systems is given in Ref. 21.

EXAMPLES REVISITED

For illustration, let us consider the five examples discussed earlier.

Example 1. For the data on unscheduled maintenance times for the USS Halfbeak, we assume the PLP. The estimates (MLEs) of β and θ are

$$\hat{\beta} = \frac{71}{\sum_{i=1}^{n-1} \log\left(\frac{25518}{t_i}\right)} \approx 2.760$$

and

$$\hat{\theta} = \frac{25518}{71^{1/2.760}} \approx 5447.3.$$

The 95% confidence interval for β is

$$\left(\frac{\chi_{0.975}^2(2(71-1)) 2.760}{2 \times 71}, \frac{\chi_{0.025}^2(2(71-1)) 2.760}{2 \times 71} \right)$$

or

$$(2.122, 3.395),$$

an interval that clearly excludes 1. Thus $\beta = 1$ is not a plausible value, so we conclude that there is deterioration in the system.

Example 2. Table 3 gives the failure times of a copy machine. Here, time is measured in numbers of copies made, which is a discrete measure, although the numbers are reported to such precision that we will take these as continuous measures. Again, assuming a PLP, we have estimates

$$\hat{\beta} = \frac{18}{\sum_{i=1}^{n-1} \log\left(\frac{196,844}{t_i}\right)} \approx 1.081$$

and

$$\hat{\theta} = \frac{196,844}{18^{1/1.081}} \approx 13,571.$$

A 95% confidence interval for β is (0.595, 1.560) which includes 1. There is no evidence of reliability improvement or deterioration.

Example 3. Consider now Balaban’s data. Here the exact failure times are unknown and we only know the operating time and the number of failures that occurred in each

interval. The observed data are then the number of failures N_i , $i = 1, 2, \dots, 10$. Under the PLP, these random variables are independent (but not identically distributed) and have distribution

$$\begin{aligned} N_i &\sim \text{POISSON} \left(\int_{t_{i-1}}^{t_i} \frac{\beta}{\theta} \left(\frac{t}{\theta} \right)^{\beta-1} dt \right) \\ &= \text{POISSON} \left(\left(\frac{t_i}{\theta} \right)^\beta - \left(\frac{t_{i-1}}{\theta} \right)^\beta \right), \end{aligned}$$

where t_i represents the right end point of the i th interval and $t_0 = 0$. The likelihood function is thus the product of Poisson likelihoods

$$L(\beta, \theta) = \prod_{i=1}^{10} \frac{1}{n_i!} \left[\left(\frac{t_i}{\theta} \right)^\beta - \left(\frac{t_{i-1}}{\theta} \right)^\beta \right]^{n_i}$$

$$\begin{aligned} &\times \exp \left[- \left(\frac{t_i}{\theta} \right)^\beta + \left(\frac{t_{i-1}}{\theta} \right)^\beta \right] \\ &= \left(\prod_{i=1}^{10} \left[\left(\frac{t_i}{\theta} \right)^\beta - \left(\frac{t_{i-1}}{\theta} \right)^\beta \right]^{n_i} \right) \\ &\times \exp \left[- \left(\frac{t_n}{\theta} \right)^\beta \right] \left(\prod_{i=1}^{10} n_i! \right)^{-1}. \end{aligned}$$

The logarithm of this likelihood can then be maximized using numerical methods. A contour plot of the log-likelihood function is shown in Fig. 6. The MLEs, found by numerically optimizing the log-likelihood function, are

$$\hat{\beta} = 0.5901 \text{ and } \hat{\theta} = 77.70.$$

Asymptotic 95% confidence intervals for β and θ , obtained using the observed Fisher information matrix, are

$$(0.5067, 0.6735)$$

and

$$(16.87, 138.5),$$

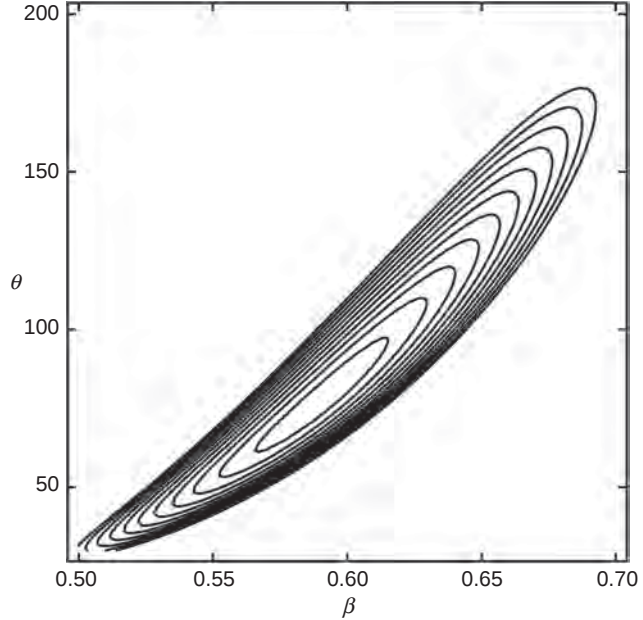


Figure 6. Contour plot for log-likelihood for Balaban's data.

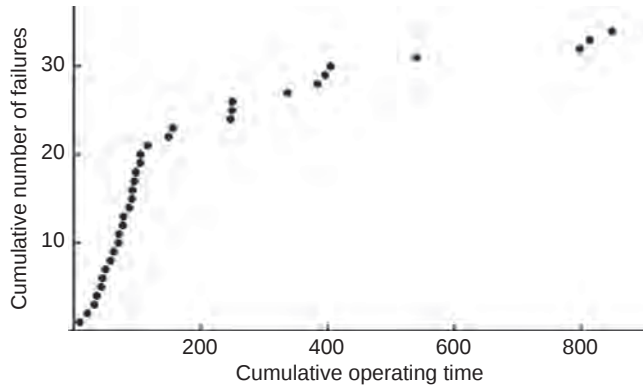


Figure 7. Jelinski and Moranda [5] software data.

respectively. Here, the confidence interval for β contains only values of β that are less than one, indicating reliability improvement.

Example 4. Next, consider the Jelinski and Moranda data on software failures. The data from Table 6 are shown in Fig. 7, where cumulative operating time is plotted against cumulative number of failures. This figure shows that initially failures occur at roughly a constant rate (approximately 20 per hundred hours). After about 100 h, the pattern of failures is much more erratic. When modeling software, we must take into account the existence of “bugs” in the system, and testing is designed to find the bugs. The reader is referred to Ref. 5 for a more detailed analysis of this data.

Example 5. Finally, consider the data from Table 5 which gives the failure times of 41 valves. Note that 17 out of the 41 valves had no failures at all, and the most failures on any system was four. With the paucity of data, we assume that the systems are identical and that the failure times are governed by a time-truncated PLP. The MLEs of β and θ are

$$\hat{\beta} = 1.420$$

and

$$\hat{\theta} = 548.9.$$

A 95% confidence interval for β is (1.070, 1.890), giving (marginal) evidence of reliability deterioration.

REFERENCES

1. Ebrahimi N. How to define system improvement and deterioration for a repairable system. *IEEE Trans Reliab* 1989;R-38:214–217.
2. Ascher H, Feingold H. Repairable systems reliability: modeling, inference, misconceptions and their causes. New York: Marcel Dekker; 1984.
3. Rigdon SE, Basu AP. Statistical methods for the reliability of repairable systems. New York: Wiley; 2000.
4. Balaban HS. Reliability growth models. *J Environ Sci* 1978;21:11–18.
5. Jelinski Z, Moranda PB. Software reliability research. In: Freiburger W, editor. Statistical computer performance evaluation. London: Academic Press; 1972.
6. Nelson W. Confidence limits for the recurrence data—applied to cost or number of product repairs. *Technometrics* 1995;37:147–157.
7. Leemis LM. Reliability: probabilistic models and statistical methods. Upper Saddle River (NJ): Prentice Hall; 1995.
8. Cox DR, Lewis PAW. The statistical analysis of series of events. London: Chapman and Hall; 1966.
9. Duane JT. Learning curve approach to reliability monitoring. *IEEE Trans Aerosp* 1964;2:563–566.
10. Rigdon SE. Properties of the Duane plot for repairable systems. *Qual Reliab Eng Int* 2002;18:1–4.
11. Finkelstein JM. Confidence bounds on the parameters of the Weibull process. *Technometrics* 1976;18:115–117.
12. Rigdon SE. Testing goodness-of-fit for the power law process. *Commun Stat* 1989;A-18:4665–4676.

13. Calabria R, Guida M, Pulcini G. Bayes estimation of prediction intervals for a power law process. *Commun Stat Theory Methods* 1990;19:3023–3035.
14. Bar-Lev S, Lavi I, Reiser B. Bayesian inference for the power law process. *Ann Inst Stat Math* 1992;44:632–639.
15. Sen A, Bhattacharyya GK. A piecewise exponential model for reliability growth and associated inferences. In: Basu AP, editor. *Advances in reliability*. Amsterdam, North-Holland Elsevier; 1993. pp. 331–355.
16. Sen A. Estimation of current reliability in a Duane-based reliability growth model. *Technometrics* 1998;40:334–344.
17. Lakey MJ, Rigdon SE. The modulated power law process. *Transactions of the 46th Annual Quality Congress. American Society for Quality Milwaukee(WI)*; 1992. pp. 559–563.
18. Black SE, Rigdon SE. Statistical inference for a modulated power law process. *J Qual Technol* 1996;28:81–90.
19. Berman M. Inhomogeneous and modulated gamma processes. *Biometrika* 1981;68: 143–152.
20. Doyen L, Gaudoin O. Classes of imperfect repair models based on reduction of failure intensity or virtual age. *Reliab Eng Syst Saf* 2004;84:45–56.
21. Pan R, Rigdon SE. Bayes inference for repairable systems. *J Qual Technol* 2009;41: 82–94.

RESOURCE MODELING ASSOCIATION

STEVEN MCKELVEY
Professor of Mathematics and
Computer Science, Center for
Integrative Studies, St. Olaf
College, Northfield, Minnesota

The Resource Modeling Association (RMA) was conceived of and founded in the early 1980s by Dr. Roland Lamberson of Humboldt State University and Dr. Robert McKelvey of the University of Montana. Working closely together, these scientists recruited a group of applied mathematicians, applied population biologists, fisheries scientists, and resource economists, primarily from the west coast of North America, to participate in annual meetings to discuss the application of models to resource management. Over the years, the RMA has grown into a truly international professional society with active members from throughout the world. North America, Europe, and Asia are particularly well represented.

The goals of this broadly interdisciplinary research-oriented society are to discuss and exhibit exemplary applications of modeling technology to questions of natural resource management. Here “natural resource” is understood in its broadest meaning, with an emphasis on renewable resources. Particular issues of concern to the society are the effects of global warming on natural resources, the creation and design of ecological preserves, population dynamics, sustainable development, the appropriate use of models in policy formulation, and water use policy.

A secondary goal of the association is to create and demonstrate professional standards for the articulation of key ideas in the design, implementation, and application of mathematical models to problems involving natural resources. In the early days of the society, the mechanisms for presenting models and their results in a

scholarly acceptable fashion were unclear. Clear standards existed for the presentation of theoretical analyses of mathematical constructs and equally clear standards existed for the presentation of quantitative results of field studies. In the early 1980s, it was not so clear how mathematical models and their results ought to be presented. A typical model would possess certain mathematical properties, equilibria of various types, chaotic behavior, quantifiable asymptotic behavior, and so on: important qualities to elucidate in professional articles. However, it was not so clear how to present the results of these models when they were applied to important “real-world” situations. The usual statistical analyses seemed inappropriate as the quantities being analyzed were not truly random in nature, but were the result of either a deterministic model or a pseudorandom computation.

This confusion, which continues today in the cases of newer styles of models such as individual-based models, is a significant issue for scholars pursuing work in the modeling field. The RMA has played, and continues to play, a key role in defining and promulgating professional standards for the description, application, and analysis of mathematical models involving natural resource issues.

The society seeks to meet its goals through two main activities: hosting an annual world conference on Natural Resource Modeling and the publication of the scholarly journal *Natural Resource Modeling* (Wiley-Blackwell), a quarterly, peer reviewed professional publication.

The RMA strives to support early career researchers in the various subdisciplines of natural resource modeling. Members are typically active in several professional organizations and the RMA seeks to nurture its members’ professional development as a way to strengthen their individual research work and visibility as well as to support the growth and strength of the RMA itself.

The society’s journal, *Natural Resource Modeling* (NRM), takes special efforts,

beyond the simple return of reviewers' comments, to work with authors to strengthen their writing and organizational skills to ensure a productive publication career into the future. If desired by an author, the NRM's editorial staff is willing to engage in extensive discussions with the author about a submitted paper and changes in the work that would result in a stronger presentation. This detailed and personal feedback, useful for any author, is particularly valuable for those early in their publication career and seeking to establish a solid presence in the profession.

The RMA's annual meeting, the World Conference on Natural Resource Modeling (WCNRM), is designed to be inviting and accessible to scholars at all points in their careers. The recent practice has been to hold meetings in North America in odd-numbered years and outside North America in even-numbered years. The meetings are typically held in locations with some connection to issues in natural resource management. Great care is taken to use inexpensive accommodations and meeting facilities, often university classrooms and residence halls, to keep meeting expenses as reasonable as possible. Paper presentations are organized to avoid parallel sessions, giving the entire membership a chance to hear papers representing a wide variety of perspectives.

The meetings are scientifically serious, representing cutting edge research in all areas of natural resource modeling. Contributed papers are screened by a science committee before acceptance onto the meeting program. At the same time, the social atmosphere of meetings is quite informal, with ample opportunities to meet and chat with other members, cementing bonds that persist past the meeting's close. It is a WCNRM tradition to incorporate an afternoon educational excursion to a nearby location of particular natural beauty or historical interest. It is also frequent practice for the hosts of the WCNRM to organize a premeeting multiday outdoor excursion of some type, be it kayaking, backpacking, or some other adventure to remind members

why they have chosen careers in the wise use and management of natural resources.

The RMA was begun in the early 1980s to promote the use and understanding of mathematical models applied to issues of natural resource management, with particular emphasis on endangered species, fisheries, and forest management on the western coast of North America. The first meeting of the society was held in 1982 at Humboldt State University in Arcata, California. The name of the conference was the First Pacific Coast Resource Modeling Conference (PCRMC). The proceedings of this meeting were published in journal form under the title *Mathematical Models of Renewable Resources*, beginning the RMA's role as a producer of scholarly publications concerning resource modeling.

As the society's membership grew, both in number and geographical breadth, the regional focus of the organization diminished. The RMA cosponsored the First Atlantic Coast Resource Modeling Conference in 1988, and in 1989 began the annual Interdisciplinary Conference on Natural Resource Modeling and Analysis (ICNRMA). The latter drew scientists from around the world and established the RMA as an international presence in the scholarly development, application, and analysis of natural resource models.

In 1994, the decision was made to consolidate these meetings into a single international conference. The first RMA WCNRM was held in 1995 in the Republic of South Africa. These world meetings continue, alternating between North American sites and sites on other continents. Aside from South Africa, the WCNRM has been held in the United States, Canada, Norway, Poland, Spain, Italy, England, Australia, Greece, the Netherlands, and Mexico.

The society's peer reviewed journal, *Natural Resource Modeling* (NRM), succeeded the proceedings of the PCRMC in 1986. The scholarly content of NRM is controlled by an editor and editorial board chosen by the RMA. The actual publication of the journal was undertaken by the Rocky Mountain Mathematics Consortium. The journal grew in both circulation and scholarly reputation.

In 2008, publication rights to the journal were transferred to Wiley-Blackwell. The RMA continues to exercise editorial control of the journal's content. The journal's contents are included in the listings created by Science Citation Index.

The association is a not for profit 501(c)(3) corporation which is incorporated under the laws of South Carolina. It is governed by a Board of Directors comprising four officers and five at-large directors. The President, the Vice President, and the five at-large members of the board are elected directly by the membership. The remaining two officers, the Secretary, and the Treasurer, are nominated by the President and elected by the board. The terms of office for the President, Secretary, and Treasurer is two years. The Vice President's position is filled by either the President-Elect (1 year) or the most recent Past President (1 year), allowing for an overlap in leadership and the preservation of institutional memory.

An Executive Secretary is also an important part of the RMA's leadership. This position carries an indeterminate length and is occupied by a long standing active member of the organization. The Executive Secretary is the public face of the organization, answering queries from the general public and providing important advice for the current officers and directors. The Executive Secretary serves at the pleasure of the Board

of Directors for whatever term the Board feels is advisable.

Lastly, the association has established an honorary title which is given, rarely, to members whose service to the RMA and the broader modeling community has been truly exceptional. RMA Fellows are elected from time to time by the Board of Directors. It is an honor given infrequently, having been given five times in the first 25 years of the RMA's existence.

Since its inception, the RMA's leadership reflects the diversity of its membership, both scholarly and geographically. The list of past Presidents includes: Roland Lamberson (Humboldt State University, 1982–1988), Robert McKelvey (University of Montana, 1989–1990), Colin Clark (University of British Columbia, 1991–1992), John Beddington (Imperial College London, 1993–1994), Wayne Getz (University of California—Berkeley, 1995–1996), Gordon Swartzman (University of Washington, 1997–1998), John Hearne (University of Natal, South Africa, 1999–2000), Anthony Charles (Dalhousie University, Canada, 2001–2002), John Conrad (Cornell University, 2002–2004), William Smith (US Forest Service, 2004–2006), Robert Fray (Furman University, 2006–2008), Steven McKelvey (St. Olaf College, 2008–2010), Keith Cridle, (University of Alaska—Fairbanks, 2010–2011).

RETRIAL QUEUES

JESÚS R. ARTALEJO
Faculty of Mathematics,
Complutense University of
Madrid, Madrid, Spain

INTRODUCTION

Retrial queues consist of a service facility and a pool of unsatisfied customers, called the *orbit*. Customers who cannot receive service leave the service area but they may join the orbit and reattempt for service after some random time. The general mechanism of a retrial queue is shown schematically in Fig. 1. The following motivating examples illustrate some situations that can be modeled as retrial queues.

Example A. Telephone Systems. A telephone subscriber, who obtains a busy signal repeats the call until the demand is satisfied. As a result, the total flow of calls circulating in a telephone network is integrated by the superposition of the flow of primary calls, which reflects the real wishes of the telephone subscribers, and the flow of repeated calls, which is the consequence of the lack of success of previous attempts.

Example B. Random Access Protocols in Communication Networks. Consider a communication line with slotted time which is shared by several stations. If two or more stations are transmitting packets simultaneously, then a collision takes place. As a result, all the involved packets are destroyed and must be retransmitted. The stations would try to retransmit in the nearest slot, but then a collision occurs with certainty. To avoid this, each station independently of the other stations involved in the conflict, may either retransmit the packet with probability p or delay actions until the next slot with probability $1 - p$. In other words, each station

introduces a random delay before the next attempt to transmit the packet.

Example C. Daily Life Situations. We often observe that a person finding a long waiting line prefers to balk and return later. A similar behavior is demonstrated by those impatient customers, who initially entered the line, but they abandon when their patience expires.

In these examples, the access from the orbit to the service facility is governed by the classical retrial policy, where each blocked customer (or unit) retries independently of the rest of customers in the orbit. The literature also includes a few applications operating under other generalized retrial policies. Most retrial queues arise from computer and communication applications, where the repeated attempts occur due to blocking in a system with limited service capacity. This is the case of Examples A and B. In contrast, Example C illustrates the case where the retrial phenomenon is caused by balking or impatience. We also mention the existence of models where retrials are due to breakdowns of the unreliable servers [1].

Apart from their practical applications, the retrial queues raise interesting mathematical problems. Because of the retrial feature, the underlying stochastic process that represents the state of the system is not space-homogeneous. This creates severe analytical problems and allows explicit formulae only in a few special cases. By contrast, the retrial literature is rich in approximation and numerical methods.

Among the vast literature in retrial queues, the existence of two specific books needs to be noted. The monograph by Falin and Templeton [2] provides an excellent account of analytical results for the basic $M/M/c$ and $M/G/1$ retrial queues and some extensions. On the other hand, the main focus of the book by Artalejo and Gomez-Corral [3] is on the application of approximate techniques and algorithmic methods for solving the analytically intractable retrial queues.

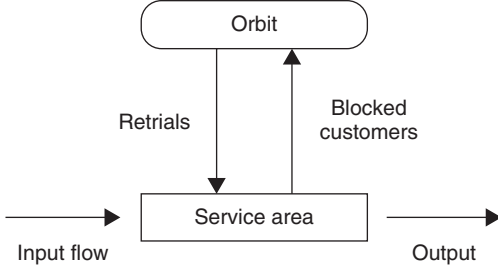


Figure 1. General structure of a retrial queue.

For an overview of the standard $M/M/c$ and $M/G/1$ queues without retrials, we refer to the articles titled *The M/M/s Queue* and *The M/G/1 Queue* in this encyclopedia.

THE BASIC MODELS AND THEIR SOLUTION

In this section, we describe the $M/M/c$ and $M/G/1$ retrial queues and summarize some natural variants and extensions.

We consider a multiserver queue to which primary customers arrive according to a Poisson process with rate λ . Service is rendered by c identical servers with service times exponentially distributed with rate ν . The service facility does not have a waiting space. Customers who find at least one server free upon arrival immediately occupy a position and leave the system after service. In contrast, any arriving customer finding all servers busy will enter the orbit. From there, the retrial customers will compete for service. The inter-retrial times are assumed to be exponentially distributed with rate μ . Moreover, we assume that the process of primary arrivals, services, and retrial times are mutually independent.

At any time t , the state of the process is represented by the bidimensional process $(C(t), N(t))$, where $C(t)$ is the number of busy servers and $N(t)$ denote the number of customers in an orbit. The process $\{(C(t), N(t)); t \geq 0\}$ is an irreducible, regular, continuous-time Markov chain with space state $S = \{0, 1\} \times \{0, 1, \dots\}$. Figure 2 illustrates the transitions among the states for the case $c = 3$. Its infinitesimal transition rates $q_{(i,j),(m,n)}$ are as follows. For

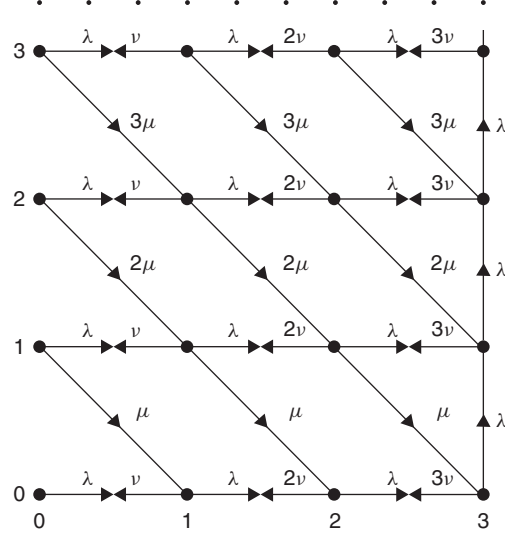


Figure 2. State space and transitions.

$0 \leq i \leq c - 1$, we have

$$q_{(i,j),(m,n)} = \begin{cases} \lambda, & \text{if } (m,n) = (i+1,j), \\ i\nu, & \text{if } (m,n) = (i-1,j), \\ j\mu, & \text{if } (m,n) = (i+1,j-1), \\ -(\lambda + i\nu + j\mu), & \text{if } (m,n) = (i,j), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

and for $i = c$

$$q_{(c,j),(m,n)} = \begin{cases} \lambda, & \text{if } (m,n) = (c,j+1), \\ c\nu, & \text{if } (m,n) = (c-1,j), \\ -(\lambda + c\nu), & \text{if } (m,n) = (c,j), \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In the $M/G/1$ retrial queue, the service distribution is generalized to follow a general law with probability distribution function $B(x)$ ($B(0) = 0$), Laplace–Stieltjes transform $\beta(s)$ and moments β_k . Now, the mathematical model can be viewed as a Markov regenerative process.

The retrial literature is rich so it is possible to find a vast number of variants

and generalizations including batch arrivals, discrete-time models, finite population, multiclass queues, nonpersistent customers, priorities, server breakdowns, and so on [3, Section 2.3]. In investigating variants, it is important to distinguish between basic properties, which follow straightforward extensions of the derivations for the basic $M/M/c$ and $M/G/1$ retrial queues, and specific significant properties of each extended model. Here, we point out the important role played by the matrix-analytic formalism, which provides useful tools to analyze many retrial queues with structured Markov chains (QBD , $GI/M/1$ and $M/G/1$ structures) [3, Chapters 7–9]. In particular, the Markovian arrival process (MAP) extends the Poisson process allowing the correlation between arrivals. On the other hand, the distributions of phase-type (PH) generalize the exponential law, but keeping many interesting properties and without losing computational tractability.

LIMITING DISTRIBUTION OF THE SYSTEM STATE

To ensure the existence of a steady-state distribution, we assume that $\rho = \lambda/cv < 1$ ($M/M/c$ retrial queue) and $\rho = \lambda\beta_1 < 1$ ($M/G/1$ retrial queue). Then, the limiting probabilities $p_{ij} = \lim_{t \rightarrow \infty} P\{C(t) = i, N(t) = j\}$, for $0 \leq i \leq c$ and $j \geq 0$, exist and they are positive.

The limiting probabilities for the $M/M/1$ retrial queue are given by

$$p_{0j} = \frac{\rho^j}{j! \mu^j} (1 - \rho)^{1 + \frac{\lambda}{\mu}} \prod_{k=0}^{j-1} (\lambda + k\mu), \quad j \geq 0, \quad (3)$$

$$p_{1j} = \frac{\rho^{j+1}}{j! \mu^j} (1 - \rho)^{1 + \frac{\lambda}{\mu}} \prod_{k=1}^j (\lambda + k\mu), \quad j \geq 0, \quad (4)$$

and the partial factorial moments $M_k^i = \sum_{j=k}^{\infty} j(j-1)\dots(j-k+1)p_{ij}$, for $i = 1, 2$

and $k \geq 0$, are

$$M_k^0 = (1 - \rho) \left(\frac{\rho}{\mu(1 - \rho)} \right)^k \prod_{j=0}^{k-1} (\lambda + j\mu), \quad k \geq 0, \quad (5)$$

$$M_k^1 = \rho \left(\frac{\rho}{\mu(1 - \rho)} \right)^k \prod_{j=1}^k (\lambda + j\mu), \quad k \geq 0. \quad (6)$$

In the case $c = 2$, the limiting probabilities can be expressed in terms of hypergeometric functions. No explicit formulae for p_{ij} are obtained for the case $c \geq 3$. However, we have the following mean values:

$$E[C] = \frac{\lambda}{v}, \quad (7)$$

$$E[N] = \frac{(v + \mu)(\lambda - v \text{Var}(C))}{\mu(cv - \lambda)}, \quad (8)$$

where C and N denote the steady-state number of busy servers and customers in orbit, respectively. Although an explicit formula for $\text{Var}(C)$ is unknown, formula (8) reduces the computation of $E[N]$ to the variance of the number of busy servers, which could be easily estimated.

A number of interesting theoretical papers are devoted to the analytical solution of the model with an arbitrary number of servers. Since the stationary distribution of the system state is expressed in terms of contour integrals or as limit of extended continued fractions, it is not convenient for practical purposes. More useful is the implementation of approximations and truncated models. This investigation can be classified along three broad lines: approximations, finite truncated models, and generalized truncated models [3, Section 3.4].

The following formulae provide a recursive scheme for the computation of the limiting

probabilities in the $M/G/1$ retrial queue:

$$p_{0j} = \frac{\lambda}{\lambda + j\mu} \pi_j, \quad j \geq 0, \quad (9)$$

$$p_{1j} = \frac{(j+1)\mu}{\lambda} p_{0,j+1}, \quad j \geq 0, \quad (10)$$

$$\pi_j = \sum_{i=0}^j \pi_i \frac{\lambda}{\lambda + i\mu} a_{j-i} + \sum_{i=1}^{j+1} \pi_i \frac{i\mu}{\lambda + i\mu} a_{j-i+1}, \quad j \geq 0, \quad (11)$$

$$\pi_0 = (1 - \rho) \exp \left\{ -\frac{\lambda}{\mu} \int_0^1 \frac{1 - \beta(\lambda - \lambda u)}{\beta(\lambda - \lambda u) - u} du \right\}, \quad (12)$$

where $a_j = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^j}{j!} dB(t)$, for $j \geq 0$, is the probability that exactly j customers arrive during a service time. The sequence $\{\pi_j; j \geq 0\}$ corresponds to the distribution of the embedded Markov chain at service completion epochs.

The two first partial factorial moments are given by

$$M_0^0 = 1 - \rho, \quad (13)$$

$$M_0^1 = \rho, \quad (14)$$

$$M_1^0 = \frac{\lambda\rho}{\mu}, \quad (15)$$

$$M_1^1 = \frac{\lambda^2\beta_2}{2(1-\rho)} + \frac{\lambda\rho^2}{\mu(1-\rho)}, \quad (16)$$

$$M_2^0 = \frac{\lambda}{\mu} M_1^1, \quad (17)$$

$$M_2^1 = \frac{\lambda^3\beta_3}{3(1-\rho)} + \frac{\lambda^4\beta_2^2}{2(1-\rho)^2} + \frac{3\lambda^3\rho\beta_2}{2\mu(1-\rho)^2} + \frac{\lambda^2\rho^3}{\mu^2(1-\rho)^2}. \quad (18)$$

WAITING TIME

We now consider the waiting time, $W(t)$ of a customer who arrives at the system at time t . According to this definition, $W(t)$ means the time that the customer spends waiting at the orbit excluding the service time. We assume that the system has reached the steady-state regime. Then, $W(t)$ is replaced by W , which denote the steady-state spent in orbit. The

distribution of W depends on the service discipline. Retrial queues arising from teletraffic applications operate under a random order access discipline. Random servicing is complicated due to the overtaking phenomena; that is, it is necessary to consider not only the number of customers present in the system at the time of arrival of a customer whose delay is to be investigated, but also the possibility that customers arriving at a later time will compete for free servers.

The waiting time of the $M/M/c$ retrial queue can be approximated by employing the truncated model with finite orbit capacity [3, Section 5.2.1].

For the waiting time of the $M/G/1$ retrial queue with random order of service, the Laplace–Stieltjes transform of W is given by Falin and Templeton [2, Section 1.8]

$$W^*(s) = 1 - \rho + \frac{\lambda(1-\rho)}{s} \int_{L_\infty^*(s)}^1 \frac{(1-u)(\beta(\lambda - \lambda u) - \beta(s + \lambda - \lambda u))}{(\beta(\lambda - \lambda u) - u)(u - \beta(s + \lambda - \lambda u))} \times \exp \left\{ \frac{1}{\mu} \int_u^1 \left(\frac{s + \mu + \lambda - \lambda v}{\beta(s + \lambda - \lambda v) - v} - \frac{\lambda - \lambda v}{\beta(\lambda - \lambda v) - v} \right) dv \right\} du, \quad (19)$$

where $L_\infty^*(s) = \beta(s + \lambda - \lambda L_\infty^*(s))$ is the Laplace–Stieltjes transform of the busy period in the standard $M/G/1$ queue without retrials.

The mean waiting time can be easily obtained with the help of Little's formula:

$$E[W] = \frac{E[N]}{\lambda} = \frac{\lambda\beta_2}{2(1-\rho)} + \frac{\rho}{\mu(1-\rho)}. \quad (20)$$

A direct method of calculation for the second moment yields

$$E[W^2] = \frac{2\lambda\beta_3}{3(1-\rho)(2-\rho)} + \frac{\lambda^2\beta_2^2}{(1-\rho)^2(2-\rho)} + \frac{\lambda\beta_2}{\mu} \left(\frac{2}{(1-\rho)^2(2-\rho)} + \frac{\rho}{(1-\rho)^2} \right) + \frac{2\rho}{\mu^2(1-\rho)^2}. \quad (21)$$

Unfortunately, it does not seem possible to numerically invert the density function of W because the above solution from Equation (19) is derived when s is a real number, but algorithmic methods of numerical inversion typically require evaluation of the Laplace–Stieltjes transform at any desired complex s . In Ref. 3, Section 5.1.1, maximum entropy, truncated and gamma approaches are discussed for the approximate analysis of W in the $M/G/1$ retrial queue.

OTHER PERFORMANCE DESCRIPTORS

The limiting distribution of the system state and the waiting time are the most extensively studied characteristics in queueing theory. However, many other performance descriptors are also of interest. In this section, we collect some results about the length of the busy period, the number of customers served, the number of retrials made by a customer, and the maximum orbit length.

For the study of other, more specific descriptors such as the number of attempts since the last service completion, the successful and blocked events, the server idle periods, and the time to reach a certain orbit limit, the reader is referred to the material in Ref. 3, Chapter 6.

Length of the Busy Period

The busy period of a retrial queue, L , is defined as the elapsed time starting with the arrival of a primary customer, finding an empty system, and ending at the first service completion epoch at which the system is empty again. In the case of the $M/M/c$ retrial queue, the algorithmic analysis of the length of the busy period and computation of its moments can be based on the generalized truncated models [3, Section 4.2.1]. In what follows, we summarize the main explicit results for the $M/G/1$ retrial queue.

The Laplace–Stieltjes transform of L in the $M/G/1$ retrial queue is given by the

following equation [2, Section 1.6]:

$$L^*(s) = \frac{\int_0^{L_\infty^*(s)} \frac{\beta(s + \lambda - \lambda u)}{e(s, u)(\beta(s + \lambda - \lambda u) - u)} du}{\int_0^{L_\infty^*(s)} \frac{du}{e(s, u)(\beta(s + \lambda - \lambda u) - u)}}, \quad s > 0, \quad (22)$$

where $e(s, u)$ is

$$e(s, u) = \exp \left\{ \frac{1}{\mu} \int_0^u \frac{s + \lambda - \lambda \beta(s + \lambda - \lambda v)}{\beta(s + \lambda - \lambda v) - v} dv \right\}, \quad 0 \leq u < L_\infty^*(s).$$

Methods based on the principle of maximum entropy and on the truncation of the orbit are still helpful for the approximate computation of the density function of L [3, Section 4.1.1].

The two first moments are

$$\begin{aligned} E[L] &= \frac{1}{\lambda} \left(\frac{1}{p_{00}} - 1 \right), \quad (23) \\ E[L^2] &= \frac{1}{p_{00}} \left(\frac{1}{(1 - \rho)^2} \left(\frac{2\rho\beta_1}{\mu} + \beta_2 \right) \right. \\ &\quad - \int_0^1 \frac{2}{\lambda\mu(\beta(\lambda - \lambda u) - u)} \\ &\quad \times \left(1 - \frac{\lambda(1 - u)\beta'(\lambda - \lambda u)}{\beta(\lambda - \lambda u) - u} - \frac{1}{1 - \rho} \right. \\ &\quad \left. \left. \times \exp \left\{ \frac{\lambda}{\mu} \int_u^1 \frac{1 - \beta(\lambda - \lambda v)}{\beta(\lambda - \lambda v) - v} dv \right\} \right) du \right). \quad (24) \end{aligned}$$

Number of Customers Served

The number of customers served, I , during a busy period is the discrete counterpart of the length of the busy period. The expectations of both characteristics are related by the simple expression $E[I] = 1 + \lambda E[L]$. The algorithmic analysis for I in the $M/M/c$ retrial queue is basically parallel to the existing approach for L .

In the case of the $M/G/1$ retrial queue, we know the following explicit formulae for the first elements of the probability mass

function:

$$P(I = 1) = a_0, \quad (25)$$

$$P(I = 2) = a_0 a_1 \frac{\mu}{\lambda + \mu}, \quad (26)$$

$$P(I = 3) = a_0^2 a_1 \frac{\lambda \mu}{(\lambda + \mu)^2} + a_0 a_1^2 \frac{\mu^2}{(\lambda + \mu)^2} + a_0^2 a_2 \frac{2\mu^2}{(\lambda + \mu)(\lambda + 2\mu)}. \quad (27)$$

A recursive scheme for the computation of the distribution function $P(I \leq k)$, for $k \geq 1$, can be found in Ref. 3, Section 4.1.2.

The two first moments are given by

$$E[I] = \frac{1}{p_{00}}, \quad (28)$$

$$E[I^2] = \frac{1}{p_{00}} \left(\frac{1 - \rho^2 + 2\frac{\lambda}{\mu}\rho + \lambda^2\beta_2}{(1 - \rho)^2} + \int_0^1 \frac{2\lambda}{\mu(\beta(\lambda - \lambda u) - u)} \times \left(\frac{1}{1 - \rho} \times \exp \left\{ \frac{\lambda}{\mu} \int_u^1 \frac{1 - \beta(\lambda - \lambda v)}{\beta(\lambda - \lambda v) - v} dv \right\} - \frac{(1 - u)\beta(\lambda - \lambda u)}{\beta(\lambda - \lambda u) - u} \right) du \right). \quad (29)$$

Number of Retrials Made by a Customer

The number of retrials made by a customer before entering service, R , is the natural discrete counterpart of W . The distribution of this descriptor has been investigated in Ref. 3, Sections 5.1.2 and 5.2.2, for the truncated versions of the $M/M/c$ and $M/G/1$ retrial queues.

Since $W = \sum_{k=1}^R \xi_k$, where ξ_k denotes the time between k th and $(k-1)$ th retrial attempts made by an arbitrary tagged customer, we find that

$$E[R] = \mu E[W], \quad (30)$$

and, in the $M/G/1$ case, we have

$$E[R] = \frac{\rho}{1 - \rho} + \frac{\lambda \mu \beta_2}{2(1 - \rho)}. \quad (31)$$

Maximum Orbit Length

The maximum orbit size, $N_{\max}$, observed during a busy period is a system descriptor clearly connected to the system congestion. Its algorithmic analysis can be performed for the original $M/M/c$ and $M/G/1$ retrial queues with infinite orbit capacity [3, Sections 4.1.3 and 4.2.3].

In the $M/M/1$ retrial queue, the distribution of $N_{\max}$ has the following explicit expression:

$$P(N_{\max} \leq k) = \frac{\sum_{m=0}^k \frac{m!}{\rho^m \prod_{n=1}^m \left(\frac{\lambda}{\mu} + n \right)}}{\rho + \sum_{m=0}^k \frac{m!}{\rho^m \prod_{n=1}^m \left(\frac{\lambda}{\mu} + n \right)}}, \quad k \geq 0. \quad (32)$$

FURTHER READING

Many textbooks include chapters or sections devoted to retrial queues [4–11] and for a general survey of the main retrial features and literature, the reader is referred to the review papers [12–15]. More concretely, the papers by Falin [12,13] provide a general overview not only of the main models but also of many variants. Yang and Templeton [15] focus on the embedded Markov chain technique and the stochastic decomposition property. The paper by Kulkarni and Liang [14] emphasizes the stochastic monotonicity and optimal control problems. Artalejo and Falin [16] concentrate on the comparative analysis between standard and retrial queues. Finally, Refs 17–19 give bibliographical information.

REFERENCES

1. Sherman NP, Kharoufeh JP, Abramson MA. An $M/G/1$ retrial queue with unreliable server for streaming multimedia applications. *Probab Eng Inf Sci* 2009;23:281–304.
2. Falin GI, Templeton JGC. *Retrial queues*. London: Chapman and Hall; 1997.

3. Artalejo JR, Gomez-Corral A. Retrial queueing systems: a computational approach. Berlin: Springer; 2008.
4. Bocharov PP, D'Apice C, Pechinkin AV, *et al.* Queueing theory. Utrech: Brill Academic Publishers; 2004.
5. Bolch G, Greiner S, De Meer M, *et al.* Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. New Jersey: Wiley-Interscience; 2006.
6. Gross D, Shortle J, Thompson J, *et al.* Fundamentals of queueing theory. 4th ed. New York: John Wiley & Sons, Inc.; 2008.
7. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman and Hall; 1995.
8. Medhi J. Stochastic models in queueing theory. Amsterdam: Academic Press; 2003.
9. Riordan J. Stochastic service systems. New York: Wiley; 1962.
10. Syski R. Introduction to congestion theory in telephone systems. Amsterdam: Elsevier Science Publishers; 1986.
11. Wolff RW. Stochastic modeling and the theory of queues. New Jersey: Prentice-Hall; 1989.
12. Falin GI. Single-line repeated orders queueing systems. Optimization 1986;17:649–667.
13. Falin GI. A survey of retrial queues. Queueing Syst 1990;7:127–167.
14. Kulkarni VG, Liang HM. Retrial queues revisited. In: Dshalalow JH, editor. Frontiers in queueing. Boca Raton (FL): CRC Press; 1997. pp. 19–34.
15. Yang T, Templeton JGC. A survey on retrial queues. Queueing Syst 1987;2:201–233. Correction: Queueing Syst 1989;4:94.
16. Artalejo JR, Falin GI. Standard and retrial queueing systems: a comparative analysis. Rev Math Complut 2002;15:101–129.
17. Artalejo JR. Accessible bibliography on retrial queues. Math Comput Model 1999;30(3-4): 1–6.
18. Artalejo JR. A classified bibliography of research on retrial queues: progress in 1990–1999. Top 1999;7:187–211.
19. Gomez-Corral A. A bibliographical guide to the analysis of retrial queues through matrix-analytic techniques. Ann Oper Res 2006;141:163–191.

REVENUE MANAGEMENT IN THE TRAVEL INDUSTRY

WARREN LIEBERMAN
Veritec Solutions Belmont,
California

Yield Management¹ could easily turn out to be the most important pricing breakthrough of the 20th Century—Professor James Makens, Wake Forest University [1]
... look to yield management at the airlines as the model for future pricing in fields as diverse as health care, telecommunications, consumer financial services, insurance, hotels and government services. Yield management techniques will rewrite the supply-and-demand equations that determine service availability and price [2]
[Yield Management] promises to spearhead a revolution in pricing strategies during this decade [3]
Yield Management is the single most important technical development in transportation management since we entered the era of airline deregulation in 1979—Bob Crandall, former CEO and President of AMR and American Airlines [4]²
We estimate that yield management has generated \$1.4 billion in incremental revenue in the last three year years [for American Airlines]... We expect yield management to generate at least \$500 million annually for the foreseeable future—Bob Crandall [4]
...our competitors used... Yield Management in every one of our markets, and they pushed

us straight towards bankruptcy... Where did we go wrong?... we didn't get our hands around the Yield Management—Don Burr, former CEO of People Express Airlines [5–8]
If I were in charge of the antitrust division, I would damn well see if I could build a case that yield management is predatory pricing—Alfred Kahn, former Chairman of the Civil Aeronautics Board under President Carter and known as the father of airline deregulation [9]
Yield management as it applies to airlines is the control and management of reservations inventory in way that increases (maximizes, if possible)... profitability, *given the flight schedule and fare structure*. [emphasis added] [4]
Revenue Management is the art and science of predicting real-time customer demand at the micromarket level and optimizing the price and availability of products [8].

INTRODUCTION

The methods and information used by firms in the travel industry that govern the range of products they offer for sale, the number of sales they make, and the prices they charge have undergone dramatic change in the past 30 years. As noted in the above quotes, these changes have revolutionized the travel industry and are penetrating other industries as well.

Whereas some of these decisions were once based on intelligent “rules-of-thumb,” they now draw on sophisticated mathematical models. In some cases, the advances in decision support tools in combination with advances in electronic distribution and management of inventory have enabled travel firms to offer and manage products that would not otherwise have been possible. Further, the financial gains made possible by these developments, initially known as yield management (YM) and later as revenue management (RM), have been extraordinary.

This article provides an historical overview of why and how this transformation occurred. Much of the discussion on the

¹As the practice of yield management spread beyond the airline industry, the term changed to revenue management. Except for providing background on the term's evolution, and in keeping with the title of this article, we generally refer to the discipline of forecasting customer demand and available supply in real-time and using that information to control product availability and dynamically set prices so as to maximize profits, as revenue management.

²This quote has become well known and is often cited; see, for example, Refs 8, 41.

early stages of yield management is oriented towards the airline industry, as it provided the initial setting for the early innovations—innovations that soon spread to other industries. As yield management evolved into other industries and became known as *revenue management*, the discipline broadened from optimizing capacity controls (e.g., the maximum number of sales that were allowed at previously established prices) to include pricing optimization (e.g., determining what prices should be). The latter portions of this article are oriented around the expanding practice of the discipline and its adoption by other segments of the travel industry.

Often misunderstood, yield or revenue management has now attained a prominent position among the strategies and tactics travel companies use to obtain a competitive advantage over their rivals. As might be gleaned from the quotes above, it is a relatively new way of doing business and many believe it to be an extraordinarily powerful technology. CEOs of companies that invested in it have extolled its virtues as well as its importance for achieving better financial returns. This article provides an overview of what revenue management is, how it evolved in the travel industry, and why it attracts strong interest.

FROM YIELD TO REVENUE: A HISTORICAL OVERVIEW

It is frequently suggested that the science, if not the practice of yield management originated in the US airline industry in the early 1980s, shortly after it was deregulated in 1978. In fact, the early stages of yield management date back quite a bit further, probably to the 1960s. Until 1985, however, the practice attracted limited attention, was narrow in scope, and remained within the airline industry.

For many years, it would not be unusual for 15% or more of those who purchased airline tickets to fail to show up for their flights. In part, this occurred because unused

tickets were fully refundable.³ For flights in high demand, more tickets would be sold than there were seats on the plane. Known as *overbooking*, this practice was routinely carried out by the airlines but it did not attract widespread attention until Allegheny Airlines did not allow Ralph Nader to board his flight in 1972. Nader sued Allegheny and the lawsuit eventually went to the US Supreme Court in 1976 [10]. Overbooking decisions based on mathematical models that seek to maximize financial returns subject to customer service constraints constitute a core element of many airline revenue management programs. Indeed, the first phase of revenue management essentially consisted solely of overbooking.

Until 2000 or perhaps even later, most airline yield management departments did not set prices, but rather focused on controlling the availability of reservation or seat inventory at different prices with the objective of maximizing net revenue. The actual prices at which tickets could be sold were determined by Pricing Departments. Although Yield Management had no responsibility for setting prices, it determined which fares and how many tickets at each fare could be purchased on each flight. In part, this resulted from the organizational structure of most airlines, whereby the Pricing Department and the Inventory Control or Yield Management Department, were separate.⁴ Indeed, in 1985 when American Airlines launched its Ultimate Super Savers fares—discounts of up to 70% with restrictions such as 30-days advance purchase, 7-day minimum stay, and only partially refundable—enabling the airlines to compete for passengers in radically new ways, its yield management department was not made aware of the new fares until they were announced to the public. This organizational decision resulted

³Today, the frequency with which passengers fail to show up for their flights is much less, as many of the tickets sold are nonrefundable.

⁴During the late 1990s, some airlines began to restructure their organizations, consolidating pricing and inventory functions or having these staff work in a more coordinated manner.

in significant revenue dilution. The fares stimulated so much demand, so quickly, even for flights 6–12 months in the future, that American's yield managers could not react fast enough to reset fare and booking controls on all flights to limit the sale of Ultimate Super Saver fares to appropriate levels. Seats that could have been sold to business travelers who were willing to purchase tickets at higher fares, but would not do so until closer to the date of their flight's departure, were instead occupied by those who purchased fares for up to 70% less. Many flights generated less revenue and profit than they otherwise might have.

Prior to the late 1980s and the adoption of revenue management by travel firms other than airlines, yield management decisions were essentially limited to two decisions:

- the number of reservations to accept for a flight;
- the fares, previously established by the pricing department, that should be offered for purchase to potential customers for a flight at various times prior to departure (the actual setting of fares was thus done outside of yield management; yield management only controlled which of the fares were available for purchase).

As we will soon discuss, one of the most innovative as well as revenue enhancing elements of airline yield management programs was that by introducing purchase restrictions on airline tickets—often termed fences—a

single product could be sold, simultaneously, at multiple prices.

Yield had a specific meaning in the airline industry, being related to revenue in two ways:

1. Revenue per available seat-mile or revenue per available seat-kilometer (RASM and RASK) provided a measure of revenue per unit of capacity.
2. Revenue per passenger-mile or revenue per passenger-kilometer (RPM or RPK) offered a normalized measure of fares paid.

Because the two types of reservation controls noted above had clear and direct impacts on both measures of yield, the term *yield management* seemed apt.

In the late 1980s and early 1990s, as the practice of yield management spread to hotels, car rental firms, passenger railroads, cruise lines, and other segments of the travel industry, the phrase “yield management” was often viewed as airline jargon. Other companies typically focused on revenue and profit, not yield. The term became increasingly controversial. Even as they embraced its concepts, many sought a different label for the practice of revenue maximization through inventory allocation controls.

The search for a new label may well have been accelerated by the airline industry's financial difficulties in 1990, 1991, and 1992. Table 1 highlights the financial difficulties experienced by the largest seven US airlines during these years. With some exceptions,

Table 1. Net Earnings of the Seven Largest US Airlines, 1990–1992 (\$000,000)

Airline	1990	1991	1992	Total
American Airlines	(76.8)	(165)	(735)	(976.8)
Continental Airlines	(273.8) ^a	(340.9)	(229.4)	(844.1)
Delta Air Lines	(259.4) ^b	(239.5)	(564.8)	(1,063.7)
Northwest Airlines	—	(310)	(386.2)	(696.2)
TWA Airlines	(162.3) ^c	48.2	(317.7)	(431.8)
United Airlines	95.8	(494.1)	(956.8)	(1,355.1)
USAir	(427.2)	(284)	(1,223.4)	(1,934.6)
Total	(1,103.7)	(1,785.3)	(4,413.3)	(7,302.3)

^aOperating loss.

^bFor the six months ending December 31, 1990.

^cOperating loss.

the airlines generally lost at least several hundred million dollars each year and sometimes much more than that [11,12]. The seven airlines combined for a loss of more than \$7 billion during these three years.

Revenue management, a name far more industry neutral, began to gain favor. References to revenue enhancement and revenue management replacing yield management can be found as early as 1990 and began to proliferate thereafter [13–14].

For many, the desire to “rebrand” yield management dovetailed nicely with the desire to broaden its scope to include a wider array of reservation inventory controls than used by airlines, explicit recognition of variable costs, and price optimization (that is, the determination of the best prices to offer). For example, cruise ships might offer 8–12 cabin categories; cruise line yield management inventory controls explicitly considered the financial impacts of alternative upgrading (and upselling) policies when taking reservations, such as allowing some customers to make a reservation for a less expensive cabin category but guaranteeing them accommodation in a more expensive cabin category. Yield management also led cruise lines to reevaluate and modify the price differentials between cabin categories, as supply and demand imbalances at the cabin category level were analyzed with greater precision. The wider array of variable costs associated with hotel stays led hotel yield management efforts to consider profit, rather than revenue. Yield management thus expanded to include practices beyond those adopted by airlines.

By 1993, *revenue management* was well on its way to replacing *yield management*. Presentations at multiindustry conferences on this topic included sessions in which speakers differentiated revenue management from yield management. Perhaps as testament to the strength of this conversion, the annual International Air Transport Association (IATA) Conference on Yield Management, initiated in 1988, was rebranded as the IATA Conference on Revenue Management in 1993; in 2001 it experienced further rebranding, becoming the IATA Revenue Management and Pricing Conference. There now seems to be growing

support for relabeling revenue management as revenue and pricing optimization (RPO).

Firms in virtually every segment of the travel industry now possess revenue management programs. And each year, these programs improve. But let us not get ahead of ourselves. Every story has its beginning and this is certainly true of revenue management. To tell it, we begin with the airline industry.

YIELD MANAGEMENT BEGINS

The total number of reservations that can be accepted for a flight is generally greater than the number of seats on a plane. It is known as the *overbooking level*, although there was a time, probably short-lived, when airline staff preferred the term “revenue coordination” due to the negative connotations of overbooking [15]. Airlines generally overbook a flight in anticipation of reservations being cancelled, passengers with reservations who fail to show up for their flights (no-shows), and other reasons. While overbooking has been a standard airline practice for over 50 years, the science of overbooking took a dramatic step forward in the mid-1960s. In that respect, this might well mark the beginning of yield management in the travel industry.⁵

As noted by Rothstein, “[w]hen our American Airlines O[perations] R[esearch] Group initiated its research in 1964, the scant literature on overbooking yielded no models or methods that had been implemented.” A year or two later, however, a newly developed overbooking program was implemented at American, dramatically improving the airline’s ability to forecast oversales (customers with tickets that show up for a flight but for whom no seats are available) and set overbooking levels that allowed the airline to achieve higher load factors on high demand flights, without increasing the number of passengers who were denied boarding [15].

While working in the airline industry in the 1980s, it was the author’s experience that

⁵Our decision to use the developments in the 1960s as the starting point for yield management reflects our emphasis on the science, not simply the practice, of the discipline.

the overbooking models used by some airlines remained relatively rudimentary while those of others had evolved considerably since the 1960s. At American Airlines, the overbooking model developed by Rothstein and his colleagues in the operations research group was replaced in 1976 by one which recognized the revenues and costs associated with overbooking levels and attempted to maximize profitability. This model was further enhanced in 1987 [4].

The science of overbooking involves as many as five forecasts:

1. the probability of reservation cancellations prior to departure;
2. the probability of passengers with reservations failing to show up for their flight;
3. incremental demand when cancellations occur before departure;
4. the cost of oversales (this may not be linear for a flight and also, it varies based on a variety of flight characteristics and the ability of the airline to retain the passenger on a different flight that it operates);
5. the incremental revenue obtained from additional reservations.

For operational reasons, it may be desirable to limit the number of oversales on a flight to a value less than what would be financially optimal. Such constraints have been incorporated into overbooking models.

Forecasting incremental demand (additional demand up to a flight's departure) allows implementing better overbooking levels prior to a flight's departure. The overbooking levels for an optimally managed flight will generally decrease over time; that is, as the number of days before a flight's departure decreases, so too will its overbooking level. The more likely it is that incremental demand will back-fill cancellations, the lower the levels of pre-departure overbooking. If all goes well, the number of reservations holding at departure will be equal to the desired departure-time booking level.

The financial benefits of overbooking were substantial for airlines, as it probably

allowed some of them to increase their load factors on high demand flights by as much as 10 percentage points (e.g., from 85%–95%), and possibly more. An airline with 750 flights per day, an average fare of \$175 per passenger per flight, average flight capacity of 120 seats and an annual load factor of 61% would have annual revenues of approximately \$3.5 billion. If 10–15% of its flights sold out, selling just one additional seat on these flights increases annual revenue by approximately \$4.8–7.2 million. Increasing the load factor of these flights by 10 percentage points could mean an annual increase of \$57–86 million. Putting this into context, in 1981 and 1982, on passenger revenues of approximately \$3.4 billion, American Airlines had operating income of \$44 million in 1981 and lost \$18 million in 1982 [16]. To say that an airline's revenue management capabilities, in terms of the accuracy of its overbooking levels could be the difference between an annual loss or profit, is not much of a stretch.

While the substantial financial benefits to the airlines from overbooking are clear, less visible is the consumer benefit, which can also be substantial. When done well, overbooking allows more consumers the ability to purchase and use their product of choice. When advance notice is given that a flight, hotel, or other service is sold out, but actualized demand turns out to be less than capacity, some consumers will have been deprived of their ability to use that service and will have had to choose something less desirable. They will have been inconvenienced unnecessarily.

Without a doubt, airline overbooking would not yield nearly the benefits it does had the airlines not adopted a business process they initially heavily resisted. On heavily booked flights, it is relatively common for passengers to inquire whether volunteers are needed to give up their seat on the flight in exchange for some form of compensation and a seat on a later flight. Many show clear disappointment when the airline's Gate Agent responds that all passengers with reservations can be accommodated on the flight and no such volunteers are needed.

This was not always the case. Prior to 1978, not only did airlines not seek volunteers, airline executives did not believe that

asking for volunteers would work when more passengers with reservations showed up for a flight than there were seats on the plane. Each airline developed its own rules and procedures for determining who would fly and who would not.

In 1966 and 1967, Professor Julian Simon contacted airlines with an idea for how they could obtain volunteers in exchange for an agreed upon level of compensation. None of the airlines thought much of his proposal, some even being derisive in their replies to him. Rather than let the concept die, he persisted in his efforts to obtain support for it [17]. More than 10 years later, his ideas finally took hold when economist Alfred Kahn became head of the Civil Aeronautics Board. The CAB mandated that airlines adopt a plan for obtaining voluntary oversales [18].

Airlines varied in the types of compensation they offered, some giving travel vouchers for use on future trips while others gave free tickets for a future flight.⁶ The success of each program varied, but overall, the volunteer program enabled airlines to try more aggressively to fill each seat. Between 1978 and 1991, the number of involuntary oversales fell from 6.4 per 10,000 passengers boarded to 1.1 per 10,000, while the number of voluntary oversales during that same period increased from approximately 1 to 15 per 10,000 passengers boarded. As you might imagine, the frequency with which passengers complained about being denied boarding also dropped dramatically [18].

Although the practice of overbooking in the airlines certainly continues, the introduction and proliferation of nonrefundable fares has significantly decreased its benefits. The need for overbooking has also declined as cancellation and no-show rates have declined dramatically.

⁶Offering a credit for future travel may increase an airline's revenue as it can stimulate trips that would not have been taken and the cost of the trip may be for more than the voucher. Compensating passengers with a free trip does not allow for this. Travel credits are now more typical.

YIELD MANAGEMENT EXPANDS

Prior to the mid-1970s, airline fares were highly regulated. Restriction-based discounted fares, so familiar today, were not available. The discounted fares that were available tended to be oriented to specific groups such as children, stand-by passengers, senior citizens, and other class-based categories. The availabilities of these fares were based solely on eligibility.

When the US airline industry was deregulated in 1978, the proponents and the detractors of deregulation envisioned major industry changes. It is safe to say, however, that of the changes that were made possible by deregulation, the fare-related impacts of revenue management were far from anyone's expectations.

In 1977, just prior to deregulation, American Airlines introduced discounted fares known as Super Savers on transcontinental flights and then throughout their system in 1978. These fares were developed in response to an alarming development that began during the summer of 1976. Businesses had sprung up that organized group charters, buying large blocks of seats from airlines at significant discounts and then re-selling them for prices that were lower than the regular fare. Discussing the situation with his staff, Bob Crandall, who was to become president of American Airlines, but was then head of marketing, said, "We're already selling 40 percent of our seats at regular fares, so why can't we figure out a way to sell the other 60 percent at a fare cheaper than what a charter operator can get?" [19].

Super Savers were American's response and other airlines followed American's lead. Super Savers required advance purchase and had a 14-day minimum-stay requirement (a round-trip purchase was thus required). Recognizing that business travelers were willing and able to pay higher fares, the restrictions were designed to prevent business travelers from purchasing them. In addition, on many flights the number of Super Saver tickets that could be purchased was controlled (i.e., only a limited number of sales were permitted) to avoid displacing demand from later-booking business travelers who were willing to pay

higher fares. For the most part, the airlines relied on various “rules-of-thumb” and the judgment and intuition of inventory control staff to set and manage discount allocations. Revenue management was expanding from overbooking to *discount allocation* control, although controlling the number of discount fares sold so as to maximize flight revenue was not carried out with the accuracy or mathematical support that was soon to come.

Nonetheless, the introduction of Super Savers was truly a watershed event in the history of revenue management. It was a unique concept. Not only did it allow the same product to be offered for sale *at multiple prices simultaneously*, but consumers of the products, not the firms that were offering them, were empowered to trade-off purchase and usage restrictions versus price and decide which product to purchase. The number of Super Savers that could be sold on any flight, however, was up to the discretion of the airline.

At this point, it is useful to examine two key points in more detail:

1. Large financial benefits are possible when a seemingly homogeneous product is sold for multiple prices.
2. The extent to which these gains are achieved depends on how well these sales are controlled.

Suppose a flight has a seating capacity of 120, the full fare is \$380 and the discounted fare is \$210, a 45% discount from full fare. As shown in Table 2, if offering discounted fares leads to incremental passengers without reducing the number of full fare passengers, the incremental revenues can be significant. If, in addition to selling 70 full fare tickets, 20 discounted fares can be sold, the flight's revenue increases by

more than 15%. But if some of these discounted fares are sold to those who would have purchased a full fare, the gains are quickly eroded. Even worse, *revenue may decline*. If 5 of the 20 discounted fare tickets are purchased by passengers who would have purchased a full fare ticket, the incremental revenue is only 8.6%. If the discounted fares are sold to 12 of these passengers, revenue actually declines by more than one percent, although the flight's load factor, the percentage of seats occupied, increases by more than 11%. Had the discounted fare been \$125, a 67% reduction, revenue declines when only 5 of the 20 discounted fares are sold to those who would have purchased a full fare. As we shall discuss, this was soon to be a real scenario. Profit levels erode even faster when the variable costs of transporting and servicing passengers are included.

The potential value of optimally managing discount fare sales may be appreciated more easily by analyzing the potential impacts of “small mistakes.” In the example above, the incremental revenue drops by approximately 9% for each passenger that purchases a discounted fare rather than the full fare (i.e., a flight with 69 full fare passengers and 20 discounted fare passengers earns 9% less revenue than a flight with 70 full fare passengers and 20 discounted fare passengers). Taken from that perspective, the financial impact of one less full fare sale on a flight, let alone several, is sufficiently great as to merit careful management.

Now, consider those flights for which demand is sufficiently high to fill every seat. Further, suppose that none of the passengers willing to pay full fare purchases a discount fare. What is the financial impact of displacing a potential full fare passenger by a passenger purchasing a discounted fare? That is, how much less incremental revenue

Table 2. Potential Revenue Impacts of Offering Two Fares

Full Fare Passengers	Discounted Fare Passengers	Load Factor (%)	Revenue (\$)	Change in Revenue (%)
70	0	58.3	26,600	N.A.
70	20	75.0	30,800	15.8
65	20	70.8	28,900	8.6
58	20	65.0	26,240	-1.4

Table 3. Potential Revenue Impact of Displacing Full Fare Passengers

Full Fare Passengers	Discounted Fare Passengers	Load Factor (%)	Revenue (\$)	Incremental Revenue (\$)	Revenue Reduction (%)
70	0	58	26,600	Baseline	N.A.
70	50	100	37,100	10,500	Baseline
69	51	100	36,930	10,330	1.6
65	55	100	36,250	9,650	8.1

is earned if there are 69 full fare passengers and 51 discount fare passengers on the flight, rather than 70 full fare passengers and 50 discount fare passengers?

As shown in Table 3, the flight's incremental revenue decreases by approximately 1.6% for each full fare passenger that is displaced by a discount fare passenger. Perhaps, this does not seem like much. After all, is not 98.4% of perfect still pretty good? Is it really worth the effort to pursue an extra 1.6%? If this airline has 1000 flights per day and 15% of its flights sell out, this "allocation error" results in \$9.3 million less revenue than the airline could have earned. Allocation errors of only several seats per flight on sold out flights lowers its annual revenues by tens of millions of dollars. And while these numbers are only illustrative, they are sufficiently accurate to demonstrate that the financial difference between optimally managing discount allocations and only managing them well is extraordinary. This is why travel companies have been willing to invest heavily in achieving marginal improvements in their revenue management programs. It was a lesson that was quickly learned and acted upon by some, but not all, airlines in the early 1980s.

Although Super Savers provided the needed innovation in the pricing structure, a few more years and an impending financial crisis would be required before the airlines fully recognized and embraced the practice of offering multiple prices with a variety of purchase restrictions to obtain incremental sales and profits. Indeed, it was not just the pricing structure, but also the ability to expertly manage the discount allocations that was also needed; but this was not yet in place.

Fortunately, the work to do this had already begun. Although not aware of how important it would ultimately be, in 1972 the Operational Research Branch of British

Overseas Airways Corporation (BOAC) laid out the underlying theory and insights for using operations research techniques to set discount control allocations to maximize flight revenue when multiple fares are sold. The ideas were presented by Kenneth Littlewood, a member of the team, at a meeting of the Airline Group of the International Federation of Operations Research Societies (AGIFORS). At the end of a paper that was written primarily to document BOAC's efforts to implement mathematical and statistical methods to set overbooking levels to maximize revenue, Littlewood also described a method for using demand forecasts and optimality conditions to set limits on the number of discounted fares that should be sold for a flight to maximize flight revenues. According to Littlewood, the ideas on discount allocation were "still in its infancy," but the introduction of new low fares by Skytrain for regular London–New York service highlighted the need to control the allocation of low yield fares [20,21]⁷. He could not have envisioned that this method, eventually termed *Littlewood's Rule*, was to become the basis for discount allocation algorithms employed in revenue management systems at many airlines.

Littlewood's key observation (somewhat simplified here) was that to maximize a flight's revenue when two fares are offered (e.g., let Y be the full fare and D the discount fare), the discounted fares should be sold so

⁷The paper was entitled "Forecasting and control of passenger bookings" and was published in the AGIFORS 12th Annual Symposium Proceedings, October 1972, pp. 95–117. Fortunately, this ground-breaking and quite readable paper has been reprinted and the reference listed at the end of this article is far easier to access.

long as

$$D \geq P \times Y \text{ or, equivalently, } P \leq D/Y$$

where P is the probability that selling an additional discounted fare will result in one less full fare ticket sale.

This implies that discounted fares should be sold so long as the ratio between the discount fare and the full fare (D/Y) is at least as great as the likelihood that the overbooking limit will be reached if no more discounted fares are sold. Essentially, it is a comparison between the revenue received from a discounted ticket and the expected marginal revenue resulting from not selling the discounted fare in anticipation of selling a full fare in its place. It is interesting to observe that the decision as to whether or not to accept an additional reservation for a discount fare does not depend on the demand distribution for discount fares, but it does depend on the demand distribution for full fares.

In 1980, Bill Swan, a member of the American Airlines Operations Research Department, modified and extended Littlewood's methodology for implementation at American Airlines. Such a system was implemented in 1982 [4]. American was probably the first US airline to implement such an advanced method for determining discount allocations. This method was extended to multiple fare classes and has come to be known as the expected marginal seat revenue (EMSR) approach. The earliest published work on this method, including extensions and modifications of the approach can be found in Refs 22–25.

Discount allocation controls were designed to work within airline reservation system control structures. The mechanisms by which reservation systems controlled fare availability differed from that of many other industries. Each fare would be mapped into one of perhaps five or eight fare classes for a flight, depending on the reservation system. Fare classes were designated by letter codes. For example, F, C, and Y were commonly used for first class, business class, and full coach fares, respectively. Discounted coach fare classes included letter codes M, B, V, H, and Q. When

a fare class was available, all fares that were mapped into it could be sold.⁸ When a fare class was closed, none of the fares mapped into that fare class could be sold. Typically, many fares were assigned to each fare class [4,22]. Over time, the number of fare classes has increased considerably. A flight may now have as many as 16 different fare classes.

Consequently, yield management assumed an indirect responsibility for price setting. While others determined the actual fares that *could* be offered to the public, yield management controlled when and which fares were *actually* offered by determining when fare classes were open for sale and when they were closed. Fare classes opened and closed multiple times during the booking period reflecting bookings, cancellations, and revisions to discount allocation decisions based on changes in incremental demand forecasts.

As you might imagine, demand forecasts play a critical role in setting inventory allocations. Demand was forecast at the fare class level. Time series models, such as exponential smoothing and booking profiles that provided estimates of how many reservations were typically received during various time periods prior to departure, were commonly relied upon to forecast demand by fare class. Consequently, the forecasts of future or incremental demand were independent of potential yield management or pricing actions (except to the extent that analysts implemented manual overrides). Even with such simplifications, the forecast models performed well.

With a pricing structure in place that allowed airlines to target fares at multiple market segments and the ability to control the availability of these fares based on science rather than human intuition and judgment, the airline industry, or perhaps more specifically American Airlines, was prepared to

⁸In some cases, fare-specific restrictions prevented a fare from being sold when its fare class was available. For example, if a fare with a 21-day advance purchase restriction was one of the fares that was mapped into the V fare class, that fare could not be sold 18 days prior to the flight, even if the V fare class was still open.

demonstrate that revenue management could transform the competitive landscape in ways that no one had previously imagined. All that was needed was the catalyst for action.

That catalyst came in the form of People Express, a low-cost airline that had begun operations in 1981 by offering low-fare short-haul service between cities that had limited air service. For \$23, you could fly from Newark, New Jersey to Buffalo, New York; for \$35 you could fly from Newark to Norfolk; these fares were 50–80% lower than those offered by other carriers. Over time, however, People Express expanded its route structure and by 1984 was competing on the same routes that were flown by major carriers such as American Airlines. Faced with such competition, the major carriers faced a dilemma: matching the fares offered by People Express would enable them to keep their customer base but not allow them to cover their costs, whereas not matching the low fares would result in the loss of too many passengers [26].

American's competitive response came in January 1985, in the form of its Ultimate Super Savers, restricted fares that were discounted as much as 70% so as to be competitive with those of People Express. The fares provided far greater discounts than Super Savers. Also, new restrictions such as limited refundability levels were included. Although discount allocation controls in the reservation system were now automated and benefited from data feeds from decision support systems, American's initial ability to optimally manage the discount allocation levels left much to be desired simply because its historical passenger demand data did not reflect the level of demand for the new fares. With each passing month, however, American's ability to maximize its revenues by controlling and managing the availability of these fares would improve. Other airlines would also ramp up their investments in methods and systems to better control discount fare availability. The stage was set for the introduction of a wide variety of discounted, restricted fares whose availability would seem to appear and disappear as if by magic and for no valid reason. With that, the

discipline known as yield management would soon enter the public lexicon.

The business community did not yet understand the potential benefits of the new pricing structure. As reported by Time, "To Wall Street, American's move looked like an attempt to get out in front of the competition by jumping off a cliff. Many airlines are already losing money, and discount fares may mean bigger losses. Stock prices of all the large airlines that joined in last week's fare war fell sharply" [27]. In fact, however, American enjoyed great financial success during the next few years; revenue management was only one of the reasons for this, but it was an important factor.

Although it had a more gradual impact on revenue management, the Airline Deregulation Act of 1978 also transformed airline routes. Given much greater flexibility in determining their routes and schedules, major airlines developed the hub-and-spoke system in the 1980s. Major airlines reduced their frequency of nonstop service between small and mid-sized cities (sometimes dropping these routes), forcing more passengers to take connecting flights. For example, whereas about 10% of American's passengers took connecting flights in 1980, by the mid-1980s about two-thirds of the passengers going into a hub airport were connecting to another flight [4]. By the late 1980s, airlines began to make fare allocation decisions that reflected passenger itineraries: this aspect of revenue management is known as *traffic management*.

For airlines that experienced significant increases in their volume of connecting passengers, the traffic management aspect of revenue management was significant. When most passengers fly point-to-point, airlines that attempt to maximize their revenues/profits by setting discount allocation levels on a flight-by-flight basis perform rather well. As connecting traffic increases, however, this practice becomes more problematic. Indeed, as we shall soon see, in some industries such network-related issues can dominate revenue management decisions. Traffic management, that is the practice of setting allocation levels (and not necessarily only discounted fare allocations)

that reflect the incremental value of a passenger's revenue to the airline's flight network, became increasingly important and eventually became incorporated into airline revenue management systems.

The importance of traffic management can be seen in the following example. Consider a person wanting to fly from Austin, Texas to London, England on a discounted fare of \$329. As there is no nonstop service from Austin to London, this passenger might fly from Austin, Texas to Dallas/Forth Worth, and then from Dallas/Fort Worth to London. Suppose someone else wants to fly from Austin to Dallas/Fort Worth (DFW) on the same flight and is willing to pay the Austin to DFW full fare of \$189. Further, suppose there is only one seat still available for sale on the flight from Austin to DFW. Which reservation should the airline accept if it wants to maximize its revenues?

Excluding variable costs and the likelihood of either passenger booking a different flight on the same airline, the answer is, it depends.⁹ Specifically, if the flight from DFW to London becomes full and demand must be turned away, it is highly likely that the airline earns greater revenue by transporting the passenger flying from Austin to DFW and turning down the passenger flying from Austin to London. If, however, the flight from DFW to London departs with empty seats, the airline is likely to earn greater revenues by transporting the passenger flying from Austin to London.

Compared to overbooking and discount allocations, traffic management is significantly more difficult to address and generates more modest incremental revenue (although certainly significant) than setting flight-independent discount allocation levels. Although the benefit of overbooking has declined in recent years as nonrefundable fares have become more prevalent, in the

1980s and 1990s, approximately 40–50% of the potential benefits of revenue management could be attributed to overbooking, 30–40% to discount allocation, and 10–20% to traffic management.¹⁰ While 10% of a \$500 million annual benefit is certainly significant, it made sense for many airlines to initially focus on overbooking and discount fare allocation, the more valuable areas, prior to traffic management. For other industries, however, this was not the case.

FROM YM TO RM: BEYOND THE AIRLINE INDUSTRY

As revenue management spread to the hotel industry and then elsewhere, its focus remained on inventory control. Prices continued to be set independently of revenue management. Marriott began investing in revenue management as early as 1985 [28] and other hotel companies soon followed. According to Sheraton's Director of Marketing Geoff Ballotti, Sheraton began implementing a hotel revenue management system offered by Control Data Corporation (CDC) in 1987, implementing the system at 28 of its hotels [5–7]. Hilton also implemented a revenue management system in the late 1980s.

Some hotel chains began to experiment with offering discount rates with restrictions such as advance purchase requirements, but for the most part, it would take several more years until hotels integrated advance purchase requirements and refundability restrictions into their pricing structures. Rather, discounted rates were targeted to specific dates (e.g., weekends, holidays) or groups (e.g., children, travel agents, senior citizens) or were unrestricted rates off the full rate. The discounted rates were tied to the full rate (known as the rack rate) by some pre-determined percentage or value. Consequently, the early hotel revenue management systems served a slightly different function than airline systems, as they were used to

⁹Such factors should, and are likely to, be considered in an airline's revenue management system although it was more likely than not for such factors to be ignored or only marginally considered in the initial revenue management systems implemented by the airlines.

¹⁰Personal communication with Barry Smith, former Chief Science Officer for SABRE and also the author's experience.

control which of several rate tiers should be offered, rather than controlling discount allocations. Each rate tier included a set of rates.

The revenue management systems measured demand pressure for each check-in date using four primary criteria: reservations on the books, time until check-in date, a forecast of incremental demand for that date, and a forecast of remaining rooms. As the magnitude of the demand pressure increased, these systems would recommend raising the rate tier offered. Consequently, many of the various rates offered by the hotel to the general public would increase, or decrease, simultaneously, by pre-determined amounts. Some discounted rates would not be included in a rate tier, so these rates would not be available when that rate tier was offered.

The initial application of revenue management in the hotel industry suffered from missteps that concerned some hotel executives sufficiently to delay their adoption of revenue management systems. Borrowing from the airline industry's revenue management "playbook," perhaps a little too much, some of the initial hotel revenue management systems incorporated a design simplification that served the airlines well; they essentially ignored traffic management, the network aspect of revenue management.¹¹ Revenue managing a flight independently of others is similar to revenue managing arrival dates at a hotel independently of other dates. Unfortunately, this simplification was apt to yield poor rate recommendations during peak periods for hotels with a high proportion of multiple night stays. When guest stays span multiple dates, managing dates independently is apt to lead to pricing mistakes.

As it turns out, compared to increasing rates on dates that ultimately sell out, hotels generally earn far greater revenues and profits by limiting the number of shorter stays so that they can accept a greater number of longer stays that span these nights. This is

typically referred to as optimizing by *length-of-stay* (length-of-rent in the car rental industry). It is far more profitable to earn \$149 per night for four nights than \$229 for a one-night stay. Recommendations to close the lower rate tiers for a peak night to all guests, regardless of whether they wanted to stay for one, four or seven nights, could easily discourage those guests who wanted to stay several nights or longer. Unfortunately, the early hotel revenue management systems were not designed to analyze demand patterns across dates and address this issue. Nor were many hotel reservation systems designed with the controls needed to implement such decisions. The systems were not designed to control room inventory in ways that would stimulate or accept demand from guests wanting longer-stays, while simultaneously limiting the number of shorter-stays on peak nights.

A simplified example illustrates this trade-off. Consider a hotel with a rate of \$149/night, for which there is sufficient demand to sell-out on Wednesday night, even if the rate increases to \$229/night. As shown in Table 4, increasing the rate from \$149 to \$229 on Wednesday, an increase of more than 50%, results in 1135 occupied rooms during the week and the weekly revenue exceeds \$193,000. Further, every room is occupied on Wednesday night.

Conversely, if the hotel implements length of stay (LOS) controls so that more of the reservations it accepts that span Wednesday night include the surrounding days, and does not increase its average rate on Wednesday, the number of rooms occupied during the week increases by 195. The weekly revenue for the hotel increases to \$198,000, about a 2.6% revenue increase. In this scenario, not all of the hotel rooms are occupied on Wednesday night, reflecting the uncertainty associated with holding back rooms for longer stay reservation requests. In actual implementation of length of stay controls, some hotels have claimed revenue increases of 8–10% or even more when compared to increasing rates on peak nights [29]. Indeed, in the above example, if the Wednesday rate had only increased to \$189, an increase of approximately 25%, the hotel would earn

¹¹Nor were overbooking recommendations incorporated into these systems, although this was more likely for business process reasons. Hotel executives tend to be very cautious about systematically overbooking.

Table 4. Comparison of Impacts from LOS Controls versus Increasing Price

	Mon	Tue	Wed	Th	Fri	Total
Available rooms	300	300	300	300	300	1500
Occupied rooms if Wednesday rate increases to \$229	190	235	300	220	190	1135
Revenue if Wednesday rate increases to \$229 (\$000)	28.3	35.0	68.7	32.8	28.3	193.1
Occupied rooms with LOS controls and rate of \$149	245	280	295	275	235	1330
Revenue with LOS controls and rate of \$149 (\$000)	36.5	41.7	44.0	41.0	35.0	198.2

9.4% more revenue by using length-of-stay controls rather than increasing its rate.

In practice, combining a rate increase for shorter stays while allowing longer guest stays to access lower rates is generally the most profitable pricing strategy for hotels during peak periods. The benefits of this strategy are recognized within the hotel industry. Hotel systems and practices have been redesigned to enable this strategy to be effectively implemented: no longer can guests make a reservation for a longer stay to obtain a lower rate and then cancel some of the days in the reservation or simply check-out early.

Aggressively increasing rate tiers on peak nights also had a negative unintended side effect. Able to provide hotels with high volumes of demand over the year, many corporations negotiated fixed rates with hotels that were significantly lower than the unrestricted rates. The corporate negotiated rates were not affected by the hotel's pricing actions. Raising the hotel's rates on peak nights might reduce hotel profitability as it could result in accepting more guests with lower corporate rates; because the rate increase slowed down the pace of bookings for publically available rates, it enabled corporations with the lower negotiated rates to access more rooms on peak nights, which was exactly the opposite of what the hotels wanted.

By 1990 these shortcomings were recognized and soon thereafter, addressed. Marriott's first generation yield management system with length-of-stay optimization was implemented in 1991. Prior to this, Marriott's yield management systems had gone through two generations. A test of the system at the Munich Marriott during a high demand

week helped the hotel obtain a 12.3% year-over-year revenue increase, despite an 11.7% reduction in average daily rate [5–7].

The adoption of yield management by the hotel industry in the late 1980s and early 1990s is well documented. Less information is available about the passenger railroad and cruise line initiatives at this time. Given permission to provide consulting and system development expertise to other firms, the American Airlines Operations Research Group began developing a yield management system for Amtrak in 1987 and a year later was doing the same for Royal Caribbean Cruises Ltd. (RCCL). Focused on enhancing the inventory control capabilities of these firms, these systems did not provide any pricing recommendations.

Among the first areas to be addressed by the RCCL effort was improving the accuracy of pre-departure cancellation rate estimates for individual and group reservations [30]. At the time, cruise lines typically estimated cancellation rates based on fixed projection rates that reflected how much money had been received by the cruise line and possibly one or two other factors. Known as a handicapping formula, the methods generally provided reasonable average estimates over the course of the year, but were subject to "catastrophic" breakdowns on individual sailings, leading to undesirable pricing actions. Over-forecasting cancellations could lead to selling cabins at highly discounted rates close to departure date. Under-forecasting cancellations could require upgrading passengers into much more expensive cabins, or even worse, fully or partially refunding their purchases while also moving them to other departures. Segmenting reservations based on a variety

of attributes (e.g., travel agency, itinerary, and type of group), and containing updating mechanisms to capture trend and seasonality, the new method provided more accurate cancellation rate estimates [31]. Having more accurate estimates of the number of additional cabins that needed to be sold on a departure enabled RCCL to make more profitable pricing decisions for each cruise.

Other elements of the RCCL yield management initiative included a group evaluator model and a variety of forward-looking reports directing analyst attention to those departures most in need of oversight and action. While the scope of RCCL's initial yield management system might be considered rather modest, it was a key factor in enabling RCCL to better manage and price its inventory, including implementing a variety of pricing innovations. In 1992, Brian Rice, then RCCL's director of revenue planning and analysis, and now CFO, conservatively estimated the incremental revenue from these efforts at over \$20 million/year, or about 2–3% of RCCL's annual revenue [5–7].

Approximately two years after Amtrak had initiated its revenue management effort, SNCF, the French national railway, did the same [32,33]. By 1989, revenue management efforts were also initiated at Avis, a rental car company. By the early 1990s all of the major car rental firms had implemented revenue management systems or were preparing to do so. The ability to evaluate and control reservations based on length-of-rent was included in all the first generation car rental revenue management systems. The revenue management system implemented at Hertz also included fleet planning and fleet deployment capabilities enabling Hertz to better determine where and when it was financially advantageous to transfer cars between rental car locations due to projected supply and demand imbalances [34]. Doing so probably made Hertz the first company to integrate revenue management with operational decisionmaking.

Revenue management continued to spread to other segments of the travel industry, including tour operators, time-share exchange, theme parks, and even yacht rentals as well as industries other

than travel. Within a few years, additional revenue enhancement techniques were included in revenue management initiatives. Perhaps the most critical development was the inclusion of price optimization capabilities; no longer would revenue management simply control the availability of previously set prices. Instead, revenue management efforts would include determining the best price levels. In some cases, dynamically re-pricing a product to maximize profits based on current and forecast supply and demand extended the underlying mathematical models to include demand elasticity estimates: forecasts of how demand would change in response to offering alternative prices. Some (i.e., published) accounts of revenue management systems that incorporated dynamic pricing capabilities include National Car Rental [35], The Moorings [36], and Princess Cruises [37].

The revenue management initiative for Princess Cruises, begun in 1997, was implemented in 2000. While the system included typical inventory controls, it also included a pricing analytics engine providing insights on the extent to which promotional prices were needed to stimulate sales on a departure. Of all the capabilities of the revenue management system, Princess staff believed this capability to be the most beneficial.¹²

Price optimization continues to be of great interest to many travel industry firms. Only a relatively few, however, have invested significant resources. At least for some of these firms, however, the returns have been extraordinary, easily eclipsing the 4% to 7% returns typically ascribed to the inventory allocation controls of revenue management. Anecdotally, price optimization capabilities appear to be capable of generating revenue increases of 10–15%.¹³

For many years, pricing has been treated as mostly art and very little science. For the most part, the science of pricing focused on

¹²Personal communication with Princess Cruises staff.

¹³These estimates reflect the author's experience as well as personal communication with Barry Smith.

estimating the costs of production and building in a reasonable profit margin. The product's price was simply the result. While many firms continue to engage in that practice, it occurs less often as the principles of revenue management obtain a greater foothold. The systematic integration of supply and demand forecasting with product design and market segmentation concepts has proven successful, although there have certainly been instances of failure and setbacks along the way. As was the case when revenue management was initially applied to the hotel industry, many of these failed efforts resulted when techniques that proved successful in one industry were used in another, without sufficient regard for the differences in the way business was conducted.

Over the past 20 years, the financial benefits of value-based pricing have become more widely accepted and communicated. The science of revenue management provides a powerful mechanism that enables value-based pricing to succeed. Indeed, with restriction-free pricing becoming more and more prevalent in the airline industry, the traditional discount allocation controls of revenue management are now giving way to an increasingly strong focus on maximizing revenue via dynamic pricing to achieve pricing optimality.

At the heart of any dynamic pricing optimization system is forecasting demand at alternative prices and then estimating the prices that will maximize profit. From a practical perspective, forecasting demand at different prices can be both complex as well as data intensive, as demand is affected by a number of factors other than price. While many companies have attempted to incorporate such variables into regression-based forecasts, such efforts often produce less than satisfactory results. Data sparsity and the inability to model many key relationships among the variables are some of the reasons that price-demand relationships are less robust or accurate than desired. Manual overrides and factors designed to prevent forecasts and recommendations from being too extreme are often integral to such systems. While some of these methods have proven useful, others have produced

sufficiently unreliable recommendations that firms stopped using the systems.

More recently, however, alternative modeling approaches such as disjunctive mapping are being developed to address these issues [38,39]. The widespread use of robotic pricing tools that capture competitor prices enables price-sensitive demand forecasts to explicitly reflect different competitive situations and support what-if analyses, as historical data has become available to estimate the likelihood of alternative competitor reactions. Indeed, one of the challenges of price-elasticity modeling is that in practice, demand curves may not be static; they depend on the directional shift of price and possibly other factors. For example, when price increases from A to B , the magnitude of the change in demand level may not be equivalent to the magnitude of the change in demand when price decreases from B to A . Such modeling and estimation complexities are more easily addressed when competitor prices are retained. Given the potential benefits of dynamic pricing, and the recent availability of data to support demand forecast modeling, we anticipate that dynamic pricing will be a strong focus of travel companies during the next decade or two.

CONCLUSION

Written for academics and experienced practitioners of revenue management as a "single-source reference for the major theory and application issues," a comprehensive treatment of revenue management was published in 2004 by Kalyan Talluri and Garrett Van Ryzin. The book is intended for those who have advanced degrees in mathematically-oriented disciplines such as operations research, statistics, or economics, although portions of the book are accessible to those without such skills and are interested in obtaining a more detailed understanding of the theory underlying revenue management [40].

From a relatively obscure business practice only 50 years ago, revenue management has come a long way. Dominated by practitioners until 1990 and perhaps longer,

the number of academic researchers may now surpass the number of practitioners. Indeed, the subject now boasts two dedicated journals: *The Journal of Revenue and Pricing Management* was initiated in 2002 and was joined by the *International Journal of Revenue Management* in 2007. In addition, articles on revenue management appear in a wide variety of journals and books.

Will the term revenue management be retained? Will it be replaced by pricing and revenue optimization or something else? What is your forecast?

Acknowledgments

I greatly appreciate the time given to me by both Kenneth Littlewood and Barry Smith to discuss some of the ideas in this article. They were able to provide me with key pieces of information that filled in gaps in the historical record and allowed me to better understand the motivation behind industry developments. As my mentor in revenue management at American Airlines, I am deeply indebted to Barry for both, the time he has given to me and value the friendship we have developed. I also owe a large debt of gratitude to Tom Cook, who headed the American Airlines Operations Research Department in 1984. Without his support and trust, I would never have had the opportunities in revenue management that have captivated me for more than 25 years.

REFERENCES

- Orkin EB. Yield management. Hotel & Catering Technology; 1988. (Reprinted by Hotel Information Systems Ltd., Heathrow, Hounslow, England.)
- Schrage M. New corporate pricing tool could be costly to consumers. California (CA): Los Angeles Times. 1991 Dec 5: D1.
- Lieberman WH. Yield management: the pricing revolution that goes beyond the travel industry. Presentation at the 4th Annual US Pricing Conference, The Pricing Institute, a division of The Institute for International Research; 1991 April 10–12. New York: The Grand Hyatt Hotel; 1991. p. 3.
- Smith BC, John L, Ross D., Yield Management at American Airlines. Interfaces 1992;22(1):8–31.
- Scorecard. Marriott's Millions. Atlanta (GA): Aeronomics Incorporated; Premier Issue, Second Quarter. 1992a. pp. 3–6.
- Scorecard. Royal Caribbean breaks through. Atlanta (GA): Aeronomics Incorporated; Third Quarter. 1992b. pp. 5, 11.
- Scorecard. Sheraton's technological evolution. Atlanta (GA): Aeronomics Incorporated; Fourth Quarter. 1992c. pp. 3–5.
- Cross RB. Revenue management: hard-core tactics for market domination. New York: Broadway Books; 1997. p. 276.
- Uchitelle L. Airlines off course. New York: New York Times Magazine; 1991. p. 612.
- Time. The law: a big bump for bumping. 1976. Available at <http://www.time.com/time/magazine/article/0,9171,918209,00.html>. Accessed 1976 June 21.
- Gowen MJ. Airline and cargo industry update. Murray Hill (NJ): DUNS Analytical Services, Dun & Bradstreet, Inc.; 1991. p. 33.
- Business Travel News. Airlines. 1993. pp. 40, 42.
- Lieberman WH. A revolution is brewing in pricing. Los Angeles: Los Angeles Times; 1990. (Looking Ahead), June 6, 1990.
- Cross RG, Lieberman WH. Yield management. Futurescope, management and society. Cambridge (MA): Decision Resources Inc.; 1991. pp. IV–1–IV–4.
- Rothstein M. OR and the airline overbooking problem. Oper Res 1985;33(2):237–248.
- AMR Corporation. 1984 Annual Report. 1985. p. 44.
- Simon JL. An almost-practical solution to airline overbooking. J Trans Econ Policy 1968;2(2):201–202.
- Simon JL. The airline oversales auction plan: the results. J Trans Econ Policy 1994;28(3):319–323.
- Sterling R. Eagle: The Story of American Airlines. New York: St. Martin's Press; 1985, p. 482.
- Littlewood K. Forecasting and control of passenger bookings. J Revenue Pricing Manage 2005;4(2):111–123.
- Littlewood K. Personal communications. 2010;
- Belobaba P. Air travel demand and airline seat inventory management [Doctoral dissertation]. Flight Transportation Laboratory. Report R87-7. Cambridge (MA): MIT; 1987; p. 214.

23. Belobaba PP. Survey paper – airline yield management: an overview of seat inventory control. *Trans Sci* 1987b;21(2):63–73.
24. Belobaba PP. Application of a probabilistic decision model to airline seat inventory control. *Oper Res* 1989;37(2):183–197.
25. Curry RE. Optimal airline seat allocation with fare classes nesting by origins and destinations. *Trans Sci* 1990;24(2):193–204.
26. Phillips RL. Pricing and revenue optimization. Stanford (CA): Stanford University Press; 2005. p. 355.
27. Time. Sky wars: airfares take a dive. 1985. Available at <http://www.time.com/time/magazine/article/0,9171,959288,00.html>. Accessed 1985 Jan 28.
28. Hsu CH, Powers TF. Marketing hospitality. New York: John Wiley & Sons, Inc.; 2002. p. 360.
29. Aeronomics. A conversation with Don Burr. Fourth Quarter. Atlanta (GA): Scorecard; 1992. pp. 6–7.
30. Fisher J, Mongalo Marco. Integrating decision support. SAS Users Groups International Conference. New York; 1993. pp. 619–623.
31. Lieberman WH. Pricing tactics and strategies in the cruise industry. *Handbook of pricing management*. Oxford University Press; 2010. In press.
32. Daudel S, Georges V. Le yield management: la face encore cachée du marketing des services. InterEditions. Paris: 1989.
33. Mitev NN. The globalisation of transport? Computerised reservation systems at American Airlines and French Railways. In: Lynch P, Trischler H, Lyth J, *et al.* editors. *Wiring Prometheus: globalization, history and technology*. Denmark: Aarhus University Press; 2004. pp. 193–216.
34. Carroll WJ, Grimes RC. Evolutionary change in product management: experiences in the car rental industry. *Interfaces* 1995;25(5): 84–104.
35. Geraghty MK, Johnson E. Revenue management saves national car rental. *Interfaces* 1997;27(1):107–127.
36. Bermudez R, Dieck T, Lieberman W. Revenue management basics in the charter boat industry. *Revenue management and pricing: case studies and applications*. London: Thomson; 2004. pp. 1–8, 184–188.
37. Scorecard. Making history. Mountain View (CA): Talus Solutions; 2000. pp. 5–8.
38. Raskin M, Lieberman W, Mullin J. Disjunctive mapping: changing the way we understand and predict customer behavior (part one). *J Revenue Pricing Manage* 2010. In press.
39. Raskin M, Lieberman W, Mullin J. Disjunctive mapping: changing the way we understand and predict customer behavior (part two). *J Revenue Pricing Manage* 2010b. In press.
40. Talluri KT, Van Ryzin GJ. The theory and practice of revenue management. Boston (MA): Kluwer Academic Publishers; 2004. p. 712.
41. Cowan AL, Gargan EA. Mirage of discount air fares is frustrating to many fliers. *New York: New York Times*; 1991. pp. A1, C3.
42. Simon J. Origins of the airline oversales auction system. *Cato Rev Bus Gov*. Available at <http://www.cato.org/pubs/regulation/regv17n2/reg17n2-simon.html>.

REVENUE MANAGEMENT WITH INCOMPLETE DEMAND INFORMATION

VICTOR F. ARAMAN
Olayan School of Business,
American University of Beirut,
Beirut, Lebanon

Stern School of Business, New
York University, New York,
New York

RENÉ CALDENTY
Stern School of Business, New
York University, New York,
New York

Consider a seller who is endowed with a fixed number of units of a product that he/she can sell to a price-sensitive and stochastically arriving stream of consumers during a given selling season. The seller's problem is to dynamically adjust the product's price to maximize the revenues he/she can collect if no replenishment is possible during the sales horizon. This setting is quite typical of many industries including among others, airlines (selling seats on a specific flight), hotels (booking rooms on a particular night), and retailers (selling seasonal merchandise). Often, these problems are labeled as *revenue management* problems since operational decisions are driven solely by revenues; inventory and capacity costs are sunk and incurred independently of changes in prices and/or number of units sold. The assumption of a fixed capacity is by no means critical if we consider that in most of these industries capacity is flexible only in the long run. Moreover, capacity decisions and price decisions take place on different timescales. Issues regarding the type of airplanes to schedule on a particular route, or the number of rooms to build in a given hotel, or the amount of seasonal merchandise to purchase from an overseas supplier are decided long before demand is realized and price policies are implemented.

In this article, we discuss, specifically, the revenue management problem alluded to in the previous paragraph using a stylized mathematical formulation. We focus on the research that considers the case in which the seller has incomplete demand information. That is, there are some characteristics of the demand process (e.g., the arrival rate or the price elasticity) that the seller does not know with certainty. Through this discussion we aim at summarizing some of the fundamental theoretical results of the existing literature. The model is simplistic but, we believe that the methods used and the insights drawn are quite representative of more general models and various other settings related to revenue management and, more broadly, to inventory management. Our exposition is by no means exhaustive and we refer the reader to Refs 1 and 2 for a comprehensive review of the literature on dynamic pricing and revenue management, and to Ref. 3 for a detailed exposition on point processes and their optimal intensity control. We start reviewing the basic mathematical model under complete demand information.

DYNAMIC PRICING WITH COMPLETE DEMAND INFORMATION

Consider a stream of potential customers arriving according to a time-homogeneous and price-independent Poisson process with rate Λ . We refer to Λ as the *market size*. Upon arrival, a consumer buys the product with probability $\bar{F}(p)$, where p is the price of the product listed at this time. We can interpret $\bar{F}(p)$ as follows. We associate with each consumer a *reservation price* (that is, a maximum willingness to pay), which is distributed according to $F(\cdot)$ among the population of consumers (see Ref. 4 for more details), so that $\bar{F}(p) = 1 - F(p)$. We will assume that F admits a density f . The effective demand process or the sales process that results from

the above description is a nonhomogeneous Poisson process with rate $\lambda : \mathbb{R}_+ \rightarrow [0, \Lambda]$, a continuous, bounded, and decreasing function such that at each time t , $\lambda_t = \lambda(p_t) = \Lambda \bar{F}(p_t)$, where p_t is the price charged at time t . We will refer to $\lambda(\cdot)$ as the *demand function*. Let $\mathcal{A}$ be the set of *admissible* price processes $p = (p_t : t \geq 0)$, that is, processes for which the value of p_t depends exclusively on the history of sales and prices up to time t (but not on future unobserved events). We will denote by $\mathcal{F}_t$ this history up to time t .

Let $\mathcal{N} = (\mathcal{N}_t : t \geq 0)$ be a standard (rate 1) Poisson process. For any $p \in \mathcal{A}$ and demand function λ , we define the cumulative demand process $N_\lambda^p(t) := \mathcal{N}\left(\int_0^t \lambda(p_s) ds\right)$ and its corresponding cumulative revenue

$$\mathcal{R}(p; x_0; \lambda) := \int_0^T p_t \mathbb{1}(N_\lambda^p(t) < x_0) dN_\lambda^p(t),$$

where T is the length of the sales horizon,

x_0 is the seller's initial inventory and $\mathbb{1}(E)$ is the indicator function of the event E . The seller's dynamic pricing problem is given by the following (intensity control) problem

$$\sup_{p \in \mathcal{A}} \mathbb{E}[\mathcal{R}(p; x_0; \lambda)]. \quad (\text{P})$$

For $p \in \mathcal{A}$, Problem (P) can be rewritten as follows (see Chapter 2 in Ref. 3 for details).

$$\begin{aligned} \sup_{p \in \mathcal{A}} \mathbb{E} \left[\int_0^T r(p_t) dt \right] \\ \text{subject to} \quad N_\lambda^p(T) \leq x_0 \quad (\text{a.s.}), \end{aligned}$$

where $r(p) := p \lambda(p) = p \Lambda \bar{F}(p)$ is the instantaneous revenue rate. Let us define $p^* := \arg\max\{r(p)\}$, $\lambda^* := \lambda(p^*)$ and $r^* := r(p^*)$. It is worth noticing that despite the fact that p^* maximizes the instantaneous revenue rate $r(p)$ choosing $p_t = p^*$ for all t is in general suboptimal. The reason is that this choice could deplete the available

inventory too fast (before T) and the seller would be better off charging a higher price without necessarily sacrificing sales. On the other hand, if $x_0 = \infty$ then the seller can never deplete his/her inventory in finite time and $p_t = p^*$ for all $t \leq T$ is indeed an optimal policy.

A standard way to solve Problem (P) is by using dynamic programming. For this, let $J^*(x, t)$ be the seller's optimal revenue-to-go if the remaining sales horizon is $t > 0$ and the current available stock is x units. It follows that $J^*(x, t)$ satisfies the Hamilton–Jacobi–Bellman (HJB) equation (see Chapter 7 in Ref. 3 for details)

$$\begin{aligned} \frac{\partial}{\partial t} J^*(x, t) \\ = \max_{p \geq 0} \left[r(p) - \lambda(p) [J^*(x, t) - J^*(x-1, t)] \right], \end{aligned}$$

with boundary conditions $J^*(x, 0) = J^*(0, t)$ for all $x, t \geq 0$. One can show that there exists a nondecreasing function ζ such that the optimal price satisfies $p^*(x, t) = \zeta(J^*(x, t) - J^*(x-1, t))$. Similarly, we define a nonnegative and decreasing function Ψ such that the right-hand side of the HJB equation is equal to $\Psi(J^*(x, t) - J^*(x-1, t))$. From the first-order optimality condition for $p^*(x, t)$ we get that the HJB equation leads to the following set of identities.

$$\begin{aligned} \frac{\partial}{\partial t} J^*(x, t) &= \Psi(J^*(x, t) - J^*(x-1, t)) \\ &= \Lambda \frac{\bar{F}(p^*(x, t))}{h(p^*(x, t))}, \end{aligned} \quad (1)$$

where $h(p) := f(p)/\bar{F}(p)$ is the hazard function of $F(p)$. It is worth noticing that the first equality defines an algorithm to solve the seller's optimization problem.

Algorithm.

Step 0: Initialization. Set $J^*(0, t) = 0$ for all $t \in [0, T]$ and $n = 1$.

Step 1: Iteration. Given $J^*(n-1, t)$ for all $t \in [0, T]$ compute $J^*(n, t)$ as the solution of the first-order ordinary differential equation.

$$\frac{\partial}{\partial t} J^*(n, t) = \Psi(J^*(n, t) - J^*(n-1, t)) \quad \text{with border condition} \quad J^*(n, 0) = 0.$$

Step 2: If $n = x$ then stop. Otherwise, set $n = n + 1$ and go to Step 1.

Closed-form solutions for $J^*(x, t)$ are not generally available. A notable exception is the case in which the reservation price is exponentially distributed, $\bar{F}(p) = \exp(-p/\theta)$. In this case, it can be shown that $J^*(x, t) = \theta \ln(\sum_{n=0}^x (\lambda^* t)^n / n!)$. For the general case, Gallego and van Ryzin [5] obtain structural results for problem (P) using the HJB optimality condition.

Theorem 1 [5]. *Suppose that $r(p)$ is a concave function. The revenue-to-go, $J^*(x, T)$, is strictly increasing and strictly concave in both T and x . Furthermore, there exists an optimal price $p^*(x, T) \geq p^*$ that is strictly decreasing in x and strictly increasing in T .*

It follows that inventory and time (the seller's primary resources) have diminishing marginal returns. The following result establishes the equivalence between the market size Λ and the sales horizon T and will be useful in our discussion on how to handle uncertainty on the market size.

Theorem 2. *Let $J_\Lambda^*(x, t)$ be the seller's revenue-to-go given a market size Λ . Then, $J_\Lambda^*(x, T) = J_1^*(x, \Lambda T)$ for all $x \geq 0$ and $T \geq 0$. It follows that, under the assumptions of Theorem 1, $J_\Lambda^*(x, T)$ is an increasing and concave function of Λ .*

Proof. For any admissible pricing policy $p = (p_t : 0 \leq t \leq T)$, we define $\tilde{p} = (\tilde{p}_t : 0 \leq t \leq \Lambda T)$ such that $\tilde{p}_t = p_{t/\Lambda}$. The result follows by noticing that $\int_0^\tau \Lambda (1 - F(p_t)) dt = \int_0^{\Lambda\tau} (1 - F(\tilde{p}_t)) dt$ for all $\tau \in [0, T]$. Hence, the pricing policy p_t in $[0, \tau]$ generates (pathwise) the same revenue and depletes the same amount of inventory than the pricing policy $\tilde{p}_t$ in $[0, \Lambda\tau]$.

Another key contribution from Ref. 5 is the observation that a fixed-price policy can be asymptotically optimal as the scale of business (inventory and demand) increases. Denote by $p^D(x_0, T) := \max\{p^*, p^0\}$, where $\lambda(p^0) := x_0/T$. One can show that $p_t = p^D$ for all $t \leq T$, is an optimal policy for the deterministic version of Problem (P) (i.e., one that replaces the stochastic increments dN_t^p by its deterministic counterpart $\lambda(p_t) dt$). We let $J^{\text{FP}}(x, t)$ be the seller's expected payoff if he/she selects the price $p^D(x, t)$ for the remaining sale horizon t .

Theorem 3 [5]. *Suppose that $r(p)$ is a concave function. Then,*

$$1 - \frac{1}{2\sqrt{\min\{x, \lambda^* T\}}} \leq \frac{J^{\text{FP}}(x, T)}{J^*(x, T)} \leq 1.$$

It follows that when x and $\lambda^* T$ are both large a fixed-price policy is sufficient to achieve an almost optimal revenue. This is a very useful result that limits the seller's needs for a possibly complex dynamic pricing policy.

An important limitation of the previous result is the assumption that the seller knows the specific demand function $p \mapsto \lambda(p)$. In practice, it is rarely the case that the seller can correctly determine this function. In some cases, some incomplete information is available based on historical data. In others cases, this demand function is completely unknown (i.e., for new products or new markets). Furthermore, with incomplete demand information it is no longer true that structural properties of an optimal solution to Problem (P) will still hold. For example, with incomplete demand information there is no guarantee that a simple fixed-price policy will produce asymptotically optimal results (i.e., Theorem 3 is not guaranteed

to hold). Indeed, as the following stylized example reveals an optimal fixed-price policy can perform very poorly if the seller does not know the true demand function.

Example. Consider Problem (P) and suppose that at the beginning of the selling season the seller is uncertain about the true demand function $\lambda(p)$. He/She only knows that there is a 50% chance that $\lambda(p) = \lambda_i(p)$, where $\lambda_i(p) = \Lambda \mathbb{1}(p \leq p_i)$, $i = 1, 2$ for two fixed prices p_1 and p_2 such that $0 < p_1 < p_2 = 2p_1$. With full information, the seller knows the value of i and chooses an optimal pricing policy $p_t = p_i$ for all t with expected payoff $J_i^*(x, T) = p_i (\Lambda T - \mathbb{E}[(N(\Lambda T) - x_0)^+])$. Hence, it follows that with incomplete information the seller's payoff $V(x, t)$ is bounded by

$$\begin{aligned} V(x, t) &\leq \frac{1}{2} J_1^*(x, T) + \frac{1}{2} J_2^*(x, T) \\ &= \frac{3}{4} p_2 (\Lambda T - \mathbb{E}[(N(\Lambda T) - x_0)^+]). \end{aligned}$$

One can show that an optimal fixed-price policy (one that maximizes the seller's ex ante expected revenue) is to set $p_t = p_2$ for all $t \leq T$. The corresponding expected payoff is given by

$$V^{\text{OFP}}(x_0, T) = \frac{p_2}{2} (\Lambda T - \mathbb{E}[(N(\Lambda T) - x_0)^+]).$$

Under this optimal fixed-price policy there is 50% chance that the seller would sell no unit during the entire selling season (i.e., when $\lambda(p) = \lambda_1(p)$). Hence, the uncertainty on the true demand function and the lack of flexibility of a fixed-price policy to adjust prices over time make this fixed-price policy a risky and suboptimal strategy. To assess the degree of suboptimality, consider the following simple adaptive pricing policy designed to learn the true demand function from early sales. For a fixed $\alpha \in (0, 1)$, let $p_t = p_2$ for all $t \leq \alpha T$. If at least one sale occurs in $[0, \alpha T]$ then set $p_t = p_2$ for all $t \in [\alpha T, T]$, otherwise set $p_t = p_1$ for all $t \in [\alpha T, T]$. The expected

payoff of this pricing policy is given by

$$\begin{aligned} V(x_0, T, \alpha) &= \frac{(1 + e^{-\Lambda \alpha T})}{2} V^{\text{OFP}}(x_0, (1 - \alpha)T) \\ &\quad + \frac{(1 - e^{-\Lambda \alpha T})}{2} \\ &\quad \left(p_2 + 2 \mathbb{E}_\tau [V^{\text{OFP}}(x_0 - 1, T - \tau)] \right), \end{aligned}$$

where $\tau \in [0, \alpha T]$ is a random time with distribution $F_\tau(s) = (1 - e^{-\Lambda s})/(1 - e^{-\Lambda \alpha T})$.

It follows that

$$\begin{aligned} \lim_{x_0 \rightarrow \infty} \frac{V(x_0, T, \alpha)}{V^{\text{OFP}}(x_0, T)} &= \\ 1 + \frac{(1 - \alpha)}{2} (1 - e^{-\Lambda \alpha T}) &\xrightarrow{T \uparrow \infty} 1 + \frac{(1 - \alpha)}{2}. \end{aligned}$$

According to Theorem 3, under full information, a fixed-price policy is asymptotically optimal as x_0 and T grow large. However, in this case with incomplete information, the optimal fixed-price policy can deviate from an optimal strategy by as much as 50% (letting $\alpha \downarrow 0$) as x_0 and T grow large. At the same time, from the bound on $V(x_0, T)$ and the value of $V(x_0, T, \alpha)$ above, it follows that $V(x_0, T, \alpha)/V(x_0, T)$ converges to 1 as x_0 and T grow large and $\alpha \downarrow 0$. This result suggests that under incomplete demand information there could be a version of Theorem 3 stating that an almost everywhere fixed priced policy is asymptotically optimal, that is, a policy in which the seller needs to spend a small amount of time learning and then implementing a fixed-price policy. We will see in the section titled "Minimax and the Regret Function" that such a theorem does indeed exist under rather general conditions.

The previous example reveals the risks that a seller could take if he/she does not fully incorporate the effects of incomplete demand information when selecting optimal pricing strategies. The simple adaptive pricing policy in this example also highlights the new informational role that pricing policies must play when there is uncertainty about the true demand function. With incomplete demand information the seller faces an *exploration-exploitation* trade-off, where the

pricing decision the seller implements incorporates a learning component as part of the revenue maximization problem. In the following section, we summarize a set of theoretical frameworks that have been proposed to incorporate this incomplete demand information in the context of revenue management and dynamic pricing.

DYNAMIC PRICING WITH INCOMPLETE DEMAND INFORMATION

In most (if not all) practical situations, the seller has only partial information about the true value of the demand function $\lambda(p)$. As a result, the dynamic pricing problem (P) needs to be modified to incorporate this additional *ambiguity* in the model formulation. Alternative solutions have been proposed to tackle this problem, which differ in two main characteristics: (i) the representation of the unknown function λ and (ii) the criteria that is used to resolve this model ambiguity.

Regarding the issue of how to model λ , two distinctive approaches have been studied: *parametric* and *nonparametric* models. As suggested by their names, the choice between these two alternatives largely depends on the capacity that the seller has to represent the demand function in terms of a finite number of parameters. To be concrete, suppose that the seller knows that $\lambda \in \mathcal{H}$, where $\mathcal{H}$ is a given family of real-valued functions. If $\mathcal{H}$ is the family of polynomials of degree K (for some fixed K) then $\lambda(p) = \beta_0 + \beta_1 p + \dots + \beta_K p^K$ and the seller's learning problem reduces to estimate the "best" set of parameters $\{\beta_k\}$. On the other hand, if $\mathcal{H} = \mathcal{C}_+[0, \infty)$, the set of nonnegative continuous functions in $[0, \infty)$, then selecting the "best" λ involves searching in this infinite-dimensional space; a problem that does not admit a parametric representation.

Independent of whether the seller uses a parametric or a nonparametric model to represent λ , the question of how to select an optimal pricing strategy needs to be addressed. This problem entails choosing a particular criteria to model how the seller's ambiguity about the true value of λ should impact his/her choice of prices.

The Bayesian approach is probably the most popular alternative used to model sequential demand learning for the case in which λ has a parametric description. Under this approach, model ambiguity on the demand function $\lambda(p)$ is captured by a set of unknown random parameters that characterize this function and for which a prior (at time 0) probability distribution is postulated. As time goes by and the evolution of the demand process is observed, this prior distribution is updated based on the new incoming information using Bayes rule. To be more specific, let θ be a random vector of parameters taking values in a set Θ and let $F(x)$ be the seller's prior probability distribution of θ . For each $\theta \in \Theta$ there is a corresponding demand function $\lambda_\theta \in \mathcal{H}$ and the seller's optimal dynamic pricing problem (P) takes, now, the following form

$$\sup_{p \in \mathcal{A}} \int_{\theta \in \Theta} \mathbb{E} [\mathcal{R}(p; x_0; \lambda_\theta)] dF(\theta).$$

From a modeling perspective, the Bayesian method relies heavily on the good judgment and expertise of the decision maker to select (i) the "right" functional form for the demand function and (ii) the "right" prior distribution for the unknown parameters. There are positive and negative sides to this approach. On the negative side, the parametric nature of the Bayesian model confines the demand learning within the boundaries specified by the proposed functional form and prior distribution and so any misspecification on these quantities will persist throughout the entire learning process. On the positive side, the Bayesian approach benefits from any prior knowledge that the decision maker may have about the demand process. This knowledge is particularly beneficial in those cases where the planning horizon is relatively short and there is limited time to learn through experimentation. Finally, Bayesian models can be mathematically tractable (with computationally efficient solution methods) for some specific family of distributions of the prior (those families of distributions for which their *conjugate* is tractable). In the section titled "Dynamic Pricing with Bayesian Learning", we review some representative papers that use this Bayesian method.

As mentioned above, one valid criticism to the Bayesian approach is its strong dependence on the existence of a prior probability distribution for θ . Rather than modeling these unknown parameters as random variables, an alternative approach is to use a point estimate $\hat{\theta} \in \Theta$, where the value of $\hat{\theta}$ is determined by optimizing a particular criteria that depends on θ and all available information. For example, at time $t = 0$, the *maximum likelihood estimation* of $\hat{\theta}$ solves

$$\hat{\theta} := \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta, \mathcal{F}_0),$$

where $\mathcal{L}(\theta, \mathcal{F}_0)$ measures the likelihood of observing a history $\mathcal{F}_0$ for a given value of $\theta \in \Theta$. In other words, the seller's model ambiguity is resolved by mean of an optimization problem in terms of a likelihood objective function. Alternative estimation criteria have been proposed, essentially by replacing $\mathcal{L}$ by other objective functions such as *least square estimation* (or more generally $\|\cdot\|_q$ -norm estimation), or *minimum entropy estimation*. As opposed to the Bayesian approach in which the exploration and exploitation stages are conducted simultaneously, in these models both steps are often decoupled. That is, given an estimate $\hat{\theta}$, the seller determines an optimal pricing strategy by maximizing expected revenue, $\mathbb{E}[\mathcal{R}(p; x_0; \lambda_{\hat{\theta}})]$. On the other hand, demand learning is achieved by periodically recomputing the value of $\hat{\theta}$ as new information is gathered, that is, replacing $\mathcal{F}_0$ by $\mathcal{F}_t$ as time goes by. In the section titled "Dynamic Pricing for Linear Demand with Unknown Coefficients", we described a particular example that uses a least square estimation approach.

Another popular approach that has also been used extensively to handle model ambiguity is the class of *robust* formulations. The distinguishing feature of this approach is that it makes no probabilistic assumptions about which function $\lambda \in \mathcal{H}$ is more or less likely to be the true demand function. This *uncertainty set* $\mathcal{H}$ (as it is usually called in this context) captures all the knowledge that sellers have about this demand. A robust pricing strategy is one that guarantees the best possible level of performance (e.g., in a maximin,

competitive ratio, or minimax regret criteria, among others) uniformly over all possible values of $\lambda \in \mathcal{H}$. For example, in the maximin version of problem (P), the seller chooses a dynamic pricing policy that solves for

$$\sup_{p \in \mathcal{A}} \inf_{\lambda \in \mathcal{H}} \mathbb{E}[\mathcal{R}(p; x_0; \lambda)];$$

that is, a pricing policy that maximizes the *worst-case* performance with respect to all demand functions in the uncertainty set $\mathcal{H}$. The popularity of this approach lies in the fact that it captures in a parsimonious way the seller's model ambiguity. In addition, and depending on the characteristics of $\mathcal{H}$ (e.g., polyhedral or ellipsoidal uncertainty sets), the maximin formulation can be solved efficiently. Among the disadvantages, the maximin criteria generally produce solutions that are too conservative, specially if $\mathcal{H}$ is "big." For instance, suppose that $\mathcal{H}$ contains a function λ_{ϵ} such that $\lambda_{\epsilon}(p) = 0$ for all $p \geq \epsilon$ for some $\epsilon > 0$ small. Then, an optimal maximin pricing policy satisfies $p_t^* < \epsilon$ for all t . It follows from this example that one needs to take special care in selecting the set $\mathcal{H}$ under a maximin objective in order to avoid trivial or nonrealistic solutions. To address this issue of conservatism, other robust criteria have been proposed. For example, under an absolute minimax regret approach the seller selects a pricing policy that minimizes the difference in performance with respect to the full information case

$$\inf_{p \in \mathcal{A}} \sup_{\lambda \in \mathcal{H}} \left\{ \sup_{\tilde{p} \in \mathcal{A}} \mathbb{E}[\mathcal{R}(\tilde{p}; x_0; \lambda)] - \mathbb{E}[\mathcal{R}(p; x_0; \lambda)] \right\}.$$

Note that in this case, it is not generally true that an optimal pricing policy is conditioned by the presence of λ_{ϵ} in $\mathcal{H}$.

Another disadvantage of the robust approach—one that is particularly relevant in our present discussion—is that it does not explicitly incorporate the possibility of demand learning. In the section titled "Minimax and the Regret Function", however, we review some recent results that show that in certain cases a simple policy that learns for a relatively small period of time and then myopically sets prices is asymptotically optimal in a robust sense.

In the following sections, we discuss some concrete models and examples within the class of parametric and nonparametric models for different demand learning criteria.

PARAMETRIC MODELS

Parametric models are based on some fundamental knowledge that the seller has about the true demand function. For example, the seller might know that $\lambda(p)$ belongs to a specific family of functions, which is characterized by a finite set of parameters (e.g., the family of linear functions as in Ref. 6). The seller parametric sequential learning problem is to identify the “best” value of those parameters using realized market information (prices and sales). As we mentioned above, there are different ways that one can use to identify the best set of parameters that we review in what follows.

Dynamic Pricing with Bayesian Learning

In the context of revenue management and dynamic pricing, there is a growing literature that uses Bayesian models to characterize optimal pricing policies when there is ambiguity about the true demand function. One particular type of uncertainty that has received significant attention is *market size uncertainty*. This model is well suited for those instances of the problem in which the seller knows relatively well consumers’ reservation price distribution $\bar{F}(p)$ but has limited information about the actual size of the population of potential buyers, Λ . Some representative papers that have studied this model are found in Refs 7–9 (see also Refs 10 and 11). In the context of Bayesian learning, Aviv and Pazgal [7] consider the special case in which Λ has a gamma prior distribution and $\bar{F}(p) = \exp(-\alpha p)$ (i.e., exponentially distributed reservation price). The advantage of using a gamma distribution for Λ is that it is a conjugate distribution for the Poisson demand process, which simplifies enormously the use of Bayes rule. In an infinite horizon setting, Farias and Van Roy [8] generalize the demand model in Ref. 7 to the family of finite mixture of gamma distributions and propose a special heuristic (*decay balancing*)

that shows a good numerical performance compared to other heuristics proposed in the literature. In the context of a retail operations, Araman and Caldentey [9] consider an infinite horizon model for an arbitrary $\bar{F}(p)$ and where Λ has a finite support. A distinguishing feature in this model is that the seller has the option to optimally stop selling the product at any given time to switch to a different assortment. This is a particularly valuable option in the context of demand learning since the seller is initially uncertain about how profitable the product really is. For example, if the true value of Λ is small then the seller is better off removing the product.

In what follows, we review the main results of the research discussed in the previous paragraph. The underlying assumption that we make in the reminder of this section is that Λ is the seller’s sole source of uncertainty. Having this in mind, we rewrite Problem (P). We denote by $G_0(\Lambda)$ the seller’s (prior) cumulative probability distribution of Λ at time 0 and by $\mathbb{P}_0(\mathbb{E}_0)$ the conditional probability measure (expectation operator) given G_0 . We define similarly G_t , $\mathbb{P}_t$, and $\mathbb{E}_t$ conditional on $\mathcal{F}_t$, the available information at time t . Using a slight abuse of notation, let $r(p) = p \bar{F}(p)$. The seller’s problem is given by

$$\sup_{p \in \mathcal{A}} \mathbb{E}_0 \left[\int_0^T \Lambda r(p_t) dt \right]$$

subject to and

$$\int_0^T dN_t^p \leq x_0 \quad (a.s.). \quad (\text{B})$$

Similar to Problem (P), we can solve Problem (B) using dynamic programming but with an enlarged state space (x, t, G_t) . The parametric description of G_t comes in handy at this point to ensure the tractability of the resulting dynamic programming (DP) algorithm.

The stochastic evolution of the new state variable G_t is derived from Bayes rule. For the sake of concreteness, let us suppose that Λ has a finite support $\{\Lambda_k\}_{k=1}^K$ (where $\{\Lambda_k\}$ is an increasing sequence) so that G_t is piecewise constant. Let $g_t(k) = G_t(\Lambda_k) - G_t(\Lambda_{k-1})$ with $\Lambda_0 = 0$. Then, the Bayes rule implies

that (see Proposition 3 and Section 6.1 in Ref. 9 for details)

$$g_t(k) = \mathbb{P}_0(\Lambda = \Lambda_k | \mathcal{F}_t) = \frac{g_0(k) \cdot (\Lambda_k I_t^p)^{N_t} \exp(-\Lambda_k I_t^p)}{\sum_j g_0(j) \cdot (\Lambda_j I_t^p)^{N_t} \exp(-\Lambda_j I_t^p)}.$$

It follows from its definition that $\{g_t, \mathcal{F}_t\}$ is a martingale. Application of Itô's lemma in the previous equation leads to the following stochastic differential equation (SDE)

$$dg_t(k) = \eta_k(g_t) (\lambda_t \bar{\Lambda}(g_t) dt - dN_t), \quad 1 \leq k \leq K,$$

$$\text{where } \bar{\Lambda}(g) := \sum_k g(k) \Lambda_k \text{ and}$$

$$\eta_k(g) := g(k) \left(\frac{\bar{\Lambda}(g) - \Lambda_k}{\bar{\Lambda}(g)} \right).$$

The resulting HJB optimality condition for Problem (B) is given by

$$\begin{aligned} \frac{\partial}{\partial t} V(x, t, g) = \max_p \left[\bar{\Lambda}(g) \bar{F}(p) [V(x-1, t, g - \eta(g)) - V(x, t, g) + \eta(g) \right. \\ \left. \cdot \nabla_g V(x, t, g)] + \bar{\Lambda}(g) r(p) \right], \quad (2) \end{aligned}$$

where $\nabla_g V(x, t, g)$ is the gradient of $V(x, t, g)$ with respect to g . The boundary conditions are $V(0, t, g) = V(x, 0, g) = 0$ and $V(x, t, e_k) = J_1^*(x, \Lambda_k t)$ where e_k is the distribution that gives probability one to the event $\{\Lambda = \Lambda_k\}$ and $J_1^*(x, t)$ is the full information value function for problem (P) when the market size is normalized to one. Note that the latter border condition follows from Theorem 2. Solving Equation (2) is usually a very difficult task because of the singularities at the boundary points $g = e_k$ and the jumps in x and g that make the HJB a delayed difference-differential equation. Despite the fact that an analytical solution is not immediately available, this optimality condition provides enough information to derive some useful properties that we can use to approximate the value function and the corresponding pricing strategy.

Theorem 4. *The value function $V(x, t, g)$ satisfies the following properties:*

- (i) *It is increasing in x and t . It is also convex in g .*
- (ii) *It is bounded by: $J^*(x, \Lambda_1 t) \leq V(x, t, g) \leq \sum_k g(k) J^*(x, \Lambda_k t) \leq J^*(x, \bar{\Lambda}(g) t)$.*
- (iii) *It satisfies $\lim_{x \rightarrow \infty} V(x, t, g) = \bar{\Lambda}(g) r(p^*)$ $t = \lim_{x \rightarrow \infty} \sum_k g(k) J^*(x, \Lambda_k t)$. The optimal price satisfies $\lim_{x \rightarrow \infty} p^*(x, t) = p^*$.*
- (iv) *A fixed-price policy is asymptotically optimal in the following sense. Let $V^{\text{FP}}(x, t, g)$ be the seller's expected revenue using an optimal fixed-price for the entire sale horizon. Then,*

$$\begin{aligned} 1 - \frac{1}{2\sqrt{\min\{x, \bar{\Lambda}(g) \bar{F}(p^*) t\}}} \\ \leq \frac{V^{\text{FP}}(x, t, g)}{V(x, t, g)} \leq 1. \end{aligned}$$

Proof. Part (i). The monotonicity of $V(x, t, g)$ on t and x follows trivially from its definition. The convexity on g follows from

$$\begin{aligned} V(x, t, g) = \sup_p \mathbb{E} \left[\sum_k g(k) \int_0^T p_t dN_t^p(k) \right], \\ \text{where } N_t^p(k) := \mathcal{N} \left(\Lambda_k \int_0^t \bar{F}(p_s) ds \right). \end{aligned}$$

Part (ii). The first inequality (lower bound) follows from implementing on the incomplete information case, the optimal pricing policy for the case of full information with $\Lambda = \Lambda_1$. The second inequality follows from interchanging the sup and the summation in the equation above. The third inequality follows from Theorem 2. Part (iii) follows from noticing that for $x = \infty$, the optimal pricing strategy is $p_t = p^*$, which maximizes the instantaneous revenue rate (see the discussion in the section titled "Dynamic Pricing with Complete Demand Information" after the definition of Problem (P)) independently of the value of Λ . Finally, Part (iv) follows from combining the upper bound on Part (ii) and the result in Theorem 3.

Note that $\mathbb{E}[J^*(x, \Lambda, t)] = \sum_k g(k) J^*(x, \Lambda_k, t)$ and so point (ii) asserts that an upper bound for the value function is given by the expected value (over the unknown market size Λ) of the full information value function. Point (iv) extends Gallego and van Ryzin's [5] result in Theorem 3 to the case with uncertain market size. This result suggests that the lack of asymptotic optimality of a fixed-price policy that we identified in Example 1 is mainly due to the seller's uncertainty about the buyer's reservation price distribution rather than the actual market size.

Approximations and Heuristics. One approach that has been used to approximate the optimal pricing policy is to first get an approximation of the value function and then plug it in the HJB equation (2) to derive an approximating pricing policy. For any approximation $V^{\text{aprx}}(x, t, g)$ of the value function, the corresponding pricing policy is given by

$$p^{\text{aprx}}(x, t, g) = \zeta \left(V^{\text{aprx}}(x-1, t, g - \eta(g)) - V^{\text{aprx}}(x, t, g) + \eta(g) \cdot \nabla_g V^{\text{aprx}}(x, t, g) \right), \quad (3)$$

the function ζ was defined in the section titled "Dynamic Pricing with Complete Demand Information". (It should be clear that the pair $(V^{\text{aprx}}, p^{\text{aprx}})$ does not solve Equation (2).) This approach was used in Ref. 9 to derive an *asymptotic approximation*. This policy is based on the observation that points (ii) and (iii) in Theorem 4 suggest the use of the following approximation for the value function $\tilde{V}(x, t, g) = \sum_{k=1}^K g(k) J^*(x, \Lambda_k, t)$, which is asymptotically optimal as x grows large. Let us denote by $\tilde{p}(x, t, g)$ the pricing policy that results from using $\tilde{V}(x, t, g)$ in Equation (3).

An alternative approach to approximate an optimal pricing policy is to solve a modified version of the HJB equation. A popular example is the so-called *naïve policy*. This approximation assumes that at every state (x, t, g) the seller solves the full information HJB in Equation (1) replacing the unknown Λ by its expected value $\bar{\Lambda}(g)$ (see Ref. 7 for details). Another example is the *decay balancing heuristic* proposed by Farias and Van

Roy [8]. Although this policy was derived in an infinite horizon setting, the main idea can be directly extrapolated to the finite horizon case. The decay balance policy combines the asymptotic policy $\tilde{V}(x, t, g)$ and the HJB equation (2) (see also Equation (1)) to propose a pricing policy $p^{\text{BD}}(x, t, g)$ solution to

$$\frac{\partial}{\partial t} V(x, t, g) = \bar{\Lambda}(g) \frac{\bar{F}(p^{\text{BD}}(x, t, g))}{h(p^{\text{BD}}(x, t, g))}.$$

Under the additional assumption that the reservation price distribution, $F(p)$, has increasing failure rate, the solution $p^{\text{BD}}(x, t, g)$ is unique. A set of numerical experiments comparing the *asymptotic approximation*, the *naïve policy*, and the *decay balancing heuristic* is reported in Ref. 8.

Dynamic Pricing for Linear Demand with Unknown Coefficients

The Bayesian method, as discussed in the previous section, relies heavily on the seller's prior distribution of the unknown parameters. In those cases, when no prior exists, the seller must rely on an alternative statistical estimation method, such as the least square estimator (LSE), the maximum likelihood estimator (MLE), the minimum entropy estimator (MEE), and others. Some representative examples are Ref. 6 (this model is discussed in more detail below), Ref. 12, which considers also a linear price demand function and obtains approximate solutions using convex programming methods, and Ref. 13. Carvalho and Puterman [13] consider a binomial demand, which is relevant in the internet environment representing those that visited a web site and ended up buying the item. Under such a demand model, some parameters are unknown and a Monte Carlo simulation is used to quantify the trade-off between learning and revenue maximization. This simulation allows to measure the performance of few suggested heuristics among others, the one-step-look-ahead policy. Because of the numerical flavor, the method could apply to various estimation techniques in particular maximum likelihood and Bayes.

Bertsimas and Perakis [6] consider a discrete variant of problem (P) in which the demand in every period n is given by

$$d_n = \beta^0 + \beta^1 p_n + \epsilon_n,$$

where the ϵ_n 's are iid $(0, \sigma^2)$ normal random variables. The parameters β_0, β_1 , and σ are unknown. The estimates at period n of these parameters are computed using the LSE method

$$\begin{aligned} (\hat{\beta}_n^0, \hat{\beta}_n^1) &= \arg \min_{\mathbf{r} \in \mathbb{R}^2} \sum_{s=1}^{n-1} (d_s - \mathbf{x}'_s \mathbf{r})^2 \quad \text{and} \\ \hat{\sigma}_n^2 &= \sum_{\tau=1}^{n-1} \frac{(d_\tau - \hat{\beta}_n^0 - \hat{\beta}_n^1 p_\tau)^2}{n-3} \\ n &= 4, \dots, T, \end{aligned}$$

where, $\mathbf{x}'_s = [1, p_s]$. At the beginning of period n , the decision variable p_n is selected to maximize the total expected revenue-to-go. To compute this revenue-to-go, the seller needs to estimate the distribution of the demand in period $s \geq n$ (denoted by $\hat{d}_{s,n}$) conditional on the current estimates $(\hat{\beta}_n^0, \hat{\beta}_n^1, \hat{\sigma}_n)$. It follows that $\hat{d}_{s,n} = \hat{\beta}_{s,n}^0 + \hat{\beta}_{s,n}^1 p_s + \hat{\epsilon}_{s,n}$, where, $\hat{\epsilon}_{s,n}$ is a normally distributed random variable with mean 0 and standard deviation $\hat{\sigma}_{s,n}$; we use the notation $\hat{y}_{s,n}$ to denote the current period n estimate of a parameter y for a future period s , $s = n, \dots, T$. The most critical step in this approach (similar to the Bayesian case) is to obtain an iterative process that, given a state space, computes future estimates. It can be shown that the evolution of these estimates is given by

$$\hat{\beta}_{s+1,n}^i = \hat{\beta}_{s,n}^i + \hat{\epsilon}_{s,n} \cdot h_i \left(p_s, \sum_{\tau=1}^{s-1} p_\tau, \sum_{\tau=1}^{s-1} p_\tau^2 \right),$$

for $i \in \{0, 1\}$, and

$$\begin{aligned} \hat{\sigma}_{s+1,n}^2 &= \mathcal{H} \left(p_s, \hat{\beta}_{s,n}^0, \hat{\beta}_{s,n}^1, \sum_{\tau=1}^{s-1} p_\tau^2, \sum_{\tau=1}^{s-1} p_\tau, \right. \\ &\quad \left. \sum_{\tau=1}^{s-1} p_\tau d_\tau, \sum_{\tau=1}^{s-1} d_\tau, \hat{\sigma}_{s,n}^2 \right), \end{aligned}$$

for some real functions h_i and $\mathcal{H}$ (see Ref. 6 for details). These estimates are used in a DP

formulation where the state space is given by the vector

$$\mathbf{X}_s = \left(x_s, \hat{\beta}_{s,n}^0, \hat{\beta}_{s,n}^1, \sum_{\tau=1}^{s-1} p_\tau^2, \sum_{\tau=1}^{s-1} p_\tau, \sum_{\tau=1}^{s-1} p_\tau d_\tau, \sum_{\tau=1}^{s-1} d_\tau, \hat{\sigma}_{s,n}^2 \right),$$

where x_n is the inventory available in period n . At each period n , one solves the following DP with

$$\begin{aligned} J_T(c_T, \hat{\beta}_{T,n}^0, \hat{\beta}_{T,n}^1, \hat{\sigma}_{T,n}^2) &= \max_{p_T} \mathbb{E}_{\epsilon_{T,n}} p_T \\ &\quad \min \left\{ (\beta_{T,n}^0 + \beta_{T,n}^1 p_T + \epsilon_{T,n}), c_T \right\}, \end{aligned}$$

and for $s = n, \dots, T-1$,

$$\begin{aligned} J_s(\mathbf{X}_s) &= \max_{p_s} \mathbb{E}_{\epsilon_n} \left[p_s \min \left\{ \hat{\beta}_{s,n}^0 + \hat{\beta}_{s,n}^1 p_s \right. \right. \\ &\quad \left. \left. + \hat{\epsilon}_{s,n}, c_s \right\} + J_{s+1}(\mathbf{X}_{s+1}) \right], \quad (4) \end{aligned}$$

where the components of $\mathbf{X}_{n+1}$ get updated based on the selected value of p_s , the corresponding demand estimate and the iterative processes described above (e.g., $c_{s+1} = c_s - \min\{\hat{\beta}_{s,n}^0 + \hat{\beta}_{s,n}^1 p_s + \hat{\epsilon}_{s,n}, c_s\}$).

A few observations are in order. First, we note that for a fixed n , the sequence $(\hat{\beta}_{s,n}^i : s = n, \dots, T)$ made of the estimates at time n for future values of the parameter is a stochastic process and in particular a martingale. This is an expected fact as it measures the learning process. A valid question at this point is how to possibly quantify the learning. It will be hard to obtain a tractable formulation of the gap between the optimal values and those obtained through the estimation process, but a Monte Carlo simulation can help in this regard. Similar to Ref. 13, one can solve for the optimal pricing policy under known parameters β_0, β_1 and σ and then implement the previous pricing policy obtained through the learning process.

The main characteristic of this problem is the concurrent parameter estimation and the pricing optimization (active learning). If we decouple these two actions (move to passive learning), whereby the pricing

decision made at time n is assumed not to affect the parameter estimates at time $n + 1$, then the state space of the DP can be reduced to one dimension, that of the available capacity.

The setting of Ref. 6 is similar to the initial problem (P). The difference, beyond the fact that time is discrete, is the choice itself of a linear demand rate with a normally distributed error term independent of the price. From a practical standpoint, this model fits well with the popular approach of forecasting demand using linear regressions. Another advantage of this model compared to those discussed in the section titled “Dynamic Pricing with Bayesian Learning” is that the learning process includes the price elasticity (i.e., an unknown reservation price distribution) as well as the market size. From a modeling perspective, the choice of a normally distributed demand could potentially lead to estimation errors if the probability of a negative demand is not negligible (e.g., low moving items). One possible way to take care of this problem without losing tractability is to consider a log-normal where $d_n = a \exp(\beta^0 + \beta^1 p_n + \epsilon_n)$ so that by taking logarithm we recover a linear model (see Ref. 14 for a discussion of related demand model in the context of retail operations).

NONPARAMETRIC MODELS: MAXIMIN AND MINIMAX METHODS

Nonparametric models are concerned with settings where the form of the demand function is not known. We distinguish two classes of nonparametric models. A first one is known as the *maximin approach*. Such criterion induces the optimizer, to maximize the worst-case profit. It is a very conservative approach. This conservatism can be reduced by introducing a carefully chosen “budget” of uncertainty, which limits the set of possible distributions or demand functions [15,16]. Many papers have applied this method in operations management contexts [17], while Ref. 18 applied it specifically to the setting discussed in this article. The other class of nonparametric models is known as the

minimax approach whereby a regret function is introduced that measures the performance of a policy relative to the performance of the optimal policy under a fully known demand function. This approach is less conservative than the maximin and has also been applied to various settings including, recently, to revenue management. We refer the reader to Refs 19–21 and references therein. The first two papers consider a single resource allocation problem and the latter analyzes a network revenue management. In the last section of this article, we discuss in some detail some of these approaches.

Learning is a dimension that is often missing in robust models. Indeed, most of the papers cited above, disregard learning and study an optimization problem that constrains the unknown demand function to some static, prespecified, (nonparametrized) set. Having said that, there exist nonparametric models that do rely on exploiting demand realizations in order to learn the demand function and shape the solution of the optimization problem. Understandably, this approach happens to have more of an algorithmic flavor; see for instance Refs 22 and 23 and (more relevant to our setting) [24,25]. We discuss in more detail the model of Besbes and Zeevi [25], which considers, specifically, problem (P), and relies on learning while using a minimax formulation. The approach used in Ref. 18 will also be discussed below as an application of the maximin criteria applied to problem (P).

Consider again the general formulation (P) of the model. As mentioned in the introduction, the expected value is taken with respect to the probability measure defined on the Poisson process filtration. In both the Bayesian and the demand estimation approaches, the unknown parameters were replaced by their estimate and the probability measure was adjusted accordingly (e.g., in the Bayesian case, $\mathbb{P}$ is replaced by $\mathbb{P}_g$, while in the demand estimate case, $\mathbb{E}$ is replaced by $\mathbb{E}_\epsilon$). Consider now the case where the demand belongs to a nonparametric class of functions. The fact that the demand intensity is unknown makes the seller’s maximization problem (P) ill-defined. In order to

overcome this first challenge, a maximin or a minimax formulation has been proposed, which transforms the problem into a stochastic game. The seller still has to choose a pricing policy and in response, nature (or a clairvoyant adversarial agent) picks the “worst” possible demand function given the seller’s choice. The (real) demand function that is unknown is replaced by the solution of an optimization problem. In other words, the seller selects the pricing policy to guarantee, somehow, the best worst-case performance.

In mathematical terms, the seller’s maximin formulation of problem (P) is given by

$$\begin{aligned} & \sup_{p \in \mathcal{A}} \inf_{\lambda \in \mathcal{H}} \mathbb{E} \left[\int_0^T p_t \lambda(p_t) dt \right] \\ & \text{subject to} \quad N_\lambda^p(T) \leq x_0 \quad (\text{a.s.}), \quad (\text{M}) \end{aligned}$$

where $\mathcal{H}$ is the (possibly infinite-dimensional) set of functions that contains the true demand function.

Maximin and the Relative Entropy Formulation

A maximin approach as presented in the section titled “Dynamic Programming with Incomplete Demand Information”, induces the seller to select the price that will maximize the worst-case expected total revenues: a worst case with respect to the demand functions and the pricing policy adopted. Reducing the conservatism of this approach entails reducing the uncertainty set or the set of possible demand functions. We follow here the approach in Ref. 18, to suggest a robust formulation of Problem (P).

In the setting of problem (P), there is a one-to-one relationship between the demand function $\lambda(\cdot)$ of the Poisson process and the probability measure $\mathbb{P}$ governing the system. Hence, the set of demand functions can be replaced by a set of probability measures and it is in the interest of the seller to pick carefully this set as it will dictate how conservative the solution will be.

Often it is the case that the seller has collected information (through historical data) on the “real” demand function he/she is or will be facing. This information is represented as

a demand function, $\beta(\cdot)$, that is not necessarily the real (and still unknown) demand function $\lambda(\cdot)$. However, with some confidence one can assume that both demand rates are not too “far” apart. We denote by $\mathbb{P}$ (respectively $\mathbb{Q}$) the probability measure induced by a demand rate $\lambda(\cdot)$ (respectively $\beta(\cdot)$). The set of probability measures $\mathbb{P}$ close to $\mathbb{Q}$, from which a clairvoyant will draw the worst possible measure, is defined as follows:

- $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, denoted by $\mathbb{P} \ll \mathbb{Q}$.
- $\mathbb{P}$ satisfies $\mathcal{E}(\mathbb{P} | \mathbb{Q}) \leq \gamma$, where $\mathcal{E}(\mathbb{P} | \mathbb{Q}) := \mathbb{E}_{\mathbb{P}} \ln \eta(T)$ and $\eta(T)$ is the Radon–Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$ [3]. The function $\mathcal{E}$ is known as the *relative entropy* of $\mathbb{P}$ with respect to $\mathbb{Q}$ and γ is some positive constant (sometimes referred to as the *budget of uncertainty*) that captures the seller’s confidence level about how close the true probability measure is to the nominal measure $\mathbb{Q}$ (in a relative entropy sense).

These two conditions define an *uncertainty set* of probability measures. Note that for $\gamma \equiv 0$, this set reduces to $\{\mathbb{P} \equiv \mathbb{Q}\}$.

We consider the following transformation of problem (P) and a variant of problem (M) under a maximin criterion

$$\begin{aligned} J(x, T) &= \sup_p \inf_{\mathbb{P} \ll \mathbb{Q}} \mathbb{E}_{\mathbb{P}} \int_0^T p(t) dN(t) \\ & \text{subject to} \quad \mathcal{E}(\mathbb{P} | \mathbb{Q}) \leq \gamma. \end{aligned}$$

Using Girsanov’s Theorem applied to point processes [3], one can prove that there is a stronger correspondence between probability measures in the uncertainty set defined above and their intensities. In particular, the Poisson process with rate $(\lambda(p_t) : t \geq 0)$ under the (unknown) probability measure $\mathbb{P}$ is a Poisson process under the nominal probability measure $\mathbb{Q}$ but with an intensity $(\beta(p_t)\kappa(t) : t \geq 0)$ where $\kappa(\cdot)$ is some (unknown) nonnegative process. Hence, instead of picking $\mathbb{P}$, it is enough for nature to pick κ under $\mathbb{Q}$. With this alternative formulation, one can solve the seller’s

problem using dynamic programming. The corresponding HJB equation is given by

$$\begin{aligned} \frac{\partial}{\partial t} J(x, t) \\ = \max_p \min_{\kappa} \left\{ \lambda(p) [p\kappa + \theta (\kappa \log \kappa + 1 - \kappa)] \right. \\ \left. + \lambda(p)\kappa [J(x-1, t) - J(x, t)] \right\}, \end{aligned}$$

with border conditions $J(x, 0) = J(0, t) = 0$ and where the constant θ is fully characterized given γ (see Ref. 18 for details). One interesting feature of this equation is the interchangeability of min and max under strong concavity of the nominal revenue rate. Moreover, as in Ref. 5, a closed-form solution can be obtained for an exponential demand function (see Ref. 18 for details).

Minimax and the Regret Function

The recent work of Besbes and Zeevi [25] studies a minimax variant of problem (M). In particular, it considers a rather general nonparametric uncertainty set $\mathcal{H}$, which is the set of uniformly bounded and Lipschitz continuous functions with bounded support. A relative regret function, $\mathcal{M}$, is defined as one minus the ratio between the value function, $J^p(x, t; \lambda)$ (generated by any pricing policy p) and the value function in the deterministic case, $J^D(x, t; \lambda)$ (conditioned on knowing the true demand function λ),

$$\mathcal{M}^p(x, t; \lambda) := 1 - \frac{J^p(x, t; \lambda)}{J^D(x, t; \lambda)}.$$

This regret function is inspired by two results in Ref. 5 (see Theorem 3) that establish that (i) the value function under a deterministic demand upper bounds the value function under a Poisson demand and (ii) a fixed-price policy, solution of the deterministic problem, is asymptotically optimal. Hence, the regret function is nonnegative and bounded by 1 (independently of λ), and measures the performance of a pricing policy p for a specific demand function λ . The smaller the regret is, the better the pricing policy is, taking the value zero only when the pricing policy is able to achieve the value of the deterministic

setting. At this point, it remains to transform the initial problem formulation into a stochastic game, where the seller chooses the “optimal” pricing policy taking into consideration that nature will then pick the worst possible demand curve (i.e., by maximizing the regret). The minimax *relative* regret formulation is written as follows

$$\inf_{p \in \mathcal{P}} \sup_{\lambda \in \mathcal{H}} \mathcal{M}^p(x, t; \lambda).$$

For any pricing policy p the inner maximization problem is (at least in theory) well defined. Solving the previous problem is in general not possible. One approach would be to construct a pricing policy that performs well. As we recalled in Theorem 3, when the demand function is known, say $\tilde{\lambda}$, a fixed-price policy $p_t = p^D(\tilde{\lambda})$ is asymptotically optimal in the context of Ref. 5. In the case where the demand curve is unknown, Besbes and Zeevi [25] propose a policy that aims at estimating, first, the demand function λ and then fix the price to be equal to $p^D(\lambda)$. Specifically, they explore, first, how demand reacts to a set of selected prices (passive learning phase) and then exploit this learning to get approximations of both the run-out price, p^0 and the maximizer of the demand function, p^* . (See the discussion that precedes Theorem 3 for the definitions of p^0 and p^* .) In the second phase, they adopt a fixed price policy, which is the minimum of the two approximated prices. The longer the learning phase is, the more accurate the price approximations are; but the higher the opportunity cost is (due to the nonoptimal pricing during the learning phase). Algorithm 1 is sensitive to the length of the learning period τ and to a granularity parameter κ and follows the following steps. (i) In the *initialization* step, κ prices are picked equidistantly from the interval of possible prices $[\underline{p}, \bar{p}]$. The interval $[0, \tau]$ is divided in κ equal time intervals, $\Delta = \tau/\kappa$. (ii) In the *learning/experimentation* step, each of the κ preselected prices is applied on one of the κ intervals of time of length Δ . Demand gets realized accordingly and at the end of this phase, the total demand realized in each of these intervals is computed; each demand value is divided by Δ thus representing an

approximation of the demand rate function $\hat{d}(p_i)$ at the price selected for that interval. The approximated demand rate is in turns multiplied by the price and the result, $\hat{d}(p_i) \cdot p_i$, is an approximation of the revenue rate at that price. At the end of these two steps, κ points approximating the demand and the revenue rate functions have been gathered. (iii) The *optimization* step identifies two specific prices

$$\hat{p}^u = \arg \max_{1 \leq i \leq \kappa} \{\hat{d}(p_i) \cdot p_i\}, \quad \text{and}$$

$$\hat{p}^c = \arg \min_{1 \leq i \leq \kappa} \{\hat{d}(p_i) - x/T\}.$$

(iv) In the last *pricing* step, the price $\hat{p} = \hat{p}^u \wedge \hat{p}^c$ is applied on the interval $(\tau, T]$. We denote by $p(\tau, \kappa)$ the output of Algorithm 1. In terms of performance, this policy is shown to be asymptotically optimal in the following sense.

Proposition 1. *Set $\tau_x = x^{-1/4}$, $\kappa_x = x^{1/4}$ and let $p_x := p(\tau_x, \kappa_x)$ be given by Algorithm 1. Then the sequence $\{p_x\}$ is asymptotically optimal, and for all $x \geq 1$*

$$\frac{C'}{x^{1/2}} \leq \sup_{\lambda \in \mathcal{H}} \mathcal{M}_x^p(x; \lambda) \leq \frac{C(\log x)^{1/2}}{x^{1/4}},$$

for some positive constant C and C' .

The lower bound (in fact obtained under slightly stronger assumptions) measures the decreasing gap between Algorithm 1 and the optimal solution.

Clearly, this approach can also be applied to the case where the demand function belongs to a parameterized set $(\lambda(\cdot; \eta) \in \mathcal{H}_\eta)$, as it was the case in both the Bayesian and the demand estimate methods. It is easily shown that the pricing policy generated by Algorithm 1 remains asymptotically optimal. The example in the section titled “Dynamic Pricing with Complete Demand Information” is a good illustration of such a setting and how effective is a pricing policy driven by a short learning period followed by a fixed price policy. Moreover and as expected, Algorithm 1’s performance improves in the parameterized case. The corresponding upper bound of Proposition 1 is tighter and the number

of pricing test points during the learning phase is set equal to the number of unknown parameters (i.e., bounded and generally small). Such an interesting result requires that the parameterized demand function satisfies additional regularity assumptions. It needs to be “identifiable” based on a set of observations, which in particular means that for any vector $d = (d_1, \dots, d_k)$, the system of equations $\{\lambda(p_i; \eta) = d_i, i = 1, \dots, k\}$ has a unique solution in η .

Nonparametric approaches relying on robust controls, as we saw above, do not follow so much a general framework or methodology as much as their parametric counterparts do (e.g., Bayesian or statistical estimation methods). Nonparametric approaches are stemmed essentially from the specific model under study. In order to have a broader sense of the different techniques that could be used, we discuss in the next section another revenue management setting that has recently been analyzed in the context of robust controls. Again, the main drawback of these approaches is the lack of learning. They do induce robust decisions, however the realizations of demand are not being incorporated in any way to reduce the uncertainty set.

Robust Single-Leg Revenue Management Model

The so-called single-leg model is to revenue management what the newsboy model is to inventory management. (A broad literature is available on this subject and is very well summarized in Ref. 2.) The setting in this single-leg revenue management model is almost identical to the dynamic pricing model that we have discussed so far with one notable difference; items can be sold at different prices at the same time. The canonical example corresponds to seats in an airplane that, depending on the different restrictions that the airline imposes on the products (e.g., refundable vs nonrefundable tickets), are sold at different fares. These fare classes are typically announced at time 0 and remain fixed throughout the entire selling season. In this setting, the seller, rather than controlling a single price over time, selects dynamically which fare classes to *open* and which ones to *close* as a function of the available inventory

and remaining sale horizon. Nonparametric methods have been used recently in the literature to formulate and solve this “cousin” version of Problem (P).

As before, let x_0 be the seller’s initial inventory and let $f_1 > f_2 > \dots > f_m$ be the fares associated to the m different products offered by the seller. We assume that every product consumes one unit of inventory. (In the general setting of *network revenue management*, the seller is endowed with a collections of different resources and different products consume a combination of resources in different quantities.) The seller’s objective is to select an admission control policy π that specifies which fare classes to accept (open) and which ones to reject (close) based on the state of the system at every point in time. A renowned strategy is the *standard nesting* policy, which works as follows. The seller selects *booking limits* $(\theta_1, \theta_2, \dots, \theta_m)$ associated to the m fare classes, such that $\theta_1 \geq \theta_2 \geq \dots \geq \theta_m$. A class- j request is accepted (i.e., to whom one unit of resource is sold at f_j) only if the number of customers of classes j to m (i.e., corresponding to fare f_j or lower) already accepted is less than or equal to θ_j .

Ball and Queyranne [19] analyze this problem in an environment where demand information is incomplete using critical ratio ideas, which are popular in the context of competitive analysis for on-line algorithms (see Ref. 26 for a survey on the subject). We denote by $R^\pi(I)$ the revenues generated by applying some policy π to a stream of arrivals, I . We call this an “on-line algorithm.” We let $R^*(I)$ be the revenues obtained by applying the best possible policy for that particular stream of arrivals I . This is the “off-line algorithm,” that is, revenues that can be achieved if the seller selects the policy after he/she sees the entire stream of arrivals. An on-line algorithm can be deterministic (i.e., for a specific stream of demand the output is deterministic) or randomized (i.e., for a specific stream of arrivals, the output depends on the outcome of some random variable). We refer the reader to Ref. 20 for a brief discussion of randomized policies.

We let Ω_π be the set of all possible demand streams I that a policy π can face and define the competitive ratio as

$$\Upsilon(\pi) := \inf_{I \in \Omega_\pi} \frac{\mathbb{E}R^\pi(I)}{R^*(I)}.$$

We denote by c^* the supremum of Υ on all possible (deterministic or randomized) policies π . It is not hard to see that by maximizing the competitive ratio we are minimizing the worst-case relative regret $\sup [1 - \mathbb{E}R^\pi(I)/R^*(I)]$. We should also note that computing c^* requires a pathwise deterministic optimization, as the expected value is only with respect to the possibly randomized algorithm π . This relative regret formulation is different from the one defined in Ref. 25. In the latter, the denominator of the regret function was the value function under a deterministic demand. In a competitive ratio setting, the denominator is a function of the specific stream of arrivals. These two denominators are asymptotically the same in the case of Problem (P) when the demand function is known.

The competitive ratio of a particular policy π , measures its Performance, which is to be compared to c^* . In particular, a policy that achieves c^* is optimal. Ball and Queyranne [19] use this method to evaluate c^* for a multifare single-leg problem and show that

$$c^* = \left(m - \sum_{j=2}^m \frac{f_j}{f_j - 1} \right)^{-1}.$$

It is also shown that standard nesting policies achieve c^* (and hence are optimal under a competitive ratio criterion) when the (nested) booking limits, θ_j ’s are given by

$$\theta_j = x_0 - \frac{x_0}{c^*} \left(j - \sum_{i=1}^j \frac{f_{i+1}}{f_i} \right).$$

The approach adopted to prove such a result is as follows. One first recognizes the optimal policy, (e.g., the standard nesting) and calculates its optimal competitive ratio (e.g., find the best possible booking limits). It remains then to show that this competitive ratio is the best achievable among all

other (deterministic or randomized) policies. The first step is proven by dissecting all possible arrival streams with respect to the candidate policy. The most critical step is the second one that often relies on a powerful result, the Neuman–Yao principle, which states that

$$c^* \geq \sup_{\pi_d} \mathbb{E}_Q \frac{R^{\pi_d}(I_Q)}{R^*(I_Q)},$$

where I_Q is a random stream of arrivals chosen according to any probability distribution Q , and π_d is any deterministic on-line algorithm. Basically, it shifts the search for the optimum from a pathwise search under both randomized and deterministic algorithms to a probability distribution on the stream of arrivals under deterministic algorithms (see Ref. 27 that provides a framework to prove bounds using the aforementioned principle). Lan *et al.* [20] build on the competitive ratio approach of Ball and Queyranne [19] to solve, among other things, for the case where the demand for each fare has known bounds.

Instead of using a competitive ratio criterion (equivalently a minimax relative regret) one can use other criteria such as the absolute minimax regret or the maximin criteria as defined in the previous section. As a matter of fact, for the multifare single-leg problem, Lan *et al.* [20] prove under an absolute minimax regret criterion that nested policies are again optimal. They propose a unified framework for the competitive ratio and the minimax absolute regret and prove that the absolute regret booking limits are more aggressive than those obtained through the competitive ratio criterion. Finally, Perakis and Roels [21] apply both a maximin and the absolute minimax criteria to a network revenue management problem where demand belongs to a polyhedral uncertainty set (generalizing the single-leg model in Ref. 20). In this more complex setting, they restrict their attention to a set of control policies that include the nested booking limits and develop robust formulations.

SUMMARY

In this article, we have reviewed a representative model of revenue management and discussed different methods to deal with incomplete demand information. Each method depends on the specific context and the level of information available. Parametric models open the way for active learning approaches in which demand realizations are used to update estimates of the unknown parameters over time. These estimates are generally embedded within a dynamic programming formulation. The solution of the DP characterizes the pricing policy to adopt. Nonparametric models are less tractable. A first method that we discussed is based on the idea of replacing the demand uncertainty by a stochastic game using a maximin approach. In order to reduce the conservatism of this formulation, robust optimization is used limiting the feasible set. Eventually, the pricing policy is obtained by solving a complex ODE. An alternative nonparametric model uses passive learning during a specified interval of time, followed by a fixed-price policy for the remaining of the horizon. The pricing solution is simple to implement and asymptotically optimal. We finally discussed the competitive ratio criteria applied to the setting of a capacity allocation revenue management problem and compared that to the other robust criteria available in the literature.

Acknowledgment

The research of the second author was supported in part by FONDECYT grant #1100712.

REFERENCES

1. Bitran GR, Caldentey R. An overview of pricing models for revenue management. *Manuf Serv Oper Manage* 2003;5:203–229.
2. Talluri K, van Ryzin G. The theory and practice of revenue management. Boston (MA): Kluwer Academic Press; 2004.
3. Brémaud P. Point processes and queues, martingale dynamics. New York: Springer; 1980.

4. Bitran GR, Caldentey R, Mondschein SV. Periodic pricing of seasonal products in retailing. *Manage Sci* 1997;43(1):64–79.
5. Gallego G, van Ryzin G. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Manage Sci* 1994;40: 999–1020.
6. Bertsimas D, Perakis G. Dynamic pricing: a learning approach. In: Hearn D, Lawphongpanich S, editors. Volume 101, Mathematical and computational models for congestion charging, Book series: applied optimization. New York: Springer; 2006. pp. 45–79.
7. Aviv Y, Pazgal A. Pricing of short life-cycle products through active learning. Working Paper, St. Louis (MO): Olin School of Business, Washington University; 2002.
8. Farias VF, Van Roy B. Dynamic pricing with a prior on market response. Working Paper, Stanford University, California; 2007.
9. Araman V, Caldentey R. Dynamic pricing for non-perishable products with demand learning. *Oper Res* 2009;57(5):1169–1188.
10. Aviv Y, Pazgal A. A partially observed markov decision process for dynamic pricing. *Manage Sci* 2005;51(9):1400–1416.
11. Lin KY. Dynamic pricing with real-time demand learning. *Eur J Oper Res* 2006;174(1): 522–538.
12. Lobo MS, Boyd S. Pricing and learning with uncertain demand. Working Paper, Duke University; 2003.
13. Carvalho AX, Puterman ML. Dynamic pricing and learning over short time horizons. Working Paper, Vancouver: Statistics Department, University of British Columbia; 2004.
14. Smith SA, Achabal D. Clearance pricing and inventory policies for retail chains. *Manage Sci* 1998;44(3):285–300.
15. Ben-Tal A, Nemirovski A. Robust convex optimization. *Math Oper Res* 2003;23: 769–805.
16. Bertsimas D, Sim M. The price of robustness. *SIAM J Optim* 2005;15:780–804.
17. Bertsimas D, Thiele A. Robust convex optimization. *Math Oper Res* 1998;23: 769–805.
18. Lim AEB, Shanthikumar JG. Relative entropy, exponential utility, and robust dynamic pricing. *Oper Res* 2006;55(2): 198–214.
19. Ball M, Queyranne M. Toward robust revenue management: competitive analysis of online booking. *Oper Res* 2009;57(4): 950–963.
20. Lan Y, Huina G, Ball OB, *et al.* Revenue management with limited demand information. *Manage Sci* 2009;54(9):1594–1609.
21. Perakis G, Roels G. Robust controls for network revenue management. *Manuf Serv Oper Manage* 2010;12(1):56–76.
22. McGill JI, van Ryzin GJ. Revenue management without forecasting or optimization: an adaptive algorithm for determining airline seat protection levels. *Manage Sci* 2000;46(6):760–775.
23. Eren S, Maglaras C. Pricing without market information. Working Paper, Columbia University; 2006.
24. Levin Y, Levina T, McGill J, *et al.* Dynamic pricing with online learning of general reservation price distribution. Refereed Proceedings of INCOM 2006:12th IFAC Symposium on Information Control Problems in Manufacturing. Saint Etienne: Elsevier; 2006.
25. Besbes O, Zeevi A. Dynamic pricing without knowing the demand function: risk bounds and near-optimal algorithms. *Oper Res* 2009; 57:1407–1420.
26. Albers S. Online algorithms: a survey. *Math Program* 2003;97(1-2):3–26.
27. Seiden SS. A guessing game and randomized online algorithms. Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC2000); 2000; Portland (OR). pp. 592–601.

REVENUE MANAGEMENT

GUILLERMO GALLEGO
IEOR, Columbia University, New
York, New York

INTRODUCTION

Revenue Management (RM) refers to techniques to optimally or near-optimally allocate capacity among different fare classes to maximize expected revenues from perishable resources. RM originated in the airline industry but its use has spread to hotels, car rentals, restaurants and other industries. In the airline industry, the resources are the seats in the airline network. There may be several fares associated with a flight itinerary. A fare is a combination of a price and a set of restrictions such as advance purchase, Saturday night stay, and limited seat selection. This is in essence service versioning where the quality of the service is modified to allow sales to multiple market segments at different prices. Allocation capacity among fares is difficult because demands are random and the allocation to lower fares often needs to be done before observing demand for higher fares. In addition, the very realization of demand depends on the capacity allocation policy. This is because as fares are closed, customer demand may shift to other fares or may be lost. Cancellations and no-shows complicate the problem and are mitigated by the careful and systematic use of capacity overbooking. Excellent books on the subject include Talluri and van Ryzin [1] and Phillips [2]. Previous surveys include McGill and van Ryzin [3], Bitran and Caldentey [4], and Elmaghraby and Keskinocak [5]. The aim of this article is to provide a brief and updated synthesis of RM covering both, single resource and network models and both independent and dependent demands. The article also provides suggestions for research at a more strategic level to help expand the practice of RM to other industries. Due to space

limitations, overbooking and forecasting are not covered. We refer the readers to the corresponding chapters in Talluri and van Ryzin [1] and Phillips [2] for these important topics. The article is organized as follows: the section titled “Single Resource RM with Independent Demands” covers the single resource problem with independent demands; the section titled “Network RM with Independent Demands” covers network models with independent demands; the section titled “Network RM with Discrete Choice Models” covers network models under discrete choice models, and the section titled “The Future of Revenue Management” provides suggestions for future research.

SINGLE RESOURCE RM WITH INDEPENDENT DEMANDS

We assume that the capacity provider has c units of perishable capacity of a single resource, such as seats on a specific flight, to be sold using n differentiated fares. More precisely, we assume n fixed fares $p_n < \dots < p_1$ with lower fares having more severe purchasing and traveling restrictions. Given the advance purchase restrictions associated with low fares, it is natural to assume that demand for fare classes arrive in a low-to-high pattern with fare class n arriving first, followed by $n - 1$, with fare class 1 arriving last. Let D_j denote the random demand for fare class j with finite mean $\mu_j = E[D_j]$. We assume that capacity provider is risk-neutral, and her objective is to allocate capacity among fares to maximize expected revenues. We assume that $D_1, \dots, D_n$ are independent random variables for the given fares. The independence assumption is approximately valid in situations where fares are well spread and there are alternative sources of capacity. A customer who finds that his/her preferred fare is unavailable, perhaps because the provider closed down the fare, is more likely to buy a similar fare from a competitor than to buy the next higher fare. The case

of dependent demands, where fare closures may result in demand recapture, is the subject of much active research and will be covered later in this article.

Given that low fares book first, it is important to decide how many units of capacity to reserve for later higher fare bookings. This is best done using dynamic programming (DP). Let $V_j(x)$ denote the maximum expected revenue that can be obtained from $x \in \{0, 1, \dots, c\}$ units of capacity from fare classes $\{j, \dots, 1\}$. The sequence of events for stage j are as follows: (i) observe the remaining inventory x ; (ii) decide the protection level, say $y \leq x$, for fares $\{j-1, j-2, \dots, 1\}$, leaving $x-y$ units available for sale at fare j ; (iii) D_j occurs, resulting in sales $\min(D_j, x-y)$ and revenue $p_j \min(D_j, x-y)$; and (iv) proceed to stage $j-1$ with capacity $\max(x-D_j, y)$.

Let $W_j(y, x) = p_j E \min(D_j, x-y) + EV_{j-1}(\max(x-D_j, y))$ and consider the DP recursion

$$V_j(x) = \max_{y \leq x} W_j(y, x). \quad (1)$$

The recursion can be started with $V_0(x) = 0, x \geq 0$ or with $V_1(x) = p_1 E \min(D_1, x)$ for $x \geq 0$ since it is optimal to make all the remaining capacity available for sale at p_1 . Assuming discrete distributions, the computational requirement to solve the DP for n stages is of order $O(nc^2)$. One can, in the process of solving the dynamic program, find an optimal policy that is, for every stage j , record for each x the largest protection level y maximizing Equation (1).

An optimal protection policy for fares $\{j, \dots, 1\}$ is a function, say $y_j(x) \leq x$, determining how many units of capacity to protect for each $x \in \{1, \dots, c\}$. Fortunately, an optimal can be characterized by a single critical protection level y_j^* such that $y_j(x) = \min(x, y_j^*)$ for each $j = 1, \dots, n-1$. It helps to think of y_j^* as the desired protection level for fares $\{j, \dots, 1\}$, so if $x \geq y_j^*$ then $y_j(x) = y_j^*$ units are protected and if $x < y_j^*$ then all $x = y_j(x)$ units are protected. Moreover, the critical integers y_j^* can be selected so that they are nondecreasing in j .

The key to the establishment of the form of the optimal policy and the determination of the critical protection levels is through

marginal analysis. For $y \in \mathcal{N}_+ = \{1, 2, \dots\}$ let $\Delta V_j(y) = V_j(y) - V_j(y-1)$, denote the marginal value of the y th unit of capacity at stage $j = 1, \dots, n$. It is possible to show that $\Delta V_j(y)$ is decreasing in y and increasing in j and that

$$\begin{aligned} \Delta W_j(y, x) &= W_j(y, x) - W_j(y-1, x) \\ &= [\Delta V_{j-1}(y) - p_j] Pr(D_j > x-y), \end{aligned}$$

for all $y \in \{1, \dots, x\}$. This implies that $W_j(y, x)$ is nondecreasing on $\Delta V_{j-1}(y) > p_j$ and nonincreasing on $\Delta V_{j-1}(y) \leq p_j$.

Let

$$\begin{aligned} y_j^* &= \max \{y \in \mathcal{N}_+ : \Delta V_j(y) > p_{j+1}\}, \\ j &= 1, 2, \dots, n-1, \end{aligned} \quad (2)$$

where y_j^* is set to zero if the set is empty. Notice that y_j^* is independent of demands $D_k, k > j$. It follows that at stage j , when we are maximizing $W_j(y, x)$ subject to $y \leq x$, we should select $y = \min(y_{j-1}^*, x)$. Indeed, if $x \leq y_{j-1}^*$ then $W_j(y, x)$ is nondecreasing in y up to x and if $x > y_{j-1}^*$ then $W_j(y, x)$ is nondecreasing up to y_{j-1}^* and nonincreasing beyond y_{j-1}^* . Moreover, it follows that $y_j^* \geq y_{j-1}^*$ since $\Delta V_j(y_{j-1}^*) \geq \Delta V_{j-1}(y_{j-1}^*) > p_j > p_{j+1}$ and $\Delta V_j(y)$ is decreasing in y . It is possible to relax the assumption that the fares are monotone $p_1 > p_2 > \dots > p_n$ but continue to assume that demands arrive sequentially in the order $D_n, D_{n-1}, \dots, D_1$. The dynamic program is unchanged, but there are some additional insights that can be gleaned. In particular, the monotonicity of the protection levels no longer holds as it is possible to show that $y_{j-1} = 0$ whenever $p_j > \max(p_1, \dots, p_{j-1})$. Dynamic programs for the multifare problem were developed independently by Brumelle and McGill [6] and Wollmer [7], with refinements by Robinson [8] who allows for nonmonotonic fares and proposes a Monte Carlo simulation approach to determining booking limits. van Ryzin and McGill [9] exploit the structure in Brumelle and McGill [6] to propose a method for setting protection levels that do not require forecasting or optimization.

Example 1. Suppose capacity is 50 and there are four fares $p_1 = 100, p_2 = 90$,

$p_3 = 80, p_4 = 70$ with independent Poisson demands with parameters $\lambda_1 = 10, \lambda_2 = 15, \lambda_3 = 25$, and $\lambda_4 = 15$. The optimal protection levels are $y_1 = 6, y_2 = 20$, and $y_3 = 44$. To see how the policy works assume that realization of demands for the four fares turn out to be $(d_4, d_3, d_2, d_1) = (13, 23, 14, 12)$. Since $c = 50$, and $y_3 = 44$, only 6 units of capacity are available to fare 4, so 4 units are sold at this fare, with residual capacity now at 44. Since $y_2 = 20$, we are allowed to book up to 24 units at fare 3, and since the demand for fare 3 is 23 we accept them all. Our residual capacity is now 21. Since we protect 6 units for fare 1, we allow up to 15 units to book at fare 2. Since the fare 2 demand is 14 we also accept them all. The residual capacity now drops to 7 and since demand for fare 1 is 12 we sell 7 units at fare 1.

Littlewood's Rule and Commonly Used Heuristics

For $n = 2$, $\Delta V_1(y) = p_1 Pr(D_1 \geq y)$ implying that

$$y_1^* = \max \{y \in \mathcal{N}_+ : p_1 Pr(D_1 \geq y) > p_2\}, \quad (3)$$

where y_1^* is set to zero if the set is empty. This is known as Littlewood's rule, [10], and it protects capacity for sale at fare p_1 as long as the marginal value exceeds p_2 . Several heuristics based on Littlewood's formula for the multifare problem were developed in the late 1980s before an optimal policy based on DP was discovered in the 1990s. For a while, some of these heuristics were thought to be optimal by some people. The heuristics are easy to understand and to compute so they were readily coded and embedded in RM systems. To this day, heuristics remain widely used, as they perform very well resulting in expected revenues that are very close to optimal. The two most popular heuristics are based on different versions of what is called the expected marginal seat revenue (EMSR).

Version *a* of the heuristic, EMSR-a, is based on the idea of adding protection levels produced by applying Littlewood's rule to successive pairs of fare classes. More precisely, $y_j^a = \sum_{k=1}^j y_{k,j}$ units are

protected for fares $\{j, \dots, 1\}$ where $y_{k,j} = \max\{y \in \mathcal{N} : p_k Pr(D_k \geq y) > p_{j+1}\}$. Version *b* is based on a single call to Littlewood's rule. To find the protection level for fares $\{j, \dots, 1\}$, aggregate demands $D[1, j] = \sum_{k=1}^j D_k$, use the weighted average fare $\bar{p}_j = \sum_{k=1}^j p_k \mu_k / \mu[1, j]$, where $\mu[1, j] = \sum_{k=1}^j \mu_k$, and apply Littlewood's rule to obtain $y_j^b = \max\{y \in \mathcal{N} : \bar{p}_j Pr(D[1, j] \geq y) > p_{j+1}\}$. Both of these heuristics perform very well with losses, relative to the DP solution, that are rarely above 0.5%, but EMSR-b is more commonly used. These heuristics were first published by Belobaba [11,12], but were probably used earlier at American Airlines.

Upper and Lower Bounds on the Value Function

It is sometimes useful to have bounds on the optimal expected revenue $V_n(c)$ as bounds are often used in revenue opportunity models (ROMs) developed for the purpose of assessing how well a given RM system performs relative to a readily computable upper bound. Consider the perfect foresight problem where the demand vector $D = (D_1, \dots, D_n)$ is known in advance. Perfect foresight allows us to optimally allocate capacity by solving a knapsack problem where x_k represents the capacity allocated to fare class k :

$$\begin{aligned} V_n(c, D) &= \max \sum_{k=1}^n p_k x_k, \\ \text{s.t. } x_k &\leq D_k \quad \forall \quad k = 1, \dots, n \\ \sum_{k=1}^n x_k &\leq c \\ x_k &\geq 0 \quad \forall \quad k = 1, \dots, n. \end{aligned} \quad (4)$$

The solution to Equation (4) is given by $x_k(D) = \min(D_k, (c - D[1, k-1])^+)$, $k = 1, \dots, n$ and corresponds to having fares arrive high-to-low with zero protection levels for lower fares. A little algebra shows that

$$V_n(c, D) = \sum_{k=1}^n (p_k - p_{k+1}) \min(D[1, k], c),$$

where for convenience we define $p_{n+1} = 0$. Because the allocations are done with perfect foresight, for every realization of D , $V_n(c, D)$

Table 1. Normal Demand Parameters and Protection Levels for Example 2

j	p_j	Demand Parameters		Protection Levels (y_j)		
		μ_j	σ_j	OPT	EMSR-a	EMSR-b
1	1050	17.3	5.8	9.7	9.8	9.8
2	950	45.1	15.0	54.0	50.4	53.2
3	699	39.6	13.2	98.2	91.6	96.8
4	520	34.0	11.3			

is at least as large as the revenue under the optimal dynamic policy for low-to-high arrivals. Thus, $V_n(c) \leq EV_n(c, D) \equiv V_n^h(c)$. Moreover, $V_n^h(c) \leq V_n(c, \mu)$, where $\mu = E[D]$, on account of Jensen's inequality and the fact that $V_n(c, D)$ is concave in D . The upper bound $V_n^h(c)$ is rarely computed, but it can be significantly sharper than $V_n(c, \mu)$ over a significant range of c values. A lower bound can be obtained by assuming a low-to-high arrival pattern with zero protection levels. This gives rise to sales $\min(D_k, (c - D[k+1, n])^+)$ at fare $k = 1, \dots, n$ and expected revenue

$$V_n^l(c) = \sum_{k=1}^n (p_k - p_{k-1}) E \min(D[k, n], c),$$

where for convenience we define $p_0 = 0$. In summary, $V_n^l(c) \leq V_n(c) \leq V_n^h(c) \leq V_n(c, \mu)$.

Example 2 [From Wollmer [7]]. There are four classes and the demand is assumed to be normally distributed. Table 1 shows the demand data and the protection levels under optimal, EMSR-a, and EMSR-b policies. The expected revenues of the optimal policy and the suboptimality gaps of the heuristics were estimated through a simulation study with 500,000 replications. The results are shown in Table 2 where we also include the bounds on $V_n(c)$.

All of the models reviewed so far treat time implicitly through the stages associated with different fare bookings. Significant work has also been done for single-resource problems where time is modeled explicitly; see Lee and Hersh [13], Feng and Gallego [14,15],

Feng and Xiao [16], and Weng and Zheng [17] for research involving the optimal time to switch between fares. See also Bitran and Mondschein [18] and Gallego and van Ryzin [19] for the related dynamic pricing problem where the price menu is not set in advance.

NETWORK RM WITH INDEPENDENT DEMANDS

Consider a network with several resources that can be combined in different ways to produce different products. In the context of an airline, the resources are the flight legs that are combined to form itineraries or routes from an origin to a destination. An origin-destination fare (ODF) represents an itinerary and a fare, so there may be many different fares corresponding to an itinerary with lower fares often being more restricted. Suppose that there are m resources in the network and let $c \in \mathbb{Z}^m$ denote the vector of initial capacities. We will assume that resources cannot be replenished during the sales horizon. Unlike the single-resource case, where time was modeled implicitly, we model time explicitly for network models. We will measure time backwards, so t will represent the time-to-go. Initially, $t = T$ and at the date of departure $t = 0$. For any time-to-go t we will let (t, x) represent the state of the system where x represents the vector of remaining inventories. The initial state is (T, c) . Let $V(t, x)$ denote the maximum expected revenue from state (t, x) . The main goal is to determine optimal or near-optimal admission control policies that tells us which ODF to accept at state (t, x) . Other goals may include computing $V(T, c)$ itself, or bounds and approximations on $V(T, c)$. We will assume that there are K itineraries, also known as origin destinations (ODs) and for each OD $k \in \{1, \dots, K\}$ there are n_k ODFs: $p_{k1} > p_{k2} > \dots > p_{kn_k}$. Corresponding to each of these fares there is a time-dependent arrival rate λ_{tkj} , so demand for fare p_{kj} over $[0, t]$ is Poisson with parameter $\int_0^t \lambda_{skj} ds$. Let A_k be the vector of resources utilized by OD k . Then A_k is an m -dimensional vector that corresponds to OD k . In the context of an airline, A_k is a vector of zeros and ones with

Table 2. Optimal Revenue and Bounds for Example 2

c	$V_n^l(c)$ (\$)	$V_n(c)$ (\$)	$V_n^h(c)$ (\$)	$V_n(c, \mu)$ (\$)	Sub-a (%)	Sub-b (%)
80	52,462	67,505	72,717	73,312	0.10	0.00
90	61,215	74,003	79,458	80,302	0.06	0.00
100	70,136	79,615	85,621	87,292	0.40	0.02
110	78,803	84,817	91,122	92,850	0.35	0.02
120	86,728	89,963	95,819	98,050	0.27	0.01
130	93,446	94,869	99,588	103,250	0.15	0.01
140	98,630	99,164	102,379	106,370	0.06	0.01
150	102,209	102,418	104,251	106,370	0.01	0.00
160	104,385	104,390	105,368	106,370	0.00	0.00

a one for each resource consumed by OD k . Consider now a time increment δt that is small enough so that we can think of $\lambda_{tkj}\delta t$ as the probability that an arrival occurs at time t for ODF kj . Then,

$$\begin{aligned}
V(t, x) &= \sum_{k=1}^K \sum_{j=1}^{n_k} \lambda_{tkj} \delta t \max\{V(t - \delta t, x), p_{kj} \\
&\quad + V(t - \delta t, x - A_k)\} + \left(1 - \sum_{k=1}^K \sum_{j=1}^{n_k} \lambda_{tkj} \delta t\right) \\
&\quad \times V(t - \delta t, x) + o(\delta t),
\end{aligned}$$

where $o(\delta t)$ is a quantity that goes to zero faster than δt . A discrete-time dynamic program can be obtained by ignoring the $o(\delta)$ term. Rearranging and taking limits we obtain the partial differential equation

$$\frac{\partial V(t, x)}{\partial t} = \sum_{k=1}^K \sum_{j=1}^{n_k} \lambda_{tkj} [p_{kj} - \Delta_k V(t, x)]^+, \quad (5)$$

with boundary conditions $V(t, 0) = V(0, x) = 0$ for all $x \geq 0$. Here $\Delta_k V(t, x) = V(t, x) - V(t, x - A_k)$ if $x \geq A_k$. Formulation (5) indicates that the marginal value of time is equal to the net revenue in excess of displacement costs. An optimal admission control rule at state (t, x) is to admit requests for which $p_{kj} \geq \Delta_k V(t, x)$, showing that if p_{kj} is admitted then so are all higher fares associated with OD k .

It is conceptually useful to think of $\lambda_{tk} = \sum_{j=1}^{n_k} \lambda_{tkj}$ as the arrival rate of requests for OD k and to write the arrival rates of specific

fares $\lambda_{tkj} = \lambda_{tk} q_{tkj}$ as thinned by probabilities q_{tkj} . This is equivalent to having random fare P_{tk} at time t for OD k , where P_{tk} takes value p_{kj} with probability q_{tkj} . Writing Equation (5) in terms of random fares yields the equivalent formulation

$$\frac{\partial V(t, x)}{\partial t} = \sum_{k=1}^K \lambda_{tk} E[P_{tk} - \Delta_k V(t, x)]^+. \quad (6)$$

Formulation (6) makes explicit the arrival rate for each itinerary and the fact that distribution of fares for an itinerary may change over time to reflect time preferences or fare restrictions. Formulation (6) is actually more general than the motivation that led to it as it is valid for any random fare P_{tk} with finite means. Viewing P_{tk} as a general random variable may be helpful to deal with net fares that depend on channel distribution costs. While formulation (6) is rich in detail, it is sometimes convenient to work with a more streamlined formulation that aggregates ODFs into a single index with $n = \sum_{k=1}^K n_k$ ODFs. Abusing notation slightly, the single index formulation results in

$$\frac{\partial V(t, x)}{\partial t} = \sum_{j=1}^n \lambda_{tj} (p_j - \Delta_j V(t, x))^+, \quad (7)$$

where in this formulation λ_{tj} and p_j denote respectively the arrival rate and the fare of ODF j . Note that in this formulation there may be multiple identical columns corresponding to different fare classes for the same OD. This formulation has the virtue of simplicity and generality at the cost of abstractness as it distances the model a little from its origins.

The simplicity of formulation (7) allows us to more easily treat the compound Poisson case, to discuss the nested-by-fare structure of optimal solutions and randomness in the arrival rates. If demand for an ODF j is a compound Poisson process with arrival rate λ_{tj} and demand size X_j with $q_j^l = P(X_j = l)$, then we can model this by using columns of the form $A_j^l = lA_j$ with corresponding fare p_j^l and arrival rate $\lambda_{tj}^l = \lambda_{tj}q_j^l$ for all l such that $q_j^l > 0$ and treating each column with corresponding fare and arrival rate as a different ODF with unit requests. This exercise essentially brings back the formulation to the Poisson case. Notice that the total fare p_j^l need not be lp_j thus allowing for economies or diseconomies of scale. Suppose ODFs j and j' consume the same vector of resources, that is $A_j = A_{j'}$ so that they represent the same itinerary but are sold at different fares, say $p_{j'} > p_j$. Then, since $\Delta_j V(t, x) = \Delta_{j'} V(t, x)$, $p_j \geq \Delta_j V(t, x)$ implies $p_{j'} > \Delta_{j'} V(t, x)$, so if it is optimal to accept ODF j it is also optimal to accept all higher fares for the same itinerary. Conversely, if $p_{j'}$ is rejected, then so are all ODFs for the same itinerary that carry a lower fare. If the arrival rates are random for each t , say λ_{tj}^i with probability θ_j^i , then the Hamilton–Jacobi–Bellman (HJB) equation (7) is of the same form with $\lambda_{tj} = \sum_i \lambda_{tj}^i \theta_j^i$.

The Deterministic Linear Program (DLP)

Formulations (6) and (7) are hard to solve. The quest to find good heuristics requires approximating the value function. The simplest approximation is the upper bound that emerges from the perfect foresight model for the network problem. Letting $\Lambda_{Tkj} = \int_0^T \lambda_{skj} ds$ denote the expected demand for ODF kj over the entire sales horizon $[0, T]$, we obtain a linear program whose value function is an upper bound on $V(T, c)$:

$$\begin{aligned} \bar{V}(T, c) = \max & \sum_{k=1}^K \sum_{j=1}^{n_k} p_{kj} y_{kj} \\ \text{subject to} & \sum_{k=1}^K A_k \sum_{j=1}^{n_k} y_{kj} \leq c, \\ & 0 \leq y_{kj} \leq \Lambda_{Tkj} \quad \forall \quad k, j. \end{aligned} \quad (8)$$

This formulation, due to Simpson [20], determines the number of units y_{kj} allocated to ODF kj subject to capacity and *expected* demand constraints. A proof that $\bar{V}(T, c)$ is indeed an upper bound can be found in Chen [21]. Since this problem is based on expected demands, it is known as the *deterministic linear program* (DLP) for RM. We will let $\bar{z}$ denote the vector of optimal dual variables to the capacity constraints.

Example 3. Consider a network with two connecting flights and assume that two fares are available for each individual flight, as well as for the connecting flight. Using the single index model we have 6 ODFs with incident matrix

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The fare vector and arrival rates are as follows:

$$p = (150 \ 100 \ 120 \ 80 \ 250 \ 170),$$

$$\lambda_t = (0.06 \ 0.00 \ 0.04 \ 0.00 \ 0.06 \ 0.00),$$

$$0 < t \leq 500,$$

$$\lambda_t = (0.00 \ 0.12 \ 0.00 \ 0.16 \ 0.00 \ 0.08),$$

$$500 < t \leq 1000,$$

with capacity vector $c = (90, 90)$. Notice that the arrival rates are such that low fare demands arrive before high fare demands. The vector of $\Lambda_{Tj} = \int_0^T \lambda_{tj} dt$ of aggregate intensities is given by $\Lambda_T = (30 \ 60 \ 20 \ 80 \ 30 \ 40)$. The solution to Equation (8) is given by $\bar{y} = (30 \ 30 \ 20 \ 40 \ 30 \ 0)$, the vector of dual prices is given by $\bar{z} = (100, 80)$, and $\bar{V}(T, c) = \$20,600$.

Heuristics Derived from the DLP

Heuristics for the stochastic system can be constructed from solutions to the DLP and its dual. Any vector $z \in \mathbb{R}_+^m$ can be used to describe a *bid-price* heuristic admission control policy for the stochastic network RM problem. The heuristic works as follows: accept request for fare p_{kj} at state (t, x) if and only if $p_{kj} \geq z' A_k$ and $A_k \leq x$. In words, accept

a request if its fare equals or exceeds the sum of bid prices for the resources consumed and there is available capacity.

Talluri and van Ryzin [1] show that if the fares P_{tk} s in formulation (6) are *continuous* random variables that are *uniformly* bounded, then the bid-price heuristic based on $z = \bar{z}$ is asymptotically optimal as the capacity and the demands are scaled up by the same factor. This is an important result for systems that satisfy the stated conditions. In practice, the distribution of P_{tk} is often discrete and leads to both, fully accepted fares $F(p_{kj} > \bar{z}A_k)$ and partially accepted fares $P(p_{kj} = \bar{z}A_k)$. When partially accepted fares arrive before fully accepted fares, the heuristic may admit too many marginal fares. To illustrate this, consider the data of Example 3 and notice that $P = \{2, 4\}$ and $F = \{1, 3, 5\}$ with fares in P arriving before fares in F . The expected revenue of the bid-price heuristic, estimated via simulation, resulted in \$17,736.49 or 14% below the upper bound of \$20,600. This illustrates how bid-price heuristic without limits on marginal fares can cannibalize sales of higher, more profitable, fares.

To illustrate how the asymptotic optimality results fail in the case of discrete fares, consider the two-fare single-resource problem: $p_1 > p_2$, with expected demands $\Lambda_{T1} = (1 - \epsilon)c$ and $\Lambda_{T1} + \Lambda_{T2} > c$. Then $\bar{z} = p_2$ and the bid-price heuristic accepts all requests. If arrivals are low-to-high and $\Lambda_{T2} > \epsilon c$ then the entire capacity will be consumed by low-fare customers when in fact most of the capacity could have been sold at the high fare. One may try to correct things by redefining bid-price heuristics to accept fares only if $p_{kj} > \bar{z}A_k$. For our example, this is equivalent to accepting only fares in $F = \{1\}$, but then things go wrong if $\Lambda_{T1} = \epsilon c$ and $\Lambda_{T1} + \Lambda_{T2} > c$ as most of the capacity gets spoiled when in fact it could have been sold at the low fare. Note that neither version (accepting fares in $F \cup P$ nor F) of the bid-price heuristic improves as we consider scaled systems with capacity ac and aggregate demands $a\Lambda_{Tj}, j = 1, 2$, for values of $a > 1$.

There is fortunately an easy way to restore the asymptotic optimality by modifying the

admission control rule for ODFs in P . The modification is to accept ODFs in P with probability $\gamma_{kj} = \bar{y}_{kj}/\Lambda_{Tkj}$. Fares in F have $\gamma_{kj} = 1$ while fares in $R = \{kj : p_{kj} < \bar{z}A_k\}$ have $\gamma_{kj} = 0$. We will refer to this heuristic as the probabilistic acceptance rule (PAR) heuristic. It can be shown that the PAR heuristic is asymptotically optimal [22,23]. To see how the PAR heuristic works, consider again Example 3. Here we have $\gamma_2 = \gamma_4 = 0.5$. Modifying the heuristic to admit requests for these ODFs with probability 50% results in \$19,426.71 in expected revenues, estimated via simulation, compared to \$17,736.49 from the bid-price heuristic.

In practice, bid-price heuristics perform better if the DLP is resolved frequently during the horizon; for example, at reading dates when forecasts may also be updated. This helps because corrections can be made before too much of the demand is cannibalized by fares in P . Resolving the DLP has good asymptotic properties, see Maglaras and Meissner [24], in the context of a single resource. However, resolving may hurt rather than help performances in some instances; see Cooper [25] for details, particularly toward the end of the horizon when the fluid approximation implied by the DLP suffers the most from ignoring randomness. Resolving four times during the horizon in Example 3 improves expected revenues, estimated through simulation, to \$18,597.21 for the bid-price heuristic, and to \$19,561.88 for the PAR heuristic. Resolving ten times further improves expected revenues to \$19,792.74 and \$19,742.64, respectively, for the bid price and the PAR heuristics. Note that with frequent resolving, the bid-price heuristic does better than the PAR heuristic. This is not unusual as the bid-price heuristic tends to gradually tighten capacity for fares in P . Jasin and Kumar [26] show that using the PAR heuristic with sufficiently frequent resolving results in a *bounded* loss of optimality relative to $V(T, c)$.

Bid prices can also benefit from refinements that make them more sensitive to demand randomness. Talluri and van Ryzin [27] suggest repeatedly solving the DLP problem with simulated, rather than average, demands and then averaging the resulting

bid prices. Likewise, one can obtain average γ values for the probability of accepting requests for the PAR heuristic. Topaloglu [28] shows that randomizing the DLP preserves the asymptotic optimality properties in Talluri and van Ryzin [1]. Adding this refinement to Example 3 and resolving four times during the horizon results in estimated expected profits \$19,525.29 and \$19,406.08, respectively for the bid price and the PAR heuristic. Resolving ten times resulted in estimated expected profits \$19,658.77 and \$19,580.36. Note that while the refinement helped with four resolves, it did not with ten. This observation is consistent with other reports such as de Boer *et al.* [29] and Williamson [30].

Heuristics that Take Randomness into Account

Heuristic capacity allocation methods for ODFs in P that take demand randomness into account are often based on the EMSR-b heuristic. We will explain the most common form of this heuristic using the double index notation, let $F_k = \{kj : p_{kj} > \bar{z}'A_k\}$, $P_k = \{kj : p_{kj} = \bar{z}'A_k\}$ and $R_k = \{kj : p_{kj} < \bar{z}'A_k\}$. Let D_{kF} be a Poisson random variable with mean $\Lambda_{kF} = \sum_{j \in F_k} \Lambda_{kj}$ and let $\bar{p}_{kF} = \sum_{j \in F_k} p_{kj} \Lambda_{kj} / \Lambda_{kF}$. Now consider the itinerary $j \in P_k$ with the highest fare p_{kj} and apply an EMSR-b heuristic to find protection level y_{kF} for fare $\bar{p}_{kF}$ with Poisson demand Λ_{kF} against fare p_{kj} . This EMSR-b type analysis may be of help even for itineraries with $P_k = \emptyset$ by finding protection level y_{kF} for fares in F_k against the largest fare p_{kj} in R_k . Applying this idea to Example 3 results in booking limits 32, 42, and 3 for itineraries 2,4, and 6 respectively. Notice that even itinerary 6 which is rejected by the DLP gets some capacity under this scheme. The expected revenues for this heuristic, estimated via simulation, are \$19,641.80.

A more popular alternative is to use a resource decomposition approach by aggregating ODFs into buckets and to use EMSR-b or a similar heuristic for each resource. Using the single index model, let $I_j = \{i : A_{ij} > 0\}$ be the set of resources utilized by ODF j . The net contribution of ODF j to resource i is defined by subtracting from the fare the

value of the capacity used by ODF j on all other resources. For any vector of bid prices $z \geq 0$, let $p_{ij}(z) = p_j - \sum_{l \neq i} A_{lj} z_l$ if $i \in I_j$ and $p_{ij}(z) = 0$ otherwise. For each resource $i = 1, \dots, m$, we can now solve a single resource RM problem with fares $p_{ij}(z)$ and arrival rates $\lambda_{ij} = \lambda_j$ for all $j \in J_i = \{j : A_{ij} > 0\}$. Note that we no longer assume that the fares will arrive low-to-high, but a robust heuristic is used to compute protection levels as if the arrivals are low-to-high and then use standard, rather than theft nesting¹ [[31], p. 30]. The low-to-high protection levels can be computed using DP or a heuristic such as EMSR-b. If a mixed arrival model is used to compute protection levels, then it is important that it is implemented using theft, rather than standard, nesting. When a request for ODF j arrives at state (t, x) , the request is accepted if the net fare is accepted for each resource i such that $A_{ij} > 0$. The problem with a direct implementation of this approach is that there may be too many different itineraries mapped to resources. Some of these itineraries will have very similar net contributions and some will have very small demands. A natural idea is to aggregate the itineraries into buckets of similar net contributions and aggregate the demands within each bucket. This exercise results in more manageable single-resource allocation problems. This is known as displacement-adjusted virtual nesting (DAVN) [32] and combines ideas of bid prices and single-leg controls.

Consider Example 3 one more time and suppose that three buckets are defined such that bucket 1 contains all net fares greater than \$120, bucket 2 all fares between \$60 and \$120, and bucket 3 all fares below \$60. The itineraries that utilize resource 1, under the single index model, are 1,2,5, and 6 and the net fares for resource 1 are $p_{11}(\bar{z}) = 150$, $p_{12}(\bar{z}) = 100$, $p_{15}(\bar{z}) = 170 = 250 - 80$, $p_{16}(\bar{z}) = 90 = 170 - 80$. Accordingly, fares 1 and 5 belong to bucket 1, with corresponding expected demands 30 and

¹Under theft nesting bookings for higher fare classes, count against the booking limits of lower fare classes.

30. Fares 2 and 6 belong to bucket 2 with corresponding expected demands 60 and 40. We can aggregate fares 1 and 5 of bucket 1 into a single weighted average fare of 160 with aggregate demand 60. Similarly, for bucket 2 we can aggregate fares 2 and 6 into a single weighted average fare of 96 with aggregate demand 100. We can repeat the procedure for resource 2 and obtain net fares $p_{23}(\bar{z}) = 120$, $p_{24}(\bar{z}) = 80$, $p_{25}(\bar{z}) = 150$, and $p_{26}(\bar{z}) = 70$, with corresponding demands 20, 80, 30, and 40. Clearly, fares 3 and 5 map into bucket 1 while fares 4 and 6 map into bucket 2. We can aggregate fares 3 and 5 of bucket 1 into a single weighted average fare of 138 with aggregate demand 50. Similarly, for bucket 2 we can aggregate fares 4 and 6 into a single weighted average fare of 76.67 with aggregate demand 120. The expected revenues for this heuristic, estimated via simulation, are \$19,770.65. It is reasonable to expect even better performance with frequent resolving.

Dynamic Bid-Price Heuristics

A number of authors have developed more refined heuristics for the network RM problem. The main aim of these heuristics is to obtain bid prices that are responsive to changes in inventory and time-to-go. Adelman [33] proposes a heuristic based on approximate dynamic programming (ADP) with time-varying linear approximations. Bertsimas and Popescu [34] propose a heuristic based on accepting requests when the fare p_j exceeds the displacement cost $\bar{V}(t, x) - \bar{V}(t, x - A_j)$ generated from solving the DLP. One particularly successful heuristic is presented by Topaloglu [35] who proposes a Lagrangian relaxation (LR) approach to the network problem that decomposes the problem by flight, see also Talluri [36]. His method provides an approximation to the marginal value of each of the resources for every vector x of residual capacities at every time-to-go t . In his numerical study, Topaloglu reports significant improvements over a number of well-known heuristics including the bid-price heuristic based on the DLP and the refinements proposed by Adelman [33] and Bertsimas and Popescu [34]. The LR

heuristic is particularly good at exploiting time-varying arrival rates and in situations where there is a significant difference between adjacent fares. Data requirements, fare structures, and scalability are things to consider before implementing dynamic bid-price heuristics. If, for example, good estimates of time-varying arrival rates are not available and the fare structure is fairly dense, then the advantages of these methods may not justify the additional complexity.

Another stream of research tries to adjust bid prices and booking limits using simulation-based optimization, see Bertsimas and de Boer [37] who use simulation-based numerical approximations to first finite differences, to obtain improved booking limits. Their numerical studies show that their method improves booking limits generated either by the DLP or by the DAVN. As is usually the case, the advantage over more primitive methods diminishes with the number of times the system is resolved during the horizon.

NETWORK RM WITH DISCRETE-CHOICE MODELS

We now turn our attention to models that allow for dependent demands. In particular, we will be reviewing research based on discrete-choice models [38]. We will assume there are n different ODFs and there are L market segments. We will denote by $N_l \subset \{1, \dots, n\}$ the consideration set of ODFs that are of potential interest to customers in market segment $l \in \mathcal{L} = \{1, \dots, L\}$. At any state (t, x) , the set of ODFs offered for sale must be subsets $S_l \subset N_l, l \in \mathcal{L}$. If the consideration sets are nonoverlapping, that is, $N_l \cap N_{l'} = \emptyset$ for all $l' \neq l$, then there is no need to add further restrictions to the offered sets. This is also true even when consideration sets overlap, if the capacity provider is free to independently decide what fares to offer to each market. As an example, different fares are offered for round-trips from the United States to Asia depending on whether the customer buys the fare in the United States or in Asia. If the provider cannot make independent decisions by market segment then he

needs to post a single offer set $S \subset \{1, \dots, n\}$. In this case, market segment $l \in \mathcal{L}$ will have available fares in $S_l = S \cap N_l$. In addition, S_l customers arriving to market segment l have the no-purchase alternative so they are really selecting from the set S_{l+} that contains S_l and the no-purchase alternative for market segment l . We will follow tradition and abuse notation slightly and write $\pi_{lj}(S_l)$ where it would be more proper to write $\pi_{lj}(S_{l+})$ to denote the probability of selecting $j \in S_{l+}$ from market segment l with $\pi_{lj}(S_l) = 0$ if $j \notin S_{l+}$. One of the most commonly used discrete-choice models is the basic attraction model (BAM):

$$\pi_{lj}(S_l) = \frac{a_{lj}}{a_{l0} + \sum_{k \in S} a_{lk}} \quad j \in S_{l+}, \quad (9)$$

where a_{lk} is the attraction value of ODF $k \in N_l$ and a_{l0} is the attraction of the no-purchase alternative.

For ease of exposition we will assume that customers arrive to market segment l as a Poisson process with time-homogeneous rate λ_l . Let p_{lj} be the fare associated with $j \in N_l$ and let p_l be the vector with components $p_{lj}, j \in N_l$. We will denote by $V(t, x)$ the maximum expected revenues that can be obtained from state (t, x) where t is the time-to-go and x is the vector of remaining resource inventories. The HJB equation is given by

$$\begin{aligned} \frac{\partial V(t, x)}{\partial t} = & \sum_{l=1}^L \lambda_l \max_{S_l \subset N_l} \sum_{j \in S_l} \pi_{lj}(S_l) \\ & \times (p_{lj} - \Delta_{lj} V(t, x)), \end{aligned} \quad (10)$$

and $\Delta_{lj} V(t, x) = V(t, x) - V(t, x - A_{lj})$ if $A_{lj} \leq x$, and $\Delta_{lj} V(t, x) = \infty$ otherwise where A_{lj} is the j th column of the resource matrix A_l associated with market segment l , $j \in N_l$. Note that in this formulation all requests j in the offered set S_l are accepted as there is no positive part on the term $p_{lj} - \Delta_{lj} V(t, x)$. For this reason, the only candidate ODFs in an optimal set S_l are those such that $A_{lj} \leq x$. Talluri and van Ryzin [39] developed the first stochastic choice-based dependent demand model for the single-resource case in discrete time. The first network formulation in continuous time is due to Gallego *et al.* [40].

As in the independent demand case, the problem of solving the HJB equations (10) is intractable and for this reason, we resort to heuristics. As in the independent demand case, we can use the perfect foresight model to develop an upper bound by solving a linear program. Let $\lambda_l(S_l)$ be the n_l -dimensional vector with components $\lambda_l \pi_{lj}(S_l), j = 1, \dots, n_l$ and $r_l(S_l) = \lambda_l \sum_{j \in S_l} p_{lj} \pi_{lj}(S_l)$ be the revenue rate associated with the offer set S_l . Notice that we can write $r_l(S_l)$ as $p'_l \lambda_l(S_l)$. The decision variables will be for each market segment l the proportion of time, say $\alpha_l(S_l)$, during the sales horizon that we offer set $S_l \in \mathcal{N}_l$ where $\mathcal{N}_l$ is the collection of all subsets of N_l . The DLP will be called the choice-based linear program (CBLP). The first CBLP formulation appeared in Gallego *et al.* [40] who also show that the value function $\bar{V}(T, c)$ of the CBLP is an upper bound on $V(T, c)$. The formulation has $\sum_{l=1}^L 2^{n_l}$ variables $\alpha_l(S_l)$ representing the fraction of time set S_l is offered for market segment $l \in \mathcal{L}$ and is given by

$$\begin{aligned} \bar{V}(T, c) = & \max T \sum_{l=1}^L \sum_{S_l \in \mathcal{N}_l} r_l(S_l) \alpha_l(S_l) \\ \text{subject to } & T \sum_{l=1}^L A_l \sum_{S_l \in \mathcal{N}_l} \lambda_l(S_l) \alpha_l(S_l) \leq c; \\ & \sum_{S_l \in \mathcal{N}_l} \alpha_l(S_l) = 1, \quad l = 1, \dots, L; \\ & \alpha_l(S_l) \geq 0 \quad \forall \quad S_l \in \mathcal{N}_l, \\ & l = 1, \dots, L. \end{aligned} \quad (11)$$

For certain discrete-choice models, including the BAM, it is possible to efficiently use column generation to solve the CBLP to optimality. This line of attack was first used by Gallego *et al.* [40] and later by Liu and van Ryzin [41] and others to solve the CBLP for the case where each market segment can make independent decisions. Bront *et al.* [42] studied the problem with overlapping market segments and show that the column generation problem is NP-hard. Once the CBLP is solved, it is possible to apply a variety of heuristics.

Table 3. Data for Example 4

j	0	1	2	3	4	5
p_{1j}		800	700	650	600	550
a_{1j}	1.95	0.30	0.36	0.39	0.43	0.47
p_{2j}		400	375	360	340	225
a_{2j}	5.56	0.67	0.70	0.71	0.73	0.91
p_{3j}		1,000	900	800	700	600
a_{3j}	28.82	0.20	0.24	0.29	0.35	0.43

Example 4. Consider a network with two flights that connect to serve three markets: market 1 consists of the first flight, market 2 consists of the second flight, and market 3 consists of the two connecting flights with respective capacities 80 and 120. Demand for each of the three markets is driven by an Multinomial Logit (MNL) model. We will assume that the aggregate arrival rates for the three market segments are respectively $\lambda_1 T = 150$, $\lambda_2 T = 300$, $\lambda_3 T = 400$ where we select for convenience a timescale with $T = 1$. Each market has five fares and an outside alternative. In Table 3 we summarize the fares p_{lj} , and the attraction coefficients a_{lj} of the BAM for each market $l = 1, 2, 3$ and each fare $j = 1, \dots, 5$ with a_{l0} representing the attractiveness of the outside alternative for market segment $l = 1, 2, 3$.

Let $S_{lj} = \{lk : k \leq j\}$. Solving CBLP results in $\alpha_1(S_{14}) = 0.946$, $\alpha_1(S_{15}) = 0.054$, $\alpha_2(S_{24}) = 0.754$, $\alpha_2(S_{25}) = 0.246$, and $\alpha_3(S_{34}) = 1$, value function $\bar{V}(T, c) = \$94,064.63$ and dual vector $\bar{z} = (452.38, 152.84)$. Let F_{lk} denote the set of fares that is always offered and let P_{lk} denote the set of fares that are offered only part of the time. Then, $F_l = S_{l4}$ for $l = 1, 2, 3$, while fares $P_1 = \{15\}$, $P_2 = \{25\}$ and $P_3 = \emptyset$.

Heuristics Based on the CBLP

A bid-price heuristic based on the dual variables $\bar{z}$ of the capacity constraint would offer the set of fares $S_l = F_l \cup P_l = \{lj : p_{lj} \geq \bar{z}'A_l\}$ to market segment $l \in \mathcal{L}$. As with the bid-price heuristic for the independent demand model (IDM), this incurs the risk of admitting too many sales of marginal fares $P_l = \{lj : p_{lj} = \bar{z}'A_l\}$ that may cannibalize more profitable sales in the set $F_l = \{lj : p_{lj} > \bar{z}'A_l\}$. For

Example 4, the bid-price heuristic based on $\bar{z} = (452.38, 152.84)$ offers set $S_{15} \times S_{25} \times S_{34}$ while capacity is available, closing fares as capacity is exhausted. The expected revenues for this heuristic, estimated via simulation, are \$90,131.29. Frequent resolution helps mitigate this risk as the sets P_l then to shrink when the CBLP is resolved at reading dates. It is possible to show that a solution to the CBLP would be such that $\alpha_l(S_l) > 0$ only for subsets of the form $S_l = F_l \cup X_l$ with $X_l \subset P_l$. Given a solution $\alpha_l(S_l), l \in \mathcal{L}$ to the CBLP, the counterpart of the PAR heuristic would be to offer set S_l with probability $\alpha_l(S_l)$. The expected revenues for this heuristic, estimated via simulation, are \$90,020.27. Research by Jasin and Kumar [26] shows that applying the PAR heuristic with frequent resolves *bounds the difference* between the performance of the PAR heuristic and the value function $V(T, c)$, so PAR with frequent resolves is an asymptotically optimal heuristic.

Heuristics that Take Randomness into Account

It is possible, but difficult, to incorporate randomness into discrete-choice models. As in the IDM, most work in this area is in the context of the single-resource problem. Finding an effective procedure to solve this problem became a priority in the airline industry as competition from low-cost carriers forced airlines to redesign fares and to reduce fare restrictions. Belobaba and Weatherford [43] propose a variant of the expected marginal seat revenue (EMSR) formula that adjusts the fare ratio to take into account the probability that a customer would buy the next highest fare. While this formula provided a significant improvement over the unadjusted EMSR heuristic, there was plenty of opportunity for improvement as evidenced by refinements in Fiig *et al.* [44]. Brumelle *et al.* [45] develop a conditional probability formula for the two-fare problem with dependent demands. Their formula is in terms of conditional probabilities but no algorithm to actually compute protection levels as provided. Gallego *et al.* [46] provide a recursive procedure to numerically compute the conditional probabilities in Brumelle *et al.* [45] and obtain optimal protection levels. In addition,

they develop a heuristic that extends to the multifare problem. Numerical experiments show that their heuristic is close to optimal and performs significantly better than those developed in Belobaba and Weatherford [43] and Fiig *et al.* [44]. The heuristic provides protection levels for a set of fares against a larger set; for example, protection for fares $S_4 = \{1, \dots, 4\}$ against fares in $S_5 = \{1, \dots, 5\}$ suggests that fare 5 should be shut down when the remaining inventory falls down to or below the protection level. This is a form of *theft* nesting where fare 5 is closed down even after a certain number of bookings under S_5 even if those bookings happen to be for fares other than 5. It is possible to transform the protection levels and booking levels to work under standard nesting; see Talluri and van Ryzin [[31], p. 30], for a discussion of standard and theft nesting. As in the IDM, heuristics that take into account randomness can be applied judiciously to network models by projecting into resources or into itineraries.

Dynamic Bid-Price Heuristics

As with the IDM, it is possible to open the hood and work more closely with the value function to obtain refined heuristics. One of the first such heuristics was proposed by Liu and van Ryzin [41] who use the bid prices from the CBLP to net out fares for each resource. They then solve a netted-fare dynamic program for each resource. Armed with approximate value functions at the resource level, they develop a heuristic admission control rule and show that it is asymptotically optimal. Research by Kunnumkal and Topaloglu [47] goes further by using LR in an attempt to capture time-variant bid prices with good results. Zhang and Adelman [48] use ADP also in an attempt to capture time-variant bid prices. In addition, Zhang [49] uses a nonlinear and nonseparable functional approximation to the problem with good results. Chaneton and Vulcano [50] take a different approach and use simulation-based sample path gradients to construct a stochastic steepest ascent algorithm with impressive computational results. Caveats in terms of data requirements, fare distribution, and scalability are as valid here as in the IDM.

Generalized Attraction Model and Reformulation of CBLP

Most of the research for dependent demand models use a form of the BAM. Gallego *et al.* [51] introduced the generalized attraction model (GAM) where

$$\pi_{lj}(S_l) = \frac{a_{lj}}{a_{l0} + \sum_{k \in S_l} a_{lk} + \sum_{k \notin S_l} b_{lk}} \quad j \in S_{l+}, \quad (12)$$

and $b_{lk} \leq a_{lk}$ represents the attractiveness of purchasing product k from an alternative source. The b_{lk} s are discounted by the inconvenience cost and by the availability probability. The special case $b_{lk} = 0$ recovers the BAM, while the case $b_{lk} = a_{lk}$ recovers the IDM. The parsimonious model $b_{lk} = \theta a_{lk}$ for some $\theta \in [0, 1]$ can help test the hypothesis $H_o : \theta = 0$ or $H_o : \theta = 1$ against the obvious alternatives. Carefully fitting a GAM so that it performs well out of sample can help mitigate the upsell optimism of the BAM and the upsell pessimism of the IDM.

Gallego *et al.* [51] also present a sales-based linear program (SBLP) that is equivalent to the CBLP for discrete-choice models governed by the GAM. The advantage of this model is that it involves only $\sum_{l \in \mathcal{L}} (n_l + 1)$ variables instead of $\sum_{l \in \mathcal{L}} 2^{n_l}$ variables, where $n_l = |\mathcal{N}_l|$ is the cardinality of the consideration set of market segment $l \in \mathcal{L}$. This facilitates the development of more sophisticated heuristics as it now becomes practical to simulate aggregate demands for the market segments; solve the SBLP; and average the value function, the bid-price vector $\bar{z}$, and the admission probability vector γ to refine the upper bound, the bid-price heuristic, and the PAR heuristic.

THE FUTURE OF REVENUE MANAGEMENT

RM is likely to remain an interesting research topic both in terms of the development of new models and new algorithms. There is much work to be done to develop models with dependent demands including the study of correlations across products and across time; see Stefanescu [52]. Competition

has been an area that has seen lots of interest but can benefit from more research; see Gallego and Hu [53], Netessine and Shumsky [54], Perakis and Sood [55], and Xu and Hoop [56]. Forecasting under choice models with censored data is an important topic with ramifications to areas beyond RM; see Talluri [57] and Vulcano *et al.* [58]. Pricing and capacity allocation under demand uncertainty is yet another area that can benefit from future research and have broad impact; see Besbes and Zeevi [59] and Harrison *et al.* [60].

Most of the RM research done till date is at the operational level, with fares and restrictions taken as given. In practice, fares are designed with the purpose of versioning services, by imposing usage or purchasing restrictions, so they can be sold to different market segments at different prices. Effective fares coupled with RM allowed airlines and other industries to significantly improve their revenues with most of the improvements dropping directly to the bottom line. Over the last decade, however, the effectiveness of imposing fare restrictions has been eroded by internet fare transparency and in the case of the airline industry, by the presence of low-cost carriers.

At a more strategic level, we see two important challenges for RM. One is to find effective market segmentation tools. The other is to broaden RM to other industries. One could tinker with the idea of product versioning with the goal of improving ways of crimping services. However, versioning is but one tool to design and price derivative services to appeal to broader market segments and to improve resource utilization. As in financial engineering, it is possible to use options and bundling/unbundling ideas to modify a core service into derivative services; see Gallego and Stefanescu [61] and Gallego and Sahin [62]. These services help customers to manage consumption risk and capacity providers to broker flexibility among customers. Derivative services can be designed and priced to segment the market and benefit companies and their customers, but also to induce new demand from market segments that otherwise would not be interested in the company's offer. In

short, derivative services can help to improve demand management [63]. The engineering of derivative services is a strategic tool that can lead to significant increases in profits and market share in many industries including travel, leisure, hotel/hospitality, internet advertising, technology, transportation, supply chain, financial products and services, and club membership, among others.

Acknowledgements

I am grateful to Georgia Perakis for inviting me to write the article and to the editors for constructive feedbacks. I am also grateful to Lin Li, Richard Ratliff, Catalina Stefanescu, and Ozge Sahin for their helpful comments.

REFERENCES

1. Talluri K, van Ryzin G. An analysis of bid-price controls for network revenue management. *Manage Sci* 1998;44(11):1577–1593.
2. Phillips R. Pricing and revenue optimization. Stanford (CA): Stanford Business Books; 2005.
3. McGill JI, van Ryzin G. Revenue management: research overview and prospects. *Transp Sci* 1999;33:233–256.
4. Bitran G, Caldentey R. An overview of pricing models and revenue management. *Manuf Serv Oper Manage* 2003;5:203–229.
5. Elmaghraby WJ, Keskinocak P. Dynamic pricing in the presence of inventory considerations. *Manage Sci* 2003;49:1287–1309.
6. Brumelle SL, McGill JI. Airline seat allocation with multiple nested fare classes. *Oper Res* 1993;41:127–137.
7. Wollmer RD. An airline seat management model for a single leg route when lower fare classes book first. *Oper Res* 1992;40:26–37.
8. Robinson LW. Optimal and approximate control policies for airline booking with sequential nonmonotonic fare classes. *Oper Res* 1995;43:252–263.
9. van Ryzin G, McGill JI. Revenue management without forecasting or optimization: an adaptive algorithm for determining seat protection levels. *Manage Sci* 2000;46:760–775.
10. Littlewood K. Forecasting and control of passengers bookings. *Proceedings XII AGIFORS Symposium*. Nathanya, Israel; 1972.

11. Belobaba PP. Airline yield management an overview of seat inventory control. *Transp Sci* 1987;21:63–73.
12. Belobaba PP. Application of a probabilistic decision model to airline seat inventory control. *Oper Res* 1989;37:183–197.
13. Lee TC, Hersh M. A model for dynamic airline seat inventory control with multiple seat bookings. *Transp Sci* 1993;27:252–265.
14. Feng Y, Gallego G. Optimal starting times of end-of-season sales and optimal stopping times for promotional fares. *Manage Sci* 1995;41:1371–1391.
15. Feng Y, Gallego G. Perishable asset revenue management with Markovian time-dependent demand intensities. *Manage Sci* 2000;46:941–956.
16. Feng Y, Xiao B. Optimal policies of yield management with multiple predetermined prices. *Oper Res* 2000;46:332–343.
17. Weng Z, Zheng Y-S. A dynamic model for airline seat allocation with passenger diversion and no-shows. *Transp Sci* 2001;35:80–98.
18. Bitran GR, Mondschein SV. Periodic pricing of seasonal products in retailing. *Manage Sci* 1997;43:64–79.
19. Gallego G, van Ryzin G. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Manage Sci* 1994;40:999–1020.
20. Simpson RW. Using network flow techniques to find shadow prices for market and seat inventory control. Memorandum M89-1, Cambridge (MA): Flight Transportation Laboratory, Massachusetts Institute of Technology; 1989.
21. Chen VCP, G  nther D, Johnson EL. Airline yield management: optimal bid prices for single-hub problems without cancellations. Working paper. Atlanta, GA: The Logistics Institute, Georgia Institute of Technology; 1999.
22. Gallego G. Lecture notes on revenue management. New York (NY): Columbia University; 2007.
23. Reiman M, Wang Q. An asymptotically optimal policy for a quantity-based network revenue management problem. Working paper. Alcatel-Lucent Bell Labs; 2007.
24. Maglaras C, Meissner J. Dynamic pricing strategies for multiproduct revenue management problems. *Manuf Serv Oper Manage* 2006;8(2):136–148.
25. Cooper W. Asymptotic behavior of an allocation policy for revenue management. *Oper Res* 2002;50:720–727.
26. Jasin S, Kumar S. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. Working paper. Stanford University; 2010.
27. Talluri K, van Ryzin G. A randomized linear programming method for computing network bid prices. *Transp Sci* 1999;33:207–216.
28. Topaloglu H. On the asymptotic optimality of the randomized linear program for network revenue management. Technical Report. Ithaca (NY): School of IEOR, Cornell University; 2007.
29. de Boer SV, Freling R, Piersma N. Mathematical programming for network revenue management revisited. *Eur J Oper Res* 2002;137:72–92.
30. Williamson EL. Airline network seat inventory control: methodologies and revenue impacts [PhD thesis]. Cambridge (MA): Flight Transportation Laboratory, Massachusetts Institute of Technology; 1992.
31. Talluri K, van Ryzin G. The theory and practice of revenue management. Boston/Dordrecht/London: Kluwer Academic Publishers; 2004.
32. Smith BC, Penn CW. Analysis of alternative origin-destination control strategies. Volume 28, AGIFORS Symposium Proceedings; New Seabury (MA). 1988.
33. Adelman D. Dynamic bid-prices in revenue management. *Oper Res* 2007;55:647–661.
34. Bertsimas D, Popescu I. Revenue management in a dynamic network environment. *Transp Sci* 2003;37:257–277.
35. Topaloglu H. Using Lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Oper Res* 2009;57:637–649.
36. Talluri K. On bounds for network revenue management. Technical Report WP-1066. UPF; 2008.
37. Bertsimas D, de Boer S. Simulation-based booking limits for airline revenue management. *Oper Res* 2005;53:90–106.
38. Ben-Akiva M, Lerman S. Discrete choice analysis: theory and applications to travel demand. 6th ed. Cambridge (MA): The MIT Press; 1994.
39. Talluri K, van Ryzin G. Revenue management under a general discrete choice model of consumer behavior. *Manage Sci* 2004;50:15–33.

40. Gallego G, Iyengar G, Phillips R. Managing flexible products on a network. CORC Technical Report TR-2004-01. New York: Department of Industrial Engineering and Operations Research, Columbia University; 2004.
41. Liu Q, van Ryzin G. On the choice-based linear programming model for network revenue management. *Manuf Serv Oper Manage* 2008;10:288–310.
42. Bront J, Mendez-Diaz I, Vulcano G. A column generation algorithm for choice-based network revenue management. *Oper Res* 2009; 57:769–784.
43. Belobaba PP, Weatherford LR. Comparing decision rules that incorporate customer diversion in perishable asset revenue management situations. *Decis Sci J* 1996;27:343–363.
44. Fiig T, Isler K, Hopperstad C, *et al.* Davn-MR: a unified theory of OD optimization in a mixed network with restricted and unrestricted fare products. AGIFORS Revenue Management and Distribution Study Group Meeting; Cape Town, South Africa. 2005.
45. Brumelle SL, McGill JI, Oum TH, *et al.* Allocation of airline seats between stochastically dependent demands. *Transp Sci* 1990; 24:183–192.
46. Gallego G, Lin L, Ratliff R. Choice based EMSR methods for single-resource revenue management with demand dependencies. *J Revenue Pricing Manage* 2009;8:207–240.
47. Kunnumkal S, Topaloglu H. A refined deterministic linear program for the network revenue management problem with customer choice behavior. *Naval Res Logist* 2008;55(6):563–580.
48. Zhang D, Adelman D. An approximate dynamic programming approach to network revenue management with customer choice. *Transp Sci* 2009;43:381–394.
49. Zhang D. An improved dynamic programming decomposition approach for network revenue management. *Manuf Serv Oper Manage* 2010. In press.
50. Chaneton J, Vulcano G. Computing bid-prices for revenue management under customer choice behavior. Working paper. New York University; 2009.
51. Gallego G, Ratliff R, Shebalov S. An efficient formulation for the choice-based linear program under a general attraction model. Working paper. Columbia University; 2010.
52. Stefanescu C. Multivariate customer demand: modeling and estimation from censored sales. Working paper. LBS; 2009.
53. Gallego G, Hu M. Dynamic pricing of perishable assets under competition. Working paper. University of Toronto; 2007.
54. Netessine S, Shumsky RA. Revenue management games: horizontal and vertical competition. *Manage Sci* 2005;51:813–831.
55. Perakis G, Sood A. Competitive multiperiod pricing for perishable products: a robust optimization approach. *Math Program* 2006; 107:295–335.
56. Xu X, Hopp WJ. A monopolistic and oligopolistic stochastic flow revenue management model. *Oper Res* 2006;54:1098–1109.
57. Talluri K. A finite-population revenue management model and a risk-ratio procedure for the joint estimation of population size and parameters. Technical Report WP-1141. UPF; 2010.
58. Vulcano G, van Ryzin G, Ratliff R. Estimating primary demand for substitutable products from sales transaction data. Working paper. New York University; 2008.
59. Besbes O, Zeevi A. Price dynamics under limited information. Working paper. Columbia University; 2009.
60. Harrison JM, Keskin BN, Zeevi A. Bayesian dynamic pricing policies: learning and earning under a binary prior distribution. Working paper. Columbia University; 2009.
61. Gallego G, Stefanescu C. Service engineering: design and pricing of service features. In: Ozer O., Phillips R., editors. *Handbook of pricing management*. Oxford University Press; 2010. In press.
62. Gallego G, Sahin O. Revenue management with partially refundable fares. *Oper Res* 2010;58:817–833.
63. Anderson CK, Carroll B. Demand management: beyond revenue management. *J Pricing Revenue Manage* 2007;6:260–263.

REVERSIBILITY IN QUEUEING MODELS

MASAKIYO MIYAZAWA
Tokyo University of Science,
Noda, Japan

INTRODUCTION

Stochastic models for queues and their networks are usually described by stochastic processes, in which random events evolve in time. Here, time is continuous or discrete. We can view such evolution in reverse for each stochastic model. A process for this backward evolution is referred to as a *time-reversed process*. It also is called a reversed time process or simply a reversed process. We refer to the original process as a *time-forward process* when it is required to distinguish it from the time-reversed process.

It is often greatly helpful to view a stochastic model from two different time directions. In particular, if some property is invariant under change of the time directions, that is, under time reversal, then we may better understand that property. A *concept of reversibility* is invented for this invariance.

A comprehensive discussion on reversibility in queueing networks and related models is presented in the book by Kelly [1]. The reversibility used is in the sense that the time-reversed process is stochastically identical to the time-forward process, but there are many discussions about stochastic models with time reversal having the same modeling features. In this article, we give a unified view to them, using a broader concept of reversibility. We first consider some examples.

1. A *stationary Markov process* (continuous- or discrete-time) is again a stationary Markov process under time reversal. This is detailed in the section titled “Markov Chain and Its Time Reversal.” In this case, the property

that transition probabilities from a given state are independent of past history and the present time, called the time homogeneous *Markov property*, is reversible for a stationary Markov processes.

2. A *random walk* is a discrete-time process with independent increments, and again a random walk under time reversal; that is, the time-reversed process also has independent increments. Hence, the independent increment property is reversible for a random walk. A random walk is a special case of a discrete-time Markov process, but it does not have a stationary distribution.
3. A *queueing model* is a system to transform an arrival stream to a departure stream. If the queueing model can be reversed in time, the roles of arrival and departure streams are exchanged, but the input–output structure may be considered to be unchanged. Thus, the stream structure is reversible for the queueing model. It is notable that a stochastic process for this model does not need to have a stationary distribution. We propose a self-reacting system for this type of models in the section titled “Self-Reacting System and Balanced Departure.”
4. A *birth-and-death process* is a continuous-time Markov process taking nonnegative integers that go up or down by 1. This process is typically used to represent population dynamics of living things. It is also used to model a queueing system called a *Markovian queue* (see section titled “Queue and Reacting System”).

These examples may be too general or too specific to be of some use, but they suggest that there may be different levels of reversibility depending on characteristics of interest and on the class of stochastic processes (or models). In particular, (4) is a

special case of (1) if it has a stationary distribution. As we will see later, this Markov chain with a stationary distribution is reversible in time, that is, the time-reversed Markov chain is stochastically identical to the time-forward process. Thus, (1) and (4) are two extremes concerning the time-reversed processes that have stationary distributions.

How can we define reversibility of a stochastic process for a queueing model? Obviously, we need a time-reversed dynamics. How can we construct it? Is the stationary distribution required for this? The answer is “no” because of examples (2) and (3). To make the problems concrete, we consider discrete- and continuous-time Markov processes with countable state spaces, which are referred to as *discrete- and continuous-time Markov chains*, respectively, because they are widely used in queueing and their network applications.

When we are interested in the stationary distributions of the Markov chains, we have to verify or assume their existence. In this case, there should be a useful notion of reversibility between extreme cases such as (1) and (4). However, structural properties such as arrivals and departures may not depend on whether or not underlying Markov chains for them have stationary distributions. We aim to devise such reversibility, which is discussed in the section titled “Local Balance and Reversibility in Structure” under *reversibility in structure*. We show how this concept arises in queueing models and their networks and how it is useful to obtain the stationary distributions in a unified way in the subsequent sections up to the section titled “Queueing Network and Product Form.” Reversibility in structure is discussed for other types of Markov processes in the section titled “Other Stochastic Processes” and some remarks for further study are given in the section titled “Concluding Remarks.”

Before we start these discussions, we recall basic facts about Markov chains and their time reversals in the section titled “Markov Chain and Its Time Reversal.” The reader who is familiar with those topics may skip this section. However, it may be noted that we define a transition rate function in a

slightly extended manner. Thus, we recommend that the reader check this definition.

MARKOV CHAIN AND ITS TIME REVERSAL

In this section, we construct time-reversed processes for discrete- and continuous-time Markov chains. We first consider the discrete-time case because it is simpler than the continuous-time case, however, the arguments are similar. A discrete-time stochastic process $\{X_n; n = 0, 1, \dots\}$ with S is called a Markov chain if, for any $n \in \mathbb{Z}_+ \equiv \{0, 1, 2, \dots\}$ and any $i, j, i_0, \dots, i_{n-1} \in S$,

$$\begin{aligned} \mathbb{P}(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_{n-1} \\ = i_{n-1}, X_n = i) = \mathbb{P}(X_{n+1} = j | X_n = i). \end{aligned}$$

If the right-hand side is independent of n , then it is said to have a *time-homogeneous transition*, which is denoted by p_{ij} . That is,

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i), \quad i, j \in S.$$

This p_{ij} is called a *transition probability* from state i to state j . Throughout this article, we always assume Markov chains to be time homogeneous. To exclude trivial exceptions, we further assume that there is a path from i to j for each $i, j \in S$, that is, there are states $i_1, i_2, \dots, i_{n-1} \in S$ for some $n \geq 1$ such that

$$p_{ii_1} p_{i_1 i_2} \times \dots \times p_{i_{n-2} i_{n-1}} p_{i_{n-1} j} > 0. \quad (1)$$

This condition means that any state is commutative with any other state; this is known as *irreducibility*.

If a probability distribution $\pi \equiv \{\pi(i)\}$ on S , that is, $\pi(i) \geq 0$ for all i such that $\sum_{i \in S} \pi(i) = 1$, satisfies

$$\pi(j) = \sum_{i \in S} \pi(i) p_{ij}, \quad j \in S, \quad (2)$$

then π is called a *stationary distribution* of the Markov chain $\{X_n\}$. Note that, $\pi(j) > 0$ for all $j \in S$ by the irreducibility. If the Markov chain $\{X_n\}$ has a stationary distribution π and if X_0 is subject to this π , then X_n has the

distribution π for each $n \geq 1$. Define the time-shifted process $\{X_n^{(-\ell)}; n = -\ell, -(\ell-1), \dots\}$ by

$$X_n^{(-\ell)} = X_{n+\ell}, \quad n = -\ell, -(\ell-1), \dots$$

Then, $\{X_n^{(-\ell)}; n \geq m\}$ for each $m \geq -\ell$ has the same joint distribution as $\{X_n; n \geq 0\}$ as long as X_0 has the stationary distribution π , and therefore, by letting $\ell \rightarrow \infty$, the Markov chain $\{X_n^{(-\infty)}; n \in \mathbb{Z}\}$ is well defined, where $\mathbb{Z}$ is the set of all integers. Thus, we can consider the Markov chain $\{X_n\}$ to be a process starting from $-\infty$. This allows us to define a time-reversed process $\{\tilde{X}_n\}$ by

$$\tilde{X}_n = X_{-n}, \quad n \in \mathbb{Z}. \quad (3)$$

Since $\mathbb{P}(X_{-(n+1)} = j) = \pi(j)$, we have

$$\begin{aligned} \mathbb{P}(\tilde{X}_0 = i_0, \tilde{X}_1 = i_1, \dots, \tilde{X}_n = i, \tilde{X}_{n+1} = j) \\ &= \mathbb{P}(X_0 = i_0, X_{-1} = i_1, \dots, X_{-n} = i, X_{-(n+1)} = j) \\ &= \mathbb{P}(X_{-(n+1)} = j) \mathbb{P}(X_0 = i_0, X_{-1} = i_1, \dots, \\ &\quad X_{-(n-1)} = i | X_{-(n+1)} = j) \\ &= \pi(j) p_{ji} p_{ii_{n-1}} \times \dots \times p_{i_1 i_0}, \end{aligned}$$

and therefore

$$\begin{aligned} \mathbb{P}(\tilde{X}_{n+1} = j | \tilde{X}_0 = i_0, \tilde{X}_1 = i_1, \dots, \tilde{X}_n = i) \\ &= \frac{\mathbb{P}(\tilde{X}_0 = i_0, \tilde{X}_1 = i_1, \dots, \tilde{X}_n = i, \tilde{X}_{n+1} = j)}{\mathbb{P}(\tilde{X}_0 = i_0, \tilde{X}_1 = i_1, \dots, \tilde{X}_n = i)} \\ &= \frac{\pi(j) p_{ji} p_{ii_{n-1}} \times \dots \times p_{i_1 i_0}}{\pi(i) p_{ii_{n-1}} \times \dots \times p_{i_1 i_0}} \\ &= \frac{\pi(j) p_{ji}}{\pi(i)} = \mathbb{P}(\tilde{X}_{n+1} = j | \tilde{X}_n = i). \end{aligned}$$

Hence, the time-reversed process $\{\tilde{X}_n\}$ is also a Markov chain with transition probability $\tilde{p}_{ij}$ given by

$$\tilde{p}_{ij} = \frac{\pi(j)}{\pi(i)} p_{ji}, \quad i, j \in S. \quad (4)$$

We have defined the time-reversed process $\{\tilde{X}_n\}$ by Equation 3, but we also can define it as a Markov chain with transition probabilities $\{\tilde{p}_{ij}\}$ of Equation 4. Furthermore, for the latter definition, it is not necessary for $\{\pi(i)\}$ to be a probability distribution, but $0 <$

$\pi(i) < \infty$ for all $i \in S$ satisfying Equation 2 is sufficient. This $\{\pi(i)\}$ is called a *stationary measure*. This also explains why a random walk can be reversed in time because $\pi(i) = a$ for any constant $a > 0$ is its stationary measure. We refer to the measure π to define $\tilde{p}$ of Equation 4 as a reference measure (see Definition 2 and Section 3 of [2] for related topics).

It is notable that we can use a reference measure other than a stationary measure. For example, we can consider a time-reversed process $\{X_{n_0-n}; 0 \leq n \leq n_0\}$ starting at $X_{n_0} = i_0$ for fixed time $n_0 > 0$ and state $i_0 \in S$. In this case, the reference measure is concentrated on i_0 . It can be shown that this process is again a Markov chain, but does not have time-homogeneous transitions. Thus, it is different from $\{\tilde{X}_n\}$. As a reference measure, a stationary distribution (or measure) has the excellent feature that the transition probabilities have a simple form.

Thus, the choice of a reference measure is crucial for the time-reversed process. Throughout this article, we take a stationary distribution (a stationary measure in a few instances) as a reference measure for a time-reversed process. However, this does not mean that we also take it for reversibility in structure, which will be considered in the section titled ‘‘Local Balance and Reversibility in Structure,’’ because the structural property should be independent of the existence of the stationary distribution.

We next consider a continuous-time Markov chain, for which irreducibility is assumed. This is quite easy because most arguments are parallel to those of a discrete-time Markov chain.

A stochastic process $\{X(t)\}$ with state space S is called a *Markov chain* if, for each $n \in \mathbb{Z}_+$, any $s > 0$, any $0 < t_1 < \dots < t_n < s$ and any $i, j, i_0, \dots, i_n \in S$,

$$\begin{aligned} \mathbb{P}(X(s+t) = j | X(0) = i_0, X(t_1) = i_1, \dots, X(t_n) \\ &= i_n, X(s) = i) \\ &= \mathbb{P}(X(s+t) = j | X(s) = i). \end{aligned}$$

We denote the right-hand side of this equation by $p_{ij}(t)$ if it is independent of s , which we always assume. We also assume

that, for each $i, j \in S$, there exists $t > 0$ such that $p_{ij}(t) > 0$, which is called *irreducibility*.

Then, it is known that there exist $a(i) \geq 0$ and $q(i, j) \geq 0$ such that

$$\lim_{h \downarrow 0} \frac{1}{h} p_{ij}(h) = q(i, j), \quad i \neq j, i, j \in S,$$

$$\lim_{h \downarrow 0} \frac{1}{h} (p_{ii}(h) - 1) = -a(i), \quad i \in S.$$

In words, $a(i)$ is a rate going out from state i , and $q(i, j)$ is a transition rate from i to j under the condition that the transition occurs. Throughout this article, we assume that $a(i)$ is finite for all $i \in S$. It is not hard to see that

$$a(i) = \sum_{j \neq i, j \in S} q(i, j), \quad i \in S,$$

and the Markov chain $\{X(t)\}$ is determined by $\{q(i, j); i, j \in S\}$. For $X(t)$ to be well defined for all $t \geq 0$, a certain regularity condition is required (e.g., see Section 2.2 of [3]). However, it is always satisfied in queueing applications.

This continuous-time Markov chain is slightly extended by adding dummy transitions by letting $q(i, i) \geq 0$. In this case, we redefine $a(i)$ as

$$a(i) = \sum_{j \in S} q(i, j), \quad i \in S.$$

In standard text books (e.g., Ref. 3), $q(i, i)$ is usually defined as $q(i, i) = -a(i)$. Thus, it is notable that our $q(i, i)$ is different from these. As a stochastic process, $q(i, i) > 0$ is not meaningful because it does not change the state. However, it is important for describing a queueing model because it enables to correctly count arriving customers who cannot enter a system. We refer to q as a *transition rate function*.

Since the Markov chain is irreducible, its stationary distribution π is uniquely determined by

$$a(j)\pi(j) = \sum_{i \in S} \pi(i)q(i, j), \quad j \in S. \quad (5)$$

This equation is slightly different from Equation 2 of the discrete-time Markov chain.

This is because the transition rate function $\{q(i, j); j \in S\}$ is not a probability distribution for the given i . Assume that the Markov chain has the stationary distribution π , then all arguments for the discrete-time Markov chain are valid for the continuous-time Markov chain if we replace p_{ij} by $q(i, j)$ and make small modifications as in Equation 5. Denote the time-reversed Markov chain by $\{\tilde{X}(t)\}$, then, similar to Equation 4, its transition rate function $\tilde{q}$ is given by

$$\tilde{q}(i, j) = \frac{\pi(j)}{\pi(i)} q(j, i), \quad i, j \in S. \quad (6)$$

From Equations 6 and 5, we immediately have the next fact.

Lemma 1. A probability distribution π on S is the stationary distribution of the Markov chain with transition rate function q if and only if there exists a nonnegative function $\tilde{q}$ on S such that

$$\pi(i)\tilde{q}(i, j) = \pi(j)q(j, i), \quad i, j \in S, \quad (7)$$

$$\sum_{j \in S} \tilde{q}(i, j) = \sum_{j \in S} q(i, j), \quad i \in S. \quad (8)$$

A message from this lemma is that the stationary distribution can be found if we can guess $\tilde{q}$ from the time-reversed model. This idea is extensively used in the book [1], and Lemma 1 is called *Kelly's lemma* (e.g., see Ref. 4).

LOCAL BALANCE AND REVERSIBILITY IN STRUCTURE

From now on, we only consider the continuous-time Markov chain $\{X(t)\}$ with state space S and transition probabilities $\{X(t)\}$ because the corresponding results for a discrete-time Markov chain are similarly obtained. We also change the notation for state from $i, j \in S$ to $x, x' \in S$ for convenience of our discussions.

We aim to define reversibility in structure for this Markov chain. For this, we first consider the strongest case that the Markov chain $\{X(t)\}$ itself is reversible. Assume that

it has the stationary measure π , and define the time-reversed Markov chain $\{\tilde{X}(t)\}$ with transition rate functions $\tilde{q}$. Then, $\{X(t)\}$ is stochastically identical to its time-reversed process $\{\tilde{X}(t)\}$ if and only if

$$q(x, x') = \tilde{q}(x', x), \quad x, x' \in S.$$

From Equation 6, this condition is equivalent to

$$\pi(x)q(x, x') = \pi(x')q(x', x), \quad x, x' \in S. \quad (9)$$

This Markov chain is said to be *reversible*. It is notable that Equation 9 yields, for fixed x and the sequence of the states $x, x_1, \dots, x_{n-1}, x' \in S$,

$$\pi(x') = \pi(x) \frac{q(x, x_1)q(x_1, x_2) \times \dots \times q(x_{n-1}, x')}{q(x', x_{n-1}) \times \dots \times q(x_2, x_1)q(x_1, x')}.$$

as long as the denominator in the right-hand side is positive. Hence, we have the stationary distribution π if $\sum_{x \in S} \pi(x) < \infty$. Otherwise, this π is a stationary measure. Thus, the reversibility enables us to obtain the stationary distribution.

A typical example of a reversible Markov chain is a *birth-and-death process*, which is defined in (4) of the section titled “Introduction.” Recall that its state space $S = \mathbb{Z}_+$, and the transition probability $q(x, x') > 0$ for $x, x' \in S$ if and only if $x' = x + 1, x' = x - 1 \geq 0$ or $x = x' = 0$. For this Markov chain, let

$$\begin{aligned} \pi(x) &= \pi(0) \frac{q(0, 1)}{q(x, x-1)} \times \frac{q(1, 2)}{q(x-1, x-2)} \\ &\quad \times \dots \times \frac{q(x-1, x)}{q(1, 0)}, \quad x \in S, \end{aligned} \quad (10)$$

then Equation 9 is satisfied. Hence, the birth-and-death process is reversible.

However, a Markov chain for a queueing model may not be reversible. In particular, this reversibility is too strong for queueing networks. Motivated by this, the following notion has been studied, which is weaker than reversibility.

Definition 1. [Local balance]. Assume that the Markov chain $\{X(t)\}$ with transition rate function q has a stationary distribution

π , and let $\tilde{q}$ be the transition rate function of its time-reversed process. Let W be a set of disjoint subsets of $S \times S$. If the Markov chain satisfies

$$\sum_{(x, x') \in C} \pi(x)q(x, x') = \sum_{(x, x') \in C} \pi(x)\tilde{q}(x, x'), \quad C \in W, \quad (11)$$

or equivalently,

$$\sum_{(x, x') \in C} \pi(x)q(x, x') = \sum_{(x, x') \in C} \pi(x')q(x', x), \quad C \in W, \quad (12)$$

then transition rate function q is said to be *locally balanced* with respect to W .

We can weaken the local balance by using the so-called test functions. Let $\mathcal{M}_+(S \times S)$ be the set of all nonnegative-valued functions on $S \times S$. For a subset $\mathcal{G}$ of $\mathcal{M}_+(S \times S)$, q is said to be *locally balanced* in test functions of $\mathcal{G}$ if

$$\begin{aligned} &\sum_{x, x' \in S} \pi(x)q(x, x')f(x, x') \\ &= \sum_{x, x' \in S} \pi(x')q(x', x)f(x, x'), \quad f \in \mathcal{G}. \end{aligned} \quad (13)$$

Let $\mathcal{G}$ be the set of indicator functions $1((x, x') \in C)$ for $C \in W$ of definition 1, where $1(\cdot)$ is the function of the statement “.” such that it equals 1 if the statement is true and 0 otherwise, then Equation 13 is equivalent to Equation 12. Hence, Equation 13 certainly generalizes Equation 12.

However, the local balance in test functions still has limitations. A problem is that transition rate functions q and $\tilde{q}$ are directly related through balance equations. This limits flexibility of $\tilde{q}$ to be chosen. Furthermore, it also requires that q has a stationary distribution. Instead of balance equations and stationary distribution, we use a set of pairs of transition rate functions q and measure σ on S to support q for a weaker sense of reversibility, where σ is said to be a measure to support q or a *supporting measure* of q if $\sigma(x) \geq 0$ for all $x \in S$, and, for each $x \in S$, $\sigma(x) > 0$ if $q(x', x) > 0$ for some $x' \in S$.

We denote the sets of all measures on S and of all supporting measures of q by $M(S)$ and $M_{\text{sup}}(q)$, respectively.

For $\sigma \in M(S)$, let $\mathcal{T}(\sigma)$ be the set of all transition rate functions on $S \times S$ which are supported by σ , and let

$$\mathcal{T} = \cup_{\sigma \in M(S)} \{(q, \sigma); q \in \mathcal{T}(\sigma)\}.$$

That is, $\mathcal{T}$ is the set of all pairs (q, σ) of transition rate function q and measure $\sigma \in M_{\text{sup}}(q)$. Note that $\sum_{x' \in S} q(x, x') < \infty$ for all $x \in S$ is always assumed, but σ is not necessarily the stationary distribution of q . Furthermore, a Markov chain with transition rate function q may not be irreducible.

Definition 2. [Reversible in structure]. A subset $\mathcal{Q}$ of $\mathcal{T}$ is said to be reversible in structure if $(\tilde{q}, \sigma) \in \mathcal{Q}$ for each $(q, \sigma) \in \mathcal{Q}$, where

$$\tilde{q}(x, x') = \begin{cases} \frac{\sigma(x')}{\sigma(x)} q(x', x), & \sigma(x) > 0, x, x' \in S, \\ 0 & \sigma(x) = 0, x, x' \in S. \end{cases}$$

This σ is called a *reference measure* to get $\tilde{q}$ from q . In particular, a Markov chain with transition rate function q is said to be *reversible in structure* with respect to $\mathcal{Q}$ if there is a $\sigma \in M_{\text{sup}}(q)$ such that $(q, \sigma) \in \mathcal{Q}$ implies $(\tilde{q}, \sigma) \in \mathcal{Q}$.

Remark 1. (a) For $\mathcal{Q} \subset \mathcal{T}$ and $\sigma \in M(S)$, let $\mathcal{Q}(\sigma) = \{q \in \mathcal{T}(\sigma); (q, \sigma) \in \mathcal{Q}\}$, then we can write the condition $(q, \sigma) \in \mathcal{Q}$ as $q \in \mathcal{Q}(\sigma)$. The latter is convenient when σ is specified. This is often the case when σ is the stationary distribution π .

(b) For the stationary distribution π of q , $q \in \mathcal{T}(\pi)$, and π is also the stationary distribution of $\tilde{q}$. However, the reversibility in structure does not require σ to be a stationary distribution. In fact, the notion of reversibility in structure is independent of its existence.

Note that $\tilde{\tilde{q}} = q$ for $(q, \sigma) \in \mathcal{Q}$. Hence, q is reversible in structure concerning $\mathcal{Q}$ if and only if $\tilde{q}$ is also so. The structure reversibility obviously includes the local balance in test functions with respect to $\mathcal{G}$. To see this, we

only need to define $\mathcal{Q}$ as the set of (q, π) such that Equation 13 holds for all $g \in \mathcal{G}$ for the stationary distribution π of q . However, $\mathcal{Q}$ of Definition 2 may be still too strong for queueing applications.

Because we are interested in queueing models, we have arrivals and departures to a system, and their roles are exchanged under time reversal. This suggests partially decomposing q into a family of transition rate functions q_u for $u \in U$, where U is a finite set, in such a way that

$$\sum_{u \in U} q_u(x, x') \leq q(x, x'), \quad x, x' \in S, \quad (14)$$

and to define a one-to-one mapping Γ from U to itself such that $\Gamma(\Gamma(u)) = u$. This Γ is said to be a *reflective permutation* on U . We refer to $\{q_u; u \in U\}$ satisfying Equation 14 as a subtransition family of q .

For example, if u represents arrivals, that is, q_u is the transition rate function for arrivals, then $\Gamma(u)$ may be considered to represent departures. Since arrivals are changed to departures under time reversal, we define the transition rate function $\tilde{q}_{\Gamma(u)}$ for $u \in U$ and supporting measure $\sigma \in M_{\text{sup}}(q)$ by

$$\tilde{q}_{\Gamma(u)}(x, x') = \begin{cases} \frac{\sigma(x')}{\sigma(x)} q_u(x', x), & \sigma(x) > 0, \\ & x, x' \in S, \\ 0 & \sigma(x) = 0, x, x' \in S, \end{cases} \quad (15)$$

that is, $\tilde{q}_{\Gamma(u)}$ is obtained from q_u by the reference measure σ . We may consider $\tilde{q}_{\Gamma(u)}$ to be the transition rate function for departures under time reversal with respect to σ . Note that Equation 15 can be used to define $\tilde{q}_u$ if we use $\Gamma^{-1}(u)$ instead of u . These facts motivate us to define the following reversibility in structure.

Definition 3. [Γ -reversible in structure]. For a finite set U and a reflective permutation Γ on U , let $\mathcal{Q}_u$ be a subset of $\mathcal{T}$ for $u \in U$, then $\{\mathcal{Q}_u; u \in U\}$ is said to be Γ -reversible in structure if, for each $\sigma \in M(S)$,

$$q_u \in \mathcal{Q}_u(\sigma) \text{ implies } \tilde{q}_{\Gamma(u)} \in \mathcal{Q}_{\Gamma(u)}(\sigma) \text{ for all } u \in U, \quad (16)$$

where $\mathcal{Q}_u(\sigma) = \{q \in \mathcal{T}(\sigma); (q, \sigma) \in \mathcal{Q}_u\}$ and $\tilde{q}_{\Gamma(u)}$ is defined by Equation 15. In particular, a Markov chain which has transition rate function q and sub-transition family $\{q_u; u \in U\}$ is said to be Γ -reversible in structure with respect to $\{\mathcal{Q}_u; u \in U\}$ if there is a $\sigma \in M_{\text{sup}}(q)$ for which Equation 16 holds. Furthermore, if $\sigma \in M_{\text{sup}}(q)$ is specified, then the Markov chain is said to be Γ -reversible in structure with respect to $\{\mathcal{Q}_u(\sigma); u \in U\}$ if Equation 16 holds.

Remark 2. The Γ -reversibility in structure of $\{\mathcal{Q}_u; u \in U\}$ is a class property for a subset of $\mathcal{T}$, while the Γ -reversibility in structure of a Markov chain with q and $\{q_u; u \in U\}$ is defined for a Markov chain. For the latter, $\{\mathcal{Q}_u; u \in U\}$ itself is not necessarily Γ -reversible in structure. However, the Γ -reversibility in structure of $\{\mathcal{Q}_u; u \in U\}$ is convenient to construct a class of models satisfying such reversibility.

We restate the definition of Γ -reversibility in structure in a convenient form for applications.

Lemma 2. Let Γ be a reflective permutation on a finite set U . Then, $\{\mathcal{Q}_u; u \in U\}$ is Γ -reversible in structure if and only if, for each $\sigma \in M(S)$,

$$\mathcal{Q}_{\Gamma(u)}(\sigma) = \{q_{\Gamma(u)} \in \mathcal{T}(\sigma); \tilde{q}_u \in \mathcal{Q}_u(\sigma)\}, u \in U. \quad (17)$$

Proof. Assume that $\{\mathcal{Q}_u; u \in U\}$ is Γ -reversible in structure. Then, for each $\sigma \in M(S)$, $q_{\Gamma(u)} \in \mathcal{Q}_{\Gamma(u)}(\sigma)$ implies $\tilde{q}_{\Gamma^2(u)} \in \mathcal{Q}_{\Gamma^2(u)}(\sigma)$, which is identical to $\tilde{q}_u \in \mathcal{Q}_u(\sigma)$ because $\Gamma^2(u) = u$. Similarly, $\tilde{q}_u \in \mathcal{Q}_u(\sigma)$ implies $\tilde{q}_{\Gamma(u)} \in \mathcal{Q}_{\Gamma(u)}(\sigma)$, which is identical to $q_{\Gamma(u)} \in \mathcal{Q}_{\Gamma(u)}(\sigma)$. Hence, we have Equation 17. Conversely, Equation 17 and $\sigma \in M(S)$ imply that $\tilde{q}_u \in \mathcal{Q}_u(\sigma)$ leads $\tilde{q}_{\Gamma(u)} \in \mathcal{Q}_{\Gamma(u)}(\sigma)$. Because $\tilde{q}_u = q_u$, $\{\mathcal{Q}_u; u \in U\}$ is Γ -reversible in structure.

We will show how Γ -reversibility in structure and Lemma 2 work for queueing

and their network models in the following sections. In the rest of this section, we briefly discuss a Poisson process that is frequently used in queueing applications as a counting process of customers. We are interested in how it arises in a Markov chain describing a queueing model and how it is connected to structure reversibility.

Definition 4. [Poisson process]. A counting process $N(t)$ is called a *Poisson process* with rate $\lambda > 0$ if it satisfies the following conditions.

- (a1) It has independent increments, that is, for any $n \geq 2$ and any time sequence $t_0 \equiv 0 < t_1 < t_2 < \dots < t_n$, the increments $N(t_\ell) - N(t_{\ell-1})$, $\ell = 1, 2, \dots, n$, are independent.
- (a2) It has stationary increments, that is, for any $s, t > 0$, the distribution of $N(s+t) - N(s)$ is independent of s .
- (a3) It is simple, that is, the increments $N(t) - N(t-)$ at time t is 0 or 1.
- (a4) $\lambda \equiv \mathbb{E}(N(1)) < \infty$.

It is not hard to see that this definition is equivalent to stating that the interoccurrence times for counting are independent and identically distributed subject to the exponential distribution with mean $1/\lambda$.

We note here two important facts related to reversibility. Let q be the transition rate function of the continuous-time Markov chain with a stationary distribution π such that $\pi(x) > 0$ for all $x \in S$. Let q_* be a transition rate function satisfying

$$q_*(x, x') \leq q(x, x'), \quad x, x' \in S. \quad (18)$$

Let $N_*(t)$ be the number of transitions of this Markov chain in the time interval $(0, t]$ generated by q_* . N_* is called a *counting process* of q_* .

Lemma 3. The counting process N_* is the Poisson process with rate $\lambda > 0$ whose occurrences are independent of the state of Markov

chain just before their counting instants if and only if

$$\sum_{x' \in S} q_*(x, x') = \lambda, \quad x \in S. \quad (19)$$

Lemma 3 easily follows from the following facts. The sojourn time at a given state of the Markov chain has an exponential distribution, the interoccurrence times of N_* are exponentially distributed, and mutually independent by the Markov property (see Ref. 4 for a complete proof).

For these q_* and π , recalling that $\pi(x) > 0$ for all $x \in S$, define $\tilde{q}_*$ as

$$\tilde{q}_*(x, x') = \frac{\pi(x')}{\pi(x)} q_*(x', x), \quad x, x' \in S$$

and let $\tilde{N}_*$ be the point process generated by $\tilde{q}_*$. A question is when $\tilde{N}_*$ is a Poisson process with rate λ . This is equivalent for $\tilde{q}_*$ to satisfy Equation 19 instead of q_* , namely,

$$\sum_{x' \in S} \pi(x') \tilde{q}_*(x', x) = \pi(x) \lambda, \quad x \in S. \quad (20)$$

In particular, if $q_* = q$, then Equation 20 automatically holds. This is intuitively clear because $\tilde{N}_*$ is the time-reversed process of N_* under π and a Poisson process is unchanged by time reversal. However, if $q_* \neq q$, then we do need the condition (Eq. 20).

We also can answer this question in terms of reversibility in structure. For this, let q be a transition rate function which has a stationary distribution π , and define $\mathcal{Q}(\pi)$ as

$$\mathcal{Q}(\pi) = \left\{ g \in \mathcal{T}(\pi); \exists c > 0, \right. \\ \left. \sum_{x' \in S} g(x, x') = c, \forall x \in S, g \leq q \right\},$$

where $g \leq q$ means $g(x, x') \leq q(x', x)$ for $x, x' \in S$. Define $\tilde{g}$ as $\tilde{g}(x, x') = \frac{\pi(x')}{\pi(x)} g(x', x)$. Then, the condition that $\tilde{q}_* \in \mathcal{Q}(\pi)$ for $q_* \in \mathcal{Q}(\pi)$ is equivalent to Equation 20. Thus, we have the following result.

Lemma 4. Assume that the Markov chain $\{X(t)\}$ has a stationary distribution π . For each $q_* \in \mathcal{Q}(\pi)$, the counting process $\tilde{N}_*$ generated by $\tilde{q}_*$ is a Poisson process with rate $\lambda > 0$ whose occurrences are independent of the state of Markov chain just after their counting instants if and only if q_* is reversible in structure with respect to $\mathcal{Q}(\pi)$.

QUEUE AND REACTING SYSTEM

We now describe queueing models by a continuous-time Markov chain. As a simplistic example, we first consider a single-server queue. Assume that customers arrive according to the Poisson process with rate $\lambda > 0$, customers are served in the manner of first-come and first-served with independently and exponentially distributed service times with mean $1/\mu$ ($\mu > 0$), which are also independent of the arrival process. This model is called an $M/M/1$ queue.

Let $X(t)$ be the number of customers at time t of the $M/M/1$ queue. Since the remaining times to the next arrivals and to the service completions of a customer being served are independent and subject to exponential distributions, $X(t)$ is a continuous-time Markov chain with state space $S = \{0, 1, \dots\}$, and its transition rate function q is given by

$$q(x, x') = \begin{cases} \lambda, & x' = x + 1, \\ \mu, & x' = x - 1 \geq 0, \\ 0, & \text{otherwise,} \end{cases} \quad x, x' \in S.$$

Hence, $\{X(t)\}$ is a special case of a birth-and-death process, and therefore reversible if it has a stationary distribution. From Equation 10, the stationary measure π is given by

$$\pi(x) = \pi(0) \rho^x, \quad x \in S,$$

where $\rho = \lambda/\mu$, and satisfies (Eq. 9). Hence, the stationary distribution exists if and only if $\rho < 1$.

The $M/M/1$ queue is a simple model, but it has key ingredients of reversibility. For

this, we introduce transition rate functions for arrivals and departures. Let

$$\begin{aligned} q_a(x, x') &= \lambda 1(x' = x + 1), \\ q_d(x, x') &= \lambda 1(x' = x - 1 \geq 0). \end{aligned}$$

Then,

$$\sum_{x' \in S} q_a(x, x') = \lambda, \quad x \in S, \quad (21)$$

$$\sum_{x' \in S} \pi(x') q_d(x', x) = \pi(x) \lambda, \quad x \in S. \quad (22)$$

Hence, by Lemmas 3 and 4, both N_a and N_d are Poisson processes with the same rate λ . Also, by these lemmas, N_a is independent of the past history of the Markov chain while N_d is independent of its future history. These facts are also valid for the $M/M/s$ queue that increases the number of servers in the $M/M/1$ queue from one to s as the Markov chain for the queue is also a birth-and-death process. These facts are known as *Burke's theorem* [5].

We now closely look at the $M/M/1$ queue and Burke's theorem; we can see that Poisson arrivals and Poisson departures occur as long as Equations 21 and 22 are satisfied, and it may not be necessary for the Markov chain $\{X(t)\}$ to be a birth-and-death process. In this situation, we need to have a larger state space S and to suitably choose $(q_a, \pi) \in \mathcal{T}$ for arrival and $(q_d, \pi) \in \mathcal{T}$ for departures, where π is the stationary distribution of $\{X(t)\}$. Another issue is that we may not be able to define the arrival process *a priori* because it may be a superposition of departure processes of other queues in network applications.

For resolving these issues, we separate the exogenous arrival process from the queueing system, and introduce the following formulation according to [4]. Let S be the state space of the queueing system removing the arrival process. Assume the following dynamics.

- (b1) A customer under state x departs with rate $q_d(x, x')$, changing x to x' .
- (b2) An arriving customer who finds state $x \in S$ changes it to $x' \in S$ with probability $p_a(x, x')$, where $\sum_{x' \in S} p_a(x, x') = 1$ for each $x \in S$.

- (b3) The system state x changes to x' with rate $q_1(x, x')$ without departure. This state transition is said to be *internal*.

This model is fairly general as a queueing system, but has nothing to do with a customer arrival process. We refer to this model as a *reacting system*.

We feed Poisson arrivals with rate $\alpha > 0$ to this reacting system, and define transition rate functions by

$$\begin{aligned} q(x, x') &= \alpha p_a(x, x') + q_d(x, x') \\ &\quad + q_i(x, x'), \quad x, x' \in S. \end{aligned} \quad (23)$$

Denote the Markov chain with transition rate function q by $\{X(t)\}$. This Markov chain represents the dynamics given by (b1), (b2), and (b3), where all actions except for (b1) occur with exponentially distributed interoccurrence times.

Assume that this Markov chain has a stationary distribution π . Then, the transition rate function $\tilde{q}$ of the time-reversed Markov chain $\{\tilde{X}(t)\}$ is given by Equation 6. A question is under what conditions $\{\tilde{X}(t)\}$ represents a reacting system with Poisson arrivals.

This problem can be stated using Γ -reversibility in structure. For this, define q_a as

$$q_a(x, x') = \alpha p_a(x, x'), \quad x, x' \in S,$$

and let $U = \{a, d\}$ and let Γ be the permutation on U such that $\Gamma(a) = d$ and $\Gamma(d) = a$. This Γ is clearly reflective. Define $\mathcal{Q}_a(\sigma)$ for $\sigma \in M(S)$ as

$$\begin{aligned} \mathcal{Q}_a(\sigma) &= \left\{ g_a \in \mathcal{T}(\sigma); \exists c > 0, \sum_{x' \in S} g_a(x, x') \right. \\ &\quad \left. = c, \forall x \in S \right\}, \end{aligned} \quad (24)$$

and let $\mathcal{Q}_a = \cup_{\sigma \in M(S)} \mathcal{Q}_a(\sigma)$. Then, $q_a \in \mathcal{Q}_a(\pi)$ for the stationary distribution π of q . In other words, $\{X(t)\}$ is a reacting system if and only if $(q_a, \pi) \in \mathcal{Q}_a$. The question is what is $\mathcal{Q}_d$ for q with $\{q_a, q_d\}$ to be Γ -reversible in structure with respect to $\{\mathcal{Q}_a, \mathcal{Q}_d\}$. The answer is immediate from Lemma 2; namely, we must have

$$\mathcal{Q}_d(\sigma) = \{g_d \in \mathcal{T}(\sigma); \tilde{g}_a \in \mathcal{Q}_a(\sigma)\}, \quad \sigma \in M(S),$$

and therefore, recalling that $\tilde{g}_a(x, x') = \frac{\pi(x')}{\pi(x)} g_d(x', x)$ for $\pi(x) > 0$,

$$\begin{aligned} \mathcal{Q}_d(\sigma) &= \left\{ g_d \in \mathcal{T}(\sigma); \exists c > 0, \sum_{x' \in S} \sigma(x') g_d(x', x) \right. \\ &= c \sigma(x), \forall x \in S \left. \right\}, \sigma \in M(S). \end{aligned} \quad (25)$$

Thus, we put $\mathcal{Q}_d = \cup_{\sigma \in M(S)} \mathcal{Q}_d(\sigma)$.

Let q_a and q_d be a subtransition rate function of q , and assume that q has its stationary distribution π , then the condition for $q_d \in \mathcal{Q}_d(\pi)$ of Equation 25 can be written as, for some $\beta > 0$,

$$\sum_{x' \in S} \pi(x') q_d(x', x) = \beta \pi(x), \quad x \in S, \quad (26)$$

which is called *quasi-reversibility* in the literature (see, e.g., Ref. 4). Thus, we have the following theorem with help of Lemmas 2 and 4.

Theorem 1. $\{\mathcal{Q}_a, \mathcal{Q}_d\}$ defined through Equations 24 and 25 is Γ -reversible in structure. Let $\{X(t)\}$ be the Markov chain with transition rate functions (Eq. 23) for the reacting system with Poisson arrivals, and assume that this Markov chain has a stationary distribution π . Then, the following conditions are equivalent.

- (c1) $\{\tilde{X}(t)\}$ is also a reacting system.
- (c2) $\{X(t)\}$ with $\{q_a, q_b\}$ is Γ -reversible in structure with respect to $\{\mathcal{Q}_a(\pi), \mathcal{Q}_d(\pi)\}$.
- (c3) $\{X(t)\}$ is quasi-reversible, that is, Equation 26 holds for some $\beta > 0$.
- (c4) The departure process N_d from the reacting system is the Poisson process with rate β that is independent of the future evolution of the Markov chain $\{X(t)\}$.

Example 1. [M/M/1 batch service queue]. Modify the M/M/1 queue in such a way that a batch of customers depart at the service completion instant subject to the following rule. All customers depart if a requested batch size is greater than

the number of customers in the system at that moment. Otherwise, the requested number of customers depart. It is assumed that the batch sizes to be requested are independently and identically distributed with a finite mean. We call this model M/M/1 batch service queue. This queue does not have Poisson departures if all batch departures are counted as departures, but it does if we only count full batches, that is, the departing batches whose sizes meet those to be requested. Thus, this reacting system is reversible in structure if full batches are only counted as departures. Details on this model can be found in Sect. 2.6 of [4].

In Theorem 1, we have assumed Poisson arrivals. It may be questioned what will occur if the arrival process is not Poisson. To answer it, we consider an external source to produce arrivals as another reacting system that is independent of the original reacting system, where departures from the latter system become arrivals to the external source but do not cause any state changes of the source. Thus, we have the closed system in which the two reacting systems are cyclically connected. If the time-reversed model has the same structure, it is not hard to see that the external source produces Poisson arrivals and departures from the original system are also Poisson. Hence, the Poisson arrivals arise under the reversibility in structure of this cyclic model.

A reacting system with Poisson arrivals can model not only a single-node queue but also a network of queues if the state space S is appropriately chosen. A problem is whether the reversibility enables us to get the stationary distribution. Obviously, this is easier for a smaller system. This motivates us to consider reacting systems as nodes of a network of queues and to construct the stationary distribution of the network from those of reacting systems of Poisson arrivals. This program is executed in the section titled "Queueing Network and Product Form."

Another question on the reacting system is whether or not we can consider arrival processes depending on the state of the reacting system. We solve this problem incorporating

exogenous arrivals as one of departures suitably extending the state space S . This is done in the section titled “Self-Reacting System and Balanced Departure.”

Before looking at those problems, we will discuss a concrete example of quasi-reversibility.

SYMMETRIC SERVICE

The examples in the section titled “Queue and Reacting System” are reversible in structure, but have no internal state transitions. In this section, we introduce the reacting system that has internal state transitions. For this, we classify customers by types, and split the transition rate functions by them.

Let T be the set of customer types and let arrivals and departures be classified by customer types. We split p_a, q_d, α and β into p_{au}, q_{du}, α_u and β_u with $u \in T$, respectively. Thus, the transition rate function of the Markov chain is changed to

$$q(x, x') = \sum_{u \in T} (\alpha_u p_{au}(x, x') + q_{du}(x, x')) + q_1(x, x'), \quad x, x' \in S.$$

Assume that q has a stationary distribution π . Then, similar changes are made for the time-reversed model, and all the arguments are parallel for the reversibility. For example, the quasi-reversibility condition (Eq. 26) is changed to

$$\sum_{x' \in S} \pi(x') q_{du}(x', x) = \beta_u \pi(x), \quad x \in S, u \in T. \quad (27)$$

Let $U = \{au, du; u \in T\}$, and let Γ be the reflective permutation such that $\Gamma(au) = du$ for all $u \in T$. We define $Q_{au}(\pi)$ and $Q_{du}(\pi)$ similarly to $Q_a(\sigma)$ and $Q_d(\sigma)$ of Equations 24 and 25.

To describe the service discipline, we assume that there are infinitely many service positions numbered as $1, 2, \dots$, and customers in the system are allocated to positions 1 to n when n customers are in system. We further assume that each x in the

state space S has the following form when n customers are in system.

$$x = (n, \{(u_\ell, w_\ell); \ell = 1, 2, \dots, n\}), \\ u_\ell \in T, w_\ell \in \{1, 2, \dots\},$$

where u_ℓ and w_ℓ are the type and remaining service stage, respectively, of a customer in position ℓ . Let k_u be the service stage of a newly arriving customer of type u , and let η_{uj} be the completion rate of the stage j , that is, the sojourn time at the stage j is exponentially distributed with mean $1/\eta_{uj}$. We assume the following dynamics under the state x .

- (d1) A type u customer arrives and chooses position i among $\{1, 2, \dots, n+1\}$ with probability $\delta(i, n+1)$, and customers in positions $i, i+1, \dots, n$ move to $i+1, i+2, \dots, n+1$. That is, x is changed to x' :

$$x' = (n+1, \{(u_\ell, w_\ell); \ell = 1, 2, \dots, i\}, (u, k_u), \{(u_\ell, w_\ell); \ell = i+1, i+2, \dots, n\}).$$

- (d2) A type u_i customer in position i with $w_i = 1$ departs with rate $\gamma(i, n)\eta_{u_i 1}$, and customers in positions $i+1, i+2, \dots, n$ move to $i, i+1, \dots, n-1$. That is, x is changed to x' :

$$x' = (n-1, \{(u_\ell, w_\ell); \ell = 1, 2, \dots, i-1\}, \{(u_\ell, w_\ell); \ell = i+1, i+2, \dots, n\}).$$

- (d3) A type u_i customer in position i with $w_i \geq 2$ decreases its remaining service stage by 1 with rate $\gamma(i, n)\eta_{u_i w_i}$, and customers in positions $\ell \neq i$ are unchanged. That is, x is changed to x' :

$$x' = (n, \{(u_\ell, w_\ell); \ell = 1, 2, \dots, i-1\}, (u_i, w_i - 1), \{(u_\ell, w_\ell); \ell = i+1, i+2, \dots, n\}).$$

Note that (d1), (d2), and (d3) uniquely specify p_a, q_d , and q_1 . This reacting system is fairly general, but it is hard to get the stationary distribution even for Poisson

arrivals. Thus, we consider the simpler situation assuming that, for each $n \geq 1$,

$$\phi(n)\delta(\ell, n) = \gamma(\ell, n), \quad \ell = 1, 2, \dots, n, \quad (28)$$

where $\phi(n) = \sum_{\ell=1}^n \gamma(\ell, n)$. A service discipline satisfying this condition is referred to as *symmetric service* (see, e.g., Ref. 1). This service discipline includes processor sharing and preemptive last-come last-served, but cannot be first-come and first-served except for exponentially distributed service times.

It is well known that the reacting system with Poisson arrivals and symmetric service, which is called a *symmetric queue*, has the stationary distribution π :

$$\begin{aligned} \pi(n, \{(u_\ell, w_\ell); \ell = 1, 2, \dots, n\}) \\ = c^{-1} \prod_{\ell=1}^n \frac{\alpha_{u_\ell}}{\phi(\ell)\eta_{u_\ell w_\ell}}, \quad 1 \leq w_\ell \leq k_{u_\ell}, \end{aligned}$$

if it exists, where c is the normalizing constant (e.g., see Ref. 4). Furthermore, the quasi-reversibility (Eq. 27) holds, and therefore $\tilde{q}_{ud} \in \mathcal{Q}_{du}(\pi)$, where

$$\tilde{q}_{du}(x, x') = \frac{\pi(x')}{\pi(x)} \alpha p_{au}(x', x), \quad x, x' \in S.$$

Thus, the symmetric queue is Γ -reversible in structure with respect to $\{\mathcal{Q}_{au}(\pi), \mathcal{Q}_{du}(\pi); u \in T\}$. Furthermore, we can verify that

$$q_I(x, x') = \tilde{q}_I(x, x'), \quad x, x' \in S.$$

These facts are particularly useful in incorporating symmetric queues as nodes of a network.

In symmetric service, the service amount of type u customer is the sum of k_u independent exponentially distributed random variables with means $1/\eta_{uk_u}, 1/\eta_{u(k_u-1)}, \dots, 1/\eta_{u1}$. Hence, if we let $\eta_{uj} = k_u \zeta_u$, then the service amount distribution of a type u customer is a k_u -order Erlang distribution with mean ζ_u . Furthermore, if type u is split to finitely many subtypes and if each subtype has finitely many stages, then the amount of work for type u customer is subject to a mixture of Erlang distributions. Thus, we can approximate any service amount distribution by this finitely many stage model.

SELF-REACTING SYSTEM AND BALANCED DEPARTURE

We modify the reacting system so that it includes the exogenous source as a part of the system, and departures are immediately transferred to arrivals. We allow having different types for arrivals and departures. Let T be the set of types. In this section, an element of T may represent characteristics other than a type of customer. For example, it may represent batch sizes of arrivals and departures and their vectors. We slightly change the dynamics (b1) and (b2) while retaining (b3).

- (e1) An entity $u \in T$ is released under state x with rate $q_{du}(x, y)$ changing state x to y .
- (e2) A released entity $u \in T$ is transferred to entity u' under state y with probability $r(u, y, u')$, where

$$\sum_{u' \in T} r(u, y, u') = 1, \quad u \in T, y \in S.$$

- (e3) A transferred entity u' under state $y \in S$ changes the state from y to $x' \in S$ with probability $p_{au'}(y, x')$, where $\sum_{x' \in S} p_{au'}(y, x') = 1$ for each $u' \in T$ and $y \in S$.

We assume that (e2) and (e3) instantaneously follow after (e1). Then, we can define the transition rate function q of the Markov chain $\{X(t)\}$ as

$$\begin{aligned} q(x, x') = \sum_{u, u' \in T, y \in S} q_{(u, y, u')}^{(D \rightarrow A)}(x, x') \\ + q_i(x, x'), \quad x, x' \in S, \end{aligned} \quad (29)$$

where q_I is a transition rate function for internal transitions, and

$$q_{(u, y, u')}^{(D \rightarrow A)}(x, x') = q_{du}(x, y) r(u, y, u') p_{au'}(y, x'). \quad (30)$$

We refer to the model described by this Markov chain as a *self-reacting system*. One may think of r as a routing probability matrix. For each u, u', y , we denote the set of pairs (g, σ) of a transition rate function g of the

form (Eq. 30) for some q_{du} , r and $p_{qu'}$ and measure σ on S to support q by $\mathcal{Q}_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})}$.

By ν , we denote the solution of the following traffic equation for each $y \in S$,

$$\sum_{u' \in T} \nu(u', y) r(u', y, u) = \nu(u, y), u \in T, y \in S. \quad (31)$$

We always assume that the solution ν exists and $\nu(u, y) > 0$ if $r(u', y, u) > 0$. This existence is always the case if T is a finite because $\{r(u, y, u'); u, u' \in S\}$ is a transition probability matrix for each $y \in S$. However, ν may not be unique not only because any multiplication with a function of y is again a solution but also because the transition probability matrix may not be irreducible. Nevertheless, it will be uniquely determined under certain reversibility in structure (see Remark 3 below).

Let σ be a supporting measure for the transition rate function q , that is, $\sigma \in M_{\text{sup}}(q)$. Of course, the stationary distribution of q is such a measure, but we do not restrict σ to be so. Define

$$\begin{aligned} & \tilde{q}_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})}(x, x') \\ &= \begin{cases} \frac{\sigma(x')}{\sigma(x)} q_{(u',y,u)}^{(\mathcal{D} \rightarrow \mathcal{A})}(x', x), & \sigma(x) > 0, x, x' \in S, \\ 0, & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$\begin{aligned} & \tilde{q}_I(x, x') \\ &= \begin{cases} \frac{\sigma(x')}{\sigma(x)} q_I(x', x), & \sigma(x) > 0, x, x' \in S, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Let $U = \{(u, y, u'); u, u' \in T, y \in S\}$, and let Γ be the permutation such that $\Gamma((u, y, u')) = (u', y, u)$ for $u, u' \in T$ and $y \in S$. This Γ is clearly reflective. We would like to find under what conditions we have $(\tilde{q}_{(u',y,u)}^{(\mathcal{D} \rightarrow \mathcal{A})}, \sigma) \in \mathcal{Q}_{(u',y,u)}^{(\mathcal{D} \rightarrow \mathcal{A})}$ for all $(u, y, u') \in U$.

We first note that $(\tilde{q}_{(u',y,u)}^{(\mathcal{D} \rightarrow \mathcal{A})}, \sigma) \in \mathcal{Q}_{(u',y,u)}^{(\mathcal{D} \rightarrow \mathcal{A})}$ is equivalent to stating that there are some $\tilde{q}_{du'}$, $\tilde{r}$ and $\tilde{p}_{au}$ satisfying

$$\begin{aligned} & \sigma(x) q_{du}(x, y) r(u, y, u') p_{au'}(y, x') \\ &= \sigma(x') \tilde{q}_{du'}(x', y) \tilde{r}(u', y, u) \tilde{p}_{au}(y, x). \end{aligned} \quad (32)$$

Define β , $\tilde{\beta}$, $\tilde{r}$, $\tilde{p}_{au}$, and $\tilde{q}_{au}$ as

$$\begin{aligned} \beta(u, y) &= \sum_{x \in S} \sigma(x) q_{du}(x, y), \\ \tilde{\beta}(u', y) &= \sum_{u \in T} \beta(u, y) r(u, y, u'), \\ \tilde{r}(u', y, u) &= \frac{\beta(u, y)}{\tilde{\beta}(u', y)} r(u, y, u'), \\ \tilde{q}_i(x, x') &= \frac{\sigma(x')}{\sigma(x)} q_i(x', x), \\ \tilde{p}_{au}(y, x) &= \frac{\sigma(x)}{\beta(u, y)} q_{du}(x, y), \\ \tilde{q}_{au}(x', y) &= \frac{\beta(u', y)}{\sigma(x')} p_{au'}(y, x'). \end{aligned}$$

Then, we can verify (Eq. 32), that is, we always have $(\tilde{q}_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})}, \sigma) \in \mathcal{Q}_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})}$ for all $(u, y, u') \in U$. Thus, no condition is needed for q of the self-reacting system to be Γ -reversible in structure except for the existence of a supporting measure σ . This fact exactly explains the reversibility in (3) of the section titled "Introduction." The problem is that the reversibility in structure gives no clue to compute σ as the stationary distribution.

To get out this situation, we consider a stronger reversibility condition. For this, we assume that the routing function $\tilde{r}$ is identical to the time reversal of r with respect to its stationary measure ν of Equation 31; that is, we assume that

$$\tilde{r}(u', y, u) = \frac{\nu(u, y)}{\nu(u', y)} r(u, y, u'), u, u' \in T, y \in S. \quad (33)$$

This is equivalent to $\beta = \tilde{\beta} = \nu$, which is further equivalent to

$$\sum_{x' \in S} \sigma(x') q_{du}(x', x) = \nu(u, x), u \in T, x \in S. \quad (34)$$

We denote the subset of $\mathcal{Q}_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})}$ which satisfies (Eq. 34) for some ν of Equation 31 by $\mathcal{Q}_{(u,y,u')}^{(\mathcal{D} \rightleftharpoons \mathcal{A})}$. We refer to a self-reacting system with $q_{(u,y,u')}^{(\mathcal{D} \rightarrow \mathcal{A})} \in \mathcal{Q}_{(u,y,u')}^{(\mathcal{D} \rightleftharpoons \mathcal{A})}(\sigma)$ for all $(u, y, u') \in U$ and some $\sigma \in M_{\text{sup}}(q)$ as a self-reacting system with a reversible routing mechanism

because Equation 33 is equivalent to stating that $\tilde{r}$ has an invariant measure $\tilde{v}$ such that

$$v(u, y) = \tilde{v}(u, y), \quad u \in T, x \in S,$$

where $\mathcal{Q}_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})}(\sigma) = \{g \in \mathcal{T}(\sigma); (g, \sigma) \in \mathcal{Q}_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})}(\sigma)\}$.

Thus, we have the following theorem.

Theorem 2. $\{\mathcal{Q}_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})}; (u, y, u') \in T \times S \times S\}$ is always Γ -reversible in structure. If the Markov chain $\{X(t)\}$ with transition rate function q given by Equation 29 has a stationary distribution π and $q_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})} \in \mathcal{Q}_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})}(\pi)$, that is, it is a self-reacting system, then the time-reversed Markov chain $\{\tilde{X}(t)\}$ is also a self-reacting system, and the following conditions are equivalent.

- (f1) $\{X(t)\}$ is Γ -reversible in structure with respect to $\{\mathcal{Q}_{(u, y, u')}^{(\mathcal{D} \rightarrow \mathcal{A})}(\pi); (u, y, u') \in U\}$; that is, Equation (34) is satisfied for $\sigma = \pi$.
- (f2) $\{X(t)\}$ represents the self-reacting system with a reversible routing mechanism.
- (f3) $\{\tilde{X}(t)\}$ is a self-reacting system with a routing function $\tilde{r}$ of Equation 33.

Remark 3. v is uniquely determined by Equation (34) if the stationary distribution π is given.

We now construct the stationary distribution from the solution of the traffic equation 31. From the stationary equation 5, the stationary distribution π must satisfy

$$\pi(x) = \frac{1}{a(x)} \sum_{x' \in S} \pi(x') q(x', x),$$

where $a(x) = \sum_{x' \in S} q(x', x)$. Substituting q of Equation 29 into this equation and using the traffic equation 31 and Equation 34 with

$\sigma = \alpha\pi$ for some constant $\alpha > 0$, we have

$$\begin{aligned} \pi(x) &= \frac{1}{a(x)} \sum_{x' \in S} \pi(x') \left(\sum_{y \in S} \sum_{u, u' \in T} q_d(x', y, u') \right. \\ &\quad \left. r(u', y, u) p_a(u, y, x) + q_i(x', x) \right) \\ &= \frac{1}{a(x)} \left(\alpha \sum_{y \in S} \sum_{u, u' \in T} v(y, u') r(u', y, u) \right. \\ &\quad \left. p_a(u, y, x) + \sum_{x' \in S} \pi(x') q_i(x', x) \right) \\ &= \frac{1}{a(x)} \left(\alpha \sum_{y \in S} \sum_{u \in T} v(y, u) p_a(u, y, x) \right. \\ &\quad \left. + \sum_{x' \in S} \pi(x') q_i(x', x) \right). \end{aligned}$$

We assume

$$(e4) \quad \tilde{q}_I(x, x') = q_I(x, x') \text{ for } x', x \in S \text{ and } b(x) \equiv \sum_{x' \in S} q_I(x, x') < a(x) \text{ for all } x \in S.$$

Then π is uniquely determined by

$$\pi(x) = \frac{\alpha}{a(x) - b(x)} \sum_{y \in S} \sum_{u \in T} v(y, u) p_a(u, y, x), \quad x \in S. \quad (35)$$

Thus we can construct the stationary distribution π from v in this case.

Example 2. [Balanced departure under symmetric service]. Suppose that a self-reacting system satisfies the condition (e4). We consider conditions to have Equation (34) for this network. Define Φ as

$$\Phi(x) = \frac{1}{a(x) - b(x)} \sum_{y \in S} \sum_{u \in T} v(y, u) p_a(u, y, x), \quad x \in S. \quad (36)$$

We assume that state x is uniquely determined by y and u when entity u is released under state $x \in S$. Denote this x by $g(y, u)$, and q_d as

$$q_d(x, y, u) = \frac{v(y, u)}{\Phi(x)} \mathbf{1}(x = g(y, u)), \quad x, y \in S, u \in T. \quad (37)$$

We refer to this departure rate as Φ -balanced according to the terminology used in Ref. 6 (see also Example 4 below). Then, if $\sum_{x \in S} \Phi(x) < \infty$, we choose $\alpha = (\sum_{x \in S} \Phi(x))^{-1}$, and let

$$\pi(x) = \alpha \Phi(x), \quad x \in S. \quad (38)$$

It is easy to check that this π is indeed the stationary distribution and the reversibility condition (34) is satisfied. Thus, by Theorem 2, if q_d is given by Equation 37 and if (e4) is satisfied, then we have the stationary distribution π of Equation 38 for Φ of Equation 36. This self-reacting system has a reversible routing mechanism.

Example 3. [Batch movement network]. We further specialize the self-reacting system with a reversible routing mechanism of Example 2 assuming that there are no internal state transitions, that is, $b(x) \equiv 0$. Let $S = \mathbb{Z}_+^n$ and $T = \mathbb{Z}_+^{n+1}$, that is, its state is an n -dimensional vector with nonnegative integer entries. We assume that $q_d(x, y, u) > 0$ if and only if $y = x - u^+$, where $u^+ = (u_1, u_2, \dots, u_n)$ for $u = (u_0, u_1, \dots, u_n) \in T$. We further assume that $r(u, y, u') > 0$ if and only if $|u| = |u'|$, where $|u| = u_0 + u_1 + \dots + u_n$.

This system can be considered as an open queueing network with n nodes and batch movements, where the i th entry of $x \in S$ represents the number of customers at node i , and the j th entry of $u \in S$ is a departing batch for $j \neq 0$ while u_0 is a departing batch from an external source for $q_d(x, y, u) > 0$. Furthermore, the u are transferred to u' under each state y not changing their total numbers. It is notable that, if $u_0 = 0$ always holds, then T can be reduced to $\mathbb{Z}^n$, and the model is considered as a closed network. This model is referred to as a *queueing network with batch movements*.

We assume that $q_d(x, y, u) > 0$ only if $y = x - u^+$ and $p_a(u, y, x) > 0$ only if $x = y + u^+$. These assumptions are reasonable for batch movements. Then, $q_d(x, y, u) > 0$ if and only if

$p_a(u, y, x) > 0$ and Equation 37 becomes

$$q_d(x, y, u) = \frac{v(x - u^+, u)}{\Phi(x)} 1(y = x - u^+), \quad x, y \in S, u \in T. \quad (39)$$

We show that we can choose an arbitrary Φ . For this, we need to verify (Eq. 36). For the batch movement network, Equation 36 becomes

$$\Phi(x) = \frac{1}{a(x)} \sum_{u \in T} v(x - u^+, u) 1(x - u^+ \in S), \quad x \in S.$$

However, this automatically holds because

$$\begin{aligned} a(x) &= \sum_{x' \in S} \sum_{y \in S} \sum_{u, u' \in T} q_d(x, y, u) \\ &\quad \times r(u, y, u') p_a(u', y, x') \\ &= \frac{1}{\Phi(x)} \sum_{u \in T} v(x - u^+, u) \\ &\quad \times 1(x - u^+ \in S), \quad x \in S. \end{aligned}$$

Hence, we can choose an arbitrary Φ in Equation 39 as long as it is a positive-valued function. This model is called a *linear network* in Ref. 7. We refer to it as a *batch movement network with balanced departure*. There is significant literature on this class of networks (see, e.g., Refs 8–10).

If all batch sizes are units, then the batch movement network is reduced to the standard queueing network with single arrivals and departures.

Example 4. [State independent routing]. We consider a special case of the batch movement network with balanced departure of Example 3 in which $r(u, y, u')$ is independent of y . This model is said to have *independent routing*.

In this network, $v(y, u)$ can be independent of y , so it can be written as $\Psi(y)v(u)$ for an arbitrary positive-valued function Ψ on S and the solution v of the traffic equation 31 which

is a function on T . Thus, Equation 39 becomes

$$q_d(x, y, u) = \frac{\Psi(x - u^+)v(u)}{\Phi(x)} 1(y = x - u^+),$$

$$x, y \in S, u \in T, \quad (40)$$

where Ψ and Φ in Equation 39 can be arbitrary as long as they are positive-valued functions. Furthermore, if $v(u)$ is of the form $\prod_{i=1}^n w_i^{u_i}$ for positive constants $w_1, w_2, \dots, w_n$, then v is included in $\Psi(x - u^+)/\Phi(x)$ as

$$v(u) = \prod_{i=1}^n w_i^{u_i} = \frac{v(x)}{v(x - u^+)}.$$

Thus, Equation 40 is further simplified as

$$q_d(x, y, u) = \frac{\Psi(x - u^+)}{\Phi(x)} 1(y = x - u^+),$$

$$x, y \in S, u \in T, \quad (41)$$

and, if the stationary distribution exists, then it is given by

$$\pi(x) = C\Phi(x) \prod_{i=1}^n w_i^{x_i}, \quad x \in S,$$

where C is the normalizing constant. Otherwise, this π with $C = 1$ is a stationary measure. This model has been widely discussed in the literature. For example, Serfozo [6] called it the *Whittle network* with Φ -balanced departure intensities (see Theorem 1.48 of [6]). There are a lot of applications of this special case to telecommunication networks (e.g., see Ref. 11).

QUEUEING NETWORK AND PRODUCT FORM

We return to reacting systems. We use them here as nodes of a queueing network with Markovian routing, where the external source is included as one of those nodes. Consider a queueing network with n nodes, numbered as $1, 2, \dots, n$. We also have an external source as node 0. Let

$$J = \{0, 1, \dots, n\}.$$

Each node i is a reacting system with state space S_i , departure rate $q_d^{(i)}$, internal transition rate $q_I^{(i)}$, and arrival probability function $p_a^{(i)}$. Let T_i be the set of customer types arriving at and departing from node i . We assume the following dynamics.

- (g1) A departure of type $u_i \in T_i$ customer from node i occurs under state $x_i \in S_i$ with rate $q_{du_i}^{(i)}(x_i, x'_i)$ which changes the state x_i to $x'_i \in S_i$.
- (g2) A type u_i customer departing from node i goes to node j as type u_j customer with probability $r(iu_i, ju_j)$ independently of everything else. It is assumed that

$$\sum_{j \in J} \sum_{u_j \in T_j} r(iu_i, ju_j) = 1, \quad (i, u_i) \in J \times T_i.$$

- (g3) An arriving type u_j customer at node j under state $x_j \in S_j$ changes the state to $x'_j \in S_j$ with probability $p_{au_j}^{(j)}(x_j, x'_j)$. It is assumed that

$$\sum_{x'_j \in S_j} p_{au_j}^{(j)}(x_j, x'_j) = 1,$$

$$(u_j, x_j) \in T_j \times S_j, j \in J.$$

- (g4) The state $x_i \in S_i$ at node i changes to $x'_i \in S_i$ with rate $q_I^{(i)}(x_i, x'_i)$ not producing any departure.

We model this network by a continuous-time Markov chain $\{X(t)\}$. Its state space S is given by

$$S = S_0 \times S_1 \times \dots \times S_n$$

and its transition rate function q is given by

$$q(x, x') = \sum_{i, j \in J} \sum_{u \in T_i} \sum_{v \in T_j} q_{iu, jv}^{(d \rightarrow a)}(x, x') + \sum_{i \in J} q_I^{(i)}(x_i, x'_i),$$

where

$$q_{iu, jv}^{(d \rightarrow a)}(x, x') = \begin{cases} q_{du}^{(i)}(x_i, x'_i) r(iu, jv) p_{av}^{(j)}(x_j, x'_j) \\ \prod_{k \neq i, j} 1(x_k = x'_k), & i \neq j, \\ \sum_{y_i \in S_i} q_{du}^{(i)}(x_i, y_i) r(iu, iv) \\ p_{av}^{(j)}(y_j, x'_j) \prod_{k \neq i} 1(x_k = x'_k), & i = j. \end{cases} \quad (42)$$

We refer to the model described by this transition rate as a *network with reacting nodes and Markovian routing*.

We aim to construct a time-reversed process of this network model. For this, we consider the reacting system of node i with Poisson arrivals with rate $\alpha_u^{(i)}$. We assume

- (g5) There exist positive $\{\alpha_{u_i}^{(i)}\}$ and $\{\beta_{u_i}^{(i)}\}$ such that the Markov chain with transition rate q_i given by

$$\begin{aligned} q_i(x_i, x'_i) = & \sum_{u_i \in T_i} \left(\alpha_{u_i}^{(i)} p_{au_i}^{(i)}(x_i, x'_i) \right. \\ & \left. + q_{du_i}^{(i)}(x_i, x'_i) \right) + q_i^{(i)}(x_i, x'_i), \\ & x_i, x'_i \in S_i, i \in J, \end{aligned} \quad (43)$$

has the stationary distribution π_i for each $\{\alpha_{u_i}^{(i)}\}$, and

$$\begin{aligned} \sum_{x'_i \in S_i} \sum_{u_i \in T_i} \pi_i(x'_i) q_{du_i}^{(i)}(x'_i, x_i) &= \beta_{u_i}^{(i)} \pi_i(x_i), \\ x_i, x'_i \in S_i, i \in J, \end{aligned} \quad (44)$$

$$\begin{aligned} \alpha_{u_j}^{(j)} &= \sum_{i \in J} \sum_{u_i \in T_j} \beta_{u_i}^{(i)} r(iu_i, ju_j), \\ j \in J, u_j \in T_j. \end{aligned} \quad (45)$$

Then, similarly to the reacting system with Poisson arrivals in the section titled “Queue and Reacting System,” we define the time-reversed reacting system and routing function in the following way:

$$\begin{aligned} \tilde{p}_{au_i}^{(i)}(x_i, x'_i) &= \frac{1}{\beta_{u_i}^{(i)}} \frac{\pi_i(x'_i)}{\pi_i(x_i)} q_{du_i}^{(i)}(x'_i, x_i), \\ & x_i, x'_i \in S_i, u_i \in T_i, \\ \tilde{q}_{du_i}^{(i)}(x_i, x'_i) &= \frac{\pi_i(x'_i)}{\pi_i(x_i)} \alpha_{u_i}^{(i)} p_{au_i}^{(i)}(x'_i, x_i), \\ & x_i, x'_i \in S_i, u_i \in T_i, \\ \tilde{q}_i^{(i)}(x_i, x'_i) &= \frac{\pi_i(x'_i)}{\pi_i(x_i)} q_i^{(i)}(x'_i, x_i), \quad (x_i, x'_i) \in S_i, \\ \tilde{r}(iu_i, ju_j) &= \frac{\beta_{u_j}^{(j)}}{\alpha_{u_i}^{(i)}} r(ju_j, iu_i). \end{aligned}$$

Let $U_i = \{(a, u); u \in T_i\}, \{(d, u); u \in T_i\}$ for each $i \in J$, and let $U = \prod_{i \in J} U_i$. Define the permutation Γ_i on U_i such that $\Gamma_i(\{(a, u); u \in T_i\}) = \{(d, u); u \in T_i\}$ for each $i \in J$, and define the permutation on U such that $\Gamma(\{v_i; v_i \in U_i, i \in J\}) = \{\Gamma(v_i); v_i \in U_i, i \in J\}$. Clearly, Γ is reflective. Let $\mathcal{Q}_{au}^{(i)}(\pi_i)$ and $\mathcal{Q}_{su}^{(i)}(\pi_i)$ be similarly defined on $\mathcal{Q}_a(\pi)$ and $\mathcal{Q}_d(\pi)$, respectively, of a reacting system (Eqs 24 and 25).

The following facts can be proved using Lemma 1 (see Ref. 4 for its proof).

Theorem 3. *If the network with reacting nodes and Markovian routing satisfies (g5), that is, each node i has a stationary distribution π_i and is quasi-reversible in separation, that is, the Markov chain with transition rate function q_i is Γ_i -reversible in structure with respect to $\{\mathcal{Q}_{au}^{(i)}(\pi_i), \mathcal{Q}_{du}^{(i)}(\pi_i); u \in T_i\}$ for each i , then it has the stationary distribution π given by $\pi = \prod_{i=1}^n \pi_i$, that is,*

$$\pi(x) = \prod_{i=0}^n \pi_i(x_i), \quad x \equiv (x_0, x_1, \dots, x_n) \in S. \quad (46)$$

Furthermore, this Markov chain is Γ -reversible in structure with respect to $\{\mathcal{Q}_{(i,u),(j,v)}^{(D=A)}(\pi); u \in T_i, v \in T_j, i, j \in J\}$, where $\mathcal{Q}_{(i,u),(j,v)}^{(D=A)}(\pi)$ is defined as the set of transition rate functions of the form (Eq. 42) satisfying Equation 44.

Example 5. A Jackson network is well known to have the product form stationary distribution, provided it is stable. This network is a special case of the network of Theorem 3. To see this, let $S = \mathbb{Z}_+^{n+1}$ and let

$$\begin{aligned} p_a^{(i)}(x, x') &= \begin{cases} 1(x = x')i = 0, \\ 1(x'_i = x_i + 1 \geq, x'_j = x_j, i \neq j), \\ i \neq 0, \end{cases} \\ q_d^{(i)}(x, x') &= \begin{cases} \lambda i = 0, \\ \min(x_i, s_i) \mu_i 1(x'_i = x_i - 1 \geq 0, \\ x'_j = x_j, i \neq j), i \neq 0, \end{cases} \\ r(iu, ju) &= p_{ij}. \end{aligned}$$

Then, the reacting system with $p_a^{(i)}, q_d^{(i)}$ and Poisson arrivals is reversible in structure, and it is not hard to see that (g4) is satisfied with $\alpha_{u_i}^{(i)} = \beta_{u_i}^{(i)}$ for all i and u_i . If we replace the i th reacting system for $i \neq 0$ in the Jackson network by a reacting system with symmetric service, (g4) is also satisfied. This network is called a *Kelly network* [1]. Its special case limited to some symmetric service is called a *BCMP network* [12], where the workload for service to be requested is generally distributed, but any distribution can be approximated by the distribution with finitely many stages.

There are examples for Theorem 3 in the other direction. A customer is said to be negative if it removes one customer without any other change at its arriving node; this was introduced by Gelenbe [13]. Such a customer is said to be a negative signal if it instantaneously moves around in a network according to a given Markovian routing until it meets an empty node or leaves the network. It is known that Theorem 3 can be applied to the *Jackson network with negative signals*. In this case, we may have $\alpha_{u_i}^{(i)} \neq \beta_{u_i}^{(i)}$ for some i and u_i . See Ref. 4 for a full description of those networks.

OTHER STOCHASTIC PROCESSES

We have mainly considered a continuous-time Markov chain. Reversibility in structure also may be considered for other stochastic processes. A discrete-time Markov chain is one of them. There is much literature for local balance, but those results are mostly parallel to those of the continuous-time case. They also can be reformulated by reversibility in structure.

Both Markov chains have discrete state spaces. To directly consider service times, we need continuous-valued components in the state of the system. A good model for this is a generalized semi-Markov process (GSMP) and its variants such as reallocatable GSMP (see, e.g., Refs 14 and 15). Roughly speaking, the states of those processes are the pairs of discrete state and positive real numbers, called *remaining lifetimes* (or *attaining*

lifetimes). If one of the remaining lifetimes expires, then it causes the transition of the discrete state, which may produce new lifetimes.

The GSMPs are particularly useful in proving the so-called insensitivity, which means that the stationary distribution depends on the lifetime distributions only through their means. This insensitivity holds if and only if certain partial balance holds, which is a spatial case of reversibility in structure. Under this condition, we again have the symmetric queues and their networks. There are some related results here if there are service interruptions that force customers to leave (see, e.g., Ref. 14). However, no new development has been made since the nineties.

There are some other stochastic processes related to queueing networks. For example, they are often considered using reflecting random walk or semi-Martingale reflecting Brownian motion (SRBM). Both are multidimensional and take values in the nonnegative orthant. Furthermore, the time-reversed process has been studied for general Markov processes (see, e.g., Ref. 16). It is known that SRBM on a polyhedral domain, which includes the nonnegative orthant as a special case, has the product form stationary distribution if and only if the so-called skew symmetric condition holds, which implies that the time-reversed process is also SRBM (see Theorem 1.2 of [17]).

CONCLUDING REMARKS

As we have seen, available models for queues and their networks are limited under reversibility in structure. Those models are still actively applied, but theoretical study is less active now. What can we do to gain more applicability? There are some comments here.

1. The reversibility described in this article has two ingredients. One is how strongly we set it up. In this article, we mainly considered this aspect for queueing network processes. Another is which class of stochastic processes

we take into consideration. We briefly discussed this issue in the section titled “Other Stochastic Processes.” If we take an appropriate class, say, not too small but not too large, then we may find more tractable models for queueing networks. This direction of study seems to be not well developed.

2. The existing studies for reversibility in queues have been intended to get the stationary distribution in closed form. This target may be too big, and it may be better to have smaller target. For example, tail asymptotic of the stationary distribution is one of them (see, e.g., Ref. 18). This may include reconsideration of the definition of reversibility in structure.

ACKNOWLEDGMENTS

The author is grateful to an anonymous referee for their thoughtful comments. The definition of reversibility in structure was greatly changed from its original form due to them.

REFERENCES

1. Kelly P. *Reversibility and Stochastic Networks*. New York: John Wiley & Sons; 1979.
2. Miyazawa M, Zwart B. Wiener-hopf factorizations for a multidimensional Markov additive process and their applications to reflected processes. *Stoch Syst* 2012;2(1):67–114.
3. Bremaud P. *Markov chains: Gibbs field, Monte Carlo simulation and queues*. Berlin: Springer; 1999.
4. Chao X, Miyazawa M, Pinedo M. *Queueing networks: customers, signals and product form solutions*. Chichester: John Wiley & Sons; 1999.
5. Burke PJ. The output of a stationary $m/m/s$ queueing system. *Ann Math Stat* 1968;39:1144–1152.
6. Serfozo R. *Introduction to stochastic networks*. Volume 44, *Applications of Mathematics* (New York). New York: Springer-Verlag; 1999.
7. Miyazawa M. Structure-reversibility and departure functions of queueing networks with batch movements and state dependent routing. *Queueing Syst* 1997;25:45–75.
8. Boucherie RJ, van Dijk NM. Product forms for queueing networks with state-dependent multiple job transitions. *Adv Appl Probab* 1991;23:152–187.
9. Henderson W, Taylor PG. Some new results of queueing networks with batch movement. *J Appl Probab* 1991;28:409–421.
10. Serfozo R. Queueing networks with dependent nodes and concurrent movements. *QUESTA* 1993;13:143–182.
11. Bonald T. Insensitive traffic models for communication networks. *Discrete Event Dyn Syst* 2007;17:405–421.
12. Baskett F, Chandy KM, Muntz RR, Palacios FG. Open, closed and mixed networks of queues with different classes of customers. *J Assoc Comput Mach* 1975;22:248–260.
13. Gelenbe E. Product-form queueing networks with negative and positive customers. *J Appl Probab* 1991;28:656–663.
14. Miyazawa M. Insensitivity and product-form decomposability of reallocatable GSMP. *Adv Appl Probab* 1993;25:415–437.
15. Schassberger R. Tow remarks on insensitive stochastic models. *Adv Appl Probab* 1986;18:791–814.
16. Gettoor RK, Sharpe MJ. Naturality, standardness, and weak duality for markov processes. *Probab Theory Relat Fields* 1984;67:1–62.
17. Williams RJ. Reflected Brownian motion under skew symmetric condition. *Probab Theory Relat Fields* 1987;75:459–485.
18. Masakiyo M. Light tail asymptotics in multidimensional reflecting processes for queueing networks. *TOP* 2011;19(2): 233–299.

REVIEWS OF MAINTENANCE LITERATURE AND MODELS

FRANCIS K. N. LEUNG
Department of Management
Sciences, City University of Hong
Kong, Hong Kong

INTRODUCTION

Modeling of the maintenance (i.e., inspection, overhaul, repair, replacement, etc.) of deteriorating systems has been of significant importance for the past 70 years, especially because of its applications in industry and medicine. There is an extensive literature on various types of preventive maintenance (PM) policies for stochastically failing single-unit and multiunit systems. In this article, we review part of the literature that deals with the problem of PM of a stochastically failing single-unit system. Notice that a unit has several or many components but is regarded as a single entity in either case, and is, respectively, called a *simple item* (e.g., a ball bearing) or a *complex system* (e.g., a car) subsequently. We emphasize how certain features of the various models affect their analysis and the conclusions that can be drawn from them. It should be noted that what we present in this article is only a very small portion of the maintenance field; so many mathematically outstanding papers have been omitted because of space considerations. However, many good review articles exist that provide an extensive overview of the field and can be helpful as guidance through the literature. We mention surveys by McCall [1], Pierskalla and Voelker [2], Sherif and Smith [3], Valdez-Flores and Feldman [4], and Wang [5] that detail much of the literature on single-unit deteriorating systems up to 2002; while surveys by Thomas [6], Cho and Parlar [7], and Dekker *et al.* [8] concentrate on multiunit deteriorating systems. These combined sets of papers list over

2000 references altogether. We also mention monographs written by Barlow and Proschan [9], Jardine [10], Nakagawa [11], and Wang and Pham [12], and handbooks edited by Ben-Daya *et al.* [13], Osaki [14], Pham [15], and Kobbacy and Murthy [16].

In the past, maintenance problems received little attention and research in this area did not have much impact. This is changing because of the increasing importance of the role of maintenance in the new industrial environment. Maintenance, if optimized, can be used as a key factor in an organization's efficiency and effectiveness. It also enhances the organization's ability to be competitive and meets its stated objectives.

Research in the area of maintenance management and engineering is on the rise. Over the past few decades, there has been tremendous interest and a great deal of research in the area of maintenance modeling and optimization. Models have been developed that cover a wide variety of maintenance problems. Although the subject of maintenance modeling is a late developer compared to other areas like production systems, the interest in this area is growing at an unprecedented rate.

MAINTENANCE POLICIES

This section provides a brief overview of the maintenance modeling research areas. It is not meant to be a complete review of maintenance models but rather act as an informative introduction to important maintenance modeling areas. Areas covered include replacement, repair and replacement, imperfect PM, condition-based maintenance (CBM), and inspection.

Maintenance strategies can be classified into three categories:

1. Corrective maintenance (CM), where replacement or repair is performed only at the time of failure. This may be

the appropriate strategy in some cases, such as when the hazard (also called *failure rate*) is constant and/or when the failure has no serious cost or safety consequences, or it is low on the priority list.

2. PM, where maintenance is performed on a scheduled basis with scheduled intervals often based on manufacturers' recommendations and experience with the equipment. This may involve replacement, repair, or both.
3. CBM, where maintenance decisions are based on the current condition of the equipment; thus avoiding unnecessary maintenance and performing maintenance activities only when they are needed to avoid failure. CBM relies on condition-monitoring techniques such as oil analysis, vibration analysis, and other diagnostic techniques to make maintenance decisions.

Implementing these strategies involves familiar maintenance activities like inspection, replacement, and repairs. In this section, we discuss representative models and modeling techniques dealing with maintenance strategies, including replacement, repair and replacement, imperfect PM, CBM, and inspection decisions. Replacement models are discussed first. They are of particular interest since they were among the earliest models developed. They were extended to include repairs and later to incorporate condition-monitoring information.

Replacement Models

In this section, we develop several basic replacement models. There are two kinds of replacement: failure replacement (a type of CM) and scheduled replacement (a type of PM). We are concerned with the latter whereby availability is increased and/or expected cost is reduced. Throughout this section, we assume that the lifetime of a simple item is governed by the known continuous-time distribution $F(x)$ for $x \geq 0$ with a finite mean μ , and further assume that there is an infinite number of new identical items available for replacement.

We are interested in minimizing the long-run expected cost rate (i.e., the expected cost per unit time in the steady state) or maximizing the limiting availability.

Age Replacement Models. Consider a renewal process with interarrival time distribution $F(x)$, which corresponds to the process of the item with only failure replacement. We define two scheduled replacement policies.

Definition 1. Age replacement: a simple item is replaced upon failure or at age T_a , whichever comes first.

Definition 2. Block replacement: a simple item or a complex system is replaced upon failure and at times $T_b, 2T_b, 3T_b, \dots$

First we discuss age replacement models, and then we develop block replacement models. Under an age replacement policy, an item is replaced upon failure or at age T_a , whichever comes first. Such a renewal process becomes a truncated renewal process with interarrival times $\{\min(X_n, T_a), n = 1, 2, \dots\}$. The stochastic behavior of such a process can be analyzed using the results of the renewal process [17].

We are only interested in the long-run expected cost rate. Let c_1 and c_2 denote the costs of a failure and a scheduled replacement, respectively. It is natural to assume that $c_1 > c_2$.

The long-run expected cost rate can be given by the expected cost per cycle divided by the expected duration per cycle, where a cycle terminates whenever a renewal (an item is replaced upon failure or at age T_a , whichever comes first) takes place. The long-run expected cost rate, using the renewal reward theorem (see, e.g., Theorem 3.6.1 of Ross [17]), is given by

$$C(T_a) = \frac{c_1 \Pr(X_n \leq T_a) + c_2 \Pr(X_n > T_a)}{E[\min(X_n, T_a)]}.$$

A cycle terminates at $\{\min(X_n, T_a), n = 1, 2, \dots\}$ and repeats itself again and again. Note that

$$\begin{aligned} E[\min(X_n, T_a)] &= \int_0^{T_a} x dF(x) + T_a \bar{F}(T_a) \\ &= \int_0^{T_a} \bar{F}(x) dx, \end{aligned}$$

where $\bar{F}(x) \equiv 1 - F(x) = \Pr(X_n > x)$; so we have

$$C(T_a) = \frac{c_1 F(T_a) + c_2 \bar{F}(T_a)}{\int_0^{T_a} \bar{F}(x) dx}. \quad (1)$$

Assume that there exists the density $f(x)$ of $F(x)$. Differentiating $C(T_a)$ with respect to T_a and equating the derivative to zero, we have

$$h(T_a) \int_0^{T_a} \bar{F}(x) dx - F(T_a) = \frac{c_2}{c_1 - c_2}, \quad (2)$$

where $h(x) = \frac{f(x)}{\bar{F}(x)}$, the hazard (or failure rate) function. Equation (2) is a nonlinear equation with respect to $T_a \geq 0$. Differentiating the left-hand side of Equation (2), we have

$$h'(T_a) \int_0^{T_a} \bar{F}(x) dx,$$

which is nonnegative (i.e., the left-hand side of Equation (2) is a monotonically increasing function) if we assume $\frac{dh(T_a)}{dT_a} \equiv h'(T_a) \geq 0$ (i.e., $F(T_a)$ is in the class of increasing failure rate (IFR) distributions). Because the left-hand side of Equation (2) equals zero at $T_a = 0$ and approaches $h(\infty)\mu - 1$ as $T_a \rightarrow \infty$ if $F(T_a)$ is IFR, we have the following theorem (Fig. 1).

Theorem 1.

- (i) If $h(x)$ is monotonically increasing and $h(\infty) > \frac{c_1}{\mu(c_1 - c_2)}$, then there exists a finite and unique T_a^* which satisfies Equation (2) and the resulting minimum cost rate is

$$C(T_a^*) = (c_1 - c_2)h(T_a^*). \quad (3)$$

- (ii) If $h(x)$ is monotonically increasing and $h(\infty) \leq \frac{c_1}{\mu(c_1 - c_2)}$, or $h(x)$ is nonincreasing, then the optimal policy is $T_a^* = \infty$, that is, there is no scheduled replacement and $C(\infty) = \frac{c_1}{\mu}$.

The proof of Theorem 1 is given in Osaki and Nakagawa [18].

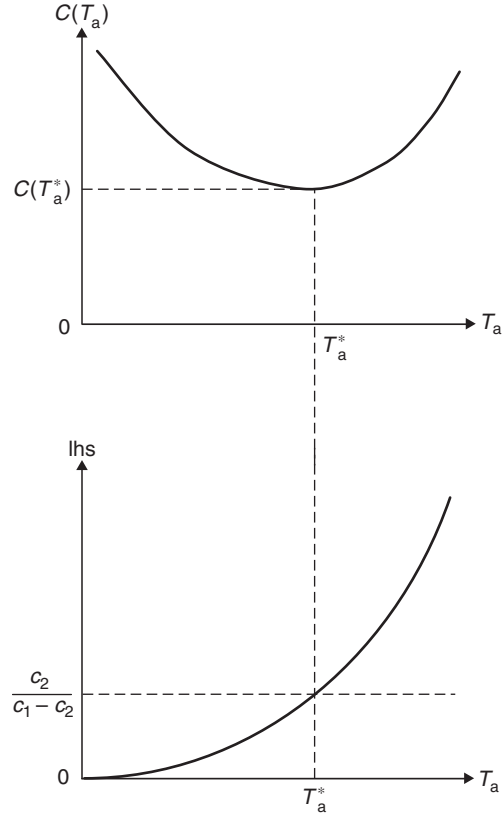


Figure 1. Behavior of $C(T_a)$ and left-hand side (lhs) of Equation (2) for the IFR distribution.

Example 1.

Case (a):

The lifetime follows a 2-Erlang distribution with parameter λ (which is a special case of the gamma distribution with scale parameter $\alpha = \frac{1}{\lambda}$ and shape parameter $\beta = 2$), that is, $F(x) = 1 - (1 + \lambda x) e^{-\lambda x}$, where $\mu = \alpha\beta = \frac{2}{\lambda}$ is the mean lifetime. The hazard function is $h(x) = \frac{\lambda^2 x}{\lambda x + 1}$ and $h(\infty) = \lambda$.

If $h(\infty) = \lambda > \frac{\lambda c_1}{2(c_1 - c_2)}$, that is, $c_1 > 2c_2$, then there exists a finite and unique T_a^* satisfying Equation (2) and $C(T_a^*) = \frac{\lambda^2 T_a^* (c_1 - c_2)}{\lambda T_a^* + 1}$. If $c_1 \leq 2c_2$, the optimal policy is $T_a^* = \infty$, that is, there is no scheduled replacement and $C(\infty) = \frac{\lambda c_1}{2}$.

Case (b):

Consider the Weibull distribution $F(x) = 1 - e^{-\lambda x^\beta}$ with $\lambda > 0$ and $\beta > 1$. The corresponding hazard function is $h(x) = \lambda \beta x^{\beta-1}$ and the

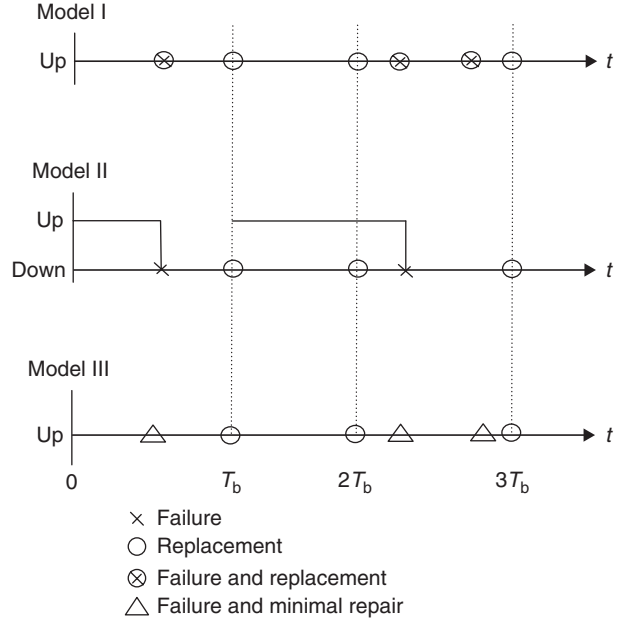


Figure 2. Interpretations of the behavior for Models I, II, and III.

optimal replacement age is the unique solution to

$$h(T_a) \int_0^{T_a} e^{-\lambda x^\beta} dx - 1 + e^{-\lambda T_a^\beta} = \frac{c_2}{c_1 - c_2}.$$

The resulting minimum cost rate is $C(T_a^*) = (c_1 - c_2)h(T_a^*)$.

Block Replacement Models. For age replacement models, we have to observe the age of an item, which is sometimes difficult administratively. However, for block replacement models, we do not need to observe the age of an item, but replace at $T_b, 2T_b, 3T_b, \dots$, which is easier to administer in general (Definition 2). Here, we develop the following three variations of block replacement models (see Fig. 2, which is adapted from Osaki [19, p. 203]).

Model I: a failed item is replaced instantaneously at failure.

Model II: a failed item remains inoperable until the next scheduled replacement.

Model III (often called *minimal repair model*): a failed system undergoes minimal repair.

In general, Model I represents a typical block replacement model and is well known. Model II is a simple variation of Model I. Model III represents periodic replacement with minimal repair upon failure, where we assume that after each failure only minimal repair is performed so that the rate of occurrence of failure (ROCOF), denoted by $r(t)$, and the effective age of a system are not disturbed.

Minimal repair means that a failed system will function, after repair, with the same ROCOF and the same effective age as at the time of the last failure. For a minimal repair model where repair time is assumed to be negligible, a nonhomogeneous Poisson process (NHPP) in which the ROCOF is monotonic can provide at least a good first-order model for a deteriorating system [20]. In other words, failures constitute an NHPP with a suitable parametric form for ROCOF. Such a model can be justified since a system (e.g., a car) has many components that can fail so that replacing some failed components has very little effect on future system reliability.

Balaban and Singpurwalla [21, pp. 66–67] provided examples of minimal repair actions:

1. A TV set has stopped functioning because of the failure of an integrated circuit panel. The set functions as soon as the failed panel is replaced. If the other components are left alone, then the set is not like a new one and a minimal repair action has been performed.
2. A tire that has run several miles is punctured by a nail and goes flat; so the car using the tire is considered to have failed. A repair of the puncture restores the car to an operational status. If we assume that the puncture patch has not strengthened the tire by a significant amount, then a minimal repair action has been performed on the car.
3. A coronary occlusion may cause heart failure. Cardio pulmonary resuscitation (CPR) may revive the patient. Assuming revival without damage to vital organs, we may view CPR as a minimal repair action.

Model I. A failed item is replaced by a new identical one during the replacement interval T_b and the scheduled replacement for the nonfailed item is performed at $T_b, 2T_b, 3T_b, \dots$. Then the long-run expected cost rate is

$$C_1(T_b) = \frac{c_1 M(T_b) + c_2}{T_b} = \frac{c_1 \int_0^{T_b} m(t) dt + c_2}{T_b}, \quad (4)$$

where $M(t) = \sum_{n=1}^{\infty} F^{(n)}(t)$ is the renewal function with the underlying distribution $F(x)$, $F^{(n)}(t)$ is the n -fold convolution of $F(x)$ with itself, $m(t) = \frac{dM(t)}{dt}$ is the renewal density, c_1 is the cost of replacement for the failed item, and c_2 is the cost of scheduled replacement for the nonfailed item. To minimize the long-run expected cost rate in Equation (4), we obtain the following equation by differentiating $C_1(T_b)$ and equating the derivative to zero:

$$T_b m(T_b) - \int_0^{T_b} m(t) dt = \frac{c_2}{c_1}. \quad (5)$$

This is a necessary condition for the existence of a finite T_b^* and the resulting minimum cost

rate is

$$C_1(T_b^*) = c_1 m(T_b^*). \quad (6)$$

In general, by specifying the underlying distribution $F(x)$ we can obtain the renewal function $M(t)$ and renewal density $m(t)$, by which we can solve the nonlinear Equation (5).

Example 2. The lifetime follows a 2-Erlang distribution with parameter λ , that is, $F(x) = 1 - (1 + \lambda x)e^{-\lambda x}$.

The Laplace–Stieltjes (L–S) transform of $M(t)$ is

$$\tilde{M}(s) = \frac{\tilde{F}(s)}{1 - \tilde{F}(s)},$$

where $\tilde{F}(s) = \int_0^{\infty} e^{-sx} dF(x)$ is the L–S transform of $F(x)$.

The L–S transform of $F(x) = 1 - (1 + \lambda x)e^{-\lambda x}$ is

$$\tilde{F}(s) = \left(\frac{\lambda}{s + \lambda} \right)^2.$$

Therefore,

$$\begin{aligned} \tilde{M}(s) &= \frac{\left(\frac{\lambda}{s + \lambda} \right)^2}{1 - \left(\frac{\lambda}{s + \lambda} \right)^2} = \frac{\lambda^2}{s(s + 2\lambda)} \\ &= \frac{\lambda}{2s} - \frac{1}{4} \left(\frac{2\lambda}{s + 2\lambda} \right). \end{aligned}$$

Inversion obtains

$$M(t) = \frac{\lambda t}{2} - \frac{1 - e^{-2\lambda t}}{4},$$

and differentiation yields

$$m(t) = \frac{\lambda}{2} - \frac{\lambda e^{-2\lambda t}}{2}.$$

Equation (5) becomes

$$-\frac{\lambda T_b e^{-2\lambda T_b}}{2} + \frac{1 - e^{-2\lambda T_b}}{4} = \frac{c_2}{c_1}$$

or

$$e^{-2\lambda T_b} + 2\lambda T_b e^{-2\lambda T_b} = \frac{c_1 - 4c_2}{c_1}.$$

Since the left-hand side of the above equation is always positive, a finite solution exists

provided $c_1 > 4c_2$, that is, a failure replacement is at least four times more expensive than a scheduled replacement. We can draw the following conclusion: if $\frac{c_2}{c_1} \geq \frac{1}{4}$, we should perform only failure replacement. If $\frac{c_2}{c_1} < \frac{1}{4}$, there exists a finite and unique T_b^* that minimizes Equation (4) and the resulting minimum cost rate is $C_1(T_b^*) = \frac{\lambda c_1}{2}(1 - e^{-2\lambda T_b^*})$.

Model II. In the first model, we assumed that a failed item is detected instantaneously just after failure. In Model II, we assume that failure is detected only at $T_b, 2T_b, 3T_b, \dots$. An item is always replaced at $T_b, 2T_b, 3T_b, \dots$, but is not replaced at the time of failure and the item remains inoperable for the duration of time from the occurrence of failure until its detection. The expected duration from the occurrence of failure until its detection per cycle is given by

$$\int_0^{T_b} (T_b - x) dF(x) = \int_0^{T_b} F(x) dx.$$

The long-run expected cost rate is

$$C_2(T_b) = \frac{c_1 \int_0^{T_b} F(x) dx + c_2}{T_b}, \quad (7)$$

where c_1 is the cost of failure per unit time, and c_2 is the cost of scheduled replacement for the failed or nonfailed item. Comparing Equation (7) with Equation (4), we find a similar long-run expected cost rate, outside of the integrand.

Referring to Equation (5), we have

$$\begin{aligned} T_b F(T_b) - \int_0^{T_b} F(x) dx &= \int_0^{T_b} [F(T_b) - F(x)] dx \\ &= \frac{c_2}{c_1}. \end{aligned} \quad (8)$$

Assuming $T_b \rightarrow \infty$, we have

$$\lim_{T_b \rightarrow \infty} \int_0^{T_b} [F(T_b) - F(x)] dx = \int_0^{\infty} \bar{F}(x) dx = \mu.$$

It is easy to show that if $\mu > \frac{c_2}{c_1}$, there exists an optimal T_b^* which is a finite and unique solution to Equation (8).

Example 3. The lifetime follows a 2-Erlang distribution with parameter λ , that is, $F(x) =$

$1 - (1 + \lambda x) e^{-\lambda x}$, where $\mu = \frac{2}{\lambda}$ is the mean lifetime.

If $\lambda \geq \frac{2c_1}{c_2}$, we should not perform any scheduled replacement. If $\lambda < \frac{2c_1}{c_2}$, there exists a unique T_b^* satisfying

$$\frac{2 - e^{-\lambda T_b}(2 + 2\lambda T_b + \lambda^2 T_b^2)}{\lambda} = \frac{c_2}{c_1},$$

and the resulting minimum cost rate is

$$C_2(T_b^*) = c_1[1 - (1 + \lambda T_b^*)e^{-\lambda T_b^*}].$$

Model III (i.e., Minimal Repair Model). We assume that minimal repair is performed when a system fails, and the ROCOF, $r(t)$, and the effective age are not disturbed by each repair. If we consider a stochastic process $\{N(t), t \geq 0\}$ in which $N(t)$ represents the number of minimal repair actions up to time t , the process $\{N(t), t \geq 0\}$ is governed by an NHPP with mean value function

$$R(t) = \int_0^t r(s) ds.$$

Noting this fact, we have the following long-run expected cost rate for Model III:

$$C_3(T_b) = \frac{c_1 \int_0^{T_b} r(t) dt + c_2}{T_b}, \quad (9)$$

where c_1 and c_2 are the costs of minimal repair for the failed components and of replacement for the system, respectively. Comparing Equation (9) with Equation (4), we also have the same long-run expected cost rate except for the integrand.

Example 4. If we assume that the ROCOF has the same analytic form as the hazard function given in Example 1, we have the following two cases:

Case (a):

$$r(t) = \frac{\lambda^2 t}{\lambda t + 1}.$$

There exists a unique T_b^* satisfying

$$\frac{(\lambda T_b)^2}{\lambda T_b + 1} - \int_0^{T_b} \frac{\lambda^2 t}{\lambda t + 1} dt = \frac{c_2}{c_1},$$

or

$$\frac{(\lambda T_b)^2}{\lambda T_b + 1} - \lambda \int_0^{T_b} \left(1 - \frac{1}{\lambda t + 1}\right) dt = \frac{c_2}{c_1},$$

or

$$\ln(\lambda T_b + 1) - \frac{\lambda T_b}{\lambda T_b + 1} = \frac{c_2}{c_1},$$

and the resulting minimum cost rate is

$$C_2(T_b^*) = \frac{c_1 \lambda^2 T_b^*}{\lambda T_b^* + 1}.$$

Case (b):

Consider the power law process (PLP), denoted by $r(t) = \lambda \beta t^{\beta-1}$ for $t \geq 0$ with $\lambda > 0$ and $\beta > 1$. The optimal replacement interval is the unique solution to

$$(\beta - 1)\lambda T_b^\beta = \frac{c_2}{c_1},$$

or

$$T_b^* = \left[\frac{c_2}{c_1 \lambda (\beta - 1)} \right]^{\frac{1}{\beta}}.$$

The resulting minimum cost rate is $C(T_b^*) = c_1 r(T_b^*)$.

The popularity of the PLP is twofold: first, it can model deteriorating or improving systems; second, point estimators for the parameters have simple closed-form expressions and tests of hypotheses can be undertaken using existing tables. The PLP denoted and given by $r(t) = \lambda \beta t^{\beta-1}$ for $t \geq 0$ and $\lambda, \beta > 0$ is the most important ROCOF parametric form in an NHPP model. If $\beta > 1$, the ROCOF increases with time, as often happens with aging machinery. This is one of the two main conditions for determining if it is worth carrying out a scheduled replacement (the other condition is that the cost c_2 of a system replacement is much greater than the cost c_1 of a minimal repair action). But if $0 < \beta < 1$, the ROCOF decreases with time; hence, the PLP can model reliability growth as well. The homogeneous Poisson process (HPP), which has constant ROCOF, is a special case of the PLP with $\beta = 1$. Rigdon and Basu [22] gave a detailed discussion on the PLP.

There are three notes on using the PLP, or any other model, for event times, namely:

1. The PLP, often misleadingly called the *Weibull process* (this name is a misnomer since only the time to first failure has a Weibull distribution), is a useful and simple model for describing the failure times of a repairable complex system. The term PLP, a phrase describing the functional form of the ROCOF function, was introduced by Ascher and Feingold [20].
2. Note that the ROCOF, which is sometimes called the *failure rate*, should not be confused with the hazard, which is also sometimes known as the *failure rate*; see Ascher and Feingold [20, Chapter 8] for a discussion of poor terminology.
3. Rigdon and Basu [22, p. 259] have reached the conclusion that “using the PLP or any other model for event times, one clearly indicates the time that data collection started and the time that it ceased. This is necessary so that the appropriate analysis, that is, an analysis based on failure-truncated or time-truncated data, can be applied and maximum information can be obtained from the data. For time-truncated data, the time between the last failure and the termination of the test contains some information that should not be wasted.”

In addition, we add two notes on the absence of sufficient data relating to the maintenance problem that are of interest when finding plausible models to be fitted and validated:

1. To overcome the problem of data shortage, we can make use of the judgment of the engineers for the elicitation of a statistical point process. Methods for eliciting expert opinions, incorporating one's judgment about the quality of the expert opinions, and combining the opinions of several experts are given by Lindley [23] and Van Noortwijk *et al.* [24]. Kypris and Singpurwalla [25], and Guida *et al.* [26] individually considered the PLP, $r(t) = \lambda \beta t^{\beta-1}$, with λ and β being unspecified and subject

to Bayesian revision. A Bayesian approach with meaningfully chosen priors would bring improvement to the analysis because of its ability to incorporate expert knowledge and engineering information into the inferential mechanism [27]; however, the approach requires much more burdensome computation.

2. Baker and Scarf [28, p. 9] have indicated that “although large sample sizes are required to estimate the optimal value of a decision variable, such as the optimal inspection interval or replacement period, to a high accuracy, it has been shown that the cost (or downtime) which is optimized can be reduced to near its true minimum value even for small sample sizes, because of the quadratic nature of a minimum. In fact, the expected excess cost over the minimum cost is inversely proportional to the sample size. This means that even modest sample sizes would be sufficient for the practical use of sensible models for determining cost- (or downtime-) based maintenance policies.”

Holland and McLean [29], Aven [30], Kumar and Klefsjo [31], and Leung and Cheng [32] provided case studies which show that minimal repair models are a good means of achieving both effective and efficient maintenance.

The seminal ideas of age replacement and minimal repair models were first introduced in the classic paper by Barlow and Hunter [33]. Twenty years later, Nguyen and Murthy [34] extended both models to a hazard and a ROCOF that are reduced by each repair, but their subsequent growths are increasing functions of the number of previous repairs. Other extensions and variations of age replacement and block replacement are discussed in Dohi *et al.* [35].

Repair-Replacement Models using Geometric Processes

The work in this section is substantially based on Lam [36].

Definition 3. Given a sequence of random variables $G_1, G_2, \dots$. If for some $r > 0$, $\{G_n r^{(n-1)}, n = 1, 2, \dots\}$ forms a renewal process (RP), then $\{G_n, n = 1, 2, \dots\}$ is a geometric process (GP). The parameter r is called the *common ratio of the GP*.

Three specializations of a GP are given below:

If $r > 1$, then the GP is called a *decreasing GP*.

If $0 < r < 1$, then the GP is called an *increasing GP*.

If $r = 1$, then the GP reduces to an RP.

Therefore, for a deteriorating system, it is reasonable to assume that the successive operating times of the system form a decreasing GP, whereas the corresponding consecutive repair times constitute an increasing GP (Fig. 3). However, the replacement times for the system are usually stochastically the same, no matter how old the used system is; hence, they will form an RP. This is the motivation behind the introduction of the GP approach.

Two Types of GP Models. Before deriving the repair-replacement models, the following assumptions are stated:

1. At the beginning, a new system is used.
2. Whenever the system fails, it can be repaired. Let X_n be the survival time after the $(n - 1)$ th repair, then a sequence $\{X_n, n = 1, 2, \dots\}$ forms a decreasing GP with parameter $r_a > 1$, and $E(X_1) \equiv \mu_{X_1} > 0$.
3. Let Y_n be the repair time after the n th failure, then a sequence $\{Y_n, n = 1, 2, \dots\}$ forms an increasing GP with parameter $0 < r_b < 1$, and $E(Y_1) \equiv \mu_{Y_1} \geq 0$. Note that $\mu_{Y_1} = 0$ means that the repair time is negligible.
4. A sequence $\{X_n, n = 1, 2, \dots\}$ and a sequence $\{Y_n, n = 1, 2, \dots\}$ are independent.

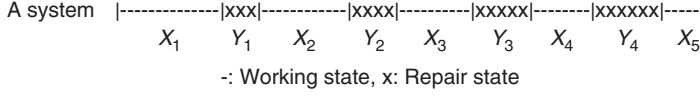


Figure 3. A realization of the system.

5. An average operating cost rate is c_o , an average repair cost rate is c_f , and an average revenue rate of a working system is w .
6. The system may sometime be replaced by a new and identical one. An average replacement cost rate under policy T or N is c_{RT} or c_{RN} , respectively, and an average replacement downtime under policy T or N is u_{RT} or u_{RN} , respectively. Two kinds of replacement policies are considered in this model.
 - a. A replacement policy T is a policy in which we replace the system whenever the working age of the system reaches T , a continuous decision variable. The working age T of a system at time t is the cumulative survival time by time t , that is,

$$T = \begin{cases} t - V_n & \text{for } U_n + V_n \leq t \\ & < U_{n+1} + V_n, \\ U_{n+1} & \text{for } U_{n+1} + V_n \leq t \\ & < U_{n+1} + V_{n+1}, \end{cases} \quad (10)$$

where $U_n = \sum_{i=1}^n X_i$, $V_n = \sum_{i=1}^n Y_i$, and $U_0 = 0$, $V_0 = 0$.

- b. A replacement policy N is a policy in which we replace the system at the time of N th failure since the last replacement, a discrete decision variable.

Under replacement policy T or N , the problem is to determine an optimal replacement policy T^* or N^* , respectively, such that the long-run expected loss per unit total time or per unit operation time is minimized.

Let T_n be the time between the $(n-1)$ th replacement and the n th replacement with $T_0 = 0$, then $\{T_n, n = 1, 2, \dots\}$ forms an RP. Applying the renewal reward theorem, the long-run expected loss per unit time is

obtained by

$$l(T) \text{ or } l(N) = \frac{E(\text{loss incurred in a cycle})}{E(\text{length of a cycle})}, \quad (11)$$

where the loss is defined as the total cost minus total revenue and a cycle is the time between two consecutive replacements.

Under the replacement policy T , denote the length of a cycle by W , then

$$W = T + V_n \text{ and } U_n < T \leq U_{n+1} \text{ for } n = 0, 1, \dots \quad (12)$$

From Definition 3 and Assumption 3, it follows that $E(Y_n) = \frac{\mu_{Y_1}}{r_b^{(n-1)}}$. Then,

$$\begin{aligned} E(W) &= E\left(T + \sum_{j=1}^n Y_j\right) \\ &= E\left[T + \sum_{j=1}^{\infty} Y_j I(n \geq j)\right], \end{aligned}$$

where I is the indicator random variable defined as

$$I(n \geq j) = \begin{cases} 1 & \text{if } n \geq j, \\ 0 & \text{if } n < j. \end{cases}$$

Since $n \geq j \Leftrightarrow U_j \leq T$, we have

$$\begin{aligned} E(W) &= E\left[T + \sum_{j=1}^{\infty} Y_j I(U_j \leq T)\right] \\ &= E(T) + \sum_{j=1}^{\infty} E(Y_j) E[I(U_j \leq T)] \\ &\quad \text{as } Y_j \text{ and } U_j \text{ are independent} \\ &= T + \sum_{j=1}^{\infty} \left(\frac{\mu_{Y_1}}{r_b^{(j-1)}}\right) [1 \times \Pr(U_j \leq T) \\ &\quad + 0 \times \Pr(U_j > T)] \\ &= T + \mu_{Y_1} \sum_{j=1}^{\infty} \frac{F_j(T)}{r_b^{(j-1)}}, \end{aligned}$$

where F_j is the distribution function of U_j .

Thus, from Equation (11) the long-run expected loss per unit total time $l(T)$ under policy T is given by

$$l(T) = \frac{(c_o - w)T + c_f \mu_{Y_1} \sum_{j=1}^{\infty} \frac{F_j(T)}{r_b^{(j-1)}} + c_{RT} u_{RT}}{T + \mu_{Y_1} \sum_{j=1}^{\infty} \frac{F_j(T)}{r_b^{(j-1)}} + u_R T}, \quad (13)$$

and also from Equation (11), the long-run expected loss per unit total time $l(N)$ under policy N is given by

$$l(N) = \frac{(c_o - w) \mu_{X_1} \sum_{j=1}^N \frac{1}{r_a^{(j-1)}} + c_f \mu_{Y_1} \sum_{j=1}^{N-1} \frac{1}{r_b^{(j-1)}} + c_{RN} u_{RN}}{\mu_{X_1} \sum_{j=1}^N \frac{1}{r_a^{(j-1)}} + \mu_{Y_1} \sum_{j=1}^{N-1} \frac{1}{r_b^{(j-1)}} + u_{RN}}. \quad (14)$$

In practice, we prefer to adopt the optimal policy N^* rather than use the optimal policy T^* , because of the much simpler form of $l(N)$. Moreover, under some mild conditions, Lam [37] proved that the optimal policy N^* is better than any policy T ; in particular, it is better than the optimal policy T^* .

To conclude this section, we mention two comprehensive reviews [38,39] of modeling some maintenance practices using GPs and statistical inference of model parameters, and a latest paper using the GP approach by Leung *et al.* [40].

Imperfect Preventive Maintenance

In many PM models, a system is assumed to be as good as new after each PM action. However, a more realistic situation is one in which the failure pattern of a preventively maintained system changes. One way to model this is to assume that, after PM, the ROCOF of the system is somewhere between as good (same) as new and as bad (same) as old. This concept is called *imperfect maintenance*.

To illustrate these types of models, consider the following PM policy proposed by Nakagawa [41]. PM is done at fixed intervals $x_k (k = 1, 2, \dots, N-1)$ and the system is replaced at the N th PM. If the system

fails between PMs, it undergoes only minimal repair. The PM is imperfect. The age after the k th PM reduces to $b_k t$, while it was t before PM (b_k is constant and $0 \leq b_k \leq 1$).

The problem is to find the optimal interval lengths $x_k (k = 1, 2, \dots, N-1)$ and the optimal number N of PMs when a replacement is done so that total expected cost rate is minimized. Let c_p be the cost of a scheduled PM, c_r be the cost of replacement, and c_m be the cost of minimal repair at failure. Also, let $r(t)$ denote the ROCOF function.

The imperfect PM model is developed under the following assumptions:

1. The age after the k th PM reduces to $b_k t$, while it was t before PM, that is, the system with age t becomes $t(1 - b_k)$ units of time younger at the k th PM, where $0 = b_0 < b_1 < \dots < b_N < 1$.
2. The system undergoes only minimal repair at failures between PMs. The ROCOF remains unchanged by minimal repair (by the definition of minimal repair).
3. The ROCOF function $r(t)$ is continuous and strictly increasing.
4. The times for PM, minimal repair, and replacement are negligible.
5. Time returns to zero after replacement, that is, the system is as good as new at replacement.

Let y_k be the effective age of the system immediately before the k th PM. Note that $y_k = x_k + b_{k-1} y_{k-1} (k = 1, 2, \dots, N)$. Hence, during the k th PM interval the age of the system changes from $b_{k-1} y_{k-1}$ to y_k . Note that $x_k = y_k - b_{k-1} y_{k-1}$.

It is more convenient to express the cost function in terms of y_k than in terms of x_k . The expected total cost per cycle is the sum of the expected cost of minimal repairs, the expected cost of PMs, and the cost of replacement. Thus,

$$E(C) = c_m \sum_{k=1}^N \int_{b_{k-1} y_{k-1}}^{y_k} r(t) dt + (N-1)c_p + c_r. \quad (15)$$

The expected cycle length is

$$E(T) = \sum_{k=1}^{N-1} (1 - b_k)y_k + y_N. \quad (16)$$

The problem is to find the optimal number N of PMs and the optimal interval lengths $x_k (k = 1, 2, \dots, N - 1)$ that minimize the total expected cost rate given by

$$C(y_1, y_2, \dots, y_N) = \frac{E(C)}{E(T)}. \quad (17)$$

Pham and Wang [42] provided a review of imperfect PM models. The reader is referred to this extensive review for more information on these models. In addition, some recent papers on imperfect PM modeling for a complex system with two dependent/independent failure modes: maintainable (e.g., the ROCOF of a mechanical component can be reduced by PM actions) and nonmaintainable (e.g., an electronic component) include Lin *et al.* [43], El-Ferik and Ben-Daya [44], Zequeira and Berenguer [45], Bartholomew-Biggs *et al.* [46], and Castro [47].

Condition-Based Maintenance

CBM consists of deciding whether or not to maintain a system according to its state using condition-monitoring techniques. There are three types of decisions that need to be made:

1. Selecting the parameters to be monitored. This depends on several factors such as the type of equipment and technology available.
2. Determining the inspection frequency.
3. Establishing the warning limits that trigger appropriate maintenance action, which could be static or dynamic.

Representative models for CBM, based on Cox's [48] proportional hazards (PH) model, are discussed here. Other modeling schemes such as the state space model and the Kalman filter for prediction [49] can also be used as modeling tools.

Cox's PH model differs from other models in the way the equipment reliability is estimated. In this case, equipment reliability is not estimated based only on its age as is usually the case. In addition to age, reliability is estimated also based on concomitant information such as the information obtained from condition-monitoring techniques (vibration analysis, oil analysis, etc.). Essentially the procedure is to modify the hazard function $h(x)$ by multiplying it by a function $g[z(x)]$ where $z(x)$ is the concomitant variable (sometimes called *covariate* or *explanatory variable*).

The PH model can be defined by

$$h(x) = h_0(x)g[z(x)] \text{ for } x \geq 0, \quad (18)$$

where $h_0(x)$ is called the *baseline hazard*, which is the hazard that the system would experience if the effects of all covariates are equal to zero.

A popular choice for the link function $g[z(x)]$ is the log-linear form

$$g[z(x)] = e^{\gamma z(x)}, \quad (19)$$

where γ is the regression coefficient. In fact, Cox's PH model is a multiple regression model. In the case of several covariates (several monitored variables), γ and $z(x)$ can be made vectors

$$h(x) = h_0(x)e^{\gamma_1 z_1(x) + \gamma_2 z_2(x) + \dots}. \quad (20)$$

The advantage of this method is that the hazard function is determined on the basis of both the age of equipment and its condition. This makes it very useful in a CBM context.

There are situations when the hazard increases with time. Then, it is reasonable to schedule a preventive replacement when the hazard becomes too high. To determine the optimal hazard level (threshold level), it is necessary to predict the future behavior of covariates. Makis and Jardine [50,51] assumed a widely accepted Markov process model to describe the stochastic behavior of covariates and presented optimal replacement decision models. One of these models was included in a CBM software package [52]. For more details, refer to Kumar and Klefsjo [53], Kumar [54], and Maillart [55].

Inspection

The inspection problem in maintenance tends to be centered around determining the optimal interval between checks on equipment to assess its status.

The problem of determining an optimal inspection policy has been widely discussed in the literature [2]. Two literature reviews of inspection are given in Leung [56] and Nakagawa [11, Chapter 8].

Some of these models combine inspection and replacement decisions. It is assumed that repairs at a modest cost will bring the production process to the in-control state. However, the residual life after these repairs will depend on the age of the system. After a certain number of repairs, replacement is performed. The problem is to determine an optimal inspection policy and the optimal replacement time that will minimize the expected cost per unit time. Models in these areas differ in the following features:

1. failure process mechanism (e.g., Markovian versus non-Markovian);
2. types of repair (e.g., minimal repairs);
3. Assumptions about inspection intervals (e.g., constant interval length and constant interval hazard).

Examples of this model include the work of Barlow and Proschan [9], Munford and Shahani [57], Munford [58,59], and Banerjee and Chiu [60].

Examples of models involving inspection decisions only include the work of Baker [61], Ben-Daya and Hariga [62], and Hariga [63]. More information on these models can be found in Hariga and Al-Fawzan [64]. In many cases, the inspection problem is jointly considered with repair and/or replacement decisions [65,66].

The other situation in which the inspection problem arises is when the purpose of the inspection is to find out whether the system has shifted to an out-of-control state, in which case a restoration action of the system to the in-control state is needed. Lee and Rosenblatt [67] considered an economic lot-sizing problem involving inspections aimed at assessing the status of the production process. In a

similar context, Rahim [68] and Ben-Daya [69] considered quality control inspections to assess the production system's status.

The inspection problem also arises in the context of CBM. As discussed earlier, CBM consists of deciding whether or not to maintain a system according to its state. The state of the system can be represented by an indicator variable that measures its deterioration. Once indicators, such as bearing wear and gauge readings, which are used to describe the state, have been specified, inspections are made to determine the values of these indicators. Appropriate maintenance action may then be taken, depending on the state. The state of the system is represented by the value taken by a random variable X . This random variable is observed at times t_i ($i = 1, 2, \dots$), and x_i is the value of X at time t_i . This variable is equal to zero if the system is new. The greater x_i , the closer the system is to a breakdown. The system breaks down if the value of X exceeds a given limit L . Taylor [70] proposed one of the earliest research works in this area. The model assumes that the system is subject to shocks that follow a Poisson distribution. The magnitude of system deterioration resulting from a shock follows an exponential distribution and the magnitudes of deterioration are independent of each other. The maintenance and repair costs are assumed to be constant. Under these assumptions, Taylor showed that the optimal policy is a control limit policy (CLP). In other words, there exists a limit x^* such that if the state of deterioration of the system is greater than x^* but less than a breakdown limit, then maintenance is required. However, no action is needed if the state of deterioration is below x^* .

Bergman [71] introduced a more general model in which the state of deterioration of the system is a nondecreasing random variable. He also proved that the CLP is an optimal policy. Zuckerman [72] provided the properties of the system (conditions) under which the CLP is optimal.

Several inspection models [73] use the delay-time concept. Delay time is defined as the time lapse from when a fault could first be noticed until the time when its repair can no longer be delayed because of unacceptable

failure consequences [74]. A repair can be undertaken during the delay-time period to avoid breakdown.

A basic delay-time inspection model is presented here. This model is based on the following assumptions:

1. The probability density function $f(h)$ of the delay time h is known.
2. Inspections take place at regular intervals of time of equal length T ; each inspection costs c_i and its duration is d_p .
3. Inspections are perfect.
4. The time of origin of the fault is uniformly distributed over time since the last inspection and is independent of h .
5. Faults arise at the rate of λ per unit time.
6. A breakdown repair costs c_f and has an average duration d_f , and an inspection repair costs c_p and has an average duration d_p .

The problem is to determine the length of the inspection interval T that minimizes the expected cost or downtime rate.

Suppose that a fault arising in the interval $(0, T)$ has a delay time in the interval $(h, h + dh)$ with probability $f(h)dh$. A breakdown repair will be performed if the fault arises in the interval $(0, T - h)$, otherwise an inspection repair will be carried out. The probability that the fault arises before $(T - h)$, given that a fault will occur, is $\frac{T-h}{T}$ (because of Assumption 4). The probability of a breakdown repair is given by

$$p(T) = \int_0^T \frac{T-h}{T} f(h) dh. \quad (21)$$

The expected downtime rate $D(T)$ is given by

$$D(T) = \frac{\lambda T d_f p(T) + d_p}{T + d_p}. \quad (22)$$

The expected cost rate $C(T)$ is given by

$$C(T) = \frac{\lambda T \{c_f p(T) + c_p [1 - p(T)]\} + c_i}{T + d_p}. \quad (23)$$

These models have been used in preventive maintenance, condition monitoring, and other settings. For a review of delay-time models, the reader is referred to the papers by Baker and Christer [75], Christer [73], and Wang [76].

REFERENCES

1. McCall JJ. Maintenance policies for stochastically failing equipment: a survey. *Manage Sci* 1965;11:493–524.
2. Pierskalla WP, Voelker JA. A survey of maintenance models: the control and surveillance of deteriorating systems. *Nav Res Logist Q* 1976;23:353–388.
3. Sherif YS, Smith ML. Optimal maintenance models for systems subject to failure: a review. *Nav Res Logist Q* 1981;28:47–74.
4. Valdez-Flores C, Feldman RM. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Nav Res Logist* 1989;36:419–446.
5. Wang HZ. A survey of maintenance policies of deteriorating systems. *Eur J Oper Res* 2002;139:469–489.
6. Thomas LC. A survey of maintenance and replacement models for maintainability and reliability of multi-item systems. *Reliab Eng* 1986;16:297–309.
7. Cho DI, Parlar M. A survey of maintenance models for multi-unit systems. *Eur J Oper Res* 1991;51:1–23.
8. Dekker R, Van Der Duyn Schouten FA, Willemann R. A review of multi-component maintenance models with economic dependence. *Math Methods Oper Res* 1997;45:411–435.
9. Barlow RE, Proschan F. *Mathematical theory of reliability*. New York: John Wiley & Sons; 1965.
10. Jardine AKS. *Maintenance, replacement and reliability*. London: Pitman; 1973.
11. Nakagawa T. *Maintenance theory of reliability*. London: Springer; 2005.
12. Wang HZ, Pham H. *Reliability and optimal maintenance*. Germany: Springer; 2006.
13. Ben-Daya M, Duffuaa SO, Raouf A. *Maintenance, modeling and optimization*. Boston (MA): Kluwer Academic Publishers; 2000.
14. Osaki S. *Stochastic models in reliability and maintenance*. Berlin: Springer; 2002.
15. Pham H. *Reliability modeling, analysis and optimization*. Singapore: World Scientific; 2006.

16. Kobbacy KAH, Murthy DNP. Complex system maintenance handbook. Germany: Springer; 2008. eBook.
17. Ross SM. Stochastic processes. New York: John Wiley & Sons; 1996.
18. Osaki S, Nakagawa T. A note on age replacement. *IEEE Trans Reliab* 1975;R-24:92–94.
19. Osaki S. Applied stochastic system modeling. Berlin: Springer; 1992.
20. Ascher HE, Feingold H. Repairable systems reliability. New York: Marcel Dekker; 1984.
21. Balaban HS, Singpurwalla ND. Stochastic properties of a sequence of inter-failure times under minimal repair and under revival. In: Abdel-Hameed MS, Cinlar E, Quinn J, editors. Reliability theory and models. Orlando (FLA): Academic Press; 1984. pp. 65–80.
22. Rigdon SE, Basu AP. The power law process: a model for the reliability of repairable systems. *J Qual Technol* 1989;21: 251–260.
23. Lindley DV. Reconciliation of probability distributions. *Oper Res* 1983;31:866–880.
24. Van Noortwijk JM, Dekker R, Cooke RM, *et al.* Expert judgment in maintenance optimization. *IEEE Trans Reliab* 1992;R-41:427–432.
25. Kyparisis J, Singpurwalla ND. Bayesian inference for the Weibull process with applications to assessing software reliability growth and predicting software failures. In: Billard L, editor. Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface. Atlanta (GA): North-Holland; 1985. pp. 57–64.
26. Guida M, Calabria R, Pulcini G. Bayes inference for a non-homogeneous Poisson process with power intensity law. *IEEE Trans Reliab* 1989;R-38:603–609.
27. Campodonico S, Singpurwalla ND. Inference and predictions from Poisson point processes incorporating expert knowledge. *J Am Stat Assoc* 1995;429:220–226.
28. Baker RD, Scarf PA. Can models fitted to small data samples lead to maintenance policies with near-optimum cost? *IMA J Math Appl Bus Ind* 1995;6:3–12.
29. Holland CW, McLean RA. Applications of replacement theory. *AIIE Trans* 1975;7: 42–47.
30. Aven T. Determination/Estimation of an optimal replacement interval under minimal repair. *Optimization* 1985;16:743–754.
31. Kumar U, Klefsjo B. Reliability analysis of hydraulic systems of LHD machines using the power law process model. *Reliab Eng Syst Saf* 1992;35:217–224.
32. Leung KNF, Cheng LMA. Determining replacement policies for bus engines. *Int J Qual Reliab Manage* 2000;17:771–783.
33. Barlow RE, Hunter LC. Optimum preventive maintenance policies. *Oper Res* 1960;8: 90–100.
34. Nguyen DG, Murthy DNP. Optimal preventive maintenance policies for repairable systems. *Oper Res* 1981;29:1181–1194.
35. Dohi T, Kaio N, Osaki S. Basic preventive maintenance policies and their variations. In: Ben-Daya M, Duffuaa SO, Raouf A, editors. Maintenance, modeling and optimization. Boston (MA): Kluwer Academic Publishers; 2000. Chapter 7. pp. 155–183.
36. Lam Y. A note on the optimal replacement problem. *Adv Appl Probab* 1988;20:479–482.
37. Lam Y. A repair replacement model. *Adv Appl Probab* 1990;22:494–497.
38. Leung KNF. Arithmetic and geometric processes. In: Pham H, editor. Handbook of engineering statistics. Germany: Springer; 2006. Chapter 49. pp. 931–955.
39. Lam Y. The geometric process and its applications. Singapore: World Scientific; 2007.
40. Leung KNF, Zhang YL, Lai KK. A bivariate optimal replacement policy for a cold standby repairable system with repair priority. *Nav Res Logist* 2010;57:149–158.
41. Nakagawa T. Sequential imperfect preventive maintenance policies. *IEEE Trans Reliab* 1988;R-37:295–298.
42. Pham H, Wang HZ. Imperfect maintenance. *Eur J Oper Res* 1996;94:425–438.
43. Lin D, Zuo MJ, Yam CMR. Sequential imperfect preventive maintenance models with two categories of failure modes. *Nav Res Logist* 2001;48:172–183.
44. El-Ferik S, Ben-Daya M. Age-based hybrid model for imperfect preventive maintenance. *IIE Trans* 2006;38:365–375.
45. Zequeira RI, Berenguer C. Periodic imperfect preventive maintenance with two categories of competing failure modes. *Reliab Eng Syst Saf* 2006;91:460–468.
46. Bartholomew-Biggs M, Zuo MJ, Li XH. Modeling and optimizing sequential imperfect preventive maintenance. *Reliab Eng Syst Saf* 2009;94:53–62.
47. Castro IT. A model of imperfect preventive maintenance with dependent failure modes. *Eur J Oper Res* 2009;196:217–224.

48. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B* 1972;34:187–220.
49. Christer AH, Wang WB, Sharp JM. A state space condition monitoring model for furnace erosion prediction and replacement. *Eur J Oper Res* 1997;101:1–14.
50. Makis V, Jardine AKS. Optimal replacement in the proportional hazards model. *INFOR* 1992;30:172–183.
51. Makis V, Jardine AKS. Computation of optimal policies in replacement models. *IMA J Math Appl Bus Ind* 1992;3:169–175.
52. Jardine AKS, Makis V, Banjevic D, *et al.* A decision optimization model for condition-based maintenance. *J Qual Mainten Eng* 1998;4:115–121.
53. Kumar D, Klefsjo B. Proportional hazards model: a review. *Reliab Eng Syst Saf* 1994;44:177–188.
54. Kumar D. Maintenance scheduling using monitored parameter values. In: Ben-Daya M, Duffuaa SO, Raouf A, editors. *Maintenance, modeling and optimization*. Boston (MA): Kluwer Academic Publishers; 2000. Chapter 14. pp. 345–374.
55. Maillart LM. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Trans* 2006;38:463–475.
56. Leung KNF. Inspection schedules when the lifetime distribution of a single-unit system is completely unknown. *Eur J Oper Res* 2001;132:106–115.
57. Munford AG, Shahani AK. A nearly optimal inspection policy. *Oper Res Q* 1972;23:373–379.
58. Munford AG. Comparison among certain inspection policies. *Manage Sci* 1981;27:260–267.
59. Munford AG. Optimal inspection policies where penalty costs are proportional to the duration of the inspection interval containing the failure. *IMA J Math Appl Bus Ind* 1990;2:247–254.
60. Banerjee PK, Chuiv NN. Inspection policies for repairable systems. *IIE Trans* 1996;28:1003–1010.
61. Baker MJC. How often should a machine be inspected? *Int J Qual Reliab Manage* 1990;7:14–18.
62. Ben-Daya M, Hariga M. A maintenance inspection model: optimal and heuristic solutions. *Int J Qual Reliab Manage* 1998;15:481–488.
63. Hariga MA. A maintenance inspection model for a single machine with general failure distribution. *Microelectron Reliab* 1996;36:353–358.
64. Hariga M, Al-Fawzan MA. Discounted models for the single machine inspection problem. In: Ben-Daya M, Duffuaa SO, Raouf A, editors. *Maintenance, modeling and optimization*. Boston (MA): Kluwer Academic Publishers; 2000. Chapter 10. pp. 215–243.
65. Lam Y. An optimal inspection-repair-replacement policy for standby systems. *J Appl Probab* 1995;32:212–223.
66. Lam Y. An inspection-repair-replacement model for a deteriorating system with unobservable state. *J Appl Probab* 2003;40:1031–1042.
67. Lee HL, Rosenblatt MJ. Simultaneous determination of production cycles and inspection schedules in a production system. *Manage Sci* 1987;33:1125–1136.
68. Rahim MA. Joint determination of production quantity, inspection schedule and control chart design. *IIE Trans* 1994;26:2–11.
69. Ben-Daya M. Integrated production, maintenance and quality model for imperfect processes. *IIE Trans* 1999;31:491–501.
70. Taylor HM. Optimal replacement under additive damage and other failure models. *Nav Res Logist Q* 1975;22:1–18.
71. Bergman B. Optimal replacement under a general failure model. *Adv Appl Probab* 1978;10:431–451.
72. Zuckerman D. Replacement models under additive damage. *Nav Res Logist Q* 1977;24:549–558.
73. Christer AH. Developments in delay-time analysis for modeling plant maintenance. *J Oper Res Soc* 1999;50:1120–1137.
74. Christer AH, Waller WM. Delay-time models of industrial inspection maintenance problems. *J Oper Res Soc* 1984;35:401–406.
75. Baker RD, Christer AH. Review of delay-time operational research modeling of engineering aspects of maintenance. *Eur J Oper Res* 1994;73:407–422.
76. Wang WB. Delay time modeling. In: Kobbacy KAH, Murthy DNP, editors. *Complex system maintenance handbook*. Germany: Springer; 2008. Chapter 14. pp. 345–370. eBook.

RISK ASSESSMENTS AND BLACK SWANS

TERJE AVEN

Department of Industrial
Economics, Risk Management,
and Planning, University of
Stavanger, Stavanger, Norway

INTRODUCTION

Risk assessments are based on probabilities and probability models. On the basis of these probabilities and models, assessments and predictions of unknown quantities are made. The probabilities and models are, to a large extent, based on the historical data observed. Such assessments have been used successfully in many types of industries, for example, nuclear and oil and gas. However, in the case of large uncertainties, the quality of these tools has been questioned. Probability models are difficult to justify and the probability estimates are subject to large uncertainties. It is acknowledged that a probability is not a “perfect tool” for measuring uncertainties: The assigned probabilities are conditional on specific background knowledge, which includes assumptions and suppositions. This knowledge could be poor, and the assumptions and suppositions may turn out to be erroneous, the result being that the probabilities assigned lead to poor predictions. Surprises relative to the assigned probabilities may occur; and by just addressing the probabilities, such surprises may be overlooked.

Let us look at an example. Consider the risk, seen through the eyes of a risk analyst in the 1970s, related to future health problems for divers working on offshore petroleum projects. An assignment is to be made for the probability that a diver would experience health problems (properly defined) during the coming 30 years due to the diving activities. Let us assume that an assignment of 1% is

made. This number is based on the available knowledge at that time. There are no strong indications that the divers will experience health problems. However, we know today that these probabilities led to poor predictions. Many divers have experienced severe health problems ([1], p. 7). By restricting risk to the probability assignments alone, we see that aspects of uncertainty and risk are hidden. There is a lack of understanding about the underlying phenomena, but the probability assignments alone are not able to fully describe this status.

Taleb [2] draws a similar conclusion using black swan logic. The inability to predict outliers (black swans) implies the inability to predict the course of history. An outlier lies outside the realm of regular expectations, because nothing in the past can convincingly point to its occurrence. The standard probabilistic methods and models used for analyzing uncertainties are not able to predict black swans.

These standard methods are suitable when there are more relevant data available. To apply the methods, probability models such as the normal distribution and the linear regression model need to be specified. The analysis is based on the assumption of a specific distribution and a specific trend curve (for example linear), which governs the future performance. But it is by no means obvious what the appropriate distribution and model should be. No one knows about the future. The historical data may indicate a specific distribution and a trend, more or less excluding extreme observations, but this does not preclude such observations occurring in the future. Statistical analysis is based on the idea of a huge number of similar situations and if “similar” is limited to the historical data, the population considered could be far too small. However, the statistician needs models to be able to perform the statistical analysis, and he/she will base his/her analysis on the data available. Taleb [2] refers to the world of *mediocristan* and *extremistan* to explain the

difference between the standard probability model context and the more extended population required to reflect surprises that occur in the future, respectively. Without explicitly formulating the thesis, Taleb [2] says that we have to see beyond the historical-based probability models, to be able to adequately reflect future observations.

The standard methods focus on the parameters, for example, $\mu_i = \alpha + i\beta$ in regression analysis. But why be concerned about μ_i , the average value when considering an infinite number of similar activities to the one studied? Risk is about the possible occurrence of surprise, and in particular negative outcomes in the future. An analysis that focuses on averages fails to reveal this in two ways:

1. The key quantity of interest is the outcome Y (e.g., future cost, number of fatalities, etc.), but the standard analysis does not express uncertainties about Y . Instead, uncertainties are expressed about fictional parameters, such as EY interpreted as the average value of Y when considering an infinite number of similar activities to the one studied.
2. The analysis is, to a large extent, based on the historical data and models explaining these, and gives little weight to possible extreme observations. The data may be more or less relevant for the future.

Some would respond to this by stating that using the historical data is the best we can do, and the analysis must always be seen in relation to its assumption. The analysis provides insights, although the assumption may be invalid. An analysis must always be seen in the light of its assumptions, but to fully express risk we can and should do better than restrict attention to historical data and estimate averages.

Many researchers have addressed the problem of using risk assessment in the case of large uncertainties [3–5]. Reid [3] argues that the claims of objectivity in risk assessments are simplistic and unrealistic. Risk estimates are subjective, and there is a common tendency of underestimation of the

uncertainties. According to Stirling [6], using risk assessment when strong knowledge about the probabilities and outcomes does not exist, is irrational, unscientific, and potentially misleading. Renn [5] summarizes the critique drawn from the social sciences over many years and concludes that technical risk analyses represent a narrow framework that should not be the single criterion for risk identification, evaluation, and management.

The answer to this type of critique is, according to O'Brien [7], to look for an alternative to risk assessments. But in our view there is no alternative to risk assessments. The critique of risk assessment is to a large extent justified, but to support the decision making we need to assess risk. The right move forward is not to reject risk assessment, but to improve the tool and its use. This also seems to be the conclusion made by most researchers in the field. The challenge is how decision making on risk can be informed by the best available technical and scientific knowledge [4, p. 100; 8]. To represent and describe this knowledge, judgments about uncertainties are required. Properly conducted risk assessments make these judgments visible and explicit. For the present study we are particularly concerned about the ability of the risk assessments to describe risk contributions from potential surprises (black swans). As argued above, the traditional probabilistic-based analysis has too narrow a risk perspective and should be replaced by a broader uncertainty characterization.

Uncertainty is a frequently occurring topic in risk assessment research, but its scope is often very technical. The literature is characterized by some prevailing ways of thinking on how to deal with uncertainties in risk assessments [9]. We seldom see fundamental discussion about uncertainty in a risk assessment context that extends beyond this thinking. It is acknowledged that risk assessments provide decision support—the decisions are risk-informed and not risk-based [8]—but what are the key aspects to consider outside the realm of the traditional risk assessment results? We argue in this article for the need for assessing

uncertainty factors that are “hidden” in the background knowledge of the assigned probabilities and expected values. The results of the risk assessments must be extended beyond the probabilistic world. In line with this thinking, we must also redefine the concept of risk: uncertainty should replace probability in the definition of risk. The common definitions of risk in the contexts we address in this article are based on the probabilities of events and their consequences [10]. In this article, we describe a framework for such extended risk assessments. But first we give a brief review of the basic features of risk assessment.

BASIC FEATURES ABOUT RISK ASSESSMENTS

In risk assessment we study an activity that may lead to events A with consequences C . Think of A as, for example, the occurrence of a gas leakage or an attack, and C as the consequences of A with respect to human life, the environment, and other attributes that humans value, measured, for example, by the number of fatalities. The consequences may take values in a space S when such a space can be formulated. The occurrence of A and the consequences C are subject to uncertainties U . Probability P is used as a measure of the uncertainties. To make this setup more precise, we need to clarify what we mean by U and P .

First, we need to clarify what we are uncertain about. Is it about aspects of the world expressed by quantities (A, C) ? Or, are the uncertainties related to the uncertainties about the world and the risk description? See Fig. 1. In the latter case, relative

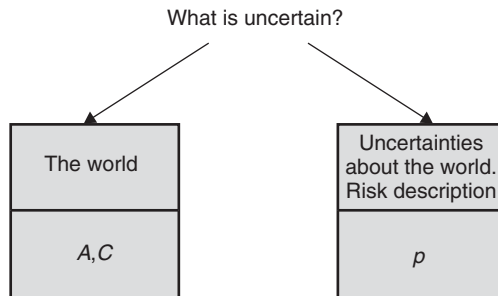
frequency-interpreted probabilities (chances) p are used to describe the risk related to A and C . The probabilities (chances) express variations within mentally constructed infinite populations.

To respond to these questions, we consider two possible objectives of the risk assessments: The purpose of the risk assessment is to (i) accurately estimate the risk, that is, the probabilities (chances) p ; or (ii) to describe the uncertainties about the world (Fig. 2).

The Objective of the Risk Assessment is to Accurately Estimate the Risk (Probabilities)

Suppose that a risk assessment investigates the probability p of an accidental event A . The probability p is defined as the fraction of times the accidental event A occurs if the activity considered were repeated (hypothetically) an infinite number of times. It expresses variations within this mentally constructed infinite population. Since the true value of p is unknown, it must be estimated. The estimation accuracy (precision) can be described in different ways, but the common tool is confidence intervals. If we have a substantial amount of data available and these are considered relevant for the estimation of the statistical parameters p , that is, the observations are considered similar to those of the population generating p , statistical theory shows that the estimation error becomes negligible. However, in a practical risk assessment setting, data are often scarce. If the statistical analysis is based on few data, the estimation accuracy would be poor and the confidence intervals would be wide. Hence, if the goal of the risk assessment is to obtain

Figure 1. Illustration of the issue: what is uncertain in risk assessments? Here A and C are events and their associated consequences, and p frequentist probabilities (chances) linked to A and C .



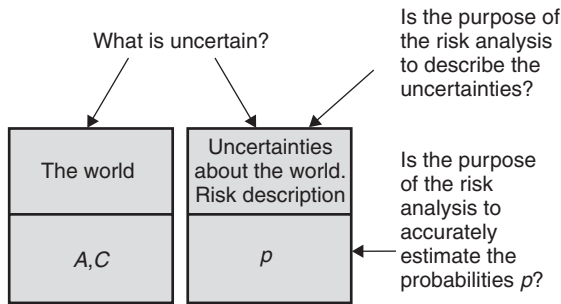


Figure 2. Different aims of the risk assessment.

accurate estimates of some true risk, we can quickly conclude that risk assessment fails. The uncertainties are, in practical cases, often large.

To increase the amount of data, we often extend the relevant population of observations to cover situations that to a varying degree are similar to the one being studied. This reduces the quality, that is, the relevancy of the data, but it would not be possible to describe this aspect by the statistical analysis. If the data are not considered relevant, the statistical analysis cannot be used to check whether the produced risk numbers are accurate compared to the underlying true risk.

Hence, we can focus on risk assessments being a tool to describe uncertainties.

The Objective of the Risk Assessment is to Describe our Uncertainties about the World

For many activities, there are large uncertainties about the world, about the occurrence of events and their consequences, as well as about the underlying factors that cause these events and consequences to occur. The aim of the assessment is to describe these uncertainties. This means that we have to (i) identify what is uncertain and (ii) who is uncertain.

Often, we are able to reduce (i) to a set of quantities, such as the occurrence of events, and the number of fatalities. We denote these events by A and the quantities by C . In most cases, in a risk assessment the answer to (ii) is the analysts. However, in some situations it could be essential to present different uncertainty assessments reflecting the different views among experts and various

stakeholders. Integrating all views into one aggregated figure could camouflage important aspects of the risk picture.

Next, we address the issue on how to express the uncertainties. Probability is the common tool, but a probability has different interpretations.

However, to describe the uncertainties, only one interpretation is suitable: the probabilities must be knowledge-based (subjective) probabilities with reference to a standard expressing the analysts' uncertainty about A and C . Following this interpretation, the assessor compares his/her uncertainty about the occurrence of the event A with the standard of drawing at random a favorable ball from an urn that contains $P(A) \cdot 100\%$ favorable balls.

Relative frequency interpreted probabilities (chances) p are not measures of uncertainty. They exist as proportions of infinite or very large populations of similar units to those considered. They are treated as unknown quantities C .

The knowledge-based probabilities express *epistemic uncertainties*. The variation in the populations of similar units to the one (those) studied that, for example, generates the true value of p , is often referred to as *aleatory (stochastic) uncertainty*. This uncertainty is, however, not an uncertainty for the analysts.

But a knowledge-based (subjective) probability can be given different interpretations: the betting interpretation and the reference to a standard defined above. Among economists and decision analysts, and the earlier probability theorists, a subjective probability is linked to betting: the probability of an event is the price at

which the person assigning the probability is neutral between buying and selling a ticket that is worth one unit of payment if the event occurs, and worthless if not [11].

But this interpretation is not appropriate to describe uncertainties as it extends beyond the realm of uncertainty assessments—it reflects the assessor's attitude to money and the gambling situation, which means that the analysis (evidence) is mixed with values. The scientific basis for risk assessment is founded on the idea that professional analysts describe risk separated from how we (the assessor, the decision maker, or other stakeholders) value the consequences and the risk.

Different probability-based measures are used to describe risk, such as probability distributions, the expected value, the variance, and quantiles. If relevant data are available, these data are used as the basis for the assignments of these measures. Models are often introduced to ease the assignments, in particular, probability models. The probability models coherently and mechanically facilitate the assignments and updating of probabilities. A Bayesian framework is adopted allowing both hard data and expert judgment to be taken into account in the assessment.

But a full risk description needs to see beyond these P measures. All probabilities are conditional on a background knowledge

K , which includes assumptions and suppositions, and in particular, the models used in the assessment. This background knowledge is an integral part of the results of the analysis and all probabilities need to be considered in relation to K . We explain this in more detail in the coming section.

A FRAMEWORK FOR AN EXTENDED RISK AND UNCERTAINTY ASSESSMENT

In this section, we present a framework for risk assessment that highlights uncertainties and in this way the potential for surprises (black swans). The framework is based on Aven [12].

Figure 3 presents the overall risk assessment framework. The key quantities of interest are denoted Z . To assess Z , a model $G(X)$ is introduced, which links a set of input quantities X to Z . Using the model G , an uncertainty description is obtained for Z . The tool used for this purpose could be an analytical approach or Monte Carlo simulation.

The framework is based on the building blocks summarized in the section titled “The Objective of the Risk Assessment is to Describe our Uncertainties about the World.” Hence, all probabilities are knowledge-based (subjective) probabilities with reference to a standard. The framework is based on a

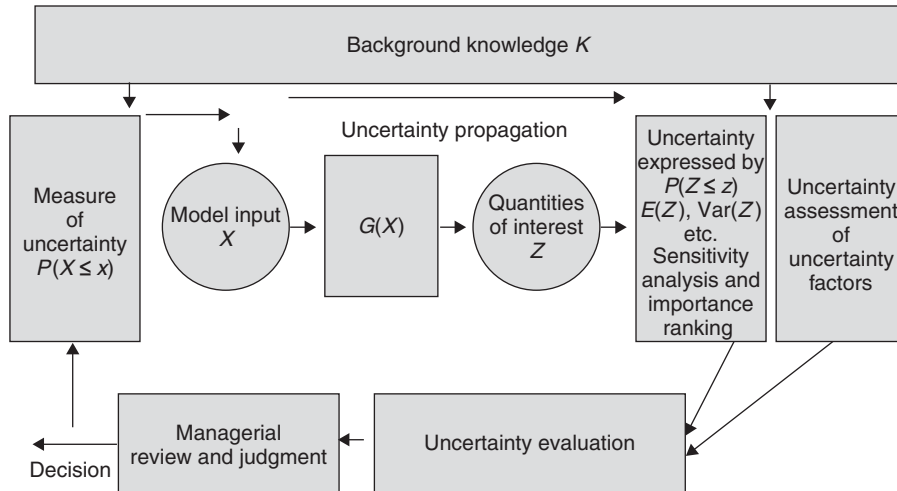


Figure 3. Structure of the suggested framework [12].

risk perspective, where the risk concept captures both consequences and uncertainties [12–14].

In addition to the probability-based measures, the framework requires a separate identification and assessment of uncertainty factors. These factors are generated from the background knowledge K and in particular the assumptions made.

The uncertainty assessments of these factors could, for example, take the following form [12,13]: Each factor's importance is measured using a sensitivity analysis. Is changing the factor important for the risk indices considered, for example, the probability that the loss exceeds a specific number? If this is the case, we next address the uncertainty of this factor. Are there large uncertainties about this factor? If the uncertainties are assessed as large, the factor is given a high risk score. Hence, to obtain a high score in this system, the factor must be judged as important for the risk indices considered and the factor must be subject to large uncertainties. This uncertainty assessment goes beyond the probabilistic analysis.

As an example, think of an assessment related to a process plant that is based on the following assumptions:

- The use of a gas leakage database L that includes data from plants operated under quite different operating conditions than the one considered.
- The use of a new technology for operating a system (S) is not changing the overall performance of the system compared to average industry data.

From this list we identify two uncertainty factors:

1. the number of gas leakages,
2. the performance of the system S .

An assessment of these factors is performed according to the ideas outlined above. The risk indices are sensitive to changes in these factors, and both factors are subject to considerable uncertainties. Special concern is related to the use of the database L , and based on the risk evaluation it was decided to

look closer into this factor and try to reduce the uncertainties and avoid surprises.

In the uncertainty evaluation, a broad uncertainty description is provided, covering probabilities, related background knowledge, and the results from sensitivity analyses, as well as the assessment results for the uncertainty factors. The evaluation provides input to a broader managerial review and judgment, which concludes on the implications of the analysis and balances different concerns (e.g., safety and costs). The result is, for example, an acceptance of the risk and uncertainties related to an activity, the need for design changes, the choice of an alternative, and so on.

It is common in risk assessments to compare the results of the assessments with relevant decision criteria, such as requirements of the form $P < p_0$, where p_0 is a fixed number [1,15]. These requirements could, for example, express that the unreliability level of a system should not exceed a specified level.

Now, given that P is a knowledge-based probability, we cannot base our decision on a direct comparison of the form $P < p_0$. There is a need for a process that extends beyond the probabilistic analysis. The probabilities are dependent on the background knowledge, the assumptions, and suppositions made, including the model G . What is required is a broader process, an uncertainty evaluation and management review and judgment, which sees the results of the assessment in a larger context, taking into account the limitations of the model, the difficulties in specifying probabilities for some quantities, and so on. This process extends beyond the domain of the traditional probability-based uncertainty analysis. The sensitivity analyses constitute an important input to a broad review and judgment process.

DISCUSSION AND CONCLUSIONS

The historical data provides insights about risk. Assuming that the future will be as the history shows, we may obtain good predictions about the future. However, there is in principle, a huge step from the history to risk, as any assumption transforming the

data to the future may be challenged. To fully express risk, we need to look beyond the historical-based data. Care should be shown when using traditional statistical analysis to describe risk, as this analysis is based on strong assumptions. Risk is, to a large extent, about the aspects not included in such analyses, namely surprises. Risk assessments go one step further by identifying hazards/threats/scenarios that could happen, not necessarily based on historical data. Many techniques have been developed to reveal such hazards/threats/scenarios and analyze and describe the associated risk. Probabilities are used to quantify the likelihood and uncertainties of the events and consequences identified. However, these analyses are not particularly suited for revealing surprises, as discussed above. The black swans are often hidden in the background knowledge of the assessments. To highlight these we suggest an approach that extends the traditional probabilistic analysis to include assessment of uncertainty factors. What we suggest is a broad risk assessment of the uncertainty factors, not a detailed quantitative risk assessment but a more qualitative approach. Crude probabilities may be assigned for the uncertainty factors but these do not replace the qualitative judgments. The search for quantitative, explicit approaches for expressing the uncertainties, even beyond the knowledge-based (subjective) probabilities, may seem to be a possible way forward. However, such an approach is rejected. Trying to be precise and accurately expressing what is extremely uncertain does not make sense. Instead, we recommend a more open qualitative approach for revealing such uncertainties, as described above.

Such an approach would not be able to guarantee that all surprises would be identified. However, the approach provides a framework for highlighting uncertainties. By seeing beyond the probabilities, an extended search basis is established. Some decision makers, managers, and politicians may find this approach unattractive as it means that the decision basis gives less clear answers. The decision maker has to weigh

the uncertainties and balance them against other concerns, such as profit. For some this would be hard. They expect science to be able to produce clear answers. But science cannot eliminate the uncertainties. The challenge is, therefore, to reveal the uncertainties to obtain an improved decision basis. The presented framework provides guidance on how to obtain such a basis.

Acknowledgments

The author is grateful to two anonymous reviewers for valuable comments and suggestions to an earlier version of this article.

REFERENCES

1. Aven T, Vinnem JE. Risk management, with applications from the offshore oil and gas industry. New York: Springer; 2007.
2. Taleb NN. The black swan: the impact of the highly improbable. London: Penguin; 2007.
3. Reid SG. Acceptable risk. In: Blockley DI, editor. Engineering safety. New York: McGraw-Hill; 1992. pp. 138–166.
4. Stirling A. Risk at a turning point? *J Risk Res* 1998;1:97–109.
5. Renn O. Three decades of risk research: accomplishments and new challenges. *J Risk Res* 1998;1(1):49–71.
6. Stirling A. Science, precaution and risk assessment: towards more measured and constructive policy debate. *Eur Mol Biol Organ Rep* 2007;8:309–315.
7. O'Brien M. Making better environmental decisions. Cambridge (MA): The MIT Press; 2000.
8. Apostolakis GE. How useful is quantitative risk assessment? *Risk Anal* 2004;24: 515–520.
9. Paté-Cornell ME. Uncertainties in risk analysis: six levels of treatment. *Reliab Eng Syst Saf* 1996;54(2–3):95–111.
10. Bedford T, Cooke R. Probabilistic risk analysis. Cambridge (MA): Cambridge University Press; 2001.
11. Singpurwalla N. Reliability and risk. A Bayesian perspective. Chichester: Wiley; 2006.
12. Aven T. A new scientific framework for quantitative risk assessments. *Int J Bus Continuity Risk Manage* 2009;1(1):67–77.

13. Aven T. Risk analysis. Chichester: Wiley; 2008.
14. Aven T, Renn O. On risk defined as an event where the outcome is uncertain. *J Risk Res* 2009;12:1–11.
15. de Rocquigny E, Devictor, N, Tarantola S, editors. Uncertainty in industrial practice. a guide to quantitative uncertainty management. New Jersey: Wiley; 2008.

RISK AVERSE MODELS

ALEXANDER SHAPIRO
School of Industrial and Systems
Engineering, Georgia
Institute of Technology,
Atlanta, Georgia

PORTFOLIO SELECTION PROBLEM

Let us consider the following portfolio selection problem. Suppose that we want to invest capital W_0 in n assets, by investing an amount x_i in asset i , for $i = 1, \dots, n$. Suppose, further, that each asset has a respective return rate R_i . We address now the question of how to distribute our wealth W_0 in an optimal way. The total wealth resulting from our investment, after one period of time, is $W_1 = W_0 + R^\top x$, where $R = (R_1, \dots, R_n)^\top$ and $x = (x_1, \dots, x_n)^\top$ are the respective (column) vectors and $R^\top x = \sum_{i=1}^n R_i x_i$ is their scalar product (for the sake of simplicity we do not take into account here the transaction costs). The problem is that at the time we have to make the investment decision, the return rates R_i are *unknown*.

Suppose that we can model the unknown returns R_i as random variables, and hence to consider their expected values $\bar{r}_i := \mathbb{E}[R_i]$, $i = 1, \dots, n$ (the notation “ $:=$ ” means equal by definition). Then we can try to maximize the expected return on an investment subject to the budget constraint $\sum_{i=1}^n x_i = W_0$. This leads to the following optimization problem

$$\text{Max}_{x \geq 0} \mathbb{E}[W_1] \text{ subject to } \sum_{i=1}^n x_i = W_0. \quad (1)$$

Since $\mathbb{E}[R^\top x] = \sum_{i=1}^n \bar{r}_i x_i$, Problem (1) has a simple optimal solution of investing everything into an asset with the largest expected return rate $\max_{1 \leq i \leq n} \bar{r}_i$. Of course, from the practical point of view, such a solution of

putting everything into one asset is too risky.

There are several ways how one can try to control the risk by making a more even distribution of the wealth among n assets. One approach is to maximize expected utility of the wealth represented by a concave nondecreasing function $U(\cdot)$. This leads to the problem of maximizing $\mathbb{E}[U(W_1)]$ subject to the “no borrowing” constraint $x \geq 0$ and the budget constraint. This approach requires specification of the utility function $U(\cdot)$, which could be quite arbitrary and not very intuitive. It could also be noted that calculation of the expectation $\mathbb{E}[U(W_1)]$ requires specification (knowledge) of the probability distribution of the random vector R , while $\mathbb{E}[W_1]$ involves only the first-order moments of the random returns.

Another possible approach is to maximize the expected return while controlling the involved risk of the investment. There are several ways how the concept of risk could be formalized. For instance, we can evaluate risk by variability of W_1 measured by its *variance* $\text{Var}[W_1]$. Since W_1 is a linear function of the random variables R_i , we have that $\text{Var}[W_1] = x^\top \Sigma x$, where Σ is the covariance matrix of the random vector R . This leads to the optimization problem of maximizing the expected return subject to the additional constraint $\text{Var}[W_1] \leq \nu$, where $\nu > 0$ is a specified constant. This problem can be written as

$$\begin{aligned} & \text{Max}_{x \geq 0} \sum_{i=1}^n \bar{r}_i x_i, \\ & \text{subject to } \sum_{i=1}^n x_i = W_0, \quad x^\top \Sigma x \leq \nu. \end{aligned} \quad (2)$$

Since the covariance matrix Σ is positive semidefinite, the function $x \mapsto x^\top \Sigma x$ is convex quadratic, and hence Equation (2) is a convex problem.

By duality arguments there exists Lagrange multiplier $\lambda \geq 0$ such that Problem

(2) is equivalent to the problem

$$\begin{aligned} & \text{Max}_{x \geq 0} \sum_{i=1}^n \bar{r}_i x_i - \lambda x^\top \Sigma x, \\ & \text{subject to } \sum_{i=1}^n x_i = W_0. \end{aligned} \quad (3)$$

We can view the objective function of the above Problem (3) as a compromise between the expected return and its variability measured by its variance. This approach was suggested by Markowitz [1], more than 50 years ago, and is still routinely used in portfolio selections.

Let us make the following observations. Problem (3) is a (convex) quadratic programming problem for which efficient numerical algorithms are available (such algorithms were already available in 1952, the time of publication of Markowitz [1] paper). Problem (2) is a convex conic programming problem which also can be efficiently solved by modern interior point algorithms. Both Problems (2) and (3) do not require specification of the whole probability distribution of the random vector R , only its first- and second-order moments. It is also worthwhile to mention that Problems (2) and (3) depart from Problem (1) in that these problems utilize possible correlations between random returns of different assets.

Let us point now to the following conceptual deficiencies of Markowitz's approach. The variance, as a measure of risk, penalizes deviations from the average (expected value) in both directions—positive and negative. However, we are not really concerned here if a random realization of the return is bigger than the average; we do not want it to be significantly smaller. The other deficiency is that the function $\rho(W) := \mathbb{E}[W] - \lambda \text{Var}[W]$, of random variable W , is not necessarily monotone for $\lambda > 0$. That is, it can happen that for any possible realization of two random variables W and W' (defined on the same probability space), variable W is always bigger than W' while $\rho(W)$ is smaller than $\rho(W')$. This could lead to a paradoxical situation when for two different decision vectors x and x' any realization of the random return $R^\top x$

is bigger than $R^\top x'$, and yet we prefer x' as a solution of Problem (3).

We can also approach risk control by imposing chance constraints. Consider the problem

$$\begin{aligned} & \text{Max}_{x \geq 0} \sum_{i=1}^n \bar{r}_i x_i, \\ & \text{subject to } \sum_{i=1}^n x_i = W_0, \text{ Prob}\{R^\top x \leq b\} \leq \alpha. \end{aligned} \quad (4)$$

That is, we impose the additional constraint that probability of the return falling below a chosen constant b is small, that is, is no more than the specified level $\alpha \in (0, 1)$. Such constraints are called *chance* or *probabilistic* constraints.

Suppose, for the moment, that vector R of random returns has a multivariate normal distribution with mean vector $\bar{r}$ and covariance matrix Σ . Then $R^\top x$ has normal distribution with mean $\bar{r}^\top x$ and variance $x^\top \Sigma x$. Consequently, if $x^\top \Sigma x \neq 0$, then

$$\text{Prob}\{R^\top x \leq b\} = \Phi\left(\frac{b - \bar{r}^\top x}{\sqrt{x^\top \Sigma x}}\right), \quad (5)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Therefore, we can write the chance constraint of Problem (4) in the form

$$b - \bar{r}^\top x + z_\alpha \sqrt{x^\top \Sigma x} \leq 0, \quad (6)$$

where $z_\alpha := \Phi^{-1}(1 - \alpha)$ is the $(1 - \alpha)$ -quantile of the standard normal distribution (for example, $z_{0.025} = 1.96$). The function $x \mapsto \sqrt{x^\top \Sigma x}$ is convex, and if $\alpha \leq 0.5$, then $z_\alpha \geq 0$, and hence the Constraint (6) is convex.

The Constraint (6) can be interpreted as that the ratio $\frac{\bar{r}^\top x - b}{\sqrt{x^\top \Sigma x}}$ of the expected return minus b to the standard deviation of the return, should be not smaller than the (positive) constant z_α . It was possible to write the chance constraint of the Problem (4) in the closed form (Eq. 6) because of the assumption that distribution of R is normal. In general, it

could be quite difficult to handle such chance constraints numerically.

COHERENT RISK MEASURES

It was argued in the previous section that in some situations the approach of optimizing the random objective on average could be too risky. This raises the question of what could be a good measure of risk of a random variable. In order to formalize this problem we proceed as follows. Consider a sample space $(\Omega, \mathcal{F})$, equipped with sigma algebra $\mathcal{F}$, and a set (space) $\mathcal{Z}$ of $\mathcal{F}$ -measurable functions $Z : \Omega \rightarrow \mathbb{R}$. For instance, if the set Ω is finite, we can equip it with sigma algebra of all its subsets. In that case any real valued function $Z(\omega)$, $\omega \in \Omega$, is $\mathcal{F}$ -measurable and we can take $\mathcal{Z}$ to be the space of all functions $Z : \Omega \rightarrow \mathbb{R}$. In general, the space of all $\mathcal{F}$ -measurable functions could be too large for a meaningful development of the theory, we will discuss this later.

With random variable $Z \in \mathcal{Z}$ we associate a number $\rho(Z)$ called its *risk measure*. That is, risk measure is a mapping $\rho : \mathcal{Z} \rightarrow \mathbb{R}$ from the space $\mathcal{Z}$ into the real line. It is also possible to consider risk measures which can take values $+\infty$ or $-\infty$ as well. We avoid this by making a suitable choice of the space $\mathcal{Z}$ and assume that $\rho(Z)$ is real valued. An example of risk measure is

$$\rho(Z) := \mathbb{E}[Z] + \lambda \text{Var}[Z]. \quad (7)$$

This risk measure was used in the formulation (Eq. 3) of the portfolio selection problem. For the sake of notational convenience, we deal from now on with minimization, rather than maximization, formulations of optimization problems, therefore the constant λ in Equation (7) is assumed to be non-negative.

Few remarks are now in order. The term *risk measure* is somewhat misleading. It could be more appropriate to call it “mean-risk measure,” since it typically involves a compromise between minimizing the mean (expectation) and controlling variability of the objective function. Some authors use the term *acceptability functionals* [2]. Anyway,

the terminology of *risk measures* became standard, so we will follow it here.

Another question is what space $\mathcal{Z}$ to use. In order for the expectation $\mathbb{E}[Z]$ and variance $\text{Var}[Z]$ to make sense, we need to equip the sample space $(\Omega, \mathcal{F})$ with a probability measure P . Then we can consider the space $\mathcal{Z} := L_2(\Omega, \mathcal{F}, P)$ of all random variables Z with finite second-order moments. In that case, the risk measure $\rho(Z)$ defined in Equation (7), is finite valued for all $Z \in \mathcal{Z}$. In the subsequent analysis, we assume that $\mathcal{Z} := L_p(\Omega, \mathcal{F}, P)$ with $p \in [1, +\infty)$, that is, $\mathcal{Z}$ is the space of random variables, defined on the probability space $(\Omega, \mathcal{F}, P)$, with finite p th order moments. Note that $\mathcal{Z} = L_p(\Omega, \mathcal{F}, P)$ is a linear space, and equipped with the norm $\|Z\|_p := (\int_{\Omega} |Z(\omega)|^p dP(\omega))^{1/p}$ becomes a Banach space.

We refer to P as the *reference* probability measure. If $Z, Z' \in \mathcal{Z}$ and $Z(\omega) \geq Z'(\omega)$ for P -almost every $\omega \in \Omega$, we write this as $Z \succeq Z'$. Unless stated otherwise, we assume that the expectations, such as $\mathbb{E}[Z]$, are taken with respect to the reference probability measure P . In case of doubt, we write $\mathbb{E}_P[Z]$ to emphasize what probability measure is employed.

It was suggested in Ref. 3 that a “good” risk measure should satisfy the following axioms, and such risk measures were called *coherent*.

- (A1) *Monotonicity*: If $Z, Z' \in \mathcal{Z}$ and $Z \succeq Z'$, then $\rho(Z) \geq \rho(Z')$.
- (A2) *Convexity*:

$$\rho(tZ + (1-t)Z') \leq t\rho(Z) + (1-t)\rho(Z')$$

for all $Z, Z' \in \mathcal{Z}$ and all $t \in [0, 1]$.

- (A3) *Translation Equivariance*: If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$.
- (A4) *Positive Homogeneity*: If $t > 0$ and $Z \in \mathcal{Z}$, then $\rho(tZ) = t\rho(Z)$.

We already argued in the previous section that the monotonicity property (axiom A1) is a natural condition that a risk measure should satisfy. Convexity property is also a natural one. Because of (A4) the convexity property (A2) holds iff $\rho(Z + Z') \leq \rho(Z) + \rho(Z')$ for all $Z, Z' \in \mathcal{Z}$. That is, risk of the sum

of two random variables is not bigger than the sum of risks. Axioms (A3) and (A4) postulate position and scale properties, respectively, of risk measures.

The risk measure ρ defined in Equation (7) with $\lambda \geq 0$, satisfies axioms (A2) and (A3), but not (A1) and (A4) unless $\lambda = 0$. If the variance $\text{Var}[Z]$ in the definition of that risk measure is replaced by the standard deviation $\sqrt{\text{Var}[Z]}$, then the property (A4) will also hold, but not the monotonicity property (A1). Another example of risk measure, popular in financial engineering, is the so-called value-at-risk measure

$$\begin{aligned} \text{V@R}_\alpha(Z) &:= \inf \{t : \text{Prob}(Z \leq t) \geq 1 - \alpha\} \\ &= F^{-1}(1 - \alpha), \end{aligned} \quad (8)$$

where $F(t) := \text{Prob}(Z \leq t)$ is the cumulative distribution function of Z . That is, $\text{V@R}_\alpha(Z)$ is the left side $(1 - \alpha)$ -quantile of the distribution of Z . This risk measure satisfies axioms (A1), (A3), and (A4), but not (A2). We will discuss examples of coherent risk measures later.

We have the following basic duality result associated with coherent risk measures. With each space $\mathcal{Z} := L_p(\Omega, \mathcal{F}, P)$, $p \in [1, +\infty)$, is associated its dual space $\mathcal{Z}^* := L_q(\Omega, \mathcal{F}, P)$, where $q \in (1, +\infty]$ is such that $1/p + 1/q = 1$. For $Z \in \mathcal{Z}$ and $\zeta \in \mathcal{Z}^*$ their scalar product is defined as

$$\langle Z, \zeta \rangle := \int_{\Omega} Z(\omega) \zeta(\omega) dP(\omega). \quad (9)$$

We denote by

$$\mathfrak{P} := \left\{ \zeta \in \mathcal{Z}^* : \int_{\Omega} \zeta(\omega) dP(\omega) = 1, \zeta \geq 0 \right\} \quad (10)$$

the set of probability density functions in the dual space $\mathcal{Z}^*$.

Theorem 1. *Let $\mathcal{Z} := L_p(\Omega, \mathcal{F}, P)$ and $\rho : \mathcal{Z} \rightarrow \mathbb{R}$ be a coherent risk measure. Then ρ is continuous (in the norm topology of $\mathcal{Z}$) and there exists a bounded convex set $\mathfrak{A} \subset \mathfrak{P}$ such that*

$$\rho(Z) = \sup_{\zeta \in \mathfrak{A}} \langle Z, \zeta \rangle, \quad \forall Z \in \mathcal{Z}. \quad (11)$$

Conversely if the representation (Eq. 11) holds for some nonempty bounded convex set $\mathfrak{A} \subset \mathfrak{P}$, then ρ is a (real valued) coherent risk measure.

The dual representation (Eq. 11) follows from the classical Fenchel–Moreau Theorem [4]; in a certain form it was derived directly in Ref. 3. The present formulation is taken from Ref. 5. Note that for $\zeta \in \mathfrak{P}$ the scalar product $\langle Z, \zeta \rangle$ can be understood as the expectation $\mathbb{E}_Q[Z]$ taken with respect to the probability measure $dQ = \zeta dP$. Therefore, the representation (Eq. 11) can be written as

$$\rho(Z) = \sup_{Q \in \mathfrak{Q}} \mathbb{E}_Q[Z], \quad \forall Z \in \mathcal{Z}, \quad (12)$$

where $\mathfrak{Q} := \{Q : dQ = \zeta dP, \zeta \in \mathfrak{A}\}$.

Let us consider some examples. The following risk measure is called the *mean upper-semideviation risk measure* of order $p \in [1, +\infty)$:

$$\rho(Z) := \mathbb{E}[Z] + \lambda \left(\mathbb{E} \left[[Z - \mathbb{E}[Z]]_+^p \right] \right)^{1/p}, \quad (13)$$

where we use notation $[a]_+ := \max\{0, a\}$. In the second term of the right-hand side of Equation (13), the excess of Z over its expectation is penalized (recall that we deal here with minimization formulation of optimization problems). In order for this risk measure to be real valued it is natural to take $\mathcal{Z} := L_p(\Omega, \mathcal{F}, P)$. It is possible to show [6, Section 6.3.2] that for any $\lambda \in [0, 1]$ this risk measure is coherent and has the dual representation (Eq. 11) with the set

$$\begin{aligned} \mathfrak{A} &= \{ \zeta' \in \mathcal{Z}^* : \zeta' = 1 + \zeta - \mathbb{E}[\zeta], \\ &\quad \|\zeta\|_q \leq \lambda, \zeta \geq 0 \}. \end{aligned} \quad (14)$$

Another example of coherent risk measure is the conditional value-at-risk measure [7]

$$\begin{aligned} \text{CV@R}_\alpha(Z) &:= \inf_{t \in \mathbb{R}} \left\{ t + \alpha^{-1} \mathbb{E}[Z - t]_+ \right\}, \\ \alpha &\in (0, 1). \end{aligned} \quad (15)$$

It is natural to take here $\mathcal{Z} := L_1(\Omega, \mathcal{F}, P)$. This risk measure is also known under the names average value-at-risk, expected short-fall, and expected tail loss; these names are

motivated by Formulas (16) and (17) below. It is possible to show that the set of minimizers of the right-hand side of Equation (15) is formed by $(1 - \alpha)$ -quantiles of the distribution of Z , in particular, $t^* = V@R_\alpha(Z)$ is such a minimizer, and hence $CV@R_\alpha(Z) \geq V@R_\alpha(Z)$.

The conditional value-at-risk can be also written as

$$CV@R_\alpha(Z) = \frac{1}{\alpha} \int_0^\alpha V@R_\tau(Z) d\tau, \quad (16)$$

and, moreover, if the cumulative distribution function $F(t)$ of Z is continuous at $t^* = V@R_\alpha(Z)$, then

$$\begin{aligned} CV@R_\alpha(Z) &= \frac{1}{\alpha} \int_{V@R_\alpha(Z)}^{+\infty} z dF(z) \\ &= \mathbb{E}[Z | Z \geq V@R_\alpha(Z)]. \end{aligned} \quad (17)$$

The dual representation (Eq. 11) for $\rho(Z) = CV@R_\alpha(Z)$ holds with the set

$$\begin{aligned} \mathfrak{A} = \left\{ \zeta \in L_\infty(\Omega, \mathcal{F}, P) : \right. \\ \left. \zeta(\omega) \in [0, \alpha^{-1}] \text{ a.e. } \omega \in \Omega, \mathbb{E}[\zeta] = 1 \right\}. \end{aligned} \quad (18)$$

The conditional value-at-risk measures appear naturally in convex approximations of chance-constrained problems [7], and in a sense give a best conservative convex approximation of chance-constrained problems [8]. For a thorough discussion of coherent risk measures and extensions the interested reader is referred to Föllmer and Schied [9].

RISK AVERSE OPTIMIZATION

Consider now a generic problem of optimizing (minimizing) random objective function $F(x, \omega)$ over decision vector $x \in \mathcal{X}$. Here $\mathcal{X} \subset \mathbb{R}^n$ and $\omega \in \Omega$ with $(\Omega, \mathcal{F}, P)$ being a probability space. In a risk averse approach, the corresponding optimization problem is written in the form

$$\text{Min}_{x \in \mathcal{X}} \rho[F(x, \omega)], \quad (19)$$

where $\rho : \mathcal{Z} \rightarrow \mathbb{R}$ is an appropriate risk measure. For $x \in \mathcal{X}$, the quantity $\rho[F(x, \omega)]$ is

understood as the risk measure of the random variable $F_x : \Omega \rightarrow \mathbb{R}$, where $F_x(\omega) = F(x, \omega)$. It is assumed, of course, that for every $x \in \mathcal{X}$, the random variable F_x belongs to the space $\mathcal{Z}$ associated with the risk measure ρ .

If the risk measure ρ is coherent, then by using representation (Eq. 12) we can write Problem (19) in the following minimax form

$$\text{Min}_{x \in \mathcal{X}} \sup_{Q \in \Omega} \mathbb{E}_Q[F(x, \omega)], \quad (20)$$

where Ω is the corresponding family of probability distributions. That is, Problem (19) can be considered as a minimax stochastic programming problem with the uncertainty set Ω of probability distributions. Minimax stochastic problems of the form (Eq. 20) were also studied directly in the stochastic programming literature [10,11].

Consider, for example, the following convex combination of the expectation and $CV@R_\alpha$ risk measures

$$\rho(Z) := (1 - \gamma)\mathbb{E}[Z] + \gamma CV@R_\alpha(Z), \quad \gamma \in [0, 1].$$

In that case, by using Equation (15), Problem (19) can be written as

$$\text{Min}_{(x,t) \in \mathcal{X} \times \mathbb{R}} \mathbb{E}[H(x, t, \omega)], \quad (21)$$

where $H(x, t, \omega) := (1 - \gamma)F(x, \omega) + \gamma\alpha^{-1}[F(x, \omega) - t]_+ + \gamma t$. That is, the corresponding risk averse problem is reduced to the expected value minimization with one additional variable. Note that if $F(x, \omega)$ is (piecewise linear) convex in x , then $H(x, t, \omega)$ is (piecewise linear) convex in (x, t) . Also, the dual (minimax) representation (Eq. 20) holds with Ω being the family of probability measures Q such that

$$1 - \gamma \leq \frac{dQ}{dP} \leq 1 + \gamma(\alpha^{-1} - 1).$$

An important property of coherent risk measures is that they preserve convexity of the function $F(\cdot, \omega)$. That is, properties (A1) and (A2) of a coherent risk measure ρ imply that if $F(x, \omega)$ is convex in x for a.e. $\omega \in \Omega$, then the function $\phi(x) := \rho[F(x, \omega)]$ is also convex.

Another important property of coherent risk measures is that for them the following interchangeability principle holds [6], [12, Section 6.4.1]. Suppose that

$$F(x, \omega) := \inf_{y \in \mathcal{Y}(x, \omega)} G(x, y, \omega), \quad (22)$$

where $G: \mathbb{R}^n \times \mathbb{R}^m \times \Omega \rightarrow \mathbb{R}$ and $\mathcal{Y}: \mathbb{R}^n \times \Omega \rightrightarrows \mathbb{R}^m$ is a multifunction mapping (x, ω) into a subset of $\mathbb{R}^m$. The right-hand side of Equation (22) can be viewed as the optimal value of the second-stage problem of minimization of $G(x, y, \omega)$ subject to the feasibility constraint $y \in \mathcal{Y}(x, \omega)$. For example, in the linear case

$$\begin{aligned} G(x, y, \omega) &:= c^\top x + q(\omega)^\top y, \\ \mathcal{Y}(x, \omega) &:= \{y : T(\omega)x + W(\omega)y = h(\omega), y \geq 0\}. \end{aligned} \quad (23)$$

For coherent risk measures ρ we have that

$$\rho[F(x, \omega)] = \inf_{y(\cdot) \in \mathcal{Y}(x, \cdot)} \rho[G(x, y(\omega), \omega)], \quad (24)$$

where now $y(\cdot)$ is an element of an appropriate functional space and the notation $y(\cdot) \in \mathcal{Y}(x, \cdot)$ means that $y(\omega) \in \mathcal{Y}(x, \omega)$ for a.e. $\omega \in \Omega$. Consequently, the first-stage Problem (19) can be written as one large optimization problem:

$$\text{Min}_{x \in \mathcal{X}, y(\cdot) \in \mathcal{Y}(x, \cdot)} \rho[G(x, y(\omega), \omega)]. \quad (25)$$

In particular, suppose that the set Ω is finite, say $\Omega = \{\omega_1, \dots, \omega_K\}$. Then Problem (25) takes the form

$$\begin{aligned} &\text{Min}_{x, y_1, \dots, y_K} \rho[(G(x, y_1, \omega_1), \dots, G(x, y_K, \omega_K))] \\ &\text{subject to } x \in \mathcal{X}, y_k \in \mathcal{Y}(x, \omega_k), k = 1, \dots, K, \end{aligned} \quad (26)$$

where $y_k = y(\omega_k)$, $k = 1, \dots, K$. In the linear case (Eq. 23), Problem (26) becomes

$$\begin{aligned} &\text{Min}_{x, y_1, \dots, y_K} c^\top x + \rho[(q_1^\top y_1, \dots, q_K^\top y_K)] \\ &\text{subject to } x \in \mathcal{X}, T_k x + W_k y_k = h_k, \\ &\quad y_k \geq 0, k = 1, \dots, K. \end{aligned} \quad (27)$$

If $\rho(Z) := \mathbb{E}[Z]$ and hence $\rho[(q_1^\top y_1, \dots, q_K^\top y_K)] = \sum_{k=1}^K p_k q_k^\top y_k$, where $p_1, \dots, p_K$ are the respective probabilities, then Equation (27) becomes a large-scale formulation of a linear two-stage stochastic programming problem with recourse.

It could be emphasized that the monotonicity condition A1 is essential for the above interchangeability property and preservation of convexity of $F(\cdot, \omega)$. For instance, the risk measure defined in Equation (7) does not preserve convexity and does not allow the equivalent formulation (Eq. 27) for two-stage linear programming with recourse [13].

MULTISTAGE RISK AVERSE OPTIMIZATION

It is not completely obvious how to extend the above approach of risk measures to a multistage setting where decisions are made at several periods of time $t = 1, \dots, T$, based on information available at the time of the decision. In that respect several approaches were suggested. It is convenient to represent the evolution of the underlying data process by a scenario tree. Such a scenario tree is constructed by enumerating possible successive states (children nodes) of a current state of the system represented by its node at stage (time period) $t = 1, \dots, T-1$. A scenario is a path along this tree starting at the (unique) root node at stage $t = 1$ and ending at the last stage $t = T$.

The basic idea of multistage stochastic programming is that if we are currently at a state of the system at stage t , represented by a node of the scenario tree, then our decision at that node is based on our knowledge about the next possible realizations of the data process, which are represented by its children nodes at stage $t+1$. In the risk neutral approach one optimizes the corresponding conditional expectation of the objective function. This allows to write the associated dynamic programming equations.

This idea can be extended to a nested formulation of multistage risk averse optimization. That is, with every node at stage

$t = 1, \dots, T - 1$, is associated a risk measure defined on the space of its children nodes. Formally this can be described by the so-called *conditional risk mappings* $\rho_{t+1|\xi_{[t]}} : \mathcal{Z}_{t+1} \rightarrow \mathcal{Z}_t$, where $\xi_1, \dots, \xi_T$ is the random data process and $\xi_{[t]} := (\xi_1, \dots, \xi_t)$ is the history of the process [12,14,15]. For example, to the mean upper-semideviation risk measure (Eq. 13), corresponds the following conditional risk mapping

$$\rho_{t+1|\xi_{[t]}}(Z) := \mathbb{E}_{|t}[Z] + \lambda \left(\mathbb{E}_{|t} \left[[Z - \mathbb{E}_{|t}[Z]]_+^p \right] \right)^{1/p}, \quad (28)$$

where $\mathbb{E}_{|t}[\cdot] := \mathbb{E}[\cdot | \xi_{[t]}]$ denotes the respective conditional expectation.

The corresponding multistage risk averse optimization is formulated in accordance with the conditional structure of the constructed risk measures:

$$\begin{aligned} \text{Min } & f_1(x_1) + \rho_{2|\xi_{[1]}}[f_2(x_2, \xi_2) + \dots \\ & + \rho_{T|\xi_{[T-1]}}[f_T(x_T, \xi_T)]] \\ \text{s.t. } & x_1 \in \mathcal{X}_1, x_t \in \mathcal{X}_t(x_{t-1}, \xi_t), t = 2, \dots, T, \end{aligned} \quad (29)$$

where $f_t(x_t, \xi_t)$ are cost functions and $\mathcal{X}_t(x_{t-1}, \xi_t)$ are feasibility sets. Note that the optimization in Equation (29) is performed over feasible *policies* $x_t = x_t(\xi_{[t]})$ and the constraints should hold for almost every realization of the random process. If we set $\rho_{t+1|\xi_{[t]}}(\cdot) := \mathbb{E}[\cdot | \xi_{[t]}]$, then Equation (29) becomes a nested formulation risk neutral multistage stochastic programming problem.

The dynamic programming equations for the nested formulation (Eq. 29) can be written similar to the risk neutral case as follows

$$\begin{aligned} Q_t(x_{t-1}, \xi_{[t]}) \\ = \inf_{x_t \in \mathcal{X}_t(x_{t-1}, \xi_t)} \left\{ f_t(x_t, \xi_t) + Q_{t+1}(x_t, \xi_{[t]}) \right\}, \end{aligned} \quad (30)$$

where

$$Q_{t+1}(x_t, \xi_{[t]}) = \rho_{t+1|\xi_{[t]}}[Q_{t+1}(x_t, \xi_{[t+1]})]. \quad (31)$$

For a formal development of these ideas we refer to Refs 6 and 12, Section 6.7 of the latter.

An alternative approach is based on considering multiperiod risk measures of the form $\rho : \mathcal{Z}_1 \times \dots \times \mathcal{Z}_T \rightarrow \mathbb{R}$, where $\mathcal{Z}_t = L_p(\Omega, \mathcal{F}_t, P)$, $t = 1, \dots, T$, with $\mathcal{F}_1 \subset \dots \subset \mathcal{F}_T$ being a filtration (sequence of sigma algebras) representing information state of the system. A particular class of such multiperiod risk measures, convenient for a linear programming formulation, was suggested in Ref. 16 under the name *polyhedral risk measures*. For a further discussion of that approach we can refer to Ref. 2.

Acknowledgments

Research of this author was partly supported by the NSF award DMS-0914785.

REFERENCES

1. Markowitz HM. Portfolio selection. J Finance 1952;7:77–91.
2. Pflug GCh, Römisch W. Modeling, measuring and managing risk. Singapore: World Scientific; 2007.
3. Artzner P, Delbaen F, Eber J-M, *et al.* Coherent measures of risk. Math Finance 1999;9: 203–228.
4. Rockafellar RT. Conjugate duality and optimization: regional conference series in applied mathematics. Philadelphia: SIAM; 1974.
5. Ruszczyński A, Shapiro A. Optimization of convex risk functions. Math Oper Res 2006;31:433–452.
6. Shapiro A, Dentcheva D, Ruszczyński A. Lectures on stochastic programming: modeling and theory. Philadelphia (PA): SIAM; 2009.
7. Rockafellar RT, Uryasev S. Optimization of conditional value at risk. J Risk 2000;2: 21–41.
8. Nemirovski A, Shapiro A. Convex approximations of chance-constrained programs. SIAM J Optim 2006;17:969–996.
9. Föllmer H, Schield A. Stochastic finance: an introduction in discrete time, de gruyter studies in mathematics. 2nd ed. Berlin: Walter de Gruyter; 2004.
10. Dupačová J. The minimax approach to stochastic programming and an illustrative application. Stochastics 1987;20:73–88.

11. Žáčková J. On minimax solutions of stochastic linear programming problems. Čas Pěst Mat 1966;91:423–430.
12. Ruszczyński A, Shapiro A. Conditional risk mappings. Math Oper Res 2006;31:544–561.
13. Takriti S, Ahmed S. On robust optimization of two-stage systems. Math Program 2004;99:109–126.
14. Frittelli M, Rosazza Gianin E. Dynamic convex risk measures. In: Szegö G, editor. Risk measures for the 21st century. Chichester: Wiley; 2004. pp. 227–248.
15. Riedel F. Dynamic coherent risk measures. Stoch Processes Appl 2004;112:185–200.
16. Eichhorn A, Römisch W. Polyhedral risk measures in stochastic programming. SIAM J Optim 2005;16:69–95.

ROBUST EXTERNAL RISK MEASURES

STEVEN KOU

Department of Industrial
Engineering and Operations
Research, Columbia
University, New York,
New York

XIANHUA PENG

Department of Mathematics,
Hong Kong University of
Science and Technology,
Hong Kong

CHRISTOPHER C. HEYDE

Department of Statistics,
Columbia University,
New York, New York

Choosing a proper risk measure is an important regulatory issue, as exemplified in governmental regulations such as Basel II accord [1] and its recent revision [2], which use Value-at-Risk (VaR) with *scenario analysis* as the risk measure for setting capital requirement for market risk. The main motivation of this article is to investigate whether VaR with scenario analysis are good risk measures for external regulation. By using the notion of *comonotonic* random variables studied in decision theory literatures such as Refs 3–8, we shall propose a new class of risk measures satisfying a new set of axioms. The new class of risk measures includes VaR with scenario analysis, in particular the current and recently revised Basel II risk measures [1,2] as special cases, and therefore provides a theoretical basis for using VaR with scenario analysis as robust risk measures for the purpose of external, regulatory risk measurement.

Broadly speaking, a risk measure attempts to assign a single numerical value to a random financial loss. Obviously, it can be problematic to use one number to summarize the whole statistical distribution of the financial loss. Therefore, one shall avoid doing this if it is at all possible. However, in many cases there is no other choice.

Examples of such cases include regulatory capital requirements, insurance risk premiums, and margin requirements in financial trading. Consequently, how to choose a good risk measure becomes a problem of great practical importance.

In this article, we only study static risk measures, that is, one period risk measures. Mathematically, let Ω be the set of all the possible states of nature at the end of an observation period, $\mathcal{X}$ be the set of financial losses under consideration, in which each financial loss is a random variable defined on Ω . Then a risk measure ρ is a mapping from $\mathcal{X}$ to the real line $\mathbb{R}$. The multiperiod or dynamic risk measures are related to dynamic consistency or time consistency for preferences [9–13].

An important property of risk measures is *robustness*. The concept of robustness has been well developed in economics and statistics. In the theory of decision or policy making in economics, the robustness of preferences or policy rules refers to the property that they can accommodate model misspecification error and would perform well across a set of alternative models. This notion of robustness has a close connection with ambiguity aversion and model uncertainty. A widely used class of robust preferences that model ambiguity and explain the famous Ellsberg Paradox [14] are the multiple priors preferences, also known as maxmin expected utility preferences, proposed by Gilboa and Schmeidler [15]. An agent with this kind of preference ranks a pay off profile Y by $\min_{P \in \mathcal{P}} \{E^P[u(Y)]\}$, where $\mathcal{P}$ is the set of subjective probabilities, or a set of priors, held by the agent, and $E^P[u(Y)]$ is the expected utility of Y under the subjective probability P . The ambiguity is reflected by the set of priors $\mathcal{P}$. Multiple priors preferences have been used to incorporate ambiguity in asset pricing [16,17]. An alternative class of robust preferences are the multiplier preferences [18–20], which also postulate that agents have multiple prior probability measures. In statistics, robustness has been studied comprehensively in the robust statistics

literature, which is mainly concerned with the distributional robustness, that is, a statistic is robust if it is insensitive to small deviation of the true underlying distribution from the assumed distribution [21,22]. Loosely speaking, a robust statistic is insensitive to small changes in the data, which are either small changes in all the observations or large changes in a few observations (outliers).

A risk measure is said to be robust if (i) it can accommodate model misspecification (possibly by incorporating multiple priors); and (ii) it is insensitive to small changes in the data (by using robust statistics).

Another important issue related to risk measures, which has not been well addressed in the existing literature, is the objective of a risk measure. In terms of objectives, risk measures can be classified into two categories: *internal* risk measures used for internal risk management at individual institutions, and *external* risk measures used for external regulation and imposed for all the relevant institutions. Internal risk measures are proposed for the interests of institutions' shareholders or managers, while external risk measures are used by regulatory agencies to maintain safety and soundness of the financial system. One risk measure may be suitable for internal management, but not for external regulation, and vice versa. There is no reason to believe that one unique risk measure can fit all needs. For example, variance is an internal risk measure that is commonly used for hedging and asset allocation in institutions. However, variance is not a good risk measure for calculating capital requirements in external regulation, because it does not help to achieve the regulators' objective of ensuring that with very high confidence level, for example, 99.9% institutions have enough capital to absorb their losses. In contrast, VaR, which summarizes the worst loss of a portfolio at a given confidence level, better suits the needs of regulators and, therefore, is more suitable for external regulation than variance.

Another reason for the existence of internal and external risk measures lies in the fact that the information available for institutions and regulators are different.

Institutions have more information and hence could tailor their risk measures to incorporate all of it. External regulators have access to a subset of information only. Therefore, their way of measuring risk is different. No amount of disclosure requirements can eliminate private information that institutions have.

In this article, we shall focus on external risk measures from the viewpoint of regulatory agencies. We will show in the sections titled "Legal Motivation" and "The Main Reason to Relax Subadditivity: Robustness" that an external risk measure should be robust but *coherent risk measures* are generally not robust. We will then propose a new class of risk measures that includes a subclass of robust ones in the section titled "Main Results: Natural Risk Statistics."

VaR WITH SCENARIO ANALYSIS AND RISK STATISTICS

Value-at-Risk

One of the most widely used risk measures in financial regulation and risk management is VaR, which is a quantile at some predefined probability level. More precisely, let $F(\cdot)$ be the distribution function of X , then for given $\alpha \in (0, 1)$, VaR of the loss variable X at level α is defined as the α -quantile of X , that is,

$$\text{VaR}_\alpha(X) := \inf\{x \mid F(x) \geq \alpha\} = F^{-1}(\alpha). \quad (1)$$

In practice, VaR is usually estimated from samples of the random loss X , that is, a data set $\tilde{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. Let $(x_{(1)}, \dots, x_{(n)})$ be the sample order statistics and $F_n(x) := \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \leq x\}}$ be the empirical distribution function. The most commonly used estimator for $\text{VaR}_\alpha(X)$ is the sample quantile $x_{([\alpha n])} = F_n^{-1}(\alpha)$, where $[\alpha n]$ is the smallest integer that is no less than αn . The sample quantile is strongly consistent and asymptotically normal [23]. See Refs 24 and 25 for more comprehensive discussion of VaR.

VaR with Scenario Analysis and Basel II Risk Measures

VaR with scenario analysis is the class of risk measures that involve the calculation and

comparison of VaR under different scenarios. A simple example of VaR with scenario analysis is

$$\max\{\text{VaR}_1, \text{VaR}_2\}, \quad (2)$$

where VaR_i is the value of VaR calculated under scenario i , $i = 1, 2$.

Basel II Accord [1] uses VaR with scenario analysis to calculate capital requirement for market risk. More precisely, Basel II Accord specifies that the market risk capital requirement at any particular day t for banks using the internal models approach is calculated as

$$c_t = \max \left\{ \text{VaR}_{t-1}, k \cdot \frac{1}{60} \sum_{i=1}^{60} \text{VaR}_{t-i} \right\}, \quad (3)$$

where k is a constant that is no less than 3; VaR_{t-i} is the 10-day VaR at 99% confidence level calculated on day $t-i$, $i = 1, \dots, 60$. VaR_{t-i} is usually estimated from a data set $\tilde{x}^i = (x_1^i, x_2^i, \dots, x_{n_i}^i) \in \mathbb{R}^{n_i}$, which are based on the most recent one year observation of market variables up to day $t-i$ and can be generated by historical simulation or Monte Carlo simulation [24,25]. Hence, VaR_{t-i} corresponds to the scenario based on the information available on the day $t-i$. Therefore, the risk measure defined in Equation (3) is a VaR with scenario analysis that involves 60 scenarios.

During the financial crisis that started in late 2007, it was found that most banks' actual losses were significantly higher than the capital requirement calculated by Equation (3) [2]. In addition, the risk measure (Eq. 3) has been criticized for its procyclicality [26]. Therefore, the Basel committee recently revised the Basel II market risk framework [2]. The revised market risk capital requirement is defined as

$$c_t = \max \left\{ \text{VaR}_{t-1}, k \cdot \frac{1}{60} \sum_{i=1}^{60} \text{VaR}_{t-i} \right\} + \max \left\{ \text{sVaR}_{t-1}, \ell \cdot \frac{1}{60} \sum_{i=1}^{60} \text{sVaR}_{t-i} \right\}, \quad (4)$$

where VaR_{t-i} is defined the same as in Equation (3); k and ℓ are constants no less than 3; sVaR_{t-i} is called the *stressed* VaR on day $t-i$, which is calculated under the scenario that the financial market is under significant stress such as the one that happened during the period from 2007 to 2008. The additional capital requirement based on stressed VaR helps to reduce the procyclicality of the original risk measure (Eq. 3).

Interestingly, the risk measures defined in Equations (2), (3), and (4) do not belong to any existing theoretical framework of risk measures in the literature. In this article, we are going to define a new class of risk measures that include these risk measures as special cases and thus provide theoretical basis for the use of VaR with scenario analysis in external regulation.

Risk Statistics: Data-based Risk Measures

In external regulation, the random loss X under consideration corresponds to the loss of the whole trading book of a financial institution, which contains all kinds of assets across many sectors. Specifying accurate models for X (under different scenarios) involves modeling the joint distribution of a large number of correlated risk factors and is hence usually very difficult. Therefore, the behavior of X under different scenarios is usually represented by different sets of data observed or generated under different scenarios, and the risk measurement of the loss is directly estimated from the data.

More precisely, suppose the behavior of the random loss X is represented by a collection of data $\tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n$, where $\tilde{x}^i = (x_1^i, \dots, x_{n_i}^i) \in \mathbb{R}^{n_i}$ is the data subset that corresponds to the i -th scenario; n_i is the sample size of $\tilde{x}^i$; $n_1 + n_2 + \dots + n_m = n$; for each $i = 1, \dots, m$, $\tilde{x}^i$ can be a data set based on historical observations, or a data set simulated according to a well-defined procedure or model, or a mixture of historical and simulated data. The random loss X can be discrete or continuous. For example, the data used in the calculation of the revised Basel II risk measure defined in Equation (4) comprise 120 data subsets that correspond to 120 different scenarios.

A *risk statistic* $\hat{\rho}$ is simply a mapping from $\mathbb{R}^n$ to $\mathbb{R}$. It is a data-based risk measure that maps $\tilde{x}$, the data representation of the random loss X , to $\hat{\rho}(\tilde{x})$, the risk measurement of X .

In this article, we will define a new set of axioms for risk statistics instead of risk measures, because risk statistics directly measure risk from data without specifying subjective models, which greatly reduces model misspecification error.

REVIEW OF EXISTING RISK MEASURES

There are mainly two families of static risk measures suggested in the literature, coherent and convex risk measures, and insurance risk measures. Both are based on axiomatic approaches, by which some axioms are postulated first, and then all the risk measures satisfying the axioms are identified.

Coherent and Convex Risk Measures

Axioms. Coherent risk measures [27] satisfy the following axioms:

Axiom A1. Translation invariance: $\rho(X + b) = \rho(X) + b$, $\forall b \in \mathbb{R}, \forall X \in \mathcal{X}$.

Axiom A2. Positive homogeneity: $\rho(aX) = a\rho(X)$, $\forall a \geq 0, \forall X \in \mathcal{X}$.

Axiom A3. Monotonicity: $\rho(X) \leq \rho(Y)$, if $X \leq Y$.

Axiom A4. Subadditivity: $\rho(X + Y) \leq \rho(X) + \rho(Y)$, $\forall X, Y \in \mathcal{X}$.

What is questionable lies in the subadditivity requirement in Axiom A4, which basically means that “a merger does not create extra risk” [27, p. 209]. We will discuss the controversies related to this axiom in the section titled “Other Reasons to Relax Subadditivity.”

It is shown in Refs 22, 27, and 28 that a risk measure ρ is coherent if and only if there exists a family $\mathcal{Q}$ of probability measures, such that $\rho(X) = \sup_{Q \in \mathcal{Q}} \{E^Q[X]\}$, $\forall X \in \mathcal{X}$, where $E^Q[X]$ is the expectation of X under the probability measure Q . Therefore, measuring risk by a coherent risk measure amounts to computing maximal expectation

under a set of prior probability measures. An equivalent approach of defining coherent risk measures via acceptance sets is also presented in Refs 27 and 28.

A special case of coherent risk measures is the *tail conditional expectation* (TCE) [27, 29–31]. More precisely, for a loss X with a continuous distribution, TCE at level α of X is defined by

$$\text{TCE}_\alpha(X) := E[X | X \geq \text{VaR}_\alpha(X)]. \quad (5)$$

For X with general distributions, $\text{TCE}_\alpha(X)$ is defined as a regularized version of $E[X | X \geq \text{VaR}_\alpha(X)]$ [32].

Another property of risk measures is called *law invariance*, as stated in the following Axiom A5.

Axiom A5. Law invariance: $\rho(X) = \rho(Y)$, if X and Y have the same distribution.

A risk measure is called a *law-invariant coherent risk measure*, if it satisfies Axiom A1–A5. Kusuoka [33] gives a representation for the law-invariant coherent risk measures.

Convex risk measures proposed in Refs 34 and 35 are risk measures that satisfy Axiom A1, A3, and the following convexity axiom A6, which relaxes the positive homogeneity and subadditivity axioms in coherent risk measures.

Axiom A6. Convexity: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$, $\forall X, Y \in \mathcal{X}, \forall \lambda \in [0, 1]$.

Main Drawbacks of Coherent and Convex Risk Measures. A main drawback of coherent risk measures is that in general, they are not robust (see the section titled “Comparison with Coherent Risk Measures”). For example, TCE is too sensitive to the modeling assumption for the tails of loss distributions, and it is sensitive to outliers in the data (see the section titled “Tail Conditional Median: A Robust Risk Measure”). However, robustness is an indispensable requirement for external risk measures, as we will point out in the section titled “The Main Reason to Relax Subadditivity: Robustness.” Hence, coherent

risk measures are not suitable for external regulation.

Another drawback of coherent and convex risk measures is that they rule out VaR, which does not satisfy subadditivity or convexity universally [27]. This posts a serious inconsistency between the academic theory and governmental practice. The inconsistency is due to the subadditivity Axiom A4. By relaxing this axiom and requiring subadditivity only for comonotonic random variables, we are able to define a new class of risk measures that include VaR and more generally, VaR with scenario analysis, thus eliminating the inconsistency.

Insurance Risk Measures

Insurance risk measures [8] satisfy the following axioms:

Axiom B1. Law invariance: the same as Axiom A5.

Axiom B2. Monotonicity: $\rho(X) \leq \rho(Y)$, if $X \leq Y$.

Axiom B3. Comonotonic additivity: $\rho(X + Y) = \rho(X) + \rho(Y)$, if X and Y are comonotonic. (X and Y are *comonotonic* if $(X(\omega_1) - X(\omega_2))(Y(\omega_1) - Y(\omega_2)) \geq 0$ holds almost surely for ω_1 and ω_2 in Ω .)

Axiom B4. Continuity: $\lim_{d \rightarrow 0} \rho((X - d)^+) = \rho(X^+)$, $\lim_{d \rightarrow \infty} \rho(\min(X, d)) = \rho(X)$, and $\lim_{d \rightarrow -\infty} \rho(\max(X, d)) = \rho(X)$, $\forall X$. ($x^+ := \max(x, 0)$, $\forall x \in \mathbb{R}$).

Axiom B5. Scale normalization: $\rho(1) = 1$.

The notion of comonotonic random variables is studied in Refs 3, 5, and 36. The psychological motivation of comonotonic random variables comes from alternative models to expected utility theory including the prospect theory (see the section titled “Theory of Choice under Uncertainty and Risk”). Dhaene *et al.* [37] give a recent review of risk measures and comonotonicity.

VaR with scenario analysis, such as the Basel II risk measures (Eqs 3 and 4), are not insurance risk measures, although VaR itself is an insurance risk measure (see Ref. 36 for the proof that VaR satisfies Axiom B3). An insurance risk measure does not necessarily satisfy subadditivity.

A main drawback of insurance risk measures is that neither can they incorporate multiple priors of the probability distributions on the set of scenarios, nor can they incorporate multiple approaches to measure risk under each scenario (see the section titled “Comparison with Insurance Risk Measures” for more detailed discussion).

The main reason that insurance risk measures cannot incorporate multiple priors is that they require comonotonic additivity. To incorporate multiple priors, we shall relax the comonotonic additivity to comonotonic subadditivity in the section titled “Main Results: Natural Risk Statistics.”

The mathematical concept of comonotonic subadditivity is also studied independently in Refs 38 and 39, which give representations of the functionals satisfying comonotonic subadditivity or comonotonic convexity from a mathematical perspective. There are several differences between Refs 38 and 39 and this article. See Ref. 40 for the details.

LEGAL MOTIVATION

In this section, we shall discuss two basic concepts of law, that is, legal realism and legal positivism, which motivate us to argue that: (i) External risk measures should be robust, because robustness is essential for law enforcement; (ii) External risk measures should have consistency with people’s behavior, because law should reflect society norms.

Legal Realism and Robustness of Law

Legal realism is the viewpoint that the legal decision of a court regarding a case is determined by the actual practices of judges, rather than the law set forth in statutes and precedents. All the legal rules contained in statutes and precedents have uncertainty due to the uncertainty in human language and that human beings are unable to anticipate all the possible future circumstances [41, p. 128]. Hence, a law is only a guideline for judges and enforcement officers [41, pp. 204–205], that is, a law is only intended to be the average of what judges and officers will decide. This concerns the robustness of law, that is, a law should be established in

such a way that different judges will reach similar conclusions when they implement the law.

In particular, external risk measures imposed in banking regulation should be robust with respect to underlying models and data. However, coherent risk measures generally lack robustness, as manifested in Theorem 3.

Legal Positivism and Social Norm

Legal positivism is the thesis that the existence and content of law depend on social norms and not on their merits. It is based on the argument that if a system of rules are to be imposed by force in the form of law, there must be a sufficient number of people who accept it voluntarily; without their voluntary cooperation, the coercive power of law and government cannot be established [41, pp. 201–204].

Therefore, external risk measures imposed in banking regulations should also reflect most people's behavior. However, decision theory suggests that most people's decision can violate the subadditivity Axiom A4 (see the section titled "Theory of Choice under Uncertainty and Risk" for details).

THE MAIN REASON TO RELAX SUBADDITIVITY: ROBUSTNESS

Robustness is Indispensable for External Risk Measures

In determining capital requirement, regulators impose a risk measure and allow institutions to use their own internal risk models and private data in the calculation. For example, Basel II's internal models approach allows institutions to use their own internal models to calculate the capital requirement for market risk.

There are two issues rising from the use of internal models and private data in external regulation: (i) the data can be noisy, flawed, or unreliable; (ii) there can be several statistically indistinguishable models for the same asset or portfolio due to limited amount of available data.

For example, the heaviness of tail distributions cannot be identified in many

cases. Although it is accepted that stock returns have tails heavier than those of normal distributions, one school of thought believes that the tails are exponential-type, while another believes that the tails are power-type. Heyde and Kou [42] show that it is very difficult to distinguish between exponential-type and power-type tails with 5,000 observations (about 20 years of daily observations). This is mainly because the quantiles of exponential-type distributions and power-type distributions may overlap. Hence, the tail behavior may be a subjective issue depending on people's modeling preferences.

To address the two issues above, external risk measures should demonstrate robustness with respect to model misspecification and small changes in the data. From a regulator's viewpoint, an external risk measure must be unambiguous, stable, and be implemented consistently across all the relevant institutions, no matter what internal beliefs or internal models each individual institution may have. In situations when the correct model cannot be identified, two institutions that have exactly the same portfolio can use different internal models, both of which can obtain the approval of the regulator. From the regulator's viewpoint, the same portfolio should incur the same or almost the same amount of regulatory capital. Therefore, the external risk measure should be robust, otherwise different institutions can be required to hold very different regulatory capital for the same risk exposure, which makes the risk measure unacceptable to most institutions.

In addition, if the external risk measure is not robust, institutions can take regulatory arbitrage by choosing a model that significantly reduces the capital requirement or by manipulating the input data.

Coherent Risk Measures are not Robust

The robustness of coherent risk measures is questionable in two aspects.

First, the theory of coherent risk measures suggests to use TCE to measure risk. However, as we will show in the section titled "Tail Conditional Median: A Robust Risk Measure," TCE is sensitive to modeling

assumptions of heaviness of tail distribution and to outliers in the data.

Second, in general, coherent risk measures are not robust. We will prove in Theorem 3 that every empirically law-invariant coherent risk measure puts larger weights on larger observations, and hence is obviously sensitive to small changes in the data.

Tail Conditional Median: A Robust Risk Measure

We propose a more robust risk measure, tail conditional median (TCM), which is a special case of the new class of risk measures to be defined in the section titled “Main Results: Natural Risk Statistics.” TCM of loss X at level α is defined as

$$\text{TCM}_\alpha(X) = \text{VaR}_{\frac{1+\alpha}{2}}(X).$$

It is called conditional median, because for X with a continuous distribution,

$$\begin{aligned} \text{TCM}_\alpha(X) &= \text{VaR}_{\frac{1+\alpha}{2}}(X) \\ &= \text{median}[X|X \geq \text{VaR}_\alpha(X)], \quad (6) \end{aligned}$$

that is, the conditional median of X given that $X \geq \text{VaR}_\alpha(X)$. For X with a general distribution involving discontinuity, $\text{VaR}_{\frac{1+\alpha}{2}}(X)$ can be viewed as a regularized version of $\text{median}[X|X \geq \text{VaR}_\alpha(X)]$.

The equality (Eq. 6) shows that if one wants to measure the size of loss beyond the confidence level α , one can use VaR at a higher level $\frac{1+\alpha}{2}$, which gives the conditional median of the size of loss beyond level α . This contradicts the criticism on VaR that VaR cannot provide information about the size of loss beyond certain confidence level.

There are many examples in the existing literature showing that VaR violates subadditivity, for example, the examples on p. 216 and p. 217 of Ref. 27. However, it is straightforward to verify that in those examples subadditivity will not be violated if one replaces VaR_α by TCM_α .

TCM can be shown to be more robust than TCE by at least four standard tools in robust statistics [21,22]: (i) influence functions, (ii)

asymptotic breakdown points, (iii) continuity of statistical functional, and (iv) finite sample breakdown points. For example, TCE has an unbounded influence function but TCM has a bounded influence function, which implies that TCM is more robust than TCE with respect to outliers in the data. See Ref. 40 for more detailed discussion.

TCE is highly model-dependent, in particular, sensitive to modeling assumption of extreme tails of loss distribution, as illustrated in Fig. 1. The left panel of the figure shows the value of TCE_α with respect to $\log(1-\alpha)$ for Laplace and T distributions with degree of freedom 3, 5, and 12, which are normalized to have mean 0 and variance 1, where α is in the range [95%, 99.9%]. The right panel of the figure shows the value of TCM_α for those distributions. Apparently, the range of variation of TCM_α with respect to change of modeling assumption is much smaller than that of TCE_α . For example, with $\alpha = 99.6\%$, the variation of TCE_α is 1.44, whereas the variation of TCM_α is only 0.75. In the context of capital requirement, suppose a bank's trading portfolio loss is σX , where $\sigma = \$1$ billion and X has mean 0 and variance 1, which can be modeled to have one of the four distributions. At confidence level $\alpha = 99\%$, if TCE_α is used to calculate capital requirement, the capital requirement for banks using one of the four models will be in the range of [2.95, 4.05] billion; in contrast, if TCM_α is used, the range will be [2.79, 3.38] billion. Hence, TCM results in much more comparable capital requirements across banks when they may use different models.

It is worth noticing that TCE is not necessarily larger than TCM. Eling and Tibiletti [43] compare TCE and TCM for a set of capital market data and find that although TCE is on average about 10% higher than TCM at standard confidence levels, TCM is higher than TCE in about 10% of the studied cases.

Robust Risk Measures Versus Conservative Risk Measures

A risk measure is said to be more conservative than another, if it generates higher risk measurement than the other for the

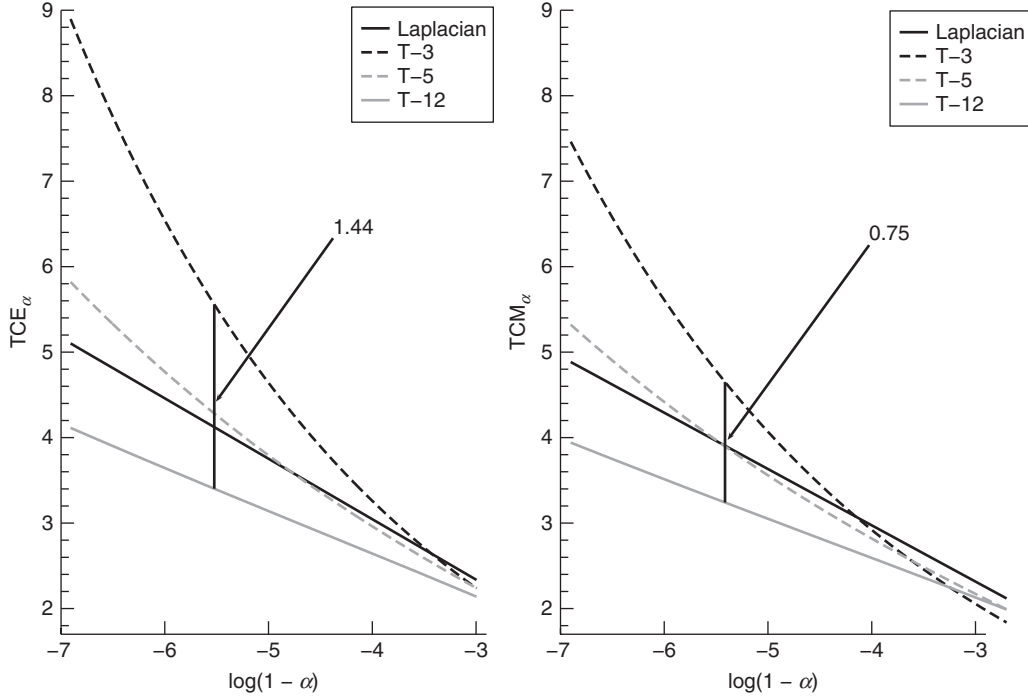


Figure 1. The left panel shows TCE for Laplacian distribution and T distributions with degree of freedom 3, 5, 12, which are normalized to have mean 0 and variance 1. The right panel shows TCM for the same set of distributions. The x-axis is $\log(1 - \alpha)$ where $\alpha \in [0.95, 0.999]$. TCM is less sensitive to modeling assumption of the underlying distribution than TCE, as the change of its value with respect to change of underlying distributions has a narrower range.

same underlying risk exposure. The use of more conservative risk measures in external regulation is desirable from a regulator's viewpoint, since it generally increases the safety of the financial system (Of course, risk measures should not be too conservative to retard the economic growth).

There is no contradiction between the robustness and the conservativeness of external risk measures. Robustness addresses the issue of whether a risk measure can be implemented consistently, so it is a requisite property of an external risk measure. Conservativeness addresses the issue of how stringently an external risk measure should be implemented, given that it can be implemented consistently. In other words, being more conservative is a further desirable property of an external risk measure. An external risk measure should be robust in the first place before one can consider the

issue of how to implement it in a conservative way.

By the following approaches, one can construct an external risk measure that is both robust and conservative:

1. More scenarios can be included to incorporate the possible effects of extreme events that lie outside normal market conditions. For example, the revised Basel II risk measure Equation (4) takes into account 60 more stressed scenarios compared to the Basel II risk measure Equation (3).
2. Regulators can simply multiply the original risk measure by a larger constant to obtain a more conservative risk measure. For example, one can use larger k in Equation (3). This is equivalent to choosing a larger s in Axiom C1 for the new risk measures (see the

section titled “Main Results: Natural Risk Statistics”).

3. One can use more weights or more prior probability measures to define the external risk measure. See the section titled “Main Results: Natural Risk Statistics” for more detailed discussion.

OTHER REASONS TO RELAX SUBADDITIVITY

Diversification and Tail Subadditivity of VaR

The subadditivity axiom for coherent risk measures and the convexity axiom for convex risk measures conform to the idea that diversification does not increase risk. There are two main motivations for diversification. One is based on the simple observation that $SD(X + Y) \leq SD(X) + SD(Y)$, for any two random variables X and Y with finite second moments, where $SD(\cdot)$ denotes standard deviation. The other is based on expected utility theory. Samuelson [44] shows that any investor with a strictly concave utility function will uniformly diversify among independently and identically distributed risks with finite second moments (see Refs 45–48 for the discussion on whether diversification is beneficial when the asset returns are dependent). Both of the two motivations require finiteness of second moments of the risks.

Is diversification still preferable for risks with infinite second moments? The answer can be no. Ibragimov and Walden [49] show that diversification is not preferable for unbounded extremely heavy-tailed distributions, in the sense that the expected utility of the diversified portfolio is smaller than that of the undiversified portfolio. They also show that, investors with certain S-shaped utility functions would prefer nondiversification, even for bounded risks. An S-shaped utility function is convex in the domain of losses, which is supported by experimental results and prospect theory [6,50]. See the section titled “Theory of Choice under Uncertainty and Risk” for more discussion on prospect theory.

VaR has been criticized because it does not satisfy subadditivity universally, but this

criticism is unreasonable because even diversification is not universally preferable. In fact, Danielsson *et al.* [51] show that VaR is subadditive in the tail region, provided that the tail of the joint distribution is not extremely fat (with tail index less than one). The simulations that they carry out also show that VaR_α is indeed subadditive when $\alpha \in [95\%, 99\%]$ for most practical applications.

To summarize, there is no conflict between the use of VaR and diversification. When the risks do not have extremely heavy tails, diversification is preferred and VaR satisfies subadditivity in the tail region; when the risks have extremely heavy tails, diversification may not be preferable and VaR may fail to satisfy subadditivity.

Asset returns with tail index less than one have very fat tails. They are hard to find and easy to identify. Danielsson *et al.* [51] argue that they can be treated as special cases in financial modeling. Even if one encounters an extremely fat tail and insists on tail subadditivity, Garcia *et al.* [52] show that when tail thickness causes violation of subadditivity, a decentralized risk management team may restore the subadditivity for VaR by using proper conditional information.

Does A Merger Always Reduce Risk

Subadditivity basically means that “a merger does not create extra risk” [27, p. 209]. However, Dhaene *et al.* [53] point out that a merger may increase risk, particularly due to bankruptcy protection for institutions. For example, it is better to split a risky trading business into a separate subsidiary institution. This way, even if the loss from the subsidiary institution is enormous, the parent institution can simply let the subsidiary institution go bankrupt, thus confining the loss to that one subsidiary institution. Therefore, creating subsidiary institutions may incur less risk and a merger may increase risk.

For example, the collapse of Britain’s Barings Bank in February 1995, due to the failure of a single trader (Nick Leeson) in Singapore clearly indicates that a merger may increase risk. Had Barings Bank setup a separate institution for its Singapore unit, the bankruptcy in that unit would not have sunk the entire bank.

Theory of Choice under Uncertainty and Risk

Risk measures have a close connection with the psychological and economic theory of people's choices under uncertainty and risk. Tversky and Kahneman [6,50] point out that people's choices under risky prospects are inconsistent with the basic tenets of expected utility theory and propose an alternative model, called *prospect theory*, which can explain a variety of preference anomalies. The axiomatic analysis of prospect theory is presented in Refs 6 and 7. Many other people have studied alternative models to expected utility theory, such as Refs 3–5, 54.

Kou *et al.* [40] provide a simple example showing that risks associated with noncomonotonic random variables can violate subadditivity, because people are risk seeking instead of risk averse in choices between probable and sure losses, as implied by prospect theory. The example also shows that most people's behavior can violate the subadditivity axiom.

Schmeidler [4] points out that risk preferences for comonotonic random variables are easier to justify than those for arbitrary random variables and accordingly relaxes the independence axiom in expected utility theory to the comonotonic independence axiom that applies only to comonotonic random variables. Tversky and Kahneman [6] and Wakker and Tversky [7] also find that prospect theory is characterized by the axioms that impose conditions only on comonotonic random variables and are thus less restrictive than their counterparts in expected utility theory.

Motivated by the prospect theory, we think it may be appropriate to relax the subadditivity to comonotonic subadditivity. In other words, we impose $\rho(X + Y) \leq \rho(X) + \rho(Y)$ only for comonotonic random variables X and Y when defining the new risk measures in the section titled “Main Results: Natural Risk Statistics.”

MAIN RESULTS: NATURAL RISK STATISTICS

In this section, we will define and fully characterize the new data-based risk measures, which we call the natural risk statistics.

As discussed in the section titled “Risk Statistics: Data-Based Risk Measures,” we are going to measure the risk from a collection of data $\tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n$ that represents the behavior of the random loss X , where $\tilde{x}^i = (x_1^i, x_2^i, \dots, x_{n_i}^i)$ is the data subset that corresponds to scenario i . The risk measurement will be given by $\hat{\rho}(\tilde{x})$, where $\hat{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ is a risk statistic.

Axioms and Representations for Natural Risk Statistics

At first, we define the concept of scenario-wise comonotonicity for two data sets.

Definition 1. Two data sets $\tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n$ and $\tilde{y} = (\tilde{y}^1, \tilde{y}^2, \dots, \tilde{y}^m) \in \mathbb{R}^n$ are scenario-wise comonotonic, if for $\forall i$, $\forall 1 \leq j, k \leq n_i$, $(x_j^i - x_k^i)(y_j^i - y_k^i) \geq 0$.

Then we postulate the following axioms for a risk statistic $\hat{\rho}$.

Axiom C1. Positive homogeneity and translation scaling: $\hat{\rho}(a\tilde{x} + b\mathbf{1}) = a\hat{\rho}(\tilde{x}) + s \cdot b$, $\forall \tilde{x} \in \mathbb{R}^n$, $\forall a \geq 0$, $\forall b \in \mathbb{R}$, where $s > 0$ is a constant, $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^n$.

Axiom C2. Monotonicity: $\hat{\rho}(\tilde{x}) \leq \hat{\rho}(\tilde{y})$ for any $\tilde{x} \leq \tilde{y}$, where $\tilde{x} \leq \tilde{y}$ means $x_j^i \leq y_j^i$, $\forall j, i$.

The two axioms above (with $s = 1$) have been proposed for coherent risk measures.

Axiom C3. Scenario-wise comonotonic subadditivity: $\hat{\rho}(\tilde{x} + \tilde{y}) \leq \hat{\rho}(\tilde{x}) + \hat{\rho}(\tilde{y})$, for any $\tilde{x}$ and $\tilde{y}$ that are scenario-wise comonotonic.

In Axiom C3, we relax the subadditivity requirement in coherent risk measures so that the subadditivity is only required for comonotonic random variables. This also relaxes the comonotonic additivity requirement in insurance risk measures.

Axiom C4. Scenario-wise permutation invariance:

$$\hat{\rho}((\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m)) = \hat{\rho}((x_{p_1,1}^1, \dots, x_{p_1,n_1}^1, x_{p_2,1}^2, \dots, x_{p_2,n_2}^2, \dots, x_{p_m,1}^m, \dots, x_{p_m,n_m}^m)),$$

for any permutation $(p_{i,1}, \dots, p_{i,n_i})$ of $(1, 2, \dots, n_i)$, $i = 1, \dots, m$.

This axiom can be considered as the counterpart of the law invariance Axiom A5 in the empirical setting. It means that if two data set $\tilde{x}$ and $\tilde{y}$ have the same empirical distribution under each scenario, then $\tilde{x}$ and $\tilde{y}$ should give the same measurement of risk.

Definition 2. A risk statistic $\hat{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ is called a natural risk statistic if it satisfies Axiom C1–C4.

The following representation theorem fully characterizes the natural risk statistics.

Theorem 1. (i) For a given constant $s > 0$ and an arbitrarily given set of weights $\mathcal{W} = \{\tilde{w}\} \subset \mathbb{R}^n$ with each $\tilde{w} = (w_1^1, \dots, w_{n_1}^1, \dots, w_1^m, \dots, w_{n_m}^m) \in \mathcal{W}$ satisfying the following conditions:

$$\sum_{j=1}^{n_1} w_j^1 + \sum_{j=1}^{n_2} w_j^2 + \dots + \sum_{j=1}^{n_m} w_j^m = 1, \quad (7)$$

$$w_j^i \geq 0, j = 1, \dots, n_i; i = 1, \dots, m, \quad (8)$$

define a risk statistic $\hat{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ as follows:

$$\begin{aligned} \hat{\rho}(\tilde{x}) := s \cdot \sup_{\tilde{w} \in \mathcal{W}} & \left\{ \sum_{j=1}^{n_1} w_j^1 x_{(j)}^1 + \sum_{j=1}^{n_2} w_j^2 x_{(j)}^2 + \dots \right. \\ & \left. + \sum_{j=1}^{n_m} w_j^m x_{(j)}^m \right\}, \\ \forall \tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) & \in \mathbb{R}^n, \quad (9) \end{aligned}$$

where $(x_{(1)}^i, \dots, x_{(n_i)}^i)$ is the order statistics of $\tilde{x}^i = (x_1^i, \dots, x_{n_i}^i)$ with $x_{(n_i)}^i$ being the largest, $i = 1, \dots, m$. Then the risk statistic defined in Equation (9) is a natural risk statistic.

(ii) If $\hat{\rho}$ is a natural risk statistic, then there exists a set of weights $\mathcal{W} = \{\tilde{w} = (\tilde{w}^1, \tilde{w}^2, \dots, \tilde{w}^m)\} \subset \mathbb{R}^n$ with each $\tilde{w} \in \mathcal{W}$

satisfying condition (7) and (8), such that

$$\begin{aligned} \hat{\rho}(\tilde{x}) = s \cdot \sup_{\tilde{w} \in \mathcal{W}} & \left\{ \sum_{j=1}^{n_1} w_j^1 x_{(j)}^1 + \sum_{j=1}^{n_2} w_j^2 x_{(j)}^2 + \dots \right. \\ & \left. + \sum_{j=1}^{n_m} w_j^m x_{(j)}^m \right\}, \\ \forall \tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) & \in \mathbb{R}^n. \end{aligned} \quad (10)$$

Proof. See Ref. 40 (for the case of $m = 1$, see also Ref. 55). Q.E.D.

The natural risk statistics can also be characterized via acceptance sets as in the case of coherent risk measures. See Ref. 40 for the details.

The Basel II risk measures, and other VaR with scenario analysis, are special cases of natural risk statistics. More precisely, we have the following theorem:

Theorem 2. The Basel II risk measure defined in Equation (3) and the revised Basel II risk measure defined in Equation (4) are both natural risk statistics.

Proof. See Ref. 40. Q.E.D.

Robust Natural Risk Statistics

Natural risk statistics include a subclass of risk statistics that are robust in two aspects: (i) they can accommodate model misspecification by incorporating multiple prior probability measures on the set of scenarios; (ii) they are insensitive to small changes in the data by using robust statistics for each scenario.

Let $\hat{\rho}$ be a natural risk statistic defined in Equation (9) that corresponds to the set of weights $\mathcal{W}$. Define a map $\phi : \mathcal{W} \rightarrow \mathbb{R}^m \times \mathbb{R}^n$ such that for each $\tilde{w} \in \mathcal{W}$, $\phi(\tilde{w}) := (\tilde{p}, \tilde{q})$, where $\tilde{p} := (p^1, \dots, p^m)$, $p^i := \sum_{j=1}^{n_i} w_j^i$,

and $\tilde{q} := (q_1^1, \dots, q_{n_1}^1, \dots, q_1^m, \dots, q_{n_m}^m)$, $q_j^i := 1_{\{p_i > 0\}} w_j^i / p_i$. Then $\hat{\rho}$ can be written as

$$\hat{\rho}(\tilde{x}) = s \cdot \sup_{(\tilde{p}, \tilde{q}) \in \phi(\mathcal{W})} \left\{ \sum_{i=1}^m p^i \hat{\rho}^i(\tilde{x}^i) \right\}, \text{ where}$$

$$\hat{\rho}^i(\tilde{x}^i) := \sum_{j=1}^{n_i} q_j^i x_{(j)}^i. \quad (11)$$

Thus, each weight $\tilde{w} \in \mathcal{W}$ specifies: (i) the prior probability measure $\tilde{p}$ on the set of scenarios; and (ii) the subsidiary risk statistic $\hat{\rho}^i$ for each scenario i , which measures risk from data subset $\tilde{x}^i$.

Hence, $\hat{\rho}$ can be robust to model misspecification by incorporating multiple prior probability measures on the set of scenarios. And $\hat{\rho}$ can be robust to small changes in the data if each subsidiary risk statistic $\hat{\rho}^i(\tilde{x}^i)$ is a robust statistic, for example, VaR, or other robust L-statistics [22, Chapter 3].

The Basel II risk measures defined in Equations (3) and (4) incorporate multiple priors of probability distributions on the set of scenarios and use VaR as subsidiary risk statistics. Hence, they are robust natural risk statistics.

COMPARISON WITH COHERENT AND INSURANCE RISK MEASURES

Comparison with Coherent Risk Measures

To compare with coherent risk measures, we first define the law-invariant coherent risk statistics, the data-based versions of law-invariant coherent risk measures.

Definition 3. A risk statistic $\hat{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ is called a law-invariant coherent risk statistic, if it satisfies Axiom C1, C2, C4, and the following axiom E3.

Axiom E3. Subadditivity: $\hat{\rho}(\tilde{x} + \tilde{y}) \leq \hat{\rho}(\tilde{x}) + \hat{\rho}(\tilde{y})$, $\forall \tilde{x}, \tilde{y} \in \mathbb{R}^n$.

We have the following representation theorem for law-invariant coherent risk statistics.

Theorem 3. (i) For a given constant $s > 0$ and an arbitrarily given set of weights $\mathcal{W} = \{\tilde{w}\} \subset \mathbb{R}^n$ with each $\tilde{w} = (w_1^1, \dots, w_{n_1}^1, \dots, w_1^m, \dots, w_{n_m}^m) \in \mathcal{W}$ satisfying the following conditions:

$$\sum_{j=1}^{n_1} w_j^1 + \sum_{j=1}^{n_2} w_j^2 + \dots + \sum_{j=1}^{n_m} w_j^m = 1, \quad (12)$$

$$w_j^i \geq 0, j = 1, \dots, n_i; i = 1, \dots, m, \quad (13)$$

$$w_1^i \leq w_2^i \leq \dots \leq w_{n_i}^i, i = 1, \dots, m, \quad (14)$$

define a risk statistic

$$\hat{\rho}(\tilde{x}) := s \cdot \sup_{\tilde{w} \in \mathcal{W}} \left\{ \sum_{j=1}^{n_1} w_j^1 x_{(j)}^1 + \sum_{j=1}^{n_2} w_j^2 x_{(j)}^2 + \dots + \sum_{j=1}^{n_m} w_j^m x_{(j)}^m \right\},$$

$$\forall \tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n, \quad (15)$$

where $(x_{(1)}^i, \dots, x_{(n_i)}^i)$ is the order statistics of $\tilde{x}^i = (x_1^i, \dots, x_{n_i}^i)$ with $x_{(n_i)}^i$ being the largest, $i = 1, \dots, m$. Then the risk statistic defined in Equation (15) is a law-invariant coherent risk statistic.

(ii) If $\hat{\rho}$ is a law-invariant coherent risk statistic, then there exists a set of weights $\mathcal{W} = \{\tilde{w}\} \subset \mathbb{R}^n$ with each $\tilde{w} \in \mathcal{W}$ satisfying Equations (12), (13), and (14), such that

$$\hat{\rho}(\tilde{x}) = s \cdot \sup_{\tilde{w} \in \mathcal{W}} \left\{ \sum_{j=1}^{n_1} w_j^1 x_{(j)}^1 + \sum_{j=1}^{n_2} w_j^2 x_{(j)}^2 + \dots + \sum_{j=1}^{n_m} w_j^m x_{(j)}^m \right\},$$

$$\forall \tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n. \quad (16)$$

Proof. See Ref. 40 (for the case of $m = 1$, see also Ref. 55). Q.E.D.

By Theorem 1 and 3, we see the main differences between natural risk statistics and coherent risk measures:

1. Coherent risk measures are in general not robust, because every law-invariant

coherent risk statistic assigns larger weights to larger observations, but assigning larger weights to larger observations is apparently sensitive to small changes in the data. In fact, the finite sample breakdown point (see Ref. 22 for the definition) of any law-invariant coherent risk statistic is equal to $1/(1+n)$, that is, one single contamination sample can cause unbounded bias.

2. VaR and VaR with scenario analysis are not law-invariant coherent risk statistics, but are natural risk statistics.
3. Law-invariant coherent risk statistics, which exclude the use of robust statistics, are a subclass of natural risk statistics.

Comparison with Insurance Risk Measures

Similarly, we can define the insurance risk statistics, the data-based versions of insurance risk measures, as follows:

Definition 4. A risk statistic $\hat{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ is called an insurance risk statistic, if it satisfies the following Axiom F1–F4.

Axiom F1. Scenario-wise permutation invariance: the same as Axiom C4.

Axiom F2. Monotonicity: $\hat{\rho}(\tilde{x}) \leq \hat{\rho}(\tilde{y})$, if $\tilde{x} \leq \tilde{y}$.

Axiom F3. Scenario-wise comonotonic additivity: $\hat{\rho}(\tilde{x} + \tilde{y}) = \hat{\rho}(\tilde{x}) + \hat{\rho}(\tilde{y})$, if $\tilde{x}$ and $\tilde{y}$ are scenario-wise comonotonic.

Axiom F4. Scale normalization: $\hat{\rho}(\mathbf{1}) = 1$.

We have the following representation theorem for insurance risk statistics.

Theorem 4. $\hat{\rho}$ is an insurance risk statistic if and only if there exists a single weight $\tilde{w} = (w_1^1, \dots, w_{n_1}^1, \dots, w_1^m, \dots, w_{n_m}^m) \in \mathbb{R}^n$ with $w_j^i \geq 0$ for $\forall i, j$ and $\sum_{i=1}^m \sum_{j=1}^{n_i} w_j^i = 1$, such that

$$\hat{\rho}(\tilde{x}) = \sum_{j=1}^{n_1} w_j^1 x_{(j)}^1 + \sum_{j=1}^{n_2} w_j^2 x_{(j)}^2 + \dots + \sum_{j=1}^{n_m} w_j^m x_{(j)}^m, \\ \forall \tilde{x} = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^m) \in \mathbb{R}^n, \quad (17)$$

where $(x_{(1)}^i, \dots, x_{(n_i)}^i)$ is the order statistics of $\tilde{x}^i = (x_1^i, \dots, x_{n_i}^i)$, $i = 1, \dots, m$.

Proof. See Ref. 40. Q.E.D.

By Theorem 1 and 4, we see the main differences between natural risk statistics and insurance risk statistics:

1. Insurance risk statistics are less robust to model misspecification than natural risk statistics, because they do not incorporate multiple priors of probability distributions on the set of scenarios, and they do not incorporate multiple subsidiary risk statistics for each scenario.
2. Insurance risk statistics do not include VaR with scenario analysis such as the ones defined in Equations (2), (3), and (4), although they include VaR as a special case. However, VaR with scenario analysis are special cases of natural risk statistics.
3. Insurance risk statistics, which incorporate only one prior, are a subclass of natural risk statistics.

CONCLUSION

We propose new data-based risk measures called *natural risk statistics* that are characterized by a new set of axioms. The new axioms only require subadditivity for comonotonic random variables, which is consistent with prospect theory in economics. Natural risk statistics include (i) TCM, which is more robust than TCE suggested by coherent risk measures; (ii) VaR with scenario analysis, in particular, the Basel II risk measures, as special cases. Therefore, natural risk statistics provide a theoretical basis for using VaR with scenario analysis in external regulation.

We emphasize that an external risk measure should be robust with respect to model misspecification and small changes in the data in order for the consistent implementation of the risk measure across all the

relevant institutions. In contrast to coherent risk measures which are generally not robust, natural risk statistics include a subclass of robust risk statistics that are suitable for external regulation.

REFERENCES

1. Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: a revised framework (Comprehensive Version); 2006.
2. Basel Committee on Banking Supervision. Revisions to the Basel II market risk framework; 2009.
3. Schmeidler D. Integral representation without additivity. *Proc Am Math Soc* 1986; 97(2):255–261.
4. Schmeidler D. Subjective probability and expected utility without additivity. *Econometrica* 1989;57(3):571–587.
5. Yaari ME. The dual theory of choice under risk. *Econometrica* 1987;55(1):95–115.
6. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 1992;5(4): 297–323.
7. Wakker P, Tversky A. An axiomatization of cumulative prospect theory. *J Risk Uncertain* 1993;7(2):147–175.
8. Wang SS, Young VR, Panjer HH. Axiomatic characterization of insurance prices. *Insur Math Econ* 1997;21(2):173–183.
9. Koopmans TC. Stationary ordinal utility and impatience. *Econometrica* 1960;28(2): 287–309.
10. Epstein LG, Zin SE. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica* 1989;57(4):937–969.
11. Duffie D, Epstein LG. Stochastic differential utility. *Econometrica* 1992;60(2): 353–394.
12. Wang T. Conditional preferences and updating. *J Econ Theory* 2003;108(2):286–321.
13. Epstein LG, Schneider M. Recursive multiple-priors. *J Econ Theory* 2003;113(1):1–31.
14. Ellsberg D. Risk, ambiguity, and the Savage axioms. *Q J Econ* 1961;75(4):643–669.
15. Gilboa I, Schmeidler D. Maxmin expected utility with non-unique prior. *J Math Econ* 1989;18(2):141–153.
16. Epstein LG, Wang T. Intertemporal asset pricing under Knightian uncertainty. *Econometrica* 1994;62:283–322.
17. Chen Z, Epstein LG. Ambiguity, risk, and asset returns in continuous time. *Econometrica* 2002;70:1403–1443.
18. Hansen LP, Sargent TJ. Wanting robustness in macroeconomics. Mimeo, University of Chicago and Stanford University; 2000.
19. Hansen LP, Sargent TJ. Robust control and model uncertainty. *Am Econ Rev* 2001; 91:60–66.
20. Hansen LP, Sargent TJ. Robustness. Princeton (NJ): Princeton University Press; 2007.
21. Hampel FR. The influence curve and its role in robust estimation. *J Am Stat Assoc* 1974;69:383–393.
22. Huber PJ, Ronchetti EM. Robust statistics. 2nd ed. New York: Wiley; 2009.
23. Shao J. Mathematical statistics. 2nd ed. New York: Springer; 2003.
24. Jorion P. Value at risk: the new benchmark for managing financial risk. 3rd ed. Boston (MA): McGraw-Hill; 2007.
25. Hull J. Risk management and financial institutions. 2nd ed. New York: Prentice Hall; 2009.
26. Adrian T, Brunnermeier MK. CoVaR. Federal Reserve Bank of New York Staff Reports; 2008. p. 348.
27. Artzner P, Delbaen F, Eber JM, *et al.* Coherent measures of risk. *Math Finance* 1999;9(3): 203–228.
28. Delbaen F. Coherent risk measures on general probability spaces. In: Sandmann K, Schönbucher PJ, editors. *Advances in Finance and Stochastics-essays in Honour of Dieter Sondermann*. New York: Springer; 2002.
29. Pflug G. Some remarks on the value-at-risk and the conditional value-at-risk. In: Uryasev S, editor. *Probabilistic constrained optimization: methodology and applications*. Boston (MA): Kluwer Academic Publishers; 2000.
30. Rockafellar RT, Uryasev S. Optimization of conditional value-at-risk. *J Risk* 2000;2(3): 21–41.
31. Acerbi C, Tasche D. On the coherence of expected shortfall. *J Bank Finance* 2002;26(7): 1487–1503.
32. Rockafellar RT, Uryasev S. Conditional value-at-risk for general loss distributions. *J Bank Finance* 2002;26(7):1443–1471.
33. Kusuoka S. On law invariant coherent risk measures. *Adv Math Econ* 2001;3:83–95.

34. Föllmer H, Schied A. Convex measures of risk and trading constraints. *Finance Stoch* 2002;6(4):429–447.
35. Frittelli M, Gianin ER. Putting order in risk measures. *J Bank Finance* 2002;26(7):1473–1486.
36. Denneberg D. Non-additive measure and integral. Boston (MA): Kluwer Academic Publishers; 1994.
37. Dhaene J, Vanduffel S, Goovaerts MJ, *et al.* Risk measures and comonotonicity: a review. *Stoch Models* 2006;22(4):573–606.
38. Song Y, Yan J-A. The representations of two types of functionals on $L^\infty(\Omega, \mathcal{F})$ and $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. *Sci China Ser A Math* 2006;49:1376–1382.
39. Song Y, Yan J-A. Risk measures with comonotonic subadditivity or convexity and respecting stochastic orders; 2006. Working paper, Chinese Academy of Sciences.
40. Kou SG, Peng X, Heyde CC. What is a good external risk measure: bridging the gaps between robustness, subadditivity, and insurance risk measures; 2006. Working paper, Columbia University.
41. Hart HLA. The concept of law. 2nd ed. Oxford: Clarendon Press; 1994.
42. Heyde CC, Kou SG. On the controversy over tailweight of distributions. *Oper Res Lett* 2004;32(5):399–408.
43. Eling M, Tibiletti L. Internal vs. external risk measures: how capital requirements differ in practice; 2008. Working paper, University of St. Gallen.
44. Samuelson PA. General proof that diversification pays. *J Financ Quant Anal* 1967;2(1):1–13.
45. Brumelle SL. When does diversification between two investments pay? *J Financ Quant Anal* 1974;9(3):473–482.
46. McMin RD. A general diversification theorem: a note. *J Finance* 1984;39(2):541–550.
47. Hong C-S, Herk LF. Increasing risk aversion and diversification preference. *J Econ Theory* 1996;70(1):180–200.
48. Kijima M. The generalized harmonic mean and a portfolio problem with dependent assets. *Theory Decis* 1997;43(1):71–87.
49. Ibragimov R, Walden J. The limits of diversification when losses may be large. *J Bank Finance* 2007;31(8):2551–2569.
50. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;47(2):263–291.
51. Daniélsson J, Jorgensen BN, Samorodnitsky G, *et al.* Subadditivity re-examined: the case for Value-at-Risk; 2005. Working paper, London School of Economics. Available at: <http://ideas.repec.org/p/fmg/fmgdps/dp549.html>.
52. Garcia R, Renault E, Tsafack G. Proper conditioning for coherent VaR in portfolio management. *Manage Sci* 2007;53(3):483–494.
53. Dhaene J, Goovaerts MJ, Kaas R. Economic capital allocation derived from risk measures. *North Am Actuar J* 2003;7(2):44–59.
54. Quiggin JC. A theory of anticipated utility. *J Econ Behav Org* 1982;3(4):323–343.
55. Ahmed S, Filipović D, Svindland G. A note on natural risk statistics. *Oper Res Lett* 2008;36:662–664.

ROBUST OFFLINE SINGLE-MACHINE SCHEDULING PROBLEMS

BITA TADAYON

University of Florida,
Gainesville, FL

J. COLE SMITH

Clemson University, Clemson, SC

The single-machine scheduling problems (SMSPs) that we examine in this article seek to schedule nonpreemptive jobs on a single machine, which is capable of performing only one task at a time. Defining J as the set of all jobs, each job $j \in J$ is typically associated with data attributes such as processing time (p_j), weight (w_j), and due date (d_j). The objective of a schedule depends on the completion time (C_j) of each job $j \in J$. Define L_j (lateness) as $C_j - d_j$, T_j (tardiness) as $\max\{0, C_j - d_j\}$, and U_j as a binary indicator value equal to 1 if and only if $C_j - d_j > 0$. Common single-machine scheduling objectives include minimizing total completion time ($\sum_{j \in J} C_j$), total weighted completion time ($\sum_{j \in J} w_j C_j$), and the number of late jobs ($\sum_{j \in J} U_j$). In addition, defining $L_{\max} = \max_{j \in J} \{L_j\}$ (and $T_{\max}$ and $C_{\max}$ analogously), other common objectives include minimizing $L_{\max}$, $T_{\max}$, and $C_{\max}$. (The latter objective is also referred to as *makespan*.)

Deterministic scheduling problems assume that the exact values of all parameters are known. As a result, it is straightforward to calculate the job completion times and objective function value corresponding to each sequence. Moreover, each problem listed above is polynomially solvable using deterministic data [1–3], assuming that the release times for all jobs are zero.

Owing to the variability of process and environmental data, uncertainty is common in many practical scheduling problems. Therefore, several researchers have developed methods to hedge against data uncertainty in scheduling problems.

Reactive (online) scheduling deals with adjusting job schedules as data is realized in order to reduce the effect of disruptions and unpredicted delays. Using this approach, the scheduler generates initial schedules without considering uncertainty, and then revises the schedule as disruptions occur. Online scheduling is useful for situations in which limited information about uncertain data is available in advance, and it is practical for the scheduler to dynamically adjust the schedules. This reactive approach can accommodate a wide variety of disruptions (such as machine breakdowns, process interruptions, or parameter value variations) and is suitable for problems in which disruptions are difficult to predict. For example, process scheduling in an operating system can be modeled as an online scheduling problem as the operating system typically does not know the execution time of a process before its completion, and it is possible for the operating system to interleave the execution of different tasks by temporarily interrupting a task and resuming its execution at a later time (preemption). Another example regards the scheduler component of a web server, where the number of users and the length of each user's tasks is difficult to predict. See Ref. 4 for an overview of online scheduling and a survey of results obtained in this area, and [5] for an example of online single-machine scheduling.

However, it is not always realistic to assume that schedules can be revised after disruptions. In these situations, a predictive (offline) scheduling approach prescribes solutions that are relatively insensitive to changes in input data. An implicit assumption made in offline scheduling is that all uncertain data values are realized after the decisions have been made. Stochastic programming and robust optimization are two important offline modeling schemes that have been applied for solving several scheduling problems under uncertainty. Stochastic optimization, which was first

introduced by Dantzig [6], assumes knowledge of data probability distributions. This method seeks to optimize the expected objective function value, while ensuring that constraints are satisfied at least with sufficiently high probability. We refer the interested reader to textbooks [7–9] and the references therein for stochastic programming background. Pinedo and Schrage [10] present a survey of stochastic optimization applications in solving scheduling problems.

The expected objective function value is not the only metric by which schedules are evaluated when data is uncertain. Sarin *et al.* [11] study the problem of selecting schedules based both on the expectation of their objective function values and the variation of their objectives, given distributions on the uncertain parameter values. Stochastic scheduling studies also consider optimization metrics such as value-at-risk (VaR) [12] or conditional-value-at-risk (CVaR) [13]. A related strategy may instead associate so-called chance constraints with stochastic scheduling problems. The work of van den Akker and Hoogeveen [14] examines problems with due dates and requires that the probability of finishing each job by its due date be sufficiently large (or else the job is counted as late). They seek to determine sequences that minimize the number of late jobs—as defined in the chance-constrained sense—given various assumptions on the job processing time distributions. Deng and Shen [15] consider a related problem in a multiserver setting, where each server has some maximum time past which it should not operate. If a server operates past its maximum time, the server is said to be working overtime. The challenge is to schedule jobs, each of which has a certain release time and stochastic processing time, to servers in order to minimize scheduling costs, while ensuring that the probability that a server works overtime is sufficiently small.

Robust optimization is an alternative approach for dealing with uncertain data [16, 17]. In robust optimization, data uncertainty is usually represented by continuous or discrete uncertainty sets, and the solution must remain feasible with respect to any

possible data outcome within the uncertainty set. Two main factors motivate the use of robust optimization. First, it is not always possible to estimate data probability distributions with desired precision. Second, in stochastic programming, the problem size increases drastically with the number of uncertain parameters, which induces substantial computational challenges. Another essential difference between the stochastic programming and robust optimization approaches is that stochastic programming models typically aim at optimizing expected system performance, while robust optimization models focus on guaranteeing a minimum quality for the solution in the worst case.

In the deterministic scheduling literature, problems are classified into different categories based on the machine environment (α), job characteristics (β), and the optimization criterion (γ), and we assume that the exact values of all parameters are known. Single machine (1), m -machine flow shop (F_m), and m -machine job shop (J_m) are examples of common machine environments (α), while having precedence relations between jobs (prec) and having families of jobs with similar characteristics (fml) are examples of job characteristics (β). Any minimization criterion, such as $C_{\max}$, $\sum C_j$, and $\sum U_j$, is an example of γ . Accordingly, we can specify every deterministic scheduling problem using Graham’s notation $\alpha|\beta|\gamma$ suggested in Ref. 18. In robust optimization, however, we assume that different job-related parameters such as p_j , d_j , and w_j are uncertain, and we optimize with respect to the worst-case data realization using a min–max objective. Therefore, we also specify the uncertain parameters in the notation, denoted by η . We specify the robustness measure for the problem, ν , which we explain in detail in the section titled “Robustness Measures” and describe a robust scheduling problem using the notation $\text{MinMax}_\nu(\alpha|\beta|\gamma, \eta)$.

In this article, we review robust optimization techniques that have been used for solving standard scheduling problems under uncertainty. Note that the majority of research in this area is focused on SMSPPs. Thus, we explore and classify the literature

of robust SMSP, while addressing some other standard scheduling problems that have been investigated.

The remainder of this article is organized as follows. In the section titled “Robustness and Uncertainty Definitions,” we describe different robustness criteria and uncertainty representations used in the robust optimization literature and address the use of each method in scheduling. Then in the section titled “Robust Single-Machine Scheduling Literature Classification,” we discuss the details of existing research in the area of robust SMSP, classify them according to the categories presented in section titled “Robustness and Uncertainty Definitions,” and explore the existing gaps and open problems. We then address related robust scheduling results in the section titled “Robust Optimization in Other Scheduling Problems” and conclude in the section titled “Conclusion and Open Research Questions.”

ROBUSTNESS AND UNCERTAINTY DEFINITIONS

To define a robust scheduling problem, we need to represent the potential values of uncertain parameters in the problem and specify a measure by which we evaluate the robustness of a particular solution. In this section, we present the most common robustness measures and uncertainty representations for robust optimization in general and for robust scheduling in particular. We also discuss the benefits, limitations, and potential applications for each scheme.

Define a sequence, π , as a permutation of jobs and denote the set of all possible permutations by Π . A scenario, ϕ , is a particular realization of uncertain parameters, where Φ represents the (possibly infinite-cardinality) set of all possible scenarios. Let Z_π^ϕ be the objective function value of job sequence π under data realization ϕ . Define Z^ϕ as the best achievable (optimal) objective function value when data scenario ϕ is realized. We use this notation to present robustness and uncertainty definitions in the following two subsections.

Robustness Measures

We discuss the three most common robustness measures that arise in the scheduling field: absolute robustness, robust deviation, and relative robust deviation [19]. (See also Sabuncuoglu and Goren [20] for a more detailed classification for possible robustness and stability measures, along with applications of some measures in solving robust SMSPs.)

Absolute robustness seeks to minimize the maximum objective function value over all scenarios. That is, when dealing with the absolute robustness measure, we associate a worst-case scenario with each possible job sequence and select a sequence whose worst-case objective value is minimum among all feasible sequences. We can mathematically state absolute robustness as $\min_{\pi \in \Pi} \max_{\phi \in \Phi} Z_\pi^\phi$. This measure is particularly useful for situations in which one seeks to guarantee at least a certain quality for the solution over all possible scenarios. In our modified Graham’s notation, the subscript ν is blank for the absolute robustness case, as this problem directly minimizes the maximum possible scheduling objective that can occur (and as such is the “default” robustness measure).

To understand the *robust deviation* measure, we first introduce the concept of regret. In the context of robust scheduling, a decision-maker chooses a schedule before observing the data values. After the data is observed, the decision-maker’s regret is given by the difference between his/her chosen schedule’s objective and the objective of the retrospective optimal solution, that is, the optimal solution given knowledge of the data outcome. Robust deviation seeks to minimize the largest possible regret and can be stated as $\min_{\pi \in \Pi} \max_{\phi \in \Phi} (Z_\pi^\phi - Z^\phi)$. Minimizing robust deviation yields a schedule whose performance, compared to the corresponding optimal performance, is relatively insensitive to data realization. Solutions that minimize robust deviation can also be interpreted as *uniformly suboptimal* solutions, that is, ϵ -optimal solutions for all data realizations, with ϵ as small as possible. In our scheduling problem notation, $\nu = \text{dev}$ for this case.

Relative robust deviation minimizes the maximum relative (or percentage) deviation from optimality, that is, $\min_{\pi \in \Pi} \max_{\phi \in \Phi} ((Z_{\pi}^{\phi} - Z^{\phi})/Z^{\phi})$. In fact, relative robust deviation is a normalized robust deviation measure that seeks to minimize the relative regret in the problem. As the value of relative regret only depends on the ratio of objective values, it can be used to create benchmarks to compare the quality of different problems' solutions. In the case of relative robust deviation, we denote $\nu = \text{rel}$.

Robust deviation and relative robust deviation measures are appropriate for environments in which the quality of solutions are evaluated after the data is realized. In such cases, the (relative) deviation of the selected decision from the optimal decision for the realized scenario is a plausible quality measure. These measures may be particularly important in the context of highly competitive markets, where a firm needs to have a satisfactory performance compared to its competitors, under any potentially realizable scenario [19].

Before proceeding, it is useful to emphasize the robust optimization paradigm, and how the objectives and uncertainty sets interact in the optimization scheme. These problems can be described as Stackelberg games [21], which involve a leader and a follower. In the robust optimization context, the leader determines a schedule with the objective of minimizing an objective. The follower, with knowledge of the leader's selected job permutation, selects data values to make the solution as poor as possible, either by maximizing the objective function (absolute robustness) or regret (robust deviation or relative robust deviation). The leader anticipates the follower's optimal action and selects a schedule that minimizes the largest objective the follower can obtain.

Example 1 To illustrate the difference between robustness measures, consider an example in which we seek to find a two-job schedule that minimizes the total completion time, where $p_1 \in [4, 5]$ and $p_2 \in [1, 6]$. If we apply the absolute robustness measure, the worst-case scenario for any sequence occurs when $p_1 = 5$ and $p_2 = 6$. Therefore,

the robust optimal solution is obtained by scheduling job 1 before job 2 (sequence (1, 2)). However, when robust deviation is applied, the largest regret value equals 4 for sequence (1, 2) (when $p_1 = 5$ and $p_2 = 1$) and equals 2 for sequence (2, 1) (when $p_1 = 4$ and $p_2 = 6$). Therefore, scheduling job 2 before job 1 optimizes robust deviation. Note that using relative robust deviation for this instance also results in sequence (2, 1) being optimal (the largest relative regret value equals $4/7$ for sequence (1, 2) and $2/14$ for sequence (2, 1)). However, robust deviation and relative robust deviation measures may result in different optimal solutions in other instances. For example, suppose that we modify the above example by letting $p_1 \in [4, 5]$ and $p_2 \in [2, 8]$. Given sequence (1, 2), the largest absolute regret value (3) is obtained when $p_1 = 5$ and $p_2 = 2$. The objective of sequence (1, 2) is 12, while the optimal objective value for this scenario is 9. Sequence (2, 1), on the other hand, results in a maximum regret value of 4, and therefore sequence (1, 2) is preferable with respect to the robust deviation measure. However, the relative regret value is $3/9$ for sequence (1, 2) and $4/16$ for sequence (2, 1). Thus, sequence (2, 1) is optimal under the relative robust deviation measure.

Note in Example 1 that the p_1 and p_2 values are confined to simple, independent intervals, that is, the follower maximizes over a rectangular feasible region. More research has recently been performed in this area to explore the impact of using alternative feasible regions for the follower, as described in the section titled "Uncertainty Representation."

Uncertainty Representation

In the robust optimization literature, several methods of expressing uncertain parameter values have been proposed. In this section, we discuss four main methods of uncertainty representation, present benefits and drawbacks of each method, and address some applications of each method in robust scheduling.

The original method of representing uncertainty in robust optimization problems

was developed by Soyster [22]. Soyster assumes that each uncertain parameter, independent of all the others, can take on any values within a continuous interval and proposes a linear optimization model to solve the problem. These uncertainty sets are given the label *interval uncertainty*. The simplicity of this method and its resulting formulations motivate its extensive application in robust scheduling problems. See Refs 23–26 as examples of using interval uncertainty to represent the values of SMSP parameters having different optimization criteria.

Note that if data is represented by interval uncertainty, and we seek to minimize absolute robustness, setting all parameters to their worst-case values will generate a worst-case scenario corresponding to any sequence of jobs. Therefore, robust SMSPs having the absolute robustness criterion under interval uncertainty is equivalent to a deterministic SMSP in which all data elements take on their worst-case values. However, generating a worst-case scenario when either the robust deviation or the relative robust deviation measure is applied is less obvious in general. Recall that in Example 1, robust deviation and relative robust deviation were maximized by setting either p_1 to its maximum value and p_2 to its minimum value, or vice versa. This is more complex than simply setting processing times to their largest possible values.

Although interval uncertainty results in simple robust formulations for several problems, it tends to generate schedules that are too conservative in the case of absolute robustness. Ben-Tal and Nemirovski [27–29] and El-Ghaoui and Lebret [30] and El-Ghaoui *et al.* [31] address this issue and introduce a method to reduce the level of conservatism by confining the data to belong to uncertainty sets in the form of ellipsoids. They propose efficient algorithms to solve the resulting convex optimization problems.

Bertsimas and Sim [32, 33] propose *budgeted uncertainty* as an alternative approach to control the level of conservatism for general robust optimization problems. This method assumes that while all uncertain parameters are bounded within a specified

interval, the total amount of variation in parameters from their “ideal values” is limited. Using this method to represent uncertain objective function and constraint coefficients of an optimization problem, they propose a robust programming problem of moderately larger size. The authors prove that using budgeted uncertainty, a robust linear programming problem can still be solved as a linear program, and the robust counterpart of a polynomially solvable 0–1 linear optimization problem remains polynomially solvable.

To the best of our knowledge, the budgeted uncertainty model is applied to robust SMSP in only one paper [34], which we discuss in the section titled “Robust Single-Machine Scheduling Literature Classification.” However, budgeted uncertainty has been directly applied in several other scheduling problems in the area of chemical process scheduling, as we discuss in the section titled “Robust Optimization in Other Scheduling Problems.”

A disadvantage of interval uncertainty models is that they do not capture correlations among data. Budgeted or ellipsoidal uncertainty can implicitly capture some correlations by limiting total parameter deviation from their norms. Moreover, because these modeling strategies essentially employ strong duality principles to the follower’s maximization problem, it is also generally easy to obtain robust formulations for the case in which uncertainty sets can be expressed as a convex set. Still, when such a convex set is not easy to obtain, or correlations among parameters yield a nonconvex uncertainty set, an alternative way of modeling uncertainty is needed. One way to represent these correlations is to enumerate all possible scenarios and represent uncertain parameters as a set of discrete scenarios for their numerical values. We refer to this method as *scenario uncertainty*. In the robust SMSP literature, using scenario uncertainty is common [35–38]. This method allows the decision-maker to consider the relationship between all uncertain factors in the scheduling environment and include all possible cases in the model. It also maintains control over the level of conservatism in the problem. However, this approach potentially

requires enumeration of a large number of data outcomes, which is computationally intractable in some cases.

Example 2 To illustrate the foregoing concepts, consider a three-job problem in which we seek to minimize weighted completion time under the absolute robustness objective, with uncertainty in the processing times (problem $\text{MinMax}(1 \parallel \sum w_j C_j, p_j)$). This example focuses on the case of budgeted uncertainty. In particular, defining $\hat{p}_1 = 8$, $\hat{p}_2 = 7$, and $\hat{p}_3 = 4$, the uncertainty set constraints bounding the p values are given by:

$$|p_j - \hat{p}_j| \leq 2, \forall j = 1, 2, 3, \text{ and } \sum_{j=1}^3 \frac{|p_j - \hat{p}_j|}{\hat{p}_j} \leq 0.6. \quad (1)$$

That is, absolute deviation from the “ideal” processing times ($\hat{p}$) is no more than 2, and the total percentage deviation is limited by 0.6. (This is one particular example of budgeted uncertainty; it is also possible to have individual deviations limited as a percentage of the ideal processing times, and/or the left-hand-side of the total deviation constraint refer to absolute deviation rather than percent deviation.) The weights for these jobs are deterministic and given by

$$w_1 = 9, w_2 = 7, w_3 = 4. \quad (2)$$

For the case of absolute robustness, Table 1 shows the six feasible sequences. In addition, for each sequence, the associated worst-case p values are depicted, alongside the resulting objective. Note that there are two sequences—(1, 2, 3) and (1, 3, 2)—that would be optimal in the deterministic version of the problem, where $p_j = \hat{p}_j$ for all j . Only the latter is optimal when minimizing absolute robustness.

Table 1 shows that each worst-case p vector is an extreme point to the uncertainty set given by Equation 1. This is as expected, because the follower (inner maximization) problem seeks to optimize a linear function over a polyhedral set. Another conjecture might be that an optimal sequence must correspond to some schedule generated

Table 1. Minimization of Absolute Robustness in Example 2

Sequence	Worst-Case Values			Objective
	p_1	p_2	p_3	
(1, 2, 3)	10	9	4.257	316.029
(1, 3, 2)	10	9	4.257	309.829
(2, 1, 3)	10	9	4.257	327.029
(2, 3, 1)	10	9	4.257	325.343
(3, 1, 2)	10	7	5.4	317
(3, 2, 1)	8	9	5.257	321.143

Optimal sequence: (1, 3, 2), objective 309.827

according to the weighted shortest processing time (WSPT) rule, where processing times correspond to the ideal processing times $\hat{p}$. (In the case above, the optimal sequence of (1, 3, 2) corresponds to one of the two WSPT sequences.) However, suppose that we now take $w_3 = 4 + \Delta$. As Δ becomes negative, the WSPT rule would schedule the jobs in order (1, 2, 3), which is suboptimal for the absolute robustness problem above. It turns out that for any nonpositive Δ , the worst-case p values for sequences (1, 2, 3) and (1, 3, 2) remain as depicted in Table 1. Thus, for $\Delta \leq 0$, the worst-case costs for these sequences are given by:

$$\text{Cost for sequence (1, 2, 3)}: 316.028 - 23.257\Delta$$

$$\text{Cost for sequence (1, 3, 2)}: 309.827 - 14.257\Delta. \quad (3)$$

For $\Delta < -0.689$, sequence (1, 2, 3) replaces (1, 3, 2) as the optimal robust solution.

As a final observation, note that the worst-case p values do change as Δ becomes positive. For sequence (1, 3, 2), the worst-case p values change from $p = (10, 9, 4.257)$ for $\Delta \leq 1.25$ to $p = (10, 7, 5.4)$ for $\Delta \geq 1.25$. (However, at $\Delta = 1.25$, sequence (3, 1, 2) becomes optimal, as sequence (1, 3, 2) is optimal only up to $\Delta < 0.81$.)

Example 3 We next reconsider Example 2 for problem $\text{MinMax}_v(1 \parallel \sum w_j C_j, p_j)$, for v corresponding to robust deviation and relative robust deviation. All data for this example is the same as in Example 2. Table 2 lists all six possible sequences.

Table 2. Minimization of Robust Deviation in Example 3

Sequence	Worst-Case Values			Objective	Retrospective Optimal		Robust Deviation
	p_1	p_2	p_3		Sequence	Objective	
(1, 2, 3)	10	7	2.6	287.4	(3, 2, 1)	254	33.4
(1, 3, 2)	10	5	3.743	276.173	(2, 3, 1)	238.659	37.514
(2, 1, 3)	6	9	3.743	272.972	(1, 3, 2)	224.173	48.799
(2, 3, 1)	6	9	4.257	289.341	(1, 3, 2)	229.827	59.514
(3, 1, 2)	8	5	5.257	268.14	(2, 1, 3)	225.028	43.112
(3, 2, 1)	6	7	5.4	274	(1, 2, 3)	218.6	55.4

Optimal sequence: (1, 2, 3), objective 33.4

For each sequence, a worst-case set of p values is given that maximizes robust deviation. In column *objective*, the resulting objective from the chosen sequence is given. The *retrospective optimal* columns depict the optimal sequence corresponding to the worst-case p values, along with the objective corresponding to that sequence. The *robust deviation*, or regret, of the solution is shown in the last column, which is given by the difference in the objective corresponding to the original sequence and that corresponding to the retrospective optimal solution. Note that in this case, some components of the worst-case p values may be smaller than their ideal values, in order to force a maximum difference between the objective of the initial sequence and that of the retrospective optimal solution.

For the relative robust deviation case, Table 3 provides an optimization analysis similar to that in Table 2. Sequence (1, 2, 3) remains optimal in this case. Interestingly, however, both the worst-case set of p values and the retrospective optimal solution

corresponding to (1, 2, 3) changes in the relative robust deviation case. With respect to sequence (2, 3, 1), p_3 equals 4.257 in the robust deviation case and equals 3.743 in the relative robust deviation case. All worst-case p values and retrospective sequences remain unchanged for the other four possible sequences.

Finally, while it is not a surprise that sequence (1, 2, 3) would be optimal for both the robust deviation and relative robust deviation measures, the same sequence need not optimize both of these measures in general. For instance, suppose that we revise $w_3 = 4.35$. In this case, sequence (1, 2, 3) is no longer optimal under either objective. Table 4 shows that robust deviation is minimized by sequence (3, 1, 2), while relative robust deviation is minimized by sequence (1, 3, 2). Importantly, for these two sequences, the same p vector that maximizes robust deviation also maximizes relative robust deviation. Hence, we only report one worst-case set of p values in Table 4 for each sequence, even though as demonstrated above, different p values may correspond to the same initial

Table 3. Minimization of Relative Robust Deviation in Example 3

Sequence	Worst-Case Values			Objective	Retrospective Optimal		Robust Deviation
	p_1	p_2	p_3		Sequence	Objective	
(1, 2, 3)	8	7.7	2	252.7	(3, 1, 2)	221.9	0.139
(1, 3, 2)	10	5	3.743	276.173	(2, 3, 1)	238.659	0.157
(2, 1, 3)	6	9	3.743	272.972	(1, 3, 2)	224.173	0.218
(2, 3, 1)	6	9	3.743	282.659	(1, 3, 2)	224.173	0.261
(3, 1, 2)	8	5	5.257	268.14	(2, 1, 3)	225.028	0.192
(3, 2, 1)	6	7	5.4	274	(1, 2, 3)	218.6	0.253

Optimal sequence: (1, 2, 3), objective 0.139

Table 4. Example Illustrating Different Optimal Solutions to Robust Deviation and Relative Robust Deviation Objectives when $w_3 = 4.35$

Sequence	Worst-Case Values			Objective	Retrospective Optimal		Robust Deviation	Relative Robust Deviation
	p_1	p_2	p_3		Sequence	Objective		
(1, 3, 2)	10	5	3.743	280.983	(2, 3, 1)	241.719	39.264	0.162
(3, 1, 2)	8	5	5.257	269.98	(2, 1, 3)	231.42	38.56	0.167

Optimal sequence: (3, 1, 2) for robust deviation and (1, 3, 2) for relative robust deviation

sequence depending on whether robust deviation or relative robust deviation is being considered.

ROBUST SINGLE-MACHINE SCHEDULING LITERATURE CLASSIFICATION

In this section, we review the robust single-machine scheduling literature and classify existing research in this area according to their robustness measure definition and uncertainty representation. Kasperski [39] summarizes some of the results obtained in the robust SMSP literature for different cases (specified by a robustness measure and an uncertainty representation) and introduces open cases in this area. A short survey of the results is also presented in Ref. 40. Here, we provide an updated discussion of the results obtained in each category of SMSPs.

As stated in the section titled “Uncertainty Representation,” the case of absolute robustness when dealing with continuous intervals of uncertainty can be equivalently solved as a deterministic problem and therefore is out of the scope of this study. However, when parameter values are constrained beyond their simple interval bounds, as in budgeted uncertainty, there is no obvious equivalent deterministic formulation for the problem. Tadayon and Smith [34] define three alternative uncertainty sets for the absolute robust SMSP when processing time values are presented as independent continuous intervals. In uncertainty sets 1, 2, and 3, they, respectively, require the total delay, the number of delayed jobs, and the total ratio by which the processing times are increased to be no more than a

budget value. They study the complexity of the problem $\text{MinMax}(1 \| Z, p_j)$, where $Z \in \{\sum C_j, \sum w_j C_j, L_{\max}, T_{\max}, \sum U_j\}$ under each uncertainty set, propose exact algorithms for polynomially solvable problems, and present mixed-integer programming formulations for the NP-hard problems and the ones having an unknown complexity.

Absolute robust SMSP in presence of scenario uncertainty has been studied in several papers. Yang and Yu [41] prove that problem $\text{MinMax}_v(1 \| \sum C_j, p_j)$ with scenario uncertainty is NP-hard for all three robustness measures (even in the special case having only two scenarios). They also present an exact dynamic programming algorithm with exponential complexity and two polynomial-time heuristics to solve this robust SMSP.

Aloulou and Della Croce [35] study several other absolute robust SMSPs with scenario uncertainty in different job-related parameters. When the scheduler seeks to minimize the number of late jobs, they prove that problems $\text{MinMax}(1 \| \sum U_j, p_j)$ and $\text{MinMax}(1 \| \sum U_j, p_j, d_j)$ are NP-hard, but the case of $\text{MinMax}(1 \| \sum U_j, d_j)$ is still open. In addition, when the objective function seeks to minimize $f_{\max} \in \{C_{\max}, L_{\max}, T_{\max}\}$ and precedence relations between jobs in the sequence are allowed, they propose polynomial algorithms for solving the problem in which p_j , d_j , or both are expressed as a set of discrete scenarios. In addition, they prove that problem $\text{MinMax}(1 \| \sum w_j C_j, w_j)$ with scenario uncertainty is NP-hard. (The NP-hardness of $\text{MinMax}(1 \| \sum w_j C_j, p_j)$ also follows from the NP-hardness of $\text{MinMax}(1 \| \sum C_j, p_j)$, which was proved in Ref. 41.)

Mastrolilli *et al.* [38] seek to generate approximation schemes for problem $\text{MinMax}(1 \parallel \sum w_j C_j, w_j, p_j)$. They prove that the problem cannot be approximated within $\mathcal{O}(\log^{1-\epsilon} n)$ for any $\epsilon > 0$, unless quasi-polynomial algorithms exist for NP. For the special case of unweighted jobs, they propose a two-approximation algorithm for solving $\text{MinMax}(1 \parallel \sum C_j, p_j)$ and prove that the problem is NP-hard to approximate within a factor less than 6/5.

In contrast to the extensive application of absolute robustness in the robust optimization literature, robust deviation has been investigated more widely in the SMSP literature. (This difference may in part be due to the ease of solving absolute robustness problems under the interval uncertainty model.) When independent continuous intervals are used to represent uncertain processing times, Lebedev and Averbakh [42] prove that the general case of problem $\text{MinMax}_{\text{dev}}(1 \parallel \sum C_j, p_j)$ is NP-hard; however, they show that when all intervals of uncertainty have the same center, the problem can be solved polynomially (in $\mathcal{O}(n \log n)$ time) if the number of jobs is odd, and is NP-hard otherwise. Montemanni [26] presents the first mixed-integer linear programming formulation for problem $\text{MinMax}_{\text{dev}}(1 \parallel \sum C_j, p_j)$ with interval uncertainty and translates some preprocessing rules into valid inequalities.

Kasperski and Zielinski [43] prove that any minmax robust deviation SMSP with total completion time criterion and interval uncertainty, whose equivalent deterministic problem is polynomially solvable, can be approximated within a factor of 2. A more general case of this problem with weighted sum of completion time criterion ($\text{MinMax}_{\text{dev}}(1 \parallel \sum w_j C_j, p_j)$) has been studied in Ref. 44 in which some dominance relations are introduced and a heuristic algorithm is proposed to find a good nondominated sequence.

Robust deviation in the case of interval uncertainty is also considered in Refs 23, 24, and 25 for different SMSPs. Averbakh [23] proves that $\text{MinMax}_{\text{dev}}(1 \parallel \max\{w_j T_j\}, w_j)$ is polynomially solvable by presenting an $\mathcal{O}(n^3)$ algorithm for the problem. Kasperski

[24] proposes a polynomial algorithm (with $\mathcal{O}(n^4)$ complexity) for optimally solving $\text{MinMax}_{\text{dev}}(1 \parallel \text{prec}|L_{\max}, p_j, d_j)$. Lu *et al.* [25] define a specific SMSP with total completion time objective function in which jobs are grouped into families and a sequence-dependent family setup time exists. They prove that when both job processing times and family setup times are presented as intervals of uncertainty, minimizing robust deviation in the corresponding problem is NP-hard. They then propose a simulated annealing-based algorithm to find good quality solutions for the problem.

Several researchers study robust deviation in SMSPs in which parameters are presented as discrete scenarios. Daniels and Kouvelis [36] examine problems $\text{MinMax}_{\text{dev}}(1 \parallel \sum C_j, p_j)$ and $\text{MinMax}_{\text{rel}}(1 \parallel \sum C_j, p_j)$ when the uncertainty set consists of discrete scenarios for the p_j values. They prove that both problems are NP-hard and present some dominance relations that can be used to determine the relative job positions in a robust schedule. Then, they use these results to develop an exact branch-and-bound algorithm for solving the robust problem. They also present intuitive heuristic approaches and evaluate their efficiency and accuracy through a set of randomly generated instances. In addition, Daniels and Kouvelis prove that even when processing times are presented as independent continuous intervals, the worst-case scenario for each sequence π belongs to the finite set of extreme point scenarios (scenarios in which each parameter takes on its lower- or upper-bound values). According to this result, they claim that the formulations presented for scenario uncertainty are also valid for interval uncertainty.

Yang and Yu [41] present a different NP-hardness proof for the problem $\text{MinMax}_{\text{dev}}(1 \parallel \sum C_j, p_j)$. A more general problem with the weighted sum of completion time criterion (using the robust deviation measure and scenario uncertainty in job processing times) is studied in Ref. 37, where a set of valid inequalities for the convex hull of its feasible region is presented and is then utilized to design a cutting-plane algorithm for solving the problem.

To the best of our knowledge, relative robust deviation has been explicitly considered in only three SMSP papers so far. Averbakh [45] studies this measure of robustness for robust combinatorial optimization problems in general and for SMSP as a special case and proves that the problem $\text{MinMax}_{\text{rel}}(1 \parallel \sum C_j, p_j)$ with interval uncertainty is NP-hard. Daniels and Kouvelis state that the analysis, results, and solution methods presented in Ref. 36 for the case of robust deviation can be applied for the relative robust deviation measure with a slight modification. Finally, Yang and Yu [41] prove the NP-hardness of problem $\text{MinMax}_{\text{rel}}(1 \parallel \sum C_j, p_j)$ with scenario uncertainty.

ROBUST OPTIMIZATION IN OTHER SCHEDULING PROBLEMS

Although robust optimization has been most frequently studied for the single-machine scheduling environment, other scheduling problems have been also addressed in the literature. Moreover, the concept of robustness in some studies extends beyond the standard definitions presented earlier in this article and includes (i) the restriction that a schedule remains feasible with a certain probability [46], (ii) that a schedule guarantees a certain performance level [47, 48]), or (iii) the incorporation of different uncertainties in scheduling environments (e.g., unpredictable production interruptions [49]). In this section, we introduce some of the other areas of scheduling in which different variations of robust optimization have been applied to hedge against uncertainty in production environments.

Kouvelis *et al.* [50] apply a robustness definition similar to the one presented in this article to comply with uncertainty in a two-machine flow shop environment. They prove that problem $\text{MinMax}_{\text{dev}}(F_2 \parallel C_{\text{max}}, p_j)$ is NP-hard in the ordinary sense. They also discuss the properties of robust schedules and develop exact and heuristic solution approaches for the problem under both interval uncertainty and scenario uncertainty in job processing times. Kasperski *et al.* [51] strengthens the result obtained in Ref. 50

by proving that $\text{MinMax}_{\text{dev}}(F_2 \parallel C_{\text{max}}, p_j)$ in the case of scenario uncertainty is strongly NP-hard and inapproximable, even for two scenarios. They also prove that the absolute robust version of the problem is strongly NP-hard, but approximable with performance ratio of 2 and not $(4/3 - \epsilon)$ -approximable for any $\epsilon > 0$ if the number of scenarios is a known constant.

Although ellipsoidal and budgeted uncertainty have not yet been widely addressed in the robust single-machine scheduling literature, it has been applied to other scheduling problems, mainly in the area of chemical process scheduling. Production scheduling, as one of the most important problems in industrial plant operations, has been studied in several research papers. Short-term production scheduling seeks to determine an optimal assignment of available resources to production tasks over time while satisfying the production requirements at due dates. Depending on the plant layout and the sequence of tasks, process scheduling is usually in the form of a standard flow shop or job shop scheduling problem. Ierapetritou and Floudas [52] propose a mixed-integer linear programming formulation for this problem.

Li and Ierapetritou [53] consider uncertainty in parameter values in the formulation presented in Ref. 52. They examine interval, ellipsoidal, and budgeted uncertainty to formulate the robust process scheduling problem given uncertainties in product unit price, task processing times, and product demand values. They compare the three formulations by solving test instances from linear formulations using CPLEX and the ones from nonlinear formulations using the DICOPT solver via GAMS. The results of their computations show that the budgeted uncertainty model may yield the most appropriate formulation, owing to the fact that it does not increase the problem size substantially and maintains the linearity of the model.

Lin *et al.* [54] and Janak *et al.* [46] also consider data uncertainty in the process scheduling problem as formulated in Ref. 52. Uncertain parameters (price, processing times, and

demands) are presented as independent continuous intervals in Ref. 54 and as random variables with known probability distributions in Ref. 46. Janak *et al.* [46] define a robust solution to be one, that is, feasible in all constraints with a certain probability. Mathematical formulations for the robust problems are presented in both papers and the efficiency of models are investigated by solving test instances from linear formulations using CPLEX and from nonlinear models using DICOPT.

Li and Ierapetritou [55] review the process scheduling under uncertainty literature. They introduce robust optimization as one of the alternative approaches in dealing with uncertainty in process scheduling and address the main advances and challenges in this area of research. Verderame *et al.* [56] provide a broader overview of planning and scheduling problems under uncertainty, across multiple sectors. Although the objective function and constraints of the model vary greatly among different sectors, they address the common attribute of requiring a standard robustness definition and uncertainty representation across all areas.

Leon *et al.* [49] define a different type of uncertainty in a standard job shop environment in which disruptions occur randomly during the process. They seek to find a sequence of jobs, that is, robust under various disruption scenarios, assuming that preemption is not allowed (when a job's process is interrupted, it has to be restarted), the performance measure is makespan, and the order of jobs cannot be revised after interruptions occur. They define the robustness measure as a randomly weighted sum of the expected makespan and the expected difference between the actual makespan and the deterministic makespan (without interruptions). Because the effect of each disruption depends on the outcome of all previous disruptions, they compute the robustness measure only in the case of a single interruption. For the case of several disruptions, they develop a surrogate method and embed it in a genetic algorithm to generate relatively robust schedules.

A robust solution in some applications is defined as a solution that guarantees a certain level of performance under all possible data realizations. Daniels and Carrillo [47] define a β -robust solution as a job sequence that maximizes the likelihood of achieving a total completion time value no greater than a constant value in a SMSP with uncertain processing times. They assume that job processing times are independent random variables with known means and variances that can be represented as discrete scenarios. They prove that the problem is NP-hard and develop an exact branch-and-bound algorithm and a polynomial heuristic for solving the problem. Wu *et al.* [48] consider a similar problem with independent normally distributed processing time values and present three models: primal (maximizing the probability of obtaining a certain level of performance), dual (minimizing the level that can be achieved with a fixed probability), and hybrid. They determine the feasibility of each model for a set of randomly generated problems and study the effect of some dominance rules in solving the problems.

COMPLEXITY SUMMARY

Table 5 provides a summary of the results obtained in the robust SMSP literature. In this table, we specify the complexity of robust SMSP problems under each robustness measure (absolute robustness “abs. rob.,” robust deviation “rob. dev.,” and relative robust deviation “rel. rob. dev.”) and each uncertainty representation (budgeted uncertainty “budg. uncert.,” interval uncertainty “interval,” and scenario uncertainty “scenario”), as well as the reference from which the result is extracted. Each problem is specified by its objective function and uncertain parameter in the first column labeled as “(obj., param.).” Note that budgeted uncertainty has been considered only for the absolute robustness criterion. Among the results obtained in Ref. 34, we only include the ones about the first uncertainty set in this table, for the sake of brevity. The cells whose corresponding problems have not been yet studied in the literature are marked by “–.”

Table 5. Complexity Results for Robust SMSP

(obj., param.)	abs. rob.		rob. dev.		rel. rob. dev.	
	budg. uncert.	Scenario	Interval	Scenario	Interval	Scenario
$(\sum C_j, p_j)$	$\mathcal{O}(n \log n)$ [34]	NP-hard [41]	NP-hard [42]	NP-hard [36, 41]	NP-hard [45]	NP-hard [36, 41]
$(\sum w_j C_j, p_j)$	Open [34]	NP-hard [41]	NP-hard [42]	NP-hard [36, 41]	NP-hard [45]	NP-hard [36, 41]
$(\sum w_j C_j, w_j)$	–	NP-hard [35]	–	–	–	–
$(\sum w_j C_j, \{p_j, w_j\})$	–	NP-hard [41]	NP-hard [42]	NP-hard [36, 41]	NP-hard [45]	NP-hard [36, 41]
$(\sum U_j, p_j)$	$\mathcal{O}(n^2)$ [34]	NP-hard [35]	–	–	–	–
$(\sum U_j, d_j)$	–	Open [35]	–	–	–	–
$(\sum U_j, \{p_j, d_j\})$	–	NP-hard [35]	–	–	–	–
$(L_{\max}, p_j)$	$\mathcal{O}(n \log n)$ [34]	$\mathcal{O}(n^2 S)$ [35]	$\mathcal{O}(n^4)$ [24]	–	–	–
$(L_{\max}, d_j)$	–	$\mathcal{O}(n^2 + n S)$ [35]	$\mathcal{O}(n^4)$ [24]	–	–	–
$(L_{\max}, \{p_j, d_j\})$	–	$\mathcal{O}(n^2 S)$ [35]	$\mathcal{O}(n^4)$ [24]	–	–	–
$(T_{\max}, p_j)$	$\mathcal{O}(n \log n)$ [34]	$\mathcal{O}(n^2 S)$ [35]	–	–	–	–
$(T_{\max}, d_j)$	–	$\mathcal{O}(n^2 + n S)$ [35]	–	–	–	–
$(T_{\max}, \{p_j, d_j\})$	–	$\mathcal{O}(n^2 S)$ [35]	–	–	–	–
$(C_{\max}, p_j)$	–	$\mathcal{O}(n^2 S)$ [35]	–	–	–	–
$(\max w_j T_j, w_j)$	–	–	$\mathcal{O}(n^3)$ [23]	–	–	–

CONCLUSION AND OPEN RESEARCH QUESTIONS

We categorize robust scheduling problems according to three main criteria: The deterministic scheduling problem (given by the Graham notational parameters α, β, γ) from which it is derived, the robustness measure (absolute robustness, robust deviation, or relative robust deviation), and the uncertainty sets that confine the uncertain parameters. In particular, the uncertainty sets typically consist either of a discrete set of scenarios that list every possible simultaneous data outcome, or a continuous region (described by a system of inequalities) to which the data can belong. In the latter case, interval uncertainty refers to the situation in which each data value is independently limited by a minimum and/or maximum value, whereas budgeted or ellipsoidal uncertainty are special cases in which inequalities confining the data parameters are more complex than simple upper or lower bounds.

From a modeling perspective, scenario uncertainty allows a decision-maker to capture correlations among data, but may require the enumeration of many scenarios to become practically useful. Budgeted or ellipsoidal uncertainty sets are convex and thus lose some modeling generality afforded by scenario uncertainty sets. However, these models implicitly capture many scenarios and often lead to more tractable robust optimization models.

Our work shows that many problems remain open for study, even under relatively simple continuous uncertainty sets. The challenge in solving robust optimization problems having continuous uncertainty sets is that the decision-maker must anticipate, and account for, data that depends on the chosen scheduling decision. In terms of Stackelberg (leader-follower) games, when the follower (inner maximization) problem is trivial, combining the outer minimization and inner maximization problems is sometimes straightforward, and the scheduling problems remain easy to solve. In other

cases, the inner maximization problem is more difficult, and leads to a far more complicated optimization problem. The complexity of several such problems remains open, even under budgeted or ellipsoidal uncertainty sets, with the absolute robustness objective. The theoretical investigation of these robust optimization problems, along with exact and heuristic schemes for solving them, constitutes a major challenge for scheduling researchers.

One of the most significant gaps in the literature of robust SMSP is the lack of results on continuous (noninterval) uncertainty representations, like budgeted uncertainty. Despite the extensive use of interval uncertainty and scenario-based uncertainty, there are disadvantages in using both. The assumption of having independent intervals for uncertain parameters results in extreme worst-case scenarios that rarely happen in practice. On the other hand, enumerating all possible cases in the case of scenario-based uncertainty may be computationally intensive. As budgeted uncertainty eliminates the shortcomings of these two methods, it has been widely applied in several robust optimization problems. Therefore, investigation of this uncertainty representation method in robust SMSP appears to be an important area for future investigation. In particular, when dealing with absolute robustness, several open problems have been listed in Ref. 34. In addition, the effect of uncertainty on the value of any parameter other than processing time is yet to be studied. For other robustness measures, budgeted uncertainty has not been addressed, which provides a great opportunity for a wide variety of research in this area.

Another opportunity for future research, which is evident from Table 5, is the application of robustness measures other than absolute robustness. Discovering the problems that remain polynomially solvable under these two robustness measures will shed light on the boundary between polynomially solvable and NP-hard robust scheduling problems. In addition to complexity analysis for unaddressed problems, designing efficient solution methods and

developing single-stage mathematical formulations for robust SMSP problems will be very beneficial. This effort could be further extended to address problems with more challenging scheduling objectives, such as total tardiness ($\sum_{j \in J} T_j$) and total weighted tardiness ($\sum_{j \in J} w_j T_j$), as well as objectives based on earliness and tardiness.

REFERENCES

1. Lawler EL. Optimal sequencing of a single machine subject to precedence constraints. *Manage Sci* 1973;19(5):544–546.
2. Moore JM. An n job, one machine sequencing algorithm for minimizing the number of late jobs. *Manage Sci* 1968;15(1):102–109.
3. Smith WE. Various optimizers for single-stage production. *Nav Res Logist Q* 1956;3(1–2):59–66.
4. Pruhs K, Sgall J, Torng E. Online scheduling. In: Leung JY-T, editor. *Handbook of scheduling: algorithms, models, and performance analysis*. Chapter 15. Boca Raton (FL): CRC Press; 2004. p 15 1–15 41.
5. Liu M, Chu C, Xu Y, *et al.* An optimal online algorithm for single machine scheduling to minimize total general completion time. *J Comb Optim* 2012;23(2):189–195.
6. Dantzig G. Linear programming under uncertainty. *Manage Sci* 1955;1(3–4):197–206.
7. Birge JR, Louveaux FV. *Introduction to stochastic programming*. New York: Springer-Verlag; 1997.
8. Heyman DP, Sobel MJ. *Stochastic models in operations research: stochastic optimization*. Volume II. Mineola (NY): Dover Publications; 2004.
9. Kall P, Wallace SW. *Stochastic programming*. Chichester: John Wiley & Sons; 1994.
10. Pinedo ML, Schrage L. Stochastic shop scheduling: a survey. In: Dempster MAH, Lenstra JK, Rinnooy Kan AHG, editors. *Deterministic and stochastic scheduling*. Volume 84. Dordrecht: Springer; 1982. p 181–196.
11. Sarin SC, Nagarajan B, Liao L. *Stochastic scheduling: expectation-variance analysis of a schedule*. New York: Cambridge University Press; 2010.
12. Atakan S, Bülbül K, Noyan N. Minimizing value-at-risk in single-machine scheduling; 2013. Available at http://www.optimization-online.org/DB_HTML/2013/10/4072.html. Accessed July 2015.

13. Sarin SC, Sherali HD, Liao L. Minimizing conditional-value-at-risk for stochastic scheduling problems. *J Scheduling* 2014;17(1):5–15.
14. van den Akker M, Hoogeveen H. Minimizing the number of late jobs in a stochastic setting using a chance constraint. *J Scheduling* 2008;11(1):59–69.
15. Deng Y, Shen S. Decomposition algorithm for optimizing multi-server appointment scheduling with chance constraints; 2014. Available at http://www.optimization-online.org/DB_HTML/2014/02/4238.html. Accessed July 2015.
16. Ben-Tal A, El-Ghaoui L, Nemirovski A. Robust optimization. Princeton (NJ): Princeton University Press; 2009.
17. Smith JC, Ahmed S. Introduction to robust optimization. In: Cochran J, editor. *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken (NJ): John Wiley & Sons; 2010. p 2574–2581.
18. Graham RL, Lawler EL, Lenstra JK, *et al.* Optimization and approximation in deterministic machine scheduling: a survey. *Ann Discrete Math* 1979;5:287–326.
19. Kouvelis P, Yu G. Robust discrete optimization and its applications. Norwell (MA): Kluwer Academic Publishers; 1997.
20. Sabuncuoglu I, Goren S. Hedging production schedules against uncertainty in manufacturing environment with a review of robustness and stability research. *Int J Comput Integr Manuf* 2009;22(2):138–157.
21. von Stackelberg H. The theory of the market economy. London: William Hodge and Co.; 1952.
22. Soyster AL. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper Res* 1973;21:1154–1157.
23. Averbakh I. Minmax regret solutions for minimax optimization problems with uncertainty. *Oper Res Lett* 2000;27(2):57–65.
24. Kasperski A. Minimizing maximal regret in single machine sequencing problem with maximum lateness criterion. *Oper Res Lett* 2005;33(4):431–436.
25. Lu CC, Lin SW, Ying KC. Robust scheduling on a single machine to minimize total flow time. *Comput Oper Res* 2012;39(7):1682–1691.
26. Montemanni R. A mixed integer programming formulation for the total flow time single machine robust scheduling problem with interval data. *J Math Model Algorithms* 2007;6(2):287–296.
27. Ben-Tal A, Nemirovski A. Robust convex optimization. *Math Oper Res* 1998;23(4):769–805.
28. Ben-Tal A, Nemirovski A. Robust solutions to uncertain linear programs. *Oper Res Lett* 1999;25:1–13.
29. Ben-Tal A, Nemirovski A. Robust solutions of linear programming problems contaminated with uncertain data. *Math Program* 2000;88(3):411–424.
30. El-Ghaoui L, Lebret H. Robust solutions to least-square problems with uncertain data matrices. *SIAM J Matrix Anal Appl* 1997;18:1035–1064.
31. El-Ghaoui L, Oustry F, Lebret H. Robust solutions to uncertain semidefinite programs. *SIAM J Optim* 1998;9:33–52.
32. Bertsimas D, Sim M. Robust discrete optimization and network flows. *Math Program* 2003;98:49–71.
33. Bertsimas D, Sim M. The price of robustness. *Oper Res* 2004;52(1):35–53.
34. Tadayon B, Smith JC. Algorithms and complexity analysis for robust single-machine scheduling problems. *J Scheduling* 2015;18(6):575–592.
35. Aloulou MA, Della Croce F. Complexity of single machine scheduling problems under scenario-based uncertainty. *Oper Res Lett* 2008;36(3):338–342.
36. Daniels RL, Kouvelis P. Robust scheduling to hedge against processing time uncertainty in single-stage production. *Manage Sci* 1995;41(2):363–376.
37. de Farias IR Jr., Zhao H, Zhao M. A family of inequalities valid for the robust single machine scheduling polyhedron. *Comput Oper Res* 2010;37(9):1610–1614.
38. Mastrolilli M, Mutsanas N, Svensson O. Single machine scheduling with scenarios. *Theor Comput Sci* 2013;477:57–66.
39. Kasperski A. Discrete optimization with interval data: minmax regret and fuzzy approach (studies in fuzziness and soft computing). Volume 228. Berlin: Springer-Verlag; 2008.
40. Aissi H, Bazgan C, Vanderpooten D. Min-max and min-max regret versions of combinatorial optimization problems: A survey. *Eur J Oper Res* 2009;197(2):427–438.
41. Yang J, Yu G. On the robust single machine scheduling problem. *J Comb Optim* 2002;6(1):17–33.

42. Lebedev V, Averbakh I. Complexity of minimizing the total flow time with interval data and minmax regret criterion. *Discrete Appl Math* 2006;154(15):2167–2177.
43. Kasperski A, Zieliński P. An approximation algorithm for interval data minmax regret combinatorial optimization problems. *Inf Process Lett* 2006;97(5):177–180.
44. Sotskov Y, Egorova N. Sequencing jobs with uncertain processing times and minimizing the weighted total flow time. *Algorithmic and Mathematical Foundations of the Artificial Intelligence*. Sofia, Bulgaria: Institute of Information Theories and Applications FOI ITHEA; 2008.
45. Averbakh I. Computing and minimizing the relative regret in combinatorial optimization with interval data. *Discrete Optim* 2005;2(4):273–287.
46. Janak SL, Lin X, Floudas CA. A new robust optimization approach for scheduling under uncertainty: II. Uncertainty with known probability distribution. *Comput Chem Eng* 2007;31(3):171–195.
47. Daniels RL, Carrillo JE. β -Robust scheduling for single-machine systems with uncertain processing times. *IIE Trans* 1997;29(11):977–985.
48. Wu CW, Brown KN, Beck JC. Scheduling with uncertain durations: modeling β -robust scheduling with constraints. *Comput Oper Res* 2009;36(8):2348–2356.
49. Leon VJ, Wu SD, Storer RH. Robustness measures and robust scheduling for job shops. *IIE Trans* 1994;26(5):32–43.
50. Kouvelis P, Daniels RL, Vairaktarakis G. Robust scheduling of a two-machine flow shop with uncertain processing times. *IIE Trans* 2000;32(5):421–432.
51. Kasperski A, Kurpisz A, Zieliński P. Approximating a two-machine flow shop scheduling under discrete scenario uncertainty. *Eur J Oper Res* 2012;217(1):36–43.
52. Ierapetritou MG, Floudas CA. Effective continuous-time formulation for short-term scheduling. 1. Multipurpose batch processes. *Ind Eng Chem Res* 1998;37(11):4341–4359.
53. Li Z, Ierapetritou MG. Robust optimization for process scheduling under uncertainty. *Ind Eng Chem Res* 2008;47(12):4148–4157.
54. Lin X, Janak SL, Floudas CA. A new robust optimization approach for scheduling under uncertainty: I. Bounded uncertainty. *Comput Chem Eng* 2004;28(6–7):1069–1085.
55. Li Z, Ierapetritou MG. Process scheduling under uncertainty: review and challenges. *Comput Chem Eng* 2008;32(4–5):715–727.
56. Verderame PM, Elia JA, Li J, *et al.* Planning and scheduling under uncertainty: a review across multiple sectors. *Ind Eng Chem Res* 2010;49(9):3993–4017.

ROBUST ORDINAL REGRESSION

SALVATORE CORRENTE

University of Catania,

Department of Economics and
Business, Corso Italia,
Catania, Italy

SALVATORE GRECO

University of Catania,

Department of Economics and
Business, Corso Italia,
Catania, Italy

Operations & Systems

Management, Portsmouth
Business School, Portsmouth,
UK

MIŁOSZ KADZIŃSKI

Poznań University of Technology,

Institute of Computing
Science, Poznań, Poland

ROMAN SŁOWIŃSKI

Poznań University of Technology,

Institute of Computing
Science, Poznań, Poland

Systems Research Institute,

Polish Academy of Sciences,
Warsaw, Poland

INTRODUCTION

Making any type of decision, from buying a car to siting a nuclear plant, from choosing the best student deserving a scholarship to ranking the cities of the world according to their liveability, involves the evaluation of several alternatives with respect to different aspects, technically called *evaluation criteria*. Multiple Criteria Decision Aiding (MCDA) (see [1, 2]) provides methodologies to recommend the Decision Maker (DM) a decision that best fits the DM's preferences. Formally, in MCDA, a set of n alternatives $A = \{a_1, \dots, a_n\}$ is evaluated with respect to a consistent family of m criteria $G = \{g_1, \dots, g_m\}$ [3]. In general, each criterion $g_j \in G$ can be considered as a real-valued function $g_j : A \rightarrow$

$\mathcal{I}_j \subseteq \mathbb{R}$, where the elements of $\mathcal{I}_j$ are real numbers having either the meaning of quantities for quantitative criteria or the meaning of ordered identifiers for qualitative criteria, for example, 1 = "bad", 2 = "medium", 3 = "good". Each criterion g_j can have an increasing or a decreasing direction of preference. In the first case, the higher the evaluation $g_j(a)$, the better a is with respect to criterion g_j ; in the other case, the higher the evaluation $g_j(a)$, the worse a is with respect to criterion g_j . For example, evaluating a car involves both quantitative and qualitative criteria having increasing or decreasing direction of preference. Price and acceleration are typical quantitative criteria, whereas comfort and safety are qualitative criteria. Among these, acceleration, comfort and safety have an increasing direction of preference, whereas price has a decreasing direction of preference. In the following, without loss of generality, we shall suppose that criteria have increasing direction of preference.

According to Roy [4], in MCDA, the following three most important decision problems are considered:

- the *choice* problem, requiring to select a small number (as small as possible) of "good" alternatives in such a way that a single alternative may finally be chosen;
- the *sorting* problem, requiring to assign each alternative to some predefined and ordered categories;
- the *ranking* problem, requiring to define a complete or partial order on A ; this preorder is the result of a procedure allowing to put together in classes alternatives that can be judged indifferent and to rank these classes.

Given two alternatives $a, b \in A$ and considering their evaluations on the m criteria belonging to family G , very often a will be better than b for some of the criteria, whereas b will be better than a for the remaining criteria. For this reason, in order to cope with the three above-mentioned problems, it

is necessary to aggregate the evaluations of the alternatives, considering the preferences of the DM. In the literature, the three best known ways of aggregation are the following:

- assigning to each alternative $a \in A$ a real number synthesizing the evaluations of a on the m criteria and being representative of the desirability of a with respect to the problem at hand, as it is the case in MAUT—multiattribute utility theory [5, 6],
- building some outranking preference relation S on A , such that for any $a, b \in A$, aSb means that a is at least as good as b , as it is the case in outranking methods [3, 7, 8],
- using a set of “if..., then...” decision rules induced from the DM’s preference information through dominance-based rough set approach (DRSA, see Refs 9–12).

The above-mentioned three ways of aggregation lead to three models of DM’s preferences, called shortly *preference models*. They have been compared at an axiomatic level with respect to their capacity of representation. The comparison leads to a conclusion that the decision rule model is the most general and able to represent the most complex interactions among criteria [13, 14].

In the case of the first model, in order to assign a real number to each alternative, we consider a value function $U : \prod_{j=1}^m \mathcal{I}_j \rightarrow \mathbb{R}$, such that for any $a, b \in A$, a is at least as good as b ($a \succeq b$) if $U(g_1(a), \dots, g_m(a)) \geq U(g_1(b), \dots, g_m(b))$. The simplest form of the value function is the additive form, defined as $U(g_1(a), \dots, g_m(a)) = \sum_{j=1}^m u_j(g_j(a))$, where $u_j(g_j(a))$ are nondecreasing functions of their arguments. In the following, for simplicity of notation, we shall use $U(a)$ instead of $U(g_1(a), \dots, g_m(a))$ for all $a \in A$.

In the case of the second model, we consider a function $S : \prod_{j=1}^m \mathcal{I}_j \times \prod_{j=1}^m \mathcal{I}_j \rightarrow \mathbb{R}$, non-decreasing in its first m arguments and nonincreasing in its last m arguments, such that for each $a, b \in A$, we have

- $S((g_1(a), \dots, g_m(a)), (g_1(b), \dots, g_m(b))) = 1$ if aSb , and $S((g_1(a), \dots, g_m(a)),$

$(g_1(b), \dots, g_m(b))) = 0$ otherwise, in case a crisp outranking relation is considered,

- a outranks b with a credibility $S((g_1(a), \dots, g_m(a)), (g_1(b), \dots, g_m(b))) \in [0, 1]$ in case a fuzzy outranking relation is considered.

In the case of the third model, starting from some preference information provided by the DM in the form of decision examples, the aim is to express relationships between the comprehensive decision concerning an alternative and its evaluations on relevant criteria, using “if..., then...” decision rules, such as

- “if maximum speed of a car is at least 175 km/h, and its price is at most 12,000 euro, then this car is comprehensively at least medium.”

The choice of a multicriteria decision-aiding method well adapted to the decision context depends on several aspects of the decision process and the cooperation between the analyst and the DM [15].

ORDINAL REGRESSION

Each decision model requires specification of some parameters. For example, using MAUT, the parameters are related to the formulation of the marginal value functions $u_j(g_j(a))$, $j = 1, \dots, m$. Eliciting direct preference information from the DM can be counterproductive in real-world decision-making situations because of a high cognitive effort required. Consequently, within MCDA, many methods have been proposed to determine the parameters characterizing the considered decision model in an indirect way, that is, inducing the values of such parameters from some holistic preference comparisons of alternatives given by the DM. This indirect preference elicitation is less demanding of cognitive effort and it is mainly used in the ordinal regression paradigm.

The most well-known ordinal regression methodology is the UTA (UTilités Additives) method proposed by Jacquet-Lagrèze and

Siskos [16], which aims at inferring one or more additive value functions from a given complete ranking on a reference set of alternatives A^R . The method considers a piecewise additive value function $U(a) = \sum_{j=1}^m u_j(g_j(a))$ having marginal value functions $u_j(\cdot), j = 1, \dots, m$, being piecewise

$$\begin{aligned} \max \quad & \varepsilon, \text{ s.t.} \\ & U(a^*) \geq U(b^*) + \varepsilon \quad \text{if } a^* \succ b^*, \text{ with } a^*, b^* \in A^R, \\ & U(a^*) = U(b^*) \quad \text{if } a^* \sim b^*, \text{ with } a^*, b^* \in A^R, \\ & \sum_{j=1}^m u_j(\beta_j) = 1, \quad u_j(\alpha_j) = 0, \quad j = 1, \dots, m, \\ & u_j(g_j(a)) \geq u_j(g_j(b)) \quad \text{if } g_j(a) \geq g_j(b), \forall a, b \in A, j = 1, \dots, m, \end{aligned} \quad \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \begin{array}{l} E^A \\ \\ E \end{array}$$

where

- β_j and α_j are the best and the worst considered values of criterion $g_j, j = 1, \dots, m$,
- $\succ$ and $\sim$ are the asymmetric and the symmetric part of the binary relation $\succsim$, representing the DM's preference information, that is, $a^* \succsim b^*$ means that a^* is at least as good as b^* for the DM,
- here, as always in the following, ε is considered without any constraint on the sign.

If the set of constraints E^A is feasible and $\varepsilon^* > 0$, then there exists at least one additive value function compatible with the DM's preferences. If there is no compatible value function, that is, if the preferences of the DM cannot be represented by an additive value function with pre-defined number of linear pieces, [16] suggests either to increase the number of linear pieces in some marginal value functions or to select the utility function U that gets the sum of deviation errors close to minimum and minimizes the number of ranking errors in the sense of Kendall or Spearman distance.

The ordinal regression paradigm has been applied within the two main MCDA approaches: those using a value function as preference model [16–20], and those using an outranking relation as preference model [21, 22].

linear, with a predefined number of linear pieces. UTA uses linear programming to assess the additive value function compatible with the preference information provided by the DM. Technically, we have to solve a linear programming problem of the type:

ROBUST ORDINAL REGRESSION

Usually, from among many sets of parameters of a preference model representing the preference information given by the DM, only one specific set is selected and used to work out a recommendation.

As the selection of one from among many sets of parameters compatible with the preference information given by the DM is rather arbitrary, *Robust Ordinal Regression* (ROR) proposes taking into account all the sets of parameters compatible with the preference information, in order to give a recommendation in terms of necessary and possible consequences of applying all the compatible preference models on the considered set of alternatives: the *necessary* weak preference relation holds for any two alternatives $a, b \in A$ ($a \succsim^N b$) if and only if a is at least as good as b for all compatible preference models, whereas the *possible* weak preference relation holds for this pair ($a \succsim^P b$) if and only if a is at least as good as b for at least one compatible preference model.

Although UTA^{GMS} [23] is the first method applying the ROR concepts, in the following, we shall describe the GRIP method [24] being its generalization taking into account intensity of preference. Then, we shall mention the other applications of the ROR that have been published later in several papers.

GRIP

In the UTA^{GMS} method [23], which initiated the stream of further developments in ROR,

the ranking of reference alternatives does not need to be complete as in the original UTA method [16]. Instead, the DM may provide pairwise comparisons just for those reference alternatives, he or she really wants to compare. Precisely, the DM is expected to provide a partial preorder $\succsim$ on A^R .

Obviously, one may also refer to the relations of strict preference $>$ or indifference $\sim$, which are defined as, respectively, the asymmetric and symmetric part of $\succsim$. The transition from a reference preorder to a value function is done in the following way: for $a^*, b^* \in A^R$,

$$\left. \begin{aligned} U(a^*) &\geq U(b^*) + \varepsilon, \text{ if } a^* > b^*, \\ U(a^*) &= U(b^*), \text{ if } a^* \sim b^*, \end{aligned} \right\} E_1$$

where ε is a (generally small) positive value.

In some decision-making situations, the DMs are willing to provide more information than a partial preorder on a set of reference alternatives, such as “ a^* is preferred to b^* at least as much as c^* is preferred to d^* .” The information related to the intensity of preference is accounted by the GRIP method [24]. It

$$\left. \begin{aligned} U(a^*) - U(b^*) &\geq U(c^*) - U(d^*) + \varepsilon, \text{ if } (a^*, b^*) > (c^*, d^*), \\ U(a^*) - U(b^*) &= U(c^*) - U(d^*), \text{ if } (a^*, b^*) \sim (c^*, d^*), \\ u_j(a^*) - u_j(b^*) &\geq u_j(c^*) - u_j(d^*) + \varepsilon, \text{ if } (a^*, b^*) \succ_j (c^*, d^*) \text{ for } g_j \in G, \\ u_j(a^*) - u_j(b^*) &= u_j(c^*) - u_j(d^*), \text{ if } (a^*, b^*) \sim_j (c^*, d^*) \text{ for } g_j \in G. \end{aligned} \right\} E_2$$

In order to check if there exists at least one model compatible with the preferences of the DM, we solve the following linear programming problem:

$$\begin{aligned} \varepsilon^* &= \max \varepsilon, \text{ s.t.} \\ E \cup E_1 \cup E_2 &= E^{\text{DM}} \end{aligned}$$

If the set of constraints E^{DM} is feasible and $\varepsilon^* > 0$, then there exists at least one additive value function compatible with the preference information provided by the DM, otherwise no additive value function is compatible with the provided information. In this case, the analyst can decide to check for the

may refer to the comprehensive comparison of pairs of reference alternatives on all criteria or to a particular criterion only. Precisely, in the holistic case, the DM may provide a partial preorder $\succsim^*$ on $A^R \times A^R$, whose meaning is: for $a^*, b^*, c^*, d^* \in A^R$,

$$(a^*, b^*) \succsim^* (c^*, d^*) \iff a^* \text{ is preferred to } b^* \text{ at least as much as } c^* \text{ is preferred to } d^*.$$

When referring to a particular criterion $g_j \in G$, rather than to all criteria jointly, the meaning of the expected partial preorder $\succsim_j^*$ on $A^R \times A^R$ is the following: for $a^*, b^*, c^*, d^* \in A^R$,

$$(a^*, b^*) \succsim_j^* (c^*, d^*) \iff a^* \text{ is preferred to } b^* \text{ at least as much as } c^* \text{ is preferred to } d^* \text{ on criterion } g_j.$$

In both cases, the DM is allowed to refer to the strict preference and indifference relations rather than to weak preference only. The transition from the partial preorder expressing intensity of preference to a value function is the following: for $a^*, b^*, c^*, d^* \in A^R$,

cause of the incompatibility [25] or can continue the decision-aiding process accepting the incompatibility.

Denoting by $\mathcal{U}_{AR}$ the set of value functions compatible with the preference information provided by the DM, in the GRIP method, three types of possible and necessary preference relations can be defined:

- $a \succsim^N b$ iff $U(a) \geq U(b)$ for all $U \in \mathcal{U}_{AR}$, with $a, b \in A$,
- $a \succsim^P b$ iff $U(a) \geq U(b)$ for at least one $U \in \mathcal{U}_{AR}$, with $a, b \in A$,
- $(a, b) \succsim^{*N} (c, d)$ iff $U(a) - U(b) \geq U(c) - U(d)$ for all $U \in \mathcal{U}_{AR}$, with $a, b, c, d \in A$,

- $(a, b) \succsim^{*P}(c, d)$ iff $U(a) - U(b) \geq U(c) - U(d)$ for at least one $U \in \mathcal{U}_{AR}$, with $a, b \in A$,
- $(a, b) \succsim_j^{*N}(c, d)$ iff $u_j(a) - u_j(b) \geq u_j(c) - u_j(d)$ for all $U \in \mathcal{U}_{AR}$, with $a, b, c, d \in A$, $g_j \in G$,

- $(a, b) \succsim_j^{*P}(c, d)$ iff $u_j(a) - u_j(b) \geq u_j(c) - u_j(d)$ for at least one $U \in \mathcal{U}_{AR}$, with $a, b \in A$, $g_j \in G$.

Given alternatives $a, b, c, d \in A$, and the sets of constraints

$$\begin{aligned}
 & \left. \begin{array}{l} U(b) \geq U(a) + \varepsilon \\ E^{\text{DM}} \end{array} \right\} E^N(a, b), & \left. \begin{array}{l} U(a) \geq U(b) \\ E^{\text{DM}} \end{array} \right\} E^P(a, b) \\
 & \left. \begin{array}{l} U(c) - U(d) \geq U(a) - U(b) + \varepsilon \\ E^{\text{DM}} \end{array} \right\} E^N(a, b, c, d), & \left. \begin{array}{l} U(a) - U(b) \geq U(c) - U(d) \\ E^{\text{DM}} \end{array} \right\} E^P(a, b, c, d) \\
 & \left. \begin{array}{l} u_j(c) - u_j(d) \geq u_j(a) - u_j(b) + \varepsilon \\ E^{\text{DM}} \end{array} \right\} E_j^N(a, b, c, d), & \left. \begin{array}{l} u_j(a) - u_j(b) \geq u_j(c) - u_j(d) \\ E^{\text{DM}} \end{array} \right\} E_j^P(a, b, c, d)
 \end{aligned}$$

we get that:

- $a \succsim^N b$ iff $E^N(a, b)$ is infeasible or it is feasible and $\varepsilon^N(a, b) \leq 0$, where $\varepsilon^N(a, b) = \max \varepsilon$, s.t. $E^N(a, b)$;
- $a \succsim^P b$ iff $E^P(a, b)$ is feasible and $\varepsilon^P(a, b) > 0$, where $\varepsilon^P(a, b) = \max \varepsilon$, s.t. $E^P(a, b)$;
- $(a, b) \succsim^{*N}(c, d)$ iff $E^N(a, b, c, d)$ is infeasible or it is feasible and $\varepsilon^N(a, b, c, d) \leq 0$, where $\varepsilon^N(a, b, c, d) = \max \varepsilon$, s.t. $E^N(a, b, c, d)$;
- $(a, b) \succsim^{*P}(c, d)$ iff $E^P(a, b, c, d)$ is feasible and $\varepsilon^P(a, b, c, d) > 0$, where $\varepsilon^P(a, b, c, d) = \max \varepsilon$, s.t. $E^P(a, b, c, d)$;
- $(a, b) \succsim_j^{*N}(c, d)$ iff $E_j^N(a, b, c, d)$ is infeasible or it is feasible and $\varepsilon_j^N(a, b, c, d) \leq 0$, where $\varepsilon_j^N(a, b, c, d) = \max \varepsilon$, s.t. $E_j^N(a, b, c, d)$;
- $(a, b) \succsim_j^{*P}(c, d)$ iff $E_j^P(a, b, c, d)$ is feasible and $\varepsilon_j^P(a, b, c, d) > 0$, where $\varepsilon_j^P(a, b, c, d) = \max \varepsilon$, s.t. $E_j^P(a, b, c, d)$;

As to properties of $\succsim^N$ and $\succsim^P$ on A , let us remind after [23] that

- $\succsim^N$ is a partial preorder on A ,
- $\succsim^N \subseteq \succsim^P$,
- $a \succsim^N b$ and $b \succsim^P c \Rightarrow a \succsim^P c$, $\forall a, b, c \in A$,

- $a \succsim^P b$ and $b \succsim^N c \Rightarrow a \succsim^P c$, $\forall a, b, c \in A$,
- $a \succsim^N b$ or $b \succsim^P a$, $\forall a, b \in A$.

The above-mentioned properties are the minimal ones characterizing $\succsim^N$ and $\succsim^P$ [26]. Other interesting properties of $\succsim^N$ and $\succsim^P$ are the following [23]:

- $\succsim^P$ is strongly complete and negatively transitive,
- $\succ^P$ is complete, irreflexive, and transitive.

FURTHER DEVELOPMENTS

When looking at the final ranking, the DM is mainly interested in the position that is attained by a given alternative and, possibly, in its comprehensive score. Therefore, in the RUTA method [27], the kind of preference information that may be supplied by the DM have been extended by information referring to the desired rank of reference alternatives, that is, final positions and/or scores of these alternatives. Indeed, people are used to refer to the desired ranks of the alternatives in their judgments. In many real-world decision situations (e.g., evaluation of candidates for a certain position), they use statements such as a^* should be among the 5% of best/worst alternatives or b^* should be ranked in the

second ten of alternatives. These statements refer to the range of allowed ranks that a particular alternative should attain. When using such expressions, people do not confront “one vs” as in pairwise comparisons or “pair vs” as in statements concerning intensity of preference, but rather rate a given alternative individually, comparing it with all the remaining alternatives jointly. Extreme ranking analysis [28] provides the worst and the best positions that each alternative can obtain considering the whole set $\mathcal{U}_{AR}$ of compatible value functions. See Refs 29–31 for some recent applications of the extreme ranking analysis methodology.

A great majority of methods designed to support the multiple criteria decision process, assuming that all evaluation criteria are considered at the same level. However, practical applications are often imposing a hierarchical structure of criteria. For example, in economic ranking, alternatives may be evaluated on indicators that aggregate evaluations on several subindicators, and these subindicators may aggregate another set of subindicators, etc. In this case, the marginal value functions may refer to all levels of the hierarchy, representing the values of particular scores of the alternatives on indicators, subindicators, sub-sub-indicators, etc. In multiple criteria hierarchy process (MCHP)[32–34], the DM may provide a partial preorder of reference alternatives in each node of the hierarchy and obtain a necessary and a possible preference relation on the whole set of alternatives in each node. MCHP permits to decompose a complex decision problem into a series of simpler subproblems.

UTADIS^{GMS} method [35] applies the ROR to sorting problems. The DM can provide preference information in terms of assignment of alternatives to intervals of classes obtaining as result, for each alternative, intervals of classes to which it can be assigned necessarily and possibly considering the whole set of compatible value functions. This approach has been subsequently extended to the DIS-CARD method, which additionally admits specification of desired class cardinalities [36].

ROR has been applied to extend the two best known families of outranking methods, which are ELECTRE [37] and PROMETHEE [28], giving birth to the ELECTRE^{GKMS} and the PROMETHEE^{GKS} methods. In these methods, the DM can provide preference information in terms of outranking of an alternative over another and/or can give comparisons among weights, possible intervals of weights, and intervals of thresholds. As a result, ROR gives necessary and possible outranking relations between alternatives considering the whole family of sets of preference model parameters (weights, indifference, preference and veto thresholds, and concordance cutting level) compatible with preference information.

Even if the additive model is among the most popular ones, some critics have been addressed to this model because it has to obey an often unrealistic hypothesis about preferential independence among criteria. In consequence, it is not able to represent *interactions* among criteria [38]. For example, consider evaluation of cars using such criteria as maximum speed, acceleration, and price. In this case, there may exist a negative interaction (*redundancy*) between maximum speed and acceleration because a car with a high maximum speed also has a good acceleration, so, even if each of these two criteria is very important for a DM who likes sport cars, their joint impact on reinforcement of preference of a more speedy and better accelerating car over a less speedy and worse accelerating car will be smaller than a simple addition of the two impacts corresponding to each of the two criteria considered separately in validation of this preference relation. In the same decision problem, there may exist a positive interaction (*synergy*) between maximum speed and price because a car with a high maximum speed usually also has a high price, and thus a car with a high maximum speed and relatively low price is very much appreciated. Thus, the comprehensive impact of these two criteria on the strength of preference of a more speedy and cheaper car over a less speedy and more expensive car is greater than the impact of the two criteria considered separately in validation of this preference relation.

To handle the interactions among criteria, one can consider *non-additive integrals*, such as Choquet integral and Sugeno integral (see, e.g., Ref. 39). ROR has been applied to such types of preference models originating the *NonAdditive Robust Ordinal Regression* (NAROR) [40]. However, the non-additive integrals suffer from limitations within MCDA [41]; in particular, they need that the evaluations on all criteria are expressed on the same scale. This means that in order to apply a nonadditive integral it is necessary, for example, to estimate if the maximum speed of 200 km/h is as valuable as the price of €35,000. Thus, in order to consider positive and negative synergies, UTA^{GMS} method has been extended to the UTA^{GMS}-INT method [42], which considers a value function, composed not only of the sum of marginal nondecreasing value functions but also of penalty and bonus functions representing negative and positive synergies between criteria, respectively. On the basis of a value function similar to that one used in UTA^{GMS}-INT, the customer satisfaction method MUSA^{INT} [43] applies the ROR concept to compare different customer profiles considering all the value functions compatible with the customers' preferences.

ROR has also been adapted to aid a group of DMs, $\mathcal{D} = \{d_1, \dots, d_p\}$, to cooperate in view of taking a collective decision. In UTA^{GMS}-GROUP and UTADIS^{GMS}-GROUP methods [44], the collective results account for the preferences expressed by each DM. For example, based on the necessary ($\succsim_{d_r}^N$) and possible ($\succsim_{d_r}^P$) preference relations for each DM d_r , in the UTA^{GMS}-GROUP method, four relations can be defined:

- $a \succsim_{\mathcal{D}}^{N,N} b : a \succsim_{d_r}^N b \text{ for all } d_r \in \mathcal{D},$
- $a \succsim_{\mathcal{D}}^{N,P} b : a \succsim_{d_r}^N b \text{ for at least one } d_r \in \mathcal{D},$
- $a \succsim_{\mathcal{D}}^{P,N} b : a \succsim_{d_r}^P b \text{ for all } d_r \in \mathcal{D},$
- $a \succsim_{\mathcal{D}}^{P,P} b : a \succsim_{d_r}^P b \text{ for at least one } d_r \in \mathcal{D}.$

Considering results of four different types permits to indicate what would happen always (for all compatible functions), sometimes (for at least one compatible function), or never (for none of the compatible

functions) with respect to a subset or to the whole set of DMs. In this way, one can investigate spaces of consensus and disagreement between the DMs.

Even if the recommendations obtained using ROR are “more robust” than a recommendation made using an arbitrarily chosen compatible model, in some decision-making situations, a score is needed to be known for different alternatives; for this reason, some users would like to see the “representative” model among all the compatible ones. The motto underlying this proposal is “one for all, all for one.” The representative value function represents all compatible value functions, which also do contribute to its definition. On the basis of the ROR concept, the representative model is the compatible model maximizing the difference of values between alternatives a and b for which a is necessarily preferred to b but b is not necessarily preferred to a and minimizing the difference of values between alternatives a and b for which neither a is necessarily preferred to b nor b is necessarily preferred to a . The representative model concept has been introduced for the first time in Ref. 45 and then applied to deal with ranking and choice problems [46, 47], outranking methods [48], sorting problems [49], and group decision making [50].

In order to explain the necessary and possible preference relations given by ROR in terms of conditions on evaluation criteria, [51] proposes to couple ROR with DRSA [52]. Applying DRSA to the necessary and possible preference supplied by ROR, we get decision rules stating, for example, that the preference, either necessary or possible, of alternative a over alternative b is explained by a strong preference on criterion g_{j_1} and at least mild preference on criterion g_{j_2} . In this case, the strong preference on criterion g_{j_1} and the at least mild preference on criterion g_{j_2} become the arguments suggested by the rule for the preference of a over b . In a learning perspective, decision rules supplied by DRSA can be the starting point for an interactive procedure for analyzing and constructing the DM's preferences. It enables the DM's understanding of the conditions for the

suggested recommendation and provides useful information about the role of particular criteria or their subsets.

ROR has been applied in multiple objective optimization (MOO; for an exhaustive collection of surveys on MOO, see Ref. 53) in Refs 54 and 55. In the first paper, GRIP method is used for an interactive exploration of the Pareto optimal set of a MOO. In the latter paper, pairwise comparisons provided by the DM are used to systematically contract a cone formed by the directions of the isoquants of all compatible achievement scalarizing functions. This cone is focusing the DM's attention on a subregion of the non-dominated set that better corresponds to his or her preferences.

It is also worth mentioning that ROR has been applied to guide evolutionary multiobjective optimization (EMO), that is, a procedure that approximates the Pareto front of a multiobjective optimization problem evolving an initial population of solutions through multiple generation by breeding and mutation, toward the set of solutions most preferred by the DM [56, 57].

Recent extensions of the ROR approach are presented in Refs 58 and 59 where the Stochastic multiobjective acceptability analysis (SMAA) [60] and the ROR are put together under a unified decision support framework.

CONCLUSION

In this article, we have presented the ROR being a family of multicriteria decision-aiding methods aiming at considering not only one compatible preference model but also the whole set of models compatible with some preference information provided by the DM in terms of pairwise comparison of some alternatives, intensities of preference, rank-related requirements, or statements concerning interaction between criteria. We have presented the GRIP method being the extension of the first ROR method UTA^{GMS} . We have also mentioned other applications of the ROR, including hierarchy of criteria (MCHP), sorting problems ($UTADIS^{GMS}$), outranking methods ($ELECTRE^{GKMS}$ and $PROMETHEE^{GKS}$), non-additive integrals

(NAROR), group decisions (UTA^{GMS} -GROUP and $UTADIS^{GMS}$ -GROUP), and MOO.

ACKNOWLEDGMENTS

The third and the fourth authors wish to acknowledge financial support from the Polish National Science Centre, grant no. N N519 441939.

REFERENCES

1. Ehrgott M, Figueira J, Greco S, editors. Trends in Multiple Criteria Decision Analysis. New York: Springer; 2010.
2. Figueira J, Greco S, Ehrgott M, editors. Multiple Criteria Decision Analysis: State of the Art Surveys. Berlin: Springer; 2005.
3. Roy B. Multicriteria Methodology for Decision Aiding. Dordrecht: Kluwer Academic; 1996.
4. Roy B. Paradigm and challenges. In: Figueira J, Greco S, Ehrgott M, editors. Multiple Criteria Decision Analysis: State of the Art Surveys. Berlin: Springer; 2005. p 3–24.
5. Dyer JS. MAUT-Multiattribute Utility Theory. In: Figueira J, Greco S, Ehrgott M, editors. Multiple Criteria Decision Analysis: State of the Art Surveys. Berlin: Springer; 2005. p 265–292.
6. Keeney RL, Raiffa H. Decisions with multiple objectives: Preferences and value tradeoffs. New York: John Wiley, Inc.; 1976.
7. Figueira J, Greco S, Roy B, Słowiński R. An overview of ELECTRE methods and their recent extensions. *J Multicriteria Decis Anal* 2013;20(1-2):61–65.
8. Figueira J, Mousseau V, Roy B. ELECTRE methods. In: Figueira J, Greco S, Ehrgott M, editors. Multiple Criteria Decision Analysis: State of the Art Surveys. Berlin: Springer; 2005. p 133–153.
9. Greco S, Matarazzo B, Słowiński R. Rough sets theory for multicriteria decision analysis. *Eur J Oper Res* 2001;129(1):1–47.
10. Greco S, Matarazzo B, Słowiński R. Decision rule approach. In: Figueira J, Greco S, Ehrgott M, editors. Multiple Criteria Decision Analysis: State of the Art Surveys. Berlin: Springer; 2005. p 507–562.
11. Słowiński R, Greco S, Matarazzo B. Rough sets in decision making. In: Meyers RA, editor. Encyclopedia of Complexity and Systems Science. New York: Springer; 2009. p 7753–7786.

12. Słowiński R, Greco S, Matarazzo B. Rough set and rule-based multicriteria decision aiding. *Pesqui Oper* 2012;32(2):213–269.
13. Greco S, Matarazzo B, Słowiński R. Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. *Eur J Oper Res* 2004;158:271–292.
14. Słowiński R, Greco S, Matarazzo B. Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. *Control Cybern* 2002;31(4):1005–1035.
15. Roy B, Słowiński R. Questions guiding the choice of a multicriteria decision aiding method. *EURO J Decis Process* 2013;1(1):1–29.
16. Jacquet-Lagrèze E, Siskos J. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. *Eur J Oper Res* 1982;10(2):151–164.
17. Charnes A, Cooper WW, Ferguson RO. Optimal estimation of executive compensation by linear programming. *Manage Sci* 1955;1(2):138–151.
18. Jacquet-Lagrèze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. *Eur J Oper Res* 2001;130(2):233–245.
19. Pekelman D, Sen SK. Mathematical programming models for the determination of attribute weights. *Manage Sci* 1974;20(8):1217–1229.
20. Srinivasan V, Shocker AD. Estimating the weights for multiple attributes in a composite criterion using pairwise judgments. *Psychometrika* 1973;38(4):473–493.
21. Mousseau V, Słowiński R. Inferring an ELECTRE TRI model from assignment examples. *J Glob Optimiz* 1998;12(2):157–174.
22. Mousseau V, Słowiński R, Zieliński P. A user-oriented implementation of the ELECTRE-TRI method integrating preference elicitation support. *Comput Oper Res* 2000;27(7-8):757–777.
23. Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *Eur J Oper Res* 2008;191(2):415–435.
24. Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. *Eur J Oper Res* 2009;195(2):460–486.
25. Mousseau V, Figueira J, Dias L, et al. Resolving inconsistencies among constraints on the parameters of an MCDA model. *Eur J Oper Res* 2003;147(1):72–93.
26. Giarlotta A, Greco S. Necessary and possible preference structures. *J Math Econ* 2013;49(2):163–172.
27. Kadziński M, Greco S, Słowiński R. RUTA: a framework for assessing and selecting additive value functions on the basis of rank related requirements. *Omega* 2013;41(4):735–751.
28. Kadziński M, Greco S, Słowiński R. Extreme ranking analysis in robust ordinal regression. *Omega* 2012;40(4):488–501.
29. Androulaki S, Psarras J, Angelopoulos D. Evaluating long term potential natural gas supply alternatives for Greece with multicriteria decision analysis. *Proceedings of the 77th EURO Working Group on MCDA*; April 2013; Rouen.
30. Siskos E, Askounis D, Psarras J. Robust e-government evaluation based on multiple criteria decision analysis techniques. *Proceedings of the 77th EURO Working Group on MCDA*; April 2013; Rouen.
31. Siskos E, Malafekas M, Askounis D, Psarras J. E-government benchmarking in European Union: a multicriteria extreme ranking approach. In: Douligieris C, Polemi N, Karantias A, et al. editors. *Collaborative, Trusted and Privacy-Aware e/m-Services*. New York: Springer; 2013. p 338–348.
32. Angilella S, Corrente S, Greco S, Słowiński R. Multiple Criteria Hierarchy Process for the Choquet integral. In: Purshouse RC, Fleming PJ, Fonseca CM, et al., editors. *EMO 2013*. Volume 7811 of LNCS. Berlin Heidelberg: Springer; 2013. p 475–489.
33. Corrente S, Greco S, Słowiński R. Multiple Criteria Hierarchy Process in Robust Ordinal Regression. *Decis Support Syst* 2012;53(3):660–674.
34. Corrente S, Greco S, Słowiński R. Multiple Criteria Hierarchy Process with ELECTRE and PROMETHEE. *Omega* 2013;41:820–846.
35. Greco S, Mousseau V, Słowiński R. Multiple criteria sorting with a set of additive value functions. *Eur J Oper Res* 2010;207(3):1455–1470.
36. Kadziński M, Słowiński R. DIS-CARD: a new method of multiple criteria sorting to classes with desired cardinality. *J Glob Optimiz* 2013;56(3):1143–1166.
37. Greco S, Kadziński M, Mousseau V, Słowiński R. *ELECTRE^{GKMS}*: Robust ordinal regression

- for outranking methods. *Eur J Oper Res* 2011;214(1):118–135.
38. Wakker PP. Additive Representations of Preferences: A New Foundation of Decision Analysis. Kluwer Academic Publishers: Dordrecht, 1989.
 39. Grabisch M. The application of fuzzy integrals in multicriteria decision making. *Eur J Oper Res* 1996;89:445–456.
 40. Angilella S, Greco S, Matarazzo B. Non-additive robust ordinal regression: A multiple criteria decision model based on the Choquet integral. *Eur J Oper Res* 2010;201(1):277–288.
 41. Roy B. A propos de la signification des dépendances entre critères: quelle place et quels modes de prise en compte pour l'aide à la décision? *RAIRO, Oper Res* 2009;43:255–275.
 42. Greco S, Mousseau V, Słowiński R. $UTA^{GMS}-INT$: robust ordinal regression of value functions handling interacting criteria. *Eur J Oper Res* 2011, Submitted. Available at <http://www.lgi.ecp.fr/Biblio/PDF/CR-LGI-2012-01.pdf> Accessed 2013 Oct 22.
 43. Angilella A, Corrente S, Greco S, Słowiński R. MUSA-INT: Multicriteria customer satisfaction analysis with interacting criteria. *Omega* 2014;42:189–200.
 44. Greco S, Kadziński M, Mousseau V, Słowiński R. Robust ordinal regression for multiple criteria group decision: UTA^{GMS} -GROUP and $UTADIS^{GMS}$ -GROUP. *Decis Support Syst* 2012;214(1):118–135.
 45. Figueira J, Greco S, Słowiński R. Identifying the “most representative” value function among all compatible value functions in the GRIP. Proceedings of the 68th EURO Working Group on MCDA; October 2008; Chania.
 46. Angilella S, Greco S, Matarazzo B. The most representative utility function for non-additive robust ordinal regression. 13th International Conference on Information Processing and Management of Uncertainty; Dortmund; 2010. p 220–229.
 47. Kadziński M, Greco S, Słowiński R. Selection of a representative value function in robust multiple criteria ranking and choice. *Eur J Oper Res* 2012;217(3):541–553.
 48. Kadziński M, Greco S, Słowiński R. Selection of a representative set of parameters for robust ordinal regression outranking methods. *Comput Oper Res* 2012; 39(11):2500–2519.
 49. Greco S, Kadziński M, Słowiński R. Selection of a representative value function in robust multiple criteria sorting. *Comput Oper Res* 2011;38:1620–1637.
 50. Kadziński M, Greco S, Słowiński R. Selection of a Representative Value Function for Robust Ordinal Regression in Group Decision Making. *Group Decis Negot* 2013;22(3):429–462.
 51. Greco S, Słowiński R, Zielniewicz P. Putting Dominance-based Rough Set Approach and robust ordinal regression together. *Decis Support Syst* 2013;54(2):891–903.
 52. Greco S, Matarazzo B, Słowiński R. Dominance-based rough set approach to decision under uncertainty and time preference. *Ann Oper Res* 2010;176(1):41–75.
 53. Branke J, Deb K, Miettinen K, Słowiński R. editors. Multiobjective Optimization: Interactive and Evolutionary Approaches. Volume 5252 of LNCS. Berlin: Springer; 2008.
 54. Figueira J, Greco S, Mousseau V, Słowiński R. Interactive multiobjective optimization using a set of additive value functions. In: Branke J, Deb K, Miettinen K, Słowiński R. editors. Multiobjective Optimization: Interactive and Evolutionary Approaches. State-of-the-Art Survey Series, LNCS 5252, Chapter 4. Berlin: Springer; 2008. p 97–120.
 55. Kadziński M, Słowiński R. Interactive robust cone contraction method for multiple objective optimization problems. *Int J Inform Technol Decis Making* 2012;11(2):327–357.
 56. Branke J, Greco S, Słowiński R, Zielniewicz P. Interactive evolutionary multiobjective optimization using robust ordinal regression. In: Ehrgott M, Fonseca CM, Gandibleux X, et al., editors. International Conference on Evolutionary Multi-Criterion Optimization. Volume 5467 of LNCS. Berlin: Springer; 2009. p 554–568.
 57. Branke J, Greco S, Słowiński R, Słowiński R.. Interactive evolutionary multiobjective optimization driven by robust ordinal regression. *Bull Pol Acad Sci Tech Sci* 2010;58(3):347–358.
 58. Kadziński M, Tervonen T. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. *Eur J Oper Res* 2013;228(1):169–180.
 59. Kadziński M, Tervonen T. Stochastic ordinal regression for multiple criteria sorting problems. *Decis Support Syst* 2013;55(11):55–66.
 60. Lahdelma R, Salminen P. Stochastic multicriteria acceptability analysis (SMAA). In: Ehrgott M, Figueira J, Greco S, editors. Trends in Multiple Criteria Decision Analysis. New York: Springer; 2010. p 321–354.

ROBUST PORTFOLIO SELECTION

DESSISLAVA
A. PACHAMANOVA
Associate Professor of
Operations Research,
Babson College, Wellesley,
Massachusetts

In the practice of finance, *robust portfolio selection* refers to the allocation of assets in such a way that the resulting portfolio is less sensitive to errors in the predicted behavior of the assets. Most generally, there are two stages to the quantitative robust portfolio selection process. At the *forecasting stage*, tools from robust statistics are utilized for the estimation of various parameters that are inputs to portfolio optimization models. At the portfolio *allocation stage*, simulation, robust optimization, and certain types of stochastic programming methods are used to minimize the effect of errors made at the forecasting stage. In the operations research literature, the term *robust portfolio selection* is used to refer mostly to the allocation stage. Hence, while some background for the most widely used robust forecasting techniques will be provided, the focus of this entry will be on robust portfolio optimization.

In order to place robust portfolio selection methods in context, the first section provides a brief summary of classical asset allocation models. The following sections discuss forecasting, simulation, and optimization methods used to make the classical models more robust or to improve their theoretical and computational characteristics.

ASSET ALLOCATION MODELS

The classical mean-variance portfolio allocation problem, first formulated by Markowitz [1], has the form

$$\begin{array}{ll}\max_{\mathbf{w}} & \mu' \mathbf{w} - \lambda \mathbf{w}' \Sigma \mathbf{w} \\ \text{s.t.} & \mathbf{w}' \mathbf{e} = 1\end{array}$$

where μ is the vector of expected returns (alphas) for N assets in the investment universe, Σ is the asset–asset covariance matrix, $\mathbf{w}$ is the N -dimensional vector of portfolio weights, λ is the risk aversion coefficient, and $\mathbf{e}$ is a vector of ones. This optimization problem formulation states that the optimal portfolio weights should be chosen so that the expected portfolio return less the portfolio risk (where risk is defined as the portfolio variance), scaled by the risk aversion coefficient, is as large as possible. The equality constraint ensures that the portfolio weights add up to one.

The above-mentioned portfolio optimization formulation is written in a utility maximization form. (The investor utility function corresponding to mean-variance optimization is the quadratic utility function.) The utility maximization framework is a parallel theory to the framework of mean-risk portfolio selection, and has its roots in the economic literature (e.g., [2]). Two alternative formulations used to describe the mean-variance portfolio allocation framework are (i) to minimize the portfolio variance subject to a constraint on the portfolio expected return and (ii) to maximize the portfolio expected return subject to a constraint on the portfolio variance. Solving these optimization problems for a range of values of the expected return (in the case of portfolio variance minimization) or the portfolio variance (in the case of expected return maximization) allows one to trace the efficient frontier of a portfolio; that is, to find the portfolios that dominate all other portfolios for any given level of expected return or risk.

The mean-variance portfolio optimization problem is a special case of the general mean-risk portfolio allocation framework, in which a portfolio risk measure is minimized subject to a constraint on the expected portfolio return. Advanced measures of risk include value-at-risk or VaR [3,4], conditional value-at-risk or CVaR [5], and lower partial moments [6,7], among others.

In their classical form, portfolio optimization problems are not robust with respect to small changes in their inputs, such as

expected returns or covariance matrices. As documented in multiple studies [8–13], a small change in the input parameters in the mean-variance optimization formulation can result in large swings in the optimal portfolio allocation. The effect is particularly strong in the case of errors in the estimation of the expected returns. In the mean-variance framework, the relative importance depends on the investor’s risk aversion, but as a general rule of thumb, errors in the expected returns are about 10 times more important than errors in the covariance matrix, and errors in the variances are about twice as important as errors in the covariances [14].

ROBUST PARAMETER ESTIMATION

Inputs to portfolio optimization models need to be estimated from data. In classical statistics, this is done by using maximum likelihood estimators (MLEs). For example, the sample mean and the sample covariance matrix are the MLEs for the true mean and covariance matrix when the underlying distribution is normal. When the sample distribution deviates even slightly from the assumed underlying distribution, however, the efficiency of these classical estimators may be drastically reduced. One approach to rectify the issues with unstable estimates is to use robust estimators. Examples of robust estimators are the median (in place of the sample mean) and the mean absolute deviation (MAD) (in place of the sample standard deviation). Advanced robust estimators include M- and S-estimators. Robust estimators may not be as efficient as MLEs when the underlying model is correct, but their properties are not as sensitive to deviations from the assumed distribution.

Robust estimation techniques for minimum-risk portfolios are discussed, for example, in [15–18]. These articles compute robust estimates of the covariance matrix of asset returns and solve a minimum-variance portfolio optimization problem, in which the covariance matrix is replaced by its robust estimate. Lauprete *et al.* [19] as well as DeMiguel and Nogales [20] study

portfolios for which robust parameter estimation and portfolio optimization are done in a single step.

Another issue is that inputs to portfolio optimization problems are often estimated from *historical* data. This can lead to counterintuitive, unstable, or merely “wrong” portfolios. Bayesian and shrinkage estimators are widely used for combining historical estimates with investor sentiments. (For a review of such estimators, see, e.g., Chapters 6 and 8 in Fabozzi *et al.* [21] and Chapter 9 in Pachamanova and Fabozzi [22].)

Bayesian estimation approaches are based on subjective interpretations of the probability that a particular event will occur. A probability distribution, called the *prior distribution*, is used to represent the investor’s knowledge about the probability before any data are observed. After more information is gathered (e.g., data are observed), Bayes’ rule is used to compute the new probability distribution, called the *posterior distribution*.

In the portfolio parameter estimation context, a posterior distribution of expected returns is derived by combining the forecast from the empirical data with a prior distribution. One of the most well-known examples of the application of the Bayesian framework in this context is the Black–Litterman model [23], which produces an estimate of future expected returns by combining the market equilibrium returns (i.e., returns that are derived from pricing models and observable data) with the investor’s subjective views. The investor’s views are expressed as absolute or relative deviations from the equilibrium together with confidence levels of the views (as measured by the standard deviation of the views).

Shrinkage is a form of averaging different estimators. The shrinkage estimator typically consists of three components: (i) an estimator with little or no structure (such as the sample mean); (ii) an estimator with a lot of structure (the shrinkage target); and (iii) a coefficient that reflects the shrinkage intensity. Probably the most well-known estimator for expected returns in the financial literature was proposed by Jorion [24]. The shrinkage target in Jorion’s model is a

vector array with the return on the minimum variance portfolio from the Markowitz portfolio optimization model, and the shrinkage intensity is determined from a specific formula. Shrinkage estimators are used for estimates of the covariance matrix of returns as well [25], although equally weighted portfolios of covariance matrix estimators have been shown to be as effective as shrinkage estimators [26].

Regardless of how sophisticated the estimation and forecasting methods are, there is always estimation error. With advances in computational capabilities and new research in the area of optimization under uncertainty, practitioners in recent years have been able to incorporate considerations for uncertainty not only at the estimation, but also at the portfolio optimization stage. Methods for considering inaccuracies in the inputs to the portfolio optimization problem include simulation (resampling) and robust optimization. The next two sections explain these robust portfolio selection techniques.

PORTFOLIO RESAMPLING

Portfolio resampling involves generating different scenarios for the values of the input parameters in a portfolio optimization problem by simulation, and finding weights that remain stable for small changes in those parameters [27–29]. An illustration of the resampling technique as applied to portfolio mean-variance optimization is as follows.

Given an initial vector of expected stock returns, $\hat{\mu}$, and covariance matrix, $\hat{\Sigma}$, for the N stocks in the portfolio, it is estimated based on T observations of the vector of asset returns:

1. Simulate S samples of size T either from a multivariate normal distribution with mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ or by sampling with replacement from the empirical data set;
2. Use the S samples from step 1 to compute S new estimates of vectors of expected returns $\hat{\mu}_1, \dots, \hat{\mu}_S$ and covariance matrices $\hat{\Sigma}_1, \dots, \hat{\Sigma}_S$;

3. Solve S portfolio optimization problems, one for each estimated pair of expected returns and covariances $(\hat{\mu}_s, \hat{\Sigma}_s)$ and save the weights for the N assets in a vector array $\mathbf{w}^{(s)}$, where $s = 1, \dots, S$;
4. To find robust portfolio weights, average out the weight for each asset over the S weights found for that asset in each of the S optimization problems.

The probability distribution for the portfolio weights obtained from the portfolio resampling process is a valuable piece of information. By plotting the weights for each asset obtained over the S iterations, one can get a sense for how variable this stock weight is in the portfolio.

A disadvantage to calculating robust portfolio weights by averaging the results of a simulation *after* the optimization stage, however, is that the final weights may not actually satisfy some of the constraints in the optimization formulation. In general, only convex constraints are guaranteed to be satisfied by the averaged final weights. Turnover constraints, for example, may not be satisfied. This fact, combined with the fact that the procedure is time-consuming and computationally intensive, can be serious limitations of the resampling approach for practical applications.

ROBUST PORTFOLIO OPTIMIZATION

Robust optimization incorporates uncertainty directly into the portfolio optimization process, and can be used in conjunction with robust portfolio estimation methods. The robust optimization paradigm (e.g., [30]) is to deal with uncertainty in optimization model inputs by requiring that the optimal solution to the optimization problem remains feasible for any realization of the uncertain parameters within prespecified uncertainty sets. Robust optimization can therefore be viewed as a max-min (or a min-max) approach depending on the structure of the constraints in the optimization problem because the constraints need to be satisfied for the worst-case values of the uncertain

parameters within the uncertainty sets. (The max-min/min-max interpretation for the robust optimization approach is similar in the case of uncertainty in the parameters of the objective function in an optimization problem—one optimizes the worst-case objective function value as the uncertain parameters vary in their prespecified uncertainty sets.)

The uncertainty sets themselves may contain infinitely many elements; however, by suitable reformulation of the problem for special kinds of uncertainty sets, a computationally tractable *robust counterpart* of the original optimization problem can be obtained that sometimes can also be linked to probabilistic guarantees on the feasibility of the optimal solution.

Robust Formulations of Classical Portfolio Allocation Frameworks

Early work in robust portfolio optimization focused on modeling uncertainty sets for inputs to the classical mean-variance problem. As an example, consider a confidence interval around each expected return forecast, which results in a “box” uncertainty set for the expected returns in the mean-variance portfolio optimization formulation,

$$U_\delta(\hat{\mu}) = \left\{ \mu \mid |\mu_i - \hat{\mu}_i| \leq \delta_i, i = 1, \dots, N \right\}.$$

The robust formulation of the portfolio selection problem is

$$\begin{aligned} \max_{\mathbf{w}} \quad & \min_{\mu \in U_\delta(\hat{\mu})} \{ \mu' \mathbf{w} \} - \lambda \mathbf{w}' \Sigma \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}' \mathbf{e} = 1 \end{aligned}$$

and it can be shown (e.g., [31] or [22]) that the robust counterpart of the problem under this uncertainty set is equivalent to

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}' \mu - \delta' |\mathbf{w}| - \lambda \cdot \mathbf{w}' \Sigma \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}' \mathbf{e} = 1 \end{aligned}$$

where $|\mathbf{w}|$ denotes the vector of component-wise absolute values of the entries of the vector of weights $\mathbf{w}$. Note that if the weight of

stock i in the portfolio is negative, the worst-case expected return for stock i is $\mu_i + \delta_i$ (the investor loses the largest amount possible). If the weight of stock i in the portfolio is positive, the worst-case expected return for stock i is $\mu_i - \delta_i$ (the investor gains the smallest amount possible). Observe that $\mu_i w_i - \delta_i |w_i|$ equals $(\mu_i - \delta_i)w_i$ if the weight w_i is positive and $(\mu_i + \delta_i)w_i$ if the weight w_i is negative. Hence, the robust counterpart minimizes the worst-case expected portfolio return. Stocks whose mean return estimates are less accurate (i.e., have a larger estimation error δ_i) are penalized in the objective function, and tend to have a smaller weight in the optimal portfolio allocation.

The effect of this particular “robustification” of the mean-variance portfolio optimization formulation can be viewed in two different ways. On the one hand, the values of the expected returns for the different stocks are adjusted downwards in the objective function of the optimization problem. In other words, the robust optimization model “shrinks” the expected return of stocks with large estimation error, that is, in this case the robust formulation is related to the shrinkage methods discussed earlier. On the other hand, the additional term in the objective function can be interpreted as a “risk-like” term that represents penalty for estimation error. The size of the penalty is determined by the investor’s aversion to estimation risk, and is reflected in the magnitude of the deltas.

Consider now an ellipsoidal uncertainty set for the expected return estimates,

$$U_\delta(\hat{\mu}) = \{ \mu \mid (\mu - \hat{\mu})' \Sigma_\mu^{-1} (\mu - \hat{\mu}) \leq \delta^2 \},$$

where Σ_μ is the covariance matrix of estimation errors for the vector of expected returns μ . This uncertainty set captures the idea of a joint confidence region, and represents the requirement that the scaled sum of squares (scaled by the inverse of the covariance matrix of estimation errors) between all elements in the set and the point estimates $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N$ can be no larger than δ^2 . In practical applications, the covariance matrix of estimation errors is often assumed to be diagonal. In the latter case, the set contains all vectors of expected returns that

are within a certain number of standard deviations from the point estimate of the vector of expected returns, and the resulting robust portfolio optimization problem would protect the investor if the vector of expected returns is within that range.

The robust counterpart of the mean-variance portfolio optimization problem with an ellipsoidal uncertainty set for the expected return estimates is the following optimization problem formulation:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \delta\sqrt{\mathbf{w}'\boldsymbol{\Sigma}_\mu\mathbf{w}} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{e} = 1. \end{aligned}$$

This is a second-order cone problem, which is computationally tractable and has a similar interpretation to the robust counterpart in the case of the box uncertainty set. Garlappi *et al.* [31] provide an interpretation of this robust model in terms of optimizing over a set of multiple priors as advocated by Gilboa and Schmeidler [32], thus pointing out links between the robust optimization approach, Bayesian estimators for expected returns, and decision theory.

A practical issue when applying the above-mentioned robust optimization model is as follows. This simple robust portfolio optimization model, which is also the one most widely considered in practice, assumes that the mean returns of the assets lie within an uncertainty set that reflects the estimation errors, whereas the covariances are known. If S independent historical asset returns are available, each of which is assumed to follow the same distribution with known covariance matrix, then, by the central limit theorem, the most natural uncertainty set for the unknown vector of expected returns is the ellipsoidal uncertainty set presented earlier in this section. As shown earlier, the robust portfolio optimization model based on this uncertainty set yields a portfolio optimization framework analogous to the classical one by Markowitz [1]. The reason for this is that the error covariance matrix for the mean returns is proportional to the covariance matrix of the returns themselves. This does not eliminate the problem observed in practice when

applying the classical Markowitz framework, which is that Markowitz portfolios are highly sensitive to small changes in input parameters (expected returns and covariance matrices).

Non-Markowitz portfolios can only be obtained if the error covariance matrix is no longer proportional to the covariance matrix of returns. One possible remedy is the zero-net α adjustment proposed by Ceria and Stubbs [33], in which the ellipsoidal set for the expected returns is intersected with a plane. On an intuitive level, the result is that the net adjustment in the total expected portfolio return can be controlled (as opposed to the case of the simple robust optimization formulation, in which the net adjustment in the expected portfolio return is always downwards). The structured uncertainty set proposed by Ceria and Stubbs [33] changes the characteristics of the optimal portfolios and also reduces the conservativeness of the optimal portfolio allocations that are often observed when using robust portfolio optimization in practice.

A variety of different uncertainty sets have been considered in the literature. Lobo and Boyd [34] consider not only interval uncertainty sets for the expected returns vector, but also the element-wise interval uncertainty sets for the covariance matrix. They show that the corresponding robust counterpart is a semi-definite programming problem. Tütüncü and Koenig [35] also consider box uncertainty sets for the expected returns and the covariance matrix. They show that the robust counterpart is a saddle-point problem, which can be solved efficiently. Costa and Paiva [36] consider uncertainty sets for the expected returns and the covariance matrices that are polytopes with known vertices and formulate the robust counterpart of a portfolio optimization model of the mean-variance type (minimum volatility of tracking error) over linear matrix inequalities. Goldfarb and Iyengar [37] show how uncertainty sets for the expected returns and covariance matrices can be mapped statistically to factor models for estimation of expected returns that are widely used in the financial industry and derive the robust counterparts of a number of portfolio optimization problems in the

factor model framework, including Sharpe ratio maximization and VaR minimization. Erdogan *et al.* [38] extend the latter approach to robust index tracking. Robust min-max portfolio selection over a set of competing scenarios is discussed in [39].

Outside the mean-variance portfolio optimization framework, Pinar and Tütüncü [40] use robust portfolio optimization in the context of finding minimum risk robust profit opportunities. Pachamanova [41] evaluates the performance of a robust formulation of the sample CVaR (using the d -norm uncertainty sets introduced by Bertsimas *et al.* [42]) for a portfolio of simulated and US market data. Quaranta and Zaffaroni [43] present a similar model, based on “box” uncertainty sets for the expected returns, and study its performance in the Italian market. Kawas and Thiele [44] introduce a single-period log-robust portfolio optimization approach, in which uncertainty sets are considered for the continuously compounded rates of return.

A comprehensive overview of the early work on robust portfolio optimization models is provided by Fabozzi *et al.* [21], and a more recent review of applications of robust optimization models in practice can be found in [45].

Distributional Uncertainty

In more recent developments, research in robust portfolio optimization has focused less on robust estimation of the parameters in classical portfolio allocation models from finance and more on finding the best allocation given some limited information about the distributions of the underlying uncertainties (asset returns), which can be, but is not limited to, first and the second moments. This research in robust optimization complements work on more general models of uncertainty that have a long history and span different areas, such as stochastic programming, economics, and finance ([46–49], among others). The uncertainty in the probabilistic model is often referred to as *ambiguity*.

El Ghaoui *et al.* [50] propose a framework for optimization of worst-case portfolio VaR based on information about first and second moments of the distribution of asset returns.

The minimum portfolio VaR problem is in fact an optimization problem with a chance constraint, and has the form

$$\begin{aligned} \min_{\gamma_0, \mathbf{w}} \quad & \gamma_0 \\ \text{s.t.} \quad & P(V_0 - \tilde{\mathbf{r}}'\mathbf{w} > \gamma_0) \leq \varepsilon \\ & \mathbf{w}'\mathbf{e} = 1 \end{aligned}$$

where V_0 is the value of the portfolio today and $\tilde{\mathbf{r}}$ is the vector of (uncertain) future asset returns. $V_0 - \tilde{\mathbf{r}}'\mathbf{w}$ is the change in portfolio value between today and a predetermined date in the future. In words, the portfolio VaR problem is to find the minimum value γ_0 so that the probability that the future portfolio loss exceeds γ_0 is less than or equal to a small constant ε (typically $\varepsilon = 5$ or 1%). If we introduce a new variable $\gamma = \gamma_0 - V_0$, the VaR optimization problem can be reformulated as

$$\begin{aligned} \min_{\gamma, \mathbf{w}} \quad & \gamma + V_0 \\ \text{s.t.} \quad & P(\gamma + \tilde{\mathbf{r}}'\mathbf{w} \geq 0) \geq 1 - \varepsilon \\ & \mathbf{w}'\mathbf{e} = 1. \end{aligned}$$

Using a multidimensional version of the Chebyshev inequality, El Ghaoui *et al.* [50] show that the robust counterpart of the VaR portfolio optimization problem (i.e., the optimization formulation for the worst-case portfolio VaR over all possible probability distributions with a given vector of means and covariance matrix) is a second-order cone problem of the kind

$$\begin{aligned} \min_{\gamma, \mathbf{w}} \quad & \gamma + V_0 \\ \text{s.t.} \quad & -E[\tilde{\mathbf{r}}'\mathbf{w}] + \alpha\sqrt{\mathbf{w}'\Sigma\mathbf{w}} \leq \gamma \\ & \mathbf{w}'\mathbf{e} = 1 \end{aligned}$$

where $E[\cdot]$ denotes expectation and α is a constant related to ε . Note that this formulation is reminiscent of the mean-variance portfolio optimization framework.

Natarajan *et al.* [51] generalize the model of El Ghaoui *et al.* to take advantage of information about asymmetry in the underlying probability distributions. They study the worst-case portfolio, that is, the VaR portfolio

optimization problem, given the information about the mean returns and the forward and backward deviations of factors that drive asset returns. (See [52] for a definition of forward and backward deviations.)

Zhu and Fukushima [53] derive the worst-case CVaR portfolio optimization formulation for special cases of probability distributions, such as a box uncertainty set and a mixture distribution of some possible distribution scenarios. Huang *et al.* [54] consider a related problem of portfolio selection under distributional uncertainty in a relative robust CVaR approach, where relative CVaR is related to the worst-case risk of the portfolio relative to a benchmark. Natarajan *et al.* [55] derive the worst-case CVaR portfolio optimization problem for all distributions of uncertain returns with given moment information.

Mean-variance and mean-absolute deviation portfolio allocation problems under distributional uncertainty are studied by Calafiore [56], who derives the worst-case portfolio allocation if the true probability distribution of the uncertain returns is assumed to lie within a distance d from an estimated nominal distribution, with distance measured according to the Kullback–Leibler divergence measure.

Popescu [57] introduces the idea of worst-case portfolio allocation with expected utility maximization given partial information about the distribution of uncertain returns. She considers the single-period robust (maximin) expected utility problem

$$\max_{\mathbf{w}} \min_{\tilde{\mathbf{r}}} E[u(\tilde{\mathbf{r}}' \mathbf{w})]$$

under the assumption that the first and the second moments of the uncertain returns are known exactly. The author suggests a parametric quadratic programming algorithm for solving the problem that is efficient for a large class of utility functions. The idea is further developed by Delage and Ye [58], who propose an approach that works for a smaller family of utility functions but considers moment uncertainty, and by Natarajan *et al.* [59], who assume that the supports and the first two moments of distributions of the random returns are known and also address the issue of moment uncertainty. Chen *et al.* [60]

maximize the expected utility or minimize the disutility over infinitely many possible ambiguous distributions of the investment returns fitting a given mean and variance estimates. In the case of a single-stage portfolio optimization problem, they show that when the disutility function is in the form of lower partial moments, VaR or CVaR, the problem can be solved analytically. They also consider the so-called S-shaped utility function that plays a key role in the prospect theory of Kahneman and Tversky [61] and show an efficient procedure for solving the robust counterpart of the corresponding portfolio selection problem.

Wozabal [62] introduces a framework for dealing with ambiguity of the distribution of random asset losses in the portfolio selection problem. He defines “robustified” risk measure versions of several portfolio risk measures, including the standard deviation, the CVaR, and the general class of distortion functionals that measure the worst-case portfolio risk over an ambiguity set of loss distributions, where an ambiguity set is defined as a neighborhood around a reference probability measure representing an investor’s beliefs about the distribution of asset losses. A numerical study shows that in most instances the robustified risk measures perform significantly better out-of-sample than their non-robust variants in terms of risk, expected losses, and turnover.

Zymler *et al.* [63] deal with the problem of portfolio optimization under distributional uncertainty in the presence of assets with nonlinear payoffs such as European options. They develop two tractable conservative approximations for the VaR of a derivative portfolio by evaluating the worst-case VaR over all return distributions of the derivative underlies with given first- and second-order moments.

A connection between weak and strong robustnesses guarantees and portfolio insurance is discussed in [64]. The authors propose a model for designing portfolios that trades off weak and strong guarantees on the worst-case portfolio return. The weak guarantee applies as long as the asset returns are realized within the prescribed uncertainty set,

while the strong guarantee applies for all possible asset returns.

Robust Optimization and Risk Measures

Consider the following portfolio selection problem, in which asset returns are assumed to be uncertain parameters:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \min_{\tilde{\mathbf{r}} \in U_\lambda(\tilde{\mathbf{r}})} \{\tilde{\mathbf{r}}' \mathbf{w}\} \\ \text{s.t.} \quad & \mathbf{w}' \mathbf{e} = 1. \end{aligned}$$

where the uncertainty set for the asset returns is defined using the expected returns $\boldsymbol{\mu}$ and the covariance matrix of returns $\boldsymbol{\Sigma}$ as

$$U_\lambda(\tilde{\mathbf{r}}) = \{\mathbf{r} \mid (\mathbf{r} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{r} - \boldsymbol{\mu}) \leq \lambda^2\}.$$

As observed by Ben-Tal and Nemirovski [65], the above-mentioned problem is equivalent to the classical mean-standard deviation portfolio allocation problem in finance,

$$\begin{aligned} \max_{\mathbf{w}} \quad & \boldsymbol{\mu}' \mathbf{w} - \lambda \sqrt{\mathbf{w}' \boldsymbol{\Sigma} \mathbf{w}} \\ \text{s.t.} \quad & \mathbf{w}' \mathbf{e} = 1. \end{aligned}$$

This perspective has led to interesting findings about the relationship between robust portfolio selection and decision theory, and has provided economic meaning to the selection of uncertainty sets in robust optimization.

The correspondence between the shape of uncertainty sets in robust optimization and some popular portfolio risk measures is explored in more detail by Natarajan *et al.* [55]. One of the most applicable aspects of this work for finance practice is to use robust optimization to modify traditional risk measures from finance in a way that improves their theoretical properties. For example, the authors present guidelines for selecting the uncertainty set in such a way as to create coherent (in the sense of Artzner *et al.* [66]) versions of traditionally noncoherent risk measures such as VaR. Specifically, if the uncertainty set for the returns is specified so that it is a subset of the convex hull of the actual set of possible values of the uncertainties in the portfolio optimization problem, then the risk

measure defined by that uncertainty set is coherent.

These results are related to a duality result that can be traced back to the robust statistics literature [67]. There is now a large body of literature on the relationship between coherent risk measures and worst-case expected portfolio return over a family of probability distributions. The so-called *representation theorem* (e.g., [68]) states that a risk measure $\rho(\cdot)$ is coherent if and only if there exists a family of probability measures $\mathbb{Q}$ such that

$$\rho(\tilde{v}) = \sup_{Q \in \mathbb{Q}} E_Q[-\tilde{v}]$$

where $E_Q(\tilde{v})$ denotes the expectation of the random variable $\tilde{v}$ under the measure Q (as opposed to the measure of $\tilde{v}$ itself). In the portfolio optimization context, $\tilde{v}$ is the portfolio return. Bertsimas and Brown [69] examine the connections between risk management tools, distributional ambiguity, and robust optimization using this result as a starting point. Specifically, with risk preferences specified by a coherent risk measure, they examine the implications for uncertainty set structure in robust linear optimization problems.

The correspondence between uncertainty sets in robust optimization and risk measures in finance has been explored in interesting ways that go beyond typical portfolio optimization. For example, using results obtained by Natarajan *et al.* [55], Bertsimas *et al.* [70] derive a generalization of the Black–Litterman approach for estimating expected returns in traditional mean-variance optimization and illustrate how to construct “BL”-type estimators for more general notions of risk such as coherent risk measures. Computationally efficient multiperiod portfolio allocation models, discussed in the next section, also take advantage of this relationship.

Multiperiod Portfolio Optimization

Portfolio management is an inherently multiperiod problem. By incorporating long-term views on asset behavior, portfolio managers can take advantage of opportunities better

and can reduce their transaction costs, as the portfolio will not get rebalanced unnecessarily often. Gulpinar and Rustem [71] discuss robust multiperiod mean-variance optimization over a scenario tree, assuming ambiguity in the estimation of the mean and the covariance structure of unknown returns. Robust portfolio optimization over scenarios in a multiperiod setting is also studied by Pinar [72] who focuses on downside risk measures.

Multiperiod portfolio optimization schemes are generally difficult to handle computationally because the optimization problems are of large dimension. Robust optimization techniques with roots in the robust control literature, as well as adjustable robust optimization techniques, have been suggested to create a portfolio optimization formulation that can be more easily applied in practice.

Ben-Tal *et al.* [73] were the first to suggest a generalization of mean-standard deviation optimization for multiple time periods as a robust optimization problem that can be written as a second-order cone problem of small dimension, thus reducing substantially the computational difficulties in solving the portfolio allocation problem over multiple stages. They study the performance of their method versus stochastic programming methods for multistage portfolio management and find that the method is not only preferable in terms of computational tractability, but also in terms of realized portfolio performance according to a variety of measures. The main idea of Ben-Tal *et al.*'s approach is actually related to a technique for modeling dynamic programming problems, the certainty equivalent controller (CEC), that has been well known in engineering applications for many years. The CEC technique applied at stage t of a particular dynamic programming problem finds the optimal policy by assuming that all uncertain parameters for subsequent stages are fixed at their expected values; that is, it eliminates the uncertainty from the optimization problems. An obvious drawback of this approach is that risk is not considered in the solution to the problem. Ben-Tal *et al.* extend the CEC and incorporate risk by considering ellipsoidal uncertainty sets around the random parameters at all subsequent

time periods. (The random parameters in the case of the multiperiod portfolio optimization problem are the future asset returns.) Consequently, if there are T time periods, the multiperiod portfolio optimization formulation becomes a single optimization problem with T balance constraints that imitate solving single-period mean-standard deviation portfolio problems linked across time.

A subsequent study by Bertsimas and Pachamanova [74] explores other aspects of the robust optimization approach to multiperiod portfolio optimization, such as introducing different uncertainty sets to make the resulting optimization problem linear and comparing the approach to single-period mean-standard deviation optimization under variety of assumptions for the stochastic behavior of the asset returns. The general conclusion from computational studies is that incorporating views on the direction of asset prices in the future through multiperiod robust optimization models tends to reduce transaction costs and improve overall portfolio performance characteristics such as realized average return, median return, and probability of loss.

Adjustable robust optimization [75], an active research area in robust optimization, is a promising new framework for solving multiperiod portfolio optimization problems. It accommodates models, for example, in which information reveals itself gradually over time and decisions in the future can be made with knowledge of events that have happened until the current time period. However, problems defined as adjustable robust optimization problems can be severely computationally intractable, and so far applications have been limited to special cases. For example, Chen *et al.* [60] find a multistage adjustable robust optimization solution to the portfolio optimization problem when an investor has a disutility function that is the second-order lower partial moment.

FURTHER READING

For an overview of simulation and optimization methods in finance, see Pachamanova and Fabozzi [22]. For a practical perspective

on the use of robust optimization in quantitative portfolio management, see Fabozzi *et al.* [21]. For more background on robust estimation, see Huber [67]. For a theoretical treatment of robust optimization, see Ben-Tal *et al.* [30]. For recent surveys of robust optimization, see Bertsimas *et al.* [76] and Gabrel *et al.* [77]. For recent surveys of applications of robust optimization in finance, see Fabozzi *et al.* [45] and Kim *et al.* [78].

REFERENCES

1. Markowitz H. Portfolio selection. *J Finance* 1952;1(7):77–91.
2. Huang C-F, Litzenberger R. Foundations for financial economics. Upper Saddle River (NJ): Prentice Hall; 1998.
3. Jorion P. Value at risk: the new benchmark for managing financial risk. 3rd ed. New York (NY): McGraw-Hill; 2006.
4. Gaivoronsky A, Pflug G. Value-at-risk in portfolio optimization: properties and computational approach. *J Risk* 2005;7(2):1–31.
5. Rockafellar RT, Uryasev S. Conditional value-at-risk for general loss distributions. *J Bank Finance* 2002;26(7):1443–1471.
6. Bawa VS. Admissible portfolio for all individuals. *J Finance* 1976;31(4):1169–1183.
7. Fishburn PC. Mean-risk analysis with risk associated with below-target returns. *Am Econ Rev* 1977;67(2):116–126.
8. Black F, Litterman R. Global portfolio optimization. *Financ Anal J* 1992;48(5):28–43.
9. Best M, Grauer R. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Rev Financ Stud* 1991;4(2):315–342.
10. Best M, Grauer R. The analytics of sensitivity analysis for mean-variance portfolio problems. *Int Rev Financ Anal* 1992;1(1):17–37.
11. Broadie M. Computing efficient frontiers using estimated parameters. *Ann Oper Res* 1993;45:21–58.
12. Chan LKC, Karceski J, Lakonishok J. On portfolio optimization: Forecasting covariances and choosing the risk model. *Rev Financ Stud* 1999;12(5):937–974.
13. Jagannathan R, Ma T. Risk reduction in large portfolios: why imposing the wrong constraints helps. *J Finance* 2003;58(4):1651–1684.
14. Chopra V, Ziemba W. The effect of errors in means, variances, and covariances on optimal portfolio choice. *J Portf Manage* 1993;19(2):6–11.
15. Cavadini F, Sbuelz A, Trojani F. A simplified way of incorporating model risk, estimation risk and robustness in mean variance portfolio management. Working paper. Tilburg, The Netherlands: Tilburg University; 2001.
16. Vaz-de Melo B, Camara RP. Robust modeling of multivariate financial data. Coppead Working Paper Series 355. Rio de Janeiro, Brazil: Federal University at Rio de Janeiro; 2003.
17. Perret-Gentil C, Victoria-Feser M-P. Robust mean-variance portfolio selection. FAME Research Paper 140. Geneva: International Center for Financial Asset Management and Engineering; 2004.
18. Welsch RE, Zhou X. Application of robust statistics to asset allocation models. *Revstat Stat J* 2007;5(1):97–114.
19. Laureprete GJ, Samarov AM, Welsch RE. Robust portfolio optimization. *Metrika* 2002;25:139–149.
20. DeMiguel V, Nogales FJ. Portfolio selection with robust estimation. *Oper Res* 2009;57:560–577.
21. Fabozzi F, Kolm P, Pachamanova D, Focardi S. Robust portfolio optimization and management. Hoboken (NJ): John Wiley & Sons; 2007.
22. Pachamanova D, Fabozzi F. Simulation and optimization in finance: modeling with MATLAB, @RISK, or VBA. Hoboken (NJ): John Wiley & Sons; 2010.
23. Black F, Litterman R. Asset allocation: combining investor views with market equilibrium. New York: Goldman, Sachs & Co. Fixed Income Research; 1990.
24. Jorion P. Bayes-Stein estimator for portfolio analysis. *J Financ Quant Anal* 1986;21(3):279–292.
25. Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix. *J Portf Manage* 2004;30(4):110–117.
26. Disatnik D, Benninga S. Shrinking the covariance matrix. *J Portf Manage* 2007;33(4):55–63.
27. Michaud RO. Efficient asset management: a practical guide to stock portfolio optimization and asset allocation. Oxford: Oxford University Press; 1998.
28. Jorion P. Portfolio optimization in practice. *Financ Anal J* 1992;48(1):68–74.

29. Scherer B. Portfolio re-sampling: review and critique. *Financ Anal J* 2002;58(6):98–109.
30. Ben-Tal A, El Ghaoui L, Nemirovski A. Robust optimization. Princeton (NJ): Princeton University Press; 2009.
31. Garlappi L, Uppal R, Wang T. Portfolio selection with parameter and model uncertainty: a multi-prior approach. *Rev Financ Stud* 2007;20(1):41–81.
32. Gilboa I, Schmeidler D. Maxmin expected utility theory with non-unique prior. *J Math Econ* 1989;18:141–153.
33. Ceria S, Stubbs R. Incorporating estimation errors into portfolio selection: portfolio construction. *J Asset Manage* 2006;7: 109–127.
34. Lobo M, Boyd S. The worst-case risk of a portfolio. 2000. Unpublished manuscript, available from http://www.stanford.edu/~boyd/papers/pdf/risk_bnd.pdf. Accessed November 2012.
35. Tütüncü RH, Koenig M. Robust asset allocation. *Ann Oper Res* 2004;132:157–187.
36. Costa O, Paiva A. Robust portfolio selection using linear-matrix inequalities. *J Econ Dyn Control* 2002;26:889–909.
37. Goldfarb D, Iyengar G. Robust portfolio selection problems. *Math Oper Res* 2003;28(1):1–38.
38. Erdogan E, Goldfarb D, Iyengar G. Robust portfolio management. CORC Technical Report TR-2004-11. New York: Department of Industrial Engineering and Operations Research, Columbia University; 2004.
39. Rustem B, Becker RG, Marty W. Robust min-max portfolio strategies for rival forecast and risk scenarios. *J Econ Dyn Control* 2000;24(11):1591–1621.
40. Pinar M, Tütüncü RH. Robust profit opportunities in risky financial portfolios. *Oper Res Lett* 2005;33(4):331–340.
41. Pachamanova D. Handling parameter uncertainty in portfolio risk minimization. *J Portf Manage* 2006;32(4):70–78.
42. Bertsimas D, Pachamanova D, Sim M. Robust linear optimization under general norms. *Oper Res Lett* 2004;32:510–516.
43. Quaranta AG, Zaffaroni A. Robust optimization of conditional value-at-risk and portfolio selection. *J Bank Finance* 2008; 32:2046–2056.
44. Kawas B, Thiele A. A log-robust optimization approach to portfolio management. *OR Spectrum* 2011;33:207–233.
45. Fabozzi F, Huang D, Zhou G. Robust portfolios: contributions from operations research and finance. *Ann Oper Res* 2010;176:191–220.
46. Chen Z, Epstein L. Ambiguity, risk, and asset returns in continuous time. *Econometrica* 2002;70:1403–1443.
47. Dupačova J. The minimax approach to stochastic program and illustrative application. *Stochastics* 1987.; 20:73–88.
48. Shapiro A, Kleywegt AJ. Minimax analysis of stochastic problems. *Optim Methods Softw* 2002.; 17:523–542.
49. Dentcheva D, Ruszczyński A. Portfolio optimization with stochastic dominance constraints. *J Bank Finance* 2006;30(2): 433–451.
50. El Ghaoui L, Oks M, Oustry F. Worst-case value-at-risk and robust portfolio optimization: a conic programming approach. *Oper Res* 2003;51(4):543–556.
51. Natarajan K, Pachamanova D, Sim M. Incorporating asymmetric distributional information in robust value-at-risk optimization. *Manage Sci* 2008;54(3):573–585.
52. Chen X, Sim M, Sun P. A robust optimization perspective on stochastic programming. *Oper Res* 2007;55(6):1058–1071.
53. Zhu S, Fukushima M. Worst-case conditional value-at-risk with applications to robust portfolio management. *Oper Res* 2009;57:1155–1168.
54. Huang D, Zhu S, Fabozzi FJ, Fukushima M. Portfolio selection under distributional uncertainty: a relative robust CVaR approach. *Eur J Oper Res* 2010;203:185–194.
55. Natarajan K, Pachamanova D, Sim M. Constructing risk measures from uncertainty sets. *Oper Res* 2009;57(5):1129–1141.
56. Calafiore G. Ambiguous risk measures and optimal robust portfolios. *SIAM J Optim* 2007;18(3):853–877.
57. Popescu I. Robust mean-covariance solutions for stochastic optimization. *Oper Res* 2007;55(1):98–112.
58. Delage E, Ye Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper Res* 2009;58(3):596–612.
59. Natarajan K, Sim M, Uichanco J. Tractable robust expected utility and risk models for portfolio optimization. *Math Finance* 2010;20(4):695–731.

60. Chen L, He S, Zhang S. Tight bounds for some risk measures, with applications to robust portfolio selection. *Oper Res* 2011;59(4): 847–865.
61. Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. *Econometrica* 1979;43:263–291.
62. Wozabal D. Robustifying convex risk measures: a non-parametric approach. Manuscript; 2012. Retrieved from Optimization Online, http://www.optimizationonline.org/DB_FILE/2011/11/3238.pdf, November 2012.
63. Zymler S, Kuhn D, Rustem B. Worst-case value-at-risk of nonlinear portfolios. *Management Science* 2013;59(1):172–188.
64. Zymler S, Rustem B, Kuhn D. Robust portfolio optimization with derivative insurance guarantees. *Eur J Oper Res* 2011;210(2): 410–424.
65. Ben-Tal A, Nemirovski A. Lectures on modern convex optimization: analysis, algorithms, and engineering applications, MPR-SIAM Series on Optimization. Philadelphia (PA): SIAM; 2001.
66. Artzner P, Delbaen F, Eber J-M, Heath D. Coherent measures of risk. *Math Finance* 1999;9:203–228.
67. Huber P. Robust statistics. New York: John Wiley & Sons; 1981.
68. Fölmer H, Schied A. Robust preferences and convex measures of risk. In: Sandmann K, Schönbucher P, editors. *Advances in finance and stochastics*. Berlin: Springer-Verlag; 2002. p 39–56.
69. Bertsimas D, Brown D. Constructing uncertainty sets for robust linear optimization. *Oper Res* 2009;57(6):1483–1495.
70. Bertsimas D, Gupta V, Paschalidis I. Inverse optimization: a new perspective on the Black-Litterman model. *Oper Res* 2012. Forthcoming.
71. Gulpinar N, Rustem B. Worst-case optimal robust decisions for multi-period portfolio optimization. *Eur J Oper Res* 2007;183(3):981–1000.
72. Pinar MC. Robust scenario optimization based on downside-risk measure for multi-period portfolio selection. *OR Spectrum* 2007;29:295–309.
73. Ben-Tal A, Margalit T, Nemirovski A. Robust modeling of multi-stage portfolio problems. In: Frenk H, Roos K, Terlaky T, Zhang S, editors. *High performance optimization*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2000. p 303–328.
74. Bertsimas D, Pachamanova D. Robust multi-period portfolio management in the presence of transaction costs. *Comput Oper Res* 2008;35(1):3–17.
75. Ben-Tal A, Goryashko A, Guslitzer E, Nemirovski A. Adjustable robust solutions of uncertain linear programs. *Math Program* 2004;99:351–376.
76. Bertsimas D, Brown D, Caramanis C. Theory and applications of robust optimization. *SIAM Rev* 2011;53(3):464–501.
77. Gabrel V, Murat C, Thiele A. Recent advances in robust optimization and robustness: an overview. Manuscript; 2012.
78. Kim JH, Kim WC, Fabozzi F. Recent developments in robust portfolios with a worst-case approach. Manuscript; 2012.

ROBUSTNESS ANALYSIS

JONATHAN ROSENHEAD
Management Science Group,
Department of Management,
London School of Economics,
London, UK

ROBUSTNESS IN HISTORICAL AND CONCEPTUAL CONTEXT

This entry deals with a specific sense of the term *robustness*, that was coined by Gupta and Rosenhead [1]. The word was borrowed from common usage: the meanings recorded in the Oxford Dictionary include “strong and hardy of body or constitution” and “strongly and stoutly built.” However, robustness has also been used and has currency in a number of other technical senses close enough in meaning to run the risk of generating confusion.

The most well-established previous use of the term is in the field of statistics, where it is used to denote a statistical estimator or procedure that is relatively insensitive to departure from the distributional assumptions made when calibrating it. Huber [2] attributes the first use of the word robustness in this sense to Box [3] in 1953.

Bayesians give the term a rather more specific meaning. A Bayesian application is robust if the posterior distribution for an unknown parameter is not unduly affected by the choice either of the prior distribution or of the form of model taken to be generating the data [4]. More recently the field of *robust optimization* [5] has become of research and practical interest. However, as will be seen, the concept of optimization sits uneasily with that of robustness in the sense of this entry.

A more nearly compatible use of the term *robustness* is due to Roy and his collaborators [6]. Somewhat analogously to sensitivity analysis, Roy's approach seeks to incorporate

the real-world experience of uncertainty into the understanding and interpretation of mathematically derived results. Sensitivity analysis is well known and widely used. It is a systematic procedure that can be used to explore how an optimal solution responds to changes in inputs—typically these are either predicted values which might vary in the future or parameters whose values are open to question. The analysis aims at discovering how sensitive the “optimal” solution is to changes in crucial factors.

Roy's sense of “robustness” differs from sensitivity analysis in two principal ways. The first is that it aims to handle not only optimization but a range of other computational results, for example, that a certain solution is feasible, or that it is near optimal. The second difference is that its perspective is virtually the mirror image of that of sensitivity analysis. Its thrust is, rather, to identify the domain of points in the solution space for which a particular result continues to hold.

However, Roy's robustness also differs in two ways from robustness analysis as described in this article. Uncertainty is attached to parameter values, rather than to the swathe of possibly intangible “unknown unknowns” [7] that are likely to be resistant to credible quantification. And as with sensitivity analysis, the idea of exploiting sequentiality to achieve flexibility is absent.

The “robustness” of robustness analysis is a measure of the degree of flexibility of future decision possibilities that an act of commitment makes or leaves available. A commitment is an action with consequences, to be implemented without unnecessary delay following the act of decision. These consequences make any commitment irreversible, in the sense that it would take a further decision with *its* associated consequences and costs to “reverse” it. A decision not to give effect to any of the possible actions under consideration is also a commitment.

Any commitment is taken as acting upon and changing an existing system, with the

intention of improving or maintaining its performance. The subsequent decisions, which will or may be taken at points in the future, will also have that aim, though in new circumstances.

“Robustness” is an alternative or supplementary criterion to that of unmediated performance. “Performance,” here, is used to signify the directly useful or harmful attributes or outputs of the system of interest. It may be taken as uni- or multi-dimensional, depending on the focus of the evaluation. Robustness, by contrast, is a metalevel concept derived from estimates of future performance in order to evaluate the extent of useful flexibility left for the interim system which would result from alternative initial commitments.

FLEXIBILITY

Flexibility becomes of interest in decision making when (i) the degree of uncertainty about the future behavior of the system, or about its future context, is sufficient to make the establishment of an optimal decision or decision path a dubious proposition; and (ii) there will be later opportunities or indeed requirements to take further decisions that have impacts on the system. If both of these conditions hold, then flexibility consists in having the ability to choose between a range of meaningful options at these future decision points. The greater the range of this choice, the greater the remaining flexibility. The freedom to make choices is perhaps the most basic of all the freedoms that people, whatever their position or role, can have. However this freedom is empty if decisions taken earlier, by you or some others, have deprived you of meaningful alternatives from which to choose.

Operational research/management science (OR/MS) has tended to downplay the decisional issues arising from uncertainty. The emphasis, rather, has been on optimization, which requires as its precondition a belief in the certainty of estimates of performance or at least in the reliability of the probability distribution of such estimates.

However, these conditions of certainty of outcome (or of probabilities of outcomes) are at best approximated, and then in only a subset of decision problems—usually those with a relatively limited forward time span. Among important issues debated in the public domain that fail or have failed these tests are climate change, financial regulation, international conflict resolution, crime prevention, and epidemic management. In some of these an approach based on valuing flexibility could in principle have (or have had) more traction. The more strategic problems of private enterprises, communities, or individuals also tend to be characterized by uncertainties for which even probabilistic estimation is problematic.

Although not widely known within OR/MS, a range of approaches to identify and characterize flexibility have been developed. These include [8]:

adaptability: the ability to accommodate successfully to a new environment;

versatility: the possession of the necessary capabilities and the ability to apply them to the challenges of new circumstances that develop; and

resilience: the ability to withstand shocks without permanent damage and to rebound or recoil.

What distinguishes robustness analysis from these approaches is (i) it incorporates the MS/OR orientation toward choice between alternatives; and (ii) it provides an operational definition of flexibility, namely, *robustness*.

ROBUSTNESS ANALYSIS

The definition of the robustness of an initial commitment is

the number of acceptable options at the planning horizon that are compatible with that commitment, as a ratio of the total number of acceptable options.

The logic underlying this definition is illustrated in Fig. 1.

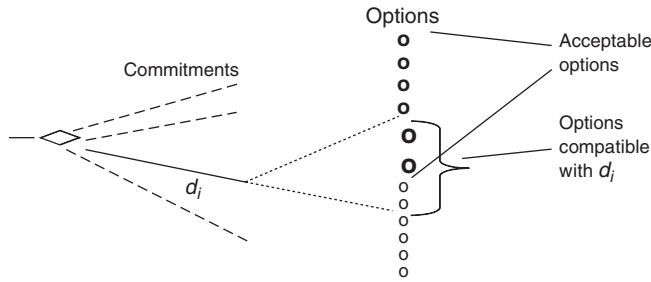


Figure 1. Schematic for robustness definition. d_i , a possible commitment; $\circ$, an unacceptable option; $\bullet$, an acceptable option; $\bullet$, an acceptable option compatible with d_i .

In slightly more formal notation, if S is the set of all options at the planning horizon with acceptable performance, S_i is the subset of S that is compatible with initial commitment d_i , and $n(\cdot)$ denotes cardinality or the number of elements in a set, then the robustness of commitment d_i is given by

$$r(d_i) = n(S_i)/n(S).$$

If none of the acceptable options are compatible with the selection of a particular commitment, its robustness score will be zero. If all the acceptable options are still available, its robustness score is 1. Thus, all robustness scores lie in the range $[0, 1]$. Higher scores are preferred to lower ones.

Robustness is an attribute of the initial commitment, and not of any plan (the options at the planning horizon). These plans are hypothetical constructs that are used to form judgments between alternative commitments. Robustness analysis can thus be seen as bifocal—one lens to focus on the next step, the other to survey the longer-term possibilities. However, the relationship between these two views is the reverse of that in conventional analytic approaches to planning, in which the immediate commitment is seen as a first step to the achievement of the identified long-term solution. In robustness analysis, rather, the longer-term is employed as a means to inform current decisions.

It is not suggested that decision makers should automatically select and implement the commitment that has the highest robustness score. There are factors other than flexibility to consider—cost, short-term benefits,

and so on. The balance between these elements will depend on circumstances. The advantage that robustness analysis brings to decision making is a language in which flexibility can participate systematically in the conversation.

Among the applications of robustness analysis are

- brewery locations across the United States [1];
- chemical plant expansion in the Dead Sea, Israel [9];
- strategic planning for hospital development for an Australian new town [10]; regional health planning for Ottawa-Carleton, Canada [11]; and
- sequential development of food production by a poor Brazilian community [12].

To employ robustness analysis it is necessary to

- identify the initial commitments to be evaluated;
- select a planning horizon (typically 5–10 years, depending on the speed of change in the environment and the time lag in implementing decisions);
- clarify the range and variety of future options that could result at the planning horizon;
- establish which commitments are compatible with, that is, leaving available, which options; and
- determine the acceptability of future options.

These steps provide the necessary elements for calculating the robustness scores for candidate initial commitments.

There is an irreducible element of judgment in preparing a problem for robustness analysis. In particular, “acceptability” may be based on quantitative information derived from a model plus an agreed threshold; or it may be based on judgments made by those who will be responsible for operating the system. “Compatibility” will sometimes be of a logical nature—if a particular commitment is not made, then certain avenues forward, which depend upon it, become inaccessible. Or the argument may be more nuanced—along the lines of “is or is not commitment C a reasonable first step if we were to think of option O as our destination?”

Where compatibility is based on logic, it will often be possible to build a graph showing the retention and attenuation of options through a series of future decision stages. In other cases, it will not be possible to identify the set of options at the planning horizon as simply consisting of the feasible combinations of intermediate decisions, so that no staged graphical representation will be possible. A method for generating options and establishing compatibility links between options and commitments in such cases is given in Rosenhead [13, pp. 198–200].

The judgments involved in a robustness analysis formulation may be made by the analyst/consultant. More commonly there will be a need for inputs from the client organization. It may be advantageous to elicit these in a workshop format, so that the inputs have the benefit of discussion among the client group. Used in this participative mode, robustness analysis is a member of the family of problem structuring methods (see Rosenhead and Mingers [14], where a fuller account of robustness analysis is also given).

Where judgmentally based inputs have an arbitrary element or potentially contested nature, the issue can often be resolved by running the analysis with alternative versions, in effect a form of sensitivity analysis. If similar patterns of robustness scores then

emerge, commitment choices based on the analysis will be reinforced; if not, attention should be focused on the sensitive aspects of the problem formulation.

VARIANTS

Multiple Future Robustness

The account of robustness analysis given above has assumed that an option is definitively acceptable or unacceptable. However, in some cases the acceptability of options is expected to depend on exogenous events that may or may not occur. In this case it may be useful to conduct a multiple future robustness analysis. This requires the identification of a (usually small) number of Futures $\{F_j\}$. The acceptability of the performance of any option (and possibly its compatibility with some commitments) could now depend on which future is considered. The robustness of any commitment d_i within Future F_j can now be calculated as

$$r(d_{ij}) = n(S_{ij})/n(S_j)$$

where all terms are defined analogously to the single future case. Any commitment now has a set of robustness scores, one for each future. Analysis of the robustness of alternative commitments can often progress by the elimination of commitments through dominance.

Debility

The rationale of measuring robustness is to keep open desirable options. The same basic logic can be used where there is a particular concern to *avoid* access to future options that are assessed as likely to perform unacceptably. However, the logic now is reversed. We can define the *debility* of an initial commitment as

the number of *unacceptable* options at the planning horizon that are compatible with that commitment, as a ratio of the total number of unacceptable options.

As with robustness, debility is constrained by its definition to lie in $[0, 1]$. However, in a mirror image of robustness, it is low debility scores that are advantageous. (See Caplin and Kornbluth [9] for more details on debility.) It is possible to conduct a combined robustness and debility analysis, providing two flexibility related measures for each commitment for any given future. This can, in principle, provide an enriched web of information for discussion about how to manage decision making under uncertainty.

Robustness and Long-Term Policy Analysis

A distinctive application of the concept of robustness has been developed by the RAND Corporation [15]. The focus of this application is on “long-term policy analysis,” which is concerned with the identification, assessment, and choice among near-term actions that will shape the options available to future generations, 30 or more years ahead. This choice will need to be made under conditions of “deep uncertainty,” where those charged with decision making do not know the relationships that will shape the long-term future, cannot rely on probability distributions of the key variables and parameters, and have no reliable way of valuing the desirability of alternative options.

RAND’s method operates through the generation of potentially very large numbers of plausible futures (i.e., scenarios), which are then employed using special-purpose software to identify near-term strategies that are robust, that is, perform reasonably well across a wide range of scenarios.

REFERENCES

1. Gupta SK, Rosenhead J. Robustness in sequential investment planning. *Manage Sci* 1968;15(2):18–29.
2. Huber PJ. Robust statistics: a review. *Ann Math Stat* 1972;43:1041–1067.
3. Box GEP. Non-normality and tests on variances. *Biometrika* 1953;49:419–432.
4. Rios Insua D, Ruggeri F, editors. *Robust Bayesian analysis*. New York: Springer; 2000.
5. Ben-Tal A, El Ghaoui L, Nemirovski A. *Robust optimization*. Princeton (NJ): Princeton University Press; 2009.
6. Roy B. Robustness in operational research and decision aiding: a multi-faceted issue. *Eur J Oper Res* 2009;200:629–638.
7. Rumsfeld D. Statement as US Secretary of state for defence. February 2002. Quoted inter alia by BBC, 2003.
8. Bahrami H, Evans S. *Super-flexibility for knowledge enterprises*. Berlin: Springer; 2005.
9. Caplin DA, Kornbluth JSH. Multiple investment planning under uncertainty. *Omega-Int J Manage Sci* 1975;3:423–441.
10. Parston G. *Planners, politics and health services*. London: Croom Helm; 1980.
11. Best G, Parston G, Rosenhead J. Robustness in practice: the regional planning of health services. *J Oper Res Soc* 1986;37:463–478.
12. Bornstein CT, Namen AA, Rosenhead J. Robustness analysis for sustainable community development. *J Oper Res Soc* 2007;58:588–601.
13. Rosenhead J. Robustness analysis: keeping your options open. In: Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley; 2001. pp. 181–207.
14. Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley; 2001.
15. Lempert RJ, Popper SW, Bankes SC. *Shaping the next one hundred years: new methods for quantitative, long-term policy analysis*. Santa Monica (CA): RAND; 2003.

ROBUSTNESS FOR OPERATIONS RESEARCH AND DECISION AIDING

BERNARD ROY
LAMSADE, Paris-Dauphine
University, Paris, France

WHAT IS OR-DA?

During World War II, the Allied Forces called upon several scientists (mainly mathematicians) to enlighten the strategic and tactical decisions in the theaters of operations. This was called *operations research* (OR) (operational research at the time). After the war, the scientists transposed what they had done in the battlefield to the corporate world, thus explaining the development of linear programming, the theory of games, the queuing theory, and more. By the late 1940s, in England, the United States, and later in France, consulting firms had the term OR in their corporate name. Aside from its purely mathematical component, OR is closely linked to decision aiding (DA), as can be clearly seen in the name [*Société Française de Recherche Opérationnelle et Aide à la Décision*, French Society for Operations Research and Decision Aiding] of the association whose purpose is to bring together researchers and practitioners in the field.

The men and women in business or the media who use the term *DA* are not necessarily referring to a scientific activity. Conversely, when talking about DA, operations research and decision aiding (OR-DA) necessarily adopts a scientific approach to the issues involving decision-making. The following definition is commonly accepted and often quoted:

Decision Aiding is the activity of the man or women who, relying on clearly explicit, but not necessarily formalized models, helps a stakeholder to reach the elements of answers to the questions that he or she may ask during a decision-making process. The elements should contribute to enlightening the decision,

communicating about the decision, drafting recommendations, or promoting behavior that would increase the coherence between the development of the decision-making process on the one hand, and the goals and value systems that the stakeholder is serving, on the other (Refs 1 and 2).

However, I cannot provide a definition for OR because there is no widely accepted definition of the terms. The first definition was “*operations research is what the Operations Research Company of New Jersey does*” (for more details, readers may refer to Refs 3 and 4).

Several fields of applications of OR-DA, among others, can be found below:

- assignment scheduling in a workshop or on a construction site;
- organizing delivery or pick-up rounds;
- drafting service schedules or timetables in a transport company;
- dam water management;
- the assignment of agricultural plots of land to classes of risk;
- the assignment of cancer cells to types of pathologies;
- routing communications within a telephone network;
- designing integrated circuits; and more.

Usefully, the reader's attention should be drawn to the following points. A person commonly called an *analyst* who practices DA relies on working hypotheses and models to reach *results* based on scientific foundations. The results may have various outcomes: satisfactory or optimal solutions, compromise proposals, acknowledgment of incoherence or incompatibility, the statements of conclusions, or even recommendations. Importantly, one should always keep in mind that the purpose of DA is not to dictate the “right decision” to the decision maker. DA does not claim to uncover hidden truths. DA must be designed in such a way to mesh with the decision-making process, thus enabling it to

evolve in accordance with the goals of the different stakeholders and facilitate mutual communication.

TWO CONCRETE EXAMPLES

Determination of the Characteristics of a Water Supply Network in an Urban Development Zone

I had to deal with this problem at a time when robustness in OR-DA was not yet an issue. I was then the scientific director of SEMA (Metra International). René Loué, an engineer at the Road and Bridges Authority, asked me what I thought of how the Roads and Bridges engineers were using an OR model to determine the structure and main characteristics of a water supply network for an urban development zone. The purpose of the very sophisticated model was to help define the optimal network that would meet a clearly defined demand for water, at the lowest cost, year after year over 30 years, in each neighborhood of the zone. On the basis of the land use plan scheduled for each neighborhood and the forecasts of household and business water consumption, the engineers had calculated the most probable volume of water that the network would have to supply, year after year, in each neighborhood.

According to René Loué, the land use plans were likely to change. How much the demand for water might change was also imperfectly known. Considering the imperfect knowledge and vague approximations in the modeling, René Loué felt that the optimum the Roads and Bridges engineers had found was totally misleading. If, several years later, actual demand sharply exceeded what they had determined as the most probable demand (to which the network had been finely adjusted) in certain areas, the streets would have to be torn up to change water lines, install new lines, and alter some other network features. On the other hand, if actual demand turned out to be much lower than the forecast probability, they would have made wasteful investments. Clearly, the best way of formulating the problem was not to search for the characteristics of a network that, for a given demand (considered as known: year

after year and neighborhood per neighborhood), would incur minimum investment and operational costs. The focus should rather be on gaining a grasp on what type of network would be the most likely to withstand a range of probable demands, considering the costs incurred by adjustments to actual demand. First, this meant that the approach could not be limited to a single scenario of the most probable demand, but had to introduce a range of plausible scenarios. Second, the model had to consider what it would cost to adjust the initially designed network to the scenario that would actually happen.

Choosing the Best Bid after a Call for Tenders

I will deal with this type of problem by referring to a concrete case where I helped find a solution, some 20 years ago. At the time, the French Post had decided to fit its regional mail-sorting centers with package sorting machines, which were still far from perfect. Therefore, it issued an international call for tenders for the construction of a prototype. The call for tenders had very detailed specifications and limited the cost of the mass-produced machine to a "specified objective cost." It received 15 bids.

Twelve criteria were used for the evaluation of each bid: prototype cost, stated deadline, the number of sorting directions, the number of sorted packages per hour, the rate of rejected or improperly sorted packages, sound levels, bag handling ease, occupied area, trust in the supplier to meet commitments, and three more technical criteria. Once each bid had been assessed by its 12 evaluation criteria, how should the best one be chosen? In this case, there was not one criterion to optimize. Considering the heterogeneity of the magnitudes serving for bid evaluation, designing a single criterion encompassing the 12 criteria was unrealistic. Regardless of the procedure selected to enlighten the choice, two aspects of the problem had to be considered:

1. Evaluating the bids on each criterion would inevitably comprise some uncertainty or arbitrariness.

2. The decision maker wanted some criteria to play a more important role than others.

I will deal later with each of the two aspects that are not specific to this particular problem.

WHAT IS ROBUSTNESS IN OR-DA?

As can be guessed by the two above examples, vague approximations and areas of ignorance, which affect modeling as well as what is commonly called the data, limit the scientific approach in OR-DA. Several examples of vague approximations and areas of ignorance can be found below:

- Modeling involves simplifying reality: several different ways are possible and only one is selected.
- One particular piece of information whose value is debatable is considered, as if it were certain.
- A random variable is supposed Gaussian, mainly for the purpose of facilitating calculations.
- Another source of vague approximations, which is often overlooked, involves the way of processing numbers, as if they represented a quantity: in the calculation, a deadline that is twice as long will be counted as being twice as penalizing, whereas it might actually be much more, or much less so.
- When modeling a system of (technical, economic, and social) constraints, one often overlooks the uppermost limit that is allowable within the (pressure, budget, and capacity) thresholds that mark (or will mark) the boundaries of the possible or the feasible in the future.
- When there are multiple criteria, the metaphor of weight is generally used to differentiate the role that the different criteria must play. However, often the weights do not refer to any objective reality, and the way they operate within the model is usually far from clear.

Importantly, the vague approximations and areas of ignorance (often referred to under the umbrella term *uncertainties*) have to be considered if useful answers are to be provided to the decision maker. Actually, not doing so may lead to misleading outcomes.

In OR-DA, *robust* is a qualifier that refers to an aptitude to withstand “vague approximations” or “areas of ignorance” to provide protection against deplorable impacts, such as results that are much worse than expected, or damage to properties to maintain (for more details see Refs 5–9).

The qualifier “robust” applies to various aspects, that is, solutions, conclusions, recommendations, and methods. Robustness primarily involves a *concern*. The concern must be present throughout DA, including at the time when the problem is formulated. This issue, which was raised at the end of the section titled “Determination of the Characteristics of a Water Supply Network in an Urban Development Zone,” is also highlighted in the following didactic example, which also underscores the share of subjectivity found in the concern for robustness.

Mrs. *D* is wavering between three solutions for her trip from *A* to *B*:

- *The subway*: This solution involves a change on the line. However, as the trains run regularly, she can estimate the travel time with a small margin of uncertainty.
- *A bus and a train*: The bus to the train station does not run very regularly. She also does not know the train timetable. However, the trains travel much faster than the subway does. Therefore, this solution is faster than the subway solution, but it might also take much longer if she has to wait for a bus and then for a train.
- *A car*: If there are no traffic jams, it is the fastest solution.

If the trip from *A* to *B* is a one-time event and if Mrs. *D* absolutely does not want to be late, she will find that the subway solution, albeit longer, is the most robust. If, on the other hand, it is a regular trip and if Mrs. *D*

considers that slight delays (as long as they do not occur too often) are acceptable, she might feel that, on average, subway travel time will be much longer than the travel time on the two other means of transportation. Therefore, the most robust solution has to be chosen from the two other solutions: even if traffic jams seldom occur, Mrs. *D* may fear that they would have unacceptable outcomes, something that would be less likely with the bus–train solution. In this case, the latter would seem to her as the most robust.

In the following paragraphs, I will call *the model on which DA relies and the procedures for processing the model to reach results*: formal representation (FR). I will call *the context, in which decisions will be made, executed, and judged*, a real-life context (RLC).

The decisions for which DA is performed will be implemented in a real-life situation that, undeniably, will not strictly conform to the model on which the DA relied. Furthermore, the decisions may be assessed in light of a value system, which may not necessarily be in perfect accordance with the system that was used to design and process the model. Therefore, seamless conformity between FR and RLC seems virtually impossible. This lack of conformity stems from extant vague approximations and areas of ignorance. The concern for robustness means that they have to be considered in the RF as best as possible; or the FR should consider at least those vague approximations or areas of ignorance that, if overlooked, would make DA unable to protect the decision maker from any deplorable impacts.

In a given decision-making situation, if the analyst is to meet the concern for robustness properly, he or she should design the FR while considering the two requirements that I will now explain.

TWO REQUIREMENTS TO MEET THE CONCERN FOR ROBUSTNESS PROPERLY

First requirement: Carefully inventory and properly consider all frailty points in the FR.

By definition, a *frailty point* is a place in the model, or in a procedure processing the model, where vague approximations or

areas of ignorance can be found. Considering a frailty point means retaining either a single option or a series of possible options. An *option* means a particular, precise way of dealing with vague approximations or areas of ignorance. For instance, an option may be:

- a given value assigned to imperfectly known data, or to a poorly defined parameter;
- a set of values setting the weight assigned to each of the relevant criteria;
- a particular probability distribution to characterize a random variable;
- a simplifying hypothesis making it possible to consider a complex phenomenon;
- a particular way of building a criterion to consider a point of view;
- a possible value assigned to a technical parameter involved in a processing procedure;
- the choice of a solution among the solutions closest to an optimum solution.

Given an identified frailty point, the analyst must ask himself or herself whether—by retaining only one option for the frailty point—he or she is running the risk that outcomes of model processing might conceal impacts that the decision maker may deem undesirable. If this is true, he or she should design a set of options likely to highlight the different eventualities. The set may be discrete (i.e., a few possible values, several sets of weights, two or three probability distributions, and two or three plausible hypotheses) or continuous (i.e., intervals of possible values and families of probability distributions characterized by parameters with values within the given intervals).

Second requirement: Design forms of answers likely to help the decision maker protect himself or herself from undesirable impacts that may result from vague approximations or areas of ignorance in the FR.

The forms of answers have to consider the decision maker's expectations of the said protection. Are there impact levels that he or she

is ready to accept in certain circumstances? Are there other levels that he or she considers unacceptable; levels for which he or she wants protection, whatever the circumstances or cost? Depending on the real-life context, and because of the vague approximations and areas of ignorance in the FR, the decision maker's decision could, to some extent, reconcile his or her need for protection on the one hand, with his or her desire to optimize the performance criterion or criteria serving to evaluate the solution, on the other. The forms of the analyst's answers will only be truly useful for the decision maker if enough information is provided, so that the latter can, in light of his or her own subjectivity, come to a decision. Given the two conflicting risks:

- Finding himself or herself poorly protected in the case of very poor performance revealing undesirable impacts.
- Finding himself or herself in a position whereby he or she forsakes any hope of good, or even very good performance.

The way both requirements are considered in current journals prompts me to make two observations, which I will present briefly (for more details, see Ref. 5).

First observation: The frailty points that are usually considered solely relate to the "data" whose purpose is to bring into play an aspect of RLC, which is called *uncertain*, into the FR.

This approach leads to design a set of *scenarios* to describe all the realities that are likely to happen and that must be considered by the concern for robustness. A scenario is defined by a single option chosen with regard to each frailty point within a set of related options. The set of scenarios may be finite or infinite.

Therefore, a proper formulation of the problem present in the section titled "Determination of the Characteristics of a Water Supply Network in an Urban Development Zone" means involving other narratives, alongside the narrative of the most probable water volumes (neighborhood per neighborhood), with each plausible narrative making up a scenario. This example also shows that

the relevant frailty points may also come from vague approximations and areas of ignorance, which are not necessarily related to the data on uncertain aspects of the RLC. Modeling water supply networks that will meet a given demand in the relevant zone involve various technical parameters. Several options for the value that will be assigned to each of the said parameters should be examined. The optimization criterion making it possible to process the model can also be formulated in different ways. Consequently, using the term *scenario* to describe a choice of options relating to the said frailty points may be inappropriate when communicating with the decision maker, who may assign a different meaning to the term. To ensure that all the frailty points are considered without restriction, I find it preferable to replace the concept of *scenario* with the concept of *version* (of the formulation of the problem). A version is still defined as a combination of options considering all the relevant frailty points (for more details, see Ref. 5). However, I will be using the term *scenario* in the rest of my article. Importantly, the frailty points that may be found in the way the model is processed should not be overlooked. This may lead to consider all the procedures in set P , wherein each p procedure is still defined by a precise option characterizing the procedure for each of the frailty points contained therein, rather than merely considering a single processing procedure.

Second observation: Most stakeholders are only interested in a single type of answer: find one (possibly several) solution(s) that can be qualified as "robust." To give the term meaning, they refer to an FR model where the performance of a solution is defined by a single criterion and they give preference to the performance of the solution in the worst-case scenario, to measure the robustness of the solution.

This type of answer relies on a fairly particular concept of a *robust solution* that fits within in a fairly restrictive schema that I will now address.

A FRAMEWORK SCHEMA TO MEET THE CONCERN FOR ROBUSTNESS

The schema may be defined by the three following characteristic traits.

Characteristic trait no. 1: In RF, the decisions, which have to be studied, are modeled as element a of a set A of the so-called feasible solutions. A single criterion g associates a performance $g(a)$ with each element. This criterion clears the way for defining the optimal solution or solutions, regardless of any concern for robustness.

Set A may be defined in *comprehension* by the set of solutions of a system of constraints, as can be seen in the example in the section titled “Determination of the Characteristics of a Water Supply Network in an Urban Development Zone.” In this example, criterion g is a cost criterion that has to be minimized. Set A can also be defined in *extension* by a list, as is the case in the example in the section titled “Choosing the Best Bid after a Call for Tenders.” However, in the latter example, it is not possible to characterize the performance of a bid by a single criterion to optimize.

Characteristic trait no. 2: A single measurement is defined to give meaning to the assertion “solution x is at least as robust as solution y .” This—and only this—measurement is involved when qualifying a solution as robust.

Three robustness measurements have been chiefly used since Kouvelis and Yu’s famous work [10]. I present them below in the event criterion g has to be minimized. On this subject, I will talk about cost (the cost being worth plus infinite for solution x , which is not feasible in scenario $s \in S$).

The first measurement relies on the notion of guaranteed cost. The guaranteed cost of solution x is the highest cost for the solution in the model. Therefore, it is the highest value of $g(x)$ in every scenario. Taking this guaranteed cost as the definition of the robustness measurement is the same as admitting that solution x is considered to be at least as robust as solution y , if, and only if, the guaranteed cost in x is, at most, equal to the cost guaranteed in y .

The two other robustness measurements rely on the concept of regret. One involves absolute regret and the other relative regret. Solution x having been selected, when the actual scenario is known, the decision maker may compare the $g(x)$ cost that he or she actually incurs in this scenario with the cost he or she would have incurred had the selected solution been the best in the scenario. Absolute regret is, by definition, the difference between the value of the $g(x)$ cost in the actual scenario and the cost incurred by the optimal solution in this scenario. Relative regret is defined by dividing absolute regret by the cost of solution x in the relevant scenario. Admittedly, the decision maker does not know which scenario will happen, but the maximum value for (absolute or relative) regret within all the scenarios clears the way for associating what can be called guaranteed regret with each solution x . Taking guaranteed regret as the definition of the robustness measurement is the same as admitting that solution x is considered at least as robust as solution y if, and only if, the guaranteed regret in x is, at most, equal to the guaranteed regret in y .

Characteristic trait no. 3: The single selected robustness measurement only involves the scenarios where it leads to the worst performance, or greatest regret, wherein what happens in the other scenarios does not play any role. Robust solutions are those that optimize the robustness measurement in every scenario.

Coguaranteed robustness measurements, guaranteed (absolute or relative) regret, involve considering only what happens in the most unfavorable scenarios (in terms of performance or regret), regardless of what happens in the other scenarios, to determine whether solution x deserves to be qualified as robust. This is a very particular way of addressing the concern for robustness. It cannot correspond to the decision maker’s picture of robustness. A solution thus qualified as robust may incur costs or cause regrets that are very mediocre in large number of scenarios. Basing the concern for robustness on the optimization of a cost or a guaranteed regret means admitting that one should primarily protect oneself as best

as possible against what might happen if the actual scenario is the worst (in terms of cost or regret) *vis-à-vis* the selected solution. This way of addressing the concern for robustness is not appropriate for the decision maker unless he or she has a very strong aversion to risk.

Characteristic trait no. 3 assigns a decisive role to the scenario (or scenarios) that determines the optimum of the robustness measurement in the definition of a robust solution (or solutions). The scenario (or scenarios) may be highly unlikely (for instance, corresponding to combinations of extreme options). Removing the scenario (or scenarios) from set S may lead to defining very different solutions as robust. This dependence of the solution qualified as robust on the presence of this or that scenario in S can be clearly seen in the case of the design of a water supply network (the section titled “Determination of the Characteristics of a Water Supply Network in an Urban Development Zone”), when all the scenarios were designed as was suggested in the section titled “Two Requirements to Meet the Concern for Robustness Properly” (see the end of the first observation).

OTHER WAYS OF DEFINING THE CONCEPT OF “ROBUST SOLUTIONS”

The above considerations explain why several authors have suggested definitions of “robust solutions” by avoiding characteristic trait no. 3 (Refs 5 and 11). I have suggested three new robustness measurements (*bw*-absolute robustness, *bw*-absolute deviation, and *bw*-relative deviation). Their purpose is to allow the decision maker to improve the way he or she would like to reconcile the two conflicting risks (see the end of the section titled “Two Requirements to Meet the Concern for Robustness Properly”). To do so, he or she must assign a value to two parameters, b and w ; b refers to a goal (in terms of cost or regret) that has to be reached or improved in the highest possible number of scenarios, whereas w sets the value of the worst cost or greatest regret that he or she would like to guarantee, regardless of which scenario actually happens ($w > b$). The decision maker who

agrees to assign a higher value to w than to w^* , guaranteed by solution x that optimizes the measurement of the related cost or guaranteed regret, obviously runs the risk that solution z (which is robust from the standpoint of the new selected measurement) may lead to a cost or regret higher than w^* , in a few scenarios. However, with solution z , the decision maker can trust that the costs or regrets will be better than what they would have been in solution x , in a very high number of scenarios (for more details, see Ref. 6).

Very few authors have attempted to define the robustness of a solution by involving several, rather than a single, robustness measurement (in other words, by avoiding characteristic trait no. 2). There are also very few authors who have attempted to address the concern for robustness, by avoiding characteristic trait no. 1. Readers may refer to the above-mentioned references for further details on both points.

Robustness can also be addressed by looking for types of answers that do not necessarily highlight solutions qualified as “robust.” Providing the decision maker with the so-called robust solutions is actually a very particular type of response, which does not always properly meet the second requirement (the section titled “Two Requirements to Meet the Concern for Robustness Properly”). I will now deal with a more general type of response.

THE CONCEPT OF “ROBUST CONCLUSIONS”

I will begin by introducing several notations. Let us take scenario $s \in S$ that has a perfectly defined model $M(s)$ in FR, the said model translating the way that the DA problem was formulated in the case where the real-life context corresponds to scenario s . In other words, $p \in P$ (the section titled “Two Requirements to Meet the Concern for Robustness Properly”) is a procedure considered for processing $M(s)$. The processing provides a set of outcomes $R(p, s)$ that may have various forms: optimal solution(s), uncontrolled solution(s), rejected solution(s), and the establishment of an impossibility or inconsistency. The analyst

relies on the said results to answer the decision maker's questions. For various reasons, the analyst may only focus on the outcomes involving pairs (p, s) that belong to a given subset Q of the Cartesian product $P \times S$.

By definition, here the term *conclusion* means an assertion that considers a certain set of outcomes of the $R(p, s)$ type. This kind of assertion is called a *robust conclusion* when it includes the statement of the conditions governing the establishment of its validity. The conditions involve the relevant subset Q of $P \times S$ and the conditions and hypotheses that give meaning to, and warrant the assertion. Therefore, the robustness of a conclusion is contingent on the field of validity defined by the said conditions.

The most commonplace statement of robust conclusions is the following, "solution x is robust, robust meaning that x is a solution that optimizes the measurement of robustness defined as follows (...) in all the outcomes $R(p, s)$ stemming from the relevant set Q ." Here, the selected robustness measurement and the relevant set Q define the field of validity of the robust conclusion.

There are less trivial forms of robust conclusions that do not refer to "robust solution." Several examples are given below.

"Except for the following procedures (...), all the $(p, s) \in Q$ pairs highlight the following evaluations (...) as feasible with the following evaluations (...)."

" $\forall (p, s) \in Q, x$ is a solution whose gap with the optimum solution never exceeds . . . %."

"There is no feasible solution that ensures that the following goals (...) (associated with different criteria) will be guaranteed, regardless of scenario $s \in S$."

"All the solutions $x_1, x_2, \dots, x_k$ have the following properties (...), except perhaps in some scenarios that are deemed as highly unlikely."

I will briefly summarize the role that the concept of "robust conclusions" has actually played in drafting recommendations in the example in the section titled "Choosing the Best Bid after a Call for Tenders," to highlight the practical contribution of "robust conclusions."

The set of feasible solutions was defined in extension by the 15 selected bids. Each

bid was evaluated on 12 criteria. Six of the solutions could be initially eliminated. Therefore, the FR model relied on what is called the *performance table*, which provides the performance for each of the nine solutions on each of the 12 criteria. To evaluate the performance on a concrete scale appropriate to each criterion, a group of experts of the French Post was set up. The frailty points related to the share of uncertainty or even arbitrariness, which might affect some of the evaluations, were inventoried. Basically, their impact could be considered by means of the concept of thresholds of indifference or preference, which do not have to be known to understand the rest of the example.

The selected processing procedure was ELECTRE IS. In this procedure, a set of weights characterizing the relative importance that the decision maker wanted to assign to each criterion had to be defined. For some of the criteria, a veto threshold could be introduced. In this case, the decision maker was a committee comprising several post office managers. The role that had to be assigned to each of the 12 criteria was not perceived in the same way by the technical manager, personnel manager, sales & marketing manager, or financial manager. Therefore, weights and vetoes were the frailty points in the way the model was processed.

To understand the meaning of the robust conclusions that were drawn, I do not have to go into the details of what set Q of the relevant (p, s) pairs covered. For each pair, the ELECTRE IS method highlighted a subset of solutions that seem the most satisfactory. The said subset had to be as small as possible, with the following constraint: any solution that was not in the subset could be eliminated, as it was not as good as at least one of the solutions in the subset. The said subset [that here play the role of the $R(p, s)$ outcome] corresponded to what is called a *kernel* in graph theory. Consequently, the method associated a kernel with each pair $(p, s) \in Q$.

A summarized version of the robust conclusions that were drawn on these bases can be found below.

Conclusion no. 1: Bid a_1 was found in almost all the kernels (this assertion was completed by the list of (p, s) pairs where a_1 was not in the matching kernel).

Conclusion no. 2: Bids a_5, a_6 , and a_8 could not be found in any kernel.

Conclusion no. 3: Bid a_9 could be found in some kernels; when this was the case, lowering the veto power of criterion 12 was enough to remove a_9 from the kernel (resulting in a_1 becoming as good as a_9).

Conclusion no. 4: The three bids a_2, a_3 , and a_4 formed a subset C of actions, which were indifferent in each pairwise comparison, which could be found in a large number of kernels.

Saying that bids a_2, a_3 , and a_4 were indifferent in each pairwise comparison meant that, with the available data at time, they could not be differentiated. Each bid of C could be considered as at least as good as each of the two others, without being substantially better.

Conclusion no. 5: Bid a_7 could be found in a large number of kernels; the actions of C were usually incomparable with a_7 .

Affirming this incomparability meant that, with the available data at the time, a_7 could not be considered as at least as good as any of the actions of C and that, vice versa, none of these actions could be considered as at least as good as a_7 .

These robust conclusions (completed by an appropriate analysis) led to the following recommendations that were submitted to the post general directorate (for more details, readers may refer to Ref. 12, Chapter 8).

1. In the case where no more than one prototype could be built, bid a_1 should be retained (conclusion no. 1).
2. In the case where more than one prototype could be built, bids a_5, a_6, a_8 , and a_9 should definitely be eliminated (conclusion nos. 2 and 3).
3. In the case where three prototypes could be built, bids a_1 and a_7 and one of the three bids of C should be retained (conclusion nos. 4 and 5). The 12 relevant criteria did not clearly differentiate bids a_2, a_3 , and a_4 (to do so, one could examine considerations that were not

(or poorly) considered by this family of criteria). If a_1 was actually retained, selecting a_3 at the same time would be more justified than selecting a_2 or a_4 because unlike the two latter, a_3 was usually incomparable with a_1 .

4. In the case where two prototypes could be built, it would be appropriate, in addition to a_1 , to chose either a_7 or one of the C bids, as a_7 was hard to compare with C bids, an option which basically brought the "decision maker's" value system into play.

CONCLUSION

To conclude, I would like to underscore two points.

1. In OR-DA, the notion of robustness has relatively tight connections with other notions such as flexibility, stability, reliability, adaptability, sensitivity, and sometimes even equity.
2. Meeting the concern for robustness supposes considering the eminently subjective way that the decision maker wants to protect himself or herself from catastrophic performance without having to forsake the hope for good, or even very good performance.

REFERENCES

1. Roy B. Paradigms and challenges. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis - state of the art surveys Springer; 2005. p 3–24.
2. Roy B. Decision-aiding today: what should we expect? In: Gal T, Stewart TJ, Hanne T, editors. Multicriteria decision making—advances in MCDM models, algorithms, theory, and applications. Kluwer Academic Publishers; 1999.
3. Morse PM, Kimball GE. Methods of operations research. New York: John Wiley and Sons; 1951.
4. Miller DW, Starr MK. Executive decisions and operations research. 2nd ed. Englewood Cliffs: Prentice-Hall; 1969.
5. Roy B. To better respond to the robustness concern in decision aiding: four proposals based

- on a twofold observation. In: Zopounidis C, Pardalos PM, editors. *Handbook of multicriteria analysis*. Springer; 2010a. p 3–24.
6. Roy B. Robustness in operational research and decision aiding: a multi-faceted issue. *Eur J Oper Res* 2010b;200:629–638.
 7. Billaut JC, Moukrim A, Sanlaville E. *Flexibility and robustness in scheduling*. New York: Wiley-ISTE; 2008.
 8. Newsletter 2002–2008 of the European Working Group Multiple Criteria Decision Aiding No. 6–13. see <http://www.inescc.pt/~ewgmcd/Articles.html>. Accessed 2013 July 25.
 9. Rosenhead J. Robustness analysis: keeping your options open. In: Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflicts*. Chichester: John Wiley and Sons; 2001. p 181–207.
 10. Kouvelis T, Yu G. *Robust discrete optimization and its applications*. Boston (MA): Kluwer Academic Publishers; 1997.
 11. Aissi H, Roy B. Robustness in multi-criteria decision aiding. In: Ehrgott M, Figueira JR, Greco S, editors. *Trends in multiple criteria decision analysis*. Volume 142. Springer, Science + Business Media; 2010. pp. 87–121.
 12. Roy B, Bouyssou D. *Aide multicritère à la décision: méthodes et cas*. Paris: Economica; 1993.

RULE DEVELOPING EXPERIMENTATION IN CONSUMER-DRIVEN PACKAGE DESIGN

ALEX GOFMAN
Moskowitz Jacobs Inc., White
Plains, New York
Marketing Department, Pace
University, Manhattan,
New York

The consumers make about 73% of their purchase decisions at the point of sale where more attractive packaging frequently leads to purchase [1]. Particularly, when the consumer is undecided, the package becomes a critical factor in the purchase choice because it communicates to consumers at the decision making time [2]. In this case, the package functions as a silent salesman [3,4], which some also call *the first moment of truth* [5]. Silayoi and Speece [2], and McDaniel and Baker [6] suggest that how consumers perceive the subjective entity of products, as presented through communication elements in the package, influences their choice. In turn, this perception may be the key to success for many marketing strategies [7].

While on the shelf, the product competes with many other offerings. Although occasionally minimalist design is appropriate for some brands and packages [8], in the majority of cases the products must figuratively jump out at the consumer in order for the consumer to select it from the shelf [9]. Orth and Malke-witz [10], Schoormans and Robben [11], and Meyers and Lubliner [12] underscore the importance of package design and suggest that even the same products produced by different brands should use different package designs to appeal to consumers.

Underwood and Ozanne [4], Underwood *et al.* [13], and Ampuero and Vila [14] demonstrated that visuals on the package can be a strategic method to differentiate one's product as pictures are more effective

stimuli compared to words. Furthermore, consumers process visual information faster and easier, particularly in situations with low involvement with the product [15,16]. For example, a part of the visual image of a package, its colors, plays a very important role in consumer preferences. Grossman and Wisenblit [17] and Madden *et al.* [18] suggested that colors often create potentially strong associations for consumers, driving their brand preferences.

One of the ways to design winning packages involves *experimentation with consumers*. Experiments lie at the very heart of new product development (NPD) and thus are connected to corporate values, habits, strategies, and organizational structures [19–23].

A traditional way to experiment with package design usually involves acceptance testing on the late stages of design when consumers just choose between a few complete package executions created by a designer, usually in focus groups [24–27]. Silayoi and Speece [2] and Moskowitz *et al.* [28,29] criticize this approach as suffering from subjectivism of designer opinion leading to only very few final packages to be chosen for testing. They suggest that experimentation should start at early stages of development.

One of the forms of experimentation with consumers in the early stages to find their preferences is conjoint analysis [30–32]. Whereas, substantial research has been done in using conjoint analysis for package design [2,33–36], the field is still developing.

This article shows how structured experimentation (based on conjoint analysis) in the form of rule developing experimentation (RDE) provides designers with a disciplined approach in early stage development to choose from the possible design features and their combinations. These designs, created out of the designer's talent, are then systematically explored to discover what works for a package design and what does not [37–39].

RULE DEVELOPING EXPERIMENTATION AND ITS APPLICATION TO PACKAGE DESIGN

RDE [40–42] is a structured methodology of consumer-based experimentation with conceptual prototypes which is applicable to NPD and other areas. RDE is derived from the consumer-driven proactive approaches to structured experimentation researching consumer preferences on the conceptual level utilizing partial profile conjoint analysis with incomplete concepts (concepts having zero conditions [28]) based on a special type of fractional main effects, experimental designs—isomorphic permuted experimental designs (IPED) [43]. RDE was gradually built in parts by multiple parties and formalized into a cohesive approach in Gofman and Moskowitz [38,42,43], Moskowitz and Gofman [40,44,45], and Moskowitz *et al.* [46–48] in cooperation with Prof. J. Wind (Wharton Business School).

In a wider business meaning, RDE is a systematized solution-oriented business process of experimentation that designs, tests, and modifies alternative ideas, packages, products, or services in a disciplined way using experimental design, so that the developer and marketer discover rules and patterns showing what appeals to the customer, even if the customer cannot articulate the need, much less the solution [40].

In RDE, as in other conjoint analysis-based approaches, the input materials are organized into attributes—groups of similar themed elements [49]. RDE usually utilizes Plackett Burman [50] or similar experimental designs with the elements arranged into experimentally designed prototypes (concepts) based on IPED. The IPED facilitates individual (for each respondent) unique sequences of prototypes (subject to statistically possible limits depending on the number of the categories and elements), creating a platform of conceptual prototypes testing in widely varied contexts of different combinations. This helps to overcome multiple problems of the traditional conjoint analysis approaches identified in Moskowitz *et al.* [28] and Gofman [41]. Particularly, due to having zero conditions, the IPED

enable conjoint analysis to estimate the absolute values of the utilities rather than the relative values, which are utilized in the majority of the existing methods [41,43,51]. These conceptual prototypes are presented to a representative group of consumers (respondents) to elicit their preferences (e.g., purchase intent or liking).

The solicited ratings are combined with the underlying IPED and are analyzed through a regression analysis, imputing the contribution of each individual element to predicted consumer preferences, and thus creating individual sets (for each respondent) of the utilities (part-worth contributions) of the elements also called *impacts*.

The resulting quantitative relationships between the elements and consumer preferences (rules) could be utilized by corporations to provide directions of how to increase consumer liking (purchase intent) under different scenarios. RDE quantifies the most likely combinations of ideas (features or messages) that may be acceptable to the consumer.

RDE application to graphics design is an extension of RDE methods involving just text and individual pictures, which has been widely used for the last three decades [46,48,52,53]. In package design, RDE does not replace the creative involvement. Instead, it allows a designer to sift through multiple options to facilitate the creation of new and attractive packages based on actionable consumer insights. To find a right package design and graphics, one follows the same six RDE steps used in RDE message optimization [40,54].

Step 1: Preparing Raw Materials. Identification of groups of features that constitute the target product (different options of the package features).

Step 2: Creating Test Stimuli. Process of mixing and matching the elements according to an experimental design to create a set of conceptual prototypes.

Step 3: Collecting Data with the Consumers. The actual experimentation with the consumers soliciting their responses to experimentally designed prototypes.

Step 4: Analyzing Results. Individual models for each respondent allow discovery of patterns in the data.

Step 5: Identifying Pattern-Based Latent Segments and Interactions. People's preferences differ regardless of the demographic group [55]. RDE identifies naturally occurring pattern-based latent segments of the population that show similar patterns of the utilities as well as detects any and all interactions between the elements [42,45,51].

Step 6: Applying the Generated Rules to Create New Products, Services, Offerings. Utilization of the results of RDE to optimize the ideas in the prototypes, creating a set of quantitative relationships between the elements of RDE and consumer preferences, and directions of how to increase consumer liking [54,56–60].

EMPIRICAL STUDY

The case study demonstrates this approach using the example of a package for yogurt. Rather than create a single package for evaluation, the designer creates a template with various features (e.g., picture of the flavor ingredient), and several options for each specific feature in the template. The goal then is to identify what drives consumer interest, and what does not.

RDE requires the creation of multiple packages based on an experimental design. In the example described, the target package is divided into four areas (Fig. 1): main picture (flavor), fat content, health message, and live cultures. All of the alternative executions of a specific feature should match the outlines of each other, and fit the outlining package (in the center of Fig. 1).

Each feature of the package can be thought of as a transparency layer, somewhat similar to Adobe Photoshop layers. The features of the project are transparent everywhere except for the key object of the layer. For example, in Fig. 2 you can see the yogurt package without variable features (template) and four layers with individual features positioned on each other. The outlines of the

package in each layer in Fig. 2 are shown just to demonstrate the component's positioning on the background package. In actual layers, the area outside a particular component is completely transparent.

Each option was created on a transparent layer with the feature positioned properly to allow correct matching during overlaying. All the surface bending was consistently captured inside each option based on a single template, shown in the middle of Fig. 1. The result is a set of graphically realistic pictures, assembled by the principle of RDE.

The computer (browser) superimposes these transparencies according to the RDE design, thus creating different executions of the experimental packages in real-time. Each new combination defined by the RDE design corresponds to a different package. Over the course of the interview, the participant evaluates many different combinations of options. The participant never sees the individual transparencies, but only complete packages (as on the right of Fig. 2).

A very important difference from a text-based RDE is the need for a filler background image for the situations when the package on the screen occasionally must be without an option, as required by the experimental design having zero conditions. The currently accepted solution [37,39,61,62] places a bare package (package without variable features) in the back of each screen, behind all the layers. This way, a zero condition (the absence of an option in the design) does not create a disturbing image with "holes" in it. The background image of the bare package makes the test stimulus on the computer screen be meaningful to consumers, even if it occasionally has an element missing.

Analysis

The objective of RDE analysis is to estimate the impact of each individual element (and optionally the interactions between them) along with the additive constant which could be interpreted as the baseline interest in the product (in the context of the specific experiment). Adding the individual impacts of the selected elements to the corresponding baseline allows a designer to compare the estimated interest (or purchase intent) of

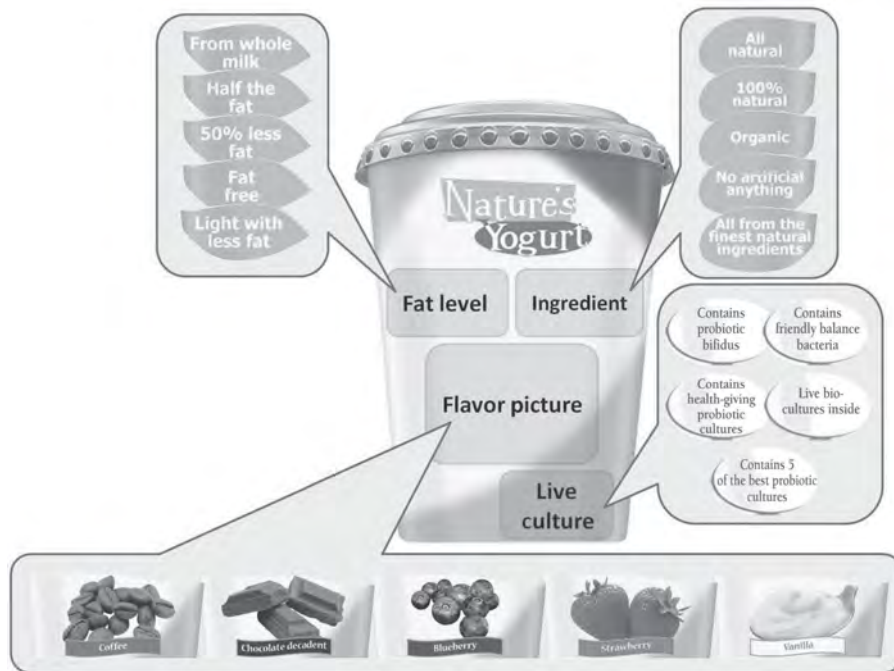


Figure 1. The layout of the yogurt package with placeholders for four attributes and individual elements (not to scale).

the packages optimized for different groups of consumers.

RDE generates individual equations or models relating the presence/absence of each of the design features to the respondent's rating. The model, thus, shows how every package element drives the response (e.g., a rating "1 = Not interested → 9 = Very interested").

Different RDE applications might use various scales and recordings. In many RDE, the respondent ratings are assigned on a 9-point scale. In this case-study, the 1–9 ratings are recoded into binary values: ratings of 1–6 are recoded to 0 (the respondent was not interested, or marginally interested); ratings of 7–9 are recoded to 100 (the respondent was interested).



Figure 2. The yogurt package without variable features (left), and four layers with individual features positioned on each other and the resulting package (right) (the outlines of the package on the layers are shown just to demonstrate the component's positioning on the background package. Gray area is transparent).

Table 1. Norms for the Additive Constant and the Utilities.

<i>Additive</i>	
<i>Constant</i>	<i>Interpretation</i>
>60	Respondents are very predisposed to the product
50–60	About half of the respondents are very positive
40–50	Respondents accept the idea
30–40	The elements need to do the work
<30	The product is a commodity and the elements must do the work
<i>Utility score</i>	
	<i>Interpretation</i>
>15	The element performs exceptionally well
10–15	The element performs well
5–10	The element breaks through the clutter
0–5	The element is barely effective, if at all
<0	The element actively detracts from acceptance

[Source: Moskowitz *et al.* [28]]

The key to understanding the data for product and package development is the impacts of the individual elements. The norms for interpreting utility scores and additive constants are listed in Table 1 (assuming the above conversion to 0–100 rating scale). Here the numbers are conditional probability (percentage) of people being interested in buying the product described. Additive constant in conjoint analysis with incomplete concepts is a baseline interest of respondents in the idea of the product (without any elements present).

Table 2 contains the results of the regression analysis of the data collected for the case study.

Total Panel (Average across All Respondents). For the total panel, the constant (a general predisposition to purchase the product if no elements are shown) is very low (+8), which is normal for graphical RDE projects (a package without elements makes little sense) [28,37,56]. The best performing elements (by far) are the flavor pictures with Strawberry and Blueberry on the top of the list (+32 and +29, respectively) with Chocolate slightly trailing them (still a very

impactful +21). Among the flavors, Coffee is the lowest performing element yet with a respectable +10 impact.

The rest of the elements have impacts in the single digits with “contains health-giving probiotic cultures,” “all from the finest natural ingredients,” and “fat free” on the top of the list (excluding the flavors). Figure 3 shows a graphical representation of the impacts of the elements (sorted by the value of the impacts).

Optimizing for the Genders. There are some notable differences between the genders in how they perceive the elements in this study, although the genders were not represented equally in the panel. The constant is low and almost the same for males and females. The biggest differences are observed in how flavors influence their purchase intent. Whereas Coffee has almost the same moderate impact (+9 and +10 for males and females, respectively), the rest of the flavors show much bigger differentiation. The most popular overall flavor, Strawberry, increases men’s purchase intent by a whopping +41% and women’s by +30%. Similarly, Blueberry has +38% impact on men versus +27% on women. Chocolate is more liked by females (+22 vs +16). The least exciting is Vanilla (+8 for men vs +15 for women). Other differences are less expressed (possibly due to an uneven number of respondents in each of the groups).

Subgroups Analysis. At the end of the survey, after the RDE part, respondents answered some demographics and self-profiling questions. One of the questions was how they define their diet—“Not at all healthy,” “Somewhat healthy,” “Healthy,” and “Very healthy.” For the purposes of the analysis, we combined the first two options into one “Less healthy diet” and the last two options into “More healthy diet.”

The constants are quite different for the groups. Less healthy-oriented dieters have a much higher predisposition to buy yogurt as compared to more healthier-oriented ones. Possibly, people on a good diet are not looking for a new yogurt as they might have a preferred brand already, whereas people

Table 2. Performance of Options for the Yogurt Packaging Study (Sorted by Element Code).

		Gender			Diet		Segments	
		Total	Male	Female	Less Healthy	More Healthy	Seg 1	Seg 2
	Base size:	140	27	113	76	62	67	73
	Constant:	8	6	8	18	-4	7	8
A1	From whole milk	2	-3	4	2	4	1	4
A2	Half the fat	4	4	4	3	5	1	7
A3	50% less fat	7	6	7	5	10	2	12
A4	Fat free	8	9	8	5	13	9	7
A5	Light with less fat	4	3	4	5	4	0	8
B1	All natural	5	5	5	4	8	-2	12
B2	100% natural	7	12	6	5	10	0	13
B3	Organic	2	2	2	1	4	-7	10
B4	No artificial anything	6	9	6	4	11	-1	14
B5	All from the finest natural ingredients	8	7	8	5	12	1	15
C1	Chocolate	21	16	22	20	24	35	8
C2	Coffee	10	9	10	4	17	21	-1
C3	Blueberry	29	38	27	21	40	48	13
C4	Strawberry	32	41	30	28	36	52	14
C5	Vanilla	14	8	15	10	18	23	5
D1	Contains friendly balance bacteria	5	8	4	2	8	2	8
D2	Contains probiotic bifidus	7	10	6	4	10	4	10
D3	Contains health-giving probiotic cultures	9	9	9	5	13	4	13
D4	Live bio-cultures inside	3	8	2	1	6	2	5
D5	Contains five of the best probiotic cultures	6	7	5	4	8	1	10

Notes: The numbers are conditional probability of the elements increasing purchase intent of respondents if shown. The constant is the baseline purchase intent if no elements are shown. "Less healthy" diet is a combination of respondents self-reported having a nonhealthy and somewhat healthy diets. "More healthy" diet is a combination of respondents self-identified having health and very health diets. Seg 1 is Segment 1 "Flavor-centric." Seg 2 is Segment 2 "Health oriented."

on a less healthy diet might think about improving it. Furthermore, as compared to less healthy dieters, more healthy eaters prefer many healthier options such as fat free yogurt (+13 vs +5); 100% natural (+10 vs +5); all natural (+8 vs +4); 50% less fat (+10 vs +5). In fact, all the elements in the ingredient and fat level attribute are preferred by healthier eaters versus less healthy dieters, which make sense.

The biggest differences are in the flavor category. In absolute values, more healthy dieters prefer every flavor more versus unhealthy eaters. By far, berries increase purchase intent in a more healthy diet (Blueberry is +40 vs +21 and Strawberry +36 vs +28). Chocolate seems to be liked

very much by both groups but not to the level of berries (+24 vs +20).

In addition to insights obtained from this particular analysis, the latter tabulation of the data provides a common sense validation of the approach.

Mind-Set Segmentation. Consumers are not all created the same. People's preferences differ substantially. Most traditional approaches divide people based on some demographic, lifestyle, or purchase behavior criteria [63–65]. A more effective approach for design and development divides people based on their mind-sets emerging from the patterns of their utilities in conjoint analysis projects [49]. People in the same mind-set

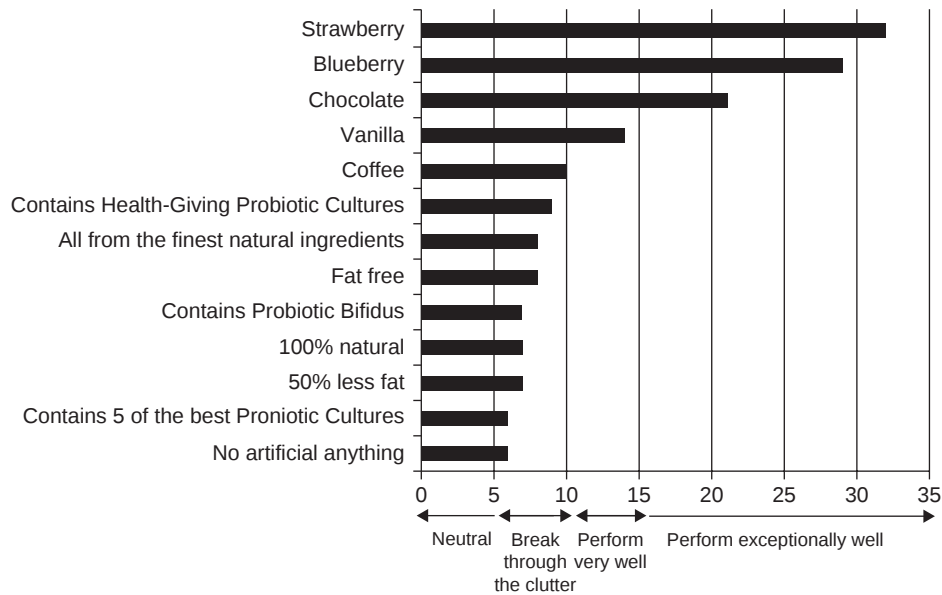


Figure 3. Performance of the elements for the total panel sorted by the element code (only elements that are not neutral are shown).

segment like similar package design features or products. We can cluster respondents based on the patterns of their responses and try to optimize the package for each segment [28].

The segmentation analysis revealed two substantially different mind-sets of the consumers. Both segments have virtually the same constant (+7 and +8). Yet, there the similarities end.

Segment 1: Flavor-Centric. likes virtually all flavors (which explains its name)—from a phenomenal +52 (Strawberry) and +42 (Blueberry) to superior +35 (Chocolate) to excellent +23 (Vanilla) and +21 (Coffee). By far, this segment prefers berry-flavored yogurt. Nothing else seems to interest this segment in yogurt packages. Except for a moderately positive “fat free” feature (+9), the rest of the elements are neutral for this segments, except for a surprising (−7) for “organic.” Possibly, this segment is skeptical to this type of marketing messages. This might explain its indifference to the rest of them as well.

Segment 2: Health-Oriented. is a very different group of people. Although it does not show such huge impacts as Segment 1 does for flavors, the segment is very consistent in its preferences. Everything that communicates health scores positively with the people from this segment. All elements from the “ingredients” (attribute B) scored high: “All from the finest natural ingredients” (+15) to “organic” (+10). Although Segment 1 does not pay as much attention to flavors, two of the elements from attribute C (“Flavors”) scored high: “Strawberry” (+14) and “Blueberry” (+13). Those are the only flavors that are associated with healthy diets. Healthy ingredients (attribute D) are also important for this segment. Four out of five elements scored high: “contains health-giving probiotic cultures” (+13) to “contains friendly balance bacteria” (+8). Fat level (attribute A) is also an important issue for them. “50% less fat” adds +12 points to the purchase intent. Except for the “from whole milk” which has a neutral impact for this segment, the

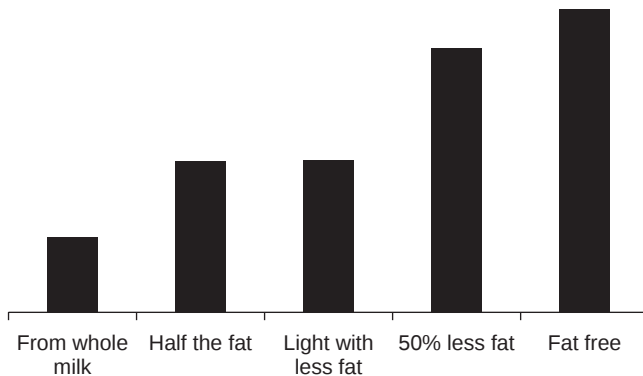


Figure 4. Performance of the elements related to fat content (shown for the total panel; consistently representative for all other subgroups).

rest of the elements from attribute A are positive (+7 to +8).

This type of segmentation afforded by individual models in RDE are actively utilized by businesses. In a typical case, a manufacturer would consider creating a separate product for each segment to maximize the appeal to the different mind-sets of the consumers. Otherwise, it will still do well creating a single product optimized for the total panel. In that case, although capturing fewer buyers than with individually optimized packages for each segment, it will create a more appealing product than one based on a pure guess.

Effect of the Fat Content Messages Wording.

A product developer could use the data to find better manufacturing options, while a marketer explores more efficient messages. For example, Fig. 4 shows the findings related to the fat content messages.

The relative values of impacts in Fig. 4 are consistent across all the subgroups. Whereas "Half the fat" is virtually identical to "Light with less fat," "50% less fat" is clearly a more efficient driver of purchase intent than the semantically identical "Half the fat" message for all the subgroups. "Fat free" is only marginally better performing comparing to the latter messages.

In the context of the current study, the respondents clearly prefer concrete, numeric presentation of the fat content over its equivalent verbal presentation. Yogurt manufacturers could choose the "50% less fat" option to capture more buyers.

SUMMARY

The utilization of RDE to graphical elements allows applying the power of conjoint analysis to understand consumer preferences of packages in the early stages of development. Following the idea of layers and applying individual permuted experimental designs to sets of the layers, allows for dynamic creation and evaluation of multiple experimentally designed packages, thus reducing the subjectivity of the designer through deeper consumer involvement in the process in the early stages.

The process should be viewed as providing important and timely input about consumer preferences to designers. It does not replace the artistic talent in any way. Quite to the contrary, it helps to narrow the multiple choices available for the designer. Particularly, this approach could be utilized as a parsimonious way to create appealing products in mature categories where differentiation is difficult and the budgets are small.

This approach resolves the problems related to consumers being unable to describe what they need. At the same time, it overcomes the complex and intervened statistical problems related to traditional methods of conjoint analysis utilizing a single experimental design for a project.

REFERENCES

1. Rettie R, Brewer C. The verbal and visual components of package design. *J Prod Brand Manage* 2000;9(1):56–70.

2. Silayoi P, Speece M. The importance of packaging attributes: a conjoint analysis approach. *Eur J Mark* 2007;41(11/12):1495–1517.
3. Pilditch J. The silent salesman: how to develop packaging that sells. New York, (NY): Random House Business Books; 1973.
4. Underwood R, Ozanne J. Is your package an effective communicator? A normative framework for increasing the communicative competence of packaging. *J Mark Commun* 1998;4(4):207–220.
5. Lofgren M, Witell L, Gustafsson A. Customer satisfaction in the first and second moments of truth. *J Prod Brand Manage* 2008;17(7):463–474.
6. McDaniel C, Baker R. Convenience food packaging and the perception of product quality. *J Mark* 1977;41(4):57–58.
7. Cunningham R, Kyle K. The case for plain packaging. *Br Med J* 1995;4(1):80–86.
8. Acevedo J. Message in a bottle. *Brand Packaging Magazine* 2008;12(7):18.
9. Mc Loone C. A design process success story. *PackagePrinting* 2009;56(6):20, 22–24.
10. Orth U, Malkewitz K. Holistic package design and consumer brand impressions. *J Mark* 2008;72(3):64–81.
11. Schoormans J, Robben H. The effect of new package design on product attention, categorization and evaluation. *J Econ Psychol* 1997;18(2–3):271–287.
12. Meyers H, Lubliner M. The marketer's guide to successful package design. New York (NY): McGraw-Hill; 1998.
13. Underwood R, Klein N, Burke R. Packaging communication: attentional effects of product imagery. *J Prod Brand Manage* 2001;10(7):403–422.
14. Ampuero O, Vila N. Consumer perceptions of product packaging. *J Consum Mark* 2006;23(2):100–112.
15. Pieters R, Warlop L. Visual attention during brand choice: the impact of time pressure and task motivation. *Int J Res Mark* 1999;16(1):1–16.
16. Pieters R, Wedel M. Attention capture and transfer in advertising: brand, pictorial, and text-size effects. *J Mark* 2004;68(2):36–50.
17. Grossman R, Wisenblit J. What we know about consumers' color choices. *J Mark Pract: Appl Mark Sci* 1999;5(3):78–88.
18. Madden T, Hewett K, Roth M. Managing images in different cultures: a cross-national study of color meanings and preferences. *J Int Mark* 2000;8(4):90–107.
19. Von Hippel E. Democratizing innovation. Cambridge (MA): MIT Press; 2005.
20. Von Hippel E, Katz R. Shifting innovation to users via toolkits. *Manage Sci* 2002;48(7):821–833.
21. Thomke S. Enlightened experimentation: the new imperative for innovation. *Harv Bus Rev* 2001;79(2):66–75.
22. Thomke S. Experimentation matters: unlocking the potential of new technologies for innovation. Cambridge (MA): Harvard Business Press; 2003.
23. Kahn K, Barczak G, Moss R. Perspective: establishing an NPD best practices framework. *J Prod Innov Manage* 2006;23(2):106–116.
24. DuPuis S, Silva J, Braue K. Package design workbook: the art and science of successful packaging. Minneapolis (MN): Rockport Publishers; 2008.
25. Kees J, *et al.* Tests of graphic visuals and cigarette package warning combinations: implications for the framework convention on tobacco control. *J Public Pol Mark* 2006;25(2):212–223.
26. Krueger R, Casey M. Focus groups: a practical guide for applied research. Newbury Park (CA): Pine Forge Press; 2008.
27. Wever R, *et al.* Sales performance of packaging for consumer electronics products. In the Proceedings of the 23rd IAPRI symposium on packaging, September 3–5, 2007, Windsor, UK; 2007.
28. Moskowitz H, Porretta S, Silcher M. Concept research in food product design and development. Ames (IO): Blackwell Pub.; 2005.
29. Moskowitz H, *et al.* Packaging research in food product design and development. Ames (IA): Wiley-Blackwell; 2009.
30. Green P, Srinivasan V. Conjoint analysis in consumer research: issues and outlook. *J Consum Res* 1978;5(2):103.
31. Cattin P, Wittink D. Commercial use of conjoint analysis: a survey. *J Mark* 1982;46(3):44–53.
32. Carroll J, Green P. Guest editorial: psychometric methods in marketing research: Part I, conjoint analysis. *J Mark Res* 1995;32(4):385–391.
33. Green P, Srinivasan V. Conjoint analysis in marketing: new developments with implications for research and practice. *J Mark*

- 1990;54(4):3–19.
34. Balakrishnan P, Jacob V. Genetic algorithms for product design. *Manage Sci* 1996;42:1105–1117.
35. Green P, Krieger A, Wind Y. Thirty years of conjoint analysis: reflections and prospects. *Interfaces* 2001;31:56–73.
36. Kotri A. Analyzing customer value using conjoint analysis: the example of a packaging company. University of Tartu-Faculty of Economics and Business Administration Working Paper Series. 2006.
37. Gofman A. Getting the package and website graphics right with consumer co-creation. In: Moskowitz H, Saguy S, Straus T, editors. *An Integrated approach to new food product development*. Boca Raton (FL): CRC Press/Taylor & Francis Group; 2009. pp. 369–386.
38. Gofman A, Moskowitz H. Developing from the ground up: self-authoring systems for text and package concepts in concept research. In: Moskowitz H, Porretta S, Silcher M, editors. *Concept research in food product design and development*. Ames (IA): Blackwell Pub.; 2005. pp. 283–322.
39. Gofman A, Moskowitz H, Mets T. Accelerating structured consumer-driven package design. *J Consum Mark* 2010;27(2):157–168.
40. Moskowitz H, Gofman A. *Selling blue elephants: how to make great products that people want before they even know they want them*. Philadelphia (PA): Wharton School Publishing; 2007.
41. Gofman A. Experimentation-based product development in mature food categories: advancing conjoint analysis approach. Tartu, Estonia: Tartu University Press; 2009.
42. Gofman A, Moskowitz H. Steps towards a consumer-driven innovation machine for ‘ordinary’ product categories in their later lifecycle stages. *Int J Technol Manage* 2009;45(3):349–363.
43. Gofman A, Moskowitz H. Isomorphic permuted experimental designs and their application in conjoint analysis. *J Sens Stud* 2010;25(1):127–145.
44. Moskowitz H, Gofman A. System and method for content optimization. Google Patents. 2003.
45. Moskowitz H, Gofman A. System and method for performing conjoint analysis. Google Patents. 2005.
46. Moskowitz H, Gofman A, Beckley J. Using high-level consumer-research methods to create a tool-driven guidebook and database for product development and marketing. *J Sens Stud* 2006;21(1):54–100.
47. Moskowitz H, *et al.* Founding a new science: mind genomics. *J Sens Stud* 2006;21(3):266–307.
48. Moskowitz H, *et al.* Rapid, inexpensive, actionable concept generation and optimization: the use and promise of self-authoring conjoint analysis for the food service industry. *Food Serv Technol* 2001;1(3):149–167.
49. Krieger A, Green P, Wind Y. *Adventures in conjoint analysis: a practitioner’s guide to trade-off modeling and applications*. Philadelphia (PA): The Wharton School of Business; 2004.
50. Kotz S, Johnson N, Read C, editors. *Multivariate analysis—Plackett and Burman designs*, *Encyclopedia of statistical sciences*. New York: John Wiley & Sons, Inc.; 1985.
51. Gofman A. Emergent scenarios, synergies and suppressions uncovered within conjoint analysis. *J Sens Stud* 2006;21(4):373–414.
52. Moskowitz H, Gofman A. The DesignLab paradigm for product creation: merging consumers, creatives, computers for the design of a children’s cereal product. In: Porretta S, editor. *Consumer preference & sensory analysis*. The Netherlands: Miller Freeman Technical Ltd.; 1996. pp. 21–39.
53. Moskowitz H, Gofman A, Tungaturthy P. Multi-media development and optimization of concepts for a fast food restaurant. In: Edwards J, editor. *Culinary arts and science: global and national perspectives*. Southampton: Computational Mechanics Publications; 1996. pp. 101–116.
54. Gofman A, Bevolo M, Moskowitz H. Role of corporate leadership and innovation claims on consumer perception of premium products. *Int J Innov Manage* 2009;13(04):665–681.
55. Green P, Krieger A. Segmenting markets with conjoint analysis. *J Mark* 1991;55(4):20–31.
56. Gofman A, Moskowitz H, Mets T. Integrating science into web design: consumer-driven web site optimization. *J Consum Mark* 2009;26(4):286–298.
57. Gofman A. Consumer driven multivariate landing page optimization: overview, issues, and outlook. *IPSI BgD Trans Internet Res* 2007;3(2):7–9.
58. Gofman A, Moskowitz H. Improving targeting of customers with rule developing experimentation and short intervention testing. *Int J*

- Innov Manage 2010;14(3):435–448.
59. Moskowitz H, *et al.* Effective and confident communications in the midst of a major crisis: an experiment in the pharmaceutical context. *Int J Pharm Healthc Mark* 2007;1(4):318–348.
60. Moskowitz H, *et al.* Research, politics and the web can mix: considerations, experiences, trials, tribulations in adapting conjoint measurement to optimizing a political platform as if a consumer product. In: Brookes R, editor. *Marketing research in a. com environment*. Amsterdam, The Netherlands: ESOMAR; 2000. pp. 223–243.
61. Gofman A. Combining eye tracking with experimental design. In: Moskowitz H, *et al.*, editors. *Packaging research in food product design and development*. Ames (IA): Wiley-Blackwell; 2009. pp. 230–246.
62. Gofman A, *et al.* Extending rule developing experimentation to perception of food packages with eye tracking. *Open Food Sci J* 2009;3:66–78.
63. Plummer J. The concept and application of life style segmentation. *J Mark* 1974;38(1):33–37.
64. Diamantopoulos A, *et al.* Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation. *J Bus Res* 2003;56(6):465–480.
65. Orth U, *et al.* Promoting brand benefits: the role of consumer psychographics and lifestyle. *J Consum Mark* 2004;21(2):97–108.

RUSSIAN SCIENTIFIC OPERATIONS RESEARCH SOCIETY

YURY MALASHENKO
Operations Research Department,
Computing Center of Russian
Academy of Science, Moscow,
Russia

DEVELOPMENT OF OPERATIONS RESEARCH SURVEY

Russian Scientific Operations Research Society (RSORS) was organized at the First International Moscow Conference on Operations Research (OR) in 1996. This event was an important stage in rapid successful development of OR in the Soviet Union and consequently in the Russian Federation (RF).

In 1964, under the patronage of the President of the Academy of Sciences (AS) of the Soviet Union, a department of applied problems was created. It was headed by Popov and Pospelov, who initiated investigations on OR within the AS. The investigations were conducted in three AS institutes: Computing Center Academy of Sciences (CCAS) in Moscow (where this direction was headed by Moiseev), Mathematical Institute of AS Siberian branch in Novosibirsk (where Zhuravlev led the research), and Cybernetics Institute of Ukrainian AS in Kiev (founded by Glushkov). Additionally, the AS council, which was headed by Glushkov was established, and publications of OR periodicals began.

The department of OR was established in CCAS in 1966. It was run by Germeier. He also led the OR department created at Lomonosov Moscow State University (MSU) as a division of Computation Mathematics and Cybernetics Faculty in 1970. The technical committee of the OR society of the Soviet Union was founded at the CCAS in 1972 and registered in IFORS. Moiseev was appointed its chairman.

The collapse of the USSR caused scientific links between the former Soviet Republics to disappear resulting in the destruction of many scientific societies. Some OR scientists went on to solve Russia's organizational and technical problems while others further continued testing the OR theory. Many of these scientists soon realized the necessity of broader scientific contacts and restoration of the former scientific ties. Russian scientists desired to unite their scientific efforts and enter into international exchange of ideas, research results, and experiences encompassing OR.

ESTABLISHMENT OF THE SOCIETY, ORGANIZATIONAL STRUCTURE, AND PUBLICATIONS

In 1996, the 50th year since the formulation of OR, the first Moscow International Conference on OR (10–13, April 1996) was held in Russia under the sponsorship of the Russian Foundation for Basic Research and the Russian Academy of Sciences (RAS). The motto of the conference was "History. Beginnings. Unification." The RSORS was established at this conference, and its governing and controlling bodies were elected. The conferences' documents and theses were published and are now available on the Internet [1]. In particular, they contain protocols, the charter, the list of founders, and other documents of the RSORS. Moiseev was elected the Honorary President, and Krasnoschekov was elected President of the RSORS.

The Moscow International Conference on OR became triennial. Together with the Conference, RSORS holds the meetings of its members and elections. Thus, every third year the Conference elects the Council and the Council elects the President of the RSORS. In the conferences held so far (second—November 17–20, 1998; third—April 4–6, 2001; fourth—September 21–25, 2004; fifth—April 10–14, 2007), Krasnoschekov was reelected President; he has thus been

working continuously in the same location for more than 13 years.

At present, the organizational structure of RSORS is as follows: The President of RSORS is Krasnoschekov, the Vice-Presidents are Savin and Malashenko, the Chairman of the Council is Zhuravlev, the Chairman of the Auditing Committee is Flerov, the Scientific Secretary of RSORS is Pospelova, and Nazarova is a secretary. Office of RSORS: off. 230, Vavilov str, 40, Moscow, 119 333, Russia. Phone: +7(499)135-01-70. Fax: +7(499)135-54-29. Organizational web site of RSORS is <http://www.ccas.ru/RSORS>

Both individual scientists and scientific groups hope that the society will effectively aid in the international exchange of ideas and new results accumulated during the recent years. OR specialists in Russia have a wide range of interests and the new scientific society should reflect this variety and use it for the benefit of the science.

Scientific works in the various spheres of OR are mostly published in the following scientific journals of RAS: *Journal of Applied and Industrial Mathematics* [2] and *Journal of Computer and Systems Sciences International* [3]. In addition, the *Journal of Mathematical Modeling* [4] contains a special OR subdivision. The *Journal of Computational Mathematics and Mathematical Physics* [5] calls attention to the works on optimization methods. *Moscow University Computational Mathematics and Cybernetics* [6] is another journal containing a special OR section. Most fundamental OR works are published by the CCAS in the sequential issues of *Communications on Applied Mathematics* [7]. Besides the Moscow International Conference, RSORS permanently holds conferences on OR under the sponsorship of the Siberian branch of AS [8].

SCIENTIFIC ACHIEVEMENTS OF THE RUSSIAN SCIENTIFIC OPERATIONS RESEARCH SOCIETY

Germeier contributed greatly to OR theory through his formulation and development of the principle of guaranteed result [9].

He investigated nonantagonistic games, where players have incomplete information on payoff functions of each other. Together with Moiseev, he set the foundation of hierarchical games [10] (see also the sections titled "Game Theory" and "Game Theory: Extension and Development" in this encyclopedia).

The famous Russian scientist, Moiseev made significant contribution to OR through the development of methods of analysis of earth's biosphere and environmental protection. The estimation of consequences of nuclear war obtained on the basis of his model and under his leadership was recognized worldwide as the nuclear winter phenomenon [11]. This result can be considered as one of the most outstanding examples of practical applications of the OR theory (see also the sections titled "Environmental Planning and Energy" and "Military Applications" in this encyclopedia).

The design of complex engineering systems gives rise to multicriterion tasks. Solving these tasks is a difficult mathematical problem. The well-known scientist, Krasnoschekov and his coworkers proposed mathematical methods of effective (Pareto optimal) solution comparison [12]. The new approach permits to select the most promising options from the set of Pareto optimal solutions. This was a fundamental contribution to OR methods. Krasnoschekov has also contributed immensely to OR theory through his principles of mathematical modeling [13]. These ideas were successfully implemented into computer-aided design systems for aircraft, financial markets, semantic search, and analysis of national language texts (information technology "Keys to Texts" [14]), gas and oil field development (see also the section titled "Simulation Modeling and Analysis" in this encyclopedia).

A new direction of research emerged under the leadership of Petrov, in the CCAS, in 1975. The system analysis of developing economy [15] provides a unique platform for synergy of methods of mathematical modeling of complex systems and achievements of the modern economic theory. These models were

similar to computable general equilibrium models, which became popular in the 1990s, and paid more attention to singularities of economic mechanisms of the system under consideration. The interval logic of seemingly paradoxical economic processes became evident because of these models. These works converted the “annals” of the Russian economic reforms into the language of mathematical models [16,17] (see also the section titled “Simulation Modeling and Analysis” in this encyclopedia).

The unified model of analysis of efficiency and vitality of multicommodity networks was developed in the OR department of CCAS recently. Classical OR constructions and methodological approaches (flows in networks, game theory, max-min principle, and the principle of guaranteed result) were exploited and developed in the framework of the model for the first time. On the basis of this model, selected problems of analysis and synthesis, reliability and vulnerability of telecommunication, and power networks were successfully solved [18]. One of the most recent developments is application of the OR methodology in the power energy sector of the RF [19] (see also the sections titled “Graph Characterizations,” “Computer and Network Reliability” and the article in this encyclopedia).

OR models often result in rather complicated and sophisticated problems for which special numerical methods are required. Many RSORS members specialize in developing such methods (see Refs 20 and 21). Germeier’s works in game theory are now being developed further by Kukushkin and his other pupils [22]. The present chair of the MSU OR department, Krasnoschekov and his collaborators explore evolutionary models of population behavior and their applications in political science (social behavior), auction theory for homogeneous goods markets, option theory, tax enforcement problems, actuarial mathematics, and optimization theory and techniques. Some of the results obtained are published in the above-mentioned journals and the conference materials; also see Refs 23–29, and the sections titled “Collective Choice

and Social Utility Theory” and “Financial Engineering” in this encyclopedia.

The work of the OR department at the MSU is supported by RF President Grant from the “Program for Support of Leading Scientific Schools.” In addition to scientific research, the department educates professional specialists for the entire country under leadership by Krasnoschekov. Under the supervision of RSORS members, dozens of PhD theses were defended and OR science received strong positive impact during the last 13 years.

RSORS members continue active investigation in all OR divisions. They are developing fundamental ideas and approaches proposed by both Russian and foreign scientists, and contribute to further advancement of the contemporary science. The next Moscow International conference on OR will take place in October, 2010 [30]. It will summarize the new achievements in OR development in Russia and other countries.

REFERENCES

1. <http://www.ccas.ru/OR-Conferences> (ISBN print 5-201-09750-2, 5-201-09749-9).
2. J Appl Indus Math. ISSN: 1990–4789 [Original Russian Texts are published in Diskretnyy analiz i issledovanie operaciy. Seriya 2].
3. J Comput Syst Sci Int. ISSN: 1064–2307 [Original Russian Texts are published in Izvestiya Akademii Nauk. Teoriya i Sistemy Upravleniya].
4. Math Model. ISSN: 0234–0879 [Matematicheskoe modelirovanie, in Russian].
5. Comput Math Math Phys. ISSN: 0965–5425 [Original Russian Texts are published in Zhurnal Vychislitel'noy Matematiki i Matematicheskoy Fiziki].
6. Moscow University Computational Mathematics and Cybernetics. ISSN: 0278–6419 [Original Russian Texts are published in Vestnik Moskovskogo Universiteta. Seriya 15. Vychislitel'naya Matematika i Kibernetika].
7. <http://www.ccas.ru>
8. <http://math.nsc.ru/conference/scor98/scorrus.html>; <http://math.nsc.ru/conference/door07/>
9. Germeyer YB. Introduction to the theory of operations research. Moscow: Nauka; 1971 (in Russian).

10. Germeier YB. Non-antagonistic games. Volume 46, Series: theory and decision library. Hingham (MA): Reidel Publications Company; 1986.
11. Moiseev NN. Man, nature and the future of civilization: "nuclear winter" and the problem of a "permissible threshold". Moscow: Novosti Press Agency Publications House; 1986.
12. Krasnoshchekov PS, Fedorov VV, Flerov YA. Elements of mathematical decision making theory. *Avtom Proektir* 1997;1:15–23 (in Russian).
13. Krasnoshchekov PS, Petrov AA. Principles of model design. Moscow: Fazis; 2000 (in Russian).
14. Kreines MG. Models and Technologies for the extraction of aggregated knowledge to control processes of the retrieval of non-structured information. *J Comput Syst Sci Int* 2009;48(2):272–281.
15. Petrov AA. On the foundations of mathematical modeling of economy. *Discrete Dyn Nat Soc* 2001;6(4):313–316.
16. Petrov AA, Pospelov IG, Shananin AA. From Gosplan to ineffective market: mathematical analysis of development of Russian economic structures. Lewiston, Queenston, Lampeter: The Edwin Mellen Press; 1999.
17. Demberel S, Olenov N, Pospelov I. An interaction model for livestock farming and steppe ecosystem. *Math Comput Simul* 2004;67(4–5):335–342.
18. Malashenko YE, Novikova NM. Uncertainty models in many-user systems. Moscow: Editorial URSS; 1999 (in Russian).
19. Davidson MR, Dogadushkina YV, Kreines EM, *et al.* Mathematical model of power system management in conditions of a competitive wholesale electric power (capacity) market in Russia. *J Comput Syst Sci Int* 2009;48(2):243–253.
20. Lotov A, Bushenkov V, Kamenev G. Interactive decision maps. Approximation and visualization of the Pareto frontier. Boston (MA): Kluwer Academic Publishers; 2004.
21. Novikova NM, Pospelova II. Multicriteria decision making under uncertainty. *Math Program* 2002;92(3):537–554.
22. Driessen TS, Van Der Laan G, Vasilev VA, editors. Russian contributions to game theory and equilibrium theory. New York: Springer; 2006.
23. Krasnoshchekov PS. What Bill Gates held back. *Herald Russian Acad Sci* 1998;68(6):440–444.
24. Vasin A. On stability of mixed equilibria. *Nonlinear Anal* 1999;38:793–802.
25. Krasnoshchekov PS, Morozov VV, Popov NM, *et al.* Hierarchical design schemes and decomposition numerical methods. *J Comput Syst Sci* 2001;40(5):754–763.
26. Vasin A, Navidi K. Optimal tax inspection strategy. *Comput Math Model* 2003;4:160–172.
27. Vasin A, Morozov V. Investment decision under uncertainty and evaluation of American options. *Int J Math Game Theory Algebra* 2006;15(3):323–336.
28. Vasin A, Stepanov D. Endogenous formation of political parties. *Math Comput Model* 2008;48(9–10):1519–1526.
29. Izmailov AF, Solodov MV. Numerical methods of optimization. Moscow: Fizmatlit/Nauka; 2008 (in Russian; also in Portuguese: Optimization. 2005, 1. 2007, 2. IMPA, Rio de Janeiro, Brazil).
30. <http://io.cs.msu.su>

SADDLEPOINTS AND VON NEUMANN MINIMAX THEOREM

BEHZAD HEZARKHANI
Faculty of Business
Administration, Memorial
University, St. John's,
Newfoundland, Canada

Two-person zero-sum (TPZS) games are win-lose games, whatever is won by one player is lost by the other player. In this respect, TPZS games are completely competitive games. That is, players cannot hope to improve their situation by any cooperation. Instances of such games are pervasive in a variety of environments: from games played for fun; to business, political, and military decision making environments. In this vein, analysis of TPZS games can provide useful insights into the actual circumstances.

TPZS games can be abstracted into simple models. This makes their analysis a reasonable starting point in studying game theory. Coincidentally, game theory as it is known today has been founded after the exposition of revealing characteristics of TPZS games in 1926 by John von Neumann [1] and the seminal book in 1946 by John von Neumann and Oskar Morgenstern [2].

This section introduces the basic properties of TPZS games and presents their general solution concepts. The structure of this section is as follows. Some preliminary issues about TPZS games, including main assumptions and simplifications, have been addressed in the section titled "General Characteristics of TPZS Games," which will set the ground for the upcoming analysis. The section titled "Minimax and Maximin Strategies as the Security Strategies" introduces the maximin and minimax strategies as the security strategies for players. In the section titled "Strictly Determined Games and Saddlepoints" the strictly determined TPZS games have been discussed. The section titled "Nonstrictly Determined Games and Mixed Strategies" addresses the nonstrictly

determined TPZS games and the concept of mixed strategies along with the von Neumann's minimax theorem that provides a general solution concept for TPZS games.

GENERAL CHARACTERISTICS OF TPZS GAMES

A game is the situation when some persons (i.e., players) are faced with decision making problems, while the outcomes to each person are also affected by the decisions made by the others. In two-person games, this interrelation of decisions exists only between two players. If the outcomes of the game to players are exactly opposite in each case, the game is zero-sum. For example, if the outcomes of a game are the dollar amounts that each player gains or loses, then the game is zero-sum if the amount that one player gains is exactly equal to the amount which the other player loses.

In TPZS games, it is assumed that when players are making their decisions, each has a finite number of different moves to choose from. The attitudes of players are also such that they prefer higher gains over lower gains or equivalently, lower losses over higher losses. Moreover, they make their decisions in accordance with their preferences; otherwise, they would not be *rational*.

The players have certain information about the game they are playing. In particular, they are assumed to know the moves available to the opponent in addition to their own available moves. The players are also fully aware of the outcomes of the game in all possible combinations of their joint moves. Additionally, the players know that they are in equal situations in terms of their knowledge about the game. These assumptions are referred to as *complete knowledge*.

The difficulty of a game situation is due to the fact that each player has to choose his or her moves without knowing the other player's

actual choice. Therefore, the question faced by each player is which move to choose. Hence, analysis of the game should elaborate on the players' choices of their moves. That is the concept of *solution* to a game.

The above characteristics are assumed to be present in all TPZS games even though they are not explicitly stated. Let us consider some examples of TPZS games.

Example [Rock–Scissor–Paper]. The game of rock–scissor–paper (R–S–P) is played as follows. Each player should choose a hand posture from three possibilities: (i) fist representing a piece of rock, (ii) flat hand representing a piece of paper, and (iii) fingers held in a “V,” representing a scissor. Then, the hand postures are shown at the same time. The winner is determined according to these rules: Rock breaks scissor, scissor cuts paper, and paper wraps rock. For example, if a player chooses rock while the other chooses scissor, rock wins, and so on. If both choices are the same, it is a draw. The winner gets one dollar from the loser.

Example [Marketing Strategy in Duopoly]. Two high-tech companies offer wireless phones in a regional market. They used to hold equal shares of the market just before the network provider installed more advanced network equipment. Consequently, both companies designed new products which took advantage of the current technology. However, they need to decide their marketing strategies. The companies have three options: to advertise (which requires significant amounts of investments), to offer discounts (which require less amounts of investment), or not to invest in marketing. If they take the same strategies, their market shares remain the same. However, if they take different strategies, the company that invests more gains more shares of the market.

We return to these examples in greater detail in forthcoming sections.

Formal Description of the Game

Despite the variety of game situations, certain aspects of TPZS games can be represented in a condensed and standard

manner that lends itself to formal analysis. In TPZS games, the participating persons are referred to as *players* 1 and 2, respectively. For referencing convenience, we refer to player 1 with male pronouns and player 2 with female pronouns throughout the rest of this section.

The sets of possible moves that players can choose from constitute an important characteristic of the game. A convenient way of representing the players' moves is through the concept of *strategy*. In games that require each player to make just one move, a strategy is identical to a singular move. However, in multistage games—where the players are required to make several moves—a strategy represents a sequence of moves, one for each stage, for a player.

The collection of a player's strategies is called the *strategy profile* for that player. We denote the strategy profiles for players 1 and 2 with S and T , respectively. Since the total number of strategies for each player is finite, we denote the strategy profiles as

$$S = \{s_1, \dots, s_n\},$$

$$T = \{t_1, \dots, t_m\},$$

where n is the total number of strategies for player 1 and m is the total number of strategies for player 2. Every individual element of a strategy profile is referred to as a *pure strategy*.

In our R–S–P example, the strategy profiles for both players are simply rock, scissor, and paper hand postures. In the marketing strategy example, the strategy profiles are comprised of three pure strategies: to advertise, to offer discounts, and to do nothing. Note that in general, players can have different strategy profiles.

Next step is to consider the outcomes of the game. The outcomes of the game—based on the players' choices of their pure strategies—is determined through a *payoff function* denoted by π . The payoff function then determines the values of pairs of pure strategies to the players. In general, this payoff is measured in units of *utility*. Therefore, the payoff of a pair of strategies shows the units of utility gained or lost by the players under the choice of those strategies.

In this vein, $\pi_1(s_i, t_j)$ shows the payoff to player 1 if he chooses strategy s_i and, at the same time, player 2 chooses strategy t_j . A game represented by the sets of strategies (strategy profiles) and the corresponding payoff function is a *game in strategic form*.

A game in strategic form can be shown by a $n \times m$ matrix, where the rows represent the pure strategies for player 1 and the columns represent the pure strategies for player 2. Each element of the matrix is a pair showing the respective payoffs to both players. Such a matrix is called the *game matrix*. Table 1 depicts the general game matrix in two-person games.

In zero-sum games, we can further simplify the presentation of game matrix. In TPZS games, by definition, the sum of payoffs to the players corresponding to each matrix element is zero; that is,

$$\pi_1(s_i, t_j) + \pi_2(s_i, t_j) = 0$$

for every s_i and t_j in S and T , respectively. Therefore, in TPZS games, for all pairs of strategies, the payoff to player 2 is the opposite of player 1's payoff; that is, for every s_i and t_j in S and T .

$$\pi_2(s_i, t_j) = -\pi_1(s_i, t_j).$$

Thus, having either player's payoffs, we can immediately infer those of the other player. Consequently, we conform to the following convention in presenting the TPZS game matrices: we exclude the player 2's payoffs from the game matrix and only depict those of player 1. We also skip the superscript 1 from the notation of payoff function. The corresponding game matrix is given in Table 2.

Table 1. General Game Matrix in Two-Person Games

	t_1	$\dots$	t_m
s_1	$\pi_1(s_1, t_1), \pi_2(s_1, t_1)$	$\dots$	$\pi_1(s_1, t_m), \pi_2(s_1, t_m)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_n	$\pi_1(s_n, t_1), \pi_2(s_n, t_1)$	$\dots$	$\pi_1(s_n, t_m), \pi_2(s_n, t_m)$

Table 2. The Game Matrix in TPZS Games

	t_1	$\dots$	t_m
s_1	$\pi(s_1, t_1)$	$\dots$	$\pi(s_1, t_m)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_n	$\pi(s_n, t_1)$	$\dots$	$\pi(s_n, t_m)$

Note that the matrix elements in Table 2 represent the payoffs to player 1. Accordingly, the game matrix in TPZS games can be depicted by the $n \times m$ matrix Π , whose element in row i and column j shows the payoff to player 1 if he chooses his i th strategy and player 2 chooses her j th strategy. The game matrix for our marketing strategy example is presented in Table 3.

A straightforward generalization of TPZS games is two-person constant-sum (TPCS) games. In TPCS games, the sums of players' payoffs under all strategy pairs are constant. In fact, every constant-sum game can be reduced to a zero-sum game. Consider a K -sum TPCS game. By definition, for every i and j , we have

$$\pi_1(s_i, t_j) + \pi_2(s_i, t_j) = K.$$

Then, using the transformation $\tilde{\pi}_2(s_i, t_j) = \pi_2(s_i, t_j) - K$, we can construct an equivalent zero-sum game (since $\pi_1(s_i, t_j) + \tilde{\pi}_2(s_i, t_j) = 0$). The validity of this transformation is due to that fact that linear transformation of players' payoff functions does not alter the structure of their preferences.

MINIMAX AND MAXIMIN STRATEGIES AS THE SECURITY STRATEGIES

Consider the game matrix for our marketing strategy example in Table 3. Obviously, player 1 obtains the highest payoff when he

Table 3. The Game Matrix for Marketing Strategy Example

	Advertise	Discount	Do Nothing
Advertise	0	5	10
Discount	-5	0	5
Do Nothing	-10	-5	0

chooses to advertise and his opponent does nothing. This would result in 10 additional market shares for player 1. But should player 1 choose his first strategy hoping to achieve this payoff? Well, not necessarily. This is because player 2 is also taking her situation into consideration. Nevertheless, the strategy pair that gives player 1 the maximum gain costs player 2 the maximum loss of 10 less market shares. How should players select their strategies then?

Since the players cannot simply aim for their best payoffs, one appropriate alternative for them is to select the strategies, which guarantee a minimum amount of gain, or equivalently a maximum amount of loss. These strategies are called *security strategies*.

In the above example, player 1 can verify that if he chooses to advertise, the minimum payoff he could get is to acquire no additional market shares: that happens if player 2 also advertises. So, not knowing the choice of player 2, the worst payoff to player 1 under his first strategy is 0. With the same logic, the minimum payoff to player 1 if he chooses to offer discounts is losing five market shares to player 2, and if he chooses to do nothing, the minimum payoff is losing 10 market shares to player 2. Then, by choosing his first strategy, player 1 could maximize his minimum payoff regardless of player 2's strategy ($\max\{0, -5, -10\} = 0$). Therefore, the security strategy for player 1 is to advertise. The payoff to player 1 in this case is called the *value of the game to player 1*. The latter can also be interpreted as the "gain-floor" for player 1 and he must not win less than that.

To depict this line of reasoning, we define v_1 as the value of the game to player 1 such that

$$v_1 = \max_{s_i} \min_{t_j} \pi(s_i, t_j).$$

In other words, v_1 is the payoff that player 1 gets by choosing the strategy s_i that maximizes the minimum payoff to him under the choice of that strategy.

Considering the game from player 2's view point, we can define her security strategy in a similar way. However, note that the values in the game matrix are, in fact, player 2's antipayoffs or simply losses. Therefore, security strategy for player 2 is the strategy

that minimizes her maximum loss, that is, her "loss-ceiling" (she must not lose more than that). We define v_2 as the value of the game to player 2 such that

$$v_2 = \min_{t_j} \max_{s_i} \pi(s_i, t_j).$$

In the marketing strategy example, the maximum loss for player 2 if she chooses to advertise is zero. The same values for player 2 under her second and third strategies are 5 and 10, respectively. Hence, the security strategy for player 2 is to select her first strategy that results in the loss-ceiling of 0 ($\min\{0, 5, 10\} = 0$).

In TPZS games, the values of the game for the players are related in a particular way. Theorem 1 expresses this relationship.

Theorem 1. *In all TPZS games, $v_1 \leq v_2$. That is*

$$\max_{s_i} \min_{t_j} \pi(s_i, t_j) \leq \min_{t_j} \max_{s_i} \pi(s_i, t_j).$$

This relationship is quite intuitive. It expresses that player 1's secured gain (gain-floor) is always less than or equal to player 2's secured loss (loss-ceiling). The interested reader is referred to p. 219 in Ref. 3 for the proof of this theorem.

STRICTLY DETERMINED GAMES AND SADDLEPOINTS

Consider a class of games for which the players' security strategies result in equal payoffs, that is $v_1 = v_2$. In these games we define v^* as the value of the game such that

$$v^* = v_1 = v_2.$$

The strategy pair corresponding to v^* is called the *saddlepoint*. Thus, if a game has the value, then it has a saddlepoint and *vice versa*. In TPZS games with the value, the choices of security strategies are *optimal* for the players. The optimality of these strategies in these games is due to the fact that none of the players can hope to improve his/her payoff by unilaterally deviating from the choice

of security strategies. In a broader sense, a saddlepoint is also a *Nash equilibrium* [see *Nash Equilibrium (Pure and Mixed)* for more on this concept].

Games with this property are referred to as *strictly determined games*. This is because the players' optimal choices of pure strategies can be determined for these games. In other words, TPZS games with the value are *solvable* in pure strategies. A saddlepoint, then, is *the solution* of a TPZS game in pure strategies.

Once again, consider the marketing strategy example. In the previous section, we calculated the individual security strategies for the players as s_1 and t_1 . The values of the game to both players are zero, that is, $v_1 = v_2 = 0$. Since these values are equal, the game is strictly determined with the value $v^* = 0$, and (s_1, t_1) is the saddlepoint, so both players should advertise. In this situation, if player 1 anticipates that player 2 is going to advertise, the optimal choice for player 1 is still to advertise. Hence, even when the players could foresee the strategy choice of their opponent in terms of security strategies, they cannot improve their payoffs by selecting any strategies other than their security strategies.

It is possible that a strictly determined game has multiple saddlepoints. In this case, more than one pair of strategies corresponds to v^* and the choices among them are equally desirable for both players.

For example, in the game matrix presented in Table 4, there are two saddlepoints (s_2, t_1) and (s_2, t_3) . The corresponding value of game is 1. In this situation, we can determine the game by stating that the player 1 chooses his second strategy while the player 2 selects either her first or third strategy. The outcome, however, is the same in both cases.

Table 4. A Strictly Determined Game with Multiple Saddlepoints

	t_1	t_2	t_3
s_1	$\begin{bmatrix} -1 & 5 & 0 \end{bmatrix}$		
s_2	$\begin{bmatrix} 1 & 3 & 1 \end{bmatrix}$		

NONSTRICTLY DETERMINED GAMES AND MIXED STRATEGIES

In the previous section, we addressed a special class of games with the property of value. However, not all TPZS games are as such. The games without the value—and consequently no saddlepoint—are called *nonstrictly determined games*. In these games, the players' choices of strategies cannot be formulated as clearly as those in the strictly determined games, that is, in terms of pure strategies.

Consider the example of R–S–P game in the section titled “General Characteristics of TPZS Games” and its game matrix in Table 5. In this game, player 1's gain-floor, v_1 , is -1 and player 2's loss-ceiling, v_2 , is 1 . The game does not have a value and consequently, there is no saddlepoint. In terms of security strategies, the choices among the three strategies do not differ for the players. Hence, the game does not have a solution in pure strategies. In other words, the optimal choices of players' pure strategies cannot be specified.

Let us discuss this issue in more detail. The game matrix in Table 6 presents a hypothetical game. Accordingly, the security strategy for player 1 is s_2 with $v_1 = 1(\max\{-1, 1\} = 1)$ and the security strategy for player 2 is t_2 with $v_2 = 3(\min\{5, 3\} = 3)$. This game has no saddlepoint because the values of the game to the players are not equal. Therefore, the choices of security strategies are not optimal in this case. To see this, suppose that player 1 decides to play his security strategy. If player 2 suspects this, she will be better off by playing her first strategy instead of her second (security) strategy. This, on the other hand, gives player 1 a chance to play his first strategy and achieve his best possible outcome. Having speculated this, player 2 can justify playing her second strategy that

Table 5. The Game Matrix for R–S–P Game

	Rock	Scissor	Paper
Rock	$\begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$		
Scissor	$\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$		
Paper	$\begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$		

Table 6. A Game with No Saddlepoint

	t_1	t_2
s_1	5	-1
s_2	1	3

combined with the choice of player 1, leads to her best outcome. It seems that when there is no saddlepoint, players could improve their payoffs by selecting some strategies other than security strategies, but at the same time, this puts them in an inconclusive situation that stems from their iterative speculations. Thus, we need to find a solution concept for the nonstrictly determined games that sheds light on these situations.

Although the solution of nonstrictly determined games cannot be determined in terms of pure strategies, we can still make further inferences about them; nevertheless, these games are still called *determined*. However, in order to develop a solution concept for these games, we need to employ an alternative view of strategy choice. The concept of *mixed strategies* provides such a view. Then, the possible solutions to nonstrictly determined games can be characterized in terms of mixed strategies.

A mixed strategy is essentially a choice of probabilities that a player assigns to the selection of his/her pure strategies. By providing a probabilistic view of strategy choices, the concept of mixed strategies considerably expands the players' strategy choice spaces. When playing a mixed strategy, a player chooses his/her strategies randomly. However, the player can decide the probabilities of these random selections. In order to provide an intuitive understanding of mixed strategies, suppose that the players have devices that can randomly suggest to them which pure strategy to choose. In the R-S-P example, assume that player 1 rolls a die in order to decide which hand posture to show. A mixed strategy for him then is to play rock if the die is either 1 or 2, play scissor in cases of 3 and 4, and play paper in cases of 5 and 6.

We can formally demonstrate a mixed strategy by a vector of probability measures that correspond to all the pure strategies of

a player. In TPZS games, mixed strategies can be shown with two separate vectors. We denote a mixed strategy for player 1 by

$$p = (p_1, \dots, p_n),$$

such that $p_i \geq 0$ for all i , and $\sum_{i=1}^n p_i = 1$. Note that the subscript i corresponds to player 1's pure strategies. A mixed strategy for player 2 is

$$q = (q_1, \dots, q_m),$$

such that $q_j \geq 0$ for all j , and $\sum_{j=1}^m q_j = 1$. Again, j indexes player 2's pure strategies. The k th element of either of these vectors depicts the probability of playing the k th pure strategy by the respective player. The additional conditions make sure that p and q are probability vectors. If all the elements of a mixed strategy are positive, then it is a *totally mixed strategy*. Note that pure strategies can now be defined as special forms of mixed strategies, where only one vector element equals to one and all the others are zero.

Returning to the R-S-P example, a mixed strategy $p = (3/4, 0, 1/4)$ implies that player 1 plays his first and third pure strategies with respective probabilities of 3/4 and 1/4, while he does not play his second pure strategy at all. The totally mixed strategy $q = (1/3, 1/3, 1/3)$ implies that player 2 will play her pure strategies with equal probabilities of 1/3.

While the payoffs with pairs of pure strategies are given in the game matrix, at this point we need to devise a measure for quantifying the payoffs with pairs of mixed strategies. Since the mixed strategies are expressed in probabilistic form, a straightforward such measure in this setting is the *expected* payoff. The expected payoff with a pair of mixed strategies, $J(p, q)$, is defined as

$$J(p, q) = \sum_{i=1}^n \sum_{j=1}^m p_i q_j \pi(s_i, t_j).$$

In this equation, $p_i q_j$ is the joint probability that pure strategies i and j are selected by players 1 and 2, respectively. The summation over the players' probability vectors

provides the expected payoff. Note that this expected payoff is in fact to player 1 (since $\pi(s_i, t_j)$ is his), and the expected payoff to the second player is the opposite, that is, $-J(p, q)$. We can also demonstrate the latter in matrix form:

$$J(p, q) = p \times \prod \times q^T.$$

The superscript T denotes the matrix transpose operation. If we assume that both the players are indifferent between a sure payoff of S and a risky payoff with the equal expected payoff (i.e., both are risk-neutral), we can use the expected payoff function as the criterion for measuring the efficiency of players' choices of mixed strategies.

Similar to the section titled "Strictly Determined Games and Saddlepoints," we can construct the security strategies for the players, but this time in mixed form. Player 1 can reason that if player 2 discovers his choice of mixed strategy p once he chooses it, she selects her mixed strategy q so that $J(p, q)$ is minimized (so that she would maximize her expected payoff). In other words, given the mixed strategy of player 1, the worst expected payoff that he could obtain is when player 2 plays $\min_q J(p, q)$. Therefore, player 1's secure strategy is to choose his mixed strategy in a way that the latter is maximized. Thus, the security strategy for player 1 in mixed strategies is

$$\dot{v}_1 = \max_p \min_q J(p, q).$$

The reasoning process for player 2 is similar. She chooses her mixed strategy so that she could minimize losses that result from the first player's payoff maximization attitude. The security strategy for player 2 in mixed strategies is

$$\dot{v}_2 = \min_p \max_q J(p, q).$$

The dots over $\dot{v}_1$ and $\dot{v}_2$ differentiate the values of the game to the players with mixed strategies from those with pure strategies. The general relationship between the $\dot{v}_1$ and $\dot{v}_2$, that is, $\dot{v}_1 \leq \dot{v}_2$, still holds.

If $\dot{v}_1 = \dot{v}_2$, then the respective mixed strategies for the players are optimum in the

sense similar to that of the pure strategies case: none of the players could improve his/her expected payoffs by deviating from the mixed security strategies. One might suspect that similar to the case with pure strategies, $\dot{v}_1$ and $\dot{v}_2$ are not necessarily equal. In fact, Luitzen Brouwer—the great mathematician whose fixed point theorem set the ground for development of game theory—thought so. However, John von Neumann in his 1926 paper proved otherwise.

Theorem 2 [Von Neumann's Minimax Theorem]. *Given a TPZS game matrix, $\prod$, there always exists a pair of mixed strategies, p and q for players 1 and 2 respectively, such that $\dot{v}_1 = \dot{v}_2$.*

The von Neumann minimax theorem provides a very strong statement for the TPZS games; it shows that all TPZS games have optimal solutions in terms of mixed strategies. The proof of this theorem requires advanced mathematical knowledge and the interested reader may refer to p. 15 in Ref. 4. This theorem is also called *Von Neumann's maximin* theorem.

In the R-S-P game, the optimal mixed strategies for both players are to play all their pure strategies with equal probabilities of 1/3. This would result in the expected payoffs of 0 for both players. In general, the pair of optimal mixed strategies might not be unique. Then again, the choice among them can be made arbitrary because, by definition, they all have the same expected payoffs. The upcoming sections provide comprehensive methods for finding the actual solutions to TPZS games.

REFERENCES

1. Von Neumann J. Zur theorie der gesellschaftsspiele. Math Ann 1928;100(1):295–320.
2. Von Neumann J, Morgenstern O. Theory of games and economic behavior. New Jersey: Princeton University Press; 1944.
3. Binmore KG. Playing for real: a text on game theory. New York: Oxford University Press; 2007.
4. Owen G. Game theory. San Diego: Academic Press; 1995.

SALES OPTIMIZATION MODELS—SALES FORCE TERRITORY PLANNING

SÖNKE ALBERS
Department of Innovation and
Marketing,
Christian-Albrechts
University at Kiel, Kiel,
Germany

MURALI K. MANTRALA
Department of Marketing,
University of
Missouri-Columbia,
Columbia, Missouri

INTRODUCTION

Personal selling is an important marketing instrument in many industries. Zoltners *et al.* [1] report that the US economy has been estimated to spend \$800 billion on sales forces, almost three times the amount spent on advertising (\$276 billion) in 2006. Personal selling is the most effective way, especially in business markets, to learn about and assess a customer's needs, inform customers of standard and/or customized solutions, detail and demonstrate complex high-value products, handle objections, close sales and provide long-term continuing service (see also **Customer Relationship Management: Maximizing Customer Lifetime Value** in the section titled "Applications and History"). However, the high impact of sales forces on firms' sales comes at a high cost. Dartnell's *30th Sales Force Compensation Survey: 1998–1999* reports that the average company spends 10% and some companies spend as much as 40% of their total sales revenues on sales force costs.

A recent meta-analysis of previous empirical studies of personal selling sales response relationships by Albers *et al.* [2] indicates that the mean personal selling sales elasticity (corrected for methodological biases) is about 0.31, significantly larger than the

mean advertising sales elasticity of about 0.06–0.22. Thus, changes in the deployment of personal selling efforts can have large impacts on companies' topline as well as bottom lines.

OVERVIEW OF SELLING EFFORT DECISIONS

Figure 1 displays how sales resource allocation and territory design are related to other problems of sales management. Size, structure, and compensation are strategic decisions that are usually set by top management for several years. They guide lower-level, shorter-term operational planning by senior and middle management with respect to *sales territory design* and *compensation plan design*. These decisions will impact tactical selling effort allocation optimization made by front-line management (e.g., district managers) as well as individual sales representatives. Conceptually, these disaggregate decisions will be guided by local managers' and/or sales representatives' knowledge of sales response functions, that is the relationships between selling effort and sales generated from the relevant sales entities figuring in the decision; for example, market segments or individual customers/accounts. Consequently, 'response function' is shown as the root of Figure 1 because forming assessments of these relationships is a prerequisite for optimizing any of the higher-level decisions.

Figure 1 also indicates that all the decision areas are, in principle, interdependent. In particular, the selected sales force size and structure have implications for the sales territory design. Further, sales territory design impacts the actual allocation of selling effort by an individual representative. Thus, optimization of decisions at each of these levels depends on the assumptions made with respect to lower-level call allocation decision rules as well as higher-level policy constraints. These interdependencies considerably complicate the development of models for globally optimizing sales force

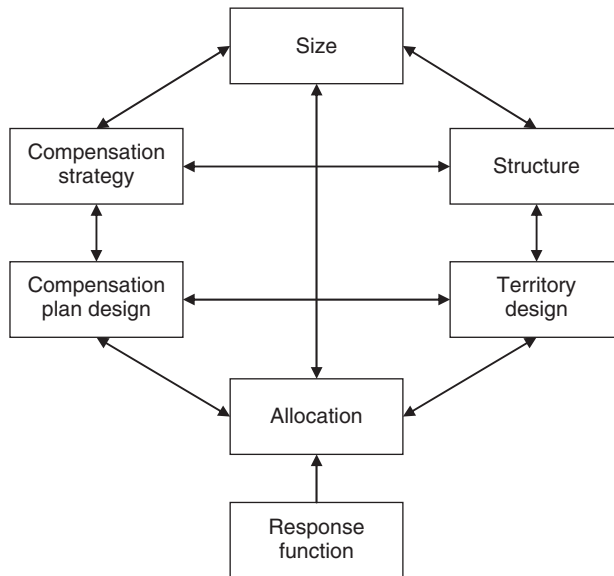


Figure 1. Selling effort decisions. [Source: Ref. 3, “Models for Sales Management Decisions,” in Berend Wierenga (ed.): *Handbook of Marketing Decision Models*, Springer Science + Business Media, International Series in Operational Research and Management Science, p. 166].

decisions. Not surprisingly, therefore, sales force optimization research has progressed by focusing on compartmentalized models. Examples of such models are Lodish’s [4] CALLPLAN for sales calls allocation across customers (described in more detail later in this article) assuming sales force size, structure, and territory design are fixed; and models for sales territory design that assume a fixed sales force size and structure such as those of Lodish [5], Zoltners and Sinha [6], and Skiera and Albers [7].

In the next section we provide a brief overview of research with respect to specification and estimation of sales response functions that lie at the heart of all sales force optimization models.

ESTIMATION OF SALES RESPONSE FUNCTIONS

Decision Calculus Approaches

In the early years of sales force decision modeling, the lack of appropriate objective data constrained the calibration of sales response functions. Therefore, researchers worked with the information that was available, namely subjective or judgmental data. This *decision calculus* approach rested on the logic that managers and salespeople were

already making intuitive decisions based on subjective estimates. The best known example of the decision calculus approach is CALLPLAN proposed by Lodish [4]. Lodish and his colleagues validated CALLPLAN in a field study at United Airlines [8], and reported successful applications at firms such as the Syntex Laboratories [9].

Econometric Approaches

From the late 1980s, the econometric estimation of response functions based on historical data has gained ground primarily because of the increasing availability of transactional panel data with large numbers of observations at the disaggregate or individual customer level. The bulk of the published econometric applications come from the pharmaceutical industry. Aside from data availability, another reason for the ascendancy of the econometric approach is the development of powerful econometric estimation methods and software that can handle many thousands of observations and come up with idiosyncratic estimates per response or sales coverage unit (SCU).

Functional Form

The question of the appropriate sales response functional form remains unresolved.

Should the sales response function be S-shaped as used in early decision calculus models like CALLPLAN or just concave in form?

A popular S-shaped functional form for any sales entity or control unit i in a set of I units as expressed below, was used in Lodish's [4] CALLPLAN. In Equation (1), S_i represents sales of i th entity, h_i the number of calls, and the four parameters, α_i , β_i , γ_i , and δ_i are each > 0 .

$$S_i = \alpha_i + (\beta_i - \alpha_i) \frac{h_i}{\gamma_i + h_i^{\delta_i}} \quad (i \in I). \quad (1)$$

Researchers do agree that sales response functions used in effort allocation models must eventually exhibit diminishing returns to increasing effort. Inherently nonlinear functional forms posed a challenge for econometric estimation in the early years. Thus, until about 15 years ago, econometric applications were limited to the estimation of the parameters of nonlinear aggregate sales response functions such as the multiplicative (or Cobb-Douglas) form which can be linearized by taking logarithms of both sides of the function with the coefficient representing the selling effort elasticity. An alternative to this "double-log" form consists of relating sales to the logarithm of the independent variable, that is the semilogarithmic function. However, the double-log and semilog functions have different implications for optimal allocations. In the former case, the selling effort should be allocated proportional to its elasticity while effort allocation should be proportional to the regression weight in the case of the semilogarithmic function [10].

For response function estimation at the individual customer level however, the output variable data such as physician-level number of prescriptions, is frequently neither continuous nor normally distributed, but is rather Poisson-distributed count data. Therefore, econometricians have modeled prescriptions as a probability function with selling effort (or detailing) as a covariate. To obtain a function with diminishing returns for which an optimal detailing level can be determined, Manchanda and Chintagunta

[11] modeled the Poisson distribution parameter λ (the mean prescription rate, Eq. 2) as an exponential function of a quadratic function of selling effort. Subsequently, Dong *et al.* [12] modeled the impact of selling effort x_{it} with parameters β_{0i} and β_{1i} as shown in Equation (3) in such a way that either a concave or convex or S-shaped relationship may evolve:

$$P_i(y_{it}|\lambda_{it}) = \frac{\lambda_{it}^{y_{it}} \cdot \exp(-\lambda_{it})}{y_{it}!} \quad (i \in I) \quad (2)$$

$$\lambda_{it} = \exp\left(\beta_{0i} + \frac{\beta_{1i}}{1 + x_{it}}\right) \quad (i \in I). \quad (3)$$

Carryover Effects of Selling Effort

With regard to the duration of the effect of personal selling, there is agreement that sales effort has substantial carryover effects [13]. These dynamic effects can be taken into account by either lagged dependent variables, leading to the well-known Koyck model for the case of geometrically decaying memory effects, or by lagged independent or "stock" variables where the effort is cumulated over time but with a certain depreciation rate per period [14].

Dealing with Heterogeneity

Econometric estimation of sales response models calls for fitting a model to historical observations from several time intervals and/or response units. The simplest model assumes that parameter values are the same across cross-sectional units; this is often unrealistic. A simple approach of realizing heterogeneous parameter values for the sales response function is to directly incorporate sources of observed heterogeneity into the model. A more sophisticated approach relies on the estimation of distributions for the response parameter values. This is especially appropriate if longitudinal customer panel data are available, providing a sufficient number of observations for each customer. One could estimate parameter values that follow certain discrete or continuous mixed distributions across the sample of observations. Finite mixture (or latent class) models

allow for the estimation of differing response functions across latent classes or segments [15]. Another option for capturing heterogeneity in parameter values is to derive distributions of parameter estimates applying either a random coefficients approach [16] or a hierarchical Bayes method [11]. Both methods provide posterior estimates per customer.

Dealing with Endogeneity

An unbiased estimation of the parameter values is only possible when the values of the independent variables are randomly chosen. However, the chosen levels of selling effort are often based on the subjective knowledge of managers regarding sales response and are thus nonrandom. This leads to the problem of *endogeneity* of independent variables. Structural approaches to estimation have been proposed to deal with this problem. For example, Manchanda *et al.* [17] propose that the parameter values be estimated with the help of a maximum likelihood approach fitting sales depending on selling effort, as well as selling effort depending on some prior knowledge of the estimated parameter values. However, such a structural approach relies on strong assumptions and it is not clear how meaningful the resulting estimates are when salespeople have not allocated their efforts as assumed. In addition, this approach offers no room to come up with recommendations for improving the allocation of effort which often is the main reason for undertaking the response function estimation in the first place (18, p. 606).

MODELS FOR SALES RESOURCE ALLOCATION AND SALES TERRITORY DESIGN

Sales Resource Allocation Models

Zoltners and Sinha [19] review models that handle basic single-period, single-resource, deterministic resource allocation problem assuming separable sales entities. Only a few models deal with more complicated but realistic problems such as *nonseparable* or interacting sales entities [4,20,21], and

multiple resource-targets; for example, a firm's joint allocation of a team of selling and service reps across its key accounts [22].

A good example for a sales resource allocation model is the influential CALLPLAN by Lodish [4]. This model addressed the question of how a salesperson's time should be allocated between customers and prospects in his/her territory while considering the time required to travel to different subareas of the territory. As shown below, CALLPLAN maximizes the sum of contributions from all customers less the associated travel costs.

Objective function: maximization of profit contribution

$$\sum_{i \in I} \underbrace{g_i \cdot f_i(h_i)}_{\text{profit contribution from sales depending on the calling time}} - \sum_{j \in J} \underbrace{k_j^R \times y_j}_{\text{travel cost multiplied by the number of tours into the subareas}} \Rightarrow \text{Max!} \quad (4)$$

Available working time is constrained to

$$\sum_{i \in I} t_i \cdot h_i + \sum_{j \in J} I_j \cdot y_j \leq T. \quad (5)$$

Number of tours to subarea j must be higher or equal to the maximum number of calls to customer i :

$$h_i \leq y_j \quad (i \in IS_j, j \in J). \quad (6)$$

Lower and upper bound for number of calls:

$$l_j \leq y_j \leq u_j \quad (j \in J). \quad (7)$$

Nonnegativity of number of tours to subarea j :

$$y_j \geq 0 \quad (j \in J). \quad (8)$$

In the above model, the objective (Eq. 4) is to find the optimal allocation of calls h_i across the set of $i \in I$ accounts or account groups that maximizes the sum of profit contributions from calls to accounts (sales $f_i(h_i)$ multiplied with gross margin g_i) less the

total travel costs (number of tours y_j multiplied with average travel cost k_j^R of a tour to subarea $j \in J$). The total time constraint (Eq. 5) ensures that salesperson's time T is not exceeded by the allocated time necessary for calls (h_i weighted with duration of a call t_i summed up over all accounts $i \in I$) and tours to subareas (number of tours y_j multiplied with average duration t_j of a tour to subarea $j \in J$). The constraint with reference to number of trips (Eq. 6) ensures that calls h_i to a particular account cannot exceed the number of trips y_j made to that account's geographic region. Constraint (7) specifies upper and lower bounds on the allocated calls h_i to each account.

As the optimization involving S-shaped response functions is a rather difficult combinatorial problem to solve [23], solutions are obtained using integer programming (see **Formulating Good MILP Models** in the section titled "Optimization Models, Techniques, and Algorithms"). Lodish in fact applies a simpler heuristic in which the sales response function is replaced by its linear concave envelope which provides solutions sufficiently close to the true optimum for most realistic problems. The famous field test at United Airlines by Fudge and Lodish [8] indicated that salespeople using CALLPLAN achieved a significantly higher level of sales increases than the control group. Zoltners and Sinha [19] propose a general modeling approach toward a salesperson call planning system which is conceptually similar to CALLPLAN. Zoltners *et al.* [21] propose an integer programming solution procedure. However, nowadays powerful nonlinear optimization algorithms as implemented in Microsoft's Excel Solver [24] can be more readily used [25,26].

Lastly there have been developments which alter the CALLPLAN model's complicating travel time constraints (Eqs 5 and 6). Specifically, in their COSTA model, Skiera and Albers [7] are able to replace these by a single total selling time constraint, where selling time equals calling time plus travel time, as they directly incorporate selling time into the SCU-level response functions used in their model. This leads to a much simpler optimization structure that can

be solved with standard software such as Microsoft's Excel Solver.

Sales Territory Design

Sales territory design deals with the problem of assigning sets of customers to salespersons. While there are some industries where field sales agents can call on any customer, for example insurance salespeople, most industries work with geographically specialized salesforces, that is salespeople assigned to geographic sales territories. Being assigned his/her own sales territory makes a salesperson feel more responsible and may result in better morale and performance [27]. Geographic sales territories are typically formed by combining SCUs like zip codes or political areas (counties, states, provinces). Then, the sales territory design problem is to determine the combination of SCUs that form each territory.

Hess and Samuels [28] proposed a model-based approach "GEOLINE" to develop *balanced territories*; that is, determine sets of contiguous SCUs which are equalized in terms of workload by minimizing a measure of compactness. "Balancing" is an often sought-after goal to make all salespeople feel they are being treated equitably. Hess and Samuels determined the optimal solution with the help of linear programming (LP) (see **An Introduction to Linear Programming**). However, as an LP's solution is noninteger, and SCUs can only be assigned in their entirety to a territory, GEOLINE works with a rounding procedure that often leads to significant imbalances of workload across territories (of up to 15%). Therefore, Shanker *et al.* [29] as well as Zoltners [30] presented integer programming approaches for the territory alignment task. Lodish [5], on the other hand, focused on a procedure for equalizing the marginal profit contribution across territories. Subsequently, Skiera and Albers [7] showed this is suboptimal as the optimal territories can have different marginal profit contribution values and only have to be equal for SCUs within territories. In another development, Zoltners and Sinha [6] provide a generalization of the earlier models by allowing for any objective function

and achieving almost equality for potential or other balancing criteria through the introduction of lower and upper bounds. However, there remains the question of how to specify the interval between lower and upper bounds. If one tries to minimize travel time then the scarce resource of time becomes more and more irrelevant until there is only one solution satisfying all constraints. However if one relaxes the constraints, then travel time considerations become increasingly important. Consequently, this model would be more helpful to users if it were set up as a multiobjective criteria problem with explicit weights for all criteria.

The model by Zoltners and Sinha [6] also offered some other important enhancements from the viewpoint of implementation. First, territories should not encompass natural boundaries like mountain ranges and rivers that make it nearly impossible to get from one part of the territory to another. This is taken into account by working with actual travel times via a network of major roads. Second, territories should be contiguous. This is ensured by the additional constraint that all SCUs had at least one road connecting to another SCU over which the base location of the salesperson could be reached.

The problem with balancing approaches is that there is only an indirect relationship between profit and territory design. Zoltners and Sinha [27] claim that equal potential increases the utilization of the salesperson's selling time. However, why not try to directly maximize profit contribution when aligning sales territories? This question is addressed by the model COSTA [7] which utilizes a new concept for incorporating travel time effects directly into the sales response function per SCU, depending on the assignment given to a certain salesperson (i.e., territory). This formulation of the sales response function allows for a simultaneous solution of the call planning problem and the assignment of SCUs to territories to maximize overall profit.

Zoltners and Sinha [27] argue that the COSTA approach is too complex and managers implement modified solutions anyway so that the optimality from a profit viewpoint is no longer guaranteed. On the

other hand, substantial progress has been made in reliably estimating sales response functions. And even if the response functions are incorrectly specified, Skiers and Albers [7] show that solutions of COSTA are superior on profit-maximizing grounds to those derived from balancing approaches.

Sales Force Size Decision-Focused Models

In practice, companies often use heuristic or rule-based approaches to setting the size of their sales force, that is the number of salespeople they need over a planning horizon, such as "*same as last year*" or "*pay as you go*" or "*earn-your-way*" strategies that add salespeople as sales grow. Most commonly, sales force sizes are based on a top-down "breakdown" approach under which the total sales force budget (or "costs") is set as a "cost-of-sales" percentage of a sales forecast. This dollar budget divided by the average salesperson salary leads to the selected sales force size. Alternatively, a more data-intensive approach to sales force sizing used by some sales managers is the "bottom-up" *workload buildup* or *activity-based method* [31]. Essentially, these approaches classify customers into segments, set norms for the numbers of sales calls to be made, and fix the number of accounts to be covered in each segment (i.e., the reach). The total workload (sales calls or hours) is then computed and divided by the average rep's call capacity to determine the required sales force size.

The above approaches are likely to produce non-profit-maximizing sales force sizes because first, they do not really focus on profits, and second, they do not fully assess the underlying market sales response functions and resource allocation decisions [32,33].

Recognizing these limitations, a number of normative models for sales force sizing based on sales response functions and marginal analysis have been proposed. More specifically, Dorfman and Steiner's [34] classic theorem establishes that the budgets for marketing instruments can be stated as percentages of profit contribution before marketing costs that equal the respective elasticities. Albers [25] notes that either of the following ratios (Eqs 9 and 10)

provide intuitive insights for managers when concerned with budgeting:

$$\frac{\text{Optimal sales force size (budget)}}{\text{Profit contribution before sales force cost}} = \text{Sales force effort elasticity} \quad (9)$$

$$\frac{\text{Optimal sales force size (budget)}}{\text{Revenue}} = \frac{\text{Sales force effort elasticity}}{\text{Price elasticity}} \quad (10)$$

Integrated Allocation and Territory Design Models

Several models integrate allocation and territory design solutions. One of the earliest was developed by Lodish [5] who extended the CALLPLAN model to incorporate the territory design issue, assuming a fixed sales force size. He maximized profits by equalizing the marginal profit of an additional hour of sales effort across territories with the help of a heuristic. First, he solved the problem of allocating selling time across accounts by assuming one superterritory with the total selling time being equal to the sum of selling times available to all salespersons and applying the CALLPLAN heuristic procedure to determine the optimal number of calls for each account. Then, the decision maker had to reassign accounts step by step intuitively, so that the individual selling time constraints were met and the marginal profit of selling time had become as equal as possible. However, as already noted, the equal marginal profits of selling time achieved were only optimal for the allocation of selling time across accounts or SCUs per territory, but not across territories.

Integrated Sizing and Allocation Models

An early example of this approach is provided by the first stages of Beswick and Cravens' [35] multistage model for sales force management which use sales response functions for small geographic subareas (groups of counties). Subsequently, Lodish [36] argued that developing sales response functions for numerous geographic subareas would be a formidable undertaking in practice. He,

therefore, proposed a decision calculus model for sales force sizing that utilized *product-by-market (customer) segment* sales response functions (where "segments" are mutually exclusive, collectively exhaustive subgroups of the total possible set of customer accounts) and simultaneously optimized product-by-market segment allocations and total sales force size. Lodish *et al.* [9] applied a version of this model at Syntex Laboratories. Later, Lodish [37] reported that Syntex credited this model with increasing their profits by over 20%.

Over 2000 successful applications of similar sales force sizing and resource allocation models have also been claimed by Sinha and Zoltners [13]. According to them, typically, the models applied were nonlinear programming models that utilize product-market segment sales effort response functions and maximize 3- to 5-year profitability for alternative sales force sizes, and product and market allocations.

Empirical Insights, Lessons from Applications. A number of the models described in this section have been applied in practice multiple times mostly by Lodish and his collaborators, and Zoltners, Sinha, and their collaborators at the consulting company ZS Associates Inc. Fortunately, both groups have made efforts over the years to document and publish the results of these model applications and associated experiences [13,27,37]. In general, both groups report that choosing the right model is an art, requiring a good balance between model complexity and ease of understanding, and sensitivity to corporate management goals. Zoltners *et al.* [38] report mature companies boosted their margins by 4.5% when they resized their sales forces and allocated resources better. While 29% of those gains came because the companies corrected the size of their sales forces, 71% of the gains were the results of changes in resource utilization. Based on 1500 successful implementations of sales territory alignments, Zoltners and Sinha [27] state that an improved alignment may provide a profit increase of 2–7%.

CONCLUSIONS

Models for sales force optimization and territory management to date have addressed and even quite successfully resolved some of the traditional strategic, operational, and tactical issues facing sales managers. Some refinements and extensions in solutions in these problem areas are needed but the core ideas are in place. However, the selling environment is rapidly evolving and a number of new sales management challenges are surfacing requiring solutions and offering significant opportunities for new optimization models-based research. These include sales force optimization in a multichannel marketing world where the focus is on customer relationship management and integrated marketing communications.

REFERENCES

1. Zoltners AA, Sinha P, Lorimer SE. Sales force effectiveness: a framework for researchers and practitioners. *J Pers Selling Sales Manag* 2008;28(2):115–131.
2. Albers S, Mantrala MK, Sridhar S. A meta-analysis of sales force response elasticities. Report No. 08-100. Marketing Science Institute; 2008 [to appear in *J Market Res*].
3. Albers S, Mantrala MK. Models for sales management decisions. In: Wierenga B, editor. *Handbook of marketing decision models*. Berlin: Springer; 2008. pp. 163–210.
4. Lodish LM. CALLPLAN: an interactive salesman's call planning system. *Manag Sci* 1971;18:P25–P40.
5. Lodish LM. Sales territory alignment to maximize profit. *J Mark Res* 1975;12:30–36.
6. Zoltners AA, Sinha P. Sales territory alignment: a review and model. *Manag Sci* 1983;29:1237–1256.
7. Skiera B, Albers S. COSTA: contribution optimizing sales territory alignment. *Market Sci* 1998;18:196–213.
8. Fudge WK, Lodish LM. Evaluation of the effectiveness of a model based salesman's planning system by field experimentation. *Interfaces* 1977;8(1, Part 2):97–106.
9. Lodish LM, Curtis E, Ness M, Kerry Simpson M. Sales force sizing and deployment using a decision calculus model at syntex laboratories. *Interfaces* 1988;18(1):5–20.
10. Skiera B, Albers S. Prioritizing sales force decision areas for productivity improvements using a core sales response function. *J Pers Selling Sales Manag Special Issue on Enhancing Sales Force Productivity* 2008;28(2):145–154.
11. Manchanda P, Chintagunta PK. Responsiveness of physician prescription behavior to salesforce effort: an individual level analysis. *Market Lett* 2004;15(2–3):129–145.
12. Dong X, Manchanda P, Chintagunta PK. Quantifying the benefits of individual level targeting in the presence of firm strategic behavior. *J Market Res* 2009;46(2):207–221.
13. Sinha P, Zoltners AA. Sales-force decision models: insights from 25 years of implementation. *Interfaces* 2001;31(3):S8–S44.
14. Hanssens DM, Parsons LJ, Schultz RL. *Market response models: econometric and time series analysis*. 2nd ed. Boston (MA): Kluwer Academic Publishers; 2003.
15. Wedel M, Kamakura WA. *Market segmentation: conceptual and methodological foundations*. Boston (MA): Kluwer Academic Publishers; 1999.
16. Narayanan S, Desiraju R, Chintagunta PK. Return on investment implications for pharmaceutical promotional expenditures: the role of marketing-mix interactions. *J Market* 2004;68:90–105.
17. Manchanda P, Rossi PE, Chintagunta PK. Response modeling with nonrandom marketing-mix variables. *J Market Res* 2004;41:467–478.
18. Chintagunta PK, Erdem T, Rossi PE, *et al*. Structural modeling in marketing: review and assessment. *Market Sci* 2006;25:604–616.
19. Zoltners AA, Sinha P. Integer programming models for sales resource allocation. *Manag Sci* 1980;26:242–260.
20. Montgomery DB, Silk AJ, Zaragoza CE. A multiple-product sales force allocation model. *Manag Sci* 1971;18:P3–P24.
21. Zoltners AA, Sinha P, Chong PSC. An optimal algorithm for sales representative time management. *Manag Sci* 1979;29:1237–1256.
22. Armstrong GM. The SCHEDULE model and the salesman's effort allocation. *Calif Manag Rev* 1976;18(4):43–51.
23. Freeland JR, Weinberg CB. S-shaped response functions: implications for decision models. *J Oper Res Soc* 1980;31:1001–1007.
24. Fylstra D, Lasdon L, Watson J, *et al*. Design and use of the microsoft excel solver. *Interfaces* 1998;28(5):29–55.

25. Albers S. Impact of types of functional relationships, decisions, and solutions on the applicability of marketing models. *Int J Res Mark* 2000;17:169–175.
26. Lilien G, Rangaswamy A. *Marketing engineering: computer-assisted marketing analysis and planning*. 2nd ed. State College: DecisionPro, Inc.; 2004.
27. Zoltners AA, Sinha P. Sales territory design: thirty years of modeling and implementation. *Market Sci* 2005;24:313–331.
28. Hess SW, Samuels SA. Experiences with a sales districting model: criteria and implementation. *Manag Sci* 1971;18:P41–P54.
29. Shanker RJ, Turner RE, Zoltners AA. Sales territory design: an integrated approach. *Manag Sci* 1975;22:309–320.
30. Zoltners AA. Integer programming models for sales territory alignment to maximize profit. *J Market Res* 1976;13:426–430.
31. Zoltners AA, Sinha P, Lorimer SE. *Sales force design for strategic advantage*. Houndmills: Palgrave Macmillan; 2004.
32. Mantrala MK, Sinha P, Zoltners AA. Impact of resource allocation rules on marketing investment-level decisions and profitability. *J Market Res* 1992;29:162–175.
33. Mantrala MK. Allocating marketing resources. In: Weitz BA, Wensley R, editors. *Handbook of marketing*. London, Thousand Oaks (CA), New Delhi: Sage; 2002. pp. 409–435.
34. Dorfman R, Steiner PO. Optimal advertising and optimal quality. *Am Econ Rev* 1954;44:826–836.
35. Beswick CA, Cravens DW. A multistage decision model for salesforce management. *J Market Res* 1977;14:135–144.
36. Lodish LM. A user oriented model for sales force size, product and market allocation decisions. *J Market* 1980;44:70–78.
37. Lodish LM. Building marketing models that make money. *Interfaces* 2001;31(3):45–55.
38. Zoltners AA, Sinha P, Lorimer SE. Match your sales force structure to your business cycle. *Harv Bus Rev* 2006;84:81–89.

SAMPLING METHODS

ALEXANDER SHAPIRO
School of Industrial and Systems
Engineering, Georgia
Institute of Technology,
Atlanta, Georgia

TWO-STAGE STOCHASTIC PROGRAMMING

Consider the following stochastic programming problem

$$\text{Min}_{x \in \mathcal{X}} \{f(x) := \mathbb{E}[F(x, \xi)]\}. \quad (1)$$

Here $\mathcal{X} \subset \mathbb{R}^n$, ξ is a random vector, whose probability distribution P is supported on set $\Xi \subset \mathbb{R}^d$ and $F: \mathcal{X} \times \Xi \rightarrow \mathbb{R}$. We assume that for every $x \in \mathcal{X}$ the expectation $\mathbb{E}[F(x, \xi)]$, taken with respect to the probability distribution of ξ , is well defined and finite valued. The notation “ $:=$ ” means “equal by definition.”

For example, in a standard formulation of two-stage linear stochastic programming with recourse, Problem (1) has the setting of $\mathcal{X} := \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ and $F(x, \xi)$ being equal to the first-stage cost $c^\top x$ plus the optimal value $Q(x, \xi)$ of the second-stage problem

$$\begin{aligned} \text{Min}_{y \in \mathbb{R}^m} \quad & q^\top y \\ \text{s.t.} \quad & Tx + Wy = h, y \geq 0, \end{aligned} \quad (2)$$

depending on the data $\xi = (q, h, T, W)$. By definition, $Q(x, \xi) = +\infty$ if the constraint system of Problem (2) is infeasible, and $Q(x, \xi) = -\infty$ if the Problem (2) is unbounded from below. Of course, in order for $F(x, \xi)$ to have a finite expectation, it is necessary for $F(x, \xi)$ to be finite valued for almost every (a.e.) $\xi \in \Xi$. If for every $x \in \mathcal{X}$, the optimal value $Q(x, \xi) < +\infty$ (i.e., the constraint system has a nonempty set of solutions) for a.e. realization of the random data, then it is said that the two-stage problem has a *relatively complete recourse*.

The expectation $\mathbb{E}[F(x, \xi)]$ is given by the multidimensional integral $\int_{\Xi} F(x, \xi) dP(\xi)$,

which can be evaluated in a closed form, only in rather specific cases. Therefore, a standard approach is to make a discretization of the probability distribution of ξ . That is, a finite number of realizations $\{\xi_1, \dots, \xi_K\}$, called *scenarios*, with respective probabilities $p_1, \dots, p_K$, is constructed. In the case of finite number of scenarios, it follows that $\mathbb{E}[F(x, \xi)] = \sum_{k=1}^K p_k F(x, \xi_k)$, and the two-stage linear stochastic program with recourse can be written as one large linear programming problem:

$$\begin{aligned} \text{Min}_{x, y_1, \dots, y_K} \quad & c^\top x + \sum_{k=1}^K p_k q_k^\top y_k \\ \text{s.t.} \quad & Ax = b, x \geq 0, \\ & T_k x + W_k y_k = h_k, y_k \geq 0, \\ & k = 1, \dots, K, \end{aligned} \quad (3)$$

by making one copy of the second-stage decision vector for every realization (scenario) of the data.

The point is, however, that even a crude discretization of the corresponding probability distribution typically results in an exponential growth of the number of scenarios. Suppose, for example, that components of the d -dimensional random vector ξ are independent of each other and one uses discretization of each marginal distribution with r points. Then the total number of scenarios is $K = r^d$. That is, while the input data (associated with discretization points) grows linearly in d , the number of scenarios grows exponentially. And, indeed, it is shown in Dyer and Stougie [1] that, under the assumption that the stochastic parameters are independently distributed, two-stage linear stochastic programming problems are #P-hard.

This raises the question whether stochastic programming problems can be solved numerically. A possible approach is to consider a moderate number of scenarios and to view our goal as just to design clever algorithms adjusted to a particular structure of the obtained problem. An obvious question with this approach is: “who is going to give us

these scenarios?” Almost surely, the actual future will be different from the future described by a finite number of scenarios. Therefore, only the first-stage decision vector x could be of practical interest in a solution of the two-stage problem (Eq. 3). Note also that in applications, quite often, it does not make much sense to try to solve the corresponding stochastic problem with a high accuracy, say of order 10^{-3} or 10^{-4} , since the involved inaccuracies resulting from inexact modeling, distribution approximations, and so on, could be far greater.

Since deterministic approximations of stochastic programming problems are computationally intractable, one can think about randomization type approaches. We are going to discuss this approach in the next section.

MONTE CARLO SAMPLING METHODS

Suppose that it is possible to generate a random sample $\xi^1, \dots, \xi^N$ of N realizations of the random vector ξ , say by Monte Carlo sampling techniques. We assume now that the sample is Independent Identically Distributed (i.i.d.). Then the expectation function $f(x)$ can be approximated by the corresponding average and hence the “true” problem (Eq. 1) can be approximated by

$$\min_{x \in \mathcal{X}} \left\{ \hat{f}_N(x) := \frac{1}{N} \sum_{j=1}^N F(x, \xi^j) \right\}. \quad (4)$$

By the Law of Large Numbers, we have that for a given $x \in \mathcal{X}$, the sample average $\hat{f}_N(x)$ converges w.p.1 to its expected value $f(x)$ as $N \rightarrow \infty$. Therefore, it is natural to expect that the optimal value and an optimal solution of Problem (4) will converge w.p.1 to their counterparts of the true problem (Eq. 1). And indeed, this can be proved under mild regularity conditions (for a survey of this type of results, we refer the reader to Ref. 2, Section 5.1.1, and references therein).

The above approach to construct Monte Carlo sampling-based approximations of the true problem is a natural one and was used by many authors in slightly different forms, under various names in different contexts. It is difficult to point to an original

introduction of that approach; in stochastic optimization literature, we can mention, for example, Refs 3 and 4, where this approach was called *stochastic counterpart method* and *sample-path optimization*, respectively. Currently it is often referred to as the *sample average approximation* (SAA) method. Once the random sample is generated, the obtained SAA problem can be considered as a stochastic programming problem with N scenarios $\xi^1, \dots, \xi^N$, each taken with equal probability $1/N$. It should be mentioned that the SAA method is not an algorithm, the constructed problem (Eq. 4) still has to be solved by an appropriate numerical procedure.

Unfortunately, for a fixed $x \in \mathcal{X}$, the convergence of the sample average $\hat{f}_N(x)$ to its expected value $f(x)$ is quite slow. By the central limit theorem, we have that $N^{1/2}[\hat{f}_N(x) - f(x)]$ converges in distribution to normal distribution with zero mean and variance equal to the variance $\sigma_x^2 := \mathbb{V}\text{ar}[F(x, \xi)]$, provided that this variance is finite. That is, $\hat{f}_N(x)$ has approximately a normal distribution with mean $f(x)$ and standard deviation $\sigma_x/\sqrt{N}$, and consequently $\hat{f}_N(x)$ converges to $f(x)$ at a (stochastic) rate of $O_p(N^{-1/2})$. In other words, in order to estimate $f(x)$ by its sample average with an accuracy $\varepsilon > 0$, one needs a sample of size $N = O(\varepsilon^{-2})$. Although various techniques, for example, variance reduction techniques and quasi-Monte Carlo methods were developed in simulation literature in order to enhance the accuracy of such estimates (it is far beyond the scope of this article to give a review of these methods, we may refer the interested reader to Refs 5 and 6), it is basically impossible to evaluate multivariate integrals with a high precision. Therefore, there is a little hope that stochastic programming problems can be solved with a high precision. It could be somewhat surprising that the SAA approach turns out to be reasonably efficient for solving certain classes of two-stage stochastic programming problems. This was verified in numerical experiments and is supported by a theory discussed below.

Let us remark that values of the sample average function $\hat{f}_N(x)$ can be computed in two somewhat different, although equivalent, ways. The generated sample $\xi^1, \dots, \xi^N$ can be stored in the computer memory, and called

every time a new value (at a different point x) of the sample average function should be computed. Alternatively, the same sample can be generated by using a common seed number in an employed pseudorandom numbers generator (this is why this approach is called the *common random number generation method*).

The idea of common random number generation is well known in simulation and gives an intuitive explanation why the SAA method may work reasonably well in practical applications. That is, suppose that we want to compare values of the objective function at two points $x_1, x_2 \in \mathcal{X}$. In that case, we are interested in the difference $f(x_1) - f(x_2)$ rather than in the individual values $f(x_1)$ and $f(x_2)$. We can estimate the expectations $f(x_1), f(x_2)$ by their respective sample averages $\hat{f}_N(x_1), \hat{f}_N(x_2)$ using independent samples, both of size N , or the same sample. In both cases, $\mathbb{E}[\hat{f}_N(x_1) - \hat{f}_N(x_2)] = f(x_1) - f(x_2)$, and

$$\begin{aligned} \text{Var}[\hat{f}_N(x_1) - \hat{f}_N(x_2)] \\ = \text{Var}[\hat{f}_N(x_1)] + \text{Var}[\hat{f}_N(x_2)] \\ - 2 \text{Cov}(\hat{f}_N(x_1), \hat{f}_N(x_2)). \end{aligned} \quad (5)$$

Now, in the case of independent samples, $\hat{f}_N(x_1)$ and $\hat{f}_N(x_2)$ are uncorrelated and hence the covariance term in Equation (5) vanishes. On the other hand, in the case of the same sample, the estimators $\hat{f}_N(x_1)$ and $\hat{f}_N(x_2)$ tend to be positively correlated, especially if x_1 and x_2 are close to each other, in which case the variance in Equation (5) becomes smaller.

The basic question of the SAA method is how large should the sample size N be, in order to solve the “true” problem (Eq. 1) with a given precision $\varepsilon > 0$. If the required sample size N is very large, say comparable with the number of scenarios generated by the above discussed discretization procedure, then there is little point in using the SAA approach. Recall that a (feasible) point $\bar{x} \in \mathcal{X}$ is said to be an ε -optimal solution of Problem (1) if $f(\bar{x}) \leq \vartheta^* + \varepsilon$, where $\vartheta^* := \inf_{x \in \mathcal{X}} f(x)$ is the optimal value of the true problem.

By employing large deviations-type probabilistic bounds, it is possible to give an estimate of the sample size N such that any ε' optimal solution, with $\varepsilon' \in (0, \varepsilon)$, of the SAA problem gives an ε optimal solution of the

true problem (Eq. 1) with a probability of at least $1 - \alpha$, where $\alpha \in (0, 1)$ is a small number (significance level). That is, by solving the SAA problem with precision, say, $\varepsilon' = \varepsilon/2$, we are guaranteed with confidence of say, 99% that we solve the true problem with precision ε . Under certain regularity conditions, the following estimate of the sample size guarantees this property [2,7]; see also Ref. 8, Section 5.3.2 for a discussion of such estimates:

$$N = O(1) \left(\frac{DL}{\varepsilon} \right)^2 \left[n \log \left(\frac{DL}{\varepsilon} \right) + \log \left(\frac{1}{\alpha} \right) \right]. \quad (6)$$

Here $O(1)$ denotes a generic constant independent of the data, n is the dimension of the decision vector x , D is the diameter of the feasible set $\mathcal{X}$ and L is Lipschitz constant of $F(\cdot, \xi)$ on $\mathcal{X}$ for all $\xi \in \Xi$. Of course, the estimate (Eq. 6) makes sense only if constants D and L are finite.

A discussion of the above estimate is now in order. In a sense, the bound (Eq. 6) gives an estimate of complexity of solving the true problem; the larger the sample size N , the more time consuming it is to solve the corresponding SAA problem. In case of linear two-stage programming, the computational effort in solving the SAA problems is more or less proportional to N (i.e., this is an empirical observation). Formula (6) suggests that the sample size N is “almost” proportional to ε^{-2} (the additional dependence on $\log(\varepsilon^{-1})$ is practically constant). Such dependence of the sample size on ε is typical for Monte Carlo-type methods and cannot be significantly changed unless the considered class of problems has a specific structure. The significance level constant α is under the logarithm sign. This means that increasing confidence, say, from 99% to 99.9% does not significantly change the sample size estimate.

The above discussion suggests that the true problem (Eq. 1) can be solved by the SAA method in a reasonable time with a reasonable accuracy provided that the following conditions are satisfied: (i) the required sample size is manageable; (ii) it is possible to solve the constructed SAA problem with a reasonable efficiency. From this point of view, the number of scenarios of the true problem

(i.e., cardinality of the support Ξ of the distribution of ξ) is irrelevant and can be infinite. The condition (ii) holds in the case of two-stage linear stochastic programming with recourse. The condition (i) holds if for every $x \in \mathcal{X}$ variability of $F(x, \xi)$ is “not too large.” In particular, $F(x, \xi)$ should be finite valued for a.e. realization of ξ . In case of two-stage stochastic programming, this means that the recourse should be relatively complete and the second-stage problem should be bounded from below.

It should be mentioned that Formula (6) has a theoretical value and is not useful for practical applications. This is because it typically gives an estimate of the required sample size far bigger than is actually needed. In the next section, we discuss a practical way of evaluating a given solution.

VALIDATION ANALYSIS

Suppose that we are given a feasible point $\bar{x} \in \mathcal{X}$ as a candidate for an optimal solution of the true problem. For example, $\bar{x}$ can be an output of a run of the corresponding SAA problem. In this section, we discuss a way to evaluate the quality of this candidate solution by estimating the optimality gap $f(\bar{x}) - \vartheta^*$ (recall that ϑ^* denotes the optimal value of the true problem). Since $\bar{x} \in \mathcal{X}$ is feasible, we have, of course that $f(\bar{x}) \geq \vartheta^*$. The following statistical procedure was developed by Mak *et al.* [9].

We can estimate $f(\bar{x})$ by the sample average $\hat{f}_{N'}(\bar{x})$, based on an i.i.d. sample of size N' . Note that since the point $\bar{x}$ is fixed, there is no need to solve the corresponding stochastic problem and hence one can use a large sample. Note also that the employed sample should be independent of the sample used in construction of the point $\bar{x}$. Together with $\hat{f}_{N'}(\bar{x})$, one can compute the sample estimate $\hat{\sigma}_{N'}^2(\bar{x})$ of the variance $\text{Var}[\hat{f}_{N'}(\bar{x})]$, and hence to construct the following $100(1 - \alpha)\%$ confidence upper bound

$$U_{N'}(\bar{x}) := \hat{f}_{N'}(\bar{x}) + z_\alpha \hat{\sigma}_{N'}(\bar{x}), \quad (7)$$

for $f(\bar{x})$ [here z_α is the $(1 - \alpha)$ -quantile of the standard normal distribution].

In order to construct a lower bound for ϑ^* , we proceed as follows. Denote by $\hat{\vartheta}_N$, the

optimal value of the SAA problem (Eq. 4). Note that $\hat{\vartheta}_N$ is random, since it is a function of the sample used in construction of the SAA problem. It is not difficult to show that $\vartheta^* \geq \mathbb{E}[\hat{\vartheta}_N]$, that is, $\mathbb{E}[\hat{\vartheta}_N]$ gives a lower bound for the true optimal value ϑ^* . The expectation $\mathbb{E}[\hat{\vartheta}_N]$ can be estimated by averaging. That is, one can solve M times SAA problems based on independently generated samples each of size N . Let $\hat{\vartheta}_N^1, \dots, \hat{\vartheta}_N^M$ be the computed optimal values of these SAA problems. Then,

$$\bar{v}_{N,M} := \frac{1}{M} \sum_{m=1}^M \hat{\vartheta}_N^m \quad (8)$$

is an unbiased estimator of $\mathbb{E}[\hat{\vartheta}_N]$. Since the samples, and hence $\hat{\vartheta}_N^1, \dots, \hat{\vartheta}_N^M$, are independent and have the same distribution, we have that $\text{Var}[\bar{v}_{N,M}] = M^{-1} \text{Var}[\hat{\vartheta}_N]$, and hence we can estimate variance of $\bar{v}_{N,M}$ by

$$\hat{\sigma}_{N,M}^2 := \frac{1}{M(M-1)} \sum_{m=1}^M (\hat{\vartheta}_N^m - \bar{v}_{N,M})^2. \quad (9)$$

Consequently,

$$L_{N,M} := \bar{v}_{N,M} - t_{\alpha, M-1} \hat{\sigma}_{N,M} \quad (10)$$

can be used as an approximate $100(1 - \alpha)\%$ confidence lower bound for the expectation $\mathbb{E}[\hat{\vartheta}_N]$. Note that the sample size M does not need to be large. Already with, say $M = 5$, one often can get an idea about variability of $\hat{\vartheta}_N$. The α -critical value $t_{\alpha, M-1}$, of t -distribution with $M - 1$ degrees of freedom is slightly bigger than the corresponding standard normal critical value z_α , and is used in order to compensate for the approximate nature of the lower confidence bound $L_{N,M}$, when M is small.

We have

$$\mathbb{E}[\hat{f}_{N'}(\bar{x}) - \bar{v}_{N,M}] = f(\bar{x}) - \mathbb{E}[\hat{\vartheta}_N] \geq f(\bar{x}) - \vartheta^*,$$

and hence

$$\hat{f}_{N'}(\bar{x}) - \bar{v}_{N,M} + z_\alpha \sqrt{\hat{\sigma}_{N'}^2(\bar{x}) + \hat{\sigma}_{N,M}^2} \quad (11)$$

provides a $100(1 - \alpha)\%$ confidence upper bound for the optimality gap. This upper bound is somewhat “conservative” since

$\vartheta^* \geq \mathbb{E}[\hat{\vartheta}_N]$, that is, $\bar{v}_{N,M}$ is a downward biased estimate of ϑ^* (see Ref. 2, Section 5.6.1, for a discussion of this bias problem and Ref. 10 for some ideas on reducing the computational burden of solving M problems).

CHANCE-CONSTRAINED PROBLEMS

Consider the following chance-constrained problem

$$\min_{x \in \mathcal{X}} f(x) \text{ subject to } p(x) \leq \alpha, \quad (12)$$

where $\alpha \in (0, 1)$ is a given significance level, $\mathcal{X} \subset \mathbb{R}^n$ is a closed set, $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $C: \mathcal{X} \times \Xi \rightarrow \mathbb{R}$, and $p(x) := \text{Prob}\{C(x, \xi) > 0\}$. The chance (probabilistic) constraint $p(x) \leq \alpha$ makes sure that the constraint $C(x, \xi) \leq 0$, depending on the random vector $\xi \in \Xi$, holds with probability at least $1 - \alpha$. For a general discussion of chance-constrained problems, we refer to Prékopa [11].

We can write the probability $p(x)$ in the form of the expectation,

$$p(x) = \mathbb{E}[\mathbf{1}_{(0, \infty)}(C(x, \xi))],$$

where $\mathbf{1}_{(0, \infty)}(\cdot)$ is the indicator function of $(0, \infty)$, that is, $\mathbf{1}_{(0, \infty)}(t) = 1$ if $t > 0$, and $\mathbf{1}_{(0, \infty)}(t) = 0$ if $t \leq 0$. Consequently $p(x)$ can be estimated by the corresponding sample average function

$$\hat{p}_N(x) = N^{-1} \sum_{j=1}^N \mathbf{1}_{(0, \infty)}(C(x, \xi^j)). \quad (13)$$

That is, $\hat{p}_N(x)$ is equal to the proportion of times that $C(x, \xi^j) > 0$, $j = 1, \dots, N$. An associated SAA problem can be written as

$$\min_{x \in \mathcal{X}} f(x) \text{ subject to } \hat{p}_N(x) \leq \gamma. \quad (14)$$

Note that we allow here for the significance level $\gamma \in [0, 1)$ of the SAA problem to be different from the significance level α of the true problem (Eq. 12).

As before, once the random sample is generated, the SAA problem can be viewed as a chance-constrained problem with scenarios $\xi^1, \dots, \xi^N$, each taken with probability $1/N$. For $\gamma > 0$, Problem (14) could be a

difficult combinatorial problem. Therefore, from a computational point of view it is advantageous to take $\gamma = 0$, in which case Problem (14) can be written as

$$\min_{x \in \mathcal{X}} f(x) \text{ subject to } C(x, \xi^j) \leq 0, j = 1, \dots, N. \quad (15)$$

If the functions $f(\cdot)$, $C(\cdot, \xi)$, $\xi \in \Xi$, are convex (linear), then Problem (15) is a convex (linear) programming problem and could be solved by efficient optimization procedures, provided that N is not too large.

Denote by $B(\cdot; \alpha, N)$, the cumulative distribution function of binomial distribution with probability of success α and number of trials N , that is,

$$B(k; \alpha, N) = \sum_{i=0}^k \binom{N}{i} \alpha^i (1-\alpha)^{N-i}, \quad k=0, \dots, N.$$

It is shown by Campi and Garatti [12], building on the work of Calafiore and Campi [13], that in the convex case under mild regularity conditions, if the sample size N satisfies the condition $B(n-1; \alpha, N) \leq \beta$, then with probability of at least $1 - \beta$, an optimal solution $\bar{x}$ of Problem (15) is a *feasible* point of the true problem (Eq. 12). This leads to an estimate of the sample size of order $N = O(\alpha^{-1})$, which guarantees with the specified confidence $1 - \beta$, that by solving Problem (15), one computes a feasible point of the true problem.

The above result, however, does not say much about optimality of the point $\bar{x}$. And, indeed, as the sample size N increases, the feasible set of Problem (15) will shrink to the set defined by constraints $C(x, \xi) \leq 0$ for all $\xi \in \Xi$, and hence eventually, the feasible set of Problem (15) becomes smaller than the feasible set of the true problem (Eq. 12). An alternative approach will be to take $\gamma = \alpha$. In that case, under mild regularity conditions, the optimal value and optimal solutions of the SAA problem (Eq. 14) will converge w.p.1 to their counterparts of the true problem (Eq. 12). For an approach to solving a certain class of chance-constrained SAA problems, we refer to the recent paper [14]. It is also possible to construct statistical upper and lower bounds for the optimal

value of the true problem (Eq. 12), by solving SAA problems of the form (Eq. 14) with $\gamma \in [0, 1]$ [15]. For some numerical experiments with various sampling-based approaches to chance-constrained problems see Ref. 16.

CONCLUDING REMARKS

For *convex* stochastic problems, an alternative to the SAA approach is the stochastic approximation (SA) method going back to Robbins and Monro [17]. The classical SA algorithm generates iterations by the formula

$$x_{j+1} = \Pi_{\mathcal{X}}(x_j - \gamma_j G(x_j, \xi^j)), \quad j = 1, \dots, \quad (16)$$

where $G(x, \xi) \in \partial_x F(x, \xi)$ is a subgradient of $F(x, \xi)$, $\Pi_{\mathcal{X}}$ is the metric projection onto the set $\mathcal{X}$, and $\gamma_j > 0$ are chosen step sizes. We may refer to Ref. 18 for a discussion of convergence properties of this method. The standard choice of the step sizes is $\gamma_j = \theta/j$ for some constant $\theta > 0$. For an *optimal* choice of the constant θ , the estimates of rates of convergence of this method are similar to the respective estimates of the SAA method. However, the method is very sensitive to the choice of the constant θ and often does not work well in practice (see, e.g., a simple example in Ref. 19, p. 1578, where it is demonstrated that a wrong choice of θ can result in a disastrously slow rate of convergence). It is somewhat surprising that a robust version of the SA algorithm, taking its origins in the mirror descent method of Nemirovski and Yudin [20], can significantly outperform SAA based algorithms for certain classes of convex stochastic problems [19].

So far we discussed static (two-stage and chance-constrained) stochastic programming problems. In principle, it is possible to extend the SAA methodology (including construction of upper and lower statistical bounds) to dynamic (multistage) stochastic programming by employing randomly generated scenarios. It should be noted, however, that the number of scenarios required to solve the true problem with a given accuracy $\varepsilon > 0$ grows exponentially with increase in the number of stages [21,22]. This makes

the SAA method practically inapplicable to multistage stochastic programming with a number of stages, say $T \geq 4$, unless the considered problems have a specific structure.

Acknowledgments

Research of this author was partly supported by the NSF award DMS-0914785.

REFERENCES

1. Dyer M, Stougie L. Computational complexity of stochastic programming problems. *Math Program* 2006;106:423–432.
2. Shapiro A, Dentcheva D, Ruszczyński A. *Lectures on stochastic programming: modeling and theory*. Philadelphia (PA): SIAM; 2009.
3. Rubinstein RY, Shapiro A. Optimization of static simulation models by the score function method. *Math Comput Simul* 1990;32:373–392.
4. Robinson SM. Analysis of sample-path optimization. *Math Oper Res* 1996;21:513–528.
5. Fishman GS. *Monte Carlo, concepts, algorithms and applications*. New York: Springer; 1999.
6. Niederreiter H. *Random number generation and quasi-Monte Carlo methods*. Philadelphia (PA): SIAM; 1992.
7. Kleywegt AJ, Shapiro A, Homem-de-Mello T. The sample average approximation method for stochastic discrete optimization. *SIAM J Optim* 2001;12:479–502.
8. Shapiro A. Monte Carlo approach to stochastic programming. In: Peters BA, Smith JS, Medeiros DJ, *et al.*, editors. *Proceedings of the 2001 Winter simulation conference*. 2001. pp. 428–431.
9. Mak WK, Morton DP, Wood RK. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Oper Res Lett* 1999;24:47–56.
10. Bayraksan G, Morton DP. Assessing solution quality in stochastic programs. *Math Program* 2006;108:495–514.
11. Prekopa A. *Stochastic programming*. Dordrecht, Boston (MA): Kluwer Academic Publishers; 1995.
12. Campi MC, Garatti S. The exact feasibility of randomized solutions of robust convex programs. *SIAM J Optim* 2008;19:1211–1230.

13. Calafiore G, Campi MC. The scenario approach to robust control design. *IEEE Trans Autom Control* 2006;51:742–753.
14. Luedtke J, Ahmed S. A sample approximation approach for optimization with probabilistic constraints. *SIAM J Optim* 2008;19:674–699.
15. Nemirovski A, Shapiro A. Convex approximations of chance constrained programs. *SIAM J Optim* 2006;17:969–996.
16. Pagnoncelli B, Ahmed S, Shapiro A. Computational study of a chance constrained portfolio selection problem. *J Optim Theory Appl* 2009;142:399–416.
17. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22:400–407.
18. Kushner HJ, Clark DS. *Stochastic approximation methods for constrained and unconstrained systems*. Berlin: Springer; 1978.
19. Nemirovski A, Juditsky A, Lan G, *et al.* Robust stochastic approximation approach to stochastic programming. *SIAM J Optim* 2009;19:1574–1609.
20. Nemirovski A, Yudin D. Volume 15, Problem complexity and method efficiency in optimization. *Wiley-Interscience Series Discrete Mathematics*. New York: John Wiley & Sons, Inc.; 1983.
21. Shapiro A, Nemirovski A. On complexity of stochastic programming problems. In: Jeyakumar V, Rubinov AM, editors. *Continuous optimization: current trends and applications*. Springer; 2005. pp. 111–144.
22. Shapiro A. On complexity of multistage stochastic programs. *Oper Res Lett* 2006;34:1–8.

SCENARIO GENERATION

WERNER RÖMISCH
Humboldt-University Berlin,
Institute of Mathematics,
Berlin, Germany

Many stochastic programming models may be rewritten into the form

$$\min \left\{ \int_{\Xi} f_0(x, \xi) P(d\xi) : x \in X, \right. \\ \left. \int_{\Xi} f_k(x, \xi) P(d\xi) \leq 0, k = 1, \dots, K \right\}, \quad (1)$$

where X is a closed subset of $\mathbb{R}^m$, Ξ a closed subset of $\mathbb{R}^d$, the functions f_k map from $\mathbb{R}^m \times \Xi$ to the extended real numbers $\mathbb{R}$ for $k = 0, \dots, K$, and P is a probability distribution on Ξ . The set X is used to describe all constraints not depending on P , and the set Ξ to contain the support of P . The integrands f_k are assumed to be lower semicontinuous jointly in (x, ξ) implying that all integrals in problem (1) are well defined (although possibly infinite).

Classical examples are *linear two-stage stochastic programs* and *optimization models with probabilistic constraints*. Linear two-stage models appear for $K := 0$ and f_0 having the representation

$$f_0(x, \xi) := \langle c(\xi), x \rangle + \inf \{ \langle q(\xi), y \rangle : W(\xi)y \\ = h(\xi) - T(\xi)x, y \geq 0 \} \quad (2)$$

by means of the infimum of a second stage linear program where some of the coefficients are affine functions of the d -dimensional random vector ξ and the variable x is the first stage decision. Models with probabilistic constraints appear, for example, for $K = 1$, $f_0(x, \xi) = \langle c, x \rangle$ and

$$f_1(x, \xi) = p - \mathbf{1}_{\{\xi \in \Xi: T(\xi)x \geq h(\xi)\}}(\xi),$$

where $\mathbf{1}_B$ denotes the characteristic function of a set B in $\mathbb{R}^d$ and $p \in (0, 1)$ is a probability level (see also **Models and Basic Properties**).

APPROXIMATION

Stability results for problem (1) with respect to approximations Q of the original probability distribution P (see Ref. 1 for a survey) state that infimal values $v(P)$ and $v(Q)$ and solution sets $S(P)$ and $S(Q)$ of the stochastic programs (1) with distributions P and Q , respectively, get close if the (uniform) distance of P and Q

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \int_{\Xi} f(\xi) P(d\xi) - \int_{\Xi} f(\xi) Q(d\xi) \right|, \quad (3)$$

with $\mathcal{F} = \{f_k(x, \cdot) : x \in X, k = 0, \dots, K\}$ gets small, the set X is compact, the objective function $x \mapsto \int_{\Xi} f_0(x, \xi) P(d\xi)$ is Lipschitz continuous on X and a metric regularity condition for the constraint set is satisfied. The latter two conditions are only needed if $K \geq 1$. As we have seen before, typical integrands f in stochastic programs are nondifferentiable or even discontinuous. (For simplicity, we assume here that all integrals in problem (1) are finite for every $x \in X$ and for the probability distribution P and its approximations Q .)

The most important way to approximate P consists in utilizing discrete probability measures Q_n having finite support

$$\text{supp}(Q_n) = \{\xi^1, \dots, \xi^n\} \subset \Xi,$$

for some $n \in \mathbb{N}$. The elements of $\text{supp}(Q_n)$ are often called *scenarios*. When replacing P by Q_n in problem (1), the (multivariate) integrals in problem (1) reduce to weighted sums and one obtains

$$\min \left\{ \sum_{i=1}^n q_i f_0(x, \xi^i) : x \in X, \right. \\ \left. \sum_{i=1}^n q_i f_k(x, \xi^i) \leq 0, k = 1, \dots, K \right\}, \quad (4)$$

where $q_i = Q_n(\{\xi^i\}) > 0$ is the probability that scenario ξ^i occurs ($i = 1, \dots, n$). Clearly,

we have $\sum_{i=1}^n q_i = 1$. The structure of the approximate stochastic programming problem (4) is very close to a standard (linear, nonlinear, and integer) optimization model. The only remaining difficulty consists in the need of many evaluations of the functions f_k at the pairs (x, ξ^i) , $i = 1, \dots, n$, if the number n of scenarios gets large. But, large n are often unavoidable when recalling numerical integration even in dimension $d = 1$ and all the more for large d as in many applied stochastic programming models, in production, energy, transportation, and finance. According to our stability considerations, the number n , the scenarios $\xi^i \in \Xi$ and their probabilities q_i for $i = 1, \dots, n$ should be selected such that for given $\varepsilon > 0$ the (absolute) error satisfies

$$e(Q_n) := \sup_{f \in \mathcal{F}} \left| \int_{\Xi} f(\xi) P(d\xi) - \sum_{i=1}^n q_i f(\xi^i) \right| \leq \varepsilon, \quad (5)$$

with $\mathcal{F}$ defined earlier. Another and more feasible condition consists in utilizing the relative error such that given $\varepsilon \in [0, 1]$, we look for a probability measure Q_n such that

$$e(Q_n) \leq \varepsilon e(\bar{Q}_1), \quad (6)$$

where the measure $\bar{Q}_1$ consists of only one distinguished scenario $\bar{\xi}$ with probability 1. A more advanced requirement consists in looking for Q_n with the smallest number $n = n_{\min}(\varepsilon, Q_n) \in \mathbb{N}$ of scenarios such that Equation (6) holds. Sometimes even $Q_0 = 0$ is considered in the literature [2] and, thus, the criterion (6) is of the form

$$e(Q_n) \leq \varepsilon \sup_{f \in \mathcal{F}} \left| \int_{\Xi} f(\xi) P(d\xi) \right|. \quad (7)$$

The behavior of $e(Q_n)$ with respect to $n \in \mathbb{N}$ and of $n_{\min}(\varepsilon, Q_n)$ with respect to ε is of considerable interest. In both cases, the dependence on the dimension d of P is crucial, too.

It is not surprising that the behavior of the two quantities depends heavily on the set $\mathcal{F}$ of integrands as well as on the probability distribution P . Since the set $\mathcal{F}$ in its present form is not very convenient to handle,

it might be an alternative to enlarge $\mathcal{F}$ in estimates (5–7). But, one has to be careful in this process as is shown next.

If $\mathcal{F}$ is the unit ball in the Banach space $\text{Lip}(\mathbb{R}^d)$ of Lipschitz continuous functions on $\Xi = \mathbb{R}^d$ and if $q_i = \frac{1}{n}$, $i = 1, \dots, n$, one obtains

$$e(Q_n) = Cn^{-\frac{1}{d}}, \quad (8)$$

for some constant C depending only on P under the weak assumptions that P is not singular with respect to the Lebesgue measure on $\mathbb{R}^d$ and that the moment $\int_{\mathbb{R}^d} \|\xi\|^{1+\delta} P(d\xi)$ is finite for some $\delta > 0$ [3, Theorem 6.2].

The rate (8) suggests that the unit ball in $\text{Lip}(\mathbb{R}^d)$ is too large and one should look for function classes $\mathcal{F}$ satisfying more restrictive conditions. One possibility is offered by the classical Koksma–Hlawka inequality [4,5]. It states for an integrand f on $\Xi = [0, 1]^d$ that is of bounded variation $V(f)$ in the sense of Hardy and Krause that the estimate

$$\left| \int_{\Xi} f(\xi) d\xi - \sum_{i=1}^n q_i f(\xi^i) \right| \leq V(f) \alpha^*(\lambda^d, Q_n), \quad (9)$$

holds with $q_i = \frac{1}{n}$, $i = 1, \dots, n$ and with the so-called *star-discrepancy* α^* defined by

$$\alpha^*(P, Q) = \sup_{\xi \in [0, 1]^d} |P([0, \xi]) - Q([0, \xi])|,$$

where $[0, \xi] = \times_{i=1}^d [0, \xi_i]$. For problem (1) with $K = 0$ and $\Xi = [0, 1]^d$, the Koksma–Hlawka inequality 9 leads to the estimate

$$\begin{aligned} e(Q_n) &\leq \sup_{f \in \mathcal{F}} V(f) \alpha^*(P, Q_n) \\ &= \sup_{x \in X} V(f_0(x, \cdot)) \alpha^*(P, Q_n). \end{aligned}$$

If $P = \lambda^d$ is the uniform distribution on $\Xi = [0, 1]^d$ and $q_i = \frac{1}{n}$, $i = 1, \dots, n$, it is known [4, Chapter 3.1] that there exist sequences $(\xi^i)_{i \in \mathbb{N}}$ with

$$\alpha^*(P, Q_n) = O(n^{-1}(\log n)^d). \quad (10)$$

Hence, compared with Equation (8), a much better convergence rate for $e(Q_n)$ can be

obtained under stronger assumptions on $\mathcal{F}$. Initiated by the pioneering work of Sloan and Woźniakowski [2], the convergence rate is further improved if the set $\mathcal{F}$ is bounded in the weighted tensor product space

$$W_2^{(1,\dots,1)}([0,1]^d) = \bigotimes_{i=1}^d W_2^1([0,1]), \quad (11)$$

where $W_2^1([0,1])$ is the Sobolev space of absolutely continuous real functions whose first derivatives belong to $L_2([0,1])$. The space (11) contains all real functions f on $[0,1]^d$, for which all mixed partial derivatives $\frac{\partial^{|u|}}{\partial \xi_u} f(\xi_u, 1)$ exist for almost every $\xi_u \in [0,1]^{|u|}$ and $u \subseteq D = \{1, \dots, d\}$ and for which the weighted norm

$$\|f\|_{d,\gamma} = \left(\sum_{u \subseteq D} \prod_{j \in u} \gamma_j^{-1} \int_{[0,1]^{|u|}} \left| \frac{\partial^{|u|}}{\partial \xi_u} f(\xi_u, 1) \right|^2 d\xi_u \right)^{\frac{1}{2}}$$

is finite. Here, we denote by $|u|$ the cardinality of $u \subseteq D$, and by $\xi_u \in [0,1]^{|u|}$, the vector containing the components of $\xi \in [0,1]^d$ whose indices are in u . The weights (γ_j) are positive and monotonically decreasing with $\gamma_1 = 1 = \prod_{j \in \emptyset} \gamma_j^{-1}$.

By utilizing the weighted Koksma–Hlawka inequality,

$$\left| \int_{\Xi} f(\xi) d\xi - \frac{1}{n} \sum_{i=1}^n f(\xi^i) \right| \leq \text{disc}_{\gamma}((\xi^i)) \|f\|_{d,\gamma}$$

with $\text{disc}(\xi) = \prod_{j=1}^d \xi_j - n^{-1} \sum_{\xi^i \in [0,\xi]} i$ and the weighted L_2 -star-discrepancy

$$\text{disc}_{\gamma}((\xi^i)) = \left(\sum_{\emptyset \neq u \subseteq D} \prod_{j \in u} \gamma_j \int_{[0,1]^{|u|}} \text{disc}^2(\xi_u, 1) d\xi_u \right)^{\frac{1}{2}}$$

instead of $\alpha^*(\lambda^d, Q_n)$ in Equation (9), it became possible in Refs 2 and 6 to prove the existence of a sequence $(\xi^i)_{i \in \mathbb{N}}$ such that for any $\delta > 0$ the estimate

$$e(Q_n) \leq C(\delta) n^{-1+\delta} \quad (12)$$

holds, where $C(\delta)$ is independent of n and d if the weights (γ_j) satisfy the condition $\sup_d \sum_{j=1}^d \gamma_j < \infty$. While the results in Refs 2 and 6 were nonconstructive, it is reported in Ref. 7 that certain shifted lattice rules attain the optimal order (12) of convergence.

SCENARIO GENERATION TECHNIQUES

We briefly discuss here about four different scenario generation techniques for stochastic programs without nonanticipativity constraints:

1. Monte Carlo simulation methods,
2. Quasi–Monte Carlo (QMC) methods,
3. Quadrature rules using sparse grids,
4. Optimal quantization (or discretization) of probability measures.

Monte Carlo Simulation Methods

Monte Carlo methods are based on drawing independent identically distributed (iid) Ξ -valued random samples $\xi^1(\cdot), \dots, \xi^n(\cdot), \dots$ [defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$] from an underlying probability distribution P (on Ξ) and on using the law of large numbers to obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(\xi^i(\omega)) = \int_{\Xi} f(\xi) P(d\xi) \quad \mathbb{P}\text{-almost surely}$$

for every real continuous and bounded function f on Ξ . Practically, iid samples are approximately obtained by pseudorandom number generators as uniform samples in $[0,1]^d$ and later transformed to more general sets and distributions. This technique may be applied, for example, to certain time series models (calibrated to the available statistical data) or to the statistical data directly. For details, refer to the article **Sampling Methods** in this encyclopedia.

Quasi–Monte Carlo Methods

The basic idea of QMC methods is to replace random samples in Monte Carlo methods by

deterministic points that are *uniformly distributed* in $[0, 1]^d$. One possibility for defining the latter property for a sequence $(\xi^i)_{i \in \mathbb{N}}$ is to require that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(\xi^i) = \int_{[0,1]^d} f(\xi) d(\xi) \quad (13)$$

holds for all real continuous functions f on $[0, 1]^d$. Basic references for QMC methods are Refs 4 and 8. With $P = \lambda^d$ (the d -dimensional Lebesgue measure) and $Q = Q_n$ denoting the probability measure with support $\text{supp}(Q_n) = \{\xi^1, \dots, \xi^n\}$ and with identical probabilities $\frac{1}{n}$ for all atoms, the estimate (9) highlights the importance of the star-discrepancy of the point set $\{\xi^1, \dots, \xi^n\}$

$$\text{disc}^*(\xi^1, \dots, \xi^n) := \alpha^*(\lambda^d, Q_n)$$

in this respect. Hence, one should look for *low discrepancy sequences*, that is, sequences $(\xi^i)_{i \in \mathbb{N}}$ such that $\text{disc}^*(\xi^1, \dots, \xi^n)$ is low for all n . Such sequences have the property

$$\frac{1}{2} n^{-1} \leq \text{disc}^*(\xi^1, \dots, \xi^n) = O(n^{-1}(\log n)^d).$$

In particular, we refer to the sequences discussed in Refs 8 (Section 5.4) and 9, namely, *Faure, Sobol, and Niederreiter sequences*. The latter are special cases of so-called (t, m, d) -sequences which in turn are based on (t, m, d) -nets [4, Chapter 4]. The latter (t, m, d) -nets and lattices [4, Chapter 5; 8, Section 5.3] represent (finite) *low discrepancy point sets* and are, presently, the most important sources for QMC methods.

If $G : [0, 1]^d \rightarrow \Xi$, $\Xi \subseteq \mathbb{R}^d$, is almost everywhere continuous, then Equation (13) means that the sequence (Q_n) converges weakly to λ^d and the continuous mapping theorem [10, Chapter 5] implies that the sequence $(Q_n G^{-1})$ converges weakly to $\lambda^d G^{-1}$. This fact allows to apply QMC methods to many situations in stochastic programming.

Quadrature Rules Using Sparse Grids

Again, we consider first the unit cube $[0, 1]^d$ in $\mathbb{R}^d$. Let nested sets of grids in $[0, 1]$ be given

$$\Xi^i = \{\xi_1^i, \dots, \xi_{m_i}^i\} \subset \Xi^{i+1} \subset [0, 1] \quad (i \in \mathbb{N}),$$

for example, the dyadic grid $\Xi^i = \{\frac{j}{2^i} : j = 0, 1, \dots, 2^i\}$. Then, the point set suggested by Smolyak [11]

$$H(q, d) := \bigcup_{\sum_{j=1}^d i_j = q} \Xi^{i_1} \times \dots \times \Xi^{i_d} \quad (14)$$

is called a *sparse grid* in $[0, 1]^d$ and points in $H(q, d)$ are also called *hyperbolic cross points*. In case of dyadic grids in $[0, 1]$, $H(q, d)$ consists of all d -dimensional dyadic grids with product of mesh size given by $\frac{1}{2^q}$ [12].

The corresponding tensor product quadrature rules for $q \geq d$ on $[0, 1]^d$ with respect to the Lebesgue measure λ^d are of the form

$$\begin{aligned} I(q, d) = & \sum_{q-d+1 \leq |\mathbf{i}| \leq q} (-1)^{q-|\mathbf{i}|} \binom{d-1}{q-|\mathbf{i}|} \\ & \times \sum_{j_1=1}^{m_{i_1}} \dots \sum_{j_d=1}^{m_{i_d}} f(\xi_{j_1}^{i_1}, \dots, \xi_{j_d}^{i_d}) \prod_{l=1}^d a_{j_l}^{i_l}, \end{aligned} \quad (15)$$

where $|\mathbf{i}| = \sum_{j=1}^d i_j$ and the coefficients $a_{j_l}^{i_l}$ ($j = 1, \dots, m_i$, $i = 1, \dots, d$) are weights of one-dimensional quadrature rules. Even if the one-dimensional weights are positive, some of the weights in Equation (15) are negative. Hence, an interpretation as scenario-based (discrete) probability measure is no longer possible. We note that the results in Ref. 13 (as extension of Ref. 2, see Section 1) apply to such tensor product quadrature rules.

Optimal Quantization of Probability Measures

Let D be a metric distance of probability measures on $\mathbb{R}^d$, for example, the Fortet–Mourier metric ζ_r of order r (to be used in the next section), the minimal L_r -metric ℓ_r used in Ref. 3 or some other metric [14] such that the underlying stochastic program behaves stable with respect to D . Furthermore, let P be a given probability distribution on $\mathbb{R}^d$. We are looking for a discrete probability measure Q_n with support $\text{supp}(Q_n) = \{\xi^1, \dots, \xi^n\}$,

$Q_n(\{\xi^i\}) = p_i, i = 1, \dots, n$, such that it is the best approximation to P in the sense

$$\begin{aligned} D(P, Q_n) &= \min\{D(P, Q) : |\text{supp}(Q)| \\ &= n, Q(\mathbb{R}^d) = 1\}. \end{aligned} \quad (16)$$

In many cases, the distance $D(P, Q)$ may be reformulated as a real function Φ defined on $[0, 1]^n \times \mathbb{R}^{dn}$. It attains its global minimum subject to the standard simplex constraint for the first n variables at $(p_1, \dots, p_n; \xi^1, \dots, \xi^n)$. Unfortunately, the function Φ is nonconvex and often nondifferentiable, hence, minimizing it is not an easy task.

We refer to Refs 15–17 for developing algorithmic procedures for minimizing Φ globally, for example, stochastic gradient algorithms, stochastic approximation methods, and stochastic branch-and-bound techniques.

The methodology of optimal quantization may be extended to multistage stochastic programs by incorporating constraints describing the tree structure [15,16].

SCENARIO REDUCTION

Let P be a probability measure on $\mathbb{R}^d$ having N scenarios ξ^i with probabilities $p_i, i = 1, \dots, N$. We consider P as approximation of the original probability distribution of a stochastic program that has to be solved computationally. Owing to running time requirements N might be too large and we have to look for an approximation Q_n of P whose support consists of only $n < N$ scenarios out of $\{\xi^1, \dots, \xi^N\}$. Then two questions arise: (i) Which of the N scenarios should be deleted and (ii) which probabilities should be assigned to the n remaining scenarios?

We review below the stability-based approach for (optimal) scenario reduction developed in the articles [18–20]. It is based on considering the distance $d_{\mathcal{F}}(P, Q_n)$ (see Eq. 3 in the section titled “Approximation”) and in determining Q_n as best approximation of P with respect to $d_{\mathcal{F}}$ among all probability measures whose supports consist of n scenarios out of $\{\xi^1, \dots, \xi^N\}$. An equivalent formulation of this best approximation problem is as follows: Let Q_J denote a probability measure

on $\mathbb{R}^d$ with $\text{supp}(Q_J) = \{\xi^i : i \in I \setminus J\}$ for some index set $J \subset I := \{1, \dots, N\}$ and let q_i be the probability of the scenario indexed by i for every $i \in I \setminus J$. Then the minimization problem

$$\begin{aligned} \min \Big\{ d_{\mathcal{F}}(P, Q_J) : J \subset I, |J| = N - n, \\ q_i \geq 0, i \in I \setminus J, \sum_{i \in I \setminus J} q_i = 1 \Big\} \end{aligned} \quad (17)$$

determines some index set J_n and weights $\bar{q}_i \in [0, 1]$ such that the probability measure $Q_n := Q_{J_n}$ with scenarios ξ^i and probabilities $\bar{q}_i$ for $i \in I \setminus J_n$ solves the best approximation problem. The formulation (17) of optimal scenario reduction leads immediately to a decomposition into an *inner* and an *outer* minimization problem, namely,

$$\begin{aligned} \min_J \Big\{ \inf_q \Big\{ d_{\mathcal{F}}(P, Q_J) : q_i \geq 0, i \in I \setminus J, \\ \sum_{i \in I \setminus J} q_i = 1 \Big\} : J \subset I, |J| = N - n \Big\}. \end{aligned} \quad (18)$$

In particular, the approach becomes powerful if solutions and the infimum $D_J(\mathcal{F}, P)$ of the inner problem may be determined explicitly for any index set $\emptyset \neq J \subset I$. In that case, the *optimal redistribution* of the N probabilities $p_i, i \in I$, to the n scenarios indexed by $I \setminus J$ is known and it remains to solve the outer (combinatorial) optimization problem

$$\min\{D_J(\mathcal{F}, P) : J \subset I, |J| = N - n\} \quad (19)$$

at least approximately. Problem (19) is known as *n-median problem* and as *NP-hard*.

Unfortunately, for many function classes $\mathcal{F}$, it is impossible to determine $D_J(\mathcal{F}, P)$ as well as the optimal redistribution explicitly. In particular, this is true for models with probabilistic constraints and for two-stage mixed-integer stochastic programs in which case (rectangular, polyhedral) discrepancies appear as distances $d_{\mathcal{F}}$ [21,22]. For discrepancies, however, the inner problem may be solved by linear programming [21]. For two-stage linear stochastic programs, the

integrands $f_0(x, \cdot)$ in Equation (2) are often proportional to certain elements of

$$\mathcal{F}_r(\Xi) = \{f : \Xi \rightarrow \mathbb{R} : |f(\xi) - f(\tilde{\xi})| \leq c_r(\xi, \tilde{\xi}), \\ \forall \xi, \tilde{\xi} \in \Xi\},$$

for some $r \in \mathbb{N}$ and with the (cost) function c_r defined by

$$c_r(\xi, \tilde{\xi}) := \max\{1, \|\xi\|^{r-1}, \|\tilde{\xi}\|^{r-1}\} \|\xi - \tilde{\xi}\| \\ (\xi, \tilde{\xi} \in \Xi)$$

(see Refs 1 and 23). The corresponding distance $\zeta_r := d_{\mathcal{F}_r}$ is known as *Fortet–Mourier metric* of order r . In the latter case the inner problem is explicitly solvable, and it holds

$$D_J(\mathcal{F}_r, P) = \sum_{j \in J} p_j \min_{i \notin J} \hat{c}_r(\xi^i, \xi^j) \quad (20)$$

$$\bar{q}_i = p_i + \sum_{\substack{j \in J \\ i(j)=i}} p_j \text{ and}$$

$$i(j) \in \arg \min_{i \notin J} \hat{c}_r(\xi^i, \xi^j), i \in I \setminus J, \quad (21)$$

that is, the redistribution rule consists in adding the probability of a deleted scenario indexed by $j \in J$ to the probability of a remaining scenario that is nearest to ξ^j with respect to the distance $\hat{c}_r$ on $\text{supp}(P)$. The so-called reduced cost $\hat{c}_r$ has the representation

Forward algorithm for scenario reduction

Step 0: $J^{[0]} := \{1, \dots, N\}$.

Step k: $u_k \in \arg \min_{u \in J^{[k-1]}} \sum_{j \in J^{[k-1]} \setminus \{u\}} p_j \min_{i \notin J^{[k-1]} \setminus \{u\}} \hat{c}_r(\xi^i, \xi^j),$
 $J^{[k]} := J^{[k-1]} \setminus \{u_k\}.$

Step n+1: Redistribution with $J := J^{[n]}$ via (21).

Similarly, the idea of the backward algorithm is based on the second special case of problem (19) with expression (20) for $n = N - 1$.

SCENARIO TREES FOR MULTISTAGE STOCHASTIC PROGRAMS

The special feature of multistage stochastic programs (see *Two-Stage Stochastic*

$$\hat{c}_r(\xi^i, \xi^j) := \min \left\{ \sum_{k=1}^{\ell-1} c_r(\xi^{i_k}, \xi^{i_{k+1}}) : \ell \in \mathbb{N}, \right. \\ \left. i_k \in I, k = 1, \dots, \ell, i_1 = i, i_\ell = j \right\}.$$

Two simple heuristic algorithms for solving Equation (19) are proposed in Refs 19 and 20: the *forward (selection)* and the *backward (reduction) heuristic*. To give an idea of the heuristics, we give a short description of the forward algorithm. Its basic idea originates from the simple structure of Equation (19) with Equation (20) for the special case $n = 1$. It is of the form

$$\min_{u \in \{1, \dots, N\}} \sum_{\substack{j=1 \\ j \neq u}}^N p_j \hat{c}_r(\xi^u, \xi^j).$$

If the minimum is attained at u^* , the index set $J = \{1, \dots, N\} \setminus \{u^*\}$ solves Equation (19) for $n = 1$. The scenario ξ^{u^*} is taken as the first element of $\text{supp}(Q)$. Then the separable structure of D_J is exploited to determine the second element of $\text{supp}(Q)$ while the first element is fixed (see Fig. 1). The process is continued until n elements of $\text{supp}(Q)$ are selected.

Programs: Introduction and Basic Properties consists in imposing information constraints on the decisions. The information flow is modeled by a filtration of σ -fields $\mathcal{A}_t$, $t = 1, \dots, T$, which is associated to the stochastic input process $\xi = (\xi_t)_{t=1}^T$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Typically, it is required that the σ -field $\mathcal{A}_t$ is generated by the random vector $(\xi_1, \dots, \xi_t)$. Then the information or *nonanticipativity* constraint

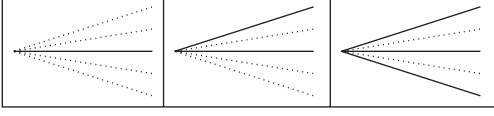


Figure 1. Illustration of selecting the first, second, and third scenario out of $N = 5$.

means measurability of the decisions x_t with respect to $\mathcal{A}_t$ for every $t = 1, \dots, T$. Mostly, $t = 1$ refers to the present. Thus, ξ_1 is deterministic and $\mathcal{A}_1 = \{\emptyset, \Omega\}$.

Clearly, any scenario-based approximation of the underlying probability distribution P of ξ has to reflect the growth of the σ -fields. Hence, the scenarios need to be *tree-structured*. In general, there are two ways to generate scenario trees, namely, (i) a tree-structure is prescribed and scenarios are generated via conditional distributions for increasing t starting with a root at $t = 1$, or (ii) in a first step, a number of scenarios is generated for the whole horizon $t = 1, \dots, T$ based on the distribution P and according to some method reported in the section titled “Scenario Generation Techniques.” Secondly, a tree structure is generated successively by bundling scenarios.

Several specific techniques for generating scenario trees are known from the literature. We refer to the survey [24] and the more recent articles [15–17, 25–34]. Most of them are discussed in the introduction of Ref. 26.

Finally, we review an application of scenario reduction techniques (see the section titled “Scenario Reduction”) to generate scenario tree. We consider an original (nonstructured) scenario set $\{\xi^1, \dots, \xi^N\}$, satisfying the root condition $\xi_1^i = \xi_1^*$, $i = 1, \dots, N$, and the parameter sets $\{1, \dots, t\}$ for increasing t and start at $t = 2$ with the reduction of the original scenario set restricted to $t = 2$. This leads to (say) k_2 remaining scenarios and certain clusters C_2^k of scenarios for $k = 2, \dots, k_2$ that are associated to one of the remaining scenarios via the next neighbor property. In general, we obtain (disjoint) partitions or clusters

$$C_t := \{C_t^1, \dots, C_t^{k_t}\}, \quad (k_t \in \mathbb{N})$$

of the index set $I = \{1, \dots, N\}$ for every $t = 2, \dots, T$. The following *forward algorithm* leads to a scenario tree process ξ_{tr} whose structure may be controlled by certain tolerances ε_t , $t = 2, \dots, T$.

Algorithm: *Forward scenario tree generation*

Step 1: Initialization

Set $C_1 = \{I\}$, $k_1 = 1$ and $t := 2$.

Step 2: Cluster computation Let $C_{t-1} = \{C_{t-1}^1, \dots, C_{t-1}^{k_{t-1}}\}$. Perform scenario reduction w.r.t. the Fortet–Mourier metric ζ_r for every scenario subset $\{\xi_t^i\}_{i \in C_{t-1}^k}$ separately for every $k \in \{1, \dots, k_{t-1}\}$ and only with respect to the t th component. Define index sets J_t^k and I_t^k of deleted and remaining scenarios and mappings $i_t^k : J_t^k \rightarrow I_t^k$ such that

$$i_t^k(j) \in \arg \min_{i \in I_t^k} \hat{c}_r(\xi_t^i, \xi_t^j), \quad (j \in J_t^k),$$

according to the rule (21). Define the partition C_t and the mapping $\alpha_t : I \rightarrow I$ by

$$C_t := \left\{ \alpha_t^{-1}(i) \mid i \in I_t^k, k = 1, \dots, k_{t-1} \right\} \quad \text{and} \quad \alpha_t(j) = \begin{cases} i_t^k(j), & j \in J_t^k, \\ j, & \text{else.} \end{cases} \quad (22)$$

C_t is a refinement of the partition C_{t-1} . If $t < T$ perform the cluster computation for $t = t + 1$ and go to Step 2.

Step 3: Tree generation According to the partition C_T and the mapping α_T (see Eq. 22) the scenario tree process ξ_{tr} is defined such that its k th scenario is

$$\xi_{\text{tr}}^k = \left(\xi_1^*, \xi_2^{\alpha_2(i)}, \dots, \xi_t^{\alpha_t(i)}, \dots, \xi_T^{\alpha_T(i)} \right) \quad \text{for some } i \in C_T^k$$

with probability $q_k := \sum_{i \in C_T^k} p_i$ for every $k = 1, \dots, k_T$. The error at step t is given by

$$\text{err}_t := \sum_{k=1}^{k_t-1} \sum_{j \in J_t^k} p_j \min_{i \in I_t^k} \hat{c}_r(\xi_t^i, \xi_t^j).$$

Hence, the generation procedure may be controlled by the condition $\text{err}_t \leq \varepsilon_t$ for every $t = 2, \dots, T$. We refer to Ref. 26 for further information and some computational experience.

Acknowledgment

The work of the author is supported by the DFG Research Center MATHEON “Mathematics for key technologies” in Berlin.

REFERENCES

1. Römisch W. Stability of stochastic programming problems. In: Ruszczyński A, Shapiro A, editors. Volume 10, Stochastic programming, Handbooks in operations research and management science. Amsterdam: Elsevier; 2003. pp. 483–554.
2. Sloan IH, Woźniakowski H. When are Quasi Monte Carlo algorithms efficient for high-dimensional integration. *J Complex* 1998;14:1–33.
3. Graf S, Luschgy H. Foundations of quantization for probability distributions. Volume 1730, Lecture notes in mathematics. Berlin: Springer; 2000.
4. Niederreiter H. Random number generation and Quasi-Monte Carlo methods. Philadelphia (PA): SIAM; 1992.
5. Götz M. Discrepancy and the error in integration. *Monatsh Mathe* 2002;136:99–121.
6. Hickernell FJ, Woźniakowski H. Integration and approximation in arbitrary dimensions. *Adv Comput Math* 2000;12:25–58.
7. Kuo FY, Sloan IH. Lifting the curse of dimensionality. *Not AMS* 2005;52:1320–1328.
8. Lemieux C. Monte Carlo and Quasi-Monte Carlo Sampling, Springer Series in Statistics. New York: Springer; 2009.
9. Pennanen T, Koivu M. Epi-convergent discretizations of stochastic programs via integration quadratures. *Numer Math* 2005;100:141–163.
10. Billingsley P. Convergence of probability measures. New York: Wiley; 1968.
11. Smolyak SA. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Sov Math Dokl* 1963;4:240–243.
12. Novak E, Ritter K. High dimensional integration of smooth functions over cubes. *Numer Math* 1996;75:79–97.
13. Novak E, Woźniakowski H. Intractability results for integration and discrepancy. *J Complex* 2001;17:388–441.
14. Rachev ST, Rüschendorf L. Volume I, Theory, Mass transportation problems. New York: Springer; 1998.
15. Pflug GCH. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Math Program* 2001;89:251–271.
16. Hochreiter R, Pflug GCh. Financial scenario generation for stochastic multi-stage decision processes as facility location problem. *Ann Oper Res* 2007;152:257–272.
17. Pagès G, Pham H, Printems J. Optimal quantization methods and applications to numerical problems in finance. In: Rachev ST, editor. Handbook of computational and numerical methods in finance. Boston (MA): Birkhäuser; 2004. pp. 253–297.
18. Dupačová J, Gröwe-Kuska N, Römisch W. Scenario reduction in stochastic programming: an approach using probability metrics. *Math Program* 2003;95:493–511.
19. Heitsch H, Römisch W. Scenario reduction algorithms in stochastic programming. *Comput Optim Appl* 2003;24:187–206.
20. Heitsch H, Römisch W. A note on scenario reduction for two-stage stochastic programs. *Oper Res Lett* 2007;35:731–738.
21. Henrion R, Küchler C, Römisch W. Discrepancy distances and scenario reduction in two-stage stochastic mixed-integer programming. *J Ind Manag Optim* 2008;4:363–384.
22. Henrion R, Küchler C, Römisch W. Scenario reduction in stochastic programming with respect to discrepancy distances. *Comput Optim Appl* 2009;43:67–93.
23. Römisch W, Wets RJ-B. Stability of ε -approximate solutions to convex stochastic programs. *SIAM J Optim* 2007;18:961–979.

24. Dupačová J, Consigli G, Wallace SW. Scenarios for multistage stochastic programs. *Ann Oper Res* 2000;100:25–53.
25. Casey M, Sen S. The scenario generation algorithm for multistage stochastic linear programming. *Math Oper Res* 2005;30:615–631.
26. Heitsch H, Römisch W. Scenario tree modeling for multistage stochastic programs. *Math Program* 2009;118:371–406.
27. Heitsch H, Römisch W. Scenario tree reduction for multistage stochastic programs. *Comput Manag Sci* 2009;6:117–133.
28. Kuhn D. Generalized bounds for convex multistage stochastic programs. Volume 548, *Lecture notes in economics and mathematical systems*. Berlin: Springer; 2005.
29. Kuhn D. Aggregation and discretization in multistage stochastic programming. *Math Program* 2008;113:61–94.
30. Høyland K, Kaut M, Wallace SW. A heuristic for moment-matching scenario generation. *Comput Optim Appl* 2003;24:169–185.
31. Pflug GCh, Mirkov R. Tree approximations of dynamic stochastic programs. *SIAM J Optim* 2007;18:1082–1105.
32. Pennanen T. Epi-convergent discretizations of multistage stochastic programs via integration quadratures. *Math Program* 2009;116:461–479.
33. Shapiro A. Inference of statistical bounds for multistage stochastic programming problems. *Math Meth Oper Res* 2003;58:57–68.
34. Frauendorfer K. Barycentric scenario trees in convex multistage stochastic programming. *Math Program Ser B* 1996;75:277–293.

SCHEDULING SEASIDE RESOURCES AT CONTAINER PORTS

FRANK MEISEL
Institute of Business Studies,
Martin-Luther-University
Halle-Wittenberg, Halle, Germany

INTRODUCTION

Container terminals (CTs) in seaports are logistics facilities that handle containerized cargo by serving vessels, trucks, and trains. They link the different modes of transportation for providing access of a terminal's hinterland to international trade routes. The general layout and the available equipment alternatives of seaport terminals are sketched in Fig. 1. At the seaside, vessels moor at the berths of a terminal where quay cranes (QCs) transship the containers between the vessels and the terminal area. The transport of containers between the terminal areas is usually performed by yard trucks or by straddle carriers. If yard trucks are used for the transport operations, cranes are deployed within the yard for stacking and retrieving containers. Automated variants of yard trucks and straddle carriers are called automated guided vehicles (AGVs) and automated lifting vehicles (ALVs), respectively. These vehicles are guided completely by a terminal's control system and need no driver onboard. For details on the technical equipment, the interested reader is referred to the studies of various authors [1–5] and to the websites of equipment manufacturers [6–11].

From the point of view of vessel operators, the service time of a vessel at a terminal is unproductive and should be kept as short as possible. They therefore expect that terminals provide sufficient handling capacity in terms of quay space, technical equipment, and workforce. This led to substantial extensions of existing terminals, new building of terminals, and upgrading of handling equipment throughout terminals worldwide in recent years. For details on

such projects; see Ilmer [13]. Strategic and tactical decisions that come along with these projects are, for example, to determine locations and layouts of new terminals and to decide on the equipment to use.

Next to the provision of service capacity, the operational performance of available terminal resources is a further driver of service quality. Fast and reliable services require minimizing delays before the service of a vessel starts and avoiding large handling times due to the low productivity of the equipment [14,15]. Hence, the best possible utilization of the available resources through effective planning and scheduling of equipment and workforce enables fast service operations and creates competitive advantage for a terminal. Therefore, various CT planning problems have been identified and research increasingly supports practitioners by developing models and algorithms for solving these management tasks. Table 1 lists the most often addressed planning tasks. This article puts focus on the operational planning and scheduling problems related to the berth space and quay crane resources of a terminal, namely the berth allocation problem, the quay crane assignment problem, and the quay crane scheduling problem.

The article is organized as follows. For supporting readers with a comprehensive introduction to seaport operations planning, the section titled “An Overview of Operational Planning and Scheduling Problems” briefly summarizes those planning problems of Table 1 that are not the focus of this article. A more detailed version of these explanations is given by Meisel [12]. Further overviews on CT operations planning are provided by various authors [16–22]. The sections titled “Berth Allocation and Quay Crane Assignment” and “Quay Crane Scheduling” provide in-depth studies on the planning problems related to the seaside resources. The section titled “Berth Allocation and Quay Crane Assignment” explains the problems of berth allocation and quay crane assignment in detail, provides a comprehensive literature survey based on Bierwirth and Meisel [23], and presents two models proposed in recent studies. The section titled “Quay

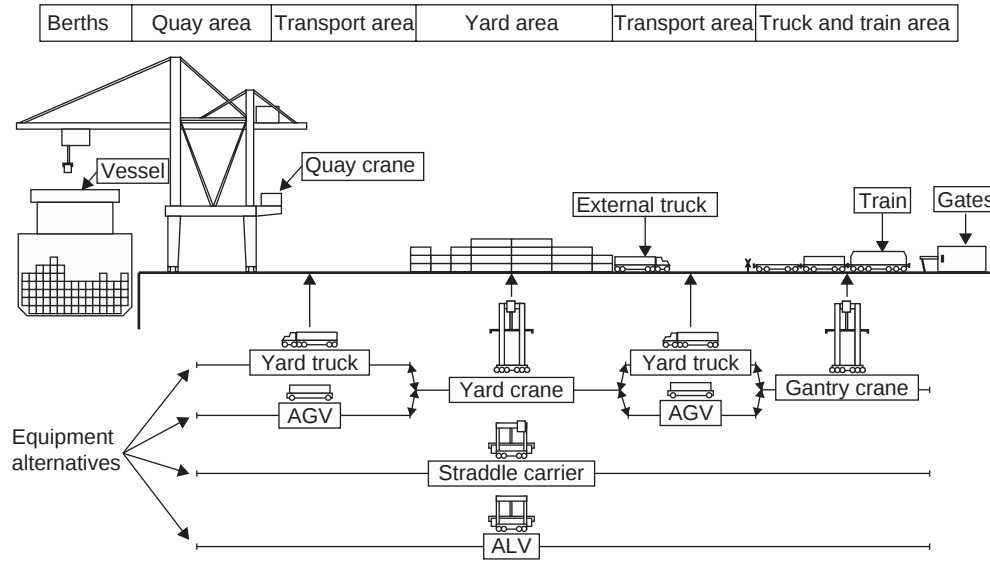


Figure 1. Schematic cross-sectional view of a container terminal [12].

Table 1. Major Operational Planning Tasks of a Container Terminal

Planning Task	Decisions to Make
Berth allocation	Berthing positions and berthing times of vessels to be served
Quay crane assignment	Number of QCs that serve a vessel Specific cranes to use for the service
Quay crane scheduling	Scheduling of loading and unloading operations for a vessel
Stowage planning	Slot positions of export containers within a vessel
Yard management	Reservation of yard capacity for liner services Selection of storage locations for individual containers Remarshalling operations of containers
Yard crane scheduling	Deployment of cranes to yard blocks Scheduling of stackings and retrievals of containers
Horizontal transport management	Assignment of vehicles to QCs Scheduling of transport orders on vehicles
Hinterland operations planning	Service order of trucks and trains Sequencing loading and unloading operations
Workforce planning	Provision of workforce capacity Scheduling of labor tasks

Crane Scheduling” presents the quay crane scheduling problem at the same level of detail. The final section concludes this article.

AN OVERVIEW OF OPERATIONAL PLANNING AND SCHEDULING PROBLEMS

Stowage planning means to arrange containers onboard a vessel by assigning a slot position to every container that has to be

transported from a considered terminal to successive terminals on a vessel’s route [24]. Stowage planning is often made on two levels of aggregation, where a coarse stowage plan is generated by assigning container classes to slots, and, then, a precise stowage plan is generated by assigning individual containers to slots of their container class [25]. The assignment of containers to slots has to respect container attributes like size and

weight as well as stability constraints of a vessel [26,27]. A typical objective is the minimization of reshuffling of containers, which occurs whenever a container dedicated to a late terminal in the vessel's route is stowed on top of a container that has to be unloaded earlier [24,28].

Yard management comprises reserving yard blocks for liner services, selecting storage locations for individual containers, and planning remarshalling operations. Yard blocks have to be reserved for buffering import and export containers of a calling vessel. For maximizing the utilization of yard space over time, space is reserved dynamically for a calling vessel [29]. When selecting storage locations of individual containers, minimization of transfer time of vehicles or minimization of reshuffles is often the pursued objective [30]. Dekker *et al.* [31] show that a stacking policy in which export containers of a same class are stored in the same yard stack reduces reshuffling at the time of the vessel service. Yard remarshalling means to reposition containers if they are not stored within their dedicated area, or if their stacking order requires reshuffling. These operations are executed ahead of the service process of a vessel in periods of low workload. The objective is to minimize the number of moves performed to restack and reposition containers [32,33].

Yard crane scheduling comprises the deployment of cranes at yard blocks and the scheduling of container stacking and retrieval operations. Deployment of yard cranes (YCs) means to dynamically assign cranes to those yard blocks where stacking and retrieval operations have to be performed. The pursued objective is the minimization of an unfinished workload, which needs to be carried from one period to the subsequent planning period [3,34]. The order of stacking and retrieving containers in the yard is usually preset by the schedules of the QCs. Scheduling YC operations is still required, for example, to retrieve the containers of a yard bay at a minimum number of reshuffle operations [35,36]. If two YCs operate within 1 yard block, crane interference needs to be considered [37,38]. Since YC operations are interrelated with further terminal

operations, scheduling of these cranes is often investigated in combination with stowage planning or QC scheduling [39–41].

Horizontal transport management comprises the assignment of vehicles to cranes and the scheduling of transport orders on the vehicles. With the first decision, vehicles are either assigned exclusively to QCs or they are pooled where each vehicle serves different QCs [42–44]. The pooling strategy requires fewer vehicles and reduces empty movement but entails more complicated dispatching. Scheduling transport orders on terminal vehicles is made on a short planning horizon of a few minutes to respect the actual progress of QC operations. Since most vehicle types carry only a single container at a time, pickup-and-delivery routing problems have to be solved here. One exception is dual-load AGVs, which can transport two 20' containers at a time [45,46]. If automated vehicles are used, routing and traffic control must be managed thoroughly to avoid blockages of concurrent loading and unloading processes (so-called *deadlocks*) [47–49].

Hinterland operations concern the service of trains and external trucks. Planning these service operations is often similar to planning the service of vessels with the exception of modified technical and organizational constraints and objectives. For example, since containers are not stacked on trains, terminal operators are more flexible in sequencing the loading and unloading operations. Consequently, a reduction potential exists for empty travel of terminal vehicles and for reshuffling of containers in the yard. Further objectives are the minimization of shunting activities of trains within the terminal [18], the minimization of waiting time of external trucks [50], and the minimization of the number of cranes used for serving external trucks and trains [51].

Workforce planning comprises the provision of sufficient workforce capacity and the scheduling of labor tasks. To decide on an appropriate workforce capacity, workforce requirements must be estimated, for example, by considering the projected workload of the horizontal transport system within the coming work shifts [42,43]. At the seaside of a terminal, the required

workforce capacity is defined by the QCs that must be manned within a shift. Therefore, Bierwirth and Meisel [52] minimize the number of manned QCs per work shift within a combined berth allocation and QC assignment problem. For a given workforce, Kim *et al.* [53] investigate the assignment of individual operators to QCs, YCs, and yard trucks with respect to work plans and the laborer's skills. Scheduling of labor tasks is often covered by the already described equipment scheduling problems but it is sometimes treated as an individual planning problem, for example, if operators are required to connect reefer containers to electric supply [54]. Objectives of labor task scheduling include the minimization of tardy task completions [54] and the minimization of the number of required laborers [55].

BERTH ALLOCATION AND QUAY CRANE ASSIGNMENT

Problem Description

Container vessels call at a terminal according to fixed liner schedules. From these schedules, the set of vessels to be served within a planning horizon is known to the terminal management, together with relevant information like expected times of arrival and an approximate number of the containers to load and unload. In the *berth allocation problem* (BAP), a berthing position and a berthing time are assigned to each vessel that is to be served within the planning horizon. A typical planning horizon is a few days up to one week. The decisions made have to respect the layout of a CT, which means that vessels must be moored within the boundaries of the quay and alongside the straight segments of the quay wall; see Fig. 2. Differing from naval ports and general cargo terminals where vessels may be repositioned

to other berthing positions [56–58], container vessels stay at the assigned position during the entire service to keep the handling time as short as possible. The projected berthing time, handling time, and berthing position of a vessel determine the temporal occupation of the allocated berth space, that is, the time span within which no other vessel can be moored at this position. Figure 3 provides an illustration of a berth plan for 29 vessels at a terminal with a quay length of 1000 m and a planning horizon of 114 h. The plan shows the projected berthing position and berthing time for each vessel. The most frequently pursued objective in berth allocation models is the minimization of vessel waiting time in order to achieve earliest possible departure times [59–61].

Handling times of vessels are important input for the BAP. Since vessel handling times depend strongly on the number of serving cranes, the problem of assigning quay cranes to vessels is closely interrelated with berth allocation and both problems are often solved jointly [62]. The plan shown in Fig. 3, therefore, also comprises an assignment of cranes to vessels. It illustrates the number of cranes assigned to each vessel on an hourly basis and the utilization of the 10 cranes available at this terminal for each period of the planning horizon. The task of the *quay crane assignment problem* (QCAP) is to allocate sufficient QC capacity to every vessel for fulfilling the required loading and unloading operations of containers. The major decision to make is the number of cranes to assign to each of the vessels that are served simultaneously, which is sometimes also referred to as the *crane split* [18]. This decision clearly impacts the handling time of a vessel, because a larger number of cranes at a vessel can fulfill the loading and unloading operations more quickly. However, owing to mutual crane interference, additional cranes may lower

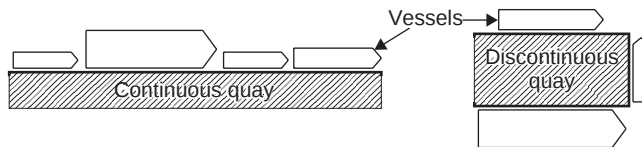


Figure 2. Berthing vessels with respect to the terminal layout.

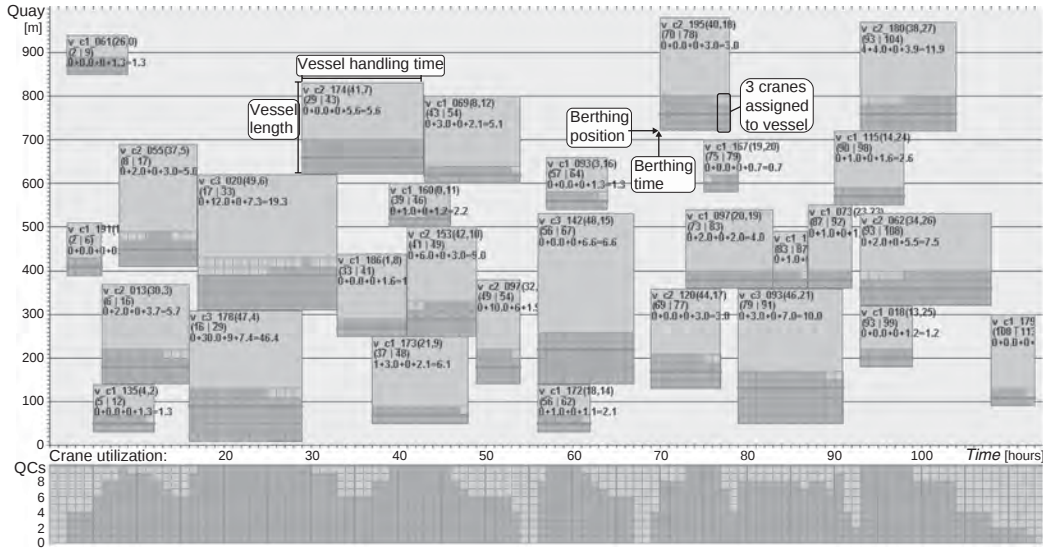


Figure 3. An example of a berth plan and crane assignment.

the average crane productivity, such that a decreasing marginal productivity is often observed at CTs. Therefore, the maximum number of cranes to assign to a vessel is usually bounded, for example, to two to six cranes per vessel depending on its size [63]. Obviously, the total number of QCs available at the terminal must be considered when distributing cranes among vessels. Moreover, a feasible assignment of cranes to vessels has to respect the berthing times preset by the berth plan. If crane interference is considered within crane assignment, one objective is to minimize the total handling time of vessels by keeping crane productivity loss as low as possible. A further objective can be to find an assignment that allows serving vessels without exceeding the departure times projected in the berth plan [64]. The second decision of the QCAP is to decide on the specific QCs to assign to a vessel. Here, one must take into account that QCs are mounted on a rail track alongside the quay and, thus, they cannot pass each other. Therefore, a specific assignment is only feasible if the relative positions of cranes are preserved at all times. When deciding on the specific QCs to serve a vessel, the number of crane setups at vessels

[62] or the movement of cranes between vessels [65] can be minimized.

A Review of Models Proposed in the Literature

Models for berth allocation are distinguished depending on the layout of the berths into discrete problems, continuous problems, and hybrid problems [23,66]. In discrete problems, the quay is partitioned into a number of sections, called *berths*. Only one vessel can be served at a berth at a time. In a continuous problem, there is no partitioning of the quay, that is, vessels can berth at arbitrary positions within the boundaries of the quay. In hybrid layouts, the quay is partitioned into berths, but large vessels may occupy more than one berth while small vessels may share a berth. Moreover, in the so-called *static problems*, vessels would have already arrived at the terminal when berth planning is to be done, whereas in dynamic problems the arrival times of vessels within the planning horizon must be respected [67].

The discrete BAP has been studied in the static variant by Imai *et al.* [67,68], Hansen and Oğuz [69], and Lee *et al.* [70]. The objective pursued in the four papers is to minimize the vessels' port stay time, which is the total of waiting and handling time of a vessel at

the terminal. Dynamic variants of the discrete BAP have been investigated in a larger number of studies but with different objectives. The minimization of port stay times of vessels within a discrete dynamic BAP has been investigated by various authors [59–61,67,69,71–76]. The minimization of tardy departures of vessels is investigated in Refs 59, 77 and 78. Imai *et al.* [79] minimize the weighted number of vessels that cannot be served until a given due date. In the work of Giallombardo *et al.* [80,81] and Hansen *et al.* [82] also, the service costs of vessels depend on their assigned berth and, hence, berthing positions are taken up in the objective function. Most of the above studies assume that the handling time of a vessel depends on its assigned berth. The studies of Lee *et al.* [70], Imai *et al.* [75], Liang *et al.* [78], and Giallombardo *et al.* [80,81] additionally incorporate assignment or scheduling decisions for the QCs into berth allocation models to determine more realistic handling times for the vessels.

The continuous static BAP has been investigated by various authors [62,83–86]. Guan *et al.* [84] propose to minimize the total weighted completion time of vessels, whereas Lee *et al.* [83] and Oğuz *et al.* [86] propose to minimize the maximum completion time among the vessels. In the models of Park and Kim [62] and Rashidi [85] a combination of port stay times, tardiness penalties, and berthing position related cost is minimized. As observed for the discrete BAP, also the continuous BAP attracted much more research in its dynamic variant [52,64–66,87–104]. The studies of various authors [64–66,87,88,97,101,103,104] pursue the minimization of waiting and/or handling time of vessels. Tardiness objectives are addressed by various authors [65,89–92,98,100,104]. In the continuous BAP also the berthing positions of vessels are considered in the objective function, for example, to minimize the deviation from a desired berthing position for a vessel or to minimize the total berth space occupied by the vessels [62,85,88–96,104]. In several studies on the continuous BAP, it is assumed that vessel handling times are given, fixed, and independent of the decisions made within

berth allocation. This assumption is relaxed in Imai *et al.* [66] and Meisel and Bierwirth [98], where vessel handling times depend on the berthing position of a vessel, and in Refs 52, 62, 64, 65, 85, 86 and 98–104 where QC assignment or QC scheduling decisions are incorporated into berth allocation models.

BAP models for hybrid layouts, where large vessels can occupy several berths or small vessels can share the same berth are presented in Refs 61, 105–112. All papers investigate the dynamic variant of a hybrid BAP. The pursued objectives are the minimization of vessel waiting and/or handling time [61,105,106,108–110,112] and the minimization of a measure related to the berthing position of vessels [107,111]. The handling times of vessels are assumed to be given and fixed [105–107], or assumed to depend on the berthing position of vessels [61,108–111], or they are determined by an additional assignment of cranes to vessels [112].

Two Examples of Models Proposed in the Literature

The following subsections present two examples of berth allocation models that are proposed in the recent literature. The models of Imai *et al.* [79] and Meisel and Bierwirth [98] have been selected to illustrate the broad range of different issues that are covered by the published studies. The study of Imai *et al.* [79] is of interest, because in their model, a certain service level is guaranteed for the vessels through a restricted maximum waiting time. The assignment of quay cranes to vessels is not in the scope of this model. In Meisel and Bierwirth [98] continuous berth allocation and crane assignment are solved jointly to respect the impact of berth space availability and crane utilization on the service quality of vessels.

A Model for Discrete Berth Allocation with a Waiting Time Limit for Vessels. Discrete dynamic berth allocation with a waiting time limit for vessels has been introduced by Imai *et al.* [79]. In this problem, a terminal is considered where the quay is partitioned into Q berths. Berths $1, 2, \dots, Q-1$ are owned by the terminal operator who decides on the berth allocation. Berth Q serves as a

subcontracted external buffer, which handles vessels in situations of excessive workload at terminals 1 to $Q - 1$. $B = \{1, 2, \dots, Q\}$ is the set of all berths. For each berth $i \in B$, S_i is the point in time when the berth becomes available, for example, after finishing the service of previously assigned vessels.

A set of vessels $V = \{1, 2, \dots, n\}$ is projected to be served at the terminal, where n is the total number of vessels. For each vessel $j \in V$, an arrival time A_j is given and also a handling time C_{ij} that is needed for serving vessel j at berth $i \in B$. Moreover, a waiting time limit L_j is given for vessel j to ensure a certain level of service quality. Vessels that cannot be served at one of the berths 1 to $Q - 1$ without exceeding the waiting time limit are assigned to the external berth Q . It is assumed that the external berth shows sufficient capacity for handling vessels quickly within acceptable waiting times. Further assumptions are as follows:

- Each berth can handle one vessel at a time.
- Each berth shows sufficient water depth to berth arbitrary vessels.
- The handling time of a vessel depends solely on its allocated berth.
- Vessels are served without preemption, that is, once the process has been started to serve a vessel it is not interrupted until the service is completed.

The decisions to make are to assign a berth to each calling vessel and to find a sequence for serving the vessels at the berths. Both decisions are represented by the binary variable x_{ijk} , which is set to 1, if vessel j is served as the $(n - k + 1)$ th last vessel at berth i . Further decision variables y_{ijk} measure the idle time of berth i between the departure of the $(n - k + 2)$ th last vessel and the arrival of the $(n - k + 1)$ th last vessel when j is served as the $(n - k + 1)$ th last vessel. The objective of the berth allocation process is the minimization of the total handling time of vessels at the external berth Q , which corresponds to a minimum payment of service charges to the external berth operator for handling vessels.

The discrete BAP with a waiting time limit is formulated as follows:

minimize

$$Z = \sum_{j \in V} \sum_{k \in U} C_{Qj} \cdot x_{Qjk} \quad (1)$$

s.t.

$$\sum_{i \in B} \sum_{k \in U} x_{ijk} = 1 \quad \forall j \in V, \quad (2)$$

$$\sum_{j \in V} x_{ijk} \leq 1 \quad \forall i \in B, k \in U, \quad (3)$$

$$S_i - A_j + \sum_{l \in V} \sum_{m \in P_k} (C_{il}x_{ilm} + y_{ilm}) + y_{ijk} + (x_{ijk} - 1) \cdot M \leq L_j \quad \forall i \in B \setminus \{Q\}, \\ j \in V, k \in U, \quad (4)$$

$$\sum_{l \in V} \sum_{m \in P_k} (C_{il}x_{ilm} + y_{ilm}) + y_{ijk} - (A_j - S_i)x_{ijk} \geq 0 \quad \forall i \in B, j \in W_i, k \in U, \quad (5)$$

$$x_{ijk} \in \{0, 1\} \quad \forall i \in B, j \in V, k \in U, \quad (6)$$

$$y_{ijk} \geq 0 \quad \forall i \in B, j \in V, k \in U. \quad (7)$$

Objective (1) minimizes the total handling time of vessels that are allocated to the external terminal. Here, $U = \{1, 2, \dots, n\}$ denotes a set of service orders that is used for indexing the vessels within the service sequence at a berth. Constraint set (2) ensures that every vessel gets assigned a berth and a position within the service order of that berth. Constraints (3) enforce that every berth serves up to one ship at any time. Constraints (4) assure that vessels do not wait longer than the time limit if they are served at one of the berths 1 to $Q - 1$. For this inequality, P_k is a subset of U such that $P_k = \{p \mid p < k \in U\}$, that is, $P_k = \{1, 2, \dots, k - 1\}$, and M is a sufficiently large value. In these constraints, the waiting time of a vessel is evaluated as a function of the ready time of the assigned berth, the arrival time of the vessel, the handling time of preceding vessel services, and berth idle times caused by late arriving vessels in the service sequence. Constraints (5) enforce that the service of a vessel cannot start before the arrival time. In this constraint, W_i is the subset of those vessels $j \in V$ that arrive after the

ready time of berth i , that is, $A_j \geq S_i$. The constraints (6) and (7) define domains for the decision variables.

A Model for Continuous Berth Allocation and Quay Crane Assignment. The joint problem of continuous berth allocation and crane assignment is formally described as follows. A terminal with a quay of length L is considered, where Q QCs are available for serving vessels. The planning horizon of the problem is H hours, where T is a corresponding set of 1 h time periods, that is, $T = \{0, 1, \dots, H - 1\}$.

A set of vessels $V = \{1, 2, \dots, n\}$ is projected to be served within the planning horizon, where n is the total number of vessels. The crane capacity demand of vessel $i \in V$ to fulfill all loading and unloading operations is m_i QC-hours. The minimum and maximum numbers of QCs to assign to the vessel are denoted by $r_i^{\min}$ and $r_i^{\max}$, yielding the range $R_i = [r_i^{\min}, r_i^{\max}]$. The length of vessel i is denoted by l_i and a desired berthing position b_i^0 is specified for every vessel within the vicinity of those yard areas that have been reserved for storing its import and export containers. Furthermore, an expected time of arrival ETA_i is given. Berthing the vessel earlier than ETA_i is possible by a speedup on its journey to the terminal. The realizable speedup, however, is bounded to an earliest starting time $EST_i \leq ETA_i$, that is, the vessel cannot be berthed earlier than EST_i . Finally, an expected finishing time EFT_i and a latest finishing time LFT_i are given for the vessel.

The following assumptions are made:

- Each quay position shows sufficient water depth to berth arbitrary vessels.
- Vessels are served without preemption.
- It takes no time to move a QC from one vessel to another vessel.
- Every crane has the technical capability to serve every vessel. Furthermore, the cranes are identical, that is, they show the same maximum productivity.

The decisions to make are to determine a berthing time s_i , a berthing position b_i , and the number of QCs to assign to each vessel $i \in V$ in its service periods such that a cost

measure is minimized. The berthing time s_i of a vessel corresponds to the start of the first period with cranes assigned, whereas its departure time e_i is defined by the end of the last period with cranes assigned. The time span between s_i and e_i defines the handling time of vessel i . The assignment of cranes to vessels is represented by a binary decision variable r_{itq} . It is set to 1 if and only if exactly q QCs are assigned to vessel i at time t . To evaluate a solution, the deviation from the desired berthing position $\Delta b_i = |b_i^0 - b_i|$, the necessary speedup $\Delta ETA_i = (ETA_i - s_i)^+$, and the tardiness $\Delta EFT_i = (e_i - EFT_i)^+$ are determined for each vessel i . Figure 4 illustrates the vessel data and variables. The following binary decision variables are additionally introduced for the mathematical formulation of the outlined problem: r_{it} is set to 1 if at least one QC is assigned to vessel i at time t . u_i is set to 1 if the finishing time of vessel i exceeds LFT_i . y_{ij} is set to 1 if vessel i is berthed below vessel j , that is, $b_i + l_i \leq b_j$. Finally, z_{ij} is set to 1 if the service of vessel i ends not later than when the service of vessel j starts.

The continuous BAP with integrated QC assignment is formulated as follows:

minimize

$$Z = \sum_{i \in V} \left(c_i^1 \cdot \Delta ETA_i + c_i^2 \cdot \Delta EFT_i + c_i^3 \cdot u_i + c^4 \cdot \sum_{t \in T} \sum_{q \in R_i} (q \cdot r_{itq}) \right) \quad (8)$$

s.t.

$$\sum_{t \in T} \sum_{q \in R_i} (q^\alpha \cdot r_{itq}) \geq (1 + \beta \cdot \Delta b_i) \cdot m_i \quad \forall i \in V, \quad (9)$$

$$\sum_{i \in V} \sum_{q \in R_i} (q \cdot r_{itq}) \leq Q \quad \forall t \in T, \quad (10)$$

$$\sum_{q \in R_i} r_{itq} = r_{it} \quad \forall i \in V, \quad \forall t \in T, \quad (11)$$

$$\sum_{t \in T} r_{it} = e_i - s_i \quad \forall i \in V, \quad (12)$$

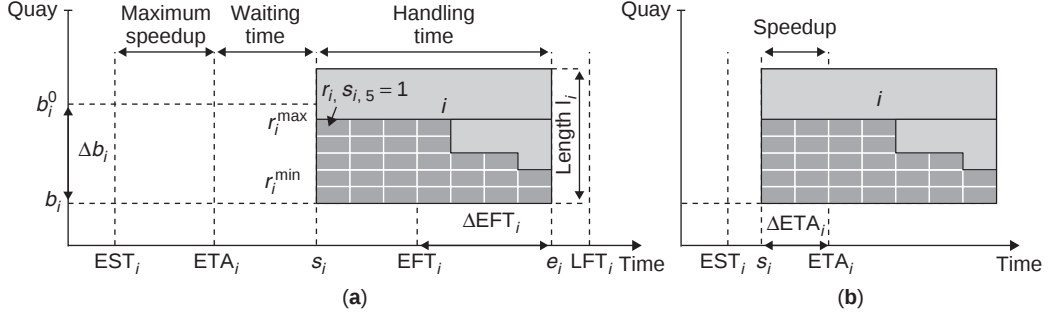


Figure 4. Vessel data—waiting before berthing (a) and speedup case (b).

$$(t+1) \cdot r_{it} \leq e_i \quad \forall i \in V, \forall t \in T, \quad (13)$$

$$t \cdot r_{it} + H \cdot (1 - r_{it}) \geq s_i$$

$$\forall i \in V, \forall t \in T, \quad (14)$$

$$\Delta b_i \geq b_i - b_i^0 \quad \forall i \in V, \quad (15)$$

$$\Delta b_i \geq b_i^0 - b_i \quad \forall i \in V, \quad (16)$$

$$\Delta ETA_i \geq ETA_i - s_i \quad \forall i \in V, \quad (17)$$

$$\Delta EFT_i \geq e_i - EFT_i \quad \forall i \in V, \quad (18)$$

$$M \cdot u_i \geq e_i - LFT_i \quad \forall i \in V, \quad (19)$$

$$b_j + M \cdot (1 - y_{ij}) \geq b_i + l_i$$

$$\forall i, j \in V, i \neq j, \quad (20)$$

$$s_j + M \cdot (1 - z_{ij}) \geq e_i \quad \forall i, j \in V, i \neq j, \quad (21)$$

$$y_{ij} + y_{ji} + z_{ij} + z_{ji} \geq 1 \quad \forall i, j \in V, i \neq j, \quad (22)$$

$$s_i, e_i \in \{EST_i, \dots, H\} \quad \forall i \in V, \quad (23)$$

$$b_i \in \{0, 1, \dots, L - l_i\} \quad \forall i \in V, \quad (24)$$

$$\Delta ETA_i, \Delta EFT_i \geq 0 \quad \forall i \in V, \quad (25)$$

$$r_{itq}, r_{it}, u_i, y_{ij}, z_{ij} \in \{0, 1\}$$

$$\forall i, j \in V, \forall t \in T, \forall q \in R_i. \quad (26)$$

The optimization model pursues the minimization of the total cost arising from the service of all vessels within the planning horizon. For every vessel, speedup cost for berthing earlier than ETA_i , tardiness cost for exceeding the expected finishing time EFT_i , and penalty cost for exceeding the latest allowed finishing time LFT_i are summed up in Equation (8). The corresponding cost rates are denoted as c_i^1 , c_i^2 , and c_i^3 . While speedup cost and tardiness cost grow constantly in time, penalty cost is incurred only once, if the departure of the vessel is beyond the latest allowed finishing time LFT_i . A fourth cost type is added to evaluate the utilized QC-hours within a berth plan. The cost rate per QC-hour is denoted as c^4 . This objective accounts for the decreasing marginal productivity of QCs and the resulting trade-off between accelerating the handling of a vessel and the operational cost of QCs.

Constraints (9) ensure that every vessel receives the required QC capacity with respect to productivity losses by QC interference and the chosen berthing position. Interference among QCs leads to reduced marginal productivity. Several authors proposed to model this productivity loss using a so-called *interference exponent* [86,113–115]. For a given interference exponent α ($0 < \alpha \leq 1$), the productivity obtained from assigning q cranes to a vessel for 1 h is given by a total of q^α QC-hours. This effect is incorporated on the left-hand side of constraint (9). The productivity of a terminal is also affected by the workload of horizontal transport means. This workload is minimal if a vessel berths at its desired berthing position b_i^0 .

If the actually chosen berthing position is apart from the desired position, the load of the horizontal transport increases, which can also lead to a productivity loss of cranes due to starvation. This productivity loss is modeled here by an increase in the vessel's QC capacity demand. Let $\beta \geq 0$ denote the relative increase of QC capacity demand per unit of berthing position deviation. Hence, a vessel positioned Δb_i quay segments away from its desired berthing position requires $(1 + \beta \cdot \Delta b_i) \cdot m_i$ QC-hours to be served as is modeled on the right side of constraint (9).

Constraints (10) enforce that at most Q cranes are utilized in a period. In every period a certain number of QCs is assigned to every vessel, which is either zero or taken from the range R_i . A consistent setting of the corresponding variables r_{it} and r_{itq} is ensured by Equation (11). Constraints (12–14) set the starting times and ending times for serving vessels without preemption. Constraints (15–18) determine the deviations from the desired berthing position, expected arrival time, and expected finishing time for each vessel. Variable u_i indicates whether the handling of vessel i ends later than LFT_i . It is set by constraints (19) where M denotes a large positive number. Constraints (20) and (21) set the variables y_{ij} and z_{ij} , which are used to avoid overlapping the handling of vessels in the space-time diagram in constraints (22). The arrival of a vessel can be expedited to at most the earliest starting time EST_i . Moreover, the planning horizon H defines a limit on the departure time of the vessels. Both aspects are reflected in constraints (23). Constraints (24) ensure that each vessel is positioned within the quay boundaries. Constraints (25) and (26) define domains for the remaining decision variables.

Discussion of the Models. The models presented cover different issues of berth allocation. When facing a practical terminal situation, one model will likely fit better than the other. For example, if the quay must be partitioned into distinct berths due to the layout of the terminal, then a discrete BAP model is needed such as that discussed in the section titled “A Model for Discrete Berth Allocation with a Waiting Time Limit for

Vessels.” Moreover, when each berth holds a set of dedicated cranes, the assumption that vessel handling times depend solely on the allocated berth is also justified. In this case, all cranes at a berth will jointly serve an allocated vessel and, therefore, the resulting vessel handling time at a berth can be estimated precisely. An explicit assignment of cranes to vessels is not required. In contrast, when the quay of a terminal is linear, terminal operators will exploit the flexibility in choosing berthing positions and assigning cranes for improved resource utilization and better service quality. Using a model for continuous berth allocation with integrated crane assignment such as that in the section titled “A Model for Continuous Berth Allocation and Quay Crane Assignment” is essential for such a terminal. Here, handling times are adjustable from assigning cranes to vessels. The resulting variable occupation of the quay space impacts on the berth allocation decisions. The flexibility gained in service operations is accounted for by a more complex measurement of service quality in the objective function. Moreover, operational cost for utilizing the crane resource should also be considered in such a situation as is done in the model presented.

Both continuous and discrete dynamic BAP models are in general $\mathcal{NP}$ -hard optimization problems. Therefore, heuristics are proposed in the literature for solving problem instances of practical size. For the models investigated here, a genetic algorithm, a squeaky wheel optimization, and a tabu search are proposed by Imai *et al.* [79] and Meisel and Bierwirth [98].

QUAY CRANE SCHEDULING

Problem Description

In the quay crane scheduling problem (QCSP), a work plan is generated for the quay cranes that are assigned to a vessel. Differing from berth allocation, the QCSP is solved shortly before the service process of a vessel starts. One reason is that the set of containers whose loading and unloading operations have to be scheduled on the cranes is not known until a vessel leaves its

previous port of call. Therefore, input data for the crane scheduling can be provided by vessel operators to a terminal's management only on a short-term basis. For the QCSP, the containers that have to be loaded and unloaded are considered as a set of tasks to be scheduled on the cranes. The tasks describe the granularity in which the workload of a vessel is considered in a QCSP model. As illustrated in Fig. 5, a task can be defined by the loading and unloading operations of containers of a bay area, a single bay, a container stack, a container group, or an individual container, respectively [116–120]. Precedence relations among tasks can be given to ensure that unloading operations precede loading operations and to represent the stacking of containers as defined by a stowage plan. A solution to the problem, called a *QC schedule*, defines a starting time for every task on a crane. Usually, the minimization of the makespan of the QC schedule is pursued, because a minimum makespan corresponds to a shortest possible vessel handling time [100,117,119]. Another objective of crane scheduling is the

maximization of crane utilization [116,121]. To avoid the crossing of cranes in a schedule and to comprise the compliance with safety margins between adjacent cranes, the QCSP requires so-called interference constraints [100,119,122,123]. Figure 6 shows a crane schedule for a vessel with 10 bays and 18 container groups that is served by 3 cranes. Here, the cranes have to keep a safety margin of two bays at all times. The latest completion time among all tasks is 233 time units, which corresponds to the makespan of the schedule and to the handling time of the vessel.

A Review of Models Proposed in the Literature

Quay crane scheduling models are distinguished into models defining tasks to be scheduled on the basis of bay areas, single bays, container groups, container stacks, or individual containers. In the simplest case, tasks refer to bay areas, where each area is exclusively served by one QC. If the bay areas are nonoverlapping, the problem is to partition the bays of a vessel such that the

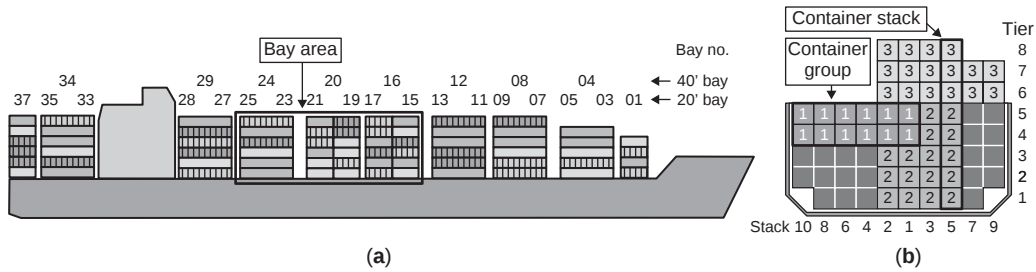


Figure 5. Storage location structure of a vessel (a) and a bay (b).

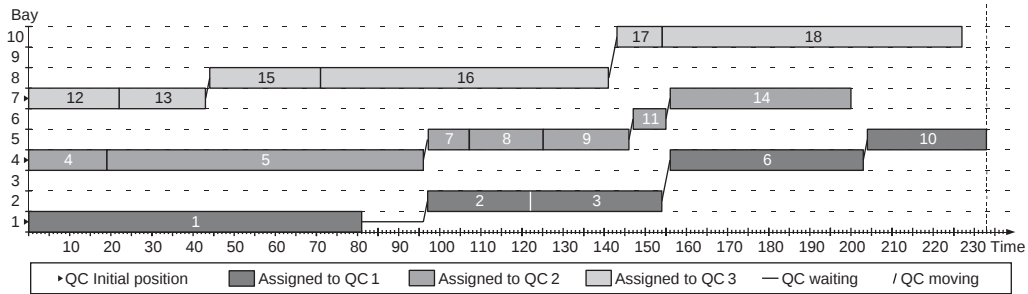


Figure 6. An example of a quay crane schedule.

workload is uniformly distributed among the cranes [116,121]. The partitioning ensures the noncrossing of the rail-mounted QCs. In addition, safety margins can be incorporated to avoid that adjacent cranes come too close to each other in a solution [122,124].

A better workload distribution is often achievable, if tasks are defined by single bays instead of bay areas. Since the operating areas of cranes can overlap, a noncrossing constraint must be incorporated to ensure feasibility of schedules. Except for the early models of Daganzo [125] and Peterkofsky and Daganzo [126], noncrossing constraints are a standard in crane scheduling models for complete bays [65,100,117,127–133]. Safety margins are additionally incorporated in the models of Liu *et al.* [100] and Choo *et al.* [133]. The pursued objective is the minimization of the schedule makespan [127,100,117,128–131,133], the minimization of the total weighted completion time of tasks [132] or the minimization of the crane travel time [65]. In the models of Daganzo [125], Peterkofsky and Daganzo [126], and Choo *et al.* [133] the service is planned for several vessels simultaneously and the pursued objective is the minimization of the total weighted completion time of vessels.

Enabling the cranes to share the workload of bays is the most typical feature of models where tasks are defined by container groups. Models for crane scheduling with container groups incorporate precedence relations among tasks to ensure that certain container groups are unloaded before other groups are loaded and to respect the stacking of containers within a bay [12,119,123,134–140]. Except for Ng and Mak [137], who consider only the noncrossing of cranes, all models for crane scheduling with container groups respect the noncrossing and safety margins to avoid crane interference. All studies propose the minimization of the schedule makespan as an objective. When cranes are scheduled for several vessels in parallel, minimization of the total finishing time of QCs and minimization of total vessel tardiness can also serve as objectives [136].

Stack-based crane scheduling has been investigated by various authors [118,141–146]. In these studies, a single

crane is considered that processes the container stacks of one bay. The focus is on the parallelization of loading and unloading of different stacks in a crane schedule. This so-called *crane double cycling* can considerably decrease the service time of a vessel bay. In Zhang and Kim [146] the consideration is extended toward the handling of hatches that cover adjacent stacks of containers. In the model of Meisel and Wichmann [120] crane scheduling is investigated on the basis of single containers. Besides the realization of crane double cycling in a schedule, reshuffled containers can be repositioned within a bay instead of temporarily unloading them. Such internal reshuffles further accelerate the service of a vessel bay. All the mentioned studies pursue a minimization of the schedule makespan.

Quay crane scheduling issues have been furthermore incorporated in some advanced approaches for stowage planning, horizontal transport operations planning, and yard crane scheduling [41,27,147–151].

Two Examples of Models Proposed in the Literature

The following sections present two models for QC scheduling. The model of Goodchild and Daganzo [143] is remarkable because it incorporates crane double cycling into the service operations planning for a crane at a bay of a container vessel. The model of Bierwirth and Meisel [123] considers multiple QCs that jointly serve one container vessel. This model is taken up here, because it contains elaborate constraints for avoiding crossing of cranes and for ensuring safety margins between adjacent cranes.

A Model for Quay Crane Scheduling with Container Stacks. The QCSP for container stacks has been formulated by Goodchild and Daganzo [143]. In this problem, a single bay of a container vessel is considered. One QC must unload the container stacks in the bay and load new containers into the cleared stack. The set of stacks is denoted by S . For each stack $i \in S$, the number of containers to unload is given by u_i and the number of containers to load is given by l_i .

The decisions to be made concern finding an order for the unloading operations and the loading operations. These decisions are represented by binary variables x_{ij} , 1 if $i \in S$ is unloaded before $j \in S$ and 0 otherwise, and y_{ij} , 1 if $i \in S$ is loaded before $j \in S$ and 0 otherwise. Furthermore, FU_i denotes the completion time for unloading stack i and FL_i denotes the completion time for loading stack i . The objective is the minimization of the handling time w of the considered bay, which corresponds to the maximum completion time among all operations. This goal is achieved through simultaneously performing loading and unloading operations, which means to apply crane double cycling in a solution whenever possible.

The following assumptions are made for the problem:

- No hatch covers separate above-deck and below-deck containers in the bay.
- The processing time for loading or unloading a stack is directly proportional to the number of containers to transship.
- Loading or unloading a stack is performed continuously without any interruption, that is, the operations are performed nonpreemptive.
- Crane operations cannot lead to an instable load configuration of the vessel.

The QCSP for container stacks is formulated as follows:

$$\text{minimize } w \quad (27)$$

s.t.

$$w \geq FL_i \quad \forall i \in S, \quad (28)$$

$$FL_i - FU_i \geq l_i \quad \forall i \in S, \quad (29)$$

$$FU_i - FU_j + M \cdot x_{ij} \geq u_i \quad \forall i, j \in S, \quad (30)$$

$$FU_j - FU_i + M \cdot (1 - x_{ij}) \geq u_j \quad \forall i, j \in S, \quad (31)$$

$$FL_i - FL_j + M \cdot y_{ij} \geq l_i \quad \forall i, j \in S, \quad (32)$$

$$FL_j - FL_i + M \cdot (1 - y_{ij}) \geq l_j \quad \forall i, j \in S, \quad (33)$$

$$FU_i \geq u_i \quad \forall i \in S, \quad (34)$$

$$x_{ij}, y_{ij} \in \{1, 0\} \quad \forall i, j \in S. \quad (35)$$

The objective (27) is to minimize the handling time of the bay, which is determined (28) by the maximum completion time of all loading operations of stacks. Constraints (29) ensure that a stack is loaded after it has been unloaded. Constraints (30) and (31) ensure that stacks are unloaded one after the other, where either stack i is unloaded before stack j if $(x_{ij} = 1)$ or j is unloaded before i ($x_{ij} = 0$). Accordingly, constraints (32) and (33) handle the loading operations of stacks. Constraints (34) and (35) define feasible domains of decision variables.

A Model for Quay Crane Scheduling with Container Groups.

In the QCSP for container groups, a set of tasks $\Omega = \{1, 2, \dots, n\}$ and a set of QCs $Q = \{1, 2, \dots, q\}$ are given. Each task $i \in \Omega$ represents a loading or unloading operation of a certain container group. The tasks have individual processing times p_i and bay positions l_i . In addition, dummy tasks 0 and $T = n + 1$ with processing times $p_0 = p_T = 0$ are given to indicate the start and end of the service of the vessel. Further task sets are defined by $\Omega^0 = \Omega \cup \{0\}$, $\Omega^T = \Omega \cup \{T\}$, and $\overline{\Omega} = \Omega \cup \{0, T\}$. Precedence relations may exist between pairs of tasks that are located within the same bay. Let Φ denote the set of precedence constrained task pairs. Furthermore, let $\Psi \supseteq \Phi$ denote the set of all task pairs for which it is known in advance that they cannot be processed simultaneously. For each crane $k \in Q$ a ready time r^k and an initial bay position l_0^k are given. Without loss of generality, it is assumed that the cranes are indexed sequentially according to their initial positioning alongside the vessel. All QCs can move between two adjacent bays in an identical travel time $\hat{t} > 0$. The travel time of QC k to traverse from its initial position l_0^k to l_j ($j \in \Omega$) is defined as $t_{0j}^k = \hat{t} \cdot |l_0^k - l_j|$. The travel time between bay positions l_i and l_j ($i, j \in \Omega$) is defined as $t_{ij} = \hat{t} \cdot |l_i - l_j|$. If i or j or both belong to $\{0, T\}$ the travel time is set to $t_{ij} = 0$. It is supposed that cranes are not allowed to cross each

other and that they have to keep a safety margin δ , measured in units of bays. The problem is to find completion times c_i for all tasks $i \in \overline{\Omega}$ with respect to the constraints such that the completion time c_T of the final task T (i.e., the makespan) is minimized.

The assumptions of the QCSP with container groups are as follows:

- Container groups are predefined from given stowage plans.
- Processing of tasks is nonpreemptive.
- All QCs show an identical, deterministic transshipment productivity. For this reason fixed processing times are given for the tasks.
- Crane operations cannot lead to an instable load configuration of the vessel.
- There is sufficient space beside the vessel to place idle QCs outside of the vessel area.

The QCSP for container groups is formulated as follows:

$$\text{minimize } c_T \quad (36)$$

s.t.

$$\sum_{j \in \Omega^T} x_{0j}^k = 1 \quad \forall k \in Q, \quad (37)$$

$$\sum_{j \in \Omega^0} x_{jT}^k = 1 \quad \forall k \in Q, \quad (38)$$

$$\sum_{k \in Q} \sum_{j \in \Omega^T} x_{ij}^k = 1 \quad \forall i \in \Omega, \quad (39)$$

$$\sum_{j \in \Omega^0} x_{ji}^k - \sum_{j \in \Omega^T} x_{ij}^k = 0 \quad \forall i \in \Omega, \forall k \in Q, \quad (40)$$

$$c_i + t_{ij} + p_j - c_j \leq M(1 - x_{ij}^k) \quad \forall i, j \in \overline{\Omega}, \quad \forall k \in Q, \quad (41)$$

$$c_i + p_j - c_j \leq 0 \quad \forall (i, j) \in \Phi, \quad (42)$$

$$c_i + p_j - c_j \leq M(1 - z_{ij}) \quad \forall i, j \in \Omega, \quad (43)$$

$$c_j - p_j - c_i \leq Mz_{ij} \quad \forall i, j \in \Omega, \quad (44)$$

$$z_{ij} + z_{ji} = 1 \quad \forall (i, j) \in \Psi, \quad (45)$$

$$\sum_{u \in \Omega^0} x_{ui}^v + \sum_{u \in \Omega^0} x_{uj}^w \leq 1 + z_{ij} + z_{ji} \quad \forall (i, j, v, w) \in \Theta, \quad (46)$$

$$c_i + \Delta_{ij}^{vw} + p_j - c_j \leq M \left(3 - z_{ij} - \sum_{u \in \Omega^0} x_{ui}^v - \sum_{u \in \Omega^0} x_{uj}^w \right) \quad \forall (i, j, v, w) \in \Theta, \quad (47)$$

$$c_j + \Delta_{ij}^{vw} + p_i - c_i \leq M \left(3 - z_{ji} - \sum_{u \in \Omega^0} x_{ui}^v - \sum_{u \in \Omega^0} x_{uj}^w \right) \quad \forall (i, j, v, w) \in \Theta, \quad (48)$$

$$r^k + t_{0j}^k + p_j - c_j \leq M(1 - x_{0j}^k) \quad \forall j \in \Omega, \quad \forall k \in Q, \quad (49)$$

$$c_i \geq 0 \quad \forall i \in \overline{\Omega}, \quad (50)$$

$$x_{ij}^k \in \{0, 1\} \quad \forall i, j \in \overline{\Omega}, \forall k \in Q, \quad (51)$$

$$z_{ij} \in \{0, 1\} \quad \forall i, j \in \Omega. \quad (52)$$

The objective (36) pursued is to minimize the handling time of the vessel. It is defined by the completion time of dummy task T because every crane is forced to visit this task after processing its assigned nondummy tasks. Constraints (37) and (38) ensure that every QC starts with the initial dummy task 0 and ends up with the final dummy task T . Constraints (39) ensure that each nondummy task is processed exactly once. Constraints (40) ensure that every nondummy task has a preceding task and a succeeding task. The completion times of the tasks are computed in constraints (41) where M is again a sufficiently large positive number. Note that for $j = T$ the makespan is computed by this constraint. The precedence relations are included in constraints (42) with respect to the task completion times. Constraints (43) and (44) set the variables z_{ij} . On this basis the nonsimultaneity condition of tasks is represented in constraints (45). Constraints (46–48) ensure that cranes do not cross and that safety margins between adjacent cranes are not violated. For this purpose,

the constraints insert buffer time between the processing of two tasks that enables a safe movement and repositioning of cranes. More formally, Δ_{ij}^{vw} denotes the minimum time span to elapse between the processing of two tasks i and j , if processed by QCs v and w respectively. For all combinations of tasks $i, j \in \Omega$ and QCs $v, w \in Q$ the minimum time span to elapse between the processing of the tasks is defined as

$$\Delta_{ij}^{vw} = \begin{cases} (l_i - l_j + \delta_{vw}) \cdot \hat{t}, & \text{if } v < w \text{ and } i \neq j \\ \quad \text{and } l_i > l_j - \delta_{vw}, \\ (l_j - l_i + \delta_{vw}) \cdot \hat{t}, & \text{if } v > w \text{ and } i \neq j \\ \quad \text{and } l_i < l_j + \delta_{vw}, \\ 0, & \text{otherwise.} \end{cases} \quad (53)$$

In the first case of Equation (53), crane v is deployed at a lower quay position than crane w ($v < w$), whereas the second case considers the reverse. A positive buffer time is required if the tasks are located at bays that are too close together to be processed simultaneously by the cranes. This is determined through δ_{vw} , which is the smallest allowed difference between the bay positions of cranes v and w , that is, $\delta_{vw} = (\delta + 1) \cdot |v - w|$. Otherwise, no buffer time is required between the processing of the tasks, which is indicated through setting $\Delta_{ij}^{vw} = 0$ in the third case of Equation (53). Θ denotes the set of all combinations of tasks and QCs that potentially lead to crane interference. It can be defined as

$$\Theta = \{(i, j, v, w) \in \Omega^2 \times Q^2 \mid (i < j) \wedge (\Delta_{ij}^{vw} > 0)\}. \quad (54)$$

Owing to the symmetry of the temporal distances Δ_{ij}^{vw} the consideration can be restricted to pairs of tasks with $i < j$. Actually, a certain combination in Θ will cause interference only if its task-to-QC assignment is selected in the schedule. Since this is unknown in advance, every element of Θ must be treated by a constraint in the QCSP model. In constraints (46) those assignments of tasks to QCs are identified that are realized in the schedule. Here, $\sum_{u \in \Omega^0} x_{ui}^v = 1$ if and only if task i

is processed by QC v and $\sum_{u \in \Omega^0} x_{uj}^w = 1$ if and only if task j is processed by QC w . If both assignments take place, the left side reveals a value of two and the tasks are not allowed to be processed simultaneously, that is, either $z_{ij} = 1$ or $z_{ji} = 1$. In the case of $z_{ij} = 1$, constraints (47) insert the minimum temporal distance calculated by Equation (53) between the completion time of task i and the starting time of task j . The corresponding case of $z_{ji} = 1$ is handled in constraints (48). The ready times of QCs are handled in Equation (49) and the feasible domains of the decision variables are defined in Equations (50–52).

Discussion of the Models. The models presented approach crane scheduling in quite different ways. The model in the section titled “A Model for Quay Crane Scheduling with Container Stacks” is a pioneering one, because it incorporates crane double cycling into QC scheduling. Crane double cycling provides a cheap opportunity for decreasing vessel handling times of terminal operators. In Goodchild and Daganzo [144] the economic advantages of QC double cycling are proven in terms of increased crane productivity, berth utilization, and vessel utilization. Since the scope of the model is restricted to a single bay, it must be solved once for every bay of a vessel in order to produce work plans for the cranes. Even then, an appropriate allocation of bays to the cranes and the interaction of the cranes at a vessel have not been taken care of. These issues are the focus of QCSP models that schedule the complete handling operations of a vessel. The model in the section titled “A Model for Quay Crane Scheduling with Container Groups” enables to distribute the workload among the cranes assigned to a vessel and to schedule the tasks while considering precedence relations, crane movements, the noncrossing condition, and safety margins between cranes. Although the model considers practical issues of crane operations in high detail, it does not support crane double cycling in a schedule.

When it comes to solving the problems presented, an efficient solution method is available for the model in the section titled

“A Model for Quay Crane Scheduling with Container Stacks.” The problem formulation corresponds to a two-machine flow shop scheduling problem that can be solved to optimality using the rule of Johnson [152]. In contrast, the problem in the section titled “A Model for Quay Crane Scheduling with Container Groups” is an $\mathcal{NP}$ -hard parallel machine scheduling problem. To cope with the complexity, Bierwirth and Meisel [123] propose a branch-and-bound algorithm that searches a heuristically reduced solution space of the problem.

CONCLUSIONS

This article provides an overview of operational planning problems in seaport terminals. It describes the problems from a decision support perspective by outlining the decisions to make and the objectives that are typically pursued. For berth allocation, quay crane assignment, and quay crane scheduling, detailed literature surveys and examples of recent optimization models have been presented.

Despite the progress achieved in research on container terminal operations planning, there are several open issues that should be a subject of future research. A general need is to enhance models and solution methods to consider terminal operations in higher detail and to generate operational plans of higher quality. A particular need for research is found in a better integration of the decision fields. Many problems described above are closely interrelated. Often, the output of one planning problem serves as the input for another problem. For example, in stowage planning the positions of containers onboard a vessel are decided on and in quay crane scheduling the fastest possible way to accomplish the required loading and unloading operations is searched for. From an integration of these fields, stowage plans that support fast service operations of quay cranes can be generated.

A challenging field of research is to investigate dynamic and stochastic aspects of terminal operations. Little has been done

on robust optimization and stochastic programming in terminal operations planning although various sources of uncertainty exist, for example, from unpunctual arrivals of vessels, unknown destination ports for arriving transshipment containers, or breakdown of equipment. Inclusion of these aspects into the operational planning processes will be a valuable step toward more reliable terminal operations.

REFERENCES

1. Tozer DR, Penfold A. Ultra-large container ships (ULCS) - designing to the limit of current and projected terminal infrastructure capabilities. Proceedings of Boxship 2001 - Future Evolution of the Containership; London. 2001.
2. Goedhart GJ. Criteria for (un)-loading container ships. Technical Report. Netherlands: Technical University Delft; 2002.
3. Linn R, Liu JY, Wan YW, *et al.* Rubber tired gantry crane deployment for container yard operation. *Comput Ind Eng* 2003;45(3):429–442.
4. Jordan MA. Quay crane productivity 2002. Presented at TOC 2002 Americas; 2002 Nov 19-21; Miami (FL).
5. Brinkmann B, *et al.* Seaports - planning and design (in German). Berlin: Springer; 2005.
6. Kalmar Industries. Internet presence. 2010. Available at <http://www.kalmarind.com/>. Accessed 2010 Apr 14.
7. Shanghai Zhenhua Heavy Industry Co., Ltd. Internet presence. 2010. Available at <http://www.zpmc.com>. Accessed 2010 Apr 14.
8. Gottwald Port Technology. Internet presence. 2010. Available at <http://www.gottwald.com>. Accessed 2010 Apr 14.
9. Liftech Consultants Inc. Internet presence. 2010. Available at <http://www.jwdliftech.com>. Accessed 2010 Apr 14.
10. Noell Mobile Systems GmbH. Internet presence. 2010. Available at <http://www.noellmobilesystems.com/>. Accessed 2010 Apr 14.
11. Siemens AG. Internet presence. 2010. Available at <http://www.automation.siemens.com/>. Accessed 2010 Apr 14.
12. Meisel F, Seaside operations planning in container terminals. Berlin: Physica; 2009.

13. Imer M. Beating congestion by building capacity: an overview of new container terminal developments in Northern Europe. *Port Technol Int* 2005;PT28-06/2:1–5.
14. Wiegman BW, Rietveld P, Nijkamp P. Container terminal services and quality, Serie Research Memoranda 0040. Amsterdam Free University Amsterdam; 2001.
15. Notteboom T. The time factor in liner shipping service. *Marit Econ Log* 2006;8(1): 19–39.
16. Meersmans PJM, Dekker R. Operations research supports container handling. *Econometric Institute Report* 234. Erasmus University Rotterdam; 2001.
17. Vis IFA, de Koster R. Transshipment of containers at a container terminal: an overview. *Eur J Oper Res* 2003;147(1):1–16.
18. Steenken D, Voß S, Stahlbock R. Container terminal operation and operations research - a classification and literature review. *OR Spectrum* 2004;26(1):3–49.
19. Vacca I, Bierlaire M, Salani M. Optimization at container terminals: status, trends and perspectives. *Proceedings of the Swiss Transport Research Conference (STRC); Monte Verità/Ascona*. 2007. pp. 1–21.
20. Stahlbock R, Voß S. Operations research at container terminals: a literature update. *OR Spectrum* 2008;30(1):1–52.
21. Crainic TG, Kim KH, *et al.* *Intermodal transportation*. Amsterdam: Elsevier; 2007. pp. 467–537.
22. Kim KH, *et al.* Models and methods for operations in port container terminals. In: Langevin A, Riopel D, editors. *Logistics systems: design and optimization*, Gerad 25th anniversary series. Berlin: Springer; 2005. pp. 213–243.
23. Bierwirth C, Meisel F. A survey of berth allocation and quay crane scheduling problems in container terminals. *Eur J Oper Res* 2010;202(3):615–627.
24. Avriel M, Penn M. Exact and approximate solutions of the container ship stowage problem. *Comput Ind Eng* 1993;25(1-4):271–274.
25. Wilson ID, Roach PA. Container stowage planning: a methodology for generating computerised solutions. *J Oper Res Soc* 2000; 51(11):1248–1255.
26. Ambrosino D, Sciomachen A, Tanfani E. Stowing a containership: the master bay plan problem. *Transp Res Part A* 2004;38(2): 81–99.
27. Imai A, Sasaki K, Nishimura E, *et al.* Multi-objective simultaneous stowage and load planning for a container ship with container rehandle in yard stacks. *Eur J Oper Res* 2006;171(2):373–389.
28. Kang JG, Kim YD. Stowage planning in maritime container transportation. *J Oper Res Soc* 2002;53(4):415–426.
29. Fu Z, Li Y, Lim A, *et al.* Port space allocation with a time dimension. *J Oper Res Soc* 2007;58(6):797–807.
30. Preston P, Kozan E. An approach to determine storage locations of containers at seaport terminals. *Comput Oper Res* 2001;28(10):983–995.
31. Dekker R, Voogd P, van Asperen E. Advanced methods for container stacking. *OR Spectrum* 2006;28(4):563–586.
32. Lee Y, Hsu NY. An optimization model for the container pre-marshalling problem. *Comput Oper Res* 2007;34(11):3295–3313.
33. Kim KH, Bae JW. A heuristic rule for relocating blocks. *Comput Ind Eng* 1998;35(3-4):655–658.
34. Cheung RK, Li CL, Lin W. Interblock crane deployment in container terminals. *Transp Sci* 2002;36(1):79–93.
35. Kim KH, Hong GP. A heuristic rule for relocating blocks. *Comput Oper Res* 2006; 33(4):940–954.
36. Caserta M, Schwarze S, Voß S. A mathematical formulation for the blocks relocation problem. Working paper. University of Hamburg; 2008.
37. Jung SH, Kim KH. Load scheduling for multiple quay cranes in port container terminals. *J Intell Manuf* 2006;17(4):479–492.
38. Ng WC. Crane scheduling in container yards with inter-crane interference. *Eur J Oper Res* 2005;164(1):64–78.
39. Imai A, Nishimura E, Papadimitriou S, *et al.* The containership loading problem. *Int J Marit Econ* 2002;4(2):126–148.
40. Ambrosino D, Sciomachen A. Impact of yard organisation on the master bay planning problem. *Marit Econ Log* 2003;5(3):285–300.
41. Kim KH, Kang JS, Ryu KR. A beam search algorithm for the load sequencing of outbound containers in port container terminals. *OR Spectrum* 2004;26(1):93–116.
42. Murty KG, Liu J, Wan YW, *et al.* A decision support system for operations in a container terminal. *Decis Support Syst* 2005;39(3):309–332.

43. Murty KG, Wan YW, Liu J, *et al.* Hongkong international terminals gains elastic capacity using a data-intensive decision-support system. *Interfaces* 2005;35(1):61–75.
44. Böse J, Reiners T, Steenken D, *et al.* Vehicle dispatching at seaport container terminals using evolutionary algorithms. Volume 2, Proceedings of the 33th Hawaii International Conference on System Sciences (HICSS 33). Los Alamitos (CA): IEEE Computer Society; 2000. pp. 1–10.
45. Grunow M, Günther HO, Lehmann M. Strategies for dispatching AGVs at automated seaport container terminals. *OR Spectrum* 2006;28(4):587–610.
46. Lehmann M. Operations planning for automated transport systems in seaport container terminals (in German) [PhD thesis]. Technical University Berlin; 2006.
47. Kim CW, Tanchoco JMA, Koo PH. Deadlock prevention in manufacturing systems with AGV systems: Banker's algorithm approach. *J Manuf Sci Eng* 1997;119(4B):849–854.
48. Kim KH, Jeon SM, Ryu KR. Deadlock prevention for automated guided vehicles in automated container terminals. *OR Spectrum* 2006;28(4):659–679.
49. Lehmann M, Grunow M, Günther HO. Deadlock handling for real-time control of AGVs at automated container terminals. *OR Spectrum* 2006;28(4):631–657.
50. Kim KH, Lee KM, Hwang H. Sequencing delivery and receiving operations for yard cranes in port container terminals. *Int J Prod Econ* 2003;84(3):283–292.
51. Froyland G, Koch T, Megow N, *et al.* Optimizing the landside operation of a container terminal. *OR Spectrum* 2008;30(1):53–75.
52. Meisel F, Bierwirth C. Integration of berth allocation and crane assignment to improve the resource utilization at a seaport container terminal. In: Haasis HD, Kopfer H, Schönberger J, editors. *Operations Research Proceedings 2005*. Berlin: Springer; 2006. pp. 105–110.
53. Kim KH, Kim KW, Hwang H, *et al.* Operator-scheduling using a constraint satisfaction technique in port container terminals. *Comput Ind Eng* 2004;46(2):373–381.
54. Hartmann S. A general framework for scheduling equipment and manpower at container terminals. *OR Spectrum* 2004;26(1):51–74.
55. Lim A, Rodrigues B, Song L. Manpower allocation with time windows. *J Oper Res Soc* 2004;55(11):1178–1186.
56. Brown GG, Lawphongpanich S, Thurman KP. Optimizing ship berthing. *Nav Res Log* 1994;41(1):1–15.
57. Brown GG, Cormican KJ, Lawphongpanich S, *et al.* Optimizing submarine berthing with a persistence incentive. *Nav Res Log* 1997;44(4):301–318.
58. Lee Y, Chen CY. An optimization heuristic for the berth scheduling problem. *Eur J Oper Res* 2008;196(2):500–508.
59. Imai A, Zhang JT, Nishimura E, *et al.* The berth allocation problem with service time and delay time objectives. *Marit Econ Log* 2007;9(4):269–290.
60. Monaco MF, Sammarra M. The berth allocation problem: a strong formulation solved by a lagrangean approach. *Transp Sci* 2007;41(2):265–280.
61. Cordeau JF, Laporte G, Legato P, *et al.* Models and tabu search heuristics for the berth-allocation problem. *Transp Sci* 2005;39(4):526–538.
62. Park YM, Kim KH. A scheduling method for berth and quay cranes. *OR Spectrum* 2003;25(1):1–23.
63. Chu CY, Huang WC. Aggregates cranes handling capacity of container terminals: the port of kaohsiung. *Marit Policy Manage* 2002;29(4):341–350.
64. Legato P, Gulli D, Trunfio R. The quay crane deployment problem at a maritime container terminal. Proceedings of the 22th European Conference on Modelling and Simulation (ECMS 2008). Nicosia. 2008. pp. 53–59.
65. Ak A, Erera AL. Simultaneous berth and quay crane scheduling for container ports. Working paper. Atlanta: H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology; 2006.
66. Imai A, Sun X, Nishimura E, *et al.* Berth allocation in a container port: using a continuous location space approach. *Transp Res Part B* 2005;39(3):199–221.
67. Imai A, Nishimura E, Papadimitriou S. The dynamic berth allocation problem for a container port. *Transp Res Part B* 2001;35(4):401–417.
68. Imai A, Nagaiwa K, Tat CW. Efficient planning of berth allocation for container terminals in Asia. *J Adv Transp* 1997;31(1):75–94.
69. Hansen P, Oğuz C. A note on formulations of static and dynamic berth allocation problems. *Les Cahiers du GERAD* 2003;30:1–17.

70. Lee DH, Song L, Wang H. Bilevel programming model and solutions of berth allocation and quay crane scheduling. Proceedings of 85th Annual Meeting of Transportation Research Board (CD-ROM), Annual Meeting of Transportation Research Board; Washington (DC). 2006.
71. Imai A, Nishimura E, Papadimitriou S. Berth allocation with service priority. *Transp Res Part B* 2003;37(5):437–457.
72. Theofanis S, Boile M, Golias M. An optimization based genetic algorithm heuristic for the berth allocation problem. *IEEE Congress on Evolutionary Computation 2007 (CEC 2007)*. Washington (DC): IEEE Computer Society; 2007. pp. 4439–4445.
73. Mauri GR, Oliveira ACM, Lorena LAN, *et al.* A hybrid column generation approach for the berth allocation problem. In: van Hemert J, Cotta C, editors. Volume 4972, 8th European Conference on Evolutionary Computation in Combinatorial Optimisation (EvoCOP 2008), LNCS. Berlin: Springer; 2008. pp. 110–122.
74. Han M, Li P, Sun J. The algorithm for berth scheduling problem by the hybrid optimization strategy GASA. Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision (ICARCV'06). Washington (DC): IEEE Computer Society; 2006. pp. 1–4.
75. Imai A, Chen HC, Nishimura E, *et al.* The simultaneous berth and quay crane allocation problem. *Transp Res Part E* 2008;44(5):900–920.
76. Zhou P, Kang H, Lin L. A dynamic berth allocation model based on stochastic consideration. Volume 2, Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA 2006). Washington (DC): IEEE Computer Society; 2006. pp. 7297–7301.
77. Golias M, Boile M, Theofanis S. The berth allocation problem: a formulation reflecting time window service deadlines. Proceedings of the 48th Transportation Research Forum Annual Meeting. Boston (MA): Transportation Research Forum; 2006.
78. Liang C, Huang Y, Yang Y. A quay crane dynamic scheduling problem by hybrid evolutionary algorithm for berth allocation planning. *Comput Ind Eng* 2008;56(3):1021–1028.
79. Imai A, Nishimura E, Papadimitriou S. Berthing ships at a multi-user container terminal with a limited quay capacity. *Transp Res Part E* 2008;44(1):136–151.
80. Giallombardo G, Moccia L, Salani M, *et al.* The tactical berth allocation problem with quay crane assignment and transshipment-related quadratic yard costs. Proceedings of the European Transport Conference (ETC). Noordwijkerhout. 2008. pp. 1–27.
81. Giallombardo G, Moccia L, Salani M, *et al.* Modeling and solving the tactical berth allocation problem. *Transp Res Part B* 2010;44(2):232–245.
82. Hansen P, Oğuz C, Mladenovic N. Variable neighborhood search for minimum cost berth allocation. *Eur J Oper Res* 2008;191(3):636–649.
83. Li CL, Cai X, Lee CY. Scheduling with multiple-job-on-one-processor pattern. *IIE Trans* 1998;30(5):433–445.
84. Guan Y, Xiao WQ, Cheung RK, *et al.* A multiprocessor task scheduling model for berth allocation: heuristic and worst-case analysis. *Oper Res Lett* 2002;30(5):343–350.
85. Rashidi H. Dynamic scheduling of automated guided vehicles [PhD thesis]. Colchester: University of Essex; 2006.
86. Oğuz C, Błazewicz J, Cheng TCE, *et al.* Berth allocation as a moldable task scheduling problem. Proceedings of the 9th International Workshop on Project Management and Scheduling (PMS 2004); Nancy. 2004. pp. 201–205.
87. Guan Y, Cheung RK. The berth allocation problem: models and solution methods. *OR Spectrum* 2004;26(1):75–92.
88. Wang F, Lim A. A stochastic beam search for the berth allocation problem. *Decis Support Syst* 2007;42(4):2186–2196.
89. Moon K. A mathematical model and a heuristic algorithm for berth planning [PhD thesis]. Pusan: Pusan National University; 2000.
90. Park KT, Kim KH. Berth scheduling for container terminals by using a sub-gradient optimization technique. *J Oper Res Soc* 2002;53(9):1054–1062.
91. Kim KH, Moon KC. Berth scheduling by simulated annealing. *Transp Res Part B* 2003;37(6):541–560.
92. Briano C, Briano E, Bruzzone AG. Models for support maritime logistics: a case study for improving terminal planning. In: Merkuryev Y, Zobel R, Kerckhoffs E, editors. Proceedings of the 19th European Conference on Modelling and Simulation (ECMS). Riga. 2005. pp. 199–203.

93. Lim A. The berth planning problem. *Oper Res Lett* 1998;22(2):105–110.
94. Lim A. An effective ship berthing algorithm. In: Thomas D, editors. *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99-Vol-1)*. San Francisco (CA): Morgan Kaufmann Publishers; 1999. pp. 594–599.
95. Tong CJ, Lau HC, Lim A, *et al.* Ant colony optimization for the ship berthing problems. In: Thiagarajan PS, Yap R, editors. *Volume 1742, 5th Asian Computing Science Conference (ASIAN'99)*, LNCS. Berlin: Springer; 1999. pp. 359–370.
96. Goh KS, Lim A. Combining various algorithms to solve the ship berthing problem. *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)*. Los Alamitos (CA): IEEE Computer Society; 2000. pp. 370–373.
97. Chang D, Yan W, Chen CH, *et al.* A berth allocation strategy using heuristics algorithm and simulation optimisation. *Int J Comput Appl Technol* 2008;32(4):272–281.
98. Meisel F, Bierwirth C. Heuristics for the integration of crane productivity in the berth allocation problem. *Transp Res Part E* 2009;45(1):196–209.
99. Theofanis S, Golias M, Boile M. Berth and quay crane scheduling: a formulation reflecting service deadlines and productivity agreements. *Proceedings of the International Conference on Transport Science and Technology (TRANSTEC 2007)*. Prague: Czech Technical University; 2007. pp. 124–140.
100. Liu J, Wan YW, Wang L. Quay crane scheduling at container terminals to minimize the maximum relative tardiness of vessel departures. *Nav Res Log* 2006;53(1):60–74.
101. Meier L, Schumann R. Coordination of interdependent planning systems, a case study. In: Koschke R, Otthein H, Rödiger KH, *et al.*, editors. *Lecture Notes in Informatics (LNI) P-109*. Bonn: Köllen Druck+Verlag GmbH; 2007. pp. 389–396.
102. Hendriks MPM, Laumanns M, Lefebvre E, *et al.* Robust periodic berth planning of container vessels. In: Kopfer H, Günther HO, Kim KH, editors. *Proceedings of the 3rd German-Korean Workshop on Container Terminal Management: IT-based Planning and Control of Seaport Container Terminals and Transportation Systems*. Bremen. 2008. pp. 1–13.
103. Meisel F, Bierwirth C. The berth allocation problem with a cut-and-run option. In: Fleischmann B, Borgwardt KH, Klein R, *et al.*, editors. *Operations Research Proceedings 2008*. Berlin: Springer; 2009. pp. 283–288.
104. Zhang C, Zheng L, Zhang Z, *et al.* The allocation of berths and quay cranes by using a sub-gradient optimization technique. *Comput Ind Eng* 2010;58(1):40–50.
105. Moorthy R, Teo CP. Berth management in container terminal: The template design problem. *OR Spectrum* 2006;28(4):495–518.
106. Dai J, Lin W, Moorthy R, *et al.* Berth allocation planning optimization in container terminals. In: Tang CS, Teo CP, Wei KK, editors. *Supply chain analysis: a handbook on the interaction of information, system and optimization*. New York: Springer; 2008. pp. 69–105.
107. Chen CY, Hsieh TW. A time-space network model for the berth allocation problem 1999. 19th IFIP TC7 Conference on System Modeling and Optimization; Cambridge. 1999.
108. Imai A, Nishimura E, Hattori M, *et al.* Berth allocation at indented berths for mega-containerships. *Eur J Oper Res* 2007;179(2):579–593.
109. Nishimura E, Imai A, Papadimitriou S. Berth allocation planning in the public berth system by genetic algorithms. *Eur J Oper Res* 2001;131(2):282–292.
110. Cheong CY, Tan KC, Liu DK, *et al.* Multi-objective and prioritized berth allocation in container ports. *Ann Oper Res* 2010;1–41. In press. DOI: 10.1007/s10479-008-0493-0.
111. Hoffarth L, Voß S. Berth allocation in a container terminal - development of a decision support system (in German). In: Dyckhoff H, Derigs U, Salomon M, *et al.*, editors. *Operations Research Proceedings 1993*. Berlin: Springer; 1994. pp. 89–95.
112. Lokuge P, Alahakoon D. Improving the adaptability in automated vessel scheduling in container ports using intelligent software agents. *Eur J Oper Res* 2007;177(3):1985–2015.
113. Schonfeld P, Sharafeldien O. Optimal berth and crane combinations in container-ports. *J Waterway Port Coast Ocean Eng* 1985;111(6):1060–1072.
114. Silberholz MB, Golden BL, Baker EK. Using simulation to study the impact of work rules on productivity at marine container terminals. *Comput Oper Res* 1991;18(5):433–452.

115. Dragovic B, Park NK, Radmilovic Z. Ship-berth link performance evaluation: simulation and analytical approaches. *Marit Policy Manage* 2006;33(3):281–299.
116. Winter T. Online and real-time dispatching problems [PhD thesis]. Technical University Braunschweig; 1999.
117. Lim A, Rodrigues B, Xu Z. A M-parallel crane scheduling problem with a non-crossing constraint. *Nav Res Log* 2007;54(2): 115–127.
118. Goodchild AV, Daganzo CF. Reducing ship turn-around time using double-cycling. Research report UCB-ITS-RR-2004-4. Berkeley: University of California; 2004.
119. Kim KH, Park YM. A crane scheduling method for port container terminals. *Eur J Oper Res* 2004;156(3):752–768.
120. Meisel F, Wichmann M. Container sequencing for quay cranes with internal reshuffles. *OR Spectrum* 2010;32(3):569–591.
121. Steenken D, Winter T, Zimmermann UT. Stowage and transport optimization in ship planning. In: Grötschel M, Krumke S, Rambau J, *et al.* editors. Online optimization of large scale systems. Berlin: Springer; 2001. pp. 731–745.
122. Lim A, Rodrigues B, Xiao F, *et al.* Crane scheduling using tabu search. Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02). Washington (DC): IEEE Computer Society; 2002. pp. 146–153.
123. Bierwirth C, Meisel F. A fast heuristic for quay crane scheduling with interference constraints. *J Schedul* 2009;12(4):345–360.
124. Lim A, Rodrigues B, Xiao F, *et al.* Crane scheduling with spatial constraints. *Nav Res Log* 2004;51(3):386–406.
125. Daganzo CF. The crane scheduling problem. *Transp Res Part B* 1989;23(3):159–175.
126. Peterkofsky RI, Daganzo CF. A branch and bound solution method for the crane scheduling problem. *Transp Res Part B* 1990;24(3):159–172.
127. Lim A, Rodrigues B, Xu Z. Approximation schemes for the crane scheduling problem. In: Hagerup T, Katajainen J, *et al.* editors. Volume 3111, 9th Scandinavian Workshop on Algorithm Theory (SWAT 2004), LNCS. Berlin: Springer; 2004. pp. 323–335.
128. Lim A, Rodrigues B, Xu Z. Solving the crane scheduling problem using intelligent search schemes (extended abstract). In: Wallace M, *et al.* editors. Volume 3258, 10th International Conference on Principles and Practice of Constraint Programming (CP 2004), LNCS. Berlin: Springer; 2004. pp. 747–751.
129. Zhu Y, Lim A. Crane scheduling with non-crossing constraint. *J Oper Res Soc* 2006;57(12):1464–1471.
130. Lee DH, Wang HQ, Miao L. An approximation algorithm for quay crane scheduling with non-interference constraints in port container terminals. Presented at Tristan VI; 2007 June 10-15; Phuket.
131. Lee DH, Wang HQ, Miao L. Quay crane scheduling with non-interference constraints in port container terminals. *Transp Res Part E* 2008;44(1):124–135.
132. Lee DH, Wang HQ, Miao L. Quay crane scheduling with handling priority in port container terminals. *Eng Optim* 2008;40(2):179–189.
133. Choo S, Klabjan D, Simchi-Levi D. Multiship crane sequencing with yard congestion constraints. *Transp Sci* 2010;44(1): 98–115.
134. Moccia L, Cordeau JF, Gaudioso M, *et al.* A branch-and-cut algorithm for the quay crane scheduling problem in a container terminal. *Nav Res Log* 2006;53(1): 45–59.
135. Sammarra M, Cordeau JF, Laporte G, *et al.* A tabu search heuristic for the quay crane scheduling problem. *J Schedul* 2007;10(4-5):327–336.
136. Tavakkoli-Moghaddam R, Makui A, Salahi S, *et al.* An efficient algorithm for solving a new mathematical model for a quay crane scheduling problem in container ports. *Comput Ind Eng* 2009;56(1):241–248.
137. Ng WC, Mak KL. Quay crane scheduling in container terminals. *Eng Optim* 2006;38(6):723–737.
138. Jung DH, Park YM, Lee BK, *et al.* A quay crane scheduling method considering interference of yard cranes in container terminals. In: Gelbukh AF, García CAR, editors. VOLUME 4293, Proceedings of the 5th Mexican International Conference on Artificial Intelligence (MICAI 2006), LNCS. Berlin: Springer; 2006. pp. 461–471.
139. Tavakkoli-Moghaddam R, Taheri F, Bazzazi M, *et al.* Two efficient algorithms for a quay crane scheduling and assignment problem. *J Intell Manuf* 2010;1–14. In press. DOI: 10.1007/s10845-010-0381-8.

140. Wang Y, Kim KH. A quay crane scheduling algorithm considering the workload of yard cranes in a container yard. *J Intell Manuf* 2010;1–12. In press. DOI: 10.1007/s10845-009-0303-9.
141. Goodchild AV, Daganzo CF. Crane double cycling in container ports: affect on ship dwell time. Research report RR20055. Berkeley: University of California; 2005.
142. Goodchild AV, Daganzo CF. Performance comparison of crane double-cycling strategies 2005. Working paper UCB-ITS-WP-2005-2. Berkeley: University of California.
143. Goodchild AV, Daganzo CF. Double-cycling strategies for container ships and their effect on ship loading and unloading operations. *Transp Sci* 2006;40(4):473–483.
144. Goodchild AV, Daganzo CF. Crane double cycling in container ports: planning methods and evaluation. *Transp Res Part B* 2007;41(8):875–891.
145. Goodchild AV. Port planning for double cycling crane operations. Proceedings of the 85th Annual Meeting of Transportation Research Board (CD-ROM), Annual Meeting of Transportation Research Board; Washington (DC). 2006.
146. Zhang H, Kim KH. Maximizing the number of dual-cycle operations of quay cranes in container terminals. *Comput Ind Eng* 2009;56(3):979–992.
147. Gambardella LM, Mastrolilli M, Rizzoli AE, *et al.* An optimization methodology for inter-modal terminal management. *J Intell Manuf* 2001;12(5):521–534.
148. Bish EK. A multiple-crane-constrained scheduling problem in a container terminal. *Eur J Oper Res* 2003;144(1):83–107.
149. Lee YH, Kang KR, Ryu J, *et al.* Optimization of container load sequencing by a hybrid of ant colony optimization and tabu search. *Lect Notes Comput Sci* 2005;3611:1259–1268.
150. Chen L, Bostel N, Dejax P, *et al.* A tabu search algorithm for the integrated scheduling problem of container handling systems in a maritime terminal. *Eur J Oper Res* 2007;181(1):40–58.
151. Canonaco P, Legato P, Mazza RM, *et al.* A queuing network model for the management of berth crane operations. *Comput Oper Res* 2008;35(8):2432–2446.
152. Johnson S. Optimal two- and three-stage production schedules with setup times. *Nav Res Log Q* 1954;1(1):61–68.

SCORING RULES

ROBERT L. WINKLER
Fuqua School of Business, Duke
University, Durham, North
Carolina

VICTOR RICHMOND R. JOSE
McDonough School of Business,
Georgetown University,
Washington, D.C.

INTRODUCTION

Uncertainty is a pervasive feature of our world, and fields such as decision analysis and statistics provide methods to help us make decisions, forecasts, and inferences in the face of uncertainty. Our everyday language includes many terms that relate to the degree of uncertainty in a situation: for example, rain is unlikely today, the chances are good that a surgical procedure will be successful, the prospects for an improved economic situation are not favorable, and so on. As the mathematical language of uncertainty, probability theory provides a structure to quantify uncertainty. Probabilities are encountered in the media (e.g., the probability of rain this afternoon is 20%) and widely used in modeling.

Although probability forecasts are formulated and used extensively, very often they are never evaluated after the event or variable of interest is observed. Scoring rules provide such evaluations by giving a numerical score based on the probabilities and on the actual observation. For example, a probability of rain of 40% in a simple two-state setting of rain versus no rain will receive a higher score than a probability of 20% if it rains, and a lower score if it does not rain. In this manner, we can use scoring rules to compare the sources of the probabilities, which might be experts, models, or simply past data. The first scoring rule used on a regular basis was a quadratic rule developed

by Brier [1] to evaluate probabilistic weather forecasts. Indeed, weather forecasting is the area in which scoring rules have been used most extensively.

The presence of such an *ex post* evaluation using suitably designed scoring rules also provides *ex ante* incentives for careful formulation of probability forecasts. Much of the early development of scoring rules emphasized this *ex ante* role of scoring rules [e.g., [2–6]]. Attention was focused on strictly proper scoring rules, for which a forecaster can maximize his or her expected score only by honestly reporting the probabilities and also has the incentive to obtain further information to increase the accuracy of the probabilities. This *ex ante* motivation yields rules that reward probabilities that have good characteristics *ex post*, as we shall see. For a general discussion of scoring rules and reviews of the scoring rule literature, see Winkler [7] and Gneiting and Raftery [8].

We discuss some basic properties of scoring rules in the second section, focusing on the aspects related to *ex ante* incentives, and present some commonly encountered rules. In the next section, we turn to *ex post* evaluation, showing how some notions involving strictly proper scoring rules relate to *ex post* evaluation. The next two sections involve scoring rules with special characteristics, namely those that provide evaluations of probabilities relative to baseline distributions and those that take into account any ordering of the events of interest. A brief summary and discussion, including some connections with other fields, is presented in the final section.

STRICTLY PROPER SCORING RULES

We begin by considering the simplest possible situation, that of a single event A and its complement. Suppose that an expert is assessing a probability for A and is being evaluated with a scoring rule S . If a probability r is reported for A , then the score will be

$S(r, e)$, where $e = 1$ if A occurs and $e = 0$ if A does not occur. Furthermore, assume that the expert's best judgment is that the probability of A is denoted by p . Then the expected score is $S(r, p) = pS(r, 1) + (1 - p)S(r, 0)$. The scoring rule S is said to be *strictly proper* if

$$S(p, p) > S(r, p) \text{ for any } r \neq p. \quad (1)$$

To maximize the expected score with a strictly proper rule, the expert should set $r = p$, thereby reporting the probability honestly. The scoring rules discussed here are oriented such that a higher score is better. Some rules in the literature, such as the Brier score [1], are oriented with a negative score being better, in which case the expert should set $r = p$ to minimize the expected score. Rules such as the Brier score can be converted to a positive orientation by changing the sign of the score, so the focus on scores with a positive orientation here is not restrictive.

The expected score $S(p, p)$ for honest reporting from a strictly proper scoring rule is strictly convex, and conversely, a strictly proper scoring rule can be generated from any strictly convex function of p that is taken as the expected score function $S(p, p)$ for honest reporting [6]. Thus, there are an infinite number of rules satisfying Equation (1). Three commonly used rules are as follows:

Quadratic:

$$S(r, e) = 1 - 2(e - r)^2, \quad (2)$$

Logarithmic:

$$S(r, e) = \log[re + (1 - r)(1 - e)], \quad (3)$$

Spherical:

$$S(r, e) = [re + (1 - r)(1 - e)] / [r^2 + (1 - r)^2]^{1/2}. \quad (4)$$

These and any other strictly proper rules can be scaled as desired (e.g., to avoid negative scores), because any positive affine transformation of a strictly proper rule is itself strictly proper. Figure 1 shows $S(r, 1)$, $S(r, 0)$, and $S(p, p)$ for the quadratic scoring rule. Note that for this simple two-event setting, $S(r, 1)$ and $S(r, 0)$ are mirror images of each

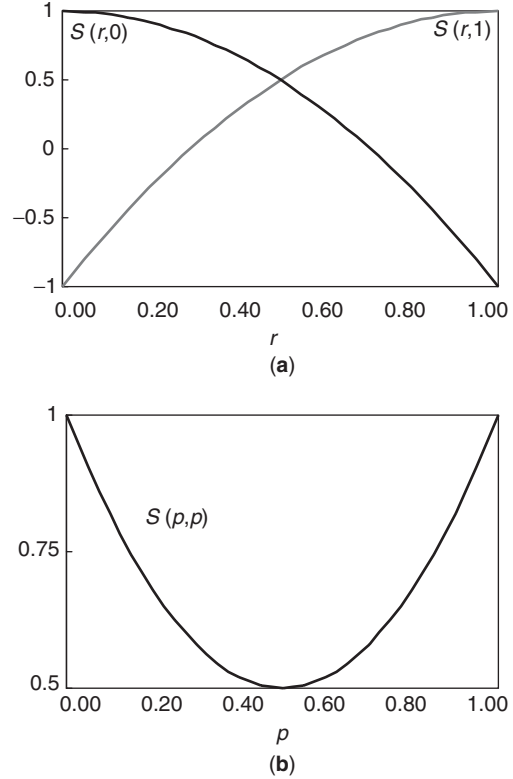


Figure 1. (a) Score functions $S(r, 1)$ and $S(r, 0)$ and (b) expected score $S(p, p)$ under honest reporting for the quadratic scoring rule.

other, with $S(r, 1)$ increasing in r and $S(r, 0)$ decreasing in r .

With the general concept of a strictly proper scoring rule established for the case of a single event (and its complement), we next generalize to the case of a set of mutually exclusive and exhaustive events $\{A_1, \dots, A_k\}$, for which the expert's probabilities are given by the vector $\mathbf{p} = (p_1, \dots, p_k)$ and the reported probabilities are $\mathbf{r} = (r_1, \dots, r_k)$. With a scoring rule S , the expert's score is $S(\mathbf{r}, \mathbf{e}_i)$ if A_i occurs, where $\mathbf{e}_i$ is a vector with the i th element equal to 1 and the other elements all equal to 0. The expected score from the perspective of the expert is $S(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^k p_i S(\mathbf{r}, \mathbf{e}_i)$, and the scoring rule is strictly proper if $S(\mathbf{p}, \mathbf{p}) > S(\mathbf{r}, \mathbf{p})$ for any $\mathbf{r} \neq \mathbf{p}$. Quadratic, logarithmic, and spherical

rules for this case are

$$S(\mathbf{r}, \mathbf{e}_i) = 2r_i - \sum_{j=1}^k r_j^2, \quad (5)$$

$$S(\mathbf{r}, \mathbf{e}_i) = \log r_i, \quad (6)$$

$$\text{and } S(\mathbf{r}, \mathbf{e}_i) = r_i / \left(\sum_{j=1}^k r_j^2 \right)^{1/2}, \quad (7)$$

respectively. Note that this setup could be used when we are considering a discrete distribution of a random variable, which could include a discretization of a continuous random variable into a set of intervals.

Finally, we present scoring rules for probability distributions of a continuous random variable $\tilde{x}$. Let p denote the expert's probability density function for $\tilde{x}$, and let r denote the corresponding reported density function. Then, for a scoring rule that gives a score $S(r, x)$ when $\tilde{x} = x$, the expert's expected score is $S(r, p) = \int_{-\infty}^{\infty} S(r, x)p(x)dx$. Quadratic, logarithmic, and spherical scoring rules in the continuous case are

$$S(r, x) = 2r(x) - \int_{-\infty}^{\infty} r^2(x) dx, \quad (8)$$

$$S(r, x) = \log r(x), \quad (9)$$

$$\text{and } S(r, x) = r(x) / \left(\int_{-\infty}^{\infty} r^2(x) dx \right)^{1/2}. \quad (10)$$

Our focus has been on strictly proper scoring rules, which have been developed with the goal of providing the expert with an incentive to report honestly. But if the expert is not well-informed with respect to the situation, reporting probabilities honestly may not mean reporting "good" probabilities. Not all probability forecasts are necessarily "good" forecasts. Fortunately, assuming honest forecasting, strictly proper scoring rules will reward forecasts by providing higher expected scores to forecasts for which p is closer to 0 or 1. To see how this works for the quadratic, logarithmic, and spherical rules that have been presented here, we note that they share an important characteristic: they are symmetric. In the single-event case, that means that the expected score

for an honestly reported probability of r is the same as the expected score for an honestly reported probability of $1-r$. As noted earlier, reporting $r = p$ under a strictly proper scoring rule results in an expected score function $S(p, p)$ that is strictly convex in p . This convexity, combined with the symmetry, means that $S(p, p)$ is minimized at $p = 0.5$ and increases as $p \rightarrow 0$ or $p \rightarrow 1$ from $p = 0.5$. These features of $S(p, p)$ are illustrated for the quadratic rule in Fig. 1. That means that under honest reporting, the expected score is higher for probability forecasts that are sharper, where sharpness refers to the degree that p is closer to 0 or 1. For example, a probability of 0 or 1 is perfectly sharp, whereas a probability of 0.5 admits a lot of uncertainty about the outcome.

To illustrate the incentives from strictly proper scoring rules for both honesty and sharpness, consider a decomposition of the expected score for the quadratic rule in the case of a probability for a single event. The expected score is $S(r, p) = p[1 - 2(1 - r)^2] + (1 - p)(1 - 2r^2)$. Expanding, adding and subtracting p^2 , and rearranging yields

$$S(r, p) = 1 - 2(p - r)^2 - 2p(1 - p). \quad (11)$$

The second term on the right-hand side of Equation (11) can be viewed as a penalty (because of the negative sign) for not setting $r = p$, and it thus provides an incentive for honesty. The last term is a penalty for lack of sharpness, because $p(1 - p)$ is maximized at $p = 0.5$ and decreases as $p \rightarrow 0$ or $p \rightarrow 1$. The best possible expected score is one, and dishonesty ($r \neq p$) or lack of perfect sharpness ($0 < p < 1$) will reduce the expected score. Other strictly proper rules (e.g., the logarithmic and spherical rules) can be decomposed in a similar manner.

Keep in mind that to maximize expected score, the expert has to report honestly, so attempting to have probabilities look sharp artificially (i.e., sharp reported probabilities that are not consistent with the expert's judgments) will decrease the expected score, not increase it. Note from Equation (11) that the sharpness term relates to the sharpness of p , not the sharpness of r . The primary

aspect of strictly proper scoring rules is to encourage honesty, and the nature of strictly proper scoring rules is such that honest reporting by experts who have sharper probabilities will yield higher expected scores than honest reporting by experts who have probabilities that are not so sharp. In the final analysis, then, strictly proper scoring rules reward both honesty and sharpness.

Strictly proper scoring rules differ in some characteristics. We will present different types of strictly proper rules in later sections and discuss some of their characteristics. As we shall see, not all strictly proper rules are symmetric in the sense discussed above, and in some cases it may be desirable to use a rule that is not symmetric. Another characteristic of note is that the logarithmic rule is the only rule for which the score depends only on the probability or density that has been assigned to the event or value of the variable that actually occurs. It does not depend on the probability or density assigned to other events or values. For example, if we consider scoring rules for k mutually exclusive and exhaustive events and event A_i occurs, the logarithmic score, $\log r_i$, depends only on r_i , whereas the quadratic score, $2r_i - \sum_{j=1}^k r_j^2$, depends on all of the probabilities $r_1, \dots, r_k$. This property, unique to the logarithmic rule when $k > 2$, is called *locality*, and it is consistent in spirit with the likelihood principle that plays a major role in statistics.

Although the quadratic, logarithmic, and spherical rules given above are the usual suspects when we think about scoring rules, they are special cases of two rich families of strictly proper scoring rules [9]. When probabilities for a set of k mutually exclusive and exhaustive events are reported, scores for the pseudospherical and power families are given by

$$S_\beta^S(\mathbf{r}, \mathbf{e}_i) = \frac{1}{\beta - 1} \left[\left(\frac{r_i}{E_{\mathbf{r}}(\mathbf{r}^{\beta-1})^{1/\beta}} \right)^{\beta-1} - 1 \right] \quad (12)$$

and

$$S_\beta^P(\mathbf{r}, \mathbf{e}_i) = \frac{r_i^{\beta-1} - 1}{\beta - 1} - \frac{E_{\mathbf{r}}(\mathbf{r}^{\beta-1}) - 1}{\beta}, \quad (13)$$

respectively, where $E_{\mathbf{r}}(\mathbf{r}^{\beta-1}) = \sum_{i=1}^k r_i(r_i^{\beta-1})$ and $-\infty < \beta < \infty$. When $\beta = 2$, Equations (12) and (13) yield the spherical and quadratic rules, respectively. When $\beta \rightarrow 1$, both families converge to the logarithmic rule.

In summary, the most important characteristic of scoring rules in an *ex ante* sense related to probability assessment is that they should be strictly proper, and there are many such rules from which to choose. If a rule is indeed strictly proper, then it should provide incentives for an expert to honestly report probabilities and to invest effort in an attempt to make those probabilities sharper. Such effort might be directed toward such things as gathering more data, using more powerful methods to analyze the data, and learning more about the processes affecting the events in question or about forecasts provided by others.

EX-POST EVALUATION WITH STRICTLY PROPER SCORING RULES

We shift now from the *ex ante* perspective of the previous section to consider the *ex post* evaluation of probability forecasts. For a given situation, we assume that an expert has probabilities that are based on the information available and consistent with the expert's best judgment. The *ex ante* viewpoint involves rules that provide incentives for the expert to attempt to come up with "good" probabilities and to report those probabilities honestly. Many of the characteristics discussed in the preceding section have counterparts when we consider their use for evaluation purposes.

As is the case with statistical analysis of data in general, a single observation is not very informative. To have a reliable evaluation (when comparing experts or models in terms of their probabilities, for example), we would like to have a large number of observations. Thus, instead of considering a given situation with a single probability or set of probabilities, we have a set of data consisting of many situations with different probabilities.

Suppose that we have a sample of probability forecasts and the corresponding observations for the occurrence of an event. For example, this might consist of probabilities of default for loans (generated by a model or assessed by a bank officer) or probabilities of rain for different days at a given location. First, we can look at all of the occasions in the data set for which a particular value of the reported probability r (say, 0.30) was used, and determine the relative frequency of occurrence of the event of interest on those occasions. Denote this relative frequency by f_r . With the quadratic scoring rule, the average score on all of the occasions with this value of r is $\bar{S}(r, f_r) = f_r[1 - 2(1 - r)^2] + (1 - f_r)(1 - 2r^2)$. *Ex ante*, the expected score from the perspective of the expert is a function of r and p , with p being known only to the expert. *Ex post*, the average score is a function of r and f_r , where we are able to observe f_r :

$$\bar{S}(r, f_r) = 1 - 2(f_r - r)^2 - 2f_r(1 - f_r). \quad (14)$$

Note that this is simply Equation (11) with f_r used in place of p .

The second term on the right-hand side of Equation (14) is a measure of *calibration*, which involves the correspondence between the reported probability and the relative frequency of occurrence of the event when that probability is used. If $f_r = r$, then the reported probabilities of r are perfectly calibrated. The more the relative frequency deviates from r , the worse the calibration is. The last term on the right-hand side of Equation (14) is a measure of *sharpness*, which is better as $f_r \rightarrow 0$ or $f_r \rightarrow 1$. Poorer calibration and less sharpness lead to lower average scores.

In data with reported probabilities and outcomes, different probabilities will be used on different occasions. The overall average score $\bar{S}$ for an expert is found by aggregating the average scores for the different values of r . If we let n_r represent the number of times a reported probability of r is used in the data set and let $n = \sum_r n_r$ represent the overall sample size, the overall average score can be

expressed as

$$\begin{aligned} \bar{S} &= \sum_r (n_r/n) \bar{S}(r, f_r) \\ &= 1 - 2 \sum_r (n_r/n) (f_r - r)^2 \\ &\quad - 2 \sum_r (n_r/n) f_r (1 - f_r). \end{aligned} \quad (15)$$

The first summation on the right-hand side of Equation (15) is an overall measure of calibration for the data set, and the second summation is an overall measure of sharpness. This decomposition into calibration and sharpness components can be generalized beyond the quadratic rule to any strictly proper scoring rule [10].

One convenient way to think about calibration and sharpness is to think of the probability assessment process for a single event as a two-step process. First, the expert puts forecast situations in equivalence classes, or “boxes,” such that the expert feels that the events in a given box have roughly the same probability of occurrence. Second, the expert assigns numbers (probability values) to the boxes. Calibration is then an evaluation of how well the expert assigns the numbers. Sharpness, on the other hand, is unrelated to the probability values. Instead, it measures how effective the expert is in creating boxes for which the relative frequency of occurrence of the events is close to 0 or 1. This is unlikely to be the way an expert really thinks about the forecasting process, but it is a convenient way to emphasize key differences between calibration and sharpness. For one thing, we can always attempt to correct for miscalibration. If an expert always gives probabilities that are too high, for example, a decision maker using those probabilities can reduce reported probabilities from that expert. Correcting for poor sharpness is a much trickier business.

We have illustrated the notion of decomposition of an average score using the quadratic scoring rule, but the same idea can be applied to other rules. Also, an *ex post* average score or an *ex ante* expected score can be decomposed in different ways. The decomposition into terms measuring calibration and sharpness is the most frequently

used decomposition, and it is arguably the most important decomposition. Gneiting and Raftery [8] comment that “the goal of probabilistic forecasting is to maximize the sharpness of the (probabilities) subject to calibration.” Although that seems reasonable, we feel that in reported evaluations, too much emphasis is typically given to calibration and not enough to sharpness.

Ex post evaluation with strictly proper scoring rules involves most of the same ideas encountered in the role of scoring rules in terms of *ex ante* incentives. *Ex ante* incentives for honesty translate into *ex post* evaluations of calibration, and *ex ante* measures of sharpness based on the probabilities translate into *ex post* measures based on the relative frequencies of occurrence of the events of interest for given probability values (given boxes). The use of scoring rules and their decompositions to evaluate probabilities *ex post* can be thought of as exploratory data analysis. Such evaluations can be used to compare experts or models. In the case of a single expert or model, they can be used to learn more about that expert’s or model’s characteristics and abilities as a probability forecaster. Feedback can also help the expert understand his or her own characteristics as a probability forecaster and attempt to improve them in the future.

SCORING RULES WITH BASELINE DISTRIBUTIONS

Scoring rules such as the quadratic, logarithmic, and spherical rules given earlier can be thought of as providing an “absolute” evaluation of probabilities. Often, we would like to have a relative evaluation by comparing how good a probability or probability distribution is, relative to some baseline. When assessing probabilities of rain, for example, it is easier to get a high score in a location where rain seldom occurs than it is in a location where it rains reasonably often. Does that mean that the probability forecasts in the drier area are better?

As noted earlier, the most commonly used scoring rules are symmetric in the sense that any permutation of the labels on the events

and their associated probabilities does not change the expected score. One implication of this symmetry, when combined with the convexity of the expected score function, is that the expected score is minimized for a uniform distribution that gives a probability of $1/k$ to each event in the k -event case. Thus, these scores are implicitly being evaluated relative to a uniform baseline distribution.

To avoid comparison with a uniform distribution, we could consider the percentage improvement in average scores over the scores for a baseline distribution. In assessing a probability of rain, for example, we might use climatology, which is the long-term relative frequency of rain in a given location at a specific time of year, as a baseline. However, a percentage improvement in the score over the baseline, which is called a skill score, is not strictly proper. For a strictly proper rule with a baseline distribution that is not uniform, we can choose a desired convex expected score function and generate a strictly proper rule that yields that expected score function [11]. For example, we might choose a function that is minimized at what might be viewed as a “least skillful” forecast. In forecasting rain, climatology might be considered “least skillful” among forecasts that seem reasonable, since it just involves looking up some past data and does not require any weather-forecasting expertise. In contrast, although a uniform distribution requires no expertise, it may not seem at all reasonable, as in the case of a dry location with a very low climatological relative frequency of rain.

Asymmetric rules can be generated from symmetric rules. For example, in the single-event case for which it is felt that the expected score with honest reporting should be minimized at a probability of q , we can take any symmetric strictly proper scoring rule S and create a new rule $S^*(r, e|q) = [S(r, e) - S(q, e)]/T(q)$, where $T(q) = S(1, 1) - S(q, 1)$ if $r \geq q$ and $S(0, 0) - S(q, 0)$ if $r \leq q$.

More generally, the families of scoring rules given by Equations (12) and (13) can be generalized to pseudospherical and power families of strictly proper scoring rules that allow for the incorporation of baseline distributions [9]. If the baseline

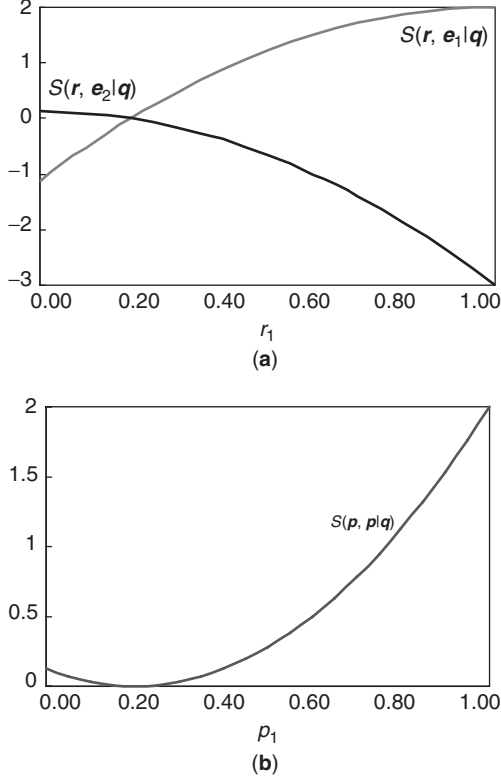


Figure 2. (a) Score functions $S(\mathbf{r}, \mathbf{e}_1|\mathbf{q})$ and $S(\mathbf{r}, \mathbf{e}_2|\mathbf{q})$ and (b) expected score $S(\mathbf{p}, \mathbf{p}|\mathbf{q})$ under honest reporting for the power scoring rule with $\beta = 2$ and $\mathbf{q} = (0.2, 0.8)$.

distribution for a set of k mutually exclusive and collectively exhaustive events is denoted by $\mathbf{q} = (q_1, \dots, q_k)$, then we can define the pseudospherical and power families of scoring rules with baselines as follows:

$$S_\beta^S(\mathbf{r}, \mathbf{e}_i|\mathbf{q}) = \frac{1}{\beta - 1} \left[\left(\frac{r_i/q_i}{E_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}]} \right)^{\beta-1} - 1 \right] \quad (16)$$

and

$$S_\beta^P(\mathbf{r}, \mathbf{e}_i|\mathbf{q}) = \frac{(r_i/q_i)^{\beta-1} - 1}{\beta - 1} - \frac{E_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}] - 1}{\beta}, \quad (17)$$

where $E_{\mathbf{r}}[(\mathbf{r}/\mathbf{q})^{\beta-1}] = \sum_{i=1}^k r_i(r_i/q_i)^{\beta-1}$ and $-\infty < \beta < \infty$. These scoring rules are scaled so that they yield scores of 0 when $\mathbf{r} = \mathbf{q}$. Thus, a positive score represents improvement over the baseline and a negative score indicates a forecast worse than the baseline. (The expert's *expected* score with honest reporting is positive except at $\mathbf{r} = \mathbf{q}$, where it is 0.) As with Equations (12) and (13), $\beta = 2$ corresponds to spherical and quadratic rules, respectively, in Equations (16) and (17), and both families converge to a logarithmic rule when $\beta \rightarrow 1$. Figure 2 shows $S_\beta^S(\mathbf{r}, \mathbf{e}_1|\mathbf{q})$, $S_\beta^S(\mathbf{r}, \mathbf{e}_2|\mathbf{q})$, and $S(\mathbf{p}, \mathbf{p}|\mathbf{q})$ for the power scoring rule with $\beta = 2$.

The consideration of baseline distributions provides a relative evaluation as opposed to an absolute evaluation, and relative evaluations are often of great interest. In addition, evaluations with baseline distributions can be useful in evaluating probabilities (and evaluating the forecasters providing the probabilities) that are made under different circumstances. For example, if one weather forecaster assesses probabilities of rain in a very dry climate (say, with a climatology of 0.05) and another forecaster assesses probabilities in a more moist climate (climatology 0.40), then it is much easier for the first forecaster to obtain higher scores in an absolute evaluation. This is because of the fact that the first forecaster is able, on average, to make sharper forecasts. If we use a relative evaluation with climatology as the baseline, then we are comparing the two forecasters in terms of how effective they are at improving upon a forecast based solely on climatology, thereby adjusting for the differences in the forecast situations. While not perfect, this will tend to even the playing field somewhat and make for fairer comparisons.

SCORING RULES THAT ARE SENSITIVE TO DISTANCE

In some situations, the events of interest are ordered. For example, in a soccer match, a team can win, lose, or tie. A win is better than a tie, which in turn is better than a loss,

so there is an ordering. If we are giving probabilities for $\tilde{x}$, the amount of rain in inches on a given day, we might assess probabilities for $\tilde{x} = 0$, $0 < \tilde{x} \leq 0.5$, $0.5 < \tilde{x} \leq 1$, and $\tilde{x} > 1$. These four events are ordered. The scoring rules for multiple events, discussed in the preceding sections, ignore any ordering. Suppose that two experts report probabilities of (0.3, 0.4, 0.2, 0.1) and (0.3, 0.1, 0.2, 0.4). If there is no rain, then the experts will receive the same score. They both gave a probability of 0.3 for the event that occurred, and they gave probabilities of 0.4, 0.2, and 0.1 for the other three events. The fact that the latter three probabilities were in different orders does not affect the score because the scoring rule does not take ordering of the events into account. Some might argue that the first expert gave more probability on the event closest to the event that occurred and should therefore receive a higher score. Scoring rules have been developed that would take ordering into account in this way, and we say that such rules are sensitive to distance. Informally, this means that for the probability not assigned to the event that occurs, a higher score will result if more probability is given to events closer to the event that occurs and less probability to “more distant” events.

The first strictly proper sensitive-to-distance scoring rule was a quadratic rule called the *ranked probability score* [12]: $S(\mathbf{r}, \mathbf{e}_i) = -\sum_{j=1}^{i-1} R_j^2 - \sum_{j=i}^{k-1} (1 - R_j)^2$, where $R_j = \sum_{l=1}^j r_l$ is a cumulative probability. By connecting the score to cumulative probabilities, the rule is able to take sensitivity to distance into consideration. As probability “moves” from events more distant from the event that occurs to events closer to the event that occurs, the cumulative probabilities change accordingly and result in an increase in the score.

The same idea that is used to generate the ranked probability score from the quadratic score can be used to obtain a sensitive-to-distance scoring rule S^* based on any strictly proper rule S for a single event and its complement:

$$S^*(\mathbf{r}, \mathbf{e}_i) = \sum_{j=1}^{i-1} S(R_j, 0) + \sum_{j=i}^{k-1} S(R_j, 1). \quad (18)$$

The corresponding expected score is $S^*(\mathbf{r}, \mathbf{p}) = \sum_{i=1}^{k-1} [P_i S(R_i, 1) + (1 - P_i) S(R_i, 0)]$, using a vector $\mathbf{P}$ of cumulative probabilities $P_j = \sum_{l=1}^j p_l$ based on $\mathbf{p}$. Note that Equation (18) can be used to generate new pseudospherical and power families of strictly proper scoring rules, with or without baseline distributions [13]. If baseline distributions are used, they will be expressed in cumulative form, with a vector $\mathbf{Q}$ of cumulative probabilities $Q_j = \sum_{l=1}^j q_l$ representing the cumulative baseline distribution.

It is important to mention that properties of the scoring rule S in Equation (18), other than the fact that it is strictly proper, are not necessarily inherited by S^* . For example, if S is logarithmic, S^* will not inherit the property of locality mentioned earlier. The score is based on the cumulative probabilities $R_1, \dots, R_{k-1}$, so it clearly depends on more than just r_i . Also, if S^* is determined from Equation (18) using a symmetric, strictly proper S without a baseline distribution, then the expected score $S^*(\mathbf{p}, \mathbf{p})$ is minimized at $\mathbf{p} = (0.5, 0, \dots, 0, 0.5)$. That is, if a baseline distribution is not chosen, the default baseline distribution is $(0.5, 0, \dots, 0, 0.5)$, not $(1/k, \dots, 1/k)$, and the score for a uniform distribution will not be the same for all events because the ordering of the events is relevant and some events are more distant from others. Note that the default baseline distribution of $(0.5, 0, \dots, 0, 0.5)$ for S^* translates to $(0.5, \dots, 0.5)$ when expressed in terms of the cumulative probabilities $(R_1, \dots, R_{k-1})$. Since the relevant probabilities are cumulative in a score that is sensitive to distance, the baseline distribution is uniform in the cumulative probabilities. Furthermore, this distribution will give the same score S^* regardless of which event occurs, because $R_j = 0.5$ in each of the $S(R_j, 0)$ and $S(R_j, 1)$ terms in Equation (18).

Commonly encountered scoring rules ignore any ordering of the events. When the events of interest are ordered, however, the ordering may be important in terms of the underlying real-world situation. For instance, with forecasts of returns on an investment, high probabilities for values that are not identical to the returns that actually occur but are close to those returns would

seem to be more valuable for investment purposes than high probabilities for values that are quite distant from the actual returns. In such a setting, a scoring rule that is sensitive to distance might provide incentives and an *ex post* evaluation that are more consistent with the decision making problem.

SUMMARY AND DISCUSSION

Probability forecasts are important inputs to quantify uncertainty in inferential and decision making problems. It is therefore important to have appropriate incentives for careful formulation of probability forecasts and to have measures to evaluate the forecasts once the uncertainty is resolved and we see what actually happens. Those are exactly the roles that are played by scoring rules. In particular, strictly proper scoring rules provide incentives for making good forecasts (i.e., sharp forecasts) and reporting them honestly. In terms of *ex post* evaluation, the incentive for honest reporting *ex ante* translates into measures of the calibration of the forecasts, and given good calibration, sharper forecasts will earn higher scores on average. A few scoring rules tend to be used most often, but rich families of strictly proper scoring rules have been developed. Beyond the basic rules, there are options that add more flexibility while maintaining the strictly proper nature of the rules. Some rules allow the evaluation of probabilities relative to a chosen baseline distribution. Among other things, this makes scores for probabilities made in different situations more comparable. Other rules take into account any ordering of the events and are sensitive to distance in the sense of giving higher scores to probability distributions assigning higher probabilities to events near the event that occurs, all other things being equal. This feature is relevant when being “close” with the probability forecast can lead to better decisions or inferences.

How might a user choose a scoring rule in a given situation? Among the basic rules, some have different properties from others and different rules can lead to different scores and different rankings of experts [14]. Thus,

the choice of a rule might depend on how one feels about those properties and about the situation at hand. For example, with probabilistic answers for a multiple-choice test, the locality property of the logarithmic rule might have strong appeal. At the same time, the possibility of a score of negative infinity with the logarithmic score is a potential concern; some claim that the logarithmic rule has undesirable properties [15], and in certain settings, locality becomes less important. For example, in a two-event setting, all rules satisfy locality, and in a setting where sensitivity to distance is considered important, a sensitive-to-distance logarithmic rule no longer has the locality property. There is no general agreement on a single “best” rule for all situations.

The use of a scoring rule that involves the choice of a baseline distribution depends on whether a relative evaluation is desired and whether there is a specific baseline distribution against which to compare probabilities. An important thing to keep in mind is that using the basic rules without choosing a baseline distribution means that probabilities are being evaluated relative to the default distribution, which is a uniform distribution. Another choice when events are ordered is whether to use a rule that is sensitive to distance, and this choice is related to whether giving higher probabilities to events close to the event that occurs is viewed as important for the situation at hand.

What about practical issues in using the rules? The use of the rules *ex post* to evaluate probabilities is straightforward, just involving the computation of scores using the formula for any scoring rule that is chosen. Scores can then be used as feedback to enable experts and modelers to see their performance and perhaps learn from it. The use of scoring rules *ex ante* means that they should be part of a general probability assessment process that might include some training regarding probability if necessary. For many experts, the connection of the scoring rule formulas with the incentives is too opaque to make it valuable to dwell on the formulas. Discussing the incentives in an intuitive fashion is generally more effective. One option for relatively simple cases (e.g., a probability

for a single event) is to present the possible scores in graphical or tabular form.

The incentive to maximize expected score with strictly proper scoring rules is probably reasonable in most cases. In the context of thinking of the expert as wanting to maximize expected utility, it implies that the utility function is linear in the score (or linear in money if the score is translated into a monetary reward). If the expert's utility function $U(S)$ for the score is known, a modification of the score to $S' = U^{-1}(S)$ with a strictly proper S will adjust for U , since $U(S') = U[U^{-1}(S)] = S$, and will thereby encourage honest reporting. A practical problem with this is that we are not likely to know U , and eliciting it from the expert is not easy. In many cases, we feel that the importance of the score is probably not large enough to cause major violations of such linear utility due to risk aversion or risk taking, for example. However, if there are other stakes related to the probability forecasts, those stakes might cause significant shifting of the probabilities away from honest reporting. For example, if the situation is viewed as a contest against other experts (rightly or wrongly), the consideration of strategic play might lead the expert to give more extreme probabilities (i.e., probabilities closer to and often equal to 0 or 1) than justified by the expert's best judgments, in order to try to win the contest [16]. In most cases, however, we would expect that experts will try to come up with the best set of probabilities given the information that is available to them, and will not think strategically. In any event, strictly proper scoring rules can always be used for *ex post* evaluation purposes, and any hedging of reported probabilities can be expected to lead to lower average scores.

In closing, we note that work on scoring rules has interesting connections to other fields. Scoring rules are closely connected to decision theory/decision analysis. A decision maker may hire an expert to report probabilities for events related to the decision and might like to tailor a scoring rule to the decision-making problem, in the spirit of Savage's "share of the business" notion [6]. The expert's reported probabilities can

be viewed as new information by the decision maker, and connections between scores and the value of that information are of interest. These notions are related to the literature on incentives and mechanism design in economics and especially to agency theory. On a different tack, expected scores from strictly proper scoring rules are related to information measures from signal processing and information theory [9]. For example, the expected score for honest reporting under a logarithmic scoring rule is the negative Shannon entropy of the expert's probability distribution $\mathbf{p}$ and is the Kullback–Leibler divergence of $\mathbf{p}$ with respect to $\mathbf{q}$ if the baseline distribution $\mathbf{q}$ is chosen. Finally, extensive experimental work by psychologists has investigated the degree to which individuals' probability assessments are well calibrated and has led to various theories of the calibration of subjective probabilities [17]. Judgments about others' probabilities are important in competitive situations, and economics and psychology are both relevant for this issue. Given the importance of probability forecasts in decision modeling and statistics as well as the connections with these different (and somewhat disparate) fields, we expect the interest in scoring rules and the application of such rules in practice to grow.

REFERENCES

1. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950;78(1):1–3.
2. Good IJ. Rational decisions. *J R Stat Soc [Ser B]* 1952;14(1):107–114.
3. McCarthy J. Measures of the value of information. *Proc Natl Acad Sci USA* 1956;42(9):654–655.
4. de Finetti B. Does it make sense to speak of "good probability appraisers"? In: Good IJ, editor. *The scientist speculates: an anthology of partly-baked ideas*. New York: Wiley; 1962. pp. 357–363.
5. Winkler RL, Murphy AH. "Good" probability assessors. *J Appl Meteorol* 1968;7(5):751–758.
6. Savage LJ. Elicitation of personal probabilities and expectations. *J Am Stat Assoc* 1971;66(336):783–801.

7. Winkler RL. Scoring rules and the evaluation of probabilities. *Test* 1996;5(1):1–60.
8. Gneiting T, Raftery A. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 2007;102(477):359–378.
9. Jose VRR, Nau RF, Winkler RL. Scoring rules, generalized entropy, and utility maximization. *Oper Res* 2008;56(5):1146–1157.
10. De Groot MH, Fienberg SE. Assessing probability assessors: calibration and refinement. In: Gupta SS, Berger JO, editors. *Statistical decision theory and related topics*. New York: Academic Press; 1982. pp. 291–314.
11. Winkler RL. Evaluating probabilities: asymmetric scoring rules. *Manage Sci* 1994;40(11):1395–1405.
12. Epstein ES. A scoring system for probability forecasts of ranked categories. *J Appl Meteorol* 1969;8(6):985–987.
13. Jose VRR, Nau RF, Winkler RL. Sensitivity to distance and baseline distributions in forecast evaluation. *Manage Sci* 2009;55(4):582–590.
14. Bickel JE. Some comparisons among quadratic, spherical, and logarithmic rules. *Decis Anal* 2007;4(2):49–65.
15. Selten R. Axiomatic characterization of the quadratic scoring rule. *Exp Econ* 1998;1(1):43–62.
16. Lichtendahl KC, Winkler RL. Probability elicitation, scoring rules, and competition among forecasters. *Manage Sci* 2007;53(11):1745–1756.
17. O'Hagan A, Buck CE, Daneshkhah A, *et al.* *Uncertain judgements: eliciting experts' probabilities*. Chichester: Wiley; 2006.

SEARCH GAMES

SHMUEL GAL
Department of Statistics,
University of Haifa,
Haifa, Israel

The first significant contribution to operations research in the area of searching is the work of Koopman and his colleagues “Search and Screening” [1]. The work in these early years of search theory emphasized the problem of optimal distribution of effort spent in search. That topic is the main subject of the classic work of Stone [2], “Theory of Optimal Search,” which was awarded the 1975 Lanchester Prize by the Operations Research Society of America. The early work on Search Theory has been surveyed in Ref. 3. The later survey [4] shows how the determination of search trajectories has come to be studied more extensively.

The present article is focused on *search games* with players that move along continuous trajectories. (Many interesting search games of discrete nature, not covered by our article, can be found in the further reading list.) These continuous search games were introduced by Isaacs [5] in the last chapter of his book and were developed by Gal [6]. That work, and the updated version by Alpern and Gal [7], has stimulated much research, with applications in computer science, economics, and biology. The present article gives an overview of the area of search games including some recent development in this area.

A situation of “hide and seek” is quite common, ranging from children’s play to military conflicts. This situation is analyzed as a two-person zero-sum game which is called a *search game*. Such a game is characterized by a search space Q which is either a compact set in a Euclidean space or an unbounded connected set, for example, the real line. The searcher usually starts moving from a specified point O called the *origin* but choosing the starting point arbitrarily is also considered. It is assumed that the searcher is free

to choose any continuous trajectory inside Q , subject to a maximal velocity constraint which is normalized to 1. The hider can be either stationary or mobile. It will always be assumed that neither the searcher nor the hider has any knowledge about the movement of the other player until their distance apart is less than or equal to the discovery radius r and at this very moment, capture occurs. The discovery radius r is assumed to be small with respect to the dimension of Q . For one-dimensional sets, r is usually assumed zero. A search trajectory will be denoted by S and a hiding trajectory by H (a single point if the hider is immobile). The cost function (payoff to the maximizing hider) $c(S, H)$, is the time spent until the hider is found (the search time). A mixed strategy s (or h) of the searcher (or hider) is a probability measure on the set of the searcher’s (or hider) pure strategies. To rigorously present such strategies, one has to introduce a substantial amount of measure-theoretic machinery for these sets. Such a construction is presented in Ref. 6 including the result that $c(S, H)$ is Borel measurable in both variables, so that we can define the payoff c in the case that the searcher uses s and the hider uses h as the expected value of c with respect to the product measure $s \times h$:

$$c(s, h) = \int c(S, H) d(s \times h).$$

The existence of a value, v , (minimax theorem) for such problems, and an optimal searcher mixed strategy are established in Refs 6 and 8. (An optimal hiding strategy need not exist, only ε -optimal strategies always exist.) The existence theorem holds for both, immobile and mobile hider.

IMMOBILE HIDER

Searching a Graph

Assume that the search space Q is a finite connected graph in which each arc has a given length. The sum of these lengths will

be denoted by μ , called the *total length of the graph*. The graph Q can be viewed as a metric space, where the distance $d(.,.)$ between any two points (not just nodes) is the length of the shortest path between them. A pure strategy for the hider is simply a (hiding) point H in Q , which can be a node or a point on one of the arcs. A pure strategy $S = S(t)$ for the searcher is a continuous trajectory in Q with speed 1, that is, a function satisfying $d(S(t), S(t')) \leq |t - t'|$, for all $t, t' \geq 0$. The payoff c corresponding to pure strategies S and H is the search time $c(S, H) = \min\{t : S(t) = H\}$.

It is easy to see that, for any graph the value of the search game satisfies $v \geq \mu/2$ attained by the hider if he chooses a uniform hiding strategy h_μ . Since any search trajectory S has a unit discovery rate, then for any time $t > 0$, $P(c(S, h_\mu) \geq t) \geq 1 - \frac{t}{\mu}$. Integrating this inequality yields $E(c(S, h_\mu)) \geq \mu/2$.

Fixed Start. A natural way to search a graph is to use a *Chinese postman tour* [9] that is, a closed trajectory that visits all the arcs of Q and has minimal length. Such a tour will be denoted by S^* . Obviously, if Q is Eulerian, then S^* can be chosen as any Eulerian tour (that visits each arc just once). In this case, if $S^*(t)$, $0 \leq t \leq \mu$ is an Eulerian tour, then visiting it with probability $1/2$ in the forward direction (using $S^*(t)$ as the search trajectory) and probability $1/2$ in the backward direction (using $S^*(\mu - t)$ as the search trajectory) is an optimal search strategy because it assures expected capture time at most $\mu/2$. Since the hider can assure at least $\mu/2$ by the uniform hiding strategy, it follows that $v = \mu/2$ for Eulerian graphs. It can also be shown that $v = \mu/2$ holds *only* for Eulerian graphs so, for any non-Eulerian graph $v > \mu/2$. An upper bound for v is μ which is attained *if and only if* Q is a tree.

The search strategy used for Eulerian graphs can be extended to non-Eulerian graphs by using a *random Chinese postman tour* that encircles S^* equiprobably in each direction. It was proven by Gal [10] that a random Chinese postman tour is an optimal search strategy *if and only if* Q is weakly Eulerian (a set of Eulerian subgraphs connected in a tree-like fashion). The optimal

hiding strategy can be constructed by a simple recursive algorithm.

A graph which is not weakly Eulerian is likely to have a very complicated solution. The following misleadingly simple example illustrates the expected difficulties: Assume that Q consists of just three unit arcs b_1 , b_2 , and b_3 connecting two nodes, the origin O and another node A . It is obvious that an optimal search strategy has to pick an arc b_i at random and go from O to A . Surprisingly, upon reaching A , it is *not* optimal to pick another arc at random and fully traverse it (and later the remaining arc). It turns out that the optimal strategy, when reaching A , is to pick an arc, b_j ($j \neq i$) at random and go only a distance x along b_j , where x is chosen by a specific random distribution [7, p. 33] return to A , fully traverse b_k (k different from i and j) and only then complete the traversal of the arc b_j . This strategy was suggested by Donald J. Newman in the late 1970s but the rigorous proof of its optimality is very complicated and took about 15 years to establish [11].

In general, the problem of finding the optimal search strategy is NP-hard as shown by Von Stengel and Werchner [12]. However, they also showed that if the time of search is limited by a (fixed) bound, which is logarithmic in the number of nodes, then the optimal strategy can be found in polynomial time. The search game on a graph can, in general, be formulated as an infinite-dimensional linear program. This formulation with an algorithm for obtaining its (approximate) solution is presented by Anderson and Aramendia [13].

Arbitrary Start. Removing the assumption of a fixed starting point of the searcher, assuming, instead, that the searcher can *choose* the starting point of his trajectory leads to an interesting problem which has been only recently investigated. All the other assumptions, made for the fixed-start games, are the same.

The *Chinese postman path* $\tilde{S}$ is a path, not necessarily closed, which visits all the points of Q and has minimum length. The searcher can use the following strategy which resembles the random Chinese postman tour (used in the fixed-start games).

Simple: Choose the starting point randomly with probability $1/2$ for each end of $\tilde{S}$ and go to the other end along $\tilde{S}$.

Sometimes *Simple* is optimal. For example, if $\tilde{S}$ is an Eulerian path (not necessarily closed) having total length μ , then *Simple* guarantees search time not exceeding $\mu/2$. On the other hand, the hider can guarantee expected search time $\geq \mu/2$ by using a uniform hiding strategy. Thus, the value of this game is $\mu/2$ which implies that *Simple* is optimal. Such an example is the three-arcs graph which is very difficult for the fixed start but very easy for the arbitrary start search game.

Except for the Eulerian path case, described above, *Simple* is optimal for any tree [14], and also for any tree with one or several Eulerian graphs attached [15]. However, Alpern *et al.* [16] have shown that topological characterization of the graphs, for which *Simple* is optimal, is *impossible*. This fact is quite surprising because it contradicts the result for the fixed-start search games.

Searching an Unknown Graph. If the structure of the graph is not known in advance, then the optimal strategy to find the hider is to explore the graph using a randomized version of a well-known search method in computer science called *depth-first search*. This randomized version enables the searcher to find the hider in expected time μ , as shown in Ref. 17, while the usual application of depth-first search can only guarantee 2μ search time.

Search in the Plane

Searching an immobile hider in any two-dimensional convex region (actually a much weaker condition is needed, see Ref. 6, p. 39) with area μ , is similar to searching an Eulerian graph. The searcher has to find a closed curve S^* that covers all the points of the region Q (in the sense that for any $x \in Q$ there exists a $t \geq 0$ such that $d(S^*(t), x) \leq r$). The length of this covering curve can be made less than $(1 + \varepsilon)\mu/2r$, where $\varepsilon \rightarrow 0$ as $r \rightarrow 0$. By encircling this curve equiprobably in each direction the searcher can assure about $\mu/4r$ expected search time. On the other hand, it can easily be shown that the hider can keep

the expected search time above $\mu/4r$ by using a uniform hiding distribution. Thus, $v \sim \mu/4r$ as expected. (The relation $\sim$ means that the ratio between the two terms tends to one as $r \rightarrow 0$.)

MOBILE HIDER

Isaacs' Princess and Monster Game

In his 1965 book [5], Rufus Isaacs presented the following game which he named *princess and monster*. The monster searches for the princess, the time required being the pay-off. They are both in a totally dark room (of any shape) but they are each cognizant of its boundary. Capture means that the distance between the princess and the monster is less than or equal to r the capture radius, which is assumed to be small in comparison with the dimension of the room. The monster, supposed to be highly intelligent, moves at a known speed. We permit the princess full freedom of locomotion.

This game remained an open problem until it was solved by Gal [18], under the condition that the search space Q is convex. Actually a much weaker condition is needed [6, p. 39]. The value, v of the princess and monster game satisfies $v \sim \frac{\mu}{2r}$, where μ is the area of the room (twice the value obtained for an immobile hider). Gal's optimal strategy for the princess is especially interesting: Go to a random location in the room. Stay still for a time interval which is not too short but not too long (the exact formula is given in Ref. 18); go to another, independent, random location; stay still again, and so on. His proposed optimal search strategy is based on the following scheme: Subdivide Q into many *narrow* rectangles. Pick a rectangle at random and search it in a specific way (for details see Refs 6 and 7); after some time pick another rectangle randomly, independently, etc. Note that this type of strategy would not be efficient if the rules of the game were changed in such a way that the hider is informed about the position of the searcher from time to time. A search strategy that is robust to such a change of rules has been presented by Lalley and Robbins [19] as follows: Choose the starting point

uniformly on the boundary of Q . Choose a random direction (from the tangent at that point) θ_1 with density $\frac{1}{2} \sin \theta$, $0 \leq \theta \leq \pi$. Continue with unit speed until you reach the boundary, and choose an i.i.d. direction θ_2 etc. This strategy is robust to partial information: even if the hider is given the position and direction of the searcher from time to time (not too often though) then the hider will not be able to predict its course for very long. It should be noted, though, that Lalley and Robbins' strategy would be ineffective for nonconvex regions but Gal's search strategy is also effective for nonconvex regions and can be adapted to more general problems; for example, a nonconstant capture radius, multidimensional regions, more general cost functions, and search game with several searchers. Several other extensions have been presented by Garnaev (e.g., Ref. 20 in which the probability of detection depends on the distance between the players).

Mobile Hider in a Graph

The princess and monster game on a circle has been suggested by Isaacs [5] as a stepping stone for the game in the region described above. This open problem was (independently) solved by Alpern [21] and Zelikin [22]. (A discrete model has been solved by Wilson [23].) Their optimal search strategy is Alpern's *coin half tour* as follows: Denote the searcher's starting point by O and its antipode by A . Start by tossing a symmetric coin and use the outcome to decide whether to turn left or right as you go (with unit speed) to the antipode; upon reaching A , toss the coin again (independently) and go back to O using the outcome of the toss-up to decide on the direction and so on. This solution was obtained under the assumption that at the beginning of the game, the locations of the searcher and the hider on the circle are uniformly and independently distributed. However, the same search strategy is optimal for a fixed or for an arbitrary starting point. The strategy of the hider is identical to the strategy of the searcher under the assumption of uniformly distributed starting point and also under arbitrary starting point. However, if the searcher starts from a fixed (known)

point, then the hider can take advantage of it by staying at A until time $\mu/2$ (half the circumference) $-\varepsilon$, and starting the coin half tour then.

Except for the circle, solving the princess and monster game on a graph is usually very difficult. For example, even solving the search game on the unit interval (a graph with two nodes and a unit arc connecting them) with an arbitrary start is very difficult. It might be expected that starting at one end (chosen at random) and sweeping as fast as possible the whole interval is an optimal search strategy, but this is not so. The hider's best response to that is to start at distance ε from a random end, and move to the other end with unit speed. This leads to $0.75 (-\varepsilon)$ expected search time but the searcher can do better and the value of the game is about 0.7. This unexpected behavior, and an approximate solution of the game on the interval has been presented by Alpern *et al.* [24].

SEARCH IN UNBOUNDED DOMAINS

Linear Search Problem

The *linear search problem* was introduced by Bellman [25] and, independently, by Beck [26] as follows: A target is located on the real line, at a point H , with a known probability distribution. A searcher, whose maximal velocity is 1, starts from the origin O and wishes to discover the target in minimal expected time. Usually, it is assumed that changing the direction of motion can be done instantaneously.

Much research has been carried out by Beck and others [27,28] but the linear search problem for a general probability distribution is still unsolved. However, a dynamic programming algorithm can produce a solution for any discrete distribution [29], and also an approximate solution, with any desired accuracy, for any probability distribution [7].

The seminal paper by Beck and Newman [30] considers the linear search problem within the framework of a search game. In order to have a meaningful game, they allow only hiding strategies with expected distance (from O) at most λ . The constant λ is assumed to be known to the searcher

but it turns out that the optimal search strategies do not depend on λ . Within this framework they show that the minimax search trajectory is to oscillate around O , *doubling* the amplitude at each step. The doubling trajectory guarantees capture by time $9|H|$ so that the expected capture time is at most 9λ and this constant is the best that can be achieved by a fixed trajectory.

They also show that the game has a value equal to $(1 + \bar{g})\lambda$ ($\doteq 4.6\lambda$) where $\bar{g}$ minimizes $\frac{g+1}{\ln g}$. An optimal (mixed) search strategy that achieves this value uses a geometric trajectory, with the generator $\bar{g}$, multiplied by a random constant $c_u = \bar{g}^u$, where u is chosen at random by the uniform distribution in the interval $[0, 2)$. Thus, the n th step size of the search strategy is $c_u \bar{g}^n$.

Rather than restricting the hider, the approach used by Gal [31] is to normalize the cost function by the distance of the target from the origin. The two approaches lead to equivalent results [6]. This type of normalization was new at that time but is now called the *competitive ratio* in computer science literature and is a standard tool in analyzing online algorithms [32].

The online solution with a positive turn cost d , is given by Demaine *et al.* [33]. The presented trajectory guarantees paying at most $9|H| + 2d$, having the (minimal) competitive ratio 9 with respect to $|H|$ plus the minimum corresponding additive term. The corresponding step size is $d(2^n - 1)/2$ for the n th step, $n = 1, 2, \dots$. It also turns out that optimal randomized strategies for searching the line in the presence of turn cost have the same (minimal) competitive ratio of about 4.6, as in the model without turn cost.

Finding Minimax Search Trajectories

Search games in unbounded domains often use a minimax solution for a homogenous unimodal functional with two arguments representing a search trajectory and a hiding point. It was shown by Gal and Chazan [34] that a minimax search trajectory for such a functional is always a geometric sequence (or exponential functions for continuous problems). This result has been developed in Ref.

6 as a general tool. In order to find the minimax trajectory all that needed is to minimize a simple function over a single parameter (the generator of this sequence) instead of searching over the whole trajectory space. This tool has been used for the linear search problem and also in the problems, to be discussed, of searching a set of rays (star search) and searching in the plane using exponential spirals. It has also been used for minimax search on the boundary of a region consisting of two rays in the plane. That problem has been presented by Baeza-Yates [35], and solved by Gal using the general tool [7, pp. 154–157].

Geometric trajectories are useful in other fields too in order to obtain effective algorithms, as shown by Chrobak and Kenyon-Mathieu [36]. That survey describes the “doubling” method for designing online and offline approximation algorithms. The rough idea is to use geometrically increasing estimates on the optimal solution to produce fragments of the algorithm’s solution. This method is illustrated by discussing several applications where doubling is used. These areas, in addition to searching, include bidding, minimum latency tours, scheduling, and clustering. This work raises the hope that doubling will become a standard tool for designing approximation algorithms.

Star Search

The natural extension of the linear search problem to $M > 2$ unbounded rays radiating from the origin, has been presented and solved by Gal [37]. It was later rediscovered in computer science literature [35] and has been called the *cow path search* in Ref. 38 which also presented several applications.

Gal showed that the optimal trajectory is *periodic* (visiting each ray every M th time) with *monotone*-increasing step size and, using the general tool described in the section titled “Finding Minimax Search Trajectories,” demonstrated that the minimax trajectory has a geometric step size with generator $\frac{M}{M-1}$. The minimax trajectory achieves the competitive ratio $1 + 2 \frac{M^M}{(M-1)^{M-1}} \sim 2eM$ ($\doteq 5.4M$) for large M .

Similar to the linear search problem, the (expected) competitive ratio can be improved

by using mixed search strategies. Specifically, the geometric trajectory with the generator $\bar{g}$, that minimizes $\frac{2(\bar{g}^M - 1)}{M(\bar{g} - 1) \ln \bar{g}}$ multiplied by a random constant $c_u = \bar{g}^u$ where u is uniformly distributed in the interval $[0, M)$, is optimal: using a periodic trajectory with step size $c_u \bar{g}^u$, leads to competitive ratio of about $3.09M$. In contrast to the pure strategy case, however, the optimality proof applies only to strategies that use periodic and monotone trajectories [6, 39–41].

López-Ortiz and Schuierer have solved several extensions of the star search. If a bound, D , is given for the distance of the hiding point from the origin then they showed that the competitive ratio is $1 + 2 \frac{M^M}{(M-1)^{M-1}} - O(1/\log D)$. A nice feature of the presented trajectory is that it is not necessary to know an upper bound on D in advance. A similar phenomenon occurs in the Beck and Newman strategies that are independent on the bound for the expected hiding distance λ .

Online star search by several agents who move in parallel has been solved in Ref. 43. If there are p searchers, then the competitive ratio is $1 + 2 \left(\frac{M}{p} - 1 \right) \left(\frac{M}{M-p} \right)^{M/p}$.

Searching in the Plane

Optimal Sweeping. The continuous version of the general tool described in the section titled “Finding Minimax Search Trajectories” is useful to find the minimax search trajectory for the following search problem in the plane. An (immobile) hider is located at a point in the plane with unknown distance and direction from the origin O . The searcher chooses a trajectory, starting from O and discovers the hider at the moment when the hider is covered by the area swept by the radius vector $\mathbf{R}$ of his trajectory. Natural search trajectories have the following *periodic and monotonic* property: the angle θ of the radius vector is always strictly increasing, and the trajectory does not intersect itself. Under this assumption, it can be proven that a minimax search trajectory can be found in the family of exponential spirals, that is, $|\mathbf{R}| = e^{b\theta}$, where $b > 0$ is a constant. The competitive ratio is $e^{2\pi b} \sqrt{1 + 1/b^2}$ with minimal value of about 17.3, attained

at $b \doteq 0.16$. It was shown by Gal [6] that this strategy is the best among all monotone and periodic strategies. Since then, it has been an open problem whether the given strategy is optimal in general. A recent work of Langetepe [44] has settled this old open problem and showed that the spiral search described above is indeed the globally optimal trajectory.

“Swimming-in-a-Fog” Problem. The following problem has been proposed by Bellman [45]. A person has been shipwrecked in a fog (at the point O) and wishes to minimize the maximum time required to reach the shore. The shape of the shore is known (say a straight line) and some information is given about its distance from O .

If the distance, D , to the shore is known exactly, then the minimax trajectory is given by Isbell [46] as follows: For convenience assume that $D = 1$. Imagine a clock face around O ; walk towards 1 o'clock for $\sqrt{4/3}$ units; then, turn on the tangent that strikes the unit circle at 2 o'clock; follow the circle to 9 o'clock and, finally, continue on a tangent for one more unit (until reaching the tangent to the unit circle at 12 o'clock). The length of this minimax trajectory is $(7\pi/6 + \sqrt{3} + 1)D \doteq 6.4D$. However, this trajectory is not effective unless D is known exactly.

An online framework for the general problem is presented in Ref. 6. The goal is to minimize the ratio between the distance traveled until reaching the shore and the unknown distance to the shore, D . It has been suggested that the solution could be an exponential spiral but the problem seems difficult and is still open.

REFERENCES

1. Koopman BO. Search and screening, Operations Evaluations Group, Report. No. 56. Rosslyn (VA): Center for Naval Analysis; 1946.
2. Stone LD. Theory of optimal search. 2nd ed. Arlington (VA): Operations Research Society of America; 1989.
3. Dobbie JM. A survey of search theory. Oper Res 1968;16:525–537.

4. Benkoski S, Monticino M, Weisinger J. A survey of the search theory literature. *Nav Res Log* 1991;38:469–494.
5. Isaacs R. *Differential games*. New York: Wiley; 1965.
6. Gal S. *Search games*. New York: Academic Press; 1980.
7. Alpern S, Gal S. *The theory of search games and rendezvous*. Boston: Springer; 2003.
8. Alpern S, Gal S. A mixed strategy minimax theorem without compactness. *SIAM J Control Optimiz* 1988;26:1357–1361.
9. Edmonds J, Johnson EL. Matching Euler tours and the Chinese postman problem. *Math Program* 1973;5:88–124.
10. Gal S. On the optimality of a simple strategy for searching graphs. *Int J Game Theor* 2000;29:533–542.
11. Pavlovic L. Search game on an odd number of arcs with immobile hider. *Yugosl J Oper Res* 1993;3(1):11–19.
12. Von Stengel B, Werchner R. Complexity of searching an immobile hider in a graph. *Discrete Appl Math* 1997;78:235–249.
13. Anderson EJ, Aramendia MA. The search game on a network with immobile hider. *Networks* 1990;20(7):817–844.
14. Dagan A, Gal S. Network search games, with arbitrary searcher starting point. *Networks* 2008;52(3):156–161.
15. Alpern S. Hide-and-seek games on a tree to which Eulerian networks are attached. *Networks* 2008;52(3):109–178.
16. Alpern S, Baston V, Gal S. Network search games with immobile hider, without a designated searcher starting point. *Int J Game Theor* 2008;37:281–302.
17. Gal S, Anderson EJ. Search in a maze. *Probab Eng Inf Sci* 1990;4:311–318.
18. Gal S. Search games with mobile and immobile hider. *SIAM J Control Optimiz* 1979;17:99–122.
19. Lalley SP, Robbins HE. Stochastic search in a convex region. *Probab Theor Relat Fields* 1988;77(1):99–116.
20. Garnaev A. A remark on the princess and monster search game. *Int J Game Theor* 1992;20:269–276.
21. Alpern S. The search game with mobile hider on the circle. In: Roxin EO, Liu PT, Sternberg RL, editors. *Differential games and control theory*. New York: Dekker; 1974. pp. 181–200.
22. Zelikin MI. On a differential game with incomplete information. *Sov Math Dokl* 1972;13:228–231.
23. Wilson DJ. Isaacs' princess and monster game on the circle. *J Optimiz Theor Appl* 1970;9:265–288.
24. Alpern S, Fokkink R, Lindelauf R, *et al.* The princess and monster game on an interval. *SIAM J Control Optimiz* 2008;47:1178–1190.
25. Bellman R. An optimal search problem. *SIAM Rev* 1963;5:274.
26. Beck A. On the linear search problem. *Isr J Math* 1964;2:221–228.
27. Beck A. More on the linear search problem. *Isr J Math* 1965;3:61–70.
28. Beck A, Beck M. The linear search problem rides again. *Isr J Math* 1986;53(3):365–372.
29. Bruce TF, Robertson JB. A survey of the linear-search problem. *Math Sci* 1988;13:75–89.
30. Beck A, Newman DJ. Yet More on the linear search problem. *Isr J Math* 1970;8:419–429.
31. Gal S. Minimax solution for linear search problems. *SIAM J Appl Math* 1974;27:17–30.
32. Borodin A, El-Yaniv R. *Online computation and competitive analysis*. Cambridge University Press; 1998.
33. Demaine ED, Fekete S, Gal S. Online searching with turn cost. *Theor Comput Sci* 2006;361:342–355.
34. Gal S, Chazan D. On the optimality of the exponential functions for some minimax problems. *SIAM J Appl Math* 1976;30:324–348.
35. Baeza-Yates RA, Culberson JC, Rawlins GJE. Searching in the plane. *Inf Comput* 1993;106(2):234–252.
36. Chrobak M, Kenyon-Mathieu C. Competitiveness via doubling. *SIGACT News* 2006;37(4):115–126.
37. Gal S. Minimax solutions for linear search problems. *SIAM J Appl Math* 1974;27:17–30.
38. Kao M, Reif JH, Tate SR. Searching an unknown environment: an optimal randomized algorithm for the cow-path problem. *Inf Comput* 1996;131(1):63–79.
39. Kella O. Star search—a different show. *Isr J Math* 1993;81:145–159.
40. Kao MY, Ma Y, Sipser M, *et al.* Optimal construction of hybrid algorithm. *J Algorithms* 1998;29:142–164.
41. Schuierer S. A lower bound for randomized searching on m rays. In: Klein R, Six HW, Wegner L, editors. *Volume 2598, Computer science in perspective: essays dedicated to Thomas Ottmann, Lecture notes computer science*. Berlin: Springer; 2003. pp. 264–277.

42. López-Ortiz A, Schuierer S. The ultimate strategy to search on m rays? *Theor Comput Sci* 2001;261(2):267–295.
 43. López-Ortiz A, Schuierer S. Online parallel heuristics, processor scheduling, and robot searching under the competitive framework. *Theor Comput Sci* 2004;310:527–537.
 44. Langetepe E. On the Optimality of Spiral Search. *Proceeding of SODA*. Philadelphia (PA): SIAM; 2010. pp. 1–12.
 45. Bellman R. Minimization problem. *Bull Am Math Soc* 1956;62:270.
 46. Isbell JR. An optimal search pattern. *Nav Res Log Q* 1957;4:357–359.
- FURTHER READING**
- Adler M, Racke H, Sivadasan N, *et al.* Randomized pursuit-evasion in graphs. *Comb Probab Comput* 2003;12(3):225–244.
- Ahlsweide R, Wegener I. *Search problems*. New York: Wiley; 1987.
- Chudnovsky DV, Chudnovsky GV, editors. *Search theory: some recent developments*. New York: Marcel Dekker.
- Foreman JG. The princess and the monster on the circle. In: *Differential games and control theory*. In: Roxin EO, Liu PT, Sternberg RL, editors. New York: Dekker; 1974. pp. 231–240.
- Foreman JG. Differential search game with mobile hider. *SIAM J Control Optimiz* 1977;15:841–856.
- Garnaev AY. *Search games and other applications of game theory*. Berlin: Springer; 2000.
- Haley BK, Stone LD, editors. *Search theory and applications*. New York: Plenum Press; 1980.
- Iida K. Volume 70, *Studies in the optimal search plan*, *Lecture Notes in Statistics*. Springer-Verlag; 1992.
- Ruckle WH. *Geometric games and their applications*. Boston (MA): Pitman; 1983.
- Finch S. *Lost in a forest*; 1999. Available at <http://www.mathsoft.com/asolve/forest/forest.html>.

SELECTIVE SUPPORT VECTOR MACHINES

ONUR SEREF

Department of Business Information
Technology, Virginia Polytechnic
Institute and State University,
Blacksburg, Virginia

Support vector machines (SVMs) are among the increasingly popular and widely used supervised machine learning methods [1]. SVMs are generally used to classify pattern vectors which are assumed to belong to two linearly separable sets of positive and negative labeled pattern vectors. The classification function is defined by a hyperplane that separates these two classes. SVM classifier finds the hyperplane that maximizes the distance between the two convex hulls formed by the pattern vectors in each class. SVM can be extended to classification of nonlinear data by implicitly embedding the original data in a nonlinear space using kernel functions [2]. SVM training is based on optimizing a quadratic convex function that is subject to linear constraints, which is a well-known problem with many efficient solutions [3]. Due to the strong statistical learning background and efficient implementations, SVM methods have found a wide spectrum of application areas recently, ranging from pattern recognition [4] and text categorization [5] to biomedicine [6–8] and financial applications [9,10].

A generalization of SVM classification is selective classification [11]. In selective classification, n sets of positive and negative labeled pattern vectors with t pattern vectors in each set are given. All pattern vectors in a set share the same label. The objective is to select a single pattern vector from each of the n sets such that the selected positive and negative pattern vectors produce the best possible solution for a binary classification problem. In the SVM context, this classification problem is the quadratic optimization problem that maximizes the margin between

positive and negative pattern vectors. The standard SVM problem can be considered as a special case of selective classification, where $t = 1$.

Selective classification is similar to the multiple-instance classification problem [12] with regard to its input. However there is a significant difference in the classification scheme. In selective classification, the selection process is symmetrical with respect to the positive and negative sets of pattern vectors with the objective of selecting a single pattern vector from each set. In multiple-instance classification, the objective is to classify positive and negative sets by classifying all pattern vectors in each set. According to the asymmetrical classification scheme in multiple-instance classification, a set is classified as positive if one or more pattern vectors in it are classified as positive whereas a set is classified negative only if all of the pattern vectors in it are classified as negative.

Selective classification poses a hard combinatorial optimization problem. It is shown that the SVM formulation for selective classification is NP-hard [13]. An alternative approach is developed based on a restricted amount of free slack that decreases the influence of misclassified pattern vectors on the separating hyperplane. The optimization problem in this approach is also a convex quadratic optimization problem with linear constraints, and can be kernelized for nonlinear classification problems. The solution to this problem has nice properties that allow simple and efficient algorithms to be developed. In one possible algorithm, the pattern with a small margin can be identified by providing a small amount of restricted free (unpenalized) slack, and such pattern vectors can be iteratively eliminated until one pattern vector remains in each set. In another algorithm, one can provide a larger amount of restricted free slack to identify and eliminate all but one pattern vector from each set in a single step, which would be equivalent to directly selecting the final pattern vector with a large margin in

each set. Both algorithms produce successful results in simulated data as well as real-life neural data from a visuomotor categorical discrimination task.

Selective SVM requires a basic understanding of the standard SVM formulation. The concepts to develop kernelized formulations for selective SVM through Lagrangian duality are similar to those involved in the standard SVM literature. Therefore, it is appropriate to start with a short review of SVMs in the next section.

SUPPORT VECTOR MACHINES

Let $\mathbf{X}$ be a set of pattern vectors $\mathbf{x}_i \in \mathbb{R}^d$ with class labels $y_i \in \{1, -1\}$ which either belong to the positive or negative labeled sets of pattern vectors. The problem of classifying these pattern vectors is equivalent to finding a function $f(\cdot)$ which correctly assigns a class label for a given pattern vector $\mathbf{x}$. A hyperplane separates positive and negative pattern vectors, and it is represented as $(\mathbf{w}, b)$, where $\mathbf{w}$ is the normal vector to the hyperplane and b is the bias. The distance γ between the hyperplane and the closest pattern vectors (support vectors) is called the *margin of the separating hyperplane*. This hyperplane is represented by the support vectors, which give the classifier its name, SVM. The hyperplane with the maximum margin is called the *maximum margin hyperplane*. Note that there is an inherent degree of freedom to represent the same hyperplane by multiplying $\mathbf{w}$ and b with $\lambda > 0$. Depending on the value of λ , the distance between $\mathbf{x}$ and the hyperplane changes. Assuming that this value is greater than a constant value, such as a unit distance of one, it is easy to show that the geometric margin becomes $1/\|\mathbf{w}\|$ [14]. The pattern vectors in the positive and negative classes may not be separable. In this case, introducing slack variables ξ_i for misclassified pattern vectors and penalizing them by $C/2$ for a chosen value of $C > 0$ leads to the following formulation:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{2} \sum_{i=1}^n \xi_i^2, \quad (1a)$$

$$\begin{aligned} \text{subject to: } & y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \\ & i = 1, \dots, n. \end{aligned} \quad (1b)$$

Although the formulation (1) finds a linear separating hyperplane, its *Lagrangian dual* can be modified to use nonlinear maps to embed the pattern vectors in a higher dimensional space in such a way that a hyperplane can separate the mapped pattern vectors in the embedded space. This embedding is done via the *kernel trick*. The mapping is defined over dot product *Hilbert spaces*. The Lagrangian dual of the soft margin classification problem provides a formulation, in which dot products can be introduced. The dot product can be replaced with kernels to introduce nonlinear mapping. The Lagrangian dual of formulation 1 is given as follows:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \\ & - \frac{1}{2C} \sum_{i=1}^n \alpha_i^2 \end{aligned} \quad (2a)$$

$$\text{subject to: } \sum_{i=1}^n y_i \alpha_i = 0, \quad (2b)$$

$$\alpha_i \geq 0, i = 1, \dots, n. \quad (2c)$$

The classification function can be written as $f(\mathbf{x}) = \text{sign}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b)$ for a pattern vector $\mathbf{x}$ with unknown class label. From complementary slackness, the bias b can be calculated using a pattern vector $\mathbf{x}_i$ which is also a support vector, that is $\alpha_i > 0$ as $b = y_i(1 - \alpha_i) - \sum_{i=1}^n y_i \alpha_i \langle \mathbf{w} \cdot \mathbf{x}_i \rangle$. One can use the equivalence $\mathbf{w} = \sum_{i=1}^n \sum_{k=1}^t y_i \alpha_i \mathbf{x}_i$ to calculate $\mathbf{w}$ from the solution to Equation (2).

The convenience of the dual formulation in Equation (2) is that it has only one constraint, and the dot products allow nonlinear kernels $K(\mathbf{x}_i, \mathbf{x}_j)$ to be incorporated in the formulation by replacing the dot product $\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$. For example, the Gaussian kernel is among the most commonly used kernels, and is given as follows:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\{ \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma} \right\}. \quad (3)$$

SELECTIVE SUPPORT VECTOR MACHINE FORMULATIONS

Selective SVM is a generalization of the standard SVM in which sets of pattern vectors replace individual pattern vectors. The main objective is to select a single pattern vector from each set such that the margin is maximized between the selected positive and negative pattern vectors. However this problem is highly combinatorial and finding the optimal solution is very hard even for small problems. Alternative efficient formulations such as a convex quadratic formulation that features a restricted amount of free slack can find good solutions.

Combinatorial Formulation

Consider that each pattern vector in a standard SVM problem is replaced with a set of pattern vectors sharing the same class label. Then, the selective margin maximization problem can be defined as follows: Let $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ be sets of pattern vectors with t pattern vectors $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,t}$ in each set $\mathbf{X}_i$. Let $y = \{y_1, \dots, y_n\}$ be the corresponding labels for each set $\mathbf{X}_i$ with each pattern vector $\mathbf{x}_{i,k}$ sharing the same label y_i . Select exactly one pattern vector $\mathbf{x}_{i,k}$ from each set $\mathbf{X}_i$ such that the margin between the selected positive and negative pattern vectors is maximized. This problem can be formulated as a quadratic mixed 0–1 programming problem as follows:

$$\min_{\mathbf{w}, b, \xi, v} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{2} \sum_{i=1}^n \sum_{k=1}^t \xi_{i,k}^2, \quad (4a)$$

subject to:

$$y_i(\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b) \geq 1 - \xi_{i,k} - M(1 - v_{i,k}), \quad (4b)$$

$$i = 1, \dots, n; k = 1, \dots, t;$$

$$\sum_{k=1}^t v_{i,k} = 1 \quad i = 1, \dots, n; \quad (4c)$$

$$v_{i,k} \in \{0, 1\} \quad i = 1, \dots, n; k = 1, \dots, t. \quad (4d)$$

Note that this formulation is similar to Equation (1), except for the extra term $M(1 - v_{i,k})$ in Equation (4b) and the new constraints (4c and 4d). M is a sufficiently

large positive number. Binary variables $v_{i,k}$ indicate whether the k th pattern vector from set i is selected or not. Note that when $v_{i,k} = 0$, the right side of Equation (4b) becomes sufficiently small such that the constraint is always satisfied, which is equivalent to removing that pattern vector. The constraint in Equation (4c) ensures that only one pattern vector is included from each set. It is clear that for sufficiently high penalty C , the selective SVM formulation can be considered as a *hard-selection* problem without the slack variables $\xi_{i,k}$, whose solution would provide a hyperplane that can completely separate the selected positive and negative pattern vectors.

Now, consider the following decision version of the problem. Let $\mathbf{X}_i$ denote a set of d -dimensional t pattern vectors. Assume that there are n such sets and all vectors $\mathbf{x}_{i,k}$ in each set $\mathbf{X}_i$ are labeled with the same label $y_i \in \{1, -1\}$. Let v^* denote a selection, where a single pattern vector $\mathbf{x}_{i,k^*}$ is selected from each set $\mathbf{X}_i$. Is there a selection v^* , in which all positive and negative pattern vectors can be separated by a hyperplane? This decision problem is shown to be NP-complete [13]. The reduction for the NP-completeness proof is from the PARTITION problem [15]. This result implies that the original combinatorial formulation (4) for the selective SVM formulation is NP-hard.

Alternative Formulation with Restricted Free Slack

The combinatorial problem is intractable even for small to medium-sized problems. An alternative convex quadratic formulation can be considered, in which a restricted amount of unpenalized slack is provided. Free slack gives the separating hyperplane some flexibility to adjust itself in order to diminish the influence of those pattern vectors with small or negative (misclassified) distance from the hyperplane. Given this flexibility, the hyperplane realigns itself with respect to the further pattern vectors with larger margin.

Consider a total, restricted, free slack amount of Υ , which is available for all pattern vectors. Note that a very small amount of free slack would make a very small difference compared to the hyperplane

obtained by the standard SVM formulation, whereas a very large free slack amount would yield trivial solutions. Depending on the selection scheme, the amount of total free slack may vary. The corresponding formulation is given as follows:

$$\min_{\mathbf{w}, b, \xi, v} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{c}{2} \sum_{i=1}^n \sum_{k=1}^n \xi_{i,k}^2, \quad (5a)$$

subject to:

$$y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_{i,k} - v_{i,k} \\ i = 1, \dots, n; k = 1, \dots, t; \quad (5b)$$

$$\sum_{i=1}^n \sum_{k=1}^n v_{i,k} \leq \Upsilon; \quad (5c)$$

$$v_{i,k} \geq 0 \quad i = 1, \dots, n; k = 1, \dots, t. \quad (5d)$$

Note that this formulation is similar to the standard SVM formulation with a convex quadratic objective function and linear constraints. The Lagrangian dual of this formulation can also be derived for nonlinear classification. The dual formulation is given as follows:

$$\max_{\alpha, \beta} \sum_{i=1}^n \sum_{k=1}^t \alpha_{i,k} - \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^t \sum_{j=1}^n \\ \times \sum_{l=1}^t y_i y_j \alpha_{i,k} \alpha_{j,l} \langle \mathbf{x}_{i,k} \cdot \mathbf{x}_{j,l} \rangle \\ - \frac{1}{2C} \sum_{i=1}^n \sum_{k=1}^t \alpha_{i,k}^2 - \Upsilon \beta \quad (6a)$$

subject to:

$$\sum_{i=1}^n \sum_{k=1}^t y_i \alpha_{i,k} = 0 \quad (6b)$$

$$0 \leq \alpha_i \leq \beta \quad i = 1, \dots, n; k = 1, \dots, t. \quad (6c)$$

The bias is equal to $b = y_i(1 - \alpha_{i,k}/C) - \langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle$, if there exist $v_{i,k} = 0$ for some i and k . Otherwise, $\alpha_{i,k} = \beta$ for all $i = 1, \dots, n$ and $k = 1, \dots, t$, in which case the bias becomes $b = (\sum_{(i,k) \in A} y_i)(|A|(1 - \beta/C) - V - \sum_{(i,k) \in A} \langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle)$, where $A = \{(i, k): \alpha_{i,k} > 0\}$. As in the standard SVM formulation, one can use the equivalence $\mathbf{w} = \sum_{i=1}^n \sum_{k=1}^t y_i \alpha_{i,k} \mathbf{x}_{i,k}$

to calculate $\mathbf{w}$ from the solution to Equation (6). For the nonlinear cases, linear dot products $(\mathbf{x}_{i,k} \cdot \mathbf{x}_{j,k})$ in Equation (6) can be replaced with an appropriate kernel $K(\mathbf{x}_{i,k}, \mathbf{x}_{j,k})$.

Free slack acquired by each pattern vector is either 0 or linearly proportional to its distance from the optimal hyperplane depending on the total slack provided. This result derives from the properties one can observe about the solutions to formulations (5 and 6). The first property is that constraint (5c) is always binding. In other words, total amount of free slack Υ is always consumed, which can be shown using complementary slackness. The next property is that if the penalized slack or the free slack is positive, that is $\xi_{i,k} + v_{i,k} > 0$, then the corresponding constraint (5b) is binding, which can be inferred from the optimality conditions. Using these properties, the following relationship between the penalized and free slack can be shown. Let $\xi_{\max} = \max_{i,k} \{\xi_{i,k}\}$, then $v_{i,k} = 0$ if $\xi_{i,k} < \xi_{\max}$ and $v_{i,k} > 0$ if $\xi_{i,k} = \xi_{\max}$.

This main result basically states that all pattern vectors with a functional margin of $d_{i,k} = y_i(\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b) < 1$ incur a penalty of $\xi_{i,k} = \min(1 - d_{i,k}, \xi_{\max})$. Free slack for pattern vectors with $\xi_{i,k} = \xi_{\max}$ is equal to $v_{i,k} = 1 - \xi_{\max} - d_{i,k}$, whose sum is equal to Υ . In other words, without loss of generality, the free slack acquired by a positive labeled pattern vector is linearly proportional to its distance from the hyperplane $\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b = 1 - \xi_{\max}$. Examples of such pattern vectors are shown in Fig. 1. Further details can be found in Ref. 13. This interpretation leads to a few possible methods to maximize the margin between the selected points.

SELECTION METHODS

The solution to the alternative formulation allows pattern vectors that are close to the hyperplane to use free slack and provide more flexibility for the separating hyperplane. The selection is done with respect to the orientation of the hyperplane. Compared to the hard combinatorial formulation with an exact solution, the alternative formulation can be referred to as *soft selection*. The two methods introduced in this section are based on the

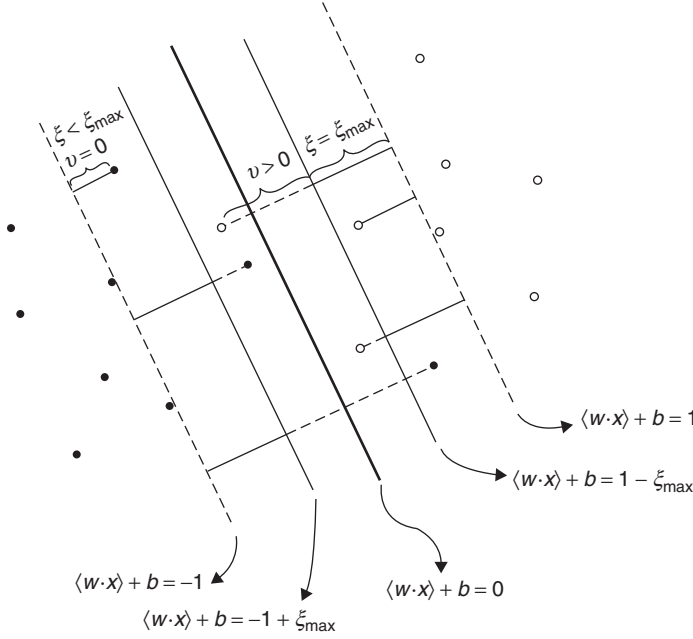


Figure 1. Example showing the relationship between penalized slack and free slack.

soft-selection formulation. The first method uses a small amount of restricted free slack for the soft-selection formulation and iteratively eliminates a single pattern vector with small or negative distance from the hyperplane from each set. The second method provides a larger amount of free slack to diminish the influence of multiple pattern vectors from each set and solves the soft-selection formulation only once to identify the pattern vector with the maximum distance within each set.

Iterative Elimination

The soft-selection formulation is mainly developed to give the separating hyperplane more flexibility and, at the same time, to

identify those pattern vectors which are misclassified or very close to the hyperplane. Such pattern vectors require more free slack among the pattern vectors within their set. One can obtain a more separated subset of positive and negative pattern vectors if such pattern vectors are removed. An intuitive basic approach is as follows: at each iteration, supply an incremental amount of free slack of n (1 unit per set), solve the soft-selection problem, identify the pattern vector with the minimum distance for each set and remove it, and repeat the iterations with the updated set of pattern vectors until only one pattern vector per set remains. This approach is summarized in Algorithm 1.

Algorithm 1. Iterative Elimination.

```

1:  $\mathbf{X} \leftarrow \mathbf{X}(0)$ 
2:  $t \leftarrow t(0)$ 
3: while  $t > 1$  do
4:    $\{\mathbf{w}, b\} \leftarrow \text{SOFT SELECTION}(\mathbf{X}, \mathbf{y}, n)$ 
5:    $r \leftarrow \emptyset$ 
6:   for  $i = 1$  to  $n$  do
7:      $k^* = \arg \min_{k=1, \dots, t} \{y_i (\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b)\}$ 
8:      $r = r \cup \mathbf{x}_{i,k^*}$ 
9:   end for

```

```

10:  $\mathbf{X} \leftarrow \mathbf{X} \setminus r$ 
11:  $t \leftarrow t - 1$ 
12: end while
13: return  $\mathbf{X}$ 

```

In Algorithm 1, $\mathbf{X}(0)$ is the original input of pattern vectors, $\mathbf{y}$ is the vector of labels for each set $\mathbf{X}_i$, n is the total free slack amount provided for the soft-selection problem, $(\mathbf{w}, b)$ is the separating hyperplane, $y_i(\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b)$ is the distance of $\mathbf{x}_{i,k}$ from the hyperplane, which can be negative if $\mathbf{x}_{i,k}$ is misclassified, $t(0)$ is the initial number of pattern vectors in each set, and r is the set of pattern vectors to be removed at each iteration.

Note that when total free slack is zero, the soft-selection problem reduces to a standard SVM problem. A naïve method can be considered, in which the total free slack is given as zero which reduces the soft-selection formulation to a standard SVM formulation. This naïve method can set a basis to compare the performance of the soft-selection formulation to the standard SVM formulation.

Direct Selection

The alternative to the iterative elimination method is to provide enough free slack to eliminate $t - 1$ pattern vectors in a single iteration. From another perspective, the extra free slack would diminish the influence of these $t - 1$ pattern vectors and let the separating hyperplane align itself with respect to the furthest pattern vectors within each set. Such pattern vectors can directly be selected after solving the soft-selection problem only once. For this algorithm, a total amount of $n(t - 1)$ free slack is provided. Direct selection is summarized in Algorithm 2 using a notation similar to Algorithm 1.

Algorithm 2. Direct Selection.

```

1:  $\{\mathbf{w}, b\} \leftarrow \text{SOFT\_SELECTION}(\mathbf{X}(0), \mathbf{y}, n(t - 1))$ 
2:  $\mathbf{X} \leftarrow \emptyset$ 
3: for  $i = 1$  to  $n$  do
4:    $k^* = \arg \max_{k=1, \dots, t} \{y_i(\langle \mathbf{w} \cdot \mathbf{x}_{i,k} \rangle + b)\}$ 
5:    $\mathbf{X} = \mathbf{X} \cup \mathbf{x}_{i,k^*}$ 
6: end for
7: return  $\mathbf{X}$ 

```

COMPUTATIONAL RESULTS

Soft-selection formulation and the two algorithms developed can be applied to simulated data as well as real-life data. The main basis of comparison for the simulated data is the naïve method for selection. The computational results included are provided to demonstrate the performance of the algorithms developed. The results regarding the first data set involve simulated data, in which the input pattern vectors are distributed uniformly in hyperspheres with respect to a number of parameters at varying levels. The second data set involves real-life data, in which electrical recordings from a macaque monkey brain are arranged as input for a selective classification task and the results are compared to those obtained with standard SVM classifiers.

Simulated Data and Performance Measure

The simulated data is generated using two parameters that determine the *dimensionality* and the *separability* of the pattern vectors. Let $S_k, k = 1, \dots, n$ denote the set of pattern vectors formed by including the k th pattern vector from each set $\mathbf{X}_i$. For each S_k, n random pattern vectors are generated, which are uniformly distributed in a hypersphere with radius r . The center of each hypersphere is also distributed uniformly in a hypersphere with radius c . Parameter r is kept constant so that c determines the separability of the data, which is the first parameter to be varied. The dimension of the data, denoted by d , is the second parameter.

The performance measure is the objective function value obtained by the standard SVM formulation for the final set of selected pattern vectors. For this purpose, the restricted total free slack can be set to zero, in which case the standard SVM formulation is obtained. The results are later normalized for each combination of dimension d and separability c with all of the results

obtained from the compared methods. The normalization is done by measuring the mean $\mu_{d,c}$ and the standard deviation $\sigma_{d,c}$ of the objective function values obtained from all of the compared methods, and normalizing each objective function value using $\mu_{d,c}$ and $\sigma_{d,c}$.

Iterative Elimination versus Naïve Elimination

Simulated data is generated for $d = 2, 4, \dots, 20$ and $c = 0, r/2, r$. Number of pattern vectors per set t is set to 6 and $p = 1$ unit of free slack per set is provided. For each combination of the parameters, 100 instances of simulated data sets are generated and tested using iterative elimination and naïve elimination, and the results are normalized. Let z_I and z_N denote the average normalized objective function values obtained from iterative elimination and naïve elimination. In Fig. 2, $z_N - z_I$ values for $d = 2, 4, \dots, 20$ are plotted for each c value. It is clear from the figure that as the dimensionality increases, the iterative elimination is significantly superior to the naïve elimination method. The difference becomes more apparent for higher levels of data separation. This result clearly shows

the success of iterative elimination due to the flexibility of the separating hyperplane incorporated by the restricted free slack.

Direct Selection

Simulated data is generated for $d = 2, 4, \dots, 20$ and $c = 0, r/2, r$. Number of pattern vectors per set t is set to 6 and $p = 1, \dots, 5$ units of free slack per set is provided. In Fig. 3, the effect of the increase in total slack is shown. The three graphs in the figure are in the order of increasing separation in the data. In each graph, the objective function values for the highest total slack parameter $p = 5$ is assumed to be the base value and the differences between the others and the base, $z_i - z_5, i = 1, \dots, 4$ are graphed. The amount of free slack does not contribute significantly for completely overlapping data in the graph 3a. However, it is clear from graph 3b that when there is some separability in the data, increasing amount of slack improves the performance of the method for higher dimensions. This difference is even more amplified in graph 3c for higher separability values. In graphs 3b and 3c, the increase in the difference between $p = 5$ and the others

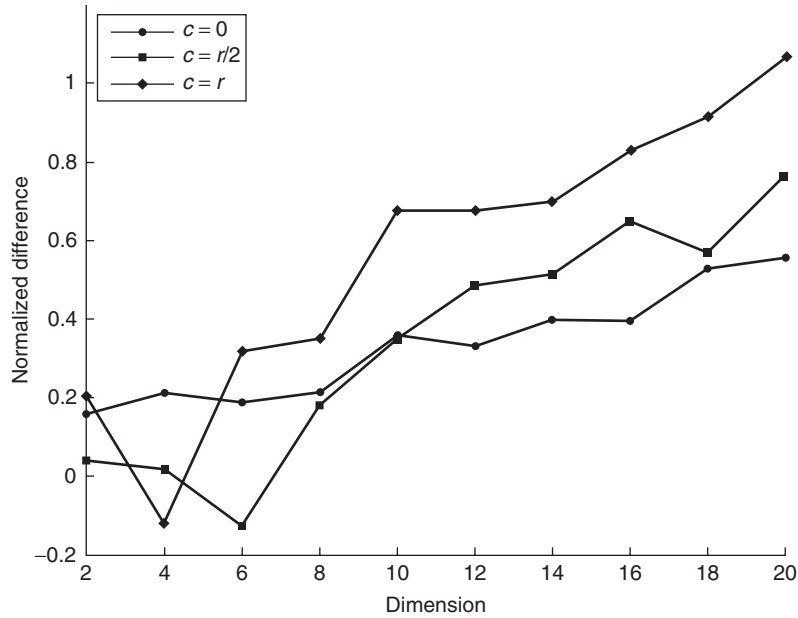


Figure 2. Normalized difference between iterative elimination and naïve elimination methods.

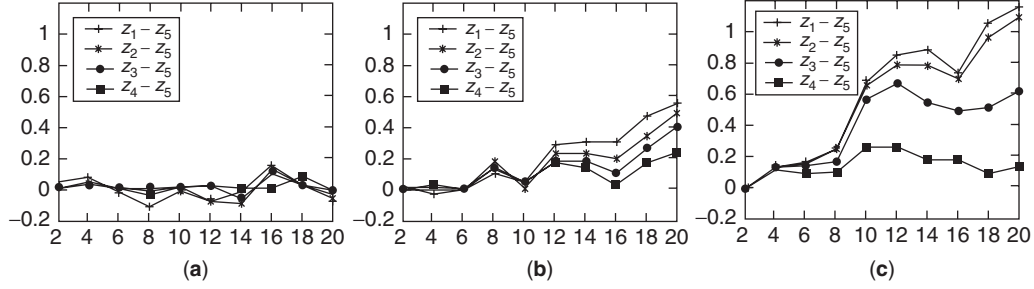


Figure 3. Effect of the amount of free slack on data with separability: (a) $c = 0$, (b) $c = r/2$, (c), $c = r$.

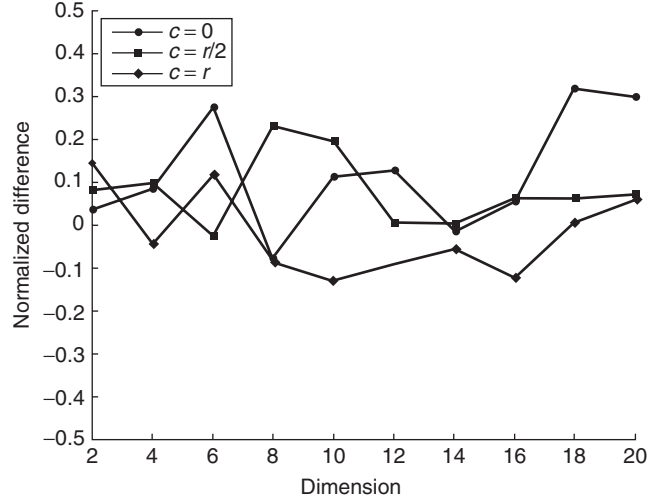


Figure 4. Comparison of iterative elimination and direct selection methods.

for higher dimensional data are also apparent. These results imply that the free slack parameter can be set as $t - 1$ (per set) for a data set with t pattern vectors in each set.

Comparison of the iterative elimination and direct selection completes the computational results on simulated data. The separability and dimensionality parameters are set in the same way as the previous computational results. Free slack parameter for direct selection is $p = t - 1$, where $t = 6$ again. In Fig. 4, the performances of iterative elimination and direct selection are shown with the values $z_D - z_I$, where z_D and z_I are normalized objective function values obtained from iterative elimination and direct selection methods, respectively. The results fluctuate and there is no significant

dominance of one method over the other. However, it can be observe from the figure that, on the average, the iterative elimination method performs slightly better than the direct elimination method.

A Real-Life Application

As an example of a real-life application, iterative elimination method is applied to a neuroscience problem. The data consists of local field potentials (LFPs) collected from multiple channels implanted in different cortical areas of a macaque monkey during a visual discrimination task. This task involves recognizing a visual “go” stimulus which is followed by a motor response. The visuomotor task is repeated multiple

times in different sessions for the same experiment with different stimuli–response combinations. Each repetition of the task is called a *trial*. The data can be arranged with respect to different stages of this task. The main objective is to be able to detect the differences between the different states of the macaque brain over the time course of the task by classifying the multidimensional and highly nonlinear neural data.

The visual stimuli are designed to create *lines* and *diamonds*. The “go” stimuli are chosen to be either lines or diamonds from one session to another. Three different stages are anticipated: (i) the detection of the visual stimulus, (ii) the categorical discrimination of the stimulus, and (iii) the motor response. The first and the third stages are relatively easy to detect; however, the second stage has not been detected in previous studies [16]. This stage involves a complex cognitive process whose onset and length vary over time.

Classification is performed with the pattern vectors collected at a specific time T^* from each trial. The classification accuracy obtained from each time point shows the time intervals when the two observed states of the monkey brain are different. However, there are temporal variations in each trial regarding the timing of the observed stages. The motivation behind the development of selective classification methods is to perform classification while accounting for these temporal variations in the underlying complex

cognitive processes. The standard SVM classifier is hindered by the noisy recordings due to these temporal variations. Given an appropriate time window, it can be assumed that one of the recordings within this window from each trial comes from the same underlying process to be detected. Selecting the most distinguishable recording from each trial at a given time window centered around T^* is a hard problem. Therefore one can use iterative elimination to detect and remove noisy recordings iteratively, to achieve better recordings for the given time window.

Nonlinear iterative elimination method is applied with a window of three recordings from each trial for each time point. This window corresponds to 15 ms. The recordings with the minimum distance are eliminated from each set, at each iteration. This procedure is repeated until only one pattern vector remains from each trial within the time window. The classification accuracy of the selected recordings from each time window is evaluated with the standard SVM classifier using 10-fold classification. In Fig. 5, the comparison of the classification accuracy results from iterative elimination, and the results from the standard SVM classification are given. The iterative elimination shows a consistent increase around 10%. This increase can be adjusted by the *baseline* approach. In order to create a baseline, class labels are randomly assigned to pattern vectors and iterative elimination is applied in order to

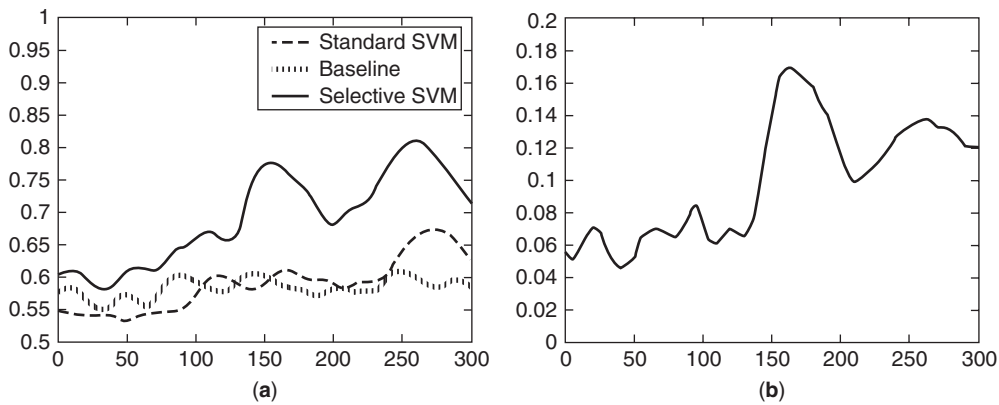


Figure 5. Comparative classification accuracy results: (a) standard SVM, baseline, and after applying selective SVM; (b) difference between the baseline and selective SVM results.

find the increase in the accuracy for random data and subtract it from the original accuracy results. The baseline is shown in Fig. 5a and the difference between the original accuracy results and the baseline results are given in Fig 5b. The peak around 160 ms in this graph is very clear. This result matches the anticipated interval of the categorical discrimination stage.

CONCLUSION

Selective SVMs are generalizations of the standard SVM classifiers. Sets of pattern vectors sharing the same label are given as input. One pattern vector is selected from each set in order to maximize the classification margin with respect to the selected positive and negative pattern vectors. The problem of selecting the best pattern vectors is referred to as the *hard-selection problem*, which is NP-hard. Soft-selection formulation is a tractable alternative which can be applied to linear and nonlinear data. Selectivity in the soft selection is controlled by the restricted free slack concept. The intuition behind this concept is to reverse the combinatorial selection problem by detecting influential pattern vectors which require free slack in order to decrease their influence on the separating hyperplane. One can find a selection with a large margin by iteratively removing such pattern vectors. The iterative elimination method is developed based on this scheme. Another alternative approach is to provide enough free slack to identify all $t - 1$ out of t pattern vectors to be removed at once, which leads to the direct selection method. Iterative elimination method using restricted free slack produces results superior to the case when no free slack is provided, in which case the classification is done using a standard SVM formulation. Iterative elimination and the direct selection methods produce comparable results.

The motivation for the development of selective classification methods comes from the classification of cognitive states in a visuomotor pattern discrimination task. Due to the temporal noise in the data, the classification results obtained are poor

with standard SVM methods. Sliding small time-windows of recordings from trials are considered as sets of pattern. Well-separated recordings are selected by the iterative elimination method. The selected recordings are evaluated with standard SVM methods, which result in a significant increase in the classification accuracy over the entire timeline of the task. The increase is adjusted by a baseline method which isolates the actual improvement peaks. These peaks clearly mark the categorical discrimination stage of the visuomotor task, which involves a complex cognitive process that has not been detected by previous studies. This result suggests that the proposed selective classification methods are capable of providing promising solutions for other classification problems in neuroscience.

REFERENCES

1. Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
2. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge: Cambridge University Press; 2004.
3. Bennet K, Campbell C. Support vector machines: Hype or hallelujah? SIGKDD Explor 2000;2(2):1–13.
4. Lee S, Verri A. Pattern recognition with support vector machines: SVM 2002. Niagara Falls: Springer; 2002.
5. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C, editors. Proceedings of the European conference on machine learning. Berlin: Springer; 1998.pp. 137–142.
6. Brown M, Grundy W, Lin D, *et al.* Knowledge-base analysis of microarray gene expression data by using support vector machines. PNAS 2000;97(1):262–267.
7. Noble WS. Kernel methods in computational biology. Support vector machine applications in computational biology. Cambridge: MIT Press; 2004. pp. 71–92.
8. Lal TN, Schroeder M, Hinterberger T, *et al.* Support vector channel selection in BCI. IEEE Trans Biomed Eng 2004;51(6):1003–1010.
9. Huang Z, Chen H, Hsu CJ, *et al.* Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis Support Syst 2004;37:543–558.

10. Trafalis TB, Ince H. Support vector machine for regression and applications to financial forecasting. International joint conference on neural networks (IJCNN'02). Como: IEEE-INNS-ENNS; 2002.
11. Seref O, Kundakcioglu OE, Pardalos PM. Selective linear and nonlinear classification. In: Pardalos PM, Hansen P, editors. Volume 45, Data mining and mathematical programming, series. CRM proceedings and lecture notes. United states of America: American Mathematical Society; 2008. pp. 211–234.
12. Dietterich TG, Lathrop RH, Lozano-Perez T. Solving the multiple instance problem with axis parallel rectangles. *Artif Intell* 1997;89:31–71.
13. Seref O, Kundakcioglu OE, Prokopyev OA, *et al.* Selective support vector machines. *J Comb Optim* 2009;17(1):3–20.
14. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
15. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-completeness. New York: W.H. Freeman; 1979.
16. Ledberg A, Bressler SL, Ding M, *et al.* Large-scale visuomotor integration in the cerebral cortex. *Cereb Cortex* 2007;17:44–62.

SELF-DUAL EMBEDDING TECHNIQUE FOR LINEAR OPTIMIZATION

KEES ROOS

Faculty of Electrical Engineering,
Mathematics and Computer Science,
Delft University of Technology, Delft,
The Netherlands

As in the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia, we consider the linear optimization (LO) problem in the standard form

$$\min \{c^T x : Ax = b, x \geq 0\}, \quad (\text{P})$$

with its dual problem

$$\max \{b^T y : A^T y + s = c, s \geq 0\}. \quad (\text{D})$$

Here, $A \in \mathbf{R}^{m \times n}$, $b, y \in \mathbf{R}^m$, $c, x, s \in \mathbf{R}^n$, and x, y, s are the vectors of variables. Without loss of generality we assume that $\text{rank}(A) = m$.

From the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia, it is clear that one can solve (P) and (D) by an interior-point method (IPM) if we know a triple (x, y, s) such that x is feasible for (P) and (y, s) is feasible for (D) and moreover this triple is close enough to the central path of (P) and (D). In general, such a triple is not a priori known and, even worse, it may not exist. The latter case occurs when one of the two problems is infeasible, or when both problems are feasible and the central path does not exist, that is, when the problems are not strictly feasible. This raises the question how IPMs can be used to solve an LO problem in such a case.

Recall that a similar issue arises when solving an LO problem with the simplex method. In order to solve (P) with the primal simplex method, one needs a vertex solution of the primal problem and, similarly, to solve (D) with the dual simplex method,

one needs a vertex solution of the dual problem. For the simplex method this problem has been resolved in several ways. The most common method is the so-called *two-phase method*. In phase 1, one solves an auxiliary problem which yields a vertex solution or detects infeasibility of the given problem. If phase 1 yields a vertex solution then phase 2 starts the simplex method at this vertex and either finds an optimal solution or detects that the problem is unbounded. In this way, a two-phase simplex method can solve any LO problem: it finds an optimal solution of the given problem or, if no optimal solution exists, it detects infeasibility or unboundedness. Other methods solve the problem in one phase (i.e., without first solving an auxiliary problem) by adding a term to the objective with a big coefficient M of an auxiliary variable that vanishes if the given problem has an optimal solution, provided that such a solution exists and M is large enough. For details, the reader could refer to the section titled "Optimization Models, Techniques, and Algorithms" in this encyclopedia. It might be mentioned that there also exists a one-phase simplex method that does not use an additional variable nor a big M . This so-called *criss-cross method* is less well-known, but interesting in itself. It uses both primal and dual pivoting operations, and solves a given problem in finitely many steps. For details, the reader could refer to Ref. 1.

The initialization problem for IPMs also has been resolved in several ways. Most common are so-called *infeasible start methods*, which are efficient and most popular. These methods start with an arbitrary infeasible triple (x^0, y^0, s^0) with $x^0 > 0$ and $s^0 > 0$ and reduce the infeasibility with about the same speed as the duality gap until both fall below some prescribed accuracy. They were developed and analyzed by Lustig [2], Kojima *et al.* [3], Mizuno [4], Mizuno *et al.* [5], Potra [6,7], Tanabe [8], Wright [9], Zhang [10]. Only few of these methods have polynomial iteration bounds, namely, Refs 4 and 10. The best of these bounds (in Ref. 4) is a factor $\sqrt{n}$ worse.

In this section, we present a different approach that was proposed first by Ye *et al.* [14] and is based on a homogeneous self-dual model introduced much earlier by Goldman and Tucker [15,16]. The given LO problem and its dual problem are embedded in such a model, while using two extra variables and two extra constraints. The model is such that it has a central path and moreover, the point corresponding to $\mu = 1$ on the central path is known. Starting from this point the model can be solved by any path-following IPM. Thus the generated optimal solution of the model yields in a simple way optimal solutions of the given problem and its dual if these exist. Otherwise, it yields a certificate that guarantees infeasibility of at least one of the two problems. Each iteration requires the solution of a system of linear equations, whose dimension is almost the same as for the primal-dual method in the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia, whereas the number of iterations is of the same order.

Notations

The set of real numbers is denoted as $\mathbf{R}$. $\mathbf{R}^n$, and $\mathbf{R}_+^n$ denote the set of real vectors of length n and its subset of nonnegative vectors, respectively. The zero vector and all-one vector in $\mathbf{R}^n$ are denoted as 0_n and e_n , respectively; if this gives no rise to confusion, we drop the subscript and simply write 0 or e . If f is a univariate function then, $f(x)$ is the vector with coordinates $f(x_i)$. This defines, for example, for any positive vector the vectors x^2 , x^{-1} , and $\sqrt{x}$. To any vector $x \in \mathbf{R}^n$, we associate the diagonal matrix X whose diagonal entries are the elements of x , in the same order. If also $s \in \mathbf{R}^n$, then Xs will be denoted shortly as xs . So xs is the so-called *Hadamard product*: its entries are obtained by componentwise multiplication of x and s . Finally, if s is positive then we denote xs^{-1} also as x/s .

REFORMULATION (P') OF (P)

The input data for the problems (P) and (D) consists of the triple (A, b, c) , where A is an

$m \times n$ matrix, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$. As before, we assume without loss of generality that $\text{rank}(A) = m$. First, we replace (P) by an equivalent problem that has only inequality constraints. In order to do this, we choose an arbitrary basis of A . So let B denote a subset of m indices in $\{1, \dots, n\}$ such that the columns of A indexed by these indices form a nonsingular $m \times m$ matrix. Since A has full row rank, this is possible. We denote the basis consisting of the columns of A indexed by B as A_B . Again, without loss of generality, we assume that A_B consists of the first m columns of A . Then we may write

$$A = [A_B \ A_N],$$

where A_N has size $m \times (n - m)$, with N consisting of the indices outside B . We then have

$$\begin{aligned} Ax = b &\Leftrightarrow A_B x_B + A_N x_N = b \\ &\Leftrightarrow x_B = A_B^{-1}(b - A_N x_N), \end{aligned} \quad (1)$$

where x_B denotes the restriction of x to its first m entries and x_N is the remaining part of x . Defining c_B and c_N in a similar way, we have

$$\begin{aligned} c^T x &= c_B^T x_B + c_N^T x_N \\ &= c_B^T A_B^{-1}(b - A_N x_N) + c_N^T x_N \\ &= c_B^T A_B^{-1} b + (c_N^T - c_B^T A_B^{-1} A_N) x_N. \end{aligned}$$

It follows that (P) can be reformulated as follows:

$$\begin{aligned} \min \{ &c_B^T A_B^{-1} b + (c_N^T - c_B^T A_B^{-1} A_N) x_N : \\ &A_B^{-1}(b - A_N x_N) \geq 0, x_N \geq 0 \}. \end{aligned}$$

Given an optimal solution x_N of this problem, an optimal solution $x = [x_B; x_N]$ of (P) is obtained by computing x_B from Equation (1). Omitting the constant term $c_B^T A_B^{-1} b$ in the objective, and redefining A , b , c , and x according to

$$\begin{aligned} A &:= -A_B^{-1} A_N, \quad b := -A_B^{-1} b, \\ c &:= (c_N^T - c_B^T A_B^{-1} A_N)^T = c_N^T - A_N^T A_B^{-T} c_B, \\ x &:= x_N, \end{aligned}$$

we can equally well solve the problem

$$\min\{c^T x : Ax \geq b, x \geq 0\}. \quad (P')$$

Now A has size $m \times (n - m)$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^{n-m}$. The dual problem is given by

$$\max\{b^T y : A^T y \leq c, y \geq 0\}. \quad (D')$$

In the next section, we embed this primal-dual pair of problems in an artificial self-dual LO problem.

EMBEDDING (P') AND (D') IN A SELF-DUAL PROBLEM (SP)

We introduce two new variables, κ and ϑ , and consider the following problem

$$\begin{aligned} \min \quad & \bar{n}\vartheta \\ \text{s.t.} \quad & Ax - b\kappa + \bar{b}\vartheta \geq 0, \\ & -A^T y + b^T y + c\kappa + \bar{c}\vartheta \geq 0, \\ & b^T y - c^T x + z\vartheta \geq 0, \\ & -\bar{b}^T y - \bar{c}^T x - \bar{z}\kappa \geq -\bar{n}, \\ & y \geq 0, x \geq 0, \kappa \geq 0, \vartheta \geq 0, \end{aligned} \quad (SP)$$

where

$$\begin{aligned} \bar{b} &= e_m + b - Ae_{n-m}, \quad \bar{c} = e_{n-m} + A^T e_m - c, \\ \bar{z} &= 1 - b^T e_m + c^T e_{n-m}, \quad \bar{n} = n + 2. \end{aligned}$$

Lemma 1. *If (SP) has a feasible solution $(y, x, \kappa, \vartheta)$ with $\kappa > 0$ and $\vartheta = 0$ then x/κ is optimal for (P') and y/κ for (D'). Otherwise, (P') and (D') have no optimal solutions.*

Proof. First, observe that if (P') has an optimal solution x^* then, by the Duality Theorem for an LO, (D') has also an optimal solution (y^*, s^*) and $b^T y^* = c^T x^*$ (see **LP Duality and KKT Conditions for LP**). Then, one may easily verify that $(y^*, x^*, 1, 0)$ satisfies the first three constraints in (SP). If $\bar{b}^T y^* + \bar{c}^T x^* + \bar{z} \leq 0$ then also the fourth constraint is satisfied. Otherwise, if $\bar{b}^T y^* + \bar{c}^T x^* + \bar{z} > 0$, we may find $\lambda > 0$ such that $\lambda(y^*, x^*, 1, 0)$ satisfies all constraints. Hence, we conclude that if (P') and (D') have optimal solutions then

(SP) has a feasible solution with $\vartheta = 0$ and $\kappa > 0$.

We now show that the converse is also true. Note that if $\kappa = 1$ and $\vartheta = 0$ then the first constraint and $x \geq 0$ represent primal feasibility and the second constraint and $y \geq 0$ represent dual feasibility. The third constraint then requires $b^T y - c^T x \geq 0$. However, since x is primal feasible and y dual feasible, we also have the converse inequality, because $c^T x \geq b^T y$ holds by the weak duality property. So it follows that $c^T x = b^T y$. Hence any feasible solution of (SP) with $\kappa = 1$ and $\vartheta = 0$ yields optimal solutions of (P') and (D'). A quite similar argument yields an even stronger result which is crucial for our purpose, namely that any feasible solution with $\kappa > 0$ and $\vartheta = 0$ yields optimal solutions of (P') and (D'). This easily follows from the fact that the first three constraints are homogeneous. Therefore, if $(y, x, \kappa, \vartheta = 0)$ is feasible for these constraints and $\kappa > 0$, then so is $(y/\kappa, x/\kappa, 1, \vartheta = 0)$, yielding that x/κ is optimal for (P') and y/κ for (D').

To proceed, we need to consider (SP) in more detail. Defining

$$M := \begin{bmatrix} 0 & A & -b & \bar{b} \\ -A^T & 0 & c & \bar{c} \\ b^T & -c^T & 0 & \bar{z} \\ -\bar{b}^T & -\bar{c}^T & -\bar{z} & 0 \end{bmatrix},$$

$$z := \begin{bmatrix} y \\ x \\ \kappa \\ \vartheta \end{bmatrix}, \quad q := \begin{bmatrix} 0 \\ 0 \\ 0 \\ \bar{n} \end{bmatrix},$$

we can reformulate (SP) as follows:

$$\min \left\{ q^T z : Mz \geq -q, z \geq 0 \right\}.$$

When writing down the dual of this problem, we get

$$\max \left\{ -q^T w : M^T w \leq q, w \geq 0 \right\}.$$

Now observe that the matrix M is skew-symmetric, that is, $M^T = -M$. Therefore, we may replace $M^T w \leq q$ by $-Mw \leq q$, which is the same as $Mw \geq -q$. This makes clear that (SP) and its dual have the same domain and the same objective. We conclude that the dual

problem is the same problem as (SP). This is expressed by saying that the problem (SP) is self-dual. Since the vector q is nonnegative and $z \geq 0$, we have $q^T z \geq 0$ showing that the objective values of (SP) are nonnegative. Since the zero vector is feasible for (SP), it follows that the optimal value is zero. We conclude that a feasible vector $z = (y, x, \kappa, \vartheta)$ is optimal if and only if $\vartheta = 0$. Combining this observation with Lemma 1, we obtain the following lemma.

Lemma 2. *If (SP) has an optimal solution $(y, x, \kappa, \vartheta)$ with $\kappa > 0$ then x/κ is optimal for (P) and y/κ for (D'). Otherwise, (P) and (D') have no optimal solutions.*

The next question is whether we can discover if an optimal solution of (SP) exists with $\kappa > 0$. As we show in the rest of this section, the answer is positive and any IPM can do this for us.

THE CENTRAL PATH OF (SP)

First, we need to show that the central path of (SP) exists. In order to do this, we need to write down the KKT system, as given in the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia, which determines the μ -center on the central path. To this end, we introduce a slack vector $s = Mz + q$ and write (SP) in the primal standard form, that is, as a minimization problem with equality constraints and nonnegative v

$$\min \left\{ \begin{bmatrix} q \\ 0 \end{bmatrix}^T \begin{bmatrix} z \\ s \end{bmatrix} : \begin{bmatrix} M & -I \end{bmatrix} \begin{bmatrix} z \\ s \end{bmatrix} = -q, \begin{bmatrix} z \\ s \end{bmatrix} \geq 0 \right\}.$$

With w denoting the vector of dual variables, the dual of this problem is given by

$$\max \left\{ -q^T w : \begin{bmatrix} -M \\ -I \end{bmatrix} w \leq \begin{bmatrix} q \\ 0 \end{bmatrix} \right\}.$$

Introducing slack vectors u and v for the dual constraints these get the form $-Mw + u = q$ and $-w + v = 0$. Now we can write down the KKT system for the current case, which

consists of the primal and dual feasibility conditions and the centering conditions. Thus, it becomes clear that the central path of (SP) exists if and only if the following system has a solution for at least one positive value of μ

$$\begin{aligned} Mz - s &= -q, & z \geq 0, s \geq 0, \\ -Mw + u &= q, & u \geq 0, \\ -w + v &= 0, & v \geq 0, \\ uz &= \mu e, \\ vs &= \mu e. \end{aligned}$$

Due to the third equation, we can eliminate w from the system. This yields the equivalent system

$$\begin{aligned} Mz - s &= -q, & z \geq 0, s \geq 0, \\ Mv - u &= -q, & u \geq 0, v \geq 0, \\ zu &= \mu e, \\ vs &= \mu e. \end{aligned}$$

Since the KKT system represents the first order optimality conditions for the primal logarithmic barrier function, and this function is strictly convex, the system has at most one solution. One easily sees that this implies that any solution will satisfy $v = z$ and $u = s$. Thus, we may conclude that the KKT system for (SP) has a solution if and only if the much simpler system

$$Mz - s = -q, \quad z \geq 0, s \geq 0, \quad (2)$$

$$zs = \mu e, \quad (3)$$

has a solution for some $\mu > 0$.

To prove the existence of the central path of (SP), it remains to show that this system has a solution for at least one positive value of μ . Due to the construction of (SP), this is now surprisingly easy. We claim that $z = e_{\bar{n}}$, $s = e_{\bar{n}}$, and $\mu = 1$ satisfy our goal. Indeed, taking $y = e_m$, $x = e_{n-m}$, $\kappa = 1$, and $\vartheta = 1$,

we get

$$\begin{aligned}
 Me_{\bar{n}} &= \begin{bmatrix} 0 & A & -b & \bar{b} \\ -A^T & 0 & c & \bar{c} \\ b^T & -c^T & 0 & \bar{z} \\ -\bar{b}^T & -\bar{c}^T & -\bar{z} & 0 \end{bmatrix} \begin{bmatrix} e_m \\ e_{n-m} \\ 1 \\ 1 \end{bmatrix} \\
 &= \begin{bmatrix} Ae_{n-m} - b + \bar{b} \\ -A^T e_m + c + \bar{c} \\ b^T e_m - c^T e_{n-m} + \bar{z} \\ -\bar{b}^T e_m - \bar{c}^T e_{n-m} - \bar{z} \end{bmatrix} = \begin{bmatrix} e_m \\ e_{n-m} \\ 1 \\ 1 - \bar{n} \end{bmatrix} \\
 &= e_{\bar{n}} - q,
 \end{aligned}$$

where we used the definitions of $\bar{b}$, $\bar{c}$, and $\bar{z}$. The following lemma summarizes the results of this section.

Lemma 3. (SP) has a central path and the μ -center is determined by the system (2)–(3). Moreover, $Me_{\bar{n}} + q = e_{\bar{n}}$, which implies that the 1-center is given by $s = z = e_{\bar{n}}$.

Of course, this lemma is of utmost importance. It means that we can solve (SP) by any IPM, because we can start any such method (primal, dual, or primal-dual) at the 1-center!

It might be no surprise that due to the self-duality the search directions in a primal-dual method for solving (SP) are obtained from a much simpler system than in the general case. They are determined by the following system:

$$M\Delta z - \Delta s = 0, \quad (4)$$

$$z\Delta s + s\Delta z = \mu e - zs. \quad (5)$$

One may easily check this by linearizing the KKT system for (SP) and its dual.

On the other hand, also due to self-duality, the primal method and the dual method are exactly the same methods. We leave it to the reader to verify that the search directions are then determined by the following system:

$$M\Delta z - \Delta s = 0, \quad (6)$$

$$(Z^{-2} + MS^{-2}M^T)\Delta z = z^{-1} - Ms^{-1} - \mu^{-1}q. \quad (7)$$

Note that we can solve Δz from Equation (7); since the matrix $Z^{-2} + MS^{-2}M^T$ is positive

definite the solution exists and is unique. Then Δs follows from Equation (6). Note that since $q = s - Mz$, the right-hand side expression in Equation (7) can be written as

$$z^{-1} - \mu^{-1}s - M(s^{-1} - \mu^{-1}z).$$

One easily verifies that this expression vanishes if $zs = \mu e$, that is, if $z = z(\mu)$ (and $s = s(\mu)$). This makes clear that at the μ -center, the primal search direction vanishes, as it should. In that case the primal direction coincides with the primal-dual direction, but otherwise they will (in general) be different.

We proceed with another important consequence of Lemma 3.

Lemma 4. If $z = (y, x, \kappa, \vartheta)$ is feasible for (SP) and $s = Mz + q$ then

$$e_{\bar{n}}^T z + e_{\bar{n}}^T s = \bar{n} + q^T z = \bar{n}(1 + \vartheta).$$

Proof. Since M is skew-symmetric, we have

$$a^T M a = 0, \quad \forall a \in \mathbf{R}^{\bar{n}}. \quad (8)$$

We call this the *orthogonality property*. By taking $a = e_{\bar{n}} - z$, it follows that

$$\begin{aligned}
 0 &= (e_{\bar{n}} - z)^T M(e_{\bar{n}} - z) = (e_{\bar{n}} - z)^T (Me_{\bar{n}} - Mz) \\
 &= (e_{\bar{n}} - z)^T (e_{\bar{n}} - s).
 \end{aligned}$$

Since $e_{\bar{n}}^T e_{\bar{n}} = \bar{n}$ and

$$z^T s = z^T (Mz + q) = q^T z = \bar{n}\vartheta, \quad (9)$$

this implies the lemma.

We just established that $q^T z = z^T s = \bar{n}\vartheta$. On the other hand, if $z = z(\mu)$ and $s = s(\mu)$ then Equation (3) implies $z^T s = \bar{n}\mu$. Thus, it follows that on the central path one has $\vartheta = \mu$. Since $\vartheta = z_{\bar{n}}$ and $z_{\bar{n}}s_{\bar{n}} = \mu$, it follows that $s_{\bar{n}} = 1$ along the central path.

OPTIMAL PARTITION OF (SP)

The question that still remains is whether we can decide if (SP) has an optimal solution with $\kappa > 0$. To answer this question, we need some extra notions and notations. First, for any $z \in \mathbf{R}^{\bar{n}}$, we define $s(z) = Mz + q$. Then z is feasible for (SP) if and only if z and $s(z)$ are nonnegative, and z is optimal if moreover $zs(z) = 0$. The latter means that z and $s(z)$ are complementary vectors, because for each index i , either $z_i = 0$ or $s_i(z) = 0$. If moreover for each i one of z_i and $s_i(z)$ is positive, we say that z and $s(z)$ are strictly complementary vectors, or shortly, that z is a strictly complementary optimal solution.

We now define index sets B and N as follows:

$$\begin{aligned} B &:= \{i : z_i > 0 \text{ for some optimal solution } z\}, \\ N &:= \{i : s_i(z) > 0 \\ &\quad \text{for some optimal solution } z\}. \end{aligned}$$

So, B contains all indices i for which an optimal solution z with positive z_i exists. We also write $z_i \in B$ if $i \in B$. Note that we certainly have $\vartheta \notin B$ because $\vartheta = 0$ for any optimal solution. The main question that we need to answer is whether $\kappa \in B$ holds or not. Because if $\kappa \in B$, then there exists an optimal solution z with $\kappa > 0$, in which case (P') and (D') have optimal solutions and not otherwise.

The next lemma implies that the sets B and N are disjoint.

Lemma 5. *Let z^1 and z^2 be feasible for (SP). Then z^1 and z^2 are optimal solutions of (SP) if $z^1 s(z^2) = z^2 s(z^1) = 0$.*

Proof. Due to the orthogonality property (8), we have

$$(z^1 - z^2)^T M(z^1 - z^2) = 0,$$

which implies

$$(z^1 - z^2)^T (s(z^1) - s(z^2)) = 0.$$

Expanding the product on the left and rearranging the terms, we get

$$\begin{aligned} (z^1)^T s(z^2) + (z^2)^T s(z^1) &= (z^1)^T s(z^1) \\ &\quad + (z^2)^T s(z^2). \end{aligned}$$

Now, z^1 is optimal if $(z^1)^T s(z^1) = 0$ and, similarly, z^2 is optimal if $(z^2)^T s(z^2) = 0$. Hence, since $z^1, z^2, s(z^1)$ and $s(z^2)$ are all nonnegative, z^1 and z^2 are optimal if

$$(z^1)^T s(z^2) + (z^2)^T s(z^1) = 0,$$

which is equivalent to

$$z^1 s(z^2) = z^2 s(z^1) = 0,$$

proving the lemma.

By a relatively old result of Goldman and Tucker [15], every LO problem has a strictly complementary optimal solution, which implies that $B \cup N = \{1, \dots, \bar{n}\}$. Owing to Lemma 5, the sets B and N are disjoint. Hence they form a partition of the index set $\{1, \dots, \bar{n}\}$. We call this the *optimal partition* of (SP). It should be clear now that the question whether there exists an optimal solution with $\kappa > 0$ is completely answered as soon as we know the optimal partition of (SP).

FINDING THE OPTIMAL PARTITION OF (SP)

We show, in this section, that if some feasible z is close enough to the μ -center, and μ is small enough, then we can recognize the optimal partition (B, N) from z . We also show that such a z can be found in a number of iterations that depends only on the dimension $\bar{n}$ and on a condition number σ_{SP} of (SP). The condition number is defined by

$$\sigma_{SP} := \min_{1 \leq i \leq \bar{n}} \max_{z \in SP^*} \{z_i + s_i(z)\},$$

where SP^* denotes the set of optimal solutions of (SP). Recall that if z is optimal then either $z_i = 0$ or $s_i(z) = 0$, for each i . For a strictly complementary solution, we have $z_i + s_i(z) > 0$ for each i . Hence, the maximum in the above definition is always positive. Therefore, being the minimum of a finite set of positive numbers, σ_{SP} is positive as well. Also, by Lemma 4, then $e^T z + e^T s(z) = \bar{n}$ for each $z \in SP^*$; this implies that the average value of $z_i + s_i(z)$ equals 1.

To proceed, we need a simple measure for the distance of any strictly feasible z to the central path. For that purpose, we define the number $\delta_c(z)$ as follows:

$$\delta_c(z) := \frac{\max(zs(z))}{\min(zs(z))}. \quad (10)$$

Observe that $\delta_c(z) = 1$ if $zs(z)$ is a multiple of the all-one vector e . Obviously, this occurs precisely if z lies on the central path. Otherwise, we have $\delta_c(z) > 1$. We consider $\delta_c(z)$ as an indicator for the “distance” of z to the central path.

Lemma 6. *Let z be a feasible solution of (SP) such that $\delta_c(z) \leq \tau$. Then, with $s = s(z)$, we have*

$$\begin{aligned} z_i &\geq \frac{\sigma_{SP}}{\tau \bar{n}}, \quad i \in B, & z_i &\leq \frac{z^T s}{\sigma_{SP}}, \quad i \in N, \\ s_i &\leq \frac{z^T s}{\sigma_{SP}}, \quad i \in B, & s_i &\geq \frac{\sigma_{SP}}{\tau \bar{n}}, \quad i \in N. \end{aligned}$$

Proof. From $\delta_c(z) \leq \tau$, we conclude that there exist positive numbers τ_1 and τ_2 such that $\tau \tau_1 = \tau_2$ and

$$\tau_1 \leq z_i s_i \leq \tau_2, \quad 1 \leq i \leq \bar{n}. \quad (11)$$

First, suppose that $i \in N$. Let $\tilde{z}$ be an optimal solution such that $\tilde{s}_i := s_i(\tilde{z})$ is maximal. Then $\tilde{z}_i = 0$ and the definition of σ_{SP} implies that $\tilde{s}_i \geq \sigma_{SP}$. Applying the orthogonality property to the points $\tilde{z}$ and z , we obtain $(\tilde{z} - z)^T(\tilde{s} - s) = 0$. Since $\tilde{z}$ is optimal, we have $\tilde{z}^T \tilde{s} = 0$. Hence we obtain

$$z^T \tilde{s} + s^T \tilde{z} = z^T s.$$

Since all involved vectors are nonnegative and $z_i \tilde{s}_i \leq z^T \tilde{s}$, we therefore may write

$$z_i \tilde{s}_i \leq z^T \tilde{s} \leq z^T s.$$

Hence, dividing by $\tilde{s}_i$ and using $\tilde{s}_i \geq \sigma_{SP} > 0$, we get

$$z_i \leq \frac{z^T s}{\sigma_{SP}}.$$

From the left inequality in Equation (11) we also have $z_i s_i \geq \tau_1$. Hence, it follows that

$$s_i \geq \frac{\tau_1 \sigma_{SP}}{z^T s}.$$

The right inequality in Equation (11) gives $z^T s \leq \bar{n} \tau_2$. Therefore

$$s_i \geq \frac{\tau_1 \sigma_{SP}}{\bar{n} \tau_2} = \frac{\sigma_{SP}}{\bar{n} \tau}.$$

This proves the second and fourth inequality in the lemma. The other inequalities are obtained in a similar way. If $i \in B$ and $\tilde{z}$ is an optimal solution such that $\tilde{z}_i$ is maximal, then $\tilde{z}_i \geq \sigma_{SP}$. Now applying the orthogonality property to the points $\tilde{z}$ and z , we obtain

$$s_i \tilde{z}_i \leq s^T \tilde{z} \leq z^T s.$$

Thus we get

$$s_i \leq \frac{z^T s}{\tilde{z}_i} \leq \frac{z^T s}{\sigma_{SP}}.$$

Using once more that $z_i s_i \geq \tau_1$ and $z^T s \leq \bar{n} \tau_2$, we obtain

$$z_i \geq \frac{\tau_1 \sigma_{SP}}{z^T s} \geq \frac{\tau_1 \sigma_{SP}}{\bar{n} \tau_2} = \frac{\sigma_{SP}}{\bar{n} \tau},$$

completing the proof of the lemma.

The results of the above lemma are summarized in Table 1.

We conclude that if $z^T s$ is so small that

$$\frac{z^T s}{\sigma_{SP}} < \frac{\sigma_{SP}}{\tau \bar{n}},$$

then we have a complete separation of the variables in B and the variables in N . Thus, we may state without further proof the following result.

Table 1. Estimates for the Variables in B and in N if $\delta_c(z) \leq \tau$

	$i \in B$	$i \in N$
z_i	$\geq \frac{\sigma_{SP}}{\tau \bar{n}}$	$\leq \frac{z^T s}{\sigma_{SP}}$
$s_i(z)$	$\leq \frac{z^T s}{\sigma_{SP}}$	$\geq \frac{\sigma_{SP}}{\tau \bar{n}}$

Lemma 7. *Let z be a feasible solution of (SP) such that $\delta_c(z) \leq \tau$. If*

$$z^T s(z) < \frac{\sigma_{SP}^2}{\tau \bar{n}},$$

then the optimal partition of (SP) follows from

$$B = \{i : z_i > s_i(z)\} \quad \text{and} \quad N = \{i : z_i < s_i(z)\}. \quad (12)$$

In the subsequent analysis, we need to find an upper bound τ for the values of $\delta_c(z)$ during the course of the algorithm. It can be shown that after each iteration of the primal and the dual method, one has $\delta_c(z) \leq (7 + \sqrt{7})/(7 - \sqrt{7}) \approx 2.215250$. Below we focus on the use of the primal-dual method. In that case, one has $\delta_c(z) \leq 4$ after each iteration, as follows from the next lemma. In this lemma, we use the vector v defined by

$$v := \frac{\sqrt{zs(z)}}{\mu} \quad (13)$$

and we use that after each iteration of the primal-dual IPM, as presented in the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia, one has $\delta(zs(z), \mu) \leq \frac{1}{2}$, where

$$\delta(zs(z), \mu) := \frac{1}{2} \|v^{-1} - v\|. \quad (14)$$

Lemma 8. *If $\delta(zs(z), \mu) \leq \frac{1}{2}$ then $\delta_c(z) \leq 4$.*

Proof. Using the vector v of z , with respect to the given $\mu > 0$, we have $zs(z) = \mu v^2$. Hence, we may write

$$\delta_c(z) = \frac{\max(\mu v^2)}{\min(\mu v^2)} = \frac{\max(v^2)}{\min(v^2)}.$$

Due to the definition (14) of $\delta(zs(z), \mu)$, it follows that

$$\|v - v^{-1}\| \leq 1.$$

Without loss of generality, we assume that the coordinates of v are ordered such that $v_1 \geq v_2 \geq \dots \geq v_n$. Then $\delta_c(z) = v_1^2/v_n^2$. Now consider the problem

$$\max \left\{ \frac{v_1^2}{v_n^2} : \|v - v^{-1}\| \leq 1 \right\}.$$

The optimal value of this problem is an upper bound for $\delta_c(z)$. One may easily verify that the optimal solution has $v_i = 1$ if $1 < i < \bar{n}$, $v_1 = \sqrt{2}$, and $v_{\bar{n}} = 1/\sqrt{2}$. Hence the optimal value is 4. This proves the lemma.

The above lemma is the basis of our next result, based on the use of the primal-dual algorithm. One easily derives a similar result for the primal method by using $\tau = 2.215250$.

Theorem 1. *After at most*

$$\left\lceil \sqrt{2\bar{n}} \log \frac{4\bar{n}^2}{\sigma_{SP}^2} \right\rceil \quad (15)$$

iterations the primal-dual algorithm yields a strictly feasible solution z of (SP) that reveals the optimal partition (B, N) of (SP) according to Equation (12).

Proof. Let us run the primal-dual algorithm with $\varepsilon = \sigma_{SP}^2/4\bar{n}$. Then, Theorem 3 in the article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia guarantees that we obtain a strictly feasible z with $z^T s(z) \leq \sigma_{SP}^2/4\bar{n}$ and $\delta(zs(z), \mu) \leq 1/2$. Lemma 8 implies that $\delta_c(z) \leq 4$. By Lemma 7, with $\tau = 4$, this z provides us the classes B and N of the optimal partition, according to (12). Using Theorem 3 in article titled **Central Path and Barrier Algorithms for Linear Optimization** in this encyclopedia again, we obtain that the required number of iterations for the given ε is at most

$$\left\lceil \sqrt{2\bar{n}} \log \frac{4\bar{n}^2}{\sigma_{SP}^2} \right\rceil,$$

which is precisely the bound given in the theorem. This completes the proof.

FINAL REMARKS

We conclude from Theorem 1 that the decision about the existence of an optimal solution with $\kappa > 0$ can be made safely if we run the primal-dual algorithm with $\varepsilon \leq \sigma_{SP}^2/(4\bar{n})$. The weakness of this result is that we do not know the condition number σ_{SP} . Its computation is at least as demanding

as solving (SP). It should be mentioned that if the entries of M are integral then we have the following easily computable lower bound for the condition number

$$\sigma_{SP} \geq \frac{1}{\prod_{j=1}^n \|M_j\|}, \quad (16)$$

where M_j denotes the j th column of M [17, Theorem I.42].

Another important consequence of the above analysis is that if ε is small enough, eventually the iterates converge to a strictly complementary solution of (SP). This follows from Table 1: the B -part of z and the N -part of $s(z)$ stay bounded away from zero. These iterate are still not exact solutions, because they remain strictly feasible. However, there exists a polynomial-time rounding procedure that yields an exact solution, provided that ε is small enough. For details, the reader could refer to Ref. 17, Section 3.3.5.

If $\kappa \in B$ then we know from Lemma 2 that, and how, optimal solutions of (P') and (D') can be obtained. If it turns out that $\kappa \in N$, then the same lemma implies that (P') and (D') do not have optimal solutions. We finally want to show that the output of the algorithm then provides a certificate for this fact.

If $\kappa \in N$, that is, if $\kappa = 0$, then the strict complementarity of z implies that the entry of $s(z)$ corresponding to κ belongs to B . This entry is the slack value of the third constraint in the original version of (SP), which means that this entry of $s(z)$ is given by $b^T y - c^T x$, since $\vartheta = 0$. So $b^T y - c^T x > 0$. It also follows from $\kappa = 0$ and $\vartheta = 0$ that $Ax \geq 0$ and $-A^T y \geq 0$. We conclude that if $\kappa \in N$ then we have x and y such that

$$\begin{aligned} b^T y - c^T x > 0, \quad Ax \geq 0, \quad A^T y \leq 0, \quad x \geq 0, \\ y \geq 0. \end{aligned} \quad (17)$$

The first inequality implies that either $b^T y > 0$ or $c^T x < 0$. If $b^T y > 0$ then (P') (and hence also (P)) is infeasible, since by assuming that $\bar{x}$ is primal feasible, we obtain

$$0 \geq \bar{x}^T A^T y = (A\bar{x})^T y = b^T y,$$

a contradiction. On the other hand, if $c^T x < 0$ then (D') is infeasible, since if $\bar{y}$ is dual feasible, we obtain

$$0 \leq \bar{y}^T Ax = (A^T \bar{y})^T x \leq c^T x,$$

again a contradiction.

Thus, we have shown that $\kappa \in N$ is a certificate for either (P') or (D') being infeasible. We know from the Duality Theorem for LO that if (P') and (D') do not have optimal solutions then there are three possibilities: (i) (P') infeasible and (D') unbounded, (ii) (P') unbounded and (D') infeasible, or (iii) (P') infeasible and (D') infeasible. Unfortunately, our certificate does not inform us about which of these three cases occur. It is still an open question whether a simple certificate exists that requires no additional computational effort and that yields information on which of the three possible cases occurs if $\kappa = 0$.

REFERENCES

1. Terlaky T. A convergent criss-cross method. *Optimization* 1985;16(5):683–690.
2. Lustig IJ. Feasible issues in a primal-dual interior-point method. *Math Program* 1990/91;67:145–162.
3. Kojima M, Megiddo N, Mizuno S. A primal-dual infeasible-interior-point algorithm for linear programming. *Math Program* 1993;61:263–280.
4. Mizuno S. Polynomiality of infeasible-interior-point algorithms for linear programming. *Math Program* 1994;67:109–119.
5. Mizuno S, Kojima M, Todd MJ. Infeasible-interior-point primal-dual potential-reduction algorithms for linear programming. *SIAM J Optim* 1995;5(1):52–67.
6. Potra FA. An infeasible-interior-point predictor-corrector algorithm for linear programming. *SIAM J Optim* 1996;6(1):19–32.
7. Potra FA. A quadratically convergent predictor-corrector method for solving linear programs from infeasible starting points. *Math Program* 1994;67(3, Ser. A):383–406.
8. Tanabe K. Centered Newton method for mathematical programming. Volume 113, *System modelling and optimization* (Tokyo, 1987), of *Lecture Notes in Control and Information Sciences*. Berlin: Springer; 1988. pp. 197–206.
9. Wright SJ. An infeasible-interior-point algorithm for linear complementarity problems. *Math Program* 1994;67(1):29–52.
10. Zhang Y. On the convergence of a class of infeasible-interior-point methods for the

- horizontal linear complementarity problem. *SIAM J Optim* 1994;4:208–227.
11. Vanderbei RJ. Linear programming: foundations and extensions. Boston (MA): Kluwer Academic Publishers; 1996.
 12. Wright SJ. Primal-dual interior-point methods. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM); 1997.
 13. Ye Y. Interior point algorithms: Wiley-Interscience Series in Discrete Mathematics and Optimization. New York: John Wiley & Sons, Inc.; 1997. Theory and analysis, A Wiley-Interscience Publication.
 14. Ye Y, Todd MJ, Mizuno S. An $O(\sqrt{n}L)$ -iteration homogeneous and self-dual linear programming algorithm. *Math Oper Res* 1994;19:53–67.
 15. Goldman AJ, Tucker AW. Theory of linear programming. In: Kuhn HW, Tucker AW, editors. Linear inequalities and related systems, *Annals of Mathematical Studies*, No. 38. Princeton (NJ): Princeton University Press; 1956. pp. 53–97.
 16. Tucker AW. Dual systems of homogeneous linear relations. In: Kuhn HW, Tucker AW, editors. Linear inequalities and related systems, *Annals of Mathematical Studies*, No. 38. Princeton (NJ): Princeton University Press; 1956. pp. 3–18.
 17. Roos C, Terlaky T, Vial J-P. Theory and algorithms for linear optimization. Chichester: Springer (Theory and Algorithms for Linear Optimization. An Interior-Point Approach. 1st ed. John Wiley & Sons; 1997); 2005.

SEMIDEFINITE OPTIMIZATION APPLICATIONS

ANTHONY MAN-CHO SO
Department of Systems Engineering
and Engineering Management,
The Chinese University of Hong
Kong, Hong Kong, China

INTRODUCTION

Recently, semidefinite programming (SDP) has received a lot of attention in many communities. Such a popularity can partly be attributed to the wide applicability of SDP, as well as recent advances in the design of provably efficient interior-point algorithms and fast heuristics. In this article, we will survey some of the recent applications of SDP and provide further pointers to the literature for interested readers. Given the versatile modeling power of SDPs and the multitude of work that applies the SDP methodology, we should emphasize that the applications discussed in this article are by no means exhaustive, and they necessarily reflect a certain degree of personal bias.

Before we begin our discussion on SDP applications, let us fix some notation. Let $\mathbb{S}^n$ be the set of $n \times n$ real, symmetric matrices, and let $\mathbb{S}_+^n \subset \mathbb{S}^n$ be the set of $n \times n$ real, symmetric, positive semidefinite matrices, that is, if $A \in \mathbb{S}_+^n$, then for any $x \in \mathbb{R}^n$, we have $x^T A x \geq 0$. We shall also write $A \succeq \mathbf{0}$ to denote the fact that $A \in \mathbb{S}_+^n$. For any $A, B \in \mathbb{S}^n$, we define the *trace inner product* $A \bullet B$ via

$$A \bullet B = \text{tr}(A^T B) = \sum_{i,j=1}^n A_{ij} B_{ij}.$$

Let $C, A_1, \dots, A_m \in \mathbb{S}^n$ and $b \in \mathbb{R}^m$. The optimization problem

$$\begin{aligned} & \inf \quad C \bullet X \\ \text{(P): subject to} \quad & A_i \bullet X = b_i \quad \text{for } 1 \leq i \leq m, \\ & X \succeq \mathbf{0} \end{aligned}$$

is called an *SDP in primal standard form*. Its dual problem, which is given by

$$\begin{aligned} & \sup \quad b^T y \\ \text{(D): subject to} \quad & \sum_{i=1}^m y_i A_i + Z = C, \\ & y \in \mathbb{R}^m, Z \succeq \mathbf{0}, \end{aligned}$$

is also an SDP and is called an *SDP in dual standard form*. Theoretically, it has been shown that both (P) and (D) can be efficiently solved (up to any desired degree of accuracy) by the ellipsoid method or interior-point algorithms [1,2]. Many techniques have also been recently developed to speed up the solution time of SDPs [3–5].

In the sequel, we shall see how (P) and (D) can be used to model a wide range of problems that arise in practice.

SEMIDEFINITE RELAXATIONS OF HARD QUADRATIC OPTIMIZATION PROBLEMS

The Basic Technique

A frequently used and powerful idea for dealing with hard optimization problems is to develop tractable convex relaxations of them. As it turns out, many difficult quadratic optimization problems can be relaxed to SDPs. To illustrate some of the basic ideas, let us consider the following class of homogeneous *quadratically constrained quadratic programs* (QCQPs):

$$\begin{aligned} & \text{minimize} \quad x^T A x \\ \text{(QCQP): subject to} \quad & x^T A_i x \succeq_i b_i, \\ & x \in \mathbb{R}^n, \\ & \text{where } i = 1, \dots, m. \end{aligned}$$

Here, we assume that $A, A_1, \dots, A_m \in \mathbb{S}^n$ and $b_1, \dots, b_m \in \mathbb{R}$, and “ $\succeq_i$ ” can represent either “ $\geq$,” “ $=$,” or “ $\leq$,” where $i = 1, \dots, m$. Problem (QCQP) is known to be NP-hard, as it captures various NP-hard problems as special cases [6–8]. To derive an SDP relaxation of (QCQP), we first observe that

$$x^T A x = A \bullet x x^T \quad \text{and} \quad x^T A_i x = A_i \bullet x x^T \\ \text{for } i = 1, \dots, m.$$

In particular, both the objective function and constraints in (QCQP) are *linear* in the matrix $x x^T$. Thus, by introducing a new variable $X = x x^T$ that denotes a rank one symmetric positive semidefinite matrix, we obtain the following *equivalent* formulation of (QCQP):

$$\begin{aligned} (\text{QCQP-RC}) : \\ \text{minimize} \quad & A \bullet X \\ \text{subject to} \quad & A_i \bullet X \geq b_i, \\ & X \succeq 0, \quad \text{rank}(X) = 1, \\ \text{where } i = & 1, \dots, m. \end{aligned}$$

The upshot of the above equivalent formulation is that it allows us to identify the fundamental difficulty in solving (QCQP). Indeed, in (QCQP-RC), the objective function and all constraints except the $\text{rank}(X) = 1$ are convex in X . Thus, we may consider dropping the trouble-causing rank constraint to obtain the following so-called *semidefinite relaxation* of (QCQP):

$$\begin{aligned} (\text{QCQP-SDR}) : \quad & \text{minimize} \quad A \bullet X \\ & \text{subject to} \quad A_i \bullet X \geq b_i, \\ & \quad \quad \quad X \succeq 0, \\ & \text{where } i = 1, \dots, m. \end{aligned}$$

It is not hard to see that (QCQP-SDR) is an instance of SDP, and hence can be solved efficiently. However, a fundamental issue that one must address is how to convert a globally optimal solution X^* to (QCQP-SDR) into a feasible solution $\tilde{x}$ to (QCQP). Note that if X^* is of rank 1, then there is nothing to do, for we can write $X^* = x^*(x^*)^T$, and x^* will be a feasible (and in fact optimal) solution to (QCQP). On the other hand, if the rank of X^* is larger than 1, then we must somehow extract from it, in an efficient manner, a vector $\tilde{x}$ that is feasible for (QCQP). There are many ways to do this. However, we must emphasize that even though the extracted solution is feasible for (QCQP), it is in general not an optimal solution (for otherwise

we would have solved an NP-hard problem in polynomial time).

We note that the above procedure has been used extensively, for example, in the signal processing community to tackle various communications problems. We refer the interested readers to Refs 9 and 10 for further details and references.

Quality of Semidefinite Relaxations

Given the SDP relaxations of various QCQP problems, it is natural to ask how good those relaxations are. In their seminal work, Goemans and Williamson [6] analyzed an SDP relaxation for the Maximum Cut (MAX-CUT) problem and provided the first provable result concerning the quality of SDP relaxations. Furthermore, it sparked off a great deal of very fruitful research on the quality of various SDP relaxations. Now, let us give an overview of Goemans and Williamson's approach. Such an approach has also been used to establish some of the latter results on the quality of SDP relaxations.

One of the motivations for Goemans and Williamson's work is the MAX-CUT Problem, which is a well-known, NP-hard combinatorial optimization problem and is defined as follows. Suppose that we are given a simple undirected graph $G = (V, E)$ with $V = \{1, \dots, n\}$, and a function $w : E \rightarrow \mathbb{R}_+$ that assigns to each edge $e \in E$ a nonnegative weight w_e . The MAX-CUT problem is that of finding a set $S \subset V$ of vertices such that the total weight of the edges in the cut $(S, V \setminus S)$, that is, sum of the weights of the edges with one end point in S and the other in $V \setminus S$, is maximized. By setting $w_{ij} = 0$ if $(i, j) \notin E$, we may denote the weight of a cut $(S, V \setminus S)$ by

$$w(S, V \setminus S) = \sum_{i \in S, j \in V \setminus S} w_{ij}, \quad (1)$$

and our goal is to choose a set $S \subset V$ such that the quantity in Equation (1) is maximized. The MAX-CUT problem can be formulated as

the following QCQP:

$$\begin{aligned} v^* = \text{maximize} \quad & \frac{1}{2} \sum_{(i,j) \in E} w_{ij}(1 - x_i x_j) \\ \text{subject to} \quad & x_i \in \{-1, 1\} \text{ for } i = 1, \dots, n, \end{aligned} \quad (2)$$

where the variable x_i indicates which side of the cut vertex i belongs to. Specifically, the cut $(S, V \setminus S)$ is given by $S = \{i \in \{1, \dots, n\} : x_i = 1\}$. Note that if vertices i and j belong to the same side of a cut, then $x_i = x_j$, whence its contribution to the objective function in problem (2) is zero. Otherwise, we have $x_i \neq x_j$, whence its contribution to the objective function is $w_{ij}(1 - (-1))/2 = w_{ij}$. Upon applying the techniques described in the section titled “The Basic Technique,” we obtain the following SDP relaxation of problem (2)

$$\begin{aligned} v_{sdp}^* = \text{maximize} \quad & \frac{1}{2} \sum_{(i,j) \in E} w_{ij}(1 - X_{ij}) \\ \text{subject to} \quad & X_{ii} = 1, \\ & X \succeq \mathbf{0}, \\ \text{where } & i = 1, \dots, n. \end{aligned} \quad (3)$$

Note that since problem (3) is a relaxation of problem (2), we have $v_{sdp}^* \geq v^*$.

Now, let X^* be an optimal solution to problem (3). In general, the matrix X^* need not be of the form xx^T , and hence it does not immediately yield a feasible solution to problem (2). However, we can extract from X^* a solution $x' \in \{-1, 1\}^n$ to problem (2) via the following randomized rounding procedure:

1. Compute the *Cholesky factorization* $X^* = U^T U$ of X^* , where $U \in \mathbb{R}^{n \times n}$. Let $u_i \in \mathbb{R}^n$ be the i th column of U . Note that $\|u_i\|_2^2 = 1$ for $i = 1, \dots, n$.
2. Let $r \in \mathbb{R}^n$ be a vector uniformly distributed on the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$.
3. Set $x'_i = \text{sgn}(u_i^T r)$ for $i = 1, \dots, n$, where

$$\text{sgn}(z) = \begin{cases} 1 & \text{if } z \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

In other words, we choose a random hyperplane through the origin (with r as its normal) and partition the vertices according to whether their corresponding vectors lie “above” or “below” the hyperplane.

It is clear that $x' = (x'_1, \dots, x'_n)$ is a feasible solution to problem (2), and hence we must have

$$v' \equiv \frac{1}{2} \sum_{(i,j) \in E} w_{ij}(1 - x'_i x'_j) \leq v^*.$$

The question now is whether there exists an $\alpha \in (0, 1)$ such that $v' \geq \alpha \cdot v^*$. The factor α is commonly known as the *approximation ratio*. More precisely, since the solution $x' \in \{-1, 1\}^n$ is produced via a randomized procedure, we are interested in its expected objective value, that is,

$$\begin{aligned} \mathbb{E}[v'] &= \frac{1}{2} \mathbb{E} \left[\sum_{(i,j) \in E} w_{ij} (1 - x'_i x'_j) \right] \\ &= \frac{1}{2} \sum_{(i,j) \in E} w_{ij} \mathbb{E} [1 - x'_i x'_j] \\ &= \sum_{(i,j) \in E} w_{ij} \Pr [\text{sgn}(u_i^T r) \neq \text{sgn}(u_j^T r)]. \end{aligned} \quad (4)$$

The following theorem provides a lower bound on $\mathbb{E}[v']$ and allows us to compare it with v_{sdp}^* :

Theorem 1. *Let $u, v \in S^{n-1}$, and let r be a vector uniformly distributed on S^{n-1} . Then, we have*

$$\Pr [\text{sgn}(u^T r) \neq \text{sgn}(v^T r)] = \frac{1}{\pi} \arccos(u^T v).$$

Furthermore, for any $z \in [-1, 1]$, we have

$$\frac{1}{\pi} \arccos(z) \geq \alpha \cdot \frac{1}{2}(1 - z) > 0.878 \cdot \frac{1}{2}(1 - z),$$

where

$$\alpha = \min_{0 \leq \theta \leq \pi} \frac{2\theta}{\pi(1 - \cos \theta)}.$$

Table 1. Quality Analyses of Various SDP Relaxations

Type of Problems	References
Combinatorial optimization	[6, 11–23]
Quadratic optimization	[7, 8, 24–36]
Communications	[37–41]
Geometry	[42–45]

As a corollary to Theorem 1, we have the following:

Corollary 1. *Given an instance (G, w) of the MAX-CUT problem and an optimal solution to problem (3), the randomized rounding procedure above will produce a cut $(S', V \setminus S')$ whose expected objective value satisfies $w(S', V \setminus S') \geq 0.878v^*$.*

Proof. Let x' be the solution obtained from the randomized rounding procedure, and let S' be the corresponding cut. By Equation (4) and Theorem 1, we have

$$\begin{aligned}
\mathbb{E}[w(S', V \setminus S')] &= \frac{1}{\pi} \sum_{(i,j) \in E} w_{ij} \cdot \arccos(u_i^T u_j) \\
&\geq 0.878 \cdot \frac{1}{2} \sum_{(i,j) \in E} w_{ij} (1 - u_i^T u_j) \\
&= 0.878v_{sdp}^* \\
&\geq 0.878v^*,
\end{aligned}$$

as desired.

Note that a crucial step in the above analysis is to study a certain probability question related to the projections of a collection of vectors onto a random vector. This turns out to be a recurring theme in many of the quality analyses of SDP relaxations. Table 1 provides pointers to further theoretical results on the quality of SDP relaxations that arise in various applications.

SEMIDEFINITE RELAXATIONS OF POLYNOMIAL OPTIMIZATION PROBLEMS

A fundamental problem in optimization is that of finding the global minimum of a

real-valued polynomial $p : \mathbb{R}^n \rightarrow \mathbb{R}$ subject to polynomial inequality constraints. Indeed, the class of polynomial functions provides a very powerful modeling tool, and many optimization problems can be formulated as polynomial optimization problems. One example is the MAX-CUT problem in the section titled “Quality of Semidefinite Relaxations,” where the constraint $x_i \in \{-1, 1\}$ can be equivalently formulated as $x_i^2 = 1$. Given its importance, polynomial optimization has been studied intensely in recent years. One approach, which is proposed by Lasserre [46] (see also the excellent survey by Laurent [47]), is to construct a *sequence* of successively tighter SDP relaxations of the given polynomial optimization problem, in such a way that the corresponding optimal values are monotone and converge to the optimal value of the original problem. To fix ideas and simplify notation, let us again focus on the case of *quadratic* optimization problems [48]. To begin, consider the problem

$$\begin{aligned}
(\text{PO}) : \quad v^* &= \text{minimize } g_0(x) \\
&\text{subject to } g_i(x) \geq 0 \\
&\text{for } i = 1, \dots, m,
\end{aligned}$$

where $g_0, g_1, \dots, g_m : \mathbb{R}^n \rightarrow \mathbb{R}$ are quadratic polynomials. Note that we may write the polynomials $g_0, g_1, \dots, g_m$ in *multiindex notation*, namely,

$$g_i(x) = \sum_{\alpha} (g_i)_{\alpha} x^{\alpha} \quad \text{for } i = 0, 1, \dots, m, \quad (5)$$

where $\alpha = (\alpha_1, \dots, \alpha_n) \geq \mathbf{0}$, $x^{\alpha} = x_1^{\alpha_1} \dots x_n^{\alpha_n}$, and $\sum_{i=1}^n \alpha_i \leq 2$. Hence, we may identify the polynomial g_i with its coefficient vector $((g_i)_{\alpha}) \in \mathbb{R}^{(n+1)(n+2)/2}$, where $i = 0, 1, \dots, m$.

Now, observe that each summand in Equation (5) is either a constant or of the form x_i or $x_i x_j$, where $1 \leq i \leq j \leq n$. For terms of the form $x_i x_j$, we may linearize them by introducing the variables y_{ij} . As a result, we have the following *linear programming (LP) relaxation* of (PO):

$$\begin{aligned}
&\text{minimize } \sum_{0 \leq i \leq j \leq n} (g_0)_{ij} y_{ij} \\
&\text{subject to } \sum_{0 \leq i \leq j \leq n} (g_k)_{ij} y_{ij} \geq 0 \quad \text{for } k = 1, \dots, m,
\end{aligned}$$

with $y_{00} = 1$ and $y_{0i} = x_i$ for $i = 1, \dots, n$. To tighten the relaxation, observe that ideally we should have $y_{ij} = y_{0i}y_{0j}$ for $1 \leq i \leq j \leq n$. Hence, we may impose the constraint

$$M_1(\mathbf{y}_1) \equiv \begin{bmatrix} 1 & y_{01} & y_{02} & \cdots & y_{0i} & \cdots & y_{0n} \\ y_{01} & y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{0j} & y_{1j} & y_{2j} & \cdots & y_{ij} & \cdots & y_{jn} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{0n} & y_{1n} & y_{2n} & \cdots & y_{in} & \cdots & y_{nn} \end{bmatrix} \succeq \mathbf{0},$$

where $\mathbf{y}_1 = (1, y_{01}, \dots, y_{0n}, y_{11}, \dots, y_{nn})$. The first SDP relaxation in the sequence can then be given by

$$\begin{aligned} (\text{PO-SDR1}) : \\ v_1^* &= \inf \sum_{0 \leq i \leq j \leq n} (g_0)_{ij} y_{ij} \\ \text{subject to } &\sum_{0 \leq i \leq j \leq n} (g_k)_{ij} y_{ij} \geq 0, \\ &M_1(\mathbf{y}_1) \succeq \mathbf{0}, \end{aligned}$$

where $k = 1, \dots, m$.

The reader may notice that (PO-SDR1) is simply the standard SDP relaxation of a quadratic optimization problem (see section titled “The Basic Technique”).

To obtain the next SDP relaxation in the sequence, consider a particular constraint

$$\sum_{0 \leq i \leq j \leq n} (g_k)_{ij} y_{ij} \geq 0,$$

where $k = 1, \dots, m$. If we multiply the left-hand side by y_{pq} (where $0 \leq p \leq q \leq n$) and linearize, we obtain

$$g_k y_{pq} = \sum_{0 \leq i \leq j \leq n} (g_k)_{ij} y_{ij} y_{pq}$$

(note that y_{pq} is an entry of $\mathbf{y}_1$). Now, it can be shown [46, 48] that the linear matrix inequality $M_1(g_k \mathbf{y}_1) = [g_k y_{pq}]_{p,q} \succeq \mathbf{0}$ is a valid constraint for the original problem (PO). To further tighten the relaxation, we impose

constraints on the terms y_{ijpq} via

$$\begin{aligned} M_2(\mathbf{y}_2) \\ \equiv \begin{bmatrix} 1 & y_{01} & \cdots & y_{0n} & y_{11} & \cdots & y_{nn} \\ y_{01} & y_{11} & \cdots & y_{1n} & y_{111} & \cdots & y_{1nn} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{0n} & y_{1n} & \cdots & y_{nn} & y_{11n} & \cdots & y_{nnn} \\ y_{11} & y_{111} & \cdots & y_{11n} & y_{1111} & \cdots & y_{11nn} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{nn} & y_{1nn} & \cdots & y_{nnn} & y_{11nn} & \cdots & y_{nnnn} \end{bmatrix} \succeq \mathbf{0}, \end{aligned}$$

where $\mathbf{y}_2 = (1, y_{01}, \dots, y_{nnnn})$. Then, the second SDP relaxation in the sequence is given by

$$\begin{aligned} (\text{PO-SDR2}) : \\ v_2^* &= \inf \sum_{0 \leq i \leq j \leq n} (g_0)_{ij} y_{ij} \\ \text{subject to } &M_1(g_k \mathbf{y}_1) \succeq \mathbf{0}, \\ &M_2(\mathbf{y}_2) \succeq \mathbf{0}, \\ &\text{where } k = 1, \dots, m. \end{aligned}$$

In general, the l th SDP relaxation in the sequence, which is given by

$$\begin{aligned} (\text{PO-SDR})_l : \\ v_l^* &= \inf \sum_{0 \leq i \leq j \leq n} (g_0)_{ij} y_{ij} \\ \text{subject to } &M_{l-1}(g_k \mathbf{y}_{l-1}) \succeq \mathbf{0} \\ &M_l(\mathbf{y}_l) \succeq \mathbf{0}, \\ &\text{for } k = 1, \dots, m, \end{aligned}$$

can be obtained by repeating the above procedure.

The following proposition states that the sequence of optimal values $\{v_l^*\}_l$ is monotone and they all provide lower bounds on the optimal value v^* of the original problem (PO).

Proposition 1. *For $l = 1, 2, \dots$, we have $v_l^* \leq v_{l+1}^* \leq v^*$.*

To prove Proposition 1, it suffices to verify that for $l = 1, 2, \dots$, the feasible region of problem (PO) is contained in that of problem (PO-SDR($l + 1$)), which in turn is contained

in the feasible region of (PO–SDR l). We refer the reader to Refs 46, 48 for details.

Although Proposition 1 implies that a better lower bound on v^* can be obtained by solving a higher order SDP relaxation (PO–SDR l) (where $l = 1, 2, \dots$), it is not clear from Proposition 1 whether the sequence of optimal values $\{v_l^*\}_l$ would converge to v^* . However, if the feasible region of (PO) is compact, then it can be shown that $v_l^* \nearrow v^*$. The proof of such result relies on representation theorems of polynomials that are positive on a compact set (see Ref. 49 for a detailed treatment). We remark that the compactness assumption is not as restrictive as it may seem. It can be guaranteed, for instance, if we know *a priori* that there exists an optimal solution x^* to (PO) with $\|x^*\|_2 \leq r$ for some $r > 0$, since then we can add the redundant polynomial inequality constraint $g_{m+1}(x) = r^2 - \|x\|_2^2 \geq 0$ to (PO) and obtain a compact feasible region. We summarize as follows:

Theorem 2. *Suppose that the feasible region of (PO) is compact. Then, we have $v_l^* \nearrow v^*$.*

There has been much research on the systematic generation of successively tighter SDP relaxations for various polynomial optimization problems. We refer the interested readers to Refs 47, 50–55 and the references therein for further details.

SAFE TRACTABLE APPROXIMATIONS OF CHANCE CONSTRAINED LINEAR MATRIX INEQUALITIES

In many practical applications, the data defining an optimization problem may not be known exactly. This can happen, for example, when we are unable to make precise measurements. Currently, there are several ways to model such data uncertainty. For instance, in the area of robust optimization, the uncertain data are assumed to lie in some bounded set S , and the goal is to find a solution that is feasible for *all* realizations of the uncertain data from the set S . On the other hand, in the

area of chance constrained optimization, the uncertain data are assumed to follow certain probability distribution, and it is assumed that certain constraints can be violated *occasionally*. Thus, the goal here is to find a solution whose constraint violation probability is below a certain threshold. It should be noted that both approaches have their own advantages and disadvantages. In this section, we will focus on the chance constrained optimization approach and see how SDP can be used to tackle a large class of chance constrained optimization problems.

To begin, consider the problem

$$\begin{aligned} & \text{minimize } c^T x \\ (\text{CCP}) : & \text{subject to } \Pr \left(\mathcal{A}_0(x) - \sum_{i=1}^h \xi_i \mathcal{A}_i(x) \geq \mathbf{0} \right) \\ & \geq 1 - \epsilon, (\dagger) \\ & x \in \mathbb{R}^n, \end{aligned}$$

where $c \in \mathbb{R}^n$ is a given objective vector; $\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_h : \mathbb{R}^n \rightarrow \mathcal{S}^m$ are affine functions in x with $\mathcal{A}_0(x) \succ \mathbf{0}$ for all $x \in \mathbb{R}^n$; $\xi_1, \dots, \xi_h$ are independent (but not necessarily identical) mean zero random variables; and $\epsilon \in (0, 1)$ is the error tolerance parameter. Problem (CCP) arises from many engineering applications, such as truss topology design and problems in control theory and communications, and has received much attention lately [30, 56–60]. Moreover, it captures chance constrained linear and conic quadratic programming as special cases. In general, the constraint $(\dagger)$ in (CCP) is computationally intractable. In an attempt to circumvent this problem, Ben-Tal and Nemirovski [30, 58] proposed a *safe tractable approximation* of $(\dagger)$ —that is, a system of constraints $\mathcal{H}$ such that (i) x is feasible for $(\dagger)$ whenever it is feasible for $\mathcal{H}$, and (ii) the constraints in $\mathcal{H}$ are efficiently computable. Specifically, their strategy is as follows. First, observe that

$$\begin{aligned} & \Pr \left(\mathcal{A}_0(x) - \sum_{i=1}^h \xi_i \mathcal{A}_i(x) \geq \mathbf{0} \right) \\ & = \Pr \left(\sum_{i=1}^h \xi_i \mathcal{A}_i'(x) \leq I \right), \end{aligned}$$

where $\mathcal{A}'_i(x) = \mathcal{A}_0^{-1/2}(x)\mathcal{A}_i(x)\mathcal{A}_0^{-1/2}(x)$. Now, suppose one can choose $\gamma = \gamma(\epsilon) > 0$ so that whenever

$$\sum_{i=1}^h (\mathcal{A}'_i(x))^2 \preceq \gamma^2 I \quad (6)$$

holds, the constraint $(\dagger)$ is satisfied. Then, the constraint (6) will be a sufficient condition for $(\dagger)$ to hold. The parameter γ can be viewed as the degree of conservatism of the approximation. In particular, a smaller γ results in a larger “gap” between the safe tractable constraint (6) and the original chance constraint $(\dagger)$. Now, the upshot of constraint (6) is that it can be expressed as a linear matrix inequality using the Schur complement:

$$\begin{bmatrix} \gamma \mathcal{A}_0(x) & \mathcal{A}_1(x) & \cdots & \mathcal{A}_h(x) \\ \mathcal{A}_1(x) & \gamma \mathcal{A}_0(x) & & \\ \vdots & & \ddots & \\ \mathcal{A}_h(x) & & & \gamma \mathcal{A}_0(x) \end{bmatrix} \succeq \mathbf{0}. \quad (7)$$

Thus, by replacing $(\dagger)$ with constraint (7), Problem (CCP) becomes tractable. Moreover, any solution $x \in \mathbb{R}^n$ that satisfies constraint (7) will be feasible for the original chance constrained problem (CCP).

Using some deep results from the functional analysis literature, it can be shown that when the random variables $\xi_1, \dots, \xi_h$ are “nice,” there indeed exists a choice of $\gamma > 0$ such that constraint (6) is a sufficient condition for $(\dagger)$ to hold. Specifically, we have the following:

Theorem 3. *Let $\xi_1, \dots, \xi_h$ be independent mean zero random variables, each of which is either (i) supported on $[-1, 1]$, or (ii) normally distributed with unit variance. Consider problem (CCP). Then, for any $\epsilon \in (0, 1/2]$, the positive semidefinite constraint (7) with $\gamma \leq \gamma(\epsilon) \equiv \left(\sqrt{8e \ln(m/\epsilon)}\right)^{-1}$ is a safe tractable approximation of $(\dagger)$.*

The proof of Theorem 3 can be found in Ref. 35; see also Ref. 58.

We refer those readers who are interested in optimization under data uncertainty to the books [61,62] for further techniques and results.

APPLICATIONS IN DATA ANALYSIS

Recently, SDP has been used to tackle various problems in data analysis and machine learning. In this section, we will study one such problem, namely, sparse principal component analysis, and show how it can be tackled using SDP.

Principal component analysis (PCA) [63] is a very important tool in data analysis. It provides a way to reduce the dimension of a given data set, thus revealing the sometimes hidden underlying structure of and facilitating further analysis on the data set. To motivate the problem of finding principal components, consider the following scenario. Suppose that we are interested in some attributes $X_1, \dots, X_n$ of a population. In order to estimate the values of these attributes, one may sample from the population. Specifically, let X_{ij} be the value of the j th attribute of the i th individual, where $i = 1, \dots, m$ and $j = 1, \dots, n$. For $j, k = 1, \dots, n$, define

$$\bar{X}_j = \frac{1}{m} \sum_{i=1}^m X_{ij}, \quad \sigma_{jk} = \frac{1}{m} \sum_{i=1}^m (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

to be the sample mean of X_j and the sample covariance between X_j and X_k , respectively. Define $\Sigma = [\sigma_{jk}]_{j,k}$ to be the sample covariance matrix. The goal is then to find the *principal components* $u_1, \dots, u_n$ such that the linear combination $\sum_{j=1}^n u_j X_j$ has *maximum* sample variance. In other words, we would like to solve the following problem:

$$\max_{\|u\|_2 \leq 1} u^T \Sigma u, \quad (8)$$

where $u = (u_1, \dots, u_n)$. The model given by problem (8) is based on the belief that principal components with a large associated variance indicate interesting features in the data. However, one of the shortcomings of this model is that the linear combination may use a large number of the original variables, that is, most of the factors $u_1, \dots, u_n$ are *non-zero*, thus posing a difficulty in interpreting which variables are the most influential. This motivates us to consider the problem of finding *sparse* principal components, that is, a set $\{u_1, \dots, u_n\}$ of principal components that

maximizes the variance while at the same time having only a few non-zero u_i 's. There are many ways to formulate such a problem. For instance, one can take the following *penalty function* approach [64]:

$$v(\rho) = \max_{\|u\|_2 \leq 1} \left\{ u^T \Sigma u - \rho \|u\|_0 \right\}, \quad (9)$$

where $\|u\|_0 = |\{i \in \{1, \dots, n\} : u_i \neq 0\}|$ is the number of non-zero elements in the vector $u \in \mathbb{R}^n$, and $\rho \in \mathbb{R}$ is a parameter controlling the level of sparsity. Note that the objective function in problem (9) is nonconvex and the associated optimization problem is hard to solve in general. However, as shown in Ref. 64, it can be relaxed to an SDP. Let us now sketch the main ideas in Ref. 64.

To begin, observe that $\Sigma \succeq \mathbf{0}$, and we may assume without loss that $\Sigma_{11} \geq \Sigma_{22} \geq \dots \geq \Sigma_{nn} \geq 0$. Let $\Sigma^{1/2} \succeq \mathbf{0}$ be such that $\Sigma = \Sigma^{1/2} \Sigma^{1/2}$. It is not hard to show that when $\rho > \Sigma_{11}$, the optimal solution to Equation (9) is given by $u = \mathbf{0}$. Thus, we only need to consider the case where $\rho \leq \Sigma_{11}$. Then, Problem (9) is equivalent to the problem

$$v(\rho) = \max_{\|u\|_2=1} \left\{ u^T \Sigma u - \rho \|u\|_0 \right\}. \quad (10)$$

For a given vector $u \in \mathbb{R}^n$, let $w(u) \in \{0, 1\}^n$ be the indicator vector of the *sparsity pattern* of u , that is, $w(u)_i = 1$ iff $u_i \neq 0$ for $i = 1, \dots, n$. Observe that

$$u^T \Sigma u = \bar{u}^T \left(\text{diag}(w(u)) \Sigma \text{diag}(w(u)) \right) \bar{u}$$

for any $\bar{u} \in \mathbb{R}^n$ such that $\bar{u}_i = u_i$ whenever $u_i \neq 0$, where $i = 1, \dots, n$. On the other hand, by the Courant–Fischer characterization, for a given sparsity pattern vector $w \in \{0, 1\}^n$, we have

$$\begin{aligned} \max_{\|u\|_2=1} u^T \left(\text{diag}(w) \Sigma \text{diag}(w) \right) u \\ = \lambda_{\max} \left(\text{diag}(w) \Sigma \text{diag}(w) \right), \end{aligned}$$

where $\lambda_{\max}(M)$ is the largest eigenvalue of the matrix M . Thus, upon letting $\sigma_i \in \mathbb{R}^n$ to be the i th column of $\Sigma^{1/2}$ (where $i = 1, \dots, n$),

we may write problem (10) as

$$\begin{aligned} v(\rho) &= \max_{w \in \{0,1\}^n} \left\{ \lambda_{\max} \left(\text{diag}(w) \Sigma \text{diag}(w) \right) - \rho e^T w \right\} \\ &= \max_{w \in \{0,1\}^n} \left\{ \lambda_{\max} \left(\text{diag}(w) \Sigma^{1/2} \Sigma^{1/2} \text{diag}(w) \right) \right. \\ &\quad \left. - \rho e^T w \right\} \\ &= \max_{w \in \{0,1\}^n} \left\{ \lambda_{\max} \left(\Sigma^{1/2} \text{diag}(w) \Sigma^{1/2} \right) - \rho e^T w \right\} \end{aligned} \quad (11)$$

$$\begin{aligned} &= \max_{\|u\|_2=1} \max_{w \in \{0,1\}^n} \left\{ u^T \left(\Sigma^{1/2} \text{diag}(w) \Sigma^{1/2} \right) u \right. \\ &\quad \left. - \rho e^T w \right\} \end{aligned}$$

$$= \max_{\|u\|_2=1} \max_{w \in \{0,1\}^n} \sum_{i=1}^n w_i \left[\left(\sigma_i^T u \right)^2 - \rho \right] \quad (12)$$

$$= \max_{\|u\|_2=1} \sum_{i=1}^n \left[\left(\sigma_i^T u \right)^2 - \rho \right]_+, \quad (13)$$

where Equation (11) follows from the fact that $\lambda_{\max}(M^T M) = \lambda_{\max}(M M^T)$ for any matrix M and $\text{diag}(w)^2 = \text{diag}(w)$ for any $w \in \{0, 1\}^n$, and Equation (12) follows from the fact that

$$\begin{aligned} u^T \left(\Sigma^{1/2} \text{diag}(w) \Sigma^{1/2} \right) u &= \left\| \text{diag}(w) \Sigma^{1/2} u \right\|_2^2 \\ &= \sum_{i=1}^n w_i^2 \left(\sigma_i^T u \right)^2 \\ &= \sum_{i=1}^n w_i \left(\sigma_i^T u \right)^2. \end{aligned}$$

Note that the objective function in problem (13) is still nonconvex. However, the upshot of problem (13) is that the objective function is linear in $u_i u_j$, where $i, j = 1, \dots, n$. In particular, let $U = u u^T \in \mathbb{R}^{n \times n}$. Then, it is easy to verify that $(\sigma_i^T u)^2 = \sigma_i^T U \sigma_i$ for $i = 1, \dots, n$. Moreover, for a symmetric matrix U , we have $U = u u^T$ and $\|u\|_2 = 1$ iff $\text{tr}(U) = 1$, $\text{rank}(U) = 1$ and $U \succeq \mathbf{0}$. Thus, we see that problem (13) is equivalent to the following problem:

$$\begin{aligned} v(\rho) &= \text{maximize} \quad \sum_{i=1}^n \left(\sigma_i^T U \sigma_i - \rho \right)_+ \\ &\text{subject to} \quad \text{tr}(U) = 1, \\ &\quad U \succeq \mathbf{0}, \text{rank}(U) = 1. \end{aligned} \quad (14)$$

It is not hard to see that the function $U \mapsto \sum_{i=1}^n (\sigma_i^T U \sigma_i - \rho)_+$ is *convex*. Unfortunately, Problem (14) is still hard to solve in general, since we are maximizing a convex function, and the constraint $\text{rank}(U) = 1$ is not convex. However, not all is lost, as we may proceed as follows. First, let us introduce a notation. Given an $n \times n$ symmetric matrix M , let $\lambda_1, \dots, \lambda_n$ be its eigenvalues. Define

$$\text{tr}(M)_+ = \sum_{i=1}^n \max\{\lambda_i, 0\}.$$

Now, observe that if $U = uu^T$ and $\|u\|_2 = 1$, then $U^{1/2} = U = uu^T$. Moreover, for any $\alpha \in \mathbb{R}$, the matrix αuu^T has one eigenvalue equal to α and $n - 1$ eigenvalues equal to 0, which implies that $\text{tr}(\alpha uu^T)_+ = \alpha_+$. It follows that

$$\begin{aligned} (\sigma_i^T U \sigma_i - \rho)_+ &= \text{tr}((\sigma_i^T uu^T \sigma_i - \rho) uu^T)_+ \\ &= \text{tr}(u(u^T \sigma_i \sigma_i^T u - \rho)u^T)_+ \\ &= \text{tr}(U^{1/2} \sigma_i \sigma_i^T U^{1/2} - \rho U)_+. \end{aligned}$$

The point of the above manipulation is that the function $U \mapsto \text{tr}(U^{1/2} \sigma_i \sigma_i^T U^{1/2} - \rho U)_+$ is *concave*. Specifically, one can prove the following:

Theorem 4. *Let $M \in \mathbb{S}^n$ and $X \in \mathbb{S}_+^n$. Then, we have*

$$\begin{aligned} \text{tr}(X^{1/2} M X^{1/2})_+ &= \max_{X \succeq P \succeq 0} \text{tr}(PM) \\ &= \min_{Y \succeq M, Y \succeq 0} \text{tr}(YX). \end{aligned} \quad (15)$$

In particular, the function $X \mapsto \text{tr}(X^{1/2} M X^{1/2})_+$ is concave on $\mathbb{S}_+^n$.

Upon applying Theorem 4 and dropping the problematic rank constraint $\text{rank}(U) = 1$ in problem (14), we obtain the following *convex relaxation*:

$$\begin{aligned} v'(\rho) &= \text{maximize} \quad \sum_{i=1}^n \text{tr}(U^{1/2} \sigma_i \sigma_i^T U^{1/2} - \rho U)_+ \\ &\text{subject to} \quad \text{tr}(U) = 1, \\ &\quad U \succeq 0. \end{aligned} \quad (16)$$

Upon letting $M_i(\rho) = \sigma_i \sigma_i^T - \rho I$ for $i = 1, \dots, n$ and using Equation (15), we see that problem (16) is equivalent to

$$\begin{aligned} v'(\rho) &= \text{maximize} \quad \sum_{i=1}^n \text{tr}(P_i M_i(\rho)) \\ &\text{subject to} \quad \text{tr}(U) = 1, \\ &\quad U \succeq P_i \succeq 0 \\ &\quad U \succeq 0, \\ &\quad \text{for } i = 1, \dots, n, \end{aligned}$$

which is an SDP.

It turns out that SDP is a very powerful tool for addressing a host of data analysis and machine learning problems. We refer interested readers to Refs 65–70 for further details.

Acknowledgments

The author would like to thank the reviewers for their detailed comments and suggestions. This research is supported by Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project No. CUHK 416908.

REFERENCES

1. Grötschel M, Lovász L, Schrijver A. Geometric algorithms and combinatorial optimization. Volume 2, Algorithms and combinatorics. 2nd ed. Berlin Heidelberg: Springer; 1993.
2. Ben-Tal A, Nemirovski A. Lectures on modern convex optimization: analysis, algorithms, and engineering applications, MPS–SIAM Series on Optimization. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2001.
3. Nesterov Y. Smoothing technique and its applications in semidefinite optimization. Math Program Ser A 2007;110(2):245–259.

4. d'Aspremont A. Subsampling Algorithms for Semidefinite Programming; 2009. In press.
5. Wen Z, Goldfarb D, Ma S, *et al.* Row by Row methods for semidefinite programming. Technical Report. New York: Department of Industrial Engineering and Operations Research, Columbia University; 2009. p. 10027.
6. Goemans MX, Williamson DP. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J ACM* 1995;42(6):1115–1145.
7. Nesterov Y. Quality of semidefinite relaxation for nonconvex quadratic optimization. CORE Discussion Paper 9719, Université Catholique de Louvain, Belgium; 1997.
8. Nemirovski A, Roos C, Terlaky T. On maximization of quadratic form over intersection of ellipsoids with common center. *Math Program Ser A* 1999;86(3):463–473.
9. Palomar DP, Eldar YC, editors. *Convex optimization in signal processing and communications*. New York: Cambridge University Press; 2010.
10. Luo Z-Q, Ma W-K, So AM-C, *et al.* Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Process Mag* 2010; 27(3):20–34.
11. Karger D, Motwani R, Sudan M. Approximate graph coloring by semidefinite programming. *J ACM* 1998;45(2):246–265.
12. Halperin E, Zwick U. A unified framework for obtaining improved approximation algorithms for maximum graph bisection problems. In: Aardal K, Gerards B, editors. *Proceedings of the 8th Conference on Integer Programming and Combinatorial Optimization (IPCO VIII)*, Volume 2081 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer; 2001. pp. 210–225.
13. Charikar M. On semidefinite programming relaxations for graph coloring and vertex cover. *Proceedings of the 13th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA 2002)*; San Francisco (CA): 2002. pp. 616–620.
14. Goemans MX, Williamson DP. Approximation algorithms for MAX–3–CUT and other problems via complex semidefinite programming. *J Comput Syst Sci* 2004;68(2):442–470.
15. Agarwal A, Charikar M, Makarychev K, *et al.* $O(\sqrt{\log n})$ approximation algorithms for Min UnCut, Min 2CNF deletion, and directed cut problems. *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC 2005)*; Baltimore (MD): 2005. pp. 573–581.
16. Alon N, Naor A. Approximating the Cut–Norm via Grothendieck's inequality. *SIAM J Comput* 2006;35(4):787–803.
17. Alon N, Makarychev K, Makarychev Y, *et al.* Quadratic forms on graphs. *Invent Math* 2006;163(3):499–522.
18. Charikar M, Makarychev K, Makarychev Y. Near-optimal algorithms for unique games. *Proceedings of the 38th Annual ACM Symposium on Theory of Computing (STOC 2006)*; Seattle (WA): 2006. pp. 205–214.
19. Chlamtac E, Makarychev K, Makarychev Y. How to play unique games using embeddings. *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*; Berkeley (CA): 2006. pp. 687–696.
20. Arora S, Lee JR, Naor A. Euclidean distortion and the sparsest cut. *J Am Math Soc* 2008;21(1):1–21.
21. Khot S, Naor A. Linear equations modulo 2 and the L_1 diameter of convex bodies. *SIAM J Comput* 2008;38(4):1448–1463.
22. Raghavendra P. Optimal algorithms and inapproximability results for every CSP?. *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC 2008)*; Victoria, British Columbia, Canada: 2008. pp. 245–254.
23. Arora S, Rao S, Vazirani U. Expander flows, geometric embeddings and graph partitioning. *J ACM* 2009;56(2):Article 5.
24. Bertsimas D, Ye Y. Semidefinite relaxations, multivariate normal distributions, and order statistics. In: Du D-Z, Pardalos PM, editors. *Handbook of combinatorial optimization*. Volume 3, Dordrecht, The Netherlands: Kluwer Academic Publishers; 1998. pp. 1–19.
25. Ye Y. Approximating quadratic programming with bound and quadratic constraints. *Math Program* 1999;84(2):219–226.
26. Nesterov Y, Wolkowicz H, Ye Y. Semidefinite programming relaxations of nonconvex quadratic optimization. In: Wolkowicz H, Sagnol R, Vandenberghe L, editors. *Handbook of semidefinite programming: theory, algorithms, and applications*, Volume 27 of *International Series in Operations Research and Management Science*. Boston (MA): Kluwer Academic Publishers; 2000. pp. 361–419.
27. Ye Y, Zhang S. New results on quadratic minimization. *SIAM J Optim* 2003;14(1):245–267.
28. Beck A, Eldar YC. Strong duality in nonconvex quadratic optimization with two

- quadratic constraints. *SIAM J Optim* 2006;17(3):844–860.
29. Luo Z-Q, Sidiropoulos ND, Tseng P, *et al.* Approximation bounds for quadratic optimization with homogeneous quadratic constraints. *SIAM J Optim* 2007;18(1):1–28.
30. Nemirovski A. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math Program Ser B* 2007;109(2–3):283–317.
31. So AM-C, Zhang J, Ye Y. On approximating complex quadratic optimization problems via semidefinite programming relaxations. *Math Program Ser B* 2007;110(1):93–110.
32. He S, Luo Z-Q, Nie J, *et al.* Semidefinite relaxation bounds for indefinite homogeneous quadratic optimization. *SIAM J Optim* 2008;19(2):503–523.
33. So AM-C, Ye Y, Zhang J. A unified theorem on SDP rank reduction. *Math Oper Res* 2008;33(4):910–920.
34. Ai W, Zhang S. Strong duality for the CDT subproblem: a necessary and sufficient condition. *SIAM J Optim* 2009;19(4):1735–1756.
35. So AM-C. Moment inequalities for sums of random matrices and their applications in optimization. *Math Program* 2009. In press.
36. Ai W, Huang Y, Zhang S. New results on hermitian matrix rank-one decomposition. *Math Program* 2010. In press.
37. Kisiailiou M, Luo Z-Q. Performance analysis of Quasi-Maximum-Likelihood detector based on semidefinite programming. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP 2005), Volume 3*; Philadelphia (PA): 2005. pp. III–433–III–436.
38. Chang T-H, Luo Z-Q, Chi C-Y. Approximation bounds for semidefinite relaxation of Max-Min-Fair multicast transmit beamforming problem. *IEEE Trans Signal Process* 2008;56(8):3932–3943.
39. So AM-C. On the performance of semidefinite relaxation MIMO detectors for QAM constellations. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*; Taipei, Taiwan: 2009. pp. 2449–2452.
40. So AM-C. Probabilistic analysis of the semidefinite relaxation detector in digital communications. *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2010)*; Austin (TX): 2010. pp. 698–711.
41. Zhang YJ, So AM-C. Optimal spectrum sharing in MIMO cognitive radio networks via semidefinite programming. 2009. In press.
42. So AM-C, Ye Y. A semidefinite programming approach to tensegrity theory and realizability of graphs. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2006)*; Miami (FL): 2006. pp. 766–775.
43. Varadarajan K, Venkatesh S, Ye Y, *et al.* Approximating the radii of point sets. *SIAM J Comput* 2007;36(6):1764–1776.
44. So AM-C, Ye Y. Theory of semidefinite programming for sensor network localization. *Math Program Ser B* 2007;109(2):367–384.
45. Zhu Z, So AM-C, Ye Y. Universal rigidity: towards accurate and efficient localization of wireless networks. *Proceedings of the 29th Conference on Computer Communications (INFOCOM 2010)*; San Diego (CA): 2010.
46. Lasserre JB. Global optimization with polynomials and the problem of moments. *SIAM J Optim* 2001;11(3):796–817.
47. Laurent M. Sums of squares, moment matrices and optimization over polynomials. In: Putinar M, Sullivant S, editors. *Emerging applications of algebraic geometry*, volume 149 of *The IMA Volumes in Mathematics and its Applications*. New York: Springer Science+Business Media, LLC; 2009. pp. 157–270.
48. Lasserre JB. Convergent LMI relaxations for nonconvex quadratic programs. *Proceedings of the 39th IEEE Conference on Decision and Control, Volume 5*; Sydney (NSW), Australia: 2000. pp. 5041–5046.
49. Prestel A, Delzell CN. Positive polynomials: from Hilbert’s 17th problem to real algebra. *Springer monographs in mathematics*. Berlin/Heidelberg: Springer; 2001.
50. Lovász L, Schrijver A. Cones of matrices and Set-Functions and 0–1 optimization. *SIAM J Optim* 1991;1(2):166–190.
51. Laurent M, Rendl F. Semidefinite programming and integer programming. In: Aardal K, Nemhauser GL, Weismantel R, editors. *Volume 12, Discrete optimization, handbooks in operations research and management science*. Amsterdam, The Netherlands: Elsevier B. V.; 2005. pp. 393–514.
52. Lasserre JB. Convex sets with semidefinite representation. *Math Program Ser A* 2009;120(2):457–477.
53. Marshall M. Representations of Non-Negative polynomials, degree bounds and

- applications to optimization. *Can J Math* 2009;61(1):205–221.
54. Putinar M, Sullivant S, editors. Emerging applications of algebraic geometry. Volume 149, The IMA Volumes in Mathematics and its Applications. New York: Springer Science+Business Media, LLC; 2009.
 55. Helton JW, Nie J. Semidefinite representation of convex sets. *Math Program Ser A* 2010;122(1):21–64.
 56. Nemirovski A, Shapiro A. Convex approximations of chance constrained programs. *SIAM J Optim* 2006;17(4):969–996.
 57. Nemirovski A, Shapiro A. Scenario approximations of chance constraints. In: Calafiore G, Dabbene F, editors. Probabilistic and randomized methods for design under uncertainty. London: Springer; 2006. pp. 3–47.
 58. Ben-Tal A, Nemirovski A. On safe tractable approximations of chance-constrained Linear Matrix inequalities. *Math Oper Res* 2009;34(1):1–25.
 59. Shenouda MB, Davidson TN. Outage-based designs for multi-user transceivers. Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009); Taipei, Taiwan: 2009. pp. 2389–2392.
 60. Li WL, Zhang YJ, So AM-C, *et al.* Slow adaptive OFDMA systems through chance-constrained programming. *IEEE Trans Signal Process* 2010;58(7):3858–3869.
 61. Shapiro A, Dentcheva D, Ruszczyński A. Lectures on stochastic programming: modeling and theory, MPS–SIAM Series on Optimization. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2009.
 62. Ben-Tal A, El Ghaoui L, Nemirovski A. Robust optimization, Princeton Series in Applied Mathematics. Princeton (NJ): Princeton University Press; 2009.
 63. Shapiro A. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika* 1982;47(2):187–199.
 64. d’Aspremont A, Bach F, El Ghaoui L. Optimal solutions for sparse principal component analysis. *J Mach Learn Res* 2008;9:1269–1294.
 65. Lanckriet GRG, Cristianini N, Barlett P, *et al.* Learning the kernel matrix with semidefinite programming. *J Mach Learn Res* 2004;5:27–72.
 66. d’Aspremont A, El Ghaoui L, Jordan MI, *et al.* A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev* 2007;49(3):434–448.
 67. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Found Comput Math* 2009;9(6):717–772.
 68. Gross D. Recovering low-rank matrices from few coefficients in any basis; 2009. In press.
 69. Candès EJ, Li X, Ma Y, *et al.* Robust principal component analysis? Technical Report 2009–13. Stanford (CA): Department of Statistics, Stanford University; 2009. pp. 94305.
 70. Wu X-M, So AM-C, Li Z, *et al.* Fast graph laplacian regularized kernel learning via semidefinite-quadratic-linear programming. In: Bengio Y, Schuurmans D, Lafferty J, *et al.*, editors. Advances in neural information processing systems 22: proceedings of the 2009 conference. Cambridge (MA): The MIT Press; 2009. pp. 1964–1972.

SEMI-INFINITE PROGRAMMING

FRANCISCO GUERRA-VÁZQUEZ
Departamento de Actuaría, Física y
Matemáticas, Universidad de las
Américas, Escuela de Ciencias,
Puebla, México

JAN-J. RÜCKMANN
The University of Birmingham
School of Mathematics,
Birmingham, UK

In this article we consider a *semi-infinite programming* (SIP) problem of the form

SIP: minimize $f(x)$ subject to $x \in M$, (1)
where the *feasible set* M is given as

$$M = \{x \in \mathbb{R}^n \mid G(x, y) \geq 0, y \in Y\}$$

and the *index set* Y is defined as

$$Y = \{y \in \mathbb{R}^r \mid u_l(y) = 0, l \in A, v_k(y) \leq 0, k \in B\}.$$

Assume that:

- The functions $f, G, u_l, l \in A = \{1, \dots, a\}$, $v_k, k \in B = \{1, \dots, b\}$ are real-valued and sufficiently smooth (i.e., they are once (twice) continuously differentiable when considering their first (second) derivatives).
- The index set $Y \subset \mathbb{R}^r$ is compact and, in general, infinite. Each element $\bar{y} \in Y$ represents a corresponding inequality constraint $G(\cdot, \bar{y}) \geq 0$. Obviously, the set Y is defined by finitely many equality and inequality constraints and, therefore, it has the structure of the feasible set of a finite programming problem.

Since there are finitely many decision variables $x = (x_1, \dots, x_n)$ and infinitely many constraints $G(x, \bar{y}) \geq 0$, $\bar{y} \in Y$, the problem is called *semi-infinite*. The objective of a solution method for SIP is to find a *local*

minimizer, that is, a *feasible point* $\bar{x} \in M$ satisfying

$$f(\bar{x}) \leq f(x) \quad (2)$$

for all feasible points x from a certain neighborhood of $\bar{x}$. Under certain assumptions, one is interested in calculating a *global minimizer* of SIP, that is, a point $\bar{x} \in M$ satisfying Equation (2) for all $x \in M$.

SIP has a wide range of applications and it became a vivid research area within mathematical programming over the last two decades; for more details we refer to the survey papers [1,2], the monographs [3–5], and the books [6,7], which contain state-of-the-art articles on theory, solution methods, and applications of SIP.

When the SIP model is applied to real-life problems, the index set Y may refer, for example

- to a (compact) time interval: at each time point of this interval a corresponding velocity is restricted (e.g., control problems in robotics);
- or to a geographic area: at each point of this area the level of allowed contamination is restricted (e.g., control problems in chemical engineering).

SIP originated partially from approximation problems: the problem of Chebyshev approximation is to approximate a continuous function $v: \mathbb{R}^r \rightarrow \mathbb{R}$ uniformly on a compact set $Y \subset \mathbb{R}^r$ by means of an n -parameter family of functions $V(x, \cdot)$ (e.g., a family of particular polynomials whose coefficients (parameters) are $x = (x_1, \dots, x_n)$).

This approximation problem reads as follows:

$$\text{minimize } p(x) := \max_{y \in Y} |v(y) - V(x, y)|$$

and it can be rewritten as the particular SIP problem (with $x_{n+1} \in \mathbb{R}$)

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} \{x_{n+1} \mid -x_{n+1} \leq v(y) \\ & \quad -V(x, y) \leq x_{n+1}, y \in Y\}. \end{aligned}$$

Compared to this special structure of an approximation problem, the more general setting (Eq. 1) allows taking into account any set of constraints of the form $G(x, y) \geq 0$, $y \in Y$.

In some applications, the feasible set contains additionally finitely many equality constraints $h_i(x) = 0$, $i \in I = \{1, \dots, m\}$ ($m < n$), where the gradients of these constraints are linearly independent at all feasible points.

The latter property means that these constraints describe locally around the feasible set an $(n - m)$ -dimensional manifold and, therefore, all the results presented in the following can be applied to this case as well. Therefore, in order to avoid unnecessary technicalities, we restrict ourselves throughout this article to the case $I = \emptyset$.

This article is organized as follows. In the section titled “First Order Optimality Conditions,” we discuss constraint qualifications, first order optimality conditions and, in particular, linear SIP problems. In the section titled “The Reduction Approach” we present the reduction approach, which allows the local transformation of an SIP problem into a finite programming problem. The next section refers to “Second Order Optimality Conditions,” and the section titled “Structural stability of SIP” considers several stability concepts including strong stability of a stationary point and structural stability of the SIP problem. Finally, in the section titled “Final Remarks,” we refer briefly to solution methods and generalized SIP problems.

Now, we introduce some notations used later:

- $C^k(U, \mathbb{R})$, $k \geq 1$ denotes the space of k -times continuously differentiable real-valued functions defined on $U \subset \mathbb{R}^n$.
- $C^{1,1}(U, \mathbb{R})$ denotes the space of continuously differentiable real-valued functions defined on $U \subset \mathbb{R}^n$ with Lipschitz continuous gradient.

- If $g \in C^1(U, \mathbb{R})$, then the row vector $Dg(x)$ ($D_{x^1}g(x)$) denotes the derivative (partial derivative with respect to the subvector x^1 of x) of g at x . Second derivatives are analogously represented.
- The space $C^k(\mathbb{R}^n, \mathbb{R})$, $k \geq 1$ can be topologized by means of the *strong* or *Whitney C^k -topology*, which is generated by allowing perturbations of the function and their derivatives up to order k and whose set of neighborhoods is controlled by means of continuous positive functions $\varepsilon(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$. This topology is denoted by C_s^k [8,9].

FIRST ORDER OPTIMALITY CONDITIONS

Throughout this article assume that the *linear independence constraint qualification* LICQ holds at each $\bar{y} \in Y$, that is, the gradients

$$\begin{aligned} & Du_l(\bar{y}), l \in A, Dv_k(\bar{y}), k \in B_0(\bar{y}) \\ & = \{k \in B \mid v_k(\bar{y}) = 0\} \end{aligned}$$

are linearly independent. Jongen *et al.* [9] have shown that LICQ at all $y \in Y$ is a generic property with respect to the C_s^1 -topology for the defining functions $u_l, l \in A, v_k, k \in B$. Furthermore, this property implies that for each $\bar{y} \in Y$ there exist neighborhoods U of $\bar{y}$ and W of the origin 0_r of $\mathbb{R}^r$ as well as a diffeomorphism $\psi : \mathbb{R}^r \rightarrow \mathbb{R}^r$ such that $\psi(\bar{y}) = 0_r$ and

$$\psi[Y \cap U] = (0_a \times \mathbb{H}^{|B_0(\bar{y})|} \times \mathbb{R}^{r-a-|B_0(\bar{y})|}) \cap W,$$

where $\mathbb{H}^t = \{z \in \mathbb{R}^t \mid z_i \geq 0, i = 1, \dots, t\}$ [9].

As for a finite programming problem, the set of active constraints at a feasible point is important for optimality conditions. We define for $\bar{x} \in M$ the set

$$Y_0(\bar{x}) = \{y \in Y \mid G(\bar{x}, y) = 0\}.$$

A significant difference between finite and semi-infinite programming problems is illustrated in the following example.

Example 1. Consider first the set Y as the feasible set of a finite programming problem. Obviously, for each $\bar{y} \in Y$ there exists a neighborhood U of $\bar{y}$ such that

$$B_0(y) \subseteq B_0(\bar{y}) \text{ for all } y \in Y \cap U, \quad (3)$$

that is, the set of active constraints at $y \in Y \cap U$ is a subset of the set of active constraints at $\bar{y}$. This property is important for many solution methods in finite programming. On the other hand, consider now the sphere $S^{n-1} \subset \mathbb{R}^n$, which has a nonempty (and infinite) boundary ∂S^{n-1} . At each of its boundary points the corresponding tangent space forms the zero-set of a linear inequality. Then S^{n-1} is described by these infinitely many linear inequality constraints (intersection of infinitely many half-spaces) and at each boundary point $\bar{x} \in \partial S^{n-1}$ there exists exactly one active constraint ($|Y_0(\bar{x})| = 1$). In particular, for any two different points $x^1, x^2 \in S^{n-1}$ we have $Y_0(x^1) \cap Y_0(x^2) = \emptyset$ and, therefore, a relationship analogous to Equation (3) is not fulfilled for SIP problems in general (obviously, S^{n-1} can be substituted by other convex sets). The control of the set of active constraints is one of the main features in SIP.

In the remainder of this article let $\bar{x} \in M$ be our point under consideration. The following simple observation will play a crucial role. Obviously, each active index $\bar{y} \in Y_0(\bar{x})$ is a *global minimizer* of the (parametric) finite programming problem

$$L(\bar{x}) : \text{minimize } G(\bar{x}, y) \text{ subject to } y \in Y.$$

This problem depends on the n -dimensional parameter $\bar{x}$ and it is called *lower level problem*. Hence, finding one (or all) active constraint(s) at $\bar{x}$ means finding one (or all) global minimizers of the corresponding lower level problem $L(\bar{x})$. In the general nonconvex case this is not possible by using deterministic methods. Now, let $\bar{y} \in Y_0(\bar{x})$. By LICQ and the well-known Karush-Kuhn-Tucker (KKT) optimality conditions for finite programming [10], there exist uniquely determined multipliers $\bar{\alpha}_l, l \in A, \bar{\beta}_k \geq 0, k \in B_0(\bar{y})$ such that

$$D_y G(\bar{x}, \bar{y}) + \sum_{l \in A} \bar{\alpha}_l D u_l(\bar{y}) + \sum_{k \in B_0(\bar{y})} \bar{\beta}_k D v_k(\bar{y}) = 0. \quad (4)$$

Furthermore, we define the set

$$\Sigma = \{x \in M \mid x \text{ is a local minimizer of SIP}\}.$$

The following theorem presents several (dual) first order optimality conditions for SIP [1,11].

Theorem 1. *First order necessary optimality condition:*

(i) *Let $\bar{x} \in \Sigma$. Then, the system*

$$Df(\bar{x})\xi < 0, D_x G(\bar{x}, y)\xi > 0, \forall y \in Y_0(\bar{x})$$

has no solution $\xi \in \mathbb{R}^n$.

(ii) *First order necessary optimality condition of Fritz John type:*

Let $\bar{x} \in \Sigma$. Then, there exist finitely many $y^j \in Y_0(\bar{x}), j = 1, \dots, q$ and $\mu_j \geq 0, j = 0, \dots, q, \sum_{j=0}^q \mu_j > 0$ ($q = 0$ is possible) such that

$$\mu_0 Df(\bar{x}) - \sum_{j=1}^q \mu_j D_x G(\bar{x}, y^j) = 0. \quad (5)$$

(iii) *First order sufficient optimality condition:*

Assume that the system

$$Df(\bar{x})\xi \leq 0, D_x G(\bar{x}, y)\xi \geq 0, \forall y \in Y_0(\bar{x})$$

has only the trivial solution $\xi = 0$. Then, $\bar{x} \in \Sigma$ and there exists a neighborhood $U \subset \mathbb{R}^n$ of $\bar{x}$ such that $f(\bar{x}) < f(x)$ for all $x \in M \cap U \setminus \{\bar{x}\}$.

In the sequel, we will use the following two constraint qualifications (which are called extended since they are extensions of the corresponding well-known constraint qualifications in finite programming).

The extended linear independence constraint qualification (ELICQ) holds at $\bar{x}$ if the vectors $D_x G(\bar{x}, y), y \in Y_0(\bar{x})$ are linearly independent.

The *extended Mangasarian Fromovitz constraint qualification* (EMFCQ) holds at $\bar{x} \in M$ if there exists a vector $\xi \in \mathbb{R}^n$ such that $D_x G(\bar{x}, y) \xi > 0$ for all $y \in Y_0(\bar{x})$.

Obviously, if ELICQ holds at $\bar{x}$, then EMFCQ holds at $\bar{x}$ as well. Furthermore, if EMFCQ holds at $\bar{x} \in \Sigma$, then we have $\mu_0 \neq 0$ in Equation (5) and we get

First order necessary optimality condition of Karush-Kuhn-Tucker type [1,11]:

Let $\bar{x} \in \Sigma$ and assume that EMFCQ holds at $\bar{x}$. Then, there exist $y^j \in Y_0(\bar{x})$, $j = 1, \dots, q$ and $\mu_j \geq 0$, $j = 1, \dots, q$ such that the gradients $D_x G(\bar{x}, y^j)$, $j = 1, \dots, q$ ($q = 0$ is possible) are linearly independent and

$$Df(\bar{x}) - \sum_{j=1}^q \mu_j D_x G(\bar{x}, y^j) = 0. \quad (6)$$

Analogously to finite programming we call a feasible point $\bar{x} \in M$ that fulfills Equation (6) a *stationary point of SIP*. A stationary point can be a local minimizer, a saddle point, or a local maximizer of SIP. In the case of a convex SIP problem (where the functions $f(\cdot)$ and $-G(\cdot, y)$, $y \in Y$ are convex) each stationary point is also a global minimizer of SIP.

Finally in this section, we will consider the particular case of a linear SIP problem, which is defined as follows:

LSIP: minimize $c^\top x$ subject to $x \in M_L$,

where the feasible set is

$$M_L = \{x \in \mathbb{R}^n \mid d(y)^\top x - e(y) \geq 0, y \in Y\}$$

with $c \in \mathbb{R}^n$ and $d: Y \rightarrow \mathbb{R}^n$, $e: Y \rightarrow \mathbb{R}$ are continuous functions and Y is defined as above. Obviously, M_L is a convex set (intersection of half-spaces). The *Slater condition* holds at M_L if there exists an $x^0 \in M_L$ such that $a(y)^\top x^0 - b(y) > 0$ for all $y \in Y$. As in finite programming, the Slater condition holds at M_L if and only if EMFCQ holds at all $x \in M_L$. Applying the above presented first order condition to LSIP we obtain

First order sufficient optimality condition for LSIP:

Let $\bar{x} \in M_L$ and $y^j \in Y_0(\bar{x})$, $j = 1, \dots, q$. Assume that there exist $\mu_j > 0$, $j = 1, \dots, q$ such that

$$c - \sum_{j=1}^q \mu_j a(y^j) = 0. \quad (7)$$

Then, $\bar{x}$ is a global minimizer of LSIP.

If, additionally, the Slater condition holds at M_L , then Equation (7) is also a necessary optimality condition.

Sometimes, solution methods for SIP use at certain iteration points particular auxiliary LSIP problems. The problem

LSIP $^{\bar{x}}$: minimize $f^{\bar{x}}(x)$ subject to $M^{\bar{x}}$

is called the *linearized problem at $\bar{x}$* where

$$f^{\bar{x}}(x) = f(\bar{x}) + Df(\bar{x})(x - \bar{x}),$$

$$M^{\bar{x}} = \{x \in \mathbb{R}^n \mid G^{\bar{x}}(x, y) \geq 0, y \in Y\}, \text{ and}$$

$$G^{\bar{x}}(x, y) = G(\bar{x}, y) + D_x G(\bar{x}, y)(x - \bar{x}), y \in Y.$$

A simple iteration procedure is to take the solution of LSIP $^{x^i}$ as the next iteration point x^{i+1} . We conclude this section by providing some relationship between the original problem SIP and a related linearized problem LSIP $^{\bar{x}}$.

Proposition 1 [1,11].

- (i) If EMFCQ holds at $\bar{x}$, then the Slater condition holds at $M^{\bar{x}}$.
- (ii) If EMFCQ holds at $\bar{x}$ and $\bar{x} \in \Sigma$, then $\bar{x}$ is a local minimizer of LSIP $^{\bar{x}}$.
- (iii) If $\tilde{x}$ is a local minimizer of LSIP $^{\bar{x}}$ and $\bar{x}$ is not a local minimizer of LSIP $^{\bar{x}}$, then we have

$$Df(\bar{x})(\tilde{x} - \bar{x}) < 0, D_x G(\bar{x}, y)(\tilde{x} - \bar{x}) \geq 0, \\ \forall y \in Y_0(\bar{x}).$$

For more details on linear SIP, we refer to Goberna and López [3].

THE REDUCTION APPROACH

One of the main challenges in SIP is the handling of the *infinite* index set Y . In this section we consider properties under which the feasible set M can be described locally around $\bar{x}$ by *finitely* many continuously differentiable inequality constraints.

Let $\bar{y} \in Y_0(\bar{x})$ and the uniquely determined multipliers $\bar{\alpha}_l$, $l \in A$, $\bar{\beta}_k \geq 0$, $k \in B_0(\bar{y})$ be given as in Equation (4). Define $B_+(\bar{y}) = \{k \in B \mid \bar{\beta}_k > 0\}$. The so-called *strong second order sufficient condition* SSOSC [12] holds at $\bar{y}$ if the matrix $Q^\top P Q$ is positive definite where

$$P = D_{yy}^2 G(\bar{x}, \bar{y}) + \sum_{l \in A} \bar{\alpha}_l D^2 u_l(\bar{y}) + \sum_{k \in B_0(\bar{y})} \bar{\beta}_k D^2 v_k(\bar{y})$$

and Q is an $(r, r - a - |B_+(\bar{y})|)$ —matrix whose columns form a basis of the tangent space

$$\{y \in \mathbb{R}^r \mid Du_l(\bar{y})y = 0, l \in A, Dv_k(\bar{y})y = 0, k \in B_+(\bar{y})\}.$$

If SSOSC holds at $\bar{y}$, then $\bar{y}$ is called a *strongly stable local minimizer* of $L(\bar{x})$ (see Kojima [13] for more details on strong stability) and we get the following statement.

Theorem 2 [13]. *Let $\bar{y} \in Y_0(\bar{x})$ be chosen as above and assume that SSOSC holds at $\bar{y}$. Then there exist an open neighborhood U of $\bar{x}$ and a uniquely determined (implicit) function*

$$\hat{y} : x \in U \mapsto \hat{y}(x) \in \mathbb{R}^r$$

with the following properties:

- (a) $\hat{y}(\bar{x}) = \bar{y}$.
- (b) $\hat{y}(x)$ is a local minimizer of $L(x)$ and SSOSC holds at $\hat{y}(x)$ for all $x \in U$.
- (c) $\hat{y}$ is a Lipschitz continuous function and $G(\cdot, \hat{y}(\cdot)) \in C^{1,1}(U, \mathbb{R})$.
- (d) If $B_+(\bar{y}) = B_0(\bar{y})$, then $\hat{y}$ is a continuously differentiable function and $G(\cdot, \hat{y}(\cdot)) \in C^2(U, \mathbb{R})$.

If $B_+(\bar{y}) = B_0(\bar{y})$ and SSOSC holds at $\bar{y} \in Y_0(\bar{x})$, then $\bar{y}$ is called a *nondegenerate* minimizer of $L(\bar{x})$. Under this latter condition, the proposition (d) in Theorem 2 is a straightforward consequence of the implicit function theorem because of the nonsingularity of the Jacobian of the system

$$\begin{aligned} G(x, y) + \sum_{l \in A} \alpha_l u_l(y) + \sum_{k \in B_+(\bar{y})} \beta_k v_k(y) &= 0 \\ u_l(y) &= 0, l \in A \\ v_k(y) &= 0, k \in B_+(\bar{y}). \end{aligned}$$

Now, we recall the so-called *reduction approach* [14–16] under which the feasible set M can be described locally around $\bar{x}$ by finitely many $C^{1,1}$ – or C^2 – constraints. The *reduction approach* RA holds at $\bar{x} \in M$ if SSOSC holds at y for all $y \in Y_0(\bar{x})$.

Theorem 3. *Assume that RA holds at $\bar{x} \in M$. Then we have:*

- (a) $Y_0(\bar{x})$ is a finite set, say $Y_0(\bar{x}) = \{y^1, \dots, y^q\}$ (Note that Y is compact).
- (b) There exists an open neighborhood U of $\bar{x}$ and uniquely determined Lipschitz continuous (implicit) functions

$$\hat{y}^j : x \in U \mapsto \hat{y}^j(x) \in \mathbb{R}^r, j = 1, \dots, q,$$

which are defined analogously to $\hat{y}$ in Theorem 2 (i.e., $\hat{y}^j(\bar{x}) = y^j$, $j = 1, \dots, q, \dots$). In particular, we have

$$G(\cdot, \hat{y}^j(\cdot)) \in C^{1,1}(U, \mathbb{R}), j = 1, \dots, q.$$

- (c) Locally around $\bar{x}$ the feasible set M is described as

$$M \cap U = \left\{ x \in U \mid G(x, \hat{y}^j(x)) \geq 0, j = 1, \dots, q \right\}.$$

Although the functions $\hat{y}^j$, $j = 1, \dots, q$ in the latter theorem are only implicitly known, the derivatives of the constraints $G(x, \hat{y}^j(x))$ can be calculated explicitly (which is important for numerical methods using first order conditions):

$$\begin{aligned} DG(x, \hat{y}^j(x)) &= D_x G(x, y) \\ &+ D_y G(x, y) D\hat{y}^j(x) \Big|_{y=\hat{y}^j(x)} = D_x G(x, y), \end{aligned}$$

where the latter equation is a consequence of the fact that $\hat{y}^j(x)$ is a local minimizer of $L(x)$. Finally, we mention that the reduction approach is a generic property [5].

SECOND ORDER OPTIMALITY CONDITIONS

In this section we present a second order optimality condition for $\bar{x} \in \Sigma$ under the assumption that RA holds at $\bar{x}$ (see Hettich and Still [17] for more details). We also refer to Rückmann [18] where second order optimality conditions are presented for the more complicated case that RA does not hold at $\bar{x}$ and, in particular, where $Y_0(\bar{x})$ can be an infinite set.

In the remainder of this section assume that $\bar{x}$ is a stationary point of SIP and that RA holds at $\bar{x}$. We will use the terminology as introduced in Theorem 3. If EMFCQ holds at $\bar{x}$, then there exist multipliers $\lambda_j \geq 0, j = 1, \dots, q$ such that

$$Df(\bar{x}) - \sum_{j=1}^q \lambda_j D_x G(\bar{x}, \hat{y}^j(\bar{x})) = 0, \lambda \geq 0, \quad (8)$$

and, by Gauvin [19], the set

$$\Lambda = \{\lambda \in \mathbb{R}^q \mid \lambda \text{ is a solution of (8)}\}$$

is nonempty and compact.

Now consider for each $y^j, j = 1, \dots, q$ the following auxiliary lower level problem

$$L^j(\bar{x}) : \text{minimize } G(\bar{x}, y) \text{ subject to } y \in Y^j,$$

where

$$\begin{aligned} Y^j &= \{y \in \mathbb{R}^r \mid u_l(y) = 0, l \in A, \\ &\quad v_k(y) = 0, k \in B_0(\bar{y}), \\ &\quad v_\rho(y) \geq 0, \rho \in B \setminus B_0(\bar{y})\}. \end{aligned}$$

Since RA holds at $\bar{x}$ the property SSOSC holds at the global minimizer y^j of $L^j(\bar{x}), j = 1, \dots, q$. By Theorem 2, there exist uniquely determined implicit functions $\hat{y}_0^j, j = 1, \dots, q$, defined on a neighborhood U_0 of $\bar{x}$ with analogous properties ($\hat{y}_0^j(\bar{x}) = y^j, j = 1, \dots, q, \dots$). Now, we can state the following result.

Theorem 4 [18,20]. *Second order sufficient optimality condition:*

Assume that $\bar{x}$ is a stationary point of SIP and that RA and EMFCQ hold at $\bar{x}$. If for each $\lambda \in \Lambda$ the matrix

$$D^2f(\bar{x}) - \sum_{j=1}^q \lambda_j D^2G(\bar{x}, \hat{y}_0^j(\bar{x}))$$

is positive definite on the tangent space

$$\begin{aligned} T(\lambda) &= \{w \in \mathbb{R}^n \mid D_x G(\bar{x}, y^j)w = 0, \\ &\quad j \in \{v \in \{1, \dots, q\} \mid \lambda_v > 0\}\}, \end{aligned}$$

then $\bar{x} \in \Sigma$ and there exists a neighborhood U of $\bar{x}$ such that $f(\bar{x}) < f(x)$ for all $x \in M \cap U \setminus \{\bar{x}\}$.

We also refer to Bonnans and Shapiro [21] and Klatte and Kummer [22] for more details on second order optimality conditions.

STRUCTURAL STABILITY OF SIP

Stability properties play an important role for SIP problems with noise or perturbed data. In this section we briefly describe the concept of structural stability of an SIP problem that, in particular, includes the (topological) stability of the feasible set M and the strong stability of stationary points (see Jongen and Rückmann [23] for more details). We start with an observation on the structure of the feasible set.

Theorem 5. *Structure of the feasible set M [24]:*

- (a) *Assume that RA holds at $\bar{x} \in M$ (and apply the terminology used in Theorem 3). If ELICQ holds at $\bar{x}$, then the functions $G(\cdot, y^j(\cdot)), j = 1, \dots, q$ can be used as a new local $C^{1,1}$ -coordinate system around $\bar{x}$. Hence, there exist neighborhoods U of $\bar{x}$ and W of the origin $0_n \in \mathbb{R}^n$ and a diffeomorphism $\phi^1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\phi^1(\bar{x}) = 0_n$ and*

$$\phi^1[M \cap U] = (\mathbb{H}^q \times \mathbb{R}^{n-q}) \cap W$$

(see also the analogous consideration for $\bar{y} \in Y$ in the section titled “First Order Optimality Conditions”).

- (b) If EMFCQ holds at $\bar{x} \in M$ and $Y_0(\bar{x}) \neq \emptyset$, then there exist neighborhoods U of $\bar{x}$ and W of 0_n and a homeomorphism $\phi^2 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\phi^2(\bar{x}) = 0_n$ and

$$\phi^2[M \cap U] = (\mathbb{H}^1 \times \mathbb{R}^{n-1}) \cap W.$$

- (c) If EMFCQ holds at all $x \in M$, then M is a Lipschitzian manifold with boundary of dimension n , where the boundary ∂M is equal to

$$\left\{ x \in \mathbb{R}^n \mid \min_{y \in Y} G(x, y) = 0 \right\}.$$

The proposition (c) in the latter theorem is obviously a generalization of the local statement in (b) to the whole set M .

The property that EMFCQ holds at all $x \in M$ is a generic property with respect to the C_s^1 -topology for the function G [24]. The following theorem shows that EMFCQ plays an important role for the topological stability of M . The feasible set M is called *topologically stable* if there exists a C_s^1 -neighborhood $F \subset C^1(\mathbb{R}^n \times \mathbb{R}^r, \mathbb{R})$ of G such that for each $\tilde{G} \in F$ the corresponding feasible set $\{x \in \mathbb{R}^n \mid \tilde{G}(x, y) \geq 0, y \in Y\}$ is homeomorphic with M .

Theorem 6. *Topological stability of the feasible set M [24]:*

- (a) If EMFCQ holds at all $x \in M$, then M is topologically stable.
 (b) If M is compact and topologically stable, then EMFCQ holds at all $x \in M$.

Next, we refer to the *strong stability of a stationary point*. This concept was introduced by Kojima [13] for finite programming problems and it refers to the existence, local uniqueness, and continuous dependence on local C^2 -perturbations of the defining functions (i.e., (f, G)) when considering a stationary point of SIP; however, we used this concept already in the context of Theorem 2). Kojima

[13] provided an algebraic characterization of the strong stability of a stationary point of a finite programming problem by using the first and second derivatives of the defining functions. Then, the concept of strong stability was extended to SIP problems and in Rückmann [18] an algebraic characterization was presented for the following three cases of a stationary point $\bar{x}$ of SIP:

- ELICQ and RA hold at $\bar{x}$,
- EMFCQ (but not ELICQ) and RA hold at $\bar{x}$, and
- EMFCQ holds at $\bar{x}$ but RA does not hold at $\bar{x}$.

We will not recall these (very technical) characterizations here; however, we will use strongly stable stationary points in the following description of structural stability of an SIP problem. We also refer to Günzel and Jongen [25], where the relationship between strong stability and MFCQ is considered. Now, we define a structurally stable SIP problem.

Definition 1.

- (a) For $\gamma \in \mathbb{R}$ define the feasible lower level set as

$$N^\gamma(f, G) = \{x \in M \mid f(x) \leq \gamma\}.$$

- (b) Consider two SIP problems $SIP(f, G)$ and $SIP(\tilde{f}, \tilde{G})$, which are defined by their objective functions f and $\tilde{f}$ as well as their constraints G and $\tilde{G}$, respectively. These problems $SIP(f, G)$ and $SIP(\tilde{f}, \tilde{G})$ are called *equivalent* (notation: $SIP(f, G) \sim SIP(\tilde{f}, \tilde{G})$) if there exist continuous mappings $\varphi^1 : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\varphi^2 : \mathbb{R} \rightarrow \mathbb{R}$ with the following properties:

- $\varphi^1(\gamma, \cdot)$ is a homeomorphism for each $\gamma \in \mathbb{R}$.
- φ^2 is a homeomorphism and monotonically increasing.
- For all $\gamma \in \mathbb{R}$ we have $\varphi^1(\gamma, [N^\gamma(f, G)]) = N^{\varphi^2(\gamma)}(\tilde{f}, \tilde{G})$.

- (c) The (original) problem $SIP = SIP(f, G)$ is called structurally stable if there exist C_s^2 - neighborhoods F^1 of f and F^2 of G such that $SIP(f, G) \sim SIP(\tilde{f}, \tilde{G})$ for all $(\tilde{f}, \tilde{G}) \in F^1 \times F^2$.

The next theorem provides an equivalent characterization of structural stability.

Theorem 7. *Structural stability of SIP [23]:*

Suppose that M is a compact set. Then, the problem $SIP(= SIP(f; G))$ is structurally stable if and only if the following three conditions are fulfilled:

- (a) *EMFCQ holds at all $x \in M$.*
- (b) *Every stationary point of SIP is strongly stable.*
- (c) *Different stationary points of SIP have different objective function values.*

FINAL REMARKS

In this article we discussed very briefly a few basic features of SIP. In contrast to finite programming an SIP problem has infinitely many inequality constraints. Such problems arise, for example, in approximation theory, optimal control, or in applications from engineering, finance, and economics that are characterized by at least one parameter varying in a compact given domain (representing e.g., time, geographic area, frequency, and price) and a corresponding inequality constraint for each feasible parameter value.

It is obvious that we could not cover all the important properties of an SIP problem here. However, we conclude with some references on two other important features.

Solution Methods

As mentioned above there are two main challenges for the design of solution methods in SIP:

- There exist infinitely many inequality constraints and, therefore, each solution approach has to transform the original problem in a (sequence of) finite programming problem(s).

- For the calculation of active constraints a corresponding global optimization problem has to be solved.

Since the 1990s, there exist solution methods that use discretization techniques or assume that RA holds at the considered iteration points [2,26,27]. Recently more advanced techniques have been developed, we refer for example to

- interval methods and adaptive convexification methods (using special tools from global optimization) [28–31],
- smoothing methods [32,33],
- interior point and bilevel methods [34,35].

Generalized Semiinfinite Programming

Finally, we refer to the broader class of so-called *generalized semi-infinite programming* (GSIP) problems, which allows that the index set depends additionally on the vector x of decision variables.

A GSIP problem is defined as

$$\text{GSIP: minimize } f(x) \text{ subject to } x \in M^{\text{GSIP}},$$

where

$$M^{\text{GSIP}} = \{x \in \mathbb{R}^n \mid G(x, y) \geq 0, y \in Y(x)\}$$

and

$$Y(x) = \{y \in \mathbb{R}^r \mid u_l(x, y) = 0, l \in A, v_k(x, y) \leq 0, k \in B\}.$$

Here, Y is substituted by $Y(x)$ and, not surprisingly, the structure of this problem can become much more complicated than that in the standard SIP case. We will illustrate this by recalling the following example

Example 2 [Example 3.3 in Guerra-Vázquez *et al.* [2]]. Consider the feasible set

$$M^{\text{GSIP}} = \{x \in \mathbb{R}^2 \mid y \geq 0, y \in Y(x)\}$$

and the index set

$$Y(x) = \{y \in \mathbb{R} \mid y \geq x_1, y \leq x_2\}.$$

Then, the feasible set can be rewritten as

$$M^{\text{GSIP}} = \left\{x \in \mathbb{R}^2 \mid x_1 \leq x_2, x_1 \geq 0\right\} \\ \cup \left\{x \in \mathbb{R}^2 \mid x_1 > x_2\right\}.$$

We see that

- M^{GSIP} is not closed; and
- it is the *union* (and not the intersection as in the standard case) of half-spaces.

The possible *nonclosedness* of the feasible set and its possible *disjunctive structure* are topological phenomena which cannot appear in standard SIP. They have a strong impact on the theory of GSIP (e.g., optimality conditions, and stability properties) and on the design of solution methods. For more information on this subject see the tutorial paper [2] and recent papers [36–38].

REFERENCES

1. Hettich K, Kortanek KO. Semi-infinite programming: theory, methods, and applications. SIAM Rev 1993;35(3):380–429.
2. Guerra-Vázquez F, Rückmann J-J, Stein O, et al. Generalized semi-infinite programming: a tutorial. J Comput Appl Math 2008;217:394–419.
3. Goberna MA, López MA. Linear semi-infinite optimization. Chichester: John Wiley and Sons, Ltd.; 1998.
4. Polak E. Optimization. Algorithms and consistent approximations. Berlin: Springer; 1997.
5. Stein O. Bilevel strategies in semi-infinite programming. Boston (MA): Kluwer Academic Publishers; 2003.
6. Reemtsen R, Rückmann J-J, editors. Semi-infinite programming. Boston (MA): Kluwer Academic Publishers; 1998.
7. Goberna MA, López MA, editors. Semi-infinite programming - recent advances. Boston (MA): Kluwer Academic Publishers; 2001.
8. Hirsch MW. Differential topology. Berlin: Springer; 1976.
9. Jongen HTh, Jonker P, Twilt F. Nonlinear optimization in finite dimensions. Boston (MA): Kluwer Academic Publishers; 2000.
10. Kuhn HW, Tucker AW. Nonlinear programming. In: Neyman J, editor. Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1951. pp. 481–492.
11. Hettich R, Zencke P. Numerische methoden der approximation und semi-infiniten optimierung. Stuttgart: Teubner; 1982.
12. Fiacco AV, McCormick GP. Nonlinear programming: sequential unconstrained minimization techniques. New York: Wiley; 1968.
13. Kojima M. Strongly stable stationary solutions in nonlinear programs. In: Robinson SM, editor. Analysis and computation of fixed points. New York: Academic Press; 1980. pp. 93–138.
14. Hettich R, Jongen HTh. Semi-infinite programming: conditions of optimality and applications. In: Stoer J, editor. Optimization techniques. Berlin: Springer; 1978. pp. 1–11.
15. Jongen HTh, Wetterling WWE, Zwier G. On sufficient conditions for local optimality in semi-infinite optimization. Optimization 1987;18:165–178.
16. Klatte D. Stable local minimizers in semi-infinite optimization: regularity and second-order conditions. J Comput Appl Math 1994; 56:137–157.
17. Hettich R, Still G. Second order optimality conditions for generalized semi-infinite programming problems. Optimization 1995;34:195–211.
18. Rückmann J-J. On existence and uniqueness of stationary points in semi-infinite optimization. Math Program 1999;86:387–415.
19. Gauvin J. A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. Math Program 1977;12:136–138.
20. Kummer B. Lipschitzian inverse functions, directional derivatives and application in $C^{1,1}$ optimization. J Optim Theory Appl 1991; 70:561–582.
21. Bonnans JF, Shapiro A. Perturbation analysis of optimization problems. New York: Springer; 2000.
22. Klatte D, Kummer B. Nonsmooth equations in optimization: regularity, calculus, methods and applications. Dordrecht: Kluwer Academic Publishers; 2002.

23. Jongen HTh, Rückmann J-J. On stability and deformation in semi-infinite optimization. In: Reemtsen R. and Rückmann J-J, editors. *Semi-Infinite Programming*. Boston (MA): Kluwer Academic Publishers; 1998. pp. 29–67.
24. Jongen HTh, Twilt F, Weber G-W. Semi-infinite optimization: structure and stability of the feasible set. *J Optim Theory Appl* 1992;72:523–552.
25. Günzel H, Jongen HTh. Strong stability implies Mangasarian-Fromovitz constraint qualification. *Optimization* 2006;55:605–610.
26. Reemtsen R, Görner S. Numerical methods for semi-infinite programming. In: Reemtsen R. and Rückmann J-J, editors. *Semi-infinite programming*. Boston (MA): Kluwer Academic Publishers; 1998. pp. 195–275.
27. Still G. Discretization in semi-infinite programming: the rate of approximation. *Math Program* 2001;91:53–69.
28. Bhattacharjee B, Green WH, Barton PI. Interval methods for semi-infinite programs. *Comput Optim Appl* 2005;30(1):63–93.
29. Mitsos A, Lemonidis P, Kun Lee C, *et al.* Relaxation-based bounds for semi-infinite programs. *SIAM J Optim* 2008;19(1):77–113.
30. Bhattacharjee B, Lemonidis P, Green WH, *et al.* Global solutions of semi-infinite programs. *Math Program* 2005;103(2):283–307.
31. Floudas CA, Stein O. The adaptive convexification algorithm: a feasible point method for semi-infinite programming. *SIAM J Optim* 2007;18:1187–1208.
32. Stein O, Tezel A. The semi-smooth approach for semi-infinite programming without strict complementarity. *SIAM J Optim* 2009;20:1052–1072.
33. Stein O. On Karush-Kuhn-Tucker points for a smoothing method in semi-infinite optimization. *J Comput Math* 2006;24:719–732.
34. Stein O, Still G. Solving semi-infinite optimization problems with interior point techniques. *SIAM J Control Optim* 2003;42:769–788.
35. Stein O, Still G. On generalized semi-infinite optimization and bilevel optimization. *Eur J Oper Res* 2002;142:444–462.
36. Günzel H, Jongen HTh, Stein O. Generalized semi-infinite programming: on generic local minimizers. *J Glob Optim* 2008;42:413–421.
37. Guerra-Vázquez F, Jongen HTh, Shikhman V. General semi-infinite programming: symmetric Mangasarian-Fromovitz constraint qualification and the closure of the feasible set. *SIAM J Optim*. 2010;20:2487–2503.
38. Jongen HTh, Shikhman V. General semi-infinite programming: critical point theory. *Optimization*. Preprint Lehrstuhl C Mathematik, Nr. 139; 2010. (In press).

SEMI-MARKOV DECISION PROCESSES

MELIKE BAYKAL-GÜRSOY
Department of Industrial and
Systems Engineering, RUTCOR,
CAIT, Rutgers University, New
Brunswick, New Jersey

INTRODUCTION

Semi-Markov decision processes (SMDPs) are used in modeling stochastic control problems arising in Markovian dynamic systems, where the sojourn time in each state is a general continuous random variable. They are powerful, natural tools for the optimization of queues [1–7], production scheduling [8–11], and reliability/maintenance [12].

For example, in a machine replacement problem with deteriorating performance over time, a decision maker, after observing the current state of the machine, decides whether to continue its usage, or initiate a maintenance (preventive or corrective) repair, or replace the machine. There are a reward or cost structure associated with the states and decisions, and an information pattern available to the decision maker. This decision depends on a performance measure over the planning horizon which is either finite or infinite, such as total expected discounted or long run average expected reward/cost with or without external constraints, and variance penalized average reward.

SMDPs are based on semi-Markov processes (SMPs) [13] (see also *Semi-Markov Processes* SMPs), that include renewal processes (see also *Definition and Examples of Renewal Processes*) and continuous-time Markov chains (CTMCs) (see also *Definition and Examples of Continuous-Time Markov Chains*) as special cases. In an SMP similar to Markov chains (DTMCs) (see also *Definition and Examples of DTMCs*), state changes occur according to the Markov

property, that is, states in the future do not depend on the states in the past given the present. However, the sojourn time in a state is a continuous random variable with distribution depending on that state and the next state, a Markov chain is a SMP in which the sojourn times are discrete (geometric) random variables independent of the next state; a CTMC is a SMP with exponentially distributed sojourn times; and a renewal process is a SMP with a single state. SMDPs, first introduced by Jewell [14] and De Cani [15], are also called as *Markov renewal programs* [16–19].

This article is organized as follows: the next section introduces basic definitions and notations. Various performance criteria are presented in the section titled “Performance Measures” and their solution methodologies are described in the sections titled “Discounted Reward Criterion,” “Average Reward Criterion,” and “Expected Time-Average Reward and Variability.”

BASIC DEFINITIONS

We consider time-homogeneous, finite state, and finite action SMDPs, and give references for the more general cases. Let $\{X_m, m \geq 0\}$ denote the state process, which takes values in a finite state space $\mathcal{S}$. We also use $\{X_m, m \in \mathcal{N}\}$ to denote the state process with $\mathcal{N}$ representing the set of nonnegative integers. At each epoch m , the decision maker chooses an action A_m from a finite action space $\mathcal{A}$. The sojourn time between the $(m - 1)$ -st and the (m) th epochs is a random variable and denoted by γ_m . The underlying sample-space $\Omega = \{\mathcal{S} \times \mathcal{A} \times (0, \infty)\}^\infty$ consists of all possible realizations of states, actions, and the transition times. Throughout, the sample space will be equipped with the σ -algebra generated by the random variables $\{X_m, A_m, \gamma_{m+1}; m \geq 0\}$. The initial state is assumed to be fixed and given. Note that we will suppress the dependence on the initial state unless given otherwise. Denote

P_{xy} , $x \in \mathcal{S}$, $a \in \mathcal{A}$, $y \in \mathcal{S}$, for the law of motion of the process, that is, for all policies $\mathbf{u}$ and all epochs m

$$P_{\mathbf{u}}\{X_{m+1} = y | X_0, A_0, \Upsilon_1, \dots, X_m = x, A_m = a\} = P_{xy}.$$

Also conditioned on the event that the next state is y , Υ_{m+1} has the distribution function $F_{xy}(\cdot)$, that is,

$$P_{\mathbf{u}}\{\Upsilon_{m+1} \leq t | X_0, A_0, \Upsilon_1, \dots, X_m = x, A_m = a, X_{m+1} = y\} = F_{xy}(t).$$

Assume that $F_{xy}(0) < 1$.

The process $\{S_t, B_t : t \geq 0\}$, where S_t is the state of the process at time t , and B_t is the action taken at time t , is referred to as the *SMDP*. Let $T_n = \sum_{m=1}^n \Upsilon_m$, that is, denote the time of n th transition. For $t \in [T_m, T_{m+1})$, clearly

$$S_t = X_m, \quad B_t = A_m.$$

Policy Types

A *decision rule* $\mathbf{u}^m$ at epoch m is a vector consisting of probabilities assigned to each available action. A decision rule may depend on all of the previous states, actions, transition times, and the present state. Let u_a^m denote the a th component of $\mathbf{u}^m$. Thus, it is the conditional probability of choosing action a at the m -th epoch, that is,

$$P_{\mathbf{u}}\{A_m = a | X_0 = x_0, A_0 = a_0, \Upsilon_1 = \tau_1, \dots, X_m = x\} = u_a^m(x_0, a_0, \tau_1, \dots, x).$$

A *policy* is an infinite sequence of decision rules $\mathbf{u} = \{\mathbf{u}^0, \mathbf{u}^1, \mathbf{u}^2, \dots\}$.

Policy $\mathbf{u}$ is called *Markov policy* if $\mathbf{u}^m$ at epoch m depends only on the current state not the past history, that is,

$$\mathbf{u}^m(x) = \mathbf{u}_a^m(x_0, a_0, \tau_1, \dots, x).$$

A policy is called *stationary* if the decision rule at each epoch is the same and it depends only on the present state of the process, $\mathbf{u} = \{\mathbf{u}, \mathbf{u}, \mathbf{u}, \dots\}$; denote f_{xa} for the probability of choosing action a when in state x . A

stationary policy is said to be *pure* if for each $x \in \mathcal{S}$ there is only one action $a \in \mathcal{A}$ such that $f_{xa} = 1$. Let U, M, F , and G denote the set of all policies, Markov policies, stationary policies, and pure policies, respectively. Clearly, $G \subset F \subset M \subset U$.

Under a stationary policy $\mathbf{f}$, the state process $\{S_t : t \geq 0\}$ is a SMP, while the process $\{X_m : m \in \mathcal{N}\}$ is the embedded Markov chain with transition probabilities

$$P_{xy}(\mathbf{f}) = \sum_{a \in \mathcal{A}} P_{xy} f_{xa}.$$

Clearly, the process $\{S_t, B_t : t \geq 0\}$ is also a SMP under a stationary policy $\mathbf{f}$ with the embedded Markov chain $\{X_m, A_m : m \in \mathcal{N}\}$.

Chain Structure

Under a stationary policy $\mathbf{f}$, state x is *recurrent* if and only if x is recurrent in the embedded Markov chain; similarly, x is *transient* if and only if x is transient for the embedded Markov chain. A SMDP is said to be *unichain* (*multichain*) if the embedded Markov chain for each pure policy is unichain (multichain), that is, if the transition matrix $P(\mathbf{g})$ has at most one (more than one) recurrent class plus (a perhaps empty) set of transient states for all pure policies $\mathbf{g}$. It is called *irreducible* if $P(\mathbf{g})$ is irreducible under all pure policies $\mathbf{g}$. Similarly, a SMDP is said to be *communicating* if $P(\mathbf{f})$ is irreducible for all stationary policies that satisfy $f_{xa} > 0$, for all $x \in \mathcal{S}$, $a \in \mathcal{A}$.

Let $\tau(x, a)$ define the expected sojourn time given that the state is x and the action a is chosen just before a transition, that is,

$$\begin{aligned} \tau(x, a) &\triangleq E_{\mathbf{u}}[\Upsilon_m | X_{m-1} = x, A_{m-1} = a] \\ &= \int_0^\infty \sum_{y \in \mathcal{S}} P_{\mathbf{u}}\{X_m = y, \Upsilon_m > t \\ &\quad | X_{m-1} = x, A_{m-1} = a\} dt \\ &= \int_0^\infty \left[1 - \sum_{y \in \mathcal{S}} P_{xy} F_{xy}(t) \right] dt. \end{aligned}$$

Let $W_t(x, a)$ denote the random variables representing the state-action intensities,

$$W_t(x, a) \triangleq \frac{1}{t} \int_0^t \mathbf{1}\{(S_s, B_s) = (x, a)\} ds,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Let U_0 denote the class of all policies $\mathbf{u}$ such that $\{W_t(x, a); t \geq 0\}$ converges. Thus, for $\mathbf{u} \in U_0$, there exist random variables $\{W(x, a)\}$ such that

$$\lim_{t \rightarrow \infty} W_t(x, a) = W(x, a).$$

Let U_1 be the class of all policies $\mathbf{u}$ such that the expected state-action intensities $\{E_{\mathbf{u}}[W_t(x, a)]; t \geq 0\}$ converge for all x and a . For $\mathbf{u} \in U_1$ denote

$$w_{\mathbf{u}}(x, a) = \lim_{t \rightarrow \infty} E_{\mathbf{u}}[W_t(x, a)].$$

From Lebesgue's Dominated Convergence Theorem $U_0 \subset U_1$.

A well-known result from renewal theory [13] is that if $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous SMP, and if the embedded Markov chain $\{X_m, m \in \mathcal{N}\}$ is unichain then, the proportion of time spent in state y , that is,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbf{1}\{Y_s = y\} ds,$$

exists. Since under a stationary policy $\mathbf{f}$, the process $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous SMP, if the embedded Markov decision process is unichain, then the limit of $W_t(x, a)$ as t goes to infinity exists and the proportion of time spent in state x when action a is applied is given as

$$W(x, a) = \lim_{t \rightarrow \infty} W_t(x, a) = \frac{\tau(x, a) Z(x, a)}{\sum_{x, a} \tau(x, a) Z(x, a)},$$

where $Z(x, a)$ denotes the associated state-action frequencies. Let $\{z_{\mathbf{f}}(x, a); x \in \mathcal{S}, a \in \mathcal{A}\}$ denote the expected state-action frequencies, that is,

$$\begin{aligned} z_{\mathbf{f}}(x, a) &= \lim_{n \rightarrow \infty} E_{\mathbf{f}} \frac{1}{n} \sum_{m=1}^n \mathbf{1}\{X_{m-1} = x, A_{m-1} = a\} \\ &= \pi_x(\mathbf{f}) f_{xa}, \end{aligned}$$

where $\pi_x(\mathbf{f})$ is the steady-state distribution of the embedded Markov chain $P(\mathbf{f})$.

The long run average number of transitions into state x when action a is applied per

unit time is,

$$\begin{aligned} v_{\mathbf{f}}(x, a) &= \frac{\pi_x(\mathbf{f}) f_{xa}}{\sum_{x, a} \tau(x, a) \pi_x(\mathbf{f}) f_{xa}} \\ &= \frac{z_{\mathbf{f}}(x, a)}{\sum_{x, a} \tau(x, a) z_{\mathbf{f}}(x, a)}. \end{aligned} \quad (1)$$

This gives $w_{\mathbf{f}}(x, a) = \tau(x, a) v_{\mathbf{f}}(x, a)$.

Reward Structure

Let R_t be the reward function at time t . R_t can be an impulse function corresponding to the reward earned immediately at a transition epoch and/or it can be a step function between transition epochs corresponding to the rate of reward as described below. The decision maker earns an immediate reward $R(X_m, A_m)$ and a reward with rate $r(X_m, A_m)$ until the $(m + 1)$ -th epoch, that is,

$$R_t = \begin{cases} R(X_m, A_m), & \text{if } t = T_m, \\ r(X_m, A_m), & \text{if } t \in [T_m, T_{m+1}). \end{cases} \quad (2)$$

Thus,

$$R_{m+1} = R(X_m, A_m) + r(X_m, A_m) \Upsilon_{m+1},$$

is the reward earned during the $(m + 1)$ -th transition [20–22].

Similarly, there is an immediate cost $C(X_m, A_m)$ and a cost with rate $c(X_m, A_m)$ with

$$C_{m+1} = C(X_m, A_m) + c(X_m, A_m) \Upsilon_{m+1}.$$

Hence, at any epoch if the process is in state $x \in \mathcal{S}$ and action $a \in \mathcal{A}$ is chosen, then the reward earned during this epoch is represented by $\bar{r}(x, a) \triangleq R(x, a) + r(x, a) \tau(x, a)$. Similarly, the cost during this epoch is represented by $\bar{c}(x, a) \triangleq C(x, a) + c(x, a) \tau(x, a)$.

Example 1. Consider the *machine replacement problem* mentioned in the section titled “Introduction” with states (1) machine is in good condition, (2) machine has some minor problems, (3) machine is down and needs to be replaced. Time to failure of the machine follows a Weibull distribution with scale parameter equal to 8000 h and shape parameter

equal to 4. The failure is minor with probability 0.95 and major requiring replacement of the machine with probability 0.05. Life time of a machine with minor problems follows a Weibull distribution with scale parameter equal to 20,000 h and shape parameter equal to 4. However, a machine with minor problems could be maintenance repaired to make it as good as new. Maintenance repair takes Weibull distributed amount of time with scale parameter equal to 8 h and shape parameter equal to 0.5. On the other hand, the machine replacement time is normally distributed with mean 72 h and variance of 8 h. Running a fully working machine earns \$100/h, and a machine with minor problem earns \$75/h profit. It costs \$40/h to repair and \$10,000 to replace a machine. Note that there is no control action available in states 1 and 3. In state 1 the decision maker needs to “wait” and in state 3 s/he needs to order a new machine. Let us denote this action as action 1. In state 2, there are two possible actions to choose: “wait” action denoted as action 1, and “initiate repair” denoted as action 2. Parameters of this model are:

$$P_{113} = 0.95, \quad P_{113} = 0.05, \quad P_{213} = 1, \\ P_{221} = 1, \quad P_{311} = 1,$$

$$\tau(1, 1) = 7252, \quad \tau(2, 1) = 1813, \\ \tau(2, 2) = 48, \quad \tau(3, 1) = 72, \\ \bar{\tau}(1, 1) = 725,200, \quad \bar{\tau}(2, 1) = 135,975, \\ \bar{\tau}(2, 2) = -1920, \quad \bar{\tau}(3, 1) = -10,000.$$

The last two reward values correspond to the incurred costs under repair and replacement, respectively.

PERFORMANCE MEASURES

We will focus on the optimality criteria over the infinite horizon, since some general results could be obtained for these models. We will first consider finding a policy $\mathbf{u}$ that

will maximize the *total discounted reward* defined as

$$\phi_\alpha(\mathbf{u}) \triangleq E_{\mathbf{u}} \left[\int_0^\infty e^{-\alpha s} R_s ds \right], \quad (3)$$

where α represents the discount factor [14,23–26]. Discounted reward optimality criterion is easier to analyze and understand than the average reward criterion, because the results for these models hold regardless of the chain structure of the embedded Markov chain. In fact, the existence of this integral is immediate under finite rewards. In addition, discounting lands itself naturally in economic problems in which the present value of future earnings is discounted as a function of the interest rate. Another interpretation of these models implies the importance of the initial decisions.

The great majority of the literature, on the other hand, is concerned with the *long run average expected reward* criterion with

$$\phi_1(\mathbf{u}) \triangleq \liminf_{t \rightarrow \infty} \frac{1}{t} E_{\mathbf{u}} \left[\int_0^t R_s ds \right], \quad (4)$$

ϕ_1 denoting the *long run average expected reward* [14,17,22,27–29]. The following alternative to ϕ_1 is given by Jewell [30], Ross [31,32], and Mine and Osaki [18] as

$$\phi_2(\mathbf{u}) \triangleq \liminf_{n \rightarrow \infty} \frac{E_{\mathbf{u}} \left[\sum_{m=1}^n R_m \right]}{E_{\mathbf{u}}[T_n]}, \quad (5)$$

referred to as the *ratio-average reward* [33]. The performance measure ϕ_2 is also used by other researchers [7,14,34–38].

Let

$$\phi_\alpha^* = \sup_{\mathbf{u} \in U} \phi_\alpha(\mathbf{u}), \quad \phi_1^* = \sup_{\mathbf{u} \in U} \phi_1(\mathbf{u}), \\ \phi_2^* = \sup_{\mathbf{u} \in U} \phi_2(\mathbf{u}).$$

A policy $\mathbf{u}$ is optimal for $\phi_\alpha(\cdot)$ if $\phi_\alpha(\mathbf{u}) = \phi_\alpha^*$. For a fixed $\epsilon > 0$, a policy $\mathbf{u}$ is ϵ -optimal for $\phi_\alpha(\cdot)$ if $\phi_\alpha(\mathbf{u}) > \phi_\alpha^* - \epsilon$. Optimality and ϵ -optimal for $\phi_1(\cdot)$, $\phi_2(\cdot)$ and the other performance measures we consider in this article are defined analogously.

The following *expected time-average reward* criterion has been considered recently by Baykal-Gürsoy and Gürsoy [39–41]

$$\psi(\mathbf{u}) \triangleq E_{\mathbf{u}} \left[\liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t R_s ds \right] \quad (6)$$

subject to the sample path constraint,

$$P_{\mathbf{u}} \left\{ \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t C_s ds \leq \gamma \right\} = 1. \quad (7)$$

This constraint requires that the long-run average costs on almost all sample paths should be bounded by γ .

More generally, they investigate the following *expected time-average variability*

$$v(\mathbf{u}) \triangleq E_{\mathbf{u}} \left[\liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t h(R_s, \frac{1}{t} \int_0^t R_q dq) ds \right], \quad (8)$$

where $h(\cdot, \cdot)$ is a continuous function of the current reward at time s and the average reward over an interval that includes time s . By letting $v^* = \sup_{\mathbf{u} \in U} v(\mathbf{u})$, the optimality and ϵ -optimality for $v(\cdot)$ are analogously defined.

DISCOUNTED REWARD CRITERION

Discounted reward can be rewritten as:

$$\begin{aligned} \phi_{\alpha}(\mathbf{u}) &= E_{\mathbf{u}} \left[\sum_{m=0}^{\infty} e^{-\alpha T_m} \left(R(X_m, A_m) \right. \right. \\ &\quad \left. \left. + \frac{r(X_m, A_m)}{\alpha} (1 - e^{-\alpha \Upsilon_m}) \right) \right] \\ &= \sum_{m=0}^{\infty} \sum_{x,a} \int_0^{\infty} e^{-\alpha t} \left[R(x, a) + \frac{r(x, a)}{\alpha} \right. \\ &\quad \left. \times \left(1 - \sum_y P_{xay} \int_0^{\infty} e^{-\alpha \tau} dF_{xay}(\tau) \right) \right] \\ &\quad P_{\mathbf{u}} \{X_m = x, A_m = a, T_m \leq t\}. \end{aligned}$$

The terms inside the second integral could be recognized as the Laplace transform of the density function $f_{xay}(\cdot)$ and will be denoted as $\tilde{f}_{xay}(\alpha)$.

The optimal discounted reward vector is represented by ϕ_{α}^{*x} for each initial state x , and

it can be shown that it satisfies the optimality equation for all $x \in S$:

$$\begin{aligned} \phi_{\alpha}^x &= \max_a \left\{ \left[R(x, a) + \frac{r(x, a)}{\alpha} \left(1 - \sum_j P_{xaj} \tilde{f}_{xaj} \right) \right] \right. \\ &\quad \left. + \sum_y P_{xay} \tilde{f}_{xay}(\alpha) \phi_{\alpha}^y \right\} \\ &= \max_a \left\{ r^{\alpha}(x, a) + \sum_y P_{xay}^{\alpha} \phi_{\alpha}^y \right\}. \quad (9) \end{aligned}$$

Second equality is obtained from the first by denoting the terms inside the square bracket as $r^{\alpha}(x, a)$ and writing $P_{xay} \tilde{f}_{xay}(\alpha)$ as P_{xay}^{α} . Note that the second equality is similar to the one obtained for the Markov Decision Processes (MDPs) (see also **Total Expected Discounted Reward MDPs: Existence of Optimal Policies**). Thus, discounted SMDPs can be reduced to discounted MDPs by using these transformations. Since $P_{xay}^{\alpha} < 1$, the right hand side of the optimality Equation (9) is a contraction mapping and the next theorem is immediate.

Theorem 1. *For SMDPs under the discounted reward criterion:*

- *There exists a unique solution to the optimality equation (9) and it is equal to ϕ_{α}^* .*
- *There exists an optimal pure policy $\mathbf{g}^*$ given by, $\phi_{\alpha}(\mathbf{g}^*) = \phi_{\alpha}^* = (I - P_{xay}^{\alpha})^{-1} r^{\alpha}(\mathbf{g}^*)$ where $r^{\alpha}(\mathbf{g}^*)$ denotes the single-period discounted reward earned under policy $\mathbf{g}^*$.*

This optimal pure policy could be obtained using the policy iteration (see also **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**), value iteration (see also **Total Expected Discounted Reward MDPs: Value Iteration Algorithm**), or linear programming (LP) algorithms [5–7,38]. The LP algorithm is discussed next. Consider the following LP

with given numbers $\beta_x > 0$ for $x \in \mathcal{S}$.

$$\begin{aligned} & \max \sum_{x \in \mathcal{S}, a \in \mathcal{A}} r^\alpha(x, a) z(x, a) \\ & \text{s.t.} \sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}^\alpha) z(x, a) = \beta_y, \quad y \in \mathcal{S} \\ & \quad z(x, a) \geq 0, \quad x \in \mathcal{S}, a \in \mathcal{A}. \end{aligned}$$

Let $\mathbf{z}^*$ be an optimum solution of the above LP. Clearly, any extreme point of this LP has $|\mathcal{S}|$ number of basic variables, where $|\cdot|$ denotes the number of elements in a given set. Thus, $\mathbf{z}^*(x, a)$ is positive only for one action a . The optimum pure policy $\mathbf{g}^*$ is then obtained by assigning $\mathbf{g}_x^*$ in such a way that $\mathbf{z}^*(x, \mathbf{g}_x^*) > 0$. Constraints defined in a similar fashion,

$$E_{\mathbf{u}} \left[\int_0^\infty e^{-as} C_s ds \right] < \gamma,$$

could be included into the LP as

$$\sum_{x \in \mathcal{S}, a \in \mathcal{A}} c^\alpha(x, a) z(x, a) < \gamma,$$

with $c^\alpha(x, a) = C(x, a) + \frac{c(x, a)}{\alpha} (1 - \sum_y P_{xay} \tilde{f}_{xay})$. Since every new constraint will increase the number of basic variables, the optimum policy will no longer be pure but randomized stationary [24].

For the countable state case, we need the assumption,

Assumption 1. There exists $\delta > 0$ and $\varepsilon > 0$, such that

$$F_{xay}(\delta) \leq 1 - \varepsilon \quad \text{for all } x \text{ and } y \in \mathcal{S} \text{ and } a \in \mathcal{A},$$

together with $|r^\alpha(x, a)| \leq M < \infty$ to ensure the existence of an optimal pure policy. Additional conditions are required for SMDPs with Borel state and action spaces, and unbounded rewards [21, 24, 38].

AVERAGE REWARD CRITERION

Average or *ratio-average* expected reward criterion is applied to systems in which the system dynamics is not slow enough to

warrant discounting. This criterion is more difficult to analyze since the existence of the optimal stationary policy depends on the chain structure. Under the condition that the SMDP is irreducible, $\phi_1(\mathbf{f}) = \phi_2(\mathbf{f})$ for every stationary policy $\mathbf{f}$ [18, 31]. However, this may not hold even for unichain SMDPs [33]. While ϕ_1 is clearly the more appealing criterion, it is easier to write the optimality equations when establishing the existence of an optimal pure policy under criterion ϕ_2 [22, 29, 42]. On the other hand, for finite state and finite action SMDPs, there exists an optimal pure policy under ϕ_1 [27, 29, 42], while such an optimal policy may not exist under ϕ_2 in a general multichain SMDP [43]. Jianyong and Xiaobo [43] investigate average reward SMDPs focusing on ϕ_2 and using a data-transformation method [19]. They show that the optimal pure policy exists in some special cases such as the unichain case and the weakly communicating case.

The optimal pure policy for the average expected reward criterion in multichain SMDPs is obtained from the optimal solution of the following LP [25] under the assumption on the sojourn times.

$$\begin{aligned} & \max \sum_{x \in \mathcal{S}, a \in \mathcal{A}} \bar{r}(x, a) v(x, a) \\ & \text{s.t.} \sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}) v(x, a) = 0, \quad y \in \mathcal{S} \\ & \quad \sum_{a \in \mathcal{A}} \tau(y, a) v(y, a) \\ & \quad + \sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}) t(x, a) = \beta_y, \quad y \in \mathcal{S} \\ & \quad v(x, a) \geq 0, \quad t(x, a) \geq 0 \quad x \in \mathcal{S}, a \in \mathcal{A}, \end{aligned}$$

where $\beta_x > 0$ for $x \in \mathcal{S}$ and $\sum_y \beta_y = 1$. The optimum average expected reward for each initial state is obtained from the dual of this LP.

In the unichain case, the average reward remains constant regardless of the initial

state, and the LP reduces to

$$\begin{aligned} \phi_1^* &= \max \sum_{x \in \mathcal{S}, a \in \mathcal{A}} \bar{r}(x, a) v(x, a) \\ \text{s.t. } &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}) v(x, a) = 0, \quad y \in \mathcal{S} \\ &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} \tau(y, a) v(y, a) = 1 \\ &v(x, a) \geq 0, \quad x \in \mathcal{S}, a \in \mathcal{A}, \end{aligned}$$

with the optimum solution denoted as $\mathbf{v}^*$. The optimum pure policy $\mathbf{g}^*$ is then obtained by assigning $\mathbf{g}_x^*$ in such a way that $v^*(x, \mathbf{g}_x^*) > 0$. The optimality equations are given for each state x by

$$\zeta_x = \max_a \left\{ \bar{r}(x, a) - g\tau(x, a) + \sum_y P_{xay} \zeta_y \right\}.$$

The solution to these equations, $\{\zeta^*, \mathbf{g}^*\}$ provides the optimum average expected reward, $\phi_1^* = \mathbf{g}^*$.

The constrained problem has been investigated for the average reward SMDPs [20,33,34]. Beutler and Ross [20,34] consider the ratio-average reward with a constraint under a condition stronger than the unichain condition. In Ref. 33, Feinberg examines the problem of maximizing both ϕ_1 and ϕ_2 , subject to a number of constraints. Under the condition that the initial distribution is fixed, he shows that for both criteria, there exist optimal mixed stationary policies when an associated LP is feasible. The mixed policies are defined as policies with an initial one-step randomization applied to a set of pure policies, hence they are not stationary. He provides an LP algorithm for the unichain SMDP under both criteria. Average expected reward SMDPs with Borel state and action spaces and unbounded rewards are considered by Schäl [42], Sennott [21], and Luque-Vásquez and Hernández-Lerma [44].

EXPECTED TIME-AVERAGE REWARD AND VARIABILITY

The expected time-average reward criterion is similar to the average expected reward criterion. Fatou's lemma immediately implies

that $\psi(\mathbf{u}) \leq \phi_1(\mathbf{u})$ holds for all policies. Baykal-Gürsoy and Gürsoy [40] show that for a large class of policies, these two rewards are equal and an ϵ -optimal randomized stationary policy can be obtained for the general (communicating, multichain) SMDP, while such a policy may not exist for the average reward problem [40]. Multiple constraints and the more general expected time-average variability criterion are also discussed. They show that an ϵ -optimal stationary policy can be obtained for the general SMDPs. If $h(x, y) = x - \lambda(x - y)^2$, then the optimal policy is a pure policy. Note that in this case maximizing $v(\mathbf{u})$ corresponds to maximizing the expected average reward penalized by the expected average variability. A decomposition algorithm to locate the ϵ -optimal stationary policy for both problems is given in Ref. 40. This algorithm utilizes an LP of the form:

$$\begin{aligned} \max &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} h \left[\bar{r}(x, a), \sum_{y \in \mathcal{S}, b \in \mathcal{A}} \bar{r}(y, b) v(y, b) \right] v(x, a) \\ \text{s.t. } &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}) v(x, a) = 0, \quad y \in \mathcal{S} \\ &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} \tau(x, a) v(x, a) = 1 \\ &\sum_{x \in \mathcal{S}, a \in \mathcal{A}} \bar{c}(x, a) v(x, a) \leq \gamma \\ &v(x, a) \geq 0, \quad x \in \mathcal{S}, a \in \mathcal{A}. \end{aligned}$$

REFERENCES

1. Hajek B. Optimal control of two interacting service stations. *IEEE Trans Automatic Control* 1984;29:491–499.
2. Stidham S, Weber RR Jr. A survey of Markov decision models for control of networks of queues. *Queueing Syst* 1993;13:291–314.
3. Sennott LI. *Stochastic dynamic programming and the control of queueing systems*. New York: Wiley and Sons; 1999.
4. Feinberg EA, Kella O. Optimality of D-policies for an M/G/1 queue with a removable server. *Queueing Syst* 2002;42:355–376.
5. Tijms HC. *Stochastic modeling and analysis: a computational approach*. Chichester: Wiley; 1986.
6. Tijms HC. *A first course in stochastic models*. Chichester, England: Wiley; 2003.

7. Heyman DP, Sobel MJ. Stochastic models in operations research-volume II. New York: McGraw-Hill Book Company; 1984.
8. Ross SM. Introduction to stochastic dynamic programming. New York: Academic Press; 1983.
9. Pinedo M. Scheduling: theory, algorithms, and systems. Englewood Cliffs: Prentice Hall; 1995.
10. Baykal-Gürsoy M, Altiok T, Danhong H. Control policies for two-stage production/inventory systems with constrained average cost criterion. *Int J Prod Res* 1994;32:2005–2014.
11. Baykal-Gürsoy M, Altiok T, Danhong H. Look-back policies for two-stage, pull-type production/inventory systems. *Ann Oper Res* 1994;48:381–400. Special Issue on Queueing Networks.
12. Hu Q, Yue W. Optimal replacement of a system according to a semi-Markov decision process in a semi-Markov environment. *Optim Methods Softw* 2003;18(2):181–196.
13. Çinlar E. Introduction to stochastic processes. New Jersey: Prentice Hall; 1975.
14. Jewell WS. Markov renewal programming I: Formulation, finite return models. *J Oper Res* 1963;11:938–948.
15. De Cani JS. A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity. *Manage Sci* 1964;10(4):716–733.
16. Denardo EV. Markov renewal programs with small interest rate. *Ann Math Stat* 1971;42:477–496.
17. Fox B. Markov renewal programming by linear fractional programming. *SIAM J Appl Math* 1966;16:1418–1432.
18. Mine H, Osaki S. Markovian decision processes. New York: Elsevier; 1970.
19. Schweitzer PJ. Iterative solution of the functional equations of undiscounted Markov renewal programming. *J Math Anal Appl* 1971;34:495–501.
20. Beutler FJ, Ross KW. Time-average optimal constrained semi-Markov decision processes. *Adv App Prob* 1986;18:341–359.
21. Sennott LI. Average cost semi-Markov decision processes and the control of queueing systems. *Prob Eng Info Sci* 1989;3:247–272.
22. Yushkevich AA. On semi-Markov controlled models with an average reward criterion. *Theory Probab Appl* 1981;26:796–802.
23. Schweitzer PJ, Puterman ML, Kindle KW. Iterative aggregation-disaggregation procedures for solving discounted semi-Markovian reward processes. *Oper Res* 1985;33:589–605.
24. Feinberg EA. Constrained discounted semi-Markov decision processes. In: How Z, Filar JA, Chen A, editors. *Markov Processes and Controlled Markov Chains*. Dordrecht, The Netherlands: Kluwer; 2002. pp. 233–244.
25. Kallenberg LCM. Linear programming and finite Markovian control problems. Volume 148, Amsterdam: Mathematical Centre Tracts; 1983.
26. Lippman SA. On dynamic programming with unbounded rewards. *Manage Sci* 1975;21(11):1225–1233.
27. Denardo EV, Fox BL. Multichain Markov renewal programs. *SIAM J Appl Math* 1968;16:468–487.
28. Lippman SA. Maximal average reward policies for semi-Markov renewal processes with arbitrary state and action spaces. *Ann Math Stat* 1971;42:1717–1726.
29. Schweitzer PJ, Federgruen AF. The functional equations of undiscounted Markov renewal programming. *Math Oper Res* 1978;3:308–321.
30. Jewell WS. Markov renewal programming II: Infinite return models, example. *J Oper Res* 1963;11:949–971.
31. Ross SM. Average cost semi-Markov processes. *J Appl Probab* 1970;7:649–656.
32. Ross SM. Applied probability models with optimization applications. San Francisco: Holden-Day; 1971.
33. Feinberg EA. Constrained semi-Markov decision processes with average rewards. *Math Meth Oper Res* 1994;39:257–288.
34. Beutler FJ, Ross KW. Uniformization for semi-Markov decision processes under stationary policies. *Adv App Probab* 1987;24:644–656.
35. Federgruen A, Tijms HC. The optimality equation in average cost denumerable state semi-Markov decision problems, recurrence conditions and algorithms. *J Appl Probab* 1978;15:356–373.
36. Federgruen A, Hordijk A, Tijms HC. Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion. *Stoch Proc Appl* 1979;9:223–235.
37. Federgruen A, Schweitzer PJ, Tijms HC. Denumerable undiscounted semi-Markov decision processes with unbounded rewards. *Math Oper Res* 1983;8(2):298–313.

38. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley and Sons; 1994.
39. Baykal-Gürsoy M. Semi-Markov decision processes: nonstandard criteria. *IEEE Regional NY/NJ Conference*, New Brunswick (NJ) 1994.
40. Baykal-Gürsoy M, Gürsoy K. Semi-Markov decision processes: Nonstandard criteria. *Prob Eng Info Sci* 2007;21:635–657.
41. Baykal-Gürsoy M, Ross KW. Variability sensitive Markov decision processes. *Math Oper Res* 1992;17:558–571.
42. Schäl M. On the second optimality equation for semi-Markov decision models. *Math Oper Res* 1992;17(2):470–486.
43. Jianyong L, Xiaobo Z. On average reward semi-Markov decision processes with a general multichain structure. *Math Oper Res* 2004;29(2):339–352.
44. Luque-Vásquez F, Hernández-Lerma O. Semi-Markov control models with average costs. *Appl Math* 1999;26(3):315–331.

SEMI-MARKOV PROCESSES AND HIDDEN MODELS

NIKOLAOS LIMNIOΣ
Laboratoire de Mathématiques
Appliquées, Université de
Technologie de Compiègne,
Compiègne, France

MARKOV RENEWAL AND SEMI-MARKOV PROCESSES

Semi-Markov processes (SMPs) were introduced independently by W.L. Smith and P. Lévy in the same Mathematical meeting in 1954. SMPs constitute a generalization of Markov and renewal processes. Jump Markov processes, Markov chains, renewal processes—ordinary, alternating, delayed, and stopped—are particular cases of SMPs. As Feller [1] pointed out, the basic theory of SMPs was introduced by Pyke [2,3]. Further significant results have been obtained by Pyke and Schaufele [4,5], Çinlar [6,7], Koroliuk [8–10], and many others. Also, see Refs 11 and 12. Currently, SMPs have achieved significant importance in probabilistic and statistical modeling. Our main references for this article are Refs 13–15.

Consider an infinite countable set, say E , and an E -valued pure jump stochastic process $Z = (Z_t)_{t \in \mathbb{R}_+}$. Let $0 = S_0 \leq S_1 \leq \dots \leq S_n \leq S_{n+1} \leq \dots$ be the jump times of Z , and $J_0, J_1, J_2, \dots$ the successive visited states of Z . Note that S_0 may also take positive values. Let $\mathbb{N}$ be the set of nonnegative integers.

Definition 1. The stochastic process $(J_n, S_n)_{n \in \mathbb{N}}$ is said to be a Markov renewal process (MRP), with state space E , if it satisfies a.s., the following equality

$$\begin{aligned} \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_0, \dots, J_n; \\ S_1, \dots, S_n) \\ = \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n) \end{aligned}$$

for all $j \in E$, all $t \in \mathbb{R}_+$ and all $n \in \mathbb{N}$. In this case, Z is called a *semi-Markov process*.

We assume that the above probability is independent of n and S_n , and, in this case, the MRP is called *time homogeneous*. The MRP $(J_n, S_n)_{n \in \mathbb{N}}$ is determined by the transition kernel $Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n = i)$, called the *semi-Markov kernel*, and the *initial distribution* α , with $\alpha(i) = \mathbb{P}(J_0 = i)$, $i \in E$. The process (J_n) is a Markov chain with state space E and transition probabilities $P(i, j) := Q_{ij}(\infty) := \lim_{t \rightarrow \infty} Q_{ij}(t)$, called the *embedded Markov chain (EMC)* of Z . It is worth noticing that here $Q_{ii}(t) \equiv 0$, for all $i \in E$, but in general we can consider semi-Markov kernels by dropping this hypothesis.

The SMP Z is connected to (J_n, S_n) by

$$\begin{aligned} Z_t = J_n, \text{ if } S_n \leq t < S_{n+1}, t \geq 0 \quad \text{and} \\ J_n = Z_{S_n}, n \geq 0. \end{aligned}$$

A Markov process with state space $E = \mathbb{N}$ and generating matrix $A = (a_{ij})_{i, j \in E}$ is a special SMP with semi-Markov kernel

$$Q_{ij}(t) = \frac{a_{ij}}{a_i}(1 - e^{-a_i t}), \quad i \neq j, \quad a_i \neq 0,$$

where $a_i := -a_{ii}$, $i \in E$, and $Q_{ij}(t) = 0$, if $i = j$ or $a_i = 0$.

Also, define $X_n := S_n - S_{n-1}$, $n \geq 1$, the interjump times, and the process $(N(t))_{t \in \mathbb{R}_+}$, which counts the number of jumps of Z in the time interval $(0, t]$, by $N(t) := \sup\{n \geq 0 : S_n \leq t\}$. Also, define $N_i(t)$ to be the number of visits of Z to state $i \in E$ in the time interval $(0, t]$. To be specific,

$$N_i(t) := \sum_{n=1}^{N(t)} \mathbf{1}_{\{J_n=i\}} = \sum_{n=1}^{\infty} \mathbf{1}_{\{J_n=i, S_n \leq t\}},$$

and also

$$N_i^*(t) = \mathbf{1}_{\{J_0=i, S_0 \leq t\}} + N_i(t).$$

If we consider the (eventually delayed) renewal process $(S_n^i)_{n \geq 0}$ of successive times

of visits to state i , then $N_i(t)$ is the counting process of renewals. Denote by μ_{ii} the mean recurrence time of (S_n^i) , that is, $\mu_{ii} = \mathbb{E}[S_2^i - S_1^i]$.

Let us denote by $Q(t) = (Q_{ij}(t), i, j \in E)$, $t \geq 0$, the semi-Markov kernel of Z . Then we can write

$$\begin{aligned} Q_{ij}(t) &:= \mathbb{P}(J_{n+1} = j, X_{n+1} \leq t \mid J_n = i) \\ &= P(i, j)F_{ij}(t), \quad t \geq 0, \quad i, j \in E, \end{aligned} \quad (1)$$

where $P(i, j) := \mathbb{P}(J_{n+1} = j \mid J_n = i)$ is the transition kernel of the EMC (J_n) , and $F_{ij}(t) := \mathbb{P}(X_{n+1} \leq t \mid J_n = i, J_{n+1} = j)$ is the conditional distribution function of the sojourn time in the state i given that the next visited state is j , ($j \neq i$). Let us also define the distribution function $H_i(t) := \sum_{j \in E} Q_{ij}(t)$ and its mean value m_i , which is the mean sojourn time of Z in state i . In general, Q_{ij} is a subdistribution, that is, $Q_{ij}(\infty) \leq 1$, hence H_i is a distribution function, $H_i(\infty) = 1$, and $Q_{ij}(0-) = H_i(0-) = 0$.

A special case of SMPs is the one where $F_{ij}(\cdot)$ does not depend on j , that is, $F_{ij}(t) \equiv F_i(t) \equiv H_i(t)$, and

$$Q_{ij}(t) = P(i, j)F_i(t). \quad (2)$$

Any general SMP can be transformed into one of the form Equation (2), (see, e.g., Ref. 13).

Let $\phi(i, t)$, $i \in E, t \geq 0$, be a real-valued measurable function and define the convolution of ϕ by Q as follows:

$$Q * \phi(i, t) := \sum_{k \in E} \int_0^t Q_{ik}(ds) \phi(k, t-s). \quad (3)$$

Now, consider the n -fold convolution of Q by itself. For any $i, j \in E$,

$$Q_{ij}^{(n)}(t) = \begin{cases} \sum_{k \in E} \int_0^t Q_{ik}(ds) Q_{kj}^{(n-1)}(t-s) & n \geq 2 \\ Q_{ij}(t) & n = 1 \\ \delta_{ij} \mathbf{1}_{\{t \geq 0\}} & n = 0, \end{cases}$$

where $\delta_{ij} = 1$, if $i = j$ and $= 0$ otherwise.

It is easy to prove (e.g., by induction) the following fundamental equality:

$$Q_{ij}^{(n)}(t) = \mathbb{P}_i(J_n = j, S_n \leq t). \quad (4)$$

Here, as usual, $\mathbb{P}_i(\cdot)$ means $\mathbb{P}(\cdot \mid J_0 = i)$, and $\mathbb{E}_i$ is the corresponding expectation.

Let us define the Markov renewal function $\psi_{ij}(t)$, $i, j \in E, t \geq 0$, by

$$\begin{aligned} \psi_{ij}(t) &:= \mathbb{E}_i[N_j^*(t)] = \mathbb{E}_i \sum_{n=0}^{\infty} \mathbf{1}_{\{J_n = j, S_n \leq t\}} \\ &= \sum_{n=0}^{\infty} \mathbb{P}_i(J_n = j, S_n \leq t) = \sum_{n=0}^{\infty} Q_{ij}^{(n)}(t). \end{aligned} \quad (5)$$

Another important function is the semi-Markov transition function

$$P_{ij}(t) := \mathbb{P}(Z_t = j \mid Z_0 = i), \quad i, j \in E, t \geq 0,$$

which is the conditional marginal law of the process. We will study this function in the next section.

Definition 2. The SMP Z is said to be regular if

$$\mathbb{P}_i(N(t) < \infty) = 1,$$

for any $t \geq 0$ and any $i \in E$.

For regular SMPs we have $S_n < S_{n+1}$, for any $n \in \mathbb{N}$, and $S_n \rightarrow \infty$. In the sequel, we are concerned with regular SMPs. The following theorem gives two criteria for regularity:

Theorem 1. A SMP is regular if one of the following conditions is satisfied:

1. for every sequence $(j_0, j_1, \dots) \in E^\infty$ and every $C > 0$, at least one of the series

$$\sum_{k \geq 0} [1 - F_{j_k j_{k+1}}(C)], \quad \sum_{k \geq 0} \int_0^C t F_{j_k j_{k+1}}(dt),$$

diverges [2];

2. there exist constants, say $\alpha > 0$ and $\beta > 0$, such that $H_i(\alpha) < 1 - \beta$, for all $i \in E$ [16].

Let us now discuss the nature of different states of an MRP. An MRP is irreducible, if, and only if, its EMC (J_n) is irreducible. A

state i is recurrent (transient) in the MRP, if, and only if, it is recurrent (transient) in the EMC. For an irreducible finite MRP, a state i is positive recurrent in the MRP, if, and only if, it is recurrent in the EMC and if for all $j \in E$, $m_j < \infty$. If the EMC of an MRP is irreducible and recurrent, then all the states are positive recurrent, if, and only if, $m := \nu m := \sum_i v_i m_i < \infty$, and null recurrent, if, and only if, $m = \infty$ (where ν is the stationary probability of EMC (J_n)). A state i is said to be *periodic* with period $a > 0$ if $G_{ii}(\cdot)$ (the distribution function of the random variable $S_2^i - S_1^i$) is discrete concentrated on $\{ka : k \in \mathbb{N}\}$. Such a distribution is also said to be *periodic*. In the opposite case, it is called *aperiodic*. Note that the term *period* has a completely different meaning from the corresponding one of the classic Markov chain theory.

MARKOV RENEWAL THEORY

As the renewal equation in the case of the renewal process theory on the half-real line, the MRE is a basic tool in the theory of SMPs.

Let us write the Markov renewal function given by Equation (5) in matrix form

$$\psi(t) = (I(t) - Q(t))^{(-1)} = \sum_{n=0}^{\infty} Q^{(n)}(t). \quad (6)$$

This can also be written as

$$\psi(t) = I(t) + Q * \psi(t), \quad (7)$$

where $I(t) = I$ (the identity matrix), if $t \geq 0$ and $I(t) = 0$, if $t < 0$.

Equation (7) is a special case of what is called a *Markov renewal equation* (MRE). A general MRE is as follows:

$$\Theta(t) = L(t) + Q * \Theta(t), \quad (8)$$

where $\Theta(t) = (\Theta_{ij}(t))_{i,j \in E}$, $L(t) = (L_{ij}(t))_{i,j \in E}$ are matrix-valued measurable functions, with $\Theta_{ij}(t) = L_{ij}(t) = 0$ for $t < 0$. The function $L(t)$ is a given matrix-valued function and $\Theta(t)$ is an unknown matrix-valued function. We may also consider a vector version of

Equation (8), that is, consider corresponding columns of the matrices Θ and L .

Let $\mathbf{B}$ be the space of all locally bounded, on $\mathbb{R}_+$, matrix functions $\Theta(t)$, that is, $\|\Theta(t)\| = \sup_{i,j} |\Theta_{ij}(t)|$ is bounded on sets $[0, \xi]$, for every $\xi \in \mathbb{R}_+$. Also, denote by $\bar{H}_i(t) := 1 - H_i(t)$.

Theorem 2. (Markov Renewal Theorem [17]). *Let the following conditions be fulfilled:*

1. *The EMC (J_n) is ergodic, that is, irreducible and positive recurrent, with stationary probability $\nu = (\nu_i, i \in E)$.*
2. *The mean sojourn time in every state is finite, i.e. that is, for every $i \in E$,*

$$m_i := \int_0^{\infty} \bar{H}_i(t) dt < \infty, \text{ and}$$

$$m := \sum_{i \in E} \nu_i m_i > 0.$$

3. *The distribution functions $H_i(t)$, $i \in E$, are nonperiodic.*
4. *The functions $L_{ij}(t)$, $t \geq 0$, are direct Riemann integrable, that is, they satisfy the following two conditions, for any $i, j \in E$:*

$$\sum_{n \geq 0} \sup_{n \leq t \leq n+1} L_{ij}(t) < \infty,$$

and

$$\lim_{\Delta \downarrow 0} \left\{ \Delta \sum_{n \geq 0} \left[\sup_{n\Delta \leq t \leq (n+1)\Delta} L_{ij}(t) - \inf_{n\Delta \leq t \leq (n+1)\Delta} L_{ij}(t) \right] \right\} = 0.$$

*Then Equation (8) has a unique solution $\Theta = \psi * L(t)$ belonging to $\mathbf{B}$, and*

$$\lim_{t \rightarrow \infty} \Theta_{ij}(t) = \frac{1}{m} \sum_{\ell \in E} \nu_{\ell} \int_0^{\infty} L_{\ell j}(t) dt. \quad (9)$$

The following result is an important application of the above theorem.

Proposition 1. *The transition function $P(t) = (P_{ij}(t))$ satisfies the following MRE*

$$P(t) = I(t) - H(t) + Q * P(t),$$

which, under conditions (1)–(3) of Theorem 2, has the unique solution

$$P(t) = \psi * (I(t) - H(t)), \quad (10)$$

and, for any $i, j \in E$,

$$\lim_{t \rightarrow \infty} P_{ji}(t) = v_i m_i / m =: \pi_i. \quad (11)$$

Here $H(t) = \text{diag}(H_i(t))$ is a diagonal matrix.

It is worth noticing that, in general, the stationary distribution π of the SMP Z is not equal to the stationary distribution v of the EMC (J_n) . Nevertheless, we have $\pi = v$ when, for example, m_i is independent of $i \in E$.

STATISTICAL INFERENCE

Statistical inference for SMPs is provided in several papers. Moore and Pyke [18] studied empirical estimators for finite semi-Markov kernels; Lagakos *et al.* [19] gave maximum likelihood estimators for nonergodic finite semi-Markov kernels; Akritas and Roussas [20] gave parametric local asymptotic normality results for SMPs; Gill [21] studied Kaplan–Meier type estimators by point process theory; Greenwood and Wefelmeyer [22] studied efficiency of empirical estimators for linear functionals in the case of a general state space; Ouhbi and Limnios [23] studied nonparametric estimators of semi-Markov kernels, nonlinear functionals of semi-Markov kernels, including Markov renewal matrices and reliability functions [15], and rate of occurrence of failure functions [24]. Also see Ref. 25. We will give here some elements of the nonparametric estimation of semi-Markov kernels.

Let us consider an observation of an irreducible SMP Z , with finite state space E , up to a fixed time T , that is, $(Z_s, 0 \leq s \leq T) \equiv (J_0, J_1, \dots, J_{N(T)}; X_1, \dots, X_{N(T)-1}, T - S_{N(T)})$, if $N(T) > 0$ and $(Z_s, 0 \leq s \leq T) \equiv (J_0)$, if $N(T) = 0$.

The empirical estimator $\hat{Q}_{ij}(t, T)$ of $Q_{ij}(t)$ is defined by

$$\hat{Q}_{ij}(t, T) = \frac{1}{N_i(T)} \sum_{k=1}^{N(T)} \mathbf{1}_{\{J_{k-1}=i, J_k=j, X_k \leq t\}}. \quad (12)$$

Then we have the following result from Moore and Pyke [18]; see also Ref. 23. We denote by $\xrightarrow{a.s.}$ and $\xrightarrow{d}$ the almost sure convergence and convergence in distribution, respectively. Denote by $N(a, b)$ the normal random variable with mean a and variance b .

Theorem 3. *For any fixed $i, j \in E$, as $T \rightarrow \infty$, we have*

1. (Strong consistency) $\max_{i,j} \sup_{t \in (0, T)} |\hat{Q}_{ij}(t, T) - Q_{ij}(t)| \xrightarrow{a.s.} 0$,
2. (Asymptotic normality) $T^{1/2}(\hat{Q}_{ij}(t, T) - Q_{ij}(t)) \xrightarrow{d} N(0, \sigma_{ij}^2(t))$,
where $\sigma_{ij}^2(t) := \mu_{ii} Q_{ij}(t)[1 - Q_{ij}(t)]$.

We define estimators of the Markov renewal function and of transition probabilities by a plug-in procedure, that is, replacing the semi-Markov kernel $Q_{ij}(t)$ in Equations (6) and (10) by the empirical estimator kernel $\hat{Q}_{ij}(t, T)$.

The following two results concerning the Markov renewal function estimator, $\hat{\psi}_{ij}(t, T)$, and the transition probability function estimator, $\hat{P}_{ij}(t, T)$, of Z were proved by Ouhbi and Limnios [23].

Theorem 4. [23]. *The estimator $\hat{\psi}_{ij}(t, T)$ of the Markov renewal function $\psi_{ij}(t)$ satisfies the following two properties:*

1. (Strong consistency) *it is uniformly strongly consistent, that is, as $T \rightarrow \infty$,*

$$\max_{i,j} \sup_{t \in (0, T)} |\hat{\psi}_{ij}(t, T) - \psi_{ij}(t)| \xrightarrow{a.s.} 0.$$

2. (Asymptotic normality) *For any fixed $t > 0$, it converges in distribution, as $T \rightarrow \infty$, to a normal random variable, that is,*

$$T^{1/2}(\hat{\psi}_{ij}(t, T) - \psi_{ij}(t)) \xrightarrow{d} N(0, \sigma_{ij}^2(t)),$$

where $\sigma_{ij}^2(t) = \sum_{r \in E} \sum_{k \in E} \mu_{rr} \{(\psi_{ir} * \psi_{kj})^2 * Q_{rk} - (\psi_{ir} * \psi_{kj} * Q_{rk})^2\}(t)$.

Theorem 5. [23]. *The estimator $\hat{P}_{ij}(t, T)$ of the transition function $P_{ij}(t)$, satisfies the following two properties:*

1. (Strong consistency) for any fixed $L > 0$, we have, as $T \rightarrow \infty$

$$\max_{i,j} \sup_{t \in [0,L]} \hat{P}_{ij}(t, T) - P_{ij}(t) \xrightarrow{a.s.} 0.$$

2. (Asymptotic normality) For any fixed $t > 0$, we have, as $T \rightarrow \infty$,

$$T^{1/2}(\hat{P}_{ij}(t, T) - P_{ij}(t)) \xrightarrow{d} N(0, \sigma_{ij}^2(t)),$$

where

$$\begin{aligned} \sigma_{ij}^2(t) = & \sum_{r \in E} \sum_{k \in E} \mu_{rr} \\ & \times [(1 - H_i) * B_{irkj} - \psi_{ij} \mathbf{1}_{\{r=j\}}]^2 \\ & * Q_{rk}(t) - [((1 - H_i) * B_{irkj} - \psi_{ij} \mathbf{1}_{\{r=j\}}) \\ & * Q_{rk}(t)]^2, \end{aligned}$$

and

$$B_{irkj}(t) = \sum_{n=1}^{\infty} \sum_{\ell=1}^n Q_{ir}^{(\ell-1)} * Q_{kj}^{(n-\ell)}(t).$$

RELIABILITY OF SEMI-MARKOV SYSTEMS

For a stochastic system with state space E described by a SMP (Z_t) , let us consider a partition U, D of E , that is, $E = U \cup D$, with $U \cap D = \emptyset$, $U \neq \emptyset$, and $D \neq \emptyset$. The set U contains the up states and D contains the down states of the system. The reliability is defined by $R(t) := \mathbb{P}(Z_s \in U, \forall s \in [0, t])$. If we define the conditional reliability by $R_i(t) := \mathbb{P}(Z_s \in U, \forall s \in [0, t] \mid Z_0 = i)$, then for any initial distribution law vector α , we have $R(t) := \sum_{i \in U} \alpha_i R_i(t)$. For the finite state space case, without loss of generality, let us enumerate first the up states and next the down states, that is, for $E = \{1, 2, \dots, d\}$, we have $U = \{1, \dots, r\}$ and $D = \{r+1, \dots, d\}$.

The conditional reliability $R_i(t)$ $i \in U$ fulfills the following MRE:

$$R_i(t) = \bar{H}_i(t) + \sum_{j \in U} \int_0^t Q_{ij}(ds) R_j(t-s), \quad (13)$$

which has the following solution

$$R_i(t) = \sum_{j \in U} (\psi_{ij} * \bar{H}_j)(t). \quad (14)$$

Integrating Equation (13) over $[0, \infty)$, we get

$$\begin{aligned} & (MTTF_1, \dots, MTTF_r)' \\ &= (I - P_{11})^{-1} (m_1, \dots, m_r)', \end{aligned} \quad (15)$$

where $MTTF_j$ is the mean time to failure starting from the state $j \in E$, and C' denotes the transpose of the vector or the matrix C .

The above reliability indicators and some additional ones are formulated in matrix form and are given in the following table.

Notation. In Table 1, we have $\bar{\mathbf{H}}_1(t) = (\bar{H}_i(t); i \in U)'$, $\bar{\mathbf{H}}_2(t) = (\bar{H}_i(t); i \in D)'$, $\bar{\mathbf{H}}_{10}(t) = (\bar{H}_{10}(t); i \in U; 0, \dots, 0)'$, $\mathbf{m}_1 = (m_i; i \in U)'$, $\mathbf{m}_2 = (m_i; i \in D)'$, $\mathbf{1}_r = (1, \dots, 1)'$ r -ones. All of the above vectors are column vectors. Furthermore, index 1 means restriction of the corresponding vector or matrix on U and index 2 means restriction of the corresponding vector or matrix on D , that is, $Q_{11}(t)$, means restriction of the matrix $Q(t)$ on $U \times U$ and α_1 restriction of the row vector α on U . By $A^{(-1)}(t)$ we mean the inverse of the matrix A in the convolution sense (see

Table 1. Closed Form Solution for Reliability and Related Measurements

Reliability	$R(t) = \alpha_1 (\mathbf{I} - \mathbf{Q}_{11})^{(-1)} * \bar{\mathbf{H}}_1(t)$
Availability	$A(t) = \alpha (\mathbf{I} - \mathbf{Q})^{(-1)} * \bar{\mathbf{H}}_{10}(t)$
Maintainability	$M(t) = 1 - \alpha_2 (\mathbf{I} - \mathbf{Q}_{22})^{(-1)} * \bar{\mathbf{H}}_2(t)$
Mean time to failure	$MTTF = \alpha_1 (\mathbf{I} - \mathbf{P}_{11})^{-1} \mathbf{m}_1$
Mean time to repair	$MTTR = \alpha_2 (\mathbf{I} - \mathbf{P}_{22})^{-1} \mathbf{m}_2$
Mean up time	$MUT = \frac{\nu_1 \mathbf{m}_1}{\nu_2 \mathbf{P}_{21} \mathbf{1}_r}$
Mean down time	$MDT = \frac{\nu_2 \mathbf{m}_2}{\nu_1 \mathbf{P}_{12} \mathbf{1}_{d-r}}$

Ref. 13). For example, if

$$A(t) = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}, \text{ then} \\ A^{(-1)}(t) = (a * d(t) - b * c(t))^{(-1)} \\ * \begin{pmatrix} d(t) & -b(t) \\ -c(t) & a(t) \end{pmatrix}.$$

Reliability estimation. From the estimator given by Equation (12), the following plug-in estimator of reliability is proposed:

$$\widehat{R}(t, T) = \widehat{\alpha}_1(\mathbf{I} - \widehat{\mathbf{Q}}_{11})^{(-1)} * \widehat{\mathbf{H}}_1(t, T).$$

The following properties are fulfilled by the above reliability estimator:

Theorem 6. [15]. *For any fixed $t > 0$ and for any $L \in (0, \infty)$, we have*

1. (Strong consistency)

$$\sup_{0 \leq t \leq L} |\widehat{R}(t, T) - R(t)| \xrightarrow{a.s.} 0, \\ \text{as } T \rightarrow \infty,$$

2. (Asymptotic normality)

$$T^{1/2}(\widehat{R}(t, T) - R(t)) \xrightarrow{d} N(0, \sigma_R^2(t)), \\ \text{as } T \rightarrow \infty,$$

where

$$\sigma_R^2(t) \\ = \sum_{i \in U} \sum_{j \in E} \mu_{ii} \left\{ \left(B_{ij} \mathbf{1}_{\{j \in U\}} - \sum_{r \in U} \alpha(r) \psi_{ri} \right)^2 \right. \\ * Q_{ij}(t) - \left[\left(B_{ij} \mathbf{1}_{\{j \in U\}} - \sum_{r \in U} \alpha(r) \psi_{ri} \right) \right. \\ \left. * Q_{ij}(t) \right]^2 \left. \right\}$$

and

$$B_{ij}(t) = \sum_{n \in E} \sum_{k \in U} \alpha(i) \psi_{ni} * \psi_{jk} * (I - H_k(t)).$$

For further results on estimation, see Refs 15, 23, 24, 26, 27.

SEMI-MARKOV CHAINS AND HIDDEN MODELS

In this section, we present semi-Markov chains (SMC), which are semi-Markov processes in discrete-time, and hidden semi-Markov models (HSMM).

Let us consider the MRP $(J_n, S_n)_{n \in \mathbb{N}}$ in discrete time $k \in \mathbb{N}$, with state space the countable set E . The semi-Markov kernel q is defined by

$$q_{ij}(k) := \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n = k \mid J_n = i), \\ \text{with } i, j \in E, \quad k, n \in \mathbb{N}. \quad (16)$$

Define also $q_{ij}(K) = \sum_{k \in K} q_{ij}(k)$, where $K \subset \mathbb{N}$. So, as in the continuous-time case, an SMC is determined by its semi-Markov kernel q and its initial distribution law α .

The process (J_n) is the EMC of the MRP (J_n, S_n) with transition kernel $P = (P(i, j))$. The semi-Markov kernel q is written as

$$q_{ij}(k) = P(i, j) f_{ij}(k),$$

where $f_{ij}(k) := \mathbb{P}(S_{n+1} - S_n = k \mid J_n = i, J_{n+1} = j)$, is the conditional distribution of the sojourn time in the state i given that the next visited state is j , ($j \neq i$).

Define the counting process of jumps $N(k) := \max\{n \geq 0 : S_n \leq k\}$. The SMC is defined as follows:

$$Z_k = J_{N(k)}, \quad k \in \mathbb{N}. \quad (17)$$

When $q_{ij}(k) = p_{ij}(p_{ii})^{k-1}$, $k = 1, 2, \dots$, for all $i, j \in E$, $i \neq j$, then the SMC Z is a Markov chain with transition matrix $p = (p_{ij})$. In this case, the transition matrix of the EMC is $P(i, j) = p_{ij}/(1 - p_{ii})$, if $p_{ii} < 1$ and $i \neq j$, and $P(i, j) = 0$, if $i = j$ or $p_{ii} = 1$, and the conditional transition function is $f_{ij}(k) = (1 - p_{ii})(p_{ii})^{k-1}$.

The Markov renewal function ψ is defined by

$$\psi_{ij}(k) := \sum_{n=0}^k q_{ij}^{(n)}(k). \quad (18)$$

The general MRE in the discrete case is as follows:

$$\Theta(k) = L(k) + q * \Theta(k),$$

where $\Theta(k), k \in \mathbb{N}$, is an unknown sequence of real matrices (of finite or of infinite dimension), $L(k), k \in \mathbb{N}$, is a known sequence of real matrices, and $q(k), k \in \mathbb{N}$, a semi-Markov kernel. The convolution in discrete time is $(q * \Theta)_{ij}(k) = \sum_{\ell \in E} \sum_{s=0}^k q_{i\ell}(s) \Theta_{\ell j}(k-s)$. The general solution of the above MRE is given by

$$\Theta(k) = \psi * L(k).$$

For example, the transition probabilities of the SMC $Z, P_{ij}(k)$, satisfy the following MRE in matrix form:

$$P(k) = I - H(k) + q * P(k),$$

where its unique solution is

$$P(k) = \psi * (I - H)(k).$$

The reliability indicators are given by the same formulae as in the continuous-time case, in Table 1, only replacing Q by q and the step function $I = I(t)$, by the Dirac matrix measure $I\delta$ at the origin (see Ref. 14). Estimation results for discrete-time SMCs can be found in Ref. 14.

An HSMM is defined by a bivariate process (Z_k, Y_k) , where Z_k is an E -valued SMC (unobserved) and Y_k is an A -valued process (observed) depending on Z_k . We suppose here that E and A are countable sets. This dependence structure in the HSMM is given by the following relation:

$$\begin{aligned} \mathbb{P}(Z_{k+1} = j, Y_{k+1} = l \mid Z_s, Y_s; s \in \{0, 1, \dots, k\}) \\ = \mathbb{P}(Z_{k+1} = j \mid Z_k, U_k) \\ \times \mathbb{P}(Y_{k+1} = l \mid Z_{k+1} = j), \end{aligned}$$

where $U_k = k - S_{N(k)}$. This is an SM-M0 type HSMM. A more general type is the SM-Mm HSMM, which satisfies the following relation:

$$\begin{aligned} \mathbb{P}(Z_{k+1} = j, Y_{k+1} = l \mid Z_s, Y_s; s \in \{0, 1, \dots, k\}) \\ = \mathbb{P}(Z_{k+1} = j \mid Z_k, U_k) \\ \times \mathbb{P}(Y_{k+1} = l \mid Z_{k+1} = j, Y_k, \dots, Y_{k-m+1}). \end{aligned}$$

The main goal here is to estimate the semi-Markov kernel and, of course,

the conditional distributions of Y , by observing only Y over $[0, M]$, that is, $(Y_0^M = y_0^M) \equiv (Y_0 = y_0, Y_1 = y_1, \dots, Y_M = y_M)$, where $(y_0, y_1, \dots, y_M) \in A^{M+1}$. The quantities to be estimated here are $\theta = (q, \Gamma)$, where q is the semi-Markov kernel, and Γ the conditional distribution of Y , that is, $\Gamma(i, l) = \mathbb{P}(Y_k = l \mid Z_k = i)$ for any $k \in \mathbb{N}, i \in E, l \in A$.

The likelihood function for the complete data set is

$$\begin{aligned} L_M(\theta; Z_0^M, Y_0^M) \\ = \alpha(Z_0) \prod_{n=1}^{N(M)} q(J_{n-1}, J_n, X_n) \bar{H}(J_{N(M)}, U_M) \\ \times \prod_{k=0}^M \Gamma(Z_k, Y_k). \end{aligned}$$

Here, for convenience, we used the notation $q(i, j, k)$ instead of $q_{ij}(k)$ for all functions.

The maximum likelihood estimators can be obtained by an algorithm for incomplete data such as expectation maximization (EM) or stochastic EM. In fact, instead of maximizing the likelihood of the observed data, that is, $L_M^0(\theta; Y_0^M) = \sum_{Z_0^M \in E^{M+1}} L_M(\theta; Z_0^M, Y_0^M)$, one maximizes the conditional expectation over θ ,

$$\mathbb{E}_{\theta^{(m)}} \left[\log L_M(\theta; Z_0^M, Y_0^M) \mid Y_0^M \right]$$

in order to obtain $\theta^{(m+1)}$, and so on. Under appropriate general conditions the sequence $\theta^{(m)}, m \geq 0$, converges to the true value of the parameter θ .

For a given system whose the temporary behavior is described by an SMC, say Z , and the values of some indicators, that is, noise, temperature, pressure, and so on, are observed, say Y , we would like to estimate the state of the system in a given present time or into the future by observing Y only. This is a typical hidden semi-Markov problem.

CONCLUDING REMARKS

Several topics concerning SMP have been omitted. For example, results on time-nonhomogeneous SMP can be found in

Refs 28–30; SM decision processes in Refs 16, 31; calculation and numerical aspects [13,29,32]; limit theorems and convergence of SMP in series scheme in Refs 8–10, 33–37; stochastic approximations in Refs 9, 10; (stochastic) additive functionals in Refs 9, 10, 13; diffusion type SMP in Ref. 38; entropy of SMP [39], and so on; and also some important applications, such as reliability and maintenance theory in Refs 13–15, 24, 40–43; queueing theory in Refs 44, 45; insurance and finance in Refs 10, 29, 46; words occurrences in Ref. 47; estimation in discrete time in Refs 14, 48; phase-type distributions in Refs 13, 43, 45, and so on. Several chapters in Refs 28 and 29 concern the above aspects. For a detailed presentation of SMCs and HSMMs see Ref. 14.

REFERENCES

1. Feller W. On semi-Markov processes. *Proc Natal Acad Sci U S A* 1964;51(2):653–659.
2. Pyke R. Markov renewal processes: definitions and preliminary properties. *Ann Math Stat* 1961;32:1231–1242.
3. Pyke R. Markov renewal processes with finitely many states. *Ann Math Stat* 1961;32:1243–1259.
4. Pyke R, Schaufele R. Limit theorems for Markov renewal processes. *Ann Math Stat* 1964;35:1746–1764.
5. Pyke R, Schaufele R. The existence and uniqueness of stationary measures for Markov renewal processes. *Ann Stat* 1967;37:1439–1462.
6. Çinlar E. Markov renewal theory. *Adv Appl Probab* 1969;1:123–187.
7. Çinlar E. *Introduction to stochastic processes*. Englewood Cliffs (NJ): Prentice-Hall; 1975.
8. Korolyuk VS, Turbin AF. *Decomposition of large scale systems*. Dordrecht: Kluwer; 1993.
9. Korolyuk VS, Swishchuk A. *Random evolution for semi-Markov systems*. Dordrecht: Kluwer; 1995.
10. Korolyuk VS, Limnios N. *Stochastic systems in merging phase space*. Singapore: World Scientific; 2005.
11. Cox DR. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proc Camb Philos Soc* 1955;51:433–441.
12. Gikhman II, Skorokhod AV. Volume 2, *Theory of stochastic processes*. Berlin: Springer; 1974.
13. Limnios N, Oprisan G. *Semi-Markov processes and reliability*. Boston (MA): Birkhäuser; 2001.
14. Barbu V, Limnios N. Volume 191, *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis, Lecture Notes in Statistics*. New York: Springer; 2008.
15. Ouhbi B, Limnios N. Nonparametric reliability estimation of semi-Markov processes. *J Stat Plan Inference* 2003;109(1/2):155–165.
16. Ross SM. *Applied probability models with optimization applications*. San Francisco (CA): Holden-Day; 1970.
17. Shurenkov VM. On the theory of Markov renewal. *Theory Probab Appl* 1984;29:247–265.
18. Moore EH, Pyke R. Estimation of the transition distributions of a Markov renewal process. *Ann Inst Stat Math* 1968;20:411–424.
19. Lagakos SW, Sommer CJ, Zelen M. Semi-Markov models for partially censored data. *Biometrika* 1978;65(2):311–317.
20. Akritas MG, Roussas GG. Asymptotic inference in continuous time semi-Markov processes. *Scand J Stat* 1980;7:73–79.
21. Gill RD. Nonparametric estimation based on censored observations of a Markov renewal process. *Z Wahrsch verw Gebiete* 1980;53:97–116.
22. Greenwood PE, Wefelmeyer W. Empirical estimators for semi-Markov processes. *Math Methods Stat* 1996;5(3):299–315.
23. Ouhbi B, Limnios N. Nonparametric estimation for semi-Markov processes based on its hazard rate functions. *Stat Inference Stoch Proc* 1999;2(2):151–173.
24. Ouhbi B, Limnios N. The rate of occurrence of failures for semi-Markov processes and estimation. *Stat Probab Lett* 2002;59(3):245–255.
25. Aalen O, Borgan O, Gjessing H. *Survival and event history analysis*. New York: Springer; 2008.
26. Limnios N, Ouhbi B. Empirical estimators of reliability and related functions for semi-Markov systems. In Lindqvist B, Doksum K, editors. *Mathematical and statistical methods in reliability*. Singapore: World Scientific; 2003.
27. Limnios N, Ouhbi B. Nonparametric estimation of some important indicators in

- reliability for semi-Markov processes. *Stat Method* 2006;3(4):341–350.
28. Janssen J, editor. *Semi-Markov models: theory and applications*. New York: Plenum Press; 1986.
 29. Janssen J, Limnios N, editors. *Semi-Markov models and applications*. Dordrecht: Kluwer; 1999.
 30. Janssen J, Manca R. *Applied semi-Markov processes*. New York: Springer; 2006.
 31. Puterman ML. *Markov decision processes: discrete stochastic programming*. New York: Wiley; 1994.
 32. Chiquet J, Limnios N. A method to compute the reliability of a piecewise deterministic Markov process. *Stat Probab Lett* 2008;78:1397–1403.
 33. Athreya KB, McDonald D, Ney P. Limit theorems for semi-Markov processes and renewal theory for Markov chains. *Ann Probab* 1978;6(5):788–797.
 34. Grigorescu S, Oprişan G. Limit theorems for J-X processes with a general state space. *Z Wahrsch verw Gebiete* 1976;35:65–73.
 35. Gyllenberg M, Silvestrov DS. *Quasi-stationary phenomena in nonlinearly perturbed stochastic systems*. Berlin: de Gruyter; 2008.
 36. Malinovskii VK. Limit theorems for recurrent semi-Markov processes and Markov renewal processes. *J Sov Math* 1987;36:493–502.
 37. Nummelin E. Uniform and ration limit theorems for Markov renewal and semi-regenerative processes on a general state space. *Ann Inst Henri Poincaré* 1978;XIV(2):119–143.
 38. Harlamov BP. *Continuous semi-Markov processes*, *Applied Stochastic Methods Series*. London: ISTE-J. Wiley; 2008.
 39. Girardin V, Limnios N. Entropy for semi-Markov processes with Borel state spaces: asymptotic equirepartition properties and invariance principles. *Bernoulli* 2006;12(3):515–533.
 40. Csenki A. Volume 90, *Dependability for systems with a partitioned state space*, *Lecture Notes in Statistics*. Berlin: Springer-Verlag; 1995.
 41. Huzurbazar AV. *Flowgraph models for multistate time-to-event data*. New York: John Wiley & Sons, Inc.; 2004.
 42. Osaki S. *Stochastic system reliability modeling*. Singapore: World Scientific; 1985.
 43. Pérez-Ocón R, Torres-Castro I. A reliability semi-Markov model involving geometric processes. *Appl Stoch Models Bus Ind* 2002;18:157–170.
 44. Anisimov VV. *Switching processes in queuing models*, *Applied Stochastic Methods Series*. London: ISTE-J. Wiley; 2008.
 45. Asmussen S. *Applied probability and queues*. New York: Wiley; 1987.
 46. Reinhard JM, Snoussi M. The severity of ruin in a discrete semi-Markov risk model. *Stoch Models* 2002;18(1):85–107.
 47. Chryssaphinou O, Karaliopoulou M, Limnios N. On discrete time semi-Markov chains and applications in words occurrences. *Commun Stat Theory Methods* 2008;37:1306–1322.
 48. Trevezas S, Limnios N. Maximum likelihood estimation for general hidden semi-Markov processes with backward recurrence time dependence. *Notes of Scientific Seminars of the St. Petersburg Department of the Steklov Mathematical Institute, Russian Academy of Sciences*, 2009; 363:105–125. and *J. Mathematical Sciences*, Special issue in honor of I. Ibragimov.

SEMI-MARKOV PROCESSES

YASEMIN SERIN
Industrial Engineering
Department, Middle East
Technical University,
Ankara, Turkey

INTRODUCTION

A semi-Markov process (SMP) is continuous-time countable state space process that is a generalization of continuous-time Markov chains in the way that the sojourn times come from general distributions. The Markovian property holds only at the transition points. Relaxing the exponential sojourn time assumption broadens the field of application of the results. For example, M/G/1 and G/M/1 queues can be modeled as SMPs and transient and limiting behavior can be analyzed using the results summarized below. This article benefitted from Refs 1–3 and is largely based on Ref. 4. The proofs of all the following theorems can be found in Ref. 4.

DEFINITIONS AND TRANSIENT BEHAVIOR

In a continuous-time Markov chain, states of a discrete-time Markov chain (DTMC) are visited with exponentially distributed sojourn times. An SMP is a generalization where sojourn times have a general distribution. So, an SMP enters a state, stays in that state for a random amount of time, and jumps into some state according to a Markov transition probability; therefore, the Markovian property holds only at the transition points. A formal definition is given next.

Let $\{X_n, n = 0, 1, 2, \dots\}$ be a discrete-time stochastic process with a countable state space S . Let Y_n be a sequence of random variables taking values on the nonnegative real line, $S_n = \sum_{k=1}^n Y_k$ and $N(t) = \max\{n : S_n \leq t\}$.

Definition 1. The continuous-time process $\{X(t), t \geq 0\}$ defined by $X(t) = X_{N(t)}$ is called a *semi-Markov process* (SMP) if the sequence $\{(X_n, Y_{n+1}), n = 0, 1, \dots\}$ satisfies

$$\begin{aligned} P(X_{n+1} = j, Y_{n+1} < y | X_n = i, Y_n, X_{n-1}, \\ Y_{n-1}, \dots, X_1, Y_1, X_0) \\ = P(X_{n+1} = j, Y_{n+1} < y | X_n = i) \\ = G_{ij}(y) \end{aligned}$$

for all $i, j \in S, n = 0, 1, \dots$ and $y \geq 0$.

The SMP stays in state X_n for Y_{n+1} units of time and jumps to X_{n+1} at time S_{n+1} . Here, $N(t)$ represents the number of transitions up to time t . Independence of G from n implies time homogeneity of the process. It must be noted that with this definition, unlike continuous-time Markov chains, a transition from a state to itself is allowed for SMPs. The matrix defined by

$$G = [G_{ij}(y)]_{i,j \in S}$$

is called the *kernel of the SMP*. Together with an initial distribution, the kernel completely describes the SMP. The cdf of the sojourn time in state i is

$$\begin{aligned} F_i(y) &= P(Y_1 < y | X_0 = i) \\ &= \sum_{j \in S} G_{ij}(y). \end{aligned}$$

The transition probability of the DTMC $\{X_n, n = 0, 1, 2, \dots\}$, that is *the embedded DTMC* is $p_{ij} = G_{ij}(\infty)$. Here, it will be assumed that finite number of transitions can occur in a finite time; that is $F_i(\infty) = 1$ for all $i \in S$. This regularity definition holds for finite state SMPs.

Example. A Series System. Consider an equipment with N components. The equipment works as long as all the components work. Component i works for an

exponential(λ_i) units of time and its repair takes R_i units of time with cdf $F_i(r)$ and mean r_i . There is a single repairman and components do not fail when one is being repaired. Failure of a component or a repair completion define a transition time. Then the process defined by

$$X_n = \begin{cases} i, & \text{if component } i \text{ fails at the } n\text{th transition} \\ 0, & \text{if a repair is over at the } n\text{th transition} \end{cases}$$

is a DTMC and the process $\{X(t), t \geq 0\}$ defined by

$$X(t) = \begin{cases} i, & \text{if component } i \text{ is being repaired at time } t \\ 0, & \text{if all components are working at time } t \end{cases}$$

with $S = \{0, 1, 2, \dots, N\}$ is an SMP with kernel G where for $i = 1, 2, \dots, N$, and $y \geq 0$,

$$G_{0i}(y) = \frac{\lambda_i}{\sum_{j=1}^N \lambda_j} \left(1 - e^{-y \sum_{j=1}^N \lambda_j} \right),$$

$$G_{i0}(y) = F_i(y),$$

$$G_{ij}(y) = 0, \text{ otherwise.}$$

The transition probabilities of the embedded DTMC are, for $i = 1, 2, \dots, N$,

$$p_{0i} = \frac{\lambda_i}{\sum_{j=1}^N \lambda_j},$$

$$p_{i0} = 1,$$

$$p_{ij} = 0, \text{ otherwise.}$$

The state transition probabilities $p_{ij}(t) = P(X(t) = j | X(0) = i)$ of the SMP for $i, j \in S$ and finite t represent the transient behavior of the process. They can be obtained as in the following theorem.

Theorem 1. *The transition probabilities of an SMP satisfies the following Markov renewal type equation*

$$p_{ij}(t) = \int_0^t \sum_{k \in S} p_{kj}(t-u) dG_{ik}(u) + \int_t^\infty \sum_{k \in S} \delta_{ij} dG_{ik}(u)$$

and the solution is

$$p_{ij}(t) = \delta_{ij}(1 - F_i(t)) + \int_0^t (1 - F_i(t-u)) dM_{ij}(u),$$

where $M_{ij}(t) = \sum_{n=1}^\infty G_{ij}^{(n)}(t)$ and $A^{(n)}$ is n -fold convolution of A and $\delta_{ij} = 1$ if $i = j$ and 0, otherwise.

STATE CLASSIFICATION AND LIMITING PROBABILITIES

The behavior of an SMP, in the long run (as t goes to ∞), can be analyzed in two parts: first, the probability of finding the process in a particular state in the long run should be investigated. Next, since the sojourn times are not memoryless, the cdf of the time until the next transition in the long run should also be investigated.

For the state classification, the states that behave in the same way are identified and the limiting probabilities are given. The *first passage time from state i to state j* is

$$T_{ij} = \min\{t \geq Y_1 : X(t-) \neq j, X(t) = j | X(0) = i\}.$$

Here, $t-$ represents the time point just before t . Let the expected time spent in i before a transition be $\mu_i = E[Y_1 | X(0) = i]$ and the expected first passage time from i to j be $\mu_{ij} = E[T_{ij}]$ for $i, j \in S$. Generally, μ_i is determined by the given sojourn time distributions. The expected first passage times can be computed as in the following theorem.

Theorem 2. *The expected first passage times satisfy the following system of linear equations:*

$$\mu_{ij} = \mu_i + \sum_{k \neq j, k \in S} p_{ik} \mu_{kj} \text{ for } i, j \in S.$$

Example. A Series System (continued). From Theorem 2, the expected first

passage times for the series system example are, for $i = 1, 2, \dots, N$,

$$\begin{aligned}\mu_{i0} &= r_i, \\ \mu_{00} &= \frac{1 + \sum_{k=1}^N r_k \lambda_k}{\sum_{k=1}^N \lambda_k}, \\ \mu_{0i} &= \frac{1 + \sum_{k=1, k \neq i}^N r_k \lambda_k}{\lambda_i}, \\ \mu_{ij} &= r_i + \frac{1 + \sum_{k=1, k \neq j}^N r_k \lambda_k}{\lambda_j}, \\ &\text{for } j = 1, 2, \dots, N.\end{aligned}$$

The usual definitions of recurrence/transience, aperiodicity, and irreducibility apply to SMPs.

Definition. A state of the SMP is recurrent (transient) if it is recurrent (transient) in the embedded DTMC.

Definition. A recurrent state i of the SMP is positive recurrent if $\mu_{ii} < \infty$ and it is null recurrent if $\mu_{ii} = \infty$.

Definition. An SMP is irreducible if the embedded DTMC is irreducible.

Definition. A state i of the SMP is aperiodic if the T_{ii} is aperiodic.

The states in an irreducible subset of S have the same classifications; these properties are class properties. Therefore, it is sufficient to study the limiting behavior of the irreducible SMPs. Let $P = [p_{ij}]$. The following theorem relates the stationary distribution of the embedded DTMC with the expected first passage times and the limiting distribution of the SMP.

Theorem 3. Let $\{X(t), t \geq 0\}$ be an irreducible SMP and let π be a positive solution to $\pi = \pi P$. If the SMP is recurrent then

$$\mu_{jj} = \frac{\sum_{i \in S} \pi_i \mu_i}{\pi_j}.$$

If the SMP is positive recurrent then

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \frac{\pi_j \mu_j}{\sum_{k \in S} \pi_k \mu_k}.$$

Example. A Series System (continued). The limiting probabilities of the embedded DTMC are

$$\begin{aligned}\pi_0 &= \frac{1}{2}, \\ \pi_i &= \frac{\lambda_i}{2 \sum_{k=1}^N \lambda_k} \quad \text{for } j = 1, 2, \dots, N.\end{aligned}$$

The probability that the series system is working in the long run is

$$\lim_{t \rightarrow \infty} p_{i0}(t) = \frac{1}{1 + \sum_{k=1}^N r_k \lambda_k}$$

and the probability that the component i is under repair in the long run is

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \frac{r_j \lambda_j}{1 + \sum_{i=1}^N r_k \lambda_k}$$

Finally, the limiting distribution of the time until the next transition is given. Keep in mind that $S_{N(t)+1}$ is the time of the first transition after t and $S_{N(t)+1} - t$ is the remaining time until the next transition.

Theorem 4. Let $\{X(t), t \geq 0\}$ be an irreducible positive recurrent and aperiodic SMP. Then

- (i) $\lim_{t \rightarrow \infty} P(X(t) = i, S_{N(t)+1} > t + x, X_{N(t)+1} = j | X(0) = k) = \frac{p_i}{\mu_i}, \int_x^\infty (p_{ij} - G_{ij}(y)) dy$
- (ii) $\lim_{t \rightarrow \infty} P(X(t) = i, S_{N(t)+1} > t + x | X(0) = k) = \frac{p_i}{\mu_i} \int_x^\infty (1 - F_i(y)) dy,$
- (iii) $\lim_{t \rightarrow \infty} P(S_{N(t)+1} > t + x | X(t) = i) = \frac{1}{\mu_i} \int_x^\infty (1 - F_i(y)) dy.$

Example. M/G/1 Queue. The arrivals to a queue are according to a Poisson process with rate λ and the iid service times have a general distribution with cdf $B(s)$ with mean $E(S)$ and second moment $E(S^2)$. Let X_n be the number of customers left in the system just after the n th departure, Y_n be the n th service time and $N(t)$ be the number of departures

from the queue in t time units. The process defined by $X(t) = X_{N(t)}$ is an SMP with

$$G_{ij}(y) = \begin{cases} \int_0^x \frac{e^{-\lambda x} (\lambda x)^{j-i+1}}{(j-i+1)!} dB(x), & \text{if } i > 0, \\ \int_0^x \frac{e^{-\lambda x} (\lambda x)^j}{j!} dB(x), & \text{if } i = 0, \\ 0, & \text{otherwise.} \end{cases} \quad j \geq i-1,$$

From Theorem 4, in the long run, the distribution of the remaining service time of the customer in service is

$$\begin{aligned} P(S_{N(t)+1} > t + x | X(t) = j) \\ = \frac{1}{E(S)} \int_x^\infty (1 - B(y)) dy \text{ if } j > 0 \end{aligned}$$

Hence if $j > 0$, the expected remaining service time is

$$\begin{aligned} E(S_{N(t)+1} - t | X(t) = j) \\ = \int_0^\infty \frac{1}{E(S)} \int_x^\infty (1 - B(y)) dy dx \\ = \frac{E(S^2)}{E(S)}. \end{aligned}$$

REFERENCES

1. Çinlar E. Introduction to stochastic processes. Englewood Cliffs, NJ: Prentice-Hall; 1975.
2. Heyman DP, Sobel MJ. Volume 1, Stochastic models in operations research. McGraw-Hill, NY; 1982.
3. Ross SM. Stochastic processes. New York: John Wiley and Sons; 1996.
4. Kulkarni VG. Modelling and analysis of stochastic systems. New York: Chapman Hall; 1995.

SENSITIVITY ANALYSIS AND DYNAMIC PROGRAMMING

CHIN HON TAN
JOSEPH C. HARTMAN
Industrial and Systems
Engineering, University of
Florida, Gainesville, Florida

Dynamic programming (DP) has been successfully applied to a wide range of problems in a variety of fields [1–4] since it was proposed by Richard Bellman in the middle of the last century [5]. One of the main advantages of DP lies in its ability to account for the dynamics of a system and it is often used by operations researchers to solve problems involving sequential decisions. In addition, unlike most optimization techniques, DP is able to handle a wide range of cost and/or reward functions. Given its versatility, it is not surprising that DP is used across various disciplines, including artificial intelligence, control, economics, and operations research. However, DP does have its drawbacks as well. The size of a dynamic program grows rapidly with the dimension of the problem. This is commonly referred to as *the curse of dimensionality* and has been the main challenge in applying DP approaches to solving real-world problems. In light of the computational difficulties that arise from the curse of dimensionality, researchers have proposed a variety of approximation approaches, including state aggregation and value function approximation, to solve large dynamic programs. The interested reader is referred to Powell [6] for a detailed presentation of approximate DP.

The validity of any DP solution depends on the accuracy of the model and the data for a given instance. However, the parameters of the model are often uncertain and estimated in practice. Hence, the robustness or stability of the solution with respect to changes in the model parameters is of interest. One approach is to solve the problem

for different parameters and try to infer the relationship between the model parameters and the optimal solution from the resulting solutions [7,8]. However, this approach has two drawbacks. First, when the size of the problem or the number of scenarios to analyze is large, this “brute force” approach can be very time-consuming and may not be feasible in practice. Second, unless the properties regarding the relationship between the model parameters and the optimal solutions are known (e.g., continuous or monotonic), the solution for a set of parameter values may not provide insights on the solution for another set of parameter values. For example, suppose that ρ is some model parameter. One cannot guarantee that the optimal objective value for $\rho = 2$ must be between the optimal objective values for $\rho = 1$ and $\rho = 3$. From a theoretical and practical perspective, more sophisticated approaches for conducting sensitivity analysis in DP models are desired.

In this article, we review the approaches that researchers have proposed for conducting sensitivity analysis in discrete dynamic programs. Broadly speaking, sensitivity analysis is the study of how changes in a model affects the resulting output. In our context, we are interested in the sensitivity of an optimal solution with respect to changes in the model parameters. In addition, we highlight that DP is related to other optimization methodologies. Relevant work in linear programming (LP) and network optimization are discussed.

In sensitivity analysis, we study the stability or robustness of an optimal solution. However, researchers have also proposed finding solutions that are robust to the variations in the model parameters in dynamic programs. Robust DP will not be discussed in this article. Recent papers on robust solution approaches include the work of Givan *et al.* [9], Iyengar [10], Nilim and El Ghaoui [11], and Itoh and Nakamura [12].

Researchers have also pointed out that the model is usually an approximation of the actual problem and hence, the structure of

the model itself is doubtful. Literature that address model uncertainties include the work of Briggs *et al.* [13], Chatfield [14], Draper [15], and Walker and Fox-Rushby [16]. In this article, we limit ourselves to uncertainties in the model parameters.

This article proceeds as follows: In the next section, we present various concepts of sensitivity analysis in the optimization literature. Then, we survey sensitivity analysis approaches and results in discrete dynamic programs. In the section titled “Linear Programming,” we highlight the relationship between DP and LP. In addition, we demonstrate how results from LP can be applied to discrete dynamic programs. We conclude with a summary and highlight directions for future research.

CONCEPTS OF SENSITIVITY ANALYSIS IN OPTIMIZATION

Let S and Π denote the set of states and the set of policies, respectively. Let $A(s)$ denote the set of actions available with $s \in S$. We consider problems of the following form:

$$V^\pi = \sum_{t=0}^T \gamma^t r^\pi(s_t^\pi), \quad (1)$$

where V^π is the value function of the system under policy $\pi \in \Pi$ through the horizon T . T can be either finite or infinite. $s_t^\pi \in S$ is the state of the system at time t under π and $\pi(s) \in A(s)$ is the action that is taken at state s under π . Note that the action that is selected depends solely on the state and is independent of t . This stationary policy assumption is not restrictive since additional states can always be declared to address the dependency of π on t . r^a is the reward associated with action a and γ is the per-period discount factor. If the state transitions are stochastic, s_t^π is random and the value function is expressed as

$$V^\pi = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r^\pi(s_t^\pi) \right]. \quad (2)$$

Let π^* denote an optimal policy and V denote the optimal value function. The objective is

to find a policy that maximizes the value function,

$$V = V^{\pi^*} = \max_{\pi \in \Pi} V^\pi. \quad (3)$$

We are interested in how V and π^* change when model parameters vary. We first consider the case where the variations in the model parameters can be expressed by a single parameter ρ (i.e., univariate). For example, suppose that γ is allowed to vary between 0.85 and 0.95. We model this by setting $\gamma = 0.9 + \rho$, where $\rho \in [-0.05, 0.05]$. This modeling approach allows for dependencies between the model parameters and is particularly useful when modeling simple variations in transition probabilities, where the sum of probabilities must add to one.

Suppose that the problem is solved for $\rho = \rho^{(0)}$ and the optimal value function and optimal policy, $V^{(0)}$ and $\pi^{(0)}$, are obtained. Sensitivity analysis aims to answer one or more of the following questions:

1. What is $\frac{dV^{\pi^{(0)}}}{d\rho} \big|_{\rho=\rho^{(0)}}$?
2. What is the range of ρ values for which $\pi^{(0)}$ is optimal?
3. What is V if $\rho = \rho^{(0)} + \delta$, for some $\delta \in \mathbb{R}$?
4. What is π^* if $\rho = \rho^{(0)} + \delta$, for some $\delta \in \mathbb{R}$?

Questions 1 and 2 are specific to $\pi^{(0)}$ and can only be answered after $\pi^{(0)}$ has been identified. Hence, these questions are addressed in a *posterior analysis* [17] or *postoptimal analysis* [18,19]. The answers to questions 3 and 4 do not require advanced knowledge of $V^{(0)}$ and $\pi^{(0)}$. However, $V^{(0)}$ and $\pi^{(0)}$ can often be used to obtain or approximate V and π^* (see the section titled “Dynamic Programming”) and knowledge of $V^{(0)}$ and $\pi^{(0)}$ may provide insights to the answers of questions 3 and 4.

Besides posterior analysis, researchers have also considered *prior analysis* where knowledge of $V^{(0)}$ and $\pi^{(0)}$ is not required. For example, inverse parametric optimization aims to find the ρ value where a policy is optimal or good [20–22]). Prior analysis will

not be discussed in this article but the interested reader is referred to Fernandez-Baca and Venkatachalam [17] for a discussion on this topic.

Alternatively, the analysis can also be classified as being either *local* or *global* [23]. Question 1 takes a local perspective and is concerned with changes in the ε -neighborhood of $\rho^{(0)}$. Wagner [24] considers question 2 to be local as well since it is concerned with the properties of $\pi^{(0)}$. In contrast, questions 3 and 4 explore the variations across a wider range of ρ values and possibly different π^* .

In the univariate problem, the stability of $V^{(0)}$ and $\pi^{(0)}$ are addressed by questions 1 and or through 4. However, the notion of stability is more complicated in the multivariate problem (i.e., multiple parameters). The typical approach to sensitivity analysis for a multivariate problem is to vary one parameter at a time while holding the other parameters constant. This is sometimes referred to as *ordinary sensitivity analysis* [25,26]. However, ordinary sensitivity analysis does not allow for simultaneous variations of different parameters. Bradley *et al.* [27] provided restrictions on the variations in the right-hand side (RHS) terms and objective-function coefficients to guarantee the optimality of the current optimal basis (i.e., optimal decision variables) in a linear program. They term these restrictions as the *100% rule*. Wendell [28] was interested in the range that each parameter was allowed to vary without violating optimality. His approach, which he termed the *tolerance approach*, is different from ordinary sensitivity analysis in that simultaneous variations of different parameters are considered. The tolerance approach has been applied to different problems by various researchers [26,28–31].

It should be clear that the topic of sensitivity analysis is wide and that this section should not be regarded as a comprehensive survey of the approaches that researchers have proposed for sensitivity analysis. Rather, this is a brief discussion of the issues in sensitivity analysis and a presentation of some of the common notions of stability that are applicable to dynamic programs.

DYNAMIC PROGRAMMING

A dynamic program comprises a set of states S , with a set of decisions $A(s)$, available at each $s \in S$. Each $a \in A(s)$ transits the system from a state s to another state s' and results in a reward of r^a . In the deterministic problem (Eq. 1), the state transitions are known with certainty. In the stochastic problem, the state transitions are uncertain (Eq. 2) and the probability of transitioning from state s to state s' under $a \in A(s)$ is denoted by $P^a(s')$. The state transitions are Markovian (i.e., only dependent on the state s' and action a) and stochastic dynamic programs are often referred to as Markov decision processes (MDPs). In this article, we focus on the stochastic problem. However, we note that the discussion and results also apply to deterministic DP since the deterministic problem is merely a special case of the stochastic problem.

Most textbooks divide MDPs into two broad categories: (i) finite horizon and (ii) infinite horizon [6,32,33]. Finite-horizon problems are usually modeled as acyclic graphs where each node corresponds to a particular s at a particular t . The common solution approach for finite-horizon problems is to define the boundary conditions $V_T(s)$ for each state at time T and recursively solve for V in a backward manner:

$$V_{t-1}(s) = \max_{a \in A(s)} \left\{ r^a + \gamma \sum_{s' \in S} P^a(s') V_t(s') \right\} \\ t = 0, 1, \dots, T-1, \forall s \in S, \quad (4)$$

where $V_t(s)$ is the value function of state s at time t . Boundary conditions do not exist for infinite-horizon problems. Infinite-horizon problems are expressed as cyclic graphs and value and/or policy iteration approaches can be used to obtain the optimal solution. These approaches are based on the optimality conditions that are commonly referred to as the *Bellman equations* [33]:

$$V(s) = \max_{a \in A(s)} \left\{ r^a + \gamma \sum_{s' \in S} P^a(s') V(s') \right\} \forall s \in S. \quad (5)$$

It is clear from Equations (1) to (3) that V is continuous in γ . Altman and Schwartz [34] showed that V remains continuous in γ in a constrained MDP. In addition, it also follows from Equations (1) to (3) that V is continuous and monotonic with respect to each individual reward. However, the rewards may not be independent from each other. For example, the cost of producing 10 items may be twice the cost of producing 5 items. In these cases, we are interested in the changes in the optimal solution with respect to the unit cost of production, rather than the cost of producing 10 items specifically. In these problems, the rewards are expressed as

$$r^a = c^a(\rho), \quad (6)$$

where ρ is a vector that represents the parameters that are of interest and $c^a : P \rightarrow \mathbb{R}$, where P is the set containing ρ . White and El-Deib [35] showed that if r^a and $V_T(s)$ are affine in ρ for all $a \in A$ and $s \in S$, the optimal value function of each state, including V , is piecewise affine and concave in ρ .

Modeling the uncertainties in the transition probabilities is complicated by the additional requirement that transition probabilities must add to one. Researchers have modeled the uncertainties in the transition probabilities in a variety of ways [9,10,36–38]. Muller [39] showed that the optimal value function is monotone and continuous when appropriate stochastic orders and probability metrics are used.

Clearly, V is dependent on T . In particular, if all the rewards are nonnegative (nonpositive), V will be monotonically nondecreasing (nonincreasing) in T . However, the former is not a necessary condition and it is possible to prove the monotonicity of V for certain problems, even if the rewards differ in sign. For example, Tan and Hartman [40] showed that the cost of owning and operating equipment is nondecreasing in T if the discount factor and maintenance cost are nonnegative and the salvage value is nonincreasing with age.

Although the value of V is sensitive to changes in the model parameters, their impact on π^* is not as apparent. In fact,

the optimal policy is often robust to minor deviations in the model parameters. White and El-Deib [35] considered an MDP with imprecise rewards and were interested in finding the set of policies that are optimal for some realization of the imprecise parameters. They provided a successive approximation procedure for an acyclic finite-horizon problem and a policy iteration procedure for the infinite-horizon problem. Harmanec [41] computed the set of nondominated policies for a finite-horizon problem with imprecise transition probabilities by generalizing the classical backward recursion approach (Eq. 4).

Researchers have highlighted that the current decision is the only decision that the decision maker has to commit to in many cases and hence, he or she might be solely interested in the optimality of the current decision (i.e., decision at time zero). Hopp [42] was interested in the tolerance on the value functions of the later stages. In particular, he derived lower bounds on the maximum allowable perturbations in the state value of the future stages such that the time zero optimal decision remain unchanged. His numerical results indicate that the tolerance of the value functions (i.e., allowable perturbation) increases geometrically with time. In addition, researchers have also proposed finding a forecast horizon such that the optimal time zero decision is not affected by data beyond that horizon. Researchers who have studied this problem include Bean and Smith [43], Hopp [44], Bean *et al.* [45], and Cheevaprawatdomrong *et al.* [46].

In most cases, it is not necessary to resolve the problem when the model parameters change. Very often, the new π^* and V can be efficiently computed or approximated from the current solution. For example, the optimal solutions to shortest path problems can be easily revised when the arc lengths change [47]. Topaloglu and Powell [48] provided an efficient approach to approximate the change in the optimal value function for a dynamic fleet management model. Their approach was based on the special structure that resulted from their value function approximation. For MDP problems with no special structure, the Bellman equation (5) can be used to verify

the optimality of the current solution. If the optimality conditions are violated, the policy iteration approach can be used to determine the new solution.

LINEAR PROGRAMMING

It has been long recognized that a dynamic program can be formulated and solved as a linear program. The LP method for DP was first proposed by Manne [49] and elaborated upon by Puterman [33] and Derman [50]. It can be shown that a dynamic program can always be formulated by the following linear program [6]:

$$\begin{aligned} & \min \sum_{s \in S} v_s, \\ \text{s.t. } & v_s - \gamma \sum_{s' \in S} P^a(s') v_{s'} \geq r^a, \quad \forall s, a \\ & v_s \in \mathbb{R}, \quad s = 1, 2, \dots, |S|. \end{aligned}$$

The optimal value function V corresponds to $v_{s_0}^*$, where s_0 is the current state (i.e., state at time 0) and π^* is defined by the constraints that are binding at optimality. Manne's observation has an important theoretical implication. It highlights that an optimal solution of a dynamic program will have the same properties as an optimal solution to its associated linear program. This is particularly insightful as sensitivity analysis is well established in LP.

For every linear program, which we will refer to as the *primal Problem*, there is another associated linear program, which is commonly referred to as the *dual problem*. The relationship between the two problems provides insight on the stability of the solution in the primal problem. For each constraint in the primal problem, there is an associated variable in the dual problem. The latter is commonly referred to as the *dual variable*. The complementary slackness theorem, which was first proposed by Goldman and Tucker [51] and extended by Williams [52], highlights the relationship between the variables and constraints of the primal and dual problems at optimality. In particular, if a variable in one problem is positive, the corresponding constraint in

the other problem must be binding. If the constraint is not binding, the corresponding variable in the other problem must be zero. The values of the optimal dual variables, also known as *shadow prices*, correspond to the marginal change in the objective value when the RHS of the associated constraint is perturbed. This result can be applied directly to DP problems where the rewards of individual actions are independent. If a small perturbation to a reward changes some state value function v_s , it follows from the complementary slackness theorem that the associated action must be optimal. If the action is not optimal, its associated reward can be perturbed by a small amount without changing the value of any v_s .

The complementary slackness theorem and economic interpretation of the shadow prices that are presented are only valid when the rewards are independent of each other. When the rewards are dependent, the individual rewards are expressed as functions of independent parameters (Eq. 5) and parametric analysis approaches can be used to analyze the sensitivity of the solution with respect to these independent parameters [53–55]. However, the applicability of LP-based sensitivity analysis approaches to dynamic programs is limited. This is because the number of constraints in the LP formulation is large [6] and hence, dynamic programs are rarely formulated and solved as linear programs in practice.

CONCLUSIONS AND FUTURE RESEARCH

DP is a versatile optimization methodology that can be used to solve a variety of problems. Conventional solution approaches assume that the model parameters are known. However, the parameters are often uncertain in practice and hence the stability of the solution is of interest to the decision maker. The typical approach is to solve the problem for different realizations of parameter values. However, we have highlighted that there are both, theoretical and practical reasons for developing more efficient sensitivity analysis approaches for dynamic programs. Sensitivity analysis is a

well-established topic in LP and can be found in most LP textbooks [55,56]. However, this topic is rarely mentioned, much less discussed, in most DP textbooks [6,32,33,57]. So there is clearly room for development. The following are a few promising areas of future research.

In this article, we have highlighted the relationship between DP and LP and demonstrated how approaches and results from the LP literature can be applied to discrete dynamic programs. However, dynamic programs are rarely solved as linear programs in practice. There is a need to consider sensitivity analysis from a DP perspective and look at approaches that exploit the DP structure of the problem (i.e., Bellman equations).

In the section titled “Dynamic Programming,” we discussed the monotonicity and continuity of V with respect to the model parameters. In particular, we have highlighted that V is not necessarily monotone in T , but it is possible to derive sufficient conditions that guarantee monotonicity. A potential area of research is to explore if more of these sufficient conditions can be derived for specific DP problems (i.e., lot-sizing problem, knapsack problem). Another interesting area of research is to identify additional properties of V and π^* with respect to the model parameters. For example, White and El-Deib [35] highlighted that V is piecewise affine and concave in ρ if r^a and $V_T(s)$ are affine in ρ .

A number of problems that are encountered in practice are large and it is not possible to compute the exact solutions for these dynamic programs. These problems are often solved by approximating the value function and/or aggregating the state space. Topaloglu and Powell [48] were able to exploit the structure of their value function approximation to efficiently revise the solution when the model parameters change. It would be interesting to explore how sensitivity analysis can be performed for different approximate dynamic programs.

Lastly, we note that researchers have typically explored problems where the parameter uncertainties are restricted to

one group of parameters, such as uncertainties solely in the rewards or transition probabilities. However, parameter uncertainty is rarely confined to one particular class of parameters in practice. Hence, researchers should also explore problems where different types of parameters are allowed to vary simultaneously.

Acknowledgments

The authors are grateful to an anonymous referee whose remarks improved this article. The authors also gratefully acknowledge National Science Foundation support under Grant No. CMMI-0813671.

REFERENCES

1. Held M, Karp RM. A dynamic programming approach to sequencing problems. *J Soc Ind Appl Math* 1962;10:196–1210.
2. Pruzan PM, Jackson JTR. A dynamic programming application in production line inspection. *Technometrics* 1967;9:73–81.
3. Elton EJ, Gruber MJ. Dynamic programming applications in finance. *J Finance* 1971;26:473–506.
4. Yakowitz S. Dynamic programming applications in water resources. *Water Resour Res* 1982;18:673–696.
5. Bellman R. The theory of dynamic programming. *Proc Natl Acad Sci U S A* 1952;38:716–7719.
6. Powell WB. *Approximate dynamic programming*. New Jersey: John Wiley & Sons, Inc.; 2007.
7. Puumalainen J. Optimal cross-cutting and sensitivity analysis for various log dimension constraints by using dynamic programming approach. *Scand J Forest Res* 1998;13:74–82.
8. Tilahun K, Raes D. Sensitivity analysis of optimal irrigation scheduling using a dynamic programming model. *Aust J Agric Res* 2002;53:339–346.
9. Givan R, Leach S, Dean T. Bounded-parameter Markov decision processes. *Artif Intell* 2000;122:71–109.
10. Iyengar GN. Robust dynamic programming. *Math Oper Res* 2005;30:257–280.
11. Nilim A, El Ghaoui L. Robust control of Markov decision processes with uncertain transition matrices. *Oper Res* 2005;53:780–798.

12. Itoh H, Nakamura K. Partially observable Markov decision processes with imprecise parameters. *Artif Intell* 2007;171:453–490.
13. Briggs A, Sculpher M, Buxton M. Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis. *Health Econ* 1994;3:95–104.
14. Chatfield C. Model uncertainty, data mining and statistical inference. *J R Stat Soc Ser A Stat Soc* 1995;158:419–466.
15. Draper D. Assessment and propagation of model uncertainty. *J R Stat Soc Ser B Methodol* 1995;57:45–97.
16. Walker D, Fox-Rushby JA. Allowing for uncertainty in economic evaluations: qualitative sensitivity analysis. *Health Policy Plann* 2001;16:435–443.
17. Fernandez-Baca D, Venkatachalam B. Sensitivity analysis in combinatorial optimization. Volume 18, *Handbook of approximation algorithms and metaheuristics*. Boca Raton (FL): Chapman & Hall/CRC; 2007. pp. 673–696.
18. Sotskov YN, Leontev VK, Gordeev EN. Some concepts of stability analysis in combinatorial optimization. *Discrete Appl Math* 1995;58:169–190.
19. Wallace SW. Decision Making under uncertainty: is sensitivity analysis of any use? *Oper Res* 2000;48:20–25.
20. Eppstein D. Setting parameters by example. *SIAM J Comput* 2003;32:643–653.
21. Sun F, Fernandez-Baca D, Yu W. Inverse parametric sequence alignment. *J Algorithms* 2004; 36–54.
22. Kim E, Kececioglu J. Inverse sequence alignment from partial examples. *Algorithms Bioinformatics* 2007;53:359–370.
23. McKay MD. Sensitivity analysis. *Proceedings of the Workshop on Validation of Computer-based Mathematical Models in Energy Related Research and Development*. Fort Worth (TX). 1979.
24. Wagner HM. Global sensitivity analysis. *Oper Res* 1995;43:948–969.
25. Wendell RE. Tolerance sensitivity and optimality bounds in linear programming. *Manage Sci* 2004;50:797–803.
26. Filippi C. A fresh view on the tolerance approach to sensitivity analysis in linear programming. *Eur J Oper Res* 2005;167:1–19.
27. Bradley SP, Hax AC, Magnanti TL. *Applied mathematical programming*. Reading (MA): Addison-Wesley; 1977.
28. Wendell RE. The tolerance approach to sensitivity analysis in linear programming. *Manage Sci* 1985;31:564–578.
29. Hansen S, Labbe M, Wendell RE. Sensitivity analysis in multiple objective linear programming: the tolerance approach. *Eur J Oper Res* 1989;38:63–69.
30. Ravi N, Wendell RE. The tolerance approach to sensitivity analysis of matrix coefficients in linear programming. *Manage Sci* 1989;35:1106–1119.
31. Wondolowski FR. Generalization of Wendell's tolerance approach to sensitivity analysis in linear programming. *Decis Sci* 1991;22:792–811.
32. White DJ. *Markov decision processes*. Chichester: John Wiley & Sons, Ltd; 1993.
33. Puterman ML. *Markov decision processes*. New York: John Wiley & Sons, Inc.; 1994.
34. Altman E, Shwartz A. Sensitivity of constrained Markov decision processes. *Ann Oper Res* 1991;32:1–22.
35. White CC, El-Deib HK. Parameter imprecision in finite state, finite action dynamic programs. *Oper Res* 1986;34:120–129.
36. Satia JK, Lave RE. Markovian decision processes with uncertain transition probabilities. *Oper Res* 1973;21:728–740.
37. White CC, Eldeib HK. Markov decision processes with imprecise transition probabilities. *Oper Res* 1994;42:739–749.
38. Kalyanasundaram S, Chong E, Shroff N. Markov decision processes with uncertain transition rates: Sensitivity and robust control. *Proceedings of the 41st IEEE Conference on Decision and Control*, Volume 4; Las Vegas (NV): 2002. pp. 3799–3804.
39. Muller A. How does the value function of a Markov decision process depend on the transition probabilities? *Math Oper Res* 1997;22:872–885.
40. Tan C, Hartman JC. Equipment replacement analysis with an uncertain finite horizon. *IIE Trans* 2010;42:342–353.
41. Harmanec D. Generalizing Markov decision processes to imprecise probabilities. *J Stat Plann Inference* 2002;105:199–213.
42. Hopp WJ. Sensitivity analysis in discrete dynamic programming. *J Optim Theory Appl* 1988;56:257–269.
43. Bean JC, Smith RL. Conditions for the existence of planning horizons. *Math Oper Res* 1984;9:391–401.

44. Hopp WJ. Identifying forecast horizons in nonhomogeneous Markov decision processes. *Oper Res* 1989;37:339–343.
45. Bean JC, Hopp WJ, Duenyas I. A stopping rule for forecast horizons in nonhomogeneous Markov decision processes. *Oper Res* 1992;40:1188–1199.
46. Cheevaprawatdomrong T, Smith RL. A paradox in equipment replacement under technological improvement. *Oper Res Lett* 2003;31:77–82.
47. Ahuja RK, Magnanti TL, Orlin JB. *Network flows*. Englewood Cliffs (NJ): Prentice Hall; 1993.
48. Topaloglu H, Powell WB. Sensitivity analysis of a dynamic fleet management model using approximate dynamic programming. *Oper Res* 2007;55:319–331.
49. Manne AS. Linear programming and sequential decisions. *Manage Sci* 1960;6:259–267.
50. Derman C. On sequential decisions and Markov chains. *Manage Sci* 1962;9:16–24.
51. Goldman AJ, Tucker AW. Theory of linear programming. In: Kuhn HW, Tucker AW, editors. *Linear inequalities and related systems*. Princeton (NJ): Princeton University Press; 1956. pp. 53–97.
52. Williams AC. Complementary theorems for linear programming. *SIAM Rev* 1970;12:135–137.
53. Ward JE, Wendell RE. Approaches to sensitivity analysis in linear programming. *Ann Oper Res* 1990;27:3–38.
54. Gal T, Greenberg HJ. *Advances in sensitivity analysis and parametric programming*. Boston (MA): Kluwer Academic; 1997.
55. Bazaraa MS, Jarvis JJ, Sherali HD. *Linear programming and network flows*. New Jersey: John Wiley & Sons, Inc.; 2005.
56. Winston WL, Venkataramanan M. *Introduction to mathematical programming: applications and algorithms*. 4th ed. Pacific Grove (CA): Duxbury Press; 2003.
57. Bather J. *Decision theory: an introduction to dynamic programming and sequential decisions*. Chichester: Wiley; 2000.

SENSITIVITY ANALYSIS IN DECISION MAKING

EMANUELE BORGONOVO
ELEUSI Research Center and
Department of Decision
Sciences, Bocconi University,
Milan, Italy

INTRODUCTION

Sensitivity analysis is one of the main steps of the modern decision-analysis process [1]. In fact, policy and decision-makers, especially in complex situations, benefit from the use of dedicated quantitative support tools for informing their decisions better. Depending on the area of application, they might be dealing with a linear programming model, a decision tree, a probabilistic risk assessment model, or a complex climate change integrated assessment model. Models capture a variety of phenomena and facets and are usually implemented through computer codes, becoming a black box to the manager. However, a manager “*wants to know why. A process starts of finding out what it was about the inputs that made the outputs come out as they did* [2, B469]”. Managerial questions find their answers by shaking the content of a decision-support model through sensitivity analysis. Sensitivity methods are an essential ingredient to draw the maximum information out of a quantitative model [3].

The rigorous application of sensitivity analysis is recommended by several agencies. The U.S. Environmental Protection Agency [4, p. 19], the Florida Commission on Hurricane Loss Projection Methodology [5], and NASA [6] explicitly include the utilization of sensitivity analysis (or importance measures) as part of best practices in their guidelines and procedures.

Several sensitivity analysis methods have been developed over time and numerous are the insights that an analyst can gain from sensitivity analysis. We will discuss them in

the remainder of this article, which is organized as follows. Section titled “Taxonomy and Notation” discusses the sensitivity analysis jargon and notation. Section titled “Local Sensitivity Analysis” presents local sensitivity methods. Section titled “Global Sensitivity Methods” presents global methods, starting with screening methods, continuing to variance-based, distribution-based, and expected value of information-based methods and estimation. Section titled “Sensitivity Analysis Settings” discusses the insights of a sensitivity analysis exercise through the concept of sensitivity analysis setting.

TAXONOMY AND NOTATION

Some taxonomy is necessary to illustrate the basic concepts of sensitivity analysis. First, the mathematical model of interest outputs a set of quantities of interest to the decision-maker. These quantities are called *decision-support criteria* (in the managerial sciences), endogenous variable (in Economics [7]) or, simply, model output (more diffuse in the mathematics and statistics terminology [8]). The quantity of interest is usually denoted by y . y depends on the set of model inputs $\{x_1, x_2, \dots, x_n\}$, which we denote by $\mathbf{x}$. Depending on the sector, we name the elements of $\mathbf{x}$ exogenous variables, factors, parameters, and model inputs. The model is the relationship that binds y to $\mathbf{x}$:

$$g : \mathbf{x} \mapsto y, \quad \Omega_{\mathbf{x}} \rightarrow \Omega_Y, \quad (1)$$

where $\Omega_{\mathbf{x}} \subseteq \mathbb{R}^k$ and $\Omega_Y \subseteq \mathbb{R}$. In Eq. (1), $\Omega_{\mathbf{x}}$ and Ω_Y are the domain of $\mathbf{x}$ and y , respectively. $g(\mathbf{x})$ is not necessarily known analytically and, most often, is the result of complex calculations performed through a computer code.

$u = \{i_1, i_2, \dots, i_k\}$ denotes a subset of indices that refer to any group of k model inputs ($2 \leq k \leq n$) and $\sim u$ the set of remaining model inputs (complementary set). $\mathbf{x}_u = \{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$ denotes the vector of k model inputs with indices in u .

A symbol of the type I_i (I_u) denotes the sensitivity measure of x_i ($\mathbf{x}_u$).

The computational cost of a sensitivity analysis exercise (C) is conventionally quantified by the number of model runs necessary to estimate the sensitivity measures. For instance, the expression $C = n^2$ is equivalent to saying that computation of the sensitivity measures requires to evaluate the model a number of times equal to the square of the number of model inputs.

LOCAL SENSITIVITY ANALYSIS

The purpose of local sensitivity analysis is to test the response of a decision-support model around some predetermined point of interest, $\mathbf{x}^0 \in \Omega_{\mathbf{x}}$. $\mathbf{x}^0$ is called *base case* or *nominal case*. The simplest form of local sensitivity analysis is represented by tornado diagrams. Their origin can be attributed to Ref. 9. A tornado diagram is a graphical representation of a series of one-factor-at-a-time (OFAT) sensitivities. A tornado diagram is obtained as follows [10]. One assigns a variation range to each factor $[\Delta x_1, \Delta x_2, \dots, \Delta x_n]$. Then, one considers the point $(x_i + \Delta x_i, \mathbf{x}_{-i}^0)$, with x_i moved to the end of its range and the remaining factors fixed at the base case. One then evaluates the differences

$$\Delta^+ y_i = g(x_i + \Delta x_i^+, \mathbf{x}_{-i}^0) - g(\mathbf{x}^0), \quad (2)$$

for all $i = 1, 2, \dots, n$. One also evaluates the differences $\Delta^- y_i = g(x_i - \Delta x_i^+, \mathbf{x}_{-i}^0) - g(\mathbf{x}^0)$.

Usually, the selected variation intervals are symmetric, so that $\Delta x_i^+ = \Delta x_i^- = \Delta x_i$ for all $i = 1, 2, \dots, n$. The results are then displayed as horizontal bars, with the factors ordered by the magnitude of $\Delta^+ y_i$ and $\Delta^- y_i$. Tornado diagrams are widely used in decision-analysis and common commercial software implements them. To illustrate tornado diagrams, we use the following simple example adapted from Ref. 11. Let

$$y = x_1 + \phi x_2 \quad (3)$$

with $0 \leq \phi \leq 1$. For generic variations in $\mathbf{x}$ one obtains

$$\Delta^+ y_i = \Delta x_1 \quad \text{and} \quad \Delta^+ y_2 = \phi \Delta x_2. \quad (4)$$

and

$$\Delta^- y_i = -\Delta x_1 \quad \text{and} \quad \Delta^- y_2 = -\phi \Delta x_2. \quad (5)$$

Selecting $\phi = \frac{1}{2}$, $\Delta x_1 = \frac{1}{2}$, and $\Delta x_2 = \frac{1}{2}$ one obtains the diagram in Fig. 1.

Tornado diagrams can be linked to finite change sensitivity indices, to the method of Morris, and to differential sensitivity [6]. This latter connection is achieved considering $\Delta x_i \rightarrow 0$, and $\lim_{\Delta x_i \rightarrow 0} \frac{\Delta y_i}{\Delta x_i} = \frac{\partial y(\mathbf{x}^0)}{\partial x_i}$. Partial derivatives are the sensitivity measures of interest in two important categories of sensitivity analysis methods: automated differentiation [12] and comparative statics [7]. Moreover, local sensitivity methods play an extremely important role in reliability analysis, field in which they are given the name of

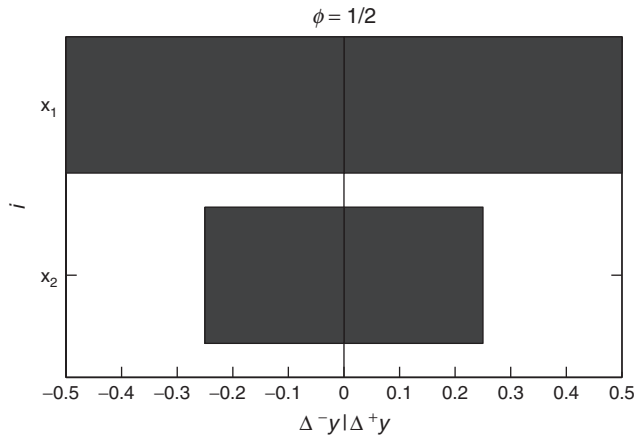


Figure 1. A possible tornado diagram associated with Eq. (3).

importance measures. Importance measures support regulators in risk-informed decision making [13] and in operational decisions concerning graded quality assurance programs, equipment prioritization for maintenance and inspections, and so on.

Reliability importance measures have had an important role also in the general development of sensitivity analysis methods. Historically, the first importance measure is because of Ref. 14. The Birnbaum sensitivity measure starts with a probabilistic interpretation in terms of system failure and represents the probability that component i is critical (we cannot dig into this subject further because of space constraints), and we refer to Ref. 14. Under the assumptions of independent failures and coherent system structure, we can write the Birnbaum importance of input x_i as

$$B_i = \frac{\partial y(\mathbf{x}^0)}{\partial x_i}. \quad (6)$$

Reference 15 shows that B_i can be seen as a particular case of the differential importance measure

$$D_i = \frac{\frac{\partial y(\mathbf{x}^0)}{\partial x_i} dx_i}{\sum_{j=1}^n \frac{\partial y(\mathbf{x}^0)}{\partial x_j} dx_j}. \quad (7)$$

For the differential importance measure the additivity property holds:

$$D_S = \sum_{s=1}^k D_{i_s} \quad (8)$$

$$D_l^T := \frac{\Delta^T G_l}{\Delta G} = \frac{B_l \Delta x_l + \sum_{k=2}^T \sum_{\substack{i_1 < i_2, \dots, i_k \\ l \in i_1 < i_2, \dots, i_k}} J_{i_1, i_2, \dots, i_k}^k(\mathbf{x}^0) \prod_{s=1}^k \Delta x_{i_s}}{\sum_{i=1}^N B_i(\mathbf{x}^0) \cdot \Delta x_i + \sum_{k=2}^T \sum_{i_1 < i_2, \dots, i_k} J_{i_1, i_2, \dots, i_k}^k(\mathbf{x}^0) \prod_{s=1}^k \Delta x_{i_s}}, \quad (11)$$

where $\Delta^T G_l$ is the total change in y provoked by a finite change in the component probabilities ($\Delta \mathbf{x}$). The sum in the extends to order T because the Taylor expansion of a reliability polynomial is exact by the multilinearity of the reliability function polynomial (see Ref. 23 for further details).

and therefore one retrieves the importance of any group of variables as sum of the individual sensitivity measures of the variables in the group. The advantages of this property are discussed in Ref. 15.

A limitation associated with the OFAT nature of the methods presented so far is their inability to display the effects of interactions. Authors have then tried to remedy this shortcoming generalizing the previous definitions including higher order derivatives ([14–22]; for an overview of this process, see Ref. 23). We recall the joint reliability importance measure [16,17], which can be defined as

$$J_{\{i,j\}} = \frac{\partial B_i}{\partial x_j} = \frac{\partial^2 y(\mathbf{x}^0)}{\partial x_j \partial x_i}. \quad (9)$$

Extending the differentiation to higher orders, one obtains the joint reliability importance measure of order k , defined as

$$J_u^k = \frac{\partial^k y(\mathbf{x}^0)}{\partial x_{i_1} \partial x_{i_2} \dots \partial x_{i_k}}. \quad (10)$$

This sensitivity measure provides the strength of the interaction between model inputs $x_{i_1}, x_{i_2}, x_{i_k}$ at $\mathbf{x}^0$.

The total order reliability importance measure [23] brings together in one indicator the Birnbaum and the joint reliability importance measures of all orders. It is defined as

For generic input–output mappings, even when differentiability does not hold, it is shown in Ref. 6 that D_l^T coincide with the total order finite change sensitivity index, which is obtained through a finite-difference decomposition. Finite change sensitivity measures are discussed in the context of inventory management applications in Ref. 24.

Other relevant reliability importance measures are the Fussell-Vesely, the risk achievement worth, and the risk reduction worth importance measures [13], all of which fall in the category of local sensitivity methods.

Local methods inspect the behavior of the model of interest around one reference point. By construction, they are not capable of accounting for uncertainty in $\mathbf{x}^0$. Hence, they ought not to be applied when uncertainty in the model inputs is considered. For this goal, researchers have developed global sensitivity methods.

GLOBAL SENSITIVITY METHODS

In the presence of uncertainty, global methods have become the golden standard of sensitivity analysis [25]. $\mathbf{x} = (x_1, \dots, x_k)$, in the presence of uncertainty becomes a random vector on probability space $(\Omega_{\mathbf{X}}, \mathcal{A}, \mathbb{P}_{\mathbf{X}})$ and is denoted by $\mathbf{X} = (X_1, \dots, X_k)$, a random vector. Correspondingly, $Y = g(\mathbf{X})$ becomes a (function of) random variable. $F_{\mathbf{X}}(\mathbf{x})$ and $F_Y = \mathbb{P}_Y(Y \leq y)$, $f_{\mathbf{X}}(\mathbf{x})$ and $f_Y(\mathbf{x})$ denote the cumulative distributions and probability densities of $\mathbf{X}$ and Y , respectively. Global sensitivity methods can be grouped in the categories of screening [26,27], variance-based [28], expected value of information-based [29], and density-based methods [30].

Screening Methods

Screening methods are designed for identifying the most relevant model inputs with a limited number of model runs. They are usually employed in an initial and exploratory phase, in which the sensitivity of the output on each factor does not need to be determined with extreme precision. The method of Morris consists of performing a series of randomized experiments evaluating the model over proper trajectories (sequences of points that inspects the model input space) and estimating the elementary effects

$$d_i = \frac{\Delta y_i}{\Delta x_i} = g(x_i - \Delta x_i, \mathbf{x}_{-i}^0) - g(\mathbf{x}^0). \quad (12)$$

The effects are then averaged over the trajectory. We refer to Ref. 26 for further details.

The method has, since then, been widely applied and studied. We underline the modifications introduced in Ref. 31.

A further screening method that has been successful in application to simulation codes is sequential bifurcation (SB) [32,33], (see also Ref. 34 for an overview). *SB is a sequential method that changes groups of inputs simultaneously* [[35], p. 1]. We refer to Ref. 35 for a thorough technical account of the method and for its extension into the so-called multi-response SB method.

NonParametric Methods

Historically, in global sensitivity analysis, the family of nonparametric techniques has been the first to be extensively investigated [36,37]. Supposing that the model output is well fitted by a linear regression (i.e., $g(\mathbf{X}) \simeq b_0 + \sum_{s=1}^n b_s X_s$), a natural measure of the sensitivity of Y on X_i is the standardized regression coefficient (SRC) [37]:

$$\text{SRC}_i = \frac{b_i \sigma_i}{\sigma_Y}, \quad (13)$$

where σ_i and σ_Y are the standard deviations of x_i and Y , respectively. If model inputs are independent, then SRC_i coincides with the Pearson regression coefficient

$$\text{PEAR}_i = \frac{\text{Cov}(Y, X_i)}{\sigma_i \sigma_Y}, \quad (14)$$

which is another notable member of the non-parametric methods family. Other representatives of this family are discussed in Refs 36 and [37].

Saltelli and Marivoet [36] demonstrates that the performance of these techniques is synthesized by the model coefficient of determination, R_Y^2 . R_Y^2 is the fraction of the model output variance explained by the regression. When R_Y^2 is low, it becomes “*unrealistic to assess influence of the input variables based on SRC's* [38].” Situations of low R_Y^2 are encountered when models present nonlinearities and interactions. The weakness of these methods can then be overcome in several ways, but two are the main ones. Remaining within the nonparametric method framework, one can define correlations on

ranks, replacing the Pearson correlation with Spearman correlation and using standardized rank regression coefficients (see Ref. 39 for a thorough overview). The second is to resort to variance-based global sensitivity measures. In fact, variance-based methods are capable of accounting for the entire variance of the model output.

Variance-Based Methods

At the basis of variance-based methods is the fundamental result known as *functional analysis of variance (ANOVA) decomposition* of Refs 40,41 (see Ref. 42 for a historical account). Efron and Stein [41] and Sobol' [43] prove that under the assumptions of (i) square integrability of $g(\mathbf{x})$, and (ii) independence of the model inputs [$f_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^k f_i(x_i)$], $g(\mathbf{x})$ can be uniquely expanded in $2^n - 1$ terms:

$$g(\mathbf{x}) = g_0 + \sum_{u \neq \emptyset} g_u(\mathbf{x}_u), \quad (15)$$

where $\sum_u$ refers to the sum over all combinations of subsets of indices and $g_u(\mathbf{x}_u)$ is computed from a series of nested integrations:

$$\begin{aligned} g_0 &= \int_{\Omega_{\mathbf{x}}} g(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \\ g_u(\mathbf{x}_u) &= \int_{\Omega_{\sim u}} \left[g(\mathbf{x}_u, \mathbf{x}_{\sim u}) - \sum_{v \subsetneq u} g_v(\mathbf{x}_v) \right] \\ &\quad f_{\sim u}(\mathbf{x}_{\sim u}) d\mathbf{x}_{\sim u}. \end{aligned} \quad (16)$$

As a result, the $g_u(\mathbf{x}_u)$ functions in Eq. (16) are orthogonal [43]. By orthogonality, $\mathbb{V}[Y] = \mathbb{E}[(g(\mathbf{x}) - g_0)^2]$ can be expressed as

$$\mathbb{V}[Y] = \sum_u V_u(\mathbf{x}_u), \quad (17)$$

with

$$V_u = \int_{\Omega_u} [g_u(\mathbf{x}_u)]^2 f_u(\mathbf{x}_u) d\mathbf{x}_u. \quad (18)$$

By Eqs (17) and (18), the functional decomposition of $g(\cdot)$ is in one-to-one correspondence with the decomposition of its variance.

From Eqs (17) and (18), Ref. 28 defines the so-called variance-based sensitivity index of order k as

$$S_u^k = \frac{V_u}{\mathbb{V}[Y]}. \quad (19)$$

Clearly, $\sum_{r=1}^k \sum_{u:|u|=r} S_u^k = 1$. Of particular relevance are the first order indices

$$S_i^1 = \frac{V_{\{i\}}}{\mathbb{V}[Y]} \quad (20)$$

and the total order indices

$$S_i^T = \frac{\sum_{u:u \ni i} V_u}{\mathbb{V}[Y]}, \quad (21)$$

where $\sum_{u:u \ni i}$ denotes a sum extended to all u that include index i . S_i^T is the portion of $\mathbb{V}[Y]$ associated with X_i and all its interactions with the remaining model inputs. Under the independence assumption, it can be proven that S_i^T is the fraction of $\mathbb{V}[Y]$ remaining after all factors are fixed, leaving the sole X_i free to vary.

From a historical viewpoint, the sensitivity measures have appeared in unrelated journal articles in different fields in the early 1990s. Iman and Hora [44] define variance-based sensitivity measures as $IH_i = \mathbb{V}[\mathbb{E}[Y|X_i]]$ in the context of reliability analysis. It can be proven that, under the independence assumption, I_i is the numerator of S_i^1 . In the same years [45] defines Sobol' sensitivity indices in Eqs (19), (20), and (21) under the independence assumption. Autonomously, Wagner [46] arrives at the definition of a non-normalized versions of both the first (S_i^1) and total order (S_i^T) variance-based measures. Wagner's definition does not rely on the independence assumption.

Variance-based measures might be subject to a misleading interpretation, if one considers them as *the expected reduction in uncertainty due to observing x_i* [47, p. 753]. In fact, what reflects a decision-maker state of knowledge about an uncertain quantity is its distribution [48]. A distribution is completely summarized by variance, under the normality assumption [48]. Similarly, risk is completely characterized by variance only

under a quadratic utility assumption [49]. In Ref. 50, a counterexample shows that stretching the interpretation of variance as representative of uncertainty leads to misleading conclusions in sensitivity analysis.

The study of variance-based sensitivity measures in the presence of dependencies among the model inputs is attracting increasing attention. The interest stems from results in Refs 47 and 51. Bedford [51] shows that the uniqueness of the decomposition in eq. (15) is lost in the presence of dependencies. Oakley and O'Hagan [47] show that terms not related to the functional structure of $g(\mathbf{x})$ emerge in Eq. (15), when correlations are present. Several authors have recently proposed alternative approaches and research is ongoing [52–55].

Density-Based Methods

A class of global sensitivity measures that considers the entire distribution without reference to a particular moment and whose sensitivity measures are well posed in the presence of correlations is the class of density-based methods [30,50,56–58]. The key intuition is to base the sensitivity measures on a comparison of the conditional model output density $f_{Y|X_i}(y)$ and the unconditional model output density $f_Y(y)$. Because of its properties, a widely studied representative of this class is the δ -sensitivity measure introduced in Ref. 30 (also known as *Borgonovo's- δ*) and defined as

$$\delta_i = \frac{1}{2} \mathbb{E} \left[\int_{\Omega_Y} |f_Y(y) - f_{Y|X_i}(y)| dy \right] dx_i, \quad (22)$$

δ_i is a method of recent introduction and attracting increasing attention both concerning its properties and estimation [59,60]. δ_i shares several convenient properties such as normalization ($0 \leq \delta_i \leq 1$), and monotonic transformation invariance. It can be shown [61] that

$$\delta_i = \frac{1}{2} \int_{\Omega_{X_i} \times \Omega_Y} |f_{Y|X_i}(y) - f_Y(y)f_{X_i}(x_i)| dy dx_i. \quad (23)$$

In fact, if Y (the model output) is independent of X_i , then the joint distribution $f_{YX_i}(y)$ is equal to the product of the marginal distributions $f_Y(y)f_{X_i}(x_i)$ and $\delta_i = 0$. The converse also holds. Thus, δ_i is a measure of statistical dependence, rather than of functional dependence.

Transformation Invariant Global Sensitivity Statistics

Monotonic transformation invariance is a property particularly convenient from both a theoretical and a numerical viewpoint. From a decision-analysis viewpoint, often the quantity of interest is not only Y (the model output), but a utility on Y , $u(Y)$. The assignment of the exact functional form to $u(Y)$ might be a cumbersome task [11]. However, if the sensitivity measure is monotonic transformation invariant, its results are unaffected by imprecision in the assessment of $u(Y)$ [11].

Perhaps the most notable advantage of monotonic transformation invariance is related to computation. When the model output is sparse (it ranges several orders of magnitudes) it becomes difficult to obtain convergence of the numerical algorithms that estimate global sensitivity measures. In these cases, analysts resort to transformations of Y for improving numerical accuracy in the estimation. The most widely used are the lognormal and the rank transformations. *The problem associated with estimating the variance of long-tailed distributions is directly attributable to the scale of the numbers involved. The scaling problem most often can be overcome by performing uncertainty importance calculations based on a logarithmic scale* [[44], p. 402]. If the sensitivity measure is transformation invariant, then analysts can exploit the advantage of the transformation avoiding the problem of transferring results of the analysis back to the original data.

A set of transformation invariant sensitivity measures are recently discussed in Ref. 11. Their definition is proposed as

$$\beta_d^Y = \mathbb{E}_i [d(F_Y, F_{Y|X_i})], \quad (24)$$

where $d(F_Y, F_{Y|X_i})$ is any distance between distribution functions which is monotonic transformation invariant. The Kuiper and Kolmogorov–Smirnov metrics are discussed in Ref. 11. Their properties are studied for decision-problem involving investment opportunities.

Value of Information-Based Sensitivity Measures

All the sensitivity measures we have discussed so far are value-based sensitivity measures. None of them captures decision-sensitivity. That is, using a nonparametric, a variance-based, and a distribution-based global sensitivity measure, an analyst is looking at uncertainty in Y and how it is apportioned by its sources. However, in several decision-analysis problems, the decision maker is selecting among competing alternatives under uncertainty. The model output is, in these cases, of the form $Y = \max_{1, 2, \dots, n}(\mathbb{E}[Y_1(\mathbf{X})], \mathbb{E}[Y_2(\mathbf{X})], \dots, \mathbb{E}[Y_n(\mathbf{X})]) = \max_j[\mathbb{E}[Y_j(\mathbf{X})]]$. In that case, getting to know X_i not only changes the distribution of Y , but might change the preferred alternative. The sensitivity measure that captures both decision-sensitivity and value sensitivity is the partial value of expected information [29,62,63].

$$\epsilon_i = \mathbb{E}_{X_i} \left[\max_j \{ \mathbb{E}[Y_j|X_i] \} \right] - \max_j \{ \mathbb{E}[Y_j(\mathbf{X})] \}. \quad (25)$$

The expected value of perfect information can be interpreted as “the difference between the expected profit that we shall obtain as a result of the clairvoyance and the expected profit that we would obtain without the clairvoyance” [64] on X_i . ϵ_i is discussed in decision-making applications with particular reference to the medical sector in Refs 29 and 62. It is also proposed as an importance measure for decision making in reliability in Ref. 65.

Estimation

The estimation of global sensitivity measures has been the subject of investigation in numerous works [28,66–68] and ongoing

research is vast. The reason is that the analytical estimation of global sensitivity measures is possible in a few cases [69]. However, in most applications, the input–output mapping $g(\mathbf{x})$ is not known analytically, and the sensitivity measures need to be computed numerically. The numerical estimation consists of the generation of an appropriate sample followed by the evaluation of the model at the corresponding points. The size of the sample is denoted by N . This step is often called (although with some vagueness in the terminology) *Monte Carlo simulation* or *uncertainty analysis*. Among the sample generation methods, we recall Latin Hypercube [70,71], Sobol’ quasi random LpTau [72], and Halton sequences [73].

The design of the Morris method is associated with a cost $C = 2 \times r \times n$, where r is the number of replicates for each factor.

No particular design is required in the estimation of nonparametric techniques, which can be obtained directly from the sample generated by a Monte Carlo simulation. Their cost is equal to $C = N$ model evaluations.

Variance-based, density-based, distribution-based, and value of information-based sensitivity measures are associated with a cost of $C = n^2N$, if a brute force estimation strategy is envisioned. For variance-based sensitivity measures, however, notable computational burden reduction results have been obtained in Refs 74 and 75, with the computation cost lowered to $C = (k + 2)N$. However, most recently, the works [61,76], and [77] show that all global sensitivity measures can be estimated at the same cost of nonparametric methods, further lowering the computational cost to $C = N$ model runs. Thus, state-of-the art allows the estimation of all first order global sensitivity measures directly following a Monte Carlo simulation, without the need of a specific design.

For computationally intensive models, the estimation might be carried in two steps, with the first step being the reduction of the model through a metamodel. Alternative types of metamodeling methods are discussed in Refs 68,78–84.

SENSITIVITY ANALYSIS SETTINGS

An important aspect of a sensitivity analysis is to clearly establish its goals. Often the goal of a sensitivity analysis is specific and tailored to answer a very particular question. This creates issues especially when the unclear statement of the question leads analysts to utilize methods not consistent with the goal or with their degree of belief. The clearest example is the utilization of local methods in the presence of uncertainty. For this reason, the literature has elaborated ways for making the use of sensitivity analysis methods systematic introducing the concept of sensitivity analysis setting [8,24]. A setting is a formulation of the sensitivity analysis problem that allows the analyst to select the most appropriate method for obtaining the desired answer. A setting must be stated before the sensitivity analysis is executed. On the basis of an investigation of different types of problems that have been encountered in the literature, we can identify the following main settings. *Robustness analysis* (For a thorough overview of robustness analysis, we refer to Refs 85–87.) In decision analysis, a relevant goal of sensitivity analysis is to find the region of Ω_X such that the optimal policy remains unchanged. In connection with influence-diagrams, methods for robustness are discussed in Refs 88–90. In problems supported by Bayesian networks, robustness analysis is discussed in Refs 91–93. The tolerance sensitivity approach of Wendell [94,95] is a robustness approach developed for linear programming problems. In a tolerance sensitivity analysis the goal is to find the variation of the parameters of a linear programming problem that insures the stability of the optimal basis.

A second setting is *factor prioritization* [8]. In a factor prioritization, the goal of the analysis is to rank factors based on their influence on the model output. The complementary setting, that is, identifying the factors that matter the least on the output variation is called *factor fixing* [96]. According to Ref. 97, screening methods find their collocation within this setting. Of interest for decision-makers is also the appreciation of the sign of

change of the model output given a change in the model inputs. This setting is at the basis of the comparative statics method in Economics [7], and is formalized as the *direction of change* setting in Ref. 24. Of further relevance to the decision-maker is to understand the model structure, that is, whether the action of a model input is individual or because of its interactions with other factors. This can be formalized in the *model structure* setting [24].

REFERENCES

1. Clemen RT. Making hard decisions: an introduction to decision analysis. Pacific Grove (CA): Duxbury Press; 1997.
2. Little JDC. Models and managers: the concept of a decision calculus. *Manage Sci* 1970;16(8):1841–1853.
3. Rabitz H. Systems analysis at the molecular scale. *Science* 1989;246:221–226.
4. US EPA. Guidance on the development, evaluation, and application of environmental models; March 2009.
5. Iman RL, Johnson ME, Watson CC Jr. Uncertainty analysis for computer model projections of hurricane losses. *Risk Anal* 2005;25(5):1299–1312.
6. Borgonovo E, Smith CL. A Study of Interactions in the Risk Assessment of Complex Engineering Systems: An Application to Space PSA. *Operations Research* 2011;59(6):1461–1476.
7. Samuelson P. Foundations of economic analysis. Cambridge (MA): Harvard University Press; 1947.
8. Saltelli A, Tarantola S. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *J Am Stat Assoc* 2002;97:702–709.
9. Howard RA. Decision analysis: practice and promise. *Manage Sci* 1988;34(6):679–695.
10. Eschenbach TG. Spiderplots versus tornado diagrams for sensitivity analysis. *Interfaces* 1992;22:40–46.
11. Baucells M, Borgonovo E. Invariant probabilistic sensitivity analysis. *Manage Sci* 2013. Forthcoming.
12. Griewank A. Evaluating derivatives, principles and techniques of algorithmic differentiation. Volume 19, *Frontiers in Applied Mathematics*. Philadelphia (PA): SIAM; 2000.

13. Cheok MC, Parry GW, Sherry RR. Use of importance measures in risk-informed regulatory applications. *Reliab Eng Syst Saf* 1998;60(3):213–226.
14. Birnbaum LW. On the importance of different elements in a multielement system. *Multivariate analysis*. Volume 2,. New York: Academic Press; 1969. p 1–15.
15. Borgonovo E, Apostolakis GE. A new importance measure for risk-informed decision making. *Reliab Eng Syst Saf* 2001;72(2):193–212.
16. Hong JS, Lie CH. Joint reliability-importance of two edges in an undirected network. *IEEE Trans Reliab* 1993;42(1):17–23.
17. Armstrong MJ. Joint reliability-importance of elements. *IEEE Trans Reliab* 1995;44(3):408–412.
18. Andrews JD, Beeson S. Birnbaum's measures of component importance for noncoherent systems analysis. *IEEE Trans Reliab* 2003;52(2):213–219.
19. Zio E, Podofillini L. Accounting for components interactions in the differential importance measure. *Reliab Eng Syst Saf* 2006;91:1163–1174.
20. Lu L, Jiang J. Joint failure importance for noncoherent fault trees. *IEEE Trans Reliab* 2007;56(3):435–443.
21. Gao X, Cui L, Li J. Analysis for joint importance of components in a coherent system. *Eur J Oper Res* 2007;182:282–299.
22. Do Van P, Barros A, Berenguer C. Reliability importance analysis of markovian systems at steady state using perturbation analysis. *Reliab Eng Syst Saf* 2008;93(1):1605–1615.
23. Borgonovo E. The reliability importance of components and prime implicants in coherent and non-coherent systems including total-order interactions. *Eur J Oper Res* 2010;204(3):485–495.
24. Borgonovo E. Sensitivity analysis with finite changes: An application to modified eq models. *Eur J Oper Res* 2010;200(1):127–138.
25. Saltelli A, Ratto M, Tarantola S, Campolongo F. Update 1 of: sensitivity analysis for chemical models. *Chem Rev* 2012;112:PR1–PR21. *Perennial Reviews*, <http://dx.doi.org/10.1021/cr200301u>. DOI: 10.1021/cr200301u.
26. Morris MD. Factorial sampling plans for preliminary computational experiments. *Technometrics* 1991;33(2):161–174.
27. Campolongo F, Saltelli A, Cariboni J. From screening to quantitative sensitivity analysis. a unified approach. *Comput Phys Commun* 2011;4:978–988.
28. Sobol' IM. Sensitivity estimates for nonlinear mathematical models. *Math Model Comput Exp* 1993;1:407–414.
29. Oakley JE. Decision-theoretic sensitivity analysis for complex computer models. *Technometrics* 2009;51(2):121–129.
30. Borgonovo E. A new uncertainty importance measure. *Reliab Eng Syst Saf* 2007;92(6):771–784.
31. Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. *Environ Model Softw* 2007;22:1509–1518.
32. Bettonvil B, Kleijnen JPC. Searching for important factors in simulation models with many factors: sequential bifurcation. *Eur J Oper Res* 1997;96(1):180–194.
33. Wan H, Ankenman BE, Nelson BL. Improving the efficiency and efficacy of Controlled Sequential Bifurcation for simulation factor screening. *INFORMS J Comput* 2010;22(3):482–492.
34. Kleijnen JPC. Factor screening in simulation experiments: review of sequential bifurcation. In: Axelopoulos C, Goldsman D, Wilson JR, editors. *Advancing the frontiers of simulation: a festschrift in Honor of George S. Fishman*. New York: Springer-Verlag; 2009. p 169–173.
35. Kleijnen JPC, Shi W, Liu Z. Factor screening for simulation with multiple responses: sequential bifurcation. *SSRN* 2012;2012-032: 1–36.
36. Saltelli A, Marivoet J. Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques. *Reliab Eng Syst Saf* 1990;28(2):229–253.
37. Helton JC. Uncertainty and sensitivity analyses techniques for use in performance assessment for radioactive waste disposal. *Reliab Eng Syst Saf* 1993;42:327–367.
38. Campolongo F, Saltelli A. Sensitivity analysis of an environmental model: an application of different analysis methods. *Reliab Eng Syst Saf* 1997;57(1):49–69.
39. Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliab Eng Syst Saf* 2006;91(10–11):1175–1209.
40. Hoeffding W. A class of statistics with asymptotically normal distribution. *Ann Math Stat* 1948;19:293–325.
41. Efron B, Stein C. The jackknife estimate of variance. *Ann Stat* 1981;9(3):586–596.

42. Owen AB. Variance components and generalized sobol indices. Stanford. 2012. Available at <http://stat.stanford.edu/owen/reports/newsobol.pdf>.
43. Sobol' IM Multidimensional quadrature formulas and Haar functions. Moscow: Nauka; 1969. (In Russian).
44. Iman RL, Hora SC. A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Anal* 1990;10(3):401–406.
45. Sobol' IM. Sensitivity analysis for non-linear mathematical models. *Math Model Comput Exp* 1993;1:407–414. English translation of Russian original paper.
46. Wagner HM. Global sensitivity analysis. *Oper Res* 1995;43(6):948–969.
47. Oakley J, O'Hagan A. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J R Stat Soc, Series B* 2004;66:751–769.
48. Huang CF, Litzenberger RH. Foundations for financial economics. Upper Saddle River (NJ): Prentice Hall; 1988.
49. Cox LA. Why risk is not variance: an expository note. *Risk Anal* 2008;28(4):925–928.
50. Borgonovo E. Measuring uncertainty importance: investigation and comparison of alternative approaches. *Risk Anal* 2006;26(5):1349–1361.
51. Bedford T. Sensitivity indices for (tree)-dependent variables. Proceedings of the Second International Symposium on Sensitivity Analysis of Model Output. Venice, Italy; 1998. p 17–20.
52. Kucherenko S, Tarantola S, Annoni P. Estimation of global sensitivity indices for models with dependent variables. *Comput Phys Commun* 2012;183(4):937–946.
53. Mara TA, Tarantola S. Variance-based sensitivity indices for models with dependent inputs. *Reliab Eng Syst Saf* 2012;107:115–121.
54. Li G, Rabitz H. General formulation of hdmr component functions with independent and correlated variables. *J Math Chem* 2012;50(1):99–130.
55. Chastaing G, Gamboa F, Prieur C. Generalized hoeffding-sobol decomposition for dependent variables: application to sensitivity analysis. *Electron J Statist* 2012;6:2420–2448.
56. Park CK, Ahn KI. A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment. *Reliab Eng Syst Saf* 1994;46:253–261.
57. Chun M-H, Han S-J, Tak N-I. An uncertainty importance measure using a distance metric for the change in a cumulative distribution function. *Reliab Eng Syst Saf* 2000;70(3):313–321.
58. Borgonovo E, Castaings W, Tarantola S. Moment independent uncertainty importance: new results and analytical test cases. *Risk Anal* 2011;31(3):404–428.
59. Liu Q, Homma T. A new computational method of a moment-independent importance measure. *Reliab Eng Syst Saf* 2009;94(7):1205–1211.
60. Ratto M, Pagano A, Young PC. Non-parametric estimation of conditional moments for sensitivity analysis. *Reliab Eng Syst Saf* 2009;94(2):237–243. feb
61. Plischke E, Borgonovo E, Smith CL. Global Sensitivity Measures from Given Data. *Eur J Oper Res* 2012;226:536–550.
62. Felli JC, Hazen G. Sensitivity analysis and the expected value of perfect information. *Med Decis Mak* 1998;18:95–109.
63. Oakley JE, Brennan A, Tappenden P, Chilcott J. Simulation sample sizes for monte carlo partial evpi calculations. *J Health Econ* 2010;29(3):468–477.
64. Howard RA. Information value theory. *IEEE Trans Syst Sci Cybern* 1966;2(1):22–26.
65. Pörn K. A decision-oriented measure of uncertainty importance for use in psa. *Reliab Eng Syst Saf* 1997;56(1):17–27.
66. Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models. *Reliab Eng Syst Saf* 1996;52(1):1–17.
67. Saltelli A. Sensitivity analysis for importance assessment. *Risk* 2002;22(3):579.
68. Oakley JE, O'Hagan A. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J R Stat Soc B* 2004;66(3):751–769.
69. Borgonovo E, Castaings W, Tarantola S. Moment independent uncertainty importance: new results and analytical test cases. *Risk Anal* 2011;31(3):404–428. In preparation.
70. Helton JC, Davis FJ, Johnson JD. A comparison of uncertainty and sensitivity analysis results obtained with random and latin hypercube sampling. *Reliab Eng Syst Saf* 2005;89:305–330.
71. Sallaberry CJ, Helton JC, Hora SC. Extension of latin hypercube samples with correlated variables. *Reliab Eng Syst Saf* 2008;93(7):1047–1059.

72. Sobol' IM. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput Math Math Phys* 1967;7:86–112.
73. Owen AB. Halton sequences avoid the origin. *SIAM Rev* 2006;48(3):487–503.
74. Saltelli A. Making best use of model evaluations to compute sensitivity indices. *Comput Phys Commun* 2002;145:280–297.
75. Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput Phys Commun* 2010;181(2): 259–270.
76. Strong M, Oakley JE, Chilcott J. Managing structural uncertainty in health economic decision models: a discrepancy approach. *J R Stat Soc, Series C* 2012;61(1):25–45.
77. Strong M, Oakley JE. An efficient method for computing partial expected value of perfect information for correlated inputs. *J Med Dec Mak* 2012. submitted to.
78. Santner TJ, Williams BJ, Notz WI. The design and analysis of computer experiments, Springer Series in Statistics. New York: Springer-Verlag; 2003.
79. Bayarri MJ, Berger J, Steinberg DM. Special issue on computer modeling. *Technometrics* 2009;51(4):353.
80. Chong Gu. Smoothing spline ANOVA models. Berlin: Springer-Verlag; 2002.
81. Ratto M, Pagano A, Young PC. State Dependent Parameter metamodeling and sensitivity analysis. *Comput Phys Commun* 2007;177(11):863–876.
82. Jack P, Kleijnen C. Design and analysis of simulation experiments. Berlin: Springer-Verlag; 2008.
83. Jack P, Kleijnen C. Kriging metamodeling in simulation: a review. *Eur J Oper Res* 2009;192(3):707–716.
84. Sudret B. Global sensitivity analysis using polynomial chaos expansion. *Reliab Eng Syst Saf* 2008;93:964–979.
85. Roy B. Robustness in operational research and decision aiding: a multi-faceted issue. *Eur J Oper Res* 2010;3:629–638.
86. Rosenhead J. Robustness analysis: keeping your options open. In: Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited: problem structuring methods for complexity, uncertainty and conflicts*. Chichester: John Wiley & Sons; 2011. p 181–1207.
87. Rosenhead J. Robustness analysis. *Wiley Encyclopedia of Operations Research and Management Science* 2011. DOI: 10.1002/9780470400531.eorms0976.
88. Nielsen TD, Jensen FV. Sensitivity analysis in influence diagrams. *IEEE Trans Man Cybern* 2003;33(1):223–234.
89. Bhattacharjya D, Shachter R. Sensitivity analysis in decision circuits. In: McAllester D, Myllymaki P, editors. *Proceedings of 24th UAI. Finland, Helsinki: AUAI Press; 2008. p 34–42.*
90. Bhattacharjya D, Shachter R. Three new sensitivity analysis methods for influence diagrams. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island (CA): AUAI Press; 2010. p 1–9.*
91. Chan H, Darwiche A. Sensitivity analysis in bayesian networks: from single to multiple parameters. *Uncertainty and Artificial Intelligence; 2004. p 67–75.*
92. Chan H, Darwiche A. A distance measure for bounding probabilistic belief change. *Int J Approx Reason* 2005;38(2):149–174.
93. Brosnan AJ. Sensitivity analysis of a bayesian belief network in a tactical intelligence application. *J Battlefield Technol* 2006;9(2): 33–40.
94. Wendell RE. The tolerance approach to sensitivity analysis in linear programming. *Manage Sci* 1985;31(5):564–578.
95. Ravi N, Wendell RE. The tolerance approach to sensitivity analysis of matrix coefficients in linear programming. *Manage Sci* 1989;35(9):1106–1119.
96. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. *Global sensitivity analysis—the primer*. Chichester: John Wiley & Sons; 2008.
97. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. *Global sensitivity analysis—the primer*. Chichester: John Wiley & Sons; 2008.

SENSITIVITY ANALYSIS IN LINEAR PROGRAMMING

CARLO FILIPPI
Department of Quantitative
Methods, University of
Brescia, Brescia, Italy

Sensitivity analysis in linear programming studies the stability of optimal solutions and the optimal objective value with respect to perturbations in the input data. This analysis is often relevant in practical applications. Usually, data are estimates of unknown parameters, such as prices of raw materials, customer demands, and travel times. In this case, the decision maker may want to know if better estimates are required before making decisions. For instance, if a parameter is not exactly known, but sensitivity analysis implies that the solution associated with the available parameter estimate remains optimal for a wide range of values around it, then a better, more expensive estimate may not be required. Conversely, if a sensitivity analysis shows that the optimal solution changes with a small perturbation of the parameter value, then a better estimate might be crucial before making any decision. It may also happen that some parameters, such as the availability of certain resources, are not exactly known simply because the decision maker can vary their values within a certain range without changing the structure of the problem. In this case, sensitivity analysis may help the decision maker to choose the most appropriate values of the parameters. Evaluating stability of the model output in front of small perturbations of the input data may be useful also for model validation.

Input data in a linear program are the constraint matrix, the right-hand side vector, and the objective vector. Once a linear program has been solved, a decision maker may ask questions such as the following:

- What happens to the obtained solution when the right-hand side/objective vector is perturbed?
- What is the rate of change in the optimal value when the right-hand side/objective vector is perturbed?
- In what region/interval may such a vector be varied while maintaining the same rate of change?

We discuss different approaches for answering questions like those. We focus on the effects of perturbations in the right-hand side and objective vectors, since such perturbations represent the most significant cases in practice. Perturbations on the constraint coefficient matrix are hard to handle analytically, and often a simple what-if analysis on different values of the sensible coefficients seems more appropriate than a sensitivity analysis.

Sensitivity analysis is often associated with parametric programming (see **Parametric LP Analysis**). In fact, these topics share many common features, but the focus is different: sensitivity analysis deals with the effects of local perturbations of parameters, whereas parametric programming is concerned with a systematic description of the optimal solution and the optimal value as explicit functions of the parameters.

PRELIMINARIES AND NOTATION

Consider the following primal and dual optimization problems in standard form (see **The Simplex Method and Its Complexity**):

$$(P) \quad \max\{c^T x : Ax = b, x \geq 0\},$$

$$(D) \quad \min\{y^T b : y^T A \geq c^T\},$$

where A is an $m \times n$ real matrix with full row rank, and c, b, x , and y are real vectors of appropriate size. Let $E = \{1, 2, \dots, n\}$ and $M = \{1, 2, \dots, m\}$; let the columns and the rows of A be indexed by E and M respectively, according to the natural ordering. For any $J \subseteq E$, let x_J be the subvector of x indexed by J and let A_J be the submatrix of A consisting of columns indexed by J . In particular, x_j is

the j th entry of vector x . A *basis* is a subset B of m elements of E such that A_B is nonsingular. A *cobasis* is the complement of a basis: $N = E \setminus B$.

The *primal basic solution* associated with basis B is the n -vector $p = (p_B, p_N) = (A_B^{-1}b, 0)$; the *dual basic solution* associated with basis B is the m -vector $d = (c_B^T A_B^{-1})^T$. The basis B is *primal feasible* if the associated primal basic solution is feasible, that is if $p_B \geq 0$; *dual feasible* if the associated dual basic solution is feasible, that is if $d^T A_N \geq c_N^T$. The basis B is *primal degenerate* if vector p_B has some zero entry; *dual degenerate* if vector $c_N^T - d^T A_N$ has some zero entry (see **Degeneracy and Variable Entering/Exiting Rules**). The basis B is *optimal* if it is both primal and dual feasible. If B is optimal, then the optimal value of the objective function is

$$z^* = c_B^T A_B^{-1} b = c_B^T p_B = d^T b.$$

Consider the following illustrative example:

$$\begin{aligned} \max \quad & 4x_1 + 5x_2 \\ \text{subject to} \quad & x_1 + x_2 \leq 10 \\ & x_1 + 3x_2 \leq 18 \\ & x_1, x_2 \geq 0. \end{aligned}$$

The feasible region is shown in Fig. 1. The unique optimal solution is the corner point $x = (6, 4)^T$. The equivalent standard form of the above model is

$$\begin{aligned} \max \quad & 4x_1 + 5x_2 \\ \text{subject to} \quad & x_1 + x_2 + x_3 = 10 \\ & x_1 + 3x_2 + x_4 = 18 \\ & x_1, x_2, x_3, x_4 \geq 0, \end{aligned}$$

where $c = (4, 5, 0, 0)^T$, $b = (10, 18)^T$, and

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 3 & 0 & 1 \end{bmatrix}.$$

The unique optimal solution $(6, 4, 0, 0)^T$ is the primal basic solution associated with basis $B = \{1, 2\}$ and cobasis $N = \{3, 4\}$. We have

$$A_B^{-1} = \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix},$$

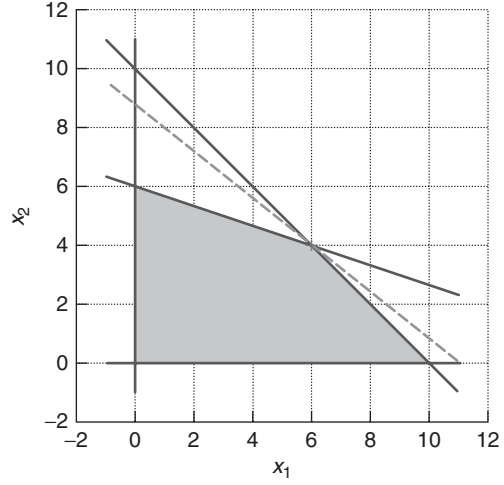


Figure 1. Feasible region of the illustrative example (shaded polygon) and slope of the objective function (dashed line).

hence $p_B = (6, 4)^T$, $p_N = (0, 0)^T$, $d = (7/2, 1/2)^T$. It is easy to check that B is optimal and that B is both primal and dual nondegenerate. The optimal value is $z^* = 44$.

NONDEGENERATE CASE

The Perturbed Problem

Consider a pair of problems (P) and (D) for fixed input A, b , and c . Let $B \subseteq E$ be an optimal basis, both primal and dual nondegenerate. Let $v \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$ be perturbation vectors for the right-hand side vector b and the objective vector c respectively, and consider the perturbed problem:

$$(P(v, w)) \quad \max\{(c + w)^T x : Ax = b + v, x \geq 0\}.$$

The standard approach to sensitivity analysis in linear programming is to characterize the perturbation vectors v and w that preserve the optimality of B . The primal basic solution associated with B in $(P(v, w))$ is

$$p(v) = (p_B(v), p_N(v)) = (p_B + A_B^{-1}v, 0),$$

the dual basic solution is

$$d(w) = d + (w_B^T A_B^{-1})^T.$$

Notice that perturbations on the right-hand side vector affect only the primal basic solution, and thus may affect the primal feasibility of the basis, but not its dual feasibility. Conversely, perturbations on the objective vector affect only the dual basic solution, and thus may affect the dual feasibility of the basis, but not its primal feasibility. Moreover, the dual basic solution is not affected by perturbations on the objective coefficients corresponding to the cobasis N .

Consider the following polyhedral cones [1]:

$$\begin{aligned} C_P &= \{v \in \mathbb{R}^m : -A_B^{-1}v \leq p_B\}, \\ C_D &= \{w \in \mathbb{R}^n : w_N^T - w_B^T A_B^{-1} A_N \\ &\leq -(c_N^T - d^T A_N)\}. \end{aligned}$$

We call C_P and C_D respectively the *primal* and *dual basic cone* associated with B . Primal nondegeneracy implies that $v = 0$ lies in the interior of C_P ; dual nondegeneracy implies that $w = 0$ lies in the interior of C_D . One can check that $p(v) \geq 0$, and thus B is primal feasible for $(P(v, w))$, if and only if $v \in C_P$. Furthermore, $d(w)^T A \geq c^T$, and thus B is dual feasible for $(P(v, w))$, if and only if $w \in C_D$.

If $v \in C_P$ and $w \in C_D$, then B is optimal and the optimal objective value is expressed as follows:

$$\begin{aligned} z(v, w) &= (c_B + w_B)^T A_B^{-1} (b + v) \\ &= z^* + d^T v + w_B^T p_B + w_B^T A_B^{-1} v. \end{aligned}$$

Notice that $z(v, w)$ is a quadratic function of v and w , in general neither convex nor concave. However, if we fix a value $v' \in P$, then $z(v', w)$ is an affine function of w , and if we fix a value $w' \in D$, then $z(v, w')$ is an affine function of v . In particular,

$$z(v, 0) = z^* + d^T v,$$

so that the dual vector d is the gradient of $z(v, 0)$ with respect to v in the interior of C_P . Similarly,

$$z(0, w) = z^* + w_B^T p_B,$$

so that the primal vector p is the gradient of $z(0, w)$ with respect to w in the interior of C_D . In fact, d is often called the vector of *shadow prices* for the entries of b , whereas p_B is often called the vector of *shadow costs* for the variables in x_B (the shadow costs of the variables in x_B are always zero).

In our example, the primal and dual basic cones are respectively

$$\begin{aligned} C_P &= \{v \in \mathbb{R}^2 : -(3/2)v_1 + (1/2)v_2 \leq 6; \\ &\quad (1/2)v_1 - (1/2)v_2 \leq 4\}, \\ C_D &= \{w \in \mathbb{R}^4 : w_3 - (3/2)w_1 + (1/2)w_2 \leq 7/2; \\ &\quad w_4 + (1/2)w_1 - (1/2)w_2 \leq 1/2\}. \end{aligned}$$

The first cone is depicted in Fig. 2. Notice that every choice of $v \in C_P$ corresponds to a translation of the constraints of the feasible region such that the corner point is still a primal feasible solution; every choice of $w \in C_D$ corresponds to a rotation of the objective vector such that the corner point is still a dual feasible solution.

Ordinary Sensitivity Analysis

From a theoretical point of view (and considering only the nondegenerate case), the cones C_P and C_D , and the expressions $p(v)$, $d(w)$, and $z(v, w)$, fully characterize the stability

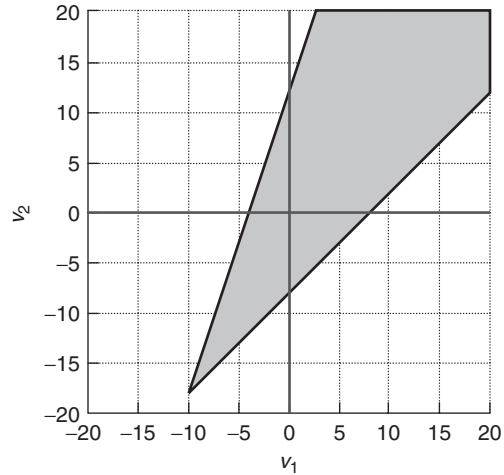


Figure 2. Optimal primal basic cone in the illustrative example.

of B . From a practical point of view, the description of C_P and C_D may be difficult to interpret for a decision maker. Even in our toy example, C_D is a four-dimensional object, impossible to visualize exactly. This drawback may be overcome by restricting attention to variations of a single entry of either the right-hand side or the objective vector at a time. In other words, for every coefficient, we look for the *critical interval* where such a coefficient may vary without affecting the optimality of the current basis, while all other coefficients are fixed to their original values. Such an approach is known as *ordinary* sensitivity analysis and it is usually implemented in commercial packages for linear optimization.

More specifically, for all $i \in M$, we let $v = \beta e^{(i)}$, where β is a scalar and $e^{(i)}$ is the m -vector with entry 1 in i th position, 0 elsewhere. In this way, we consider just the intersection of the primal basic cone with the i th coordinate axis, and we obtain an interval I_i , possibly unbounded, containing 0, and such that basis B remains primal feasible and thus optimal, if and only if $\beta \in I_i$. The computation of I_i proceeds as follows. Let α_{ki} denote the entry of A_B^{-1} on the row indexed by $k \in B$ and the column indexed by $i \in M$. By imposing $v = \beta e^{(i)}$, the conditions $-A_B^{-1}v \leq p_B$ defining C_P boil down to

$$-\beta \alpha_{ki} \leq p_k \quad \text{for all } k \in B,$$

that is,

$$\begin{aligned} \beta &\geq -p_k / \alpha_{ki} & \text{for all } k \in B \text{ with } \alpha_{ki} > 0, \\ \beta &\leq -p_k / \alpha_{ki} & \text{for all } k \in B \text{ with } \alpha_{ki} < 0. \end{aligned}$$

The last two conditions define the extrema of I_i . If $\alpha_{ki} \leq 0$ for all $k \in B$, then I_i is unbounded from below; if $\alpha_{ki} \geq 0$ for all $k \in B$, then I_i is unbounded from above. Note that all data needed for computation are directly available in the simplex tableau corresponding to B (see **The Simplex Method and Its Complexity**).

Analogously, for all $j \in E$, we let $w = \gamma e^{(j)}$, where γ is a scalar and $e^{(j)}$ is the n -vector with entry 1 in j th position, 0 elsewhere. In this way, we obtain an interval J_j , possibly unbounded, containing 0, and such that basis

B remains dual feasible and thus optimal if and only if $\gamma \in J_j$. The computation of J_j proceeds as follows. If $j \in N$, then the conditions $w_N^T - w_B^T A_B^{-1} A_N \leq -(c_N^T - d^T A_N)$ defining C_D boil down to the unique condition

$$\gamma \leq -(c_j - d^T A_j),$$

so that J_j is unbounded from below. Notice that $r_j = c_j - d^T A_j$ is often called the *reduced cost coefficient* of the primal variable x_j with respect to basis B . If $j \in B$, then the conditions defining C_D boil down to

$$-\gamma \sum_{i \in M} \alpha_{ji} a_{ik} \leq -r_k \quad \text{for all } k \in N,$$

where a_{jk} denotes the entry of A_N on the row indexed by $j \in B$ and the column indexed by $k \in N$. Thus we have

$$\begin{aligned} \gamma &\geq r_k / \left(\sum_{i \in M} \alpha_{ji} a_{jk} \right) & \text{for all } k \in N \text{ with} \\ &\sum_{i \in M} \alpha_{ji} a_{ik} > 0, \\ \gamma &\leq r_k / \left(\sum_{i \in M} \alpha_{ji} a_{jk} \right) & \text{for all } k \in N \text{ with} \\ &\sum_{i \in M} \alpha_{ji} a_{ik} < 0. \end{aligned}$$

The last two conditions define the extrema of J_j . If $\sum_{i \in M} \alpha_{ji} a_{jk} \leq 0$ for all $k \in N$, then J_j is unbounded from above; if $\sum_{i \in M} \alpha_{ji} a_{jk} \geq 0$ for all $k \in N$ then J_j is unbounded from below. Again, all data needed for computation are directly available in the simplex tableau corresponding to B . In particular, $\sum_{i \in M} \alpha_{ji} a_{jk}$ is simply the entry of the tableau on the row and column indexed by j and k respectively.

In our small example, if we let $v = (\beta, 0)^T$, then we get

$$C_P \cap \{(\beta, 0) : \beta \in \mathbb{R}\} = \{(\beta, 0) : -4 \leq \beta \leq 8\},$$

and thus $I_1 = [-4, 8]$. If we let $v = (0, \beta)^T$, then we get

$$C_P \cap \{(0, \beta) : \beta \in \mathbb{R}\} = \{(0, \beta) : -8 \leq \beta \leq 12\},$$

and thus $I_2 = [-8, 12]$. The cuts originating these intervals are evidenced in Fig. 3.

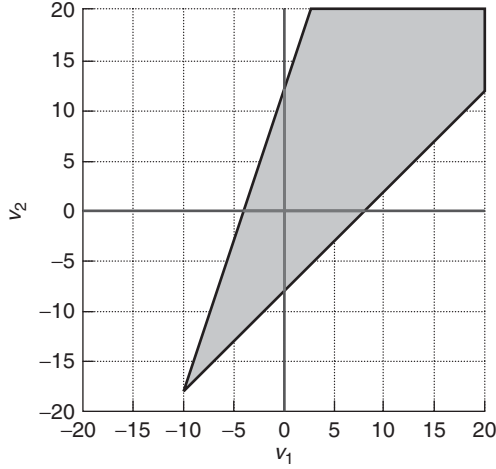


Figure 3. Ordinary sensitivity analysis: cuts of the primal basic cone originating the right-hand side intervals (bold segments).

Similarly, ordinary sensitivity analysis with respect to the objective coefficients lead to the following intervals:

$$\begin{aligned} J_1 &= [-7/3, 1], J_2 = [-1, 7], \\ J_3 &= (-\infty, 7/2], J_4 = (-\infty, 1/2]. \end{aligned}$$

As already noticed, for a nonbasic index j it always happens that the interval returned by ordinary sensitivity analysis has the form $(-\infty, -r_j]$, where r_j is the reduced cost coefficient of x_j with respect to the current basis. The objective coefficients c_j with $j \in N$ may be perturbed simultaneously and independently inside their corresponding intervals, maintaining the optimality of B . However, in general the critical intervals obtained by ordinary sensitivity analysis are not valid when different coefficients vary simultaneously, and this is a severe restriction to their applicability. The next two approaches to sensitivity analysis try to surmount this restriction.

The 100% Rule

The 100% rule exploits two facts: (i) the intervals obtained from ordinary sensitivity analysis are subsets of the basic cones, and (ii) the basic cones are convex sets [1]. Hence, any convex combination of points contained

in critical intervals is contained in the corresponding basic cone and thus, the corresponding perturbation preserves the optimality of the current basis while allowing simultaneous perturbations of the coefficients.

Let us focus on right-hand side perturbations, and assume for simplicity of exposition, that the critical interval $I_i = [s_i, f_i]$ is bounded for all $i \in M$. Then, the convex hull of the critical intervals is expressed by the inequalities

$$\sum_{i \in M} (v_i / \omega_i) \leq 1,$$

where $\omega_i \in \{s_i, f_i\}$, for all $i \in M$ [2].

Coming back to our example, the ordinary sensitivity analysis of the right-hand side coefficients has led to the intervals $I_1 = [s_1, f_1] = [-4, 8]$ and $I_2 = [s_2, f_2] = [-8, 12]$. The convex hull of these two intervals, that is the 100% region, is then described by the following inequalities:

$$\begin{aligned} v_1/(-4) + v_2/(-8) &\leq 1, \\ v_1/(-4) + v_2/12 &\leq 1, \\ v_1/8 + v_2/(-8) &\leq 1, \\ v_1/8 + v_2/12 &\leq 1. \end{aligned}$$

(Fig. 4). The description of the 100% region associated with the objective coefficients can be derived similarly.

In general, we have a set difficult to interpret: the description of the 100% region requires 2^m inequalities in case of right-hand side perturbations, 2^n in case of objective perturbations. However, the idea is to generate only the inequalities that are strictly required. For example, if we are interested only in perturbations that increase the right-hand side coefficients, we need to check only one inequality, namely $\sum_{i \in M} (v_i / f_i) \leq 1$.

The Tolerance Approach

The *tolerance approach* focuses on simultaneous and independent variations of the right-hand side and objective coefficients. Independent means that for every b or c coefficient an interval of allowable perturbation is specified, and every entry of b and c is free to vary inside its interval. In its

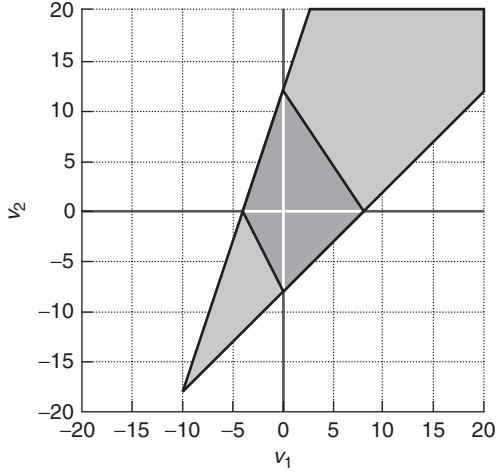


Figure 4. The 100% rule: region of allowed right-hand side perturbations (dark polygon).

most simple form, the tolerance approach returns a unique numerical value (*tolerance*) representing the maximum additive perturbation which can be applied on each b or c coefficient without affecting the optimality of the given basis. In this sense, the tolerance approach tries to resolve the trade-off between allowing multidimensional perturbations and giving clear results.

Consider right-hand side perturbations. We want to specify a unique interval $[-\rho, +\rho]$, such that if $v_i \in [-\rho, +\rho]$ for all $i \in M$, then basis B remains primal feasible (and thus optimal), and we want to get the largest possible interval. Geometrically, we look for the largest hypercube centered in the origin and contained in the primal basic cone C_P . Let an *allowable right-hand side tolerance* be a nonnegative value ρ such that

$$\{v \in \mathbb{R}^m : -\rho \leq v_i \leq +\rho \text{ for all } i \in M\} \subseteq C_P.$$

The *maximum right-hand side tolerance* ρ^* is the largest allowable right-hand side tolerance. Let again α_{ki} denote the entry of A_B^{-1} on the row indexed by $k \in B$ and the column indexed by $i \in M$. For all $k \in B$, let

$$\rho_k = p_k / \sum_{i \in M} |\alpha_{ki}| = \sum_{i \in M} \alpha_{ki} b_i / \sum_{i \in M} |\alpha_{ki}|.$$

It can be proved that [3]

$$\rho^* = \min\{\rho_k : k \in B\}.$$

Returning to our example, we have

$$\begin{aligned} \rho^* &= \min\{6/(3/2+1/2); 4/(1/2+1/2)\} \\ &= \min\{3; 4\} = 3. \end{aligned}$$

Thus, the two right-hand side perturbations v_1 and v_2 may vary simultaneously and independently in the same interval $[-3, +3]$ without affecting the primal feasibility, and hence the optimality, of the current basis. The corresponding tolerance region is depicted in Fig. 5 (dark square). A value of ρ larger than 3 would result in a tolerance region not contained in the primal basic cone C_P . However, if we relax the requirement that each coefficient has the same interval and that each interval is centered in the origin, we may obtain a larger region.

More precisely, we can define *allowable individual right-hand side tolerances* as non-negative scalars ρ^-_i and ρ^+_i such that if $-\rho^-_i \leq v_i \leq +\rho^+_i$ for all $i \in M$, then the current basis remains optimal. Let the values $\rho_k (k \in B)$ be computed as before. It can be

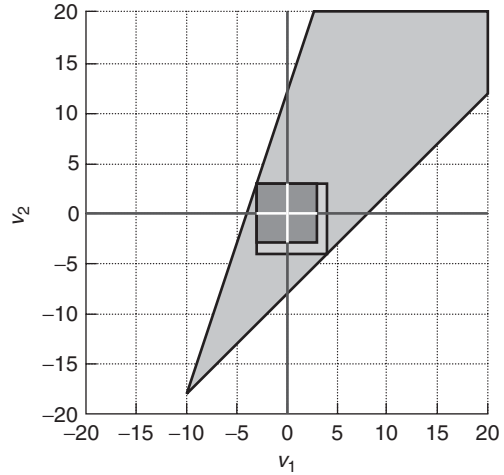


Figure 5. Tolerance approach: region of allowed right-hand side perturbations with unique tolerance (dark square) and with individual tolerances (light grey rectangle containing the dark square).

proved [4] that the following, easily computable values are allowable individual right-hand side tolerances:

$$\begin{aligned}\rho^{*-}_i &= \min\{\rho_k : k \in B; \alpha_{ik} < 0\}, \\ \rho^{*+}_i &= \min\{\rho_k : k \in B; \alpha_{ik} > 0\} \\ &\text{for all } i \in M.\end{aligned}$$

Since all ρ^{*-}_i and ρ^{*+}_i are minima over a subset of the ρ_k , whereas ρ^* is the minimum over all the ρ_k , we have $\rho^{*-}_i \geq \rho^*$ and $\rho^{*+}_i \geq \rho^*$ for all $i \in M$, so that

$$\begin{aligned}\{v \in \mathbb{R}^m : -\rho^{*-}_i \leq v_i \leq +\rho^{*+}_i \text{ for all } i \in M\} \supseteq \\ \{v \in \mathbb{R}^m : -\rho^* \leq v_i \leq +\rho^* \text{ for all } i \in M\}.\end{aligned}$$

In our example, it turns out that $\rho^{*-}_1 = \rho^{*+}_2 = \rho_1$ and $\rho^{*+}_1 = \rho^{*-}_2 = \rho_2$. The corresponding tolerance region is depicted in Fig. 5 (light gray rectangle containing the dark square).

Analogous tolerances can be computed for objective vector perturbations [3,4].

DEGENERATE CASE

When primal and/or dual degeneracy occurs in the optimal basis, then sensitivity analysis becomes harder to perform and to interpret correctly. We first analyze the difficulties related to degeneracy and then sketch some possible solutions.

Primal Degeneracy

Primal degeneracy usually implies that the same optimal solution can be associated with different optimal bases and to different dual solutions (see **Degeneracy and Variable Entering/Exiting Rules**). Geometrically, this means that the null right-hand side perturbation vector, $v = 0$, lies on the boundary of different basic cones, and that in each basic cone, the optimal value can be a different function of the perturbations. As a consequence, the optimal value function $z(v, 0)$ is not differentiable with respect to v . To clarify this point, let us modify the

right-hand side vector of the toy example considered so far:

$$\begin{aligned}\max \quad & 4x_1 + 5x_2 \\ \text{subject to} \quad & x_1 + x_2 + x_3 = 10 \\ & x_1 + 3x_2 + x_4 = 30 \\ & x_1, x_2, x_3, x_4 \geq 0.\end{aligned}$$

With the new right-hand side vector, the unique optimal solution becomes $p^* = (0, 10, 0, 0)^T$ with optimal value $z^* = 50$. The simplex method could correctly certify that p^* is associated with basis $B = \{1, 2\}$. As a consequence, the obtained dual solution is $d = (7/2, 1/2)^T$ and the obtained primal basic cone is

$$\begin{aligned}C_P = \{v \in \mathbb{R}^2 : -(3/2)v_1 + (1/2)v_2 \leq 0; \\ (1/2)v_1 - (1/2)v_2 \leq 10\}.\end{aligned}$$

As explained above, if $v \in C_P$ then B remains optimal and the optimal value is expressed as

$$z(v, 0) = z^* + d^T v = 50 + (7/2)v_1 + (1/2)v_2.$$

In particular, ordinary sensitivity analysis returns two intervals, $I_1 = [0, 20]$ and $I_2 = [-20, 0]$, that should be interpreted as follows: if $v_2 = 0$, then the slope of the objective value is 7/2 as long as $v_1 \in I_1$; if $v_1 = 0$, then the slope of the objective value is 1/2 as long as $v_2 \in I_2$. This is correct, but it is only a partial truth. Indeed, due to primal degeneracy, p^* is also associated with basis $B' = \{4, 2\}$. If we consider B' , we have the dual solution $d' = (5, 0)^T$ and the primal basic cone

$$C'_P = \{v \in \mathbb{R}^2 : 3v_1 - v_2 \leq 0; -v_1 \leq 10\}.$$

If $v \in C'_P$, then B' remains optimal and the optimal value is expressed as

$$z(v, 0) = z^* + (d')^T v = 50 + 5v_1.$$

In particular, ordinary sensitivity analysis returns other two intervals, $I'_1 = [-10, 0]$ and $I'_2 = [0, \infty)$, that should be interpreted as follows: if $v_2 = 0$, then the slope of the objective value is 5 as long as $v_1 \in I'_1$; if $v_1 = 0$, then

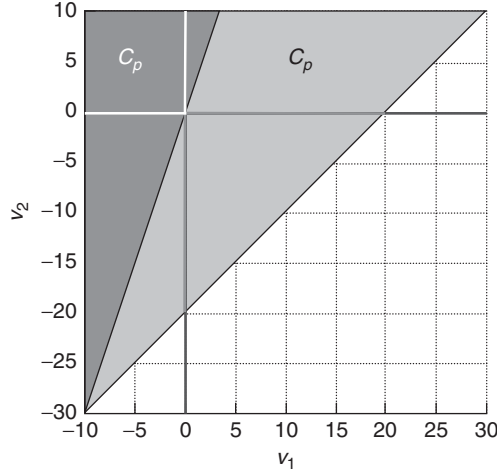


Figure 6. Primal degeneracy: the origin of the right-hand side perturbation space is on the boundary of different primal basic cones.

the slope of the objective value is 0 as long as $v_2 \in I_2'$ (Fig. 6).

Thus, in order to get a correct sensitivity analysis, we have to distinguish between positive and negative perturbations. For example, if we increase the first right-hand side coefficient, then interval I_1 applies, and the objective function increases with a rate of 7/2; this rate is sometimes called *right-hand shadow price*. If we decrease the first right-hand side coefficient, then I_1' applies, and the objective function decreases with a rate of 5; this rate is sometimes called *left-hand shadow price*.

Dual Degeneracy

Dual degeneracy usually implies multiple primal optimal solutions, all associated with the same dual solution, that is, with the same slope of the objective function value (see **Degeneracy and Variable Entering/Exiting Rules**). Geometrically, this means that there are multiple and overlapping primal basic cones containing the origin, and that the optimal value function has constant slope on the union of all these cones. To clarify this point, let us modify the objective vector of our toy example:

$$\begin{aligned} \max \quad & 8x_1 + 10x_2 + 7x_3 + x_4 \\ \text{subject to} \quad & x_1 + x_2 + x_3 = 10 \\ & x_1 + 3x_2 + x_4 = 18 \\ & x_1, x_2, x_3, x_4 \geq 0. \end{aligned}$$

One can check that the above model has four distinct primal nondegenerate optimal bases $B(h)$, with $h = 1, 2, 3, 4$. These bases and the corresponding primal solutions are listed below.

h	$B(h)$	p_1^h	p_2^h	p_3^h	p_4^h
1	{1,2}	6	4	0	0
2	{1,4}	10	0	0	8
3	{3,2}	0	6	4	0
4	{3,4}	0	0	10	18

We have

$$\begin{aligned} d^{(h)} &= (c_{B(h)}^T A_{B(h)}^{-1})^T = (7, 1)^T \quad \text{and} \\ c_{N(h)}^T - d^{(h)T} A_{N(h)} &= 0, \end{aligned}$$

for all $h = 1, 2, 3, 4$. Thus, all bases are dual degenerate and share the same dual solution. The optimal objective value is $z^* = 88$. The corresponding four primal basic cones are

$$\begin{aligned} C_P^{(1)} &= \{v \in \mathbb{R}^2 : -(3/2)v_1 + (1/2)v_2 \leq 6; \\ &\quad (1/2)v_1 - (1/2)v_2 \leq 4\}, \\ C_P^{(2)} &= \{v \in \mathbb{R}^2 : -v_1 \leq 10; v_1 - v_2 \leq 8\}, \\ C_P^{(3)} &= \{v \in \mathbb{R}^2 : -v_1 + (1/3)v_2 \leq 4; \\ &\quad -(1/3)v_2 \leq 6\}, \\ C_P^{(4)} &= \{v \in \mathbb{R}^2 : -v_1 \leq 10; -v_2 \leq 18\}. \end{aligned}$$

They are depicted in Fig. 7: notice that the origin is in the interior of every cone. Consider ordinary sensitivity analysis on the first right-hand side coefficient. Depending on the basis returned by the simplex method, four different intervals are obtained:

$$\begin{aligned} I_1^{(1)} &= [-4, 8], I_1^{(2)} = [-10, 8], \\ I_1^{(3)} &= [-4, \infty), I_1^{(4)} = [-10, \infty). \end{aligned}$$

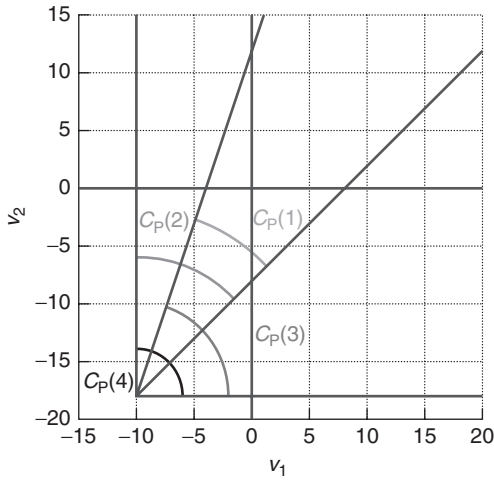


Figure 7. Dual degeneracy: the origin of the right-hand side perturbation space is in the interior of different, overlapping primal basic cones.

Notice that in this case, every interval corresponds to the perturbation of a different primal optimal solution, but the slope of the optimal objective function is always the same.

Some Guidelines

In general, under primal and/or dual degeneracy, the analysis developed for the nondegenerate case is still valid, but the interpretation of its results needs special attention. The following guidelines are useful for a correct interpretation [5].

- The optimal basis is not necessarily unique; the sensitivity analysis results depend on the particular basis used.
- The dual solution gives the rate of change in the objective value if the right-hand side vector is changed within the primal basic cone; the dual solution remains optimal in this cone.
- The primal solution gives the rate of change in the objective value if the objective vector is changed within the dual basic cone; the primal solution remains optimal in this cone.

Such guidelines should be taken into account for ordinary sensitivity analysis,

100% rule, and the tolerance approach, since all return subsets of the basic cones. A particular remark is required by the tolerance approach. Consider again, right-hand side perturbations in the example of primal degeneracy considered above. Recall that the tolerance approach looks for the largest cube centered in the origin and contained in the primal basic cone. It is easy to check that, for both optimal bases B and B' , this cube reduces to the origin itself, thus $\rho^* = 0$. This is true whenever primal degeneracy occurs. Hence, individual tolerances should always be used.

Using Optimal Values

When ordinary sensitivity analysis, that is perturbation in a single coefficient, is considered, then degeneracy issues may be faced more satisfactorily.

Let d be the dual solution associated with the returned optimal basis B . We know that d_i is the slope of the optimal value function in the interval I_i , for all $i \in M$. If dual degeneracy occurs, the slope could remain d_i also on a wider interval. We may identify the widest interval by solving two auxiliary linear programs. Let

$$\begin{aligned} s &= \min_{\beta, x} \{ \beta : Ax - \beta e^{(i)} = b, x \geq 0, \\ &\quad c^T x = z + d_i \beta \}, \\ f &= \max_{\beta, x} \{ \beta : Ax - \beta e^{(i)} = b, x \geq 0, \\ &\quad c^T x = z + d_i \beta \}. \end{aligned}$$

By construction, the slope of the optimal objective value is d_i if and only if $s \leq \beta \leq f$.

Consider now a case of primal degeneracy. Since vector b lies on the boundary of different primal basic cones, the dual optimal solution, and thus the rate of change of the optimal objective value, may change depending on the sign and direction of the perturbation. We may identify the correct dual solutions to use by solving two auxiliary linear programs. Let

$$\begin{aligned} \delta^- &= \min_y \{ y_i : y^T A \geq c^T, y^T b = d^T b \}, \\ \delta^+ &= \max_y \{ y_i : y^T A \geq c^T, y^T b = d^T b \}. \end{aligned}$$

It can be proved [6] that in case of ordinary sensitivity analysis with respect to the i th entry of the right-hand side vector, the correct slopes of the optimal objective value for negative and positive perturbations are respectively, δ^- and δ^+ .

FURTHER READING

Key references on sensitivity analysis are the book by Gal [7] and the succeeding one edited by Gal and Greenberg [8]. An excellent survey on different approaches to sensitivity analysis in linear programming is given by Ward and Wendell [9].

The tolerance approach has been proposed by Wendell [3]. The original idea has been successively extended to exploit *a priori* information in order to enlarge the tolerance intervals and to admit individual tolerances. Other extensions and applications are cited in Ref. 7. The use of tolerance regions beyond the basic cones is considered in Ref. 9.

Often, textbooks do not stress adequately the interpretation problems connected with degeneracy. This fact has been pointed out by some authors, for example, Ref. 10.

Let X^* be the set of optimal solutions of problem (P) . Consider the following sets:

$$\mathcal{B} = \{j \in E : x_j > 0 \text{ for some } x \in X^*\},$$

$$\mathcal{N} = \{j \in E : x_j = 0 \text{ for all } x \in X^*\}.$$

Sets $\mathcal{B}$ and $\mathcal{N}$ form a partition of the index set E , called *optimal partition*. Under mild hypotheses, path-following interior point methods for linear programming return an optimal solution x^* such that

$$\{j \in E : x_j^* > 0\} = \mathcal{B}$$

(see **Central Path and Barrier Algorithms for Linear Optimization**). Starting from this fact, a stream of research has originated on sensitivity analysis founded on the stability of optimal partitions of variables instead of optimal bases [11,12]. The main advantage of optimal partitions is that their analysis is not affected by degeneracy; cases where optimal partitions are useful in practice are illustrated in Ref. 13.

A related approach to sensitivity analysis is founded on the stability of the set of *active constraints*, that is, the set of inequality constraints that remain satisfied as equality in an optimal solution. See Ref. 14 and references therein.

Sensitivity analysis with respect to the matrix A of constraint coefficient is much less studied. We mention again Refs 7 and 8 and the references therein. An approach using optimal partitions is studied in Ref. 15.

Finally, a criticism on the use of sensitivity analysis to couple uncertainty on the parameters of a deterministic linear program is suggested in Ref. 16 and further elaborated in Ref. 17.

REFERENCES

1. Schrijver A. Theory of linear and integer programming. Chichester: Wiley Interscience; 1986.
2. Bradley SP, Hax AC, Magnanti TL. Applied mathematical programming. Reading (MA): Addison-Wesley; 1977.
3. Wendell RE. The tolerance approach to sensitivity analysis in linear programming. *Manage Sci* 1985;31(5):564–578.
4. Wendell RE. Sensitivity analysis revisited and extended. *Decis Sci* 1992;23(5):1127–1142.
5. Roos C, Terlaky T, Vial J-P. Interior point approach to linear optimization: theory and algorithms. New York: John Wiley & Sons, Inc.; 1997.
6. Akgül M. A note on shadow prices in linear programming. *J Oper Res Soc* 1984;35(5):425–431.
7. Gal T. Postoptimal analyses, parametric programming, and related topics. 2nd ed. Berlin: Walter de Gruyter; 1995.
8. Gal T, Greenberg HJ, editors. Advances in sensitivity analysis and parametric programming. Boston (MA): Kluwer Academic Publisher; 1977.
9. Ward JE, Wendell RE. Approaches to sensitivity analysis in linear programming. *Ann Oper Res* 1990;27(1–4):3–38.
10. Jansen B, de Jong JJ, Roos C, *et al.* Sensitivity analysis in linear programming: just be careful! *Eur J Oper Res* 1997;101(1):15–28.
11. Adler J, Monteiro RDC. A geometric view of parametric programming. *Algorithmica* 1992;8:161–176.

12. Greenberg HJ. Simultaneous primal-dual right-hand-side sensitivity analysis from a strictly complementary solution of a linear program. *SIAM J Optim* 2000;10(2):427–442.
13. Greenberg HJ. The use of optimal partition in a linear programming solution for postoptimal analysis. *Oper Res Lett* 1994;15:179–185.
14. Ghaffari Hadigheh A, Terlaky T. Sensitivity analysis in linear optimization: Invariant support set intervals. *Eur J Oper Res* 2006;169(3):1158–1175.
15. Greenberg HJ. Matrix sensitivity analysis from an interior solution of a linear program. *INFORMS J Comput* 1999;11(3):316–327.
16. Wallace SW. Decision making under uncertainty: is sensitivity analysis of any use? *Oper Res* 2000;48(1):20–25.
17. Hige JL, Wallace SW. Sensitivity analysis and uncertainty in linear programming. *Interfaces* 2003;33(4):53–60.

SENSITIVITY ANALYSIS OF SIMULATION MODELS

JACK P.C. KLEIJNEN
Department of Information
Management/Center, Tilburg University,
Tilburg, Netherlands

Sensitivity analysis of simulation models quantifies the change in the simulation output as the simulation input changes. Such an analysis may have the following goals:

- verification and validation (V & V)
- optimization
- risk or uncertainty analysis and robustness analysis.

Notice that this contribution will provide many synonyms because the terminology differs among the many scientific disciplines in which simulation is applied [1].

The simulation *output* (say) w may be either a single response (a scalar or univariate variable) or multiple responses (a vector or multivariate variable). This contribution assumes that in case of multiple outputs the sensitivity analysis is applied per output. In linear regression analysis (see the section titled “Classic Designs and Metamodels” below), ordinary least squares (OLS) ignoring the correlations among the multiple outputs suffices (no generalized least squares (GLS) needed) [2, p. 703]. In Kriging (Gaussian process modeling), however, it may be better not to ignore these correlations; see the section titled “Kriging Metamodels and Designs.”

The simulation *input* consists of (say) k input variables x_j with $j = 1, \dots, k$. These inputs may be further classified in different ways. From a *managerial* point of view, Taguchi [3] distinguishes between decision or control variables (e.g., number of machines of each type in manufacturing facility design; see also the section titled “Manufacturing Applications” in this encyclopedia) and environmental or noise factors (e.g., machine

failure rates). Inputs are called *factors* in the statistical theory on design of experiments (DOE). Realistic simulation models have multiple inputs so $k > 1$.

From a *mathematical* point of view, the inputs x_j may be divided into qualitative (nominal, categorical) and quantitative inputs; the quantitative inputs may be further divided into integer (using either an absolute or an ordinal scale) and continuous (interval or ratio scales) [4, p. 11]. The (say) $k' \leq k$ quantitative inputs may be varied either locally or globally (global sensitivity analysis is also called *What If* analysis). Local changes enable the estimation of the *gradient*, which is the vector with the first-order partial derivatives (say) $\nabla(w) = (\partial w/\partial x_1, \dots, \partial w/\partial x_{k'})$, where $w(x_1, \dots, x_{k'})$ denotes the Input/Output (I/O) function (or mapping) implicitly defined by the given simulation model (computer program or code). Gradients play an important role in simulation optimization (see the sections titled “Simulation Optimization” and “Stochastic Approximation Methods” and also the article ***Simulation Optimization in Risk Management*** in this encyclopedia). If the inputs, however, are integers (e.g., number of machines) or qualitative (e.g., priority rule), then the gradient is not defined.

The users may also be interested in output changes caused by *global* changes in an input; that is, an input changes from its minimum to its maximum in the *experimental area*. The selection of this area is not detailed in the DOE literature. This area is called the *experimental frame* in Ref. 5, emphasizing that the given simulation model may be valid for one area, but not for a different area; for example, the model may be valid for a small area concentrated around the current or “base” *scenario*—combination of factor values. The experimental area depends on the *goals* of the simulation study; that is, sensitivity analysis of a proposed completely new system has other requirements than optimization of the current system has. Various goals are discussed in Refs 6 and 7.

Sensitivity analysis is needed for any type of simulation model—be it discrete-event, continuous, or hybrid (see also *Simulation Optimization in Risk Management*), terminating or nonterminating (see *Introduction to Discrete-Event Simulation*). The sensitivity analysis in this contribution uses statistical methods (techniques) that have been developed for the design and analysis of the following three types of experiments:

- real-life (physical) systems
- deterministic simulation models
- random (stochastic) simulation models.

DOE for *real-life* systems is covered extensively by many textbooks [8,9]. DOE for *deterministic simulation* gained popularity with the increased use of “computer codes” for the design (in an engineering, not a statistical sense) of airplanes, automobiles, TV sets, chemical processes, computer chips, and so on—in computer aided engineering (CAE) and computer aided design (CAD). DOE in this domain is often called *design and analysis of computer experiments* (DACE). The classic article on DACE is Ref. 10; a classic textbook is Ref. 11. DOE for *random simulation* is the focus of several textbooks by Kleijnen; the most recent one is Ref. 4.

Note that deterministic simulation models become random if the exact input values are unknown. Hence, these values are sampled from statistical distribution functions. For example, the well-known economic order quantity (EOQ) model assumes that the demand rate is constant and known. In practice, however, the demand rate is unknown, but a distribution function of its possible values may be estimated from past data or from expert opinion. A sample may then be generated from this input distribution and fed into the EOQ model. This input sample is propagated through the EOQ model to generate a sample of values for the output (inventory cost). The resulting estimated density function (EDF) of the output may be analyzed to determine the appropriate decision variable, namely the EOQ. This EOQ example is detailed in Ref. 12. This is related to *risk analysis* discussed in the section titled “Risk Analysis” in this encyclopedia; see also Refs 4,

pp. 3, 123–130. Sensitivity analysis and risk analysis answer different questions. Sensitivity analysis answers the question: “Which are the most important factors in the simulation model of a given real system.” Risk analysis answers the question: “What is the probability of a given event happening: for example, a nuclear or a financial disaster?”

Sensitivity analysis using DOE perceives the simulation model as a *black box*: that is, only the simulation I/O data are used (it ignores the values of internal variables and specific functions implied by the simulation’s computer modules). *White-box* approaches look inside the simulation model (e.g., at service time generated by $-\ln(r)/\mu$) to estimate the gradient for local—not global—sensitivity analysis and for optimization; examples of such white-box approaches are perturbation analysis and the score function or likelihood ratio method. These latter methods require fewer simulation runs, but stricter mathematical assumptions so their application domain is more restricted. Perturbation analysis is detailed in Ref. 14, the score function in Ref. 15; see also Refs 16 and 17, the section titled “Simulation Optimization,” and the article titled *Simultaneous Perturbation and Finite Difference Methods* in this encyclopedia.

Sensitivity analysis through DOE implies a *metamodel* to select an experimental design and to analyze the experimental results; that is, *design* and *analysis* of experiments are like “chicken and egg.” The analysis uses—implicitly or explicitly—a metamodel (also called *response surface*, *surrogate*, *emulator*, etc.), which is an approximation of the I/O function implied by the underlying simulation model; that is, the experiment yields I/O data that are used to estimate this function.

There are different *types* of metamodels. The most popular type are polynomials of first or second order (degree); see the section titled “Classic Designs and Metamodels” below. Obviously, a first-order polynomial implies a gradient that remains constant over the whole experimental area. Another metamodel type are Kriging models, which enable the approximation of more general I/O functions; such a Kriging metamodel implies

a gradient that changes over the whole experimental area, as the section titled “Kriging Metamodels and Designs” will show. References to other metamodel types are given by Kleijnen [4, p. 8].

Different metamodels require different designs; for example, first-order polynomials require two values per factor, whereas Kriging requires many more values per factor—as the following sections will show. The adequacy of the design and metamodel combination depends on the *goal* of the experiment [6].

Thus, sensitivity analysis may also serve *optimization*. Besides the methods discussed in the section titled “Simulation Optimization” in this encyclopedia, there are methods closely related to the methods discussed in this contribution. The latter methods allow multiple random simulation outputs, selecting one output as the goal (objective) output while the other random simulation outputs must satisfy prespecified target values (thresholds). Generalized response surface methodology (GRSM) fits a series of local low-order polynomials to the simulation I/O data [18]. Kriging is combined with integer nonlinear programming in Ref. 19 (see also the section titled “Nonlinear Programming (NLP) and Global Optimization (GO)” in this encyclopedia). “Robust” simulation-optimization in the sense of Taguchi [3] is discussed in Ref. 12.

The *V & V* of simulation models may use expert knowledge; for example, it is well-known that the more customers arrive per time unit, the more their waiting times increase. However, this knowledge is qualitative; to obtain quantitative knowledge, a simulation model is developed. If the simulation model’s I/O behavior violates this qualitative knowledge, the model should be seriously questioned: are there programming and conceptual errors? A case study is the ecological simulation of CO₂ gas emissions in Ref. 20; the regression metamodel leads to the discovery of a computer error (one of the simulation modules should be split into two modules and called in the correct order). Another case study is the sonar simulation in Ref. 21, which consists of several modules. For some modules, the naval experts suggest that the

factors have specific signs; the corresponding estimates in the regression metamodel turn out to have these signs so the validity of these modules does not seem questionable.

Whereas in real-life experimentation it is not practical to investigate many factors, realistic simulation experiments may have hundreds of factors. Moreover, whereas in real-life experiments it is hard to vary a factor over many values, in simulation experiments this restriction does not apply [Latin hypercube sampling (LHS) is a design that has as many values per factor as it has combinations; see the section titled “Kriging Metamodels and Designs” below]; hence, a multitude of scenarios may be observed. Furthermore, simulation experiments are well-suited to the application of sequential designs instead of “one-shot” designs; see the sections titled “Factor Screening: Sequential Bifurcation” and “Kriging Metamodels and Designs” below. Thus, a change of mindset of the experimenter is necessary [22].

The following overview of the remainder of this contribution enables the readers to decide which sections they might wish to skip. Because of space restrictions this contribution aims at enabling the readers to understand the major problems and solutions in sensitivity analysis; for details they need to consult the references.

The section titled “Classic Designs and Metamodels” covers *classic* designs and their corresponding metamodels, namely, Resolution-III (R-III) designs including selected 2^{k-p} designs for first-order polynomials, Resolution-IV (R-IV) and Resolution-V (R-V) designs for two-factor interactions, and designs including central composite designs (CCDs) for second-degree polynomials. Compared with these designs, the traditional approach of changing one factor at a time is inferior; that is, the traditional approach does not enable the estimation of interactions among factors, and it gives less accurate estimators of the first-order effects of the factors [4, p. 33].

The section titled “Factor Screening: Sequential Bifurcation” reviews *factor screening*, focusing on sequential bifurcation.

Table 1. An One-Sixteenth Fractional Factorial Design for Seven Factors

Combination	1	2	3	4 = 1.2	5 = 1.3	6 = 2.3	7 = 1.2.3
1	−	−	−	+	+	+	−
2	+	−	−	−	−	+	+
3	−	+	−	−	+	−	+
4	+	+	−	+	−	−	−
5	−	−	+	+	−	−	+
6	+	−	+	−	+	−	−
7	−	+	+	−	−	+	−
8	+	+	+	+	+	+	+

Traditionally, simulation analysts experiment with only a few factors, assuming that these factors are the most important ones.

The section titled “Kriging Metamodels and Designs” reviews *Kriging* and its designs. Kriging metamodels are fitted to data that are obtained for larger experimental areas than the areas often used when fitting low-order polynomials; that is, Kriging models are *global* rather than local. Kriging is used for prediction; its final goals are sensitivity analysis and optimization.

The last section *summarizes* this contribution.

CLASSIC DESIGNS AND METAMODELS

Let $w(x_1, \dots, x_k)$ denote the *true* I/O function of the given simulation model. Sensitivity analysis using DOE approximates this $w(\cdot)$ through a relatively simple explicit function, which forms a *metamodel* of the underlying simulation model. An example metamodel is the *first-order polynomial* augmented with a residual term:

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \epsilon, \quad (1)$$

where ϵ denotes the *residual*, which is caused by the approximation error (lack-of-fit) of the metamodel plus the noise caused by the pseudo random numbers (PRNs) used in random simulation; β_j quantifies the first-order effect of factor j , which is called the *main* effect in DOE (especially in analysis of variance (ANOVA)); β_0 is the intercept; y is the

metamodel output or predictor of the true simulation output.

To estimate the coefficients of polynomial metamodels, many DOE textbooks give specific designs. To understand these designs, consider an example with $k = 7$ factors. To estimate the first-order polynomial metamodel, at least $k + 1 = 8$ combinations need to be simulated. Table 1 presents a so-called 2^{7-4}_{III} design, which should be read as follows. The columns with the symbols **1** through **7** give the values of factor j ($j = 1, \dots, 7$) in the experiment; **4 = 1.2** means that the value of factor 4 equals the product of the values of the factors 1 and 2 in the corresponding combination the symbol “−” means that the factor has the standardized value -1 —which corresponds with the lowest value in the original scale—and “+” means that the factor has the standardized value $+1$ —highest value in the original scale. For example, the element in row 1 and column **7** is -1 because $x_{1,7} = x_{1,1}x_{1,2}x_{1,3} = (-1) \times (-1) \times (-1)$.

Table 1 is an example of a *R-III design*, which is defined as a design that enables the unbiased estimation of the coefficients of the first-order polynomial—assuming that this polynomial is an adequate or “valid” approximation of the true I/O function. Table 1 specifies only 8 combinations; that is, this design requires the simulation of only a $2^{-4} = 1/16$ fraction of all $2^7 = 128$ combinations that a “full-factorial two-level” design would require.

A R-III design enables the *OLS* estimation of the coefficients of the first-order polynomial

in Equation (1):

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}, \quad (2)$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)'$ denotes the OLS estimator, $\mathbf{X}$ the $n \times q$ matrix of explanatory (independent, regression) variables; n the number of factor combinations; q the number of regression coefficients; the estimated intercept $\hat{\beta}_0$ corresponds with a column for the dummy variable (say) x_0 that has the constant value +1 in the experiment. OLS is the classic estimation method in linear regression analysis, assuming *white noise*; that is, the metamodel's residuals ϵ are normally (Gaussian), independently, and identically distributed (NIID) with zero mean.

Note that Equation (1) implies that the most important factor is the one with the maximum absolute value of the first-order effect. The estimated gradient is $(\hat{\beta}_1, \dots, \hat{\beta}_k)'$ if all k factors are continuous.

Next, the *validity* of the fitted first-order polynomial metamodel may be tested. The DOE literature presents an F lack-of-fit statistic, but this test assumes white noise. A more general procedure uses the fitted metamodel to predict the outcome of the given simulation model with an input combination that is not part of the R-III design. An example combination is any of the 2^7 combinations that is not listed in the rows of Table 1; another example is the combination that is the “center” of the experimental area, defined by $x_j = 0$ for all quantitative factors. A 2^{7-4} design is *saturated*; that is, $n = q$ (the OLS estimator defined in Equation (2) still exists). However, if $k = 6$, then Table 1 can still be used but one of the columns (e.g., the last column) is not used. In that case the design is not saturated: $n = 8 < q = 7$. Then the adequacy of the metamodel may be tested through *cross-validation*; that is, row i ($i = 1, \dots, n$) of the design is deleted; the metamodel is reestimated from the remaining $(n - 1)$ combinations; the reestimated metamodel is used to predict the expected simulation output $E(w_i)$; if the resulting prediction error is too big, then the metamodel is rejected [4, pp. 57–63].

If the fitted metamodel is found not to be valid, then an alternative metamodel may

be the first-order polynomial in Equation (1) augmented with *two-factor interactions* $\beta_{j,j'}$:

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^{k-1} \sum_{j'=j+1}^k \beta_{j,j'} x_j x_{j'} + \epsilon. \quad (3)$$

The number of effects now increases dramatically: $q = 1 + k + k(k - 1)/2$; for example, $k = 6$ implies $q = 22$. A *R-IV design* ensures that the estimated first-order effects $\hat{\beta}_j$ are *not biased* by the two-factor interactions $\beta_{j,j'}$. Such a design is easily constructed, as follows. Let $\mathbf{D}$ denote the R-III design; that is, replace the symbol “−” in Table 1 by −1, and “+” by 1 so the 8×7 design matrix $\mathbf{D}$ results. Adding $-\mathbf{D}$ to $\mathbf{D}$ gives a R-IV design (so the R-IV design doubles the number of combinations).

To get unbiased estimators of the *individual* interactions in Equation (3), a *R-V design* is needed. An example is a 2_V^{6-1} design, which can be found in any DOE textbook. Unfortunately, its $2^5 = 32$ combinations may take too much computer time if a simulation run is computationally expensive. In that case, it is better to use a saturated design; for example, a Rechtschaffner design, listed in Ref. 4, p. 49.

Simulation optimization may use a *second-degree polynomial* metamodel

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^k \sum_{j'=j}^k \beta_{j,j'} x_j x_{j'} + \epsilon, \quad (4)$$

so the number of effects further increases with k “purely quadratic” effects $\beta_{j,j}$ —assuming all k inputs are quantitative. To estimate all effects in Equation (4), a *CCD* augments the R-V design with the “central point” and $2k$ “axial points,” which change each factor one-at-a-time by $-c$ and c units where $c > 0$. Unfortunately, a CCD is rather wasteful in case of expensive simulation models, because it uses five values per factor (instead of the minimum three needed to estimate a second-order polynomial) and it is not saturated. Alternatives for CCDs are discussed in Refs 4 and 9.

Many simulation applications of these classic designs are provided by Ref. 4.

This reference also discusses multivariate simulation outputs, nonnormality of the output, variance heterogeneity, and common random numbers (CRN) (also see **Common Random Numbers**).

FACTOR SCREENING: SEQUENTIAL BIFURCATION

The *goal* of factor screening is to identify the really important factors among the many factors that may be varied in a simulation experiment. Realistic simulation models may have hundreds or thousands of factors. Unfortunately, many simulation analysts vary only a handful of factors, conjectured to be the most important ones. Screening is related to “sparse” effects, “parsimony,” “Pareto principle,” “Occam’s razor,” “20–80 rule,” and “curse of dimensionality.”

The need for screening is illustrated by the following two case studies. In Ref. 23, a greenhouse deterministic simulation model has 281 factors; Dutch politicians, however, want to reduce the release of CO₂ gasses, starting with legislation for a few factors only. In Ref. 24, a discrete-event simulation model of a supply chain of the Ericsson company in Sweden has 92 factors; screening identifies a shortlist with 10 factors after simulating only 19 combinations.

There are several *types* of screening designs. R-III designs (already discussed in the section titled “Classic Designs and Metamodels”) are called *screening designs* by some authors [25]. Actually, simulation experiments proceed sequentially (ignoring parallel computers), so it is possible to apply designs that sequentially select factor combinations to be simulated. Sequential bifurcation is an efficient design—it is *super-saturated*; that is, $n < k$ —and it is effective if the following metamodel *assumptions* are satisfied (see also [26]):

1. a first-order polynomial augmented with two-factor interactions—see Equation (3)—is an adequate approximation;
2. all its first-order effects β_j have known signs and are nonnegative;
3. it shows “strong heredity”; that is, if a factor has no important main effect (e.g., $\beta_1 = 0$), then this factor does not interact with any other factor ($\beta_{1:j} = 0$) [27].

The role of these assumptions may be demonstrated as follows. The first step of sequential bifurcation aggregates all factors into a single group, and tests whether or not that group of factors has an important effect; that is, combination 1 has $x_j = -1$ ($j = 1, \dots, k$) and combination 2 has $x_j = 1$. Substitution of these x values into Equation (3) proves that *no cancellation* of effects occurs, given assumption 2. If that group indeed has an important effect—which is very likely in the first step—then the second step splits the group into two subgroups—it bifurcates—and tests each of these subgroups for importance. In the next steps, sequential bifurcation splits important subgroups into smaller subgroups, and discards unimportant subgroups. In the final step, all individual factors that are not in subgroups identified as unimportant, are estimated and tested. Details are given in Ref. 28, including references to alternative sequential screening methods.

KRIGING METAMODELS AND DESIGNS

Originally, the South African mining engineer Danie Krige developed Kriging (also spelled as *kriging*) for interpolation in geo-statistical or spatial sampling; see the classic Kriging textbook [29]. Later on, Kriging was applied to the I/O data of deterministic simulation models [10,11]. Only recently, Kriging has also been applied to random simulation models; see the references below.

So-called *Ordinary* Kriging uses the *linear* predictor

$$y = \lambda' \mathbf{w}, \quad (5)$$

where $\mathbf{w}$ denotes the vector with the n “old” simulation outputs (assuming no replicates, which is indeed the case for deterministic simulation) and λ denotes the vector with Kriging weights. Under the assumption that

the outputs form a stationary covariance process, the *optimal Kriging weights* are

$$\lambda_o = \Gamma^{-1} \left(\gamma + 1 \frac{1 - \mathbf{1}' \Gamma^{-1} \gamma}{\mathbf{1}' \Gamma^{-1} \mathbf{1}} \right), \quad (6)$$

where $\gamma = (\text{cov}(w_i, w_{i'}))$ with $i, i' = 1, \dots, n$ denotes the $n \times n$ matrix with the covariances between the n old outputs, and $\gamma = (\text{cov}(w_i, w_{n+1}))$ denotes the n -dimensional vector with the covariances between the n old outputs w_i and the “new” output w_{n+1} to be predicted. Obviously, γ varies with the location of this new output, so these weights λ_o are not constants (the regression parameters β are). Kriging assumes that outputs are more correlated, the closer their inputs are. Actually, the most popular type of *correlation function* in Kriging of simulation I/O data is

$$\exp \left(- \sum_{j=1}^k \theta_j h_j^{p_j} \right) = \prod_{j=1}^k \exp \left(- \theta_j h_j^{p_j} \right), \quad (7)$$

where h_j denotes the Euclidean distance between the j th input x_j ($j = 1, \dots, k$) of the input combinations $\mathbf{x}_g$ and $\mathbf{x}_{g'}$ ($g, g' = 1, \dots, n+1$); θ_j denotes the importance of input j (the higher θ_j is, the less effect input j has), and p_j denotes the smoothness of the correlation function (e.g., $p_j = 2$ implies an infinitely differentiable function); $p = 1$ and $p = 2$ give the so-called *exponential* and *Gaussian correlation* functions, respectively. So, the correlation function (Eq. 7) implies that the weights (Eq. 6) decrease with the *distance* between the new input combination to be predicted $\mathbf{x}_{n+1}$ and the n old combinations $\mathbf{x}_i$.

The resulting Kriging predictor is an *exact interpolator*; that is, the Kriging predictions equal the outputs observed for the n old input combinations ($y_i = w_i$; $i = 1, \dots, n$). This is an attractive property in deterministic simulation, but not in random simulation. Only recently, Kriging has been investigated for random simulation models; the oldest publication seems [30]. Furthermore, Ref. 11, pp. 215–249 accounts for the so-called *nugget effect* or *measurement error* such that the Kriging predictor does not interpolate the n

old outputs. Recently, Ankenman *et al.* [31] introduced Kriging for random simulation, accounting for variance heterogeneity (e.g., the variance of the waiting time increases as the traffic rate in a queueing simulation increases). A similar Kriging model is developed by Yin *et al.* [32], speaking of the “modified nugget effect.” The Kriging predictors that account for nugget effects do not interpolate $\bar{w}_i$, the outputs averaged over the (say) $m_i \geq 1$ ($i = 1, \dots, n$) replicates per input combination. Finally, Ankenman [31] accounts for possible correlations between the outputs for different input combinations caused by CRN.

It can be shown that the Kriging meta-model resulting from Equations (5) and (6) implies a specific *gradient*; see Ref. 4, p. 156 (software provides these gradients; see next paragraph).

A major problem is that *the correlation function is unknown*, so both the type and the parameter values must be estimated; see again Equation (7). To estimate the parameters, the standard Kriging literature and software uses maximum likelihood estimators (MLEs). The estimation of the correlation functions, the corresponding optimal Kriging weights, the Kriging predictor, and the corresponding gradient in deterministic simulation can all be done through DACE, which is Matlab software that is well documented and free of charge [33]. Unfortunately, for random simulation there is no well-documented software yet; so, in the mean time, Kriging might apply DACE to the averages $\bar{w}_i$. For example, van Beers and Kleijnen [30] uses DACE with these averages, and reports Kriging predictions that are much better than regression predictions.

The Kriging literature virtually ignores problems caused by replacing the optimal weights λ in Equation (5) by the estimated optimal weights (say) $\hat{\lambda}_0$; see Equation (6) or a related formula that accounts for nugget effects and CRN. These estimated weights make the Kriging predictor a *nonlinear* estimator. Consequently, the literature underestimates the true variance of the Kriging predictor; moreover, this estimated variance

and the true variance do not reach their maxima for the same input combination, which is important in sequential designs [34].

There is a simple solution to this variance estimation problem, in random simulation if each input combination is replicated a number of times; this solution is *distribution-free bootstrapping*. The basics of bootstrapping are explained in Ref. 35. In Kriging, the bootstrap resamples with replacement the m_i replicates for combination i ($i = 1, \dots, n$). This sampling gives the bootstrapped average $\bar{w}_i^*$ where the superscript $*$ is the usual symbol to denote a bootstrapped observation. These n bootstrapped averages $\bar{w}_i^*$ give the bootstrapped estimated weights λ_0^* using a formula like Equation (6) with $\hat{\Gamma}^* = (\widehat{cov}(\bar{w}_i^*, \bar{w}_{i'}^*))$ with $i, i' = 1, \dots, n$ and $\hat{\gamma}^* = (\widehat{cov}(\bar{w}_i^*, \bar{w}_{n+1}^*))$. Then, Equation (5) gives the bootstrapped Kriging predictor $y^* = \lambda_0^* \bar{\mathbf{w}}^*$. This predictor gives the bootstrapped gradient $\nabla(w)^*$. To decrease sampling effects, this whole procedure is repeated B times (e.g., $B = 100$). This enables estimation of the mean and variance of the Kriging predictor and its gradient [36].

To obtain the Kriging I/O simulation data, experimenters often use LHS. LHS gives a space-filling design that is *noncollapsing*; that is, projection of a point in the k -dimensional input space onto one of the k axes gives a uniform spread on that axis. Different types of LHS and other types of space-filling designs (including references and software) are discussed in Ref. 4, p. 130.

Instead of a (one-shot) design like LHS, a *sequentialized* design may be used. In sensitivity analysis of simulation models, a sequential design implies that the simulation I/O data are analyzed—so the data generating process is better understood—*before* the next input combination is selected. This property implies that the design depends on the specific underlying simulation model; that is, the design is customized (tailored or application-driven, not generic). Moreover, simulation proceeds sequentially (again ignoring parallel computers) so sequential designs are attractive. The following sequential design for Kriging in sensitivity analysis is developed in Ref. 36:

1. Start with a *pilot* experiment, using some small generic space-filling design (e.g., LHS).
2. Fit a *Kriging* metamodel to the simulation I/O data that are available at this step (in the first pass of this procedure, these data are the data resulting from step 1).
3. Consider (but do not yet simulate) a set of *candidate* input combinations that have not yet been simulated and that are selected through some space-filling design; select as the next combination to be actually simulated, the candidate combination that has the *highest predictor variance* (estimated through bootstrapping).
4. Use the combination selected in step 3 as the input combination to be simulated, and obtain the corresponding simulation output.
5. Refit a Kriging model to the I/O data augmented with the data from step 4.
6. Return to step 3, until the Kriging metamodel is acceptable for its goal, sensitivity analysis.

The resulting design is indeed *customized*; for example, if the true (unknown) I/O function is a simple hyperplane within a subspace of the total experimental area, then this design selects relatively few points in that part of the input space. (A sequential design for constrained optimization—instead of sensitivity analysis—is presented in Ref. 19.)

Nearly all Kriging publications and software assume univariate output, whereas in practice simulation models have *multivariate output*; the latter output is discussed in Ref. 11 [pp. 101–116], 37 and 38. In practice, univariate Kriging is applied per type of simulation output.

Often the analysts know that the simulation's I/O function has certain properties, for example, as the traffic rate increases, the expected waiting time increases monotonically. *Monotonicity-preserving* bootstrapped Kriging is discussed in Ref. 39.

SUMMARY

This contribution gave an overview of sensitivity analysis through the statistical design and analysis of experiments, treating the simulation model as a black box.

It covered classic designs and their analysis through corresponding metamodels; namely, R-III designs for first-order polynomials, R-IV and R-V designs for two-factor interactions, and designs such as CCD for second-degree polynomials.

It also reviewed factor screening for simulation experiments with very many factors, focusing on sequential bifurcation.

Furthermore, it reviewed Kriging and its designs—especially LHS and sequential designs.

Now the challenge is to apply these methods to simulation in practice.

Acknowledgment

I thank the anonymous referee and the Topical Editor for their comments that has led to a major revision of the original version.

REFERENCES

1. Karplus WJ. The spectrum of mathematical models. *Perspect Comput* 1983;3(2):4–13.
2. Ruud PA. An introduction to classical econometric theory. New York: Oxford University Press; 2000.
3. Taguchi G. Volumes 1 and 2, System of experimental designs. White Plains (NY): UNIPUB/Krauss International; 1987.
4. Kleijnen JPC. Design and analysis of simulation experiments. New York: Springer; 2008 (forthcoming Chinese translation: Publishing House of Electronics Industry; 2010).
5. Zeigler BP, Praehofer H, Kim TG. Theory of modeling and simulation. 2nd ed. San Diego (CA): Academic Press; 2000.
6. Kleijnen JPC, Sargent RG. A methodology for the fitting and validation of metamodels in simulation. *Eur J Oper Res* 2000;120(1):14–29.
7. Law AM. Simulation modeling and analysis. 4th ed. Boston (MA): McGraw-Hill; 2007.
8. Montgomery DC. Design and analysis of experiments. 7th ed. Hoboken (NJ): Wiley; 2009.
9. Myers RH, Montgomery DC, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. 3rd ed. New York: Wiley; 2009.
10. Sacks J, Welch WJ, Mitchell TJ, *et al.* Design and analysis of computer experiments (includes Comments and Rejoinder). *Stat Sci* 1989;4(4):409–435.
11. Santner TJ, Williams BJ, Notz WI. The design and analysis of computer experiments. New York: Springer; 2003.
12. Dellino G, Kleijnen JPC, Meloni C. Robust simulation-optimization using metamodels. In: Rossini MD, Hill RR, Johansson B, *et al.* Proceedings of the 2009 Winter Simulation Conference. New York: Association for Computing Machinery; 2009. pp. 540–550.
13. Storlie CB, Swiler LP, Helton JC, *et al.* Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliab Eng Syst Saf* 2009;94(11):1735–1763. Accepted (also Sandia report SAND report 2008-6570).
14. Ho Y, Cao X. Perturbation analysis of discrete event dynamic systems. Dordrecht (Netherlands): Kluwer; 1991.
15. Rubinstein RY, Shapiro A. Discrete-event systems: sensitivity analysis and stochastic optimization via the score function method. New York: Wiley; 1993.
16. Fu MC. What you should know about simulation and derivatives. *Nav Res Log* 2008;55:723–736.
17. Spall JC. Introduction to stochastic search and optimization; estimation, simulation, and control. New York: Wiley; 2003.
18. Angün E, Kleijnen J, den Hertog D, *et al.* Response surface methodology with stochastic constraints for expensive simulation. *J Oper Res Soc* 2009;60:735–746.
19. Kleijnen JPC, van Beers W, van Nieuwenhuyse I. Constrained optimization in simulation: a novel approach. *Eur J Oper Res* 2010;202:164–174.
20. Van Ham G, Rotmans J, Kleijnen JPC. Techniques for sensitivity analysis of simulation models: a case study of the CO₂ greenhouse effect. *Simulation* 1992;58(6):410–417.
21. Kleijnen JPC. Case study: statistical validation of simulation models. *Eur J Oper Res* 1995;87(1):21–34.
22. Kleijnen JPC, Sanchez SM, Lucas TW, *et al.* State-of-the-art review: a user's guide to

- the brave new world of designing simulation experiments. *INFORMS J Comput* 2005;17(3):263–289.
23. Bettonvil B, Kleijnen JPC. Searching for important factors in simulation models with many factors: sequential bifurcation. *Eur J Oper Res* 1997;96:180–194.
 24. Kleijnen JPC, Bettonvil B, Persson F. Screening for the important factors in large discrete-event simulation: sequential bifurcation and its applications. In: Dean A, Lewis S, editors. *Screening: methods for experimentation in industry, drug discovery, and genetics*. New York: Springer; 2006. pp. 287–307.
 25. Yu H-F. Designing a screening experiment with a reciprocal Weibull degradation rate. *Comput Ind Eng* 2007;52(2):175–191.
 26. Xu J, Yang F, Wan H. Controlled sequential bifurcation for software reliability study. In: Henderson SG, Biller B, Hsieh M-H, *et al.* *Proceedings of the 2007 Winter Simulation Conference*. New York: 2007. pp. 281–288.
 27. Wu CFJ, Hamada M. *Experiments; planning, analysis, and parameter design optimization*. New York: Wiley; 2000.
 28. Kleijnen JPC. Factor screening in simulation experiments: review of sequential bifurcation. In: Alexopoulos C, Goldsman D, Wilson JR, editors. *Advancing the frontiers of simulation: a festschrift in honor of George S. Fishman*. New York: Springer; 2009. pp. 169–173.
 29. Cressie NAC. *Statistics for spatial data: revised edition*. New York: Wiley; 1993.
 30. Van Beers WCM, Kleijnen JPC. Kriging for interpolation in random simulation. *J Oper Res Soc* 2003;54(3):255–262.
 31. Ankenman B, Nelson B, Staum J. Stochastic kriging for simulation metamodeling. *Oper Res* 2010. In press.
 32. Yin J, Ng SH, Ng KM. A study on the effects of parameter estimation on Kriging model's prediction error in stochastic simulations. In: Rossini MD, Hill RR, Johansson B, *et al.* *Proceedings of the 2009 Winter Simulation Conference*. New York: 2009. pp. 674–685.
 33. Lophaven SN, Nielsen HB, Sondergaard J. DACE: a Matlab Kriging toolbox, version 2.0. Lyngby: IMM Technical University of Denmark; 2002.
 34. Den Hertog D, Kleijnen JPC, Siem AYD. The correct Kriging variance estimated by bootstrapping. *J Oper Res Soc* 2006;57(4):400–409.
 35. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.
 36. Van Beers W, Kleijnen JPC. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *Eur J Oper Res* 2008;186(3):1099–1113.
 37. Chilès J-P, Delfiner P. *Geostatistics: modeling spatial uncertainty*. New York: Wiley; 1999.
 38. Wackernagel H. *Multivariate geostatistics: an introduction with applications*. 3rd ed. Berlin: Springer; 2003.
 39. Kleijnen JPC, Van Beers WCM. Monotonicity-preserving bootstrapped Kriging metamodels for expensive simulations. Working Paper. 2010. Tilburg University; In press.

SEQUENTIAL DECISION PROBLEMS UNDER UNCERTAINTY

CHELSEA C. WHITE III
School of Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta,
Georgia

We consider in this article, sequential decision problems under uncertainty. We list their general features, discuss how they can be described mathematically, describe what constitutes a practical solution, discuss how the solution can be obtained, and mention several variations of the standard problem statement.

Consider a *system* that evolves between multiple *decision epochs*. At each decision epoch, an *action* (or decision) is selected based on the system's current *state*, and a reward is generated that depends on both the current state and action. The system state then makes transition (*system dynamics*) to a new state at the next decision epoch, and this transition is affected by both the current state and action. The next decision epoch arrives, and the process continues till the end of the *planning horizon*. The current action selected has both immediate impact (e.g., the reward of making the decision) and long-term impact (e.g., the impact on the system state at the next decision epoch). The *problem objective* is to maximize the *problem criterion*, a measure of value that can depend on the values of the states and actions at decision epochs over the planning horizon. A rule that selects an action based on the system's state is called a *policy*. An *optimal policy* is a policy that maximizes the criterion with respect to all (feasible) policies. A *problem solution* is an optimal policy and the value of the concomitant criterion.

For example, consider a camera store that is open for business Monday through Saturday. Inventory is counted each Saturday at the end of the work day, an order is placed to a wholesaler, and the order

arrives Monday, just as the store opens. Customers arrive through the week to make purchases. A wholesale cost is accrued when cameras are ordered; reward is accrued when a camera is sold, and holding costs are accrued for cameras that are not sold. Our objective is to maximize the average profit (reward minus cost) that we can expect to receive over a week's time by selecting the best number of cameras to order, based on current inventory count. Ordering too few cameras reduces immediate wholesale cost but may cause the inventory to stock out, reducing expected profits. Ordering too many cameras will increase immediate wholesale cost and inventory holding cost associated with the cameras that are unsold by the end of the week. Further examples of sequential decision problems under uncertainty, including routing and scheduling, finance, health care, and communications network management, can be found in Refs 1–3.

All models of sequential decision making under uncertainty have the following key ingredients:

- a set of decision epochs;
- a problem horizon;
- a set of actions;
- a definition of state;
- a definition of state observation;
- a definition of what the decision maker knows and when these data become known (called the *information pattern*);
- a description of the dynamics of the system, that is, how the state of the system evolves between decision epochs, as a function of current state and action taken;
- a value structure dependent on action and state values over the problem horizon.

Referring to the camera inventory example:

- Decision epochs occur each Saturday evening.

- Although not explicitly stated above, we will assume that the planning horizon is infinite and hence contains a countable but infinite number of decision epochs.
- The set of actions is the number of cameras that can be ordered each Saturday evening, that is, $\{0, 1, 2, \dots\}$.
- The state is the inventory level, a number in the state space $\{0, 1, 2, \dots\}$, assuming backordering is not allowed.
- We have implicitly assumed that inventory level is perfectly observed at each decision epoch (perfect observations).
- We assume that the decision maker knows the state of the inventory at each decision epoch, past and present (perfect memory).
- We assume that next Saturday's inventory level is the sum of this Saturday's inventory level and the number of items ordered, minus the number of customers who will purchase cameras through the week. In order to insure that our problem is well defined, we assume to know the probability the number of customers who will enter the store through the coming week and make a purchase. Thus, if D is the (random) number of purchases made during the week, then we assume that the probability $\text{Prob}\{D = d\}$ is given for all $d \in \{0, 1, 2, \dots\}$.
- The problem criterion is the expected profit (reward minus cost) per week, where there is a cost per item of ordering wholesale, and inventory holding cost per item for an item in inventory that is not sold, and a reward received when a camera is sold.

For our camera inventory example, a policy is a rule that determines the number of cameras to order and that this number can depend on how many cameras are currently in inventory. An optimal policy, which may not be unique, is a policy that maximizes expected weekly profit. A solution is an optimal policy and the resulting maximum expected weekly profit. At an operational level, an optimal policy tells the decision maker the number of cameras to order

each Saturday, as a function of the current inventory level. Strategically, knowing the maximum expected weekly profit is valuable if deciding whether the company is worth purchasing, assuming the camera store were for sale.

An alternative criterion is the expected total discounted cost over a finite or infinite horizon. In this case, we may need to specify a *discount factor*, which relates the value of a unit of money received at the current decision epoch to the value of a unit of money received at the next decision epoch.

We now explore the various ingredients of a model of sequential decision making under uncertainty in more detail.

Decision Epochs and Problem Horizons. Decision epochs occurred periodically (weekly) in the camera inventory problem, and the planning horizon is infinite. For the general model, decision epochs can also occur when a particular event occurs. For example, the assignment of an incoming job to a set of available machines occurs when the incoming job arrives. Further, the problem of hiring an administrative assistant [i.e., the secretary problem [4]] depends on when applications are received and when subsequent interviews can be scheduled.

The problem horizon can be finite or infinite. Typically, for a problem horizon of length T , the decision epochs are the set $\{1, 2, \dots, T\}$. Throughout, we limit our discussion to a countable number of decision epochs over the length of the problem horizon; see [1], Chapter 11, for a discussion of continuous time models. We also remark that T may be a random variable, as would be the case for optimal stopping problems. Such problems are concerned with choosing a time or decision epoch to take a particular action, such as the secretary problem.

Actions. Let $\mathbf{A}(i)$ be the set of actions available for selection at any decision epoch, given that the state of the system is i . For the camera inventory problem, $\mathbf{A}(i) = \{0, 1, 2, \dots\}$ for all levels of inventory i . In general, the set $\mathbf{A}(i)$ can be finite, countable and infinite, or uncountable. We let $a(t)$ be the action taken at epoch t .

State and State Dynamics. The state of the system is the information that one needs to know about the system in order to intelligently select a good action. For example, in order to intelligently place an order for more cameras, the decision maker needs to know how many cameras are currently in stock. We remark that the level of the camera store's inventory (the state) is dynamic. If $s(t)$ is the inventory level at epoch t , then state dynamics can be described by the equation $S(t+1) = \max\{0, s(t) + a(t) - D(t)\}$, where $S(t+1)$ is the random variable describing the inventory level at epoch $t+1$, $a(t)$ is the number of cameras ordered at epoch t and the random variable $D(t)$ is the number of cameras bought between epochs t and $t+1$.

In general, the dynamics of the state can be described by the stochastic difference equation $S(t+1) = f(s(t), a(t), D(t))$, where $D(t)$ is a random variable with given cumulative distribution function $\text{Prob}\{D(t) \leq d\}$. An alternative description of state dynamics is the conditional probability $\text{Prob}\{s(t+1) = j | s(t) = i, a(t) = a\}$.

Occasionally, other descriptions of state dynamics are given, such as set inclusion, for example, $S(t+1) \in \alpha(s(t), a(t), D(t))$, where $\alpha(s, a, D)$ is a given set dependent on the triple s, a, D . For example, let $s(t)$ be the cost of a barrel of oil at time t . Then, the statement "the cost of a barrel of oil a week in the future is within plus or minus \$10 of the current cost of a barrel of oil" can be described as $s(t) - 10 \leq S(t+1) \leq s(t) + 10$.

State Observations. The camera store example assumes that current inventory can be counted perfectly without cost. Inventory levels for large inventories, however, are not necessarily known accurately. More generally, there are many other examples of systems where the system's state is observed inaccurately and/or a state observation, perfect or imperfect, is costly. For example, any system that involves diagnosis would fall into this category, such as problems involving machine maintenance and repair and human diagnosis and treatment. With regard to the latter example, data available to the decision maker (a health-care provider) are signs, symptoms, and laboratory test results, rather than the actual underlying state of

the patient's health. As another example with an underlying network structure, consider traffic congestion monitoring. Sensors monitor traffic speed and variability, sensor data are transmitted to a traffic control center, and the level of traffic congestion is then inferred. For this example and for the diagnostic examples, corrupted state observations can result from sensor error and data transmission error.

There are two general ways of taking into account state observation error. Let $z(t)$ be the observation of the state available to the decision maker. When state dynamics are described by conditional probability, it is typically assumed that probabilities of the form $\text{Prob}\{Z(t+1) = z, S(t+1) = j | s(t) = i, a(t) = a\}$ are given. Note that

$$\begin{aligned} \text{Prob}\{Z(t+1) = z, S(t+1) = j | s(t) \\ = i, a(t) = a\} &= \text{Prob}\{Z(t+1) = z | s(t+1) \\ &= j, s(t) = i, a(t) = a\} \\ \text{Prob}\{S(t+1) = j | s(t) = i, a(t) = a\}, \end{aligned}$$

where $\text{Prob}\{Z(t+1) = z | s(t+1) = j, s(t) = i, a(t) = a\}$ is referred to as the observation probability and $\text{Prob}\{S(t+1) = j | s(t) = i, a(t) = a\}$ is referred to as the transition probability. We remark that it is usually assumed that $\text{Prob}\{Z(t+1) = z | s(t+1) = j, s(t) = i, a(t) = a\}$ is independent of $s(t) = i$.

If functional equations describe system dynamics, it is typically the case that observations also are described functionally. Hence, $S(t+1) = f(s(t), a(t), D(t))$ and $Z(t+1) = g(s(t+1), s(t), a(t), D'(t))$, where $D(t)$ and $D'(t)$ are random variables with given distribution functions. For a system with state and action spaces having finite cardinality, the conditional distribution description and the functional description of state dynamics are equivalent.

We mention another description of dynamic systems involving corrupted state observations that often has an implicit notion of state: AND/OR tree. The AND/OR tree is usually associated with decision analysis or heuristic search, an area of artificial intelligence (AI). We remark that "heuristic" has different meanings in the operations

research and AI literatures. OR nodes are called *decision nodes* and AND nodes are *consequence* (or “nature’s”) *nodes*. An OR node is the root node, AND nodes follow OR nodes, and OR nodes (other than the root node) follow AND nodes. Each link from an OR node to an AND node is associated with an action $a(t)$, and each link from an AND node to an OR node is associated with an observation $z(t+1)$.

Information Pattern. The information pattern is a description of what the decision maker knows and when the decision maker knows it. It is typically assumed that when a decision epoch occurs, the decision maker knows the current state of the system without delay. It is generally assumed that the decision maker has perfect memory. Hence, actions at epoch t are made, based on the information set $I(t)$, which comprises all past and present observations, $z(1), \dots, z(t)$, all former actions, $a(1), \dots, a(t-1)$, and the *a priori* $\text{Prob}\{s(1) = i\}$, for all i . If the system is completely observed, $s(t)$ serves as a *sufficient statistic* for $I(t)$. That is, knowing $s(t)$ is sufficient for action selection, relative to knowing $I(t)$. For the partially observed case, $x(t)$ is a sufficient statistic, where the i th element of the vector $x(t)$ is the conditional probability $\text{Prob}\{S(t) = i | I(t)\}$.

Policy. A policy is a function that maps $I(t)$ into the action set at epoch t . Thus, policy π chooses action $a(t)$ as follows: $\pi(I(t)) = a(t)$.

Criterion. Let $r(s(1), a(1), \dots, s(T), a(T))$ be the (scalar) reward that we would like to maximize. The value of this reward for policy π and given initial state $s(1)$ is $r(s(1), \pi(I(1)), S(2), \pi(I(2)), \dots, S(T), \pi(I(T)))$. We note, however, that since the $S(t)$ and the $\pi(I(t))$ are random variables for $t \geq 2$, the reward is a random variable. There are several ways of comparing two random variables; we mention two. First, map each random variable into a scalar, such as the maximum (or minimum) the random value can take or the expectation of the reward. We then prefer the larger of the resulting scalar values. Second, compare two random variables using their cumulative probability distributions. For example, let $R = r(s(1), \pi(I(1)), \dots, S(T), \pi(I(T)))$ and $R' = r(s(1), \pi'(I(1)), \dots, S(T),$

$\pi'(I(T)))$. We prefer π to π' if and only if $\text{Prob}\{R' \leq r\} \leq \text{Prob}\{R \leq r\}$ for all r . By far, the most examined criterion is the expectation of the reward.

We remark that the reward function can assume many specialized functional forms. The two most examined are the additive form

$$r(s(1), a(1), \dots, s(T), a(T)) = \sum r(s(t), a(t))$$

and the multiplicative form

$$r(s(1), a(1), \dots, s(T), a(T)) = \prod r(s(t), a(t)).$$

Both of these forms represent appropriate model structure for a variety of real world applications and enable computational simplification.

Problem Objective. Our objective is to determine an optimal policy, that is, a policy that generates a “best” criterion, justifying the above discussion on how two random variables can be compared. Assume, for example, that $v(\pi) = E[r(s(1), \pi(I(1)), \dots, S(T), \pi(I(T)))]$, where E is the expectation operator [conditioned on the initial state value, $s(1)$], and assume that we prefer π to π' if and only if $v(\pi) \leq v(\pi')$. Then, the problem objective is to find a π^* , an optimal policy, such that $v(\pi) \leq v(\pi^*)$, for all policies π . Under a robust set of circumstances, there exists an optimal policy; see [1] for details.

We also remark that both π^* and $v(\pi^*)$ are of interest. An optimal policy is a set of IF-THEN rules that tells the decision maker what actions to take under what circumstances and when [i.e., IF $I(t)$ is the information set at epoch t , THEN $\pi^*(I(t))$ is an action to select]. The scalar $v(\pi^*)$ informs the decision maker what the expected reward will be, assuming actions taken are based on an optimal policy. If this scalar is too low, then the decision maker may decide, for example, to raise prices or find ways to lower costs, or to make a more strategic decision to leave or not enter the business.

Complications can arise when the problem horizon is infinite. An example of such a problem is when $v(\pi)$ is infinite for all π . In this situation, it is typically the case where

total (bounded) costs are discounted by a discount factor strictly less than 1, insuring that $v(\pi)$ for all π is finite if $r(s, a)$ is uniformly bounded. Another approach, assumed in the camera inventory example, is for the criterion to be the average expected value accrued between epochs.

Problem Solution Determination. Thus far, we have described a mathematical model of a generic sequential decision problem under uncertainty. If the model is a reasonably accurate description of the real problem, the problem solution, if known, would be of value to the decision maker. We now show how to determine an optimal policy and the value of the concomitant criterion.

While determining a problem solution, it is an immense help if the criterion is either additive or multiplicative. If so, then the so-called *principle of optimality* can be invoked and a standard application of *dynamic programming* can be used to essentially decompose the problem into a series of single decision epoch problems [5]. Thus, the problem of determining π^* from the set of all policies is equivalent to determining what action to select at epoch t , as a function of $s(t)$, from the set of all available actions, for all t starting at T (when T is finite) and working backward to 1 (which is why this application of dynamic programming is sometimes referred to as *backwards dynamic programming*). This decomposition is referred to as *value iteration* [6]. In the parlance of decision analysis, dynamic programming and *folding back the tree* are equivalent [7]. Although somewhat different in concept, the AI-based algorithms A^* and AO^* also have analogies to dynamic programming [8].

When T is infinite and the criterion is either the expected total discounted reward or the expected average reward, the existence of an epoch-invariant optimal policy is guaranteed when the function $r(s, a)$ is epoch invariant. *Modified policy iteration* [1] is a set of computational procedures for determining an optimal policy and criterion value in this case. These procedures are distinguished by an index $k \in \{1, 2, \dots\}$, where when $k = 1$, modified policy iteration is also called *value iteration*, and when k approaches infinity,

modified policy iteration is called *policy iteration*. *Linear programming* can also be used to find a problem solution.

Structured Results. We remark that certain classes of problems have optimal policies with computationally favorable and easily implemented structural characteristics. For example, consider the camera inventory example. Under very robust circumstances, there exists an optimal policy of the form $\pi(s(t)) = \max\{s^* - s(t), 0\}$, where s^* is an “order up to” inventory number. Note that the problem is now reduced to finding the scalar s^* . Note also that the rule “order up to s^* ” is easy to understand and hence implement.

Many maintenance problems are such that it is optimal to let the machine produce when the “state” of the machine (e.g., a measure of the machine’s capacity to produce quality products) is above a particular threshold and replace the machine once its state has deteriorated below this threshold.

There exists an optimal (s, S) policy for a broad class of inventory problems that require a setup charge before delivery, where one orders up to S if the inventory falls below s ($s \leq S$). Thus, the problem is reduced to finding s and S . Further details can be found in Ref. 1.

The multiarmed bandit problem [9] requires the decision maker to decide at each decision epoch which one of several levers to pull on a multiarmed slot machine in order to maximize expected total discounted reward. A reward is accrued each time a lever is pulled, based on an initially unknown distribution associated with that particular lever. With each pull, the distribution of the selected lever is better identified. The trade-off is between pulling levers with high expected payoffs and learning more about the distributions of the other levers. We also assume that at any decision epoch, the decision maker can quit the game. Under robust assumptions, there exist optimal policies having the form of an index rule: if lever i has the highest index, relative to the indices of the other levers, and if its index is larger than the retirement reward, then pull lever i . If the highest index for all levers is less than the retirement reward, then stop.

Hence, this problem is a type of optimal stopping problem.

Violations of the Principle of Optimality. We remark that if the criterion is a weighted sum of an additive functional form and a multiplicative functional form, the principle of optimality may not hold and hence standard application of dynamic programming algorithms may not be guaranteed to produce an optimal problem solution. Such a situation is of interest since additive criteria are often used to model cost and multiplicative criteria can be used to model risk. Thus, if there is a trade-off of cost and risk under consideration (as might be the case in disruption management or hazardous materials transport routing), then problem solution determination can require specialized algorithms; see [10] for details.

Problem Variations. There are many interesting variations on the standard problem described above, and these variations produce interesting analytical challenges.

Throughout, we have assumed a single decision maker. However, multiple decision makers may exist, each having its own unique information pattern. If each decision maker has its own criterion and is essentially competing with the others, then this is a *game*. If each decision maker is trying to minimize the same criterion, then this is a *team*.

A decision maker may be trying to maximize several criteria simultaneously, which becomes a multiobjective problem [11] in which the problem objective often becomes one of determining the set of *noninferior* policies.

Models of sequential decision problems under uncertainty require an immense amount of precisely specified parameters. In reality, many of the parameters may not be precisely known. For example, $r(s, a)$ may depend on a parameter λ , where all we know is that $\lambda \in \Lambda$. This results in a *parametric problem*, which poses many interesting computational challenges.

We have assumed throughout that the state, or an observation of the state, is available to the decision maker without delay. However, in reality, especially when the system state is measured by sensors,

there are invariable (noise corrupted) data transmission and analysis *delays* that may compromise the *value of information*. An analysis of the value of information, as a function of delay and observation quality, can be found in Ref. 12.

Determining the solutions of problems with large state and action spaces can be computationally challenging. Interesting new techniques for solving such problems can be found in Ref. 13.

REFERENCES

1. Puterman ML. Markov decision processes. New York: John Wiley & Sons, Inc.; 1994.
2. White DJ. Real applications of Markov decision processes. *Interfaces* 1985;15:73–83.
3. White DJ. Further real applications of Markov decision processes. *Interfaces* 1988;18:55–61.
4. Bearden JN, Rapoport A, Murphy RO. Sequential observation and selection with rank-dependent payoffs: an experimental test. *Manage Sci* 2006;52:1437–1449.
5. Bellman R. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
6. Howard RA. Dynamic programming and Markov processes. Cambridge (MA): The MIT Press; 1960.
7. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. Cambridge: Cambridge University Press; 1993.
8. Pearl J. Heuristics: intelligent search strategies for computer problem solving. Boston (MA): Addison Wesley; 1984.
9. Gittens JC. Bandit processes and dynamic allocation indices. *J R Stat Soc [Ser A]* 1979;41:148–177.
10. White CC, Stewart BS, Carraway RL. Multiobjective preference-based search in acyclic OR graphs. *Eur J Oper Res* 1992;56(3):357–363.
11. Stewart BS, White CC. Multiobjective A*. *J Assoc Comput Mach* 1991;38:775–814.
12. Bander J, White CC. Markov decision processes with noise-corrupted and delayed state observations. *J Oper Res Soc* 1999;50(6):660–668.
13. Powell WB. Approximate dynamic programming. New York: Wiley-Interscience; 2007.

SEQUENTIAL QUADRATIC PROGRAMMING METHODS

KLAUS SCHITTKOWSKI
Department of Computer Science,
University of Bayreuth, Bayreuth,
Germany

YA-XIANG YUAN
AMSS, Academy of Mathematics and
Systems Science, Chinese
Academy of Sciences, Beijing,
China

INTRODUCTION

We consider the nonlinear, constrained optimization problem to minimize an objective function f under m nonlinear inequality constraints, that is,

$$\begin{aligned} &\text{minimize } f(x) \\ &x \in \mathbb{R}^n : g(x) \geq 0, \end{aligned} \quad (1)$$

where x is an n -dimensional parameter vector and where $g(x) := (g_1(x), \dots, g_m(x))^T$, that is, we have m nonlinear inequality constraints. Equation (1) is now called a *nonlinear program* (NLP) in abbreviated form.

Equality constraints and simple bounds for the variables are omitted to simplify the subsequent notation. They are easily introduced again whenever needed, where equality constraints are linearized in the quadratic programming subproblem. It is assumed that the objective function $f(x)$ and m constraint functions $g_1(x), \dots, g_m(x)$ are continuously differentiable on the whole $\mathbb{R}^n$.

Optimization theory for smooth problems is based on the Lagrangian function that combines the objective function $f(x)$ and the constraints $g(x)$ in a proper way. In particular, the Lagrangian function allows us to state necessary and sufficient optimality conditions.

Let problem (1) be given. First we define the *feasible region* P as the set of all feasible solutions

$$P := \{x \in \mathbb{R}^n : g(x) \geq 0\}. \quad (2)$$

and the set of *active constraints* with respect to $x \in P$ by

$$I(x) := \{j : g_j(x) = 0, 1 \leq j \leq m\}. \quad (3)$$

The *Lagrangian function* of Equation (1) is defined by

$$L(x, u) := f(x) - u^T g(x) \quad (4)$$

for all $x \in \mathbb{R}^n$ and $u = (u_1, \dots, u_m)^T \in \mathbb{R}^m, u \geq 0$. The variables u_j are called the *Lagrangian multipliers* of the nonlinear programming problem. x is also called the *primal* and u the *dual* variable of the NLP (1).

The first idea has been investigated in the Ph.D. thesis of Wilson [1]. Sequential quadratic programming (SQP) methods became popular during the late 1970s due to the articles of Han [2,3] and Powell [4,5]. Their superiority over other optimization methods known at that time, was shown by Schittkowski [6]. Since then many modifications and extensions have been published on SQP methods. Nice review articles are given by Boggs and Tolle [7] and Gould and Toint [8]. All modern optimization textbooks have chapters on SQP methods, for example, see Fletcher [9], Gill *et al.* [10] and Sun and Yuan [11]. As the presentation of even a selected overview is impossible due to the limited space here, we concentrate on a few important facts and highlights from a personal view without any attempt to be complete.

The basic ideas of SQP methods are found in the section titled “Sequential Quadratic Programming Methods.” The role of quasi-Newton updates is outlined in the section titled “Reduced Hessian SQP Steps.” Several stabilization procedures are available by which convergence is guaranteed, for example, line search, which is

discussed in the section titled “Stabilization by Line Search,” trust regions is discussed in the section titled “Stabilization by Trust Regions,” or refer to filter techniques outlined in the section titled “Stabilization by a Filter,” Some typical convergence results for global convergence, that is, convergence toward a stationary point starting from an arbitrary point, and local convergence, fast final convergence speed close to a solution, are listed in the section titled “Convergence.”

SEQUENTIAL QUADRATIC PROGRAMMING METHODS

Sequential quadratic programming or SQP methods belong to the most powerful optimization algorithms we know today for solving differentiable NLPs of the form (1). The theoretical background is described, for example, in Stoer [12], and an excellent review is given by Boggs and Tolle [7]. From a more practical point of view, SQP methods are also introduced in the books of Papalambros and Wilde [13] and Edgar and Himmelblau [14], among many others. Their excellent numerical performance has been tested and compared with other methods, see Schittkowski [6], and for many years they belonged to the set of most frequently used algorithms for solving practical optimization problems.

The basic idea is to formulate and solve a quadratic programming subproblem in each iteration, which is obtained by linearizing the constraints and approximating the Lagrangian function $L(x, u)$ (4) quadratically. Starting from any $x_0 \in \mathbb{R}^n$, suppose that $x_k \in \mathbb{R}^n$ is an actual approximation of the solution, $v_k \in \mathbb{R}^m$ an approximation of the multipliers, and $B_k \in \mathbb{R}^{n \times n}$ an approximation of the Hessian of the Lagrangian function, $k = 0, 1, 2, \dots$. Then, a quadratic program (QP) of the form

$$\begin{aligned} & \text{minimize } \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \\ & d \in \mathbb{R}^n : \nabla g(x_k)^T d + g(x_k) \geq 0 \end{aligned} \quad (5)$$

is formulated and solved in each iteration. To simplify the subsequent presentation, we assume that (Eq. 5) is always solvable. Let d_k

be the optimal solution, u_k the corresponding multiplier of this subproblem. A new iterate is then obtained by

$$x_{k+1} := x_k + d_k \quad (6)$$

and, for the moment, $v_{k+1} := u_k$.

To motivate the above approach, let us assume that there are only equality constraints of the form $g(x) = 0$. The Karush-Kuhn-Tucker (KKT) optimality conditions are then written in the form

$$F(x, u) := \begin{pmatrix} \nabla_x L(x, u) \\ g(x) \end{pmatrix} = 0.$$

In other words, the optimal solution and the corresponding multipliers are the solution of a system of $n + m$ nonlinear equations $F(x, u) = 0$ with $n + m$ unknowns $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$.

Let (x_k, v_k) be an approximation of the solution. We apply Newton’s method and get an estimate for the next iterate by

$$\nabla F(x_k, v_k) \begin{pmatrix} d_k \\ y_k \end{pmatrix} + F(x_k, v_k) = 0.$$

After insertion, we obtain the equation

$$\begin{pmatrix} B_k & : & -\nabla g(x_k) \\ \nabla g(x_k)^T & : & 0 \end{pmatrix} \begin{pmatrix} d_k \\ y_k \end{pmatrix} + \begin{pmatrix} \nabla f(x_k) - \nabla g(x_k) v_k \\ g(x_k) \end{pmatrix} = 0 \quad (7)$$

with $B_k := \nabla_{xx} L(x_k, v_k)$. Defining now $u_k := y_k + v_k$, we get

$$B_k d_k - \nabla g(x_k) u_k + \nabla f(x_k) = 0$$

and

$$\nabla g(x_k)^T d_k + g(x_k) = 0.$$

But these equations are exactly the optimality conditions for the QP formulated for equality constraints only. To sum up, we get the following conclusion.

A SQP method is identical to Newton’s method for solving the KKT optimality conditions, if B_k is the Hessian of the Lagrangian

function and if we start sufficiently close to a solution.

Now, we assume again that we have inequality constraints instead of the equality ones. A straightforward analysis based on the optimality conditions shows that if $d_k = 0$ is an optimal solution of Equation (5) and u_k the corresponding multiplier vector, then x_k and u_k satisfy the necessary optimality conditions of Equation (1).

Note that the above analysis also serves to point out the crucial role of the dual variable u_k . To get a fast convergence rate as expected for Newton's method, we have to adapt the multiplier values accordingly together with the primal ones.

To avoid the computation of second derivatives, B_k is set to a positive definite approximation of the Hessian of the Lagrangian function. For the global convergence analysis, any choice of B_k is appropriate as long as the eigenvalues are bounded away from zero. However, to guarantee a superlinear convergence rate, we update B_k by a quasi-Newton method, for example, the BFGS formula

$$B_{k+1} := B_k + \frac{y_k y_k^T}{s_k^T y_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} \quad (8)$$

with

$$\begin{aligned} y_k &:= \nabla_x L(x_{k+1}, u_k) - \nabla_x L(x_k, u_k), \\ s_k &:= x_{k+1} - x_k. \end{aligned} \quad (9)$$

Usually, we start with the unit matrix for B_0 and stabilize the formula by requiring that $a_k^T b_k \geq 0.2 b_k^T B_k b_k$. Positive definite matrices B_k guarantee unique solutions and in addition, a sufficient descent property of a merit function needed for the global convergence analysis.

Moreover, it is possible to proceed from a Cholesky factorization of B_k in the form

$$B_k = L_k L_k^T$$

with a lower triangular matrix L_k and to update L_k directly to get a factorization of B_{k+1} , see Gill *et al.* [15],

$$B_{k+1} = L_{k+1} L_{k+1}^T = L_k L_k^T + a_k a_k^T - b_k b_k^T.$$

A particular advantage is that an initial Cholesky decomposition is avoided, if a primal–dual method is applied to solve the QP (5).

Note that the above motivation proceeds from a local approximation of a solution. If starting from an arbitrary $x_0 \in \mathbb{R}^n$, $u_0 \in \mathbb{R}^m$, one has to stabilize the algorithm by introducing additional safeguards, for example, a line search, a trust region, or a filter. These three possibilities are discussed in the subsequent sections.

REDUCED HESSIAN SQP STEPS

The linear system (7) derived from the QP subproblem can be rewritten as

$$\begin{pmatrix} Y_k^T B_k Y_k & Y_k^T B_k Z_k & -R_k \\ Z_k^T B_k Y_k & Z_k^T B_k Z_k & 0 \\ -R_k^T & 0 & 0 \end{pmatrix} \times \begin{pmatrix} p_k \\ q_k \\ y_k \end{pmatrix} = \begin{pmatrix} -Y_k^T \nabla f(x_k) \\ -Z_k^T \nabla f(x_k) \\ g(x_k) \end{pmatrix}, \quad (10)$$

if we have the QR factorization

$$\nabla g(x_k) = (Y_k, Z_k) \begin{pmatrix} R_k \\ 0 \end{pmatrix},$$

and if we use the notation $p_k = Y_k^T d_k$ and $q_k = Z_k^T d_k$. Thus, it is easy to see that p_k and q_k can be obtained by

$$\begin{pmatrix} Z_k^T B_k Y_k & Z_k^T B_k Z_k \\ -R_k^T & 0 \end{pmatrix} \begin{pmatrix} p_k \\ q_k \end{pmatrix} = \begin{pmatrix} -Z_k^T \nabla f(x_k) \\ g(x_k) \end{pmatrix}, \quad (11)$$

which indicates that we only need the reduced matrix $Z_k^T B_k$ instead of the full matrix B_k to compute the QP step d_k . Hence, Nocedal and Overton [16] suggest to replace $Z_k^T B_k$ by a quasi-Newton matrix $\bar{B}_k$, and to obtain the QP step d_k by solving

$$\begin{pmatrix} \bar{B}_k \\ -\nabla g(x_k)^T \end{pmatrix} d = \begin{pmatrix} -Z_k^T \nabla f(x_k) \\ g(x_k) \end{pmatrix}. \quad (12)$$

The matrix $\bar{B}_k \in \mathbb{R}^{(n-m) \times n}$ is a quasi-Newton matrix that approximates $Z_k^T \nabla_{xx}^2 L(x_k, u_k)$,

which can be updated by the Broyden's unsymmetric Rank-1 formula

$$\bar{B}_{k+1} := \bar{B}_k + \frac{(\bar{y}_k - \bar{B}_k s_k) s_k^T}{\|s_k\|_2^2},$$

where $s_k := x_{k+1} - x_k$ and $\bar{y}_k := Z_{k+1}^T \nabla f(x_{k+1}) - Z_k^T \nabla f(x_k)$. This approach can be more efficient than the standard SQP because it reduces storage and calculation due to the fact that it uses a reduced matrix $B_k \in \mathbb{R}^{(n-m) \times n}$ instead of the full matrix B_k , particularly when m is close to n , namely, for problems there are lots of variables and constraints but less freedom in the variables. Moreover, this approach preserves the q-superlinearly convergent property of the standard SQP step, for example, see Nocedal and Overton [16].

We move along this direction further. If we replace $Z_k^T B_k Z_k$ by a quasi-Newton matrix $\hat{B}_k$ and replace $Z_k^T B_k Y_k$ by zero in the linear system (11), we obtain

$$\begin{pmatrix} 0 & \hat{B}_k \\ -R_k^T & 0 \end{pmatrix} \begin{pmatrix} p_k \\ q_k \end{pmatrix} = \begin{pmatrix} -Z_k^T \nabla f(x_k) \\ g(x_k) \end{pmatrix}. \quad (13)$$

The reason for doing so is that Powell [4] discovered that the SQP method converges two-step Q-superlinearly,

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_{k-1} - x^*\|} = 0, \quad (14)$$

provided that $Y_k^T B_k Z_k$ is bounded. An advantage of this approach, which is called *two-side reduced Hessian method*, compared to Nocedal and Overton's one-side reduced Hessian method, is that the matrix $\hat{B}_k$ is a square matrix and can be kept positive definite by updates such as the BFGS formula. Furthermore, the computation cost is also reduced. But, the price we have to pay is that we can only prove two-step Q-superlinear convergence (14) instead of the one-step Q-superlinear convergence. Indeed, examples exist to show that the two-step Q-superlinear convergence of the two-side reduced Hessian method cannot be improved to one-step Q-superlinear convergence, for example, see Byrd [17] and Yuan [18].

STABILIZATION BY LINE SEARCH

We have seen in the previous sections that one should better approximate the actual primal and dual variables simultaneously to satisfy the optimality conditions of (1) based on the Newton method. Thus, we define $z_k := (x_k, v_k)$, where $x_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are approximations of a KKT point in the k th step. A new iterate is then computed from

$$z_{k+1} := z_k + \alpha_k p_k, \quad (15)$$

where $p_k := (d_k, y_k) = (d_k, u_k - v_k)$ is the search direction obtained from solving the QP (5), see also Equation (7), and where α_k is a scalar parameter called the *step length*. The goal is to compute a sufficiently accurate step size within as few function evaluations as possible, so that the underlying algorithm converges. α_k should satisfy at least a sufficient decrease condition of a merit function

$$\phi_r(\alpha) := \psi_r \left(\begin{pmatrix} x \\ v \end{pmatrix} + \alpha \begin{pmatrix} d \\ u - v \end{pmatrix} \right) \quad (16)$$

with a suitable penalty function $\psi_r(x, v)$ and with $r = (r_1, \dots, r_m)^T \in \mathbb{R}^m$, a vector of m positive penalty parameters.

A merit function is supposed to contain information about the objective function to be minimized, and the violation of constraints. The first one studied in the context of SQP methods was the L_1 penalty function

$$\psi_r(x, v) := f(x) + \sum_{j=1}^m r_j g_j(x)^-, \quad (17)$$

where $g_j(x)^- := -\min(0, g_j(x))$ denotes violated constraints, see Han [3]. $\psi_r(x, v)$ is a so-called exact penalty function, that is, a minimizer of $\psi_r(x, v)$ subject to x for sufficiently large r is also a solution of NLP. But there are two severe drawbacks: $\psi_r(x, v)$ is nondifferentiable preventing efficient algorithms and, since $\psi_r(x, v)$ does not depend on multiplier information, it enables the Maratos effect [19], that is, the possibility of very slow convergence close to a solution.

Instead of Equation (17), augmented Lagrangian merit functions

$$\psi_r(x, v) := f(x) - v^T g(x) + \frac{1}{2} \sum_{j=1}^m r_j g_j(x)^2 \quad (18)$$

are an appropriate alternative, which differs by the choice of the multiplier approximation, see Boggs and Tolle [7] or Gill *et al.* [20]. A slightly different augmented Lagrangian function is given by

$$\begin{aligned} \psi_r(x, v) := & f(x) - \sum_{j \in J} (v_j g_j(x) \\ & - \frac{1}{2} r_j g_j(x)^2) - \frac{1}{2} \sum_{j \in K} v_j^2 / r_j \end{aligned} \quad (19)$$

with $J := \{j : 1 < j \leq m, g_j(x) \leq v_j / r_j\}$ and $K := \{1, \dots, m\} - J$, see Schittkowski [21] and Rockafellar [22], but very many other merit functions have been investigated in the past.

In all cases, the objective function is *penalized* as soon as an iterate leaves the feasible domain. The corresponding penalty parameters r_j , $j = 1, \dots, m$, which control the degree of constraint violation, must be carefully chosen to guarantee a descent direction of the merit function, see Schittkowski [21] or Wolfe [23] in a more general setting, that is, to achieve at least

$$\phi_{r_k}(0) = \nabla \psi_{r_k}(x_k, v_k)^T \begin{pmatrix} d_k \\ u_k - v_k \end{pmatrix} < 0 \quad (20)$$

when applied in the k -iteration step.

The implementation of a line search algorithm is a crucial issue when implementing a nonlinear programming algorithm, and has significant effect on the overall efficiency of the resulting code. On the one hand, we need a line search to stabilize the algorithm. On the other hand, it is not desirable to waste too many function calls. Moreover, the behavior of the merit function becomes irregular in case of constrained optimization because of very steep slopes at the border caused by large penalty terms. The implementation is more complex than shown above, if linear constraints and bounds of the variables are to be satisfied during the line search.

Usually, the step length parameter α_k is chosen to satisfy a certain Armijo [24] condition, that is, a sufficient descent condition

of a merit function, which guarantees convergence to a stationary point. However, to take the curvature of the merit function into account, we need some kind of compromise between a polynomial interpolation, typically a quadratic one, and a reduction of the step size by a given factor, until a stopping criterion is reached. Since $\phi_r(0)$, $\phi'_r(0)$, and $\phi_r(\alpha_i)$ are given, and α_i is the actual iterate of the line search procedure, we easily get the minimizer of the quadratic interpolation. We accept then the maximum of this value and the Armijo parameter as a new iterate, as shown by the subsequent code fragment.

Algorithm 1. Let β , μ with $0 < \beta < 1$, $0 < \mu < 0.5$ be given.

Start: $\alpha_0 = 1$

For $i = 0, 1, 2, \dots$ do:

1. If $\phi_r(\alpha_i) < \phi_r(0) + \mu \alpha_i \phi'_r(0)$, then stop.
2. Compute $\bar{\alpha}_i := \frac{0.5 \alpha_i^2 \phi'_r(0)}{\alpha_i \phi'_r(0) - \phi_r(\alpha_i) + \phi_r(0)}$.
3. Let $\alpha_{i+1} := \max(\beta \alpha_i, \bar{\alpha}_i)$.

The algorithm goes back to Powell [4] and corresponding convergence results are found in Schittkowski [21]. $\bar{\alpha}_i$ is the minimizer of the quadratic interpolation, and we use the Armijo descent property for checking termination. Step 3 is required to avoid irregular values, since the minimizer of the quadratic interpolation could be outside of the feasible domain $(0, 1]$. Additional safeguards are required, for example, to prevent violation of bounds. Algorithm 1 assumes that $\phi_r(1)$ is known before calling the procedure, that is, that the corresponding function values are given. We stop the algorithm if sufficient descent is not observed after a certain number of iterations. If the tested step size falls below machine precision or the accuracy by which model function values are computed, the merit function cannot decrease further.

It is possible that $\phi'_r(0)$ becomes negative due to round-off errors in the partial derivatives or that Algorithm 1 breaks down because of too many iterations. In this case, we proceed from a descent direction of the merit function, but $\phi'_r(0)$ is extremely small. To avoid interruption of the whole iteration process, the idea is to repeat the line search

with another stopping criterion. Instead of testing Equation (20), we accept a step size α_k as soon as the inequality

$$\phi_{r_k}(\alpha_k) \leq \max_{k-p(k) \leq j \leq k} \phi_{r_j}(0) + \alpha_k \mu \phi'_{r_k}(0) \quad (21)$$

is satisfied, where $p(k)$ is a nondecreasing integer with $p(k) := \min\{k, p\}$, p a given tolerance. Thus, we allow an increase of the reference value $\phi_{r_{j_k}}(0)$ in a certain error situation, that is, an increase of the merit function value.

To implement the nonmonotone line search, we need a queue consisting of merit function values at previous iterates. In case of $k = 0$, the reference value is adapted by a factor greater than 1, that is, $\phi_{r_{j_k}}(0)$ is replaced by $t\phi_{r_{j_k}}(0)$, $t > 1$. The basic idea to store reference function values and to replace the sufficient descent property by a sufficient “ascent” property in max-form, is described in Dai [25] and Dai and Schittkowski [26], where convergence proofs are presented. The general idea goes back to Grippo *et al.* [27], and was extended to constrained optimization and trust-region methods in a series of subsequent articles, see, for example, Toint [28,29].

STABILIZATION BY TRUST REGIONS

Another globalization technique for nonlinear optimization is trust region, where the next iterate requires to be in the neighborhood of the current point, namely, $\|x_{k+1} - x_k\| \leq \Delta_k$, where $\Delta_k > 0$ is the trust-region radius updated from iteration to iteration. Thus, an essential constraint of a trust-region subproblem is

$$\|d\|_2 \leq \Delta_k. \quad (22)$$

But, simply adding the trust-region constraint (22) to the QP (5) may lead to an infeasible subproblem. There are mainly three approaches in constructing trust region subproblems. To simplify our discussions, first we consider only equality constraints.

The first approach is the null-space technique, which was studied by many authors, including Vardi [30], Byrd *et al.* [31], and

Omojokun [32]. In this approach, the trial step s_k consists of a range-space step v_k and the null-space step h_k . v_k reduces the linearized constraint $\|\nabla g(x_k)^T d + g(x_k)\|$ and h_k reduces the approximate Lagrange function in the null space of $\nabla g(x_k)^T$. For example, v_k can be the least norm solution of

$$\begin{aligned} & \text{minimize } \|\nabla g(x_k)^T v + g(x_k)\|_2^2 \quad (23) \\ & v \in \mathbb{R}^n : \|v\|_2 \leq \delta \Delta_k, \end{aligned}$$

where $\delta \in (0, 1)$ is a given constant. The parameter $\delta < 1$ leaves freedom for defining h_k . Once v_k is computed, h_k can be obtained by solving

$$\begin{aligned} & \text{minimize } \frac{1}{2} (v_k + h)^T B_k (v_k + h) + \nabla f(x_k)^T h \\ & h \in \mathbb{R}^n : \nabla g(x_k)^T h = 0 \quad (24) \\ & \|h\|_2 \leq \sqrt{1 - \delta^2} \Delta_k. \end{aligned}$$

The trust-region step $s_k := v_k + h_k$ obtained by the null space technique can be regarded as the solution of the QP

$$\begin{aligned} & \text{minimize } \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \\ & d \in \mathbb{R}^n : \nabla g(x_k)^T d + \theta_k g(x_k) = 0 \quad (25) \\ & \|d\|_2 \leq \Delta_k \end{aligned}$$

for some parameter $\theta_k \in (0, 1]$.

The second approach is to replace all the linearized constraints in Equation (5) by a single quadratic constraint, which was proposed by Celis *et al.* [33],

$$\begin{aligned} & \text{minimize } \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \\ & d \in \mathbb{R}^n : \|\nabla g(x_k)^T d + g(x_k)\|_2 \leq \xi_k \quad (26) \\ & \|d\|_2 \leq \Delta_k, \end{aligned}$$

where $\xi_k \geq \min_{\|d\|_2 \leq \Delta_k} \|\nabla g(x_k)^T d + g(x_k)\|_2$ to ensure the subproblem is feasible. One particular choice for ξ_k is given by Powell and Yuan [34]:

$$\begin{aligned} & \min_{\|d\|_2 \leq \delta_1 \Delta_k} \|\nabla g(x_k)^T d + g(x_k)\|_2 \leq \xi_k \\ & \leq \min_{\|d\|_2 \leq \delta_2 \Delta_k} \|\nabla g(x_k)^T d + g(x_k)\|_2 \end{aligned}$$

where $0 < \delta_2 < \delta_1 < 1$ are two constants.

The third type of trust-region subproblems is to minimize an approximation of some penalty function. On the basis of the L_∞ exact penalty, Yuan [35] suggests

$$\begin{aligned} \text{minimize } \Phi_k(d) &:= \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \\ &+ \sigma_k \|\nabla g(x_k)^T d + g(x_k)\|_\infty \quad (27) \\ d \in \mathbb{R}^n : &\|d\|_\infty \leq \Delta_k, \end{aligned}$$

which can be converted into a standard QP. The penalty parameter $\sigma_k > 0$ may need to be updated to ensure the sufficient reduction in the norm of the residuals of linearized constraints. The L_∞ penalty function in Equation (27) can be replaced by any other penalty functions, including the augmented Lagrange function studied by Niu and Yuan [36]. For example, if we use the L_1 penalty function, it leads to the famous SL_1 QP method given by Fletcher [37].

Once the trial step s_k is computed, a trust-region algorithm for constrained optimization can be given as follows.

Algorithm 2. (A general trust-region SQP method)

Let $0 < \tau_3 < \tau_4 < 1 < \tau_1$, $0 < \tau_2 < 1$, and $x_0 \in \mathbb{R}^n$ be given.

For $k = 0, 1, 2, \dots$ until convergence do

1. Compute the trial step s_k .
2. Compute the ratio between actual reduction and predicted reduction in a merit function Ψ_k ,

$$r_k := \frac{\Psi_k(x_k) - \Psi_k(x_k + s_k)}{Pred_k}$$

3. Let

$$x_{k+1} := \begin{cases} x_k + s_k & \text{if } r_k > 0, \\ x_k & \text{otherwise,} \end{cases}$$

and

$$\Delta_{k+1} \in \begin{cases} [\tau_3 \|s_k\|_2, \tau_4 \Delta_k] & \text{if } r_k < \tau_2, \\ [\Delta_k, \tau_1 \Delta_k] & \text{otherwise.} \end{cases}$$

4. Update B_{k+1} and define $\Psi_{k+1}(x)$.

For different trust-region algorithms, $Pred_k$ and $\Psi_k(x)$ have to be chosen accordingly. For the L_∞ trust-region algorithm of Yuan [35], we have that

$$\Psi_k(x) := f(x) + \sigma_k \|g(x)\|_\infty$$

and

$$Pred_k := \Phi_k(0) - \Phi_k(s_k),$$

where $\Phi_k(d)$ is defined in Equation (27). The general requirements for $Pred_k$ are that it is always positive and it is a good approximation to the actual reduction in the merit function, particularly near the solution.

Unless the merit function $\Psi_k(x)$ is a differentiable penalty function, Maratos-effect will occur. Thus, it is necessary to try a second-order correction step $\hat{s}_k$ whenever the standard trust region trial step s_k is not acceptable. Normally, the second order correction step $\hat{s}_k$ is computed by solving another subproblem, which is a slight modification of the trust subproblem for obtaining s_k . For example, if s_k is obtained by Equation (27), we can compute $\hat{s}_k$ by solving the following Second-order correction subproblem,

$$\begin{aligned} \text{minimize } &\frac{1}{2} (s_k + d)^T B_k (s_k + d) + \nabla f(x_k)^T d \\ &+ \sigma_k \|\nabla g(x_k)^T d + g(x_k + s_k)\|_\infty \quad (28) \\ d \in \mathbb{R}^n : &\|s_k + d\|_\infty \leq \Delta_k. \end{aligned}$$

As the new subproblem differs from the original subproblem only by a term $g(x_k + s_k) - g(x_k) - \nabla g(x_k)^T s_k$ which is of magnitude of $\|s_k\|^2$, one can easily show that $\|\hat{s}_k\| = O(\|s_k\|^2)$, hence adding $\hat{s}_k$ will retain the nice superlinearly convergent property of the SQP step s_k . More important is that the second-order correction step $\hat{s}_k$ will always make the merit function accept the new point $x_k + s_k + \hat{s}_k$ when x_k is close to the solution, which was first proved by Yuan [38].

Now, we give a brief discussion about inequality constraints. The inequality constraints in Equation (1) can be transformed into equality constraints

$$g(x) - s = 0$$

by introducing slack variables $s = (s_1, \dots, s_m)^T \in \mathbb{R}^m$, with the additional nonnegative constraints:

$$s \geq 0. \quad (29)$$

The nonnegative constraints (29) can be handled by the log-barrier function or by scaled trust-region techniques, for example, see Dennis *et al.* [39], Coleman and Li [40], and Byrd *et al.* [41].

For detailed discussions about trust-region SQP methods, we recommend the related chapters in the monograph by Conn, Gould and Toint [42] and the review article of Yuan [43].

STABILIZATION BY A FILTER

Filter method was first given by Fletcher and Leyffer [44]. This approach is based on viewing the constrained optimization problem (1) as two separate minimization problems,

$$\begin{aligned} &\text{minimize } f(x) \\ &x \in \mathbb{R}^n \end{aligned}$$

and

$$\begin{aligned} &\text{minimize } h(g(x)) \\ &x \in \mathbb{R}^n, \end{aligned}$$

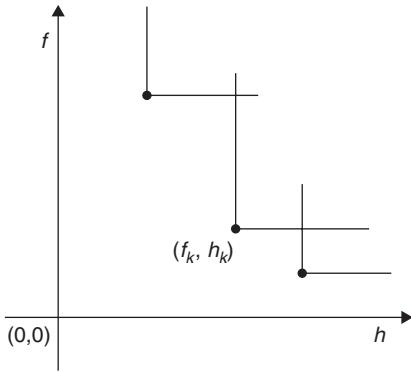
where

$$h(g(x)) := \|g^-(x)\|_1 = \sum_{j=1}^m |\min(0, g_j(x))|.$$

For each iterate point x_k , there is a corresponding pair

$$(f_k, h_k) = (f(x_k), \|g^-(x_k)\|_1),$$

see the subsequent sketch.



The concept of domination is crucial for the filter method.

Definition 1. A pair (f_k, h_k) is said to dominate another pair (f_l, h_l) if and only if both $f_k < f_l$ and $h_k < h_l$.

The idea of the filter approach is that a new point is acceptable as long as it is not dominated by any previous iterates. Such an acceptable point will be added to the filter which is used to control subsequent iterates.

Definition 2. A filter is a list of pairs (f_l, h_l) such that no pair dominates any other. A point (f_k, h_k) is said to be acceptable for inclusion in the filter if it is not dominated by any point in the filter.

At each iteration, a Filter-SQP algorithm needs to construct a QP subproblem. Fletcher and Leyffer [44] uses the subproblem

$$\begin{aligned} &\text{minimize } \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \\ &d \in \mathbb{R}^n : \nabla g(x_k)^T d + g(x_k) \geq 0 \quad (30) \\ &\|d\|_\infty \leq \Delta_k. \end{aligned}$$

The outlines of a general filter-SQP algorithm can be stated as follows.

Algorithm 3. (A general filter-SQP algorithm)

Given $x_0 \in \mathbb{R}^n$, let $k := 1$. Let the initial filter be the pair $\{(f_0, h_0)\}$.

For $k = 0, 1, 2, \dots$ until convergence do

1. Solve the QP subproblem obtaining d_k .
2. Provisionally set $x_{k+1} := x_k + d_k$.
3. If (f_{k+1}, h_{k+1}) is dominated by some pair in the filter then $x_{k+1} := x_k$. Otherwise, accept x_{k+1} , add (f_{k+1}, h_{k+1}) to the filter and remove all points dominated by (f_{k+1}, h_{k+1}) from the filter.
4. Construct the QP subproblem for the next iteration.

It is well known that when using a penalty function to ensure global convergence of nonlinear optimization algorithms, we normally require some sufficient decrease in the penalty function. Similarly, in the implementation of a filter method, the simple domination has to be replaced by some sufficient domination. Fletcher and Leyffer

[44] suggest to accept the provisional point x_{k+1} to the filter if either

$$h_{k+1} \leq 0.99h_l$$

or

$$f_{k+1} \leq f_k - \max(0.25\Delta q_l, 10^{-4}h_l\mu_l)$$

holds for all filter entries l , where $\Delta q_l := -\frac{1}{2}d_l^T B_l d_l - \nabla f(x_l)^T d_l$ is the predicted reduction of $f(x)$ based on the quadratic model, λ_l is the Lagrange multiplier of Equation (30), and μ_l is the projection of the least power of 10 larger than $\|\lambda_l\|_\infty$ in the interval $[10^{-6}, 10^6]$.

The trust-region QP subproblem (30) may become infeasible, either because the trust-region radius Δ_k is too small or due to the inconsistency of the linearized constraints. Thus, a restoration phase has to be used to move the iterate toward the feasible region. This is achieved by minimizing the norm of the residuals of the linearized constraints in the trust region.

CONVERGENCE

There remains the question whether the convergence of an SQP method can be proved in a mathematically rigorous way. In fact there exist numerous theoretical convergence results in the literature; see for example, Boggs and Tolle [7]. We give here only an impression about the type of these statements, and repeat some results that have been stated in the early days of the SQP methods.

In the first case, we consider the global convergence behavior, that is, the question, whether the SQP methods converges when starting from an arbitrary initial point. Suppose that the augmented Lagrangian merit function (18) is implemented and that the primal and dual variables are updated in the form (15).

Theorem 1. *Let $\{(x_k, v_k)\}$ be a bounded iteration sequence of the SQP algorithm with a bounded sequence of quasi-Newton matrices $\{B_k\}$ and assume that*

$$d_k^T B_k d_k \geq \gamma d_k^T d_k \quad (31)$$

for all k and a $\gamma > 0$. Then there is an accumulation point of $\{(x_k, v_k)\}$ satisfying the KKT conditions for (1).

The statement of the theorem is quite weak. But without any further information about second derivatives, we cannot guarantee that the approximated point is indeed a local minimizer.

To investigate now the local convergence speed, we start from an initial $x_0 \in \mathbb{R}^n$ that is sufficiently close to an optimal solution. Normally, we find the following assumptions for local convergence analysis.

Assumption 1. Assume that

1. $x_k \rightarrow x^*$,
2. $f(x), g_1(x), \dots, g_m(x)$ are at least twice continuously differentiable,
3. $g_i(x^*) = 0$ for $i = 1, \dots, m$, that is, we know all active constraints,
4. $\nabla g_i(x^*)$ are linearly independent for $i = 1, \dots, m$, that is, the constraint qualification is satisfied,
5. x^* is a KKT point and u^* is the corresponding Lagrange multiplier,
6. The inequality

$$d^T \nabla_{xx}^2 L(x^*, u^*) d > 0$$

holds for all non-zero vectors d satisfying $\nabla g(x^*)^T d = 0$, that is, a second-order sufficient condition is satisfied.

The following theorem is the fundamental result for local convergence analysis of SQP methods.

Theorem 2. *Assume that the conditions in Assumption 1 are satisfied. Let d_k be the solution of*

$$\text{minimize } \frac{1}{2} d^T B_k d + \nabla f(x_k)^T d \quad (32)$$

$$d \in \mathbb{R}^n : \nabla g(x_k)^T d + g(x_k) = 0.$$

Then, the limit

$$\lim_{k \rightarrow \infty} \frac{\|x_k + d_k - x^*\|}{\|x_k - x^*\|} = 0 \quad (33)$$

holds if and only if

$$\lim_{k \rightarrow \infty} \frac{\|P_k(B_k - \nabla_{xx}^2 L(x^*, u^*)d_k)\|}{\|d_k\|} = 0 \quad (34)$$

where P_k is the projection from $\mathbb{R}^n$ onto the null space of $\nabla g(x_k)^\top$, and $\nabla_{xx}^2 L(x^*, u^*)$ is the Hessian of the Lagrangian at the solution x^* .

This important result was first given by Boggs *et al.* [45] and Powell [46]. A nice proof for this result can be found in Stoer and Tapia [47].

In order to apply Theorem 2 to establish suplinear convergence results for various SQP methods, the general approach is to prove that $x_{k+1} = x_k + d_k$ holds eventually, where d_k is the solution for the QP subproblem (32). Thus, for line search SQP methods we need to prove that the unit step length $\alpha_k = 1$ can be accepted by the corresponding line search conditions. While for trust-region SQP algorithms, one has to show that the trust-region constraint $\|d\| \leq \Delta_k$ will be inactive and the d_k can be accepted by the algorithm for all large k . And, for filter-SQP algorithms, it is required to show that $x_k + d_k$ will be accepted by the filter for all large k .

However, no matter what line search or trust region is used for globalization of the SQP method, due to the Maratos effect, $x_k + d_k$ may not be accepted though Equation (33) is true, when the merit function is a nonsmooth exact penalty function. Even for filter methods, Maratos effect can also happen because a superlinearly convergent SQP step d_k can sometimes increase both the objective function f and the constraint violations $h(g)$. To overcome the Maratos effect, there are many techniques, which can be grouped into four categories. The first one is the watchdog technique, in which standard line search conditions are relaxed at some iterations. The watchdog technique was proposed by Chamberlain *et al.* [48]. The second is the second-order correction technique, where an additional subproblem is solved to obtain an accepted full step, for example, see Fletcher [49], Mayne and Polak [50], and Fukushima [51]. The third is to use a smooth exact penalty function or the augmented Lagrange function as the merit

function; for example, see Schittkowski [21], Powell and Yuan [34,52], and Ulbrich [53]. The fourth is the nonmonotone technique, for example, see Ulbrich and Ulbrich [54], and Gould and Toint [55].

From the characterization of the local convergence property of the SQP method given in Theorem 2, the superlinear convergence of a specific SQP method depends on the relation Equation (34), which is generally hard to verify directly if B_k is updated by quasi-Newton updates. The standard approach is more or less following the sophisticated and elegant technique for analyzing quasi-Newton methods for unconstrained optimization given by Dennis and Moré [56,57]. Their technique is to prove the local Q-superlinear convergence by two steps. The first step is to show that the sequence $\{x_k\}$ converges R-linearly. Then, the R-linear convergence of the sequence is used to obtain more local convergence properties, such as the following relations

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| < \infty, \quad \sum_{k=1}^{\infty} \|x_k - x^*\| < \infty,$$

the uniform boundedness of B_k and B_k^{-1} , and the limit (34), in order to establish the Q-superlinear convergence of the sequence $\{x_k\}$.

However, for a given specific quasi-Newton method, there is still a gap to apply the beautiful convergence results discussed above. For example, it is not easy to know whether condition (31) holds for a particular quasi-Newton updates. Moreover, to ensure the update formulae generate positive definite matrices, we always require $s_k^\top y_k > 0$, a condition that is normally satisfied for unconstrained optimization. However, for constrained optimization problems, as the solution (x^*, u^*) is only a saddle point of the Lagrange function, the vector y_k defined by Equation (9) cannot guarantee $s_k^\top y_k > 0$. Powell [5] suggests replacing y_k by the vector

$$\bar{y}_k = \begin{cases} y_k, & \text{if } s_k^\top y_k > 0.2s_k^\top B_k s_k \\ t_k y_k + (1 - t_k)B_k s_k, & \text{otherwise,} \end{cases} \quad (35)$$

where $t_k = 0.8s_k^\top B_k s_k / (s_k^\top B_k s_k - s_k^\top y_k)$. With this modification, Powell [5] proved the

following remarkable result for the SQP method with BFGS update.

Theorem 3. *Assume that the above assumptions are valid and $\alpha_k = 1$ for all k . Then the sequence $\{x_k\}$ converges R -superlinearly, that is,*

$$\lim_{k \rightarrow \infty} \|x_{k+1} - x^*\|^{1/k} = 0.$$

The R -superlinear convergence speed of $\{x_k\}$ is somewhat weaker than the Q -superlinear convergence rate of $\{z_k := (x_k, v_k)\}$ defined below which was first proved in the form

$$\lim_{k \rightarrow \infty} \frac{\|z_{k+1} - z^*\|}{\|z_k - z^*\|} = 0$$

for the so-called DFP update formula, that is, a slightly different quasi-Newton method. In this case, we get a sequence β_k tending to zero with

$$\|z_{k+1} - z^*\| \leq \beta_k \|z_k - z^*\|.$$

But, it should be noted that the Q -superlinear convergence of z_k does not imply the Q -superlinear convergence of x_k .

Acknowledgment

This article was written during Ya-xiang Yuan's visit at Bayreuth supported by the Alexander-von-Humboldt Foundation. Ya-xiang Yuan would like to thank Humboldt Foundation for the support and thank Klaus Schittkowski for being his host.

REFERENCES

1. Wilson RB. A simplicial algorithm for concave programming [PhD thesis]. Harvard University; 1963.
2. Han S-P. Superlinearly convergent variable metric algorithms for general nonlinear programming problems. *Math Program* 1976;11:263–282.
3. Han S-P. A globally convergent method for nonlinear programming. *J Optim Theory Appl* 1977;22:297–309.
4. Powell MJD. A fast algorithm for nonlinearly constrained optimization calculations. In: Watson GA, editor. Volume 630, *Numerical analysis, Lecture Notes in Mathematics*. Berlin: Springer; 1978.
5. Powell MJD. The convergence of variable metric methods for nonlinearly constrained optimization calculations. In: Mangasarian OL, Meyer RR, Robinson SM, editors. *Nonlinear programming 3*. London, New York: Academic Press; 1978.
6. Schittkowski K. *Nonlinear programming codes*. Volume 183, *Lecture notes in economics and mathematical systems*. Heidelberg: Springer; 1980.
7. Boggs PT, Tolle JW. Sequential quadratic programming. *Acta Numer* 1995;4:1–51.
8. Gould NIM, Toint PhL. SQP methods for large-scale nonlinear programming. *Proceedings of the 19th IFIP TC7 Conference on System Modelling and Optimization: Methods, Theory and Applications*. Cambridge; 1999. pp. 149–178.
9. Fletcher R. *Practical methods of optimization*. Chichester: John Wiley & Sons, Ltd.; 1987.
10. Gill PE, Murray W, Wright M. *Practical optimization*. London: Academic Press; 1982.
11. Sun WY, Yuan Y. *Optimization theory and methods: nonlinear programming*. New York: Springer; 2006.
12. Stoer J. Foundations of recursive quadratic programming methods for solving nonlinear programs. In: Schittkowski K, editor. Volume 15, *Computational mathematical programming, NATO ASI Series, Series F: Computer and Systems Sciences*. New York: Springer; 1985.
13. Papalambros PY, Wilde DJ. *Principles of optimal design*. Cambridge: Cambridge University Press; 1988.
14. Edgar TF, Himmelblau DM. *Optimization of chemical processes*. New York: McGraw Hill; 1988.
15. Gill PE, Murray W, Saunders MA. Methods for computing and modifying the LDL factors of a matrix. *Math Comput* 1975;29:1051–1077.
16. Nocedal J, Overton ML. Projected Hessian update algorithms for nonlinear constrained optimization. *SIAM J Numer Anal* 1985;22:821–850.
17. Byrd RH. An example of irregular convergence in some constrained optimization methods that use projected Hessian. *Math Program* 1985;32:232–237.

18. Yuan Y. An only 2-step Q-superlinear convergence example for some algorithm that used reduced Hessian approximation. *Math Program* 1985;32:269–285.
19. Maratos N. Exact penalty function algorithms for finite dimensional and control optimization problems [PhD Thesis]. Imperial College of Science Technology, University of London; 1978.
20. Gill PE, Murray W, Saunders MA, *et al.* Some theoretical properties of an augmented Lagrangian merit function. In: Pardalos PM, editor. *Advances in optimization and parallel computing*. Amsterdam, New York, Oxford: North Holland Publishing Co.; 1992. pp. 101–129.
21. Schittkowski K. On the convergence of a sequential quadratic programming method with an augmented Lagrangian search direction. *Optimization* 1983;14:197–216.
22. Rockafellar T. Augmented Lagrangian multiplier functions and duality in nonconvex programming. *SIAM J Control* 1974;12:268–285.
23. Wolfe P. Convergence conditions for ascent methods. *SIAM Rev* 1969;11:226–235.
24. Armijo L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac J Math* 1966;16:1–3.
25. Dai YH. On the nonmonotone line search. *J Optim Theory Appl* 2002;112:315–330.
26. Dai YH, Schittkowski K. A sequential quadratic programming algorithm with nonmonotone line search. *Pac J Optim* 2008;4: 335–351.
27. Grippo L, Lampariello F, Lucidi S. A nonmonotone line search technique for Newton's method. *SIAM J Numer Anal* 1986;23: 707–716.
28. Toint PL. An assessment of nonmonotone line search techniques for unconstrained optimization. *SIAM J Sci Comput* 1996;17:725–739.
29. Toint PL. A nonmonotone trust-region algorithm for nonlinear optimization subject to convex constraints. *Math Program* 1997;77:69–94.
30. Vardi A. A trust region algorithm for equality constrained minimization: convergence properties and implementation. *SIAM J Numer Anal* 1985;22:575–591.
31. Byrd RH, Schnabel RB, Shultz GA. A trust region algorithm for nonlinearly constrained optimization. *SIAM J Numer Anal* 1987;24:1152–1170.
32. Omojokun EO. Trust region algorithms for optimization with nonlinear equality and inequality constraints [PhD thesis]. University of Colorado at Boulder, USA; 1989.
33. Celis MR, Dennis JE Jr, Tapia RA. A trust region algorithm for nonlinear equality constrained optimization. In: Boggs PT, Byrd RH, Schnabel RB, editors. *Numerical optimization*. Philadelphia (PA): SIAM; 1985. pp. 71–82.
34. Powell MJD, Yuan Y. A trust region algorithm for equality constrained optimization. *Math Program* 1991;49:189–211.
35. Yuan Y. On the convergence of a new trust region algorithm. *Numer Math* 1995;70: 515–539.
36. Niu LF, Yuan Y. A new trust-region algorithm for nonlinear constrained optimization. *J Comput Math* 2010;28:72–86.
37. Fletcher R. A model algorithm for composite non-differentiable optimization problems. *Math Program* 1982;17:67–76.
38. Yuan Y. On the superlinear convergence of a trust region algorithm for nonsmooth optimization. *Math Program* 1985;31:269–285.
39. Dennis JE, Heinkenschloss M, Vicente LN. Trust-region interior-point SQP algorithms for a class of nonlinear programming problems. *SIAM J Control Optim* 1998;36:1750–1794.
40. Coleman TF, Li Y. A trust region and affine scaling interior point method for nonconvex minimization with linear inequality constraints. *Math Program* 2000;88:1–31.
41. Byrd RH, Gilbert JC, Nocedal J. A trust region method based on interior point techniques for nonlinear programming. *Math Program* 2000;89:149–185.
42. Conn AR, Gould NIM, Toint PhL. *Trust region methods*, MPS/SIAM Series on Optimization. Philadelphia (PA): SIAM; 2000.
43. Yuan Y. A review of trust region algorithms for optimization. In: Ball JM, Hunt JCR, editors. *ICM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics*. Oxford: Oxford University Press; 2000. pp. 271–282.
44. Fletcher R, Leyffer S. Nonlinear programming without a penalty function. *Math Program* 2002;91:239–269.
45. Boggs PT, Tolle JW, Wang P. On the local convergence of quasi-Newton methods for constrained optimization. *SIAM J Control Optim* 1982;20:161–171.
46. Powell MJD. Variable metric methods for constrained optimization. In: Bachem A,

- Grötschel M, Korte B, editors. Mathematical programming: the state of the art. Berlin: Springer; 1983. pp. 288–311.
47. Stoer J, Tapia RA. On the characterization of q-superlinear convergence of quasi-Newton methods for constrained optimization. Math Comput 1987;49:581–584.
48. Chamberlain RM, Lemarechal C, Pedersen HC, *et al.* The watchdog technique for forcing convergence in algorithms for constrained optimization. Math Program Study 1982;16:1–17.
49. Fletcher R. Second order correction for non-differentiable optimization. In: Watson GA, editor. Numerical analysis. Berlin: Springer; 1982. pp. 85–115.
50. Mayne DQ, Polak E. A superlinearly convergent algorithm for constrained optimization problems. Math Program Study 1982;16:45–61.
51. Fukushima M. A successive quadratic programming algorithm with global and superlinear convergence properties. Math Program 1986;35:253–264.
52. Powell MJD, Yuan Y. A recursive quadratic programming algorithm that uses differentiable exact penalty function. Math Program 1986;35:265–278.
53. Ulbrich S. On the superlinear local convergence of a filter-SQP method. Math Program 2004;100:217–245.
54. Ulbrich M, Ulbrich S. Non-monotone trust region methods for nonlinear equality constrained optimization without a penalty function. Math Program Ser B 2003;95:103–135.
55. Gould NIM, Toint PhL. Global convergence of a non-monotone trust-region filter algorithm for nonlinear programming. In: Hager WW, Huang S, Pardalos PM, *et al.*, editors. Multi-scale optimization methods and applications. New York: Springer; 2006. pp. 125–150.
56. Dennis JE, Moré JJ. A characterization of superlinear convergence and its application to quasi-Newton methods. Math Comput 1974;28:549–560.
57. Dennis JE, Moré JJ. Quasi-Newton methods, motivation and theory. SIAM Rev 1977;19:46–89.

SERIES SYSTEMS

MINJAE PARK
HOANG PHAM
Rutgers University, The State
University of New Jersey,
New Brunswick, New Jersey

INTRODUCTION

There are many ways in which a system can be configured to function properly. It could be as simple as a series system, parallel system, or as complex in structure as a network. Once the system is configured in a proper manner, then it is evaluated to test the reliability to meet the performance criteria and the acceptable reliability level. In this article, we study the series system and evaluate its reliability.

In a traditional gasoline car, for the engine to work, we need several different parts to operate in sequence. In detail, a piston in the engine starts from the top and is pushed down to intake the mixture of air and gasoline into a cylinder. Then the piston is pushed back up to compress the mixture. When the piston is at the top, the spark plug functions to ignite the gasoline, causing it to explode in the cylinder; it then pushes the piston back down. When the piston is back at the bottom, the valve opens and lets the air and gas out. As can be seen from the above process, if any of the parts fail to operate, we are not able to start the engine and hence the car does not move. Similar real life examples can be found in systems such as refrigerator, flash lights, computer, and so on. In general, a series system functions successfully only when all the components function properly. This indicates that every component should work successfully in order for the system to work. Consider a series system consisting of q components as shown in Fig. 1.

The above shows that a system with q components is connected in series such that, if any of the component fails to function, this

will result in the failure of function of the next component. Therefore, for a series system to succeed, all of its components must function well during the intended mission time of the system.

Since all components of the entire system should successfully operate in a series system, the reliability of this series system is the probability that all components succeed. This implies that all q components in the system must succeed. Let R_s be the reliability of the system, Z_i be the indicator where the i th component functions (does not fail), and $P(Z_i)$ be the probability that the i th component is operational. The reliability of the system is then expressed as

$$\begin{aligned} R_s &= P(Z_1 \cap Z_2 \cap \dots \cap Z_q) \\ &= P(Z_1) \cdot P(Z_2|Z_1) \cdot P(Z_3|Z_1 \cdot Z_2) \cdots \\ &\quad P(Z_q|Z_1 \cdot Z_2 \cdots Z_{q-1}). \end{aligned} \quad (1)$$

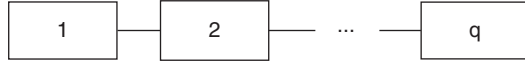
Since the above equation is expressed as a conditional probability, this also includes the case in which the previous failure impacts the following failure rates. However, if each component's failure rate is independent, then the equation can be written as

$$\begin{aligned} R_s &= P(Z_1) \cdot P(Z_2) \cdot P(Z_3) \cdots P(Z_q) \\ &= \prod_{i=1}^q P(Z_i). \end{aligned} \quad (2)$$

Additionally, for a simple series system, the reliability of the system can be calculated from the reliability of its individual components [1].

$$R_s = \prod_{i=1}^q R_i. \quad (3)$$

We are going to develop the warranty cost model for the series system. First, we briefly review the warranty policy. In general, warranty is an obligation attached to products that require the manufacturers to provide compensation for consumers according to the

Figure 1. Series system with q components.

warranty terms when the warranted products fail to perform their intended functions [2]. The warranty is nowadays important not only to the manufacturers but also to the buyers of any commercial product. The consumer is provided a resource for dealing with items that fail due to the uncertainty of product performance. The manufacturer is provided protection because warranty terms explicitly limit responsibility in terms of both time and type of product failure.

To develop the cost model for the series system, several assumptions have been made in this article as follows: (i) The repair cost and the replacement cost are constant. (ii) Repair service time and replacement service time are negligible. (iii) All warranty claims are executed and all claims are valid. (iv) Repairs are imperfect and the repair process can be modeled by a quasi-renewal process (QRP). (v) The time to perform an inspection in which to determine whether the failed component needs repair or a replacement is negligible. (vi) The warranty period is renewable. (vii) Each component's failure rate is independent in the series system.

BACKGROUND

Many researchers [3–5] have studied the optimization of system reliability using different techniques and have investigated various characteristics and reliability of the series system. Duarte *et al.* [3] developed the method to solve the problem of maintenance management of a series system based on preventive maintenance over the different system components. Nahas and Nourelfath [4] studied a reliability optimization problem for a series system to maximize the system reliability subject to the system budget. Xu *et al.* [5] proposed a computational approach for the reliability redundancy optimization problem. Warranty models and their properties have been investigated by many researchers [6–8]. In Blischke and Murthy's books [6,7], general descriptions of various

kinds of warranty policies and mathematical models can be found. Park and Pham [8] developed the warranty cost models, reliability, and other measures for k -out-of- n systems using mixed and an altered QRPs.

To analyze the combination of the series system to the warranty, we use two concepts from the classic QRP, which are the altered QRP and the mixed QRP. The classic QRP model shows that the time to failure is reduced to a fraction α of the previous repair interval and this means that it is dependent of all repair intervals from the prior period. However, it is not easy to find such examples in real life. Therefore, to overcome this unpractical preposition, we develop an altered QRP with different parameters for interfailure intervals. Another alternate QRP is the mixed QRP. This allows an immediate replacement for a malfunctioning product within a warranty period in the case of a premature and severe malfunction despite the number of failures being less than a threshold level. On the basis of the two above-mentioned concepts, cost analysis for a series system is conducted.

In case the warranty policy is renewable, the distribution function of the number of failures can be obtained. If a product fails to perform within the warranty period w , the warranty policy is renewable for the same period w .

ALTERED QUASI-RENEWAL PROCESSES

Because the repair is assumed to be imperfect, the repair process can be modeled by a QRP. We first define QRP. On the basis of Wang and Pham's QRP [9], we present two alternate QRPs: the altered QRP and mixed QRP. The altered QRP is changing a parameter in the equations, and the mixed QRP is mixing repair intervals and replacement intervals. Then we investigate the distribution of the number of failures and develop the cost model for a single-component system in terms of the expected warranty cost and

the variance of the warranty cost. Finally, we develop the cost model for the series system.

Definition [9]. Let X_n be the interfailure interval between the $(n-1)$ th and n th events of the process. The counting process $\{N(t), t > 0\}$ is a *QRP* associated with the distribution F and the constant parameter α , $\alpha > 0$, if $X_n = \alpha^{n-1} \cdot Z_n$, $n = 1, 2, \dots$ where Z_n s are i.i.d. and $Z_n \sim F$.

The probability density function (pdf), cumulative distribution function (cdf), and failure rate, respectively, for $n = 2, 3, 4, \dots$ are given by

$$\begin{aligned} f_n(x) &= \frac{1}{\alpha^{n-1}} f_1\left(\frac{1}{\alpha^{n-1}}x\right) \\ F_n(x) &= F_1\left(\frac{1}{\alpha^{n-1}}x\right) \\ h_n(x) &= \frac{1}{\alpha^{n-1}} h_1\left(\frac{1}{\alpha^{n-1}}x\right). \end{aligned} \quad (4)$$

Altered Quasi-Renewal Process

From the classic QRP, each interoccurrence interval has a common pattern like $X_2 = \alpha \cdot X_1$, $X_3 = \alpha^2 \cdot X_1$, $X_4 = \alpha^3 \cdot X_1, \dots, X_n = \alpha^{n-1} \cdot X_1$. Let X_1 represent the first failure interval. This means that the failure time will be reduced with a fraction α . In reality, such a specific pattern would be difficult to observe. We now discuss an altered QRP that does not need to assume such a pattern.

Definition [8]. A counting process $\{N(t), t > 0\}$ is said to be the altered QRP associated with the distribution F and the random parameter $\beta_n, \beta_n > 0$, a constant, if $X_n = \beta_n \cdot X_1, n = 2, 3, \dots$ where $X_1 \sim F$ and β_n , where $n = 2, 3, \dots$ are not necessarily equal.

Let f_i and F_i be the pdf and cdf of X_i . For the altered QRP, the cdf and pdf for $n = 2, 3, 4, \dots$ are, respectively, given by

$$\begin{aligned} F_n(x) &= F_1\left(\frac{1}{\beta_n}x\right) \\ f_n(x) &= \frac{1}{\beta_n} f_1\left(\frac{1}{\beta_n}x\right). \end{aligned} \quad (5)$$

Additionally, the component j 's cdfs and pdfs of interoccurrence interval i , respectively, for $i = 2, 3, 4, \dots$ are given by

$$\begin{aligned} F_{ij}(x) &= F_{1j}\left(\frac{1}{\beta_{ij}}x\right) \\ f_{ij}(x) &= \frac{1}{\beta_{ij}} f_{1j}\left(\frac{1}{\beta_{ij}}x\right). \end{aligned} \quad (6)$$

Mixed Quasi-Renewal Processes

Bai and Pham [10] proposed a warranty model addressing the relationships between the number of repairs and replacements. Let h be an upper limit of the number of repairs under the warranty. Let N_a and N_b be the number of repair and replacement within a warranty period respectively. Similarly, let $N_a(t)$ and $N_b(t)$ be the number of repairs and the number of replacements at random variable time T respectively. w denotes a warranty period assuming that N_a and N_b are correlated and $\text{cov}(N_a, N_b) \neq 0$, then

$$\begin{aligned} N_a(w) < h &\rightarrow N_b(w) = 0, \\ N_b(w) > 0 &\rightarrow N_a(w) = h. \end{aligned} \quad (7)$$

This implies that, if there are less than h failures, the replacements would not occur, instead, only repair services would be provided. If there are h failures, a replacement would occur. In other words, if there are more than h system failures within w , the failed product will be replaced instead of being repaired again. The *repair-limit risk-free* warranty policies of a fixed period w , where a replacement only happens after h failures under a warranty period, are suggested [10]. However, in this article, we consider that the replacement would happen if the system failure is premature and severe because the likelihood of it occurring in real life is higher. Therefore, Equation (7) was altered to Equation (8) for the mixed QRP as follows:

$$\begin{aligned} N_a(w) < h &\rightarrow N_b(w) \geq 0, \\ N_b(w) > 0 &\rightarrow N_a(w) \leq h. \end{aligned} \quad (8)$$

In this model, even if there are less than h failures, the replacement strategies can

happen using mixed QRP. In the case of a failure of function of a product, either repair or replacement may be provided on the basis of the failure condition, and the manufacturer should select which services would be provided. Additionally, in general, the superposition of any two renewal processes is not a renewal process. Since we assume that the repairs and replacements would not happen simultaneously, they could be the renewal processes.

COST ANALYSIS

In this section, warranty cost analyses are elaborated through computation of both the expected value and the variance of the warranty cost. Also, the distribution of N and its properties per period and several statistical properties of warranty cost function per cycle or per product sold are shown.

Distribution of $N_s(t)$

Let $N_s(t)$ be the number of system failures at random variable time T within a warranty period w . Let $f_s(\cdot)$, $F_s(\cdot)$, and $R_s(\cdot)$ be the pdf, cdf, and reliability function respectively of system failure times within a warranty period w . We assume that the warranty policy is renewable after each repair/replacement. This section derives the distribution of the number of system failures within the warranty period. Under the perfect repair assumption, we can obtain the probability mass function (pmf) of N_s as follows [11]:

$$P[N_s = n_s] = [F_s(w)]^{n_s} R_s(w), \quad \text{for } n_s = 1, 2, 3, \dots \quad (9)$$

Lemma 1 [8]. *Let $f_{is}(\cdot)$, $F_{is}(\cdot)$, and $R_{is}(\cdot)$ be the pdf, cdf, and reliability function respectively of system failure times after $(i-1)$ th repair/replacement within a warranty period w . Under the imperfect repair, every $F_{is}(w)$ is different, so the pmf of N_s is given by*

$$P[N_s = n_s] = \left(\prod_{i=1}^{n_s} F_{is}(w) \right) \left(R_{(n_s+1)s}(w) \right). \quad (10)$$

We now discuss the cost analyses for a single-component system considering the interfailure intervals approach. From Ref. 8, the reliability function of a single-component system, which has n number of failures in the warranty period w , is given by

$$R_s(w) = \prod_{i=1}^n (R_{is}(w)) = \left(1 - \int f_{1s}(x) dx \right) \times \prod_{i=2}^n \left(1 - \int \frac{1}{\beta_{is}} f_{is} \left(\frac{1}{\beta_{is}} x \right) dx \right). \quad (11)$$

Using Equations (10) and (11), the pdf of the system failure times, $f(n)$, under the warranty is given by

$$\begin{aligned} f(n) &= \left(\prod_{i=1}^n F_{is}(w) \right) \left(R_{(n+1)s}(w) \right) \\ &= \left(\int f_{1s}(x) dx \right) \left(\prod_{i=2}^n \left(\int \frac{1}{\beta_{is}} f_{is} \left(\frac{1}{\beta_{is}} x \right) dx \right) \right) \\ &\quad \times \left(1 - \int \frac{1}{\beta_{(n+1)s}} f_{1s} \left(\frac{1}{\beta_{(n+1)s}} x \right) dx \right). \end{aligned} \quad (12)$$

The expected number of warranty repair services can be obtained as follows:

$$\begin{aligned} E(N_a) &= \int n_a \cdot f(n_a) \cdot dn_a = \int n_a \cdot \left(\int f_{1s}(x) dx \right) \\ &\quad \times \prod_{i=2}^{n_a} \left(\int \frac{1}{\beta_{is}} f_{is} \left(\frac{1}{\beta_{is}} x \right) dx \right) \\ &\quad \times \left(1 - \int \frac{1}{\beta_{(n+1)s}} f_{1s} \left(\frac{1}{\beta_{(n+1)s}} x \right) dx \right) \cdot dn_a. \end{aligned} \quad (13)$$

Therefore, the variance of the number of repaired services can be obtained by

$$\text{Var}(N_a) = E(N_a^2) - [E(N_a)]^2, \quad (14)$$

where $E(N_a^2) = \int n_a^2 \cdot f(n_a) \cdot dn_a$, and $f(n_a)$ is given in Equation (12) and $E(N_a)$ is given in Equation (13).

Similar to Equation (13), we obtain the expected number of replacement warranty services

$$E(N_b) = \int n_b \cdot \left(\prod_{i=1}^{n_b} \left(\int f_{is}(y) dy \right) \right) \left(1 - \int f_{is}(y) dy \right) \cdot dn_b. \quad (15)$$

The variance of the number of replacement services could be obtained.

Let c_a and c_b be the repair cost per failure and replacement cost per failure respectively. For the warranty cost per product sold, C , the expected warranty cost is given by

$$E(C) = c_a E(N_a) + c_b E(N_b), \quad (16)$$

where $E(N_a)$ is given in Equation (13) and $E(N_b)$ is given in Equation (15).

Note that

$$\begin{aligned} E[N_a N_b] &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b P[N_a = n_a, N_b = n_b] \\ &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b P[N_s = n_s] \\ &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h n_a n_b \left(\prod_{i=1}^{n_a+n_b} (F_{is}(w)) \right) \\ &\quad \times (R_{(n_a+n_b+1)s}(w)). \end{aligned} \quad (17)$$

The covariance of the number of repair services and the replacement services can be obtained as follows:

$$\text{Cov}(N_a, N_b) = E[N_a N_b] - E[N_a]E[N_b], \quad (18)$$

where $E[N_a]$, $E[N_b]$, and $E[N_a N_b]$ are given in Equations (13), (15), and (17).

The variance of the warranty system cost is given by

$$\begin{aligned} \text{Var}(C) &= c_a^2 \cdot \text{Var}(n_a) + c_b^2 \cdot \text{Var}(N_b) \\ &\quad + 2c_a c_b \cdot \text{Cov}(N_a, N_b), \end{aligned} \quad (19)$$

where $\text{Var}(N_a)$ and $\text{Cov}(N_a, N_b)$ is given in Equations (14) and (18) and $\text{Var}(N_b)$ can be obtained by a similar way for $\text{Var}(N_a)$.

Warranty Cost Analyses for Series Systems

In this section, warranty cost analyses are conducted by computing the expected warranty cost as well as the variance of the warranty for the series systems based on the interfailure intervals.

A component's reliability function is considered first. We investigate the reliability of each component and its interfailure interval using altered QRP. Let $f_{ij}(\cdot)$, $F_{ij}(\cdot)$, and $R_{ij}(\cdot)$ be the pdf, cdf, and reliability function of component j 's failure times after $(i-1)$ th repair/replacement within a warranty period w , respectively. The reliability function of the j th component and the i th interfailure interval using the altered QRP is as follows:

$$\begin{aligned} R_{ij}(w) &= 1 - \int f_{ij}(x) dx \\ &= 1 - \int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx. \end{aligned} \quad (20)$$

Because there are q components in the series system, its system reliability can be obtained by multiplying each component's reliability. Each component's failure rate is assumed to be independent. Therefore, using Equation (3), the reliability function of the i th interfailure interval for the series system is given by

$$\begin{aligned} R_{is}(w) &= \prod_{j=1}^q R_{ij}(w) \\ &= \prod_{j=1}^q \left(1 - \int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx \right). \end{aligned} \quad (21)$$

From Equation (12), we can obtain the pdf of series systems

$$\begin{aligned} f(n) &= \left(\prod_{i=1}^{n_s} (F_{is}(w)) \right) (R_{(n_s+1)s}(w)) \\ &= \left(\left(1 - \prod_{j=1}^q \left(1 - \int f_{1j}(x) dx \right) \right) \right. \\ &\quad \left. \times \prod_{i=2}^{n_s} \left(1 - \prod_{j=1}^q \left(1 - \int \frac{1}{\beta_{ij}} f_{1j} \left(\frac{1}{\beta_{ij}} x \right) dx \right) \right) \right) \end{aligned}$$

$$\times \left(\prod_{j=1}^q \left(1 - \int \frac{1}{\beta_{(n_s+1),j}} f_{1j} \left(\frac{1}{\beta_{(n_s+1),j}} x \right) dx \right) \right). \quad (22)$$

Using Equation (22), the expected number of warranty repair services is given by

$$E(N_a) = \int n_a \cdot f(n_a) \cdot dn_a. \quad (23)$$

We can easily derive the second moment as follows:

$$E(N_a^2) = \int n_a^2 \cdot f(n_a) \cdot dn_a. \quad (24)$$

Using the first moment and the second moment, the variance of the number of repair services is given by

$$\text{Var}(N_a) = E(N_a^2) - [E(N_a)]^2, \quad (25)$$

where $E(N_a)$ and $E(N_a^2)$ are given in Equations (23) and (24).

Additionally, the covariance of the repair services and the replacement services is given as follows:

$$\text{Cov}(N_a, N_b) = E[N_a N_b] - E[N_a]E[N_b]. \quad (26)$$

Using Equation (17), we also obtain

$$\begin{aligned} E[N_a N_b] &= \sum_{n_b=0}^{\infty} \sum_{n_a=0}^h \left(n_a n_b \left(1 - \prod_{j=1}^q \left(1 - \int f_{1,j}(x) dx \right) \right) \right. \\ &\quad \times \prod_{i=2}^{n_a+n_b} \left(1 - \prod_{j=1}^q \left(1 - \int \frac{1}{\beta_{ij}} f_{1,j} \left(\frac{1}{\beta_{ij}} x \right) dx \right) \right) \\ &\quad \times \prod_{j=1}^q \left(1 - \int \frac{1}{\beta_{(n_a+n_b+1)j}} \right. \\ &\quad \left. \left. \times f_{1,j} \left(\frac{1}{\beta_{(n_a+n_b+1)j}} x \right) dx \right) \right). \end{aligned} \quad (27)$$

Similarly, we can obtain the expected number of replacement services $E(N_b)$ and the variance of number of the replacement

service $\text{Var}(N_b)$. Using $E(N_a)$ and $E(N_b)$, the expected warranty cost for the series system is given by

$$E(C) = c_a E(N_a) + c_b E(N_b). \quad (28)$$

Combining Equations (25) and (26), the variance of warranty system cost is given by

$$\begin{aligned} \text{Var}(C) &= c_a^2 \cdot \text{Var}(N_a) + c_b^2 \cdot \text{Var}(N_b) \\ &\quad + 2c_a c_b \cdot \text{Cov}(N_a, N_b). \end{aligned} \quad (29)$$

NUMERICAL EXAMPLES AND SENSITIVITY ANALYSIS

In this section, a numerical example is presented to illustrate the analyses of system cost functions for the series system. We assume that the failure time of the system follows an exponential distribution with constant failure rate $\lambda = 1$. Given $q = 2$ (see Fig. 2) and $h = 1$. We assume that the warranty costs and β values are as follows:

$$c_a = \$5,000, c_b = \$8,000 \text{ and}$$

$$\beta_{i1} = 0.99, \beta_{i2} = 0.98.$$

The results are given in Table 1 including the expected warranty cost, $E(C)$, standard deviation, $\text{SD}(C)$, and coefficient of variation (CV) for various values of w .

In Fig. 3, the expected warranty cost increases as warranty time increases. When there exists a covariance between the repair and the replacement services, it can be seen that both the expectation and the standard deviation of warranty cost functions increase monotonically over the warranty period. It would be expected that the standard deviation in this particular case is larger than the expected warranty cost since the distribution of n is coming from a geometric distribution whose standard deviation is



Figure 2. Series system with two components.

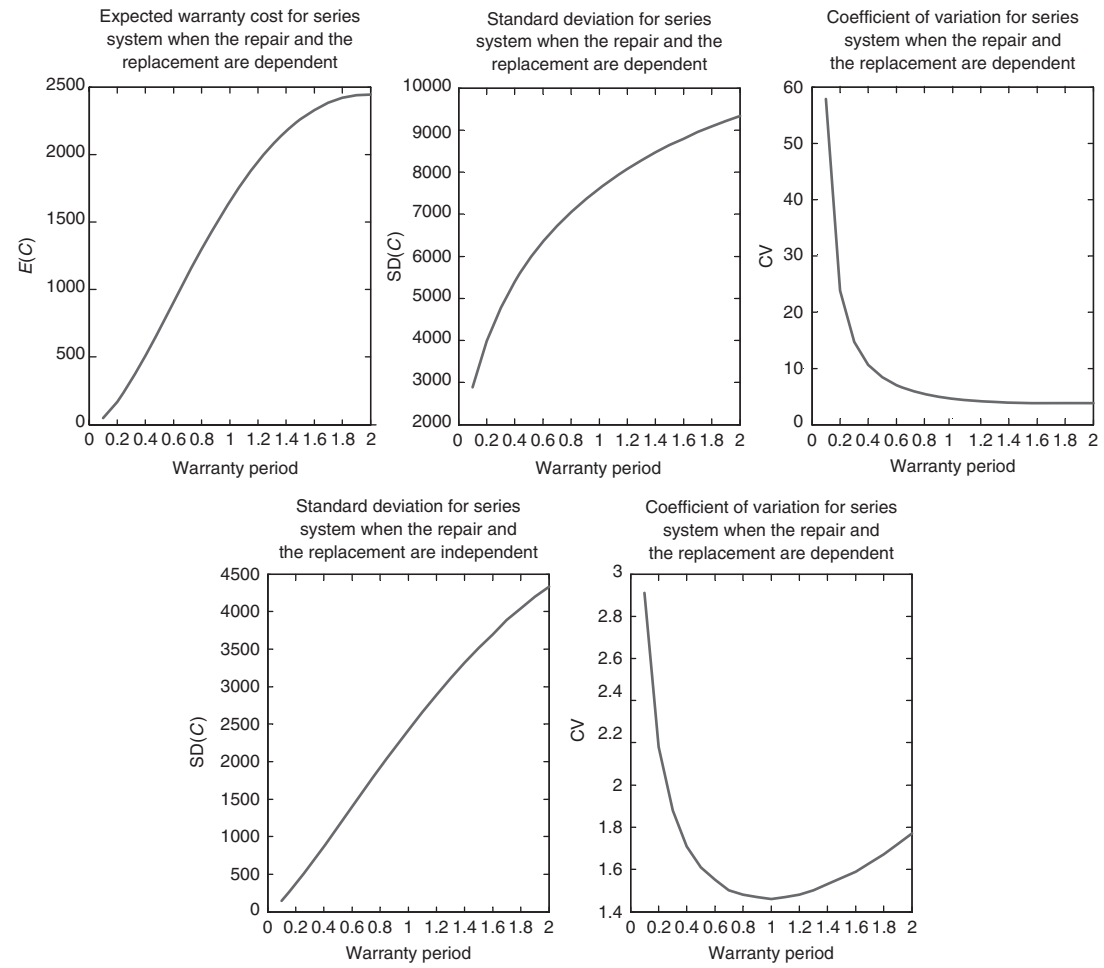


Figure 3. $E(C)$, $SD(C)$, and CV for the series system when the repair and the replacement are dependent or independent.

Table 1. $E(C)$, $SD(C)$, and CV for Series System

W	$E(C)$	SD^a	CV^a	SD^b	CV^b
0.1	50	2890	57.87	145	2.91
0.2	168	3990	23.77	366	2.18
0.3	326	4776	14.64	614	1.88
0.4	509	5396	10.61	872	1.71
0.5	704	5907	8.39	1135	1.61
0.6	905	6342	7.01	1399	1.55
0.7	1104	6719	6.09	1660	1.50
0.8	1297	7050	5.44	1660	1.28
0.9	1480	7345	4.96	2171	1.47
1	1651	7611	4.61	2418	1.46
1.1	1806	7853	4.35	2657	1.47
1.2	1946	8073	4.15	2887	1.48
1.3	2068	8276	4.00	3108	1.50
1.4	2173	8464	3.90	3319	1.53
1.5	2260	8638	3.82	3519	1.56
1.6	2329	8795	3.78	3698	1.59
1.7	2382	8948	3.76	3883	1.63
1.8	2418	9087	3.76	3883	1.61
1.9	2439	9215	3.78	4196	1.72
2	2445	9332	3.82	4332	1.77

^aWhen repair and replacement are dependent.^bWhen repair and replacement are independent.

always larger than the expected warranty cost. The CV is defined as the ratio of the standard deviation to the mean and describes the magnitude sample values and the variation within them. When they are independent, the CV has a v-shape curve as the warranty time progresses. When it decreases, it indicates smaller variations.

REFERENCES

1. http://www.weibull.com/SystemRelWeb/series_system.htm.
2. Pham H. Handbook of engineering statistics. Springer; 2006.
3. Duarte J, Craveiro J, Trigo T. Optimization of the preventive maintenance plan of a series components system. Int J Pres Ves Pip 2006;83:244–248.
4. Nahas N, Nourelfath M. Ant system for reliability optimization of a series system with multiple-choice and budget constraints. Reliab Eng Syst Saf 2005;87:1–12.
5. Xu Z, Kuo W, Lin H. Optimization limits in improving system reliability. IEEE Trans Reliab 1990;39:51–60.
6. Blischke W. Warranty Cost Analysis. CRC Press; 1994.
7. Blischke W, Murthy D. Product warranty handbook. CRC Press; 1996.
8. Park M, Pham H. Warranty system-cost analysis using quasi-renewal processes. J Oper Res Soc 2009;45:263–274.
9. Wang H, Pham H. A quasi renewal process and its applications in imperfect maintenance. Int J Syst Sci 1996;27:1055–1062.
10. Bai J, Pham H. Repair-limit risk-free warranty policies with imperfect repair. IEEE Trans Syst Man Cybern 2005;35:765–772.
11. Bai J, Pham H. Cost analysis on renewable full-service warranties for multi-component systems. Eur J Oper Res 2006;168:492–508.

SERVICE OUTSOURCING

YONG-PIN ZHOU
Foster School of Business,
University of Washington,
Seattle, Washington

ZHONG JUSTIN REN
Boston University School of
Management, Boston and
MIT Sloan School of
Management, Cambridge,
Massachusetts

INTRODUCTION

Globalization poses challenges to current businesses. To keep costs down, many business firms have turned to outsourcing. By definition, outsourcing refers to the act “to procure (as some goods or services needed by a business or organization) under contract with an outside supplier” (Merriam–Webster Dictionary). Almost any business process can be outsourced: manufacturing, product design, procurement, customer care, or other services (see ***Supply Chain Outsourcing*** for an in-depth discussion). Various benefits have been reported that fuel the outsourcing growth. Early outsourcing deals focus on cost savings [1,2], but as Kakabadse and Kakabadse [3] note, recently the goal of outsourcing has shifted from cost reduction to value creation. The value additions found by researchers have included improved speed or quality of services [1], access to outside talent, innovation, and new technology, strategic alliances with outsourcing partners [1,2], focus on core missions [4], and greater flexibility [1]. For a detailed discussion on advantages and disadvantages of outsourcing, please see the article titled ***Business Process Outsourcing*** in this encyclopedia.

With information technology, many types of tasks that are traditionally performed in-house can now be outsourced [5]. Indeed, not only is outsourcing here to stay, it is also set to grow. Those projected to grow the fastest in

2008–2009—financial, legal, supply chain, medical, and news services—are all service functions [6]. A Gartner Group study has estimated that the overall global outsourcing market will grow by 8.1% in 2008 [7]. Anand and Aron [8] cite external studies that project that the off-shore business process outsourcing (BPO) industry will grow from \$123.6 billion in 2001 to over \$230 billion in 2015. A big portion of the service outsourcing activities are service related. It is estimated that, in the total US GDP, roughly 84% of value-added activities are service related [5]. Service outsourcing, therefore, plays an important role in the current and future US economic activities.

From a research perspective, however, operations management (OM) research on service outsourcing is less visible, compared with research in manufacturing and physical systems. In this context, we review OM research in the field of service outsourcing, and then discuss some future research opportunities. In particular, we focus on the following three research areas:

1. *Capacity Planning and Supplier Coordination in Service Outsourcing.* Probably the most prominent component in service outsourcing is to plan for adequate capacity. This can be difficult considering the uncertainty involved in a service system, and the heterogeneity in customer requirements and server capabilities. In addition, it requires coordination with the service provider, which in turn is dependent upon economic incentives. We will review papers in both areas, that is, those that analyze capacity systems and aim to improve capacity planning methods, and those that focus on incentives and coordination mechanisms.
2. *Service Outsourcing under Information Asymmetry.* A frequent problem that companies face in service outsourcing is that they lack perfect information

about their potential outsourcing partners. Recent researches have been successful in dealing with such a problem, which we review in depth.

3. *Quality Concerns in Service Outsourcing.* Currently, more and more attention is being paid to the potential pitfalls in service outsourcing. Quality concern is frequently cited as one of the greatest issues in outsourcing. We review researches that study this problem and highlight their findings that can help companies avoid the “quality trap” and maintain a high level of service quality when outsourcing.

Before we delve into reviewing individual papers, a note on the distinction between outsourcing and offshoring is in order. Offshoring refers to a change in location of service facilities, and does not necessarily entail sourcing from outside vendors. For example, companies may choose to setup and operate their own customer service centers in some other foreign countries. In comparison, outsourcing means contracting out business processes to an outside supplier, which may not be located offshore. Please refer to the article titled **Business Process Outsourcing** in this encyclopedia for more discussions and examples distinguishing the two terms.

LITERATURE REVIEW

Capacity Planning and Supplier Coordination

Capacity planning is a perennial topic in the OM service literature, and it is no exception in the OM outsourcing literature. In any outsourcing configuration, the capacity (and other things like effort) at both parties need to be planned and coordinated. This is very often done via a contract mechanism. In this part, many papers have to do with contracting issues in the context of call center outsourcing, but they all use different terminologies. To expedite the discussion, it is necessary to first introduce some common notations.

In a simple one-to-one call center outsourcing setting, two parties are involved: a firm (which we call the *user*) sends all or some of

its customer calls (or emails) to an outside call center (which we call the *vendor*). The vendor hires and trains its own agents to handle the calls, and each successfully handled call provides a direct benefit to the user firm (direct revenue or customer satisfaction). The contract then determines how much payment should be made from the user to the vendor. At the least a typical contract specifies scope of work, deliverables, terms and conditions, and service level agreement (SLA) [9]. Transfer payment can be calculated in a variety of ways. For example, a typical call center outsourcing contract involves at least some forms of pay per time (PPT) or pay per call (PPC) [10]. In the former arrangement, the user pays the vendor based on the agent hours that are spent. This is also referred to as *pay-for-capacity* [11]. In the latter arrangement, the payment is based on the number of calls that are handled. This is also referred to as *piecemeal* [12] or *pay-for-job* [11]. In practice, these contracts are often combined with other incentives or penalties based on call resolution [12], waiting time, or SLA achievement [13,14].

Ren and Zhou [12] and Hasija *et al.* [13] are concerned with the design of a coordinating contract. A contract is said to be *coordinating* if it achieves two goals in the decentralized setting: (i) it induces the vendor to choose system-optimal staffing and effort level (i.e., $s^v = s^I$ and $e^v = e^I$) and (ii) it allows for an arbitrary split of system profit between the user and the vendor. With a coordinating contract, both parties can first achieve the highest total profit, and then share it in such a way that each is better off than under a noncoordinating contract.

Consider the following model of such a simple one-to-one call center outsourcing supply chain. The vendor is modeled as a multiserver queueing system with customer abandonment. The arrival rate is λ , the service rate is μ , and the customer patience (i.e., time waiting in the queue until abandonment) has continuously differentiable PDF f and CDF F . The staffing cost rate is c_s , the customer waiting cost rate is c_w , and each time a customer abandons, there is a cost of c_a . Of the served calls, only a portion, p , are satisfactorily resolved. Each resolved

call results in a revenue of (representing direct sales, customer satisfaction, etc.), but incurs a loss of goodwill c_g for each resolved call (representing the reduction in both immediate and future sales). The resolution rate p is influenced by the agents directly, but ultimately determined by the vendor's effort level, which includes agent training, infrastructure development, technical support, and so on. The vendor's resolution $p(e)$ is assumed to be concave, increasing in its effort level e . The vendor bears the effort cost at a rate of c_e .

Therefore, the vendor must make two decisions: staffing level and effort level, and the user must determine the appropriate terms for the contract to induce the vendor to make system-optimal decisions.

The vendor's $G/GI/s + GI$ queueing model is hard to analyze. Even a simpler $M/M/s + M$ system is hard to analyze [15]. Ren and Zhou [12] follow a fluid approximation approach, which is shown to be remarkably accurate by Whitt [16]. In the fluid approximation, if the vendor's staffing level is s , then the system throughput rate, abandonment rate, and steady-state wait time are respectively, $T(s) = \min(\lambda, \mu s)$, $L(s) = \lambda - T(s)$, and $W(s) = \frac{L(s)}{f(0)\lambda}$ assuming $f(0) \neq 0$.

In the centralized (i.e., first-best) setting, the objective is to maximize total system profit:

$$\begin{aligned} \max_{s,e} \pi^I(s,e) = & rp(e)T(s) - c_s\mu s - c_e e - c_a L(s) \\ & - c_w\lambda W(s) - c_g(1 - p(e))T(s), \end{aligned} \quad (1)$$

and the optimal decisions are $s^I = \lambda/\mu$ and $p'(e^I) = \frac{c_e}{(r+c_g)\lambda}$, if $\frac{c_e}{(r+c_g)\lambda} < p'_0$; $e^I = 0$ otherwise.

In the decentralized outsourcing supply chain, the user offers the vendor a contract to take all of its calls. The vendor decides its staffing and effort levels, and, depending on the service outcome (e.g., T , W , L , etc.), it receives a payment Ψ from the user as specified in the contract. The user receives a revenue r from each served and resolved customer. Because the vendor is invisible to the customers, the user also bears all the costs of customer abandonment, customer waiting,

and unresolved calls. In such a setting, the vendor and user maximize their respective individual profit π^v and π^u :

$$\begin{aligned} \pi^v(s,e) &= E(\Psi) - c_s\mu s - c_e e, \\ \pi^u(s,e) &= rp(e)T(s) - c_a L(s) - c_w\lambda W(s) \\ &\quad - c_g(1 - \bar{p}(e))T(s) - E(\Psi). \end{aligned}$$

Ren and Zhou [12] find that a simple PPC contract, where $\Psi = bT(s)$, can coordinate the vendor's staffing level (i.e., $s^v = s^I$), but it fails to coordinate the vendor's effort level. In fact, the vendor will exert no effort (i.e., $e^v = 0$) at all because it pays for all the effort cost but receives no accordant benefit. A pay-per-call-resolved (PPCR) contract, where $\Psi = bp(e)T(s)$, also coordinates the vendor's staffing level (i.e., $s^v = s^I$). Although it lifts the vendor's effort level, it still falls short of the system-optimal effort level (i.e., $0 < e^v < e^I$). This is intuitive because, while PPCR rewards the vendor for its effort, the benefit resulting from a higher effort level is shared between the user and the vendor. On the other hand, the vendor bears the total cost of effort. This incentive misalignment is similar to the "double marginalization" observed in physical-goods supply chain. To remedy this, Ren and Zhou [12] propose two additional contracts—pay-per-call-resolved plus cost sharing (PPCR+CS) and partnership (PART)—that are shown to coordinate the supply chain (i.e., $s^v = s^I$ and $e^v = e^I$). Under the PPCR+CS contract, $\Psi = bp(e)T(s) + (1 - \alpha)(c_s\mu s + c_e e) - \alpha[c_g(1 - p(e))T(s) + c_a L(s) + \frac{c_w}{f(0)}L(s)]$. It basically lets the user share the vendor's costs so that overall the vendor's incentive is proportional to that of the supply chain. Furthermore, the vendor's share of the total profit is b/r . By varying b , the two parties can achieve a desirable share of the total profit. While PPCR+CS requires more information sharing, which may be difficult to do practically, PART requires less information sharing: the vendor pays the user for each customer arrival but gets to keep all the revenue. In addition, the user shares part of the vendor's staffing cost and gives it a fix effort compensation based on the system optimal e^I : $\Psi = rp(e)T(s) - (1 - p(e))c_g T(s) - (1 - \alpha)[rp(e^I) -$

$c_g(1 - p(e^I))T(s) + (1 - \alpha)c_s\mu s + (1 - \alpha)c_e e^I - \alpha\left(c_a + \frac{c_w}{f(0)}\right)L(s)$. In essence, PART has the flavor of a franchise arrangement.

Instead of the fluid approximation, Hasija *et al.* [13] use an $M/M/N + M$ queueing system to model the vendor's call center and find that neither a PPC nor a PPT contract can coordinate the vendor's staffing level. This difference is due to the fact that, while it is advantageous to be at 100% utilization level in a fluid system, in a queueing system, system utilization must stay strictly below 100% and customer abandonment is convex increasing in the utilization. However, Hasija *et al.* [13] show that by adding additional penalty for violating service level agreements the enhanced PPC and PPT contracts can coordinate the outsourcing chain.

In addition, Hasija *et al.* [13] study the effect of information asymmetry. Specifically, the vendor's effort level is private information, and as a result the user does not know whether the vendor has a high or low productivity (i.e., service rate). In such a case, an optimal contract will have to additionally induce the vendor to self-select and reveal its productivity type. The authors show that coordinating contracts in the full-information case may no longer be optimal. Instead, the authors propose several specific combinations of PPT and/or PPC contracts with additional constraints (e.g., average service time limit, SLAs, and/or wait time penalty) and show them to be optimal. In addition to achieving maximum profit for the system, the authors also discuss how profits can be allocated between the user and the vendor under such contracts.

Milner and Olsen [17] demonstrate the importance of selecting the right SLA and setting the right parameter. It studies a call center outsourcing situation where the vendor handles work not only from the user under a contract but also from other noncontract sources. The vendor call center system is modeled as a multiserver queue and the user requires the vendor to satisfy a constraint on delay percentile (i.e., the percentage of delay that exceeds a limit). To the extent that the vendor has more capacity than necessary, it will also take noncontract work to increase revenue. The problem for

the vendor then becomes how to prioritize calls from these two different sources: the vendor wants to minimize the average wait time for the noncontract jobs, subject to the delay percentile constraint for the contract jobs. Using the heavy traffic regime, where system load approaches 1, as an approximation to the original system, the author shows that it is optimal to use a two-threshold policy where contract jobs are given priority only if the queue length is neither too long nor too short. This result has implications for both the user and the vendor. For the user, even though the delay percentile constraint in the contract is satisfied, the two-threshold policy could seem unfair because contract jobs will be given lower priority when the queue is long. The authors then suggest that the user should add more constraints to the contract, or impose a convex penalty on excess delay (which makes it optimal for the vendor to use a fairer single-threshold policy). Several other contract forms were also studied and found inadequate.

While these three papers study contract issues in a complete outsourcing setting, where the vendor handles all of the user's customer calls, Akş in *et al.* [11] study contract issues in a cosourcing setting where the user outsources only some, but not all, of its calls to the vendor for strategic (customer segmentation) or operational (call volume uncertainty) reasons (this term is also described in Willcocks and Lacity [18]). Cosourcing is often promoted as an alternative to complete outsourcing [19]. Demand at a call center—the call volume—is uncertain and fluctuating in nature, as is the demand at most services. The forecasted call volume pattern can be broken into two parts—the base volume (which is a stable volume over time) and the fluctuation (anything above the base). It is easy and cost effective to staff to the base call volume. However, to achieve any reasonable service level, call centers must plan extra staffing to deal with the fluctuation. In a cosourcing setting, an important question becomes which contract type—contracting the base or contracting the fluctuation—is a better one. Specifically, in the former setting, the user's would staff just enough to deal with the stable base demand, while the vendor

would have to carry safety capacity to handle the fluctuation. (Since the vendor is paid by the volume, this contract takes the form of PPC.) In the latter setting, the vendor is responsible for the base and the user handles any fluctuation above the base. The vendor therefore will set a stable staffing level to handle the base and be compensated by PPT.

On the surface, the user should prefer to keep the base because of the lower in-house staffing cost it entails. However, the vendor's optimal charge will be different in the two settings. Overall, Akşin *et al.* [11] are able to characterize the optimal capacity levels and partially characterize the optimal prices under both contract types. Because the vendor optimally charges different prices under different contracts, neither contract type dominates for the user. The contract choice depends on the call fluctuation patterns in addition to the revenue and cost parameters.

Two more papers, Gans and Zhou [10] and Koçağa and Ward [20], deal with the cosourcing arrangement, but they eschew the contracting issues and focus on real-time call-routing decisions instead.

Gans and Zhou [10] consider the operational-level coordination issues in a call center cosourcing arrangement where calls of high value are often kept "in-house" by the user and the low-value calls are outsourced. Since the customer service representatives often receive nested skill training, those handling high-value calls are also capable of handling low-value calls. This presents the user a chance to use idle in-house, high-value representatives to handle low-value calls opportunistically. This helps the user firm to lower the volume of outsourced calls and hence cost. The authors present four coordination schemes differing in their routing sophistication and coordination/communication requirements. Assuming identical service time distribution of low- and high-value calls, they show that a relatively simple routing scheme can perform nearly optimally in large systems. This scheme also has the added benefit of not requiring extensive coordination between user and vendor firms on a real-time basis. In small systems, however, it is necessary to

use a threshold-type routing policy—using in-house high-value representatives to take low-value calls only when idle capacity exceeds a certain threshold.

In Koçağa and Ward [20], the user operates its own call center but, when the call center is congested, customers abandon due to long waits. In such cases, it is worthwhile to route calls to an outside vendor. A natural trade-off occurs between call abandonment and outsourcing cost. The paper formulates a Markov decision process model and shows that a queue-length-based threshold policy is optimal. Because this policy is complicated, the authors also propose a diffusion approximation and calculates the optimal threshold. Numerical analysis reveals that the diffusion approximation is very accurate and optimally utilizing the outsourcing option can yield significant cost savings.

While all the papers discussed so far deal with one-level service processes, Lee *et al.* [14] study contract issues in the outsourcing of a two-level service process. When all the customers first arrive, they are handled by first-level agents (who are called *gatekeepers*). Because customers are heterogeneous, the agents first diagnose the difficulty of each request. When the difficulty level is below a threshold k , the gatekeepers further proceed to treat these customer requests. This is successful only with probability $F(k)$. Customer requests with difficulty level above k are passed on (along with those unsuccessfully treated by the gatekeepers, called *mistreatment*) to a second-level group of agents (called *experts*). The experts then successfully handle all the requests. Staffing, as well as customer waiting, on both levels incur costs. The mistreatment by the gatekeepers also incurs a further cost. Using a Halfin–Whitt approximation [21] to staff both levels, the authors find the optimal first-best staffing levels and the optimal threshold k^* . They then show that (i) when only the expert level process is outsourced, a PPC+WP contract (where WP indicates additional linear penalty for waiting time) coordinates the systems, (ii) when only the gatekeeper level process is outsourced, a PPC+PPT+WP contract coordinates the system, and (iii) when both

levels are outsourced, a PPC+PPT+WP+ST contract (where ST indicates additional linear payment for successful treatments) coordinates the system *only if* system parameters satisfy certain condition.

Service Outsourcing under Information Asymmetry

Both parties involved in an outsourcing arrangement do not share all of their private information with each other. How does this information asymmetry affect the contract design, and what is the impact of such information asymmetry? Four papers address this question.

As we previously reviewed, in Hasija *et al.* [13], the vendor's effort level is private information not shared with the user. The authors show that, to achieve coordination, the contract in this setting is more complicated than that with full information. Additional SLAs or penalties must be added and only particular combinations of PPC and PPT contracts can coordinate.

Akan *et al.* [22] examine call center outsourcing contract design under a different kind of information asymmetry: the user firm has private information about customer call demand (high or low), while the vendor only has a prior belief of its distribution. When it is the vendor who offers the contract, the authors show that the optimal screening contract offers too low an abandonment probability to the high-type user (as compared with the first best) but offers no such distortion in the abandonment probability to the low-type user. This result contradicts the "efficiency at top" results in the literature on monopolist screening, and is largely due to the use of the queueing systems. When it is the user who offers the contract, the low type will offer the same abandonment probability as in the first best case, but the high type offers an abandonment probability below the first best, and has to pay more for each call. These distortions, however, can be mitigated and even removed by a two-part tariff contract—in addition to the per call charge, the user also makes a fixed payment to the vendor.

Ren and Zhang [23] examine call center outsourcing contract design when the user

does not know the vendor's capacity cost or quality cost. In many situations, when a vendor's service quality is high—be it service speed, or problem resolution rate—it tends to be more expensive as well. In the call center industry, for instance, labor costs are typically lower for call centers located overseas, compared to those located in the United States. However, the service quality for domestic call centers is often much higher. In some situations, the two costs may be positively correlated. For example, successful quality improvement initiatives may simultaneously bring down the capacity and quality costs of a firm. Ren and Zhang [23] study service outsourcing where the vendor exhibits such correlations. In their model, a user plans to outsource with an outside vendor, but it has no information about the vendor's capacity cost or quality cost (i.e., how hard it is for the vendor to reach a certain service level). However, the user does know the correlation between the two costs. Such information is usually obtainable through market research. The authors show that the form of the optimal outsourcing contract depends critically on the relationship between capacity and quality costs. They also study the effectiveness of simple contracts, contracts that are not contingent upon information unknown to the user. They find that those noncontingent contracts perform well when the correlation between capacity cost and quality cost is positive, but they can perform much worse when the correlation is negative.

Finally, Plambeck and Zenios [24] study a dynamic principal-agent model where the principal outsources work to the agent but the agent's effort is unobservable. Their model can be motivated by the so-called *maintenance problem*. A machine generates profits, but its owner delegates its maintenance to a manager, who has to repair the machine when it breaks down. The owner cannot observe the manager's effort in repairing the machine, and can only observe if the machine is functioning or not. Therefore the owner's problem is to design a performance-based contract that will motivate the manager to take actions that maximize the owner's profits over a finite horizon. Their model is applicable to similar situations in

service outsourcing, where the actions of the outsourced services are not observable, and only their outcomes are observed.

Quality Concerns in Service Outsourcing

One of the key considerations in service outsourcing is assuring high service quality by the vendor. As we stated in the introduction, firms are increasingly focusing on the quality benefits of outsourcing rather than the cost benefits. Three papers deal with the quality aspect of outsourcing.

As we reviewed before, Ren and Zhou [12] examine how contracts can be used to effectively ensure optimal service quality level by the vendor. They show that simple contracts, even the pay-per-call-resolved contract that takes quality into consideration, do not coordinate the vendor's quality training effort level. Instead, more complex contracts are needed. In this article, quality is measured by the call resolution rate achieved by the vendor.

Benjaafar *et al.* [25] study how firms should go about outsourcing to their vendors with service quality considerations. Below, we describe their model in a service outsourcing setting, even though it applies to both service and inventory systems. They compare two scenarios. The first is to allocate demand among a set of vendors, but the proportion of demand that each vendor gets is an increasing function of the service level of each vendor. Here service level is measured by customer waiting time, abandonment probability, and so on. They call this approach supplier-allocation (SA) approach. The other, called the *supplier-selection* (SS) approach, uses a probability to select one vendor. The probability of being selected is an increasing function of service level committed by the vendor. They find that often using just one vendor (i.e., the SS approach) is advantageous, but such an advantage depends upon the form of cost functions of serving customers incurred by the vendors. Moreover, by selecting an appropriate allocation mechanism, the user can indeed induce a higher service level provided by its vendors.

Cachon and Zhang [26] study a queueing model with one user and two vendors. The

user could orchestrate a service competition by allocating demand based on the vendors' performance, which is measured by speed of service. But how? The authors consider a variety of allocation policies, starting from state-independent policies that minimize the user's lead time given fixed capacities and nonstrategic vendors. They then move on to more complex state-dependent policies, and find that the variation in performance across different allocation policies is wide. Moreover, they show that there exists an allocation policy that can induce maximum vendor speed and at the same time minimize the user's lead time. Therefore, there need not be a trade-off in incentive and efficiency in outsourcing.

In the ride share model studied by Li *et al.* [27], the service level is measured by the percentage of demand that is denied. BostonCoach Inc. is a premier provider of executive sedan, limousine, and event transportation services. Not unlike the call centers, it has a relatively inflexible fixed capacity, yet its demand is random and fluctuating: in the Boston area, on a typical day, only about 85% of its demand is booked one day in advance. If BostonCoach projects to have insufficient capacity, it can subcontract with its collaborating companies (thus this is a cosourcing arrangement), but this has to be done one day ahead. That is, the overall capacity (own capacity plus contracted capacity) must be decided one day ahead. In real time, if, on the day of service, demand is much higher than expected, then requests will be denied, resulting in significant ill-will cost. The authors develop a model that incorporates various demand features (origin-destination, time window, specific vehicle and/or driver request, and trip type) and capacity constraints (driver availability, shift constraints, driver/vehicle pairing, etc.). A math program is devised to minimize the total cost and a column-generation method is adopted to solved it. Results show that using the proposed method can significantly improve the productivity of existing fleet. This results in reduced percentage of denied services, reduced use of drivers, and reduced use of contract services.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this article, we aim to give an overview of the current status of OM research in service outsourcing. Currently, the number of papers in this area is not large but we expect the area to grow over time.

Several areas are promising for future research:

- Outsourcing vendor selection is a difficult but important topic. Owing to the uniqueness of services being outsourced and the difficulty of measuring and monitoring various aspects of service quality, switching vendors is a process that involves significant cost and time. The current service outsourcing contracting papers have focused on one-shot games. Clearly, it would be interesting and helpful to study the use of repeated-game models and other more complex contract structures (for an example of the use of repeated game in a supply chain context [28]).
- Most services are complex, multiple-stages processes. Outsourcing decisions do not have to be all-or-nothing. Two directions are worth pursuing: (i) The user can choose to outsource some stages of the service process while keeping the rest in-house, but the current papers, with the exception of Lee *et al.* [14], are all concerned with the outsourcing of a single-stage process. The choice of what to outsource, how to outsource, and how to coordinate the stages that are outsourced with those that are not are all interesting areas for future research. (ii) Even after the user has decided what to outsource, there is still the question of how much to outsource. Cooutsourcing, as we review above, is a useful arrangement. More research is needed to shed light on its effective use.
- Other types of information asymmetry, such as private demand forecasts, are worth studying. In service outsourcing, the vendor often has to make

capacity decision with only a rough estimate of future demand, while the user may have better information. Hence, depending on specific situations, the user may have an incentive to overforecast demand to the vendor, causing costly capacity overinvestment. How to induce truthful behavior from the user and achieve the maximum system efficiency is an interesting topic. Similar problems have been studied in the context of supply chain management. Interested readers are referred to Cachon and Lariviere [29].

- On a more technical level, the difference in the results obtained by Ren and Zhou [12] and Hasija *et al.* [13] is due to the different queueing models they choose (the fluid model in Whitt [16] and the diffusion model in Garnett *et al.* [15]). Therefore, much work still needs to be done to understand the impact of using different queueing models and approximation methods. Interested readers are referred to Ward Whitt's and Avi Mandelbaum's research web pages¹.

Since there already exists a large literature on information technology outsourcing (ITO), it is also beneficial for OM researchers to learn about the other aspects of outsourcing that we have not covered in this article. Dibbern *et al.* [30] give an excellent and comprehensive review of the IS outsourcing literature. We have touched upon several relevant topics in this article, but interested readers are highly encouraged to read [30]. Although they only review the ITO literature, we believe the methodologies and frameworks apply to other outsourcing settings, including service outsourcing.

Finally, we fully expect research in the areas of personal, legal, medical, and other professional service outsourcing to

¹<http://www.columbia.edu/~ww2040/allpapers.html> and <http://iew3.technion.ac.il/serveng/References/references.html> respectively.

grow rapidly, along with the increasing use of outsourcing in these areas in practice.

REFERENCES

- Dickens Johnson MM. Current trends of outsourcing practice in government and business: Causes, case studies and logic. *JoPP* 2008;2:248–268.
- Leavitt N. The changing world of outsourcing. *Comput IEEE Comput Soc* 2007;40(12): 13–16.
- Kakabadse N, Kakabadse A. Critical review – outsourcing: a paradigm shift. *J Manage Dev* 2000;19(8):670–728.
- Douglas Brown and Wilson Scott. *The Black Book of Outsourcing*. New Jersey: John Wiley & Sons, Inc.; 2005.
- Karmarkar US, Apte UM. Operations management in the information economy: information products, processes, and chains. *J Oper Manage* 2007;25(2):438–453.
- Douglas Brown and Wilson Scott. *The Black Book of Outsourcing: State of the Industry Report*. Florida: Brown-Wilson Group, Inc.; 2008.
- Gartner Group. Gartner says worldwide outsourcing market to grow 8.1 percent in 2008. Available at <http://www.gartner.com/it/page.jsp?id=578307>. Accessed 2010.
- Anand K, Aron R. Quality, incentives and inspection regimes in offshore service production: Theory and evidence. Working paper. Available at http://digital.mit.edu/wise2006/papers/7B-2_Anand-Aron.pdf. Accessed 2008.
- Engelke WD. Writing an outsourcing contract. Available at <http://www.hsv.com/writers/engel/outsou3.html>. Accessed 2009.
- Gans N, Zhou Y-P. Call-routing schemes for call-center outsourcing. *Manuf Serv Oper Manage* 2007;9(1):33–50.
- Zeynep Akşin O, de Véricourt F, Karaesmen F. Call center outsourcing contract analysis and choice. *Manage Sci* 2008;54(2):354–368.
- Justin Ren Z, Zhou Yong-Pin. Call center outsourcing: coordinating staffing level and service quality. *Manage Sci* 2008;54(2): 369–383.
- Hasija S, Pinker EJ, Shumsky RA. Call center outsourcing contracts under information asymmetry. *Manage Sci* 2008;54(4):793–807.
- Lee H-H, Pinker EJ, Shumsky RA. Outsourcing a two-level service process. University of Rochestre; 2009. Available at http://www.hsiao-huilee.com/My_Research.html. Accessed 2009.
- Garnett O, Mandelbaum A, Reiman M. Designing a call center with impatient customers. *Manuf Serv Oper Manage* 2002;4(3): 208–227.
- Whitt W. Fluid models for many-server queues with abandonments. *Oper Res* 2006;54(1): 37–54.
- Milner JM, Olsen TL. Service-level agreements in call centers: perils and prescriptions. *Manage Sci* 2008;54(2):238–252.
- Willcocks LP, Lacity MC. *Strategic sourcing of information systems: perspectives and practices*. New York: John Wiley & Sons, Inc.; 1998.
- Fuhrman F. Co-sourcing: a winning alternative to outsourcing call center operations. *Call Cent Solut* 1999;17(12):100–103.
- Koçağa YL, Ward A. Dynamic outsourcing for large-scale service systems. Working paper. 2009.
- Halfin S, Whitt W. Heavy-traffic limits for queues with many exponential servers. *Oper Res* 1981;29(3):567–588.
- Akan M, Ata B, Lariviere MA. Asymmetric information and economies-of-scale in service contracting. Working paper, COSM-07-001. 2007. Available at <http://www.kellogg.northwestern.edu/faculty/lariviere/research/CallCenterContract.htm>. Accessed 2009.
- Justin Ren Z, Zhang F. Service outsourcing: Capacity, quality and correlated costs. Working paper; 2009. Available at <http://apps.olin.wustl.edu/faculty/zhang/>. Accessed 2009.
- Plambeck EL, Zenios SA. Performance-based incentives in a dynamic principal-agent model. *Manuf Serv Oper Manage* 2000;2: 240–263.
- Benjaafar S, Elahi E, Donohue K. Outsourcing via service quality competition. *Manage Sci* 2007;53(2):241–259.
- Cachon GP, Zhang F. Obtaining fast service in a queueing system via performance-based allocation of demand. *Manage Sci* 2007;53(3): 408–420.
- Li Y, Wang X, Adams TM. Ride service outsourcing for profit maximization. *Transp Res Part E* 2009;45:138–148.
- Taylor TA, Plambeck EL. Supply chain relationships and contracts: the impact of repeated interaction on capacity investment

- and procurement. *Manage Sci* 2007;53(10):1577–1593.
29. Cachon G, Lariviere M. Contracting to assure supply: how to share demand forecasts in a supply chain. *Manage Sci* 2001;47(5):629–46.
30. Dibbern J, Goles T, Hirschheim R, *et al.* Information systems outsourcing: a survey and analysis of the literature. *Database Adv Inf Syst* 2004;35(4):6–102.

SHIPPER AND CARRIER COLLABORATION

ÖZLEM ERGUN

LUYI GUI

H. Milton Stewart School of
Industrial and Systems
Engineering, Georgia
Institute of Technology,
Atlanta, Georgia

LORI HOUGHTALEN

Babson College, Babson Park,
Massachusetts

OKAN ÖRSAN ÖZENER

Industrial Engineering
Department, Özyegin
University, Istanbul,
Turkey

Cooperation among companies is widely practiced today in different business sectors. Facing increasing pressures to improve profitability, companies are responding by seeking and implementing solutions that require strategic collaboration with external partners because these solutions afford benefits that cannot be achieved alone. In the transportation industry, cooperation occurs among both the buyers of transportation services, that is shippers, and the suppliers of transportation services, that is carriers. For example, Elemica [1] coordinates and jointly procures truckload transportation services for more than 2000 companies in the chemicals industry creating a balanced transportation network due to *economies of scope*. Less-than-truckload (LTL) operations, where the carriers (truck companies) collect and bundle freights from various shippers to transport from one breakbulk terminal to another, can also be regarded as an implicit collaboration among shippers creating savings due to *economies of scale*. Such operations enable shippers to enhance their overall shipping efficiency by sharing truck capacity, resulting in increased capacity utilization and decreased deadheading. This,

in return, reduces the shippers' individual transportation costs as well.

On the carriers' side, alliances are formed in many modes of transportation, especially those with regularly scheduled service routes. For example, the first modern air cargo alliance, SkyTeam Cargo, was formed in 2000 by Aeroméxico Cargo, Air France Cargo, Delta Air Logistics, and Korean Air Cargo. According to the published figures, the two main air cargo alliances, WOW Cargo and SkyTeam Cargo, accounted for about 20% [2] and 17.5% [3,4], respectively, of the total air freight traffic in freight ton kilometers in 2003. Members of an alliance usually collaborate by integrating their operating networks into an extended and better load-balanced one, which is more profitable and has a reduced cost of fleet idling. The infrastructure setup costs, such as equipment and administrative costs, are often shared as well, providing more incentives for individual carriers to join an alliance.

Collaboration among shippers and among carriers plays a key role in the current cargo transportation industry and will serve as a future trend as it facilitates more efficient and cost-effective operations. However, many challenges are posed in the management of carrier alliances and collaborative procurement networks as they are typically composed of entities driven by selfish interests. In this article, we provide an overview of how to design and manage collaborative initiatives so that both the overall efficiency and the stability of the collaboration is guaranteed. To this end, we discuss the following three management mechanisms, differing from each other in the levels of coordination and information sharing required among the collaborating entities as follows:

1. We first consider a *centralized* model where a central authority is assumed to run the overall business operations and then distribute the revenues or costs among the members. Individual members have the option of breaking

away from the collaboration if their net benefits are undermined. Our goal is to evaluate the maximum potential benefit from collaboration by solving for the centralized system optimal solution, and to allocate the benefit in a fair way. Such a setting is relevant, for example, in collaborative transportation procurement networks where there is a centralized entity such as Elemica that serves as the centralized overseer.

2. Most carrier alliances function as *decentralized* systems where each carrier operates its own assets and makes its own routing decisions. Therefore, we next consider mechanisms where a centralized authority only designs the rules of interaction for the individual entities in order to elicit a solution out of individual decisions that is both socially efficient and beneficial for individuals. Specifically, as most carrier alliances are formed primarily for pooling fleet capacity, one natural approach is to establish capacity-sharing mechanisms, for example, unit prices for capacity exchanges.
3. Finally, we consider *auction* mechanisms that are widely adopted for the management of procurement of transportation services, especially when there is a large number of carriers and the market is fragmented (such as in the US trucking industry [5]). In such settings, procurement auctions can be used to implicitly coordinate the actions of the individual buyers and sellers. We discuss the design of auctions in order to reveal true information from the bidders and thus, to obtain a socially efficient outcome.

Note that the level of coordination and information sharing decreases as we go from centralized operations and cost or benefit allocations, to exchange mechanisms where there is a limited centralized control amounting to establishing rules and prices for the exchanges, and finally to auction mechanisms where synergies get exploited by the evolution of user bids. However, as the coordination and information sharing

requirements decrease, the computational burden on individuals for making optimal decisions increases. Moreover, the guaranteed system efficiency may also decrease as the degree of centralized control decreases.

Since the analysis of the approaches that we will address in the rest of this article largely depends on basic game theoretic concepts, we introduce these concepts in the following section.

GAME THEORY BASICS

Cooperative Game Theory

The field of *cooperative game theory* studies the interactions of selfish participants in a collaboration and provides a set of mathematical tools and concepts to study the “fair” distribution of costs and benefits. Specifically, a *cooperative game* is defined by a set N of players, referred to as the *grand coalition*, and a *characteristic function* $\mathcal{V}: 2^N \rightarrow \mathbb{R}$. The function $\mathcal{V}$ maps a value $\mathcal{V}(S)$ to every subset/subcoalition $S \subset N$, interpreted as the total benefit that the members of S can achieve by cooperating among themselves. We assume that $\mathcal{V}(\emptyset) = 0$. The central issue in cooperative game theory is how to allocate the total gain $\mathcal{V}(N)$ among the individual players $i \in N$ in a “fair” way to sustain the long-term stability of the grand coalition. One of the most prominent and widely accepted notions of fairness is the “core” of the game. A payoff allocation in the core represents a very strong type of stability that is not threatened by the defection of subcoalitions. However, frequently the core of a general cooperative game is empty. Mathematically, a payoff vector $x \in \mathbb{R}^{|N|}$, where x^i is the payoff to player i , is said to be in the core if

$$\sum_{i \in N} x^i = \mathcal{V}(N) \quad (\text{budget balance property}), \quad (1)$$

$$\sum_{i \in S} x^i \geq \mathcal{V}(S) \quad \forall S \subset N. \quad (\text{stability property}). \quad (2)$$

Some other important concepts regarding the *fairness* of benefit allocations from cooperative game theory include the stable

set, the nucleolus, and the Shapley value. We refer to Young [6] for a thorough review of basic cost allocation methods and to Borm *et al.* [7] for a survey of cooperative games associated with operations research problems.

Mechanism Design

Given a set of players N and a set $\mathcal{O}$ of possible outcomes for these players, let $v^i(o)$ be the valuation of an outcome $o \in \mathcal{O}$ for player i . Generally, we assume that each individual player aims to optimize his/her total payment. The goal of *mechanism design* is to design an algorithm that chooses an outcome $\tilde{o} \in \mathcal{O}$, and an n -tuple of side payments $\{s^1, \dots, s^n\}$ such that the total payoff to player i , x^i , calculated as $x^i = v^i(\tilde{o}) + s^i$, makes the outcome $\tilde{o}$ attractive to the individual players. Intuitively, mechanism design is concerned with determining both an algorithmic ingredient (to obtain the desirable system outcome) and a payment ingredient that motivates the players. This field has received widespread attention recently and has been used successfully to develop algorithms and protocols for interconnected collections of computers such as on the Internet. For a detailed discussion of mechanism design we refer the reader to Nisan and Ronen [8].

CENTRALIZED COLLABORATION MECHANISMS

In the centralized collaborative settings we assume that all information, assets, and operational decisions are handed over to a centralized authority and in return, the system is able to exploit synergies fully and promises a fair allocation of benefits and costs to individuals.

Such centralized carrier and shipper collaborations can generally be represented as cooperative games called (*multicommodity network flow games*) where the individuals own the commodities to be shipped between two given nodes for a revenue, as well as the capacities on the edges which will be used for the shipments. The concept of *flow games* is introduced by Kalai and Zemel [9,10], who

study centralized cooperation on networks with a single commodity where each arc is owned by a unique player. Derks and Tijs [11] extend the earlier results by studying multiple commodities with a common source–sink pair. More recently, Markakis and Saberi [12] study the coalitional multicommodity game in which players are the nodes of the given network who own capacity in terms of the maximum flow allowed through that node. Based on flow game models, the centralized authority generally can obtain the best system solution by solving a (multicommodity) flow problem on the given integrated network. The network flow models also facilitate the allocation of the revenue generated by the system completely and fairly so that no individual or groups of individuals defect from the collaboration.

Problem Definition

We define the maximum multicommodity flow (MCF) game as follows: let $G = (V, E)$ be a directed graph on a set of nodes V and a set of edges E which represents the entire transportation network in consideration. Let N be the set of shippers or carriers serving as players of the game. Each player $i \in N$ has a demand set D^i where $d_{(o,d,i)}$ denotes the maximum amount of existing demand from node o to node d for player i and $r_{(o,d,i)}$ denotes the revenue generated by delivering a unit of this demand. In the multicommodity setting, each demand triplet (o, d, i) represents a different commodity. Each edge $e \in E$ has a capacity c_e , which denotes the maximum amount of flow, summed over all commodities, allowed on that edge. γ_e^i denotes the fraction of capacity that player i owns on edge e , and $\sum_{i \in N} \gamma_e^i = 1$. For any set $S \subset N$, including N itself, we associate with it the value $\mathcal{V}(S)$ which equals the maximum revenue generated through the network owned by players in the set S . Such a value can be computed by solving a maximum MCF problem defined by the following linear program (P^S):

$$(P^S): \quad \mathcal{V}(S) = \max \sum_{(o,d,i) \in \bigcup_{i \in S} D^i} f_{(d,o,i)}^{(o,d,i)} r_{(o,d,i)} \quad (3)$$

$$\begin{aligned}
& \text{such that } \sum_{\{e:e \in IEdges(v)\}} f_e^{(o,d,i)} \\
& - \sum_{\{e:e \in OEdges(v)\}} f_e^{(o,d,i)} \leq 0 \\
& \forall v \in V \quad \forall (o,d,i) \in \bigcup_{i \in S} D^i \quad (4) \\
& \sum_{(o,d,i) \in \bigcup_{i \in S} D^i} f_e^{(o,d,i)} \leq c_e \sum_{i \in S} \gamma_e^i \quad \forall e \in E \quad (5) \\
& f_{(d,o,i)}^{(o,d,i)} \leq d_{(o,d,i)} \quad \forall (o,d,i) \in \bigcup_{i \in S} D^i \quad (6) \\
& f \geq 0. \quad (7)
\end{aligned}$$

In (P^S) , the variables $f_e^{(o,d,i)}$ denote the amount of flow of commodity (o,d,i) on edge e for each commodity (o,d,i) owned by members within S . $IEdges(v)$ denotes the set of incoming edges into node v and $OEdges(v)$ denotes the set of outgoing edges from v . The objective function (3) maximizes the revenue obtained by satisfying commodity demand. Constraint (4) is a flow balance constraint at every node of the network with respect to each of the commodities. Constraint (5) is a capacity constraint on each edge and constraint (6) is a demand constraint for every commodity. The optimal solution to (P^S) represents the flow that maximizes the revenue generated within the subcoalition S .

An outcome of the MCF game defined above consists of a feasible multicommodity flow, subject to demand and capacity constraints that maximizes the overall revenue generated. We next discuss how solving P^S also provides an allocation, $x \in \mathbb{R}^n$, of the resulting total revenue, that is $\mathcal{V}(N)$, among the players.

An Allocation in the Core

For cooperative games where the values of subcoalitions are computed via linear programming (LP) techniques, utilizing the dual optimal solution to the grand coalition's optimization problem is a common strategy to obtain an allocation for the players. Owen [13] shows that for linear production games where the LPs that are used to solve for

values of subcoalitions follow a regular pattern, such a dual scheme provides an allocation in the core. Kalai and Zemel [10] show that for simple networks, that is single commodity networks where each edge has a single unit of capacity and is uniquely owned, every core allocation corresponds to an optimal dual solution for the corresponding optimization problem. Deng *et al.* [14] provide an extension of Kalai and Zemel [10]. In the MCF game scenario, the dual payoff vector is defined as $x^i = \sum_{e \in E} \alpha_e^* c_e \gamma_e^i + \sum_{(o,d,i) \in D^i} \beta_{(o,d,i)}^* d_{(o,d,i)}$ where $\alpha^* = \{\alpha_e^* : \forall e \in E\}$ and $\beta^* = \{\beta_{(o,d,i)}^* : \forall (o,d,i) \in \bigcup_{i \in N} D^i\}$ are the optimal dual solutions associated with constraints (5) and (6) of the program (P^N) respectively.

Theorem 1. [15]. *The core of the MCF game defined in the section titled “Problem Definition” is nonempty and a payoff allocation in the core can be constructed in polynomial time by solving the dual program to the LP (P^N) . However, the core of this game is not generally fully characterized by these dual solutions.*

The fact that the core of the MCF game is not fully characterized by the dual payoffs is largely due to the generality of the game where no specific conditions are required in terms of either the network structure or individual capacity ownership levels. This fact not only motivates further research for understanding the elements in the core that are not driven from the dual payoffs, but also poses additional challenges in understanding the mechanisms that induce fair allocations (refer to Gui and Ergun [16] for details).

Applications

An Example. Figure 1 illustrates a time-expanded air-cargo alliance network with two capacitated edges (edges (1,3) and (2,4)) that represent flights operated by carrier A. The capacity on each of these edges is one. Edge (1,2) and (3,4) are ground edges, representing the ability of a load to wait in a location over time and therefore, have unlimited capacity. The per-unit revenues and the sizes of the loads are described in Table 1. The unique

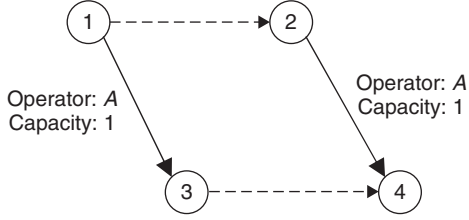


Figure 1. Alliance network for allocation example.

Table 1. Loads for Allocation Example

Demand	Per-Unit Revenue ($r^{(o,d,k)}$)	Size ($d^{(o,d,k)}$)
(1, 3, A)	2	1
(2, 4, A)	2	1
(1, 4, B)	6	1
(2, 4, C)	3	1

centralized optimal solution is to deliver load (1, 4, B) using leg (1, 3) and ground edge (3, 4), and to deliver load (2, 4, C) using leg (2, 4), which generates a total revenue of 9. The optimality of this solution can be confirmed by solving the centralized LP (P^N) for this example. A subset of allocations in the core to carriers A, B, and C can then be characterized by $[\alpha_{(1,3)}^* + \alpha_{(2,4)}^*, 6 - \alpha_{(1,3)}^*, 3 - \alpha_{(2,4)}^*]$ where $\alpha_{(1,3)}^*$ and $\alpha_{(2,4)}^*$ are the optimal dual solutions with respect to edges (1, 3) and (2, 4) respectively, and thus must satisfy $\alpha_{(1,3)}^* \in [2, 6]$ and $\alpha_{(2,4)}^* \in [2, 3]$. For example, the dual solution in which $\alpha_{(1,3)}^* = 2$ and $\alpha_{(2,4)}^* = 2$ leads to the core allocation of $x^A = 4, x^B = 4$, and $x^C = 1$.

Other Applications of Cooperative Game Theory to Network Flow Problems. Flow game models have been widely adopted to analyze interactions of selfish entities in network collaborations under various practical contexts. There exists a set of papers that join cooperative game theory and classic network flow problems, such as Ergun *et al.* [17] and Özener and Ergun [18], which consider the centralized operations and cost allocations of the shipper collaboration based on the lane covering problem (LCP); Sanchez-Soriano *et al.* [19], which studies transportation games where shippers and

carriers are disjoint sets; Granot *et al.* [20], which analyzes delivery games that are associated with the Chinese postman problem; Derks and Kuipers [21], which studies routing games; and Tamir [22], which considers continuous and discrete network synthesis games. These papers in general study the existence of cost allocations with well-studied properties from cooperative game theory and propose computational procedures for finding such allocations. In addition, Song and Panayides [23] make use of cooperative game theory and provide a quantitative study to analyze liner shipping alliances by considering two small examples (involving 3 ports and 2–3 carriers). The costs and benefits are allocated among the alliance members in proportion to the shipping capacity each provides. However, in many situations a proportional allocation of benefits is not guaranteed to provide a payoff to the alliance members that is sufficient to sustain the stability of the collaboration. Moreover, network-related games such as the network design game of Kubo and Kasugai [24], the assignment game of Shapley and Shubik [25], and facility location game of Goemans and Skutella [26] have also been studied in the literature. Here, we provide detailed descriptions of two fundamental applications to classic network problems, the traveling salesman games (TSG) and vehicle routing games (VRG), which have been extensively discussed in the literature.

The TSG considers the problem of allocating the cost of a round trip among the cities visited. Depending on the game, there may exist a home city which is not included in the set of players. Tamir [27] considers a TSG with a home city and proves that under a symmetric cost structure, the TSG always has a nonempty core when the number of players is less than or equal to four. Kuipers [28] extends this result to a TSG with five players. Note that the core may be empty when the number of players is greater than or equal to six. Potters *et al.* [29] discuss the TSG and its variant with fixed routes. For the fixed route TSG where for all coalitions the cities are visited in the same order, they prove that the game has a nonempty core if

the fixed route solution is the optimal solution to the original TSP and the cost matrix satisfies the triangle inequality. They also introduce a class of matrices that generates TSGs with nonempty cores. Engevall *et al.* [30] consider the cost allocation problem of a distribution system of a gas and oil company where the total cost of a tour is to be allocated among the customers that are visited. Besides the well-known concepts regarding fair allocations from cooperative game theory, such as the nucleolus and the Shapley value, they introduce the “demand nucleolus,” which is similar to the nucleolus except that the excess of a coalition is multiplied by the demand of that coalition. By doing this, they aim to reduce the importance of coalitions that have relatively larger demands in computing the nucleolus.

The VRG considers the problem of allocating the cost of optimal vehicle routes among the customers served. Gothe-Lundgren *et al.* [31] consider the VRG with a homogeneous fleet and present conditions when the core of a VRG is nonempty. Then, they use a constraint generation approach to compute the nucleolus of a VRG with a nonempty core. Engevall *et al.* [32] extend this result to a VRG with heterogeneous vehicles with some simplifications, for which they study the existence of allocations in the core.

Although centralized systems are very efficient as one decision maker can choose the optimal solution for the collaboration and allocate the overall benefits in a fair manner so that the collaboration is stable, in many settings designing fully centralized systems is not realistic. In fact, real-life applications of network collaborations tend to be decentralized to a certain extent, as network users typically make decisions regarding their own commodity routing and capacity management. Hence, the incentives within the system must be designed in such a way that the individual players are motivated to participate in the grand coalition and choose solutions that are collectively optimal for the collaboration; that is, the revenue obtained is close to the maximum level that can be achieved within a fully centralized system. We present a detailed discussion of the management of shipper and carrier

collaborations under partially decentralized settings in the next section.

EXCHANGE MECHANISMS

In this section we focus on mechanisms that facilitate cooperation among participants in a decentralized setting through the exchange of network resources. We will primarily focus on exchanges among carriers. Given the decentralized setting, an important consideration of work in this area is that the entities that are collaborating (i.e., carriers and/or shippers) are autonomous and self-optimizing. Mechanisms which seek to influence behavior therefore must ensure that each entity stands to benefit by making the desired decisions.

The majority of work related to exchange mechanisms in collaborative transportation has been focused on settings in which carriers collaborate by sharing capacity. In general, the goal of these mechanisms is to encourage carriers to make decisions such that the welfare of the collaboration as a whole (i.e., the alliance or grand coalition) is optimized. Agarwal and Ergun [15] explore a general multicommodity flow cooperative game in which multiple players may own the capacity on each edge in the network. The proposed mechanism solves an inverse optimization problem (i.e., a modified dual problem in which every feasible solution satisfies complementary slackness conditions with a desired primal solution determined *a priori* [33]) to establish prices that players pay for the use of capacity which they do not own. The authors show that the prices always encourage players participating in the cooperative game to make capacity and routing decisions that are optimal for the grand coalition. However, it is only for the special case in which every edge is uniquely owned that the resulting allocation is guaranteed to lie in the core of the game. Gui and Ergun [16] expand on this work by examining conditions in which a core allocation can be guaranteed even when every edge in the network is not uniquely owned. A key finding is that the relationship between allocations obtained using the mechanism developed in Agarwal and Ergun [15] and the core is influenced by the presence of excess network capacity. Specifically,

in a congested network the mechanism will always yield prices that lead to an allocation in the core, even when edges are not uniquely owned.

A number of mechanisms have been proposed to manage collaborations in a specific industry. In the passenger airline industry, a problem receiving increased scholarly attention is that of how to share revenues among code-share partners. In addition to allocation of revenue, mechanisms must address the allocation of capacity among partners. Approaches to what is commonly called *alliance revenue management* can be classified broadly as static (those approaches in which capacity and pricing decisions do not change as capacity availability changes) or dynamic (approaches in which the mechanism adjusts recommendations according to the availability of seats in the flight network). An example of a static scheme is one in which revenue is shared between code-share partners according to a negotiated proportion (e.g., the proportion of miles that a customer flies on each airline). An example of a dynamic scheme is one which allocates revenue according to the bid prices of the partners, which change over time as inventory in the network diminishes.

Vinod [34] and Wright *et al.* [35] both address the problem of revenue management in passenger alliances. In general, Vinod asserts that the more the alliance can function as a single entity, the more benefit will be gained by collaborating. He acknowledges, however, that significant legal and technical challenges generally prevent full integration. Thus, analyzing an alliance among passenger airlines from a more decentralized perspective is necessary and he briefly explores the idea of how code-share partners can share revenue using a dynamic fare proration scheme based on bid prices. A key argument is that a revenue-sharing agreement should allocate to a carrier an amount of revenue for a flight leg that is at least as great as his bid price for that leg.

Wright *et al.* [35] conduct a more formal analysis of alliance revenue management mechanisms, analyzing each for its ability to induce alliance-optimal behavior. It is shown that static mechanisms can be appropriate in

limited situations, but in general a dynamic policy will lead to decisions that are more profitable for the alliance and its members. Furthermore, mechanisms that manage capacity according to a free-sale policy are promising. (A free-sale mechanism is one in which an airline purchases seats from a partner on an as-needed basis, as opposed to purchasing a block of seats to sell over time). The authors examine three types of free-sale dynamic mechanisms, and find the alliance characteristics that will make each type of scheme particularly successful.

Alliances among airlines also exist in the cargo segment of the industry. Houghtalen *et al.* [36] propose a mechanism for managing interactions among combination carriers, which are those carriers that transport cargo in the belly of passenger aircraft. The mechanism operates under assumptions similar to those applied in Agarwal and Ergun [15] and Gui and Ergun [16]: namely, that some form of centralized coordination or control is relied upon for the computation of mechanism parameters (i.e., capacity allotments and prices), but since carriers make their own routing decisions (and do so in a self-optimizing manner), a decentralized perspective is necessary when analyzing the computation and realization of revenue allocations. The mechanism proposed by the authors allocates capacity among partner carriers in a manner that ensures an alliance-optimal routing of cargo is feasible, then establishes prices for each flight leg that encourage carriers to make alliance-optimal routing decisions. The methodology for establishing prices, similar to that proposed by Agarwal and Ergun [15], is based on solving an inverse optimization problem. Computational experiments suggest that the proposed mechanism, when applied in a realistic alliance setting in which capacity prices must be fixed for a period of time, is not only very successful at achieving near-optimal alliance revenue, but also is very robust with respect to variability in cargo demand characteristics.

Because airline flights, even in an alliance setting, are operated by a single carrier, mechanisms designed to manage interactions among airlines can assume that every edge

in the transportation network is owned by a single carrier. In the liner shipping industry, in contrast, this assumption cannot be made. In an alliance in this industry, shipping companies work together to provide a specified (i.e., weekly) frequency of service on a given shipping route. As a result, a single edge in the transportation network may be serviced by multiple shipping companies. Agarwal and Ergun [5] study alliances in the liner shipping industry, from the perspective of how to allocate the costs associated with operating the shipping network. Their proposed mechanism utilizes a similar methodology as that employed in Agarwal and Ergun [15]. However, acknowledging that a core allocation is not guaranteed when there are multiple owners on an edge, constraints are introduced to ensure, at a minimum, rationality for subgroups of a given size. Computational experiments show that the mechanism is effective in finding capacity prices that lead to cost allocations in the core provided that the core is nonempty. The assignment of ships to the service routes in the network (which is accomplished before pricing decisions are made) can influence the costs assignments made by the mechanism but does not in general, impact the ability of the mechanism to find core allocations.

Exchange mechanisms among carriers in the truckload transportation industry are the focus of Özener *et al.* [37]. In contrast to the exchange mechanisms discussed above (related to the airline and liner shipping industry), the goal of this work is to identify beneficial ways for trucking carriers to exchange shipping lanes. That is, the carrier serving a particular edge (lane) in the network changes rather than carriers sharing capacity in a given lane. Lane exchanges can reduce carrier costs primarily by reducing costs associated with repositioning. Özener *et al.* [37] examine a decentralized setting in which carriers make decisions in a myopic manner; the only centralized perspective necessary is for determining the rules under which the exchange mechanism will operate. The authors propose four different mechanisms, where each mechanism is defined primarily by whether side payments and

information sharing are allowed among carriers. They find that while all four mechanisms are able to achieve substantial cost savings for the carriers, allowing information sharing has a greater impact on the cost savings achieved by the mechanism than allowing side payments between carriers.

An interesting related application of mechanisms designed to facilitate the exchange of resources is in the field of air traffic flow management. Specifically, the resources of interest are the landing (or arrival) slots at an airport, which are typically awarded to individual airlines based in part on the number of slots the airline has historically been allotted. In the case of a necessary reduction in the arrival capacity at an airport (typically due to weather), reassigning landing slots can benefit airlines by reducing the overall costs associated with a weather delay. The goal of slot trading mechanisms is to identify attractive exchanges of landing slots between carriers, typically with the aid of a mediator to evaluate bids or offers submitted by airlines. See, for example, Vossen and Ball [38] and Madas and Zografos [39]. Slot trading mechanisms differ from those discussed in detail above in that the resources being exchanged are not carrier-owned.

We return to the example introduced in the previous section to demonstrate the ideas underlying the exchange mechanisms discussed in this section. A mechanism designed to achieve maximum benefit from collaboration should facilitate the collaborative optimal solution in which loads $(1, 4, B)$ and $(2, 4, C)$ are delivered. Clearly, carrier A will need incentive to cooperate with this solution; one possibility is for carriers B and C to pay carrier A for the use of capacity. Thus the mechanism might determine that appropriate payments for capacity on edges $(1, 3)$ and $(2, 4)$ are 2.5 and 2.5, respectively, yielding a solution in which the net revenue associated with carrier A is 5, carrier B is 3.5, and carrier C is 0.5. Note that according to this solution, all carriers receive strictly positive benefit.

It is important to note that the design of the mechanism can have a substantial

impact on the benefit experienced by the participating carriers. For example, the solution proposed above is acceptable from the standpoint that all carriers receive strictly positive benefit from collaborating. However, there exists a range of possible payments that carriers B and C might offer to carrier A that fit this criteria. It is interesting to consider the fairness of various revenue allocations that might be obtained by a given mechanism. For example, one could argue that carrier A should receive more benefit, since it is contributing all of the valuable capacity in the network. On the other hand, carriers B and C are contributing the valuable cargo. Designing a mechanism in consideration of these issues is the subject of Houghtalen *et al.* [40].

Finally, we consider the possibility that a given solution may yield unintended results. Again, referring to the example above, consider a solution in which the cost associated with capacity on edge $(1, 3)$ is 5. That is, carrier B will have to pay carrier A a price of 5 for the use of that capacity. The cost of capacity on edge $(2, 4)$ is 2. If we examine this solution in detail, we find that it gives rise to a secondary market for capacity on edge $(2, 4)$. A capacity exchange in the secondary market will occur if carrier B realizes that a unit of capacity on edge $(2, 4)$ can be bought from carrier C for the price of 4, instead of buying capacity on $(1, 3)$ directly from carrier A for the price of 5. In this case, the resulting allocations actually realized by carriers A , B , and C will be 2, 2, and 4, respectively. This solution is unsatisfactory for carrier A , who could earn 4 by choosing to work alone instead of collaborating with carriers B and C .

AUCTION MECHANISMS

Collaborations among shippers and carriers in the truckload (TL) market aim to exploit the synergies among transportation requests. In a TL network, synergies arise from economies of scope rather than economies of scale [41]. More specifically, the complementarity of the transportation requests allows carriers to execute continuous moves, thereby reducing the need for

asset repositioning. One way to realize the cost savings affiliated with reduced asset repositioning is through transportation procurement auctions.

Transportation procurement auctions are widely used in practice. The auctions, conducted via an electronic marketplace, provide a mechanism for shippers to acquire transportation services from a large number of carriers. The US truckload transportation market is large and highly fragmented [5]. In this large and dynamic market, a fast and reliable auction mechanism is not only convenient but also may deliver significant cost savings as reported in Caplice and Sheffi [41], Moore *et al.* [42], Ledyard *et al.* [43], and Elmaghraby and Keskinocak [44]. TL procurement auctions are usually conducted as a reverse auction. In this setting, a single shipper (the auctioneer) announces the loads to be auctioned and the participating carriers (the bidders) submit their bids for the loads. Finally, the shipper evaluates the bids from the carriers and assigns his loads to the carriers. TL procurement auctions possess several desirable properties. Namely, they do not require information sharing among carriers, they can include additional business considerations such as the number of carriers utilized by a shipper, the capacities of carriers, and threshold volumes [41] and finally, they can be employed as a means for a secondary market among carriers [45].

Compared to other possible auction settings, a combinatorial auction is most suitable for exploiting economies of scope in transportation procurement since it allows carriers to bid for a set of transportation requests rather than bidding only for one load at a time. We refer the reader to De Vries and Vohra [46] and Abrache *et al.* [47] for a thorough review of combinatorial auctions and the associated issues in designing a combinatorial auction mechanism. In a combinatorial truckload procurement auction, the shipper announces the set of loads and invites a set of carriers to bid on these loads. If the shipper allows carriers to bid on all possible combination of the loads, the total number of bundles is equal to 2^n where n is the number of loads being auctioned. As mentioned in Sheffi [48], handling large instances

with a combinatorial auction mechanism is quite challenging: “Real problems, which may include thousands of lanes, dozens or hundreds of carriers, and millions of combination bids, are more challenging and require special code.” To overcome this issue, the shipper may impose a restriction on the number of bundles for which a carrier may submit a bid. For example, the shipper may use a pre-specified bidding language [49]. By designing a bidding language in a certain way, the shipper can influence the carriers to submit bids on a limited number of bundles [50]. However, imposing a very strict restriction on the number of bundles decreases the possibility of exploiting the synergies and hence limits the effectiveness of the combinatorial auction. Clearly, there is a delicate trade-off between the efficiency and implementability of the auction mechanism.

After the shipper auctions the loads, the carriers (who may have preexisting commitments to other loads) decide on the bid prices for the bundles of loads. A carrier’s bid calculation process requires a precise estimation of the marginal cost of each bundle of loads. That is, a carrier has to solve an optimization problem to determine the least cost set of tours to serve its preexisting load commitments plus the bundle of loads in consideration and compute the associated bid. The carriers have to carry out this process for every bundle they anticipate acquiring, unless an implicit bidding approach is employed as described in Chen *et al.* [51] and Beil *et al.* [52]. When the underlying optimization problem is NP-hard, this bid selection process significantly limits the implementation of such auction mechanisms for practical purposes. Using approximation algorithms for the bid selection process of the carriers may improve the computational efficiency, however, this approach will produce suboptimal solutions for the carriers [53,54]. Furthermore, even if the underlying optimization problem is polynomially solvable, the bid determination process may take an exponential effort if no limit is imposed on the number of bundles for which a carrier can submit a bid.

Even if we assume that the carriers can exactly determine the marginal cost of serving a bundle of loads, they are not required to submit this value as a bid to the shippers. Depending on the incentives in a collaborative and competitive environment, the selfish carriers rarely disclose their private information and may lie about their costs to achieve higher benefits. Even in environments where there is a centralized decision maker, the trust issues between participants often threaten the collaboration and hence there is always an inherent risk of losing collaborative opportunities. However, there exist certain auction mechanisms that prevent carriers from lying about their cost information. In Vickery, or second-bid-price auctions, the carriers are guaranteed to bid their true cost value for the bundles of loads. In a single item Vickery auction, the carriers submit their bids to the shipper for the item and the item is awarded to the carrier with the lowest bid. However, instead of being paid the amount of the lowest bid, the winning carrier is paid the amount of the second-lowest bid. In the combinatorial auction extension, a winning carrier is paid its total combined bid for the awarded bundles plus an additional payment. This additional payment is calculated by excluding the particular carrier from the auction, recalculating the total cost of the auction to the shipper, and then subtracting the original cost of the auction (including the particular carrier) to the shipper. We refer to Figliozzi *et al.* [55] and Beil *et al.* [52] for discussions of the implementation of Vickery auctions in transportation procurement markets. In most other auction settings, such as first-bid-price auction, the truthfulness of the carriers is not guaranteed (as in Song and Regan [56] and Chen *et al.* [51]).

After multiple carriers submit their bids on the bundles of loads to the shipper, the shipper solves a winner determination problem to assign bundles to the carriers. The objective of the shipper is to determine the minimum cost allocation of the lanes to the bidding carriers, thereby minimizing the total transportation cost. However, as the bids from various carriers may include bids for overlapping sets of lanes, solving the winner determination problem requires

the solution of a set partitioning-type problem, which is known to be *NP-hard* [46]. Restricting the number of bundles on which a carrier is allowed to bid can reduce the size of the winner determination problem, and employing approximation heuristics for solving the set partitioning problem can further mitigate the complexity of the problem. However, both actions will significantly decrease the effectiveness of the auction mechanism in exploiting the synergies among transportation requests.

We return to the example introduced in the section titled “Centralized Collaboration Mechanisms” to demonstrate how auction mechanisms work in this context. To this end, we slightly modify the example so that the capacity on all four edges is equal to one. Furthermore, they are no longer operated by carrier A. The origin, destination, per-unit revenue, and size of each load is as described in Table 1. In this auction setting, the carriers bid on available capacity on the edges and the winner is determined by solving a winner determination problem (if necessary). If we assume a single item auction setting where each edge is auctioned sequentially, the resulting bids from the carriers are given in Table 2. Note that carrier B does not submit any bids because it cannot generate any revenue by acquiring capacity only on a single edge. The winning bids are carrier A’s bid of 2 for edge (1,3) and carrier C’s bid of 3 for edge (2,4), which generates a total revenue of 5. This example clearly shows that single item auction setting ignores the synergies between auctioned items.

Next, we consider a combinatorial auction setting for this example. Even for this small instance with only three carriers and four “items” on which carriers are bidding, each carrier has to submit a bid for 15 different bundles (if no bid limit is imposed), and the resulting winner determination problem has 45 different bids. In the optimal solution, the winning bids are carrier B’s bid of 6 for edges (1,3) and (3,4) and carrier C’s bid of 3 for edge (2,4), which generates a total revenue of 9. This corresponds to the centralized optimal solution for this instance. Note that if

Table 2. Carrier Bids in Sequential Single Item Auction

Edge	Carrier A	Carrier B	Carrier C
(1,2)	0	0	0
(1,3)	2	0	0
(2,4)	2	0	3
(3,4)	0	0	0

we limit the number of bids from each carrier in order to increase the computational efficiency, we may not achieve the system optimal solution.

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this article, we have described three generic mechanisms that allow for collaboration in the transportation industry. Each of these mechanisms requires different levels of coordination and information sharing among the participating entities while yielding different levels of efficiency. One important conclusion we derive is the fact that a fair and efficient mechanism with limited coordination requirements can be designed using analytical tools that properly take into account economies of scope and scale.

Future research directions include (i) the further study of the mathematical structures of these mechanisms to identify the best possible payoffs and how to achieve them computationally; (ii) understanding the implementation and fairness requirements in different industries and then improving the mechanisms by incorporating these additional constraints; and (iii) understanding how the discussed mechanisms perform under uncertainty and the design of robust mechanisms.

REFERENCES

1. Elemica. Available at <http://www.elemica.com/about/fact-sheet.html>, Accessed 2010.
2. Lufthansa Cargo. Wow alliance successfully established in airfreight market. 2003. Available at <http://www.lhcargo.com/content.jsp?path=0,1,19141,19155,61007,66831>, Accessed 2010.

3. Schiphol Cargo Department. Cargonews. 2004. Available at www.schiphol.com/, Accessed 2010.
4. Kupfer F, Meersman H, Onghena E, *et al.* Air cargo: the difference between success and failure? European Transport Conference. Leeuwenhorst Conference Centre, The Netherlands. Available at <http://www.etcproceedings.org/paper/air-cargo-the-difference-between-success-and-failure>, 2009.
5. Agarwal R, Ergun O. Network design and allocation mechanisms for carrier alliances in liner shipping. Working paper. 2009.
6. Young HP. Cost allocation: methods, principles, applications. Amsterdam: North-Holland Publishing Company; 1985.
7. Borm P, Hamers H, Hendrickx R. Operations research games: a survey. TOP 2001; 9(2):139–199.
8. Nisan N, Ronen A. Algorithmic mechanism design. Games Econ Behav 2001;35:166–196.
9. Kalai E, Zemel E. Generalized network problems yielding totally balanced games. Oper Res 1981;30(5):998–1008.
10. Kalai E, Zemel E. Totally balanced games and games of flow. Math Oper Res 1982; 7(3):476–478.
11. Derks JJM, Tijs SH. Stable outcomes for multicommodity flow games. Method Oper Res 1985;50:493–504.
12. Markakis E, Saberi A. On the core of the multicommodity flow game. EC'03: Proceedings of the 4th ACM Conference on Electronic Commerce. New York: ACM Press; 2003. pp. 93–97.
13. Owen G. On the core of linear production games. Math Program 1975;9(3):358–370.
14. Deng X, Ibaraki T, Nagamochi H. Combinatorial optimization games. Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithm. New York: ACM Press; 1997. pp. 720–729.
15. Agarwal R, Ergun O. Mechanism design for a multicommodity flow game in service network alliances. Oper Res Lett 2008;36:520–524.
16. Gui L, Ergun O. Dual payoffs, core and a collaboration mechanism based on capacity exchange prices in multicommodity flow games. WINE'08: Proceedings of the 4th International Workshop on Internet and Network Economics. Shanghai, China. 2008; pp. 61–69.
17. Ergun O, Kuyzu G, Savelsbergh M. Shipper collaboration. Comput Oper Res 2007; 34(6):1551–1560.
18. Özener OO, Ergun O. Allocating costs in a collaborative transportation procurement network. Transp Sci 2008;42(2):146–165.
19. Sanchez-Soriano J, Lopez MA, Garcia-Jurado I. On the core of transportation games. Math Soc Sci 2001;41(2):215–225.
20. Granot D, Hamers H, Tijs S. On some balanced, totally balanced and submodular delivery games. Math Program 1999;86(2): 355–366.
21. Derks J, Kuipers J. On the core of routing games. Int J Game Theory 1997;26(2): 193–205.
22. Tamir A. On the core of network synthesis games. Math Program 1991;50(1):123–135.
23. Song DW, Panayides PM. A conceptual application of cooperative game theory to liner shipping strategic alliances. Marit Policy Manage 2002;29(3):285–301.
24. Kubo M, Kasugai H. On the core of the network design game. J Oper Res Soc Jpn 1992;35:250–255.
25. Shapley L, Shubik M. The assignment game I: the core. Int J Game Theory 1972;1:111–130.
26. Goemans M, Skutella M. Cooperative facility location games. Annual ACM-SIAM Symposium on Discrete Algorithms; San Francisco (CA). 2000.
27. Tamir A. On the core of a traveling salesman cost allocation game. Oper Res Lett 1989;8(1):31–34.
28. Kuipers J. A note on the 5-person traveling salesman game. Math Method Oper Res 1993;38(2):131–139.
29. Potters JAM, Curiel IJ, Tijs SH. Traveling salesman games. Math Program 1992;53(1):199–211.
30. Engevall S, Gothe-Lundgren M, Varbrand P. The traveling salesman game: an application of cost allocation in a gas and oil company. Ann Oper Res 1998;82:203–218.
31. Gothe-Lundgren M, Jornsten K, Varbrand P. On the nucleolus of the basic vehicle routing game. Math Program 1996;72(1):83–100.
32. Engevall S, Gothe-Lundgren M, Varbrand P. The heterogeneous vehicle-routing game. Transp Sci 2004;38(1):71–85.
33. Ahuja R, Orlin J. Inverse optimization. Oper Res 2001;49(5):771–783.
34. Vinod B. Alliance revenue management. J Revenue Pricing Manage 2005;4(1):66–82.
35. Wright C, Groenevelt H, Shumsky R. Dynamic revenue management in airline alliances. Working paper. 2008.

36. Houghtalen L, Ergun O, Sokol J. Designing mechanisms for the management of carrier alliances. Working paper. 2009.
37. Özener OO, Ergun O, Savelsbergh M. Lane exchange mechanisms for truckload carrier collaboration. *Transp Sci* 2010 (epub ahead 7 July 2010).
38. Vossen T, Ball M. Slot trading opportunities in collaborative ground delay programs. *Transp Sci* 2006;40(1):29–43.
39. Madas M, Zografos G. Airport slot allocation: From instruments to strategies. *J Air Transp Manage* 2006;12:53–62.
40. Houghtalen L, Ergun O, Sokol J. Designing fair allocations for carrier alliances. Working paper. 2009.
41. Caplice C, Sheffi Y. Optimization-based procurement for transportation services. *J Bus Logist* 2003;24(2):109–128.
42. Moore EW, Warmke JM, Gorban LR. The indispensable role of management science in centralizing freight operations at Reynolds-metals-company. *Interfaces* 1991;21(1):107–129.
43. Ledyard JO, Olson M, Porter D, *et al.* The first use of a combined-value auction for transportation services. *Interfaces* 2002;32(5):4–12.
44. Elmaghraby W, Keskinocak P. Combinatorial auctions in procurement. The practice of supply chain management: where theory and application converge. Norwell (MA): Kluwer Academic Publishers; 2004. pp. 245–258.
45. Song J, Regan A. An auction based collaborative carrier network. *Transp Res Part E: Logist Transp Rev* 2003;1763:1–5.
46. De Vries S, Vohra RV. Combinatorial auctions: a survey. *INFORMS J Comput* 2003;15(3):284–309.
47. Abrache J, Crainic T, Gendreau M. Design issues for combinatorial auctions. *4OR: Q J Oper Res* 2004;2(1):1–33.
48. Sheffi Y. Combinatorial auctions in the procurement of transportation services. *Interfaces* 2004;34(4):245–252.
49. Nisan N. Bidding and allocation in combinatorial auctions. Proceedings of the 2nd ACM conference on Electronic commerce; Minneapolis (MN). 2000. pp. 1–12.
50. Rothkopf MH, Pekec A, Harstad RM. Computationally manageable combinatorial auctions. *Manage Sci* 1998;44(8):1131–1147.
51. Chen R, AhmadBeygi S, Beil D, *et al.* Solving truckload procurement auctions over an exponential number of bundles. Technical report. University of Michigan; 1998; Available at [Http://www-personal.umich.edu/amycohn/PAPERS/truckload.pdf](http://www-personal.umich.edu/amycohn/PAPERS/truckload.pdf).
52. Beil D, Chen R, Cohn A. Implicit bidding approach for truckload procurement and other combinatorial auctions. Technical report. Ross School of Business Paper No. 1034. University of Michigan; 2007.
53. Song J, Regan A. Approximation algorithms for the bid construction problem in combinatorial auctions for the procurement of freight transportation contracts. *Transp Res Part B: Meth* 2005;39(10):914–933.
54. Kuyzu G, Ergun O, Savelsbergh M. Bid price optimization for simultaneous truckload transportation auctions. Working paper. 2009.
55. Figliozzi MA, Mahmassani H, Jaillet P. Framework for study of carrier strategies in auction-based transportation marketplace. *Transp Res Rec: J Transp Res Board* 2003;1854:162–170.
56. Song JJ, Regan A. Combinatorial auctions for transportation service procurement - the carrier perspective. *Transp Res Rec: J Transp Res Board* 2004;1833:40–46.

SHORTEST PATH PROBLEM ALGORITHMS

ASHISH K. NEMANI

Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

RAVINDRA K. AHUJA

Innovative Scheduling, Inc., and
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

Shortest path problems (SPPs) play a central role in the field of network flows. They arise frequently in a wide variety of applications either as a stand-alone problem or as a part of more complex problem settings, thereby attracting the attention of practitioners as well as researchers. Several applications and variants of the problem have been discussed in earlier articles. SPPs typically involve a network represented as a directed graph $G = (N, A)$, where N is the set of n nodes and A denotes the set of m arcs between nodes. In a single-source shortest path problem (SSPP), the network contains a distinguished node s , called the *source node*, and each arc $(i, j) \in A$ has an associated weight (cost) w_{ij} , also referred to as the length of the arc. The length of a directed path is defined as the sum of the length of all constituent arcs. The SSPP is used to determine the directed path of the shortest length, from the source node to all other nodes (Fig. 1). In all-pairs shortest path problems (ASPPs), we determine the shortest paths between each pair of nodes. The SPP can also be viewed as a special case of the *minimum cost flow problem* [1], where one unit of flow is to be sent from the source node to all other nodes in an incapacitated network at the minimum cost possible. In this article, we make several assumptions about the network, which are not restrictive in practice and can be

easily relaxed if required. The last assumption is restrictive but all algorithms described in this article terminate after detecting the presence of a negative cycle.

- *Data Integrality.* This is necessary for those algorithms whose complexity depends on data (maximum arc length, C). If required, we can convert a rational arc length into an integer value by multiplying it with a suitably large number.
- *Directed Path from Source Node s to Every Other Node.* If there does not exist a path from s to a node i , we can add an arc (s, i) of a very large length to satisfy this assumption. For an ASPP, the graph is assumed to be strongly connected; this assumption can be satisfied by adding arcs of infinite length between the unconnected nodes.
- *Directed Network.* A network with undirected arcs and nonnegative arc lengths can be easily converted into a directed network by replacing each undirected arc $\{i, j\}$ with two directed arcs, (i, j) and (j, i) , of the same length w_{ij} . It should be observed that an undirected arc with negative length constitutes a negative cycle.
- *No Negative Cycle.* We assume that the network does not contain a directed cycle of negative length. The decision version of the SPP on graphs with negative cycles is NP-complete, and no polynomial time algorithm exists unless $P = NP$. Existing algorithms for the SPP identify shortest walks that happen to be shortest paths when no negative cycles are present, which is not true for general networks. In the presence of negative cycles, one can drive shortest walk lengths to minus infinity by traversing a reachable negative cycle infinitely many times while shortest path lengths are always lower bounded by $-nC$.

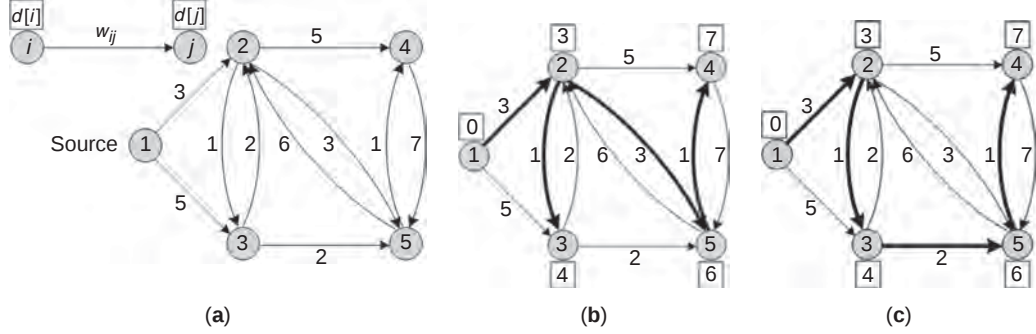


Figure 1. Illustration of a shortest path tree.

Before discussing the details of the algorithms, we briefly review the data structures used to store a shortest path from the source to all other nodes. A naive approach would be to store it in an array of size on the order of $O(n^2)$, as each path may contain at most $n - 1$ arcs. However, with the development of new data structures, such as shortest path tree, it is possible to store it in an array of size $n - 1$. In a shortest path tree, the distance between the selected root node and all other nodes is minimal. It is encoded by storing a predecessor pointer $\text{pred}[j]$ for each node j other than s . Figures 1(b) (c) show two examples of shortest path trees (highlighted arcs) for the network given in Fig. 1(a). It is based on the following property of shortest paths:

Property 1. If the path $P = i_1, i_2, \dots, i_k$ is a shortest path from node i_1 to i_k , then for every $h = 2, 3, \dots, k - 1$, the subpath $P' = i_1, i_2, \dots, i_h$ is the shortest path from i_1 to i_h .

The scope of SPPs differs on the basis of the individual scenario. We discuss five classes of SPP in this article: (i) Shortest paths in acyclic networks; (ii) SSPP with nonnegative arc lengths; (iii) SSPP with general arc lengths; (iv) ASPP with nonnegative arc lengths; and (v) ASPP with general arc lengths. The single destination SPP, which is another variant of the problem, aims to determine the shortest path length from each node to a common destination. It can be converted to the SSPP simply by reversing the direction of the arcs.

In this article, each algorithm for SSPP that we discuss maintains a distance label

$d(i)$ for each node $i \in N$ of the network during its execution. At any intermediate step these labels represent the upper bounds on the shortest paths from the source to node i . At termination, these denote the optimal shortest path lengths. We can now extend Property 1 to achieve the optimality conditions for SSPP.

Theorem 1. For every node $i \in N/\{s\}$, let $d(i)$ denote the distance label at optimality, that is, vector $\mathbf{d}$ denotes the shortest path distance with $d(s) = 0$. Then, a directed path P from the source node to node j is a shortest path if, and only if, $d(j) \leq d(i) + w_{ij} \forall (i, j) \in A$, and $d(j) = d(i) + w_{ij} \forall (i, j) \in P$.

Let us define the reduced arc length w_{ij}^d of an arc $(i, j) \in A$ with respect to the distance labels as $w_{ij}^d = w_{ij} + d(i) - d(j)$. We mention the following properties, which will be useful in algorithmic developments.

Property 2.

- For any directed cycle W , $\sum_{(i,j) \in W} w_{ij}^d = \sum_{(i,j) \in W} w_{ij}$.
- For any directed path P from node k to node l , $\sum_{(i,j) \in P} w_{ij}^d = \sum_{(i,j) \in P} w_{ij} + d(k) - d(l)$.
- If vector $\mathbf{d}$ represents the shortest path distances, $w_{ij}^d \geq 0 \forall (i, j) \in A$.

SHORTEST PATHS IN ACYCLIC NETWORKS

A network is said to be acyclic if it does not contain a directed cycle. We order the nodes

of this network so that $i < j$ for each arc $(i, j) \in A$. This ordering of nodes is called *topological ordering* and can be easily determined by traversing all arcs in $O(m)$ time. It can be observed that we need to conveniently access all arcs emanating from each node $i \in N$, and to facilitate that we maintain an adjacency list $A(i)$ of node i . We now present a simple *reaching algorithm* to find a shortest path in an acyclic network using the topological order and the adjacency list:

Step 1. Initialize $d(s) = 0$ and $d(i) = \infty$, $\forall i \in N/\{s\}$.

Step 2. Examine each node in topological order i and scan arcs in $A(i)$. If for any arc $(i, j) \in A$, $d(j) > d(i) + w_{ij}$, set $d(j) = d(i) + w_{ij}$ and $\text{pred}[j] = i$.

The algorithm is illustrated clearly in Fig. 2. The network in Fig. 2(a) is first topologically ordered as shown in Fig. 2(b). Thereafter, we apply the reaching algorithm by considering each node in topological order. The dotted arcs emanate from the node under examination, while the bold arcs represent the partial paths which are included in the final shortest path. As given in Step 1, we

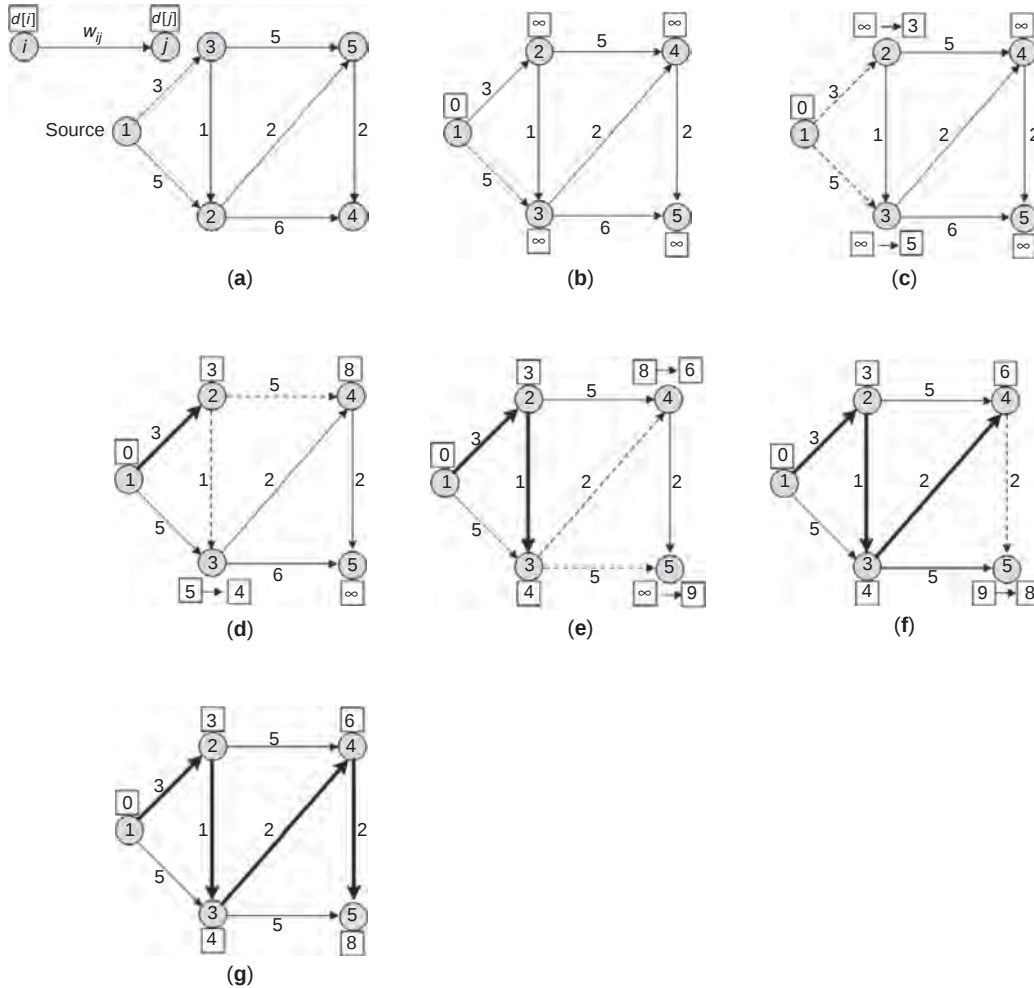


Figure 2. Illustrating the reaching algorithm for the shortest path in an acyclic network.

initialize the distance labels of the source to zero and that of all other nodes to infinity (Fig. 2(b)). Figure 2(c) through (f) examines each node one at a time and updates the distance labels of adjacent nodes. The final shortest path is shown in Fig. 2(g).

The correctness of the reaching algorithm follows from the observation that whenever a node is examined its distance label is optimal. We refer readers to Ahuja *et al.* [2] for the detailed proof. The algorithm solves the problem in $O(m)$ time, as it traverses each arc only once and spends $O(1)$ time on it. We should also note that this is the best possible complexity because any algorithm must examine each arc of the network.

Though the algorithm described above is very simple and effective, it cannot be applied in general cases that contain cycles. However, we utilize the same basic strategy and optimality conditions with some additional steps to solve more complex cases. The next class of problem is the SSPP with nonnegative arc lengths.

SINGLE-SOURCE SHORTEST PATHS WITH NONNEGATIVE ARC LENGTHS

A classical approach to solve the SSPP in a network with nonnegative arc lengths was first proposed by Dijkstra in 1959. At each step, Dijkstra's algorithm maintains a distance label $d(i)$ for each node $i \in N$ and designates it as either temporary or permanent. The distance label corresponding to the

permanent node represents a shortest path from the source, while that corresponding to the temporary node gives an upper bound on the shortest path distance from the source. We start by initiating the distance label of source node to zero and making it permanent. The distance labels of all other nodes are initialized to infinity. The basic idea is the same as in the reaching algorithm. We start from the source s , then examine each node one at a time updating the distance labels of its adjacent nodes. The only difference is that in each iteration of Dijkstra's algorithm the node with the lowest distance label is examined, while in the reaching algorithm it is chosen in topological order. The chosen node is made permanent because none of the arcs from a temporary node can reduce its distance label further due to the nonnegative arc length restriction. The operation of selecting a node with the shortest distance label is referred to as *node selection*; the operation of subsequently updating the distance labels is referred to as *distance update* or *relaxing*.

Dijkstra's algorithm maintains an outwardly directed tree T rooted at the source node that spans all nodes with finite distance labels. Every tree arc, (i, j) , satisfies the property $d(j) = d(i) + w_{ij}$ with respect to the current distance label. The arc (i, j) indicates that, in the current shortest path from source to node j , node i precedes node j . At the termination, the tree becomes the shortest path tree. Figure 3 gives the algorithmic framework of Dijkstra's algorithm. Let N_P and N_T

```

algorithm Dijkstra;
begin
   $N_P := \emptyset; N_T := N;$ 
   $d(i) := \infty \forall i \in N;$ 
   $d(s) := 0$  and  $\text{pred}(s) := 0;$ 
  while  $|S| < n$  do
    begin
      find  $i$  such that  $d(i) = \min\{d(j) : j \in N_T\};$  //Node-Selection();
       $N_P := N_P \cup \{i\};$ 
       $N_T := N_T - \{i\};$ 
      for each  $(i, j) \in A(i)$  do //Relaxing
        if  $d(j) > d(i) + w_{ij}$  then  $d(j) := d(i) + w_{ij}$  and  $\text{pred}(j) := i;$ 
    end;
  end;

```

Figure 3. Framework of Dijkstra's algorithm.

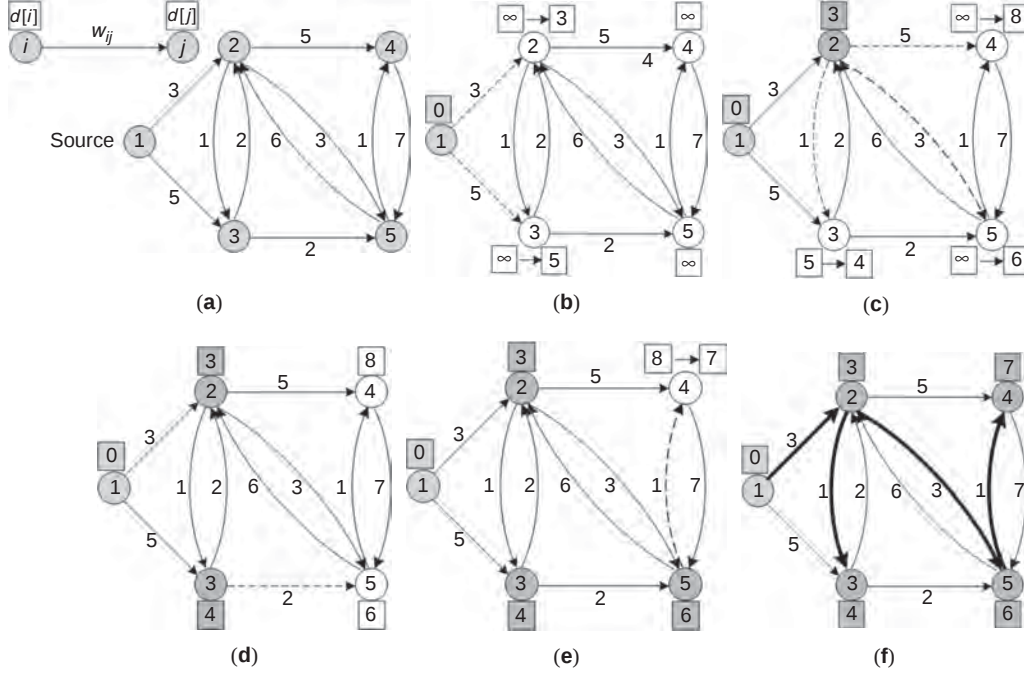


Figure 4. Illustrating Dijkstra's algorithm.

denote the permanent and temporary nodes respectively.

We illustrate Dijkstra's algorithm in Fig. 4 using the network given in Fig. 1(a). The gray nodes are permanent, while the white ones are temporary. The dotted arcs represent the arcs being examined in a particular iteration. As shown in Fig. 4(a), we initialize the distance labels of source to zero and all other nodes to infinity. Thereafter, we examine the source node in Fig. 4(b), and reduce the distance labels of node 2 and node 3 to 3 and 5, respectively. As node 2 has the lowest distance label, it is made permanent and the arcs emanating from it are examined next. The process continues as shown in Fig. 4(b) through (e), terminating with the shortest path as shown in Fig. 4(f).

The correctness of Dijkstra's algorithm can be established by showing that at each stage two properties are always maintained: (i) for each node $i \in N$, $d(i)$ is the shortest path from s to i among all paths that pass through nodes only in N_P , and (ii) for each $i \in N_P$, $j \in N_T$, $d(i) \leq d(j)$. The details

of the proof can be seen in Ref. 2. The worst-case complexity of the algorithm is governed by two main operations: (i) *Node selection*: The algorithm selects one node from the set of temporary nodes whose cardinality decreases by 1 in each iteration. Hence, the overall running time for this operation is $n + (n-1) + \dots + 1 = O(n^2)$. And, (ii) *Relaxing/distance-update*: Each arc is relaxed once, thereby adding $O(m)$ to the total running time. This complexity is based on simple data structures and is appropriate for very dense networks. Researchers have attempted several implementation approaches, and have successfully improved either the worst-case complexity or the running time in practice. We list all of these implementations in Table 1.

The applicability of Dijkstra's algorithm is limited to networks with all nonnegative length arcs. It selects the node with the lowest distance label from the set of temporary nodes and makes it permanent, based on the fact that no temporary node can reduce it any further. This may not be true if a negative arc

Table 1. Running Time of Dijkstra's Algorithm with Various Data Structures

Algorithm	Running Time
Original implementation	$O(n^2)$
Dial's implementation	$O(m + nC)^a$
d -Heap implementation	$O(m \log_d n)$, where $d = m/n$
Fibonacci heap implementation	$O(m + n \log n)$
Radix heap implementation	$O(m + n \log(nC))$

^a C , length of longest arc.

is incident to it from a temporary node. In the next section we discuss several approaches to solve the problem with general arc lengths.

SINGLE-SOURCE SHORTEST PATHS WITH GENERAL ARC LENGTHS

The presence of negative arc weights makes the problem more complex and renders the previous incremental algorithm inapplicable. It should be noted that the presence of negative arcs does not signify the presence of a negative cycle, which would make the problem NP-hard. Therefore, the algorithms in this case should either find a shortest path or terminate with the identification of a negative cycle.

Generic Label-Correcting Algorithm

The shortest path algorithms are generally divided into two groups: *label setting* and *label correcting*. Both approaches iteratively assign distance labels to nodes, but vary in labeling them either permanent or temporary. The label-setting algorithms designate

one node permanent at each iteration, while the label-correcting algorithms consider all nodes temporary until the final iteration. The label-setting algorithm is an incremental algorithm and, therefore, is applicable to only acyclic networks or networks with non-negative arc weights. Dijkstra's algorithm discussed in the section titled "Single-Source Shortest Paths with Nonnegative Arc Lengths" belongs to this class. On the other hand, the label-correcting algorithm continues checking all arcs repeatedly until no improvement is possible. Accordingly, it is applicable to all networks that include general arc lengths but no negative cycle.

The generic label-correcting algorithm, similar to Dijkstra's algorithm, maintains a distance label for every node and its predecessor at each stage. The predecessor index of a node j , $\text{pred}(j)$, represents the node prior to node j in the current shortest path of length $d(j)$. At the termination point, the graph formed by these indices is a shortest path tree if no negative cycle is encountered. The basic concept is to successively update the distance labels of each node until they satisfy the optimality conditions given by Theorem 1. We show the algorithmic framework in Fig. 5.

With Property 2 at optimality, the distance labels must satisfy the nonnegative condition on reduced arc lengths, that is, $w_{ij}^d \geq 0 \forall (i, j) \in A$. The generic label-correcting algorithm selects an arc (i, j) that violates this condition, and decreases the distance label of node j , thereby making its reduced length equal to zero. We illustrate this on the network shown in Fig. 6(a), and have added negative arcs to make it more general. If the algorithm selects the arcs (1, 2), (1, 3), (2, 5), (3, 4), (4, 2), and (2, 5) in this sequence, we obtain the distance labels

Figure 5. Framework of generic label-correcting algorithm.

```

algorithm generic-label-correcting
begin
   $d(s) := 0$  and  $\text{pred}(s) := 0$ ;
   $d(i) := \infty \forall i \in N - \{s\}$ ;
  while there exists an arc so that  $d(j) > d(i) + w_{ij}$  do
     $d(j) := d(i) + w_{ij}$  and  $\text{pred}(j) := i$ ;
  end;
end;

```

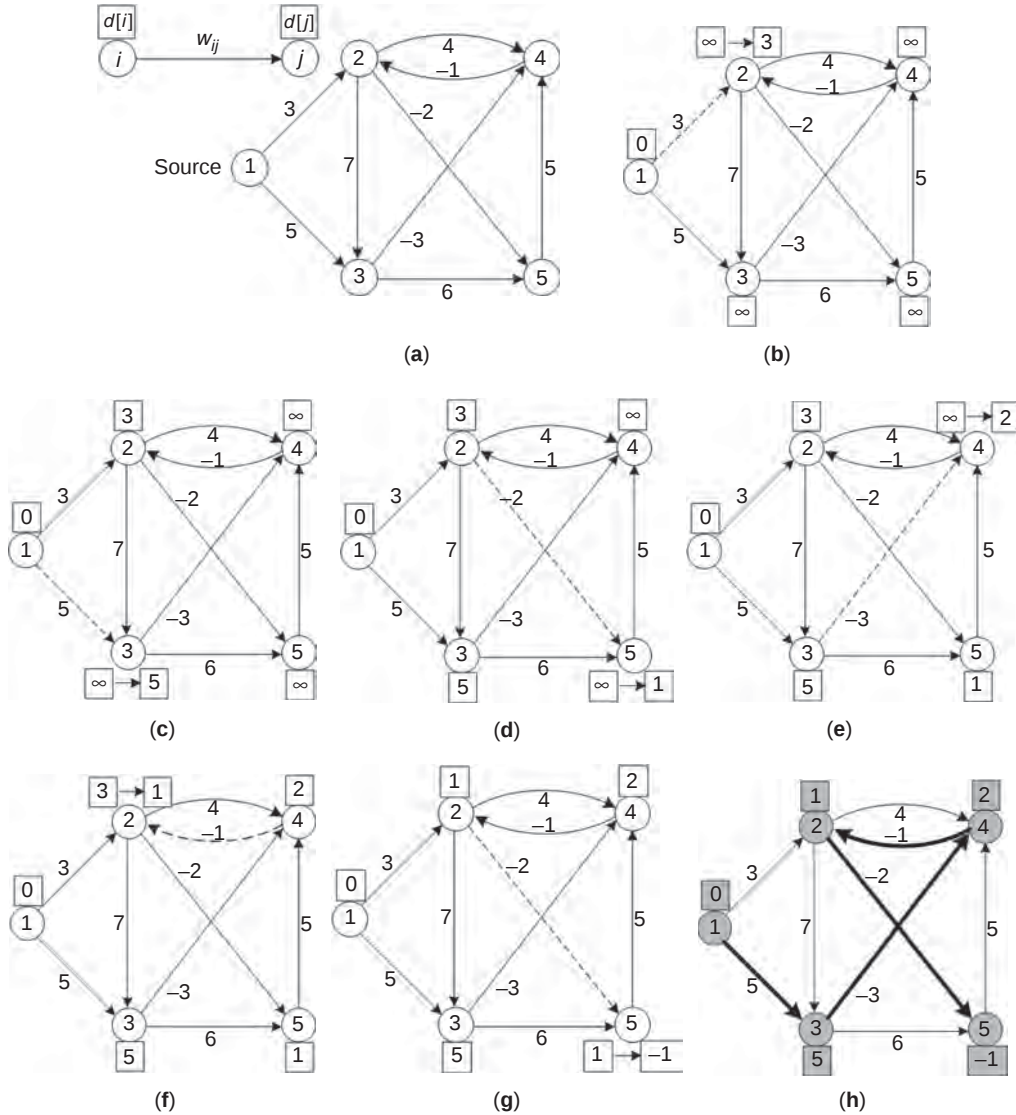


Figure 6. Illustrating the generic label-correcting algorithm.

shown in Fig. 6(b) through (g). At this point, no arc violates its optimality condition and the algorithm terminates with the shortest path tree as shown in Fig. 6(h). Please note that all nodes are temporary until the last iteration, after which all of them simultaneously become permanent.

The correctness of the algorithm follows from the fact that at its termination all conditions of Property 2 are satisfied. We briefly look at the complexity in a special

case where all data are integral. Let C be the longest arc length. It can be observed that any distance label is bounded within the range $[nC, -nC]$, as a path can contain at most $(n - 1)$ arcs of length C . Therefore, any distance label can be updated at most $2nC$ times as each update decreases it by at least one unit because all data are integers. It may also take $O(m)$ time to find the arc violating the optimality condition. Consequently, the total number of updates for all

nodes is at most $2n^2C$, which takes $O(n^2mC)$ time. Without going into detail, we state that in a network with irrational arc lengths the worst-case complexity of the generic label-correcting algorithm is $O(2^n)$ time [2].

The generic label-correcting algorithm does not specify a method for selecting an arc violating Property 2. Therefore, an inefficient approach may require a very high running time. Bellman (1957) and Ford (1956) proposed a method that makes this phase very efficient. We discuss it in the following section.

Bellman–Ford Algorithm

As discussed earlier, the Bellman–Ford algorithm solves the SSPP with general arc lengths but fails if, and only if, the network contains a negative cycle reachable from the source. It either calculates the shortest paths or reports the presence of a negative cycle. We discuss the details about the negative cycle detection later in this section. The algorithm uses the ideas of dynamic programming. Similar to the generic label-correcting algorithm, the Bellman–Ford algorithm maintains the distance labels and predecessor indices of all nodes at each stage. Dijkstra’s algorithm follows an incremental approach for distance updates, considering the arcs emanating only from a set of nodes specifically chosen. Conversely, in the Bellman–Ford algorithm, all arcs are examined and the distance labels violating the optimality conditions are updated. The algorithmic framework is given in Fig. 7.

We illustrate the Bellman–Ford algorithm in Fig. 8 using the numerical example

given in Fig. 6(a). As stated in the algorithm, we initialize the distance label of source to zero and all other nodes to infinity, as shown in Fig. 8(b). In each iteration, we check all arcs violating the optimality conditions and update the distance labels as well as the predecessor indices. In Fig. 8(b), only the arcs emanating from the source are violating the optimality conditions. Therefore, the distance labels of nodes 2 and 3 are reduced from infinity to 3 and 5 respectively. In the next three iterations [Figs 8(c) through (e)], all arcs are scanned repetitively. Finally, we determine the shortest path tree using the predecessor indices as shown in Fig. 8(f). As discussed earlier, none of the nodes are permanent until the last iteration, after which all of them become permanent simultaneously.

We next show that the Bellman–Ford algorithm performs at most $n - 1$ passes (first *for-loop* in Fig. 7) to achieve the optimal distance labels. It can be observed that in each pass all arcs are scanned, taking a total of $O(m)$ time. Therefore, the running time of the algorithm is $O(nm)$. We first show by induction that at the end of the k th pass, the algorithm computes the shortest path distances for all nodes that are connected to the source node by the shortest path consisting of k or fewer arcs. This is true for $k = 1$. Now, suppose that the assertion is true for the k th pass and $d_k(j)$ is the distance label at the k th pass. By definition, $d_k(j)$ is the shortest path from source to node j , provided that a shortest path to it contains k or fewer arcs and otherwise is an upper bound on the shortest path length. Consider a shortest path $s = i_0 - i_1 - i_2 - \dots - i_k - i_{k+1} = j$ from source s

```

algorithm Bellman–Ford;
begin
     $d(s) := 0$  and  $\text{pred}(s) := 0$ ;
     $d(i) := \infty \forall i \in N - \{s\}$ ;
    for each  $k := 1$  to  $n - 1$  do
        for each  $(i, j) \in A$  do //Relaxing
            if  $d(j) > d(i) + w_{ij}$  then  $d(j) := d(i) + w_{ij}$  and  $\text{pred}(j) := i$ ;
        for each  $(i, j) \in A$  do
            if  $d(j) > d(i) + w_{ij}$  then "Network contains negative cycle";
    end;

```

Figure 7. Framework of Bellman–Ford algorithm.

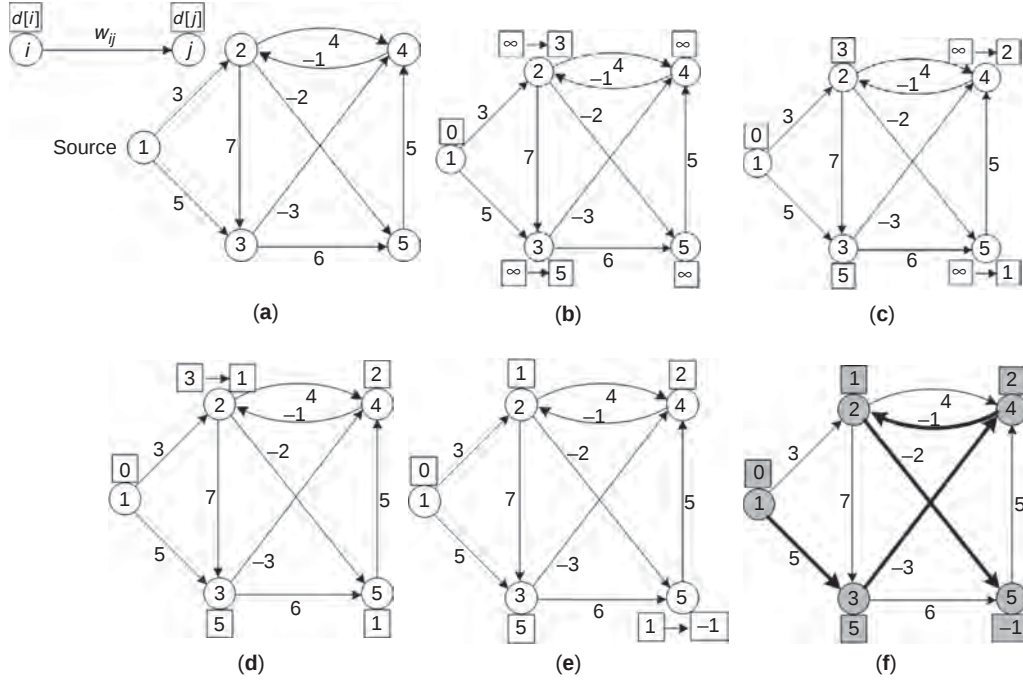


Figure 8. Illustrating Bellman–Ford algorithm.

to node j consisting of $k+1$ arcs, but no shortest path containing fewer than $k+1$ arcs. By Property 1, the path $i_0 - i_1 - i_2 - \dots - i_k$ is a shortest path from source to node i_k ; by induction hypothesis, the distance label of node i_k at the end of the k th pass must be equal to the length of this path. Therefore, in the $k+1$ th pass, we examine the arc (i_k, i_{k+1}) and set the distance label of i_{k+1} equal to the length of the path $i_0 - i_1 - i_2 - \dots - i_k - i_{k+1}$. It establishes our claim and also proves the correctness of the algorithm that, if we set $k = n - 1$ as a path, it can contain at most $n - 1$ arcs.

If a network contains a negative cycle, it can be easily detected by the Bellman–Ford algorithm as shown in the second “for-loop” of Fig. 7. As discussed earlier, each distance label must be at its optimal value after $n - 1$ iterations; therefore, additional distance label updates indicate the presence of a negative cycle. The worst-case complexity of the Bellman–Ford algorithm as shown above is $O(mn)$, as all arcs are examined

in each pass. It can be improved for practical purposes by storing the nodes in a list whose distance labels change in a pass, and then examining the list in a first-in-first-out (FIFO) order in the next pass. We refer readers to Ref. 2 for the details of the FIFO-based algorithm. We now extend the scope of our problem from single-source shortest path to ASPPs.

ALL-PAIRS SHORTEST PATH PROBLEMS

In this section, we consider the problem of finding the shortest path distances between all pairs of nodes in a network. It can be viewed as an extension of SSPP, and any algorithm described earlier can be modified to solve the ASPP. We suggest two approaches for solving this problem: (i) repetitive single-source shortest path algorithms and (ii) Floyd–Warshall’s algorithm. As stated earlier, we assume that the network is strongly connected and that it does not contain a negative cycle. However,

if negative cycles are present, all of these algorithms terminate after their detection.

Repetitive Single-Source Shortest Path Algorithm

We can solve the ASPP by running a single-source shortest path algorithm n times, once for each node as a source. If all arc lengths are nonnegative, we can use any SSPP algorithm, such as Dijkstra's algorithm, taking $f(n, m, C)$ time. Therefore, the total running time of this algorithm is $O(nf(n, m, C))$. However, if the network contains some negative arcs, we can either apply the Bellman–Ford algorithm n times taking $O(n^2m)$ time or apply any algorithm after transforming the network to one with nonnegative arc lengths as discussed next.

We select a node s and apply the Bellman–Ford algorithm to compute the shortest path distances from source to all other nodes. If a negative cycle is detected, the all-pairs shortest path also contains a negative cycle; otherwise, the shortest distance labels denote the shortest path from node s . Now, we replace the current arc lengths with the reduced arc length; that is, $w_{ij}^d = w_{ij} + d(i) - d(j)$ for each arc $(i, j) \in A$. As explained in the introductory section, since the reduced arc lengths are always positive, we can now apply any SSPP algorithm. Once the algorithm terminates, the actual shortest path distance between each pair of nodes, (k, l) , can be determined by adding $d(l) - d(k)$ to the corresponding shortest path distance in the modified network. We differentiate the distance label obtained from each source by maintaining two indices: first for the source and second for the destination. In other words, for each pair of nodes (k, l) , we maintain the distance label $d[k, l]$, which represents an upper bound on the shortest path length from k to l . This method requires $O(nm)$ time for one application of the Bellman–Ford algorithm, and $O(nf(n, m, C))$ time for solving n shortest paths problems with a different source. Therefore, the total running time equals $O(nm + nf(n, m, C)) = O(nf(n, m, C))$. For example, the running time with the Fibonacci heap implementation of Dijkstra's algorithm is $O(mn + n^2 \log n)$.

Floyd–Warshall Algorithm

The Floyd–Warshall algorithm uses the dynamic programming approach to solve the all-pairs SPPs. Let $d_k[i, j]$ represent the shortest path length from node i to node j using only the nodes $1, 2, \dots, k-1$ as the internal nodes. The algorithm repetitively computes $d_1[i, j]$, $d_2[i, j]$, and so on, until $d_{n+1}[i, j]$ for all pairs (i, j) of nodes. It is clear that $d_{n+1}[i, j]$ represents the optimal shortest path lengths, as any node can be the internal node. Therefore, the recursive relation, which is the core of any dynamic programming approach, is given by

$$d_k[i, j] = \begin{cases} w_{ij} & \text{if } k = 0 \\ \min(d_{k-1}[i, j], d_k[k, j] + d_k[k, j]) & \text{if } k \geq 1. \end{cases}$$

The relation at $k = 0$ is obvious as we do not include any internal nodes. A shortest path that uses only nodes $1, 2, \dots, k$ either does not pass through node k , in which case $d_{k-1}[i, j]$ is minimum, or it does, in which case $d_k[i, k] + d_k[k, j]$ is minimum. It also maintains the predecessor indices $\text{pred}[i, j]$, for each node pair, which denotes the last node prior to node j in the current shortest path from node i to node j . This helps in building shortest path trees for each source node. The formal description of the algorithm is given in Fig. 9.

```

algorithm Floyd–Warshall;
begin
     $d[i, j] := \infty$  and  $\text{pred}[i, j] := 0; \forall [i, j] \in N \times N;$ 
     $d[i, j] := 0 \forall i \in N;$ 
    for each arc  $(i, j) \in A$  do
         $d[i, j] := w_{ij};$  and  $\text{pred}[i, j] := i;$ 
    for each  $k := 1$  to  $n$  do
        for each  $[i, j] \in N \times N$  do
            if  $d[i, j] > d[i, k] + d[k, j]$  then
                begin
                     $d[i, j] := d[i, k] + d[k, j];$ 
                     $\text{pred}[i, j] := \text{pred}[k, j];$ 
                end;
    end;

```

Figure 9. Framework of the Floyd–Warshall's algorithm.

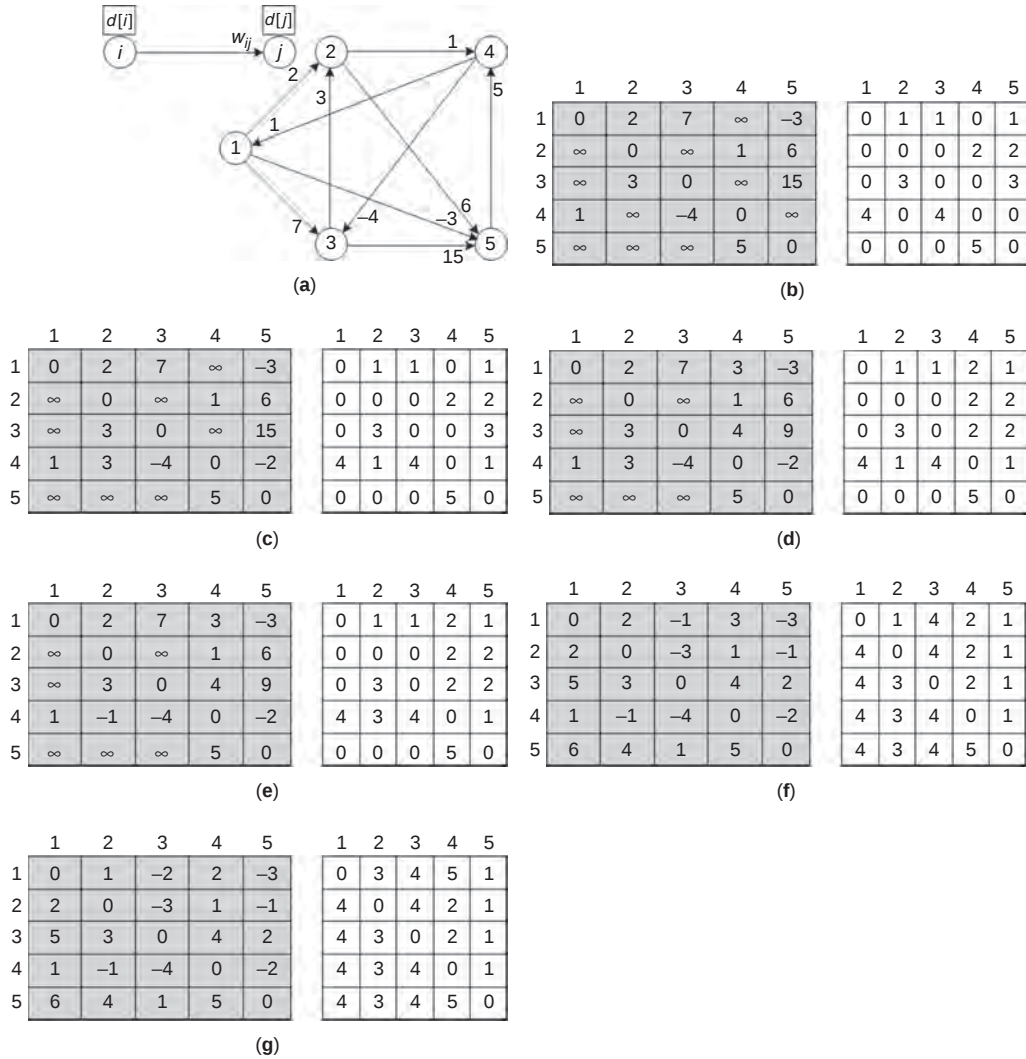


Figure 10. Illustrating the Floyd-Warshall's algorithm.

It is clear from Fig. 9 that the Floyd–Warshall's algorithm performs n passes, increasing the number of possible internal nodes in each successive pass. Each pass examines and updates the distance label of every pair of nodes requiring $O(n^2)$ time. Therefore, the total time required for the algorithm is $O(n^3)$. We explain it further using a numerical example given in Fig. 10(a), a general graph with negative length arcs but with no negative cycle. We present $d_k[i, j]$ (shaded matrix) and $\text{pred}_k[i, j]$

(plain matrix) matrices for all $[i, j] \in N \times N$ at each (k) of the $n = 5$ passes. In Fig. 10(b), $d_1[i, j]$ and $\text{pred}_1[i, j]$ (the distance label and predecessor of node j in the current path from node i) for each pair (i, j) of nodes are given when no internal node is allowed in the path. It can be observed that each value in the first distance matrix corresponds to the length of an arc. In Fig. 10(c), we allow only node 1 as the internal node in the paths. Therefore, the shortest path length from node 4 to node 2 reduces to 3 and, from node

Table 2. Recent Algorithms for SSPP

Algorithm	Running Time
Bertsekas [3]	$O(mn)$
Goldberg [5]	$O(m + n \log C)^a$
Mitra <i>et al.</i> [7]	$O(mk + n)^b$
Mitra <i>et al.</i> [9]	$O(mk/2 + n)$
Raman [11]	$O(m + n\sqrt{\log n \log \log n})$
Raman [13]	$O(m + n(w \log n)^{1/3})$
Thorup [15]	$O(m + \log \log n)$
Meyer [17]	$O(n + m)$
Moriya <i>et al.</i> [19]	$O(m)$
Thorup [21]	$O(m)$

^a C , length of the longest arc.^b k , ratio of the shortest and longest arc length.

4 to node 5, it reduces to -2 . In the next four passes Fig. 10(d)–10(g), we successively allow nodes 2, 3, 4, and 5 to be considered as the internal nodes of the paths. Finally, we achieve all-pairs shortest path distances in Fig. 10(g). We use the predecessor matrix to trace the shortest paths between each pair of nodes. For example, the shortest path between node pair (1, 2) can be traced as 1-5-4-3-2 and its total weight is 1.

As discussed earlier, the presence of a negative cycle makes the problem NP-hard. In that case, the Floyd–Warshall algorithm terminates with the negative cycle detection in $O(n^3)$ time. We incorporate two tests at each distance label update: (i) if $i = j$, check whether $d[i, i] < 0$; and (ii) if $i \neq j$, check whether $d[i, j] < -nC$. If either of these two tests returns true, the network contains a negative cycle.

Though the Floyd–Warshall’s algorithm is very efficient, several researchers have attempted to develop algorithms with an improved worst-case and average-time complexity. We list some recent developments in the single-source shortest path in Table 2 and in the all-pairs shortest path in Table 3. It should be noted that with new advances in the computational fields, the complexity of algorithms also changes.

SUMMARY

SPPs are one of the most fundamental computational problems, both from theoretical and practical perspectives. In this article, we

Table 3. Recent Algorithms for ASPP

Algorithm	Running Time
Johnson’s algorithm [4]	$O(n^2 \log n + mn)$
Pettie <i>et al.</i> [6]	$O(mn \log \alpha(m, n))^a$
Pettie [8]	$O(n^2 \log \log n + mn)$
Fredman [10]	$O(n^3 (\log \log n / \log n)^{1/3})$
Takaoka [12]	$O(n^3 \sqrt{\log \log n / \log n})$
Dobosiewicz [14]	$O(n^3 / \sqrt{\log n})$
Han [16]	$O(n^3 (\log \log n / \log n)^{5/7})$
Takaoka [18]	$O(n^3 (\log \log n)^2 / \log n)$
Zwick [20]	$O(n^3 / \sqrt{\log \log n / \log n})$
Feder and Motwani [22]	$O(n^3 / \log n)$
Chan [23]	$O(n^3 / \log n)$
Shoshan and Zwick [24]	$O(n^{2.376} C)$
Zwick [25]	$O(n^{2.575} C^{0.681})$

^a $\alpha(m, n)$, inverse Ackermann’s function.

have studied several variants of the problem, each with special characteristics: (i) shortest paths in a directed acyclic network; (ii) single-source shortest paths with nonnegative arcs; (iii) single-source shortest paths with general arcs; (iv) all-pairs shortest paths with nonnegative arcs; and (v) all-pairs shortest paths with general arcs. We start with the most restrictive class of the problem (i) and gradually evolve toward the most general version (v). We have discussed widely studied algorithms for each variant of the problems. All of these algorithms are iterative and rely on the same basic concept: distance labels (upper bound on the shortest paths). They also maintain predecessor indices for all nodes to construct the shortest path trees. Finally, we provide a brief overview of the recent advances in SPPs.

REFERENCES

1. Ahuja RK, Magnanti TL, Orlin JB. Minimum cost flows: basic algorithms. In: Network flows: theory, algorithms, and applications. New Jersey: Prentice Hall; 1993. pp 294–357.
2. Ahuja RK, Magnanti TL, Orlin JB. Shortest paths: label setting algorithms. In: Network flows: theory, algorithms, and applications. New Jersey: Prentice Hall; 1993. pp. 93–132.
3. Bertsekas DP, editor. The shortest path problems. In: Network optimization: continuous and discrete models. Massachusetts: Athena Scientific; 1998. pp. 51–114.

4. Cormen TH, Leiserson CE, Rivest RL, Stein C. All-pair shortest paths. In: *Introduction to algorithms*. 2nd ed. New York: McGraw-Hill; 2001. pp. 620–636.
5. Goldberg AV. A simple shortest path algorithm with linear average time. Technical Report STAR-TR-01-03. InterTrust Technologies Corp; 2001.
6. Pettie S, Ramachandran V. A shortest path algorithm for real-weighted undirected graphs. *SIAM J Comput* 2005;34:1398–1431.
7. Mitra PP, Hasan R, Kaykobad M. A linear time algorithm for single source shortest path problem. Bangladesh: ICCIT; 2000.
8. Pettie S. A new approach to all-pairs shortest paths on real-weighted graphs. *Theor Comput Sci* 2004;312:47–74.
9. Mitra PP, Hasan R, Kaykobad M. On linear time algorithm for SSSP Problem. Working paper. Bangladesh: Bangladesh University of Engineering & Technology.
10. Fredman ML. New bounds on the complexity of the shortest path problem. *SIAM J Comput* 1976;5:49–60.
11. Raman R. Priority queues: small, monotone and trans-dichotomous. In: 4th annual European symposium on algorithms (ESA). Volume 1136, LNCS, London (UK): Springer; 1996. pp. 121–137.
12. Takaoka T. A new upper bound on the complexity of the all pairs shortest path problem. *Inform Process Lett* 1992;43: 195–199.
13. Raman R. Recent results on the single-source shortest paths problem. *ACM SIGACT News* 1997;28(2):81–87.
14. Dobosiewicz W. A more efficient algorithm for the min-plus multiplication. *Int J Comput Math* 1990;32:49–60.
15. Thorup M. On RAM priority queues. *SIAM J Comput* 2000;30:86–109.
16. Han Y. Improved algorithm for all pairs shortest paths. *Inform Process Lett* 2004; 91:245–250.
17. Meyer U. Single-source shortest-paths on arbitrary directed graphs in linear average-case time. In: *Proceedings of the 12th Annual Symposium on Discrete Algorithms*. Philadelphia (PA): ACM- SIAM; 2001. pp. 797–806.
18. Takaoka T. A faster algorithm for the all-pairs shortest path problem and its application. In: *Proceedings of the 10th International Conference on Computing and Combinatorics*. Volume 3106, *Lecture Notes in Computer Science*. Berlin: Springer; 2004. pp. 278–289.
19. Moriya E, Tsugane K. Optimally fast shortest path algorithms for some classes of graphs. *Int J Comput Math* 1998;70:297–317.
20. Zwick U. A slightly improved sub-cubic algorithm for the all pairs shortest paths problem with real edge lengths. In: *Proceedings of the 15th International Symposium on Algorithms and Computation*. Volume 3341, *Lecture Notes in Computer Science*. Berlin: Springer; 2004. pp. 921–932.
21. Thorup M. Undirected single-source shortest paths with positive integer weights in linear time. *J ACM* 1999;46:362–394.
22. Feder T, Motwani R. Clique partitions, graph compression and speeding-up algorithms. *J Comput Syst Sci* 1995;51:261–272.
23. Chan TM. All-pairs shortest paths with real weights in $O(n^3/\log n)$ time. *Algorithmica* 2008;50:236–243.
24. Shoshan A, Zwick U. All-pairs shortest paths in undirected graphs with integer weights. *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science*, New York: 1999. pp. 605–614.
25. Zwick U. All-pairs shortest paths using bridging sets and rectangular matrix multiplication. *J ACM* 2002;49:289–317.

SIMPLEX-BASED LP SOLVERS

ISTVÁN MAROS

Department of Computing, Imperial
College London, London, UK

Department of Computer Science and
Systems Technology, University
of Pannonia, Veszprém, Hungary

Two main algorithms are used to solve linear programming (LP) problems, namely, the simplex method (SM) and the interior point methods. As both are needed in practice (not discussed here), their development has been a great challenge which initiated substantial research activities. Nowadays, proper implementations of both types of solvers are very capable and can solve large-scale LP problems reliably and efficiently. This article concentrates on the computational techniques of the SM, and discusses the structure and algorithmic components of state-of-the-art simplex-based solvers. The reader is assumed to be familiar with the theory of the SM.

SUMMARY OF THE REVISED SIMPLEX METHOD

The SM has two main versions. They are the primal and the dual SMs. The usefulness of both necessitates the inclusion of them in the toolbox of an advanced optimization system.

Modern implementations of the SM still follow the principles and logic of the original version introduced by Dantzig [1]. However, every algorithmic component has undergone a serious revision and the newly developed advanced techniques are very sophisticated and efficient.

The key to solving large-scale LP problems is the use of the *revised simplex method* (RSM) and exploiting sparsity of the emerging matrices (and vectors). It is well known that the RSM keeps the matrix of constraints in original (sparse) form. In early implementations, the simplex operations

were performed using the *product form of the inverse* (PFI) of the basis [2]. Later the LU decomposition for LP (first suggested by Markowitz [3]) proved to be superior and is used nearly exclusively in commercial and academic solvers.

To be able to discuss the advanced simplex techniques, the problem to be solved and the algorithmic steps of the RSM have to be reviewed.

The problem considered is supposed to be in a computational form:

$$\begin{array}{ll} \min & z = \mathbf{c}^T \mathbf{x}, \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}, \end{array} \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{c}, \mathbf{x}, \mathbf{l}, \mathbf{u} \in \mathbb{R}^n$, and $\mathbf{b} \in \mathbb{R}^m$. Some components of $\mathbf{l}$ and $\mathbf{u}$ can be $-\infty$ or $+\infty$, respectively. $\mathbf{A}$ contains a unit matrix $\mathbf{I}$ corresponding to the added *logical variables* that make every constraint an equality. Therefore, $\mathbf{A} = [\mathbf{I}, \bar{\mathbf{A}}]$ which implies $\text{rank}(\mathbf{A}) = m$. The original variables of the problem (that multiply the columns of $\bar{\mathbf{A}}$) are called *structural variables*.

This problem can be solved by the primal- or the dual SM. Using any of them provides solution to both the primal and dual *problems*. It is useful to emphasize that both algorithms work on problem (1) (which can be called *primal problem*) but under different assumption and use different rules for the iterations.

The *main algorithmic steps* of an iteration of the primal (dual) SM are summarized below.

Assume a primal (dual) feasible basis $\mathbf{B}$ is given. Matrix $\mathbf{A}$ is logically partitioned as $\mathbf{A} = [\mathbf{B} \ \mathbf{R}]$, where $\mathbf{B}$ contains the basic columns and $\mathbf{R}$ the rest.

1. *Pricing*. Check if the current basis is optimal. If yes, algorithm terminates. Otherwise, select an incoming (dual: outgoing) variable x_q (dual: x_{Bp} , the p th basic variable).

2. *Transform Candidate Column (Row).* Determine the updated form of column q , $\alpha_q = \mathbf{B}^{-1}\mathbf{a}_q$ (dual: nonbasic part of row p , $\rho^p = \mathbf{e}_p^T \mathbf{B}^{-1}\mathbf{R}$).
3. *Ratio Test.* Perform primal (dual) ratio test with eligible positions and determine the primal (dual) steplength. If it is $\mp\infty$ the solution is unbounded, algorithm terminates.
4. *Updating.* Perform transformations and updating as required by the algorithm used.

Remarks:

- In some steps additional computational work is necessary if specific algorithmic techniques are used (details later).
- To find a first feasible solution phase-1 algorithms can be defined that use practically the above algorithmic steps with slight modifications.

SPECIAL FEATURES OF LARGE-SCALE LP PROBLEMS

While the original version of the SM can be coded in about 30 lines of a high level programming language (Fortran, C/C++), an implementation of this type is utterly useless for solving real-life (industry strength) LP problems.

Such problems are large in size and possess some important features. In brief, they are sparsity, structure, fill in, and susceptibility to numerical troubles. To properly handle these features sophisticated data structures and algorithmic units are needed in order to successfully solve real large-scale problems of several hundreds of thousands constraints and variables.

The most important of the special features is *sparsity* (often referred to as *low density*). Sparsity means that the ratio of nonzero elements in an entity (matrix or vector) to the total number of elements is very small. An interesting experience shows that, on an average, there are 5–10 nonzeros in a column of $\bar{\mathbf{A}}$ independent of the actual size of the problem. It implies that larger problems are increasingly more sparse. According to this “rule” the average density is not more than 0.1% for problems with around 10,000 rows. There is a collection of LP problems, which is widely used for testing and comparing algorithms. It is available on the internet from netlib/lp/data.

Many large-scale LP problems possess some *structure*, like the ones that represent dynamic models, spatial distribution, or uncertainty. Such problems contain several, often identical, blocks that are just loosely connected by some constraints and/or variables having nonzero coefficients in the rows/columns of several blocks. Structure

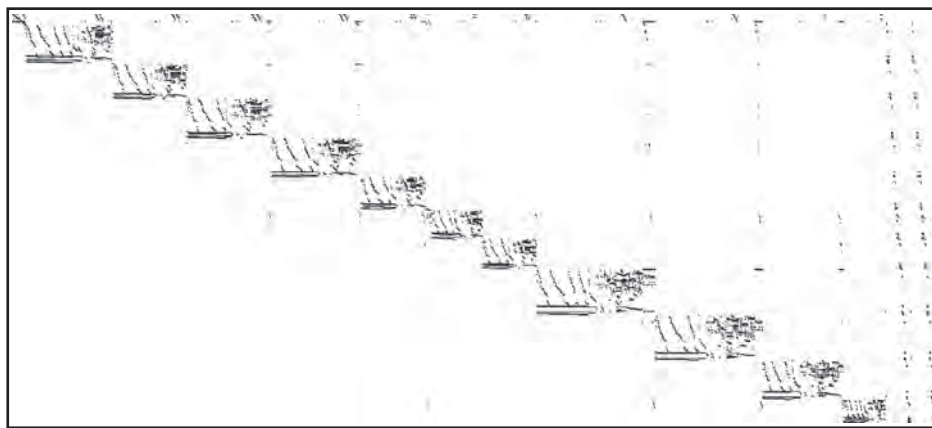


Figure 1. Nonzero pattern of GREENBEA.MPS. Problem size: 2393 rows (including the objective function), 5405 columns, 31,499 nonzeros.

can be visualized by displaying the *nonzero pattern* of the $\mathbf{A}$ matrix. There exist several software tools that enable such visualization. These viewers are very useful tools as they can reveal unknown properties of LP models and also can suggest good solution strategies. Figure 1 shows the nonzero structure of netlib problem GREENBEA.MPS.

The third characteristic feature can be observed during simplex iterations. It is called *fill in*. If the tableau form of the SM is used it is updated by an *elementary transformation matrix* (ETM) $\mathbf{E}$ after every basis change. Although the matrices of large-scale problems are sparse, they tend to fill in. It means that, as a result of premultiplying (the already transformed) $\mathbf{A}$ by an $\mathbf{E}$, nonzeros can appear in places of zeros. Nonzeros can spread quickly and soon the entire tableau can be full. It may cause serious difficulties. For instance, in case of a $10,000 \times 20,000$ matrix, the original number of nonzeros can be around 100,000 (five per column); these many numbers can easily be accommodated in the memory of a personal computer. However, if it fills up then 200,000,000 nonzeros have to be stored and transformed in subsequent iterations, which is prohibitively large. This is where the RSM comes in (see later).

The fourth important feature is *susceptibility to numerical troubles*. It is not a surprise as this phenomenon is quite well known in computational linear algebra. Since cancelation, rounding, and truncation errors can accumulate, the solution algorithm can go wrong not only numerically but also logically (e.g., feasible problem declared infeasible, wrong basis declared optimal, and solution found unbounded instead of optimal). Numerical issues have triggered substantial improvements in computational techniques of the SM.

ADVANCED IMPLEMENTATION OF THE SIMPLEX METHOD

The subsequent discussion is mostly based on Maros [4]. Readers interested in more details are referred to this research monograph.

Requirements of LP Solvers

As the development of LP software is a major effort, it is reasonable to expect that the codes possess some desirable properties. The most important ones are briefly summarized below.

Robustness. A wide range of problems can be solved successfully. If the attempt is unsuccessful, a meaningful answer is returned (e.g., unrecoverable numerical troubles). Robustness implies reliability and accuracy.

Efficiency. In simple terms, efficiency means high execution speed.

Capacity. Large and difficult problems can be solved on a given computer configuration.

Portability. The system can be used on different hardware and software platforms.

Memory Scalability. The solver can adjust to the size of the main memory and operate reasonably efficiently if at least a minimum amount of memory is available. (The solver is not critically memory-hungry.) At the same time, it can adapt well if more memory is allocated.

Extensibility. The code permits both the simple addition of new features internally to the product and the straightforward integration of that product as a building block for other code.

These properties can be ensured partly during algorithm design and partly in the implementation phase.

Representation of LP Problems

LP problems are usually created by some modeling systems that provide a file containing problem data ($\mathbf{A}$, $\mathbf{c}$, $\mathbf{l}$, $\mathbf{u}$, and $\mathbf{b}$) in a format that is understood by a solver. It is called the *external representation* of LP problems. Interestingly, the only internationally recognized *de facto* standard is the MPS input format which was created in the 1950s. Though it is

considered very outdated, every solver is prepared to accept problems in this format. Thus, it is the means of communicating problems among solvers.

The original version is called *fixed MPS format* that uses fixed format records of no more than 80 characters. It reflects the ancient punched card structure. Rows and columns of the matrix of constraints are identified by names of no more than eight characters. The file is subdivided into sections introduced by an indicator record containing a section name (NAME, ROWS, COLUMNS, RHS, RANGES, BOUNDS) or end of file (ENDATA).

Indicator records, names, and numerical values have to be given in well-defined fixed positions. Only the nonzero elements have to be supplied in a column-wise order, each column starting with a new column name. Within a column, each nonzero is identified by its position (row name). Components of vectors (**c**, **l**, **u**, and **b**) are also identified by column or row names. The objective row must be included in the matrix as a nonbinding constraint (row). Surprisingly, the sense of optimization (min or max) cannot be defined in the MPS file. It must be given as a run parameter at solution time.

The heavy restrictions of the MPS format have led to the acceptance of a bit relaxed version of it, which is usually referred to as *free MPS format*. Here, the length of names is not limited; each field is separated from the next by one or more space characters. However, data must appear in the same sequence as in the fixed format. The free format is not standardized.

The first thing a solver does is reading the MPS file and converting its contents into an *internal representation*. This is logically quite a tricky operation that must be very quick. It calls for a fast MPS input unit, which is a “must have” for every advanced solver. During conversion, different statistics such as the number of rows/columns, number of nonzeros and density, number of variables by type, and some measures of the magnitude of elements are also created.

Preprocessing

Usually the immediate internal form of the constraints is

$$L_i \leq \sum_{j=1}^n a_{ij}x_j \leq U_i, \quad i = 1, \dots, m$$

$$\ell_j \leq x_j \leq u_j \quad j = 1, \dots, n.$$

By preprocessing the problem is brought into an equivalent form that can be solved more efficiently and accurately. The main aims are to reduce the size of the problem, detect infeasibility/unboundedness of the problem without solving it, and improve numerical characteristics. Typical actions of preprocessing are presolve and scaling.

Presolve. During *presolve*, a sequence of increasingly stringent tests is performed. The most frequently used test can (i) identify contradicting individual bounds, (ii) detect empty rows, (iii) detect empty columns, (iv) identify singleton rows and convert them into individual bounds, (v) remove fixed variables, (vi) identify redundant and forcing constraints, (vii) tighten/relax individual bounds, (viii) identify implied free variables, (ix) detect singleton columns to impose constraints on their reduced costs, (x) identify dominated variables, and (xi) reduce density of the matrix.

The reduction in problem size, number of nonzeros, and density can be quite substantial.

Scaling. After presolve, it is often advisable to perform *scaling* on the problem. It can improve the numerical characteristics. However, there is no guarantee for the success but in most cases it is worth doing it.

Scaling in the LP context is a linear transformation of the matrix of constraints whereby each row and each column is multiplied by a real number called *scale factor*. It entails the adjustment of the right-hand side, objective function, and bounds. Formally, scaling is equivalent to pre- and postmultiplying **A** by diagonal matrices, such that

$$\mathbf{RAC}\bar{\mathbf{x}} = \mathbf{Rb},$$

with $\mathbf{R} = \text{diag}(R_1, \dots, R_m)$, $\mathbf{C} = \text{diag}(C_1, \dots, C_n)$, where R_i and C_j are the scale factors, and $\mathbf{C}\bar{\mathbf{x}} = \mathbf{x}$. From the latter

$$\bar{\mathbf{x}} = \mathbf{C}^{-1}\mathbf{x}.$$

Obviously, the finite individual bounds of $\bar{x}_j$ are also transformed to u_j/C_j and, similarly, the explicit lower bounds, if given, to ℓ_j/C_j . The objective row is not scaled row-wise. However, the computed column scale factors can be applied to it in which case the objective values of the scaled and unscaled problems are the same. It can be useful when the iterations are monitored.

The importance of scaling in computational linear algebra has long been known. Unfortunately, its role in LP is not that well understood. There are examples where scaling makes it possible to solve an otherwise unsolvable problem but there are other examples where just the opposite occurs: a problem cannot be solved when scaled but becomes easily solvable without scaling. The main difference between traditional linear algebra and LP is that the former works with the matrix of a fixed set of equations while in LP the set of (basic) equations keeps changing and each of them may require a different scaling. What can be done cheaply in LP is a one-off scaling of $\mathbf{A}$ which then applies to all bases selected by the SM.

There are several methods to determine generally good row and column scale factors.

Equilibration. It is a well-established standard method. Each row is scaled to make the largest absolute element equal to 1. It is achieved by multiplying the row with the reciprocal of the largest absolute element (so the row scale factors are these reciprocals). It is followed by a similar scaling for the columns (i.e., the column scale factors are the reciprocals of the absolute maximum values in a column). As a result, the largest absolute element in each row and each column will be 1.

Geometric Mean. For each row, the

$$\left(\max_j \{|a_{ij}|\} \times \min_j \{|a_{ij}|\} \right)^{\frac{1}{2}}$$

quantity is computed and its reciprocal (or its nearest power of 2 to avoid rounding error) is used to multiply the nonzeros of the row. A similar action is taken for each column.

Arithmetic Mean. Each row is multiplied by the reciprocal of the arithmetic mean of the nonzero elements in the row (or the nearest power of 2), followed by a similar action for the columns.

In practice, these simple methods are used in combination, one after the other. If round-off is not an issue, the last method in a combination is always equilibration. The reason for it is that it creates some ground to setting some absolute tolerances for the solver. As such, a typical combination is geometric mean followed by equilibration.

Curtis and Reid [5] have proposed a scaling that is optimal in some sense. It requires the solution of a quadratic optimization problem which may sound too complicated. However, in practice just a few iterations give very good approximations and thus the computational demand is not very excessive and the quality is usually very good. For details the reader is referred to Curtis and Reid [5] or Maros [4] (pp. 113–116).

After presolve and scaling, the remaining problem is augmented with a unit matrix of logical variables that make every constraint an equality, as in Equation (1).

When the problem is solved every action of preprocessing has to be undone so that the solution is obtained in terms of the original problem.

Factorization of the Basis

The two main matrix–vector operations in the SM are computing $\alpha = \mathbf{B}^{-1}\mathbf{a}$ and $\pi^T = \mathbf{r}^T\mathbf{B}^{-1}$ with some $\mathbf{a}$ and $\mathbf{r}$ vectors. In both cases the inverse of the basis is used. The required α and π vectors can also be determined as the solutions of two systems of linear equations if the above equations are rearranged: $\mathbf{B}\alpha = \mathbf{a}$ and $\mathbf{B}^T\pi = \mathbf{r}$.

In the first case, the inverse of the basis or some equivalent form is needed that can be used to perform the above operations. In practice, the PFI turned out to be the most appropriate equivalent representation. $\mathbf{B}^{-1}$ is written as $\mathbf{B}^{-1} = \mathbf{E}_s\mathbf{E}_{s-1} \cdots \mathbf{E}_1$, where each

$\mathbf{E}_i$ is an ETM that differs from the unit matrix in just one column which is usually referred to as *eta* (η) vector. Only these etas need to be stored together with their position in the matrix. In each iteration a new ETM (eta) is created from the updated incoming column which is simply added to the list of ETMs. The introduction of PFI [2] has revolutionized the SM and it became capable of solving real large sparse LP problems.

As some weaknesses of PFI have surfaced (numerical difficulties, quick growth of the number of nonzeros in the etas, need for frequent reinversion) attention has turned to the second possibility, that is, the solution of systems of equations. The breakthrough was the introduction of the LU representation of the basis, $\mathbf{B} = \mathbf{LU}$, where $\mathbf{L}$ is a lower and $\mathbf{U}$ is an upper triangular matrix. Operations with triangular matrices are very simple.

$\mathbf{B}\alpha = \mathbf{a}$ can be rewritten as $\mathbf{LU}\alpha = \mathbf{a}$. To obtain α , first $\mathbf{Ly} = \mathbf{a}$ is solved for $\mathbf{y}$ using *forward substitution* then $\mathbf{U}\alpha = \mathbf{y}$ is solved for α using *back substitution*. Both operations can be performed very efficiently once the LU form is available.

Creating the LU form is logically more complicated than that of PFI, in particular when we aim at a sparse form. However, efficient implementations exist [6]. There is one more difficulty. After every basis change the LU form is usually destroyed, and so must be restored. There are several techniques that can do this [7–11]. Details of efficient implementation of maintaining LU can be found in Suhl and Suhl [12].

Whether PFI or LU is used and updated after every basis change, the accuracy of them must be restored by periodical reinversion or refactorization, respectively. Experience shows that the advisable reinversion frequency for PFI is 30–50 iterations and the refactorization frequency is 100–150 iterations. The frequency is usually a parameter that has a good default value which, however, can be changed by the user. There are some requirements that these procedures must satisfy: (i) speedy operation to cause minimum delay in optimization, (ii) generation of sparse PFI or LU to speed up subsequent operations in the SM, and (iii) high numerical accuracy.

The overall success of a solver heavily depends on how the PFI or LU is implemented.

Initial Basis

Creation of a good initial basis is a big challenge. Even the notion of “good” is not uniquely defined. Some of the requirements are: the initial basis should (i) be as feasible as possible, (ii) have good objective value, (iii) be triangular. In practice, triangularity always takes priority followed by a combination of (i) and (ii). Obviously, determining an initial basis must be done in “no time,” that is, very quickly.

The all-logical basis corresponding to the ever present identity matrix is triangular and is immediately available. However, in most cases it is not a very good choice. Other triangular bases like the CPLEX basis [13] or CRASH (LTSF) [14] are much better candidates to determine an initial basis.

Optimization often starts from a previously saved basis or an optimal basis of a related problem. Solvers are prepared to accept such bases and even can overcome size compatibility problems.

Pricing

In the general step, the SM chooses an improving nonbasic variable to enter (dual: basic to leave) the basis. If none can be found the current solution is optimal. The rest of this discussion here refers to the primal simplex. Similar steps are valid for the dual.

Selection of an incoming variable is made by checking the optimality conditions for each nonbasic variable. This operation is often referred to as *pricing*. Any variable that violates its optimality condition is an improving candidate. This freedom has given rise to several pricing strategies. Currently none of them is known to be the best one.

From practical point of view, the best pricing strategy is the one that leads to an optimal solution with the least computational effort (shortest solution time). This would be a global strategy. However, in any specific iteration the SM can see only the neighboring bases. Therefore, we cannot expect more than a locally good selection.

It is typical that a large proportion of the total solution time is spent in pricing. Therefore, it is essential to economize this operation.

Regarding the strategies, two extremes can be identified: trying to minimize the number of iterations results in very slow iterations involving heavy computations per iteration and, on the other extreme, methods that have high iteration speed usually require a very large number of iterations until termination. There is no clear theoretical answer which is better but experience shows that usually it is worth making some extra effort to improve the quality of the choice.

In an iteration we can check all nonbasic reduced costs (*full pricing*), or only a subset of them if a sufficient number of improving candidates have already been found (*partial pricing*).

If the computed reduced costs are directly compared then we talk about *direct pricing*. If they are (perhaps approximately) brought to some common platform for comparison then it is termed *normalized pricing*.

The motivation for normalized pricing is the following.

The reduced cost of variable x_j is $d_j = c_j - \pi^T \mathbf{a}_j$, where π is the simplex multiplier. The actual value of d_j depends on the column scale. Namely, if $\hat{\mathbf{a}}_j = s_j \mathbf{a}_j$ and $\hat{c}_j = s_j c_j$, where $s_j > 0$ is the scale factor, then

$$\hat{d}_j = \hat{c}_j - \pi^T \hat{\mathbf{a}}_j = s_j(c_j - \pi^T \mathbf{a}_j) = s_j d_j.$$

Therefore, if the magnitude of the reduced cost is used in comparison the result may be misleading. However, we can obtain a scale-independent measure for d_j if we apply the following *dynamic scaling*:

$$\tilde{d}_j = \frac{d_j}{\|\alpha_j\|}, \quad (2)$$

where $\|\cdot\|$ denotes the Euclidean norm and $\alpha_j = \mathbf{B}^{-1} \mathbf{a}_j$. To see the scale independence of $\tilde{d}_j$, we note that $\hat{\alpha}_j = \mathbf{B}^{-1} \hat{\mathbf{a}}_j = s_j \mathbf{B}^{-1} \mathbf{a}_j = s_j \alpha_j$, and, therefore,

$$\|\hat{\alpha}_j\| = s_j \|\alpha_j\|.$$

So, if the scaled column $\hat{\mathbf{a}}_j$ is used instead of $\mathbf{a}_j$ then both the numerator and denominator are multiplied by s_j in Equation (2). Therefore, $\tilde{d}_j$ remains unchanged. Column selection on the basis of Equation (2) is called *steepest edge pricing* (some details later).

In *hierarchical pricing*, we rank the checked improving candidates according to some further criteria (e.g., expected steplength). This technique requires the setting up of a multiplier vector per criterion (each is equivalent to solving $\mathbf{B}^T \mathbf{v} = \mathbf{h}$ for $\mathbf{v}$) and an extra dot product for each candidate.

It is possible to select some of the most promising candidates during a pricing pass and do iterations with the current basic variables and the selected few. This method is referred to as *multiple pricing*. The selection of candidates is called a *major iteration* (actually, no iteration but selection is done) and the subsequent iterations are *minor iterations*. If the selected subproblem is solved entirely we talk about *complete suboptimization*, if the process terminates when none of the selected candidates can be entered with a positive gain we call it *partial suboptimization*.

Multiple pricing has several attractive features. First of all, from the updated few columns we can select one that makes the largest improvement in the objective function. Second, multiple pricing can be effective in coping with degeneracy. Simply, if any of the selected candidates can result in a positive step, it will be chosen automatically. Third, if the pivot element in the chosen column is not of proper size, the remaining columns can be searched to find a better one, and it comes at no additional cost.

Though multiple pricing does not fit into all pricing strategies to be discussed, it is still an important technique and worth including in an advanced solver.

Below is a brief summary of the best known pricing strategies.

Dantzig Pricing. It is a full pricing selecting a variable with the most violating reduced cost. It is not a particularly good performer for large-scale problems.

Partial Pricing. If there are many improving candidates, a good choice of them can be

found by scanning only a part of the nonbasic variables. This practice is known as *partial pricing*. There are different tactics how to partition the matrix for pricing. We mention just a few. If the partitions are determined once we talk about *static partitioning*. The other possibility is to redefine the partitions as optimization progresses, called *dynamic partitioning*. In either case, pricing starts in the partition next to the one where it stopped during the previous iteration. This goes on in a wrap-around manner after the last partition was scanned. Within partitions, it is possible to do full or partial pricing. The latter is called *nested partial pricing*.

Sectional Pricing. Real-life LP problems often exhibit some structure. In such cases the nonzeros of matrix $\mathbf{A}$ are located in clusters. The projection of the clusters on the horizontal axis forms sections. Pointers can be set up to mark the boundaries of the sections. If partial pricing is done along the sections we talk about *sectional pricing*. This concept has a deeper meaning than just being a partial pricing.

It is typical that vectors in a section are more related to each other than to vectors in other sections. Therefore, if multiple pricing is used we can obtain unrelated (or loosely related) vectors if they are picked from different sections which can lead to more meaningful minor iterations. Namely, in this case, we can expect that the inclusion of one of the candidates in the basis will not ruin the profitability of the remaining ones and they also can enter the basis.

More variations emerge if we do partial pricing within the sections and also do partial pricing with the sections themselves. It means that only a given number of sections are scanned per major iteration and partial pricing is done within the sections.

A general sophisticated pricing scheme is presented in Maros [15], which contains several known pricing strategies as special cases and also can be used to define new ones.

Steepest Edge Pricing. It is the most advanced normalized pricing. Computing dynamic scale factors based on Equation (2) would require prohibitively much

computational work. However, in Forrest and Goldfarb [16], an exact update formula has been proposed that reduces the extra computational work, though it is still substantial. At the same time, experience shows that among the known pricing strategies this one leads to the smallest number of iterations. It does not always translate into the shortest solution time. This very important pricing strategy is a “must have” for every advanced solver. It is a full pricing.

Devex Pricing. Historically, Devex [17] was the first normalized pricing algorithm. Today it is viewed as an approximate steepest edge method. The updated factors it uses gradually worsen. When the inaccuracy exceeds a user-defined level the scale factors are recomputed in a new reference framework. The dual version of Devex practically does not require extra computations. In primal the updated pivot row has to be determined which is computationally more or less equivalent with one more full pricing pass.

Ratio Test

The ratio test is specific to the algorithm (primal or dual) and the phase (phase-1 or phase-2). It is used to determine the basis change. In primal phase-1 and in both phases of dual piecewise linear functions can be defined that help determine the largest progress of the actual objective function with a selected incoming (dual: outgoing) variable. This technique has some additional advantages, like improved numerical stability, efficiency in the presence of degeneracy. For details, see Maros [4].

PROTOTYPE SIMPLEX-BASED SOLVER

In this section, the structure of a simplex-based prototype LP system is presented. An implementation must contain both the primal and the dual methods. Also, smooth transition from primal to dual iterations (and vice versa) should be made possible. Practice shows that the combination of the two main algorithms (one followed by the other to finish off or to make a correction if necessary) can result in the best performance.

Run Parameters, Parameter File

The operation of a simplex-based solver is controlled by several parameters. They can roughly be categorized as *string*, *algorithmic*, and *numerical* parameters.

String parameters are used to identify names (name of input file, output file, etc.). Algorithmic parameters describe the choice of main algorithm (primal, dual), direction of optimization (min, max), presolve strategy, starting basis, pricing, and so on. Numerical parameters define values of zero tolerances, frequencies (such as refactorization and saving of intermediate solutions), maximum number of nonimproving iterations before an antidegeneracy strategy is initiated, maximum number of iterations/solution time, and so on.

Run parameters are specified in a *parameter file*. Usually, every parameter has a default value that users can change by making modifications to the appropriate entry of this file.

Structure

The following functional units must be present in an advanced implementation.

Fast MPS Input. Converts a problem into internal representation. If memory allows it prepares not only column- but also a row-wise storage of $\mathbf{A}$ as some matrix-vector operations can be speeded up considerably with a row-wise $\mathbf{A}$. Provides statistics about the problem (m , $\bar{n}$, v_A , number of variables and constraints by type, density of $\mathbf{A}$, number of nonzeros in the objective row, right-hand side, and range).

Advanced versions of this functional unit are able to analyze the problem and suggest a solution strategy. They can also display the nonzero pattern of $\mathbf{A}$ with zoom facility and produce further statistics about the distribution of magnitudes of the nonzero elements. Visual inspection of the matrix can help select a good pricing strategy.

Preprocessing. Presolve and scaling are usually optional. The level of presolve (the actual tests to be performed) can also be controlled. Different versions of scaling can

be included to provide a choice. The IBM method of Benichou *et al.* [18] is a good candidate if the problem is stable enough to tolerate the small round-off errors generated by this method (the majority of the problems are such). However, for numerically demanding problems the Curtis-Reid scaling is better and should also be made available. Finally, it is important to allow the option of “no scaling,” as the LP relaxation of most combinatorial problems are best solved in this way.

Starting Procedures. The absolute minimum is the lower logical basis. Additionally, at least one of the crash techniques, generally CRASH (LTSF) or a variant of it, must also be available. It makes sense to implement an upper triangular crash on the basis of the same ideas as LTSF, which could be called *UTSF*. In this case, there is no lower triangular part and the LU form update can be very efficient, at least at the beginning of the iterations.

Factorization (LU), or Inversion (PFI). The LU form is highly preferred because of its proven superiority. As the PFI is easier to implement and performs quite well it can be the choice for a first working version of an LP system. Great care must be exercised when designing the data structures to enable the replacement of PFI by LU at a later stage.

Updating Triangular Factors for the LU Version. While the product form update is also a realistic alternative for the LU, the triangular update is more efficient. It leads to a slower growth of nonzeros which, in turn, reduces the need for too frequent refactorizations.

Pricing Strategies. If memory allows solvers usually maintain the vector of reduced costs, $\mathbf{d}$, and update it after every basis change. This applies to primal phase-2. As the composition of the objective function keeps changing in primal phase-1, it is reasonable to recompute the corresponding reduced costs as updating would be too complicated. For dual phase-2 the dual reduced costs (the primal basic variables)

are explicitly available. The dual phase-1 reduced costs are usually recomputed as the structure of the dual phase-1 objective function may change in every iteration.

Regarding the selection of the incoming (in primal) or outgoing (in dual) variables, usually many of the discussed pricing methods are available. Mostly some normalized pricing (Devex but possibly steepest edge) is recommended in both primal and dual. If SIMPRI of Maros [15] is implemented, its parameters can be adjusted to obtain a specific nonnormalized pricing method.

Column and Row Update. The selected column(s) (in primal) or row (in dual) have to be updated. This is a computationally demanding procedure. It is important to be able to exploit sparsity or even hyper-sparsity [19] of the vectors involved. Row- and column-wise storage of $\mathbf{A}$ and a sophisticated double storage of $\mathbf{U}$ of the LU decomposition are the main sources of speed in this part of the SM.

If either the simplex multiplier π or the vector of reduced costs $\mathbf{d}$ is kept updated it also gets transformed here.

Pivot Methods. This is actually the ratio test according to the main algorithm (primal or dual) and the feasibility (phase-1 or -2) for determining the outgoing (in primal) or incoming (in dual) variable. The quoted piecewise linear functions are the best performers as they make the greatest progress and also care for the numerical stability and are efficient in the presence of degeneracy.

Control Unit. Its first task is to interpret the parameter file, if present. It monitors and supervises the entire solution process, invokes functional units, evaluates their outcome, makes decision on the algorithmic techniques, triggers refactorization (reinversion), changes solution strategy if necessary (e.g., checks whether an antidegeneracy strategy has to be applied and makes decision on pricing strategy). It can be quite sophisticated in case of an implementation that has a considerable algorithmic richness.

Solution Output. The solution must be presented in the “unpresolved” form which is

usually obtained after the appropriate *post-solve* procedure (to undo presolve) has been applied. Complete primal and dual solutions can be sent to a file for possible further processing. Usually one line for each variable and constraint is prepared containing some replicated input data (name, variable/row type, lower and upper bound, objective coefficient/RHS value, range) and computed values (primal and dual variables). It is useful to include some additional information by marking basic variables and variables at bound. If the solution is infeasible variables at infeasible level are also marked. If a modeling system was used this file is probably retrieved by it. Otherwise, the file can be perused or submitted to a special purpose program for further analysis, processing, and presentation.

In advanced solvers, upon request, sensitivity analysis of the optimal solution can be performed. It can help make a statement about the robustness of the solution.

FURTHER READINGS

State-of-the-art optimizers are valued commercial products. Therefore, their authors are not keen to disclose details of the solvers. They usually benefit from results of academic research that appear in the open literature.

Recently, academic solvers started to catch up with the commercials in terms of speed and reliability. The commercials are still leading in algorithmic richness. They may have simplex, interior point methods, quadratic programming, integer LP, network optimization, and so on.

In addition to the already quoted source [4] which gives a relatively comprehensive treatment of the SM, and other quoted sources, the journal papers [19–27] give further insight into simplex-based solvers.

REFERENCES

1. Dantzig GB Maximization of a linear function of variables subject to linear inequalities. In: Koopmans TC, editor. Activity analysis of production and allocation. New York: Wiley; 1951. pp. 339–347.

2. Dantzig GB, Orchard-Hays W. Alternate algorithm for the revised simplex method using product form for the inverse. Santa Monica: The RAND Corporation; 1953. RM-1268.
3. Markowitz HM. The Elimination form of the inverse and its application to linear programming. *Manage Sci* 1957;3(3):255–269.
4. Maros I. Computational techniques of the simplex method. Boston: Kluwer Academic Publishers; 2003. p. 325.
5. Curtis AR, Reid JK. On the automatic scaling of matrices for Gaussian elimination. *J Inst Math Appl* 1972;10:118–124.
6. Suhl UH, Suhl LM. Computing sparse LU factorizations for large-scale linear programming bases. *ORSA J Comput* 1990;2(4):325–335.
7. Bartels RH, Golub GH. The simplex method of linear programming using LU decomposition. *Commun ACM* 1969;12(5):266–268.
8. Forrest JJH, Tomlin JA. Updating triangular factors of the basis to maintain sparsity in the product-form simplex method. *Math Program* 1972;2:263–278.
9. Saunders MA. A fast and stable implementation of the simplex method using Bartels-Golub updating. In: Bunch JR, Rose DJ, editors. *Sparse matrix computation*. New York: Academic Press; 1976. pp. 213–226.
10. Goldfarb D. On the Bartels-Golub decomposition for linear programming bases. *Math Program* 1977;13:272–279.
11. Reid JK. A sparsity-exploiting variant of the Bartels-Golub decomposition for linear programming bases. *Math Program* 1982;24(1):55–69.
12. Suhl UH, Suhl LM. A fast LU update for linear programming. *Ann Oper Res* 1993;43:33–47.
13. Bixby RE. Implementing the simplex method: the initial basis. *ORSA J Comput* 1992;4(3):267–284.
14. Maros I, Mitra G. Strategies for creating advanced bases for large-scale linear programming problems. *INFORMS J Comput* 1998;10(2):248–260.
15. Maros I. A general pricing scheme for the simplex method. *Ann Oper Res* 2003;124:193–203.
16. Forrest JJH, Goldfarb D. Steepest edge simplex algorithms for linear programming. *Math Program* 1992;57(3):341–374.
17. Harris PMJ. Pivot selection method of the devex LP code. *Math Program* 1973;5:1–28.
18. Benichou M, Gautier JM, Hentges G, *et al.* The efficient solution of large-scale linear programming problems. *Math Program* 1977;13:280–322.
19. Hall JAJ, McKinnon KIM. Hyper-sparsity in the revised simplex method and how to exploit it. *Comput Optim Appl* 2005;32(3):259–283.
20. Greenberg HJ. Pivot selection tactics. In: Greenberg HJ, editor. *Design and implementation of optimization software*. Alphen aan den Rijn: Sijthoff and Nordhoff; 1978.
21. Maros I. A general phase-I method in linear programming. *Eur J Oper Res* 1986;23(1):64–77.
22. Gill PE, Murray W, Saunders MA, *et al.* A practical anti-cycling procedure for linearly constrained optimization. *Math Program* 1989;45:437–474.
23. Maros I. A piecewise linear dual procedure in mixed integer programming. In: Gianesi F, Komlosi S, Rapcsak T, editors. *New trends in mathematical programming*. Dordrecht: Kluwer Academic Publishers; 1998. pp. 159–170.
24. Maros I. A piecewise linear dual Phase-1 algorithm for the simplex method. *Comput Optim Appl* 2003;26:63–81.
25. Maros I. A generalized dual Phase-2 simplex algorithm. *Eur J Oper Res* 2003;149(1):1–16.
26. Koberstein A, Suhl U. Progress in the dual simplex method for large scale LP problems: practical dual phase 1 algorithms. *Comput Optim Appl* 2007;37(1):49–65.
27. Koberstein A. Progress in the dual simplex method for solving large scale LP problems: techniques for a fast and stable implementation. *Comput Optim Appl* 2008;41(2):185–204.

SIMPLIFYING AND SOLVING DECISION PROBLEMS BY STOCHASTIC DOMINANCE RELATIONS

LOUIS EECKHOUDT
IESEG School of Management,
LEM, Université Catholique
de Lille, Lille, France

CORE, Université Catholique de
Louvain, Louvain-la-Neuve,
Belgium

HARRIS SCHLESINGER
Department of Finance,
University of Alabama,
Tuscaloosa, Alabama

ILIA TSETLIN
INSEAD, Fontainebleau, France
INSEAD, Singapore

INTRODUCTION

The concept of expected utility provides a powerful framework for modeling decision making under risk. To compare two alternatives, we just need to compare their expected utilities. A potential obstacle here is that eliciting utility functions might be difficult. In a consulting context, a consultant might have to suggest a course of action without having the opportunity to elicit a corporate utility function. In a group decision making context, different parties might have different utilities, but if every group member prefers the same alternative, the difference in individual utilities is irrelevant.

Therefore, if two alternatives can be compared in a way that sets some of the particulars of the utility function aside, the situation is simplified, the resulting decision is more reliable, and might be acceptable to a wider group of involved parties. The concept of stochastic dominance offers that type of comparison. It is a formal

statistical property, allowing partial ranking of probability distributions. If one alternative is preferred to another in the sense of the first-order stochastic dominance, then any decision maker with an increasing utility function would have the same preference. If one alternative is preferred to another in the sense of the second-order stochastic dominance, then any decision maker with an increasing and concave utility would have the same preference. Similar analogies hold for higher orders. In the case of multiple choices, even if stochastic dominance criterion does not provide complete ordering of all possible decisions, it might help to eliminate the dominated ones, so that more thorough analysis will focus on the subset of the most promising alternatives. The inferior alternative(s) can be formally identified via the concept of convex stochastic dominance [1,2].

Fishburn and Vickson [3], Whitmore and Findlay [4], and Levy [5] provide comprehensive surveys of stochastic dominance literature. Empirical tests for stochastic dominance efficiency using the notion of convex stochastic dominance are described in, e.g., Refs 6 and 7. Our article focuses on more recent results that highlight the link between stochastic dominance of different orders.

The next section presents several equivalent definitions of stochastic dominance, and explains the connection with the expected utility framework. After stating formal general results, we discuss and illustrate them at the intuitive level. The section titled “Stochastic Dominance as the Preference for Combining Good with Bad” shows that stochastic dominance of different orders can be connected via the preference for combining good with bad, while the next section presents the applications.

STOCHASTIC DOMINANCE AND Nth-DEGREE RISK

We start with a definition of stochastic dominance [8–10]: Assume that all random

variables have bounded supports contained within the interval $[a, b]$. Let F denote the cumulative distribution function (cdf) for such a random variable. Define $F^{(0)}(x) \equiv F(x)$ and define $F^{(i)}(x) \equiv \int_a^x F^{(i-1)}(t) dt$ for $i \geq 1$.

Definition 1. Distribution F weakly dominates Distribution G in the sense of N th-order stochastic dominance if

1. $F^{(N-1)}(x) \leq G^{(N-1)}(x)$ for all $a \leq x \leq b$,
2. $F^{(i)}(b) \leq G^{(i)}(b)$ for $i = 1, \dots, N-2$.

We write F NSD G to denote that F dominates G via N th-order stochastic dominance. If the random variables $\tilde{X}$ and $\tilde{Y}$ have cdfs $\tilde{F}$ and $\tilde{G}$ respectively, we will also say that $\tilde{X}$ NSD $\tilde{Y}$.

Integrals of cdfs, appearing in Definition 1, can be expressed via lower partial moments:

$$\int_a^x (x-t)^k dF(t) = (k!)F^{(k)}(x), \quad k = 1, 2, \dots \quad (1)$$

Equation (1) follows from integration by parts:

$$\begin{aligned} \int_a^x (x-t)^k dF(t) &= (x-t)^k F(t) \Big|_{t=a}^{t=x} \\ &\quad + k \int_a^x (x-t)^{k-1} dF^{(1)}(t), \end{aligned}$$

where the first term is zero since $F^{(k)}(a) = 0$ and $(x-t)^k \Big|_{t=x}^{t=x} = 0$. Repeating the integration by parts $(k-1)$ times yields Equation (1).

The following theorem expresses the well-known links between stochastic dominance and expected utility. [Hadar and Russell [11] and Hanoch and Levy [12], introduced this notion into the economics literature for second-order stochastic dominance. See Refs 9 and 13, as well as [8], for extensions to higher orders of stochastic dominance.] Here $u(x)$ denotes the individual's utility function. For notational convenience, we use $u^{(n)}(x)$ to denote $\frac{d^n u(x)}{dx^n}$.

Theorem 1. The following statements are equivalent:

1. F dominates G in the sense of N th-order stochastic dominance,
2. $\int_a^b u(x) dF(x) \geq \int_a^b u(x) dG(x)$, for all functions u such that $\text{sgn } u^{(n)}(x) = (-1)^{n+1}$ for $n = 1, \dots, N$.

Proof. Integrating by parts, the difference in expected utility for the distributions F and G can be expressed as

$$\begin{aligned} \int_a^b u(x) dF(x) - \int_a^b u(x) dG(x) &= u(x)[F(x) - G(x)] \Big|_{x=a}^{x=b} \\ &\quad + (-1) \int_a^b u^{(1)}(x) d[F^{(1)}(x) - G^{(1)}(x)] \\ &= (-1) \int_a^b u^{(1)}(x) [F^{(0)}(x) - G^{(0)}(x)] dx. \end{aligned}$$

Repeating the integration by parts $(N-1)$ times yields

$$\begin{aligned} \int_a^b u(x) dF(x) - \int_a^b u(x) dG(x) &= \sum_{k=1}^{N-1} (-1)^k u^{(k)}(b) [F^{(k)}(b) - G^{(k)}(b)] \\ &\quad + (-1)^N \int_a^b u^{(N)}(x) [F^{(N-1)}(x) - G^{(N-1)}(x)] dx. \end{aligned}$$

Therefore, the first claim of the theorem implies the second.

Equation (1) is useful to prove that the second claim of the theorem implies the first. Consider $u(t) = -(b-t)^k$, $k = 1, \dots, N-2$. Then

$$\begin{aligned} \int_a^b u(t) dF(t) &\geq \int_a^b u(t) dG(t) \\ \Leftrightarrow \int_a^b (b-t)^k dG(t) &\geq \int_a^b (b-t)^k dF(t) \\ \Leftrightarrow F^{(k)}(b) &\leq G^{(k)}(b). \end{aligned}$$

Similarly, consider

$$u(t) = \begin{cases} -(x-t)^{N-1} & \text{if } t \leq x \\ 0 & \text{if } t > x \end{cases}.$$

Then

$$\begin{aligned} \int_a^b u(t) dF(t) &\geq \int_a^b u(t) dG(t) \\ \Leftrightarrow \int_a^x (x-t)^{N-1} dG(t) &\geq \int_a^x (x-t)^{N-1} dF(t) \\ \Leftrightarrow F^{(N-1)}(x) &\leq G^{(N-1)}(x). \end{aligned}$$

Stochastic dominance of order N implies stochastic dominance of any higher order. To isolate the N th-order effect from the ones of lower orders, Ekmann [14] introduced the concept of N th-degree risk. The special case where $N = 2$ was introduced much earlier by Rothschild and Stiglitz [15], who call this case a “mean-preserving increase in risk.”

Definition 2. Distribution F has more N th-degree risk than Distribution G if

1. $F^{(N-1)}(x) \leq G^{(N-1)}(x)$ for all $a \leq x \leq b$,
2. $F^{(i)}(b) = G^{(i)}(b)$ for $i = 1, \dots, N-1$.

The following corollary to Theorem 1 relates the concept of N th-degree risk to the expected utility framework.

Corollary 1. The following statements are equivalent:

1. G has more N^{th} -degree risk than F ,
2. $\int_a^b u(x) dF(x) \geq \int_a^b u(x) dG(x)$, for all functions u such that $\text{sgn } u^{(N)}(x) = (-1)^{N+1}$.

As follows from Equation (1), G has more N th-degree risk than F if and only if F NSD G and the first $N-1$ moments of F and G are identical.

How could we visualize those stochastic dominance relations at the intuitive level? To construct a pair of random variables ordered by the first-order stochastic dominance, we can take any random variable $\tilde{X}$ and add to it a nonpositive random variable $\tilde{Z}$. Then $\tilde{X} + \tilde{Z}$ is first-order dominated by $\tilde{X}$. For the second-order dominance, $\tilde{Z}$ can be a random variable with nonpositive conditional mean (i.e., $\mathbb{E}(\tilde{Z}|\tilde{X} = x) \leq 0$ for all x). Then $\tilde{X} + \tilde{Z}$ is second-order dominated by $\tilde{X}$.

In the examples above, $\tilde{X}$ can also be a degenerate random variable, that is, $\tilde{X} \equiv c$. Then, for any random variable $\tilde{Y}$, $\tilde{Y} \leq c$, we can say that $\tilde{Y}$ is first-order dominated by $\tilde{X} \equiv c$, and $\tilde{Y}$ is second-order dominated by $\tilde{X} \equiv \mathbb{E}(\tilde{Y})$. However, higher orders of stochastic dominance (or ordering by N th degree risk with $N > 2$) cannot be constructed with one random variable being degenerate. Indeed, suppose that $\tilde{X}$ and $\tilde{Y}$ can be ordered by N th degree risk with $N > 2$, while $\tilde{X} \equiv c$ is degenerate. By Definition 2, the first $N-1$ moments of $\tilde{X}$ and $\tilde{Y}$ are identical. Thus, $\text{Var}(\tilde{Y}) = \text{Var}(\tilde{X}) = 0$, and $\tilde{Y} = c$ almost surely. As an extension, if $\tilde{X}$ and $\tilde{Z}$ are two independent random variables, then a pair $(\tilde{X} + \tilde{Z}, \tilde{X})$ is either ranked by the first-order or second-order stochastic dominance, or cannot be ranked by stochastic dominance of any order.

The above discussion suggests that for stochastic dominance of order three and higher, we cannot have a degenerate probability distribution; both lotteries must be risky. A good illustration is the result of an experiment by Mao [16], which was used by Menezes *et al.* [17] to motivate the concept of aversion to downside risk (i.e., prudence that requires positive third-order derivative of the utility function). Consider the following two lotteries. Lottery A pays 1000 with a probability of $\frac{3}{4}$ and pays 3000 with a probability of $\frac{1}{4}$. Lottery B pays zero with a probability of $\frac{1}{4}$ and pays 2000 with a probability of $\frac{3}{4}$. Note that both lotteries exhibit the same first two moments in their distributions, and lottery A third-order stochastically dominates B. Individuals who prefer lottery A to lottery B exhibit “downside risk aversion.”

The concept of downside risk is related to the following decomposition: Define $\tilde{X}_1 \equiv 2000$, $\tilde{Y}_1 \equiv 1000$, $\tilde{X}_2 \equiv 0$, and $\tilde{Y}_2 \equiv [-1000, +1000]$, a 50–50 lottery. Note that $\tilde{Y}_1$ is a first-order increase in risk over $\tilde{X}_1$, $\tilde{Y}_2$ is a second-order increase in risk over $\tilde{X}_2$, A is the 50–50 lottery $[\tilde{X}_1 + \tilde{Y}_2, \tilde{Y}_1 + \tilde{X}_2]$, and B is the 50–50 lottery $[\tilde{X}_1 + \tilde{X}_2, \tilde{Y}_1 + \tilde{Y}_2]$. Thus, we can think of the following story for lotteries A and B: there are two equally likely states, and we receive 2000 in one state and 1000 in the other. Then, we are required to add a zero-mean risk $[-1000, 1000]$ to one of

the states. Lottery A (B) adds this risk to the state where we receive 2000 (1000). Third-order stochastic dominance (or downside risk aversion) is consistent with the preference to add risk to the higher-wealth state. An alternative view is related to the concept of “preceding changes in risk” [18]. We can start from lottery B, $[\tilde{X}_1 + \tilde{X}_2, \tilde{Y}_1 + \tilde{Y}_2]$, and make second-order improvement (replace $\tilde{Y}_2$ with $\tilde{X}_2$) at lower wealth level ($\tilde{Y}_1 \equiv 1000$), followed by second-order deterioration of the same size (replace $\tilde{X}_2$ with $\tilde{Y}_2$) at a higher-wealth level ($\tilde{X}_1 \equiv 2000$). That would yield lottery A. Here the second-order improvement precedes (in terms of wealth) second-order deterioration, and thus leads to the third-order improvement.

To conclude, the example above is consistent with the preference for combining good with bad, resulting in the lottery $[\tilde{X}_1 + \tilde{Y}_2, \tilde{Y}_1 + \tilde{X}_2]$, as opposed to combining good with good and bad with bad, leading to the lottery $[\tilde{X}_1 + \tilde{X}_2, \tilde{Y}_1 + \tilde{Y}_2]$, where $\tilde{X}_i$ dominates $\tilde{Y}_i$ in the sense of i th-order stochastic dominance, $i = 1, 2$. The next section states and proves how stochastic dominance of different orders (and different orders of risk) can be characterized by a particular preference for combining good with bad.

STOCHASTIC DOMINANCE AS THE PREFERENCE FOR COMBINING GOOD WITH BAD

Let $[A, B]$ denote a lottery that pays either A or B , each with probability one-half. Consider the mutually independent random variables $\tilde{X}_N, \tilde{Y}_N, \tilde{X}_M$, and $\tilde{Y}_M$, and assume that $\tilde{X}_i$ dominates $\tilde{Y}_i$ via i th-order stochastic dominance for $i = M, N$. We wish to compare the 50–50 lotteries $[\tilde{X}_N + \tilde{Y}_M, \tilde{Y}_N + \tilde{X}_M]$ and $[\tilde{X}_N + \tilde{X}_M, \tilde{Y}_N + \tilde{Y}_M]$.

Theorem 2. [19]. *Suppose that $\tilde{X}_i$ dominates $\tilde{Y}_i$ via i th-order stochastic dominance for $i = M, N$. The lottery $[\tilde{X}_N + \tilde{Y}_M, \tilde{Y}_N + \tilde{X}_M]$ dominates the lottery $[\tilde{X}_N + \tilde{X}_M, \tilde{Y}_N + \tilde{Y}_M]$ via $(N + M)$ th-order stochastic dominance.*

Proof. Let T be a positive integer and define $\mathbb{U}_T \equiv \{u \mid \text{sgn } u^{(n)}(w) = (-1)^{n+1} \text{ for}$

$n = 1, \dots, T\}$. For an arbitrary function $u \in \mathbb{U}_{N+M}$, define $v(w) \equiv \mathbb{E}u(\tilde{Y}_M + w) - \mathbb{E}u(\tilde{X}_M + w)$, where $\mathbb{E}$ denotes the expectation operator. We first show that $v \in \mathbb{U}_N$. To see this, consider any integer k , $1 \leq k \leq N$. Observe that $(-1)^k u^{(k)} \in \mathbb{U}_{N+M-k} \subset \mathbb{U}_M$, and therefore $\text{sgn } v^{(k)}(w) = \text{sgn } [\mathbb{E}u^{(k)}(\tilde{Y}_M + w) - \mathbb{E}u^{(k)}(\tilde{X}_M + w)] = (-1)^{k+1}$. The second equality above follows from $(-1)^k u^{(k)} \in \mathbb{U}_M$ and $\tilde{X}_M$ MSD $\tilde{Y}_M$. Thus, $v \in \mathbb{U}_N$.

The condition that $\tilde{X}_N$ dominates $\tilde{Y}_N$ via N th-order stochastic dominance, together with $v \in \mathbb{U}_N$, implies that $\mathbb{E}v(\tilde{X}_N) \geq \mathbb{E}v(\tilde{Y}_N)$, which by the definition of v is equivalent to

$$\begin{aligned} \mathbb{E}u(\tilde{X}_N + \tilde{Y}_M) - \mathbb{E}u(\tilde{X}_N + \tilde{X}_M) \\ \geq \mathbb{E}u(\tilde{Y}_N + \tilde{Y}_M) - \mathbb{E}u(\tilde{Y}_N + \tilde{X}_M). \end{aligned}$$

Rearranging terms above, this inequality is equivalent to

$$\begin{aligned} \frac{1}{2} \{ \mathbb{E}u(\tilde{X}_N + \tilde{Y}_M) + \mathbb{E}u(\tilde{Y}_N + \tilde{X}_M) \} \\ \geq \frac{1}{2} \{ \mathbb{E}u(\tilde{X}_N + \tilde{X}_M) + \mathbb{E}u(\tilde{Y}_N + \tilde{Y}_M) \}, \end{aligned}$$

which is precisely the lottery preference claimed in the theorem.

Theorem 2 expresses a preference for lotteries that combine “relatively good” assets with “relatively bad” ones, where relatively good and bad are defined via stochastic dominance. It is analogous to the notion of “disaggregating the harms” as discussed by Eeckhoudt and Schlesinger [20], if we interpret the “harms” as sequentially replacing each of the $\tilde{X}$ random variables with a $\tilde{Y}$ random variable in the sum $\tilde{X}_N + \tilde{X}_M$.

These preferences lead to a partial ordering of the four alternative sums of random variables, based upon stochastic dominance criteria:

$$\begin{aligned} \tilde{X}_N + \tilde{X}_M > \tilde{X}_i + \tilde{Y}_j > \tilde{Y}_N + \tilde{Y}_M \\ \text{for } (i, j) \in \{(M, N), (N, M)\}. \end{aligned} \quad (2)$$

Note that the sums $\tilde{X}_M + \tilde{Y}_N$ and $\tilde{X}_N + \tilde{Y}_M$ cannot be ordered via stochastic dominance.

In the spirit of Menezes and Wang [21], we can refer to these two sums as the “inner risks” and the sums $\tilde{X}_N + \tilde{X}_M$ and $\tilde{Y}_N + \tilde{Y}_M$ as the “outer risks.” Theorem 2 thus expresses a preference for a 50–50 lottery over the two inner risks as opposed to a 50–50 lottery over the two outer risks. If all lotteries in Equation (2) are degenerate (in which case $M = N = 1$), preference for a lottery over the two inner risks as opposed to a lottery over the two outer risks reduces to the standard notion of risk aversion.

The following Corollary extends Theorem 2 to Ref. 14’s ordering by N th-degree risk.

Corollary 2. *Suppose that $\tilde{Y}_i$ has more i^{th} -degree risk than $\tilde{X}_i$ for $i = M, N$. Then the lottery $[\tilde{X}_N + \tilde{X}_M, \tilde{Y}_N + \tilde{Y}_M]$ has more $(N + M)$ th-degree risk than the lottery $[\tilde{X}_N + \tilde{Y}_M, \tilde{Y}_N + \tilde{X}_M]$.*

Stochastic Dominance and Correlation Aversion

Theorem 2 is also consistent with the notion of correlation aversion [22], as discussed in Ref. 23. Denote by I and I' two binary random variables, with marginal distributions taking either 0 or 1 with probability 1/2. Lottery $[\tilde{X}_N, \tilde{Y}_N]$ can be thought of as the random variable $I\tilde{X}_N + (1 - I)\tilde{Y}_N$, and lottery $[\tilde{X}_M, \tilde{Y}_M]$ can be thought of as the random variable $I'\tilde{X}_M + (1 - I')\tilde{Y}_M$. Summing these two random variables gives

$$I\tilde{X}_N + (1 - I)\tilde{Y}_N + I'\tilde{X}_M + (1 - I')\tilde{Y}_M. \quad (3)$$

Theorem 2 compares the 50–50 lotteries $A = [\tilde{X}_N + \tilde{Y}_M, \tilde{Y}_N + \tilde{X}_M]$ and $B = [\tilde{X}_N + \tilde{X}_M, \tilde{Y}_N + \tilde{Y}_M]$, where A combines good with bad, while B combines good with good and bad with bad. Note that Lottery A corresponds to Equation (3) with $I' = 1 - I$ and Lottery B corresponds to Equation (3) with $I' = I$. In that sense, A (B) corresponds to the perfect negative (positive) correlation between events I and I' , that specify lotteries $[\tilde{X}_N, \tilde{Y}_N]$ and $[\tilde{X}_M, \tilde{Y}_M]$. More generally, denote correlation between I and I' by ρ . Then Equation (3) becomes

$$\begin{aligned} & I\tilde{X}_N + (1 - I)\tilde{Y}_N + I'\tilde{X}_M + (1 - I')\tilde{Y}_M \\ &= \begin{cases} \tilde{X}_N + \tilde{X}_M & \text{with probability } \frac{1}{4}(1 + \rho), \\ \tilde{X}_N + \tilde{Y}_M & \text{with probability } \frac{1}{4}(1 - \rho), \\ \tilde{Y}_N + \tilde{X}_M & \text{with probability } \frac{1}{4}(1 - \rho), \\ \tilde{Y}_N + \tilde{Y}_M & \text{with probability } \frac{1}{4}(1 + \rho); \end{cases} \\ &= \begin{cases} [\tilde{X}_N + \tilde{X}_M, \tilde{Y}_N + \tilde{Y}_M] & \text{with probability } \frac{1}{2}(1 + \rho), \\ [\tilde{X}_N + \tilde{Y}_M, \tilde{Y}_N + \tilde{X}_M] & \text{with probability } \frac{1}{2}(1 - \rho). \end{cases} \end{aligned}$$

As this expression shows, increasing ρ increases the probability of the lottery that combines good with good and bad with bad (i.e., Lottery B), and decreases the probability of the lottery that combines good with bad (Lottery A). Therefore, by Theorem 2, increasing the correlation ρ is a deterioration in the sense of $(N + M)$ th-order stochastic dominance, consistent with the notion of correlation aversion.

The next section discusses how the concept of stochastic dominance, and Theorem 2 in particular, can be applied to decision making under uncertainty.

APPLICATIONS TO DECISION MAKING

As shown in the previous section, preferring more of an attribute to less and preferring to combine good lotteries with bad ones implies preferences consistent with stochastic dominance. Therefore, if the decision is about the alternatives that can be ordered by stochastic dominance, the choice can be made without computing the expected utility, and thus without knowing the particulars of the utility function. First, we discuss several contexts where the decision is about allocating good or bad random variables to a particular state. In such situations, we show how Theorem 2 can be used to determine the optimal allocation. Secondly, we discuss the connection between stochastic dominance (of arbitrary high order) and exponential utility.

Allocating Risks

The main building block in Theorem 2 is a lottery $[\tilde{X}, \tilde{Y}]$, resulting in either random

variable $\tilde{X}$ or random variable $\tilde{Y}$, with equal chances. There are several contexts where such a lottery might occur. One is a two-period consumption/saving model, where labor income in the first period is represented by the random variable $\tilde{X}$, and in the second period is represented by the random variable $\tilde{Y}$, as illustrated in the section titled “Precautionary Effects.” A second is a corporation with two headquarters and taxable profits, as illustrated in the section titled “Corporation with Two Headquarters.” A third is more “direct,” where the decision maker receives either $\tilde{X}$ or $\tilde{Y}$ with equal chances. Such a context is very useful to construct examples of stochastic dominance of higher orders, and to consider the effect of an independent additive background risk, as illustrated in the section titled “Tempering Effects of Risk.”

Precautionary Effects. Consider a simple two-period model of consumption and saving. An individual with time-separable preferences has a random labor income of $\tilde{X}$ at date $t = 0$ and income $\tilde{Y}$ at date $t = 1$. The individual decides to save some of her income at date $t = 0$ and to consume the rest. She must decide on how much to save before learning the realized value of $\tilde{X}$. Thus, her consumption at date $t = 0$ is $\tilde{X} - s$, where s is the amount saved. If $s < 0$, the consumer is borrowing money (i.e., negative savings) and consuming more than the realized value of $\tilde{X}$ at date $t = 0$. We assume that the interest rate for borrowing or lending is zero and that there is no time-discounting for valuing consumption at date $t = 1$. At this date, the individual consumes her income plus any savings, $\tilde{Y} + s$. Let s^* denote the individual’s optimal choice for savings.

Suppose that $\tilde{X}$ dominates $\tilde{Y}$ via N th-order stochastic dominance. For any nonnegative scalar $\varphi \geq 0$, since φ first-order dominates $-\varphi$, it follows from Theorem 2 that the 50–50 lottery $[(\tilde{X} - \varphi), (\tilde{Y} + \varphi)]$ dominates $[(\tilde{X} + \varphi), (\tilde{Y} - \varphi)]$ by $(N + 1)$ th-order stochastic dominance. Reinterpreting the “50–50 lottery” $[A, B]$ as sequential consumption of A at $t = 0$ and B at $t = 1$, Theorem 2 implies that saving an arbitrary amount $\varphi \geq 0$ always dominates saving the amount

$-\varphi$, whenever preferences satisfy $(N + 1)$ th-order stochastic dominance preference. It follows that we must have $s^* \geq 0$ for this individual.

For example, suppose that $\tilde{X} \equiv \mathbb{E}(\tilde{Y})$ and that preferences are given by expected utility. Then this result coincides with the classical case examined by Leland [24] and Sandmo [25], for the case where preferences display prudence, $u''' \geq 0$. This basic result extends in several directions:

- (i) Suppose that one’s labor income is risky in both periods, but it is riskier in the sense of a second-degree increase in risk at date $t = 1$. Anyone with preferences exhibiting an aversion to downside risk (i.e., prudence) would prefer to save money rather than to borrow money. If the second-period wealth is riskier in the sense of second-order stochastic dominance, a precautionary saving demand then requires both aversion to downside risk and aversion to mean-preserving spreads, as defined in Ref. 15.
- (ii) As opposed to (i), suppose that income in the second period is stochastically lower in the sense of first-order stochastic dominance. In that case, aversion to downside risk is no longer required to generate a demand for precautionary saving. We only need someone to be risk averse, that is, averse to risk increases of degree two.
- (iii) We obtain the general case stated above if first-period wealth dominates second-period wealth via N th-order stochastic dominance: A precautionary demand is generated whenever preferences satisfy $(N + 1)$ th-degree stochastic dominance preference.

Finally, suppose that there are two sources of future income risk (e.g., labor income and capital income) which might be correlated. How does an increase in correlation affect savings? In the setting described in section titled “Stochastic Dominance and Correlation Aversion,” increasing correlation leads to $(N + M)$ th-order deterioration. In turn, that will increase savings if preferences are

consistent with $(N + M + 1)$ th-order stochastic dominance.

Corporation with Two Headquarters. We consider here an example applying Theorem 2 with $N = 1$ and $M = 2$, but one that is not an example of a precautionary effect. Consider a risk-neutral corporation with taxable profit $\tilde{X}$ in country A and taxable profit $\tilde{Y}$ in country B . We assume that the tax schedule is identical in both countries and that $\tilde{X}$ second-order dominates $\tilde{Y}$. The tax owed on realized profit π is denoted by $t(\pi)$. The tax schedule is assumed to be increasing with a marginal tax rate that is also increasing, but at a decreasing rate. This assumption is realistic since the marginal rate is often bounded by some maximum, such as 50% of additional profit, and certainly is strictly bounded by 100%. After-tax profits can thus be written as $u(\pi) = \pi - t(\pi)$. If $t(\cdot)$ is differentiable, our assumptions about t imply that $u''(\pi) < 0$ and $u'''(\pi) > 0$. Moreover, since we should also have $t'(\pi) < 1$ for any profit level π , it follows that $u'(\pi) > 0$ as well.

Suppose now that the corporation has a new project with a pre-tax distribution of profit $\tilde{Z}$, where $\tilde{Z} > 0$ almost surely. The corporation must decide whether to locate the project in country A or in country B . The after-tax total profit of the corporation is given by $u(\pi_A) + u(\pi_B)$, where π_i denotes the realized pre-tax profit in country i , for $i = A, B$. Since $\tilde{Z}$ first-order dominates zero, it follows from Theorem 2 that $E[u(\tilde{X}) + u(\tilde{Y} + \tilde{Z})] > E[u(\tilde{X} + \tilde{Z}) + u(\tilde{Y})]$ because the valuation function u (i.e., after-tax profits) satisfies third-order stochastic dominance preference. Thus, the firm should locate the new project in country B in order to maximize its global after-tax profit.

Tempering Effects of Risk. Kimball [26] defines a key behavior consequence of temperance; namely, that an unavoidable risk will “lead an agent to reduce exposure to another (independent) risk.” From Theorem 2, we obtain a variant of Kimball’s tempering effect: an unavoidably higher level of risk in one state will lead one to reduce exposure to a second risk in that state. Here,

higher or lower “risk” is characterized via second-order stochastic dominance.

The notion of tempering effects extends to higher orders of risk. For example, Lajeri-Chaherli [27] studies the effects of background risks on precautionary savings and in doing so, examines the condition of decreasing absolute temperance. A necessary condition for this property within her expected utility framework is $u(5) > 0$, which she labels as “edginess.” By choosing $M = 2$ and $N = 3$, we can interpret edginess as implying that a decrease in one risk (via second-order stochastic dominance) helps to temper the effects of an increase in downside risk of another additive risk.

Interestingly, we can obtain a second equivalence for temperance that has nothing to do with tempering effects, by letting $N = 3$ and $M = 1$. Let $\tilde{Y}_3$ be an increase in downside risk over $\tilde{X}_3$, and let $\tilde{X}_1$ first-order dominate $\tilde{Y}_1$. The (stochastically) higher wealth in $\tilde{X}_1$ helps to mitigate the effects of the increased downside risk in $\tilde{Y}_3$. Thus, we prefer to pair $\tilde{X}_1$ together with $\tilde{Y}_3$ in our lottery preference. In the special case, where $\tilde{X}_1$ and $\tilde{Y}_1$ are both constants, we get a precautionary effect as in the section titled “Precautionary Effects.” But note how this equivalence for “temperance” does not describe any type of tempering effect. Instead, it implies that the property of temperance yields a precautionary effect in protecting against increase in downside risk of future labor income. Similarly, by choosing $M = 1$ and $N = 4$, edginess can be interpreted as implying a precautionary effect against increase in fourth-degree risk for future labor income.

Stochastic Dominance and Exponential Utility

A useful implication is that preference for the dominant lottery in Theorem 2 restricts potential forms of the utility function. Tsetlin and Winkler [28] extend this approach to a multiattribute setting. Consider a decision maker with a nondecreasing utility for wealth, $u'(x) \geq 0$. That means, for example, that \$200 is better than \$100. Consider a lottery [\$100, \$200], and suppose that we combine two such lotteries by adding outcomes.

This can be done by combining good with good and bad with bad, yielding [\$200, \$400] or by combining good with bad, yielding [\$300, \$300] which is \$300 for sure. For the choice between \$300 for sure and an uncertain outcome with a mean of \$300, a preference for combining good with bad is consistent with risk aversion: $u''(x) \leq 0$.

Now, for someone who is risk averse, we can say that \$300 is good and [\$200, \$400] is bad. Combining these two lotteries with an independent [\$100, \$200] lottery gives [[\$300, \$500], \$500] if we combine good with good and bad with bad, and [[\$400, \$600], \$400] if we combine good with bad. Up to a linear transformation, these are the same lotteries as A and B at the end of the first section (there we had [[0, 2000], 2000] and [[1000, 3000], 1000]), and combining good with bad is preferred (i.e., consistent with prudence) if $u'''(x) \geq 0$. As discussed in the section titled “Tempering Effects of Risk,” the next steps would yield temperance ($u(4)(x) \leq 0$), and then edginess ($u(5)(x) \geq 0$). Overall, preference for combining good with bad implies that $u \in \mathbb{U}_\infty$, where $\mathbb{U}_\infty = \{u | (-1)^{k-1} u^{(k)} \geq 0, k = 1, 2, \dots\}$.

Utility functions from $\mathbb{U}_\infty$ have been studied quite extensively [13, 29–32], as they are consistent with stochastic dominance of any order. Whitmore (13, p. 78) mentions that “No compelling economic rationale for this extension is apparent although it can be shown that all functions in the class exhibit nonincreasing absolute risk aversion. Moreover, $\mathbb{U}_\infty$ can be shown to contain many of the families of utility functions which are used routinely by investigators, such as the linear, exponential, logarithmic, and power families. Furthermore, one might argue that a perfectly rational economic decision maker will have preferences which undergo the smooth and systematic change with wealth level which is implied by the conditions defining $\mathbb{U}_\infty$.” The preference for combining good with bad, as opposed to combining good with good and bad with bad (Theorem 2), offers such economic rationale for $\mathbb{U}_\infty$.

Utility functions from $\mathbb{U}_\infty$ are also very tractable, since they can be characterized via Laplace transforms. Consider $u(x)$ defined on $(0, \infty)$. By Bernstein’s lemma (33, p. 439,

Theorem 1a), $u \in \mathbb{U}_\infty$ if and only if its first derivative (i.e., the marginal utility) is a Laplace transform of a (not necessarily finite) measure ψ on $[0, \infty)$: $u'(x) = \int_0^\infty e^{-rx} d\psi(r)$. Then (29, Theorem 1)

$$u \in \mathbb{U}_\infty \text{ if and only if } u(x) = u(x_0) + \int_0^\infty [(1 - e^{-r(x-x_0)})/r] d\psi(r). \quad (4)$$

By Equation (4), any utility function from $\mathbb{U}_\infty$ is a mixture (weighted average) of exponential utilities. This provides a quick way to check whether one alternative dominates another with respect to $\mathbb{U}_\infty$. These results are summarized in the theorem below, which is based on Ref. 31 (Proposition 1 and Proposition 4).

Theorem 3. *The following statements are equivalent:*

1. $\int_a^b u(x) dF(x) \geq \int_a^b u(x) dG(x)$ for all $u \in \mathbb{U}_\infty$,
2. $\int_a^b e^{-rx} dF(x) \leq \int_a^b e^{-rx} dG(x)$ for all $r \in [0, \infty)$,
3. *There exists N such that F weakly dominates G in the sense of N th-order stochastic dominance.*

By Theorem 3, exponential utility [constant absolute risk aversion (CARA)] is a main building block for ranking risky alternatives. (Note that its second condition is equivalent to ranking moment-generating functions for distributions F and G .) Even if the second condition does not hold for all r , but for a wide enough range of risk aversion levels considered to be “reasonable,” choosing F over G would be justified. This is consistent both with Kirkwood’s observation [34] that most decision analysis applications use exponential utility function and with his suggested framework of using exponential utility for sensitivity analysis with respect to risk tolerance levels. This is good news for decision analysts, since exponential utility is, arguably, the easiest to use, in particular because the certainty equivalent of a project does not depend on the presence of independent additive background risk.

SUMMARY

Probability distributions can be partially ranked via stochastic dominance (Definition 1). Within an expected utility framework, preferences are consistent with stochastic dominance of order N if the first N derivatives of the utility function alternate in sign (Theorem 1). If preferences are consistent with stochastic dominance of arbitrary high order, the utility function is a weighted average (mixture) of exponential utilities (Eq. 4).

If preferences are consistent with stochastic dominance, the choice between alternatives might be easier to determine and be somewhat independent of the particulars of the decision maker's utility function, as discussed in the section titled "Stochastic Dominance and Exponential Utility." But why should anybody's preferences satisfy stochastic dominance? Alternatively, how could we check for that?

Theorem 2 shows that the decision maker's preferences are consistent with stochastic dominance of any order if the decision maker prefers more of the attribute to less, and prefers to combine good random variables with bad ones, as opposed to combining good with good and bad with bad. This condition is similar in spirit to risk aversion and correlation aversion in the sense that it reflects an aversion to the combination of bad with bad. Given the partial ordering in Equation (2), a 50–50 gamble between two intermediate outcomes is preferred to a gamble between two extreme outcomes. Preference for combining good with bad seems intuitive in many instances, especially in the wealth or consumption context. In the assessment process, the decision maker should be presented with choices that highlight the issue of preferences for combining good lotteries with bad versus combining good with good and bad with bad.

Theorem 2 also can be used to add intuition to many other extant concepts in the literature, such as skewness preference [18] and transfer principles in income redistribution [35,36]. Furthermore, in some instances, decisions are indeed about how to allocate risks across two states (or two time periods, or two firms, or two groups of people). The

section titled "Allocating Risks" shows how such decisions can often be made independent of a particular utility function or even a particular preference function, as long as we believe that valuations are consistent with stochastic dominance criteria.

Acknowledgments

The authors thank the anonymous referee for very helpful and constructive comments. Ilia Tsetlin's research was supported in part by the Center for Decision Making and Risk Analysis at INSEAD.

REFERENCES

1. Fishburn PC. Convex stochastic dominance with continuous distribution functions. *J Econ Theory* 1974;7:143–158.
2. Fishburn PC. Convex stochastic dominance. In: Whitmore GA, Findlay MC, editors. *Stochastic dominance*. Massachusetts: Lexington Books; 1978. pp. 337–351.
3. Fishburn PC, Vickson RG. Theoretical foundations of stochastic dominance. In: Whitmore GA, Findlay MC, editors. *Stochastic dominance*. Massachusetts: Lexington Books; 1978. pp. 37–113.
4. Whitmore GA, Findlay MC. *Stochastic dominance*. Massachusetts: Lexington Books; 1978.
5. Levy H. Stochastic dominance and expected utility: survey and analysis. *Manage Sci* 1992;38:555–593.
6. Bawa VS, Bodurtha JN, Rao MR, *et al.* On determination of stochastic dominance optimal sets. *J Finance* 1985;40:417–431.
7. Post T. Empirical tests for stochastic dominance efficiency. *J Finance* 2003;58:1905–1931.
8. Ingersoll JE. *Theory of financial decision making*. New Jersey: Rowman & Littlefield; 1987.
9. Jean WH. The geometric mean and stochastic dominance. *J Finance* 1980;35:151–158.
10. Jean WH. The harmonic mean and other necessary conditions for stochastic dominance. *J Finance* 1984;39:527–534.
11. Hadar J, Russell WR. Rules for ordering uncertain prospects. *Am Econ Rev* 1969;59:25–34.
12. Hanoch G, Levy H. Efficiency analysis of choices involving risk. *Rev Econ Stud* 1969;36:335–346.

13. Whitmore GA. Stochastic dominance for the class of completely monotone utility functions. In: Thomas BF, Tae KS, editors. *Studies in the economics of uncertainty in honor of Josef Hadar*. New York: Springer; 1989. pp. 77–88.
14. Ekern S. Increasing N th degree risk. *Econ Lett* 1980;6:329–333.
15. Rothschild M, Stiglitz JE. Increasing risk: I. A definition. *J Econ Theory* 1970;2:225–243.
16. Mao JCT. Survey of capital budgeting: theory and practice. *J Finance* 1970;25:349–360.
17. Menezes CF, Geiss C, Tressler J. Increasing downside risk. *Am Econ Rev* 1980;70:921–932.
18. Chiu H. Skewness preference, risk aversion, and the precedence relations on stochastic changes. *Manage Sci* 2005;51:1816–1828.
19. Eeckhoudt L, Schlesinger H, Tsetlin I. Apportioning of risks via stochastic dominance. *J Econ Theory* 2009;144:994–1003.
20. Eeckhoudt L, Schlesinger H. Putting risk in its proper place. *Am Econ Rev* 2006;96:280–289.
21. Menezes CF, Wang XH. Increasing outer risk. *J Math Econ* 2005;41:875–886.
22. Epstein LG, Tanny SM. Increasing general correlation: a definition and some economic consequences. *Can J Econ* 1980;13:16–34.
23. Denuit M, Eeckhoudt L, Rey B. Some consequences of correlation aversion in decision sciences. *Ann Oper Res* 2009. In press DOI 10.1007/s10479-008-0446-7.
24. Leland HE. Saving and uncertainty: the precautionary demand for saving. *Q J Econ* 1968;82:465–473.
25. Sandmo A. The effect of uncertainty on saving decisions. *Rev Econ Stud* 1970;37:353–360.
26. Kimball MS. Precautionary motives for holding assets. In: Newman P, Milgate M, Falwell J, editors. *The new Palgrave dictionary of money and finance*. London: Macmillan; 1992.
27. Lajeri-Chaherli F. Proper prudence, standard prudence, and precautionary vulnerability. *Econ Lett* 2004;82:29–34.
28. Tsetlin I, Winkler R. Multiattribute utility satisfying a preference for combining good with bad. *Manage Sci* 2009;55(12):1942–1952. In press.
29. Brockett P, Golden L. A class of utility functions containing all the common utility functions. *Manage Sci* 1987;33:955–964.
30. Pratt JW, Zeckhauser R. Proper risk aversion. *Econometrica* 1987;55:143–154.
31. Thistle PD. Negative moments, risk aversion, and stochastic dominance. *J Financ Quant Anal* 1993;28:301–311.
32. Caballe J, Pomansky A. Mixed risk aversion. *J Econ Theory* 1996;71:485–513.
33. Feller W. Volume 2, *An introduction to probability theory and its applications*. 2nd ed. New York: Wiley; 1971.
34. Kirkwood CW. Approximating risk aversion in decision analysis applications. *Decis Anal* 2004;1:51–67.
35. Fishburn PC, Willig RD. Transfer principles in income redistribution. *J Public Econ* 1984;25:323–328.
36. Moyes P. Stochastic dominance and the Lorenz curve. In: Silber J, editor. *Handbook of income inequality measurement*. Boston (MA): Kluwer Academic Publishers; 1999. pp. 199–222.

SIMULATED ANNEALING

SŁAWOMIR T. WIERZCHOŃ
Institute of Computer Science,
Polish Academy of Sciences,
Sopot, Poland

Institute of Informatics, Gdańsk
University, Gdańsk, Poland

INTRODUCTION

A minimization problem is defined as

$$\min_{s \in S} f(s), \quad (1)$$

where S denotes a set of feasible configurations, and $f: S \rightarrow \mathbb{R}$, called *cost* or *objective function*, is a real valued function evaluating each configuration $s \in S$. If S is a discrete set, we are faced with combinatorial (or discrete) optimization; otherwise we refer to it as *continuous optimization*.

Combinatorial optimization, covering more or less practical problems like job shop scheduling, timetabling, graph coloring, traveling salesman problem (TSP), or vehicle routing, poses serious challenges, as most of such problems is NP-hard [1–3]. Similarly, continuous optimization of nonconvex functions of many variables makes real troubles even if the entire function is continuous and twice differentiable. Finding stationary points of f and checking definiteness of the hessian of f in these points is a time-consuming problem when the number of variables increases [4]. That is why we are looking for methods providing simple, cheap, and rather fast means for coping with Equation (1).

Such means are offered by the metaheuristics, that is, iterative processes that “guide and modify the operations of subordinate heuristics to efficiently produce high quality solutions” (5, p. 270). Simulated annealing (SA) [6,7] is an example of a metaheuristic

inspired by the process of physical annealing relying upon slowly cooling of a solid until a minimum energy configuration is attained. SA combines two strategies: random walk and iterative improvement. It can be viewed as a local search algorithm in which there are occasional “upward” moves that lead to the cost increase. In effect, the current locally optimal solution can be directed toward a global minimum.

As noted by Ingber [8], the most attractive features of SA are as follows: (i) it can process quite arbitrary objective functions and constraints; (ii) it is easy to implement; and (iii) it converges in probability to a global minimum, that is, it is possible to obtain a solution arbitrarily close to this minimum, with a probability arbitrarily close to one. Because the algorithm almost did not use any knowledge concerning the optimization problem on which it is run, it is classified—similarly as evolutionary algorithms [9]—as a “black-box” optimization algorithm. But being so universal, it is not free from disadvantages. Ingber [8] lists also its negative features, such as (i) rather low speed of convergence; (ii) difficulty with fine-tuning to specific problems; and (iii) the fact that the statistical convergence does not provide any guarantee of optimality in practical cases.

In this article, we present mathematical and implementational details concerning the SA algorithm. We start our exposition with a concise description of the Metropolis algorithm [10]. It is not only referred to as “one of top 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century” [11] but it is also the main ingredient of the SA algorithm. Such a presentation allows also gentle introduction of basic notions from the field of Markov chains used to analyze properties of the SA. Next, in the section titled “SA Algorithm”, a basic variant of the SA algorithm is presented and its specialization to the problems from the field of combinatorial and continuous optimization are discussed, respectively, in the sections titled “SA for

Combinatorial Optimization Problems” and “SA for Continuous Optimization Problems”. Remarks on the convergence are also given here. The section titled “Other Variants of the SA Algorithm” describes further modifications of the SA algorithm. Mainly we focus here on the problems with noise or imprecise measurements of the cost function and on the problems with vector-valued cost function (i.e., multiobjective optimization). Finally, the section titled “Further Reading” concludes the article pointing to the literature devoted to the topics discussed in the previous sections.

METROPOLIS ALGORITHM

Formally, the heart of the SA is a sampling method invented by Metropolis *et al.* [10] to simulate direct draws from complicated multivariate distributions. In the literature, it is usually referred to as the *Metropolis algorithm*, and sometimes the term “M(RT)² algorithm” (referring to the initials of these authors) is used. Because of the generalization proposed by Hastings [12], we also refer to it as the Metropolis–Hastings, or M–H, algorithm. Here, we shortly review basic concepts needed to design such a sampler. A reader interested in more rigorous and fast presentation is referred, for example, to Ref. 13.

Suppose for simplicity that X is a random variable (rv) taking its values in a finite set S . In this case, S may be identified with the set $\{1, 2, \dots, n\}$. The distribution function (df) π , also called *target distribution*, which characterizes X is a vector $\pi = (\pi_1, \dots, \pi_n)$, where $\pi_i = \mathbb{P}(X = i)$ denotes the probability that X takes the value $i \in S$.

The sampling procedure can be identified with a (possibly infinite) sequence of trials $\{X_k\}$, where the probability of an outcome of a given trial depends only on the outcome of previous trial, that is, $p_{ij}(k) = \mathbb{P}(X_k = j | X_{k-1} = i)$. Such a “memoryless” sequence of trials is said to be a Markov chain [14], and $p_{ij}(k)$ are its transition probabilities. When these probabilities are independent of the trial number k , we say that $\{X_k\}$ is a homogeneous Markov chain (otherwise it is said to

be nonhomogeneous). In this case, the transition probabilities are represented by the so-called stochastic matrix $\mathbf{P} = [p_{ij}]$, that is, a real valued matrix such that

$$\sum_{j=1}^n p_{ij} = 1, \quad i = 1, \dots, n, \quad (2)$$

$$p_{i,j} \in [0, 1], \quad i, j = 1, \dots, n.$$

A nice property of such a representation is that if $a_j(k) = \mathbb{P}(X_k = j)$ stands for the probability of an outcome $j \in S$ at the k th trial, then

$$a_j(k) = \sum_{i=1}^n a_i(k-1) p_{ij}$$

$$= \sum_{i=1}^n a_i(0) p_{ij}^{(k)}, \quad k = 1, 2, \dots, \quad (3)$$

where $p_{ij}^{(k)}$ is (i, j) th entry of the matrix $\mathbf{P}^{(k)} = \mathbf{P}^k$ (i.e., $\mathbf{P}$ to the power k) representing probability of transition from i to j in k steps. Hence, a homogeneous finite Markov chain is completely determined by the pair $(\mathbf{a}(0), \mathbf{P})$, where $\mathbf{a}(0) = (a_1(0), \dots, a_n(0))$ is the initial distribution of states.

Under certain conditions (given below) on the transition probabilities, there exists a unique stationary distribution $\mathbf{s} = (s_1, \dots, s_n)$, whose components are given by

$$s_j = \lim_{k \rightarrow \infty} \mathbb{P}(X_k = j | X_0 = i)$$

$$= \lim_{k \rightarrow \infty} p_{ij}^{(k)}, \quad i = 1, \dots, n. \quad (4)$$

This equation can be rewritten in the form (here $\mathbf{1}$ stands for the column vector of ones)

$$\lim_{k \rightarrow \infty} \mathbf{P}^k = \mathbf{1} \mathbf{s} = \mathbf{S}, \quad (5)$$

that is, each row of the matrix $\mathbf{S}$ equals the row vector $\mathbf{s}$. If so, we can check that for any stochastic row vector $\mathbf{p}$ there holds

$$\lim_{k \rightarrow \infty} \mathbf{p} \mathbf{P}^k = \mathbf{p} \mathbf{S} = \mathbf{s}, \quad (6)$$

that is, if $\mathbf{s}$ is the stationary distribution of a Markov chain then, after some time, the

chain follows this distribution whatever the initial distribution $\mathbf{p}$ is.

Metropolis and his collaborators used this fact to construct a Markov chain in such a way that its stationary distribution equals a target distribution π .

Let us recall [14] that a Markov chain with transition matrix $\mathbf{P}$ is

1. irreducible, if any two states communicate with each other, that is, for any $i, j \in S$ there exists such a number $n(i, j)$ that $p_{ij}^{(n(i, j))} > 0$,
2. aperiodic, if it is not forced into some cycle of fixed length between certain states (in case of a finite chain it suffices that $p_{ii} > 0$ for some $i \in S$).

A Markov chain satisfying these two conditions has a unique stationary distribution obeying the equation $\mathbf{s} = \mathbf{s}\mathbf{P}$, that is, $\mathbf{s}$ is the left eigenvector of the matrix $\mathbf{P}$ associated with the eigenvalue $\lambda = 1$.

In the Metropolis algorithm the transition probabilities p_{ij} are defined as follows:

$$p_{ij} = \begin{cases} 0 & \text{if } j \notin N(i) \text{ and } j \neq i \\ g_{ij} a_{ij} & \text{if } j \in N(i) \text{ and } j \neq i \\ 1 - \sum_{l=1, l \neq i}^n g_{il} a_{il} & \text{if } i = j \end{cases} \quad (7)$$

Here $N(i)$ is the set of states accessible from i and g_{ij} is the so-called generation probability responsible for proposing a state j to which the chain can move in the next step. It defines an important part of the algorithm—the neighborhood structure on S as well as the size of such a neighborhood. The acceptance probability a_{ij} is the probability of accepting such a move. In case of optimization problem, the g_{ij} represents a conditional probability of moving from current solution i to a neighboring solution j ; this new solution is accepted with probability a_{ij} and rejected with probability $1 - a_{ij}$. The only problem is how to choose these probabilities.

To answer it, note that a sufficient condition for a unique stationary distribution $\mathbf{s}$ is that the detailed balance equation,

$$s_i p_{ij} = s_j p_{ji}, \quad (8)$$

holds for all i, j in S . It may be considered as a mathematical counterpart of the thermal equilibrium of a physical system. The Markov chain satisfying this condition is said to be reversible.

As we want drawing samples from the target density π , we should replace $\mathbf{s}$ by this π in the above equation. Now, if the generation matrix $\mathbf{G} = [g_{ij}]$ is irreducible (i.e., there exists such an integer k^* that all the entries of $\mathbf{G}^{k^*}$ are positive) and the next equality holds [14]

$$\frac{a_{ij}}{a_{ji}} = \frac{\pi_j g_{ji}}{\pi_i g_{ij}}, \quad (9)$$

then the Markov chain with transitions p_{ij} defined in Equation (7) is irreducible, aperiodic, and reversible. In general, to guarantee the condition (9), we define acceptance probabilities as (consult also Ref. 15 for alternative formulation)

$$a_{ij} = \varphi\left(\frac{\pi_j g_{ji}}{\pi_i g_{ij}}\right) = \varphi(\xi), \quad (10)$$

where $\varphi: [0, \infty) \rightarrow [0, 1]$ is a nondecreasing function satisfying the condition

$$\frac{\varphi(\xi)}{\varphi(1/\xi)} = \xi. \quad (11)$$

If the generation matrix is symmetric, then Equation (10) reduces to

$$a_{ij} = \varphi\left(\frac{\pi_j}{\pi_i}\right). \quad (12)$$

It is this case considered by Metropolis [10], while Equation (10) is the generalization introduced by Hastings [12]. Two examples of the mapping (Eq. 12) are frequently used

$$\begin{aligned} \varphi_1(\xi) &= \min(1, \xi), \\ \varphi_2(\xi) &= \xi/(1 + \xi). \end{aligned} \quad (13)$$

A physical system in thermal equilibrium at temperature T has its energy probabilistically distributed among all different energy states E . The law governing the distribution of the energies is the so-called Boltzmann

distribution

$$\pi_i = \frac{e^{-E(i)/(k_B T)}}{\sum_{j=1}^n e^{-E(j)/(k_B T)}} = \frac{1}{Z} e^{-E(i)/(k_B T)}, \quad (14)$$

where $E(i)$ stands for the energy of the configuration $i \in S$, k_B is the Boltzmann constant, and T denotes temperature. Choosing units where $k_B = 1$, and using φ_1 from Equation (13) we obtain standard Metropolis acceptance criterion

$$\begin{aligned} a_{ij} &= \min \left\{ 1, \exp \left[-\frac{E(j) - E(i)}{T} \right] \right\} \\ &= \min \left(1, e^{-\Delta E_{ij}/T} \right), \end{aligned} \quad (15)$$

while using φ_2 we obtain the logistic (or Barker) criterion [16]

$$a_{ij} = \frac{1}{1 + e^{\Delta E_{ij}/T}}. \quad (16)$$

Note an important difference between these criteria. The first of them accepts unconditionally any movement to the state with lowered energy, while the second accepts the states with “slightly” lowered or increased energy with the probability close to 1/2. As shown by Pescun [17], the criterion (15) is optimal among other possible choices in the sense that it rejects candidate states less often than other possible criteria. Interestingly, the matrix $\mathbf{A} = [a_{ij}]$ generated in both cases is not stochastic. And indeed it need not be a stochastic matrix—see Chapter 8 in Ref. 3 for numerical examples illustrating the way of forming the transition matrix.

Taking into account Equation (6) we see that from a physical standpoint the above procedure “simulates the thermal motion of atoms in thermal contact with a heat bath at temperature T ” [6] and allows identify ground states, that is, configuration with lowest energy in the limit of low temperature.

SA ALGORITHM

To adopt the Metropolis algorithm to the minimization problem 1, one follows these four steps:

1. Treat the set of states S as the set of admissible solutions.
2. Treat the energy E as the objective function f , and its changes ΔE_{ij} as $\Delta f_{ij} = f(j) - f(i)$.
3. Treat the temperature T as a control parameter, and develop a cooling schedule specifying how it is lowered from high to low values.
4. Define a way in which current solution $i \in S$ can be modified to generate a next (neighboring) solution j . The neighbors of a given solution correspond to the set $N(i)$ of states accessible from a given state i .

The minimization procedure can be summarized as follows:

1. **Initialization:** Set the counter $k \leftarrow 0$, choose initial temperature T_k , and an initial solution $i_k \in S$. Set $f_k \leftarrow f(i_k)$, $i_{\text{best}} \leftarrow i_k$, $f_{\text{best}} \leftarrow f_k$.
2. **while** not done
 - (2a) **repeat** sufficient number of times
 - (2a-1) choose randomly (with probability $g_{i_k j}$) a neighbor $j \in N(i_k)$ and evaluate it, $f_j \leftarrow f(j)$
 - (2a-2) if $r < a_{i_k j}(T_k)$, where r is a random number drawn from the uniform distribution $U(0, 1)$, then $i_k \leftarrow j$ else do not change i_k
 - (2a-3) if $f_j < f_{\text{best}}$ then set $i_{\text{best}} \leftarrow j$, $f_{\text{best}} \leftarrow f_j$
 - (2b) lower the temperature, $T_{k+1} \leftarrow c(T_k)$
 - (2c) $k \leftarrow k + 1$
3. **return** $i_{\text{end}}, f_{\text{end}}$

Note that the “pure” SA algorithm does not keep track of the best solution found so far, that is, it does not use the variables i_{best} and f_{best} ; however, it is a standard practice used when implementing optimization algorithm allowing to focus on substantial improvements only.

To use the SA algorithm the next ingredients must be detailed: a neighborhood

structure for implementing step (2a-1), a transition mechanism for implementing step (2a-2) and a cooling schedule requested in step (2b). Obviously a proper, that is, concise and easy-to-manage representation of the entire problem must be established first.

Also, it would be nice to have recipes allowing to choose initial temperature and initial configuration. The first of these problems is very subtle, and some hints on how to choose T_0 can be found, for example, in Section 1.7 of Ref. 1. The initial configuration i_0 is usually chosen randomly from S . However, a good initial configuration results in faster convergence. Thus it is possible, for instance, to run a simple heuristic (e.g., hill-climbing algorithm) to find a tentative configuration and use it then as i_0 in the algorithm.

To fully specify the algorithm, we also need to know the number of repetitions of the inner and outer loops. The **repeat** loop—obeying steps (2a-1)–(2a-3) implement the Metropolis algorithm. The number of repetitions of this loop at fixed temperature depends heavily on the cooling schedule realized in step (2b). This schedule also affects a stopping criterion. Formally, the algorithm stops when the temperature T_k is close to zero, although more practical criteria are often used.

We discuss these topics in the following sections. Although SA is a general-purpose algorithm, some of its implementational details depend on a problem representation. That is why we distinguish between the variants designed for combinatorial and continuous optimization.

A reader interested in further details concerning computer implementation is referred, for example, to Section 10.9 in Ref. 18. Its first part describes a program for solving the TSP with possibly additional constraints, and in the second part of this section continuous minimization is considered.

SA FOR COMBINATORIAL OPTIMIZATION PROBLEMS

Specialization of the SA Algorithm

Neighborhood Structure. Formally, the neighborhood structure is defined by a func-

tion $N: S \rightarrow 2^S$ that assigns to every admissible solution $i \in S$ a set of neighbors $N(i) \subseteq S$, that is, a set of “slight” modifications of the entire solution i [2].

To be illustrative, consider a symmetric TSP problem with n cities. Typically, the cities are labeled by subsequent integers $1, 2, \dots, n$, and j th city has coordinates (x_j, y_j) . A configuration i is a permutation $i_{(1)} i_{(2)} \dots i_{(n)}$ of the set $\{1, \dots, n\}$ interpreted as the order in which the cities are visited. Thus assuming that the salesperson starts from a fixed city, the space S consists of $(n-1)!$ permutations representing possible tours. Note that such a representation although simple has one disadvantage: each tour is represented by n different permutations. For instance, if $n = 8$ then, for example, $\{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\{8, 1, 2, 3, 4, 5, 6, 7\}$ represent the same tour. A neighbor of a tour i can be obtained by using three different operations explained in Fig. 1. Two of them were invented by Lin [19], while the third is an operator used in evolutionary computations.

An important property which the neighborhood structure must possess is that for any two states $i_s, i_f \in S$ there exists a finite sequence of $K \geq 2$ states $j_1, \dots, j_K$ such that $j_k \in N(j_{k-1}), k = 2, \dots, K$, and $i_s = j_1, i_f = j_K$. Such a sequence is referred to as the *path*, and the neighborhood structure with which any two states are joined by a path is said to be well defined. Surely, the neighborhoods induced by the operators from Fig. 1 are well defined. Detailed discussion on how to define effective neighborhood structures can be found, for example, in Refs 2 and 3.

Tuning Parameters of the Algorithm. With the neighborhood structure, the generation probability is defined in standard SA as

$$g_{ij} = \begin{cases} \frac{1}{|N(i)|} & \text{if } j \in N(i) \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

where $|N(i)|$ is the cardinality of the set $N(i)$, and the acceptance probability is defined as in Equation (15). In other words, all the neighbors are of equal importance.

The temperature T_k is lowered from some initial value T_0 until the final value

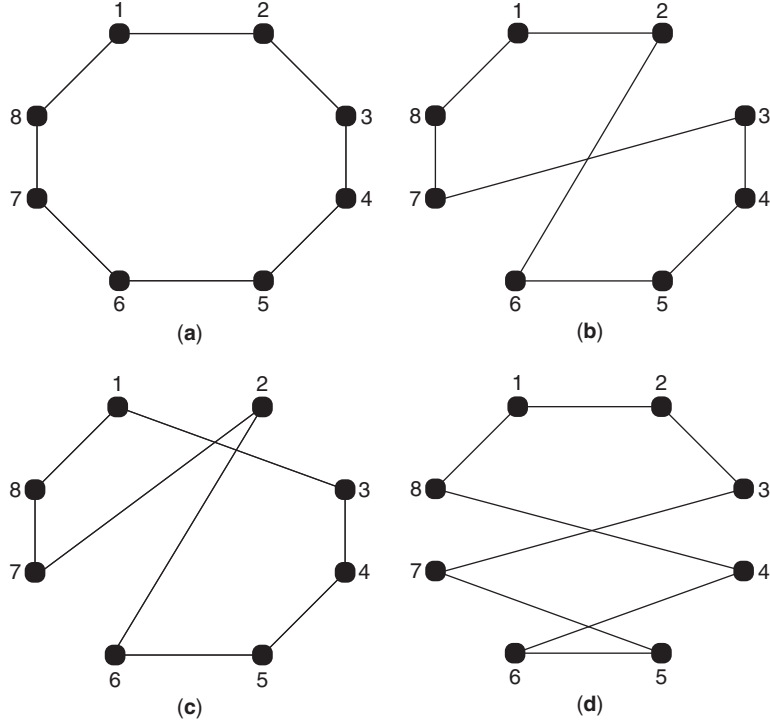


Figure 1. Operators for generating neighbors of a tour in a TSP: (a) original tour $i = \{1, 2, 3, 4, 5, 6, 7, 8\}$, (b) 2-change, or inversion: reverses the order in randomly chosen subtour $3, 4, 5, 6$, that is, $i' = \{1, 2, 6, 5, 4, 3, 7, 8\}$, (c) translation: places randomly chosen subtour, for example, $3, 4, 5, 6$ after randomly chosen city, for example, city no. 1: $i'' = \{1, 3, 4, 5, 6, 2, 7, 8\}$, (d) switching: interchanges the cities from two randomly chosen positions in the tour, here fourth and seventh positions, that is, $i''' = \{1, 2, 3, 7, 5, 6, 4, 8\}$.

T_f is reached. The problem of choosing proper initial and final values as well as the problem of how fast this T_k should be lowered requires deeper reflection, and, frequently, trial-and-error experiments. Detailed discussion of these topics can be found, for example in Refs 20 and 21. Some suggestions on how to choose the value of T_0 are given in Section 1.7 in Ref. 1.

The most frequently used cooling schedules are geometric,

$$\begin{aligned} c_1(T_k) &= \alpha T_k = T_0 \alpha^k, \quad \alpha \in [0.8, 0.99], \\ k &= 1, 2, \dots, \end{aligned} \quad (18)$$

or hyperbolic schedule,

$$c_2(T_k) = \frac{T_k}{1 + \beta T_k}, \quad \beta \approx 0, k = 1, 2, \dots \quad (19)$$

Although the geometric schedule does not, in general, guarantee convergence to the minimum, it is popular because it allows convergence in reasonable time. Other cooling schedules are discussed and reviewed, for example, in Refs 1, 22, and 23.

It is suggested—see Ref. 1, Section 1.7 of Ref. 18—that each new value of T_k in Equation (18) should be constant for $100n$ reconfigurations, or for $12n$ successful reconfigurations, whichever comes first. Here n stands for the number of parameters (degrees of freedom) of the problem to be solved.

The **while**-loop (2) in the SA algorithm stops after few (say three to five) successive temperature stages without any acceptance.

Convergence of the Algorithm

Under certain conditions, the SA algorithm converges in probability to the set of optimal

solutions $S^* \subseteq S$, that is,

$$\lim_{k \rightarrow \infty} \mathbb{P}(X_k \in S^*) = 1, \quad (20)$$

where X_k denotes the random variable being the outcome of k th iteration of the algorithm. To prove this fact, we can view the SA algorithm as a sequence of infinitely long homogeneous Markov chains, or as an inhomogeneous Markov chain with transition probabilities depending on current temperature T_k (and thus on iteration number k). The second approach is much more complicated, and hence we briefly sketch the first one only. Detailed discussion of this subject can be found, for example, in Chapter 8 of Ref. 3.

In the first (rather unrealistic) case, these “certain conditions” are such that the Markov chain controlled by the transition probability matrix $\mathbf{P}(T_k)$ is irreducible and aperiodic for a fixed value of T_k . For instance, if the neighbor structure N is well defined, the generation probabilities g_{ij} are given by Equation (17) and the acceptance probabilities a_{ij} by Equation (15), then the matrix $\mathbf{P}$ with entries (7) is regular, and such that $p_{jj} > 0$ for at least one $j \in S$ (see, e.g., Theorem 8.4 in Ref. 3). Thus, there exists the unique stationary distribution $\pi(T_k)$ with the components (cf. Theorem 8.5 in Ref. 3)

$$\pi_i(T_k) = \frac{e^{-f(i)/T_k}}{\sum_{j \in S} e^{-f(j)/T_k}} = \frac{e^{-f(i)/T_k}}{Z(T_k)}. \quad (21)$$

Denoting $f^* = \min_{i \in S} f(i)$, these components may be rewritten in the form

$$\begin{aligned} \pi_i(T_k) &= \frac{e^{-f(i)/T_k}}{Z(T_k)} \cdot \frac{e^{f^*/T_k}}{e^{f^*/T_k}} \\ &= \frac{\exp[-(f(i) - f^*)/T_k]}{\sum_{j \in S} \exp[-(f(j) - f^*)/T_k]}. \end{aligned} \quad (22)$$

Defining $\delta = \min_{j \in S \setminus S^*} f(j) - f^*$ and observing that the denominator of this last expression is greater than one, we can write (cf. Lemma 8.2 in Ref. 3)

$$\pi_i(T_k) \leq \exp\left(-\frac{\delta}{T_k}\right), \quad (23)$$

which implies that

$$\lim_{T_k \rightarrow 0} \pi_i(T_k) = \begin{cases} 0 & \text{if } i \in S \setminus S^* \\ \frac{1}{|S^*|} & \text{if } i \in S^* \end{cases} \quad (24)$$

According to Lemma 8.4 in Ref. 3 the Equation (24) holds also for nonsymmetric generation probabilities g_{ij} .

Thus, the SA algorithm converges to a global minimum if for each value of $T_k, k = 0, 1, \dots$, the corresponding Markov chain is infinitely long and the sequence of the temperatures T_k converges to zero as $k \rightarrow \infty$.

Geman and Geman [24] is perhaps the most popular reference establishing conditions on the cooling schedule for formal convergence of a variant of SA algorithm designed for image restoration problem (consult Section 8.6 in Ref. 25 for a deeper discussion). More general result, formulated under nonhomogeneous Markov chain assumption, states that (consult [26]) the SA algorithm converges if and only if $\lim_{k \rightarrow \infty} T_k = 0$ and

$$\sum_{k=1}^{\infty} \exp(-h^*/T_k) = \infty, \quad (25)$$

where h^* is the smallest number such that every state $i \in S$ communicates with S^* at height h^* . Recall from Ref. 26 that a state $i \in S$ communicates with S^* at height h if there exists a path $j_1, \dots, j_K$ in S such that $j_1 = i$ and $j_K \in S^*, K > 1$, and such that $f(j_k) \leq f(j_1) + h$ for all $k = 2, \dots, K$. This result shows that certain annealing schedules guarantee convergence to S^* . Particularly, we can use a logarithmic schedule

$$c_3(T_k) = \frac{d}{\log(1+k)}, \quad (26)$$

where d is a positive constant such that $d \geq h^*$. The only problem relies upon estimation of the value h^* . Gidas [27] gives a similar condition together with a discussion on how to choose d .

Note also that for continuous state spaces there are results in the form of stochastic differential equations—the reader is referred to section titled “SA with noisy cost function” for further details.

Some Extensions

An intriguing generalization of the Boltzmann distribution (14) is obtained by considering generalized entropy (consult Ref. 28 for a quick introduction)

$$\text{Ent}_q = k \frac{1 - \sum_{i \in S} p_i^q}{q - 1}, \quad q \in \mathbb{R}. \quad (27)$$

The principle of maximum entropy constraining the energy to average value T leads to the distribution

$$\begin{aligned} p_i &= \frac{[1 - \beta(1 - q)E(i)]^{1/(1-q)}}{\sum_{j \in S} [1 - \beta(1 - q)E(j)]^{1/(1-q)}} \\ &= \frac{[1 - \beta(1 - q)E(i)]^{1/(1-q)}}{Z_q}, \end{aligned} \quad (28)$$

where $\beta = 1/(k_B T)$ is the “inverse temperature”. Particularly, in the $q \rightarrow 1$ limit we recover the Boltzmann distribution (14). The distribution (28) generalizes large families of standard distributions like Lévy α -stable distributions L_α [29], and Student’s t - and r -distributions [30]. The paper [29] provides a simple and efficient procedure for generating deviates from such a distribution. It is a generalization of Box–Müller method for generating Gaussian deviates.

Penna [31] was the first who proposed using such a distribution in solving the TSP problem. Here the Metropolis criterion (15) generalizes to

$$a_{ij}(q; T) = \min \left\{ 1, [1 - (1 - q)\Delta E_{ij}/T]^{1/(1-q)} \right\}, \quad (29)$$

and obviously it approaches Equation (15) in the $q \rightarrow 1$ limit. Numerical experiments reported in Ref. 31 show that for $q < 0$, the number of annealing steps is considerably smaller than in the classical SA algorithm.

Further modification of Equation (29) was proposed by Franz and Hoffmann [32].

In conclusion of this section, let us mention an important problem of adapting the size of the neighborhood. As noted in Ref. 33 “limitation of SA with a fixed neighborhood size is its

inability to perform search at different scales in different stages of search”. This problem is naturally solved in the case of continuous optimization (described in the next section), but in combinatorial optimization it is rarely considered, although varying the neighborhood size improves search abilities (see, e.g., the VNS algorithm described in Refs 2 and 5). The papers [33,34] are attempts to solve this problem efficiently.

SA FOR CONTINUOUS OPTIMIZATION PROBLEMS

At present, we assume that $S \subset \mathbb{R}^n$, that is, we search for such a vector $\mathbf{x} = (x_1, \dots, x_n)$, which minimizes a real valued function f of n -variables.

The simple distribution (17) is not sufficient in this case. Chib and Grindberg [13] mention five methods of drawing candidates to implement the Metropolis algorithm and they conclude that precise answer to the question how these choices should be made is far from complete. The most popular method relies upon generating the candidate $\mathbf{y}$, given current solution $\mathbf{x}$, according to the equation

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \quad (30)$$

where $\mathbf{z}$ is called the *increment rv*, or simply, a disturbance, and follows a multivariate density $p(\mathbf{z})$. Now the generation density is written $g(\mathbf{x}, \mathbf{y})$ and $g(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} - \mathbf{x})$.

Below, we briefly review the most popular variants of the SA algorithm devoted to unconstrained optimization problems. Note, however, that using random walk (30) we can go beyond the set S , hence some “repair” strategies must be eventually designed to prevent such a situation.

An extensive review of other variants of SA can be found, for example, in Chapters 1 and 5 of Ref. 1 and in Ref. 35.

Classical Simulated Annealing

In the field of evolution strategies, we use the term *mutation probability* to name the pdf p . Symmetric multivariate Gaussian pdf is a most natural candidate for generating random vectors $\mathbf{z}$. It perfectly fits to the

assumptions of the Metropolis algorithm and to the three postulates formulated by Beyer and Schwefel in Ref. 9. These postulates are as follows:

1. *Reachability*: Any state $\mathbf{x} \in S$ should be reachable from any other state $\mathbf{y} \in S$ in finite time.
2. *Scalability*: The length of search step (mutation strength) should be tunable in order to adapt to the properties of the fitness landscape.
3. *Absence of Biases*: The distribution $p(\mathbf{z})$ should be chosen according to the maximum entropy principle. If $S = \mathbb{R}^n$ this requirement leads naturally to the normal distribution, and if $S = \mathbb{Z}^n$ —to the geometrical distribution.

Assuming that the standard deviation σ , referred to as the *step size*, is proportional to the temperature T_k at iteration k , that is, $\sigma = \sqrt{T_k}$, the density function used to generate isotropic disturbances takes the form

$$p_k(\mathbf{z}) = \frac{1}{(\pi T_k)^{n/2}} \exp\left(-\frac{\|\mathbf{z}\|^2}{2T_k}\right). \quad (31)$$

Such a choice allows vigorous search of S with large steps in high temperatures, and very careful improvements of the solution realized with small steps in lower temperatures. In other words, the size of the neighborhood decreases with time steps. Note also that it is easy to generate variate $\mathbf{z}$: all its components have the form $\sqrt{T_k} \cdot N(0, 1)$, where $N(0, 1)$ stands for the standard Gaussian random number.

Geman and Geman [24] proved that the SA with the Metropolis acceptance criterion and Gaussian disturbances converges to the global minimum if and only if the temperature T_k in k th iteration is inversely proportional to the logarithmic function of iterations number, given a sufficiently high initial temperature T_0 , that is,

$$T_k = \frac{T_0}{\log(1+k)}. \quad (32)$$

A simple heuristic justification of this fact can be found in Refs 8 and 36, and a rigorous

treatment of this topic is provided by Gidas [27]. Other studies on the convergence of such an algorithm are reviewed in Section 4 of Ref. 35.

Fast Simulated Annealing

The random walk employed to generate candidate solutions is an important factor influencing the rate of convergence of SA. If it allows reasonable exploration, supported by a smart exploitation of “most interesting” distant regions of the search space, then the chance of fast identification of the global optimum grows radically.

This observation inspired Szu and Hartley [36] to replace multinormal Gaussian disturbance occurring in Equation (30) by isotropic multivariate Cauchy disturbance with the pdf

$$p_k(\mathbf{z}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\pi^{\frac{n+1}{2}}} \frac{T_k}{(\|\mathbf{z}\|^2 + T_k^2)^{\frac{n+1}{2}}}. \quad (33)$$

Under such a setting, and the choice of the logistic acceptance criterion (16), the cooling schedule takes the form

$$T_k = \frac{T_0}{1+k}, \quad (34)$$

that is, it is inversely linear in the number of iterations.

Important feature of the multivariate Cauchy distribution $C(0, T_k)$ is that it realizes occasionally long jumps among essentially local sampling over the search space. This is in contrast with local sampling controlled by multinormal distribution; here the expected distance between the current point $\mathbf{x}_k$ and its neighbor $\mathbf{y}_{k+1}$ equals $\sigma_k \sqrt{n}$.

Cauchy distribution is a special case of α -stable Lévy distribution L_α (mentioned earlier), where $\alpha \in (0, 2]$ is a parameter. Particularly, $\alpha = 2$ corresponds to Gaussian distribution and $\alpha = 1$ to Cauchy distribution. The tails of such a distribution vanish as $x^{-(1+\alpha)}$ what results in the variance divergence when $\alpha \neq 2$ and—finally—in much larger jumps called, also, *Lévy flights*. Further information on this family can be found in Ref. 37 (an interested reader can visit Nolan’s web page <http://academic2.american.edu/~jhnolan/>).

edu/~jpnolan/ to find links to many papers on this subject as well as some software). Other good source of recipes showing how to generate random vectors with such distributions is Ref. 38.

Very Fast Simulated Annealing

Further quickening of the cooling process can be gained by tuning search intensity along each dimension individually. As the step size should be proportional to the temperature T_k , the requirement of individual tuning implies that the temperature varies along each dimension individually. More precisely, i th component z_i of the vector $\mathbf{z}$ is generated according to the density [8]

$$p_k(z_i) = [\kappa (|z_i| + T_{k,i}) \log(1 + 1/T_{k,i})]^{-1}, \quad (35)$$

where $T_{k,i}$ is the temperature in dimension i at time k , and κ is a normalizing constant. The joint pdf $p_k(\mathbf{z})$ is the product of all the densities $p_k(z_i)$, $i = 1, \dots, n$. Contrary to the Gaussian and Cauchy distributions already considered, this pdf is anisotropic. In this case, we obtain exponential cooling schedule

$$T_{k,i} = T_{0,i} \exp(-b_i k^{1/n}), \quad (36)$$

where k, b_i , and $T_{0,i}$ are appropriate constants. Further details of this algorithm—now called *adaptive simulated annealing* (ASA)—can be found at Ingber’s web page <http://www.ingber.com/#ASA>. Apart from scientific papers, the source code in C language can be downloaded from this page.

The method leading to the ASA algorithm was further developed by Yao [39] who proposed to replace the pdf (35) by

$$p_k(z_i) = [\kappa (|z_i| + 1/\log(1/T_{k,i})) \log(1 + \log(1/T_{k,i}))]^{-1}. \quad (37)$$

This time the cooling schedule takes the form

$$T_{k,i} = T_{0,i} \exp\left[-\exp(b_i k^{1/n})\right]. \quad (38)$$

Although these results sound very promising, it should be noted that the heuristic proof used in Ref. 36 is based on the assumption of generating and not visiting any state infinitely often in time—consult [39] for a deeper discussion of this topic.

Generalized Simulated Annealing

Tsallis distribution, called also *q-Gaussian distribution* in Ref. 29, and discussed in the section titled “Some Extensions” can be used for solving continuous optimization problems also. Tsallis and Stariolo [40] are the first who proposed such an algorithm. Its convergence is studied by Nishimoro and Inoue in Ref. 41.

OTHER VARIANTS OF THE SA ALGORITHM

Up to this moment, it was assumed that the cost of any configuration i (if S is discrete) or $\mathbf{x}$ (if $S \subset \mathbb{R}^n$) can be univocally measured and expressed as a single scalar. In this section we relax this assumption. Allowing imprecise measurements we obtain so-called stochastic annealing, and while allowing vector valuation of f we are faced with multiobjective optimization.

SA with Noisy Cost Function

Assume that the evaluation of the cost function is imprecise or is disturbed by random noise. The second case is easier to process and can be treated as an approximation of the former one. In general, there are two ways to cope with such a problem. The first adapts SA to the requirement of imprecise measurements, while the second combines it with the algorithms of stochastic optimization. Brief review of these approaches can be found in Ref. 42 and Section 8.2.2 in Ref. 25, while the second approach, elaborated, among others, by Gelfand and Mitter [43], is described in Section 8.4 and 8.6 in Ref. 25 and in Section 3 in Ref. 35.

Let $\tilde{f}(\mathbf{x})$ denote the evaluation of the true and unknown value $f(\mathbf{x})$ disturbed by a Gaussian noise, that is, $\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Obviously, the observed difference $\Delta \tilde{f}_{\mathbf{xy}} = \tilde{f}(\mathbf{x}) - \tilde{f}(\mathbf{y})$ is also Gaussian rv $N(\delta_{\mathbf{xy}}, \tilde{\sigma}^2)$ where $\delta_{\mathbf{xy}} = f(\mathbf{x}) - f(\mathbf{y})$ is the

true difference, and $\tilde{\sigma}^2 = 2\sigma^2$. The true value of $\delta_{\mathbf{xy}}$ can be estimated by averaging its multiple observations. Denoting j th such an observation as $\Delta \tilde{f}_{\mathbf{xy}}^{(j)}$, we estimate $\delta_{\mathbf{xy}}$ as

$$\hat{\delta}_{\mathbf{xy}}^m = \frac{1}{m} \sum_{j=1}^m \Delta \tilde{f}_{\mathbf{xy}}^{(j)}. \quad (39)$$

Now the only problem is to estimate the acceptance probability $a(\hat{\delta}_{\mathbf{xy}}^m)$. It was observed in Ref. 44 that the cumulative normal distribution with appropriately chosen variance looks very similar to the logistic acceptance criterion (16). This allows to derive a correlation between the number of samples m and the temperature T :

$$T \approx \sqrt{\frac{8m}{\pi \tilde{\sigma}}}.$$

Assuming further the deterministic rule to accept the candidate solution if $\hat{\delta}_{\mathbf{xy}}^m \leq 0$, we obtain so-called stochastic simulation algorithm [44]. A number of drawbacks of this approach was noted in Ref. 42, where a more efficient variant with effective sampling mechanism was proposed.

The second approach can be summarized as a discretized version of the Langevin equation [35]

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \tilde{G}_k(\mathbf{x}_k) + b_k \mathbf{w}_k, \quad (40)$$

where $\tilde{G}_k(\mathbf{x}_k)$ represents a direct (noisy) gradient $\partial f / \partial \mathbf{x}$ if f is a function of class C^2 or its approximation estimated from noisy measurements. $\{a_k\}$ is a sequence of positive step sizes, $\{b_k\}$ is a sequence of “temperatures” such that $\lim_{k \rightarrow \infty} b_k = 0$, and $\{\mathbf{w}_k\}$ is a sequence of independent, identically distributed Gaussian vectors $N(\mathbf{0}, \mathbf{I}_n)$. Adding random disturbance $b_k \mathbf{w}_k$ simulates the decision step of SA algorithm, where an upward move is sometimes accepted in the hope of leaving a local minimum attained by the algorithm.

Generally the algorithm slowly converges to the global minimum, that is, as $1 / \log \log k$. Further details about this approach can be found in the original paper [43] or in Section 8.4 of Ref. 25, while Section 8.6 clarifies the convergence theory of such an algorithm.

SA for Multiobjective Optimization

In multiobjective optimization, each configuration $\mathbf{x}$ is evaluated using few, say $v > 1$, different criteria, that is, $f(\mathbf{x})$ is now a vector consisting of v values $[f_1(\mathbf{x}), \dots, f_v(\mathbf{x})]$.

A simplest way of comparing different configurations is to associate a weight $w_j \geq 0, j = 1, \dots, v$ with each criterion and define a single objective function $f'(\mathbf{x}) = \sum_{j=1}^v w_j f_j(\mathbf{x})$. Such a procedure is rather trivial and frequently controversial: there is no unique recipe how to choose the weights in advance.

A more serious approach makes use of the notion of nondominance, or min-dominance in our case. Namely, we say that a configuration $\mathbf{x}$ min-dominates other configuration $\mathbf{y}$ (or, that $\mathbf{y}$ is min-dominated by $\mathbf{x}$) only if $f_j(\mathbf{x}) \leq f_j(\mathbf{y}), j = 1, \dots, v$, and $f_{j^*}(\mathbf{x}) < f_{j^*}(\mathbf{y})$ for at least one $1 \leq j^* \leq v$.

Now the real challenge in extending SA to handle multiple objectives lies in determining how to compute the probability of accepting an individual $\mathbf{y}$ min-dominated by current configuration $\mathbf{x}$. A simple solution is to consider the difference in the number of criteria that they min-dominate configuration $\mathbf{y}$. Particularly, in Ref. 45 so-called amount of domination was proposed. Apart from the number of poorly satisfied criteria, this measure takes into account relative differences between the evaluations of these criteria.

A number of specific approaches is reviewed in Section 10.2 in Refs 45–47.

Parallelization of the SA Algorithm

As the computing time is a critical factor in many disciplines, parallel versions of the algorithm were proposed. Two main approaches to the parallelization are proposed: (i) using P elementary processors P Markov chains are executed and (ii) only several states of the same Markov chain are executed on a single processor. These approaches are briefly characterized in Section 1.4 in Ref. 1, and the book by Azencott [48] is a good source of information on this subject.

Parallel SA can exploit functional parallelism: the use of multiple processors to evaluate different phases of a single move,

implementing several Markov chain computations in parallel [48], carrying out parallel computations for several states of the same Markov chain [1]. The data parallelism in SA is employed by the use of different processors or group of processors that propose and evaluate moves by multiple independent runs [11].

FURTHER READING

There is a vast literature devoted to the theory and application of SA. Concluding the article, we propose some guide to the literature.

The book [14] presents rigorously important facts from the field of Markov chains while the paper [13] offers readable introduction to the M–H algorithm.

Chapter 8 in Ref. 25 gives a short and detailed description of main directions developed within the annealing algorithms. Many interesting and practical information on SA and its applications can be found in Chapter 1 in Ref. 1, while new variants of this algorithm (like simulated diffusion or microcanonic annealing) are described in Chapter 5. The book also describes other metaheuristic and fast introduction to this field can be found in the paper [5]. Formal framework for local search algorithms and easy accessible considerations on the asymptotic convergence of SA is provided by the book [3]. Another variant of SA termed *quantum annealing*, that is, an optimization method inspired by a process of quantum fluctuations, and its relation to SA is discussed in the book [49].

Section 4 of Ref. 8 contain a review of various applications of SA in quite complex problems, while in Section 1 of this paper, advantages and disadvantages of SA are given. In papers [50,51], the performance of SA on three problems: graph partitioning, graph coloring, and numbers partitioning, is discussed. The general conclusion is such that the performance of SA is mixed; in some problems it outperforms the best-known tailored heuristics, and in other cases specialized heuristics performs better. In our opinion, this is not a pessimistic conclusion. It rather confirms the “no free lunch”

theorem presented in Ref. 52. According to this theorem, if algorithm *A* outperforms algorithm *B* on some cost functions, then broadly speaking there must exist exactly as many other functions where *B* outperforms *A*. The study [53] examines the performance of SA on another combinatorial problem: maximum matching problem.

Lastly, let us note that using stable distributions when generating neighbors of a given configurations opens new possibilities for designing efficient metaheuristics. Exhaustive bibliography concerning theory and application of Tsallis statistics (including SA) can be downloaded from <http://tsallis.cat.cbpf.br/biblio.htm>.

Acknowledgments

The author appreciates the comments of three anonymous reviewers and the comments of Topical Editor on an earlier version of this article.

REFERENCES

1. Dréo J, Pétrowski A, Siarry P, *et al.* Metaheuristics for hard optimization. Methods and case studies. Berlin: Springer; 2006. p. 369.
2. Hoos HH, Stutzle T. Stochastic local search. San Francisco (CA): Morgan Kaufmann; 2004. p. 672.
3. Michiels W, Aarts E, Korst J. Theoretical aspects of local search. Berlin: Springer; 2007. p. 238.
4. Schwefel HP. Evolution and optimum seeking. New York: John Wiley & Sons, Inc.; 1995. p. 456.
5. Blum C, Roli A. Metaheuristics in combinatorial optimization: overview and conceptual comparison. ACM Comput Surv 2003;35(3): 268–308.
6. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 13;220:671–680. DOI: 10.1126/science.220.4598.671.
7. Černý V. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. J Optim Theory Appl 1985;45(1):41–51. DOI: 10.1007/BF00940812.
8. Ingber L. Simulated annealing: practice versus theory. Math Comput Model 1993;18(11): 29–57.

9. Beyer HG, Schwefel HP. Evolution strategies – a comprehensive introduction. *Nat Comput* 2002;1(1):3–52. DOI: 10.1023/A:10150599-28466.
10. Metropolis N, Rosenbluth AW, Rosenbluth MN, *et al.* Equation of state calculations by fast computing machines. *J Chem Phys* 1953; 21(6):1087–1092.
11. Dongarra J, Sullivan F. Introduction to the top 10 algorithms. *Comput Sci Eng* 2000;2(1): 22–23. DOI: 10.1109/MCISE.2000.814652.
12. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;57(1):97–100. DOI: 10.1093/biomet/57.1.97.
13. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *Am Stat* 1995;49(4):327–335. DOI: 10.1234/12345678.
14. Iosifescu M. Finite Markov processes and their applications. 2nd ed. New York: Dover Publications; 2007. p. 304.
15. Kesidis G, Wong E. Optimal acceptance probability for simulated annealing. *Stoch Int J Probab Stoch Proc* 1990;29(2):221–226. DOI: 10.1080/17442509008833615.
16. Barker AA. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Aust J Phys* 1965;18(2): 119–133.
17. Pescun PH. Optimum Monte-Carlo sampling using Markov chains. *Biometrika* 1973;60(3): 607–612. DOI: 10.1093/biomet/60.3.607.
18. Press WH, Flannery BP, Teukolsky SA, *et al.* Numerical recipes in C. The art of scientific computing. 2nd ed. Cambridge: Cambridge University Press; 1992. p. 1020. DOI: 10.2277/0521431085.
19. Lin S. Computer solutions of the traveling salesman problem. *Bell Syst Tech J* 1965;44: 2245–2269.
20. Lundy M, Mees A. Convergence of an annealing algorithm. *Math Program* 1986;34(1): 111–124. DOI: 10.1007/BF01582166.
21. Reeves CR, editor. Modern heuristic techniques for combinatorial problems. Oxford: Blackwell Scientific Publishing; 1993. p. 320.
22. Nourami Y, Andresen B. A comparison of simulated annealing cooling strategies. *J Phys A Math Gen* 1998;31(41):8373–88385. DOI: 10.1088/0305-4470/31/41/011.
23. Stander J, Silverman BW. Temperature schedules for simulated annealing. *Stat Comput* 1994;4(1):21–32. DOI: 10.1007/BF00-143921.
24. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 1984;6(6):721–741. DOI: 10.1080/02664769300000058.
25. Spall JC. Introduction to stochastic search optimization: estimation, simulation and control. Hoboken (NJ): Wiley; 2003. p. 618.
26. Hajek B. Cooling schedules for optimal annealing. *Math Oper Res* 1988;13(2): 311–329. DOI: 10.1287/moor.13.2.311.
27. Gidas B. Nonstationary Markov chains and convergence of the annealing algorithm. *J Stat Phys* 1985;39(1/2):73–131. DOI: 10.1007/BF01007975.
28. Tsallis C. Levy distributions. *World Phys* 1997;10(7):42–46.
29. Thistleton W, Marsh JA, Nelson K, *et al.* Generalized Box-Muller method for generating q-gaussian random deviates. *IEEE Trans Inf Theory* 2007;53(12):4805–4810.
30. Souza AMC, Tsallis C. Student's *t*- and *r*-distributions: unified derivation from an entropic variational principle. *Physica A* 1997;236(1):52–57.
31. Penna TJP. Traveling salesman problem and Tsallis statistics. *Phys Rev E* 1995;51(1):R1-R3. DOI: 10.1103/PhysRevE.51.R1.
32. Franz A, Hoffmann KH. Threshold accepting as a limit case for a modified Tsallis statistics. *Appl Math Lett* 2003;16(1):27–31.
33. Yao X. Comparison of different neighbourhood sizes in simulated annealing. *Proceedings of the 4th Australian Conference on Neural Networks*. Sydney, Australia; 1993. pp. 216–219.
34. Cheh KM, Goldberg JB, Askin RG. A note on the effect of neighborhood structure in simulated annealing algorithm. *Comput Oper Res* 1991;18(6):537–547. DOI: 10.1016/0305-0548(91)90059-Z.
35. Locatelli M. Simulated annealing algorithms for continuous global optimization. In: Pardalos PM, Romeijn HE, editors. Volume 2, Handbook of global optimization. Boston (MA), Dordrecht, London: Kluwer Academic Publishers; 2002. pp. 179–230.
36. Szu H, Hartley R. Fast simulated annealing. *Phys Lett A* 1987;122(3, 4):157–162.
37. Nolan JP. Stable distributions—models for heavy tailed data. Boston (MA): Birkhauser; 2010.
38. Devroye L. Non-uniform random variate generation. New York: Springer; 1986. p. 843.

39. Yao X. A new simulated annealing algorithm. *Int J Comput Math* 1995;56(3):161–168. DOI: 10.1080/00207169508804397.
40. Tsallis C, Stariolo DA. Generalized simulated annealing. *Physica A* 1996;233(1):395–406. DOI: 10.1016/S0378-4371(96)00271-3.
41. Nishimori H, Inoue J. Convergence of simulated annealing using the generalized transition probability. *J Phys A Math Gen* 1998; 31(26):5661–5672. DOI: 10.1088/0305-4470/31/26/007.
42. Branke J, Meisel S, Schmidt C. Simulated annealing in the presence of noise. *J Heuristics* 2008;14(6):627–654. DOI: 10.1007/s10732-007-9058-7.
43. Gelfand S, Mitter S. Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$. *SIAM J Control Optim* 1993;31(1):111–131.
44. Ball RC, Fink TMA, Bowler NE. Stochastic annealing. *Phys Rev Lett* 91(3):030201. DOI: 10.1103/PhysRevLett.91.030201.
45. Bandyopadhyay S, Saha S, Maulik U, *et al.* A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Trans Evol Comput* 2008;12(3):269–283.
46. Coello Coello CA, Lamont GB, Veldhuizen DA. *Evolutionary algorithms for solving multi-objective problems*. 2nd ed. New York: Springer; 2007. p. 800.
47. Suman B, Kumar P. A survey of simulated annealing as a tool for single and multiobjective optimization. *J Oper Res Soc* 2006;57(10): 1143–1160.
48. Azencott R, editor. *Simulated annealing: parallelization techniques*. New York: Wiley; 1992. p. 256.
49. Das A, Chakrabarti BK, editors. *Quantum annealing and related optimization methods*. Berlin: Springer; 2005. p. 377. DOI: 10.1007/11526216.
50. Johnson DS, Aragon CR, McGeoch LA, *et al.* Optimization by simulated annealing: an experimental evaluation; Part I, Graph partitioning. *Oper Res* 1989;37(6):865–892. DOI: 10.1287/opre.37.6.865.
51. Johnson DS, Aragon CR, McGeoch LA, *et al.* Optimization by simulated annealing: an experimental evaluation; Part II, Graph coloring and number partitioning. *Oper Res* 1991;39(3):378–406. DOI: 10.1287/opre.39.3.378. 37(6):865–892.
52. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997;1(1):67–682.
53. Sasaki G, Hajek B. The time complexity of maximum matching by simulated annealing. *J Assoc Comput Mach* 1988;35(2):387–403. DOI: acm.org/10.1145/42282.46160.

SIMULATION OPTIMIZATION IN RISK MANAGEMENT

MARCO BETTER
FRED GLOVER
OptTek Systems, Inc., Boulder,
Colorado

INTRODUCTION

Without uncertainty there is no risk. However, very few, if any, real-world situations are risk free. In the world of deterministic optimization, we often choose to “ignore” uncertainty in a given situation in order to come up with a unique and objective solution to a problem. But in situations where uncertainty is at the core of the problem, as it is in those that involve risk management, a different strategy is required.

The field of optimization offers various approaches to cope with uncertainty. In this context, the exact values that the parameters (e.g., the data) of the optimization problem will have are not known with absolute certainty, but may vary to a greater or lesser extent depending on the nature of the factors they represent. In other words, there may be many possible “realizations” of the parameters, each of which is called a *scenario*.

The following sections titled “Scenario Optimization” and “Robust Optimization” briefly describe traditional scenario-based approaches to optimization, such as *scenario optimization* and *robust optimization*. The section titled “Simulation Optimization” then delves in greater detail into a recently emerging approach called *simulation optimization* that is having a profound impact on risk management.

SCENARIO OPTIMIZATION

Dembo [1] offers an approach to solving stochastic programs based on a method for

solving deterministic scenario subproblems (SPs) and combining the optimal scenario solutions into a single feasible decision.

Imagine a situation in which we want to minimize the cost of producing a set J of finished goods. Each good j ($j = 1, \dots, n$) has a per-unit production cost c_j associated with it, as well as an associated utilization rate a_{ij} of resources for each finished good. In addition, the plant that produces the goods has a limited amount of each resource i ($i = 1, \dots, m$), denoted as b_i . We can formulate a deterministic mathematical program for a single scenario s (SP) as follows:

SP:

$$z^s = \min \sum_{j=1}^n c_j^s x_j \quad (1)$$

Subject to:

$$\sum_{j=1}^n a_{ij}^s x_j = b_i^s \quad \text{for } i = 1, \dots, m \quad (2)$$

$$x_j \geq 0 \quad \text{for } j = 1, \dots, n, \quad (3)$$

where c^s , a^s , and b^s , respectively, represent the realization of the cost coefficient, the resource utilization, and the resource availability data under scenario s . Consider, for example, a company that manufactures a certain type of Maple door. Depending on the weather in the region where the wood for the doors is obtained, the costs of raw materials and transportation will vary. The company is also considering whether to expand production capacity at the facility where doors are manufactured, so that a total of six scenarios must be considered. The six possible scenarios and associated parameters for Maple doors are shown in Table 1.

Therefore, model SP needs to be solved once for each of the six scenarios.

The scenario optimization approach can be summarized in two steps:

Table 1. Possible Scenarios for Maple Doors

Scenario	Capacity	$P(C)$ (%)	Weather	$P(W)$ (%)	P (Scenario)	Cost, c_j	Utilization, a_{ij}	Availability, b_j
1	Current	50	Dry	33	1/6	Low	High	Low
2			Normal	33	1/6	Medium	Low	Low
3			Wet	33	1/6	High	Low	Low
4	Expanded	50	Dry	33	1/6	Low	High	High
5			Normal	33	1/6	Medium	Low	High
6			Wet	33	1/6	High	Low	High

1. Compute the optimal solution to each deterministic SP.
2. Solve a tracking model to find a single, feasible decision for all scenarios.

The key aspect of scenario optimization is the function of the tracking model in step 2. For illustration we introduce a simple form of tracking model. Let p^s denote the estimated probability for the occurrence of scenario s . Then, a simple tracking model for our problem can be formulated as follows:

$$\begin{aligned} \text{Minimize } & \sum_s p^s \left(\sum_j c_j^s x_j - z^s \right)^2 \\ & + \sum_s p^s \left(\sum_{ij} a_{ij}^s x_j - b_i^s \right)^2 \end{aligned} \quad (4)$$

$$\text{Subject to : } x_j \geq 0 \quad \text{for } j = 1, \dots, n. \quad (5)$$

The purpose of this tracking model is to find a solution that is feasible under all the scenarios, and penalize solutions that differ greatly from the optimal solution under each scenario. The two terms in the objective function are squared to ensure nonnegativity. More sophisticated tracking models can be used for various different purposes. In risk management, for instance, we may select a tracking model that is designed to penalize performance below a certain target level.

ROBUST OPTIMIZATION

Robust optimization may be used when the parameters of the optimization problem are known only within a finite set of values. The robust optimization framework gets its

name because it seeks to identify a robust decision—that is, a solution that performs well across many possible scenarios.

In order to measure the robustness of a given solution, different criteria may be used. Kouvelis and Yu identify three criteria: (i) absolute robustness, (ii) robust deviation, and (iii) relative robustness. We illustrate the meaning and relevance of these criteria, by describing the robust optimization approach described in Ref. 2.

Consider an optimization problem where the objective is to minimize a certain performance measure such as cost. Let S denote the set of possible data scenarios over the planning horizon of interest. Also, let X denote the set of decision variables and P the set of input parameters of our decision model. Correspondingly, we let P^s identify the value of the parameters belonging to scenario s and let F^s identify the set of feasible solutions to scenario s . The optimal solution to a specific scenario s is then

$$z^s = f(X_s^*, P^s) = \min_{X \in F^s} f(X, P^s). \quad (6)$$

We assume here that f is convex. The first criterion, absolute robustness, also known as *worst-case optimization*, seeks to find a solution that is feasible for all possible scenarios and optimal for the worst possible scenario. In other words, in a situation where the goal is to minimize the cost, the optimization procedure will seek the robust solution, z^R , that minimizes the cost of the maximum-cost scenario. We can formulate this as an objective function of the form

$$z^R = \min \left\{ \max_{s \in S} f(X, P^s) \right\}. \quad (7)$$

The second criterion, robust deviation, is the one that minimizes the largest deviation from optimality. In other words, we seek to minimize the worst-case deviation that separates the robust solution from the optimal solution to each scenario. The third criterion, relative robustness, is similar to the second, but instead of a direct deviation measure, we use a relative deviation measure with respect to optimality. Suppose, for convenience, we reformulate the SP introduced in the previous section as follows (based on Ref. 2, pp. 26–29):

$$z^s = \min c^s x \quad (8)$$

$$\text{Subject to : } A^s x = b^s \quad (9)$$

$$x \geq 0. \quad (10)$$

Then, the robust deviation problem would be formulated as

$$\text{Minimize } y \quad (11)$$

$$\text{Subject to : } c^s x \leq y + z^s \quad (12)$$

$$A^s x = b^s \quad (13)$$

$$x \geq 0. \quad (14)$$

Variations to this basic framework have been proposed in Ref. 3 to capture the risk-averse nature of decision makers, by introducing higher moments of the distribution of z^s in the optimization model, and implementing weights as penalty factors for infeasibility of the robust solution with respect to certain scenarios.

SIMULATION OPTIMIZATION

Until quite recently, the methods available for finding optimal decisions have been unable to effectively handle the complexities and uncertainties posed by many real-world situations of the form treated by simulation. The areas of scenario optimization and robust optimization described in the previous sections attempt to deal with some of these, but the modeling framework limits the range of situations that can be tackled with such technology.

The complexities and uncertainties that render real-world systems mathematically intractable are the primary reason that simulation is often chosen as a basis for handling decision problems associated with those systems. On the one hand, the set of possible scenarios is often unknown *a priori*. On the other hand, the possible combinations of parameters is too numerous to be handled efficiently by methods such as those previously described. Consequently, decision makers must deal with the dilemma that many important types of optimization problems can be treated only by the use of simulation models, but once these problems are submitted to simulation there are no optimization methods that can adequately cope with them.

Recent developments are changing this picture. The field of metaheuristics—the domain of optimization that augments traditional mathematics with artificial intelligence and methods derived from physical, biological, or evolutionary analogs—has witnessed advances yielding effective engines for guiding a series of complex evaluations in the quest for optimal values for the decision variables, as described in Refs 4–9.

Modern simulation optimization tools are designed to solve optimization problems of the form

$$\text{Minimize } F(x) \quad (\text{objective function})$$

$$\text{Subject to : } Ax \leq b \quad (\text{constraints})$$

$$g_l \leq G(x) \leq g_u \quad (\text{requirements})$$

$$l \leq x \leq u \quad (\text{bounds}),$$

where the vector x of decision variables includes variables that range over continuous values and variables that only take on discrete values (both integer values and values with arbitrary step sizes).

The objective function $F(x)$ is typically highly complex; under the context of simulation optimization $F(x)$ represents a performance measure output by the

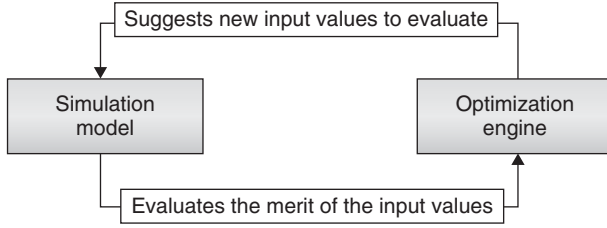


Figure 1. Coordination between optimizer and simulator.

simulation, and may be any mapping from a vector x of real values to a real scalar value. The constraints represented by the inequality $Ax \leq b$ are linear (given that non-linearity in the model is embedded within the simulation itself), and the coefficient matrix A and the right-hand-side values corresponding to vector b must be known. The requirements $g_l \leq G(x) \leq g_u$ impose simple upper and/or lower bounds on a function that can be linear or nonlinear, and is an output of the simulation. The values of the bounds g_l and g_u must be known constants. All the variables are bounded and some may be restricted to be discrete, as previously noted. Each evaluation of $F(x)$ and $G(x)$ requires a simulation of the system. By combining simulation and optimization, a powerful design tool results.

Optimizers designed for simulation embody the principle of separating the method from the model. In such a context, the optimization problem is defined outside the complex system. Therefore, the evaluator (i.e., the simulation model) can change and evolve to incorporate additional elements of the complex system, while the optimization routines remain the same. Hence, there is a complete separation between the model that represents the system and the procedure that is used to solve optimization problems defined within this model.

The optimization procedure uses the outputs from the system evaluator, which measures the merit of the inputs that were fed into the model. On the basis of both current and past evaluations, the method decides upon a new set of input values (Fig. 1).

Provided that a feasible solution exists, the optimization procedure ideally carries out a special search where the successively

generated inputs produce varying evaluations, not all of them improving, but which over time provide a highly efficient trajectory to the globally best solutions. The process continues until an appropriate termination criterion is satisfied (usually based on the user's preference for the amount of time devoted to the search).

The simulation optimization framework is very flexible in terms of the performance measures the decision maker wishes to evaluate. In fact, the only limitation is not on the side of the optimization engine, but on the simulation model's ability to evaluate performance based on specified values for the decision variables. In order to provide in-depth insights into the use of simulation optimization in the context of risk management, we present a practical application through the use of an illustrative example in the context of optimal portfolio selection.

Example: Selecting Optimal Portfolios of Projects. The energy industry uses project portfolio optimization to manage investments in exploration and production, as well as power plant acquisitions [10,11]. Decision makers typically wish to maximize the return on invested capital, while controlling the exposure of their portfolio of projects to various risk factors.

The following example involves a company that has 61 potential projects in its investment funnel. Each type of project requires an initial investment and a certain number of business development, engineering, and earth sciences personnel. The company has a budget limit for its investment opportunities, and a limited number of personnel of each skill category. Projects may start in different time periods, but there is a restricted window of opportunity of up to three years for

each project. The company must select a set of projects to invest in that will best further its corporate goals.

Probably the best-known model for portfolio optimization is rooted in the work of Nobel laureate Harry Markowitz. The model, called the *mean-variance model* [12], is based on the assumption that the expected portfolio returns will be normally distributed. The model seeks to balance risk and return in a single objective function, as follows.

Given a vector of portfolio returns r and a covariance matrix Q of returns, we can formulate the model as follows:

$$\text{Maximize} \quad r^T w - k w^T Q w \quad (15)$$

$$\text{Subject to :} \quad \sum_i c_i w_i = b \quad (16)$$

$$w \in \{0, 1\}, \quad (17)$$

where k represents a coefficient of the firm's risk aversion, c_i represents the initial investment in project i , w_i is a binary variable representing the decision whether to invest in project i , and b is the available budget. We will use the mean-variance model as a base case for the purpose of comparing with other selected models of portfolio performance.

To facilitate our analysis, we make use of the OptFolio® software that combines simulation and optimization into a single system specifically designed for portfolio optimization [13–15].

We examine three cases to demonstrate the flexibility of this method to enable a variety of decision alternatives that significantly improve upon traditional mean-variance portfolio optimization, and illustrate the flexibility afforded by simulation optimization approaches in terms of controlling risk. The results also show the benefits of managing and efficiently allocating scarce resources like capital, personnel, and time. The weighted average cost of capital, or

annual discount rate, used for all cases was 12%.

Case 1: Mean-Variance Approach. In this first case, we implement the mean-variance portfolio selection method of Markowitz described above. The decision is to determine participation levels (0 or 1) in each project with the objective of maximizing the expected net present value (NPV) of the portfolio while keeping the standard deviation of the NPV below a specified threshold of \$140 million. We denote the expected value of the NPV as μ_{NPV} and the standard deviation of the NPV as σ_{NPV} . This case can be formulated as follows:

$$\text{Maximize} \quad \mu_{\text{NPV}} \quad (\text{objective function}) \quad (18)$$

$$\text{Subject to:} \quad \sigma_{\text{NPV}} < \$140\text{M} \quad (\text{requirement}) \quad (19)$$

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (20)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j \quad (\text{personnel constraints}) \quad (21)$$

$$\text{All projects must start in year 1} \quad (22)$$

$$x_i \in \{0, 1\} \forall i \quad (\text{binary decisions}). \quad (23)$$

The optimal portfolio has the following performance metrics:

$$\mu_{\text{NPV}} = \$394\text{M},$$

$$\sigma_{\text{NPV}} = \$107\text{M},$$

$$P(5)_{\text{NPV}} = \$176\text{M},$$

where $P(5)_{\text{NPV}}$ denotes the fifth percentile of the resulting NPV probability distribution (i.e., the probability of the NPV being lower than the $P(5)$ value is 5%).

The bound imposed on the standard deviation in Equation (19) does not seem binding. However, due to the binary nature of the decision variables, no project additions are possible without violating the bound. Figure 2 shows a graph of the probability distribution of the NPV obtained for 1000 replications of this base model. The thin line represents the expected value.

Case 2: Risk Controlled by Fifth Percentile. In the context of risk management, statistics such as variance or standard deviation of returns are not always easy to interpret and other measures may be more intuitive and useful. For example, it is clearer to say “there is a 95% chance that the portfolio return is above some value X ” than to say that the standard deviation is \$107 million. This measure can easily be implemented in a simulation optimization procedure by imposing a requirement on the fifth percentile of the resulting distribution of returns. In this case, the decision is to determine participation levels (0, 1) in each project with the objective of maximizing the expected NPV of the portfolio, while keeping the fifth percentile of the NPV distribution above the value determined in Case 1. This is achieved by imposing the requirement in Equation (25) on the model below. In other words, we want to find the portfolio that produces the maximum average return, as long as no more than 5% of the observations fall below \$176 million. Also, in this case we allow for delays in the start dates of projects, according to windows of opportunity defined for each project. In order to achieve this, we have created copies of a project that are shifted by one, two, or three periods into the future (according to the windows of opportunity defined for each project). Mutual exclusivity clauses ensure that only one start date for each project is selected. For example, to represent the fact that Project A can start at time $t = 0, 1$, or 2,

we include the following mutual exclusivity clause as a constraint:

$$\text{Project } A_0 + \text{Project } A_1 + \text{Project } A_2 \leq 1.$$

The subscript following the project name corresponds to the allowed start dates for the project, and the constraint allows at most only one of these to be chosen.

$$\text{Maximize} \quad \mu_{\text{NPV}} \quad (24)$$

Subject to :

$$P(5)_{\text{NPV}} > \$176\text{M} \quad (\text{requirement}) \quad (25)$$

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (26)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j \quad (\text{personnel constraints}) \quad (27)$$

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i \quad (\text{mutual exclusivity}) \quad (28)$$

$$x_i \in \{0, 1\} \quad \forall i \quad (\text{binary decisions}), \quad (29)$$

where m_i denotes the set of mutually exclusive projects related to project i .

In this case, we have replaced the standard deviation with the fifth percentile as a

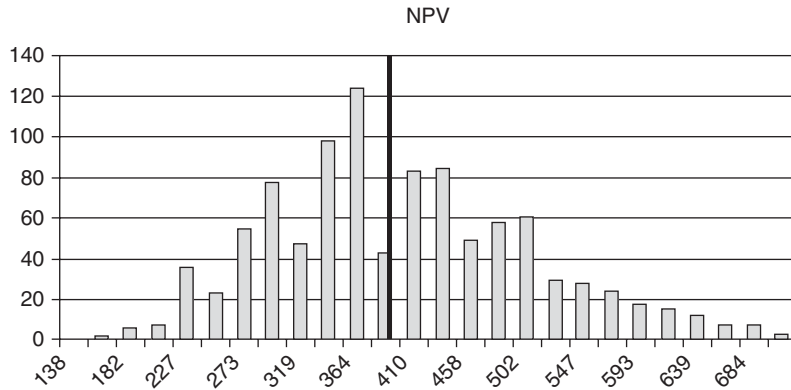


Figure 2. Mean-variance portfolio.

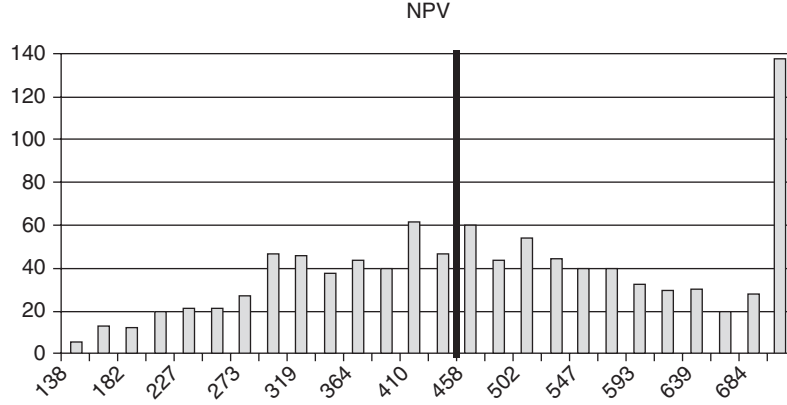


Figure 3. Portfolio risk controlled by $P(5)$.

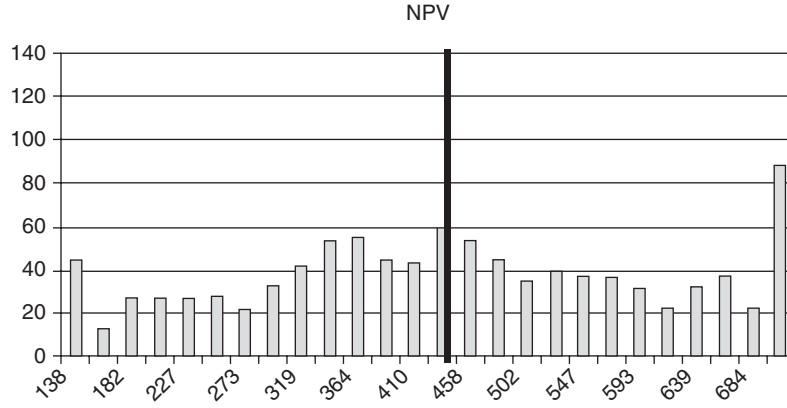


Figure 4. Probability maximizing portfolio.

measure of risk containment. The resulting portfolio has the following attributes:

$$\begin{aligned}\mu_{NPV} &= \$438\text{M}, \\ \sigma_{NPV} &= \$140\text{M}, \\ P(5)_{NPV} &= \$241\text{M}.\end{aligned}$$

By using the fifth percentile instead of the standard deviation as a measure of risk, we are able to obtain an outcome that shifts the distribution of returns to the right, compared to Case 1, as shown in Fig. 3.

This case clearly outperforms Case 1. Not only do we obtain significantly better financial performance but we also achieve a higher personnel utilization rate and a more diverse portfolio.

Case 3: Probability Maximizing and Value-at-Risk. In Case 3, the decision is to determine participation levels (0, 1) in each project with the objective of maximizing the probability of meeting or exceeding the mean NPV found in Case 1. This objective is expressed in Equation (30) of the following model:

$$\text{Maximize } P(NPV \geq \$394\text{M}) \quad (30)$$

Subject to :

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (31)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j \quad (\text{personnel constraints}) \quad (32)$$

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i \quad (\text{mutual exclusivity}) \quad (33)$$

$$x_i \in \{0, 1\} \forall i \quad (\text{binary decisions}). \quad (34)$$

This case focuses on maximizing the chance of achieving a goal and essentially combines performance and risk containment into one metric. The probability in Equation (30) is not known *a priori*, so we must rely on the simulation to obtain it. The resulting optimal solution yields a portfolio that has the following attributes:

$$\mu_{\text{NPV}} = \$440\text{M},$$

$$\sigma_{\text{NPV}} = \$167\text{M},$$

$$P(5) = \$198\text{M}.$$

Although this portfolio has a performance similar to the one in Case 2, it has a 70% chance of achieving or exceeding the NPV goal. As can be seen in the graph of Fig. 4, we have succeeded in shifting the probability distribution even further to the right, therefore increasing our chances of exceeding the returns obtained with the traditional Markowitz approach. In addition, in Cases 2 and 3, we need not make any assumption about the distribution of expected returns.

As a related corollary to this last case, we can conduct an interesting analysis that addresses Value-at-Risk (VaR). In traditional (securities) portfolio management, VaR is defined as the worst expected loss under normal market conditions over a specific time interval and at a given confidence level. In other words, VaR measures how much the holder of the portfolio can lose with probability $= \alpha$ over a certain time horizon [16]. In the case of project portfolios, we can define VaR as the probability that the NPV of the portfolio will fall below a specified value. Going back to our present case, the manager may want to limit the probability of incurring negative returns. In this example, we formulate the problem in a slightly different way: we still want to maximize the expected return, but we limit the probability that we incur a loss to $\alpha = 1\%$ by using the requirement shown in Equation (36) as

follows:

$$\text{Maximize} \quad \mu_{\text{NPV}} \quad (35)$$

Subject to :

$$P(\text{NPV} < 0) \leq 1\% \quad (\text{requirement}) \quad (36)$$

$$\sum_i c_i x_i \leq b \quad (\text{budget constraint}) \quad (37)$$

$$\sum_i p_{ij} x_i \leq P_j \quad \forall j \quad (\text{personnel constraints}) \quad (38)$$

$$\sum_{m_i \in M} x_i \leq 1 \quad \forall i \quad (\text{mutual exclusivity}) \quad (39)$$

$$x_i \in \{0, 1\} \forall i \quad (\text{binary decisions}). \quad (40)$$

The portfolio performance under this scenario is

$$\mu_{\text{NPV}} = \$411\text{M},$$

$$\sigma_{\text{NPV}} = \$159\text{M},$$

$$P(5) = \$195\text{M}.$$

These results from the VaR model turn out to be slightly inferior to the case where the probability was maximized. This is not a surprise, since the focus is on limiting the probability of downside risk, whereas before, the goal was to maximize the probability of obtaining a high expected return. However, this last analysis could prove valuable for a manager who wants to limit the VaR of the selected portfolio. As shown here, for this particular set of projects, a very good portfolio can be selected with that objective in mind.

CONCLUSIONS

Practically every real-world situation involves uncertainty and risk, creating a need for optimization methods that can handle uncertainty in model data and input parameters. We have briefly described two popular methods, *scenario optimization* and *robust optimization*, that seek to overcome limitations of classical optimization approaches for dealing with uncertainty, and which undertake to find high-quality solutions that are feasible under as many scenarios as possible. However, these methods are unable

to handle problems involving moderately large numbers of decision variables and constraints, or involving significant degrees of uncertainty and complexity. In these cases, simulation optimization is becoming the method of choice. The combination of simulation and optimization affords all the flexibility of the simulation engine in terms of defining a variety of performance measures as desired by the decision maker. In addition, as we demonstrate through a project portfolio selection example, modern optimization engines can enforce requirements on one or more outputs from the simulation, a feature that scenario-based methods cannot handle. Finally, simulation optimization produces results that can be conveyed and grasped in an intuitive manner, providing an especially useful tool for identifying improved business decisions under risk and uncertainty.

REFERENCES

1. Dembo R. Scenario optimization. *Ann Oper Res* 1991;30:63–80.
2. Kouvelis P, Yu G. Robust discrete optimization and its applications. Dordrecht: Kluwer Academic Publishers; 1997. pp. 8–29.
3. Mulvey JM, Vanderbei RJ, Zenios SA. Robust optimization of large-scale systems. *Oper Res* 1995;43(2):264–281.
4. Campos V, Laguna M, Martí M. Scatter search for the linear ordering problem. *New methods in optimization*. New York: McGraw-Hill; 1999. pp. 331–339.
5. Glover F. A template for scatter search and path relinking. Volume 1363, *Artificial evolution. Lecture notes in computer science*. New York: Springer; 1998. pp. 13–54.
6. Glover F, Laguna M. *Tabu search*. Norwell (MA): Kluwer Academic Publishers; 1997.
7. Glover F, Laguna M, Martí R. Fundamentals of scatter search and path relinking. *Control Cybern* 2000;29(3):653–684.
8. Glover F, Laguna M, Martí R. *Scatter search, advances in evolutionary computing: theory and applications*. New York: Springer; 2003. pp. 519–537.
9. Laguna M, Martí R. *Scatter search: methodology and implementations in C*. Norwell (MA): Kluwer Academic Publishers; 2003. pp. 217–254.
10. Haskett WJ. Optimal appraisal well location through efficient uncertainty reduction and value of information techniques. *Proceedings of the Society of Petroleum Engineers Annual Technical Conference and Exhibition*; Denver (CO). 2003.
11. Haskett WJ, Better M, April J. Practical optimization: dealing with the realities of decision management. *Proceedings of the Society of Petroleum Engineers Annual Technical Conference and Exhibition*; Houston (TX). 2004.
12. Markowitz H. Portfolio selection. *J Finance* 1952;7(1):77–91.
13. April J, Glover F, Kelly JP. Portfolio optimization for capital investment projects. *Proceedings of the 2002 Winter Simulation Conference*. San Diego (CA). 2002. pp. 1546–1554.
14. April J, Glover F, Kelly JP. Optfolio - a simulation optimization system for project portfolio planning. *Proceedings of the 2003 Winter Simulation Conference*. New Orleans (LA). 2002. pp. 301–309.
15. Kelly JP. Simulation optimization is evolving. *INFORMS J Comput* 2002;14(3):223–225.
16. Benninga S, Wiener Z. Value-at-risk (VaR). *Math Educ Res* 1998;7(4):1–7.

SIMULTANEOUS ASCENDING AUCTIONS

PETER CRAMTON
Department of Economics,
University of Maryland,
College Park, Maryland

INTRODUCTION

One of the most successful methods for auctioning many related items is the simultaneous ascending auction. This auction form was first developed for the US Federal Communications Commission's (FCC) spectrum auctions, beginning in July 1994, and has subsequently been adopted with slight variation for spectrum auctions worldwide, resulting in revenues in excess of \$250 billion. The method, first proposed by Paul Milgrom, Robert Wilson, and Preston McAfee, has been refined with experience, and extended to the sale of divisible goods in electricity, gas, and environmental markets. This article describes the method and its extensions, provides evidence of its success, and suggests what we can learn from the years of experience conducting simultaneous ascending auctions.

A key feature of the simultaneous ascending auction, at least in its basic form, is the requirement that bids be for individual items, rather than packages of items. As a result, the simultaneous ascending auction exposes bidders to the possibility that they will win some, but not all, of what they desire. In contrast, combinatorial auction methods [1] eliminate this exposure problem by allowing bidders to bid on packages of items. Nonetheless, the simultaneous ascending auction has not been supplanted by more powerful combinatorial methods, but is an essential method that any auction designer should have in his toolkit. The simultaneous ascending auction (and its variants) will remain the best method for auctioning many related items in a wide range of

circumstances, even settings where some of the goods are complements for some bidders, so the exposure problem is a real concern.

True combinatorial auctions—those that allow package bids—are also important, and indeed essential in the most complex auction environments. Still there are many situations where the simultaneous ascending auction will do a sufficiently good job in limiting the exposure problem that its other advantages (especially simplicity and price discovery) make it a preferred auction method. In other settings, where complementarities are both strong and varied across bidders, package bids are needed to improve the efficiency of the auction mechanism.

The simultaneous ascending auction is a natural generalization of the English auction when selling many goods. The key features are that all the goods are on the block at the same time, each with a price associated with it, and the bidders can bid on any of the items. The bidding continues until no bidder is willing to raise the bid on any of the items. Then the auction ends with each bidder winning the items on which it has the highest bid, and paying its bid for any items won.

The reason for the success of this simple procedure is the excellent price discovery it affords. As the auction progresses, bidders see the tentative price information and condition subsequent bids on this new information. Over the course of the auction, bidders are able to develop a sense of what the final prices are likely to be, and can adjust their purchases in response to this price information. To the extent that the price information is sufficiently good and bidders retain sufficient flexibility to shift toward their best package, the exposure problem is mitigated—bidders are able to piece together a desirable package of items, despite the constraint of bidding on individual items rather than packages.

Perhaps even more importantly, the price information helps the bidders focus their valuation efforts on the relevant range of the price space. This benefit of price discovery is ignored in the standard auction models,

which assume that each bidder costlessly knows his valuation for every possible package (or at least knows how his valuation depends on the information of others in the case of nonprivate value models). However, in practice, determining values is a costly process. When there are many items, determining precise values for all possible packages is not feasible. Price discovery focuses the bidders' valuation efforts to the relevant parts of the price space, improving efficiency and reducing transaction costs.

To further mitigate the exposure problem, most simultaneous ascending auctions allow bidders to withdraw bids. This enables bidders to back out of failed aggregations, shifting bids to more fruitful packages. However, bid withdrawals often facilitate undesirable gaming behavior, and thus the ability to withdraw bids needs to be constrained carefully. Price discovery—not bid withdrawal—is the more effective way to limit the exposure problem in simultaneous ascending auctions.

Since July 1994, the FCC has conducted dozens of spectrum auctions with the simultaneous ascending format, raising tens of billions for the US Treasury. The auctions assigned thousands of licenses to hundreds of firms. These firms provide a diversity of wireless communication services, now enjoyed by a vast majority of the population. The FCC is not alone. Countries throughout the world are now using auctions to assign spectrum. Indeed, the early auctions in Europe for third-generation (3G), mobile wireless licenses raised nearly \$100 billion. Auctions have become the preferred method of assigning spectrum and most have been simultaneous ascending auctions. (See Refs 2 and 3 for a history of the auctions.)

There is now substantial evidence that this auction design has been successful. Resale of licenses has been uncommon, suggesting that the auction assignment is nearly efficient. Revenues have often exceeded industry and government estimates. The simultaneous ascending auction may be partially responsible for the large revenues. By revealing information in the auction process, bidder uncertainty is reduced, and bidders can safely bid more aggressively. Also, revenues may increase to the extent

that the design enables bidders to piece together more efficient packages of items.

Despite the general success, the simultaneous ascending auctions have experienced a number of problems from which one can draw important lessons. One basic problem is the simultaneous ascending auction's vulnerability to revenue-reducing strategies in situations where competition is weak. Bidders have an incentive to reduce their demands in order to keep prices low, and to use bid signaling strategies to coordinate on a split of the items.

This article begins by motivating the design choices in a simultaneous ascending auction and discusses an important extension, the simultaneous clock auction, for the sale of many divisible goods. Then it describes typical rules including many important details. Next the performance of the simultaneous ascending auction is examined, identifying both strengths and weaknesses of the approach. Many refinements to the auction method, which have been introduced from early experience, are discussed.

AUCTION DESIGN

The critical elements of the simultaneous ascending auction are (i) open bidding, (ii) simultaneous sale, and (iii) no package bids. These features create a Walrasian pricing process that yields a competitive equilibrium provided (i) items are substitutes, (ii) bidders are price takers, and (iii) bid increments are negligible [3]. Of course, these conditions do not hold in practice. Some degree of market power is common, at least some items are complements, and bid increments in the 5–10% range are required to get the auction to conclude in a manageable number of rounds.

Still, the simultaneous ascending auction does perform well in practice largely because of the benefits of price discovery that come from open bidding and simultaneous sale. These benefits take two forms. First, in situations where bidder values are interdependent, price discovery may reduce the winner's curse and raise revenue [4]. Bidders are able to bid more aggressively, since they have better information about the item's value. More

importantly, when many items are for sale, the price discovery lets the bidders adapt their bidding and analysis to the price information, which focuses valuation efforts and facilitates the aggregation of a complementary package of items.

The alternative of sequential sale limits information available to bidders and limits how the bidders can respond to information. With sequential auctions, bidders must guess what prices will be in future auctions when determining bids in the current auction. Incorrect guesses may result in an inefficient assignment when item values are interdependent. A sequential auction also eliminates many strategies. A bidder cannot switch back to an earlier item if prices go too high in a later auction. Bidders are likely to regret having purchased early at high prices, or not having purchased early at low prices. The guesswork about future auction outcomes makes strategies in sequential auctions complex, and the outcomes less efficient.

Almost all the simultaneous ascending auctions conducted to date do not allow package bids. Bids are only for individual items. The main advantages of this approach are simplicity and anonymous linear prices. The auction is easily implemented and understood. The disadvantage is the exposure problem. With individual bids, bidding for a synergistic combination is risky. The bidder may fail to acquire key pieces of the desired combination, but pay prices based on the synergistic gain. Alternatively, the bidder may be forced to bid beyond its valuation in order to secure the synergies and reduce its loss from being stuck with the dogs. Bidding on individual items exposes bidders seeking synergistic combinations to aggregation risk. See Ref. 5 for an equilibrium analysis of the exposure problem in the simultaneous ascending auction (SAA), and Ref. 6 for experimental results.

Not allowing package bids can create inefficiencies. For example, suppose there are two bidders for two adjacent parking spaces. One bidder with a car and a trailer requires both spaces. He values the two spots together at \$100 and a single spot is worth nothing; the spots are perfect complements. The second bidder has a car, but no trailer. Either spot

is worth \$75, as is the combination; the spots are perfect substitutes. Note that the efficient outcome is for the first bidder to get both spots for a social gain of \$100, rather than \$75 if the second bidder gets a spot. Yet any attempt by the first bidder to win the spaces is foolhardy. The first bidder would have to pay at least \$150 for the spaces, since the second bidder will bid up to \$75 for either one. Alternatively, if the first bidder drops out early, he will “win” one spot, losing an amount equal to his highest bid. The only equilibrium is for the second bidder to win a single spot by placing the minimum bid. The outcome is inefficient and fails to generate revenue. In contrast if package bids are allowed, then the outcome is efficient. The first bidder wins both with a bid of \$75 for the package.

Unfortunately, allowing package bids creates other problems. Package bids may favor bidders seeking large aggregations due to a variant of the threshold problem. Continuing with the last example, suppose that there is a third bidder, who values either spot at \$40. Then the efficient outcome is for the individual bidders to win both spots for a social gain of $75 + 40 = \$115$. But this outcome may not occur when values are privately known. Suppose that the second and third bidders have placed individual bids of \$35 on the two spots, but these bids are topped by a package bid of \$90 from the first bidder. Each bidder hopes that the other will bid higher to top the package bid. The second bidder has an incentive to understate his willingness to push the bidding higher. He may refrain from bidding, counting on the third bidder to break the threshold of \$90. Since the third bidder cannot come through, the auction ends with the first bidder winning both spaces for \$90.

Package bidding also adds complexity. Unless the complementarities are large and heterogeneous across bidders, a simultaneous ascending auction without package bids may be preferred.

Clock Auctions

An important variation of the simultaneous ascending auction is the simultaneous clock auction. The critical difference is that bidders simply respond with quantities desired at prices specified by the auctioneer. Clock

auctions are especially effective in auctioning many divisible goods. There is a clock for each divisible good indicating its tentative price per unit quantity. Bidders express the quantity desired at the current prices. For those goods with excess demand the price is raised and bidders again express their desired quantities at the new prices. This process continues until supply just equals demand. The tentative prices and assignments then become final.

If we assume no market power and bidding is continuous, then the clock auction is efficient with prices equal to the competitive equilibrium [7].

Discrete, rather than continuous rounds, means that issues of bid increments, ties, and rationing are important. This complication is best handled by allowing bidders in each round to express their demand curves for all points along the line segment between the start of round prices and the end of round prices. Allowing a rich expression of preferences within a round makes bid increments, ties, and rationing less important. Since preferences for intermediate prices can be expressed, the efficiency loss associated with the discrete increment is less, so the auctioneer can choose a larger bid increment, resulting in a faster and less costly auction process.

TYPICAL RULES

The simultaneous ascending auction works as follows. A group of items with strong value interdependencies are up for auction at one time. A bidder can bid on any collection of items in any round, subject to an activity rule that determines the bidder's current eligibility. The auction ends when a round passes with no new bids on any item. This auction form was believed to give the bidders flexibility in expressing values and building packages of items. Common rules are described below:

Quantity Cap. To promote competition in the aftermarket, a bidder is often limited in the quantity that it can win.

Payment Rules. Typically payments are received at least at two times: a refundable deposit before the bidding begins

and a final payment for items won. The refundable deposit often defines the bidder's initial eligibility, that is, the maximum quantity of items that the bidder can bid for. A bidder interested in winning a large quantity of items would have to submit a large deposit. The deposit provides some assurance that the bids are serious. Some auctions require bidders to increase their deposit as bids increase.

Minimum Bid Increments. To assure that the auction concludes in a reasonable amount of time, minimum bid increments are specified. Bid increments are adjusted in response to bidder behavior. Typically, the bid increments are between 5 and 20%.

Activity Rule. The activity rule is a device for improving price discovery by requiring a bidder to bid in a consistent way throughout the auction. It forces a bidder to maintain a minimum level of activity to preserve its current eligibility. A bidder desiring a large quantity at the end of the auction (when prices are high) must bid for a large quantity early in the auction (when prices are low). As the auction progresses, the activity requirement increases, reducing a bidder's flexibility. The lower activity requirement early in the auction gives the bidder greater flexibility in shifting among packages early on when there is the most uncertainty about what will be obtainable.

Number of Rounds per Day. A final means of controlling the pace of the auction is the number of rounds per day. Typically, fewer rounds per day are conducted early in the auction when the most learning occurs. In the later rounds, there is much less bidding activity, and the rounds can occur more quickly.

Stopping Rule. A simultaneous stopping rule is used to give the bidders maximum flexibility in pursuing backup strategies. The auction ends if a single round passes in which no new bids

are submitted on any item. In auctions with multiple stages (later stages have higher activity requirements), the auction ends when no new bids are submitted in the final stage. Few or no bids in an earlier stage shifts the auction to the next stage.

Bid Information. The most common implementation is full transparency. Each bidder is fully informed about the identities of the bidders and the size of the deposits. High bids and bidder identities are posted after each round. In addition, all bids and bidder identities are displayed at the end of each round, together with each bidder's eligibility.

Bid Withdrawal. To limit the exposure problem, high bidders can withdraw their bids subject to a bid withdrawal penalty. If a bidder withdraws its high bid, the auctioneer is listed as the highest bidder and the minimum bid is the second-highest bid for that item. The second-highest bidder is in no way responsible for the bid, since this bidder may have moved on to other items. If there are no further bids on the item, the auctioneer can reduce the minimum bid. To discourage insincere bidding, there are penalties for withdrawing a high bid. The penalty is the larger of 0 and the difference between the withdrawn bid and the final sale price. This penalty is consistent with the standard remedy for breach of contract. The penalty equals the damage suffered by the seller as a result of the withdrawal.

PERFORMANCE OF THE SIMULTANEOUS ASCENDING AUCTION

Since we do not observe the values that the firms place on items, it is impossible to directly assess the efficiency of the simultaneous ascending auction from field data. Nonetheless, we can indirectly evaluate the auction design from the observed behavior in spectrum auctions [2,8].

Revenue is a first sign of success. Auction revenues have been substantial.

A second indicator of success is that the auctions tend to generate market prices. Similar items tend to sell for similar prices. Limited substitution, however, can cause some price anomalies [9].

A third indicator of success is the formation of efficient license aggregations. Bidders did appear to piece together sensible license aggregations in the early auctions.

An analysis of the US auction data in early auctions shows evidence of local synergies [10]. Consistent with local synergies, this study finds that bidders did pay more when competing with a bidder holding neighboring licenses. Hence, bidders did bid for synergistic gains and, judging by the final footprints, often obtained them.

DEMAND REDUCTION AND COLLUSIVE BIDDING

Despite the apparent success of these auctions, an important issue limiting both efficiency and revenues is demand reduction and collusive bidding. These issues become important in multiple-item auctions.

For example, suppose there are two identical items and two bidders (Large and Small). Large has a value of \$100 per item (\$200 for both); whereas Small has a value of \$90 for one item and values the second at \$0. In a simultaneous ascending auction with a reserve price less than \$80, the unique equilibrium outcome is for each to bid for a single item, so that the auction ends at the reserve price. Large prefers ending the auction at the reserve price, winning a single item for a profit of more than \$20, since to win both items Large would have to raise the price on both items above \$90 (it is a dominant strategy for Small to bid up to \$90 on each), which results in a profit of less than \$20. The demand reduction by Large harms both efficiency and seller revenues.

This logic is quite general. In multiunit uniform-price auctions, typically every equilibrium is inefficient [11]. Bidders have an incentive to shade their bids for multiple units, and the incentive to shade increases with the quantity being demanded. Hence, large bidders will shade more than small

bidders. This differential shading creates an inefficiency.

The simultaneous ascending auction can be viewed as an ascending-bid version of a uniform-price auction. For items that are close substitutes, the simultaneous ascending auction has generated near uniform prices. However, the incentives for demand reduction and collusive bidding are more likely pronounced in an ascending version of the uniform-price auction [12]. To illustrate this, consider a simple example with two identical goods and two risk-neutral bidders. Suppose that for each bidder the marginal value of winning one item is the same as the marginal value of winning a second item. These values are assumed independent and private, with each bidder drawing its marginal value from a uniform distribution between 0 and 100. First consider the sealed-bid uniform-price auction where each bidder privately submits two bids and the highest two bids secure units at a per-unit charge equal to the third-highest bid. There are two equilibria to this sealed-bid auction: a demand-reducing equilibrium where each bidder submits one bid for \$0 and one bid equal to its marginal value; and a sincere equilibrium where each bidder submits two bids equal to its marginal value. The sincere equilibrium is fully efficient in that both units will be awarded to the bidder, who values them more. The demand-reducing equilibrium, however, raises zero revenue (the third-highest bid is zero) and is inefficient since the bidder with the higher value wins only one unit.

Next consider the same setting, but where an ascending version of the auction is used. Specifically, view the ascending auction as a two-button auction where there is a price clock that starting from price 0 increases continuously to 100. The bidders depress the buttons to indicate the quantity they are bidding for at every instant. The buttons are “nonre-pushable” meaning a bidder can decrease its demand but cannot increase it. Each bidder observes the price and can observe how many buttons are being depressed by its opponent. The auction ends at the first price such that the total number of buttons depressed is less

than or equal to two. This price is the clearing price. Each bidder will win the number of units he demands when the auction ends, and is charged the clearing price for each unit he wins. In this game, if dominated strategies are eliminated, there is a unique equilibrium in which the bidding ends at a price of 0, with both bidders demanding just a single unit. The reason is that each bidder knows that if it unilaterally decreases its bidding to one unit, the other bidder will instantaneously end the auction, as argued above. But since the bidder prefers the payoff from winning one unit at the low price over its expected payoff of winning two units at the price high enough to eliminate the other bidder from the auction, the bidder will immediately bid for just one unit, inducing an *immediate* end to the auction. Thus, the only equilibrium here is analogous to the demand-reducing equilibrium in the sealed-bid uniform-price auction. The efficient equilibrium does not exist. This example shows that the incentives to reduce demand can be more pronounced in an open auction, where bidders have the opportunity to respond to the elapsed bidding. The 1999 German GSM spectrum auction, which lasted just two rounds, illustrates this behavior [13]; see Ref. 14 for experimental evidence.

For an auction format like the FCC’s, where the bidding occurs in rounds and bidding can be done on distinct units, there exist equilibria where the bidders coordinate a division of the available units at low prices relative to their own values [15,16]. Bidders achieve these low-revenue equilibria by threatening to punish those bidders, who deviate from the cooperative division of the units. The idea in these articles is that bidders have the incentives to split up the available units thereby ending the auction at low prices.

Collusive bidding strategies in the FCC’s DEF auction are examined in Ref. 12. During the DEF auction, the FCC and the Department of Justice observed that some bidders used bid signaling to coordinate the assignment of licenses. Specifically, some bidders engaged in *code bidding*. A code bid uses the trailing digits of the bid to tell other bidders on which licenses to bid or not to bid. Since bids were often in millions of dollars, yet were

specified in dollars, bidders at negligible cost could use the last three digits—the trailing digits—to specify a market number. Often, a bidder (the sender) would use these code bids as retaliation against another bidder (the receiver), who was bidding on a license desired by the sender. The sender would raise the price on some market that the receiver wanted, and use the trailing digits to tell the receiver on which license to cease bidding. Although the trailing digits are useful in making clear which market the receiver is to avoid, *retaliating bids* without the trailing digits can also send a clear message. The concern of the FCC is that this type of coordination may be collusive and may dampen revenues and efficiency.

Following the experience in the DEF auction, the FCC restricted bids to a whole number of bid increments above the standing high bid. This eliminates code bidding, but it does nothing to prevent retaliating bids. Retaliating bids may be just as effective as code bids in signaling a split of the licenses when competition is weak.

The auctioneer has many instruments to reduce the effectiveness of bid signaling. These include the following:

- *Concealing Bidder Identities.* This prevents the use of targeted punishments against rivals. Unless there are strong efficiency reasons for revealing identities, anonymous auctions may be preferable.
- *Setting High Reserve Prices.* High reserve prices reduce the incentive for demand reduction in a multiple-item auction, because as the reserve price increases the benefit from reducing demands falls. Moreover, higher reserve prices reduce the number of rounds that the bidders have to coordinate a split of the licenses and still face low prices.
- *Offering Preferences for Small Businesses and Nonincumbents.* Competition is encouraged by favoring bidders that may otherwise be disadvantaged *ex ante*. In the DEF auction, competition for the D and E licenses could have been increased by extending small business

preferences to the D and E blocks, rather than restricting the preferences to the F block.

- *Offering Larger Licenses.* Many small licenses are more easily split up. At the other extreme, a single nationwide license is impossible to split up. Such extreme bundling may have negative efficiency consequences, but improves revenues.

The inefficiencies of demand reduction can be eliminated with a Vickrey auction or more practically with a clock-proxy auction [9,17].

CONCLUSION

A simultaneous ascending auction is a sensible way to auction many items. This approach has been used with great success to auction spectrum, energy, and pollution permits. The process yields a competitive equilibrium in simple settings. Evidence from practice suggests that the desirable properties of the design are largely robust to practical complications, such as mild complementarities. However, when competition is weak, there is a tendency for collusive bidding strategies to appear, especially in the fully transparent implementation in which all bids and bidder identities are revealed after each round.

When some items are complements, the simultaneous ascending auction suffers from the exposure problem, since package bids are not allowed. To limit this problem, the auction rules typically allow the withdrawal of standing high bids. This allows a bidder to back out of a failed aggregation. Interestingly, an examination of withdrawals in the FCC spectrum auctions reveals that the withdrawals typically were not used for this desired purpose, but rather were mostly for undesirable bid signaling. As a result, current auctions limit a bidder's withdrawals to occur in just a few rounds of the auction.

Recent experience with simultaneous ascending auctions suggests that there are many advantages to the clock variation [7]. Most importantly, the ability of bidders to engage in tacit collusion through bid signaling is greatly reduced. Demand reduction,

however, remains a problem in the standard clock auction. However, demand reduction can be eliminated with the package clock auction [9].

REFERENCES

1. Cramton P, Shoham Y, Steinberg R. Combinatorial auctions. Cambridge (MA): MIT Press; 2006.
2. Cramton P. The FCC spectrum auctions: an early assessment. *J Econ Manag Strat* 1997;6(3):431–495.
3. Milgrom P. Putting auction theory to work. Cambridge (MA): Cambridge University Press; 2004.
4. Milgrom P, Weber RJ. A theory of auctions and competitive bidding. *Econometrica* 1982;50:1089–1122.
5. Goeree JK, Lien Y. An equilibrium analysis of the simultaneous ascending auction. Working Paper, University of Zurich, 2009.
6. Brunner C, Goeree JK, Holt CA, *et al.* An experimental test of flexible combinatorial spectrum auction formats. *Am Econ J Microecon* 2010;2:39–57.
7. Ausubel LM, Cramton P. Auctioning many divisible goods. *J Eur Econ Assoc* 2004;2:480–493.
8. McAfee RP, McMillan J. Analyzing the airwaves auction. *J Econ Perspect* 1996;10:159–176.
9. Cramton P. Spectrum auction design. Working Paper, University of Maryland, 2009.
10. Ausubel LM, Cramton P, Preston McAfee R, *et al.* Synergies in wireless telephony: evidence from the broadband PCS auctions. *J Econ Manag Strat* 1997;6(3):497–527.
11. Ausubel LM, Cramton P. Demand reduction and inefficiency in multi-unit auctions. University of Maryland, Working Paper 9607, revised July 2002.
12. Cramton P, Schwartz J. Collusive bidding in the FCC spectrum auctions. *Contrib Econ Anal Policy* 2002;1:1. Available at www.bepress.com/bejeap/contributions/vol1/iss1/art11.
13. Grimm V, Riedel F, Wolfstetter E. Low price equilibrium in multi-unit auctions: the GSM spectrum auction in Germany. *Int J Ind Organ* 2002;21:1557–1569.
14. Goeree JK, Offerman T, Sloof R. Demand reduction and preemptive bidding in multi-unit license auctions. Working Paper, University of Zurich, 2006.
15. Engelbrecht-Wiggans R, Kahn CM. Multi-unit auctions with uniform prices. *Econ Theor* 1998;12(2):227–258.
16. Brusco S, Lopomo G. Collusion via signalling in simultaneous ascending bid auctions with heterogeneous objects, with and without complementarities. *Rev Econ Stud* 2002;69:407–436.
17. Ausubel LM, Cramton P, Milgrom P. The clock-proxy auction: a practical combinatorial auction design. In: Cramton P, Shoham Y, Steinberg R, editors. *Combinatorial auctions*. Chapter 5. Boston (MA): MIT Press; 2006. pp. 115–138.

SIMULTANEOUS PERTURBATION AND FINITE DIFFERENCE METHODS

SHALABH BHATNAGAR
Department of Computer
Science and Automation,
Indian Institute of
Science, Bangalore,
India

INTRODUCTION

Optimization methods play an important role in many disciplines such as signal processing, communication networks, neural networks, economics, operations research, manufacturing systems, and so on. As examples, consider a communication network where the goal is to optimally allocate link bandwidth amongst competing flows, or a manufacturing plant where the goal is to decide the optimal order in which to allocate machine capacity for manufacturing various products. These are two specific instances of innumerable problems across various disciplines that fall within the broad category of optimization problems. Such problems are typically characterized via an objective function whose optimum constitutes the desired solution. These problems can be deterministic or stochastic, or they can be static or dynamic.

A general optimization problem (that we shall be concerned about in this article) has the following form:

$$\text{Find } \theta^* \text{ for which } J(\theta^*) = \min_{\theta \in C} J(\theta), \quad (1)$$

where $J: \mathcal{R}^N \rightarrow \mathcal{R}$ is called the *objective function*, θ is a tunable N -dimensional parameter and $C \subset \mathcal{R}^N$ is the constraint set in which θ takes values. If one has complete information about the function J , its derivatives, and so on, and about the set C , then Equation (1) is a deterministic optimization problem. If on the other hand, J is obtained as $J(\theta) = E_{\xi}[h(\theta, \xi)]$, where

$E_{\xi}[\cdot]$ is an expected value of observations or samples $h(\theta, \xi)$ with *random noise* ξ , and one is allowed to observe only these samples (without really knowing J), then one is in the realm of stochastic optimization. As we shall see subsequently, many times one resorts to search algorithms in order to find a minimum, that is, a solution to Equation (1). In stochastic optimization algorithms, it is not uncommon to make a random choice in the search direction. Thus, a second distinction between deterministic and stochastic optimization problems lies in the way in which search progresses: a random search algorithm results in the optimization setting being stochastic.

Suppose now that J has the form $J(\theta) = \sum_{i=1}^N E[h_i(X_i)]$, where X_i is the “state” of an underlying process in stage i (we have N stages in all here) and h_i denotes a stage- and state-dependent cost function. Let $\theta = (\theta_1, \dots, \theta_N)^T$, with each θ_j being a scalar ($j = 1, \dots, N$) and let X_i depend on $\theta_1, \dots, \theta_i$, $i = 1, \dots, N$. The idea is that optimization can be done one stage at a time over N stages after observing the state X_i in each stage i . The value θ_i in stage i (of the parameter) has a bearing on the cost of all subsequent stages $i+1, \dots, N$. This is a problem of dynamic optimization. Approaches such as dynamic programming are often used to solve dynamic optimization problems. Other manifestations of dynamic optimization, say over an infinite number of stages or in continuous time, also exist. In static optimization on the other hand, a one-shot optimization is performed even when there are multiple stages; that is, the parameters $\theta_1, \dots, \theta_N$ are optimized all at once in the first stage itself. There may also be only a single stage instead of multiple stages. Broadly speaking while in dynamic optimization one makes decisions “on the fly” as states are revealed, in static optimization, there is no explicit notion of time or perhaps even state as all decisions can be made at once. In this article, we shall be concerned with certain gradient estimation techniques for

stochastic optimization problems. We shall primarily consider the static optimization setting involving just a single stage. It may be worth noting that in many optimization problems, one is concerned with maximizing an objective unlike Equation (1) that considers minimization. However, such problems can be recast as Equation (1) by considering the objective function to be $-J(\theta)$ instead of $J(\theta)$.

We now provide an overview of the remainder of this article. In the next section, we briefly discuss well-known local search algorithms, mainly from the viewpoint of deterministic optimization. The algorithms that we review subsequently such as the finite-difference stochastic approximation (FDSA), simultaneous perturbation stochastic approximation (SPSA), higher-order algorithms, smoothed functional algorithms, and so on, are all local search algorithms, though for the stochastic optimization setting. These algorithms can be viewed as stochastic analogs of the deterministic algorithms in the next section.

A fundamental stochastic algorithm is the Robbins–Monro (R–M) algorithm [1] that estimates the zeros of a function. Most stochastic search algorithms can be viewed as variants of the R–M algorithm. In particular, the convergence analysis of stochastic search algorithms follows along the lines of that of the R–M algorithm. Hence in the third section, we give an overview of the R–M algorithm and discuss issues related to its convergence.

Amongst the first stochastic gradient search algorithms based on estimating the gradient of the objective function, is the FDSA algorithm [2] due to Kiefer and Wolfowitz. We review this algorithm in the section titled “Finite-Difference Stochastic Approximation.” As we shall see, FDSA is not efficient under high-dimensional parameters since the number of function measurements required to estimate the gradient using FDSA grows linearly with the parameter dimension.

Spall invented the SPSA algorithm [3,4] that is remarkable in that it requires only two function measurements at each instant, regardless of the parameter dimension by

randomly perturbing all parameter components. This algorithm has been found to be very efficient in various settings. Its development has resulted in a flurry of research activity during the last two decades with a significant amount of research focused around this algorithm. This research includes not just theoretical analyses and applications of SPSA in various disciplines, but also the development of similar algorithms. We review the SPSA algorithm in the section titled “Simultaneous Perturbation Stochastic Approximation.”

In the section titled “Variations to the Basic Scheme,” we review the development of algorithms that have a flavor similar to SPSA and possess a finite-difference structure in most cases. We discuss in particular, deterministic perturbations SPSA, higher-order algorithms that is those that estimate not just the gradient but also the Hessian, and the smoothed functional algorithm. We also review developments in multitimescale algorithms that are seen to be useful in areas such as simulation optimization and adaptive control. Further, we discuss developments involving applications of SPSA and similar algorithms on an optimization problem that is similar to Equation (1) but with functional (inequality) constraints.

Finally, in the last section, we provide an overview of the many diverse applications where SPSA and some of the SPSA-type algorithms have been successfully applied. Our aim in doing so is to bring to the fore the fact that SPSA and the related algorithms constitute powerful methods for stochastic optimization that have been found useful in many disciplines of science and engineering. Excellent (though less recent) surveys on the SPSA algorithm are available in Refs 5 and 6. The current survey is a more up-to-date account of the developments related to SPSA and as mentioned, also covers the algorithms developed along the lines of SPSA.

ALGORITHMS FOR LOCAL SEARCH

Search algorithms can be broadly classified into two categories: global search and local search algorithms. Global (resp. local) search

algorithms look for the global (resp. local) minimum. For the optimization problem (Eq. 1), we say that $\theta^* \in C$ is a global minimum of the function J if $J(\theta^*) \leq J(\theta) \forall \theta \in C$. On the other hand, we say that $\theta^* \in C$ is a local minimum of J if there exists an $\epsilon > 0$ such that $J(\theta^*) \leq J(\theta) \forall \theta \in C$ with $\|\theta - \theta^*\| < \epsilon$. Many times, the norm $\|\cdot\|$ is chosen to be the Euclidean norm. A well-known example of a global search technique is simulated annealing [7,8]. Even while it is desirable to obtain a global minimum, global search techniques are known to be very slow. Hence, quite often, the only acceptable option is to use local search methods.

A typical local search algorithm (ignoring noise effects for now) has the form [9–11]

$$\theta_{n+1} = \theta_n - a_n D(\theta_n)^{-1} \nabla J(\theta_n), \quad (2)$$

where a_n , $n \geq 0$ is called the *step-size* sequence; that is, a sequence of positive real numbers that asymptotically diminish to zero, $D(\theta_n)$ is a positive definite $N \times N$ matrix, and $\nabla J(\theta_n)$ denotes the gradient of $J(\theta)$ evaluated at $\theta = \theta_n$. Here $J(\cdot)$ is the objective function to be minimized (Eq. 1). Note that if $\nabla J(\theta_n)$ and $D(\theta_n)$ are analytically known quantities, then recursion Equation (2) can proceed as is and we are in the domain of deterministic optimization. On the other hand, if $J(\theta_n)$ is of the form $J(\theta_n) = E_\xi[h(\theta_n, \xi)]$ and we only have access to noisy cost samples $h(\theta_n, \xi)$, then quantities $\nabla J(\theta_n)$ and, in many cases, $D(\theta_n)$ have to be estimated. The algorithms for stochastic optimization inherently have randomness because of the presence of noise in the cost samples. In addition, the estimators of $\nabla J(\theta_n)$ and $D(\theta_n)$ may introduce additional randomness (as happens for instance in the SPSA algorithm, see the section titled “Simultaneous Perturbation Stochastic Approximation,” as well as the higher-order and SF algorithms in the section titled “Variations to the Basic Scheme”), resulting in the search direction being random as well. In order to bring out ideas clearly, we assume in the remainder of this section that $\nabla J(\theta_n)$ and $D(\theta_n)^{-1}$ are analytically known quantities for the updates as in Equation (2); that is, we have a deterministic optimization

framework. In the later sections, however, we shall be primarily concerned with the stochastic optimization setting. The algorithms there shall correspond to the stochastic analogs of Equation (2).

Given $\theta \in \mathcal{R}^N$ such that $\nabla J(\theta) \neq 0$, any $x \in \mathcal{R}^N$ satisfying $x^T \nabla J(\theta) < 0$ is called a *descent* direction since the directional derivative $x^T \nabla J(\theta)$ along the direction x is negative and thus by a Taylor’s expansion, one obtains

$$J(\theta + a_n x) = J(\theta) + a_n x^T \nabla J(\theta) + o(a_n). \quad (3)$$

Here, given two real-valued sequences $\{x_n\}$ and $\{y_n\}$, we say that $x_n = o(y_n)$ if $x_n/y_n \rightarrow 0$ as $n \rightarrow \infty$. From the above, one can see that $J(\theta + a_n x) < J(\theta)$ for a_n sufficiently small. Now since $D(\theta_n)$ is a positive definite matrix, so is $D(\theta_n)^{-1}$. Hence, $x = -D(\theta_n)^{-1} \nabla J(\theta_n)$ is a descent direction as well. Algorithms that update along descent directions are also called *descent algorithms*.

The following well-known algorithms are special cases of Equation (2):

1. *Gradient Algorithm*. This is the most commonly used descent algorithm. Here $D(\theta_n) = I$ (the N -dimensional identity matrix). This is also called the *steepest descent* algorithm since its updates are strictly along the direction of negative gradient.
2. *Jacobi Algorithm*. In this algorithm, $D(\theta_n)$ is set to be an $N \times N$ -diagonal matrix with its i th diagonal element $\nabla_{i,i}^2 J(\theta_n)$, that is, the second partial derivative of $J(\theta)$ with respect to the i th component of θ evaluated at $\theta = \theta_n$. The latter also corresponds to the (i,i) th element of the Hessian $\nabla^2 J(\theta_n)$. For $D(\theta_n)$ to be a positive definite matrix in this case, it is easy to see that all elements $\nabla_{i,i}^2 J(\theta_n)$, $i = 1, \dots, N$ need be positive.
3. *Newton Algorithm*. Here $D(\theta_n)$ is chosen to be $\nabla^2 J(\theta_n)$, the Hessian of $J(\theta_n)$.

The $D(\theta_n)$ matrices in the Jacobi and Newton algorithms by themselves need not be

positive definite (for all n) in general as they vary with θ_n and hence, should be projected appropriately after each parameter update so as to ensure that the resulting matrices are positive definite (pp. 88–98, [9]). With proper scaling provided by the $D(\theta_n)$ matrix, the descent directions obtained using Jacobi and Newton algorithms are preferable to the one using the gradient algorithm. However, obtaining estimates of the Hessian in addition to the gradient, in general requires much more computational effort. Our main focus in this paper is on gradient-based algorithms and subsequently we shall also review some development on higher-order (Newton/Jacobi) algorithms.

THE ROBBINS–MONRO ALGORITHM

We shall from now on be primarily concerned with the stochastic optimization problem. Given the function J , a first order necessary condition for a local minimum θ^* in the interior of C is $\nabla J(\theta^*) = 0$. The local search algorithms of the previous section aim to achieve this by converging to a point θ^* for which $\nabla J(\theta^*) = 0$ if θ^* is in the interior of C . Further, if θ^* is a local minimum that is on the boundary of C , the search algorithm can converge to θ^* regardless of whether or not $\nabla J(\theta^*) = 0$. The condition $\nabla J(\theta^*) = 0$ is not a sufficient condition for a local minimum. Thus, in principle, θ^* could be a maximum or a saddle point (minimum along some components of the function and maximum along some others) if there is one as well. It is normally the case, however, that stochastic search algorithms under sufficiently general noise conditions do indeed tend to converge to a local minimum.

The development in the area of root-finding stochastic algorithms (not necessarily for optimization) started in a seminal paper by Robbins and Monro [1]. The problem of finding the roots of a function $L : \mathcal{R}^N \rightarrow \mathcal{R}^N$ under noisy observations was considered in Ref. 1. The algorithms of the previous section in the stochastic optimization case can be considered to be special cases of this algorithm. For instance, in the case of the gradient algorithm, one can consider

$L(\theta) = \nabla J(\theta)$. All the algorithms that we review in the subsequent sections are based on the R–M algorithm. Because of its importance, we review the R–M algorithm in detail here.

The R–M Algorithm

Let $g(\theta, \xi_n)$ denote the n th observed sample of $L(\theta)$ when the parameter is $\theta \in \mathcal{R}^N$ and the sample is corrupted by noise $\xi_n \in \mathcal{R}^K$. Here $g : \mathcal{R}^N \times \mathcal{R}^K \rightarrow \mathcal{R}^N$ is such that $E_{\xi}[g(\theta, \xi_n)] = L(\theta)$. The noise random vectors $\xi_n, n \geq 1$ could be independent and identically distributed (i.i.d.) or more generally, may depend on the underlying system state. As we explain below, this setting is more general than the original setting considered in Ref. 1.

Let $\{a_n, n \geq 1\}$ be a deterministic sequence of positive real numbers that satisfy the property

$$\sum_n a_n = \infty, \sum_n a_n^2 < \infty. \quad (4)$$

The first condition above is required to ensure that the algorithm does not prematurely converge. The second condition is required to reduce the effect of noise. We discuss issues related to the convergence of the algorithm in the next section. Common examples of $\{a_n, n \geq 1\}$ include (a) $a_n = c_1/(n + c_2)$, (b) $a_n = c_1/(n + c_2)^\alpha$, $\alpha \in (0.5, 1)$, (c) $a_n = (\ln(n + c_1))/(n + c_2)$, (d) $a_n = c_1/(c_2 + n \ln n)$, and so on, for some suitably chosen constants c_1 and $c_2 > 0$.

Note that $g(\theta, \xi_n)$ can be written as $g(\theta, \xi_n) = L(\theta) + M_{n+1}$, where $M_{n+1} = g(\theta, \xi_n) - L(\theta)$, $n \geq 0$. The R–M stochastic approximation algorithm is as follows: for $n \geq 0$,

$$\theta_{n+1} = \theta_n - a_n(L(\theta_n) + M_{n+1}), \quad (5)$$

where θ_n is the n th update of the parameter θ . If θ is constrained to take values within a prescribed set $C \subset \mathcal{R}^N$ as with Equation (1), one can project after each iterate the value of θ_{n+1} to the set C . The new value of θ_{n+1} then corresponds to its projected value after the update.

The original R–M algorithm in Ref. 1 considered the form of Equation (5) of the

updates by assuming that the random vectors M_{n+1} have the zero-vector as their mean. Further, these vectors are independent and have a common distribution. In Ref. 1, it is not assumed that the samples have the form $g(\theta_n, \xi_n)$. The latter is a more general, nonadditive noise form. When the samples are obtained in the form $g(\theta_n, \xi_n)$, the random vectors M_{n+1} form a “martingale difference” sequence and are no longer i.i.d.. We say M_{n+1} , $n \geq 0$ form a martingale difference sequence if the conditional expectation $E[M_{n+1} | \theta_i, M_{i+1}, 0 \leq i \leq n-1, \theta_n] = 0$ with probability one. Analyses for the more general case are available for instance in Refs 12 and 13. Next, we briefly discuss the convergence of the algorithm 5.

Convergence of the Algorithm

We consider the form (Eq. 5) of the algorithm by assuming M_{n+1} , $n \geq 0$ to be a martingale difference sequence. The form of M_{n+1} , $n \geq 0$ is considered in Ref. 1 that is, of independent random vectors having a common distribution and with the zero-vector as their mean is a special case of them being a martingale difference sequence. Convergence in the mean square (m.s.) sense was shown in Ref. 1. The general case has been covered, for instance, in Refs 12 and 13 where convergence in the almost sure (a.s.) sense is shown. A good account of probability theory, including the various notions of convergence for a sequence of random vectors, can be found in Ref. 14. In order to prove convergence of recursions such as Equation (5), one needs to first ensure that the parameter updates in these recursions remain stable or uniformly bounded; that is, do not blow up. We briefly discuss this aspect below. However, before we proceed, it is important to make the point that assuming this is true; that is, the parameter updates are uniformly bounded and suppose they converge in the a.s. sense, then it can be shown that they also converge in the m.s. sense. The converse is however not true even when updates are bounded.

If the set C in which θ takes values is a bounded well-defined subset of $\mathcal{R}^N$, one would normally project the parameter after each update (as mentioned above) to the set

C , thereby ensuring that the obtained parameters are both feasible (i.e., take values in the set where they are allowed to take values in) and remain bounded. Even if C is unbounded but one roughly knows the region of the space where the “asymptotically stable” equilibria lie, one could choose a large bounded set (a subset of C) that contains the above region and use projection to ensure that the updates remain bounded. This would also imply that the remainder of the space is not visited by the algorithm which may, in fact, be good since the algorithm in such a case would not waste its resources in exploring the region of the space that does not contain the equilibria. The projection technique is often used to ensure the stability of iterations. Other approaches to prove stability of the iterations (when the constraint region is unbounded) include the stochastic Lyapunov technique [13] and the recently proposed approach in Refs 12 and 15, whereby one does a scaling of the original iteration (Eq. 5) to approximate the same with a deterministic process in a manner similar to the construction of the fluid model of Refs 16 and 17. This approach is remarkable in that using just an ordinary differential equation (ODE)–based analysis, one can prove stability of the original random recursion. Another approach [18] is to define a constraint region for the updates, use projection as explained above, but gradually increase the size of the constraint region as iterations progress.

An ODE-based argument whereby one shows that the noise effects vanish asymptotically under the step-size conditions (Eq. 4) and that the resulting iteration follows asymptotically a limiting ODE, once stability of the updates (discussed above) is ensured, has been followed in Refs 12, 13, 19–21. We remark here that the ODE used to prove stability of the stochastic updates in Ref. 15 is different from the ODE used to show convergence once stability is established. The former ODE corresponds to the scaled stochastic recursion there while the latter is described below and is commonly used to establish convergence given that the updates are stable. The ODE associated with Equation (5) is

$$\dot{\theta}(t) = -L(\theta(t)), \quad (6)$$

where $\dot{\theta}(t)$ is the derivative of $\theta(t)$ relative to t . If θ^* is a unique asymptotically stable equilibrium of Equation (6), then one can prove under certain conditions a.s. convergence to θ^* that is, $\theta_n \rightarrow \theta^*$ a.s. as $n \rightarrow \infty$, [12,13,21]. In the case of multiple isolated equilibria, the algorithm shall converge to one amongst them, depending on the noise and initial condition. By isolated equilibria, we mean that one can construct certain sufficiently small open neighborhoods such that exactly one equilibrium is contained within each neighborhood. Results for the case when the set of stable fixed points comprises of certain “invariant sets” that are not necessarily isolated equilibria, are available in Refs 12, 13 and 22.

The R–M algorithm, when applied in the context of gradient-based schemes, aims at finding the root of the function gradient when it is not analytically known but only noisy samples are available. In what follows, we consider the stochastic optimization problem of finding roots of the function gradient under noisy samples.

FINITE-DIFFERENCE STOCHASTIC APPROXIMATION

In the R–M algorithm (Eq. 5), suppose that $L(\theta_n) = \nabla J(\theta_n)$. Then one can show that under certain conditions, Equation (5) converges to a local minimum of J . We are thus in the domain of stochastic gradient algorithms; that is, gradient algorithms with noise (see the section titled “Algorithms for Local Search”). Suppose $J(\theta) = E_\xi[h(\theta, \xi)]$, that is, we are in the stochastic optimization setting. If $\nabla h(\theta, \xi)$ exists for any given noise sample ξ , then under certain regularity conditions, for instance as given in Refs 23–25, one can interchange the expectation and gradient operators to obtain $\nabla J(\theta) = E[\nabla h(\theta, \xi)]$. In such a case, Equation (5) can be applied with $M_{n+1} = (\nabla h(\theta_n, \xi_n) - L(\theta_n))$ as the martingale difference term to obtain asymptotic convergence to a local minimum of J . Infinitesimal perturbation analysis and its variants are largely based on this idea 23–27. In practice, however, it is difficult to have access to direct gradient measurements. It

is also possible that while the function J is continuously differentiable with bounded higher-order derivatives, the function h is not (differentiable). One therefore, requires other gradient estimation techniques.

The first approach in this direction goes under the name of FDSA [2]. The basic algorithm here is still Equation (5), except with $L(\theta) = \nabla J(\theta)$ approximated by a finite-difference estimate (Eq. 7) with effective noise $M_{n+1} = (M_{n+1}^i, i = 1, \dots, N)^T$, where $M_{n+1}^i = (\xi_{n+1,i}^+ - \xi_{n+1,i}^-)/2c_n$, $n \geq 0$. We assume that the vectors $(\xi_{n+1,i}^+ - \xi_{n+1,i}^-, i = 1, \dots, N)^T$, $n \geq 0$ form a martingale difference sequence. The algorithm is thus in the form of Equation (5). The quantity multiplying $a(n)$ on the RHS of Equation (5), that is, $(\nabla J(\theta_n) + M_{n+1})$ (assuming $L(\theta_n) = \nabla J(\theta_n)$) is approximated here by the N -vector $Y_n = (Y_{n,1}, \dots, Y_{n,N})^T$, where

$$Y_{n,i} = \frac{(J(\theta_n + c_n e_i) + \xi_{n+1,i}^+) - (J(\theta_n - c_n e_i) + \xi_{n+1,i}^-)}{2c_n} \quad (7)$$

$$= \frac{J(\theta_n + c_n e_i) - J(\theta_n - c_n e_i)}{2c_n} + \frac{\xi_{n+1,i}^+ - \xi_{n+1,i}^-}{2c_n}, i = 1, \dots, N. \quad (8)$$

Here e_i denotes the unit vector with 1 in the i th place and 0 elsewhere. In Equation (7), $J(\theta_n + c_n e_i) + \xi_{n+1,i}^+$ and $J(\theta_n - c_n e_i) + \xi_{n+1,i}^-$ correspond to independent noise-corrupted measurements of the objective function $J(\cdot)$ under parameters $\theta_n + c_n e_i$ and $\theta_n - c_n e_i$, respectively. In the following, we simply represent the other gradient estimates in a similar manner, as in Equation (8) instead of Equation (7), to suit the form $(L(\theta_n) + M_{n+1})$ in Equation (5).

The scalar parameters c_n , $n \geq 0$ should be carefully chosen so that $c_n \rightarrow 0$ (as $n \rightarrow \infty$) at a slow rate in order that the variance in the FDSA estimates does not blow up. We explain conditions on c_n , $n \geq 0$ in the case of the SPSA algorithm in the next section. Similar conditions hold for the FDSA algorithm as well. Quite often, it makes sense to simply set $c_n = \delta$ for a ‘small’ $\delta > 0$ as has been done in Refs 28–30 (see also p. 15 of Ref. 13 for a discussion along these lines). In Refs 31 and

32, it is seen that the use of common random numbers, that is $\xi_{n,i}^+ = \xi_{n,i}^-$ reduces the estimator variance. This may be possible in some simulation-based settings with random variables obtained from the same pseudorandom sequence. However, in most practical applications, this is difficult to achieve even with simulation. FDSA [2] is also often called the *Kiefer–Wolfowitz algorithm* in honor of its inventors.

A variant of FDSA that is based on one-sided differences is described next. The gradient estimates in Equation (8) are also called *two-sided finite-difference estimates* while those that we describe below in Equation (9) are called *one-sided estimates*. In one-sided FDSA, one uses Equation (5) with the following estimate of $(\nabla J(\theta_n) + M_{n+1})$ in place of Equation (8) (assuming $L(\theta_n) = \nabla J(\theta_n)$):

$$Y_{n,i} = \frac{J(\theta_n + c_n e_i) - J(\theta_n)}{c_n} + \frac{\xi_{n+1,i}^+ - \xi_{n+1,i}^-}{c_n}, \quad 1, \dots, N. \quad (9)$$

Note that the estimates (Eq. 8) require $2N$ function measurements in order to obtain one estimate Y_n of the gradient. With Equation (9), one requires $(N+1)$ function measurements to obtain a gradient estimate. In the setting of simulation optimization [11,23,26,27,30,33,34], where one does not have access to function measurements but simulates the whole system, one requires in effect $2N$ (resp. $(N+1)$) parallel simulations of the entire system when using two-sided (resp. one-sided) estimates. This makes the procedure highly computationally expensive when N becomes large and therefore, one requires more efficient methods for gradient estimation.

SIMULTANEOUS PERTURBATION STOCHASTIC APPROXIMATION

Spall [3,4] invented a remarkable algorithm that has become popular by the name simultaneous perturbation stochastic approximation (SPSA). It is remarkable in that it requires only two function measurements for a parameter of any dimension

(i.e., any $N \geq 1$) and exhibits fast convergence (normally faster than FDSA). The two function measurements are obtained from simultaneously perturbing all components of the parameter vector randomly. The two perturbed parameters correspond to $\theta_n + c_n \Delta_n$ and $\theta_n - c_n \Delta_n$, respectively, where $\Delta_n = (\Delta_{n,1}, \dots, \Delta_{n,N})^T$ is a vector of independent, mean-zero, symmetric random variables that satisfy an inverse moment bound. Most often, one assumes these random variables to be distributed according to the symmetric Bernoulli distribution with $\Delta_{n,i} = \pm 1$ w.p. $1/2$, $i = 1, \dots, N$, $n \geq 0$. In fact, it is found in Ref. 35 that under certain conditions, the optimal distribution on components of the simultaneous perturbation vector is a symmetric Bernoulli distribution. The authors in Ref. 35 obtain this result under two separate objectives: (i) minimize the mean square error of the estimate and (ii) maximize the likelihood that the estimate remains in a symmetric bounded region around the true parameter.

As with the previous section, we assume here as well that $L(\theta)$ in Equation (5) has the form $L(\theta) = \nabla J(\theta)$. In the SPSA algorithm, the quantity $(L(\theta_n) + M_{n+1})$ in Equation (5) is replaced by $Y_n = (Y_{n,1}, \dots, Y_{n,N})^T$ with

$$Y_{n,i} = \frac{J(\theta_n + c_n \Delta_n) - J(\theta_n - c_n \Delta_n)}{2c_n \Delta_{n,i}} + \frac{\xi_{n+1,i}^+ - \xi_{n+1,i}^-}{2c_n \Delta_{n,i}}, \quad (10)$$

where the random vectors $(\xi_{n+1,i}^+ - \xi_{n+1,i}^-)$, $n \geq 0$ are assumed to form a martingale difference sequence. Further, $\Delta_{n,i}$, $i = 1, \dots, N$ are independent of the above random vectors.

An interesting observation here is that while in FDSA, given θ_n , the estimate of $\nabla J(\theta_n)$, ignoring the noise term $(\xi_{n+1,i}^+ - \xi_{n+1,i}^-)/2c_n$ (Eq. 8), is deterministic; it is not so in SPSA. The randomness in the gradient estimate (even without the noise term) in SPSA arises due to the randomness in the search direction that in turn, is caused by the random perturbations Δ_n . Thus, even if the noise term in the SPSA estimate (Eq. 10) $((\xi_{n+1,i}^+ - \xi_{n+1,i}^-)/2c_n \Delta_{n,i})$ is not present, one would still be in the stochastic optimization

setting because of the randomness in search direction. This aspect was briefly discussed in the first two sections of this article. Note that Equation (10) has both terms stochastic, while this is not the case with Equation (5) where the term multiplying $a(n)$ has a deterministic part ($L(\theta_n)$, given θ_n) and a stochastic part (M_{n+1}). Through suitable Taylor series expansions, however (see Eq. 11), the first term on the RHS of Equation (10) can be further split into a deterministic term (i.e., precisely the gradient of J) and a stochastic term (together with an asymptotically diminishing error). The latter stochastic term can be combined with the second stochastic term in Equation (10) to obtain a dominant stochastic term, that again turns out to be a martingale difference. We describe the SPSA algorithm in detail below.

The SPSA Algorithm.

Step 0 (Initialize): Set $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,N})^T$ and obtain $\Delta_0 = (\Delta_{0,1}, \dots, \Delta_{0,N})^T$, where $\Delta_{0,i}$ are independent samples obtained according to $\Delta_{0,i} = \pm 1$ w.p. 1/2, $i = 1, \dots, N$. Select parameter sequences $\{a_n, n \geq 0\}$ and $\{c_n, n \geq 0\}$. Make two sample measurements $Y_0^+ = J(\theta_0 + c_0 \Delta_0) + \xi_1^+$ and $Y_0^- = J(\theta_0 - c_0 \Delta_0) + \xi_1^-$ with parameters $(\theta_0 + c_0 \Delta_0)$ and $(\theta_0 - c_0 \Delta_0)$, respectively. Select a large finite integer M . Set $n := 0$.

Step 1: Update parameter $\theta_n = (\theta_{n,1}, \dots, \theta_{n,N})^T$ according to: For $i = 1, \dots, N$,

$$\theta_{n+1,i} = \theta_{n,i} - a_n \left(\frac{Y_n^+ - Y_n^-}{2c_n \Delta_{n,i}} \right).$$

Step 2: Set $n := n + 1$.

- If $n < M$, obtain new samples $\Delta_{n,i}$ independently according to $\Delta_{n,i} = \pm 1$ w.p. 1/2. Form the vectors $\theta_n = (\theta_{n,1}, \dots, \theta_{n,N})^T$ and $\Delta_n = (\Delta_{n,1}, \dots, \Delta_{n,N})^T$. Make two measurements $Y_n^+ = J(\theta_n + c_n \Delta_n) + \xi_{n+1}^+$ and $Y_n^- = J(\theta_n - c_n \Delta_n) + \xi_{n+1}^-$, respectively. Go to Step 1.

- If $n = M$, terminate algorithm and output $\theta_n = (\theta_{n,1}, \dots, \theta_{n,N})^T$ as the final value of the parameter.

We assume above that the algorithm terminates after M iterations. Asymptotic convergence is then achieved as $M \rightarrow \infty$. More sophisticated stopping criteria may however be used as well. Here the parameters a_n, c_n are such that $a_n, c_n \rightarrow 0$ as $n \rightarrow \infty$. Because of the presence of c_n in the denominator of the martingale difference term, one requires here that $\sum_n \left(\frac{a_n}{c_n} \right)^2 < \infty$ in addition to $\sum_n a_n = \infty$, that is similar to the first condition in Equation (4). The former condition above is however stronger than the second condition in Equation (4) and is required in order to use the martingale convergence theorem (Theorem 11, Appendix C of Ref. 12) to show that an associated martingale sequence converges. The above assumption on the parameters $a_n, c_n, n \geq 0$ has been made in Ref. 3. In Ref. 18, a relaxation is made whereby it is assumed that $a_n, c_n \rightarrow 0$ as $n \rightarrow \infty$ and that $\sum_n a_n = \infty$, $\sum_n a_n^p < \infty$ for some $p \in (1, 2]$. Typically $a_n, c_n, n \geq 0$ can be chosen according to $a_n = a/(A + n + 1)^\alpha$ and $c_n = c/(n + 1)^\gamma$, for $a, A, c > 0$. The suggested values of α and γ given in Refs 36 and 37 are 1 and 1/6 respectively. In Ref. 38, it is observed that the choices $\alpha = 0.602$ and $\gamma = 0.101$ perform better in practical settings.

The reason why SPSA is a valid gradient algorithm is easy to see from Taylor expansions of $J(\theta_n + c_n \Delta_n)$ and $J(\theta_n - c_n \Delta_n)$ around the parameter θ_n as under

$$\begin{aligned} J(\theta_n + c_n \Delta_n) &= J(\theta_n) + c_n \Delta_n^T \nabla J(\theta_n) \\ &\quad + \frac{c_n^2}{2} \Delta_n^T \nabla^2 J(\theta_n) \Delta_n + o(c_n^2), \\ J(\theta_n - c_n \Delta_n) &= J(\theta_n) - c_n \Delta_n^T \nabla J(\theta_n) \\ &\quad + \frac{c_n^2}{2} \Delta_n^T \nabla^2 J(\theta_n) \Delta_n + o(c_n^2). \end{aligned}$$

From the above two equations, we obtain

$$\begin{aligned} \frac{J(\theta_n + c_n \Delta_n) - J(\theta_n - c_n \Delta_n)}{2c_n \Delta_{n,i}} &= \nabla_i J(\theta_n) \\ &\quad + \sum_{j=1, j \neq i}^N \frac{\Delta_{n,j}}{\Delta_{n,i}} \nabla_j J(\theta_n) + o(c_n). \end{aligned} \quad (11)$$

Now observe that the conditional expectation of the estimate given the parameter θ_n equals

$$\begin{aligned} E \left[\frac{J(\theta_n + c_n \Delta_n) - J(\theta_n - c_n \Delta_n)}{2c_n \Delta_{n,i}} \mid \theta_n \right] \\ = \nabla_i J(\theta_n) + o(c_n), \end{aligned} \quad (12)$$

from the properties on the perturbations $\Delta_{n,i}$. Thus, while the search direction is randomly chosen and need not follow a descent path in a given update, the algorithm is seen to follow asymptotically the right path.

In Ref. 39, a one-measurement version of SPSA has been presented. The simulation here is run with the parameter $\theta_n + c_n \Delta_n$ with Y_n having the form

$$Y_{n,i} = \frac{J(\theta_n + c_n \Delta_n)}{c_n \Delta_{n,i}} + \frac{\xi_{n+1}^+}{c_n \Delta_{n,i}}, \quad (13)$$

$i = 1, \dots, N$ in place of Equation (10) with Δ_n as before. A calculation similar to the two-measurement SPSA, based on a Taylor series expansion, shows that

$$\begin{aligned} \frac{J(\theta_n + c_n \Delta_n)}{c_n \Delta_{n,i}} &= \frac{J(\theta_n)}{c_n \Delta_{n,i}} + \nabla_i J(\theta_n) \\ &+ \sum_{j=1, j \neq i}^N \frac{\Delta_{n,j}}{\Delta_{n,i}} \nabla_j J(\theta_n) \\ &+ \frac{c_n}{2} \frac{\Delta_n^T \nabla^2 J(\theta_n) \nabla_n}{\Delta_{n,i}} + o(c_n). \end{aligned} \quad (14)$$

As has been noted in Ref. 39 and other references, for instance Ref. 40, that even while a similar conclusion as Equation (12) can be seen to be valid for the one-measurement version, its performance primarily suffers because of the presence of the first term on the RHS of Equation (14) as the factor c_n in the denominator there can make the term excessively large for large n (because $c_n \rightarrow 0$ as $n \rightarrow \infty$). In particular, two-simulation SPSA (Eq. 10) has been seen to be much superior in comparison to one-simulation SPSA (Eq. 13).

A one-sided version of SPSA with two measurements has been considered in Ref. 18. Here the two simulations are

run with parameters $\theta_n + c_n \Delta_n$ and θ_n , respectively, and Y_n above has the form

$$Y_{n,i} = \frac{J(\theta_n + c_n \Delta_n) - J(\theta_n)}{c_n \Delta_{n,i}} + \frac{\xi_{n+1}^+ - \xi_{n+1}^-}{c_n \Delta_{n,i}}, \quad (15)$$

$i = 1, \dots, N$ with Δ_n as before. A Taylor series expansion here shows that a second-order term that cancels in the case of two-sided SPSA, remains in the one-sided version. This term vanishes as $n \rightarrow 0$, however, it adds slightly to the overall bias in the estimate. The two-sided form, Equation (10) is the most studied and used in applications.

It is interesting to note that in the expression (Eq. 10) [also Eqs (15) and (13)] for the $Y_{n,i}$, the numerator is the same across all i , $i = 1, \dots, N$, while the denominator has a quantity $\Delta_{n,i}$ that depends on i . This is unlike FDSA where the denominator is the same while the numerator is different for different i , see Equations (8) and (9). Thus, in going from FDSA to SPSA, the complexity in estimating the gradient shifts (in a way) from the numerator of the estimator to its denominator. It should be noted that simulating N independent symmetric Bernoulli random variables, is in general, far less computationally expensive than obtaining $2N$ or $(N+1)$ objective function measurements or simulations, particularly when N is large. It has been seen from both, theory and experiments [3,30,41], that two-sided, two-measurement SPSA is computationally far superior to FDSA. In Ref. 3, asymptotic normality results for SPSA and FDSA are used to establish the relative efficiency of SPSA. The asymptotic analysis for the R-M algorithm can be adapted to prove a.s. convergence of the iterates in the SPSA algorithm [3]. Assuming that there is a unique globally asymptotically stable equilibrium θ^* for the associated ODE (i.e., a global minimum for the basic algorithm), the asymptotic normality result in Ref. 3 essentially says that

$$n^{r/2}(\theta_n - \theta^*) \xrightarrow{D} N(\mu, \Sigma)$$

as $n \rightarrow \infty$, where $\xrightarrow{D}$ denotes convergence in distribution, $N(\mu, \Sigma)$ is a multivariate Gaussian with mean μ and covariance matrix Σ that depends on the Hessian at θ^* . In general, $\mu \neq 0$. The quantity r depends upon the choice of the gain sequences $a_n, c_n, n \geq 0$.

Many interesting analyses of the SPSA algorithm have been reported in the literature. In Ref. 3, the above asymptotic normality result is used to argue the relative asymptotic efficiency of SPSA over FDSA. In particular, it is argued that SPSA results in an N -fold computational savings over FDSA. In Ref. 18, a projected version of SPSA, where the projection region is gradually increased, has been presented. This is a novel approach to take care of the issue of iterate stability in general. In Refs 42 and 43, application of SPSA for optimization in the case of nondifferentiable functions is considered. A detailed analysis of SPSA under general conditions can also be found in Ref. 44. An analysis of SPSA and FDSA when common random numbers are used in the simulations is given in Ref. 32. Different ways of gradient estimation in SPSA using past measurements have been reported in Refs 45 and 46. In Ref. 47, iterate averaging for stability of the SPSA recursions and improved algorithmic behavior is explored. A case of weighted averaging of the Kiefer–Wolfowitz and SPSA iterates is considered in Ref. 48. In Refs 49 and 50, SPSA is proposed for use as a global search algorithm. In Ref. 36, SPSA is compared with a two-sided smoothed functional algorithm (see the section titled “The Smoothed Functional Algorithm”) and it is found that SPSA is the better of the two algorithms. This algorithm is referred to as *random directions stochastic approximation* (RDSA) in Ref. 36. In Ref. 38, the general technique for implementing SPSA and the choice of gain sequences is discussed.

The SPSA algorithm has evoked significant interest due to its good performance, ease of implementation, and wide applicability. Moreover, it is observed to be scalable in that the computational effort does not increase enormously with parameter dimension. This is not the case with FDSA (see for instance, the experiments in Ref. 30). In the next two sections, we shall

review some of the recent research directions related to the SPSA algorithm and many of the various applications, in fact application areas/domains where the original algorithm and also its variants have been applied and found useful.

VARIATIONS TO THE BASIC SCHEME

In this section, we describe some of the variations to the basic scheme.

Multitimescale Algorithms

Consider the following variation to the R–M scheme:

$$\theta_{n+1} = \theta_n + a_n(f(\theta_n, \omega_n) + M_{n+1}^1), \quad (16)$$

$$\omega_{n+1} = \omega_n + b_n(g(\theta_n, \omega_n) + M_{n+1}^2). \quad (17)$$

Here, Equations (16) and (17) are coupled updates that involve a parameter ‘ ω ’ in addition to θ . Also, $M_{n+1}^1, M_{n+1}^2, n \geq 0$ are certain martingale difference sequences w.r.t. the coupled recursions. Further, $a_n, b_n, n \geq 0$ are two step-size schedules that satisfy the regular step-size conditions (Eq. 4). In addition, we also require $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$. An example of $a_n, b_n, n \geq 0$ is $a_n = c_1/(n+1), b_n = c_2/(n+1)^\alpha$, for any $c_1, c_2 > 0$ and $\alpha \in (0.5, 1)$. Recursions governed by $\{a_n\}$ would appear to be quasi-static when viewed from the “timescale of $\{b_n\}$ ” while from the “timescale of $\{a_n\}$,” the other recursions would appear to be equilibrated. Thus recursions (Eq. 16) are called the “slower recursions” while recursions 17 are the “faster ones.” A general analysis of two-timescale stochastic approximation using an ODE approach is provided in Ref. 51; see also Ref. 12.

Multitimescale algorithms are helpful in cases when in between two successive updates of the algorithm, one typically has to perform an inner-loop procedure recursively until it converges. Thus, one would in practice, have to wait for a long time before updating the algorithm once. By using a multitimescale algorithm as in Equations (16) and (17), both recursions (for

the inner and outer loops) can run together, and convergence to the desired point can be achieved. Key application areas where this procedure has been successfully applied are simulation optimization and adaptive control. We review these application areas in the next section. There is another reason why multitimescale algorithms are interesting. In Ref. 52, averaging of stochastic approximation iterates in the case of one-timescale algorithms such as Equation (5) has been seen to improve the rate of convergence. The same procedure can be accomplished using a two-timescale algorithm such as Equations (16) and (17) wherein the “averaging” is performed along the faster timescale.

In Ref. 53, a smoothed version of SPSA that is seen to improve performance is presented. The idea there is similar to the averaging in Ref. 52 and the resulting algorithm has a multitimescale character. In Refs 54 and 55, the step-sizes a_n , $n \geq 0$ are adaptively set according to the objective function value obtained. Since the update direction in SPSA is random, a move in the descent direction in Refs 54 and 55 is rewarded by a slightly higher step-size in the next update step. A move in the ascent direction, on the other hand, attracts a penalty. Moreover, if the objective function value gets worse, a certain blocking mechanism is enforced whereby starting from the previous estimate, a new gradient evaluation is made with a reduced step-size a_n . The procedure of Refs 54 and 55 is performed for the smoothed version of SPSA making the overall scheme again of the multitimescale type.

The convergence theory of multitimescale methods has been dealt with in detail in Refs 11, 12, 28–30, 34, 40, 51, 56, and 57 and many other references.

Deterministic Perturbations SPSA

In the basic SPSA framework, the perturbations $\Delta_i(n)$ were considered to be random variables. In Ref. 58, certain deterministic conditions on the perturbation vectors and noise sequences in SPSA are obtained under which the algorithm is shown to converge. In Refs 40 and 59, certain deterministic constructions for the perturbations in SPSA have

been described. Each deterministic construction in the above references relies on obtaining a suitable set in which the $\{\pm 1\}^N$ -valued vectors Δ_n take values in, and assigning in a cyclic manner, the elements in the set to the vectors Δ_n at each of the instants $n \geq 0$.

It is observed in Ref. 40 that the perturbation vectors in a deterministic construction (assuming they are $\{\pm 1\}^N$ -valued) should satisfy the following property in the case of two-sided SPSA (Eq. 10): There exists an integer $P > 1$ such that for all $i, j \in \{1, \dots, N\}$, $i \neq j$ and for any $s \geq 1$,

$$\sum_{n=s}^{s+P} \frac{\Delta_i(n)}{\Delta_j(n)} = 0. \quad (18)$$

For one-measurement SPSA, one requires the following property in addition to Equation (18): there exists a $P > 1$ such that for every $k \in \{1, \dots, N\}$ and any $s \geq 1$,

$$\sum_{n=s}^{s+P} \frac{1}{\Delta_k(n)} = 0. \quad (19)$$

The idea behind using deterministic perturbations is to try and reduce the bias in the gradient estimates by cancelling bias terms in a regular/deterministic manner.

One of the constructions in Ref. 40 that we briefly describe is observed to significantly improve the performance of one-simulation SPSA (Eq. 13). The construction is via “normalized Hadamard matrices.” Let $K = \lceil \log_2(N+1) \rceil$. Obtain a $2^K \times 2^K$ -matrix H_{2^K} in a recursive manner as follows:

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\text{and } H_{2^l} = \begin{bmatrix} H_{2^{l-1}} & H_{2^{l-1}} \\ H_{2^{l-1}} & -H_{2^{l-1}} \end{bmatrix}, \quad l > 1.$$

Having obtained H_{2^K} , form a new matrix Q of dimension $2^K \times N$ by removing the first column of H_{2^K} and selecting any N of its remaining columns. Let $Q(1), \dots, Q(2^K)$ denote the rows of Q . Now set $\Delta_0 = Q(1)$, $\Delta_1 = Q(2), \dots, \Delta_{2^K-1} = Q(2^K)$, $\Delta_{2^K} = Q(0)$ etc. Thus, the perturbations are obtained by cyclically passing through the rows of Q . SPSA with deterministic perturbations

has been successfully applied in some applications [60–62].

Higher-Order SPSA

The algorithms FDSA and SPSA described in the previous two sections are gradient algorithms (see the section titled “Algorithms for Local Search”). Adaptive stochastic approximation algorithms based on computing the Hessian (in addition to the gradient) estimates typically require many more samples of the objective function than those that estimate only the gradient. For instance, in Ref. 37, the Hessian is estimated using finite differences that are in turn based on finite-difference estimates of the gradient. This requires $O(N^2)$ samples of the objective function at each update epoch. In Ref. 63, an adaptive multivariate version of the Robbins-Monro [1] algorithm is obtained for the case where the objective function gradients are known and the Hessian is estimated using finite gradient differences.

Recently in Ref. 64, a simultaneous perturbation version of Newton’s algorithm was presented. The Hessian estimates here are obtained using four objective function samples at each update epoch in cases where the gradient estimates are not known, and three samples in cases where these are known. This is achieved by using two independent perturbation sequences with random variables in each assumed bounded, zero-mean, symmetric, having a common distribution and mutually independent of one another, and with finite inverse moments. This method is an extension of regular SPSA [3] that estimates only the gradient and requires only one perturbation sequence. The Hessian updates are projected to the set of positive definite matrices (necessary for a *descent* direction). In Ref. 65, a similar algorithm is presented where the projected Hessian is replaced by the geometric mean of the projected eigenvalues of it. Certain enhancements to the Jacobian or Hessian estimates [64] that improve their quality have recently been presented in Ref. 66.

In Ref. 11, four adaptive multitimescale Newton-based algorithms in the setting of simulation optimization have been proposed.

These are all based on the simultaneous perturbation method and require four, three, two, and one simulation(s) respectively. The four-simulation algorithm has a similar Hessian estimate as the one in Ref. 64. In addition, different estimates for the Hessian have been derived in Ref. 11. The algorithms in Ref. 11 are developed in the setting of simulation optimization and use three timescales where on the fastest scale the Hessian is averaged, on the middle timescale data averaging is done, and on the slowest scale, the optimization parameter θ is updated. We briefly describe below the algorithm of Ref. 64.

Let $\Delta_{n,k}, \hat{\Delta}_{n,k}, k = 1, \dots, N, n \geq 0$ denote the two independent perturbation sequences. Let $\Delta_n = (\Delta_{n,1}, \dots, \Delta_{n,N})^T$ and $\hat{\Delta}_n = (\hat{\Delta}_{n,1}, \dots, \hat{\Delta}_{n,N})^T$, respectively. Consider four independent simulations or function measurements with parameters $(\theta_n + c_n \Delta_n), (\theta_n - c_n \Delta_n), (\theta_n + c_n \Delta_n + \hat{c}_n \hat{\Delta}_n)$, and $(\theta_n - c_n \Delta_n + \hat{c}_n \hat{\Delta}_n)$, respectively. Here $c_n, \hat{c}_n \rightarrow 0$ as $n \rightarrow \infty$. The estimate of the (k, l) th component of the Hessian ($\hat{H}_n$) has the form $\hat{H}_n^{k,l} = (4c_n \hat{c}_n)^{-1} [(\Delta_{n,k} \hat{\Delta}_{n,l})^{-1} + (\Delta_{n,l} \hat{\Delta}_{n,k})^{-1}] [(J(\theta_n + c_n \Delta_n + \hat{c}_n \hat{\Delta}_n) + \xi_{n+1}^{++}) - (J(\theta_n + c_n \Delta_n) + \xi_{n+1}^+) - (J(\theta_n - c_n \Delta_n + \hat{c}_n \hat{\Delta}_n) + \xi_{n+1}^{+-}) + (J(\theta_n - c_n \Delta_n) + \xi_{n+1}^-)]$, $k, l = 1, \dots, N$. Here, $\xi_{n+1}^+, \xi_{n+1}^-, \xi_{n+1}^{++}$ and ξ_{n+1}^{+-} , $n \geq 0$ are certain martingale difference sequences that are in general independent of each other. Then

$$\theta_{n+1} = \theta_n - a_n \bar{H}_n^{-1} \left(\frac{J(\theta_n + c_n \Delta_n) - J(\theta_n - c_n \Delta_n)}{2c_n} + \frac{\xi_{n+1}^+ - \xi_{n+1}^-}{2c_n} \right) \Delta_n^{-1}, \quad (20)$$

$$H_n^1 = \frac{n}{n+1} H_{n-1}^1 + \frac{1}{n+1} \hat{H}_n, \quad (21)$$

$$\bar{H}_n = f_n(H_n^1), \quad (22)$$

$n \geq 0$. Here $\Delta_n^{-1} = (\Delta_{n,1}^{-1}, \dots, \Delta_{n,N}^{-1})^T$. Also, f_n is a map that projects the Hessian update to the set of positive definite matrices. In practice, performing this step and taking inverse of the projected Hessian matrix are time-consuming operations, more so when N is large. The Jacobi version of the algorithm [11,64], whereby only the diagonal entries of the Hessian are updated and the off-diagonal

terms are set to zero, is an efficient way to use some of the Hessian information to improve algorithm behavior while the computational effort increases only marginally (over gradient SPSA) in doing so. This procedure is usually recommended for large N [11,64]. Here a_n , c_n , and $\hat{c}_n$ can be chosen according to $a_n = a/(n+A)^\alpha$, $c_n = c/n^\gamma$ and $\hat{c}_n = \hat{c}/n^\gamma$, respectively, with $a, c, \hat{c} > 0$ and $A \geq 0$. In Ref. 64, the values recommended for α and γ are $\alpha = 0.602$ and $\gamma = 0.101$, respectively.

The Smoothed Functional Algorithm

In Ref. 67, a one-simulation “smoothed functional (SF)” algorithm was presented. The idea there is to convolve the gradient of the objective function with a multivariate Gaussian density and then by an appropriate integration by parts argument, shows that it is the same as a convolution of the objective itself with a scaled Gaussian. Hence, by slowly diminishing the variance parameter in the Gaussian, one recovers the desired gradient using one simulation. This algorithm is called the *smoothed functional algorithm* because the procedure of convolving with a Gaussian smooths the objective. We explain the key idea here. Given a constant $\beta > 0$, let

$$D_\beta J(\theta) \triangleq \int G_\beta(\theta - \eta) \nabla_\eta J(\eta) d\eta \quad (23)$$

represent the convolution of the gradient of average cost ($J(\cdot)$) with the N -dimensional multivariate normal p.d.f.

$$G_\beta(\theta - \eta) = \frac{1}{(2\pi)^{N/2} \beta^N} \times \exp\left(-\frac{1}{2} \sum_{i=1}^N \frac{(\theta_i - \eta_i)^2}{\beta^2}\right),$$

where $\theta \triangleq (\theta_1, \dots, \theta_N)^T$ and $\eta = (\eta_1, \dots, \eta_N)^T$, respectively. Integrating by parts in Equation (23), it is easy to see that

$$\begin{aligned} D_\beta J(\theta) &= \int \nabla_\theta G_\beta(\theta - \eta) J(\eta) d\eta \\ &= \int \nabla_\eta G_\beta(\eta) J(\theta - \eta) d\eta. \end{aligned} \quad (24)$$

One can easily check that $\nabla_\eta G_\beta(\eta) = \frac{-\eta}{\beta^2} G_\beta(\eta)$ [34,56]. Substituting $\eta' = \frac{\eta}{\beta}$ in Equation (24), one then obtains

$$D_\beta J(\theta) = \frac{1}{\beta} \int -\eta' G_1(\eta') J(\theta - \beta\eta') d\eta', \quad (25)$$

where $G_1(\eta') = G_\beta(\eta')|_{\beta=1}$. Replacing $-\eta'$ by η , one gets the following SF estimate for the gradient [67]:

$$\nabla J(\theta_n) \approx \frac{1}{\beta} \frac{1}{M} \sum_{n=1}^M \eta_n J(\theta_n + \beta\eta_n), \quad (26)$$

for $\beta > 0$ small and M large. Here η_n is an N -vector of independent $N(0, 1)$ -distributed random variables.

The following symmetric two-sided finite-difference form of SF estimates has been studied in Refs 34, 36, 68, and 69:

$$\begin{aligned} \nabla J(\theta_n) &\approx \frac{1}{\beta} \frac{1}{M} \sum_{n=1}^M \eta_n (J(\theta_n + \beta\eta_n) \\ &\quad - J(\theta_n - \beta\eta_n)), \end{aligned} \quad (27)$$

for small β and large M as before. The balanced estimate (Eq. 27) has less bias as compared to Equation (26). This can be seen using suitable Taylor series expansions as in the case of the SPSA estimates in the section titled “Simultaneous Perturbation Stochastic Approximation.” A detailed convergence analysis of this algorithm in the multiscale setting is presented in Ref. 34.

The SF algorithm falls within the category of finite-difference methods, one that is also based on the idea of random perturbations. In Ref. 36, it is observed that SF is not as computationally efficient as SPSA. Further, generating Gaussian random variables requires more computational effort than generating Bernoulli random variables. However, the objective smoothing that results because of the convolution can, in fact, help the algorithm to converge to a global minimum. This fact has been observed in Ref. 68. In experiments on an admission control problem with dependent service times [70], where the objective function is seen to

be almost flat with a sharp dip in between, it is observed that many times SF converges to the global minimum. However, there is no theory currently available that proves convergence of SF to a global minimum. It would be interesting if guidelines can be suggested for guaranteed convergence of this and other algorithms to a global minimum. Perturbations in SF need not be confined to normal random variables but more generally, the uniform and Cauchy distributions can also be used.

The above ideas have been explored further in Ref. 34 and two SF estimates for the Hessian have been derived for the first time. Using these estimates for the Hessian, and the above estimates for the gradient, two Newton SF algorithms have been developed in Ref. 34. Thus one of the algorithms there requires only one measurement or simulation (with parameter $\theta_n + \beta\eta_n$) while the other requires two of them (with parameters $\theta_n + \beta\eta_n$ and $\theta_n - \beta\eta_n$, respectively). The Hessian estimates have been derived using the (above) integration by parts argument twice. An advantage with this approach is that the same simulations that are used to estimate the gradients are also used to estimate the Hessians. Thus, with a small increase in computational effort (using the Jacobi case say), one obtains a good second-order algorithm that has the potential to converge to a global minimum.

Optimization under Inequality Constraints

We consider here the following variation to the basic problem (Eq. 1).

$$\begin{aligned} \text{Find } \theta^* \text{ for which } J(\theta^*) = \min_{\theta \in C} \{J(\theta) \mid G_i(\theta) \\ \leq \alpha_i, \quad i = 1, \dots, p\}. \end{aligned} \quad (28)$$

Here $G_i(\cdot)$ and $\alpha_i, i = 1, \dots, p$ are certain prescribed functions and constants, respectively, that constitute the functional constraints. Such problems arise for instance in communication networks, where the objective function J could correspond to the negative of the average throughput (because the problem is one of minimization and not maximization, see

the first section), and a constraint may specify that the average delay must be below a threshold. Another constraint may similarly specify that the probability of packet loss during transmission must not be more than a small constant, say 0.01.

There is a vast literature on optimization problems with functional constraints. More recently, SPSA has been applied on these problems as well. In Ref. 71, a constrained optimization setting in which the inequality constraints are explicit functions of the optimization parameters is considered and an SPSA-based algorithm is presented in which the parameter is projected to the constraint set after each update. In Ref. 72, the penalty function approach is used together with SPSA for an optimization problem with inequality constraints. As with Ref. 71, it is assumed in Ref. 72 that the constraint functions are analytically known as *functions of the underlying parameter*.

For a constrained optimization problem on a discrete set, certain penalty function-based SPSA algorithms are presented in Ref. 73 that differ from each other in the manner in which the projection is done in order to obtain a parameter within the discrete set of parameters. Applications to a problem of resource allocation are also studied in Ref. 73.

In Ref. 74, a problem of optimization under inequality constraints is considered, where both the objective and constraint functions are certain long-run averages. Thus, neither the objective nor the constraints are known analytically and so projection after each parameter update to the constraint region is ruled out. A Lagrange multiplier approach is used in Ref. 74 and the problem is formulated in the multitimescale stochastic approximation framework. Algorithms based on (gradient and Newton) SPSA and SF are presented for this problem in Ref. 74. Experiments on a communication network scenario show good performance using these algorithms.

APPLICATIONS OF SPSA

SPSA has been successfully applied in a wide range of applications and in a

variety of disciplines such as electrical and communications engineering, mechanical engineering, industrial engineering, civil engineering, management science, physics, operations research, chemical engineering, and biology. We describe here some of the representative work on applications of SPSA and algorithms developed along the lines of SPSA. Reference 75 provides an up-to-date web site on SPSA that contains many of the papers included in this survey. This web site also includes many other papers on SPSA that are not included here and which are by no means any less interesting.

Discrete Parameter Optimization

It is well known that discrete optimization problems are hard combinatorial problems. SPSA, which is primarily a continuous optimization algorithm, has been successfully adapted in many applications involving discrete optimization problems [70,76–80]. In Refs 70, 76 and 77, discrete parameter SPSA has been applied on problems in admission control in communication networks.

Estimating the Fisher Information Matrix

There has been work on using the simultaneous perturbation approach for estimating the Fisher information matrix that is widely used in system identification problems. Reference 81 is an early work on natural gradient estimation using SPSA, where the Fisher information matrix plays a key role. In Ref. 82, a resampling based method to estimate the Fisher information matrix has been proposed. In Ref. 83, an extension to the approach in Ref. 82 is presented for the case when parts of the matrix are known *a priori*.

Simulation Optimization

Simulation optimization refers to a class of techniques that are applicable when the system model is not known, however, the system can be simulated and the outcomes from the simulation can be used to optimize the system parameters [28,30,33]. Most often, one is interested in optimizing an objective function $J(\cdot)$ that is a long-run average quantity.

For instance, in a queueing system, think of minimizing the long-run average waiting time of a customer. A difficulty with such an objective function is that in order to use an algorithm such as SPSA, one needs to know the value of $J(\theta_n)$ at each update θ_n of the parameter. Many of the perturbation analysis-based approaches [23,25–27] use only one simulation but require strict regularity conditions on the system and performance functions. Moreover, quite often parameter updates here are performed only at regeneration epochs of a certain underlying process.

In Refs 28 and 29, two-timescale stochastic approximation (see section titled “Multitimescale Algorithms”) is used for this purpose where on the faster scale, data is aggregated and on the slower one, the parameter is updated. The FDSA gradient estimates were used in Refs 28 and 29. In Ref. 30, SPSA-based estimates were used that were seen to significantly improve performance over the algorithms in Refs 28 and 29. More recently, in Refs 11 and 34, Newton-based algorithms have been developed in the multitimescale framework. In Ref. 69, one of the algorithms of Ref. 34 has been applied for a problem of optimal trajectory estimation in parameterized stochastic differential equations.

Optimal Control and Reinforcement Learning

There has been significant work in the areas of optimal control and reinforcement learning where SPSA has been applied. For instance, in Refs 84 and 85 the SPSA algorithm has been applied for solving infinite-horizon Markov decision processes (MDPs) when model information is not known, under the discounted cost and long-run average cost criteria respectively. In Refs 86 and 87, the finite-horizon MDP problem is similarly considered. The case when the underlying state-space of the system could be potentially large is also considered because of which certain feature-based representations are used. In Ref. 62, a reinforcement learning algorithm using deterministic perturbations SPSA has been developed and studied for the problem of optimal mortgage refinancing. In Ref.

88, certain multitimescale Q-learning algorithms in the area of reinforcement learning have been developed. In Ref. 60, an algorithm for long-run average cost control of Markov chains conditioned on rare events has been developed. Also, in Ref. 89, two actor-critic algorithms for hierarchical control of Markov decision processes have been developed. All the above algorithms are based on some form of SPSA (with randomized or deterministic perturbations and one or two simulations). For a problem of adaptive control, Spall and Cristion [90] use SPSA and show convergence for a time-varying objective. An application of SPSA in an optimal control setting has also been studied in Ref. 91.

Communication and Wireless Networks

SPSA has been applied in problems of control and optimization in communication networks. In Ref. 57, SPSA has been applied for finding optimal policies in available bit rate (ABR) service in asynchronous transfer mode (ATM) networks. In the context of the random early detection (RED) scheme for the Internet, Vaidya and Bhatnagar [92] developed and applied a variant of SPSA that works with only the sign of the objective (this algorithm can be of independent interest in communications engineering where it is desirable to make decisions based on one or at most a few bits of information). In Ref. 61, a higher-order-SPSA-based algorithm has been applied for tuning RED parameters. This algorithm shows better performance than most well-known algorithms in the RED literature [61]. The problem of admission control using SPSA has been studied in Refs 70, 76, and 77. Application of an SPSA-based control scheme in dynamic routing has been studied in Ref. 88. In Ref. 93, SPSA has been applied for finding an optimal slot assignment policy in bluetooth wireless networks. In Ref. 94, a measurement-based routing algorithm derived from SPSA has been proposed for several unicast sessions. In Ref. 95, SPSA has been applied on a problem of optimal placement of acoustic sensors for a system.

Neural Networks

SPSA has been applied in many applications related to neural networks. For instance, in

Ref. 96, SPSA is used to train the weights of a neural network controller for controlling the light timings at a traffic junction in a traffic network. A recursive learning scheme for recurrent neural networks using SPSA has been proposed in Ref. 97. In Ref. 98, an on-line training algorithm based on SPSA is used in a neural tracking control system. A modified SPSA algorithm with certain adaptive gain sequences and smoothing of iterates across recursions is used in the design of a neural network controller in Ref. 54. In Ref. 99, SPSA is used for training in robust neural network tracking controllers in nonlinear systems. In Ref. 53, a smoothed SPSA estimation procedure has been used for modeling the underlying control mechanism in a neural network that is used for the control of dynamic systems [100]. In Ref. 101, SPSA is applied for cascade learning in single hidden layer neural networks. Certain modifications to the SPSA algorithm using gradient smoothing and adaptive gains with an application to training neural networks have been studied [55].

Pattern Recognition and Image Processing

SPSA has been applied in pattern recognition and signal/image processing applications as well. In Ref. 102, an SPSA-based algorithm for feature selection in a nearest neighbor classifier is proposed. In Ref. 103, SPSA is applied for feature-relevance estimation in video retrieval applications. In Ref. 104, a combination of Tsallis entropy and SPSA is used to register images efficiently. In Ref. 105, SPSA has been successfully applied for efficient image restoration. A nonlinear inversion procedure for images has also been proposed [106].

Traffic Optimization

SPSA has been applied to various aspects of traffic optimization, for both airline and land-based traffic. In Ref. 107, the problem of minimizing airtraffic delay is treated as a constrained optimization problem and a penalty function approach is used together with SPSA. In Ref. 108, an optimal gate-holding schedule to minimize cost of delays and air congestion in airtraffic networks is

obtained using SPSA. In Ref. 109, an application to aerodynamic design using SPSA is studied. In Ref. 110, an SPSA-based model has been proposed for the problem of seat allocation in airlines.

In Ref. 111, SPSA is used in a function approximation-based scheme to control the timing of signals in a traffic network. In Ref. 112, SPSA is applied to find optimal traffic signal timings for light rail transit (LRT) vehicles, see also Ref. 113 where SPSA is used in an application of traffic signal control as well. In Ref. 114, SPSA with a perturbation distribution over a smaller set than the one obtained from Bernoulli sequences, is used to optimize fuel consumption in engine operations.

Other Interesting Applications

In some other applications, SPSA has been applied in game programming [115]. In Ref. 116, an implementation of SPSA on an analog very large scale integration (VLSI) chip has been discussed. In Ref. 117, an SPSA-based approach is used for dynamic optimization in systems requiring continuous measurements. In Ref. 118, an application of SPSA to electrocardiogram (ECG) analysis has been studied. In Ref. 119, a “system of systems” view of military tasks/operations is presented and the first- and second-order SPSA algorithms are used for optimizing costs under given constraints. In Ref. 41, FDSA and SPSA have been applied for on-line estimation of parameters in nonlinear/non-Gaussian state-space models with particle filters. In Ref. 120, a simulation optimization framework with SPSA estimates has been considered for inventory control applications in semiconductor manufacturing. In Ref. 121, SPSA has been applied for automatic controller tuning for electrohydraulic valves. In Ref. 122, SPSA has been applied in the modeling of batch animal cell cultures. In Ref. 123, a “grid infrastructure” is used for determining the optimal locations of oil wells.

The applications described above, apart from being interesting in themselves, also illustrate the fact that SPSA and SPSA-type algorithms are widely used stochastic optimization procedures that have been successfully applied in a variety of disciplines.

REFERENCES

1. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22:400–407.
2. Kiefer E, Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Ann Math Stat* 1952;23:462–466.
3. Spall JC. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans Auto Contr* 1992;37(3):332–341.
4. Spall JC. Introduction to stochastic search and optimization. New York: John Wiley & Sons, Inc.; 2003.
5. Spall JC. Stochastic optimization, stochastic approximation and simulated annealing. In: Webster JG, editor. Volume 20, Wiley Encyclopedia of Electrical and Electronics Engineering. New York: John Wiley & Sons, Inc.; 1999. pp. 529–542.
6. Spall JC. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins APL Tech Dig* 1998;19:482–492.
7. Kirkpatrick S, Gelatt CD, Vecchi M. Optimization by simulated annealing. *Science* 1983;220:621–680.
8. Gelfand SB, Mitter SK. Recursive stochastic algorithms for global optimization in $\mathcal{R}^{d*}$. *SIAM J Contr Optim* 1991;29(5):999–1018.
9. Bertsekas DP. Nonlinear programming. Belmont (MA): Athena Scientific; 1999.
10. Bertsekas DP, Tsitsiklis JN. Parallel and distributed computation. New Jersey: Prentice Hall; 1989.
11. Bhatnagar S. Adaptive multivariate three-timescale stochastic approximation algorithms for simulation based optimization. *ACM Trans Model Comput Simul* 2005;15(1):74–107.
12. Borkar VS. Stochastic approximation: a dynamical systems viewpoint. New Delhi: Hindustan Book Agency; 2008.
13. Kushner HJ, Yin GG. Stochastic approximation algorithms and applications. New York: Springer; 1997.
14. Chow YS, Teicher H. Probability theory: independence, interchangeability and martingales. 3rd ed. New York: Springer; 1997.
15. Borkar VS, Meyn SP. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J Contr Optim* 2000;38(2):447–469.

16. Dai JG. On positive Harris recurrence for multiclass queueing networks: a unified approach via fluid limit models. *Ann Appl Probab* 1995;5:49–77.
17. Dai JG, Meyn SP. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans Automat Contr* 1995;40:1889–1904.
18. Chen HF, Duncan TE, Pasik-Duncan B. A Kiefer–Wolfowitz algorithm with randomized differences. *IEEE Trans Automat Contr* 1999;44(3):442–453.
19. Ljung L. Analysis of recursive stochastic algorithms. *IEEE Trans Automat Contr* 1977;AC-22:551–575.
20. Kushner HJ, Clark DS. Stochastic approximation methods for constrained and unconstrained systems. New York: Springer; 1978.
21. Benveniste A, Métivier M, Priouret P. Adaptive algorithms and stochastic approximations. Berlin: Springer; 1990.
22. Benaim M. A dynamical systems approach to stochastic approximations. *SIAM J Contr Optim* 1996;34(2):437–472.
23. Chong EKP, Ramadge PJ. Optimization of queues using an infinitesimal perturbation analysis-based stochastic algorithm with general update times. *SIAM J Contr Optim* 1993;31(3):698–732.
24. Chong EKP, Ramadge PJ. Stochastic optimization of regenerative systems using infinitesimal perturbation analysis. *IEEE Trans Automat Contr* 1994;39(7):1400–1410.
25. Fu MC. Convergence of a stochastic approximation algorithm for the $GI/G/1$ queue using infinitesimal perturbation analysis. *J Optim Theor Appl* 1990;65:149–160.
26. Ho YC, Cao XR. Perturbation analysis of discrete event dynamical systems. Boston (MA): Kluwer; 1991.
27. Cassandras CG. Discrete event systems: modeling and performance analysis. Boston (MA): Aksen Associates; 1993.
28. Bhatnagar S, Borkar VS. Multiscale stochastic approximation for parametric optimization of hidden Markov models. *Probab Eng Inform Sci* 1997;11:509–522.
29. Bhatnagar S, Borkar VS. A two timescale stochastic approximation scheme for simulation based parametric optimization. *Probab Eng Inform Sci* 1998;12:519–531.
30. Bhatnagar S, Fu MC, Marcus SI, *et al.* Two timescale algorithms for simulation optimization of hidden Markov models. *IIE Trans* 2001;33(3):245–258.
31. L'Ecuyer P, Glynn PW. Stochastic optimization by simulation: convergence proofs for the $GI/G/1$ queue in steady-state. *Manag Sci* 1994;40(11):1562–1578.
32. Kleinman NL, Spall JC, Naiman DQ. Simulation-based optimization with stochastic approximation using common random numbers. *Manag Sci* 1999;45:1570–1578.
33. Fu MC, Hill SD. Optimization of discrete event systems via simultaneous perturbation stochastic approximation. *IIE Trans* 1997;29(3):233–243.
34. Bhatnagar S. Adaptive Newton-based smoothed functional algorithms for simulation optimization. *ACM Trans Model Comput Simul* 2007;18(1):2:1–2:35.
35. Sadegh P, Spall JC. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. In: *Proceedings of the American Control Conference*. Albuquerque (NM): 1997. pp. 3582–3586.
36. Chin DC. Comparative study of stochastic algorithms for system optimization based on gradient approximation. *IEEE Trans Syst Man Cybern Part B* 1997;27(2):244–249.
37. Fabian V. Stochastic approximation. In: Rustagi JJ, editor. *Optimizing methods in statistics*. New York: Academic Press; 1971. pp. 439–470.
38. Spall JC. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans Aerosp Electron Syst* 1998;34:817–823.
39. Spall JC. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica* 1997;33:109–112.
40. Bhatnagar S, Fu MC, Marcus SI, *et al.* Wang. Two-timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Trans Model Comput Simul* 2003;13(2): 180–209.
41. Poyiadjis G, Singh SS, Doucet A. Gradient-free maximum likelihood parameter estimation with particle filters. In: *Proceedings of the American Control Conference*. Minneapolis (MN); 2006. pp. 3052–3067.
42. Bartkute V, Sakalauskas L. Application of stochastic approximation in technical design. *Eur J Oper Res* 2007;181(3):1174–1188.
43. He Y, Fu MC, Marcus SI. Convergence of simultaneous perturbation stochastic

- approximation for nondifferentiable optimization. *IEEE Trans Automat Contr* 2003;48:1459–1463.
44. Gerencsér L. Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans Automat Contr* 1999;44(5):894–906.
 45. Abdulla MS, Bhatnagar S. SPSA with measurement reuse. In: *Proceedings of Winter Simulation Conference*. Monterey (CA); 2006. pp. 320–328.
 46. Maeda Y. Time difference simultaneous perturbation method. *Electron Lett* 1996; 32:1016–1018.
 47. Maryak JL. Some guidelines for using iterate averaging in stochastic approximation. In: *Proceedings of the IEEE Conference on Decision and Control*. San Diego (CA); 1997; pp. 2287–2290.
 48. Dippon J, Renz J. Weighted means in stochastic approximation of minima. *SIAM J Contr Optim* 1997;35:1811–1827.
 49. Chin DC. A more efficient global optimization algorithm based on Styblinski and Tang. *Neural Netw* 1994;7:573–574.
 50. Maryak JL, Chin DC. Global random optimization by simultaneous perturbation stochastic approximation. *IEEE Trans Automat Contr* 2008;53:780–783.
 51. Borkar VS. Stochastic approximation with two timescales. *Syst Contr Lett* 1997;29:291–294.
 52. Polyak BT, Juditsky AB. Acceleration of stochastic approximation by averaging. *SIAM J Contr Optim* 1992;30(4):838–855.
 53. Spall JC, Cristion JA. Nonlinear adaptive control using neural networks: estimation with a smoothed form of simultaneous perturbation gradient approximation. *Stat Sin* 1994;4:1–27.
 54. Renotte C, Wouwer Vande A, Remy M. Neural modeling and control of a heat exchanger based on SPSA techniques. In: *Proceedings of the American Control Conference*. Chicago (IL); 2000. pp. 3299–3303.
 55. Wouwer Vande A, Renotte C, Remy M. Application of stochastic approximation techniques in neural modelling and control. *Int J Syst Sci* 2003;34:851–863.
 56. Bhatnagar S, Borkar VS. Multiscale chaotic SPSA and smoothed functional algorithms for simulation optimization. *Simulation* 2003;79(10):568–580.
 57. Bhatnagar S, Fu MC, Marcus SI, *et al.* Optimal structured feedback policies for ABR flow control using two-timescale SPSA. *IEEE/ACM Trans Netw* 2001;9(4):479–491.
 58. Wang I-J, Chong EKP. A deterministic analysis of stochastic approximation with randomized directions. *IEEE Trans Automat Contr* 1998;43(12):1745–1749.
 59. Xiong X, Wang I-J, Fu MC. An asymptotic analysis of stochastic approximation with deterministic perturbation sequences. In: *Proceedings of the Winter Simulation Conference*. San Diego (CA); 2002. pp. 285–291.
 60. Bhatnagar S, Borkar VS, Madhukar A. A simulation based algorithm for ergodic control of Markov chains conditioned on rare events. *J Mach Learn Res* 2006;7: 1937–1962.
 61. Patro RK, Bhatnagar S. A probabilistic constrained nonlinear optimization framework to optimize RED parameters. *Perform Eval* 2009;66(2):81–104.
 62. Chinthalapati VLR, Bhatnagar S. A simultaneous deterministic perturbation actor–critic algorithm with an application to optimal mortgage refinancing. In: *Proceedings of the IEEE Conference on Decision and Contr*. San Diego (CA); 2006. pp. 4151–4156.
 63. Ruppert D. A Newton–Raphson version of the multivariate Robbins–Monro procedure. *Ann Stat* 1985;13:236–245.
 64. Spall JC. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans Automat Contr* 2000;45:1839–1853.
 65. Zhu X, Spall JC. A modified second-order SPSA optimization algorithm for finite samples. *Int J Adapt Contr Signal Process* 2002;16:397–409.
 66. Spall JC. Feedback and weighting mechanisms for improving Jacobian estimates in the adaptive simultaneous perturbation algorithm. *IEEE Trans Automat Contr* 2009;54(6):1216–1229.
 67. Katkovnik VY, Kulchitsky Y. Convergence of a class of random search algorithms. *Automat Remote Contr* 1972;8:1321–1326.
 68. Styblinski MA, Tang T-S. Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing. *Neural Netw* 1990;3: 467–483.
 69. Bhatnagar S, Karmeshu, Mishra VK. Optimal parameter trajectory estimation in

- parameterized SDEs: an algorithmic procedure. *ACM Trans Model Comput Simul* 2009;19(2):8:1–8:27.
70. Mishra V, Bhatnagar S, Hemachandra N. Discrete parameter simulation optimization algorithms with applications to admission control with dependent service times. In: *Proceedings of the IEEE Conference on Decision and Control*. New Orleans (LA); 2007. pp. 2986–2991.
 71. Sadegh P. Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation. *Automatica* 1997;33:889–892.
 72. Wang I-J, Spall JC. Stochastic optimization with inequality constraints using simultaneous perturbations and penalty functions. *Int J Contr* 2008;81(8):1232–1238.
 73. Whitney JE, Hill SD. Constrained optimization over discrete sets via SPSSA with application to non-separable resource allocation. In: *Proceedings of the Winter Simulation Conference*. Arlington (VA); 2001. pp. 313–317.
 74. Bhatnagar S, Hemachandra N, Mishra V. Stochastic approximation algorithms for constrained optimization via simulation. *ACM Trans Model Comput Simul* 2009. In press.
 75. Spall JC. Selected references on theory, applications and numerical analysis. Available at URL <http://www.jhuapl.edu/SPSA/>.
 76. Bhatnagar S, Reddy IBB. Optimal threshold policies in communication networks via discrete parameter stochastic approximation. *Telecommun Syst* 2005;29:9–31.
 77. Bhatnagar S, Kowshik HJ. A discrete parameter stochastic approximation algorithm for simulation optimization. *Simulation* 2005;81(11):757–772.
 78. Gerencsér L, Hill SD, Vágó Z. Optimization over discrete sets via SPSSA. *Proceedings of the IEEE Conference on Decision and Control*. Phoenix (AZ): IEEE press; 1999. pp. 1791–1795.
 79. Gerencsér L, Hill SD, Vágó Z. Discrete optimization via SPSSA. In: *Proceedings of the American Control Conference*. Arlington (VA); 2001. pp. 1503–1504.
 80. Hill SD, Gerencsér L, Vágó Z. Stochastic approximation on discrete sets using simultaneous difference approximations. In: *Proceedings of the American Control Conference*. Boston (MA); 2004. pp. 2795–2798.
 81. Maeda Y, Maruyama T. Natural gradient using simultaneous perturbation without probability densities for blind source separation. In: *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*. Nara, Japan; 2003. pp. 439–443.
 82. Spall JC. Monte Carlo computation of the Fisher information matrix in non-standard settings. *J Comput Graph Stat* 2005;14(4):889–909.
 83. Das S, Spall JC, Ghanem R. An efficient calculation of Fisher information matrix: Monte Carlo approach using prior information. In: *Proceedings of the IEEE Conference on Decision and Control*. New Orleans (LA); 2007. pp. 963–968.
 84. Bhatnagar S, Kumar S. A simultaneous perturbation stochastic approximation based actor–critic algorithm for Markov decision processes. *IEEE Trans Automat Contr* 2004;49(4):592–598.
 85. Abdulla MS, Bhatnagar S. Reinforcement learning based algorithms for average cost Markov decision processes. *Discrete Event Dyn Syst* 2007;17(1):23–52.
 86. Bhatnagar S, Abdulla MS. Simulation-based optimization algorithms for finite horizon Markov decision processes. *Simulation* 2008;84(12):577–600.
 87. Abdulla MS, Bhatnagar S. Parameterized actor–critic algorithms for finite-horizon MDPs. In: *Proceedings of the American Control Conference*. New York; 2007. pp. 534–539.
 88. Bhatnagar S, Babu K. New algorithms of the Q-learning type. *Automatica* 2008;44(4):1111–1119.
 89. Bhatnagar S, Panigrahi JR. Actor–critic algorithms for hierarchical Markov decision processes. *Automatica* 2006;42(4):637–644.
 90. Spall JC, Cristion JA. Model-free control of nonlinear stochastic systems with discrete-time measurements. *IEEE Trans Automat Contr* 1998;43(9):1198–1210.
 91. Yin G, Zhang Q, Yan HM, *et al.* Random direction optimization algorithms with applications to threshold controls. *J Optim Theor Appl* 2001;110:211–233.
 92. Vaidya R, Bhatnagar S. Robust optimization of random early detection. *Telecommun Syst* 2006;33(4):291–316.
 93. Reddy GR, Bhatnagar S, Rakesh V, *et al.* An efficient algorithm for scheduling in blue-tooth piconets and scatternets. *Wirel netw.*

- DOI: 10.1007/s11276-009-0229-3, 2009. In press.
94. Guven T, La RJ, Shayman MA, *et al.* Measurement-based optimal routing on overlay architectures for unicast sessions. *Comput Netw* 2006;50(12):1938–1951.
 95. Sadegh P, Spall JC. Optimal sensor configuration for complex systems. In: *Proceedings of the American Control Conference*. Philadelphia (PA); 1998. pp. 3575–3579.
 96. Chin DC, Smith RH. A traffic simulation for mid-Manhattan with model-free adaptive signal control. In: *Proceedings of the Summer Computer Simulation Conference*. San Diego (CA); 1994. pp. 296–301.
 97. Maeda Y, Wakamura M. Simultaneous perturbation learning rule for recurrent neural networks and its FPGA implementation. *IEEE Trans Neural Netw* 2005; 16(6):1664–1672.
 98. Ni J, Song Q. Dynamic pruning algorithm for multilayer perceptron-based neural control systems. *Neurocomputing* 2006; 69:2097–2111.
 99. Song Q, Spall JC, Soh YC. Robust neural network tracking controller using simultaneous perturbation stochastic approximation. *IEEE Trans Neural Netw* 2008; 19(5):817–835.
 100. Spall JC, Cristion JA. A neural network controller for systems with unmodeled dynamics with applications to wastewater treatment. *IEEE Trans Syst Man Cybern B* 1997;27:369–375.
 101. Thangavel P, Kathirvalavakumar T. Simultaneous perturbation for single hidden layer networks – cascade learning. *Neurocomputing* 2003;50:193–209.
 102. Johannsen DA, Wegman EJ, Solka JL, *et al.* Simultaneous selection of features and metric for optimal nearest neighbor classification. *Commun Stat Theor Methods* 2004;33:2137–2157.
 103. Sudha V, Bhatnagar S, Basavaraja SV, *et al.* SPSA based feature relevance estimation for video retrieval. In: *Proceedings of IEEE Workshop on Multimedia Signal Processing*. Queensland, Australia; 2008. pp. 598–603.
 104. Maeda S, Nailon W, Durrani T. Fast and accurate image registration using Tsallis entropy and simultaneous perturbation stochastic approximation. *Electron Lett* 2004;40:595–597.
 105. Maryak JL, Spall JC. Simultaneous perturbation optimization for efficient image restoration. *IEEE Trans Aerosp Electron Syst* 2005;41:356–361.
 106. Chin DC. Efficient identification procedure for inversion processing. In: *Proceedings of the IEEE Conference on Decision and Control*. Kobe, Japan; 1996. pp. 3129–3130.
 107. Hutchison DW, Hill SD. Simulation optimization of airline delay with constraints. In: *Proceedings of the Winter Simulation Conference*. Arlington (VA); 2001. pp. 1017–1022.
 108. Kleinman NL, Hill SD, Ilenda VA. SPSA/SIMMOD optimization of air traffic delay cost. In: *Proceedings of the American Control Conference*. Albuquerque (NM); 1997. pp. 1121–1125.
 109. Xing XQ, Damodaran M. Application of simultaneous perturbation stochastic approximation method for aerodynamic shape design optimization. *AIAA J* 2005;43:284–294.
 110. Gosavi A, Ozkaya E, Kahraman A. Simulation optimization for revenue management of airlines with cancellations and overbooking. *OR Spectr* (special issue on revenue management) 2007;29:21–38.
 111. Spall JC, Chin DC. Traffic-responsive signal timing for system-wide traffic control. *Transp Res C* 1997;5:153–163.
 112. Koch MI, Chin DC, Smith RH. Network-wide approach to optimal signal light timing for integrated transit vehicle and traffic operations. Volume 2, *Proceedings of the 7th National Conference on Light Transit*. Baltimore: National Academy of Sciences Press; 1995. pp. 126–131.
 113. Srinivasan D, Choy MC, Cheu RL. Neural networks for real-time traffic signal control. *IEEE Trans Intell Transp Syst* 2006;7(3):261–272.
 114. Popovic D, Jankovic M, Magner S, *et al.* Extremum-seeking methods for optimization of variable cam timing engine operation. *IEEE Trans Contr Syst Techn* 2006;14(3): 398–407.
 115. Kocsis L, Szepesvári C. Universal parameter optimisation in games based on SPSA. *Mach Learn* 2006;63(3):249–286.
 116. Cauwenberghs G. An analog VLSI recurrent neural network learning a continuous-time trajectory. *IEEE Trans Neural Netw* 1996;7(2):346–361.
 117. Nusawardhana N, Zak SH. Simultaneous perturbation extremum-seeking method for

- dynamic optimization problem. In: Proceedings of the American Control Conference. Boston (MA); 2004. pp. 2805–2810.
118. Gerencsér L, Kozmann G, Vágó Z, *et al.* The use of SPSA method in ECG analysis. IEEE Trans Biomed Eng 2002;49: 1094–1101.
 119. Luman RR. Upgrading complex systems of systems: A CAIV methodology for warfare area requirements allocation. Mil Oper Res 2000;5(2):53–75.
 120. Schwartz JD, Wang W, Rivera DE. Simulation-based optimization of process control policies for inventory management in supply chains. Automatica 2006; 42(8):1311–1320.
 121. Yuan QH. A model-free automatic tuning method for a restricted structured controller by using simultaneous perturbation stochastic approximation (SPSA). In: Proceedings of the American Control Conference; 2008 11–13 June. Seattle (WA); 2008. pp. 1539–1545.
 122. Wouwer AV, Renotte C, Bogaerts PH. A short note on SPSA techniques and their use in nonlinear bioprocess identification. Math Comput Model Dyn Syst 2006; 12(5):415–422.
 123. Bangerth W, Klie H, Matossian V, *et al.* An autonomic reservoir framework for the stochastic optimization of well placement. Cluster Comput 2005;8:255–269.

SINGLE MACHINE SCHEDULING

NEIL GEISMAR

Department of Information and
Operations Management, Mays
Business School, Texas A&M
University, College Station, Texas

Instances of the single machine scheduling problem arise often in business and in everyday life. A person faced with several tasks, each of which must be completed by its *due date* (e.g., a student with several assignments), should schedule his/her time so that each is completed on time or with the least possible lateness for each. In scheduling theory, this objective is called *minimizing maximum lateness*. If his/her instructors do not accept assignments late, then the due dates are called *deadlines* and his/her objective changes to *minimizing the weighted number of tardy jobs*, where each task has a *weight* that represents its value, in this case its contribution to the final grade. Companies typically have multiple projects to deliver to customers with contractually stated due dates and penalties assessed for each time unit (e.g., day) by which the job is late. The scheduler of such a company seeks to *minimize total weighted tardiness*. Each project's weight may signify its daily penalty or be a subjective measure of the value of the customer. A supplier providing components for a just-in-time manufacturer (described in ***Just-in-Time/Lean Production Systems***) must minimize both earliness and tardiness of his deliveries.

The single machine scheduling problem is foundational to the study of production scheduling in general because results for a single machine lead to insights into all other scheduling environments. Many exact and heuristic solutions to problems with parallel machines or multiple processing stages (e.g., job shops, flow shops, open shops) are derived by decomposing the system into separate single machines. Furthermore, because it is a special case of the multimachine models (see

the article titled ***Job Shop Scheduling***, and the section titled "Supply Chain Scheduling" in this encyclopedia), if scheduling a single-machine system is shown to be NP-hard [1] for a given objective, then the general scheduling problem for that objective is NP-hard. NP-hard problems are described in depth in the section titled "Network Flows," in this encyclopedia. For this article's purposes, it suffices to know that, given an NP-hard problem, it is highly unlikely that an efficient algorithm—stated precisely, one whose running time is upper bounded by a polynomial in the size of the input for the algorithm—can be developed that will solve a general instance of that problem.

This article summarizes results in the single-machine environment that provide an introduction to the field. These form a basis for deeper study of the topic. The first sections consider the regular performance measurements: the sections titled "Total Weighted Completion Time," "Maximum Lateness," "Maximum of Nondecreasing Cost Functions with Precedence Constraints," "Weighted Number of Tardy Jobs," and "Total Weighted Tardiness." In this context, *regular* means that all jobs are available when processing begins (often called *time zero*) and that the objective function is nondecreasing in the finishing times of the jobs. For such problems, we need to consider schedules without preemption and without idle time [2]. A *preemption* is the interruption of a job's processing to run other jobs on the machine, followed by returning the interrupted job to the machine for its remaining processing time; the preempted job's previous processing is not lost. Excluding preemption and idle time implies that a solution consists simply of a *sequence* of jobs, since each job begins without delay after its immediate predecessor concludes. Later sections describe problems in which both earliness and tardiness are minimized (see the section titled "Earliness and Tardiness Problems") and those in which some jobs are not available when processing

begins (nonzero release dates, defined below) and that allow preemptions (see the section titled “Release Dates and Preemptions.”)

The standard Makespan (completion time of last job) objective is generally not considered in this environment because its value is simply the sum of all processing times, independent of the order in which jobs are performed. One exception is the Makespan with sequence-dependent setup times problem, which can be solved by the Gilmore–Gomory [3] algorithm for the traveling salesman problem (see **Combinatorial Traveling Salesman Problem Algorithms**).

Notation used herein follows that traditionally used in the scheduling literature [4]. The parameters describe characteristics of the jobs to be processed:

- n : number of jobs to be scheduled on the machine
- p_j : processing time of job j
- w_j : weight (value) of job j
- d_j : due date of job j
- r_j : release date of job j . This is the earliest date at which work can begin on job j . By default, $r_j = 0, \forall j$. How nonzero values for r_j may complicate a problem is demonstrated in the section titled “Maximum Lateness.”

The term *date* is used generically in the scheduling literature. It may indicate any point in time, be it a day, an hour, a minute, and so on.

The problems’ variables are performance measurement criteria that are used to compute the objective values of the schedules:

- C_j : completion time of job j
- L_j : lateness of job j : $L_j = C_j - d_j$
- T_j : tardiness of job j : $T_j = \max\{0, C_j - d_j\}$
- U_j : indicator variable for tardiness of job j :
 $U_j = 1$, if $C_j > d_j$; $U_j = 0$, otherwise

We shall use the traditional scheduling classification scheme of Graham *et al.* [5] to describe each problem in the form $\alpha|\beta|\gamma$. The first field indicates the machine environment; hence, $\alpha = 1$ throughout this article. The second field describes processing constraints or

special characteristics. For example, $\beta = r_j$ indicates that the jobs have nonzero release dates, $\beta = s_{jk}$ denotes sequence-dependent setup times, $\beta = prmp$ implies job preemption is allowed, and $\beta = prec$ means precedence constraints apply to some of the jobs. Note that this field may be empty or may contain multiple descriptors. The γ field contains the objective to be minimized. Examples include $\sum w_j C_j$, $L_{\max}$, $\sum w_j U_j$, and $\sum w_j T_j$.

TOTAL WEIGHTED COMPLETION TIME

The total completion time objective provides a measure of work-in-process inventory (see the section titled “Inventory Management and Control” in this encyclopedia). The weights are used to assign different values to the jobs; a weight may represent the actual revenue derived from the job, the holding cost of the job’s inventory, or the importance of its customer. In the scheduling classification scheme [5], this problem is represented by $1||\sum w_j C_j$. The sequence that minimizes this objective is the *weighted shortest processing time (WSPT)* rule [6], which schedules jobs in nondecreasing order of p_j/w_j . Because the method used to prove this result—adjacent pairwise interchange—is commonly used in single machine scheduling research, we include the proof of this result.

Theorem 1. *The WSPT rule is optimal for problem $1||\sum w_j C_j$.*

Proof. Given a set of n jobs, suppose there is some optimal sequence S that is not in WSPT order. There must be two consecutive jobs i then k such that $p_i/w_i > p_k/w_k$. Define T to be the start time of job i and C_j to be the completion time of each job $j = 1, \dots, n$, for sequence S .

Let sequence S' be identical to S except that i and k are reversed. Let C'_j be the completion time of each job $j = 1, \dots, n$, for sequence S' . Obviously, job k starts at time T in sequence S' . Therefore, since every job except i and k complete at the same time in each sequence, the difference in the total weighted completion times for the two

sequences is

$$\begin{aligned}
& \sum w_j C_j - \sum w_j C'_j \\
&= w_i C_i + w_k C_k - (w_i C'_i + w_k C'_k) \\
&= 2T + w_i p_i + w_k (p_i + p_k) \\
&\quad - (2T + w_k p_k + w_i (p_i + p_k)) \\
&= w_k p_i - w_i p_k > 0,
\end{aligned}$$

by the assumption that $p_i/w_i > p_k/w_k$. Hence, sequence S cannot be optimal.

The total completion time problem with release dates ($1|r_j|\sum C_j$) is NP-hard in the strong sense [7]. However, with preemption, the problem ($1|r_j,prmp|\sum C_j$) is solvable by the shortest remaining processing time (*SRPT*) rule [8]. Though the *SRPT* result is generally not feasible for the nonpreemptive problem, it has been used as an effective lower bound for branch-and-bound approaches to this problem [9,10].

It follows from the preceding paragraph that the total weighted completion time problem with release dates ($1|r_j|\sum w_j C_j$) is NP-hard in the strong sense, too. However, a strong lower bound cannot be found by allowing preemption in this case because the corresponding preemptive problem ($1|r_j,prmp|\sum w_j C_j$) is also NP-hard [11]. Nevertheless, effective branch-and-bound schemes have been developed for this problem [12,13].

MAXIMUM OF NONDECREASING COST FUNCTIONS OF C_j

Many due-date related objectives for single machine problems can be expressed as

$$\begin{aligned}
& \text{Minimize } f_{\max} \\
&= \max \{f_1(C_1), f_2(C_2), \dots, f_n(C_n)\},
\end{aligned}$$

where $f_j(C_j)$ is a nondecreasing cost function, $j = 1, \dots, n$. We first look at a specific case—the section titled “Maximum Lateness”—and then at a problem with a general objective of this form that has precedence constraints (see the section titled “Maximum of Nondecreasing Cost Functions with Precedence Constraints”).

Maximum Lateness

Maximum lateness is a measure of the service level. In fact, it measures the satisfaction of the customer who receives the *worst* service. Unsurprisingly, the maximum lateness is minimized by the *earliest due date (EDD)* rule [14], which sequences jobs in nondecreasing order of d_j .

Theorem 2. *The EDD rule is optimal for problem $1||L_{\max}$.*

To illustrate how minor changes to a problem can significantly affect its solution, note that the maximum lateness problem with jobs released at different times ($1|r_j|L_{\max}$) is NP-hard in the strong sense [7].

There are special cases of problem $1|r_j|L_{\max}$ that have polynomial solutions:

- If all release dates are the same, $r_j = r$ for $j = 1, \dots, n$, then the problem reduces to $1||L_{\max}$, which is solved by the *EDD* rule.
- If all due dates are the same, $d_j = d$ for $j = 1, \dots, n$, then an optimal schedule can be found by scheduling the jobs in nondecreasing order of release dates.
- If all jobs have unit processing time, $p_j = 1$ for $j = 1, \dots, n$, then as each job completes, schedule the available job with the smallest due date [15].

Each of these special cases and Theorem 2 can be proven by using adjacent pairwise interchange.

Maximum of Nondecreasing Cost Functions with Precedence Constraints

A precedence constraint requires that a job or set of jobs be completed before a certain job begins. Such constraints are common in scheduling problems. Examples include construction (plumbing and electrical work must be completed before the walls are finished) and curriculum design (an advanced course requires certain prerequisites). The single machine problem of minimizing the maximum cost function if the jobs have arbitrary precedence constraints is denoted $1|prec|f_{\max}$.

Lawler's algorithm [16] to find an optimal sequence uses backward dynamic programming (see the articles in the section titled "Dynamic Programming" of this encyclopedia). Let J be the set of jobs that have not been scheduled and $J' \subseteq J$ be the set of jobs that are currently eligible to be scheduled,

that is, J' is the set of jobs that have no successors currently in J . By starting at the sequence's completion time $C_{\max} = \sum_{j=1}^n p_j$ and working backward, the algorithm schedules in each iteration a job that will end its processing at time $\sum_{j \in J} p_j$. The job so chosen is $j^* = \operatorname{argmin}_{i \in J'} \{f_i(\sum_{j \in J} p_j)\}$.

Algorithm L1973

Step 1: Let $J = \{1, \dots, n\}$, J' be the set of all jobs with no successors in J , and $T = \sum_{j=1}^n p_j$.

Step 2: Let $j^* = \operatorname{argmin}_{i \in J'} \{f_i(\sum_{j \in J} p_j)\}$.

Step 3: Schedule job j^* to be processed during interval $[T - p_{j^*}, T]$.

Step 4: Let $T = T - p_{j^*}$ and $J = J \setminus \{j^*\}$. Update J' by adding those jobs which may now be scheduled, that is, those for which j^* was the only successor.

Step 5: If $J = \emptyset$, then STOP. Otherwise, go to Step 2.

Theorem 3. *Algorithm L1973 produces an optimal schedule for problem 1|prec| $f_{\max}$.*

later in σ than in σ' , and $f_k(T_k) < f_{k'}(T_k)$. Hence, σ' cannot be optimal.

Proof. Suppose σ' is an optimal sequence in which job k' concludes processing at time T_k , that $f_k(T_k) < f_{k'}(T_k)$, and that job $k \in J'$ during the iteration in which job k' was scheduled. Figure 1 displays the order in which the jobs are processed in sequence σ' ; each block represents the time used for its job. A and B are (possibly empty) collections of jobs.

Now suppose that sequence σ' is modified to create sequence σ by scheduling job k during interval $[T_k - p_k, T_k]$ and shifting backward by p_k time units all those jobs that were between jobs k and k' , including k' ; See Fig. 2.

If σ' observes the precedence constraints, then so does σ . Furthermore, $f_{\max}(\sigma) \leq f_{\max}(\sigma')$: job k is the only job that finishes



Figure 1. Sequence σ' for the proof of Theorem 3.



Figure 2. Sequence σ for proof of Theorem 3.

WEIGHTED NUMBER OF TARDY JOBS

The weighted number of tardy jobs metric is commonly used to evaluate managers as it, too, measures the service level (it is equivalent to the weighted number of on time completions) and is easy to track. Recall that a job's weight represents its value to the company (e.g., penalty paid, importance of customer), so this objective quantifies part of a manager's contribution to revenue. The general weighted number of tardy jobs problem is NP-hard [17]; however, we present a polynomial algorithm that solves the unweighted case and a pseudopolynomial algorithm that solves the general problem.

The algorithm of Moore [18] optimally solves the unweighted problem 1|| $\sum U_j$. It begins with all jobs in *EDD* order; without loss of generality, we assume that *EDD* order is $1, 2, \dots, n$. In iteration j , $j = 1, \dots, n$, job j is appended to the ordered set of on-time jobs G^e . If job j 's calculated completion time is larger than its due date ($C_j > d_j$), then the job in G^e with the largest processing time ($\operatorname{argmax}_{k \in G^e} \{p_k\}$) is moved to the set of tardy jobs G^t . This process is continued until all

Algorithm M1968

Input: ordered job set $G = \{1, 2, \dots, n\}$ in *EDD* order, $p_j, d_j, j = 1, 2, \dots, n$

Output: G^e is the ordered set of early jobs, G^ℓ is the set of late jobs

Step 1: Set $G^e = \{1\}$, $G^\ell = \emptyset$, $j = 1$, and $C_1 = p_1$.

Step 2: If $C_j > d_j$, then

Move job $q = \operatorname{argmax}_{k \in G^e} \{p_k\}$ from set G^e to set G^ℓ

Let $C_r = \sum_{i \in G^e} \{p_i\}$

Step 3: Set $j = j + 1$.

If $j \leq n$ then

Set $G^e = G^e \cup \{j\}$

$C_j = C_r + p_j$

Go to Step 2

Step 4: Output schedule $G^e \cup G^\ell$. STOP.

A pseudopolynomial dynamic programming algorithm solves problem $1||\sum w_j U_j$ [19] via recursion. As above, it progresses through jobs $j = 1, 2, \dots, n$, in *EDD* order. For each job j , the algorithm finds the minimum weighted number of tardy jobs incurred when sequencing jobs $\{1, \dots, j\}$ such that job j is completed no later than time $t = p_j, \dots, P_j$, where $P_j = \sum_{k=1}^j p_k$; this quantity is denoted

$F_j(t)$. $F_j(t)$ is calculated as follows: if job j has been scheduled to complete before time t , then $F_j(t) = F_j(t - 1)$. If the objective is minimized by having job j added to the current schedule so that it completes at time t , then $F_j(t) = F_{j-1}(t - p_j)$. Otherwise, if $F_{j-1}(t) + w_j < F_{j-1}(t - p_j)$, then job j is designated as a tardy job. The optimum is $F_n(P_n)$:

Algorithm LM1969

Input: A set of n jobs arranged in *EDD* order, their processing times p_j , their weights w_j , and their due dates d_j .

Initial conditions: $F_0(t) = 0, t \geq 0$; $F_j(t) = \infty, t < 0, j = 1, \dots, n$.

Recursive relation:

$$F_j(t) = \min \begin{cases} F_j(t - 1) \\ F_{j-1}(t - p_j) \\ F_{j-1}(t) + w_j \end{cases}$$

for $j = 1, \dots, n; t = p_j, \dots, P_j$.

Optimum function value: $F_n(P_n)$.

This algorithm runs in $O(nP_n)$ time.

TOTAL WEIGHTED TARDINESS

The total weighted tardiness objective serves as a valuable complement to the number of tardy jobs objective. Monitoring the total weighted tardiness can ensure that a small

number of tardy jobs is not achieved by having those jobs that are tardy be unacceptably late. The total tardiness problem ($1||\sum T_j$) is NP-hard [20], and the total weighted tardiness problem ($1||\sum w_j T_j$) is NP-hard in the strong sense [21].

The pseudopolynomial algorithm that solves the unweighted total tardiness problem [21] also solves to problem 1|| $\sum w_j T_j$ if there are agreeable weights. A problem is said to have *agreeable weights* if $p_i < p_j$ implies $w_i \geq w_j$.

EARLINESS AND TARDINESS PROBLEMS

There are many circumstances in which schedulers wish to avoid both earliness and tardiness. Obviously tardiness may reduce efficiency and customer goodwill, but earliness can lead to unnecessary expense through the accumulation of inventory. Either type of deviation can disrupt a just-in-time manufacturer [22]. Computer systems may want to minimize the range of retrieval times for records in a file in order to provide uniform response to users' requests [23]. Another example is a chemical manufacturer that produces two rapidly deteriorating chemicals that are combined to form a final product. Earliness or lateness

in production of either of these components could diminish productivity [24]. We now consider two approaches to this problem: one minimizes the average deviation of completion times about a common due date, the other minimizes the maximum penalty incurred by any job either for its early start or tardy completion.

Any sequence that minimizes the average deviation of completion times for a set of jobs with a common due date must be V-shaped [25]. Intuitively, the jobs in such a sequence are arranged so that the processing times decrease, reach a minimum, and then increase. Formally, a sequence $\{j_1, j_2, \dots, j_n\}$ is *V-shaped* if there exists some integer $k \in [1, n]$ such that $p_{j_1} \geq \dots \geq p_{j_k} \leq \dots \leq p_{j_n}$. The following algorithm [26] finds an optimal V-shaped sequence by ordering the jobs in nonincreasing order of their processing times then alternatively assigns them to two subsets: those processed before the due date (B), and those processed after the due date (A).

Algorithm K1981

Input: ordered job set $U = \{1, 2, \dots, n\}$ in *longest processing time first* order, common due date d

Output: A sequence that minimizes the average deviation of completion times

$B = A = \emptyset$

For $j = 1$ to n , Step 2

 Assign job j to the last position in B

 Remove j from U

 If $U \neq \emptyset$ then

 Assign job $j + 1$ to the first position in A

 Remove $j + 1$ from U

 End If

Next j

Schedule B so that its last job completes at time d

Schedule A so that its first job starts at time d

Output schedule $B \cup A$. STOP.

A related problem type minimizes the maximum penalty for either earliness or lateness [24]. Each job j has its own target start time (a_j) and target finish time (b_j); no penalty is imposed if the job's processing occurs within the interval

$[a_j, b_j]$. A common example of such a system would be a project described by a critical path method (CPM) network in which each job is assigned an earliest start date (a_j) and a latest finish date (b_j).

A job's earliness is defined by $E_j = \max\{0, a_j - s_j\}$, where s_j is job j 's starting time. Similarly, job j 's lateness is $L_j = \max\{0, s_j + p_j - b_j\}$. The penalty functions for earliness and lateness, respectively, are $g(E_j)$ and $h(L_j)$; they need not be equal, but each is continuous and non-decreasing, with $g(0) = h(0) = 0$. Thus, the objective to be minimized is $z = \max\{g(\max\{E_j\}), h(\max\{L_j\})\}$.

We assume that $a_i < a_j$ implies that $b_i \leq b_j$, which is common for service providers who guarantee a fixed duration for job completion. It follows that an optimal schedule will order jobs so that $a_1 \leq a_2 \leq \dots \leq a_n$ and $b_1 \leq b_2 \leq \dots \leq b_n$; hence, the main task is to determine each job's starting time.

A maximal lower bound on the sum $E_i + L_j$ over all pairs $i < j$ is used to find the minimum objective value. The easily derived lower bound is

$$E_i + L_j \geq \sum_{k=i}^j p_k - (b_j - a_i).$$

It is demonstrated by Fig. 3.

It can be deduced that there is an optimal schedule whose objective value is $g(E^*)$, where E^* and T^* are nonnegative numbers that satisfy the following simultaneous system:

$$\begin{aligned} E^* + T^* &= \max_{i < j} \left\{ \sum_{k=i}^j p_k - (b_j - a_i) \right\} \\ g(E^*) &= h(T^*) \end{aligned}$$

The first equation guarantees that the lower bound on $E_i + L_j$ is feasible, and the second minimizes the objective for this sum. Once E^* is found, the jobs' starting times can be derived as follows:

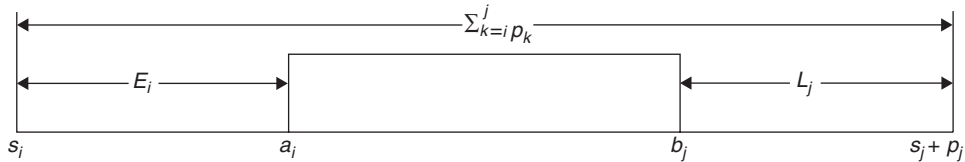


Figure 3. Demonstration of lower bound for $E_i + L_i$.

Table 1. Example Data for $L_{\max}$ Problem with Nonzero Release Dates and Preemption

	a	b
p_j	5	8
r_j	0	2
d_j	12	10

1. $s_1 = a_1 - E^*$
2. For $j = 2$ to n : $s_j = \max\{s_{j-1} + p_{j-1}, a_j - E^*\}$.

RELEASE DATES AND PREEMPTIONS

The introduction of nonzero release dates implies that an optimal schedule may have idle time. Allowing job preemptions may reduce this idle time and, therefore, the objective value. This ability provides the scheduler with much flexibility. Consider the simple example of $1|r_j, prmp|L_{\max}$ with two jobs given in Table 1.

The standard *EDD* rule would yield the schedule in Fig. 4 with $L_{\max} = 3$.

The *preemptive earliest due date* rule [27] yields the optimal schedule in Fig. 5 with $L_{\max} = 1$.

Now consider the minimization of the maximum of nondecreasing cost functions of C_j , $j = 1, \dots, n$, in this environment, both with and without precedence constraints. Recall that these functions are denoted $f_j(C_j)$ and that this problem with all release dates equal—in which case there is no advantage to preemption—was solved in the section titled “Maximum of Nondecreasing Cost Functions of C_j .” We first solve the general problem then address the minimization of the weighted number of tardy jobs, which

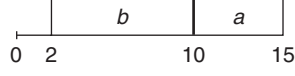


Figure 4. EDD schedule for data in Table 1.

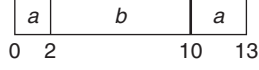


Figure 5. Preemptive EDD schedule for data in Table 1.

was first discussed in the section titled “Weighted Number of Tardy Jobs.”

The concept of *blocks* of jobs is useful for scheduling problems with release dates and preemptions. We first define notation for the total processing time for all jobs within a block B : $p(B) = \sum_{j \in B} p_j$. A *block* B is a minimal set of jobs processed without idle time from $r(B) = \min_{j \in B} \{r_j\}$ until $t(B) = r(B) + p(B)$, such that each job $j \notin B$ is either completed no later than $r(B)$ or not released before $t(B)$.

Given a set N of jobs with no precedence constraints, their release dates r_j , and their processing times p_j , an optimal schedule for the minimization of the maximum of nondecreasing cost functions $(1|r_j, prmp, |f_{\max})$ can be computed as follows [27]:

1. Order the jobs so they are nondecreasing in r_j .
2. Determine the initial block structure by scheduling them in this order, that is, determine where idle time will be imposed by the release dates.
3. For each block B , define $\ell = \operatorname{argmin}_{j \in B} \{f_j(t(B))\}$. The optimal schedule will have $C_\ell = t(B)$, but job ℓ may start earlier than $t(B) - p_\ell$, that is, it may be preempted.
4. Determine the block structure of $B \setminus \{\ell\}$ as in Step 2. Job ℓ will be processed in the times not occupied by $B \setminus \{\ell\}$, that is, if $B \setminus \{\ell\}$ is composed of intervals (subblocks) $B_1, \dots, B_b$, then job ℓ will be processed during the intervals $[r(B), t(B)] - \cup_{i=1}^b [r(B_i), t(B_i)]$.
5. Repeat this process for each subblock.

This algorithm can easily be modified to allow for precedence constraints. Let $j \rightarrow k$ signify that job j must complete before job k can start. This requirement is enforced by adjusting the release date of any such k so that $r_j + p_j \leq r_k$ whenever $j \rightarrow k$. Furthermore, the job ℓ selected in Step 3 must have no successors within its block B .

The weighted number of tardy jobs problem with release dates and preemptions $(1|r_j, prmp, |\sum w_j U_j)$ is NP-hard, but Lawler [28] developed a pseudopolynomial algorithm for it. This algorithm operates in polynomial time when applied to the unweighted problem $(1|r_j, prmp, |\sum U_j)$. Furthermore, when applied to the weighted problem without release dates, it produces an optimal schedule without preemptions. It also solves the problem with release dates but without preemptions if release dates and due dates are similarly ordered, that is, if

$$r_1 \leq r_2 \leq \dots \leq r_n \iff d_1 \leq d_2 \leq \dots \leq d_n.$$

CONCLUDING REMARKS

There are several additional single machine scheduling problems beyond those discussed here. A scheduler may face two objectives, one primary and one secondary. One approach is to solve for the primary one first by applying the appropriate scheduling rule, then, if possible, to address the second by breaking any ties with the scheduling rule for the second. Emmons [29] considered this problem when the primary objective is a nondecreasing cost function of completion times and the secondary is total completion time. Chen and Bulfin [30] developed algorithms to find the schedule with the fewest number of tardy jobs among all those having minimum maximum tardiness. Others [31,32] have sought the set of Pareto Optimal schedules for multicriteria problems. A schedule is *Pareto Optimal* if decreasing the value of one objective necessarily increases the value of another.

Batch scheduling problems in single machine systems fall into two major groups [2,33]. In one, the jobs belong to different families. Jobs within the same family may

be processed consecutively with no setup required between them, but a setup is required when the machine switches families. Another type of batch system allows for multiple jobs to be processed simultaneously on the machine. The machine may have a batch-size limit. Both types of batch scheduling problems are generally solved by dynamic programming.

REFERENCES

1. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-Completeness. San Francisco (CA): W.H. Freeman and Company; 1979.
2. Brucker P. Scheduling algorithms. 5th ed. Berlin: Springer; 2007.
3. Gilmore P, Gomory R. Sequencing a one-state variable machine: a solvable case of the traveling salesman problem. *Oper Res* 1964;12:675–679.
4. Pinedo M. Scheduling: theory, algorithms, and systems. 3rd ed. New York: Springer; 2008.
5. Graham RL, Lawler EL, Lenstra JK, *et al.* Optimization and approximation in deterministic sequencing and scheduling: a survey. *Ann Discrete Math* 1979;5:287–326.
6. Smith WE. Various optimizers for single stage production. *Nav Res Log Q* 1956;3:59–66.
7. Lenstra JK, Rinnooy Kan AHG, Brucker P. Complexity of machine scheduling problems. *Ann Discrete Math* 1977;1:343–362.
8. Schrage L. A proof of the shortest remaining processing time processing discipline. *Oper Res* 1968;16:687–690.
9. Ahmadei RH, Bagchi U. Lower bounds for single-machine scheduling problems. *Nav Res Log Q* 1990;37:967–979.
10. Chu C. A branch-and-bound algorithm to minimize total flow time with unequal release dates. *Nav Res Log Q* 1992;39:859–875.
11. Labetoulle J, Lawler EL, Lenstra JK, *et al.* Preemptive scheduling of uniform machines subject to release dates. In: Pulleyblank WR, editor. Progress in combinatorial optimization. New York: Academic Press; 1984. pp. 245–261.
12. Hariri AMA, Potts CN. An algorithm for single machine sequencing with release dates to minimize total weighted completion time. *Discrete Appl Math* 1983;5:99–109.
13. Belouadah H, Posner ME, Potts CN. Scheduling with release dates on a single machine to minimize total weighted completion time. *Discrete Appl Math* 1992;36:213–231.
14. Jackson JR. Scheduling a production line to minimize maximum tardiness. Research Report 43. Los Angeles (CA): Management Science Research Project, University of California; 1955.
15. Horn WA. Some simple scheduling algorithms. *Nav Res Logist* 1974;21:177–185.
16. Lawler EL. Optimal sequencing of a single machine subject to precedence constraints. *Manage Sci* 1973;19(5):544–546.
17. Karp RM. Reducibility among combinatorial problems. Complexity of computer computations (Proc. Sympos. IBM Thomas J. Watson Research Center, Yorktown Heights (NY)). New York: Plenum; 1972. pp 85–103.
18. Moore JM. An n job, one machine sequencing algorithm for minimizing the number of late jobs. *Manage Sci* 1968;15:102–109.
19. Lawler EL, Moore JM. A functional equation and its application to resource allocation and sequencing problems. *Manage Sci* 1969;16(1):77–84.
20. Du J, Leung JY-T. Minimizing total tardiness on One machine is NP-hard. *Math Oper Res* 1990;15:483–495.
21. Lawler EL. A pseudopolynomial time algorithm for sequencing jobs to minimize total tardiness. *Ann Discrete Math* 1977;1:331–342.
22. Ohno T, Mito S. Just-in-time for today and tomorrow. New York: Productivity Press; 1988.
23. Merten AG, Muller ME. Variance minimization in single machine sequencing problems. *Manage Sci* 1972;18(9):518–528.
24. Sidney JB. Optimal single-machine scheduling with earliness and tardiness penalties. *Oper Res* 1977;25(1):62–69.
25. Eilon S, Chowdhury IG. Minimising waiting time variance in the single machine problem. *Manage Sci* 1977;23(6):567–575.
26. Kanet JJ. Minimizing the average deviation of job completion times about a common due date. *Nav Res Logist Q* 1981;28(4):643–651.
27. Baker KR, Lawler EL, Lenstra JK, *et al.* Preemptive scheduling of a single machine to minimize maximum cost subject to release dates and precedence constraints. *Oper Res* 1983;31(2):381–386.
28. Lawler, EL. A dynamic programming algorithm for preemptive scheduling of a single

- machine to minimize the number of late jobs. *Ann Oper Res* 1990;26:125–133.
29. Emmons H. A note on a scheduling problem with dual criteria. *Nav Res Logist Q* 1975;22:615–616.
30. Chen C-L, Bulfin RL. Scheduling a single machine to minimize two criteria: maximum tardiness and number of tardy jobs. *IIE Transactions* 1994;26:76–84.
31. van Wassenhove LN, Gelders F. Solving a bicriterion scheduling problem. *Eur J Oper Res* 1980;4:42–48.
32. Nelson RT, Sarin RK, Daniels RL. Scheduling with multiple performance measures: the one machine case. *Manage Sci* 1986;32:464–479.
33. Potts CN, Kovalyov MY. Scheduling with batching: a review. *Eur J Oper Res* 2000;120:228–249.

SINGLE-DIMENSIONAL SEARCH METHODS

EDWIN K. P. CHONG
Department of Electrical and
Computer Engineering, Colorado
State University, Fort Collins,
Colorado

STANISLAW H. ŻAK
School of Electrical and Computer
Engineering, Purdue University,
West Lafayette, Indiana

OVERVIEW

We are interested here in the problem of minimizing a function $f: \mathbb{R} \rightarrow \mathbb{R}$ (i.e., a single-dimensional problem). We refer to f as the *objective function*. The approach is to use an iterative search algorithm, also called a *line-search method*. Single-dimensional search methods are of interest for the following reasons. First, they are special cases of search methods used in multivariable problems. Second, they are used as part of general multivariable algorithms (as described later in the section titled “Line Search in Multidimensional Optimization”).

In an iterative algorithm, we start with an initial candidate solution $x^{(0)}$ and generate a sequence of iterates $x^{(1)}, x^{(2)}, \dots$. For each iteration $k = 0, 1, 2, \dots$, the next point $x^{(k+1)}$ depends on $x^{(k)}$ and the objective function f . The algorithm may use only f , and perhaps its first derivative f' , and even its second derivative f'' . In this article, we study several algorithms:

- Golden section method (uses only f)
- Fibonacci method (uses only f)
- Bisection method (uses only f')
- Secant method (uses only f')
- Newton's method (uses f' and f'').

The exposition here is based on that in Chong and Żak [1, Chapter 7].

GOLDEN SECTION SEARCH

The search methods we discuss in this section and the next allow us to determine the minimizer of an objective function $f: \mathbb{R} \rightarrow \mathbb{R}$ over a closed interval, say $[a_0, b_0]$. The only property that we assume of the objective function f is that it is *unimodal*, which means that f has only one local minimizer. An example of such a function is depicted in Fig. 1.

The methods we discuss are based on evaluating the objective function at different points in the interval $[a_0, b_0]$. We choose these points in such a way that an approximation to the minimizer of f may be achieved in as few evaluations as possible. Our goal is to narrow the range progressively until the minimizer is “boxed in” with sufficient accuracy.

Consider a unimodal function f of one variable and the interval $[a_0, b_0]$. If we evaluate f at only one intermediate point of the interval, we cannot narrow the range within which we know the minimizer is located. We have to evaluate f at two intermediate points, as illustrated in Fig. 2. We choose the intermediate points in such a way that the reduction in the range is symmetric, in the sense that

$$a_1 - a_0 = b_0 - b_1 = \rho(b_0 - a_0),$$

where

$$\rho < \frac{1}{2}.$$

We then evaluate f at the intermediate points. If $f(a_1) < f(b_1)$, then the minimizer must lie in the range $[a_0, b_1]$. If, on the other hand, $f(a_1) \geq f(b_1)$, then the minimizer is located in the range $[a_1, b_0]$ (Fig. 3).

Starting with the reduced range of uncertainty, we can repeat the process and similarly find two new points, say a_2 and b_2 , using the same value of $\rho < \frac{1}{2}$ as before. However, we would like to minimize the number of objective function evaluations while reducing the width of the uncertainty interval. Suppose, for example, that $f(a_1) < f(b_1)$, as in Fig. 3. Then, we know that $x^* \in [a_0, b_1]$.

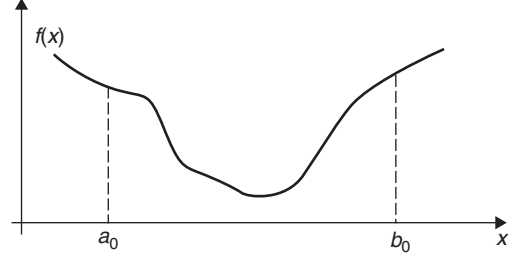
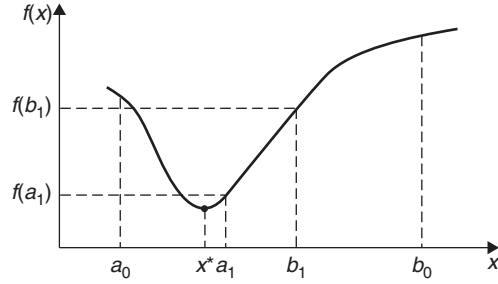
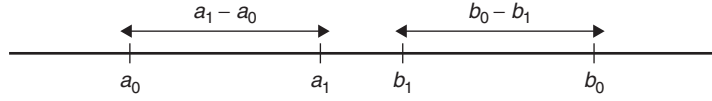


Figure 1. Unimodal function.

Figure 2. Evaluating the objective function at two intermediate points.


 Figure 3. The case where $f(a_1) < f(b_1)$; the minimizer $x^* \in [a_0, b_1]$.

Because a_1 is already in the uncertainty interval and $f(a_1)$ is already known, we can make a_1 coincide with b_2 . Thus, only one new evaluation of f at a_2 would be necessary. To find the value of ρ that results in only one new evaluation of f , see Fig. 4. Without loss of generality, imagine that the original range $[a_0, b_0]$ is of unit length. Then, to have only one new evaluation of f it is enough to choose ρ so that

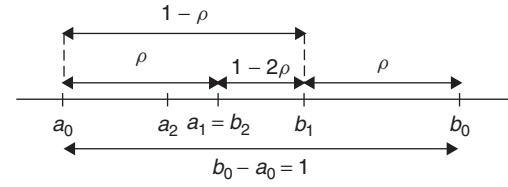
$$\rho(b_1 - a_0) = b_1 - b_2.$$

Because $b_1 - a_0 = 1 - \rho$ and $b_1 - b_2 = 1 - 2\rho$, we have

$$\rho(1 - \rho) = 1 - 2\rho.$$

We write the above quadratic equation as

$$\rho^2 - 3\rho + 1 = 0.$$


 Figure 4. Finding value of ρ resulting in only one new evaluation of f .

The solutions are

$$\rho_1 = \frac{3 + \sqrt{5}}{2}, \quad \rho_2 = \frac{3 - \sqrt{5}}{2}.$$

Because we require that $\rho < \frac{1}{2}$, we take

$$\rho = \frac{3 - \sqrt{5}}{2} \approx 0.382.$$

Observe that

$$1 - \rho = \frac{\sqrt{5} - 1}{2},$$

and

$$\frac{\rho}{1 - \rho} = \frac{3 - \sqrt{5}}{\sqrt{5} - 1} = \frac{\sqrt{5} - 1}{2} = \frac{1 - \rho}{1},$$

that is,

$$\frac{\rho}{1 - \rho} = \frac{1 - \rho}{1}.$$

Thus, dividing a range in the ratio of ρ to $1 - \rho$ has the effect that the ratio of the shorter segment to the longer equals the ratio of the longer to the sum of the two. This rule was referred to by ancient Greek geometers as the *golden section*.

If the golden section rule is used it means that at every stage of the uncertainty range reduction (except the first), the objective function f need only be evaluated at one new point. The uncertainty range is reduced by the ratio $1 - \rho \approx 0.61803$ at every stage. Hence, N steps of reduction using the golden section method reduces the range by the factor

$$(1 - \rho)^N \approx (0.61803)^N.$$

Example 1. Suppose that we wish to use the golden section search method to find the value of x that minimizes

$$f(x) = x^4 - 14x^3 + 60x^2 - 70x$$

in the interval $[0, 2]$ (this function comes from an example in Bunday [2]). We wish to locate this value of x to within a range of 0.3.

After N stages, the range $[0, 2]$ is reduced by $(0.61803)^N$. Hence, we choose N so that

$$(0.61803)^N \leq 0.3/2.$$

It requires four stages of reduction; that is, $N = 4$.

Iteration 1. We evaluate f at two intermediate points a_1 and b_1 . We have

$$a_1 = a_0 + \rho(b_0 - a_0) = 0.7639,$$

$$b_1 = a_0 + (1 - \rho)(b_0 - a_0) = 1.236,$$

where $\rho = (3 - \sqrt{5})/2$. We compute

$$f(a_1) = -24.36,$$

$$f(b_1) = -18.96.$$

Thus, $f(a_1) < f(b_1)$, so the uncertainty interval is reduced to

$$[a_0, b_1] = [0, 1.236].$$

Iteration 2. We choose b_2 to coincide with a_1 , and so f need only be evaluated at one new point,

$$a_2 = a_0 + \rho(b_1 - a_0) = 0.4721.$$

We have

$$f(a_2) = -21.10,$$

$$f(b_2) = f(a_1) = -24.36.$$

Now, $f(b_2) < f(a_2)$, so the uncertainty interval is reduced to

$$[a_2, b_1] = [0.4721, 1.236].$$

Iteration 3. We set $a_3 = b_2$ and compute b_3 :

$$b_3 = a_2 + (1 - \rho)(b_1 - a_2) = 0.9443.$$

We have

$$f(a_3) = f(b_2) = -24.36,$$

$$f(b_3) = -23.59.$$

So $f(b_3) > f(a_3)$. Hence, the uncertainty interval is further reduced to

$$[a_2, b_3] = [0.4721, 0.9443].$$

Iteration 4. We set $b_4 = a_3$ and

$$a_4 = a_2 + \rho(b_3 - a_2) = 0.6525.$$

We have

$$f(a_4) = -23.84,$$

$$f(b_4) = f(a_3) = -24.36.$$

Hence, $f(a_4) > f(b_4)$. Thus, the value of x that minimizes f is located in the interval

$$[a_4, b_3] = [0.6525, 0.9443].$$

Note that $b_3 - a_4 = 0.292 < 0.3$.

FIBONACCI METHOD

Recall that the golden section method uses the same value of ρ throughout. Suppose now that we are allowed to vary the value ρ from stage to stage, so that at the k th stage in the reduction process we use a value ρ_k , at the next stage we use a value ρ_{k+1} , and so on. As in the golden section search, our goal is to select successive values of ρ_k , $0 \leq \rho_k \leq 1/2$, such that only one new function evaluation is required at each stage. To derive the strategy for selecting evaluation points, consider Fig. 5. From this figure we see that it is sufficient to choose ρ_k such that

$$\rho_{k+1}(1 - \rho_k) = 1 - 2\rho_k.$$

After a few manipulations, we obtain

$$\rho_{k+1} = 1 - \frac{\rho_k}{1 - \rho_k}.$$

There are many sequences $\rho_1, \rho_2, \dots$ that satisfy the above law of formation and the condition that $0 \leq \rho_k \leq 1/2$. For example, the sequence $\rho_1 = \rho_2 = \rho_3 = \dots = (3 - \sqrt{5})/2$ satisfies the above conditions and gives rise to the golden section method.

Suppose that we are given a sequence $\rho_1, \rho_2, \dots$ satisfying the conditions above and we use this sequence in our search algorithm. Then, after N iterations of the algorithm, the uncertainty range is reduced by a factor of

$$(1 - \rho_1)(1 - \rho_2) \dots (1 - \rho_N).$$

Depending on the sequence $\rho_1, \rho_2, \dots$ we get a different reduction factor. The natural question is as follows: What sequence

$\rho_1, \rho_2, \dots$ minimizes the above reduction factor? This problem is a constrained optimization problem that can be stated formally as

$$\text{minimize } (1 - \rho_1)(1 - \rho_2) \dots (1 - \rho_N)$$

$$\text{subject to } \rho_{k+1} = 1 - \frac{\rho_k}{1 - \rho_k},$$

$$k = 1, \dots, N - 1$$

$$0 \leq \rho_k \leq \frac{1}{2}, k = 1, \dots, N.$$

Before we give the solution to the optimization problem above, we need to introduce the *Fibonacci sequence* $F_1, F_2, F_3, \dots$. This sequence is defined as follows. First, let $F_{-1} = 0$ and $F_0 = 1$ by convention. Then, for $k \geq 0$,

$$F_{k+1} = F_k + F_{k-1}.$$

Some values of elements in the Fibonacci sequence are as follows:

$$\begin{array}{cccccccc} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 & F_7 & F_8 \\ 1 & 2 & 3 & 5 & 8 & 13 & 21 & 34 \end{array}$$

It turns out that the solution to the above optimization problem is

$$\rho_1 = 1 - \frac{F_N}{F_{N+1}},$$

$$\rho_2 = 1 - \frac{F_{N-1}}{F_N},$$

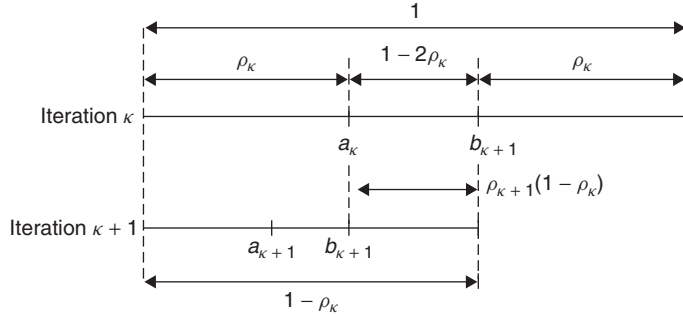


Figure 5. Selecting evaluation points.

$$\begin{aligned}
& \vdots \\
\rho_k &= 1 - \frac{F_{N-k+1}}{F_{N-k+2}}, \\
& \vdots \\
\rho_N &= 1 - \frac{F_1}{F_2},
\end{aligned}$$

where the F_k are the elements of the Fibonacci sequence. The resulting algorithm is called the *Fibonacci search method*. We present a proof for the optimality of the Fibonacci search method later in this section.

In the Fibonacci search method, the uncertainty range is reduced by the factor

$$\begin{aligned}
& (1 - \rho_1)(1 - \rho_2) \dots (1 - \rho_N) \\
&= \frac{F_N}{F_{N+1}} \frac{F_{N-1}}{F_N} \dots \frac{F_1}{F_2} = \frac{F_1}{F_{N+1}} = \frac{1}{F_{N+1}}.
\end{aligned}$$

Because the Fibonacci method uses the optimal values of $\rho_1, \rho_2, \dots$ the above reduction factor is less than that of the golden section method. In other words, the Fibonacci method is better than the golden section method in that it gives a smaller final uncertainty range.

We point out that there is an anomaly in the final iteration of the Fibonacci search method, because

$$\rho_N = 1 - \frac{F_1}{F_2} = \frac{1}{2}.$$

Recall that we need two intermediate points at each stage, one that comes from a previous iteration and another that is a new evaluation point. However, with $\rho_N = 1/2$, the two intermediate points coincide in the middle of the uncertainty interval, and, therefore, we cannot further reduce the uncertainty range. To get around this problem, we perform the new evaluation for the last iteration using $\rho_N = 1/2 - \epsilon$, where ϵ is a small number. In other words, the new evaluation point is just to the left or right of the midpoint of the uncertainty interval. This modification to the Fibonacci method is, of course, of no significant practical consequence.

As a result of the modification above, the reduction in the uncertainty range at the last

iteration may be either

$$1 - \rho_N = \frac{1}{2}$$

or

$$1 - (\rho_N - \epsilon) = \frac{1}{2} + \epsilon = \frac{1 + 2\epsilon}{2},$$

depending on which of the two points has the smaller objective function value. Therefore, in the worst case, the reduction factor in the uncertainty range for the Fibonacci method is

$$\frac{1 + 2\epsilon}{F_{N+1}}.$$

Example 2. Consider the function

$$f(x) = x^4 - 14x^3 + 60x^2 - 70x.$$

Suppose that we wish to use the Fibonacci search method to find the value of x that minimizes f over the range $[0, 2]$, and locate this value of x to within the range 0.3.

After N steps the range is reduced by $(1 + 2\epsilon)/F_{N+1}$ in the worst case. We need to choose N such that

$$\frac{1 + 2\epsilon}{F_{N+1}} \leq \frac{\text{final range}}{\text{initial range}} = \frac{0.3}{2} = 0.15.$$

Thus, we need

$$F_{N+1} \geq \frac{1 + 2\epsilon}{0.15}.$$

If we choose $\epsilon \leq 0.1$, then $N = 4$ is sufficient.

Iteration 1. We start with

$$1 - \rho_1 = \frac{F_4}{F_5} = \frac{5}{8}.$$

We then compute

$$a_1 = a_0 + \rho_1(b_0 - a_0) = \frac{3}{4},$$

$$b_1 = a_0 + (1 - \rho_1)(b_0 - a_0) = \frac{5}{4},$$

$$f(a_1) = -24.34,$$

$$f(b_1) = -18.65,$$

$$f(a_1) < f(b_1).$$

The range is reduced to

$$[a_0, b_1] = \left[0, \frac{5}{4}\right].$$

Iteration 2. We have

$$\begin{aligned} 1 - \rho_2 &= \frac{F_3}{F_4} = \frac{3}{5}, \\ a_2 &= a_0 + \rho_2(b_1 - a_0) = \frac{1}{2}, \\ b_2 &= a_1 = \frac{3}{4}, \\ f(a_2) &= -21.69, \\ f(b_2) &= f(a_1) = -24.34, \\ f(a_2) &> f(b_2), \end{aligned}$$

so the range is reduced to

$$[a_2, b_1] = \left[\frac{1}{2}, \frac{5}{4}\right].$$

Iteration 3. We compute

$$\begin{aligned} 1 - \rho_3 &= \frac{F_2}{F_3} = \frac{2}{3}, \\ a_3 &= b_2 = \frac{3}{4}, \\ b_3 &= a_2 + (1 - \rho_3)(b_1 - a_2) = 1, \\ f(a_3) &= f(b_2) = -24.34, \\ f(b_3) &= -23, \\ f(a_3) &< f(b_3). \end{aligned}$$

The range is reduced to

$$[a_2, b_3] = \left[\frac{1}{2}, 1\right].$$

Iteration 4. We choose $\epsilon = 0.05$. We have

$$\begin{aligned} 1 - \rho_4 &= \frac{F_1}{F_2} = \frac{1}{2}, \\ a_4 &= a_2 + (\rho_4 - \epsilon)(b_3 - a_2) = 0.725, \\ b_4 &= a_3 = \frac{3}{4}, \\ f(a_4) &= -24.27, \\ f(b_4) &= f(a_3) = -24.34, \\ f(a_4) &> f(b_4). \end{aligned}$$

The range is reduced to

$$[a_4, b_3] = [0.725, 1].$$

Note that $b_3 - a_4 = 0.275 < 0.3$.

We now turn to a proof of the optimality of the Fibonacci search method. Skipping the rest of this section does not affect the continuity of the presentation.

To begin, recall that we wish to prove that the values of $\rho_1, \rho_2, \dots, \rho_N$ used in the Fibonacci method, where $\rho_k = 1 - F_{N-k+1}/F_{N-k+2}$, solve the optimization problem

$$\begin{aligned} &\text{minimize} \quad (1 - \rho_1)(1 - \rho_2) \dots (1 - \rho_N) \\ &\text{subject to} \quad \rho_{k+1} = 1 - \frac{\rho_k}{1 - \rho_k}, \\ &\quad \quad \quad k = 1, \dots, N-1 \\ &\quad \quad \quad 0 \leq \rho_k \leq \frac{1}{2}, \quad k = 1, \dots, N. \end{aligned}$$

It is easy to check that the above values of $\rho_1, \rho_2, \dots$ for the Fibonacci search method satisfy the feasibility conditions in the optimization problem above. Recall that the Fibonacci method has an overall reduction factor of $(1 - \rho_1) \dots (1 - \rho_N) = 1/F_{N+1}$. To prove that the Fibonacci search method is optimal, we show that for any feasible values of $\rho_1, \dots, \rho_N$, we have $(1 - \rho_1) \dots (1 - \rho_N) \geq 1/F_{N+1}$.

It is more convenient to work with $r_k = 1 - \rho_k$ rather than ρ_k . The optimization problem stated in terms of r_k is

$$\begin{aligned} &\text{minimize} \quad r_1 \dots r_N \\ &\text{subject to} \quad r_{k+1} = \frac{1}{r_k} - 1, \quad k = 1, \dots, N-1 \\ &\quad \quad \quad \frac{1}{2} \leq r_k \leq 1, \quad k = 1, \dots, N. \end{aligned}$$

Note that if $r_1, r_2, \dots$ satisfy $r_{k+1} = \frac{1}{r_k} - 1$, then $r_k \geq 1/2$ if and only if $r_{k+1} \leq 1$. Also, $r_k \geq 1/2$ if and only if $r_{k-1} \leq 2/3 \leq 1$. Therefore, in the constraints above, we may remove the constraint $r_k \leq 1$, because it is implied implicitly by $r_k \geq 1/2$ and the other constraints.

Therefore, the constraints above reduce to

$$r_{k+1} = \frac{1}{r_k} - 1, \quad k = 1, \dots, N-1,$$

$$r_k \geq \frac{1}{2}, \quad k = 1, \dots, N.$$

To proceed further, we need the following technical lemmas. In the statements of the lemmas, we assume that $r_1, r_2, \dots$ is a sequence that satisfies

$$r_{k+1} = \frac{1}{r_k} - 1, \quad r_k \geq \frac{1}{2}, \quad k = 1, 2, \dots$$

Lemma 1. For $k \geq 2$,

$$r_k = -\frac{F_{k-2} - F_{k-1}r_1}{F_{k-3} - F_{k-2}r_1}.$$

Proof. We proceed by induction. For $k = 2$, we have

$$r_2 = \frac{1}{r_1} - 1 = \frac{1 - r_1}{r_1} = -\frac{F_0 - F_1r_1}{F_{-1} - F_0r_1},$$

and hence the lemma holds for $k = 2$. Suppose now that the lemma holds for $k \geq 2$. We show that it also holds for $k + 1$. We have

$$\begin{aligned} r_{k+1} &= \frac{1}{r_k} - 1 \\ &= \frac{-F_{k-3} + F_{k-2}r_1}{F_{k-2} - F_{k-1}r_1} - \frac{F_{k-2} - F_{k-1}r_1}{F_{k-2} - F_{k-1}r_1} \\ &= -\frac{F_{k-2} + F_{k-3} - (F_{k-1} + F_{k-2})r_1}{F_{k-2} - F_{k-1}r_1} \\ &= -\frac{F_{k-1} - F_kr_1}{F_{k-2} - F_{k-1}r_1}, \end{aligned}$$

where we used the formation law for the Fibonacci sequence.

Lemma 2. For $k \geq 2$,

$$(-1)^k(F_{k-2} - F_{k-1}r_1) > 0.$$

Proof. We proceed by induction. For $k = 2$, we have

$$(-1)^2(F_0 - F_1r_1) = 1 - r_1.$$

But $r_1 = 1/(1 + r_2) \leq 2/3$, and hence $1 - r_1 > 0$. Therefore, the result holds for $k = 2$. Suppose now that the lemma holds for $k \geq 2$. We show that it also holds for $k + 1$. We have

$$\begin{aligned} &(-1)^{k+1}(F_{k-1} - F_kr_1) \\ &= (-1)^{k+1}r_{k+1} \frac{1}{r_{k+1}}(F_{k-1} - F_kr_1). \end{aligned}$$

By Lemma 1,

$$r_{k+1} = -\frac{F_{k-1} - F_kr_1}{F_{k-2} - F_{k-1}r_1}.$$

Substituting for $1/r_{k+1}$, we obtain

$$\begin{aligned} &(-1)^{k+1}(F_{k-1} - F_kr_1) \\ &= r_{k+1}(-1)^k(F_{k-2} - F_{k-1}r_1) > 0, \end{aligned}$$

which completes the proof.

Lemma 3. For $k \geq 2$,

$$(-1)^{k+1}r_1 \geq (-1)^{k+1} \frac{F_k}{F_{k+1}}.$$

Proof. Because $r_{k+1} = \frac{1}{r_k} - 1$ and $r_k \geq \frac{1}{2}$, we have $r_{k+1} \leq 1$. Substituting for r_{k+1} from Lemma 1, we get

$$-\frac{F_{k-1} - F_kr_1}{F_{k-2} - F_{k-1}r_1} \leq 1.$$

Multiplying the numerator and denominator by $(-1)^k$ yields

$$\frac{(-1)^{k+1}(F_{k-1} - F_kr_1)}{(-1)^k(F_{k-2} - F_{k-1}r_1)} \leq 1.$$

By Lemma 2, $(-1)^k(F_{k-2} - F_{k-1}r_1) > 0$, and, therefore, we can multiply both sides of the

inequality above by $(-1)^k(F_{k-2} - F_{k-1}r_1)$ to obtain

$$(-1)^{k+1}(F_{k-1} - F_k r_1) \leq (-1)^k(F_{k-2} - F_{k-1}r_1).$$

Rearranging yields

$$(-1)^{k+1}(F_{k-1} + F_k)r_1 \geq (-1)^{k+1}(F_{k-2} + F_{k-1}).$$

Using the law of formation of the Fibonacci sequence, we get

$$(-1)^{k+1}F_{k+1}r_1 \geq (-1)^{k+1}F_k,$$

which upon dividing by F_{k+1} on both sides gives the desired result.

We are now ready to prove the optimality of the Fibonacci search method and the uniqueness of this optimal solution.

Theorem 1. *Let $r_1, \dots, r_N$, $N \geq 2$, satisfy the constraints*

$$\begin{aligned} r_{k+1} &= \frac{1}{r_k} - 1, \quad k = 1, \dots, N-1, \\ r_k &\geq \frac{1}{2}, \quad k = 1, \dots, N. \end{aligned}$$

Then,

$$r_1 \cdots r_N \geq \frac{1}{F_{N+1}}.$$

Furthermore,

$$r_1 \cdots r_N = \frac{1}{F_{N+1}}$$

if and only if $r_k = F_{N-k+1}/F_{N-k+2}$, $k = 1, \dots, N$. In other words, the values of $r_1, \dots, r_N$ used in the Fibonacci search method form a unique solution to the optimization problem.

Proof. By substituting expressions for $r_1, \dots, r_N$ from Lemma 1 and performing the appropriate cancellations, we obtain

$$\begin{aligned} r_1 \cdots r_N &= (-1)^N(F_{N-2} - F_{N-1}r_1) \\ &= (-1)^N F_{N-2} + F_{N-1}(-1)^{N+1}r_1. \end{aligned}$$

Using Lemma 3 we have

$$\begin{aligned} r_1 \cdots r_N &\geq (-1)^N F_{N-2} + F_{N-1}(-1)^{N+1} \frac{F_N}{F_{N+1}} \\ &= (-1)^N (F_{N-2}F_{N+1} - F_{N-1}F_N) \frac{1}{F_{N+1}}. \end{aligned}$$

It is readily checked that the following identity holds: $(-1)^N(F_{N-2}F_{N+1} - F_{N-1}F_N) = 1$. Hence,

$$r_1 \cdots r_N \geq \frac{1}{F_{N+1}}.$$

From the above we see that

$$r_1 \cdots r_N = \frac{1}{F_{N+1}}$$

if and only if

$$r_1 = \frac{F_N}{F_{N+1}}.$$

This is simply the value of r_1 for the Fibonacci search method. Note that fixing r_1 determines $r_2, \dots, r_N$ uniquely.

For further discussion on the Fibonacci search method and its variants, see Wilde [3].

BISECTION METHOD

Again we consider the problem of finding the minimizer of an objective function $f: \mathbb{R} \rightarrow \mathbb{R}$ over an interval $[a_0, b_0]$. As before, we assume that the objective function f is unimodal. Further, suppose that f is continuously differentiable and that we can use values of the derivative f' as a basis for reducing the uncertainty interval.

The *bisection method* is a simple algorithm for successively reducing the uncertainty interval based on evaluations of the derivative. To begin, let $x^{(0)} = (a_0 + b_0)/2$ be the midpoint of the initial uncertainty interval. Next, evaluate $f'(x^{(0)})$. If $f'(x^{(0)}) > 0$, then we deduce that the minimizer lies to the *left* of $x^{(0)}$. In other words, we reduce the uncertainty interval to $[a_0, x^{(0)}]$. On the other hand, if $f'(x^{(0)}) < 0$, then we deduce that the minimizer lies to the *right* of $x^{(0)}$. In this case, we reduce the uncertainty interval to $[x^{(0)}, b_0]$.

Finally, if $f'(x^{(0)}) = 0$, then we declare $x^{(0)}$ to be the minimizer and terminate our search.

With the new uncertainty interval computed, we repeat the process iteratively. At each iteration k , we compute the midpoint of the uncertainty interval. Call this point $x^{(k)}$. Depending on the sign of $f'(x^{(k)})$ (assuming that it is nonzero), we reduce the uncertainty interval to the left or right of $x^{(k)}$. If at any iteration k we find that $f'(x^{(k)}) = 0$, then we declare $x^{(k)}$ to be the minimizer and terminate our search.

Two salient features distinguish the bisection method from the golden section and Fibonacci methods. First, instead of using values of f , the bisection method uses values of f' . Second, at each iteration, the length of the uncertainty interval is reduced by a factor of $1/2$. Hence, after N steps, the range is reduced by a factor of $(1/2)^N$. This factor is smaller than in the golden section and Fibonacci methods.

Example 3. Recall Example 1 where we wish to find the minimizer of

$$f(x) = x^4 - 14x^3 + 60x^2 - 70x$$

in the interval $[0, 2]$ to within a range of 0.3 . The golden section method requires at least four stages of reduction. If, instead, we use the bisection method, we would choose N so that

$$(0.5)^N \leq 0.3/2.$$

In this case, only three stages of reduction are needed.

NEWTON'S METHOD

Suppose again that we are confronted with the problem of minimizing a function f of a single real variable x . We assume now that at each measurement point $x^{(k)}$ we can determine $f(x^{(k)})$, $f'(x^{(k)})$, and $f''(x^{(k)})$. We can fit a quadratic function through $x^{(k)}$ that matches its first and second derivatives with that of the function f . This quadratic has the form

$$q(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + \frac{1}{2} f''(x^{(k)})(x - x^{(k)})^2.$$

Note that $q(x^{(k)}) = f(x^{(k)})$, $q'(x^{(k)}) = f'(x^{(k)})$, and $q''(x^{(k)}) = f''(x^{(k)})$. Then, instead of minimizing f , we minimize its approximation q . The first-order necessary condition for a minimizer of q yields

$$0 = q'(x) = f'(x^{(k)}) + f''(x^{(k)})(x - x^{(k)}).$$

Setting $x = x^{(k+1)}$, we obtain

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}.$$

Example 4. Using Newton's method, we will find the minimizer of

$$f(x) = \frac{1}{2}x^2 - \sin x.$$

Suppose that the initial value is $x^{(0)} = 0.5$, and that the required accuracy is $\epsilon = 10^{-5}$, in the sense that we stop when $|x^{(k+1)} - x^{(k)}| < \epsilon$.

We compute

$$f'(x) = x - \cos x, \quad f''(x) = 1 + \sin x.$$

Hence,

$$\begin{aligned} x^{(1)} &= 0.5 - \frac{0.5 - \cos 0.5}{1 + \sin 0.5} \\ &= 0.5 - \frac{-0.3775}{1.479} \\ &= 0.7552. \end{aligned}$$

Proceeding in a similar manner, we obtain

$$\begin{aligned} x^{(2)} &= x^{(1)} - \frac{f'(x^{(1)})}{f''(x^{(1)})} \\ &= x^{(1)} - \frac{0.02710}{1.685} = 0.7391, \\ x^{(3)} &= x^{(2)} - \frac{f'(x^{(2)})}{f''(x^{(2)})} \\ &= x^{(2)} - \frac{9.461 \times 10^{-5}}{1.673} = 0.7390, \\ x^{(4)} &= x^{(3)} - \frac{f'(x^{(3)})}{f''(x^{(3)})} \\ &= x^{(3)} - \frac{1.17 \times 10^{-9}}{1.673} = 0.7390. \end{aligned}$$

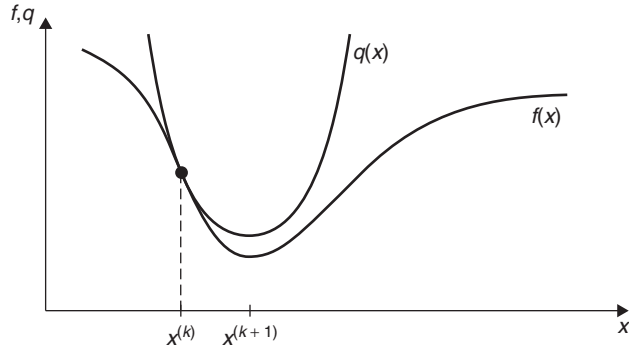


Figure 6. Newton's algorithm with $f''(x) > 0$.

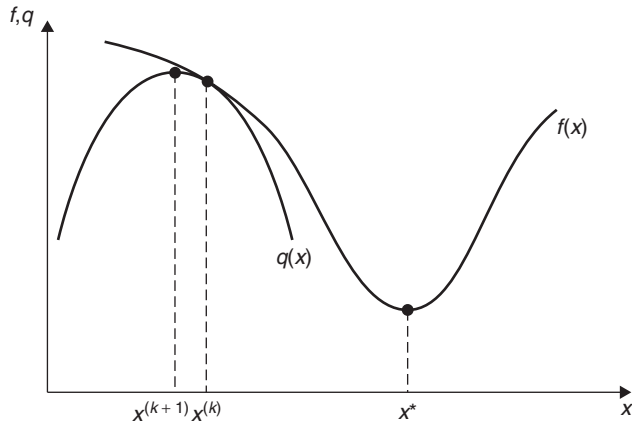


Figure 7. Newton's algorithm with $f''(x) < 0$.

Note that $|x^{(4)} - x^{(3)}| < \epsilon = 10^{-5}$. Furthermore, $f'(x^{(4)}) = -8.6 \times 10^{-6} \approx 0$. Observe that $f''(x^{(4)}) = 1.673 > 0$, so we can assume that $x^* \approx x^{(4)}$ is a strict minimizer.

Newton's method works well if $f''(x) > 0$ everywhere (Fig. 6). However, if $f''(x) < 0$ for some x , Newton's method may fail to converge to the minimizer (see Fig. 7).

Newton's method can also be viewed as a way to drive the first derivative of f to zero. Indeed, if we set $g(x) = f'(x)$, then we obtain a formula for iterative solution of the equation $g(x) = 0$:

$$x^{(k+1)} = x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})}.$$

In other words, we can use Newton's method for zero finding.

Example 5. We apply Newton's method to improve a first approximation, $x^{(0)} = 12$, to

the root of the equation

$$g(x) = x^3 - 12.2x^2 + 7.45x + 42 = 0.$$

We have $g'(x) = 3x^2 - 24.4x + 7.45$.

Performing two iterations yields

$$x^{(1)} = 12 - \frac{102.6}{146.65} = 11.33,$$

$$x^{(2)} = 11.33 - \frac{14.73}{116.11} = 11.21.$$

Newton's method for solving equations of the form $g(x) = 0$ is also referred to as *Newton's method of tangents*. This name is easily justified if we look at a geometric interpretation of the method when applied to the solution of the equation $g(x) = 0$ (Fig. 8).

If we draw a tangent to $g(x)$ at the given point $x^{(k)}$, then the tangent line intersects the

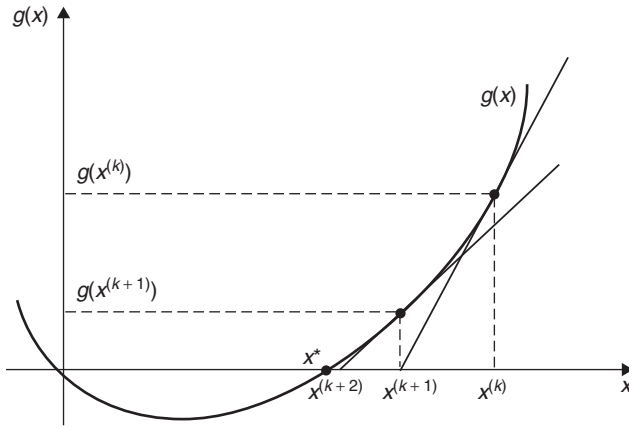


Figure 8. Newton's method of tangents.

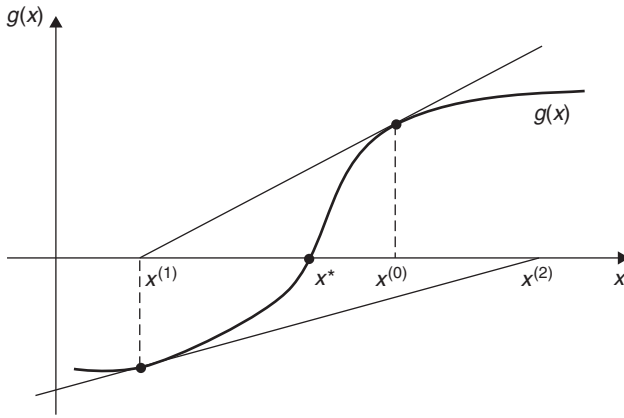


Figure 9. Example where Newton's method of tangents fails to converge to the root x^* of $g(x) = 0$.

x axis at the point $x^{(k+1)}$, which we expect to be closer to the root x^* of $g(x) = 0$. Note that the slope of $g(x)$ at $x^{(k)}$ is

$$g'(x^{(k)}) = \frac{g(x^{(k)})}{x^{(k)} - x^{(k+1)}}.$$

Hence,

$$x^{(k+1)} = x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})}.$$

Newton's method of tangents may fail if the first approximation to the root is such that the ratio $g(x^{(0)})/g'(x^{(0)})$ is not small enough (Fig. 9). Thus, an initial approximation to the root is very important.

SECANT METHOD

Newton's method for minimizing f uses second derivatives of f :

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}.$$

If the second derivative is not available, we may attempt to approximate it using first derivative information. In particular, we may approximate $f''(x^{(k)})$ above with

$$\frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}.$$

Using the foregoing approximation of the second derivative, we obtain the

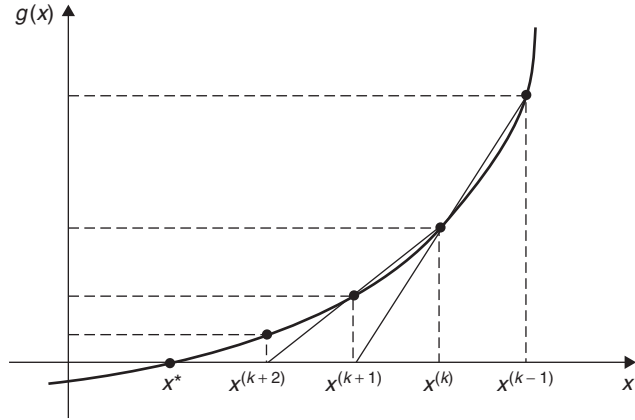


Figure 10. Secant method for finding the root.

algorithm

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f'(x^{(k)}) - f'(x^{(k-1)})} f'(x^{(k)}),$$

called the *secant method*. Note that the algorithm requires two initial points to start it, which we denote as $x^{(-1)}$ and $x^{(0)}$. The secant algorithm can be represented in the following equivalent form:

$$x^{(k+1)} = \frac{f'(x^{(k)})x^{(k-1)} - f'(x^{(k-1)})x^{(k)}}{f'(x^{(k)}) - f'(x^{(k-1)})}.$$

Observe that, like Newton's method, the secant method does not directly involve values of $f(x^{(k)})$. Instead, it tries to drive the derivative f' to zero. In fact, as we did for Newton's method, we can interpret the secant method as an algorithm for solving equations of the form $g(x) = 0$. Specifically, the secant algorithm for finding a root of the equation $g(x) = 0$ takes the form

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{g(x^{(k)}) - g(x^{(k-1)})} g(x^{(k)}),$$

or, equivalently,

$$x^{(k+1)} = \frac{g(x^{(k)})x^{(k-1)} - g(x^{(k-1)})x^{(k)}}{g(x^{(k)}) - g(x^{(k-1)})}.$$

The secant method for finding the root is illustrated in Fig. 10 (compare this with Fig. 8). Unlike Newton's method, which uses the slope of g to determine the next

point, the secant method uses the “secant” between the $(k-1)$ th and k th points to determine the $(k+1)$ th point.

Example 6. We apply the secant method to find the root of the equation

$$g(x) = x^3 - 12.2x^2 + 7.45x + 42 = 0.$$

We perform two iterations, with starting points $x^{(-1)} = 13$ and $x^{(0)} = 12$. We obtain

$$x^{(1)} = 11.40,$$

$$x^{(2)} = 11.25.$$

Example 7. Suppose that the voltage across a resistor in a circuit decays according to the model $V(t) = e^{-Rt}$, where $V(t)$ is the voltage at time t and R is the resistance value.

Given measurements $V_1, \dots, V_n$ of the voltage at times $t_1, \dots, t_n$, respectively, we wish to find the best estimate of R . By the *best estimate* we mean the value of R that minimizes the total squared error between the measured voltages and the voltages predicted by the model.

We derive an algorithm to find the best estimate of R using the secant method. The objective function is

$$f(R) = \sum_{i=1}^n (V_i - e^{-Rt_i})^2.$$

Hence, we have

$$f'(R) = 2 \sum_{i=1}^n (V_i - e^{-Rt_i}) e^{-Rt_i} t_i.$$

The secant algorithm for the problem is

$$\begin{aligned} R_{k+1} &= R_k - \frac{R_k - R_{k-1}}{\frac{\sum_{i=1}^n (V_i - e^{-R_k t_i}) e^{-R_k t_i} t_i}{-(V_i - e^{-R_{k-1} t_i}) e^{-R_{k-1} t_i} t_i}} \\ &\quad \times \sum_{i=1}^n (V_i - e^{-R_k t_i}) e^{-R_k t_i} t_i. \end{aligned}$$

For further reading on the secant method, see Conte and Boor [4]. Newton's method and the secant method are instances of *quadratic fit* methods. In Newton's method, $x^{(k+1)}$ is the stationary point of a quadratic function that fits f' and f'' at $x^{(k)}$. In the secant method, $x^{(k+1)}$ is the stationary point of a quadratic function that fits f' at $x^{(k)}$ and $x^{(k-1)}$. The secant method uses only f' (and not f'') but needs values from *two* previous points. We leave it to the reader to verify that if we set $x^{(k+1)}$ to be the stationary point of a quadratic function that fits f at $x^{(k)}$, $x^{(k-1)}$, and $x^{(k-2)}$, we obtain a quadratic fit method that uses only values of f :

$$\begin{aligned} x^{(k+1)} &= \frac{\sigma_{12}f(x^{(k)}) + \sigma_{20}f(x^{(k-1)}) + \sigma_{01}f(x^{(k-2)})}{2(\delta_{12}f(x^{(k)}) + \delta_{20}f(x^{(k-1)}) + \delta_{01}f(x^{(k-2)}))}, \end{aligned}$$

where $\sigma_{ij} = (x^{(k-i)})^2 - (x^{(k-j)})^2$ and $\delta_{ij} = x^{(k-i)} - x^{(k-j)}$. This method does not use f' or f'' , but needs values of f from *three* previous points. Three points are needed to initialize the iterations.

A similar approach to fitting (or interpolation) based on higher order polynomials is possible. For example, we could set $x^{(k+1)}$ to be a stationary point of a *cubic* function that fits f' at $x^{(k)}$, $x^{(k-1)}$, and $x^{(k-2)}$.

LINE SEARCH IN MULTIDIMENSIONAL OPTIMIZATION

Single-dimensional search methods play an important role in multidimensional optimization problems. In particular, iterative algorithms for solving such optimization problems [1] typically involve a *line search* at every iteration. To be specific, let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that we wish to minimize. Iterative algorithms for finding a minimizer of f are of the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)},$$

where $\mathbf{x}^{(0)}$ is a given initial point and $\alpha_k \geq 0$ is chosen to minimize $\phi_k(\alpha) = f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)})$. The vector $\mathbf{d}^{(k)}$ is called the *search direction* and α_k is called the *step size*. Figure 11 illustrates a line search within a multidimensional setting. Note that choice of α_k involves a single-dimensional minimization. This choice ensures that under appropriate conditions,

$$f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)}).$$

We may, for example, use the secant method to find α_k . In this case we need the derivative of ϕ_k , which is

$$\phi'_k(\alpha) = \mathbf{d}^{(k)\top} \nabla f(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}).$$

This is obtained using the chain rule. Therefore, for applying the secant method for the line search the gradient ∇f , the initial line-search point $\mathbf{x}^{(k)}$, and the search direction $\mathbf{d}^{(k)}$ are required. Of course, other single-dimensional search methods may be used for line search [5,6].

Line-search algorithms used in practice involve considerations that we have not yet discussed. First, determining the value of α_k that exactly minimizes ϕ_k may be computationally demanding; even worse, the minimizer of ϕ_k may not even exist. Second, practical experience suggests that it is better to allocate more computational time on iterating the multidimensional optimization algorithm rather than performing exact line searches. These considerations led to the development of conditions for terminating

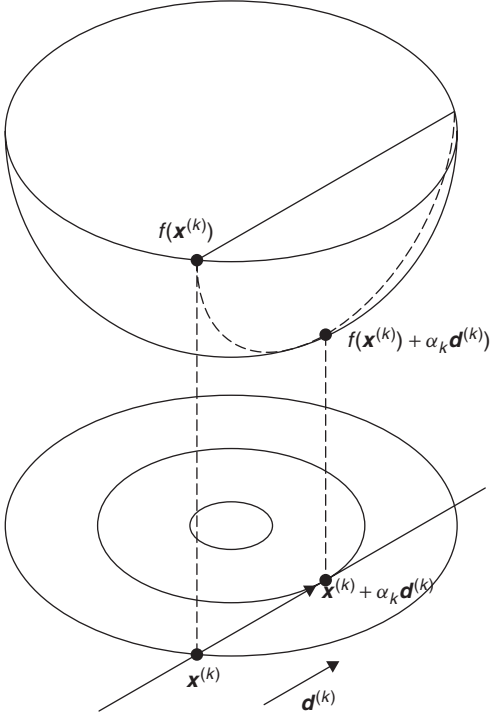


Figure 11. Line search in multidimensional optimization.

line-search algorithms that would result in low-accuracy line searches while still securing a sufficient decrease in the value of the f from one iteration to the next. The basic idea is that we have to ensure that the step size α_k is not too small or too large.

Some commonly used termination conditions are as follows. First, let $\varepsilon \in (0, 1)$, $\gamma > 1$, and $\eta \in (\varepsilon, 1)$ be given constants. The *Armijo condition* ensures that α_k is not too large by requiring that

$$\phi_k(\alpha_k) \leq \phi_k(0) + \varepsilon \alpha_k \phi'_k(0).$$

Further, it ensures that α_k is not too small by requiring that

$$\phi_k(\gamma \alpha_k) \geq \phi_k(0) + \varepsilon \gamma \alpha_k \phi'_k(0).$$

The *Goldstein condition* differs from Armijo in the second inequality:

$$\phi_k(\alpha_k) \geq \phi_k(0) + \eta \alpha_k \phi'_k(0).$$

The first Armijo inequality and the Goldstein condition are often jointly called the *Armijo–Goldstein condition*. The *Wolfe condition* differs from Goldstein in that it involves only ϕ'_k :

$$\phi'_k(\alpha_k) \geq \eta \phi'_k(0).$$

A stronger variation of this is the *strong Wolfe condition*:

$$|\phi'_k(\alpha_k)| \leq \eta |\phi'_k(0)|.$$

A simple practical (inexact) line-search method is the *Armijo backtracking algorithm*, described as follows. We start with some candidate value for the step size α_k . If this candidate value satisfies a prespecified termination condition (usually the first Armijo inequality), then we stop and use it as the step size. Otherwise, we iteratively *decrease* the value by multiplying it by some constant factor $\tau \in (0, 1)$ (typically $\tau = 0.5$) and recheck the termination condition. If $\alpha^{(0)}$ is the initial candidate value, then after m iterations the value obtained is $\alpha_k = \tau^m \alpha^{(0)}$. The algorithm *backtracks* from the initial value until the termination condition holds. In other words, the algorithm produces a value for the step size of the form $\alpha_k = \tau^m \alpha^{(0)}$ with m being the smallest value in $\{0, 1, 2, \dots\}$ for which α_k satisfies the termination condition.

For more information on practical line-search methods, we refer the reader to Fletcher [5, pp. 26–40], Nash and Sofer [7, Section. 10.5], Bertsekas [8, Appendix C], Gill and Murray [9], and Gill *et al.* [10].

Acknowledgments

We thank Dennis M. Goodman for providing us with references [9] and [10].

REFERENCES

1. Chong EKP, Żak SH. An introduction to optimization. 3rd ed. New York: John Wiley and Sons, Inc.; 2008.
2. Bunday BD. Basic optimization methods. London: Edward Arnold; 1984.

3. Wilde DJ. Optimum seeking methods. Englewood Cliffs (NJ): Prentice Hall; 1964.
4. Conte SD, de Boor C. Elementary numerical analysis: an algorithmic approach. 3rd ed. New York: McGraw-Hill Book Co.; 1980.
5. Fletcher R. Practical methods of optimization. 2nd ed. Chichester: Wiley; 1987.
6. Luenberger DG, Ye Y. Linear and nonlinear programming. 3rd ed. New York: Springer Science + Business Media; 2008.
7. Nash SG, Sofer A. Linear and nonlinear programming. New York: McGraw Hill; 1996.
8. Bertsekas DP. Nonlinear programming. 2nd ed. Belmont (MA): Athena Scientific; 1999.
9. Gill PE, Murray W. Safeguarded steplength algorithms for optimization using descent methods. Technical Report NPL NAC 37. Teddington: National Physical Laboratory, Division of Numerical Analysis and Computing. 1974 Aug.
10. Gill PE, Murray W, Saunders MA, *et al.* Two step-length algorithms for numerical optimization. Technical Report SOL 79-25. Stanford (CA): Systems Optimization Laboratory, Department of Operations Research, Stanford University; 1979 Dec.

SINGLE-SEARCH-BASED HEURISTICS FOR MULTIOBJECTIVE OPTIMIZATION

MATTHIAS EHROGOTT
Department of Engineering
Science, The University of
Auckland, Auckland, New
Zealand

INTRODUCTION

Multiobjective optimization (MO) is the subfield of operations research that deals with the optimization of several conflicting objectives over some set of feasible solutions. An MO problem can be formulated as

$$\min \{(z_1(x), \dots, z_p(x)) : x \in X\}, \quad (1)$$

where $X \subset \mathbb{R}^n$ is a feasible set in decision or variable space and $z : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is a vector valued objective function consisting of the individual objective functions $z_k, k = 1, \dots, p$. The outcome set $Z = z(X) := \{z(x) : x \in X\} \subset \mathbb{R}^p$ is the image of the feasible set in objective or outcome space. Multiobjective optimization requires optimization over vectors, and because of the absence of a canonical total order of vectors, a definition is needed. A feasible solution $x \in X$ is called efficient (or Pareto optimal) if there does not exist $x' \in X$ such that $z_k(x') \leq z_k(x)$ for all $k = 1, \dots, p$ and $z_j(x') < z_j(x)$ for some j . If x is efficient, then $z(x) = (z_1(x), \dots, z_p(x))$ is called nondominated (sometimes also efficient). The set of efficient solutions is denoted by X_E , the set of nondominated outcomes is denoted Z_N . Solving an MO problem means finding the set Z_N (and for each $z \in Z_N$ at least one $x \in X_E$ such that $z = z(x)$).

Single-search-based heuristics are often applied to combinatorial optimization problems (see the section titled “Combinatorial Optimization/Integer Programming” of this

encyclopedia). A multiobjective combinatorial optimization (MOCO) problem is formulated as

$$\min \{Cx : Ax \geq b, x \in \mathbb{Z}^n\}. \quad (2)$$

Here C is a $p \times n$ objective function matrix, where c^k denotes the k th row of C . A is an $m \times n$ matrix of constraint coefficients and $b \in \mathbb{R}^m$. Usually the entries of C, A and b are integers and the feasible set $X = \{x : Ax \geq b, x \in \mathbb{Z}^n\}$ describes a combinatorial structure such as, for example, spanning trees of a graph, paths, and matchings, in which case $X \subset \{0, 1\}^n$ and is a finite set.

A heuristic (see the section titled “Heuristics and Meta-heuristics” of this encyclopedia) for an MO problem is an algorithm that computes a set of solutions $S \subset \mathbb{R}^n$ without necessarily guaranteeing that any element of S is feasible or efficient. An element of S is locally potentially efficient (with respect to S) if there is no other element of S dominating it. The heuristic proceeds in iterations and stops as soon as some criterion is met, at which stage it is assumed that the image of the current set of locally potential efficient solutions $z(S_{PE})$ is a good approximation of the nondominated set Z_N . Most multiobjective heuristics follow one of two ideas. They either repeatedly apply a single search to find some potentially efficient solutions with each run of the single search. At the end of the procedure the solution sets are merged to define the final approximation. These heuristics are the topic of this article. Population-based heuristics work on a set of solutions and try to improve the whole set of solutions in every iteration.

An important issue is the quality of the heuristic, i.e., how well the final solution set approximates Z_N . Two main aspects are the distance between the elements of $z(S_{PE})$ and Z_N and the distribution of the elements of $z(S_{PE})$ along Z_N . There is a substantial amount of literature on this topic and further references can be found in Zitzler *et al.* [1]. It

is in general difficult to compare two different sets of potentially nondominated outcomes.

In the following, single-search heuristic approaches for MO are explained, covering in detail simulated annealing, tabu search, and (Pareto) local search. Following these sections, some ideas underlying hybrid heuristics are briefly mentioned. At the end of the article some conclusions and suggestions for further reading are given.

SIMULATED ANNEALING

Simulated annealing is a local search technique that avoids getting trapped in local optima by accepting nonimproving solutions with some probability that decreases with the number of iterations. Most simulated annealing-based heuristics for MO are closely related to the original idea. However, it is always necessary to adapt the algorithm to deal with the notion of efficient solutions when defining the scheme for decreasing the temperature, the rule for accepting a new solution with some probability depending on the temperature, the mechanism for covering the whole nondominated set, and the (possible) use of information drawn from a set of solutions.

An Example of a Multiobjective Simulated Annealing (MOSA) Algorithm

A typical and popular simulated annealing algorithm is MOSA proposed by Ulungu [2].

Algorithm 1. Multiobjective Simulated Annealing (MOSA).

```

Input:  $\Lambda$  // Set of weights
 $T, \alpha, T_{\text{stop}}, N_{\text{step}}, N_{\text{stop}}$  // SA parameters
 $X_{\text{PE}} \leftarrow \emptyset$ 
for all  $\lambda \in \Lambda$  do
   $T_0 \leftarrow T; N_{\text{count}} \leftarrow 0; n \leftarrow 0$ 
  randomly draw  $x^n \in X; X_{\text{PE}\lambda} \leftarrow \{x^n\}$ 
  repeat
    randomly draw  $x \in \mathcal{N}(x^n)$ 
    if  $\text{isBetter}(x, x^n)$  or  $\text{isAccepted}(x, x^n, n, T_n, \lambda)$  then
       $X_{\text{PE}\lambda} \leftarrow \text{archive}(X_{\text{PE}\lambda}, x); x^{n+1} \leftarrow x; N_{\text{count}} \leftarrow 0$ 
    else
       $x^{n+1} \leftarrow x^n; N_{\text{count}}++$ 
    end if
     $n++$ ;  $\text{updateParameters}(\alpha, n, T_n)$ 
  until  $\text{isFinished}(N_{\text{count}}, T_n)$ 
end for

```

It uses a predefined set of weight vectors Λ to define search directions in outcome space. The algorithm starts with an initial, randomly generated solution x^0 and employs a neighborhood structure $\mathcal{N}(x)$. MOSA selects a neighbor $x \in \mathcal{N}(x^n)$ and compares x with x^n as follows: If $z_k(x) < z_k(x^n)$ for all $k = 1, \dots, p$, then all the objectives have improved and x is accepted. If there exist k and k' such that $z_k(x) < z_k(x^n)$ and $z_{k'}(x) > z_{k'}(x^n)$, then x and x^n are incomparable and both are potentially efficient. If $z_k(x) \geq z_k(x^n)$ with at least one strict inequality, then x is dominated by x^n . In the latter two cases, x is accepted with decreasing probability, depending on the current “temperature” of the cooling schedule (procedure `isAccepted`) based on the values $\sum_{k=1}^p \lambda_k z_k(x)$ and $\sum_{k=1}^p \lambda_k z_k(x^n)$. The probability of accepting a solution is

$$\min \left\{ 1, \exp \left(\sum_{k=1}^p - \frac{\lambda_k (z_k(x) - z_k(x^n))}{T_n} \right) \right\}.$$

When a neighbor is accepted, the set of potentially efficient solutions $X_{\text{PE}\lambda}$ in direction λ is updated. The search stops after a certain number of iterations, or when a predetermined temperature is reached (procedure `isFinished`). At the end, MOSA combines the sets of potentially efficient solutions in a single set X_{PE} by merging the sets $X_{\text{PE}\lambda}$ (procedure `merge`) for all $\lambda \in \Lambda$.

```

 $X_{PE} \leftarrow \text{merge}(X_{PE\lambda})$ 
return  $X_{PE}$  // set of potentially efficient solutions

```

Other Multiobjective Simulated Annealing Algorithms

The Pareto simulated annealing (PSA) method of Czyzak and Jaskiewicz [3,4] combines simulated annealing principles with ideas from genetic algorithms. A set $S \subset X$ of solutions is determined and used as initial solutions. Each solution in this set is improved iteratively following the same simulated annealing principle, using some of the acceptance rules of Serafini [5]. For a given solution $\bar{x} \in S$, the weights are changed in order to increase the probability of moving it away from its closest neighbor in S denoted by $\bar{x}'$.

The method of Parks and Suppaitnarm [6] does not use search directions or an aggregation of the objectives. Each objective is considered separately, applying only the non-domination definition to select potentially efficient solutions. It uses a set of solutions to ensure the exploration of the complete nondominated set. Suppaitnarm *et al.* [6] give recommendations about the tuning of the parameters (acceptance rule, annealing schedule).

Suman [7] proposes an algorithm that starts with an initial randomly selected solution and stops if at any value of the temperature no solution has been added to X_{PE} for a predetermined number of iterations. In the iterations, a new neighbor solution is selected randomly. If it is potentially efficient, it becomes the new solution; otherwise it is accepted with probability $\exp(-\Delta/T)$, where T is the temperature and Δ is the difference in fitness between the new and the current solution (measured as 1 plus the number of solutions dominating it). Diversification is achieved by a restart from a randomly selected element of X_{PE} .

Smith *et al.* [8] also avoid the use of a scalarizing function and evaluate the fitness of a solution in a similar way, but divided by the size of X_{PE} . Moreover, they do not restrict the size of the archive; in fact, they allow dominated solutions to remain in the archive. In contrast, Bandyopadhyay *et al.*

[9] keep the archive to a limited size by using clustering. Their algorithm stops when the temperature drops to a predefined level, with a specified number of iterations at each level. Once again, acceptance of a solution is based on dominance and the number of solutions in the archive dominating new solutions.

TABU SEARCH

The idea of tabu search is also based on local search. Tabu search algorithms try to avoid revisiting local minima by forbidding moves that lead to solutions visited before. By keeping a list of such tabu moves, the search is driven toward areas of the feasible set that have not been explored before. Despite this general idea, moves that lead to exceptionally good solutions are usually accepted and have their tabu status overridden. In multiobjective tabu search algorithms, it is necessary to adapt the tabu criterion to deal with multiple objectives, to use diversification of the search to ensure coverage of the nondominated set, and possibly to exploit information obtained from a set of solutions.

An Example of a Multiobjective Tabu Search Algorithm

The algorithm of Gandibleux *et al.* [10] carries out a series of searches in the objective space using a reference-point-based scalarizing function. These searches are guided by the current approximation of the nondominated set. The algorithm also employs two tabu lists: list $TmemX$ in decision space prevents a return to previously visited solutions and $TmemY$ in objective space is used for updating objective weights between consecutive searches.

In iteration n , the algorithm searches the neighborhood $\mathcal{N}(x^n)$ of the current solution x^n (procedure `exploreNeighborhood`). The successor $\bar{x}$ of x^n is selected from the list of neighbor solutions $\mathcal{L} \subset \mathcal{N}(x^n)$ as the best

according to the scalarizing function

$$s(z(x), y^U, \lambda) = \min_{1 \leq k \leq p} \left\{ \lambda_k (z_k(x) - y_k^U) \right\} + \rho \sum_{k=1}^p \lambda_k (z_k(x) - y_k^U),$$

where $\rho > 0$ and λ_k are nonnegative weights. The number of candidates in list $\mathcal{L}$ is limited to K solutions. The value of this parameter depends on the neighborhood size ($1 \leq K \leq |\mathcal{N}(x^n)|$). The reference point y^U is the locally determined utopian point $y^U = (y_1^U, \dots, y_p^U)$ over $\mathcal{L}$, where $y_k^U < \min\{z_k(x) : x \in \mathcal{L}\}$ (procedure `utopian`). This point dominates the ideal point given by the lowest objective function value of each objective among the solutions in the neighborhood of the current solution.

The tabu list $TmemX$ is used to avoid cycling (procedures `isTabu` and `updateTabuMemoryX`). The selected solution $\bar{x} \in \mathcal{L}$, which minimizes $s(z(x), y^U, \lambda)$ over $\mathcal{L}$ such that the move $x^n \rightarrow \bar{x}$ is not tabu, becomes the new current solution x^{n+1} . The *tabu daemon* overrides the tabu status of a solution $x' \in \mathcal{N}(x^n)$, if $s(z(x'), y^U, \lambda) \leq s(z(\bar{x}), y^U, \lambda) - \Delta$, with Δ being a threshold value (procedure `TabuDaemon`). Each iteration ends with the identification of the potentially efficient solutions $X_{PE\mathcal{L}}$ in $\mathcal{L}$, which represents a local approximation of the nondominated frontier (procedure `archive`).

The tabu searches are adjusted according to a comparison of the local set of potentially efficient solutions $X_{PE\mathcal{L}}$ compared to the global approximation X_{PE} . If $\bar{x} \in X_{PE\mathcal{L}}$ seems to be a promising solution (defined by the number of solutions in X_{PE} that are dominated by $\bar{x}$ being greater than a threshold η), the current search continues for more iterations (procedure `isPromising`). If, however, $X_{PE\mathcal{L}}$ does not contribute any solutions to X_{PE} , the current neighborhood is said to be sterile (routine `isSterile`) and the number of iterations in the current search process is decreased.

The scalarizing function leads to the exploration of the nondominated frontier in the direction given by the weight vector λ . The diversification strategy updates the weights by changing the values of λ_k according to the change in objective function $z_k(x) - z_k(\bar{x})$, distinguishing the case of deterioration, indifference, weak or strong improvement. If the objective k is weakly or strongly improved, the weight λ_k is either slightly or strongly decreased, whereas in the other cases the weight is increased (procedure `setSearchDirection`). To ensure that all the areas of the nondominated frontier are visited, an objective space tabu list $TmemY$ is used. In case of weak or strong improvement, objective k is declared tabu for a predetermined number of iterations. Thus, its weight cannot be increased for the duration of the tabu tenure.

Algorithm 2. Multiobjective Tabu Search MOTS.

Input: TTX, TTY, μ // tabu tenure in decision and objective space, threshold
 Δ, K // tabu daemon parameter and size of candidate list
 $IterInit$ // initial number of iterations for a tabu search
 $stopConditions$ // conditions to stop the search
 $TmemX, TmemY \leftarrow \emptyset$
 $n \leftarrow 1$
select $x^n \in X; X_{PE} \leftarrow \{x^n\}$
repeat
 // Diversification in Z
 $\lambda \leftarrow setSearchDirection(TmemY, X_{PE}); Iter \leftarrow IterInit$
 repeat
 // Local exploration of the nondominated set by tabu search
 $\mathcal{L} \leftarrow exploreNeighborhood(x^n); \bar{z} \leftarrow utopian(\mathcal{L})$
 $\bar{x} \leftarrow \operatorname{argmin}_{x \in \mathcal{L}} \{s(z(x), y^U, \lambda) : \neg isTabu(move(x^n \rightarrow x), TmemX)\}$
 $x^* \leftarrow TabuDaemon(\bar{x}, \mathcal{L}, TmemX)$

```

if  $x^* = \emptyset$  then
   $x^{n++} \leftarrow \bar{x}$ 
else
   $x^{n++} \leftarrow x^*$ 
end if
// Adjustment of the intensification and update of the approximation
 $TmemX \leftarrow \text{updateTabuMemory}(\text{move}(x^{n-1} \rightarrow x^n), TmemX)$ 
 $X_{PE\mathcal{L}} \leftarrow \{x^n\}$ 
for all  $x \in \mathcal{L}$  do
   $X_{PE\mathcal{L}} \leftarrow \text{archive}(X_{PE\mathcal{L}}, x)$ 
end for
if  $\text{isPromising}(x^n, X_{PE})$  then
  increase  $Iter$ 
end if
if  $\text{isSterile}(X_{PE\mathcal{L}}, X_{PE})$  then
  decrease  $Iter$ 
end if
 $X_{PE} \leftarrow \text{merge}(X_{PE}, X_{PE\mathcal{L}})$ 
until  $Iter$  is performed
until  $\text{stopConditions}$  are fulfilled
return  $X_{PE}$ 

```

Other Tabu Search Methods

The method of Hansen [11] starts with a randomly selected set of solutions that are dispersed throughout the objective space in order to allow searches in different areas of the nondominated set. Weights are defined for each solution with the aim of forcing the search into a certain direction of the non-dominated set and away from other current solutions that are potentially efficient. The algorithm then starts a tabu search from each of the initial solutions. The algorithm has been applied to the knapsack problem and the resource constrained project scheduling problem.

The algorithm of Armentano and Arroyo [12] starts with a set of potentially efficient solutions, constructs their neighborhoods, and selects potentially efficient solutions from the union of these neighborhoods. A limited number K of these solutions that are not tabu are selected to be included in the set of locally potentially efficient solutions, and the others are stored as candidate solutions for possible exploration. To aid the selection of solutions, they are grouped into K clusters. The algorithm uses a diversification scheme that penalizes frequently used moves.

Choobineh *et al.* [13] start with p randomly selected solutions and conduct p tabu

searches in parallel, one for each objective function.

Kulturel-Konak *et al.* [14] begin with a randomly selected feasible solution. In each iteration, they first determine an objective function following a multinomial probability distribution. Then the candidate solution is found according to the single objective tabu search principle for the selected objective and the list of potentially efficient solutions is updated. If this does not happen for a specified number of iterations, any element of X_{PE} is selected for a restart.

The algorithm of Jaeggi *et al.* [15], designed for continuous problems, starts from a single solution, but adapts the use of short term memory (tabu list), medium term memory (list of potentially efficient solutions), and long term memory (regions of Z that have been explored) to the multiobjective case.

PARETO LOCAL SEARCH

Simple local search heuristics for single objective optimization start from an initial solution, proceed to investigate the neighborhood of the current solution to find the best solution in the neighborhood, and stop as soon as no further improvement can be made. Naturally such algorithms stop with a local

optimum and no attempt is made to escape from it. In MO it is necessary to define a local optimum as a set of solutions. $S \subset X$ is locally efficient if each $x \in S$ is such that x is locally efficient with respect to neighborhood N , that is, if for each $x \in S$ there is no $x' \in N(x)$ such that $z(x')$ dominates $z(x)$ and such that no element of S is dominated by another one, that is, S is a set of potentially efficient solutions, as discussed by Paquete *et al.* [16].

Very few algorithms based on the local search idea with locally efficient sets have been proposed. We present the generic algorithm given by Paquete *et al.* [16]. The algorithm starts with a single solution, maintains

Algorithm 3. Pareto Local Search PLS.

```

Select initial solution  $x$ 
 $X_T \leftarrow \{x\}; X_V \leftarrow \emptyset$ 
repeat
  Select  $x$  in  $X_T$ ;  $X_V \leftarrow X_V \cup \{x\}$ ;  $W \leftarrow \emptyset$ 
  for all  $x' \in N(x)$  do
    if there is no  $x'' \in W \cup X_V \cup X_T$  with  $z(x'') \leq z(x')$  then
       $W \leftarrow W \cup \{x'\}$ 
    end if
   $X_V \leftarrow \{x' \in X_V : \text{there is no } x'' \in W \text{ dominating } x'\}$ 
   $X_T \leftarrow \text{merge}(V_F \setminus \{x\}, W)$ 
end for
until  $X_T = \emptyset$ 
return  $X_{PE} \leftarrow X_V$ 

```

HYBRID METAHEURISTICS

While originally most single-search-based algorithms were devised along the lines of the single objective versions of the algorithmic paradigms of simulated annealing and tabu search with the main difference being whether a scalarization function was applied or whether the neighborhood search was applied to several starting points using an adaptation of the acceptance principle of new solutions to the efficiency principle, the trend in new algorithms then changed to the development of hybrid heuristics. Hybrid metaheuristics combine ideas from single-search-based heuristics and population-based heuristics. Hybrid metaheuristics can be grouped into several categories.

- Algorithms that include local search within a population-based method. An

archive of potentially efficient solutions, and iteratively investigates all neighbors of all solutions in the archive. X_{PE} is divided into X_V , containing solutions whose neighborhood has been explored, and X_T , solutions whose neighborhood remains to be investigated. In each iteration, the algorithm visits all neighbors of a solution x from X_T . If $x' \in N(x)$ is not dominated by any solution in $N(x) \cup X_V \cup X_T$, then it is added to a set W and will be chosen later unless it becomes dominated by some element of $N(x) \cup X_V \cup X_T$.

example is an algorithm of Jaszkievicz [17], which is a combination of a multiobjective genetic algorithm and local search.

- Algorithms managing coverage of the whole nondominated set using a population of solutions. Some of the algorithms mentioned in the sections titled “Other Multiobjective Simulated Annealing Algorithms” and “Other Tabu Search Methods” actually fall under this category.
- Because population-based methods are particularly well suited for managing a varied set of solutions, they are often strong in covering the whole nondominated set, whereas single-search-based methods can provide a better focus on improving individual solutions. Hybrid methods can exploit the complementary strengths of population-based

and single-search-based algorithms. Examples are the algorithms described by Ben Adbelaziz *et al.* [18], combining a genetic algorithm with tabu search, and López-Ibáñez *et al.* [19], combining an ant colony approach with tabu search.

- Approaches combining heuristics with exact methods and constraint programming have been proposed. In general, a trend is emerging toward custom-made heuristics for specific problems that combine various ingredients from a number of heuristics as well as the exact method. Examples include those of Barichard and Hao [20] and Gandibleux and Fréville [21].

CONCLUSION AND SUGGESTIONS FOR FURTHER READING

In this article, the main approaches for single-search-based heuristics for MO have been explained. These are MOSA, MOTS, and PLS. For each approach one complete algorithm has been presented together with brief explanations of other ideas presented in the literature over time. A brief summary of the trend toward hybrid heuristics has been given.

The related literature is very extensive and a few suggestions for further reading are presented now. Jones *et al.* [22] analyzed the literature until 2002. Ehrgott and Gandibleux [23] present a survey of multi-objective metaheuristics up to 2004. In Ref. 24 the same authors cover hybrid heuristics in depth. All of these survey articles contain many references. Books covering the subject include Gandibleux *et al.* [25] and Talbi [26].

REFERENCES

1. Zitzler E, Thiele L, Laumanns M, *et al.* Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Trans Evol Comput* 2003;7:117–132.
2. Ulungu EL, Teghem J, Fortemps P. Heuristics for multi-objective combinatorial optimisation problem by simulated annealing. In: Wei Q, Gu J, Chen G, *et al.* editors, *MCDM: theory and applications*. Windsor, UK: SCI-TECH Information Services; 1995. pp. 228–238.
3. Czyzak P, Jaskiewicz A. Pareto simulated annealing. In: Fandel G, Gal T, editors. *Multiple criteria decision making. Proceedings of the XIIth international conference, Hagen (Germany)*. Volume 448, *Lecture notes in economics and mathematical systems*, Berlin: Springer; 1997. pp. 297–307.
4. Czyzak P, Jaskiewicz A. Pareto simulated annealing—a metaheuristic technique for multiple objective combinatorial optimization. *J Multi-Criteria Decis Anal* 1998;7:34–47.
5. Serafini P. Simulated annealing for multiobjective optimization problems. *Proceedings of the 10th international conference on multiple criteria decision making. Volume I*, Taipei-Taiwan; 1992. pp. 87–96.
6. Suppakitnarm A, Seffen K, Parks G, *et al.* A simulated annealing algorithm for multiobjective optimization. *Engineering Optimization* 2000;33:59–85.
7. Suman B. Study of self-stopping PDMOSA and performance measure in multiobjective optimization. *Comput Chem Eng* 2005;29:1131–1147.
8. Smith KI, Everson RM, Fieldsend JE, *et al.* Dominance-based multiobjective simulated annealing. *IEEE Trans Evol Comput* 2008;12:323–342.
9. Bandyopadhyay S, Saha S, Maulik U, *et al.* A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Trans Evol Comput* 2008;12:269–283.
10. Gandibleux X, Mezdaoui N, Fréville A. A tabu search procedure to solve multiobjective combinatorial optimization problems. In: Caballero R, Ruiz F, Steuer R, editors. *Advances in multiple objective and goal programming. Volume 455, Lecture notes in economics and mathematical systems*, Berlin: Springer; 1997. pp. 291–300.
11. Hansen MP. Tabu search for multiobjective combinatorial optimization: TAMOCO. *Control Cybern* 2000;29:799–818.
12. Armentano VA, Arroyo JEC. An application of a multi-objective tabu search algorithm to a bicriteria flowshop problem. *J Heuristics* 2004;10:463–481.
13. Choobineh FF, Mohebi E, Khoo H. A multi-objective tabu search for a single-machine scheduling problem with sequence-dependent setup times. *Eur J Oper Res* 2006;175:463–481.

14. Kulturel-Konak S, Smith AE, Norman BA. Multi-objective tabu search using a multinomial probability mass function. *Eur J Oper Res* 2006;169:918–931.
15. Jaeggi DM, Parks GT, Jipuras T, *et al.* The development of a multi-objective tabu search algorithm for continuous optimisation problems. *Eur J Oper Res* 1008;185: 1192–1212.
16. Paquete L, Schiavinotto T, Stützle T. On local optima in multiobjective combinatorial optimization problems. *Ann Oper Res* 2007;156:83–97.
17. Jaszkiwicz A. Multiple objective genetic local search algorithm. In: Köksalan M, Zionts S, editors. Multiple criteria decision making in the new millennium. Volume 507, Lecture notes in economics and mathematical systems. Berlin: Springer; 2001. pp. 231–240.
18. Ben Abdelaziz F, Chaouachi J, Krichen S. A hybrid heuristic for multiobjective knapsack problems. In: Voss S, Martello S, Osman I, *et al.* editors, Meta-Heuristics: advances and trends in local search paradigms for optimization, Dordrecht: Kluwer Academic Publishers; 1999. pp. 205–212.
19. López-Ibáñez M, Paquete L, Stützle T. Hybrid population-based algorithms for the biobjective quadratic assignment problem. *J Math Model Algorithms* 2006;5:111–137.
20. Barichard V, Hao JK. A population and interval constraint propagation algorithm. In: Fonseca CM, Fleming PJ, Zitzler E, *et al.* editors, Evolutionary multi-criterion optimization—second international conference, EMO 2003, Faro, Portugal, April 8–11, 2003. Proceedings. Volume 2632, Lecture notes in computer science. Berlin: Springer; 2004. pp. 88–101.
21. Gandibleux X, Fréville A. Tabu search based procedure for solving the 0/1 multiobjective knapsack problem: the two objective case. *J Heuristics* 2000;6:361–383.
22. Jones D, Mirrazavi SK, Tamiz M. Multi-objective meta-heuristics: An overview of the current state-of-the-art. *Eur J Oper Res* 2002;137:1–9.
23. Ehrgott M, Gandibleux X. Approximative solution methods for multiobjective combinatorial optimization. *TOP* 2004;12:1–88.
24. Ehrgott M, Gandibleux X. Hybrid metaheuristics for multi-objective combinatorial optimization. In: Blum C, Blesa-Aguilera MJ, Roli A, *et al.* editors, Hybrid metaheuristics. Volume 114, Studies in computational intelligence, Berlin: Springer; 2008. pp. 221–259.
25. Gandibleux X, Sevaux M, Sörensen K *et al.*, editors. Metaheuristics for multiobjective optimisation. Volume 535, Lecture notes in economics and mathematical systems. Berlin: Springer; 2004.
26. Talbi EG. Metaheuristics: from design to implementation. John Wiley & Sons; 2009.

SLOVAK SOCIETY FOR OPERATIONS RESEARCH

ZLATICA IVANICOVA
Department of Operations Research and
Econometrics, Faculty of Economic
Informatics, University of Economics,
Bratislava, Slovak Republic

Slovak Society for Operations Research (Slovenská spoločnosť pre operačný výskum) was established in 1993 after splitting the Czechoslovak Society for Operations Research into the Slovak and Czech Societies. These days, the society has 75 members of which 14 are guests from the Czech Society for Operations Research.

The aims of the Slovak Society for Operations Research are as follows:

1. *To Support the Development of Operations Research as a Scientific Discipline.* For this purpose, the Society organizes meetings, conferences, and seminars; and supports publishing scientific and applied results in operations research. It organizes scientific study trips of its members at institutions, the work of which is oriented to the theoretical development or applications of operations research both in the Slovak republic and abroad.
2. *To Support Educational and Enlightenment Activities Devoted to the Area of Operations Research.* For this purpose, the Society organizes lectures and panel discussions for people working in the management sphere of industry, commerce, and agriculture; in the central management institutions; and education with the aim of informing them about the possibilities of applications of operations research in the spheres of their professional activity. It carries out the advisory activity concerning the possibilities of gaining qualification in operations research in all public and private institutions. The Society

is the custodian of the standardization of operations research terminology in Slovak languages.

3. *To Inform Others about Useful Applications of Operations Research Methods.* For this purpose, the Society provides free-of-charge orientation consultations for private firms and public institutions. The consultations are provided by those members of the Society who have the best professional abilities for the successful solution of the given problem. The Society supports the consultation activity in operations research, which is carried out on a commercial basis, and gives objective information about its level.
4. *To Organize and Coordinate Contacts with Partner Institutions of a Similar Type Abroad.* For this purpose, the Society supports within the framework of its budget, the participation of its members at specialized conferences abroad and organizes the working trips of foreign operations research specialists to the Slovak republic.

The history of individuals who served as presidents of the Slovak Society for Operations Research after splitting from the Czech Society for Operations Research is as follows:

1993–1995	Prof. Ing. Jozef Sojka, CSc
1995–1997	Asst. Prof. Ing. Michal Fendek, CSc
1997–1999	Asst. Prof. Ing. Vladimír Mlynarovič, CSc
1999–2001	Prof. Ing. Zlatica Ivaničová, Ph. D
2001–2004	Prof. Ing. Mgr. Jaroslav Husár, CSc
2004–2007	Asst. Prof. Ing. Ivan Brezina, Ph. D
2007–2010	Prof. Ing. Michal Fendek, Ph. D

Current Representatives of Slovak Society for Operations Research

President: Prof. Ing. Michal Fendek, Ph. D.

Vice-President: Prof. Ing. Zlatica Ivaničová, Ph. D.

Secretary: Ing. Marian Reiff, Ph. D.

Treasurer: Ing. Kvetoslava Surmanová, Ph. D.

ACTIVITIES OF THE SLOVAK SOCIETY FOR OPERATIONS RESEARCH

International Workshop on Quantitative Methods in Science and Practice

This workshop is organized every second year for the members of the Slovak Society for Operations Research and representatives from the Czech Republic (namely, University of Economics, Prague and Technical University in Ostrava). It deals with study programs oriented toward operations research for doctoral, graduate, and undergraduate studies. It is important to note that the basic ideas of the lectures and seminars aimed at operations research were established at the Department of Economic and Mathematical Methods in the year 1961, and later known as the Department of Operations Research and Econometrics, at the University of Economics in Bratislava. Professors of the Slovak Society for Operations Research are, in principle, the supervisors of the lectures and seminars focused on the operations research offered in other universities in Slovakia. Besides the University of Economics in Bratislava, aspects of operations research are also taught at Comenius University in Bratislava, Slovak Agricultural University in Nitra, and University of Zilina, and partly at the technical universities in Bratislava and Košice. Subjects of operations research are modified according to the specific study programs of the above-mentioned universities.

International Conferences on Quantitative Methods in Economics

The Slovak Society for Operations Research organizes the international conferences on “Quantitative Methods in Economics” (multiple criteria decision making), every

second year. Presentations are focused on the methodology and construction of the models and applications of the methods of operations research and econometrics. We would like to stress that each conference is specially devoted to the actual problems of economic policy, on macro and micro levels, solved by models and methods of operations research and econometrics.

The first six international conferences were organized under the Czechoslovak Society for Operations Research. After splitting, the Slovak Society for Operations Research organized the following conferences:

- 14th International Conference—High Tatras, Slovakia 2008
Quantitative Methods in Economics (Multiple criteria decision making XIV)
- 13th International Conference—Bratislava, Slovakia 2006
Quantitative Methods in Economics (Multiple criteria decision making XIII)
- 12th International Conference—Vrútky, Slovakia 2004
Quantitative Methods in Economics (Multiple criteria decision making XII)
- 11th International Conference—Nitra, Slovakia 2002
Quantitative Methods in Economics (Multiple criteria decision making XI)
- 10th International Conference—High Tatras, Slovakia 2000
Quantitative Methods in Economics (Multiple criteria decision making X)
- 9th International Conference—Bratislava, Slovakia 1998
Quantitative Methods in Economics (Multiple criteria decision making IX)
- 8th International Conference—Bratislava, Slovakia 1996
Quantitative Methods in Economics (Multiple criteria decision making VIII)
- 7th International Conference—Bratislava, Slovakia 1994
Quantitative Methods in Economics (Multiple criteria decision making VII).

Proceedings of all international conferences on “Quantitative Methods in Economics” are on the web page <http://www.fhi.sk/en/katedry-fakulty/kove/ssov/papers/>.

SLOVENIAN SOCIETY INFORMATIKA (SSI)–SECTION FOR OPERATIONS RESEARCH (SOR)

LIDIJA ZADNIK STIRN
Biotechnical Faculty,
University of Ljubljana,
Ljubljana, Slovenia

GENERAL DATA

The Slovenian Society INFORMATIKA—Section for Operations Research (SSI-SOR) is a nonprofit professional association of experts and researchers in the field of operations research. The SSI-SOR has its headquarters in Ljubljana, Slovenia. Its postal address is: Vožarski pot 12, 1000 Ljubljana, Slovenia, Phone: +386124153440, Fax: +38612415344, e-mail: sdrobne@fgg.uni-lj.si (Mr. Samo Drobne, Secretary), web page <http://www.drustvo-informatika.si/sekcije/sor>. The President of SSI is Mr. Niko Schlamberger, Slovenian Society INFORMATIKA, Vožarski pot 12, 1000 Ljubljana, Slovenia, Phone: +38641735054; Fax: 38612415344, e-mail: niko.schlamberger@gmail.com. The President of SSI-SOR is Prof. Dr Lidija Zadnik Stirn, University of Ljubljana, Biotechnical Faculty, Chair of Applied Mathematics, Večna pot 83, 1000 Ljubljana, Slovenia, Phone: +386 13205003, Fax: +386 1 2571169, e-mail: lidija.zadnik@bf.uni-lj.si. The secretary of SSI-SOR is Mr. Samo Drobne, M. Sc., University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova 2, 1000 Ljubljana, Slovenia, Phone: +386 1 4768649, Fax: +386 1 4250681, e-mail: sdrobne@fgg.uni-lj.si.

The SSI regularly edits the newsletter *Uporabna informatika* (in Slovene language), which is also at the web page [1]: <http://www.drustvo-informatika.si/posta>.

GENERAL FACTS ABOUT SOR

Till the end of 1991, Slovenian researchers and experts from the field of operations

research were active members of the Yugoslav Operations Research Society [2,3].

The Slovenian Section (Chapter) for Operations Research (SOR) at the Slovenian Society INFORMATIKA (SSI) was established in the new state of the Republic of Slovenia in December 1992. The leading persons in the Constitution Committee of SOR were Janez Grad, Viljem Rupnik, Janez Barle, Ludvik Bogataj, Marija Bogataj, Lidija Zadnik Stirn, Samo Drobne, Stane Indihar, Vesna Omladič, and Ivan Meško. There were nearly 40 founding members of the SOR. The first President of SOR was Prof. Ddr. Viljem Rupnik (1992–1997). Mr. Samo Drobne, M.Sc. (1992–) was elected Secretary General of SOR.

The SOR group today consists of 77 members, mostly from universities, (research) institutes, as well as management (firms) in Slovenia. The goal of the group is to pursue, support, and facilitate research, development, application, and education in the operations research area where mathematics, economics, computer science, statistics, environmental economics, and system theory, as well as some other disciplines, come together. Therefore, the interdisciplinary and applied science nature of OR is one of the main concerns of the SOR group. Thus, SOR is a forum for scientists and practitioners from all areas of operations research and allied fields, across the disciplines and programs in resource management, networks, tools, information, and education. In the last few years, its main concern has been to increase its visibility and influence, especially for closer collaboration with educational institutions (universities) and praxis (industry), international collaboration, and publicity.

SOR is managed by an Executive Board elected from the members of the section (chapter). The body has 22 members, of which one is the President of SOR (since 1997, Prof. Dr. Lidija Zadnik Stirn has been the President of SOR), one Vice-President, one Secretary (Mr. Samo Drobne, since 1992), and one Vice-Secretary. SOR has

five honorary members. Operation of SOR is regulated by SSI statute and bylaws. In 1999, the SSI code of professional ethics was proposed and passed at an Extraordinary General Assembly. The operation of SOR is monitored by the SSI Supervisory Board.

Since the inception of SOR, its members have been active in several fields. Among the most important activities is the organization of 10 international symposia on operations research, known as SOR'93, SOR'94, SOR'95, SOR'97, SOR'99, SOR'01, SOR'03, SOR'05, SOR'07, and the jubilee one, the 10th International Symposium on Operational Research SOR'09 [4]. At first, these symposia were organized annually, and later, after an agreement was reached with the Croatian Operational Research Society, they are organized biannually, each year in one country, in Slovenia and Croatia. Operations research symposia in Slovenia consist of plenary sessions, where

several (usually six) prominent scientists present invited papers and about 60 contributed papers in sessions on networks, stochastic and combinatorial optimization, algorithms, multicriteria decision making, scheduling and control, location theory and transport, environment and human resource management, duration models, finance and investment, production and inventory, education and statistics, and OR communications. The presented papers are reviewed by two independent reviewers according to international standards and are published in the form of proceedings. The editors, leading members of SOR, are proud of all 10 SOR Proceedings, especially the last six, which are cited in international publications, such as INSPEC, MathSci, [see the list of SSI-SOR publications below (Fig. 1)].

Further, the members of SOR have written three monographs, two of which,

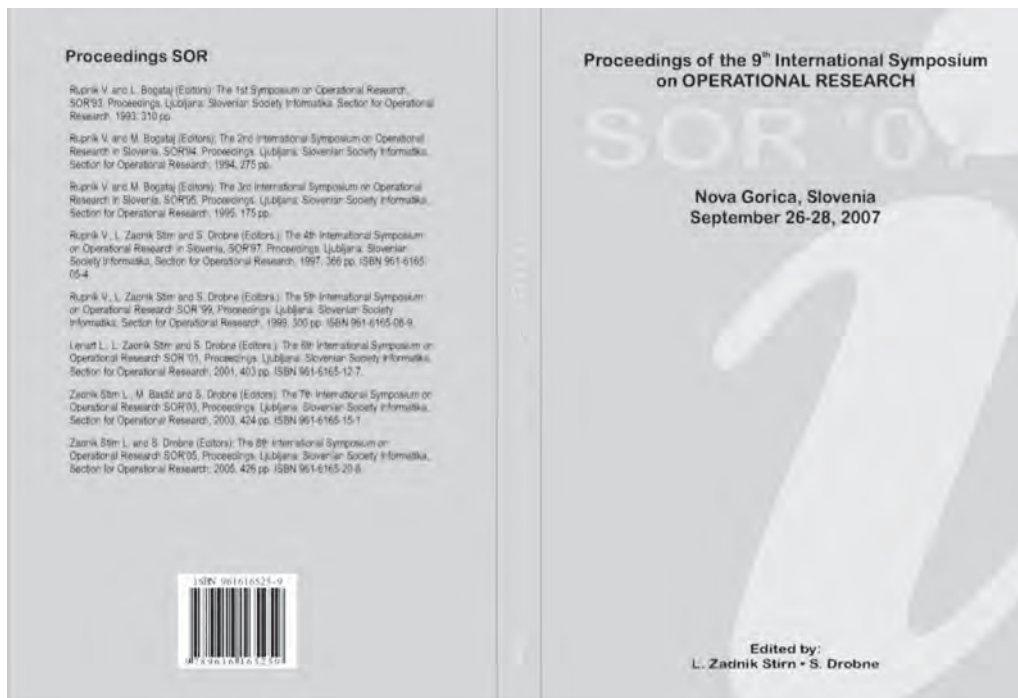


Figure 1. Cover page of the *Proceedings of the Ninth International Symposium on Operational Research* in Slovenia, SOR'07.

titled *Solutions to Production Problems*, and *Selected Decision Support Models for Production and Public Policy Problems*, are in English [5].

Since 1997, the members of SOR are coeditors of the *Central European Journal of Operations Research* (CEJOR), which is published by Physica-Verlag [6]. In the year 2010, they will be bringing out a special issue of CEJOR devoted to 10 highly professional international symposia organized by SOR.

The members of SOR are also coorganizers of annual meetings in informatics, which take place in Slovenia, and of the biannually organized symposia on Operations Research in Croatia; they actively participate in the INFORMS, EURO, IFORS, KOI, and GOR conferences.

SOR is the 47th member of IFORS, admitted to IFORS at the end of 2007 [7].

Finally, SOR became the 30th member of EURO on July 13, 2008. SOR was admitted to EURO in Sandton during the IFORS conference.

SOR PUBLICATIONS

Proceedings

The SOR published 10 proceedings, which present reviewed articles from the International Operational Research symposia organized by SOR in Slovenia. All the articles in the proceedings are in English [5]. These proceedings are given below:

- Proceedings of the First International Symposium on Operational Research, Slovenia, SOR'93. Edited by V. Rupnik and L. Bogataj, SSI-SOR, Ljubljana 1993, pp. 300.
- Proceedings of the Second International Symposium on Operational Research, Slovenia, SOR'94. Edited by V. Rupnik and M. Bogataj, SSI-SOR, Ljubljana 1994, pp. 280.
- Proceedings of the Third International Symposium on Operational Research, Slovenia, SOR'95. Edited by V. Rupnik and M. Bogataj, SSI-SOR, Ljubljana 1995, pp. 170.

- Proceedings of the Fourth International Symposium on Operational Research, SOR'97. Edited by V. Rupnik, L. Zadnik Stirn, and S. Drobne, SSI-SOR, Ljubljana, 1997, pp. 300.
- Proceedings of the Fifth International Symposium on Operational Research, SOR'99. Edited by V. Rupnik, L. Zadnik Stirn, and S. Drobne, SSI-SOR, Ljubljana, 1999, pp. 370.
- Proceedings of the Sixth International Symposium on Operational Research, SOR'01. Edited by L. Lenart, L. Zadnik Stirn, and S. Drobne, SSI-SOR, Ljubljana, 2001, pp. 403.
- Proceedings of the Seventh International Symposium on Operational Research, SOR'03. Edited by L. Zadnik Stirn, M. Bastiš, and S. Drobne, SSI-SOR, Ljubljana, 2003, pp. 424.
- Proceedings of the Eighth International Symposium on Operational Research, SOR'05. Edited by L. Zadnik Stirn and S. Drobne, SSI-SOR, Ljubljana, 2005, pp. 425.
- Proceedings of the Ninth International Symposium on Operational Research, SOR'07. Edited by L. Zadnik Stirn and S. Drobne, SSI-SOR 2007, pp. 460 (the cover page is presented in Fig. 1).
- Proceedings of the 10th International Symposium on Operational Research, SOR'09. Edited by L. Zadnik Stirn, J. Žerovnik, S. Drobne, and A. Lisec, SSI-SOR 2009, pp. 584.

Monographs

Most of the members of SOR are attached to universities or institutes. Thus, the members of SOR are mostly researchers in the field of operations research or related fields. Some of them, together with their colleagues from abroad, mostly from Croatia, have compiled their work in the form of monographs. Since 1998, SOR has published three monographs:

- V. Rupnik. *Teorija faktorjev integrabilnosti gospodarstva in njihovo praktično*

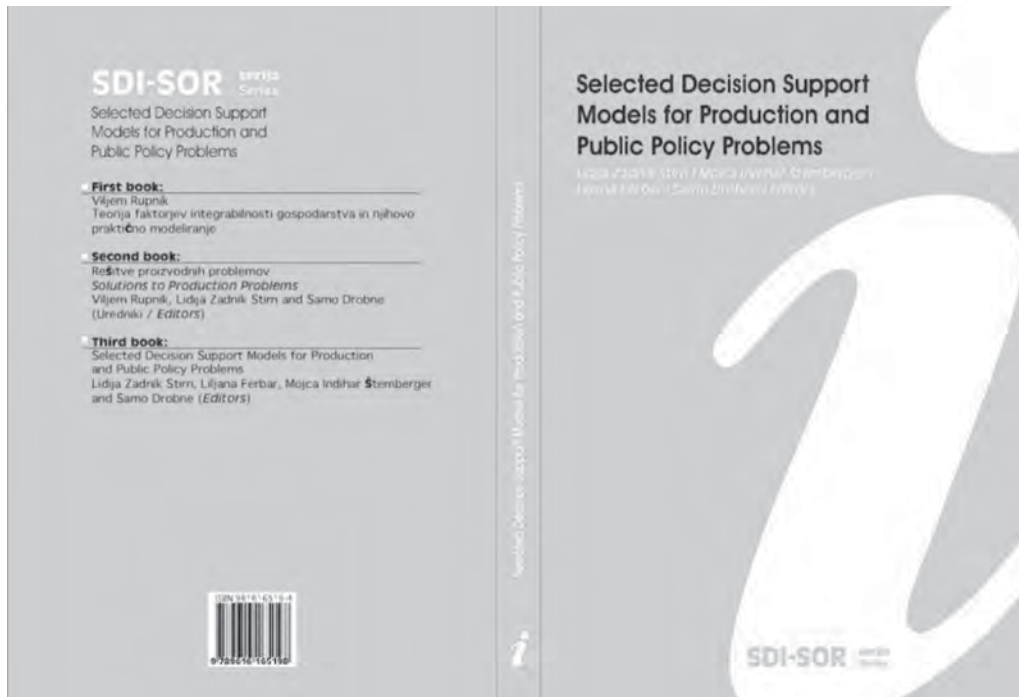


Figure 2. Cover page of the third book in SDI-SOR series (monograph): *Selected Decision Support Models for Production and Public Policy Problems*.

modeliranje. SSI-SOR, Ljubljana, 1998, pp. 720 (in slovene language).

- *Solutions to Production Problems*. Edited by V. Rupnik, L. Zadnik Stirn, and S. Drobne. SSI-SOR, Ljubljana, 2000, ISBN 961–6165–10–0, pp. 290 (15 chapters in Slovene language, 12 chapters in English).
- *Selected Decision Support Models for Production and Public Policy Problems* (Fig. 2). Edited by L. Zadnik Stirn, L. Ferbar, M. Indihar Štemberger, and S. Drobne. SSI-SOR, Ljubljana 2005 ISBN 961–6165–19–4, pp. 243 (in English).

SOME FACTS ABOUT SSI

The Slovenian Society INFORMATIKA (SSI) was established in 1976 [8]. It is a non-profit association of individuals who work in various areas of information technology, information sciences in business, operations

research, management science, and similar fields in universities, institutes, firms, and administration and who, among other concerns, care to exchange their experience and findings among themselves as well as make it publicly available. SSI members are citizens of Slovenia. In 2006, it was officially recognized as a society that operates in public interest.

SSI is managed by an Executive Board elected from among the members of the Society. The body has 12 members, 4 of whom are the President (for the next four years, Mr. Niko Schlamberger again), President Honorary, and two Vice-Presidents (for next four years it is again Prof. Dr. Lidija Zadnik Stirn and Prof. Dr. Marko Bajec). The operation of the Society is regulated by its statutes and bylaws. In 1999, a Code of Professional Ethics had been proposed and passed at an Extraordinary General Assembly. The operation of SSI is monitored by a Supervisory Board.

The main areas of activities to support the SSI mission are increasing its visibility and influence; closer collaboration with educational institutions, government, and industry; and international activities. The Society has established five chapters headed by Chapter Presidents: Information Systems Research, Operational Research (SOR—established in 1992), Language, History of Computing, and, the most recent, Seniors, which was established in 2007. As of 2000, it is the holder of ECDL (European Computer Driving Licence) in the Republic of Slovenia. International activities are subject to SSI affiliations, which are CEPIS (Council of European professional informatics societies) (1998), IFIP (International federation for information processing) (1998), ECDL Foundation 2000, IT STAR (2001), IFORS (2007), and EURO (2008).

Since 1994, SSI has been awarding recognitions for contributions to computer and information science and for improvement of its work and visibility. So far 44 recognitions have been awarded, of which three in 2008 were for individual achievements, but the recipients may be corporate entities as well. The award ceremony is public and normally takes place during the opening ceremony of the SSI annual conferences.

SSI runs regular and special publications. The regular ones are a quarterly professional journal *Uporabna informatika* (Applied Informatics) in Slovenian language where the Slovenian translation of the Bangemann Report was published, and a scientific periodical *Informatica* in the English language [9]. Special publications are proceedings of conferences and special issues of the journals. Special publications are proceedings, monographs (especially are to be stressed those of SOR, mentioned above), and others.

The main event is the annual open conference Dnevi slovenske informatike (The Days of Slovenian Informatics). It started in 1993 and has since earned the recognition as the most comprehensive event of the year in the field of information sciences in Slovenia. It is a national conference with international participation, and every year noted local and foreign guest speakers are

invited. Other events are international scientific biannual conferences organized by the Chapter of Operational Research (SOR) and other conferences and meetings organized in collaboration with various entities.

The SSI home pages are accessible at www.drustvo-informatika.si.

Publications of SSI

The main publication of SSI is the journal *Uporabna informatika* (Applied Informatics). The journal was started in 1992 and it is the only Slovenian professional informatics journal. Occasionally, there are also papers from foreign authors in the English language. Yearly, there are four regular issues and in the last few years there have been a few special issues each year. Its mission is to inform the professional public and users with up-to-date achievements in informatics in Slovenia and worldwide. A special merit of the journal is its presentation of Slovenian projects and achievement. The journal also informs professionals and the general public on European developments and on important documents, which are the basis of trends in informatics in the European Union and inevitably influence our environment. *Uporabna informatika* was the first to publish the Slovenian translation of the Bangemann Report. It has published the European Ministerial Declaration; the content of a special issue was the Slovenian translation of *The White Book on Information Society in Europe*. A special issue in 2000 titled *The Blue Book—Slovenia as an Information Society* has been a major contribution to the respective national strategy. The journal is edited by the Editorial Board appointed by the Executive Board of the Society from among notable Slovenian professionals—members of the Society. All the articles are reviewed and evaluated regarding their suitability for publication to maintain professional quality and editorial standards. The journal has regular columns: editorial, reports, reviews, solutions, new technologies, and events. The response the editors receive from the readers is an indication that the editorial policy matches the expectations of the readers. The journal

is distributed freely to the members of the Society.

Further, SSI regularly publishes the scientific journal *Informatica*. This journal was started in 1976 as a quarterly international publication for computer sciences and informatics. It is an interdisciplinary journal, the editorial policy of which is oriented toward possibilities of finding and establishing multidisciplinary approaches in newly appearing research and technology niches. The journal has the status of a reference publication in international citation databases such as SCI, SSCI, and TCI. The activity of the Editorial Board indicates an orientation toward increase in the value of the citation indices. Such orientation is strongly supported by the society, as *Informatica* is the only Slovenian scientific journal for informatics that has international recognition. The Editorial Board has set up a worldwide network of editors and reviewers who communicate mainly via electronic mail. In the Board there are about 50 editors and numerous reviewers from about 20 countries all over the world. Occasionally, a special issue is brought out covering a particular topic. Individual numbers are edited by specific domains and columns. The journal covers the domains of computer sciences and informatics, which provide a framework for new fields of research such as artificial intelligence, information theory, robotics, cybernetics and theory of systems, quantum theory of information, criticism of the artificial intelligence, new formal systems, theory of chaos, and others. It also presents reports on strategic projects on leading-edge technologies, on research on information and language, on electronic dictionaries, on archives of knowledge, on microprocessor technology, and similar themes.

For professional Slovene language, that is, the Slovene terminology in informatics and operations research, the publication *Islovar*

is indispensable and of great importance. The Internet interactive IT terminology dictionary (*Islovar*; www.islovar.si) started in 2000 and has been improved continually since. The dictionary is well visited (on average >5000 visits per month) and is recognized as a reference dictionary for government public tenders. There are close to 5000 entries, most of which are edited and reviewed. *Islovar* is open and available at no cost. The plan is to enhance it with the pronunciation.

REFERENCES

1. <http://www.uporabna-informatika.si/>. Accessed 2010 June 18.
2. Janez B, Janez G. Sekcija za operacijske raziskave (Section for Operations Research). In: Rupnik V, editor. Proceedings of the 1st International Symposium on Operational Research. Ljubljana: SSI-SOR; 1993.
3. Rupnik V. Ob 30 letnici operacijskih raziskav v sloveniji (30 years of Operations Research in Slovenia). In: Rupnik V, editor. Proceedings of the 10th International Symposium on Operational Research. Ljubljana: SSI-SOR; 2009.
4. Stirn LZ. On occasion of the 10th OR symposium in Slovenia. In: Stirn LZ, Zerovnik J, Drobne S, Lisee A, editors. Proceedings of the 10th International Symposium on Operational Research. Ljubljana: SSI-SOR; 2009.
5. <http://www.drustvo-informatika.si/sekcije/sor/publikacije>. Accessed 2010 June 18.
6. <http://www.springer.com/business+&+management/operations+research/journal/10100>. Accessed 2010 June 18.
7. http://ifors.org/newsletter/IFORS_Sept08.pdf. Accessed 2010 June 18.
8. Schlamberger N, drustvu O. <http://www.drustvo-informatika.si/drustvo/o-drustvu/>. Accessed 2010 June 18.
9. <http://www.drustvo-informatika.si/publikacije/revija-informatika>. Accessed 2010 June 18.

SOCCER/WORLD FOOTBALL

DAVID R. BRILLINGER
Statistics Department,
University of California,
Berkeley, California

Soccer/world football, also known as *association*, is a sport involving two teams of players kicking and heading a medium-sized ball about on a large rectangular field. Each team has a goal that is quite wide and high and seeks to shoot the ball into the other team's goal. The basic objective is to see which team of the two can score the most goals and thereby win the game. The sport is subject to laws set down by the Fédération Internationale de Football Association (FIFA). Typically, there is a group of teams playing against each other in a league or tournament, with a formal schedule of games. A team's intention then is to win a tournament or championship by winning the final game or scoring the most points, where 3 points are awarded for a win, 1 for a tie, and 0 for a loss.

The sport of soccer has a long and full history possibly going back to 2500 BC [1]. In the present time, the teams each have at most 11 players. One, the goalie, may handle the ball but only within a restricted area in front of the goal. There is a neutral referee whose decisions are final. The grandest tournament, the World Cup, moves amongst different countries and occurs nominally every 4 years. The past world champions are Uruguay, Argentina, Brazil, England, France, Germany, and Italy. The sport has become a multibillion dollar business.

Tactics are basic to the games, with many styles and plating formations available. A team's choice has been based, in part, on studies of match statistics. The pioneer in the field of match performance analysis is Charles Reep [2–4].

DATA COLLECTION AND DESCRIPTIVE ANALYSES

Data collection and analysis have been basic to soccer for many years. The most common statistics are a game's result and the number of goals scored by each team. Outputs of the data analyses include reports, tables, graphs, talks, images, and videos.

Many things go on during a game. A basic question is what is to be recorded for the analysis and how to do so. Charles Reep [2,3] developed solutions starting in the early 1950s. The study of basic events in a game has been called both *match performance analysis* and *match analysis* [2,5–7].

A recent addition to the type of information available for individual games is near-continuous high frequency digital spatial-temporal data. Companies come to a stadium and set up an array of video cameras. By signal processing, they then develop the spatial-temporal coordinates of the changing locations of the players on the field, the ball, the linesmen and the referee. Companies doing this include Match Analysis, Prozone Holdings, and Sport-Universal SA. Di Salvo *et al.* [8] have validated one such system. The basic quantities that may now be tabled directly include counts of tackles, crosses, distances covered, key moments and actions, possessions, passes, interceptions, runs with the ball, incorrect referee decisions, fouls, penalty kicks, challenges, entries into the opponent's area, ball touches, blocks, forward passes, long balls, high intensity running, ball velocity and accuracy, and balls received. Individual player's performances may be evaluated directly. The data can lead to changes in tactics during a game [9,10]. Substantial soccer databases have been available for many years. Websites include www.uefa.com, www.fifa.com [11,12], www.soccerbase.com, www.soccerway.com, and www.soccerpunter.com.

One early investigation of basic data is that of Moroney ([13], pp. 101–103). He studied the numbers of goals scored in 480 games and prepared a histogram of the counts. In another early study, the number of successful passes in passing movements were studied for the English First Division during the years 1957–1958 and 1961–1962 [14]. The results were presented in tabular form. Reep and Benjamin found a stable correspondence between shots on goal and goals scored of about 10 to 1. Distances traveled by individual players during a game are graphed by player position in Ref. 5. Hughes and Franks [15] present scatter and time series plots. One early discovery was the existence of a home advantage. Schedules are often set up so that the two teams meet both, at home and away to deal with this advantage. Jochems [16,17], while addressing the question of whether the Dutch football pools were breaking the gambling laws, found the percentages to be 46.6 for a home win, 31.0 for a visitor win, and 22.4 for a draw/tie. Managers make use of such data in their decision-making, more so every year. Stochastic models are important in this connection.

STOCHASTIC MODELING

Stochastic models are pertinent to soccer statistics, because in part there is much uncertainty in what happens and what may happen. During the preceding 50 years, stochastic models have been constructed to address soccer questions.

Between Game Modeling

Some models may be distinguished as to concerning final goals, win–tie–loss, or points. Specific distributions and models that have been employed include bivariate Poisson, exponential, extreme value, GARCH, generalized linear, logistic, Markov, negative binomial, ordinal, regression, prior, point process, and state space. Many analyses are based on the Poisson distribution whose probability function is given by

$$\text{Prob}\{Y = y\} = \mu^y e^{-\mu} / y! \text{ for } y = 0, 1, 2, \dots$$

Here Y might represent the number of goals a team scores in a future game. In an early work, Moroney [13], using the data mentioned above, compared the number of goals scored by a team in a game to their estimated expected frequency assuming that a Poisson distribution held. On examining the result, he was led to seek to improve it and fit a negative binomial distribution. This is a generalization of the Poisson. This fit was found satisfactory. However, Greenhough *et al.* [18] found that when data from 169 countries were pooled, extreme value distributions were needed.

Reep *et al.* [14,19] work on the number of passes in successful passing movements. They fit the negative binomial and found it was good when 0-length cases were excluded.

Suppose that team i at home is playing against team j visiting. Denote the final score by (X_{ij}, Y_{ij}) with X referring to the home team. Various authors have assumed that X_{ij} and Y_{ij} are independent Poissons with respective means,

$\alpha_i \beta_j, \gamma_i \delta_j$	Ref. 20
$\exp\{\alpha + \eta + \gamma_i + \delta_j\},$ $\exp\{\alpha + \gamma_j + \delta_i\}$	Refs 21 and 22
$\alpha_i \beta_j \eta, \alpha_j \beta_i$	Ref. 23
$\exp\{\alpha + \eta + \gamma_i - \delta_j - \phi$ $X(\varepsilon_i - \varepsilon_j)\}, \exp\{\alpha + \gamma_j - \delta_i$ $+ \phi(\varepsilon_i - \varepsilon_j)\}$	Ref. 24

here η is the home advantage, ϕ a psychological effect, and $\varepsilon_i = \gamma_i + \delta_i$. In their fitting, Dixon and Coles modify the low score model probabilities. The probability that team i wins the game may be estimated having estimated the model parameters.

Maher [20] estimates $\alpha, \beta, \gamma, \delta$ by maximum likelihood and mentions fitting the bivariate Poisson. Karlis and Ntzoufras [25] give details of fitting a bivariate Poisson, studying the data for 24 leagues. They find that the assumption of independence is not rejected in 15 out of the 24 cases. They also consider the negative binomial distribution and the inclusion of interaction terms.

The goal difference $X_{ij} - Y_{ij}$ is also important. It determines whether a game result is

win, tie or loss (W, T , or L) and is used to resolve conflicts when the points are equal. Stefani [26] considers the model

$$X_{ij} - Y_{ij} = \eta + \rho_i - \rho_j + \text{random error}$$

with η the home advantage and ρ_i the rankings. Karlis and Ntzoufras [27] fit the Poisson difference distribution to observed goal differences including a home effect, attacking, and defensive parameters.

Fahrmeir and Tutz [28] employ a logistic latent variable-cutpoint setup to model an ordinal-valued variable. Their model is

$$\begin{aligned} \text{Prob}\{Y_{ij} = r\} &= F(\theta_r + \alpha_i - \alpha_j) \\ &\quad - F(\theta_{r-1} + \alpha_i - \alpha_j) \\ &\text{for } r = W, T, L, \end{aligned}$$

with F the logistic and α_i the ability of team i . In contrast, Brillinger [29–33] employs an extreme value variable-cutpoint approach. It leads to the model

$$\begin{aligned} \text{Prob}\{i \text{ wins at home playing } j\} &= 1 - \exp\{-\exp\{\beta_i + \gamma_j + \theta_2\}\}, \\ \text{Prob}\{i \text{ ties at home against } j\} &= \exp\{-\exp\{\beta_i + \gamma_j + \theta_2\}\} \\ &\quad \times \exp\{-\exp\{\beta_i + \gamma_j + \theta_1\}\}, \\ \text{Prob}\{i \text{ loses at home against } j\} &= \exp\{-\exp\{\beta_i + \gamma_j + \theta_1\}\}, \end{aligned}$$

with $\theta_2 > \theta_1$ cutpoints and the standardizations $\sum \beta_i, \sum \gamma_i = 0$. The parameters β_i and γ_i represent home and away effects of team i . The extreme value distribution employed is longer tailed to the right. The result of the preceding game and the distances between the cities involved were considered as explanatories. The $\beta_i - \gamma_i$ can be interpreted as the advantage of team i when playing at home. These values were also considered in Ref. 34. These authors model the goal differences as

$$X_{ij} - Y_{ij} = \eta + \alpha_i + \beta_j + \text{random error},$$

calling the α_i and β_j “forces.” Least squares estimation is employed. Including past results did not improve things. Goodard [35] fits an ordered probit model, that is, F above relates to the normal, with explanatories.

In a linear regression analysis, Panaretos [36] studies the number of points collected in the course of a double round robin tournament and the effect of the explanatories goals scored, goals conceded, and time of ball possession. The regression model

$$\begin{aligned} \text{points} &= \alpha \\ &+ \beta \log(\text{goals scored}/\text{goals conceded}) \\ &+ \gamma \text{ball possession} + \text{random noise} \end{aligned}$$

is fitted and a proportion of variance explained of 0.971 is found.

Barnett and Hilditch [37] study whether the field being artificial turf affects the game result. They find that the home team’s advantage was increased. The use of artificial turf was abandoned.

Attention now turns to the case where time or round, t , plays an essential role in the modeling. Jochems [16,17] defines, as a measure of strength of team i at home and j visiting, $\lambda_{ij} = W_i - W_j$, with W_i the average points of team i in its preceding games. The prediction of the result for team i is

$$\begin{aligned} i \text{ wins if } \lambda_{ij} &> 0, \\ i \text{ wins if } \lambda_{ij} &= 0 \text{ home advantage,} \\ i \text{ loses if } \lambda_{ij} &< 0. \end{aligned}$$

Jochem’s study had been commissioned by the Netherlands Lottery Commission to see if any skill was involved on the part of tipsters.

The models listed above can be turned into forecasting procedures by simply adding a t to the subscripts of the parameters and estimating them by using the data up to time t . One then uses the schedule of remaining games to develop forecasts. For example, Stefani [26] considers the model

$$X_{ijt} - Y_{ijt} = \eta_{ijt} + \rho_i - \rho_j + \text{random error},$$

with ρ_i a rating based on i ’s past games and $\eta_{ijt} = \pm \eta$ represents a home advantage. The

quantity ρ_i represents team i 's ability. The fit is by conditional least squares.

Dixon and Coles [23] employ independent Poissons and introduce an exponentially decaying time function to damp out the effects of previous rounds. The Poisson means are $\alpha_{i(t)}\beta_{j(t)}\eta$ and $\alpha_{j(t)}\beta_{i(t)}$, respectively. The variable t is time or round number and η stands for home advantage. Estimation is by maximizing the likelihood.

Fahrmeir and Tutz [28] work with a latent variable-motivated ordinal model,

$$\text{Prob}\{R_{ij}^t = r\} = F(\theta_{tr} - \alpha_{ti} - \alpha_{tj}) \\ - F(\theta_{tr-1} - \alpha_{ti} - \alpha_{tj}),$$

with R the result, and $r = 0, 1, 2$ a win-tie-loss indicator. Further F is the logistic, and α_{ti} are random walks in t . The computations involve a Kalman filter and simulation. Rue and Salvesen [24] employ a Bayesian dynamic generalised linear model with independent normals having conditional means given the past

$$\exp\{\alpha + \eta + \gamma_i^t - \delta_j^t - \phi(\varepsilon_i^t - \varepsilon_j^t)\}, \\ \exp\{\alpha + \gamma_j^t - \delta_i^t + \phi(\varepsilon_i^t - \varepsilon_j^t)\},$$

to predict next weekend's results. They employ a directed graph to describe the causal structure of the model as a function of time.

Brillinger [32,33], for each round, t , uses previous round results. He fits a latent variable-based model and then uses simulation to make projections of the final points and standings of the teams using the knowledge of the remaining games in the schedule. Estimates are then available for future rounds.

Harvey and Fernandes [38] study the series of goals scored by England against Scotland in a biyearly match. In a state space approach, they assume a negative binomial model with mean an exponentially decaying function of past values.

Within Game Modeling

Next, consider modeling the progress of a single game. One approach breaks the course

of a game into states corresponding to zones in the field. Pollard and Reep [39] carry out a logistic regression analysis to model the probability of scoring from various positions on the field. In a series of papers [40–42], Markov chains with a finite number of states are employed. Hirotsu and Wright [40] use a four-state model to determine the optimal timing of substitution and tactical decisions. The expected number of points to be gained from a change in tactics is considered an objective function, and dynamic programming is employed. The model includes transition rates of scoring and conceding, and rates of gaining and losing possession. Hirotsu and Wright found evidence for replacing a defender with an attacker when losing.

A common problem is to identify change taking place. Croucher [43] studied the effect of changing the points awarded. The object of this change is to generate more attacking play. Ridder *et al.* [44] study the effect of a red card in game. They infer that the scoring rate does change after the ejection, and has a negative effect on the team with fewer players. Other studies that deal with the effect are given in Refs 45–47. The latter take a point process approach. When team i is playing team j , the times of goals are modeled by Poisson processes with log rates

$$v_{ij}(t) = \lambda_i(t) \exp\{\alpha_j + \beta z_{ij}(t)\} \text{ and} \\ v_{ji}(t) = \lambda_j(t) \exp\{\alpha_i + \beta z_{ji}(t)\},$$

respectively, t being time into the game. The function, $\lambda_i(t)$, is referred to as the *attack intensity* for team i and α_i as the defense parameter. The z_{ij} are the explanatories

$$z_{ij1}(t) = 1 \text{ when rival player off with red} \\ \text{card, 0 otherwise;} \\ z_{ij2}(t) = 1 \text{ if actual score positive;} \\ z_{ij3}(t) = 1 \text{ if actual score negative;}$$

where t is time, β and z_{ij} are the 3 vectors. The parameters are estimated using data from the 2006 World Cup. The fitted model may be employed to generate simulations. Dixon and Robinson [48] also take a point

process approach, employing a birth process, to model the scoring.

The match video data being collected these days may be processed to determine trajectories of players and the ball. Then models can be developed. Xie *et al.* [49] take a video of a game, break it into segments, for example, “play” or “break.” They deal with camera pans and zooms using the green grass ratio and motion intensity. Their analysis uses dynamic programming and spatial-temporal hidden Markov processes. Min *et al.* [50] work on the same problem making use of rule-based reasoning, Bayesian inference, simulation, and expert systems. Palavi *et al.* [51] develops a graph-based multiplayer detection and tracking system. In the simplest case of a notable goal scoring movement in one game, Brillinger [31] develops a potential function approach to obtain a stochastic model.

Berument *et al.* [52] look for evidence of soccer results affecting the Turkish stock market. They do find an effect for one team as well as a day of the week effect.

Garvia [53] asks, “Is soccer dying?” He uses a Markov switching model to look for a break in the series of yearly average goals per game. He finds a drop for England, Italy, and Spain around 1965.

RANKING

Rankings/ratings/seedings are numerical values meant to describe the relative performances of teams in a league. They are used for a variety of specific purposes. One is the preparation of schedules for knock-out tournaments. Another is for the maximization of the probability that the best competitors meet in a tournament’s later stages. If there are N teams the collection of ranking values might be the sequence of integers 1 to N . A famous example is the FIFA/Coca-Cola world ranking [www.fifa.com.worldfootball/ranking/procedure/men.html [54]], where the values derive from a formula involving major games during the previous 4 years, strength of opponent, $W-T-L$ points, and several other items. It orders the teams of all the countries that are

members of the FIFA and is updated steadily as more game results become available.

Jochems [16,17] employed the differences of average game scores and compared this to tipsters’ predictions. Hill [55] uses Kendall’s tau to compare tipster results with the end of season standings and finds association. Stefani [26] suggests the model

$$X_{ij} - Y_{ij} = \eta + \rho_i - \rho_j + \text{random error},$$

with the ρ_i ’s rankings. Stefani [56] surveyed the major world sports rating systems. Bassett [57] suggests improving the estimates by employing a robust estimation method to handle long tails. Forrest and Simmons [58,59] investigate the predictive quality of newspaper tipsters match results. Stefani and Pollard [60] provide a critical survey of ranking procedures. Gelade and Dobson [61] relate the international standing of a country’s team to various social factors including number playing regularly, wealth, and climate. McHale and Davies [62] focus on the FIFA rankings.

There are papers on ratings for other sports; for example, Refs 63–65.

TOURNAMENTS AND SCHEDULING

Scheduling is a basic step involved when games need to be organized amongst the members of a group of teams. Two basic types of tournament design are the round robin and the knock-out. In the round robin, each pair of teams plays against each other the same number of times, equally at home and away. The champion of the group is the team with the most points at the end of all the games. In the knock-out case, there is a seeding order and the first round is based on it. The tournament progresses with two teams playing in pairs setup by the seeding order with the winner going on to the following round. This continues until only one team is left, the champion.

Sometimes there are numerical-valued objective functions that can be employed in developing schedules. Their forms can include distance traveled, total expenses,

numbers of consecutive home or away games, and referees' time. There are often constraints that need to be taken note of in the scheduling. For example, there may be multiple venues, television schedules, desired dates for two teams from the same city. Analytic tools employed in an optimization include integer programming, maximization, knapsack algorithms, simulation and stochastic models.

Kuonen [21,66] employs a logistic regression model and shows how output can be used to estimate probabilities of interest for a knock-out tournament. Unlike a round robin, future opponents for the following rounds are only known probabilistically. Kuonen works with European Cup Winners Cup data. He investigates three methods for calculating seeding coefficients based on the data for the 3 preceding years.

Urban and Russell [67] consider scheduling competitions on multiple venues, venues not associated with any of the participants. Della Croce and Oliveri [68] present an integer linear programming approach for scheduling the Italian League. They had to deal with cable TV and with games between teams in the same city.

Scheduling the Chilean League has also posed special problems. As described in Refs 69 and 70, these were dealt with via using mathematical programming, recognized existing weaknesses, stadiums availability, international requirements, and reducing travel distance.

Objectives may be described probabilistically and then stochastic models are involved. When the goal is to determine a champion, it is natural to ask questions like: what is the probability that the "best" team actually wins the tournament, or how much better is the first team than the second? Appleton [71] provides definitions and reviews many tournament structures. He employs simulations.

Scarf *et al.* [72] provide an extensive review including discussion of: tournament metrics, how the design influences outcome uncertainty, the use of simulation, robustness (to teams dropping out), effects of rule changes, and probability model employed. He defines the parameter

$$P_{qR} = \text{Prob}\{\text{team in top } 100q \text{ pretournament rank percentile progresses advances to round } R + 1\};$$

here R is round achieved and $0 < q < 1$. To understand the P_{qR} values, he graphs them against q .

It is necessary to schedule the referees also. Yavuz *et al.* [73] provide an extensive review. The topics covered include leagues in different countries requiring different objective functions and constraints, tension between referees and clubs caused by past incidents, reducing frequent assignment of the same referee to a team's games. In amateur leagues, there may be several games the same day for the same referee, perhaps at different places even.

GAME THEORY

The expression "game theory" has both a general and a technical meaning. The "technical" comes from the classic Von Neumann–Morgenstern [74] setup. The "general" refers to other things, particularly tactics. Both meanings provide general frameworks with which to study soccer. A difficulty arising in the theories' applications to soccer is that the game is highly dynamic. Jones and Trantner's [75] book is unusual in that, it starts by emphasizing the defensive formations. Wilkinson's [76] book is devoted to the topic of tactics.

The formation selected would be controlled by the coach. The 2-3-5 was very common for many years. In it, there are two defensive specialists, five attacking, and three midfielders in between. Newer formations include 3-5-2, 5-4-1, 4-5-1, 4-3-3, and even 4-1-4-1 with an extra row. The formation 4-4-2 is favored by various famous teams. The two attackers can drop back to assist in obtaining the ball or be on the sides up front. When the team has possession of the ball, the two outside midfielders can move forward to increase pressure on the opponent. Wilson [77] provides a review and insightful analysis of many of the formations employed in games dating from 1878 to 2008.

Pollard and Reep [39] seek to quantify the effectiveness of different playing strategies.

They break the action of a game down into discrete events, for example, pass, center, shoot. They define the “yield” as the probability a goal is scored minus the probability one is conceded. This quantity is used to evaluate the expected outcome of a team possession.

The book, *The Science of Soccer*, has a chapter on “game theory” [78]. It discusses how the strength and deployment of the team moderates the apparent random motion and discusses the fact that low scoring benefits the weaker team.

One of the momentous occasions in a soccer game is the awarding of a penalty shot. There have been a number of formal studies. Greenless *et al.* [79] study how the penalty takers’ uniform color and prepenalty gaze affect the impressions formed by the goal-keeper. When penalties take place to break a tie, McGarry and Franks [80] identify an optimal order for a team to take the shots in. Their conclusion is to begin by ranking the players 1 (best) to 11 (worst). Then use the order: 5, 4, 3, 2, 1, and if necessary after that 6, 7, 8, 9, 10, and 11. Jordet *et al.* [81] considers the roles of stress, skill, and fatigue for answering why the English are poor at penalties. He analyzed 200 shots from World Cup and European Championship. In an analysis of many penalties, Franks *et al.* [82] learned that in 80% of the cases studied, if the nonkicking foot points to the left, the ball will be shot to the left. They developed a training program to test this discovery.

Larsen [2] describes the arrival of Reep’s ideas in Norway and the success that teams adopting it enjoyed.

The theory of von Neumann and Morgenstern provides general definitions and leads to random strategies. Haigh [83] provides the following intuitive example. The table shows estimated percentages of successful penalty shots based on 1876 attempts.

Goalie	Kicker		
	Left	Center	Right
Left	60	90	93
Center	100	30	100
Right	94	85	60

These data concern the goalie’s direction of motion and the direction of the Kicker’s shot.

The situation can be considered is a two-person zero-sum game. If the kicker wins by scoring, the goalie loses. If the goalie wins by saving, the kicker loses. Haigh finds that Nash equilibrium theory leads to the goalie’s choosing to move left, center or right with the respective probabilities of 0.44, 0.13, and 0.43. For the kicker, they are 0.37, 0.29, and 0.34, respectively. If either player uses their strategy, 80% of the shots would be scored in the long run. This turns out to be close to the actual percentage.

Other references to applications of the formal theory include Refs 84–86. Hirotsu and Wright [85] consider a zero-sum game focusing on who wins the game, while Hirotsu *et al.* [87] focus on points gained. Here one has a nonzero-sum game. Bennett *et al.* [88] use hypergame analysis, involving the ranking of various forms of hooligans’ actions and authorities’ reactions followed by hooligan’s reactions and so on. They explore the implications of such scenarios.

ECONOMICS AND MANAGEMENT

Reilly and Williams [89] write, “In the 1980s it became apparent that the football (soccer) industry and professionals in the game could no longer rely on the traditional methods of previous decades. Methods of management science were applied to organizing the big soccer clubs and the training of players could be formulated on a systematic basis.” Desbordes [90, Chapter 10] advocates the introduction of modern management tools into soccer. This makes sense if for no other reason than the fact that billions of dollars are now involved in the sport. One does need to remember that at the youth and amateur levels, the amount of money involved is little. Really, all that is needed for a game is a “ball” and a field or a beach. This is one charm of the game.

Management refers to a sport’s structure, owners, sponsors, marketers, financiers, schedulers, and others with some form of control. Management provides the superstructure. The fans expect both their team’s players and its management to be successful

in acquiring players, providing facilities, and hiring the coach. Management assists by supporting the enterprise generally. It provides the infrastructure. Other things that management is responsible for include cash inflows–outflows, advertising, manpower planning and hiring, planning new and keeping up old stadiums, and signing players. The stadium owners provide facilities for security, ticket collection, food, souvenirs, and parking. Management studies the performance of their players and others [10,91].

Sometimes, there is a numerical-valued objective function for the management to work with. Then optimization methods may be employed to good advantage. Other concerns are loss functions, earnings/revenue, prestige/honor, recruitment, salary negotiation, scouting, referees, merchandise, commercialism, productivity, profitability, travel, and television rights.

There is a common wish to improve individual player's and a team's performance. An accompanying question is how to assess performance and efficiency. One paper on the topic for English soccer is Ref. 92. Others are Refs 93 and 94. The latter concerns the Spanish Soccer League. There have been analyses. Audas *et al.* [95] study the impact of managerial change on team performance using the data of English football match results since the 1970s, while Partovi and Corredoir [96] investigate models for prioritizing and designing rule changes.

In a study of competitive balance in the Dutch League, Koning [97] employs the values C_{ij} , corresponding to W, T, L , generated via

$$\begin{aligned} D_{ij} &= \varphi_i - \varphi_j + \varepsilon_{ij} \text{ with the } \varepsilon\text{'s mean 0,} \\ &\quad \text{common variance normals,} \\ C_{ij} &= 1 \text{ if } D_{ij} > c_2 \\ &= 0 \text{ if } D_{ij} = 0 \\ &= -1 \text{ if } D_{ij} < c_1. \end{aligned}$$

It is used to study changes in competitive balance in Holland.

A novel application of programming theory is presented in Ref. 98. It applies an integer program to the generalized knapsack

problem to find a "Dream Team." Player values were established by a public poll with the participants given a budget of 2.5 million dollars each. A knapsack algorithm was then employed to establish the best team at the lowest price.

Calster *et al.* [99] study an unwanted happening in soccer, the scoreless draw. The occurrence is related to indices of a team's offensive performance including total goals and earned points per game.

SOME REMAINING TOPICS

The book, *The Inner Game of Soccer*, provides details of the game from a referee's standpoint [100]. It discusses the laws and the mechanics and talks of the characteristics of the best referees and the ones who can turn benign games into ugly unsportsman-like confrontations.

Reference 78 lays out some of the physics of soccer. For example, the chapter titled "The ball in flight" analyzes the movement of a ball spinning in flight and how players can make it bend. Hall [101] elaborates on that describing the physical forces at work in a double banana shot, which curves in one direction then swerves in another. Such a shot is very difficult for the goalie to handle.

Sports biomechanics concerns physical actions that occur in sport that can be observed and improved, for example, to reduce injuries. The Refs 102 and 103 are on the topic of biomechanics in soccer. Grimpampi *et al.* [104] concern computational biomechanics. Bradley *et al.*'s study [105] is a human biology study. It refers to results on high-intensity running using data collected via an array of television cameras.

There are many articles on employing stochastic models in evaluating bets. Most of the articles referred to above go on to study the efficiency of various gambling strategies.

What is ahead for the game? Wright [106] writes on 50 years of operations research in sport to provide a background. The future is discussed in Ref. 107 and by the sports writer Garner [108] in *World Soccer*. He suggests the following changes: bigger goals, smaller penalty area, a revised offside rule,

fouls denoted contact or technical, red-carded player punished but replacement allowed. Reviewing the literature and practice suggest that one can expect to see much more in the way of the collection and analysis of the spatial-temporal data collected in a game.

REFERENCES

1. Henshaw R. The encyclopedia of world soccer. Washington, DC: New Republic; 1979.
2. Larsen O. Charles Reep: a major influence on British and Norwegian football. *Soccer Soc* 2001;2:55–78.
3. Pollard R. Charles Reep (1904–2002): pioneer of notational and performance analysis in football. *J Sports Sci* 2002;20:853–855.
4. Pollard R. Home advantage in soccer: variations in its magnitude and a literature review of the inter-related factors associated with its existence. *J Sport Behav* 2006;29:169–189.
5. Reilly T, Thomas V. A motion analysis of work-rate in differential positional roles in professional football match-play. *J Hum Mov Stud* 1976;2:87–97.
6. McGarry T, Franks IM. The science of match analysis. In: Reilly T, Williams AM, editors. *Science and soccer*, 2nd ed. New York: Routledge; 2003. pp. 265–275.
7. Hughes M. Notational analysis. In: Reilly T, Williams AM, editors. *Science and soccer*, 2nd ed. London: Routledge; 2003. pp. 245–264.
8. Di Salvo V, Collins A, McNeill B, Cardinale M. Validation of Prozone: a new video-based performance analysis system. *Int J Perform Anal Sport* 2006;6:108–119.
9. Carling C, Reilly T, Williams AM. *Handbook of soccer match analysis*. London: Routledge; 2005.
10. Carling C, Reilly T, Williams AM. *Performance assessment for field sports*. London: Routledge; 2009.
11. FIFA. Laws of the game. (2008/2009). www.fifa.com.
12. FIFA. Disciplinary code. 2009. www.fifa.com/mm/50/02/75/disco_2009_en.pdf.
13. Moroney M. *Facts from Figures*. London: Pelican; 1951.
14. Reep C, Benjamin B. Skill and chance in association football. *J R Stat Soc A* 1968;131:581–585.
15. Hughes M, Franks I. Analysis of passing sequence, shots and goals in soccer. *J Sports Sci* 2005;23:509–514.
16. Jochems DB. Voetbalprognoses. *Stat Neerl* 1958;12:17–31.
17. Jochems DB. Forecasting the outcomes of soccer matches: a statistical appraisal of the Dutch experience. *Metrika* 1962;5:194–207.
18. Greenhough J, Birch PC, Chapman SC, Rowlands G. Football goal distributions and extremal statistics. *Phys A* 2002;316:615–624.
19. Reep C, Pollard R, Benjamin B. Skill and chance in ball games. *J R Stat Soc A* 1971;134:623–629.
20. Maher MJ. Modelling association football scores. *Stat Neerl* 1982;36:109–118.
21. Kuonen D. Modelling the success of soccer teams in European championships. Technical Report 96.1, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne; 1996.
22. Lee AJ. Modelling soccer in the Premier League: is Manchester United really the best? *Chance* 1997;15–19.
23. Dixon MJ, Coles SC. Modelling association football scores and inefficiencies in the football betting market. *Appl Stat* 1997;46:265–280.
24. Rue H, Salvesen O. Prediction and retrospective analysis of soccer matches in a league. *J R Stat Soc D* 2000;49:399–418.
25. Karlis D, and Ntzoufras I. Analysis of sports data using bivariate Poisson models. *Statistician* 2003;52:381–393.
26. Stefani RT. Improved least squares football, basketball, and soccer predictions. *IEEE Trans Syst, Man, Cybern* 1980;10:116–123.
27. Karlis D, Ntzoufras I. Bayesian modelling of football outcomes: using the Skellam's distribution for goal difference. *IMA J Manag Math* 2009;20:133–145.
28. Fahrmeir L, Tutz G. Dynamic stochastic models for time-dependent ordered paired comparison systems. *J Am Stat Assoc* 1994;89:1438–1449.
29. Brillinger DR. An analysis of an ordinal-valued time series. In: *Athens conference on applied probability and time series analysis*. Volume 2, *Lecture Notes in Statistics* Volume 115. Springer; 1996. pp. 73–87.
30. Brillinger DR. Modelling some Norwegian soccer data. In: Nair V, editor. *Advances*

- in statistical modelling and inference. Singapore: World Scientific; 2006. pp. 3–20.
31. Brillinger DR. A potential function approach to the flow of play in soccer. *J Quant Anal Sports* 2007;3(1):3.
 32. Brillinger DR. Modelling game outcomes of the Brazilian 2006 Series A Championship as ordinal-valued. *Brazilian J Probab Stat* 2008;22:89–104.
 33. Brillinger DR. An analysis of Chinese Super League partial results. *Sci China A: Math* 2009;52:1139–1156.
 34. Leeflang PSH, van Praag BMS. A procedure to estimate relative powers in binary contacts and an application to Dutch Football League results. *Stat Neerl* 1971;25:203–208.
 35. Goodard J. Regression models for forecasting goals and match results in association football. *Int J Forecast* 2005;21:331–340.
 36. Panaretos VM (2002). A statistical analysis of the European Soccer Champions League. *ASA Proceedings of the Joint Statistics Meetings*. pp. 2600–2602.
 37. Barnett V, Hilditch S. The effect of an artificial pitch surface on home team performance in football (soccer). *J R Stat Soc A* 1993;156:39–50.
 38. Harvey AC, Fernandes C. Time series models for count data of qualitative observations. *J Bus Econ Stat* 1989;7:407–422.
 39. Pollard R, Reep C. Measuring the effectiveness of playing strategies at soccer. *J R Stat Soc D* 1997;46:541–550.
 40. Hirotsu N, Wright M. Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. *J Oper Res Soc* 2002;53:88–96.
 41. Hirotsu N, Wright M. An evaluation of characteristics of teams in association football by using a Markov process model. *Statistician* 2003;52:591–602.
 42. Hirotsu N, Wright M. Determining the best strategy for changing the configuration of a football team. *J Oper Res Soc* 2003;54:878–887.
 43. Croucher JS. The effect of changing competition points in the English football league. *Teach Stat* 1984;6:39–42.
 44. Ridder G, Cramer JS, Hopstaken P. Down to ten: estimating the effect of a red card in soccer. *J Am Stat Assoc* 1994;89:1124–1127. [see also Albert J, Bennett J, Cochran JJ. *Anthology of statistics in sports*. Philadelphia: SIAM; 2005. pp. 299–302.]
 45. Wright M, Hirotsu N. The professional foul in football: tactics and deterrents. *J Oper Res Soc* 2003;54:213–221.
 46. Vecer J, Kopriva F, Ichiba T. Estimating the effect of a red card in soccer: when to commit an offense in exchange for preventing a goal opportunity. *J Quant Anal Sports* 2009;5(1):8.
 47. Volf P. A random point process model for score in matches. *IMA J Manag Math* 2009;20:121–131.
 48. Dixon MJ, Robinson ME. A birth process model for association football matches. *Statistician* 1998;47:523–538.
 49. Xie L, Xu P, Chang S-F, Divarkanm A, Sun H. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognit Lett* 2004;25:76–775.
 50. Min B, Kim J, Choe C, Eom H, McKay RI. A compound framework for sports prediction: a football case study. *Knowledge-Based Syst* 2008;21:551–562.
 51. Palavi V, Mukherjee J, Majumdar AK, Sural S. Graph-based multiplayer detection and tracking in broadcast soccer videos. *IEEE Trans Multimedia* 2008;10:794–805.
 52. Berument H, Ceylan NB, Gozpınar E. Performance of soccer on the stock market: evidence from Turkey. *Social Sci J* 2006;43:695–699.
 53. Garvia A. Is soccer dying? A time series approach. *Appl Econ Lett* 2007;7:275–278.
 54. FIFA. Ranking. 2006. [www.fifa.com.worldfootball/ranking/procedure/men.html](http://www.fifa.com/worldfootball/ranking/procedure/men.html).
 55. Hill ID. Association football and statistical inference. *Appl Stat* 1974;23:203–208.
 56. Stefani RT. Survey of the major world sports rating systems. *J Appl Stat* 1997;24:635–646.
 57. Bassett GW. Robust sports ratings based on least absolute errors. *Am Stat* 1997;51(2).
 58. Forrest D, Simmons R. Making up the results: the work of the Football Pools Panel, 1963–1997. *Statistician* 2000;49:253–260.
 59. Forrest D, Simmons R. Forecasting sport: the behaviour and performance of football tipsters. *Int J Forecast* 2000;16:317–331.
 60. Stefani R, Pollard R. Football rating systems for top-level competition: a critical survey. *J Quant Anal Sports* 2007;3(3):3.
 61. Gelade GA, Dobson P. Predicting the comparative strengths of national football teams. *Social Sci Q* 2007;88:244–258.

62. McHale I, Davies S. Statistical analysis of the effectiveness of the FIFA world rankings. In: Albert J, Koning RH, editors. *Statistical thinking in sports*. New York: Chapman and Hall; 2008. pp. 77–90.
63. Leake RJ. A method for ranking teams with an application to 1974 college football. *North Holland: Management Science in Sports*; 1976.
64. Harville D. The use of linear model methodology to rate high school or college football teams. *J Am Stat Assoc* 1977;72:278–289.
65. Stern H. On the probability of winning a football game. *Am Stat* 1992;45:179–183. [see also Albert J, Bennett J, Cochran JJ. *Anthology of statistics in sports*. Philadelphia: SIAM; 2005. pp. 53–57].
66. Kuonen D. Statistical models for knockout soccer tournaments. Technical Report 97.3. Lausanne: Department of Mathematics, Swiss Federal Institute of Technology; 1997.
67. Urban TL, Russell RA. Scheduling sports competitions on multiple venues. *Eur J Oper Res* 2003;148:302–311.
68. Della Croce F, Oliveri D. Scheduling the Italian Football League: an ILP-based approach. *Comput Oper Res* 2006;33:1963–1974.
69. Noronha TF, Ribeiro CC, Duran G, Souyris S, Weintraub A. A branch-and-cut algorithm for scheduling the highly-constrained Chilean soccer tournament. *Lect Notes Comput Sci* 2007;3867:174–186.
70. Duran G, Guajardo M, Weintraub A, Wolf R. Scheduling the Chilean League using mathematical programming. *OR/MS Today* 2009.
71. Appleton DR. May the best man win? *Statistician* 1995;44:529–538.
72. Scarf P, Yusof MMN, Bilbao M. A numerical study of designs for sporting contests. *Eur J Oper Res* 2009;198:190–198.
73. Yavuz M, Inan UH, Figlah A. Fair referee assignments for professional football leagues. *Comput Oper Res* 2008;35:2937–2951.
74. Von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton: Princeton University Press; 1944.
75. Jones R, Trantner T. *Soccer Strategies: defensive and attacking tactics*. Reedswain: Spring City; 1979.
76. Wilkinson WHG. *Soccer tactics: top team strategies explained*. Crowood: Marlborough; 1988.
77. Wilson J. *Inverting the pyramid: a history of football tactics*. London: Orion; 2008.
78. Wesson J. *The science of soccer*. New York: Taylor and Francis; 2002.
79. Greenless I, Leyland A, Thelwell R, Filby W. Soccer penalty takers' uniform colour and pre-penalty gaze affects the impressions formed by opposing goal-keepers. *J Sports Sci* 2008;26:569–576.
80. McGarry T, Franks IM. On winning the penalty shoot-out in soccer. *J Sports Sci* 2000;18:401–409.
81. Jordet G, Hartman E, Vischer C, Lemmink KAPM. Kicks from the penalty mark in soccer: the roles of stress, skill, and fatigue for kick outcomes. *J Sport Sci* 2007;25:121–129.
82. Franks I, McGarry T, Hanvey T. From notation to training: analysis of the penalty kick. *Insight* 1999;2:24–26.
83. Haigh J. Uses and limitations of mathematics in sport. *IMA J Manag Math* 2009;20:97–108.
84. Pacacios-Huerta I. Professionals play minimax. *Rev Econ Studies* 2003;70:395–415.
85. Hirotsu N, Wright M. Modelling tactical changes of formation in association football as a zero-sum game. *J Quant Anal Sports* 2006;2(2):4.
86. Coloma G. Penalty kicks in soccer: an alternative methodology for testing mixed-strategy equilibria. *J Sports Econ* 2007;8:530–545.
87. Hirotsu N, Ito M, Miyaji C, Hamano K, Taguchi A. Modelling tactical changes of formation in association football as a non-zero-sum game. *J. Quant Anal Sports* 2009;5(3):2.
88. Bennett PG, Dando MR, Sharp RG. Using hypergames to model difficult social issues: an approach to the case of soccer hooliganism. *J Oper Res Soc* 1980;31:621–635.
89. Peiser B, Minten J. Soccer violence. In: Reilly T, Williams AM, editors. *Science and soccer*. 2nd ed. Abingon: Routledge; 2003. pp. 230–242.
90. Desbordes M. *Marketing and football: an international perspective*. Portsmouth: Butterworth-Heinemann; 2007.
91. Covell D, Walker S, Siciliano J, Hess PW. *Managing sports organizations: responsibility for performance*, 2nd ed. Portsmouth: Butterworth-Heinemann; 2007.

92. Dawson P, Dobson S, Gerrard B. Stochastic frontiers and the temporal structure of managerial efficiency in English soccer. *J Sports Econ* 2000;1–4:341–362.
 93. Haas DJ. Technical efficiency of Major League Soccer. *J Sports Econ* 2003;04–03: 203–215.
 94. Sala-Garrido R, Carrion VL, Esteve AM, Bosca JE. Analysis and evolution of efficiency in the Spanish Soccer League. *J Quant Anal Sports* 2009;5(1):3.
 95. Audas R, Dobson S, Goddard J. The impact of managerial change on team performance in professional sports. *J Econ Bus* 2002;54:633–650.
 96. Partovi FY, Corredoir RA. Quality deployment for the good of soccer. *Eur J Oper Res* 2002;137:642–656.
 97. Koning RH. Balance in competition in Dutch soccer. *Statistician* 2000;49:419–431.
 98. Mosheiov G. The solution of the soccer ‘Dream League’ game. *Math Comput Model* 1998;27:79–83.
 99. Calster BV, Smits T, Van Huffel S. The curse of scoreless draws in soccer: the relationship with a team’s offensive, defensive, and overall performance. *J Quant Anal Sports* 2008;4(1):4.
 100. Sellin E. The inner game of soccer. Mountain View: World Publications; 1976.
 101. Hall LE. Laws of motion. New York: Rosen; 2005.
 102. Lees A, Nolan L. The biomechanics of soccer: a review. *J Sports Sci* 1998;16:211–234.
 103. Lees A. Biomechanics applied to soccer skills. In: Science and soccer, 2nd ed, Reilly T, Williams AM, editors. New York: Routledge; 2003. pp. 109–119.
 104. Grimpampi E, Pasculli A, Sacripanti A. Computational biomechanics, stochastic motion and team sports. 2008. arxiv.org/pdf/0811.3659
 105. Bradley PS, Sheldon W, Wooster B, Olsen P, Boanas P, Krstrup P. High-intensity running in English FA Premier League soccer matches. *J Sports Sci* 2009;27:159–168.
 106. Wright MB. 50 years of OR in Sport. *J Oper Res Soc* 2009;60:S161–S168.
 107. The Orange Future of Football Report. 2008. http://www.orange.co.uk/sport/football/pics/3395_1.htm
 108. Garner P. The future of football. World Soccer; London; 2009.
- FURTHER READING**
- Archontakis F, Osborne E. Playing it safe? A Fibonacci strategy for soccer betting. *J Sports Econ* 2007;8:295–308.
- Baxter M, Stevenson R. Discriminating between the Poisson and negative binomial distributions: an application to goal scoring in association football. *J Appl Stat* 1988;15:347–354.
- Bellos A. Futebol, the Brazilian Way of Life. London: Bloomsbury; 2002.
- Boeri T, Severgnini B. The Italian job: match rigging, career concerns and media concentration in ‘Series A’. IZA Discussion Paper No. 3745. 2008.
- Brimson D. March of the hooligans: soccer’s bloody fraternity. London: Virgin; 2007.
- Brocas I, Carrillo JD. Do the three-point victory and golden goal rules make soccer more exciting? *J Sports Econ* 2004;5:169–185.
- Cantor A, Arcucci D. Gooool: a celebration of soccer. New York: Fireside; 1997.
- Carmichael F, Thomas D, Ward R. Team performance: the case of English Premiership Football. *Manag Dec Econ* 2000;21:31–45.
- Comeron M. The prevention of violence in sport. Strasbourg: Council of Europe; 2002.
- Cordeiro SE. El futbol y la corrupcion. Volume 5. In: Carrion F, editor. Biblioteca del Futbol Ecuatoriano, 5 Volumes. Quito: Flacso; 2006. pp. 233–250.
- Crowder M, Dixon M, Ledford A, Robinson M. Dynamic modelling and prediction of English Football League matches for betting. *Statistician* 2002;51:157–168.
- David FN. Games, gods and gambling. London: Griffin; 1962.
- Dobson S, Goddard J. The economics of football. Cambridge: Cambridge University Press; 2001.
- Dobson S, Goddard J. Forecasting scores and results and testing the efficiency of the fixed odds betting market in Scottish League football. In: Albert J, Koning RH, editors. Statistical thinking in sports. New York: Chapman and Hall; 2008. pp. 91–109.
- Dyte D, Clarke SR. A ratings based Poisson model for World Cup soccer simulation. *J Oper Res Soc* 2000;51:993–998.
- Eastaway R, Haigh J. Beating the odds: the hidden mathematics of sport, London: Robson; 2007.
- Emonet B. Revisiting statistical aspects of soccer, STS Report. Lausanne: Swiss Technical University; 2000.

- England C. No more budha, only football, London: Sceptre; 2003.
- Espita-Escuer M, Garcia-Cebrianm LI. Measuring the efficiency of Spanish First-Division soccer teams. *J Sports Econ* 2004;04-03: 203-215.
- Fernandez-Cantelli E, Meeden G. An improved award system for soccer. *Chance* 2002;7: 275-278.
- Foer F. How soccer explains the world: an unlikely theory of globalization, New York: Harper Collins; 2004.
- Giulianotti R. Football: a sociology of the global game, Polity: Bodmin; 1999.
- Giulianotti R, Bonney N, Hepworth M. Football, violence and social identity, London: Routledge; 1994.
- Goddard J, Asimakopoulos I. Forecasting football results and the efficiency of fixed-odds betting. *J Forecast* 2004;23:51-66.
- Holey AJ, Johnson BS. Menaces to management: a developmental view of British Soccer Hooligans, 1961-1986. *Sports J* 1998;1(1).
- Hornby N. Fever pitch. Riverhead trade; London: 1992.
- Hughes M, Bartlett R. The use of performance indicators in performance analysis. *J Sports Sci* 2002;20:739-754.
- Hurley WJ. Special issue on operations research in sport. *Comput Oper Res* 2006;33:1871-1873.
- Jackson DA. Index betting on sports. *Statistician* 1994;43:309-315.
- Karlis D, Ntzoufras I. On modelling soccer data. *Student* 2000;3:229-244.
- Kern M, Sussmuth B. Managerial efficiency in German top league soccer: an econometric analysis of club performances on and off the pitch. *German Econ Rev* 2005;6:485-506.
- Knorr-Held L. Dynamic rating of sports teams. *Statistician* 2000;49:261-276.
- Kristan M, Pers J, Perse M, Kovacic S. Towards fast and efficient methods for tracking players in sports. In: Pers J, Magee DR, editors. *Proceedings of ECCV Workshop on Computer Vision Based Analysis in Sport Environments*. 2006. pp. 14-25.
- Kuonen D, Roehrl ASA. Was Frances's World Cup win pure chance? *Student* 2000;3:153-166.
- Kuper S, Szymanski S. *Soccernomics*, New York: Nation; 2009.
- Kuypers T. Information and efficiency: an empirical study of a fixed odds betting market. *Appl Econ* 2000;32:1353-1363.
- Ladany SP, Machol RE, editors. *Optimal strategies in sports*. Amsterdam: North-Holland; 1977.
- Lewis MM. *Moneyball: the art of winning an unfair game*. New York: Norton; 2003.
- Lucena CJP, Noronha S, Urrutia S, Ribeiro CC. A multi-agent framework to retrieve and publish information on qualification and elimination data in sports tournaments. Rio de Janeiro, Brazil: *Monografias em Ciência da Computação*; 2006.
- Machol RE, Ladany SP, Morrison DG. *Management science in sports*. Amsterdam: North-Holland; 1976.
- Magnum. *Soccer*. London: Phaidon; 2002.
- Mason T. *Association football and english society 1863-1915*. Brighton: Harvester; 1980.
- McCullough P, Nelder JA. *Generalized linear models*, 2nd ed. London: Chapman and Hall; 1989.
- McGuire EJ, Courneya KS, Widmeyer WN, Carron AV. Aggression as a potential mediator of the home advantage in professional hockey. *J Sport Exerc Physiol* 1992;14:148-158.
- Morris D. *The soccer tribe*. Jonathan Cape; London: 1981.
- Needham CJ. *Tracking and modelling of team game interactions*. Ph.D. Thesis, University of Leeds; 2003.
- Osborne MJ. *An introduction to game theory*. Oxford: Oxford University Press; 2004.
- Ribeiro CC, Urrutia S. OR on the ball: applications in sports scheduling and management. *OR/MS Today* 2004; (June).
- Sadovskii AL, Sadovskii LE. *Mathematics and sports*. Providence: American Mathematical Society; 1993.
- Simons R. *Bamboo goalposts*. London: MacMillan; 2008.
- Sloane PJ. The economics of professional football: the football club as a utility maximiser. *Scottish J. Polit Econ* 1971;8:121-146.
- Sloane PJ editor. *The economics of sport*. *Econ Affairs* 1997;17:Special issue.
- Stefani R. Measurement and interpretation of home advantage. In: Albert J, Koning RH, editors. *Statistical thinking in sports*. New York: Chapman and Hall; 2008. pp. 203-216.
- Stefani RT. Predicting score difference versus score total in rugby and soccer. *IMA J Manag Math* 2009;20:147-158.
- SUP AMISCO PRO. <http://213.30.139.108/sport-universal/uk/amiscopro.htm>; 2010.

- Turnbull J, Raab A, Satterlee T. The global game: writers on soccer. Bison; London: 2008.
- Wright DB. Football standings and measurement levels. *Statistician* 1997;46:105–110.
- Yue Z, Broich H, Seifritz F, Mester J. Mathematical analysis of a soccer game. Part 1: individual and collective behaviors. *Studies Appl Math* 2008;121:223–243.
- Yue Z, Broich H, Seifritz F, Mester J. Part 2: energy, spectral and correlation analyses. *Studies Appl Math* 2008;121:245–261.

SOCIEDAD PERUANA DE INVESTIGACIÓN OPERATIVA Y DE SISTEMAS (SOPIOS)

ROSA DELGADILLO
DAVID MAURICIO
Universidad Nacional
Mayor de San Marcos,
Lima, Peru

Operations research (OR) has been promoted in Peru by Eugen Blum, a Swiss national and professor of the Mathematics Faculty of the Universidad Nacional Mayor de San Marcos (UNMSM) and the other universities. Eugen Blum promoted the creation of an association for OR in 1985, bringing together students, alumni, and teachers of the OR school (the OR school of Mathematics Faculty was created in 1969). He died in 1993.

ANTECEDENTS AND SOCIETY CREATION

In the year 2006, David Mauricio Sanchez, professor of the Systems Engineering, Faculty of the UNMSM in the Latin Iberoamerican Congress on Operational Research (CLAIO), under the organization of the Association Latin Iberoamerican Operations Research (ALIO), proposed that Peru be the headquarters of the XIII Latin American Summer School on Operations Research (XIII ELAVIO). The proposal was based on the following: (i) it was necessary to encourage the development of the OR field in Peru, and researchers and practitioners needed to be motivated and encouraged; (ii) an ELAVIO had never been held in Peru, so this would be an excellent opportunity to gather national professionals in the field and encourage them to research. The proposal was accepted by the Managerial Committee of ALIO, and in February 2008, the first summer school on OR in Peru was organized, which took place in the district of Chosica, a few kilometers from Lima. The XIII ELAVIO was organized by two universities: the

UNMSM and the Universidad Inca Garcilaso de la Vega. In this academic event, students, professionals, and organizers of the school got together and decided to form the Peruvian Society of Operations Research, with the aim of bringing researchers, practitioners, teachers, graduates, and undergraduates of all universities of Peru interested in developing the OR field. That is, it was thought that the Peruvian Society would bring together professionals from different Peruvian universities with different areas of training with emphasis on the development of the OR field; an *ad hoc* committee was established for the creation of the Peruvian Society of Operations Research, which was composed of professionals from different universities.

In October 25, 2008 officially the Peruvian Society for Operations Research and Systems (SOPIOS) was created, in a general assembly, carried out in the Faculty of Systems Engineering of the UNMSM; in this assembly, the first Managerial Committee constituted by a President, a Vice President, a Secretary, the Administrative Financier Director, a Director of Events and Institutional Image, a Director of Research and Development, and the Director of Education and Publications were chosen. The first managerial committee was chaired by Rosa S. Delgadillo Avila.

The founding members of the SOPIOS were 42 university teachers from UNMSM (Faculty of Systems Engineering, Faculty of Mathematics), Universidad Inca Garcilaso de la Vega, Universidad Nacional de Ingeniería, Universidad Católica Santo Toribio de Mogrovejo of Chiclayo, and Universidad Cesar Vallejo of Trujillo.

SOPIOS, from the legal point of view, is a nonprofit civil association registered in the registry zone No. IX in Lima, with item number 12266253 in the Lima Registry Office. The Managerial and Government organs of SOPIOS are the General Assembly and the Managerial Committee, whose mandate is for two years, similar to the fiscal committee. The associates' General Assembly is summoned at least once a year; however, the Directors

have monthly meetings and constant communication with their members through phone and e-mail.

The purposes of the SOPIOS are as follows [1]:

1. congregating professionals, students, and institutions with interests in OR, systems, and related areas;
2. stimulating activities such as teaching, research and development in OR in Peru;
3. Spreading the applicability of OR and related areas in the sectors of production and service, as a factor of competitiveness and productivity for the national development;
4. supporting the educational entities in subjects related to academic formation in operations research and systems;
5. exchanging experiences and information among professionals, academic and managerial level personnel in the country and abroad, so that it permits progress in OR; and
6. contributing to the technological and scientific research in the country, promoting research, and technology transfer activities.

The following activities are considered for SOPIOS purposes:

1. organizing and promoting OR and systems events at the local, regional, national, and international level periodically;
2. editing publications in OR, systems, and related areas;
3. establishing communication channels and relationship with other national and international organizations, which are interested in the application and development of OR;
4. promoting the integration and application of OR in other areas;
5. advising the implementations of solutions to problems of OR, systems, and related areas, and document the successful cases; and
6. programming the acquisition of funds for accomplishing these objectives.

ACHIEVEMENTS

The group of professionals that participates in SOPIOS, was the group that organized the XIII ELAVIO and the first ELAVIO in Peru, which took place in February 2008 at Chosica-Lima (Fig. 1), with a very good international representation [2]. This event had participants from Peru, Brazil, Cuba, Colombia, Chile, Uruguay, Canada, The Netherlands, Philippines, Israel, and Dominican Republic, with a total of 55 participant students. This ELAVIO was organized by the UNMSM and the Universidad Inca Garcilaso de la Vega and was sponsored by ALIO, International Federation of Operational Research Societies (IFORS), and other institutions such as Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC), and Centro Latinoamericano de estudios en Informática (CLEI).

In June 2008, taking advantage of the visit to Peru by Professor Milan Zeleny of Fordham University, New York, and with the purpose of maintaining the society actively, the International conference “Integración de Conocimiento y gestión de auto-producción” was organized in the System Engineering Faculty of the UNMSM (Fig. 2). This conference stirred up the interest of the professionals and students.

SOPIOS, with their objective to diffuse and to stimulate the research in OR in Peru, has considered to organize national congresses every year, similar to its regional partners, the SPBO (Brazilian Symposium on Operations Research organized by the Brazilian Society of Operational Research (SOBRAPO)) and OPTIMA (Chilean Congress on Operations Research organized by the Chilean Institute of Operations Research (ICHIO)). With this purpose and since there are a lot of universities in the city of Lima (the population is nearly 8 million people), in January 2009, the SOPIOS invited different national universities (Universidad Nacional de Ingeniería, Universidad Nacional del Callao, UNMSM) and private universities (Universidad de Lima, Universidad Ricardo Palma, Universidad de Ciencias Aplicadas, Pontificia Universidad Católica del Perú, Universidad Inca Garcilaso de la



Figure 1. XIII ELAVIO, February 2008.



Figure 2. Conference of Milan Zeleny, June 2008.

Vega) to organize the congress. The Faculty of Industrial Engineering of the UNMSM, was proposed to be the organizer of the Peruvian Congress on Operations Research and Systems (COPIOS 2009), which was accepted by the directive committee of SOPIOS.

COPIOS 2009, will be held in November, and the objective of the congress is not limited to the diffusion and transmission of knowledge, but rather to also attempt to integrate the academia with the industry and services, and the government, so as to foster a sustained development of the OR, systems, and related areas, with benefits for the managerial sector as well as the academic, and for the feedback of knowledge and experiences that could be utilized in both directions. This congress is being sponsored by UNMSM and national organizations such as CONCYTEC, colegio de Ingenieros del Perú, Ministerio de Transporte y Comunicaciones, Sociedad Nacional de Industrias, Pontificia Universidad Católica del Perú, Universidad Nacional del Callao, Universidad Andina del Cusco, Diario la Primera, and international organizations such as IEEE and CLEI. The publicity given to this congress has been to such an extent in the national and international levels, that we will have conference coordinators of the first order, such as Tomas Leibling, Nelson Maculan, Phillipe Michelon, Ann Marie Maculan, and Irene Loiseau.

PERSPECTIVES

At present, the number of associates has increased to 56 members; they are professors

of the universities: Pontificia Universidad Católica del Peru, Universidad Nacional del Callao, and UNMSM (Faculty of Industrial Engineering). However, it is expected that with the congress to be held in November 2009, other professionals and professors of other universities in the interior of Peru will be integrated.

SOPIOS hopes to congregate professionals and academics from the different regions of the country (universities and institutions) so as to become a representative institution that is able to promote the development of OR, systems, and related areas, as much as possible, in each region of Peru.

SOPIOS hopes to integrate the organizations ALIO and IFORS, as representative members of Peru, for which, we have already begun requesting enrolment of applicants at the end of the 2008.

We hope to get closer to our regional partners, so that we can make combined activities for the promotion, dissemination, and application of OR to the problems that our countries face. Finally, we hope to get closer with our nonregional partners also.

REFERENCES

1. Statutes of SOPIOS. Available at www.sopios.org. Accessed 2008 Oct.
2. IFORS. Bulletin of IFORS. Volume 2, No. 1. International Federation of Operational Research Societies; Mar 2008.

SOFTWARE FOR NONLINEARLY CONSTRAINED OPTIMIZATION

SVEN LEYFFER

ASHUTOSH MAHAJAN

Mathematics and Computer Science

Division, Argonne National Laboratory,
Argonne, Illinois

INTRODUCTION AND BACKGROUND

Software for nonlinearly constrained optimization (NCO) tackles problems of the form

$$\begin{cases} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & l_c \leq c(x) \leq u_c \\ & l_A \leq A^T x \leq u_A \\ & l_x \leq x \leq u_x, \end{cases} \quad (1)$$

where the objective function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and the constraint functions $c_i: \mathbb{R}^n \rightarrow \mathbb{R}$, for $i = 1, \dots, m$ are twice continuously differentiable. The bounds $l_c, l_A, l_x, u_c, u_A, u_x$ can be either finite or infinite. Equality constraints are modeled by setting $l_j = u_j$ for some index j . Maximization problems can be solved by multiplying the objective by -1 (most solvers handle this transformation internally). Solvers often take advantage of special constraints such as simple bounds and linear constraints.

NCO solvers are typically designed to work well for a range of other optimization problems such as solving a system of nonlinear equations (most methods reduce to Newton's method in this case), bound-constrained problems, and linear programming (LP) or quadratic programming (QP) problems. In this survey, we concentrate on solvers that can handle general NCO problems possibly involving nonconvex functions. There is software such as MOSEK [1] and CVXOPT [2] that are designed specifically for problems with only convex functions.

Methods for solving Equation (1) are iterative. That is, given an estimate x_0 of the solution, they compute a sequence of iterates $\{x_k\}$ converging to a stationary point, using the following four basic components:

1. *Approximate Subproblem.* The subproblem is an approximation of Equation (1) around the current iterate x_k that can be solved (approximately), and produces a step or search direction that improves the current estimate x_k . Examples of such subproblems are linear or QP approximation, the barrier problem, and an augmented Lagrangian; see the article titled **Foundations of Constrained Optimization**.
2. *Global Convergence Strategy.* The (local) subproblem alone cannot guarantee convergence to a stationary point and optimization methods need a convergence strategy that forces the sequence $\{x_k\}$ to converge. Examples of convergence strategies are penalty functions, augmented Lagrangian, filter and funnel methods; see the article titled **Foundations of Constrained Optimization**.
3. *Global Convergence Mechanism.* The global convergence strategy is enforced by using a mechanism that improves the subproblem by shortening the step that is computed. There are two classes of convergence mechanism: line search and trust region.
4. *Convergence Test.* A convergence test is needed to stop the algorithm once the error in the Karush–Kuhn–Tucker (KKT) (or Fritz–John (FJ)) conditions is below a user-specified tolerance. In addition, solvers may converge only to a local minimum of the constraint violation.

Solvers for NCO are differentiated by how each of these key ingredients is implemented. In addition there are a number of

secondary distinguishing factors such as licensing (open-source versus commercial or academic), API and interfaces to modeling languages, sparse or dense linear algebra, programming language (Fortran/C/MATLAB), and compute platforms on which a solver can run.

The next sections list some solvers for NCOs, summarizing their main characteristics. A short overview can be found in Table 1. We distinguish solvers by the definition of

their subproblem. An alternative characterization could be based on the way in which solvers handle constraints.

INTERIOR-POINT SOLVERS

Interior-point methods approximately solve a sequence of perturbed KKT systems, driving a barrier parameter to zero. Interior-point methods can be regarded as perturbed

Table 1. NCO Software Overview

Name	Model	Global Method	Interfaces	Language
ALGENCAN	Augmented Lagrangian	Penalty	AMPL, C/C++, CUTer, Java, Matlab, Octave, Python, R	f77
CONOPT	GRG/SLQP	Line search	AIMMS, GAMS	Fortran
FilterSQP	SQP	Filter/trust region	AMPL, CUTer, f77	Fortran 77
GALAHAD	Augmented Lagrangian	Nonmonotone/augmented Lagrangian	CUTer, fortran	Fortran95
Ipfilter	IPM	Filter/line search	AMPL, CUTer, f90	Fortran 90
IPOPT	IPM	Filter/line search	AMPL, CUTer, C, C++, f77	C++
KNITRO	IPM	Penalty/trust region line search	AIMMS, AMPL, GAMS, Mathematica, MATLAB, MPL, C, C++, f77, Java, Excel	C
KNITRO	SLQP	Penalty/trust region	s.a.	C
LANCELOT	Augmented Lagrangian	augmented Lagrangian/trust region	SIF, AMPL, f77	f77
LINDO	GRG/SLP	Only convex	C, MATLAB, LINGO	
LOQO	IPM	Line search	AMPL, C, MATLAB	C
LRAMBO	SQP	ℓ_1 exact penalty/line search	C	C/C++
NLPQLP	SQP	Augmented Lagrangian/line search	C, f77, MATLAB	f77
MINOS	Augmented Lagrangian	None	AIMMS, AMPL, GAMS, MATLAB, C, C++, f77	f77
PATH	LCP	Line search	AMPL	C
PENNON	Augmented Lagrangian	Line search	AMPL, MATLAB	C
NPSOL	SQP	Penalty Lagrangian/line search	AIMMS, AMPL, GAMS, MATLAB, C, C++, f77	f77
SNOPT	SQP	Penalty Lagrangian/line search	AIMMS, AMPL, GAMS, MATLAB, C, C++, f77	f77
SQPlab	SQP	Penalty Lagrangian/line search	MATLAB	MATLAB

Newton methods applied to the KKT system in which the primal/dual variables are kept positive.

Ipfilter

Algorithmic Methodology. Ipfilter is based on the interior-point filter method of Ulbrich *et al.* [3]. A step is first generated by solving the normally used KKT system. It then uses a specially designed filter for selecting the step size in the direction of the calculated step. The first component of the filter is the sum of “constraint violation” and “centrality,” while the second component is the norm of the gradient of the Lagrangian. This method can sometimes converge to a local maximum because the filter does not explicitly check the objective function value. In the implementation of Ipfilter, the second component of the filter is modified to overcome this problem [4].

Software and Technical Details. Ipfilter version 0.3 is written in Fortran90. It requires a user to have MA57 routine from the HSL library. Ipfilter is available at no cost for academic purposes and can be obtained by contacting the developers at the Ipfilter web page [5]. It has interfaces in AMPL and CUTEr. The user can also write their own functions and derivatives in Fortran90.

IPOPT

Algorithmic Methodology. IPOPT is a line-search filter interior-point method [6–8]. The outer loop approximately minimizes a sequence of nonlinearly (equality-) constrained barrier problems for a decreasing sequence of barrier parameters. The inner loop uses a line-search filter SQP method to approximately solve each barrier problem. Global convergence of each barrier problem is enforced through a line-search filter method, and the filter is reset after each barrier parameter update. Steps are computed by solving a primal–dual system, corresponding to the KKT conditions of the barrier problem. The algorithm controls the inertia of this system by adding δI ($\delta > 0$) to the Hessian of the Lagrangian, ensuring descent properties. The inner iteration includes second-order correction steps and mechanisms for switching to a feasibility

restoration if the step size becomes too small. The solver has an option for using limited-memory BFGS updates to approximate the Hessian of the Lagrangian.

Software and Technical Details. The solver is written in C++ and has interfaces to C, C++, Fortran, AMPL, CUTEr, and COIN-OR’s NLPAPI. It requires BLAS, LAPACK, and a sparse indefinite solver (MA27, MA57, PARDISO, or WSMP). The user (or the modeling language) must provide function and gradient information, and possibly the Hessian of the Lagrangian (in sparse format). By using the PARDISO [9] parallel linear solver, IPOPT can solve large problems on shared-memory multiprocessors. IPOPT is available on COIN-OR under Common Public License at its website [10].

KNITRO

KNITRO includes both IPM and SLQP methods. The IPM version is described in this section.

Algorithmic Methodology. KNITRO implements both trust region–based and line-search-based interior-point/barrier methods [11–13]. It approximately solves a sequence of barrier subproblems for a decreasing sequence of barrier parameters. It uses a merit or a penalty function for accepting a step and for ensuring global convergence. The barrier subproblems are solved by a sequence of linearized primal–dual equations. KNITRO has two options for solving the primal–dual system: direct factorization of the system of equations and preconditioned conjugate-gradient (PCG) method. The PCG method solves the indefinite primal–dual system by projecting onto the null space of the equality constraints. KNITRO also includes modules for solving mixed-integer nonlinear optimization problems and optimization problems with simple complementarity constraints. In addition, it contains crossover techniques to obtain an active set from the solution of the IPM solve and a multistart heuristic for nonconvex NCOs.

Software and Technical Details. KNITRO is written in C. It has interfaces to a range

of modeling languages, including AIMMS, AMPL, GAMS, Mathematica, MATLAB, and MPL. In addition, KNITRO has interfaces to C, C++, Fortran, Java, and Excel. It offers callback and reverse communication interfaces. KNITRO makes use of MA27/MA57 to solve the indefinite linear systems of equations, and these routines are included in the package. KNITRO is available from Ziena Optimization, Inc., and the user's manual is available on-line [14].

LOQO

Algorithmic Methodology. LOQO [15] uses an infeasible primal–dual interior-point method for solving general nonlinear problems. The inequality constraints are added to the objective by using a log-barrier function. Newton's method is used to obtain a solution to the system of nonlinear equations that is obtained by applying first-order necessary conditions to this barrier function. The solution of this system provides a search direction, and a filter or a merit function is used to accept the next iterate obtained by performing a line search in this direction. The filter is specially modified for interior-point methods [16]. It allows a point that improves either the feasibility or the barrier objective as compared to the current iterate only. The merit function is the barrier function and an additional ℓ_2 norm of the violation of constraints. The exact Hessian of the Lagrangian is used in the Newton's method. When the problem is nonconvex, this Hessian is perturbed by adding to it the matrix δI , where I is an identity matrix and $\delta > 0$ is chosen so that the perturbed Hessian is positive definite. This perturbation ensures that if the iterates converge, they converge to a local minimum. LOQO can also handle complementarity constraints by using a penalty function [17].

Software and Technical Details. LOQO can be used to solve problems written in AMPL. The user can also implement functions in C and link to the LOQO library. LOQO can also read files in MPS format for solving LPs. A MATLAB interface for LOQO is available for LPs, QPs, and second-order cone programs. The library and binary files are

available for a license fee at its homepage [18], while a limited student version is available freely. A user manual provides instructions on installing and using LOQO.

SEQUENTIAL LINEAR/QUADRATIC SOLVERS

Sequential linear/quadratic solvers solve a sequence of LP and/or QP subproblems. This class of solvers is also referred to as active-set methods, because they provide an estimate of the active set at every iteration. Once the active set settles down, these methods become Newton methods on the active constraints (except SLP).

CONOPT

Algorithmic Methodology. CONOPT [19,20] implements three active-set methods. The first is a gradient projection method that projects the gradient of the objective onto a linearization of the constraints and makes progress toward the solution by reducing the objective. The second variant is an SLP method, and the third is an SQP method. CONOPT includes algorithmic switches that automatically detect which method is preferable. It also exploits triangular parts of the Jacobian matrix and linear components of the Jacobian to reduce the sparse factorization overhead. CONOPT is a line-search method.

Software and Technical Details. CONOPT has interfaces to AMPL, GAMS, and AIMMS and is available from any of the three companies.

FilterSQP

Algorithmic Methodology. FilterSQP [21] implements a trust-region SQP method. Convergence is enforced with a filter, whose components are the ℓ_1 -norm of the constraint violation, and the objective function. The solver starts by projecting the user-supplied initial guess onto the linear part of the feasible set and remains feasible with respect to the linear constraints. It uses an indefinite QP solver that finds local solutions to problems with negative curvature. The solver switches to a feasibility restoration phase

if the local QP model becomes inconsistent. During the restoration phase, a weighted sum of the constraint violations is minimized by using a filter SQP trust-region approach. The restoration phase either terminates at a local minimum of the constraint violation or returns to the main algorithm when a feasible local QP model is found. FilterSQP computes second-order correction steps if the QP step is rejected by the filter.

Software and Technical Details. The solver is implemented in Fortran77 and has interfaces to AMPL, CUTer, and Fortran77. The code requires subroutines to evaluate the problem functions, their gradients, and the Hessian of the Lagrangian (provided by AMPL and CUTer). It uses BQPD [22] to solve the possibly indefinite QP subproblems. BQPD is a null-space active-set method and has modules for sparse and dense linear algebra with efficient and stable factorization updates. The code is licensed by the University of Dundee.

KNITRO

Algorithmic Methodology. In addition to the interior-point methods (see the section titled “KNITRO” of the section titled “Interior-Point Solver”), KNITRO implements an SLQP algorithm [23,24]. In each iteration, an LP that approximates the ℓ_1 exact-penalty problem is solved to determine an estimate of the active set. This LP also has an additional infinity-norm trust-region constraint. Those constraints in the LP that are satisfied as equalities are marked as active and are used to set up an equality-constrained quadratic programming (EQP), whose objective is a quadratic approximation of the Lagrangian of Equation (1) at the current iterate. An ℓ_2 -norm trust-region constraint is also added to this QP. The solution of the LP and the EQP are then used to find a search direction [13,24]. A projected conjugate-gradient method is used to solve the EQP. The penalty parameter ρ_k is updated to ensure sufficient decrease toward feasibility.

Software and Technical Details. The SLQP method of KNITRO requires an LP solver.

KNITRO includes CLP as well as a link to CPLEX as options. For other details, the section titled “KNITRO” of the section titled “Interior-Point Solver”. This active-set algorithm is usually preferable for rapidly detecting infeasible problems and for solving a sequence of closely related problems.

LINDO

Algorithmic Methodology. LINDO provides solvers for a variety of problems including NCOs. Its nonlinear solver implements an SLP algorithm and a generalized gradient method. It provides options to randomly select the starting point and can estimate derivatives by using finite differences.

Software and Technical Details. LINDO [25] provides interfaces for programs written in languages such as C and MATLAB. It can also read models written in LINGO. The full version of the software can be bought on-line, and a limited trial version is available for free [26].

LRAMBO

Algorithmic Methodology. LRAMBO is a *total quasi-Newton* SQP method. It approximates both the Jacobian and the Hessian by using rank-1 quasi-Newton updates. It enforces global convergence through a line search on an ℓ_1 exact-penalty function. LRAMBO requires as input only an evaluation program for the objective and the constraints. It combines automatic differentiation (AD) and quasi-Newton updates to update factors of the Jacobian. And it computes updates of the Hessian. Details can be found in Griewank *et al.* [27].

Software and Technical Details. LRAMBO uses the NAG subroutine E04NAF to solve the QP subproblems. It also requires ADOL-C [28,29] for the AD of the problem functions. LRAMBO is written in C/C++.

NLPQLP

Algorithmic Methodology. NLPQLP is an extension of the SQP solver NLPQL [30] that

implements a nonmonotone line search to ensure global convergence. It uses a quasi-Newton approximation of the Hessian of the Lagrangian, which is updated with the BFGS formula. A nonmonotone line search is used to calculate the step length that minimizes an augmented Lagrangian merit function.

Software and Technical Details. NLPQLP is implemented in Fortran. The user or the interface is required to evaluate the function values of the objective and constraints. Derivatives, if unavailable, can be estimated by using finite differences. NLPQLP can evaluate these functions and also the merit function on a distributed memory system. A user guide with documentation, algorithm details, and examples is available [31]. NLPQLP can be obtained under academic and commercial license from its website.

NPSOL

Algorithmic Methodology. NPSOL [32] solves general nonlinear problems by using an SQP algorithm with a line search on the augmented Lagrangian. In each major iteration, a QP subproblem is solved whose Hessian is a quasi-Newton approximation of the Hessian of the Lagrangian using a *dense* BFGS update.

Software and Technical Details. NPSOL is implemented in Fortran, and the library can be called from Fortran and C programs and from modeling languages such as AMPL, GAMS, AIMMS, and MATLAB. The user or the interface must provide routines for evaluating functions and their gradients. If a gradient is not available, NPSOL can estimate it by using finite differences. It treats all matrices as dense and hence may not be efficient for large-sparse problems. NPSOL can be warm started by specifying the active constraints and multiplier estimates for the QP. NPSOL is available under commercial and academic licenses from Stanford Business Software Inc.

PATHNLP

Algorithmic Methodology. PATH [33,34] solves mixed complementarity problems

(MCPs). PATHNLP automatically formulates the KKT conditions of an NLP, Equation (1), specified in GAMS as an MCP and then solves this MCP using PATH. The authors note that this approach is guaranteed only to find stationary points and does not distinguish between local minimizers and maximizers for nonconvex problems. Thus, it works well for convex problems. At each iteration, PATH solves a linearization of the MCP problem to obtain a Newton point. It then performs a search in the direction of this point to find a minimizer of a merit function. If this direction is not a descent direction, it performs a steepest descent step in the merit function to find a new point.

Software and Technical Details. The PATHNLP is available only through GAMS [35]. The PATH solver is implemented in C and C++ [36].

SNOPT

Algorithmic Methodology. SNOPT [37] implements an SQP algorithm much like NPSOL, but it is suitable for large, sparse problems as well. The Hessian of the Lagrangian is updated by using limited-memory quasi-Newton updates. SNOPT solves each QP using SQOPT [38], which is a reduced-Hessian active-set method. It includes an option for using a projected conjugate-gradient method rather than factoring the reduced Hessian. SNOPT starts by solving an “elastic program” that minimizes the constraint violation of the linear constraints of Equation (1). The solution to this program is used as a starting point for the major iterations. If a QP subproblem is found to be infeasible or unbounded, then SNOPT tries to solve an elastic problem that corresponds to a smooth reformulation of the ℓ_1 exact-penalty function. The solution from a major iteration is used to obtain a search direction along which an augmented Lagrangian merit function is minimized.

Software and Technical Details. SNOPT is implemented in Fortran77 and is compatible with newer Fortran compilers. All functions

in the SNOPT library can be used in parallel or by multiple threads. It can be called from other programs written in C and Fortran and by packages such as AMPL, GAMS, AIMMS, and MATLAB. The user or the interface has to provide routines that evaluate function values and gradients. When gradients are not available, SNOPT uses finite differences to estimate them. SNOPT can also save basis files that can be used to save the basis information to warm start subsequent QPs. SNOPT is available under commercial and academic licenses from Stanford Business Software Inc.

SQPlab

Algorithmic Methodology. SQPlab [39] was developed as a laboratory for testing algorithmic options of SQP methods [40]. SQPlab implements a line-search SQP method that can use either exact Hessians or a BFGS approximation of the Hessian of the Lagrangian. It maintains positive definiteness of the Hessian approximation using the Wolfe or Powell condition. SQPlab uses a Lagrangian penalty function, $p_\sigma(x, y) = f(x) + y_c^T c(x) + \sigma \| \max\{0, c(x) - u_c, l_c - c(x)\} \|_1$, to promote global convergence. SQPlab has a feature that allows it to treat discretized optimal control constraints specially. The user can specify a set of equality constraints whose Jacobian has uniformly full rank (such as certain discretized optimal-control constraints). SQPlab then uses these constraints to eliminate the state variables. The user needs to provide only routines that multiply a vector by the inverse of the control constraints.

Software and Technical Details. SQPlab is written in MATLAB and requires `quadprog.m` from the optimization toolbox. The user must provide a function simulator to evaluate the functions and gradients. A smaller version is also available that does not require `quadprog.m` but instead uses `qpal.m`, an augmented Lagrangian QP solver for medium-sized convex QPs. All solvers are distributed under the Q public license from INRIA, France [41].

AUGMENTED LAGRANGIAN SOLVERS

Augmented Lagrangian methods solve Equation (1) by a sequence of subproblems that minimize the augmented Lagrangian, either subject to a linearization of the constraints or as a bound-constrained problem.

ALGENCAN

Algorithmic Methodology. ALGENCAN is a solver based on an augmented Lagrangian-type algorithm [42,43] in which a bound-constrained problem is solved in each iteration. The objective function in each iteration is the augmented Lagrangian of the original problem. This subproblem is solved by using a quasi-Newton method. The penalty on each constraint is increased if the previous iteration does not yield a better point.

Software and Technical Details. ALGENCAN is written in Fortran77, and interfaces are available for AMPL, C/C++, CUTer, JAVA, MATLAB, Octave, Python, and R. The bound-constrained problem in each iteration is solved by using GENCAN, developed by the same authors. ALGENCAN and GENCAN are freely available under the GNU Public License from the TANGO project web page [44].

GALAHAD

Algorithmic Methodology. GALAHAD [45] contains a range of solvers for large-scale nonlinear optimization. It includes LANCELOT B, an augmented Lagrangian method with a nonmonotone descend condition; FILTRANE, a solver for feasibility problems based on a multidimensional filter [46]; and interior-point and active-set methods for solving large-scale quadratic programs (QPs). GALAHAD also contains a presolve routine for QPs [47] and other support routines.

Software and Technical Details. GALAHAD is a collection of thread-safe Fortran95 packages for large-scale nonlinear optimization. It is available in source form at <http://galahad.rl.ac.uk/> [48]. GALAHAD has links to CUTer and Fortran.

LANCELOT

Algorithmic Methodology. LANCELOT [49] is a large-scale implementation of a bound-constrained augmented Lagrangian method. It approximately solves a sequence of bound-constrained augmented Lagrangian problems by using a trust-region approach. Each trust-region subproblem is solved approximately: first, the solver identifies a Cauchy-point to ensure global convergence, and then it applies conjugate-gradient steps to accelerate the local convergence. LANCELOT provides options for a range of quasi-Newton Hessian approximations suitable for large-scale optimization by exploiting the group-partial separability of the problem functions.

Software and Technical Details. LANCELOT is written in standard ANSI Fortran77 and has been interfaced to CUTer [50] and AMPL. The distribution includes installation scripts for a range of platforms in single and double precision. LANCELOT is available freely from Rutherford Appleton Laboratory, UK [51].

MINOS

Algorithmic Methodology. MINOS [52] uses a projected augmented Lagrangian for solving general nonlinear problems. In each “major iteration,” a linearly constrained nonlinear problem is solved where the linear constraints constitute all the linear constraints of Equation (1) and also linearizations of nonlinear constraints. This problem is in turn solved iteratively by using a reduced-gradient algorithm along with a quasi-Newton algorithm. The quasi-Newton algorithm provides a search direction along which a line search is performed to improve the objective function and reduce the infeasibilities. MINOS does not guarantee convergence from a remote starting point. The user should therefore specify a starting point that is “close enough” for convergence. This issue will be addressed in a new software Knossos that will use a stabilized version of the linearly constrained Lagrangian method [53]. The user can also modify other parameters such as the penalty parameter

in augmented Lagrangian to control the algorithm.

Software and Technical Details. MINOS is implemented in Fortran. The library can be called from Fortran and C programs and interfaces such as AMPL, GAMS, AIMMS, and MATLAB. The user must input routines for the function and gradient evaluations. If a gradient is not available, MINOS estimates it by using finite differences. MINOS is available under commercial and academic licenses from Stanford Business Software Inc.

PENNON

Algorithmic Methodology. PENNON [54] is an augmented Lagrangian penalty-barrier method. In addition to standard NCOs, it can handle semidefinite NCOs that *include a semidefiniteness constraint on matrices of variables*. PENNON represents the semidefiniteness constraint by computing an eigenvalue decomposition and adds a penalty-barrier for each eigenvalue to the augmented Lagrangian. It solves a sequence of unconstrained optimization problems in which the inequality constraints appear in barrier functions and the equality constraints in penalty functions. Every unconstrained minimization problem is solved with Newton’s method. The solution of this problem provides a search direction along which a suitable merit function is minimized. Even though the algorithm has not been shown to converge for nonconvex problems, it has been reported to work well for several problems.

Software and Technical Details. PENNON can be called from either AMPL or MATLAB. The user manual [55] provides instructions for input of matrices in sparse format when using AMPL. PENNON includes both sparse factorizations using the algorithm of Ng and Peyton [56] and dense factorization using LAPACK. Commercial licenses and a free academic license are available from PENOPT-GbR.

TERMINATION OF NCO SOLVERS

Solvers for NCOs can terminate in a number of ways. Normal termination

corresponds to a KKT point, but solvers can also detect (locally) infeasible NCOs and unboundedness. Some solvers also detect failure of constraint qualifications. Solvers also may terminate because of errors, and we provide some simple remedies.

Normal Termination at KKT Points

Normal termination corresponds to termination at an approximate KKT point, that is, a point at which the norm of the constraint violation, the norm of the first-order necessary conditions, and the norm of the complementary slackness condition are less than a user-specified tolerance. Some solvers divide the first-order conditions by the modulus of the largest multiplier. If the problem is convex (and feasible), then the solution corresponds to a global minimum of Equation (1). Many NCO solvers include sufficient decrease conditions that make it less likely to converge to local maxima or saddle points if the problem is nonconvex.

Termination at Other Critical Points

Unlike LP solvers, NCO solvers do not guarantee global optimality for general nonconvex NCOs. Even deciding whether a problem is feasible or unbounded corresponds to a global optimization problem. Moreover, if a constraint qualification fails to hold, then solvers may converge to critical points that do not correspond to KKT points.

Fritz–John (FJ) Points. This class of points corresponds to first-order points at which a constraint qualification may fail to hold. NCO solvers adapt their termination criterion by dividing the stationarity condition by the modulus of the largest multiplier. This approach allows convergence to FJ points that are not KKT points. We note that dividing the first-order error is equivalent to scaling the objective gradient by a constant $0 \leq \alpha \leq 1$. As the multipliers diverge to infinity, this constant converges to $\alpha \rightarrow 0$, giving rise to an FJ point.

Locally Inconsistent Solutions. To prove that an NCO is locally inconsistent requires the (local) solution of a feasibility problem,

in which a norm of the nonlinear constraint residuals is minimized subject to the linear constraints. A KKT point of this problem provides a certificate that Equation (1) is locally inconsistent. There are two approaches to obtaining this certificate. Filter methods switch to a feasibility restoration phase if either the stepsize becomes too small or the LP/QP subproblem becomes infeasible. Penalty function methods do not switch but drive the penalty parameter to infinity.

Unbounded Solutions. Unlike the case of LP where a ray along the direction of descent is necessary and sufficient to prove that the instance is unbounded, it is difficult to check whether a given nonlinear program is unbounded. Most NCO solvers use a user-supplied lower bound on the objective and terminate if they detect a feasible point with a lower objective value than the lower bound.

Remedies If Things Go Wrong

If a solver stops at a point where the constraint qualifications appears to fail (indicated by large multipliers) or where the nonlinear constraints are locally inconsistent or at a local and not global minimum, then one can restart the solver procedure from a different initial point. Some solvers include automatic random restarts.

Another cause for failure is errors in the function, gradient, or Hessian evaluation (e.g., IEEE exceptions). Some solvers provide heuristics that backtrack if IEEE exceptions are encountered during the iterative process. In many cases, IEEE exceptions occur at the initial point, and backtracking cannot be applied. It is often possible to reformulate the nonlinear constraints to avoid IEEE exceptions. For example, $\log(x_1^3 + x_2)$ will cause IEEE exceptions if $x_1^3 + x_2 \leq 0$. Adding the constraint $x_1^3 + x_2 \geq 0$ does not remedy this problem, because nonlinear constraints may not be satisfied until the limit. A better remedy is to introduce a nonnegative slack variable, $s \geq 0$ and the constraint $s = x_1^3 + x_2$, and then replace $\log(x_1^3 + x_2)$ by $\log(s)$. Many NCO solvers will honor simple bounds (and interior-point solvers guarantee $s > 0$), so this formulation avoids some IEEE exceptions.

CALCULATING DERIVATIVES

All solvers described above expect either the user or the modeling environment (AMPL, GAMS, etc.) to provide first-order and sometimes second-order derivatives. Some solvers can estimate first-order derivatives by finite differences. If the derivatives are not available or if their estimates are not reliable, one can use several AD tools that are freely available. We refer the readers to Griewank [57] for principles and methods of AD and to More [58] for using AD in nonlinear optimization. AD tools include ADOL-C [29]; ADIC, ADIFOR, and OpenAD [59]; and Tapenade [60].

WEB RESOURCES

In addition to a range of NCO solvers that are available on-line, there exist an increasing number of web-based resources for optimization. The NEOS server for optimization [61,62] allows users to submit optimization problems through a web interface, using a range of modeling languages. It provides an easy way to try out different solvers. The NEOS wiki [63] provides links to optimization case studies and background on optimization solvers and problem classes. The COIN-OR project [64] is a collection of open-source resources and tool for operations research, including nonlinear optimization tools (IPOPT and ADOL-C). A growing list of test problems for NCOs includes the CUTer collection [65], Bob Vanderbei's collection of testproblems [66], the COPS testset [67], and the GAMS model library [68]. The NLP-FAQ [69] provides on-line answers to questions on nonlinear optimization.

Acknowledgments

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, US Department of Energy, under Contract DE-AC02-06CH11357. This work was also supported by the US Department of Energy through the grant DE-FG02-05ER25694.

REFERENCES

1. Andersen ED, Jensen B, Jensen J, *et al.* Mosek version 6. Technical Report TR-2009-3. MOSEK; 2009.
2. Dahl J, Vandenberghe L. CVXOPT user's guide. 2010. Available at <http://abel.ee.ucla.edu/cvxopt/userguide/>.
3. Ulbrich M, Ulbrich S, Vicente L. A globally convergent primal-dual interior-point filter method for nonconvex nonlinear programming. *Math Program* 2004;100:379–410.
4. Ulbrich M, Ulbrich S, Vicente LN. A globally convergent primal-dual interior-point filter method for nonlinear programming: new filter optimality measures and computational results. Technical Report 08-49. Department of Mathematics, University of Coimbra; 2008.
5. Silva R, Ulbrich M, Ulbrich S, *et al.* Ipfilter. 2010. Available at <http://www.mat.uc.pt/ipfilter/>.
6. Wächter A, Biegler L. Line search filter methods for nonlinear programming: motivation and global convergence. *SIAM J Optim* 2005;16(1):1–31.
7. Wächter A, Biegler L. Line search filter methods for nonlinear programming: local convergence. *SIAM J Optim* 2005;16(1):32–48.
8. Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program* 2006;106(1):25–57.
9. Schenk O, Waechter A, Hagemann M. Matching-based preprocessing algorithms to the solution of saddle-point problems in large-scale nonconvex interior-point optimization. *J Comput Optim Appl*. 2007;36(2-3):321–341. Available at <http://www.pardiso-project.org/>.
10. Wächter A, Biegler L. IPOPT project. 2010. Available at <https://projects.coin-or.org/Ipopt>.
11. Byrd RH, Hribar ME, Nocedal J. An interior point algorithm for large scale nonlinear programming. *SIAM J Optim* 1999;9(4):877–900.
12. Byrd RH, Gilbert JC, Nocedal J. A trust region method based on interior point techniques for nonlinear programming. *Math Program* 2000;89:149–185.
13. Byrd RH, Nocedal J, Waltz RA. Knitro: an integrated package for nonlinear optimization. In: di Pillo G, Roma M, editors. *Large-scale nonlinear optimization*. New York: Springer; 2006. pp. 35–59.
14. Waltz R, Plantenga T. KNITRO user's manual. 2009. Available at <http://ziena.com/documentation.htm>.

15. Vanderbei R, Shanno D. An interior point algorithm for nonconvex nonlinear programming. *COAP* 1999;13:231–252.
16. Benson H, Shanno D, Vanderbei R. Interior-point methods for nonconvex nonlinear programming: filter-methods and merit functions. *Comput Optim Appl* 2002;23(2): 257–272.
17. Benson H, Sen A, Shanno D, *et al.* Interior-point algorithms, penalty methods, and equilibrium constraints. *Comput Optim Appl* 2006;34(2):155–182.
18. Vanderbei R. LOQO. 2010. Available at <http://www.princeton.edu/~rvdb/loqo/LOQO.html>.
19. Drud A. A GRG code for large sparse dynamic nonlinear optimization problems. *Math Program* 1985;31:153–191.
20. Drud A. CONOPT. Bagsvaerd, Denmark: ARKI Consulting and Development, A/S; 2007. Available at <http://www.gams.com/dd/docs/solvers/conopt.pdf>.
21. Fletcher R, Leyffer S. User manual for filter SQP. Numerical Analysis Report NA/181. University of Dundee; 1998.
22. Fletcher R. Stable reduced Hessian updates for indefinite quadratic programming. Numerical Analysis Report NA/187. Scotland, UK: University of Dundee, Department of Mathematics; 1999.
23. Fletcher R, de la Maza ES. Nonlinear programming and nonsmooth optimization by successive linear programming. *Math Program* 1989;43:235–256.
24. Byrd RH, Gould NIM, Nocedal J, *et al.* An algorithm for nonlinear optimization using linear programming and equality constrained subproblems. *Math Program, B* 2004;100(1): 27–48.
25. LINDO Systems Inc. LINDO API 6.0 user manual. 2010.
26. LINDO Systems Inc. 2010. Available at <http://www.lindo.com/>.
27. Griewank A, Walther A, Korzec M. Maintaining factorized KKT systems subject to rank-one updates of Hessians and Jacobians. *Optim Methods Softw* 2007;22(2):279–295.
28. Griewank A, Juedes D, Mitev H, *et al.* ADOL-C: a package for the automatic differentiation of algorithms written in C/C++. *ACM TOMS* 1996;22(2):131–167.
29. Griewank A, Walther A. ADOL-C a package for automatic differentiation of algorithms written in C/C++. 2004. <https://projects.coin-or.org/ADOL-C>.
30. Schittkowski K. NLPQL: a Fortran subroutine for solving constrained nonlinear programming problems. *Ann Oper Res* 1985;5(2): 485–500.
31. Schittkowski K. NLPQLP: A Fortran implementation of a sequential quadratic programming algorithm with distributed and non-monotone line search—user's guide, Version 3.1. Report. Department of Computer Science, University of Bayreuth; 2009. Available at <http://www.math.uni-bayreuth.de/~kschittkowski/nlpqlp.htm>.
32. Gill P, Murray W, Saunders M, *et al.* User's guide for NPSOL Version 5.0: a Fortran package for nonlinear programming. Report SOL 86-1. San Diego (CA): Department of Mathematics, University of California; 1998.
33. Dirkse S, Ferris M. The PATH solver: a non-monotone stabilization scheme for mixed complementarity problems. *Optim Methods Softw* 1995;5:123–156.
34. Ferris M, Munson T. Interfaces to PATH 3.0: design, implementation and usage. *Comput Optim Appl* 1999;12:207–227.
35. Dirkse S, Ferris M. PATHNLP. GAMS. 2010. Available at <http://www.gams.com/dd/docs/solvers/pathnlp.pdf>.
36. Dirkse S, Ferris M. PATH-AMPL binaries. 2010. <ftp://ftp.cs.wisc.edu/math-prog/solvers/path/AMPL/>.
37. Gill P, Murray W, Saunders M. User's guide for SNOPT Version 7: software for large-scale nonlinear programming. Report. San Diego: Department of Mathematics, University of California; 2006.
38. Gill P, Murray W, Saunders M. User's guide for SQOPT Version 7: software for large-scale nonlinear programming. Report. San Diego (CA): Department of Mathematics, University of California; 2006.
39. Gilbert JC. SQPlab—A MATLAB software for solving nonlinear optimization problems and optimal control problems. Technical Report. Le Chesnay Cedex, France: INRIA; 2009.
40. Bonnans J, Gilbert J, Lemaréchal C, *et al.* Numerical optimization: theoretical and practical aspects, 2nd ed. Berlin: Springer; 2006.
41. Gilbert JC. SQPlab. 2010. Available at <http://www-rocq.inria.fr/~gilbert/modulopt/optimization-routines/sqplab/sqplab.html>.
42. Andreani R, Birgin E, Martinez J, *et al.* On augmented Lagrangian methods with general lower-level constraints. *SIAM J Optim* 2007;18:1286–1309.

43. Andreani R, Birgin E, Martinez J, *et al.* Augmented Lagrangian methods under the constant positive linear dependence constraint qualification. *Math Program* 2008;111: 5–32.
44. Martinez J, Birgin E. TANGO: trustable algorithms for nonlinear general optimization. 2010. Available at <http://www.ime.usp.br/~egbirgin/tango/index.php>.
45. Gould NIM, Orban D, Toint PL. GALAHAD, a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. *ACM Trans Math Softw* 2004;29(4):353–372.
46. Gould NIM, Leyffer S, Toint PL. A multidimensional filter algorithm for nonlinear equations and nonlinear least squares. *SIAM J Optim* 2004;15(1):17–38.
47. Gould NIM, Toint PL. Presolving for quadratic programming. *Math Program* 2004;100(1): 95–132.
48. Gould NIM, Orban D, Toint PL. GALAHAD. Rutherford Appleton Laboratory; 2002. Available at <http://galahad.rl.ac.uk/>.
49. Conn AR, Gould NIM, Toint PL. LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A). Heidelberg: Springer; 1992.
50. Bongartz I, Conn AR, Gould NIM, *et al.* CUTE: constrained and unconstrained testing environment. *ACM Trans Math Softw* 1995;21:123–160. <http://cuter.rl.ac.uk/cuter-www/interfaces.html>.
51. Conn AR, Gould NIM, Toint PL. LANCELOT. 2010. Available at <http://www.cse.scitech.ac.uk/nag/lancelot/lancelot.shtml>.
52. Murtagh B, Saunders M. MINOS 5.4 user's guide. Report SOL 83-20 R. Department of Operations Research, Stanford University; 1998.
53. Friedlander MP, Saunders MA. A globally convergent linearly constrained Lagrangian method for nonlinear optimization. *SIAM J Optim* 2005;15(3):863–897.
54. Kocara M, Stingl M. PENNON—a code for convex nonlinear and semidefinite programming. *Optim Methods Softw* 2003;18(3): 317–333.
55. Kocara M, Stingl M. PENNON user's guide (version 0.9). 2008. Available at <http://www.penopt.com>.
56. Ng E, Peyton B. Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM J Sci Comput* 1993;14(5):1034–1056.
57. Griewank A. Evaluating derivatives: principles and techniques of algorithmic differentiation. Philadelphia (PA): SIAM; 2000.
58. Moré J. Automatic differentiation tools in optimization software. Technical Report ANL/MCS-P859-1100. Mathematics and Computer Science Division, Argonne National Laboratory; 2000.
59. Hovland P, Lyons A, Narayanan SHK, *et al.* Automatic differentiation at Argonne. 2009. Available at <http://wiki.mcs.anl.gov/autodiff/>.
60. Hascoët L, Pascual V. TAPENADE 2.1 user's guide. Technical Report 0300. INRIA; 2004.
61. Czyzyk J, Mesnier M, Moré J. The NEOS server. *IEEE J Comput Sci Eng* 1998;5:68–75. Available at www-neos.mcs.anl.gov/neos/.
62. NEOS. The NEOS server for optimization. Argonne National Laboratory; 2003. Available at <http://www-neos.mcs.anl.gov/neos/>.
63. Leyffer S, Wright S. NEOS wiki. Argonne National Laboratory; 2008. Available at <http://wiki.mcs.anl.gov/NEOS/>.
64. COINOR. COIN-OR: computational infrastructure for operations research. 2009. <http://www.coin-or.org/>.
65. Gould N, Orban D, Toint P. CUTER. 2010. Available at <http://cuter.rl.ac.uk/cuter-www/>.
66. Vanderbei R. Nonlinear optimization models. 2010. Available at <http://www.orfe.princeton.edu/~rvdb/ampl/nlmodels/>.
67. Dolan ED, Moré J, Munson TS. COPS large-scale optimization problems. 2010. Available at <http://www.mcs.anl.gov/~more/cops/>.
68. GAMS. The GAMS model library index. 2010. Available at <http://www.gams.com/modlib/modlib.htm>.
69. Fourer R. Nonlinear programming frequently asked questions. 2010. Available at http://wiki.mcs.anl.gov/NEOS/index.php/Nonlinear_Programming_FAQ.

SOFTWARE FOR SOLVING NONCOOPERATIVE STRATEGIC FORM GAMES

THEODORE L. TUROCY
School of Economics,
University of East Anglia,
Norwich, UK

INTRODUCTION AND OVERVIEW

Game theory is an elegant tool for modeling and analyzing strategic interaction among multiple agents. Over the last half-century, it has proven its worth and flexibility across disciplines from economics to computer science to operations research. This flexibility comes with some costs. It can be unwieldy to write down a formal specification of a noncooperative strategic form game, even for relatively small numbers of players and strategies available to those players. Furthermore, once the game is specified, computing predictions for how the game will or ought to be played, especially the Nash equilibrium (or equilibria) of the game, is tedious to do by hand for even small games, and completely impractical for games of moderate size. To move beyond toy examples and investigate models that are useful in operations and management contexts, software tools for specifying and solving noncooperative strategic form games are essential.

The good news for the practitioner interested in solving particular games is that the Nash equilibrium of a game can be formulated in terms of a variety of mathematical programming problems. These programming problems, in turn, can be solved using a combination of off-the-shelf general-purpose libraries, as well as implementations that may take advantage of the special structure of the programming problems that arise specifically from games.

This article first gives a high-level overview of Gambit, an integrated package

for doing computation on finite games. The main body of the article provides a “field guide” to computing equilibria. This guide mixes an overview of the available algorithms and implementations with general advice on techniques and reasonable expectations for solving games. In a brief conclusion, some current and anticipated future directions for research in this area over the next few years are highlighted.

GAMBIT

The most comprehensive package available for computing Nash equilibria in finite games is Gambit, available at <http://www.gambit-project.org>. The Gambit project offers a unified suite of tools for working with finite games, available for Windows, OS X, and Linux. Gambit is fully open source, and is licensed under the GNU GPL.

The Gambit project was initiated in the early 1990s by Richard McKelvey and developed extensively in the mid-1990s via a National Science Foundation grant to McKelvey and Andrew McLennan. The impetus behind the conception of the project was to provide a framework for developing and manipulating representations of finite games in extensive and strategic form, and providing implementations of the emerging array of algorithms for computing Nash equilibria and other solution concepts for games to researchers, students, and practitioners.

For most users, the front door to Gambit is the graphical user interface. This tool offers the ability to construct games of small to medium size interactively, including the specification of players, move structures, information structures, and payoffs. The interface also allows for the interactive elimination of dominated strategies and actions and conversion of an extensive game to its reduced strategic form equivalent. Finally, it serves as a front end to the implementation of methods for computing

one or more Nash equilibria of the game, and for inspecting the implications of the resulting equilibria.

As of this writing in early 2010, Gambit is set for a significant overhaul. The current architecture of Gambit dates back to the mid-1990s, and is written primarily in C++. The architecture is modular, in that there is a self-contained library for representing games, entities within games such as players and outcomes, and concepts related to games such as mixed strategy profiles. Each algorithm for computing Nash equilibria in games is implemented separately, and wrapped as a stand-alone command-line tool for external scripting use. This current architecture makes Gambit modestly scriptable.

The state of the architecture also reflects the time at which it was written. In the mid-1990s, the term *open source* had yet to be coined officially, and libraries for scientific computing with licenses that encouraged reuse and redistribution were few. In addition, advances in computing power and algorithmic development have significantly expanded the scope of the type and size of game that can be analyzed using off-the-shelf hardware. The planned next iteration of Gambit will rebuild the architecture to be more friendly to modern scripting languages like Python, and to interface more cleanly with other tools relevant to the practice of game theory and scientific computing. Several of these changes are already available on an experimental basis in the current version. In addition, the current Gambit architecture represents all games by an explicit table of payoffs in memory. Decreases in the cost of computation, combined with improvements in algorithms, have now made more ambitious computational projects feasible. For example, one might explore numerically how equilibrium predictions depend on a parameter such as a player's attitude toward risk. When analyzing such a family of games, customized calculation of payoffs can offer significant advantages in representation space, and in running times. Future iterations of Gambit will offer facilities for supporting programming

game representations targeted at specific applications.

In the discussion in this article, the focus is on general-purpose tips on how to go about solving finite games. The organization will generally follow the structure of tools available in Gambit. Where other software resources exist that are not well-integrated into the Gambit framework, these will also be indicated.

FIELD GUIDE TO SOLVING GAMES

The first piece of advice for the practitioner interested in computing Nash equilibria in finite games is to have reasonable expectations. Computing a Nash equilibrium is algorithmically a challenging problem; see, for instance, Daskalakis *et al.* [1] and the references therein. In addition, there may be many Nash equilibria of a game (for example, McKelvey and McLennan [2]). Taken together, these imply that games that do not seem to be particularly large at first glance may be surprisingly difficult to solve. This suggests that a good strategy in approaching a game-theoretic model in a field application is to begin with a simple game structure.

Basic results in game theory show that the set of Nash equilibria is a subset of the strategy profiles that survive iterated elimination of strictly dominated strategies. Because running times are exponential in the number of strategies in the game, it is valuable to reduce the number of strategies being considered, whenever possible. Experience indicates checking for and eliminating dominated strategies pays off on average in improved overall running times. Gambit offers routines for automating this process.

The focus of this article is on solving games in strategic form. Historically, methods for computing equilibria in the strategic form of a game were well developed earlier. The last decade has seen significant algorithmic advancements in computing equilibria in extensive games (e.g., the sequence form of Koller *et al.* [3]). The representation of a game in extensive form is generally smaller than its strategic form counterpart, particularly when players have many information

sets in the extensive version of the game. When possible, consider using an algorithm that exploits the extensive structure of the game; the discussion in this article indicates where extensive form versions of equilibrium computation methods exist.

Complexity results such as the ones cited above are worst-case guarantees. Many games of interest in practice may have structure that can be exploited. The researcher may have prior beliefs that an equilibrium of a certain type might exist, or might be interested primarily in equilibria involving certain actions being played. Organizing the search for equilibria around this knowledge can help speed up the determination of answers to specific research questions, rather than waiting on an exhaustive analysis of the set of equilibria. There has been recent research in the direction of using heuristics to optimize the time to compute one Nash equilibrium.

For any game, an excellent place to begin the analysis is a simple enumeration of pure strategy equilibria. While the number of possible pure strategy profiles increases exponentially in the size of the game, checking whether a given strategy profile is a Nash equilibrium is in practice fast, and since it is a brute-force search, it is straightforward to parallelize. The rest of this section gives an overview of the available tools for computing Nash equilibria in general, including those involving mixed strategies. These methods divide into those that are applicable for games with two players and those that are applicable for games with arbitrary numbers of players. We now turn to each of those classes individually.

Two-Player Games

In a strategic form game with two players, the expected payoff to each strategy is a linear function of the probabilities that specify the other player's mixed strategy. This special structure admits algorithms for computing the set of Nash equilibria that exploit linearity and convexity. In a general two-player game, the set of Nash equilibria can be expressed as the union of convex sets of mixed strategy profiles. Furthermore, when the game is constant-sum, the set of

equilibria is itself one convex component. In both cases, the extreme points of the convex components are rational numbers when the payoffs specifying the game are rational. Gambit takes the perspective that all payoffs and probabilities in specifying a game are rational.

For two-player constant-sum games, the Nash equilibrium problem can be formulated as a standard linear programming problem (Dantzig [4]). Therefore, any linear programming solver can be brought to bear on computing equilibria. Gambit provides a self-contained linear programming solver in its distribution, wrapped in the command-line tool `gambit-lp`. In addition, there is experimental support for formulating the linear programming problem based on any game and passing the problem to external solvers. Since there are many strong options for both free and commercial linear programming solvers, external solvers may be the appropriate choice for many applications.

For two-player games that are not constant-sum, the equilibrium problem is a linear complementarity problem (Lemke and Howson [5]). There are two approaches to solving the linear complementarity problem. Lemke and Howson provided a linear path-following algorithm based on pivoting, reminiscent of the simplex method for linear programs. Experience has shown that this method, implemented as `gambit-lcp` in the Gambit package, is often efficient in finding an equilibrium of the game. However, the method of Lemke and Howson [5] is not guaranteed to find all equilibria; see Shapley [6] for an example. In addition, Savani and von Stengel [7] show how to construct pathological examples of games in which Lemke–Howson method takes exponentially long to arrive at an equilibrium.

Instead, to guarantee computation of all equilibria, the reverse pivoting algorithm of Avis and Fukuda [8] can be used. This method, implemented in Gambit as `gambit-ennummixed`, systematically enumerates all possible extreme points of the components of the set of equilibria, and determines whether or not they satisfy the Nash equilibrium conditions. An independent implementation of the algorithm, named `lrslib`, has been

made available under the GPL by Avis at <http://cgm.cs.mcgill.ca/~avis/C/lrs.html>. Avis' implementation includes a driver program for using the library to compute the set of Nash equilibria. For access to `lrslib` without the need to download and compile it, Rahul Savani has made a Web-based frontend to `lrslib` for computing equilibria [9]. As of this writing, there is an experimental wrapping of `lrslib` available in the `gambit-ennummixed` tool. The Gambit wrapping is often a version or two behind the latest `lrslib`, and therefore it does not always include the latest features; for instance, the most recent `lrslib` has support for executing the algorithm on multiple processors, which is not yet available via Gambit. A good self-contained resource describing the reverse pivoting algorithm and other recent developments in computing all equilibria of two-player strategic form games is Avis *et al.* [10], which includes a table summarizing how expected running times scale with the number of strategies in the game.

Because both the linear program and the reverse pivoting algorithm are of independent interest outside of their applications to the equilibrium computation problem, external solvers generally offer better support than the native Gambit implementations. One strength of the Gambit native implementation not always found in external solvers is the ability to carry out the entire calculation in exact rational arithmetic. Rational arithmetic is significantly slower than floating-point arithmetic, but pivoting calculations can be done exactly without rounding error, and exact representations of the extreme points of the equilibrium set can be found. While the set of Nash equilibria is a collection of isolated points in general, games that are of interest in practice often have components of equilibria that have positive dimension. Rational arithmetic handles the implied degeneracies exactly. In addition, Gambit offers support in the `gambit-ennummixed` tool for identifying the structure of the equilibrium components.

The existence of positive-dimensional components of equilibria in strategic games often arises because they are the reduced strategic form representations of extensive games.

The sequence form (Koller *et al.* [3]) allows for using linear programming and linear complementarity programming to compute equilibria while exploiting the extensive structure of the game.

General-Player Games

Games with more than two players lack the linear structure in payoffs present in games with only two players. Therefore, the problem of computing equilibria becomes more challenging. The development of effective methods for solving these games has occurred more recently than for two-player games, and remains an active area for further development in algorithms and implementations.

There is one existing method for finding all equilibria in a generic game with more than two players. This method relies on enumerating all the possible supports (sets of strategies played with positive probability) in the game. On a given support, the Nash equilibrium conditions can be expressed as a collection of equalities that are polynomial in the probabilities with which strategies are played, and a collection of inequalities that are also polynomial in the same probabilities. Therefore, for a given support, one can use an external solver to compute all solutions of the system of polynomial equalities, then check the candidate solutions, and keep only those that satisfy the inequalities and form valid probability distributions. A version of this approach was implemented in Gambit by Andrew McLennan ca. 1995, and the algorithm was formally defined and proved correct by Herings and Peeters [11].

Implementations of algorithms for finding the solutions to systems of polynomial equations are a recent development in the field of algebraic geometry. Two leading packages as of this writing are `PHCpack` by Jan Verschelde (at <http://www.math.uic.edu/~jan/download.html>) and `Bertini` by Bates *et al.* (at <http://www.nd.edu/~sommese/bertini/>). Turcoy [12] describes preliminary results of integrating `PHCpack` with Gambit. The experimental interface in Gambit to this method, currently distributed separately from the main Gambit package, also includes support for using `Bertini` as the backend solver.

Carrying out this algorithm is a formidable task, because the number of potential supports is exponential in the number of strategies a player can choose. If player i has k_i strategies, then there are $2^{k_i} - 1$ potential supports for that player, and $\prod_{i=1}^n (2^{k_i} - 1)$ possible supports overall. As Herings and Peeters [11] and Turocy [12] show, many of these supports can be ruled out as candidates by checks that require little processing time, so in practice the number of supports that have to be passed to the polynomial solver is significantly less than the theoretical maximum. In addition, because each support is an independent problem, the solution process is embarrassingly parallelizable; it is straightforward to utilize multiple independent threads of computation if multiple processors are available.

For some applications, exhaustive enumeration of all equilibria may not be necessary. Given that multiplicity of Nash equilibria is the rule rather than the exception, one may simply be interested in obtaining a sample Nash equilibrium. There exist several approaches for computing a single equilibrium of a game, all based on homotopy methods. The leading approaches in this area are surveyed by Herings and Peeters [13]. Loosely speaking, the basic idea behind these methods is, given a game of interest, to construct a sequence of games converging to that game. The sequence is constructed in such a way that the first game in the sequence is easy to solve, perhaps because it has an equilibrium in strictly dominant strategies. Then, because the set of Nash equilibria changes in a continuous way along the sequence of games, it is possible to “trace” an equilibrium back to the original game of interest.

In fact, the path-following algorithm of Lemke and Howson [5] can be interpreted as implementing just such a homotopy, for the special case of two-player games. As noted, the linear structure of expected payoffs in two-player games means that techniques for following the path of equilibria are exactly available. In games with more than two players, numerical procedures must be used to follow generally nonlinear paths

approximately. The Lemke-Howson method is generalized to n -player games by the global Newton method of Govindan and Wilson [14,15]. This has been implemented in the GameTracer package by Blum, Koller, and Shelton, and is available at <http://dags.stanford.edu/Games/gametracer.html>, distributed under the GNU GPL. Gambit incorporates the GameTracer implementation in the gambit-gnm and gambit-ipa command-line tools. The operation of the global Newton method is parameterized by specifying an initial perturbation vector, which specifies how the payoffs of the game are adjusted to find the initial starting game that is easy to solve. The resulting equilibria found by the procedure in general depend on the choice of this initial vector.

This last observation, that the equilibria found by an algorithm are a subset of the full set of equilibria, suggests the idea of selecting from the set of equilibria using an appropriately chosen algorithm. The multiplicity of Nash equilibria has always been a vexing problem in using the concept to predict behavior in a game, and the number of equilibria in a game tends to grow as the number of strategies and number of players are increased. This idea of using an algorithm to systematically select an equilibrium with desirable properties was made explicit in the form of the tracing procedure of Harsanyi [16] and Harsanyi and Selten [17]. The tracing procedure imagines players arriving at a Nash equilibrium of a game by beginning with initial beliefs about how other players will play the game. On the basis of this initial belief, through a process of introspection, the players adjust their intended actions based on their beliefs, while also adjusting their beliefs by realizing that other players are undergoing the same adjustment process. The tracing procedure results in paths that may not be differentiable at some points, thus complicating the numerical implementation of the tracing. Herings and Peeters [18] provide an implementation of an algorithm for following a differentiable path through the space of mixed strategy profiles to find a Nash equilibrium. While the path of strategy profiles followed is not exactly that of the tracing procedure, Herings and Peeters show that the

equilibrium the method computes coincides with the equilibrium selected by the tracing procedure. A FORTRAN implementation of the algorithm is available from the authors at <http://www.personeel.unimaas.nl/r.peeters>.

Another equilibrium computation and selection approach based on a process of behavioral adjustment uses the quantal response equilibrium of McKelvey and Palfrey [19]. In a quantal response equilibrium, the payoffs to each strategy are perturbed by an additive, independent stochastic shock. Under specified conditions, as the variance of this random shock goes to zero, the set of quantal response equilibria converge to a subset of the set of Nash equilibria. Turocy [20,21] demonstrated the method of using the results of McKelvey and Palfrey for the case when the shock follows the extreme value distribution to compute an arbitrarily good approximation of a Nash equilibrium by following a differentiable path of quantal response equilibria. This path can be interpreted as having the same conceptual structure of introspection and adjustment as the tracing procedure, but, since it uses a different dynamical system, it does not necessarily result in the same equilibrium as selected by the tracing procedure. This is implemented in Gambit as the gambit-logit command-line tool. This approach has an independent interest in laboratory experiments in game theory, where the quantal response equilibrium is used to organize the behavior of subjects playing games in controlled experiments (see, e.g., Goeree and Holt [22]).

Individually, none of the preceding three methods (global Newton, tracing procedure, quantal response equilibria) are guaranteed to be able to find all equilibria, even in games where the equilibria are isolated points. However, taken together, they are still useful elements of a toolkit for analyzing strategic games. Since they are approximation-based methods, the intermediate output they generate has a meaningful game-theoretic interpretation either as the exact equilibrium of an approximation to the original game or an approximate equilibrium to the exact original game. Also, often they find different equilibria when applied to a game.

Therefore, some questions, such as whether a game has a unique equilibrium, may in practice be answered most easily by applying a combination of these methods, rather than setting up an exhaustive search involving the solution to many systems of polynomial equations.

A final word of caution is in order when working with games with more than two players. Unlike two-player games, the set of Nash equilibria of a game with more than two players need not be expressible as the union of a finite number of convex sets. For a simple three-player example that illustrates this possibility, see Nau *et al.* [23]. Since, as noted before, many games that arise in practical applications may have positive-dimensional sets of equilibria, this possibility should be kept in mind. As of this writing, there is no tool that will reliably find and report the existence of such positive-dimensional sets. There is some support in the Bertini package for exploring positive-dimensional solution sets that could be brought to bear on the problem. In many cases, the existing PHCpack or Bertini implementations may fail to report any equilibria on a support, when in fact a positive-dimensional set of equilibria exists there. The global Newton, tracing procedure, and quantal response methods, if they find equilibria in such a component, will generally report only individual points from the component, and give no indication of the existence of equilibria nearby.

SUMMARY AND OUTLOOK

There exist a good set of tools for the practitioner to create, manipulate, and solve non-cooperative strategic form games. Advances in algorithm development, particularly in the solution of systems of polynomial equations, and the steady decrease in the price of computational power, have combined to significantly expand the size of games that can be analyzed on commodity hardware. For small applications, these tools may be able to provide essentially off-the-shelf, black-box solutions, particularly in the case of games with only two players.

In aiming to be a field guide for computing equilibria, this article omits many details of

the algorithms mentioned, as well as many other approaches to computing Nash equilibria. The survey of McKelvey and McLennan [24] is a good, if becoming dated, deeper introduction to the topic. A good resource for updates is the January 2010 issue of the journal *Economic Theory*, which was devoted entirely to the computation of Nash equilibria in finite games, and included the surveys of Avis *et al.* [10] and Herings and Peeters [13] mentioned earlier.

In view of the complexity results cited, no amount of algorithmic cleverness or hardware capacity will make the problem of computing the full set of equilibria of a strategic game fully tractable. Some current directions of research in this area involve heuristics which improve the average-time performance in, for instance, finding one equilibrium in particular classes of games; some progress in this direction has been reported by Porter *et al.* [25]. Another direction in the analysis of games is being advanced by behavioral game theory. If a Nash equilibrium is difficult to find, as the results on computational complexity of finding equilibria indicate, then it seems unsatisfactory to expect players to arrive at a Nash equilibrium in practice. Instead, predictions based on simpler heuristics such as level- k reasoning (Stahl [26]) or cognitive hierarchy models (Camerer *et al.* [27]) have been shown to organize how subjects in controlled laboratory experiments play games better than Nash equilibrium, and as such might be superior models for how players make strategic decisions in novel situations.

REFERENCES

1. Daskalakis C, Goldberg PW, Papadimitriou CH. The complexity of computing a Nash equilibrium. Proceedings of the 38th Annual ACM Symposium on Theory of Computing. Seattle (WA): 2006. pp. 71–78.
2. McKelvey RD, McLennan A. The maximal number of regular totally mixed Nash equilibria. *J Econ Theory* 1997;72(2):411–425.
3. Koller D, von Stengel B, Megiddo N. Efficient computation of equilibria for extensive two-person games. *Games Econ Behav* 1996;14(2):247–259.
4. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
5. Lemke CE, Howson JT Jr. Equilibrium points of bimatrix games. *J Soc Ind Appl Math* 1964;12(2):413–423.
6. Shapley LS. A note on the Lemke–Howson algorithm. In: Balinski M, editor. Mathematical programming study, 1: pivoting and extensions. Amsterdam: North-Holland Publishing co.; 1974. pp. 175–189.
7. Savani R, von Stengel B. Hard-to-solve bimatrix games. *Econometrica* 2006;74(2):397–429.
8. Avis D, Fukuda K. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete Comput Geom* 1992;8(3):295–313.
9. Savani R. Solve a bimatrix game. 2005. Available at <http://banach.lse.ac.uk/form.html>. Accessed in 2010.
10. Avis D, Rosenberg GD, Savani R, *et al.* Enumeration of Nash equilibria for two-player games. *Econ Theory* 2010;42(1):9–38.
11. Herings PJ-J, Peeters R. A globally convergent algorithm to compute all Nash equilibria of n -person games. *Ann Oper Res* 2005;137(1):349–368.
12. Turocy TL. Towards a black-box solver for finite games: finding all Nash equilibria with Gambit and PHCpack. In: Stillman ME, Takayama N, Verschelde J, editors. Proceedings of the IMA Software for Algebraic Geometry Workshop. New York: Springer; 2008.
13. Herings PJ-J, Peeters R. Homotopy methods to compute equilibria in game theory. *Econ Theory* 2010;42(1):119–156.
14. Govindan S, Wilson R. A global Newton method to compute Nash equilibria. *J Econ Theory* 2003;110(1):65–86.
15. Govindan S, Wilson R. Computing Nash equilibria by iterated polymatrix approximation. *J Econ Dyn Control* 2004;28(7):1229–1241.
16. Harsanyi J. The tracing procedure: a Bayesian approach to defining a solution for n -person noncooperative games. *Int J Game Theory* 1975;4(2):61–94.
17. Harsanyi J, Selten R. A general theory of equilibrium selection in games. Cambridge (MA): MIT Press; 1988.
18. Herings PJ-J, Peeters R. A differentiable homotopy to compute Nash equilibria of n -person games. *Econ Theory* 2001;18(1):159–185.

19. McKelvey RD, Palfrey TR. Quantal response equilibria for normal form games. *Games Econ Behav* 1995;10(1):6–38.
20. Turocy TL. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games Econ Behav* 2005;51(2):243–263.
21. Turocy TL. Using quantal response to compute Nash and sequential equilibria. *Econ Theory* 2010;42(1):255–269.
22. Goeree JK, Holt CA. Ten little treasures of game theory and ten intuitive contradictions. *Am Econ Rev* 2001;91(5):1402–1422.
23. Nau RF, Gomez Canovas S, Hansen P. On the geometry of Nash equilibria and correlated equilibria. *Int J Game Theory* 2004;32(4):443–453.
24. McKelvey RD, McLennan A. Computation of equilibria in finite games. In: Amman HM, Kendrick DA, Rust J, editors. Volume 1, Handbook of computational economics. Amsterdam: Elsevier; 1996.
25. Porter RW, Nudelman E, Shoham Y. Simple search methods for finding a Nash equilibrium. *Games Econ Behav* 2008;63(2):642–662.
26. Stahl DO. Evolution of smart- n players. *Games Econ Behav* 1993;5(4):604–617.
27. Camerer CF, Ho T-H, Chong J-K. A cognitive hierarchy model of games. *Q J Econ* 2004;119(3):861–898.

SOLVING INFLUENCE DIAGRAMS: EXACT ALGORITHMS

ROSS D. SHACHTER
Management Science and Engineering,
Stanford University, Stanford,
California

DEBARUN BHATTACHARJYA
Business Analytics and Math Sciences,
IBM T. J. Watson Research Center,
Yorktown Heights, New York

Influence diagrams [1] are popular graphical models in decision analysis for representing, solving, and analyzing a single decision maker's decision situation. Many of the important relationships among uncertainties, decisions, and values can be captured in the structure of these models, explicitly revealing irrelevance and the timing of observations. The phrase "solving influence diagrams" refers to computing the optimal strategy and the value at the optimal strategy, based on the norms of decision analysis. In this article, we highlight some of the most popular exact methods for solving influence diagrams.

Influence diagrams were initially conceived as tools primarily to be used in the formulation stage to assess the beliefs of a decision maker. However, research has shown that they are also efficient computational tools for the evaluation stage, and not merely graphical depictions for communicating ideas between the decision maker and the analyst [2,3]. There have been many developments in solution and analysis since then [4–11,12,13].

Methods for the solution of decision analysis problems have mainly been based on three graphical representations: decision trees, influence diagrams, and cluster trees. Decision trees are commonly taught in a first course in decision analysis. They are able to capture the asymmetries in real-world decision problems, and are easy to understand since they capture the informational order of variables in decision situations. On the other

hand, influence diagrams capture the conditional independencies in the model and are able to represent the model as conveniently assessed by the decision maker. Cluster trees are a solution representation derived from influence diagrams and will be presented in the section titled "Solving Influence Diagrams: Cluster Trees." In the next section we introduce the terminology and notation for influence diagrams used in this article.

INFLUENCE DIAGRAMS

Node and Arc Semantics

To describe a decision situation, a decision maker defines a number of *variables*. When a variable is *uncertain*, it represents a feature in the decision situation that the decision maker has no control over, and its possibilities are the *states* of the uncertainty. When a variable is a *decision*, the decision maker can choose to pursue any of its *alternatives*. In this article, we will deal with decision situations where there are a finite number of states or alternatives. We allow uncertainties to be *observed* to take specific states for certain and these observations are known as *evidence*.

An influence diagram is a directed graphical model for reasoning under uncertainty with nodes corresponding to the variables. Variables are denoted by upper case letters (X) and their states by lower case letters (x). A bold faced letter indicates a set of variables ($\mathbf{X}$). We will refer interchangeably to a node in the diagram and the corresponding variable. If there is an arc directed from node X to node Y , X is known as a *parent* of Y , and Y is a *child* of X . A node along with its parents forms the *family* for that node. The parents for node X are denoted $\mathbf{Pa}(X)$, and the family is $X\mathbf{Pa}(X)$. If there is a directed path from node X to node Y , we say that X is an *ancestor* of Y , and that Y is a *descendent* of X . Influence diagrams are *acyclic* graphs, meaning that there is no node X in the graph that is its own ancestor.

Influence diagrams have three kinds of nodes, decision, *uncertain (or chance)*, and *value* nodes, which together represent choices, beliefs, and preferences. Decision nodes are drawn as rectangles, uncertain nodes as ovals, and value nodes as rounded rectangles (or polygons). Decision nodes correspond to decision variables that are under the control of the decision maker and uncertain and value nodes correspond to variables that are not.

The semantics of an arc in an influence diagram depends on the type of node it is directed toward. Arcs directed into an uncertain or value node X are *conditional arcs*, and the distribution assessed for X is conditioned on its parents, $P\{X|\mathbf{Pa}(X)\}$. For value variables, we assume that this distribution can be represented as a deterministic function of the parent variables. *Informational arcs* directed toward decision nodes indicate those variables that will be observed before the decision is made. An influence diagram is *completely specified* if states (or alternatives) are assigned for each node, as well as conditional probability distributions for all uncertain nodes.

We will demonstrate some solution algorithms with the help of an influence diagram example from Jensen *et al.* [7], shown in Fig. 1. This influence diagram has four decision nodes $D1$ through $D4$, four value nodes $V1$ through $V4$, and twelve uncertain nodes A through L .

Maximizing Expected Utility

The criterion for decision making is represented by the value nodes. In this article, we assume that there may be multiple value nodes in the influence diagram, which are summed to obtain the total value. Influence diagrams with multiple additive or multiplicative value nodes were introduced and solved by Tatman and Shachter [14]. We assume that the value nodes characterize the value of the attributes in terms of a single numeraire, which we assume is dollars, and that there is a von Neumann–Morgenstern utility function $u(\cdot)$ applied to their total [15]. The utility function $u(\cdot)$ is typically assumed to be strictly increasing and continuously differentiable. The *certain equivalent* (CE) of an

uncertain $\mathbf{V}$ represents the certain payment that the decision maker finds indifferent to $\mathbf{V}$, and is given by $u^{-1}(E[u(\sum_{V \in \mathbf{V}} V)])$. The most common utility functions are *linear*, $u(v) = av + b$, and *exponential* functions of the sum of the values, $u(v) = -ae^{-v/\rho} + b$, where $a > 0$ and $\rho > 0$, both of which allow us to express the *value of information* [16,17] exactly in closed form. For simplicity, in this article we will assume a linear utility function. Note that the linear utility function corresponds to a risk-neutral decision maker, and in this situation the certain equivalent is equal to the maximum expected total value.

Thus the general problem we describe is to determine the decisions $\mathbf{D}$ that maximize the expected sum of values when the variables $\mathbf{E}$ have already been observed and the parents of each decision D are observed before it is made. In other words, the goal is to find the *optimal strategy*, given that evidence $\mathbf{E} = \mathbf{e}$ has been observed. Therefore, the solution of an influence diagram should reveal the optimal strategy and the certain equivalent of the decision situation, informing the decision maker about what to do in his/her decision situation, and what the decision situation is worth to him/her.

Other Standard Assumptions and Requirements

It is inconsistent to make observations that tell us anything about the decisions we are

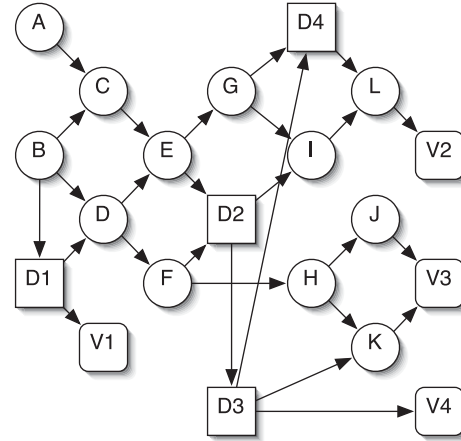


Figure 1. Influence diagram example from Jensen *et al.* [7].

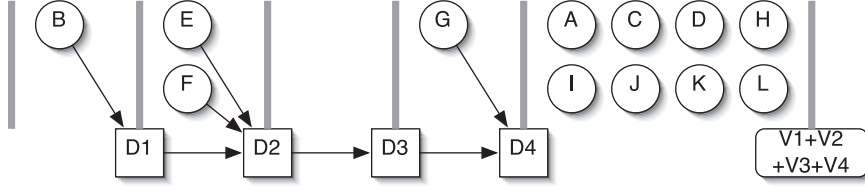


Figure 2. Decision windows for the influence diagram from Fig. 1.

yet to make, so we require that evidence variables **E** not be *responsive* in any way to **D** [18]. This is enforced by not allowing the nodes in **E** to be descendants of the nodes in **D**.

We also require that there be a total ordering of the decisions. In the case of the diagram in Fig. 1, the decisions are made in the order $D1, D2, D3$, and $D4$. We assume *no forgetting*, that all earlier observations and decisions are observed before any later decisions are made. No forgetting is a standard assumption for influence diagrams, and the no forgetting arcs are usually not drawn in the diagram when they can be inferred implicitly. For example, since there is a directed path from $D1$ to $D2$, there does not need to be an explicit arc from $D1$ to $D2$, but there should be directed arcs from $D2$ to $D3$ and $D3$ to $D4$. That way, it is clear that F is observed before $D3$.

No forgetting can also be understood using the notion of *decision windows* [19]. The total ordering of the decisions in the no forgetting ordering imposes a partial order on the other variables. Figure 2 shows decision windows for our influence diagram example, which indicate what variables are known at the time of each decision. The windows are one-way in the sense that future decisions can observe anything from earlier windows but nothing from the future. From Fig. 2, note that B is observed before $D1$ and therefore is known at the time all decisions are made. Any evidence would also be in this window. Uncertainties A, C, D, H, I, J, K, L are not known at the time of any decision; hence they are placed in the decision window after the last decision $D4$.

We assume that the influence diagram is a *single component*, that is, there is at least one path between any two nodes if we ignore

the direction of the arcs. Otherwise, we can solve each component independently. We also assume that there are no nodes which can be simply removed. For example, some non-responsive variables might become relevant if observed.

Requisite Observations

We make a quick note about *requisite* observations [5,20,21]. Evaluating an influence diagram entails finding the optimal alternative for every decision under the given observations available for the decision. However, based on the structure of the graph, only some of the observations are needed for a particular decision and there can be significant efficiency gains by recognizing which of the available observations are *requisite* for each of the decisions. The requisite observations for each decision can be computed, in reverse order, by considering which observations are relevant to the value descendants of the decision [20–22].

Therefore, a preprocessing step can be carried out on the influence diagram under consideration. In this preprocessing step, if an observation is not requisite for a decision, the arc from that observation into the decision is removed. Figure 3 presents the influence diagram example where preprocessing has been carried out and shows the requisite observations for every decision. Note that, although Fig. 3 has more arcs than Fig. 1, it actually reflects a significant reduction in graph complexity since the no forgetting arcs were hidden in Fig. 1. For instance, B is available at the time of all decisions, but it is not requisite for any decision other than $D1$, hence there are no arcs from B to $D2, D3$ or $D4$ in Fig. 3.

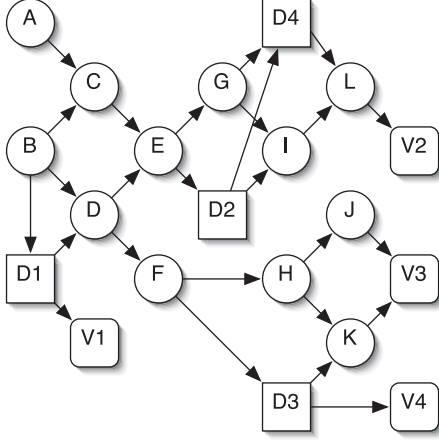


Figure 3. Requisite observations for the influence diagram from Fig. 1.

SOLVING INFLUENCE DIAGRAMS: DECISION TREES

One way to solve an influence diagram is to convert it into a decision tree [1] and then solve the tree. To be able to do this, the influence diagram must be a *decision tree network*, where all the ancestors of any decision node are parents of that node. Equivalently, if we call an arc between uncertain nodes *nonsequential* if the parent is in a later decision window than the child, a decision tree network is an influence diagram without nonsequential arcs.

Any influence diagram satisfying our assumptions can be transformed into a decision tree network through a sequence of *arc reversals* [19]. Arc reversal applies Bayes' rule to reverse the topological order of uncertain nodes [2,3]. An arc from uncertain node X to uncertain node Y with parent sets $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ can be reversed, as shown in Fig. 5a, unless there is another directed path from X to Y ; if there were, then reversing the arc would create a directed cycle. When the arc (X, Y) can be reversed, the joint distribution of X, Y conditioned on $\mathbf{A}, \mathbf{B}, \mathbf{C}$ can be factored two ways:

$$\begin{aligned} P\{X, Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\} &= P\{X | \mathbf{A}, \mathbf{B}\} P\{Y | X, \mathbf{B}, \mathbf{C}\} \\ &= P\{X | Y, \mathbf{A}, \mathbf{B}, \mathbf{C}\} P\{Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\}. \end{aligned}$$

The new distribution for Y given $\mathbf{A}, \mathbf{B}, \mathbf{C}$ is obtained by marginalizing X from the joint

$$\begin{aligned} P\{Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\} &= \sum_x P\{X = x, Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\} \\ &= \sum_x P\{X = x | \mathbf{A}, \mathbf{B}\} P\{Y | X = x, \mathbf{B}, \mathbf{C}\}, \end{aligned}$$

and it is used to obtain the new distribution for X given $Y, \mathbf{A}, \mathbf{B}, \mathbf{C}$:

$$\begin{aligned} P\{X | Y, \mathbf{A}, \mathbf{B}, \mathbf{C}\} &= \frac{P\{X, Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\}}{P\{Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\}} \\ &= \frac{P\{X | \mathbf{A}, \mathbf{B}\} P\{Y | X, \mathbf{B}, \mathbf{C}\}}{P\{Y | \mathbf{A}, \mathbf{B}, \mathbf{C}\}}. \end{aligned}$$

The influence diagram shown in Fig. 1 is not a decision tree network, because the arcs (C, E) , (D, E) , and (D, F) are nonsequential. We can obtain a decision tree network by reversing those arcs. Note that in the process nonsequential arcs would be created from A to E and F that would also need to be reversed. We simplify matters by reversing (A, C) before reversing the nonsequential arcs. Figure 4 presents the decision tree network for the example, after the four arc reversals. Note that in Fig. 1, D is an ancestor but not a parent of decisions nodes $D2$,

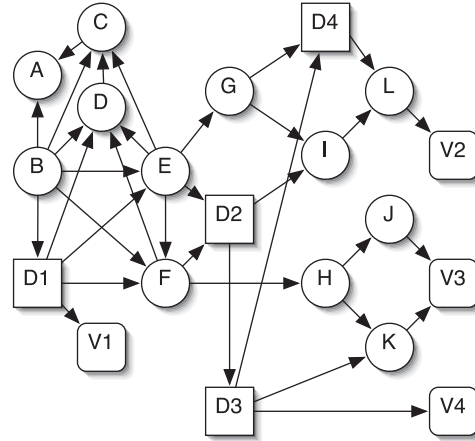


Figure 4. Decision tree network for the influence diagram from Fig. 1.

$D3$, and $D4$, while in Fig. 4, it is no longer an ancestor of any decision nodes.

Decision trees are easy to understand but they become intractable for decision situations involving a large number of variables since they are exponential in the number of nodes. Decision trees do have some advantage in that they may be able to better represent asymmetric decision situations such as the used car buyer problem [23]. We briefly discuss asymmetry later in the article. Larger decision situations may be more amenable for representation and solution through other means. In the following sections, we review techniques that are able to exploit the conditional independence of the influence diagram, and therefore reduce the complexity of the solution.

SOLVING INFLUENCE DIAGRAMS: MODIFICATIONS TO THE DIAGRAM

A *variable elimination algorithm* directly evaluates the original influence diagram through a sequence of node elimination steps [2,3,14,24]. The algorithm iteratively removes nodes from the diagram until all that remains is a value node and any evidence nodes, representing the optimal value of the influence diagram and the observations already made. The optimal strategy is determined in terms of an optimal policy for each decision before that decision is eliminated.

The algorithm recognizes a variety of conditions under which nodes can be eliminated, and at least one node can be removed in any proper influence diagram until all that remains are value and evidence nodes. For example, *barren nodes* are decision or unobserved uncertain nodes that have no children. These variables have no effect upon the value and can be simply removed from the diagram. These might be present in the original diagram or created in the process of variable elimination.

There are four other graphical transformations needed: arc reversal between uncertain nodes discussed in the previous section, and a removal operation for each type of node: decision, uncertain, and value.

When an uncertain node has only one child, and that child is a value node, *uncertain node removal* eliminates the node from diagram by taking expectation, and the value node inherits its parents as shown in Fig. 5b,

$$\begin{aligned} E[V|\mathbf{A}, \mathbf{B}, \mathbf{C}] \\ = \sum_x P[X = x|\mathbf{A}, \mathbf{B}] E[V|X = x, \mathbf{B}, \mathbf{C}]. \end{aligned}$$

When a decision node has only one child, that child is a value node, and all other parents of the value node are also parents of the decision node, *optimal policy determination* replaces the decision node with a (deterministic) uncertain one, representing the optimal policy for each possible combination of the other parents of the value node as shown in Fig. 5c,

$$E[V|\mathbf{B} = \mathbf{b}] = \max_d E[V|D = d, \mathbf{B} = \mathbf{b}],$$

for all possible $\mathbf{b}$. That policy node can then be removed, but the optimal policy should be saved.

When there are multiple value nodes, *value node removal* replaces two value nodes by one value node equal to their sum with the union of their parents, as shown in Fig. 5d,

$$E[V_1 + V_2|\mathbf{A}, \mathbf{B}, \mathbf{C}] = E[V_1|\mathbf{A}, \mathbf{B}] + E[V_2|\mathbf{B}, \mathbf{C}].$$

To exploit the decomposition of the value, this operation should be delayed as long as possible, ideally until the parents of one of the value nodes are all parents of the other, or when both value nodes are descendants of the latest decision.

Given these operations, an influence diagram with no directed cycles and ordered decisions with no forgetting can be solved by the following algorithm [3,14]. First, if any of the value nodes has a child, turn it into an uncertain node and give it a value child equal to itself.

Repeat, until all that remains is a single value node and the evidence nodes:

1. If there is a barren node, remove it.
2. Otherwise, if the latest decision node has exactly one child, a value node,

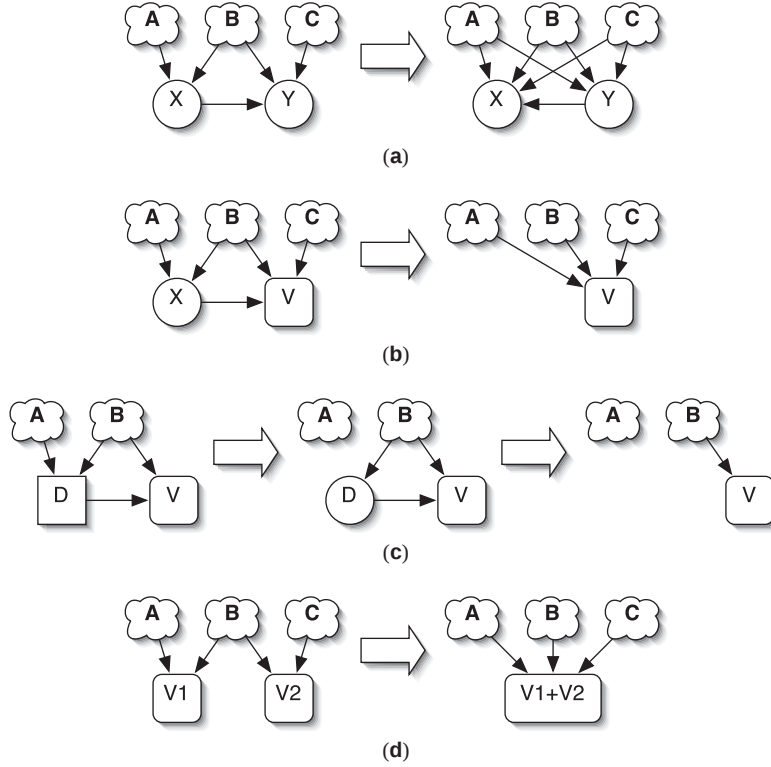


Figure 5. The four graphical transformations to the influence diagram.

and all of the other parents of the value are observed before the decision, then determine its optimal policy and remove it.

3. Otherwise, if there is an uncertain node that is not observed before the latest decision and has [at most] one value child, then remove it after reversing any arcs to its nonvalue children in graph order.
4. Otherwise, a value node needs to be removed, preferably a descendant of the latest decision or when its parents are all parents of another value node.

Let us demonstrate the algorithm with the help of value reduction steps shown in Fig. 6. Figure 3 presents the original influence diagram with the requisite observations

for all decisions. Figure 6a shows the diagram after the removal of five unobserved uncertain nodes, L , I , J , K , and H . The optimal policy for $D4$ can now be determined, and Fig. 6b shows the diagram after the decision node $D4$ has been removed and the value nodes $V3$ and $V4$ have been summed so that the policy for $D3$ can be determined. Uncertain node G and decision node $D3$ have been removed in Fig. 6c. Decision node $D2$ and uncertain node F have been removed in Fig. 6d. Uncertain nodes E , A , and C have been removed in Fig. 6e. Value nodes $V2$ and $V3 + V4$ have been summed in Fig. 6f so that uncertain node D can be removed in Fig. 6g. The two value nodes are summed in Fig. 6h so that decision node $D1$ can be removed in Fig. 6i. Finally, uncertain node B is removed in Fig. 6j, so that only the total value remains.

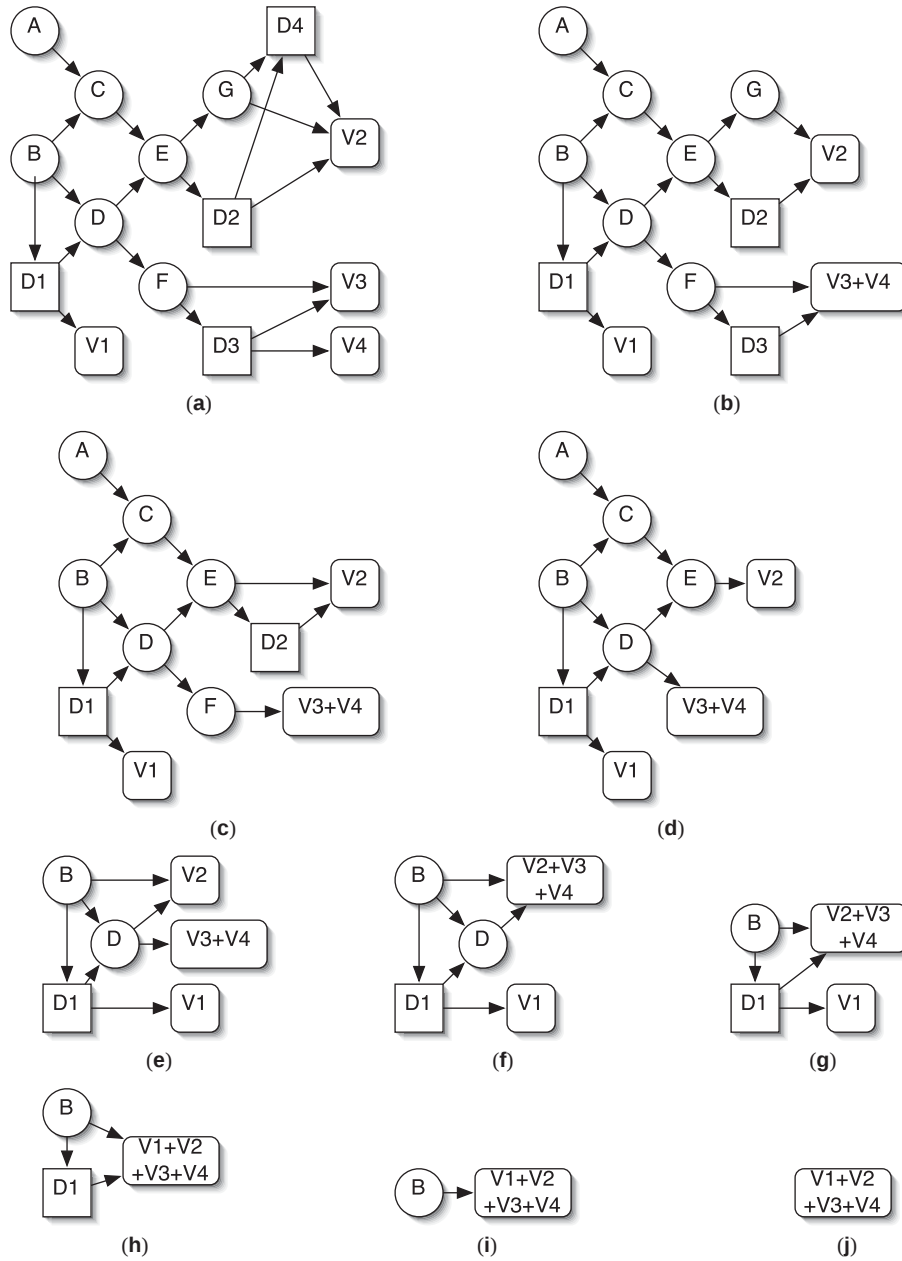


Figure 6. Node reduction steps for the influence diagram from Fig. 1.

SOLVING INFLUENCE DIAGRAMS: CLUSTER TREES

Much of the popularity of belief networks arises from the improvement in computational power through the development of

efficient algorithms, many of which use graphical structures called *chordal graphs* and *cluster trees*, abstracted from the original diagram [25–27,13]. Several modified versions of these algorithms have been able to tackle the problem of solving influence

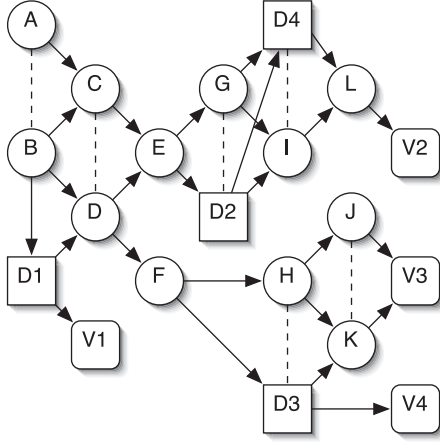


Figure 7. Requisite observations and moral edges for the influence diagram from Fig. 1.

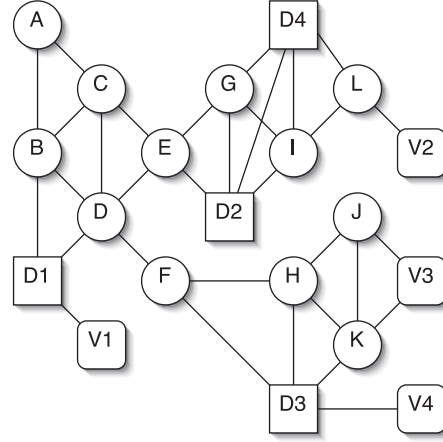


Figure 8. Undirected graph for the moralized diagram from Fig. 7.

diagrams [6–8,21]. Much has been written about the construction of chordal graphs and cluster trees, and we will present a simplified version, enough to show how they can be used to solve influence diagrams.

We start with a variable elimination ordering similar to the one we presented in the previous section. For the diagrams in Fig. 6, the order of uncertain and decision node removal was $L, I, J, K, H, D4, D3, G, D2, F, E, A, C, D, D1$, and B . We treat all of the value nodes as if they were eliminated before any other variables.

The first step is to ensure that each node's family will be together to represent its conditional distribution. We add *moralizing edges*, undirected edges between the parents of a node, if there is not already an arc between them. Starting with the requisite observations in Fig. 3, we add moralizing edges to obtain the diagram shown in Fig. 7. For example, A and B are parents of C with no arc between them so an undirected moralizing edge is added between them.

The next step is to create an undirected version of the graph shown in Fig. 7 as shown in Fig. 8. We would ideally like this graph to be *chordal* or *triangulated*, where any undirected circuit of length four or more has a chord, as this one does, but we will add these extra edges in the next step, if necessary.

Now we construct a directed version again, using the variable elimination order. Arcs are directed toward the earlier variables to be eliminated from the later ones, as shown in Fig. 9. This graph must also satisfy an important condition to be a *directed chordal graph*: there must be a directed arc between any two nodes with a common child [13]. Unlike the moralizing edges, these new arcs can themselves create the need for additional arcs. When we add them, we use the elimination order to determine the direction of the arcs. In this example, because the graph in Fig. 8 is chordal and this elimination order is “perfect,” we have no new arcs to add. However, if C were eliminated before A , then there would be an arc from A to C and we would have needed to add an arc from D to A since both of them would be parents of C . The best elimination order is one that requires no extra arcs in this diagram.

We are now prepared to construct a *rooted cluster tree* [6] (also called a *strong junction tree* [7]), a directed tree with each node corresponding to a *cluster* or set of variables, and satisfying four important properties: (i) there is one root cluster with no child clusters, and all other clusters have exactly one child; (ii) there are arcs in the directed chordal graph between all of the variables in each cluster; (iii) if the same variable appears in two clusters, it appears in every cluster in between them in the tree; and (iv)

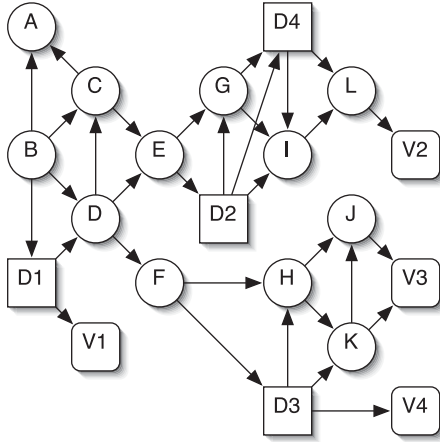


Figure 9. Directed chordal graph for the influence diagram from Fig. 1.

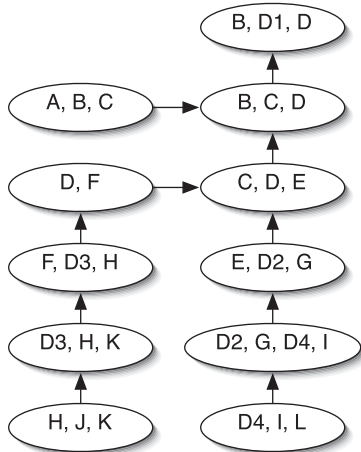


Figure 10. Rooted cluster tree for the influence diagram from Fig. 1.

if one cluster follows another in the tree, then the variables that are only in the parent cluster will be eliminated before any of the variables in the child cluster.

There can be many different rooted cluster trees for the influence diagram shown in Fig. 1, depending on the elimination order, and even with the same elimination order, but here is an algorithm to construct the rooted cluster tree shown in Fig. 10, from a directed chordal graph such as Fig. 9.

1. Select the variable having no parents, X , mark all other nonvalue variables as *unselected*, set the *current* cluster to be $\{X\}$, and make it the root cluster.
2. While there is at least one unselected variable X , with all of its parents comprising the current cluster, select the latest such X in the elimination order and add it to the current cluster.
3. If there are no more unselected variables, we are done. Otherwise, select the unselected variable X latest in the elimination order, creating a new current cluster with its family, $X\text{Pa}(X)$. Make this current cluster parent to one of the clusters containing all of X 's parents; we suggest that the child cluster should be chosen as close to the root cluster as possible. Now go to step 2.

In our example, variables B , $D1$, and D would be selected to form the root cluster. Variable C is selected next and the cluster $\{B, C, D\}$ is created as a parent of cluster $\{B, D1, D\}$. Variable A is selected next, and the cluster $\{A, B, C\}$ is created as a parent of $\{B, C, D\}$. This proceeds until all of the variables, except the value variables, have been included in clusters, and those clusters have been added to the tree. Note that for our purposes, we do not need to have the value variables in clusters.

The first step in solving the influence diagram with the cluster tree is to associate each uncertain variable X with a cluster in the cluster tree that contains the family of X , $X\text{Pa}(X)$. (For value variables, the family does not need to include the value variable itself, just its parents.) The addition of moralizing edges guarantees that there will be at least one such cluster for each uncertain variable. In the cluster tree shown in Fig. 10, the variable B could be associated with any of the three clusters containing B , but C must be associated with A, B, C because that is the only cluster containing its family in Fig. 1.

For each cluster, we compute a *utility potential* as the product of all of the conditional probabilities (and evidence likelihoods) of the uncertain variables associated with that cluster. For example, if B is associated with cluster B, C, D , then its

utility potential would be $P\{B\}$; otherwise, if no variables were associated with the cluster its utility potential would be 1. Cluster H, J, K is associated with variables, J, K , and $V3$, so its utility potential would be $P\{J|H, K\}P\{K|H\}E[V3|J, K]$.

For each cluster, we also compute a *probability potential*, similar to the utility potential, except that it does not include expectation factors for value variables. For example, the probability potential for B, C, D would be the same as its utility potential, but the probability potential for H, J, K would be $P\{J|H, K\}P\{K|H\}$.

We then visit each cluster in graph order. We multiply the cluster potentials by any corresponding potentials we received from parent clusters. We then eliminate, in order, those variables not present in its child cluster. For uncertain variables, we sum over all states of the variable in both potentials. For decision variables, we need to maximize over the utility potential for each state of the requisite variables, all of which will be in that cluster, and select the corresponding element from the probability potential. (For value variables, there is no need to do anything.) The resulting potentials should be a function of variables in the child cluster and they should be passed along to it.

At the end, in the root cluster, we have a single number left in each potential. The probability potential equals the probability of the evidence, and the utility potential divided by the probability potential equals the maximal expected value. The optimal strategy is represented by the maximization choices we made for the decision variables.

For example, in cluster H, J, K we sum the potentials over the states of J and pass along the new potentials, as a function of H, K to cluster $D3, H, K$. We multiply that cluster's own potentials, $P\{K|D3, H\}E[V4|D3]$, by the passed potential, sum over the states of K , and pass the resulting potential as a function of $D3, H$ to cluster $F, D3, H$. We multiply that cluster's own potential, $P\{H|F\}$, by the passed potential, sum over the states of H , and maximize over $D3$ given F , saving the optimal choices. We pass the resulting potential as a function of F to cluster D, F . We continue processing each cluster this way until the root

cluster's potentials are reduced to a single number each.

The cluster tree method uses the conditional independence in the diagram and organizes the computations so that summation and maximization operations occur in the appropriate order. We provided a brief overview of solving influence diagrams with cluster trees, leaving out many of the details. Shachter and Peot [6], Jensen *et al.* [7], Shachter [21], and Jensen and Nielsen [28] are some references for further reading on cluster tree construction and evaluation with message passing.

OTHER RELATED TOPICS

The literature on influence diagrams is vast and varied; here we provide a quick overview of other related topics.

Asymmetry

The methods we have described for solving influence diagrams exploit the *global structure* of the influence diagram, that is, the topology of the graph, and in particular through conditional independence and separable values. The complexity of the influence diagram at the level of global structure is studied in terms of the number of nodes and the *treewidth*, which is a measure of the connectivity of the network. On the other hand, *local structure* refers to the specific numbers used for building the graphical model. Here we discuss the issue of local structure in influence diagrams.

Real-world decision situations often exhibit a significant amount of *asymmetry*, that is, decision situations where some possible combinations of variables have no meaning. In an influence diagram representation, this typically means that the full rectangular table does not need to be evaluated. Asymmetry in the decision situation is represented through the numbers in the model, and mainly in the form of *determinism*, or the presence of zeros in conditional probability distributions, and *context specific independence* [29], which refers to independence between variables given a particular instantiation of some

other variables. In most real-world decision situations, there is a tremendous opportunity to also exploit the specific numbers in the model so that evaluation and analysis can be made even more efficient. Previous research has shown the potential benefits from tapping local structure for improving evaluation efficiency [30,31].

Influence diagrams face some difficulty in representing asymmetric decision situations. Some of the shortcomings of influence diagrams for representing asymmetric situations are documented in Smith *et al.* [31], Bielza and Shenoy [32], and Shenoy [33]. Several researchers have worked on trying to extend and augment the influence diagram representation to capture asymmetry directly, see for instance Covaliu and Oliver [34], Shenoy [35], and Jensen *et al.* [36]. Another direction of research has been to focus on identifying ways to exploit the asymmetry while solving the influence diagram. This has been actively pursued in the belief network literature, with the help of graphical representations such as arithmetic circuits [37,38]. Decision circuits have been developed to perform efficient evaluation of influence diagrams [9,13], building on the advances in arithmetic circuits for belief network inference. The benefit of the decision circuit framework is that compact circuits are compiled in an offline phase so that subsequent on-line operations for evaluation and analysis can be more efficient. This can be particularly useful in situations for real-time decision making and in decision systems [39], that is, for decision situations under uncertainty that recur frequently and involve considerable automation.

Other Related Graphical Models

In the previous section, we mentioned a few graphical models for specifically dealing with the issue of asymmetry. There are various other graphical models closely related to the influence diagram literature.

There are other graphical models for representing and solving decision problems, such as game trees [40] and valuation networks [41]. The fusion algorithm for solving valuation networks is closely related to the cluster tree methods described here.

There have also been several extensions to influence diagrams that relax some of the assumptions, such as limited memory influence diagrams (LIMIDs) [42], unconstrained influence diagrams [43], and so on. Many of these graphical models are harder to solve than standard influence diagrams and, like the graphical models for representing asymmetry, are rather complex. There is also a considerable amount of literature on influence diagrams with continuous uncertain variables, starting with Shachter and Kenley [44]. Another direction of research has been on extending influence diagrams to problems with multiple decision makers [42,45,46].

Sensitivity Analysis and Value of Information

The phrase *sensitivity analysis* refers, in general, to understanding how the output for a system varies as a result of changes in the system's inputs. For influence diagrams, one may be concerned about how the optimal solution and the value change with respect to a change in the parameters, that is, the probabilities and the utilities, or a change in the informational assumptions of the problem. Sensitivity analysis has been an essential aspect of decision analysis throughout the field's development. Sensitivity analysis aids in identifying the model's critical elements, forming the basis for iterative refinement of the model, and can also be used after the analysis for defending a particular strategy to the decision maker [47].

Sensitivity analysis is an essential technique to support influence diagram modeling and it provides valuable insight about the critical assessments. Sensitivity analysis for influence diagrams has been studied in Nielsen and Jensen [10] and Bhattacharya and Shachter [11,12]. Decision circuits are particularly suitable for performing sensitivity analysis on influence diagrams since partial derivatives are easy to compute with compiled circuits, and therefore analysis on the model can be performed conveniently within this framework.

One of the most powerful sensitivity analysis techniques of decision analysis is the computation of *value of information* (or clairvoyance), which is the difference in value obtained by changing the decisions by

which some of the uncertainties are observed [16,17]. The value of information for a particular uncertainty specifies the maximum that the decision maker should be willing to pay to observe the uncertainty before making a particular decision. Value of information can be computed on the cluster tree used to solve the original problem [21] or in other ways [11,48].

REFERENCES

- Howard R, Matheson J. Influence diagrams. In: Howard R, Matheson J, editors. Volume II, The principles and applications of decision analysis. Menlo Park (CA): Strategic Decisions Group; 1984. pp. 721–762.
- Olmsted S. On solving influence diagrams [PhD dissertation]. Stanford University, CA, USA; 1983.
- Shachter R. Evaluating influence diagrams. *Oper Res* 1986;34:871–882.
- Cooper G. A method for using belief networks as influence diagrams. In: Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence; Minneapolis (MN). 1988. pp. 55–63.
- Shachter R. Probabilistic inference and influence diagrams. *Oper Res* 1988;36:589–605.
- Shachter R, Peot M. Decision making using probabilistic inference methods. In: Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence; Stanford (CA). 1992. pp. 276–283.
- Jensen F, Jensen FV, Dittmer S. From influence diagrams to junction trees. In: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence; Seattle (WA). 1994. pp. 367–373.
- Madsen A, Jensen FV. Lazy propagation in junction trees. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence; Madison (WI). 1998. pp. 362–369.
- Bhattacharjya D, Shachter R. Evaluating influence diagrams with decision circuits. In: Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence; Vancouver, Canada. 2007. pp. 9–16.
- Nielsen T, Jensen FV. Sensitivity analysis in influence diagrams. *IEEE Trans Syst Man Cybern Part A Syst Hum* 2003;33(1):223–234.
- Bhattacharjya D, Shachter R. Sensitivity analysis in decision circuits. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence; Helsinki, Finland. 2008. pp. 34–42.
- Bhattacharjya D, Shachter R. Three new sensitivity analysis methods for influence diagrams. Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence; Catalina Island (CA). 2010. pp. 56–64.
- Shachter R, Bhattacharjya D. Dynamic programming in influence diagrams with decision circuits. Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence; Catalina Island (CA). 2010. pp. 509–516.
- Tatman J, Shachter R. Dynamic programming and influence diagrams. *IEEE Trans Syst Man Cybern* 1990;20(2):365–379.
- von Neumann J, Morgenstern O. Theory of games and economic behavior. 2nd ed. Princeton (NJ): Princeton University Press; 1947.
- Howard R. Information value theory. *IEEE Trans Syst Sci Cybern* 1966;2:22–26.
- Raiffa H. Decision analysis: introductory lectures on choices under uncertainty. Reading (MA): Addison-Wesley; 1968.
- Heckerman D, Shachter R. A definition and graphical representation for causality. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence; Montreal, Canada. 1995. pp. 262–273.
- Shachter R. An ordered examination of influence diagrams. *Networks* 1990;20:535–563.
- Nielsen T, Jensen FV. Well-defined decision scenarios. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence; Stockholm, Sweden. 1999. pp. 502–511.
- Shachter R. Efficient value of information computation. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence; Stockholm, Sweden. 1999. pp. 594–601.
- Shachter R. Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence; Madison (WI). 1998. pp. 480–487.
- Howard R. The used car buyer. In: Howard R, Matheson J, editors. Volume II, Readings on the principles and applications of decision analysis. Menlo Park (CA): Strategic Decisions Group; 1962. pp. 689–718.
- Shachter R. Model building with belief networks and influence diagrams. In: Edwards W, Miles R, von Winterfeldt D, editors. Advances

- in decision analysis: from foundations to applications. New York: Cambridge University Press; 2007. pp. 177–201.
25. Lauritzen S, Spiegelhalter D. Local computations with probabilities on graphical structures and their application to expert systems. *J R Stat Soc Ser B* 1988;50(2):157–224.
 26. Jensen FV, Lauritzen S, Olesen KG. Bayesian updating in causal probabilistic networks by local computation. *Comput Stat Q* 1990;4: 269–282.
 27. Shenoy P, Shafer G. Axioms for probability and belief-function propagation. In: Shachter R, Levitt T, Lemmer J, *et al.*, editors. *Uncertainty in artificial intelligence 4*. Amsterdam: North-Holland; 1990. pp. 169–198.
 28. Jensen FV, Nielsen TD. *Bayesian networks and decision graphs*. 2nd ed. New York: Springer; 2007.
 29. Boutilier C, Friedman F, Goldszmidt M, *et al.* Context-specific independence in Bayesian networks. In: *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*; Portland (OR): 1996. pp. 115–123.
 30. Gomez M, Cano A. Applying numerical trees to evaluate asymmetric decision problems. In: Nielsen T, Zhang N, editors. *Symbolic and quantitative approaches to reasoning with uncertainty*, *Lecture Notes in Artificial Intelligence* 2711. Berlin/Heidelberg: Springer; 2003. pp. 196–207.
 31. Smith J, Holtzman S, Matheson J. Structuring conditional relationships in influence diagrams. *Oper Res* 1993;41(2):280–297.
 32. Bielza C, Shenoy P. A comparison of graphical techniques for asymmetric decision problems. *Manag Sci* 1999;45(11):1552–1569.
 33. Shenoy P. A comparison of graphical techniques for decision analysis. *Eur J Oper Res* 1994;78(1):1–21.
 34. Covaliu Z, Oliver R. Representation and solution of decision problems using sequential decision diagrams. *Manag Sci* 1995;41(12): 1860–1881.
 35. Shenoy P. Valuation network representation and solution of asymmetric decision problems. *Eur J Oper Res* 2000;121(3):579–608.
 36. Jensen FV, Nielsen TD, Shenoy P. Sequential influence diagrams: a unified asymmetry framework. *Int J Approx Reason* 2006;42(1):101–118.
 37. Darwiche A. A differential approach to inference in Bayesian networks. In: *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. 2000. pp. 123–132.
 38. Darwiche A. A differential approach to inference in Bayesian networks. *J ACM* 2003;50(3):280–305.
 39. Holtzman S. *Intelligent decision systems*. Reading (MA): Addison-Wesley; 1989.
 40. Shenoy P. Game trees for decision analysis. *Theory Decis* 1998;44(2):149–171.
 41. Shenoy P. Valuation based systems for Bayesian decision analysis. *Oper Res* 1992;40: 463–484.
 42. Lauritzen S, Nilsson D. Representing and solving decision problems with limited information. *Manag Sci* 2004;47:1235–1251.
 43. Jensen FV, Vomlelova M. Unconstrained influence diagrams. In: *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*; Edmonton, Canada. 2002. pp. 234–241.
 44. Shachter R, Kenley C. Gaussian influence diagrams. *Manag Sci* 1989;35:589–605.
 45. Koller D, Milch B. Multi-agent influence diagrams for representing and solving games. *Games Econ Behav* 2003;45:181–1221.
 46. Detwarasiti A, Shachter R. Influence diagrams for team decision analysis. *Decis Anal* 2005;2(4):207–228.
 47. Howard R. The evolution of decision analysis. In: Howard R, Matheson J, editors. *The principles and applications of decision analysis*. Menlo Park (CA): Strategic Decisions Group; 1983.
 48. Matheson J. Using influence diagrams to value information and control. In: Oliver RM, Smith JQ, editors. *Influence diagrams, belief networks and decision analysis*. New York: John Wiley & Sons, Inc.; 1990. pp. 25–63.

SOLVING STOCHASTIC PROGRAMS

GÜZİN BAYRAKSAN
University of Arizona,
Department of Systems and
Industrial Engineering,
Tucson, Arizona

Stochastic programming extends deterministic optimization by including random parameters and probabilistic statements in model formulation. It has been successfully applied for decision making under uncertainty in diverse fields such as finance, energy, manufacturing, transportation, supply chain, and environmental management among others [1]. A general stochastic program can be formulated as follows:

$$\min_{x \in X} \left\{ \mathbb{E}f_0(x, \tilde{\xi}) : \mathbb{E}f_k(x, \tilde{\xi}) \leq 0, k = 1, 2, \dots, K \right\}, \quad (\text{SP})$$

where f_k , $k = 0, 1, \dots, K$ are extended real-valued functions with inputs being the decision vector x and a realization of the random vector $\tilde{\xi}$. $X \subset \mathbb{R}^{d_x}$ represents the set of deterministic constraints that x must satisfy, and $\Xi \subset \mathbb{R}^{d_\xi}$ denotes the support of $\tilde{\xi}$, where d_x and d_ξ are the dimensions of the vectors x and $\tilde{\xi}$, respectively. We assume that $\tilde{\xi}$ has distribution, $\mathbb{P}$, that is independent of x , and the expectations in (SP), taken with respect to the distribution of $\tilde{\xi}$, are well-defined and finite for all $x \in X$.

A wide variety of problems can be cast as (SP) depending on K , X , and f_k , $k = 0, 1, 2, \dots, K$. For example, in a two-stage stochastic linear program with recourse, $K = 0$, $X = \{Ax = b, x \geq 0\}$, and $f_0(x, \tilde{\xi}) = cx + h(x, \tilde{\xi})$, where $h(x, \tilde{\xi})$ is the optimal value of the linear program

$$\begin{aligned} h(x, \tilde{\xi}) &= \min_y \tilde{q}y \\ \text{s.t.} \quad &\tilde{W}y = \tilde{r} - \tilde{T}x, \quad y \geq 0. \end{aligned}$$

Here, $\tilde{\xi}$ is a random vector that is comprised of random elements of $\tilde{q}$, $\tilde{W}$, $\tilde{R}$, and

$\tilde{T}$. Stochastic programs with recourse model time-staged decisions under uncertainty that are common in practice. For instance, decisions on where to locate factories and warehouses must be made without a full knowledge of the uncertainties like demands. When uncertainties become known, recourse decisions such as minimum cost transportation of goods to demand points, and overtime and subcontracting to meet excess demand are made. These recourse decisions are captured in the second stage of the above problem. In contrast, in a stochastic program with a single probabilistic constraint, $K = 1$, $f_0(x, \tilde{\xi}) = cx$, and

$$f_1(x, \tilde{\xi}) = \mathbb{I}(\tilde{A}'x \geq \tilde{b}') - \alpha,$$

where $\mathbb{I}(\cdot)$ denotes the indicator function that takes value 1 if its argument is true and zero otherwise, and $\alpha \in (0, 1)$ is a desired probability level. In this case, $\tilde{\xi}$ is comprised of random elements of $\tilde{A}'$ and $\tilde{b}'$. Here, a single stage decision is made by taking into account the probability of an “event” of interest. For instance, this event can be “not meeting demand” and its probability can be limited to a desired level.

Stochastic programs constitute one of the hardest classes of optimization problems to solve. A major challenge in algorithmic design for stochastic programs stems from the need to calculate multidimensional expectations that appear in (SP). When d_ξ , the dimension of $\tilde{\xi}$, is moderate-to-high and $\tilde{\xi}$ is continuous or has a large number of realizations, calculating $\mathbb{E}f_k(x, \tilde{\xi})$, $k = 0, 1, \dots, K$ might not be possible even for a fixed x . Algorithms typically require these function evaluations, along with other more complicated function evaluations such as subgradients of these expectations, in an iterative fashion for many different x . Even if the expectations can be calculated for a fixed x , optimization over the set of feasible solutions can be quite challenging. For example, two-stage stochastic integer programs with integer second stage can

result in a discontinuous and nonconvex objective function in x . Many stochastic programs with probabilistic constraints have nonconvex feasible regions. Therefore, a second major difficulty in algorithm design for certain classes of stochastic programs is having to deal with a challenging nonconvex optimization problem.

Despite the aforementioned difficulties, there has been a steady progress toward developing solution methods for stochastic programs. Algorithm design for stochastic programs depends heavily on the class of problem being addressed and relies on exploiting the special structures of that class. That said, certain algorithmic ideas, such as decomposition, are repeatedly used in solving stochastic programs. Below, we review some of these commonly used algorithmic ideas. To illustrate the concepts, we use two-stage stochastic linear programs as an example and point out references for other classes of problems. We note that our treatment is brief in nature and serves as an introduction; as such, we refer the interested readers for further reading in the section titled “Concluding Remarks”. We assume a basic familiarity with optimization. Before we begin a more detailed look into these algorithmic ideas, we first give a broad overview.

Solution methods for stochastic programs can be classified into two categories: (i) exact and (ii) approximate solution methods. Exact solution methods are typically used when the number of realizations of $\tilde{\xi}$ is small-to-moderate. They also allow solution of approximations of stochastic programs. For example, Monte Carlo sampling-based approximations that use a sample size of n turn (SP) into a problem with a discrete distribution that has at most n realizations. Then, the resulting approximate problem can be solved via one of the exact solution methods. Given that it is not possible to solve many stochastic programs exactly, approximate solution methods are paramount in stochastic programming. The main idea here is to approximate the stochastic program with a much smaller, computationally tractable one. A desirable property of these approximations and the solutions obtained

via these approximations is that they converge to the true problem and its solution set in some sense. We now examine the two categories in more detail.

EXACT SOLUTION METHODS

An underlying theme of exact solution methods for stochastic programs is the exploitation of the special structures found in many stochastic programs to achieve tractability. For example, consider a two-stage stochastic linear program with recourse described above and assume $\tilde{\xi}$ has a finite number of realizations $\{\xi^1, \xi^2, \dots, \xi^S\}$, called *scenarios*, with respective probabilities $p^1, p^2, \dots, p^S$. Then, the two-stage stochastic linear program with recourse can be written in extensive form as a large scale linear program

$$\begin{aligned} \min_{x, y^1, \dots, y^S} \quad & cx + \sum_{s=1}^S p^s q^s y^s \\ \text{s.t.} \quad & Ax = b, \\ & T^s x + W^s y^s = r^s, \quad s = 1, 2, \dots, S, \\ & x \geq 0, \quad y^s \geq 0, \quad s = 1, 2, \dots, S. \end{aligned} \tag{1}$$

Note that in (SP) optimization is with respect to x ; while in the extensive form in Equation (1), optimization is with respect to x and $y^1, y^2, \dots, y^S$. Decisions x , called the *first stage decisions*, do not depend on a particular scenario s and must be made before knowing the realization of the random parameters. In contrast, decisions $y^1, y^2, \dots, y^S$, called the *second stage decisions*, depend on a particular scenario s . The constraint matrix of this linear program has a special structure, called the *block structure* [2, Chapter 5], which can be exploited to design efficient algorithms. A class of exact solution methods modifies well-known optimization algorithms to be efficiently used for stochastic programs by exploiting this structure. These include simplex-based algorithms [3] and interior point method implementations specific to stochastic programs [4–7].

Decomposition is a commonly used technique in stochastic programming as many stochastic programs are decomposable by

scenario and/or by stage. It is a classic “divide and conquer” technique that replaces a large scale, difficult problem with a number of smaller, tractable problems. Solving these smaller problems is more efficient than solving one large scale problem. Decomposition-based algorithms have been successfully used to solve two and multistage stochastic programs—some with risk-averse objectives [8,9]—as well as aiding efficient solution of other classes of (SP) such as stochastic programs with probabilistic constraints [10]. In the rest of this section, we discuss two common decomposition techniques in detail.

Primal Decomposition and the L-Shaped Method

Overview. We can rewrite Equation (1) as

$$\begin{aligned} z^* &= \min_x cx + \mathbb{E}h(x, s) \\ \text{s.t. } Ax &= b, x \geq 0, \end{aligned} \quad (2)$$

where

$$h(x, s) = \min_y \{q^s y : W^s y = r^s - T^s x, y \geq 0\}. \quad (\text{SUB}^s)$$

For ease of exposition, assume $\{\pi : \pi W^s \leq q^s\} \neq \emptyset$, $s = 1, 2, \dots, S$ and Equation (2) has *relatively complete recourse*, that is, for all $x \in X = \{Ax = b, x \geq 0\}$, (SUB^s) is feasible, for all $s = 1, 2, \dots, S$ and $X \neq \emptyset$ and compact. Let’s make the following observations: (i) for a fixed $x \in X$, Equation (2) decomposes into S subproblems given by (SUB^s) , $s = 1, 2, \dots, S$, (ii) it is well known that $\mathbb{E}h(x, s)$ is convex on X [11, Chapter 2, Proposition 13]. These observations constitute the heart of the primal decomposition algorithms for two-stage stochastic linear programs. In stochastic programming, this type of decomposition is referred to as the *L-shaped* method, owing to the shape of the block structure of a two-stage stochastic program [12]. However, the same algorithm with minor technical modifications is known as *Benders’ Decomposition* in integer programming [13] and as *Kelley’s cutting plane algorithm* in convex programming [14,15].

The idea of primal decomposition is to replace the *recourse function*, $\mathbb{E}h(x, s)$, by a lower approximating function. Since $\mathbb{E}h(x, s)$ is convex on X , such a lower approximation can be easily obtained via subgradients. The resulting problem is called the *master problem*, given by

$$\begin{aligned} \underline{z} &= \min_{x, \theta} cx + \theta \\ \text{s.t. } Ax &= b, x \geq 0, \\ G_i x + \theta &\geq g_i, i = 1, 2, \dots, I. \end{aligned} \quad (\text{MP}) \quad (3)$$

Here, θ is a variable representing the lower approximation of $\mathbb{E}h(x, s)$ and this lower approximation is given by the “cuts” in Equation (3). G_i are the cut gradients and g_i are the cut intercepts. The solution of the master problem provides two pieces of information. First, the optimal value of (MP) provides a lower bound on z^* , denoted by $\underline{z}$. Second, the optimal solution of the master problem, $\hat{x}$, yields a feasible ($A\hat{x} = b, \hat{x} \geq 0$) but typically suboptimal solution to Equation (2). This candidate solution, $\hat{x}$, is then passed to the subproblems (SUB^s) , $s = 1, 2, \dots, S$. Similar to the master problem, solution of the subproblems at $x = \hat{x}$ yields two pieces of information. First, the value of $\mathbb{E}h(\hat{x}, s)$ is obtained through $\sum_{s=1}^S p^s h(\hat{x}, s)$. This, together with the first stage cost at $\hat{x}$, that is, $\hat{z} = c\hat{x} + \mathbb{E}h(\hat{x}, s)$, provides an upper bound to z^* . Notice that since the lower approximation is sequentially refined throughout the algorithm, $\underline{z}$ is monotonically increasing. However, the upper bound, $\hat{z}$, can bounce as $\hat{x}$ changes. Therefore, we need to keep track of the best upper bound found so far, denoted as $\bar{z}$. The second piece of information that comes from solving the subproblems is a subgradient on $\mathbb{E}h(x, s)$. This information is then passed back to the relaxed master problem via a “cut” and the lower approximation is updated. With this more refined lower approximation, the master problem is solved again, a new candidate solution is obtained, and the algorithm continues in the same way until upper and lower bounds are sufficiently close, that is, $\bar{z} - \underline{z} \leq \text{toler} \cdot \max\{\|\bar{z}\|, \|\underline{z}\|\}$, where *toler* is a user specified tolerance.

To see how the subproblems yield a cut for the master problem, consider the simple

case of one scenario, $S = 1$. In this case, $\mathbb{E}h(x, s) = h(x, 1)$. The dual of (SUB¹) is $\max_{\pi} \{\pi(r^1 - T^1x) : \pi W^1 \leq q^1\}$. The feasible region of the dual problem is a polyhedron in standard form and is nonempty by our assumption, so, it has a finite number of extreme points, $\pi^1, \pi^2, \dots, \pi^L$. Furthermore, again by our assumption of relatively complete recourse, there exists an optimal solution at one these extreme points. Hence, we can equivalently write the dual as

$$h(x, 1) = \max_{l=1, \dots, L} \pi^l(r^1 - T^1x).$$

This is a piecewise linear convex function in x . There are an exponential number of extreme points; therefore, identifying all of them might not be efficient. The L-shaped method identifies only a subset of these that are needed to solve the problem. When (SUB¹) is solved at $\hat{x}$, this yields $\hat{\pi}$, one of the extreme dual points, which results in the cut

$$\theta \geq \hat{\pi}r^1 - \hat{\pi}T^1x. \quad (4)$$

Here, $-\hat{\pi}T^1$ is a subgradient of $h(x, 1)$ at $\hat{x}$ and Equation (4) is the subgradient inequality. The right-hand side of Equation (4) is equal to $\mathbb{E}h(\hat{x}, s)$ when evaluated at $x = \hat{x}$ and is less than or equal to $\mathbb{E}h(x, s)$ as we move away from $\hat{x}$. Relating Equation (4) to Equation (3), the cut gradient is $G = \hat{\pi}T^1$ and the cut intercept is $g = \hat{\pi}r^1$.

When there are a number of scenarios, that is, $S > 1$, using the fact that subgradient of $\mathbb{E}h(x, s)$ is equal to the expectation of the individual subgradients in our case [11, Chapter 2, Proposition 13], we can obtain a cut with

$$G_i = \sum_{s=1}^S p^s \pi_i^s T^s, \text{ and } g_i = \sum_{s=1}^S p^s \pi_i^s r^s, \quad (5)$$

where $\pi_i^s, s = 1, 2, \dots, S$ are the optimal duals obtained from (SUB^s)'s at iteration i of the algorithm. Cuts of the form of Equation (3) are called the *optimality cuts*. If we relax the assumption of relatively complete recourse, it is possible to obtain $\hat{x} \in X$ when (MP) is solved, which causes infeasibility in at least

one of the subproblems. Such a first stage decision must be eliminated. This can be done via the so-called *feasibility cuts*. Feasibility cuts are obtained in a similar fashion to Equations (4) and (5) but using the Phase-I problem corresponding to (SUB^s)'s. For brevity, we skip the details and refer the readers to Birge and Louveaux [2] and Ruszczyński and Shapiro [11].

Enhancements and Extensions. We presented a basic version of the L-shaped method above with only optimality cuts and what is called a *single cut version*. A *multicut* version, analyzed by Birge and Louveaux [16], replaces θ in (MP) by $\sum_{s=1}^S p^s \theta^s$, where θ^s is a lower approximation of $h(x, s)$ for each $s = 1, 2, \dots, S$. In this case, instead of the aggregate cuts in Equation (3), S cuts of the form of Equation (4) are generated from each subproblem and added to the master problem. This can speed up the convergence of the L-shaped method. However, it can also result in a larger master problem, especially in later iterations. Variations in between single and multicut versions are possible, where several scenarios can be bundled together to form a cut for this group.

In the early iterations of the L-shaped method, solutions can oscillate wildly, slowing convergence. The *regularized decomposition* of Ruszczyński [17] and Ruszczyński and Świetanowski [18] adds a quadratic regularizing term to the objective function of the master problem that encourages the new solution to remain close to the current solution. This controls the size of the steps taken by the algorithm and also limits the number of cuts that need to be stored in the master problem. As a result, significant savings in total computation time can be achieved, even though a quadratic master problem needs to be solved. In an effort to speed up convergence, *trust regions* similarly aim to control the size of the steps taken by the algorithm. Trust regions are constraints in the master problem that restrict the new solution to be within a region close to the current solution. Typically, l_1 or l_∞ norms are used as these can be represented as linear constraints. Successful implementations of trust regions in stochastic programming can be

found in Keller and Bayraksan [19], Linderoth and Wright [20], and Santoso *et al.* [21]. Finally, L-shaped method is amenable to *parallel* implementation [20,22].

The L-shaped method has been extended to work for a larger class of stochastic programs. For multistage stochastic programs, *nested L-shaped* or *nested decomposition* method [23,24] decomposes multistage stochastic programs by stage and works in a similar fashion. In the forward pass of the algorithm, candidate solutions for the future stages are passed on, and in the backward pass of the algorithm, cuts are provided to the previous stages. When the first stage variables are integer and the second stage is continuous, L-shaped method works exactly like the Benders decomposition. However, when there are integrality restrictions in the second stage, the recourse function is often discontinuous and nonconvex in x and this complicates the L-shaped method. The *integer L-shaped* method of Laporte and Louveaux [25] overcomes this difficulty by providing cuts that are linear in x in the case of pure binary first stage variables. The second stage (mixed) integer programs need to be solved to optimality to obtain these cuts, which can be computationally undesirable. Also, these cuts are known to be weak. The *disjunctive decomposition* method of Sen and Hingle [26] expands on the L-shaped method and allows efficient solution of the subproblems as linear programming (LP) relaxations. The LP relaxations are sequentially refined throughout the algorithm by the use of disjunctive cuts that eliminate fractional solutions in the second stage. For computational experience an extensions on disjunctive decomposition see, Ntamo [27], Ntamo and Sen [28,29], and Sen and Sherali [30].

Dual Decomposition

Let us again rewrite Equation (1), this time by duplicating the first stage variables for each scenario. The resulting equivalent reformulation of Equation (1) is

$$\min_{\substack{x^1, \dots, x^S, \\ y^1, \dots, y^S}} \sum_{s=1}^S p^s (cx^s + q^s y^s)$$

$$\text{s.t. } Ax^s = b, \quad s = 1, 2, \dots, S, \quad (6)$$

$$T^s x^s + W^s y^s = r^s, \quad s = 1, 2, \dots, S, \quad (7)$$

$$x^s \geq 0, y^s \geq 0, \quad s = 1, 2, \dots, S, \quad (8)$$

$$x^1 = x^2 = \dots = x^S. \quad (9)$$

Constraints (9) are called the *nonanticipativity constraints*, which force the first stage variables to be identical across scenarios. Nonanticipativity constraints can be represented in different ways and Equation (9) is one possible way. The idea of the dual decomposition methods is to relax the nonanticipativity constraints and solve the resulting Lagrangian dual problem. In the linear case, this can be done by the classic Dantzig–Wolfe decomposition [31]. Let $Gx = 0$ represent the nonanticipativity constraints. For example, when there are two scenarios ($S=2$), $G = [I; -I]$, where I is an identity matrix of size d_x . Given Lagrange multipliers λ , the Lagrangian is

$$\mathcal{L}(x, y, \lambda) = \sum_{s=1}^S p^s (cx^s + q^s y^s) + \sum_{s=1}^S \langle \lambda, G^s x^s \rangle. \quad (10)$$

The relaxed problem $D(\lambda) = \min_{x,y} \{\mathcal{L}(x, y, \lambda)\}$: Equations (6)–(8) decomposes into S separable scenario subproblems. Each of the resulting subproblems is a two-stage stochastic program with exactly one scenario, s , and decisions x^s and y^s . The Lagrangian dual

$$\max_{\lambda} D(\lambda)$$

can then be solved by standard optimization techniques. For example, it is well known that $D(\lambda)$ is concave in λ and a master problem can be created that gives an upper approximation to $D(\lambda)$, whose solution yields a $\hat{\lambda}$ and an upper bound, $\bar{z}$. $\hat{\lambda}$ is then passed on the subproblems, that is, $D(\hat{\lambda}) = \min_{x,y} \{\mathcal{L}(x, y, \hat{\lambda})\}$: Equations (6)–(8) is solved by solving the S separable subproblems. This in turn provides a better upper approximation by providing a “cut” to the master problem, again by the use of subgradient inequality—but for a concave

function—and a possible lower bound. Primal solutions, x , are also updated and the best lower bound found so far is stored as $\underline{z}$. The algorithm continues until the upper and lower bounds are sufficiently close, that is, $\bar{z} - \underline{z} \leq \text{toler} \cdot \max\{|\bar{z}|, |\underline{z}|\}$, where toler is a user-specified tolerance.

We note a difference between the L-shaped method and the dual decomposition method. In the L-shaped method, the master problem decisions are x and θ , which represents the lower approximation of $\mathbb{E}h(x, s)$, and in the dual decomposition, the master problem decisions are λ and θ , which, this time, represents the upper approximation of $D(\lambda)$. The size of λ grows as the number of scenarios grow but the size of x remains the same. Therefore, solving the dual master can become harder as S grows. Nevertheless, dual decomposition methods have been successfully applied to multistage stochastic programs, especially when the subproblems have a special structure, enabling efficient solution.

In particular, *augmented Lagrangian* methods have proven useful in multistage stochastic programs. These augment the Lagrangian equation (10) with a quadratic term that penalizes deviations in the nonanticipativity constraints. Two of the well-known dual decomposition methods for stochastic programs, the *progressive hedging* algorithm of Rockafeller and Wets [32] and the *scenario decomposition* method of Mulvey and Ruszczyński [33], are augmented Lagrangian methods. They differ in the way the nonanticipativity constraints are written, the way the subproblems are solved, and the way the Lagrange multipliers are updated. For stochastic integer programs, a duality gap exists, so, a Lagrangian dual method can be combined with branch and bound [34]. Progressive hedging has been applied to mixed integer multistage stochastic programs [35–37] without convergence guarantees due to loss of convexity but empirical convergence has been observed. For a computational comparison of nested L-shaped and an augmented Lagrangian implementation on a class of multistage stochastic mean-variance portfolio problems, see Parpas and Rustem [38].

APPROXIMATE SOLUTION METHODS

As d_ξ , the dimension of the random vector, grows stochastic programs can quickly become intractable. For instance, for discrete random vectors the problem size can grow exponentially in this dimension. For continuous random vectors, taking expectations—even for a fixed x —gets very difficult as the dimension of the integral grows. Furthermore, the size of the multistage stochastic programs grows exponentially in the number of stages. Some of these approximations exploit the structural properties of the underlying stochastic programs—such as the bounds based on convexity of the recourse function—and some are general, for example, Monte Carlo sampling-based approximations. Once a computationally tractable approximation of (SP) is formed, the idea of approximate solution methods is to sequentially refine this approximation until a solution of desired quality is obtained. Sometimes, the tractable approximation is used within one of the exact solution methods described above, providing estimates of the expectations and subgradients required throughout the algorithm. Approximate solution methods mainly differ in the way the approximations are formed, the way refinements to these approximations are made, and if they are used internally to other solution procedures or not. We review two common types of approximations used for solving stochastic programs below.

Monte Carlo Sampling-Based Approximations

Let $\tilde{\xi}^1, \tilde{\xi}^2, \dots, \tilde{\xi}^n$ be independent and identically distributed (i.i.d.) random variables as $\tilde{\xi}$. A sampling approximation of Equation (1) or Equation (2) is

$$z_n^* = \min_{x \in X} cx + \frac{1}{n} \sum_{i=1}^n h(x, \tilde{\xi}^i). \quad (\text{SP}_n)$$

This is referred to as a *sample average approximation* (SAA). Let x_n^* denote an optimal solution to (SP_n) . The pair (z_n^*, x_n^*) provides statistical estimators of the optimal value and optimal solution of (SP), respectively. Monte Carlo sampling-based

approximations are typically justified by providing conditions under which z_n^* converges to z^* and the limit points of x_n^* solve (SP) [39–42]. There is also work on the rate of convergence of these estimators. For certain classes of stochastic programs such as stochastic integer programs with integer first stage, and for two-stage stochastic linear programs with discrete distributions, the rate of convergence can be exponential [43,44]. While the above uses i.i.d. sampling, sampling can be done in different ways, for instance, stratified sampling or antithetic variates can be used to reduce variance. For the analysis and computational experiments with alternate sampling methods and variance reduction techniques in the context of stochastic programming see [45–48].

(SP_n) is essentially the same as Equation (1) with $S = n$ and $p^1 = p^2 = \dots = p^n = \frac{1}{n}$. Of course, in the case of finite Ξ , when S is prohibitively large to allow an exact solution method, sample size can be chosen such that $n \ll S$ to achieve tractability. Then, (SP_n) can be solved by any of the exact solution methods described above. Such an approach is referred to as *external sampling*. We note that this is not an approximate “solution method” but provides approximations, (z_n^*, x_n^*) , that are statistical in nature. Sampling is done externally to construct an instance of (SP_n) and the resulting problem is solved via an exact solution method. For (SP) with $K \geq 1$, such as stochastic programs with probabilistic constraints, feasibility of the solutions found by SAA becomes an issue. Sample sizes that guarantee feasibility with a certain confidence level have been analyzed for such problems [49,50].

Sampling approximations can also be used *internal* to an exact solution method. For example, sampling can be done within a branch and bound algorithm [51]. We collectively refer to these approximate solution methods as *internal sampling* methods. One set of such algorithms are rooted in the L-shaped method. The *stochastic decomposition* method of Hige and Sen [52] uses sampling-based cutting planes within the L-shaped method. Pereira and Pinto [53] provide an extension in the multistage setting and Infanger [54] and Dantzig and Infanger

[45] embed importance sampling within the L-shaped method.

Another class of sampling-based solution methods are the stochastic versions of steepest descent where the (sub)gradients used within the algorithm are replaced by their sampling-based estimators. *Stochastic approximation* (SA) [55] and the *stochastic quasi-gradient* (SQG) [56] belong to this category. They are rooted in the Robbins and Monro [57] and Kiefer and Wolfowitz [58] stochastic approximation methods, where the former uses unbiased estimators of the (sub)gradients and the latter uses finite differences to estimate the (sub)gradients. One advantage of these methods is that they have the potential to handle decision-dependent stochasticity; however, they do not take full advantage of the special structures found in many stochastic programs. They can also be sensitive to the step sizes used and can have slower convergence. Recent modifications to the SA method use longer step sizes to improve convergence and averages solutions to reduce noise. The resulting robust mirror descent SA has promising numerical performance [59].

Bounding Approximations

Another way to handle the difficult expectations in (SP) is by replacing them with bounds that are easier to calculate. These bounds can be embedded in a solution method that sequentially refines them. Bounds in stochastic programming can be derived via: (i) appropriate approximations of the distribution of ξ , or (ii) functional approximations that replace the complicated expectations with an easy function to calculate, and typically for multistage problems by (iii) aggregating the time stages and appropriately changing the information structure and the timing of the data revelation. For details on bounds of the third type, we refer the readers to [60–63]. Bounds for stochastic programs with probabilistic constraints can be obtained by replacing the probabilistic constraint(s) with expressions using probability inequalities, such as generalizations of Chebyshev’s inequality. These can be viewed as bounds of type two above. For details, we refer the readers to Section 9.4 of Birge and

Louveaux [2] for earlier developments and the article by Nemirovski and Shapiro [64] for more recent developments in this area.

Bounds of the first type, that is, that approximate the probability distribution, have been used most often in an algorithmic setting. These include bounds based on Jensen's [65] and Edmundson-Madansky (E-M) [66–68] inequalities. Bounds obtained through Jensen's inequality replace the distribution of $\tilde{\xi}$ by a single point, $\mathbb{E}\tilde{\xi}$, the expected value of $\tilde{\xi}$. Then, if $h(x, \tilde{\xi})$ is convex on Ξ , a lower bound is obtained by $h(x, \mathbb{E}\tilde{\xi}) \leq \mathbb{E}h(x, \tilde{\xi})$. Calculating $h(x, \mathbb{E}\tilde{\xi})$ for a fixed x is quite easy; it requires only one function evaluation at $\mathbb{E}\tilde{\xi}$. Bounds obtained through E-M inequality, on the other hand, replace the distribution of $\tilde{\xi}$ with an extremal distribution that puts mass on the extreme points of the supports of the elements of $\tilde{\xi}$ (assuming Ξ is bounded). Under convexity of $h(x, \tilde{\xi})$ on Ξ , this provides an upper bound. Since $h(x, \tilde{\xi})$ has to be evaluated at the extreme points of Ξ , the computational effort grows exponentially in $d_{\tilde{\xi}}$. Bounds based on Jensen's and E-M inequalities have been improved over the years to allow for correlations, use of more distributional information such as second moments, and also to allow for more general, possibly unbounded Ξ [69–72]. An alternative to the E-M bound is the piecewise linear bound [73–75], which replaces $\mathbb{E}h(x, \tilde{\xi})$ by a sum of $d_{\tilde{\xi}}$ one-dimensional expectations. Thus, the computational effort grows linearly in $d_{\tilde{\xi}}$. These are bounds of the second type, that is, they are based on functional approximations. Note that bounds based on functional approximations and extensions of bounds based on Jensen's and E-M inequalities can be formed by *generalized moment problems*, which have been first explored in the context of stochastic programming by Dupačová [76].

Bounds can be used in an algorithmic way to find approximate solutions to (SP). These methods are referred to as *sequential approximation methods* [77,78]. They sequentially partition the support of the random vector Ξ and form upper and lower bounds on the conditional expectations of each partition. As the partitions become more refined, bounds become tighter and

the method terminates when upper and lower bounds are sufficiently close to each other. Another way to refine the bounds is to tighten the generalized moment problems.

Convergence and Quality of Approximate Solutions

Convergence of approximations is a desired property. While convergence can be shown in a variety of ways, *epiconvergence* provides an appropriate framework for approximations of optimization problems. A sequence of functions g_v is said to epiconverge to g if the epigraphs of g_v converge to that of g . The epigraph of g is defined as the set $\text{epi } g = \{(x, \alpha) : \alpha \geq g(x)\}$. Epiconvergence is convenient for optimization because of the property that if $\hat{x}$ is an accumulation point of $x_v^* \in \arg \min g_v(x)$, then it solves the original problem, that is, $\hat{x} \in \arg \min g(x)$. For instance, in (SP_n) , $g_v(x) = \mathbb{E}_n \bar{f}_0(x, \tilde{\xi})$ and $g(x) = \mathbb{E} \bar{f}_0(x, \tilde{\xi})$, where $\bar{f}_0$ is similar to f_0 but captures the feasibility constraints by taking value $+\infty$ when x is not a feasible solution, and $\mathbb{E}_n$ is the expectation taken with respect to the empirical measure $\mathbb{P}_n = \sum_{i=1}^n \frac{1}{n} \delta_{\tilde{\xi}^i}$, where $\delta_{\tilde{\xi}^i}$ denotes the unit mass on $\tilde{\xi}^i$. A similar setup is possible, for example, when a bounding approximation replaces the probability distribution with $\mathbb{P}_v$. Then, if $\mathbb{P}_v$ converges weakly to $\mathbb{P}$ almost surely, and f satisfies certain regularity conditions, we can achieve epiconvergence. For a comprehensive treatment of epiconvergence see Rockafellar and Wets [79], for its implications on approximate optimization problems see Kall [80] and for its use in stochastic programming see Birge and Wets [77].

While convergence is a desired property, in practice, given finite computational power, approximations of (SP) yield only approximate solutions. Thus, the quality of the obtained solutions needs to be determined. Such quality statements can sometimes be made within an approximate solution method, for example, by providing the difference between upper and lower bounds. When such a quality statement is not available, Monte Carlo sampling-based methods can be used to provide point and interval estimators of the optimality gap of a candidate solution

found by any method. Such procedures for (SP) with $K = 0$ under fairly general conditions have been developed in Mak *et al.* [81], see also the tutorial by Bayraksan and Morton [82]. Note that these statistical estimators can be used in conjunction with any approximate method that eventually produces solutions that solve (SP) to design an approximate solution procedure. For instance, external or internal sampling methods can generate x_n^* using sample size n . Then, a statistical estimator of its optimality gap can be generated. If the optimality gap is sufficiently small, then this solution provides a high-quality solution with a desired confidence level. Otherwise, the sample size is increased. For rules to increase sample sizes and to stop such a *sequential sampling procedure* see Bayraksan and Morton [83].

Solution quality for (SP) with $K \geq 1$ involves two components: first, to ensure that the approximate solution is feasible (depending on the type of approximation) and second, to provide bounds on the optimal value. If approximation is done via Monte Carlo sampling, then, as mentioned above, feasibility becomes an issue. For bounding approximations, feasibility can be automatically satisfied as the constraints that replace, for instance, a probabilistic constraint guarantee that this constraint is satisfied. For methods that provide probabilistic statements on the feasibility of approximate solutions and the objective values for this class of problems, see [50,64,84].

CONCLUDING REMARKS

This article provides an overview of solution methods for stochastic programs. In particular, decomposition-based solution techniques are discussed. Two common classes of approximate solution methods, one rooted in Monte Carlo sampling-based approximations and the other in deterministically valid bounding approximations are introduced. Convergence and solution quality of approximate solutions are discussed.

As this is intended as an overview article, and the topic is too extensive to cover in detail, for more details on algorithmic approaches to different classes of

stochastic programs, we refer the reader to the section titled “Optimization under Uncertainty” of this encyclopedia. In our treatment of the subject, we mainly focused on stochastic programs with recourse, so, for solution methods for stochastic programs with probabilistic constraints. We note that discretization can be done in other ways besides Monte Carlo sampling, for details on different scenario generation and reduction techniques see the article titled **Scenario Generation** in this encyclopedia. Finally, for information about stochastic programming software. For a deeper understanding of the methods discussed above, see textbooks Birge and Louveaux [2] and Shapiro *et al.* [85] and also the handbook [11]. Interested readers can also look at the stochastic programming community website, www.stoprog.org, which contains a wealth of information including introductory materials. A searchable online bibliography on stochastic programming can be found at <http://www.eco.rug.nl/mally/spbib.html>.

We end by noting that developing efficient solution methods for stochastic programs is an exciting and ever growing area of research. Some of the current research includes solving different classes of (SP), such as multistage stochastic mixed integer programs, possibly with nonlinear constraints/objectives, and stochastic programs with stochastic dominance and probabilistic constraints. Another challenge is to develop continuous versions of stochastic programs and enable their solutions. For instance, stochastic programs with (stochastic) differential equations still await modeling and efficient solution techniques.

Acknowledgments

This work was supported, in part, by the National Science Foundation under Grant EFRI-0835930.

REFERENCES

1. Wallace SW, Ziemba WT, editors. Applications of stochastic programming. MPS-SIAM Series in Optimization. Philadelphia (PA). 2005.
2. Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.

3. Kall P. Computational methods for solving two-stage stochastic linear programming problems. *J Appl Math Phys* 1979;30: 261–271.
4. Berkelaar A, Dert C, Oldenkamp B, *et al.* A primal-dual decomposition-based interior point approach to two-stage stochastic linear programming. *Oper Res* 2002;50(5):904–915.
5. Birge JR, Holmes DF. Efficient solution of two-stage stochastic linear programs using interior point methods. *Comput Optim Appl* 1992;1:245–276.
6. Czyzyk J, Fourer R, Mehrotra S. A study of the augmented system and column-splitting approaches for solving two-stage stochastic linear programs by interior-point methods. *ORSA J Comput* 1995;7(4):474–490.
7. Lustig IJ, Mulvey JM, Carpenter TJ. Formulating two-stage stochastic programs for interior point methods. *Oper Res* 1991;39:757–770.
8. Ahmed S. Convexity and decomposition of mean-risk stochastic programs. *Math Program* 2006;106:433–446.
9. Schultz R, Tiedemann S. Conditional value-at-risk in stochastic programs with mixed-integer recourse. *Math Program* 2006; 105:365–386.
10. Luedtke J. An integer programming and decomposition approach to general chance-constrained mathematical programs. In: Eisenbrand F, Shephard B, editors. *Proceedings of IPCO 2010, Lecture Notes in Computer Science*. Berlin Heidelberg: Springer; 2010. pp. 271–284.
11. Ruszczyński A, Shapiro A, editors. *Handbooks in operations research and management science*. Volume 10, *Stochastic programming*. Amsterdam, The Netherlands: Elsevier; 2003.
12. Van Slyke RM, Wets RJ-B. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM J Appl Math* 1969;17:638–663.
13. Benders JF. Partitioning procedures for solving mixed-variable programming problems. *Numer Math* 1962;4:238–252.
14. Kelley JE. The cutting plane method for solving convex programs. *SIAM J Ind Appl Math* 1960;8:703–712.
15. Hiriart-Urruty JB, Lemaréchal C. *Convex analysis and minimization algorithms II. Comprehensive studies in mathematics*. Berlin Heidelberg, Germany: Springer; 1993.
16. Birge JR, Louveaux FV. A multicut algorithm for two-stage stochastic linear programs. *Eur J Oper Res* 1988;34:384–392.
17. Ruszczyński A. A regularized decomposition method for minimizing a sum of polyhedral functions. *Math Program* 1986;35:309–333.
18. Ruszczyński A, Świetanowski A. Accelerating the regularized decomposition method for two stage stochastic linear problems. *Eur J Oper Res* 1997;101:328–342.
19. Keller BD, Bayraksan G. Scheduling jobs sharing multiple resources under uncertainty: a stochastic programming approach. *IEE Trans* 2010;42:16–30.
20. Linderoth JT, Wright SJ. Decomposition algorithms for stochastic programming on a computational grid. *Comput Optim Appl* 2003;24:207–250.
21. Santoso T, Ahmed S, Goetschalckx M, *et al.* A stochastic programming approach for supply chain network design under uncertainty. *Eur J Oper Res* 2005;167:96–115.
22. Nielsen SS, Zenios SA. Scalable parallel benders decomposition for stochastic linear programming. *Parallel Comput* 1997;23(8):1069–1088.
23. Birge JR. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper Res* 1985;33:989–1007.
24. Noël M-C, Smeers Y. Nested decomposition of multistage nonlinear programs with recourse. *Math Program* 1987;37:131–152.
25. Laporte G, Louveaux FV. The integer L-shaped method for stochastic integer programs with complete recourse. *Oper Res Lett* 1993;13:133–142.
26. Sen S, Hige J. The C^3 theorem and a D^2 algorithm for large scale stochastic mixed-integer programming: set convexification. *Math Program* 2005;104:1–20.
27. Ntamo L. Disjunctive decomposition for two-stage stochastic mixed-binary programs with random recourse. *Oper Res* 2010;58:229–243.
28. Ntamo L, Sen S. The million variable “march” for stochastic combinatorial optimization. *J Glob Optim* 2005;32:385–400.
29. Ntamo L, Sen S. A comparative study of decomposition algorithms for stochastic combinatorial optimization. *Comput Optim Appl* 2008;40:299–319.
30. Sen S, Sherali H. Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming. *Math Program* 2006;106:203–222.
31. Dantzig GB, Wolfe P. Decomposition principle for linear programs. *Oper Res* 1960;

- 8:101–111.
32. Rockafellar RT, Wets RJ-B. Scenarios and policy aggregation in optimization under uncertainty. *Math Oper Res* 1991;16:119–147.
 33. Mulvey JM, Ruszczyński A. A new scenario decomposition method for large-scale stochastic optimization. *Oper Res* 1995;43:477–490.
 34. Carøe C, Schultz R. Dual decomposition in stochastic integer programming. *Oper Res Lett* 1999;24:37–54.
 35. Løkketangen A, Woodruff D. Progressive hedging and Tabu search applied to mixed integer (0,1) multistage stochastic programming. *J Heuristics* 1996;2:111–128.
 36. Takriti S, Birge JR. Lagrangian solution techniques and bounds for loosely coupled mixed-integer stochastic programs. *Oper Res* 2000;48:91–98.
 37. Takriti S, Birge JR, Long E. A stochastic model for the unit commitment problem. *IEEE Trans Power Syst* 1996;11:1497–1508.
 38. Parpas P, Rustem B. Computational assessment of nested Benders and augmented Lagrangian decomposition for mean-variance multistage stochastic problems. *INFORMS J Comput* 2007;19(2): 239–247.
 39. Dupačová J, Wets RJ-B. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann Stat* 1988;16:1517–1549.
 40. King AJ, Rockafellar RT. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math Oper Res* 1993;18:148–162.
 41. Shapiro A. Asymptotic analysis of stochastic programs. *Ann Oper Res* 1991;30:169–186.
 42. Shapiro A. Monte Carlo sampling methods. In: Ruszczyński A, Shapiro A, editors. Volume 10, *Handbooks in operations research and management science: stochastic programming*. Amsterdam, The Netherlands: Elsevier; 2003. pp. 353–425.
 43. Kleywegt AJ, Shapiro A, Homem-de-Mello T. The sample average approximation method for stochastic discrete optimization. *SIAM J Optim* 2001;12(3):479–502.
 44. Shapiro A, Homem-de-Mello T. On rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM J Optim* 2000;11:70–86.
 45. Dantzig GB, Infanger G. Large-scale stochastic linear programs—importance sampling and benders’ decomposition. In: Brezinski C, Kulisch U, editors. *Computational and applied mathematics I*. Amsterdam: North-Holland Publishing Co.; 1991. pp. 111–120.
 46. Higle JL. Variance reduction and objective function evaluation in stochastic linear programs. *INFORMS J Comput* 1998; 10:236–247.
 47. Homem-de-Mello T. On rates of convergence for stochastic optimization problems under non-i.i.d. sampling. *SIAM J Optim* 2008;19:524–551.
 48. Linderoth J, Shapiro A, Wright S. The empirical behavior of sampling methods for stochastic programming. *Ann Oper Res* 2006; 142:219–245.
 49. Campi MC, Garatti S. The exact feasibility of randomized solutions of robust convex programs. *SIAM J Optim* 2008;19:1211–1230.
 50. Luedtke J, Ahmed S. A sample approximation approach for optimization with probabilistic constraints. *SIAM J Optim* 2008;19:674–699.
 51. Norkin VI, Pflug GCh, Ruszczyński A. A branch and bound method for stochastic global optimization. *Math Program* 1998; 83:425–450.
 52. Higle JL, Sen S. *Stochastic decomposition: a statistical method for large scale stochastic linear programming*. Dordrecht: Kluwer Academic Publishers; 1996.
 53. Pereira MVF, Pinto LMVG. Multi-stage stochastic optimization applied to energy planning. *Math Program* 1991;52:359–375.
 54. Infanger G. Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Ann Oper Res* 1992;39:69–95.
 55. Kushner HJ, Yin GG. *Stochastic approximation algorithms and applications*. New York: Springer; 1997.
 56. Ermoliev Y. Stochastic quasigradient methods. In: Ermoliev Y, Wets RJ-B, editors. *Numerical techniques for stochastic optimization*. Berlin: Springer; 1988. pp. 141–185.
 57. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951; 22:400–407.
 58. Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. *Ann Math Stat* 1952;23:462–466.
 59. Nemirovski A, Juditsky A, Lan G, *et al.* Robust stochastic approximation approach to stochastic programming. *SIAM J Optim* 2009;19:1574–1609.
 60. Birge JR. Aggregation bounds in stochastic linear programming. *Math Program* 1985;31:25–41.

61. Kuhn D. Aggregation and discretization in multistage stochastic programming. *Math Program* 2008;113:61–94.
62. Morton DP, Wood RK. Restricted-recourse bounds for stochastic linear programming. *Oper Res* 1999;47:943–956.
63. Wright SE. Primal-dual aggregation and disaggregation for stochastic linear programs. *Math Oper Res* 1994;19:893–908.
64. Nemirovski A, Shapiro A. Convex approximations of chance constrained programs. *SIAM J Optim* 2006;17:969–996.
65. Jensen JL. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math* 1906;30:175–193.
66. Edmundson HP. Bounds on the expectation of a convex function of a random variable. Technical report. The Rand Corporation Paper 982. Santa Monica (CA); 1956.
67. Madansky A. Bounds on the expectation of a convex function of a multivariate random variable. *Ann Math Stat* 1959;30:743–746.
68. Madansky A. Inequalities for stochastic linear programming problems. *Manage Sci* 1960;6:197–204.
69. Dokov SP, Morton DP. Second-order lower bounds on the expectation of a convex function. *Math Oper Res* 2005;30:662–677.
70. Frauendorfer K. Solving slp recourse problems with arbitrary multivariate distributions—the dependent case. *Math Oper Res* 1988;13:377–394.
71. Edirisinghe NCP, Ziemba WT. Bounding the expectation of a saddle function with application to stochastic programming. *Math Oper Res* 1994;19:314–340.
72. Edirisinghe NCP, You G-M. Second-order scenario approximation and refinement in optimization under uncertainty. *Ann Oper Res* 1996;64:143–178.
73. Birge JR, Wallace SW. A separable piecewise linear upper bound for stochastic linear programs. *SIAM J Control Optim* 1988;26:725–739.
74. Birge JR, Wets RJ-B. Sublinear upper bounds for stochastic programs with recourse. *Math Program* 1989;43:131–149.
75. Birge JR, Dulá JH. Bounding separable recourse functions with limited distribution information. *Ann Oper Res* 1991;30:277–298.
76. Dupačová (as Žáčková) J. On minimax solutions of stochastic linear programming problems. *Časopis pro Pěstování Matematiky* 1966;91:423–429.
77. Birge JR, Wets RJ-B. Designing approximation schemes for stochastic optimization problems, in particular, for stochastic programs with recourse. *Math Program Study* 1986;27:54–102.
78. Kall P. Stochastic programming with recourse: upper bounds and moment problems—a review. In: Guddat J, *et al.* editors. *Advances in mathematical optimization*. Berlin: Akademie-Verlag; 1988. pp. 86–103.
79. Rockafellar RT, Wets RJ-B. *Variational analysis*. Berlin: Springer; 1998.
80. Kall P. Approximation to optimization problems: an elementary review. *Math Oper Res* 1986;11:9–18.
81. Mak WK, Morton DP, Wood RK. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Oper Res Lett* 1999;24:47–56.
82. Bayraksan G, Morton DP. Assessing solution quality in stochastic programs via sampling. *Tutorials in Operations Research* 2009;5:102–122.
83. Bayraksan G, Morton DP. A sequential sampling procedure for stochastic programming. *Oper Res* 2010. In press.
84. Wang W, Ahmed S. Sample average approximation of expected value constrained stochastic programs. *Oper Res Lett* 2008;36:515–519.
85. Shapiro A, Dentcheva D, Ruszczyński A. *Lectures on stochastic programming: modeling and theory*, MPS-SIAM series on optimization. Philadelphia (PA): Society for Industrial and Applied Mathematics (SIAM) and the Mathematical Programming Society (MPS); 2009.

SOME OPTIMIZATION MODELS AND TECHNIQUES FOR ELECTRIC POWER SYSTEM SHORT-TERM OPERATIONS

ANDY SUN

Georgia Institute of Technology,
H. Milton Stewart School of
Industrial and Systems
Engineering, Atlanta, Georgia

DZUNG T. PHAN

IBM T.J. Watson Research
Center, Business Analytics
and Mathematical Sciences
Department, Yorktown
Heights, New York

INTRODUCTION

It is not an exaggeration to say that modern electric power systems are built on optimization models and solution techniques. Figure 1 shows a picture of the major optimization problems in power system operations with a timescale to indicate the horizon of the different types of decisions involved.

As a simple description, the long-term system planning studies the question of how much new generation and transmission capacity is needed and where in the power system they should be built in order to satisfy a projected future demand in a time horizon of 5–15 years or longer; the maintenance scheduling problem aims to optimally schedule the maintenance of generators, transformers, and other equipments so that the system operation will be disturbed as little as possible at a low cost, where the time horizon of decision is usually 1 year; the short-term operation problems determine daily or intraday generation schedules to match demand and satisfy various constraints.

All the major decision problems described earlier are built on two fundamental problems, the so-called unit commitment (UC) and economic dispatch based on optimal

power flow (OPF). In this article, we focus on these two problems and introduce their optimization models, important variations, and solution methods. We review both the basic formulations and the most recent advances in the field. In the remaining of the introduction, we give a brief overview of the basic structure of power systems with an introduction to common notations used in the literature. We then describe the UC and OPF problems.

A typical power system involves three major subsystems: generation, transmission network, and the distribution system. Figure 2 shows a simple illustration. The generation subsystem usually consists of industrial scale power plants, each of which normally has several generators, also called *generating units*. The voltage of the power produced by the generating units is stepped up by transformers to voltage levels at 138, 230 kV, or higher, before connecting to the transmission network. The high voltage levels help to reduce the thermal loss incurred during long distance transfer of power along transmission lines. The nodes of the transmission network, also called *buses* in the power system literature, represent physical points where the transmission lines connect to each other, and generators and/or loads are linked to the network. Transmission network is usually sparse and meshed. Before the power is served to the load, voltage is again stepped down to low levels of a few kilo volts and ultimately to 120 or 240 V in North America. This low voltage network is called the *distribution system*, which directly connects to the end users of electricity.

The generation scheduling problem considered in this article mainly deals with the high voltage transmission network, although with the increase of distributed generators, the distribution system is changing from a largely consumer network to a hybrid of power consumers and suppliers. The models we present in this article can be extended to the distribution system as well.

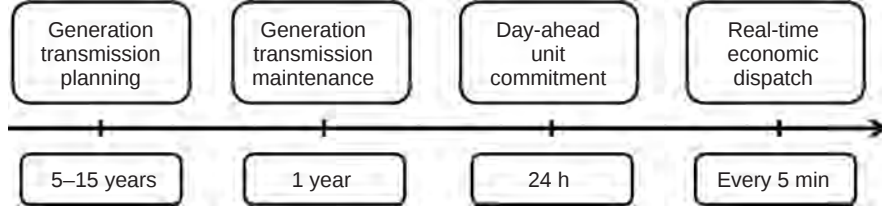


Figure 1. Major problems in power systems operation. The time arrow points to the start of real-time operation. The locations of the problems on the time arrow indicate their relative distance to the real-time operation. The typical decision horizon of each problem is given under each problem.

Figure 3 shows a simple example of a power system, which has 14 buses, 20 transmission lines, 5 generators, and 11 loads. Transformers, breakers, and other equipments are ignored. Denote the set of buses in the system as $\mathcal{N} = \{1, 2, \dots, 13\}$, and the subset of buses connected to generators as $\mathcal{G} = \{1, 2, 3, 6, 8\}$. Denote the set of transmission lines as $\mathcal{L} = \{(1, 2), (1, 5), (2, 5), \dots, (13, 14)\}$, where each transmission line is specified by its two end buses. Each transmission line (i, j) is characterized by its electrical properties, such as the conductance g_{ij} and the susceptance b_{ij} .

Each generation unit can be either turned on to produce electricity or turned off. When the generation unit is turned on, it is called *committed*; when it is turned off, it is called *decommitted*. The on/off state of a generator is called its *commitment status*. The unit commitment problem determines the commitment status and generation output of each generator to satisfy electricity demand over a time horizon of a day or a few days. There are various constraints in the UC problem. For example, the minimum up (down) time constraints require the generators remain in the on (off) state for a minimum amount of time once they are turned on (off); the ramping constraints require that the change in generation output between two consecutive time periods of a generator cannot exceed a limit; there are also constraints on the maximum and minimum generation output levels that each generator can achieve, and transmission network constraints that involve power flow equations and power flow limits on transmission lines; there are also constraints on the voltage limits on buses. The commitment statuses of generators are modeled as

binary decision variables in the UC problem, and the generation output levels and voltage levels are modeled as continuous variables. As we will discuss in details later, the power flow equations and network constraints involve nonlinear nonconvex functions of the continuous variables. Therefore, the unit commitment is a mixed-integer nonlinear nonconvex optimization problem. In practice, linear approximations of the nonconvex power flow constraints are widely used, which simplifies UC to a mixed-integer linear optimization problem.

The UC problem is solved a day before the real-time operation in order to accommodate long starting and shutting down times of generators. Once the UC problem is solved, the commitment decision for the next day is fixed. When real-time operation starts, the system operator dispatches the committed generators to meet realized demand at a minimum total generation cost. This procedure is called *economic dispatch*. The underlying optimization problem is called *OPF*, which has a similar set of constraints as in the UC problem, except that it only involves continuous decision variables, that is, generation output and voltages, and very often only has one decision period (therefore, the ramping constraints are not considered in this case). The OPF problem is solved every 5–15 min.

Owing to the nonlinear and nonconvex nature of the UC and OPF, it is still a challenge to obtain fast, robust solutions to these problems in large-scale power systems. New challenges are emerging too. For example, unconventional generators such as wind and solar power are intrinsically stochastic; how to deal with such uncertainties poses

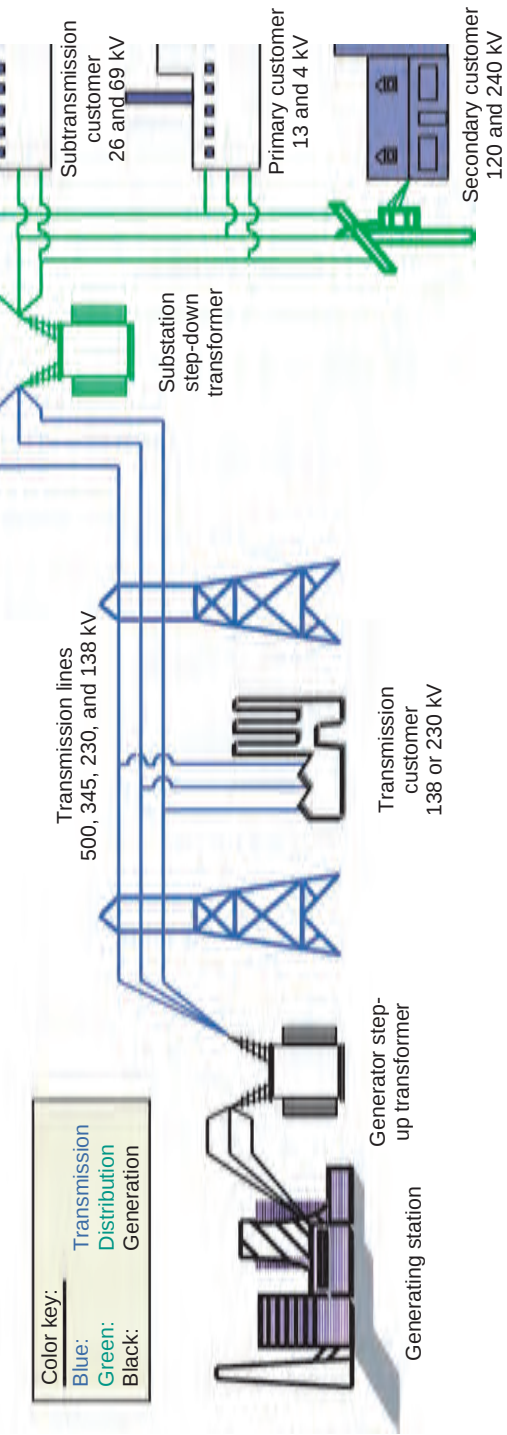


Figure 2. The basic structure of a typical electric power system consists of the generation, the transmission network, and the distribution system. *Source:* FERC [1].

challenges for large-scale integration of variable generation resources into the power grid. In this article, we introduce fundamental models and discuss some of the most recent advances in meeting these challenges.

The rest of this article is organized as follows. The section titled “Optimal Power Flow Problem” introduces power flow equations and OPF models. The section titled “Global Optimization Methods for AC OPF” introduces global optimization methods to exactly solve the OPF problem. Two lower bounding techniques, the Lagrangian relaxation and the semidefinite relaxation methods, are discussed in detail. The section titled “Distributed Algorithms for AC OPF” motivates decentralized optimization schemes for solving large-scale OPF problems and reviews the literature. The section titled “Security-Constrained Optimal Power Flow Problem” reviews the important problem of security-constrained OPF and solution methods. The UC and its optimization formulation are introduced in the section titled “Security-Constrained Unit Commitment Problem.” The recent development of the adaptive robust UC model is reviewed in the section titled “Adaptive Robust Optimization for SCUC under Uncertainty.”

OPTIMAL POWER FLOW PROBLEM

As introduced in the previous section, OPF is a nonconvex, nonlinear optimization problem. The most defining constraints of the problem are the ones involving alternating current (AC) power flow equations, which describe the relationship between generator power output, bus voltages, and power flows on the transmission lines. In this section, we first introduce two equivalent AC OPF formulations with AC power flow equations. Then, we discuss a linear approximation of the AC power flow equations, which leads to the so-called DC OPF problem.

AC OPF Models

In the AC model, voltage and power are expressed as complex numbers. In the rectangular coordinates, the voltage V_i at a bus

$i \in \mathcal{N}$ can be expressed as

$$V_i = e_i + jf_i,$$

where $j = \sqrt{-1}$ is the imaginary unit and e_i and f_i are the real and imaginary parts of the complex voltage, respectively. Similarly, the power injected into a bus i can be also decomposed into the real part, the active power P_i , the imaginary part, and the reactive power Q_i . In power flow equations, we are concerned with the *net* power injection at a bus i , which is the difference between the power injected from the generator at bus i expressed as $p_i^g + jq_i^g$ and the power consumed by the load $p_i^d + jq_i^d$; therefore, $P_i = p_i^g - p_i^d$, and $Q_i = q_i^g - q_i^d$.

The AC OPF problem of minimizing the total generation cost is given as

$$\min_{e, f, p^g, q^g} c(p^g) = \sum_{i \in \mathcal{G}} (c_{i0}(p_i^g)^2 + c_{i1}p_i^g + c_{i2}) \quad (1a)$$

$$\text{s.t. } p_i^g - p_i^d = \sum_{j \in \mathcal{N}} (G_{ij}(e_i e_j + f_i f_j) - B_{ij}(e_i f_j - e_j f_i)), \quad \forall i \in \mathcal{N} \quad (1b)$$

$$q_i^g - q_i^d = \sum_{j \in \mathcal{N}} (-G_{ij}(e_i f_j - e_j f_i) - B_{ij}(e_i e_j + f_i f_j)), \quad \forall i \in \mathcal{N} \quad (1c)$$

$$| -G_{ij}(e_i^2 + f_i^2 - e_i e_j - f_i f_j) | - B_{ij}(e_i f_j - e_j f_i) \leq p_{ij}^{\max}, \quad \forall (i, j) \in \mathcal{L} \quad (1d)$$

$$(v_i^{\min})^2 \leq e_i^2 + f_i^2 \leq (v_i^{\max})^2, \quad \forall i \in \mathcal{N} \quad (1e)$$

$$p_i^{\min} \leq p_i^g \leq p_i^{\max}, \quad \forall i \in \mathcal{G} \quad (1f)$$

$$q_i^{\min} \leq q_i^g \leq q_i^{\max}, \quad \forall i \in \mathcal{G}, \quad (1g)$$

where Equations (1b and 1c) are the AC power flow equations; G_{ij} and B_{ij} are defined as

$$G_{ij} = \begin{cases} -g_{ij} & \text{if } i \neq j \\ g_{ii} + \sum_{k \neq i} g_{ik} & \text{if } i = j \end{cases}$$

$$B_{ij} = \begin{cases} -b_{ij} & \text{if } i \neq j \\ b_{ii} + \sum_{k \neq i} b_{ik} & \text{if } i = j \end{cases}$$

As mentioned earlier, g_{ij} and b_{ij} are the conductance and susceptance of the transmission line (i, j) , respectively, and g_{ii} and

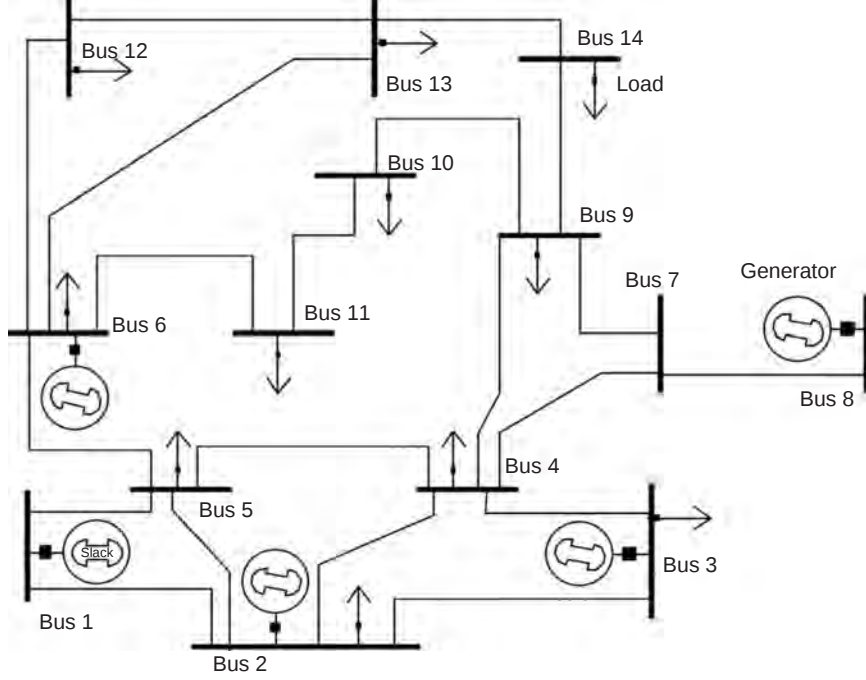


Figure 3. An example diagram of a power system with 14 buses; 20 transmission lines; 5 generators (the circle symbols) at buses 1, 2, 3, 6, and 8; and 11 loads (the arrows pointing out from buses) [2].

b_{ii} are the conductance and susceptance from the node i to the ground; constraint (Eq. 1d) ensures that the active power flow on each transmission line is within its limit $p_{ij}^{\max}$; constraint (Eq. 1e) imposes upper and lower bounds, $v_i^{\max}$ and $v_i^{\min}$, on the voltage magnitude at each bus; constraints (Eqs. 1f and 1g) set the upper and lower bounds on the active and reactive power output of generators, $p_i^{\max}, p_i^{\min}, q_i^{\max}, q_i^{\min}$, respectively. The objective (1a) models the generation cost as a quadratic function of the active power output of the generators, which is a common practice in the power industry. Sometimes, other objective functions such as the total power loss in transmission lines or other constraints such as apparent power flow limits are considered.

The above-mentioned AC OPF formulation (Eq. 1) is a nonconvex quadratic optimization model. Another equivalent formulation can be derived by expressing the complex bus voltages in the polar coordinates as

$$V_i = v_i e^{j\theta_i} = v_i (\cos \theta_i + j \sin \theta_i),$$

where v_i is the magnitude and θ_i is the phase angle of the complex voltage. In this representation, the AC OPF problem can be written as

$$\begin{aligned} \min_{v, \theta, p^g, q^g} \quad & \sum_{i \in \mathcal{G}} (c_{i0} (p_i^g)^2 + c_{i1} p_i^g + c_{i2}) \\ \text{s.t.} \quad & p_i^g - p_i^d = \sum_{j \in \mathcal{N}} G_{ij} v_i v_j \cos(\theta_i - \theta_j) \\ & + B_{ij} v_i v_j \sin(\theta_i - \theta_j), \quad i \in \mathcal{N} \quad (2a) \end{aligned}$$

$$\begin{aligned} q_i^g - q_i^d = \sum_{j \in \mathcal{N}} G_{ij} v_i v_j \sin(\theta_i - \theta_j) \\ - B_{ij} v_i v_j \cos(\theta_i - \theta_j), \quad i \in \mathcal{N} \quad (2b) \end{aligned}$$

$$\begin{aligned} & | -G_{ij} (v_i^2 - v_i v_j \cos(\theta_i - \theta_j)) \\ & + B_{ij} v_i v_j \sin(\theta_i - \theta_j) | \\ & \leq p_{ij}^{\max}, \forall (i, j) \in \mathcal{L} \quad (2c) \end{aligned}$$

$$v_i^{\min} \leq v_i \leq v_i^{\max}, \quad \forall i \in \mathcal{N} \quad (2d)$$

(1f), (1g)

The rectangular formulation (1) and polar formulation (2) are equivalent in the way that an optimal solution of one formulation can be easily transformed to an optimal solution of the other. From the optimization modeling and computation perspective, one advantage of the polar formulation (2) is that the voltage magnitude constraints (Eq. 2d) are simple linear bounding constraints, whereas in the rectangular representation model (Eq. 1e), the voltage magnitude constraints (Eq. 1e) are nonconvex quadratic constraints. Furthermore, sometimes the system operation requires the voltage magnitude at certain buses be fixed, which leads to a reduction of number of variables in the polar formulation, but an increase in the number of quadratic equality constraints in the rectangular formulation. However, there are also compelling reasons to use the rectangular formulation. First, Equation (1) is a quadratic optimization problem, whose Hessian matrix is constant; this is convenient for the application of interior-point methods. Second, the elimination of trigonometric functions also speeds up function evaluations. Third, the rectangular model is commonly used in practice because they facilitate the use of certain fast-decoupled solution methods for solving the power flow problem [3].

DC OPF Model

The AC OPF model involves nonconvex quadratic or trigonometric functions. It is still a computational challenge to solve large-scale AC OPF problems in a fast and robust way. In practice, the linear approximation model with linearized power flow equations has been widely used. It is usually called the *DC OPF model*, although it should not be confused with direct current. The DC power flow equations are obtained from the polar representation of power flow equations (Eqs 2a and 2b). The basic steps of the approximation are outlined here:

- We assume $G_{ij} = 0$ for all $(i, j) \in \mathcal{L}$, because the line resistance is usually very small.
- Under normal operating conditions, the voltage angle difference between two

connected buses is usually small, that is, $(\theta_i - \theta_j) \approx 0$, which implies $\sin(\theta_i - \theta_j) \approx (\theta_i - \theta_j)$ and $\cos(\theta_i - \theta_j) \approx 1$.

- Under normal operating conditions, the magnitudes of bus voltages are also close to each other, that is, if we normalize the bus voltages with respect to a common voltage unit, then the product of two bus voltages is again close to the unit 1.

Combining these three steps, the DC OPF model is given as:

$$\begin{aligned} & \min_{\mathbf{p}^g, \theta} \sum_{i \in \mathcal{G}} (c_{i2}(p_i^g)^2 + c_{i1}p_i^g + c_{i0}) \\ \text{s.t. } & \sum_{j \in \mathcal{N}} B_{ij}(\theta_i - \theta_j) = p_i^g - p_i^d, \quad \forall i \in \mathcal{N} \quad (3a) \\ & B_{ij}(\theta_i - \theta_j) \leq p_{ij}^{\max}, \quad \forall (i, j) \in \mathcal{L} \\ & p_i^{\min} \leq p_i^g \leq p_i^{\max}, \quad i \in \mathcal{G}, \quad (3b) \end{aligned}$$

where the AC power flow equations (Eqs 2a and 2b) are approximated by the linear equalities (Eq. 3a) and the power flow constraints (Eq. 2c) are approximated by linear inequalities (Eq. 3b).

Under normal operating conditions, the DC OPF model gives quite accurate approximation for the active power output of generators, and the solution methods for the convex DC OPF problem (Eq. 3) are much faster and robust. However, it does not account for reactive powers, and the approximation can lead to significant errors for stressed systems.

In the early 1960s, Carpentier originally introduced the AC OPF problem and made the first attempt to solve it [4]. Since then, many solution approaches have been studied, including successive linear/quadratic programming [5], trust-region-based methods [6], Lagrangian–Newton methods [7], and interior-point methods [8]. See [3] for a recent review. These approaches usually require a centralized implementation and aim to find a local solution satisfying first-order optimality conditions. Global optimization algorithms and distributed algorithms are also proposed in the literature. In this article, we focus on these two types of algorithms.

GLOBAL OPTIMIZATION METHODS FOR AC OPF

In this section, we introduce exact methods to solve the AC OPF problem (1) to global optimality. First, we discuss two methods to obtain lower bounds on the global optimal solution, namely the Lagrangian relaxation in the section titled “Lagrangian Relaxation” and the semidefinite programming relaxation in the section titled “Semidefinite Programming Relaxation.” Then, the global optimization methods for AC OPF, when these relaxations have a nonzero duality gap, are introduced in the section titled “Branch-and-Bound Algorithms.”

Lagrangian Relaxation

To present the Lagrangian relaxation method, we first formulate the Lagrangian dual function and the corresponding Lagrangian dual problem. Then, we show an efficient decomposition method for evaluating the dual function. This is essential in making the Lagrangian relaxation method practically useful.

The Lagrangian dual problem of the primal AC OPF (Eq. 1) is formed by dualizing the quadratic constraints (Eqs 1b–1e), while retaining simple bounding constraints (1f and 1g). The Lagrangian function is defined as

$$\begin{aligned} L(\mathbf{e}, \mathbf{f}, \mathbf{p}^g, \mathbf{q}^g, \lambda, \beta, \mu, \gamma, \eta) = & c(\mathbf{p}^g) + \sum_{i \in \mathcal{N}} \left(\lambda_i \left(\sum_{j \in \mathcal{N}} (e_i(G_{ij}e_j - B_{ij}f_j) + f_i(G_{ij}f_j + B_{ij}e_j)) + p_i^d \right) \right. \\ & + \beta_i \left(\sum_{j \in \mathcal{N}} (f_i(G_{ij}e_j - B_{ij}f_j) - e_i(G_{ij}f_j + B_{ij}e_j)) + q_i^d \right) + \sum_{(i,j) \in \mathcal{L}} \mu_{ij}(-(e_i^2 + f_i^2)G_{ij} \\ & + (e_ie_j + f_if_j)G_{ij} - (e_if_j - e_jf_i)B_{ij}) + \sum_{i \in \mathcal{N}} (\eta_i - \gamma_i)(e_i^2 + f_i^2) - \sum_{i \in \mathcal{G}} (\lambda_i p_i^g + \beta_i q_i^g). \end{aligned}$$

$$\begin{aligned} \min_{\mathbf{e}, \mathbf{f}, \mathbf{p}^g, \mathbf{q}^g} \quad & L(\mathbf{e}, \mathbf{f}, \mathbf{p}^g, \mathbf{q}^g, \lambda, \beta, \mu, \gamma, \eta) \\ - \sum_{(i,j) \in \mathcal{L}} \mu_{ij} p_{ij}^{\max} + \sum_{i \in \mathcal{N}} (\gamma_i (v_i^{\min})^2 - \eta_i (v_i^{\max})^2) \end{aligned} \quad (4a)$$

$$\text{s.t. } (1f), (1g) \quad (4b)$$

$$\sum_{i \in \mathcal{N}} (v_i^{\min})^2 \leq \sum_{i \in \mathcal{N}} (e_i^2 + f_i^2) \leq \sum_{i \in \mathcal{N}} (v_i^{\max})^2. \quad (4c)$$

Notice that the sphere constraints (Eq. 4c) are redundant because of the bounding constraints (Eq. 1e) on nodal voltages. However, the inclusion of Equation (4c) will help solve Equation (4) analytically, as shown in the following discussion.

It is known that for any feasible value of the Lagrange multipliers $\{(\lambda, \beta, \mu, \gamma, \eta) : \mu \geq 0, \gamma \geq 0, \eta \geq 0\}$, the value of the dual objective function, $g(\lambda, \beta, \mu, \gamma, \eta)$, provides a lower bound on the global optimal objective function value of Equation (1). Therefore, to

obtain the best lower bound for all feasible multipliers, we need to solve the following Lagrangian dual problem:

$$\begin{aligned} \max_{\lambda, \beta, \mu, \gamma, \eta} \quad & g(\lambda, \beta, \mu, \gamma, \eta) \\ \text{s.t. } \quad & \mu \geq 0, \\ & \gamma \geq 0, \eta \geq 0. \end{aligned} \quad (5)$$

This is a concave maximization problem, so an optimal solution can be obtained by convex optimization techniques. However, the practical use of the Lagrangian duality method greatly depends on how efficiently the function value and a subgradient of $g(\lambda, \beta, \mu, \gamma, \eta)$ can be evaluated, which is related to how efficiently the inner nonconvex problem (4) can be solved. It is shown in the following equations that Equation (4) is tractable because of its special structure. In particular, problem (4) can be decomposed into two subproblems: a convex separable quadratic optimization problem with simple box constraints

$$\begin{aligned}
\min_{\mathbf{p}^g, \mathbf{q}^g} \quad & c(\mathbf{p}^g) - \sum_{i \in \mathcal{G}} (\lambda_i p_i^g + \beta_i q_i^g) \\
\text{s.t.} \quad & p_i^{\min} \leq p_i^g \leq p_i^{\max}, \quad q_i^{\min} \leq q_i^g \leq q_i^{\max}, \\
& \forall i \in \mathcal{G}
\end{aligned} \tag{6}$$

and a nonconvex quadratic optimization problem with two sphere constraints

$$\begin{aligned}
\min_{\mathbf{x}} \quad & \mathbf{x}^T \mathbf{A}(\lambda, \beta, \mu, \gamma, \eta) \mathbf{x} \\
\text{s.t.} \quad & \sum_{i \in \mathcal{N}} (v_i^{\min})^2 \leq \mathbf{x}^T \mathbf{x} \leq \sum_{i \in \mathcal{N}} (v_i^{\max})^2, \tag{7}
\end{aligned}$$

where $\mathbf{x} = [\mathbf{e}^T \mathbf{f}^T]^T$ and the matrix $\mathbf{A}(\lambda, \beta, \mu, \gamma, \eta)$ are given as

$$\begin{aligned}
\mathbf{A}(\lambda, \beta, \mu, \gamma, \eta) = & \left(\begin{bmatrix} \lambda_1 \mathbf{G}_1; -\beta_1 \mathbf{B}_1; \\ \vdots \\ \lambda_{|\mathcal{M}|} \mathbf{G}_{|\mathcal{M}|}; -\beta_{|\mathcal{M}|} \mathbf{B}_{|\mathcal{M}|}; \\ \beta_1 \mathbf{G}_1; +\lambda_1 \mathbf{B}_1; \\ \vdots \\ \beta_{|\mathcal{M}|} \mathbf{G}_{|\mathcal{M}|}; +\lambda_{|\mathcal{M}|} \mathbf{B}_{|\mathcal{M}|}; \end{bmatrix} \begin{bmatrix} -\beta_1 \mathbf{G}_1; -\lambda_1 \mathbf{B}_1; \\ \vdots \\ -\beta_{|\mathcal{M}|} \mathbf{G}_{|\mathcal{M}|}; -\lambda_{|\mathcal{M}|} \mathbf{B}_{|\mathcal{M}|}; \\ \lambda_1 \mathbf{G}_1; -\beta_1 \mathbf{B}_1; \\ \vdots \\ \lambda_{|\mathcal{M}|} \mathbf{G}_{|\mathcal{M}|}; -\beta_{|\mathcal{M}|} \mathbf{B}_{|\mathcal{M}|}; \end{bmatrix} \right) + \begin{pmatrix} \mathbf{G}(\mu), & -\mu \circ \mathbf{B} \\ \mu \circ \mathbf{B}, & \mathbf{G}(\mu) \end{pmatrix} + \text{diag}([\eta - \gamma; \eta - \gamma]),
\end{aligned}$$

where $\mathbf{G}_i$ and $\mathbf{B}_i$ are the i th rows of matrices $\mathbf{G}$ and $\mathbf{B}$, respectively; $\text{diag}(\mathbf{x})$ is a diagonal matrix with $\mathbf{x}$ on the diagonal; “ $\circ$ ” denotes the componentwise product; and $\mathbf{G}(\mu)$ has the ij th component as $\mu_{ij} G_{ij}$ for off-diagonal element and $\sum_{j \neq i} -\mu_{ij} G_{ij}$ for the i th diagonal element.

Owing to the special structure, a global optimal solution $\mathbf{x}^*$ of Equation (7) has the following closed form

$$\mathbf{x}^* = \begin{cases} \mathbf{v} \sqrt{\frac{\sum_{i \in \mathcal{N}} (v_i^{\max})^2}{\|\mathbf{v}\|}}, & \text{if } \lambda_{\min} < 0 \\ \mathbf{v} \sqrt{\frac{\sum_{i \in \mathcal{N}} (v_i^{\min})^2}{\|\mathbf{v}\|}}, & \text{otherwise,} \end{cases}$$

where $\mathbf{v}$ is the eigenvector associated with the smallest eigenvalue $\lambda_{\min}$ of the matrix $\mathbf{A} + \mathbf{A}^T$. Therefore, solving Equation (7) amounts to computing the extreme eigenvector of the matrix. The conductance $\mathbf{G}$ and susceptance $\mathbf{B}$ matrices of real-world power systems are often sparse with graph density less than 0.1, thus $\mathbf{A} + \mathbf{A}^T$ is also sparse. Hence, sparse techniques can be applied to further reduce the computational burden.

The objective function value and a subgradient of $g(\lambda, \beta, \mu, \gamma, \eta)$ can be computed from the solutions of Equations (6 and 7). Thus, the dual problem (5) as a convex non-differentiable optimization problem can be

efficiently solved by a subgradient projection method, which gives a lower bound on the global optimal function value. One can use the optimal solution found in the previous period as a warm start for the next problem. Furthermore, it is observed that the Lagrangian multiplier from the primal problem is often a good initial guess for the dual problem [9].

Semidefinite Programming Relaxation

In this section, we briefly discuss another relaxation technique to find a lower bound for the AC OPF problem, which is first proposed in Ref. 10. The observation is that as the AC OPF in the rectangular coordinate is a quadratically constrained quadratic program (QCQP), and it is well known that any QCQP can be relaxed to a semidefinite programming (SDP) problem, then the AC OPF problem can also be relaxed to an SDP problem, which is convex and thus solvable in polynomial time by interior-point methods.

The basic procedure of formulating the SDP relaxation can be described as follows. First, the AC OPF problem, as a nonconvex QCQP problem, can be reformulated as an equivalent optimization problem involving matrix variables, where the matrix variables are constrained to be positive semidefinite and have rank one. Besides the rank constraint, which is a nonconvex constraint and

difficult to deal with, all other constraints of the reformulated optimization problem are linear in the matrix variables. Therefore, by removing the rank constraint, we come to a linear SDP problem, given in the following equation

$$\min_{W, \alpha} \sum_{i \in \mathcal{G}} \alpha_i \quad (8a)$$

$$\text{s.t.} \quad \begin{cases} c_{i1}(\text{trace } \mathbf{Y}_i W + p_i^d) + c_{i2} - \alpha_i \\ \sqrt{c_{i0}}(\text{trace } \mathbf{Y}_i W + p_i^d) \\ \sqrt{c_{i0}}(\text{trace } \mathbf{Y}_i W + p_i^d) \\ -1 \end{cases} \pi 0, \quad \forall i \in \mathcal{G} \quad (8b)$$

$$p_i^{\min} - p_i^d \leq \text{trace } \mathbf{Y}_i W \leq p_i^{\max} - p_i^d, \quad \forall i \in \mathcal{N} \quad (8c)$$

$$q_i^{\min} - q_i^d \leq \text{trace } \bar{\mathbf{Y}}_i W \leq q_i^{\max} - q_i^d, \quad \forall i \in \mathcal{N} \quad (8d)$$

$$\text{trace } \mathbf{Y}_{ij} W \leq p_{ij}^{\max}, \quad \forall (i, j) \in \mathcal{L} \quad (8e)$$

$$(v_i^{\min})^2 \leq \text{trace } \mathbf{M}_i W \leq (v_i^{\max})^2, \quad \forall i \in \mathcal{N} \quad (8f)$$

$$W \geq 0. \quad (8g)$$

The parameters $\mathbf{Y}_i$, $\bar{\mathbf{Y}}_i$, and $\mathbf{Y}_{ij}$ are matrices formed from conductance and susceptance matrices of the electric network. $\mathbf{M}_i$ is a square matrix of size $2|\mathcal{N}|$ with a 1 in the i th and $(i + |\mathcal{N}|)$ th diagonal entries and 0's elsewhere [10], and W is the relaxation of matrix $(\mathbf{e}^T; \mathbf{f}^T)(\mathbf{e}^T; \mathbf{f}^T)^T$.

The dual problem of the above-mentioned SDP is again an SDP, and there is no duality gap between the two. Furthermore, solving the dual SDP has certain advantages over solving the primal problem (8). For example, the number of decision variables in the dual SDP is essentially linear with respect to the number of buses $|\mathcal{N}|$, as opposed to being quadratic in Equation (8). In addition, it is easier to work with the dual SDP to detect whether the solution of the relaxation problem recovers a global optimal solution [10].

Branch-and-Bound Algorithms

Numerical experiments on IEEE test systems show that the Lagrangian relaxation and the SDP relaxation work surprisingly well

to recover the global optimization solutions for the AC OPF problems [9, 10]. However, in general, there is no theoretical guarantee that these relaxation methods find a global optimum. Indeed, it has been shown that the relaxations can give a strictly lower bound to the global optimal objective value on some simple electric networks [11]. This raises the need to develop algorithms that guarantee to find a global optimum for the AC OPF problems. It is still a relatively less explored field. The recent work in Refs 9 and 11 proposes such algorithms based on the branch-and-bound principle. We note that the SDP and the Lagrangian dual relaxations give the same lower bounds, but they could differ in computational performance.

In the following, we briefly describe the general branching and pruning procedures of these algorithms. First, the branch-and-bound algorithm starts from a root node that corresponds to a set of the decision variables that contains the feasible region of the original AC OPF problem (1), denoted as Ω . For example, the following set would suffice to be the root node

$$\begin{aligned} \mathcal{B}_0 = \{(\mathbf{p}^g, \mathbf{q}^g, \mathbf{e}, \mathbf{f}) : \mathbf{p}^{\min} \leq \mathbf{p}^g \leq \mathbf{p}^{\max}, \\ \mathbf{q}^{\min} \leq \mathbf{q}^g \leq \mathbf{q}^{\max}, (v_i^{\min})^2 \leq e_i^2 + f_i^2 \leq (v_i^{\max})^2, \\ \forall i \in \mathcal{N}\}. \end{aligned}$$

Any subsequent node in the search tree corresponds to a subset of $\mathcal{B}_0$, generated by branching from its parent node. The branching corresponds to a partition of the search set associated with the parent node, which can be implemented by either a rectangular bisection over the upper and lower bounds on the real and reactive power, $\mathbf{p}^g$ and $\mathbf{q}^g$ as proposed in Ref. 9, or a radial bisection over the voltage magnitude as proposed in Ref. 11.

A search node is called *active*, if it has not been pruned. At an active node of iteration k , denoted as $\mathcal{B}_k$, the algorithm solves the Lagrangian or SDP relaxation of a modified AC OPF problem with a feasible region defined as $\Omega \cap \mathcal{B}_k$. This feasible region can be easily constructed from the original AC OPF by appending the constraints in $\mathcal{B}_k$, which are bounds on $\mathbf{p}^g, \mathbf{q}^g$, and voltage magnitudes. Recall that such bounding constraints

are amiable for the decomposition described in the section titled “Lagrangian Relaxation” for the Lagrangian relaxation. Furthermore, both the rectangular partition and the radial partition preserve this bounding structure of the search sets. Therefore, the Lagrangian relaxation can be efficiently solved, and the SDP relaxation can be formulated and solved just as for the original AC OPF.

The obtained lower bound is compared to the smallest objective function value of the incumbent solution, which is based on the best feasible solution found so far in the branch-and-bound process and also provides an upper bound on the global optimum. The current search node is pruned if the obtained lower bound is higher than the best objective value of the incumbent solution or the lower bound is $+\infty$, which means the modified AC OPF is infeasible; otherwise, the incumbent is replaced by the obtained lower bound and the current node remains active.

The feasible region for $\mathbf{p}^g$ shrinks at every step of the branch-and-bound process; therefore, a nested sequence of bisections shrinks to a singleton. Because the objective function is a function of only $\mathbf{p}^g$, for any sequence of infinitely decreasing partition elements, the upper and lower bounds converge in the limit. Thus, the bounding scheme is consistent. The selection rule for choosing partition elements also improves bounds. Hence, the branch-and-bound algorithm is convergent [9].

DISTRIBUTED ALGORITHMS FOR AC OPF

Previous sections introduce models and solution methods for the AC OPF problems, which require a central control entity to gather system-wide information and to implement the solution algorithms. This centralized operation scheme is widely used in today’s power industry.

However, in some situations, decentralized decision-making is more appealing or even necessary. For instance, in a large power system involving multiple interconnected control areas, global information pooling and sharing may be limited by institutional arrangements. Owing to this, a centralized OPF algorithm for the entire interconnection may not be properly implementable. Instead,

a decentralized operation scheme can be utilized to achieve a coordinated dispatch, which often improves the economics of the entire system comparing to individual dispatch without coordination.

A decentralized operation scheme usually relies on an iterative process of computation and communication. Within an iteration, each subsystem operator solves a modified OPF problem for its own control area, which involves parameters that will be updated from the solutions of its adjacent subsystems. The parameter update is accomplished by communication between subsystems. Baldick and his colleagues in a series of papers first demonstrated the viability of the decentralized scheme for coordination of subcontrol areas in a large power system [12, 13]. Since then, several papers have contributed to the topic. In the following, we summarize the existing approaches based on the OPF models, decomposition methods, and algorithmic frameworks.

- **OPF Models.** Both AC and DC OPF problems have been studied. In particular, [12–15] have proposed distributed algorithms for AC OPF problems, whereas [16, 17] have proposed for DC OPF problems. In [17], the authors use a nonlinear DC model to account for network losses.
- **Decomposition Methods.** A large power system is usually divided into subcontrol regions interconnected by tie lines. The couplings between subsystems arise from constraints such as power flow balance equations and thermal limits on transmission lines. In order to apply a distributed algorithm, the coupling between subsystems need to be broken down. A common idea behind the proposed decomposition methods is to introduce redundant variables in the overlapping regions between subsystems. These redundant variables can be power injections and voltages on fictitious buses introduced on tie lines [12, 13, 15, 17]; or, redundant power flow variables on tie lines [14, 16].

• **Algorithmic Framework.** After redundant variables are introduced, the centralized OPF model is decomposed into subproblems. A common approach is to relax coupling constraints using Lagrangian relaxation or some of its variants. In particular, the existing proposals can be divided into two groups. The first group uses the exact framework of Lagrangian or augmented Lagrangian relaxation, where the coupling constraints are relaxed by introducing Lagrange multipliers, the resulting Lagrangian problem is solved by a distributed algorithm, and the Lagrange multipliers (the shadow prices) are updated by a subgradient scheme. For example, the auxiliary problem principle (APP) is used to solve the Lagrangian problem in Ref. 12. The alternating direction method and the predictor–corrector proximal method are applied and compared to the APP method in Ref. 13. These three methods have comparable performance in terms of number of iterations and computation time. The other group of methods uses a variant of Lagrangian relaxation framework, where the KKT system of the subproblem is solved approximately in each iteration [15, 16]. The key observation is that stacking together the KKT conditions of subsystems gives the KKT condition of the entire system. Therefore, if the distributed algorithm converges, the solution thus satisfies the overall KKT condition and is at least a stationary point of the centralized problem. It is worth mentioning that the distributed algorithm frameworks such as the APP or the alternating direction methods are originally proposed for convex optimization problems. The convergence is not guaranteed for general nonconvex problems such as the AC OPF problems.

SECURITY-CONSTRAINED OPTIMAL POWER FLOW PROBLEM

Contingency analysis is routinely performed in the process of economic dispatch. The

goal is to ensure that demand can be satisfied both in the normal case of operation and in the case of contingencies. By contingency, we mean the situation where any one or more components in the power system, such as generators, transmission lines, transformers, or other equipments, experience unexpected failure. The OPF problem with contingency constraints is often referred to as the *security-constrained optimal power flow (SCOPF)*. There are two major types of SCOPF models: the preventive model [18] and the corrective model [19].

The preventive model finds a minimum cost normal case dispatch solution that is also feasible for all prespecified contingency conditions. A general formulation is given in the following equation:

$$\begin{aligned} \min_{\mathbf{x}_0, \dots, \mathbf{x}_C, \mathbf{u}_0} \quad & c(\mathbf{x}_0, \mathbf{u}_0) \\ \text{s.t.} \quad & \mathbf{g}_k(\mathbf{x}_k, \mathbf{u}_0) = 0, \quad k = 0, \dots, C \\ & \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_0) \leq 0, \quad k = 0, \dots, C, \end{aligned}$$

where $\mathbf{x}_0$ and $\mathbf{x}_k$ represent the state variables in the normal case and the k th contingency, respectively. The state variables include nodal voltage magnitudes and phase angles. $\mathbf{u}_0$ represents the control variable in the normal case, such as the active and reactive powers. As stated earlier, the preventive model requires the normal case control $\mathbf{u}_0$ to be feasible for all C contingencies.

The second type of the SCOPF model is the so-called corrective model, which allows the system operator to readjust control variables after the contingency occurs. The rationale is that the power system components, such as transmission lines and transformers, can usually sustain a short period of overloading without being damaged, which gives the system operator a time window to adjust control variables to eliminate any violations caused by the contingency [19]. The corrective model can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{x}_0, \dots, \mathbf{x}_C, \mathbf{u}_0, \dots, \mathbf{u}_C} \quad & c(\mathbf{x}_0, \mathbf{u}_0) \\ \text{s.t.} \quad & \mathbf{g}_k(\mathbf{x}_k, \mathbf{u}_k) = 0, \quad k = 0, \dots, C \\ & \mathbf{h}_k(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad k = 0, \dots, C \\ & |\mathbf{u}_k - \mathbf{u}_0| \leq \bar{\mathbf{u}}_k^{\max}, \quad k = 1, \dots, C. \end{aligned}$$

Here, the system operator has the flexibility to choose a potentially different control variable $\mathbf{u}_k$ for each contingency k . The last constraint imposes the maximum deviation between the normal case control and the postcontingency control. Note that these constraints should be understood component-wise as $|u_{k,i} - u_{0,i}| \leq \bar{u}_{k,i}^{\max}$ for every generator $i \in \mathcal{G}$.

One of the major challenges of SCOPF is its huge dimensionality. For C contingency scenarios, the problem size of the SCOPF problem is roughly $C + 1$ times larger than that of the normal OPF problem. For large-scale power systems involving numerous contingencies, centralized solution algorithms may encounter prohibitive memory usage and long execution times. In the literature, solution approaches include iterative contingency selection schemes, decomposition methods, and network compression methods. It is well known that many postcontingency constraints are redundant, contingency filtering techniques identify and include only those most critical contingencies into the formulation. Benders decomposition algorithms decompose the SCOPF into a master problem and subproblems, where subproblems check the solution feasibility and possibly generate a linear cut for the master problem. The network compression technique aims to reduce the size of the problem based on the observation that the impact of a contingency is in general confined to a localized area of the power grid. See [20] for a recent review on these techniques.

SECURITY-CONSTRAINED UNIT COMMITMENT PROBLEM

Previous sections introduce the fundamental OPF problems. For these OPF problems, the commitment status of generators is assumed to be known and fixed over the dispatch period. The other fundamental problem in power systems operation is the UC problem, where the commitment status of generators is the critical decision to be made, together with the dispatch decision for generation power output.

The UC problem is usually solved in a day-ahead time framework for the following

24 h of the day and thus involves a multiperiod decision-making, whereas the OPF-based economic dispatch is usually solved in real time for one single period (for example, every 5 min in the ISO New England system). Thus, in addition to all the constraints in the economic dispatch problem, the multiperiod unit commitment also requires intertemporal constraints such as constraints on ramping capabilities of generators and constraints on the minimum time that a generator has to remain on (off) after a starting up (a shutting down).

In the process of UC, it is also important to account for various uncertainties such as forecast error for load and renewable energy supply and unexpected outages of generators and transmission lines. To deal with unexpected outages, the security constraints as described in the section titled “Security-Constrained Optimal Power Flow Problem” are incorporated in the UC problem. To account for the load and supply uncertainty, a widely adopted approach is to keep a certain amount of generation capacity available for fast response to sudden unexpected changes in load and supply. This amount of standby capacity is called *reserves*, which are further classified into different categories based on the speed of its response, such as ten-minute spinning reserve (TMSR), ten-minute nonspinning reserve (TMNSR), and thirty-minute operating reserve (TMOR).

Incorporating all of the above-mentioned modeling considerations, the security-constrained unit commitment (SCUC) problem is formulated as the following mixed-integer optimization problem [21, 22]:

$$\min_{\mathbf{x}, \mathbf{u}, \mathbf{v}, \mathbf{p}, \mathbf{r}} \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{G}} (f_i^t(x_i^t, u_i^t, v_i^t) + c_i^t(p_i^t)) \quad (9a)$$

$$\text{s.t. } x_i^{t-1} - x_i^t + u_i^t \geq 0, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \quad (9b)$$

$$x_i^t - x_i^{t-1} + v_i^t \geq 0, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \quad (9c)$$

$$\begin{aligned} x_i^t - x_i^{t-1} &\leq x_i^{\tau}, \\ \forall \tau &\in [t + 1, \min\{t + \text{MinUp}_i - 1, T\}], \\ t &\in [2, T], i \in \mathcal{G}, \end{aligned} \quad (9d)$$

$$\begin{aligned}
x_i^{t-1} - x_i^t &\leq 1 - x_i^t, \\
\forall \tau \in [t+1, \min\{t + \text{MinDw}_i - 1, T\}], t \in [2, T], \\
i \in \mathcal{G}
\end{aligned} \quad (9e)$$

$$\sum_{i \in \mathcal{G}} p_i^t = \sum_{j \in \mathcal{D}} \bar{d}_j^t, \quad \forall t \in \mathcal{T}, \quad (9f)$$

$$\begin{aligned}
-RD_i x_i^t - SD_i v_i^t &\leq p_i^t - p_i^{t-1} \leq RU_i x_i^{t-1} \\
&+ SU_i u_i^t, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \quad (9g)
\end{aligned}$$

$$\begin{aligned}
-f_{l,k}^{\max} &\leq \mathbf{a}_{l,k}^T (\mathbf{p}^t - \mathbf{d}^t) \leq f_{l,k}^{\max}, \\
\forall t \in \mathcal{T}, l \in \mathcal{C}_k, k \in \mathcal{L}, \quad (9h)
\end{aligned}$$

$$p_i^{\min} \leq p_i^t + \sum_{a \in \mathcal{A}} r_{i,a}^t \leq p_i^{\max}, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \quad (9i)$$

$$p_i^{\min} x_i^t \leq p_i^t \leq p_i^{\max} x_i^t, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, \quad (9j)$$

$$\sum_{i \in \mathcal{G}, a \in \mathcal{A}_k} r_{i,a}^t \geq \bar{r}_k^t, \quad \forall t \in \mathcal{T}, k \in \mathcal{N}_r, \quad (9k)$$

$$\sum_{a \in \mathcal{A}_k} r_{i,a}^t \leq \bar{r}_{i,k}^t, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}, k \in \mathcal{N}_r, \quad (9l)$$

$$x_i^t, u_i^t, v_i^t \in \{0, 1\}, \quad \forall i \in \mathcal{G}, t \in \mathcal{T}. \quad (9m)$$

The decision variables include binary variables x_i^t, u_i^t, v_i^t for commitment decisions and continuous variables p_i^t and $r_{i,a}^t$ for generation and reserve decisions, where $x_i^t = 1$ if generator i is producing electricity, that is, in the on state, at time t , and $x_i^t = 0$ otherwise; $u_i^t = 1$ if generator i is turned on from the off state at time t ; $v_i^t = 1$ if generator i is turned off at time t ; p_i^t is the amount of electricity produced by generator i at time t ; and $r_{i,a}^t$ is the amount of reserve provided by generator i for reserve type a at time t .

The parameters in the above-mentioned model are explained later. The function $f_i^t(x_i^t, u_i^t, v_i^t)$ is the total fixed cost of generator i at time t in the horizon $\mathcal{T}$, including startup cost, shutdown cost, and other fixed costs. $c_i^t(p_i^t)$ is the operating cost, which is usually

approximated as a convex piece-wise linear function of the active power output p_i^t . $\bar{d}_j^t$ is the forecast nodal load at bus j , time t . RU_i and RD_i are, respectively, the ramp-up and ramp-down rates of generator i , and SU_i and SD_i are the ramp-up and ramp-down rates when generator i is starting up and shutting down, respectively.

$\mathbf{a}_{l,k} \in \mathbb{R}^{|\mathcal{V}|}$ is the shift factor vector of line l in the k th contingency where line k is offline. The i -th component of $\mathbf{a}_{l,k}$ is defined as the change of power flow on line l if a unit of power is injected at bus i and taken out at the reference bus. $\mathcal{C}_k$ is the set of remaining lines when line k is tripped. $\bar{r}_k^t$ is the system level reserve requirement for reserve type k at time t . $\mathcal{N}_r$ is the set of all reserve types such as TMSR, TMNSR, and TMOR. $\bar{r}_{i,a}^t$ is the maximum amount of reserve type a that generator i can provide at time t . $\mathcal{A}_k$ is the set of reserve types that generator k can provide.

Constraints (Eqs 9b and 9c) represent logic relations between on and off status and the turn-on and turn-off actions. Constraints (Eqs 9d and 9e) restrict the minimum up and down times for each generator. Constraint (Eq. 9f) enforces energy balance in each time period. Constraint (Eq. 9g) is the ramping constraint, which enforces the ramp-up and ramp-down limits of each generator. The ramping constraint says that when the generator is already on at time period t , the ramp-up and ramp-down rates at that time cannot exceed RU_i and RD_i , respectively; when the generator is being turned on or turned off at time period t , the ramping rates cannot exceed another set of limits SU_i and SD_i . Constraint (Eq. 9h) is the linearized power flow equations using shift factors for the i th contingency, which means the i th transmission line is tripped off, and $i = 0$ represents the normal case.

Constraint (Eq. 9i) limits the total production and reserve for each generator. Constraint (Eq. 9j) limits the production level of each generator. Constraint (Eq. 9k) is the system reserve requirement for reserve category k at time t . Constraint (Eq. 9l) provides an upper bound on generator i 's reserve capacity.

ADAPTIVE ROBUST OPTIMIZATION FOR SCUC UNDER UNCERTAINTY

As discussed in the section titled “Security-Constrained Unit Commitment Problem,” there are various uncertainties in the daily operation of power systems, among which an increasingly significant source of uncertainty is from the generation side. In particular, renewable energy resources using wind and solar power are intrinsically intermittent and uncertain. The forecast of these renewable resources is much more challenging than the forecast of traditional load. Furthermore, the penetration level of renewable energy in the power grid is increasing rapidly in many major power systems in the world, which makes the associated supply uncertainty an emerging threat to reliable system operations.

The SCUC model introduced in the section titled “Security-Constrained Unit Commitment Problem” uses the reserve approach to deal with uncertainties. This approach relies on predetermined reserve requirements, which do not reflect the fast changing uncertainty in the operational environment. A popular alternative is the stochastic-programming-based unit commitment model. Stochastic UC models explicitly consider scenarios of uncertainties and generate a UC decision with a minimum expectation cost over the considered scenarios [23]. A significant challenge of the stochastic programming approach is the representation of uncertainty, in particular, the difficulty of scenarios selection in a large power system; the other challenge, perhaps even more formidable, is the computational burden of the resulting large-scale mixed-integer optimization problem.

In this section, we introduce a new direction of study on UC under severe uncertainty using the robust optimization-based approach. Robust optimization has recently emerged as a rich and powerful paradigm for decision-making under uncertainty [24]. The basic idea of robust optimization is based on two concepts: uncertainty sets and robust constraints. An uncertainty set is a deterministic set that describes the uncertain parameters. The simplest example is an interval of

the uncertainty, for instance wind power output of a wind farm. This interval describes all possible values of the wind power. As wind power is uncertain, any constraint in the original deterministic UC model (9) that involves wind power needs special attention. In particular, we want a robust version of these constraints, which enforces that the UC and dispatch decisions satisfy these constraints for any realization of wind power in the uncertainty set. This is the so-called robust constraints. A UC model with uncertainty sets and robust constraints is a robust UC model.

There is much more to the above-mentioned simplistic setting of robust UC models. In the section titled “Adaptive Robust UC Model,” we introduce an adaptive robust optimization-based UC model, which has more sophisticated optimization structure and more realistic uncertainty sets. Then, we present an efficient solution method for this model in the section titled “Solution Method to Solve Adaptive Robust Model.”

Adaptive Robust UC Model

In this section, we consider the net nodal load, which combines the traditional load and wind power output as negative load. The net nodal load, therefore, is highly uncertain. The interval uncertainty set is a very simplistic example of uncertainty sets. Consider the following uncertainty set of net load at each time period t :

$$\begin{aligned} \mathcal{D}^t(\bar{\mathbf{d}}_i, \hat{\mathbf{d}}_i^t, \Delta^t) &:= \{\mathbf{d}^t \in \mathbb{R}^{N_d} : \\ &\sum_{i \in \mathcal{N}_d} \frac{|d_i^t - \bar{d}_i^t|}{\hat{d}_i^t} \leq \Delta^t, \\ &d_i^t \in [\bar{d}_i^t - \hat{d}_i^t, \bar{d}_i^t + \hat{d}_i^t], \forall i \in \mathcal{N}_d\}, \quad (10) \end{aligned}$$

where $\mathcal{N}_d$ is the set of nodes that have uncertain injections, N_d is the number of such nodes, $\mathbf{d}^t = (d_i^t, i \in \mathcal{N}_d)$ is the vector of uncertain net injections at time t , $\hat{d}_i^t$ is the nominal value of the net injection of node i at time t , $\bar{d}_i^t$ is the deviation from the nominal net injection value of node i at time t , the interval $[\bar{d}_i^t - \hat{d}_i^t, \bar{d}_i^t + \hat{d}_i^t]$ is the range of the uncertain d_i^t , and the inequality in Equation (10) controls

the total deviation of all injections from their nominal values at time t . The parameter Δ^t is the “budget of uncertainty,” taking values between 0 and N_d . When $\Delta^t = 0$, the uncertainty set $\mathcal{D}^t = \{\bar{\mathbf{d}}\}$ is a singleton, corresponding to the nominal deterministic case. As Δ^t increases, the size of the uncertainty set $\mathcal{D}^t$ enlarges. This means that larger total deviation from the expected net injection is considered, so that the resulting robust UC solutions are more conservative and the system is protected against a higher degree of uncertainty. When $\Delta^t = N_d$, $\mathcal{D}^t$ equals to the entire hypercube defined by the intervals for each d_i^t for $i \in \mathcal{N}_d$.

Now, we formulate the adaptive robust UC model as follows:

$$\begin{aligned} \min_{\mathbf{x}} \quad & (\mathbf{c}^T \mathbf{x} + \max_{\mathbf{d} \in \mathcal{D}} \min_{\mathbf{y} \in \Omega(\mathbf{x}, \mathbf{d})} \mathbf{b}^T \mathbf{y}) \\ \text{s.t.} \quad & \mathbf{F}\mathbf{x} \leq \mathbf{f}, \mathbf{x} \text{ binary}, \end{aligned} \quad (11)$$

where $\mathbf{x}$ is the vector of commitment-related decisions, $\mathbf{y}$ is the vector of dispatch-related variables, $\mathcal{D}$ is the Cartesian product of nodal load uncertainty sets over time, $\Omega(\mathbf{x}, \mathbf{d}) = \{\mathbf{y} : \mathbf{H}\mathbf{y} \leq \mathbf{h}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \leq \mathbf{g}, \mathbf{I}_u \mathbf{y} = \mathbf{d}\}$ is the set of feasible dispatch solutions for a fixed commitment decision $\mathbf{x}$ and nodal net injection realization $\mathbf{d}$, the constraints $\mathbf{H}\mathbf{y} \leq \mathbf{h}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \leq \mathbf{g}, \mathbf{I}_u \mathbf{y} = \mathbf{d}$ are a matrix representation of constraints (Eqs 9f–9l), and constraints $\mathbf{F}\mathbf{x} \leq \mathbf{f}$ is a matrix representation of commitment constraints (Eqs 9b–9e).

From this formulation, we can see that the commitment decision $\mathbf{x}$ considers all possible future net load represented in the uncertainty set $\mathcal{D}$. Such a commitment solution remains feasible, thus *robust*, to any realization of the uncertain net load. In comparison, the traditional UC solution only guarantees feasibility for a single nominal net load, whereas the stochastic optimization UC solution only considers a finite set of preselected scenarios of the uncertain net load. Furthermore, in this formulation, the optimal second-stage decision $\mathbf{y}$ is a function of the uncertain net load $\mathbf{d}$, where the functional form $\mathbf{y}(\mathbf{d})$ is implicitly determined by the economic dispatch problem $\min_{\mathbf{y} \in \Omega(\mathbf{x}, \mathbf{d})} \mathbf{b}^T \mathbf{y}$. The dispatch solution is thus *fully adaptive* to any realization of the uncertainty.

Solution Method to Solve Adaptive Robust Model

This section outlines an efficient algorithm to solve the adaptive robust UC model (Eq. 11) proposed in Ref. 25; similar algorithms are also proposed in Refs 26 and 27. Observe that the adaptive robust UC model (11) is a two-stage problem, where the first stage finds the optimal commitment decision and the second stage finds the worst-case dispatch cost under a fixed commitment solution. The proposed solution method is a two-level algorithm. The outer level uses a modified Benders decomposition type method that uses cuts generated from the inner level algorithm. The inner level solves the second-stage problem $\max_{\mathbf{d} \in \mathcal{D}} \min_{\mathbf{y} \in \Omega(\mathbf{x}, \mathbf{d})} \mathbf{b}^T \mathbf{y}$, which can be reformulated either as a bilinear optimization problem by dualizing the inner dispatch problem $\min_{\mathbf{y} \in \Omega(\mathbf{x}, \mathbf{d})} \mathbf{b}^T \mathbf{y}$ [25] or as a mixed-integer linear optimization problem by introducing binary variables and big- M constants to linearize the bilinear terms [26].

The modification to the traditional Benders decomposition proposed in Ref. [25] is the following: after solving the inner problem, the constraints of the inner dispatch problem associated with the obtained worst-case net load are added to the outer problem. This approach turns out to significantly speed up the convergence of the algorithm. It is formalized and rigorously analyzed in Ref. [27].

Modified Benders Decomposition Algorithm:

Step 1: Solve the following outer level master problem:

$$\begin{aligned} \min_{\mathbf{x}, \alpha, \mathbf{y}^i, \forall i=1, \dots, k-1} \quad & \mathbf{c}^T \mathbf{x} + \alpha \\ \text{s.t.} \quad & \alpha \geq \lambda_l^T (\mathbf{A}\mathbf{x} - \mathbf{g}) - \varphi_l^T \mathbf{h} + \eta_l^T \mathbf{d}_l, \\ & \forall l \leq k, \\ & \mathbf{F}\mathbf{x} \leq \mathbf{f}, \mathbf{x} \text{ binary}, \\ & \mathbf{y}^i \in \Omega(\mathbf{x}, \mathbf{d}^i), \forall i = 1, \dots, k-1. \end{aligned} \quad (12)$$

Let $(\mathbf{x}_k, \alpha_k)$ be the optimum. Set $L^{\text{BD}} = \mathbf{c}^T \mathbf{x}_k + \alpha_k$. The inequality (Eq. 12) is the cut generated from the second-stage problem.

Step 2: For the fixed commitment decision $\mathbf{x}_k$, solve the second-stage problem. Denote its optimal objective value as $R(\mathbf{x}_k)$, and the obtained worst-case net load as $\mathbf{d}^k$. Add the constraints with $\mathbf{d}^k$, namely $\mathbf{y}^k \in \Omega(\mathbf{x}, \mathbf{d}^k)$, to the outer problem. Set $U^{\text{BD}} = \mathbf{c}^T \mathbf{x}_k + R(\mathbf{x}_k)$.

Step 3: If $U^{\text{BD}} - L^{\text{BD}} < \varepsilon$, stop and return $\mathbf{x}_k$. Otherwise, let $k = k + 1$, and go to Step 1.

Extensive computational experiments have been conducted on some large-scale power systems, such as the ISO New England's power system [25]. Comparing to the reserve adjustment approach used in the current practice, the adaptive robust UC model demonstrates better economic efficiency in terms of lower average commitment and dispatch costs, significantly reduced volatility of the dispatch cost, and robustness against different uncertainty distributions.

CONCLUSIONS

In this article, we have discussed basic optimization models and the most recent advances for the OPF and the unit commitment problem, two of the fundamental decision-making problems in short-term power system operations. In particular, advanced global optimization techniques for solving the OPF problem are presented. An adaptive robust optimization model and solution methods for the UC problem are introduced, which provide a new, effective framework for power system operations under uncertainty. We have also discussed contingency analysis for large-scale power systems and distributed optimization schemes for decentralized control of the power systems. In conclusion, modern optimization methods play a central role in power system operations. Further advances are needed to meet new challenges arising in the fast-evolving modern power systems.

REFERENCES

1. U.S.-Canada Power System Outage Task Force. 2004. Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations. Available at <http://www.ferc.gov/industries/electric/indus-act/reliability/blackout/ch1-3.pdf>. Accessed 2014 Jan 13.
2. Christie R. 1999. Power System Test Archive. <http://www.ee.washington.edu/research/pstca>. Accessed 2014 Jan 13.
3. Frank S, Steponavice I, Rebennack S. Optimal power flow: a bibliographic survey I - formulations and deterministic methods. *Energy Syst* 2012;3(3):221–258.
4. Carpentier J. Contribution to the economic dispatch problem. *Bull Soc Fr Electric* 1962;8(3):431–447.
5. Contaxis GC, Delkis C, Korres G. Decoupled optimal power flow using linear or quadratic programming. *IEEE Trans Power Syst* 1986;PWRS-1:1–7.
6. Sousa AA, Torres GL. Robust optimal power flow solution using trust region and interior-point methods. *IEEE Trans Power Syst* 2011;26(2):487–499.
7. Dommel H, Tinney W. Optimal power flow solutions. *IEEE Trans Power App Syst* 1968;PAS-87(10):1866–1876.
8. Torres G, Quintana V. An interior-point method for nonlinear optimal power flow using voltage rectangular coordinates. *IEEE Trans Power Syst* 1998;13(4):1211–1218.
9. Phan DT. Lagrangian duality and branch-and-bound algorithms for optimal power flow. *Oper Res* 2012;60(2):275–285.
10. Lavaei J, Low S. Zero duality gap in optimal power flow problem. *IEEE Trans Power Syst* 2012;27(1):92–107.
11. Gopalakrishnan A, Raghunathan AU, Nikovski D, Biegler LT. Global optimization of optimal power flow using a branch and bound algorithm. 50th Annual Allerton Conference on Communication, Control, and Computing; 2012. p 609–616.
12. Kim BH, Baldick R. Coarse-grained distributed optimal power flow. *IEEE Trans Power Syst* 1997;12(2):932–939.
13. Kim BH, Baldick R. A comparison of distributed optimal power flow algorithms. *IEEE Trans Power Syst* 2000;15(2):599–604.
14. Biskas PN, Bakirtzis AG. Decentralized opf of large multiarea power systems. *IEE Proc Gener Transm Distrib* 2006;153(1):99–105.
15. Nogales FJ, Prieto FJ, Conejo AJ. A decomposition methodology applied to the multi-area optimal power flow problem. *Ann Oper Res* 2003;120:99–116.

16. Biskas PN, Bakirtzis AG, Macheras NI, *et al.* A decentralized implementation of dc optimal power flow on a network of computers. *IEEE Trans Power Syst* 2005;20(1):25–33.
17. Conejo AJ, Aguado JA. Multi-area coordinated decentralized dc optimal power flow. *IEEE Trans Power Syst* 1998;13(4):1272–1278.
18. Alsac O, Stott B. Optimal load flow with steady state security. *IEEE Trans Power App Syst* 1974;PAS-93(3):745–751.
19. Monticelli A, Pereira MVF, Granville S. Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Trans Power Syst* 1987;2(1):175–180.
20. Capitanescu F, Martinez Ramos JL, Panciatici P, *et al.* State-of-the-art, challenges, and future trends in security constrained optimal power flow. *Electric Power Syst Res* 2011;81:1731–1741.
21. Guan X, Luh PB, Amalfi JA. An optimization-based method for unit commitment. *Int J Electric Power Energy Syst* 1996;14:9–17.
22. Wang S, Shahidehpour S, Kirschen D, *et al.* Short-term generation scheduling with transmission constraints using augmented Lagrangian relaxation. *IEEE Trans Power Syst* 1995;10:1294–1301.
23. Takriti S, Birge JR, Long E. A stochastic model for the unit commitment problem. *IEEE Trans Power Syst* 1996;11(3):1497–1508.
24. Ben-Tal A, Ghaoui LE, Nemirovski A. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton (NJ): Princeton University Press; 2009.
25. Bertsimas D, Litvinov E, Sun XA, *et al.* Adaptive robust optimization for the security constrained unit commitment problem. *IEEE Trans Power Syst* 2013;28(1):52–63.
26. Jiang R, Wang J, Guan Y. Robust unit commitment with wind power and pumped storage hydro. *IEEE Trans Power Syst* 2012;27(2):800–810.
27. Zeng B, Zhao L. Solving two-stage robust optimization problems by a column-and-constraint generation method. *Oper Res Lett* 2013;41(5):457–461.

SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH

IGNACIO GARCIA JURADO
Department of Mathematics,
Coruna University, Coruna,
Spain

SEIO (www.seio.es), which stands for Sociedad de Estadística e Investigación Operativa, is the Spanish Society of Statistics and Operations Research. It was founded in February 12, 1962, in Madrid, in the facilities of CSIC, Consejo Superior de Investigaciones Científicas (the Spanish Science Foundation). Fifty-two persons, including academics and practitioners, participated in the inaugural meeting. Originally, SEIO was only concerned with operations research and its promotion within Spain. In fact, from 1962 to 1984, SEIO stood for Sociedad Española de Investigación Operativa (Spanish Society of Operations Research), and only in December 20, 1984, SEIO changed its name, aims and statutes, and included statistics in its list of key fields. Nowadays, it is a scientific society, which aims to promote and develop the theory, methodology, and applications of statistics and operations research mainly, but not only, within Spain.

The first President of SEIO was Fermín de la Sierra (from February 12, 1962, to March 31, 1969) and, after him until today, the Presidents have been Sixto Ríos (1969–1984), Pilar Ibarrola (1984–1986), Marco A. López Cerdá (1986–1989), Daniel Peña (1989–1992), Elías Moreno (1992–1994), Jesús Pastor (1994–1998), Rafael Infante (1998–2001), Pedro Gil (2001–2004), Domingo Morales (2004–2007), and Ignacio García Jurado (the current president). According to the statutes, a president is elected for a term of three years. Immediately after the election, he or she serves as the elected president for one and a half years; then he or she is the president for three years, and finally he or she serves as the past president for one and a half years.

Currently, the elected president of SEIO is José M. Angulo, who will be the president in September 2010.

SEIO has also a vice president of operations research and a vice president of statistics. It is governed by an Executive Committee formed by the president, the past president or the elected president, the two vice presidents, the secretary, and five other elected members. The executive committee is advised by two academic committees when dealing with scientific issues regarding operations research or statistics. Each academic committee is chaired by the corresponding vice president and has two other elected members. SEIO also has commissions which are nominated by the executive committee to promote activities in the key areas. Nowadays, there are three such commissions: the commissions of education and dissemination, cooperation and development, and universities.

Currently, SEIO has about 700 ordinary members. It has also 17 institutional members including the Instituto Nacional de Estadística (the Statistical Office of Spain).

SEIO organizes a national conference every one and a half years. In 2009 it organized the 31st National Conference in Murcia (Spain). In the last national conferences, the number of participants was about 400 and the number of lectures delivered was about 300.

SEIO publishes two journals: *Test*, a journal of statistics, and *Top*, a journal of operations research. The history of these two journals is as follows. In 1950, Sixto Ríos founded the journal *Trabajos de Estadística*. In 1963, *Trabajos de Estadística* changed its name to *Trabajos de Estadística e Investigación Operativa* and became the official journal of SEIO. In 1986 *Trabajos de Estadística e Investigación Operativa* was split into two different journals: *Trabajos de Estadística*, focused on statistics, and *Trabajos de Investigación Operativa*, focused on operations research. In 1992 *Trabajos de Estadística* became *Test* and started

publishing papers only in English. In 1993 *Trabajos de Investigación Operativa* went through a similar process: it became *Top* and started publishing papers only in English. Since 2007 *Test* and *Top* are published and distributed by Springer, though they continue being the two official journals of SEIO. The editors of *Test* have been José M. Bernardo (1992–1996), Antonio Cuevas and Wenceslao González Manteiga (1997–2001), Enrique Castillo and Juan A. Cuesta (2002–2004), and M. Ángeles Gil and Leandro Pardo (2005–2008). Since January 2009 its editors are Ricardo Cao and Domingo Morales. The editors of *Top* have been Jaume Barceló and Laureano Escudero (1993–2000) and Marco A. López Cerdá and Ignacio García Jurado (2001–2006). Since January 2007 its editors are Emilio Carrizosa and Justo Puerto. Nowadays *Test* and *Top* are well-established scientific journals in their respective fields; for instance, in the science edition of the Journal Citation Reports 2008, produced by Thomson Reuters, *Top* has an impact factor 0.694 and occupies position 45 in the list of 64 journals of operations research and management science, whereas *Test* has an impact factor 0.930 and occupies position 50 in the list of 92 journals of statistics and probability.

SEIO also distributes information and promotes both dissemination and applications of statistics and operations research. To that aim, it currently maintains a web site, whose address is www.seio.es, produces the electronic bulletin of general information *InfoSEIO*, and publishes *BEIO*, a bulletin which mainly collects papers on applications of statistics and operations research. *BEIO* is indexed in MathScinet and Zentralblatt MATH. Its current editor is M. Carmen Pardo.

Since 1984, SEIO and Fundación Ramiro Melendreras have jointly awarded a prize for young researchers. The Ramiro Melendreras Prize is given at the closing session of the national conference and it has become an important prize in its field within Spain. The winners of this prize have been Wenceslao González Manteiga (1984), Domingo Morales (1985), David Ríos (1986), Ignacio García Jurado (1988), Ricardo Cao (1989), Luciano Méndez Naya (1991), José A. Vilar (1992), J.J. Salazar (1995), M. Carmen Pardo (1997), Joaquín Sánchez Soriano (1998), Jacobo de Uña (2000), Isabel Molina (2001), Javier Toledo (2003), Alberto Olivares (2004), Pedro Terán (2006), Enrique Chacón (2007), and Sergio García Quiles (2009).

SEIO encourages the existence of working groups to support the research activities of its members. Currently SEIO has six working groups in the following fields: location theory, game theory, multicriteria decision analysis, classification and multivariate analysis, teaching of statistics and operations research, and functional data analysis.

SEIO has well-established relations with the most relevant international scientific organizations in the fields of statistics and operations research. In particular, SEIO is a member of the International Federation of Operational Research Societies (IFORS), Association of European Operational Research Societies (EURO), Association of Latin–Iberoamerican Operational Research Societies (ALIO), International Statistical Institute (ISI), and European Courses on Advances Statistics (ECAS). SEIO is also a member of the European Mathematical Society (EMS), and the Spanish Committee of Mathematics (CEMAT).

Today, SEIO is a very active scientific society and the most important Spanish organization which promotes operations research.

SPLIT CUTS

KENT ANDERSEN
Department of Mathematics,
Otto-von-Guericke
Universität, Magdeburg,
Germany

INTRODUCTION

We consider the following form of a mixed integer program (MIP):

$$\begin{aligned} \min \quad & c^T y \\ \text{s.t.} \quad & y \in \{x \in P : x_j \text{ integer for all } j \in N_I\}, \end{aligned}$$

where $P \subseteq \mathbb{R}^n$ is a polyhedron describing the set of feasible solutions to the linear program (LP) relaxation of MIP, $N := \{1, 2, \dots, n\}$ is used to index the variables, and $N_I \subseteq N$ denotes those variables that are required to be integers. The set $P_I := \{x \in P : x_j \text{ integer for all } j \in N_I\}$ denotes the set of mixed integer points in P . We assume that P is a *rational* polyhedron, that is, that P can be described with inequalities that are defined using only rational numbers.

The following concept is key in the remainder. A *disjunction* with q terms is an expression of the form $D^1 x \leq d^1 \vee D^2 x \leq d^2 \vee \dots \vee D^q x \leq d^q$, where for every $i \in \{1, 2, \dots, q\}$, the matrix D^i and the vector d^i define the linear system $D^i x \leq d^i$ in n variables. A vector $y \in \mathbb{R}^n$ satisfies this disjunction if $D^i y \leq d^i$ for some $i \in \{1, 2, \dots, q\}$.

Let $(\pi, \pi_0) \in \mathbb{Z}^{n+1}$ be arbitrary. If we assume $\pi_j = 0$ for all continuous variables $j \in N \setminus N_I$, then the *split disjunction* $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ is satisfied by all mixed integer points $x \in P_I$. The set of all split disjunctions will be denoted:

$$\Pi := \{(\pi, \pi_0) \in \mathbb{Z}^{n+1} : \pi_j = 0 \text{ for } j \in N \setminus N_I\}.$$

The idea of using disjunctions to derive cutting planes for MIP was first considered by

Balas [1]. Balas showed how to use an optimal basis for the LP relaxation of MIP, together with a disjunction to obtain a cut that was named the *intersection cut*. In a later paper [2], Balas introduced *disjunctive programs*, where disjunctions were considered together with polyhedra to obtain valid cuts for MIP. A disjunctive program is an optimization problem defined from a polyhedron and a disjunction. The problem is to find a point in the polyhedron that satisfies the disjunction and has the best possible objective.

The special case of split disjunctions was first studied by Cook *et al.* [3], where the term *split cuts* was introduced. Given the mixed integer program (MIP), the *split closure* with reference to P and N_I was analyzed. The split closure with reference to P and N_I is the set obtained from P by adding *all* split cuts that can be derived by considering *all* split disjunctions in Π , and it was shown in Cook *et al.* [3] that the split closure with reference to P and N_I is a polyhedron. Since then, several different proofs have been given to show this fact [4–6].

Strong relationships have been established between split cuts and other classes of cutting-planes; see Cornuéjols and Li [7] for an excellent description of the relationships between different classes of cutting planes. Split cuts have been shown to be equivalent to mixed integer Gomory cuts and mixed integer rounding cuts [8]. Furthermore, it has been shown that every split cut is an intersection cut, and that every split cut can be obtained from a basis of the LP relaxation [4,9].

Recently, computational experiments have been performed to measure how tight an approximation the split closure gives to the mixed integer hull on a number of test problems. Balas and Saxena [10] formulated a parametric MIP for computing a split cut that is violated by a given point; that is, for solving the separation problem for split cuts. Although the separation problem for split cuts is known to be NP-complete [11], Balas and Saxena managed to compute the split

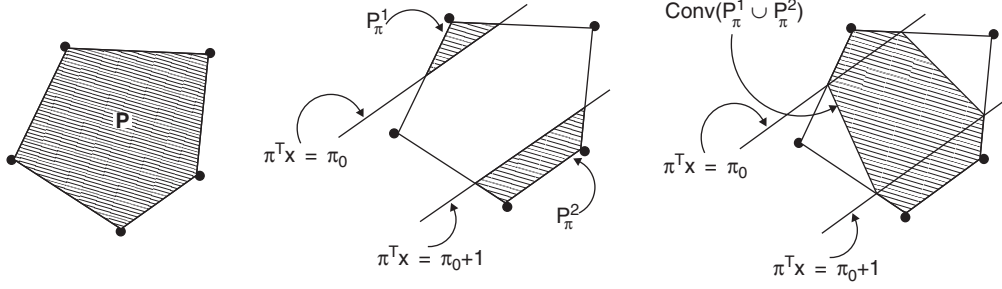


Figure 1. Effect of applying a split disjunction to a polytope. (a) Polytope P , (b) The sets P_π^1 and P_π^2 , and (c) The disjunctive hull.

closure on several test problems and they demonstrated that the split closure often provides a very tight approximation. Dash *et al.* [5] provided a different formulation of the separation problem and they compared the strength of the split closure to the strength of other classes of cutting-planes.

The remainder is organized as follows: In the section titled “Split Disjunctions and Polyhedra,” we describe the basic operation of applying a split disjunction to obtain a tighter approximation of P_I than P . In the section titled “Split Cuts from Bases of the LP Relaxation,” we derive split cuts from split disjunctions and bases of the LP relaxation of MIP. Finally, in the section titled “Computational Results on the Split Closure,” we present computational results to measure the quality of the split closure as an approximation to the set of mixed integer point in P_I .

SPLIT DISJUNCTIONS AND POLYHEDRA

We now present the operation of using a split disjunction to construct a tighter approximation of a mixed integer set. To illustrate the construction, we first give an example in dimension two.

Example 1. Consider the polyhedron P in $\mathbb{R}^2$ shown in Fig. 1(a), and suppose x_1 and x_2 are both required to be integer; that is, suppose $N = N_I = \{1, 2\}$.

In Fig. 1(b), the two parallel lines $\pi_1 x_1 + \pi_2 x_2 = \pi_0$ and $\pi_1 x_1 + \pi_2 x_2 = \pi_0 + 1$ corresponding to a split disjunction $\pi^T x \leq$

$\pi_0 \vee \pi^T x \geq \pi_0 + 1$ are shown. This split disjunction defines the two subsets $P_\pi^1 := \{x \in P : \pi^T x \leq \pi_0\}$ and $P_\pi^2 := \{x \in P : \pi^T x \geq \pi_0 + 1\}$ of P that are shown by the two shaded regions in Fig. 1(b). Observe that since no integer point $x \in \mathbb{Z}^2$ satisfies $\pi_0 < \pi^T x < \pi_0 + 1$, every integer point in P belongs to the set $P_\pi^1 \cup P_\pi^2$. Therefore, the valid inequalities for $P_\pi^1 \cup P_\pi^2$ are valid for the integer points in P .

Since an inequality is valid for $P_\pi^1 \cup P_\pi^2$ if and only if it is valid for $\text{Conv}(P_\pi^1 \cup P_\pi^2)$, the valid inequalities for $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ are valid inequalities for the integer points in P . The set $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ is shown by the shaded polygon in Fig. 1(c), and we call this set the *disjunctive hull* defined by the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. Observe that there are two valid inequalities for $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ that are not valid for P , and these inequalities can therefore be used as cuts to cut off fractional points in P .

We now generalize the construction presented in Example 1 to any mixed integer set P_I . Let $(\pi, \pi_0) \in \Pi$ be arbitrary. We have that every mixed integer point $x \in P_I$ satisfies $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. As in Example 1, the disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ defines two subsets $P_\pi^1 := \{x \in P : \pi^T x \leq \pi_0\}$ and $P_\pi^2 := \{x \in P : \pi^T x \geq \pi_0 + 1\}$ of P , and every mixed integer point in P_I belongs to $P_\pi^1 \cup P_\pi^2$. It follows that the valid inequalities for $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ are valid for P_I . As in Example 1, we call the set $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ the *disjunctive hull* defined by $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. We will not present the proof here, but it can be shown that

the disjunctive hull defined from any split disjunction is a polyhedron.

Observe that *all* split disjunctions $(\pi, \pi_0) \in \Pi$ potentially can tighten the LP relaxation of MIP; that is, we have $P_I \subseteq \text{Conv}(P_\pi^1 \cup P_\pi^2) \subseteq P$ for all $(\pi, \pi_0) \in \Pi$. The intersection of *all* disjunctive hulls over *all* split disjunctions $(\pi, \pi_0) \in \Pi$ therefore gives the tightest possible relaxation of P_I that can be obtained with split cuts

$$\text{SC} := \bigcap_{(\pi, \pi_0) \in \Pi} \text{Conv}(P_\pi^1 \cup P_\pi^2).$$

The set SC is called the *split closure* with reference to P and N_I . Observe that the split closure with reference to P and N_I is defined as the intersection of an infinite number of polyhedra. It is therefore not obvious whether SC is a polyhedron. It was first shown by Cook *et al.* that SC is a polyhedron [3].

SPLIT CUTS FROM BASES OF THE LP RELAXATION

The objective in this section is to characterize the geometry of the split cuts that can be derived from a basis of the LP relaxation. Central to this objective is the special case when P is a translate of a polyhedral cone, which we will consider in the section titled “Split Cuts from Translates of Polyhedral Cones.” In the section titled “Split Cuts from Bases and Split Cuts from Polyhedra,” we apply the results of the previous section to obtain cuts from a basis, and we relate these

cuts to the split closure with reference to P and N_I . We show how to strengthen a split disjunction in the section titled “Strengthening Split Disjunctions.” In the section titled “Mixed Integer Gomory Cuts from Rows of the Simplex Tableau,” we show how mixed integer Gomory cuts can be viewed as split cuts.

Split Cuts from Translates of Polyhedral Cones

We now consider the special case when the LP relaxation P is defined from a translate of a polyhedral cone. The reason is, as we will explain later, that every split cut can be obtained from an object of this type. Furthermore, the formulas for computing a split cut from a simplex table will follow directly from the results presented in this section.

A *polyhedral cone* is a cone that is finitely generated, that is a set of the form $C := \{x \in \mathbb{R}^m : x = \sum_{j \in J} s_j r^j \text{ and } s \geq 0\}$, where J is a finite index set and $r^j \in \mathbb{R}^m$ for $j \in J$. In other words, C is the set of points $x \in \mathbb{R}^m$ that can be written as a nonnegative combination of the vectors r^j for $j \in J$. A *translate* of a polyhedral cone C is a set that can be written as the sum of a point $\bar{x} \in \mathbb{R}^m$ and a polyhedral cone; that is, a set of the form $\bar{x} + C$. In the following, we will call the vectors r^j for $j \in J$ for *rays*. Figure 2(a) gives an example of a translate of a polyhedral cone defined from two rays r^1 and r^2 and a point $\bar{x} \in \mathbb{R}^2$.

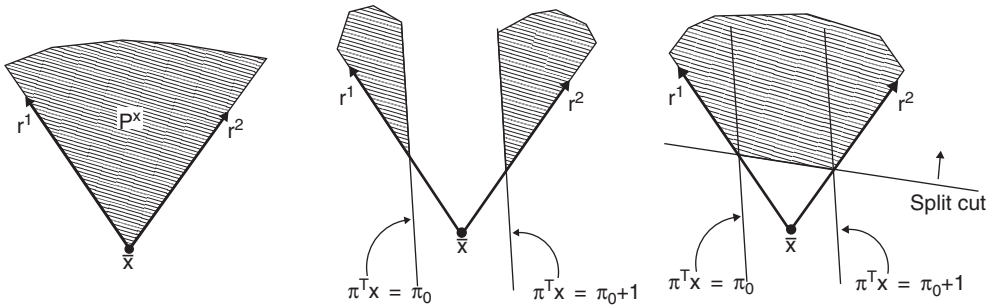


Figure 2. Deriving a split cut from a translated cone. (a) A polyhedron P^x of the form $P^x = \bar{x} + \text{Cone}(r^1, r^2)$, (b) The polyhedron P^x and a split disjunction, and (c) The disjunctive hull and the split cut.

In this section we consider the special case when the LP relaxation P has the form

$$P = \{(x, s) \in \mathbb{R}^m \times \mathbb{R}^{|J|} : x = \bar{x} + \sum_{j \in J} s_j r^j \text{ and } s \geq 0\}$$

and $P_I = \{(x, s) \in P : x \text{ is an integer}\}$. Observe that the projection of P onto the space of the x variables is a translate of a polyhedral cone. The variables s_j for $j \in J$ are the multipliers on the rays in the conic combinations, and they are considered continuous variables.

The geometry of the sets P and P_I will now be explained by considering their projection onto the space of the x variables, that is by considering the sets

$$P^x := \{x \in \mathbb{R}^m : \exists s \in \mathbb{R}^{|J|} \text{ s.t. } (x, s) \in P\}$$

$$\text{and } P_I^x := P^x \cap \mathbb{Z}^m.$$

Observe that P_I^x is a pure integer set; that is, all variables are required to be integers. Furthermore, the linear relaxation P^x of P_I^x is the translate $P^x = \bar{x} + C$ of the polyhedral cone $C = \{x \in \mathbb{R}^m : x = \sum_{j \in J} s_j r^j \text{ and } s \geq 0\}$.

We next show how split disjunctions can be used to derive split cuts for P_I . To illustrate the construction, we first give an example where the sets P^x and P_I^x are two dimensional.

Example 2. Consider the translated polyhedral cone P^x in Fig. 2(a). Two parallel lines $\pi^T x = \pi_0$ and $\pi^T x = \pi_0 + 1$ corresponding to a split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ are drawn in Fig. 2(b). As in Example 1, the split disjunction defines two subsets $\{x \in P^x : \pi^T x \leq \pi_0\}$ and $\{x \in P^x : \pi^T x \geq \pi_0 + 1\}$ of P^x (the two shaded areas in Fig. 2(b)). The disjunctive hull corresponding to $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ is the shaded area in Fig. 2(c).

Observe that only one inequality is needed to describe the disjunctive hull (the split cut given in Fig. 2(c)). Also observe that, if we let x^1 denote the point on the halfline $\{\bar{x} + \alpha r^1 : \alpha \geq 0\}$ that satisfies $\pi^T x^1 = \pi_0$, and if we let x^2 denote the point on the halfline $\{\bar{x} + \alpha r^2 : \alpha \geq 0\}$ that satisfies $\pi^T x^2 = \pi_0 + 1$, then the split cut in Fig. 2(c) corresponds to the line that passes through x^1 and x^2 . In other words, the split cut in the figure can be

obtained by first identifying the two points x^1 and x^2 , and then computing the line that passes through x^1 and x^2 .

We next generalize the construction of a split cut given in Example 2. Let $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ be a valid split disjunction. We assume $\bar{x}$ does not satisfy the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$, that is we assume $\pi_0 < \pi^T \bar{x} < \pi_0 + 1$ (we will show below that no split cut can be derived from a split disjunction that is satisfied by $\bar{x}$). Define $\epsilon_\pi := \pi^T \bar{x} - \pi_0$ to be the amount by which $\bar{x}$ violates the first term of the split disjunction.

Now let $j \in J$ be arbitrary. For $\alpha \geq 0$, define $x^j(\alpha) := \bar{x} + \alpha r^j$. We are interested in the smallest value $\alpha \geq 0$ for which $x^j(\alpha)$ satisfies the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$.

There are three cases: either the halfline $\{x^j(\alpha) : \alpha \geq 0\}$ intersects the hyperplane $\pi^T x = \pi_0$, or the halfline $\{x^j(\alpha) : \alpha \geq 0\}$ intersects the hyperplane $\pi^T x = \pi_0 + 1$, or the halfline $\{x^j(\alpha) : \alpha \geq 0\}$ does not intersect any of these two hyperplanes. In Example 2, the halfline $\{x^1(\alpha) : \alpha \geq 0\}$ intersects the hyperplane $\pi^T x = \pi_0$, and the halfline $\{x^2(\alpha) : \alpha \geq 0\}$ intersects the hyperplane $\pi^T x = \pi_0 + 1$.

The specific value of $\alpha \geq 0$ for which $x^j(\alpha)$ satisfies the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ can now be computed. If $\pi^T r^j < 0$, then $x^j(\alpha)$ is on the hyperplane $\pi^T x = \pi_0$ for $\alpha = -\frac{\epsilon_\pi}{\pi^T r^j}$, if $\pi^T r^j > 0$, then $x^j(\alpha)$ is on the hyperplane $\pi^T x = \pi_0 + 1$ for $\alpha = \frac{1 - \epsilon_\pi}{\pi^T r^j}$, and if $\pi^T r^j = 0$, then $x^j(\alpha)$ is not on any of the hyperplanes $\pi^T x = \pi_0$ and $\pi^T x = \pi_0 + 1$ for any $\alpha \geq 0$. Hence, if we define

$$\alpha_\pi^j := \begin{cases} -\frac{\epsilon_\pi}{\pi^T r^j} & \text{if } \pi^T r^j < 0, \\ \frac{1 - \epsilon_\pi}{\pi^T r^j} & \text{if } \pi^T r^j > 0, \\ +\infty & \text{otherwise,} \end{cases} \quad (1)$$

for every $j \in J$, then α_π^j is the smallest value $\alpha \geq 0$ for which $x^j(\alpha)$ satisfies the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. Observe that we define $\alpha_\pi^j = +\infty$ when the halfline $\{x^j(\alpha) : \alpha \geq 0\}$ does not intersect any of the hyperplanes $\pi^T x = \pi_0$ and $\pi^T x = \pi_0 + 1$. When $\alpha_\pi^j < +\infty$, the point $x^j(\alpha_\pi^j) = \bar{x} + \alpha_\pi^j r^j$ is

called an *intersection point* [1]. In Example 2, the points x^1 and x^2 are intersection points.

Given the numbers α_π^j for $j \in J$, a split cut associated with the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ can be defined as follows:

$$\sum_{j \in J} \frac{s_j}{\alpha_\pi^j} \geq 1. \quad (2)$$

Observe that for every $j \in J$ for which $\alpha_\pi^j < +\infty$, the point (x^j, s^j) defined by $x^j := x^j(\alpha_\pi^j)$, $s_j^j := \alpha_\pi^j$ and $s_k^j := 0$ for $k \in J \setminus \{j\}$ is in P and satisfies Equation (2) with equality. In other words, every intersection point gives a point in P that satisfies Equation (2) with equality.

Recall that the sets $P_\pi^1 := \{(x, s) \in P : \pi^T x \leq \pi_0\}$ and $P_\pi^2 := \{(x, s) \in P : \pi^T x \geq \pi_0 + 1\}$ give the points in P that satisfy the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. The validity of Equation (2) for the disjunctive hull $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ was proven by Balas [1]. In that paper Equation (2) was called the *intersection cut* due to the fact that the intersection points give points in P that satisfy Equation (2) with equality. The following lemma shows that, in fact, Equation (2) gives a complete description of the disjunctive hull:

Lemma 1. *Let $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ be a split disjunction. We have*

- (i) *If $\pi^T \bar{x} \notin [\pi_0, \pi_0 + 1]$, then $\text{Conv}(P_\pi^1 \cup P_\pi^2) = P$.*
- (ii) *If $\pi^T \bar{x} \in [\pi_0, \pi_0 + 1]$, then $\text{Conv}(P_\pi^1 \cup P_\pi^2) = \{(x, s) \in P : \text{satisfies Equation (2)}\}$.*

Lemma 1(i) shows that no split cut can be derived from a split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ that is satisfied by $\bar{x}$. Secondly, Lemma 1(ii) shows that inequality (2) is exactly the cut needed to describe the disjunctive hull when $\bar{x}$ violates the split disjunction.

Split Cuts from Bases and Split Cuts from Polyhedra

We now show how the results in the section titled “Split Cuts from Translates of Polyhedral Cones” can be used to derive split cuts from a basis of the LP relaxation of a MIP.

This basis can, for example, be an optimal basis for the LP relaxation. However, the results presented in this section apply to *any* basis of the LP relaxation. In this section we assume the LP relaxation of MIP is given by

$$P = \{x \in \mathbb{R}^n : Ax = b \text{ and } x \geq 0\},$$

where A is an $m \times n$ matrix, and b is an $m \times 1$ vector. Recall that the set $N := \{1, 2, \dots, n\}$ is used to index the coordinates of x , and that $N_I \subseteq N$ is used to index the variables that are integer constrained. We assume P has a basis; that is, we assume the rank of A is m . The j th column of A is denoted a_j .

Let $B \subseteq N$ denote a subset of the variables that define a basis of the LP relaxation; that is, suppose $|B| = m$ and that the vectors $a_j \in \mathbb{R}^m$ for $j \in B$ are linearly independent. Also let A_B denote the submatrix of A defined from the columns of A indexed by B , and let $J := N \setminus B$ index the nonbasic variables. Finally, let $\bar{x}(B) \in \mathbb{R}^n$ denote the basic solution corresponding to the basis B .

We now use the basis B to give a different representation of P . Firstly, by multiplying the equality system $Ax = b$ with the basis-inverse $(A_B)^{-1}$, we obtain

$$P = \{x \in \mathbb{R}^n : x_i + \sum_{j \in J} \bar{a}_{i,j} x_j = \bar{x}_i(B), \text{ for } i \in B \text{ and } x \geq 0\}, \quad (3)$$

where $\bar{a}_{i,j} \in \mathbb{R}^m$ for $j \in J$ is defined by $\bar{a}_{i,j} := (A_B)^{-1} a_j$. We next add the redundant equalities $x_j = x_j$ for $j \in J$ to the representation (3) of P . We can now write P in the following form:

$$P = \{x \in \mathbb{R}^n : \begin{aligned} x_i &= \bar{x}_i(B) - \sum_{j \in J} \bar{a}_{i,j} x_j, & \text{for } i \in B, \\ x_j &= x_j, & \text{for } j \in J, \\ x &\geq 0. \end{aligned} \quad (4)$$

Finally, by defining the vectors $r^j \in \mathbb{R}^n$ for $j \in J$

$$r_k^j := \begin{cases} -\bar{a}_{k,j} & \text{if } k \in B, \\ 1 & \text{if } k = j, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

the representation (4) of P can be rewritten in the form

$$P = \left\{ x \in \mathbb{R}^n : x = \bar{x}(B) + \sum_{j \in J} r^j x_j \text{ and } x \geq 0 \right\}. \quad (6)$$

Observe that the vectors r^j for $j \in J$, and the basic solution $\bar{x}(B)$, are directly available from the simplex table corresponding to the basis B .

We now construct a relaxation of P corresponding to the basis B . If we skip the nonnegativity constraints on the basic variables x_i for $i \in B$, we obtain the following relaxation $P(B)$ of P :

$$\begin{aligned} P(B) &:= \{x \in \mathbb{R}^n : Ax = b \text{ and } x_j \geq 0 \text{ for } j \in J\} \\ &= \{x \in \mathbb{R}^n : x = \bar{x}(B) \\ &\quad + \sum_{j \in J} r^j x_j \text{ and } x_j \geq 0 \text{ for } j \in J\} \\ &= \bar{x}(B) + C(B), \end{aligned}$$

where $C(B)$ is the polyhedral cone generated by the vectors r^j for $j \in J$. Since $P \subseteq P(B)$, the valid inequalities for the mixed integer points in $P(B)$ are also valid inequalities for the mixed integer points in P . In particular, the split cuts that can be derived from $P(B)$ are also satisfied by the points in P_I .

We can therefore apply the results in the section titled “Split Cuts from Translates of Polyhedral Cones” to obtain valid split cuts for P_I from a split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ and $P(B)$. We now specialize the results in the section titled “Split Cuts from Translates of Polyhedral Cones” to the polyhedron $P(B)$.

Let $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ be an arbitrary split disjunction. First observe that Lemma 1(i) shows that if $\bar{x}(B)$ satisfies the split disjunction, then no split cut can be derived from $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ and $P(B)$. In the following, we therefore assume $\pi_0 < \pi^T \bar{x}(B) < \pi_0 + 1$. Let $\epsilon_\pi := \pi^T \bar{x}(B) - \pi_0$ denote the amount by which $\bar{x}(B)$ violates the first term of the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$.

As in the section titled “Split Cuts from Translates of Polyhedral Cones,” for every $j \in J$, we need to compute the smallest value $\alpha \geq 0$ for which $\bar{x}(B) + \alpha r^j$ satisfies the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. For every $j \in J$, let $\alpha_\pi^j(B)$ denote this value. Applying formula (1), we obtain that for every $j \in J$, the value

$$\alpha_\pi^j(B) := \begin{cases} \frac{\epsilon_\pi}{\sum_{i \in B} \pi_i \bar{a}_{i,j} - \pi_j} & \text{if } \pi_j - \sum_{i \in B} \pi_i \bar{a}_{i,j} < 0, \\ \frac{1 - \epsilon_\pi}{\pi_j - \sum_{i \in B} \pi_i \bar{a}_{i,j}} & \text{if } \pi_j - \sum_{i \in B} \pi_i \bar{a}_{i,j} > 0, \\ +\infty & \text{otherwise,} \end{cases} \quad (7)$$

is the smallest value of $\alpha \geq 0$ for which the point $\bar{x}(B) + \alpha r^j$ satisfies the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$. The split cut that can be derived from $P(B)$ and $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ is now given by

$$\sum_{j \in J} \frac{x_j}{\alpha_\pi^j(B)} \geq 1.$$

We now summarize what we have achieved in this section. Let $\mathcal{B}^*$ denote the set of *all* bases of P ; that is, let $\mathcal{B}^*$ be the set of m -subsets B of N such that the vectors a_j for $j \in B$ are linearly independent. Also, for every $B \in \mathcal{B}^*$ and split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$, let $P_\pi^1(B) := \{x \in P(B) : \pi^T x \leq \pi_0\}$ and $P_\pi^2(B) := \{x \in P(B) : \pi^T x \geq \pi_0 + 1\}$ denote the points in $P(B)$ that satisfy the split disjunction.

Lemma 1(ii) shows that, for any basis $B \in \mathcal{B}^*$ such that $\bar{x}(B)$ violates the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$, the disjunctive hull defined by $P(B)$ and $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ is given by

$$\begin{aligned} & \text{Conv}(P_\pi^1(B) \cup P_\pi^2(B)) \\ &= \left\{ x \in P(B) : \sum_{j \in J} \frac{x_j}{\alpha_\pi^j(B)} \geq 1 \right\}. \end{aligned}$$

In other words, the disjunctive hull defined by $P(B)$ and the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ is described by the split cut that can be obtained from $P(B)$ and the split disjunction. If $\bar{x}(B)$ satisfies the split disjunction,

then Lemma 1(i) shows that $\text{Conv}(P_\pi^1(B) \cup P_\pi^2(B)) = P(B)$; that is, the disjunctive hull defined by $P(B)$ and the split disjunction is given by $P(B)$, and no split cut can be derived.

Recall that $P_\pi^1 := \{x \in P : \pi^T x \leq \pi_0\}$ and $P_\pi^2 := \{x \in P : \pi^T x \geq \pi_0 + 1\}$ denote the points in P that satisfy the split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$, and that $\text{Conv}(P_\pi^1 \cup P_\pi^2)$ is the disjunctive hull defined by P and this split disjunction. A natural question is whether there are split cuts that can be obtained from P and a split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ which cannot be obtained from $P(B)$ and the split disjunction for some basis $B \in \mathcal{B}^*$. A negative answer to this question was given in Andersen *et al.* [4] by showing the formula

$$\text{Conv}(P_\pi^1 \cup P_\pi^2) = \cap_{B \in \mathcal{B}^*} \text{Conv}(P_\pi^1(B) \cup P_\pi^2(B)).$$

This formula shows that the disjunctive hull defined by P and a split disjunction $\pi^T x \leq \pi_0 \vee \pi^T x \geq \pi_0 + 1$ can be described by the split cuts that can be obtained from the split disjunction and bases B of P . We note that it was demonstrated in Andersen *et al.* [4] that infeasible bases are also needed for the formula to hold.

The formula also implies that the split closure can be rewritten as follows. Let $P_I(B) := \{x \in P(B) : x_j \text{ integer for all } j \in N_I\}$ denote the mixed integer points in $P(B)$, and let $\text{SC}(B)$ denote the split closure with reference to $P(B)$ and N_I . We have

$$\begin{aligned} \text{SC} &= \cap_{(\pi, \pi_0) \in \Pi} \text{Conv}(P_\pi^1 \cup P_\pi^2) \\ &= \cap_{(\pi, \pi_0) \in \Pi} \cap_{B \in \mathcal{B}^*} \text{Conv}(P_\pi^1(B) \cup P_\pi^2(B)) \\ &= \cap_{B \in \mathcal{B}^*} \cap_{(\pi, \pi_0) \in \Pi} \text{Conv}(P_\pi^1(B) \cup P_\pi^2(B)) \\ &= \cap_{B \in \mathcal{B}^*} \text{SC}(B). \end{aligned}$$

In other words, the split closure with reference to P and N_I is the intersection of the split closures with reference to the sets $P_I(B)$ and N_I over all bases B of P .

Strengthening Split Disjunctions

The idea of strengthening a cut $c^T x \geq c_0$ for P_I is to modify $c^T x \geq c_0$ to an inequality $\bar{c}^T x \geq \bar{c}_0$ that dominates $c^T x \geq c_0$ on P ; that is, to modify $c^T x \geq c_0$ to an inequality $\bar{c}^T x \geq \bar{c}_0$ that

cuts off more points of P . A procedure for strengthening split cuts was first presented by Balas and Jeroslow [12].

Let $B \in \mathcal{B}^*$ be an arbitrary basis of the LP relaxation of MIP, and let $(\pi, \pi_0) \in \Pi$ denote a split disjunction that is violated by $\bar{x}(B)$. We now show how to construct a split disjunction $(\pi', \pi'_0) \in \Pi$ such that the split cut derived from $P(B)$ and (π', π'_0) dominates the split cut derived from $P(B)$ and (π, π_0) .

The procedure we present only modifies coordinates of π corresponding to nonbasic and integer-constrained variables. Let $J_I := J \cap N_I$ index those coordinates, and let $k \in J_I$ be arbitrary. For simplicity, let $\bar{a}^k(\pi) := \sum_{i \in B} \pi_i \bar{a}_{i,k}$.

By using Equation (7) of the section titled “Split Cuts from Bases and Split Cuts from Polyhedra,” the coefficient on x_k in the split cut defined by $P(B)$ and (π, π_0) is given by $\frac{1}{\alpha_\pi^k(B)}$, where

$$\alpha_\pi^k(B) := \begin{cases} \frac{\epsilon_\pi}{\bar{a}^k(\pi) - \pi_k} & \text{if } \pi_k - \bar{a}^k(\pi) < 0, \\ \frac{1 - \epsilon_\pi}{\pi_k - \bar{a}^k(\pi)} & \text{if } \pi_k - \bar{a}^k(\pi) > 0, \\ +\infty & \text{otherwise.} \end{cases} \quad (8)$$

Observe that the value π_k in the split disjunction (π, π_0) is not involved in the formula for $\epsilon_\pi = \sum_{i \in B} \pi_i \bar{x}_i(B) - \pi_0$. Hence the violation of the split disjunction is not changed if we change π_k to another integer. Furthermore, π_k is not involved in the formula for $\alpha_\pi^j(B)$ for any $j \in J \setminus \{k\}$.

It follows that we can choose a new value π'_k for π_k so as to minimize the coefficient $\frac{1}{\alpha_{\pi'}^k(B)}$ on x_k without changing the coefficients on the other nonbasic variables in the resulting split cut.

It is easy to see that the optimal value π'_k for π_k is either $\pi'_k = \lfloor \bar{a}^k(\pi) \rfloor$ or $\pi'_k = \lceil \bar{a}^k(\pi) \rceil$. If we choose $\pi'_k = \lfloor \bar{a}^k(\pi) \rfloor$, then the new value of $\alpha_{\pi'}^k(B)$ will be $\frac{\epsilon_\pi}{\bar{a}^k(\pi) - \lfloor \bar{a}^k(\pi) \rfloor}$ and if we choose $\pi'_k = \lceil \bar{a}^k(\pi) \rceil$, the new value of $\alpha_{\pi'}^k(B)$ will be $\frac{1 - \epsilon_\pi}{\lceil \bar{a}^k(\pi) \rceil - \bar{a}^k(\pi)}$. Since we would like to make the coefficient $\frac{1}{\alpha_{\pi'}^k(B)}$ on x_k as small as possible, it is best to choose $\pi'_k = \lceil \bar{a}^k(\pi) \rceil$ if $\frac{1 - \epsilon_\pi}{\lceil \bar{a}^k(\pi) \rceil - \bar{a}^k(\pi)} > \frac{\epsilon_\pi}{\bar{a}^k(\pi) - \lfloor \bar{a}^k(\pi) \rfloor} \iff \epsilon_\pi < \bar{a}^k(\pi) - \lfloor \bar{a}^k(\pi) \rfloor$. Otherwise, it is best to choose $\pi'_k = \lfloor \bar{a}^k(\pi) \rfloor$.

We therefore have the following lemma which shows how to obtain a split disjunction $(\pi', \pi'_0) \in \Pi$ from (π, π_0) such that the split cut derived from (π', π'_0) and $P(B)$ dominates the split cut derived from (π, π_0) and $P(B)$.

Lemma 2. *Let $B \in \mathcal{B}^*$ be a basis of P and let $(\pi, \pi_0) \in \Pi$ be a split disjunction violated by $\bar{x}(B)$. Define the split disjunction $(\pi', \pi'_0) \in \Pi$ by $\pi'_0 := \pi_0$ and*

$$\pi'_j := \begin{cases} \pi_j & \text{if } j \in N \setminus J_I, \\ \lfloor \bar{a}^j(\pi) \rfloor & \text{if } j \in J_I \text{ and } \bar{a}^j(\pi) \\ & - \lfloor \bar{a}^j(\pi) \rfloor \leq \epsilon_\pi, \\ \lceil \bar{a}^j(\pi) \rceil & \text{if } j \in J_I \text{ and } \bar{a}^j(\pi) \\ & - \lfloor \bar{a}^j(\pi) \rfloor > \epsilon_\pi. \end{cases} \quad (9)$$

Then, the split cut obtained from (π', π'_0) and $P(B)$ dominates the split cut obtained from (π, π_0) and $P(B)$.

Mixed Integer Gomory Cuts from Rows of the Simplex Table

Given a row of a simplex table, a mixed integer Gomory cut [13] can be derived. Throughout this section $B \in \mathcal{B}^*$ denotes an arbitrary basis of the LP relaxation. Consider the row of the simplex table where the variable x_i , $i \in B$, is basic:

$$x_i + \sum_{j \in J} \bar{a}_{i,j} x_j = \bar{x}_i(B).$$

Assume $\bar{x}_i(B)$ is fractional. Define $f_0 := \bar{x}_i(B) - \lfloor \bar{x}_i(B) \rfloor$ and $f_j := \bar{a}_{i,j} - \lfloor \bar{a}_{i,j} \rfloor$ for the nonbasic and integer-constrained variables $j \in J_I$. Gomory showed that the following inequality is satisfied by the mixed integer points in P_I :

$$\sum_{j \in J_I: f_j \leq f_0} \frac{f_j}{f_0} x_j + \sum_{j \in J_I: f_j > f_0} \frac{1-f_j}{1-f_0} x_j + \sum_{j \in J \setminus J_I: \bar{a}_{i,j} \geq 0} \frac{\bar{a}_{i,j}}{f_0} x_j - \sum_{j \in J \setminus J_I: \bar{a}_{i,j} < 0} \frac{\bar{a}_{i,j}}{1-f_0} x_j \geq 1. \quad (10)$$

We will derive inequality (10) as a split cut in two steps. First consider the split disjunction $(\pi, \pi_0) \in \Pi$ defined by $\pi_0 := \lfloor \bar{x}_i(B) \rfloor$ and $\pi = e^i$, where e^i denotes the i^{th} unit vector.

Observe that $\epsilon_\pi = f_0$. Applying formula (7) of the section titled “Split Cuts from Bases and Split Cuts from Polyhedra,” we obtain the following split cut from $P(B)$ and the split disjunction (π, π_0) :

$$\sum_{j \in J: \bar{a}_{i,j} > 0} \frac{\bar{a}_{i,j}}{f_0} x_j - \sum_{j \in J: \bar{a}_{i,j} \geq 0} \frac{\bar{a}_{i,j}}{1-f_0} x_j \geq 1. \quad (11)$$

Inequality (11) is also known as *Gomory’s fractional cut*. The second step is to apply Lemma 2 to strengthen the split disjunction (π, π_0) . Formula (9) now gives the following strengthened split disjunction (π', π'_0) of (π, π_0) :

$$\pi'_j = \begin{cases} \pi_j & \text{if } j \in N \setminus J_I, \\ \lfloor \bar{a}_{i,j} \rfloor & \text{if } j \in J_I \text{ and } \bar{a}_{i,j} - \lfloor \bar{a}_{i,j} \rfloor \leq f_0, \\ \lceil \bar{a}_{i,j} \rceil & \text{if } j \in J_I \text{ and } \bar{a}_{i,j} - \lfloor \bar{a}_{i,j} \rfloor > f_0. \end{cases}$$

Computing the split cut associated with (π', π'_0) and $P(B)$ then gives the mixed integer Gomory cut (Eq. 10).

COMPUTATIONAL RESULTS ON THE SPLIT CLOSURE

Balas and Saxena [10] optimized over the split closure for a number well-known test problems. Specifically, a MIP was considered, and the best solution in the split closure was computed. We now present their approach, and some of the computational results that were obtained. In this section, we assume the LP relaxation P of P_I is of the form $P = \{x \in \mathbb{R}^n : Ax \geq b\}$.

In order to optimize over the split closure, the *separation problem* for the split closure needs to be solved. This separation problem receives as input a (current) point $\bar{x} \in P$, and addresses the question of whether or not there exists a split cut that is violated by $\bar{x}$. If such a split cut exists, this split cut must be returned.

Let $\bar{x} \in P$ denote a point that must be separated from SC, if possible, and let $\bar{s} := A\bar{x} - b$ denote the corresponding values of the surplus variables. Balas and Saxena formulated

the separation problem for the split closure with reference to P and N_I as the following parametric mixed integer linear program (PMILP):

$$\begin{aligned}
& \min z(\theta) := (1 - \theta)u^T \bar{s} + \theta v^T \bar{s} - \theta(1 - \theta) \\
& \text{s.t.} \\
& (u - v)^T A - \pi = 0, \\
& (v - u)^T b + \pi_0 = \theta - 1, \\
& u, v \geq 0, \\
& (\pi, \pi_0) \in \Pi, \\
& 0 \leq \theta \leq 1.
\end{aligned}$$

Balas and Saxena showed that solving PMILP solves the separation problem for the point $\bar{x}$ and the split closure with reference to P and N_I . Specifically, it was shown that the point $\bar{x}$ belongs to the split closure with reference to P and N_I if and only if the optimal objective value to PMILP is nonnegative. If the optimal objective of PMILP is negative, then an optimal solution $(u^*, v^*, \pi^*, \pi_0^*, \theta^*)$ to PMILP gives a split cut $\alpha^T x \geq \beta$ defined by

$$\begin{aligned}
\alpha &= (u^*)^T A - \theta^* \pi^* \text{ and} \\
\beta &= (u^*)^T b - \theta^* \pi_0^*,
\end{aligned}$$

which is violated by $\bar{x}$.

Observe that PMILP is a nonlinear mixed integer program. However, also observe that if the variable $\theta \in [0, 1]$ is fixed to a specific value, then the remaining problem is a mixed integer linear program. This observation can therefore be exploited by discretizing the interval $[0, 1]$ and only solve PMILP for

Table 2. Summary of Results for Pure Integer Instances

Total Number of Pure Integer Instances	23
Average % gap closed by split closure	72.47%
98–100% gap closed	9
75–98% gap closed	4
25–75% gap closed	6
< 25% gap closed	4

a finite set of carefully chosen values of $\theta \in [0, 1]$.

The above observation is then used by Balas and Saxena to design a cutting-plane algorithm for optimizing over the split closure with reference to P and N_I . First, the LP relaxation of MIP is solved, and an optimal solution $\bar{x}$ is obtained. The separation problem PMILP is then solved for $\bar{x}$ and a number of selected values of $\theta \in [0, 1]$. If a violated split cut is found, the cut is added to the (current) LP relaxation, the problem is reoptimized and a new point $\bar{x}$ is obtained to be separated. This procedure is then iterated. The cutting-plane algorithm terminates when it is concluded that enough values of $\theta \in [0, 1]$ have been tested without finding a violated split cut. Observe that although a cutting-plane is added in every iteration, this cut is *not* included in the formulation of the separation problem PMILP; that is, the separation problem PMILP is always formulated in terms of the original constraints $Ax \geq b$.

Balas and Saxena measures the performance of the above algorithm by the amount of integrality gap that is closed. Table 1 and Table 2 below are from Balas and Saxena [10], and they give a short summary of the performance of split cuts. There are a total of 61 test problems. Table 1 gives results for the mixed integer instances, and Table 2 gives results for the pure integer instances (when there are no continuous variables). As can be seen from these tables, the split cuts close a substantial amount of the integrality gap on most problems. Furthermore, in Balas and Saxena [10] and Dash *et al.* [5], the split closure is compared with other classes of cutting-planes, and it is concluded that split cuts most often close substantially more of the integrality gap.

Table 1. Summary of Results for Mixed Integer Instances

Total Number of Mixed Integer Instances	38
Average % gap closed by split closure	72.78%
98–100% gap closed	15
75–98% gap closed	10
25–75% gap closed	7
< 25% gap closed	6

REFERENCES

1. Balas E. Intersection cuts - a new type of cutting plane for integer programming. *Oper Res* 1971;19:19–39.
2. Balas E. Disjunctive programming. *Ann Discrete Math* 1979;5:3–51.
3. Cook WJ, Kannan R, Schrijver A. Chvátal closures for mixed integer programming problems. *Math Program Ser A* 1990;47:155–174.
4. Andersen K, Cornuéjols G, Li Y. Split closure and intersection cuts. *Math Program Ser A* 2005;102:457–493.
5. Dash S, Gunluk LO, Lodi A. Mir closures of polyhedral sets. *Math Program Ser A* 2010;121:33–60.
6. Vielma JP. A constructive characterization of the split closure of a mixed integer linear program. *Oper Res Lett* 1998;35:29–35.
7. Cornuéjols G, Li Y. Elementary closures for integer programs. *Oper Res Lett* 2001;28:1–8.
8. Nemhauser G, Wolsey LA. A recursive procedure to generate all cuts for 0–1 mixed integer programs. *Math Program Ser A* 1990;46:379–390.
9. Balas E, Perregaard M. A precise correspondence between lift-and-project cuts, simple disjunctive cuts, and mixed integer Gomory cuts for 0–1 programming. *Math Program Ser B* 2003;94:221–245.
10. Balas E, Saxena A. Optimizing over the split closure. *Math Program Ser A* 2008;113:219–240.
11. Caprara A, Letchford A. On the separation of split cuts and related inequalities. *Math Program Ser A* 2003;94:279–294.
12. Balas E, Jeroslow RG. Strengthening cuts for mixed integer programs. *Eur J Oper Res* 1980;4:224–234.
13. Gomory RE. An algorithm for the mixed integer problem. Santa Monica (CA): The Rand Corporation; 1960. Research Memorandum – 2597.

SPREADSHEET MODELING FOR OPERATIONS RESEARCH PRACTICE

THOMAS GROSSMAN
University of San Francisco,
San Francisco, California,
USA

This article presents the essential information for operations research practitioners to make effective use of spreadsheets, with particular attention to issues that are important to success but not widely understood.

Spreadsheets have been an important platform for analysis ever since VisiCalc launched in 1979. Spreadsheets put the power of mathematical modeling and computer programming into the hands of anyone with a PC. Microsoft Excel is as much a part of the modern office as a desk, and spreadsheets have a high degree of acceptance among managers. Empirical research shows many examples where spreadsheets are used for managerial decision making or are otherwise essential to finance, operations, general business, and science [1–4].

Spreadsheets are a valuable resource in the practitioner's toolbox. The spreadsheet is a flexible fourth-generation programming language with high acceptance by consulting clients, endowed with remarkable power and flexibility for operations research techniques. Practitioners can enhance their effectiveness by integrating the spreadsheet into their practice

EXAMPLES OF HIGH IMPACT OPERATIONS RESEARCH USING SPREADSHEETS

Here are some examples of spreadsheets having an impact on operations research practice.

Oggier *et al.* [5] explain how Nestlé deployed operations research to end users around the world. They developed, deployed, and integrated into their global business

practices sensitivity analysis, forecasting, simulation, and optimization. Distributed analysts use them for evaluating the economic profitability of Nestlé's projects and more generally evaluating the value of the Nestlé group and its many businesses.

A special issue of Interfaces [6] contains 12 papers highlighting the rich usage of spreadsheets in operations research practice. Farasyn *et al.* [7] describe how Procter & Gamble (P&G) used spreadsheet models across the company to determine optimal inventory levels for required customer service levels, subject to the characteristics and constraints of particular supply chains. The models have contributed to inventory reductions exceeding \$350 million and significant intangible benefits to P&G. Because the spreadsheet models are easy to use and share, they became the standard tool for inventory setting at P&G and are used by hundreds of supply chain planners worldwide.

Olavson and Fry [8] discuss their experiences in building and applying spreadsheet-based decision-support tools. Using examples from Hewlett-Packard (HP) in forecasting, planning, procurement, and product management, they discuss lessons learned for creating reusable spreadsheet tools to support fact-based decision making, and using spreadsheets as the front- and back-end user interface for models coded in a procedural language. This article is notable for providing a high level road map for approaching operations research-driven change projects (independent of spreadsheets) with options ranging from “no analytics” because the organization is not ready, through to “transferable tool.”

Amaral and Kuettner [9] review spreadsheet techniques and approaches that they used to analyze supply chains at HP and to manage spreadsheet limitations. They discuss how they employed many of Excel's advanced capabilities while minimizing unnecessary complexity.

Pasupathy and Medina-Borja [10] describe the use of data envelopment analysis to make resource-allocation recommendations to help American Red Cross managers evaluate the performance of different chapters. They integrate Microsoft Excel, Premium Solver, Visual Basic for Applications, and Microsoft Access to formulate, automatically program, and solve 4000 different linear-programming models.

IMPLEMENTING OPERATIONS RESEARCH MODELS IN A SPREADSHEET

The spreadsheet is suitable for programming high quality operations research models. As discussed later, the spreadsheet can be tolerant of poor programming practices that work initially but prove unmaintainable over time. Therefore, it is incumbent on the operations research practitioner to perform a spreadsheet programming task with the same rigor and diligence as with a traditional language such as C#. The operations research programmer will benefit from understanding the characteristics of the spreadsheet programming language, the principles of “spreadsheet engineering,” and the findings of research on spreadsheet errors.

The Spreadsheet as a Programming Language

This section is based on Grossman *et al.* [4]. Operations research practitioners select a traditional programming language (e.g., FORTRAN, C#, Java) based on its characteristic strengths and weaknesses for a given task. Let us examine the characteristics of the dominant spreadsheet language, Microsoft Excel.

First, Excel is a “fourth-generation language” (4GL). Second, Excel is in the class of “rapid development languages” (RDLs). Third, Excel functions as an “integrated development environment” (IDE).

A 4GL is a nonprocedural programming language with natural language-like features that can be employed directly by less skilled programmers, with a simple interface. Analysts can use a 4GL, such as Excel, to develop computer applications more rapidly than conventional programming languages and enjoy

dramatic boosts in productivity. In addition, the spreadsheet has a higher level of abstraction than do third-generation languages, with automatic definition of variables by grid location, automatic variable typing, and no concept of registers or storage locations; indeed the underlying hardware is invisible.

The spreadsheet is a member of a class of programming languages called *rapid-development languages* that yield faster development than traditional languages. Using the “function point” measure of the work accomplished by a computer program, spreadsheets require on average only six statements to implement a function point that would require 30 statements in Visual Basic, 50 statements in C++, or 110 statements in FORTRAN, and hence “spreadsheets boost productivity through the roof” [11].

The spreadsheet functions as an IDE, which is a software tool that helps programmers create applications by simplifying or automating programming tasks. The spreadsheet provides a seamlessly integrated source code editor and interpreter, and powerful report creation and printing capabilities. The spreadsheet itself provides a simple graphical user interface (GUI), and Excel’s Forms toolbar provides convenient tools for building more advanced GUI features that allow spreadsheet application software to be developed. It is notable that the spreadsheet grants the programmer real-time observation of programming statement results; this powerful feature of the spreadsheet computer programming language is typically not available in other languages, and provides much of the functionality of a traditional debugger. Excel also provides helpful productivity tools such as Trace Precedents/Dependents, Display Formulas, R1C1 cell referencing, as well as traditional tools such as Watch Window and Evaluate Formula.

In the final analysis, the spreadsheet programming language is “highly functional, extendible, easy to integrate, and scalable” [12].

Spreadsheet Engineering

Professional programmers understand the necessity of employing a rigorous software engineering methodology when implementing an operations research model in any programming language. Spreadsheets are no different. The principles and practices of “spreadsheet engineering”—software engineering for spreadsheets—are necessary to write effective, maintainable code in a spreadsheet programming language. However, because the spreadsheet is an RDL, it is tolerant of poor coding practices [11]; hence spreadsheet programs are accommodating of poorly structured, unmaintainable code. This does not mean that the spreadsheet is a poor programming language, but rather that the programmer must be diligent to employ good practices.

Several operations research organizations have proprietary spreadsheet engineering methodologies. These are rigorous, refined, documented, and vital to their success. However, published spreadsheet engineering methodologies are underdeveloped relative to methodologies for traditional languages. The operations research professional who writes spreadsheet software is well-advised to familiarize himself/herself with the existing spreadsheet engineering methodologies, and should be aware that the general principles of traditional software engineering apply to writing programs in a spreadsheet language.

The domain of large-scale spreadsheet financial planning models has several sophisticated methodologies ([13, 14], and Operis in [15]), which are compared in [16]. There have been a few sets of detailed spreadsheet engineering recommendations for creating more general spreadsheet models, especially financial models. The most complete are book-length treatments by Read and Batson [17] and Tennent and Friend [18]. Valuable contributions can be found in Chapter 3 of [19] and [20].

A particular risk when using a spreadsheet language is the failure to discard a prototype model implementation. An initial proof-of-concept model programmed in any language is normally a poor foundation for further programming effort. Just as in a traditional language, it is essential to discard a

prototype spreadsheet, and start over with a structured design, similar to a painter who discards an initial sketch and uses a fresh canvas for the final work.

Research Findings on Spreadsheet Errors

A spreadsheet containing cell formulas is a computer program, and all computer programs are susceptible to errors. A corpus of spreadsheet errors is available [21]. Several scoping-level studies (summarized in [22, 23]) suggest that spreadsheets are excessively susceptible to errors. This prompted more rigorous follow-on research [24–27] which finds that the definition of errors is problematic; past studies are insufficiently documented and suspected; careful auditing identifies many potential errors, many of which cause no numerical impact; and some documented errors are large. Most important, they find that some organizations seem to be able to build error-free spreadsheets.

There is evidence that professional programmers use spreadsheets at a high level of accuracy [4] including a massive (70 MB) spreadsheet that has been demonstrated to be error-free. This, plus the many examples of successful spreadsheet application software belie any special error risk when spreadsheets are written by people with programming skill.

The professional literature on programming [28] suggests that the distinction between unskilled “amateur” programmers trumps any special error vulnerabilities of the spreadsheet compared to other languages. In Weinberg’s nomenclature, amateur programmers have domain skill but not programming skill, and generally write bad code in whatever language they use. In contrast, professional programmers have programming skill independent of any domain skill, and generally write good code.

Hence, poor programmers are likely to write poor spreadsheets, and good programmers are likely to write good spreadsheets. This is consistent with the findings of the spreadsheet error literature. The hypothesis that spreadsheets are unusually error-prone or unsafe is at best unproven.

As a practical matter, an operations research team that has the sophistication

to write quality software in a procedural language can undoubtedly write quality software in a spreadsheet language provided they approach the task with equivalent diligence.

FOUR APPROACHES TO USING SPREADSHEETS IN OPERATIONS RESEARCH PRACTICE

There are four approaches to using spreadsheets for operations research practice: Direct programming, VBA to generate spreadsheet cell formulas, Direct VBA, and Spreadsheet for I/O. (Note: VBA, or “Visual Basic for Applications,” is a popular programming language that is bundled free with Excel. VBA is a simplified version of the programming language Visual Basic that is used routinely by professional programmers in many disciplines. VBA has strong integration with the Excel spreadsheet.)

The Direct programming approach is writing spreadsheet cell formulas in the usual way. Direct programming is by far the most common approach. Amaral and Kuettner [9] show how some lesser-known Excel features can be used to enhance the user interface and program flexible scenarios.

The VBA to generate spreadsheet cell formulas approach enables the programmer to surmount the limitations of direct programming. Direct programming can be very challenging for large-scale models, and models with a high degree of dimensionality. Although it is easy to use the spreadsheet Copy-Paste and Fill features to create cell formulas, it becomes impractical when there are variations in formulas that must be typed by hand. To solve this problem, the analyst can use VBA to populate a spreadsheet with cell formulas; in essence, one writes a program to write a program. Details can be found in [29]. This approach is powerful and surprisingly straightforward.

The Direct VBA approach requires coding the model in VBA, and using the spreadsheet proper primarily for data management. In this approach, the spreadsheet itself is relatively simple, and the complexity of the model is handled behind the scenes by VBA code. VBA can make calls to sophisticated

operations research algorithms that are custom-written or provided by third-party vendors. Pasupathy and Medina-Borja [10] use a spreadsheet and VBA to automate the creation and solution of thousands of linear programs.

The Spreadsheet for I/O approach requires the use of traditional operations research software, such as a matrix generator coded in a procedural language, or a modeling language such as GAMS or AMPL. Data is pulled in from a spreadsheet and results pushed out into a spreadsheet.

Regardless of the approach, the operations research practitioner can take advantage of Excel’s effective reporting and graphing capabilities, and the widespread availability of Excel on desktops around the globe.

SPREADSHEET TOOLS FOR OPERATIONS RESEARCH

More information on these products can be found in [30] and a compendium [31]. Operations research algorithms are provided in spreadsheets via third party “add-ins.” Spreadsheet add-ins are products that add capability to the spreadsheet. Add-ins can provide new features accessed through menus, new functions, or be precoded modules in the form of templates. There is a sizable ecosystem of commercial add-ins for operations research and other analytic needs. Add-ins for operations research fall into three distinct groups: model-driven analysis (including Monte Carlo simulation and optimization), statistics and data-driven analysis, and business function-specific analysis. Operations research practitioners should also be aware of certain spreadsheet productivity tools.

Productivity and Spreadsheet Development Tools

There are add-ins to aid spreadsheet development including error checking, spreadsheet comparison tools, automation, security, version control, and more [32].

Model Handoff to Clients and Spreadsheet Application Software

Operations research practitioners sometimes need to transfer a model to a client, or to convert a model into a standalone application. With a spreadsheet, model transfer is challenging because unless appropriate steps are taken, the source code can be modified by the user with unpredictable results. This issue is discussed in [9, 33].

There exist tools to remove access to the source code before deployment to users, by converting a spreadsheet into a standalone application (i.e., an executable file) or web-based HTML page. One can also use Excel features to convert a spreadsheet into an add-in [34].

Model-Driven Analysis

Model-driven analysis is the application of an analytical technique to a spreadsheet model to obtain insight or decision guidance. There is a tutorial paper on performing sensitivity analysis in native Excel [35], and on scenario analysis in native Excel and integrating it with optimization and simulation [36]. There are several commercial tools for sensitivity analysis that quickly compute and chart the effect of changing an input on a designated output, or generate a Tornado Chart, including SensIt, TopRank, and Risk Solver Platform, and the academic tool Sensitivity ToolKit [37].

There are tools that perform iterative search, extending the Goal Seek feature of Excel that searches on a model input to obtain a specified value of a model output. These handle multiple goal seeks at once and can remember and repeat them. They include Spreadsheet Goal Seeker, and multicell Goal Seeker [38].

There are several commercial add-ins for building and computing decision trees, including TreePlan, PrecisionTree, and Risk Solver Platform [39].

Monte Carlo Simulation. There are several feature-rich commercial add-ins for Monte Carlo simulation. In addition to Oracle's Crystal Ball and Palisade's @Risk, recent entrants include Vose Software's

ModelRisk and Frontline Systems' Analytic Solver Platform with innovative "simulation optimization" capability [40].

Optimization. Optimization add-ins to Excel are the mostly highly developed, and are distinctive for their power. Optimization in a spreadsheet can be done using Excel's built-in Solver feature. Befitting a free product made available to the mass market, the built-in Solver is accessible to users with little or no operations research training, has restrictions on model size, and provides a limited feature set. The built-in Solver has proven valuable for countless users, is employed routinely in some business processes, and has introduced the power of optimization to a mass audience.

Operations research practitioners with more demanding requirements will want to utilize one of the high performance commercial optimization add-ins that are available [41]. The leading commercial products are the Premium Solver and Analytic Solver Platform by Frontline Systems (the vendor of the built-in Solver), and What's Best! by Lindo Systems. Spreadsheet solvers can handle problems of large scale and complexity. These add-ins can integrate with third-party, world-class optimization tools such as CPLEX and GUROBI, and have the ability to extract the underlying optimization mode out of the spreadsheet, feed it to an external algorithm that runs independent of the spreadsheet, and then populate the spreadsheet with the optimal solution.

The level of performance of commercial spreadsheet optimization add-ins can be very high. For example (Straver and Fylstra, personal communication, June 14, 2012) describe the use of the Analytic Solver Platform add-in to Excel using the GUROBI solver, to solve to optimality a real-world linear programming model with 2.4 million decision variables and 50,000 constraints in 6 min and 52 s using a \$1000 personal computer. This includes 1:00 min to parse the spreadsheet model, 5:20 min for GUROBI to set it up, and 24 s for GUROBI to optimize it.

Spreadsheet optimization is suitable for routine use. For example, a utility company has a spreadsheet linear program with

250,000 decision variables that is optimized on a regular basis (Straver and Fylstra, personal communication, June 14, 2012).

There are commercial add-ins for metaheuristics, including Analytic Solver Platform and Evolver. Parametric optimization can be performed using the commercial product Analytic Solver Platform, and also academic products Sensitivity ToolKit, and SolverTable [42].

Spreadsheet Design for Optimization. For most operations research techniques, an add-in has obvious spreadsheet design requirements (e.g., decision tree add-ins) or can handle arbitrary spreadsheet design (e.g., simulation add-ins). In contrast, optimization models are sensitive to spreadsheet design. When constructing a spreadsheet model that will be subject to optimization, it is important to code the model so as to preserve linearity (which is challenging to verify in a spreadsheet and hence must be built-in at design time), and enable easy identification of constraints to the optimization add-in. Guidelines and examples for using spreadsheets for optimization can be found in [19, 43–45].

Statistics and Data-Driven Analysis

Statistics and data-driven analysis are tools that perform statistics and business intelligence functions, including forecasting. Excel has a free statistics add-in called the *Analysis ToolPak* that in older releases has exhibited quality problems and lost the confidence of the statistics community; these issues have largely been addressed. In addition, there are many commercial statistics add-ins available at a variety of price points, including StatTools, Lumenaut Statistics, and Analyse-it [46].

Commercial tools for spreadsheet data mining include XLMiner, Analytic Solver Platform, and NeuroXL [47].

Spreadsheet time-series forecasting has around a dozen add-in products. These generate time-series forecasts using smoothing, regression, and neural network approaches. They include XLMiner, ForecastX Wizard, and PeerForecaster [48].

Business Function-Specific Analysis

There are several commercial add-ins designed for specific functional areas of business. In finance, there are add-ins for pricing various types of derivatives, to support trading decisions (with integration to brokers), and a tool for automatic data feeds. There is a tool for spreadsheet marketing engineering, templates for spreadsheet marketing plans, and a spreadsheet advertising keyword pricing tool. There are spreadsheet tools for Six Sigma greenbelts, for statistical quality control, and even spreadsheet tools intended for use on the factory floor [31].

The Queueing ToolPak add-in [49] offers add-in Excel functions that allow the practitioner to easily program queueing theory formulas directly in a spreadsheet.

FURTHER READING

This site provides a list of EuSpRIG papers over the years. Most are available for download: European Spreadsheet Risks Interest Group Conference Papers, Abstracts, and Indexes, <http://www.eusprig.org/conference-abstracts.htm>, accessed July 23, 2014.

RELATED ARTICLES

Introduction to Linear Programming Models; Newton-Type Methods; The Need for Simulation Optimization; Monte Carlo Simulation for Quantitative Risk Assessment; Describing Decision Problems by Decision Trees; Model-Based Forecasting; Operations Research in Data Mining; The M/M/s queue; The M/M/s/c queue

REFERENCES

1. Coles S, Rowley J. Spreadsheet modelling for management decision making. *Ind Manag Data Syst* 1996;96(7):17–23.
2. Ragsdale CT. Teaching management science with spreadsheets: from decision models to decision support. *INFORMS Trans Educ* 2001;1(2):68–74.
3. Croll, G. The importance and criticality of spreadsheets in the city of London. *Proceedings of the European Spreadsheet*

- Risks Interest Group, 2005. <http://arxiv.org/abs/0709.4063>, accessed July 18, 2014.
4. Grossman, TA, Mehrotra, V, Özlük, Ö. Lessons from mission-critical spreadsheets. *Commun Assoc Inf Syst* 2007; 20(60):1009–1042. <http://aisel.aisnet.org/cais/vol20/iss1/60/>, accessed July 18, 2014.
 5. Oggier C, Fragnière E, Stuby J. Nestlé improves its financial reporting with management science. *Interfaces* 2005;35(4):271–280.
 6. LeBlanc LJ, Grossman TA. The use of spreadsheet software in the application of management science and operations research. *Interfaces* 2008;38(4):225–227.
 7. Farasyn I, Perkoz K, van de Velde W. Spreadsheet models for inventory target setting at procter & gamble. *Interfaces* 2008;38(4):241–250.
 8. Olavson T, Fry C. Spreadsheet decision-support tools: lessons learned at Hewlett-Packard. *Interfaces* 2008;38(4):300–310.
 9. Amaral J, Kuettner D. Analyzing Supply Chains at HP Using Spreadsheet Models. *Interfaces* 2008;38(4):228–240.
 10. Pasupathy KS, Medina-Borja A. Integrating excel, access, and visual basic to deploy performance measurement and evaluation at the American Red Cross. *Interfaces* 2008;38(4):324–337.
 11. McConnell S. Rapid development. Redmond, WA: Microsoft Press; 1996.
 12. Ragsdale CT, Power DJ, Bergey PK. AMCIS 2002 Panels and Workshops II: Spreadsheet-based DSS curriculum issues. *Commun Assoc Inf Syst* 2002;9(21):356–365.
 13. FAST. The FAST Standard: Practical, structured design rules for financial modeling, Version FAST01b; 2012. <http://www.fast-standard.org/>, accessed July 18, 2014.
 14. SSRB (2011). Best Practice Spreadsheet Modeling Standards, version 7.0. <http://www.ssrb.org/>, accessed July 18, 2014.
 15. Swan J. Practical financial modelling: a guide to current practice. 2nd ed. London: CIMA Publishing; 2008.
 16. Grossman, TA, Özlük, Ö. Spreadsheets Grow Up: Three Spreadsheet Engineering Methodologies for Large Financial Planning Models. European Spreadsheet Risks Interest Group 11th Annual Symposium, Greenwich, England, 2010 July.
 17. Read, N, Batson, J. Spreadsheet modeling best practice; 1999. <http://www.eusprig.org/smbp.pdf>, accessed June 14, 2012.
 18. Tennent J, Friend J. Guide to business modelling. London: Profile Books; 2001.
 19. Powell SG, Baker KR. Management science, the art of modeling with spreadsheets 4/e. Hoboken: Wiley; 2013.
 20. Raffensperger, JF. New guidelines for spreadsheets. Proceedings of the European Spreadsheet Risks Interest Group, Amsterdam, 2001 July; pp. 61–76; 2001.
 21. EuSpRig. EuSpRIG Original Horror Stories; 2009. <http://www.eusprig.org/stories.htm>, accessed July 18, 2014.
 22. Panko RR. What we know about spreadsheet errors. *J End User Comput* 1998;10(2):15–21.
 23. Panko, RR. What we know about spreadsheet errors [Revised], 2008 May. <http://panko.shidler.hawaii.edu/My%20Publications/Whatknow.htm>, accessed July 18, 2014.
 24. Powell SG, Baker KR, Lawson B. A critical review of the literature on spreadsheet errors. *Decis Support Syst* 2008a;46:128–138.
 25. Powell SG, Baker KR, Lawson B. An auditing protocol for spreadsheet models. *Inform Manag* 2008b;45(5):312–320.
 26. Powell SG, Baker KR, Lawson B. Errors in operational spreadsheets. *J End User Org Comput* 2009a;21(3):24–36.
 27. Powell SG, Baker KR, Lawson B. Impacts of errors in operational spreadsheets. *Decis Support Syst* 2009b;47(2):126–132.
 28. Weinberg G. The psychology of computer programming, Silver Anniversary Edition. New York: Dorset House; 1998.
 29. LeBlanc L, Galbreth M. Implementing large-scale optimization models in excel using VBA. *Interfaces* 2007a;37(4):370–382.
 30. Grossman, TA. Resources for Spreadsheet Analysts. *Analytics* 2010 May–June, pp. 8–13; 2010. <http://www.analytics-magazine.org/may-june-2010/139-software-survey-resources-for-spreadsheet-analysts.html>, accessed July 18, 2014.
 31. Spreadsheet Analytics Website; 2014. <http://www.usfca.edu/management/spreadsheet-analytics>, accessed July 18, 2014.
 32. Development. Spreadsheet Analytics Website; 2014. <http://www.usfca.edu/management/spreadsheet-analytics/development-management>, accessed July 23, 2014, accessed July 23, 2014.
 33. Grossman, TA. Source code protection for applications written in Microsoft Excel and Google Spreadsheet. Proceedings of the European Spreadsheet Risks Interest

- Group, 2007, pp. 81–92; 2007. <http://arxiv.org/abs/0801.4774>, accessed July 23, 2014.
34. Deployment & Compilers. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/development-management/deployment-compilers>, accessed October 6, 2014.
35. Grossman, TA. A primer on spreadsheet analytics. Proceedings of the European Spreadsheet Risks Interest Group, 2008, pp. 129–140; 2008. <http://arxiv.org/abs/0809.3586>, accessed July 23, 2014.
36. Grossman TA, Özlük Ö. A spreadsheet scenario analysis technique that integrates with optimization and simulation. *INFORMS Trans Educ* 2009;10(1):18–33.
37. Sensitivity Analysis. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/sensitivity-analysis>, accessed October 6, 2014.
38. Goal-Seeking. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/goal-seeking>, accessed October 6, 2014.
39. Decision Trees. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/decision-trees>, accessed October 6, 2014.
40. Monte Carlo Simulation. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/monte-carlo-simulation>, accessed October 6, 2014.
41. Optimization Solvers. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/solvers/optimization>, accessed October 6, 2014.
42. Parametric Optimization. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/model-driven-analytics/solvers/parametric-optimization>, accessed October 6, 2014.
43. Conway DG, Ragsdale CT. Modelling optimization problems in the unstructured world of spreadsheets. *Omega* 1997;25(3):313–322.
44. LeBlanc L, Hill J, Greenwell G, *et al.* Nukote's spreadsheet linear programming models for optimizing transportation. *Interfaces* 2004;34(2):139–46.
45. LeBlanc L, Galbreth M. Designing large-scale supply chain linear programs in spreadsheets. *Commun ACM* 2007b;50(8):59–64.
46. Statistics. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/data-driven-analytics/statistics>, accessed October 7, 2014.
47. Data Mining. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/data-driven-analytics/data-mining>, accessed October 6, 2014.
48. Time-Series Forecasting. Spreadsheet Analytics Website; 2014. <https://sites.google.com/a/usfca.edu/business-analytics/business-function-analytics/forecasting>, accessed October 6, 2014.
49. QTP. Queueing ToolPak; 2014. <http://queueingtoolpak.org/>, accessed July 18, 2014.

STAKEHOLDER PARTICIPATION

ROBERT L. HEATH
School of Communication,
University of Houston,
Houston, Texas

Consideration of stakeholder and stakeholder participation is as old as discussions of society and practices of governance. In the past forty years, however, the term has become widely and sometimes loosely used in academic and professional literature, and even in media coverage and popular jargon. It is widely discussed in management, public relations, issues management, marketing, business ethics, governance, and even product/service advertising. References to triple bottom line thinking can suggest some very heavy lifting to engage with various stakeholders to mutual benefit or serve as mere platitude that can be distorted by market positioning.

Reflecting the seriousness of stakeholder engagement, Renn *et al.* [1] cautioned, “Without public involvement, environmental policies are doomed to fail” (p. xiii). Concerns about sustainability, corporate social responsibility, social marketing, and community relations make stakeholder participation an important topic for management and operations. The topic requires theoretical examination, monitoring processes, understanding of negotiation and collaborative decision making, as well as examination of the ethics of community engagement.

Interest in stakeholders takes a descriptive/pragmatic tack as well as a normative approach based on the ethics of citizenship. The post-World War II era, especially during the civil rights and anti-Vietnam war battles, moved a peripheral topic into prominence. Riding this wave, Freeman [2] helped popularize stakeholder theory and participation as a driver of strategic management. Topics such as decision analysis, control in the face of uncertainty, and risk management and communication have grown variously

from the logic he and others articulated. Looking at stakeholders from a private sector perspective, he advised, “A situation where a solution to a stakeholder problem is imposed by a government agency or the courts must be seen as a managerial failure” (pp. 74–75). Thus, he reasoned, “Managers in organizations with high Stakeholder Management Capability think in ‘stakeholder-serving’ terms” (p. 80). Among the many reports emphasizing the ethical (as risk democracy) and functional value of stakeholder participation were the National Research Council’s *Improving Risk Communication* (1989), *Understanding Risk* (1996), and *Decision Making for the Environment* (2005). Similarly relevant is the logic of Kahneman *et al.* [3] that judgment often is made under uncertainty which depends on individual reference points, varied willingness to risk, and sense of utilities to be gained or lost.

STAKEHOLDER LOGIC

Stakeholder theory addresses rights and obligations (both inside and outside of an organization regardless of whether it is for profit, nonprofit, or governmental) as power resources applied by competing interests. The topic is inseparable from related considerations of efficiency, power, risk, and control. Competing interests and conflicting issue positions are fundamental to stakeholder theory. Interest is best conceptualized as something desired to be gained or lost—something that can be risked or is at risk. Conflict results as power is wielded by each person’s or any group’s ability and willingness to use resources to maximize its outcomes even if that means (and in conflict it can) diminishing other groups’ or persons’ ability to achieve their goals.

What then is the rationale for the popularity of the term, and even the argument that there is a stakeholder theory? As de Bussy [4,5] observed, “The theory raises

a highly contentious question: in whose interest should a business corporation or other type of organization be run?" (p. 4815). One answer is that the interest served is that of the organization (as a collective of many interests). It would presume that the organization through various power resources makes the stakeholder dependent on it for serving the interest of the stakeholder. Another view suggests that the stakeholder and the organization have mutual, and even aligned, interests, in fact mutually beneficial interests and objectives.

A pragmatic and even cybernetic logic reasons that each organization depends on the stakeholders who hold resources that the organization needs to achieve its mission and vision. Pragmatically, the focal organization might be a dominant force in the relationship, the stakeholder might be dominant, or the relationship might be mutually beneficial. In fact, the matter may often be more a matter of "grey" rather than any one of the three options purely applied.

A stake is something of value (tangible or intangible, material or symbolic—or both) that one organization or entity holds that another wants or needs: A risk of gain or loss. Mitchell *et al.* [6] defined stake as "something that can be lost. The use of risk to denote stake appears to be a way to narrow the stakeholder field to those with legitimate claims, regardless of their power to influence the firm or the legitimacy of their relationship to the firm" [7, 8]. Wages or salary, by this definition, is a stake, but so is trust. Labor is a resource, as is wage or salary. The quality of each product is a stake. A buyer needs more stakes to buy a higher quality item—all other factors being the same. A coal mining company needs stakes held (application of regulatory standards) by regulators and legislators (environmental protection standards translated into regulation) as well as stakes of employee goodwill.

Stakeholders and stakeholders are bound to one another by relationships defined by exchange processes and the meaning that defines the stakes and those processes. Stakeholders have stakes that under various conditions they grant or hold. For instance, workers hold stakes (their time, skill, and

effort) that they can grant to or hold from various companies that need labor; employment contracts, labor relations, and marketplace conditions define such exchange. Stakes (salary, wages, and other benefits) held by the company can be granted to workers in this exchange. Workers are stakeholders of wages/salaries and benefits—as well as working conditions, such as safety. Safety may be assured by the employer or required by legislation and regulation. The rationale for resource dependency is that each definable entity is variously a stakeholder and a stakeholder who have interests more or less aligned.

Freeman and Gilbert [9] advised the following implications of effective understanding and engagement with stakeholders: "Corporate strategy must reflect an understanding of the values of organizational members and stakeholders." Such understanding can and must inform "the ethical nature of strategic choice" (p. 7).

Each organization, a business for instance, has multiple stakeholders (and stakeholders), on which they are likely to be variously dependent and interdependent through complex stakeholder/stakeholder networks. Thus, for instance, the workers who seek the stakes of safe working conditions hold votes that legislators need. Thus, the votes of the workers are exchanged for the legislative position of the legislators who hold the stake as to whether the regulation produces (or does not) a safe working environment. This notion of multiple stakes, multiple stakeholders (and the reciprocation) of stakeholders is the essence of stakeholder participation. Stakeholder theory and participation best practices center on rights, obligations, legitimacy, reputation, power resources, and the quality of relationships, particularly as interests are aligned badly or well. As a consequence, strategic management relies on stakeholder/holding and granting; strategic participation depends on knowing who the stakeholders are, what stakes they hold, and their willingness and ability to use them as resources to add value in efforts to align interests and achieve objectives.

MONITORING AS STRATEGIC MANAGEMENT

The strategic management challenge is to identify those persons, groups, and organizations who can affect the organization and who are affected by it in ways that pose opportunity or peril. Strategic management depends on effective stakeholder monitoring: Identifying stakeholders, knowing their interests and relationships (dependence, interdependence, and independence) with the organization and other stakeholders, and their issues positions. Monitoring must be guided by vigilant regard for stakeholders' willingness and ability to participate and to grant or withhold stakes the organizations wants and needs. In such matters, two strategic management alternatives are salient: risk analysis and decision analysis. "A decision analysis can include a risk analysis component, and the design of a risk management plan may require decision analysis support" [10].

The years following World War II witnessed increased activism that peaked during the 1960s and 1970s. During this era, issues management and related disciplines developed to propose systems for monitoring and engaging stakeholders. Organizations developed means for scanning, identifying, tracking, interpreting, and analyzing issues and their advocates [7]. Strategic management, operations, decisions about corporate social responsibility, marketing especially social marketing, and public relations developed functional and ethical responses to critics.

Mitchell *et al.* [6] aided efforts to identify and understand stakeholders by addressing not only who or what really counts, but also the "maddening variety of signals on how questions of stakeholder identification might be answered" (p. 853). They addressed ways of knowing "which groups are stakeholders deserving or requiring management attention, and which are not" (p. 854). They reasoned, "Persons, groups, neighborhoods, organizations, institutions, societies, and even the natural environment are generally thought to qualify as actual or potential stakeholders" (p. 855). Within the limits of

available resources and their efficient use, who and what should be given attention? To that end, these researchers distinguished claimants versus influencers, actual versus potential relationships, and an array of factors of power, dependence, and reciprocity relevant to the conditions and qualities of relationship networks.

Mitchell *et al.*'s quest was to determine the added value stakeholder identification and participation could yield. The best theory and set of best practices yield the most benefit—add the most value to the relationship—at the most expedient cost. Upon this logic, Mitchell *et al.* [6] offered a taxonomy for stakeholder identification. It features *interdependence* of power (within each relationship and over resources), the *legitimacy* of the power, and the claims it is used to make. In addition, their discussion features the *urgency* of stakeholder claims and the *salience* of any of many claims among competing stakeholders. Viewed in this way, stakeholder theory features an array of variously interdependent stakeholders, which in idiosyncratic ways support or oppose various measures and choices relevant to an organization's management and operations.

Even a superficial review of the literature on this topic can suggest that stakeholders in practice are typically viewed as members of the "general public." Closer review reveals a variety of generic categories: activist publics, intraindustry players, interindustry players, potential activist publics, customers, employees, legislators, regulators, judiciary, investors, neighbors, standard media, and social media [7]. Discussions of stakeholder participation can limit attention to concern over activists, customers, and concerned citizens—and even citizens who are "not concerned but should be." In fact, stakeholder theory and participation reasons that intraindustry or interindustry players (or governmental agencies, as well as other activist groups to activists) may be as important or more important, as might be investors; and they might be customers, or potential customers; voters or potential voters.

The sophisticated trend is to engage in thoughtful consideration of who counts, why

they count, and the nature and quality of the relationship and motives to grant or hold resources are keys to understanding stakeholder participation. The central question regarding who counts is to look for interests and who can add value to discussions leading to better alignment of those interests.

At one level of analysis, the key theme of stakeholder participation centers on how various structures and functions can engage productively with the organization's stakeholders so that the organization can succeed, especially in a competitive marketplace where other organizations (with business, nonprofit, or governmental) compete for tangible and intangible resources. As well, power and the dynamics of participation depend on how power resources are defined and interpreted by various stakeholders in ways that give advantage to some players and disadvantage to others.

To determine who and what really counts—and why, Mitchell *et al.* [6] proposed a *normative theory of stakeholder identification*, which can be used “to explain logically why managers should consider certain classes of entities as stakeholders” (p. 853). The other key option, the *descriptive theory of stakeholder salience*, can offer insights “to explain the conditions under which managers do consider certain classes of entities as stakeholders” (p. 853).

Stakeholder analysis coupled to an interest in effective management brings about traditional concerns for effective management philosophies and relationship management as both a problematic for organizational effectiveness as well as the systemic and ethical or normative qualities of society. In essence, effective stakeholder participation not only helps organizations to be more successful but it also can enrich societies by bringing various interests to bear constructively on one another. This conclusion is based on the logic that what advances the interests of one entity at the disadvantage of others may hurt all; and, if each entity works for the mutual good of all of the various societal interests in society, it therefore operates more successfully as a means for collective risk management.

In defining the nature of the relationship, focus can be given to basic resource dependency, independency, or interdependency. For instance, a seller cannot exist without a buyer (as stakeholder). So too, a buyer (as stakeholder) needs a seller (stakeholder). Likewise, the interdependency extends to matters of issues management, the essential tug of war over policies that define the public and private sectors and their interconnectedness. Each entity needs resources held by the other. This paradigm is central to resource dependency theory which focuses attention on the question: What must organization A do to get stakeholder group A to grant stakes held (as resources)?

Stakeholder participation can be dysfunctional as in the case when an organization can achieve a closed-loop exchange by which interdependence becomes dependence; such conditions ultimately foster attempts by the dependent party to achieve independence. Interdependence becomes distorted when buyers are dependent on the seller for income. That model operated in the United States when company stores dominated mining or timbering towns and during the practice of sharecropping. Miners or loggers earned script which could only be used at the company store. As sharecroppers, farmers borrowed against their crops and sold their crops to the owner who was able to keep the farmers in perpetual debt. Thus, in answer to the question that de Bussy [4] raised, this system operated because one participant could control the stakes and the conditions of their exchange. But, is it sustainable?

Stakeholder participation requires knowing who they are and what they seek and hold, the positions they take on key issues. Does that align with or conflict with the organization's interests, positions, and policies? Who or what must change to reduce conflict and advance alignment? How can discourse advance and hinder the willingness and ability to exchange stakes?

ENGAGING STAKEHOLDERS: NEGOTIATING OR PLAYING HARDBALL

Conflict theory and stakeholder theory interconnect by the assumption that as one entity

works to achieve some goal, other entities can exert power resources that help or hinder (control or constrain) that effort. Cooperation and collaboration are characterized by one entity reciprocally granting power resources the other needs (whether voluntarily or under pressure) to achieve its goals. Competition and conflict result from efforts to use resources to constrain, and even defeat the other organization, or person. In this equation, the resources held, the skilled use of the resource, and the ethics that motivate granting or withholding become important factors of engagement.

Game theory presumes a zero-sum outcome where one entity does better at the expense of another. The contrary option is non-zero-sum game. Thus, there could be lose-lose outcomes where the conflict (or game) as played out leads both sides to a net loss. Another option is win-loss. And the optimal strategy is win-win, typically featured as the process of collaboration, joint problem solving where each side seeks to advance its interests but does so by working to advance (rather than deny) the interests of the other entity or entities engaged in power resource endeavors.

Stakeholder participation rests on pragmatism and ethical principle. Stressing the pragmatics of participation, Freeman [2] offered two options: Negotiation or playing hardball. At the heart of stakeholder participation is the identification of stakeholders, the stakes that each stakeholder holds and those sought by each stakeholder, and the conditions (however voluntary) that influence the willingness and ability of each to engage in stake exchanges, however, mutually beneficial. Freeman [2] stressed the importance of featuring the organization as the focal hub surrounded by stakeholders as each of many spokes.

Rowley [11] reasoned that theorists and practitioners should move beyond addressing stakeholder participation as dyadic ties or one-to-one interconnectiveness to determine which stakeholders count and what drives participation with them. "Since stakeholder relationships do not occur in a vacuum of dyadic ties, but rather in a network of influences, a firm's stakeholders are

likely to have direct relationships with one another" [11]. Stakeholder influence is best approached through social network analysis which "accommodates multiple, interdependent stakeholder demands and predicts how organizations respond to the simultaneous influence of multiple stakeholders" (p. 887). To this end, "the density of the stakeholder network surrounding an organization and the organization's centrality in the network influence its degree of resistance to stakeholder demand" (p. 888). By this logic, four archetypal responses are likely depending on pressures related to density: Commander, compromiser, subordinate, and solitary. As an aside, it's worth noting how in the past decade people have become obsessed with terrorism; the role of solitary has become much more salient than it previously was in stakeholder discussions.

This line of reasoning stresses the dynamics operating between a focal organization and its stakeholders as well as those between various or all other stakeholders. At least two factors are made relevant by this kind of analysis. First, stakeholder participation is driven by *collective expectations* of what each stakeholder wants or desires from others, and how each interprets the expectations on it by others in the exchange network. Second, *density* becomes a factor, at least because "dense networks furnish stakeholders with the capacity to monitor the focal organization's actions more efficiently" [11]. *Density* predicts the ease of information exchange between stakeholders and how readily they create and use coalitions. In addition to density, *centrality* features each stakeholder's position in networks and its ability and willingness to relate to and influence, as well as be influenced by others. Centrality predicts an organization's ability to resist or engage stakeholder pressure. By this logic, Rowley advanced the archetype-driven logics that compromisers are in high density/high centrality conditions, whereas commanders are in high centrality but low density. Subordinates are in high density networks but are low in centrality. The solitary is characterized by low density and low centrality.

Negotiation, and even hardball, become stakeholder participation options when relationships are dense and players are central to networks as well as when collective expectations guide willingness and ability to exchange stakes systematically.

Preferring to avoid characterizing conflict as functional/dysfunctional, Behfar and Thompson [12] preferred a quasifunctional framework “that recognizes boundary conditions and the temporal or contextual aspects of group functioning” (p. 3). Their analysis cautioned against stereotyped assumptions about how functions succeed or fail merely because of presumed motives and conditions of engagement. Thus, they recognize that some who are attentive to improving processes of stakeholder engagement see within-industry conflict as grounds for effective problem solving whereas the conflict with issue activist groups (such as those opposing gender bias or environmental policy) is characterized as flawed and dysfunctional. “Whereas the potential of between-group negotiation is highlighted theoretically, groups often fail behaviorally to fully realize their potential. The ‘quasi’ nature of conflict again demands careful procedures and management to determine how the conflict will manifest between groups” (p. 26). Battles between industry and regulatory stakeholders may thrive when high technical expertise prevails over social skills. Those factors conversely are likely to lead to dysfunctional conflict during issue conflicts with nonindustry groups. With those groups, technical risk assessment and management battles without strong social skills may produce dysfunction between companies, governmental agencies, activist publics, and lay publics.

This line of argument becomes even clearer and more operational under the claim that “relationships shape negotiations, and negotiations shape relationships through socially driven and situated definition, enactment, and interpretation of conflict and exchange” [13]. Playing out this theme, McGinn addressed the importance of attributions and emotions, knowledge structures (shared technical and experiential knowledge), procedural rules and norms,

and coordinated actions, and information sharing. These conditions predict in various ways and in context the distribution and integration of resources (stakes) that result in productive or unproductive negotiations and other modes of conflict resolution.

However much biases, emotions, shared (or lack of) knowledge, shared (or conflicting) interpretation schema, objectives, points of reference, and expected utilities drives conflict resolution in general, it is an inherent problematic of stakeholder participation. To these, one can add factors of centrality, density, shared and collective expectations, claimants versus influencers, actual versus potential relationships, and an array of factors of power, dependence, and reciprocity relevant to the conditions and qualities of relationships.

ENGAGING STAKEHOLDERS: PARTICIPATION THROUGH COLLABORATION

Scholarship features descriptive and normative approaches to stakeholder identification and participation. Negotiation arises as key players known to one another exercise their stakes within interdependent relationships. As a normative challenge, the organization may decide to know descriptively who the players are/should be ethically within dependent or independent relationships. Reflecting this normative theme, Gregory *et al.* [14] observed, “Multiparty deliberative processes have become a popular way to increase public participation in public policy choices” (p. 4). Instead of resulting through interdependence, processes may need to be created by organizations seeking to build relationships and resolve issues through collaborative decision making. This logic drove the creation of, for instance, citizen’s advisory councils. Such engagement is both a pragmatic/functional strategy and an ethical choice on the part of organizations engaging in stakeholder participation. The legitimacy of such processes “depends on participant’s ability, first, to understand the issues facing them and, then, to form and express their own positions on them.

These tasks pose significant cognitive and emotional challenges” (p. 4).

Clarkson [15] recommended differentiating stakeholder issues, which are relationship specific, to societal issues which are more generally matters addressed even well beyond the specific points of stakeholder/stakeseeker interaction. Rather than viewing corporate social responsibility as something based on absolute principles, for instance, he proposed “that corporate social performance can be analyzed and evaluated more effectively by using a framework based on the management of a corporation’s relationships with its stakeholders than by using models and methodologies based on concepts concerning corporate social responsibilities and responsiveness” (p. 92).

Stakeholders operate in relationships predicated on the likelihood that stakes are more likely to be willfully exchanged with organizations that meet or exceed their expectations as opposed to organizations that fall short of expectation. These expectations can take the form of preferred positions on issues that might variously be stakeholders’ issues or social issues. These are not independent, but take on idiosyncratic qualities as they are applied to define stakeholder relationships and justify the meeting of expectations. Corporate (social) responsibility, in its many forms, becomes a meeting place for the discussion and enactment of standards of performance that meet expectations and help rather than harm relationships needed for productive stake exchange.

Whereas other aspects of stakeholder engagement presume that stakeholders knowingly hold stakes and are willing and able to employ them as power resources, Arnstein [16] worried four decades ago that community engagement often failed because participation was attempted without a redistribution of power. The era of her concern was that in which ordinary citizens were learning the power resource ropes of community activism. But the caution she raised is important for stakeholder participation best practices today. Can engagement occur because the right people are asked to become engaged and are they or can they

be convinced that engagement counts? Her ladder of citizen participation postulated citizen power through citizen control, delegated power, and partnership. Tokenism results in the presence of placation, consultation, and information (especially, one might add, when risk communication entails data dumps or discourse involving highly technical knowledge). Lastly, she postulated that nonparticipation resulted when engagement was seen as therapy or therapy was substituted for engagement and when publics were manipulated or coopted. In the lower rungs on the ladder, support might be engineered rather than being the genuine result of ethical and functional systems.

Considering definable targets of engagement, McComas *et al.* [17] noted that complex decisions based on highly technical information can pit those who understand such topics against those who worry about their risk level and who challenge their more knowledgeable adversaries’ right to make decisions that defy public understanding. How can concerned citizens mature into policy makers (recall the ladder of citizen participation)? Strategies may differ when engaging in stakeholder participation with those who believe they have the knowledge and ethical judgment to work for negotiation, and even play hardball.

McComas *et al.* suggested a four-step decision process should begin by finding concerned citizens or those who can become concerned citizens before their opinions are tightly held. With both general types of citizens, engagement might also introduce other discussants who not only understand and facilitate the discourse but also appreciate the value and interpretative disparities that might be at play. Efforts and commitments can be made to foster understanding and seek information that is needed to address concerns and might not be readily available to the collaboration. The second step requires highlighting “the importance of thinking carefully about the problem at the forefront of the consultation” (p. 371). “Helping participants to distinguish between *means* and *ends* objectives also allows for greater creativity in identifying alternatives for evaluation in decision making” (p. 372). This moment in stakeholder participation can be troubling

for those organizations whose intention with stakeholder participation is to engineer consent for their preferred solution (recall the ladder of citizen participation). Finally, one key ingredient in effective stakeholder participation is to select alternatives and address trade-offs as a means for discovering alignment of interests that lead to productive stake exchanges. "This analysis is conducted by carefully analyzing the predicted consequences of each alternative against the participants' stated objectives" (p. 373). In such collaboration, procedural justice matters not only as a means for evaluating the quality of participation but as motivation for its continuation. In such matters, there can be rational, socioeconomic, and relational incentives.

Confronted with its mission/vision and traditional sets of operations, organizations can encounter their own mind set as a hindrance to collaborative stakeholder engagement. If stakes are expressions of values and interests, a multiple-objective decision analysis is required to achieve effective stakeholder participation [18].

CONCLUSION

Discussion of stakeholder participation can be approached in descriptive, normative, instrumental, and managerial fashions [19]. From whatever perspective, stakeholder participation raises a universal concern: the challenge to determine what factors shape the quality of relationships between and among various organizations, corporations, and governmental agencies and the conditions of exchange. As Donaldson and Preston [19] advised, "Stakeholder management requires, as its key attribute, simultaneous attention to the legitimate interests of all appropriate stakeholders, both in the establishment of organizational structures and general policies and in case-by-case decision making" (p. 67).

Featuring a dynamic conception of management and operations, Mumby [20] reasoned, "Organizations are not stable, fully integrated structures. Rather, they are the product of various groups with competing

goals and interests. An organization services a group's interests to the extent that it is able to produce, maintain, and reproduce those organizational practices that sustain that group's needs" (p.166). The logic of that statement can be clarified and expanded by noting that the satisfaction of a group's needs interlock and align with the satisfactions of multiple, networked, and rationalized groups' needs, wants, and interests. By this logic, markets and public policies result from the structures of stake exchange and the meaning that makes them enactable interests.

REFERENCES

1. Renn O, Webler T, Wiedemann P. Foreword. In: Renn O, Webler T, Wiedemann P, editors. *Fairness and competence in citizen participation: evaluation models for environmental discourse*. Dordrecht: Kluwer Academic Publishers; 1995. pp. xiii–xixv.
2. Freeman RE. *Strategic management: a stakeholder approach*. Boston (MA): Pitman; 1984.
3. Kahneman D, Slovic P, Tversky A. *Judgment under uncertainty: heuristics and biases*. New York: Cambridge University Press; 1982.
4. de Bussy N. Stakeholder theory. In: Donsbach W, editor. *International encyclopedia of communication*. Malden (MA): Blackwell Publishing; 2008. pp. 4815–4817.
5. de Bussy N. Applying stakeholder thinking to public relations: an integrated approach to identifying relationships that matter. In: van Ruler B, Vercic AT, Vercic D, editors. *Public relations metrics: research and evaluation*. New York: Routledge; 2005. pp. 282–300.
6. Mitchell RK, Angle BR, Wood DJ. Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. *Acad Manag Rev* 1997;22:853–886.
7. Heath RL, Palenchar MJ. *Strategic issues management: organizations and public policy challenges*. 2nd ed. Thousand Oaks (CA): Sage Publications; 2009.
8. Ulmer RR, Seeger MW, Sellnow TL. Stakeholder theory. In: Heath RL, editor. *Encyclopedia of public relations*. Thousand Oaks (CA): Sage Publications; 2004. pp. 808–811.
9. Freeman RE, Gilbert DR Jr. *Corporate strategy and the search for ethics*. Englewood Cliffs (NJ): Prentice Hall; 1988.

10. Pate-Cornell ME, Dillon RL. The respective roles of risk and decision analyses in decision support. *Decis Anal* 2006;3:220–232.
11. Rowley TJ. Moving beyond dyadic ties: a network theory of stakeholder influences. *Acad Manag Rev* 1997;22:887–910.
12. Behfar KJ, Thompson LL. Conflict within and between organizational groups: functional, dysfunctional, and quasifunctional perspectives. In: Behfar KJ, Thompson LL, editors. *Conflict in organizational groups: new directions in theory and practice*. Evanston (IL): Northwestern University Press; 2007. pp. 3–35.
13. McGinn KL. Relationships and negotiations in context. In: Thompson LL, editor. *Negotiation theory and research*. New York: Psychology Press; 2006. pp. 129–143.
14. Gregory R, Fischhoff B, Daniels T. Acceptable input: using decision analysis to guide public policy deliberations. *Decis Anal* 2005;2(1):4–16.
15. Clarkson MBE. A stakeholder framework for analyzing and evaluating corporate social performance. *Acad Manag Rev* 1995;20:92–117.
16. Arnstein SR. A ladder of citizen participation. *J Am Inst Plann* 1969;35/4:216–224.
17. McComas KA, Arvai J, Besley JC. Linking public participation and decision making through risk communication. In: Heath RL, O'Hair HD, editors. *Handbook of risk and crisis communication*. New York: Routledge; 2009. pp. 364–385.
18. Merrick JRW, Parnell GS, Barnett J, *et al.* A multiple-objective decision analysis of stakeholder values to identify watershed improvement needs. *Decis Anal* 2005;2:44–57.
19. Donaldson T, Preston LE. The stakeholder theory of the corporation: concepts, evidence, and implication. *Acad Manag Rev* 1995;20:65–91.
20. Mumby DK. *Communication and power in organizations: discourse, ideology, and domination*. Norwood (NJ): Ablex; 1988.

STANDBY REDUNDANT SYSTEMS

SALVATORE DISTEFANO
Department of Mathematics,
University of Messina,
Messina, Italy

REDUNDANCY

“Redundancy (in reliability, maintainability, and availability) is the existence of a means in addition to the means which would be sufficient for a functional unit to perform a required function or, for data, to represent information” [1]. Redundancy is used primarily to improve reliability, availability, and/or performance. Reliability and performance requirements are often in contrast: redundancy goals focus on implementing *fault tolerance*, that is *“the ability of a system or unit to continue normal operation despite the presence of hardware or software faults”* [2], while performance requirements mainly involve reduction of the system response time and optimization–maximization of the resources/components throughput and utilization. Thus, in the former case, in order to optimize the system reliability, redundant resources/components, units or subsystems are managed by applying specific redundancy policies, by which few units are operating ones and the others act as spares (*active/standby redundancy*). While, in case of performance redundancy, all the redundant components are operating, working in parallel and thus, sharing the workload (*load sharing*).

There are essentially two kinds of redundancy techniques employed in fault-tolerant designs, *space redundancy* and *time redundancy*. Space redundancy provides separate physical copies of a resource, function, or data item. Time redundancy, used primarily in digital systems, involves the process of storing information to handle transients, or encoding information that is shifted in time

to check for unwanted changes. Space, or hardware redundancy (hereafter only redundancy), is the approach most commonly associated with fault-tolerant design.

Redundancy does not just imply a duplication of hardware because it can also often be implemented by coding, at the software level. However, to avoid common mode failures, redundant elements should be realized (designed and produced) independently from each other. Irrespective of the failure mode (e.g. shorts or opens), redundancy still appears in parallel on the reliability block diagram.

From the operating point of view, one distinguishes between active and standby redundancy [3,4]:

- *Active Redundancy*. Redundant elements are subjected from the beginning to the same load as operating elements and load sharing is possible. External components are not required to perform the function of detection, decision, and switching when an element or path in the structure fails. The redundant units are always operating and automatically pick up the load for a failed unit.
- *Standby Redundancy*. Redundant elements can be energized or not until one of the operating elements fail, no load sharing is possible; failure rate in the reserve state can be lower than in the operating state or also assumed to be zero. External elements are required to detect, make a decision, and switch to another element or path as a replacement for a failed element or path.

Figure 1 introduces an overview of the most important techniques developed in order to achieve redundancy. The distinction between Active and Standby redundancy is there clearly depicted, even if the *k-out-of-n* redundancy technique can be classified as hybrid, implementing aspects of both active and standby redundancy.

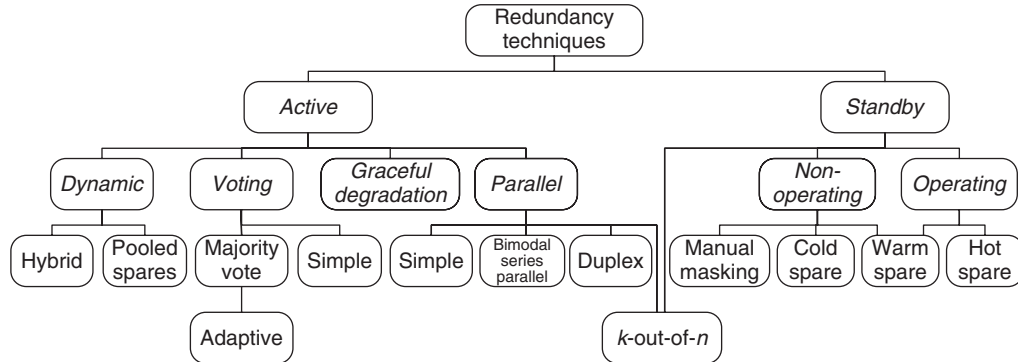


Figure 1. Redundancy technique taxonomy.

According to this classification, *standby redundant systems* are those systems composed of two or more units/components/sub-systems with similar functionalities and that each of them can substitute one of the other redundant units in the system functioning, acting as a spare. The main characteristic of standby redundant systems is the *standby policy* applied to manage the redundant units, usually with the aim to maximize/optimize the system reliability/availability, and/or to reduce the down time.

In order to better investigate standby redundancy, an interesting characterization of standby redundant systems reliability and availability that can be useful in the following is made in Ref. 5, distinguishing among three classes:

- **Basic Reliability/Availability.** A system which is designed, implemented, and deployed with sufficient components (hardware, software, and procedures) to satisfy the system's functional requirements, has basic reliability/availability. Such a system will deliver the correct functionality so long as no faults occur and no maintenance operations are performed. Whenever a fault occurs or a maintenance operation is performed, however, an outage may be observed. Typically, basic reliability/availability systems are deployed as simplex (nonreplicated) systems.

- **High Reliability/Availability.** A system which is designed, implemented, and deployed with sufficient components to satisfy the system's functional requirements, but which also has sufficient redundancy in components (hardware, software, and procedures) to mask certain defined faults, has high reliability/availability.

"Sufficient" and "certain" are terms related to the masking concept. "Masking" a fault implies shielding the external observation of the failure, this does not ensure that the fault will not occur, but that it is unobservable. This is invariably achieved through a replication mechanism appropriate to the component, a redundancy strategy. When a component fails, the redundant component replaces it. In this context, sufficient redundancy implies that hardware, software, and procedure faults must all be masked; that is, adequate replication mechanisms or redundancy strategies have to be implemented, also being aware that certain defined faults can be covered, since not all faults can be masked.

The degree of "transparency" in which this replacement occurs can lead to a wide variation of systems which call themselves high reliability/availability, that can be identified as the following:

- **Hot Standby.** Active and redundant components are tightly coupled in

groups and (logically) indistinguishable to the users of the component group. Indeed, the user does not “see” the group, but only the component behavior it requires. Following a component fault, users of the component are not disconnected and do not observe the fault in any way.

- *Warm Standby.* Following a component fault, the redundant system is unreliable/unavailable till a warm standby spare (partially loaded/energized) becomes fully operating. Users of the component are disconnected and may lose *some* of their work in progress. An automatic fault detection and recovery mechanism detects the fault, and notifies the redundant component to take over. This redundant component has been actively running, and is partially initialized.
- *Cold Standby.* Following a component fault, the redundant system becomes unreliable/unavailable till a cold standby spare (starting from scratch) becomes operating. Users of the component are disconnected and lose *any* work in progress. An automatic fault detection and recovery mechanism then detects the fault, and brings into service the redundant component.
- *Manual Masking.* Some manual action is required to put the redundant component into service after a fault, during which time the system is unavailable for use. It could be considered as a standby management policy as above with a manual fault detection and recovery mechanism.
- *Continuous Reliability/Availability.* This extends the definition of high reliability/availability and applies it to planned outages as well. High reliability/availability, as defined above, compensates for unplanned outages. Continuously reliable/available systems must also have a masking strategy that deals with planned outages.

High availability identifies and classifies four *standby policies* for managing standby redundant systems. Automatic techniques, implementing hot, warm, and cold standby policies, are of particular interest in the specific literature and thus are studied in depth in the following sections.

From the reliability point of view a component in hot standby could be considered as active, since it is always characterized by the same reliability CDF, also after changing from standby to active and vice versa. This behavior needs further investigation. In order to do that, let us compare a simple active redundant system composed of two independent and identically distributed units connected in parallel, against a similar configuration in which, instead, a hot standby redundancy policy is applied. It is possible to observe that, in general, a significant difference between the two redundant systems exists: in the former case, the load can be shared between the two units, while a hot standby spare replicates the primary unit task or is fully synchronized to this latter in order to be ready to substitute it in case of failure. Thus, in the standby redundant configuration, the load is not shared among the redundant units; the hot standby unit is kept readily available in order to minimize, ideally to annul, the downtime. If the primary unit implements a stateful process, to achieve the goal of hot standby becomes more complex since it is necessary to constantly refresh the state of the hot standby unit. With this aim, mainly referring to software contexts, specific techniques, grouped together as *active replication*, have been implemented.

The term *active replication* is used to describe a set of techniques used to achieve hot standby, by actively and completely sharing processing state between replicas. In such cases, sharing state implies not only transferring information between replicas, but also coordinating and synchronizing the information. In fact, “state” is a dynamic entity in a distributed system, and changes frequently. Replicas present their state externally, and process it internally, in a consistent manner. “Synchronizing” a system’s state really means ensuring that operations are applied in the same order to all participant replicas.

Active replication techniques are thus largely concerned with ordering operations (an example is the virtual synchrony model [6]). The various active replication techniques available (causal ordering, total ordering, etc.) are attempts to trade-off more or less rigorous ordering guarantees with their respective performance overheads (in general, the more rigorous the order, the more degraded is the performance). In addition to providing an ordered communications vehicle, active replication systems need mechanisms to support the notions of “groups” of replicas, and have membership services which identify, in a distributed sense, which processes are members of a group at any time. Mechanisms are needed to allow members to join a group and “catch up” with the processing state of the group, and to be removed from a group either voluntarily or in response to a detected failure in a member. Hence, active replication systems invariably include failure detection and heartbeat mechanisms as part of their group membership services. Active replication is frequently associated with the notion of “make forward progress” systems, since by maintaining consistent server replicas, clients are not interrupted in the event of a failure, and are always able to advance the state of their work.

Passive or lazy replication describes a set of techniques used to achieve warm standby. Recall that warm standby was defined as a replicated system in which some amount of initialization and state sharing occurs between replicas. In practice, warm standby solutions are often used as a relaxed form of “hot standby” high availability, because active replication can be quite costly in terms of performance and system design complexity. The amount of “relaxation” from the stringent active replication varies, however, so that there are many approaches which can legitimately call themselves warm standby, but which have widely differing amounts of initialization and state sharing. Lazy replication is frequently associated with the notion of “Rollback & Recover” systems, since even with some state maintenance, clients are interrupted in the event of a failure, and must explicitly rebound to the standby.

STANDBY REDUNDANT SYSTEMS

A system is said to have units which are reliability-wise in standby when there is an active unit or subsystem to which are attached units or subsystems which stand by idly during the mission either in a quiescent/nonoperating or warm-up mode, until they are called upon to operate. A sensing subsystem has to detect a failure in the active unit or subsystem and to command the switching subsystem to switch in the standby unit or subsystem. This is unlike the reliability-wise parallel system in which all units start to operate simultaneously at the beginning of the mission, and the system succeeds when at least one unit completes the mission successfully.

Summarizing, a standby redundant system can be composed of three main parts as shown in Fig. 2:

- *Redundant Subsystems, Connected in Parallel.* In this, one or more of them are operating (e.g., by applying specific k -out-of- m voting mechanism), the others are in standby, acting as spare units that substitute the operating ones on faults in order to ensure the redundant system reliability/availability.
- *Switching Subsystems.* A switch for each spare is necessary in order to transition the spare from the standby to the operating state. Such subsystems contain both, the switch device (switches, valves, etc.) and the switching actuator.
- *Monitoring–Sensing Subsystem.* A sensor monitoring the standby redundant system is necessary. It drives the corresponding switch, triggering the switching actuators.

In order to provide a quantitative evaluation of a standby redundant system, we investigate the reliability of a two-unit standby configuration as the one shown in Fig. 2. Let us assume a warm standby policy is applied to S2, so that it is in a warm standby till unit S1 is operating. Following the reasoning applied in Ref. 7, such a system succeeds in its mission with mission time T , when

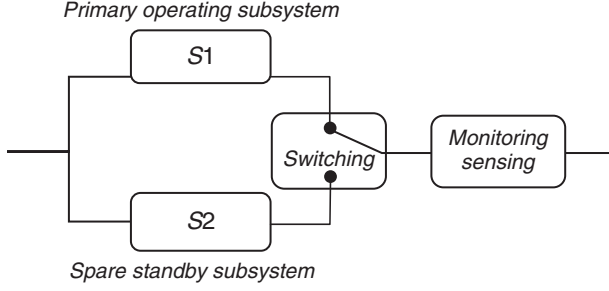


Figure 2. Two-unit standby redundant system.

1. the active unit does not fail ($R_1(T)$) and the sensing subsystem does not fail ($R_{SE}(T)$) and the switch does not fail open during the mission ($R_{SWO}(T)$), in total $R_{Ev1}(T) = R_1(T) R_{SE}(T) R_{SWO}(T)$, or
2. the active unit fails before the end of the mission at $0 < T_1 \leq T$ ($\int_0^T f_1(T_1) dT_1$; $f_1(T) = -dR_1(T)/dT$) and the sensing ($R_{SE}(T_1)$) and switching subsystems do not fail in both open ($R_{SWO}(T)$) and close ($R_{SWC}(T)$); moreover, the switch ($R_{SWC}(T_1)R_{SWO}(T)$) and the actuator ($R_{ACT}(T_1)$), and the standby unit, not having already failed, succeeds for the remainder of the mission.

Thus, from Ref. 7, we obtain that $R_{Ev2}(T) = \int_0^T f_1(T_1) R_{SE}(T_1) R_{SWO}(T) R_{SWC}(T_1) R_{ACT}(T_1) R_{20}(R_{20}^{-1}(R_{2S}(T_1)) + T - T_1) dT_1$ and therefore, the overall reliability of the two-unit standby redundant system is

$$\begin{aligned}
 R_{2SRS}(T) &= R_{Ev1}(T) + R_{Ev2}(T) \\
 &= R_1(T) R_{SE}(T) R_{SWO}(T) \\
 &\quad + R_{SWO}(T) \int_0^T f_1(T_1) R_{SE}(T_1) \\
 &\quad \times R_{SWC}(T_1) R_{ACT}(T_1) R_{20} \\
 &\quad \times (R_{20}^{-1}(R_{2S}(T_1)) + T - T_1) dT_1.
 \end{aligned} \tag{1}$$

In the hot standby case, $R_{20}(T) = R_{2S}(T) = R_2(T)$ Equation (1) becomes

$$\begin{aligned}
 R_{2SRS}^H(T) &= R_1(T) R_{SE}(T) R_{SWO}(T) \\
 &\quad + R_2(T) R_{SWO}(T) \int_0^T f_1(T_1) R_{SE}(T_1) \\
 &\quad \times R_{SWC}(T_1) R_{ACT}(T_1) dT_1.
 \end{aligned} \tag{2}$$

In the cold standby case, $R_{2S}(T) = 1$ and $R_{20}(T) = R_2(T)$ and therefore,

$$\begin{aligned}
 R_{2SRS}^C(T) &= R_1(T) R_{SE}(T) R_{SWO}(T) + R_{SWO}(T) \\
 &\quad \times \int_0^T f_1(T_1) R_{SE}(T_1) R_{SWC}(T_1) \\
 &\quad \times R_{ACT}(T_1) R_2(T - T_1) dT_1.
 \end{aligned} \tag{3}$$

Equation (1) can be simplified by making the assumption that the switching and the sensing subsystems cannot fail. In this case, Equations (1), (2), and (3) become respectively:

$$\begin{aligned}
 R_{2SRS}(T) &= R_1(T) + \int_0^T f_1(T_1) R_{20} \\
 &\quad \times (R_{20}^{-1}(R_{2S}(T_1)) + T - T_1) dT_1,
 \end{aligned} \tag{4}$$

$$\begin{aligned}
 R_{2SRS}^H(T) &= R_1(T) + R_2(T) \int_0^T f_1(T_1) dT_1 \\
 &= R_1(T) + R_2(T)(1 - R_1(T)),
 \end{aligned} \tag{5}$$

$$\begin{aligned}
 R_{2SRS}^C(T) &= R_1(T) + \int_0^T f_1(T_1) R_2(T - T_1) dT_1 \\
 &= R_1(T) + f_1(T) * R_2(T),
 \end{aligned} \tag{6}$$

where $*$ is the symbol of convolution. Equations (5) and (6) can be generalized to a standby redundant system composed of $n - 1$ spare components (n in total):

$$\begin{aligned}
 R_{n\text{SRS}}^H(T) &= R_1(T) + R_2(T)(1 - R_1(T)) + \dots \\
 &\quad + R_n(T) \prod_{i=1}^{n-1} (1 - R_i(T)), \quad (7) \\
 R_{n\text{SRS}}^C(T) &= R_1(T) + f_1(T) * R_2(T) + \dots \\
 &\quad + (((f_1(T) * R_2(T))' * R_3(T))' * \dots)' \\
 &\quad * R_n(T) = R_1(T) + f_1(T) * R_2(T) \\
 &\quad + \dots + f_1(T) * R_2(T) * (-f_3(T)) \\
 &\quad * \dots * (-f_n(T)). \quad (8)
 \end{aligned}$$

The terms of Equation (8) are obtained as a generalization of Equation (6) by recursively applying this latter at each failure. Moreover, the final formula is obtained by exploiting the derivative convolution rule establishing that $d(f * g)/dx = df/dx * g = f * dg/dx$.

The exponential distribution is often used in reliability theory because of its mathematical tractability. In case all the components, switches, and actuators of the system are characterized by exponentially distributed reliability, Equation (1) becomes

$$\begin{aligned}
 R_{2\text{SRS}}(T) &= e^{-(\lambda_1 + \lambda_{\text{SE}} + \lambda_{\text{SWO}})T} + e^{-\lambda_{\text{SWO}}T} \\
 &\quad \times \int_0^T \lambda_1 e^{-(\lambda_1 + \lambda_{\text{SE}} + \lambda_{\text{SWC}} + \lambda_{\text{ACT}})T_1} \\
 &\quad \times e^{-\lambda_{2\text{O}}((\lambda_{2\text{S}} - \lambda_{2\text{O}})/\lambda_{2\text{O}})(T_1 + T)} dT_1 \\
 &= e^{-(\lambda_1 + \lambda_{\text{SE}} + \lambda_{\text{SWO}})T} \\
 &\quad + \frac{\lambda_1 (e^{-(\lambda_{\text{SWO}} + \lambda_{2\text{O}})T} - e^{-(\lambda_1 + \lambda_{\text{SE}} + \lambda_{\text{SWO}} + \lambda_{\text{SWC}} + \lambda_{\text{ACT}} + \lambda_{2\text{S}})T})}{\lambda_1 + \lambda_{\text{SE}} + \lambda_{\text{SWC}} + \lambda_{\text{ACT}} + \lambda_{2\text{S}} - \lambda_{2\text{O}}}. \quad (9)
 \end{aligned}$$

In case of perfect switching and sensing systems, $\lambda_{\text{SE}}, \lambda_{\text{SWC}}, \lambda_{\text{SWO}}, \lambda_{\text{ACT}} = 0$, and thus $R_{2\text{SRS}}(T) = e^{-\lambda_1 T} + \lambda_1 (e^{-\lambda_{2\text{O}} T} - e^{-(\lambda_1 + \lambda_{2\text{S}})T}) / (\lambda_1 + \lambda_{2\text{S}} - \lambda_{2\text{O}})$, that in case of cold standby ($\lambda_{2\text{S}} = 0, \lambda_{2\text{O}} = \lambda_2$) is $R_{2\text{SRS}}^C(T) = e^{-\lambda_1 T} + \lambda_1 (e^{-\lambda_2 T} - e^{-\lambda_1 T}) / (\lambda_1 - \lambda_2)$. If $\lambda_1 = \lambda_2 = \lambda$, such a formula is indeterminate, but

from Equation (6), it has $R_{2\text{SRS}}^C(T) = e^{-\lambda T} + \int_0^T \lambda e^{-\lambda T} dT_1 = e^{-\lambda T} + \lambda e^{-\lambda T} T = e^{-\lambda T} (1 + \lambda T)$. Such a formula can be generalized for a system composed of $n + 1$ redundant elements. A similar standby system can be considered as a system composed of a single unit that can tolerate n consecutive faults before fails. The n th fault will drive to the failure of the system. In this way, the reliability of a system tolerating $n - 1$ faults is [8,9]

$$\begin{aligned}
 R_{n\text{SRS}}^C(T) &= e^{-\lambda T} \left(1 + \lambda T + \frac{(\lambda T)^2}{2!} + \dots + \frac{(\lambda T)^{n-1}}{(n-1)!} \right) \\
 &= e^{-\lambda T} \sum_{i=0}^{n-1} \frac{(\lambda T)^i}{i!}. \quad (10)
 \end{aligned}$$

STANDBY REDUNDANT SYSTEMS MANAGEMENT

Once the main components and parts of a standby redundant system have been identified, along with the way, the management policies that can be applied to spares (cold, warm, hot), specific configurations, techniques and strategies to apply to the standby redundant system in order to improve its reliability and availability are investigated. In particular, issues related to *k-out-of-n* standby redundancy will be studied in depth while aspects such as repairability, standby priority, and optimization will be briefly discussed/outlined due to the extent of such topics.

k-out-of-n Redundancy

A problem that may arise in such systems regards the monitoring and sensing subsystem in the making of a decision: how to establish the correctness of a response? A possible solution is the *voting redundancy* technique [10]. Assume that the same input is applied to two identical digital elements and compare the outputs. If each bit agrees, then either they are both working properly (likely), or they have both failed in an identical manner (unlikely). If a difference between the

two outputs is detected, then there is an error, although it is not possible to establish which element is in error. If a third unit is taken into consideration, in case all three outputs agree, then either all three are working properly (most likely) or all three have failed in the same manner (most unlikely). If two of the element outputs agree, then most likely the element with a different output has failed and the redundant system can rely on the output of the other two elements. Thus with three elements, it is possible to correct one error. If two errors have occurred, it is very possible that they will fail in the same manner, and the comparison will agree (vote along) with the majority. By extending such an idea to N elements, the so-called *N-modular redundancy* technique is obtained. *Majority redundancy* is a special case of a *k-out-of-n* redundancy, used primarily in digital circuits.

Generally, a system consisting of $n \geq 2$ components or subsystems, of which only $1 \leq k \leq n$ need to be functioning for system success, is called a *k-out-of-n* configuration. More specifically, such a configuration is referred as *k-out-of-n:G* (G stands for good) or, alternatively, as *(n-k)-out-of-n:B* (B as bad). In the former case (*k-out-of-n:G*), when $k = n$, the system has a series structure, whereas for $k = 1$, it reduces to a parallel structure, vice versa for *k-out-of-n:B* systems. Thus, a *k-out-of-n* system is more general than either a series or a parallel structure. Hence, the result for this system holds for both series and parallel structures. It is important to remark that all the components of a *k-out-of-n* system are initially operating. Moreover, assuming that all the components of the *k-out-of-n* system have independent and identically distributed lifetime with CDF $F(t)$, pdf $f(t)$ and reliability $R(t) = 1 - F(t)$, the *k-out-of-n* system reliability is

$$R_S(t) = \begin{cases} 1 - F^n(t), & k = 1; \\ \sum_{i=k}^n \binom{n}{i} R^i(t) F^{n-i}(t), & k \geq 1. \end{cases} \quad (11)$$

Equation (11) is obtained by expanding $(R(t) + F(t))^n$. $k = 1$ identifies parallel

systems ($R_s(t) = 1 - F^n(t)$), while $k = n$ corresponds to series ($R_s(t) = R^n(t)$).

As stated above, majority redundancy is a special case of *k-out-of-n* systems, in which $2n + 1$ outputs are fed to a voter whose output represents the majority of its $2n + 1$ input signals. The majority redundancy realizes in a very simple way a fault-tolerant structure without the need for control or switching elements, the required function is performed without interruption until a second failure occurs, the first failure will automatically be masked by the majority redundancy.

The *k-out-of-n* redundancy can be considered as a particular policy that mixes active redundancy for some subsystems and standby redundancy for other subsystems. As in *k-out-of-n* systems, in a *k-out-of-n* redundant configuration (*k-out-of-n:G*), k components are necessary for the system to provide the required function and $n - k$ stay in the reserve-standby state acting as spares, vice versa in *k-out-of-n:B*. Each time a component among the k operating ones fails, a spare is activated, that is switched from standby to operating mode, until there are spares available. When no more spare components are available, the *k-out-of-n* standby redundant system fails.

Assuming, as above, the n lifetime components identically distributed (no more independent due to the standby) having a common CDF $F(t)$, pdf $f(t)$, and reliability $R(t) = 1 - F(t)$, and moreover, ideal failure detection and switching, two different standby are investigated: hot and cold. In the former case, the reliability of a *k-out-of-n* hot standby redundant system corresponds to the one reported in Equation (11), since no changes in the reliability CDF occur for the standby spares at changing points.

More complex is the case of cold standby. A way to solve such a problem is to view it as a particular case of the cold standby system, and thus apply Equation (8). Following such reasoning, it is necessary to identify as primary subsystem of the *k-out-of-n* cold standby redundant system, the series of the k operating components, while the others $n - k$ components are cold spares. In this way, due to the assumption of identically distributed

components, Equation (8) becomes

$$\begin{aligned}
R_{k/n\text{SRS}}^C(T) &= R_1(T) + f_1(T) * R_2(T) + \dots \\
&\quad + f_1(T) * (-R_2(T)) * (-f_3(T)) \\
&\quad * \dots * (-f_n(T)) \\
&= R(T)^k - (1 - R(T)^k) * f(T) + \dots \\
&\quad + (1 - R(T)^k) * (-f(T)) * \dots * \\
&\quad (-f(T)) \\
&= R(T)^k + \sum_{i=k+1}^n (-1)^{i-k} (1 - R(T)^k) \\
&\quad * f(T)^{*i-k}. \tag{12}
\end{aligned}$$

In case of exponentially distributed lifetime components, referring to the equations in Ref. 11 for the i -fold convolution power of exponential pdf $f(T)^{*i} = \lambda(\lambda T)^{i-1}e^{-\lambda T}/(i-1)!$, Equation (12) becomes

$$\begin{aligned}
R_{k/n\text{SRS}}^C(T) &= e^{-k\lambda T} + \sum_{i=k+1}^n (-1)^{i-k} (1 - e^{-k\lambda T}) \\
&\quad * \frac{\lambda(\lambda T)^{i-k-1}e^{-\lambda T}}{(i-k-1)!} = e^{-k\lambda T} + \sum_{i=k+1}^n \\
&\quad \times \left((-1)^{i-k} \left(1 - e^{-\lambda T} \sum_{j=0}^{i-k-1} \frac{(\lambda T)^j}{j!} \right) - (k-1)^{i-k} \right. \\
&\quad \times \left. \left(e^{-k\lambda T} - E^{-\lambda T} \sum_{j=0}^{i-k-1} \frac{((1-k)\lambda T)^j}{j!} \right) \right). \tag{13}
\end{aligned}$$

Repairability

Standby redundant systems can be broadly classified as having nonrepairable or repairable elements [12]. System reliability can be improved by having repairable elements. Maintenance can be either corrective or preventive; corrective maintenance occurs only after the failure of an element. On the other hand, if maintenance and repair are performed before an element fails, then the unit undergoes preventive maintenance. This classification holds for scheduled preventive work as well as for maintenance that is performed as a result of inspection. A good reference on the topic is Ref. 12, that though dating back to 1986, provides an actual,

valid, and interesting classification of the specific topic.

With regard to repairable standby systems, an interesting aspect to consider is *priority management*. The concept of priority in standby redundancy is mainly used to introduce a distinction between the primary/working units and the spare/standby ones: when a primary unit is repaired, the spares activated in substitution of such primary unit, due to the priority of this latter, come back into standby. Otherwise, if priorities are not specified, the primary could become a spare of the operating units, also of the standby ones substituting it.

In Ref. 13, a redundant system consisting of a repairable priority unit and a standby unit is considered. Usually, the priority unit performs the desired system function. On failure of the priority unit, the standby unit is switched into service. Meanwhile, the priority unit is repaired. As soon as repair of the priority unit is completed, it is switched back into service. Priority standby can be contrasted with usual standby with repair, which has been analyzed by Gnedenko [14] and Srinivasan [15]. In usual standby, switchover from one unit to the other does not occur until the operating unit fails, irrespective of which unit is operating. Gnedenko and Srinivasan used simple renewal theory arguments and obtained Laplace–Stieltjes transforms of the time between units. The preemptive priority results are relatively insensitive to the form of successive switchovers. From these transforms, the availability distributions are obtained.

More recently, in Ref. 16, Goel and Kuma Tyagi considered and studied in depth, a priority redundant system with two dissimilar units and three modes: one of the units has a priority operative mode and the other has a priority repair mode. Another interesting work on the topic has been developed by Vanderperre and Makhanov in Ref. 17. In this, the authors consider a two-unit cold standby system attended by two repairmen and subjected to a priority rule. In order to describe the random behavior of the twin system, they employ a stochastic process endowed with state probability functions satisfying coupled Hokstad-type differential equations. Since an

explicit evaluation of the exact solution is in general quite intricate, they proposed a numerical solution of the equations to analyze the long-run availability of the T-system.

Optimization

Another challenge/issue of great importance in standby redundancy and standby redundant system, is the optimization. Several articles address either the provision of redundant components in parallel or the combination of structural redundancy with the enhancement of component reliability. These are called *redundancy allocation problems* and *reliability redundancy allocation problems*, respectively.

The redundancy allocation problem involves the simultaneous selection of components and a system-level design configuration, which can collectively meet all design constraints in order to optimize some objective functions such as system cost and/or reliability [18]. In this problem, there are several different component types with different levels of cost, reliability, weight, and other characteristics, and the components redundant within the subsystem are of the same type. Redundancy allocation problems are well documented in Refs 19 and 20, which employ a special case of reliability redundancy allocation problems without exploring the alternatives of component-combined improvement.

Recently, much of the effort in optimal reliability design has been placed on general reliability redundancy allocation problems, rather than redundancy allocation problems. In practice, only two redundancy schemes are mainly considered: active and cold standby. As discussed above, cold standby redundancy provides higher reliability, but it is hard to implement due to the difficulty of failure detection. Reliability design problems have generally been formulated considering only active redundancy; however, an actual optimal design may include active redundancy or cold standby redundancy or both. However, any effort for improvement usually requires resources. Quite often, it is hard for a single objective to adequately describe a real problem for which an optimal design is required.

For this reason, multiobjective system design problems always deserve a lot of attention.

Optimal reliability design problems are known to be NP-hard [21]. Thus, finding efficient optimization algorithms is always a hot spot in this field. In the past few decades, optimization problems including those of redundancy allocation and reliability apportionment, have been treated by many researchers as nonlinear integer programming (nIP) problems with one or several resource constraints. Furthermore, the practicality in recent process synthesis and process optimization of highly reliable systems leads to an extended system reliability design with the consideration of finding the optimum levels of component reliability and the number of redundant components at each subsystem level. This problem, also identified as optimal reliability assignment/redundant allocation problem, is more difficult than reliability apportionment or redundancy allocation problems, because it should be simultaneously determined by two different types of decision variables.

During the past two decades, numerous reliability design techniques have been proposed. These techniques can be classified as *implicit enumeration*, *dynamic programming*, *branch-and-bound method*, *linear programming*, *Lagrangian multiplier method*, *heuristic methods*, and so on. Tillman *et al.* [22] and Kuo *et al.* [20] have well reviewed the field of optimization techniques for system reliability design in their books. Dhingra [23] researched the reliability apportionment problem for a series system with time-dependent reliability. Hikita *et al.* [24] presented an application of the surrogate-constraints algorithm for an optimal reliability assignment/redundant allocation problem. Jacobson and Arora [25] employed a method combining the simplex searching method and the heuristic approach for a similar problem with the consideration of different starting values. Shelokar *et al.* [26] recently developed an ant colony algorithm to solve both continuous function and combinatorial optimization problems in reliability engineering.

Most of these techniques transform the original problems formulated as an NIP problem into a 0–1 linear programming problem. However, this transformation leads the original problems to an increase in the number of variables and constraints to be treated. This requires much more computation time and larger computation memory space so that researchers in the reliability optimization field have placed more emphasis on the heuristic methods, which yield reasonably good solutions with little computation, even though heuristic methods have some weaknesses in several specific optimization problems. Genetic algorithm (GA) is one of metaheuristic optimization techniques, which include simulated annealing, tabu search, and evolutionary strategies.

GA has been demonstrated to converge to the optimal solution for many diverse and difficult problems as a powerful and stochastic tool, based on principles of natural evolution. For specific engineering problems, however, GA sometimes can suffer from the premature convergence situation of its solution because of its fundamental requirements, that is, not using *a priori* knowledge and not exploiting local search information. Also, it is necessary to adequately tune the unknown GA parameters such as population size, maximum generation, crossover probability, and mutation probability, because they strongly affect a balance between exploitation and exploration in the search space.

Metaheuristic methods, especially GAs, have been widely and successfully applied in optimal system design because of their robustness and efficiency, even though they are time consuming, especially for large problems. To improve computation efficiency, hybrid optimization algorithms have been increasingly used to achieve an effective combination of GAs with heuristic algorithms, simulation annealing methods, NN techniques, and other local search methods.

In this context, Coit and Smith [27] presented a new formulation and solution method for the redundancy allocation problem with mixing components redundant in subsystems. They used a GA to solve this problem in k -out-of- n :G system. Chen and You [28] also presented the use of an immune

algorithm to solve such a problem in a series system. In general, this reliability design problem has been formulated by considering active redundancy. However, the choice of redundancy strategies for each subsystem is much more realistic and it provides a better tool for the designers. This becomes an additional decision variable in a redundancy allocation problem with mixing components redundant in subsystems.

Coit and Liu [29] presented a new formulation and solution method for the redundancy allocation problem when a system design includes multiple subsystems designed with either active or cold standby redundancy. This solution method assumes that the redundancy strategy (i.e., active or cold standby) for each subsystem is predetermined. Coit [30] presented a new formulation and solution method to the RAP when there are some subsystems using active redundancy and cold standby redundancy, or selecting the best redundancy strategy. Furthermore, Tavakkoli-Moghaddam *et al.* [31] used the genetic algorithm to solve the abovementioned problem. Unfortunately, the redundancy allocation problem with mixing components in subsystems is not considered when the redundancy strategy can be chosen for individual subsystems. Thus, the problem is to select the best redundancy strategy, combination of components, and redundancy level for each subsystem in order to maximize the system reliability under system-level constraints.

Most of the above cited works are based on the assumption that the element lifetimes are random variables. Although this assumption has been adopted and accorded with the facts in widespread cases, it is not reasonable in a vast range of situations such as a space shuttle system, where the estimations of probability distribution functions or density functions of element lifetimes are impossible due to the imprecise nature of the data. In this case, fuzzy theory [32–34], which is proposed as a branch of nonadditive measure theory for dealing with subjective uncertain phenomena, is widely employed in term of characterizing the element lifetime as fuzzy variables, and fuzzy redundancy optimization models are then widely studied [35–37].

Beyond fuzzy theory, uncertainty theory initialized by Liu [33] is another branch of nonadditive measure theory for describing subjective uncertain phenomena; this has also been well developed [38,39]. An in-depth investigation of the standby redundancy optimization problem within the framework of uncertainty theory can be found in Ref. 40.

REFERENCES

1. International Standard Organization(ISO)/ International Electrotechnical Commission (IEC). ISO/IEC 2382-14:1997 Information technology - Vocabulary - Part 14: Reliability, maintainability and availability. 1997.
2. Institute of Electrical and Electronics Engineers (IEEE). IEEE 610-1991 - IEEE Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries. 1991. ISBN:1559370793.
3. Birolini A. Reliability engineering: theory and practice. 5th ed. Berlin: Springer; 2007. p. 593. ISBN 3-540-40287-X.
4. U.S. Department of Defense. MIL-HDBK-338B Military Handbook Electronic Reliability Design Handbook. 1998.
5. Resnick RI. A modern taxonomy of high availability. 1996. Available at http://www.verber.com/mark/cs/systems/A_Modern_Taxonomy_of_High_Availability.htm
6. Birman K, Joseph T. Exploiting virtual synchrony in distributed systems. SIGOPS Oper Syst Rev 1987;21(5):123–138. DOI: <http://doi.acm.org/10.1145/37499.37515>.
7. Dimitri Kececioglu. Volumes 1&2, Reliability engineering handbook. Lancaster (PA): DEStech Publications; 1991. ISBN Volume 1: 1932078002, ISBN Volume 2: 1932078010.
8. Barlow RE, Proschan F. Mathematical theory of reliability. New York: John Wiley & Sons, Inc.; 1965.
9. Bazovsky I. Reliability theory and practice. Englewood Cliffs (NJ): Prentice-Hall; 1961–2004.
10. Shooman ML. Reliability of computer systems and networks: fault tolerance, analysis and design. New York: John Wiley & Sons, Inc.; 2002. p. 528. Hardcover. plus XXII. ISBN: 978-0-471-29342-2.
11. Cox DR. Renewal theory. London: Methuen; 1962. p. 142. plus IX.
12. Yearout RD, Reddy P, Grosh DL. Standby redundancy in reliability - a review. IEEE Trans Reliab 1986;35(3):285–292. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4335434&isnumber=4335413>
13. Osaki S. Reliability analysis of a two-unit standby redundant system with priority. Can J Oper Res 1970;8(1):60–62.
14. Gnedenko BV. Mathematical theory of reliability. New York: Academic Press; 1969.
15. Srinivasan VS. The effect of standby redundancy in system's failure with repair maintenance. Oper Res 1966;14:1024–1036.
16. Goel LR, Tyagi VK. A two unit priority redundant system with three modes and correlated failures and repairs. Microelectron Reliab 1992;32(10):1367–1372. Elsevier.
17. Vanderperre EJ, Makhnov SS. Long-run availability of a priority system: a numerical approach. Math Probl Eng 2005;1:75–85.
18. Coit DW, Smith A. Optimization approaches to the redundancy allocation to the redundancy allocation problem for series-parallel systems. In: Proceedings of the Fourth Industrial Engineering Research Conference; 1995 May; Nashville (TN). pp. 342–349.
19. Kuo W, Prasad V. An annotated overview of system-reliability optimization. IEEE Trans Rel 2000;49(2):176–187.
20. Kuo W, Prasad VR, Tillman FA, *et al.* Optimal reliability design: fundamentals and applications. Cambridge: Cambridge University Press; 2001.
21. Chern MS. On the computational complexity of reliability redundancy allocation in a series system,. Oper Res Lett 1992;11(5):309–315.
22. Tillman F, Hwang C, Kuo W. Optimization of systems reliability. New York: Marcel Dekker; 1980.
23. Dhingra AK. Optimal apportionment of reliability and redundancy in series systems under multiple objectives. IEEE Trans Reliab 1992;41(4):576–582.
24. Hikita MY, Nakagawa Y, Nakashima K, *et al.* Reliability optimization of system by a surrogate constraints algorithm. IEEE Trans Reliab 1978;7(5):325–328.
25. Jacobson DW, Arora SR. Simultaneous allocation of reliability and redundancy using simplex search. In: Proceedings of Annual Reliability and Maintainability Symposium; Las Vegas (NV). 1996. pp. 243–250.
26. Shelokar PS, Jayaraman VK, Kulkarni BD. Ant algorithm for single and multi-objective reliability optimization problems. Qual Reliab Eng 2002;18:497–514.

27. Coit DW, Smith A. Reliability optimization of series-parallel systems using a genetic algorithm. *IEEE Trans Reliab* 1996;45:254–260.
28. Chen T, You P. Immune algorithms-based approach for redundant reliability problems with multiple component choices. *Comput Ind* 2005;56:195–205.
29. Coit DW, Liu J. System reliability optimization with k-out-of-n subsystems. *Int J Reliab Qual Saf Eng* 2000;7:129–143.
30. Coit DW. Maximization of system reliability with a choice of redundancy strategies. *IIE Trans* 2003;35:535–544.
31. Tavakkoli-Moghaddam R, Safari J, Sassani F. Reliability optimization of series-parallel systems with a choice of redundancy strategies using a genetic algorithm. *Reliab Eng Syst Saf* 2008;93(4):550–556. DOI: 10.1016/j.res.2007.02.009.
32. Liu B, Liu Y. Expected value of fuzzy variable and fuzzy expected value models. *IEEE Trans Fuzzy Syst* 2002;10(4):445–450.
33. Liu B. Uncertainty theory. 1st, 2nd, and 3rd ed. Berlin: Springer; 2004–2007–2009.
34. Zadeh LA. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1978;1:3–28.
35. Cai KY, Wen CY, Zhang ML. Fuzzy variables as a basis for a theory of fuzzy reliability in the possibility context. *Fuzzy Sets Syst* 1991;42:145–172.
36. Utkin LV. Optimization by fuzzy reliability and cost of system components. *Microelectron Reliab* 1994;34(1):53–59.
37. Zhao R, Liu B. Standby redundancy optimization problems with fuzzy lifetimes. *Comput Ind Eng* 2005;49:318–338.
38. Gao X. Some properties of continuous uncertain measure. *Int J Uncertain Fuzz Knowl Based Syst* 2009;17(3):419–426.
39. You CL. Some convergence theorems of uncertain sequences. *Math Comput Model* 2009;49(3–4):482–487.
40. Liu B. Theory and practice of uncertain programming. 2nd ed. Berlin: Springer; 2009.

STANDBY SYSTEMS

SALVATORE DISTEFANO
Department of Mathematics,
University of Messina,
Messina, Italy

WHAT IS STANDBY?

According to several dictionaries [1–3] the adjective standby mainly means “ready for emergency use,” or “kept readily available for use as substitute/spare in an emergency, shortage, or the like, of or pertaining to a waiting period.” The adjective derives from the verb stand by [3]: “1. To be ready or available to act. 2. To wait for something, such as a broadcast, to resume. 3. To remain uninvolved; refrain from acting: stood by and let him get away. 4. To remain loyal to; aid or support: stands by her friends. 5. To keep or maintain: stood by her decision.”

From an in-depth analysis of such definitions we can identify a least common multiple for standby, that is the concept of putting aside or shelving, waiting for the occurrence of a specific event. The concept of wait is therefore of primary importance in standby systems: a standby system waits, ready for use, till it will be solicited to act or operate.

Another concept that often, but not always, characterizes standby is the one of *substitute/spare*: in such cases the wait is related to the failure, shortage or outage of a primary system, and the standby system is ready to substitute this latter in its main functionalities, the so-called *standby redundancy*.

Why Standby Systems?

Standby systems are those systems characterized by a double mode of acting, working or operating: full and standby. In the full mode, the system provides its service and functionalities to the environment, while in

the standby mode it does not provide any service and/or functionality although it could do that, waiting for an external input, signal, call, alert that is able to switch the system from standby to full mode.

There are several motivations that justify the introduction and the use of standby systems:

- optimizing cost and resources management, by introducing power-saving policies;
- optimizing the reliability/availability of complex systems, applying standby redundancy policies to their components;
- reducing the environmental impact, as also imposed and specified by several official laws and standards [4–8];
- the necessity of managing a population, a group of systems larger than needed to cover the real demand, introducing specific group management policies in order to keep busy, working the systems' group or to have a large amount of standby resources available.

Examples of standby systems can be found in different contexts and fields of application, such as standby commitments, standby cash, standby letters of credit, standby works in economy, the military standby reserve and standby force, the travel standby list, the standby assistance, the standby juror system, standby actors, standby alarms, and so on. But it is in technological and engineering fields that standby systems find the most varied and valuable applications, stimulating systematic and in-depth studies on the topic.

Standby Systems in Technological Contexts

In technological contexts, the concept of standby is mainly associated with and refers to the energy and/or to the power consumption of the system. Standby power, standby power systems, standby batteries, standby generators, and standby fuel are becoming

popular words increasingly used in plans, projects, data sheets, diagrams, schemes, and technical notes. For this reason standby and related terms have been included in many technical glossaries.

A term of particular interest in technological context (borrowed from electronics) is *sleep mode* and its synonymous *standby mode* [9]: a mode in which electronic appliances are turned off but under power and ready to activate on command. The *standby power* is instead the power used while an electrical device is in its lowest power mode (sometimes called the *off-mode* consumption or the *no-load* power use) [6]. This typically occurs when the product is switched off or not performing its primary purpose [7].

The introduction and the specification of the sleep/standby mode has given rise to a wide spread of standby systems, and consequently has driven the systematic and in-depth study of the topic. This trend has consolidated in the last decade due to the increasing sensitivity toward environment, pollution, and energy-related problems.

From this pulse, many government [6–8,10] and nongovernment [4,5,11] initiatives and projects were born. This has also revolutionized the approach in designing energy-efficient devices and systems, moving toward low-power electronics and focusing on standby management policies in order to optimize the power consumption and/or the reliability/availability of the overall system. For example, in computing contexts, such a trend has led to green computing, that is *the study and practice of designing, manufacturing, using, and disposing of computers, servers, and associated subsystems—such as monitors, printers, storage devices, and networking and communications systems—efficiently and effectively with minimal or no impact on the environment* [12]. Standby systems play a strategic role in green computing.

Remaining in the ICT context, a definition of standby worthy of mention has been provided by the ATIS Telecom glossary 2007 [10] (incorporating and superseding the ANSI T1.523-2001—Telecom Glossary 2000 standard—formerly known as Federal Standard 1037C); it characterizes three cases:

1. *In computer and communications systems operations, pertaining to a power-saving condition or status of operation of equipment that is ready for use but not in use.* (An example of a standby condition is a radio station operating condition in which the operator can receive but is not transmitting.)
2. *Pertaining to a dormant operating condition or state of a system or equipment that permits complete resumption of operation in a stable state within a short time.*
3. *Pertaining to spare equipment that is placed in operation only when other, in-use equipment becomes inoperative.* (Standby equipment is usually classified as (i) *hot standby equipment*, which is warmed up, i.e., powered and ready for immediate service, and which may be switched into service automatically upon detection of a failure in the regular equipment, or (ii) *cold standby equipment*, which is turned off or not connected to a primary power source, and which must be placed into service manually.)

Even though this latter definition is specific for computer and communications systems, it can be easily extended to electronic devices and appliances and, more in general, it can also include energized devices/systems with sleep/standby modes and corresponding power management facilities. In this way, a class of *standby powered systems* can be identified, characterized by a *phantom-load* sleep/standby mode as an alternative to the normal full-load functioning. In the sleep mode, a standby powered system can however be energized although not operating on a load. Such a busy powered wait characterizes the standby state, as also specified in case 3 of the above definition. According to that, the sleep-state condition is associated with both, the power supplied to the system while in standby and the time to resume the system from standby. The lower is the standby power (*hot/warm standby*) the longer is the time to resume the system from standby, since a greater number of its functions/components must be restarted/reinitialized. In case no

power is supplied in standby (*cold*), the system is switched off and therefore its restart can be very long, since the system must restart from scratch. Moreover, sometimes such operations cannot be performed automatically as discussed in the definition.

Such characterization can be extended to the sleep mode of standby powered systems without relating it to the restart/resuming feature, specifying the hot/warm standby in case the system standby mode is fully or partially powered, respectively, and cold standby if the system is not powered at all when in sleep mode.

It is important to note that the term *standby powered system* here introduced should not be confused with standby power system that is *an independent reserve source of electric energy that, upon failure or outage of the normal source, provides electric power of acceptable quality so that the user's facilities may continue in satisfactory operation* [11].

STANDBY SYSTEMS IN RELIABILITY

The standby is a hot topic in reliability, availability, and dependability contexts. It is usually associated with the concept of *redundancy*, implementing techniques and policies for optimizing the system reliability/availability through an adequate management of the redundant components' sleep/standby modes, the so-called standby redundancy; that is, *the use of redundant elements that are left inoperative until a failure occurs in a primary element* [13].

But the notion of standby in reliability/availability is a more general and complex concept that needs to be adequately investigated from a general and abstract point of view, without restricting the discussion to redundancy. In order to provide such a general abstract overview of standby systems in reliability, in the following sections the standby mode will be investigated from both internal/isolated and external/system reliability perspectives; that is, the behavior of a standby system will first be evaluated from an intrinsic point of view, without considering the interactions of

the standby system with the external environment, and then by taking into account such interactions into a more general system reliability viewpoint in which the standby system is included as a component.

From the Inside

In order to evaluate the effect of standby modes into standby systems, it is necessary to characterize the standby state in terms of reliability/availability, specifying the operating condition of a standby system and all the possible states it can assume. Once the standby systems' states are identified, a finer classification of standby can be applied, resulting in three classes.

Standby State. First of all, it is necessary to clearly specify what the standby/sleep mode or state is and means from the reliability/availability point of view. In other words, it is necessary to identify the state a standby system can assume. In system reliability/availability, a system, subsystem, or component can assume two states: up if reliable/available and down if unreliable/unavailable. Such kind of systems, subsystems, and/or components are thus classified as *Boolean*. Standby systems cannot be easily included in the class of Boolean systems, since they are a "borderline" state; it is not easy to classify a standby mode as up or down state. A more specific classification is required.

In Ref. 13, the two main states a system can assume are thus specified:

- *Up*. this pertains to a system or component that is operable and in service. Such a system is either operating or idle.
 - *Operating*. This pertains to a system or component that is operable, in service, and in use.
 - *Idle*. This pertains to a system or component that is operable and in service, but not in use.
- *Down*. This pertains to a system or component that is not operable or has been taken out of service.

Table 1. Standby System State Characterization

State	Feature					
	Operable		Service		Use	
	Yes	Not	In	Not In	In	Not In
Up						
Operating	✓		✓		✓	
Idle	✓		✓		✓	
Dormant	✓			✓		✓
Down						
Failed		✓		✓		✓

In this way, the *up time* is the period of time during which a system or component is operable and in service; that is, the sum of idle time and operating time. The idle time is the period of time during which a system or component is operable and in service, but not in use. The operating time is the period of time during which a system or component is operable, in service, and in use. Finally, the *down time* is the period of time during which a system or component is not operable or has been taken out of service.

Starting from such specification [13], three main features by which the states a system can assume are identified are as follows:

- *operable*—if the system is ready for use in its intended environment;
- *in service*—if the system is ready to provide its functions to the environment;
- *in use*—if the system is performing its intended functions.

While operable and in-use features mainly refer to the standby system from its internal point of view, the in-service feature is instead more specific to the external point of view, and thus it will be discussed in greater detail in the next subsection.

In this way, by applying the above characterization and making all the possible combinations among such features, four states are obtained for a standby system: the three states already identified in Ref. 13 plus a new state, summarized in Table 1: *operating*, *idle*, *dormant*, and *failed*. The operating, idle, and failed (down) states thus correspond

to the one defined in Ref. 13. The new dormant state represents a condition in which, the operable standby system is in its sleep mode waiting for an external trigger event occurrence. Such a new state, from the internal, isolated point of view has to be classified as an up state since the standby system is operable, as also in operating and idle states. The classification can change by considering a different, external, point of view, taking into consideration whether the standby system is in service or not, as done in the next subsection.

In this way, the idle and the dormant states can also be interpreted as two cases of a unique statement, and consequently they can be merged into a unique standby state. The merging of idle and dormant states into standby has a physical interpretation: the former condition (idle) represents the case in which the standby system automatically/self jumps from the operating state into the standby one (and vice versa) if no or reduced load is applied. The state change in such cases exclusively depends on the internal status and on the evolution of the standby system and does not affect the external environment. The latter condition (dormant) describes the case in which the standby switching is driven from external events. In this way, the standby system depends on the external environment, that usually is another component, subsystem or system.

Therefore, since often it is not possible to split the standby behaviors of a system into idle and dormant states due to the nature of the underlying physics phenomena, only a unique standby state is usually specified by merging such states, as already introduced and also confirmed by the specific literature [14–16]. Figure 1 pictorially depicts this fact by specifying the standby system state machine with all the possible states above characterized (rounded rectangles) and the corresponding state changes or events (arrows) a standby system can assume or perform. Four events have been identified for a standby system in the state machine of Fig. 1: *failure*, *resume*, *sleep*, and *repair*. The *failure* event represents transits driving to the failed state: generally systems in both operating and standby states can fail so failure

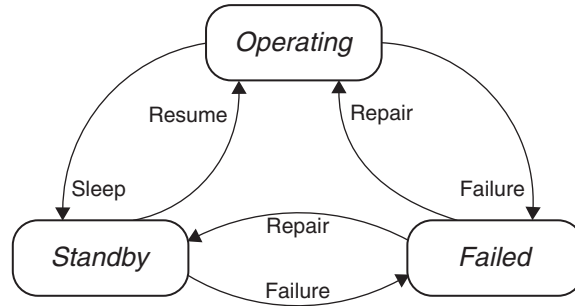


Figure 1. State machine of a standby system.

events from both these states are defined. The *resume* event switches from standby to operating states, representing the transition from standby mode to full-loaded operating conditions. The *sleep* event switches from operating to standby states, performing the reverse transition of resume events. Finally, the *repair* event implements the reverse path of failure events, representing the repair of the standby system that can be performed from the failed state to both standby or operating states, according to the condition applied to the system after the repair.

Types of Standby. In the section titled “Standby Systems in Reliability,” and more specifically in case 3, we refer to a characterization of standby mode made in Ref. 10 distinguishing between hot and cold standbys, there applied to a more general class of standby powered systems. Other interesting definitions are provided in Ref. 17:

- *Cold Standby.* This is a configuration in which a redundant functional unit can be brought into service with some delay should the primary functional unit fail.
- *Hot Standby.* This is a configuration in which a redundant functional unit can be immediately brought into service should the primary functional unit fail.

The above definitions characterize the standby mode/state by considering the power supplied to a standby system and/or the delay in bringing this latter into service: cold if no power is supplied to the system, introducing a delay for restarting in service, and hot if the system is powered and no delay elapses

for coming into service. In order to translate such classification in reliability terms, it can be observed that usually the power supplied to a system is in direct relationship with the unreliability of the system, that is the greater the energy supplied to the system the higher is its unreliability since the system created more work. This fact is confirmed in particular by the cold standby, since a system that is not powered cannot fail on its own; it can fail only in case of external causes (not considered in this subsection).

Thus, starting from this reasoning, a classification of the standby state in terms of reliability is possible by exploiting such relationships between reliability and power. In this way, three types of standby can be characterized [15,16,18,19]:

- *Cold Standby.* This is a condition/state in which the standby system, subsystem or component cannot fail autonomously.
- *Warm Standby.* This is a condition/state in which the standby system, subsystem or component can fail autonomously, but with reduced failure rate and/or delayed reliability distribution referred to the operating state’s ones.
- *Hot Standby.* This is a condition/state in which the standby system, subsystem or component can fail autonomously and with the same failure rate and/or reliability distribution of the operating state’s ones.

In case of cold standby, the system remains (intrinsically) reliable while in standby. In the other cases (warm and hot), the system can also fail from its standby

state. The warm standby is a case not included in the previous definitions, but it is of particular interest in reliability, and, thus, it has been dealt with here. Warm standby implies that the standby system changes the law or the distribution governing its reliability when the sleep mode is triggered, an effect that has to be adequately represented in terms of reliability. Such behavior is sometimes identified in the specific literature as *delayed aging process* [18].

The hot standby can be considered as a special case of warm standby, in which the standby system is energized as the system in operating condition, and therefore in terms of (intrinsic) reliability, there is no difference among the two conditions. However, it is important to remark that the above characterization of standby is made from an internal perspective, by considering the standby system in an isolated context. Dependent, on demand, cascading, common cause failures, and other similar external events can trigger the failure of the standby system, also in case of cold standby. By changing the perspective, the different concepts expressed here, especially the hot standby one, assume different meanings, as investigated in the following section.

From the Outside

By changing the point of view, the meaning of concepts and their interpretations can change; as a consequence it is necessary to review what is specified in the previous section by observing the standby system from the outside. Till now, we have dealt with standby systems in an isolated context, only discussing about a generic relationship with the external environment. Such a relationship can be considered as a dependent-dynamic interaction between the standby system and the external environment. However, in the system reliability/availability perspective, a standby system is considered as a part or component of a more complex system. More specifically, the main goal of system reliability/availability is the construction of a model (life distribution) that represents the time-to-failure of the entire system, based on life distributions and maintenance policies of

the components, subassemblies, and/or assemblies by which the system is composed.

In our specific case, it is necessary to adequately identify and evaluate the interactions among standby systems, subsystems or components (components hereafter) when they are included or compose more complex systems. Thus, the standby is here represented as a dependent interaction among components, subsystems, or systems. Such interaction involves two sides: the drive component triggering the dependency and the standby component that suffers this.

In order to adequately model and evaluate the standby, specific information regarding and quantifying dependent and dynamic relationships among the components are needed. The reliability/availability distributions in such case become bivariate or multivariate and specific models, such as proportional hazard models [20], frailty models [21], copula-based models, and so on, can be used for expressing the correlation among the components involved in dependent behaviors.

According to such considerations, from the system reliability point of view, the characterization made in Table 1 has to be revised and reinterpreted. Two standby states have been identified there, the idle and the dormant states; then, they have been merged into a unique standby state in the state machine of Fig. 1 and evaluated from an intrinsic point of view, as up states.

The dormant state represents the case in which the standby component depends on other components of the systems driving the standby. In such cases, since the standby component in the dormant state is operable but not in service till the trigger event occurs, from an external perspective the standby component has to be considered as down and out of service. In other words, the down state is characterized by both the failed state and the dormant one. But, while in the former case (failed) an external time-consuming action (repair, replacement, etc.) is needed to restore the operational conditions of the component, in the other case (dormant) the standby component is not physically failed: it is in a dormant state, waiting for an input from its driver component switching the standby component in service.

With regard to the characterization of Fig. 1, where the standby state groups idle and dormant states, the classification of dormant state as down state conflicts with the one identifying idle state as up state. In such cases, in order to classify the overall standby state as up or down in terms of reliability/availability, it is necessary to study in depth, case by case, the underlying physical phenomena and evaluate whether the standby is driven by another component of the system (dormant-down) or not (idle-up).

This can be considered as a drawback of the one-standby-state characterization implemented, but usually it is the only way to represent standby components, since they are characterized by reliability/availability distributions that can change from operating to standby conditions, as discussed in the next section. This fact becomes evident with the specification of three types of standby state (hot, warm, and cold) made in the section titled “Summary.” In particular, the warm and cold standbys are characterized by their own reliability/availability distributions that differ from the operating ones. Therefore, operating and standby states cannot be merged and have to be modeled into separate states, regardless of whether the standby state is classified among up or down states since the reliability/availability distributions of operating and standby states are different.

From such a perspective, the hot standby state that has the same distribution as the operating state, is meaningfully represented only if it is associated with down states since it is clearly distinguished from the operating state that is an up state. In case such a distinction cannot be established, that is the hot standby is identified as an up state, the standby characterization becomes meaningless. Moreover, a component in hot or warm standby state can obviously fail in case the standby is characterized both as an up state and as a down state. Instead, no failures are possible from cold standby states in any case.

Assessing Reliability of Standby Systems

A standby system switches from sleep to operating modes and vice versa if a specific

external event occurs. In terms of probability, this means that the unreliability of the system can be represented by a bivariate CDF $F_{X,Y}(x,y)$ depending on two events: the system time-to-failure X and the conditioning event Y . The problem here is how to obtain the conjunct $F_{X,Y}(x,y)$ CDF by knowing the marginal CDFs $F_X(x)$ and $F_Y(y)$. But, in this way, the problem is ill posed because in general it is not possible to obtain the conjunct CDF from the marginal CDFs; it is necessary to deeply know the dependency between X and Y . Thus since a solution for the generic problem does not exist, some specific cases, fixing precise and circumscribed assumptions on the dependency, have been analyzed.

A representation widely spread in such a context is the *proportional hazard models* one [20], also known as *Cox models*, and in common, cause failure context as *β -factor models* and characterization of the more general *frailty models* [21]. Proportional hazard models express the covariance of the system time-to-failure random variable before and after the occurrence of a specific event influencing the system (the standby trigger) in terms of hazard-failure rate. Discursively, in proportional hazard models, the treatment under study has a multiplicative effect on the subject's hazard rate, that is, in the simplest case, the two hazard/failure rates are proportional.

Thus, the proportional hazard model assumes changing a stress variable (or explanatory variable) has the effect of multiplying the hazard rate by a constant. In more formal terms, a proportional hazard model can be expressed by the following equation [20,22]:

$$h_z(t) = g(z) h_0(t), \quad (1)$$

where $z = \{x, y, \dots\}$ is a vector of one or more explanatory variables believed to affect lifetime, $h_0(t)$ is the hazard rate for a nominal (or baseline) set $z_0 = (x_0, y_0, \dots)$ of these variables, and $h_z(t)$ is the function describing the hazard rate for a new value of z .

In other words, changing z , the explanatory variable vector, results in a new hazard function that is proportional to the nominal

hazard function and the proportionality constant is a function of z , $g(z)$, independent of the time variable t . A common and useful form for $f(z)$ is the log linear model which has the equation $g(x) = e^{ax}$ for one variable, $g(x, y) = e^{ax+by}$ for two variables, and so on. For a more detailed specification of proportional hazard models and their applications to reliability and availability of standby systems, see Refs 20 and 22.

Proportional hazard models are the simplest models for representing dependent-correlated events such as the standby. Anyway, other different and more complex forms of correlation can be used for representing dependent events (copula-based models, Marshall–Olkin models, and variants). The reader is directed to Chapter 8 of Ref. 23 for an overview of such models.

Coming back, suppose we apply the proportional hazard model of Equation (1) to a standby system initially in fully operating mode, the system hazard rate of the first change point (operating sleep mode switch at time y) is

$$h_y(t) = \begin{cases} h_X(t), & t < y, \\ \beta h_X(t), & t \geq y, \end{cases} \quad (2)$$

where $h_X(t)$ and $\beta h_X(t)$ are the hazard rate of the system in fully operating and sleep modes, respectively. β is a constant weighting the impact of standby in terms of reliability and is known as *dormancy factor* [15] or *dependency rate* [16]. In standby systems, the values of β range from 0 to 1: $\beta = 0$ identifies cold standby and $0 < \beta < 1$ characterizes warm standby while, in case of hot standby, $\beta = 1$.

Given the CDF of the trigger event $YF_Y(y)$, $h_X(t)$, and β , the goal of this section is to provide an equation describing the unreliability CDF of the standby system $F_{X,Y}(t)$. Since given the hazard rate $h(t)$, the CDF $F(t)$ can be expressed as $F(t) = 1 - e^{-\int_0^t h(x) dx}$, and the PDF $f(t)$ is $f(t) = h(t)e^{-\int_0^t h(x) dx}$, by applying the law of total probability it is possible to obtain the following formula for

$F_{X,Y}(t)$:

$$\begin{aligned} F_{X,Y}(t) &= \int_{-\infty}^{+\infty} \Pr(X \leq t | Y = y) f_Y(y) dy \\ &= \int_0^t \Pr(X \leq t | Y = y) f_Y(y) dy \\ &\quad + \int_t^{+\infty} \Pr(X \leq t | Y = y) f_Y(y) dy \\ &= \int_0^t (1 - \Pr(X > t | Y = y)) f_Y(y) dy \\ &\quad + F_X(t) F_Y(y) \Big|_t^{+\infty}. \end{aligned} \quad (3)$$

In order to evaluate the term $\Pr(X > t | Y = y)$, it is necessary to observe that the standby system is firstly, till y , in the sleep mode, and then, at y , switches in the fully operating state. In terms of probability this represents a change point, in which the system lifetime changes its distribution from the standby ($F_X^S(t) = 1 - e^{-\beta \int_0^t h_X(x) dx}$) to the fully operating ($F_X^O(t) = 1 - e^{-\int_0^t h_X(x) dx}$) mode. Applying the *conservation of reliability principle* [24], it is possible to quantify such effect as time, into the *delay* τ , by obtaining the *time equivalent*, so that at change point y it has

$$R_X^S(y) = R_X^O(y + \tau) \longrightarrow \tau = R_X^{O-1}(R_X^S(y)) - y, \quad (4)$$

where $R(t) = \Pr(T > t) = 1 - F(t)$, assuming that the CDF $R_X^O(\cdot)$ is *strictly decreasing* and therefore *invertible*. It is important to remark that the time equivalent delay τ could also be negative, that means that the system at change point increases its hazard rate.

In this way, the probability $\Pr(X > t | Y = y)$ becomes

$$\begin{aligned} \Pr(X > t | Y = y) &= R_X^O(t + \tau) \\ &= R_X^O\left(t + R_X^{O-1}(R_X^S(y)) - y\right). \end{aligned} \quad (5)$$

By substituting this term in Equation (3), we have

$$\begin{aligned} F_{X,Y}(t) &= F_X^S(t)(1 - F_Y(t)) \\ &\quad + \int_0^t (1 - R_X^O) \left(t + R_X^{O-1} \left(R_X^S(y) \right) - y \right) \\ &\quad \times f_Y(y) dy. \end{aligned} \quad (6)$$

The one reported in Equation (6) is a general formula valid for any type of standby. It could be of interest to characterize it for hot and cold standby. In the hot standby case, the system never changes its reliability, that is $R_X^O(t) = R_X^S(t) = R_X(t)$ (and obviously $F_X^O(t) = F_X^S(t) = F_X(t)$), thus it is expected that $F_{X,Y}^{HS}(t) = F_X(t)$. In order to verify Equation (6) it is necessary to apply such a condition to this latter thus obtaining

$$\begin{aligned} F_{X,Y}^{HS}(t) &= F_X(t)(1 - F_Y(t)) \\ &\quad + \int_0^t (1 - R_X(t)) f_Y(y) dy \\ &= F_X(t)(1 - F_Y(t)) + F_X(t)F_Y(t) \\ &= F_X(t). \end{aligned} \quad (7)$$

A system in cold standby is initially inactive, thus its reliability is $R_X^S(t) = 1$ or $F_X^S(t) = 0$, and therefore

$$\begin{aligned} F_{X,Y}^{CS}(t) &= \int_0^t (1 - R_X^O(t - y)) f_Y(y) dy \\ &= \int_0^t F_X^O(t - y) f_Y(y) dy, \end{aligned} \quad (8)$$

that is the convolution among events X and Y as expected.

The formulas reported in Equations 6–8 are valid for any generally distributed CDF. In the particular case that the standby system lifetime is exponentially distributed, due

to the age/memory-less property of exponential CDF, Equation (6) becomes an *exponential polynomial*:

$$\begin{aligned} F_{X,Y}(t) &= (1 - e^{-\lambda_X^S t}) e^{-\lambda_Y t} + (1 - e^{-\lambda_Y t}) \\ &\quad - \int_0^t e^{-\lambda_X^O \left(t + (\lambda_X^S - \lambda_X^O) / \lambda_X^O y \right)} \lambda_Y e^{-\lambda_Y y} dy \\ &= (1 - e^{-\lambda_X^S t}) e^{-\lambda_Y t} + (1 - e^{-\lambda_Y t}) \\ &\quad - \lambda_Y \int_0^t e^{-\lambda_X^O t - (\lambda_X^S - \lambda_X^O + \lambda_Y) y} dy \\ &= (1 - e^{-\lambda_X^S t}) e^{-\lambda_Y t} + (1 - e^{-\lambda_Y t}) \\ &\quad - \frac{\lambda_Y}{\lambda_X^S - \lambda_X^O + \lambda_Y} e^{-\lambda_X^O t} \\ &\quad \times \left(1 - e^{-(\lambda_X^S - \lambda_X^O + \lambda_Y) t} \right). \end{aligned} \quad (9)$$

SUMMARY

Standby systems are of fundamental importance in actual technologies since they are a way of reducing the environmental impact, extending the systems' lifetime, and especially of reducing and optimizing costs. It is important to fix the concept of standby such that it identifies a specific condition that differs from the operating one. Here, such distinction is mainly investigated from the reliability perspective providing both the intrinsic, isolated viewpoint and the external, system reliability viewpoint. In this way, a standby system can be characterized as an up or down state according to the underlying physical process that therefore must be adequately evaluated before classifying the standby.

In technological fields, the concept of standby is strictly related to the power. According to the power consumption of a standby system, the standby state can be further characterized. Following the relationship between reliability and power, three types of standby can be characterized as hot, if the system in standby has the same power consumption and reliability distribution of the fully operating condition; warm, if the system in standby is partially powered and

consequently has higher reliability than in the operating state; and cold, if the system is not powered and therefore cannot fail from standby.

REFERENCES

1. Merriam-Webster Online Dictionary. Available at <http://www.merriam-webster.com/>. Accessed 2009 Sept.
2. Dictionary.com Online Dictionary. Available at <http://dictionary.reference.com/>. Accessed 2009 Sept.
3. YourDictionary.com Online Dictionary. Available at <http://www.yourdictionary.com/>. Accessed 2009 Sept.
4. International Energy Agency (IEA). IEA Standby power initiative. Task Force 1: Definitions and Terminology of Standby Power; 17–18 Nov 1999; Washington (DC). 1999.
5. International Electrotechnical Commission (IEC). IEC 62301 standard: household electrical appliances - Measurement of standby power. Edition 2.0. Available at <http://www.iec.ch/cgi-bin/procgi.pl/www/iecwww.p?wwwlang=E&wwwprog=pro-det.p&He=IEC&Pu=62301&Pa=&Se=&Am=&Fr=&TR=&Ed=2>. Accessed 2009 Sept.
6. Australian Government, Department Of Environment, Water, Heritage and the Arts. Australian standby power program. Available at <http://www.energyrating.gov.au/standby.html>. Accessed 2009 Sept.
7. U.S. Environmental Protection Agency and U.S. Department of Energy. ENERGY STAR program. Available at <http://www.energystar.gov/>. Accessed 2009 Sept.
8. The European Commission. The directive 2005/32/EC on the eco-design of energy-using products (EuP). Available at http://ec.europa.eu/enterprise/eco_design/index_en.htm. Accessed 2009 Sept.
9. Meier A. Standby power use - definitions and terminology. First Workshop on Reducing Standby Losses; 1999 Jan; Paris.
10. Alliance for Telecommunications Industry Solutions (ATIS). American national standard ATIS telecom glossary 2007. Available at <http://www.atis.org/glossary/>. Accessed 2009 Sept.
11. Institute of Electrical and Electronics Engineers (IEEE). IEEE Std 446-1995 - IEEE recommended practice for emergency and standby power systems for industrial and commercial applications. 1995.
12. Murugesan S. Harnessing green IT: principles and practices. *IT Prof* 2008;10(1):24–33. DOI: 10.1109/MITP.2008.10.
13. Institute of Electrical and Electronics Engineers (IEEE). IEEE 610-1991 - IEEE standard computer dictionary. A compilation of IEEE standard computer glossaries. 1991. ISBN: 1559370793.
14. Lešanovský A. Analysis of a two-unit standby redundant system with three states of units. *Appl Math* 1982;27(3):192–208.
15. Dugan JB, Bavuso S, Boyd M. Dynamic fault tree models for fault-tolerant computer systems. *IEEE Trans Reliab* 1992;41(3):363–377.
16. Distefano S, Puliafito A. Dependability evaluation with dynamic reliability block diagrams and dynamic fault trees. *IEEE Trans Dependable Secure Comput* 2009;6(1):4–17.
17. International Standard Organization (ISO)/International Electrotechnical Commission (IEC). ISO/IEC 2382-14:1997 information technology - Vocabulary - Part 14: Reliability, maintainability and availability. 1997.
18. Cha JH, Mi J, Yun WY. Modelling a general standby system and evaluation of its performance. *Appl Stoch Models Bus Ind* 2008;24(2):159–169.
19. Vanderperre EJ. Long-run availability of a warm standby system. *Math Notes* 2008;85(5):623–630. in Russian: *Matematicheskie zametki* 2008;84(5):667–675).
20. Cox DR. Regression models and life tables. *J R Stat Soc [Ser B]* 1972;34(2):187–220. JSTOR: 2985181.
21. Aalena OO, Hjorth NL. Frailty models that yield proportional hazards. *Stat Probab Lett* 2002;58(4):335–342.
22. Li X, Yan R, Zuo MJ. Evaluating a warm standby system with components having proportional hazard rates. *Oper Res Lett* 2009;37(1):56–60.
23. Bedford T, Cooke RM. Probabilistic risk analysis: foundations and methods. Cambridge: Cambridge University Press; 2001. pp. xx+393. ISBN: 0-521-77320-2.
24. Kececioglu D. Volumes 1 & 2, Reliability engineering handbook. Lancaster (PA): DEStech Publications; 1991. ISBN: 1932078002 (Volume 1), ISBN: 1932078010 (Volume 2).

STATISTICAL ANALYSIS OF CALL-CENTER OPERATIONAL DATA: FORECASTING CALL ARRIVALS, AND ANALYZING CUSTOMER PATIENCE AND AGENT SERVICE

HAIPENG SHEN
Department of Statistics and
Operations Research,
University of North Carolina
at Chapel Hill, Chapel Hill,
North Carolina

Call centers are modern service networks in which agents provide services to customers via telephones. They have become a primary contact point between service providers and their customers. *Inbound* call centers take calls that are initiated by customers. Examples include phone operations that support reservation and sales for hotels, airlines, and car rental companies, as well as service centers of retail banks and brokerage firms. *Outbound* call centers initiate calls to outside parties, for example, those in organizations such as telemarketers, bill collection agencies, and marketing research and polling companies.

The increasing importance and complexity of call-center operations have rendered call centers into a fertile ground for academic research, both empirical and theoretical. Two survey papers [1,2] and an on-line bibliography [3] have well documented the current state of relevant research on call-center operations and directions for future research. This article focuses on statistical research about inbound call centers, in the areas of call arrival forecasting and analysis of customer patience and agent service times. There has been empirical research on call centers that is done outside the operations community and covers behavioral and psychological aspects [2] and human resource issues [4].

Brown *et al.* [5] described a simplified path that each call follows through a typical inbound call center. A customer calls the telephone number associated with

the center. Once connected, the customer usually first enters an interactive voice response unit (IVR) and identifies himself or herself. In the IVR, the customer can perform some self-service transactions, and sometimes conclude the service there. In case the customer still wants to speak with an agent, if an agent is available and is capable of performing the desired service, then the customer is connected to the agent to receive service immediately; otherwise, the customer joins an invisible queue of waiting callers and starts to wait. It is often the case that informational or noninformational announcements such as music, commercials, or updates about any progress in the queue are provide to the customer while waiting. Eventually, the customer is either served by an agent or becomes impatient and hangs up (i.e., abandons) before an agent is available to serve him or her. Sometimes, the customer may call again after getting a busy signal, or abandoning the system, or even finishing a service.

The above simple description suggests that call centers can be modeled as queueing systems. Call-center managers then use queueing models to manage their center operations, seeking to appropriately balance agent utilization and service quality according to quality of service measures such as *average speed of answer* (ASA) or *fraction of abandoning customers*.

Existing queueing models usually make specific theoretical assumptions about various system primitives, such as the call arrival process, service durations, or customer abandonment behavior. The classical Erlang-C queueing model, for example, makes three assumptions: (i) the call arrival process is Poisson; (ii) service durations are exponentially distributed; and (iii) customers never abandon the queue before getting served. The third assumption turns out to be a serious shortcoming for modeling call centers, because of the prevalence of customer abandonment in call centers. The more recent so-called Erlang-A model is a more realistic

model that explicitly takes into account abandonments but assumes the *time to abandonment* distribution is exponential [6].

Gans *et al.* [1] assert that “the modeling and control of call centers must necessarily start with careful data analysis.” The queueing models all involve certain parameters for the system primitives, for example, the Erlang-A model needs one to specify call arrival rates, service rates, and customer abandonment rates. To apply these models in practice, the associated parameters need to be estimated or forecasted through statistical analysis of historical call-center operational data. Furthermore, Brown *et al.* [5] showed that statistical analysis could be used to validate and calibrate queueing theoretical models.

Despite the importance of statistics, however, the empirical research on call centers remained very scarce until recently. The recent emerging growth, especially in the area of call forecasting, is primarily driven by increased access to call-center data and by widened recognition of the importance of empirical research.

This article aims to provide an introductory review of some empirical research about inbound call centers. It starts with a brief description of call-center operational data and an analysis tool, continues in the sections titled “Call Forecasting” and “Customer Patience and Agent Service Times” with general overviews of statistical research in call forecasting, and analysis of customer patience and agent service times respectively, and ends with a discussion of several directions for future research in the section titled “Future Challenges.”

CALL-CENTER OPERATIONAL DATA

Call-center workforce management systems (WFM) routinely collect call-center operational data. The data record the physical process by which calls are handled. For example, among many other things, they consist of the time stamps of the various stages that a call goes through and the associated operational variables for each stage, which provide necessary starting points for most of the empirical research reported below.

A collaborative team of operations management researchers and statisticians performed the first thorough statistical analysis of a (then) unique call-by-call operational database collected at a call center of a small Israeli bank [5]. The analysis finds that several common queueing model assumptions are rejected empirically; in addition, some queueing-theoretic results are shown to be robust against such violations while a few others are not. The data and an earlier empirical analysis are publicly accessible from the Service Enterprise Engineering (SEE) Center at the Technion.

One key lesson the team learned is that, for successful and comprehensive statistical analysis, one needs call-by-call (or transaction-by-transaction) data that record the detailed event-history of the individual calls. For example, the team used the call-by-call data to develop statistical models for the distributions of call arrivals, agent service times and customer patience, which, in turn, allows one to make rigorous and valid statistical inference, or identify interesting empirical observations. Unfortunately, in practice call centers typically only use the call-by-call data to calculate simple summary statistics for the calls that arrive within fixed intervals of time, often 15 or 30 min, and then discard the transaction-level data. As demonstrated in Ref. 5, the aggregated summaries are too simple to provide enough details for meaningful statistical modeling, several examples of which are discussed later.

In addition, the WFMs are designed to collect the operational data in an “engineering-efficient” way; hence the collected data in most cases are not directly amenable for statistical analysis. The experience of analyzing the pilot study helped the team to define and develop a *DATA MOdel for Call-Center Analysis* (DATA-MOCCA) [7]. The mission of the DATA-MOCCA project is to collect, preprocess, organize, analyze, and distribute transactional data from call centers. The core of DATA-MOCCA is a database system that is useful for the analysis of the call-by-call data that are gleaned from call centers’ telephone switches and information systems. The latest distributable implementation includes

two large databases, one of a medium sized US bank and the other of an Israeli telecom company, along with a user-friendly query engine, SEESTAT, that facilitates empirical exploration and statistical analysis of the data. A public version of SEESTAT can be downloaded from the Technion SEE Center.

CALL FORECASTING

Seventy percent of call-center operating expenses are salary costs for human resource [1]. Effective management of a call center requires its manager to match up the center resource with future workload during any planning period. The workload or *offered load* is defined to be the product of the call arrival rate and the mean service time of the arrivals. One can interpret workload as the expected time units of work that arrives to the system per unit of time. For example, consider a stationary 30-min interval during which the arrival rate is 4 and the mean service time is 15 min, then the workload to the system would be $4 \times 15/30 = 2$. The workload serves as a lower bound for the staffing level for an Erlang-C system, which does not allow blocking or abandonment. Ideally, for staffing purposes, a manager wants to forecast workload accurately. However, most of the existing research has focused on forecasting future call arrival rates, with the exception of Refs 5 and 8.

In this section, we first describe several common characteristics of call arrivals in the section titled “Common Characteristics of Call Arrivals”: time inhomogeneity, arrival rate uncertainty, and interday and intraday dependence. These features are observed empirically and are useful for developing effective forecasting methods. We then review existing call-forecasting methods in the section titled “Call-Forecasting Methods.”

Common Characteristics of Call Arrivals

In most inbound call centers, both the arrival rate of calls and the mix of types of calls entering the system vary over time. Call arrivals have both *predictable* and *unpredictable* components: over very short periods

of time, seconds or minutes, the numbers of arriving calls are viewed as stochastic and unpredictable, typically as some form of a Poisson process; however, over longer periods of time—hours, days, weeks, or months—the numbers of arrivals are traditionally viewed as more predictable (see Fig. 5 of Ref. 1 for a hierarchical view of arrival rates).

Call forecasting starts with historical call arrival data, which record the time stamps at which calls arrive to the center and their corresponding service types, in cases where different types of calls are handled by different groups of agents. The common practice in the literature is as follows: first, the center operating period is divided into short time intervals, usually 15 or 30 min, during which the arrival rates are assumed to remain constant; then, the raw data of arrival times are aggregated into numbers of arrivals within the short intervals. The collection of such aggregated call volumes for a given day is referred as the *intraday call volume profile* of that day [9].

Common call-center models and practice then assume that the intraday call volume profile is a realization from a time-inhomogeneous Poisson process, where the underlying arrival rate function varies across different intervals. Brown *et al.* [5] constructed a statistical test, and empirically validated the time-inhomogeneous Poisson assumption using the Israeli call-center data. In addition, they performed a statistical test and claimed that the arrival rate function remains random (or uncertain) given the available covariates of time-of-day, day-of-week, and type of service, and hence is hard to predict. A Poisson process with uncertain arrival rate is referred as a *doubly stochastic Poisson process*, or a *Cox process*.

The phenomenon of arrival rate uncertainty can be manifested empirically through the *overdispersion* of the call volumes during the same time interval across multiple days. Here overdispersion refers to the fact that the variance of the volumes is much larger than the mean. Note that, if the volumes were independent and identically distributed according to a Poisson random variable, then the variance would be equal to the mean. This phenomenon has been recognized before

by various authors. Maman [10] recently proposed a theoretical model for overdispersion in call centers and hospital emergency departments, and suggested procedures to empirically estimate the model parameters.

After some initial processing, historical call arrival data, or historical intraday call volume profiles, can be stored as a two-way data matrix, where each row corresponds to the intraday call volume profile for a particular day, and each column records the call volumes during a particular time interval across the days. Alternatively, one could concatenate all the intraday profiles one after another to form a single long profile.

Viewing the historical arrival data as a matrix, there exists a two-way dependence structure that is useful for forecasting future arrival rates. First, there is *interday dependence* among the call volumes across days. For example, Brown *et al.* [5], Shen and Huang [9], and Weinberg *et al.* [11] all reported empirical evidence of dependence within the time series of daily total call volumes. Such dependence can depend on seasonal factors such as day-of-week, month-of-year, and is helpful for forecasting arrival rates several days, weeks, or months ahead.

In addition, there is *intraday dependence* in that call volumes during different time intervals within a given day are positively correlated. For example, Avramides *et al.* [12] developed doubly stochastic Poisson models that reproduced three empirically observed features of call arrival data, *overdispersion*, *time-varying rate*, and *intraday dependence*. Intraday dependence makes it possible to perform within-day dynamic updating, for example, to adjust the forecasts for the afternoon arrival rates on the basis of the call volumes in the morning of the same day. This updating is operationally both feasible and beneficial. As shown in Refs 9, 11, and 13, within-day updating can substantially reduce the forecast error for the rest of the day, and it can be done as early as a couple of hours into the working day to still have a significant effect. In turn, operational benefits would follow from the ability to change agent schedules according to the updated forecast. For example, to adjust for a revised forecast, call-center managers may send people home early

or have them take training sessions or work on alternative tasks; or they may arrange agents to work overtime or call in part-time or work-from-home agents. There is some ongoing work that formulates stochastic programming models with recourse actions to actually quantify the practical benefits of within-day updating for scheduling agents, using real call-center data.

Call-Forecasting Methods

Early work on forecasting of call arrivals usually ignores intraday dependence, and focuses on interday dependence among daily total call volumes. Part of the reason is due to lack of relevant data. Forecasting is usually performed through time series forecasting methods such as using autoregressive integrated moving average (ARIMA) models. Transfer functions and exogenous variables are included to incorporate special effects such as advertisement, holidays, and sales promotion. In addition, only point forecasts for future arrival rates are produced. Workforce management in call centers traditionally assumes that these point forecasts are certain. However, the arrival rate uncertainty reported earlier in the section titled “Common Characteristics of Call Arrivals” suggests that the forecasted rates are not realized with probability one.

Recently, with the availability of call-by-call data, several important developments in the call-forecasting literature have been made [5,8,9,11,13–15] that take into account both interday and intraday dependence. The forecasting approaches can be roughly grouped into two categories: model-driven and data-driven. Most of the recent developments consider both point forecasts and distributional forecasts, which capture forecasting uncertainty.

Model-Driven Approaches. The model-driven approaches include Refs 5, 8, 11, 13, and 14. They start with various stochastic models of call arrivals, often as some variations of Poisson processes.

Both Brown *et al.* [5] and Weinberg *et al.* [11] model the call arrivals using a time-inhomogeneous Poisson process with a random rate function. In both papers, a

square-root transformation is first applied to the original arrival counts. The motivation is twofold: first, the square-rooted counts roughly have a constant variance; second, they are approximately normally distributed. Brown *et al.* [5] then build a multiplicative two-way mixed-effects model on the transformed counts. The square-root of the arrival rate for each interval is modeled as the product of the daily total of the square-rooted rates and the proportion of arrivals during that time interval. They then assume that the arrival proportion for each interval remains the same for different days of the week, and the daily total square-rooted rate can be forecasted using the total of the square-rooted counts of the previous day.

Weinberg *et al.* [11] extended the work of Brown *et al.* [5] to develop forecasting models for both day-to-day forecasting and within-day updating. As an example of the various forecasting methods, the statistical model proposed by Weinberg *et al.* [11] is presented below. Suppose N_{ij} is the number of arrivals during the j th time interval of the i th day, which is modeled as a Poisson random variable with a random rate λ_{ij} that depends on both day-of-week and time-of-day. A two-way multiplicative Bayesian model is then proposed as follows:

$$\begin{cases} x_{ij} = \sqrt{N_{ij} + 1/4} = \sqrt{\lambda_{ij}} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2), \\ \sqrt{\lambda_{ij}} = y_i g_{d_i}(t_j), \\ y_i - \alpha_{d_i} = \beta(y_{i-1} - \alpha_{d_{i-1}}) + \eta_i, \eta_i \sim N(0, \phi^2), \\ \frac{d^2 g_{d_i}(t_j)}{dt_j^2} = \tau_{d_i} \frac{dw_{d_i}(t_j)}{dt_j}, \end{cases}$$

where t_j is the center of the j th time period, $g_{d_i}(\cdot)$ models the *smooth* intraday call arrival pattern that depends on day-of-week $d_i (= 1, \dots, 5)$, y_i is a random effect with an autoregressive structure adjusting for d_i , and $W_{d_i}(t)$ are independent Wiener processes with $W_{d_i}(0) = 0$, and $\text{var}\{W_{d_i}(t)\} = t$. The model essentially assumes that the time series of daily total square-rooted rates follows a first-order autoregressive model with day-of-week effect, and the arrival proportion changes smoothly across the time intervals and can depend on day-of-week. The authors then choose suitable prior distributions for

the model parameters, and develop a Markov chain Monte Carlo (MCMC) algorithm for parameter estimation and forecasting. They also propose a within-day learning algorithm to efficiently perform within-day updating of arrival rates.

Without using the square-root transformation, Shen and Huang [13] directly modeled the historical arrival count matrix as observations from a time series of inhomogeneous Poisson processes, where the time series dependence is across the sequence of days (or rows of the data matrix). The count matrix is then viewed as a multivariate Poisson time series with the dimension being the number of columns in the matrix, or the number of time periods within a day. The dimensionality of the time series is usually high, for example, a 17-h working day consists of 68 time intervals of 15 min each. Hence, the authors start with a Poisson factor analysis on the count matrix to achieve dimension reduction. The estimated factor score series are then modeled using univariate time series models. The forecasts of future factor scores are combined with the extracted factor loadings to yield forecasts of future arrival rate functions. Penalized Poisson regressions on the factor loadings are used to generate dynamic within-day updating of arrival rates.

Unlike in the previous papers, Aldor-Noiman *et al.* [8] considered a univariate response vector by concatenating the square-root-transformed daily volume profiles together. This “long” volume profile (instead of the “fat” count matrix analyzed in the earlier studies) is then modeled using a normal linear mixed-effects model that treats both day and period effects as random effects, both of which are assumed to have a first-order autoregressive correlation structure, and includes exogenous variables such as day-of-week and billing cycle information as fixed effects. A two-stage model estimation algorithm is proposed in order to avoid convergence issues. The authors consider the problem of forecasting both arrival rate as well as workload, under various forecast lead times. In addition, they study the effect of interval length on forecasting accuracy and

the effect of the model results on call-center operational performance.

Similar to Weinberg *et al.* [11], Soyer and Tarimcilar [14] also used a Bayesian approach to model call arrivals at an inbound sales call center. For the purpose of developing good marketing strategies, the authors are interested in advertisement-specific analysis of call arrivals to evaluate the efficiency of various advertisements and promotions. Different from the previous studies, the call arrival data are counts of incoming calls during periods of days over the life cycle of an advertisement. The counts are modeled using a modulated time-inhomogeneous Poisson process. A proportional rate model is used to model the underlying arrival rate function as the product of a baseline rate function and an advertisement effect consisting of exogenous variables such as the advertisement characteristics. Similar models have been considered before in the survival analysis literature. Random effects are used to incorporate potential heterogeneity among various advertisements.

Data-Driven Approaches. The data-driven forecasting approaches include Refs 9 and 15. The approaches have advantages in cases where no specific stochastic models for the call arrival processes are made. For example, there exists empirical evidence that calls entering the service queue after a congested IVR did not follow a time-inhomogeneous Poisson process.

Shen and Huang [9] combine the ideas of dimension reduction and time series forecasting; hence the method is similar to the one proposed by them in Ref. 13. Two differences between Refs 9 and 13 are (i) the former does not make any assumption on the arrival process, while the latter assumes that the call arrivals follow a time-inhomogeneous Poisson process; (ii) the former produces forecasts for future call volumes while the latter generates future rate forecasts. The authors [9] treat the square-rooted intraday call volume profiles as a high dimensional vector time series. They first reduce the dimensionality by applying singular value decomposition to the square-rooted count matrix. The dimension reduction allows

them to obtain several pairs of “most important” interday and intraday features, which naturally decouple interday and intraday dependence. Once these features are obtained, in a manner similar to Ref. 13, time series models and penalized regression techniques are used to perform interday forecasting and intraday updating. Since no distributional assumptions are made on the arrival counts, nonparametric bootstrapping is proposed to derive distributional forecasts.

Similar to Aldor-Noiman *et al.* [8], Taylor [15] views the historical data as a “long” univariate time series, and compares the empirical performance of several univariate time series methods for forecasting intraday call arrivals. The methods studied include seasonal ARIMA time series models, periodic autoregressive models, exponential smoothing models with two seasonal cycles (for intraday and intraweek), robust exponential smoothing, and dynamic harmonic regression. The author focuses on point forecasting, and considers forecasting lead times ranging from one half-hour ahead to two weeks ahead. No clear winning method is identified, although seasonal ARIMA and exponential smoothing with two seasonal cycles appear to perform better for lead times up to two days ahead.

CUSTOMER PATIENCE AND AGENT SERVICE TIMES

Call-center operations are complicated partly due to the human factors involved in particular, customers and agents. In call centers, it is people, rather than “jobs,” that require service and it is people, rather than equipments, that provide the service. Call-by-call operational data also provide opportunities for researchers to empirically study customer and agent behaviors in a call-center environment. Although research in this area is still lacking, interesting and important developments have been made recently. The review below focuses on statistical analysis of customer patience and agent service times.

Customer Abandonment and Patience

As discussed earlier, for customers who want to speak with agents, chances are that they

need to wait in the service queue before getting served. Some of them will become impatient, and abandon the system before an agent becomes available. Hence, it is important to understand customer abandonment or patience behavior. In addition, a wrong model of customer patience can lead to wrong staffing [16].

Traditional queueing theory has been naive in its modeling of customer patience, which is often ignored. If at all acknowledged, customers are usually assumed to arrive to the system with independent and identically distributed patience. Furthermore, the patience distribution is most often assumed to be exponential as in the Erlang-A model [6].

Brown *et al.* [5] analyze customer patience empirically at the Israeli call center. They propose to use *the time that a customer is willing to wait before hanging up* as an ancillary measure for the customer's patience. Note that this measure is observed only for a customer who does hang up. For a customer who gets served, we only know that he or she is willing to wait longer than his or her actual waiting time; however, we do not know exactly how much longer, in which case his or her patience is considered to be *right censored*. For statistical analysis of customer patience, the censoring has to be dealt with using survival analysis techniques.

Several interesting empirical observations are revealed by the analysis. First, a clear stochastic ordering emerges among customers seeking different services. For example, stock-trading customers are more patient than customers seeking regular services. Similarly, high priority customers appear to be more patient than regular customers [1]. Second, customers are more likely to abandon the queue after waiting for either only a few seconds or about 60 s. One plausible explanation is that the call-center system plays a "Please wait" message after a customer waits a few seconds and again at 1 min, which apparently prompts the customers to abandon the queue. This also suggests that the patience distribution is not exponential. If it were, the probability for a customer to abandon would remain the same, no matter how long she had waited.

Furthermore, the authors demonstrate empirically that the Erlang-A model is very robust against the violation of the exponential assumption of customer patience. The observation of the robustness motivates further research to explore for theoretical justification, which is later provided by Zeltyn and Mandelbaum [16].

Adaptivity of Customer Patience

Traditional queueing theory also assumes that customer patience is independent of system performance, unrelated to past experience or future anticipation of the system. However, Zohar *et al.* [17] provide some empirical evidence to support an inconsistent hypothesis: experienced customers "adapt" their patience on the basis of their expectations regarding system waiting time.

Using the Israeli call-center data, for every 15-min interval during weekdays between 7.00 a.m. and midnight, the *percentage of abandonment* of those customers who wait as well as the *average waiting time* for the same customers are calculated. Surprisingly, for experienced customers, the percentage of abandonment remains at about 38%, regardless of what the average waiting time is, which ranges between 90 and 250 s; on the other hand, for novice customers, the two measures are strongly positive linear dependent. In addition, for experienced customers, the expected patience increases linearly when the system is more congested (and hence needs the customers to wait longer). This also suggests that the experienced customers know what to expect on the basis of their past experience, and when the system needs them to wait longer, they are willing to do so.

Agent Service Times

As providers of service to customers via telephones, call-center agents are another important human component of call-center operations. However, even less empirical research has been devoted to study them.

Traditionally, queueing models assume that service times within a given interval are independent and identically distributed according to an exponential distribution. This

assumption is imposed mainly for theoretical simplicity due to the lack-of-memory property of exponential distributions. Although some empirical evidence exists for the exponentiality, service times are not always exponentially distributed. For example, Brown *et al.* [5] analyze the individual call service times in the Israeli call center, and show that they are lognormally distributed, that is, their natural logs are normally distributed. This fact makes the natural logs appealing for statistical analysis. Furthermore, the lognormal distribution provides an excellent fit not only for the overall service time but also when restricted to service types, individual agents, time-of-day, and so on.

Using the call-by-call data, the authors also find out that, in the early portion of the data, a significant amount of calls (7%) are handled in less than 10 s, comparing to 2% of such calls during the latter portion. Service times shorter than 10 s are questionable. As a matter of fact, those short service times are primarily caused by agents who simply hang up on customers to obtain extra rest time. This phenomenon of agents “abandoning” customers is often due to distorted incentive schemes, especially those that overemphasize short average service time or the total number of calls handled by an agent.

FUTURE CHALLENGES

We conclude this article with a brief discussion of several challenging research problems. Empirical analysis of call-by-call operational data should shed light on all of them, influencing how they are modeled by researchers, as well as how they are managed in practice.

The first challenge is about workload forecasting. Most research reviewed in the section titled “Call Forecasting” focuses on forecasting future arrival rates. However, to support agent staffing, a call-center manager requires forecasts of future workload, which combines arrivals with the service times of these arrivals. This is an important but difficult problem for time-varying arrivals, because there exists a time lag between peak arrivals and peak congestion [18].

More work is needed to develop methods that forecast arrival rates and service rates simultaneously. Some interesting contribution has been made recently at the Technion, where a procedure is proposed to estimate the service-time distribution of those abandoning customers.

The second challenge concerns the use of IVRs, which enable customers to use self-service without speaking to an agent. Little is formally known about the statistical or operational properties of IVRs and how they should be managed. At a high level, customer interaction with an IVR can be thought of as a service time, to which queueing-theoretic concepts can be applied. However there has been very limited theoretical work on the use of IVRs. To summarize, the time a customer spends in the IVR, as well as how that time affects subsequent agent service time, are important elements of call-center operations about which little is known.

The third challenge has to do with customer retrieval behavior. More specifically, many of the impatient abandoning customers then become so-called retrials, calling again at a later time. Similarly, customers that receive busy signals, as well as those that are served, may become retrials. While there are classical theoretical models of retrieval queues, again little is known about the actual retrieval behavior of customers. Careful analysis of retrieval data should help to shape how retrials are modeled and managed in the future. Some ongoing research at the Technion is analyzing customer retrials using frailty models.

The analyses of IVRs and customer retrials are interesting in and of themselves. They are also important because of their potential impact on the statistical properties of the stream of calls that arrives to agents. Both IVR and retrieval behavior have the potential to change important statistical properties of the pattern of arrivals to agents. For example, the calls exiting a congested IVR may not be Poisson.

While waiting, customers usually listen to music, commercials, or updates of their positions in the queue. These messages are referred to as *waiting-time fillers* in consumer psychology. How customers react to such

time fillers is another important but under-explored research problem. Brown *et al.* [5] observe empirically that a “Please wait” message actually prompts customers to hang up. Recently, Munichor and Rafaeli [19] used lab experiments to study caller reactions to three types of telephone waiting-time fillers: music, apologies, and information about location in the queue. On the basis of the callers’ evaluations after getting their services, the authors conclude that callers are happier when they are given a sense of progress through periodic updates of their position in the queue.

Furthermore, a better empirical understanding of service-time heterogeneity is needed. Empirical research has shown that differences among service times are associated with service types and time-of-day. There is much more work that could and should be done to better understand the drivers of service times. For example, it is interesting to understand how the heterogeneity is caused by agent-specific factors such as learning, forgetting, shift fatigue, and cross-training. Such research findings can complement and benefit recent theoretical work addressing agent heterogeneity in call centers.

Finally, the existence of heterogeneous customer classes and agent types and the employment of skills-based routing introduce additional complications and present many interesting questions. For example, in the context of workload prediction, the prediction depends in some sense on how customers are routed to agents; if one routes more customers to the slow agents, then the workload will be higher. Skills-based routing decisions also affect customers’ waiting experience and their patience.

REFERENCES

1. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. *Manuf Serv Oper Manage* 2003;5:79–141.
2. Akşın Z, Armony M, Mahrotra V. The modern call-center: a multi-disciplinary perspective on operations management research. *Prod Oper Manage* 2007;16:665–688.
3. Mandelbaum A. Call centers: research bibliography with abstracts. Version 7, 2006 May. Technical Report, Technion–Israel Institute of Technology. Available at http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf.
4. Batt R, Doellgast V, Kwon H. US call center industry report 2004: national benchmarking report strategy, HR practices & performance. CAHRS Working Paper No. 05–06. Ithaca (NY): Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies; 2005.
5. Brown LD, Gans N, Mandelbaum A, *et al.* Statistical analysis of a telephone call center: a queueing science perspective. *J Am Stat Assoc* 2005;100:36–50.
6. Garnett O, Mandelbaum A, Reiman M. Designing a call center with impatient customers. *Manuf Serv Oper Manage* 2002; 4: 208–227.
7. Trofimov V, Feigin P, Mandelbaum A, *et al.* DATA-MOCCA: Data Model for Call Center Analysis. Volume 1: model description and introduction to user interface. Technical Report, Technion–Israel Institute of Technology; 2006. Available at <http://iew3.technion.ac.il/serveng/References/DataMOCCA.pdf>.
8. Aldor-Noiman S, Feigin PD, Mandelbaum A. Workload forecasting for a call center: methodology and a case study. *Ann Appl Stat* 2009. In press.
9. Shen H, Huang JZ. Interday forecasting and intraday updating of call center arrivals. *Manuf Serv Oper Manage* 2008;10:391–410.
10. Maman S. Uncertainty in the demand for service: the case of call centers and emergency departments [Masters Thesis]: Technion–Israel Institute of Technology; 2009. Available at http://iew3.technion.ac.il/serveng/References/Thesis_Shimrit.pdf.
11. Weinberg J, Brown LD, Stroud JR. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J Am Stat Assoc* 2007;102:1185–1199.
12. Avramidis AN, Deslauriers A, L’Ecuyer P. Modeling daily arrivals to a telephone call center. *Manage Sci* 2004;50:896–908.
13. Shen H, Huang JZ. Forecasting time series of inhomogeneous Poisson processes with applications to call center workforce management. *Ann Appl Stat* 2008;2:601–623.
14. Soyer R, Tarimcilar MM. Modeling and analysis of call center arrival data: a Bayesian approach. *Manage Sci* 2008;54:266–278.
15. Taylor J. A comparison of univariate time series methods for forecasting intraday

- arrivals at a call center. *Manage Sci* 2008; 54:253–265.
16. Zeltyn S, Mandelbaum A. Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst* 2005;51:361–402.
17. Zohar E, Mandelbaum A, Shimkin N. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Manage Sci* 2002;48:566–583.
18. Feldman Z, Mandelbaum A, Massey WA, *et al.* Staffing of time-varying queues to achieve time-stable performance. *Manage Sci* 2009;54:324–338.
19. Munichor N, Rafaeli A. Numbers or apologies? Customer reactions to tele-waiting time fillers. *J Appl Psychol* 2007;92:511–518.

STATISTICAL METHODS FOR OPTIMIZATION

L. JEFF HONG

Department of Industrial
Engineering and Logistics
Management, The Hong Kong
University of Science and
Technology, Clear Water Bay,
Hong Kong, China

INTRODUCTION

In this chapter, we consider the following stochastic optimization problem

$$\text{maximize } \{g(\mathbf{x}) = \mathbb{E}_{\mathbf{x}} [Y(\mathbf{x})]\}, \quad \mathbf{x} \in \Theta \subset \mathbb{R}^d, \quad (1)$$

where $\mathbf{x}$ is the vector of decision variables (also called a solution) in $\mathbb{R}^d$ and Y is a random variable whose distribution depends on $\mathbf{x}$. In other words, we try to maximize an objective function that is defined as the expected value of a random variable whose distribution depends on the decision variable $\mathbf{x}$. To evaluate the objective function, we can only run experiments at a particular value of $\mathbf{x}$. In this article, we only consider simulation experiments that are conducted in a computer. The output of an experiment is only a realization of $Y(\mathbf{x})$. Furthermore, the distribution of the output may vary significantly over the feasible region, and little is known about its properties (e.g., differentiability and convexity) of the (implied) objective function.

Depending on the structures of the solution $\mathbf{x}$ and feasible region Θ , many statistical methods have been proposed to solve Problem 1. When Θ is a continuous subset of $\mathbb{R}^d$, we may solve the problem using stochastic approximation algorithms or response surface methodologies; when $\mathbf{x}$ is discrete and integer ordered, $|\Theta|$ may be large or even infinite and we may use random search algorithms. All these methods are introduced in

other articles in this encyclopedia. In this article, we focus on the ranking and selection (R&S) problem, where Θ has a small number of solutions (often less than 500); we may simulate all solutions and select the best among them.

Many practical problems can be formulated as an R&S problem. Here are some examples:

- choice of product variants to stock in inventory to maximize expected profit when demands are stochastic;
- allocation of buffer capacity along an auto assembly line to minimize the sum of work-in-process of all stations;
- determination of the number of automatic guided vehicles (AGVs) to balance the investment in AGVs and the average storage-and-retrieval time of a material handling system.

R&S problems are classical statistical problems. They have been studied extensively in the statistics and simulation literature. For a more comprehensive review of the topics, readers are referred to the monograph of Bechhofer *et al.* [1] and the review article by Kim and Nelson [2].

BASIC FORMULATIONS

Suppose that there are totally $k \geq 2$ solutions in Θ , denoted as $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$. We let $Y_j(\mathbf{x}_i)$ denote the j th observation from simulating solution $\mathbf{x}_i$, and assume that $Y_j(\mathbf{x}_i)$ follows a normal distribution of mean $g(\mathbf{x}_i)$ and variance σ_i^2 , where $g(\mathbf{x}_i)$ is unknown. Then, solving Problem 1 is equivalent to selecting the solution among $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ that has the largest mean. To further simplify the notation, we let $\mu_i = g(\mathbf{x}_i)$ and $Y_{ij} = Y_j(\mathbf{x}_i)$. Without loss of generality, we let $\mu_k \geq \mu_2 \geq \dots \geq \mu_1$ and the goal is to find which solution is $\mathbf{x}_k$.

Because Y_{ij} , $i = 1, \dots, k$, are normal random variables, no statistical methods

can guarantee to find the best solution with probability 1 using a finite number of observations. Therefore, all R&S procedures can only guarantee to find the best solution with a certain predetermined and large probability $1 - \alpha$, where α is often set to 0.1, 0.05 or 0.01.

Furthermore, because the difference between μ_k and μ_{k-1} can be arbitrarily close to zero, it imposes both theoretical and practical difficulties on the problem. Theoretically, without knowing the difference between μ_k and μ_{k-1} , it is difficult to decide the (finite) sample sizes for all solutions that guarantee to identify $\mathbf{x}_k$ with a probability $1 - \alpha$. Practically, if μ_k and μ_{k-1} are close to each other, it may require a large amount of simulation effort to differentiate them. In many practical situations, however, we may think both $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ are acceptable if their means are close to each other and hence, the effort is essentially wasted. To resolve these difficulties, two formulations were proposed to soften the goal of finding $\mathbf{x}_k$ with a probability $1 - \alpha$. In the first formulation, the goal is softened to finding a subset of solutions that contains $\mathbf{x}_k$, which is known as the “subset selection formulation”. In the second formulation, the goal is softened to finding $\mathbf{x}_k$ if $\mu_k - \mu_{k-1} \geq \delta$, which is known as the “indifference-zone (IZ) formulation”.

SUBSET-SELECTION FORMULATION

The subset selection formulation was proposed by Gupta [3]. It supposes that there are $n_i > 0$ i.i.d. observations from solution $\mathbf{x}_i$ for all $\mathbf{x}_1, \dots, \mathbf{x}_k$. The goal is to use these data to obtain a subset $I \subset \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ such that

$$\Pr\{\mathbf{x}_k \in I\} \geq 1 - \alpha. \quad (2)$$

As an example of a subset selection procedure, we consider the classical Gupta’s procedure, where each solution has n observations, $\sigma_1^2 = \dots = \sigma_k^2 = \sigma^2$, and σ^2 is known.

Gupta’s Procedure.

Step 1. Determine the constant h , which satisfies $\Pr\{Z_i \leq h, i = 1, 2, \dots, k-1\} =$

$1 - \alpha$ where $(Z_1, Z_2, \dots, Z_{k-1})$ have a multivariate normal distribution with means 0, variances 1, and common pairwise correlations $1/2$.

Step 2. For all $i = 1, 2, \dots, n$, calculate $\bar{Y}_i(n) = \frac{1}{n} \sum_{j=1}^n Y_{ij}$.

Step 3. Let the set I include all solutions $\mathbf{x}_\ell$ such that

$$\bar{Y}_\ell(n) \geq \max_{i \neq \ell} \bar{Y}_i(n) - h\sigma\sqrt{2/n}. \quad (3)$$

Note that

$$\begin{aligned} \Pr\{\mathbf{x}_k \in I\} &= \Pr\left\{\bar{Y}_k(n) \geq \bar{Y}_i(n) - h\sigma\sqrt{\frac{2}{n}}, \right. \\ &\quad \left. i = 1, 2, \dots, k-1\right\} \\ &= \Pr\left\{\frac{\bar{Y}_i(n) - \bar{Y}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} \right. \\ &\quad \left. \leq h - \frac{\mu_i - \mu_k}{\sigma\sqrt{2/n}}, i = 1, 2, \dots, k-1\right\} \\ &\geq \Pr\{Z_i \leq h, i = 1, 2, \dots, k-1\} \\ &= 1 - \alpha, \end{aligned} \quad (4)$$

where Equation (4) follows from the facts that $\mu_i - \mu_k \leq 0$ and $\bar{Y}_i(n) - \bar{Y}_k(n)$ is a normal random variable with mean $\mu_i - \mu_k$ and variance $2\sigma^2/n$ for all $i = 1, \dots, k-1$. Therefore, Gupta’s procedure is valid, that is it guarantees Equation (2).

Gupta’s procedure includes many of the elements that are essential to a subset selection procedure. It determines the screening criteria, that is Equation (3), based on the variances of the solutions and the fact that $\mu_i - \mu_k \leq 0$ for all $i = 1, \dots, k-1$. At the end of the procedure, it gives a subset that contains the optimal solution $\mathbf{x}_k$ with a probability greater than or equal to $1 - \alpha$. However, there is no guarantee on the size of the subset. It may contain multiple solutions.

INDIFFERENCE-ZONE FORMULATION

The IZ formulation was proposed by Bechhofer [4]. It supposes that $\delta > 0$ is the smallest difference that the experimenter feels

is worth detecting, where δ is known as the IZ parameter. The goal is to guarantee that

$$\Pr\{\text{selecting } \mathbf{x}_k\} \geq 1 - \alpha \quad (5)$$

if $\mu_k - \mu_{k-1} \geq \delta$.

As an example of an IZ selection procedure, we consider the classical Bechhofer's procedure, where each solution has n observations, $\sigma_1^2 = \dots = \sigma_k^2 = \sigma^2$, and σ^2 is known.

Bechhofer's Procedure.

Step 1. Determine the constant h , which is the same as the one in Gupta's procedure. Let

$$n = \left\lceil \frac{2h^2\sigma^2}{\delta^2} \right\rceil.$$

Step 2. Take n i.i.d. observations from each solution and calculate $\bar{Y}_i(n)$ for all $i = 1, 2, \dots, k$.

Step 3. Select the solution with the largest sample mean $\bar{Y}_i(n)$ as the best.

Note that

$$\begin{aligned} & \Pr\{\text{selecting } \mathbf{x}_k\} \\ &= \Pr\{\bar{Y}_k(n) \geq \bar{Y}_i(n), i = 1, 2, \dots, k-1\} \\ &= \Pr\left\{\frac{\bar{Y}_i(n) - \bar{Y}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} \leq -\frac{\mu_i - \mu_k}{\sigma\sqrt{2/n}}, i = 1, 2, \dots, k-1\right\} \\ &\geq \Pr\{Z_i \leq h, i = 1, 2, \dots, k-1\} \quad (6) \\ &= 1 - \alpha, \end{aligned}$$

where Equation (6) follows from the fact that $\mu_k - \mu_i \geq \delta$ for all $i = 1, \dots, k-1$. Therefore, Bechhofer's procedure is valid, that is, it guarantees Equation (5).

Bechhofer's procedure includes many of the elements that are essential to an IZ selection procedure. It determines the sample sizes required for all k solutions based on the variances of the solutions and the

fact that $\mu_k - \mu_i \geq \delta$ for all $i = 1, \dots, k-1$. At the end of the procedure, it selects the solution with the largest sample mean after all solutions acquired the required number of samples. Furthermore, because the procedure is designed for the least favorable configuration (LFC), where $\mu_k - \mu_i = \delta$ for all $i = 1, \dots, k-1$, it is typically conservative, achieving a probability of correct selection of more than $1 - \alpha$, when the configuration of the solutions is not the LFC.

DESIGNING R&S PROCEDURES FOR COMPUTER EXPERIMENTS

Classical R&S procedures were designed for real experiments, for example agricultural or pharmaceutical experiments, which often take a long time to complete. When solving Problem (1), however, we typically consider computer experiments that can be completed in minutes or even seconds. The nature of computer experiments introduces some new features to R&S problems. In this section, we consider several of these features and discuss how R&S procedures may take these features into consideration.

UNKNOWN AND UNEQUAL VARIANCES

Both Gupta's and Bechhofer's procedures assume that variances are known and equal. When conducting computer experiments, however, we often explore a relatively large number of solutions which may differ significantly. The assumption of equal known variances is no longer reasonable. Much work has been done to relax this assumption and to allow unknown and unequal variances.

For subset selection procedures, samples from all solutions are available before applying the procedure. Therefore, one may use these samples to estimate the variances first and then use the estimated variances to determine a screening criterion. Note that when variances are estimated, procedures need to be modified or adjusted so that they account for the variability of estimated variances. The procedure of Boesel *et al.* [5] is an

example of such a procedure. It first calculates the sample variances S_i^2 , where

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

for all $i = 1, \dots, k$. It then includes in the subset all $x_i, i = 1, \dots, k$, such that

$$\bar{Y}_i(n_i) \geq \max_{j \neq i} \left\{ \bar{Y}_j(n_j) - \left(t_i^2 \frac{S_i^2}{n_i} + t_j^2 \frac{S_j^2}{n_j} \right)^{1/2} \right\},$$

where $t_i = t_{(1-\alpha)^{1/(k-1)}, n_i-1}$ is the $(1-\alpha)^{1/(k-1)}$ -quantile of a t -distribution with $n_i - 1$ degrees of freedom.

For IZ selection procedures, one needs to determine the sample sizes for all solutions based on the variances. Therefore, all of these procedures need a first stage to collect some samples from every solution to estimate the variances. They then use the estimated variances to determine the sample size of every solution in the second stage. After collecting the second-stage samples, the procedures select the solution with the largest overall sample mean, calculated from both the first- and second-stage samples. Rinott's procedure [6] is an example of such a procedure. In the first stage, it takes n_0 samples from all solutions, calculates their sample variances S_i^2 , and determines the final sample sizes

$$N_i = \max \left\{ n_0, \left\lceil \frac{h^2 S_i^2}{\delta^2} \right\rceil \right\}$$

for all $i = 1, \dots, k$, where h is a constant that solves Rinott's integral for n_0, k , and α ([7]). In the second stage, it takes $N_i - n_0$ samples from $\mathbf{x}_i$ and selects the solution with largest $\bar{Y}_i(N_i)$.

COMMON RANDOM NUMBERS

All procedures considered so far use independent samples for different solutions. However, it is not necessary. The randomness in computer experiments is created by using pseudorandom numbers [8]. This allows the experimenter to use the same stream of pseudorandom numbers for all solutions and thus, to introduce positive correlations between

$Y_{i\ell}$ and $Y_{j\ell}$, and to sharpen the comparisons between any pair of solutions $\mathbf{x}_i$ and $\mathbf{x}_j$ (for instance, the examples 14.10 and 14.11 in Ref. 9 show that common random numbers (i.e., CRNs) can lead to dramatic data savings).

When common random numbers are used, we typically calculate the sample variances S_{ij}^2 of $Y_{i\ell} - Y_{j\ell}$ for all $i \neq j$ and use the variances to conduct subset selection or IZ selection. As an example, we consider Clark and Yang's IZ selection procedure [10]. The procedure first takes n_0 observations from all solutions in the first stage, calculates the sample variances

$$S_{ij}^2 = \frac{1}{n_0 - 1} \sum_{\ell=1}^{n_0} \left[(Y_{i\ell} - Y_{j\ell}) - (\bar{Y}_i(n_0) - \bar{Y}_j(n_0)) \right]^2$$

for all $i \neq j$, and computes the final sample size

$$N = \max \left\{ n_0, \max_{i \neq j} \left\lceil \frac{t^2 S_{ij}^2}{\delta^2} \right\rceil \right\},$$

where $t = t_{1-\alpha/(k-1), n_0-1}$. In the second stage, it takes $N - n_0$ samples from all x_i and selects the solution with largest $\bar{Y}_i(N)$. The procedure ensures its validity by using the Bonferroni inequality that is typically very conservative when the number of solutions is large. Nelson and Matejcek [11] also considered this problem for equal-variance cases. They introduced the concept of sphericity to avoid using the Bonferroni inequality.

SEQUENTIAL NATURE OF COMPUTER EXPERIMENTS

In many real experiments, samples are taken in a batch. For instance, when testing the effectiveness of a drug, a sample of n patients may take the drug at the same time. It is often not practical to give the drug to only one patient at a time and wait to see the result before giving the drug to another. However, computer experiments are often conducted on just a single computer. Then, the samples are taken one at a time. In this situation, it makes sense to conduct subset selections

sequentially until only one solution remains in the subset. This is essentially the basic idea behind many sequential IZ selection procedures.

Many of the sequential IZ selection procedures were designed based on a Brownian motion approximation. Let

$$Z_{ij}(n) = \frac{n}{\sigma_{ij}^2} [\bar{Y}_i(n) - \bar{Y}_j(n)],$$

where $\sigma_{ij}^2 = \text{Var}(Y_{i\ell} - Y_{j\ell})$, which is $\sigma_i^2 + \sigma_j^2$ if $Y_{i\ell}$ and $Y_{j\ell}$ are independent, and let $\mathcal{B}_\Delta(t)$ denote a Brownian motion process with a drift Δ evaluated at time $t > 0$. Then, one can show that $Z_{ij}(n), n = 1, 2, \dots$, have the same joint distribution as $\mathcal{B}_{\mu_i - \mu_j}(n/\sigma_{ij}^2), n = 1, 2, \dots$

Given the Brownian motion approximation, we can then apply the results of Brownian motion process to design sequential IZ selection procedures. An oft-used result is the Fabian's bound [12]. Let Λ denote a triangular continuation region formed by $-a + \lambda t$ and $a - \lambda t$, where $a > 0, \lambda = \Delta/2 > 0$ and $t \geq 0$. Let $T = \inf\{t > 0 : \mathcal{B}_\Delta(t) \notin \Lambda\}$, T is the random time that $\mathcal{B}_\Delta(t)$ first exists Λ . Then, Fabian [12] showed that

$$\Pr\{\mathcal{B}_\Delta(T) < 0\} = \frac{1}{2} e^{-a\Delta}. \quad (7)$$

Note that it is the probability that a Brownian motion process with a positive drift Δ exists in the triangular continuation region from the negative side.

The KN procedure of Kim and Nelson [13] uses Equation (6) to select appropriate triangular region parameters (a, λ) to ensure the probability of correct selection. Suppose that $\mathbf{x}_i$ remains until the $n - 1$ sample. Then, after taking an additional sample for every remaining solution, $\mathbf{x}_i$ will be eliminated if

$$\frac{n}{S_{ij}^2} [\bar{Y}_i(n) - \bar{Y}_j(n)] < -a + \lambda \frac{n}{S_{ij}^2}$$

for any remaining $\mathbf{x}_j$. Note that the elimination criterion corresponds to the existing triangular region from the negative side if $\mu_i - \mu_j \geq \delta$. The KN procedure allows the use of common random numbers, is statistically

valid, and is much more efficient than typical two-stage procedures, especially when some of the differences between μ_k and $\mu_i, i = 1, \dots, k - 1$, are significantly larger than δ . Kim and Nelson [14] extended the procedure to allow weak dependence among the samples that are taken from the same solution. Batur and Kim [15] designed a fully sequential IZ procedure with a continuation region that is formed by parabolic boundaries.

Hong [16] further extended the Brownian motion approximation and showed that

$$Z_{ij}(m, n) = \left(\frac{\sigma_i^2}{m} + \frac{\sigma_j^2}{n} \right)^{-1} [\bar{Y}_i(m) - \bar{Y}_j(n)]$$

and $\mathcal{B}_{\mu_i - \mu_j} \left(\left(\sigma_i^2/m + \sigma_j^2/n \right)^{-1} \right)$ have the same joint distribution. This result can be used to design sequential IZ selection procedures that allows different sample sizes for different solutions. It is especially helpful when σ_i^2 and σ_j^2 are significantly different.

SAMPLING VERSUS SWITCHING

Typical fully sequential IZ selection procedures, such as the KN procedure, take a sample at a time from every remaining solution and reduce the total sampling cost by allowing early termination. It also requires repeated switching among the computer models of different solutions. Unfortunately, the computational overhead of switching can be significant, sometimes orders of magnitude more than that of sampling. Therefore, these sequential procedures may actually be slower than the two-stage procedures (e.g., Rinott's procedure) that require at most $2k$ switchings, even though they may require a smaller number of samples to make the selection decision.

To design sequential selection procedures that account for switching costs, Hong and Nelson [17] proposed two strategies. In both strategies, the procedures have a first stage that estimates the sample variances of all solutions, computes the parameters of the continuation region, and decides the maximal sample size that a solution may have before the selection decision is made. In the

first strategy, which is known as the *minimum switching strategy*, the procedure sorts the first-stage sample means and allocates the maximal sample size to the solution with the highest first-stage sample mean. It then allocates one sample at a time to the second-best solution and compares the second-best with the best to see if either one may be eliminated. If the second-best is eliminated, the best remains to be the current best; otherwise, the second-best becomes the current best and we allocate the maximal sample size to it. Then, we take a sample at a time from the third-best, compare it to the current best, and so on until only one solution remains. We declare the remaining solution as the best solution. In this strategy, the current sample best solution always gets the maximal sample size and we switch at most k times after the first stage (which switches k times). Therefore, it has a minimal $2k$ switchings. In the second strategy, Hong and Nelson [17] suggested using dynamic decision making for the number of switchings, based on the sampling and switching costs.

To implement both strategies, Hong and Nelson [17] design sequential IZ selection procedures based on Fabian's bound. Osogami [18] further extended the minimal switching procedure and used Anderson's bound instead of Fabian's bound, and the resultant procedure is more efficient than the one used by Hong and Nelson [17].

LARGE NUMBER OF SOLUTIONS

In many optimization problems, the number of solutions may be large. In such a case, IZ selection procedures are typically not very efficient. When the number of solutions is large, the differences between the best solution and some of the not-so-good solutions can be quite large. Then, these solutions may be easily eliminated by a subset selection procedure.

Nelson *et al.* [19] proposed a decomposition approach that uses a subset selection in the first stage to screen out clearly inferior solutions and applies an IZ selection to select the best solution in the second stage. They showed that, if the first stage has an error probability α_1 and the second stage

has an error probability α_2 , the error probability of the entire procedure is bounded above by $\alpha_1 + \alpha_2$. Therefore, to ensure that the procedure selects the best solution with a probability at least $1 - \alpha$, they suggested designing both stages with a $1 - \alpha/2$ probability of correct decision and developed such a procedure.

As we mentioned in the section titled "Sequential Nature of Computer Experiments," sequential procedures are much more efficient than typical two-stage procedures when there are clearly inferior solutions. Therefore, they also work well when the number of solutions is large.

EMBEDDING R&S IN OTHER OPTIMIZATION ALGORITHMS

To use R&S procedures to solve stochastic optimization problems, all solutions need to be sampled and compared. When the number of solutions is very large (or even infinite), R&S procedures become infeasible. However, many other stochastic optimization algorithms can solve such problems. The basic idea of many such algorithms is to use problem structure to gradually find the optimal solution by evaluating only a small fraction of the solutions. In this subsection, we summarize methods that embed R&S procedures in different levels of those algorithms to improve their performances.

Boesel *et al.* [5] proposed the "cleanup" idea, which applies an R&S procedure at the end of the optimization process to select the best solution among all solutions that are visited by the optimization algorithm. Because an optimization algorithm may visit many solutions before it stops, their procedure implements the decomposition approach of the previous section, uses the subset selection procedure of the section titled "Unknown and Unequal Variances" to screen out clearly inferior solutions by using the samples collected in the optimization process, and then applies a Rinott-type procedure of the same section to select the best solution. Their procedure guarantees that the solution found at the end is the best solution among all visited solutions with a probability of at least $1 - \alpha$.

Many optimization algorithms compare the solutions from a neighborhood and select

the best one. Pichitlamken *et al.* [20] proposed a sequential selection procedure to conduct this neighborhood selection. Their procedure takes into the consideration of the fact that in the neighborhood selection problem, different solutions may have different numbers of samples and only the sample means of those solutions may be available. They showed that their procedure achieves the required probability of correct selection under the typical IZ formulation and works well when embedded in optimization algorithms. Similar to the neighborhood selection problem, Xu *et al.* [21] proposed a sequential procedure that tests the local optimality of a solution among its neighboring solutions.

In the optimization process, the solutions are visited sequentially and we often do not know in advance how many solutions the algorithm will visit. This imposes a difficulty in the design of selection procedures, because the number of solutions (k) is often a necessary design parameter. Hong and Nelson [22] proposed two strategies to solve this problem of not knowing k in advance: the single-elimination strategy and the stop-and-go strategy. In the first strategy, a solution that is eliminated will not be considered again. In the second strategy, an eliminated solution will be considered again after more solutions are visited. Hong and Nelson [22] designed statistically valid fully sequential procedures to implement both strategies.

So far, we have only considered those optimization problems where the randomness exists only in the objective function. In many optimization problems, however, there is also randomness in the constraints. Then, identifying the feasibility of a solution is as important as selecting the best solution. Batur and Kim [15] considered how to find feasible solutions among a set of solutions in the presence of multiple stochastic constraints; Andradóttir and Kim [23] considered how to select the best feasible solution when there is a stochastic constraint.

CONCLUDING REMARKS

1. In this article, we consider the selection of the solution with the best mean

and the output of the computer experiments at the solution is modeled as a normal random variable. There are also other types of selection problems. For instance, Bernoulli selection problem selects the solution with the highest probability of success [24], multinomial selection problem selects the solution that is most likely to be the best [25], and comparison-with-a-standard problem selects the best solution only if it is clearly better than the standard [26,27].

2. The formulations and procedures introduced in this article are all frequentist's procedures. There are also Bayesian formulations of R&S problems; for example, the optimal computing budget allocation of Chen *et al.* [28] and the Bayesian procedures of Chick and Inoue [29,30]. They solve the problem of allocating a fixed number of samples to maximize the posterior probability of correct selection. Branke *et al.* [31] compared different procedures, frequentist's and Bayesian, through comprehensive numerical study. They concluded that Bayesian procedures are typically less conservative, and thus more efficient.
3. Selection of the best is related to another statistical problem known as *multiple comparisons*, which compares the means of all k solutions [1]. For instance, the problem of all pairwise comparisons aims to provide a joint confidence interval for $\mu_i - \mu_j$ for all $i \neq j$. The Tukey–Kramer test ([32]) and the Bonferroni approach ([33]) are often used to solve the problem.
4. R&S procedures have been implemented by some commercial simulation and optimization software packages. For instance, the Process Analyzer of Arena® simulation software implements the subset selection procedure of Nelson *et al.* [19] to compare different scenarios; the simulation optimization software OptQuest® implements the IZ selection procedure of Pichitlamken *et al.* [20] to conduct a “cleanup”, that is selecting the best among all visited

solutions after the optimization process finishes; and the Scenario Selector of AweSim® simulation software implements a number of R&S procedures including both, subset selection procedures and IZ selection procedures to compare different scenarios [34].

Acknowledgments

The author would like to thank Barry L. Nelson for teaching him R&S and for years of collaboration on R&S. The research is supported in part by the Hong Kong Research Grants Council under the grants GRF 613706 and GRF 613907.

REFERENCES

1. Bechhofer RE, Santner TJ, Goldsman DM. Design and analysis of experiments for statistical selection, screening, and multiple comparisons. New York: Wiley; 1995.
2. Kim SH, Nelson BL. Selecting the best system. In: Henderson SG, Nelson BL, editors. Elsevier handbooks in operations research and management science: simulation. The Netherlands: Elsevier; 2006. Chapter 17.
3. Gupta SS. On some multiple decision (Selection and Ranking) rules. *Technometrics* 1965; 7:225–245.
4. Bechhofer RE. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann Math Stat* 1954;25:16–39.
5. Boesel J, Nelson BL, Kim SH. Using ranking and selection to “clean up” after simulation optimization. *Oper Res* 2003;51:814–825.
6. Rinott Y. On two-stage selection procedures and related probability-inequalities. *Commun Stat* 1978;A7:799–811.
7. Wilcox RR. A table for Rinott’s selection procedure. *J Qual Technol* 1984;16:97–100.
8. Law AM. Simulation modeling and analysis. 4th ed. New York: McGraw-Hill; 2007.
9. Spall JC. Introduction to stochastic search and optimization: estimation, simulation, and control. Hoboken (NJ): Wiley; 2003.
10. Clark GM, Yang WN. A Bonferroni selection procedure when using common random numbers with unknown variances. *Proceedings of the 1986 Winter Simulation Conference*. Washington (DC). 1986. pp. 313–315.
11. Nelson BL, Matejcek FJ. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Manage Sci* 1995;41:1935–1945.
12. Fabian V. Note on Anderson’s sequential procedures with triangular boundary. *Ann Stat* 1974;2:170–176.
13. Kim SH, Nelson BL. A fully sequential procedure for indifference-zone selection in simulation. *ACM Trans Model Comput Simul* 2001;11:251–273.
14. Kim SH, Nelson BL. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. *Oper Res* 2006;54:475–488.
15. Batur D, Kim SH. Finding feasible systems in the presence of constraints on multiple performance measures. *ACM Trans Model Comput Simul* 2009. In press.
16. Hong LJ. Fully sequential indifference-zone selection procedures with variance-dependent sampling. *Nav Res Log* 2006;56:511–525.
17. Hong LJ, Nelson BL. The trade-off between sampling and switching: new sequential procedures for indifference-zone selection. *IIE Trans* 2005;37:623–634.
18. Osogami T. Finding probably best systems quickly via simulations. *ACM Trans Model Comput Simul* 2009;19:Article 12.
19. Nelson BL, Swann J, Goldsman D, *et al.* Simple procedures for selecting the best simulated system when the number of alternatives is large. *Oper Res* 2001;49:950–963.
20. Pichitlamken J, Nelson BL, Hong LJ. A sequential procedure for neighborhood selection-of-the-best in optimization via simulation. *Eur J Oper Res* 2006;173:283–298.
21. Xu J, Nelson BL, Hong LJ. Industrial strength COMPASS: a comprehensive algorithm and software for optimization via simulation. *ACM Trans Model Comput Simul* 2010;20:Article 3.
22. Hong LJ, Nelson BL. Selecting the best system when systems are revealed sequentially. *IIE Trans* 2007;39:723–734.
23. Andradóttir S, Kim SH. Fully sequential procedures for comparing constrained systems via simulation. *Naval Research Logistics: ISyE Department, Georgia Institute of Technology*; 2010;57:403–421. Working paper.
24. Wieland JR, Nelson BL. An odds-ratio indifference-zone selection procedure for Bernoulli populations. *The 2004 Winter Simulation Conference Proceedings*. Washington (DC). 2004; pp. 630–638.

25. Miller JO, Nelson BL, Reilly CH. Efficient multinomial selection in simulation. *Nav Res Log* 1998;45:459–482.
26. Nelson BL, Goldsman D. Comparisons with a standard in simulation experiments. *Manage Sci* 2001;47:449–463.
27. Kim SH. Comparison with a standard via fully sequential procedure. *ACM Trans Model Comput Simul* 2005;15:1–20.
28. Chen CH, Lin J, Yücesan E, *et al.* Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dyn Syst Theory Appl* 2000;10:251–270.
29. Chick SE, Inoue K. New two-stage and sequential procedures for selecting the best simulated system. *Oper Res* 2001a;49: 732–743.
30. Chick SE, Inoue K. New procedures to select the best simulated system using common random numbers. *Manage Sci* 2001b;47: 1133–1149.
31. Branke J, Chick S, Schmidt C. Selecting a selection procedure. *Manage Sci* 2007; 53:1916–1932.
32. Lomax RG. Statistical concepts: a second course for education and the behavioral sciences. 2nd ed. Mahwah (NJ): Lawrence Erlbaum; 2001.
33. Banks J, Carlson JS, Nelson BL, *et al.* Discrete-event system simulation. 5th ed. Upper Saddle River (NJ): Prentice Hall; 2010.
34. Goldsman D, Nelson BL, Opicka T, *et al.* A ranking and selection project: experiences from a university-industry collaboration. *Proceedings of the 1999 Winter Simulation Conference*. Phoenix (AZ). 1999. pp. 83–92.

STATISTICAL PROCESS CONTROL

FLAVIO SANSON FOGLIATTO

MICHEL J. ANZANELLO

Department of Industrial Engineering,
Federal University of Rio Grande
do Sul, Porto Alegre, Brazil

Statistical quality control (SQC) was conceived at Bell Telephone Laboratories by Walter A. Shewart in the 1920s. Shewart proposed the use of control charts (CCs) as the primary tool in SQC. The idea was to provide constant monitoring of processes through sampling units of production and measuring important quality characteristics (QCs). Since the seminal proposition of CCs for variables, many other charts have been presented in the literature; for example, cumulative sum, exponentially weighted moving average, and multivariate charts. Although varying in terms of construction, these CCs aim at a common objective: to monitor process variation and, in the case of multivariate charts, covariation. In SQC, quality is viewed as being inversely proportional to variability.

Process variation due to different sources is transferred to QCs and detected in their output readings. Two variation sources are important in SQC; they are named *common* and *assignable*. Common sources of variation are inherent to processes, and should not compromise the quality of their outputs. Processes exposed solely to common causes of variation are deemed as being in a state of statistical control (or in-control). On the other hand, processes exposed to assignable sources of variation are said to be out of statistical control (or out-of-control). Such sources are incidental and may be related to operator errors, variation in raw material specifications, or equipment malfunctioning, among others. CCs are useful in detecting assignable sources of variation, indicating a process change in status, from in-control to out-of-control.

CCs vary according to the QC under monitoring and how they are inputted in the chart. QCs represented by continuous measurements, such as moisture content in a food product or diameter of an auto part, are plotted in variables CCs. QCs measured as counts, such as the number of defective parts in a production lot, are plotted in attributes CCs. Data informed in a CC for variables may be inputted as individual measurements, means calculated from a group of measurements or an accumulative sum of measurements, leading to different types of CCs. In all the cases, interpretation of the information inputted in the chart is similar, as explained next.

A typical CC is given in Fig. 1. The process under surveillance is sampled at predetermined time intervals, and a measurement of interest (such as the sample average in this example) is plotted in the chart. Samples points are connected through lines to help visualizing process behavior. The CC is bounded by two lines: UCL = upper control limit and LCL = lower control limit. These limits are determined such that samples taken from a process that is in-control are expected to yield measurements predominantly within its boundaries. A center line (CL) is also displayed in the chart, giving the expected measurement value in case no assignable causes are present in the process. In-control processes are characterized by sample measurements predominantly within the control limits, randomly distributed above and below the CL. Deviations from this behavior indicate the presence of assignable causes of variation, leading to an out-of-control state. A number of rules may help in detecting an out-of-control situation in a CC. For example, when 3σ control limits are used (i.e., limits positioned at a 3σ distance from the CL, where σ is the process standard deviation), 1σ and 2σ limits may be added to the chart and the following rules applied to detect an out-of-control state: (i) two out of three consecutive points positioned

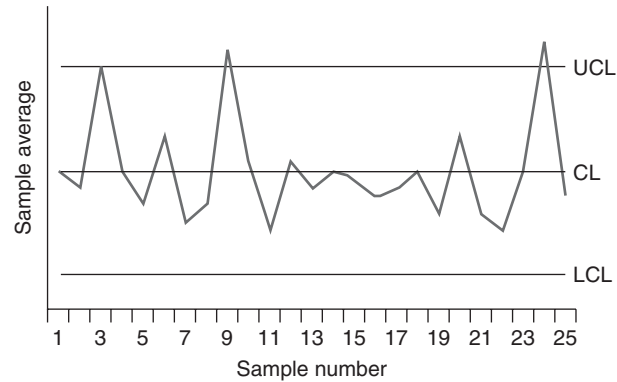


Figure 1. Representation of a typical control chart.

beyond the $\pm 2\sigma$ limits; (ii) four out of five consecutive points positioned beyond the $\pm 1\sigma$ limits; and (iii) eight consecutive points positioned in the same side of the CL.

Adequate process sampling is the key in SQC. Samples typically comprise a number of measurements gathered in subgroups from which statistics are calculated and inputted in the CC. In general, subgroup sizes should be kept small such that every sample is homogenous. The objective is to minimize the chance that when a change occurs in the process, it occurs between sampling times and not within the extraction of one sample. Three to eight data points are typical in the case of variable data; for attributes, larger samples sizes are recommended. Data points may be continuously sampled from the process and consolidated in subgroups at fixed time intervals, or collected only near the time the sampling is done. Frequent sampling of small subgroups from the process tends to give better results if compared to taking few large samples.

Another key issue in SQC is the CC design. Data inputted in a chart are assumed to follow a known probability distribution, usually the Normal. Therefore, the choice of control limits results in a probability that data points will fall within its boundaries. Consequently, there may be points that, although plotted beyond the control limits, were sampled from an in-control process. This is called a *false alarm*. The opposite may also happen, that is, samples from an out-of-control process positioned within the control limits. This is called a *failure to detect*. In well-designed CCs, the

probability of these two errors should be kept as small as possible. A quantity called the *average run length* (ARL) links these two probabilities; it represents the average number of points within control limits in a chart before a point falls beyond limits. Ideally, ARL should be maximized when the process is in-control, and minimized when the process is out-of-control. In Shewart-type charts, if p is the probability of a point plotting beyond control limits, $ARL = 1/p$, since p is constant over all samples. Finally, setting the parameters in a CC requires sampling from a process that is known to be in-control. CC design and process parameter estimation are part of what is referred to as the *Phase 1* of a CC implementation; Phase 2 takes place when the CC starts to be used for process monitoring.

Process capability is a concept related to SQC. To be considered capable, a process must produce units conforming to specifications. If specification limits on a QC are tighter than the control limits set for the same QC in a CC, then the process is likely to be considered as being not capable (in cases where the statistic plotted in the CC corresponds to the QC, e.g., when process is monitored through individuals charts, then the process will be considered as being not capable). Although out-of-control processes may still yield conforming products, such situation is not desirable since such processes are unpredictable regarding their outputs. Capability is determined through indices that basically measure a scaled distance

between control and specification limits; for example, C_p and C_{pk} .

In sections that follow, we describe the fundamentals of CC for variables and attributes, as well as modifications on traditional CCs. The chapter is concluded by a brief explanation on process capability indices.

CONTROL CHART FOR VARIABLES

Several QCs used for process control can be expressed by numerical measurements, named *variables* (e.g., dimension, volume, and weight, among others). CCs for variables aim at monitoring process location and variability. The location is monitored using an $\bar{X}$ chart (see Fig. 1), while variability can be accounted by the range CC, R , or by the standard deviation CC, S . It is important to monitor both mean and variability charts in process control procedures. A process may be centered (i.e., in control regarding the mean chart) and yet present excessive variability, which may compromise its capability; the opposite may also take place. The methods for process control described here assume that the measurements follow an approximately normal distribution.

Phase 1 of process control starts constructing the CCs. The $\bar{X}$ chart depicts the process mean by plotting the averages of samples of size n collected from sample intervals of size h . The control limits and the center line of the chart are functions of the parameters $\mu_{\bar{X}}$, which gives an estimate of the process mean μ , and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, which gives an estimate of the process standard deviation σ . These parameters are usually unknown and must be estimated from process observations. Samples consisting of 20 to 25 observations are recommended, and must be collected from in-control processes [1,2].

The average of observations x_i is calculated as in Equation (1), while an estimator of $\mu_{\bar{X}}$ is given in Equation (2):

$$\bar{x} = \sum_{i=1}^n x_i / n \quad (1)$$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \bar{x}_i}{k}, \quad (2)$$

with k as the number of samples.

An estimator of $\sigma_{\bar{X}}$ is a function of sample size and sample range, as in Equation (3). The range measure is widely used due to its simplicity and high efficiency, especially with small samples (i.e., $n = 4$ or 5).

$$\hat{\sigma}_{\bar{X}} = \frac{\bar{R}}{d_2 \sqrt{n}}. \quad (3)$$

In Equation (3), d_2 is the average relative range, defined as the relation between the range and the standard deviation of a normally distributed variable (i.e., $W = R/\sigma$), and $\bar{R}$ is the average range from the k samples. d_2 values are found in statistical tables and depend on n [1,3].

The control limits for the process mean are represented by

$$UCL = \bar{\bar{x}} + \frac{3}{d_2 \sqrt{n}} \bar{R} = \bar{\bar{x}} + A_2 \bar{R} \quad (4)$$

$$CL = \bar{\bar{x}} \quad (5)$$

$$LCL = \bar{\bar{x}} - \frac{3}{d_2 \sqrt{n}} \bar{R} = \bar{\bar{x}} - A_2 \bar{R}, \quad (6)$$

where A_2 is available in statistical tables and depends on sample size [1,4]. The limits calculated using Equations (4) and (6) are commonly referred to as 3-sigma limits.

The R CC, aimed at monitoring the range of samples consisting of n observations, is obtained in a similar manner. We first estimate μ_R e σ_R as

$$\hat{\mu}_R = \bar{R} \quad (7)$$

and

$$\hat{\sigma}_R = d_3 \frac{\bar{R}}{d_2}, \quad (8)$$

with d_3 as the standard deviation of W . The control limits for the R chart are

$$\text{UCL} = \bar{R} + 3d_3 \frac{\bar{R}}{d_2} = D_4 \bar{R} \quad (9)$$

$$\text{CL} = \bar{R} \quad (10)$$

$$\text{LCL} = \bar{R} - 3d_3 \frac{\bar{R}}{d_2} = D_3 \bar{R}. \quad (11)$$

Similar to previous cases, D_3 and D_4 can be found in statistical tables [4].

Phase 2 of process control consists of collecting further samples of size n (i.e., new observations of a running process), estimating the statistics $\bar{x}$ and $\bar{R}$, and plotting the statistics on the CC. As long as $\bar{x}$ and $\bar{R}$ remain within limits and chart points are randomly distributed along the CC, the process is considered to be in statistical control. Consider the data in Fig. 1 as observations collected in Phase 2; there are clearly two points violating the UCL, denoting that the process is out-of-control.

Montgomery [1] states that the $\bar{X}$ chart is efficient in detecting large variations on the process, for example, variations in the interval from 1.5σ to 2σ . The detection of smaller variations, however, requires other CCs such as the *CUSUM* (Cumulative Sum Control Chart) and *EWMA* (exponentially weighted moving average) charts.

The *CUSUM* chart relies on the cumulative sum of deviations between the sample values and the target value, incorporating all the information in the sample sequence [see Equation (12)].

$$C_i = \sum_{j=1}^i (\bar{x}_j - \mu_0), \quad (12)$$

with C_i as the cumulative sum until sample i , $\bar{x}_j$ as the average of the i th sample, and μ_0 as the process average target. In the absence of variation in the process, C_i is randomly distributed around zero. In case the process average moves from μ_0 to μ_1 ($\mu_1 \neq \mu_0$), there is an increasing trend if $\mu_1 > \mu_0$, and a decreasing trend if $\mu_1 < \mu_0$. In addition, the *CUSUM* chart is highly recommended for monitoring the process relying on a single observation ($n = 1$).

A second alternative for detecting small variations on the chart mean is the *EWMA*. Similar to the *CUSUM* chart, the *EWMA* chart can handle single observation data ($n = 1$). The z_i statistic for the *EWMA* chart is given as [1]

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1}, \quad (13)$$

with $0 < \lambda \leq 1$ as the weighting factor of observation x_i . The metric z_0 , which is associated to observation x_1 , can be the process target or the average of previous observations (i.e., $z_0 = \bar{x}$). The *EWMA* chart gives more weight to samples recently collected from the process [see Equation (13)]. This enables fast detection of process changes, but turns the chart vulnerable to false alarms.

An important assumption underlying above charts is the consideration that observations are normally distributed. However, observations departing slightly from the normality can also be monitored by these charts due to the central limit theorem.

CONTROL CHART FOR ATTRIBUTES

Several QCs cannot be properly represented by numerical measurements. In this case, each item can be classified as conforming or nonconforming to a specification of interest. This type of QC is named *attribute*. For example, consider a manufacturing process that produces glass jars: a jar that does not present one or more required QCs (e.g., size) is classified to be nonconforming. There are three basic CCs for attributes in literature: (i) CC for fraction of nonconforming (p chart), (ii) CC for number of nonconformities (c chart), and (iii) CC for the number of nonconformities per unit (u chart). Due to space restrictions, we explain the first of these charts.

The fraction, or percentage, of nonconforming arises as a simple and efficient measurement unit in process control. Although beyond the scope of this article, it is easy to prove that the statistical principles underlying the fraction of nonconforming CCs are based on the binomial distribution [1]. Consider p as the probability of having a nonconforming unit; if p is known, one can use

the binomial distribution to construct the p CC. This is not the usual situation, so $\hat{p}$ (i.e., the estimator of the nonconforming fraction) needs to be estimated from the population as in Equation (14).

$$\hat{p}_i = \frac{D_i}{n} \quad i = 1, 2, \dots, k, \quad (14)$$

with D_i as the number of nonconformities in sample i , and n as the number of items in that sample. It is usual to collect k samples. The average of the nonconforming fraction comes as

$$\bar{p} = \frac{\sum_{i=1}^k \hat{p}_i}{k}. \quad (15)$$

The control limits for the p chart are given by

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad (16)$$

$$CL = \bar{p} \quad (17)$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}. \quad (18)$$

Phase 2 of the process control procedure is performed as previously done in CCs for variables: collect further samples of size n , estimate the statistic $\hat{p}$, and plot $\hat{p}$ on the CC. As long as $\hat{p}$ remains within control limits and randomly distributed along the CL, the process is considered to be in-control. Points that exceed the boundaries should be investigated regarding assignable causes [1,2].

OPTIMIZING PROCESS SAMPLING: ADAPTIVE CHARTS

Several modifications in the classic Shewhart CCs have been proposed in the literature. They aim at providing better tools for monitoring processes; one such modification is the adaptive CC. Charts are said to be adaptive when one or more of their parameters (sample size, sampling frequency, and control limit coefficient) are allowed to vary during process monitoring.

Adaptive CCs work as follows. Consider a chart with sample points plotted in the region between warning limits (usually set at $\mu \pm 2\sigma$) and control limits given by $\mu \pm L\sigma$, where L (> 2) is user defined and usually equal to 3. Let that region be called the *warning region*. Since points are in that region, it is possible that the process has shifted to an out-of-control state. To verify that the next sample should be taken sooner or be larger; alternatively a smaller value of L could also be adopted. On the other hand, if sample points are plotted near the center line, chart parameters may be relaxed. When the process is in-control, such policy should lead to a reduction in the mean time to signalize a special cause, and in the average inspection cost.

In theory, the area between control limits could be divided into multiple warning regions, each with different chart parameter settings. Thus, every time a sample point falls into one of these regions, one of several combinations of values for n (sample size), h (sample interval), and L (control limit coefficient) would be chosen according to the most recent information. However, works by Reynolds *et al.* [5], Reynolds and Arnold [6], Runger and Pignatiello [7], and Reynolds [8] have demonstrated that using only two values for n , h , and L leads to better results in terms of minimizing the mean time to signal a special cause.

To compare the performance of alternative adaptive CCs, Reynolds *et al.* [5] proposed two indices: (i) ANSS—average number of samples to signal, defined as the expected number of samples taken from the process until an out-of-control state is signalized; and (ii) ATS—average time to signal, corresponding to the expected time until the chart signalizes an out-of-control state. Under fixed control limits, the probability q of a sample point beyond control limits is independent of the sampling interval, be it fixed or variable. Thus, if no changes take place in the process, the random variable N representing the number of samples to signal is described by a geometric distribution with parameter q . The ANSS is then defined as $ANSS = E(N) = 1/q$. If the process state remains unchanged, the sampling intervals

$h_1, h_2, \dots$ are independent and identically distributed and the ATS is given by $ATS = E(N)E(h_i) = ANSS \times E(h_i)$. In addition to ATS and ANSS, other performance indices have been proposed to the adaptive charts; for example, ANOS—average number of observations to signal [9,10] to analyze CCs where n is variable; and ANSW—average number of switches [11] to analyze charts with adaptive sampling interval.

In addition to the statistical indices described above, determination of the best set of values for the chart parameters may also be based on economic issues. The following economic performance indices have been adopted in works on adaptive charts by authors such as Das *et al.* [12] and Prabhu *et al.* [13]: ETC—expected total cost, giving the process control costs for a production of known and finite size, and ECTU—expected cost per time unit. ECTU is suitable for processes that operate continuously; in case of small lot productions, both indices lead to similar results.

We now give details on the $\bar{X}$ CC with adaptive sampling interval. Our presentation is mainly based on Reynolds *et al.* [5], Runger and Montgomery [14], and Reynolds [15]. Other adaptive charting schemes are reviewed in Tsung and Wang [16].

The adaptive $\bar{X}$ chart works as follows. Assume a QC following a normal distribution with mean μ and known variance σ^2 , where μ_0 is the target mean value. The CC center line is given by μ_0 and the control limits are $\mu_0 \pm L\sigma/\sqrt{n}$. The area between control limits is divided in r regions $I_1, I_2, \dots, I_r$, each

corresponding to a sampling interval h . If the mean $\bar{x}_i$ of sample i is plotted in region I_j , then the next sample $i + 1$ is taken from the process after an interval h_j . The shortest sampling interval $h_{\min}$ is determined considering the minimum time needed to take a sample from the process; the longest interval $h_{\max}$ corresponds to the maximum reasonable time interval for which the process runs unobserved. Although possible, the use of r different intervals in cases where the process mean is $\mu \neq \mu_0$ leads to a larger ATS value if compared to using only two sample intervals ($h_{\min}$ and $h_{\max}$), as demonstrated by Reynolds *et al.* [5], among others. These authors propose the following optimal sampling policy:

$$\begin{aligned} h_{\min} \text{ when } & \begin{cases} \mu_0 - L\frac{\sigma}{\sqrt{n}} \leq \bar{x} < \mu_0 - w\frac{\sigma}{\sqrt{n}} \\ \text{or} \\ \mu_0 + w\frac{\sigma}{\sqrt{n}} < \bar{x} \leq \mu_0 + L\frac{\sigma}{\sqrt{n}} \end{cases} \\ h_{\max} \text{ when } & \mu_0 - w\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + w\frac{\sigma}{\sqrt{n}}, \end{aligned} \quad (19)$$

where w is the parameter that divides the area between control limits in a central region, and two warning regions close to the control limits, as depicted in Fig. 2.

The value of w may be determined to allow comparison between the performance of the adaptive and the fixed sampling interval CCs. For this, when $\mu = \mu_0$, both charts must present the same ANSS, ATS, and sampling rate. The same ANSS value is obtained using the same L when setting the control limits for the charts. The ATS value depends on the

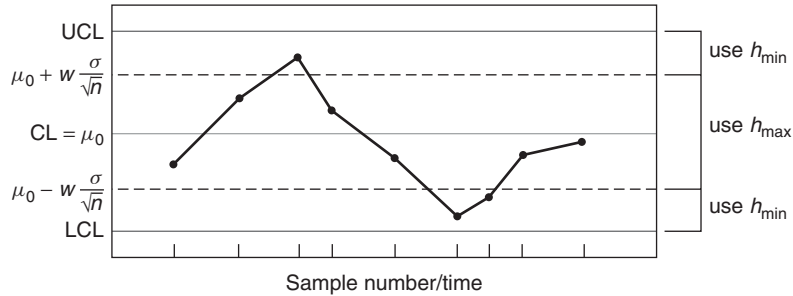


Figure 2. Example of a control chart with adaptive sampling interval.

sampling interval and w . If the interval in the fixed chart is h_f , the adaptive intervals must satisfy the condition $E(h_i) = h_f$. If two sampling intervals h_s (short) and h_l (long) are used and the sample mean is normally distributed, then w is given by

$$w = \Phi^{-1} \left[\frac{2\Phi(L)(h_f - h_s) + h_l - h_f}{2(h_l - h_s)} \right], \quad (20)$$

where $\Phi(\cdot)$ is the standard normal distribution function and $h_s < h_f < h_l$. If $h_f = 1$ (i.e., one time interval), then $E(h_i) = 1$, and the expression (20) is rewritten as

$$w = \Phi^{-1} \left[\frac{2\Phi(L)(1 - h_s) + h_l - 1}{2(h_l - h_s)} \right]. \quad (21)$$

Pairing the two CCs as proposed above allows comparison of ATS values for different values of $\mu_1 \neq \mu_0$, searching for the chart that best detects the shift in μ . It is convenient to express shifts from μ_0 to μ_1 in terms of standard deviations, that is, consider $\mu_1 = \mu_0 + k\sigma$ and then calculate the values of ATS as a function of k .

PROCESS CAPABILITY: DEFINITION AND INDICES

Process capability may be broadly defined as the ability to meet customer demands regarding process outcomes. The basic idea is to compare the admissible process variation range, represented by the specification limits, with the actual process variation. Specification limits may be unilateral or bilateral, leading to different capability assessments in each case. In the case of bilateral limits, a process is deemed as capable whenever 99.73% of its outcomes are within specification limits; the value was chosen to correspond to an area under the normal curve bounded by $\mu \pm 3\sigma$. For unilateral limits, the threshold value is 99.865%. Process capability should only be assessed in processes that are in-control.

The process capability indices (PCIs) C_p , C_{pk} , C_{pm} , and C_{pmk} were originally conceived to be used in normally distributed data. All their statistical figures, such as point estimates and confidence intervals, rely on that

assumption. Although important, the normality assumption is not mandatory for process capability assessment, since PCIs have also been proposed for nonnormal data; see Montgomery [1].

PCIs may be divided into two categories according to the loss function in which they are founded: (i) those with behavior analogous to “goal posts” in a football game and (ii) those based on Taguchi’s loss function. Indices in (i) are C_p and C_{pk} ; they associate no loss to a QC with outputs within specification limits. Indices in (ii) are C_{pm} and C_{pmk} ; they penalize outputs proportionally to their deviation from the specification target value. In what follows, we detail indices in class (i).

In the calculation of C_p and C_{pk} variation within subgroups (σ_{ST}) is considered, the process generating the data is assumed to be in-control, and data points are independent and normally distributed. An estimate of σ_{ST} was needed to set the control limits of the CCs used to monitor process behavior; the above may be used in this calculation. In processes monitored through $\bar{X}$, and R charts, the estimate is $\hat{\sigma}_{ST} = \hat{\sigma}_{\bar{R}} = \frac{\sum_{i=1}^k R_i/k}{d_2}$, where k is the number of subgroups considered and d_2 is a tabled value, available in Duncan [4]. The expression for C_p is

$$C_p = \frac{USL - LSL}{6\hat{\sigma}_{ST}},$$

where USL and LSL are the upper and lower specification limits, respectively. C_p values in the interval [1.00, 1.33] denote processes with reasonable capacity; when $C_p \geq 2.0$ process capacity is deemed as excellent. Kotz and Johnson [17] recommend $C_p \geq 1.67$ in the case of new processes and $C_p \geq 1.50$ for existing processes. For processes with unilateral specifications, C_p is modified as follows:

$$C_{PU} = \frac{USL - \mu}{3\hat{\sigma}_{ST}} \quad \text{or} \quad C_{PL} = \frac{\mu - LSL}{3\hat{\sigma}_{ST}}.$$

Whenever the process target T (represented by μ) differs from the specification interval mean point M , C_p overestimates the process capability and C_{pk} should be used:

$$C_{pk} = \text{Min}(C_{PL}, C_{PU}).$$

While C_p takes only the process standard deviation into account, C_{pk} is a function of both the process average and standard deviation. The values of C_p and C_{pk} coincide whenever $T = M$. For more details on process capability indices, see Montgomery [1] and Kotz and Johnson [17].

REFERENCES

1. Montgomery D. Introduction to statistical process control. 6th ed. New York: John Wiley & Sons Inc; 2008.
2. Navidi W. Statistics for engineers and scientists. New York: McGraw Hill; 2006.
3. Devor R, Chang T, Sutherland J. Statistical quality design and control. Macmillan: New York; 1992.
4. Duncan A. Quality control and industrial statistics. Chicago: Richard D. Irwin, Inc; 1986.
5. Reynolds M, Amin R, Arnold C, *et al.* $\bar{X}$ charts with variable sampling intervals. *Technometrics* 1988;30:181–192.
6. Reynolds M, Arnold J. Optimal one-sided Shewhart control charts with variable sampling intervals. *Seq Anal* 1989;8:51–77.
7. Runger G, Pignatiello J, Joseph J. Adaptive sampling for process control. *J Qual Technol* 1991;23:135–155.
8. Reynolds M. Evaluating properties of variable sampling interval control charts. *Seq Anal* 1995;14:59–59.
9. Costa A. $\bar{X}$ charts with variable sample size. *J Qual Technol* 1994;26:155–163.
10. Park C, Reynolds M. Economic design of a variable sample size $\bar{X}$ chart. *Commun Stat—Simul Comput* 1994;23:467–483.
11. Amin R, Letsinger W. Improved switching rules in control procedures using variable sampling intervals. *Commun Stat Simulat* 1991;20:205–230.
12. Das T, Jain V, Gosavi A. Economic design of dual sampling interval policies for $\bar{X}$ charts with and without run rules. *IIE Trans* 1997;29:497–506.
13. Prabhu S, Montgomery D, Runger G. Economic-statistical design of an adaptive $\bar{X}$ chart. *Int J Prod Econ* 1997;49:1–15.
14. Runger G, Montgomery D. Adaptive sampling enhancements for Shewhart control charts. *IIE Trans* 1993;25:41–51.
15. Reynolds M. Variable-sampling-interval control charts with sampling at fixed times. *IIE Trans* 1996;28:497–510.
16. Tsung F, Wang K. Adaptive charting techniques: literature review and extensions. In: Lenz H-J, Wilrich P-T, Schmid W, editors. *Frontiers in statistical quality control 9*. Amsterdam: Physica-Verlag; 2010. pp. 19–35.
17. Kotz S, Johnson N. Process capability indices. London: Chapman and Hall; 1993.

STOCHASTIC DYNAMIC PROGRAMMING MODELS AND APPLICATIONS

MEHMET A. BEGEN
Ivey School of Business,
University of Western
Ontario, London,
Ontario Canada

INTRODUCTION

Stochastic dynamic programming (DP) models are similar to deterministic DP models [1,2] in terms of states, stages, actions/decisions, and rewards (costs or payoffs) but differ in their state transitions.

In formulating DPs, determining the state correctly can be difficult but crucial for the solution. We can assume the state to be a description of the system that gives the information required to make conscious decisions. Actions are the decision alternatives that the decision maker can choose when the system is in a particular state and influence the future states. Stages are the epochs that a decision is made and there is a change in system's state. Rewards are the payoffs or the costs that the system incurs at each stage with a choice of action for a given state. State transition specifies (the distribution of) the next stage's state given a pair of action and state of the current stage. The functional equation is a link between the current stage's problem and the next stage's problem. Its solution gives an optimal action for each possible state at that stage.

In deterministic DP models, a specification of a state and action pair at the current stage determines the next stage's state with certainty, that is, there is no ambiguity in state transitions. Figure 1 is an example for a state transition of a deterministic DP model.

In stochastic DP models, the state in the next stage is not known deterministically in advance and can only be given probabilistically, that is, given a state and an action

pair, there is some probability that the system will be in some state in the next stage. One can represent this, for each action, in a matrix where rows and columns represent states and the entries of the matrix are probabilities. This matrix is called the *state transition probability matrix*. We can think of deterministic DP models as special cases of stochastic DP ones where the state transition probability matrix (of a state and action pair) consists of only 1 s and 0 s. This is also reflected in the functional equation of the stochastic DP. A generic functional equation for deterministic DP (of a maximization problem) can be given as $f_n(s) = \max_a \{r(s, a) + f_{n+1}(s')\}$ where s is the current state, a is the action taken while in state s , $r(s, a)$ is the expected reward obtained with the state and action pair (s, a) , s' is the state in stage $n + 1$, and finally $f_n(s)$ is the maximum expected reward that can be earned when the state is s in stage n . State s' is determined precisely by state s and action a . In stochastic DP models, we can revise this equation by redefining $r(s, a) = \sum_j \text{prob}(j|s, a) r(j|s, a)$ and $f_{n+1}(s') = \sum_j \text{prob}(j|s, a) f_{n+1}(j)$, where $\text{prob}(j|s, a)$ is the probability that the system will be in state j in the next stage given that the system is in state s and action a is chosen in the current stage, and $r(j|s, a)$ is the reward obtained in the next stage if the state is s and action a is taken in the current stage. Therefore, $r(s, a)$ and $f_{n+1}(s')$ now represent expected values. Furthermore, specifying an action a in state s does not suffice to determine the state in the next stage; the system can be in state j with probability $\text{prob}(j|s, a)$. Figure 2 shows the evolution of states in a stochastic DP model.

Rewards and state transition probabilities on stage n may be different than other stages and therefore one may need to use stage-dependent $r_n(s, a)$ and $\text{prob}_n(j|s, a)$. Throughout the article, we assume stationary rewards and state transition probabilities. Furthermore, we focus on finite-horizon models, that is, $1 \leq n \leq N < \infty$. In finite horizon problems,

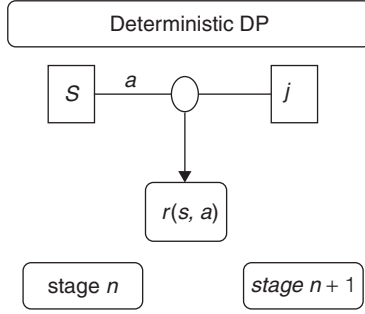


Figure 1. State transition of a deterministic DP model.

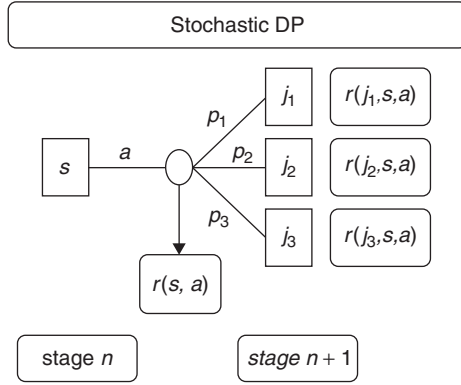


Figure 2. State transition of a stochastic DP model.

one may solve both the deterministic and stochastic DPs with backward induction [1] by solving $f_N(s) = \max_a r(s, a)$ and then solving $f_n(s) = \max_a \{r(s, a) + f_{n+1}(s')\}$ for $n = N-1, N-2, \dots, 1$. At each stage, for each possible state, an optimal action is determined. (Backward induction is the only way of solution for finite-horizon stochastic dynamic programs.) By comparing the functional equations for deterministic and stochastic DPs, we see that the number of (arithmetic) operations for a stochastic problem is more than that for a deterministic one of a similar size.

Deterministic DP models have no standard formulation and are considered to be an art [1–4]. Stochastic DP models are not an exception and have more variety of formulations. DP models are best understood by examples. Before discussing a selection

of stochastic DP applications, we start with an application of deterministic DP resource allocation model and consider its stochastic version.

MODELS AND APPLICATIONS

Resource Allocation

Resource allocation problems can be seen in a variety of contexts. We start with a deterministic example and introduce stochastic aspects to the problem later. There are a number of alternatives (e.g., investments, projects) and a certain amount of resource (e.g., money, budget, labor) that needs to be allocated to each alternative to achieve a certain objective (e.g., maximize total payoff): to make things concrete, we consider N investments with a total available budget of S dollars. Let $r_n(a)$ be the return of investment n if amount a is allocated to it and $f_n(s)$ be the maximum expected reward that is earned with investments $n, n+1, \dots, N$ when there are s dollars available for the investments $n, n+1, \dots, N$.

We first consider the allocation problem where the investment decisions have to be made simultaneously and investments will commence at the same time. The functional equation for this problem can be written as $f_n(s) = \max_{a \leq s \leq S} \{r_n(a) + f_{n+1}(s-a)\}$ for $n = 1, 2, \dots, N-1$ and $f_N(s) = \max_{a \leq s \leq S} r_N(a)$ in which the state in the next stage is $(s-a)$ (amount left available for investments $n+1, n+2, \dots, N$) when the current state is s (s dollars are the available budget for investments $n, n+1, \dots, N$) and the decision maker takes action a (amount of money allocated to investment n). The functional equations for investment n state that an allocation is made such that the current investment return, $r_n(a)$, plus the best possible return of remaining investments, $f_{n+1}(s-a)$, is maximized. If there is only a single investment alternative left, that is, $n = N$, then the allocation is just chosen to maximize the return of investment N . One can solve this problem by backward induction.

Now suppose the return of investment n can be either $K_n(a)$ with probability p_n or $k_n(a)$ with probability $q_n = 1 - p_n$. We

consider two different extensions. In the first extension, the allocation decisions are to be made simultaneously and investments start at the same time (as above in the deterministic case). In the second extension, the decisions are to be made sequentially, that is, an investment is realized and its return is acquired before we invest the next one.

The first case is almost identical to the deterministic case. We only need to redefine $r_n(a)$ as $p_n K_n(a) + q_n k_n(a)$, that is, reward is now the expected return of investment n . Note that we can still determine the next stage's state $(s - a)$ with certainty for a given current state and action pair (s, a) . Thus, we can classify this as a deterministic DP. On the other hand, in the second extension, the state in the next stage can only be given probabilistically for a given action and state pair (s, a) ; it can be either $s - a + K_n(a)$ with probability p_n or $s - a + k_n(a)$ with probability q_n . Therefore, the second case is an example of a stochastic DP and its functional equation becomes $f_n(s) = \max_{a \leq s} \{p_n K_n(a) + q_n k_n(a) + p_n f_{n+1}(s - a + K_n(a)) + q_n f_{n+1}(s - a + k_n(a))\}$ for $n = 1, 2, \dots, N - 1$ and $f_N(s) = \max\{s, \max_{a \leq s} K_N(a)p_N + k_N(a)q_N\}$. Again we can solve these two extensions with backward induction.

We now discuss several more applications of stochastic DP models.

Secretary Selection (Accepting the Best Offer)

Suppose we are hiring for a secretary position. There are N potential candidates and we interview each candidate one at a time. We assume that N is fixed and known to the decision maker, and that there exists a ranking among the candidates. After each interview, we can either choose to hire the person or continue by interviewing the next candidate in line. If we do not hire the current person we interview, they will go and find work elsewhere. Thus, they can be no longer hired. Ideally, we would like to hire the best candidate out of N people. However, we do not know the rank of the candidates, and we can only rank candidates that we have seen so far, for example, we can decide whether the current candidate is better than

any of the previous candidates that we have interviewed so far. Note that if the current candidate is not the best so far then the probability that this person is the best among all N candidates is zero, and if there are any people left to interview then there is still a possibility that the best candidate is yet to come. On the other hand, if the current interviewee is the best candidate seen so far, then we may choose to hire that person. We seek a decision rule that will maximize the probability of hiring the best candidate.

We start with determining the probability that the current candidate n is the best of N candidates if this person n is the best so far. We can think of this as conditional probability:

$$\begin{aligned} &\text{Prob}\{\text{the current candidate is the best of } N | \\ &\quad \text{the current candidate is best of } n\} \\ &= 1/N \div 1/n = n/N. \end{aligned}$$

To make a decision, all we need to know is whether the current candidate is the best so far or not. Therefore, we define the states as $\{B, NB\}$. (B means that the current candidate is the best seen so far and NB denotes the candidate is not the best.) We have only two decisions; hire or do not hire. If we choose to hire the current candidate n , the reward is the probability that person is the best of N , that is, n/N . On the other hand, if we do not hire the current candidate n , then we do not receive any immediate reward and interview the candidate $n + 1$. In this case, the probability that the next candidate will be the best seen so far is $1/(n + 1)$. Now we can write the functional equation for this problem. Let $f_n(s)$ be the maximal probability of selecting the best candidate after interviewing the candidate n and when the state is s (B or NB). Then $f_N(B) = 1$ and $f_N(NB) = 0$ and for $n = 1, \dots, N - 1$

$$\begin{aligned} f_n(NB) &= n/(n+1)f_{n+1}(NB) + 1/(n+1)f_{n+1}(B) \\ f_n(B) &= \max\{f_n(NB), n/N\}. \end{aligned}$$

In $f_n(NB)$, we do not have a choice but to interview the next candidate $(n + 1)$ since the current candidate n is not the best one

so far (and therefore not the best of N). In this case, with probability $n/n + 1$, the candidate $(n + 1)$ will not be the best (among the first $n + 1$ candidates) and with probability $1/n + 1$, the candidate $n + 1$ will be the best (among the first $n + 1$ candidates). On the other hand, in $f_n(B)$, we do have a choice since the current candidate is the best seen so far (and therefore there is a chance of n/N that the candidate is the best of N); we hire if the $n/N \geq n/(n + 1)f_{n+1}(NB) + 1/(n + 1)f_{n+1}(B)$, and do not hire otherwise.

The optimal policy is of the form—interview the first m candidates, and then hire the first candidate that is the best seen so far [3]. We can see this from the functional equation of the model: n/N is increasing in n and $f_n(NB)$ is decreasing in n , therefore there will be a value m after which n/N will be larger than $f_n(NB)$ and we will hire the first best candidate after seeing m candidates. (To see why $f_n(NB)$ is decreasing consider the following: the probability of selecting the best candidate when we rejected the first n candidates is at least as good as probability of selecting the best candidate when we rejected the first $n + 1$ candidates, therefore $f_n(NB)$ is decreasing in n .) As the number of candidates N increases, both the optimal proportion of candidates to interview before hiring m/N and probability of selecting the best candidate $f_1(\cdot)$ is reaching $1/e$ (or 36.8%). Hence, for large N , the optimal thing to do is to interview the first (approximately) 37% candidates and then hire the first candidate who has been the best so far.

This problem can also be thought of as accepting the best offer where a decision maker is presented with N offers in a sequential order [4]. After seeing an offer, the decision maker has to either accept or reject the offer. If an offer is rejected, then it is lost, and the decision maker knows only the relative rank of those offered. The objective is again to maximize the likelihood of choosing the best offer among N possible ones.

The secretary selection problem (or accepting the best offer) is a type of optimal stopping problem. The following is another example for the optimal stopping problem.

Selling an Asset

Suppose we are a real estate agent trying to sell a house. At the beginning of each week, if the house is still on the market, we must advertise it for another week. Advertising costs C dollars per week. Offers for the house will come in from buyers during the week, and at the end of week we must choose to either sell the house to the highest bidder or keep it on the market. If we do not sell the house, then we need to advertise it again next week. The bids received during the week will expire at the end of week and we receive new bids every week. The probability of receiving a bid b is $\text{prob}(b)$ and we can receive at most $m + 1$ different bids including a zero bid (i.e., receiving no bids in the week). If we sell the house then we will receive $r\%$ of the bid as commission; however, if the house is not sold after N weeks, then we will receive nothing (even though we have paid all the advertisement costs).

If the house is already sold, then there is no decision to make, and if it is still on the market then we need to know the maximum bid to decide whether to accept it (and sell the house) or not to sell it. Therefore, the state is given by the highest bid, and state space is $\{b_0, b_1, b_2, \dots, b_m, \text{sold}\}$. We assume $b_0 = 0$ and $b_i > 0$ for $m + 1 > i > 0$. We only have two actions: *sell* or *do not sell*. If the house is already sold then there is no decision to be made. Otherwise, we can choose *sell* or *do not sell*. If we sell, then we receive a commission of $r\%$, the process ends, and the system state becomes *sold*. If we do not sell (and the house is still in the market), then we pay C dollars (except in the week of N) and the system state will be determined by arrival of bids in the next week, that is, with $\text{prob}(b)$ s. Let $f_n(s)$ be the maximal expected value (of [commission - advertisement costs]) obtained after receiving the highest bid of s in week n . Then for all $s \neq \text{sold}$, we can represent the functional equations as

$$\begin{aligned} f_N(s) &= \max\{sr/100, 0\} \\ f_n(s) &= \max\{sr/100, -C + \sum_m \text{prob}(b_m) \\ &\quad \times f_{n+1}(b_m)\} \text{ for } n = 1, 2, \dots, N - 1. \end{aligned}$$

We can solve these equations by backward induction and determine an optimal action for each state in each week. In this problem, there is a threshold bid $L(n)$ for every week n . If $L(n)$ is larger than the highest bid, then we do not sell the house, and we sell it otherwise. This $L(n)$ number decreases as n gets larger (i.e., as it gets closer to week N) since there will be less number of weeks to sell the house.

Consider the following numerical example: $r = 15\%$, $N = 3$, $C = \$1000$, $b_0 = 0$ and $b_1 = \$150,000$, $b_2 = \$250,000$, $\text{prob}(0) = 0.3$, $\text{prob}(150,000) = 0.4$, $\text{prob}(250,000) = 0.3$. Then, $f_3(0) = 0$, $f_3(150,000) = 22,500$, $f_3(250,000) = \max\{0, 150,000 \times 15/100\}$, $f_3(250,000) = 37,500$, $f_2(0) = 19,250 = \max\{0.3 \times f_3(0) + 0.4 \times f_3(150,000) + 0.3 \times f_3(250,000), 0\}$, $f_2(150,000) = 22,500 = \max\{19,250, 150,000 \times 15/100\}$, $f_2(250,000) = 37,500$, $f_1(0) = 25,025$, $f_1(150,000) = \max\{25,025, 150,000 \times 15/100\}$, $f_1(250,000) = 37,500$. With these calculations and by keeping track of optimal action at this week (i.e., sell or do not sell), we can identify the optimal policy as follows: sell in Week 1 if the highest bid is $\$250,000$ and sell in Weeks 2 and 3 if the highest bid is at least $\$150,000$, that is, $L(1) = \$250,000$ and $L(2) = L(3) = \$150,000$. For more on this problem and its variants see Puterman [3] and Ross [4].

A Betting Game

Consider a game in which a player can bet any nonnegative amount up to her current wealth at each round of the game [4]. After the bet is placed, she will either win the bet amount with probability p or lose this amount with probability $q = 1 - p$. The game will be played N times, that is, she will bet N times. We assume the player is risk averse and her objective is to maximize her expected utility given by the $\log$ of her wealth at the end of the game, that is, after N bets. Then the question is how the bets should be placed.

We first have to determine the state, that is, the information needed to decide a bet amount. Obviously, she needs to know how much wealth she has before she places any bets. Suppose she has a current wealth of s and bet an amount of as ($0 \leq a \leq 1$), that is, a is the ratio of the bet to her current wealth s . Then, with probability p , her wealth level will

be $(s + as)$ and with probability $q = 1 - p$, the wealth level will be $(s - as)$. If there was only a single bet, $N = 1$, then the expected payoff would be $p \log(s + as) + q \log(s - as)$ and she would choose a such that this expected value would be maximized. Now, we extend this idea to $N > 1$. Let $f_n(s)$ be the maximal expected value of the objective if the current wealth is s and there are n more bets left. (The final stage, no more bets, is numbered as 0 in this formulation.) We have already defined the action as the ratio of the bet to the current wealth. Then, we can represent the functional equation for this stochastic DP model as

$$\begin{aligned} f_0(s) &= \log s \\ f_n(s) &= \max_{(0 \leq a \leq 1)} \{pf_{n-1}(s + as) + qf_{n-1}(s - as)\} \\ &\text{for } n = 1, 2, \dots, N. \end{aligned}$$

We solve these equations by using backward induction. We start with $n = 1$ (i.e., there is only one more bet remaining): $f_1(s) = \max_{(0 \leq a \leq 1)} \{p \log(s + as) + q \log(s - as)\}$. If $p \leq q$ (i.e., $p \leq 1/2$), then the optimal action is to bet nothing, that is, $a^* = 0$. (Rewrite the terms as $p \log(s + as) + q \log(s - as) + p \log(s - as) - p \log(s - as)$ and organize them as $p \log(s^2(1 - a^2)) + (q - p) \log(s(1 - a))$, both terms are maximized if $a^* = 0$.) This holds not only for $n = 1$ but also for all n . Suppose $p > q$ (i.e., $p > 1/2$), then we can determine $f_1(s)$ with calculus and obtain $a^* = p - q$. Therefore, $f_1(s) = p \log(s(1 + p - q)) + q \log(s(1 - p + q)) = p \log(2ps) + q \log(2qs) = K + \log s$, where $K = \log 2 + p \log p + q \log q$. Next, we find $f_2(s) = \max_{(0 \leq a \leq 1)} \{p \log(s + as) + q \log(s - as)\} + K$. We note that $f_2(s) = f_1(s) + K$ hence $f_2(s) = 2K + \log s$. Similarly, we see that $f_n(s) = nK + \log s$. Therefore, the optimal strategy for all rounds of the game is to bet a^*s , that is, $a^* = (p - q)s$ when $p > q$.

Hunting Preference of a Lion

We consider the hunting decisions of a lion 3. The objective of the lion is to maximize the probability of its survival over the next 20 days. Each day the lion can choose whether to hunt or not to hunt. We assume (for simplicity) that the lion considers only one type

of prey. Besides hunting on its own, the lion can also choose to hunt in a group of lions up to eight. The hunt success chance increases with the number of lions in the group. The hunt success probability is estimated by $p(k) = 0.04 + 0.05k$, where k ($1 \leq k \leq 8$) is the number of lions in the group. If a hunt is successful, the lions share their prey equally. Furthermore, lions do not make more than one hunt attempt per day.

The lion needs approximately 6 kg of meat a day and has a gut capacity of about 30 kg. Hunting requires an additional 0.5 kg of food energy and this number does not change with the number of lions involved in a hunt. If the energy reserve of the lion reaches zero, it dies. On the other hand, preys have average edible meat of 247 kg, and so can satisfy the food demand of several lions.

We will build a stochastic DP model for this problem. Stages are naturally given as days; we add a day to determine the survival probability at the end of 20 days. The actions are whether to hunt or not, and if to hunt, in what size group. We may represent the actions as $\{0, 1, \dots, 8\}$ where 0 represents no hunt (NH). However, in this problem, we may consider only actions: hunt with a group of 8 or NH. This is because as the group size k increases, the success rate of a hunt increases (given by $p(k)$), the food consumption for a hunt does not change (it is 0.5 kg of food energy for any k , and in fact we would expect the lion to consume less as k increases), and a hunt with a group of eight results in more than $(247/8 > 30)$ 30 kg of meat for each lion. Therefore, with these numbers, lions will always choose to hunt in a group of eight (if it chooses to hunt) and we represent the actions as H (hunt with a group of eight) and NH. Next, we need to think about the state for the problem. What is the information that the lion requires to decide whether to hunt or not? It is the amount of food energy level at the beginning of a day; since the total capacity is 30 kg and a hunt consumes 0.5 kg, we can represent the state space as $\{0, 0.5, 1, 1.5, \dots, 30\}$, where 0 denotes the death state. If the lion has an energy level of s and chooses to hunt, then its energy level the next day will be 30 with probability $p(8)$ and be $\max(0,$

$s - 6.5)$ with probability $1 - p(8)$. If the lion does not hunt, then its energy level the next day will be $\max(0, s - 6)$. Furthermore, $\text{prob}\{0 | 0, \cdot\} = 1$. We are now ready to write down the functional equations for this problem. Let $f_n(s)$ be the probability of survival (at the end of 20 days) if there are n days to go and the lion's energy reserve is s . Then

$$\begin{aligned} f_n(0) &= 0 \text{ for } n = 0, 1, 2, \dots, 20 \\ f_n(s) &= \max\{f_{n-1}(\max\{0, s - 6\}), p(8)f_{n-1}(30) \\ &\quad + (1 - p(8))f_{n-1}(\max\{0, s - 6.5\})\} \\ &\text{for } n = 1, 2, \dots, 20 \text{ and} \\ &\quad s = 0, 0.5, 1, 1.5, \dots, 30. \end{aligned}$$

We can solve this problem by backward induction and determine $f_{20}(s)$, probability of the lion's survival after 20 days if its current energy level is s . By tracking down the optimal action (NH or H) for each day, we can determine the lion's optimal decisions daily over the horizon of 20 days. For instance, if the lion's energy level is 6.5, then its survival probability at the end of 20 days is about 30% and if the energy level is 7 the survival probability jumps to 46%. With an energy level of 6.5 at the beginning of Day 1, the lion chooses not to hunt on Day 1 but hunts on Day 2; and with an energy level of 7 (or more), the lion chooses to hunt on Day 1.

Hockey Goalie or No Goalie

In ice hockey, there are usually (e.g., no power play) five players and one goalie playing for each team. The goalie is the most important defense player. However, if toward the end of the game the team is losing, the coach of the losing team may decide to pull his goalie, that is, replace the goalie with another player. This decision gives the pulling team a 6-to-5 skater advantage over the opposing team and increases the pulling team's chance to score. However, at the same, it increases the likelihood of the opponent to score as well and at a larger factor. However, the pulling team may score and tie the game with this decision. (In the 2010 Winter Olympic Games in Vancouver, we have seen examples of this

strategy.) Then the question is to determine when (if ever) should a team pull the goalie? This problem is studied in literature [5,6] and our example is from Puterman [3].

We build a stochastic DP model for a hockey coach whose team is losing near the end of the game. The coach estimates the following: pulling the goalie will increase the probability that the coach's team scores the only goal in the next t seconds from p_G to p_{NG} ; however, the probability that the opponent team scores the only goal in this t -second time interval rises from q_G to q_{NG} . The coach assumes that the events in successive t -second intervals are independent. The coach's objective is to maximize the probability of having at least a tie.

The most difficult part of this problem is to determine a state variable. One good choice is the score difference between the teams; the score of the coach's team - the score of the opposing team, $\{\dots, -2, -1, 0, 1, 2, \dots\}$. Given such a state definition, it is obvious that, if the score difference were zero or greater, then the coach would not pull the goalie. Stages are every t -second intervals. Suppose there are Nt seconds left in the game (then we can simply assume $N + 1$ stages and work with $n = 0, 1, 2, \dots, N$). The coach has two actions: let the goalie play (G) or pull the goalie (NG). Let s be the current state, then state transitions are as follows: suppose the coach chooses action G , then with probability p_G the coach's team will score in the next t seconds so the new state will be $s + 1$, with probability q_G , the opposing team will therefore score and the new state will be $s - 1$, and with remaining probability $1 - p_G - q_G$, there will be no scoring so it will be s . The same structure holds for the action of NG except the probabilities change. Thus we have: $\text{prob}(s + 1|s, G) = p_G$, $\text{prob}(s|s, G) = 1 - p_G - q_G$, $\text{prob}(s - 1|s, G) = q_G$, and $\text{prob}(s + 1|s, NG) = p_{NG}$, $\text{prob}(s|s, NG) = 1 - p_{NG} - q_{NG}$, $\text{prob}(s - 1|s, NG) = q_{NG}$.

Now we write down the functional equations. Let $f_n(s)$ be the probability of having at least a tie after Nt seconds when the current score difference is s and there

are n more t -seconds to go. Then

$$\begin{aligned} f_0(s) &= 1 \text{ if } s \geq 0 \text{ and } f_0(s) = 0 \text{ if } s < 0 \\ f_n(s) &= \max\{R(G), R(NG)\} \text{ for } n = 1, 2, \dots, N \\ &\text{and } s = \dots, -2, -1, 0, 1, 2, \dots, \\ &\text{where } R(X) = p_X f_{n-1}(s + 1) \\ &\quad + q_X f_{n-1}(s - 1) + (1 - p_X - q_X) f_{n-1}(s) \\ &\text{for } X = R, NG. \end{aligned}$$

This model can once again be solved by backward induction and we can determine an optimal decision for each t -second interval by starting from $f_0(s)$ to $f_N(s)$ for all s .

OTHER APPLICATIONS

In this section, we briefly mention and list a few more applications of stochastic DP from literature. In Goto *et al.* [7], a stochastic DP model is developed to determine the number of (airline) meals to load on an airplane on various time intervals before its departure. The meals have to be prepared in advance; however, the exact number of passengers changes until the last minute. Authors show that the use of the model can improve customer service (by decreasing meal shortages) and save money. Another application [8] is from electrical engineering in which researchers use stochastic DP modeling to optimize streaming of media packetized over a lossy packet network (such as Internet). The trade-off is between the number of transmissions (cost) and probability of a packet being lost or delayed (quality). The model takes the current status of the network to be the state and determines whether a retransmission should be sent or not. There are many other applications of stochastic DP in communication networks [9]. Revenue management [10,11] is another important area where stochastic DP models play an important role. For example, in Bitran and Mondschein [12], authors use stochastic DP to investigate hotel room booking decisions with different customer segments and dynamic arrivals. In Green *et al.* [13], researchers use stochastic DP to study the problem of admitting different class of

patients to a diagnostic facility when emergency patients are present. In a similar study [14], authors also model patient preferences (e.g., choice of a physician and time of day for an appointment) in determining which appointment requests to admit to maximize revenue in a primary-care clinic.

It is worthwhile to mention that stochastic DP is used (recently more often) in health care, both in clinical and operational problems. For example, Patrick *et al.* [15] considers a multipriority patient scheduling model for a diagnostic resource (such as a CT scanner) to achieve priority-level target wait times in minimal cost. In Su and Zenios [16], researchers study the effect of patient choice on a kidney transplant system and the effect of the queueing discipline on organ allocation efficiency to reduce the rate of kidney refusals by patients. In another study [17], a stochastic DP model is developed to determine the number of surgeries booked for a day in the presence of both elective and emergency surgery demand. Other examples include liver transplantation decisions [18], optimal timing to start HIV therapy [19], the influenza vaccine selection [20], and optimal motion plan for the steering of an image-guided medical needle [21].

Besides the applications discussed in this article, there are many other applications of stochastic DP models in health care [22,23], finance [24], scheduling and sequencing [25], production, inventory and vehicle routing [26,27], sports [28], game theory and reinforcement learning [29–31], reliability and maintenance [32], energy [33], water reservoir systems [34], and natural resources (e.g., forestry, fisheries, and agriculture) [35]. See White [36] and Puterman [37] for more applications.

CONCLUDING REMARKS

In developing stochastic DP models, identification of the state variable is crucial. The state description should provide all the relevant information necessary for making a decision. Normally, we do not include any information that does not change from one stage to another in the state description (e.g., data) as well as the number of stages

remaining because this is implicitly present in the specification of rewards and transition probabilities.

In this article, we have focused on finite-horizon problems ($N < \infty$); however, for some applications, it makes more sense to consider infinite horizon. For infinite-horizon DPs, we need solution methods other than backward induction, although the formulations for both finite and infinite horizon models are virtually the same. For more on infinite horizon problems, refer to Puterman [3], Ross [4], and Bertsekas [38].

A major challenge in solving stochastic DPs in practice has been computational; that is, the size of the state space is too large to solve even with modern computing capabilities. This is known as *curse of dimensionality* [1,38,39]. For example, consider an inventory problem where the objective is to determine how much to keep (for a family of products) in stock to satisfy a random demand in the next N months where unsatisfied demand is lost. The objective is to maximize the expected total profit (sales revenue – inventory holding costs – ordering costs – lost sales cost). The expected revenue and cost computations depend on how much of each product type is in the inventory since some products are substitute (or complementary) of others. The decision is how much (if any) to order in each month to replenish the inventory. The state variable for this problem should include the inventory on hand for each product i . Suppose the capacity (of the storage facility per product family) is C , then the state is vector $s_n = (q_1, q_2, \dots, q_m)_n$ where $0 \leq q_i \leq C$ for $i = 1, 2, \dots, m$ and $n = 1, 2, \dots, N$. Suppose $m = 7$ and $C = 10$, then the state space contains 10^7 terms. (In recent decades, a field of research called *approximate dynamic programming* [38] has developed in an attempt to solve larger models.)

One of the advantages of the DP framework is that it is often possible to obtain a structure for an optimal policy. Establishing such a structure for the optimal policy is desirable because it provides managerial and intuitive insight into the problem. Furthermore, with structure, computation of the optimal policy can often be simplified.

REFERENCES

1. Moshe S. Dynamic programming: introductory concepts in EORMS. Wiley; 2010.
2. Smith DK. Dynamic programming, deterministic models in EORMS. Wiley; 2010.
3. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: Wiley; 1994.
4. Ross SM. Introduction to stochastic dynamic programming. New York: Academic Press; 1983.
5. Washburn A. Still more on pulling the goalie. Interfaces 1991;21(2):59–64.
6. Ingolfsson A. Ice hockey in EORMS. Wiley; 2010.
7. Goto JH, Lewis ME, Puterman ML. Coffee, tea, or ...?: a markov decision process model for airline meal provisioning. Trans Sci 2004;38(8):107–118.
8. Chou PA, Miao Z. Rate-distortion optimized streaming of packetized media. IEEE Trans Med 2006;8(2):390–404.
9. Altman E. Applications of Markov decision processes in communication networks. In: Feinberg EA, Shwartz A, editors. Handbook of Markov decision processes, methods and applications. Boston (MA): Kluwer Academic Publishers; 2002. pp. 489–536.
10. MacGill JI, Van Ryzin GJ. Revenue management: research overview and prospects. Trans Sci 1999;33(2):233–256.
11. Weatherford LR, Samuel EB. A taxonomy and research overview of perishable-asset revenue management: yield management, overbooking, and pricing. Oper Res 1992;40(5):831–844.
12. Bitran GR, Mondschein SV. An application of yield management to the hotel industry considering multiple day stays. Oper Res 1995;43(3):427–444.
13. Green L, Savin S, Wang B. Managing patient service in a diagnostic medical facility. Oper Res 2006;54(1):11–25.
14. Gupta D, Wang L. Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 2008;56(3):576–592.
15. Patrick J, Puterman ML, Queyranne M. Dynamic multi-priority patient scheduling. Oper Res 2008;56(6):1507–1527.
16. Su X, Zenios S. Patient choice in kidney allocation: the role of the queueing discipline. MSOM 2004;6(4):280–301.
17. Gerchak Y, Gupta D, Henig M. Reservation planning for elective surgery under uncertain demand for emergency surgery. Manage Sci 1996;42(3):321–334.
18. Sandikci B, Maillart LM, Schaefer AJ, *et al.* Estimating the patient's price of privacy in liver transplantation. Oper Res 2008;56(6):1393–1410.
19. Shechter SM, Bailey MD, Schaefer AJ, *et al.* The optimal time to initiate HIV therapy under ordered health states. Oper Res 2008;56(1):20–33.
20. Wu JT, Wein LM, Perelson AS. Optimization of influenza vaccine selection. Oper Res 2005;53(3):456–476.
21. Ron Alterovitz and Ken Goldberg Motion Planning in Medicine: Optimization and Simulation Algorithms for Image-Guided Procedures, Springer Tracts in Advanced Robotics, Berlin: Springer; 2008.
22. Schaefer AJ, Bailey MD, Shechter SM, *et al.* Modeling medical treatment using markov decision processes. In: Brandeau M, Sainfort F, Pierskalla W, editors. Handbook of operations research/management science applications in health care. Norwell (MA): Kluwer Academic Publishers; 2004. pp. 597–616.
23. Patrick J, Begen MA. Markov decision processes and its applications in healthcare. In: Yin Y, editor. Handbook of healthcare delivery systems. CRC Press. in press
24. Schäl M. Markov decision processes in finance and dynamic options. In: Feinberg EA, Shwartz A, editors. Handbook of Markov decision processes, methods and applications. Boston (MA): Kluwer Academic Publishers; 2002. pp. 461–487.
25. Pinedo M. Scheduling: theory, algorithms and systems. New York (NY): Springer; 2008.
26. Denardo EV. Dynamic programming: models and applications. Englewood Cliffs (NJ): Dover; 2003.
27. Kleywegt AJ, Nori VS, Savelsbergh MVP. Dynamic programming approximations for a stochastic inventory routing problem. Trans Sci 2004;38(1):42–70.
28. Norman JM. Dynamic programming in sport: a survey of applications. IMA J Math Appl Bus Ind 1995;6:171–176.
29. Leyton-Brown K, Shoham Y. Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press; 2009.

30. Sutton RS, Barto AG. Reinforcement learning. Cambridge (MA): MIT Press; 2000.
31. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming. Belmont (MA): Athena Scientific; 1996.
32. Pham H. Handbook of reliability engineering. London: Springer; 2003.
33. Song H, Liu C, Lawarrée J, *et al.* Optimal electricity supply bidding by Markov decision process. IEEE Trans Power Syst 2000;15(2):618–624.
34. Lamond BF, Boukhtouta A. Water reservoir applications of Markov decision processes. In: Feinberg EA, Schwartz A, editors. Handbook of Markov decision processes, methods and applications. Kluwer Academic Publishers; 2002. pp. 537–558.
35. Kennedy JOS. Dynamic programming applications to agriculture and natural resources. New York (NY): Elsevier; 1986.
36. White DJ. A survey of applications of markov decision processes. J Oper Res Soc 1993;44(11):1073–1096.
37. Puterman ML. Dynamic programming. In: Meyers RA, editor. Encyclopedia of physical science and technology. New York: Academic Press; 1987. pp. 438–463.
38. Bertsekas D. Dynamic programming and stochastic control. Belmont (MA): Athena Scientific; 2001.
39. Powell WB. Approximate dynamic programming: solving the curses of dimensionality. Hoboken (NJ): John Wiley & Sons, Inc.; 2007.

STOCHASTIC GRADIENT METHODS FOR SIMULATION OPTIMIZATION

MICHAEL C. FU
University of Maryland,
College Park, Maryland

INTRODUCTION

Stochastic gradient methods refer to techniques for estimating gradients from sample paths or simulation replications. Often, a distinction is made between what are known as direct versus indirect methods. Indirect methods estimate an approximation to the gradient—usually a form of secant based on finite differences, whereas direct methods refer to approaches that estimate the actual gradient—geometrically, the tangent vector. Direct methods generally provide an unbiased estimator for the gradient, whereas indirect methods are biased, with the bias generally vanishing as the difference used in the estimator goes to zero. However, the efficiency of the estimators depends not only on the bias but also the amount of computation required and the precision of the estimate, usually measured in terms of variance. In settings in which they are applicable, direct estimators can be far more efficient than indirect estimators. This article deals only with direct estimators.

Two areas in which gradients play a central role are sensitivity analysis and optimization. For simulation optimization, gradient estimators can be used in algorithms that exploit gradient information, such as those based on stochastic approximation. For example, if we are trying to minimize a function $J(\theta)$, as a function of the parameter (vector) θ , then the stochastic approximation algorithm takes the following general form:

$$\theta_{n+1} = \theta_n - a_n \widehat{\nabla J}(\theta),$$

where θ_n denotes the n th iterate of the parameter, $\{a_n\}$ is a sequence of step-size multipliers, and $\widehat{\nabla J}(\theta)$ denotes an estimate of the gradient $\nabla J(\theta)$. Clearly, the effectiveness of the algorithm rests heavily on the quality of the gradient estimate. In general, the two desirable properties are those of any statistical estimator: low bias and low noise, sometimes combined into a single mean-squared error measure. Specifically, here we seek unbiased estimators

$$E[\widehat{\nabla J}(\theta)] = \nabla J(\theta),$$

with low variance.

This article describes the two most well-known approaches for stochastic gradient estimation:

- perturbation analysis (PA),
- likelihood ratio method (LRM), which is also called the score function method.

Specifically, the focus is on presenting the ideas of infinitesimal perturbation analysis (IPA) and the LR method (LRM); they are illustrated with examples from the main application areas of queueing, inventory, and finance. Other approaches not described here are based on Malliavin calculus, weak derivatives or WD (also called measure-valued differentiation), simultaneous perturbations, and frequency domain experimentation; the latter two approaches are more akin to indirect methods since they involve actual changes in the underlying parameter values. *Neither IPA nor LRM estimators require any alteration in the parameter settings of the simulation model.*

IPA AND LRM

Let $J(\theta)$ denote a *performance measure* that depends on a vector of *parameters* θ , where it is assumed that there is some *sample* performance $L(\theta, \omega)$ such that $J(\theta) = E[L(\theta, \omega)]$, with ω denoting a sample path or simulation replication. In terms of simulation, ω

could represent all of the random number seeds required to generate a single sample path. Common sample performance measures include

- queueing—average time in queue, average number in system, percentage of customers who wait in queue for more than a certain prescribed limit;
- inventory—long-run average holding, ordering and/or backlogging costs; percentage lost sales, percentage demand met from on-hand inventory;
- finance—profit/loss profile or terminal wealth in investment/portfolio management; (discounted) payoff from a derivatives contract.

Parameters of interest can be divided into two main classes:

- distributional—for example, the mean of interarrival and service times in a queueing system, mean of demand in an inventory system, the mean return and volatility of a stock in a finance problem;
- structural—for example, the upper bound target wait time in a queueing system, the reorder point and order quantities in inventory control, the knock-in or knock-out values in barrier options (finance).

The type of parameter is usually clear from the context but in some cases, the distinction is unclear or arbitrary, and depending on the formulation (or reformulation) a parameter could be viewed as either type. The focus of the exposition here will be on distributional parameters.

Although most performance measures of interest are expectations, including probabilities as special cases by using indicator functions, quantiles

$$q_\alpha = \min_x \{P(L \leq x) \geq \alpha\}, \quad 0 \leq \alpha \leq 1,$$

cannot be captured by expectations. Quantiles will not be treated here, but the reader is referred to the final section titled “Probing

Further” for references and applications of PA to this setting.

By the definition of expectation, the calculation of $E[L]$ requires knowing the distribution of L . However, the *raison d’être* for using stochastic simulation is that this distribution is not easy to get a handle on. Basically, the distribution is available *implicitly*, in that the simulation model generates samples from the distribution of the output performance measure using samples from the input distributions/processes. In the dynamic setting, these are called sample paths, or in simulation terminology, simulation replications or sometimes, realizations. For example, in queueing, the inputs are generally input and service time processes, for which the simulation model generates sample paths of the number in queue and waiting time processes.

To be more precise, the expectation is written in two ways, first using the formal definition of expectation and then using the elementary result for the expectation of a function of random variables:

$$E[L(X)] = \int y \, dF_L(y) = \int \dots \int L(x) \, dF_X(x), \quad (1)$$

where X is a *vector* of all the input random variables used in the simulation, F_L is the distribution of the univariate output sample performance measure L , F_X is the (joint) distribution of X , and the second integral has dimension corresponding to the number of input random variables (dimension of the vector X). As mentioned already, F_L is not known explicitly; else, simulation would not be needed. However, in simulation, F_X is usually known, because it is used to generate the input process to the simulation model; for example, interarrival and service times in a queueing system. In probability theory, this is a basic result of computing the expectation of a function of random variables, called the “law of the unconscious statistician” in the well-known textbook, “Stochastic Processes,” by Sheldon Ross ([1], although the terminology seems to have been removed in the second edition; see also Ross [2]), evidently reflecting that many people use the result as if it were the definition of expectation rather than

an elementary result. In any case, stochastic simulation can be viewed as carrying out the “law of the unconscious statistician” using a simulation model to generate output samples $L(X)$ from input samples X .

Note the parameter θ has not been placed anywhere in expression (1). In fact, where it is placed is what “differentiates” IPA and LRM. To simplify the exposition, henceforth θ will be assumed to be a scalar since in the vector case, the gradient can be obtained by finding each of the partial derivatives individually, so there is no loss of generality in making this assumption. Thus, the problem is to estimate

$$\frac{dE[L]}{d\theta} \quad (2)$$

using simulation.

A “brute-force” finite difference (indirect) estimate is obtained by doing additional simulations at parameter value $(\theta + \Delta\theta)$, subtracting from these samples of L the simulation estimates of L at the original parameter value θ , and then dividing by the perturbation $\Delta\theta$. However, if the number of parameters is large (e.g., large number of decision variables in an optimization problem), then this could entail a large number of additional simulations. Moreover, selecting the value for $\Delta\theta$ entails a trade-off between variance and bias, where larger $\Delta\theta$ generally means lower variance but higher bias when the goal is to estimate the value of Equation (2). The order of the bias can be decreased by using symmetric differences, that is, simulate at both $(\theta + \Delta\theta)$ and $(\theta - \Delta\theta)$, but the downside is that this would require basically doubling the number of simulations.

Research starting from the 1980s has resulted in developing accurate and efficient *direct* ways to estimate derivatives from simulation. Returning to the placement of θ , consider the following two cases:

$$E[L(X)] = \int_0^1 \dots \int_0^1 L(X(\theta; u)) du, \quad (3)$$

$$E[L(X)] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x) dF_X(x; \theta). \quad (4)$$

The first case—corresponding to IPA—views the underlying randomness as a vector of random numbers generating the input random variables X depending on the parameter θ , whereas the second case—corresponding to LRM—takes θ as a parameter in the (joint) distribution F_X of all of the input random variables. Another version of the first case (IPA) not considered explicitly here is

$$E[L(X)] = \int_0^1 \dots \int_0^1 L(X(u); \theta) du,$$

where the parameter does not enter into the input random variables but directly into the sample performance. This is the structural parameter case mentioned earlier, with an example being the upper bound target wait time in a queueing system.

To illustrate the ideas more concretely, consider a specific example involving just two input random variables, assumed to be generated independently, with $X_1 \sim \exp(\theta)$ and $X_2 \sim U(0, 1)$; that is, the first random variable is exponentially distributed with mean θ and the second is uniformly distributed over $[0, 1]$. Then Equations (3) and (4) are written respectively as

$$E[L(X)] = \int_0^1 \int_0^1 L(X_1(u_1; \theta), u_2) du_1 du_2, \quad (5)$$

where $X_1(u; \theta) = -\theta \ln u$ is the common method of generating samples from the distribution $\exp(\theta)$, and

$$E[L(X)] = \int_0^1 \int_0^{\infty} L(x_1, x_2) \frac{1}{\theta} e^{-x_1/\theta} dx_1 dx_2. \quad (6)$$

Differentiating Equations (5) and (6), assuming that the differentiation operator can be brought inside the integrals (conditions under which this is permitted are discussed in the section titled “Unbiasedness: Validity of the Interchange”):

$$\begin{aligned} \frac{dE[L(X)]}{d\theta} &= \int_0^1 \int_0^1 \frac{\partial L(X_1(u_1; \theta), u_2)}{\partial X_1} \\ &\quad \times \frac{dX_1(\theta; u_1)}{d\theta} du_1 du_2, \end{aligned} \quad (7)$$

and for the second case,

$$\begin{aligned} \frac{dE[L(X)]}{d\theta} &= \int_0^1 \int_0^\infty L(x_1, x_2) \left[\frac{1}{\theta} \left(\frac{x_1}{\theta} - 1 \right) \right] \\ &\quad \times \frac{1}{\theta} e^{-x_1/\theta} dx_1 dx_2, \end{aligned} \quad (8)$$

with respective IPA and LRM estimators corresponding to Equations (7) and (8):

$$\frac{\partial L(X_1, X_2)}{\partial X_1} \frac{dX_1}{d\theta}, \quad L(X_1, X_2) \left[\frac{1}{\theta} \left(\frac{X_1}{\theta} - 1 \right) \right].$$

To finalize the first estimator requires knowing something about the *specific form* of L (to calculate the first factor) and defining the *derivative of a random variable* with respect to a parameter (to calculate the second factor); intuitively, though, differentiating $X_1(u; \theta) = -\theta \ln u$ gives $\frac{dX_1(u; \theta)}{d\theta} = -\ln u = \frac{X_1(u; \theta)}{\theta}$. The second estimator can be implemented *without any further knowledge or simulation* over what is required to estimate the original performance measure by $L(X_1, X_2)$; that is, the original performance measure estimator is simply multiplied by a “weight” function. This simple example contains the main ingredients of the general case.

Returning to the more general settings of Equations (3) and (4), again differentiating each, assuming an interchange of integration and differentiation is permissible,

$$\frac{dE[L(X)]}{d\theta} = \int_0^1 \dots \int_0^1 \frac{dL(X(\theta; u))}{d\theta} du, \quad (9)$$

and

$$\frac{dE[L(X)]}{d\theta} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x) \frac{\partial f(x; \theta)}{\partial \theta} dx, \quad (10)$$

where f is the (joint) density corresponding to F_X . For expositional ease in introducing the approaches, begin by assuming that the parameter only appears in X_1 , which is generated *independently* of the other input random variables. So for the case of Equation (9), use

the chain rule to write

$$\begin{aligned} \frac{dE[L(X)]}{d\theta} &= \int_0^1 \dots \int_0^1 \frac{dL(X_1(\theta; u_1), X_2, \dots)}{d\theta} du \\ &= \int_0^1 \dots \int_0^1 \frac{\partial L}{\partial X_1} \frac{dX_1(\theta; u_1)}{d\theta} du. \end{aligned} \quad (11)$$

In other words, the IPA estimator takes the same general form as before:

$$\frac{\partial L(X)}{\partial X_1} \frac{dX_1}{d\theta}, \quad (12)$$

where the parameter appears in the transformation from random number to random variate, and the derivative is expressed as the product of a sample path derivative and derivative of a random variable.

Assume that X_1 has marginal p.d.f. $f_1(\cdot; \theta)$ and that the joint p.d.f. for the remaining input random variables $(X_2, \dots)$ is given by f_{-1} , which has no (functional) dependence on θ . Then the assumed independence gives $f = f_1 f_{-1}$, and the expression (10) involving differentiation of a density (measure) can be further manipulated using the product rule of differentiation to yield the following:

$$\begin{aligned} \frac{dE[L(X)]}{d\theta} &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x) \frac{\partial f_1(x_1; \theta)}{\partial \theta} f_{-1}(x_2, \dots) dx \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x) \frac{\partial \ln f_1(x_1; \theta)}{\partial \theta} f_1(x_1; \theta) \\ &\quad \times f_{-1}(x_2, \dots) dx \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x) \frac{\partial \ln f_1(x_1; \theta)}{\partial \theta} f(x) dx. \end{aligned}$$

In other words, the LRM estimator takes the form

$$L(X) \frac{\partial \ln f_1(X_1; \theta)}{\partial \theta}, \quad (13)$$

where the second factor in Equation (13) is called the score function in statistics.

The IPA estimator (12) requires the notion of a derivative of a random variable, which intuitively coincides with the definition of a derivative in the usual sense by defining the

limiting difference on a common probability space and fixing the sample outcome. Several useful cases are now presented, where here X represents a single random variable.

- Assuming the existence of a density:

$$\frac{dX}{d\theta} = -\frac{\partial F(x; \theta)/\partial \theta}{\partial F(x; \theta)/\partial x} \Big|_{x=X}, \quad (14)$$

where F is the cumulative distribution function of X , so the denominator is simply the probability density function [3], [4].

- For a *location*, *scale*, or *generalized scale* parameter, defined here in terms of the random variable rather than the distribution:
- θ is a *scale* parameter for X if there exists a random variable Z that has no dependence on θ such that $X = Z\theta$ with probability 1. Then,

$$\frac{dX}{d\theta} = \frac{X}{\theta}.$$

An example is the exponential distribution, for which the mean is a scale parameter, verifying the result $\frac{dX}{d\theta} = \frac{X}{\theta}$ that was intuitively derived earlier through direct differentiation of the expression $X(u, \theta) = -\theta \ln u$, and which could also be derived using Equation (14). Another scale parameter example is $U(0, \theta)$.

- θ is a *location* parameter for X if there exists a random variable Z that has no dependence on θ such that $X = Z + \theta$ with probability 1. Then,

$$\frac{dX}{d\theta} = 1.$$

An example is the mean of a uniform distribution, that is $U(\theta - \delta, \theta + \delta)$. Note that the mean is not a location parameter in the previous example where θ is parameterized as $U(0, \theta)$.

- θ is a *generalized scale* parameter for X if there exist a constant $\bar{\theta}$ and a random variable Z that have no

dependence on θ such that $X = \bar{\theta} + Z\theta$ with probability 1. Then,

$$\frac{dX}{d\theta} = \frac{X - \bar{\theta}}{\theta}.$$

An example is the spread of a uniform distribution, that is $U(\bar{\theta} - \theta, \bar{\theta} + \theta)$.

Nearly every distribution has at least one parameter that is location or (generalized) scale, and many have both (e.g., normal, Cauchy, Gumel, logistic, general uniform).

To illustrate the ideas concretely, consider two sample performance functions of two random variables, this time without specifying the actual distributions.

Example 1. $L(X) = X_1 + X_2$.

- IPA estimator:

$$\frac{dX_1}{d\theta} + \frac{dX_2}{d\theta},$$

regardless of any dependence between X_1 and X_2 .

- LRM estimators:

$$(X_1 + X_2) \frac{\partial \ln f(X_1, X_2; \theta)}{\partial \theta},$$

for the general bivariate case, and

$$(X_1 + X_2) \left(\frac{\partial \ln f_1(X_1; \theta)}{\partial \theta} + \frac{\partial \ln f_2(X_2; \theta)}{\partial \theta} \right),$$

for the independent case.

Example 2. $L(X) = \max(X_1, X_2)$.

- IPA estimator:

$$\frac{dX_1}{d\theta} \mathbf{1}_{\{X_1 > X_2\}} + \frac{dX_2}{d\theta} \mathbf{1}_{\{X_1 < X_2\}},$$

regardless of any dependence between X_1 and X_2 , but generally unbiasedness requires $P(X_1 = X_2) = 0$ (e.g., when both are continuous random variables), due to the discontinuity at $X_1 = X_2$.

- LRM estimators:

$$\max(X_1, X_2) \frac{\partial \ln f(X_1, X_2; \theta)}{\partial \theta},$$

for the general bivariate case, and

$$\max(X_1, X_2) \left(\frac{\partial \ln f_1(X_1; \theta)}{\partial \theta} + \frac{\partial \ln f_2(X_2; \theta)}{\partial \theta} \right),$$

for the independent case.

In both examples, notice that the LRM estimators do not depend on the form of the sample performance, but on the number of random variables involved. Also, the simple examples illustrate that where the parameter appears is in some sense under the control of the modeler/analyst. So a service time parameter can be viewed as a parameter of the underlying distribution generating the service time random variate, or it can be viewed as a parameter directly influencing the value of the service time random variable itself, where the underlying uncertainty is simply a stream of random numbers. In some cases, the parameter could even appear in both places at the same time.

Generalizing these two examples leads to the following general forms of estimators, where the parameter can possibly occur in every input random variable (otherwise, that term is simply zero):

- IPA estimator:

$$\sum_i \frac{\partial L(X)}{\partial X_i} \frac{dX_i}{d\theta},$$

- LRM estimators: (multivariate, independent)

$$L(X) \frac{\partial \ln f(X; \theta)}{\partial \theta}, \quad L(X) \sum_i \frac{\partial \ln f_i(X_i; \theta)}{\partial \theta}.$$

PRACTICAL GUIDELINES

This section provides brief guidelines on the challenges and implementation of the two approaches.

Challenges

- LRM is likely to encounter computational challenges in the case where the same parameter recurs; for example, if the simulation model involves i.i.d. input random variables where θ is a parameter in the common distribution. Examples include service times or interarrival times in queueing and customer demands in inventory control. More details are provided below when implementation is discussed.
- The main requirement for IPA to succeed is that the sample performance be continuous with respect to the parameter of interest. Although technically this is only a sufficient and not necessary condition (i.e., it is not difficult to construct examples where IPA works for discontinuous sample performances), for all practical purposes this is also a necessary condition. In practice, a discontinuity manifests itself in two principal forms: inherently discontinuous sample performances, for example probability performance measures (where the sample performances are indicator functions), or any performance measure whose sample performance is discrete-valued (except in the trivial case where it is constant as a function of the parameter of interest); and simulation models where small changes in the parameter of interest could cause an abrupt change in system behavior such as a different event to occur; for example, a customer in a queueing system balking rather than being served.
- LRM may encounter difficulties in handling *structural* parameters, such as inventory control parameters, although a change of variables sometimes can be used to move the parameter into a distribution. The ability to move the parameter into or out of the underlying input distributions are sometimes referred to as “push-out” and “push-in” methods.
- LRM cannot handle parameters that change the underlying support of the distribution.

- For *discrete* distributions, IPA works if the parameter occurs in the support *values*, whereas LRM works if the parameter occurs in the support *probabilities*.
- *Higher derivative* estimates are generally easier to derive using LRM, though the variance of the resulting estimators may be problematic. A combination of two approaches might lead to an estimator with improved variance; for example, IPA for the first derivative and then LRM for the second derivative, or vice versa.

Implementation

- Implementation of the LRM estimator is usually the most straightforward. However, if the input process involves i.i.d. random variables whose common distribution depends on the parameter of interest, then the *variance of the estimator will grow linearly* with the number of times that random variable is used in the simulation, so the estimator can quickly become practically useless. This difficulty can be mitigated by batching; for example, rather than averaging over 1000 customers, take an average of 100 samples of 10 customers each. Using regenerative cycles is another way around the problem.
- Implementation of the IPA estimator generally requires *further knowledge of the dynamics of the system*, essentially to be able to implement the $\partial L / \partial X_i$ portion of the chain rule, where X_i is the input random variable through which the parameter dependence enters. In many cases of practical interest, this is not difficult to implement, and the additional computational burden is relatively minimal.

UNBIASEDNESS: VALIDITY OF THE INTERCHANGE

Justifying the exchange of differentiation (limit) and expectation operators required in Equations (9) or (10) can be equated with the concept of uniform integrability,

which is a necessary and sufficient condition. However, since uniform integrability is impossible to verify in practical applications, the key result used in the theoretical proofs of unbiasedness is the (Lebesgue) dominated convergence theorem.

For IPA, a sufficient condition is that the sample performance be continuous with respect to the parameter, which translates into requirements on the form of the performance measure and on the dynamics of the underlying stochastic system. A nice set of sufficient conditions that is often used requires Lipschitz continuity. In the framework of generalized semi-Markov processes (GMSPs) as a model for stochastic discrete-event simulation, the so-called “commuting condition” provides a useful set of structural conditions for checking the dynamics of the underlying system.

For LRM, the bounding condition in the dominated convergence theorem is applied to the (joint) density (or mass) function, where the bound is for $f(x; \theta)$ with respect to the parameter θ and not its usual argument x . This is more intuitively viewed as a requirement of absolute continuity of $f(\cdot; \theta + \Delta\theta)$ with respect to $f(\cdot; \theta)$, whereby for all x , if $f(x; \theta) = 0$, then $f(x; \theta + \Delta\theta) = 0$. A simple example illustrates how difficulties may arise. Consider the $U(0, \theta)$ probability density function

$$f(x; \theta) = \frac{1}{\theta} \mathbf{1}\{0 < x < \theta\},$$

where LRM does not apply. In this case, f viewed as a function of θ for fixed x has a discontinuity at $\theta = x$; in terms of absolute continuity, for $\Delta\theta > 0$, at the point $x = \theta + \Delta\theta/2$, the function fails to be absolutely continuous, since $f(\theta + \Delta\theta/2; \theta) = 0$, but $f(\theta + \Delta\theta/2; \theta + \Delta\theta) = 1/(\theta + \Delta\theta)$.

Smoothed Perturbation Analysis: An Extension of IPA

When IPA fails, numerous extensions have been proposed; thus, the term “perturbation analysis” actually refers to all of these different approaches and not just IPA; references for other PA techniques are given in the final section. This section briefly introduces the

most developed extension called smoothed perturbation analysis (SPA), which is based on conditional Monte Carlo, a well-known variance reduction technique in stochastic simulation in which the idea is to choose a random variable (or set of random variables), represented here by Z , such that

- (a) $E[L|Z]$ is easily computable;
- (b) $\text{Var}(E[L|Z])$ is small.

In variance reduction applications, it can be shown that $\text{Var}(E[L|Z]) \leq \text{Var}L$, so conditional Monte Carlo estimators are guaranteed to do no worse in terms of variance (this does not consider the additional computation that might be required).

Analogously, the main idea of SPA is to choose Z such that

- (a) $\frac{d}{d\theta}E[L|Z]$ is easily computable;
- (b) $E\left[\frac{d}{d\theta}E[L|Z]\right] = \frac{d}{d\theta}E[E[L|Z]] = \frac{dE[L]}{d\theta}$.

Thus, the main challenges in applying SPA are (i) the choice of what to condition on and (ii) how to compute (estimate) the resulting conditional expectation derivative. Challenge (ii) may lead to many additional simulations.

To demonstrate how this works for the two simple examples, consider a probability performance measure, that is $P(L(X) \leq x)$ for some fixed x , for which IPA fails because the sample performance is an indicator function. For simplicity of exposition, take the case where the two random variables are independent and assume θ enters into the distribution of the first random variables X_1 . For $L(X) = X_1 + X_2$, noting that $P(X_1 + X_2 \leq x) = E[\mathbf{1}\{X_1 + X_2 \leq x\}]$, simply condition on the other random variable:

$$\begin{aligned} E[\mathbf{1}\{X_1 + X_2 \leq x\}|X_2] &= P(X_1 \leq x - X_2|X_2) \\ &= F_1(x - X_2; \theta), \end{aligned}$$

where F_1 is the cumulative distribution function for X_1 . Differentiating gives the density as the SPA estimator:

$$\frac{d}{d\theta}E[\mathbf{1}\{X_1 + X_2 \leq x\}|X_2] = f_1(x - X_2; \theta).$$

Similarly, for the other sample performance $L(X) = \max(X_1, X_2)$,

$$\begin{aligned} E[\mathbf{1}\{\max(X_1, X_2) \leq x\}|X_2] \\ &= P(X_1 \leq x|X_2 \leq x)\mathbf{1}\{X_2 \leq x\} + 0 \cdot \mathbf{1}\{X_2 > x\} \\ &= F_1(x; \theta)\mathbf{1}\{X_2 \leq x\}, \end{aligned}$$

and differentiating gives the SPA estimator:

$$f_1(x; \theta)\mathbf{1}\{X_2 \leq x\}.$$

APPLICATIONS

As mentioned earlier, the three main operations research/management science application areas for stochastic gradient estimation have been queueing systems, production/inventory systems, and finance. Some simple illustrative examples are provided for each. Chronologically, the first application area was queueing systems, which dominated the research focus for almost two decades; since queueing networks can be used to model so many different classes of large-scale systems, including communications networks, manufacturing systems, supply chains, or transportation network/systems, queueing network simulation models cover a huge range of applications. Applications to inventory control and financial engineering began in the 1990s.

Queueing

Application of IPA and LRM will be illustrated using the simplest queueing example—a first-come, first-served single-server queue, taking the (sample) performance measure as the waiting time (in queue) of the n th customer ($n > 1$), denoted by W_n , with the parameter θ in the service time distribution, where X_i denotes the service time of the i th customer. To be concrete, take θ as the mean of an exponential distribution.

First consider the case where the parameter θ affects only a single service time, specifically $X_{n-1} \sim \exp(\theta)$, giving the following estimators for $\frac{dE[W_n]}{d\theta}$:

- IPA

$$\frac{dX_{n-1}}{d\theta}\mathbf{1}\{W_n > 0\} = \frac{X_{n-1}}{\theta}\mathbf{1}\{W_n > 0\},$$

which reflects the notion that the service time of a customer only affects the waiting times of subsequent customers in the same busy period, where the condition $W_n > 0$ implies that the n th and $(n - 1)$ th customer are in the same busy period.

- LRM

$$\frac{W_n}{\theta} \left(\frac{X_{n-1}}{\theta} - 1 \right).$$

Now consider the case where the parameter affects all of the service times, specifically θ is the common mean for the i.i.d. exponentially distributed service times $\{X_i\}$, which leads to the following estimators for $\frac{dE[W_n]}{d\theta}$:

- IPA

$$\sum_{n^* \leq i < n} \frac{dX_i}{d\theta} = \sum_{n^* \leq i < n} \frac{X_i}{\theta},$$

where n^* denotes the index of the customer that begins the busy period of customer n , so the sum is empty if customer n begins the busy period.

- LRM

$$\frac{W_n}{\theta} \sum_{i < n} \left(\frac{X_i}{\theta} - 1 \right),$$

which basically has variance proportional to n .

Although the straightforward LRM estimator is computationally prohibitive for large n , by utilizing the independence of W_n from service times before the start of customer n 's busy period, which is done implicitly in the IPA estimator, the variance problem can be ameliorated. In other words, for $i < n^*$, $E[W_n X_i] = E[W_n]E[X_i] \implies E[W_n(X_i/\theta - 1)] = 0$, so a much-improved LRM estimator with the same expected value is given by

$$\frac{W_n}{\theta} \sum_{n^* \leq i < n} \left(\frac{X_i}{\theta} - 1 \right).$$

This illustrates one way in which regeneration can be used to mitigate potential variance problems in LRM estimators.

Inventory

Inventory control is the second class of applications to which derivative estimation has been successfully applied. When simulating inventory control systems utilizing base-stock policies, IPA estimators can be easily implemented, and if the sample performance is continuous with respect to the base-stock level—often the case in practice—the resulting IPA estimators are unbiased. Difficulties arise when setup costs are included in the model, requiring a two-parameter order policy with a reorder point and order quantity or order-up-to level, because small changes in one of the parameters can cause discontinuities in the sample path. To illustrate the challenge, consider an (s, S) inventory control example, where s is the reorder point and S is the order-up-to point. Rather than the usual parameters s and S , the parameters of interest will be s and $q = S - s$, where the latter can be viewed as the (approximate) order quantity. The sample paths will show intuitively the continuity with respect to changes in s for q held constant, whereas there are discontinuities with respect to changes in q with s held constant. As a result, IPA estimators will be unbiased for performance measures with respect to s (with q held constant), whereas they would be biased for performance measures with respect to q , and SPA estimators would be required.

Figure 1 shows two sample paths of the inventory position, where the vertical lines indicate the order decision points at the end of each period. Thus, if the inventory position is below the reorder point s , then an order up to S is placed. The bolder path is the original sample path—called the *nominal path* in PA literature, whereas the lighter path—called the *perturbed path*—is the path after the perturbation by Δs with q held constant, which results in S also being increased by Δs . The key in the sample path analysis is that discontinuities in going from the nominal path to the perturbed path can only occur at the reorder points if an order decision becomes a no-order decision or vice versa. In Figure 1, there is an order placed at the end of period n in the nominal path, and there is also an

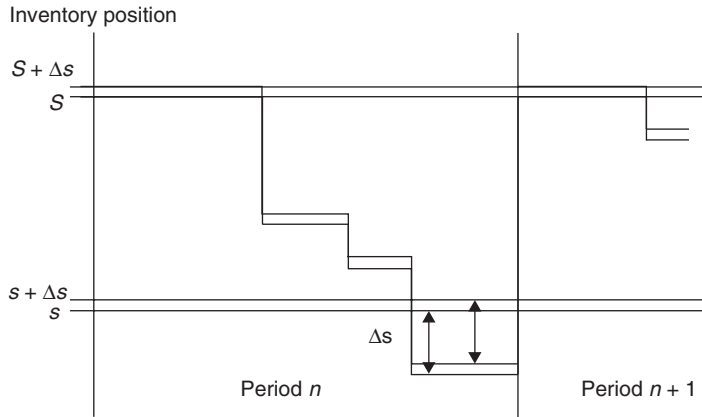


Figure 1. IPA is easy for (s, S) inventory system for the derivative with respect to s , assuming q held fixed, because the distance of the inventory position from the reorder point is the same in the perturbed (lighter) and the original (bolder) sample paths, as indicated by arrows.

order placed at the end of period n in the perturbed path. Most importantly, the distance of the inventory position from the reorder point is unchanged before and after the perturbation, *regardless of the size (and sign) of the perturbation*, and thus the perturbed path is simply the nominal path changed by Δs .

Figure 2 shows two analogous sample paths when s is kept constant, but q is increased by $\Delta q > 0$. In the case shown, Δq is sufficiently large so that the perturbed path ends up not ordering at the end of period n , which results in a drastic divergence from the original (nominal) sample path, resulting in a discontinuity in a sample performance defined on the process.

Finance

In financial risk management, hedging of contracts is a key focus, and gradients play the central role in executing any hedging strategy. These gradients are called the “Greeks” in finance, because Greek letters are used to represent the various partial derivatives (sensitivities). For example, the most famous ones include the “delta,” “vega,” and “rho,” which are the sensitivities of the price with respect to the underlying asset price, the underlying asset’s volatility, and the interest rate, respectively. As pointed out in the Lancaster Prize-winning book by Glasserman, *Monte Carlo Methods in Financial Engineering* (p. 377):

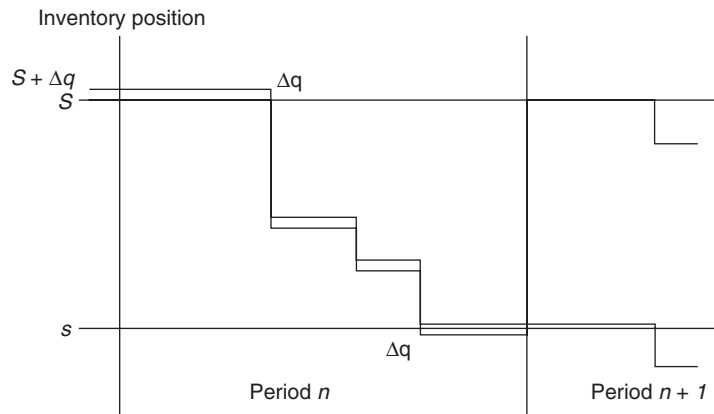


Figure 2. IPA fails for (s, S) inventory system for the derivative with respect to q (or S), assuming s held fixed, because a positive perturbation can cause an order decision in the original (bolder) path to become a no-order decision in the perturbed (lighter) path.

“Whereas the prices themselves can often be observed in the market, their sensitivities cannot, so accurate calculation of sensitivities is arguably even more important than calculation of prices.”

As mentioned previously, application of gradient estimation methods in simulation to the finance setting did not occur until the mid 1990s, and probably did not become known in the finance community until after the turn of the century, as indicated by the fourth (2000) edition of a very popular textbook by John Hull [5] on derivatives pricing, *Options, Futures, and Other Derivatives Securities* (p. 410):

“Suppose that we are interested in the partial derivative of f with q , where f is the value of the derivative and q is the value of an underlying variable or a parameter. First, Monte Carlo simulation is used in the usual way to calculate an estimate, f , for the value of the derivative. A small increase, Δq , is then made in the value of q , and a new value for the derivative, f^* , is calculated. An estimate for the hedge parameter is given by

$$\frac{f^* - f}{\Delta q}$$

In order to minimize to [sic] standard error of the estimate of the Greek letter, the number of time intervals, N , the random number streams, and the number of trials, M , should be the same for estimating both f and f^* .”

Simulation is widely used on Wall Street and throughout the financial industry worldwide, and the techniques discussed in this article can be used to obtain estimates of the sensitivities with respect to various parameters without resimulation. In both pricing and hedging of derivatives contracts, the final problem usually comes down to an optimization problem. For hedging, one usually tries to minimize some (expected or probability of) deviation from a target payoff (or path of payoffs), whereas for pricing, the optimization problem is generally to maximize the expected payoff in a classical optimal stopping problem. With regards to pricing, the second (1993) edition of the previously mentioned textbook by John Hull had this to say:

“Monte Carlo simulation can only be used for European-style options.”

However, research advances in developing simulation algorithms for pricing American-style derivatives has led to this statement being tempered to “... cannot easily handle situations where there are early exercise opportunities” and actually including material describing some of the approaches. One of the simulation approaches that has been developed uses gradient-based stochastic optimization, where the decision of whether or not to exercise an option is made according to a decision rule that is defined by a parameterized exercise boundary, and thus gradients with respect to the parameters are central to the method.

To illustrate the ease with which these estimators can be implemented, consider a simple European call stock option with payoff given by

$$(S_T - K)^+,$$

where S_t denotes the stock price at time t , T is the expiration time of the option, and K is the strike price. The option price is given by the expectation of the discounted payoff, where here the discounting assumes a constant risk-free continuous interest rate r . The delta of this option is the sensitivity with respect to the current stock price S_0 :

$$\frac{d}{dS_0} E \left[e^{-rT} (S_T - K)^+ \right].$$

The IPA estimator is straightforward to obtain, although the final form will depend on the form of the underlying stock price model. To apply LRM, the “parameter” of interest S_0 has to be incorporated into the underlying distribution of S_T . How this is done also depends on the underlying stock price model. For the Black–Scholes (geometric Brownian motion process, lognormally distributed) stock price model with volatility σ , the respective IPA and LRM estimators for the option delta are given by

$$e^{-rT} \frac{S_T}{S_0} \mathbf{1}_{\{S_T > K\}}, \quad e^{-rT} (S_T - K)^+ \frac{Z}{S_0 \sigma \sqrt{T}},$$

where $Z \sim N(0, 1)$ is the standard normal random variate used to generate S_T from S_0 via

$$S_T = S_0 e^{(r - \sigma^2/2)T + \sigma\sqrt{T}Z}.$$

The form of the IPA estimator remains unchanged for stock price models in which the current stock price acts as a scale parameter for the future stock price, for example exponential Lévy processes.

SUMMARY

The main message is that when conducting stochastic simulation, efficient gradient estimates can often be easily obtained from the original simulation just by doing a little extra calculation and *without having to alter any model parameter settings*. These gradient estimates can be used in gradient-based simulation optimization, which can take many different forms, from stochastic approximation to sequential response surface methodology to sample average approximation based on nonlinear deterministic optimization formulations.

LRM requires knowledge of the distributions that serve as inputs to the simulation model. In most settings, these are known and thus, it is relatively easy to obtain sensitivity estimates. IPA is more restrictive, because its applicability depends on properties of both the system model and the performance measure(s) of interest. In cases where both techniques can be applied, the IPA estimate usually has lower variance.

Because of the more limited applicability of IPA, there has been an abundance of research to extend its applicability, resulting in numerous other techniques that are more specialized and problem dependent and sometimes requiring additional simulations. The most generally applicable of these is SPA, which is based on conditional Monte Carlo, relying on the “smoothing” property of conditional expectation to allow the interchange in Equation (9).

PROBING FURTHER

A large part of the exposition here is excerpted and condensed from the cover article, [6], which also discusses estimators based on weak derivatives (WD); for a recent survey of the latter, see Heidegott [7]. A further in-depth discussion can be found in the handbook chapter by Fu [8]. Both references include a table for implementing many commonly used input probability distributions in IPA, LRM, or WD estimators. Treatment at a similar level for the more general PA is presented in Fu *et al.* [9] and for the more general “score functions” in Rubinstein *et al.* [10].

Recent books on simulation containing extensive coverage of gradient estimation include Glasserman [11], Chapter 7, which treats both IPA and LRM in the finance setting, although IPA is called the “pathwise” method there, which seems to have become the generally accepted nomenclature in finance; Chapter VII of Asmussen and Glynn [12], which also includes both IPA and LRM estimators; and Chapter 11 of the broader discrete-event systems textbook by Cassandras and Lafortune [13], which includes coverage of IPA, SPA, and finite perturbation analysis (FPA). Earlier books focusing on specific gradient estimation approaches include Ho and Cao [14], Glasserman [4], and Cao [15] for IPA, and Rubinstein and Shapiro [16] for LRM, where it is called the score function method; see also Rubinstein and Melamed [17]. A detailed discussion of the “commuting condition” for IPA can be found in Glasserman [4].

As pointed out early in the discussion, IPA is just one method—albeit the most common and easily implementable technique—in the vast PA arsenal that has been developed over many decades. The main idea behind the PA approach is to study the effect on the sample performance when a parameter value is perturbed slightly; however, in most PA methods, the parameter is never actually altered in either the analysis or in the simulation, for example IPA can be viewed as the limiting case when the size of the perturbation vanishes. Using conditional Monte Carlo to extend IPA’s applicability

in the framework of SPA was introduced by Gong and Ho [18], though the conditioning idea was also used in Suri and Zazanis [3], and the book by Fu and Hu [19] covers SPA in full generality, including numerous application areas. Other PA techniques include rare perturbation analysis (RPA), originally based on the thinning of point processes [20]; structural infinitesimal perturbation analysis (SIPA), dealing specifically with structural parameters [21]; discontinuous perturbation analysis (DPA), based on the use of generalized functions (the Dirac-delta function) to model discontinuities in the sample performance function [22]; and augmented infinitesimal perturbation analysis (APA), another extension of IPA different from SPA [23]. Techniques to estimate the effect of a *finite* perturbation in the parameter value include FPA [24]; extended perturbation analysis (EPA) [25]; and the augmented chain method [26]. A related technique is the standard clock (SC) method, based on the uniformization of Markov chains [27].

As pointed out several times, it is theoretically possible for θ to appear simultaneously in two places, for example,

$$E[L(X)] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(x; \theta) dF(x; \theta).$$

A theoretical framework for such a “unified” view of IPA and LRM was presented in L’Ecuyer [28], where a few simple examples are presented. Recently Wang *et al.* [29,30] have demonstrated practical examples arising naturally in finance applications where the “push-in/push-out” change of variables technique described in Rubinstien and Shapiro [16] for LRM can be used to derive a hybrid IPA/LRM estimator in a situation where neither IPA nor LRM is applicable by itself.

Besides queueing, inventory, and finance, other applications include stochastic activity networks, preventive maintenance, statistical process control, and traffic light control; for specific references, see Fu [8], which also contains sources for those approaches mentioned in the introduction not covered in this article: Malliavin calculus, WD, simultaneous perturbations, and frequency domain

experimentation. Not included there are the more recent developments for quantiles, which include Hong [31] for IPA and Fu *et al.* [32] for SPA, and also the connection between LRM and the Malliavin calculus approach in Chen and Glasserman [33].

For more on gradient-based simulation optimization applications and issues, see Fu [34], Spall [36], Andradottir [37], Rubinstein and Melamed [38], Pflug [39], Rubinstein and Shapiro [16], as well as other chapters in the Elsevier Handbook on Simulation edited by Henderson and Nelson [40].

Finally, a useful source for following the state-of-the-art research developments in stochastic gradient estimation and simulation optimization are the annual Proceedings of the Winter Simulation Conference, available on-line at <http://www.wintersim.org>.

Brief Historical Notes

PA was “invented” by Ho *et al.* [41], when the first author was consulting on a real-world buffer design problem for a Fiat Motor Company serial production line. The LR/SF method was introduced by Glynn [42], Reiman and Weiss [43], and Rubinstein [44]. The first application of PA to inventory systems was by Fu [35] for (s,S) policies, where both IPA and SPA estimators are derived. Glasserman and Tayur [45] treat the easier base-stock setting where IPA is applicable. The first applications of stochastic gradient estimation to finance are contained in Fu and Hu [46] and Broadie and Glasserman [47].

REFERENCES

1. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1982.
2. Ross SM. Introduction to probability models. New York: Academic Press; 1972.
3. Suri R, Zazanis MA. Perturbation analysis gives strongly consistent sensitivity estimates for the M/G/1 queue. *Manage Sci* 1988;34: 39–64.
4. Glasserman P. Gradient estimation via perturbation analysis. Boston (MA): Kluwer Academic; 1991.
5. Hull JC. Options, futures, and other derivative securities. 4th ed. Englewood Cliffs (NJ): Prentice Hall; 2000.

6. Fu MC. What you should know about simulation and derivatives (Cover Story). *Nav Res Log* 2008;55:723–736.
7. Heidergott B. Gradient estimation for discrete event systems by measure-valued differentiation. *ACM Trans Model Comput Simul* 2010;20(1).5:1–5:28.
8. Fu MC. Gradient estimation. In: Henderson SG, Nelson BL, editors. *Handbooks in operations research and management science: simulation*. New York: Elsevier; 2006. chapter 19. pp. 575–616.
9. Fu MC. Perturbation analysis. In: Gass S, Harris C, editors. *Encyclopedia of Operations Research and Management Science*. 2nd ed. Boston: Kluwer Academic Publishers; 2001. pp. 608–611.
10. Rubinstein RY, Shapiro A, Uryasev S. Score functions. In: Gass S, Harris C, editors. *Encyclopedia of Operations Research and Management Science*. 2nd ed. Boston: Kluwer Academic Publishers; 2001. pp. 614–617.
11. Glasserman P. *Monte Carlo methods in financial engineering*. New York: Springer; 2004.
12. Asmussen S, Glynn P. *Stochastic simulation: algorithms and analysis*. New York: Springer; 2007.
13. Cassandras CG, Lafortune S. *Introduction to discrete event systems*. New York: Springer; 2008.
14. Ho YC, Cao XR. *Discrete event dynamics systems and perturbation analysis*. Boston (MA): Kluwer Academic; 1991.
15. Cao XR. *Realization probabilities: the dynamics of queueing systems*. New York: Springer; 1994.
16. Rubinstein RY, Shapiro A. *Discrete event systems: sensitivity analysis and stochastic optimization by the score function method*. New York: John Wiley & Sons, Inc.; 1993.
17. Rubinstein RY, Melamed B. *Modern simulation and modeling*. New York: John Wiley & Sons, Inc.; 1998.
18. Gong WB, Ho YC. Smoothed perturbation analysis of discrete-event dynamic systems. *IEEE Trans Automat Control* 1987;32: 858–867.
19. Fu MC, Hu JQ. *Conditional Monte Carlo: gradient estimation and optimization applications*. Boston (MA): Kluwer Academic; 1997.
20. Brémaud P, Vázquez-Abad FJ. On the path-wise computation of derivatives with respect to the rate of a point process: The phantom RPA method. *Queueing Syst Theory Appl* 1992;10:249–270.
21. Dai L, Ho YC. Structural infinitesimal perturbation analysis for derivative estimation in discrete event dynamic systems. *IEEE Trans Automat Control* 1995;40:1154–1166.
22. Shi LY. Discontinuous perturbation analysis of discrete event dynamic systems. *IEEE Trans Automat Control* 1996;41:1676–1681.
23. Gaivoronski A, Shi L, Sreenivas RS. Augmented infinitesimal perturbation analysis: an alternate explanation. *Discrete Event Dyn Syst Theory Appl* 1992;2:121–138.
24. Ho YC, Cao XR, Cassandras CG. Infinitesimal and finite perturbation analysis for queueing networks. *Automatica* 1983;19:439–445.
25. Ho YC, Li S. Extensions of infinitesimal perturbation analysis. *IEEE Trans Automat Control* 1988;33:827–838.
26. Cassandras CG, Strickland SG. On-line sensitivity analysis of Markov chains. *IEEE Trans Automat Control* 1989;AC-34:76–86.
27. Vakili P. Using a standard clock technique for efficient simulation. *Oper Res Lett* 1991;10: 445–452.
28. L'Ecuyer P. A unified view of the IPA, SF, and LR gradient estimation techniques. *Manage Sci* 1990;36:1364–1383.
29. Wang Y, Fu MC, Marcus SI. A new stochastic gradient estimator for American option pricing. *Eur Control Conf* 2009:1191–1196.
30. Wang Y, Fu MC, Marcus SI. A new stochastic derivative estimator for discontinuous payoff functions with application to financial derivatives. *Oper Res* 2010.
31. Hong LJ. Estimating quantile sensitivities. *Oper Res* 2009;57(1):118–130.
32. Fu MC, Hong LJ, Hu JQ. Conditional Monte Carlo estimation of quantile sensitivities. *Manage Sci* 2009;55(12):2019–2027.
33. Chen N, Glasserman P. Malliavin Greeks without Malliavin calculus. *Stoch Processes Appl* 2007;117:1689–1723.
34. Fu MC. Optimization for simulation: Theory vs. practice (Feature Article). *INFORMS J Comput* 2002;14(3):192–215.
35. Fu MC. Sample path derivatives for (s, S) inventory systems. *Oper Res* 1994;42(2):351–364.
36. Spall JC. *Introduction to stochastic search and optimization: estimation, simulation, and control*. Hoboken (NJ): Wiley; 2003.
37. Andradóttir S. Simulation optimization. In: Banks J, editor. *Handbook of simulation: principles, methodology, advances, applications,*

- and practice. New York: John Wiley & Sons, Inc.; 1998. Chapter 9. pp. 307–333.
38. Rubinstein, Melamed. *Modern Simulation and Modeling* John Wiley & Sons: New York; 1996.
39. Pflug GC. *Optimization of stochastic models*. Boston (MA): Kluwer Academic; 1996.
40. Henderson SG, Nelson BL, editors. *Handbooks in operations research and management science: simulation*. New York: Elsevier; 2006.
41. Ho YC, Eyler MA, Chien TT. A gradient technique for general buffer storage design in a serial production line. *Int J Prod Res* 1979; 17:557–580.
42. Glynn PW. Likelihood ratio gradient estimation: an overview. *Proceedings of the 1987 Winter Simulation Conference*. Piscataway (NJ): IEEE; 1987. pp. 366–375.
43. Reiman MI, Weiss A. Sensitivity analysis for simulations via likelihood ratios. *Oper Res* 1989;37:830–844.
44. Rubinstein RY. Sensitivity analysis of computer simulation models via the score efficient. *Oper Res* 1989;37:72–81.
45. Glasserman P, Tayur S. Sensitivity analysis for base-stock levels in multi-echelon production-inventory systems. *Manage Sci* 1995;41:263–281.
46. Fu MC, Hu JQ. Sensitivity analysis for Monte Carlo simulation of option pricing. *Probab Eng Inform Sci* 1995;9:417–446.
47. Broadie M, Glasserman P. Estimating security price derivatives using simulation. *Manage Sci* 1996;42:269–285.

STOCHASTIC HAZARD PROCESS

SÜLEYMAN ÖZEKİCİ
Department of Industrial
Engineering, Koç University,
Istanbul, Turkey

The reliability of a device or system is represented by the survival probability

$$R(t) = P\{L > t\} = e^{-H(t)} \quad (1)$$

for $t \geq 0$ where L is the lifetime and H is the hazard function. This may also be rewritten as

$$H(t) = -\ln R(t) \quad (2)$$

and the hazard function is increasing in $t \geq 0$. Moreover, if it is absolutely continuous and the hazard rate function $h(t) = dH(t)/dt$ exists almost everywhere, then the hazard at time t is $H(t) = \int_0^t h(s)ds$.

In reliability theory, perhaps the most widely used model to represent the lifetime is the first passage time,

$$L = T_S = \inf \{t \geq 0; X_t \geq S\} \quad (3)$$

when a wear process $X = \{X_t; t \geq 0\}$ exceeds a critical threshold $S \geq 0$. This implies that the item fails as soon as the total wear X exceeds the critical level S . The probabilistic structure of failure is then embedded in the stochastic structure of the wear process X and the threshold S . The literature is full of a variety of models with continuous wear and discontinuous shocks as they occur randomly in time. We refer the reader to Abdel-Hameed *et al.* [1] and Aven and Jensen [2] for examples.

In this article, we focus on the hazard function and process. Our terminology will call H the hazard function when it is deterministic, and hazard process when it is stochastic. To make this distinction clear, we use the notation $H(t)$ and H_t to denote the deterministic and stochastic hazard accumulated until time

t , respectively. Needless to say, as implied by the title, the discussion is on the stochastic case for the most part. However, we first take a look at some of the implications behind the simple deterministic representation (1) in the section titled “The Hazard Function”. The section titled “The Hazard Process” discusses some stochastic representations for H , the next section discusses briefly “Modulation via Random Environments”, and we take a more detailed look at Markov modulated models in the section titled “Intrinsic Aging Models” after a brief discussion on modulation via random environments in the third section. Finally, this article is intended to provide the reader a clear and useful understanding and interpretation of the hazard function and process. Technical details and proofs are omitted and we will try to keep the theoretical derivations at the minimal level. In our notation, we let $\mathbb{R}_+ = [0, +\infty)$ and $\mathbb{R}_0 = (0, +\infty)$ denote the set of nonnegative and positive real numbers, respectively.

THE HAZARD FUNCTION

We now consider some models where the hazard function H has meaningful interpretations. An interesting characterization of the lifetime distribution implied by (1) is that

$$P\{L > t\} = e^{-H(t)} = P\{\hat{L} > H(t)\}, \quad (4)$$

where $\hat{L}$ is a random variable that has the exponential distribution with rate 1. In view of the passage time representation (3), we can also write $L = T_{\hat{L}}$ as the passage time of the deterministic function $X_t = H(t)$ when the threshold $S = \hat{L}$ is exponentially distributed with rate 1. One can therefore think of the increasing process H as an “intrinsic” aging process such that the system fails as soon as its intrinsic age exceeds its intrinsic lifetime $\hat{L}$. This observation allowed Çınlar and Özekici [3] to measure the age of complex systems by an intrinsic clock given by the hazard

function so that the intrinsic lifetime is simply exponentially distributed. This implies that any system is initially endowed by an exponentially distributed lifetime with rate 1 and it fails when its intrinsic age, or hazard, exceeds this threshold. Singpurwalla [4], in a recent closely related article, calls this endowment the hazard potential that the system consumes by time and it fails once it is depleted. Most of the stochastic hazard models that we consider later in this article will be extensions of this idea where intrinsic aging depends on some random environmental process that affects the deterioration or failure properties of the system. This provides an intuitive and natural construction of stochastic hazard processes.

The structure in (1) also holds in shock models where the system is subject to shocks at times $\{U_i\}$ with random magnitudes $\{Y_i\}$ such that the pairs $\{(U_i, Y_i)\}$ form a Poisson random measure M on $\mathbb{R}_+ \times \mathbb{R}_0$ with some mean measure $m(dt, dy) = \lambda(t)dtF(dy)$, where λ is the shock arrival rate function and F is the distribution function of the shock magnitudes. If the system fails as soon as the magnitude of a shock at time $t \geq 0$ exceeds a critical threshold $a(t) \geq 0$, then

$$\begin{aligned} P\{L > t\} &= P\{M\{(u, y); 0 \leq u \leq t, y > a(t)\} = 0\} \\ &= e^{-H(t)}, \end{aligned} \quad (5)$$

where

$$\begin{aligned} H(t) &= m\{(u, y); 0 \leq u \leq t, y > a(t)\} \\ &= \int_0^t \lambda(u) du \int_{a(u)}^{+\infty} F(dy) \\ &= \int_0^t \lambda(u)(1 - F(a(u))) du, \end{aligned} \quad (6)$$

and the hazard rate function becomes $h(t) = dH(t)/dt = \lambda(t)(1 - F(a(t)))$. The deterministic function a depicts the shock tolerance of the item and it will decrease in time if the item gets less tolerant to shocks as it ages. A shock with the same magnitude at a later time is then more likely to cause failure.

A special case of (6) is when the threshold level is $a = 0$ and $Y_i = 1$ trivially so that $L = U_1$ and the system fails at the time of the

first shock. This gives

$$H(t) = \int_0^t \lambda(u) du \quad (7)$$

by (6). Now, the stochastic process

$$N_t = \int_0^t \int_{\mathbb{R}_0} y M(du, dy) = \sum_{U_i \leq t} Y_i = \sum_i I_{\{U_i \leq t\}} \quad (8)$$

is a nonstationary Poisson process with mean function $H(t) = \int_0^t \int_{\mathbb{R}_0} y m(du, dy) = \int_0^t \lambda(u) du$ and L is the time of first arrival of N . This provides a very simple interpretation of the hazard function H since the hazard rate $h(t) = dH(t)/dt = \lambda(t)$ is simply the arrival rate of the nonstationary Poisson process at any time t . One can therefore think of the failure time as the time of the first arrival of a nonstationary Poisson process with expectation function given by the hazard function.

THE HAZARD PROCESS

The hazard functions discussed in the previous section are implicitly assumed to be deterministic. In this article, we focus on an interesting extension of this idea that has attracted a lot of attention by supposing that $H = \{H_t; t \geq 0\}$ is indeed an increasing stochastic process taking values in $\mathbb{R}_+$. We provided two interpretations of the hazard function as an intrinsic aging process by (4), and as an arrival expectation function by (7). A natural extension is to randomize them via modulation by an external stochastic process. We focus on this issue later in more detail in the third and fourth sections. The concept of random hazard functions is also used in earlier papers by Gaver [5] and Arjas [6]. The intrinsic aging model of Çinlar and Özekici [3] is another example where the stochastic hazard process is modulated by an external environmental process that describes the conditions under which the system operates. Lindley and Singpurwalla [7] discuss the effects of the random environment on the reliability of a system consisting of components which share the same environment. Although the initial state

of the environment is random, they assume that it remains constant in time and components have exponential life distributions in each possible environment. Kebir [8] models the hazard rate as a stochastic process with independent increments. In case the hazard rate process is increasing, this leads to an explicit computational formula for the reliability function using its Itô decomposition. In a recent article, Singpurwalla [4] introduces a novel approach to model lifetime distribution through the hazard potential. Stochastic hazard functions are also used in different contexts that do not necessarily involve devices. For example, Ludkovski and Young [9] suppose that the hazard or mortality rate of an individual's lifetime is given by a diffusion process in their financial model on the pricing of mortality contingent claims. Cathcart and Al-Jahel [10] use a doubly stochastic hazard rate process for the default time of bonds where the primary objective is to price them.

The stochastic hazard process H is an increasing right-continuous process and Çinlar [11] gives an excellent description of such processes with Markov and semi-Markov structures. Although his discussions are on the stochastic structure of wear processes, the results also apply to our hazard process. For example, if H is continuous and strong Markov, then it must be deterministic and this case indeed corresponds to a deterministic hazard function discussed in the first section for which (4) holds. An interesting model of sufficient generality for our purposes is when H is a right-continuous strong Markov process that is also quasi-left-continuous. Then, we have the stochastic representation

$$H_t = H_0 + \int_0^t b(H_u) du + \int_{[0,t] \times \mathbb{R}_0} k(H_{u-}, y) M(du, dy), \quad (9)$$

where $b: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, $k: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are deterministic functions and M is a Poisson random measure with some mean measure $m(du, dy) = duv(dy)$ for some σ -finite, diffuse measure v on $\mathbb{R}_+$. The structure is slightly more complicated since the actual hazard process is obtained from a process

that satisfies (9) after a random time change. But, we think that the model in (9) is of sufficient interest, generality, and computational tractability. In fact, the infinitesimal generator

$$Gf(x) = \lim_{t \downarrow 0} \left(\frac{E[f(H_t) - f(H_0) | H_0 = x]}{t} \right) \quad (10)$$

of the Markov process H is given by

$$Gf(x) = f'(x)b(x) + \int_{\mathbb{R}_0} [f(x+k(x,y)) - f(x)] v(dy) \quad (11)$$

for bounded and differentiable functions $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ with derivative $f'(x) = df(x)/dx$. If the points of the Poisson random measure M are labeled as $\{(U_i, Y_i)\}$, we can also write

$$H_t = H_0 + \int_0^t b(H_u) du + \sum_{U_i \leq t} k(H_{U_i-}, Y_i) \quad (12)$$

in a simplified format. Here, $b(H_u)$ is the continuous rate of increase in the hazard depending on the hazard level H_u at any time u . Moreover, a shock occurring at time U_i with magnitude Y_i increases the hazard by an amount $k(H_{U_i-}, Y_i)$ depending on the hazard level H_{U_i-} just before the shock time U_i and the shock magnitude Y_i .

When the hazard process is stochastic, we can express the reliability function as

$$R(t) = P\{\hat{L} > H_t\} = E\left[E\left[\hat{L} > H_t | H_t\right]\right] = E\left[e^{-H_t}\right], \quad (13)$$

in view of (4) by supposing that the intrinsic lifetime $\hat{L}$ is independent of the hazard process H . In other words, the system fails as soon as the hazard process exceeds an exponential threshold with rate 1.

For the increasing Markov process H in (9), we can use potential theory and the generator (11) to write

$$E[f(H_t) | H_0 = x] = f(x) + \int_0^t E[Gf(H_s) | H_0 = x] ds. \quad (14)$$

In particular, when $f(x) = e^{-x}$, we trivially have $f'(x) = -e^{-x} = -f(x)$ and (14) becomes

$$\begin{aligned} E[e^{-H_t} | H_0 = x] \\ = e^{-x} - \int_0^t E[e^{-H_s} c(H_s) | H_0 = x] ds, \end{aligned} \quad (15)$$

where $c(x) = b(x) + \int_{\mathbb{R}_0} [1 - e^{-k(x,y)}] v(dy)$. Therefore, given any initial hazard $H_0 = x$, the reliability function $R(t) = E[e^{-H_t} | H_0 = x]$ satisfies the differential equation

$$\frac{dE[e^{-H_t} | H_0 = x]}{dt} = -E[e^{-H_t} c(H_t) | H_0 = x] \quad (16)$$

with the boundary condition $R(0) = E[e^{-H_0} | H_0 = x] = e^{-x}$. One has to solve (16) in order to determine the reliability function corresponding to the stochastic hazard process given by (9). As a special case, if $b(x) = b$, and $k(x, y) = k(y)$ independent of x , then $c(x) = b + \int_{\mathbb{R}_0} [1 - e^{-k(y)}] v(dy) = c$ independent of x . Therefore, (16) clearly implies that $dR(t)/dt = -cR(t)$ and this leads to the exponential distribution since the solution is $R(t) = R(0)e^{-ct}$.

A reasonable example for stochastic hazard processes is an increasing Lévy process which is basically a process that has independent and stationary increments. This is a very useful process in modeling stochastic phenomena occurring in many fields of sciences and engineering. Various properties and applications of Lévy processes are discussed in Barndorff-Nielsen *et al.* [12]. Gjessing *et al.* [13] use stochastic hazard processes based on Lévy processes and discuss applications in biological sciences. An increasing Lévy process has the representation

$$H_t = bt + \int_{[0,t] \times \mathbb{R}_0} yM(du, dy) \quad (17)$$

with $b \geq 0$, where M is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{R}_0$ with mean measure $m(du, dy) = duv(dy)$ for some Lévy measure v that satisfies $\int_{\mathbb{R}_0} \min\{1, y\}v(dy) < +\infty$. This is clearly a special case of the general Markov model in (9) obtained by taking

$b(x) = b, k(x, y) = y$ independent of x , and setting $H_0 = 0$. The structure in (17) implies that the expected hazard is

$$\begin{aligned} E[H_t] &= bt + \int_{[0,t] \times \mathbb{R}_0} ym(du, dy) \\ &= \left(b + \left(\int_{\mathbb{R}_0} yv(dy) \right) \right) t \end{aligned} \quad (18)$$

and reliability becomes

$$R(t) = E[e^{-H_t}] = e^{-\left(b + \left(\int_{\mathbb{R}_0} yv(dy)(1 - e^{-y})\right)\right)t} = e^{-ct}, \quad (19)$$

where $c = b + \int_{\mathbb{R}_0} (1 - e^{-y}) v(dy)$. Note that (19) corresponds to the exponential distribution with hazard rate c . It is also clear that (19) satisfies (16) since $c(x) = c$ independent of x .

Among the increasing Lévy processes, there are three that are worth mentioning at this time. They all lead to the exponential distribution with different hazard rates through different stochastic hazard processes. The first one is the compound Poisson process when $v(dy) = \lambda F(dy)$ where λ is the arrival rate and F is the distribution of the jumps. In this case, the number of jumps of the hazard process H during any interval is finite and

$$P\{H_t \leq x + bt\} = \sum_{n=0}^{+\infty} \frac{e^{-\lambda t} (\lambda t)^n}{n!} F^{*n}(x), \quad (20)$$

where F^{*n} is the n -fold convolution of F . Moreover,

$$E[H_t] = (b + \lambda E[Y]) t, \quad (21)$$

where Y is a generic random variable that represents the jump magnitude of the compound Poisson process with distribution F . The hazard rate is

$$\begin{aligned} c &= b + \lambda \int_0^{+\infty} F(dy) (1 - e^{-y}) \\ &= b + \lambda E[1 - e^{-Y}]. \end{aligned} \quad (22)$$

The second one is the gamma process when $v(dy) = (\beta e^{-\alpha y}/y) dy$ where $\alpha > 0$ is the scale

parameter and $\beta > 0$ is the shape parameter. This leads to a process H that has an infinite number of jumps in any finite interval with the gamma distribution

$$P\{H_t \leq x + bt\} = \int_0^x \left(\frac{\alpha e^{-\alpha y} (\alpha y)^{\beta t - 1}}{\Gamma(\beta t)} \right) dy \quad (23)$$

and finite expectation

$$E[H_t] = (b + (\beta/\alpha))t. \quad (24)$$

The corresponding hazard rate is

$$\begin{aligned} c &= b + \int_0^{+\infty} \left(\frac{\beta e^{-\alpha y}}{y} \right) (1 - e^{-y}) dy \\ &= b + \beta \ln(1 + (1/\alpha)). \end{aligned} \quad (25)$$

Dykstra and Laud [14] proposed an extension of the gamma process to represent the hazard rate process. This representation leads to some desirable properties in terms of Bayesian analysis of reliability models. Gamma processes have been used widely to model random phenomena in many fields. In an earlier work by Moran [15] on storage models, for example, the random flow of water in dams is depicted by a gamma process.

Finally, we have a stable process when $v(dy) = (\alpha\beta y^{-(1+\beta)} / \Gamma(1-\beta)) dy$ where $\alpha > 0$ is the scale parameter and $0 < \beta < 1$ is the shape index. Here, $\Gamma(z) = \int_0^{+\infty} s^{z-1} e^{-s} ds$ is the gamma function. The stable process H also has an infinite number of jumps in any finite interval, but H_t is finite almost surely with the Laplace transform

$$E[e^{-\gamma H_t}] = e^{-(b + \alpha\gamma^\beta)t} \quad (26)$$

and infinite expectation $E[H_t] = +\infty$. The hazard rate now becomes

$$c = b + \int_0^{+\infty} \left(\frac{\alpha\beta y^{-(1+\beta)}}{\Gamma(1-\beta)} \right) (1 - e^{-y}) dy = b + \alpha. \quad (27)$$

Stochastic hazard processes can also be constructed using continuous stochastic processes. However, not all of them are suitable candidates. The increasing Lévy process

in (17), for example, reduces to the deterministic hazard function $H(t) = bt$ when the sample path is restricted to be continuous. A continuous Lévy process, on the other hand, is a Brownian motion which does not have increasing sample paths. We can obtain another stochastic hazard processes through the maximum process

$$H_t = \max_{0 \leq u \leq t} W_u, \quad (28)$$

where W is the standard Brownian motion, or the Wiener process. It is well known that

$$\begin{aligned} P\{H_t \leq x\} &= 1 - 2P\{W_t > x\} \\ &= 1 - 2 \int_{x/\sqrt{t}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \end{aligned} \quad (29)$$

since W_t has the normal distribution with mean 0 and variance t . This distribution implies that the reliability function is

$$\begin{aligned} R(t) &= E[e^{-H_t}] = \int_0^{+\infty} e^{-x} P\{H_t \in dx\} \\ &= \int_0^{+\infty} e^{-x} \left(\frac{2e^{-x^2/2t}}{\sqrt{2\pi t}} \right) dx = 2e^{t/2} \Phi(-\sqrt{t}), \end{aligned} \quad (30)$$

where Φ is the standard normal distribution function. Another increasing process is obtained by the inverse

$$\mathcal{H}_t = \inf\{s \geq 0 : H_s > t\}. \quad (31)$$

This is, in fact, a special case that is already covered in our previous discussion on Lévy processes since it is well known that the process $\mathcal{H}$ in (28) is also a Lévy, in fact stable, process with scale parameter $\alpha = \sqrt{2}$ and shape index $\beta = 0.5$ so that $v(dy) = (1/\sqrt{2\pi y^3}) dy$.

A suitable candidate for the hazard process is an increasing Markov process. For computational tractability, if the Markov process H is defined on some finite state space $E \subseteq \mathbb{R}_+$, then using uniformization results under reasonable assumptions we have the representation

$$H_t = X_{N_t}, \quad (32)$$

where $X = \{X_n; n = 0, 1, \dots\}$ is an increasing Markov chain with an upper triangular transition matrix P , and $N = \{N_t; t \geq 0\}$ is a Poisson process with some rate λ . Now,

$$P\{H_t = j | H_0 = i\} = \sum_{n=0}^{+\infty} \frac{e^{-\lambda t} (\lambda t)^n}{n!} P^n(i, j) \quad (33)$$

and the reliability function is

$$\begin{aligned} R(i, t) &= E \left[e^{-H_t} | H_0 = i \right] \\ &= e^{-\lambda t} \sum_{j \in E} e^{-j} \sum_{n=0}^{+\infty} \frac{(\lambda t)^n}{n!} P^n(i, j) \\ &= e^{-\lambda t} \sum_{j \in E} e^{-j} e^{\lambda t P}(i, j) \end{aligned} \quad (34)$$

conditional on the initial hazard $H_0 = i$. In (34), $e^{\lambda t P}$ is the exponential matrix

$$e^{\lambda t P} = \sum_{n=0}^{+\infty} \frac{s^n}{n!} A^n = \lim_{n \rightarrow +\infty} \left(I + \frac{sA}{n} \right)^n \quad (35)$$

defined for any matrix A and scalar $s \geq 0$. The computation of the matrix exponential (35) can be done in many ways and we refer the interested reader to Moler and Loan [16] for various methods. MATLAB, for example, uses the scaling and squaring method employing Padé approximants.

Another approach to model the stochastic hazard process H is through its stochastic hazard rate process $h_t = dH_t/dt$. An interesting model in finance is used by Ludkovsky and Young [9] where the stochastic hazard is given by the nonnegative diffusion

$$dh_t = \mu(h_t, t)h_t dt + \sigma(t)h_t dW_t, \quad (36)$$

where μ and σ are the drift and volatility terms respectively while W is the Wiener process. The hazard rate is for the lifetimes of individuals or insurers on which mortality contingent claims or derivatives are written. Clearly, stochastic calculus is used to analyze issues related to the pricing of such financial instruments where the risk is realized when mortality happens at the lifetime of the individual before maturity.

MODULATION VIA RANDOM ENVIRONMENTS

Most of the complex stochastic models in operations research and management sciences operate in randomly changing environments. Stochastic and deterministic model parameters are often modulated by this random environment. This is quite evident in reliability applications where deterioration and aging happens intrinsically depending on the state of the environment. However, the idea has far more general applicability in a variety of problems and we will provide a brief discussion in this section before we focus on reliability applications in more detail. We use the term “environment” in a generic sense so that it represents any set of conditions that affect the stochastic structure of the model investigated. The concept of an “environmental” process, in one form or another, has been used in the literature for various purposes. Neveu [17] provides an early reference to paired stochastic processes where the first component is a Markov process while the second one has conditionally independent increments given the first. Ezhov and Skorohod [18] refer to this as a Markov process with homogeneous second component. In a more modern setting, Çınlar [19] introduced Markov additive processes and provided a detailed description on the structure of the additive component. The environment is modeled as a Markov process in all these cases and the additive process represents the stochastic evolution of a quantity of interest.

In reliability models, a complex device generally consists of a large number of components with stochastically dependent lifetimes. In the literature, random environments are used to provide a tractable model of stochastic dependence among the components where the environment is an external process that depicts all physical, structural, operational, and other conditions which affect the deterioration, aging, and failure of the components. Since all components are subjected to the same environmental conditions, their lifetimes are dependent via their common environment. Thus, the environmental process is actually a factor of variation in the

failure structure of the components. These ideas were introduced by Çınlar and Özekici [3] who propose to construct an intrinsic clock which ticks differently in different environments to measure the intrinsic age of the device. The environment is modeled by a semi-Markov jump process and the intrinsic age is represented by the cumulative hazard accumulated in time during the operation of the device in the randomly varying environment. This is a rather stylish choice which envisions that the intrinsic lifetime of any device has an exponential distribution with parameter 1. The intrinsic aging model of Çınlar and Özekici [3] is studied further by Çınlar *et al.* [20] to determine the conditions that lead to associated component lifetimes. The association of the lifetimes of components subjected to a randomly varying environment is discussed by Lefèvre and Milhaud [21]. In a recent article, Singpurwalla [4] provides a review by discussing hazard potentials in reliability modeling.

Random environments are also used in analyzing mission-based or phased-mission systems. In this setting, a device is designed to perform missions consisting of a possibly random sequence of phases with random durations. The environmental process is now the mission process and the failure rate depend on the phase of the mission performed at any time. Most of the literature on phased-mission systems assume that the sequence of the phases are deterministic, and that the durations are either deterministic or exponentially distributed. The main reason is that random durations and sequences make the analysis more complex. There are only a few papers which consider dynamic phase sequences that are modified according to the states of the system. Deterministic phase durations may be realistic for some applications such as aerospace applications, discussed by Mura and Bondavalli [22] where phases are preplanned on the ground. However, Alam and Al-Saggaf [23] state that random phase durations are more realistic in many systems such as real-time control for aircraft and space vehicles in which different sets of computational tasks are executed during different phases of a control process.

Since the exponential distribution is not suitable to model all phase durations due to its long tail behavior, as stated by Mura *et al.* [24], phase durations should be modeled by general random variables. The main concern is the computation of the reliability function. However, there may be several definitions of reliability in these systems depending on the objectives of the mission. The classical definition, of course, is one of *system reliability* which is the probability that the system will function without failure for a given amount of time. Other definitions focus on the phases of the mission as well as the whole of the mission. *Mission reliability* is the probability that a given number of phases or the whole mission will be completed without system failure. Finally, *phase reliability* will refer to the probability that a given critical phase of the mission will be completed without failure by a given time. We refer the reader to Çekyay and Özekici [25,26] for the reliability of semi-Markov and Markov missions with different component failure structures.

There is also an enormous amount of literature on so-called condition-based reliability systems where the condition of the system is observed first before a decision is made on maintenance. Here, the condition of the system is often represented by some stochastic process of an endogenous nature. In other words, the condition is not depicted by some external environmental process, but it is a process that reflects the condition of the system and the time to failure depends on the current condition. The modulation of the reliability model is therefore done by this process. Chen and Trivedi [27] consider the optimal conditional-maintenance problem using semi-Markov decision processes, while Wang *et al.* [28] provide a recent work on condition-based replacement problem with spare parts provisioning. An extended discussion and survey on condition-based maintenance models can be found in Çekyay and Özekici [29].

A similar modulation holds in software reliability systems as well. In this case, the way that the user operates the system, or the so-called operational profile, plays a key role in software reliability assessment. Failure

probabilities of the modules and system reliability, all depend on the random sequence of operations it performs. Musa [30] defines the operational profile of any software as the set of all operations that it is designed to perform with a probability distribution defined on them. The operational profile in this setting provides the random environment for the software system. This approach is used in Özekici and Soyer [31] to analyze probabilistic and statistical issues related to software reliability.

Özekici [32] discusses the use of random environments in complex models in operations research with applications in reliability, inventory, and queueing. In inventory models, the stochastic structure is depicted by the demand and the lead-time processes. Song and Zipkin [33] argue that the demand for the product may be affected by a randomly changing “state-of-the-world,” which we choose to call the “environment” in our exposition. A periodic review model in a random environment with uncertain supply is analyzed in Özekici and Parlar [34]. Erdem and Özekici [35] consider inventory models with a demand process that fluctuates with respect to stochastically changing economic conditions. Queueing models also involve stochastic and deterministic parameters that are subject to variations depending on some environmental factors. The customer arrival rate as well as the service rate are not necessarily constants that remain intact throughout the entire operation of the queueing system. A comprehensive discussion on Markov modulated queueing systems can be found in Prabhu and Zhu [36] where customer arrival and service rates are modulated by a Markov process. More recently, Çanakoğlu and Özekici [37] and Çelikyurt and Özekici [38] applied the idea to various portfolio optimization problems where the correlation among returns of risky assets is formulated by a stochastic market representing the underlying financial, economic, social, and other factors. In the valuation model of Cathcart and Al-Jahel [10] for defaultable bonds, the time to default has a hazard rate that depends on a latent state variable and stochastic interest rate. Interested readers

are referred to Rolski *et al.* [39] for further applications in queueing, insurance, and finance.

INTRINSIC AGING MODELS

Although the discussions in the previous section clearly illustrate the use of random environments in operations research and management sciences, the concept offers more applicability in reliability and maintenance models. It is generally assumed that a device always works in a given fixed environment. The probability law of the deterioration and failure process thus remains intact throughout its useful life. The life distribution and the corresponding failure rate function is taken to be the one obtained through statistical life testing procedures that are usually conducted under ideal laboratory conditions by the manufacturer of the device. Data on lifetimes may also be collected while the device is in operation to estimate the life distribution. In any case, the basic assumption is that the prevailing environmental conditions either do not change in time or, in case they do, they have no effect on the deterioration and failure of the device. Therefore, statistical procedures in estimating the life distribution parameters, and decisions related with replacement and repair are based on the calendar age of the item.

There has been growing interest in recent years in reliability and maintenance models where the main emphasis is placed on the so-called intrinsic age of a device rather than its real age. This is necessitated by the fact that devices often work in varying environments during which they are subject to varying environmental conditions with significant effects on performance. The deterioration and failure process therefore depends on the environment, and it no longer makes much sense to measure the age in real time without taking into consideration the different environments that the device has operated in. There are many examples where this important factor cannot be neglected or overlooked. Consider, for example, the jet engine of

an airplane which is subject to varying atmospheric conditions such as pressure, temperature, humidity, and mechanical vibrations during take-off, cruising, and landing. The changes in these conditions cause the engine to deteriorate, or age, according to a set of rules which may well deviate substantially from the usual one that measures the age in real time irrespective of the environment.

As a matter of fact, the intrinsic age concept is being used routinely in practice in one form or another. In aviation, the calendar age of an airplane since the time it was actually manufactured is not of primary importance in determining maintenance policies. Rather, the number of take-offs and landings, total time spent cruising in fair conditions or turbulence, or total miles flown since manufacturing or since the last major overhaul are more important factors. Another example is a machine or a workstation in a manufacturing system which may be subject to varying loading patterns depending on the production schedule. In this case, the atmospheric conditions do not necessarily change too much in time, and the environment is now represented by varying loading patterns so that, for example, the workstation ages faster when it is overloaded, slower when it is underloaded, and not at all when it is not loaded or kept idle. Therefore, the term *environment* is used in a loose sense here so that it represents any set of conditions that affect the deterioration and aging of the device.

Reliability models in random environments provide a perfect opportunity in order to model stochastic hazard processes. This is because the failure rate at any time is random since the state of the environment that determines it is random. Intrinsic aging in Çinlar and Özekici [3] is described by the simple relationship

$$\frac{dH_t}{dt} = r(Y_t, H_t), \quad (37)$$

where H_t is the intrinsic age or hazard of the system at time t and $Y = \{Y_t; t \geq 0\}$ is an environmental process with state space E that reflects the states of various environmental factors and r is the hazard rate

function. Clearly, $r(i, a)$ is the hazard rate when the hazard is a in environment i .

Intrinsic Aging in a Fixed Environment

Suppose, for now, that the environment remains fixed at some state i so that $Y_t = i$ for all $t \geq 0$. During any environment i , the reliability function

$$R(i, t) = P\{L > t | Y = i\} = e^{-H(i, t)} \quad (38)$$

is known, where $H(i, t)$ is the corresponding deterministic hazard function. Now, (38) allows us to construct an intrinsic clock to measure the intrinsic age of the device at time t as $H_t = H(i, t)$ and the real lifetime is characterized by

$$L = \inf\{s \geq 0; H_s \geq \hat{L}\}, \quad (39)$$

where $\hat{L}$ is a random variable representing the intrinsic lifetime. Moreover, it has an exponential distribution with parameter 1 since

$$P\{L > t | Y = i\} = P\{\hat{L} > H_t | Y = i\} = e^{-H_t}. \quad (40)$$

The intrinsic age at time t is defined as $H_t = H(i, t)$ provided that the environment is in state i throughout $[0, t]$. Then,

$$\tau(i, a) = \inf\{t \geq 0 : H(i, t) > a\} \quad (41)$$

is the amount of operation time it takes to age a brand new item to intrinsic age a in environment i . We now define

$$h(i, a, t) = H(i, \tau(i, a) + t) \quad (42)$$

so that this gives the intrinsic age at time t in state i given that the intrinsic age is a at the beginning. As long as the environment is fixed, aging occurs deterministically according to $h(i, a, t)$ and the item fails as soon as the age exceeds the exponential threshold.

Intrinsic Aging in a Random Environment

Suppose now that the environmental process is not fixed but described as the minimal semi-Markov process associated with a Markov renewal process. Let T_n denote the time of the n th environment change and X_n denote the n th environmental state for $n \geq 0$ with $T_0 \equiv 0$. The main assumption is that the process $(X, T) = \{(X_n, T_n); n \geq 0\}$ is a Markov renewal process on the state space $E \times \mathbb{R}_+$ with some semi-Markov kernel Q . Moreover, $Y = \{Y_t; t \geq 0\}$ is the minimal semi-Markov process associated with (X, T) . More precisely, $Y_t = X_n$ whenever $T_n \leq t < T_{n+1}$. For any $i, j \in E$ and $t \geq 0$,

$$Q(i, j, t) = P[X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i] \quad (43)$$

and it is well known that X is a Markov chain on E with transition matrix $P(i, j) = P[X_{n+1} = j | X_n = i] = Q(i, j, +\infty)$. We further suppose for technical reasons that the Markov renewal process has infinite lifetime so that $\sup_n T_n = +\infty$. Note that if the state space E is finite, this condition is always satisfied and the assumption is required only when E is infinite. We refer the reader to Çinlar [40] for a more rigorous and detailed treatment of Markov renewal processes and theory. The probabilistic structure of T is given by the conditional distributions

$$P\{T_{n+1} - T_n \leq t | X_n = i, X_{n+1} = j\} = \frac{Q(i, j, t)}{P(i, j)} \quad (44)$$

for $P(i, j) > 0$. Moreover, it is easy to see that

$$G_i(t) = P\{T_1 \leq t | X_0 = i\} = \sum_{j \in E} Q(i, j, t) \quad (45)$$

denotes the distribution of the duration of phase i . We let $\bar{G}_i = 1 - G_i$ denote the corresponding survival function for each state i .

An intuitive choice to extend the construction of the intrinsic aging process in this setting is to measure the age by the hazard that accumulates during the operation of the device in the randomly varying environment. If the environment is not fixed, then

the intrinsic aging process $H = \{H_t; t \geq 0\}$ of the item is a continuously increasing stochastic hazard process constructed recursively by

$$H_{T_n+t} = h(X_n, H_{T_n}, t) = H(X_n, \tau(X_n, H_{T_n}) + t) \quad (46)$$

provided that $t \leq T_{n+1} - T_n$ and the item is functioning at time $T_n + t$.

This intrinsic aging model simply combines the hazard functions of the components in the environmental states. Given the deterministic hazard functions $\{h(i, \cdot); i \in E\}$ and a realization of the environmental process Y , the age process H is completely defined by (46) recursively by taking $n = 0, 1, 2, \dots$. If the initial age is $H_0 = a$ and the initial environment is $X_0 = i$ with $T_0 \equiv 0$, then the initial real age of the component is $\tau(i, a)$ and it ages deterministically as $H_s = h(i, a, s)$ for $s \leq T_1$. At some time $T_1 = u$, the environment jumps to state $j \in E$ with some probability $Q(i, j, du)$ and the age is now the hazard given by $H_{u+s} = h(j, H_u, s)$ for $u + s \leq T_2$. The sample path of H is constructed similarly in time as the environmental process evolves so that, in general, if the environment jumps to some state $k \in E$ at the n th jump time $T_n = t$, then the age evolves as $H_{t+s} = H(k, H_t, s)$ so long as $t + s \leq T_{n+1}$.

To simplify our notation, we use $P_{ia}(D) = P(D | X_0 = i, H_0 = a, \hat{L} > a)$ and $E_{ia}[Z] = E[Z | X_0 = i, H_0 = a, \hat{L} > a]$ to denote conditional probabilities and expectations for any event D and random variable Z , respectively. The construction of the stochastic hazard process allows us to write the reliability function as

$$\begin{aligned} R(i, a, t) &= P_{ia}\{L > t\} = P_{ia}\{\hat{L} > H_t\} \\ &= E_{ia}\left[e^{-(H_t - a)}\right] \end{aligned} \quad (47)$$

since the intrinsic lifetime has the exponential distribution with rate 1. The computation of (47), however, is not a simple task in general. In case $T_1 > t$ is given, then the conditional probability is simply

$$P_{ia}\{L > t | T_1 > t\} = e^{-(h(i, a, t) - a)} \quad (48)$$

trivially since $H_t = h(i, a, t)$ whenever $X_0 = i$, $H_0 = a$, and $T_1 > t$. We now consider some interesting cases in decreasing order of complexity and generality.

General Model in a Semi-Markov Environment

We can determine the reliability of the item using a Markov renewal argument for the system reliability function $f(i, a, t) = P_{ia}\{L > t\}$. By conditioning, we can write

$$\begin{aligned} f(i, a, t) &= P_{ia}\{L > t, T_1 > t\} + P_{ia}\{L > t, T_1 \leq t\} \\ &= \bar{G}_i(t) e^{-(h(i, a, t) - a)} + \sum_{j \in E} \int_{[0, t]} Q(i, j, ds) \\ &\quad \times e^{-(h(i, a, s) - a)} f(j, h(i, a, s), t - s) \\ &= \bar{G}_i(t) e^{-(h(i, a, t) - a)} \\ &\quad + \sum_{j \in E} \int_{\mathbb{R}_+ \times [0, t]} \tilde{Q}(i, a; j, db; ds) f(j, b, t - s), \end{aligned} \quad (49)$$

where $\tilde{Q}$ is a semi-Markov kernel on $E \times \mathbb{R}_+$ defined by

$$\tilde{Q}(i, a; j, B; ds) = Q(i, j, ds) 1_B(h(i, a, s)) \quad (50)$$

for all $i, j \in E, a \in \mathbb{R}_+, B \subseteq \mathbb{R}_+$, and $s \geq 0$. Here, $1_B(a)$ is the indicator function in $a \in \mathbb{R}_+$ which is equal to 1, if and only if $a \in B$. Moreover, it is obviously a measure in $B \subseteq \mathbb{R}_+$. The first part of (49) follows from (45) and (48). Thus, (49) is a Markov renewal equation $f = g + \tilde{Q} * f$ where $g(i, a, t) = \bar{G}_i(t) e^{-(h(i, a, t) - a)}$, and it has a unique solution $f = \tilde{R} * g$ so that the reliability function is

$$\begin{aligned} R(i, a, t) &= P_{ia}\{L > t\} \\ &= \sum_{j \in E} \int_{\mathbb{R}_+ \times [0, t]} \tilde{R}(i, a; j, db; ds) \bar{G}_j(t - s) \\ &\quad \times e^{-(h(j, b, t - s) - b)}, \end{aligned} \quad (51)$$

where $\tilde{R} = \sum_{n=0}^{+\infty} \tilde{Q}^n$ is the Markov renewal kernel corresponding to $\tilde{Q}$.

Exponential Model in a Semi-Markov Environment

A special case with some simplification is when the aging in each state happens at a constant rate so that the item fails exponentially. If the hazard rate in state i is λ_i , then $h(i, a, t) = a + \lambda_i t$ and $e^{-(h(i, a, t) - a)} = e^{-\lambda_i t}$ independent of the initial age a . For $f(i, t) = P_i\{L > t\}$, the Markov renewal (49) now simplifies to

$$f(i, t) = \bar{G}_i(t) e^{-\lambda_i t} + \sum_{j \in E} \int_{[0, t]} \tilde{Q}(i, j, ds) f(j, t - s), \quad (52)$$

where $\tilde{Q}(i, j, ds) = Q(i, j, ds) e^{-\lambda_i s}$. The solution of (52) has a simpler structure which gives the reliability function

$$\begin{aligned} R(i, t) &= P_i\{L > t\} \\ &= \sum_{j \in E} \int_{[0, t]} \tilde{R}(i, j, ds) \bar{G}_j(t - s) e^{-\lambda_j(t - s)}. \end{aligned} \quad (53)$$

Exponential Model in a Markovian Environment

A computationally tractable model is obtained when the environmental process Y is a Markov process so that $\bar{G}_i(t) = e^{-\mu_i t}$ and $Q(i, j, t) = P(i, j)(1 - e^{-\mu_i t})$. In this case, by stopping the Markov process Y at the system lifetime L , we obtain another Markov process

$$Z_t = \begin{cases} Y_t & \text{if } t < L \\ \Delta & \text{if } t \geq L \end{cases} \quad (54)$$

with state space $E \cup \{\Delta\}$ where Δ is an absorbing state and its generator is

$$A(i, j) = \begin{cases} -(\mu_i + \lambda_i) & \text{if } j = i \\ \mu_i P(i, j) & \text{if } j \neq i \end{cases} \quad (55)$$

for $i, j \in E$. Moreover, $A(i, \Delta) = \lambda_i$ and $A(\Delta, j) = 0$ for all $i \in E$ and $j \in E \cup \{\Delta\}$. The solution of (53) can now be obtained explicitly by noting that

$$R(i, t) = P_i\{L > t\} = P_i\{Z_t \in E\} = \sum_{j \in E} e^{tA}(i, j), \quad (56)$$

where e^{tA} is the matrix exponential defined by (35). This follows by observing that the transition function of a Markov process is given by the matrix exponential of its generator. Çekyay and Özekici [26] provide theoretical derivations and numerical illustrations on the reliability, mean time to failure, and availability of Markovian systems discussed in this section.

REFERENCES

1. Abdel-Hameed MS, Çınlar E, Quinn J, editors. Reliability theory and models, Notes and reports in computer science and applied mathematics. Orlando (FL): Academic Press; 1984.
2. Aven T, Jensen U. Stochastic models in reliability. Berlin: Springer; 1999.
3. Çınlar E, Özekici S. Reliability of complex devices in random environments. Probab Eng Inf Sci 1987;1:97–115.
4. Singpurwalla ND. The hazard potential: introduction and overview. J Am Stat Assoc 2006;101:1705–1717.
5. Gaver DP. Random hazard in reliability problems. Technometrics 1963;5:211–226.
6. Arjas E. The failure and hazard process in multivariate reliability systems. Math Oper Res 1981;6:551–562.
7. Lindley DV, Singpurwalla ND. Multivariate distributions for the lifetimes of components of a system sharing a common environment. J Appl Probab 1986;23:418–431.
8. Kebir Y. On hazard rate processes. Nav Res Log Q 1991;38:865–867.
9. Ludkovsky M, Young VR. Indifference pricing of pure endowments and life annuities under stochastic hazard and interest rates. Insur Math Econ 2008;42:14–30.
10. Cathcart L, Al-Jahel L. Pricing defaultable bonds: a middle-way approach between structural and reduced-form models. Quant Finance 2006;6:243–253.
11. Çınlar E. Markov and semimarkov models of deterioration. In: Abdel-Hameed MS, Çınlar E, Quinn J, editors. Reliability theory and models. Orlando (FL): Academic Press; 1984. pp. 3–41.
12. Barndorff-Nielsen OE, Mikosch T, Resnick SI, editors. Lévy processes: theory and applications. Boston (MA): Birkhäuser; 2001.
13. Gjessing HK, Aalen OO, Hjort NL. Frailty models based on Lévy processes. Adv Appl Probab 2003;35:532–550.
14. Dykstra RL, Laud P. A Bayesian nonparametric approach to reliability. Ann Stat 1981;9:356–367.
15. Moran PAP. The theory of storage. New York: John Wiley & Sons, Inc.; 1959.
16. Moler C, van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev 2003; 45(1):3–49.
17. Neveu J. Une généralisation des processus á accroissements positifs indépendants. Abh aus den Mathematischen Semin der Univ Hamburg 1961;25:36–61.
18. Ezhov II, Skorohod AV. Markov processes with homogeneous second component: I. The theory Probab Appl 1969;14:3–14.
19. Çınlar E. Markov additive processes: I. Z Wahrscheinlichkeitstheorie verw Geb 1972; 24:85–93.
20. Çınlar E, Shaked M, Shanthikumar JG. On lifetimes influenced by a common environment. Stoch Processes Appl 1989;33:347–359.
21. Lefèvre C, Milhaud X. On the association of the lifelenghts of components subjected to a stochastic environment. Adv Appl Probab 1990;22:961–964.
22. Mura I, Bondavalli A. Markov regenerative stochastic Petri nets to model and evaluate phased mission systems dependability. IEEE Trans Comput 2001;50:1337–1351.
23. Alam M, Al-Saggaf UM. Quantitative reliability evaluation of repairable phased-mission systems using Markov approach. IEEE Trans Reliab 1986;35:498–503.
24. Mura I, Bondavalli A, Zang X, *et al.* Dependability modeling and evaluation of phased mission systems: a DSPN approach. Proceedings of IFIP International Conference on Dependable Computing for Critical Applications (DCCA-7) San Jose, California: . 1999. pp. 299–318.
25. Çekyay B, Özekici S. Reliability of semi-Markov missions. Technical report. Istanbul, Turkey: Department of Industrial Engineering, Koç University; 2008.
26. Çekyay B, Özekici S. Reliability, MTTF, and availability of systems with Markovian missions and aging. Technical report. Istanbul, Turkey: Department of Industrial Engineering, Koç University; 2009.
27. Chen D, Trivedi KS. Optimization for condition-based maintenance with semi-Markov decision process. Reliab Eng Syst Saf 2005;90:25–29.

28. Wang L, Chu J, Mao W. A condition-based replacement and spare provisioning policy for deteriorating systems with uncertain deterioration to failure. *Eur J Oper Res* 2009;194:184–205.
29. Çekyay B, Özekici S. Condition-Based Maintenance under Markovian deterioration. Technical report. Istanbul, Turkey: Department of Industrial Engineering, Koç University; 2009.
30. Musa JD. Operational profiles in software reliability engineering. *IEEE Softw* 1993; 10:14–32.
31. Özekici S, Soyer R. Reliability of software with an operational profile. *Eur J Oper Res* 2003;149:459–474.
32. Özekici S. Complex systems in random environments. In: Özekici S, editor. Volume F154, Reliability and maintenance of complex systems, NATO ASI Series. Berlin: Springer; 1996. pp. 137–157.
33. Song J-S, Zipkin P. Inventory control in a fluctuating demand environment. *Oper Res* 1993;41:351–370.
34. Özekici S, Parlar M. Inventory models with unreliable suppliers in a random environment. *Ann Oper Res* 1999;91:123–136.
35. Erdem A, Özekici S. Inventory models with random yield in a random environment. *Int J Prod Econ* 2002;78:239–253.
36. Prabhu NU, Zhu Y. Markov-modulated queueing systems. *Queueing Syst* 1989;5:215–246.
37. Çanakoğlu E, Özekici S. Portfolio selection in stochastic markets with exponential utility functions. *Ann Oper Res* 2009;166:281–297.
38. Çelikyurt U, Özekici S. Multiperiod portfolio optimization models in stochastic markets using the mean-variance approach. *Eur J Oper Res* 2007;179:186–202.
39. Rolski T, Schmidli H, Schmidt V, *et al.* Stochastic processes for insurance and finance. Chichester: John Wiley & Sons, Ltd.; 1999.
40. Çinlar E. Markov renewal theory. *Adv Appl Probab* 1969;1:123–187.

STOCHASTIC MIXED-INTEGER PROGRAMMING ALGORITHMS: BEYOND BENDERS' DECOMPOSITION

SUVRAJEET SEN
Data Driven Decisions Lab,
ISE Department, Ohio
State University,
Columbus, Ohio

INTRODUCTION

As discussed elsewhere in this volume, stochastic programming models provide an entrée into decision-making problems in which data uncertainty is best viewed through the lens of random variables or stochastic processes. There are numerous applications in which the need to address strategic issues in decision making leads to stochastic mixed-integer programming (SMIP) models. The breadth of examples considered in the literature include applications in inventory management [1], manufacturing capacity planning [2], process engineering [3], production scheduling [4,5], server location [6], supply chain planning [7], vehicle routing [8], and many more. These models arise in situations where the “lumpy” nature of decisions are coupled with forecast uncertainty. Failure to accommodate uncertainty can lead to plans that are “brittle,” in the sense that minor excursions from the norm may bring such systems to a grinding halt. Unfortunately, the study of SMIP problems is relatively recent, and many open issues remain. It is our hope that this article will kindle an interest in SMIP, and give readers the starting point to explore new properties and algorithms in this area.

While deterministic mixed-integer linear programming (MILP) methodology has made significant strides in the past several decades (see other articles in this volume), it is worth mentioning that these methods are often

inadequate for the direct solution of SMIP problems due to the explosion in the size of the deterministic equivalent MILP. The key to solving SMIP problems is the development of new decomposition algorithms for specially structured MILP problems. In this article, we will cover SMIP algorithms that may be looked upon as being in the same genre as Benders’ decomposition [9]. However, the reader will soon recognize that the mathematical and computational differences are significant.

SMIP problems involve models in which decisions and observations form an interleaving sequence, with decisions in earlier stages affecting future choices. As one might expect, the key to designing effective algorithms for solving such problems lies in the ability to assess the consequences of early decisions on future costs, and in this sense, value function approximations are important. Hence these models and algorithms share some affinity with dynamic programming [10].

Problem Setting and Classification

In this article, we will restrict ourselves to two-stage models, leaving the multistage SMIPs to more extensive surveys [11,12]. Moreover, to keep our focus on issues arising from integer restrictions, we will discuss the most common SMIP model, which uses “Expectation” minimization as a criterion. We do note, however, that the ability to include inequality constraints within an optimization model allows one to model more general risk measures without significant changes to the underlying structure of the problem.

To state the problem of interest, let $\tilde{\omega}$ denote a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and X denote a constraint set specified by linear equations and inequalities. We use $\mathbf{X}$ to denote either the set of binary vectors $\mathcal{B}$, or integer vectors $\mathcal{Z}$ or even mixed-integer vectors $\mathcal{M} = \{x | x \geq 0, x_j \text{ integer}, j \in J_1\}$, where J_1 is a given index set consisting of some or all of the first-stage

variables $x \in \mathbb{R}^{n_1}$. Then the two-stage SMIP model may be stated as given in Equations (1–2c).

$$\text{Min}_{x \in X \cap \mathbf{X}} c^\top x + E[h(x, \tilde{\omega})], \quad (1)$$

where

$$h(x, \omega) = \text{Min } g(\omega)^\top y \quad (2a)$$

$$W(\omega)y \geq r(\omega) - T(\omega)x \quad (2b)$$

$$y \geq 0, y_j \text{ integer}, j \in J_2, \quad (2c)$$

where J_2 is an index set that may include some or all the variables listed in $y \in \mathbb{R}^{n_2}$. Given x , the above notation implies that the random variable $\tilde{\omega}$ induces another random variable $h(x, \tilde{\omega})$ whose outcomes are denoted by $h(x, \omega)$. Throughout this article, we will assume that all realizations $W(\omega)$ are rational matrices of size $m_2 \times n_2$. Whenever J_2 is nonempty, and $|J_2| \neq n_2$, Equation (2) is said to provide a model with mixed-integer recourse. We will also assume that the random variables have finite support, so that the expectation in Equation (1) reduces to a summation. Although Equation (2) is stated as though the random variable influences all data, most applications lead to models that lead to only some data uncertainty, which in turn leads to certain specialized models.

The degree of difficulty associated with SMIP varies significantly with the nature (binary or integer) of the “lumpy” decisions, as well as, the stages in which they arise. Accordingly, we begin by classifying SMIP problems according to the kinds of decisions that appear in different stages of the problem. Let

$B \equiv \{\text{The set of stages in which the model has Binary decisions}\},$

$C \equiv \{\text{The set of stages in which the model has Continuous decisions}\},$

$D \equiv \{\text{The set of stages in which the model has Discrete decisions}\}.$

For stochastic linear programs (SLPs), the sets B and D are empty, whereas, SMIP problems are devoted to instances in which these sets are nonempty.

The first observation regarding SMIP models relates to instances that satisfy $B = D = \{1\}$. Such instances can be solved by using the original version of Benders' decomposition. However, for instances in which $2 \in B \cup D$, Benders' decomposition in its original form is no longer applicable, and such SMIP instances require the use of more recently developed mathematical and computational machinery. In the interest of space, we will restrict our attention to algorithms that are applicable to two-stage SMIP problems for which $B = \{1, 2\}$, $C = \{2\}$, and either $D = \emptyset$ or $D = \{2\}$. While the case of $B = D = \{1\}$, and $C = \{1, 2\}$ is adequately covered through standard Benders' decomposition, we mention Norkin *et al.* [13], which discusses an approach in which the second-stage value functions are approximated using sampling methods within a branch-and-bound (B&B) search. This approach is particularly relevant for instances that require the second stage value function to be approximated by sampling [14,15].

MATHEMATICAL PROPERTIES AND COMPUTATIONAL COMPLEXITY

Because the value function of integer programming problems are notorious for their lack of convexity, and even the lack of continuity, it is interesting to identify circumstances under which one may obtain continuity of the expected recourse function. Schultz has shown that the expected recourse function satisfies the following [16].

Proposition 1. *Assume that Equation (2) has randomness only in $r(\tilde{\omega})$, and let the probability space of this random variable, denoted $(\Omega, \mathcal{A}, \mathcal{P})$, be such that $\mathcal{P}$ is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^{m_2}$. Moreover, suppose that the following hold.*

- (a) *(Dual Feasibility). There exists $\theta \geq 0$ such that $W^\top \theta \leq g$.*
- (b) *(Complete Recourse). For any choice of $r(\omega)$, the MILP feasible set is nonempty.*
- (c) *(Finite Expectation). $E[\|r(\tilde{\omega})\|] < \infty$.*

Then, the expected recourse function is continuous.

It so happens that the requirement of absolute continuity (with respect to the Lebesgue measure in $\Re^{m_2}$) is rather restrictive from the point of view of practical optimization models. In order to appreciate this, observe that many practical LP/IP models have constraints that are entirely deterministic; for example, flow conservation/balance constraints often have no randomness in them. Formulations of this type (where some constraints are completely deterministic) fail to satisfy the requirement that the measure $\mathcal{P}$ is absolutely continuous with respect to the Lebesgue measure in $\Re^{m_2}$.

Intuitively, one recognizes that SMIP problems are at least as hard as MILP problems because any instance of the latter can be reduced to an SMIP. Formal analysis of the complexity of both stochastic linear as well as stochastic combinatorial optimization problems (i.e., $B = \{2\}$) rely on reduction of graph reliability problems to stochastic programming [17]. The reduction used in Dyer and Stougie [17] allows one to conclude that these optimization problems are $\#P$ -hard, where the class $\#P$ denotes counting problems for which membership to a set of items to be counted can be decided in polynomial time. Indeed, Dyer and Stougie [17] actually shows that the evaluation of the expected recourse function is itself NP-hard for a given first-stage vector x . Nevertheless, for certain special structures (assuming certain similarities between the first- and second-stage variables and objective functions) fully polynomial time approximations can be obtained for two-stage stochastic linear programming problems [18]. These ideas have also been applied to SMIP by using rounding to get integer solutions. However, it is an open question whether such methods can be extended to standard applications of SMIP where the stages ordinarily represent different aspects of strategic (e.g., location) planning and tactical (e.g., assignment) responses.

A GENRE OF SMIP ALGORITHMS

One of the more important requirements for stochastic programming algorithms is their ability to accommodate a large number of scenarios so that realistic applications can be addressed. For SLP, this observation has led to novel decomposition algorithms, some of which are deterministic procedures [19], while some others are stochastic [14,20]. On the basis of current evidence for SMIP problems, it appears that decomposition approaches look most promising [21,22].

A Unifying Framework for SMIP Benders' Decomposition

We begin by reiterating that our focus will be on cases for which $2 \in B \cup D$, so that the standard version of Benders' decomposition is not applicable because the expected recourse function (MILP value function) is nonconvex. Nevertheless, a conceptual extension of Benders' decomposition [23] has been designed for SMIP. While this scheme does not lead to a practical algorithm, it helps to highlight some of the challenges raised by these problems.

In order to maintain simplicity in this presentation, we assume that the second-stage problem satisfies the complete recourse property. Assuming that the random variable modeling uncertainty is discrete, with finite support ($\Omega = \{\omega^1, \dots, \omega^N\}$), a two-stage SMIP may be stated as

$$\text{Min } c^\top x + \sum_{\omega \in \Omega} p(\omega) g(\omega)^\top y(\omega) \quad (3a)$$

$$\text{s.t. } Ax \geq b \quad (3b)$$

$$T(\omega)x + Wy(\omega) \geq r(\omega), \forall \omega \in \Omega \quad (3c)$$

$$(x, \{y(\omega)\}_{\omega \in \Omega}) \geq 0, x_j \text{ integer}, j \in J_1, \\ \text{and } y_j(\omega) \text{ integer}, \forall j \in J_2. \quad (3d)$$

Despite the fact that there are several assumptions underlying Equation (3), it is somewhat general from the IP point of view since both the first- and second-stage variables allow general mixed-integer problems.

Suppose now that we wish to apply a resource directive decomposition method (i.e.,

Benders' decomposition). At iteration k of such a method, we solve one second-stage subproblem for each outcome ω , and assuming that there is an appropriate solution method for the second stage, we can obtain a nondecreasing price function $h_k^\omega(\cdot)$ for each outcome $\omega \in \Omega$ [24]. Consequently, we obtain a "cut" of the form

$$\eta \geq \bar{h}_k(x) := \sum_{\omega \in \Omega} p(\omega) h_k^\omega(r(\omega) - T(\omega)x). \quad (4)$$

Hence, as iterations proceed, one obtains a sequence of relaxed master programs of the following form.

$$\text{Min } c^\top x + \eta \quad (5a)$$

$$\text{s.t. } Ax \geq b \quad (5b)$$

$$\eta \geq \bar{h}_t(x), \quad t = 1, \dots, k \quad (5c)$$

$$x \geq 0, x_j \text{ integer}, j \in J_1. \quad (5d)$$

As with Benders' (or L-shaped) decomposition, each iteration augments the first-stage approximation with one additional collection of price functions as shown in Equation (5c). The rest of the procedure also mimics Benders' decomposition in that the sequence of objective values of Equation (5) generate an increasing sequence of lower bounds, whereas, the subproblems at each iteration provide values used to compute an upper bound. The method stops when the upper and lower bounds are sufficiently close. Provided that the second-stage problems are solved using B&B, it is not difficult to show that the method must terminate in finitely many steps. Of course, finiteness also presumes that Equation (5) can be solved in finite time.

Unfortunately, there are several algorithmic and computational questions that this conceptual method leaves unanswered. From an algorithmic point of view, there is no recommendation regarding the choice of price functions, and from a computational point of view, it is not clear that such a master program, which is nonconvex, can be solved to global optimality in reasonable time. Yet, the general structure of the scheme is appealing because it generalizes traditional

Benders' decomposition. The remaining algorithms presented in this article will differ from the one above in the manner in which Equation (4) is approximated, thus simplifying the solution of Equation (5).

SMIP Benders' Decomposition with Combinatorial Approximations

Unlike the price functions used in the previous section, Laporte and Louveaux [25] devised a method that uses combinatorial approximations. The algorithmic setting within which the inequalities of Laporte and Louveaux [25] are used follows the basic outline of Benders' decomposition. That is, at each iteration k , we solve one master program, and as many subproblems as there are outcomes of the random variable.

For this section, we allow $B = \{1, 2\}$, $C = \{2\}$, and $D = \{2\}$ in Equations (1) and (2). The approximation presented here can be applied to a wide class of expected recourse functions, so long as the first-stage decisions are binary, and we have a known lower bound on the expected recourse function. Interestingly, despite the nonconvexity of value functions of second-stage problems, the valid inequality provided by Laporte and Louveaux [25] is affine.

At iteration k , let the first-stage decision x^k be given, and let

$$I_k = \{i | x_i^k = 1\}, \quad Z_k = \{1, \dots, n_1\} - I_k.$$

Next define the affine function

$$\delta_k(x) = |I_k| - \left[\sum_{i \in I_k} x_i - \sum_{i \in Z_k} x_i \right]. \quad (6)$$

It is easily seen that when $x = x^k$ (assumed binary), $\delta_k(x) = 0$; whereas, for all other binary vectors $x \neq x^k$, at least one of the components must switch "states." Hence, for $x \neq x^k$, we have

$$\left[\sum_{i \in I_k} x_i - \sum_{i \in Z_k} x_i \right] \leq |I_k| - 1, \text{ i.e., } \delta_k(x) \geq 1. \quad (7a)$$

Next suppose that a lower bound on the expected recourse function, denoted $\bar{L}$, is

available. Let $\bar{h}_k$ denote the value of the expected recourse function for a given x^k . If $\bar{h}_k = \infty$ (i.e., the second-stage is infeasible), then Equation (7a) can be used to delete x^k . On the other hand, if $\bar{h}_k$ is finite, then the following inequality is valid:

$$\eta \geq \bar{h}_k(x) := \bar{h}_k - \delta_k(x)[\bar{h}_k - \bar{L}]. \quad (7b)$$

Here the term *valid* implies that the inequality does not delete any point of the epigraph of the expected recourse function. This is the “optimality” cut of Laporte and Louveaux [25], and is used in place of Equation (4). Finally, we should mention that such cuts can be strengthened using any special structure that can be exploited [22].

SMIP Benders' Decomposition with Disjunctive Programming: Set Convexification

This algorithm combines disjunctive programming ideas within the framework of Benders' decomposition, and is therefore referred to as *disjunctive decomposition*. In the previous two algorithms, each iteration requires the solution of multiple NP-hard subproblems (one for each scenario). The algorithm described in this section requires the solution of several LP subproblems, and as a result, one has greater control on the amount of computational effort in any given iteration [26]. This algorithm imposes the following structure on Equations (1) and (2): a fixed recourse matrix (W), $B = \{1, 2\}$, $C = \{2\}$, $D = \{2\}$. The methodology here is one of sequential convexification of the mixed-integer recourse problem. Such an approach (sequential convexification) provides stronger relaxations in later iterations, thus easing the computational burden. Note that cuts for this class of methods can also be derived using the RLT framework [27].

We start this development with the assumption that by using appropriately penalized continuous variables, the subproblem remains feasible for any restriction of the integer variables $y_j, j \in J_2$. Let x^k be given, and suppose that matrices $W_k, T_k(\omega)$, and $r_k(\omega)$ are given. Initially (i.e., $k = 1$) these matrices are simply $W, T(\omega)$, and $r(\omega)$, and recall that in our notation, we include

the constraints $-y_j \geq -1, j \in J_2$ explicitly in $Wy \geq r(\omega) - T(\omega)x$. (Similarly, the constraint $-x \geq -1$ is also included in the constraints $x \in X$.) During the course of solving the 0-1 MILP subproblem for outcome ω , suppose that we solve the following LP relaxation:

$$\text{Min } g^\top y \quad (8a)$$

$$\text{s.t. } W_k y \geq r_k(\omega) - T_k(\omega)x \quad (8b)$$

$$y \in \mathbb{R}_+^{n_2}. \quad (8c)$$

Whenever the solution to this problem is fractional, we will be able to derive a valid inequality that can be used in all subsequent iterations. Let $y^k(\omega)$ denote a solution to Equation (8), and let $j(k)$ denote an index $j \in J_2$ for which $y_j^k(\omega)$ is noninteger for one or more $\omega \in \Omega$. To eliminate this noninteger solution, a disjunction of the following form will be used:

$$\mathcal{S}_k(x^k, \omega) = \mathcal{S}_{0,j(k)}(x^k, \omega) \cup \mathcal{S}_{1,j(k)}(x^k, \omega),$$

where

$$\mathcal{S}_{0,j(k)}(x^k, \omega) = \{y \in \mathbb{R}_+^{n_2} \mid W_k y \geq r_k(\omega) - T_k(\omega)x^k, -y_{j(k)} \geq 0\}, \quad (9a)$$

$$\mathcal{S}_{1,j(k)}(x^k, \omega) = \{y \in \mathbb{R}_+^{n_2} \mid W_k y \geq r_k(\omega) - T_k(\omega)x^k, y_{j(k)} \geq 1\}. \quad (9b)$$

The index $j(k)$ is referred to as the *disjunction variable* for iteration k . This is precisely the disjunction used in lift-and-project cuts [28]. However for SMIP problems, the objective function used for cut generation will be different (see Remark 1).

Let $\lambda_{0,1}$ denote the vector of multipliers associated with the rows of W_k in Equation (9a), and $\lambda_{0,2}$ denote the scalar multiplier associated with the variable $y_{j(k)}$ in Equation (9a). Let $\lambda_{1,1}$ and $\lambda_{1,2}$ be similarly defined for Equation (9b). Then the theory of disjunctive programming implies that if $(\pi, \pi_0(x^k, \omega), \omega \in \Omega)$ satisfy Equation (10), then $\pi^\top y \geq \pi_0(x^k, \omega)$ is a valid inequality for $\mathcal{S}_k(x^k, \omega)$. To simplify the notation in Equation (10) we drop the dependence on x^k , and will reintroduce it

and recognize the dependence on x following Remark 1.

$$\pi_j \geq \lambda_{0,1}^\top W_{jk} - I_j^k \lambda_{0,2} \forall j, \quad (10a)$$

$$\pi_j \geq \lambda_{1,1}^\top W_{jk} + I_j^k \lambda_{1,2} \forall j, \quad (10b)$$

$$\begin{aligned} \pi_0(\omega) &\leq \lambda_{0,1}^\top (r_k(\omega) - T_k(\omega)x^k), \\ \forall \omega &\in \Omega, \end{aligned} \quad (10c)$$

$$\begin{aligned} \pi_0(\omega) &\leq \lambda_{1,1}^\top (r_k(\omega) - T_k(\omega)x^k) + \lambda_{1,2} \\ \forall \omega &\in \Omega, \end{aligned} \quad (10d)$$

$$\begin{aligned} -1 &\leq \pi_j \leq 1, \forall j, -1 \leq \pi_0(\omega) \leq 1, \\ \forall \omega &\in \Omega, \end{aligned} \quad (10e)$$

$$\lambda_{0,1}, \lambda_{0,2}, \lambda_{1,1}, \lambda_{1,2} \geq 0, \quad (10f)$$

where

$$I_j^k = \begin{cases} 0, & \text{if } j \neq j(k) \\ 1, & \text{otherwise.} \end{cases}$$

Remark 1. Several objectives have been proposed in the disjunctive programming literature for choosing cut coefficients [29]. One possibility for SMIP problems is to maximize the expected value of the depth of cut: $E[\pi_0(\omega)] - E[y^k(\omega)]\pi$. We should note that the optimal objective value of the resulting LP can be zero, which implies that the inequality generated by the LP does not delete some of the fractional points $y^k(\omega)$, $\omega \in \Omega_k$. Here Ω_k denotes those $\omega \in \Omega$ for which $y^k(\omega)$ does not satisfy mixed-integer feasibility. So long as the cut deletes a fractional $y^k(\omega)$ for some ω , we may proceed with the algorithm. However, if we obtain an inequality such that $(\pi^k)^\top y^k(\omega) \geq \pi_0^k(\omega)$, for all $\omega \in \Omega_k$, then one such outcome should be removed from the expectation operation $E[y^k(\tilde{\omega})]$, and this vector should be replaced by a conditional expectation over the remaining vectors $y^k(\omega)$. Since the rest of the LP remains unaltered, the reoptimization should be carried out using a “warm start.” Other objective functions can also be used for the cut-generation process. For instance, we could maximize the function $\text{Min}_{\omega \in \Omega} \pi_0(\omega) - y^k(\omega)^\top \pi$.

For vectors $x \neq x^k$, the cut may need to be modified in order to maintain its validity. It can be shown that for any other x , one only

needs to modify the right-hand side scalar π_0 ; in other words, the vector π^k provides valid cut coefficients so long as the recourse matrix is fixed. This result is known as the *Common Cut Coefficients (C^3) Theorem* [26].

The LP used to obtain the common cut coefficients is known as the C^3 LP, and its solution $(\pi^k)^\top$ is appended to W_k in order to obtain W_{k+1} . In order to be able to use these coefficients in subsequent iterations, we will also calculate a new row to append to $T_k(\omega)$, and $r_k(\omega)$ respectively. These new rows will be obtained by solving some other LPs, which we will refer to as RHS-LPs. These calculations are summarized next.

Let $\lambda_{0,1}^k, \lambda_{0,2}^k, \lambda_{1,1}^k, \lambda_{1,2}^k \geq 0$ denote the values obtained from C^3 LP in iteration k . Since these multipliers are nonnegative, the C^3 Theorem [26] allows us to use these multipliers for any choice of (x, ω) . Hence by using these multipliers, the right-hand side function $\pi_0(x, \omega)$ can be written as

$$\begin{aligned} \pi_0(x, \omega) &= \text{Min}\{(\lambda_{0,1}^k)^\top r_k(\omega) \\ &\quad - (\lambda_{0,1}^k)^\top T_k(\omega)x, (\lambda_{1,1}^k)^\top r_k(\omega) \\ &\quad + \lambda_{1,2}^k - (\lambda_{1,1}^k)^\top T_k(\omega)x\}. \end{aligned}$$

For notational convenience, we put

$$\begin{aligned} \bar{v}_0(\omega) &= (\lambda_{0,1}^k)^\top r_k(\omega), \\ \bar{v}_1(\omega) &= (\lambda_{1,1}^k)^\top r_k(\omega) + \lambda_{1,2}^k, \end{aligned}$$

and

$$[\bar{\gamma}_\ell(\omega)]^\top = (\lambda_{\ell,1}^k)^\top T_k(\omega), \quad \ell \in \{0, 1\},$$

so that

$$\begin{aligned} \pi_0(x, \omega) &= \text{Min}\{\bar{v}_0(\omega) - [\bar{\gamma}_0(\omega)]^\top x, \bar{v}_1(\omega) \\ &\quad - [\bar{\gamma}_1(\omega)]^\top x\}. \end{aligned}$$

Being the minimum of two affine functions, the epigraph of $\pi_0(x, \omega)$ can be represented as the union of the two half-spaces. Hence the epigraph of $\pi_0(x, \omega)$, restricted to the set X will be denoted as $\Pi_X(\omega)$, and represented as

$$\Pi_X(\omega) = \cup_{\ell \in L} \mathcal{E}_\ell(\omega),$$

where $L = \{0, 1\}$ and

$$\mathcal{E}_\ell(\omega) = \{(\eta, x) \mid \eta \geq \bar{v}_\ell(\omega) - \bar{y}_\ell(\omega)^\top x, \\ x \in X\}.$$

Here $X = \{x \in \mathbb{R}^{n_1} \mid Ax \geq b, x \geq 0\}$, and we assume that the inequality $-x \geq -1$ is included in the constraints $Ax \geq b$. It follows that the closure of the convex hull of $\Pi_X(\omega)$ provides the appropriate convexification of $\pi_0(x, \omega)$. This process is carried out by solving a linear program patterned after a cut-generation LP of disjunctive programming [27]. However, to avoid confusion between cut generation and this particular scheme of convexification, we refer to the resulting LP as the RHS-LP. Unlike problem C^3 -LP however, this LP is solved for every outcome ω .

Given the limitations of this article, we refer the reader to Sen and Higle [26] for further details, including a specific formulation of RHS-LP, and a proof of finite convergence. However, it is important to emphasize one detail, which has a significant impact on algorithmic efficiency. The observation is that C^3 -LP is itself an SLP, and solving this problem using decomposition provides very encouraging computational results [21]. Some of these computational results will be summarized subsequently. The main iterative steps are summarized below. The stopping rule is the same as in standard Benders' decomposition.

Disjunctive Decomposition.

0. *Initialize.* Assume that x^1 is given, and $k \leftarrow 1$.
1. *Sequential Convexification in Stage 2.*
 - a. *(Solve the LP relaxation for all ω).* Given x^k , solve the LP relaxation of each subproblem, $\omega \in \Omega$.
 - b. *(Solve C^3 -LP).* Use Equation (10) to generate a cut [26]. Append the solution $(\pi^k)^\top$ to the matrix W_k to obtain W_{k+1} .
 - c. *(Solve RHS-LP).* Solve one RHS-LP per $\omega \in \Omega$, and derive $r_{k+1}(\omega)$, $T_{k+1}(\omega)$.
 - d. *(Solve a cut-enhanced LP relaxation for all ω).* Using $x = x^k$, and the

updated matrices $W_{k+1}, r_{k+1}(\omega)$, $T_{k+1}(\omega)$, solve an LP relaxation for each $\omega \in \Omega$.

2. *Derive a Benders' Cut for Stage 1.* Using the dual variables obtained from the solution of the LP relaxations in step 1(d), derive a Benders' cut (also L-shaped cut [30]) of the form $\eta \geq \bar{h}_k(x) := \alpha + \beta^\top x$. This affine function replaces the nonconvex function 4, and is appended to the master program whose objective function retains piecewise linearity and convexity of standard Benders' decomposition.
3. *Solve the Master Program.* Obtain x^{k+1} by solving the Benders' master program; $k \leftarrow k + 1$ and repeat from 1.

SMIP Benders' Decomposition with Disjunctive Programming: Value Function Convexification

This method allows disjunctive decomposition to be used with a branch-and-cut framework, and is abbreviated as D^2 -BAC [31]. While the subtitle above emphasizes value function convexification, we look upon this as an added feature which will complement set convexification due to D^2 . The D^2 -BAC method is designed to be applicable for instances in which $B = \{1, 2\}$, $C = \{2\}$, and $D = \{2\}$. Because D^2 -BAC uses a B&B search in the second stage, it allows an "approximate solve" of the second-stage problem by truncating the number of nodes explored in the B&B search tree. Moreover, note that a B&B search tree that is obtained for one scenario (ω) can be used to approximate the second-stage MILP associated with another scenario. Let $Q(\omega)$ denote the set of nodes of the B&B tree that have been explored for the subproblem associated with scenario ω . For any node $q \in Q(\omega)$, let $z_{q\ell}(\omega)$ and $z_{qu}(\omega)$ denote vectors whose elements are used to define lower and upper bounds, respectively, on the second-stage (integer) variables. In some cases, an element of z_{qu} may be $+\infty$, and in this case, the associated constraint may be ignored, implying that the associated dual multiplier is fixed at 0.

In any event, suppose that for a given $x \in X$, we wish to construct a lower bounding

approximation of the value function of the second-stage problem. The LP relaxation for node q may be written as

$$\begin{aligned} h_q^\omega(x) &= \text{Min } g^\top y \\ Wy &\geq r(\omega) - T(\omega)x \\ y &\geq z_{q\ell}(\omega), \quad -y \geq -z_{qu}(\omega). \end{aligned} \quad (11)$$

The LP in Equation (11) can be used to derive a lower bounding function for node q , that is, we obtain an inequality of the form

$$h_q^\omega(x) \geq \alpha_q(\omega) + \beta_q(\omega)^\top x.$$

Let $\mathcal{E}_q(\omega)$ denote the epigraph of the above affine approximation. Since the optimal value of the second-stage problem for a particular outcome (ω) is at least as large as the minimal lower bound, we obtain a lower bound on the subproblem value function for an outcome ω by defining a function

$$h^\omega(x) := \text{Min}_{q \in Q(\omega)} \alpha_q(\omega) + \beta_q(\omega)^\top x. \quad (12)$$

Since Equation (12) is a nonconvex function, using it directly in stage one would still result in a Benders' master program that is nonconvex. Nevertheless we can convexify Equation (12), thus leading to a more convenient master program. We, therefore, use disjunctive programming to derive a facet of the set

$$\mathcal{E}(\omega) := \cup_{q \in Q(\omega)} \text{clconv } \mathcal{E}_q(\omega).$$

For each $\omega \in \Omega$, a facet of $\text{clconv } \mathcal{E}(\omega)$ provides an affine lower bounding approximation of Equation (12). Taking expectations then yields an affine approximation, which replaces Equation (4), and once again, the objective function of Equation (5) remains piecewise linear and convex.

Convergence properties of the above process are discussed in Sen and Sherali [31] and computational results appear in Yuan and Sen [21]. Table 1 summarizes computational results that have appeared in the latter paper. These problem instances are stochastic server location problems (SSLPs) in which servers have to be located to meet demands, which may or may not be realized

Table 1. Comparative Computational Results [21]

SSLP	CPU (s)		
	D^2	D^2 -BAC	D^2 -BAC+SLP
5_25_50	1.64	0.70	0.36
5_25_100	2.15	1.73	0.89
5_50_100	7.10	3.70	1.56
5_50_500	34.50	23.05	12.36
5_50_1000	140.47	64.17	22.77
5_50_2000	603.37	274.40	42.74
10_50_50	295.95	373.98	262.13
10_50_100	396.76	452.31	486.99
10_50_500	1902.2	2772.22	1313.38
10_50_1000	5410.1	5677.80	2139.47
10_50_2000	9055.29	>10,800	3916.47
15_45_5	110.34	232.30	211.79
15_45_10	1494.89	222.41	153.41
15_45_15	7210.63	1988.26	803.56

in the future [6]. If demand is realized, then servers have to be allocated in such a way that capacity restrictions are satisfied and certain operating rules (e.g., servers can be only a specified distance from demand) must be satisfied, while meeting as much demand as possible. As shown in Ntamo and Sen [32] SSLP problems are essentially unsolvable without decomposition. (Here *unsolvable* refers to the inability to obtain an optimal solution within 10,000 s of computing on a Sun Fire 400 workstation.) However, Table 1 demonstrated that these are being solved to optimality using various versions of disjunctive decomposition (D^2). The earliest version, referred to as D^2 in Table 1, uses only set convexification, whereas the D^2 -BAC algorithm uses both value function and set convexification. The first two columns represent ideas that were implemented in Ntamo and Sen [32], and the column labeled D^2 -BAC+SLP refers to the implementation that uses a specialized SLP decomposition [21] to solve the cut formation LP (C^3 -LP). The improvements reported in Yuan and Sen [21] are clearly remarkable, considering that most of these instances are unsolvable using state-of-the-art deterministic MILP solvers [32].

The instances in Table 1 were first reported in Ntamo and Sen [6], and subsequently used as a test-bed for evaluating advances for this genre of decomposition algorithms. To appreciate the significance of some of the instances, we note that if one were to use a deterministic equivalent formulation of SSLP_10_50_2000, then the resulting problem would have at least 1 million binary variables ($10 \times 50 \times 2000$). In the decomposition framework, however, the number of first-stage variables is 10 and the number of second-stage variables is 500 (per scenario). Because of decomposition, the 2000 subproblems are solved independently, and as a result we are able to solve the high-dimensional deterministic equivalent by solving a collection of lower-dimensional problems. This is why decomposition is a winning strategy.

FURTHER CONNECTIONS WITH THE LITERATURE AND CONCLUSIONS

We have restricted this article to methods that can be viewed as the next generation of Benders' decomposition algorithms. While most of the methods presented in this article require the first-stage decisions to be binary, Sherali and Zhu [33] propose a search algorithm, which accommodates continuous variables in the first stage. The case of random recourse matrices has also been addressed in Ntamo [34].

In order to give the reader a brief picture of the SMIP landscape surrounding the class models stated in Equation (1)–(2) we mention the simple and/or complete recourse models studied in Klein Haneveld and van der Vlerk [35] and van der Vlerk [36]. Methods for this class fully exploit the structure of these problems by establishing connections with continuous recourse models. At the other end of the SMIP spectrum, we mention the pure SMIP models that have been studied using global optimization [37,38]. Finally, another important version of two-stage SMIP problems is the case of infinitely many scenarios (i.e., continuous random variables) [15]. This approach is based on an extension of stochastic decomposition [14], which

was originally designed for two-stage SLP problems.

The literature also makes several recommendations for multistage SMIPs as well [1,5,39]. This class of problems arises in several engineering applications in which system requirements evolve over time, and capacity/schedules must be adjusted to respond in an optimal way. The method discussed in Caroe and Schultz [5] was motivated by one such application in electricity operations scheduling. Despite the availability of basic methodology for multistage SMIP problems, there is clearly a paucity of computational work with these problems. It is likely that the basic multistage algorithms proposed in Lulli and Sen [1], Caroe and Schultz [5], and Alonso-Ayuso *et al.* [39] will need further enhancements using valid inequalities, such as those suggested in Guan *et al.* [40].

SMIP problems represent the next frontier of optimization research. Indeed, this area represents the union of applications that are addressed by SLP, and MILP, and the resulting impact cannot be overstated. It is just as clear that the computational challenges that need to be overcome are also significant. At this point in its development, theoretical as well as computational results look promising. As computational power grows over the next few decades, and further algorithmic advances are realized, we expect to see further successes and applications of SMIP models.

Acknowledgments

This research was supported, in part, by Air Force Office of Scientific Research (AFOSR) Grant FA9550-08-1-0154, and NSF grant CMMI - 0900070. This paper was started as part of a tutorial series given to Ph.D. students in the state of Ohio, under the banner "OR-Ohio." It was completed due to the patience and perseverance of Cole Smith and Shabbir Ahmed. Thanks are also due to the referees for their careful reading of the paper.

REFERENCES

1. Lulli G, Sen S. A branch and price algorithm for multi-stage stochastic integer programs with applications to stochastic lot sizing problems. *Manage Sci* 2004;50:786–796.
2. Barahona F, Berman S, Gunluk O, *et al.* Robust capacity planning in semiconductor manufacturing. *Nav Res Log* 2005;52: 459–468.
3. Goel V, Grossman I. A class of stochastic programs with decision dependent uncertainty. *Math Program Ser B* 2006;108:355–394.
4. Keller B, Bayraksan G. Scheduling job sharing multiple resources under uncertainty: a stochastic programming approach. *IIE Trans* 2010;42:16–30.
5. Caroe CC, Schultz R. Dual decomposition in stochastic integer programming. *Oper Res Lett* 1999;24:37–45.
6. Ntaimo L, Sen S. The million variable “March” for stochastic combinatorial optimization. *J Glob Optim* 2005;32(3):385–400.
7. Alonso-Ayuso A, Escudero LF, Garin A, *et al.* An approach for strategic supply chain planning under uncertainty based on stochastic 0-1 programming. *J Glob Optim* 2003;26: 97–124.
8. Laporte G, Van Hamme L, Louveaux FV. An integer L-shaped algorithm for the capacitated vehicle routing problem with stochastic demands. *Oper Res* 2002;50:415–423.
9. Benders JF. Partitioning procedures for solving mixed-variable programming problems. *Numer Math* 1962;4:238–252.
10. Bertsekas DP. *Dynamic programming and optimal control*. 3rd ed. Athena Scientific; 2005.
11. Schultz R. Stochastic programming with integer variables. *Math Program Ser B* 2003;97:285–309.
12. Sen S. Algorithms for stochastic mixed-integer programming models. In: Aardal K, Nemhauser GL, Weismantel R, editors. *Handbook of discrete optimization*. Amsterdam: North-Holland Publishing Co.; 2005. pp. 515–558.
13. Norkin VI, Ermoliev YM, Ruszczyński A. On optimal allocation of indivisibles under uncertainty. *Oper Res* 1998;46(3):381–395.
14. Higle JL, Sen S. Stochastic decomposition: an algorithm for two-stage linear programs with recourse. *Math Oper Res* 1991;16: 650–669.
15. Yuan Y. *Algorithmic advances in stochastic combinatorial optimization and applications* [PhD dissertation]. ISE Department, Ohio State University; 2010 May.
16. Schultz R. Continuity properties of expectation functions in stochastic integer programming. *Math Oper Res* 1993;18:578–589.
17. Dyer M, Stougie L. Computational complexity of stochastic programming problems. *Math Program* 2006;106:423–432.
18. Schmoys DB, Swamy C. An approximation scheme for stochastic linear programming and its application to stochastic integer programming. *J ACM* 2006;53:978–1012.
19. Rockafellar RT, Wets RJB. Scenario and policy aggregation in optimization under uncertainty. *Math Oper Res* 1991;16:119–147.
20. Infanger G. Monte Carlo (Importance) sampling within a Benders’ decomposition algorithm for stochastic linear programs. *Ann Oper Res* 1992;39:69–95.
21. Yuan Y, Sen S. Enhanced cut generation methods for decomposition-based branch-and-cut algorithms for two-stage stochastic mixed-integer programs. *INFORMS J Comput* 2009;21:480–487.
22. Penuel J, Smith JC, Yuan Y. An integer decomposition algorithm for solving a two-stage facility location problem with second-stage activation costs. *Nav Res Log* 2010;57:391–402.
23. Caroe CC, Tind J. L-shaped decomposition of two-stage stochastic programs with integer recourse. *Math Program* 1998;83(3): 139–152.
24. Tind J, Wolsey LA. An elementary survey of general duality theory in mathematical programming. *Math Program* 1981;21:241–261.
25. Laporte G, Louveaux FV. The integer L-shaped methods for stochastic integer programs with complete recourse. *Oper Res Lett* 1993;13:133–142.
26. Sen S, Higle JL. The C^3 theorem and a D^2 algorithm for large scale stochastic integer programming. *Math Program* 2005;104: 1–20.
27. Sherali HD, Fraticelli BMP. A modification of Benders’ decomposition algorithm for discrete subproblems: an approach for stochastic programs with integer recourse. *J Glob Optim* 2002;22:319–342.
28. Balas E, Ceria S, Cornuéjols S. A lift-and-project cutting plane algorithm for mixed 0-1 programs. *Math Program* 1993;58: 295–324.

29. Serali HD, Shetty CM. Optimization with disjunctive constraints. Volume 181, Lectures notes in economics and mathematical systems., Berlin: Springer; 1980
30. Van Slyke R, Wets RJB. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM J Appl Math* 1969;17:638–663.
31. Sen S, Serali HD. Decomposition with branch-and-cut approaches for two-stage stochastic integer programming. *Math Program* 2006;106:203–223.
32. Ntamo L, Sen S. A comparative study of decomposition algorithms for stochastic combinatorial optimization. *Comput Optim Appl* 2008;40:299–319.
33. Serali HD, Zhu X. On solving discrete two-stage stochastic programs having mixed-integer first- and second-stage variables. *Math Program* 2006;108:597–616.
34. Ntamo L. Disjunctive decomposition for two-stage stochastic mixed-binary programs with Random recourse. *Operations Research*. In press.
35. Klein Haneveld WK, van der Vlerk MH. Stochastic integer programming: general models and algorithms. *Ann Oper Res* 1999;85:39–57.
36. van der Vlerk MH. Convex approximations for complete integer recourse models. *Math Program* 2004;99:297–310.
37. Ahmed S, Tawarmalani M, Sahinidis NV. A finite branch and bound algorithm for two-stage stochastic integer programs. *Math Program* 2004;100:355–377.
38. Kong N, Schaefer AJ, Hunsaker B. Two-stage integer programs with stochastic right-hand sides: a superadditive dual approach. *Math Program* 2006;108:275–296.
39. Alonso-Ayuso A, Escudero LF, Ortuño MT. BFC, A Branch and Fix coordination algorithmic framework for solving some types of stochastic pure and mixed 0-1 programs. *Eur J Oper Res* 2003;151:503–519.
40. Guan Y, Ahmed S, Nemhauser GL. Cutting planes for multi-stage stochastic integer programs. *Oper Res* 2009;57:287–298.

STOCHASTIC MODELING AND OPTIMIZATION IN BASEBALL

BRAD NULL
AOL Research, Santa
Barbara, California

Baseball players and teams have many opportunities to use operations research to maximize their chances of winning games and pennants. For example, a pitcher must determine what pitch to throw, whether to pitchout, or whether to intentionally walk the hitter. A batter must decide whether to swing, bunt, or hit and run. Fielders must determine where to position themselves, for example, whether or not to “play in” to attempt to prevent a runner on third base from scoring, and runners must know whether to steal, run on contact, or take the extra base on a hit. Beyond this, the manager has to pick a lineup, place them in a batting order, decide when and whom to substitute, and set a pitching rotation, while the general manager must decide who to draft, who to trade, who to release, who to pursue in free agency and arbitration, and how much money to offer them. Even higher up, decisions must be made about when to hire and fire the manager or general manager and whom to replace them with, how to set payroll and ticket prices, and other more general business decisions.

Optimized solutions to these problems can have a significant impact on the bottom line. For example, researchers have estimated that optimal pitch selection could be worth as many as 2 additional wins a year [1], optimal batting order selection as much as 3 additional wins [2], and optimal closer usage up to 4.5 wins [3]! The value of a single win for a Major League team has been estimated to be between \$750,000 and \$4.5 million [4]. Thus, it is no surprise that these teams have begun to take notice of these problems and look for better solutions [5].

This article reviews operations research techniques that can be and are used to address these problems, focusing on those geared toward improving a team's chances of winning games and championships. We organize these problems into three groups:

1. *On-Field Decisions*: determining what a batter, pitcher, runner, or fielder will do at any point in the game.
2. *Player Usage Decisions*: determining which players to use at which points in the game.
3. *Personnel Decisions*: determining who should be on the team.

In addressing these problems, this article assumes a basic understanding of the game of baseball. For a general reference about the game, please see one of the many fine reference books in this area, such as Ref. 6, or just head to the local ballpark. (Note: the focus of this article is not intended to diminish the importance of several other important avenues of baseball research relating to such topics as scheduling, draft management, and stadium and operations planning and management; see Refs 7 and 8 for an introduction to some of these areas.)

ON-FIELD DECISION MAKING

By on-field decision, we are referring to a decision made by (or at least executed by) a player or players on the field. In this section, we review approaches used to address the various on-field decisions faced by pitchers, batters, fielders, and runners. Given the large number of such decisions, and the significant overlap in techniques used to address them, this section is organized according to analytical methods as opposed to particular decision problems. (While sample applications are discussed herein, references abound with more focused discussion of modeling related to the particular decision problems faced by

pitchers [1,9–11], batters [12–14], fielders [9,15], and baserunners [9,16].)

There are two key requirements to adequately model such on-field decisions (as there are with any decision analysis problem). These are as follows:

1. A reliable assessment of the probability of any outcome given any decision.
2. An understanding of the value of each possible outcome. These values can be in terms of the change in win probability of the action, the change in expected runs scored due to the action, or some other metric.

Given these two sets of information, determination of the optimal strategy is straightforward. As a simple example, consider a batter deciding whether or not to sacrifice bunt with a runner on first base and no outs in the bottom of the ninth inning. Assume that if the batter bunts he is always successful, resulting in the runner being advanced to second and the batter being thrown out. In this case, the batter's team will win 80% of the time. If the batter does not bunt, he gets a hit 30% of the time, resulting in the runner advancing to second base and the batter going to first base with no outs, which would result in the batter's team winning the game 90% of the time. Otherwise, the batter gets out and the runner remains at first base, in which case the batter's team wins 75% of the time. Thus, in the first case, the batter's team wins 80% of the time, and in the latter it wins 79.5% ($0.30 \times 0.90 + 0.70 \times 0.75$) of the time, so it would be slightly preferable for the batter to bunt. Of course, this is a simplified example, and there are significantly more play result possibilities and several ways of assessing these probabilities and values. In the remainder of this section, we focus on the various stochastic modeling techniques of use in assessing these quantities, as well as how to optimize decisions given these values.

Assessing Play Result Probabilities

We separate the primary methods of assessing play result probabilities into three categories according to their specificity: (i) average case probabilities, (ii) player-based probabilities, and (iii) probabilities assessed for specific plays. The main difference among the three lies in whether or not the abilities of the specific players involved are accounted for in the estimates. Average case methods do not account for specific player abilities; player-based methods generally account for either the batter or pitcher, and play-based methods account for both.

Average Case. Average case evaluation is the simplest approach to estimating play result probabilities. The basic approach aggregates historical data from a large set of games or plate appearances (associated with a large set of players) in which a particular action was taken—counting the historical frequency of the different possible results over these plate appearances. For example, to estimate the transition probabilities of “swinging away” in each of the 24 possible base out states in an inning (3 possible out states—0, 1, or 2, multiplied by eight possible combinations of runners on base— $2 \times 2 \times 2$), we might look at the total number of times a batter attempted to get a hit in each state over the course of a season, and the number of times each possible successor state was visited immediately afterward. This sort of analysis can also be used with respect to other actions. For example, in analyzing the sacrifice bunt, Tango *et al.*, look at all bunt attempts between 2000 and 2004, segmenting them in various ways, for example looking at bunt attempts by nonpitchers before the seventh inning with a runner on first base and no outs, to estimate the general probabilities of various results given such a strategy [9].

Player-Level Estimates. While average case analysis can be useful when pursuing general insight into decisions, the actual abilities of specific players vary wildly across numerous dimensions, and it is generally insightful to account for this when assessing optimal strategies for specific teams. While the most

obvious and direct method to estimate these probabilities would simply be to use a player's historical statistics to estimate his personal probabilities, this approach suffers from two common problems. First, there is often insufficient data to generate a reliable estimate. For example, most players attempt to bunt no more than a handful of times in a season, providing too little data with which to form any sort of estimate of ability. In addition, player performance generally suffers from the well-documented regression to the mean effect, meaning that players who have historically performed much better than the average player will not tend to maintain this high level of performance in the future (although they do frequently perform somewhat better than average), while players who perform worse than average will tend to improve somewhat in the future (if given the chance). For more on this phenomenon and its origins, see Ref. 17.

Precisely because of these sorts of issues, much energy has been devoted to developing models and projections of player ability and performance (see Ref. 18 for a comparison of the performance of a representative sample of them), many of which are quite complicated and estimate not just a player's underlying ability at a specific point in time but also factors that might affect a player's performance over time such as age, stadium, and year-to-year variance in performance.

Of particular interest for our applications are Bayesian approaches, which generate a prior distribution of player ability for specific populations of Major League players, meaning that any randomly selected player in this population could be viewed as having been drawn from this underlying ability prior distribution. Given this prior, adding evidence about a player (historical data) allows us to update the distribution and derive a posterior distribution for that specific player. These models have several attractive properties. First, they provide a framework for attempting to separate the inherent randomness in performance from the underlying ability; that is, account for the aforementioned regression to the mean effect. They

also make it very easy to deal with new players by simply assigning them the population prior distribution, and are often particularly easy to update as new information is gathered (particularly those based on Beta distributions such as in Refs 19 and 20, which are conjugate to multinomial data). In addition, they can be used to estimate abilities and likely results on a granular level (i.e., with respect to any specific possible outcome such as a strikeout or a home run) given any set of data. Most importantly though, such models have been shown to fit the data very well compared to other approaches, as well as possess excellent predictive power [20].

These sorts of models can also be applied to estimate baserunning and other abilities and can be modified to incorporate other environmental effects. Of course, very sparse data (such as a few observations per player) may lead to little separation in the estimates of the abilities of different players, but this is better than the alternatives of either assigning the same estimate to all players or using very sparse data to generate a point estimate.

Play-Level Estimates. Even more specific than player-level models are models that estimate the probabilities of the various outcomes of a specific game situation incorporating the influences of both the batter and the pitcher. One early and elegant model to estimate the probability of a particular result given a specific batter and pitcher (and their associated likelihoods) is the log5 model of James [21]. The general form of this model is

$$p(y) = \frac{\frac{x_b x_p}{x_l}}{\frac{x_b x_p}{x_l} + \frac{(1-x_b)(1-x_p)}{1-x_l}},$$

where $p(y)$ represents the probability of a specific outcome (e.g., probability of getting a hit), x_b and x_p represent the batter's and pitcher's individual probabilities with respect to the result in question, and x_l represents the league average. One key benefit of this model is that no parameterization is required. Given an assessment of the ability of the batter and pitcher, we can

immediately produce an assessment of the expected performance when they face each other. Further, many researchers have also commented on the good fit of this model to actual data [22].

In these models, as well as average case and player-based models, it is often important to account for other situational factors, such as the impact of the particular stadium, home field advantage, and handedness effects. Some of these effects have been estimated in isolation via various methods such as regression-based approaches (see Ref. 23 and its references for a comparison of such efforts to estimate stadium effects). Regression-based or curve-fitting approaches can also be used to attempt to estimate the impact of several effects at once, as well as generate an alternative model to the log5 for estimating the probability of particular play outcomes under certain conditions. An implementation of such a linear model is in Ref. 14, including a cautionary note about when such modeling of particular effects may or may not be appropriate. Finally, an extension of the log5, referred to as the *odds ratio method* has also been proposed to incorporate additional factors as well [24].

Estimating Outcome Values

Given some estimate of the probability of any play outcome, the question remains as to what is the value of that outcome. As mentioned above, the general goal of on-field decisions is to maximize the probability of winning the game. Thus, we would ideally like to determine the value of any play outcome in terms of the resulting probability or probability added of winning the game, although at times it may be useful to also quantify the value in terms of expected runs generated or some other metric. We discuss these various approaches below.

Markov Chains and Win Probability. One valuable concept that has been employed extensively for this task (since at least Howard in 1960 [25]) is the use of Markov models. One typical implementation of such (for example, in Refs 10 and 26) defines the state of a game as a 7-tuple composed of the following pieces of information:

- Inning
- Half inning
- Outs
- Orientation of the runners on base
- Lineup spot of the next batter for the visiting team
- Lineup spot of the next batter for the home team
- Home team runs scored minus visiting team runs scored.

Capping the run differential at ± 10 runs, modeling any extra inning as the ninth (i.e., if the teams are tied at the end of the ninth inning, we assume the next inning is also the ninth—an equivalent representation), and assuming a fixed lineup, gives 734,832 possible states. Thus, such models are generally very tractable for typical Markov chain analysis.

Using one of the methods discussed in the previous section we can either directly calculate the transition probabilities for this chain (i.e., the probability of any successor state as defined by the above 7-tuple given the current state) or else assess the likelihood of more conventional play results (e.g., strikeout or home run). In the latter case, we would then need a mapping from event types to resulting states given the current state. Such a mapping can be generated using historical data as in Ref. 14, a simple heuristic as in Ref. 27, or via another method.

In such chains, there are a limited number of possible successor states for any given state. Also, the states can be ordered chronologically (except for the potentially recurring ninth inning). These facts make calculating values of interest (e.g., the probability of either team winning, the distribution of the final score, etc.) with respect to this Markov chain considerably easier, and faster (see Ref. 14 for a thorough discussion of these points). To calculate these values, we can begin with all states that represent the end of the game and assign each a one or zero according to whether or not a selected team would have won the game in that state. Then, move backward among the states chronologically, determining the value of each state by multiplying the probability of transitioning

to each possible successor state from that state by the previously determined value of said successor state. This is exactly the backwards induction technique of dynamic programming, and thus different alternatives throughout the game can also be easily incorporated within this approach to convert this decision analysis approach into a full-scale dynamic program. (See the section titled “Dynamic Programming” in this encyclopedia for more on this general subject.)

This methodology is applied in Refs 10 and 14 in conjunction with player- and play-based transition models respectively to formulate optimal policies with respect to intentional walks and bunts for each of the 700,000 + possible states for specific matchups and games. The results in Ref. 14 also showed that context matters; that is, the optimal policy (with respect to bunting) can vary wildly from game to game as can the value of such optimization. Solving the dynamic program over 100 games from the 2007 season with transition probabilities derived using a Bayesian model of player ability, the percentage of states in which bunting was optimal varied between 7 and 17%, depending upon the specific team and opponent. A modified version of this approach was also used in Ref. 9 to construct average case Win Expectancy tables, which were then used to develop rules of thumb for optimal strategies with respect to bunts, steals, and intentional walks, incorporating some key elements of the associated players’ abilities.

Until now, we have glossed over the relatively strong Markov assumption that this approach makes, which requires that the transition probabilities are known and fixed from the beginning of the game. However, if player abilities are unknown, as represented in the Bayesian models of ability discussed in the section titled “Player Level Estimates,” and/or changing, this assumption does not hold. The impact of the first of these issues (unknown ability) was tested in Ref. 14 on actual games by comparing results generated via the Markov chain approach with results generated via simulation on both player ability and play results. While this research did show pitfalls of using Markov-based analysis

to evaluate season level results (different game outcomes are not independent!), it concluded that there is no discernable difference in game level predictions for the two models beyond the expected simulation error. The latter issue (changing ability) is a thornier one, particularly with respect to changes correlated with in-game performance; i.e. momentum. Debates have raged for years about whether or not momentum even exists in sports at all, and thus we leave that issue to another forum [28].

Run Expectancy. By slightly altering the state space and the value function, the Markov model from the previous section can also be used to estimate the expected number of runs scored in the remainder of a game or inning given a current state and transition matrix (as in Ref. 9 or 29). Bellman hypothesized that maximizing run expectancy may be a good proxy for maximizing win probability early in the game [30], although results in Ref. 10 indicate that optimal decisions under the two criteria can vary considerably. Nonetheless, while the output of such a model should be used with care, this approach significantly reduces the complexity of both the analysis and the ultimate output. As a result, the approach is quite common. For example, Ref. 9 takes this approach to create more general rules of thumb than those mentioned in the previous section.

Alternate Metrics. Simplifying things even further, some researchers have chosen to bypass the Markov chain framework altogether and use other metrics to evaluate possible outcomes. For example, in Ref. 1, optimal pitch selection is evaluated using the newly popular OPS statistic (on-base percentage plus slugging percentage) as the metric of choice. As discussed therein, OPS has been shown to be highly correlated with run production, which in turn has been closely tied to win percentage (namely, via the oft used Pythagorean theorem of baseball [31,32]). Thus, the hope of such an approach is that selecting the alternative to maximize the expected OPS of a given at bat (or from the defense’s point of view to minimize expected

OPS) will in turn maximize the team's probability of winning the game. These sorts of approaches are especially useful for decisions that need to be made in many contexts and without much time to plan, like pitch selection and base running.

Game Theory

Thus far, we have focused exclusively on a single team's strategy within a game. However, every game has two teams with opposing objectives. Thus, there is a natural game theoretic component to any decision problem faced therein. In other words, any analysis of what one team might do must account for the actions the other team may or may not take. On one level, the impact of the opposing team is not terribly significant. Each team is still solving its own optimization problem and to do so needs only to assess the probability of different outcomes, which is based only on the likelihood that the opponent will take certain actions (not whether this behavior is optimal). This is a reasonable approach for a one-off decision. However, the sort of decision analysis discussed in this article generally only proves truly useful when applied to large numbers of games and situations. If such analysis led a team to radically alter its behavior, their opponents would likely respond by altering their behavior as well. For example, if a team determines that it is generally optimal for a specific player to steal when he is on first base, and the player implements this strategy by always stealing immediately, then soon that team's opponents will realize this and always pitchout on these occasions, increasing the probability that runner will be caught stealing and likely eliminating the advantage in this strategy. Thus, it is worthwhile to examine equilibrium strategy in baseball to better understand the optimal long term strategy with respect to situations like these. Owing primarily to data limitations, existing research in this area generally stops at examining whether players appear to be using optimal mixed strategies. Such work has examined such decisions as pitch selection [1,11] and bunting [9], frequently concluding that players do not seem to be mixing optimally.

Fortunately, many decisions in baseball are sequential, for example, first the pitcher decides whether or not to intentionally walk the batter, and then if the pitcher does not, the batter decides whether or not to bunt. These can be integrated directly into the dynamic programming framework outlined above and all optimal decisions can be solved for sequentially. Further, comparing the results of such optimizations to those derived from one sided optimization can illuminate the "path to equilibrium." For decisions that are truly simultaneous, such as in the steal/pitchout sub-game, equilibrium strategies can be solved for using standard techniques (see the section titled "Game Theory" for more on this topic), the respective optimal strategies, and more importantly the likelihoods of the different outcomes of the play, can then be modeled as a single decision point within a larger dynamic program modeling the entire game.

PLAYER USAGE DECISIONS

While on-field decisions account for a large proportion of the in-game decisions that must be made by baseball teams, we have intentionally avoided the discussion of one significant class of in-game decisions thus far, substitution. This is because this and other player usage decisions can be considerably more complex than the decisions we have dealt with to this point. Thus, it is often difficult to solve for optimal strategies for these decisions, requiring the search for approximately optimal strategies and use of heuristics. There are two key reasons for this additional complexity.

First, the objective often requires a wider scope; that is, we generally have to look beyond the current game to see the impact of these decisions. For example, choosing to use a player in a particular game can affect his availability (especially pitchers) in subsequent games. Also, the risk of fatigue and injury rise the more a player is used. Finally, given that we do not have perfect information about all players and their abilities, the more we use a player, the more information we gather on that player, the value of which

can be significant when deciding whether or not to use this player in future games.

Second, the model of a single game becomes necessarily more complex when we allow for substitution. Consider the Markov chain previously discussed with just over 700,000 states. If we allow for each team to use six relief pitchers and try to incorporate these possibilities into such a chain, the state space expands by a factor of 37,249 (see Ref. 14 for details on the calculations). Allowing for five possible pinch hitters multiplies the state space by a factor of 37,295,449! Given these issues, various approaches have been taken to limit the scope of these problems and derive insights. Below we highlight some recent results with respect to several player usage decisions including substitution, batting order selection, and setting of pitching rotations.

Substitution

The most straightforward way to analyze substitution alternatives is to implement the dynamic programming approach laid out in the section titled “Markov Chains and Win Probability” with respect to a pared down set of substitution possibilities. Hirotsu and Wright [26] took this approach, solving such a problem limited to three substitutes. Even with this limitation, their approach, implemented in 2003, took approximately 12h to solve one game! Despite the fact that processor speeds continue to improve, this highlights the importance of approximate techniques.

The authors in Refs 3 and 9 both took heuristic approaches utilizing the concept of leverage (defined as the ratio between how much a single run scored changes the probability of a team winning in the current situation vs how it would have changed their probability of winning at the beginning of the game). They used this concept to greatly reduce the complexity of the state space by defining every possible state by its leverage (and inning) and attempting to find an optimal benchmark strategy of the form that a particular type of substitute should be brought in if the current leverage is above a certain amount. For example, Woolner [3]

estimated that if a team were to bring its closer in any time, the leverage rises above 2.55 (with slightly different rules if it is the ninth inning or the closer is well rested); this would be worth an average of 1.6 wins per team over the course of a season.

Batting Order

Another important problem a manager faces daily is selecting a batting order. This problem has been studied in operations research circles since at least 1966 [12], and frequently with the incorporation of a Markov-chain-based analysis [29]. Again, the main problem with this approach is computational. There are $9!$ possible lineups (assuming the nine starters are preselected, otherwise there can be significantly more). Attempts to exhaustively search for the optimum among these many choices took 5.5 days for researchers in 1997 and 2h in 2003 [2]. While improvements in processor speed will continue to reduce this time, this is certainly another area where fast, nearly optimal heuristics could be of value. Two such approaches that have been taken recently include the application of decision rules [29] and classification of batters and logical class ordering [2]. The latter was shown to yield lineups with an expected run production within 0.1 runs of optimum on average, while running thousands of times faster than a program to solve for the optimal solution.

Other Usage Decisions

Other usage decisions a manager faces over the course of a season include actually picking which players to start in a game and setting a pitching rotation. One oft-debated question in this area concerns whether teams are benefited by the five man rotation that has become ubiquitous since the late 1980s versus the four man rotation common in the previous era. This decision differs from optimizing substitution or the batting order in that the issues here are not computational, but rather primarily a problem of estimating true fatigue effects. The difficulty of such estimation is highlighted by the fact that, in two recent analyses of the subject, Refs 9 and

33, one concluded that a five man rotation was optimal, the other recommended four!

All in all, while the work in this area to date has been informative, it is obvious that there is still much room for improved models and approaches to address these challenging problems. In particular, such approaches may benefit from more integration with Markov-chain- or simulation-based models of entire games (as much of the other research discussed herein has), perhaps applying approximate dynamic programming techniques (see the section titled Approximate Dynamic Programming in this encyclopedia for more on this topic) to address these much larger problems.

PERSONNEL DECISIONS

The final decisions we discuss are those involving the acquisition and dismissal of player personnel. How do teams decide who they want on their team and how much to pay them? Of course, the foremost consideration in determining whether or not to acquire a player is to estimate how well he will perform and the value of this performance. As mentioned in the section titled "Player-Level Estimates," much research has been devoted to creating systems to forecast player performance. Further, several metrics have been invented to value players (particularly in recent years), most of which value players in terms of expected runs created. See, for example, the Linear Weights methods detailed in Ref. 34. Given such values, follow-on analysis can use the commonly accepted convention that 10 runs is approximately worth 1 win to estimate the impact on wins and losses of given players. For example, this approach is taken in Ref. 4 to determine the value of Alex Rodriguez.

This value can be highly context specific. For instance, that same article calculated different values for different wins, determining that wins more likely to get a team into the playoffs are approximately six times more valuable than those with little effect on a team's pennant chances. Further, different players may be more valuable to certain teams due to other characteristics that cause

them to have a better "fit" with that team. For example, one classic hypothesis, discussed in Ref. 29, asserts that a team with a mix of players with a high propensity to reach base as well as players with a high propensity to hit home runs can generate more runs than a team with players of just one type or the other. Additionally, given that we may be more certain about the ability of certain players (i.e., those with a large amount of Major League experience), there may be a team-varying value of this certainty. For example, a high-variance player ("high upside") may be more valuable to team just on the outside of making the playoffs, while a dominant team may prefer the known quality ("proven veterans") even though they both have the same expected ability. Initial analysis to attempt to measure such effects [14] indicates that different players do have significantly different effects on different teams. In a small sample of experiments, this value for particular players ranged by up to 2 wins.

FRONTIERS

This article has highlighted just some of the operations research problems, methods, and results in baseball today. In the coming years, such models will certainly be expanded upon and potentially rewritten. Of particular interest in the near future will be models derived from new data that is providing much more granular information about the game, such as the precise location and movement of individual pitches [35], or the precise positioning and movement of the fielders on each play. Such data is opening up additional problems to quantitative analysis, such as player positioning and pitch selection, as well as new methods to address these problems. This, as well as the ongoing leveraging of more and more advanced operations research techniques, will continue to expand the horizons of our knowledge of optimal strategies in baseball.

REFERENCES

1. Kovash K, Levitt SD. Professionals do not play minimax: evidence from major

- league baseball and the national football league. Working Paper 15347. 2009. www.nber.org/papers/w15347.
2. Sokol JS. A robust heuristic for batting order optimization under uncertainty. *J Heuristics* 2003;9:353–370.
 3. Woolner K. Are teams letting their closers go to waste? In: Jonah K, editor. *Baseball between the numbers*. New York: Basic Books; 2006. pp. 253271.
 4. Silver N. Is alex rodriguez overpaid? In: Jonah K, editor. *Baseball between the numbers*. New York: Basic Books; 2006. pp. 253271.
 5. Lewis M. *Moneyball*. New York: W.W. Norton and Co.; 2004.
 6. Hample Z. *Watching baseball smarter: a professional fan's guide for beginners, semi-experts, and deeply serious geeks*. Vintage; 2007.
 7. Trick MA. *Integer and constraint programming approaches for round-robin tournament scheduling*. Heidelberg: Springer Berlin; 2003. DOI: 10.1007/b11828.
 8. Noll RG, Zimbalist AS, editors. *Sports, jobs, and taxes: the economic impact of sports teams and stadiums*. Harrisonburg (VA): Donelley and Sons; 1997.
 9. Tango T, Lichtman M, Dolphin A. *The book - playing the percentages in baseball*. Dulles (VA): Potomac; 2007.
 10. Null B. *Baseball decision optimization models and the intentional walk*. 2005. A tutorial project. www.baseballcalculus.com/wp-content/uploads/2010/02/BaseballTutorial.pdf.
 11. Weinstein-Gould J. Keeping the hitter off balance: mixed strategies in baseball. *J Quant Anal Sports* 2009;5(2):Article 7. DOI: 10.2202/1559-0410.1173.
 12. Cook E. *Percentage baseball*. Cambridge (MA): M.I.T. Press; 1966.
 13. Trueman R. Analysis of baseball as a markov process. In: Ladany SP, Macol RE, editors. *Optimal strategies in sports*. New York: Elsevier-North Holland; 1977.
 14. Null B. *Stochastic modeling and optimization in baseball [Thesis paper]*. Stanford University; 2009.
 15. Jensen S, Shirley K, Wyner A. Bayesball: a Bayesian hierarchical model for evaluating fielding in major league baseball. *Ann Appl Stat* 2009;3:491–520.
 16. Baumer B. Using simulation to estimate the impact of baserunning ability in baseball. *J Quant Anal Sports* 2009;5(2):Article 8. DOI: 10.2202/1559-0410.1174.
 17. Schall E, Smith G. Do baseball players regress toward the mean? 2007 Available at <http://www.economics.pomona.edu/GarySmith/BBr-egress/baseball.html>.
 18. Silver N. Hitter projection roundup. 2007 Available at <http://www.baseballprospectus.com/unfiltered/?p=564>.
 19. Albert J. A breakdown of a batter's plate appearance - four hitting rates. By the Numbers. 2006. pp. 23–30.
 20. Null B. Modeling baseball player ability with a nested dirichlet distribution. *J Quant Anal Sports* 2009;5(2):Article 5. Available at <http://www.bepress.com/jqas/vol5/iss2/5>.
 21. James B. *The bill james baseball abstract* 1983. New York: Ballantine; 1983.
 22. Levitt D. *The Batter/Pitcher Matchup*. 1999 Available at http://www.baseballthinkfactory.org/btf/scholars/levitt/articles/batter_pitcher_matchup.htm.
 23. Acharya R, Ahmed A, D'Amour A, *et al.* Improving major league baseball park factor estimates. *J Quant Anal Sports* 2008; 4(4):Article 2.
 24. Tango T. The odds ratio method. 2006 Dec 18. Available at www.insidethebook.com/ee/index.php/site/comments/the_odds_ratio_method/.
 25. Howard RA. *Dynamic programming and markov processes*. Cambridge (MA): M.I.T. Technology Press and Wiley; 1960. pp. 49–54.
 26. Hirotsu N, Wright M. Modeling a baseball game to optimize pitcher substitution strategies incorporating handedness of players. *IMA J Manag Math* 2005;16(2): 179–194. DOI: 10.1093/imaman/dpi009.
 27. D'Esopo DA, Lefkowitz B. The distribution of runs in the game of baseball. In: Ladany SP, Macol RE, editors. *Optimal strategies in sports*. New York: Elsevier-North Holland; 1977.
 28. Reifman A. The hot hand in sports. 2006–2010 Available at <http://thehothand.blogspot.com/>
 29. Bukiet B, Harold ER, Palacios JL. A markov chain approach to baseball. *Oper Res* 1997; 45:14–23.
 30. Bellman R. *Dynamic programming and markovian decision processes, with application to baseball*. In: Ladany SP, Macol RE, editors. *Optimal strategies in sports*. New York: Elsevier-North Holland; 1977.

31. James B. The bill james baseball abstract 1982. New York: Ballantine; 1982.
32. Miller S. A derivation of James' Pythagorean projection. *Chance Mag* 2007;20(1): 40–48.
33. Woolner K. Five starters of four? On pitching and stamina. In: Jonah K, editor. *Baseball between the numbers*. New York: Basic Books; 2006. pp. 253271.
34. Tango T. Linear weights. 2008. Available at www.tangotiger.net/wiki/index.php?title=Linear_Weights
35. Nathan A. Tracking baseball pitches using video technology: The PITCHf/x system. 2010. Available at <http://webusers.npl.illinois.edu/~a-nathan/pob/pitchtracker.html>

STOCHASTIC NETWORK INTERDICTION

DAVID P. MORTON
Graduate Program in Operations
Research, The University of
Texas at Austin, Austin,
Texas

INTRODUCTION

A prototypical interdiction model involves two adversaries, a leader and a follower. In this article, we begin by illustrating ideas through a shortest-path network interdiction model, where the follower seeks a shortest path in a directed network. Before the follower does so, the leader can first lengthen (interdict) some of the arcs in the network, under limited resources, to maximize the length of the follower's shortest path. We begin with *deterministic* shortest-path interdiction before turning to the stochastic case. This serves to set ideas and notation in a simpler setting. More importantly, it allows us to distinguish formulation techniques and algorithmic tools that work well in both settings from those specific to the deterministic case. With the same motivation, we also consider deterministic and stochastic models for interdicting a maximum-reliability path.

We focus on the interdiction of shortest-path, and maximum-reliability-path, networks for concreteness, but these should be viewed as proxies for a more general model of operations. For example, instead of a shortest-path problem, the operations problem can involve critical infrastructure such as a municipal water system [1], a communications network [2], an electric power system [3,4], or a petroleum reserve [5]. Alternatively, the operations problem can involve transport of illicit nuclear material, or a weapon, across a transportation network [6,7] or can involve a rogue country's program to build a first nuclear weapon [8]. In the

former cases, "we" may be the follower, that is, we operate the infrastructure. And, in this setting, we seek to understand our system's vulnerability to attack by formulating an interdiction model so that we may thwart such an attack. Alternatively, we may seek to disrupt operation of an adversary's critical infrastructure, and so we may be the leader. Indeed, in the latter set of examples, we are the leader, seeking to interdict such an adversary.

Several fundamental ideas in optimization modeling and algorithms come to the fore in this article. Our core interdiction problem is naturally formulated as a nested "max-min" model with the outer decision variables being discrete and the inner minimization being a linear program. Such a model cannot be solved by standard optimization software. That said, fixing the variables of the outer maximization, and taking the dual of the inner linear program (LP) leads to a nonnested mixed-integer program (MIP), which can be solved by standard software. We find it appealing that duality plays such an integral role in simply formulating a tractable optimization model. We also describe an interdiction model where direct application of this idea fails to yield a linear MIP. Here, we instead achieve a linear MIP by using an exact-penalty reformulation of the inner LP. We develop special-purpose Benders' decomposition algorithms for the interdiction models we consider. The network nature of our problems leads to particularly transparent master programs, that is, models that can be written down immediately as being obviously correct formulations. This basic decomposition algorithm is enhanced by valid inequalities that use a simple logical construct, that is, the inequalities condition on whether the leader interdicts at least one arc on paths attractive to the follower.

The section titled "Deterministic Shortest-Path Network Interdiction" relies heavily on Israeli and Wood [9], and describes deterministic interdiction of a shortest path model in which the follower's source (s)

and terminus (t) are known to the leader. The section titled “Stochastic Shortest-Path Network Interdiction” generalizes this to the case where the follower’s s - t pair is known to the leader only through a probability distribution. The sections titled “Interdicting a Maximum-Reliability Path: Known s - t Pair” and “Interdicting a Maximum-Reliability Path: Stochastic s - t Pair” consider a variant in which the leader minimizes the follower’s maximum-reliability path, first with a certain s - t pair and then with a stochastic s - t pair. The former is shown to be captured by the model of the section titled “Deterministic Shortest-Path Network Interdiction,” but the latter requires another approach. The last section summarizes.

To avoid excessive problem development with diminishing returns on modeling and algorithmic insights, in this article, we restrict attention to four related interdiction problems on shortest-path and maximum-reliability networks. However, before proceeding, we point to work on interdiction of another classic network-flow problem, a maximum-flow network. Deterministic maximum-flow network interdiction is studied by Phillips [10] and Wood [11].

In stochastic maximum-flow interdiction, both arc capacities and the success of interdiction attempts can be random. The leader can (attempt to) reduce the capacity of a subset of arcs, possibly to zero. The follower maximizes flow in the residual network, realizing the arc capacities. The leader’s goal is to minimize the expected value of the resulting maximum flow. Both Cormican *et al.* [12] and Janjarassuk and Linderoth [13] consider interdiction of a stochastic maximum-flow network in which the sample space is large—even infinite, because, for example, arc capacities are continuous random variables. Cormican *et al.* use deterministically valid upper and lower bounds on the optimal value, that is, on the minimal value of the expected maximum flow. These bounds are defined on a partition of the sample space, which is iteratively refined, tightening the bounds. Janjarassuk and Linderoth instead form upper and lower bounds on the optimal value using Monte Carlo sampling. While it is beyond the scope of this article, these types

of bounds can also be used in the stochastic network interdiction problems that we consider here when the sample space is large.

DETERMINISTIC SHORTEST-PATH NETWORK INTERDICTION

Formulation and Reformulation

In our shortest-path interdiction model, the timing of the decisions and the information available to the two adversaries are as follows. The adversaries agree on the network topology $G(N, A)$, the follower’s source, $s \in N$, and terminus, $t \in N$, and they further agree on the network’s arc lengths. The leader knows the mechanism by which the follower selects an s - t path, that is, that the follower selects a shortest path. Decisions are made sequentially. The leader first lengthens some subset of the network’s arcs. The follower then selects a shortest s - t path, knowing which arcs were lengthened. The leader’s goal is to maximize the length of the follower’s shortest path. We formulate the deterministic shortest-path network interdiction model as follows:

Network and sets:

$G(N, A)$ = directed network with nodes N
and arcs A
 $FS(i)$ = set of arcs leaving node i
 $RS(i)$ = set of arcs entering node i

Data:

R = resource matrix limiting interdiction decisions
 r = resource vector limiting interdiction decisions
 c_{ij} = original length of arc $(i, j) \in A$; $c_{ij} > 0$
 d_{ij} = additional length if arc $(i, j) \in A$ is interdicted; $d_{ij} > 0$

Leader’s decision variables:

$x_{ij} = 1$ if arc $(i, j) \in A$ is lengthened and 0 otherwise

Follower’s decision variables:

$y_{ij} = 1$ if arc $(i, j) \in A$ is traversed in follower’s shortest path and 0 otherwise

Deterministic Shortest-Path Interdiction Formulation:

$$z^* = \max_{x \in X} \left[\begin{array}{l} \min_y \sum_{(i,j) \in A} (c_{ij} + d_{ij}x_{ij})y_{ij} \\ \text{s.t.} \sum_{(i,j) \in FS(i)} y_{ij} - \sum_{(j,i) \in RS(i)} y_{ji} \\ = \begin{cases} 1 & i = s \\ 0 & i \in N \setminus \{s, t\} \\ -1 & i = t \end{cases} \\ y_{ij} \geq 0, (i,j) \in A \end{array} \right], \quad (1)$$

where $X = \{x : Rx \leq r, x \in \{0, 1\}^{|A|}\}$. Here, x and y denote the $|A|$ -dimensional binary vectors with respective components x_{ij} and y_{ij} , $(i, j) \in A$. We assume that G has an s - t path. The set X can, for example, simply constrain the number of arcs the leader can lengthen by setting $R = (1, 1, \dots, 1)$, $x \in X$ can represent multiple knapsack constraints across different interdiction resources, and/or X can incorporate more general constraints. The leader may be precluded from interdicting certain arcs by the choice of R and r , and X can also include logical inequalities to exclude nonoptimal combinations of interdictions [14].

With $x \in X$ fixed, the inner optimization problem in model (1) is simply the follower's shortest-path problem, formulated as a minimum-cost flow problem, with one unit of flow originating at s and terminating at t . Arc (i, j) has length c_{ij} if $x_{ij} = 0$ and $c_{ij} + d_{ij}$ if $x_{ij} = 1$.

Owing to its max-min structure, model (1) does not lend itself to direct solution by a general purpose solver for MIPs. However, the inner problem is an LP and so we can (i) fix $x \in X$, (ii) take the dual of the inner LP, and (iii) release x . With $-\pi_i$, $i \in N$, denoting the dual variables associated with the flow-balance constraints of model (1), doing so leads to the following MIP:

$$z^* = \max_{x, \pi} \pi_t \quad (2a)$$

$$\text{s.t. } x \in X \quad (2b)$$

$$\pi_j - \pi_i - d_{ij}x_{ij} \leq c_{ij}, (i, j) \in A \quad (2c)$$

$$\pi_s = 0. \quad (2d)$$

When x is fixed we have the dual of a shortest-path problem, with π_i , $i \in N$,

denoting the length of the shortest path from s to i , here with arc lengths in the interdicted network as dictated by x . Constraint (2c), coupled with maximization of the objective function, captures Bellman's equations for the shortest-path problem

$$\pi_j = \min_{i \in RS(j)} \{\pi_i + (c_{ij} + d_{ij}x_{ij})\} \forall j \in N.$$

Having expressed our model as a single, nonnested, optimization problem, we can attempt to solve model (2) using a branch-and-bound algorithm for general MIPs. This approach may succeed, and if so, is arguably the simplest way to solve model (1). However, one reason this approach may fail is that the LP relaxation of model (2) can be weak, especially when the values of the d_{ij} parameters are large relative to the original arc lengths, c_{ij} . Roughly speaking, this weakness arises because the LP solution's fractional values of x are such that the set of attractive paths for the follower have equal length. Valid inequalities have been proposed to tighten the LP relaxation and we discuss these further below.

Remarks.

1. Model (1) captures discrete interdiction of arcs, but if we relax the binary constraints on x to $0 \leq x_{ij} \leq 1$, $(i, j) \in A$, then we allow continuous lengthening of arc (i, j) from c_{ij} up to $c_{ij} + d_{ij}$ as considered by Fulkerson and Harding [15] and Golden [16]. In this case, we can solve the desired interdiction model by solving the LP relaxation of model (2).
2. Model (1) is a bilevel Stackelberg game with the leader's decision transparent to the follower. An alternative that instead captures secrecy is a Cournot game in which the adversaries effectively announce their decisions simultaneously. In our setting, this leads to a two-person zero-sum game involving so-called *mixed strategies*. Player II (formerly the follower) places a probability distribution on an exponentially large collection of s - t paths, and Player I (formerly the leader) places a probability distribution on the exponentially

large collection of feasible solutions of X . See Washburn and Wood [17] for such a model and Nehme [18] for a hybrid model in which some of the leader's decisions are transparent to the follower but others are concealed.

3. We assume above that the leader and follower agree on the arc lengths. An alternative is to model the two adversaries as having differing perceptions of the arc lengths. Bayrak and Bailey [19] study shortest-path network interdiction under such information asymmetries.
4. Exchanging the order of the decisions in model (1) to form " $\min_y \max_x$ " yields the robust optimization model of Bertsimas and Sim [20], where the decision-maker (the leader) first selects an s - t path under *uncertain* arc lengths. Their model of uncertainty is adversarial, with nature (the follower) lengthening arcs in a worst-case manner, given y 's s - t path and constrained by X . In the interdiction model (1), the path-selection decision is made in a *certain* environment. That said, model (1) can provide insight regarding vulnerabilities in a system we operate by allowing a resource-constrained adversary to optimally disrupt our system. Presumably, we seek to understand such vulnerabilities so that we can harden our system prior to the potential disruption.
5. The special case in which $X = \{x : \sum_{(i,j) \in A} x_{ij} \leq k\}$ and d_{ij} is sufficiently large for all $(i,j) \in A$, leads to a cardinality-constrained shortest-path interdiction problem in which interdiction of an arc effectively removes it from the network. This problem is known as the *k-most-vital arcs problem* [21,22] and the associated decision problem is strongly NP-complete [23,24].
6. We assume that the leader knows the follower will select a shortest s - t path. If the leader is unsure how the follower will behave, then model (1) is arguably appropriately conservative, that is, it

plans for the worst case. If it is known, for example, that the follower will select a random s - t path, or traverse the network according to a Markov decision process (MDP), then a different interdiction model may be in order, depending on the leader's goal. We return briefly to a random-walk model in the next section. See Bailey *et al.* [25] for interdiction of a follower who solves an MDP.

7. Model (1) can be used to interdict a follower who selects a maximum-reliability s - t path in a network in which each arc has some probability of evading detection and a reduced probability of evading detection when a sensor is installed on the arc. We return to this in greater detail in the section titled "Interdicting a Maximum-Reliability Path: Known s - t Pair."

Algorithms

We can employ Benders' decomposition [26] to solve model (2). Here, we partition the decision variables into binary variables, x , and continuous variables, π . We can view model (2) as

$$z^* = \max_{x \in X} h(x), \quad (3)$$

where

$$\begin{aligned} h(x) &= \max_{\pi} \pi_t \\ \text{s.t. } & \pi_j - \pi_i \leq c_{ij} + d_{ij}x_{ij}, (i,j) \in A \\ & \pi_s = 0 \\ & = \min_y \sum_{(i,j) \in A} (c_{ij} + d_{ij}x_{ij})y_{ij} \end{aligned} \quad (4a)$$

$$\begin{aligned} \text{s.t. } & \sum_{(i,j) \in FS(i)} y_{ij} - \sum_{(j,i) \in RS(i)} y_{ji} \\ & = \begin{cases} 1 & i = s \\ 0 & i \in N \setminus \{s, t\} \\ -1 & i = t \end{cases} \end{aligned} \quad (4b)$$

$$y_{ij} \geq 0, (i,j) \in A, \quad (4c)$$

and where the first linear program, with optimal value $h(x)$, is again dual to that defined in model (4).

The function $h(\cdot)$ is concave on the convex hull of X . Benders' decomposition iteratively builds a piecewise-linear outer approximation of $h(\cdot)$ as follows. Let Y denote the feasible region of model (4) and let $y^{(\ell)}$, $\ell \in \mathcal{L}$, denote all of Y 's extreme points, that is, each $y^{(\ell)}$ denotes an s - t path that the follower can select. Denote these paths $P^{(\ell)}$, $\ell \in \mathcal{L}$, that is, $(i, j) \in P^{(\ell)}$ if and only if $y_{ij}^{(\ell)} = 1$. The so-called *full master program* for Benders' decomposition is then

$$\begin{aligned} z^* = \max_{x, \theta} \quad & \theta \\ \text{s.t.} \quad & \theta \leq \sum_{(i,j) \in A} (c_{ij} + d_{ij}x_{ij})y_{ij}^{(\ell)}, \quad \ell \in \mathcal{L} \\ & x \in X, \end{aligned}$$

where the inequalities defining θ are known as optimality cuts, or simply cuts. Equivalently, using our path notation, we can write the full master program as

$$\begin{aligned} z^* = \max_{x, \theta} \quad & \theta \\ \text{s.t.} \quad & \theta \leq \sum_{(i,j) \in P^{(\ell)}} (c_{ij} + d_{ij}x_{ij}), \quad \ell \in \mathcal{L} \\ & x \in X. \end{aligned} \tag{5}$$

Algorithm 1. Benders' decomposition algorithm for model (1).

Input: an instance of model (1), error tolerance $\epsilon \geq 0$

Output: interdiction plan $\bar{x}$ with $z^* - h(\bar{x}) \leq \epsilon$

find at least one s - t path; initialize L and program (6)'s set of cuts

$\underline{z} \leftarrow 0$

repeat

 solve model (6) to obtain $\hat{x}$ and $\bar{z}$

 solve model (4) with $x = \hat{x}$ to obtain s - t path P and optimal value $h(\hat{x})$

 update L and model (6)'s set of cuts with path P

if $h(\hat{x}) > \underline{z}$ **then**

$\bar{x} \leftarrow \hat{x}$; $\underline{z} \leftarrow h(\hat{x})$

end if

until $(\bar{z} - \underline{z} \leq \epsilon)$

return $\bar{x}$ and $\underline{z} = h(\bar{x})$

Israeli and Wood [9] report that direct application of Algorithm 1 is ineffective in solving model (1). It is more effective to simply apply a general purpose MIP solver to formulation (2). We detail Benders' decomposition for two reasons. First, Israeli and Wood [9] report that a version of Algorithm 1 is quite effective when enhanced

In this way, the full master program (5) is an "obviously correct" formulation that we could have written down immediately following the problem description. (This is arguably not the case for all applications of Benders' decomposition.)

Of course, enumerating all s - t paths is out of the question for all but the smallest of networks. Benders' decomposition, also known as the *L-shaped method* [27], iteratively updates a subset of paths $L \subset \mathcal{L}$ in a relaxed master program

$$\begin{aligned} \bar{z} = \max_{x, \theta} \quad & \theta \\ \text{s.t.} \quad & \theta \leq \sum_{(i,j) \in P^{(\ell)}} (c_{ij} + d_{ij}x_{ij}), \quad \ell \in L \\ & x \in X. \end{aligned} \tag{6}$$

In particular, we solve the master program (6) to obtain an interdiction plan $\bar{x}$ and an optimal value satisfying $\bar{z} \geq z^*$. We then solve the shortest-path problem (4) with $x = \bar{x}$ to obtain a new path P with optimal value $h(\bar{x}) \leq z^*$. If $\bar{z} - h(\bar{x})$ is sufficiently small, we terminate. Otherwise, we add P to the collection of paths indexed by L and repeat. This is formalized in Algorithm 1.

in a manner we describe shortly. Second, in the stochastic case, even when Benders' decomposition is not enhanced, it is more effective than attempting to solve directly the stochastic analog of model (2).

Solving the MIP master program (6) at each iteration requires much more computational effort than solving the

shortest-path subproblem (4). Israeli and Wood [9] find that enumerating *all* shortest s - t paths when solving model (4) for a fixed x , or even enumerating near-shortest paths, and adding the associated collection of cuts to model (6) improves performance.

We now turn to an idea that can be used on its own or, better, to enhance Algorithm 1. Suppose that P is a path associated with a cut of model (6) and suppose the lower bound of Algorithm 1 satisfies $\underline{z} > \sum_{(i,j) \in P} c_{ij}$. Then, we may add a path-covering inequality

$$\sum_{(i,j) \in P} x_{ij} \geq 1, \quad (7)$$

to the relaxed master program. The intuition behind inequality (7) is simple: The leader has, in the course of the algorithm, already forced the follower to a path with length $\underline{z}$. If not interdicted, path P has length $\sum_{(i,j) \in P} c_{ij} < \underline{z}$. So, the leader must interdict at least one arc on P , as enforced by inequality (7).

This idea can be pushed further. Suppose that P is a path associated with a cut of model (6) and suppose the lower bound of Algorithm 1 satisfies $\underline{z} > \sum_{(i,j) \in P} c_{ij} + \max_{(i,j) \in P} d_{ij}$. Then, we may add

$$\sum_{(i,j) \in P} x_{ij} \geq 2, \quad (8)$$

and so on.

To simplify the discussion, we now assume that (i) an interdicted arc is removed from the network and (ii) there is an s - t path in the residual network for all $x \in X$, that is, it is impossible to disconnect s and t . While arc

removal can be handled in our original formulation by selecting d_{ij} to be sufficiently large, this can lead to very weak LP relaxations, and hence, we seek another approach. Moreover, the assumption that interdiction removes an arc is without loss of generality because an arc (i,j) can be replaced by two arcs in parallel (albeit with appropriate extensions of notation): one arc from i to j has length c_{ij} , the second arc from i to j has length $c_{ij} + d_{ij}$, and we disallow interdiction of the latter arc through constraints $x \in X$.

We can partition $\mathcal{L}$ via $\mathcal{L} = \underline{\mathcal{L}} \cup \mathcal{L}^* \cup \bar{\mathcal{L}}$, where

$$\begin{aligned} \underline{\mathcal{L}} &= \left\{ \ell : \sum_{(i,j) \in P^{(\ell)}} c_{ij} < z^* \right\} \\ \mathcal{L}^* &= \left\{ \ell : \sum_{(i,j) \in P^{(\ell)}} c_{ij} = z^* \right\} \\ \bar{\mathcal{L}} &= \left\{ \ell : \sum_{(i,j) \in P^{(\ell)}} c_{ij} > z^* \right\}. \end{aligned}$$

Consider

$$x \in X, \quad \sum_{(i,j) \in P^{(\ell)}} x_{ij} \geq 1, \quad \ell \in L. \quad (9)$$

An interdiction plan x satisfies system (9) with $L = \underline{\mathcal{L}}$ if and only if x solves model (1). Moreover, system (9) with $L = \underline{\mathcal{L}} \cup \mathcal{L}^*$ is infeasible. Of course, we do not know z^* and hence cannot partition $\mathcal{L}$ *a priori*, but we can do so algorithmically as detailed in Algorithm 2.

Algorithm 2. Covering-based decomposition algorithm for model (1).

Input: an instance of model (1)

Output: optimal solution x^* and optimal value z^* to model (1)

$L \leftarrow \emptyset$

$\hat{x} \leftarrow 0$

$\underline{z} \leftarrow 0$

repeat

 solve model (4) with $x = \hat{x}$ to obtain s - t path P and optimal value $h(\hat{x})$

 update L and system (9)'s set of inequalities with path P

if $h(\hat{x}) > \underline{z}$ **then**

```

 $\bar{x} \leftarrow \hat{x}; \bar{z} \leftarrow h(\hat{x})$ 
end if
  find  $\hat{x}$  satisfying system (9) or determine that system (9) is infeasible
until (system (9) is infeasible)
return  $x^* = \bar{x}$  and  $z^* = \bar{z}$ 

```

An attractive aspect of Algorithm 2 is that we repeatedly solve the “covering problem,” system (9), which tends to be much easier to solve than the relaxed master program (6) of Algorithm 1. On the other hand, Algorithm 2 tends to require more iterations than Algorithm 1. Moreover, Algorithm 2 does not have an upper bound that can be used to terminate with an ϵ -optimal solution. We note that inequalities (7) and (8) are not standard valid inequalities. Rather, they are guaranteed not to cut off all optimal solutions unless one has already been found and are termed supervalid inequalities [9]. While we do not detail it here, Israeli and Wood [9] describe an effective algorithm for solving model (1) that combines the key ideas of Algorithms 1 and 2.

STOCHASTIC SHORTEST-PATH NETWORK INTERDICTION

In this section, we generalize the deterministic interdiction model (1) to the stochastic setting. The source-terminus pair and the arc lengths that the follower faces are uncertain to the leader. Let $\xi = (c, d, s, t)$ denote these uncertain elements, and let $\omega \in \Omega$ index a finite set of realizations, $\xi^\omega = (c^\omega, d^\omega, s^\omega, t^\omega)$, that are assumed known to the leader only through a probability mass function, ϕ^ω , $\omega \in \Omega$. We assume that Ω is of modest size. The timing of the decisions and realizations of the uncertainty are as follows. First, the leader interdicts some subset of arcs via $x \in X$. Then, a source-terminus pair and arc lengths are revealed, $\xi^\omega = (c^\omega, d^\omega, s^\omega, t^\omega)$. The follower solves what is now a deterministic problem to select a shortest s^ω - t^ω path, knowing the source, terminus, and the arc lengths dictated by ω and $x \in X$.

We formalize this generalization by extending model (3) to this stochastic setting

$$z^* = \max_{x \in X} \sum_{\omega \in \Omega} \phi^\omega h(x, \xi^\omega), \quad (10)$$

where we have also extended the definition of $h(\cdot)$ to

$$\begin{aligned}
 h(x, \xi^\omega) &= \min_y \sum_{(i,j) \in A} (c_{ij}^\omega + d_{ij}^\omega x_{ij}) y_{ij} \\
 \text{s.t.} \quad &\sum_{(i,j) \in FS(i)} y_{ij} - \sum_{(j,i) \in RS(i)} y_{ji} \\
 &= \begin{cases} 1 & i = s^\omega \\ 0 & i \in N \setminus \{s^\omega, t^\omega\} \\ -1 & i = t^\omega \end{cases} \quad (11) \\
 &y_{ij} \geq 0, (i,j) \in A.
 \end{aligned}$$

Parallel to our development above, we can reformulate model (10) as a large-scale MIP,

$$\begin{aligned}
 z^* &= \max_{x, \pi} \sum_{\omega \in \Omega} \phi^\omega \pi_{t^\omega}^\omega \\
 \text{s.t.} \quad &x \in X \\
 &\pi_j^\omega - \pi_i^\omega - d_{ij}^\omega x_{ij} \leq c_{ij}^\omega, \\
 &\quad (i,j) \in A, \omega \in \Omega \\
 &\pi_{s^\omega}^\omega = 0, \omega \in \Omega,
 \end{aligned} \quad (12)$$

and attempt to solve model (12) directly using an MIP solver.

The same potential shortcomings that we describe above apply to direct solution of model (12) and so we extend Algorithm 1 to the stochastic setting. The master program is

$$\begin{aligned}
 \max_{x, \theta} \quad &\sum_{\omega \in \Omega} \phi^\omega \theta^\omega \\
 \text{s.t.} \quad &\theta^\omega \leq \sum_{(i,j) \in P(\ell, \omega)} (c_{ij}^\omega + d_{ij}^\omega x_{ij}), \\
 &\ell \in L^\omega, \quad \omega \in \Omega \\
 &x \in X.
 \end{aligned} \quad (13)$$

Extending our notation from the previous section, $P^{(\ell, \omega)}$, $\ell \in L^\omega$, denotes all of the s^ω - t^ω paths, $\omega \in \Omega$. If $L^\omega = L^\omega$, then model (13) is the full master program that is equivalent to model (10) with optimal value z^* . If $L^\omega \subset L^\omega$,

$\omega \in \Omega$, then model (13) has only a subset of the cut constraints and is a relaxed master program, whose optimal value we denote $\bar{z}$. Algorithm 3 iteratively grows the sets L^ω

by alternately solving model (13) to obtain $\bar{x}$ and solving the collection of shortest-path problems (11) indexed by $\omega \in \Omega$ with $x = \bar{x}$.

Algorithm 3. Benders' decomposition algorithm for model (10).

Input: an instance of model (10), error tolerance $\epsilon \geq 0$
Output: interdiction plan $\bar{x}$ with $z^* - \sum_{\omega \in \Omega} \phi^\omega h(\bar{x}, \xi^\omega) \leq \epsilon$

for $\omega \in \Omega$ **do**
 find at least one s^ω - t^ω path; initialize L^ω and program (13)'s set of cuts
end for
 $\underline{z} \leftarrow 0$
repeat
 solve model (13) to obtain $\hat{x}$ and $\bar{z}$
 for $\omega \in \Omega$
 solve model (11) with $x = \hat{x}$ to obtain s^ω - t^ω path P^ω and $h(\hat{x}, \xi^\omega)$
 update L^ω and model (13)'s set of cuts with path P^ω
 end for
 if $\sum_{\omega \in \Omega} \phi^\omega h(\hat{x}, \xi^\omega) > \underline{z}$ **then**
 $\bar{x} \leftarrow \hat{x}$; $\underline{z} \leftarrow \sum_{\omega \in \Omega} \phi^\omega h(\hat{x}, \xi^\omega)$
 end if
until $(\bar{z} - \underline{z} \leq \epsilon)$
return $\bar{x}$ and $\underline{z} = \sum_{\omega \in \Omega} \phi^\omega h(\bar{x}, \xi^\omega)$

Pan and Morton [28] report that Algorithm 3 outperforms direct application of an MIP solver to a variant of model (12). This contrasts with the situation in the deterministic setting because of the stochastic nature of the problem. Specifically, when x is fixed we can solve model (12) by solving $|\Omega|$ separate shortest-path problems. Consistent with application of the decomposition algorithm to the deterministic problem, Pan and Morton [28] find that valid inequalities and generating multiple cuts (for each ω) at each iteration can significantly improve the decomposition algorithm's performance. When we do not employ the valid inequalities that we discuss next, the linear-programming relaxation of the master program can be weak, precluding solution of all but the smallest of problem instances.

For simplicity, we again assume that interdiction of an arc amounts to effectively removing that arc from the network, and that there is an s^ω - t^ω path for all $\omega \in \Omega$ and $x \in X$. Let h_ℓ^ω be the length of the path associated with cut constraint $\ell \in L^\omega$ in master program (13) at some iteration of Algorithm 3. Let $T^\omega = \{\ell_1, \ell_2, \dots, \ell_k\} \subset L^\omega$,

where these indices satisfy the ordering

$$\min_{\ell \in L^\omega} h_\ell^\omega = h_{\ell_1}^\omega \leq h_{\ell_2}^\omega \leq \dots \leq h_{\ell_k}^\omega \leq h_{\ell_{k+1}}^\omega \equiv \bar{z}^\omega, \quad (14)$$

where $\bar{z}^\omega$ is an upper bound on the length of follower ω 's shortest path under an optimal interdiction plan. A *step inequality* is of the following form

$$\theta^\omega \leq h_{\ell_1}^\omega + \underbrace{(h_{\ell_2}^\omega - h_{\ell_1}^\omega)v_{\ell_1}^\omega + (h_{\ell_3}^\omega - h_{\ell_2}^\omega)v_{\ell_2}^\omega + \dots + (h_{\ell_{k+1}}^\omega - h_{\ell_k}^\omega)v_{\ell_k}^\omega}_{\sum_{\ell_i \in T^\omega} (h_{\ell_{i+1}}^\omega - h_{\ell_i}^\omega)v_{\ell_i}^\omega}, \quad (15)$$

where

$$v_\ell^\omega \leq \sum_{(i,j) \in P^{(\ell,\omega)}} x_{ij}, \ell \in T^\omega \quad (16a)$$

$$0 \leq v_\ell^\omega \leq 1, \ell \in T^\omega. \quad (16b)$$

Consider path $P^{(\ell_1,\omega)}$, that is, follower ω 's shortest path (among the paths in L^ω), which has length $h_{\ell_1}^\omega$. The idea of the step inequality

is that if the leader fails to interdict at least one arc on $P^{(\ell_1, \omega)}$ then the follower's path will have length $\theta^\omega = h_{\ell_1}^\omega$. If at least one arc on $P^{(\ell_1, \omega)}$ is interdicted, then θ^ω telescopes to $h_{\ell_2}^\omega$, unless an arc on $P^{(\ell_2, \omega)}$ is also interdicted, and so on.

Given a solution to the LP relaxation of model (13), we can efficiently identify the most violated step inequality for each $\omega \in \Omega$. This amounts to finding the index set $T^\omega \subset L^\omega$ that minimizes the right-hand side of inequality (15) with x , and hence v via inequality (16a), fixed. This can be accomplished by solving a separation problem involving a shortest-path problem on an acyclic network whose nodes correspond to elements of L^ω [28]. Addition of the step inequalities can dramatically tighten the LP relaxation of model (13) and reduce the running time of Algorithm 3 [28]. Step inequalities still apply when we relax the assumption that an interdicted arc is removed from the network. In this case, we must replace the index set, $P^{(\ell, \omega)}$, on the right-hand side of inequality (16a) with the set of uninterdicted arcs on $P^{(\ell, \omega)}$.

Remarks.

1. That (s^ω, t^ω) , $\omega \in \Omega$, is assumed to follow a probability distribution is an important *modeling* choice. It contrasts sharply with the way we model the follower selecting a path in the network. Instead of the worst-case path-selection model that we consider, the follower could instead take a random walk from source to terminus [29,30]. So, we can conceive four different interdiction models based on whether the follower selects a source-terminus pair and a path, respectively, in a worst-case or random manner.
2. Randomness in d_{ij} can model uncertainty in the leader's ability to lengthen arc (i, j) . For example, d_{ij} can be a binary random variable in which an interdiction can fail, and d_{ij} takes value zero, or succeed, and d_{ij} takes some lengthening value. Such models of stochastic binary effectiveness of

interdiction attempts in maximum-flow interdiction have also been developed [12,13].

3. Algorithm 3 is a multicut version of Benders' decomposition. A single-cut variant, in which cuts are aggregated across $\omega \in \Omega$, has a smaller master program with fewer cut constraints and fewer decision variables. But, the single-cut algorithm typically requires more iterations than its multicut counterpart. For more on multicut and single-cut algorithms, see Birge and Louveaux [31].
4. In implementing Algorithm 3, when we update L^ω and program (13)'s set of cuts, we should ensure that we do not add a redundant cut. In general, this can be done by seeing whether the current pair (x, θ^ω) obtained when solving model (13), is "cut off," that is, violates, the proposed cut. However, in the current setting, it is perhaps preferable to simply ensure that the associated path P^ω is indeed new.
5. The path-covering inequalities (7) used in the deterministic case differ from the step inequalities (15) and (16) we describe here. That said, both share the key idea of conditioning on whether at least one arc on an attractive path for the follower is interdicted and are related to the star inequalities of Atamtürk *et al.* [32].

INTERDICTING A MAXIMUM-RELIABILITY PATH: KNOWN s - t PAIR

Suppose the follower seeks an s - t path in G that maximizes the probability of evading detection. If the follower traverses arc $(i, j) \in A$, the probability of evading detection on that arc, by the leader's *indigenous* sensing capability, is p_{ij} . The leader can augment the indigenous capability by interdicting that arc, that is, installing a sensor on $(i, j) \in A$, and reducing the evasion probability to q_{ij} , where $0 < q_{ij} < p_{ij} \leq 1$. Assuming detection events on distinct arcs in the network are mutually independent, and that the leader's sensors are transparent to the follower, we

can formulate this interdiction model as

$$\min_{x \in X} \left[\begin{array}{ll} \max_P & \prod_{(i,j) \in P} [p_{ij}(1-x_{ij}) + q_{ij}x_{ij}] \\ \text{s.t.} & P \text{ is an } s\text{-}t \text{ path} \end{array} \right]. \quad (17)$$

The notation in model (17) is identical to that developed above with the addition of the evasion probabilities p_{ij} and q_{ij} just introduced and the fact that the inner maximization over P is over all s - t paths. Carrying out a standard logarithmic transformation [33] we obtain the following objective function:

$$\begin{aligned} & \ln \left(\prod_{(i,j) \in P} [p_{ij}(1-x_{ij}) + q_{ij}x_{ij}] \right) \\ &= \sum_{(i,j) \in P} \ln [p_{ij}(1-x_{ij}) + q_{ij}x_{ij}] \\ &= \sum_{(i,j) \in P} [\ln(p_{ij})(1-x_{ij}) + \ln(q_{ij})x_{ij}] \\ &= \sum_{(i,j) \in P} [\ln(p_{ij}) + \ln(q_{ij}/p_{ij})x_{ij}], \end{aligned}$$

where the second inequality holds only because x_{ij} is binary.

So, model (17) is equivalent to

$$\min_{x \in X} \left[\begin{array}{ll} \max_y & \sum_{(i,j) \in A} [\ln(p_{ij}) + \ln(q_{ij}/p_{ij})x_{ij}] y_{ij} \\ \text{s.t.} & \sum_{(i,j) \in FS(i)} y_{ij} - \sum_{(j,i) \in RS(i)} y_{ji} \\ &= \begin{cases} 1 & i = s \\ 0 & i \in N \setminus \{s, t\} \\ -1 & i = t \end{cases} \\ & y_{ij} \geq 0, (i,j) \in A \end{array} \right]. \quad (18)$$

Model (18), in turn, is equivalent to model (1), with $\max_x \min_y$, where $c_{ij} = -\ln(p_{ij})$ and $d_{ij} = -\ln(q_{ij}/p_{ij})$.

INTERDICTION A MAXIMUM-RELIABILITY PATH: STOCHASTIC s - t PAIR

The previous section shows that interdicting a maximum-reliability path with a known s - t pair can be captured by our deterministic shortest-path interdiction model (1). It

is natural to then hope that interdicting a maximum-reliability path, where s - t is uncertain but governed by a known probability mass function, can be captured by our stochastic shortest-path interdiction model (1). Unfortunately, as we see shortly, another modeling approach is required in the case of a stochastic s - t pair.

Analogous to our development in the section titled “Stochastic Shortest-Path Network Interdiction,” let $\xi = (p, q, s, t)$ denote the uncertain elements and let $\omega \in \Omega$ index a finite set of realizations, $\xi^\omega = (p^\omega, q^\omega, s^\omega, t^\omega)$, known to the leader only through a probability mass function, ϕ^ω , $\omega \in \Omega$. First, the leader installs sensors on a subset of arcs via $x \in X$. Then, a source-terminus pair and evasion probabilities are revealed, $\xi^\omega = (p^\omega, q^\omega, s^\omega, t^\omega)$. The follower solves what is now a deterministic problem to select an s^ω - t^ω path with maximum evasion probability, knowing the source, terminus, and per-arc evasion probabilities as dictated by ω and by the locations of the sensors, x .

We formulate this generalization as

$$\min_{x \in X} \sum_{\omega \in \Omega} \phi^\omega h(x, \xi^\omega), \quad (19)$$

where we now define

$$\begin{aligned} h(x, \xi^\omega) &= \max_P \prod_{(i,j) \in P} [p_{ij}^\omega(1-x_{ij}) + q_{ij}^\omega x_{ij}] \\ \text{s.t. } & P \text{ is an } s^\omega\text{-}t^\omega \text{ path.} \end{aligned} \quad (20)$$

Here, $h(x, \xi^\omega)$ is the follower’s *conditional* evasion probability, conditioned on ξ^ω . Hence, the objective function in model (19) is the (unconditional) probability the follower evades detection, and that is what the leader minimizes.

When $|\Omega| = 1$ we know that model (19) is equivalent to model (1) via the logarithmic transformation described in the previous section. Unfortunately, this fails when $|\Omega| > 1$. In other words, we cannot replace model (20) with the inner maximization in model (18), at least when trying to solve model (19). In short, the reason is that for positive functions, while $\min_x f(x)$ is equivalent to $\min_x \ln(f(x))$, $\min_x [f_1(x) + f_2(x)]$ is *not* equivalent to $\min_x [\ln(f_1(x)) + \ln(f_2(x))]$.

The following model is equivalent to model (19):

$$\min_{x \in X} \sum_{\omega \in \Omega} \phi^\omega \exp[\ln(h(x, \xi^\omega))]. \quad (21)$$

And, $h(x, \xi^\omega)$ can be represented by a shortest-(or rather longest-) path formulation, as in the inner maximization in model (18), or via its dual. Using the latter, the following is the analog of model (12) for interdicting a stochastic maximum-reliability path

$$\begin{aligned} z^* = \min_{x, \pi} \quad & \sum_{\omega \in \Omega} \phi^\omega e^{\pi t^\omega} \\ \text{s.t.} \quad & x \in X \\ & \pi_j^\omega - \pi_i^\omega - \ln(q_{ij}^\omega/p_{ij}^\omega)x_{ij} \\ & \geq \ln(p_{ij}^\omega), (i, j) \in A, \omega \in \Omega \\ & \pi_{s^\omega}^\omega = 0, \omega \in \Omega. \end{aligned} \quad (22)$$

Model (22) is a nonlinear MIP. Its continuous relaxation is a convex nonlinear program, and so it may be computationally tractable. Still, we much prefer a linear MIP, and so we pursue another approach.

In a *generalized* network-flow model, flow on an arc can experience a gain or loss. If an arc's gain factor is 0.9 and 0.8 units of flow are initiated at that arc's tail, then 0.72 units depart its head. The following recasting of $h(\cdot)$, as defined in model (20), uses this idea:

$$\begin{aligned} h(x, \xi^\omega) \\ = \max_{y, z} y_{t^\omega} \end{aligned} \quad (23a)$$

$$\text{s.t.} \quad \sum_{(i, j) \in FS(s^\omega)} (y_{ij} + z_{ij}) = 1 \quad (23b)$$

$$\begin{aligned} \sum_{(i, j) \in FS(i)} (y_{ij} + z_{ij}) - \sum_{(j, i) \in RS(i)} (p_{ji}^\omega y_{ji} + q_{ji}^\omega z_{ji}) = 0, \\ i \in N \setminus \{s^\omega, t^\omega\} \end{aligned} \quad (23c)$$

$$y_{t^\omega} - \sum_{(j, i) \in RS(t^\omega)} (p_{ji}^\omega y_{ji} + q_{ji}^\omega z_{ji}) = 0 \quad (23d)$$

$$0 \leq y_{ij} \leq 1 - x_{ij}, \quad (i, j) \in A \quad (23e)$$

$$0 \leq z_{ij} \leq x_{ij}, \quad (i, j) \in A. \quad (23f)$$

We view each arc that can receive a sensor as two arcs in parallel, and if $x_{ij} = 1$ then flow may occur only on the “sensor” arc via z_{ij} , and otherwise only on the “no sensor” arc via y_{ij} . Flow on arc (i, j) is multiplied by that arc's gain factor, either p_{ij}^ω or q_{ij}^ω . Constraint (23b) forces one unit of flow out of s^ω , flow is conserved at all intermediate nodes in constraint (23c), and constraint (23d) defines the flow that reaches t^ω as y_{t^ω} . That value is maximized in the objective function (23a). Flow occurs on the appropriate arc due to the leader's decision variable x_{ij} in constraints (23e) and (23f). We note that the upper bound in the latter constraint can be dropped since $p_{ij}^\omega > q_{ij}^\omega$.

Model (19) with h defined in model (23) now has a linear subproblem but is in min-max form. We have dealt with this in the sections above by taking the dual of the inner linear program. Attempting to do so here leads to a new complication. Namely, the resulting model is *not* a linear MIP because dual variables on constraint (23e) multiply x_{ij} . The latter variables are binary and hence these nonlinear terms can be linearized [6]. That said, we view what follows as a more attractive approach, rooted in an exact-penalty result [34, Lemma 2]:

Theorem 1. *Suppose the following linear program has a finite optimal solution:*

$$\begin{aligned} h(x) = \max_y \quad & fy \\ \text{s.t.} \quad & Dy \leq d \\ & 0 \leq y \leq U(e - x) : \pi(x). \end{aligned} \quad (24)$$

Here, e is the vector of all ones, $U = \text{diag}(u)$, where $u = (u_1, \dots, u_n)$, and the domain of h is $X = \{x : Rx \leq r, x \in \{0, 1\}^n\}$. Assume $\bar{\pi}$ satisfies $\bar{\pi} \geq \pi^*(x)$, for all $x \in X$, where $\pi^*(x)$ is an optimal dual (sub)vector to model (24). Let $\bar{\Pi} = \text{diag}(\bar{\pi})$ and consider

$$\begin{aligned} \bar{h}(x) = \max_y \quad & (f - x^T \bar{\Pi})y \\ \text{s.t.} \quad & Dy \leq d \\ & 0 \leq y \leq u. \end{aligned}$$

Then $h(x) = \bar{h}(x)$ for $x \in X$.

Using Theorem 1, and the fact that $p_{ij}^\omega - q_{ij}^\omega, (i,j) \in A$, provides an upper bound on optimal dual prices for constraint (23e), we can rewrite $h(\cdot)$ from model (23) as follows, with the understanding that x is restricted to X :

$$\begin{aligned}
& h(x, \xi^\omega) \\
&= \max_{y,z} y_{t^\omega} - \sum_{(i,j) \in A} (p_{ij}^\omega - q_{ij}^\omega) x_{ij} y_{ij} \\
&\text{s.t.} \quad \sum_{(i,j) \in FS(s^\omega)} (y_{ij} + z_{ij}) = 1 \\
&\quad \sum_{(i,j) \in FS(i)} (y_{ij} + z_{ij}) \\
&\quad - \sum_{(j,i) \in RS(i)} (p_{ji}^\omega y_{ji} + q_{ji}^\omega z_{ji}) \\
&\quad = 0, \quad i \in N \setminus \{s^\omega, t^\omega\} \\
&\quad y_{t^\omega} - \sum_{(j,i) \in RS(t^\omega)} (p_{ji}^\omega y_{ji} + q_{ji}^\omega z_{ji}) = 0 \\
&\quad y_{ij} \geq 0, z_{ij} \geq 0, \quad (i,j) \in A. \quad (25)
\end{aligned}$$

The constraints of model (25) differ from those of model (23) only in that the simple upper bounds in constraints (23e) and (23f) have been removed. Removal of the latter is discussed above. Removal of the former simple bound allows the follower to traverse a “no sensor” arc (y_{ij}) even if the leader installs a sensor on (i,j) via $x_{ij} = 1$. However, the follower has a new penalty in the objective function of model (25) that is sufficiently large so that there is no incentive to do so. Theorem 1 ensures that the two formulations are equivalent. The advantage of the reformulation (25) is that we can now follow the three-step process we have successfully employed before, that is, (i) fix $x \in X$, (ii) take the dual of the inner LP, and (iii) release x . Doing so leads to the following *linear* MIP:

$$\begin{aligned}
& \min_{x,\pi} \sum_{\omega \in \Omega} \phi^\omega \pi_{s^\omega}^\omega \\
& \text{s.t.} \quad x \in X \\
& \quad \pi_i^\omega - p_{ij}^\omega \pi_j^\omega + (p_{ij}^\omega - q_{ij}^\omega) x_{ij} \geq 0, \quad (i,j) \in A, \omega \in \Omega \quad (26) \\
& \quad \pi_i^\omega - q_{ij}^\omega \pi_j^\omega \geq 0, \quad (i,j) \in A, \omega \in \Omega \\
& \quad \pi_{t^\omega} = 1, \quad \omega \in \Omega.
\end{aligned}$$

The dual variables, π_i , in our shortest-path formulation (2) had the interpretation of being the length of the shortest s - i path. In model (26), the dual variables instead represent the conditional evasion probability given that the follower has successfully reached node i without being detected. If $x_{ij} = 1$ then $p_{ij}^\omega - q_{ij}^\omega$ is sufficiently large so that the first of the paired constraints in model (26) defining π_i^ω is vacuous.

While it required a bit more effort than the development in the previous sections, we have achieved a large-scale MIP formulation for interdicting a maximum-reliability path when the follower’s s - t pair is uncertain. We could go on to describe a decomposition algorithm for model (26) and to develop step inequalities for that formulation. Such a development largely parallels that of the section titled “Stochastic Shortest-Path Network Interdiction.” We do not detail it here, and instead refer the reader to Pan and Morton [28].

Remarks.

1. We could use ideas behind Kelley’s cutting-plane method [35], which is closely related to Benders’ decomposition and the L-shaped method, to construct a decomposition algorithm for the nonlinear MIP of model (22), in which the master program is a linear MIP and there are $|\Omega|$ shortest-path subproblems. However, the linear MIP of model (26) can also be decomposed using Benders’ decomposition, and its master program is at least as strong.
2. Often it may be more natural to place a sensor on a node, for example, at a security checkpoint, rather than on an arc. While alternatives are possible, this can be captured in our arc-based formulation by using a standard transformation that splits the node into two nodes and adds an arc to represent travel through the physical node.
3. Theorem 1 says that $h(x) = \bar{h}(x)$. Yet, we know $h(\cdot)$ is concave and $\bar{h}(\cdot)$ is convex on the convex hull of X . (In general, neither function is linear.) This apparent contradiction is resolved by noting

that Theorem 1 only ensures equality of h and $\bar{h}$ at binary values of x permitted by $x \in X$. The functions can differ at fractional points in the convex hull of X .

4. Pushing the previous remark further, model (19), with $h(\cdot)$ defined in model (23), involves minimization of a function that is concave on the convex hull of X . Such a function is not amenable to an outer-linearization decomposition algorithm. In contrast, model (19), with $h(\cdot)$ defined in model (25) involves minimization of a convex function on X 's convex hull, and this model lends itself to Benders' decomposition. This observation is consistent with the fact that we could directly reformulate the latter model into a single large-scale linear MIP.
5. Obtaining tight "big M " values in an integer program is key to avoiding a loose LP relaxation and hence key to computational tractability. The value $p_{ij}^\omega - q_{ij}^\omega$ in model (26) represents such a value, and it can be tightened to $\psi_j^\omega(p_{ij}^\omega - q_{ij}^\omega)$, where ψ_j^ω denotes an upper bound on the value of maximum-reliability path from j to t^ω . We can compute such a bound by finding all maximum-reliability paths for the system in which no sensors are installed. It is possible to further tighten ψ_j^ω , for example, at certain nodes in the branch-and-bound tree where some x_{ij} values have been fixed to one.

SUMMARY

We have described interdiction of shortest-path, and maximum-reliability-path, networks both with deterministic and stochastic source-terminus pairs. Duality plays a key role in reformulating a nested max-min (or, min-max) model as a nonnested MIP, at least when the leader's decision variables parameterize the objective function coefficients of the follower's model. When the leader's decision variables instead affect the right-hand side of the follower's linear

program, we made use of an exact-penalty result, along with duality, to construct an optimization model appropriate for solution by a standard MIP solver. The resulting MIP can have a weak linear-programming relaxation, and so we employed Benders' decomposition, coupled with valid inequalities, to allow solution of large-scale instances.

We have pointed to a number of variants of our interdiction model. A robust optimization model arises when we exchange the order of the optimization operators in our initial max-min model. Our model is a bilevel Stackelberg game in which the leader's decisions are transparent to the follower. When instead the leader's decisions are concealed, a two-person zero-sum Cournot game arises. We assume that the source-terminus pair is governed by a known probability mass function, but we could instead have an adversarial formulation for this model primitive. Similarly, we assume that the follower optimizes to select a worst-case path (from the leader's perspective), but we could have assumed instead that the follower takes a random walk from the source to terminus. Another model variant we pointed to represents information asymmetries between the leader and follower regarding arc lengths. Still further model variants were not touched on in this article. For example, the leader may need to prioritize, that is, rank-order, interdiction decisions because of budget uncertainty or because the leader's activities are deployed over time [18,36,37].

Acknowledgments

This work has been supported by the National Science Foundation through grants CMMI-0653916 and CMMI-0800676, the Defense Threat Reduction Agency through grant HDTRA1-08-1-0029, and the US Department of Homeland Security under Grant Award Number 2008-DN-077-ARI021-03. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Homeland Security.

REFERENCES

1. Watson JP, Murray R, Hart WE. Formulation and optimization of robust sensor placement problems for drinking water contamination warning systems. *J Infrastructure Syst* 2009;15:330–339.
2. Armbruster B, Smith JC, Park K. The optimization of packet filter placements to combat distributed denial of service attacks. *Eur J Oper Res* 2007;176:1283–1292.
3. Baldick R, Salmerón J, Wood RK. Analysis of electric grid security under terrorist threat. *IEEE Trans Power Syst* 2004;19:905–912.
4. Bienstock D, Verma A. The $N - k$ problem in power grids: New models, formulations and numerical experiments. Working paper. Industrial Engineering and Operations Research, Columbia University; 2009. Available at <http://www.columbia.edu/~dano/papers/papers.html> Accessed August 6, 2010.
5. Brown GG, Carlyle M, Salmerón J, *et al.* Defending critical infrastructure. *Interfaces* 2006;36:530–544.
6. Morton DP, Pan F, Saeger KJ. Models for nuclear smuggling interdiction. *IIE Trans Oper Eng* 2007;38:3–14.
7. Wein LM, Wilkins AH, Baveja M, *et al.* Preventing the importation of illicit nuclear materials in shipping containers. *Risk Anal* 2006;26:1377–1393.
8. Brown GG, Carlyle WM, Harney R, *et al.* Interdicting a nuclear-weapons project. *Oper Res* 2009;57:866–877.
9. Israeli E, Wood RK. Shortest-path network interdiction. *Networks* 2002;40:97–111.
10. Phillips CA. The network inhibition problem. Proceedings of the 25th Annual ACM Symposium on the Theory of Computing. San Diego: 1993. pp. 776–785.
11. Wood RK. Deterministic network interdiction. *Math Comput Model* 1993;17:1–18.
12. Cormican K, Morton DP, Wood RK. Stochastic network interdiction. *Oper Res* 1998;46:184–197.
13. Janjarassuk U, Linderoth JT. Reformulation and sampling to solve a stochastic network interdiction problem. *Networks* 2008;52:120–132.
14. Salmerón J, Wood RK, Baldick R. Worst-case interdiction analysis of large-scale electric power grids. *IEEE Trans Power Syst* 2009;24:96–104.
15. Fulkerson DR, Harding GC. Maximizing the minimum source-sink path subject to a budget constraint. *Math Program* 1977;13:116–118.
16. Golden B. A problem in network interdiction. *Nav Res Log Q* 1978;25:711–713.
17. Washburn AR, Wood RK. Two-person zero-sum games for network interdiction. *Oper Res* 1994;43:243–251.
18. Nehme MV. Two-person games for stochastic network interdiction: models, methods, and complexities [PhD thesis]. The University of Texas at Austin. 2009.
19. Bayrak H, Bailey MD. Shortest path network interdiction with asymmetric information. *Networks* 2008;52:133–140.
20. Bertsimas DJ, Sim M. Robust discrete optimization and network flows. *Math Program* 2003;98:49–71.
21. Corley HW, Sha DY. Most vital links and nodes in weighted networks. *Oper Res Lett* 1982;1:157–160.
22. Malik K, Mittal AK, Gupta SK. The k -most vital arcs in the shortest path problem. *Oper Res Lett* 1989;8:223–227.
23. Ball MO, Golden BL, Vohra RV. Finding the most vital arcs in a network. *Oper Res Lett* 1989;8:73–76.
24. Bar-Noy A, Khuller S, Schieber B. The complexity of finding most vital arcs and nodes. Technical Report CS-TR-35-39. Computer Science Department, University of Maryland. 1995.
25. Bailey MD, Shechter SM, Schaefer AJ. SPAR: stochastic programming with adversarial recourse. *Oper Res Lett* 2006;34:307–315.
26. Benders JF. Partitioning procedures for solving mixed-variables programming problems. *Numer Math* 1962;4:238–252.
27. Van Slyke RM, Wets RJ-B. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM J Appl Math* 1969;17:638–663.
28. Pan F, Morton DP. Minimizing a stochastic maximum-reliability path. *Networks* 2008;52:111–119.
29. Berman O, Krass D, Xu CW. Locating flow-intercepting facilities: new approaches and results. *Ann Oper Res* 1995;60:121–143.
30. Gutfraind A, Hagberg A, Pan F. Optimal interdiction of unreactive Markovian evaders. Proceedings of the 6th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. Berlin: Springer; 2009. pp. 102–116.

31. Birge JR, Louveaux FV. A multicut algorithm for two-stage stochastic linear programs. *Eur J Oper Res* 1988;34:384–392.
32. Atamtürk A, Nemhauser GL, Savelsbergh MWP. The mixed vertex packing problem. *Math Program* 2000;89:35–53.
33. Ahuja RK, Magnanti TL, Orlin JB. *Network flows*. Upper Saddle River (NJ): Prentice Hall; 1993.
34. Morton DP, Wood RK. Restricted-recourse bounds for stochastic linear programming. *Oper Res* 1999;47:943–956.
35. Kelley JE. The cutting plane method for solving convex programs. *SIAM J Ind Appl Math* 1960;8:703–712.
36. Hochbaum DS. Dynamic evolution of economically preferred facilities. *Eur J Oper Res* 2009;193:649–659.
37. Michalopoulos DP. Prioritization and optimization in stochastic network interdiction problems [PhD thesis]. The University of Texas at Austin. 2008.

STOCHASTIC OPTIMAL CONTROL FORMULATIONS OF DECISION PROBLEMS

KUNAL SRIVASTAVA
DUŠAN M. STIPANOVIĆ
Department of Industrial and
Enterprise Systems Engineering,
University of Illinois at
Urbana-Champaign, Urbana,
Illinois

This article presents some of the most commonly used stochastic optimal control formulations. Stochastic optimal control has proved to be a valuable tool for sequential decision making under uncertainty. In fairly general terms, a stochastic optimal control problem consists of a system state equation which models the evolution of the system's trajectories under the control action in the presence of noise and a cost function which assigns a real-valued cost to the state and control trajectory over a time horizon. The time horizon associated with the optimal control problem is the interval of interest over which we need to optimize the cost function. Different formulations of the optimal control problem deal with finite, infinite, or random time horizons.

A crucial aspect of the sequential decision making problems is the fact that a greedy method of choosing a control action which minimizes present cost at each time instant may not be an optimal policy since such a choice might affect the system state so that future control actions will yield high cost. Thus, we need to balance the trade-off between finding a low present cost and the undesirability of a high future cost. The dynamic programming principle (in the section titled "Dynamic Programming," this encyclopedia) is the central idea which is used to solve such problems. Stochastic optimal control problems in discrete time are also known as *Markov decision processes*. See the section titled "Markov Decision

Processes" in this encyclopedia for further details on a discrete time formulation. We restrict ourselves to the continuous time formulations when the state evolution is the solution of a stochastic differential equation (SDE). In this case, the dynamic programming principle yields a partial differential equation which characterizes the optimal control trajectory. This partial differential equation is known as the Hamilton–Jacobi–Bellman (HJB) equation of optimal control.

The rest of the article is organized as follows: in the section titled "Preliminaries", we give a brief introduction to the mathematical setup behind the problem formulation and introduce key concepts which are useful in stating the results later. In the section titled "Formulation of the Optimal Control Problem", we give a classification of the optimal control problem in terms of the different classes of admissible control. The case of control under full state information is dealt with in the section titled "Optimal Control under Full State Information" and control under partial information is dealt with in the next section. Finally, we provide the conclusion and suggest some further reading.

PRELIMINARIES

In this section, we review some basic facts about SDEs which form the foundation of the analysis to follow. Our presentation closely follows the approach taken in Ref. 1. An SDE is an equation of the form

$$\begin{aligned}\frac{dX_t}{dt} &= b(t, X_t) + \sigma(t, X_t)W_t \\ X_t &\in \mathbb{R}^n, b(t, x) \in \mathbb{R}^n, \sigma(t, x) \in \mathbb{R}^{n \times m},\end{aligned}\quad (1)$$

where W_t is a vector whose each component is a *white noise*. An equivalent representation of the SDE is given in the following form:

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dB_t, \quad (2)$$

where B_t is an m -dimensional Brownian motion and can be heuristically interpreted as $W_t = \frac{dB_t}{dt}$. This is just a formal representation as the sample paths of B_t are nowhere differentiable. The term $b(t, x)$ is known as the *drift coefficient* and $\sigma(t, x)$ is the *diffusion coefficient*. This is known as the Itô interpretation of Equation (1). Also, a solution X_t of the SDE satisfies the integral form of the equation, that is,

$$X_t = X_0 + \int_0^t b(t, X_\tau) d\tau + \int_0^t \sigma(t, X_\tau) dB_\tau.$$

The solution X_t is known as an Itô diffusion. Let $\mathcal{F}_t = \sigma(B_s : s \leq t)$. Then X_t is known to be adapted to $\mathcal{F}_t$. The expression $I_t = \int_0^t f(\tau, \omega) dB_\tau$ is known as an *Itô integral*. It is well defined for a *predictable* or *nonanticipating* [1] stochastic process $f(t, \omega)$. The Itô integral has the following two useful properties:

1. *Martingale Property*. The integral has a continuous version and is a martingale, that is

$$\mathbb{E}(I_t | \mathcal{F}_s) = I_s, \quad s \leq t.$$

This implies that $\mathbb{E}[\int_0^t f(\tau, \omega) dB_\tau] = 0$.

2. *Itô Isometry*.

$$\begin{aligned} & \mathbb{E} \left[\left(\int_0^T f(t, \omega) dB_t \right)^2 \right] \\ &= \mathbb{E} \left[\int_0^T f(t, \omega)^2 dB_t \right]. \end{aligned}$$

An Itô process is a stochastic process which has a representation of the following form:

$$\begin{aligned} dX_t &= u(t, \omega) dt + v(t, \omega) dB_t \\ X_t &\in \mathbb{R}^n, u(t, \omega) \in \mathbb{R}^n, v(t, \omega) \in \mathbb{R}^{n \times m}, \end{aligned}$$

where B_t is an m -dimensional Brownian motion. An important result is the Itô's formula, which states that for a given, twice-differentiable function $g(t, x) \in C^2([0, \infty) \times \mathbb{R}^n)$, the process

$Y_t = g(t, X_t)$ is also an Itô process whose k th coordinate has a representation

$$\begin{aligned} dY_k &= \frac{\partial g_k}{\partial t}(t, X) dt + \sum_i \frac{\partial g_k}{\partial x_i}(t, X) dX_i \\ &+ \frac{1}{2} \sum_{i,j} \frac{\partial^2 g_k}{\partial x_i \partial x_j}(t, X) dX_i dX_j, \end{aligned}$$

where $dB_i dB_j = \delta_{ij} dt$, $dB_i dt = dt dB_i = 0$. The Itô's formula generalizes the chain rule of elementary calculus as can be seen by substituting $v = 0$ in the above equation. Another important concept is the notion of an *infinitesimal generator* of an Itô diffusion. If X_t is a time-homogeneous Itô diffusion, then the infinitesimal generator A of the process X_t is defined as

$$Af(x) = \lim_{t \rightarrow 0} \frac{\mathbb{E}^x[f(X_t)] - f(x)}{t},$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Here, we have used the notation $\mathbb{E}^x[f(X_t)] = \mathbb{E}[f(X_t) | X_0 = x]$. After applying Itô's formula to $f(X_t)$, where X_t satisfies the homogeneous equation

$$\begin{aligned} dX_t &= b(X_t) dt + \sigma(X_t) dB_t \\ X_t &\in \mathbb{R}^n, b(x) \in \mathbb{R}^n, \sigma(x) \in \mathbb{R}^{n \times m} \end{aligned}$$

with $X_0 = x$ we arrive at the following:

$$\begin{aligned} f(X_t) - f(x) &= \int_0^t \left(\sum_{i=1}^n b_i(X_s) \frac{\partial f}{\partial x_i} \right. \\ &+ \frac{1}{2} \sum_{i,j} [\sigma \sigma^T]_{ij}(X_s) \frac{\partial^2 f}{\partial x_i \partial x_j} \Big) ds \\ &+ \sum_{i,k} \int_0^t \sigma_{ik} \frac{\partial f}{\partial x_i} dB_k \end{aligned}$$

On taking expectations and letting $t \rightarrow 0$, we arrive at the action of the infinitesimal generator of the diffusion as

$$Af(x) := \langle \nabla f(x), b(x) \rangle + \frac{1}{2} \text{tr}(\sigma(x) \sigma^T(x) \nabla^2 f(x)),$$

where ∇ denotes the gradient operator and ∇^2 is the Hessian matrix. Infinitesimal generators arise naturally in the optimal control problem and serve as an important link

between problems in stochastic control and partial differential equations.

FORMULATION OF THE OPTIMAL CONTROL PROBLEM

In this section, we provide a classification of the optimal control problems based on the information available to the controller. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with the filtration $\{\mathcal{F}_t\}$ and an m -dimensional $\{\mathcal{F}_t\}$ -Brownian motion $\{B_t\}$. $\mathcal{F}_t$ denotes the total information available to the controller at time $t \geq 0$. The state evolution is given by the following SDE:

$$\begin{aligned} dX_t &= dX_t^u = b(t, X_t, u_t) dt + \sigma(t, X_t, u_t) dB_t, \\ X_s &= x \in \mathbb{R}^n. \end{aligned}$$

The notation X_t^u refers to the state trajectory obtained when the control action is u_t . The control action u_t is restricted to the Borel set $U \subset \mathbb{R}^k$. Given the state evolution equation, an optimal control problem is characterized by specifying the class of *admissible control* $\mathcal{A}$ and the *cost function* $J(s, x, u)$. The different classes of admissible controls which are typically considered, are as follows:

Open-Loop Control. In this class, the control actions are functions of the form $u(t, \omega) = u(t)$. This specifies a deterministic control action which does not depend on any observation except maybe the initial condition $X_0 = x$.

Feedback Control. This class consists of control functions which are adapted to $\{\mathcal{F}_t^X\}$, where $\mathcal{F}_t^X$ is the σ -algebra generated by $\{X_r; s \leq r \leq t\}$; that is at each time t , the control $u(t)$ is a function of the observed trajectory $X([s, t])$. In the context of game theory, such control strategies are also known as *behavioral strategies*.

Markov Control. This is a subclass of feedback control actions with the additional property that the control action is of the form $u(t, X(t))$; that is the control action is only dependent on the current state. If further, we have that the control action is of the form $u(X(t))$ then

it is known as the *stationary Markov control*.

Partial Information. This class of controls refer to the case when the controller has partial information about the system state. In many instances, this partial information Z_t is modeled as an outcome of an SDE of the following form:

$$dZ_t = a(t, X_t) dt + \gamma(t, X_t) d\hat{B}_t.$$

The control action is required to be adapted to the σ -algebra generated by the observation process Z_t .

OPTIMAL CONTROL UNDER FULL STATE INFORMATION

In this section, we discuss various formulations where the controller has access to the state vector. As expected, the analysis of such problems tend to be simpler compared to problems when the controller has access to partial information about the state.

Infinite Horizon Discounted Cost Problem

Here, we provide an informal derivation of the HJB equation for the case of infinite horizon discounted cost problem. We assume that the state equation follows the homogeneous SDE

$$\begin{aligned} dX_t &= dX_t^u = b(X_t, u_t) dt + \sigma(X_t, u_t) dB_t, \\ X_0 &= x \in \mathbb{R}^n. \end{aligned}$$

The cost structure of the infinite horizon discounted time problem is given as

$$J(x, u) = \mathbb{E} \left[\int_0^\infty e^{-\gamma t} c(X_t, u_t) dt \mid X_0 = x \right].$$

Here $c(X_t, u_t)$ denotes the running cost and γ is a discount factor which is included to ensure integrability conditions on the cost and the value function. In certain economic applications, γ also has the interpretation of interest rates which is used to discount the cost to get its present value. A high value of

γ tends to make the control action myopic in nature since it discounts the future more heavily. The optimal control problem is to find the control action in the admissible set which minimizes the cost function. The optimal cost achieved is known as the *value function* $V(x)$ which has the following representation:

$$V(x) = \inf_{u \in \mathcal{A}} J(x, u).$$

We define the set of admissible controls $\mathcal{A}$ as the set of Markov controls. The *principle of optimality* can be used to arrive at the following equation which is satisfied by the optimal cost function:

$$V(x) = \inf_u \mathbb{E} \left[\int_0^t e^{-\gamma\tau} c(X_\tau, u(X_\tau)) d\tau + e^{-\gamma t} V(X_t) \right]. \quad (3)$$

Here, we make the simplifying assumption that there exists an optimal control u^* for which the value function is achieved, that is, $V(x) = J(x, u^*)$. The HJB equation is a partial differential equation which is obtained when we let the time approach zero. The derivation of the HJB equation for the discounted cost problem is as follows: By applying Itô's rule to the term $e^{-\gamma t} V(X_t)$, we arrive at

$$\begin{aligned} & \mathbb{E} \left[\int_0^t e^{-\gamma\tau} c(X_\tau, u(X_\tau)) d\tau + e^{-\gamma t} V(X_t) \right] \\ &= V(x) + \mathbb{E} \int_0^t e^{-\gamma\tau} (-\gamma V(X_s) + A^u V(X_\tau) \\ & \quad + c(X_\tau, u(X_\tau))) d\tau, \end{aligned}$$

where A^u is the infinitesimal generator

$$\begin{aligned} A^u V(x) &:= \langle \nabla V(x), b(x, u) \rangle \\ & \quad + \frac{1}{2} \text{tr}(\sigma(x, u) \sigma^T(x, u) \nabla^2 V(x)). \end{aligned}$$

We assume that $V(x)$ is twice differentiable so that the infinitesimal generator is well defined. Now from Equation (3) we arrive at the following equation which holds for all $u \in \mathcal{A}$:

$$\mathbb{E} \int_0^t e^{-\gamma\tau} (-\gamma V(X_s) + A^u V(X_\tau) + c(X_\tau, u(X_\tau))) d\tau \geq 0,$$

where equality holds with the control u^* . Now taking limit as $t \rightarrow 0$ and assuming that the limit and expectation are interchangeable, we get

$$-\gamma V(x) + A^u V(x) + c(x, u(x)) \geq 0,$$

where equality holds for $u = u^*$ or equivalently,

$$\inf_{u \in U} (-\gamma V(x) + A^u V(x) + c(x, u(x))) = 0. \quad (4)$$

This is the HJB partial differential equation for the infinite horizon discounted cost problem. This equation reduces the problem of the determination of an optimal control in the class of Markov policies to the solution of a partial differential equation. An important thing to note is that like the deterministic optimal control problem, the infinite horizon optimal control problem gives rise to optimal stationary Markov policies.

Optimal Control till a Stopping Time

For discussing the optimal control problem till a stopping time, we will assume that the system dynamics evolve as a nonhomogeneous SDE

$$dX_t = dX_t^u = b(t, X_t, u_t) dt + \sigma(t, X_t, u_t) dB_t,$$

with the initial condition $t_0 = s$ and $X_s = x$. Given the algebra generated by X_t as $\mathcal{F}_t^X$, a stopping time τ is an extended real-valued random variable which satisfies $\{\tau \leq t\} \in \mathcal{F}_t^X$ for all t . Let G be a fixed domain in $\mathbb{R} \times \mathbb{R}^n$ and τ_G be the first exit time after s by the process, that is,

$$\tau_G = \inf\{t > s; (t, X_t) \notin G\}.$$

It can be checked that τ_G is a stopping time. The optimal control problem is to minimize the cost function

$$\begin{aligned} J(s, x, u) &= \mathbb{E} \left[\int_s^{\tau_G} c(r, X_r) dr \right. \\ & \quad \left. + h(\tau_G, X_{\tau_G}) \chi_{\{\tau_G < \infty\}} \mid X_s = x \right], \end{aligned}$$

where $h(\tau_G, X_{\tau_G})$ is the terminal cost. It can be easily seen that by choosing $G = [0, T) \times \mathbb{R}^n$, the problem reduces to a finite time horizon problem with horizon T . Clearly, in this case, the value function is dependent on the initial time as well, that is,

$$V(s, x) = \inf_{u \in \mathcal{A}} J(s, x, u).$$

The HJB equation associated with this problem can be written as

$$\frac{\partial V(t, x)}{\partial t} + \inf_{u \in U} [A^u(t)V(t, x) + c(t, x, u)] = 0, \quad (5)$$

with the boundary condition $V(t, x) = h(t, x)$ for all $(t, x) \in \partial G$, where ∂G denotes the boundary of the set G . Various conditions under which the solution of the HJB equation is guaranteed to exist are given in Ref. 2. One of them is a *uniform ellipticity* condition which requires that there exists a k such that the following holds:

$$\sum_{i,j} (\sigma \sigma')_{i,j} \xi_i \xi_j \geq \|\xi\|^2 \quad \forall \xi \in \mathbb{R}^n.$$

This condition essentially means that noise affects every component of the SDE regardless of the coordinate system. We now present a *verification theorem* [2] which gives sufficient condition for optimality.

Theorem 1. *Let $V(s, x)$ be a solution of the HJB equation (Eq. 5) which satisfies the boundary condition mentioned above. If in addition the solution $V(s, x)$ has well-defined partial derivatives $V_s, V_{x_i}, V_{x_i x_j}$; is continuous on the closure $\bar{G}$; and satisfies the polynomial growth condition for some D, k such that $|V(t, x)| \leq D(1 + |x|^k)$ when $(t, x) \in G$, then we have*

1. $V(s, x) \leq J(s, x, u)$ for any admissible control u and initial condition $(s, x) \in G$;
2. if u^* is an admissible feedback control such that

$$\begin{aligned} & A^{u^*}(s)V(s, x) + c(s, x, u^*) \\ &= \min_{u \in U} [A^u(s)V(s, x) + c(s, x, u)] \end{aligned}$$

for all $(s, x) \in G$, then $V(s, x) = J(s, x, u^*)$ for all $(s, x) \in G$. Thus, u^* is optimal.

The following example is an interesting application of the HJB equation to the problem of portfolio selection. This problem was first formulated by Merton and is one of the few problems for which an analytical solution is available.

Example. Optimal Portfolio Selection [3]: Consider an investor whose wealth at time t is denoted by X_t . At each instant, the investor needs to make a decision about the fraction of wealth to invest in a stock which is denoted as $\alpha^1(t)$ and the amount of wealth to consume. The amount of wealth consumed by time t is denoted $\alpha^2(t)$. The unconsumed and uninvested stock can grow at a risk-free interest rate of r . The price of the risky asset (stock) denoted as $S(t)$, changes according to

$$dS = RS dt + \sigma S dW,$$

where r, R , and σ are constants and moreover, $R > r > 0$ with $\sigma \neq 0$. The evolution of the total wealth is given by the SDE

$$\begin{aligned} dX = & (1 - \alpha^1(t))Xr dt + \alpha^1(t)X(R dt + \sigma dW) \\ & - \alpha^2(t) dt. \end{aligned}$$

Define $G := \{(t, x) \mid 0 \leq t \leq T, x \geq 0\}$ and the stopping time τ is the first exit time from G . The control action is denoted by $u(t) = (\alpha^1(t), \alpha^2(t))$. The payoff functional to be maximized is

$$J(x, u) = \mathbb{E} \left[\int_t^\tau e^{-\gamma s} F(\alpha^2(s)) ds \right],$$

where F is a given utility function with $\gamma > 0$ as the discount rate. The HJB equation associated with the problem is given as

$$\begin{aligned} \frac{\partial V}{\partial t} + \max_{0 \leq a_1 \leq 1, a_2 \geq 0} & \times \{ (a_1 x \sigma)^2 / 2 V_{xx} + ((1 - a_1) x r \\ & + a_1 x R - a_2) V_x + e^{-\gamma t} F(a_2) \} = 0 \end{aligned}$$

with the boundary condition $V(0, t) = 0$, $V(x, T) = 0$. The maximum value can be easily computed to be

$$\alpha^{1*} = \frac{-(R-r)V_x}{\sigma^2 x V_{xx}}, \quad F'(\alpha^{2*}) = e^{\gamma t} V_x,$$

given the constraints $0 \leq \alpha^{1*} \leq 1, 0 \leq \alpha^{2*} \leq x$ are valid. Thus, if we know the value function V then we can compute the optimal control policies. It is assumed that the function F is of the form $F(\alpha) = \alpha^\theta$, $0 < \theta < 1$. We proceed by guessing the functional form of the value function as

$$V(x, t) = g(t)x^\theta$$

where $g(t)$ is some function which needs to be determined. The optimal control takes the form

$$\alpha^{1*} = \frac{R-r}{\sigma^2(1-\theta)}, \quad \alpha^{2*} = [e^{\gamma t} g(t)]^{1/(\theta-1)} x.$$

By substituting the value of the optimal control in the HJB equation along with the guessed form of the value function, we arrive at

$$(g'(t) + v\theta g(t) + (1-\theta)g(t)(e^{\gamma t} g(t))^{1/(\theta-1)} x^\theta) = 0, \quad (6)$$

where the constant v is given as

$$v := \frac{(R-r)^2}{2\sigma^2(1-\theta)} + r.$$

Equation (6) is an Ordinary Differential Equation (ODE) which can be solved for $g(t)$ as follows:

$$g(t) = e^{-\gamma t} \left[\frac{1-\theta}{\gamma - v\theta} \left(1 - e^{\frac{-(\gamma - v\theta)(T-t)}{1-\theta}} \right) \right]^{1-\theta}.$$

If $R-r \leq \sigma^2(1-\theta)$, then $0 \leq \alpha^{1*} \leq 1$ and $\alpha^{2*} \geq 0$ as required.

Ergodic Control

Here, we give a brief introduction to ergodic control and we refer the reader to Ref. 4 for a detailed analysis. Ergodic control problems deal with minimization of a time-averaged cost over admissible controls. This cost criterion formulation is useful in scenarios when transients are fast and the effective choice is between different steady states. The cost to be minimized is given as

$$J(x, u) = \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T c(X_s u_s) ds \right],$$

and the value function is given as

$$V(x) = \inf_{u \in \mathcal{A}} J(x, u).$$

Even though we represent the value functions as a function of the initial state, it turns out that asymptotically the effect of the initial state vanishes and the cost $J(x, u)$ and the value function $V(x)$ are independent of the initial state. The HJB equation associated with the ergodic control problem is obtained by taking limit as $\gamma \rightarrow 0$ in the HJB equation corresponding to the infinite horizon discounted cost problem (Eq. 4). The equation we arrive at is as follows:

$$\inf_{u \in U} (A^u \phi(x) + c(x, u) - \eta) = 0.$$

This equation needs to be solved for both $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and the constant η . The constant η is the optimal value of the ergodic control problem, that is, $V(x) = \eta$. The function $\phi(x)$ is used to derive the optimal stationary Markov control action as follows:

$$u^*(x) \in \operatorname{argmin}_{u \in U} (A^u \phi(x) + c(x, u)).$$

Risk-Sensitive Control

In this section we provide a brief introduction to the problem of risk-sensitive control as presented in Ref. 5. The risk-sensitive control problem deals with a cost function of the form

$$J(x, u) = \mathbb{E} [\phi(Y)],$$

where $Y = \int_t^T c(s, X_s, u_s) dt + h(X(T))$. The function $\phi(x)$ is an increasing function. A typical choice of the function $\phi(x)$ is of the form $\phi(x) = \rho \exp(\rho x)$. If we assume that $\phi(x)$ is differentiable, then applying Taylor expansion of ϕ around $\mathbb{E}[Y]$, we get

$$\begin{aligned} \phi(Y) &\approx \phi(\mathbb{E}(Y)) + \phi'(\mathbb{E}(Y))(Y - \mathbb{E}(Y)) \\ &\quad + \frac{1}{2}\phi''(\mathbb{E}(Y))(Y - \mathbb{E}(Y))^2. \end{aligned}$$

On taking expectation, we arrive at

$$\begin{aligned} J(s, x, u) &= \mathbb{E}(\phi(Y)) \approx \phi(\mathbb{E}(Y)) \\ &\quad + \frac{1}{2}\phi''(\mathbb{E}(Y))\text{Var}(Y). \end{aligned}$$

Clearly, if $\phi(x)$ is strictly convex around $\mathbb{E}(Y)$ then $\phi''(\mathbb{E}(Y)) > 0$ and the variance term adds to the cost. The optimal control action then tends to minimize the variance and hence, is termed *risk averse*. Similarly when the $\phi(x)$ is strictly concave, the control action is termed *risk seeking*. In the special case when $\phi''(\mathbb{E}(Y)) = 0$, the control is called *risk neutral*. Risk-sensitive control has interesting links to problems in robust control theory. For a detailed analysis of the problem of risk-sensitive control, we refer the reader to Ref. 6.

OPTIMAL CONTROL UNDER PARTIAL INFORMATION

Optimal control problems with partial information arise in a variety of cases when the complete state information is not available to the controller. The partial information is modeled as the outcome of an SDE of the following form:

$$dZ_t = a(t, X_t) dt + \gamma(t, X_t) d\hat{B}_t.$$

The cost function to be minimized is given as

$$\begin{aligned} J(s, x, u) &= \mathbb{E} \left[\int_s^T c(r, X_r, u_r) dr \right. \\ &\quad \left. + h(X_T) \mid X_s = x \right]. \quad (7) \end{aligned}$$

The main idea behind optimal control with partial information is to convert the problem into a problem with full state information by suitably defining a new state vector. In this section, we will restrict ourself to the special case when the state and observation equations are linear. Our presentation in the following section is based on the presentation in Ref. 2. The state and observation equations for the case under consideration is given as

$$dX_t = [A(t)X_t + B(t)u(t)] dt + \sigma(t) dB_t, \quad (8)$$

$$dZ_t = H(t)X_t dt + \gamma(t) d\hat{B}_t. \quad (9)$$

The initial state is a Gaussian random vector X_s and $Z_s = 0$. $\hat{B}_t$ is a standard Brownian motion which is independent of B_t and X_s . At time t , the controller has the information $I_t = \{Z_r : s \leq r \leq t\}$. Such an information pattern is known in the literature as a *classical information pattern*. Let $\mathcal{F}_t$ be an increasing family of σ -algebras such that X_s is $\mathcal{F}_s$ measurable and B_t and $\hat{B}_t$ are $\mathcal{F}_t$ measurable. The control u_t satisfies the following conditions: u_t belongs to the feasible set U and is nonanticipative with respect to $\mathcal{F}_t$. Moreover, it is assumed that $E \int_s^T |u(t)|^k dt < \infty$ for every $k > 0$. Under this assumption, there exists a unique solution of the state equations X_t and the observation equation Z_t . A control is called *nonanticipative* if it assigns the same value for all scenarios with the same past and present. In other words, a nonanticipative control does not anticipate the future evolution. The set of admissible controls satisfying these conditions is denoted by $\mathcal{U}$. Let us define $\mathcal{G}_t^u$ as the σ -algebra generated by the observation process,

$$\mathcal{G}_t^u = \mathcal{F}\{Z_r, s \leq r \leq t\}.$$

Then clearly, we have $\mathcal{G}_t^u \subset \mathcal{U}$. Also, let us define the uncontrolled system as the SDE where $u_t = 0$ as

$$dX_t^0 = A(t)X_t^0 dt + \sigma(t) dB_t,$$

$$dZ_t^0 = H(t)X_t^0 dt + \gamma(t) d\hat{B}_t.$$

We can now apply the Kalman–Bucy filter to arrive at the SDE for the mean square optimal estimate of the the state and the covariance of the error matrix. The mean square

optimal estimate of the state is given by the conditional expectation $\hat{X}_t^0 = \mathbb{E}[X_t^0 | \mathcal{G}_t^0]$, where $\mathcal{G}_t^0 = \mathcal{F}(Z_r^0, r \leq t)$. The Kalman–Bucy filter gives the following SDE for $\hat{X}_t^0$ [2]:

$$\begin{aligned} d\hat{X}_t^0 &= A(t)\hat{X}_t^0 dt + F(t)[dZ_t^0 - H(t)\hat{X}_t^0 dt], \\ \hat{X}_s^0 &= \mathbb{E}[X(s)]; \end{aligned}$$

where $F(t) = R(t)H'(t)[\gamma\gamma'(t)]^{-1}$. The error covariance matrix $R(t) = \mathbb{E}[(X_t^0 - \hat{X}_t^0)(X_t^0 - \hat{X}_t^0)']$ satisfies the matrix differential equation

$$dR = (AR + RA' - RH'(\gamma\gamma')^{-1}HR + \sigma\sigma') dt. \quad (10)$$

We additionally have $Z_t^0 - \int_s^t H(r)\hat{X}_r^0 dr = \int_s^t \gamma(r) dB_r^1$, where B_t^1 is a standard Brownian motion adapted to $\mathcal{G}_t^0$. The term $\int_s^t \gamma(r) dB_r^1$ is known as the *innovation* [7]. Let us also define the deterministic processes X_t^+ and Z_t^+ as the outcome of the ODE:

$$\begin{aligned} dX_t^+ &= [A(t)X_t^+ + B(t)u(t)] dt, \\ dZ_t^+ &= H(t)X_t^+ dt. \end{aligned}$$

Then by linearity of the SDE, we have $X_t = X_t^0 + X_t^+$ and $Z_t = Z_t^0 + Z_t^+$. The following lemma from Ref. 2 characterizes the estimator of the controlled SDE given by Equation (8).

Lemma 1. *If $u(t)$ is $\mathcal{G}_t^0$ measurable and the equality $\mathcal{G}_t^0 = \mathcal{G}_t^u$ holds between the σ -fields of measurements of the uncontrolled and controlled processes, then*

1. *the conditional distribution of X_t given $\mathcal{G}_t^u$ is Gaussian, with mean $\hat{X}_t = \hat{X}_t^0 + X_t^+$ and covariance matrix $R(t)$.*
2. *the conditional mean $\hat{X}_t$ obeys the SDE*

$$\begin{aligned} d\hat{X}_t &= [A(t)\hat{X}_t + B(t)u(t)] dt \\ &\quad + F(t)\gamma(t)dB_t^1, \end{aligned} \quad (11)$$

where B_t^1 is the Brownian motion in the definition of the innovation process.

One of the most important ideas in the control of partially observed systems is the separation principle.

Separation Principle

Before we explain the separation principle, we need to redefine the cost function in terms of the conditional density of the state estimate. Thus, we define

$$p(t, x) = (2\pi)^{\frac{n}{2}} (\det R(t))^{-\frac{1}{2}} \exp \left[-\frac{1}{2} x' R(t)^{-1} x \right].$$

The conditional density of the estimate is given by $p(t, x - \hat{X}_t)$. We define a modified cost function as follows:

$$\begin{aligned} \hat{c}(t, \hat{X}_t, u) &= \int_{\mathbb{R}^n} c(t, x, u) p(t, x - \hat{X}_t) dx, \\ \hat{h}(\hat{X}) &= \int_{\mathbb{R}^n} h(x) p(T, x - \hat{X}), \\ \hat{J}(u) &= \mathbb{E} \left[\int_s^T \hat{c}(t, \hat{X}_t, u(t)) dt + \hat{h}(\hat{X}_T) \right]. \end{aligned} \quad (12)$$

The separation principle states that the optimal control with partial information with classical information pattern can be decoupled into an estimation problem and an optimal control problem with full information. Under a suitable transformation of the state evolution and cost, the problem is converted into a completely observed optimal control problem. The state evolution equation is replaced by the state estimate evolution (Eq. 11). The modified cost is given by Equation (12). The following lemma from Ref. 2 provides a link between the original and modified cost.

Lemma 2. *Under the condition of Lemma 4.1, $J(u) = \hat{J}(u)$.*

Let us define two relevant classes of controls as

$$\begin{aligned} \mathcal{Y}_0 &= \{u \in \mathcal{U} : u(t) \text{ is } \mathcal{G}_t^0 \text{ measurable}\}, \\ \mathcal{Y} &= \{u \in \mathcal{Y}_0 : \mathcal{G}_t^0 = \mathcal{G}_t^u \text{ for } s \leq t \leq T\}. \end{aligned}$$

Then clearly, if we minimize $\hat{J}(u)$ on $\mathcal{Y}_0$ and the minimum $u^* \in \mathcal{Y}$, then by Lemma 4.2 u^* also minimizes $J(u)$ on $\mathcal{Y}$. Wonham in Ref. 8 gives a characterization of control laws

which satisfy the condition $\mathcal{G}_t^0 = \mathcal{G}_t^u$. Let us define $\mathcal{C}^k = \mathcal{C}^k[s, T]$ as the space of continuous functions from $[s, T]$ to $\mathbb{R}^k$ endowed by the sup norm $\|\cdot\|$. Let Γ denote the space of functions γ from $[s, T] \times \mathcal{C}^k$ into the control set U which satisfy the following conditions:

1. If $g(r) = h(r)$ for $s \leq r \leq t$, then $\gamma(t, g) = \gamma(t, h)$.
2. For each $\gamma \in \Gamma$, there exists a constant K such that $\|\gamma(t, g) - \gamma(t, h)\| \leq K\|g - h\|$ for all $g, h \in \mathcal{C}^k$ and $s \leq t \leq T$.
3. $\gamma(\cdot, \cdot)$ is Borel measurable.
4. $\gamma(t, 0)$ is bounded.

Define further a class of controls $\mathcal{Y}_1$ as $\mathcal{Y}_1 := \{u : u(t) = \gamma(t, \eta)\}$. It was shown by Wonham [8] that $\mathcal{Y}_1 \subset \mathcal{Y}$. Thus, if we find an optimal control for $\hat{J}(u)$ in the class $\mathcal{Y}_1$ then the control is also optimal for the cost function $J(u)$ in the class $\mathcal{Y}_0$. We now turn our attention to the linear regulator problem where $U = \mathbb{R}^n$ and the cost function is given by

$$J(x, s, u) = \mathbb{E} \left[\int_s^T (X(t)' M(t) X(t) + u(t)' N(t) u(t)) dt + X(T)' D X(T) \mid X(s) = x \right]. \quad (13)$$

The state estimate evolution is the same as the one mentioned earlier in Equation (11). The modified cost is given by

$$\begin{aligned} \hat{J}(u) &= \mathbb{E} \left[\int_s^T (\hat{X}(t)' M(t) \hat{X}(t) + u(t)' N(t) u(t)) dt + \hat{X}(T)' D \hat{X}(T) \right] + Z, \\ Z &= \mathbb{E} \left[\int_s^T ((X(t) - \hat{X}(t))' M(t) (X(t) - \hat{X}(t))) dt + ((X(T) - \hat{X}(T))' D (X(T) - \hat{X}(T))) \right]. \end{aligned}$$

It can be checked that the term Z is unaffected by control and hence, the term

getting affected by control is the same as in the full information case $J(s, x, u)$ with the state replaced by the state estimate. Thus, the optimal control under partial information has the same functional form as the optimal control under full information with the state vector replaced by the state estimate $u^* = -N^{-1}(t)B(t)'K(t)\hat{X}^*(t)$. The gain $K(t)$ is given by the following matrix Riccati equation:

$$\begin{aligned} \dot{K}(s) &= -K(s)A(s) - A'(s)K(s) \\ &\quad + K(s)B(s)N^{-1}(s)B(s)'K(s) - M(s), \\ K(T) &= D \end{aligned}$$

and the state estimate satisfies the SDE

$$\begin{aligned} d\hat{X}^*(t) &= [A(t)\hat{X}^*(t) + B(t)u^*(t)] dt \\ &\quad + F(t)[dZ^*(t) - H(t)\hat{X}^*(t)dt], \quad (14) \end{aligned}$$

with the initial data $\hat{X}^*(s) = \mathbb{E}[\hat{X}^*(s)]$, where $X^*(t)$ and $Z^*(t)$ are given by the solution of Equations of (2) and (3) respectively, when the optimal control is applied; that is, $u(t) = u^*(t)$.

CONCLUSION AND FURTHER READING

In this article we presented an overview of the basic formulations used in stochastic optimal control. The HJB equation, which is the central equation characterizing optimal control under different cases, was presented. One of the main research problems outside the scope of the current article is characterizing conditions for the existence of a solution to the HJB equation. Some of these conditions are presented in Refs 1, 2, and 6. In Ref. 6, the concept of viscosity solutions to the HJB equation is discussed. It can be shown that for even simple problems, the HJB equation may fail to have a value function which is differentiable. Thus, we need to expand the notion of what a solution to the HJB equation means. A viscosity solution is a generalization of the solution concept for the HJB equation. In this article, we focused on the case where the state evolution is continuous; however, optimal control problems have been formulated in cases when the state evolution is given by a jump diffusion process [9].

REFERENCES

1. Oksendal B. Stochastic differential equations an introduction with applications. 6th ed. New York: Springer; 2007.
2. Fleming WH, Rishel RW. Deterministic and stochastic optimal control. New York: Springer; 1975.
3. Evans LC. An introduction to mathematical optimal control theory. Lecture Notes. Berkeley: University of California. <http://math.berkeley.edu/~evans/>
4. Borkar VS. Controlled diffusion processes. Probab Surv 2005;2:213–244.
5. Ross K. Stochastic control in continuous time. Lecture notes. Stanford University. <http://www.swarthmore.edu/NatSci/kross1/teaching.html>
6. Fleming WH, Soner HM. volume 25, Controlled Markov processes and viscosity solutions. 2nd ed. Stochastic modeling and applied probability. New York: Springer; 2006.
7. Kailath T. An innovations approach to least square estimation. Part I: linear filtering in additive white noise. IEEE Trans Automat Contr 1968;13:646–655.
8. Wonham WM. On the separation theorem of stochastic control. SIAM J Contr 1968;6:312–326.
9. Oksendal B, Sulem A. Applied stochastic control of jump diffusions. 2nd ed. New York: Springer; 2007.

STOCHASTIC ORDERS FOR STOCHASTIC PROCESSES

ALFRED MÜLLER
Universität Siegen, Department
Mathematik, Siegen, Germany

MOSHE SHAKED*
University of Arizona, Department of
Mathematics, Tucson, Arizona

Stochastic orders are useful tools for comparing random variables or random vectors, as well as for stochastic processes. Stochastic orders have been applied in all areas where probability is utilized, such as reliability theory, queueing theory, and insurance. Extensive coverage of stochastic orders can be found, for instance, in the monographs by Müller and Stoyan [1] and Shaked and Shanthikumar [2].

We first recall a couple of common stochastic orders for random variables and random vectors. In this article, “increasing” and “decreasing” are used in the weak sense. Expectations are assumed to exist whenever they are mentioned. The real line is denoted by $\mathbb{R}$.

Let X and Y be two random variables with the distribution functions F and G . A stochastic order relation is usually a partial order relation among the corresponding distribution functions. However, with some abuse of terminology, we often say that a stochastic order “compares” random variables, rather than their distribution functions. One of the most common stochastic orders is the so-called the ordinary (or the first order) stochastic order, according to which X is smaller than Y if

$$E\phi(X) \leq E\phi(Y) \quad \text{for all real increasing functions } \phi. \quad (1)$$

We denote this by $X \leq_{\text{st}} Y$, but, as noted earlier, formally this definition really compares

the underlying distribution functions F and G . Another common stochastic order is the so-called the increasing convex stochastic order, according to which X is smaller than Y if

$$E\phi(X) \leq E\phi(Y) \quad \text{for all real increasing convex functions } \phi. \quad (2)$$

We denote this by $X \leq_{\text{icx}} Y$.

Let now $\mathbf{X} = (X_1, X_2, \dots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ be two random vectors. The stochastic orders $\leq_{\text{st}}$ and $\leq_{\text{icx}}$ defined in Equations 1 and 2 can be extended to $\mathbf{X}$ and $\mathbf{Y}$ as follows: We say that $\mathbf{X}$ is smaller than $\mathbf{Y}$ with respect to the ordinary stochastic order $\leq_{\text{st}}$ if

$$E\phi(\mathbf{X}) \leq E\phi(\mathbf{Y}) \quad \text{for all real } n\text{-dimensional increasing functions } \phi. \quad (3)$$

We say that $\mathbf{X}$ is smaller than $\mathbf{Y}$ with respect to the increasing convex order $\leq_{\text{icx}}$ if

$$E\phi(\mathbf{X}) \leq E\phi(\mathbf{Y}) \quad \text{for all real } n\text{-dimensional increasing convex functions } \phi. \quad (4)$$

MONOTONICITY AND COMPARABILITY OF STOCHASTIC PROCESSES

We now proceed to stochastic comparisons of stochastic processes. A good reference for this is Chapter 5 in Müller and Stoyan [1], which is based, in turn, on Chapter 4 in Stoyan [3].

Let $\mathbf{X} = (X_t : t \in T)$ be a stochastic process. In this article, the parameter set T (the “time”) can be any subset of the real line, but usually it will be $T = \{0, 1, 2, \dots\}$ or $T = (0, \infty)$. Let $\mathbf{Y} = (Y_t : t \in T)$ be another stochastic process with the same parameter set T as $\mathbf{X}$. We can “think” about $\mathbf{X}$ and $\mathbf{Y}$ as “infinite” random vectors with values in $\mathbb{R}^T$, and then extend Equations 3 or 4 to these infinite vectors. This is usually called *complete comparability*. We now define

*Supported by NSA grant H98230-12-1-0222.

a seemingly weaker notion called *strong comparability*. We say that $\mathbf{X} = (X_t : t \in T)$ is *strongly smaller* than $\mathbf{Y} = (Y_t : t \in T)$ with respect to the order $\leq_{\text{st}}$ (denoted as $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$) if

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \leq_{\text{st}} (Y_{t_1}, Y_{t_2}, \dots, Y_{t_n})$$

for all choices of $\{t_1, t_2, \dots, t_n\} \subseteq T$. (5)

In the case of $\leq_{\text{st}}$, complete comparability and strong comparability are equivalent—see the comment after Theorem 1 in the following. Replacing Equation 5 by

$$X_t \leq_{\text{st}} Y_t \quad \text{for all } t \in T,$$

however, yields an order on $\mathbf{X}$ and $\mathbf{Y}$ that is weaker than $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$.

For a single stochastic process, we can also consider the question of internal monotonicity. We say that a process $\mathbf{X} = (X_t : t \in T)$ is monotone with respect to $\leq_{\text{st}}$, if $X_s \leq_{\text{st}} X_t$ for all $s < t$, and similarly for other orders.

An often successful procedure of establishing $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$ for two stochastic processes is by a construction of a proper coupling; see Lindvall [4] or Thorisson [5]. In the context of stochastic comparisons of stochastic processes, coupling means the joint construction, on the same probability space, of two processes that are conveniently related almost surely. Using couplings, it often may be simple to obtain the desired conclusion of stochastic ordering. For the order $\leq_{\text{st}}$, in the case where the state space of $\mathbf{X}$ and $\mathbf{Y}$ is the real line, the following version of the celebrated Strassen's theorem ensures the desired conclusion (below, “ $=_{\text{st}}$ ” denotes equality in law).

Theorem 1. *The two stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ satisfy $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$ if, and only if, there exists a coupling of two stochastic processes $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ that satisfy*

$$\begin{aligned} \hat{\mathbf{X}} &=_{\text{st}} \mathbf{X}, \\ \hat{\mathbf{Y}} &=_{\text{st}} \mathbf{Y}, \end{aligned}$$

and

$$\hat{\mathbf{X}} \leq \hat{\mathbf{Y}} \text{ almost surely.}$$

In some situations, the coupling is easy to find, whereas in other cases, it may be quite tricky.

It follows from the earlier-given Strassen's theorem that strong comparability and complete comparability are equivalent for $\leq_{\text{st}}$ ordering. This also implies that the order $\leq_{\text{st}}$ is closed under increasing transformations. This is formally stated next.

Theorem 2. *If the two stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ satisfy $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$ then*

$$\phi(\mathbf{X}) \leq_{\text{st}} \phi(\mathbf{Y})$$

for any (univariate or multivariate) increasing functional ϕ .

A useful consequence of Theorem 2 involves first passage times of the compared processes $\mathbf{X}$ and $\mathbf{Y}$. That is, for some real a consider

$$T_{\mathbf{X}} = \inf\{t : X_t < a\}$$

and

$$T_{\mathbf{Y}} = \inf\{t : Y_t < a\}.$$

Note that $T_{\mathbf{X}}$ and $T_{\mathbf{Y}}$ are increasing functionals of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Thus, we have the following result.

Theorem 3. *If the two stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ satisfy $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$, then*

$$T_{\mathbf{X}} \leq_{\text{st}} T_{\mathbf{Y}}.$$

For discrete-time processes $\mathbf{X} = (X_n : n \in T)$ and $\mathbf{Y} = (Y_n : n \in T)$, with $T = \{0, 1, 2, \dots\}$, we have the following sufficient conditions for the stochastic comparability of $\mathbf{X}$ and $\mathbf{Y}$.

Theorem 4. *Let $\mathbf{X}$ and $\mathbf{Y}$ be two discrete-time stochastic processes. If*

$$X_0 \leq_{\text{st}} Y_0,$$

and if

$$\begin{aligned} & [X_n | X_1 = x_1, \dots, X_{n-1} = x_{n-1}] \\ & \leq_{\text{st}} [Y_n | Y_1 = y_1, \dots, Y_{n-1} = y_{n-1}] \\ & \text{whenever } x_i \leq y_i, \quad i = 1, 2, \dots, n-1, \\ & n = 1, 2, 3, \dots, \end{aligned}$$

then $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$.

This result is a special case of a result from Kamae *et al.* [6].

Complete and strong comparability can also be defined for other stochastic orders, such as the icx order. We say that $\mathbf{X} = (X_t : t \in T)$ is strongly smaller than $\mathbf{Y} = (Y_t : t \in T)$ with respect to the order $\leq_{\text{icx}}$ (denoted as $\mathbf{X} \leq_{\text{icx}} \mathbf{Y}$) if

$$\begin{aligned} & (X_{t_1}, X_{t_2}, \dots, X_{t_n}) \leq_{\text{icx}} (Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}) \\ & \text{for all choices of } \{t_1, t_2, \dots, t_n\} \subseteq T. \end{aligned}$$

A result that is similar to Theorem 2, holds for the order $\leq_{\text{icx}}$, but because of the required convexity that is involved, some continuity restrictions have to be made; see Theorem 5.1.5 in Müller and Stoyan [1]. However, it is shown in Müller and Stoyan (Example 5.1.7 in Ref [1]) that for the order $\leq_{\text{icx}}$, in general, strong comparability and complete comparability are not equivalent.

Regarding internal monotonicity for the order $\leq_{\text{icx}}$, we say that a process $\mathbf{X} = (X_t : t \in T)$ is monotone with respect to $\leq_{\text{icx}}$, if $X_s \leq_{\text{icx}} X_t$ for all $s < t$. It follows easily from Jensen's inequality that any martingale is monotone with respect to $\leq_{\text{icx}}$.

MONOTONICITY AND COMPARABILITY OF MARKOV PROCESSES

First, let us consider a discrete-time homogeneous Markov chain $\mathbf{X}$. That is, $\mathbf{X} = (X_n : n \in T)$ with $T = \{0, 1, 2, \dots\}$. Denote the transition kernel that is associated with $\mathbf{X}$ by $Q_{\mathbf{X}}$, that is, $Q_{\mathbf{X}} = \{q_{i,j}\}$, where

$$\begin{aligned} & q_{i,j} = P\{X_{n+1} = j | X_n = i\}, \\ & i \text{ and } j \text{ vary over the state space of } \mathbf{X}. \end{aligned}$$

The i th row of $Q_{\mathbf{X}}$ is a discrete probability vector $\mathbf{q}_i = (q_{i,j}, j \in \text{state space of } \mathbf{X})$. If the distribution that corresponds to $\mathbf{q}_i$ is stochastically increasing in i (this can be denoted as $\mathbf{q}_i \leq_{\text{st}} \mathbf{q}_j$, whenever $i < j$) then $Q_{\mathbf{X}}$ is said to be stochastically increasing. When $Q_{\mathbf{X}}$ is stochastically increasing, we have the following monotonicity property of the corresponding Markov chain.

Theorem 5. Consider a homogeneous Markov chain $\mathbf{X}$ with transition kernel $Q_{\mathbf{X}}$. If

$$X_0 \leq_{\text{st}} X_1 \quad (6)$$

and if $Q_{\mathbf{X}}$ is stochastically increasing, then $\mathbf{X}$ is stochastically increasing in the sense that

$$X_n \leq_{\text{st}} X_{n+1}, \quad n = 0, 1, 2, \dots \quad (7)$$

Condition 6 is not very stringent, and it holds in many applications. For instance, if the state space of the process has a minimal state in which the process $\mathbf{X} = \{X_0, X_1, \dots, X_n, \dots\}$ starts at time $n = 0$, then Equation 6 evidently holds. Here is an interesting, useful, and simple example.

Example 6. Let $\mathbf{X}$ be a homogeneous birth and death chain, that is, $\mathbf{X}$ has the state space $\{0, 1, 2, \dots\}$, and its transition probabilities are all equal to 0 except perhaps for

$$\begin{aligned} & q_{0,0}, \quad q_{0,1}, \quad q_{i,i-1}, \quad \text{and} \\ & q_{i,i+1}, \quad i = 1, 2, \dots \end{aligned}$$

It is easy to verify that $Q_{\mathbf{X}}$ is stochastically increasing, if, and only if,

$$q_{i-1,i} \leq q_{i,i+1}, \quad i = 1, 2, \dots$$

If the chain starts at $X_0 = 0$ then, by Theorem 5, X_n is increasing in n with respect to $\leq_{\text{st}}$.

If the inequality in Equation 6 in Theorem 5 is reversed, and if $Q_{\mathbf{X}}$ is assumed to be stochastically decreasing, then it can be

shown that $\mathbf{X}$ is decreasing with respect to $\leq_{\text{st}}$.

If the process $\mathbf{X}$ has a stationary distribution, and Equation 7 holds, then the stationary distribution is an upper bound, in the sense $\leq_{\text{st}}$, on the distribution of any X_n .

An analog of Theorem 5 for a continuous-time Markov process and also for the ordering $\leq_{\text{icx}}$ can be stated and proved [[1], Section 5.2].

Applying Theorem 4 to the case of Markov chains yields the following result.

Theorem 7. *Let $\mathbf{X}$ and $\mathbf{Y}$ be two homogeneous Markov chains. If*

$$X_0 \leq_{\text{st}} Y_0,$$

and if

$$[X_{n+1}|X_n = i] \leq_{\text{st}} [Y_{n+1}|Y_n = j] \\ \text{whenever } i \leq j,$$

then $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$.

For two homogeneous continuous-time Markov processes with state space $\{0, 1, 2, \dots\}$, an analog of Theorem 7 also holds [[1], Section 5.2].

Example 8. Recall that a discrete-time homogeneous Markov chain with state space $\{0, 1, 2, \dots\}$ is called *skip-free positive*, if it does not have positive jumps of magnitude more than one. Now, let $\mathbf{X}$ and $\mathbf{Y}$ be two homogeneous Markov chains, both with the state space $\{0, 1, 2, \dots\}$. Suppose that

$$X_0 \leq_{\text{st}} Y_0,$$

$$[X_{n+1}|X_n = i] \leq_{\text{st}} [Y_{n+1}|Y_n = i] \quad \text{for all } i,$$

and

$$[Y_{n+1}|Y_n = i] \geq i \quad \text{for all } i,$$

and that $\mathbf{X}$ is skip-free positive. Then $\mathbf{X} \leq_{\text{st}} \mathbf{Y}$.

ORDERING ELEMENTS OF MARKOV CHAINS

A function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be supermodular if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ it satisfies

$$\phi(\mathbf{x}) + \phi(\mathbf{y}) \leq \phi(\mathbf{x} \wedge \mathbf{y}) + \phi(\mathbf{x} \vee \mathbf{y}),$$

where the operators $\wedge$ and $\vee$ denote coordinatewise minimum and maximum, respectively. Let $\mathbf{X}$ and $\mathbf{Y}$ be two n -dimensional random vectors such that $E[\phi(\mathbf{X})] \leq E[\phi(\mathbf{Y})]$ for all supermodular functions $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$, provided the expectations exist. Then $\mathbf{X}$ is said to be smaller than $\mathbf{Y}$ in the supermodular order (denoted by $\mathbf{X} \leq_{\text{sm}} \mathbf{Y}$). The supermodular order is often used to compare the strength of positive dependence among the components of the compared random vectors. Only random vectors with the same marginals can be compared with respect to $\leq_{\text{sm}}$. Hence, we consider below a stationary Markov chain, that is, all the X_n have the same marginal distribution. The next result shows that if the transition kernel of the chain is stochastically increasing, then dependence (in the sense of the supermodular order) is decreasing with the lag.

Theorem 9. *Let $\mathbf{X}$ be a stationary homogeneous Markov chain with a stochastically increasing transition kernel. Then*

$$(X_1, X_2) \geq_{\text{sm}} (X_1, X_3) \geq_{\text{sm}} \dots \geq_{\text{sm}}$$

$$(X_1, X_n) \geq_{\text{sm}} \dots \geq_{\text{sm}} (X_1, X^+)$$

where X^+ is an independent copy of X_1 .

The above result is a special case of a result of Hu and Pan [7]. Further results of this type can be found in Kulik and Wichelhaus [8] and Li and Xu [9].

Another interesting result that compares elements of Markov chains involves the hazard rate order. Recall that a random variable X is said to be smaller than another random variable Y in *hazard rate order* if

$$P\{Y > x\}/P\{X > x\} \quad \text{is increasing in}$$

$$x \in (-\infty, \max(u_X, u_Y)),$$

where u_X and u_Y are the right end points of the supports of X and Y , respectively. We denote this as $X \leq_{\text{hr}} Y$. If X and Y are absolutely continuous with density functions f_X and f_Y , distribution functions F_X and F_Y , and

failure rate functions $r_X \equiv f_X/(1 - F_X)$ and $r_Y \equiv f_Y/(1 - F_Y)$, then $X \leq_{\text{hr}} Y$ if, and only if, $u_X \leq u_Y$, and

$$r_X(x) \geq r_Y(x), \quad x \in (-\infty, u_X), \quad (8)$$

Note that if $X \leq_{\text{hr}} Y$ then $X \leq_{\text{st}} Y$. The following result is taken from Ross *et al.* [10].

Theorem 10. *Let $\{X_n^{(i)}, n = 0, 1, \dots\}$ be a homogeneous Markov chain with state space $\{1, 2, \dots, M\}$ (M can be infinity), which starts at state i , that is, $X_0^{(i)} = i$. If*

$$X_1^{(i)} \leq_{\text{hr}} X_1^{(i')} \quad \text{for all } i \leq i'$$

then

- (a) $I_1 \leq_{\text{hr}} I_2$ implies that $X_n^{(I_1)} \leq_{\text{hr}} X_n^{(I_2)}$ for all $n \geq 0$, and
- (b) $X_n^{(1)} \leq_{\text{hr}} X_{n'}^{(1)}$ whenever $n \leq n'$.

COMPARISONS OF POINT PROCESSES

Let $\{K(t) : t \geq 0\}$ and $\{N(t) : t \geq 0\}$ be two point processes. That is, for each $t \geq 0$, $K(t)$ and $N(t)$ are the numbers of jumps that the corresponding processes have experienced over the time interval $(0, t]$. In addition to the possible relationship $\{K(t) : t \geq 0\} \leq_{\text{st}} \{N(t) : t \geq 0\}$ between these processes, according to the definition in Equation 5, we will also consider the following two other stronger relationships. In these comparisons, the “greater” process is the one with “more” or “denser” points. The discussion that follows is based on Shaked and Szekli [11] and Szekli [12].

For any positive integer n , let $B_1, B_2, \dots, B_n$ be bounded Borel sets of $[0, \infty)$. Let $K(B_i)$ and $N(B_i)$ denote the number of jumps of the corresponding processes with jump times in the set B_i , $i = 1, 2, \dots, n$. Suppose that for all choices of an integer n and bounded Borel sets $B_1, B_2, \dots, B_n$, it holds that

$$(K(B_1), K(B_2), \dots, K(B_n)) \leq_{\text{st}} (N(B_1), N(B_2), \dots, N(B_n)),$$

where, also here, the relation $\leq_{\text{st}}$ is in the sense of 3. Then $\mathbf{K} = \{K(t) : t \geq 0\}$ is said to be smaller than $\mathbf{N} = \{N(t) : t \geq 0\}$ in the usual stochastic order over $\mathcal{N}$, where $\mathcal{N}$ denotes the space of integer-valued Radon measures. This is denoted by

$$\mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N}.$$

This “usual” stochastic order over $\mathcal{N}$ gives a “global” comparison of the point processes $\mathbf{K}$ and $\mathbf{N}$. It is easy to see that

$$\mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N} \implies \mathbf{K} \leq_{\text{st}} \mathbf{N}.$$

Next, let $X_1, X_2, \dots$ be the sequence of interpoint distances of the process $\mathbf{K}$, and let $Y_1, Y_2, \dots$ be the sequence of interpoint distances of the process $\mathbf{N}$. We assume that the X_i s and that the Y_i s are almost surely positive. Moreover, we assume that the processes are nonexplosive in the sense that $\lim_{n \rightarrow \infty} \sum_{i=1}^n X_i = \infty$ and $\lim_{n \rightarrow \infty} \sum_{i=1}^n Y_i = \infty$ almost surely. Suppose that, for all choices of an integer n and indices $i_1, i_2, \dots, i_n$, it holds that

$$(X_{i_1}, X_{i_2}, \dots, X_{i_n}) \geq_{\text{st}} (Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}),$$

where, also here, the relation $\leq_{\text{st}}$ is in the sense of 3. Then $\mathbf{K}$ is said to be smaller than $\mathbf{N}$ in the usual stochastic order over $\mathbb{R}^\infty$, and is denoted by

$$\mathbf{K} \leq_{\text{st}-\infty} \mathbf{N}.$$

The usual stochastic order over $\mathbb{R}^\infty$ gives a “local” comparison of the point processes $\mathbf{K}$ and $\mathbf{N}$. Also here it is easy to see that

$$\mathbf{K} \leq_{\text{st}-\infty} \mathbf{N} \implies \mathbf{K} \leq_{\text{st}} \mathbf{N}.$$

It has been shown that, in general,

$$\mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N} \not\Rightarrow \mathbf{K} \leq_{\text{st}-\infty} \mathbf{N}$$

and also that

$$\mathbf{K} \leq_{\text{st}-\infty} \mathbf{N} \not\Rightarrow \mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N}.$$

The most interesting applications of the previous orderings of point processes are

stochastic comparisons of renewal processes. We describe some of these in the following. The proofs of the following results can be found in Shaked and Szekli [11] and Szekli [12].

Theorem 11. *Consider two nondelayed renewal processes $\mathbf{K}$ and $\mathbf{N}$ with generic interpoint distances X and Y , respectively. The following three statements are equivalent.*

- (i) $Y \leq_{\text{st}} X$,
- (ii) $\mathbf{K} \leq_{\text{st}} \mathbf{N}$,
- (iii) $\mathbf{K} \leq_{\text{st}-\infty} \mathbf{N}$.

Theorem 12. *Consider two delayed renewal processes $\mathbf{K}^d = \{K^d(t) : t \geq 0\}$ and $\mathbf{N}^d = \{N^d(t) : t \geq 0\}$, with the corresponding delays X^d and Y^d and with the same interrenewal distribution after the delay. The following statements are equivalent.*

- (i) $Y^d \leq_{\text{st}} X^d$,
- (ii) $\mathbf{K}^d \leq_{\text{st}} \mathbf{N}^d$,
- (iii) $\mathbf{K}^d \leq_{\text{st}-\infty} \mathbf{N}^d$.

The next result gives conditions for $\leq_{\text{st}-\mathcal{N}}$ in terms of the failure rates.

Theorem 13. *Consider two nondelayed renewal processes $\mathbf{K}$ and $\mathbf{N}$ with generic interpoint distances X and Y , respectively. Let r_X and r_Y denote the failure rate functions that correspond to X and Y , respectively. If*

$$r_X(t) \leq r_Y(s) \quad \text{for all } 0 \leq s \leq t \quad (9)$$

then

$$\mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N}.$$

Note that if $Y \leq_{\text{hr}} X$ (in the sense of Eq. 8) and if X is DFR (i.e., has decreasing failure rate) or if Y is DFR, then Equation 9 holds.

Theorem 14. *Consider two delayed renewal processes $\mathbf{K}^d$ and $\mathbf{N}^d$, with the corresponding delays X^d and Y^d and with*

the same interrenewal distribution after the delay. Let r_d denote the failure rate function corresponding to X^d . If $Y^d \leq_{\text{hr}} X^d$ and if

$$r_{X^d}(t) \leq r(s) \quad \text{for all } 0 \leq s \leq t, \quad (10)$$

where r is the failure rate function associated with the common interrenewal distribution function, then

$$\mathbf{K}^d \leq_{\text{st}-\mathcal{N}} \mathbf{N}^d.$$

Denote by X a generic interpoint distance in Theorem 14. We note that Equation 10 holds if $X \leq_{\text{hr}} X^d$, and if X is DFR or if X^d is DFR.

Finally, we give conditions for two nonhomogeneous Poisson processes to be ordered according to the earlier-mentioned orders. Recall that for a nonhomogeneous Poisson process $\mathbf{K} = \{K(t) : t \geq 0\}$, the mean function $M_{\mathbf{K}}$ is defined by $M_{\mathbf{K}}(t) = E[K(t)]$, $t \geq 0$, and the intensity function $\lambda_{\mathbf{K}}$ is defined by $\lambda_{\mathbf{K}}(t) = (d/dt)M_{\mathbf{K}}(t)$, $t \geq 0$.

Theorem 15. *Let $\mathbf{K}$ and $\mathbf{N}$ be two nonhomogeneous Poisson processes with mean functions $M_{\mathbf{K}}$ and $M_{\mathbf{N}}$, respectively, and with intensity functions $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{N}}$, respectively.*

- (i) *If $M_{\mathbf{K}}(t) \leq M_{\mathbf{N}}(t)$, $t \geq 0$, then $\mathbf{K} \leq_{\text{st}} \mathbf{N}$.*
- (ii) *If $\lambda_{\mathbf{K}}(t) \leq \lambda_{\mathbf{N}}(t)$, $t \geq 0$, then $\mathbf{K} \leq_{\text{st}-\mathcal{N}} \mathbf{N}$.*
- (iii) *If $M_{\mathbf{K}}^{-1}(M_{\mathbf{N}}(t)) - t$ is increasing in $t \geq 0$, then $\mathbf{K} \leq_{\text{st}-\infty} \mathbf{N}$.*

There is also a vast literature on stochastic comparison of more general stochastic processes, for which we give only some references. An important paper on the comparison of solutions of stochastic differential equations in the sense of $\leq_{\text{st}}$ is Ikeda and Watanabe [13]. For a similar ordering result in the increasing convex ordering sense, Hajek [14] is a good reference. Many interesting comparison results for semimartingales and Lévy processes can be found in Bergenthum and Rüschendorf [15]. Finally, there are also a number of papers on the comparison of generalizations of Poisson

processes, trying to formalize in queuing theory the intuition that more variable arrival processes lead to worse performance of the queuing system; see Miyoshi and Rolski [16] for an overview.

REFERENCES

1. Müller A., Stoyan D. Comparison methods for stochastic models and risks. Chichester: John Wiley & Sons, Ltd.; 2002.
2. Shaked M., Shanthikumar JG. Stochastic orders. New York: Springer; 2007.
3. Stoyan D. Comparison methods for queues and other stochastic models. Chichester: John Wiley & Sons, Ltd.; 1983.
4. Lindvall T. Lectures on coupling method. New York: John Wiley & Sons, Inc.; 1992.
5. Thorisson H. Coupling, stationarity, and regeneration. New York: Springer; 2000.
6. Kamae T., Krengel U., O'Brien GL. Stochastic inequalities on partially ordered spaces. *Ann Probab* 1977;5:899–912.
7. Hu T., Pan X. Comparisons of dependence for stationary Markov processes. *Probab Eng Inform Sci* 2000;14:299–315.
8. Kulik R., Wichelhaus C. Dependence in lag for Markov chains on partially ordered state spaces with applications to degradable networks. *Stoch Models* 2007;23:683–696.
9. Li H., Xu SH. On the dependence structure and bounds of correlated parallel queues and their applications to synchronized stochastic systems. *J Appl Probab* 2000;37:1020–1043.
10. Ross SM., Shanthikumar JG., Zhu Z. On increasing-failure-rate random variables. *J Appl Probab* 2005;42:797–809.
11. Shaked M., Szekli R. Comparison of replacement policies via point processes. *Adv Appl Probab* 1995;27:1079–1103.
12. Szekli R. Stochastic ordering and dependence in applied probability. New York: Springer; 1995.
13. Ikeda N., Watanabe S. A comparison theorem for solutions of stochastic differential equations and its applications. *Osaka J Math* 1977;14:619–633.
14. Hajek B. Mean stochastic comparison of diffusions. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* 1985;68:315–329.
15. Bergenthum J., Rüschendorf L. Comparison of semimartingales and Lévy processes. *Ann Probab* 2007;35:228–254.
16. Miyoshi N., Rolski T. Ross-type conjectures on monotonicity of queues. *Aust N Z J Stat* 2004;46:121–131.

STOCHASTIC SEARCH METHODS FOR GLOBAL OPTIMIZATION

ZELDA B. ZABINSKY
Department of Industrial and
Systems Engineering,
University of Washington,
Seattle, Washington

Stochastic search methods, or random search algorithms, incorporate some form of randomness in the execution of the algorithm. They are frequently applied to ill-structured global optimization problems, in which the objective function and constraints may be black-box functions and only the function value or membership oracle (constraint satisfaction) is available. The decision variables may be continuous and/or discrete. There are many such global optimization problems that arise in engineering applications of complex systems as well as economic and biological systems.

In contrast to deterministic methods, stochastic search methods sacrifice a strong convergence property for a weaker probabilistic result. However, some stochastic methods can provide a relatively good solution with polynomial time complexity, on the average [1], for many problems where deterministic methods are known to be NP-hard. For example, estimating the volume of a convex body using a deterministic method requires an exponential number of function evaluations, but a stochastic search method can estimate the volume with a high probability of being correct in polynomial time [2,3]. In an optimization context, Zabinsky and Smith [4,5] showed that, if one could sample uniformly from improving regions, then the expected number of such iterations to achieve a solution arbitrarily close to the global optimum with high probability

increases at most *linearly* in dimension. This theoretical sampling was called *pure adaptive search* (PAS). A theory on when linear complexity, on average, may be achieved was developed through the analysis of PAS and subsequent constructs of annealing adaptive search (AAS), and hesitant adaptive search (HAS) (see the article titled ***Random Search Algorithms***, or Zabinsky [4] for an overview). These analyses suggest that stochastic search methods can be efficient and effective and have motivated the development of recent algorithms.

In this article, we discuss several algorithms that were motivated by the ideal performance of PAS and subsequent analyses of AAS and HAS, and utilize hit-and-run as a Markov chain Monte Carlo sampler to develop approximating algorithms. There is a companion article in this encyclopedia (see ***Random Search Algorithms***), which describes PAS, AAS, and HAS in more detail and gives an overview of algorithms, but does not give a detailed description of the algorithms, as found in this article. This article emphasizes algorithms utilizing Markov chain Monte Carlo samplers, including improving hit-and-run (IHR) and hide-and-seek, which fall into the category of simulated annealing algorithms. It also summarizes how the analysis of AAS motivated an analytically derived cooling schedule and in the category of multistart methods, the analysis of HAS motivated a stopping and restarting criterion. Hit-and-run has also been used in the context of an interacting particle algorithm. The assumption in this article is that the objective function can be evaluated, even though it may be a black-box function, whereas the article titled ***Random Search Algorithms*** contains a discussion of algorithms that allow “noisy” objective functions that include randomness and may be available only through a discrete-event simulation or experiment.

GENERAL STOCHASTIC SEARCH ALGORITHM

We consider a general global optimization problem (P) defined as,

$$(P) \quad \min_{x \in S} f(x),$$

where x is a vector of n decision variables (continuous and/or discrete), f is a real-valued function defined over S , and S is an n -dimensional feasible region. The objective function f may be a black-box function and no explicit mathematical expression is required. The feasible set S may be represented by a system of equations, such as $g_j(x) \leq 0$ for $j = 1, \dots, m$, or there may simply be a membership oracle that indicates whether x is feasible. Several types of oracles are discussed in Grötschel *et al.* [6], and the membership oracle is the weakest among them, supporting the use of stochastic search algorithms on ill-structured problems. Notice that the feasible region may include both continuous and discrete variables.

We assume that S is nonempty, and that a unique global optimum exists, using the

General Random Search Algorithm

- Step 0.** Initialize algorithm parameters Θ_0 , initial points $X_0 \subset S$ and iteration index $k = 0$.
- Step 1.** Generate a collection of candidate points $V_{k+1} \subset S$ according to a specific *generator* and associated *sampling distribution*.
- Step 2.** Update X_{k+1} based on the candidate points V_{k+1} , previous iterates and algorithmic parameters. Also update algorithm parameters Θ_{k+1} .
- Step 3.** If a stopping criterion is met, stop. Otherwise increment k and return to Step 1.

The performance of the algorithm is often characterized by properties of the stochastic processes of points $\{X_k\}$ and objective function values $\{Y_k = f(X_k)\}$. Under certain conditions, they converge in probability to the global optima. Thus, stochastic search methods relax a deterministic convergence result to a probabilistic one.

EXAMPLES OF STOCHASTIC SEARCH ALGORITHMS

The stochastic search algorithms discussed in this article are grouped into categories that are commonly recognized. The first category is simulated annealing algorithms that

notation

$$x_* = \arg \min_{x \in S} f(x) \quad \text{and} \quad y_* = f(x_*) = \min_{x \in S} f(x).$$

If S is compact and f is continuous, or if S is finite and f is finite, the existence of a global optimum is guaranteed. Other regularity conditions on (P) may also ensure the existence of a global minimum. Multiple optima are allowed in many algorithms. We let

$$S(y) = \{x \in S : f(x) \leq y\}$$

denote the level set at y . Typically, the goal of a stochastic search method is to find a point $x \in S(y_* + \epsilon)$ for $\epsilon > 0$.

We describe a general random search algorithm that produces a sequence of iterates $\{X_k\}$ indexed by $k = 0, 1, \dots$ which may depend on previous points and algorithmic parameters. The current iterate X_k may represent a single point, or a collection of points, to include population-based algorithms. The iterates are also capitalized to denote that they are random variables, reflecting the probabilistic nature of the random search algorithm.

broadly include algorithms that maintain and generate a single point at each iteration. The second category is population-based algorithms where a collection of points are maintained and generated at each iteration. The well-known genetic and evolutionary algorithms fall into the second category. The third category is multistart and clustering algorithms that use a sequence of single points in a run, but keep some form of history regarding a collection of points to define a cluster. The fourth category is interacting particle algorithms that combine simulated annealing algorithms with population-based algorithms. The fifth category differs from the others in that it is based on an explicit model

for sampling, and includes cross-entropy as a method.

Simulated Annealing

Simulated annealing algorithms are motivated by the physical process of annealing and were originally proposed for optimization in Kirkpatrick *et al.* [7].

A simulated annealing algorithm has three main components. First a candidate point is generated, in Step 1, according to a transition probability. Then the candidate point is accepted or rejected, in Step 2, according to an acceptance probability. The third component is to update the temperature parameter, which is an important parameter for the acceptance probability. The update procedure for the temperature parameter is referred to as the *cooling schedule*. The most common acceptance probability is the Metropolis criterion [8], where a candidate point V_{k+1} generated from X_k is accepted with probability

$$\min \{1, \exp((f(X_k) - f(V_{k+1})) / T)\}.$$

The sequence of points $\{X_k\}$ can be described as a filtered random walk, governed by the combination of transition probability and acceptance probability distributions. If the temperature parameter T is fixed and the random walk satisfies certain conditions, then its stationary distribution [9,10] is shown to be a Boltzmann distribution with density

$$g_T(x) = \frac{e^{-f(x)/T}}{\int_S e^{-f(z)/T} dz}.$$

Bélisle [11] proved that the sequence of objective function values generated by simulated annealing converges in probability to y_* when the candidate generator has a positive probability of sampling anywhere in the feasible set (as with hit-and-run), and the temperature converges to zero in probability. Ghate and Smith [12] extended this result to generalizations of simulated annealing.

Many candidate point generators have been explored and implemented in simulated annealing algorithms [13,14]. In the continuous domain, the candidate point is often

expressed as

$$V_{k+1} = X_k + S_k \cdot D_k, \quad (1)$$

where X_k is the current point, S_k is a step size, and D_k is a direction vector. In some cases, the direction vector is generated by a stochastic gradient at X_k [15], but there are many algorithms that choose D_k according to a uniform distribution on an n -dimensional hypersphere. The step size S_k is also randomly generated, with limits that may shrink or expand depending on previously sampled points. Typically, the range on step size defines the local neighborhood of the candidate point generator.

The generator called hit-and-run [16] also chooses D_k uniformly on a hypersphere, but the step size is generated uniformly on the set of step lengths that provide a feasible solution along direction D_k , that is, uniformly on $\{s : X_k + sD_k \in S\}$. A feature of hit-and-run is that it is globally reaching, that is, there is a positive probability of sampling anywhere in S from any point X_k . Algorithms that only generate candidate points in a local neighborhood of the current point rely on an acceptance probability to allow the algorithm to accept worse (in terms of objective function value) points and thus escape local minima. Simulated annealing with hit-and-run as the candidate point generator always has a positive probability of sampling within ϵ of the global optimum, no matter what the acceptance probability is, although a carefully chosen acceptance probability can be used to speed up convergence.

The first use of hit-and-run as a candidate point generator had a simple acceptance probability—only accept points with improving objective function values. This is called *IHR*, and it has been successfully applied to realistic problems [17]. The idea of IHR is to approximate the ideal linear performance of PAS, and in Zabinsky *et al.* [18], it was shown that the expected number of function evaluations of IHR to get within ϵ of the optimum is $O(n^{5/2})$ on a certain class of convex programs.

Adding acceptance probabilities according to the Metropolis criterion to the hit-and-run generator was first done in Romeijn and

Smith [19,20], and the algorithm was called *hide-and-seek*. A property of hide-and-seek is that it converges to a Boltzmann T distribution for a fixed temperature [21]. Thus, as the temperature follows a cooling schedule, the algorithm can be viewed as approximating a sequence of Boltzmann distributions parameterized by T . The choice of cooling schedules has been studied extensively. An exponential cooling schedule, $T_k = T_0(\gamma)^k$, is commonly used [7] and Hajek [22] suggested a logarithmic cooling schedule, $T_k = T_0/\ln(k+1)$, that was motivated by an analysis using the maximum “depth” of local minima. For a general cooling schedule, Romeijn and Smith [19,20] showed that if hide-and-seek ran long enough at each temperature value to converge to its stationary Boltzmann distribution, then it would dominate the performance of PAS, and the number of these temperature values would be linear in dimension. This led to an analytically derived adaptive cooling schedule in Romeijn and Smith [20]. The analysis of AAS provided conditions for when a linear complexity on the average number of function evaluations could be achieved. This led to a more comprehensive cooling schedule in Shen *et al.* [23], again motivated by the linear performance, on average, of PAS.

Even though the acceptance probability for simulated annealing is interpreted as aiding the algorithm to escape local optima, simulated annealing has also been successfully applied to convex programs. Bertsimas and Vempala [24] and Kalai and Vempala [25] used hit-and-run as a candidate generator in a simulated annealing-type algorithm for solving convex programs with a membership oracle. In Kalai and Vempala [25], simulated annealing is shown to converge quickly and under certain conditions, the Boltzmann distribution is proven to be optimal for annealing on convex problems.

Using Markov chain methods, Ghate and Smith [26] use dynamic programming to determine the number of iterations of simulated annealing at each decreasing temperature value. Also using results about Markov chains, Nolte and Schrader [27] obtained bounds on the finite-time performance of simulated annealing on various combinatorial problems.

Simulated annealing on finite combinatorial problems uses a candidate point generator that is specifically chosen for that problem. For example, a k -city swap is a candidate point generator for the traveling salesperson problem (TSP) that maintains feasibility of a tour. Another generator for integer variables chooses randomly amongst the nearest neighbors of the current point along coordinate directions. The tabu search [28,29] is a modification of a nearest neighbor generator, where close points are considered “tabu” based on previous evaluations, and only unevaluated points are considered as possible candidates. At the other extreme, very large-scale neighborhood (VLSN) generators are successful for other types of combinatorial problems [30].

When the neighborhood generator is restricted to be local, certain conditions on the acceptance probability and cooling schedule are needed to provide convergence properties. Typically, a slow cooling schedule is required which implies a lot of computation. Jerrum and Sorkin [31] determined that simulated annealing can, with high probability and in polynomial time, find the optimal bisection of a random graph. They used a neighborhood structure specific to their problem. A new version of hit-and-run for discrete [32] or mixed continuous–discrete domains, called *pattern hit-and-run* [33], retains the same property of continuous hit-and-run of sampling globally. Thus, it has the same convergence properties for any fixed temperature and cooling schedule as continuous hit-and-run. Specifically, the cooling schedule in Shen *et al.* [23] is derived for integer as well as continuous domains, but its combination with pattern hit-and-run remains for future research.

The proposal distributions of simulated annealing algorithms, as approximations of Boltzmann distributions that concentrate on the global optima, have the potential to provide efficient algorithms as theoretically shown by PAS, AAS, and other analyses. The difficulty is in finding the neighborhood generator and cooling schedule that preserve the desired convergence properties for specific problems. The use of hit-and-run and more recently, pattern hit-and-run are Markov chain Monte Carlo samplers that serve for a

general global optimization problem. More research is needed to incorporate constraints efficiently.

Population-Based Algorithms

Genetic and evolutionary algorithms are population-based random search methods that have been successfully applied toward global optimization. There has been much written on the topic [34,35], and there are several annual conferences and journals dedicated to their study and application. These algorithms are motivated by biological processes, where the concept is to produce “off-spring” of the current population using parameters related to reproduction. Then a selection criterion is applied to create the next generation.

In the context of the general random search where X_k consists of a population of feasible points, Step 1 uses the current population to generate the population V_{k+1} of candidate points, or off-spring. The mechanism used to generate these candidate points consists of cross-over (or recombination) and mutation. The original genetic algorithm of Holland [36] was designed for binary variables and the cross-over mechanism combined two “parents” in X_k and generated two “children” in V_{k+1} by swapping the trailing values in an analogy to the physical cross-over of DNA. The mutation mechanism randomly chooses a binary variable and flips its value. The recombination and mutation mechanisms have many variations and have been extended to continuous variables. Whenever possible, the recombination and mutation mechanisms are tailored to a specific problem to achieve feasibility in a manner similar to how the k -city swap maintains feasibility for the TSP. The mutation element is critical to ensure that the entire feasible region is reachable.

The selection mechanism, as in Step 2, uses the objective function values of the parents and children to mimic “survival of the fittest.” Typically, each candidate point is assigned a selection probability that is used to create the future generation in X_{k+1} . The points with the best objective function values have a higher probability of being selected, which is referred to as *elitism*. However, a

pure elitist strategy converges prematurely and it is recognized that some diversity is needed to converge to a global optimum. An appropriate adaption of the selection probabilities is analogous to a sufficiently slow cooling schedule for simulated annealing. Researchers have used Markov chain analyses to study the behavior of the method [37–39], but tuning of the parameters still requires computational experimentation.

Particle swarm optimization [40,41] is also a population-based stochastic method that is based on the social behavior of birds and schools of fish to locate food. The individuals in the swarm, or population of points in X_k , are typically modeled by position and velocity. The candidate points in V_{k+1} are generated by updating the position based on a scaled step in the direction of the velocity. This resembles the step-size generator of simulated annealing in Equation (1). The updated velocity for each individual is influenced by the overall trend toward the minimum objective function value. In some cases, gradient information is used to update the velocities. In Step 2, the candidate points are all taken for the next population of points, $X_{k+1} = V_{k+1}$, but the updating of velocities can be viewed as a type of selection where better points influence the parameters for the next population. There is also a constraint to maintain a minimum distance between points, so that the swarm does not converge too quickly to a single location.

Ant colony optimization [42] is also a population-based stochastic method based on biology, where the population of ants represent a collection of feasible points. The ants move individually, but communicate through a pheromone that provides feedback on locations with good objective function values. Another population-based stochastic method is differential evolution [43,44] that modifies the genetic cross-over and mutation rules to incorporate a weighted difference between two points to generate a third point. Several of these stochastic algorithms have been numerically tested in Ali *et al.* [45].

A drawback to these algorithms is that the methods of generating new points (e.g., recombination and mutation, or location and velocity) are problem dependent, and require

computational experiments to identify appropriate parameter values. However, the population-based algorithms tend to have efficiencies that single-point algorithms lack, due to the sharing of information during the selection mechanism. Future research is needed to clearly show the impact of various selection mechanisms on efficiency and effectiveness.

Multistart and Clustering Algorithms

Another class of stochastic algorithms is based on the idea of combining a local search with a global search. A natural approach to global optimization, and perhaps the oldest one [46,47], is to apply a local, gradient-type search method with different starting points to determine a local optimum associated with each starting point. Multistart is the term used when a series of different starting points are used to initiate local searches. Typically the starting points are randomly generated, independently and uniformly distributed on S . Although the idea of multistart was originally used with gradient-type searches, multistart has been extended to be used with other local searches as well as globally reaching simulated annealing algorithms.

Multistart is successful in finding many local optima, but there is an inefficiency due to discovering the same local solution from many different starting points. This motivated a class of clustering methods, where clusters of starting points are formed and local searches are only initiated when it is predicted that a new local optimum is likely to be found. Instead of statistically clustering starting points, the well-known multilevel single linkage [48,49] algorithm viewed clusters as trees and “linked” points that led to the same local optima. Several variants include Ali and Storey [50] and Locatelli [51], and a related algorithm called *random linkage* [52,53] was developed. Multistart methods are discussed in Törn and Žilinskas [54] and Laguna and González-Velarde [55], a more recent multistart heuristic in Ugray *et al.* [56], and a successful application to the molecular distance problem in Grosso *et al.* [57].

An important issue in multistart methods is to determine on each single run of an algorithm (local search or perhaps simulated annealing), whether to stop that run and restart with a new run, or whether to terminate completely. Sequential stopping rules for multistart algorithms with pure random search and stratified pure random search were developed in Hart [58] using probability estimates. In Zabinsky *et al.* [59], a stopping and restarting strategy was developed that considered trade-offs between the computational effort and the probability of obtaining the global optimum. The analysis used the theoretical development of HAS, and modeled the behavior of the algorithm to estimate the trade-offs. A framework called dynamic multistart sequential search was developed and numerical tests demonstrate its viability for global optimization.

Multistart and clustering methods provide a means to combine the single-point generation algorithms as in simulated annealing with a collection of points that changes throughout the algorithm.

Interacting Particle Algorithm

The interacting particle algorithm combines the ideas of simulated annealing with population-based algorithms, and is theoretically based on Feynman–Kac systems in statistical physics [60,61]. In Del Moral [60], a general Markov kernel is used as a candidate point generator in Step 1, and a general selection criterion is used in Step 2. A special case of an interacting particle algorithm was developed in Molvalioglu *et al.* [62–64] where hit-and-run was used as the candidate point generator, and a modification of the Metropolis criterion to a population of points (as in Del Moral [60]) was used for the selection mechanism. Specifically, if the candidate points in V_{k+1} are denoted $v_1^{k+1}, v_2^{k+1}, \dots, v_N^{k+1}$, then the point x_i^{k+1} is set to v_j^{k+1} with probability p_j where

$$p_j = \frac{e^{-f(v_j^{k+1})/T}}{\sum_{n=1}^N e^{-f(v_n^{k+1})/T}}.$$

The underlying mathematical model of this interacting particle algorithm is a Boltzmann–Gibbs transformation, which is a special case of the Feynman–Kac measures in Del Moral [60]. As the temperature parameter in the selection criterion goes to zero, the interacting particle algorithm converges to the global optima in terms of the sampling probability distribution. As in simulated annealing with a single particle, the temperature parameter must be chosen carefully to achieve computational benefits.

A numerical study was performed in Molvalioglu *et al.* [62], comparing the interacting particle algorithm to simulated annealing with multistart. Making multiple restarts can also be viewed as having multiple particles instead of a single particle; however, when the restarts are independent, there is no interaction between the particles. The numerical results indicated that there is computational benefit to the interaction.

As in simulated annealing, the temperature parameter T must approach zero for the population to converge. In Molvalioglu *et al.* [63,64], a metacontrol approach was developed to control the temperature parameter in order to speed convergence with a high probability of accuracy. The metacontrol methodology dynamically determines the changes in the temperature parameter by adapting it to the state of the sampling process. The methodology controls the evolution of the probability density function of the particles with the temperature parameter to make the algorithm's search process satisfy user-defined performance criteria. Our performance criteria are directly related to the probability density function of the particles. An interesting feature of the metacontrol approach is that a cooling schedule is not defined *a priori*, but the temperature parameter is adapted to the observed function values. The use of the metacontrol methodology demonstrated that, in contrast to a nonincreasing cooling schedule of temperature in simulated annealing, a dynamic heating and cooling of temperature in the interacting particle algorithm yields enhanced performance Molvalioglu *et al.* [63].

The combination of the interacting particle algorithm with the metacontrol methodology reduces the necessity to have a cooling schedule that works for all problems. The metacontrol methodology can adapt the cooling schedule to the specific problem based on observed function evaluations. More research is needed to adaptively control the parameters of the generation mechanism as well as the selection mechanism.

Cross-Entropy and Other Model-Based Methods

The stochastic search algorithms presented so far generate new candidate points based on the current point or population of points and rely on a set of parameters that govern the transition and acceptance probabilities. Zlochin [65] calls these algorithms *instance-based* methods because they use specific instances of points and function evaluations in determining the next iteration. In contrast, he discusses *model-based* methods that rely on an explicit sampling distribution and update parameters of the probability distribution. These methods are also summarized in another article in this encyclopedia (see **Random Search Algorithms** and **Cross-Entropy Method**) and are just briefly mentioned here.

The most well known of these model-based methods is the cross-entropy method [66,67]. Whereas instance-based methods can be interpreted as a Markov chain sampling method that implies a sampling distribution, the cross-entropy method uses a family of explicit distributions that can be sampled from directly. The idea is based on sequential importance sampling where points are sampled from a sequence of parametrized distributions, and the parameter is updated to minimize the “distance” between the sampling distribution and the target distribution. The target distribution, in the context of global optimization, concentrates its probability on the set of global optima. The distance between the current sampling distribution and the target distribution is estimated by the Kullback–Leibler divergence, also referred to as the *cross-entropy* between two probability distributions. The parameter for the next iteration is chosen to

minimize the Kullback–Leibler divergence between the sampling distribution and the target distribution. Variations of the cross-entropy method have been applied to continuous global optimization problems and combinatorial optimization problems such as the TSP, the quadratic assignment problem, and many other applications. Details of how the update parameters are chosen can be found in Rubinstein and Kroese [66].

Another model-based method is called *model reference adaptive search* (MRAS), and it similarly uses the idea of selecting parameters to bias the sampling distribution toward a degenerate distribution on the global optima. Global convergence properties for MRAS on continuous and discrete domains with noisy objective functions are given in Hu *et al.* [68,69].

Finally, an algorithm aimed at approximating PAS directly, without using a Markov chain Monte Carlo sampler, uses the model of quantum computing for optimization. Baritompä *et al.* [70,71] developed Grover’s adaptive search for optimization using quantum computing. They showed that Grover’s algorithm (originally proposed for database search) could be modified to efficiently generate PAS iterates and thus achieve the ideal linear performance on a quantum computer. Quantum global optimization methods are an exciting area and enlarge the class of stochastic search methods.

SUMMARY

In summary, stochastic search methods have been shown to be effective methods for challenging global optimization problems. The fundamental mechanisms for generating new points are varied and motivated by physical and biological systems, and can be tailored to specific problem formulations. The theoretical treatments for the sampling distributions (as a Markov kernel or a particular form of distribution) provide properties that lead to convergence results. Relating the selection mechanism to a Boltzmann distribution also allows theoretical analyses that can improve algorithm performance. Hit-and-run satisfies

many properties that lead to convergence in probability and approximate desirable complexity results. The model-based methods of cross-entropy and MRAS are also relating sampling methods with target distributions to provide effective algorithms. The idea of metacontrol to incorporate observed function values into modifying parameter values of the generating and selection mechanisms has also shown to improve algorithm performance. Lastly, there is a bright future for quantum global optimization methods.

REFERENCES

1. Mitzenmacher M, Upfal E. Probability and computing: randomized algorithms and probabilistic analysis. New York: Cambridge University Press; 2005.
2. Dyer ME, Frieze AM. Computing the volume of convex bodies: a case where randomness provably helps. *Proc Symp Appl Math* 1991; 44:123–169.
3. Dyer M, Frieze A, Kannan R. A random polynomial time algorithm for approximating the volume of convex bodies. *J ACM* 1991; 38:1–17.
4. Zabinsky ZB. Stochastic adaptive search for global optimization. Boston (MA): Kluwer Academic Publishers; 2003.
5. Zabinsky ZB, Smith RL. Pure adaptive search in global optimization. *Math Program* 1992;53:323–338.
6. Grötschel M, Lovász L, Schrijver A. Geometric algorithms and combinatorial optimization. Berlin: Springer; 1998.
7. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. *Science* 1983;20:671–680.
8. Metropolis N, Rosenbluth A, Rosenbluth M, *et al.* Equation of state calculations by fast computing machines. *J Chem Phys* 1953;21: 1087–1090.
9. Aarts E, Korst J. Simulated annealing and Boltzmann machines - a stochastic approach to combinatorial optimization and neural computers. New York: John Wiley & Sons; 1989.
10. Locatelli M. Convergence of a simulated annealing algorithm for continuous global optimization. *J Global Optim* 2000;18:219–234.
11. Bélisle CJP. Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *J Appl Probab* 1992;29:885–895.

12. Ghate A, Smith RL. Adaptive search with stochastic acceptance probabilities for global optimization. *Oper Res Lett* 2008;36(3):285–290.
13. Horst R, Pardalos PM. Handbook of global optimization. Netherlands: Kluwer Academic Publishers; 1995.
14. Pardalos PM, Romeijn HE. Volume 2, Handbook of global optimization. Netherlands: Kluwer Academic Publishers; 2002.
15. Spall JC. Introduction to stochastic search and optimization: estimation, simulation and control. Hoboken (NJ): John Wiley & Sons; 2003.
16. Smith RL. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper Res* 1984;32:1296–1308.
17. Zabinsky ZB, Graesser DL, Tuttle ME, *et al.* Global optimization of composite laminate using improving hit and run. In: Floudas CA, Pardalos PM, editors. Recent advances in global optimization. Princeton (NJ): Princeton University Press; 1992. pp. 343–365.
18. Zabinsky ZB, Smith RL, McDonald JF, *et al.* Improving hit and run for global optimization. *J Global Optim* 1993;3:171–192.
19. Romeijn HE, Smith RL. Simulated annealing for constrained global optimization. *J Global Optim* 1994;5:101–126.
20. Romeijn HE, Smith RL. Simulated annealing and adaptive search in global optimization. *Probab Eng Inform Sciences* 1994;8:571–590.
21. Bélisle CJP, Romeijn HE, Smith RL. Hit-and-run algorithms for generating multivariate distributions. *Math Oper Res* 1993;18:255–266.
22. Hajek B. Cooling schedules for optimal annealing. *Math Oper Res* 1988;13:311–329.
23. Shen Y, Kiatsupaibul S, Zabinsky ZB, *et al.* An analytically derived cooling schedule for simulated annealing. *J Global Optim* 2007;38:333–365.
24. Bertsimas D, Vempala S. Solving convex programs by random walks. *J ACM* 2004;51(4):540–556.
25. Kalai A, Vempala S. Convex optimization by simulated annealing. *Math Oper Res* 2006;31(2):253–266.
26. Ghate A, Smith RL. A dynamic programming approach for efficient Boltzmann sampling. *Oper Res Lett* 2008;36(6):665–668.
27. Nolte A, Schrader R. A note on the finite time behavior of simulated annealing. *Math Oper Res* 2000;25(3):476–484.
28. Glover F, Kochenberger GA. Handbook of metaheuristics. Volume 57, International series in operations research & management science. Boston (MA): Kluwer Academic Publishers; 2003.
29. Glover F, Laguna M. Tabu search. In: Reeves CR, editor. Modern heuristic techniques for combinatorial problems. New York: Halsted Press; 1993. pp. 70–150.
30. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123:75–102.
31. Jerrum M, Sorkin GB. The Metropolis algorithm for graph bisection. *Discrete Appl Math* 1998;82(1-3):155–175.
32. Baumert S, Ghate A, Kiatsupaibul S, *et al.* Discrete hit-and-run for generating multivariate distributions over arbitrary finite subsets of a lattice. *Oper Res* 2009;57(3):727–739.
33. Mete HO, Shen Y, Zabinsky ZB, *et al.* Pattern discrete and mixed hit-and-run for global optimization. *J Global Optim* 2010; DOI: 10.1007/s10898-010-9534-8.
34. Bäck T, Fogel D, Michalewicz Z. Handbook of evolutionary computation. New York: IOP Publishing and Oxford University Press; 1997.
35. Davis L. Handbook of genetic algorithms. New York: Van Nostrand Reinhold; 1991.
36. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor (MI): University of Michigan Press; 1975.
37. De Jong KA, Spears WM, Gordon DF. Using Markov chains to analyze GAFOs. In: Whitley D, Vose M, editors. Foundations of genetic algorithms 3. San Francisco: Morgan Kaufmann; 1995. pp. 115–137.
38. Koehler GJ. New directions in genetic algorithm theory. *Ann Oper Res* 1997;75:49–68.
39. Koehler GJ. Computing simple GA expected waiting times. Proceedings of the Genetic and Evolutionary Computation Conference; 1999 Jul 13–17; Orlando (FL). 1999. p. 795.
40. Kennedy J, Eberhart RC. Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, Volume 4. 1995. pp. 1942–1948.
41. Kennedy J, Eberhart RC, Shi Y. Swarm intelligence. San Francisco (CA): Morgan Kaufmann Publishers; 2001.
42. Dorigo M, Stützle T. Ant colony optimization. Cambridge (MA): MIT Press; 2004.

43. Ali MM, Törn A. Topographical differential evolution algorithm using pre-calculated differentials. In: Zilinskas A, Dzemyda G, Saltenis V, editors. *Stochastic and global optimization*. Netherlands: Kluwer Academic Publishers; 2002. pp. 1–17.
44. Storn R, Price K. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J Global Optim* 1997;11(4):341–359.
45. Ali MM, Khompatraporn C, Zabinsky ZB. A numerical evaluation of several global optimization algorithms on selected benchmark test problems. *J Global Optim* 2005;31(4): 635–672.
46. Dixon LCW, Szegö GP. *Towards global optimization*. Amsterdam: North-Holland; 1975.
47. Dixon LCW, Szegö GP. *Towards global optimization 2*. Amsterdam: North-Holland; 1978.
48. Rinnooy Kan AHG, Timmer GT. Stochastic global optimization methods; part I: clustering methods. *Math Program* 1987;37:27–56.
49. Rinnooy Kan AHG, Timmer GT. Stochastic global optimization methods; part II: multi level methods. *Math Program* 1987;37:57–78.
50. Ali MM, Storey C. Topographical multilevel single linkage. *J Global Optim* 1994;5:349–358.
51. Locatelli M. Relaxing the assumptions of the multilevel single linkage algorithm. *J Global Optim* 1998;13:25–42.
52. Locatelli M, Schoen F. Random linkage: a family of acceptance / rejection algorithms for global optimization. *Math Program* 1999;85 (2):379–396.
53. Schoen F. Two-phase methods for global optimization. In: Pardalos PM, Romeijn HE, editors. *Volume 2, Handbook of global optimization*. Netherlands: Kluwer Academic Publishers; 2002. pp. 151–177.
54. Törn A, Žilinskas A. *Global optimization*. Germany: Springer; 1989.
55. Laguna, M, González-Velarde, JL, editors. *Computing tools for modeling, optimization, and simulation*. Boston (MA): Kluwer Academic Press; 2000.
56. Ugray Z, Lasdon L, Plummer J, *et al.* Scatter search and local NLP solvers: a multistart framework for global optimization. *Inf J Comput* 2007;19(3):328–340.
57. Grosso A, Locatelli M, Schoen F. Solving molecular distance geometry problems by global optimization algorithms. *Comput Optim Appl* 2009;43(1):23–37.
58. Hart WE. Sequential stopping rules for random optimization methods with applications to multistart local search. *SIAM J Optim* 1998;9(1):270–290.
59. Zabinsky ZB, Bulger D, Khompatraporn C. Stopping and restarting strategy for stochastic sequential search in global optimization. *J Global Optim* 2010;46(2):273–286.
60. Del Moral P. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. New York: Springer; 2004.
61. Del Moral P, Miclo L. On the convergence and applications of generalized simulated annealing. *SIAM J Control Optim* 1999;37(4): 1222–1250.
62. Molvalioglu O, Zabinsky ZB, Kohn W. Multi-particle simulated annealing. In: Törn A, Zilinskas J, Zilinskas A, editors. *Models and algorithms for global optimization*. New York: Springer; 2007. pp. 215–222.
63. Molvalioglu O, Zabinsky ZB, Kohn W. The interacting-particle algorithm with dynamic heating and cooling. *J Global Optim* 2009; 43:329–356.
64. Molvalioglu O, Zabinsky ZB, Kohn W. Meta-control of an interacting-particle algorithm. *Nonlin Anal: Hybrid Syst* 2010; DOI: 10.1016/j.nahs.2010.04.004.
65. Zlochin M, Birattari M, Meuleau N, *et al.* Model-based search for combinatorial optimization: a critical survey. *Ann Oper Res* 2004;131:373–395.
66. Rubinstein RY, Kroese DP. *The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation, and machine learning*. New York: Springer; 2004.
67. Rubinstein RY, Kroese DP. *Simulation and the Monte Carlo method*. 2nd ed. New Jersey: John Wiley & Sons; 2008.
68. Hu J, Fu MC, Marcus SI. A model reference adaptive search method for global optimization. *Oper Res* 2007;55:549–568.
69. Hu J, Fu MC, Marcus SI. A model reference adaptive search method for stochastic global optimization. *Commun Inform Syst* 2008;8(3):245–276.
70. Baritomba WP, Bulger DW, Wood GR. Grover's quantum algorithm applied to global optimization. *SIAM J Optim* 2005;15(4): 1170–1184.
71. Bulger D, Baritomba WP, Wood GR. Implementing pure adaptive search with Grover's quantum algorithm. *J Optim Theory Appl* 2003;116:517–529.

STRATEGIC AND OPERATIONAL PREPOSITIONING IN CASE OF SEASONAL NATURAL DISASTERS: A PERSPECTIVE

ARUNA APTE
Naval Postgraduate School,
Monterey, CA, USA

INTRODUCTION

The classification of disasters based on time and location and the lifecycle of the disaster are critical for consideration when planning humanitarian assistance [1, 2]. Planners must consider the lifecycle of the disaster that is based on the stages of preparedness, response, and recovery when determining a strategy for disaster relief [3]. Effective logistics in the case of a natural disaster entails getting the necessary emergency supplies and services to the affected population in time. Approximately, 80% of disaster relief involves logistics activities [4]. In the aftermath of a disaster, the distribution and delivery of critical supplies and services to the affected population is a major undertaking. Practitioners [5–7], media, organizational reports, and scholars [3, 8–10] agree that preparation and planning is a significant part of any relief effort.

The objective of all of the organizations involved in humanitarian assistance for these disasters is to reduce human suffering and casualties. Providing relief is largely dependent upon the speed and scope of the response, which is often a function of the level of preparedness that has been established before the disaster event. “Famines occur not because there is not enough food in the world but because the food is not where it is needed” ([11], p. 213). This concept can be applied to disasters, as well: Suffering and casualties often increase owing to disasters because the emergency supplies and services do not reach the affected population in time.

Emergency supplies delivered and services provided depend on not just the provisions and capacities but also the demand estimation [3, 12]. Incorrect demand estimation can force the relief providers to depend on surge, a fairly expensive strategy [2, 13–15]. Forecasting the quantity of demand is a difficult proposition but estimating where and when is even harder [1, 16, 17].

The primary strategies for addressing disasters are prepositioning, proactive and phased deployment, and surge [2, 18–20]. *Prepositioning* is usually understood as the storage of essential commodities to be utilized in disaster relief. However, prepositioning also addresses the issues of capacity planning, facility location for delivering commodities and dispensing emergency services, and strengthening the physical infrastructure. *Proactive deployment* refers to moving supplies into an area in advance of a demand request to reduce lead time. *Phased deployment* normally occurs after the disaster has struck. It refers to pushing inventory into a disaster area as it is needed and in the quantity in which it is needed. A *surge* is the deployment of manpower and supplies from locations outside the disaster area. A surge does not rely on prepositioned inventory but instead relies on established excess capacity. Prepositioned capacity and commodities, a coordinated proactive or phased deployment of assets that leads up to and continues past the time of the disaster, and a surge to supplement these three strategies can eventually attempt to meet the demand spikes and mitigate suffering in all disasters.

The challenges in response supply chains, such as demand surges, the uncertainty of supplies, and the critical time windows in the face of vulnerable infrastructure (both physical and social), raise several serious questions. Finding answers to these questions will result in a more effective and efficient disaster relief. How can demand quantities be estimated [3, 12, 21]? How can

adequate prepositioned disaster relief inventory be established and maintained [22–24]? How reliable are the potential, virtual, and real supply lines [25–32]? Should the supplies be sourced locally (but outside the disaster zone) or should they be imported [33–35]? In addition, the following questions are of special interest for this study: Is prepositioning the best strategy for all types of disasters? Or do there exist certain types of disasters where prepositioning is a better strategy? These are some of the critical decisions faced by humanitarian organizations.

Although all of these questions and strategies are worth researching to facilitate decision-making, in this article, we focus on prepositioning for the following reasons. First, the United Nations [13] and the World Meteorological Organization [15] estimate that every dollar invested in preparing for a disaster saves seven dollars in disaster response. From an economic standpoint, this is a major motivation. In addition, lessons from the military organizations [1, 36] emphasize the importance of prepositioning supplemental resources in or near an incident's location to support operations during extreme conditions. The success of the military in using prepositioned stocks to support operations other than war helps in understanding prepositioning. US Army prepositioned stocks in Southwest Asia and Korea are good examples. One of the significant and often repeated lessons learned from various disasters in the recent past is to be prepared through prepositioning. The analysis of data, based on Action Request Forms issued after Hurricane Katrina made landfall, revealed that regional prepositioning would have expedited the lead time of delivery. In case of a disaster, such prepositioning could be the first wave of relief supplies until more could arrive to the affected area. In addition, such prepositioning could be cost effective if about-to-be-outdated supplies from the safety stock of stockpiles (stored by organizations involved in relief efforts) from nearby locations could be used for such relief [10, 19].

In order to understand prepositioning in humanitarian logistics, we can also derive lessons from commercial supply chains [8].

In the commercial sector, this is similar to prepositioning spare parts and reserve equipment because of the uncertainty of the locations in which parts and equipment will be needed and the uncertainty of the time in which they will be requested. It is also similar to positioning seasonal inventories close to the retail stores during the season and moving them back to warehouses in the off season [37]. In both these cases, forecasting the demand is much easier than in the case of a war or disaster.

We, first, want to clarify what we mean by prepositioning in this article. Prepositioning is divided into two categories: *Strategic* prepositioning occurs long before a disaster strikes, with readiness in mind; and *operational* prepositioning occurs when the disaster is imminent (a response is anticipated within a short lead time) and during the actual response. Strategic prepositioning predominantly involves capacity planning, reinforcing infrastructure, and investigating the needs and availability of various facilities [38–41]. Both the approaches definitely involve the dimension of time in the life cycle of a disaster. However, it needs to be clarified further that the operational prepositioning is also an activity-based approach as it focuses on actually locating the facilities with the appropriate level of inventory, such as warehouses with emergency supplies or triage facilities with medicines [14, 39, 42–48]. Prepositioning, in either approach, is most effective when relatively longer lead times are possible and the frequent location of affected areas such as the “tornado alley” in the southwestern parts of the United States are known or can be estimated.

Natural disasters such as hurricanes, typhoons, and flooding usually occur frequently with certainty during certain seasons in similar locations. For example, hurricanes are most likely to occur in late summer and early fall in the southeastern United States, and typhoons and flooding occur mostly during the monsoon season in Asia. We call these seasonal natural disasters. Although earthquakes occur at a higher frequency in certain locations, they cannot be deemed seasonal because of their

sudden onset, resulting in unpredictability in terms of time.

From the perspective of prepositioning for seasonal natural disasters, only the seasonality offers some insight into planning the disaster relief. In this article, we focus on natural, seasonal disasters, because natural disasters are plentiful, and thus easier to draw inferences from, and because seasonal disasters are relatively more predictable. The prepositioning approach is particularly useful in the case of seasonal natural disasters because of the predictability of both location and timing. In this research, we outline and demonstrate our position on operational and strategic prepositioning. We do this through a literature review in the section titled “Literature Review”, a presentation of our perspective in the section titled “A Perspective”, and a description of caveats in the section titled “The Caveats”. In the section titled “Conclusion”, we offer the conclusions.

LITERATURE REVIEW

Prepositioning, for either approach, has been studied by the scholars [39, 42–50]. Recently, this research has garnered even more interest from academics [51]. See Operations Research to Improve Disaster Supply Chain Management, the humanitarian community, and military organizations—particularly the US military—because of the strategic importance of global humanitarian assistance and disaster relief. More importantly, the operational capabilities of relief providers have improved, perhaps because of the scrutiny and lessons learned from previous disasters and of numerous research and case studies in humanitarian operations. Prepositioning supplies is appropriate when the lead time to respond exceeds the time line for the need. It is also a good solution when critical transportation has to be preserved [34]. Emergency agencies realize the value of designing and reconfiguring disaster relief systems by locating the assets and the facilities based on various objectives. The fact that analytical tools are essential in this effort is well documented in the literature.

Strategic prepositioning, as defined in this article, takes place in the predisaster, preparedness phase in the lifecycle of a disaster [1]. This includes capacity planning of manpower, assets, and resources [3, 50]. Preparedness activities are predominantly strategic and comprise setting the stage for an appropriate response. Operational prepositioning, as defined in this article, mostly takes place before the disaster strikes, is imminent, or even after during the response phase of the lifecycle of a disaster. However, this category is also activity based. For seasonal natural disasters, where the type and location of the disaster can be forecasted results in certain amount of lead time for the relief operations that benefit from both strategic and operational prepositioning.

Facility Location

There are a variety of optimization models available in the literature for facility location in the response phase or operational prepositioning [14, 38–41, 52, 53]. (We would like to note that although all the articles study facility location models, some of the articles also incorporate inventory decisions made way in advance, which does not necessarily conform to our definition of operational prepositioning.) The objective function involves locating the facilities by minimizing the average distance or time and minimizing the maximum distance of the demand points from a facility [14, 40]. Some models involve minimizing the total number of facilities [42, 43], whereas some consist of maximizing reliability [45, 54, 55].

We focus on those models that are developed to locate facilities for disaster relief. Two types of optimization models, the set covering problem (SCP) ([56], pp. 6–7) and the facility location problem (FLP), have been explored by researchers to find solutions to the problem of locating facilities to aid in disaster relief. Surveys of SCPs can be found in Church and ReVelle [57], Brandeau and Chiu [58], and Klose and Drexl [59]. There have been various modifications of the SCP model, as reviewed extensively by Marianov and ReVelle [60], for emergency services. Further studies can be found in the recent

review by Caunhye *et al.* [61] and the references therein.

In the SCP, each facility is associated with a region. This is a binary problem about whether the facility is to be “built/located” or not. In this model, there are no “how to locate” and “transport” decisions to be made, which are typically critical in disaster relief operations (because of the uncertainty surrounding the available manpower, existing infrastructure, and nature of the demand). The FLP is a network-based problem in which one chooses from a candidate list of facilities, making the decision to close or open each individual facility and how much to transport. Certain extensions of these problems also incorporate expansion of the facilities. Both types of models, SCP and FLP, deal with the issue of location, but they address different issues. Traditionally, the FLP is known for facility location in supply chains. However, for emergency services, the SCP or its variants are often used [59].

Approaches developed by researchers are based on the minimum cost covering problem, better known as the location set covering problem (LSCP). The LSCP seeks the minimum number of facilities and their locations so that all demand points are covered. The constraints normally require that all points of demand be covered within a certain time or distance. Satisfying these constraints makes the LSCP a substantial planning tool in disaster relief for locating emergency services because all demand points have to be covered when it comes to an affected population. The models most applicable to emergency service facilities incorporate the maximum time or distance that exists between the user and the service [57, 62–68]. See Introduction to facility location. Such models provide substantial analytical tools for planning for disaster relief.

The maximal covering location problem (MCLP) does not require the coverage of all demand points owing to economic constraints and hence is modeled using a candidate list to pick the best p facilities for maximizing coverage. The MCLP requires finding a subset of p facilities so that maximum (not all) demand can be covered within a given maximal distance. When there is a budget involved, and

there most often is, the MCLP is the natural choice in planning for warehouses and triages in the response phase.

The facility location models have evolved over time, incorporating different constraints based on specific situations of prepositioning discussed in our research. For example, Hale and Moberg [42] specify constraints for locating storage areas for critical emergency equipment and supplies. Whether it is a response to a natural disaster or preparation for famine relief, critical supplies have to be stored in a safe and secure location. Storing these items at every distribution center, manufacturing facility, transportation hub, and office within the supply chain can be unreasonably expensive. Hale and Moberg [42] developed a model with the usual maximum distance constraints found in the generalized SCP. However, they also incorporated constraints for minimum distance between each secured location from each facility. These constraints not only balance operational effectiveness and cost efficiency but also make the solution space smaller and hence more manageable.

Some applications of facility location models with real data can be found in the humanitarian logistics literature. However, such examples that combine rigor and relevance for prepositioning are few. In this article, we offer two succinct examples of establishing facilities for prepositioning in case of disasters.

Balcik and Beamon [39] developed a facility location model to incorporate inventory decisions for a disaster, determining the number and location of distribution centers and the inventory levels at each distribution center. In their study, the authors applied the model to 286 scenarios, each defined by disaster location and its impact; they also considered 45 candidate distribution centers, based on data gathered from the National Geophysical Data Center.

CARE International is one of the largest humanitarian organizations in the world. CARE has adopted disaster preparedness models for providing relief to the affected population of a disaster-struck area by prepositioning items in advance of a disaster

to reduce the response time [14]. Improving agility for acquiring and distributing resources is critical in a disaster situation. Duran *et al.* [14] developed a model evaluating the effects of prepositioning relief supplies on reducing average response time. The findings were applied by CARE when it established its first prepositioning facility in 2008.

Capacity Planning

In addition to the actual storage of commodities or location of emergency facilities, there is also a need to plan for the capacity to deliver essential supplies and services. This type of planning entails the anticipation of a surge in human resources and the availability of an adequate physical infrastructure. This type of prepositioning can be described as capacity planning.

Using a two-stage stochastic optimization model [69] to evaluate disaster response scenarios—through notional scenarios [20] based on public source data—offers viable strategic options. Tean [69] and Heidtke [20] uncovered the important humanitarian logistics questions of “What and how much do they need?” as opposed to “How much can we send?” Salmeron and Apte [50] further investigated this problem of strategic allocation of resources and capacity planning for humanitarian responses to future disasters. We discuss this study in detail in an example later.

Mobile facilities are frequently used to respond to diverse and remote requests for relief. Although transportation has grown to be the second largest operating cost for humanitarian operations since the 1990s [70–72], professional fleet management is almost nonexistent in smaller nongovernmental organizations (NGOs). Research studies have looked at the management of fleets of vehicles for use in capacity planning for a disaster [73, 74], exposing a vein of research for capacity planning in transportation.

A PERSPECTIVE

The literature we have discussed up to this point is divided into two parts because we

believe that it is useful to consider the two types of prepositioning—strategic and operational.

Operational Prepositioning

We have defined operational prepositioning as activity-based prepositioning (such as locating warehouses with stocking of necessary inventory of supplies) that is mostly carried out in the imminent or response phase. For example, this includes the facility location of assets such as warehouses for the storage of consumable supplies, the location of triage facilities for consumable medicines, or the military stocks that can be used for operations other than war. In addition, facility location is necessary for the inventory that is needed for virtually all seasonal natural disasters. The decision of where to preposition supplies must be made with consideration of the trade-offs between reducing the distribution time by placing stocks “near enough” a potentially affected area and placing stocks so close to (or within) a potential disaster area that they are adversely impacted by the disaster when it occurs.

A desirable logistics strategy for disaster events such as the 2004 Indian Ocean tsunami, 2005 Hurricanes Katrina and Rita, and 2010 Pakistan floods could have been locating the supplies outside the potential disaster zone. Prepositioning positively influences the speed of emergency response that is a critical factor in disaster relief [14, 35, 39, 52, 75, 76]. Specifically, Ichoua [76] discussed a model based on scenarios to incorporate locating facilities with inventory and routing decisions, thus encompassing operational and strategic prepositioning, whereas Ukkusuri and Yushimito [75] focus on operational prepositioning by considering disruptions in transportation network. Studies in logistics [50, 77–81] concluded that infrastructure such as distribution centers, warehouses, and medical clinics should be established in locations based on the density of the population and transportation hubs. Location is even more critical when uncertainty is considered [25–32]. Galindo and Batta [82] develop a model specifically taking into account destruction of such locations. An

important, if not *the* most important, stage in the response supply chain is the last mile distribution [22, 83, 84]. In this and many other complex humanitarian logistics systems, the effective and efficient deployment of relief hinges on facility location.

An Example. An example of what can be achieved if academics, practitioners, and emergency planners collaborate in solving the problem of operational prepositioning using analytical tools can be found in the study done by Dekle *et al.* [43]. This study was carried out in the southeastern United States, specifically Florida, where hurricanes from the Gulf of Mexico and the Caribbean often land in late summer and early fall. Federal Emergency Management Agency (FEMA) opens Disaster Recovery Centers (DRCs) in affected areas where such disasters strike. The question is not whether the hurricanes will strike, but where and when they will strike. Dekle *et al.* [43] helped FEMA identify at least three (as required by FEMA) DRC sites in Alachua County, Florida. The decision-making process was designed to be efficient so that it could be implemented based on the parameters of an imminent hurricane during the response phase. The researchers chose from a number of objectives: minimize the average travel distance to the closest DRC, minimize the maximum travel distance to a DRC, minimize the total number of DRCs needed such that all population locations are within a specified distance of the nearest DRC, and maximize the probability that at least one DRC can be useable after the disaster. FEMA and the authors agreed to focus on minimizing the total number of DRCs for three different distances.

The authors formulated the problem as an SCP and solved the original model using aggregate data with Microsoft Excel by solving p -center problems in succession. The emergency management coordinator of the county believed that by having a predetermined list of DRC locations, the agency could provide relief sooner, because of being able to quickly open the facilities and make them operational for upcoming demand of the affected population.

It is important to note that the study's authors accepted the loss of accuracy in reducing the problem size by aggregating the data [85–87]. The 6600 original demand points were reduced to 198 aggregate demand points using Pick-the-Farthest Algorithm. Similarly, 3900 original potential DRC sites were aggregated into 162 sites. They demonstrated that using similar aggregation, the planners can improve solvability and, furthermore, can increase the simplicity of the problem so as to make it more user-friendly. These two attributes of this project made it swiftly executable for operations. It should be noted that this simplification does not capture the entire problem but it is effective. However, the main advantage of this case study was not just the location of DRCs but that the users may not have a single global objective. Other important lessons include the following: The choice of a model may be subjective, data collection is most of the work, and, last but not least, simplification by aggregation of data may be indispensable.

Strategic Prepositioning

Strategic prepositioning involves acquiring services and supplies through capital planning, contracting, or collaboration. It is this prepositioned capacity that leads to response in the time of need. One of the lessons learned after recent seasonal natural disasters, specifically Hurricane Katrina, was that effective and efficient operations cannot be accomplished without strategic prepositioning [10]. Therefore, in addition to storing critical supplies, FEMA has planned evacuation routes and temporary shelters designated in the hurricane-prone region. In other words, FEMA has done some capacity planning.

If a public agency does not have adequate resources to stage prepositioned supplies, it might use networks of volunteers to serve as prepositioned capacity for manpower. FEMA's Response and Recovery Division Director for Region IX noted that the missions such as medical evacuation, planning for commodity distribution, and temporary housing have to be planned before

disaster strikes [6]. Similarly, the US California National Guard recommends that such a disaster response system for prepositioning supplies needs to be in place [7], and the Asia-Pacific Task Force for Emergency Preparedness [88] supports a strategic plan for a more effective emergency preparedness, risk reduction, and disaster response. A discussion of evacuation planning can be found in the studies by Regnier [89, 90] See Evacuation Planning.

An Example. Strategic prepositioning deals with questions such as what level of capacity is needed and can be maintained for warehouses, medical facilities, airfield ramp space, and temporary shelter space? These decisions have to be made substantially ahead of when a disaster strikes. The main objective of strategic prepositioning is to minimize the expected number of casualties. Salmeron and Apte [50] extended the two-stage stochastic optimization model developed by Tean [69] and Heidtke [20] to address vital issues in strategic prepositioning. First-stage decisions involved the expansion of these resources significantly before disaster strikes. In the second stage, these allocated resources were utilized to minimize expected casualties. The study focused on setting up capacity and resources for executing efficient relief operations.

The model handled multiple affected areas that could be affected by disasters for which there may be inadequate probabilistic information. This was possible because of the stochastic element of their model. In order to assess the consequences of prepositioning strategies, the methodology of Salmeron and Apte [50] modeled some of the operational specifics. In the study, the two-stage probabilistic model guided the expansion of capacities: warehouses for critical supplies, medical staff, airstrips, and temporary shelters to minimize expected casualties (their first objective). These second-stage decisions were based on the concept of recourse in which supplies and personnel were deployed after the disaster had occurred and assets had already been allocated. A secondary objective was to minimize the expected

population that would not be transported to the shelters.

The test case used to solve the model and evaluate the results [50], although hypothetical, was based on data using public sources [20, 69] and consultation with the humanitarian community [5, 6]. Computational results on this notional test case provided insights into the problem complexity.

The research study also demonstrated the benefit of using optimization to guide budget allocation. The authors observed that there was a trend in the allocation of the budget to different categories of expansion. They attributed the trend to “(i) the mismatch between the initial MoT capacity and that of health providers, warehouses and ramp spaces, (ii) the budget level, and (iii) planners’ estimates for survival rates and penalties (casualties) for undelivered commodities” [50, p. 573]. The sensitivity analysis performed for different levels of the budget offered interesting insights for emergency planners.

Varying the budget in increments while maintaining expenditure persistence offered insights to the emergency managers. First, as allocation levels increased in warehouses and shelters, the expansion of ramp space and health facilities remained somewhat constant. Results showed that the warehouse capacity was the dominating constraint for minimizing the casualties. Decreasing unsatisfied demand of supplies was explained by the bottleneck shift to warehouse expansion, resulting in the significantly higher expense of rescuing the critical population that was left behind. As the expansion capacity for the shelters diminished, the budget allocation also decreased. One of the significant findings for allocating the budget to different resources was that after a certain level of budget all the allocations evened out with no reduction in the casualties. The research [50] also assessed the benefit of the stochastic model using a perfect information model and averaging scenarios model.

The take-away lessons from Salmeron and Apte [50], for the instance discussed in their article, were as follows: with lower budgets, existing transportation capacity and health capacity for critical populations had to be in line except in the case where the survival rate

was very low or penalties for not meeting the demand were very high. However, as more budgets became available, the expansion of warehouses and meeting the demand became significant because of high cost of special transportation and the exorbitant expense for health facilities for remaining critical population. In general, the research demonstrated that stochastic optimization allows the planner to enter a variety of scenarios by severity or location of the disaster. In addition, their solution accounted for all of the scenarios concurrently without giving any specific importance to some. In addition, logisticians can derive insights from the interaction between increased budgets, their distribution, and the overall effect on the objective.

THE CAVEATS

As demonstrated by the literature in humanitarian logistics and practices in the humanitarian community, prepositioning (strategic and operational) is a strategy that provides significant benefits for effective and efficient humanitarian operations. However, there are some caveats.

One of the caveats about prepositioning is the need to understand the risk associated with the storage of supplies in an area that might be affected by the very disaster for which the relief is aimed [35]. This risk has many facets. Locating the commodities outside the potential disaster area is necessary for timely deployment, but the investment in such inventories could be prohibitively expensive. In addition, the feasibility of maintaining stocks of critical items such as food, water, and medicines for any period of time in countries that are poor, that have high levels of corruption, or that have low governance should be carefully weighed. Any of these situations will affect the security of the prepositioned supplies. Prepositioning medicines adds another wrinkle to the implementation of the strategy. Storing vaccines, antibiotics, and antidotes is appropriate but only when the biological agents created in the aftermath of the disasters to be combated are stable and not mutating.

It is important to note that the success of prepositioning by the military in case of

war cannot be translated to success in prepositioning for humanitarian assistance and disaster relief, as demonstrated in the case of the US Navy described by Apte *et al.* [17]. The assets in the US Navy have capabilities mostly for supporting conflicts and not humanitarian relief. Therefore, deploying them for operations other than war may not accomplish the objective of humanitarian relief. In other words, having prepositioned assets and then deploying them in case of a disaster does not necessarily constitute a good strategy, unless the assets selected are appropriate.

Another difficulty of prepositioning is funding. A significant percentage of donations to humanitarian organizations are earmarked. Such earmarked funding has a negative effect on the prepositioning strategy. Most donors like to see the immediate impact of their contributions following a particular disaster and not sometime in the future. In addition, the earmarked funding cannot be easily used by the organization for any other disaster. To some extent, prepositioning is also a kind of insurance, and there is nothing to “show” if it is not utilized. Although most decision-makers agree on the virtues of prepositioning, the strategy cannot be implemented until the issues around funding are resolved.

Demand estimation constitutes a significant part of prepositioning. The issues to be addressed are quantity, location, and timing of the emergency supplies to be delivered and services to be provided [3, 12]. The mismatch between demand and supply can lead to an inefficient and ineffective disaster relief [2, 13–15]. Demand of the affected population in the future, therefore, can be termed as essential to estimate yet hard to forecast. We, in this article, owing to scope, do not focus on estimation of demand but acknowledge the significant role it plays in prepositioning. Future research in forecasting of demand can play a substantial role.

CONCLUSION

Given the scope and depth of the issues in humanitarian logistics, there are still many unresolved issues—such as the

adequate establishment and maintenance of prepositioned disaster relief inventory; reliability of the potential, virtual, and real supply lines; and decisions about local (but outside the disaster zone) or imported sourcing of supplies—that can be perceived as opportunities to extend the existing literature cited in the first section with further studies. Owing to the scope of this article, there are many details of the prepositioning approach that cannot be addressed here, especially in regard to the inventory to be prepositioned—the type, quantity, and rotation from shelf-life for perishability—as well as the pros and cons of prepositioning. Each one of these topics deserves an article in its own right.

There are strategies other than prepositioning for humanitarian assistance and disaster relief. For reasons discussed at the beginning of this article, we have focused on prepositioning for seasonal natural disasters and have offered our perspective. Further studies could consider how organizations mentioned in this research deal with all types of disasters and not just seasonal natural disasters. Such studies could also enhance the understanding of the prepositioning approaches used for the Strategic National Stockpile, as carried out by the Centers for Disease Control and Prevention.

The scope of the operations, the high uncertainty, and the urgency in the face of inadequate infrastructure are a few of the challenges in humanitarian operations that can be mitigated by sound strategies, specifically prepositioning. Prepositioning can manage the breadth of these challenges through the near-enough location of actual supplies and the availability of capacities for various resources. High uncertainty of demand due to a disaster can be dealt with by prepositioning the majority of relief supplies that are imperishable and commonly necessary in most disasters. Urgency can be eased by reducing the lead time because of the proximity of stockpiles or attention to capacity planning. Strategic prepositioning can improve the inadequate infrastructure.

We outlined and demonstrated our position to shed some light on prepositioning

approaches for relief in case of natural seasonal disaster by addressing two key parts of it: operational and strategic. We described examples to illustrate the development and solution processes that can be used to analyze similar situations. There are many considerations in developing a prepositioning approach, as we explained in the caveats section. These challenges could be explored further.

After studying the literature—research articles and case studies—our findings indicate that, when both the modeling and data employed have been based on input and feedback from experts, the prepositioning models—either operational or strategic—are extremely beneficial for humanitarian relief in the case of seasonal natural disasters. This in turn helps with critical decision-making in humanitarian organizations such as CARE [14] and FEMA [43].

REFERENCES

1. Apte A. Humanitarian logistics: a new field of research and action. *Found Trends© Technol Inf OM* 2009;3(1):1–100.
2. Apte A, Yoho K.. Strategies for Logistics in Case of a Natural Disaster, 40th Annual Meeting of Western Decision Sciences Institute, 2011.
3. Çelik M, Ergun Ö, Johnson B, *et al.* Humanitarian logistics. In: Smith JC, editor-Chapter 2. *Tutorials in operations research*. Hanover: INFORMS; 2012. pp. 18–49. <http://dx.doi.org/10.1287/educ.1120.0100>.
4. Trunick JA. Logistics when it counts. *Logist Today* 2005;46(2):38.
5. Eisner, R.. Fritz Institute, San Francisco, CA, Private Communication. 2007
6. Fenton, R.. Regional Response and Recovery Director of Federal Emergency Management Agency, USA. Private Communication. 2008.
7. Nelan D. Terrorism & disaster response. *Homeland Security Second Annual Conference and Showcase*, Naval Postgraduate School, Monterey, CA; 2008.
8. Van Wassenhove LN. Humanitarian aid logistics: supply chain management in high gear. *J Oper Res Soc* 2006;57(5):475–489.

9. Kovacs G, Spens KM. Humanitarian logistics in disaster relief operations. *Int J Phys Distrib Logist Manag* 2007;37(2):99–114.
10. Holguin-Veras JH, Perez N, Ukkusuri SV, *et al.* Emergency logistics issues in Hurricane Katrina: a synthesis and preliminary suggestions for improvement. *Transp Res Rec J Transp Res Board* 2007;2022:76–82.
11. Long D, Wood D. The logistics of famine relief. *J Bus Logist* 1995;16(1):213–229.
12. Cort J, Jain N, Kapadia R, Knepper M, Knocke J, McCutchen T.. CARE USA-Demand Estimation and Emergency Procurement, Final report, ISyE 4106 Senior Design, Fall 2009, Georgia Institute of Technology; 2009.
13. United Nations. United Nations Human Development Report 2007/2008—Fighting Climate Change: Human Solidarity in a Divided World United Nations Human Development Program. Palgrave Macmillan, New York; 2007.
14. Duran S, Gutierrez MA, Keskinocak P. Pre-positioning of emergency items for CARE international. *Interfaces* 2011;41(3):223–237.
15. World Meteorological Organization. 2009. Fact sheet: Climate information for reducing disaster risk. http://www.wmo.int/wcc3/documents/WCC3_factsheet1_disaster_EN.pdf. Accessed June 2013.
16. McCoy JH. Humanitarian response: improving logistics to save lives. *Am J Disaster Med* 2008;3(5):283–293.
17. Apte A, Yoho K, Greenfield C, Ingram C. Selecting Maritime Disaster Response Capabilities, *Journal of Supply Chain Management*, 06(02), July–December; 2013, 40–58.
18. Kent J Jr, Flint D. Perspectives on the evolution of logistics thought. *J Bus Logis* 1997;18(2):15–29.
19. Townsend FF. The federal response to hurricane Katrina: lessons learned. Washington, D.C.: The White House; 2006. <http://georgewbush-whitehouse.archives.gov/reports/katrina-lessons-learned/letter.html> Accessed June 2013.
20. Heidtke CL. Reducing the ‘gap of pain’: a strategy for optimizing federal resource availability in response to major incidents. [M.A. Thesis in Security Analysis (Homeland Security and Defense)]. Naval Postgraduate School, Monterey, CA, 2007.
21. Rawls CG, Turnquist MA. Pre-positioning and dynamic delivery planning for short-term response following a natural disaster. *Socio-Econ Plan Sci* 2012;46:46–54.
22. Beamon BM, Kotleba SA. Inventory management support systems for emergency humanitarian relief operations in South Sudan. *Int J Logist Manag* 2006;17(2):187–212.
23. Bravata DM, Zaric GS, Holty JC, *et al.* Reducing mortality from anthrax bioterrorism: strategies for stockpiling and dispensing medical and pharmaceutical supplies. *Biosecur Bioterror Biodef Strategy Pract Sci* 2006;4(3):244–262.
24. Lodree EJ Jr., Taskin S. Supply chain planning for hurricane response with wind speed information updates. *Comput Oper Res* 2009;36(1):2–15.
25. Feng T, Keller LR. A multiple objective decision analysis for terrorism protection: potassium iodide distribution in nuclear incidents. *Decis Anal* 2006;3(2):76–93.
26. Dani S. Predicting and managing supply chain risks. In: Zsidisin GA, Ritchie B, editors. *Supply chain risk: a handbook of assessment, management, and performance*. New York: Springer 2008. pp. 53–66. ISBN: 978-0-387-79934-6.
27. Mulla A. Risk management system—a conceptual model. In: Zsidisin GA, Ritchie B, editors. *Supply chain risk: a handbook of assessment, management, and performance*. New York: Springer; 2008.
28. Khan O, Christopher M, Burnes B. The role of product design in global supply chain risk management. In: Zsidisin GA, Ritchie B, editors. *Supply chain risk: a handbook of assessment, management, and performance*. New York: Springer; 2008.
29. Wagner SM, Bode C. Dominant risks and risk management practices in supply chains. In: Zsidisin GA, Ritchie B, editors. *Supply chain risk: a handbook of assessment, management, and performance*. New York: Springer 2008; 2008.
30. Hendricks KB, Singhal VR. Managing disruptions in contemporary supply chains. In: Gattorna J, editor. *Dynamic supply chain alignment*. Aldershot: Gower; 2009. pp. 85–96.
31. Khan S, Richter A. Pilot model: judging alternate modes of dispensing in Los Angeles county. *Interfaces* 2009;39:228–240.
32. Liang C, Wang C, Luh H, *et al.* Disaster avoidance mechanism for content-delivering service. *Comput Oper Res* 2009;36:27–39.
33. Brown R, Schank J, Dahlman C, *et al.* Assessing the potential for using reserves in

- operations other than war *MR796*. Santa Monica, CA: *The RAND Corporation*; 1997.
34. Welser W IV., Yoho K, Robbins M, *et al.* Assessment of the USCENTCOM medical distribution structure *MG-929-A*. Santa Monica, CA: *The RAND Corporation*; 2010.
 35. Campbell A, Jones P. Pre-positioning supplies in preparation for disasters. *Eur J Oper Res* 2011;209(2):156–165.
 36. Pettit SJ, Beresford AK. Emergency relief logistics: an evaluation of military, non-military and composite response models. *Int J Logist Res Appl* 2005;8(4):313–331.
 37. Karmarkar U. 2011. LA Times Professor of Technology and Strategy, Anderson School of Business, University of California Los Angeles, USA. *Private Communication*.
 38. Chang MS, Tseng YL, Chen JW. A scenario planning approach for the flood emergency logistics preparation problem under uncertainty. *Transp Res E* 2007;43:737–754.
 39. Balcik B, Beamon B. Facility location in humanitarian relief. *Int J Logist Res Appl* 2008;11(2):101–121.
 40. Mete H, Zabinsky Z. Stochastic optimization of medical supply location and distribution in disaster management. *Int J Prod Econ* 2010;46(2):273–286.
 41. Horner MW, Downs JA. Optimizing hurricane disaster relief goods distribution: model development and application with respect to planning strategies. *Disasters* 2010;34(3):821–844.
 42. Hale T, Moberg CR. Improving supply chain disaster preparedness: a decision process for secure site location. *Int J Phys Distrib Logist Manag* 2005;35(3):195–207.
 43. Dekle J, Lavieri MS, Martin E, *et al.* A Florida county locates disaster recovery centers. *Interfaces* 2005;35(2):133–139.
 44. McCall VM.. Designing and prepositioning humanitarian assistance pack-up kits (HA PUKs) to support pacific fleet emergency relief operations [Master's Thesis]. Naval Postgraduate School, Monterey, California; 2006.
 45. Cataldi M, Cho C, Guterriez C, Hull J, Kim P, Park A, Pickering J, and Swann J. 2009. Operations research bites back: improving malaria interventions in Africa, Working paper, Georgia Institute of Technology.
 46. Lee EK, Smalley HK, Zhang Y, *et al.* Facility location and multi-modality mass dispensing strategies and emergency response for biodefense and infectious disease outbreaks. *Int J Risk Assess Manag Biosecur* 2009;12(2/3/4):311–351.
 47. Rawls CG, Turnquist MA. Pre-positioning of emergency supplies for disaster response. *Transp Res B* 2010;44:521–534.
 48. Rawls CG, Turnquist MA. Pre-positioning planning for emergency response with service quality constraints. *OR Spectr* 2011;33(3):481–498.
 49. Owen SH, Daskin MS. Strategic facility location: a review. *Eur J Oper Res* 1998;111:423–447.
 50. Salmeron J, Apte A. Stochastic optimization for natural disaster asset pre-positioning. *Prod Oper Manag* 2010;19(5):561–575.
 51. Ergun O, Karakus G, Keskinocak P, *et al.* Operations research to improve disaster supply chain management. In: Cochran JD, editor. *Encyclopedia of operations research and management science*. New York: John Wiley & Sons, Inc.; 2011. pp. 3802–3810.
 52. Tzeng GH, Cheng HJ, Huang TD. Multi-objective optimal planning for designing relief delivery systems. *Transp Res E Logist Transp Rev* 2007;43(6):673–686.
 53. Döyen A, Aras N, Barbarosoğlu G. A two-echelon stochastic facility location model for humanitarian relief logistics. *Optim Lett* 2012;6(6):1123–1145.
 54. Berman O, Krass D, Menezes MBC. Facility reliability issues in network p -median problems: strategic centralization and co-location effects. *Oper Res* 2007;55(2):332–350.
 55. Shen ZJM, Zhan RL, Zhang J. The reliable facility location problem: formulations, heuristics, and approximation algorithms. *INFORMS J Comput* 2011;23(3):470–482.
 56. Nemhauser GL, Wolsey LA. *Integer and combinatorial optimization*. Upper Saddle River, NJ: John Wiley & Sons; 1999.
 57. Church RL, ReVelle CS. The maximal covering location problem. *Pap Reg Sci Assoc* 1974;32:101–118.
 58. Brandeau ML, Chiu SS. An overview of representative problems in location research. *Manag Sci* 1989;35(6):645–674.
 59. Klose A, Drexl A. Facility location models for distribution system design. *Eur J Oper Res* 2005;162(1):4–29.
 60. Marianov V, ReVelle C. Siting emergency services. In: Drezner Z, editor. *Facility location: a survey of applications and methods*. New York, NY: Springer-Verlag; 1995. pp. 199–223.

61. Caunhye A, Nie X, Pokharel S. Optimization models in emergency logistics: a literature review. *Socio-Econ Plan Sci* 2012;46:4–13.
62. Hakimi S. Optimum distribution of switching centers in a communications network and some related graph theoretic problems. *Oper Res* 1965;13:462.
63. Cabot A, Francis R, Strary M. A network flow solution to a rectilinear distance facility location problem. *AIIE Transact* 1970;2:132–141.
64. Toregas CR, Swain R, ReVelle CS, *et al.* The location of emergency service facilities. *Oper Res* 1971;19:1363–1373.
65. Fitzsimmons J. A methodology for emergency ambulance deployment. *Manag Sci* 1973;19:627–636.
66. Tansel BC, Francis RL, Lowe TJ. Location on networks: a survey. Part I: the p-center and p-median problems. *Manag Sci* 1983;29(4):482–497.
67. Shmoys DB, Tardos E, Aardal KI. Approximation algorithms for facility location problems. *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pp. 265–274; 1997.
68. Snyder L. Introduction to facility location. In: Cochran JD, editor. *Encyclopedia of operations research and management science*. New York: John Wiley & Sons, Inc.; 2011. pp. 2457–2474.
69. Tean ES. Optimized positioning of pre-disaster relief force and assets [M.S. Thesis in Operations Research]. Naval Postgraduate School, Monterey, CA; 2006.
70. Disparte D. The postman parallel. *CarNation* 2007;2:22–27.
71. International Business Times, 2010. <http://www.ibtimes.com/humanitarian-operations-challenges-fleet-management-191668#>, Accessed 2013 June.
72. Pedraza-Martinez, AJ, Hasija S and Van Wassenhove LN. 2013. Fleet Management Coordination in Decentralized Humanitarian Operations, Working Paper.
73. Pedraza-Martinez AJ, Van Wassenhove LN. Vehicle replacement in the international committee of the red cross. *Prod Oper Manag* 2009;22(2):365–376.
74. Pedraza-Martinez AJ, Stapleton O, Van Wassenhove LN. Field vehicle fleet management in humanitarian operations: a case-study based approach. *J Oper Manag* 2011;29(5):404–421.
75. Ukkusuri SV, Yushimito WF. Location routing approach for the humanitarian prepositioning problem. *Transp Res Rec J Transp Res Board* 2008;2089:18–25.
76. Ichoua S. Humanitarian logistics network design for an effective disaster response. *Proceedings of the 7th International ISCRAM Conference Seattle*, Vol. 1; 2010.
77. Tong H-M. Plant location decisions of foreign manufacturers. Ann Arbor: University Microfilms International Press; 1979.
78. Jungthirapanich C, Benjamin C. A knowledge base decision support system for locating a manufacturing facility. *IIE Trans.* 1995;27:789–99.
79. MacCarthy B, Atthirawong W. Factors affecting location decisions in international operations—a Delphi Study. *Int J Oper Prod Manag* 2003;23(7):794–818.
80. Mazzarol T, Choo S. A study of the factors influencing the operating location decisions of small firms. *Proper Manag* 2003;21(2):190–208.
81. Ozdamar L, Ekinci E, Kucukyazici B. Emergency logistics planning in natural disasters. *Ann Oper Res* 2004;129:217–245.
82. Galindo G, Batta R. Prepositioning of supplies in preparation for a hurricane under potential destruction of prepositioned supplies. *Socio-Econ Plan Sci* 2013;47(1):20–37.
83. Thornton B. Disaster preparedness, response, and post-disaster operations Presentation at Conference on Humanitarian Logistics Conference. Atlanta, United States: H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech; 2009.
84. Balcik B, Beamon BM, Smilowitz K. Last mile distribution in humanitarian relief. *J Intell Transp Sys* 2008;12(2):51–63.
85. Daskin MS, Haghani AE, Khanal C, *et al.* Aggregation effects in maximum covering models. *Ann Oper Res* 1989;18:115–139.
86. Current JR, Schilling DA. Analysis of errors due to demand data aggregation in the set covering and maximal covering location problems. *Geogr Anal* 1990;22(2):116–126.
87. Francis RL, Lowe TJ, Rushton G, *et al.* A synthesis of aggregation methods for multi-facility location problems: strategies for containing errors. *Geogr Anal* 1999;31(1):67–87.
88. Asia-Pacific Task Force for Emergency Preparedness. 2009. <http://www.apec.org/Groups/SOM-Steering-Committee-on-Economic-and-Technical->

- Cooperation/Working-Groups/Emergency-Preparedness.aspx, Accessed 2013 June.
89. Regnier E. Public evacuation decisions and hurricane track uncertainty. *Manag Sci* 2008;54(1):16–28.
90. Regnier E. Evacuation planning. In: Cochran JD, editor. *Encyclopedia of operations research and management science*. New York: John Wiley & Sons, Inc.; 2011. pp. 1745–1756.

STRATEGIC CUSTOMER BEHAVIOR IN A SINGLE SERVER QUEUE

MOSHE HAVIV

Department of Statistics, The Hebrew
University of Jerusalem,
Jerusalem, Israel

INTRODUCTION

Usually when one considers a queueing model, such as an M/M/1 queue, both arrival and service rates are given. Sometimes an optimization question regarding these parameters is also posed. For example, denote the arrival and service rates by λ and μ , respectively. Suppose the service rate of μ comes with a cost of $c(\mu)$ per unit of time whenever the server is available (and not only while he/she is actually utilized) and it costs C per customer per unit of time in the system. Assume λ to be fixed. Then, one looks for the value of $\mu > \lambda$, which minimizes $c(\mu) + \lambda C/(\mu - \lambda)$, the total social cost per unit of time. In this article we move one step further: it is the individual customers, rather than a central planner, who decide on their actions. Moreover, they do that by optimizing their selfish needs, while ignoring the effect of their moves on others, known as *externalities*. For example, for a given μ , they may decide whether or not to join the queue. It is only their aggregate behavior, which leads to the completion of the model via determining the value of λ . Commencing with the seminal paper [1], this branch of research got much attention in the literature. A state-of-the-art summary (until 2003) on it is given in Ref. 2. We describe below, some of the main results and ideas in this field while considering a few key examples. We limit ourselves to the M/M/1 model with homogeneous (with respect to their cost and rewards parameters) customers. In fact, in this case, strategic behavior is more subtle than in the heterogeneous case.

Customers who consider joining a queue or who are already waiting, may face various types of dilemmas: to join or not to join a queue, to pay or not to pay a fee in order to belong to a high priority class, or whether or not to renege after waiting for a while. We assume that they are selfish and hence maximize their own utility, not minding the externalities that their moves impose on others. While deliberating, they are aware of the fact that other customers may face the same type of dilemma. An individual customer wants to select his/her optimal action, but which action is best depends on what others are doing. For example, if all do not join the queue, it would be better for one to join as this will come with no need to wait, while if all do join, one may face too much waiting and may do best by not joining. Such decision-making problems are best described as noncooperative games. We do not attempt to review, thoroughly, this model. The interested reader is referred to any text in game theory, such as Refs 3 and 4.

In noncooperative games one's utility for any given strategy, is a function of the strategies selected by others. If one's best strategy is not a function of what others do, one's strategy is called a *dominant strategy*. However, in many cases such strategies do not exist and we look for a *pure Nash equilibrium* strategy profile. This refers to a set of strategies, one for each individual, such that each one of them is the best for the corresponding individual, given that all others follow their part in the profile. Yet, the problem may persist: there may not be such a profile.¹ Enlarging the strategy space by allowing mixing (namely, players select a lottery or a probability distribution over their set of pure strategies and assess their utility given their behavior and those of others by the corresponding expected value, when all lotteries

¹As we see below, there is another possible problem: multiple pure equilibria profile may exist.

are carried out independently), may guarantee the existence of a Nash equilibrium. Of course, a pure equilibrium is a special case of a mixed equilibrium.

In all models dealt with below, we assume a symmetric game. This indicates that all customers share the same set of strategies and the same utility function. In particular, customers are homogeneous with respect to their cost/reward parameters. Under a symmetric strategy profile all use the same strategy. Next, for any given such strategy, all assume steady-state conditions under this behavior and all assess their utilities accordingly. Then, we look for a *symmetric Nash equilibrium*, namely an equilibrium profile under which all use the same strategy. More formally, let S be the common set of mixed strategies and let $u(x, y)$ be one's utility if one selects strategy $x \in S$, given that all others select strategy $y \in S$.² Thus, x_e is defined as a symmetric Nash equilibrium if

$$x_e \in \arg \max_{y \in S} u(y, x_e).$$

From now on when we refer to an equilibrium, we mean a symmetric Nash equilibrium.

Let $v(x)$ be the social aggregate utility when all use the strategy x . Typically, $v(x)$ is defined as the sum (or integral) over the individual utilities. A social optimizer looks for an x^* such that

$$x^* \in \arg \max_{x \in S} v(x).$$

Usually, x_e and x^* do not coincide. A mechanism design question is how to manipulate the system so as the resulting new x_e in the manipulated system coincides with the old x^* . This can sometimes be achieved by charging prices on various behaviors. In other cases, more complicated schemes are called for.

Our basic model is that of an M/M/1 queue, namely customers arrive in accordance to a Poisson process with rate Λ and their service times are independent and exponentially

distributed with mean μ^{-1} . Unless stated otherwise, service is granted on a first-come first-served (FCFS) basis. Customers suffer a cost of C per unit of time in the system (inclusive of service) and are rewarded by R due to service completion. In order to avoid trivialities, assume that $R > C/\mu$. Without loss of generality, assume that not joining comes with zero utility.

We are not claiming to supply a comprehensive survey on the field of customers' behavior in queues. Rather, we would like to give the flavor of the issues dealt with, while considering some of the most important queueing decision models. Note that we restrict ourselves to the case where customers are homogeneous with respect to their parameters C and R . Many models where this assumption is removed are also surveyed in Ref. 2. We point out that as opposed to strict optimization problems, the homogeneous case is not a special case of the nonhomogeneous case. In fact, it is usually more difficult for analysis as it calls for the use of mixed strategies under equilibrium, while in the fully heterogeneous case (where no two customers share the same parameter values such as R and/or C), the equilibrium is usually pure under which all use a parameter-dependent pure strategy.

TO QUEUE OR NOT TO QUEUE

The most basic question customers ask is whether or not to join the queue. We next deal with two different scenarios. In the first, customers do not observe the queue while making their decisions and once they join, they cannot change their mind if upon arrival they observe a longer queue than expected. In the second, they observe the queue length upon their arrival and decide whether or not to join based on this information. In both cases we also look at the socially optimal behavior and look for ways to regulate the system by levying appropriate tolls so that the resulting equilibrium behavior coincides with the socially optimal behavior. Finally, we discuss the "avoid the crowd" phenomenon. The results from the unobservable model appear originally in Ref. 5, while

²By utility here and throughout the rest of the article we refer to the expected utility when the two lotteries are carried out independently.

those from the observable case are from Ref. 1. More comprehensive summaries appear, respectively, in Chapters 3 and 2 of Ref. 2.

The Unobservable Case

Equilibrium Behavior. If $\Lambda < \mu$ and all join, then everybody's utility from joining is $C/(\mu - \Lambda) - R$. Thus, if this utility is positive, then joining is a dominant strategy. Indeed, no matter which fraction p of customers join, resulting in a Poisson arrival process with a rate of Λp and with a mean waiting time of $1/(\mu - p\Lambda)$, one's best action is to join. This is not the case where $\Lambda \geq \mu$ or where $\Lambda < \mu$ but $R < C/(\mu - \Lambda)$. Here, a dominant strategy does not exist: if all join, one better not join, and if none of them join, one better join since, as assumed, $R > C/\mu$. The equilibrium strategy is to join with probability p for p obeying $R - C/(\mu - p\Lambda) = 0$. Indeed, under such a p , $0 < p < 1$, if used by all, one is indifferent between joining and not joining, and hence mixing between joining and not joining with probabilities p and $1 - p$, respectively, is a best response for an individual. Such a p is unique. Denoting it by p_e , then

$$p_e = \frac{\mu - \frac{C}{R}}{\Lambda}.$$

Also note that when $p_e < 1$ the utility of all customers, those who join and those who do not, equals zero. No other symmetric equilibrium exists.

To Avoid or to Follow the Crowd. Many decision problems stemming from queueing models possess exactly one of the following two phenomena, that is, when an individual should or should not follow the crowd. Specifically, suppose the set of strategies can be presented by an interval, that is, each strategy corresponds to a number in this interval. This is certainly the case when only two pure strategies are available and the probability with which one option is selected, describes the mixed strategy completely. In particular, the unit interval can be identified with the set of strategies. Let $u(y, x)$ be as defined in the introduction, but now x and y have a numerical value. Define $y^*(x) \in \arg \max_y u(y, x)$. In other words, $y^*(x)$

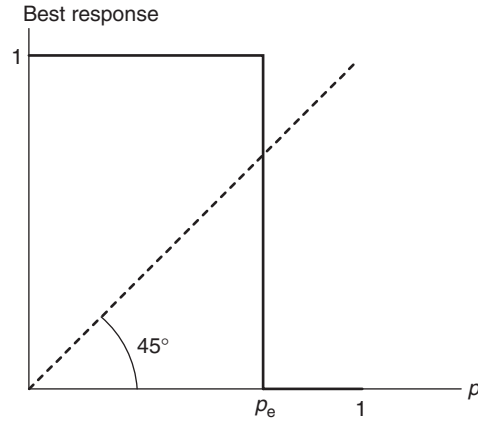


Figure 1. Avoid the crowd.

is one's *best response* (ties are possible) when all use strategy x . We say that the phenomenon of avoid the crowd (ATC) and follow the crowd (FTC) prevail, when $y^*(x)$ is monotone nonincreasing and nondecreasing, respectively.

The situation we face in the “to queue or not to queue” decision problem is that of ATC. Specifically, suppose all join with probability p . Then $u(q, p)$ is one's utility when one joins with probability q . Note that

$$u(q, p) = \begin{cases} q \left(R - \frac{C}{\mu - p\Lambda} \right), & p\Lambda < \mu \\ -\infty, & p\Lambda \geq \mu. \end{cases}$$

It is easy to see that $q^*(p) = 0$ when $p > p_e$ and $q^*(p) = 1$ when $p < p_e$.³ When $p = p_e < 1$, all strategies are best responses against p_e . As it turns out, symmetric equilibria are the points where the best response function $q^*(p)$ intersects the line $q = p$. By definition, in the ATC case, the best response function is not increasing. The function $q = p$ is of course increasing with p . Hence, the two continuous functions, $q^*(p)$ and p , intersect at exactly one point. In other words, a unique symmetric equilibrium exists in the ATC case (Fig. 1). Things are more involved in the FTC cases as we exemplify later.

³In the case where $p = 1$ is dominant, $p_e = 1$.

Socially Optimal Behavior. In case of a joining probability of p and assuming that $p\Lambda < \mu$, the social gain per unit of time equals

$$p\Lambda \left(R - \frac{C}{\mu - p\Lambda} \right). \quad (1)$$

Thus, a social optimizer looks for the value of p , $0 \leq p \leq 1$, which maximizes Equation (1). Denote this value by p_s . It is easy to see that

$$p_s = \begin{cases} 1, & \Lambda < \mu - \sqrt{\frac{C\mu}{R}} \\ \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\Lambda}, & \text{otherwise.} \end{cases}$$

In the case where $p_s = 1$, the social gain is as stated in Equation (1) with $p = 1$. When $p_s < 1$, it equals

$$(\sqrt{R\mu} - \sqrt{C})^2. \quad (2)$$

It is also possible to see that unless $p_s = 1$, $p_s < p_e$. In other words, left for themselves customers tend to join at a rate, which is higher than is socially desired. This is the case since joining customers ignore the *negative externalities*, in terms of added waiting time to future arrivals, that those who join inflict on each other.

In the case where $p_s < p_e$, by artificially reducing R or increasing C , one can make the new p_e and the old p_s coincide. It is a simple exercise to see that the amount to reduce R with, is $\sqrt{CR/\mu}$, while the amount to increase C with, is $\sqrt{RC\mu} - C$. The former is achieved by imposing an entry fee of $\sqrt{CR/\mu}$ while the latter is achieved by charging customers an extra $\sqrt{RC\mu} - C$ per unit of time in the system, making customers suffer in fact a cost of $\sqrt{RC\mu}$ (instead of C) per unit of waiting time. The collected fees under both charging schemes go to the society's coffer. In both cases individual customers end up with zero utility. Under this optimal entry toll, the social planner collects $(\sqrt{R\mu} - \sqrt{C})^2$ per unit of time. In other words, all consumer surplus as appearing in Equation (2), is transferred completely from the individual customers to the social purse. Such complete transfer of consumer surplus can be achieved when individuals are not more informed than the central planner.

The Observable Case

The model dealt with here is as in the previous section, but with one key distinction: customers observe the queue length upon their arrival and decide whether or not to join accordingly. Thus, a pure strategy in this case means a recipe, which for any possible queue length, prescribes joining or not joining. Assuming that nobody leaves the system after joining it in the first place, and recalling that the service distribution is memoryless, a sensible symmetric strategy will be of the threshold type. By that we mean that there exists some integer value $n \geq 1$, such that all join if and only if the queue length they observe upon arrival is less than n [6].⁴ As a consequence of the FCFS discipline in this observable model, if an individual's utility is not a function of unknown actions taken (or to be taken) by others, a dominant strategy exists.

Equilibrium and Socially Optimal Behavior.

First, the threshold equilibrium behavior is trivial: join if and only if the number in the system is smaller than n_e where⁵

$$n_e = \left\lfloor \frac{R\mu}{C} \right\rfloor. \quad (3)$$

In fact, this strategy is dominant. As opposed to the unobservable case (when Λ is large enough), those who join end up with a positive utility. Second, finding the socially optimal threshold n_s is more cumbersome, but this was done in paper [1].⁶ In particular, n_s is the unique integer value of $n \geq 1$ obeying $g(n-1) \leq \frac{R\mu}{C} \leq g(n)$ where $g(n) = \frac{n(1-\rho) - \rho(1-\rho^n)}{(1-\rho)^2}$. It was shown in Ref. 1 that $n_s \leq n_e$ with equality if and only if $n_e = 1$. By definition, under social optimization the consumer surplus is on average larger than it is under the equilibrium behavior.

⁴Yet, things might be a bit more involved. See Ref. 6 for a formal treatment for this issue.

⁵In the case where $R\mu/C$ is an integer, both n_e and $n_e - 1$ as any mixing between them, define an equilibrium profile.

⁶See Ref. 2, pp. 27–29, for an alternative proof.

As in the unobservable case, individual and social behavior will be aligned by imposing an entry fee or by charging a fee per unit time in the system so that the new n_e coincides with the original n_s . From Equation (3), we learn that any entry fee T obeying

$$n_s = \left\lfloor \frac{(R + T)\mu}{C} \right\rfloor$$

is suitable for this purpose. In other words,

$$\frac{Cn_s}{\mu} - R \leq T < \frac{C(n_s + 1)}{\mu} - R.$$

Under any socially optimal entry fee some of the consumer surplus stays in the hands of the individuals. This is the case since those who join, even after paying the entry fee, end up with a positive utility: The shorter the queue they join, the larger is their surplus. This is the case since customers can adjust their behavior to the actual queue length, while the central planner is limited to a single, one for all price. Put differently, customers are more informed than the social planner is. On the other hand, if a queue length-dependent entry fee is imposed, then all of the consumer surplus can be transferred to the common coffer. Indeed, charging (a bit less than) p_n for one who joins the system when n customers are present, with

$$p_n = \begin{cases} R - \frac{C(n+1)}{\mu} & 0 \leq n \leq n_s - 1 \\ \infty & \text{otherwise.} \end{cases}$$

makes customers behave socially optimal while leaving no surplus in their hands. For more on this pricing see Ref. 7.

Another interesting way to elicit the socially optimal behavior was suggested in Ref. 8. Instead of an FCFS policy, a joining customer is placed at any position but the last one. A possible name for this queueing regime is *not-FCFS*. This allows preempting the one who is in service. In fact, in the case where only one customer is in the system, preempting him/her is the only option. Now all join, but they have the option to renege (i.e., to abandon) as they observe how many

are in front of them. An equilibrium strategy is to renege from the rear of the queue as soon as the total number in the system reaches $n_s + 1$. This leaves the number in the system as it is socially optimal. The reason for this alignment between individual and social behaviors is the fact that the one in the rear of the queue (the one who in fact needs to take decisions) inflicts no externalities on the others under the not-FCFS service policy. Hence, his/her interests and those of the society coincide.

The Case of Processor Sharing or Random Queues

In the case where service is granted so as the next to commence or to conclude service is decided randomly among all those in line was dealt with in Refs 9 and 10. Specifically, a random queue is one where the FCFS discipline is replaced by letting the next to commence service being decided randomly among all those present in line at the epoch of service completion. The processor sharing scheme is such that the server splits in capacity evenly among all those who are present at the system. Yet, in the case of exponential service times, this is equivalent to performing service, but deciding to whom it is granted at random among all those present only in its completion. It is clear that one should better not stay in line if it is too long. In Ref. 10 it is assumed that the one to renege is the last to arrive (who in fact balks upon arrival). Thus, the question here is in fact to queue or not to queue. As expected, an equilibrium policy prescribes joining if and only if the queue length is smaller than some threshold. Yet, since waiting times are also functions of later to arrive customers, true interaction between customers' policies exists here. Specifically, an ATC prevails here and the unique equilibrium policy can be pure or mixed. In the latter case, mixing is among two consecutive threshold queue lengths.

In Ref. 9 it is assumed that as soon as one arrives, he/she is considered as all others. Thus, in case that one needs to leave, all those in line should be treated as equals. It is shown that in this model no equilibrium exists. Yet, an ϵ -equilibrium, for any $\epsilon > 0$

exists.⁷ Under this profile, all stay as long as the total number in queue is smaller than some threshold, renege with some rate when at the threshold (meaning reneging after exponentially distributed time if no change in the number in queue took place in the meanwhile), and when the threshold is crossed, renege at a very high rate (equivalently, in the limit, to performing a lottery, which decides who among those in queue should leave immediately). This is a two-parameter policy (threshold queue length and reneging rate). Ref. 9 shows how to determine it.

TO UPGRADE OR NOT TO UPGRADE

We go on dealing with the previous model but with two key changes. First, customers have the option of belonging to a premium class. Those who belong to this class have preemptive priority over the others, called *regular customers*. Among customers belonging to the same class, FCFS is assumed. Yet, belonging to the premium class comes with a cost of θ . No payment is required for those elected to be regular customers. Second, customers do not have the option of not joining. Thus, from the decision problem point of view of paying or not for priority, the value of service is irrelevant. For stability, we need to assume that $\Lambda < \mu$. To keep our notation standard with the literature, we replace, here, Λ with λ .

The trade-off is between paying for priority and having to wait longer in case of not paying. As all other customers face the same decision problem and as one's waiting time in both classes is a function of the decisions taken by others, the decision model is that of a noncooperative game. In the next section we deal with the unobservable version of the model. Here the strategy space is simple: how to mix between paying and not paying for priority. In the subsequent section we deal with the observable version. There, a pure strategy is more involved, as, for any number of customers from the two groups one may face

upon arrival, one needs to decide if to purchase priority or not.⁸ In the final section titled "How Much To Pay for an Upgrade," we deal again with the unobservable case, but now customers can pay as much as they wish while the more one pays, the higher is one's priority level.

The Unobservable Case

The following appears in Ref. 2. Assume that the symmetric mixed strategy of purchasing priority with probability p is used by all. Then, by standard results from priority queues, the mean time in the system of premium customers is $1/(\mu - \lambda p)$ and that of regular customers is $\mu/[(\mu - \lambda)(\mu - \lambda p)]$. Hence, in this case the added value (which can be negative) from belonging to the premium class in comparison with belonging to the regular class is

$$f(p) \equiv \frac{\rho C}{\mu(1 - \rho)(1 - \lambda p)} - \theta.$$

It is easy to see that $f(p)$ is monotone increasing with p . In other words, the more customers purchase priority, the more priority is valuable to an individual. This is an example of the FTC phenomenon.⁹

The next question is what are the equilibrium values for p . There are three mutually exclusive and exhaustive possibilities:

Case 1: $f(0) \geq 0$. Since $f(p)$ is monotone increasing, it would be better to purchase priority irrespective of how many others do so. In other words, purchasing priority is a dominant strategy. In particular, $p_e = 1$. See Fig. 2a for the best response function.

⁸True, the number of premium customers present in line is irrelevant.

⁹As opposed to "to queue or not to queue" problem where it is *a priori* clear that the ATC phenomenon prevails, the situation here is more subtle. Indeed, the more purchase priority, the less is the number of regular customers to be overtaken by a premium customer (hinting toward ATC). Yet, the more purchase priority, the more one needs priority for oneself in order not to be overtaken by later to arrive premium customers (hinting toward FTC). The latter effect seems to "win" among these two antagonistic forces.

⁷An ϵ -equilibrium, is a strategy profile such that if followed by all, the one who deviates from it can gain at most ϵ .

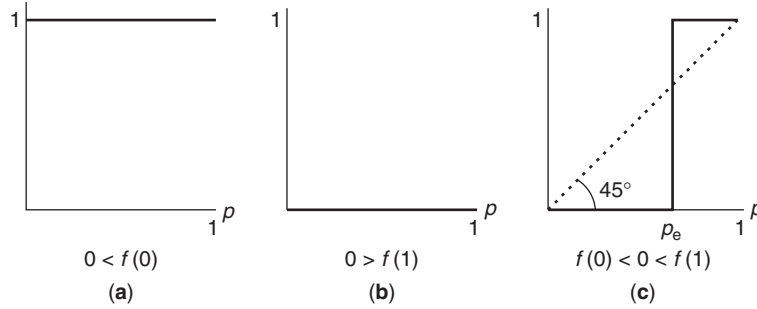


Figure 2. Follow the crowd.

Case 2: $f(1) \leq 0$. Owing to an argument similar to the one stated in the previous item, not purchasing priority is a dominant strategy. In particular, $p_e = 0$. See Fig. 2b for the best response function.

Case 3: $f(0) < 0 < f(1)$. The first thing to observe is that the two pure strategies define equilibria. Indeed, when all or none purchase priority, it is optimal for an individual to follow suit. But there is a third equilibrium. Let p_e be the unique value such that $f(p_e) = 0$. Of course, such a value exists under the assumption of this case. If all use this strategy, one is indifferent to purchasing priority or not, and hence purchasing priority with probability p_e is one's best response. No other equilibrium exists.¹⁰ The best response function is depicted in Figure 2c. Specifically, when $p < p_e$ the unique best response is 0. It is 1 when $p > p_e$, and any mixing is a best response against all using p_e . Thus, the best response function is nondecreasing. It can hence intersect the 45 degree line more than once. Indeed, this occurs three times: at 0, p_e , and 1. In other words, there are three equilibria as stated in the third item above. Indeed, multiple equilibria are the rule in FTC situations.

The Observable Case

The following model originated in Ref. 11, where the analysis given below appears in

Ref. 12. Suppose an arrival to the system observes the state (i, j) , such that $i \geq 0$ regular customers and $j \geq 0$ premium customers are present. He/she considers whether or not to purchase priority. The value of j does not effect his/her decision: they inflict on him/her a mean queueing time of j/μ , regardless of whether or not he/she purchases priority. Thus, without loss of generality assume that $j = 0$ and remove further reference to that value. It is clear that the larger the value of i , the larger the value of the priority. Priority is valuable not only in order to overtake regular customers, but also in order not to be overtaken by customers who would arrive in the future and would elect to purchase priority. Thus, one also bears in mind the purchasing strategies used by others, making this decision model a noncooperative game.

Suppose all use the pure threshold strategy of n , namely they purchase priority if and only if they see n or more regular customers upon arrival. Clearly, when all behave like that, n is the largest possible length of the regular queue. Let $W(n)$ be the mean queueing time (exclusive of service) ahead of a customer who is currently in position n in the regular customer queue where the premium queue is empty. Note that $W(n) + 1/\mu$ is his/her mean total time in the queue. Let B be the mean length of a busy period in an M/M/1 queue. Of course, $B = 1/(\mu - \lambda)$.

Theorem 1. *The pure threshold strategy n , $n \geq 1$, is an equilibrium if and only if*

$$\theta - CB \leq CW(n) \leq \theta. \quad (4)$$

¹⁰The interested reader is referred to Ref. 2, p.84, where it is shown that the equilibrium under p_e suffers from lack of stability (defined there) while the two pure equilibria are stable.

Proof. Two conditions need to be met in order for n to prescribe an equilibrium. First, an arrival facing $n - 1$ regular customers, not purchasing priority, and hence having to be in the system for time whose mean equals $W(n) + 1/\mu$ is better than purchasing priority and being there for a time whose mean is $1/\mu$. For that it is required that $CW(n) \leq \theta$. Second, an arrival facing n regular customers, purchasing priority and hence having to be in the system for time whose mean is $1/\mu$ is better than not purchasing priority and having to be in the system for a time whose mean is $B + W(n) + 1/\mu$. Note that the term B here is due to the fact that when all adopt the threshold strategy n , it requires a busy period to move from position $n + 1$ to position n in the regular queue. For this to be an optimal move, it is required that $\theta \leq C(B + W(n))$.

In Ref. 12 it is shown that there exists at least one value for n , which meets inequality 4. Equally important is that it is possible that there are a few consecutive values for such n , leading to multiple equilibria. As stated, this is typical for FTC situations as exemplified, already, in the unobservable version of this problem. Indeed, the larger the threshold used by others (namely, the less aggressive they are in terms of purchasing priority), the larger (i.e., less aggressive) should be the optimal threshold used by an individual, since the risk of being overtaken is now smaller. This intuitive argument is proved to be correct in Ref. 13. It is also shown there that if both n and $n + 1$ prescribe an equilibrium, then there exists another equilibrium, one which mixes between these two thresholds. Under such mixing the mean queueing time of the one in the worst possible position in the regular queue (under this policy) equals θ/C .¹¹ Finally, in Ref. 13, an efficient algorithm for computing the required mean queueing times is developed.

How Much to Pay for an Upgrade

We now look at a different scenario as suggested in Ref. 14. We assume an unobservable

¹¹Yet, as opposed to the pure equilibria, the mixed ones are not stable.

M/M/1 queue where customers are allowed to pay any amount they wish. The more they pay, the higher is their priority level, regardless of how much more they have paid or when they actually have arrived. We assume no preemption and hence the next to commence service among those in line is the one who paid the most. As in the previous section, the option of not joining does not exist.¹² Ties in payments are broken on an FCFS basis. Looking for a Nash equilibrium, it is clear that no pure one exists. Indeed, if all pay the same amount, we end up with an FCFS system. Yet, given this behavior, an individual better pay a bit more and overtake all others. By the same token, the mixed equilibrium should be without atoms. Again, had an atom existed, say at x , one who pays a little bit more than x , would overtake the quantum who pays x . Finally, the support of the values paid is a continuous interval. This is the case since had there been a gap, say in $[a, b]$ with values no one selects, it would have been better paying a than a bit more than b since it comes with a clear saving without actually losing any priority value. This contradicts the fact that all pure strategies in the support come with the same gain. In fact, the support should be $[0, u]$ for some value for u .

A customer who uniquely gets top priority queues up for time whose mean equals ρ/μ . On the other hand, the corresponding value for a customer who is uniquely most inferior equals $\rho/[\mu(1 - \rho)^2]$.¹³ Hence,

$$u' \equiv \frac{\rho C}{\mu(1 - \rho)^2} - \frac{\rho C}{\mu}$$

is an upper bound on the value of priority. Thus, $u \leq u'$. In fact, we next argue that $u = u'$: the one who pays zero (and is uniquely inferior to all) and the one who pays u (and is uniquely prior to all) share the same utility.

¹²For the case where this option exists and all customers share the same value for service R , see Ref. 15.

¹³Proof: A busy period comes with a mean value of $1/[\mu(1 - \rho)]$. Upon arrival, the expected number of customers in the system is $\rho/(1 - \rho)$. The product of these two values is the mean time of what this inferior customer has to spend in the queue.

Had $u < u'$, it would have been better for the former to pay a bit more than u rather than pay zero. Hence, $u = u'$.

The question that remains now is what is the distribution over payments on $[0, u]$ defining the equilibrium strategy. Its cumulative distribution function was derived in Ref. 14. Specifically, let $F(x)$ be the probability of paying up to x . Then, for any $x \in [0, u]$,

$$F(x) = \frac{1}{\rho} \left[\left(\frac{1}{(1-\rho)^2} - \frac{x\mu}{C\rho} \right)^{-\frac{1}{2}} - (1-\rho) \right].$$

In the above model it is assumed that if one pays more than another, regardless of how much more, one has priority over the other. An alternative model, with a relative priority flavor, is suggested in Ref. 16. Specifically, if upon service completion there are n customers in queue and customer i paid $x_i \geq 0$ toward priority, then customer i is the next to commence service with probability $x_i / \sum_{j=1}^n x_j$, $1 \leq i \leq n$. It is shown there that a unique equilibrium exists. Moreover, it is pure and all pay

$$\frac{\rho^2 C}{\mu(1-\rho)(2-\rho)}.$$

TO RENEGE OR NOT TO RENEGE

The phenomenon of customers who renege (abandon) the queue some time after joining, is common. The terminology of reneging hints to the possibility that customers “walked away from a commitment” or that they “change their mind”. Standard modeling does not allow for this possibility. Instead, we assume that customers join the queue while planning to leave it (abandon) later in case that something occurs or some relevant information is revealed to them.

Before giving further details on the model we pose, we make the following observation. In an unobservable M/M/1 FCFS queue, reneging is not due to learning while waiting that the queue length is in fact longer than anticipated. It was shown in Ref. 13, that irrespective of the reneging policy employed by the others, for an individual, the longer

is his/her past waiting time, the shorter is his/her expected time ahead. Thus, if one did not renege until now, it will be inconsistent from one’s side to do it now (or later). Thus, reneging needs to be attributed to increasing waiting costs per unit of time (which can be interpreted as loss of patience) or due to a deterioration in the value of service. In fact, these two factors are two sides of the same coin: the cost/reward ratio should go up while waiting in order to observe reneging.

We next describe two such situations in the unobservable M/M/1 FCFS queue. In the first, when the cost/reward is constant for some time and then it jumps abruptly so as to enforce immediate reneging. Interestingly, no reneging takes place ahead of this time. Yet, one may elect not to join at all. In the second, we let the cost/reward ratio go up smoothly, leading to a mixed equilibrium strategy regarding when to abandon the queue. Thus, reneging can take place at some continuous interval of time after waiting.

The Bang Bang Case

The model we deal with here is as our standard unobservable M/M/1 FCFS model with one key distinction: for anyone who has already stayed T units of time in the system without completing service (regardless of whether it had commenced or not), the value of service drops to zero. Assuming that customers recall the length of their past queueing time (i.e., each one of them holds a watch), a pure strategy here tells whether or not to join, and in the latter case, if and when to renege. As always, mixing is possible. In Ref. 13, it is shown that under any behavior of other, if a best response prescribes joining (not necessarily uniquely), then reneging after waiting for less than T units of time is never part of the best response. This is due to the fact that the probability of completing service in the next instant of time, goes up with the length of the past waiting time. Thus, in our quest for an equilibrium strategy we need to look for an equilibrium joining probability, while it is clear that one who joins reneges after T units of time if service is not completed by then.

In summary, the problem we face here is of finding the equilibrium joining probability.

Let $f(p)$ be the expected utility of one who joins and plans to leave after T units of time if one's service does not end by then, given that all others do the same with probability p (and do not join at all with probability $1 - p$). In Ref. 13, it is shown that

$$f(p) = \frac{R(\mu - \mu e^{-(\mu-\lambda p)T}) + \frac{\lambda p C T (\mu - \lambda p) e^{-(\mu-\lambda p)T} - C \mu (1 - e^{-(\mu-\lambda p)T})}{\mu - \lambda p}}{\mu - \lambda p e^{-(\mu-\lambda p)T}}.$$

Since $f(p)$ is monotone decreasing with p , if $f(1) \geq 0$, then joining (and then reneging at T) is a dominant strategy. Otherwise, there exists a unique value for p , $0 < p < 1$, with $f(p) = 0$. Denote this value by p_e . Then joining with probability p_e (and in case of joining, reneging after T units of time) is the unique equilibrium strategy. Again, this is a ATC situation.

Smooth Deterioration in Waiting Conditions

We now relax the assumption that the cost of waiting per unit of time is constant. So assume that $c(t)$ is the cost of waiting per unit of time for one who already waits t units of time. In other words, one who has already waited for t units of time suffered, so far, a loss of $\int_{\tau=0}^t c(\tau) d\tau$. The situation here is of course more complicated than the one described in the previous section. We describe one possibility as reported in Ref. 17.

Theorem 2. *Assume the following:*

1. $c(0) = 0$ and $c(t)$ is monotone increasing and concave. Also, $\lim_{t \rightarrow \infty} c(t) > R\mu$. Let T_2 be with $c(T_2) = \mu R$.¹⁴
2. $\beta(t) \equiv c(t)/R - c'(t)/c(t)$ is monotone increasing.
3. Let T_1 be with $\beta(T_1) = \mu - \lambda$ and assume that $T_1 < T_2$.

¹⁴It is clear that all renege not later than T_2 units of time after joining.

Define the cumulative distribution function $G(t)$ by¹⁵

$$G(t) = \begin{cases} 0, & 0 \leq t \leq T_1 \\ 1 - \frac{\mu - \beta(t)}{\lambda}, & T_1 \leq t < T_2 \\ 1, & t \geq T_2. \end{cases}$$

Then, the unique equilibrium strategy is to join and then to renege after time which follows the distribution defined by the function $G(t)$.

In Ref. 17, it is also shown that if a pure equilibrium reneging strategy exists (of course, under different assumptions than those postulated Theorem 2), it prescribes reneging at time T_2 . Sufficient conditions for this to be the case are stated there.

CONCLUDING REMARKS

A number of decision problems faced by customers in a memoryless single server queue were surveyed here as noncooperative games. This survey is by no means exhaustive. Other problems were dealt with in the literature. For example, in an M/M/1 when customers never learn the queue situation, they can try again and again, hoping to find the server available. These retrials comes with a cost. The question is then when to retry. The case where customers do not recall their past experience and do not even possess a watch, leading to identically, exponentially, and independently distributed intervals between consecutive retrials, was solved in Ref. 18. The cases of full or partial information regarding their own whereabouts (such as when they have retried last) are still open. Another question is when to arrive (given a finite period and a distribution on the number of arrivals during that period). See Ref. 19 for further details.

The assumption of stationary or steady-state conditions assumed throughout this survey, is removed in Ref. 20. Here customers know their time of arrival and need

¹⁵Note that since $c(T_2) = \mu R$, there is an atom of size $c'(T_2)/(R\lambda\mu)$ at T_2 .

to decide if to try to get service or to leave for good upon arrival to an M/M/N/N queue (when trials are costly and hence finding all servers busy after trying ends up with a loss). Assuming an empty queue at time $t = 0$, the equilibrium strategy prescribes trying if the time of arrival is smaller than some threshold value t_e . Afterwards, one tries with some time-dependent probability $p_e(t)$. Interestingly, $p_e(t_e+) < 1$. Treating similar models, for example the simplest M/M/1 to queue or not to queue model dealt with throughout this survey, is still open. Clearly it is a technically challenging problem.

For more problems dealing also with general service distributions, heterogeneous customers, and multiserver queues see Ref. 2. In particular, see Refs 21 and 22 for the observable version of the “to queue or not to queue” problem in the M/G/1 case.¹⁶

Finally, many models where *informational externalities* are involved can be designed. One example is dealt with in Ref. 23. There, a customer who joins one out of two queues sends, just by his/her own action, a message to future arrivals that this is the better queue to join according to his/her belief. Hence, it sometimes makes sense to join the longer queue. Another example appears in Ref. 24. Here, there is only one server whose utility is not known. Customers get private signals and after inspecting the queue length, decide if they prefer joining over balking. Here too, those who join affect the posterior on the value of service assessed by future arrivals. Again, longer queues can sometimes be more appealing than short ones.

REFERENCES

1. Naor P. The regulation of queue size by levying tolls. *Econometrica* 1969;37:15–24.
2. Hassin R, Haviv M. To queue or not queue: equilibrium behavior in queueing systems, Boston (MA): Kluwer's International Series; 2003. Available at <http://www.math.tau.ac.il/~hassin/book.html>.
3. Fudenberg D, Tirole J. Game theory. Cambridge (MA): The MIT Press; 1994.
4. Osborne M, Rubinstein A. A course in game theory. Cambridge (MA): The MIT Press; 1994.
5. Edelson NM, Hildebrand K. Congestion tolls for Poisson queueing processes. *Econometrica* 1975;43:81–92.
6. Hassin R, Haviv M. Nash equilibrium and subgame perfection. *Ann Oper Res* 2002;113:15–26.
7. Chen F, Frank M. State dependent pricing with a queue. *IIE Trans* 2001;33:847–860.
8. Hassin R. On the optimality of first come last served queues. *Econometrica* 1985;53:201–202.
9. Assaf D, Haviv M. Reneging from time sharing and random queues. *Math Oper Res* 1990;15:129–138.
10. Altman E, Shimkin N. Individual equilibrium and learning in processor sharing system. *Oper Res* 1998;46:776–784.
11. Adiri I, Yechiali U. Optimal priority purchasing and pricing decisions in non-monopoly and monopoly queues. *Oper Res* 1974;22:1051–1066.
12. Hassin R, Haviv M. Equilibrium threshold strategies: the case of queues with priorities. *Oper Res* 1997;45:966–973.
13. Hassin R, Haviv M. Equilibrium strategies for queues with impatient customers. *Oper Res Lett* 1995;17:41–45.
14. Glazer A, Hassin R. Stable priority purchasing in queues. *Oper Res Lett* 1986;4:285–288.
15. Lui FT. An equilibrium queueing model of bribery. *J Polit Econ* 1985;93:760–781.
16. Haviv M, van der Wal J. Equilibrium strategies for processor sharing and queues with relative priorities. *Probab Eng Inf Sci* 1997;11:403–412.
17. Haviv M, Ritov Y. Homogeneous customers renege from invisible queues after random times under deteriorating waiting conditions. *Queueing Syst Theory Appl* 2001;38:495–508.
18. Hassin R, Haviv M. Optimal and equilibrium retrial rates in a busy system. *Probab Eng Inform Sci* 1996;10:223–227.
19. Glazer A, Hassin R. M/M/1: On the equilibrium distribution of customer arrival. *Eur J Oper Res* 1983;13:146–150.
20. Haviv M, Kella O, Kerner Y. Equilibrium strategies based on time or index of arrival. *Probab Eng Inform Sci* 2010;24:13–25.

¹⁶The unobservable version is a straightforward generalization of what was presented here in section titled “The Unobservable Case”.

21. Haviv M, Kerner Y. On balking from an empty queue. Queueing Syst Theory Appl 2007;55:239–249.
22. Kerner Y. Equilibrium joining probabilities for an M/G/1 queue Games and Economic Behaviour. EURANDOM Working Paper 2009-020. 2009.
23. Veeraraghavan S, Debo LG. Joining longer queues: informational externalities in queue choice. Manuf Serv Oper Manage 2009;11:542–562.
24. Debo LG, Parlour CA, Rajan U. Inferring quality from a queue, Chicago Booth School of Business Working Paper 09-26 2008. Available at http://faculty.chicagobooth.edu/laurens/Research/Papers/debo_herding_single_queue/pdf.

STRUCTURAL RESULTS FOR POMDPs

LUDMILA V. ZHELTOVA
Department of Operations
and Supply Chain
Management,
Cleveland State
University, Cleveland,
Ohio

The partially observable Markov decision process (POMDP) is a powerful mathematical model for decision making under uncertainty. A wide range of problems such as inventory control, machine maintenance, quality control, optimal search efforts, and medical decision making could be modeled as POMDP.

POMDP is a generalization of a Markov decision process (MDP). In MDP, the state of the system is known with certainty. However, in POMDP, the underlying state cannot be directly observed. Instead, a belief state, probability distribution over the set of possible states, based on a set of observations and observation probabilities, is used to track the system state.

Owing to its complexity, a primary goal of POMDP analysis is to determine conditions under which the optimal policy has a simple structure. Examples of such policies with simple structures for POMDPs include (i, I) maintenance policy [1], control limit policy (CLP) in maintenance [2–6] and health care [7], and monotone, at-most-four region (AM4R) policy in maintenance [8–13].

A partial order of belief states is necessary in establishing the optimal policy structure. In this article, we discuss two types of partial orders: stochastic and likelihood ratio. The likelihood ratio order of belief states is preserved after conditioning on any information and along with other conditions provides monotonicity of POMDP's optimal value function and policy.

The rest of the article is organized as follows: In the section titled “Model Description”, we formulate general POMDP model.

The section titled “Structured Policies” introduces structural policies and discusses the sufficient conditions for monotonicity of the value function and optimal policy. The section titled “Structured Policies in Maintenance Optimization” discusses in detail the structured policies in the context of maintenance optimization problems when three actions replace, operate, inspect are available.

Lastly, in the section titled “Conclusion”, further reading is outlined.

MODEL DESCRIPTION

Formally, a discrete-time POMDP is defined as a tuple (S, O, A, q, T, R) , where $S = \{0, 1, \dots, N\}$ is a finite set of core states with natural order; $O = \{1, \dots, M\}$ is a finite set of observations; A is a finite set of actions; q is an observation model that specifies the probability of making an observation k in a state $i \in S$; T is a state-transition model that specifies a probability distribution over the resulting state j , given an initial state i and action a ; C is a cost function that maps each underlying state-action pair into an immediate expected cost.

At each decision epoch t , the system is in some state $i \in S$. Although the state i is not directly observable by the decision maker, the probability that the system is in a given state can be calculated. Let π be a vector of state probabilities called a *belief* or *information state*. The belief state π has components $[\pi_0, \pi_1, \dots, \pi_N]$, where π_i is the probability that the system is in the core state i . These belief states form the belief state space $\Omega = \{\pi : 0 \leq \pi_i \leq 1, \sum_{i=0}^N \pi_i = 1\}$.

At each decision epoch t , the decision maker chooses an action $a \in A$ for which the system incurs the immediate expected cost $C(i, a)$ and makes a transition to a new unobservable state $j \in S$ according to a transition probability matrix $p^a = [p_{ij}^a]$, $i, j = 0, 1, \dots, N$. The decision maker receives the observation message k with probability $q_{ij}^a = \Pr\{k | j, a\}$ and updates the new belief state as $T_j(\pi | a, k) \equiv \pi_j(t+1) = \frac{q_{jk}^a(\pi P^a)_j}{\sigma(k | \pi, a)}$, where

probabilities q_{jk} are elements of observation matrix Q and $\sigma(k|\pi, a) = \sum_{i=0}^N q_{ik}^a (\pi P^a)_i$, with $i, j = 0, 1, \dots, N$ and $k = 1, 2, \dots, M$.

To solve POMDP means finding a rule for selecting actions, that is, a policy. Policy optimizes an objective of the process performance. For some problems, as we consider in this article, such an objective is to minimize the total expected discounted cost over an infinite horizon, $V(\pi)$:

$$V(\pi) = \min_{a \in A} \left\{ \sum_{i=0}^{i=N} \pi_i C(i, a) + \alpha \sum_{k=0}^M \sigma(k|\pi, a) V[T(\pi|a, k)] \right\}, \quad (1)$$

where $\alpha \in (0, 1)$ is a discount factor. Let $h(\pi, a) = \sum_{i=0}^{i=N} \pi_i C(i, a) + \alpha \sum_{k=0}^M \sigma(k|\pi, a) V[T(\pi|a, k)]$.

STRUCTURED POLICIES

The policy for a POMDP specifies the actions for all possible belief states. However, solving POMDP, that is, computing the optimal policy, is considered a very hard problem due to continuity of the belief state space (see ‘‘Reduction of a POMDP to an MDP’’). More precisely, a POMDP is considered to be PSPACE-complete, meaning that it ‘‘can be computed using polynomial

space.’’ Numerous exact and approximation algorithms [14–26] exist to solve POMDP, but such an approach to policy computation could have limitations in terms of policy quality and computational efficiency.

Notably, classes of problems with exploitable optimal policy structures can considerably reduce the complexity of POMDP solutions and time required to calculate the optimal policy for particular instance of the problem. *A priori* knowledge of policy structure allows its fast computation and the simplicity of the policy structure provides additional benefits of effective communication and implementation.

It is known that the optimal value function for finite horizon POMDP is piecewise linear and concave function of belief state [24]. The optimal policy for POMDP with three core states and two available actions a_1 and a_2 could be presented graphically as shown in Fig. 1a. Such a policy does not have any structure compare to one that is pictured in Fig. 1b. The policy in Fig. 1b is referred as a *control limit* or *threshold policy*. Since system has three deterioration states, the belief state space Ω is represented by an equilateral triangle, with each point in the triangle corresponding to a belief state. The straight lines presented in both Fig. 1a and b separate sets of belief states where actions a_1 or a_2 are optimal.

Speaking about CLP for MDP, it is defined by the existence of the state i^* such that

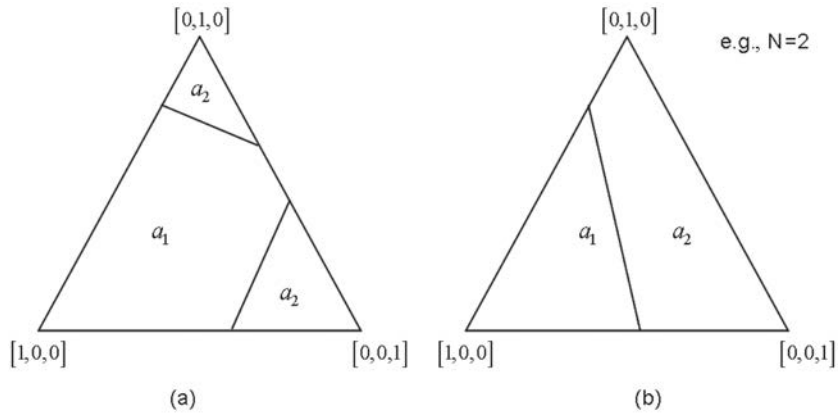


Figure 1. Nonstructured and structured POMDP policies for the system with two core states.

action a_1 is optimal in state i if $i \leq i^*$, and action a_2 is optimal for any $i > i^*$, where $i, i^* = 0, 1, \dots, N$. Note that for MDP the system's state is perfectly observable.

In case of POMDP, since the core state is not observable, CLP is defined by specifying the belief state π^* such that for the system in any belief state $\pi \in \Omega$, action a_1 is optimal if π is “smaller” than π^* and action a_2 is optimal if π is “larger” than π^* .

While for MDP, the state order is straightforward, for POMDP belief state is a probability vector and to specify when one belief state is “smaller” or “larger” than the other requires special approach. Particularly, two partial orders of belief states—stochastic and likelihood ratio—are used to define which belief state is “smaller” or “larger” than the other. These partial orders play an important role in establishing the structural results for POMDPs. We define that the belief state π^1 to be stochastically smaller than belief state π^2 , that is, $\pi^1 \prec_{st} \pi^2$, if and only if $\sum_{i=m}^N \pi_i^1 \leq \sum_{i=m}^N \pi_i^2$ for all $m = 0, 1, 2, \dots, N$. We define that the belief state π^1 to be likelihood ratio smaller than the belief state π^2 , that is, $\pi^1 \prec_{lr} \pi^2$, if and only if $\pi_i^1 \pi_j^2 - \pi_j^1 \pi_i^2 \geq 0$ for all $i \leq j$ and $i, j = 0, 1, \dots, N$. It is well known that for $N = 1$, these two partial orders are equivalent and, for $N > 1$, likelihood ratio order implies stochastic order ([27]).

Although, some results on policy structure for POMDPs are achieved by considering stochastically ordered belief states (for example, in Refs 9,13, Lovejoy in Ref. 27 argues that only likelihood ratio order could provide general results for the optimal policy and value function monotonicity because of the fact that likelihood ratio order of belief states is preserved under Bayesian update. That is, for $\pi^1 \prec_{lr} \pi^2$, $T(\pi^1 | a, k) \prec_{lr} T(\pi^2 | a, k)$, if and only if $\pi^1 P^a \prec_{lr} \pi^2 P^a$, for any $a \in A$ and $k \in O$. Likelihood ratio order “survives” the action and observation updates as well, that is, $T(\pi^1 | a, k)$ is likelihood ratio order nondecreasing in $a \in A$ and $k \in O$.

The steps required to establish the optimal policy monotonicity are articulated in Ref. 27. First, the conditions under which the optimal value function is monotone in likelihood ratio ordered belief states must

be found [4]. These conditions include a nondecreasing in core state of the system cost structure, total positivity of order 2 of transition probability and observation matrices. When matrix is totally positive of order 2 (TP_2), if all of its 2×2 determinants are nonnegative. However, the weaker than TP_2 conditions that could be imposed on observation matrices still provide the optimal value function monotonicity along with previously mentioned conditions are possible [27]. The conditions that provide the optimal policy monotonicity on any partial order are established [27] and supermodularity of function $h(\pi, a)$ are included. That is, the optimal policy is monotone along partially ordered sets of belief states, if $h(\pi, a)$ has isotone differences on $\Omega \times A$, that is, $h(\pi^1, a) - h(\pi^1, a') \leq h(\pi, a) - h(\pi, a')$ for $\pi \prec_{lr} \pi^1$ and $a' \leq a$. While details and conditions that provide super/submodularity of the function are omitted [4,28], we would like to note that it arises in many applications and has an interesting property to be preserved under some useful operations such as maximization or minimization.

Finally, sufficient conditions that provide convexity of policy regions and, hence, the monotonicity of the optimal policy are established and discussed in Ref. 29 in detail.

STRUCTURED POLICIES IN MAINTENANCE OPTIMIZATION

In this section, we discuss the conditions that guarantess optimal policy structure in the context of maintenance optimization problem. Consider a production system with a finite set of core states $S = \{0, 1, \dots, N\}$ with state 0 being the “good as new” state and state N being the most deteriorated state for which at each decision epoch three actions are available: replacing the system with a new one, operating the system without inspection, operating with perfect inspection. Further on, we refer to these actions as replacing, operating, and inspecting, respectively. The core state of the system is not observable unless the decision maker is willing to pay for cost l_i , perfect inspection.

Each action takes one time period with C_i , L_i , and $L_i + M$ being per-period replacing,

operating, and inspecting costs, respectively, that the system accrues in core state i , $i = 0, 1, \dots, N$. Then Equation (1) could be specified as follows:

$$V(\pi) = \min \left\{ \begin{array}{l} \pi C + \alpha V(e_0), \\ \pi L + \alpha V(\pi P), \\ \pi L + M + \alpha \pi P V(e) \end{array} \right\}, \quad (2)$$

where $V(e) \equiv [V(e_0), V(e_1), \dots, V(e_N)]^T$ and e_i is a vector with 1 in i th place and 0 everywhere else.

For problem described by Equation (2), the optimal policy structure has been examined in Refs 9, 11–13.

In Ref. 12, Ross considers the system with two core states 0 (“good as new”) and 1 (the worst state). Then the belief state of the process is a probability distribution $\pi \equiv [1 - \pi_1, \pi_1]$, where π_1 is a probability of being in the worst state. He assumes that the cost of operation increases with the deterioration level, while the costs of replacement and inspection are independent of the core state. To establish a policy structure, he considers the cost of operation being less than the cost of inspection and less than the cost of replacement and the transition probability matrix P being upper-triangular. These conditions allow to define the optimal policy as monotone, AM4R policy, that is, there exist three numbers p_1 , p_2 , and p_3 , where $0 \leq p_1 \leq p_2 \leq p_3 \leq 1$, such that it is optimal to replace the system if $p_3 \leq \pi_1$, to inspect the system if $p_1 \leq \pi_1 < p_2$, and to continue to operate otherwise. The numerical example illustrating this policy structure could be found in Ref. 12. Such a policy could be graphically presented as four disjoint regions (Fig. 2). The existence of the second operate region is somewhat counterintuitive, but could be explained by the reasoning that in the current state, it might not be worth inspecting the system that is about to be replaced.

In Ref. 13, White extends the results of Ref. 12 for an $N + 1$ state system such that the more that the system deteriorates, the more likely it is to deteriorate further. White establishes conditions under which the optimal policy has monotone, AM4R structure along the straight line of stochastically increasing belief states. These conditions include nondecreasing in core state-action costs with operating costs increasing faster than replacement costs.

Next, conditions that provide monotone, AM4R policy structure along sample path are established in Refs 9 and 11. We define the sample path as a sequence of belief states that the process would occupy over time if the system operates continuously and no actions are taken. For example, the sample path emanating from belief state π is $\{\pi, \pi P, \pi P^2, \dots\}$.

Assuming that transition probability matrix P is upper-triangular and TP_2 , and the same cost structure as in Ref. 13, Rosenfield establishes the monotone, AM4R structure along sample path that emanates from state e_i , $i = 0, 1, \dots, N$.

The assumptions about transition probability matrix P are of a particular interest. Upper triangularity of the transition probability matrix implies that the system could not improve on its own and TP_2 ness of this matrix implies that the more the system deteriorates, the more likely it is to deteriorate further. Note that these conditions imposed on matrix P yield likelihood ratio order increasing sample paths, that is, $e_i \prec_{lr} e_i P \prec_{lr} \dots \prec_{lr} e_i P^k$.

Finally, the results providing monotone, AM4R optimal policy structure, received in Ref. 11, are extended in Ref. 9 for the certain time-based sample paths.

If we consider the sample path emanating from any belief state π , then in order for this path to be in increasing likelihood ratio order,

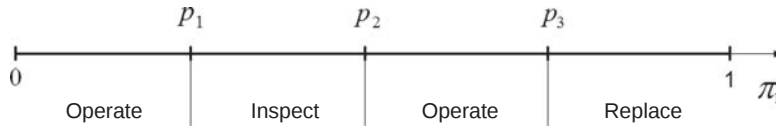


Figure 2. Monotone, AM4R policy for the system with two core states.

this path, first, should start in increasing likelihood ratio order and, second, stay in this order. While the latter is provided by the TP_2 ness of transition probability matrix P , the former requires state π to belong to a special region defined as $\Omega_{\ell r} \equiv \{\pi \in \Omega : \pi \prec_{\ell r} \pi P\}$.

Although one could argue that the system mostly starts from the known state or returns to a known state after the replacement or inspection, it is still worthwhile to establish policy structure along nonextreme sample paths in order to gain insights that might be useful in establishing similar results for systems with imperfect information.

Furthermore, interesting results concerning monotone, AM4R policy structure for maintenance optimization problems could be found in Refs 1,8,10, and 30.

CONCLUSION

This article introduces the structural results for POMDP and discusses the conditions that provide such results for general POMDP model.

In this article, we discuss in detail the monotone, AM4R policy structure and conditions that guarantee such a structure for a particular maintenance optimization problem. However, CLPs are subject of numerous papers in different areas of research. Interested reader could find the results concerning CLPs in Refs 3,31–33. Though, we consider only stochastic and likelihood ratio partial orders, some papers consider different partial orders. For example, in Ref. 34, results concerning policy structure are achieved by considering Blackwell partial order.

Interested reader could find further details about maintenance policies in Refs 35–38. Also, Monahan in Ref. 19 provides an excellent overview of POMDP models including some structural results.

REFERENCES

- Ohnishi M, Moroika T, Ibaraki T. Optimal minimal-repair and replacement problem of discrete-time Markovian deterioration system under incomplete state information. *Comput Ind Eng* 1994;27:409–412.
- Albright SC. Structural results for partially observable Markov decision processes. *Oper Res* 1979;27(5):1041–1053.
- Grosfeld-Nir A. Control limits for two-state partially observable Markov decision processes. *Eur J Oper Res* 2007;182:300–304.
- Lovejoy W. Ordered solutions for dynamic programs. *Math Oper Res* 1987b;12:269–276.
- Rosenfield D. Markovian deterioration under uncertain information — a more general model. *Nav Res Logist Q* 1976;23:389–405.
- White C. Optimal control-limit strategies for a partially observed replacement problem. *Int J Syst Sci* 1979;10:321–331.
- Alagoz O, Maillart LM, Schaefer AJ, Roberts MS. The optimal timing of living-donor liver transplantation. *Manag Sci* 2004;50(10):1420–1430.
- Maillart LM. Maintenance policies for systems with condition-monitoring and obvious failures. *IEE Trans* 2006;38:463–475.
- Maillart LM, Zheltova LV. Structured maintenance policies on interior sample paths. *Nav Res Logist* 2007;54(6):645–655.
- Ohnishi M, Kawai H, Mine H. An optimal inspection and replacement policy under incomplete state information. *Eur J Oper Res* 1986;27:117–128.
- Rosenfield D. Markovian deterioration with uncertain information. *Oper Res* 1976;24(1):141–155.
- Ross SM. Quality control under Markovian deterioration. *Manag Sci* 1971;19(9):587–596.
- White III CC. Optimal inspection and repair of a production process subject to deterioration. *J Oper Res Soc* 1978;29(3):235–243.
- Cheng H-T. Algorithms for partially observable Markov decision processes. PhD thesis, University of British Columbia, 1988.
- Hansen EA. Finite-memory control of partially observable systems. PhD thesis, University of Massachusetts, Amherst, 1998.
- Littman ML, Cassandra AR, Kaelbling LP. Efficient dynamic programming updates in partially observable Markov decision processes. Technical Report CS-95-19, Brown University Providence, RI, 1996.
- Lovejoy W. Computationally feasible bounds for partially observed Markov decision processes. *Oper Res* 1991a;39(1):162–175.
- Lovejoy W. A survey of algorithmic methods for partially observed Markov decision processes. *Ann Oper Res* 1991b;28:47–66.

19. Monahan GE. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Manag Sci* 1982;28(1): 1–16.
20. Ng AY, Jordan M. PEGASUS: A policy search method for large MDPs and POMDPs. *Proceedings of Uncertainty in Artificial Intelligence*; Stanford, CA, 2000.
21. Pineau J, Gordon G, Thrun S. Point-based value iteration: An anytime algorithm for POMDPs. *Proceedings of the International Joint Conference on Artificial Intelligence*; Acapulco, Mexico, 2003.
22. Poupart P, Boutilier C. Bounded finite state controllers. *Advances in neural information processing systems*. Vancouver: MIT Press, 2004.
23. Roy N, Gordon G, Thrun S. Finding approximate POMDP solutions through belief compression. *J Artif Intell Res* 2005;23:1–40.
24. Smallwood RD, Sondik EJ. The optimal control of partially observable Markov processes over a finite horizon. *Oper Res* 1973;21:1071–1088.
25. Zhang W. Algorithms for partially observable Markov decision processes. PhD thesis, The Hong Kong University of Science and Technology, 2001.
26. Zhou R, Hansen EA. An improved grid-based approximation algorithm for POMDPs. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*; Seattle, 2001.
27. Lovejoy WS. Some monotonicity results for partially observed Markov decision processes. *Oper Res* 1987a;35(5):736–742.
28. Topkis D. Minimizing a Submodular Function on a Lattice. *Oper Res* 1978;26:305–321.
29. Lovejoy W. On the convexity of policy regions in partially observed systems. *Oper Res* 1987c;35(4):619–621.
30. Jin L, Mashita T, Suzuki K. An optimal policy for partially observable Markov decision processes with non-independent monitors. *J Qual Mainten Eng* 2005;11(3):228–238.
31. Ghasemi A, Yacout S, Ouali MS. Optimal condition based maintenance with imperfect information and the proportional hazards model. *Int J Prod Res* 2007;45(4): 989–1012.
32. Krishnamurthy V, Djonin D. Structured threshold policies for dynamic sensor scheduling—a partially observed Markov decision process approach. *IEEE Trans Signal Proc* 2007;55(10):4938–4957.
33. Makis V, Jardine AKS. Optimal replacement in the proportional hazards model. *INFORM* 1992;30(1):172–183.
34. Rieder U. Structural results for partially observed control models. *Meth Model Oper Res* 1991;35:473–490.
35. Dekker R. Applications of maintenance optimization models: a review and analysis. *Reliab Eng Syst Saf* 1996;51:229–240.
36. Pierskalla WP, Voelker JA. A survey of maintenance models: the control and surveillance of deteriorating systems. *Nav Res Logist Q* 1976;23:353–388.
37. Valdez-Flores C, Feldman RM. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Nav Res Logist* 1989;36:419–446.
38. Simmons Ivy J, Pollock S. Marginally monotonic maintenance policies for a multi-state deteriorating machine with probabilistic monitoring, and silent failures. *IEEE Trans Reliab* 2005;54(3):489–497.

FURTHER READING

- Chhatwal J, Alagoz O, Burnside ES. Optimal policies for biopsy decision-making based on mammography findings and demographic factors. *Oper Res* 2010;58(6):1577–1591.
- Derman C. On optimal replacement rules when changes of state are Markovian. In: Bellman R, editor. *Mathematical optimization techniques R-396-PR*. Berkeley: University of California Press; 1963. pp. 201–210.
- So Kut C. Optimality of control limit policies in replacement models. *Nav Res Logist* 1992;39(5):599–739.
- Puterman ML. *Markov decision processes*. New York: John Wiley and Sons; 1994.
- Satia JK, Lave RE. Markovian decision processes with probabilistic observation of states. *Manag Sci* 1973;20(1):1–13.

STRUCTURED OPTIMAL POLICIES FOR MARKOV DECISION PROCESSES: LATTICE PROGRAMMING TECHNIQUES

ANDREEA DRAGUT
Laboratoire d'Informatique Fondamentale,
Universite Aix-Marseille II (IUT d'Aix en
Provence), France

MONOTONIC OPTIMAL POLICIES

The structure of optimal policies and/or of the value function can be used to reduce the original Markov decision process (MDP) to an optimization problem over a smaller subset of policies, namely the *structured ones*. Monotonicity of optimal policies is a well-known and useful structural property (Chapter 8 in [1], p. 368; Section 4.7.3 in [2], p. 105). It reduces the exponential growth in the dynamic programming (DP) formulation with respect to the size of an MDP, when looking for the optimal policies. It also often allows for feasible numerical solutions ([3], Section 7.9 in [4], p. 210).

Recently, the structure of the value function has been used to compare different control systems (Chapter 11 in [5], p. 61).

This article focuses on monotonic optimal policies for maximization of MDPs (see **MDP Basic Components**). It is mainly based on the monotonicity research of [6–8]: set using DP and with monotonicity entailed by the objective function shown to be supermodular/submodular. Thus, the DP algorithm proves to be useful for the theoretical investigation of the structure of the control problems. For other general approaches see [1,9–12].

The monotonicity of an optimal policy reflects the qualitative properties of the problems studied. For monotonicity we need partially ordered state and action spaces. Partial orderings are important in many fields, and have spread out from mathematics

to physics, biology, and economics. In physics, partial ordering stands beside spatial symmetries and Hamiltonian structures for dynamical systems studies; these being properties that induce specific behaviors. As pointed out in [13], partial orderings significantly restrain the behavior, yet allow for interesting (and even chaotic) trajectories.

Partial orderings affect the chaotic behavior. They avoid it for most initial conditions or they make it unstable when physically observing the system in the long run. This type of robustness is also discussed for monotonic optimal policies. Partial orderings doubled by the lattice programming techniques of [8] encompass many applications in admission and control of queues, maintenance, production planning models.

Widely known optimal monotonic behavior paths are the threshold policies. Threshold policies (on–off controls) for dynamical systems are strategies that switch the control inputs from one level to another whenever a certain measured variable crosses a predetermined single threshold value. A control limit policy or simple threshold is defined by the existence of a state x_0 such that if the observed state is $x \geq x_0$, then it is optimal to make a change, otherwise it is optimal to continue the current trend. The proof of their optimality was very often made by an induction based on value iteration (see **Total Expected Discounted Reward MDPs: Value Iteration Algorithm**), which establishes a submodular/supermodular type of property for the optimal value function. They were identified early in manufacturing and admission control of queues problems.

Example 1. Simple threshold policy in a two-server queueing system (from [14,15])

Consider a queue with Poisson arrivals (rate λ) and two exponential servers, with asymmetric service rates μ_1 and μ_2 (with $\mu_1 > \mu_2$). The customers in the queue have to be assigned non-preemptively to the servers, that is, a customer sent to the slower server cannot be sent back to the queue. A controller

decides after each event/jump (which can either be an arrival or a service completion) if it sends a customer from the queue to a free server (if available).

Lin and Kumar [14] and Koole [15] studied this model and showed that to minimize the number of customers in the system, the fast server should always be used (as long as there are customers in the queue), and that the slower server should only be used if the number of customers in queue exceeds a certain level. Thus, monotonicity means that if we move a customer to a slower server in the state $(x + 1, 0)$, then we are going to similarly move one in any state $(x + n, 0)$ with $n \geq 1$.

The state space considered was of the form (x, i) where x is the total number of customers in the queue and at the first server, and $i \in \{0, 1\}$ denotes the number of customers at the second server. In [14], $\rho(x, i) = x + i$ is taken as a direct cost, while [15] used a more general class of cost functions, including $\rho(x, i)$.

Assuming, by scaling, that $\lambda + \mu_1 + \mu_2 = 1$, in [15] the continuous-time model was reduced to a discrete one by uniformization (see **Uniformization in Markov Decision Processes**), looking just at the event/jump times. For the moments when one of the servers idles, a fictitious transition was added to make the times between each two jump times identically exponentially distributed. $u_n(x, i)$, the expected cost over the next n jumps, in the embedded discrete-time model, starting in state (x, i) , was defined as:

$$u_{n+1}(x, i) = \rho(x, i) + \lambda w_n(x + 1, i) + \mu_1 w_n((x - 1)^+, i) + \mu_2 w_n(x, 0); \quad i = 0, 1, \quad \text{where} \\ w_n(x, 0) = \min\{u_n(x, 0), u_n(x - 1, 1)\} \quad \text{if } x > 0, \\ w_n(0, i) = u_n(0, i), \quad w_n(x, 1) = u_n(x, 1), \quad u_0(x, i) = \rho(x, i).$$

The optimality of the threshold policy holds if $u_n(x + 1, 0) - u_n(x, 1) \leq u_n(x + 2, 0) - u_n(x + 1, 1)$. Indeed, if $u_n(x + 1, 0) - u_n(x, 1) \leq 0$, then it is optimal not to use the second server, and vice versa. Thus, as shown in [15], if $u_n(x + 1, 0) - u_n(x, 1)$ is increasing in x , a threshold policy is optimal.

This requirement also eliminated the theoretical possibility of moving a customer from server 1 to 2 (from the definition of $w_n(1, 0)$).

By taking that $\rho(x, i) = g(x + i)$, with g increasing and convex, both the supermodularity of u_n and the desired monotonic increase were obtained.

Other types of threshold policies are the start up policies, N-policies, D-policies, T-policies (see Chapter 11 in [1], p. 444) for the single removable server and known workload with a $M/G/1$ queue or $M^X/G/1$ queue with batch arrivals. Under the N-policy, the server idles each time the system becomes empty, that is, there is no customer in the system and the server activates itself when there are N customers in the system. The D-policy activates the server when the cumulative service times of the customers in the queue exceeds some fixed value D . Under the T-policy the server can take a vacation time T after the system becomes empty. If there are no customers at the end of the vacation, the server takes another vacation. If customers are found, the server is activated (turned on) and works until the system is empty. See [16]; Chapter 11 in [1], p. 448 for a detailed discussion.

In general under a monotonic optimal policy, the decision increases with the state of the system. This means that the greater the rank of a state is (in the state space order), the higher its impact will be. The classical examples of simple monotonic policies are optimal pricing of a single product based on sales, optimal advertising of a single product based on sales, and optimal maintenance with overhaul options of one unreliable machine. For details see Chapter 3 in [8], p. 167 and Chapter 4 in [17], p. 108. Under a series of conditions, which will be discussed in this article, the monotonic policies can be described as: “As sales decreases in the current period, it is optimal to advertise more in the next period.”, “Optimal prices in each period increase with the sales in the preceding period.”, “More radical overhaul operations are taken when the machine is in a less desirable state.”

The rest of this article is structured as follows. The second section presents generalities regarding lattices, as well as some of Topkis’s results concerning maximization of supermodular functions on lattices.

The third section presents monotonicity results for discrete-time MDP models with finite and infinite horizons, as well as for partially observed Markov decision processes (POMDPs). The fourth section gives comparable statistics for the optimal monotonic nondecreasing policies. The fifth section presents monotonicity results for continuous-time models. The sixth section succinctly presents the relationship between supermodularity and convexity, as well as ways of obtaining other types of structured policies. Finally, the seventh section presents some interesting classes of monotone policies encountered in applications.

LATTICE PROGRAMMING TECHNIQUES PRELIMINARIES

Definition 1. A set X endowed with a reflexive, antisymmetric, and transitive binary relation “ $\leq_X$ ” on X is called a partially ordered set (poset).

Definition 2. Let $(X, \leq_X)$ be a finite poset. An antichain (respectively, chain) in X is a set of pairwise incomparable (respectively, comparable) elements. The size of the longest antichain (respectively, chain) is called the partial order width (respectively, length). A chain C is called complete in X if for any $x, y \in C$ such that $x \leq_X y$, $\nexists z \in X \setminus C$ such that $x <_X z <_X y$.

Lemma 1 [Dilworth’s Lemma and its Dual [18,19]]. *The partial order width of a finite poset $(X, \leq_X)$ is equal to the minimum number of chains needed to cover X . The partial order length of a finite poset $(X, \leq_X)$ is equal to the minimum number of antichains needed to cover X . If N be the cardinality of X , W the partial order width, and L the partial order length then $N \leq LW$.*

In other words, Dilworth’s Lemma states that the nonexistence of an antichain of size $n + 1$ in $(X, \leq_X)$ is a necessary and sufficient condition for $(X, \leq_X)$ to be the union of n total orders. This union provides a visualization

of a finite poset, which allows for a lossless enumeration of its elements.

Example 2. Let $(X, \leq_X)$ be the power set of a finite set $X = \{1, 2, 3\}$ ordered by inclusion. Then the poset width is 3. The antichain description is $\{\emptyset\}, \{\{1\}, \{2\}, \{3\}\}, \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}, \{\{1, 2, 3\}\}$, while the chain description is $\{\emptyset, \{1\}, \{1, 2\}, \{1, 2, 3\}\}, \{\emptyset, \{2\}, \{2, 3\}\}, \{\emptyset, \{3\}, \{1, 3\}\}$.

Definition 3. Let K a subset of a poset $(X, \leq_X)$ and $x \in X$. We say that x is an upper bound (respectively, lower bound) of K , denoted by $x \geq_X K$ (respectively, $x \leq_X K$), if for all $y \in K$, we have $x \geq_X y$ (respectively, $x \leq_X y$).

Definition 4. A lattice is a partially ordered set $(X, \leq_X)$ where for any pair of elements there is a supremum (least upper bound) and infimum (greatest lower bound) (belonging to X). Let $Y \subseteq X$ a subset in the lattice. Then Y is a complete sublattice of X if for any $a, b \in Y$, both their maximum (denoted by $a \vee b$), and their minimum (denoted by $a \wedge b$) taken in X are elements of Y .

Definition 5. In the vector-lattice $(X_1 \times \dots \times X_N, \leq_{X_1}, \dots, \leq_{X_N})$ we say that $x, y \in X_1 \times \dots \times X_N$ are in the natural order relation “ $x \leq y$ ” if componentwise $x_i \leq_{X_i} y_i, \forall i \in \{1, \dots, N\}$.

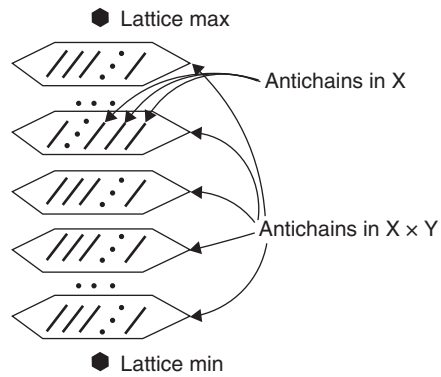


Figure 1. Representation of a vector-lattice $X \times Y$ as a set of antichains—layers, each of them in turn is also represented as a union of antichains according to $\leq_X$.

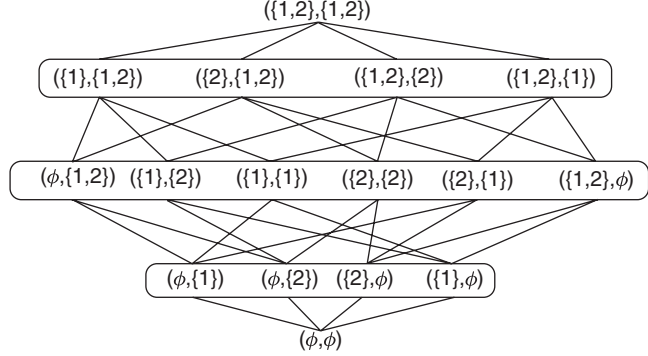


Figure 2. Representation of the vector-lattice obtained from the product with itself of the powerset of $\{1, 2\}$ ordered by inclusion.

Using Dilworth's Decomposition Lemma (i.e., Lemma 1) a vector-lattice $X \times Y$ can first be $\leq_{X \times Y}$ -decomposed and then any antichain can be further $\leq_X$ -decomposed—see Fig. 1. We notice that a complete chain in $X \times Y$ contains an element from each horizontal layer of the graphic decomposition of the lattice $X \times Y$.

Example 3. For the product with itself of the powerset of $\{1, 2\}$ ordered by inclusion we have Fig. 2.

Definition 6. Let $(X \times Y, \leq_{X \times Y})$ be a vector-lattice and $Z \subseteq X \times Y$. We say that Z is a subset without holes if for any $(x, y), (\tilde{x}, \tilde{y}) \in Z$, Z includes all the minimal length complete chains containing elements $(a, b) \in X \times Y$ such that $(x, y) \leq_{X \times Y} (a, b) \leq_{X \times Y} (\tilde{x}, \tilde{y})$.

Definition 7. Let $(X \times Y, \leq_{X \times Y})$ be a vector-lattice and $Z \subseteq X \times Y$ a subset. Then $\mathbf{Bot}(Z)$ is the bottom of Z if any $y \in \mathbf{Bot}(Z)$ is such that $\nexists x \in Z$ with $x < y$ (i.e., a minimal element of Z) and any two elements of $\mathbf{Bot}(Z)$ are not comparable. Likewise, by reversing the order we define that $\mathbf{Top}(Z)$ is the top of $Z \subseteq X \times Y$.

Definition 8. A subset K of a poset $(X, \leq_X)$ is said to be increasing if $x \in K, \tilde{x} \in X$ and $x \leq_X \tilde{x}$ implies $\tilde{x} \in K$.

Lemma 2 [Theorem 2.3.1 from [8], p. 30]. A sublattice of $\mathbb{R}^n$ is subcomplete if and only if it is compact.

Definition 9. Let X, Y be posets. A function $g : X \times Y \rightarrow \mathbb{R}$ is supermodular (sometimes called superadditive) on $X \times Y$ if it satisfies: $g(x^+, y^+) + g(x^-, y^-) \geq g(x^+, y^-) + g(x^-, y^+)$ for any $x^+ \geq_X x^-$ in X , and $y^+ \geq_Y y^-$ in Y (Section 4.7.2 in [2], p. 103). The function $-g$ is submodular if and only if g is supermodular.

Definition 10. Let X, Y be posets. Let $S \subseteq X \times Y$, we say that $S_y = \{x | (x, y) \in S\}$ is the section of S at $y \in Y$.

Definition 11. Let X, Y be posets, $S \subseteq X \times Y$, and S_y the section of S at $y \in Y$. A function $f : S \rightarrow \mathbb{R}$ has isotone or increasing differences in (x, y) on S if $f(x, y^+) - f(x, y^-)$ is increasing in x on $S_{y^-} \cap S_{y^+}$, for all $y^+ \geq_Y y^-$ in Y (Section 2.6.1 in [8], p. 42). A function g has antitone or decreasing differences on S if $-g$ has increasing differences.

For lattices, we also have the following definition, which is consistent with Definition 9:

Definition 12. Let X be a lattice. A function $f : X \rightarrow \mathbb{R}$ is supermodular on X if it satisfies $f(x) + f(\tilde{x}) \leq f(x \vee \tilde{x}) + f(x \wedge \tilde{x})$ for all $x, \tilde{x} \in X$, where $\vee$ denotes the maximum operator, and $\wedge$ denotes the minimum lattice operator (Section 2.6.1 in [8], p. 42).

The following lemma presents a series of properties encountered in the monotonicity proofs for stationary Markov control processes.

Lemma 3. *Let X and A be fixed nonempty Borel subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. For each $x \in X$, let A_x be a nonempty (measurable) subset of A . Let $S = \{(x, a) | x \in X, a \in A_x\}$ be a measurable subset of $X \times A_x$.*

Let $v, w : S \rightarrow \mathbb{R}$.

1. *If v and w have isotone/increasing differences, then so has $v + w$.*
2. *Let S be a lattice. If $w(\cdot, \cdot)$ is submodular, then $w(x, \cdot)$ is also submodular, for each $x \in X$.*
3. *Let $w_1 : X \rightarrow \mathbb{R}$ and $w_2 : A \rightarrow \mathbb{R}$, then $w(x, a) = w_1(x) \cdot w_2(a)$ has increasing differences if w_1 is an increasing monotone function and w_2 is a decreasing one.*

From a practical point of view, submodularity and supermodularity can be seen as *complementarity* and respectively *substitutability*. Loosely speaking, when optimizing some cost $f : X$ across features (elements of X), two features are substitutes (respectively complements) when having more of one of them does not induce us to choose more (respectively less) of the other.

Specifically, let us model production costs in an economic system with n kinds of inputs through a function f , supermodular on $\mathbb{R}_+^n$, as in [1], p.377. Suppose all inputs are used as much as possible. If f is supermodular on $\mathbb{R}_+^n$, then

$$f(x + \gamma e_i) - f(x) \leq f(x + \lambda e_j + \gamma e_i) - f(x + \lambda e_j),$$

for $x \in \mathbb{R}_+^n$, $\gamma > 0$, $\lambda > 0$, and $i \neq j$. This inequality states that the extra cost of using γ extra units of input i is raised by having λ extra units of input j (notice that here we have no economies of scale, but the opposite effect). In other words, having more of input j may reduce the effectiveness of using more of input i . In this sense, inputs i and j substitute each other's effectiveness. Thus, in economics, nondecreasing differences, hence supermodularity, are equivalent to commodities being *substitutes*. Similarly, submodularity, is equivalent to commodities being *complementary* (unrelated products or products that might be consumed together).

Sufficient Conditions for Increasing Solutions

In agreement with [7,8], we consider a parameterized collection of maximization problems, which includes the MDPs and stochastic games as particular cases. A stochastic game with only one player is the MDP. A decision rule is an equilibrium point in a one-player stochastic game if and only if it is an optimal policy in the MDP [20].

In this section we state sufficient conditions, in terms of supermodularity and increasing differences, for increasing/isotone optimal solutions. For an exhaustive discussion on lattice programming techniques see [8]. For the alternative minimization formulation we refer to [9] and Chapter 8, Monotone Optimal Policies, in [1], p. 368. The main results of this section are Topkis's Theorem stating the preservation of supermodularity under the maximization operation and Topkis's Theorem describing the argmax sets encountered in the DP framework under submodularity.

Theorem 1 [Topkis's Theorem 2.7.6 from [8], p. 70]. *Let X, A be lattices and S a sublattice of $X \times A$ and $A_x \subseteq A$ for any $x \in X$. If*

1. *$g(\cdot, \cdot) : X \times A \rightarrow \mathbb{R}$ is supermodular in $(x, a) \in (X \times A)$,*
2. *$f() : \prod_X S \rightarrow \mathbb{R}$ is defined by $f(x) = \sup_{a \in A_x} g(x, a)$, where $\prod_X S$ is the projection of S on X .*

Theorem 2 [Topkis's Theorem 2.8.1 from [8], p. 76]. *Let A be a lattice, X be a partially ordered set, and $A_x \subseteq A$ for any $x \in X$. If we have*

1. *A_x is increasing in $x \in X$ ($A_x \subseteq A_{\tilde{x}}$ for $x \leq_X \tilde{x} \in X$)*
2. *$g(x, \cdot) : A \rightarrow \mathbb{R}$ is supermodular in $a \in A$ for each $x \in X$, and has increasing differences in $(x, a) \in (X \times A)$*

then $\arg \max_{a \in A_x} g(x, a)$ is increasing in x on $\left\{x \in X \mid \arg \max_{a \in A_x} g(x, a) \text{ is nonempty} \right\}$.

Lemma 4 *[[21] Extension of Theorem 2 Theorem 2.8.1 from [8], p. 76, and Lemma 4.7.1 from [2], p. 104]. Let X be a poset and let $A := \cup_{x \in X} A_x$ be a poset indexed after X , and $g : X \times A \rightarrow \mathbb{R}$ a real-valued supermodular (sometimes called superadditive) function on $(\tilde{x}, \tilde{a}) \in X \times A_x$, for each $x \in X$, with $g(x, a) = 0$, for any $(x, a) \notin X \times A_x$. If we have*

1. *for each $x \in X$ there exists $\max_{a \in A} g(x, a)$*
2. *$A_x \subset A_{\tilde{x}}$ for any $x \leq_X \tilde{x} \in X$, and $A_x = A_{\tilde{x}}$, for any $x, \tilde{x} \in X$ not comparable (i.e., the family $\{A_x \mid x \in X\}$ is expanding)*
3. *for any $x \leq_X \tilde{x} \in X$ and $a \leq \tilde{a} \in A_x \cup A_{\tilde{x}}$ we have that $a \in A_x$ and $\tilde{a} \in A_{\tilde{x}}$ (i.e., the family is ascending)*

then for any $x^+ \geq_X x^-$ in X , and any $a^- \in \text{Top arg max}_{a \in A_{x^-}} g(x^-, a)$ either there exists $a^+ \in \text{Top arg max}_{a \in A_{x^+}} g(x^+, a)$ such that $a^+ \geq_A a^-$ in A , or there is no element in $\text{Top arg max}_{a \in A_{x^+}} g(x^+, a)$ comparable with a^- .

We notice that the operation “max” returns a value, while “arg max” returns a set. The Top arg max being thus the Top of the set returned by “arg max” according to Definition 7.

DISCRETE-TIME MODELS

Discrete-Time Markov Decision Problems—Definition

The results presented in this section hold for a large class of discrete-time nonstationary Markov control processes with finite horizon and can be easily extended for the infinite horizon.

We consider the following as a discrete-time nonstationary Markov control model with immediate rewards:

$$\mathcal{M} = \left\{ (X(t), A, A_t(x_t), S_t, Q_t(\cdot | x_t, a_t), \rho_t(x_t, a_t))_{t \in \mathbb{N}} \right\}.$$

In all generality the state and action spaces are standard Borel spaces (i.e., nonempty Borel subsets of a Polish space endowed with the σ -field of Borel subsets). We restrict this model to $X(t), A_t(x_t) \subset A$ subsets of $\mathbb{R}^q$ respectively $\mathbb{R}^m$. More general results are stated whenever the case occurs.

The system with state space $X(t)$ is observed at discrete time periods $t = 0, 1, \dots$. When the system is observed at time t to be at state $x_t \in X(t)$, an action a_t from the action set $A_t(x_t)$ should be chosen. Let $S_t := \{(x_t, a_t) \mid (x_t, a_t) \in X(t) \times A_t(x_t)\}$ be the set of possible pairs of state and action at each period. Let $A := \cup_{t \in \mathbb{N}} \cup_{x_t \in X(t)} A_t(x_t)$ be the union of all action sets. After $a_t \in A_t(x_t)$ is chosen, the following two things will happen:

1. the system will receive a reward $\rho_t(x_t, a_t)$ and
2. the system will transfer to state x_{t+1} at the next period. The state transition probability governed by $Q_t(\cdot | x_t, a_t)$, where Q_t is a measurable stochastic kernel on $X(t)$ given S_t ; that is, $Q_t(\cdot | x_t, a_t)$ is a probability measure on the σ -field of Borel subsets of $X(t)$ and $Q_t(B | \cdot, \cdot)$ is a measurable function on S_t for each $B \subset X(t)$.

Several optimality criteria are used to identify a decision-making policy: maximizing the expected total discounted reward over the course of the entire process **Total Expected Discounted Reward MDPs: Existence of Optimal Policies, Total Expected Discounted Reward MDPs: Policy Iteration Algorithm, Total Expected Discounted Reward MDPs: Value Iteration Algorithm**, maximizing the expected total reward or maximizing the average reward **Average Reward of a Given MDP Policy**.

Taking a discount factor into account does not affect any theoretical results in the finite horizon problems, while in the infinite horizon problems it does.

For a finite horizon problem ($T < \infty$) an optimal (deterministic) policy can be found by solving the following set of recursive equations (via backwards induction/DP

algorithm)

$$f_t(x_t) = \sup_{a_t \in A_t(x_t)} \left\{ \rho_t(x_t, a_t) + \lambda \int f_{t+1}(z) dQ_t(z|x_t, a_t) \right\}, \quad (1)$$

where $f_t(x_t)$ is the optimal MDP value function in state $x_t \in X(t)$. For discrete space (finite or countable) we use $p_t(x_t, a_t, x_{t+1})$ since $\int f_{t+1}(z) dQ_t(z|x_t, a_t) = \sum f_{t+1}(z) P\{Z_{x_t, a_t, t} = z\} = \sum f_{t+1}(z) p_t(x_t, a_t, z)$. The discrete analog of the optimality equations above can be written as

$$u^*(x_t) = \sup_{a_t \in A_t(x_t)} \left\{ \rho_t(x_t, a_t) + \lambda \sum_{x_{t+1} \in X(t+1)} \times p_t(x_t, a_t, x_{t+1}) u_{t+1}^*(x_{t+1}) \right\} \quad (2)$$

for $x_t \in X(t); t = 1, \dots, T-1$.

The above optimality equations are also referred to as *Bellman equations* or *functional equations*. The u^* is sometimes referred as *value-to-go function*.

Passing to the limit in the finite horizon optimality equations we obtain the characterization of an optimal policy that maximizes the expected total discounted reward of a stationary model in the infinite horizon case: $f(x) = \sup_{a \in A(x)} \{ \rho(x, a) + \lambda \int f(z) dQ(z|x, a) \}$.

An optimal policy that maximizes the expected total discounted reward can be found by solving the following set of recursive equations

$$f_i(x) = \sup_{a \in A(x)} \left\{ \rho(x, a) + \lambda \int f_{i+1}(z) dQ(z|x, a) \right\}, \quad (3)$$

where $f_i(x)$ is the optimal MDP value function in state $x \in X$. The optimal action $a^*(x)$ for state $x \in X$ is chosen as the argument $a \in A(x)$ that maximizes the right-hand side of the equation above. Value iteration (see **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**) and linear programming are algorithms used to solve the recursive optimality equations.

Definition 13. Let $\mathbf{D}_t = \{d_t : X(t) \rightarrow A | d \text{ measurable } d_t(x_t) \in A_t(x_t), \forall x_t \in X(t)\}$

be the set of decision rules at time t . Let $\mathbf{D}_t^* = \{d_t^* : X(t) \rightarrow A | d^* \text{ measurable } d_t^*(x_t) = a^*(x_t) \in A_t(x_t), \forall x_t \in X(t)\}$ be the set of decision rules maximizing the Bellman optimality equation (2) at stage t . A nonstationary Markovian (deterministic) policy is a sequence of decision rules, that is, $\pi = \{d_t\}_{t \in \mathbb{N}}$.

1. π is called a monotonic nondecreasing policy if for any $x_t \leq_{\mathbb{N}^q} \tilde{x}_t$ we have $d_t(x_t) \leq_{\mathbb{N}^m} d_t(\tilde{x}_t)$.
2. π is called a weakly monotonic nondecreasing policy if for any $x_t \leq_{\mathbb{N}^q} \tilde{x}_t$ either $d_t(x_t) \leq_{\mathbb{N}^m} d_t(\tilde{x}_t)$, or $d_t(\tilde{x}_t)$ and $d_t(x_t)$ are not comparable (see [22] for the weak order definition).

A stationary (deterministic) Markovian policy is a sequence of decision rules, that is, $\pi = \{d_t\}_{t \in \mathbb{N}}$, with $d_t = d$. It is denoted by d^∞ , and this type of policy is very important for the infinite horizon MDP. A stationary (deterministic) Markovian policy $d_t : X(t) \rightarrow A$ with $d(x) = a^*(x)$ is denoted by $(d^*)^\infty$.

More information on different types of policies can be found in **MDP Basic Components** as well as in Chapter 2 in [2], p. 22. For the optimality of the deterministic stationary policy $(d^*)^\infty$ we refer to Chapter 6, Section 6.2 in [2], p. 146, as well as **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**.

Monotonicity Obtained via Dynamic Programming

To prove a conjectured property of the optimal policies one needs

1. to identify the key structural properties that drive the result (supermodularity in our case) and
2. to prove by induction that these structural properties are inherited from one stage to another in the optimality DP equations.

In the case of monotonic optimal policies, during the second step, one uses the so-called closure properties of the key structural properties. They are maximization preserves

Table 1. Ways of Constructing a Submodular/Supermodular Function

h convex	h concave	v super-modular	v submodular	h decreasing	h increasing	v monotonic	$h \circ v$ super-modular	$h \circ v$ super-modular
X		X			X	X	X	
X			X	X		X	X	
	X	X		X		X		X
	X		X		X	X		X

submodularity and monotonicity, and expectation preserves monotonicity.

The monotonicity of the optimal policies for discrete-time MDPs is generally proven by induction via value iteration. To prove the increasing monotonicity of the optimal policy we first have to prove that at each stage t , a series of properties on the value-to-go (optimal return) function f_t defined by the optimality equations. The method suggested by Topkis's results in the section titled "Sufficient Conditions for Increasing Solutions" is informally the following:

1. Determine the regularity conditions for the existence of an optimal stationary policy for the infinite horizon case.
2. Determine the conditions under which, at an arbitrary stage t , the function f_{t+1} is monotone increasing, and thus the conditions for the optimal decisions increase with the state (Lemma 6) (proof by induction).
3. Define $w_t(x_t, a_t) = \rho_t(x_t, a_t) + \lambda \int f_{t+1}(z) dQ_t(z|x_t, a_t)$ being the immediate reward plus the value-to-go (optimal return) at stage $t+1$ and determine the conditions under which $w_t(x_t, a_t)$ is supermodular in $(x_t, a_t) \in S_t$ (either Theorem 3 or Theorem 4).
4. According to Topkis's Theorem 1, which states the preservation of submodularity under the maximization operation, we establish that $f_t(a_t) = \sup_{a_t \in A_t(x_t)} w_t(x_t, a_t)$ is supermodular. The supermodularity of the optimal return function is thus propagated inductively from stage to stage.
5. $w_t(x_t, a_t)$ being also nondecreasing, we are in the conditions of the Theorem 2, which establishes that the optimal decisions are increasing.

The submodularity of the value-to-go function requires, in most of the cases, the submodularity of the immediate rewards. For a detailed discussion on the propagation of properties via the DP we refer to [23].

Different ways to obtain submodularity conditions had their inspiration in Table 1 in [6], p. 312.

This table gives a variety of ways of constructing a submodular or supermodular function by taking the composition of two other functions with certain properties. In Table 1 we read across any given row, and the assumptions corresponding to the checked columns on the left of the double line imply the result corresponding to the checked column to the right of the double line. All properties of v and $h \circ v$ hold on a given lattice L , v is a real-valued function on L , and h is a real-valued function on the real line.

Definition 14. Let $K \subseteq \mathbb{R}^p$ and $X \subseteq \mathbb{R}^q$, and let $\int_K d\tilde{Q}(z|x)$ be the probability measure of K with respect to the distribution function $\tilde{Q}(z|x)$ on $\mathbb{R}^p$. $\tilde{Q}(z|x)$ is said to be stochastically increasing in x on X if $\int_K d\tilde{Q}(z|x)$ is increasing in x on X for each increasing subset K in $\mathbb{R}^p$.

On the basis of Theorem 3.9.1 in [8], p. 160, [24] on stochastically increasing distribution functions, we have the following Lemma.

Lemma 5 [General form in Theorem 3.9.1 from [8], p. 160, [24]]. Let X a subset of $\mathbb{R}^q$ and $\{\tilde{Q}(z|x)|x \in X\}$ a collection of distribution functions on $\mathbb{R}^p$. $\tilde{Q}(z|x)$ is stochastically increasing in x on X if and only if $\int h(z)d\tilde{Q}(z|x)$ is increasing in x on X for each increasing real-valued function $h(z)$ on $\mathbb{R}^p$.

Further reading on stochastic orderings is in [10,25,26]. In [27] integral stochastic orderings are used to prove the monotonicity

of optimal value functions. Their model has general state and action space and is a finite horizon, discrete time one. However, using [28] it can be extended to the infinite horizon case. In the case of convex order, see also [29,30].

Lemma 6 [Lemma 3.9.4. from [8], p. 164]. *Consider a discrete-time nonstationary Markov decision process with finite horizon T . If for any $t \in \{0, 1, \dots, T-1\}$, $\lambda \in (0, 1)$,*

1. $A_t(x_t) \subset A_t(\tilde{x}_t)$ for any $t \in \{0, 1, \dots, T-1\}$ and all states $x_t \leq_{\mathbb{R}^q} \tilde{x}_t \in X(t) \subset \mathbb{R}^q$
2. $\rho_t(\cdot, a_t), \rho_T(\cdot) : X(t) \rightarrow \mathbb{R}$ are non-decreasing in x_t , for any $(x_t, a_t) \in X(t) \times A_t(x_t)$
3. the distribution function for the state w in the period $t+1$ given the state x and the action a , $Q(w|x_t, a_t)$, is stochastically increasing in x , for any $(x_t, a_t) \in X(t) \times A_t(x_t)$.

Then the maximizer $f_t(\cdot) : X(t) \rightarrow \mathbb{R}$ is nondecreasing for any $t \in \{0, 1, \dots, T\}$ where

$$f_t(x_t) = \max_{a_t \in A_t(x_t)} \left\{ \rho_t(x_t, a_t) + \lambda \int f_{t+1}(z) dQ_t(z|x_t, a_t) \right\}.$$

The key element of the proof of Lemma 6 is Lemma 5.

Example 4. Maintenance with overhaul [Example 3.9.5. from [8], p. 168]. Consider a firm operating an equipment whose state $x \in \mathbb{R}^q$. A higher value of x indicates a more desirable state, as it is made more precise below. The firm can partially control the equipment by choosing from among decisions $a = 1, \dots, h$, where $h \geq 2$. Decision h is to do nothing to the equipment. Decision 1 is to completely replace the equipment. Decisions $2, \dots, h-1$ correspond to various overhaul operations on the equipment. Assume that the set of allowable decisions $A(x)$ in state x is increasing in x , and $A(x')$ is a subset of $A(x')$ for $x' \leq x$. For instance, the firm may require that the equipment be either

overhauled or replaced in very low states, but that any decision is feasible in higher states. The replacement and overhaul operations are thorough enough that the distribution function $Q(w|x, a)$ of the state in the next period does not depend on the present state x if the firm makes a decision a with $a \leq h-1$. If the firm makes decision h and does nothing to the equipment, then a higher present state x makes it more likely that there is a higher state in the next period. Therefore, the distribution function $Q(w|x, a)$ of the state in the next period is stochastically increasing in the present state x . The firm receives the return $\rho(x, a)$ for making decision a in state x , and $\rho(x, a)$ is decreasing in the state x . Assume that if $x' \leq x''$, then $\rho(x', a) - \rho(x'', a)$ is increasing in the decision a on $A(x')$. According to Lemma 6 (Lemma 3.9.4. from [8], p. 164) the optimal control decisions increase with the state of the system.

A series of examples can be found in Chapter 3 in [8], p. 167; Chapter 8 in [1], p. 387; Chapter 4, [2], p. 108.

Finite Horizon Case

Definition 15. Let $K \subseteq \mathbb{R}^q$ and $S \subseteq \mathbb{R}^n$, and let $\int_K d\tilde{Q}_y(z)$ be the probability measure of K with respect to the distribution function $\tilde{Q}_y(z)$ on $\mathbb{R}^q$. $\tilde{Q}_y(z)$ is said to be stochastically supermodular in y on S if $\int_K d\tilde{Q}_y(z)$ is supermodular in y on S for each increasing subset K in $\mathbb{R}^q$.

Theorem 3 [Theorem 3.9.2. from [8], p. 165]. *Consider $\mathcal{M}$, a discrete-time nonstationary Markov decision process with finite horizon T . If we have for any $t \in \{0, 1, \dots, T-1\}$*

Condition 1. $A_t(x_t)$ finite and $\rho_t(\cdot, a_t), \rho_t(\cdot) : X(t) \rightarrow \mathbb{R}$ bounded

Condition 2. $S_t := \{(x_t, a_t) | (x_t, a_t) \in X(t) \times A_t(x_t)\}$ is a sublattice of the vector-lattice $\mathbb{R}^n = \mathbb{R}^q \times \mathbb{R}^m$

Condition 3. $A_t(x_t) \subset A_t(\tilde{x}_t)$ for all states $x_t \leq_{\mathbb{R}^q} \tilde{x}_t \in X(t) \subset \mathbb{R}^q$

Condition 4. $\rho_t(\cdot, a_t), \rho_T(\cdot) : X(t) \rightarrow \mathbb{R}$ are nondecreasing in x_t , for any $(x_t, a_t) \in X(t) \times A_t(x_t)$

Condition 5. $\rho_t(\cdot, \cdot) : X(t) \rightarrow \mathbb{R}$ are supermodular in $(x_t, a_t) \in S_t$

Condition 6. the distribution function for the state w in the period $t + 1$ given the state x_t and the action a , $Q_t(w|x_t, a_t)$, is stochastically increasing in x_t , for any $(x_t, a_t) \in X(t) \times A_t(x_t)$, and is stochastically supermodular in $(x_t, a_t) \in S_t$.

Then for any $t \in \{0, 1, \dots, T - 1\}$,

1. the function

$$w_t(x_t, a_t) = \rho_t(x_t, a_t) + \lambda \times \int f_{t+1}(z) dQ_t(z|x_t, a_t) \quad (4)$$

is supermodular in $(x_t, a_t) \in S_t$

2. $f_t(x_t) = \sup_{a_t \in A_t(x_t)} \{\rho_t(x_t, a_t) + \lambda \int f_{t+1}(z) dQ_t(z|x_t, a_t)\}$ is supermodular in $x_t \in X(t)$

3. the set of optimal decisions $\arg \max_{a_t \in A_t(x_t)} \{\rho_t(x_t, a_t) + \lambda \int f_{t+1}(z) dQ_t(z|x_t, a_t)\}$ is nondecreasing in the state $x_t \in X(t)$, and

4. there is a greatest (least) optimal decision for each state x_t and this greatest (least) optimal decision is increasing in $x_t \in X(t)$ (i.e., there exists at least one monotonic nondecreasing optimal policy).

Condition 1 can be reduced to X and A to be finite, or generalized [7] by conditions leading to the existence of an optimal action at each step such as compactness with smoothness conditions on $w_t(\cdot, \cdot)$.

In the proof of Theorem 3, f_t is supermodular as a sum of supermodular functions ($\rho_t(x_t, a_t)$ and $\int f_{t+1}(z) dQ_t(z|x_t, a_t)$). In applications, $\int f_{t+1}(z) dQ_t(z|x_t, a_t)$ nondecreasing in x means that a control has more of an impact when a state is “greater,” “at a higher level” (in the order of the state space) than when it is “at a lower level.”

The proofs of Theorem 3 also hold for the case in which $X(t)$ is only a sup-lattice, while

$A_t(x_t)$ is a lattice for any $x_t \in X(t)$. The other theorems for the finite horizon case are consequences of Theorem 3 stated in forms more convenient to be used.

Corollary 1 [[21], generalization of Lemma 4.7.2 from [2], p. 106]. Let $\{y_j\}_{j \in X}$, $\{\tilde{y}_j\}_{j \in X}$, $\{v_i\}_{i \in X}$ be real-valued nonnegative numbers indexed after $X \subset \mathbb{R}^q$ countable. Suppose $\sum_{j \in K} y_j \geq \sum_{j \in K} \tilde{y}_j$ for any increasing subset $K \subseteq X$ and the sums $\sum_{j \in X} y_j$, $\sum_{j \in X} \tilde{y}_j$ are finite. If for any $i, j \in (X, \leq_{\mathbb{R}^q})$ such that $i \geq_{\mathbb{R}^q} j$ we have $v_i \geq v_j$ then $\sum_{j \in X} v_j y_j \geq \sum_{j \in X} v_j \tilde{y}_j$.

This corollary of Lemma 5 is essential in determining the extent of generality of the state space in the discrete state space model, as it will be seen in the following Theorem 4.

By summarizing the results of [7], Theorem 4.7.5 from [2], p. 108 and using a discrete version of Theorem 3 from [8], p. 165 we have the following result.

Theorem 4. Consider $\mathcal{M}$ a discrete-time nonstationary Markov decision process with finite horizon T , and finite state and action space subsets of the vector-lattices $\mathbb{N}^q \times \mathbb{N}^m$. For any $t \in \{0, 1, \dots, T - 1\}$:

1. (a) $\rho_t(\cdot, \cdot)$ is supermodular/superadditive in $(x_t, a_t) \in X(t) \times A_t(x_t)$
 (b) $\rho_t(\cdot, a_t) : X(t) \rightarrow \mathbb{R}$ is nondecreasing
 (c) $\rho_T(\cdot) : X(T) \rightarrow \mathbb{R}$ is nondecreasing
2. (a) $A_t(x_t) \subset A_t(\tilde{x}_t)$ for any $x_t \leq \tilde{x}_t \in X(t) \subset \mathbb{N}^q$
 (b) $S_t := X(t) \times A_t(x_t)$ is a lattice with respect to the componentwise partial order on the product space $\mathbb{N}^n = \mathbb{N}^q \times \mathbb{N}^m$
 and for any increasing subset $K \subseteq X(t + 1)$
3. $Q_t(\cdot, \cdot) := \sum_{x_{t+1} \in K} p_t(\cdot, \cdot, x_{t+1})$ is supermodular/superadditive in $(x_t, a_t) \in X(t) \times A_t(x_t)$
4. $Q_t(\cdot, a_t) := \sum_{x_{t+1} \in K} p_t(\cdot, a_t, x_{t+1})$ is nondecreasing in $x_t \in X(t)$, for any $a_t \in A_t(x_t)$.

Then there exists an optimal policy nondecreasing in state.

Proof. Given

$$u_t^*(x_t) = \max_{a_t \in A_t(x_t)} w_t(x_t, a_t), \quad (5)$$

$$u_T^*(x_T) = \rho_T(x_T).$$

If $w_t(\cdot, \cdot) : X(t) \times A_t(x_t)$ is supermodular and the max is attained, Theorem 2 holds and

$$a_{x_t}^* := \max \arg \max_{a_{x_t} \in A_t(x_t)} w_t(x_t, a_{x_t}).$$

Let $a_t \leq_{\mathbb{N}^q} \tilde{a}_t \in \{A_t(x_t) \times A_t(\tilde{x}_t)\} \cup \{A_t(\tilde{x}_t) \times A_t(x_t)\}$ and $K \subseteq X(t+1)$ an increasing subset. To prove by induction in t that

$$\sum_{x_{t+1} \in X(t+1)} p_t(x_t, a_t, x_{t+1}) u_{t+1}^*(x_{t+1})$$

is supermodular in $(x_t, a_t) \in X(t) \times A_t(x_t)$ for any t we use that for any t , $Q_t(\cdot, a_t)$ is nondecreasing, S_t is a sublattice, the action family is expanding, as well as Corollary 1 with

- $x_t^- \leq x_t^+ \in X(t)$
- $a_t^- \leq a_t^+$, where $a_t^- \in A_t(x_t^-)$ and $a_t^+ \in A_t(x_t^+)$
- $y_j = [p_t(x_t^+, a_t^+, j) + p_t(x_t^-, a_t^-, j)]$, where $j \in X(t+1)$
- $\tilde{y}_j = [p_t(x_t^-, a_t^+, j) + p_t(x_t^+, a_t^-, j)]$, where $j \in X(t+1)$
- $v_j = u_{t+1}^*(j)$.

Checking Theorem 4's hypotheses for each increasing set is not an easy task. In [31] for a production control problem on for vector-lattices $\mathbb{N}^q$, a useful tool was the decomposition of any increasing set into elementary increasing subsets of the type

$$K_s = \{x \in X(t) \subseteq \mathbb{N}^q \mid x \geq_{\mathbb{N}^q} s\},$$

and the fact that any increasing subset of $X(t) \subseteq \mathbb{N}^q$ can be written as $\bigcup_{j=1}^r K_{s_j}$, where $s_1, \dots, s_r \in \mathbb{N}^q$.

Also Theorem 4's hypotheses are greatly simplified if $A_t(x_t) = A$ constant.

Corollary 2 [Theorem 4.7.4 from [17], p. 107]. Suppose for $X = \mathbb{N}$, $A(x) = A$, $t = 1, \dots, T-1$ that

1. $\rho_t(\cdot, a) : \mathbb{N} \rightarrow \mathbb{R}$ is nondecreasing for all $a \in A$
2. $p_t(\cdot, a, x_{t+1})$ is nondecreasing in $x_t \in \mathbb{N}$, for any $x_{t+1} \in \mathbb{N}$ and $a \in A$
3. $\rho_t(\cdot, \cdot)$ is supermodular(submodular) in $(x_t, a_t) \in \mathbb{N} \times A$
4. $p_t(\cdot, \cdot, x_{t+1})$ is supermodular(submodular) in $(x_t, a) \in \mathbb{N} \times A$ for all $x_{t+1} \in \mathbb{N}$
5. $\rho_T(\cdot) : \mathbb{N} \rightarrow \mathbb{R}$ is nondecreasing.

Then, there exist optimal decision rules $d_t^*(x)$, which are nondecreasing (nonincreasing) in x for $t = 1, \dots, T-1$.

The conditions for minimization problems from the Corollary 2 are a particular case of the general form from Theorem 8-16 from [1], p. 427.

Dragut [21] shows the existence of weakly monotonic optimal policies under less restrictive requirements than the ones from [7,8,17].

Proposition 1 [21]. Consider a discrete-time nonstationary Markov decision process with finite horizon T , and finite state and action spaces $\mathbb{N}^q \times \mathbb{N}^m$. If for any $t \in \{0, 1, \dots, T-1\}$ Conditions 1, 3, 4 of Theorem 4 hold and

1. $X(t)$, $A_t(x_t)$, for any $x_t \in X(t)$ are bounded subsets without holes of the infinite vector-lattice $\mathbb{N}^q \times \mathbb{N}^m$
2. $A_t(x_t) \subset A_t(\tilde{x}_t)$ for any $x_t \leq_{\mathbb{N}^q} \tilde{x}_t \in X(t)$, and $A_t(x_t) = A_t(\tilde{x}_t)$ for any $x_t, \tilde{x}_t \in X(t)$ not comparable (i.e., the family $\{A_t(x_t) \mid x_t \in X(t)\}$ is expanding)
3. for any $x_t \leq_{\mathbb{N}^q} \tilde{x}_t \in X(t)$, and $a \leq_{\mathbb{N}^m} \tilde{a} \in A_t(x_t) \times A_t(\tilde{x}_t)$ or $A_t(\tilde{x}_t) \times A_t(x_t)$ we have that $a \in A_t(x_t)$ and $\tilde{a} \in A_t(\tilde{x}_t)$ (i.e. the family is ascending) and if for any increasing subset $K \subseteq X(t+1)$

4. if $(\exists)x_t \neq \tilde{x}_t \in X(t)$ such that $A_t(x_t) \neq A_t(\tilde{x}_t)$ then $\sum_{x_{t+1} \in K} p_t(x, a_t, x_{t+1})$ is non-decreasing in $a_t \in A_t(x)$, for any $x \in X(t)$,

then there exist optimal decision policies, which are weakly monotonic nondecreasing in the state, for any $t \in \{0, 1, \dots, T-1\}$.

Equivalency between submodularity and weak submodularity on $X = \mathbb{R}^2$ is discussed in Theorem 8-4 from [1], p. 378.

Backward Induction Algorithms. The basis of this backward induction algorithm is the intuitive argument called *Principle of Bellman Optimality*: “An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.” The ideas behind backwards induction are as follows.

- Envision being in the last time period for all the possible states and decide the best action for those states. This yields an optimal value (max value for a maximization problem) for that state in that period.

$$\forall x_T \in X(T), \quad u_T^*(x_T) = \rho_T(x_T).$$

- Next envision being in the next-to-last period for all the possible states and decide the best action for those states (an elements of the $\arg \max$), given you now know the optimal values of being in various states at the next time period. This is done solving the optimality equation (2) at each stage t .
- Continue this process until you reach the initial time $t_0 = 0$.

Since the optimal policy must give optimal actions for each state, in the backward induction algorithm the optimal policy chosen is $\pi = \{d_t\}_{t \in \mathbb{N}}$, where $d_t = a_t^* \in \arg \max_{a_t \in A_t(x_t)} w_t(x_t, a_t)$.

From Section 4.7.6 in [2], p. 111, a monotonic backwards induction algorithm can only

be used for very particular nonstationary MDPs with finite X and with $A(x) = A$ for all $x \in X$. The monotonic backwards algorithm reduces the complexity of finding an optimal policy. It considers only nondecreasing control actions instead of the whole $A(x)$. Thus, the algorithm looks for candidate actions for the $\arg \max$ for a given state x only in the superior cones of the elements of the $\arg \max$ of all states $\tilde{x}$, which are smaller than x . A series of examples with different action spaces can also be found in Chapter 8 in [1], p. 428.

A more general algorithm can be derived for multidimensional state and action spaces but is not computationally feasible. Even very small lattices cannot be entirely handled in the main computer memory (with all the couples state–action subsets). Moreover, in many applications the state and action space are merely bounded subsets without holes in a countable lattice. If we strengthen the assumptions and we require the action space to be a lattice instead of partially bounded subset of $\mathbb{N}^m$, for all these unordered actions, there will exist a greater action (their maximum) in the vector order. In real life, however, such “max” actions imply a large increase of the state/action space, possibly canceling the computational gains brought by monotonicity. Avoiding this loss, requirements from Lemma 4 lead to weakly monotonic nondecreasing policies [21]: $X(t)$ and A_t are covered by disjoint antichains using Lemma 1; for each segment of lattice explored by the algorithm only its bottom and top are kept, and the search is “ordered” (antichain sweep between top and bottom—Fig. 1).

Infinite Horizon Case

For an infinite horizon problem ($T = \infty$), a stationary model (i.e., a model in which the parameters do not change over time) is a standard assumption in the literature. Following this assumption, we drop the time index t for the infinite horizon models presented in the rest of this article.

We consider $\{X, A, (A(x), x \in X), Q, \rho\}$ as a discrete-time infinite horizon stationary Markov control model with immediate rewards, where X and A are Borel spaces, and $Q(B|x, a)$ a stochastic kernel on

X . Consider $S := \{(x, a) | (x, a) \in X \times A(x), A(x) \text{ is a Borel subset}\}$ and $\rho : S \rightarrow \mathbb{R}$ a measurable function. Then $Q(\cdot | x, a)$ is a probability measure on X , for every $(x, a) \in S$, and $Q(B | \cdot, \cdot)$ is a measurable function on S , for every Borel set B on X .

A policy π and an initial state $x = x_0$ determine the Markov control process.

To grasp the mechanism behind the derivation of several types of monotonicity conditions in the infinite horizon case we present the interesting open research formulation made by [11] when presenting the result of [9].

Theorem 5 [9,11]. *Assume X, A are posets. α the discount factor then f_t is weakly increasing if for an arbitrary t the following hold.*

1. S is “monotonically complete” in the sense that $x \leq x'$, $a \leq a'$ and $(x, a), (x', a') \in S$ imply that (x, a) and (x', a') belong to S (i.e. $a \in A(x)$ and $a' \in A(x')$)
- 2.

$$w_t(x, a) := \rho(x, a) + \lambda \int f_{t+1}(z) dQ(z | x, a) \quad (6)$$

is finite and has increasing differences

3. there is a maximizer f_t at stage t such that $f_t(x)$ is weakly minimal in $D_t^*(x)$, $x \in X$, where $D_t^*(x)$ is the set of all decisions in $A(x)$ maximizing the Bellman optimality equation at stage t .

Then f_t is “weakly” increasing.

By considering the order “ $\geq$ ” instead of the order “ $\leq$ ” on X and A , one obtains the criteria for f_t to be decreasing (increasing). Different criteria to fulfill the requirements of “monotonically complete” led to several variants of the Theorem 5. Surely, for applications, new less restrictive criteria can still be developed starting from this formulation. Another source of results was Table 1 in [6], p. 312, which gives ways of constructing a supermodular/submodular functions.

Discounted Reward Criterion. In the case of the expected total discounted reward we distinguish among results with bounded and unbounded rewards.

In applications with evolving systems (queueing control, production) there is rarely a preset bound to the system state: the system “grows” and/or “shrinks” and the rewards/costs follow in a nondecreasing way—prior costs/rewards bounds are restrictive. Thus, Section 6.10 in [2], p. 232 sets relative bounds for the rewards growth with respect to the system state and limits the probability of reaching distant states.

Assumption 1 [Assumptions 6.10.1, 6.10.2 from Section 6.10 in [2], p. 232]. Suppose $\varpi : X \rightarrow \mathbb{R}_+$ satisfies $\inf_{x \in X} \varpi(x) > 0$. Also assume the following.

1. There exists a constant $\mu < \infty$ such that $\sup_{a \in A(x)} |\rho(x, a)| \leq \mu \varpi(x)$, for $x \in X$.
2. There exists a constant $0 \leq \kappa < \infty$, and for each $0 \leq \lambda < 1$, there exists an $0 \leq \beta < 1$ and an integer J such that for all $\pi = (d_1, \dots, d_J)$ where $d_k \in D^{\text{MD}}$ the set of deterministic Markovian decision rules, $1 \leq k \leq J$ hold
 - (a)

$$E^\pi \{\varpi(X_{t+1}) | X_t = s, d_t(X_t) = a\} \leq \kappa \varpi(x)$$

for all $a \in A(x)$ and $x \in X$.

(b)

$$E^\pi \{\varpi(X_{t+1}) | X_t = x\} \leq \beta \lambda^{-1} \varpi(x)$$

for $x \in X, t = 1, 2, \dots$,

These assumptions ensure the convergence of value iteration without a prior bound on rewards. The finite horizon results from Corollary 2 can be extended in the more general framework of obtaining structured optimal stationary policies from Section 6.11 in [2], p. 255.

Corollary 3 [Theorem 6.11.6 from [17], p. 258]. Suppose for $X = \mathbb{N}$ and $A(x) = A$ that Assumption 1 that

1. $\rho(\cdot, a) : \mathbb{N} \rightarrow \mathbb{R}$ is nondecreasing for all $a \in A$
2. $Q(\cdot, \cdot) := \sum_{y \geq k} p(\cdot, \cdot, y)$ is supermodular/superadditive in $(x, a) \in \mathbb{N} \times A$ for all $k \in \mathbb{N}$
3. $Q(\cdot, a) := \sum_{y \geq k} p(\cdot, a, y)$ is nondecreasing in $x \in \mathbb{N}$, for any $a \in A$ for all $k \in \mathbb{N}$
4. $\rho(\cdot, \cdot)$ is supermodular/submodular in $(x, a) \in \mathbb{N} \times A$.

Then, there exists an optimal stationary policy $(d^*)^\infty$ with the property that $d^*(x)$ is nondecreasing (nonincreasing) in x .

This holds since the pointwise limit of any sequence of nondecreasing functions is nondecreasing. To ensure the existence of a nondecreasing decision, the hypothesis is stated using the elementary increasing sets on $\mathbb{N}$.

In the monotone value iteration algorithm, the maximization can be carried out only over structured action sets (Section 6.11 in [2], p. 259).

For state and action spaces intervals on $\mathbb{R}$, [32] gives alternative sets of sufficient conditions for $w_t(x, a)$ in Equation (6) to have increasing/decreasing differences. Conditions for where the states have a linear dependency (Conditions 1, 2) are given under the following common assumptions:

Assumption 2 [Assumptions 4.2 from [32]; and 4.2.1 and 4.2.2 from [33], p. 46].

1. X and A are intervals on $\mathbb{R}$.
2. $(1 - \lambda)a + \lambda a' \in A(1 - \lambda)x + \lambda x'$, for all $x, x' \in X$, $a \in A(x)$, $a' \in A(x')$, $\lambda \in [0, 1]$, $A(y) \subset A(x)$, for $x \leq y$ in X , and $A(x)$ is convex for all $x \in X$.
3. ρ is convex on S .
4. The one-stage cost $\rho : S \rightarrow \mathbb{R}$ is nonnegative, lower semicontinuous, and inf-compact on S .
5. The transition law Q is strongly continuous.
6. There is a policy π such that its expected total discounted cost $V_\alpha(\pi, x) < +\infty$ for all $x \in X$.

Condition 1.

1. Q is given by $x_{t+1} = \gamma x_t + \delta a_t + \xi_t$, $t = 0, 1, \dots$, $\gamma, \delta > 0$, and $\{\xi_t\}$ is a sequence of i.i.d. r.v., which take values in $Z \subset \mathbb{R}$.
2. (Obviously, assuming that $\gamma x + \delta a + z \in X$, for all $x \in X$, $a \in A(x)$, and $z \in Z$.)
3. $x \rightarrow A(x)$ is descending.
4. ρ is supermodular on S .

Condition 2.

1. S is a lattice.
2. Consider Requirement 1 from Condition 1, but changing $\gamma, \delta > 0$ by $\gamma > 0$, $\delta < 0$
3. $x \rightarrow A(x)$ is ascending.
4. ρ is submodular on S .

Theorem 6. Suppose Assumption 2 holds. Then there is a decreasing stationary optimal policy under Condition 1 and an increasing stationary optimal policy under Condition 2.

Flores-Hernandez and Montes-deOca [32] also draw conditions for nonlinear models. Other results on monotonicity and convexity of the optimal value function for discounted MCPs are provided in Lemmas 6.1 and 6.2 from [34].

Relevant research has also been developed by assuming that the reward function is bounded over the domain of the state space. Some monotonicity results have the submodularity of the DP operator given directly [1, 6, 9, 35–39] and some in terms of the elements of the Markov control model [1, 7, 11, 22, 40]. One popular criterion for submodularity on $X = \mathbb{R}^2$ is that w_t is twice differentiable on a convex set and has a nonnegative second partial derivative (Corollary 8-1 from [1], p. 376). The proof of this criterion led to the aforementioned economic interpretation of submodularity.

For a general real poset state space the monotonicity conditions are attained by imposing the objective function to be lower/upper semicontinuous and the state space to be a compact set such that it can achieve its minimum/maximum. Thus in

general, lattice DP is stated in Corollary 8-4 from [1], p. 380 and particularized for $X = \mathbb{R}^2$ applications in Theorem 8-5 p.381, Corollary 8-5a, Corollary 8-5b. The result on $X = \mathbb{R}^2$ is also given in a Markov control setting in Theorem 8-16 p.427.

For countable poset state space, the theorem in [7] states the existence conditions for structured optimal decision rules, with respect to the state of the system. It considers discrete MDPs with $S_t := \{(x_t, a_t) | (x_t, a_t) \in X(t) \times A_t(x_t)\}$ sublattice of an arbitrary lattice, and with summarized utility functions.

For an infinite horizon stationary discounted $\mathcal{M}$ model, Theorem 3 holds (Section 39 from [8], p. 166).

In [41,42] alternative sets of conditions are derived for achieving the existence of monotonically nondecreasing optimal policies for the discounted stationary Markov control process with state space subset of $\mathbb{Z}^N$, the admissible actions set $A(x)$ finite for any $x \in X$ and with the immediate nonnegative rewards bounded on $S := X \times A(x)$. Theorem 4 holds provided that an optimal pure stationary policy exists. Numerical consequences for the infinite-stage model are investigated in [43,44].

Average Expected Reward Criterion. For applications, structural results for the average reward models were obtained by using the limit of discounted models a long time ago [45]. Derman [46], p. 121 and Tijms [47], p. 221 provide examples of structured policy iteration algorithms.

Under the vanishing discount approach for countable-state models, results similar to the discounted infinite horizon ones from Corollary 3 are established (Section 8.11 in [2], p.225) in the general framework of obtaining structured optimal stationary policies.

Assumption 3 [Assumptions 8.10.1, 8.10.2, 8.10.3, 8.10.4 from [2], p. 415].

1. For each $x \in X \subset \mathbb{N}$, $-\infty < \rho(x, a) \leq R < \infty$.
2. For each $x \in X$ and $0 \leq \lambda < 1$, the solution of equation (5) $u_\lambda^*(x) > -\infty$.

3. There exists a $K < \infty$ such that, for each $x \in X$, $u_\lambda^*(x) - u_\lambda^*(0) \leq K$ for $0 \leq \lambda < 1$.
4. There exists a nonnegative function $M(x)$ such that
 - (a) $M(x) < \infty$
 - (b) for each $x \in X$, $u_\lambda^*(x) - u_\lambda^*(0) \geq -M(x)$ for all λ , $0 \leq \lambda < 1$ and
 - (c) there exists an $a_0 \in A_0$ for which $\sum_{y \in X} p(0, a_0, y)M(y) < \infty$.

Corollary 3 [Theorem 8.11.3 from [17], p. 426]. Suppose that $X = \mathbb{N}$ and $A(x) = A$ and Assumption 2 and the ones in Corollary 3 hold.

Then there exists a $\liminf$ average optimal stationary policy $(d^*)^\infty$ with the property that $d^*(s)$ is nondecreasing (nonincreasing) in s . Further, when X is finite, $(d^*)^\infty$ is average optimal.

A monotone policy iteration algorithm is presented for a unichain model.

For state and action spaces intervals on $\mathbb{R}$, in [32] there are derived increasing/decreasing stationary optimal policies under similar assumptions with the discounted case. Noncountable Markov control processes examples are discussed there as well as in Example 5.4.2 from [33].

Optimal Monotone Policies–Based Algorithms. In the absence of a uniqueness assumption on optimal policies, a series of approximating finite MDP problems approximate the underlying infinite MDP problem's optimal value. In [48,49] monotonicity of optimal policies is exploited to establish the existence and discovery of solution and forecast horizons in a class of nonhomogeneous MDPs. Garcia and Smith [48] used a “forecast index” such that the first-period optimal decision is monotonically nondecreasing in that index. Cheevaprawatdomrong *et al.* [49] do not require the uniqueness of an infinite horizon optimal first policy. Their assumption is called *well-posedness*, that is, a first-period optimal decision can be determined by the data associated with a finite number

of periods of problem data. The property exploited was monotonicity of the first decision with respect to problem horizons, which generalizes [50].

Other Markovian Models

Consider a POMDP model as in ***Partially Observable MDPs (POMDPs): Introduction and Examples***. A finite action/state space, POMDP model can be analyzed by transforming it into an equivalent continuous-state MDP in which the system state is a probability distribution on the unobserved states in the POMDP, and transition probabilities are derived through Bayes' rule (see ***Reduction of a POMDP to an MDP***). Therefore, standard solution techniques used for solving MDPs can also be used for solving POMDPs. Smallwood and Sondik [51] present conditions for general POMDPs under which, if there are only a finite number of control intervals remaining, the optimal payoff function is piecewise-linear and convex. Sondik [52] extends these results to the infinite horizon discounted case. In this framework for a discrete-time finite POMDP, Lovejoy [53] uses likelihood ratio ordering to provide sufficient conditions for the optimal value function to be a monotone.

For POMDPs with finite or countably infinite (core) state and observation spaces and finite action sets, see [54]. For monotonicity results on general state, action, and observation spaces we refer to [55].

In optimal inspection [56] and replacement [40], the stochastic ordering was used to derive monotonicity results on subsets of the action space. A general set of conditions was derived in [57] for countable-state problems and a survey followed in [58].

The “stochastically increasing” ordering led to more restrictive conditions due to the first-order stochastic dominance interaction with the conditioning as pointed out by [54] when investigating POMDPs with three or more states.

The monotonicity was also used for the value function approach to approximate the optimal value function with a function defined over the same information space.

Monotonic structure results are derived in the constrained MDPs for the flow control in communication networks. Extensive reviews can be found in Section 16.6 in [59], p. 503 and Chapter 5 in [60], p. 45.

Other structural results concerning randomized monotone policies in infinite horizon average-cost constrained MDP can be found in [61,62].

COMPARABLE STATISTICS FOR THE OPTIMAL MONOTONIC NONDECREASING POLICIES

DP problems provide optimal policies often too complicated to be used for deriving insights. Intuitively one might expect that the monotonicity of the decision rules would imply a stable structure of the optimal paths (i.e., an increase in a exogenous parameter θ will lead to a uniform change of the optimal path). This might be true in the unidimensional stationary case [63]. In [64] a two-dimensional example of a chaotic optimal path for a monotonic decision rule is given, as well as an analysis of connections between monotonic policies and the optimal path when $A_t(x_t)$ are sublattices of $\mathbb{R}^m$ for any $x_t \in X(t)$. The authors consider a general dynamic decision problem (DDP), where in each period, the set of allowable actions may depend on the action from the previous period, as well as of the current state, and a set of exogenous parameters $\Theta := \Theta_1 \times \dots \times \Theta_l$, where $\Theta_i \subseteq \mathbb{R}$ is locally compact. For the DDP considered it is described by: an action space $A = A_1 \times \dots \times A_n$, $A_i \subseteq \mathbb{R}$, and A_i locally compact, such that A is a vector-lattice, a finite set of random stationary Markovian processes $x_i^t \in X_i$, having a positive persistence of uncertainty, where $X = X_1 \times \dots \times X_m$, $X_i \subseteq \mathbb{R}$, and X_i locally compact, and the payoff per period $F(a^t, a^{t-1}, x^t; \theta) + \sum_i \kappa_i(a_i^t)$, where both $F(a^t, a^{t-1}, x^t; \theta)$, and $\sum_i \kappa_i(a_i^t)$ are upper semicontinuous. The assumptions considered were:

1. the value function exists and is well defined under total expected reward criterion.

2. the graph $\Gamma := \{(a, a', x; \theta) | a \in A(a', x; \theta)\}$ is a sublattice of $A \times A \times X \times \Theta$ (in the product order), where $A(a', x; \theta)$ is the set of allowable actions in state x .
3. $A(a', x; \theta)$ is a complete sublattice of A .
4. $F(a, a', x; \theta)$ is supermodular in $(a, a') \in \Gamma(x; \theta) := \{(a, a') | a \in A(a', x; \theta)\}$ and has nondecreasing differences in $((a, a'), (x; \theta))$ on $A(a', x; \theta)$ (i.e., full complementarities).

Proposition 2 [64]. *Consider a dynamic decision problem with a monotone decision rule. Then for any initial conditions s^0, x^0 and any two parameter choices $\theta^+ \geq \theta^-$, for all $t > 0$, $a^t(\theta^+) \geq a^t(\theta^-)$ along any sample path, where in every period the largest optimal action is always chosen.*

As a consequence, we have a robustness result that can also be derived independently as a direct consequence of Theorem 2.

Corollary 4. *Let A be a lattice, $\Theta \subset \mathbb{R}^p$ a sublattice, and X a partially ordered set with $A_{x,\theta} \subseteq A$ for any $(x, \theta) \in X \times \Theta$. If we have*

1. *for each $(x, \theta) \in X \times \Theta$ there exists $\max_{a \in A} g(x, \theta, a)$*
2. *$A_{x,\theta}$ is increasing in $(x, \theta) \in X \times \Theta$*
3. *$g(x, \theta, \cdot) : A \rightarrow \mathbb{R}$ is supermodular in $a \in A$ for each $(x, \theta) \in X \times \Theta$, and has increasing differences in $((x, \theta, \cdot) \in X \times \Theta \times A)$,*

then for any $x^+ \geq_X x^- \in X$, $\theta^- \leq_{\mathbb{R}^p} \theta^+ \in \Theta$, we have that

$$\begin{aligned} & \max_{a \in A} \arg \max_{a \in A} g(x^-, \theta^-, a) \\ & \leq \max_{a \in A} \arg \max_{a \in A} g(x^+, \theta^+, a). \end{aligned}$$

Similar results can be obtained for weakly nondecreasing monotonic policies with or without separable objectives [21].

CONTINUOUS-TIME MODELS

For *semi-Markov decision processes* (SMDPs) we refer to Chapter 11 in [2], p. 530. In the case of discounted criterion, the discounted semi-Markov problem is transformed in a discrete-time model with a discount given by a decision rule–dependent factor. If Assumption 3 holds in the associated discrete-time model, then the results concerning the existence of solutions of the optimality equation and optimal policies hold as well.

Assumption 4. [Assumption 11.1.1 from [2], p. 532]. Let $F(t|x, a)$ be the probability that the next decision epoch occurs within t time units of the current decision epoch; given the chosen action a from $A(x)$ in state x at the current decision epoch, there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$F(\delta|x, a) \leq 1 - \varepsilon$$

for all $x \in X$ and $a \in A(x)$.

For bounded rewards and finite $A(x)$ (or under the alternative conditions), Theorem 11.3.2 from [2], p.545 states the existence of an optimal stationary deterministic policy and the convergence of the value iteration algorithm.

For unbounded rewards, Theorem 11.3.3 from [2], p. 547 states the conditions for the convergence of the policy (modified) iteration algorithm: as a consequence of its convergence, the discrete case monotonicity results still hold.

For the average reward criterion in Section 11.4.3 from [2], p. 556, transform the model to an equivalent discrete-time model using the transformation. For the finite state case and $A(x)$ finite (or under the alternative conditions) and a unichain transition probability matrix for every stationary policy Theorem 11.4.6, p. 557 ensures the use of the discrete case results.

For general assumptions see [22,27].

For a general continuous-time Markov Decision Processes (CTMDP) the use of uniformization for analyzing Markov processes was formalized by [65] in the context of countable-state continuous-time models

with uniformly bounded transition rate. See also Section 11.5 in [2], p. 561. For both discounted and average reward criteria the discrete case monotonicity results hold in the equivalent discrete-time MDP.

The research for the existence of a monotonic optimal policy in CTMDP was related to controlled queueing systems and their threshold policies. Seminal papers were [66] for multimodular functions and [67] for switching. For the relation between multimodularity and submodularity we refer to Section 10.7 in [68], p. 218.

Going beyond problem specific proofs two general techniques are used. A DP one [69–71] generalized by discrete event systems DP in [5, 68, 72–74].

The value function is a composition of event (as arrivals controlled/fork arrivals, departure, tandem departures) operators. Since events occur only at transition times, the classical uniformization operator is replaced by a concatenation of such operators. The submodularity/increasingness properties are studied per operator. After the “uniformization” takes place, the embedded MDP is studied. Fictitious transitions might be added in some states. In event-based DP, establishing the structure of an optimal policy can then be performed by verifying that the different operators constituting the problem satisfy certain properties. This decomposition of the problem is extremely useful, since the properties of commonly used operators in queueing control problems can be investigated individually for once and indexed as in [5]. Thus, if a queueing control problem is composed of known operators, establishing the structure of the optimal policy is usually trivial by checking the already investigated operators. The event-based DP also studies the propagation of two more types of second-order properties related to convexity: multimodularity in [66] and directional convexity in [75, 76] and later in [25].

A second technique is the sample path one as in [77–81]. Compared to the technique based on propagating submodularity/supermodularity through DP optimality equations in an MDP setting, sample path analysis can solve only a subset of problems,

but it provides more insight and it requires a set of assumptions less restrictive than the Markovian ones. A review of different ways of comparing sample paths based on the ideas of vector majorization and set dominance can be found in [81]. They group the proof techniques into three general classes: forward induction, backward induction, and interchange arguments. They focus on deterministic policies, but using [82] proofs can be extended to randomized policies. Comparisons of the expectations of functions of one or more state variables are allowed only after the policies have been compared on a pathwise basis.

In order to apply the forward induction, one identifies a so-called hypothesized optimal policy and also derives the equations that capture the time behavior of the system under any policy. The proof then consists of a forward induction argument over times of all distinct events, establishing an ordering between the hypothesized optimal policy and the arbitrary policy. There might be more than one way to describe the evolution of the system considered, but as the authors point out not all of them make possible the comparisons between the optimal policy and an arbitrary policy. The technique was employed for the optimality of the μc rule for the multiclass $G/M/1$ queue and for reviewing the cases of the optimality of the *JSQ* (“join-the-shortest-queue”) rule in the control of routing to symmetric parallel queues.

As in forward induction, the backward induction compares the hypothesized optimal policy to an arbitrary policy. The proof requires more insight on the structural properties of the hypothesized optimal policy since in the induction step t , it is considered a policy that made the same decisions as the arbitrary one and then switches to the optimal one for the remaining decision stages. The technique was used for the optimality of the *Smallest Workload* rule for the multiclass $G/M/1$ queue with workload information.

The approach based on interchange arguments is generally the most popular and the most error prone. It is then essential to verify that the statistics of the input data have not been changed and that inadmissible control actions (see definition in Section 2 in [81]) are

Table 2. Structural Properties

$h(z)$	$g(x, y) =$	$g(x, y)$
Concave	$h(x + y)$	Jointly concave decreasing differences
Concave	$h(x - y)$	Jointly concave increasing differences
Convex	$h(x + y)$	Convex increasing differences
Convex	$h(x - y)$	Convex decreasing differences

not generated. This approach tries to show that the policy can be improved through the interchange of control decisions and/or the order of some events in the input sequence (typically, arrival times of customers, service times, routings). The comparison of policies is based on coupling [83], the same probability space [81], or a reduced new space followed by a proof of the optimality of the policy in the original model [84].

Among the papers that review optimal control on a sample path basis we refer also to [85] for scheduling problems and to [86] for communication networks. For a general overview see [87].

SUBMODULARITY AND CONVEXITY. GENERAL STRUCTURED OPTIMAL POLICIES

There are some qualitative analogies between submodular functions on lattices and convex functions on convex sets as emphasized by Table 1, giving ways of constructing submodular functions. Submodular functions are often considered as a discrete analog to convex functions [88].

Other structural properties relating the convexity/concavity with the monotone differences are presented in (Lemma 2.6.1 from [8], Section 2.6.2 Transformations, p. 49) Table 2. These were proven useful in the control of queues for obtaining a characterization of submodular types for the value functions (either direct [15], or indirect, in defining the operators in the event-based DP [5]).

Both submodularity and convexity are second-order properties in the sense that for twice differentiable functions on open

sets in $\mathbb{R}^n$ each property is equivalent to certain conditions on the matrix of second partial derivatives. The exact statement of Topkis's Theorem 1 holds for convex functions on convex sets, and this analogy was useful in finding certain properties of submodular functions. Among the closure (closed convex cone) properties, which are useful in propagating structural properties via DP there are

- the convex cone properties: expectation preserves convexity/concavity and monotonicity
- the single point properties: maximization preserves convexity and monotonicity as it does with the submodularity, while the maximum of jointly concave functions is concave.

In [23] the authors discuss in detail the convex cone properties and give a set of metatheorems that describe how properties of value functions are preserved and propagated through MDPs. Their main results say that the value functions satisfy property P among the convex cone ones closed under maximization or minimization if the reward functions satisfy property P and the transition probabilities satisfy a stochastic version of this property. For stationary finite horizon minimization MDP problems with real state and action space, monotone optimal policies can be characterized with convexity/concavity and submodularity, see Theorem 8.3 from Chapter 8 in [89], p. 127 as well as Theorem 3.10.2, Theorem 3.10.3 from Chapter 3 in [8], p. 172. For results in a more general context see [27].

In inventory management problems, the idea to defer orders until the inventory level was at a low enough level (s) so that the setup cost will be incurred when a large enough amount (at least $S - s$) is ordered had been appealing to researchers in the 1950s. However, its proof eluded them because the value function in the DP formulation of the problem was neither convex nor concave. Finally, [90] provided a proof by introducing a new class of functions called K -convex functions.

The easiest way of seeing it is that a K -convex function must lie below the line segment connecting $(y, h(y))$ and $(y', h(y') + K)$ on the interval $[y, y']$. Such a function can be discontinuous, but the jumps at the discontinuity point must necessarily be downwards and cannot be too large. Other seminal papers were [91,92]. This family of problems can be seen in the more general context of separable MDP as in Chapter 7 in [1], p. 307.

For general references on structured solutions for MDPs see the overall unified perspectives of [11]; Section 7.1 Structured Policies in [93], p. 147; Section 6.11, p. 255 and Section 8.11, p. 425 in [2]; Section 5.4 in [89], p. 87 (finite horizon); [94] (infinite).

For POMDPs with finite or countably infinite (core) state and observation spaces and finite action set see [95]. For countable stage decision processes see [96,97].

APPLICATIONS

This is not an extended overview, we just grouped some interesting monotonic policies according to the solution methodology. For one-dimensional problems, a monotone policy is often in the form of a control limit policy, while for two-dimensional problems, it is often in the form of a switching curve. On switching curves in the two-dimensional state space, the seminal paper is [66] discussing the optimal control of a Markov network with two service stations and linear cost. For switching curves in manufacturing we also refer to [70].

A general framework of obtaining monotone policies was the event DP (see the section titled “Continuous-Time Models”) established for the optimality of routing [80], admission control, service rate, general flow control policies that are monotone in the state, under different information structure. In [5] it was recently used to compare different systems differing in several operators, but for which the value function was the same.

The monotonicity in the case of a single queue has been studied in [9,38,50,98–102]. The monotonicity in the case of a network is treated in [67,71,80,103–105] (see also [73,74] for extensions) [106].

In corrective maintenance, a typical control limit policy is in [107] for an $M/G/1$ queueing system with an unreliable server. Classical monotonicity is derived in the corrective approach of [108] for parallel servers subject to random failure, or in the opportunistic maintenance approach of [109] for a k different types of components. For a review see [110–112].

In dynamic models, short-term information such as a varying deterioration of components or unexpected opportunities can be taken into account and decisions may change over the planning horizon. Thus, a specific type of monotonic optimal policy developed in maintenance is the *At-Most-Four-Region* (AM4R) policy introduced by [113] and further studied by [114] in the POMDP setting. A monotonic AM4R policy along an ordered set of knowledge states may change at most three times, and changes are following a pre-set sequence. It is supposed that the more deteriorated the system is, the more likely it is to deteriorate further and/or to fail. Also, if a knowledge state is stochastically less than another, then the system is “less deteriorated” in the “smaller” state, following the idea that no news is good news. The order considered was likelihood ratio ordering.

A more general type of policy called *component/marginally monotone* is found in [115,116]. Other maintenance papers with monotone optimal policies are [117–123].

For a monotonicity in binary and stochastic inventory models we refer to Chapter 8 in [89, p. 119].

In other application areas there exist direct monotonicity proofs using the general theory as in production and innovation [48,124], Chapter 5 in [48] pp. 188, [24,125–127] or staffing problems [120].

Consider an MDP problem having a monotone optimal policy in one or two dimensions, while for higher dimensions its optimal policies can only be found numerically. Then for higher dimensions, one might use heuristic methods leading to monotone policies. This is the case of a manufacturing application studied by [128,129]. For scheduling a single machine producing several different classes of items in a make-to-stock mode, the optimal policy is monotone for the case of

two products and backordering. For higher dimensions [130] derive heuristic index policies and hedging points (idleness policies). For the index policies, an index is computed for each class as a function of the inventory in that class. The class with the smallest index is produced, while if all indices are positive, the machine is idle. The concepts of monotone switching and monotone idling are used to describe the monotonic policies. This means that B_k , the set of states in which class k is preferred is bounded by an increasing switching surface and by a decreasing idling surface. For a review of such heuristics, see [131].

Acknowledgments

The author would like to thank Ad Ridder from Vrije Universiteit Amsterdam; Lisa Maillart, Jeffrey Kharoufeh from University of Pittsburgh; and the anonymous referees for their suggestions and insights, which improved this manuscript.

REFERENCES

1. Heyman SP, Sobel MJ. Volume II, Stochastic models in operations research. New York: McGraw-Hill; 1984.
2. Puterman M. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley & Sons, Inc.; 1994.
3. Conlisk J. Stability and monotonicity for interactive Markov chains. *J Math Sociol* 1992;17:127–143.
4. McCall J. The economics of search: exchangeability. New York: Routledge; 2008. ISBN:0415299926.
5. Koole G. Monotonicity in Markov reward and decision chains: theory and applications. Volume 1, Issue 1 Foundations and trends in stochastic systems. Hanover (MA): Now Publishers Inc.; 2006. pp. 1–76. ISSN:1551-3106.
6. Topkis DM. Minimizing a submodular function on a lattice. *Oper Res* 1978;26:305–321.
7. White CC. The optimality of isotone strategies for Markov decision problems with utility criterion. In: Hartley R, Thomas LC, White DJ, editors. Recent developments in Markov decision processes. New York: Academic Press; 1980. pp. 261–276. ISBN:0123284600.
8. Topkis DM. Supermodularity and complementarity, Frontiers of Economic Research series. Princeton (NJ): Princeton University Press; 1998. ISBN: 978-0-691-03244-3.
9. Serfozo RF. Monotone optimal policies for Markov decision processes. *Math Program Study* 1976;6:202–215.
10. Muller A and Stoyan D. Comparison methods for stochastic models and risks. New York: Wiley; 2002. ISBN: 0-471-49446-1.
11. Hinderer KF. On the structure of solutions of stochastic dynamic programs. Proceedings of the 7th Conference on Probability Theory Brasov; 1984; Romania. pp. 173–182.
12. Hinderer KF, Stieglitz M. On polychotomous search problems. *Eur J Oper Res* 1994;73(1):279–294.
13. Landsberg AS, Friedman EJ. Dynamical effects of partial orderings in physical systems. *Phys Rev E* 1996;54(4):3135–3141.
14. Lin W, Kumar PR. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans Automat Control* 1984;29:696–703.
15. Koole G. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Syst Control Lett* 1995;26:301–303.
16. Balachandran KR. Control policies for a single server system. *Manage Sci* 1973;19:1013–1018.
17. Puterman M. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley & Sons, Inc.; 2005.
18. Dilworth RP. A decomposition theorem for partially ordered sets. *Ann Math* 1950;51:161–166.
19. Mirsky L. A dual of Dilworth decomposition theorem. *Am Math Mon* 1971;78:876–877.
20. Sobel MJ. Myopic solutions of Markov decision processes and stochastic games. *Oper Res* 1981;29:995–1009.
21. Dragut AB. A weakly monotonic backward induction algorithm on finite bounded subsets of vector lattices. *J Comput Appl Math* 2004;164-165:295–306.
22. White DJ. Negatively isotone optimal policies for random walk type Markov decision processes. *OR Spectrum* 1982;4:41–45. ISSN: 0171-6468.
23. Smith JE, McCardle KF. Structural properties of stochastic dynamic programs. *Oper Res* 2002;50(5):796–809.

24. Athey S, Schmutzler A. Product and process flexibility in an innovative environment. *RAND J Econ* 1995;26:557–574.
25. Shaked M, Shanthikumar J. *Stochastic orders*. New York: Academic Press; 1994. ISBN-13: 978-0387329154.
26. Massey WA. Stochastic orderings for Markov processes on partially ordered spaces. *Math Oper Res* 1987;12(2):350–367. DOI: 10.1287/moor.12.2.350.
27. Müller A. How does the value function of a Markov decision process depend on the transition probabilities. *Math Oper Res* 1997;22:872–885.
28. Schäl M. Conditions for optimality in dynamic programming and for the limit of n -stage optimal policies to be optimal. *Z Wahrscheinlichkeitstheorie verw Gebiete* 1975;32:179–196.
29. Hernández-Lerma O, Runggaldier W. Monotone approximations for convex stochastic control problems. *J Math Syst Estimation Control* 1994;4:99–140.
30. Rieder U, Zagst R. Monotonicity and bounds for convex stochastic control models. *Methods Models Oper Res* 1994;39:187–207.
31. Dragut AB. A Markovian approach to the mathematical control Of NPD projects [PhD thesis]. The Netherlands: Technische Universiteit Eindhoven; 2002.
32. Flores-Hernandez R, Montes-deOca R. Monotonicity of minimizers in optimization problems with applications to MCPs. *Kybernetika* 2007;43:347–368.
33. Hernández-Lerma O, Lasserre JB. *Discrete-time Markov control processes*. Berlin, New York: Springer; 1996. ISBN: 0-387-94579-2.
34. Cruz-Suárez D, Montes-deOca R, Salem-Silva F. Conditions for the uniqueness of optimal policies of discounted Markov decision processes. *Math Methods Oper Res* 2004;60:415–436.
35. Kalin D. A note on monotone optimal policies for Markov decision processes. *Math Program* 1978;15:220–222.
36. Kalin D. Über Markoffsche Entscheidungsmodelle mit halbgeordnetem Zustandsraum. *OR-Verfahren* 1979;33:233–245.
37. Kalin D. Beiträge zj strukturierten markoffschen entscheidungsmodellen [PhD thesis]. Universität Bonn, Habilitationsschrift; 1981.
38. Serfozo R. Optimal control of random walks, birth and death processes and queues. *Adv Appl Probab* 1981;13:61–83.
39. Sobel MJ. Reservoir management models. *Water Resour Res* 1975;11:767–776.
40. White CC. Optimal control-limit strategies for a partially observed replacement problem. *Int J Syst Sci* 1979;10:321–331.
41. White DJ. Isotone policies for the value iteration method for Markov decision processes. *OR Spectrum* 1984;6:223–227.
42. White DJ. Discount-isotone policies for Markov decision processes. *OR Spectrum* 1988;10:13–22.
43. Tijms HC. An algorithm for average costs denumerable state semi-Markov decision problems with applicants to controlled production and queuing systems. In: Hartley R, Thomas LC, White DJ, editors. *Recent developments in Markov decision processes*. New York: Academic Press; 1980. pp. 143–179. ISBN: 0123284600.
44. White DJ. Isotone optimal policies for structured Markov decision processes. *Eur J Oper Res* 1981;7:392–402.
45. Taylor HM. Markovian sequential replacement processes. *Ann Math Stat* 1965;36:1677–1694.
46. Derman C. *Finite state Markovian decision processes*. New York: Academic; 1970.
47. Tijms HC. *Stochastic modelling and analysis, a computational approach*. New York: Wiley; 1986. ISBN: 0-471-90911-4.
48. Garcia A, Smith RL. Solving nonstationary infinite horizon stochastic production planning problems. *Oper Res Lett* 2000;27(3):135–141.
49. Cheevaprawatdomrong T, Schochetman IE, Smith RL, *et al.* Solution and forecast horizons for infinite horizon nonhomogeneous Markov decision processes. *Math Oper Res* 2007;32:51–72.
50. Altman E, Koole G. Control of a random walk with noisy delayed information. *Syst Control Lett* 1995;24:207–213.
51. Smallwood RD, Sondik EJ. The optimal control of partially observable Markov processes over a finite horizon. *Oper Res* 1973;21(5):1071–1088.
52. Sondik E. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs. *Oper Res* 1978;26(2):282–304. DOI: 10.1287/opre.26.2.282.
53. Lovejoy WS. Some monotonicity results for partially observed Markov decision processes. *Oper Res* 1987;35(5):736–743.

54. Albright C. Some monotonicity results for partially observed Markov decision processes. *Oper Res* 1979;27(5):1041–1953.
55. Rieder U. Structural results for partially observed control models. *Math Methods Oper Res* 1991;35:473–490.
56. White CC. Optimal inspection and repair of a production process subject to deterioration. *J Oper Res Soc* 1978;29(3):235–243.
57. White CC. Monotone control laws for noisy, countable-state Markov chains. *Eur J Oper Res* 1980;5:124–132.
58. White CC. A survey of solution techniques for the partially observed Markov decision process. *Ann Oper Res* 1991;32(1):215–230. DOI: 10.1007/BF02204836.
59. Feinberg EA, Shwartz A. *Handbook of Markov decision processes: methods and applications*. Norwell (MA): Kluwer Academic Publishers; 2002. ISBN: 978-0-7923-7459-6.
60. Altman E. *Constrained Markov decision processes*. Boca Raton (FL): Chapman and Hall/CRC Press; 1999. ISBN-13:978-0849303821.
61. Djonin D, Krishnamurthy V. *Q-learning algorithms for constrained Markov decision processes with randomized monotone policies: Application to mimo transmission control*. *IEEE Trans Signal Process* 2007;55(5):2170–2181. DOI: 10.1109/TSP.2007.893228.
62. Djonin D, Krishnamurthy V. MIMO transmission control in fading channels—a constrained Markov decision process formulation with monotone randomized policies. *IEEE Trans Signal Process* 2007;55(10):5069–5083. DOI: 10.1109/TSP.2007.897859.
63. Denekere R, Pelikan S. Competitive chaos. *J Econ Theory* 1986;40:13–25.
64. Friedman EJ, Johnson S. Dynamic monotonicity and comparative statics for real options. *J Econ Theory* 1997;75(1):104–121.
65. Serfozo RF. An equivalence between continuous and discrete time Markov decision processes. *Oper Res* 1979;27(3):616–620.
66. Hajek B. Optimal control of two interacting service stations. *IEEE Trans Automat Control* 1984;29(6):491–499. ISSN: 0018-9286.
67. Weber RR, Stidham S Jr. Control of service rates in networks of queues. *Adv Appl Probab* 1987;24:202–242.
68. Altman E, Gaujal B, Hordijk A. Volume 1829, *Discrete-event control of stochastic networks: multimodularity and regularity*, *Lecture Notes in Mathematics*. Berlin: Springer; 2003. ISBN: 978-3-540-20358-2.
69. Bartroli M, Stidham SJ. Towards a unified theory of structure of optimal policies for control of network of queues. Technical report No. UNC/OR/TR87-5. Chapel Hill (NC): Department of Operations Research, University of North Carolina; 1987.
70. Veatch MH, Wein LM. Monotone control of queueing networks. *Queueing Syst Theory Appl* 1992;12(3-4):391–408. ISSN: 0257-0130.
71. Glasserman P, Yao DD. Monotone optimal control of permutable GSMPs. *Math Oper Res* 1994;19:449–476.
72. Koole G. *Stochastic scheduling and dynamic programming*. Technical Report 113. Amsterdam, The Netherlands: Centrum Wiskunde and Informatica (CWI); 1995.
73. Koole GM. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Syst* 1998;30(3-4):323–339. DOI: 10.1023/A:1019177307418.
74. Altman E, Koole G. On submodular value functions and complex dynamic programming. *Stoch Models* 1998;14(5):1051–1072.
75. Chang CS, Shanthikumar JG, Yao DD. *Stochastic modeling and analysis of manufacturing systems*. *Stochastic convexity and stochastic majorization*. Berlin: Springer; 1994. pp. 188–231. ISBN:978-0-387-94319-0.
76. Shaked M, Shanthikumar J. *Stochastic convexity and its applications*. *Adv Appl Probab* 1988;20:427–446.
77. Nain P. Qualitative properties of the Erlang blocking model with heterogeneous user requirements. *Queueing Syst* 1990;6(1):189–206. DOI: 10.1007/BF02411473.
78. Ross K, Yao D. Monotonicity properties for the stochastic knapsack. *IEEE Trans Inf Theory* 1990;36(5):1173–1179. DOI: 10.1109/18.57223.
79. Liyanage L, Shanthikumar J. Second-order properties of single-stage queueing systems. In: Bhat BR, Basawa IV, editors. *Queueing and related models*. Oxford, UK: Oxford University Press; 1992. pp. 129–160.

80. Altman E, Stidham S Jr. Optimality of monotone policies for two-action Markovian decision processes, with applications to control of queues with delayed information. *Queueing Syst* 1995;21:267–291.
81. Liu Z, Nain P, Towsley D. Sample Path Methods in the Control of Queues. *Queueing Syst* 1995;21:293–335.
82. Nain P, Towsley D. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Trans Automat Control* 1994;39:1853–1855.
83. Shields P. The ergodic theory of discrete sample paths. Volume 13, Graduate studies in mathematics. Providence (RI): American Mathematical Society; 1996. p. 218.
84. Ephremides A, Varaiya P, Walrand J. A simple dynamic routing problem. *IEEE Trans Automat Control* 1980;25:690–693.
85. Green TC, Stidham S. Sample-path conservation laws, with applications to scheduling queues and fluid systems. *Queueing Syst Theory Appl* 2000;36(1/3):175–199.
86. Altman E. Applications of Markov decision processes in communication networks. In: Feinberg E, Shwartz A, editors. *Handbook of Markov decision processes methods and applications*. Boston (MA): Kluwer; 2002.
87. El-Taha M, Stidham S Jr. Volume 11, Sample-path analysis of queueing systems, International series in operations research and management science. Boston (MA): Springer; 1998.
88. Lovasz L. Submodular functions and convexity. In: Bachem AB, Grötschel M, Korte B, editors. *Mathematical programming: the state of the art (Bonn, 1982)*. Berlin: Springer; 1983. pp. 235–257.
89. Porteus EL. *Foundations of stochastic inventory theory*. Stanford: Stanford University Press; 2002. ISBN: 9780804743990.
90. Scarf HE. The optimality of (s, s) policies in the dynamic inventory problem. In: Sheshinski E, Weiss Y, editors. *Optimal pricing, inflation, and the cost of price adjustment*. 1993. MIT Press initially in mathematical methods in the social sciences. Arrow K, Karlin S, Suppes P, editors. Cambridge (MA): Stanford University Press; 1959.
91. Veinott AF. On the optimality of (s, s) policies: new conditions and a new proof. *SIAM J Appl Math* 1966;14:525–552.
92. Schal M. On the optimality of (s, s) policies in dynamic inventory models with finite horizon. *SIAM J Appl Math* 1976;30:528–537.
93. White DJ. *Markov decision processes*. New York: John Wiley & Sons, Inc.; 1993. ISBN: 9780471936275.
94. Porteus EL. Conditions for characterizing the structure of optimal strategies in infinite-horizon dynamic programs. *J Optim Theory Appl* 1982;36:419–432. DOI: 10.1007/BF00934355.
95. Fernandez-Gaucherand E, Arapostathis A, Marcus S. On the average cost optimality equation and the structure of optimal policies for partially observable Markov processes. *Ann Oper Res* 1991;29:d471–512.
96. Porteus EL. On the optimality of structured policies in countable stage decision processes. *Manage Sci* 1975;22(2):148–157. DOI: 10.1287/mnsc.22.2.148.
97. Porteus EL, Kreps D. On the optimality of structured policies in countable stage decision processes II: positive and negative problems. *SIAM J Appl Math* 1977;32(2):457–466.
98. Altman E, Gaujal B, Hordijk A. *Discrete-event control of stochastic networks: modularity and regularity*. Berlin, New York: Springer; 2008.
99. Altman E, Nain P. Closed-loop control with delayed information. *Perform Eval Rev* 1992;20:193–204.
100. Altman E, Stidham S Jr. Optimality of monotonic policies for two-action Markovian decision processes, including information and action delays. *Queueing Syst* 1996;12(2):307–328.
101. Stidham S Jr, Weber RR. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Queueing Syst* 1993;37:611–625.
102. Stidham S Jr, Weber RR. A survey of Markov decision models for control of networks of queues. *Queueing Syst* 1993;13:291–314.
103. Naor P. On the regulation of queueing size by levying tolls. *Econometrica* 1969;37:15–24.
104. Altman E, Konstantopoulos P, Liu Z. Stability, monotonicity and invariant quantities in general polling systems. *Queueing Syst* 1992;11:35–57.
105. Glasserman P, Yao DD. *Monotone structure in discrete event systems*. New York: John Wiley and Sons; 1994.
106. Rykov VV. Monotone control of queueing systems with heterogeneous servers. *Queueing Syst Theory Appl Arch* 2001;37(4):391–403. ISSN: 0257-0130.

107. Federgruen A, So KC. Optimal maintenance policies for single-server queueing systems subject to breakdowns. *Oper Res* 1990;38:330–343.
108. Jansen J, Van Der Duyn Schouten FA. Maintenance optimization on parallel production units. *J Math Appl Bus* 1995;6:113–134.
109. Zheng X, Fard N. A maintenance policy for repairable systems based on opportunistic failure rate tolerance. *IEEE Trans Reliab* 1991;40:237–244.
110. Dekker R, Wildeman RE, Van Der Duyn Schouten FA. A review of multi-component maintenance models with economic dependence. *Math Methods Oper Res* 1997;45:411–435. DOI: 10.1007/BF01194788.
111. Laggoune R, Chateauneuf A, Aissani D. Opportunistic policy for optimal preventive maintenance of a multi-component system in continuous operating units. *Comput Chem Eng* 2009;33(9):1499–1510. DOI: 10.1016/j.compchemeng.2009.03.003.
112. Valdez-Flores C, Feldman RM. A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Nav Res Log* 1989;36(4):419–446. DOI: 10.1002/1520-6750(198908)36:4.
113. Ross SM. Quality control under Markovian deterioration. *Manage Sci* 1971;19(9):587–596.
114. Maillart LM. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Trans* 2006;38:463–475.
115. Ozekici S. Optimal periodic replacement of multi-component reliability systems. *Oper Res* 1988;36:542–552.
116. Ivy J, Pollock S. Marginally monotonic maintenance policies for a multi-state deteriorating machine with probabilistic monitoring, and silent failures. *IEEE Trans Reliab* 2005;54(3):489–497. DOI: 10.1109/TR.2005.853443.
117. Anderson MQ. Monotone optimal preventive maintenance policies for stochastically failing equipment. *Nav Res Log* 1981;28(3):347–358. DOI: 10.1002/nav.3800280301.
118. Yao DD, Zheng S. Coordinated quality control in a two-stage system. *IEEE Trans Automat Control* 1999;44(6):1166–1179.
119. Banjevic D, Jardine AKS, Makis V, *et al.* A control-limit policy and software for condition-based maintenance optimization. *INFOR* 2001;39(1):32–50.
120. Fu MC, Marcus SI, Wang I. Monotone optimal policies for a transient queueing staffing problem. *Oper Res* 2000;48(2):327–331.
121. Pavitsos A, Kyriakidis EG. Markov decision models for the optimal maintenance of a production unit with an upstream buffer. *Comput Oper Res* 2009;36(6):1993–2006.
122. Pierskalla WP, Voelker JA. A survey of maintenance models: the control and surveillance of deteriorating systems. *Nav Res Log* 1976;23(3):353–388.
123. Chen M, Feldman RM. Optimal replacement policies with minimal repair and age-dependent costs. *Eur J Oper Res* 1997;98(1):75–84.
124. Hopenhayn H, Prescott E. Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica* 1992;60:1307–1406.
125. Athey S. Monotone comparative statistics under uncertainty. *Q J Econ* 2002;CXVII(1):187–223.
126. Dragut AB. Monotonic robust optimal control policies for the time-quality trade-offs in concurrent new product development (NPD). *Comput Math Appl* 2006;51(2):229–238.
127. Papadakia K, Powell WB. Monotonicity in multidimensional Markov decision processes for the batch dispatch problem. *Oper Res Lett* 2007;35(2):267–272.
128. Ha AY. Optimal dynamic scheduling policy for a make-to-stock production system, Working paper series B 124. Technical report School of Organization and Management, Yale University New London (CT); 1993.
129. Ha AY. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper Res* 1997;45(1):42–53.
130. Veatch MH, Wein LM. Scheduling a make-to-stock queue: index policies and hedging points. *Oper Res* 1996;44:634–647.
131. Bertsimas D. The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Syst* 1995;21:337–389.

SUBGRADIENT OPTIMIZATION

YURI NESTEROV
Center for Operations Research and
Econometrics (CORE), Université
Catholique de Louvain (UCL),
Louvain-la-Neuve, Belgium

GENERAL NONSMOOTH OPTIMIZATION PROBLEM

General nonsmooth optimization problem has the following form:

$$\min_{x \in R^n} \{f_0(x) : f_i(x) \leq 0, i = 1, \dots, m\}, \quad (1)$$

where some of functional components $f_i(x)$ are nondifferentiable. The absence of differentiability creates serious theoretical difficulties on different levels (optimality conditions, definition, and computation of search directions, etc.).

Let us start from the definition of differential characteristics of nondifferentiable functions. Its most popular version was suggested by Clarke [1] (for alternative approaches, see Refs 2–4).

Definition 1. Let f be locally Lipschitz continuous at point $x \in R^n$. The Clarke *directional derivative* of f along $v \in R^n$ is defined as

$$\hat{f}(x, v) = \limsup_{\substack{x' \rightarrow x \\ t \rightarrow +0}} \frac{1}{t} [f(x' + tv) - f(x')]$$

The Clarke *subdifferential* of f at x is the set

$$\partial_C f(x) = \{g \in R^n : \langle g, v \rangle \leq \hat{f}(x, v) \forall v \in R^n\}.$$

Then, the Clarke *subgradient* of f at x is an arbitrary element from $\partial_C f(x)$.

If f is differentiable at x , then $\partial_C f(x)$ consists of a single element, the gradient $\nabla f(x)$.

Clarke subdifferential appears to be very convenient for theoretical analysis of nonsmooth optimization problems (see Ref. 4 for recent extensions and discussions). For example, it is possible to write down the first-order necessary optimality conditions for problem (1) in the following natural form:

$$\begin{aligned} g_0 + \sum_{i=1}^m \lambda_i^* g_i &= 0, \\ g_i &\in \partial_C f_i(x^*), \quad i = 0, \dots, m, \\ \lambda_i^* g_i(x^*) &= 0, \quad \lambda_i^* \geq 0, \quad i = 1, \dots, m. \end{aligned} \quad (2)$$

This is the *generalized Karush–Kuhn–Tucker condition*. If $m = 0$, this condition becomes

$$0 \in \partial_C f(x^*). \quad (3)$$

However, from the viewpoint of numerical methods, the situation is not so good. First of all, Clarke subdifferential is not supported by a proper calculus, which should allow to arrange the computation of its elements by a sequence of operations used for computing the function's value. Let us look at the following example.

Example 1. Define $f_1(x) = \max\{x, 0\}$, $f_2(x) = \min\{x, 0\}$. Then

$$\begin{aligned} \hat{f}_1(0, v) &= f_1(v), \quad \partial_C f_1(0) = [0, 1], \\ \hat{f}_2(0, v) &= f_2(v), \quad \partial_C f_2(0) = [0, 1]. \end{aligned}$$

However, $f(x) = f_1(x) + f_2(x) \equiv x$, and there is no simple way for getting the correct answer $\partial_C f(0) = \{1\}$ using some operations with the sets $\partial_C f_1(0)$ and $\partial_C f_2(0)$. At least the natural receipt $\partial_C f_1(0) + \partial_C f_2(0)$ does not work.

Fortunately, for the majority of nonsmooth optimization methods, at each test point we need to compute only a single element from Clarke subdifferential. This can be done by

the technique of *lexicographical differentiation* [5]. Define the directional derivative

$$f'(x, h) = \lim_{\alpha \rightarrow +0} \frac{1}{\alpha} [f(x + \alpha h) - f(x)].$$

Note that it transforms a function of x into a function of h . Let us introduce the corresponding linear operator

$$\phi = H_x[f] \Leftrightarrow \phi(h) = f'(x, h).$$

For defining a lexicographic gradient of function f at point x , we need to fix a basis $U = (u_1, \dots, u_n)$ in R^n . Then we can define the sequence of functions

$$\begin{aligned} f_{x,U}^{(0)} &= H_x[f], \\ f_{x,U}^{(k)} &= H_{u_k}[f_{x,U}^{(k-1)}], \quad k = 1, \dots, n. \end{aligned}$$

It can be proved that $f_{x,U}^{(n)}$ is a linear function. Its gradient is called the *lexicographic gradient* of function f at x along the basis U .

For lexicographic gradients we have a perfect Chain Rule. Hence, these gradients can be computed even by automatic differentiation. Lexicographic gradients are some special elements of Clarke subdifferential. At the same time, they satisfy all natural properties of the usual gradient.

The majority of nonsmooth optimization methods are justified by the proof of convergence to Clarke stationary points defined by Equation (3). Unfortunately, this result is much weaker than the similar statement for the smooth optimization problems. Indeed, consider the unconstrained minimization problem (1) with $m = 0$ and the following objective function:

$$\begin{aligned} f_0(x) &= \frac{1}{4} |1 - x^{(1)}| + \sum_{i=1}^{n-1} |1 + x^{(i+1)} - 2|x^{(i)}||, \\ x^* &= (1, \dots, 1). \end{aligned} \quad (4)$$

This is a unimodular function with no saddle points. For each point $x \neq x^*$ there exists a direction of strict linear decrease of the function value. Unfortunately, this function has exponentially many Clarke stationary points. For $n \geq 4$ and the starting point $x_0 =$

$(-1, 1, \dots, 1)$, this problem remains very challenging for the modern methods of general nonsmooth minimization.

Among the existing methods for approximating Clarke stationary points, we should mention the *bundle* methods [6,7], and the *gradient sampling* methods [8]. The former methods are trying to decrease the value of the objective function by updating a piecewise linear model constructed around the current prox-center (see Ref. 9 for detailed discussion). The model is dynamically improved by incorporating new elements computed at the minimizers of the current version of the model. If a significant decrease of the objective function is achieved, the prox-center is moved to a new location. This idea admits a full theoretical justification for convex case (we discuss it in the next section). In non-convex situation, the worst-case complexity guarantees for this approach are necessarily quite weak. Let us look at the following example.

Example 2. Let vector $c \in R^n$ have positive integer coefficients. Consider the function

$$\begin{aligned} \phi_c(x) &= \left(1 - \frac{1}{\sum_{i=1}^n c^{(i)}} \right) \max_{1 \leq i \leq n} |x^{(i)}| \\ &\quad - \min_{1 \leq i \leq n} |x^{(i)}| + |\langle c, x \rangle|. \end{aligned}$$

It appears that the problem of finding $x \in R^n$ with $\phi_c(x) < 0$, or proving that this is impossible, is NP-hard. Thus, for a bundle method started at $x = 0$, we can hardly hope that it can find a direction of strict decrease of ϕ_c . This direction corresponds to the integer solution of equation $\langle c, x \rangle = 0$ with entries ± 1 .

The gradient sampling methods select the test points randomly around the current center. For the majority of practical applications, the objective function at these points is differentiable with probability 1. Hence, the gradients at these points can be used for approximating the Clarke subdifferential at

the center. Unfortunately, the random scanning in a high-dimensional space is very expensive. Consider the following example.

Example 3. Let all coefficients of the vector $\sigma \in R^n$ be ± 1 . Consider the following function:

$$f_\sigma(x) = \max_{1 \leq i \leq n} |x^{(i)}| - 2 \max \left\{ 0, \min_{1 \leq i \leq n} \sigma^{(i)} x^{(i)} \right\}.$$

This function can be used in the *resisting oracle* fighting against the gradient sampling technique. As far as the number of calls of this oracle is smaller than $2^n - 1$, it can continue returning the differential characteristics of the first part of f_σ reallocating σ to be one of the missing sign patterns among all test points. Thus, in order to discover the full picture, we need at least 2^n calls. Note that along the direction σ function f_σ has linear decrease.

CONVEX PROBLEMS

Consider now the *convex* optimization problem:

$$\min_{x \in Q} f(x), \quad (5)$$

where $Q \subseteq R^n$ is a closed convex set, and function f is closed and convex on Q . Then, at each point $x \in Q$ there exists a subdifferential $\partial f(x)$:

$$f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall g_x \in \partial f(x), \quad \forall y \in Q. \quad (6)$$

Thus, the necessary and sufficient optimality condition at point $x^* \in Q$ can be written in the following form:

$$\exists g^* \in \partial f(x^*) : \langle g^*, x - x^* \rangle \geq 0 \quad \forall x \in Q.$$

The majority of practical methods for solving Equation (5) requires computation of a single *subgradient* $g_x \in \partial f(x)$ at each test point $x \in Q$. Since for nonsmooth functions

it is difficult to ensure a monotone decrease of the objective, the computational strategies of *subgradient methods* are based on the following inequality:

$$\langle g_x, x - x^* \rangle \geq 0, \quad x \in Q. \quad (7)$$

The earliest subgradient schemes employed the fact that by Equation (7) the *Euclidean distance* between the current test point and the optimal one can be decreased by an appropriate move along the direction of *antigradient* $-g_x$. The simplest *primal* subgradient method [10,11] looks as follows:

$$x_0 \in Q, \quad x_{k+1} = \pi_Q(x_k - h_k g_{x_k}), \quad k \geq 0, \quad (8)$$

where $\pi_Q(x)$ denotes the Euclidean projection of point x onto the set Q . The appropriate choice of the step sizes are as follows:

$$h_k = \frac{1}{\|g_{x_k}\|} \cdot \frac{R}{\sqrt{k+1}}, \quad h_k = \frac{f(x_k) - f(x^*)}{\|g_{x_k}\|^2},$$

where R is an estimate for $\|x_0 - x^*\|$. In both cases, it is possible to prove the following *rate of convergence*:

$$\min_{0 \leq i \leq k} f(x_i) - f(x^*) \leq \frac{c \cdot L R}{\sqrt{k+1}}, \quad (9)$$

where $c > 0$ is an absolute constant and L is a uniform upper bound for $\|g_x\|$, $x \in Q$. It can be shown that for this class of optimization problems the rate (Eq. 9) cannot be improved uniformly in the dimension of the space of variables [12].

The *dual* subgradient methods are based on the *linear model* of the objective function [13]. Let us choose $\lambda_i \geq 0$, $0 \leq i \leq k$, with $\sum_{i=0}^k \lambda_i = 1$. Then for all $x \in Q$ we have

$$l_k(x) \stackrel{\text{def}}{=} \sum_{i=0}^k \lambda_i [f(x_i) + \langle g_{x_i}, x - x_i \rangle] \stackrel{(6)}{\leq} f(x).$$

One of the variants of the dual method, the *simple averaging scheme*, looks as follows:

$$x_{k+1} = \arg \min_{x \in Q} \left\{ \frac{1}{k+1} \sum_{i=0}^k [f(x_i) + \langle g_{x_i}, x - x_i \rangle] + \frac{L \cdot \|x - x_0\|^2}{2R\sqrt{k+1}} \right\}.$$

In this method, it is possible to approximate simultaneously the optimal solutions of the primal and the dual problems. The earliest versions of the dual methods are also known as *mirror descent schemes* [12]. They allow measurement of distances in non-Euclidean norms. This possibility is especially useful for Q having a simplex structure. All these methods ensure the optimal rate of convergence (Eq. 9).

Another big group of subgradient methods is based on a *piece-wise linear model*. They are called *bundle methods*. Define

$$L_k(x) \stackrel{\text{def}}{=} \max_{0 \leq i \leq k} [f(x_i) + \langle g_{x_i}, x - x_i \rangle] \stackrel{(6)}{\leq} f(x).$$

This function contains *all* information on the objective function accumulated after k iterations. The most straightforward way of using it is implemented in the *Kelley method*

$$x_{k+1} \in \underset{x \in Q}{\text{Argmin}} L_k(x). \quad (10)$$

Unfortunately, this natural strategy has exponentially bad lower complexity bound [14, p. 158]. A more reserved behavior is implemented in the *classical bundle method* [15]. In this scheme, starting from a prox-center x_0 , we construct the following sequence:

$$x_{k+1} = \arg \min_{x \in Q} \{L_k(x) + \frac{1}{2} \gamma_k \|x - x_0\|^2\}, \quad (11)$$

where γ_k are certain proximity parameters. This process is stopped if $f(x_k)$ is sufficiently smaller than $f(x_0)$. In this case, we choose x_k as a new prox-center and start the process again.

For the majority of bundle methods it is possible to prove convergence to the optimal solution of problem (5). However, to the best of our knowledge, no complexity estimates of the classical bundle method are published yet. There exists only one bundle-type method with justified complexity bounds. It

is called the *level method* [16]:

$$\begin{aligned} \lambda_k &= \frac{1}{2} \min_{0 \leq i \leq k} f(x_i) + \frac{1}{2} \min_{x \in Q} L_k(x), \\ Q_k &= \{x \in Q : L_k(x) \leq \lambda_k\}, \\ x_{k+1} &= \pi_{Q_k}(x_k). \end{aligned} \quad (12)$$

Each iteration of this method is quite expensive. However, usually it demonstrates very fast convergence. In the worst-case situation, it has the optimal rate of convergence (Eq. 9).

The last big group of subgradient methods consists of the *cutting plane schemes*. These methods use inequality Equation (7) as a *cutting plane* that divides the whole space on two parts, only one of them containing x^* . Consequently, they are interested in decreasing the “size” of *localization sets*:

$$\mathcal{L}_k = \{x \in Q : \langle g_{x_i}, x_i - x \rangle \geq 0, i = 0, \dots, k\}.$$

For that, we need to put new cutting plane in a properly defined central point of current localization set.

The first straightforward implementation of this idea was done in the *center-of-gravity method* [17]:

$$x_{k+1} = \mathcal{CQ}(\mathcal{L}_k). \quad (13)$$

In view of very heavy computational cost of each iteration, this method is interesting only from theoretical point of view. It was the first nonsmooth optimization method with *linear* rate of convergence:

$$\min_{0 \leq i \leq k} f(x_i) - f(x^*) \leq \left(1 - \frac{1}{e}\right)^{k/n} LR. \quad (14)$$

For a finite dimension, this rate is optimal [12]. The implementable version of this method is given by the famous *Ellipsoid Method*. Assume that $Q = \{x : \|x - x_0\| \leq R\}$ and choose $H_0 = I$, where I is $n \times n$ identity matrix. Then this method works as follows:

$$\begin{aligned} x_{k+1} &= x_k - \frac{R}{n+1} \cdot \frac{H_k g_{x_k}}{\langle H_k g_{x_k}, g_{x_k} \rangle^{1/2}}, \\ H_{k+1} &= \frac{n^2}{n^2 - 1} \left(H_k - \frac{2}{n+1} \frac{H_k g_{x_k} g_{x_k}^T H_k}{\langle H_k g_{x_k}, g_{x_k} \rangle} \right). \end{aligned} \quad (15)$$

This method was invented by Yudin and Nemirovski [18] and rediscovered later by Shor [19]. Method (15) was used by Khachiyan [20] in his famous proof of polynomial-time solvability of linear programming.

At the current moment there exists several cutting plane methods based on different notions of central points of convex sets. Among the most popular schemes are method of *inscribed ellipsoids* [21], *analytic center* method [22,23], and *volumetric center* method [24,25]. For all of them it is possible to prove suboptimal rate of convergence. However, the complexity of each iteration of these schemes is much higher than that of the Ellipsoid method.

REFERENCES

1. Clarke F. Optimization and nonsmooth analysis. New York: Wiley; 1983.
2. Demyanov V, Rubinov V. On quasi-differentiable functionals. *Sov Math Dokl* 1980;250(1):21–25.
3. Ioffe A, Tikhomirov V. Theory of extremal problems. Amsterdam: North-Holland; 1979.
4. Mordukhovich B. Variational analysis and generalized differentiation. Berlin: Springer; 2006.
5. Nesterov Y. Lexicographic differentiation of nonsmooth functions. *Math Prog* 2005;104(2-3):669–700.
6. Lemarechal C. An extension of Davidon methods to non-differentiable problems. In: Balinski M, Wolfe P, editors. Volume 3, Nondifferentiable optimization: mathematical programming study. Amsterdam: North-Holland; 1975. pp. 95–109.
7. Wolfe P. A method of conjugate subgradients for minimizing nondifferentiable functions. In: Balinski M, Wolfe P, editors. Volume 3, Nondifferentiable optimization: mathematical programming study. Amsterdam: North-Holland; 1975. pp. 145–173.
8. Burke J, Lewis A, Overton M. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J Optim* 2005;15(3):751–779.
9. Kiwiel K. Methods of descent for nondifferentiable optimization, Lecture Notes in Mathematics 1133. Berlin: Springer; 1985.
10. Shor N. Minimization methods for nondifferentiable functions. Berlin: Springer; 1985.
11. Polyak B. A general method for solving extremum problems. *Sov Math Dokl* 1967;8:593–597.
12. Nemirovsky A, Yudin D. Problems complexity and method efficiency in Optimization. New York: Wiley-Interscience; 1983.
13. Nesterov Y. Primal-dual subgradient methods for convex problems. *Math Program* 2009;120(1):261–283.
14. Nesterov Y. Introductory lectures on convex optimization. Boston (MA): Kluwer; 2004.
15. Hiriart-Urruty JB, Lemarechal C. Convex analysis and minimization algorithms. Berlin: Springer; 1993.
16. Lemarechal C, Nemirovskii A, Nesterov Y. New variants of bundle methods. *Math Program* 1995;69:111–147.
17. Levin A. On an algorithm for minimizing a convex function. *Sov Math Dokl* 1965;6:286–290.
18. Yudin D, Nemirovsky A. Informational complexity and efficient methods for solving convex extremal problems. *Ekonom i Mat Metody* 1976;12:357–369.
19. Shor N. A cutting plane method for solving convex programming problem. *Cybernetics* 1977;1:42–50.
20. Khachiyan L. A polynomial algorithm in linear programming. *Sov Math Dokl* 1979;20(1):191–194.
21. Tarasov S, Khachiyan L, Erlich I. The method of inscribed ellipsoids. *Sov Math Dokl* 1988;37:226–230.
22. Zonnevend G. New algorithm in convex programming based on a notion of “centre” (for a system of analytic inequalities) and on rational extrapolation. In: Hofman K, *et al.*, editors. Trends in mathematical optimization. Basel: Birkhauser; 1988. pp. 311–326.
23. Atkinson D, Vaidya P. A cutting plane algorithm for convex programming that uses analytic centers. *Math Program* 1995;69:1–43.
24. Vaidya P. A new algorithm for minimizing convex functions over convex sets. *Math Program* 1996;73:291–341.
25. Anstreicher K. Towards a practical volumetric cutting plane method for convex programming. *SIAM J Optim* 1998;9(1):190–206.

SUBJECTIVE PROBABILITY

NABIL I. AL-NAJJAR

LUCIANO DE CASTRO

Department of Managerial Economics and
Decision Sciences, Kellogg School of
Management, Northwestern University,
Evanston, Illinois

INTRODUCTION

What is probability? Observers of scientific progress in the last few decades will likely find this question puzzling. Modern probability theory is, by all objective measures, a runaway success in shaping modern science. In management sciences, entire fields such as finance, economics, and operations research are in part founded on probabilistic concepts and tools. Yet it is hard to think of other concepts as important as probability whose very meaning remains unclear and, often, controversial.

The formative years of the modern theory of probability, roughly from the 1920s through the 1950s, also witnessed lively debates about its nature and interpretation.¹ The arguments revolved around issues like: Is probability an objective feature of the phenomena under study, or merely a subjective judgment of the decision maker? How is probability related to frequency? If probability is an objective feature of reality, like heat or magnetism, then what scientific experiment could be devised to prove its existence and ascertain its value? If it is, on

the other hand, a decision maker's purely subjective state of mind, then is there a way to judge its reasonableness or consistency with empirical evidence?

Classic works by Doob [15] and Savage [9] sidestepped these philosophical issues by providing elegant mathematical formalisms of the concept of probability and related constructs. The phenomenal growth of modern probability theory and applications owes much to these works, which freed researchers from being bogged down with the hard conceptual issues of an earlier generation.

But, setting foundational questions aside neither implies that these questions have been answered nor that their practical and conceptual implications magically disappear. In fact, we would argue that it is in the management sciences, be it competitive strategy, finance, economics, or game theory, that these foundational questions about the meaning and interpretation of probability have potentially the greatest significance.

To illustrate, consider some elementary issues in the study of competitive strategy:

- *Overoptimism:* Do decision makers (firms, managers, investors) tend to be overoptimistic? If probability judgments are purely subjective, then in what sense could they be wrong, or overoptimistic?
- *Objective versus Subjective Probability:* Relatedly, are some probability judgments more "objective" than others? Is there a sense in which decision makers can separate the objective from the subjective parts of their judgments?
- *Risk versus Uncertainty:* Should decision makers approach one-of-a-kind, highly uncertain decisions like betting on the success of a new disruptive technology in the same way as they approach routine, well-understood risks? More broadly, can one give formal meaning to Knight's [3] distinction between risk (roughly, events with known

¹The classic works include Keynes [1], Borel [2], Knight [3], Ramsey [4], de Finetti [5,6], von Mises [7], Reichenbach [8], and Savage [9]. Galavotti [10] provides a comprehensive overview of the subject; see also Bernstein [11] for a popular account of the notions of risk, uncertainty, and probability. For the early subjectivist views of Ramsey and de Finetti, see Zabell [12], Galavotti [13], and Galavotti [14].

odds) and uncertainty (unknown, or unknowable odds)?

- *Belief Formation, Learning, and Testing*: How should firms translate the past experience into probability judgments to use in future decisions? And, is there a meaningful way to test these judgments against new evidence?

This is a sample of the questions that surface, in different guises, in every major branch of the management sciences. A major task of a formal decision theoretic framework is to provide a systematic way to address (or, as the case may be, dismiss) such questions.

The centerpiece of our survey is Savage's [9] theory of subjective probability, which remains the foundation of the subject till date. We provide an account of its main assumptions and conclusions. Our emphasis is on the interpretation of the axioms, important attempts to extend the theory, possible critiques, and potential limitations.

EXPECTED UTILITY THEORY

von Neumann–Morgenstern Representation

Before discussing Savage's framework,² it is useful to take a detour into von Neumann and Morgenstern's [19] classic representation theory when probabilities are objectively given. In their model, the primitives are a set of consequences: $C = \{\dots, x, y, z, \dots\}$, for example, monetary outcomes, and the set of probability distributions $\mathcal{P}$, or lotteries, on C with finite support. An individual has a preference $\succsim$ over $\mathcal{P}$ that is reflexive, complete, transitive, and (suitably) continuous.³ The crucial ingredient in von Neumann and Morgenstern's theory is the *independence axiom*:

for all lotteries p, q , and z and real number $\alpha \in (0, 1]$

$$p \succsim q \iff \alpha p + (1 - \alpha)z \succsim \alpha q + (1 - \alpha)z.$$

This is an additivity property of the preference: if two compound lotteries, like those on the RHS of the equivalence above, share a common component $(1 - \alpha)z$, then this component can be removed without affecting the ranking of the remaining components p and q .

von Neumann and Morgenstern show that a preference satisfies the axioms if and only if there is a utility function $u : C \rightarrow \mathbb{R}$, unique up to positive affine transformation, such that⁴

$$p \succsim q \iff \int_C u(c) dp(c) \geq \int_C u(c) dq(c). \quad (1)$$

In words, the decision maker ranks lotteries by applying the expected utility criterion with respect to the von Neumann and Morgenstern utility function u . When C is a convex set of real numbers, the utility function u will embed the decision maker's risk attitude.

von Neumann and Morgenstern work laid the foundation for the expected utility criterion, which is central to all subsequent developments in decision theory. A major drawback, however, is that a decision maker in their model is somehow presented with probability distributions to choose from. In most cases of interest, such as in games of strategy or business plans with even moderate degree of realism and complexity, decision makers are not offered the opportunity to choose between gambles with objectively known probabilities. In such situations, the question is: Under what conditions would decision makers behave as if they had "mental models" or "representations" of their environment that take the form of a probabilistic belief to use in applying the expected utility criterion? We turn to this next.

²In addition to Savage's [9] classic work, Fishburn [16] provides a textbook account while Kreps [17] is an excellent, very readable introduction to the subject. Much of our terminology and notation follows Machina and Schmeidler [18].

³This is known as the *Archimedean axiom*. See the references above for precise statement.

⁴Given our assumption that $\mathcal{P}$ consists of distributions with finite support; the integral here is, in fact, a sum. We use the integral notion to emphasize symmetry with similar expressions in other representation theorems.

Savage's Framework

The key ingredients in Savage's theory are

- States: $\Omega = \{\dots, \omega, \dots\}$.

In principle, a state of the world ω is a complete specification of every conceivable aspect of the decision problem at hand. In Savage's words, a state is "a description of the world leaving no relevant aspect undescribed." While this may be a useful conceptualization of states in a foundational work, Savage is, of course, aware of the need for a more parsimonious notion of state in practical problems.⁵

- Events: $\mathcal{E} = 2^\Omega = \{\dots, A, B, E, \dots\}$.

An event E is a set of states. Events will usually refer to the information available to the decision maker; in this case, interpret the event E as the piece of information that "the state belongs to E ." Note that the set of events is the power set 2^Ω , so there is no *a priori* restriction on what set of states can constitute an event. In Savage's framework, such restrictions are part of the objective constraints facing the decision maker, rather than inherent in the framework itself. See the section titled "Feasibility Constraints and Complexity."

- Consequences: $C = \{\dots, x, y, z, \dots\}$.

A consequence x is a complete description of all that is relevant to the decision maker's utility. This usually includes monetary rewards and other measures of material payoffs, but can also contain social and/or psychological factors (such as fairness, guilt, and the like).⁶

⁵Savage refers to these as "small worlds," which are coarsenings of the underlying complete state space.

⁶Savage writes that consequences "might in general involve money, life, state of health, approval of friends, well-being of others, the will of God, or anything at all about which the person could possibly be concerned. Consequences might appropriately be called states of the person, as opposed to states of the world."

- Acts: $\mathcal{F} = \{\dots, f, g, \dots\}$.

An act is a finite-valued function that maps states to consequences. The restriction to finite values is for technical and expository convenience, and is not essential for the theory.

Savage's Axioms

Savage's goal was to derive a representation of a decision maker's choice behavior in which uncertainty is represented by a probabilistic belief about the unknown states. This will, among other things, provide an interpretation of probability as the decision maker's implied degree of belief. Choice behavior is formalized as a preference $\succsim$ on $\mathcal{F}$ (formally, a binary relation on $\mathcal{F} \times \mathcal{F}$). Although in concrete decision problems choice is limited to some exogenously given feasible set of acts $\mathcal{B} \subset \mathcal{F}$, the framework assumes a preference that is defined on all acts and is independent of the particular feasible set.⁷

A crucial methodological aspect of Savage's framework is its focus on observable choices. Cognitive processes and other psychological aspects of decision making matter only to the extent that they have directly measurable implications on choice—at least in principle, and possibly only under idealized or hypothetical experimental or choice settings. The process of decision making—what steps a decision maker takes, or what heuristics he employs—has no formal meaning within the theory. The process is only relevant in so far as it has measurable behavioral consequences, and these are summarized by the preference $\succsim$ in the sense that one can, in principle, offer the decision maker the choice between any two acts f and g , and directly measure what choice he/she makes.

Seven axioms, numbered P1 through P7, characterize preferences that have an expected utility representation. Below we formally state and comment on the most controversial axioms, P1–P4.

⁷Other models of choice, such as the minimax regret criterion proposed by Savage [20], allow dependence on the feasible set.

Axiom P1 [Ordering]. $\succsim$ is complete, reflexive and transitive.

The key part of this axiom is completeness; it requires the decision maker to be able to rank any conceivable C -valued function on the state space, no matter how complex such function may be. Completeness implies that a rational decision maker cannot say: “I am unable to choose between f and g because the evidence available to me is insufficient and/or ambiguous.”

Completeness has been questioned by a number of authors. Bewley [21,22], for example, argues that one should not expect completeness to hold when there is ambiguity about the probabilities. Shafer [23] argued for a constructive interpretation of subjective probability, under which completeness need not hold. Savage’s response to these objections would presumably be that in any real choice in a problem, a decision must ultimately be made, even when the evidence is scant or imperfect.

Axiom P2 [Sure-Thing Principle–STP].

For all events E and acts f, g, h , and h' ,

$$\begin{aligned} \begin{bmatrix} f(\omega) & \text{if } \omega \in E \\ h(\omega) & \text{if } \omega \notin E \end{bmatrix} &\succsim \begin{bmatrix} g(\omega) & \text{if } \omega \in E \\ h(\omega) & \text{if } \omega \notin E \end{bmatrix} \\ \implies &\begin{bmatrix} f(\omega) & \text{if } \omega \in E \\ h'(\omega) & \text{if } \omega \notin E \end{bmatrix} \\ &\succsim \begin{bmatrix} g(\omega) & \text{if } \omega \in E \\ h'(\omega) & \text{if } \omega \notin E \end{bmatrix}. \end{aligned}$$

This is perhaps the most central of Savage’s Axioms. It states that the preference between two acts with a common extension outside some event E does not depend on that common extension. In the formalism above, consequences are determined according to f and g on the event E , and a common act h outside E . The STP (sure-thing principle) then says that the preference remains unaltered if we replace h by some other common extension h' . This amounts to the separability of the preference across events, a necessary

property for it to have an expected utility representation.

The STP cuts in many different ways and has been at the center of most of the controversies in this area. The two classic objections to the Savage framework by Allais [24] and Ellsberg [25] are based on violations of the STP. In the section titled “Information, Dynamic Choice, and Dynamic Consistency,” we note the close connection between the STP and the consistency of dynamic choice.

For the next axiom, we need the following definition: An event E is *null* if any pair of acts which differ only on E are indifferent. Below, identify a consequence x with the constant act that yields x in every state.

Axiom P3 [Monotonicity]. For all outcomes x and y , nonnull events E and acts g ,

$$\begin{aligned} \begin{bmatrix} x & \text{if } \omega \in E \\ g(\omega) & \text{if } \omega \notin E \end{bmatrix} &\succsim \begin{bmatrix} y & \text{if } \omega \in E \\ g(\omega) & \text{if } \omega \notin E \end{bmatrix} \\ \iff &x \succsim y. \end{aligned}$$

Roughly, if a consequence x is preferred to another consequence y , then this is so regardless of the state in which these consequences obtain. Conceptually, this axiom amounts to separating states, which are the object of uncertainty, from consequences, which is what matters for payoffs.

Objections to this axiom appeared in Karni *et al.* [26], Schervish *et al.* [27], and Karni [28], among others.

Axiom P4 [Weak Comparative Probability]. For all events A, B , and outcomes $x^* > x$ and $y^* > y$.

$$\begin{aligned} \begin{bmatrix} x^* & \text{if } \omega \in A \\ x & \text{if } \omega \notin A \end{bmatrix} &\succsim \begin{bmatrix} x & \text{if } \omega \in B \\ x^* & \text{if } \omega \notin B \end{bmatrix} \\ \implies &\begin{bmatrix} y^* & \text{if } \omega \in A \\ y & \text{if } \omega \notin A \end{bmatrix} \\ &\succsim \begin{bmatrix} y & \text{if } \omega \in B \\ y^* & \text{if } \omega \notin B \end{bmatrix}. \end{aligned}$$

Since $x^* \succ x$, the first preference ranking means that the decision maker prefers to bet on the event A rather than B . We would like to interpret this to mean that he/she, therefore, judges A to be more likely than B , an interpretation that is possible only if this likelihood judgments is independent of the specific consequences x^* and x . This is precisely what the axiom asserts.

Savage requires three more axioms, P5 (nondegeneracy), P6 (small event continuity), and P7 (uniform monotonicity).⁸ These axioms are needed primarily for technical reasons, and so they do not carry the substantive weight of the others.

Savage's Subjective Expected Utility Representation

Savage's Representation Theorem [Savage, 1954]. *A preference $\succsim$ satisfies P1-P7 if and only if there is a finitely additive probability measure⁹ P and a function $u : C \rightarrow \mathbb{R}$ such that for every pair of acts f and g :*

$$f \succsim g \iff \int_{\Omega} u(f(\omega)) dP \geq \int_{\Omega} u(g(\omega)) dP. \quad (2)$$

Moreover, P is unique and u is unique up to positive affine transformation.

In words, the theorem says that a decision maker who satisfies the axioms

- reduces all uncertainty about the states to a subjective probability measure P that reflects his beliefs;
- ranks consequences according to a utility function u that reflects his taste;

- evaluates acts according to the expected utility criterion.

Savage's theorem implicitly *defines* probability as the degree of belief implied by the decision maker's choice over uncertain prospects. Probability is not an objective property of the world, but the decision maker's subjective assessment of the likelihood of various events, as expressed in his willingness to make bets on those events. In Savage's framework, it makes no sense to talk about beliefs being "right" or "wrong," overly optimistic or pessimistic, or whether some probabilities may be more "objective" than others.¹⁰

An important implication of the theory is the *separation of tastes from beliefs*. To illustrate this, consider two decision makers who may disagree on how to rank various acts. In principle, their disagreement may have its source in how they assess the likelihood of various events, in their preference over consequences, or some complex interaction between the two. For example, imagine two leaders who may disagree on how to deal with an emerging nuclear threat by a rogue regime. Their disagreement may be the result of different assessment of how likely the regime is to develop weapon-grade fuel, and/or how undesirable it is to have such regime armed with nuclear weapons. Suppose that the two decision makers satisfy Savage's axioms, with P_1, P_2 and u_1, u_2 denoting their subjective beliefs and utilities respectively. Then Savage's theorem separates tastes from beliefs because it asserts that there can be no source of disagreement

⁸P5 asserts the existence of two nonindifferent outcomes, and P6 implies that the subjective belief is atomless (in particular, Ω must be infinite; Gul [29] provides an alternative model with a finite state space). P7 is necessary to handle infinite-valued acts.

⁹See footnote 16 for a brief discussion of the role of finite additivity.

¹⁰Anscombe and Aumann [30] offer a simpler derivation of subjective probability provided one is willing to assume the existence of objective lotteries. Specifically, their acts take values in the space of probability distribution on consequences, $\Delta(C)$, rather than "pure" consequences C as in Savage, and the decision maker is assumed to rank objective lotteries according to expected utility criterion. The key assumption here is an appropriate version of the independence axiom.

beyond differences between either P_1, P_2 and/or u_1, u_2 .¹¹

What is the significance of separating tastes from beliefs? One may argue that both are aspects of the preference $\succsim$ and both are subjective. Intuitively, however, tastes and beliefs are quite different animals. In Aumann's [31] words: "[U]tilities directly express tastes, which are inherently personal. It would be silly to talk about 'impersonal tastes,' tastes that are 'objective' or 'unbiased'. But it is not at all silly to talk about unbiased probability estimates, and even to strive to achieve them. On the contrary, people are often criticized for wishful thinking—for letting their preferences color their judgement. One cannot sensibly ask for expert advice on what one's tastes should be; but one may well ask for expert advice on probabilities."

While Aumann's reasoning is compelling, and undoubtedly shared by many (including the authors), it is important to understand that Savage's theory has nothing to say about judging whether beliefs are reasonable or to test them against evidence. De Finetti points out that a subjective proposition, such as a subjective probability assessment, is one which "no experience can prove [...] right, or wrong; nor, in general, could any conceivable criterion give any objective sense to the distinction [...] between right and wrong."

¹¹In Savage's theory, the separation between tastes and beliefs is a consequence of the fact that the representation identifies P , u , and the way in which they should be combined (the expected utility criterion). Machina and Schmeidler [18] provide an alternative axiomatization—which relaxes the STP but strengthens P4—characterizing preferences that are *probabilistically sophisticated*. As in Savage, these are preferences in which the decision maker has a probabilistic belief P over the state space, and who is indifferent between acts that induce the same distribution on consequences (under P). The difference is that lotteries over consequences are ranked according to a monotone functional; the expected utility criterion is but a special case of such functional. Machina and Schmeidler's [18] theory provides some of the weakest conditions known, separating tastes from beliefs while not requiring expected utility.

[6, p. 174]. The problem of developing criteria for testing beliefs, or to narrow the range of reasonable beliefs based on, say, criteria of simplicity, remains open.

INTERPRETATION AND IMPLICATIONS

Normative Theory and the Definition of "Rationality"

It is hard to argue that Savage's theory and his representation accurately describes all (or even most) instances of how people actually make choices. Classic examples by Allais [24] and Ellsberg [25] show that seemingly reasonable choices may violate the Savage axioms. A parallel development, pioneered by Tversky and Kahneman [32], draws on research in psychology to point out systematic deviations from Savage-style rationality. In a seminal paper, Tversky and Kahneman [32] argue that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. In general, these heuristics are quite useful, but sometimes they lead to severe and systematic errors."¹² In general, the resulting behavior is inconsistent with Savage's theory.

But, if this theory is not descriptive of the way people actually make decisions, then what is its contribution? One answer is that the theory provides an operational *definition* of what we mean by rational behavior, namely, as behavior consistent with the axioms and the expected utility representation (Eq. 2). This is a definition with far reaching consequences both in what it includes as well as what it excludes.

On the one hand, Savage's theory may be seen as too permissive a framework for defining rationality. Any probability measure on the state space, no matter how absurd, qualifies as rational belief. Believers in intelligent

¹²See Kahneman *et al.* [33] for a collection of important contributions to the literature on "heuristics and biases," and Kahneman [34] for a more recent account of the literature, especially, in how it relates to economics and related fields.

design and members of the Flat Earth Society¹³ are rational, provided only that they are consistent with the representation. Savage is simply not in the business of passing normative judgments about what is and what is not a reasonable belief, or how beliefs should be formed. Rather, he seeks criteria for coherence of beliefs “to distinguish between coherent behavior and blunder, or demonstrable incoherence in the face of uncertainty” and it is best thought of as a tool “by which a person can police his own potential decisions for incoherency.” Savage [35, p. 307].

On the other hand, Savage’s theory has powerful consequences in terms of what it rules out as irrational behavior, or as admissible models of such behavior. His framework, among other things: (i) erases any difference between risk and uncertainty; (ii) expresses dynamic decision problems as constrained static problems; and (iii) reduces differences of opinions to either unexplained difference in priors or to differences in information. These and other implications of Savage’s framework are discussed further in this article. Here we illustrate the power of the framework in a particularly simple example.

Consider the following example which appears in Machina [36]: Mom has one indivisible treat to give to one of her two children, Abigail and Ben. Let a, b denote the acts of giving the treat to Abigail and Ben, respectively, and let m denote the act where Mom flips a fair coin and decides to give the treat to Abigail if the coin turns heads and to Ben otherwise. Suppose that $a \sim b$, so Mom is indifferent between giving the treat to one child or the other. Expected utility implies that m should be indifferent to a and b , but it also seems entirely reasonable that $m \succ a \sim b$, the latter ranking indicating that Mom views a coin toss as more equitable, and thus preferable to an arbitrary assignment of the treat. This, on the surface, seems like a perfectly rational behavior that contradicts Savage’s representation.¹⁴

In Savage’s framework, instances of conflict between supposedly rational behavior and the rationality axioms must be accounted for as the result of misspecifying the state space, the set of consequences, or the constraints facing the decision maker. In Machina’s Mom example, the answer is easy: The set of consequences ought to take into account “fairness.” What matters for Mom and the kids is not just who gets the treat, but also whether the procedure is perceived as fair. The original, and apparently paradoxical description of the problem missed an important payoff-relevant aspect: Mom may not be indifferent to the different procedures for allocating the treat, and so the procedure should properly be part of the set of consequences.

Machina [36, Section 6] provides examples and a critical discussion of the approach of change-the-consequences-so-anomalies-disappear, which is presented in the last paragraph. He shows, for example, how enlarging the space of consequences can eliminate the famous “paradox” due to Allais [24]. In a related vein, Geanakoplos *et al.* [37] provide examples and an analysis of “psychological games.” These are strategic situations where payoffs may depend not just on the material consequences, but also on such considerations as fairness, guilt, vindictiveness, and so on.

The upshot of the above discussion is: (i) if the set of consequences (or any other aspect of the model, such as the state space) is misspecified, then there is no reason not to expect violations and anomalies; and (ii) Savage’s framework provides a *methodology* that gives us guidance as to which models or explanations to pursue when we encounter anomalous behavior. In the above example, the framework suggests a misspecification of the set of consequences as the source of the anomaly.

¹³<http://www.alaska.net/clund/edjublonskopf/Flatearthsociety.htm>.

¹⁴Probabilities in this example are objective, in the sense of being exogenously given as part of

the description of the problem. The example can, of course, be readily restated for subjective probability. The point of the example is to illustrate an intuitive contradiction with the reasonableness of the expected utility criterion aspect of Savage’s representation.

Feasibility Constraints and Complexity

Savage's theory assumes a preference (and produces a representation) defined over all acts. In virtually every problem of interest, the decision maker chooses from a set of feasible acts, $\mathcal{B} \subseteq \mathcal{F}$. In this case, the representation identifies the optimal acts given $\mathcal{B}$ as the set.

$$\operatorname{argmax}_{f \in \mathcal{B}} \int_{\Omega} u(f(\omega)) dP. \quad (3)$$

Although feasibility constraints do not formally appear in Savage's representation, his theory has the implication that such constraints must be introduced as exogenous and objective elements of the decision problem. Subjectivity in Savage's theory is limited to the decision maker's taste over consequences, his beliefs over states, and the expected utility form he uses to combine the two. Objective feasibility constraints cannot, within Savage's theory, influence the decision maker's tastes or subjective judgments.

This has important implications. A common, and reasonable, complaint about Savage's theory, for example, Shafer [23], is that a decision framework should take into account the complexity of describing the states, of comparing acts, and so on. At one level, complexity, however defined, can be trivially accommodated within Savage's theory: let $\mathcal{Z} \subset 2^{\mathcal{F}}$ denote the set of all subsets of acts that are consistent with our favorite notion of "simplicity" or computational feasibility, and limit the applicability of Equation (3) to those sets $\mathcal{B} \in \mathcal{Z}$.

The problem with this approach is that in Savage's framework, Equation (3) makes sense only because we start with a preference over all acts and derive a representation which is independent of any cognitive or complexity-based constraints.¹⁵ One might reasonably expect cognitive or computational limitations to constrain the decision maker's preference. But accommodating such constraints in Savage's framework seems to

require a decision maker who starts with a preference that violates them in the first place.¹⁶

Information, Dynamic Choice, and Dynamic Consistency

So what might be reasonable specifications of constraints that limit decision makers' choices in practice? Obvious sources of constraints involve such things as wealth, technology, or regulations. Call these, for lack of a better term, *physical* constraints, so they can be more easily distinguished from *informational* constraints that reflect limited information. The simplest way to introduce limited information in Savage's framework is in terms of a finite information partition $\Pi = \{\pi_1, \dots, \pi_n\}$ on the state space.¹⁷ At a state ω , the decision maker knows only that the state belongs to the partition element $\pi(\omega) \in \Pi$ that contains ω . Limited information can be viewed as constraining the decision maker's choices to the subset of acts $\mathcal{F}_{\Pi}$ that are measurable with respect to Π . An act $f \in \mathcal{F}_{\Pi}$ is a choice that is contingent on the information and can be evaluated in a completely standard way.

¹⁶A related issue is de Finetti's and Savage's insistence on assuming only finite, rather than countable, additivity. It is possible to obtain a representation like Equation 2, as Arrow [39] does, with countably additive probability, but at the cost of restricting attention to, say, a Borel σ -algebra of events, rather than the power set. Savage and de Finetti would object to this on methodological grounds: a decision framework should not blur the separation between structural assumptions about the choice setting and the feasibility constraints facing the decision maker in a particular choice problem. In Savage's framework, imposing measure theoretic or topological structures amounts to restrictions on the set of acts available to the decision maker. These restrictions should be part of the exogenous constraints facing the decision maker, and not as limitations on the scope of his subjective beliefs. See Al-Najjar [40] for further discussion of this point.

¹⁷More sophisticated models represent information via a σ -algebra and, in dynamic settings, an information filtration.

¹⁵Kopylov [38] provides an extension of Savage's theory to more general domains of events that need not be algebras.

Although Savage's original formulation is *timeless*, the above makes it possible to reinterpret it as a choice problem that takes place over time. Think of the problem as consisting of an *ex ante* stage, where acts are evaluated according to the representation (2), and an *ex post* stage, where partial information about the state is revealed. Suppose this information takes the form that the state lies in some component π of the finite partition Π . Then the STP essentially amounts to saying that the *ex ante* preference $\succsim$ gives rise to a well-defined conditional preference $\succsim_\pi$ that is dynamically consistent in the sense that, roughly, the conditional ranking over acts agrees with the *ex ante* ranking.¹⁸ Epstein and Le Breton [41] show that dynamic consistency is essentially equivalent to a weakened version of the STP. In particular, models of ambiguity aversion (e.g., multiple prior models) cannot be dynamically consistent in any plausible sense (see the section titled "Risk versus Uncertainty").¹⁹

The fact that Savage's framework can so effortlessly accommodate limited information is one of its crucial advantages. Without this it would be hard to imagine how the development of games with incomplete information [44], dynamic games of all types, and models of knowledge and interactive decision making [31,45] could have proceeded.

What makes this formulation of dynamic choice so attractive is its tractability: dynamic choice is nothing more than static choice subject to informational constraints. As noted by some authors (Kreps [17,46], for instance), this reduction to static choice seems to be missing intuitively important aspects of dynamic behavior we would care about as modelers in understanding games, political contests, or business strategy.

¹⁸Formally, the complement of π is null under $\succsim_\pi$, and for any two acts f and g that agree outside π , $f \succsim g$ if and only if $f \succsim_\pi g$.

¹⁹See also Karni and Schmeidler [42]. Wakker [43] discusses another issue related to dynamic consistency, namely aversion to information that could result when the expected utility criterion is violated.

Risk versus Uncertainty

The distinction between situations where probabilities are known and ones where there is insufficient information to form a probability judgment seems, at an intuitive level, meaningful. Most people think of coin tosses, dice throws, or routine insurance losses as well-understood actuarial risks governed by objective probabilities. This is in contrast with events like "there will be a terrorist nuclear attack within the next 10 years," or "my competitor will respond to my new product introduction with a price cut," where there does not seem to be an obvious probability assignment. Knight [3] and Keynes [47] distinguish between situations of risk, where probabilities are well-understood, and problems involving uncertainty.²⁰ While this distinction seems, at some levels, quite intuitive, it has no meaning within Savage's theory. Put differently, a decision maker in this theory reduces all uncertainties to risks, in the sense that he treats all uncertain prospects in the same way, namely by evaluating them using the expected utility criterion with respect to his subjective probability measure.

²⁰Keynes [47] explains the difference as follows: "By 'uncertain' knowledge, let me explain, I do not mean merely to distinguish what is known for certain from what is only probable. The game of roulette is not subject, in this sense, to uncertainty; nor is the prospect of a Victory bond being drawn. Or, again, the expectation of life is only slightly uncertain. Even the weather is only moderately uncertain. The sense in which I am using the term is that in which the prospect of a European war is uncertain, or the price of copper and the rate of interest 20 years hence, or the obsolescence of a new invention, or the position of private wealth-owners in the social system in 1970. About these matters there is no scientific basis on which to form any calculable probability whatever. We simply do not know. Nevertheless, the necessity for action and for decision compels us as practical men to do our best to overlook this awkward fact and to behave exactly as we should if we had behind us a good Benthamite calculation of a series of prospective advantages and disadvantages, each multiplied by its appropriate probability, waiting to be summed."

Bewley [21,22], Schmeidler [48], and Gilboa and Schmeidler [49] extend the framework to allow for such distinction. Bewley relaxes the completeness part of P1 so the decision maker may be unable to rank some pairs of acts. He assumes that von Neumann–Morgenstern independence axiom holds over the set of acts on which his preference is complete. His main theorem represents such preferences as follows: There is a compact convex set of probability measure $\mathcal{P}_B$ and a von Neumann–Morgenstern utility function u such that

$$\begin{aligned} f \succsim g &\iff \int_{\Omega} u(f(\omega)) dP \\ &\geq \int_{\Omega} u(g(\omega)) dP, \quad \forall P \in \mathcal{P}_B. \end{aligned} \quad (4)$$

In words, f is preferred to g if and only if there is unanimity among all measures in $\mathcal{P}_B$ on how to rank them. When the set $\mathcal{P}_B$ is a singleton, this reduces to standard subjective expected utility (which must, of course, be complete). A nonsingleton $\mathcal{P}_B$ represents uncertainty, or ambiguity, about the “true” probability and will lead to incompleteness. The unanimity criterion in Equation (4) is a robustness requirement under which the decision maker ranks acts only when such ranking is insensitive to how the “true” prior varies within $\mathcal{P}_B$.

Gilboa and Schmeidler [49] also develop a representation involving a convex set of priors.²¹ Rather than weakening completeness, Gilboa and Schmeidler [49] weaken the independence axiom (essentially, the STP) to obtain the following representation: there is a compact convex set of probability measure $\mathcal{P}_{GS}$ and a von Neumann–Morgenstern utility function u such that

$$\begin{aligned} f \succsim g &\iff \min_{P \in \mathcal{P}_{GS}} \int_{\Omega} u(f(\omega)) dP \\ &\geq \min_{P \in \mathcal{P}_{GS}} \int_{\Omega} u(g(\omega)) dP. \end{aligned} \quad (5)$$

²¹In Schmeidler’s [48] model, beliefs are represented as a capacity and the decision maker evaluates acts using the Choquet integral. These preferences overlap with those introduced in Gilboa and Schmeidler [49].

That is, each act f is evaluated according to the worst measure $P_f \in \mathcal{P}_{GS}$ for that act. Again, this reduces to Savage-style representation if the set $\mathcal{P}_{GS}$ is a singleton, in which case the preference must satisfy independence. However, when independence is violated (and their other axioms hold), the set $\mathcal{P}_{GS}$ is nonsingleton, and the resulting preference displays pessimism or caution. Gilboa–Schmeidler theory draws some of its motivation from the desire to reproduce choice patterns consistent with those appearing in the Ellsberg [25] paradox. Their representation has been used to fit anomalies in asset pricing (starting with Dow and Werlang [50]) and macroeconomics (see Hansen [51] for a recent example).²²

As attempts to capture ambiguity or uncertainty, Bewley’s [21,22] and Gilboa and Schmeidler’s [49] representations suffer from significant problems. Bewley’s model cannot answer the obvious question: What should the decision maker do when he faces a choice between two acts f and g that are not ranked according to his criterion? Interesting cases of ambiguity arise precisely when a decision maker has to choose between, say, two business plans about which he lacks precise priors. Presumably, the decision maker must ultimately make a choice, yet Bewley’s theory is silent on what this choice should be. It may be useful to think of Bewley’s model as corresponding to a different revealed preference experiment: while Savage envisions a setting where the decision maker is *forced* to make a choice when confronted with any feasible set $\mathcal{B}$, Bewley’s model is best thought of as recording the choice this decision maker would *voluntarily* make.

The problems with Gilboa and Schmeidler’s [49] model are different and potentially more serious. It is well known, and can be easily demonstrated by example, that their model cannot admit any plausible form of dynamic consistency of choice. See Epstein

²²See Al-Najjar and Weinstein [52] for a critique of these efforts in the special issue of *Economics & Philosophy* on ambiguity aversion, and the replies to their paper in that same issue.

and Le Breton’s [41] paper, titled “Dynamically Consistent Beliefs must be Bayesian.” Arguments concerning the tension between their maxmin criterion (Eq. 5) and consistent dynamic choice are discussed at length in Al-Najjar and Weinstein [52].

It is interesting to note that neither Bewley nor the Gilboa–Schmeidler models appear to capture Knight’s [3] intuition of what uncertainty is about: “[W]e must observe at the outset that when an individual instance only is at issue, there is no difference for conduct between a measurable risk and an unmeasurable uncertainty. The individual, as already observed, throws his estimate of the value of an opinion into the probability form [...] and ‘feels’ toward it as toward any other probability situation.” This quote of Knight, and Keynes’ description of uncertainty in footnote (20), seems to imply that uncertainty is inherently about dynamic choice, while the Bewley and Gilboa–Schmeidler’s models produce uncertainty-sensitive behavior even in a single, static choice problem. This suggests that their representations capture a different type of behavior than what either Knight or Keynes had in mind.

Exchangeability, Objectivity, and Frequencies

As pointed out earlier, probability judgments in Savage’s theory are purely subjective; there is no sense in which some probabilities are more “objective” than others. Intuitively, however, subjective probabilities should, somehow, be grounded with objective facts, such as the frequencies of outcomes obtained in repetitions of the similar experiments. De Finetti’s notion of exchangeability and his famous representation make it possible to integrate a decision maker’s subjective judgment with objective frequencies. Our exposition here follows Kreps [17] (who considers “De Finetti’s theorem as the fundamental theorem of (most) statistics.”)

Suppose that an experimental scientist or a statistician conducts (or passively observes) a sequence of experiments with outcomes in some set S (so the state space has the product structure $\Omega = S \times S \times \dots$). To pool information across experiments, the observer must assume, in one way or another, that the

experiments are, in a sense, “similar.” De Finetti’s concept of exchangeability formalizes this intuition of similarity and its role as guide in decision making. Roughly, a decision maker subjectively views a set of experiments as *exchangeable* if he treats the indices interchangeably.²³

Different experiments will typically result in different outcomes due to the influence of a multitude of poorly understood or unmodeled factors. Nevertheless, a decision maker’s subjective judgment that the experiments are exchangeable amounts to believing that they are governed by the same underlying stochastic structure. De Finetti’s celebrated theorem says that a probability distribution P on Ω is exchangeable if and only if it has the parametric form

$$P = \int_{\Theta} P^{\theta} d\mu(\theta). \quad (6)$$

Here the parameter set Θ indexes the set of all i.i.d. distributions P^{θ} with marginal θ , and μ is a probability distribution on Θ .

The decomposition (Eq. 6) says that a decision maker subjectively views the experiment as exchangeable if and only if he believes the outcomes to be i.i.d. with parameter θ whose distribution is given by a subjective belief μ . Since the i.i.d. parameters may also be interpreted as limiting frequencies, the representation can be equivalently stated as: a subjective belief P is exchangeable if and only if it can be expressed as distribution on limiting frequencies.

Chew and Sagi [53] show that a form of event exchangeability (plus other weak axioms) implies probabilistic sophistication. De Castro and Al-Najjar [54] isolate the implications of exchangeability assuming only that the preference is monotone, transitive, and continuous, but otherwise incomplete and/or fail probabilistic sophistication. Their motivation is that De Finetti’s

²³ Somewhat more formally, exchangeability means that the decision maker ranks as indifferent an act f and the act $f \circ \pi$ that pays f after a finite permutation of the coordinates π is applied. In particular, he considers an outcome s to be just as likely to appear in the i th as in the j th experiment.

representation simultaneously identifies the relevant parameters *and* Bayesian beliefs about them. On the other hand, the idea of similarity of experiments and the concept of parameters are meaningful even if expected utility is not assumed. De Castro and Al-Najjar [54] develop a subjective version of the ergodic theorem which takes as primitive preferences, rather than probabilities (as in standard ergodic theory). They use this ergodic theorem to show decomposition into i.i.d. parameters without necessarily assuming that the preference is expected utility.

CONCLUDING REMARKS

In his famous reply to calls to abandon a theory due to its anomalies and unintuitive implications, Hilbert said that “No one will drive us from the paradise which Cantor created for us.” Savage has created a “paradise” where important ideas, and even entire fields, could flourish. It is hard to conceive of modern game theory, industrial organization, mechanism design, auction theory, models of incomplete information, bargaining theory, portfolio theory, and option pricing without uncertainty and the use of probabilistic methods to reason about it. Savage has given us a broad license to reduce all uncertainty to risks that can be quantified using probability.

We argued above that one should set aside critiques of Savage that are based on misunderstanding or too narrow an interpretation of his framework. Often the failure is to recognize that Savage’s axioms are not substantive statement subject to refutations by experimental or psychological evidence. Rather, they should be viewed as an organizing methodology that guides and constrains our modeling of choice under uncertainty. By this standard, more than half century after its introduction, Savage’s theory is an unqualified success. Yet, there is a sense that something was lost in the bargain, since we must give up the ability to talk about such things as the difference between risk and uncertainty, “true” dynamic choice, or to have a meaningful theory of belief formation. These are promising direction for future research.

Acknowledgment

We thank Simone Galperti and Mallesh Pai for comments and proof-reading. This essay owes much to countless conversations we had with Jonathan Weinstein. His insights into decision theory shaped many of the ideas reported here.

REFERENCES

1. Keynes J. A treatise on probability. New York: Harper & Row; 1921.
2. Borel E. Apropos of a treatise on probability. In: Kyburg H, Smokler H, editors. Studies in subjective probability. New York: John Wiley & Sons, Inc.; 1964. pp. 46–60.
3. Knight F. Risk, uncertainty and profit. New York: Harper; 1921.
4. Ramsey F. Truth and probability. The foundations of mathematics and other logical essays. London: Routledge & Kegan Paul; 1931. pp. 156–198.
5. de Finetti B. La prévision: ses lois logiques, ses sources subjectives. Ann. Inst. Henri Poincaré 1937;7:1–68.
6. de Finetti B. Probabilism. Erkenntnis 1989; 31(2):169–223. (Originally published in 1931 as “*Probabilismo*” in Italian).
7. von Mises R. Probability, statistics and truth. London: Allen and Unwin: Macmillan; 1957.
8. Reichenbach H. The theory of probability. University of California Press; 1949.
9. Savage LJ. The foundations of statistics. New York: John Wiley & Sons Inc.; 1954.
10. Galavotti M. Philosophical introduction to probability. Stanford (CA): CSLI Publications Stanford; 2005.
11. Bernstein PL. Against the gods: the remarkable story of risk. New York: John Wiley & Sons, Inc.; 1996.
12. Zabell S. Ramsey, truth, and probability. Theoria 1991;57:211–238.
13. Galavotti M. Anti-realism in the philosophy of probability: Bruno de Finetti’s subjectivism. Erkenntnis 1989;31(2):239–261.
14. Galavotti MC. Subjectivism, objectivism and objectivity in bruno De Finetti’s Bayesianism. In: Corfield D, Williamson J, editors. Foundations of Bayesianism. Amsterdam: Kluwer Academic Publishers; 2001. pp. 161–174.
15. Doob J. Stochastic processes. New York: John Wiley & Sons, Inc.; 1953.

16. Fishburn P. Utility theory for decision making. New York: John Wiley & Sons, Inc.; 1970.
17. Kreps D. Notes on the theory of choice. Boulder, CO: Westview Press; 1988.
18. Machina MJ, Schmeidler D. A more robust definition of subjective probability. *Econometrica* 1992;60(4):745–780.
19. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton (NJ): Princeton university press; 1947.
20. Savage LJ. The theory of statistical decision. *J Am Stat Assoc* 1951;46(253):55–67.
21. Bewley T. Knightian decision theory: Part I. Cowles Foundation Discussion Paper No. 807. 1986.
22. Bewley T. Knightian decision theory. Part I. *Decis Econ Finance* 2002;25(2):79–110.
23. Shafer G. Savage revisited. *Stat Sci* 1986; 1(4):463–501. With comments and a rejoinder by the author.
24. Allais M. Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica* 1953;21(4):503–546.
25. Ellsberg D. Risk, ambiguity, and the savage axioms. *Q J Econ* 1961;75(4):643–669.
26. Karni E, Schmeidler D, Vind K. On state dependent preferences and subjective probabilities. *Econometrica* 1983;51(4):1021–1031.
27. Schervish M, Seidenfeld T, Kadane J. State-dependent utilities. *J Am Stat Assoc* 1990; 840–847.
28. Karni E. A definition of subjective probabilities with state-dependent preferences. *Econometrica* 1993;61(1):187–198.
29. Gul F. Savage's theorem with a finite number of states. *J Econ Theory* 1992;57(1):99–110.
30. Anscombe F, Aumann R. A definition of subjective probability. *Ann Math Stat* 1963;34(1): 199–205.
31. Aumann RJ. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 1987;55(1):1–18.
32. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1974;185(4157):1124–1131.
33. Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: heuristics and biases. New York: Cambridge University Press; 1982.
34. Kahneman D. Maps of bounded rationality: psychology for behavioral economics. *Am Econ Rev* 2003;93(5):1449–1–75.
35. Savage LJ. Difficulties in the theory of personal probability. *Philos Sci* 1967;34:305–10.
36. Machina M. Dynamic consistency and non-expected utility models of choice under uncertainty. *J Econ Lit* 1989;27(4):1622–1668.
37. Geanakoplos J, Pearce D, Stacchetti E. Psychological games and sequential rationality. *Games Econ Behav* 1989;1:60–79.
38. Kopylov I. Subjective probabilities on “SMALL” domains. *J Econ Theory* 2007;133 (1):236–265.
39. Arrow K. Essays in the theory of risk-bearing. London: North-Holland; 1971.
40. Al-Najjar NI. Decision makers as statisticians: diversity, ambiguity and learning. *Econometrica* 2009;77:1339–1369.
41. Epstein LG, Le Breton M. Dynamically consistent beliefs must be Bayesian. *J Econ Theory* 1993;61(1):1–22.
42. Karni E, Schmeidler D. Atemporal dynamic consistency and expected utility theory* 1. *J Econ Theory* 1991;54(2):401–408.
43. Wakker P. Nonexpected utility as aversion of information. *J Behav Decis Making* 1988; 1:169–175.
44. Harsanyi JC. Games with incomplete information played by “Bayesian” players. I. The basic model. *Manage Sci* 1967;14: 159–182.
45. Aumann R. Agreeing to disagree. *Ann Stat* 1976;4(6):1236–1239.
46. Kreps D. Anticipated utility and dynamic choice. In: Jacobs D, Kalai E, Kamien M, editors. *Frontiers of research in economic theory: the Nancy L. Schwartz memorial lectures*. Cambridge: Cambridge University Press; 1998.
47. Keynes JM. The general theory of employment. *Q J Econ* 1937;51(2):209–223.
48. Schmeidler D. Subjective probability and expected utility without additivity. *Econometrica* 1989;57(3):571–587.
49. Gilboa I, Schmeidler D. Maxmin expected utility with nonunique prior. *J Math Econ* 1989; 18(2):141–153.
50. Dow J, Werlang SdC. Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica* 1992;60:197–204.
51. Hansen L. Beliefs, doubts and learning: valuing macroeconomic risk. *Am Econ Rev* 2007;97(2):1–30.

52. Al-Najjar NI, Weinstein J. The ambiguity aversion literature: a critical assessment. *Econ Philos* 2009;25:249–284, lead paper in a special issue on ambiguity aversion, with replies and a rejoinder.
53. Chew SHC, Sagi J. Event exchangeability: probabilistic sophistication without continuity or monotonicity. *Econometrica* 2006;74:771–786.
54. De Castro L, Al-Najjar NI. A subjective foundation of objective probability. Northwestern University; 2009.

SUPPLIER SELECTION

DAMIAN R. BEIL
Stephen M. Ross School of Business,
University of Michigan, Ann Arbor,
Michigan

Today the average US manufacturer spends roughly half its revenue to purchase goods and services [1]. This makes a company's success dependent on its interactions with suppliers. The role of procurement managers (buyers) within companies has become extremely important, often involving staggering dollar values: a recent cross-industry survey of companies—in areas ranging from aerospace to semiconductors—placed companies' average total spend per procurement employee at \$115 million [2].

With so much of a company's money on the line, and increasing reliance on outsourcing of many complex services and products, the job of a buyer is not only important but also challenging. Buyers must define and measure what "best value" means for the buying organization, and execute procurement decisions accordingly. To identify best value, the buyer must interface with technical, legal, and operations experts within the buyer's company, and act as an expert negotiator and coordinator across many internal and external parties.

Supplier selection is the process by which the buyer identifies, evaluates, and contracts with suppliers. The challenges mentioned above make supplier selection a fertile topic for operations and management science disciplines. There is also a growing audience for such research, as the importance of fostering talent by employing buyers with analytical expertise, general management backgrounds, and deep knowledge in a particular purchasing category becomes widespread [3].

This article is organized around the major steps involved in supplier selection. First, the buyer must identify qualified potential suppliers, as described in the section titled "Identifying Potential Suppliers." Next, the

buyer must evaluate these suppliers. This process is initiated when the buyer formally solicits information from suppliers, as described in the section titled "Information Requests to Suppliers." Depending on the information request, suppliers respond by providing "bids" for the contract, specifying an offer on the contract terms, such as price, lead time, and quality. Various contract terms, which relate to the type of contract up for bid, are overviewed in the section titled "Contract Terms." Suppliers' offers often evolve over the course of negotiation with the buyer, and negotiation processes are touched on in the section titled "Negotiation Process." Finally, as discussed in the section titled "Supplier Evaluation and Contract Award," the buyer determines which supplier or suppliers will be awarded a contract and subsequently monitors the supplier during the life of the contract to support future supplier selection iterations. Finally, the section titled "Supplier Selection Research" discusses ORMS research on supplier selection. While this article introduces the key steps in supplier selection, further readings in this encyclopedia can provide a more detailed picture. Pointers to such readings are suggested throughout this article; most importantly, see section titled "Vehicle Routing and Scheduling" in this encyclopedia for details on procurement contracts and sections titled "Game Theory" and "Game Theory: Extension and Development" in this encyclopedia for discussions about negotiations.

For consistency in the article we refer to "buyers," but in practice the role we describe is also called *procurement manager*, *procurement agent*, or *contracts manager*. This article focuses on the complexities and frictions involved in supplier selection (verifying that suppliers are indeed qualified, using historical supplier performance in making award decisions, etc.). Such complexities are present in the supplier selection process for most goods and services. One possible exception is raw materials traded via commodity exchanges; such markets

are specifically designed to circumvent the complexities and frictions of supplier selection by instituting highly standardized contract terms, using liquid markets to find prices, and using clearinghouses to guarantee the terms of trade. As such, commodity exchanges will not be covered in this article.

IDENTIFYING POTENTIAL SUPPLIERS

To survive in the intensely competitive global economy, it is often critically important to not only develop existing suppliers but also to discover new suppliers. This section outlines the process of finding viable new suppliers. The section titled “Importance of New Suppliers” first briefly motivates why a buyer might wish to find new suppliers. The section titled “Reasons for Supplier Qualification Screening” explains why identifying suppliers is only part of the challenge—the buyer must also be cognizant of the need to ensure that such suppliers are qualified. Supplier qualification screening processes are discussed in the section titled “Supplier Qualification Screening Process.” Because identifying and qualifying potential suppliers can be time consuming and costly, buyers often develop a long-term supply base consisting of qualified suppliers, as discussed in the section titled “Creating a Supply Base.”

Importance of New Suppliers

Several factors make new suppliers important. First, there may exist new suppliers who are superior in some way to a firm’s existing suppliers. For example, a new supplier may have developed a novel production technology or streamlined process, which allows significant reduction in the production costs relative to predominate production technology or processes. Or, a new supplier may have a structural cost advantage over existing suppliers, for example, due to low labor costs or favorable import/export regulations in its home country. Second, existing suppliers may go out of business, or their costs may be increasing. Third, the buyer may need additional suppliers simply

to drive competition, reduce supply disruption risks, or meet other business objectives such as supplier diversity (see the section titled “Contract Award”). In recognition of these reasons, buyers and their internal customers may be obliged by company policy to locate a minimum number of viable, potential suppliers for every product or service procured.

Reasons for Supplier Qualification Screening

Finding a viable new supplier is challenging mainly because of the need to verify the supplier’s ability to meet the buyer’s myriad requirements [4]. Supplier nonperformance on even the most basic level, and for the simplest commodity, can have dire consequences for the buyer, as recounted in the following nursery rhyme:

For want of a nail the shoe was lost. For want
of a shoe the horse was lost.
For want of a horse the rider was lost. For
want of a rider the battle was lost.
For want of a battle the kingdom was lost. And
all for the want of a horseshoe nail.

While apocryphal, the “for want of a nail” message holds a surprising degree of relevance for today’s complex, global supply chains. Boeing’s 787 Dreamliner production schedule was significantly affected by shortages of fasteners, essentially bolts that secure sections of the fuselage together [5]. In consumer products many product safety issues have been traced back to suppliers failing to meet a buyer’s requirements, resulting in dangerous lead paint in toys [6], unsafe car tires [7], and pet food containing poisonous chemicals [8].

Production delays due to parts shortages and recalls of faulty products produced by noncompliant suppliers have cost buyer firms millions of dollars through recalls, warranty costs, and associated inventory adjustments, and have inflicted untold damage on their reputations and future sales potential. Menu Foods’ market capitalization was halved after the firm recalled its pet food in March 2007; its suppliers are suspected of mixing toxic melamine into their wheat gluten in an effort to make it appear more protein rich [9].

New Jersey-based tire importer Foreign Tire Sales traced field failures of its tires to an unauthorized design change made by its supplier, whose design engineer decided to omit gum strips, apparently unaware of their role in preventing tread separation [7]. A surprised Foreign Tire Sales was forced by US government authorities to recall a quarter of a million tires, and risked bankruptcy as a result [10].

Supplier Qualification Screening Process

To avoid the dire outcomes of supplier non-performance, buyers typically take proactive steps to verify a supplier's qualifications prior to awarding them a contract. The primary goal of "supplier qualification screening" is to reduce the likelihood of supplier nonperformance, such as late delivery, nondelivery, or delivery of nonconforming (faulty) goods. A secondary goal is simply to ensure that the supplier will be a responsible and responsive partner in the day-to-day business relationship with the buyer [4]. Supplier qualification screening involves many aspects, which are outlined below.

Reference Checks. The buyer may contact previous customers and ask about the supplier's delivery performance, adherence to contract terms, what (if any) problems arose and how they were resolved, and so on.

Financial Status Checks. The buyer may use published supplier ratings (e.g., Dunn and Bradstreet) to determine the supplier's financial status and likely financial viability in the short to medium term. For example, if the supplier has recently assumed significant debt, this may raise red flags about the possibility that the supplier will declare bankruptcy before fulfilling its obligations to the buyer.

Surge Capacity Availability. The supplier's capacity to increase delivery quantities within short lead times is important as the buyer may be uncertain about their exact quantity needs over the life of the contract. This is particularly true for long-term contracts where demand for the buyer's product may be heavily tied to unforeseen market events (e.g., demand for an airplane manufacturer's products are highly dependent on the

overall economy, which in turn periodically goes through periods of growth and contraction). Surge capacity is available when a supplier has access to second or third shifts, overtime, underutilized facilities, and so on.

Indications of Supplier Quality. The buyer might require that suppliers have ISO 9000 certification (or similar), indicating that the supplier has policies, procedures, documentation, and training in place to ensure continuous adherence to quality standards. However, in some cases the certification documents can be misleading and/or easily forged [4]. To actually see if an adequate level of quality is achievable, the buyer may have to look deeply into the supplier's organization to ensure that the supplier is capable and competent to meet the buyer's specifications.

Ability to Meet Specifications. To rigorously check the supplier's capabilities the buyer might (i) request samples of supplier products and test them to ensure conformance to the buyer's requirements; (ii) visit the supplier's production facility and interview line workers and engineers to ensure that all members of the supplier team understand the critical features of the product in their charge. For example, a buyer seeking to purchase tires from a supplier may interview the design engineers to ensure that they understand each aspect of the tire's design (for instance, the role of gum strips in preventing tread separation at high speeds); and (iii) audit the production facilities to ensure that production can and will only proceed in a manner approved by the buyer. For instance, the buyer may require the supplier to restrict their production to small batch sizes in order to prevent contamination outbreaks from spoiling the entire production run.

Buy-in from Internal Customer(s). Because the buyer typically acts on behalf of an internal customer within the buyer's organization, buy-in from this internal customer is a crucial step prior to contracting. For example, suppose the buyer is purchasing a complex circuit board on behalf of the engineering department (which owns responsibility for this component). To ensure that the internal customer has confidence in the supplier and is willing to work with the supplier, the buyer

will set up meetings between the buyer firm's engineers responsible for the part and the supplier's engineers who would be responsible for producing it.

Supplier qualification processes are costly and can be time consuming. As described above, the processes can involve travel to distant supplier sites. Interviews with suppliers and suppliers' customers are time consuming. Moreover, the entire process involves not only the buyer but also internal customers throughout the buyer organization. Consequently, qualification can take weeks or months—even for commodity-type parts such as printed circuit boards.

Creating a Supply Base

Suppliers who have passed the qualification requirements and are eligible for contract award are commonly referred to as *prequalified* suppliers. If the buyer utilizes short-term contracts and frequently reprocures the same item, it typically makes sense to establish a cohort of prequalified suppliers who will compete for these contracts. Even if the buyer uses long-term contracts for individual items (meaning contracts for individual items are infrequently rebid), it might still make sense to use a prequalified supply base: if the supply base members can potentially supply many different items, they can compete to produce whichever item's long-term contract is up for rebidding. Finally, using a supply base not only reduces qualification screening costs but also allows for the development of standardized contracts and terms and conditions for prequalified suppliers, thereby streamlining administrative processes involved in contracting.

INFORMATION REQUESTS TO SUPPLIERS

Once the buyer has identified potential suppliers, the next step in supplier selection is to formally request that the suppliers provide information about their goods or services. While there is no agreed-upon terminology, generally the buyer makes one of three types of information requests to suppliers. The request types, each appropriate for a different situation, are described below.

Request for information (RFI) is issued when the buyer seeks to gain market intelligence regarding what alternatives and possibilities are available to meet the buyer's needs. Typically, the buyer asks suppliers what goods and services they could potentially provide, what differentiates them from other vendors in the marketplace, and so on. With an RFI the buyer does not state a particular intention to award a contract. However, since responding to an RFI is time consuming for suppliers, generally, suppliers will only respond to the RFI if they expect that the buyer will eventually issue a request for proposal (RFP) or a request for quote (RFQ), which are discussed below.

Request for proposal (RFP) is issued when the buyer has a sense of the marketplace and has a statement of work containing a set of "performance" requirements that the buyer needs fulfilled. For example, the RFP may describe a formed part with certain strength, flexibility, and fire-resistance requirements, but not specify the particular composition of the material. Suppliers respond to the RFP with details on how they would satisfy the buyer's performance requirements and the price they would be willing to accept to do so. Upon learning the supplier's proposed pricing, the buyer may revise its requirements and/or negotiate exact terms with suppliers. Thus, the process is generally iterative. An RFP is appropriate for procurement of items that are nonstandard or highly complex, requiring supplier input and expertise about the best way to meet the requirements set forth in the RFP.

Request for quote (RFQ) is issued when the buyer can develop a statement of work that states the exact specifications of the goods or service needed. This is the case, for example, if the buyer seeks a part made of a particular plastic and formed to a specific set of thickness, density, and shape specifications. RFQs are often used in conjunction with highly structured competitive tendering processes. Typically there is no need for detailed negotiations with suppliers after bid receipt, as lowest price or some other objective criterion is used to evaluate bids. Owing

to their up-front specification requirements, RFQs are appropriate for procurement of items that are standard and well known in the marketplace. For example, in the electronics industry this would include commodity components such as cables, connectors, and circuit boards.

CONTRACT TERMS

The supplier selection process culminates in a contract between the buyer and one or more suppliers. The information received from suppliers via the requests described in the section titled “Information Requests to Suppliers” must ultimately be translated into formal contractual terms before contracting can occur. A contract with a supplier specifies what the supplier should do and how they will be paid by the buyer. At the highest possible level, contract terms relate to either monetary transfers (payment terms) or how the contract will be executed (nonpayment terms). Contracts can specify any number of payment and nonpayment arrangements. A few common ones are listed here to provide the reader with a sense of what types of contract terms the buyer might consider during negotiations and when making a contract award decision. The choice of the particular contract structure (long-term or short-term, fixed-cost or cost-plus, etc.) is beyond the scope of this article, but procurement contracts are covered in detail in the section titled “Vehicle Routing and Scheduling”.

Payment Terms. In a fixed-price contract, the price term specifies what the supplier will be paid regardless of the actual cost to execute its contractual obligations. In a cost-plus contract, a formula is specified, which determines how much the supplier will be paid; for example, under a cost-plus contract the supplier could receive a fixed percentage (e.g., 107%) of the total cost incurred or simply receive payments for time and materials. Various payments can also be specified as contingent on certain actions by the supplier (these can also take the form of penalties). Examples include a payment made only upon delivery or upon maintaining a certain target

inventory service level, an award fee granted for meeting budget targets in a cost-plus contract, an award fee for completing the project within a certain timeline, and so on. Liquidated damages clauses [11] can be used to specify an amount that either the buyer or supplier must pay to the counterparty upon breaching the contract.

Nonpayment Terms. The contract can specify all kinds of details related to how the contract will be executed, for instance, delivery quantities, delivery frequencies, delivery locations, service level, quality level, technical specifications, duration of the contract, and so on. Contracts where goods must be transported typically assign “incoterms” [12] defining the precise point at which the buyer takes control of the shipment (and hence the associated costs and risks).

NEGOTIATION PROCESS

As we will discuss in the section titled “Supplier Evaluation and Contract Award,” when making contract award decisions the buyer considers each supplier’s qualifications as well as the contract terms they offer (e.g., price). A supplier’s qualifications are generally considered exogenous, for example, a supplier’s reputation is based on historical performance and is not alterable in the short term. Contract terms, on the other hand, can be “negotiable” between the buyer and supplier. In a negotiation the buyer attempts to induce favorable terms from suppliers, and likewise, the suppliers attempt to induce favorable terms from the buyer. There are many different possible negotiation processes. This section overviews a few canonical negotiation processes, but a detailed discussion is reserved for the sections titled “Game Theory” and “Game Theory: Extension and Development” of this encyclopedia. For convenience we adopt the viewpoint of a buyer when discussing negotiations.

For better or worse, negotiations are often viewed as zero-sum games where the buyer gains what the supplier gives up. An extreme example of this is the *take-it-or-leave-it offer* approach whereby a powerful buyer essentially dictates the terms to the suppliers. For

instance, the buyer might demand a certain price and simply refuse to consider the supplier unless they agree to this price.

Take-it-or-leave-it offers are rather draconian, and buyers may be reluctant to utilize them for short-term gains if suppliers perceive them as unfair. Furthermore, take-it-or-leave-it offers require the buyer to credibly commit to not renegotiate with the supplier should the supplier choose to reject the buyer's offer. If the buyer cannot make such a commitment, the threat imputed in a take-it-or-leave-it offer is meaningless.

Competitive tendering is an alternative way to extract concessions from suppliers whereby suppliers are played off one another. Typically, suppliers simultaneously submit bids (in response to an RFP or RFQ). Competitive tendering approaches differ in the amount of visibility that suppliers have regarding competitors' bids. At one extreme is the dynamic open-descending bid format. In this format, suppliers see all bids submitted and can respond by lowering their own bid, until all but one bidder has dropped out [typically bidding lasts an hour or so, [13]]. At the other extreme is the sealed-bid format in which each bid is known only to the buyer and the supplier who submitted it. The US government's competitive tendering is typically done through sealed bidding.

It is also possible that the buyer can utilize neither competition nor take-it-or-leave-it offers. Instead, the buyer and a single supplier might *bargain* in some general and unstructured way. Negotiation processes in practice may combine take-it-or-leave-it offering, competitive tendering, and bargaining. For instance, the buyer could employ price-based competitive tendering with a reserve price (the reserve price imposes an upper bound on the amount the buyer is willing to pay for the contract and thereby acts like a take-it-or-leave-it offer) to home in on the most promising supplier, then bargain with this supplier to finalize the contract terms.

Negotiations do not always take a zero-sum approach. The buyer and supplier can potentially both benefit if they realize that their incentives are aligned rather than in conflict. Research to help buyers and

suppliers realize shared interests has led to numerous advances in software-enabled "expressive bidding" in combinatorial auctions; see *The Graph Model for Conflict Resolution*. For instance, in transportation auctions for truckload procurement, both the shipper (buyer) and the carrier (supplier) benefit if the shipper's lanes up for bid complement the carrier's existing transportation networks in a way that minimizes empty truck movements (e.g., see Ref. 14 and references therein).

SUPPLIER EVALUATION AND CONTRACT AWARD

This section describes how the buyer evaluates suppliers, determines the contract winner(s), and performs follow-up monitoring to inform future supplier selections. Supplier evaluation is the process by which the buyer rank orders the suppliers, as described in the section titled "Supplier Evaluation." The buyer then uses this rank ordering, along with other business considerations, to determine which supplier(s) will be awarded the contract, as described in the section titled "Contract Award." Finally, after contract award the buyer can monitor supplier performance and use this information during future supplier selection processes as described in the section titled "Supplier Monitoring."

Supplier Evaluation

The buyer begins the supplier evaluation process by identifying the "dimensions" it wishes to use when evaluating suppliers. Thanaraksakul and Phruksaphanrat [15] surveyed 76 papers on supplier selection in the purchasing literature and found that price, quality, and delivery were the most commonly listed supplier evaluation dimensions. Additional dimensions are also used. Thanaraksakul and Phruksaphanrat [15] provide an extensive list of such dimensions, categorized by prevalence in the purchasing literature. Frequently appearing dimensions include production capacity and flexibility, technical capabilities and support, information and communication systems, financial

status, and innovation and R&D. Dimensions that appear with moderate frequency in the literature include quality systems, management and organization, personnel training and development, performance history, geological location, reputation and references, packaging and handling ability, amount of past business, warranties and claim policies, procedural compliance, attitude and strategic fit, labor relations record, and desire for business. Of course, buyers often employ new dimensions in response to prevailing business issues and challenges. Dimensions that have emerged recently include environmental and social responsibility, safety awareness, domestic political stability, cultural congruence with the buyer organization, and terrorism risk. See Ref. 15 and the literature cited therein for more details.

Once suitable dimensions are identified, the ability to rank order suppliers is crucial for reaching an informed supplier selection decision. Rank ordering is simple when supplier bids are differentiated by a sole dimension such as price. This might be the case, for example, if the buyer has issued an RFQ for a highly standardized component delivered in a certain quantity by a certain date and suppliers are asked to respond with their price for the contract. However, rank ordering suppliers becomes complex when bids must be evaluated across multiple dimensions. For example, if the buyer wishes to evaluate suppliers' bids on the dimensions of price *and* lead time, the buyer must construct a trade-off between these two dimensions to determine whether it prefers, say, a bid with a high price and short lead time to a bid with a low price and long lead time. The challenge of supplier evaluation lies in constructing this trade-off in a way that accurately reflects the buyer's preferences.

Research addressing this challenge dates back by decades. A detailed review is beyond the scope of this article but is provided in Ref. 16. A variety of methodologies are proposed in this literature, but the overall approach is usually one of three types. The first approach is to capture the trade-offs with a simple relationship between the dimensions, for example, a linear function of the dimensions. This is simple, transparent,

and is often used in practice for these reasons. However, the simplicity of this approach has a downside: the buyer can find it difficult to construct weights that truly reflect its preferences. The second approach is to go beyond simple weighting in order to reflect the buyer's preferences with more detailed models. For instance, the buyer can value lead time using inventory models from ORMS and then trade this valuation off against the price offered by the supplier. This approach is less transparent, which makes it more difficult to implement in practice. Success has been enjoyed, particularly in domains in which the trade-offs are relatively clear but how to evaluate them is not—for example, a multiproduct buy in which suppliers can produce subsets of items and offer quantity discounts over the entire order. In such cases, OR tools like mathematical programming can be quite beneficial. Finally, a third approach recognizes that buyers at times find it difficult to articulate their preferences in a quantitative way and answers this difficulty by proposing ways to infer quantitative preferences over attributes by observing the buyer's qualitative choices between options involving these attributes.

Contract Award

Once the buyer has a sound methodology for evaluating suppliers, the process of contract awarding can begin. During this phase the buyer determines which supplier or suppliers to award a contract to. Supplier evaluation is a key ingredient in this process, but award decisions can hinge on more than just how the buyer evaluates the supplier.

For example, even if suppliers are closely matched, the buyer may choose to award the contract to just one of them. *Sole award* contracting may be favorable if the scope of work is best accomplished by a single supplier. For example, the contract may require significant capital investments on the part of the supplier and/or buyer, creating strong economies of scale. Sole-award contracting may also be used if it is unduly costly or risky to deal with multiple suppliers. For example, the buyer may be sourcing an item with intellectual property value (e.g., fabricating a

proprietary part) and need to closely monitor the supplier to prevent leakage of this intellectual property. A buyer outsourcing sensitive back-office operations (e.g., processing of client data) may need to invest a tremendous amount to train its supplier to ensure robust security measures.

Likewise, even if one supplier dominates another, the buyer might choose to give business to both of them. *Multiple-award* contracting can be useful if the buyer wishes to diversify its supply sources to mitigate disruption risks, or if suppliers have insufficient capacity or reverse economies of scale. There are also more strategic reasons for multi-sourcing. For example, a buyer might wish to prevent any supplier from becoming a monopolist, meaning it is the only viable supplier for a particular good or service needed by the buyer. This would happen, for example, if all the supplier's competitors exited the market because of bankruptcy. A buyer facing a monopolist supplier cannot leverage competition. To avoid this fate, the buyer may award contracts to several suppliers to keep them solvent and thereby encourage their continued presence in future contract competitions.

In general, there are many considerations which might tip the scales in favor of one supplier or another, as evidenced by the long list of potential supplier evaluation dimensions provided in the section titled "Supplier Evaluation." The buyer might deliberately favor incumbent suppliers to foster trust and loyalty or, for example, to avoid the administrative costs of training a new supplier on the buyer's invoicing and payment procedures. Supplier location may also be a concern in a way not manifested in logistics costs. For instance, if the buyer's potential customers are governments, these customers may be more likely to purchase the buyer's products when the government's home suppliers benefit from the purchase. The buyer might also consider supplier diversity objectives when making award decisions. If the buyer organization has supplier diversity goals, preference is given to historically underrepresented or disadvantaged businesses, such as small businesses, minority- and women-owned

businesses, sheltered workshops and non-profit organizations, and so on. Supplier diversity is typically not legally mandated in private industry, and the amount of preference the buyer grants to underrepresented or disadvantaged businesses in any one supplier selection event is typically situation dependent. However, specifically mandated preferences may apply to government contracts [17].

Regardless of which award criteria are used by the buyer, making such criteria transparent makes it easier for the buyer organization to monitor its contract award decisions, to ensure the reasons for contract award are sound (e.g., due to the merits of the bid). For example, a "low price wins" rule makes it difficult for procurement managers to "cheat" the buyer organization by negotiating a sweetheart deal with a supplier in return for a bribe. For these and other reasons, the US government, for example, generally favors competitive negotiations for procurement, with clearly announced rules for determining the winner [17].

Supplier Monitoring

Many contracts specify the provision of goods over an extended duration of time, ranging from weeks to years. Monitoring supplier performance during the life of the contract has several aims. For example, it supports quality if the buyer inspects incoming goods to ensure that they conform to quality specifications. Monitoring also supports cost containment—if there is a problem with quality, it can be identified and charged back to supplier. For supplier selection itself, however, monitoring is most important in so far as it helps the buyer make more informed supplier selections in the future.

In particular, during supplier evaluation the buyer may consider factors that influence the total cost of doing business with the supplier. Such costs can include, for example, the conformance and nonconformance costs, which the buyer anticipates will be incurred during the life of the contract (e.g., costs of inspections and defect correction, respectively). These fall under the supplier evaluation dimension of "past performance" listed in the section titled

“Supplier Evaluation.”) The buyer may forecast these costs for each supplier—see Ref. 18 for examples of how this is done in practice. These forecasts can be constructed using historical performance data collected through supplier monitoring. For instance, the supplier’s historical percentage of defective items can inform the buyer’s forecast for nonconformance costs during the life of a contract. [If, on the other hand, the supplier is new and thus the buyer’s protocols require more careful inspection of incoming material (conformance costs), this also needs to be taken into account by the buyer at the time of supplier evaluation.] Historical information about supplier performance can also be leveraged during the negotiation process with suppliers. The buyer may choose to directly incorporate this information into a competitive bidding process via a bid markup or some other means to send a clear signal to the supplier about the importance of performance [18].

SUPPLIER SELECTION RESEARCH

Extant research on supplier selection can be divided into two broad streams. The first stream dominates the purchasing literature and identifies appropriate criteria and methods supporting supplier evaluation. The primary goals are to help the buyer decide what its objectives are, what the dimensions to evaluate suppliers over are, and how to evaluate suppliers using these dimensions. The second stream assumes that the buyer knows what it wants and has an existing methodology for evaluating suppliers. It focuses on decisions such as what types of negotiation formats or contracts to employ and how to elicit information that suppliers may be reluctant to reveal.

ORMS research on supplier selection typically falls into the second stream of research (although powerful OR tools such as math programming are also very useful in the first stream). This research typically employs parsimonious models of the supplier evaluation process, providing clear objectives for the suppliers and buyer. As such, these models often most closely align

with RFQ processes for the award of a well-specified contract, for which supplier evaluation issues are relatively transparent. Typically the perspective taken is that of a buyer seeking to award a contract to one or more competing suppliers possessing private information about their cost to fulfill the contract. The competitive negotiation process is often modeled as an auction or auction-like procedure—see the section titled “Game Theory: Extension and Development”. To ensure a consistent model of behavior, typically all actors (buyer and suppliers) are assumed to be fully rational—see the section titled “Non-Cooperative Extensive-Form (NCEF) Games of Perfect Information.”

For illustrative purposes, in the next section we provide examples of ORMS models applicable to supplier selection. We then conclude this section, and the article, with examples of supplier selection research topics that employ such ORMS models; see the section titled “Examples of ORMS Supplier Selection Research.”

Examples of Supplier Selection Models in ORMS

As mentioned above, ORMS supplier selection models often focus on RFQ settings. The typical setting studied is one in which it is relatively easy to evaluate supplier bids, but the supplier has private information about what contract terms (e.g., payment size) it would be willing to accept. For example, the buyer would like to find the lowest-cost supplier, and each supplier i knows its cost to perform the contract is x_i , while the buyer only has a prior belief that the cost follows some probability distribution F .

Auctions are common in practice and are also convenient for modeling. They are often invoked as a structured way to model the act of soliciting information (bids) from suppliers and negotiating contract terms. An auction is no more than a set of rules determining how suppliers will pass information to the buyer, and how the buyer will use this information in determining the contract award winner(s) and payment(s).

Once auction rules are combined with assumptions on the suppliers’ behavior, the model can be used to predict the contract

winner and payment. This perspective is quite useful for analyses, as it accommodates questions such as “how will different auction formats affect the contract winner and payments?” For example, suppose all bidders are fully rational. If the auction uses a second-price format, it is a dominant strategy for bidders to continue bidding down until they either win the auction or reach their true cost. This implies that for n suppliers the auction winner will be the bidder $j = \arg \min_{i=1,\dots,n} \{x_i\}$ and the expected contract price is $E[X_{2:n}]$, where $X_{2:n}$ is the second-lowest order statistic of n draws from distribution F .

Thus far we have discussed buyer beliefs, auctions rules, and supplier behavior, comprising the core ingredients of many ORMS supplier selection models. Additional embellishments to the core model can be added to address specific research questions. For illustrative purposes we will next briefly describe a model of supplier qualification screening.

With an auction model, an ORMS researcher can answer questions such as “what is the effect of an additional bidder on the expected contract payment?” Clearly, the expected contract payment decreases with the number of suppliers, raising the natural question of how many suppliers the buyer should invite to its auction. From the section titled “Identifying Potential Suppliers,” we know that identifying suitable suppliers is costly for the buyer. This motivates ORMS supplier selection models that incorporate costs associated with supplier qualification screening. Qualification screening can be thought of as a process whereby the buyer incurs a cost to learn information about the supplier’s qualification. The following model discussion is from Ref. 19, which to our knowledge is the first paper in the ORMS and economics literature to model supplier qualification screening.

The qualification screening model begins by parameterizing qualification requirements. To this end, define a continuum of qualification requirements, with 0 representing requirements that every supplier satisfies and 1 representing requirements that virtually no supplier satisfies. Let $q_0 \in [0, 1]$

be the scalar along this continuum that represents the buyer’s preaward requirements. For each supplier i , define its qualification level q_i as the maximum qualification threshold that supplier i can pass. Owing to opaque qualification requirements set by the buyer (as discussed in the section titled “Supplier Qualification Screening Process,” e.g., gaining rapport with the buyer’s internal customer), supplier i does not precisely know its true qualification level q_i , but the buyer and supplier share the common belief that q_i is distributed according to probability distribution H . This setup could model, for instance, a buyer deciding to outsource a portion of its production currently done in-house, facing new suppliers it knows little about and who in turn know little about the buyer’s (possibly idiosyncratic) qualification requirements. The strictness of the buyer’s preaward requirements is captured by $1 - H(q_0)$, which equals the probability that $q_i \geq q_0$.

Under this model, the buyer and each supplier responding to the RFQ are equally unsure of the supplier’s qualification until costly qualification verification is undertaken by the buyer. If $q_i \geq q_0$, the buyer’s qualification process on supplier i would reveal that supplier i is qualified. The cost the buyer would incur to do so is denoted by K , the total cost to the buyer of verifying that an individual supplier meets all requirements to be deemed qualified. For example, K may include the cost of purchasing and testing supplier products, travel to supplier facilities abroad, and so on. For simplicity, this model assumes that K is the same for all suppliers. This assumption is most appropriate when suppliers are similar, at least in terms of the cost drivers of qualification, such as distance from the buyer or number of units that the buyer must purchase to run a test sample.

On the other hand, if $q_i < q_0$, supplier i would be rejected during the buyer’s qualification process after failing to meet a requirement. In this latter case, what cost would the buyer incur? Assuming that the requirements are nested (passing a larger threshold implies passing a smaller threshold, but not vice versa), the cost strictly increases with q_i and approaches K as q_i approaches q_0 . One may assume that the cost is linear and

normalized such that a threshold of 0 costs 0 to verify, implying that weeding out an unqualified supplier i costs the buyer q_i/q_0K . This is without loss of generality, because any nonlinear and strictly increasing cost function of q_i over $[0, q_0]$ can be renormalized to be linear by redefining q_i and renormalizing the distribution H . This completes the supplier qualification screening model.

This section provided a glimpse of ORMS supplier selection research models. The next section provides examples of ORMS supplier selection research topics which could be explored with such models.

Examples of ORMS Supplier Selection Research

As the above discussion in the present article indicates, there are many complexities and trade-offs involved in supplier selection, providing many opportunities for ORMS research. The following lists just a few of the complex trade-offs that need to be understood by the buyer, illustrating the types of questions that can be addressed by ORMS research.

Timing of Supplier Qualification. A buyer will typically only contract with a supplier who has passed qualification screening, but to cast a wide net a buyer may choose to consider bids from suppliers who have not yet passed qualification screening. Delaying some or all supplier qualification screening until after bids have been tendered can save time and money wasted on qualifying suppliers who are unable to offer competitive pricing. Industrial (i.e., nongovernment) buyers generally do consider bids even if the supplier is not fully qualified already [4]. The optimal timing of supplier qualification screening processes and price negotiations is studied in Ref. 19.

Scope of Negotiations (RFPs). Future ORMS research can be expected to inform practitioners about how best to include evaluation criteria that are currently considered to be too complex or subjective. These criteria will, with more advanced research, be quantified and subject to analytical comparisons. In industry, efforts at total-cost modeling date back by decades [18]. One goal for supplier selection might be to refine the total-cost

or life-cycle-cost analysis, so that it is more readily useable during negotiations with suppliers. To this end, ORMS research has studied so-called multiattribute negotiation mechanisms (also called *multiattribute auctions*, [20,21] and expressive bidding mechanisms (see **Quantum Game Theory** on combinatorial auctions), and many supplier evaluation methodologies have been developed in the purchasing literature [16].

Negotiation Formats. There is a wide body of economics literature that seeks to determine which rules of negotiation (or which “mechanism”) will work well in which situation. Owing to many complexities and the variety of situations encountered in supplier selection, a natural opportunity exists to study which rules of negotiation work well in supplier selection. For example, the buyer’s preference between total-cost open-bid and sealed-bid auction formats can be driven by how dissimilar the suppliers are in terms of their nonprice factors such as historical conformance and nonconformance costs [22].

Supply base Design. When suppliers are concentrated in the same region, the buyer is vulnerable to cost risks such as spikes in transportation costs between the buyer’s location and the supplier region, for example, due to a strike at the port of origin. Choosing suppliers in different regions lessens the correlation between suppliers’ total costs, but this helps the buyer only if she is able to prevent windfall profit-taking by low-cost suppliers. Wan and Beil [23] studied the relationship between the buyer’s bargaining power (ability to prevent suppliers from taking windfall profits) and the optimal level of diversification of the supply base across regions.

Additional research topics include understanding supplier collusion and methods to mitigate its effects, and research refining the typical assumptions of ORMS supplier selection research (such as laboratory tests of bidder rationality in auctions).

The present article provides a very basic introduction to the goals, complexities, and terminology of supplier selection. Interested readers are encouraged to refer to the section titled “Further Reading” for articles that can

serve as a point of departure for further study into supplier selection.

REFERENCES

1. U.S. Census Bureau. Statistics for industry groups and industries: 2005. Technical Report M05(AS)-1: U.S. Census Bureau; Nov 2006. Annual Survey of Manufactures.
2. Center for Advanced Purchasing Studies. Cross-industry metric report. Technical report. 31, Oct 2008.
3. Reinecke N, Spiller P, Ungerman D. The talent factor in purchasing. McKinsey Q 2007;(1):6–9.
4. Hedderich F, Giesecke R, Ohmsen D. Identifying and evaluating Chinese suppliers: China sourcing practices of german manufacturing companies. Practix 2006;9:1–8.
5. Lunsford JL, Glader P. Boeing's nuts-and-bolts problem; Shortage of fasteners tests ability to finish dreamliners. Wall Street J 2007;A8.
6. Spencer J, Casey N. Toy recall shows challenge China poses to partners. Wall Street J 2007;A1.
7. Welch D. Made in China: Faulty tires; How a Chinese supplier's bad decision turned into one importer's worst nightmare, and may mean the end of his business. Jul 12 2007. Available at http://www.businessweek.com/bwdaily/dnflash/content/jul2007/db20070711_%487422.htm. Accessed 2009.
8. Myers R. Food fights. CFO Magazine 2007 Jun.
9. Schmit J, Weise E. Three firms indicted in pet-food recall case. USA Today 2008 Feb 6.
10. Jensen C. Recalls of Chinese auto parts are a mounting concern. New York Times Wheels Blog. 2008. Available at <http://wheels.blogs.nytimes.com/2008/12/19/recalls-of-chinese-auto-parts-are-a-mounting-concern/>. Accessed 2009.
11. Corbin AL. Corbin on contracts. Newark, NJ: Matthew Bender & Company, Inc. 2007.
12. International Chamber of Commerce. INCOTERMS rules. Available at <http://www.iccwbo.org/incoterms/id3040/index.html>.
13. Ghawai D, Scheider GP. New approaches to online procurement. Proc Acad Inf Manage Sci 2004;8(2):25–28.
14. Chen R, *et al.* Solving truckload procurement auctions over an exponential number of bundles. Transp Sci 2009;43:493–510.
15. Thanaraksakul W, Phruksaphanrat B. Supplier evaluation framework based on balanced scorecard with integrated corporate social responsibility perspective. In: Proceedings of the International MultiConference of Engineers and Computer Scientists; 2009 Mar 18–20, Volume II. Hong Kong; 2009.
16. de Boer L, Labro E, Morlacchi P. A review of methods supporting supplier selection. Euro J Purch Supply Manage 2001;7:75–89.
17. Federal Acquisition Regulations. Available at <http://www.arnet.gov/far>. Accessed 2009.
18. Ellram LM. Total cost modeling in purchasing. Tempe, AZ: Center for Advanced Purchasing Studies; 1994.
19. Wan Z, Beil DR. RFQ auctions with supplier qualification screening. Oper Res 2008;57:934–949.
20. Beil DR, Wein LM. An inverse-optimization-based auction mechanism to support a multiattribute rfq process. Manage Sci 2003;49(11):1529–1545.
21. Parkes DC, Kalagnanam J. Models for iterative multiattribute procurement auctions. Manage Sci 2005;51(3):435–451.
22. Kostamis D, Beil DR, Duenyas I. Total-cost procurement auctions: Impact of suppliers' cost adjustments on auction format choice. Manage Sci 2009;55:1985–1999.
23. Wan Z Beil DR. Bargaining power and supply base diversification. Working paper 2008 Oct.

FURTHER READING

Single-source models

- *Basic auction theory.*
Krishna V. Auction theory. San Diego, CA: Academic Press; 2002. pp. 1–31.
- *Mechanism design.*
Myerson RB. Optimal auction design. Math Oper Res 1981;6:58–73.
- *Multiattribute auctions.*
Che YK. Design competition through multidimensional auctions. RAND J Econ 1993;24:668–680.

Multiple-source models

- *Split-award contracts.*
Anton JJ, Yao DA. Split awards, procurement, and innovation. RAND J Econ 1989;20: 538–552.

- *Quantity flexible contracts.*

Dasgupta S, Spulber DF. Managing procurement auctions. *Inf Econ Policy* 1990;4:5–29.

- *Entry fees.*

Seshadri S, *et al.* Multiple source procurement competitions. *Mark Sci* 1991;10:246–263.

- *Repeated sourcing events.*

Rob R. The design of procurement contracts. *Am Econ Rev* 1986;76:378–389.

Competitive bidding and moral hazard

- *Bidding competition and risk sharing.*

McAfee RP, McMillan J. Bidding for contracts: a principal-agent analysis. *RAND J Econ* 1986;17:326–338.

- *Cost-plus versus price-only contracts.*

Bajari P, Tadelis S. Incentives versus transaction costs: a theory of procurement contracts. *RAND J Econ* 2001;32:387–407.

SUPPLY CHAIN COORDINATION

FUQIANG ZHANG
Olin Business School, Washington
University, St. Louis, Missouri

A supply chain is a network of firms/entities that convert raw materials and components into final products and then deliver to consumers. If the supply chain is managed by a central planner who is able to control all decisions, then we call it a *centralized supply chain*. The set of actions that can optimize the supply chain's performance is called the *centralized optimal solution*. In practice, however, it is not uncommon for parties within a supply chain to be independent organizations that aim to maximize their own objectives. In this case, we call it a *decentralized supply chain*. The behavior of a decentralized supply chain can be characterized using the *Nash* or *Stackelberg equilibrium* concept (see ***Nash Equilibrium (Pure and Mixed)***).

Achieving *coordination* is equivalent to achieving the centralized optimal solution for the supply chain.¹ Under the centralized optimal solution, the pie for the entire system is maximized, so each player in the system can enjoy a larger size of the pie. This implies that under supply chain coordination, the players can achieve *Pareto optimal* profit allocations, that is, it is not possible to increase the profit of one player without hurting the profit of the other. Clearly, a coordinated supply chain is an ideal situation, since all players can be better off than without coordination. Unfortunately, since the players have different objectives and interests, the equilibrium outcome in a decentralized supply chain

often deviates from the centralized optimal solution, thus creating inefficiencies in the supply chain.

How to coordinate a decentralized supply chain is a central issue in supply chain management. One solution is to design a contractual scheme to align the incentives among all players so that the optimal actions coincide individually with the centralized optimal actions. In fact, many commonly observed contracts can serve as coordination devices. This article provides an introductory review on how to achieve the goal of supply chain coordination using such contractual arrangements.

A real-world supply chain could be extremely complex. For example, it may have an intricate network structure involving an arbitrary number of firms, each firm may have private information about its own cost and demand forecast, and the actions taken by firms may not be observable or verifiable. As an illustrational example, throughout this article, we consider a simple supply chain consisting of only two firms—a supplier and a retailer. The retailer is modeled as a newsvendor, that is, she faces a random demand in a single selling season and decides on the quantity to order from the supplier (see ***Newsvendor Models***). (We use “she” for the retailer and “he” for the supplier throughout this article.) Everything is common knowledge and there are no hidden actions. Also, all players are risk-neutral; so, they try to maximize their expected profits. We demonstrate how several commonly observed contracts can be used to coordinate this simple supply chain. We also provide intuitive explanation for the reason why these contracts can realign the incentives of the supply chain members. For a more comprehensive treatment of general supply chain coordination problems, readers are referred to Cachon [2].

This article is organized as follows. The section titled “Basic Model” introduces the basic model and the concept of *double marginalization* in a decentralized supply

¹This is a typical definition of supply chain coordination in the research literature. A more general definition is that all supply chain members take actions together to increase (not necessarily optimize) total supply chain performance. See Chopra and Meindl [1] for more details and practical examples.

chain. The section titled “Supply Chain Coordination Contracts” proposes several contracts to coordinate the supply chain. The section titled “Additional Issues” briefly discusses some additional issues in supply chain coordination. And, this article is concluded in the final section.

BASIC MODEL

Consider a supply chain consisting of a supplier and a retailer. The retailer sells a product in a single selling season at a fixed price p . Market demand D is random and has a distribution (density) function $F(f)$. The retailer procures the product from the supplier. The supplier incurs a cost c for each unit of product delivered to the retailer. Owing to long production and transportation lead times, the retailer has to decide on the order quantity before the market demand is realized.² In the case where the demand is less than the order quantity, leftover inventory at the end of the selling season has a unit salvage value $0 \leq v < c$. (Note that there are several additional parameters that we may add into the basic model. For instance, there could be a penalty cost for unsatisfied demand, the retailer may have to incur a unit cost as well for selling the product. However, the introduction of these parameters will not affect the understanding of the coordination contracts. So, we omit these parameters for ease of exposition.) For notational convenience, let E denotes the expectation operation and also, defines $x \wedge y = \min(x, y)$, $x^+ = \max(0, x)$, and $\bar{F} = 1 - F$.

²This supply chain setting (i.e., a supplier selling through a newsvendor retailer) is standard in the operations literature. In such a setting, the retailer faces a fixed retail price but uncertain market demand. Alternatively, we can model the retailer as a price-setting firm but with deterministic demand. The latter model setting has been widely used in studying channel coordination in the marketing literature. Although it is not the focus of this article, interested readers are referred to Jeuland and Shugan [3] and Moorthy [4] for more details on how to coordinate supply chains with deterministic market demand.

Centralized Optimal Solution

We first present the centralized optimal solution for the above supply chain. Latter we will use this solution as a benchmark for comparison. When the supplier and the retailer are controlled by a central planner, the retailer can obtain the product from the supplier at cost c . So, the supply chain’s problem reduces to the classic newsvendor problem. Observe that the supply chain’s profit is solely determined by the retailer’s order quantity. Let Q be the order quantity. Then the supply chain’s expected profit is given by

$$\begin{aligned}\pi_c(Q) &= pE(Q \wedge D) + vE(Q - D)^+ - cQ \\ &= (p - v)E(Q \wedge D) + (v - c)Q.\end{aligned}\quad (1)$$

We use subscript c for centralized supply chain. It is straightforward to show that $\pi_c(Q)$ is concave (i.e., $\pi_c''(Q) < 0$). Thus, the following first-order condition characterizes the profit-maximizing quantity Q_c^* for the centralized supply chain

$$\left. \frac{d\pi_c(Q)}{dQ} \right|_{Q=Q_c^*} = (p - v)\bar{F}(Q_c^*) + (v - c) = 0$$

or

$$F(Q_c^*) = \frac{p - c}{p - v}.\quad (2)$$

Equation (2) is the critical fractile solution to the classic newsvendor problem. Note that the centralized optimal solution Q_c^* increases in price p and salvage value v and decreases in cost c .

Wholesale Price Contract

Next, we consider the behavior of a decentralized supply chain. That is, now the supply chain members are independent entities that try to maximize their own objectives. We use a wholesale price contract to highlight how a decentralized supply chain may deviate from the centralized optimal solution. Under the wholesale price contract, the supplier charges a unit price $w > c$ for each unit of product delivered to the retailer. Assume that the wholesale price is exogenously given, that is, the wholesale price contract has already

been established at the outset. The sequence of events is as follows: The retailer places an order at the supplier; the supplier produces the product and delivers to the retailer; then market demand is realized and sales begin; finally, the selling season ends and leftover inventories are salvaged. Below we analyze the behavior of the supply chain under the wholesale price w .

The retailer's problem under the wholesale price arrangement is the same as in the centralized supply chain, except that the procurement cost is w rather than c . Given an order quantity Q , the retailer's expected profit now becomes

$$\begin{aligned}\pi_r(Q) &= pE(Q \wedge D) + vE(Q - D)^+ - wQ \\ &= (p - v)E(Q \wedge D) + (v - w)Q.\end{aligned}\quad (3)$$

We use subscript r for the retailer. Similarly, $\pi_r(Q)$ is concave and the first-order condition for the retailer's profit-maximizing order quantity Q_w^* is given by

$$F(Q_w^*) = \frac{p - w}{p - v}, \quad (4)$$

where the subscript w stands for wholesale price.

By comparing the Equations (2) and (4), we can see that $Q_w^* < Q_c^*$, since F is an increasing function and $w > c$. That is, under the wholesale price contract, the retailer tends to order less than under the

centralized optimal solution. Such a result can be explained using the standard risk analysis in the newsvendor problem. Since market demand is uncertain, the retailer essentially bets against demand when making an ordering decision. If demand realization is high, then the retailer loses $p - c$ potential profit for each unit of unmet demand. This is known as *underage cost* for a newsvendor. On the other hand, if demand realization is low, then the retailer incurs a loss of $c - v$ for each unit of leftover inventory. This is called *overage cost* for a newsvendor. It is quite intuitive that all else being equal, the retailer's optimal order quantity increases in the underage cost and decreases in the overage cost. Under the wholesale price contract, the underage cost decreases to $p - w$, while the overage cost increases to $w - v$, which leads to a lower order quantity at the retailer.

An alternative explanation of the above under-ordering result emerges from the marginal analysis (see Ref. 4 for the use of marginal analysis in channel coordination under deterministic demand). Under centralized control, the retailer's (i.e., the supply chain's) expected profit in Equation (1) can be divided into two parts: the revenue part, $pE(Q \wedge D) + vE(Q - D)^+$, and the cost part, cQ . The retailer's optimal order quantity is achieved at the point where the marginal revenue equals the marginal

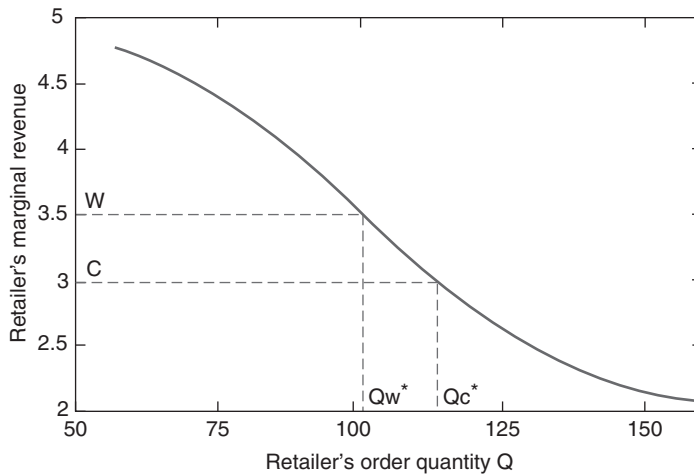


Figure 1. Retailer's marginal revenue as a function of order quantity Q ($c = 3$, $p = 5$, $v = 2$, $w = 3.5$, D is normal with $N[100, 30^2]$).

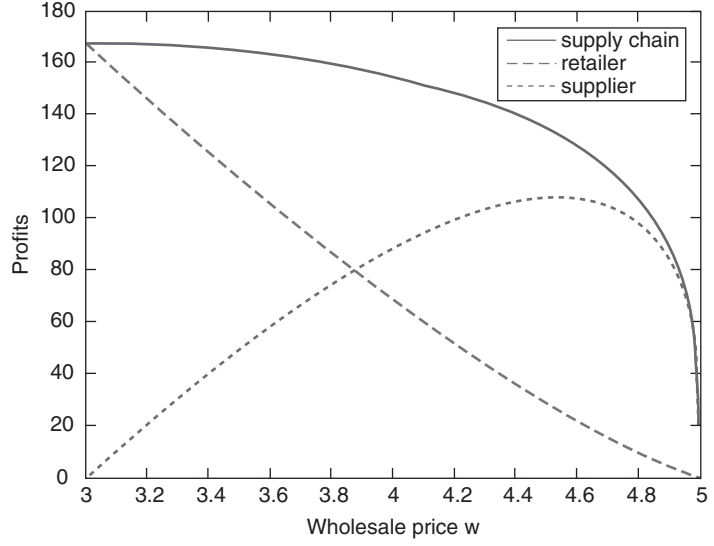


Figure 2. Supply chain members' profits as functions of the wholesale price w ($c = 3, p = 5, v = 2, D$ is normal with $N[100, 30^2]$).

cost. Note that under the wholesale price contract, the retailer's revenue function in Equation (3) is the same as the supply chain's revenue function. This means that the retailer's marginal revenues are the same in both cases. However, the retailer's marginal cost increases from c (in the centralized supply chain) to w (in the wholesale price contract). Figure 1 uses a numerical example to visualize how the retailer's optimal order quantity is determined. In this example, $c = 3, p = 5, v = 2$, and the demand D follows a normal distribution with mean 100 and standard deviation 30. In a centralized supply chain, the retailer's optimal order quantity Q_c^* is determined by equalizing the marginal revenue and the marginal cost $c = 3$. Similarly, in a decentralized supply chain, the retailer's optimal order quantity Q_w^* is determined by equalizing the marginal revenue and the marginal cost $w = 3.5$. We can see that a higher marginal cost leads to a lower order quantity for the retailer (i.e., $w > c$ leads to $Q_w^* < Q_c^*$).

Since the supply chain profit depends on the retailer's order quantity, it means that the wholesale price contract does not maximize the supply chain's performance. This is the so-called *double marginalization* problem: The wholesale price exists because

the supplier requires a unit profit margin $w - c$ on top of that for the retailer, which in turn leads to a lower order quantity by the retailer and creates inefficiencies in the supply chain. Such an effect was first identified in Ref. 5. For the same numerical example described above, Fig. 2 shows the supplier's, the retailer's, and the supply chain's expected profits as functions of the wholesale price w . As we can see, the supply chain's profit curve declines as the wholesale price increases. That is, a more severe double marginalization problem leads to higher efficiency loss for the supply chain. The retailer's profit also decreases in the wholesale price, which is intuitive. The supplier's profit curve is slightly different. It increases first and then decreases in wholesale price. This is because while a higher wholesale price means a higher unit profit margin, it also forces the retailer to order less. So, the supplier's profit is not monotone in wholesale price.

The double marginalization problem implies that more sophisticated contractual arrangements (relative to the wholesale price contract) are needed to induce the retailer to order the centralized optimal solution. From the above marginal analysis, we essentially need to design a contractual scheme in which

the retailer's marginal revenue and marginal cost curves intercept at the supply chain's optimal solution, Q_c^* . We show in the next section how to achieve this goal via carefully designed contractual arrangements.

SUPPLY CHAIN COORDINATION CONTRACTS

The above analysis has shown that the wholesale price contract induces the retailer to order less than the supply chain optimal quantity. We need to increase the retailer's order quantity in order to achieve coordination. In this section, we show that several commonly observed contracts in practice can achieve coordination by manipulating the retailer's marginal revenue and/or marginal cost curves. To save space, we focus on the following three contract formats: buyback, quantity discount, and revenue-sharing contracts.

Buyback Contract

A buyback contract contains two parameters (w, b) . The retailer pays a wholesale price w to the supplier; at the same time, the supplier agrees to buy back the leftover inventories at a buyback price b ($v < b < c$). (If $b \leq v$, then the buyback is useless because the retailer is better off salvaging the leftover inventories by herself; if $b \geq c$, then the retailer will order infinity.) This is also known as *return policies*. For example, in the book industry, many publishers offer buyback terms to retailers, that is, unsold books can be returned to the publisher at the end of the selling season for a certain price. Notice that the supplier does not have to physically take the products back in a buyback contract; rather, he only needs to subsidize the disposed excess stock. Pasternack [6] was among the first to study buyback contracts in the context of the newsvendor problem. Intuitively, buyback price b improves the retailer's marginal revenue from salvaging the leftover products (i.e., from v to b) and thus should increase her order quantity, and a large enough b should bring the retailer's order quantity back to the centralized optimal solution. (Equivalently, we may use the underage and overage cost

concepts to explain why buyback contracts work. In particular, the buyback policy can induce the retailer to order more because it reduces the retailer's overage cost from $c - v$ to $c - b$.) Next, we confirm this conjecture.

Under the buyback contract, the retailer's expected profit can be written as

$$\begin{aligned}\pi_r(Q) &= pE(Q \wedge D) + bE(Q - D)^+ - wQ \\ &= (p - b)E(Q \wedge D) + (b - w)Q,\end{aligned}\quad (5)$$

which is concave in Q . Through manipulation, we can derive the first-order condition for the retailer's profit-maximizing order quantity, Q_b^*

$$\left. \frac{d\pi_r(Q)}{dQ} \right|_{Q=Q_b^*} = (p - b)\bar{F}(Q_b^*) + (b - w) = 0,$$

which gives

$$F(Q_b^*) = \frac{p - w}{p - b}. \quad (6)$$

We use subscript b for buyback. To coordinate the supply chain, we need $Q_b^* = Q_c^*$, or

$$\frac{p - w}{p - b} = \frac{p - c}{p - v} \quad (7)$$

by comparing Equations (2) and (6). Thus, any combination of values for (w, b) that satisfy Equation (7) can coordinate the supply chain. Note that there are an infinite number of solutions to Equation (7). Now the question is: Which one should we use to coordinate the supply chain? It turns out that different solutions of (w, b) correspond to different profit allocations between the supplier and the retailer. To see this, let

$$w = p - \phi(p - c) \text{ and } b = p - \phi(p - v), \quad (8)$$

where $0 \leq \phi \leq 1$ is a constant. It is easy to verify that Equation (7) holds, so the pair (w, b) given in Equation (8) coordinates the supply chain. Further, the retailer's expected profit in Equation (5) can now be written as

$$\begin{aligned}\pi_r(Q) &= \phi(p - v)E(Q \wedge D) + \phi(v - c)Q \\ &= \phi\pi_c(Q).\end{aligned}\quad (9)$$

That is, the buyback contract in Equation (8) not only coordinates the supply chain, but also assigns a portion ϕ ($0 \leq \phi \leq 1$) of the supply chain profit to the retailer (the supplier gets a portion of $1 - \phi$). The ability to achieve arbitrary allocation of profits between supply chain members is a desirable property. In many cases, the supply chain members have unequal bargaining powers and as a result, they may require different profit allocations in order to participate in coordination. From the above analysis, it is clear that buyback contracts can serve both purposes well (i.e., coordination and arbitrary profit allocation). It follows from Equation (8) that a lower ϕ corresponds to a higher b , which implies that a higher buyback price leads to a lower retailer profit. This is mainly because a higher buyback price is associated with a higher wholesale price w .

Finally, there are some variations of the buyback contract in practice. For instance, it is common that fashion manufacturers use the so-called *markdown money* to compensate the retailers for liquidated items via clearance pricing. In a markdown money contract, the markdown money can be a percentage of the price markdown, or it can be a rebate credit for each unit of salvaged product. Both forms of the markdown money have the same effect on the retailer's decision as the buyback term in the buyback contract. Another related example is quantity flexibility (or backup agreement). In a quantity flexibility contract, similarly, the supplier charges a wholesale price but subsidizes the retailer for unsold units. To be specific, the supplier is fully responsible for a portion of the retailer's order (note that the supplier provides partial protection on the retailer's entire order in the buyback contract). See Refs 7 and 8 for more discussion of quantity flexibility contracts.

Quantity Discount Contract

The buyback contract coordinates supply chains by altering the retailer's marginal revenue curve. Similarly, to achieve coordination, we may manipulate the retailer's marginal cost curve so that it crosses the marginal revenue curve precisely at Q_c^* , the order quantity that is optimal for the supply

chain. Consider the following so-called *quantity discount* contract: The supplier charges the retailer a unit wholesale price $w(Q)$, which is a function of the order quantity Q . This is also called an *all-unit* quantity discount, because the wholesale price applies to the entire order quantity. By definition of the quantity discount, we focus on $w'(Q) < 0$, that is, wholesale price functions that are decreasing in order quantity. This type of arrangement is quite common in practice, where a buyer enjoys a lower price by buying more (e.g., many consumer packaged goods suppliers routinely offer incentives to induce large order quantities from their retailers). Next, we analyze this quantity discount contract.

Under this contract, the retailer's expected profit is given by

$$\begin{aligned}\pi_r(Q) &= pE(Q \wedge D) + vE(Q - D)^+ - w(Q)Q \\ &= (p - v)E(Q \wedge D) + (v - w(Q))Q.\end{aligned}\quad (10)$$

Similar to the analysis of buyback contracts in the section titled "Buyback Contract," we try to express the retailer's profit function $\pi_r(Q)$ as a fraction of the supply chain's profit. Let

$$\begin{aligned}w(Q) &= (1 - \phi)(p - v) \left(\frac{E(Q \wedge D)}{Q} \right) \\ &\quad + \phi(c - v) + v,\end{aligned}\quad (11)$$

where $0 \leq \phi \leq 1$ is a constant. Plugging the above $w(Q)$ into Equation (10) gives

$$\begin{aligned}\pi_r(Q) &= \phi(p - v)E(Q \wedge D) + \phi(v - c)Q \\ &= \phi\pi_c(Q).\end{aligned}\quad (12)$$

Hence, under the quantity discount scheme in Equation (11), the retailer's incentive is aligned with the supply chain's because $\pi_r(Q) = \phi\pi_c(Q)$. The retailer keeps ϕ ($0 \leq \phi \leq 1$) portion of the supply chain profit.

Since

$$\begin{aligned} \frac{d\left(\frac{E(Q \wedge D)}{Q}\right)}{dQ} &= \frac{\bar{F}(Q)Q - E(Q \wedge D)}{Q^2} \\ &= -\frac{\int_0^Q xf(x)dx}{Q^2} < 0, \end{aligned}$$

we know $w(Q)$ is a decreasing function of Q , that is, $w(Q)$ represents a quantity discount scheme. Note that $w(Q) \rightarrow \phi(c - v) + v \leq c$ as $Q \rightarrow \infty$, which means that the wholesale price the supplier charges could be lower than the cost c . However, this will never be the retailer's optimal choice because the supplier's, and therefore all parties' profits are negative when the order quantity is excessively large.

From the given Equation (12), we know that any order quantity that maximizes the retailer's profit will also maximize the supply chain's profit. Essentially, it implies that, by using a properly designed quantity discount contract, we can achieve supply chain coordination and arbitrarily allocate profits between the supply chain members.

Revenue-Sharing Contract

So far we have introduced two coordination contracts: buyback contracts and quantity discount contracts. In buyback contracts, a buyback term is used to improve the retailer's marginal revenue curve and thus induce the retailer to take the supply chain optimal action. In quantity discount contracts, a nonlinear wholesale price function is used to manipulate the retailer's marginal cost curve so that it intersects with the marginal revenue curve at the supply chain optimal quantity. What happens if we use certain contractual schemes to adjust both the marginal revenue and marginal cost curves? In this section, we study revenue-sharing contracts that can achieve supply chain coordination by manipulating both the retailer's marginal revenue and marginal cost curves. The practice of revenue-sharing has been observed; for instance, in the video rental industry, where studios offer movie copies to rental stores at relatively cheap prices, and

as a compensation, the rental stores agree to share revenue with the studios.

Let $\{w, \lambda\}$ denote a revenue-sharing contract, where w is the wholesale price and λ is the revenue share (i.e., the portion of the supply chain's expected revenue) obtained by the retailer. To be more specific, the retailer pays the supplier w for each unit of product, and meanwhile, the retailer keeps λ portion of the revenue from either selling or salvaging each unit of product. Under revenue sharing, the retailer's expected profit is given by

$$\begin{aligned} \pi_r(Q) &= \lambda[pE(Q \wedge D) + vE(Q - D)^+] - wQ \\ &= \lambda(p - v)E(Q \wedge D) + (\lambda v - w)Q. \end{aligned} \quad (13)$$

Consider the following parameters values for the revenue-sharing contract:

$$w = \phi c \text{ and } \lambda = \phi, \quad (14)$$

where $0 \leq \phi \leq 1$ is a constant. Then the retailer's profit becomes

$$\begin{aligned} \pi_r(Q) &= \phi(p - v)E(Q \wedge D) + \phi(v - c)Q \\ &= \phi\pi_c(Q). \end{aligned} \quad (15)$$

Thus, the revenue-sharing scheme in Equation (14) achieves coordination and assigns a portion of ϕ ($0 \leq \phi \leq 1$) profits to the retailer. Note that in a coordinating revenue-sharing contract, there is $w = \phi c \leq c$, which implies that the supplier should charge a wholesale price that is not greater than his cost. Therefore, instead of charging a markup when selling to the retailer, the supplier makes profits solely from sharing the revenue with the retailer.

It is worth mentioning that we may go one step further to coordinate a supply chain via profit sharing. That is, both firms agree to take actions to maximize the supply chain's objective, and then split the total profits between themselves. Another related coordination device is the so-called *two-part tariff* contract. In this type of contract, the supplier delegates all decisions to the retailer (i.e., the retailer serves as a central planner for the entire supply chain), and in return,

the retailer compensates the supplier using a lump-sum payment. Both the profit-sharing and the two-part tariff notions are essentially equivalent to vertical integration, where supply chain firms merge into one. Although intuitive, it may not always be appropriate to propose the use of vertical integration as a coordination solution. In fact, there could be numerous factors that may prevent firms from easily integrating with each other in a supply chain (e.g., firms would like to gain control of their own decisions, there may be additional administration costs due to more complex organizational structure, and different firms may be endowed with different amounts of information, etc.). More discussion of the comparison between revenue-sharing and other coordination contracts can be found in Cachon and Lariviere [9].

ADDITIONAL ISSUES

The previous section proposes three different supply chain coordination contracts (i.e., buyback, quantity discount, and revenue-sharing). All three contracts are able to induce supply chain optimal actions and allocate profits between firms in a flexible way. Although they seem to work equally well, they have different implications in practice. In particular, some contracts might be relatively easier to implement than others under different situations. A detailed discussion of the advantages and the disadvantages of the coordination contracts is beyond the scope of this short article. Moreover, we have illustrated the three contracts using the simplest supply chain setting. The most practical situations are much more complex than the one described in the previous sections. As a result, the above three contracts may or may not be able to coordinate these more complex supply chains. In this section, we briefly explain the additional issues that may complicate our supply chain coordination analysis. The purpose is to provide a preliminary introduction of more realistic problem settings, but not to give a thorough solution to these new coordination problems. Nevertheless, we list the representative references that address these additional issues.

Multiple Periods

The basic model in the section titled “Basic Model” considers a single-period problem in which the retailer makes an ordering decision only once. The single-period model is appropriate for perishable products that have short life-cycles (e.g., books, fashion items, high-tech products). For products with longer life spans, a multiperiod model might be more reasonable. In this case, the retailer can regularly replenish inventory from the supplier, and the supply chain may become a two-echelon inventory system (when the supplier makes to stock) or a production-inventory system (when the supplier makes to order). Cachon and Zipkin [10] and Caldentey and Wein [11] study how to coordinate these two different systems, respectively.

Multiple Retailer Decisions

There is a single decision in the basic model, that is, the retailer’s order quantity. Coordination is relatively easy because one only needs to make sure the retailer adopts the supply chain’s optimal order quantity. However, in practice, there are many situations where the retailer needs to make more than one decision. One of the common decisions made by a retailer is sales effort. Imagine that a retail store influences demand by using promotional displays, offering shopping assistance, and giving out gift cards and coupons. Taylor [12] and Krishnan *et al.* [13] study how to coordinate a supply chain in which the retailer makes both ordering and sales effort decisions. In particular, they investigate the so-called *sales rebate* contract where the supplier charges a wholesale price and at the same time compensates the retailer with a rebate for each unit sold beyond a certain target sales level. Another decision a retailer often makes is the retail price. How to coordinate a price-setting newsvendor? It turns out that the buyback contract does not work in this case in general. However, it can be shown that the revenue-sharing contract can still coordinate such a supply chain (see Ref. 9 for details).

Other Supply Chain Structures

A real-world supply chain may consist of a number of connected firms with various network structures. From a research point of view, we need to first understand the following three building-block structures: serial, assembly, and distribution supply chains. In a serial supply chain, there is only a single firm at each stage. The basic model studied in the section titled “Basic Model” is a typical example. An assembly supply chain consists of one manufacturer at the downstream stage, but there could be multiple suppliers at the upstream stage. The manufacturer uses the complementary components delivered by the suppliers to assemble a final product. To coordinate such an assembly system, one needs to consider both the vertical relationship (between the suppliers and the manufacturer) and also the horizontal relationship (between the suppliers). Bernstein and DeCroix [14] and Zhang [15] study the assembly inventory systems under periodic review and propose coordination schemes accordingly. Distribution supply chains are also commonly seen in practice. For example, a supplier sells his product through multiple independent retailers. Recent research that addresses the coordination problem for distribution supply chains can be found in Chen *et al.* [16], Bernstein and Federgruen [17], Narayanan *et al.* [18], and Krishnan and Winter [19].

Asymmetric Information

When supply chain members are independent organizations, it is natural that they are endowed with different types and amounts of information about the supply chain. For example, the supplier may possess private information about production cost, while the retailer may have superior information about market demand. Two types of asymmetric information have been widely investigated in the supply chain management literature: cost and demand information. When asymmetric information is present, supply chain coordination is not always achievable. This is because supply chain coordination requires information sharing among the supply chain firms, but truthful information sharing may

not be in the interest of the information holders. Chen [20] provides a review of the information-sharing literature in supply chain management (see *Information Sharing in Supply Chains*).

Strategic Customer Behavior

The traditional supply chain management literature models customer demand at an aggregate level and do not consider forward-looking customer behavior. This is clearly a simplification because demand is composed of individual consumers who may consciously react to market conditions. In the above basic model, there is a selling price p in the regular selling season, and then there is a salvage price $v < p$ in the markdown season. If a consumer anticipates a lower price after the regular season, this may affect her purchasing decision in the regular season. Specifically, if she waits, she may get the same product at a lower price, but she has to face the risk of a stockout after the regular selling season. How does this kind of consumer behavior affect supply chain management, and in particular, supply chain coordination? A recent research development is to explicitly incorporate the consumer side into the picture. Su and Zhang [21] study the impact of such strategic customer behavior on supply chain performance. Interestingly, they demonstrate that under the presence of strategic consumer behavior, a decentralized supply chain with a wholesale price contract can perform strictly better than a centralized supply chain. The reason is that a wholesale price (and thus double marginalization) actually helps the retailer to credibly commit to a low order quantity, which in turn induces the consumers to be willing to pay a high price in the regular season. This implies that the centralized optimal solution is not always the best benchmark for evaluating supply chain performance. Under strategic customer behavior, such a benchmark can be surpassed.

CONCLUSIONS

A decentralized supply chain consists of independent firms that wish to maximize their own objectives. As a result, the firms’ actions

often deviate from the actions that optimize the supply chain's performance. This article provides an introductory review on how firms can use various contractual schemes to achieve supply chain coordination. Three commonly observed contracts have been analyzed: buyback, quantity discount, and revenue-sharing. It has been shown that for a simple, two-stage supply chain, all these contracts can realign the firms' incentives to induce supply chain optimal actions. In addition, by properly choosing the contract parameters, we can achieve arbitrary profit allocations among supply chain firms.

For illustrational purpose, in this article we focus on a simple supply chain setting. Additional issues that complicate the coordination analysis may exist. We also provide a brief discussion of these issues. Some of the issues have already been addressed in the supply chain management literature, but some of them deserve further research attention. For example, the topic of information sharing has not been fully explored. Also, strategic customer behavior opens a new and promising research arena in supply chain management. Finally, there has recently been a fast-growing interest in studying the effect of supply uncertainties on firms' sourcing decisions. How to coordinate a supply chain with unreliable supply is an interesting topic for future research, too.

REFERENCES

1. Chopra S, Meindl P. Supply chain management: strategy, planning, and operations. 3rd ed. Upper Saddle River (NJ): Prentice Hall; 2007.
2. Cachon GP. Supply chain coordination with contracts. In: Graves S, de Kok T, editors. Handbooks in operations research and management science. Amsterdam, the Netherlands: Elsevier; 2003.
3. Jeuland A, Shugan S. Managing channel profits. *Market Sci* 1983;2(3):239–272.
4. Moorthy S. Managing channel profits: comment. *Market Sci* 1987;6:375–379.
5. Spengler J. Vertical integration and antitrust policy. *J Pol Econ* 1950;58:347–352.
6. Pasternack B. Optimal pricing and return policies for perishable commodities. *Market Sci* 1985;4(2):166–176.
7. Eppen GD, Iyer AV. Backup agreements in fashion buying—the value of upstream flexibility. *Manag Sci* 1997;43(11):1469–1484.
8. Tsay A. Quantity-flexibility contract and supplier-customer incentives. *Manag Sci* 1999;49:1339–1358.
9. Cachon GP, Lariviere M. Supply chain coordination with revenue sharing: strengths and limitations. *Manag Sci* 2005;51(1):30–44.
10. Cachon GP, Zipkin PH. Competitive and cooperative inventory policies in a two-stage supply chain. *Manag Sci* 1999;45(7):936–953.
11. Caldentey RA, Wein L. Analysis of a decentralized production-inventory system. *Manuf Serv Oper Manag* 2003;5(1):1–17.
12. Taylor T. Supply chain coordination under channel rebates with sales effort effects. *Manag Sci* 2002;48(8):992–1007.
13. Krishnan H, Kapuscinski R, Butz DA. Coordinating contracts for decentralized supply chains with retailer promotional effort. *Manag Sci* 2004;50(1):48–63.
14. Bernstein F, DeCroix GA. Inventory policies in a decentralized assembly system. *Oper Res* 2006;54(2):324–336.
15. Zhang F. Competition, cooperation, and information sharing in a two-echelon assembly system. *Manuf Serv Oper Manag* 2006;8(3):273–291.
16. Chen F, Federgruen A, Zheng Y-S. Coordination mechanisms for a distribution system with one supplier and multiple retailers. *Manag Sci* 2001;47(5):693–708.
17. Bernstein F, Federgruen A. Decentralized supply chains with competing retailers under demand uncertainty. *Manag Sci* 2005;51(1):18–29.
18. Narayanan VG, Raman A, Singh J. Agency costs in a supply chain with demand uncertainty and price competition. *Manag Sci* 2005;51(1):120–132.
19. Krishnan H, Winter RA. Vertical control of price and inventory. *Am Econ Rev* 2007;97(5):1840–1857.
20. Chen F. Information sharing and supply chain coordination. In: Graves S, de Kok T, editors. Handbooks in operations research and management science. Amsterdam, the Netherlands: Elsevier; 2003.
21. Su X, Zhang F. Strategic customer behavior, commitment, and supply chain performance. *Manag Sci* 2008;54(10):1759–1773.

SUPPLY CHAIN OUTSOURCING

ANDY A. TSAY

OMIS Department, Leavey School of
Business, Santa Clara University,
Santa Clara, California

This article focuses on the outsourcing of manufacturing/production/assembly, procurement/sourcing, logistics, and product design/development. The first three in the list are traditionally considered to be the major supply chain functions, to which we add the fourth to acknowledge that design decisions strongly preordain the supply base and the manufacturing processes. Also, in modern practice, the insource-versus-outsource decisions for design and manufacturing are often intertwined. Alternatively this scope can be viewed as the endeavor of stewarding a product from concept to market, and then operating the resulting supply chain.

Since the outsourcing of a supply chain activity is a special instance of business process outsourcing (BPO), readers should first review the BPO article in this encyclopedia (see ***Business Process Outsourcing***). Similar to that entry, this article is intended to serve as a tutorial, and will not provide a comprehensive review of the research literature. The BPO article ends with a high-level assessment of extant operations research and management science (ORMS) research on general outsourcing that applies to supply chain outsourcing research as well.

We will discuss the outsourcing of each of the stated activities, with the caveat that they do not segment cleanly into distinct and sequential steps. Because of cross-functional coordination issues, the resolution of the insource-versus-outsource quandary for each activity depends on how the others are conducted. Further, the service providers are increasingly blending these into a package deal for their customers.

Supply chain outsourcing is of tremendous interest in the business community. While

the potential benefits are alluring, the magnitude of the risks is illustrated by recent high-profile crises faced by Cisco [1], Boeing [2], Mattel [3], and Menu Foods [4]. Most of those included offshoring as well, but the problems could be attributed primarily to the incentive conflicts and loss of visibility symptomatic of delegation to external providers of services.

OUTSOURCING OF MANUFACTURING

For many, outsourcing in the supply chain setting first brings to mind the treatment of manufacturing/production/assembly. The most basic form is the purchase of a standard material from an outside party, for which the proper control processes are generally well-understood and efficient. The need for managerial concern grows when the buyer obtains more complex manufacturing services from a service provider, such as when a brand owner engages a contract manufacturer to produce a noncommodity product using nonstandard processes.

The following terms are used by practitioners to characterize the key players in such settings:

OEM: original equipment manufacturer

OBM: original (or own) brand manufacturer

CM: contract manufacturer

The acronym OEM has classically referred to a party that makes and sells a branded product, but (somewhat anachronistically) continues to be applied to the brand owner even if the “M” is performed by another party. However, such a firm may decide “to OEM” a component, which means that the procured part will retain the supplier’s brand identity. In this usage, the term *OEM* refers to the supplier. Emerging from the electronics industry around 2004 [5], the OBM label is used to affirm that the brand owner is also performing the manufacturing. A CM manufactures products that ultimately

bear another party's brand, and traditionally does not own the intellectual property of the design.

This article will frame outsourcing's consequences from the OEM's perspective, since the OEM controls the decision of whether to outsource in the first place. Supply chain outsourcing inherits the motives for general outsourcing (see *Business Process Outsourcing*), but here the OEM's highest priorities are typically to avoid ownership of the factory workforce, assets and infrastructure, and to tap into specialized expertise that can quickly ramp to a cost-efficient and robust stream of supply. The increasing availability of a competent CM segment serving virtually every product category makes this possible. This is a boon to start-ups which may lack the capital for funding internal production (or elements of design, logistics, etc.). The lowering of these entry barriers has profoundly affected the competitive dynamics in many industries.

The general risks also persist, especially for the sourcing of complex or nonstandard manufacturing services rather than manufactured product. Context-specific concerns include the CM's counterfeiting the product, stealing proprietary processes, or simply not following the expected procedures. Even without deliberate malfeasance, coordination problems arise from inserting company boundaries between manufacturing and other key OEM functions such as product design or sales/marketing.

The coordination between design and manufacturing that is required to generate manufacturable designs (i.e., design for manufacturability, DFM; see *Design for Manufacturing and Assembly*) and accurate bills-of-materials is already elusive when the two functions are under the same roof. Separating into different companies only adds additional obstacles, such as incompatibilities in data formats, materials nomenclature, or part-numbering conventions.

OEM sales and marketing managers can struggle to monitor quality and manufacturing status when the manufacturing has been outsourced, particularly when products are spread across multiple CMs. These managers face frustrations trying to respond

to end-customer questions like "Where is my order and when will it ship?" or "Can I still change the configuration?" [6].

These points make apparent that the performance of any outsourcing strategy will depend on the needs for flow of information along partners in the resulting extended enterprise. This underlies the argument that product architecture drives supply chain architecture. A "modular" architecture includes a one-to-one mapping from functional elements to components, and specifies decoupled interfaces between components. An "integral" architecture includes a complex (non-one-to-one) mapping from functional elements to components and/or coupled interfaces between components [7]. Because decomposability reduces the need for communication, the writing of detailed specifications, and iteration in designing the parts for which each party is responsible, modular products (e.g., personal computers) tend to be built (and designed) by modular supply chains (heavy outsourcing; many suppliers for each component) while integral products (e.g., high-performance automobiles) tend to come from integral supply chains (little outsourcing; vertically integrated industry) [8]. In short, the supply chains can be "mix and match" only to the extent that the product components (and the associated business processes and IT platforms) are "plug and play."

A vast body of practitioner literature weighs in on the insource-versus-outsource question for manufacturing, and provides operational advice on engaging a CM. A recurring topic therein is the hidden costs of coordination. The need for cross-functional integration between design and manufacturing calls for CM involvement very early in the OEM's product design timeline, or even outsourcing design and manufacturing as a bundle to qualified CMs [9]. The consequences of the latter approach will be discussed in the section titled "Outsourcing of Product Design/Development." Information technology, in areas such as enterprise resource planning (ERP) or product life-cycle management (PLM), can also enhance the coordination across chasms created by outsourcing. Of course, technical connectivity can provide only limited benefit without compatible

data formats and business processes. These are the goals of standards such as the Partner-Interface-Protocol (PIPs) framework of RosettaNet (www.rosettanet.org) or the programs of the Voluntary Inter-Industry Commerce Solutions (VICS) consortium (www.vics.org) that include collaborative planning, forecasting, and replenishment (CPFR). These require a willingness to share information with partners and, to reiterate a recurring theme, nontrivial investments of human and financial resources.

For these reasons and others, insourcing of manufacturing remains a legitimate strategy. OEMs in various industries have affirmed their commitment to this approach [10,11].

OUTSOURCING OF PRODUCT DESIGN/DEVELOPMENT

While typically not as asset-intensive as manufacturing, maintaining superior product design capability can also be imposingly expensive, especially in the specialized human capital. The coordination difficulties noted earlier may also motivate the outsourcing of design in a way that reintegrates it with outsourced manufacturing activities. [Although theoretically possible, there is no evidence that any OEMs are outsourcing significant amounts of design while retaining manufacturing in-house [9].] Consequently, design outsourcing is a growing trend in many industries [12].

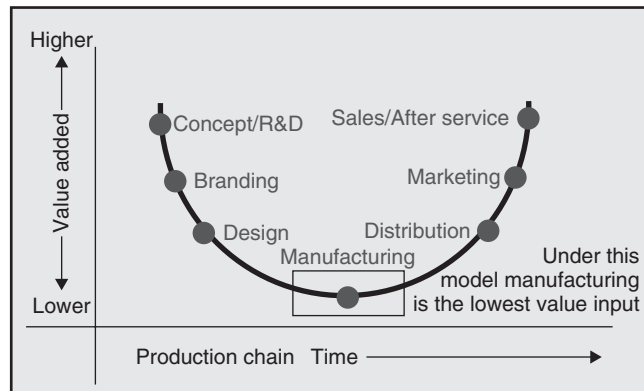
The term *design* can be applied to a vast range of activities, from generation of product concepts all the way to the creation of very precise schematics of product configuration. These pieces can be further subdivided, and any of the segments are candidates for outsourcing. Some firms even outsource the pursuit of breakthrough innovation (R&D) to specific organizations or to the open community. This could be broadened to consider innovation in processes or business models, not just designs of specific products. To invoke a recurring theme, design outsourcing is not a binary decision. Of course, a firm should be incessantly wary about outsourcing design activities that might fall among its core

competences. Indeed, in early 2009, amidst a broad economic crisis that had many of its rivals conducting massive layoffs, Apple Inc. was aggressively hiring specialized chip designers to work on key technologies. This reflected Apple's desire to get critical new features to market quickly while sharing fewer details about its technology plans with external chip suppliers [13].

Design outsourcing faces the challenges of services procurement explained in the article on BPO (see ***Business Process Outsourcing***). Communication and coordination across-company boundaries are particularly difficult because even though design can be performed much more systematically than most people think, it is still inherently a creative activity that entails working with ideas that are not fully developed, with many interdependencies among decisions. Further, the staff asked to liaison with the design service providers often, at least initially, lack the appropriate background and organizational support for the increased emphasis on project and relationship management [12,14].

Yet the stakes are high, especially for the many design decisions that have implications for manufacturing and supply chain management [9]. Among other impacts, design decisions strongly constrain the choice of materials, suppliers, and manufacturing processes. The folk wisdom is that roughly 70% or more of a product's life-cycle costs (manufacturing, supply chain, quality) are preordained at the design stage. (That design decisions are important is not in dispute, but the concrete evidence for this numeric rule of thumb is apocryphal. As Barton *et al.* [15] note, "Where authors support it by reference to published work, the references are to authors who themselves provide no substantive proof to support their claims. These referenced authors assert it themselves or quote a study by, or give a quote from, a major corporation. Examples of such corporations are Boeing, British Aerospace, General Electric, Rolls-Royce, Westinghouse, and Ford. With the exception of Rolls-Royce, the studies reportedly carried out by these major corporations cannot be easily traced as they are inadequately referenced, for example, as '...a Boeing

Figure 1. Smile Curve (Source: Version from <http://www.madeintaiwan.tv/blog/?p=10>).



study of turbine engines...’ or ‘...according to General Motors’ executives...’. Regarding Rolls-Royce, Barton *et al.* [15] emphasize that the widely quoted original study [16] actually concludes that good design decisions can reduce 80% of “unnecessary” costs rather than of total costs.)

In principle, one way to maximize the coordination between design and manufacturing decisions in an outsourced supply chain is to entrust both activities to the same service provider. The following two paradigms for such joint outsourcing have arisen in a variety of industries:

- CDM: contract (or custom or collaborative) design and manufacturing
 ODM: original (or own) design manufacturing (er)

The primary difference between CDM and ODM is in the ownership of the intellectual property (IP) in the product design. In the CDM model, the IP fully belongs to the OEM, whereas an ODM owns the product IP and is entitled to use it to create its own brands, products for other OEMs, generics, or white-box products. Additionally, ODMs take on inventory liability, while in the CDM model, the OEM generally owns the inventory. In these senses, a CDM is primarily a service company whereas an ODM is primarily a product company.

The historical evolution of both CDM and ODM reflect the reality that, in many industries and for many product types, manufacturing excellence has become a commodity. This is the notion that underlies the “(Stan

Shih) Smile Curve,” attributed to the founder of Acer who proposed it in 1992 [17]. One version of the framework appears in Fig. 1.

The Smile Curve originally addressed what semiconductor and electronic CMs (mostly Taiwan-based) experienced in the 1990’s, a phenomenon which persists in many industries today: the lion’s share of the wealth in the value chain seems to gravitate to the activities upstream and downstream of the manufacturing segment, so that the profitability of a traditional CM will be squeezed from both sides. This contributes another entry to the previous section’s list of OEM motives for outsourcing manufacturing, but here it conveys why the pure-play CM business model, even with its focus, economies of scale, and risk-pooling, seems so often to be a transient form on the way to either CDM or ODM.

Indeed, some traditional CMs have added CDM or ODM services in search of a sustainable means of differentiation, one able to command higher margins. CMs are thus able to enhance their manufacturing competitiveness by generating designs that are efficiently manufacturable.

CDM is less dramatic a transformation for a CM, so the documentation of this phenomenon seems to focus more on the ODM form. The majority of written coverage of ODM has been in the press surrounding the electronics industry. There the ODMs grew out of the motherboard companies in Taiwan, who moved into computer systems, especially notebook computers. A resulting misconception is that this organizational

format is unique to electronics. In fact, the model appeared earlier in bicycle manufacturing and possibly elsewhere. Apparel ODMs reside in various parts of Asia, especially Hong Kong and Korea.

OEMs are increasing their usage of both models. This suggests that the motives for manufacturing outsourcing are compelling, but that design and manufacturing need to go together. Beyond that the OEM still must choose a specific model, neither of which dominates the other.

An OEM benefits from the turnkey aspect of either solution, as CDM and ODM service providers both allow an OEM to tap into a supply chain possessing many of the benefits of vertical integration. An ODM provides a virtually complete product off-the-shelf to fill out an OEM's product portfolio, a fast solution that allows the OEM to reduce in-house R&D expense but with only limited customizability. However, the ODM needs to supply multiple OEMs to recover the cost of R&D for a platform and to mitigate the inventory risk. As noted, ownership of the IP entitles the ODM to supply the OEM's direct competitors, as well as to become a competitor itself via white-box or own-brand products.

This raises the ultimate question for the OEM, which is how to differentiate its brand when its ODM-supplied product is technically equivalent to many others on the market. ODMs are very protective of the identities of their clients for this reason, lest end-consumers start to view their products as a commodity. For mobile computers, a category which is predominantly supplied by Asian ODMs, OEMs such as Hewlett-Packard pursue differentiation through the industrial design of the chassis, the software bundle, and after-sale support.

ODMs have both the motivation (as indicated by the Smile Curve) and the prerogative to become OBMs. Indeed, the evolutionary path from CM to ODM to OBM is well-documented for firms in late-industrializing economies such as Taiwan, Korea, Singapore, and Malaysia [18,19]. Firms that started as CMs offering cheap labor learned design skills and market insights from sophisticated customers who needed to transfer this knowledge to preserve

DFM. Eventually these ODMs aspired to develop their own brands to increase their control and financial returns.

If an OEM does not design or manufacture, what competitive capabilities does it retain that are so hard for ODMs to learn or obtain? Earlier, the Taiwanese electronics ODMs had only limited success achieving sustained global brand awareness, forcing them to compete on price. Challenges include the need to develop sophisticated distribution channels with infrastructure for functions such as handling returns, offering credit, and providing warranty service, and the marketing expertise that provides the deep customer knowledge critical to the conception of attractive products. However in some cases, a CM or ODM can obtain these rapidly through an acquisition. In any event, an OEM outsourcing design to its manufacturing partner should proceed under the assumption that this partner will eventually become a direct competitor.

OUTSOURCING OF PROCUREMENT

Procurement already implies outsourcing, in that some good or service is being purchased from outside a firm. The buying firm has another outsourcing decision to make beyond that, regarding whether management of the procurement activity itself should also be entrusted to an outside party.

The outside party can be a focused procurement specialist, to which some apply the label "procurement service provider" (PSP). These offer not only capacity and expertise, but also assets such as supplier databases and technology for conducting reverse auctions.

Alternatively, the magnitude of the financial impact may present a valid business case for integrating the manufacturing and direct materials decisions. This approach can prevent myopic actions such as choosing the cheaper component in ignorance of the detrimental impact on assembly cost. "Turnkey" engagement of a CM, in which the CM is entirely responsible for providing the end product to the OEM, ostensibly produces tight integration with the least overhead

while retaining the benefits of manufacturing outsourcing.

CMs have financial incentives to take over the OEM's procurement of direct materials, for which they typically earn a percentage markup over the cost of materials. In addition, because financial analysts sometimes base certain metrics on revenue, the ability to count the flow of materials as revenue can elevate a CM's public stature. In the electronics sector, for example, materials can represent 75–80% of a CM's revenue. At the same time, competition has pressured the margins that CMs can earn solely by manufacturing. Some CMs now view direct manufacturing as a loss leader for driving business through the profit center that direct procurement has become.

A strategic view of procurement understands that control of the buying decision is a precious asset. The livelihood of any seller depends on making buyers happy, and sophisticated or large-scale buyers may be able to extract strategically significant treatment from suppliers. This preferential treatment can take the form of low prices (either straightforwardly or indirectly through rebates and other subsidies), short lead times, liberal return privileges, forgiveness of occasional contract noncompliance, assurance of supply in times of scarcity, influence over technology road-maps, technical support, and so forth. Thus, something more profound than markup on materials changes hands when an OEM outsources procurement; the OEM may cede control of the preferential treatment as well.

OEMs have a variety of options for retaining control of procurement while still outsourcing manufacturing. Between turnkey outsourcing, which exposes the OEM to a long list of hazards, and full insourcing of procurement, in which the OEM foregoes the potential benefits of outsourcing, are a number of procurement models that seek a compromise by incorporating various preventive and reactive business controls. Amaral *et al.* [20] summarizes the strengths and weaknesses of a number of such approaches:

In-House

With in-house procurement, OEMs buy directly from suppliers, managing storage

and transit to CMs. In electronics, OEMs began with this approach when they first outsourced production by providing prepackaged part kits to CMs for overflow assembly work. The OEM completely controls procurement in this way, which minimizes outsourcing risks.

Such control is costly. In-house procurement requires fully staffed organizations, highly integrated information systems, and distributed sites for planning, executing, and managing the inbound supply chain from suppliers to CMs. OEMs must stay abreast of technical developments and in contact with potential suppliers around the world. They must also maintain inventory storage locations (hubs) near the various CM assembly sites.

Turnkey

In the turnkey model, the CM negotiates with and buys directly from suppliers. Thus, the OEM can keep procurement overhead low while leveraging the CM's buying power and ability to break bulk, which can be a boon for small OEMs. Also the CM can pool the demand uncertainty of multiple OEMs to reduce the safety stocks. In principle, this efficiency should translate into low costs and high availability for the OEM.

The turnkey model carries many hazards, including forfeiture by the OEM of preferential treatment and loss of visibility into true procurement and material costs. For large OEMs and for noncommodity parts, the CM's procurement leverage will probably be weaker than the OEM's.

Turnkey with Audits

With this approach, the OEM retains the advantages of the turnkey model but adds auditing to detect errors and deter fraud, partially mitigating several hazards of pure turnkey. The OEM may perform the audits itself or rely on a specialist firm.

Through audits, OEMs can discover only a fraction of possible problems, may not gain full recovery of damages, and often lose the time value of money. They still forfeit preferential treatment and lose visibility into true procurement and material costs, while bearing the auditing expenses. When OEMs

believe they have greater procurement leverage than their CMs, they often choose from among the following procurement models.

Supplier Rebates

When OEMs believe they can negotiate superior prices and effectively monitor and collect private rebates from suppliers, they can obtain the same partial risk mitigation as they do with audits, because the information technology systems for tracking and collecting rebates essentially perform ongoing audits. In addition, suppliers can safely offer the OEM preferential pricing without revealing their prices to CMs and other OEMs. That is, this scheme achieves “price masking.”

The primary disadvantages of supplier rebates are the cost to the OEM for tracking and processing rebates and the costs to suppliers of negotiating prices separately with the CM and the OEM. Smaller OEMs and suppliers might find the administrative burden of the rebate scheme to be prohibitive. The CM is still ultimately the one paying the suppliers, and may use this role to enhance its own procurement leverage.

Buy–Sell

In the buy–sell model, the OEM buys directly from the supplier at a private price and immediately resells to the CM at a higher price. This achieves price masking and the benefits of maintaining direct OEM–supplier relationships. Once the buy–sell transaction is complete, the supplier delivers the materials directly to the CM.

The primary disadvantages for OEMs are the overhead required to manage procurement and any investment in systems and processes to enable buy–sell execution. In addition to maintaining supplier relationships, the OEMs must replicate the channel functions of a materials reseller.

Consignment

Consignment is an arrangement in which OEMs buy and own the inventory, which the CMs store. OEMs often use this model for parts that are unique, slow moving, proprietary, or scarce. OEMs can thus mask

prices and establish inventory buffers above those prescribed by the CMs’ standard policies. With consignment, OEMs are responsible for most of the procurement activities, reducing various risks but adding overhead costs.

An OEM need not limit itself to just one of these approaches, and many leading firms have developed the ability to execute many of them simultaneously. At the time of publication of Ref. 21, HP used buy–sell for strategic commodities (high value or coming from key suppliers), that is, the 20% of parts representing about half of its production spending. For the next 50% of parts, HP used audits (to verify pricing) and rebates (to mask pricing). HP allowed CMs to procure the remaining commodity parts in turnkey fashion.

OUTSOURCING OF LOGISTICS

Logistics, which many traditionalists view as the core activity of supply chain management, has a long history of outsourcing. A strong motive comes from the asset intensiveness of logistics (e.g., infrastructure for transportation, handling, storage, and increasingly, IT for real-time, global tracking at the individual package level of detail). At the same time, a strong enabler for the existence of logistics service providers is the fungibility of these assets across many customer and material categories. The value proposition of these service providers also includes an inventory dimension. Cycle stocks will decrease to the extent that clients feel relief from the need to completely fill containers or vehicles, greater delivery reliability reduces the need for safety stock, and pipeline inventories will be lower if the transit times are shorter. Service providers with an international footprint offer critical expertise in moving products through disparate physical, legal, and regulatory environments.

As with the broad supply chain functions discussed earlier, logistics can be divided into specialized elements, with labels such

as “freight forwarding,” “inbound logistics,” “warehousing,” “outbound logistics” (distribution), “service logistics” (for spare parts), and “reverse logistics.” All of these, as well as their smaller subtasks, are candidates for outsourcing. Some providers now offer nontraditional services such as light manufacturing, after-sale repair, packaging of products into store displays, and even financing of inventory.

Key terms for identifying the providers of logistics services are

3PL third-party logistics provider
 4PL fourth-party logistics provider
 LLP lead logistics provider

Lynch notes that “3PL” was first used in the early 1970s to identify intermodal marketing companies (IMCs) in transportation contracts [22]. Prior to that, transportation contracts involved only the shipper and the carrier. As intermediaries that accepted shipments from shippers and tendered them to rail carriers, IMCs became the third party to these contracts. Since then, the definition has broadened to refer to any company that offers a logistics service. “4PL” is generally attributed to Accenture, which registered it as a trademark in 1996 (while still known as Andersen Consulting). Accenture described the 4PL as an integrator, but today some consultants, software companies, and even 3PLs lay claim to being a 4PL. A 4PL can be thought to serve the function of general contractor. “LLP” is favored by some as a term with greater transparency than 4PL. Lynch [22] advocates using the following definitions.

3PL

According to the Council of Supply Chain Management Professionals, 3PL is a firm which provides multiple logistic services for use by customers. Preferably, these services are integrated, or “bundled” together by the provider. These firms facilitate the movement of parts and materials from suppliers to manufacturers, and finished products from manufacturers to distributors and retailers. Among the services which they provide are transportation, warehousing, cross-docking,

inventory management, packaging, and freight forwarding.

4PL

According to Accenture, 4PL is a supply chain integrator that assembles and manages the resources, capabilities, and technology of its own organization with those of complementary service providers to deliver a comprehensive supply chain solution.

LLP

Lynch defines LLP as a party that serves as the client’s primary supply chain management provider, defining processes, and managing the provision and integration of logistic services through its own organization and those of its subcontractors [22].

As with the other forms of supply chain outsourcing, a large body of practitioner literature provides advice on how to make the insource-versus-outsource decision for logistics and then how best to engage the service provider. In terms of the degree of difficulty of managing the outsourced relationship, logistics is comparable to design in being purely a service, but is comparable to manufacturing in the availability of relatively straightforward metrics such as delivery times and error rates.

THE OEM’S ROLE IN THE SUPPLY CHAIN OF THE FUTURE

In a world in which the ability to outsource nearly all supply chain functions have some joking that OEM must be an acronym for “Outsource Everything but Marketing,” one must wonder whether the OEM should still be considered the lead actor in the resulting ecosystem of codependent partners [23]. Some OEMs may evolve to a business model based on performing as supply chain “orchestrator” or “integrator,” such as the role played by the venerable Li & Fung or, increasingly in some of its electronics segments, Hewlett-Packard [24]. This resembles the function of the 4PL/LLP in logistics, but with a comprehensive scope.

Success in this role requires a core competence in managing complexity and designing mechanisms to contain transactions costs and moral hazards. As autocratic fiat does not exist in such extended enterprises, supply chain management takes on a political flavor in which negotiation and relationship-building prowess are the critical determinants of success [25]. How power, and hence profitability, will distribute across these supply chains will provide rich substrate for business research for years to come.

REFERENCES

- Berinato S. What went wrong at Cisco. CIO 2001 Aug 1.
- Lunsford JL. Boeing scrambles to repair problems with new plane. Wall St J 2007 Dec 7:A1.
- Casey N. Mattel issues third major recall. Wall St J 2007 Sep 5:A3.
- Schmit J. Tainted pet food suit settled for \$24 million. USA Today 2008 May 23.
- Pick A. The EMS and ODM monitor. iSuppli Corporation; 2004.
- Preysman V. Private supply chains. Electron News 2001;47(9):52.
- Ulrich K. The role of product architecture in the manufacturing firm. Res Policy 1995;24:419–440.
- Fine CH. Clockspeed-based strategies for supply chain design. Prod Oper Manage 2000;9(3):213–221.
- Ulrich KT, Ellison DJ. Beyond make-buy: Internalization and integration of design and production. Prod Oper Manage 2005;14(3):315–330.
- Reinhardt A. Nokia's magnificent mobile-phone manufacturing machine. Bus Week 2006 Aug 3.
- Kane YI. Sharp focuses on manufacturing. Wall St J 2008 Jul 9:B1.
- Anderson EG Jr, Davis-Blake A, Erzurumlu SS, *et al.* The effects of outsourcing, offshoring, and distributed product development organization on coordinating the NPD process. In: Loch C, Kavadias S, editors. Handbook on new product development management. Burlington (MA): Elsevier; 2007.
- Kane YI, Clark D. In major shift, Apple builds its own team to design chips. Wall St J 2009 Apr 30:B1.
- Anderson EG Jr, Davis-Blake A, Parker GG. Managing outsourced product design: The effectiveness of alternative integration mechanisms. UT Austin working paper; 2007.
- Barton JA, Love DM, Taylor GD. Design determines 70% of cost? A review of implications for design evaluation. J Eng Des 2001;12(1):47–58.
- Symon RJ, Dangerfield KJ. Application of design to cost in engineering and manufacturing. NATO AGARD Lecture Series No. 107: The Application of Design To Cost And Life Cycle Cost To Aircraft Engines; 12–13 May; Saint Louis, France, London, 15–16 May; 1980. pp.7.1–7.17.
- CompuTrade. Stan Shih, father of Taiwan branding: OEM and own-brand should be divided for better effectiveness. CompuTrade 2007 Jul 2.
- Hobday M. Innovation in East Asia: the challenge to Japan. London: Edward Elgar Publishing; 1995.
- Boulton W, Kelly M. Electronics manufacturing in the Pacific Rim. Baltimore (MD): WTEC; 1997.
- Amaral J, Billington CA, Tsay AA. Safeguarding the promise of production outsourcing. Interfaces 2006;36(3):220–233.
- Billington C, Kuper A. Trends in procurement: a perspective. In: ASCET 5. San Francisco (CA): Montgomery Research, Inc. (sponsored by Accenture); 2003. pp. 102–105.
- Lynch CF. Name your terms. DC Velocity 2005 Jul.
- Ojo B. Boxed. Electron Supply Manuf 2005;2(4):32–40.
- Anderson EG Jr, Parker GG. From buyer to integrator: the transformation of the supply chain manager in the vertically disintegrating firm. Prod Oper Manage 2002;11(1):75–91.
- Amaral J, Tsay AA. How to win 'spend' and influence partners: lessons in behavioral operations from the outsourcing game. Prod Oper Manage 2009;18(6):621–634.

SUPPLY CHAIN RISK MANAGEMENT

MANMOHAN S. SODHI
Cass Business School,
City University
London, London, UK

CHRISTOPHER S. TANG
UCLA Anderson School,
Los Angeles, California

In addition, there are *strategic performance metrics* such as profits, market share, revenue growth, return on assets (ROA), and share price performance. Because disruptions to the supply chain can play havoc with these and other metrics, risk management for the supply chain has become an important emerging area for practitioners and researchers.

SUPPLY CHAIN RISK MANAGEMENT (SCRM)

SUPPLY CHAIN MANAGEMENT

A *supply chain* is a network of organizations possibly including suppliers, manufacturers, logistics providers, wholesalers/distributors, and retailers that aims to produce and deliver products or services for the end customer. *Supply chain management* is the management of material, information, and financial flows through the supply chain. It includes the coordination and collaboration of processes and activities across different functions such as marketing, sales, production, product design, procurement, logistics, finance, and information technology within the supply chain.

Supply chain management includes tracking and seeking to improve *operational performance metrics* including

- time-related measures such as product development cycle time, time to market, production lead time, replenishment/delivery lead time, cash-to-cash cycle;
- cost-related measures such as product development cost, material cost, production cost, labor cost, inventory cost, shipping and handling cost, sales and marketing cost, fixed and overhead cost; and
- customer satisfaction-related measures such as product availability, product and service quality, reliability, after sales support, and total cost of ownership.

During the 1990s and later, many firms implemented various supply chain initiatives to (i) increase revenue, for example, more product variety, more-frequent new product introductions, more sales channels/markets and new markets, (ii) reduce cost, for example, supply base reduction, on-line sourcing including e-markets and on-line auctions, offshore manufacturing, just-in-time inventory systems, and vendor managed inventory, and (iii) reduce assets, for example, outsourcing of manufacturing, information technology and logistics. Supply chains today are more “efficient” but also more complex than those in the past. As a consequence, they are also more vulnerable to disruptions and delays with large unanticipated consequences of seemingly contained events [1].

Such disruptions can have long-term stock price effects and equity risk effects. On the basis of an analysis of 827 disruption announcements made over a 10-year period, Hendricks and Singhal [2] found that companies suffering from the occurrence of uncertain events experienced 33–40% lower stock returns relative to their industry benchmarks over a three-year time period starting one year before and ending two years after the date announcing the event. The impact need not be financial alone: in 2007, over 100 brands of tainted pet food contaminated with the toxic chemical melamine killed thousands of dogs and cats in the United States. The impact of such incidents has led to a

growing interest in the area of “supply chain risk” and its management, as evidenced in the number of practitioner conferences and consultancy reports devoted to the topic [3].

Many executives and researchers started paying more attention to responding to severe disruptions after the attacks on September 11, 2001 [4,5]. However, it is important to manage risks in the larger context of coordinating supply and demand under various types of uncertainty.

While it is tempting to think of supply chain risk management (SCRM) as extending supply chain management to include risk, it is really a multidisciplinary area with research and practice that draws on *supply chain management*, *enterprise risk management*, and *crisis management*. These domains in turn draw on a broad swathe of the academic literature in areas such as organizational behavior, psychology, decision analysis, empirical analysis, stochastic modeling, and mathematical programming.

A consensus view from two supply chain conferences—Supply Chain Thought Leaders (SCTL) meeting in Madrid, 2008 and the International Supply Chain Risk Management (ISCRiM) workshop in Trondheim, 2008—is to consider SCRM as a subset of both supply chain management and of enterprise risk management. SCRM can also be viewed as a subset of crisis management and business continuity.

However, these parent areas each have a primary purpose different from that of SCRM. For instance, enterprise risk management tends to focus on risk disclosure in financial reporting of a company so as to comply with such regulations as the Sarbanes-Oxley Act in the United States or the KonTraG requirements in Germany. Likewise, supply chain management is primarily concerned with ways to improve the operational performance of a supply chain under “normal” circumstances. Crisis management focuses on the survival of the organization or even that of society and may seem distant from worries stemming from a delayed delivery of parts at a plant.

Still, we can draw on these domains, especially on enterprise risk management, to motivate viewing SCRM. Therefore, we

can view SCRM as comprising the four steps in enterprise risk management: (i) identifying risks, (ii) assessing these risks, (iii) mitigating these risks, and (iv) responding to incidents through communication and other means. The first three entail activities that take place *before* the occurrence of an incident that generates negative (and mostly) unanticipated consequences while the last one applies to actions taken *during* and *after* the occurrence of an incident—the focus here is on time, which is also called *time-based risk management* [6].

RISK IDENTIFICATION

Basic methods for identifying risks include identifying critical uncertainties in scenario planning [7], and charting different types of risk using taxonomy-based risk identification [8]. Scholars and practitioners have proposed different taxonomies. For instance, Juttner *et al.* [9] provide a classification of risks based on sources such as environmental risk sources, network risk sources, and organizational risk sources. Chopra and Sodhi [10] talk of delays and disruptions, while [11] discuss supply risk, process risks, and demand risks. In identifying risks, the terms *uncertainty* and *risk* have been used interchangeably in the supply chain literature although the economics literature has attempted to narrow risk to only those situations where the uncertain outcomes have a known probability distribution.

Broadly speaking, we can consider three types of risk motivated by supply chain management:

- *Supply Risks.* Supply risks relate to incidents that occur at different tiers of suppliers and threaten supplies to a company. These incidents include supplier defaults or other unexpected changes in supply cost, delivery, quality, or reliability. As many manufacturers reduced the number of direct suppliers so as to foster better supplier relationships, this form of supply risks is growing in importance [12]. Risks

related to outsourcing fall in this category.

- *Process Risks.* Despite significant efforts in implementing Total Quality Management (TQM), Lean Manufacturing, and Six Sigma, many companies are still facing risks from products produced as a result of faulty design or manufacturing. For example, because of unsafe toys with magnets designed internally by Mattel, Mattel recalled over 17 million toys in 2007. There are internal and external process risks in the manufacturing and service industries as well. For example, in 2004, IBM announced that yield problems at its plant in East Fishkill, New York contributed to the \$150 million first-quarter loss by its microelectronics division [13]. The lower-than-expected yields reduced the plant's effective capacity and limited IBM's ability to meet customer demand.
- *Demand Risks.* The uncertain nature of product demand is one of the supply chain risks that all companies need to face with uncertainty surrounding both volume and mix. For example, Hewlett-Packard (HP) developed multiple versions for each model of their DeskJet printers to serve different geographical regions, that is, Asia-Pacific, Europe, and the Americas. Owing to uncertain demand in each region, HP faced the problem of simultaneously overstocking some models of printers in one region while understocking the same in other regions.

There are also risks to the enterprise itself and hence to the entire supply chain. These include:

- *Intellectual property (IP) Risks.* While outsourcing or off-shoring can result in lower manufacturing costs, it makes it difficult to protect IP. For example, even though the reform of the IP protection law has made some good progress after China's WTO entry in 2001, some incidents can still occur there.

- *Behavioral Risks.* As the number of partners increases in a global supply chain, the level of visibility and control can be reduced significantly. The low visibility level and the low control level reduce the "confidence" of each supply chain partner regarding the following information: the replenishment lead time/order status quoted by upstream partners, and demand forecasts provided by downstream partners. Christopher and Lee [14] argue that a low confidence level can cause the entire supply chain to enter a "risk spiral" so that each supply chain partner either "inflates" their order or "disguises" their on-hand inventory because of the lack of confidence in the replenishment lead time and demand forecasts. The confidence level deteriorates further as every partner starts gaming the system, and hence, the "risk spiral" continues.
- *Political/Social Risks.* A global supply chain is subjected to social/political risks when multiple countries are involved. For example, Airbus, a four-nation consortium, incurred an opportunity loss of €4.8 billion due to a two-year delay in launching the Superjumbo A380. In addition to technical problems associated with the wiring system, political issues across the four countries with manufacturing plants to make these planes may be a reason. Airbus' parent, EADS has struggled to develop a restructuring plan to replace political bargaining with industrial logic [15].

RISK ASSESSMENT

For each identified risk, assessment involves estimating the likelihood of an incident occurring and estimating the consequences in case such an incident does occur. The consequences may include short-term and long-term financial losses, loss of human lives, and loss in brand equity. Zsidisin *et al.* [16] obtain common themes on supply risk assessment techniques and examine how different purchasing companies reduce

their supply risk. Where consequences are probability distributions, they may be “measured” as value-at-risk or VaR-based or related *risk metrics* [17].

A company may conduct *risk mapping* exercises to assess the likelihood of each risk that could occur and the associated consequences in case of an occurrence. Doing so helps focus on actions to reduce the overall impact that is possible from risks with low likelihood of an incident occurring but with potentially large consequences or from risks that have small impact per occurrence but can occur quite frequently.

The likelihood of a risk refers to the “occurrence frequency” of different uncertain events and based on this frequency—whether observed or subjective—the risk may be considered “normal” or “abnormal.” Frequent fluctuations of process yields, material costs, currency exchange rates, and product demands can be viewed as normal risks, while rare events such as natural/man-made disasters, contaminated products, and system failures are abnormal. Some researchers argue that managing normal risks is part of supply chain management while managing abnormal risks is the part of SCRM; however, the issue is really about who manages these risks in the organization rather than whether SCRM has normal risks in its scope or not.

We need to draw a distinction between the occurrence of uncertain events and the consequences: the uncertain event is the risk, while the eventual consequences (and to whom) depend on the actions taken by different parties after the event occurs. For example, from the viewpoint of many manufacturers, the terrorist attack on September 11, 2001 was an uncertain event, while the subsequent supply delays were a disruption directly caused by the suspension of air transportation.

Consequences may be local or global depending on whether the impact of the risk incident is limited to a particular location within the supply chain or whether the entire supply chain is affected. *Global consequences* impact the entire supply chain. For example, due to a recent surge in gasoline prices, Ford’s supply chain for SUV manufacturing

came to a halt in 2008 as consumers switched to compact cars; and Mattel suspended its production of toys after recalling over 20 million in 2007. *Local consequences* impact a particular market or a particular site. For example, Ford closed five plants in the United States for several days after all air traffic was suspended after September 11, 2001.

However, where an incident occurs may or may not be connected to where the consequences are felt. Thus, assessing supply chain risk entails understanding where a risk incident can occur (or is currently occurring) as distinct from where the consequences might be. We can categorize supply chain risks as being related to incidents that occur globally, spanning the entire supply chain, or those that occur locally at a particular supply chain entity only.

Global risks refer to uncertain events arising from the global environment within which the supply chain operates. These uncertain events include social or political instabilities, credit crunch crises, and commodity price increases.

Local risks refer to uncertain events that occur at one or more supply chain entities. Examples include natural disasters, labor union strikes, bankruptcies, contaminated production processes, or violation of IP rights at one or more supply chain entities. Local risks also include damaging behavior of certain supply chain partners.

We can thus assess supply chain risks by the “point of occurrence” and by consequence (Table 1) to set the stage for developing mitigation strategies. Identifying the point of occurrence of each type of risk and its consequences would create shared awareness of different types of risks and their potential impact on different supply chain parties. This process can enable supply chain partners to better define their roles and responsibilities and to generate support for collaborative efforts for mitigating risks for all parties. For example, after suffering from a “global” \$2 billion loss in sales that was caused by a “local” fire that disrupted the production of microchips at a supplier’s plant, Ericsson worked with its supply chain partners

Table 1. Supply Chain Risks and their Consequences

Supply chain risk (point of possible occurrence)	Consequences (Region of possible eventual impact)	
	<i>Local</i> (impact to a particular supply chain entity or market)	<i>Global</i> (impact to the entire supply chain or multiple markets)
<i>Local</i> (originating at a supply chain entity or in a particular market)	Local risks stemming from supply, demand, or failure match supply and demand	Risks originating in a plant or region whose consequences eventually impact the entire supply chain or markets
<i>Global</i> (originating supply chain-wide or globally)	Global risks that affect a particular supply chain entity	Global risks that can affect the entire supply chain as a whole

to develop proactive plans should an incident occur [18].

Clearly, the affected supply chain entity is responsible for reducing the local risk when the consequences are also local. Arguably, some of this is very much part of supply chain management, for example, maintaining inventory to reduce the impact of unexpected delays in supplies.

At the other extreme, for global risks that have global consequences, collaborative efforts from the entire supply chain would be needed at the corporate levels of companies. For example, to improve supply chain security, the Customs-Trade Partnership against Terrorism (C-TPAT) certification program was established in 2002 by the US Customs. This certification program requires all supply chain partners to comply with a standard operating procedure to ensure security. To entice companies to comply with the best security practices, C-TPAT certified companies are allowed to clear customs faster with less inspection [19].

To handle local risk with global consequences—some would argue this is really what SCRM should focus on—we need to understand both the role of the point of occurrence and the affected parties. For example, after recognizing the global consequence (Mattel's worldwide recall of lead-tainted toys) associated with a local risk (supplier was using lead-tainted paints to produce toys), Mattel announced a safety check system to prevent the manufacture of toys with noncompliant levels of lead in paint [20].

Finally, situations entailing global risk with local consequences are usually the result of corporate policy applied inappropriately or incorrectly to a local situation. For instance, there may be local negative consequences of a corporate procurement policy. This can be mitigated only by communication between the local facilities with the corporate entity.

RISK MITIGATION

Once a particular supply chain risk is identified, assessed, and selected for mitigation, it can be mitigated by decreasing its likelihood, reducing its potential consequences or both. Paulsson and Nilsson [21] present 23 methods for mitigating risks; however, there are three basic approaches: *accept*, *avoid*, and *reduce*. *Accepting* the risk does not require doing anything other than the company bearing the entire consequence in case there is a risk incident or transferring part of the consequences to its insurance company or to its supply chain partner. Transferring risk through insurance or through financial instruments like swaps does not actually reduce the likelihood of the risk. Even if it reduces the impact of the risk to a certain extent, it may result in moral hazard whereby the company can become more risk-prone knowing that it can transfer some or all of the financial consequences. Likewise, liability insurance may offer financial compensations to customers who suffer from using unsafe products, but it does not reduce the damage to the reputation of the company

nor the suffering of the people who used these products.

Avoiding risks entails efforts to reduce the likelihood of the occurrence of certain undesirable incidents. Such efforts include the development of fail-safe systems, that is, systems that cannot fail or that can trigger corrective actions to prevent failures, and the application of quality-based principles to ensure there is no failure in the detection of a risk incident with highly negative consequences. Such approaches can be quite useful in security-related risks where preventing an incident from ever happening is the best approach. Lee and Wolfe [22] illustrate how certain technologies, say, biometric systems for positive identification of personnel and smart container systems for monitoring the internal temperature and pressure of each container, can be used to prevent containers being tampered with throughout the shipping process. The US Homeland Security has developed the Container Security Initiative (CSI) that requires all containers to be pre-screened at the port of departure before they arrive at US ports to reduce the likelihood of seaport terrorist attacks in the United States.

Reducing risk impact entails efforts to reduce the impact of risk incidents in case such incidents do occur. Broad approaches in this category are (i) alignment of supply chain partners' incentives, (ii) the use of "buffers" or redundancies, and (iii) flexibility.

1. *Alignment.* Besides long-term partnerships, there are other mechanisms to coordinate the interests of the different supply chain partners. Tang [23] reviewed different types of supply contracts to coordinate a supply chain so that all parties will act in the interest of the entire supply chain when dealing with demand risks including *wholesale price contracts*, *buyback contracts*, and *revenue sharing contracts*. A two-part tariff, that is, a fixed cost and a per unit wholesale price, can be used to entice the downstream partner to order according to the optimal order quantity (as per the newsvendor solution) for the entire supply chain. A buy-back contract is a returns policy under

which the manufacturer is required to buy back the retailer's excess inventory at a reduced price. Under certain conditions, doing so can achieve supply chain coordination. Under a revenue sharing contract, the retailer shares the revenue with the manufacturer and obtains a reduction in the wholesale price in return, achieving supply chain coordination. Narayanan and Raman [24] have studied *risk sharing* and revenue sharing to align incentives across supply chain partners.

2. *Buffers for Redundancy.* Firms can always build in some redundancies throughout the supply chain so as to reduce the cost implications of certain undesirable events associated with supply, process, and demand risks; Chopra and Sodhi [10] refer to these as *buffers* against supply chain risk. For example, extra inventory, extra back-up production capacity, and extra back-up suppliers are buffers against delays and disruption in the supply chain. However, Sheffi [4] argues that redundancies can be expensive when used against (rare) unanticipated events. Also, redundancies disguise inefficiencies in the supply chain, inhibiting the achievement of a lean supply chain. Flexibility overcomes these disadvantages: it can reduce the impact of the occurrence of certain unanticipated events and it can also be put to use with planned changes, for instance, to produce a greater variety of products.
3. *Flexibility.* Tang and Tomlin [11] analyze five different types of flexibility strategies corresponding to *multiple suppliers*, *flexible supply contracts*, *flexible manufacturing process*, *postponement*, and *responsive pricing*. The ability to shift order quantities across suppliers can be a powerful mechanism for the manufacturer to hedge against supply risks. Under flexible supply contracts, the manufacturer is allowed to adjust the order quantity within a prespecified range, say, a few percent of the order quantity. This helps to

mitigate the impact associated with demand risks [25]. The manufacturing process is flexible if different types of products can be manufactured in the same plant, enabling the manufacturer to reduce supply, process, or demand risks [26]. Postponement calls for delayed product differentiation by producing a generic product initially and then customizing it for different markets and customers later, thus allowing a company to respond to demand changes across multiple markets quickly [27]. Responsive pricing is an effective tool to mitigate supply or demand risks by manipulating demand when the supply is inflexible. For example, as the supply of certain components from Taiwan was affected by an earthquake, Dell's response was to lower the price of certain products so as to entice their on-line customers to "shift" their demands to other Dell computers that utilized components from other countries.

Similar to the above, Lee [28] recommends the principles of alignment, agility, and adaptability and provides industry examples for reducing the impact associated with different types of supply chain risks. Alignment calls for aligned interests among supply chain partners so as to facilitate close communication, cooperation, and collaboration; agility entails flexibility and responsiveness; and adaptability requires close monitoring of the environment so that one can deploy a recovery plan in a timely manner.

RISK RESPONSE AND COMMUNICATION

When a company cannot prevent a risk incident from occurring or mitigate it as described in the previous section, it has to figure out ways to respond quickly so as to contain the damage. The focus is really on time and therefore Sodhi and Tang [6] called it *time-based risk management*. Communication is a part of *crisis management* that is also needed as a response to untoward incidents.

Once the supply chain partners have identified and assessed certain types of supply chain risks and developed certain risk mitigation plans based on either the alignment or the agility strategies, it remains to develop adaptive capabilities for quick response to risk incidents, that is, reduce response lead time comprising of (i) the time to *detect* a risk incident, (ii) the time to *design* one or more solutions as well as *selecting* one solution in response to the incident, and (iii) the time to *deploy* the solution. Companies can reduce the impact of supply chain risk incidents by shortening these three elements of time and hence the response time. After deployment, the time it takes to restore the supply chain operations is the recovery time.

If the response lead time is short, the eventual impact will be lower as well—if not, the company is better off pursuing risk avoidance as discussed earlier. To understand the impact of response lead time on total impact of a risk incident, consider the case of Med fly (Mediterranean fruit fly) eradication in California in 1980. Instead of calling for an effective but costly aerial spray of Malathion in a small area of 30 square miles, Governor Brown chose to go with releasing sterile male flies and traps and within 12 months, the area of infestation expanded more than 20-fold. Local consequences became global over this time and Japan and other countries imposed import restrictions. Eventually, aerial spray had to be done over an area of 1500 square miles starting on July 14, 1981. This delayed action came at a significant cost: an expenditure of over \$100 million and Governor Brown's political career [29].

Detection time can be reduced by developing mechanisms to discover a risk incident quickly when it occurs or even to predict it before it occurs. Companies must also identify ways to share the information with their supply chain partners. Monitoring and advance warning systems can enable firms to reduce the detection lead time. For instance, many firms have various IT systems for monitoring the material flows (delivery and sales) and information flows (demand forecasts, production schedule, inventory level, quality) along the supply chain on a regular basis. For example, Nike has a "virtual

radar screen” for monitoring its supply chain operations [30]; Nokia monitors the delivery schedule of suppliers; and Seven-Eleven Japan monitors the production/delivery schedule from their vendors (suppliers) as well as the point of sales data from different stores throughout the day [31]. Such monitoring systems typically use various types of control charts (see **Supply Chain Coordination**) to monitor the operations and to issue an alert should any anomaly occur. Advance warning systems, by contrast, are intended to detect an undesirable event before it actually occurs, thus reducing detection time to a minimum. For example, smart alert systems enable GE to conduct remote-sensing and diagnostics, so that it can deploy engineers to service turbines before catastrophic failures occur. *Design time* would be reduced if the company and its partners can develop contingent recovery plans for different types of disruptions *in advance*. Many organizations seek to do so through *business continuity* efforts. For example, Li and Fung has a variety of different contingent supply plans to enable them to switch from one supplier in one country to another supplier in a different country [32]. Seven-Eleven Japan has developed contingency delivery plans to enable them to switch from one transportation mode (trucks) to another (motorcycles), depending on traffic conditions [31]. A company may find it useful to use decision analysis tools such as decision-tree analysis (see **Describing Decision Problems by Decision Trees**) to refine and select a recovery plan. Sarin [33] presents a decision-analysis framework for evaluating different earthquake-safety plans.

Deployment lead time can be shortened by improving communication and coordination among supply chain entities within the company or within supply chain partner firms, once a recovery plan has been selected. Van Wassenhove [34] suggested three forms of coordination: (i) by command (central coordination), (ii) by consensus (information sharing), and (iii) by default (routine communication). In the event of a major disruption, coordination by command seems to be more appropriate during the design phase of a

recovery plan and its deployment. Once the recovery plan is deployed, coordination by consensus would seem appropriate especially when each party has a clear role and responsibility established in advance.

Communication, an important aspect of *crisis management*, is also a time-based response approach motivated by public perception. Companies need to develop crisis management plans to manage the public perception so as to reduce the long-term impact on the company’s reputation. Immediately after a risk incident, the involved supply chain parties need to develop plans to restore customer relationship and communicate their approach for rebuilding public confidence. For example, immediately after Mattel issued multiple recalls in 2007, it created simple ways for consumers to return the recalled products using downloadable forms, and free shipping mailers. To restore trust, the company acknowledged problems to stakeholders and apologized to customers publicly in the press. In exchange for returned toys, it offered coupons for customers to buy other Mattel products of the same value. It also announced the three-stage safety check system in September 2007: test every batch of inputs (paint); test every batch of production; and test samples of finished goods from each production run. In addition to external communication, internal communication is important to facilitate organizational learning to improve better manage supply chain risk in the future [35]. At Mattel, a new position was created reporting directly to the Mattel’s CEO to ensure that potential recall risks were quickly addressed, and, if necessary, elevated within the company for improved internal communication [36].

CONCLUSION

We have described SCRM referring to its roots in the disciplines of supply chain management, enterprise risk management, and crisis management. We chose the enterprise risk management approach of viewing risk management in terms of (i) risk identification, (ii) risk assessment, (iii) risk mitigation, and (iv) risk response and communication.

While rooted in these disciplines, SCRM is not fully a part of any of these. SCRM is not a part of supply chain management although good supply chain management needs to include principles of SCRM. Likewise, SCRM is not a part of enterprise risk management, although to do effective enterprise risk management, a company needs to excel in SCRM. Finally, SCRM is not part of crisis management and business continuity, although if the supply chain experiences a crisis, then any solution to crisis management has to incorporate SCRM as we saw from Mattel's example. Thus, SCRM is important not just to supply chain professionals but to others as well. Given its growing importance to companies, it is likely it will be viewed as a discipline in its own right.

Acknowledgment

The authors are grateful to the participants of the Supply Chain Thought Leaders (SCTL) meeting in Madrid, 2008 and the International Supply Chain Risk Management (ISCRIM) workshop in Trondheim, 2008 for sharing their views on this topic.

REFERENCES

1. Craighead C, Blackhurst J, Rungtusanatham MJ, *et al.* The severity of supply chain disruptions: design characteristics and mitigation capabilities. *Decis Sci* 2007;38(1):131–156.
2. Hendricks KB, Singhal VR. An empirical analysis of the effect of supply chain disruptions on the long-run stock price performance and equity risk of the firm. *Prod Oper Manag* 2005;14(1):35–52.
3. McKinsey. Understanding supply chain risk: A McKinsey Global Survey. USA: The McKinsey Quarterly; 2006.
4. Sheffi Y. The resilient enterprise. Boston (MA): MIT Press; 2005.
5. Zsidisin G, Ritchie B. Supply chain risk: a handbook of assessment, management & performance. New York: Springer; 2008.
6. Sodhi M, Tang CS. Time-based disruption management: a framework for managing supply chain disruption. In: Blackhurst J, Wu T, editors. Managing supply chain risk and vulnerability: tools and methods for supply chain decision makers. New York: Springer; 2009.
7. Garvin DA, and Levesque LC. Meeting the Challenge of Corporate Entrepreneurship. *Harvard Business Review* 2006;84(10).
8. Carr MJ, Konda SL, Monarch I, *et al.* Taxonomy based risk identification. Research Report No. SEI-93-TR-6. USA: Carnegie Mellon University; 1993.
9. Juttner U, Peck H, Christopher M. Supply chain risk management: outlining an agenda for future research. *Int J Log* 2003;6(4): 197–210.
10. Chopra S, Sodhi M. Managing risk to avoid supply chain breakdown. *Sloan Manag Rev* 2004;46(1):53–61.
11. Tang CS, Tomlin B. The power of flexibility for mitigating supply chain risks. *Int J Prod Econ* 2008;116:12–27.
12. Tang CS. Supplier relationship map. *Int J Log Res Appl* 1999;2(1):39–56.
13. Krazit T. Trouble in East Fishkill? IBM chip group struggles. 2004. InfoWorld.com, April 21.
14. Christopher M, Lee H. Mitigating supply chain risk through improved confidence. Working paper. UK: Cranfield School of Management; 2004.
15. Gumbel P. Trying to Untangle Wires, *Time Europe Magazine*. 2006. Available at <http://www.time.com/time/europe/magazine/printout/0,13155,1543879,00.html>
16. Zsidisin G, Ellram L, Carter J, *et al.* An analysis of supply risk assessment techniques. *Int J Phys Distrib Log Manag* 2004;34(5):397–413.
17. Sodhi M. Tactical planning under demand risk for a global electronics company. *Prod Oper Manag* 2005;14(1):69–79.
18. Norrman A, Jansson U. Ericsson's proactive supply chain risk management approach after a serious sub-supplier accident. *Int J Phys Distrib Log Manag* 2004;34:434–456.
19. Tang CS. Robust strategies for mitigating supply chain disruptions. *Int J Log Res Appl* 2006;9(3):33–45.
20. Pyke D, and Tang CS. How to Mitigate Product Safety Risks Proactively? Process, Challenges and Opportunities, working paper, UCLA Anderson School, 2008.
21. Paulsson U, Nilsson CH. Potential risk handling alternatives for supply chain disruptions in liquid food production: the case of V&S Vin & Sprit AB, the Sundsvall site. Research Report no. 1015. Sweden: Lund University; 2008.

22. Lee H, Wolfe M. Supply chain security without tears. *Supply Chain Manag Rev* 2003;7(1):12–20.
23. Tang CS. Perspectives in supply chain risk management. *Int J Prod Econ* 2006;103(4): 451–488.
24. Narayanan VG, Raman A. Aligning incentives in supply chains. *Harv Bus Rev* 2004;6:94–103.
25. Tsay AA, Lovejoy WS. Quantity flexible contracts and supply chain performance. *Manuf Serv Oper Manag* 1999;1, 2:89–111.
26. Jordan W, and Graves SC. Principles on the Benefits of Manufacturing Process Flexibility *Management Science* 1995;41(4):577–594.
27. Lee HL, Tang CS. Modeling the costs and benefits of delayed product differentiation. *Manag Sci* 1997;43:40–53.
28. Lee H. The triple-a supply chain. *Harv Bus Rev* 2004;10:102–112.
29. Denardo EV. *The science of decision making: a problem-based approach using excel*. New York: Wiley Publisher; 2002.
30. Hartwigsen K. *Integrated crisis management: building a framework for unified response*. USA: Nike Internal Presentation; 2006.
31. Lee H, Whang J. Seven Eleven Japan. Stanford Business School case # GS-18. 2006.
32. St.George A. Li and Fung: Beyond ‘Filling in the Mosaic’. Boston (MA): Harvard Business School Case 9-398–092; 1998.
33. Sarin R. A social decision analysis of the earthquake safety problem. In: Kleindorfer PR, Sertel MR, editors. *Mitigation and financing seismic risks*. The Netherlands: Kluwer Publishers; 2001.
34. Van Wassenhove L. Humanitarian aid logistics: supply chain management in high gear. *J Oper Res Soc* 2006;57:475–489.
35. Zsidisin G, Melnyk S, Ragatz G. An institutional theory perspective of business continuity planning for purchasing and supply management. *Int J Prod Res* 2005;43(16): 3401–3420.
36. Pyke D, Tang CS. How to mitigate product safety risks proactively? Process, challenges and opportunities. Working paper. USA: UCLA Anderson School; 2008.

SUPPLY CHAIN SCHEDULING: ORIGINS AND APPLICATION TO SEQUENCING, BATCHING AND LOT SIZING

NICHOLAS G. HALL

Department of Management Sciences,
Fisher College of Business, The Ohio
State University, Columbus, Ohio

This article discusses supply chain coordination at the detailed, operational level. Work in this area is motivated both by the practical importance of the problems studied and the wide variety of interesting research questions. Thomas and Griffin [1] provide a detailed review of the literature of supply chain management. They indicate that over 11% of the US Gross National Product comes from nonmilitary logistics expenditures. In fact, for many products, logistics expenditures account for more than 30% of the cost of goods sold. The authors identify a need for research that addresses supply chain issues at an operational rather than a strategic level, and that uses deterministic rather than stochastic methodology to analyze them. The recent development of the supply chain scheduling research area, which is the subject of this article, addresses both these requirements.

Further motivation for this article is provided by a survey of integrated production and distribution models by Sarmiento and Nagi [2]. As they point out, the recent trend toward lean manufacturing and reduced inventory levels [3,4] has created a closer interaction between the adjacent stages of a supply chain. This is because inventory can no longer be used as a buffer between poorly coordinated schedules. This, in turn, increases the need for coordination in scheduling and other related decisions between various decision makers with different costs, constraints, and objectives.

The need for coordinated operational decision making of the type discussed here arises in applications with short time

horizons, which reduces the usefulness of models for strategic decision making. Related applications include customized and assemble-to-order products, fashion items, and seasonal products. Relevant examples of manufactured goods include electronics products, apparel, toys, and perishable food products. Among the typical characteristics of these goods are that the products are time sensitive and the orders are distinct. For a more detailed discussion of these applications, see *Coordination of Production and Delivery in Supply Chain Scheduling* in this encyclopedia.

When a set of jobs needs to be manufactured, the cost and feasibility of scheduling them often depend on the *sequence* in which they are processed. Pinedo [5] provides many examples. In some manufacturing environments, a sequence of the jobs that is both feasible and cost minimizing can be found by an indexing rule based on their characteristics. In other environments, a more complex algorithm is required, but a cost-minimizing sequence can still be computed in a reasonable amount of time. However, in a third group of environments, the sequencing problem is formally intractable [6], and hence finding a cost-minimizing sequence is likely to be computationally possible only if the number of jobs is small. In this last case, scheduling decisions for practical size problems are often made heuristically, which usually results in solutions that do not minimize cost.

An example of mutually incompatible decision making between a supplier and a manufacturer occurs in sequencing the production of jobs. Suppose that a supplier uses a particular preferred sequence of jobs that reduces its sequence-dependent setup costs. Further, suppose that, in order to reduce the supplier's inventory holding cost, these jobs are delivered individually to a manufacturer in the production sequence preferred by the supplier. Problematically, the manufacturer may have completely different sequence-dependent setup costs

that suggest a very dissimilar processing sequence. The manufacturer then has various options, none of which is necessarily favorable. First, it can follow the sequence used by the supplier, which controls holding costs but incurs high setup costs. Second, it can store the jobs as needed until enough of them have arrived that it can use its preferred sequence, which controls setup costs but incurs high holding costs. Solutions that interpolate between these two possibilities in various ways are also available. The decisions of the supplier affect the costs of the manufacturer, but the converse is not true. Therefore, the manufacturer can provide incentives to the supplier to change its job delivery sequence to be more compatible with the preferred processing sequence of the manufacturer.

Several modern extensions of classical scheduling theory address issues of conflicting objectives or competing agents. First, multicriteria scheduling [7,8] discusses problems faced by a single decision maker who measures cost in multiple ways. Although this research does not model supply chains, the use of Pareto optimality suggests some approaches that can be useful in that context. Second, integrated scheduling [9,10] considers operational issues across multiple stages of the supply chain within a single organization. Although there are no competing agents, conflicts naturally arise between decisions at different stages. For

example, the optimal production sequence may need to be modified in order to control transportation costs. Third, sequencing games [11–13] consider multiple agents within the same stage of production. If the agents can cooperate, the overall cost of the system is reduced. The focus of this research is on the existence of core solutions and related concepts within cooperative and bargaining games [14], rather than optimization of the overall system cost. Finally, supply chain scheduling [15,16] considers problems with multiple agents across multiple supply chain stages. The supply chain scheduling literature evaluates the benefit of coordination for the overall supply chain, and makes prescriptive statements about how to achieve coordination. Table 1 describes the main features of these four extensions, thereby illustrating the origins of supply chain scheduling and its relationship with other modern extensions of classical scheduling theory.

In multistage supply chains, *batching* involves processing and/or transporting several jobs contiguously. One motivation for batching is to reduce setup times or costs that occur when different types of items are produced consecutively. Examples arise in the chemical, food processing, and fertilizer industries. A second example arises in the manufacturing of engineering components, where the principles of group technology [17] can be used to organize the components into

Table 1. Modern Extensions of Classical Scheduling Theory

Topic	Multicriteria Scheduling	Integrated Scheduling	Sequencing Games	Supply Chain Scheduling
Number of agents	One	One	Many	Many
Number of stages	One or two	Two	One	Two or more
Optimization of system cost	Yes	Yes	No	Yes
Evaluation of benefit of coordination	No	Yes	Yes	Yes
Agents have distinct jobs	No	No	Yes	Yes
Prescriptive mechanisms	Pareto, optimization	Optimization, heuristics	Game theory, allocations	Incentives, game theory
Early reference	Nelson <i>et al.</i> [7]	Maggu <i>et al.</i> [9]	Curiel <i>et al.</i> [11]	Hall and Potts [14]
Recent reference	Hoogeveen [8]	Lee and Chen [10]	Agnetis <i>et al.</i> [12]	Hall and Liu [15]

families with enough similarity that setups are not needed between the processing of their components. A second motivation for batching arises from transportation cost savings. When a supplier ships to a manufacturer, a transportation cost is incurred. Batching of more jobs reduces the number of shipments and thus the total fixed cost of transportation. A third motivation occurs where several jobs are naturally processed in a batch, for example as with burn-in operations in semiconductor manufacturing [18].

Chen [19] provides a comprehensive review of batching within the integrated scheduling literature. Lee and Chen [10] study the integration of batching issues in transportation, including those related to time and capacity, with scheduling decisions. Two types of transportation are considered: the transfer of semifinished jobs in batches from one processing facility to another using automated transporters, and the delivery of finished goods in batches to warehouses or customers using trucks. Both transportation capacity and transportation times are considered. The focus of the work is on the solvability of a variety of integrated scheduling problems, including the design of efficient algorithms and proofs of intractability. In addition, the work describes a heuristic for one of the intractable problems studied, and analyzes its worst case performance. This work is motivational for the topic of supply chain scheduling, but considers neither multiple decision makers with conflicting objectives and distinct job sets nor the design of incentives that result in effective coordination of scheduling and other related decisions.

An issue that is closely related to batching is *lot sizing*, which is the decision on how and when to split a production lot of identical items into sublots. In a multistage supply chain, the use of sublots permits the overlapping of different operations on the same production lot, in order to reduce throughput time. Thus, the use of sublots by a supplier effectively reduces the arrival times of jobs at the manufacturer, which makes the scheduling problem of the manufacturer less constrained. However,

except under vendor-managed inventory [20] (see *Vendor-Managed Inventory*), this also means that the manufacturer is responsible for the inventory holding cost of the subplot at an earlier time. The resulting trade-offs depend on the relative costs of production setups, holding inventory and delivery. Depending on these trade-offs, the manufacturer may wish to provide incentives to the supplier to change its lot-sizing policy.

We discuss some examples of lot-sizing models in multistage supply chains, with the objective of minimizing the total of ordering and holding costs. Monahan [21] shows how a supplier can use a price discount to encourage a manufacturer to order in larger lot sizes, which benefits the supplier. This model is generalized by Lee and Rosenblatt [22], who introduce the practical realities of a limit on the discount that can be offered and different lot sizes for the supplier and the manufacturer. They describe an algorithm for the supplier's joint price discount and ordering problem. Sensitivity analysis results show that the amount of price discount which the supplier should offer depends on the costs of both parties. These works are motivational for coordination issues that arise in supply chain scheduling, but fail to consider timing issues and costs at the individual job level.

As with sequencing decisions, batching and lot-sizing decisions are difficult to make in some planning environments. An additional complication is that effective decision making often requires that sequencing, batching, and lot-sizing decisions be made jointly rather than separately. Within supply chains, still another complication is that the self interested sequencing, batching and lot-sizing decisions of a supplier and a manufacturer may not be mutually compatible. In such cases, the overall performance of the supply chain is adversely affected, sometimes severely, by poorly coordinated decision making.

Potts and Van Wassenhove [23] and Potts and Kovalyov [24] provide detailed reviews of the literature of scheduling in combination with batching and lot sizing. We discuss some

early work on coordination issues in production and distribution systems with multiple stages. Ow *et al.* [25] describe a multiagent scheduling system. Lee *et al.* [26] consider a scheduling problem that arises in a two-stage assembly system consisting of two suppliers and one manufacturer. They discuss the solvability of this problem, and describe both exact and approximate solution procedures. Potts *et al.* [27] consider a more general problem with several suppliers and one manufacturer. They describe approximate solution methods, and study their worst case performance. The focus of all these papers is on optimization and approximation results. Indeed, none of these papers provides a detailed analysis of cooperation between decision makers with different and possibly conflicting objectives. The supply chain scheduling papers, discussed in the next section, do achieve this.

SUPPLY CHAIN SCHEDULING LITERATURE

We first introduce a prototypical real-world supply chain scheduling problem. Boros *et al.* [28] consider a supply chain consisting of a foreign shipping company and a domestic port. The main decision to be made is the lot size with which the shipping company removes the empty containers from the domestic port, so that the joint profit of the two agents is maximized. The two agents have conflicting interests regarding this lot size. The shipping company prefers larger lot sizes and longer cycle times, which require fewer trips by its shipping vessels. However, the port prefers smaller lot sizes and shorter cycle times, which reduce the accumulation of the empty containers and their storage cost. The paper proposes an iterative heuristic to search for the lot size that maximizes the overall system profit. At each iteration, a candidate lot size is evaluated using a deterministic vehicle scheduling problem and a stochastic container storage capacity optimization problem. On the basis of the results of a computational study that includes 300 randomly generated problem instances, the heuristic delivers solutions that are usually within 1% of the optimal profit across

a wide range of parameter settings. However, the optimal lot size for the system is, in general, not optimal for either the shipping company or the port. As a result, each party requires an incentive to adopt a lot size that is different from its optimal choice. More generally, this outcome is common in supply chain scheduling problems. In most situations, cooperation by the parties generates substantial savings for the overall system. In many cases, it is possible for them to agree a distribution of these savings that ensures cooperation.

The earliest supply chain scheduling paper, by Hall and Potts [15], considers scheduling and batching problems with three different objectives in a supply chain where a supplier delivers to several manufacturers, who in turn deliver to customers. The supplier acts to minimize the total of its sequencing and batching cost. The supplier's sequencing and batching decisions define the arrival times of the jobs at the manufacturer. Then the manufacturer, given these arrival times, acts to minimize the total of its sequencing and batching cost. The decisions of both players are made using optimal dynamic programming algorithms. The optimal joint sequencing and batching decisions of the supplier and manufacturer, acting together to minimize the total supply chain cost, are also found similarly. Example instances show that the solution which results from the supplier and manufacturer acting independently can be considerably more costly than the solution of the combined problem. Further instances show that an optimal combined solution may not be a Nash equilibrium. This is because the supplier benefits from unilaterally changing its solution, even though the supply chain as a whole incurs larger cost. However, some simple mechanisms exist that can be used to obtain the supplier's cooperation in choosing the overall optimal solution. These mechanisms include the manufacturer offering an incentive for the supplier to deliver all the jobs together, and offering to share holding costs for completed jobs at the supplier.

Dawande *et al.* [29] study a real-world supply chain where a manufacturer makes

products that are shipped to customers by a distributor. The manufacturer chooses a schedule that minimizes unproductive time. This schedule uses long production runs in order to reduce the number of production setups. However, the distributor optimizes either customer service relative to due dates or inventory holding cost, and consequently prefers a different schedule to the manufacturer. If each party follows its own preferred schedule, then the overall supply chain is not well coordinated and its performance, as measured by total system cost, is typically poor. The results in the paper include measurement of each player's conflict, which is the relative increase in cost that results from using the other player's optimal schedule, relative to its own optimal schedule. This conflict is often significant, which motivates the study of coordination. How this conflict is resolved depends on the balance of power within the supply chain. Models are analyzed for situations where the manufacturer has dominant power, and for other situations where the distributor has dominant power. Finally, by solving the problem where both players cooperate fully, it is possible to evaluate the benefit of cooperation. Since achieving full cooperation typically requires the dominant player not to implement its optimal schedule, incentive mechanisms are developed to ensure its cooperation. Examples of such mechanisms include lump sum side payments, and contributions by the distributor toward the cost of additional production setups by the manufacturer.

Agnetis *et al.* [30] consider a two-stage supply chain consisting of a supplier and several manufacturers. A schedule is fully determined by a sequence. The supplier and each manufacturer have a different ideal sequence for processing the jobs, which is determined by its production changeover costs. An intermediate storage buffer of limited capacity is available to resequence the jobs after the supplier processes them, but before they reach the manufacturers. The paper considers the minimization of total changeover plus buffer storage cost. Efficient algorithms are described for the problems faced by the supplier and by each manufacturer individually. It is shown that

cooperation between the supplier and the manufacturers in sharing buffer cost can reduce the overall supply chain cost.

Chen and Hall [31] consider a two-stage assembly system where several suppliers provide parts to a manufacturer. A product cannot start processing at the manufacturer until all its parts have been supplied. The manufacturer performs nonbottleneck operations, for example, through outsourcing. The objectives considered are the minimization of the total completion time, and the optimization of a customer service measure which is either the total completion time or the maximum lateness of the latest job. The results begin with an analysis of conflict, that is, how far from optimal the best suppliers' schedule is when applied to the manufacturer's problem, and vice versa. For the total completion time problem, a lower bound on the manufacturer's conflict when using the suppliers' schedule is derived analytically and evaluated computationally at between 19% and 83%, depending on the data. Conversely, following the manufacturer's optimal schedule gives a conflict of between 3% and 37%, when used for the suppliers' problem. If the manufacturer uses maximum lateness as an objective, then following the suppliers' schedule results in very large conflict that increases with the number of suppliers. Conversely, using the manufacturer's optimal schedule results in a conflict that is between 1% and 49% in the suppliers' problem.

Motivated by these substantial levels of conflict, Chen and Hall [31] study four scenarios for bargaining power in the supply chain. First, if the suppliers dominate and the manufacturer negotiates with them by compensating the suppliers for using a different schedule, sufficient compensation can always be made to ensure that an optimal supply chain schedule is achieved. Second, if the manufacturer dominates and the suppliers negotiate, a similar result is achieved. Third, if the manufacturer dominates and does not accept negotiation from the suppliers, the manufacturer's solutions can be modeled as constraints within the suppliers' problem. Fourth, if the parties cooperate, sufficient conditions are described for compensating the

players to follow the system optimal schedule; whether those conditions are met depends on the problem data. Finally, the relative cost saving from cooperation between the suppliers and the manufacturer is evaluated. When all players use completion time as an objective, this ranges between 0% and 27%. When the manufacturer uses maximum lateness as an objective, the relative cost saving ranges between 1% and 83%. The possibility of sharing these substantial cost savings provides a strong incentive for the players to cooperate.

Manoj *et al.* [32] consider the benefit of coordinated decision making in a supply chain consisting of a manufacturer, a distributor, and several retailers. The distributor batches finished goods produced by the manufacturer, manages their inventory, and delivers them to the retailers. The costs of conflict are substantial. If a dominant manufacturer imposes its optimal schedule on the distributor, the result is typically an increase in the distributor's holding cost of between 52% and 75%. Consequently, the distributor may provide an incentive for the manufacturer to change its schedule. Conversely, if a dominant distributor imposes its optimal schedule on the manufacturer, the result is typically a several hundred percent increase in the manufacturer's production costs from changing production rates. In this case, the manufacturer may provide an incentive for the distributor to change its batching and delivery schedule. The benefit of cooperation from coordinated decision making by the manufacturer and distributor is also discussed. Where the manufacturer has dominant market power, cooperation results in a cost reduction for the distributor of between 28% and 37%. Where the distributor has dominant market power, the manufacturer's cost reduction is between 45% and 57%. These savings are typically sufficient to ensure the cooperation of both parties when incorporated in simple incentive mechanisms.

The following insights appear consistently within the supply chain scheduling literature.

1. There exist several reasons why different parties within a supply chain naturally make scheduling, lot sizing, and batching decisions that are mutually incompatible.
2. The cost of conflict imposed by a party with dominant power on other parties is usually considerable, and, in practice, quite likely to lead to the breakdown of the supply chain partnership if implemented.
3. The willingness of parties to cooperate with others depends not only on their relative market power but also on their relative positions upstream or downstream within the supply chain.
4. When the parties have approximately equal power in the supply chain, it is sometimes but not always the case that they can achieve and maintain coordination. Even for the same problem environment, whether this is possible may depend on problem data.
5. The saving in total supply chain cost that results from the coordination of sequencing, batching, and lot-sizing decisions is usually considerable.
6. The more costs are considered, the greater is the possibility for cooperation between the parties to reduce the total system cost.
7. Incentive mechanisms need not be complicated in order to ensure coordinated decision making; indeed, simple side payments are one of the most effective incentives in many supply chain scheduling problems.
8. An optimal cooperative solution is not always a Nash equilibrium, therefore verifiability may be needed to ensure stable decision making.
9. The greater the savings from cooperation between the parties, the easier it is to design incentives that will lead to a supply chain optimal solution.

SUMMARY

Research in supply chain scheduling analyzes the coordination of operational

decisions in supply chains. Conflicts over operational decisions such as sequencing, batching, and lot sizing arise naturally in many real-world supply chains, for example because of differences in relative costs between different supply chain members. Since the decisions being made are operational ones with short time horizons, more reliable data is available than in other supply chain planning environments. This permits more accurate estimation of the benefits of supply chain coordination, for example through optimization models, and therefore better decision making about when such coordination is worthwhile. Because the study of supply chain scheduling is relatively new, there are many opportunities for designing incentive schemes that encourage supply chain coordination in operational decision making. Because different decisions and costs are considered, successful incentive schemes may be different from those in traditional supply chain management. As discussed in *Capacity Allocation in Supply Chain Scheduling*, cooperative game theory models should yield valuable insights about the stability of cooperation between decision makers in various applications. The rapidly evolving literature of supply chain scheduling already includes a journal special issue devoted to this topic [33].

REFERENCES

1. Thomas DJ, Griffin PM. Coordinated supply chain management. *Eur J Oper Res* 1996;94:1–15.
2. Sarmiento AM, Nagi R. A review of integrated analysis of production-distribution systems. *IIE Trans* 1999;31:1061–1074.
3. Feld WM. *Lean manufacturing: tools, techniques and how to use them*. New York City: CRC Press; 2000. <http://www.crcpress.com>.
4. Rajagopalan S, Malhotra A. Have U.S. manufacturing inventories really decreased? An empirical study. *Manuf Serv Oper Manage* 2001;3:14–24.
5. Pinedo M. *Scheduling, theory, algorithms and systems*. 2nd ed. Englewood Cliffs (NJ): Prentice Hall; 2002.
6. Garey MR, Johnson DS. *Computers and intractability: a guide to the theory of NP-completeness*. San Francisco (CA): Freeman; 1979.
7. Nelson RT, Sarin RK, Daniels RL. Scheduling with multiple performance measures - the one-machine case. *Manage Sci* 1986;32:464–479.
8. Hoogeveen H. Multicriteria scheduling. *Eur J Oper Res* 2005;167:592–623.
9. Maggu PL, Das G, Kumar R. On equivalent - job for job - block in $2 \times n$ sequencing problem with transportation times. *J Oper Soc Jpn* 1981;24:136–146.
10. Lee C-Y, Chen Z-L. Machine scheduling with transportation considerations. *J Schedul* 2001;4:3–24.
11. Curiel I, Pederzoli G, Tijs S. Sequencing games. *Eur J Oper Res* 1989;40:344–351.
12. Curiel I, Hamers H, Klijn F. Sequencing games: a survey. Chapter 2, Chapters in game theory: in honor of Stef Tijs. Boston (MA): Kluwer Academic Publishers; 2002. pp. 27–50.
13. Agnetis A, de Pascale G, Pranzo M. Computing the Nash solution for scheduling bargaining problems. *Int J Oper Res* 2009;6:54–69.
14. Peleg B, Sudhölter P. *Introduction to the theory of cooperative games*. Boston (MA): Kluwer; 2003.
15. Hall NG, Potts CN. Supply chain scheduling: batching and delivery. *Oper Res* 2003;51:566–584.
16. Hall NG, Liu Z. Capacity allocation and scheduling in supply chains. *Operations Research* 2010. In press.
17. Ham I, Hitomi K, Yoshida T. *Group technology: applications to production management*. Boston (MA): Kluwer-Nijhoff; 1985.
18. Lee C-Y, Uzsoy R, Martin-Vega LA. Efficient algorithms for scheduling semiconductor burn-in operations. *Oper Res* 1992;40:764–775.
19. Chen Z-L. Integrated production and outbound distribution scheduling: review and extensions. *Oper Res* 2010;58:130–148.
20. Vendor Managed Inventory. Everything you need to know about vendor managed inventory. 2010. Available at <http://www.vendormanagedinventory.com>.
21. Monahan JP. A quantity discount pricing model to increase vendor profits. *Manage Sci* 1984;30:720–726.

22. Lee HL, Rosenblatt MJ. A generalized quantity discount pricing model to increase supplier's profit. *Manage Sci* 1986;32: 1177–1185.
23. Potts CN, Van Wassenhove LN. Integrating scheduling with batching and lot-sizing: a review of algorithms and complexity. *J Oper Res Soc* 1992;43:395–406.
24. Potts CN, Kovalyov MY. Scheduling with batching: a review. *Eur J Oper Res* 2002;120:228–249.
25. Ow PS, Smith SF, Howie R. A cooperative scheduling system. In: *Expert systems and intelligent manufacturing*. Amsterdam, Netherlands: Elsevier Science Publishing; 1988. pp. 43–56.
26. Lee C-Y, Cheng TCE, Lin BMT. Minimizing the makespan in the 3-machine assembly-type flowshop scheduling problem. *Manage Sci* 1993;39:616–625.
27. Potts CN, Sevast'janov SV, Strusevich VA, *et al.* The two-stage assembly scheduling problem: complexity and approximation. *Oper Res* 1995;43:346–355.
28. Boros E, Lei L, Zhao Y, *et al.* Scheduling vessels and container-yard operations with conflicting objectives. *Ann Oper Res* 2008;161:149–170.
29. Dawande M, Geismar HN, Hall NG, *et al.* Supply chain scheduling: distribution systems. *Prod Oper Manage* 2006;15:243–261.
30. Agnetis A, Hall NG, Pacciarelli D. Supply chain scheduling: sequence coordination. *Discrete Appl Math* 2006;154:2044–2063.
31. Chen Z-L, Hall NG. Supply chain scheduling: conflict and cooperation in assembly systems. *Oper Res* 2007;55:1072–1089.
32. Manoj UV, Gupta JND, Gupta SK, *et al.* Supply chain scheduling: just-in-time environment. *Ann Oper Res* 2008;161: 53–86.
33. Hall NG, Lei L, Pinedo M, editors. Volume 161, Supply chain coordination and scheduling. Norwell (MA): Springer; 2008.

SUPPORT VECTOR MACHINES FOR CLASSIFICATION

ROBIN C. GILBERT
THEODORE B. TRAFALIS
INDRA ADRIANTO
Laboratory of Optimization and
Intelligent Systems, School of
Industrial Engineering,
University of Oklahoma,
Norman, Oklahoma

INTRODUCTION

The ability of intelligent learning to apply what we have learned to new situations is called *generalization* [1–9]. Support vector machines (SVMs) are a family of learning algorithms first introduced by Vapnik [7], which are developed on this generalization concept. The SVM learning problem can be cast as a quadratic programming problem and other formulations result into a linear programming problem. In contrast to other learning networks, the resulting SVM optimization problem is convex with a global optimal solution. SVMs can discriminate between classes that are linearly *separable*, linearly inseparable, and nonlinearly separable. They work by maximizing the *margin* between two classes of training data that is, maximizing the distance between an affine hyperplane separating the two classes of data while avoiding overfitting the result by using a regularization process.

SVMs integrate the classification pattern learned from many training examples and determine a rule that maps inputs to their most probable classes. The learned rule has a good generalization behavior if it does as well as on new inputs as on the old ones. A classical result in learning theory shows that the function learned through the empirical risk minimization principle generalizes well if the *hypothesis* space from which they are chosen is *simple* [9]. The classical definition of a simple hypothesis space (a space with

low complexity) is technically involved. An example of a simple hypothesis space is the set of linear functions defined on the plane: the complexity of this set of functions or *VC dimension* is three since this is the greatest number of points in the plane belonging to two distinct classes that can be arranged so that suitable linear functions can separate them.

More recent work [10] shifts attention from the hypothesis space and concentrates on the concept of stability of learning algorithms. In simple words, a learning algorithm is stable if the removal of a training sample from a large set of samples results in a small change in the learned function. SVMs enjoy this stability property. Moreover, they are stable with respect to data perturbations. An advantage of SVM is the natural use of *kernel methods* that provide the ability to perform a mapping of the observed data into a high-dimensional (possibly infinite-dimensional) Hilbert space, thus, providing an avenue for exploring nonlinear binary classifiers [2,11].

SVMs have been applied to a wide range of problems such as computational biology [12–14], fluid mechanics [15–17], manufacturing engineering [18], meteorology [19–21], and even political science [22]. Recently, SVMs have been applied to short-term portfolio management [23], exchange rate prediction [24], and stock price prediction [25,26]. Other interesting applications are in the area of production [27], inventory transactions [28], bioinformatics and gene analysis [14,29], and web mining [30]. SVMs also have applications in manufacturing [31,32] and other domains, which are described in the special issue on Support Vector Machines and Applications of *Computational Management Science* [33].

The article is organized as follows. In the section titled “Binary Classification with Support Vector Machines,” the binary classification with SVMs is discussed. Multiclass classifications schemes are discussed in the section titled “Multiclass Classification Schemes.” The section titled “Machines

Based on Central Representations of the Version Space” provides improvements of SVMs and discusses computational issues, followed by the conclusion section.

BINARY CLASSIFICATION WITH SUPPORT VECTOR MACHINES

Consider a classification problem for which one has to find a *decision rule* that will evaluate if an *observation* represented by a vector $\mathbf{x} \in \mathbb{R}^n$ belongs to either one of two distinct *classes* arbitrarily labeled “A” and “B.” To do so, a *training set* $S = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i \in [1, m]\}$ of m pairs of n -vectors and their *label values* is initially given with the *convention* that the label value is -1 if the corresponding vector belongs to class A and 1 otherwise. The decision rule to be built from S is, therefore, a function $f : \mathbb{R}^n \rightarrow \{-1, 1\}$ that conventionally depends on some other real-valued function g over $\mathbb{R}^n$ such that

$$\begin{cases} f(\mathbf{x}) = -1 & \text{if } g(\mathbf{x}) < 0, \\ f(\mathbf{x}) = 1 & \text{otherwise.} \end{cases}$$

In SVM theory, one attempts to *map* observations into some *Hilbert space* $\mathcal{H}$, also called *feature space*, using the set S and the function g as an algebraic measure of distance from an *affine hyperplane* dividing $\mathcal{H}$ into two *half-spaces*, one half-space for mapped points belonging to class A and one half-space for mapped points belonging to class B. The mapping itself is realized with the help of the *kernel trick* for which a continuous, symmetric, and positive semidefinite function $k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, also called *kernel function* or *kernel*, can be expressed as a dot product in a Hilbert space. Since an affine hyperplane in $\mathcal{H}$ that has a normal vector $w \neq 0$ and a bias $b \in \mathbb{R}$ is the set $\{x \in \mathcal{H} : \langle w, x \rangle_{\mathcal{H}} + b = 0\}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the dot product in $\mathcal{H}$, one may define $\mathcal{H}$ as the space which dot product is *implied* by a given kernel k . If a kernel k is defined, then, by Mercer’s theorem, there exists a map $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ such that $k(\mathbf{w}, \mathbf{x}) = \langle \phi(\mathbf{w}), \phi(\mathbf{x}) \rangle_{\mathcal{H}}$ for all $(\mathbf{w}, \mathbf{x}) \in \mathbb{R}^{2n}$. Combining the kernel trick with the notion of algebraic distance from an affine hyperplane in $\mathcal{H}$, the function g then takes the form $g : \mathbf{x} \mapsto k(\mathbf{w}, \mathbf{x}) + b$,

where the vector $\mathbf{w} \in \mathbb{R}^n$ and the scalar $b \in \mathbb{R}$ are built using the training set S .

Given a Hilbert space $\mathcal{H}$ defined by a kernel k , there exists a possibility that no affine hyperplane can separate the mapped data, in which case changing k might give a new Hilbert space $\mathcal{H}$ where such a separation is possible. Now, if the mapped data coming from a source of unknown distribution is separable by an affine hyperplane in $\mathcal{H}$, then there is a null probability that only one such hyperplane exists, that is, there is almost always an infinity of such affine hyperplanes. The question is which one shall be chosen? In the SVM theory, one chooses the unique affine hyperplane that maximizes the distance between itself and the nearest mapped data. Given an affine hyperplane H in $\mathcal{H}$ with a nonzero normal vector $w \in \mathcal{H}$ and bias $b \in \mathbb{R}$, the distance between a point $x \in \mathcal{H}$ and H is $\delta(x, w, b) = |\langle w, x \rangle_{\mathcal{H}} + b| / \|w\|_{\mathcal{H}}$. Now, supposing that x^* is the mapped data nearest to H , we have $\delta(x^*, w, b) = \delta(x^*, \lambda w, \lambda b)$ for all scalar $\lambda \neq 0$. By imposing the *convention* that $|\langle w, x^* \rangle_{\mathcal{H}} + b| = 1$, the *unique* normal vector of the *optimal* affine hyperplane is found by minimizing the quantity $\|w\|_{\mathcal{H}}$ subject to the constraints that the mapped data are separated. Data separation is enforced by the set of constraints $y_i g(\mathbf{x}_i) > 0$ for all $i \in [1, m]$, although SVM theory chooses the *convention* that $y_i g(\mathbf{x}_i) \geq 1$, for all $i \in [1, m]$.

In all the following, the dot product in $\mathcal{H}$ and its corresponding norm are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. The map from $\mathbb{R}^n$ to $\mathcal{H}$ that is implied by the choice of a specific kernel k is denoted by ϕ .

Hard-Margin Support Vector Machines

The (primal) optimization problem for the so-called hard-margin SVMs is

$$\min_{w, b} \left\{ \|w\|^2 / 2 : y_i (\langle w, \phi(\mathbf{x}_i) \rangle + b) \geq 1, \forall i \in [1, m] \right\}.$$

Note that a feasible solution exists only if the data are separable by an affine hyperplane in $\mathcal{H}$. Introducing real positive Lagrange multipliers $u_1, \dots, u_m$, the Lagrangian for this problem is given by $L(w, b) = \|w\|^2 / 2 + \sum_{i=1}^m u_i (1 - y_i (\langle w, \phi(\mathbf{x}_i) \rangle + b))$. It follows that the Karush–Kuhn–Tucker

(KKT) conditions at optimality for this problem are

$$\begin{cases} \nabla L(w, b) = \mathbf{0}, \\ u_i(1 - y_i(\langle w, \phi(\mathbf{x}_i) \rangle + b)) = 0, \forall i \in [1, m]. \\ u_1, \dots, u_m \geq 0. \end{cases}$$

From the second condition, we immediately deduce that, at optimality, $u_i = 0$ or $u_i \neq 0$ and $y_i(\langle w, \phi(\mathbf{x}_i) \rangle + b) = 1$. The training vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ such that $u_i > 0$ are called *support vectors* and hence the name support vector machine. From the first KKT condition, we obtain $w = \sum_{i=1}^m u_i y_i \phi(\mathbf{x}_i)$ and $\langle \mathbf{u}, \mathbf{y} \rangle = 0$ with $\langle \cdot, \cdot \rangle$ being the Euclidean dot product [3]. When these two equalities are substituted into the Lagrangian, we obtain $L(\mathbf{u}) = \langle \mathbf{1}_m, \mathbf{u} \rangle - \mathbf{u}^T \mathbf{Q} \mathbf{u} / 2$, where $\mathbf{1}_m$ is the m -vector made of 1s and $\mathbf{Q}$ is the matrix whose (i, j) -th element is $\mathbf{Q}_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$. We then derive a dual formulation for the hard-margin SVM, which is, in fact, the canonical SVM problem

$$\min_{\mathbf{u} \in \mathbb{R}^m} \left\{ \mathbf{u}^T \mathbf{Q} \mathbf{u} / 2 - \langle \mathbf{1}_m, \mathbf{u} \rangle : \langle \mathbf{u}, \mathbf{y} \rangle = 0, \mathbf{u} \geq 0 \right\}.$$

Let $I_+ = \{i \in [1, m] : u_i > 0\}$, then, for an optimal solution $\mathbf{u}$, the function g is given by $g(\mathbf{x}) = \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$. The bias b is obtained from the second KKT condition by $b = y_j - \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}_j)$ for any $j \in I_+$.

Soft-Margin Support Vector Machines

When a feasible solution does not exist for the hard-margin SVMs, then the training set is not separable by an affine hyperplane. To enable a few outliers to exist while computing the separating hyperplane, a set of m nonnegative slack variables $v_1, \dots, v_m$ is introduced into the constraints of the primal problem, giving $y_i g(\mathbf{x}_i) \geq 1 - v_i$ for all $i \in [1, m]$. The form of these constraints is a *convention*. If there is no constraint on the sign of the slack variables, then one can choose equality constraints $y_i g(\mathbf{x}_i) = 1 + v_i$ for all $i \in [1, m]$ or any variant instead of the classical inequality constraints above. Soft-margin SVMs traditionally pick up the set of inequality constraints and therefore obtain a multiobjective optimization problem in which one has to maximize the margin

while minimizing a p -norm of the slack vector. To avoid dealing with such a difficult optimization problem, the two objectives are summed with one objective weighted by a trade-off coefficient $C > 0$.

The objective function is $\|\mathbf{w}\|^2/2 + C\|\mathbf{v}\|_p$ with $v_i \geq 0$ for all $i \in [1, m]$. When $p = 1$, we have the primal optimization problem of L_1 soft-margin SVM and we obtain the following dual quadratic programming problem

$$\min_{\mathbf{u} \in \mathbb{R}^m} \left\{ \mathbf{u}^T \mathbf{Q} \mathbf{u} / 2 - \langle \mathbf{1}_m, \mathbf{u} \rangle : \langle \mathbf{u}, \mathbf{y} \rangle = 0, \right. \\ \left. 0 \leq \mathbf{u} \leq C \right\}.$$

The sole difference between the L_1 soft-margin SVM formulation and the hard-margin SVM formulation is the upper bound on the Lagrange multipliers $u_1, \dots, u_m$. Let $I_+ = \{i \in [1, m] : u_i > 0\}$; then, for an optimal solution $\mathbf{u}$, the function g is given by $g(\mathbf{x}) = \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$. Let $I_- = \{i \in I_+ : u_i < C\}$; then the bias b is computed by $b = y_j - \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}_j)$ for any $j \in I_-$.

When $p = 2$, the resulting dual formulation for the L_2 SVM is

$$\min_{\mathbf{u} \in \mathbb{R}^m} \left\{ \mathbf{u}^T \mathbf{H} \mathbf{u} / 2 - \langle \mathbf{1}_m, \mathbf{u} \rangle : \langle \mathbf{u}, \mathbf{y} \rangle = 0, \mathbf{u} \geq 0 \right\},$$

with $\mathbf{H} = \mathbf{Q} + 2\mathbf{I}_m/C$ and $\mathbf{I}_m$ is the $m \times m$ identity matrix. This formulation is almost identical to the one for hard-margin SVMs. Let $I_+ = \{i \in [1, m] : u_i > 0\}$; then, for an optimal solution $\mathbf{u}$, the function g is given by $g(\mathbf{x}) = \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$. The bias b is computed by $b = y_j - \sum_{i \in I_+} u_i y_i k(\mathbf{x}_i, \mathbf{x}_j)$ for any $j \in I_+$.

Kernels

There is a wide variety of kernels, blossoming various families of such functions, each one having desirable properties for certain kind of problems. Kernels can be chosen from a family or built specifically for one particular problem. In the latter case, one should keep in mind that kernels represent a measure of distance and angle into some (possibly higher dimensional) Hilbert space. The considerations related to what constitutes the “closeness” of two distinct observations is

left to the person modeling the classification problem and the nature of the manipulated data. The most commonly used kernels over $\mathbb{R}^n \times \mathbb{R}^n$ are as follows:

- $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{Q} \mathbf{y}$ with $\mathbf{Q}$ being a symmetric positive semi-definite $n \times n$ -matrix. In particular, if $\mathbf{Q} = \mathbf{I}_n$, then $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ (Euclidean dot product).
- $k(\mathbf{x}, \mathbf{y}) = (\mu + \langle \mathbf{x}, \mathbf{y} \rangle)^\nu$ with $\mu \geq 0$ and $\nu \in \mathbb{N}^*$ (polynomial kernels). The associated Hilbert space is the space of all functions $\mathbf{x} \mapsto x_1^{\nu_1} x_2^{\nu_2} \dots x_n^{\nu_n}$ such that $\sum_{i=1}^n \nu_i \leq \nu$, which is a space of dimension $\binom{n+\nu}{\nu}$ [3].
- $k(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^n (1 + x_i y_i)$ (all-subsets kernels). The associated Hilbert space is the space of all functions $\mathbf{x} \mapsto x_1^{\nu_1} x_2^{\nu_2} \dots x_n^{\nu_n}$ such that $(\nu_1, \dots, \nu_n) \in \{0, 1\}^n$ [3].
- $k(\mathbf{x}, \mathbf{y}) = \sum_{1 \leq i_1 < \dots < i_\nu \leq n} \prod_{j=1}^\nu x_{i_j} y_{i_j}$ (ANOVA kernel of degree ν). The associated Hilbert space is the space of all functions $\mathbf{x} \mapsto x_1^{\nu_1} x_2^{\nu_2} \dots x_n^{\nu_n}$ such that $(\nu_1, \dots, \nu_n) \in \{0, 1\}^n$ and $\sum_{i=1}^n \nu_i = \nu$, which is a space of dimension $\binom{n}{\nu}$. A more general and flexible family of such kernels can be constructed from the so-called *graph kernel* [3].
- $k(\mathbf{x}, \mathbf{y}) = \exp(-\mu(\mathbf{x} - \mathbf{y})^T \mathbf{Q}(\mathbf{x} - \mathbf{y}))$ with $\mu > 0$; $\mathbf{Q}$ being a symmetric positive semidefinite $n \times n$ -matrix (generalized radial basis function kernels). These are the most widely used kernels, even though they are not robust to outliers [34]. The associated map $\mathbf{x} \mapsto k(\mathbf{x}, \cdot)$ has images into an infinite-dimensional Hilbert space [11].

Kernels exist for sets and structured data such as sequences and graphs [3]. They can be also built from other kernels or simply from a real-valued function f on the set that the observations belong to. If $\lambda \in \mathbb{R}_+$ and k_1 and k_2 are two kernels, then $k_1 + \lambda k_2$, $k_1 k_2$ (Schur product) and $(\mathbf{x}, \mathbf{y}) \mapsto f(\mathbf{x})f(\mathbf{y})$ are kernels. If p is a polynomial with positive coefficients, then $p(k)$ and $\exp(k)$ are also kernels.

Model Selection

Selecting a suitable kernel and an appropriate trade-off parameter C ensures the optimality of a binary classifier. However, this model selection is far from being straightforward but tools exist to help choosing effective models.

Cross-Validation. In a cross-validation, the training set S is *partitioned* into several subsets, which are alternatively used to build a classifier to be tested on the remaining subsets. In other words, the set S is iteratively divided into an actual training set, for which a classifier is built, and a validation set, for which the classifier is tested. At each iteration of the cross-validation, a *confusion matrix* $\mathbf{C}$ is constructed for the results of the validation set. For an ℓ -class problem, the matrix $\mathbf{C}$ is an $\ell \times \ell$ matrix for which each i -th row represents the number of samples belonging to class i and each j -th column represents the number of samples predicted to belong to class j . A measure of error on $\mathbf{C}$ is then devised in order to estimate the goodness of the chosen model. One such measure is $\varepsilon = \min_{i \in \ell} \{C_{ii} / (\mathbf{1}_\ell, \mathbf{C}_i)\}$ with $\mathbf{C}_i$ being the i th row of the matrix $\mathbf{C}$. In other words, this measure of error is the minimum fraction of correctly classified samples.

The repeated partition of S may be chosen randomly, although some samples might never be used for training or testing. Consequently, better cross-validation methods have been implemented and the most widely used ones are the l -fold cross-validation and its particular case, the leave-one-out (LOO) cross-validation. In an l -fold cross-validation, or more exactly in a *stratified* l -fold cross-validation, the set S is partitioned into k subsets of equal size that contains the same proportion of class labels as in S . During each iteration of the method, $l - 1$ subsets are used for training and the remaining subset is used for testing. Every subset is used for testing only once and hence a maximum of l iterations. At each iteration, the error ε can be seen as the result of a *Bernoulli trial* and its confidence interval can be easily calculated. The LOO cross-validation is an extreme case of the l -fold cross-validation in which only a single sample is taken out at each iteration

to be in the validation set. This last approach usually needs modifications in order to be efficiently implemented for large data sets.

Bootstrapping. Bootstrapping is a resampling technique for inferring sample statistics by drawing randomly with replacement several observations and testing them multiple times. An estimate of the distribution of the statistic is then obtained, which gives the possibility to validate different models.

Given a testing set of m observations and an SVM classifier, the following procedure is repeated l times (l is the number of rounds):

1. Sample n observations with replacement from the testing set.
2. Test the sample with the SVM.
3. Calculate and store a measure of error for the sample.

Once the procedure is finished, it is then possible to use the sampling distribution of the measures of error to estimate the median value and a confidence interval.

Pattern Search. If a finite number of kernels are available for a given problem, then cross-validation or bootstrapping schemes can be combined with a pattern search in order to isolate the best kernel and its best parameters, notably the trade-off parameter C . The objective is to minimize the measure of error returned by the cross-validations over the search domain. Nevertheless, this problem is no longer necessarily a convex optimization problem and deterministic pattern searches for convex problems are no longer guaranteed to successfully reach an optimal solution. Therefore, metaheuristics such as genetic algorithms, simulated annealing, tabu search, or ant colony optimization might be more suitable to minimize this objective.

Decomposition Techniques for Training SVMs

As highlighted in the previous sections, obtaining an SVM classification rule requires solving a quadratic programming problem. However, solving such problems on large data sets turns out to be a serious computational challenge. To circumvent this computational problem, Osuna *et al.* [35] were among the

first to introduce a decomposition technique, which was proved to be asymptotically convergent [36]. The outline of the decomposition method is as follows:

1. *Initialization* Set $i = 1$ and $q \in [2, m]$. Set $\mathbf{u}_1$ as initial solution of the SVM quadratic programming problem. Go to step 2.
2. If $\mathbf{u}_i$ is an optimal solution of the SVM problem then stop; otherwise, go to step 3.
3. Partition $[1, m]$ into two subsets S_1 and S_2 such that $|S_1| = q$. Let $\mathbf{v}_i$ and $\mathbf{w}_i$ be the subvectors of $\mathbf{u}_i$ corresponding to S_1 and S_2 respectively. Go to step 4.
4. Solve the L_1 soft-margin SVM programming problem for $\mathbf{v}_i$:

$$\min_{\mathbf{v}_i \in \mathbb{R}^q} \left\{ \frac{\mathbf{v}_i^T \mathbf{Q}_{11} \mathbf{v}_i}{2} - \langle \mathbf{1}_q - \mathbf{Q}_{12} \mathbf{w}_i, \mathbf{v}_i \rangle : \right. \\ \left. \langle \mathbf{y}, \mathbf{u}_i \rangle = 0, 0 \leq \mathbf{v}_i \leq \mathbf{C} \right\},$$

with $\mathbf{Q}_{\alpha\beta}$ being the submatrix of the matrix $\mathbf{Q}$ corresponding to the rows in S_α and the columns in S_β . Go to step 5.

5. Let $\mathbf{v}_{i+1}$ be the optimal solution of the problem in step 4. Set $\mathbf{w}_{i+1} = \mathbf{w}_i$; then $i \leftarrow i + 1$. Go to step 2.

Several partitioning methods have been proposed to obtain the subset S_1 . Osuna *et al.* [35] and Saunders *et al.* [37] cast observations that violate the KKT conditions into S_1 . In the sequential minimal optimization (SMO) method, Platt [38] imposes $q = 2$ so that the resulting quadratic optimization problem can be solved analytically. However, it also results that the choice of the elements of the subset S_1 is based on some heuristics. In the implementation of SVM^{light} [39], Joachims modifies the problem of step 4 and sets q to be an even number. A linear programming problem is derived so that the decision variable picks up the elements to be casted into S_1 [36].

MULTICLASS CLASSIFICATION SCHEMES

Two different family of approaches have been devised in order to deal with multiple-class

classification problems. The first family contains techniques that modify the structure of the basic SVM optimization problems in order to directly support and discriminate several classes. The all-at-once SVMs of Crammer and Singer [40] are designed in such a manner. The other family of approaches is more general in nature since the resulting multiclass classification schemes work with any type of binary classifier and not just with SVM binary classifiers alone. Vapnik proposed the one-against-all SVM [8], Platt *et al.* introduced a decision-tree-based scheme for multiclass classification [41], and Allwein *et al.* presented the use of error-correcting output codes for ℓ -class problems [42].

All-at-Once Support Vector Machines

In this framework [40], given an ℓ -class problem ($\ell \geq 2$), an observation $\mathbf{x} \in \mathbb{R}^n$ will belong to class $i \in [1, \ell]$ if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \in [1, \ell]$ such that $j \neq i$. The objective function for the primal problem is derived from the binary soft-margin SVM classifier. Assuming that $g_j(\mathbf{x}) = \langle \mathbf{w}_j, \phi(\mathbf{x}) \rangle$, then the objective function is $\sum_{j=1}^{\ell} \|\mathbf{w}_j\|^2/2 + C\|\mathbf{v}\|_1$ with $\mathbf{v} \geq \mathbf{0}$. The constraints are $(\mathbf{w}_{y_i} - \mathbf{w}_j)^T \phi(\mathbf{x}_i) \geq 1 - \delta_{y_i}^j - v_i$ for all $(i, j) \in [1, m] \times [1, \ell]$ with $\delta_{y_i}^j$ being the Kronecker symbol, that is, $\delta_{y_i}^j = 1$ if $j = y_i$ and 0 otherwise. The corresponding dual problem [43] is

$$\begin{aligned} \min_{\mathbf{u} \in \mathbb{R}^{m \times \ell}} \quad & \frac{1}{2} \mathbf{u}^T \mathbf{H} \mathbf{u} - \langle \mathbf{e}, \mathbf{u} \rangle \\ \text{subject to} \quad & \begin{cases} \langle \mathbf{1}_\ell, \mathbf{u}_i \rangle = 0, i \in [1, m], \\ 0 \leq (\mathbf{u}_i)_j \leq C \delta_{y_i}^j, (i, j) \in [1, m] \times [1, \ell]. \end{cases} \end{aligned}$$

The matrix $\mathbf{H}$ is defined by $\mathbf{H} = \mathbf{K} \otimes \mathbf{I}_\ell$ with $\otimes$ being the Kronecker product. The (i, j) th element of the matrix $\mathbf{K}$ is $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for all $(i, j) \in [1, m]^2$. The vector $\mathbf{u} \in \mathbb{R}^{m \times \ell}$ is composed of all the vectors $\mathbf{u}_i \in \mathbb{R}^\ell$, $i \in [1, m]$, in the following manner: $\mathbf{u}^T = (\mathbf{u}_1^T, \dots, \mathbf{u}_m^T)^T$. Similarly, the vector $\mathbf{e}$ is composed of all the vectors $\mathbf{e}_i \in \mathbb{R}^\ell$, where j -th elements are defined by $(\mathbf{e}_i)_j = \delta_{y_i}^j - 1$. The decision rule is given by $f(\mathbf{x}) = \arg \max_{j \in [1, \ell]} \sum_{i=1}^m (\mathbf{u}_i)_j k(\mathbf{x}_i, \mathbf{x})$. Determining the class of $\mathbf{x} \in \mathbb{R}^n$ requires ℓ decision evaluations when the decision functions are

trained all at once on a problem that has $m \times \ell$ variables.

One-Against-All Support Vector Machines

A simple multiclass classification scheme for ℓ classes was proposed by Vapnik [8] for binary classifiers having a real-valued decision function g that is strictly positive if a point $\mathbf{x} \in \mathbb{R}^n$ belongs to the class it discriminates and negative if it does not. In this scheme, a total of ℓ binary classifiers are trained, each one of them determining whether a given observation belongs to one single class or not, that is, if it belongs to the rest of the classes or not and hence the name *one-against-all*. Suppose that ℓ such decision functions $g_1, \dots, g_\ell$ are determined, then the final classification rule f for determining the class of a given observation $\mathbf{x} \in \mathbb{R}^n$ is $f(\mathbf{x}) = \arg \max_{i \in [1, \ell]} g_i(\mathbf{x})$. Like for the all-at-once SVM, determining the class of $\mathbf{x}$ requires ℓ decision evaluations but each decision function g_i is trained on $m = \sum_{i=1}^{\ell} m_i$ observations, m_i being the number of training observations belonging to class i .

Decision-Tree-Based Support Vector Machines

Platt *et al.* [41] introduced a decision-tree-based architecture for binary classifiers returning Boolean values. This architecture is the one of a *decision-directed acyclic graph* (DDAG) that is, as the name suggests, a scheme based on a directed graph with no cycles. Suppose that, for a ℓ -class classification problem, a total of $\ell(\ell - 1)/2$ pairwise binary classifiers are trained, that is, we are given decision rules f_{ij} discriminating between class $i \in [1, \ell]$ and class $j \in [1, \ell]$ with $i < j$.

The DDAG classifier has a pyramidal structure in which every node but the root node has two direct successors, and every node but the leaf nodes has two direct predecessors. A path from the root node to one of the leafs is computed by updating an ordered list of classes to be tested, each node successively removing one element from that list. The removing process works by picking up the decision rule f_{ij} discriminating the class at the top of the list, denoted by i , from the one at the bottom, denoted by j ,

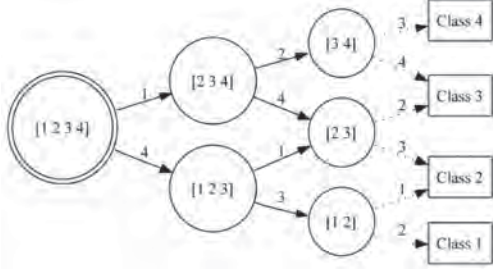


Figure 1. DDAG for a four-class problem. Each edge is labeled with the class number that has been removed from the list.

and testing whether the observation $\mathbf{x} \in \mathbb{R}^n$ belongs to class i or class j . If, for example, $\mathbf{x}$ belongs to class i , then the next node on the path would be the node which list does not contain j and reciprocally; hence, the reason for two direct successors at each node. The path starts at the root node with initial list $\{1, \dots, \ell\}$ and goes to one leaf node with a list containing only two classes. Once the leaf node has been reached, the last decision is made and the last class is removed from the list, the remaining one being the class $\mathbf{x}$ belongs to (Figure 1). Determining the class of $\mathbf{x}$ requires $\ell - 1$ decision evaluations and each decision function g_{ij} is trained on $m_i + m_j$ observations, m_i being the number of training observations belonging to class i and m_j being the number of training observations belonging to class j .

Error-Correcting Output Code SVMs

Suppose $l \geq 1$ Boolean decision functions $f_1, \dots, f_l$ are given for an ℓ -class problem. A function f_i , $i \in [1, l]$, will return either $+1$ or -1 depending on which class the input belongs to. If the function f_i was not trained to discriminate one class, then, by convention, the output will be zero for this class. An example is given in Table 1 where all outputs are arranged into a decision matrix.

Denoting an $\ell \times l$ decision matrix built from $f_1, \dots, f_l$ by $\mathbf{D}$, we can notice that each row of $\mathbf{D}$ represents a code that we seek to be unique for each class. If it is the case, then for an observation $\mathbf{x} \in \mathbb{R}^n$, the outputs for every f_i , $i \in [1, l]$, are put in order so it forms the code for $\mathbf{x}$, which is then stored

Table 1. Example of Error-Correcting Codes for a Four-Class Problem and Five Decision Functions

Class	f_1	f_2	f_3	f_4	f_5
1	0	-1	-1	-1	+1
2	+1	-1	0	+1	+1
3	+1	-1	-1	-1	0
4	-1	+1	+1	0	-1

in an $l \times 1$ vector $c(\mathbf{x})$. This code is then compared to each row of $\mathbf{D}$ by computing a distance between $c(\mathbf{x})$ and the codes on the rows of $\mathbf{D}$. The class of $\mathbf{x}$ will be the one with the minimum distance between $c(\mathbf{x})$ and the code on the row corresponding to that class. The distance proposed by Allwein *et al.* [42] between $\mathbf{x}$ and the i th class is $d(c(\mathbf{x}), \mathbf{D}_i) = \sum_{j=1}^l (1 - \text{sign}([c(\mathbf{x})]_j \mathbf{D}_{ij})) / 2$ and the final decision rule is $f(\mathbf{x}) = \arg \min_{i \in [1, \ell]} d(c(\mathbf{x}), \mathbf{D}_i)$. This approach is, in a way, a generalization of the one-against-all technique and the decision-tree-based technique for which a potentially lesser number of decision functions is needed in order to determine classes. Furthermore, depending on the codes contained in $\mathbf{D}$, the size of the programming problems to be solved can be significantly reduced. The contents of $\mathbf{D}$ also suggest that there exists optimal codes that combine a high discriminating power and a short number of decision functions trained on smaller problems.

MACHINES BASED ON CENTRAL REPRESENTATIONS OF THE VERSION SPACE

More recent approaches consider the concept of the center of a convex set together with general nonlinear programming techniques. They aim at improving the generalization performances over standard SVMs and at creating *sparse hypotheses*. The interpretation of these approaches is based on the observation that the dual of the data points in the input space become hyperplanes and separating hyperplanes become points reciprocally. From a Bayesian point of view, an SVM solution can be viewed as the center of the *largest hypersphere* that can be inscribed in the set of

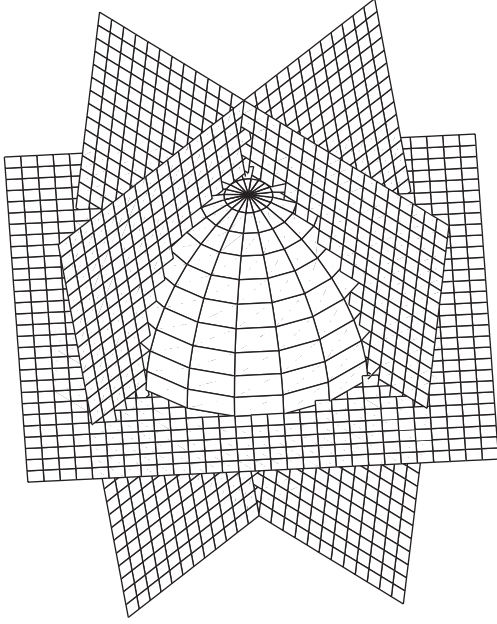


Figure 2. Example of a version space in $\mathbb{R}^3$. The version space is the surface segment of the hypersphere shown here restricted by hyperplanes in which consistent hypotheses can lie. Each hyperplane corresponds to a labeled training instance.

all consistent *hypotheses* called the *version space* (Figure 2). The version space boundaries are tangentially touched by the inscribed hypersphere, the support vectors corresponding to samples with hyperplanes touching the hypersphere [44,45]. Mathematically, given a training set $S = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i \in [1, m]\}$ of m observations and a map $\phi : \mathbb{R}^n \mapsto \mathcal{H}$ to the feature space $\mathcal{H}$, the version space $\mathcal{V}$ is defined as $\mathcal{V} = \{w \in \mathcal{H} : \|w\| = 1, y_i \langle w, \phi(\mathbf{x}_i) \rangle \geq 0, i \in [1, m]\}$.

Bayes Point Machines

If a version space is asymmetric or elongated, then the center of the largest inscribed hypersphere does not appear to be the best choice. An alternative approach with a potentially better generalization performance is to find an optimal solution using the Bayes point approximated by the center of mass of the version space. However, calculating this center requires the computation of the version space volume, which is a #P-hard problem [46]. Several methods have been proposed

to approximate the center of mass of the version space. *Bayes point machines* (BPMs) developed by Herbrich *et al.* [44], Ruján [47], and Minka [48] estimate the Bayes point or the center of mass by averaging over weight vectors drawn according to the uniform posterior distribution. BPMs construct a hypothesis based on this center of the version space and this choice can be justified by theoretical arguments [49,50] in addition to having a geometric appeal. Given a training set $S = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i \in [1, m]\}$ of m observations, a map $\phi : \mathbb{R}^n \mapsto \mathcal{H}$ to the feature space $\mathcal{H}$ and a decision rule $f(\mathbf{x}) = \text{sign}(\langle w, \phi(\mathbf{x}) \rangle)$, the Bayes classification strategy is given by

$$\begin{aligned} f_{\text{Bayes}}(\mathbf{x}) &= \text{sign} \left(\int_{\mathcal{V}} f(\mathbf{x}) p(w|S) dw \right) \\ &= \text{sign} \left(\int_{\mathcal{V}} \text{sign}(\langle w, \phi(\mathbf{x}) \rangle) p(w|S) dw \right) \\ &\approx \text{sign} \left(\text{sign} \int_{\mathcal{V}} \langle w, \phi(\mathbf{x}) \rangle p(w|S) dw \right) \\ &= \text{sign} \left(\left\langle \int_{\mathcal{V}} w p(w|S) dw, \phi(\mathbf{x}) \right\rangle \right), \end{aligned}$$

where $p(w|S)$ is the probability of the weight vector w given the training set S . The center of mass is $w_{\text{CM}} = \int_{\mathcal{V}} w p(w|S) dw$ and therefore the Bayes point can be approximated by the center of mass under the uniform posterior distribution condition that is, $w_{\text{Bayes}} \approx w_{\text{CM}}$.

In one approach [47], the center of mass is determined using a kernelized billiard algorithm in which the version space is traversed uniformly and an estimate of the center of mass is repeatedly updated. Informally, the center of mass is estimated by averaging over the trajectory of a billiard ball bouncing in version space. In another approach, the perceptron algorithm is used to obtain a large number of classifiers in the version space by learning on different permutations of the training set. Then, the Bayes point is approximated using the average of the weight vectors of those classifiers. Their results showed that BPMs have better generalization performances compared to SVMs in the hard

boundary or hard-margin case, but the performance improvement over SVMs in the soft boundary case is decreased.

Relevance Vector Machines

For a large majority of data sets, the version space diverges from sphericity and a BPM outperforms an SVM at statistically significant levels. For artificial examples with very elongated version spaces, the generalization error of a BPM can be half that of an SVM [44,51]. However, current implementations of BPMs have a number of drawbacks: the algorithm can be slow in execution and better mechanisms for a soft boundary (an imitation of a soft margin) need to be found. The BPM may exhibit good generalization performances, but it has the disadvantage that the hypothesis is dense that is, nearly all data points appear in the final hypothesis. Ideally, we would also like to derive kernel classifiers that give sparse hypotheses using a minimal number of data points. The most effective means of obtaining sparse hypotheses remains the object of research, but an excellent approach is the *relevance vector machine* (RVM) of Tipping [52]. The RVM formulation utilizes a probabilistic Bayesian learning framework in order to obtain sparse solutions and provide probabilistic outputs. Other advantages of RVMs are automatic estimation of parameters and the ability to use arbitrary basis functions. The results showed that performance of RVMs is comparable to SVMs in terms of generalization.

Analytic Center Machines

Rather than using the center of mass of version space, an alternative might be to use a hypothesis that lies toward the center of this space but which is much easier to compute. Such an example is the *analytic center machine* (ACM) (Fig. 3).

On the basis of the utilization of the analytic center of the version space, Trafalis and Malyscheff [45] proposed ACMs for binary classification. Their experimental results indicated that the ACM also outperforms SVMs in the hard boundary case. The ACM uses a logarithmic barrier function whose value on a point grows to infinity as the point approaches the boundary of the version space. The analytic center of the version space is then determined when this function achieves its minimum and can be written as

$$w_{\text{ACM}} := \arg \min_{w \in \mathcal{V}} \left\{ - \sum_{i=1}^m \ln(y_i \langle w, \phi(\mathbf{x}_i) \rangle + b) \right\}.$$

Suppose we are given a kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and $\mathbf{w} = \sum_{i=1}^m u_i \phi(\mathbf{x}_i) + v$, the ACM formulation can be defined in the space of $(\mathbf{u}, v) \in \mathbb{R}^{m+1}$ as

$$\min_{(\mathbf{u}, v) \in \mathbb{R}^{m+1}} \left\{ - \sum_{i=1}^m \ln(\langle \mathbf{u}, y_i \mathbf{K}_i \rangle + y_i v) : \|\langle \mathbf{u}, v \rangle\|^2 = 1 \right\},$$

where $\mathbf{K}$ is the kernel matrix which (i, j) -th element is $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for all $(i, j) \in [1, m]^2$. Trafalis and Malyscheff solved this optimization problem using a projected Newton descent method and proposed an online

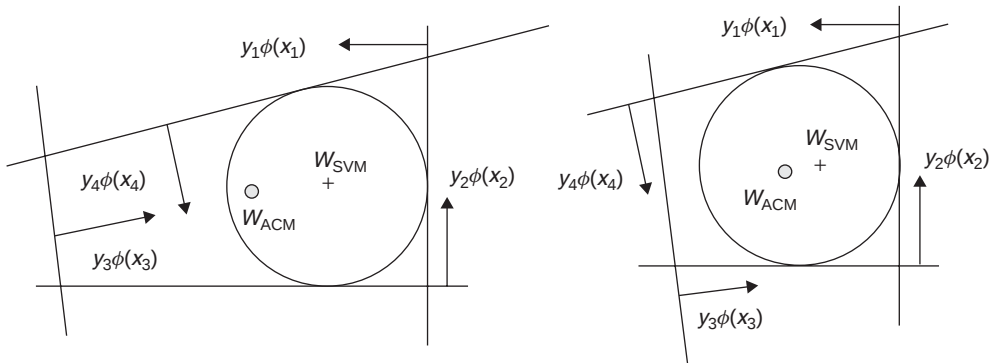


Figure 3. The analytic center of version space (w_{ACM}) and the center of the largest inscribed sphere (w_{SVM}) in an elongated version space.

approach in the implementation of ACM. They also suggested that a better classification performance can be achieved by removing redundant constraints that are not essential for the description of the version space.

p-Center Machines

Another approach proposed by Brückner [53] is based on an approximation of the *p*-center of the version space, the so-called *p*-center machine (*p*-CM). In the *p*-CM formulation, Brückner utilizes the *p*-center, which was initially proposed by Moretti [54] who introduced an efficient iterative algorithm to find the center of a polytope using a weighted projection centering method. Moretti showed that the *p*-center can be achieved in polynomial time and that the algorithm is parallelizable so that the computational effort can be improved significantly. Following this work, Brückner indicated that the *p*-center is also close to the Bayes point and that the *p*-CM has similar performances to SVMs and BPMs.

CONCLUSIONS

In this article we have provided a review of the theory of SVMs from the perspective of classification. Several modifications of SVMs, such as the ACM, the BPM, and the *p*-center machine, have been also discussed. In addition, kernel methods, model selection techniques, and computational issues have been investigated. There were several approaches that have not been discussed such as classification problems with noisy data [29,55,56], SVMs clustering and regression, as well as incremental SVMs [57,58]. SVMs and kernel methods have been found to work well in practice; however, the subject is still very much under development, but it can be expected to develop an important tool for data mining, machine learning, and their many applications.

REFERENCES

1. Abe S. Support vector machines for pattern classification. London: Springer; 2005.
2. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
3. Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge: Cambridge University Press; 2004.
4. Schölkopf B, Burges CJC, Smola AJ. Advances in kernel methods: support vector learning. Cambridge (MA): MIT Press; 1999.
5. Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge (MA): MIT Press; 2002.
6. Smola AJ. Advances in large margin classifiers. Cambridge (MA): MIT Press; 2000.
7. Vapnik VN. Estimation of dependences based on empirical data. New York: Springer; 1982.
8. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
9. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
10. Poggio T, Rifkin R, Mukherjee S, *et al.* General conditions for predictivity in learning theory. *Nature* 2004;428(6981):419–422.
11. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 1998;2(2):121–167.
12. Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 2001; 17(4):349–358.
13. Lee Y-J, Mangasarian OL, Wolberg WH. Survival-time classification of breast cancer patients. *Comput Optim Appl* 2003;25(1–3): 151–166.
14. Santosa B, Conway T, Trafalis TB. Knowledge-based clustering and application of multi-class support vector machines for genes expression analysis. In: Dagli CH, Buczak AL, Ghosh J, editors. Volume 12, Intelligent engineering systems through artificial neural networks. New York: ASME Press; 2002. pp. 391–396.
15. Oladunni OO, Trafalis TB. Single phase laminar and turbulent flow classification in annulus via a knowledge-based linear model. In: Dagli CH, editor. Volume 16, Intelligent engineering systems through artificial neural networks. New York: ASME Press; 2006. pp. 599–604.
16. Oladunni OO, Trafalis TB, Papavassiliou DV. Knowledge-based multiclass support vector machines applied to vertical two-phase flow. In: Alexandrov VN, van Albada GD, Sloot

- PM, editors. Volume 3991, Computational Science—ICCS 2006, Lecture Notes in Computer Science. Berlin: Springer; 2006. pp. 188–195.
17. Trafalis TB, Oladunni OO, Papavassiliou DV. Two-phase flow regime identification with a multiclassification support vector machine (SVM) model. *Ind Eng Chem Res* 2005;44(12):4414–4426.
18. Raman S, Trafalis TB, Gilbert RC. Adaptive methods in coordinate metrology. In: Gopalakrishnan B, editor. Volume 5999, Intelligent systems in design and manufacturing VI. Bellingham (WA): SPIE; 2005. pp. 59990A.
19. Adrianto I, Smith TM, Scharfenberg K, *et al.* Evaluation of various algorithms and display concepts for weather forecasting. Combined preprints 85th AMS Annual Meeting. Boston (MA): American Meteorological Society; 2005. pp. 351–356.
20. Mansouri H, Gilbert RC, Trafalis TB, *et al.* Ocean surface wind vector forecasting using support vector regression. In: Dagli CH, Buczak AL, Enke DL, editors. Volume 17, Smart systems engineering: computational intelligence in architecting complex engineering systems. New York: ASME Press; 2007. pp. 333–338.
21. Trafalis TB, Adrianto I, Richman MB. Active learning with support vector machines for tornado prediction. In: Shi Y, van Albada GD, Dongarra JJ, editors. Volume 4487, Computational Science—ICCS 2007, Lecture Notes in Computer Science. Berlin: Springer; 2007. pp. 1130–1137.
22. Malysheff AM, Trafalis TB. Support vector machines and the electoral college. Volume 3, Proceedings of the 2003 IEEE International Joint Conference on Neural Networks. Piscataway (NJ): IEEE; 2003. pp. 2344–2348.
23. Ince H, Trafalis TB. Kernel methods for short-term portfolio management. *Expert Syst Appl* 2006;30(3):535–542.
24. Ince H, Trafalis TB. A hybrid model for exchange rate prediction. *Decis Support Syst* 2006;42(2):1054–1062.
25. Ince H, Trafalis TB. Short term forecasting with support vector machines and application to stock price prediction. In: Dagli CH, Buczak AL, Ghö J, *et al.*, editors. Volume 13, Intelligent engineering systems through artificial neural networks. New York (NY): ASME Press; 2003. pp. 737–742.
26. Ince H, Trafalis TB. Kernel principal component analysis and support vector machines for stock price prediction. *IEE Trans* 2007;39(6):629–637.
27. Alenezi A, Moses S, Trafalis TB. Flowtime estimation for multi-resource production systems using support vector regression. In: Dagli CH, Buczak AL, Embrechts MJ, *et al.*, editors. Volume 15, Intelligent engineering systems through artificial neural networks. New York: AMSE Press; 2005. pp. 699–702.
28. Beardslee EA, Trafalis TB. Data mining methods in a metrics-deprived inventory transactions environment. In: Zanasi A, Brebbia CA, Ebecken NF, editors. Data mining VI: data mining, text mining and their business applications. Amherst (NY): WIT Press; 2005. pp. 513–522.
29. Santosa B, Trafalis TB. Robust multiclass kernel-based classifiers. *Comput Optim Appl* 2007;38(2):261–279.
30. Shung Chung W, Gruenwald L, Trafalis TB. Support vector clustering for web usage mining. In: Dagli CH, Buczak AL, Ghö J, *et al.*, editors. Volume 12, Intelligent engineering systems through artificial neural networks. New York: ASME Press; 2002. pp. 385–390.
31. Malysheff AM, Trafalis TB, Raman S. From support vector machine learning to the determination of the minimum enclosing zone. *Comput Ind Eng* 2002;42(1): 59–74.
32. Prakashvudhisarn C, Trafalis TB, Raman S. Support vector regression for determination of minimum zone. *J Manuf Sci Eng* 2003;125(4):736–739.
33. Trafalis TB. Foreword. *Comput Manag Sci* 2006;3(2):101–102.
34. Chen J-H. M-estimator based robust kernels for support vector machines. In: Kittler J, Petrou M, Nixon MS, *et al.*, editors. Proceedings of the 17th International Conference on Pattern Recognition—ICPR 2004. Los Alamitos (CA): IEEE Computer Society Press; 2004.
35. Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. In: Principe J, editor. Volume 7, Neural networks for signal processing. New York: IEEE; 1997. pp. 276–285.
36. Lin C-J. On the convergence of the decomposition method for support vector machines. *IEEE Trans Neural Netw* 2001;12(6):1288–1298.

37. Saunders C, Stitson MO, Weston J, *et al.* Support vector machine reference manual. Technical Report CSD-TR-98-03. Egham: Royal Holloway, University of London; 1998.
38. Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ, editors. *Advances in kernel methods: support vector learning*. Cambridge (MA): MIT Press; 1999. pp. 185–208.
39. Joachims T. Making large-scale support vector machine learning practical. In: Schölkopf B, Burges CJC, Smola AJ, editors. *Advances in kernel methods: support vector learning*. Cambridge (MA): MIT Press; 1999. pp. 169–184.
40. Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems. In: Cesa-Bianchi N, editor. *Proceedings of the 13th Annual Conference on Computational Learning Theory—COLT 2000*. San Francisco (CA): Morgan Kaufmann; 2000. pp. 35–46.
41. Platt JC, Cristianini N, Shawe-taylor J. Large margin DAGs for multiclass classification. In: Solla SA, Müller K-R, Leen TK, editors. *Advances in neural information processing systems 12*. Cambridge (MA): MIT Press; 2000. pp. 61–73.
42. Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *J Mach Learn Res* 2000;1:113–141.
43. Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw* 2002;13(2):415–425.
44. Herbrich R, Graepel TH, Campbell C. Bayes point machines. *J Mach Learn Res* 2001;1(4):245–279.
45. Trafalis TB, Malyscheff AM. An analytic center machine. *Mach Learn* 2002;46(1-3):203–223.
46. Kaiser MJ, Morin TL, Trafalis TB. Centers and invariant points of convex bodies. In: Gritzmann P, Sturmfels B, Klee V, editors. *Applied geometry and discrete mathematics: the victor klee festschrift: discrete mathematics and theoretical computer science*. Providence (RI): American Mathematical Society; 1991. pp. 367–385.
47. Ruján P. Playing billiards in version space. *Neural Comput* 1997;9(1):99–122.
48. Minka TP. A family of algorithms for approximate Bayesian inference [PhD thesis]. Cambridge (MA): Massachusetts Institute of Technology; 2001.
49. Oppen M, Haussler D. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Phys Rev Lett* 1991;66(20):2677–2680.
50. Watkin TLH, Rau A, Biehl M. The statistical mechanics of learning a rule. *Rev Mod Phys* 1993;65(2):499–556.
51. Herbrich R, Graepel TH, Campbell C. Robust bayes point machines. In: Verleysen M, editor. *Proceedings of ESANN '2000: 8th European Symposium on Artificial Neural Networks*. Brussels: D-Facto; 2000. pp. 49–54.
52. Tipping ME. The relevance vector machine. In: Solla SA, Müller K-R, Leen TK, editors. *Advances in neural information processing systems 12*. Cambridge (MA): MIT Press; 2000. pp. 652–658.
53. Brückner M. The p-center machine. *Proceedings of the International Joint Conference on Neural Networks: IJCNN 2005*. Piscataway (NJ): IEEE Operations Center; 2005. pp. 1000–1005.
54. Moretti AC. A weighted projection centering method. *Comput Appl Math* 2003;22(1): 19–36.
55. Trafalis TB, Gilbert RC. Robust classification and regression using support vector machines. *Eur J Oper Res* 2006;173(3):893–890.
56. Trafalis TB, Gilbert RC. Robust support vector machines for classification and computational issues. *Optim Methods Softw* 2007;22(1):187–198.
57. Son H-J, Trafalis TB. Detection of tornados using an incremental revised support vector machine with filters. In: Alexandrov VN, Dongarra JJ, Sloot PMA, *et al.*, editors. *Volume 3993, Computational Science—ICCS 2006, Lecture Notes in Computer Science*. Berlin: Springer. pp. 506–513.
58. Son H-J, Trafalis TB, Richman MB. Determination of the optimal batch size in incremental approaches: an application to tornado detection. Volume 5, *Proceedings of the IEEE International Joint Conference on Neural Networks 2005 (IJCNN'05)*. Piscataway (NJ): IEEE Operations Center; 2005. pp. 2706–2710.

SUPPORTING THE STRATEGY PROCESS: THE ROLE OF OPERATIONAL RESEARCH/ MANAGEMENT SCIENCE

FRANCES O'BRIEN
Warwick Business School,
University of Warwick,
Coventry, UK

INTRODUCTION

To explore how operational research/management science (OR/MS) can support the strategy process, it is helpful first to define what is meant by the terms strategy and strategy process. However, defining the term “strategy” is not as simple as it may seem. As De Wit and Meyer note, “. . . a common definition of the term strategy is illusive,” because there are “. . . strongly differing opinions on most key issues and disagreements run deep. . .” [1, p. 3]. The traditional view is that strategy is something an organization *has* [2], consisting of “. . . the direction and scope of an organization over the long term, which achieves advantage in a changing environment through its configuration of resources and competences with the aim of fulfilling stakeholder expectations.” [3, p. 9]. Implicit within this definition is the notion of strategic decisions that need to be made within the organization and external issues that the organization may have to face.

Every organization faces issues and decisions that are of strategic importance; examples of strategic decisions abound in every sector at every level: a retailer may wish to consider a decision to move into new markets or offer new product ranges; a manufacturer, facing decreasing demand for their products, may be forced to reduce production and to close factories. Strategic decisions have a number of characteristics. For example, Coyne and Subramanian [4] describe strategic decisions as those which

- drive or shape most of a company's subsequent actions;
- are not easily changed once made;
- have the greatest impact on whether the company's strategic objectives are met.

Pidd [5] adds to this by noting that since strategic decision making is crucial to an organization's survival, the “temperature of the debate can be very high” and that there may be considerable confusion over the objectives. He also notes that strategic decisions are often complex, involving many different interacting factors that have to be dealt with by individuals who may each see things differently. Dyson [6] notes that strategic decisions require investment of significant resources, and often involve a delay between their implementation and realization of results. In addition, Kirkwood [7] points out that it is often difficult retrospectively to determine whether an option chosen was the right one.

Organizations also face issues that are of strategic importance and where the need to make a decision is not immediately obvious, until some further understanding of the issue and its impact on the organization is gained. Consider the following example. As a result of the recent turmoil in the financial markets, a national government is considering revising the regulatory system for their financial services sector. For an organization operating in such a sector, the impact of such an environmental change may not be immediately obvious. Managers would want to understand the nature of the issues involved and how they might impact on their business both in the immediate here and now and over the longer term. An analysis of the issues may lead managers to review and revise the organization's direction, which in turn could lead to a review of its current decisions and policies and the development and evaluation of new ideas. The activities relevant to this example are those of making sense of issues, setting direction, and revising existing or developing and evaluating new strategic options. Such

activities can be connected into a process of interrelated activities that together bring about the strategic development of an organization.

Processes that facilitate the development of strategy within organizations can be differentiated by the extent to which they are deliberate or emergent. Johnson *et al.* [3] describe deliberate or intended strategy processes as those including strategic planning activities, strategy workshops, and the use of external consultants, each of which are designed to proactively explore issues and develop strategic ideas and recommendations prior to their implementation. In contrast, they describe emergent strategy processes as including experimentation and learning from partial commitments (logical incrementalism), resource allocation routines, cultural processes, and organizational politics. Under this view, strategy is not something that is planned; rather, it is something which emerges over time. A number of authors [3,8,9] note that it is helpful to see strategy as a blend of both the intended and the emergent; indeed, Hart and Banbury suggest that adopting multiple strategy processes improves organizational performance [8]. This review is primarily situated with the intended or planned view of strategy in that it describes a process for intentionally developing strategy through a set of interrelated activities and identifies tools that can be used to support such activities. However, it is important to remember that the debates involved within an activity revolve around people who may have strongly held views. Thus, in practice, supporting the strategy process at times exhibits emergent characteristics.

This article focuses on how the process of developing strategy within organizations can be supported with approaches from both the OR/MS and other fields. It begins by introducing a number of key activities that together form the strategy process. Next, it considers the issue of how the strategy process may be supported by reviewing the literature covering the use of approaches to support strategy. Using this literature, the article explains how a variety of approaches can be brought together and mapped onto the elements of

the strategy process as support mechanisms. Finally, a number of suggestions are made regarding future research directions relevant to this field.

STRATEGIC DEVELOPMENT PROCESSES

Just as there are many different definitions of strategy, there are also numerous versions of the strategy process [1,3,9]. For this article, the strategic development process originating from Dyson and Foster's empirical research [10,11] has been chosen since it contains many of the activities found in other strategy processes, but more importantly because it promotes the importance of strategy rehearsal, prior to implementation—a feature that sits comfortably within the ethos of the OR/MS discipline. At its core are a number of activities which Dyson and Foster suggest are “essential” for effective strategy development [10,11]. The following list of activities has been taken from more recent theoretical research [12], which has evolved from this earlier work:

- Strategic direction (where are we heading?):
 - setting direction, including vision and mission;
 - setting objectives, priorities, goals, targets, and performance measures—all informed by and aligned with direction.
- Current and future strategic position (what are we facing, now and in the future?):
 - assessing the external environment (social, political, economic, competitive, technological, and regulatory issues);
 - assessing the internal environment (resources and capabilities).
- Strategic choices (how can we get there?):
 - generating ideas for strategic initiatives/options;
 - rehearsing and revising ideas in a virtual feedback process that incorporates learning from virtual or projected performance;

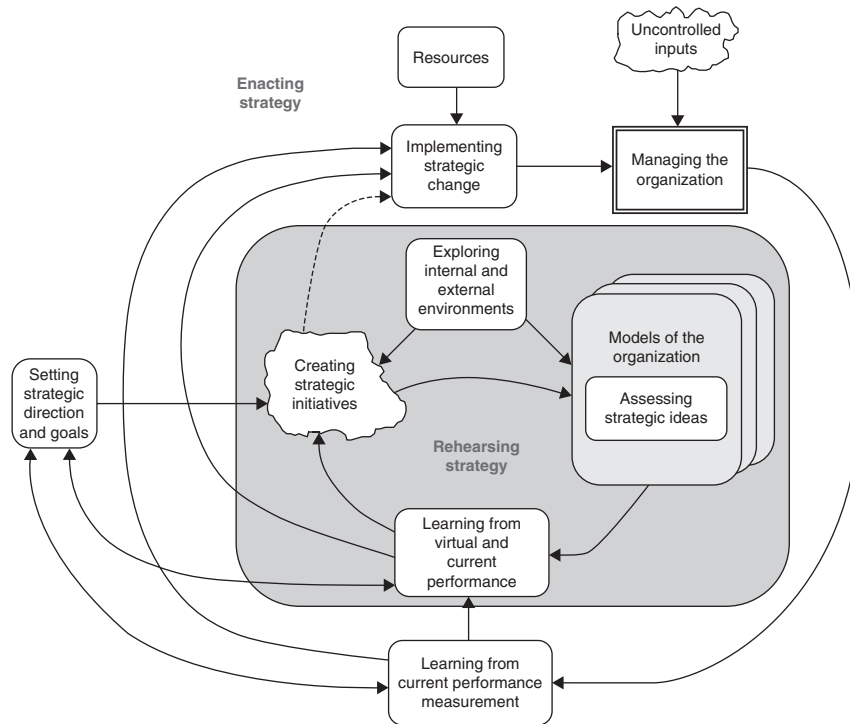


Figure 1. The strategic development process [12, Figure 1.7, p. 12].

- selecting initiatives/options.
- Making it happen (how are we progressing?):
 - learning from current organizational performance;
 - optimizing operational efficiency;
 - implementing strategic decisions/strategic change (i.e., enacting strategy).

Figure 1 [12, p.12] demonstrates how the activities listed above interrelate with one another and together form a strategy process. It is noticeable that the process is not linear but is framed around a closed, feedback loop, where the outcome of the process, learning from current performance, feeds back to influence both ongoing change and the direction-setting process. While the logical starting point from a conceptual perspective is the setting of direction, presenting the process as an interrelated system of activities

indicates that it does not have to be followed in a strictly linear fashion from one starting point. For example, managers may meet to consider how they may assess some issues in the external environment, which may lead them to reconsider the direction and initiatives of the organization.

Another feature of the process presented in Fig. 1 is the clear distinction made between rehearsing and enacting strategy, where the rehearsal of ideas and the examination of their consequences prior to implementation are seen as important aspects of the decision-making process. The figure shows that the rehearsal process itself is an iterative loop involving the use of a model or models, highlighting that a number of different modeling approaches can be used to help managers think through the consequences of their ideas.

SUPPORTING STRATEGIC DEVELOPMENT PROCESSES

A natural place to look for approaches to support the strategy process would be some of the mainstream strategy textbooks. A cursory glance of such textbooks [3,9,13] would reveal a number of classic approaches including PEST analysis (political, economic, social, technological) for exploring the organization's external environment, Porter's five forces (threat of new entrants and substitutes, bargaining power of suppliers and buyers, and intensity of rivalry) for exploring competitive forces within an industry, and SWOT analysis (strengths, weaknesses, opportunities, and threats) for exploring the organization and its environment, amongst others. One notable feature of such approaches is that individually they do not support the whole process; rather, they provide a particular perspective for part of the process or a specific activity. Thus, it is quite usual to find multiple tools being brought together to offer insights from different perspectives and levels. Another noticeable feature is that the reader is unlikely to find many approaches familiar to the OR/MS practitioner, apart from scenario planning, a tool for exploring future uncertainty in the external environment. This is not to say that the OR/MS has little to offer strategy support as the following paragraphs demonstrate.

Historically, OR/MS has been concerned with supporting decision makers through the use of "...appropriate analytical methods..." [14]. OR/MS approaches are characterized by their rational and systematic nature. Pidd [5] distinguishes between two forms of rationality and thus the nature of support that approaches can provide to strategy development: substantive rationality (rational choice) and procedural rationality (process of rational choosing). As he explains, "...substantive rationality is suited to situations in which the means to an end are uncertain, but the ends are known and the problem is finding some way to choose between these options." (p. 793) However, when the ends are not known or agreed upon and thus form the actual topic under consideration, procedural rationality

may be more helpful since it "...requires processes and tools that support the search for alternatives that encourage systematic information gathering and analysis and that help people find acceptable solutions when there is conflict." (p. 794)

Traditional OR/MS approaches, largely quantitative, typically belong to the realm of substantively rational approaches. Consider the work of Freeman who describes the use of a strategic simulation model "...the aim of which was to translate wide-ranging demands on the force into detailed operational staffing requirements." [15, p. 187]. Another example is the work of Vasko and Stott [16], who present the formulation of a linear programming (LP) model used to support the evaluation of a particular strategic option. Their work considered the issue of whether an idle facility should be opened in order to boost overall production capacity since it was believed that such an action might enable the organization to capture more of the existing market for the relevant products. In using such quantitative tools, the OR/MS practitioner is often in the role of expert modeler whose role is to undertake analysis of an issue or decision and provide interpretations and/or recommendations based on their analytical efforts. However, Eden [17] argues that strategy development would be more effective if seen as a "social process" rather than an analytical one, where the role for OR/MS is to provide facilitation, using appropriate tools, rather than to provide the actual analysis needed in a problem-solving mode. The development of problem structuring methods (PSMs) or soft OR approaches [18], evident over the past 20 years fits well with Eden's view. PSMs are largely qualitative approaches suited to problem situations characterized by

- "multiple actors,
- differing perspectives,
- partially conflicting interests,
- significant intangibles,
- perplexing uncertainties." [19].

Pidd [5] explains that "the soft approaches aim to provide decision support that is procedurally rational, although how they do this

varies...[They] all aim to help people to find enough accommodation to make progress on wicked problems without ignoring the realities of power and influence.” (p. 798) He gives the examples of soft systems methodology, which provides a combination of frameworks (CATWOE—customers, actors, transformation process, worldview, owner, environmental constraints) and approaches (rich pictures and activity modeling) to help people think through their different worldviews as well as SODA (strategic options development and analysis), using cognitive mapping, and SCA (the strategic choice approach), which both provide guidance on how to structure and manage the steps of an intervention.

Clearly, as there are a number of different activities within the strategy process, there are multiple possible roles for OR/MS in supporting strategy, since it is not possible to support the whole process with a single approach. Some aspects of the process can usefully be supported through the processes of group facilitation and negotiation, while others can be supported through the use of expert analysis that can be fed into the strategy process. Ormerod [20, p. 118] notes that the opportunities for OR/MS in supporting strategy are “multiple and varied” and may involve “...help with analysis or help with the process or both.” Dyson [6] advocates supporting the strategy process through the use of both OR/MS and tools from other fields, where the term “tool” “...is used as a generic name for any method, model, technique, tool, framework, methodology or approach used to provide decision support. Tool therefore refers to a decision aid used in a methodological manner for specific purposes in decision making or planning activities. A tool can be either quantitative or qualitative and can be manual or computerized. It can be based on OR/MS methods or methods from another discipline. A tool can also be based on one or several methods.” [21, p. 931].

Supporting the strategy process is one of three streams of endeavor that make up the OR and Strategy field [22], the other two streams being identified as strategic OR which Bell [23] defines as “Operations Research which achieves a sustainable competitive advantage” and public sector

policy analysis, attributed to Rosenhead [24], which considers issues and decisions such as public health planning [25], energy supply and planning [26,27], policy analysis for the prison service [28], privatization of a public enterprise [29], understanding social reforms [30], and modeling the impact of financial decisions for a government department [31].

A number of studies have been undertaken into the use of “tools” to support strategy [21,32–37]. These studies, reviewed in the next section, cover two sources of evidence: literature reviews covering articles describing the use of specific tools to support strategy and surveys of the practice of tool use.

Tools for Supporting Strategy: A Literature Review

Few articles exist that review the literature concerning tools to support strategy. Clark [32] provides a picture of the literature up to 1990 where his review explored the use of OR/MS tools to support strategic planning. He identified some 24 tools from the literature between 1980 and 1990 and mapped their use onto the relevant stage in the strategy process, which he organized into four stages: analyzing the environment, planning direction, planning strategy, and implementing strategy. He also noted whether the publication concerned application of a tool or theory development. Of the 766 papers classified, 75 were concerned with analyzing the environment, where forecasting was the most frequently used tool. Only three papers were concerned with planning direction, including the setting of mission and vision. Planning strategy was covered in 133 papers, 74 of which were concerned with evaluating strategy and where forecasting and simulation were the most popular tools. Implementation accounted for some 76 papers where forecasting, project management, and spreadsheets were the most popular tools. What is clear from Clark’s review is that the traditional or harder OR/MS approaches, such as simulation and forecasting, play a prominent role in the literature, whereas there were few articles reporting the use of “softer” approaches in supporting the four stages, and in particular the planning direction stage.

From 1990 onward, the volume of articles describing the use of particular tools to support strategy has increased; thus, if the literature review were repeated today, a rather different picture might emerge. To demonstrate this, Table 1 identifies a number of tools drawn from the OR/MS and other fields that have been documented as being used to support activities within the strategy process. In particular, Table 1 identifies each tool's source (OR/MS or strategy/management) and maps the tool onto the activities it can support. For the reader who may wish to further understand how a tool is used to support an activity, the table also identifies a selection of references that specifically describe the application of a tool in relation to one or more of the activities highlighted.

It is noticeable in Table 1 that the mapping between tools and activities in the process is not one-to-one; tools may be used to support multiple activities. For example, visioning approaches do not necessarily stop once the vision and direction have been defined; they may incorporate an assessment or exploration of the external and internal environments, or they may include the generation of initiatives that are needed to bring about the vision. Similarly, while scenario planning is primarily intended as a tool to explore possible future uncertainties within the external environment, it can also be used to help generate strategic initiatives and to test their robustness against plausible possible futures.

Tools for Supporting Strategy: Surveys of Practice

The literature surveying the use of tools to support strategy has investigated two practitioner groups—OR/MS practitioners and executives/MBA alumni. Table 2 summarizes the common tools that were found to be used across six surveys, four of which were conducted with executives or MBA alumni [21,34,35,37] and two of which were conducted with OR/MS practitioners [33,36]. For each of the four executive/MBA surveys, a rank was calculated indicating the extent of a tool's use amongst the respondents, with a rank of 1 indicating the tool used by the greatest proportion of respondents. However,

such a calculation was not possible for the two OR/MS surveys. Clark and Scott [33] identified the top tools for specific activities with no differentiation between their popularity. O'Brien [36] reported the regularity of tool use, thus the ranking of 1 associated with this survey indicates the tools reported as most regularly used to support strategic activities.

A few points are worth noting in Table 2. First there is no general agreement upon a definitive collection of tools that can be used to support strategy; each survey covers a different number and collection of tools, although there is some overlap as can be seen in the table; to facilitate this analysis, the table has been ordered according to the number of surveys a tool appears in, hence the first tool listed (balanced scorecard) appeared in five of the six surveys. One implication of the differences between toolsets is that there are a large number of tools included in the surveys that do not appear in Table 2, the reason being that they each were listed in only one survey; noticeable amongst this set is the top-ranked tool from Rigby and Bilodeau's survey [34], strategic planning, which is defined as "a comprehensive process for determining what a business should become and how it can best achieve that goal" [76]. Another feature of the differences between toolsets can be seen in the naming of tools. Take, for example, "resource-based view" and "resource capability analysis"; it is not clear whether these two tools are in fact the same or different. Second, with the exception of SWOT analysis, the rankings typically vary across the surveys. SWOT analysis does, however, stand out compared to the other tools since it was found to be the tool used by most respondents in three of the executive surveys and it was also reported as being used by OR/MS practitioners. Finally, the overall list of tools presented in the table spans both the OR/MS and the strategy fields, as can be seen by the column labeled "source," where "O" indicates a tool from the OR/MS field and "S" indicates a tool from the strategy or management field. A closer inspection of Table 2 reveals that the executive surveys largely cover strategy or management tools; only the surveys by

Table 1. Tools for Supporting Activities within the Strategy Process

Supporting Tool	Source of Tool (OR/MS, Strategy)	Direction		Creation			Rehearsal and Choice Assessing/ rehearsing initiatives /selecting options	Implementation		
		Setting direction	Setting strategic goals/objectives/priorities	Assessing external environment	Assessing internal environment	Generating ideas for strategic initiatives		Measuring/evaluating current performance	Optimizing operational efficiency	Implementing strategic change
Analytic hierarchy process [27,38]	OR/MS	—	—	—	—	—	X	—	—	—
Balanced scorecard [39,40]	Strategy	—	—	—	—	—	X	X	—	—
Capital investment Appraisal [41]	Strategy	—	—	—	—	—	X	—	—	—
Cognitive mapping [28]	OR/MS	X	X	X	X	X	—	—	—	—
Data envelopment analysis [42,43]	OR/MS	—	—	X	—	—	—	X	X	—
Drama theory [44,45]	OR/MS	—	—	X	X	—	—	—	—	—
Porter's Five forces [46]	Strategy	—	—	X	—	—	—	—	—	—
Forecasting [47]	OR/MS	—	—	X	—	—	—	—	—	—
Mapping competencies [48,49]	Strategy	—	—	—	X	—	—	—	—	—
Multiple criteria decision analysis [50–52]	OR/MS	—	X	—	—	—	X	—	—	—
Optimization approaches [53]	OR/MS	—	—	—	—	—	X	—	X	X
Political economic social technological (PEST) analysis [54]	Strategy	—	—	X	—	—	—	—	—	—

(continued overleaf)

Table 1. (Continued)

Supporting Tool	Source of Tool (OR/MS, Strategy)	Direction		Creation			Rehearsal and Choice Assessing/ rehearsing initiatives /selecting options	Implementation		
		Setting direction	Setting strategic goals/objectives/priorities	Assessing external environment	Assessing internal environment	Generating ideas for strategic initiatives		Measuring/evaluating current performance	Optimizing operational efficiency	Implementing strategic change
8	Project management [55]	OR/MS	—	—	—	—	—	—	—	X
	Real options [56,57]	Strategy	—	—	—	—	—	—	—	—
	Resource-based view [49,58,59]	Strategy	—	—	—	X	X	—	—	—
	Scenario planning [60–62]	Strategy	—	—	X	—	X	—	—	—
	Simulation [15,53]	OR/MS	—	—	—	—	X	—	X	X
	Soft systems methodology [63,64]	OR/MS	X	X	—	—	—	—	—	—
	Stakeholder analysis [65]	Strategy	—	—	X	X	—	—	—	—
	Strategy mapping [66]	Strategy	X	—	—	—	—	X	—	—
	Strengths weaknesses opportunities threats (SWOT/TOWS) analysis [67,68]	OR/MS	—	—	X	X	—	—	—	—
	System dynamics [58,69–72]	OR/MS	—	—	X	X	—	X	—	—
	Visioning approaches [73–75]	Strategy	X	X	X	X	—	—	—	—

Table 2. Comparing Surveys of Tool Use in Supporting Strategy, Showing Ranking of Tools

Survey Author(s)		Tapinos	Rigby and Bilodeau	Stenfors <i>et al.</i>	Gunn and Williams	Clark and Scott	O'Brien
Year of publication		2005 [35]	2007 [34]	2007 [21]	2007 [37]	1996 [33]	2010 [36]
Respondents to survey		MBA alumni	Executives	Executives	Executives	OR/MS	OR/MS
Location of respondents		Global	Global	Finland	UK	UK	UK
Tool	Source						
Balanced scorecard	S	8	11	3	7	—	9
Scenario planning	S/O	7	9	7	8	—	9
Benchmarking	S	2	4	—	2	—	3
SWOT analysis	S	1	—	1	1	—	9
Porter's five forces	S	12	—	—	14	X	24
Risk analysis	O	5	—	4	—	X	3
Value chain analysis	S	14	—	15	12	—	24
Core competencies	S	—	5	—	6	—	9
Mission and vision statements	S	14	5	—	—	—	24
Optimization	O	—	—	12	—	X	24
Brainstorming	S	—	—	9	—	X	3
Cost benefit analysis	S	3	—	—	—	X	3
Financial analysis	S/O	—	—	5	—	X	1
Project management tools	S	—	—	13	—	X	1
Simulation	O	—	—	13	—	X	9
Soft systems methodology	O	20	—	—	—	X	24
Statistical analysis	O	—	—	10	—	X	3
BCG and portfolio matrices	S	17	—	—	—	—	24
Customer relationship Management	S	—	2	—	—	—	9
Knowledge management	S	—	9	—	—	—	18
PEST analysis	S	10	—	—	—	—	18
Cognitive mapping	O	19	—	—	—	—	24
Decision-tree analysis/Decision analysis	O	18	—	—	—	—	9
Enterprise resource planning	S	—	—	17	—	—	24
Forecasting	O	—	—	—	—	X	3
Life cycle analysis	S	—	—	11	9	—	—
Quality methods	S/O	—	—	6	—	—	18
Resource-based view	S/O	14	—	—	—	—	9
Resource capability analysis	S	—	—	15	13	—	—
Spreadsheet applications	O	—	—	2	—	X	—
Stakeholder analysis	O	—	—	—	5	—	18
Tools in this table		15	7	16	10	11	28
Tools in survey		22	25	19	15	18	52

Note: Tools are ranked according to the proportion of respondents reporting their use, with a rank of 1 indicating the tool most reported as used to support strategy. The Clark and Scott survey notes which were the top tools rather than ranking tools in any specific order. O'Brien's survey reports the regularity of use of a tool, hence the ranking applied to this survey reports the most regularly used tool with a rank of 1 being the tool reported as most regularly used to support strategy. Source of tool: S, strategy; O, OR/MS.

Tapinos [35] and Stenfors *et al.* [21] include a few of the OR/MS tools, notably statistical analysis, simulation, and optimization in the case of the former and cognitive mapping and soft systems methodology in the case of the latter. Similarly, the OR/MS survey by Clark and Scott [33] covers mostly OR/MS tools, while O'Brien's survey covers a spread of OR/MS and strategy tools. These two surveys shall now be considered in greater detail to illustrate the use of OR/MS and strategy tools by OR/MS practitioners.

Clark and Scott [33] conducted an empirical study into the strategic level OR/MS tool usage in the United Kingdom by practitioner members of the UK OR Society. Their survey related tool use to specific stages in a three-phase strategy process: phase one of the process involved a situation assessment (i.e., where are we now?); phase two concerned a strategic analysis (i.e., where are we going?), and phase three related to strategic implementation (i.e., how do we get there?).

The tools used to support phase one included simulation, spreadsheets, LP, forecasting, and financial models, all used to appraise the current internal environment. In addition, for setting objectives and mission brainstorming, Porter's five forces and soft systems methodology were used. For exploring the external environment, forecasting, spreadsheets, risk analysis, financial models, and statistical analysis were reported as being used. To support phase two, which involved generating, evaluating, and choosing between alternative strategic options, spreadsheets, simulation, LP, and cost-benefit analysis were used. Finally, to support phase three, activities such as developing implementation plans and monitoring and reviewing performance, project management, simulation, and statistical analysis were used. Clark and Scott concluded that OR/MS practitioners were involved with most of the core strategic tasks within an organization. It is noticeable that in the tools reported as being used, some appearance of the softer approaches was noted (e.g., soft system methodology); In addition, it is noticeable that OR/MS practitioners also reported the use of management

and strategy tools (e.g., Porter's five forces and brainstorming), particularly to support phase one of the process. In a viewpoint relating to Clark and Scott's paper, Pidd [77] queried the lack of soft OR in responses; the authors, in their response [78], suggested that the approaches were at that time still not widely known or used.

O'Brien's [36] survey into the practice of supporting strategy within the United Kingdom OR practitioner community follows Clark and Scott's [33] previous research in that it invited participants to identify the tools that were used to support strategy within their organization, and to relate these tools to specific activities within the strategy process described earlier in this article. The work also identified which tools were combined to support strategy. The survey found that a variety of tools was used to support strategy, including tools from both within and beyond the OR/MS field. Amongst the tools cited as most used to support strategy were financial analysis, project management, benchmarking, brainstorming, cost-benefit analysis, forecasting, risk analysis, and statistical analysis. It was also noticeable that strategy tools (e.g., SWOT analysis, Porter's five forces, and PEST analysis) were reported as being used along with OR/MS tools. In addition, the awareness of softer OR/MS or problem structuring methods was greater than their reported use might suggest. Tools were also reported as being used in combination to support strategy development; popular combinations included scenario planning with simulation and SWOT with PEST analysis. The activity that most tools were associated with was that of strategy evaluation, followed by performance measurement and assessment of the external environment, whereas the activities that tools were least associated with were setting direction, generating options, and implementing strategy. When tools were mapped onto specific activities, it was clear that the relationship between tools and activities was not a one-to-one mapping; the same tool was reported as being used to support more than one activity; for example, simulation was used to support evaluation and selection of options but some also reported its use for

assessing the environment and setting goals. Similarly, brainstorming was primarily used to generate options, set direction and goals, but it was also used to assess the internal and external environments.

In reflecting about the characteristics of the emerging toolset for supporting strategy, one might expect that practitioners would restrict their toolset to those associated with their background, discipline, or training, but this does not appear to be the case—OR/MS practitioners, while drawing on tools from their own field, also draw on tools from the strategy and management fields, just as executives and MBA alumni draw on the strategy tools such as those commonly taught in MBA programs, but also utilize tools from the OR/MS field. Thus, Dyson's view [6] that tools from a wider field than OR/MS can be used to support strategy seems to be supported in practice. Further research is needed into how tools are used to support strategy; surveys simply tell us that a tool is used to support an activity, but not how, when, or with whom? In particular, a topic for research is how tools are used by the OR/MS practitioner in relation to the decision makers, for the assumption is that the OR/MS practitioner plays a support role rather than being the decision maker *per se*.

Combining Tools to Support Strategy

As demonstrated above, tools are not used in isolation but can be combined in supporting strategy. Within the OR/MS field, there have been recent developments in the use of multimethodological approaches where tools are combined in support of interventions [79]. There has been little systematic research in this area with respect to strategy development; however, some examples exist to illustrate the potential. The extant literature may be broadly divided into two groups. The first group includes tools that have been combined in order to provide complementary insights to the issues under consideration by highlighting different perspectives. The second group consists of tools that are combined either to create a new tool or whose purpose is to extend or supplement an existing tool into new areas. An example of the former is Brady's [80] use of Porter's five forces with

soft systems methodology to provide complementary insights into the issues surrounding the future of an Irish road gritting facility. An example of the second group is the extension of scenario planning through the addition of multiple criteria decision analysis to evaluate strategic options that have been generated through the development and analysis of alternative scenarios [50,51,81]. Other examples that belong in this second group include Ormerod's articles that report the combination of a number of typically soft OR/MS tools to support information systems strategy development across a range of organizations [82–84].

O'Brien's survey [36] suggests that tools from both the OR/MS and strategy toolsets are combined to support strategy. For example, the most reported combination of two tools in the survey was scenario planning with simulation, followed by SWOT analysis with PEST analysis. As with the use of individual tools, further research is needed to explore how and why practitioners are combining tools to support strategy. Further research might also usefully explore how tools from across the two fields can be combined in supporting strategy.

CONCLUSIONS AND FURTHER RESEARCH

This review began by explaining the concepts of strategy, strategic decisions, and issues. It then introduced the idea of a strategy process consisting of a number of key interrelated activities, which are needed for the effective strategic development of an organization. In reflecting on the rational nature of the OR/MS discipline, two forms of rationality were identified: substantive rationality that has characterized much of the traditional "hard" OR/MS approaches, and procedural rationality that characterizes many of the newer "soft" OR/MS approaches. In supporting the strategy process, it was noted that both substantively and procedurally rational approaches can offer support to different activities within the process. The review then summarized the literature and empirical research highlighting which tools have been used to support the activities of

a strategy process, either individually or in combination.

A key conclusion of this review is that the strategy process can be supported by a wide variety of tools, where the mapping between tools and activities in the process is not one-to-one; different tools can be used to support more than one activity and individual activities can be supported by more than one tool. Thus, in supporting the strategy process it can be beneficial to use more than one tool, since different tools typically offer varied and often complementary perspectives. A related conclusion is that tools for supporting the strategy process can be drawn from both the OR/MS and strategy fields. This field offers many opportunities for further research, some of which are identified below.

Firstly reviews and overviews of the literature documenting the variety of tools to support different activities within the strategy process are lacking both within the OR/MS and strategy fields. In particular, Clark's [32] literature review of the early 1990s could be usefully updated to provide a current picture taking account of more recent developments. For example, within the OR/MS field it would be useful to explore the literature reporting the use of specific soft OR/MS approaches or PSMs to support strategy development.

Second, this review has summarized empirical research, which has identified the tools used to support strategy by both OR/MS practitioners and executives/MBA alumni. While there is a reasonable knowledge of the tools that are used within the United Kingdom and Europe amongst the OR/MS community, little is known about the use of tools to support strategy by OR/MS practitioners within the United States and other regions of the world.

Third, the existing research has largely identified which tools are used to support strategy. However, knowledge that a tool is used does not tell the whole story; it provides a starting point for further research to explore both how and why the tools are used and combined to support strategy within organizations. Issues that are relevant to this research include, for example, the extent to which tools are used as theory promotes,

or modified to suit the particular circumstances within an organization. As Gunn and Williams [37] suggest, questions such as how and why tools are used are more difficult to research than identifying which tools are used and thus they require different research approaches. Such research could usefully be conducted adopting a strategy as practice lens [85], taking the view that strategy is an activity, something that people *do*. Such a perspective would enable the microlevel exploration of OR/MS practitioners as strategic actors engaged in strategy work. In particular, it would facilitate research into the practice of using tools, for example, how scenario planning is used within an organization, along with the detail of what actually happens during its use and which different actors are involved, say in the workshops where it is used.

Finally, there is scope to explore both theoretical and practical developments in how tools from across the two fields of OR/MS and strategy can be brought together to support strategy.

These conclusions have important implications for academics and practitioners alike. For practitioners, it suggests the need to look beyond their traditional training for tools that are of practical use in supporting strategy. For academics, it offers numerous opportunities for exploring the use and development of tools, possibly in combinations that bridge the different fields.

REFERENCES

1. De Wit B, Meyer R. *Strategy: process, content, context*. 3rd ed. London: Thomson; 2004.
2. Whittington R. Completing the practice turn in strategy. *Organ Stud* 2006;27(5):613–634.
3. Johnson G, Scholes K, Whittington R. *Exploring corporate strategy*. 7th ed. London: Prentice Hall; 2006.
4. Coyne KP, Subramaniam S. Bringing discipline to strategy. *McKinsey Q* 1996;(4):14–25.
5. Pidd M. Contemporary OR/MS in strategy development and policy-making: some reflections. *J Oper Res Soc* 2004;55:791–800.
6. Dyson RG, editor. *Strategic planning: models and analytical techniques*. Chichester: Wiley; 1990.

7. Kirkwood CW. Does operations research address strategy? *Oper Res* 1990;38(5): 747–751.
8. Hart S, Banbury C. How strategy-making processes can make a difference. *Strat Manag J* (1986–1998) 1994;15(4):251.
9. Grant RM. Contemporary strategy analysis. 5th ed. Oxford: Blackwell; 2006.
10. Dyson RG, Foster MJ. Effectiveness in strategic planning. *Eur J Oper Res* 1980; 5(3):163.
11. Dyson RG, Foster MJ. Effectiveness in strategic planning revisited. *Eur J Oper Res* 1983;12(2):146.
12. O'Brien FA, Dyson RG, editors. Supporting strategy: frameworks, methods and models. Chichester: John Wiley & Sons, Ltd.; 2007.
13. Mintzberg H, Quinn JB. The strategy process: concepts, contexts and cases. 2nd ed. New Jersey: Prentice Hall; 1991.
14. INFORMS. What is OR. The Science of Better. 2009. Available at http://www.theorsociety.com/Science_of_Better/htdocs/prospect/what/index.htm. Accessed 2009 July 24.
15. Freeman JM. Planning police staffing levels. *J Oper Res Soc* 1992;43(3):187–194.
16. Vasko F, Stott K, Jr. Strategic planning: OR to the rescue. *OR Insight* 2008;21(3): 26–32.
17. Eden C. Strategy development as a social process. *J Manag Stud* 1992;29(6):799–811.
18. Rosenhead J, Mingers J. Rational analysis for a problematic world revisited. Chichester: Wiley; 2001.
19. Rosenhead J. Past, present and future of problem structuring methods. *J Oper Res Soc* 2006;57(7):759.
20. Ormerod R. The OR/MS contribution to strategy development and policy-making. *J Oper Res Soc* 2006;57(1):117.
21. Stenfors S, Tanner L, Syrjänen M, *et al.* Executive views concerning decision support tools. *Eur J Oper Res* 2007;181:929–938.
22. Dyson RG. Strategy, performance and operational research. *J Oper Res Soc* 2000; 51(1):5–11.
23. Bell PC. Strategic operational research. *J Oper Res Soc* 1998;49(4):381–391.
24. Rosenhead J. Into the swamp: the analysis of social issues. *J Oper Res Soc* 1992; 43(4):293–305.
25. van Gennip CEG, Hulshof JAM, Lootsma FA. A multi-criteria evaluation of diseases in a study for public-health planning. *Eur J Oper Res* 1997;99(2):236–240.
26. Dijk D, Kok M. A comparison of different modelling approaches to strategic energy planning. *Eur J Oper Res* 1987;29(1): 42–50.
27. Hamalainen RP. A decision aid in the public debate on nuclear power. *Eur J Oper Res* 1990;48(1):66–76.
28. Eden C, Ackermann F. Cognitive mapping expert views for policy analysis in the public sector. *Eur J Oper Res* 2004;152(3): 615–630.
29. Sueyoshi T. Privatization of nippon telegraph and telephone: was it a good policy decision? *Eur J Oper Res* 1998;107(1):45–61.
30. Tsoukas H, Papoulias DB. Understanding social reforms: a conceptual analysis. *J Oper Res Soc* 1996;47(7):853–863.
31. Calvert D, Kaufman R. Using analysis to influence strategic-decision making: option valuation at ECGD. *J Oper Res Soc* 2008;59:198–202.
32. Clark DN. A literature analysis of the use of management science tools in strategic planning. *J Oper Res Soc* 1992;43(9): 859–870.
33. Clark DN, Scott JL. Strategic level MS/OR tool usage in the United Kingdom: an empirical survey. *J Oper Res Soc* 1995;46(9):1041–1051.
34. Rigby D, Bilodeau B. Bain's global 2007 management tools and trends survey. *Strat Leader* 2007;35:9–16.
35. Tapinos E. Strategic development process: Investigating the relationship between organisational direction and performance measurement. Warwick business school. Coventry: University of Warwick; 2005. pp. 273.
36. O'Brien F. Supporting strategy: a survey of UK OR/MS practitioners. *J Oper Res Soc* 2010; In press. pre-print version available from <http://wrap.warwick.ac.uk>.
37. Gunn R, Williams W. Strategic tools: an empirical investigation into strategy in practice in the UK. *Strat Change* 2007;16(5):201–216.
38. Arbel A, Orgler YE. An application of the AHP to bank strategic planning: the mergers and acquisitions process. *Eur J Oper Res* 1990;48(1):27–37.
39. Kaplan RS, Norton DP. The balanced scorecard - measures that drive performance. *Harv Bus Rev* 1992;70(1):71–79.

40. Wisniewski M, Dickson A. Measuring performance in Dumfries and Galloway Constabulary with the balanced scorecard. *J Oper Res Soc* 2001;52(10):1057–1066.
41. Berry RH. A financial perspective on strategic investments. In: O' Brien F, Dyson RG, editors. *Supporting strategy: frameworks, methods and models*. Chichester: Wiley; 2007.
42. Day DL, Lewin AY, Li H. Strategic leaders or strategic groups: a longitudinal data envelopment analysis of the US brewing industry. *Eur J Oper Res* 1995;80(3): 619–638.
43. Medina-Borja A, Pasupathy K, Triantis K. Large-scale data envelopment analysis (DEA) implementation: a strategic performance management approach. *J Oper Res Soc* 2007;58(8):1084–1098.
44. Bryant JW. Confrontations in health service management: Insights from drama theory. *Eur J Oper Res* 2002;142(3):610–624.
45. Bryant JW, Darwin JA. Exploring inter-organisational relationships in the health service: An immersive drama approach. *Eur J Oper Res* 2004;152(3):655–666.
46. Porter ME. The five competitive forces that shape strategy. *Harv Bus Rev* 2008; 86(1):78–96.
47. Bunn DW, Salo AA. Forecasting with scenarios. *Eur J Oper Res* 1993;68(3):291–303.
48. Eden C, Ackermann F. Mapping distinctive competencies: a systemic approach. *J Oper Res Soc* 2000;51(1):12–20.
49. Bryson JM, Ackermann F, Eden C. Putting the resource-based view of strategy and distinctive competencies to work in public organizations. *Public Adm Rev* 2007;67(4): 702–717.
50. Montibeller G, Gummer H, Tumidei D. Combining scenario planning and multi-criteria decision analysis in practice. *J Multicriteria Decis Anal* 2006;14:5–20.
51. Wright G, Goodwin P. Future-focussed thinking: combining scenario planning with decision analysis. *J Multicriteria Decis Anal* 1999;8(6):311–321.
52. Montibeller G, Belton V. Causal maps and the evaluation of decision options—a review. *J Oper Res Soc* 2006;57(7):779–791.
53. Butler TW, Karwan KR, Sweigart JR. Multi-level strategic evaluation of hospital plans and decisions. *J Oper Res Soc* 1992;43(7):665–675.
54. Burt G, Wright G, Bradfield R, *et al.* The role of scenario planning in exploring the environment in view of the limitations of PEST and its derivatives. *Int Stud Manag Organ* 2006;36(3):50–76.
55. Asrilhant B, Meadows M, Dyson RG. Exploring decision support and strategic project management in the oil and gas sector. *Eur Manag J* 2004;22(1):63–73.
56. Angelou G, Economides A. A real options approach for prioritizing ICT business alternatives: a case study from broadband technology business field. *J Oper Res Soc* 2008;59(10):1340–1351.
57. Timothy AL. Strategy as a portfolio of real options. *Harv Bus Rev* 1998;76(5):89–99.
58. Warren K. The dynamics of strategy. *Bus Strat Rev* 1999;10(3):1–16.
59. Kunc M, Morecroft J. Resource-based strategies and problem structuring: using resource maps to manage resource systems. *J Oper Res Soc* 2009;60(2):191–199.
60. Van der Heijden K. *Scenarios—the art of strategic conversation*. Chichester: Wiley; 1996.
61. van der Heijden K, Bradfield R, Burt G, *et al.* *The sixth sense: accelerating organisational learning with scenarios*. Chichester: Wiley; 2002.
62. Schoemaker PJH. Scenario planning: a tool for strategic thinking. *Sloan Manag Rev* 1995;36(2):25–40.
63. Ormerod R. Putting soft OR methods to work: Information systems strategy development at Richards Bay. *J Oper Res Soc* 1996;47(9):1083–1097.
64. Checkland P. Soft systems methodology. In: Rosenhead J, Mingers J, editors. *Rational analysis for a problematic world revisited*. Chichester: Wiley; 2001.
65. Bryson J, Cunningham G, Lokkesmoe K. What to do when stakeholders matter: the case of problem formulation for the African American men project of Hennepin County, Minnesota. *Public Adm Rev* 2002;62(5): 568–584.
66. Kaplan RS, Norton DP. Having trouble with your strategy? Then map it. *Harv Bus Rev* 2000;78(5):167–176.
67. Weihrich H. Daimler-Benz's move towards the next century with the TOWS Matrix. *Eur Bus Rev* 1993;93(1):4–11.
68. Dyson RG. Strategic development and SWOT analysis at the University of Warwick. *Eur J*

- Oper Res 2004;152(3):631–640.
69. Warren K. Improving strategic management with the fundamental principles of system dynamics. *Syst Dyn Rev* 2005;21(4): 329–350.
 70. Warren K. Why has feedback systems thinking struggled to influence strategy and policy formulation? Suggestive evidence, explanations and solutions. *Syst Res Behav Sci* 2004;21(4):331–347.
 71. Kunc MH, Morecroft JDW. Competitive dynamics and gaming simulation: lessons from a fishing industry simulator. *J Oper Res Soc* 2007;58(9):1146–1155.
 72. Dyer I. Energy modelling platforms for policy and strategy support. *J Oper Res Soc* 2000;51(2):136–144.
 73. Ackoff RL. Idealized design: creative corporate visioning. *Omega* 1993;21(4):401–410.
 74. O'Brien F, Meadows M. Developing a visioning methodology: Visioning Choices for the future of operational research. *J Oper Res Soc* 2007;58(5):557–575.
 75. Vidal RVV. The vision conference: facilitating creative processes. *Syst Pract Action Res* 2004;17(5):385–405.
 76. Rigby D. Management tools 2007: an executive guide, in *www.Bain.com*, B.a.C. Inc., editor. Boston (MA): Bain and Company, Inc.; 2007.
 77. Pidd M. OR/MS and strategic management. *J Oper Res Soc* 1996;47(3):473–474.
 78. Clark DN, Scott JL. Response to “OR/MS and strategic management”. *J Oper Res Soc* 1996;47(3):474–476.
 79. Munro I, Mingers J. The use of multimethodology in practice - results of a survey of practitioners. *J Oper Res Soc* 2002;53(4):369–378.
 80. Brady M. Analysis of a public sector organizational unit using strategic and operational analysis tools. *Knowl Process Manag* 2008;15(2):140–149.
 81. Goodwin P, Wright G. Enhancing strategy evaluation in scenario planning: a role for decision analysis. *J Manage Stud* 2001;38(1):1–16.
 82. Ormerod RJ. Information systems strategy development at Sainsbury's supermarkets using “soft” OR. *Interfaces* 1996;26(1): 102–130.
 83. Ormerod R. Putting soft OR methods to work: information systems strategy development at Palabora. *Omega* 1998;26(1):75–98.
 84. Ormerod R. Putting soft OR methods to work: the case of IS strategy development for the UK Parliament. *J Oper Res Soc* 2005;56(12):1379–1398.
 85. Jarzabkowski P, Balogun J, Seidl D. Strategizing: the challenges of a practice perspective. *Hum Relat* 2007;60(1):5–27.

SURGERY PLANNING AND SCHEDULING

S. AYCA ERDOGAN

BRIAN T. DENTON

Edward P. Fitts Department of Industrial
and Systems Engineering, North
Carolina State University, Raleigh,
North Carolina

In 2007, total health-care spending in the United States reached US \$2.3 trillion, and continues to rise at the fastest rate in US history. The design and implementation of better planning and scheduling systems is an important area to study in order to reduce high costs and improve access to services in the health-care system. Surgery scheduling, in particular, is an area with significant potential for realizing greater efficiencies. Poor scheduling prevents health-care providers from matching patient demand with available capacity, causing inefficient use of resources, decreased return on investment, and long waiting lists for patients. It has been estimated that surgery accounts for more than 40% of a hospital's total revenues and expenses. Recent studies indicate that resource utilization, overtime, and on-time start performance within surgical suites could be improved at most hospitals. These important performance measures are influenced in part by the surgery scheduling systems and policies that are used in practice.

Surgery scheduling systems impact a variety of expensive resources including operating rooms (ORs), the postanesthesia care unit (PACU), intensive care unit (ICU), hospital beds, equipment resources such as mobile diagnostic imaging devices, and human resources including surgeons, nurses, anesthesiologists, and other staff. The unpredictable nature of surgery results in uncertainty in the duration of surgery and patient recovery. This can be caused by many factors including the varying experience of surgeons and OR teams, the presence of residents or surgical fellows in

academic medical centers, or patient characteristics like age, weight, or unobservable physiological factors that only become known at the time of surgery. Unexpected *add-on cases* (urgent surgeries that arise on short notice), patient cancellations, and unanticipated staff or resource unavailability are additional sources of uncertainty that can create problems on the day of surgery.

Surgery scheduling is also difficult because there are a number of competing performance criteria. The efficiency and effectiveness of a surgery schedule may be judged differently depending on the individual's perspective. For instance, OR managers are often concerned with utilization of resources such as ORs, expensive equipment, and staff. Patients, on the other hand, are more concerned with waiting time.

In this article, we provide a general overview of the most important parts of the surgery delivery system that affects surgery scheduling. We focus on the surgical suite itself since it comprises the most expensive resources in the delivery of surgical care. We describe the most significant factors that complicate surgery scheduling, and characterize the types of uncertainty that affect scheduling. Our review focuses on the literature related to surgery scheduling that has appeared in both the operations research and the health-care literature. We particularly emphasize recent studies that consider new and important aspects of surgery scheduling systems. We provide a taxonomy of the literature based on the type of operations research methodology used. We also provide a small example to illustrate how OR methods can be used in surgery scheduling. Finally, we briefly discuss some open challenges and opportunities that exist for future research.

SURGERY SCHEDULING

Surgery may be performed on an *inpatient* or an *outpatient* basis. In the hospital outpatient setting, patients are admitted to the

hospital and leave the hospital on the same day of the surgery. In such cases the patient recovers at home. In the inpatient setting, patients either arrive on the day of surgery or they are assigned to a hospital bed ahead of the day of surgery. Inpatients stay in the hospital to recover for one or more days following surgery. Owing to increasing volume, many outpatient surgeries are now performed in ambulatory service centers (ASCs), which are distinct from hospitals (though often associated with a hospital). ASCs include an intake area, often several ORs, and a recovery area. Unlike hospitals, ASCs are not equipped for emergencies or overnight stays and most of them are open 8–10 h a day.

The overall surgery process involves several activities before, during, and after the actual surgical procedure. These include *preoperative*, *intraoperative*, and *postoperative* stages. Figure 1 illustrates these stages as well as the related activities. Preoperative care starts with the patient's and surgeon's decision to have surgery. It may involve a visit to a preoperative clinic where an examination and follow-up tests may be completed. It also typically involves some preparation prior to the day of surgery. On the day of surgery, the patient arrives at the hospital or ASC at a preassigned time. A number of administrative activities may be required,

and the patient may require a brief physical examination or medical tests (e.g., blood test) before surgery. Intraoperative care comprises the activities performed in the OR. After surgery, the postoperative stage begins and patients are admitted to a recovery area such as a PACU or ICU depending on the nature of the surgery. In the case of inpatient surgery, the patients may return directly to their hospital bed. The complete surgery process ends when a patient has fully recovered and no additional follow-up is needed by the surgeon. This may take weeks, months, or even years. In this review, we focus on activities performed in the surgical suite on the day of surgery. Owing to the high cost of surgical suite resources, this is often a bottleneck in the overall process.

Surgical suites at hospitals must be equipped to handle the regular daily surgery schedule as well as urgent add-on cases. ASCs, on the other hand, are designed to perform nonurgent outpatient surgeries with minimal resources and therefore at a lower cost. Many characteristics of the OR environment are the same whether the surgery is performed in a hospital or in an ASC. A typical surgical suite is composed of several individual ORs sharing resources such as an equipment storage area, sterilization resources, preoperative

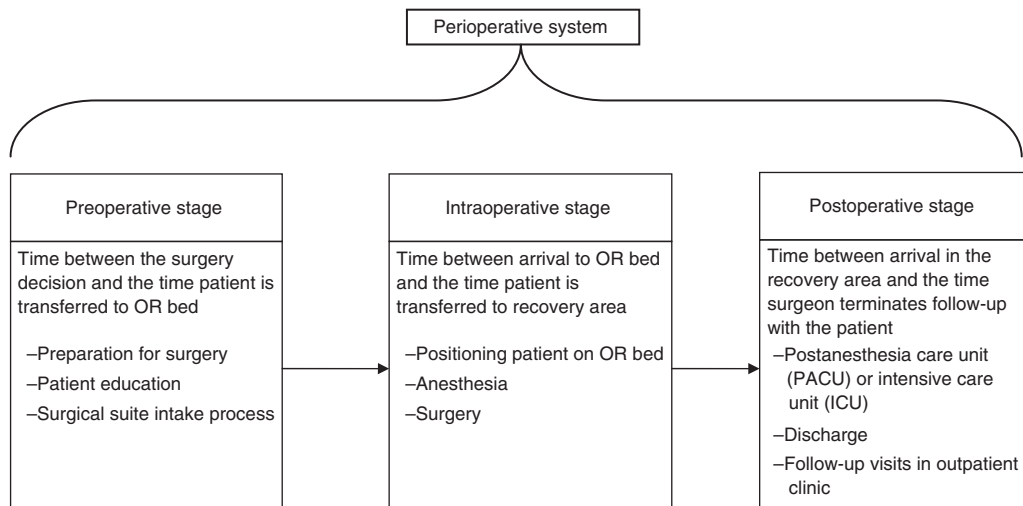


Figure 1. Stages of a perioperative system.

intake, and postoperative recovery rooms. Depending on the type of surgery, some ORs might have different technological resources. For instance, some types of cardiac surgery require cardiopulmonary bypass equipment. In some cases, these special resources may be dedicated to the ORs, while in others they may be mobile and transferred between ORs as needed.

The design and layout of a surgical suite can influence surgery scheduling. Traditional surgical suites have separate intake, surgery, and recovery areas. The patients enter the intake area, progress to surgery, and then eventually transition to a separate recovery area. Figure 2 illustrates the patient flow through this type of a surgical suite. More modern designs have experimented with a common intake and recovery area. The patient is allocated to a room, referred to as a *pre/post room*, where the intake process is carried out. The patient is subsequently taken to surgery, and then returned to their pre/post room after surgery [1].

The process for scheduling surgeries may differ from one hospital or ASC to another. One common system used by many hospitals is known as *block-booking*. In a block-booking system, blocks of uninterrupted OR time are reserved for individual surgeons or surgical specialties in a periodic (often weekly or monthly) schedule. Surgeons book cases into their allotted block time only if the surgery can be completed within the assigned time (or if an administrative exception is granted). The amount of block time is based on the surgeon's or surgical specialty's historical OR usage or some other performance criteria.

Other constraints such as surgeon preferences and resource availability are also taken into account when assigning block times.

Another scheduling approach used by some hospitals is *nonblocked* or *open-booking*. In these systems, surgeons submit requests for OR time. Surgeries are scheduled on a first-come-first-serve basis until either a maximum number of cases is reached or a predetermined OR capacity (planned available time) is reached [2]. Surgeons can submit their cases up until the day before surgery at which point a schedule is generated for the surgical listing.

In both block-booking and open-booking systems, a detailed schedule of surgeries is generated which includes surgery-to-OR assignments and estimated start times. On the day of surgery, the schedule (whether based on block-booking or open-booking) is revised by the OR manager (often an anesthesiologist or the head nurse) throughout the day to accommodate additional add-on cases, and to compensate for cases that run longer or shorter than expected. Both of these systems have advantages and disadvantages. In open-booking systems, surgeries are scheduled on short notice and uncertain demand gives rise to variable OR utilization rates [2]. Block-booking often overcomes this disadvantage by allocating blocks of OR time to surgical specialties based on historical utilization rates. However, it does not allow blocked OR time to be used for other specialties or surgeons, which can result in unused OR time.

Many hospitals use systems that are combinations of open-booking and block-booking systems. For instance, surgeons or surgical

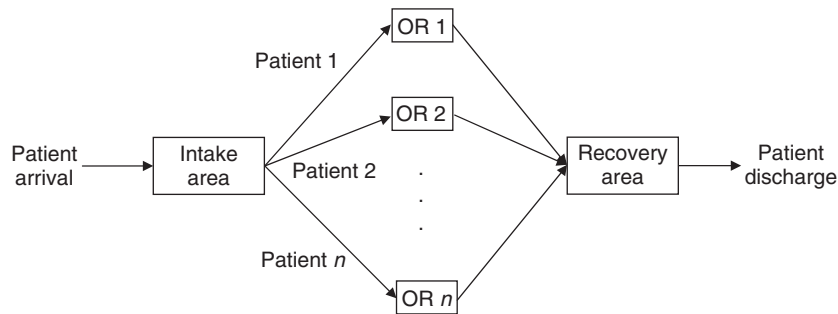


Figure 2. Patient flow through a surgical suite.

specialties may share their assigned block time if it is unutilized within a certain lead time prior to the day of surgery. Furthermore, within the same surgical suite, some ORs may strictly follow a block-booking system, whereas others may follow an open-booking system.

Managing a surgical suite is complicated because it involves many levels of decision making that affect or are affected by surgery scheduling systems [3]. Strategic decisions take place over long time periods (e.g., one year) and include decisions such as whether to use an open-booking or block-booking system, major capacity decisions such as the number of ORs to utilize, investment in new technology, and the number and type of staff to employ. Tactical level decision making involves reserving OR time and capacity for surgical specialties, finding an efficient surgery mix, and determining how much time to plan for each type of surgery. These decisions involve balancing the cost of allocating too much time which causes idle time both for the OR and the staff with the cost of allocating too little time which may result in overtime. Operational decisions involve sequencing surgeries in a particular OR, estimating start times for individual surgeries [4], and setting appointment times for patients to arrive for surgery so as to minimize on-site waiting and achieve high patient satisfaction. In addition, decisions about how to manage add-on cases, and whether to cancel scheduled cases must be made on the day of surgery [3].

COMPLICATING FACTORS IN SURGERY SCHEDULING

Each activity in the perioperative period is critical to the successful delivery of surgical services to patients. The types of activities and their durations may differ from inpatient to outpatient settings and from one hospital to another. One commonality is that most of these activities have significant uncertainty in duration. This is well known for surgical procedures themselves and has also been reported for preoperative and postoperative time in ASCs [1]. Preoperative time, for instance, can depend on

many factors including whether the patient requires some prescreening (e.g., a patient with diabetes might require a blood test prior to surgery) or if the patient requires a translator. The surgery itself can be highly uncertain in duration. Factors affecting surgery include the type of procedure, physical characteristics of the patient such as body-mass-index, age, gender, and the surgical environment (e.g., academic vs nonacademic medical center or hospital vs outpatient procedure center) [5]. Uncertainty in postoperative time can result from the outcome of surgery and the patient's response to anesthesia. Previous research indicates that surgery duration and anesthesia duration often fit a lognormal distribution [4]. Although perfect prediction of surgery durations is impossible, improved estimates can have a positive impact on costs associated with underutilization and overutilization of the surgery suite.

Another complicating factor in surgery scheduling is the uncertainty in the number of patients to be scheduled on a particular day. Surgery at hospitals can be divided into two major categories such as *elective* and *non-elective* cases [6]. For elective cases, surgery may be planned well in advance (e.g., months) to be performed on a future date. For nonelective cases, on the other hand, the surgery is unanticipated. These cases must be worked into the existing schedule, either by using intentionally reserved or otherwise available space in the schedule, or by creating space by canceling elective cases. Some hospitals allocate separate ORs for emergencies and add-ons, whereas others allow slack time in the schedule [7,8].

Other causes of uncertainty in the number of surgeries on a particular day include patient punctuality, no-shows, or cancellations [9]. Outpatient settings are more affected by patient punctuality and no-shows. In inpatient settings, due to the greater complexity and risk of surgery, cancellations are a significant source of uncertainty. For instance, a cancellation may occur due to unexpected changes in the patient's medical status.

Variation and uncertainty in daily surgery mix is another complicating factor due to the varying resource requirements for different

types of surgery. Hospitals often control the mix by allocating OR capacity to specific surgical groups (e.g., thoracic, orthopedic, and colorectal). Such allocation, often called *block-booking*, is done based on historical utilization by surgical groups.

Surgical suites have high fixed costs, the large proportion of which are associated with the labor cost of the OR team. Surgical suites also have variable costs related to scheduling. For instance, ORs typically have a planned utilization time (e.g., 8 h) beyond which overtime costs for some members of the OR team begin to accrue. Therefore, on-time surgery start performance, to the extent it affects overtime, is an important consideration. Efficient surgery scheduling also affects the amount of waiting by the OR team, and other critical resources. The recovery area can also be the bottleneck for the surgical suite at certain times of the day. Therefore, resources such as PACU, ICU, and other hospital beds required after surgery and staff should be considered during the scheduling process [10].

High fixed costs encourage scheduling a sufficient number of cases to fully utilize capacity on the day of surgery. This pressure combined with complex interdependency between activities, and the need to schedule a large number of activities with uncertain duration within a fixed time during the day makes for a very challenging scheduling problem. Not surprisingly, much of the literature has focused on stochastic models (e.g., discrete-event simulation, queueing, stochastic programming, and Markov decision processes). In the next section, we provide a detailed review of the relevant literature on surgery scheduling.

LITERATURE REVIEW ON SURGERY SCHEDULING

Surgery scheduling has been a widely studied topic in the medical and operations research fields. There are a number of previous literature reviews related to surgery scheduling. An early review [11] focuses on classifying the previous research based on various stages of decision making, performance measures such as OR utilization and costs, and

the scheduling process. A more recent review article provides detailed classifications based on surgery durations (deterministic or stochastic), patient arrivals (elective and nonelective), and operations research methodology [6]. Previous reviews of literature on general health care appointment scheduling (e.g., primary care clinics [12]) are also relevant to surgery scheduling. In our review, we emphasize recent studies that consider new and important aspects of surgery scheduling systems. We categorize the literature based on the type of OR methodology used, and use our literature review to identify opportunities for future research which we summarize in the section titled “Open Challenges and Future Research Areas.”

Queueing Models

There have been numerous queueing-based studies presented over the past several decades on the problem of scheduling patients for surgeries and outpatient clinic appointments [13–15]. Queueing research has focused on single server problems for which appointment decisions are economically significant. Although not directly applicable, scheduling problems from other contexts (e.g., manufacturing lines and other service systems) are also relevant to surgery scheduling. Much of this research is relevant to single-OR scheduling with open-booking systems. The aim of the single-OR scheduling problem is to assign start times for a certain number of surgeries to be scheduled in a particular OR on a given day. The scheduler must consider the variability in surgery durations. Clearly, the amount of waiting and idling between surgeries will be impacted by the choice of start times. Also, depending on the planned OR time, some overtime may occur at the end of the day. Figure 3 demonstrates the situations when these conditions occur in an OR. The section titled “Example” provides a specific example in this context.

In the above referenced queueing literature, the authors assume an infinite horizon and stationary service and arrival processes which are not typical in surgery scheduling environments. Some researchers propose more realistic queueing models with a finite horizon. For instance, Brahimi and

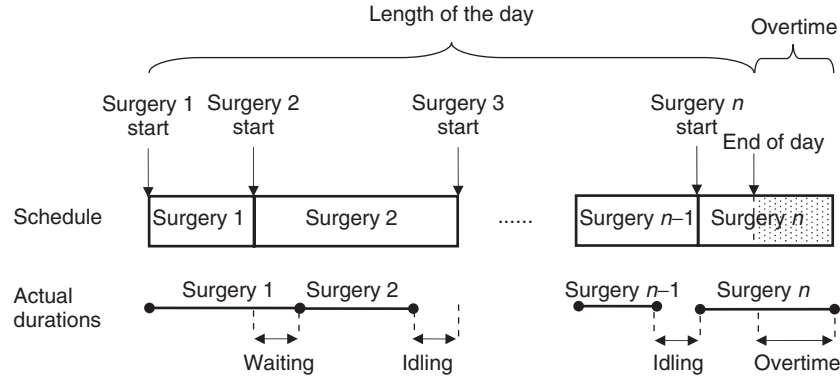


Figure 3. Single OR scheduling problem. The scheduler selects surgery start times. The solid line depicts one possible scenario which includes some waiting, idling, and overtime.

Worthington [16] develop a queueing system with a time-dependent Markovian arrival rate and discrete service time distributions ($M(t)/G/s$ queue) where the number of customers at any time is assumed to be finite. Wang [17] studies the scheduling problem both as a static scheduling problem in which a finite number of customers are scheduled at once, and as a dynamic scheduling problem in which an additional number of customers are scheduled one at a time after an initial batch of customers have been scheduled. He uses phase-type distributions to investigate the transient solution of a Markovian server with general arrival distribution ($S(n)/M/1$ queue) to determine optimal start times for each customer.

Vanden Bosch and Dietz [18] present a queueing model with phase-type service durations and deterministic arrivals where patient no-shows are also considered. In their study, patients are classified into different groups depending on their service durations. Analyzing the special structure of the model, they propose an algorithm which is guaranteed to find the optimal schedule along with the optimal sequence efficiently.

Although not directly applicable to surgery scheduling, it is worth mentioning the recent study by Zeng *et al.* [19] on overbooking in clinical appointment schedules to reduce the negative impact of no-shows. The authors develop a queueing model with patients having different no-show probabilities. They propose that overbooking is beneficial for open-booking systems, and

they demonstrate that clustering patients according to their no-show probabilities will produce better schedules.

Queueing models offer valuable general insights about the effect of uncertainty on scheduling decisions. They also have several shortcomings in the context of surgery scheduling. Some of the queueing articles referred above assume that the system reaches steady state. However, surgical suites typically operate for some specific period of time during the day (e.g., 8–12 h) after which a small number of ORs are available for emergencies. Thus, the system is terminating, and the steady-state assumption is a significant approximation that has not been carefully validated. Queueing models also often require strong assumptions such as exponentially distributed surgery durations, which is not reasonable for most types of surgery that tend to better fit a lognormal distribution [4].

Simulation

Simulation models have found considerable application to surgery planning and scheduling. Although they do not provide closed form solutions like the queueing models, simulation models are more flexible in terms of assumptions about the probability distributions of surgery or other activity durations. However, they are also more computationally intensive than queueing models.

Ho and Lau [20] used simulation in their studies to compare the performance of

simple scheduling rules in a single server system. They propose that the performance of the scheduling rules are affected by the *environmental conditions* of the operating environment such as the probability of no-shows, number of patients to schedule, and the service distribution. They consider simple rules such as scheduling two, three, or more patients simultaneously at the beginning of the session and scheduling the later patients with intervals equal to the mean of the service distribution. They also modified these rules by changing the interval length between patients. In total, they evaluated more than 50 different rules with several different combinations of the three environmental conditions. They present the Pareto set of scheduling rules with respect to expected idle time of the server and expected patient waiting time.

Many researchers have used simulation methods to analyze more complex multi-OR surgical environments with recovery areas. Schmitz and Kwak [21], for example, used simulation to analyze a multi-OR surgical suite with recovery rooms to determine the number of ORs to open on a day given that the number of surgeries to be performed is known. They also study the impact of an increase in the number of ORs on recovery room usage and determine the need for recovery room capacity given that some of the surgeries are done in the outpatient setting. Lowery [22] simulates patient flow through a hospital's critical care units including ORs, ICU, and PACU beds to determine the bed requirements for these recovery units.

Simulation methods have been widely used for testing the performance of scheduling heuristics. In [1], a discrete event simulation model is developed for a newly designed outpatient surgical suite to study the impact of scheduling and start time heuristics and the daily surgical mix on the expected patient waiting time and overtime. The authors tested the performance of seven combinations of OR allocation heuristics and sequencing rules such as scheduling surgeries according to longest processing time first (LPT), shortest processing time first (SPT), increasing variance (VAR), and increasing covariance (COV), and found

that in general LPT and combination of LPT and VAR rules (LPT-VAR) perform better than other rules. The authors also found that arrival time schedules and daily surgery mix have significant effects on the performance measures while scheduling rules have smaller effects on the performance measures. The authors also compare the performances of the heuristics with a bi-criteria genetic algorithm which finds near optimal solutions.

Examining a surgery planning from a strategic perspective, Everett [23] developed a decision support system based on a discrete event simulation model to analyze the waiting times of surgery patients in a public health-care system with multiple hospitals in Australia. His model includes a single list for all surgical patients categorized by urgency and surgery type and can be used to match availability of the hospitals' and patients' needs. This simulation model considers each hospital as an entity in a large health-care system which may reflect the publicly funded health systems, such as in Canada, United Kingdom, and Australia, where there is more centralized control over the assignment of patients to health providers. The article provides evidence that the health-care policies used in different countries may influence the process as well as the efficiency in surgery scheduling.

Simulation models have generally found wide usage in the area of surgery planning and scheduling. Their flexibility allows them to consider many important factors relevant to complex surgical suites. Furthermore, they are descriptive models, meaning that they can be used to compare alternative policies or rules for planning and scheduling. However, there are also some shortcomings. First, they are computationally intensive. Secondly, generating results can be time-consuming compared to queueing models for which analytic solutions are available. A third important shortcoming of simulation models is that they do not provide optimal solutions.

Optimization

Optimization techniques have been widely applied to surgery scheduling. Many researchers have used deterministic models

for scheduling surgeries. Some have also used stochastic optimization models in order to incorporate uncertainties related to surgery durations and patient arrivals. We begin by describing deterministic models and then move on to stochastic optimization models.

Ozkarahan [2] proposed a goal programming model for scheduling surgeries in a surgical suite where block-booking is used for scheduling. Her model includes multiple objectives such as OR utilization, OR and surgeon preferences, and intensive care capacity. For each of these objectives, a goal is specified and the objective is to minimize the deviations from these goals.

Pham and Klinkert [24] extended the job shop scheduling problem to a *multimode blocking job shop* in order to satisfy resource constraints (modes) specific to the OR environment. They use a mixed integer programming (MIP) model formulation with the objective of minimizing the weighted sum of makespan and the sum of starting times of all procedures. They propose that add-on and emergency surgeries can be scheduled by adding new constraints using *job insertion*. Job insertion inserts new surgeries into the established schedule and bumps some elective surgeries to the following day, if necessary, to perform emergent surgeries. Fei *et al.* [25] have developed an integer programming (IP) model to assign elective surgeries with deterministic durations into multiple ORs. The objective is to minimize the weighted sum of overutilization and underutilization of ORs. They solve instances optimally with a branch and price algorithm.

There are also a number of stochastic optimization models for surgery scheduling. The majority assume a single OR, and determine optimal start times for surgeries (see Fig 3). Weiss [26] provides the optimal start times for a simple two-surgery problem with uncertain surgery durations in a single OR. Balancing surgeon's idle time and waiting time, the author shows that the problem is equivalent to the well-known *news vendor* problem from inventory theory, for which the optimal solution is known in closed form. Weiss also examines the optimal sequence of two cases by examining the expected cost resulting from the sequences and shows that

the sequencing decision is related to the dispersion of the density function of surgery durations. He proposes that scheduling surgeries that have distributions with "fatter" tails first results in higher total expected costs of waiting and idling due to the impact on the following surgery.

Denton and Gupta [27] study a general two-stage stochastic linear programming formulation of the single server appointment scheduling problem. They determine the optimal start times for surgeries. The criteria in their model are expected patient waiting time, expected OR idle time, and expected overtime with respect to a defined length of day. They develop an iterative approximate method based on upper and lower bounds on the optimal solution to the problem with continuously distributed surgery durations. More recently, Begen and Queyranne [28] studied the same problem with random service durations given by a joint discrete probability distribution. They prove that there exists an optimal appointment schedule which is integer and can be found in polynomial time.

Gerchak *et al.* [7] use a stochastic dynamic programming model for a multiperiod surgery scheduling problem. In particular, they use a Markov decision process to investigate the optimality of using *cut-off policies* in an open-booking system. Cut-off policies suggest rejecting future elective cases in order to be able to accept possible emergent surgeries if a certain number of ORs are already occupied by elective cases. The authors show that these cut-off policies are not necessarily optimal; however, they are typically close to optimal. Lamiri *et al.* [29] propose a stochastic model for scheduling elective surgeries over a planning horizon, however, they consider OR capacity for both elective and add-on patients. They use simulation to generate samples for emergency surgeries. The samples are used to estimate deterministic capacity requirements for an MIP model that minimizes total overtime costs and other costs associated with hospitalization, and penalties for waiting time of the patients, violating patient's and surgeon's preferences and deadlines. Their model is solved via branch-and-bound algorithm and

then compared to the solution of a simulation optimization approach. The authors conclude that simulation optimization offers improvements even for small sample sizes.

Denton *et al.* [30] proposed a two-stage stochastic MIP formulation to find optimal allocation of surgery blocks to ORs. They also develop a *robust formulation* which aims to minimize the maximum cost (worst-case outcome) that may result from the set of uncertain surgery durations. The robust formulation is advantageous when limited data is available for surgery durations. It is also much less computationally intensive than their stochastic programming model. Batun *et al.* [31] model a multi-OR scheduling problem as a stochastic MIP to determine the number of ORs to open on a given day, allocation of surgeries to ORs, sequence of surgeries within each OR, and start times for each surgeon. The focus of their study is on parallel scheduling of surgeries in which a surgeon may be allocated two or more ORs on the day of surgery.

Lamiri *et al.* [32] used a column generation approach to find the assignments of the elective surgeries into ORs on a given day in the planning horizon. The authors formulate the problem as a stochastic IP model and reserve a random capacity for emergency surgeries. Column generation is used as a solution technique in which columns represent possible surgery-to-OR assignments in a given planning horizon.

Despite the wide literature on optimization methods in OR scheduling problems, there is still a lack of literature in optimization of multi-OR surgical suite scheduling. The current literature suggests that considering the stochastic aspects of the surgical environment is essential for solving realistic problems. However, optimization methods have disadvantages as the problem size gets larger. Finding the optimal solution becomes computationally harder and sometimes impossible for large-scale realistic scheduling problems. Developing efficient stochastic optimization methods to solve larger and more realistic problems, which consider multiple ORs and other critical resources in the surgical suite, remains as an open challenge for researchers.

Heuristic Methods

As described in the previous section, the stochastic and combinatorial nature of surgery scheduling problems gives rise to computationally challenging optimization models. Many researchers have addressed this problem by proposing fast heuristics. This enables consideration of more complex and more realistic systems such as a multi-OR surgical suite with recovery rooms.

Denton *et al.* [33] consider a two-stage stochastic MIP model of the single OR scheduling problem with the addition of surgery sequencing. Decisions in their model are the optimal start times for surgeries and the optimal sequence of surgeries. A simple interchange heuristic is used to find near-optimal sequences, and start time decisions are computed to optimality for each sequence generated by the interchange heuristic. They also prove for a two-surgery example, conditions under which a stochastic ordering defines the optimal sequence, and use numerical experiments to demonstrate that sequencing in order of increasing variance of surgery duration is typically optimal or near-optimal. Intuitively, this is because delaying highly uncertain surgeries to the end of the day limits their impact on other surgeries in the schedule. Vanden Bosch and Dietz [34] propose a local heuristic method to sequence the customers in a first-come-first-serve appointment system when customers differ by waiting cost, no-show probability, and service time distributions. They also propose a solution method to find the optimal schedule given a fixed sequence.

Beliën and Demeulemeester [35] propose a simulated annealing approach as well as several different MIP heuristics to create a master surgery schedule for a block-booking system with the objective of minimizing the expected shortage of beds. Their MIP heuristics involve solving a number of MIPs, which aim to minimize the maximum expected duration, maximum expected bed occupancy, and maximum weighted sum of mean and variance of the daily bed occupancy respectively. Their simulated annealing approach, on the other hand, starts with a random surgery block and does pairwise exchanges

of blocks to find an improved schedule with respect to bed shortage.

Bin-packing algorithms have been widely used to find the mix of surgical cases for multi-OR surgical environments [36]. In this context, the ORs are available for a fixed time during the day, and represent the bins. The surgeries, which have an estimated duration, are the items to be packed in the bins. There are two types of bin-packing problems that have been studied: *on-line* and *off-line*. In the on-line bin-packing problem, urgent surgeries are scheduled sequentially one at a time. Many heuristics have been proposed for this problem. One well-known heuristic, called *best fit*, schedules surgeries into the OR which has the least amount of remaining time that is sufficient to fit the case. Another heuristic, called *worst fit*, schedules the surgery into the emptiest OR which has enough time to fit the case.

In the off-line version of the bin-packing problem, surgeries are batched and simultaneously assigned to ORs. LPT is the best-known heuristic for off-line bin packing. To apply LPT, surgeries are sorted based on their mean durations and then assigned sequentially to the ORs. Many variants of LPT exist including *best fit descending* in which surgeries are assigned sequentially to the fullest OR after being sorted by LPT rule and *worst fit descending* in which they are assigned to the emptiest OR. Dexter *et al.* [36] compare several of these heuristics for scheduling add-on surgeries into remaining open OR times. They found that off-line algorithms are able to fit more add-on surgeries and in particular *best fit descending with special constraints* is more likely to maximize OR utilization.

Hans *et al.* [37] study the bin-packing heuristics with the *portfolio effect* to consider the uncertainty in surgery durations. Portfolio management is used to reduce the risk (minimize variance) or increase the profitability (maximize expected return) by distributing the investments into various different projects instead of investing in a single project. They propose both constructive and local search heuristics to minimize the total planned slack time between surgeries depending on the variances of different

surgery durations. They found that, as a result of the portfolio effect, surgeries with similar duration variability are often scheduled on the same day for each specialty.

In general, heuristics provide a powerful trade-off between descriptive models, such as simulation and queueing models, and prescriptive (optimization) models for problems that are very computationally challenging. The main shortcoming of heuristics is that they do not necessarily provide an optimal solution and it is often difficult to develop good bounds on their performance.

EXAMPLE

In this section, we provide a small stochastic programming example on surgery scheduling. In this example, we consider a problem similar to the one illustrated in Fig. 3 with five surgeries to be scheduled on a given day. Surgery durations are distributed uniformly between 60 and 120 min. The length of the day is assumed to be 450 min. The optimal start times for the surgeries can be found by solving the following minimization problem:

$$\min_x \left\{ E \left[c^w \sum_{i=1}^5 w_i + c^\ell \ell \right] \right\} \quad (1)$$

where $x_i, i = 1, \dots, 4$, are the time allowances between start times of each of the surgeries, $w_i, i = 1, \dots, 5$, are the waiting times, and ℓ is the overtime. Overtime cost and waiting cost are assumed to be equal, that is, $c^w = c^\ell = 1$ in this example. The model in Equation (1) can be formulated as a two-stage stochastic linear program. The x_i are the first stage decision variables, and w_i and ℓ are second stage decisions which are made after the first stage decisions [27]. It is assumed that the first surgery is scheduled at time zero. Thus the waiting time for this surgery is zero.

The optimal solution to this two-stage stochastic programming problem is $x_1 = 95.11, x_2 = 99.73, x_3 = 99.40, x_4 = 96.01$ which means that after first surgery is scheduled at time zero, second one will be scheduled at time 95.11, third one at time 194.84, fourth one at time 294.24, and the

last one at time 390.25. The total expected waiting time plus overtime cost of this solution is 65.39.

A common practice is to set x_i equal to the mean of the duration distribution estimated based on historical data. Stochastic programming takes this uncertainty into account by sampling surgery duration scenarios to find a near-optimal solution which works well on average across all the possible scenarios. For this particular example, 10,000 different duration scenarios were used to find the optimal solution. If the mean value solution was used for this example, the total expected waiting time plus overtime cost would be 79.93, which is 22% higher than the total cost of the stochastic programming solution. Note that the optimal solution trades off expected waiting time cost and idle time cost, which are competing performance measures. For instance, short allowances tend to result in longer waiting time but shorter idle time. Alternatively, long allowances tend to result in shorter waiting times and longer idle times.

OPEN CHALLENGES AND FUTURE RESEARCH AREAS

In this article, we have so far provided an overview of important aspects related to surgery scheduling and a discussion of some of the most significant complicating factors. Surgery scheduling has been a topic of investigation for several decades. However, there are many open research areas and opportunities still to be explored. We describe some of these below.

Most of the above referenced literature focuses on the design of daily schedules prior to the day of surgery. Little effort has been spent on developing rules for real time scheduling on the day of surgery. This is a common problem that OR managers have to face every day to manage the deviations caused by the surgeries that run longer or shorter than anticipated. In addition to the daily operational level scheduling problem, there are also aspects of long range surgery planning and scheduling that need more attention. For instance, the vast majority

of the current literature focuses on patient waiting time on the day of surgery, when, in fact, the time spent in waiting lists may be more important from the health outcomes perspective.

Despite the rich literature considering the uncertainty in surgery durations, demand uncertainty is still largely unexplored. For instance, uncertainty in patient no-shows, case cancellations, and the addition of add-on cases to schedules on short notice can all have significant impact on surgery schedules. Most research has assumed that a fixed number of elective cases are to be scheduled. Uncertainty in demand from day to day can have a significant impact on surgery schedules resulting in substantial delays or cancellations on the day of surgery. Other causes of uncertainty such as unpredictable internal demand on recovery room resources on the day of surgery (e.g., ICU beds) are also in need of study.

Many of the studies referenced above focus on ORs; however, other parts of the surgical suite can also influence performance. For instance, the PACU can become a bottleneck, leading to patients backing up in ORs, and causing longer than anticipated turnover times between surgeries. This could ultimately result in overtime or surgery cancellations. Uncertainty in other critical activities like recovery time have been considered in the literature and demonstrated to be important considerations in scheduling. Some of the above referenced literature on simulation and deterministic mathematical programming consider this issue [10]; however, few attempts have been made to incorporate these aspects into stochastic optimization models.

Another research area that needs further attention is the design of surgical suites and processes within and around the OR. This has emerged as a way to increase OR utilization by reducing nonoperative time and increasing throughput. For example, Sandberg *et al.* [38] proposed a new OR suite design called *OR of the future (ORF)* for technologically intensive surgeries with an intake room and early recovery room attached to the OR. They are motivated by the goal of performing some of the

preoperative and postoperative processes in parallel to increase throughput and decrease overtime. With this new structure, the goal is to perform anesthesia induction in the induction room while the OR is prepared for surgery. Early recovery is provided by an additional perioperative nurse, while the anesthesia personnel can move to the intake of the next patient. Further study is needed to analyze the impact of new designs, which will likely lead to opportunities to improve the efficiency of surgery scheduling.

In this article, we have discussed the basic concepts related to surgery planning and scheduling as well as the surgical environment and the resources. We have also presented a broad range of operations research and management science methodologies used in surgery scheduling. Though this is a very well-studied subject, there are still significant opportunities to improve.

Acknowledgments

This project was funded in part by grants CMMI-0620573 and CMMI-0844511 (Denton) from the National Science Foundation.

REFERENCES

- Gul S, Denton BT, Fowler J, *et al.* Bi-criteria scheduling of surgical services for an outpatient procedure center. Working Paper.
- Ozkarahan I. Allocation of surgeries to operating rooms by goal programming. *J Med Syst* 2000;24(6):339–378.
- Dexter F, Epstein RH, Traub RD, *et al.* Making management decisions on the day of surgery based on operating room efficiency and patient waiting time. *Anesthesiology* 2004;101(6):1444–1452.
- Strum DP, May JH, Vargas LE. Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models. *Anesthesiology* 2000;92(4):1160–1167.
- Strum DP, Sampson AR, May JH, *et al.* Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* 2000;92:1454–1466.
- Cardoen B, Demeulemeester E, Beliën J. Operating room planning and scheduling: A literature review. FED Research Report KBI 0807 Katholieke Universiteit Leuven; 2008.
- Gerchak Y, Gupta D, Henig M. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manage Sci* 1996;42(3):321–334.
- Klassen KJ, Rohleder TR. Outpatient appointment scheduling with urgent clients in a dynamic, multi period environment. *Int J Serv Ind Manage* 2003;15(2):167–186.
- Basson MD, Butler TW, Verma H. Predicting patient nonappearance for surgery as a scheduling to optimize operating room utilization in a veterans' administration hospital. *Anesthesiology* 2006;104(4):826–834.
- Marcon E, Dexter F. Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Manage Sci* 2006;9: 87–98.
- Blake JT, Carter MW. Surgical process scheduling: a structured review. *J Health Syst* 1997;5(3):17–30.
- Cayirli T, Veral E. Outpatient scheduling in health care: a review of literature. *Prod Oper Manage* 2003;12(4):519–549.
- Welch J. Appointment systems in hospital outpatient departments. *Oper Res Q* 1964;15:224–237.
- Mercer A. A queuing problem in which the arrival times of the customers are scheduled. *J R Stat Soc Ser B* 1960;22(1):108–113.
- Soriano A. Comparison of two scheduling systems. *Oper Res* 1966;14(3):388–397.
- Brahimi M, Worthington DJ. Queuing models for outpatient appointment systems: a case study. *J Oper Res Soc* 1991;42(9):733–746.
- Wang PP. Static and dynamic scheduling of customer arrivals to a single-server system. *Nav Res Logist* 1993;40:345–360.
- Vanden Bosch PM, Dietz DC. Minimizing expected waiting time in a medical appointment system. *IIE Trans* 2000;32:841–848.
- Zeng B, Turkcan A, Lin J, *et al.* Clinical scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Ann Oper Res*. DOI: 10.1007/s10479-009-0569-5 (published 12 June 2009).
- Ho C, Lau H. Minimizing total cost in scheduling outpatient appointments. *Manage Sci* 1992;38(12):1750–1764.
- Schmitz HH, Kwak NK. Monte Carlo simulation of operating-room and recovery-room usage. *Oper Res* 1997;20(6):1171–1180.
- Lowery JC. Simulation of a hospital's surgical suite and critical care area. In: *Proceedings*

- of the 1992 Winter Simulation Conference; Arlington (VA). 1992. pp. 1071–1078.
23. Everett JE. A decision support simulation model for the management of an elective surgery waiting system. *Health Care Manage Sci* 2002;5:89–95.
 24. Pham DN, Klinkert A. Surgical case scheduling as a generalized job shop scheduling problem. *Eur J Oper Res* 2008;185(3): 1011–1025.
 25. Fei H, Chu C, Meskens N, *et al.* Solving surgical cases assignment problem by a branch and price approach. *Int J Prod Econ* 2008;112:96–108.
 26. Weiss EN. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Trans* 1990;22(2): 143–150.
 27. Denton B, Gupta D. A sequential bounding approach for optimal appointment scheduling. *IIE Trans* 2003;35:1003–1016.
 28. Begen MA, Queyranne M. Appointment scheduling with discrete random durations. In: *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*; New York. 2009. pp. 845–854.
 29. Lamiri M, Xie X, Dolgui A, *et al.* A stochastic model for operating room planning with elective and emergency demand for surgery. *Eur J Oper Res* 2008;185(3):1026–1037.
 30. Denton BT, Miller AJ, Balasubramanian HJ, *et al.* Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper Res.* In press.
 31. Batun S, Denton B, Schaefer A, *et al.* Operating room pooling and parallel surgery processing under uncertainty. *INFORMS J Comput* 2010. In press.
 32. Lamiri M, Xie X, Zhang S. Column generation approach to operating theater planning with elective and emergency patients. *IIE Trans* 2008;40:838–852.
 33. Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Manage Sci* 2007;10(1):13–24.
 34. Vanden Bosch PM, Dietz DC. Scheduling and sequencing arrivals to an appointment system. *J Serv Res* 2001;4(1):15–25.
 35. Beliën J, Demeulemeester E. Building cyclic master surgery schedules with leveled resulting bed occupancy. *Eur J Oper Res* 2007; 176(2):1185–1204.
 36. Dexter F, Macario A, Traub RD. Which algorithm for scheduling add-on elective cases maximizes operating room utilization? *Anesthesiology* 1999;91:1491–1500.
 37. Hans EW, Wullink G, Van Houdenhoven M, *et al.* Robust surgery loading. *Eur J Oper Res* 2008;185(3):1038–1050.
 38. Sandberg WS, Daily B, Egan M, *et al.* Deliberate perioperative systems design improves operating room throughput. *Anesthesiology* 2005;103(2):406–418.

**SWISS OPERATIONS RESEARCH
SOCIETY (SCHWEIZERISCHE
VEREINIGUNG FÜR OPERATIONS
RESEARCH/ASSOCIATION SUISSE
POUR LA RECHERCHE
OPERATIONELLE/ASSOCIAZIONE
SVIZZERA DI RICERCA OPERATIVA)**

JÜRIG KOHLAS
Department of Informatics,
University of Fribourg,
Fribourg, Switzerland

DOMINIQUE DE WERRA
Ecole Polytechnique Fédérale de
Lausanne, EPFL,
Switzerland

LUCA MARIA GAMBARDELLA
IDSIA, Istituto Dalle Molle di
Studi sull'Intelligenza
Artificiale, USI-SUPSI
Lugano, Switzerland

HISTORY

The preconditions for founding the Swiss Operations Research Society (SVOR/ASRO) were laid by Prof. Dr. h. c. Daenzer, Director of the Institute for Management at the Swiss Federal Institute of Technology (ETH) in Zurich; Prof. Dr. h. c. Käfer, Professor for Management at the University of Zurich; and Prof. Dr. h. c. Linder, Professor for Statistics at the University of Geneva and the Swiss Federal Institute of Technology (ETH) in Zurich. They prepared the ground by a first symposium on OR, held on June 20 and 21, 1958, in Zurich. It was devoted to a first overview of research in operations research (OR), both in universities and within companies. There were already teachers at the universities active in the field: Krelle in St. Gallen, Künzi in Zurich,

Billeter in Fribourg, Ferraud in Geneva, Scheurer and Erard in Neuchatel, and Angehrn in Basel. Several young people were busy to apply OR in industry. Some of them later became professors at universities (Soom in St. Gallen, Weinberg at the Swiss Federal Institute of Technology in Zurich, and Nievergelt in St. Gallen).

Several eminent exponents of OR visited Switzerland during this period and gave talks (A.W. Tucker, L. Arnoff, R.W. Shepperd, A. Adam, R. Roy, and R. Henn, among others, who later became professor in St. Gallen). Furthermore, several talks and workshops took place at different places. Interest in OR was growing, as testified by the increasing number of participants at these events.

The desire to create an organization to coordinate and represent the activities in OR in Switzerland developed and became urgent. The Swiss Society for Statistics and Economy (Schweizerische Gesellschaft für Statistik und Volkswirtschaft, SGSV, founded 1864) had already hosted three study groups (one each for theoretical economy, statistics, and management). Therefore, it seemed appropriate to install a further, new study group devoted to OR. This was judged to be an interesting complement to the existing groups, as OR could be applied in all the three domains. Prof. Künzi, Prof. Daenzer, and Alexis Stravs, a young engineer, contacted the SGSV and were well received. Therefore, a founding committee was constituted, composed of the following members:

Universities

- E. Billeter, Fribourg
- W. Daenzer, ETH Zurich
- R. Henn, St. Gallen
- K. Käfer, Zurich
- Linder, Geneva/ETH Zurich
- J. Niehans, Zurich
- W. Wegmüller, Bern

Industry and Administration

- H. Angst, Swissair, Zurich
- A. Guisolan, Dept. of Defense, Bern
- E. Nievergelt, SBB/CFF/FFS (Swiss Federal Railways), Bern
- E. Soom, Landis and Gyr, Zug
- A. Stravs, ETHZ, Institute for Management, Zurich
- J.J. Weidman, Nestlé, La Tour de Peilz
- F. Weinberg, MFO (Maschinenfabrik Oerlikon, now ABB), Zurich

The general assembly of the Swiss Society for Statistics and Economy, which took place on May 12 and 13, 1961, in La Chaux de Fonds, approved the new study group, which formed on the name “Schweizerische Vereinigung für Operation Research/Association Suisse pour la Recherche Opérationnelle” (Swiss Association for Operations Research). Its first president was Prof. H.P. Künzi, professor at the University of Zurich and later also at the Swiss Federal Institute of Technology in Zurich. The opening symposium of the new association took place on September 15 and 16 in Zurich at the ETH featuring 14 talks mainly from the members of the founding committee. The symposium was followed by the first general assembly of the association. It confirmed H.P. Künzi as the president and elected the founding committee as its first board, with E. Stiefel, ETH Zurich, and Schraffl, Céligny, as additional members. At the time of the first general assembly the association had 261 members, among them 41 collective members (companies). They are considered as the founding members.

The association included in its activities different successful introductory courses into new hot subjects (such as dynamic programming, branch and bound methods, and decision under uncertainty). Many leading OR specialists were invited for talks (among others G.B. Dantzig, P.L. Hammer, and P. Wolfe). To advance the exchange between universities and industry, many visits of companies were organized under the title “Operations Research at Company x”. Workshops in OR were organized, first in Lenk, and later and till now together

with the postdoc program “Troisième cycle en rechercheopérationnelle” of the Universities of the French speaking part of Switzerland.

The association also engaged in international activities, especially with the German association (DGU: Deutsche Gesellschaft für Unternehmensforschung) in the beginning. The first common conference took place in 1971 in Zurich followed by another one 1975 in Interlaken (this time with the DGOR: Deutsche Gesellschaft für Operations Research). Later, similar conferences were organized with the Austrian association (ÖGOR: Österreichische Gesellschaft für Operations Research), namely, 1980 in Vienna, 1984 in Zurich, and 1987 in Graz. By its first president, H.P. Künzi, the Swiss Association for Operations Research was also involved in the well-known “Henn-Schubert-Künzi” workshops in Oberwolfach, Germany.

The journal “Industrielle Organisation,” of the Institute for Management of the ETH in Zurich, published Swiss work in OR for a long time. Of course, very soon Swiss authors started publishing in international journals. H.P. Künzi was together with W. Krelle, the first editor of the Springer series “Monographien zur Unternehmensforschung” (Monographs for Operations Research) and encouraged thereby the German literature in the field. Later, together with M. Beckman, he was again the first editor of the “Springer Lecture Notes in Operations Research and Mathematical Systems”, a series that was devoted to the fast publication of new research results.

In the early 1960s, no chair of operational research as such existed in most universities of western Switzerland; some courses were offered sparsely on an optional basis.

At EPUL (to become EPFL in 1969), research was starting in OR as parts of doctoral theses. In the early 1970s, chairs were created in some institutions, in particular, in EPFL, which played an active role under the leadership of Prof. Charles Blanc who had a crucial influence on the recognition and the subsequent introduction of OR in the engineering curricula.

From that moment, SVOR/ASRO also developed in western Switzerland (French speaking part of Switzerland), attracting

new members from academia as well as industry. Several courses on classical or emerging OR methods were organized in western Switzerland and were followed by many participants from industry and university. Doctoral students also attended these courses, and finally, a permanent organization was set up under the flag of the “Troisième cycle romand”.

The creation of the 3e cycle romand de recherche opérationnelle contributed substantially to reinforce the links between the various institutions in which OR was used and developed. Undoubtedly, the SVOR/ASRO benefited considerably from these efforts.

As a national OR Society, the SVOR/ASRO had to be present also at the European level and similarly in the world. The SVOR/ASRO was present in Brussels when EURO was created in January 1975. Our representative was Alexis Stravs, who with a great dedication attended practically all administrative meetings and never failed to report the latest developments to the current SVOR/ASRO officers. For years, he was the living memory of the Society at the European level. However, he was more than that since he was also the representative of SVOR/ASRO for IFORS, the International Federation of Operations Research Societies, which was founded in 1958. For several years A. Stravs was also a member of the IFORS External Affairs Committee.

Thanks to the efforts of A. Stravs and his successors, the SVOR/ASRO has always maintained very close links with both the European Association EURO and the International Federation IFORS.

For SVOR/ASRO, the fifth EURO Congress organized at EPFL in 1982 with the American Society TIMS was an excellent opportunity to attract more than 600 representatives of the OR community in the world to Switzerland.

In 1997 the International Conference on Mathematical Programming took place at EPFL organized by T. Liebling and D. de Werra it attracted more than a thousand participants.

It may be worth mentioning that IFORS also had between 2010 and 2012 Dominique

de Werra as the president. D. de Werra was the first OR Professor at EPFL in 1971 and the president of SVOR/ASRO from 1984 to 1986.

He was the president of EURO from 1987 to 1988, and in 2004, he was the president of the Gold Medal jury of EURO. D. de Werra received the EURO Gold Medal in 1995, and he was the IFORS Distinguished Lecturer in 2006.

As a consequence of the concrete presence of SVOR/ASRO in the world of OR, it is worth mentioning that among the seven treasurers who cover the existence of EURO, four are members of the SVOR/ASRO (J. Pasquier, J. Bovet, M. Widmer and P. Solot). Such a confidence may undoubtedly be explained by the excellent reputation of Swiss bankers at that time. In 2007, the EURO also presented Hans-Jakob Lüthi, professor of operations research at ETH in Zurich and an SVOR/ASRO committee member, and his group the EURO Excellence in Practice Award for their work on flexibility in energy risk management.

In the past 51 years, from 1961 to 2012, the SVOR/ASRO has had 10 presidents and approximately 62 committee members. Each of them has strongly contributed to the promotion and development of the OR discipline in Switzerland and abroad. The list of SVOR/ASRO presidents is presented here.

1961–1970: Hans Kuenzi, University of Zurich (deceased)

1970–1976: Erich Soom, University of St. Gallen

1976–1980: Jürg Kohlas, University of Fribourg

1980–1984: Peter Kall, University of Zurich

1984–1986: Dominique de Werra, EPFL Lausanne

1986–1996: Heinz Schiltknecht, Sandoz, Bâle

1996–2001: Alain Hertz, EPFL Lausanne

2001–2005: Philippe Solot, Aicos, Basel

2005–2011: Daniel Costa, Nestlé, Vevey

2011–: Luca Maria Gambardella, IDSIA-USI-SUPSI, Lugano

CURRENT ORGANIZATION

The main objectives of SVOR/ASRO are to promote OR in Switzerland, to bridge the gap between the theory and practice of OR, and to represent the interests of the Swiss OR community in all relevant organizations.

The SVOR/ASRO includes members of different types: individual members, collective members, and students. The society is managed by an executive committee composed of the president and other three to five members, including the treasurer, the bulletin editor, the secretary and the person responsible for the program. The committee is renewed every three years by the general assembly.

The site of the general assembly is set each year following two principles. In some cases, the general assembly is organized during the visit of industrial sites, such as logistics operators, production companies, and traffic control organizations, where a potential impact of OR solutions is foreseen. In other years, the general assembly is held during conferences that the association is organizing and sponsoring in Switzerland. In particular, the 50th SVOR/ASRO anniversary assembly was organized in Zurich, during the International Conference on Operations Research (OR 2011). This international conference of the three German-speaking OR societies GOR, ÖGOR, and SVOR was organized by the ETH and UNI Zurich under the patronage of SVOR/ASRO. The committee chairs Karl Schmedders, (University of Zurich) and Hans-Jakob Lüthi (ETH Zurich) organized the event with more than 620 lectures within 197 sessions and three plenary speakers, Dimitris Bertsimas, Kenneth Judd, and William Pulleyblank. We are proud to report that these efforts were well received by the 850 participants from around the world.

Since 2003, the SVOR/ASRO has organized 10 Swiss Operations Research Days, in collaboration with IBM Research, Zurich. Over the years, the annual Swiss Operations Research Days have established themselves as a unique opportunity for researchers from the Swiss academic community and IBM Research to learn about each other's work and interests and to exchange current research developments in OR. The aim is to

not only exchange knowledge but also promote collaboration, initiate new projects, and have a good time with inspiring discussions.

An important and appreciated tradition of SVOR/ASRO is to print two or three times per year a paper bulletin which reaches all members. The bulletin opens a window over the main initiatives and events which involve OR in Switzerland and abroad.

One of the most important activities of the association is to promote OR among students and young researchers, encouraging them to practice this fascinating discipline. In 1997, the SVOR/ASRO established an annual prize for the best Master's thesis in OR and a three-year prize for the best doctoral thesis carried out in Switzerland in OR disciplines. The SVOR/ASRO also organizes a competition among high school students to encourage them to enroll in programs of study in mathematics and OR. Since 2004, the SVOR/ASRO has organized, in collaboration with SATW (Swiss Academy of Technical Sciences), seven competitions involving more than 200 students across Switzerland.

The SVOR/ASRO has recently received very important news. The Conference of the University of Western Switzerland (CUSO) has decided to approve and support the new "doctoral program CUSO in Operations Research." This program, which replaces the "Troisième Cycle Romand de Recherche Opérationnelle," started in 1970, has the primary objective of maintaining the same level as in previous cycles and start new initiatives, encouraging the participation of PhD students. Long-term goals are promoting OR in Switzerland and increasing its visibility at the international level by creating a solid network between research groups in Switzerland. From now on two major annual seminars are planned. In addition to the traditional winter seminar in Zinal, a new summer seminar will be organized.

The following is a list of leading academic groups in OR in Switzerland.

- Decision Support Systems, Department of Informatics, University of Fribourg.
- EPFL, École Polytechnique Fédérale de Lausanne, Lausanne

- IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, USI-SUPSI, Lugano
- Haute école de gestion de Genève
- Institute for Operations Research, Chair of Logistics Management, and Centre for Energy Policy and Economics, ETH Zurich
- Institute for Operations Research and Mathematics of Economics, University of Zurich
- ior/cf-HSG, Institute for Operations Research and Computational Finance, University of St. Gallen
- Institut universitaire Kurt Bösch, Sion
- Quantitative Methods, Department of Business Administration, University of Bern

- University of Applied Sciences of Western Switzerland
- University of Geneva
- University of Neuchâtel
- Zurich University of Applied Sciences

Acknowledgment

The authors acknowledge the SVOR/ASRO members who commented on early versions of this article.

REFERENCES

1. www.svor.ch.

SYMMETRY HANDLING IN MIXED-INTEGER PROGRAMMING

JIM OSTROWSKI
Department of Management Sciences,
University of Waterloo, Waterloo,
Ontario, Canada

INTRODUCTION

A standard method for solving integer linear programs (ILPs) is *branch and bound*. In branch and bound, the set of feasible solutions is partitioned, forming more easily solved subproblems. The presence of symmetry means that many of these subspaces are equivalent in a sense we describe later. Only one member of each collection of equivalent subspaces needs to be explored. Failure to recognize that many subspaces are equivalent results in a waste of computational effort that can render an instance unsolvable by branch and bound.

In this article, we show how the presence of a high degree of symmetry can render branch-and-bound methods useless. We also discuss how to find the symmetry group of an ILP formulation as well as how to use the symmetry group to reduce the size of the feasible region. Unfortunately, it is not clear how to best use a problem's symmetry group to reduce the feasible region. We discuss three different methods geared at exploiting symmetry in this manner. Orbitopal fixing uses the symmetry present in graph partitioning problems to fix variables throughout the branch-and-bound tree. Isomorphism pruning and orbital branching consider general ILPs, exploiting symmetry by fixing variables throughout the branch-and-bound tree and, in the case of isomorphism pruning, pruning nodes.

SYMMETRY IN INTEGER PROGRAMMING

An ILP is the problem of finding values of decision variables that maximize a linear

function, subject to a set of linear constraints with the additional restriction that values of all variables should be integral. An ILP can be written as

$$Z_{\text{ILP}} = \max_{x \in \mathbb{Z}_+^n} \{c^T x \mid Ax \leq b\}, \quad (\text{ILP})$$

where $\mathbb{Z}_+^n$ denotes the set of n -dimensional nonnegative integer vectors, $A \in \mathbb{Q}^{m \times n}$, $b \in \mathbb{Q}^m$, and $c \in \mathbb{Q}^n$, where $\mathbb{Q}$ denotes the set of rational numbers. The set $\{x \in \mathbb{Z}_+^n \mid Ax \leq b\}$ is called the *feasible region* of the ILP, which we denote as $\mathcal{F}$. A point is *feasible* if it is in the feasible region. The function that is maximized is known as the *objective function*.

Consider the following ILP from Ref. 1:

$$\begin{aligned} Z_{\text{ILP}} = \max_{x \in \{0,1\}^{n+1}} \{ & x_{n+1} | 2x_1 + 2x_2 + \cdots + 2x_n \\ & + x_{n+1} = 2k + 1 \} \\ \text{where } & k \in \mathbb{Z}_+, k \leq n. \end{aligned} \quad (1)$$

The example in Equation (1) can be easily solved without using a computer for any choice of n and k . The sum $2x_1 + 2x_2 + \cdots + 2x_n + x_{n+1}$ must always be odd, forcing $x_{n+1} = 1$, and hence the objective function, to take the value 1. Despite the problem's obvious solution, traditional branch-and-bound methods cannot solve even modest-sized instances. Table 1 gives computational results obtained from solving small problem instances of Equation (1) using the commercial solver CPLEX with its cut generation features turned off.

For any $k \in \mathbb{Z}$, $0 \leq k < n$, there will be a fractional solution feasible to the LP relaxation of Equation (1) with $x_{n+1} = 0$. Thus, nodes in the branch-and-bound tree will not be pruned until either k variables are fixed to 1 or $n - k$ variables are fixed to 0. As a result, the enumeration tree will contain at least $\min(2^k, 2^{n-k})$ nodes. Inspection of the enumeration tree, however, reveals that most of the work performed by traditional branch and bound is unnecessary. Many nodes in the

Table 1. Computational effort required to solve Equation (2)

n	k	Time (s)	Nodes
20	5	3.24	54,262
20	6	6.97	116,278
20	7	12.24	203,400
20	8	17.83	293,928
20	9	21.68	352,714
20	10	21.74	352,714
25	5	13.86	23,228
25	6	38.96	657,798
25	7	92.71	1,562,273
30	5	42.7	736,284
30	6	160.15	2,629,573

tree represent isomorphic subproblems that need to be solved only once.

To illustrate this phenomenon, consider the specific instance:

$$\min_{x \in \{0,1\}^5} \{x_5 \mid 2x_1 + 2x_2 + 2x_3 + 2x_4 + x_5 = 3\}. \quad (2)$$

The tree shown in Fig. 1 is generated using traditional branch-and-bound methods, by branching on the fractional variable with the smallest index. This branching method is, in general, a poor branching strategy. However, in this instance, the branching strategy is guaranteed to produce the smallest tree. Only nodes for which the LP relaxation is feasible are included in the figure. The LP relaxations of all nonleaf nodes in the tree have optimal values of 0, while the optimal solutions to each of the leaf nodes is also an

optimal solution to the ILP (and hence has an objective value of 1).

Consider node D , the subproblem formed by fixing x_1 to 0 and x_2 to 1. We can rewrite the subproblem as

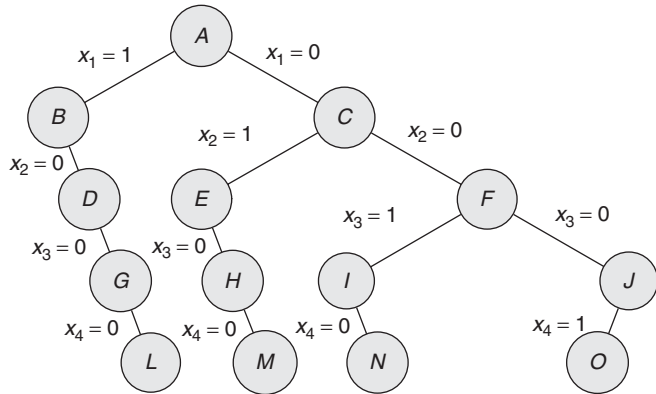
$$\min_{x \in \{0,1\}^5} \{x_5 \mid 2x_3 + 2x_4 + x_5 = 1\}. \quad (3)$$

Now consider node E , the subproblem formed by fixing x_1 to 0 and x_2 to 1. This subproblem can also be written as

$$\min_{x \in \{0,1\}^5} \{x_5 \mid 2x_3 + 2x_4 + x_5 = 1\}. \quad (4)$$

The subproblem at node D is isomorphic to the subproblem at node E . It is easy to see why traditional methods have difficulty with this type of problem as both nodes D and E are solved. It is not necessary to consider *both* nodes D and E , but traditional methods provide no way of recognizing that two nodes are isomorphic. Further inspection shows that there are many other nodes that represent isomorphic subproblems (for instance, nodes G , H , and I represent the same subproblem, as well as nodes L , M , N , and O).

Not only are isomorphic subproblems solved multiple times, but also the subproblems themselves can be difficult to solve. Note that an optimal solution is feasible in each node as shown in Fig. 1. Intuitively, nodes whose feasible region contains an optimal solution are more difficult to prune than nodes that do not because nodes with optimal solutions need more of the integrality gap to be closed.

**Figure 1.** Branching tree for example.

A good branch-and-bound algorithm should avoid solving isomorphic subproblems. Ideally, sets of isomorphic subproblems should be recognized during the branching process and all but one member of each set should be pruned. For example, node E can be pruned because it is isomorphic to node D ; nodes I and O can be pruned because they are isomorphic to nodes G and L , respectively. Pruning to avoid this repetition will significantly reduce the number of nodes in the enumeration tree, hopefully making even very large symmetric problems solvable by integer programming methods.

MATHEMATICAL PRELIMINARIES

The many isomorphic subproblems in the example arise because of the symmetry in Equation (2). To formalize the definition of symmetry, we need to briefly introduce some basic concepts from group theory. A thorough review of group theory can be found in Refs 2–4.

Let I^n be the ground set $\{1, \dots, n\}$. Let S^n be the collection of permutations of the ground set, that is, the collection of all functions π such that $\pi : I^n \rightarrow I^n$ is a one-to-one and onto (bijective) mapping. A binary operation acting on permutations is the *composition* operation, that is, $\pi_i \circ \pi_j = \pi_i(\pi_j)$. For notational convenience, a group containing permutations will be referred to only by the set of elements $\mathcal{G}$. S^n is the *symmetric group* of I^n . Any subset of S^n that satisfies the properties of a group is called a *permutation group*. To provide emphasis and to distinguish S^n from a general permutation group, S^n will often be referred to as *the complete symmetric group*.

Given a vector $\lambda \in \mathbb{R}^n$ and a permutation group $\mathcal{G}$, $\pi \in \mathcal{G}$ acts on λ by permuting its coordinates: $\pi(\lambda) = (\lambda_{\pi_1}, \lambda_{\pi_2}, \dots, \lambda_{\pi_n})$. A permutation group can also act on a collection of vectors. For $X \subset \mathbb{R}^n$, $X^{\mathcal{G}}$ is given by $\{\pi(x) | x \in X \text{ and } \pi \in \mathcal{G}\}$.

Cyclic notation is used to describe permutations. For any $i \in \{1, \dots, n\}$ and permutation π , π 's action on i can be represented by the sequence $(i, \pi(i), \pi^2(i), \dots)$, where $\pi^2(i) = \pi(\pi(i))$. Because $\mathcal{G}$ is a finite group, the elements of this sequence repeat. Let p be the

first power of π such that $\pi^p(i) = i$. The cycle $(i, \pi(i), \pi^2(i), \dots, \pi^{p-1}(i))$ can be used to describe how the permutation acts on a subset of the elements. The permutation maps i to $\pi(i)$. It also sends the element $\pi(i)$ to $\pi^2(i)$. The set I^n can be partitioned using subsets formed by cycles of π . The collection of cycles formed by the partition of I^n defines the permutation. For example, suppose π is a permutation with $\pi(1) = 1$, $\pi(2) = 3$, $\pi(3) = 2$, $\pi(4) = 5$, $\pi(5) = 4$. The permutation π can be written as $(1)(2, 3)(4, 5)$. Since π does not permute element 1, the cycle (1) need not be included. The permutation π can be written more succinctly as $(2, 3)(4, 5)$. For any element not found in a cycle, it can be assumed that the element is not changed by the permutation.

Let $\mathcal{G}$ be an arbitrary permutation group acting on $\{1, 2, \dots, n\}$. The group $\mathcal{G}$ can be extended to act on sets of elements. For $S \subseteq \{1, 2, \dots, n\}$, the set $\pi(S) = \{\pi(i) | i \in S\}$. For a point $z \in \mathbb{R}^n$, the *orbit* of z under the action of the group $\mathcal{G}$ is the set of all elements of $\mathbb{R}^n$ to which z can be mapped by permutations in $\mathcal{G}$, that is,

$$\text{orb}(\mathcal{G}, z) \stackrel{\text{def}}{=} \{z' \in \mathbb{R}^n | \exists \pi \in \mathcal{G} \text{ such that } z' = \pi(z)\} = \{\pi(z) | \pi \in \mathcal{G}\} = z^{\mathcal{G}}. \quad (5)$$

The notion of orbits can be extended to variables by considering the unit vector. We can consider variables x_i and x_j to be symmetric if $j \in \text{orb}(\{i\}, \mathcal{G})$. By definition, if $j \in \text{orb}(\{k\}, \mathcal{G})$, then $k \in \text{orb}(\{j\}, \mathcal{G})$, that is, the elements j and k share the same orbit. Therefore, the union of the orbits

$$\mathcal{O}(\mathcal{G}) \stackrel{\text{def}}{=} \bigcup_{j=1}^n \text{orb}(\{e_j\}, \mathcal{G}) \quad (6)$$

forms a partition of $I^n = \{1, 2, \dots, n\}$, which is referred to as the *orbital partition* of $\mathcal{G}$, or simply the *orbits* of $\mathcal{G}$. Any two elements of I^n whose corresponding unit vectors share an orbit under the group $\mathcal{G}$ are *equivalent* or *symmetric* with respect to $\mathcal{G}$.

The *stabilizer* of a set $S \subseteq I^n$ in $\mathcal{G}$ is the set of permutations in $\mathcal{G}$ that send S to itself:

$$\text{stab}(S, \mathcal{G}) = \{\pi \in \mathcal{G} | \pi(S) = S\}. \quad (7)$$

The stabilizer of S is a subgroup of $\mathcal{G}$. The set of permutations that stabilize each set in the collection $\{S_1, S_2, \dots, S_k\}$ is written as

$$\text{stab}(S_1, S_2, \dots, S_k, \mathcal{G}) = \bigcap_{j=1}^k \text{stab}(\{S_j\}, \mathcal{G}). \quad (8)$$

For any collection of permutations, $\mathcal{A}$, $\langle \mathcal{A} \rangle$ is defined to be the smallest group containing all permutations in $\mathcal{A}$. The group $\langle \mathcal{A} \rangle$ is the group generated by $\mathcal{A}$, and similarly the set $\mathcal{A}$ is a *generating set*, or *generator*, of $\langle \mathcal{A} \rangle$. Thus, any group can be compactly represented by a generating set.

Example 1. Let $\mathcal{A} \subset S^4$ consist of the permutations $\pi^1 = (2, 3)$ and $\pi^2 = (1, 4)$. $\langle \mathcal{A} \rangle = \{(), (1, 4), (2, 3), (1, 4)(2, 3)\}$ is the smallest group containing $\mathcal{A}$. The permutations π_1 and π_2 are generators of $\langle \mathcal{A} \rangle$. Because $\pi_1(1) = 1$ and $\pi_1(4) = 4$, $\pi_1 \in \text{Stab}(\{1\}, \langle \mathcal{A} \rangle)$, $\pi_1 \in \text{Stab}(\{4\}, \langle \mathcal{A} \rangle)$, and $\pi_1 \in \text{Stab}(\{1\}, \{4\}, \langle \mathcal{A} \rangle)$. In addition, $\pi_1 \in \text{Stab}(\{2, 3\}, \langle \mathcal{A} \rangle)$ since elements 2 and 3 are not mapped outside the set $\{2, 3\}$. Since π_1 maps element 2–3, we have $\text{Orb}(\{2\}, \langle \mathcal{A} \rangle) = \{2, 3\}$. The orbital partition, $\mathcal{O}(\langle \mathcal{A} \rangle)$, is $(\{1, 4\}, \{2, 3\})$. We can also consider orbits of sets of elements. For example, $\text{orb}(\{1, 2\}, \langle \mathcal{A} \rangle) = (\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\})$.

SYMMETRY GROUPS OF INTEGER LINEAR PROGRAMS

The *symmetry group* $\mathcal{G}(\text{ILP})$ of an integer program is the collection of permutations in S^n that map every feasible solution of (ILP) of value t to another feasible solution of (ILP) also of value t [5]. Formally,

$$\mathcal{G}(\text{ILP}) \stackrel{\text{def}}{=} \{\pi \in S^n \mid \forall x \in \mathcal{F}(\text{ILP}), \pi(x) \in \mathcal{F} \text{ and } c^T \pi(x) = c^T x\}. \quad (9)$$

Computing the symmetry group of a general ILP is NP-hard, and typically more difficult than solving the ILP itself. As a result, practical methods aimed at exploiting symmetries are forced to use a subset of the symmetry group that is found by examining the problem formulation.

Given a permutation $\pi \in S^n$ and a permutation $\sigma \in S^m$, let $A(\sigma, \pi)$ be the matrix obtained by permuting the columns of A by π and the rows of A by σ , that is, $A(\sigma, \pi) = P_\sigma A P_\pi$, where P_σ and P_π are appropriate permutation matrices. The *formulation group* $\mathcal{G}(A, b, c)$ of (ILP) is the set of permutations:

$$\mathcal{G}(A, b, c) \stackrel{\text{def}}{=} \{\pi \in \Pi^n \mid \pi(c) = c, \exists \sigma \in S^m \text{ with } \sigma(b) = b \text{ such that } A(\sigma, \pi) = A\}. \quad (10)$$

For any $\pi \in \mathcal{G}(A, b, c)$, if $\hat{x}$ is feasible for (ILP) (or the LP relaxations of (ILP)), then $\pi(\hat{x})$ is also feasible. Moreover, the solutions $\hat{x}$ and $\pi(\hat{x})$ have the same objective value. Thus, $\mathcal{G}(A, b, c) \subseteq \mathcal{G}(\text{ILP})$. As an example, we consider the following ILP:

Example 2.

$$\begin{aligned} \min \sum_{i=0}^8 x_i & \quad (11) \\ \text{subject to} & \\ x_0 + x_1 + x_2 + x_3 & \quad + x_6 \geq 1 \\ x_0 + x_1 + x_2 & \quad + x_4 \quad + x_7 \geq 1 \\ x_0 + x_1 + x_2 & \quad + x_5 \quad + x_8 \geq 1 \\ x_0 & \quad + x_3 + x_4 + x_5 + x_6 \geq 1 \\ x_1 & \quad + x_3 + x_4 + x_5 \quad + x_7 \geq 1 \\ x_2 + x_3 + x_4 + x_5 & \quad + x_8 \geq 1 \\ x_0 & \quad + x_3 \quad + x_6 + x_7 + x_8 \geq 1 \\ x_1 & \quad + x_4 \quad + x_6 + x_7 + x_8 \geq 1 \\ x_2 & \quad + x_5 + x_6 + x_7 + x_8 \geq 1 \\ x & \in \{0, 1\}^9 \end{aligned}$$

Consider the dominating set problem defined in Equation (11). The formulation group $\mathcal{G}(A, b, c)$ in this example contains 72 permutations. However, it can be generated using just four permutations. The generators are listed in Table 2 along with the corresponding σ such that $A(\sigma, \pi) = A$.

Because A is a symmetric matrix, each permutation π and its corresponding σ are identical. For general ILP problems, it is

Table 2. Generators for $\mathcal{G}(A, b, c)$

π	σ
(3, 6)(4, 7)(5, 8)	(3, 6)(4, 7)(5, 8)
(1, 2)(4, 5)(7, 8)	(1, 2)(4, 5)(7, 8)
(1, 3)(2, 6)(5, 7)	(1, 3)(2, 6)(5, 7)
(0, 1)(3, 4)(6, 7)	(0, 1)(3, 4)(6, 7)

Table 3. Generators for $\text{Stab}(\{0\}, \mathcal{G}(A, b, c))$

π
(3, 6)(4, 7)(5, 8)
(1, 2)(4, 5)(7, 8)
(1, 3)(2, 6)(5, 7)

unlikely that any such pair of permutations will be identical. Indeed, if A is not a square matrix, π and σ will not act on the same set of elements.

The orbital partition $\mathcal{O}(\mathcal{G}(A, b, c))$ consists of just one orbit, $\{0, 1, \dots, 8\}$. The stabilizer of 0 (i.e., $\text{stab}(\{0\}, \mathcal{G}(A, b, c))$) contains only eight permutations, listed in Table 3. The orbits of the stabilizer $\text{stab}(\{0\}, \mathcal{G}(A, b, c))$ are $\{0\}$, $\{1, 2, 3, 6\}$, and $\{4, 5, 7, 8\}$.

An optimal dominating set for the graph is the set $\{0, 3, 6\}$. We can find other equivalent solutions by permuting this set with any of the permutations in $\mathcal{G}(A, b, c)$. For example, consider permuting the set $\{0, 3, 6\}$ with the last generator, $(0, 1)(3, 4)(6, 7)$. In this case, $\{0, 3, 6\}$ is mapped to the set $\{1, 4, 7\}$. The cover consisting of $\{1, 4, 7\}$ is then an equivalent optimal solution.

The set of all optimal dominating sets equivalent to $\{0, 3, 6\}$, written as $\{0, 3, 6\}^{\mathcal{G}(A, b, c)}$, is the collection of sets:

$$\{0, 3, 6\}^{\mathcal{G}(A, b, c)} = \{\{0, 3, 6\}, \{1, 4, 7\}, \{2, 5, 8\}\}.$$

Not all optimal dominating sets are equivalent to the set $\{0, 3, 6\}$. For example, there is no permutation that maps the set $\{0, 3, 6\}$ to $\{0, 4, 8\}$. It is easy to see that $\{0, 3, 6\}$ and $\{0, 4, 8\}$ are not equivalent by noting that x_0 , x_3 , and x_6 appear together in a constraint (the first constraint), while the variables x_0 , x_4 , and x_8 do not. Another point of interest

can be seen by examining all sets equivalent to $\{0, 4, 8\}$:

$$\{0, 4, 8\}^{\mathcal{G}(A, b, c)} = \{\{0, 4, 6\}, \{0, 5, 7\}, \{1, 3, 8\}, \{1, 5, 6\}, \{2, 3, 7\}, \{2, 5, 6\}\}.$$

The number of solutions equivalent to a given solution is solution dependent, making it difficult to predict how many equivalent optimal solutions exist.

The equivalence of solutions induced by symmetry is a major factor that might confound the branch-and-bound process. For example, suppose that x^* is an (integral) optimal solution to a binary ILP. At the root node, the decision is made to branch on variable x_j , creating one subproblem by fixing $x_j = 0$ and another by fixing $x_j = 1$. If $\exists \pi \in \mathcal{G}(\text{ILP})$ such that $[\pi(x^*)]_j = 1 - x_j^*$, then x^* is a feasible solution for one child node and $\pi(x^*)$ is feasible for the other child node. In general, subproblems where feasible regions contain optimal solutions will be more difficult to solve because the proportion of the integrality gap that must be closed is greater than a subproblem that does not contain any optimal solution. If the cardinality of $\mathcal{G}(\text{ILP})$ is large, then there may be many such subproblems, leading to long solution times.

Note that the formulation group is very dependent on the problem formulation. For example, consider adding the constraint $\sum_{i=0}^8 2^i x_i \geq 0$ to Example 2. This constraint does not remove any feasible points from the LP relaxation and does not remove symmetry from the ILP, but it does remove all of the symmetry from the formulation.

COMPUTING SYMMETRY GROUPS AND FORMULATION GROUPS

Recall that the formulation group of an ILP is

$$\mathcal{G}(A, b, c) \stackrel{\text{def}}{=} \{\pi \in \Pi^n \mid \pi(c) = c, \exists \sigma \in S^m \text{ with } \sigma(b) = b \text{ such that } A(\sigma, \pi) = A\}. \quad (12)$$

A popular method of computing the formulation group is by constructing a graph whose automorphism group is similar to the formulation group [6–9].

In the case where A is a $(0, 1)$ -matrix and both c and b are vectors of ones, then the formulation group can be found by considering the bipartite graph, H , associated with the adjacency matrix:

$$\begin{pmatrix} A & 0 \\ 0 & A^t \end{pmatrix}. \quad (13)$$

The graph H contains vertices representing variables and vertices representing constraints. Permutations that map H to itself only make sense in the context of integer programming if they map variable-vertices to other variable-vertices (and likewise constraint-vertices to other constraint-vertices). Thus, if Γ is the symmetry group of H , then $\mathcal{G}(\text{ILP}) = \text{stab}(\{1, \dots, n\}, \Gamma)$. More general cases require the construction larger, colored graphs. Details can be found in Refs 10 and 11.

Computing the formulation group for an ILP is equivalent to the graph isomorphism problem. Graph isomorphism is neither known to be NP-hard nor to be polynomial time solvable [12], even when the graph is bipartite. However, current software such as nauty [13] and Saucy [14] are able to quickly compute the automorphism group of most reasonably sized graphs.

EXPLOITING SYMMETRY WHEN COMPUTING THE RELAXATION

Typically in integer programming, a relaxation of the ILP is formed by ignoring the integrality constraints on the variables, creating the linear program (LP). This gives a lower bound to the problem, which can then be used to prune nodes. The LP solver can use symmetry to reduce the size of the relaxation.

Let y be any feasible solution to the LP relaxation of (ILP) with symmetry group $\mathcal{G}$. By the definition of $\mathcal{G}$ and the convexity of $\mathcal{F}$, any linear combination of elements of $\{\pi(y) | \pi \in \mathcal{G}\}$ is in $\mathcal{F}$. As a result, an optimal solution, x^* , to the LP relaxation of the form $x^* = \frac{\sum_{\pi \in \mathcal{G}} \pi(x^*)}{|\mathcal{G}|}$ is guaranteed to exist.

Thus, the LP relaxation formed by adding the equalities $x_i = \frac{\sum_{\pi \in \mathcal{G}} \pi(x_i)}{|\mathcal{G}|}$ will provide an identical LP bound. For any $j \in \text{orb}(i, \mathcal{G})$, there is a permutation σ that sends j to i . Thus, $\frac{\sum_{\pi \in \mathcal{G}} \pi(x_j)}{|\mathcal{G}|} = \frac{\sum_{\pi \in \mathcal{G}} \pi(\sigma(x_i))}{|\mathcal{G}|}$. By the definition of a group, π and σ in $\mathcal{G}$ implies that $\pi(\sigma^{-1})$ is also in $\mathcal{G}$. Thus, $\frac{\sum_{\pi \in \mathcal{G}} \pi(\sigma(x_i))}{|\mathcal{G}|} = \frac{\sum_{\pi \in \mathcal{G}} \pi(x_i)}{|\mathcal{G}|}$.

Adding the constraints $x_i = \frac{\sum_{\pi \in \mathcal{G}} \pi(x_i)}{|\mathcal{G}|}$ for each i is equivalent to adding $x_i = x_j$ for all $j \in \text{orb}(i, \mathcal{G})$. Using these equalities, the number of variables can be reduced to the number of orbits in the partition $\mathcal{O}(\mathcal{G})$ (the number of constraints can be reduced in a similar way). This reduction can decrease the time required to solve the LP significantly; however, the given LP solution can be highly fractional.

Using the symmetry group to aid in the computation of the LP relaxation has not been a well-researched area. This is mostly due to the fact that most problems exhibiting a large degree of symmetry are very hard, and only instances of small size are solvable. For these problems, the LP relaxations are easy to solve. This is not the case, however, in the context of semidefinite programming (SDP). Some interesting work has been done at using symmetry to reduce the size of SDPs, particularly for the quadratic assignment problem [15].

REMOVING SYMMETRY FROM THE FEASIBLE REGION

Another class of symmetry-breaking algorithms uses the symmetry group of the problem to reduce the feasible region by removing sets of equivalent solutions. This strategy of reducing the feasible region allows for more general techniques than reformulation.

Recall Example 2, a dominating set problem on nine nodes. Let $\mathcal{G}$ be the symmetric group. At the root node, all variables in the problem share an orbit. Because it is clear that at least one variable must be equal to 1, an arbitrary variable can be chosen and fixed to 1. For example, x_0 can be fixed to 1. The reason for fixing x_0 to 1 is very intuitive. Any optimal solution must contain at least one variable with value 1. Suppose that in a given optimal solution, $\hat{x}$, there is a j with $\hat{x}_j = 1$ for some $j \in \{0, \dots, 8\}$.

Since all variables are symmetric at the root node, there is a permutation $\pi \in \mathcal{G}$ that maps element j to element 0. Then, $\pi(\hat{x})$ is feasible and optimal (by definition of $\mathcal{G}$) with $\pi(\hat{x}_0) = 1$. By setting $x_0 = 1$, solutions are removed from the feasible region, but every solution that is removed is symmetric to at least one solution remaining in the feasible region. Hence, fixing an arbitrary variable to 1 serves to remove some symmetry as well as to tighten the LP bound.

A similar technique is used in graph coloring problems. Given a graph G , a coloring is valid if no two adjacent vertices are assigned the same color. Given a valid coloring, equivalent valid colorings can be generated by relabeling colors. A typical integer program (IP) formulation for this problem contains binary variables x_i^j that represent if vertex i is colored j and binary variables y^l that represent if any vertex has color l . The mathematical formulation is

$$\begin{aligned} \min \quad & \sum_{l=1}^k y^l \\ \text{s.t.} \quad & x_i^l + x_j^l \leq y^l \quad \forall (i,j) \in E \quad \forall l \in \{1, \dots, k\} \\ & \sum_l x_i^l = 1 \quad \forall i \in V \\ & y^l, x_i^l \in \{0, 1\} \quad \forall i \in V \quad \forall l \in \{1, \dots, k\}. \end{aligned} \quad (14)$$

The symmetry group of formulation (14) contains the permutations

$$\begin{aligned} \pi_{l,m} &= (x_1^l, x_1^m)(x_2^l, x_2^m) \dots (x_n^l, x_n^m) \\ &\quad \forall l, m \in \{1, \dots, k\}, \quad l \neq m, \end{aligned}$$

simply meaning that colors can be arbitrarily relabeled.

Suppose that G contains a clique of size k . No two vertices in the clique can be assigned the same color. Given that no vertex is currently colored, each of the k vertices in the clique can be arbitrarily colored with each of the first k colors. As with fixing a variable to 1 in Example 2, this fixing of colors removes symmetric solutions from the feasible region and reduces the size of the symmetry group acting on the problem by at least a factor of $k!$.

As before, this fixing is valid because for each solution that is removed, there is at least one equivalent solution remaining in the feasible region. These intuitive fixing rules aim to reduce the size of the feasible region (as well as the symmetry) of the problem instance by removing a large set of feasible solutions. A solution can be removed as long as it is guaranteed that for each solution removed a *representative solution* (i.e., an equivalent solution) remains in the feasible region. This is formalized in Ref. 16 where Eric Friedman adapts a notion from geometry. For a permutation group Γ and a set $X \subset \mathbb{R}^n$, a *generalized fundamental domain* of X with respect to Γ is a subset F of X such that X can be constructed using the points in F along with the permutations in Γ , that is, $F^\Gamma = X$.

Formally, a set F is a generalized fundamental domain of X with respect to Γ if and only if for every $x \in X$, the set $\text{orb}(x, \Gamma) \cap F$ is nonempty. A fundamental domain is *minimal* if it does not contain a smaller fundamental domain.

Example 3. If $X = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$ and $\Gamma = \{e, \pi = (1, 2)\}$, there are three different fundamental domains. The set X is a trivial a fundamental domain because $X^\Gamma = X$. In addition, the set $F_1 = \{(0, 0), (1, 0), (1, 1)\}$ is a fundamental domain. This can be seen by noting that the only element in X that is not in F_1 is the element $(0, 1)$, but $\pi((1, 0)) = (0, 1)$, so $F_1^\Gamma = X$. For the same reason, the set $F_2 = \{(0, 0), (0, 1), (1, 1)\}$ is also a fundamental domain. No two elements in F_1 are symmetric, so F_1 is a minimal fundamental domain with respect to Γ . Similarly, F_2 is also minimal.

For a polyhedron $X \subseteq \mathbb{R}^n$, a permutation group Γ , and an n -vector c consider the set

$$F_c(X, \Gamma) = \{x \in X \mid c^T x \geq c^T \pi(x) \quad \forall \pi \in \Gamma\}. \quad (15)$$

For any $x \in \mathcal{F}$, at least one element $y \in \text{orb}(x, \Gamma)$ satisfies the constraints $c^T y \geq c^T \pi(x) \forall \pi \in \Gamma$ so $F_c(X, \Gamma)$ is a fundamental domain of X with respect to Γ . If c is such that only one feasible solution

satisfies the set of constraints, then F_c is a minimal fundamental domain. One such c comes from enforcing a lexicographic ordering of the variables, that is, $c_{lex} = (2^{n-1}, 2^{n-2}, \dots, 2, 1)$. $F_{c_{lex}}$ is a minimal fundamental domain [16] for all $X \subseteq \{0, 1\}^n$ and Γ .

Example 4. For Example 2, the constraints of $F_{c_{lex}}$ that result from the generators of $\mathcal{G}$ are given in Table 4.

There is an immediate concern with using lexicographic constraints. Adding constraints that define $F_{c_{lex}}$ will cause problems for most numerical methods because of the magnitude of the coefficients in c_{lex} . However, these constraints guarantee that the resulting fundamental domain is minimal for any $X \subseteq \{0, 1\}^n$ and Γ . Despite the numerical issues, most literature focuses on choosing fundamental domains based on lexicographic ordering (generating fundamental domains by selecting lexicographic either maximal or minimal elements).

In the context of integer programming, X represents the set of feasible solutions, $\mathcal{F}$, to a problem and Γ represents the symmetry group of the problem, $\mathcal{G}$. By definition, for any fundamental domain F of $\mathcal{F}$ with respect to $\mathcal{G}$ (ILP), any optimal solution in $\mathcal{F}$ is symmetric to a solution in F . Hence,

$$\min_{x \in F} \{c^T x\} = \min_{x \in \mathcal{F}} \{c^T x\}. \quad (16)$$

Unfortunately, restricting the feasible region to $\mathcal{F}$ does not improve the LP bound values. This is because one representative of each set of symmetric integer solutions *and* fractional solutions remain in $\mathcal{F}$.

STATIC VERSUS DYNAMIC SYMMETRY-BREAKING METHODS

The search of (ILP) can be restricted to a fundamental domain. The issue then becomes how to choose a fundamental domain. Symmetry-breaking tools that use fundamental domains can be split into two different classes. Static symmetry-breaking methods determine the fundamental domain at the start of the algorithm, and dynamic methods determine the fundamental domain during the branching process.

Static symmetry-breaking methods can be very effective at solving highly symmetric problems. Methods such as adding symmetry-breaking constraints only require an alteration of the problem formulation and do not require special software. In addition, knowing the domain explicitly at every node in the tree makes it possible to fix variables based on the (implicit or explicit) fundamental domain constraints.

Static symmetry-breaking methods have some limitations. The first, as previously mentioned, is the potentially large number of inequalities needed to remove all of the symmetry from the problem. This problem can sometimes be avoided with clever algorithms. Secondly, restricting the search to an arbitrarily chosen fundamental domain could distort the branching process by changing the relative importance of the variables. For example, restricting the feasible region to the set of lexicographically minimal solutions would significantly increase the importance of variables with small indices. Algorithms that, in theory, have no branching restrictions may require branching in a rigid way to achieve the best results.

Static methods can make finding optimal solutions more difficult, as the solution found

Table 4. Lexicographic constraints

π	Lexicographical constraint
(3,6)(4,7)(5,8)	$32x_3 + 16x_4 + 8x_5 + 4x_6 + 2x_7 + x_8 \geq 32x_6 + 16x_7 + 8x_8 + 4x_3 + 2x_4 + x_5$
(1,2)(4,5)(7,8)	$128x_1 + 64x_2 + 16x_4 + 8x_5 + 2x_7 + x_8 \geq 128x_2 + 64x_1 + 16x_5 + 8x_4 + 2x_8 + x_7$
(1,3)(2,6)(5,7)	$128x_1 + 64x_2 + 31x_3 + 8x_5 + 4x_6 + 2x_7 \geq 128x_3 + 64x_6 + 32x_1 + 8x_7 + 4x_2 + 2x_5$
(0,1)(3,4)(6,7)	$256x_0 + 128x_1 + 32x_3 + 16x_4 + 4x_6 + 2x_7 \geq 256x_1 + 128x_0 + 32x_4 + 16x_3 + 4x_7 + 2x_6$

must also be in the chosen fundamental domain. Adding symmetry-breaking constraints to the root node may, in fact, remove optimal solutions that, along with a specified branching rule, would have been easier to find than the equivalent optimal solution remaining in the fundamental domain. This has been studied in the context of constraint programming in Ref. 17, where the authors show that branching strategies can drastically alter the time required to find solutions. The problem caused by pruning these easily found optimal solutions can be overcome with more intelligent branching. For instance, if the set of representative solutions is chosen based on a lexicographic order, then branching in that order would avoid this problem. However, rules like this are not ideal, as branching is very important in every integer program, not just those that contain symmetry. Therefore, choosing a branching rule *a priori* is not desirable.

Dynamic symmetry-breaking strategies differ from static methods because they construct the fundamental domain during the solution process, not *a priori*. Because the fundamental domain is not predefined, the branching decisions can influence the fundamental domain and not vice versa.

Because the fundamental domain is defined during the branching process, dynamic symmetry-breaking methods cannot use off-the-shelf software. To date, no such software exists. In addition, because the domain is not fully defined at nodes in the tree, fewer variables can be fixed than with static methods.

Symmetry-breaking strategies designed for specific problem classes are typically static methods. Meller *et al.* [18] attack symmetry in an optimal facility layout design using constraints that reduce the size of the fundamental domain. Sherali and Smith [19] present three different applications where symmetry arises: telecommunications network-design problems, noise pollution problems, and machine procurement and operation problems. They also discuss symmetry-reducing constraints for these specific problems. Mendez-Diaz and Zabala present formulations, as well as symmetry-breaking inequalities, for the graph coloring problem

[20]. They also present computational results investigating several different ways to handle symmetry [21].

Symmetry Breaking via Orbitopes

Explicitly stating all the lexicographic inequalities (as in Table 4) is not practical for many reasonably sized problems. One avenue of research is to find classes of problems for that there are efficient ways to enforce lexicographic ordering without adding exponentially many ill-conditioned constraints to the LP formulation. This is the purpose of [22]. In their work, Kaibel and Pfetsch consider the set of all 0/1 matrices of size $p \times q$, $\mathcal{M}_{p,q}$. Given a symmetry group $\mathcal{G}$ acting on the columns of such a matrix, they define the *full orbitope* $O_{p,q}(\mathcal{G})$ to be the convex hull set of matrices of $\mathcal{M}_{p,q}$ that are lexicographically maximal within their orbits (where the ordering of the variables is row by row). For example, the 0/1 matrix

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

does not have lexicographically decreasing columns, so it is not in the orbitope $O_{5,4}(S^4)$. The permutation (3, 4) swaps columns 3 and 4, giving the matrix:

$$X' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

The matrix X' has lexicographically decreasing columns, so it is in $O_{5,4}(S^4)$.

Kaibel and Pfetsch focus their attention on both *packing* and *partitioning orbitopes*, where either the cyclic group (C^q) or the complete symmetric group (S^q) acts on the columns. The packing orbitope, $O_{p,q}^{\leq}(\mathcal{G}) \subset O_{p,q}(\mathcal{G})$, consists of all lexicographically maximal 0/1 matrices where each row contains at most a single 1. The partitioning orbitope, $O_{p,q}^=(\mathcal{G}) \subset O_{p,q}(\mathcal{G})$, consists of all lexicographically maximal matrices where

each row contains exactly a single 1. Packing and partitioning orbitopes arise in a wide array of problems, like graph coloring and partitioning, bin packing, as well as some scheduling problems. It should be noted that the set-packing/partitioning need only occurs as a substructure of the problem at hand. Additional inequalities may be present, so long as their effect on the column symmetry group is accounted for.

Kaibel and Pfetsch [22] provide linear descriptions of all facet-defining inequalities for the convex hull of orbitopes $O_{p,q}^{\leq}(C^q)$, $O_{p,q}^{\leq}(S^q)$, $O_{p,q}^=(C^q)$, and $O_{p,q}^=(S^q)$. They also provide separation algorithms for all inequalities that run in polynomial time. The facet-defining inequalities have only $\{-1, 0, 1\}$ coefficients, avoiding the numerical issues that arise when lexicographic constraints are used to eliminate symmetry. However, to completely describe the polytope $O_{p,q}^{\leq}(S^q)$ and $O_{p,q}^=(S^q)$, an exponential number of linear inequalities is required.

This technique can be applied to classes of integer programming problems as follows. Elements in a partitioning orbitope represent possible solutions to IP formulations of set-partitioning problems such as graph coloring, while the packing orbitope represents solutions to IP formulations of set-packing problems. For example, imagine a set-partitioning problem in which items are placed in one of q indistinguishable sets. The variable $x_{i,j} = 1$ if item i is placed in set j . A solution to the problem can be represented as a $p \times q$ matrix, where p is the number of items and q is the number of sets. Since the sets are indistinguishable, any permutation of the sets will map feasible partitions to other feasible partitions. Therefore, every permutation of the columns is in the symmetry group. Symmetry in the formulation can be removed by adding the additional constraint that the solution must also be an element of the orbitope $O_{p,q}(S^q)$. Kaibel and Pfetsch [22] show that enforcing membership in either $O_{p,q}^{\leq}(C^q)$ or $O_{p,q}^=(C^q)$ is simple. Ensuring that the first column is the lexicographically largest column in the matrix is enough to guarantee that the matrix is in the appropriate orbitope. Note also that C^q contains only $q - 1$ permutations, so only $q - 1$

constraints are needed to enforce a lexicographic ordering.

Enforcing membership in either $O_{p,q}^{\leq}(S^q)$ or $O_{p,q}^=(S^q)$ is not as simple as it is for the cyclic group C^q . Kaibel and Pfetsch note that the projection of $O_{p,q}^=(S^q)$ is affinely isomorphic to $O_{p-1,q-1}^{\leq}(S^{q-1})$, so they restrict their attention to the orbitope $O_{p,q}^=(S^q)$. The main result of their work is that *shifted column inequalities* (SCIs), along with row-sum constraints and lower bound constraints, provide a complete linear description of $O_{p,q}^{\leq}(S^q)$. To describe the SCIs, some notation is necessary. A *bar* is formed by indices of the matrix

$$B = \{(i,j), (i,j+1), \dots, (i,q)\} \quad (17)$$

for some $2 \leq i \leq p$, $2 \leq j \leq \min\{i, q\}$. A *shifted column* (SC) is a set of indices

$$S = \{(i_1, j_1), (i_2, j_2), \dots, (i_\eta, j_\eta)\} \quad (18)$$

where $i_1 \leq i_2 \leq i_\eta$, $j_1 \leq j_2 \leq j_\eta \leq j$, and no two elements in S share the same diagonal in the matrix. For any bar B and SC S , they call $\sum_{(i,j) \in B} x_{i,j} \leq \sum_{(i,j) \in S} x_{i,j}$ a *shifted column inequality*. The orbitope $O_{p,q}^=(S^q)$ can be completely described by nonnegativity constraints, the row-sum equations $\sum_j x_{i,j} = 1$ for all i , and the SCIs formed by all bars and SCs.

Example 5. Figures 2–4 are three different SCs and bars. A “+” denotes elements in the bar while a “−” denotes elements in the SC.

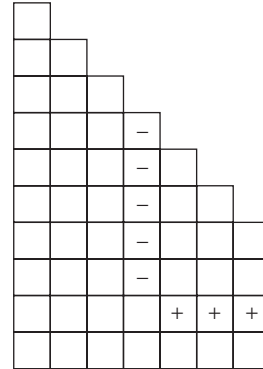


Figure 2. Shifted column 1.

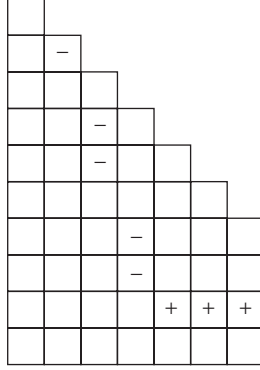


Figure 3. Shifted column 2.

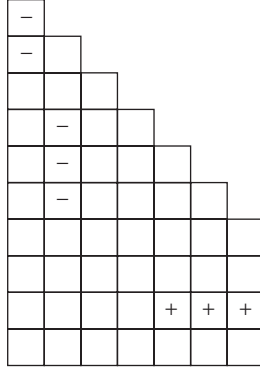


Figure 4. Shifted column 3.

The inequalities derived from the SC and the bar inequalities are listed in Table 5.

Note that in partitioning problems $\sum_{(i,j) \in B} x_{i,j} \leq 1$. If an SCI is violated, it must be because $\sum_{(i,j) \in B} x_{i,j} = 1$ and $\sum_{(i,j) \in S} x_{i,j} = 0$. Consider an example of a matrix that violates the first SCI shown above. Clearly the matrix in Figure 5 is not lexicographically maximal, as column 4 can be swapped with column 5 to create a lexicographically larger matrix.

Kaibel *et al.* [23] use the linear time algorithm to fix variables throughout the enumeration tree. They call their fixing *orbitopal fixing*. They also test orbitopal fixing on a class of graph partitioning problems and compare it to a more general symmetry breaking method, isomorphism pruning, which is detailed in the following

Table 5. Shifted column and bar inequalities in orbitopes

Figure	Shifted column inequality
2	$x_{9,5} + x_{9,6} + x_{9,7} \leq x_{4,4} + x_{5,4} + x_{6,4} + x_{7,4} + x_{8,4}$
3	$x_{9,5} + x_{9,6} + x_{9,7} \leq x_{2,2} + x_{4,3} + x_{5,3} + x_{7,4} + x_{8,4}$
4	$x_{9,5} + x_{9,6} + x_{9,7} \leq x_{1,1} + x_{2,1} + x_{4,2} + x_{5,2} + x_{6,2}$

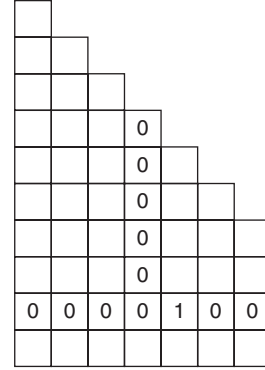


Figure 5. Matrix that does not satisfy SCI.

section. Unlike isomorphism pruning, using orbitopal fixing does not restrict branching behavior, and the results suggest that the gain in performance as a result of using orbitopal fixing over isomorphism pruning is a result of the flexible branching. Further efforts by Kaibel *et al.* [24] show that finding constraints for slightly more general orbitopes quickly becomes NP-hard. Finding orbitopal symmetries is discussed in Ref. 25.

Isomorphism Pruning

The leading approach for solving

$$\min_{x \in \{0,1\}^n} \{c^T x | Ax \leq b\}, \quad (19)$$

where Equation (19) is highly symmetric ($|\mathcal{G}(\text{BIP})|$ is large) and is isomorphism pruning, developed for integer programming by Margot [26–29]. Isomorphism pruning uses the branching decisions to generate a minimal fundamental domain dynamically. Let $\mathcal{T}$ be any branch-and-bound tree for

Equation (19) where nodes are only pruned if the LP relaxation is infeasible and not by bound. Each node at depth n in $\mathcal{T}$ has all variables fixed and, therefore, corresponds to a feasible integral solution of the ILP. The tree is oriented by letting the branch $x_i = 1$ be the left branch ($x_i = 0$ the right branch). The fundamental domain used by isomorphism pruning is defined by the branch-and-bound tree $\mathcal{T}$. The fundamental domain consists only of the leftmost element of each collection of equivalent feasible solutions. The goal of isomorphism pruning, then, is to recognize when all solutions of a given node are equivalent to solutions to the left of that node.

For any feasible solution x in node a of the tree $\mathcal{T}$, the branching history from the root node of $\mathcal{T}$ to a is needed in order to determine whether there is an equivalent solution to x that is found at a node to the left of a (to determine whether x is in the fundamental domain). This branching information is encoded in the vector M^a that ranks variables based on the order in which they were branched. Suppose that x_i was the first variable branched on at an ancestor of a (the variable chosen for branching at the root node). Index i , then, has rank “1” at node a and $M[i]^a = 1$. Similarly, if node j was branched on at the ancestor of node a with depth k in the tree, then j is assigned rank “ k ” at node a and thus $M[j]^a = k$. For the sake of completeness, variables that have not been branched on are given the rank $n + 1$. We can extend the idea of rank to work on subsets of I^n . For any set $S \subset I^n$, $M^a(S) = \{M^a(i) | i \in S\}$. Key theorem from Ref. 29 is that if $M^a(F_1^a)$ is not lexicographically minimal in the set $\text{orb}(M^a(F_1^a), \mathcal{G})$, then for no solution feasible at node a is in the fundamental domain. Thus, node a can be pruned by isomorphism if $M^a(F_1^a)$ is not lexicographically minimal in the set $\text{orb}(M^a(F_1^a), \mathcal{G})$.

The goal of isomorphism pruning is to determine when a node contains *only* solutions that are not in the fundamental domain. Unlike orbitopal fixing, there has been no attempt to use symmetry to separate those solutions in the fundamental domain from those that are not. As a result, isomorphism pruning may prune nodes much later in the

tree as would have been possible had the constraints defining the fundamental domain been included in the original problem formulation. However, there has been some effort in exploiting symmetry in order fixing variables in addition to pruning nodes. A powerful idea, called *orbit setting*, is to combine the variable fixing with symmetry considerations in order to fix many additional variables. For instance, if a variable can be set to zero as a result of any logical implication, then all variables sharing the same orbit can also be fixed to zero.

Orbital Branching

Orbital branching, discussed in Refs 30 and 31, is an intuitive way of exploiting the symmetry group $\mathcal{G}(\text{ILP})$ during branching decisions. The classic 0–1 variable branching dichotomy does not take advantage of the problem information encoded in the symmetry group. The presence of symmetry provides ways of strengthening both branching disjunctions and variable fixing algorithms. Symmetry can be exploited by fixing variables via symmetry considerations either at the node or during branching. Orbital branching fixes variables during branching. By using symmetry considerations to strengthen branching decisions, the affects of branching on a given variable are better known at the time the branching decision is made and, thus, results in better branching decisions. Orbital branching can be used as a stand-alone symmetry-breaking tool or in conjunction with other methods such as isomorphism pruning.

Let $\mathcal{G}^a$ be the symmetry group of the subproblem represented by node a in the branch-and-bound tree. Let $O = \{i_1, i_2, \dots, i_{|O|}\} \subseteq N^a$ be an orbit of the symmetry group $\mathcal{G}^a$. Given a subproblem a , the disjunction

$$x_{i_1} = 1 \vee x_{i_2} = 1 \vee \dots \vee x_{i_{|O|}} = 1 \vee \sum_{i \in O} x_i = 0 \quad (20)$$

splits the feasible region. In what follows, it is shown that for any two variables x_j and x_k with $i, j \in O$, the two children $a(j)$ and $a(k)$ of a , obtained by fixing respectively x_j and x_k to 1, have the same optimal solution value. As a

consequence, disjunction (20) can be replaced by the binary disjunction:

$$x_h = 1 \vee \sum_{i \in O} x_i = 0, \quad (21)$$

where h is an arbitrary element in O . Note that the additional variable fixings in Equation (21) can be done by the orbit-setting algorithm in isomorphism pruning and equivalent algorithms in constraint programming literature. Orbital branching can be extended to allow for branching on general disjunctions [32].

In addition to exploiting the symmetry group at the root node, $\mathcal{G}$, orbital branching can recognize and exploit symmetry that enters the problem as a result of the branching decisions. This method is similar to that of Ref. 33 for constraint programming. The results of Ref. 30, however, indicate that the time spent finding these new symmetries outweigh their benefits.

REFERENCES

1. Bertsimas D, Tsitsiklis JN. Introduction to Linear Optimization (Athena Scientific Series in Optimization and Neural Computation, 6). Belmont (MA): Athena Scientific; 1997.
2. Grove LC, Benson CT. Finite reflection groups. New York: Springer; 1985.
3. Rotman JJ. An introduction to the theory of groups. 4th ed. New York: Springer; 1994.
4. Cameron PJ. Permutation groups. London: Mathematical Society; 1999.
5. Margot F. Symmetry in integer linear programming. Tepper Working Paper, E37, 2008.
6. Aloul FA, Ramani A, Markov IL, *et al.* Solving difficult instances of boolean satisfiability in the presence of symmetry. IEEE: IEEE Trans CAD 2003;22:1117–1137.
7. Ramani A, Markov IL. Automatically exploiting symmetries in constraint programming. Lect Notes Artif Intell 2005;3419:98–112.
8. Ramani A, Aloul FA, Markov IL, *et al.* Breaking instance-independent symmetries in exact graph coloring. J Artif Intell Res 2006;26:191–224.
9. Puget J-F. Automatic detection of variable and value symmetries. Volume 3709, 11th International Conference on Principles and Practice of Constraint Programming, Lecture Notes in Computer Science. New York: Springer; 2005. pp. 475–409.
10. Salvagnin D. A dominance procedure for integer programming [Master's thesis]. University of Padova; 2005.
11. Ostrowski J. Symmetry in integer programming [PhD thesis]. Lehigh University; 2009.
12. Luks EM. Permutation groups and polynomial-time computation. In: Finkelstein L, Kantor WM, editors. Volume 11, Groups and Computation, American Mathematical Society DIMACS Series. Providence, Rhode Island: AMS; (DIMACS, 1991). 1993. pp. 139–175.
13. McKay BD. Nauty User's Guide (Version 1.5). Canberra: Australian National University; 2002.
14. Darga P, Liffiton M, Sakallah K, *et al.* Exploiting structure in symmetry generation for CNF. In: Proceedings of the 41st Design Automation Conference. San Diego: 2004. pp. 530–534.
15. De Klerk E, Sotirov R. Exploiting group symmetry in semidefinite programming relaxations of the quadratic assignment problem. Optim Online 2007.
16. Friedman EJ. Fundamental domains for integer programs with symmetries. In: Dress AWM, Xu Y, Zhu B, editors. Combinatorial Optimization and Applications, 1st International Conference, COCOA 2007; 2007 Aug 14–16; Xi'an, China. Proceedings, volume 4616 of Lecture Notes in Computer Science. New York: Springer; 2007. pp. 146–153.
17. Gent IP, Harvey W, Kelsey TW. Groups and constraints: symmetry breaking during search. Principles and practice of constraint programming - CP 2002, Volume 2470 of Lecture Notes in Computer Science. New York: Springer Berlin/Heidelberg; 2002. pp. 233–240.
18. Meller RD, Narayanan V, Vance PH. Optimal facility layout design. Oper Res Lett 1998;23(3–5):117–127.
19. Sherali HD, Smith JC. Improving zero-one model representations via symmetry considerations. Manage Sci 2001;47(10):1396–1407.
20. Méndez-Díaz I, Zabala P. A polyhedral approach for graph coloring. Electron Notes Disc Math 2001;7:1–4.
21. Méndez-Díaz I, Zabala P. A branch-and-cut algorithm for graph coloring. Disc Appl Math 2006;154(5):826–847.

22. Kaibel V, Pfetsch ME. Packing and partitioning orbitopes. *Math Program* 2008;114: 1–36.
23. Kaibel V, Peinhardt M, Pfetsch ME. Orbitopal fixing. *IPCO 2007: The 12th Conference on Integer Programming and Combinatorial Optimization*. New York: Springer; 2007. pp. 74–88.
24. Kaibel V, Faenza Y, Loos A, *et al.* Tractable and intractable orbitopes. *Personal Correspondence*; 2009.
25. Berthold T, Pfetsch ME. Detecting orbitopal symmetries. In: Fleischmann B, Borgwardt K, Klein R, *et al.*, editors. *Operations Research Proceedings 2008*. New York: Springer; 2009. pp. 433–438.
26. Margot F. Pruning by isomorphism in branch-and-cut. *Math Program* 2002;94:71–90.
27. Margot F. Exploiting orbits in symmetric ILP. *Math Program Ser B* 2003;98:3–21.
28. Margot F. Symmetric ILP: coloring and small integers. *Disc Optim* 2007;4:40–62.
29. Margot F, Ostrowski J, Linderoth J. Flexible isomorphism pruning. *Working paper*. 2009.
30. Ostrowski J, Linderoth J, Rossi F, *et al.* Orbital branching. *IPCO 2007: The 12th Conference on Integer Programming and Combinatorial Optimization*, Volume 4517 of *Lecture Notes in Computer Science*. New York: Springer; 2007. pp. 104–118.
31. Ostrowski J, Linderoth J, Rossi F, *et al.* Orbital branching. *Math Program*. In press.
32. Ostrowski J, Linderoth J, Rossi F, *et al.* Constraint orbital branching. In: *IPCO 2008: The 13th Conference on Integer Programming and Combinatorial Optimization*. Bertinoro, Italy; 2008.
33. Meseguer P, Torras C. Exploiting symmetries within constraint satisfaction search. *Artif Intell* 2001;129(1–2):133–163.

SYSTEM AVAILABILITY

DONALD P. GAVER
PATRICIA A. JACOBS
Department of Operations Research,
Naval Postgraduate School,
Monterey, California

OVERVIEW

All mechanical, electrical, nuclear power generation, propulsion, weapons systems, computer and data storage systems (both hard and software), communications, biological/medical, and many combinations thereof that are designed, manufactured and tested, operated and maintained using human skills, engineering, and scientific knowledge, *are subject to failure*. This means that at some point of time, a system becomes unable to perform the intended design function, or mission. *Reliability* is traditionally a probabilistic measure of the propensity of component, subsystem, or system (of systems) to satisfactorily perform a required designed performance function on, say, a civilian (consumer or government), or a military, or homeland security mission. This includes effectively and promptly responding to a natural disaster such as a hurricane, tornado, or earthquake. During the period of inability to perform, the item is said to be *not available* or *unavailable*. Often this propensity is quantified as a *conditional probability*, where the conditions include stressful aspects of the operational environment and usage. Note that it is not customary to include *vulnerability* to opponent action in measures of military or security system reliability, but the prevention and remediation responses to deliberate failure causation and those of failures from natural, including operator, causes are similar. In homeland security, vulnerability can be forestalled by instrumentation and personnel screening [1]. Many, even most, failed systems are subject to failure correction

or rectification, are *repairable*, which often requires a nonnegligible total *down time*, during which time interval they are completely or partially unavailable. Thus, an initial informal definition of *availability* can be an estimated probability that, at the moment of demand/need for a system's performance, it will furnish that service, and not be waiting for or being rectified/repaired/replaced. Field support of system availability and capability to accomplish a mission profile depends upon *sustainment resources*—personnel, diagnostic instrumentation, logistics/spare parts, tow vehicles, and so on. These must also be available when needed, and themselves be reliable.

It is often true that the failure propensity measure, or *hazard*, or *hazard rate*, of a component, subsystem, or system, $h(x)$, eventually increases with time or intensity of usage and “wear” or “age”, x , after being fielded. Hazard (rate) is defined as the conditional probability of failure of a system, given survival to operating age x , and conditional on past and present environmental conditions; biometric survival analysis is relevant here [2]. Typically, a new or modified/upgraded system is tested so as to identify and remove or reduce the effect of *design defects* or *failure modes*; during the early and operational testing period every attempt is made to induce the activation, thus exposure of, *design defects* or *failure modes*, that is, to *stimulate* occurrences of failures and remove their causes [3]. Some faults inevitably remain, and can be found later, desirably during routine scheduled maintenance. However, such maintenance can be unnecessary (time-scheduled maintenance is costly in time and resources, and may well find nothing, so wastefully decreases system availability and increases life cycle cost). It is currently thought to be desirable, and increasingly technically feasible, to identify and automatically track *system health* changes—failure-preceding or prognostic events (excess heat, undue vibration, wear of, say, tire treads, tires, engine oil condition,

etc.) that can signal an incipient failure and thereby suggest needed system design or operational changes before total, possibly catastrophic, failure can occur; periodic lubrication of civilian motor vehicles is a case in point. Annual physical checkups for humans is another. A specific potential payoff area is that of using aerothermodynamic signatures to predict fouling in gas turbines via sensing change in mass flow and adiabatic efficiency. Reduction of power output, and increase in fuel consumption can lead to degradation of compressor performance because of blade fouling and consequent roughness, and hence of gas turbine capacity [4]. The procedure for mitigating (“washing blades”) such conditions is known as *condition-based maintenance (CBM)*, and makes use of currently available automated embedded sensors, logistic assets, and training to conduct optimized CBM; hence CBM+. However, a CBM+ subsystem is itself subject to malfunction and failure, so false positive and false negative diagnoses must be anticipated and made rare in order for overall gains to be made. Thus the reliability/availability of CBM+ must itself be considered as part of the total system package.

Accelerated testing procedures are often used at an early (developmental) testing stage [5]. During the testing phases, both engineering and operational, it is intended, likely, and desirable to *find design faults and “fix” them*, so that they will not activate without adequate warning and cause mission-abortive events under active usage conditions. It is the usual objective to test and “fix” until time between failures during testing become a trend-free stable-in-time random process meeting specified criteria, often represented by evidence of a temporally stationary *time-between-successive-failures distribution*, with the further assumption that the time between failures are statistically independent (*iid random variables* = IIDRV, i.e., a renewal process [6]); see Gaver *et al.* [3], and several computer packages, such as Reliasoft (www.reliasoft.com). Frequently and without much validation, times to failure (measured from installation or repair termination to next failure) are assumed exponentially distributed, with the

mean value fixed—the so-called mean time between failures. Termed the “*exponential assumption*,” this is computationally convenient but worthy of critical scrutiny if possible. Appeal to lessons from analogous systems in similar usage environments is advisable.

Initial tests of a newly designed system typically reveal “infant failures” that should be removed before any approximate stability of times-between-failures occurs. The consequence of such removal is referred to as *reliability growth*, and it has been modeled and analyzed extensively, for example, by Crow [7,8], for example, in the software package Reliasoft; see Gaver *et al.* [3] for discussion of testing for growth using a sequential (run of successes) stopping rule; such testing is made more complex if the system performance is in sequential stages; failure of an early stage may mask the exposure of later stages, thus requiring longer tests so as to explore all stages for faults. After (*if*) there is evidence of failure time stochastic stability it is meaningful to speak of the *mean (operating) time between failures*; this is often specified as a reliability metric, but since systems are used to perform a task or mission, an estimate of the probability of mission success is usually more relevant, certainly if the item is a one-shot device or is usually quiescent in performance: a piece of ammunition or a communication system for emergencies. This can degrade or fail out right during idle periods, necessitating periodic inspections. Caution is required; justification for such a stable failure-time distribution must depend on actual data monitoring, and the currently appropriate particular distributional form (model) will depend on environmental conditions, including maintenance skill, and usage and environment; for example, operating times between successive failures for a particular equipment (surface motor vehicle, aircraft or helicopter-civilian or military) in the Arctic can be expected to differ from those operating under tropical, or hot, sandy, and dusty desert conditions.

Upon failure, some (but not all, e.g., expendable missiles and ammunition) system components are repaired; this requires

diagnostic and maintenance facilities and personnel, and logistics (spare parts); the system may then be again capable of performing designed-for functions. Repair or replacement is intended to restore original capability, or possibly upgrade (or in fact actually occasionally and unintentionally *downgrade*) system function. Preventive maintenance (PM), based on the observed or inferred/sensed physical condition of the system, may be used to ward off catastrophic failures or to halt prolonged degradation. As stated earlier, current attention is being focused on automated information-system-supported CBM, or *CBM+* (alternatively, integrated vehicle health management, *IVHM*). Such subsystems continuously monitor the “health condition” of various subsystems and provide warning of imminent malfunction/failure. If successful, such a capability should allow more operative mission hours, and shorter and fewer time-consuming and costly repairs. Success depends upon the trustworthiness of the *CBM+/IVHM* systems available [9–11]. Availability is enhanced under well-tuned *CBM+/IVHM*, since times between failures are lengthened and maintenance periods shortened.

This article considers that times or events between such system failures (during functional usage) are punctuated by periods of “down-time”, during which the system (copy or version thereof) is (i) awaiting maintenance, which may include moving, or being moved, to a repair facility, and awaiting the arrival of parts, instrumentation, and personnel; (ii) undergoing diagnosis of faults and subsequent maintenance actions; or (iii) awaiting further operational assignment (in idle standby); items may fail or degrade during such waiting/idle times, and so preliminary tests may aptly be performed prior to mission assignment: premission testing is often conducted, especially with aircraft and other platforms. This permits the detection of imperfect repair before the mission actually is in progress.

Availability issues also arise in the operation of modern networked computer systems [12]. Among the causes of unavailability of computer utilities such as *clouds* are service

interruption from hostile action: hacking and jamming. Service degradation by necessary security actions is also a performance or service inhibitor.

SPECIFIC DEFINITIONS

The (classical) system *operational availability* (A_0) *parameter* is a measure of the capability of a system to *begin* to perform a designed-for function or mission *on immediate “random” demand*. Nominally, this would imply that the system is in idle state (iii) above, but too long a sojourn of idleness (in “cold standby”) in a hostile environment can induce *startup unavailability which adds to down time*. *Mission availability*, measuring the capability of a system to perform its function from demand for service until mission termination is usually of far more operational significance than simple A_0 . The latter is too often treated—inappropriately—as “the long-run fraction of time” (or “constant probability”) that a component or system is “up” or completely mission capable at the time of demand for operation. In reality, such a constant value may not prevail; “the” value of A_0 may well change with time and usage, including the effect of maintenance variation between individual copies of a system design. Further, a system’s availability may be partial: for example, a multi-engine vehicle can still partially function if degraded by the loss of one out of several engines. However, loss of an engine or, one of several sensors, or a communication link sometimes allows a modified or limited version of the intended mission to proceed until a replacement is furnished. Consequently, availability need not be a binary instantaneous concept.

The object of this article is to describe methods in use for describing, measuring, testing, and improving *availability* and to provide cautionary comments similar to those above.

COMPONENT AND SYSTEM OPERATIONAL CLASSIFICATION

Engineered systems of all types are composed of numbers of components (“parts”),

subsystems, and entire systems designed to perform cooperative functions. Modern warfare has been said to be conducted by “systems of systems” (SoS). Health care systems and road and air traffic systems have the same SoS characteristics. Such systems operate efficiently and effectively according to correct design, manufacture, and usage especially if usefully guided by carefully conducting modeling, simulation, and limited tests. They are likely to fail or degrade with age and stress; some failure modes are mission and even life-threatening, and can occur with little warning. Hence condition-based monitoring/maintenance (CBM) and PM are required and invoked; time spent in such makes the entity temporarily mission unavailable, but is intended to provide greater overall net system mission availability.

Systems, and especially subsystems and components, can be roughly categorized as follows:

- *one-shot/disposable* (e.g., fuel, ammunition, missiles or mines, electronic components such as computer chips or screens, remote vulnerable sensors, etc.); *and/or*
- *repairable* (segmentally)—*replaceable* (e.g., failure-prone engines and control devices, vehicle chassis, computer hardware, elements of communication networks, etc.); an injured or wounded human can sometimes be included in this category, as being part of an entire system.

In many cases, a failed repairable system becomes less repairable as the system ages, the *particular item's* availability terminates, and it is replaced; the latter replacement can be a redesigned upgrade [3]. Alternatively, a downed military aircraft can become a “hanger queen”, and is cannibalized for spare parts; a desperate practice that is discouraged. Further, the pattern of component/system usage can vary: many items such as some sensors and alarm systems, and engines and generators are normally running or “hot” during a mission, so if failure occurs or becomes imminent, symptoms are quickly evident. Others are normally inactive

or “cold”, such as an idle aircraft parked on a flight deck, or a rescue vehicle.

Many systems are made up of complex assemblies or combinations of one-shot/disposable *and* repairable/replaceable subsystems: a vehicle or platform burns fuel and may fire ammunition, both disposable, but its chassis, propulsion, and steering and sensors, and communications are very often repairable/replaceable. Many vehicles operate in groups: convoys of trucks or small boats, task groups, aircraft squadrons, and so on; here the mission may involve all elements of the group, so the timely availability of such a subforce at a particular site can be spoiled if just one or a few of the member elements fails—the entire system may become mission-incapable, or unavailable if one or more of such elements fails, or is damaged by opponent action (when in military or homeland protection application), by environmental conditions (e.g., sandstorms), or by owner/user mishandling. The tendency for such to occur is enhanced when hostilities occur, and when the sustainment/support forces and logistics are overwhelmed and themselves unavailable.

MODELING AVAILABILITY

Modeling the availability of a system or an SoS is carried out to forecast or predict that system's condition to perform its mission. System availability has been characterized probabilistically or stochastically for many years; an initial classic is Barlow and Proschan [13], but aspects of repairable system availability go far back to Erlang, Khinchine, and to Palm in the 1930s and a further few decades, well summarized in Feller [14] as the “repairman problem” or a service system. See also Cox and Oakes [2] on survival analysis, essentially summarizing biological (e.g., animal, human) “reliability”. Failure-prone but repairable SoS resemble biological epidemics: failures of subsystems can stress, or infect, other subsystems that in turn can cause total system collapse if not quickly isolated and mitigated.

The simplest analytical model for a failure-prone but repairable system is the alternating renewal process (see, however,

practical cautions in the section titled “Overview”) [6,15]. Letting a sequence of up-times, $\{\mathbf{U}_i\}$, and a sequence of down-times, $\{\mathbf{D}_i\}$, be independent sequences of independent random variables with marginal distributions $F_U(u; \underline{\theta})$ and $G_D(v; \underline{\varphi})$, respectively, and $\underline{\theta}$ and $\underline{\varphi}$ representing parameters; then if $A(t)$ is the availability at time t , given that the subsystem starts at the beginning of an up period, $\mathbf{U}_1$, say, we have the backward recurrence renewal equation

$$A(t) = P\{\mathbf{U}_1 > t\} + \int_0^t P\{\mathbf{U}_1 + \mathbf{D}_1 \in dx\} A(t-x) \quad (1)$$

which can be solved numerically or in terms of Laplace–Stieltjes transforms. Basic renewal theory asymptotics [15, Chapter 11], shows that if $E[\mathbf{U}]$ and $E[\mathbf{D}]$ exist/are finite then

$$\lim_{t \rightarrow \infty} A(t) = E[\mathbf{U}] / (E[\mathbf{U}] + E[\mathbf{D}]) = A_o \quad (2)$$

the popular (often misused) operational availability. In cases in which the system must function throughout a mission of length M , a more informative measure is mission availability

$$A_m = A_o \frac{P\{\mathbf{U} > M\}}{E[\mathbf{U}]},$$

an asymptotic renewal theory result [15].

Complex systems, starting with series/tandem configurations, and extending to series/parallel/redundant combinations can be mathematically analyzed under the initially stated independence assumptions, and with these progressively relaxed.

STATISTICAL INFERENCE ON AVAILABILITY

If the simple assumption of mutually independent sequences of independent identically distributed (iid) up-time r.v.s $\{\mathbf{U}_i\}$, and likewise iid down-times $\{\mathbf{D}_i\}$, is (provisionally) acceptable, then alternating renewal theory shows that the long-run *point availability* is given by $A_o = E[\mathbf{U}] / (E[\mathbf{U}] + E[\mathbf{D}])$, provided expectations exist. A natural estimate of $A_o, \hat{A}_o = \bar{u} / (\bar{u} + \bar{d})$, where $\bar{u}$ and $\bar{d}$ are the

sample averages of up and down realizations; assuming no censoring [16] assesses confidence intervals for A_o using *jackknifing*, wherein the Logit transform of $\hat{A}_o$ is recomputed, omitting pairs of observations successively. The method is applied with good success to various redundant systems and several plausible distributional forms. An alternative would be the *bootstrap* [17]. A sensible (semi) nonparametric approach is the nonparametric bootstrap: resample from the empirical distributions of up- and down-times, provided that preliminary examination of the data does not wildly contradict iid assumptions [18]. In many applied situations, data may be insufficient to validate (or invalidate) such assumptions at all conclusively. System aging may decrease the mean time between failures and increase the mean system down-time. It may well be wise to adaptively smooth up- and down-time means, and employ these in the $\hat{A}_o$ formula or, often more appropriately, in $\hat{A}_m$ formula. Assumption checking is often carried out using engineering judgment, and should continue throughout system or SoS lifetime.

CONCLUSIONS

Availability quantitatively defines a system’s capability to respond to demand for its service. It combines reliability, the uninterrupted up-time to failure, and repair or down-time, during which the system cannot perform its mission. Simple parametric definitions of operational availability, $\hat{A}_o$, and mission availability, A_m , are given, and probabilistically modeled, and statistical inference procedures given.

Many more results are available from the references. For a general high-level overview, see Davis [19].

REFERENCES

1. Rios Insua D, Rios J, Banks D. Adversarial risk analysis. *J Am Stat Assoc* 2009;104: 841–854.
2. Cox DR, Oakes D. Analysis of survival data. London: Chapman and Hall; 1984.

3. Gaver DP, Jacobs PA, Glazebrook KD, *et al.* Probability models for sequential-stage system reliability growth via failure mode removal. *Int J Reliab Qual Saf Eng* 2003;10:15–40.
4. Millsaps KT, Baker J, Patterson JS. Detection and localization of fouling in a gas turbine compressor from aerothermodynamic measurements. *Proceedings ASME Turbo Expo 2004: Power for Land, Sea, and Air*; 2004 Jun 14–19; Vienna, Austria. Norcross, Georgia: ASME; 2004. Paper No. GT2004-54173. pp. 1867–1976.
5. Meeker WQ, Escobar LA. *Statistical methods for reliability data*. New York: John Wiley and Sons; 1998.
6. Cox DR. *Renewal theory*. London: Methuen & Co; 1970.
7. Crow LH. Reliability analysis for complex repairable systems. In: Proschan F, Serfling RJ, editors. *Reliability and biometry*. Philadelphia: SIAM; 1974. pp. 379–410.
8. Crow LH. A method for achieving an enhanced mission capability. *2002 Proceedings of Annual Reliability and Maintainability Symposium*. Piscataway (NJ): IEEE; 2002. pp. 153–157.
9. Macheret Y, Koehn P. Prognostics and advanced diagnostics for improving steady-state and pulse reliability. *Proceedings 2006 IEEE Aerospace Conference*. Piscataway (NJ): IEEE; 2006. pp. 1–11.
10. Williams Z. Benefits of IVHM: an analytical approach. *Proceedings 2006 IEEE Aerospace Conference*. Piscataway (NJ): IEEE; 2006. pp. 1–8.
11. Osborne B. Leveraging remote diagnostics data for predictive maintenance. In: Wilson AG, Limnios N, Keller-McNulty S, *et al.*, editors. *Modern statistical and mathematical methods in reliability*. Singapore: World Scientific Publishing Co.; 2005. pp. 353–373.
12. Gelenbe E, Finkel D, Tripathi SK. On the availability of a distributed computer system with failing components. *1985 Proc. ACM SIGMETRICS*. New York: ACM; 1985. pp. 6–13.
13. Barlow RE, Proschan F. *Mathematical theory of reliability*. New York: Wiley; 1967.
14. Feller W. Volume 1, *An introduction to probability theory and its applications*. 3rd ed. New York, NY: Wiley; 1968.
15. Feller W. Volume 2, *An introduction to probability theory and its applications*. New York: Wiley; 1966.
16. Gaver DP, Chu BB. Jackknife estimates of component and system availability. *Technometrics* 1979;21:443–450.
17. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall; 1993.
18. Gaver DP, Jacobs PA. System availability: time dependence and statistical inference by (semi) non-parametric methods. *Appl Stoch Model Data Anal* 1989;5:357–375.
19. Davis TP. Science, engineering, and statistics. *Appl Stoch Model Bus Ind* 2006;22:401–430.

SYSTEMS IN SERIES

MARCUS AGUSTIN
Department of Mathematics and
Statistics, Southern Illinois
University Edwardsville,
Edwardsville, Illinois

SYSTEMS IN PRACTICE

A number of applications in fields such as economics, engineering, higher education, and medical sciences consider units or components that are connected in some structure. The manner in which the components are connected as well as how each component is designed to function will determine if the system will function successfully or not. For example, when a package is delivered from point A to point B, it will typically go through different loading points in a systematic order. These loading points are the components in the system. Observing a failure at a loading point may result in delay in delivery of the package. Another example was considered in Ref. 1 regarding the different pathways that students pursuing a two-year degree can take. Each pathway is considered as a component and a student (in this case, the system) can take one and only one pathway. Ref. 2 assumed a committee with n members collectively making a decision on the acceptance of a project. Acceptance of a project depends on having k members agreeing to accept the project, where $k = 1, \dots, n$. The higher the value of k , the stricter the acceptance rule becomes. In the medical field, the survival of an individual depends on how organs are connected. The failure of a critical organ may result in the death of an individual.

A system that is commonly used in operations research and management science is called a *series system*. In a series system, components are connected in such a way that the failure of a single component results in system failure. From a different point of view,

a series system will function if and only if all components are functioning. A series system is also referred to as a *competing risks system*, since the failure of an individual (system) can be classified as one of n possible risks (component) that contribute to the individual's failure. An example of a series system is illustrated in Refs 3 and 4, where a series system of softwares is considered with the failure of a software (component) resulting in the failure of the system. Ref. 5 provides analysis of times to failures in the presence of competing risks with application to bone marrow transplantation for leukemia patients. Ref. 6 can be treated as an example of a series system, where an individual with a loan is the system of interest and the distinct reasons for complete loan repayment ahead of maturity date are the components connected in series. Finally, Bier *et al.* [7] and Hausken [8], among others, considered the optimal defensive strategy for a series system based on a trade-off between cost and security.

One of the goals in reliability is to ensure that at any given instant, we observe a functioning system with a high probability. The probability that the system is functioning is defined as the system reliability. In finding the system reliability, one needs to obtain the reliability of each individual component and take into account the manner in which the components are connected. In this article, we consider a series system and our main interest is to find the system reliability. In particular, we will focus on the situation where the reliability of each component is a function of the time to component failure.

SYSTEM RELIABILITY

Structure Function of a Series System

Let us consider a system of n components, where the state of a component can be classified as "functioning" or "failed." We can define a binary random variable X_i , for $i = 1, \dots, n$, as

$$X_i = \begin{cases} 1, & \text{if component } i \text{ is functioning,} \\ 0, & \text{if component } i \text{ has failed.} \end{cases}$$

The random variables $X_1, \dots, X_n$ are assumed to be mutually independent, that is, the failure of a component does not affect the probability of failure of any other component. We define the 0-1 vector $\mathbf{X} = (X_1, \dots, X_n)$ to be the state vector of a system. For any two vectors $\mathbf{X}$ and $\mathbf{Y}$, we define vector inequality to be $\mathbf{X} \leq \mathbf{Y}$ whenever $X_i \leq Y_i$ for all $i = 1, \dots, n$. In addition, $\mathbf{X} < \mathbf{Y}$, if $X_j < Y_j$ for at least one j and $X_k \leq Y_k$ for the remaining k components.

The state of a system, referred to as the *structure function* of the system, is a function that depends on the manner in which the individual components are connected and the state of each component. A system's structure function is denoted by $\phi(\mathbf{X})$ and takes a value 0 or 1, where

$$\phi(\mathbf{X}) = \begin{cases} 1, & \text{if the system is functioning,} \\ 0, & \text{if the system has failed.} \end{cases}$$

The structure function is said to be a non-decreasing function whenever $\mathbf{X} \leq \mathbf{Y}$ implies that $\phi(\mathbf{X}) \leq \phi(\mathbf{Y})$.

A primary interest in reliability is to consider systems where the action of replacing a failed component by a new component will not result in the system being "worse off." Moreover, it is of interest to examine a system where components are all deemed to be relevant to the system function. One can define a component i to be irrelevant if for all values of X_j , where $j = 1, \dots, n$ and $j \neq i$,

$$\begin{aligned} &\phi(X_1, \dots, X_{i-1}, 0, X_{i+1}, \dots, X_n) \\ &= \phi(X_1, \dots, X_{i-1}, 1, X_{i+1}, \dots, X_n). \end{aligned}$$

In other words, changing the status of an irrelevant component from failed to functioning will not have any effect on the status of the system. Systems that do not contain irrelevant components and where the structure function is nondecreasing are called *coherent systems*. For any coherent system with n components, its structure function can be bounded by

$$\prod_{i=1}^n X_i \leq \phi(\mathbf{X}) \leq 1 - \prod_{i=1}^n (1 - X_i). \quad (1)$$

A more detailed discussion of coherent systems is given in **Coherent Systems**. For an in-depth treatment of coherent systems, the reader is referred to the monographs by Barlow and Proschan [9] and Gnedenko *et al.* [10].

A series system is an example of a coherent system. It is defined to be a system that functions if and only if all components are functioning. From the point of view of a failed component, a series system fails as soon as one component has failed. Thus, the structure function of a series system is

$$\phi(\mathbf{X}) = X_1 X_2 \cdots X_n = \prod_{i=1}^n X_i. \quad (2)$$

Note that Equation (2) is the lower bound of the structure function of any coherent system given in Equation (1). On the other hand, the upper bound in Equation (1) is the structure function of a parallel system (see **Parallel Configurations**). Since the entries of $\mathbf{X}$ are all binary, the structure function of a series system can be written as $\phi(\mathbf{X}) = \min(X_1, \dots, X_n)$. Figure 1 shows a series system with three components.

As an illustration, consider a series system with $n = 4$ components where component 1 has failed, while the remaining $n - 1 = 3$ components are functioning. Thus, the state vector is $\mathbf{X} = (X_1, X_2, X_3, X_4) = (0, 1, 1, 1)$, which results in $\phi(\mathbf{X}) = \min(0, 1, 1, 1) = 0$. On the other hand, if all components are functioning, then $\phi(\mathbf{X}) = \min(1, 1, 1, 1) = 1$. Clearly, $\phi(\mathbf{X}) = 0$ whenever $X_i = 0$ for some i . To show that the series system is a coherent system, we need to consider two state vectors $\mathbf{X}$ and $\mathbf{Y}$ such that $X_i = 0$ for some $i = 1, \dots, 4$, while $Y_j = 1$ for all $j = 1, \dots, 4$. In other words, $\mathbf{X} < \mathbf{Y}$. From Equation (2), it follows that $\phi(\mathbf{X}) \leq \phi(\mathbf{Y})$, which means that ϕ is a nondecreasing function. In addition, it is easily seen that a series system with state vector $\mathbf{Y} = (1, 1, 1, 1)$ will fail as soon as one

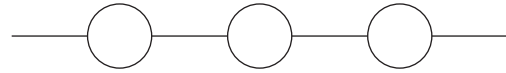


Figure 1. Diagram for a series system with three components.

component fails. Hence, each component in a series system is relevant. Thus, a series system is coherent.

Reliability of a Series System

Suppose a series system is observed up to an instant of time t . Let $X_i(t)$ be the state of component i at time t and $\mathbf{X}(t) = (X_1(t), \dots, X_n(t))$ is the resulting state vector. Let $p_i(t) = P\{X_i(t) = 1\}$ be the probability that the i th component is functioning at time t and suppose $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$. We define

$$h(\mathbf{p}(t)) = E[\phi(\mathbf{X}(t))] = P\{\phi(\mathbf{X}(t)) = 1\}$$

to be the reliability of the system at time t .

From Equation (2) and the fact that $X_i(t)$ is a 0-1 variable, the structure function of a series system can be expressed as

$$\phi(\mathbf{X}(t)) = \begin{cases} 1, & \text{if } X_i(t) = 1 \text{ for } i = 1, \dots, n, \\ 0, & \text{if at least one } X_j(t) = 0. \end{cases}$$

Since $X_1(t), \dots, X_n(t)$ are independent random variables, the reliability of a series system is

$$\begin{aligned} h(\mathbf{p}(t)) &= P\{X_1(t) = 1, \dots, X_n(t) = 1\} \\ &= P\{X_1(t) = 1\}, \dots, P\{X_n(t) = 1\} \\ &= \prod_{i=1}^n p_i(t). \end{aligned} \quad (3)$$

TIME-DEPENDENT SYSTEM RELIABILITY

Survival Function for a Series System

We will now consider a series system with n components, where the failure of a component depends on how long the component has been in use. Let T_i be the time to failure of component i and we assume that T_i can be modeled via the probability density function $f_i(t; \theta)$, parameterized by θ . We will assume that $T_1, \dots, T_n$ are independent random variables. From Ref. 11, the reliability of component i is defined to be the probability that the component survives past time t . This is also referred to as the *survival function* and is expressed as

$$S_i(t; \theta) = P\{T_i > t\} = \int_t^\infty f_i(u; \theta) du. \quad (4)$$

When dealing with failure of components, an important function to consider is the hazard function $h_i(t; \theta)$, which can be interpreted as the amount of risk of a component. In terms of the probability density function and the survival function, the hazard function can be written as

$$\begin{aligned} h_i(t; \theta) &= \lim_{\Delta t \rightarrow 0} \left[\frac{S_i(t; \theta) - S_i(t + \Delta t; \theta)}{\Delta t} \right] \\ &\times \frac{1}{S_i(t; \theta)} = -\frac{S'_i(t; \theta)}{S_i(t; \theta)} = \frac{f_i(t; \theta)}{S_i(t; \theta)}. \end{aligned} \quad (5)$$

By noting that $\frac{d}{dt} S_i(t; \theta) = -f_i(t; \theta)$, the hazard function and the survival function are related by $h_i(t; \theta) = -\frac{d}{dt} \log S_i(t; \theta)$.

Let Y be the random variable associated with the system lifetime and $S(y; \theta)$ be the system reliability. The corresponding density function, cumulative distribution function, and hazard function of the system lifetime are $f(y; \theta)$, $F(y; \theta)$, and $h(y; \theta)$, respectively. In terms of the density function and cumulative distribution function, the hazard function is

$$h(y; \theta) = \frac{f(y; \theta)}{S(y; \theta)}. \quad (6)$$

From Equation (2), the system lifetime for a series system can be expressed as $Y = \min(T_1, \dots, T_n)$. Using Equation (4), the system reliability is equal to

$$\begin{aligned} S(y; \theta) &= P\{Y > y\} = P\{\min(T_1, \dots, T_n) > y\} \\ &= P\{T_1 > y, \dots, T_n > y\} = \prod_{i=1}^n P\{T_i > y\} \\ &= \prod_{i=1}^n S_i(y; \theta). \end{aligned} \quad (7)$$

It follows that the density function of the system lifetime of a series system is

$$f(y; \theta) = -\sum_{i=1}^n S'_i(y; \theta) \left[\prod_{\substack{j=1 \\ j \neq i}}^n S_j(y; \theta) \right].$$

The hazard function is obtained by substituting the appropriate expressions for $f(y; \theta)$ and $S(y; \theta)$ into Equation (6).

Let us consider the special case when the component lifetimes are also identically distributed. In other words, the density function and survival function of the i th component are $f_*(t; \theta)$ and $S_*(t; \theta)$, respectively. From Equation (7), the system reliability simplifies to

$$S(y; \theta) = [S_*(y; \theta)]^n$$

and the corresponding density function is

$$f(y; \theta) = n f_*(y; \theta) [S_*(y; \theta)]^{n-1}.$$

Exponential Lifetimes

We will now consider the case when the individual component lifetimes are assumed to be exponentially distributed. Exponential lifetimes for a software were considered in Ref. 12, where the mean lifetime depends on the residual number of defects remaining in the software as well as the failure rate contributions of each software. Exponential lifetimes were also considered, among others, by Basu *et al.* [13] in the area of market research; Osborne and Severini [14] in goodness-of-fit testing; and Özekici and Soyer [15] and Agustin and Agustin [16] in the area of software reliability.

Let T_i be the lifetime of component i that is assumed to be exponentially distributed with density function

$$f_i(t; \theta_i) = \theta_i \exp\{-\theta_i t\}, t > 0.$$

Thus, the expected lifetime of component i is $\mathbf{E}(T_i) = 1/\theta_i$. Using Equations (4) and (5), the survival function and hazard function of T_i are

$$S_i(t; \theta_i) = \exp\{-\theta_i t\}, \quad t > 0 \quad \text{and}$$

$$h_i(t; \theta_i) = \theta_i, t > 0,$$

respectively. Thus, the hazard function of the component lifetimes is constant.

Let $Y = \min(T_1, \dots, T_n)$ denote the series system lifetime and let $\theta = (\theta_1, \dots, \theta_n)$ represent the parameter vector of interest. It follows from Equation (7) that

$$S(y; \theta) = \prod_{i=1}^n \exp\{-\theta_i y\} = \exp\left\{-y \left(\sum_{i=1}^n \theta_i\right)\right\}.$$

Moreover, the density function of Y is

$$\begin{aligned} f(y; \theta) &= \sum_{i=1}^n \theta_i \exp\{-\theta_i y\} \left[\prod_{\substack{j=1 \\ j \neq i}}^n \exp\{-\theta_j y\} \right] \\ &= \left[\sum_{i=1}^n \theta_i \right] \left[\exp\left\{-y \left(\sum_{i=1}^n \theta_i\right)\right\} \right]. \end{aligned}$$

Consequently, we note from Equation (6) that

$$h(y; \theta) = \sum_{i=1}^n \theta_i.$$

One can easily see that the system lifetime Y is exponentially distributed with mean $\mathbf{E}(Y) = [\sum_{i=1}^n \theta_i]^{-1}$.

When the component lifetimes are also identically distributed, that is, $\theta_i = \theta$ for $i = 1, \dots, n$, then the above functions take simplified forms. The system reliability and density function of the series system lifetime become

$$S(y; \theta) = \exp\{-y(n\theta)\} \quad \text{and}$$

$$f(y; \theta) = (n\theta) \exp\{-y(n\theta)\},$$

respectively, which results in the hazard function

$$h(y; \theta) = n\theta.$$

Hence, the series system lifetime is exponentially distributed with mean $\mathbf{E}(Y) = (n\theta)^{-1}$.

Pareto Lifetimes

We now consider the scenario where the component lifetimes do not have constant hazards. In particular, we will consider a family of monotone hazard functions of the form

$$h_i(t; \alpha_i, \lambda_i, \phi_i) = \alpha_i + \frac{\lambda_i}{t + \phi_i}, \quad (8)$$

with $\alpha_i > 0$, $\phi_i > 0$, and $\lambda_i \geq -\alpha_i \phi_i$. This family of hazard functions, known as a *Pareto family*, was considered in Ref. 17 when modeling the situation, where the failure of an

individual due to a certain disease depends on events that transpired in the individual's life, such as surgery. A similar form of the hazard function in Equation (8) for examining the failure of a software was also used in Ref. 18. This work was extended in Ref. 19 by assuming that the lifetime of softwares connected in series has hazard function similar to Equation (8), with the additional assumption that the value of one of the parameters depends on the number of system failures observed.

From Equation (8) and using the fact that $S_i(t; \theta) = \exp \left\{ - \int_0^t h_i(u; \theta) du \right\}$, the survival function of the i th component is

$$S_i(t; \alpha_i, \lambda_i, \phi_i) = \exp\{-\alpha_i t\} \left(1 + \frac{t}{\phi_i}\right)^{-\lambda_i}.$$

Note that if $\lambda_i = 0$ for $i = 1, \dots, n$, then the component lifetimes become exponentially distributed. Thus, the case considered in section titled "Exponential Lifetimes" becomes a special case of Pareto lifetimes.

Let $\theta = (\alpha_1, \lambda_1, \phi_1, \dots, \alpha_n, \lambda_n, \phi_n)$ denote the parameter vector and $Y = \min(T_1, \dots, T_n)$ represent the lifetime of a series system with n components. Using Equation (7), we can find the survival function of Y to be

$$S(y; \theta) = \left[\exp \left\{ - \left(\sum_{i=1}^n \alpha_i \right) y \right\} \right] \times \left[\prod_{i=1}^n \left(1 + \frac{y}{\phi_i} \right)^{-\lambda_i} \right].$$

As a special case, if $\phi_i = \phi$ for $i = 1, \dots, n$, the survival function of Y becomes

$$S(y; \theta) = \left[\exp \left\{ - \left(\sum_{i=1}^n \alpha_i \right) y \right\} \right] \times \left[\left(1 + \frac{y}{\phi} \right)^{-\sum_{i=1}^n \lambda_i} \right],$$

which implies that the system lifetime is Pareto with parameters $(\sum_{i=1}^n \alpha_i, \sum_{i=1}^n \lambda_i, \phi)$. In addition, if $\alpha_i = \alpha$ and $\lambda_i = \lambda$ for $i = 1, \dots, n$, that is, the component lifetimes are independent and identically distributed, then

$$S(y; \theta) = \exp\{-(n\alpha)y\} \left(1 + \frac{y}{\phi}\right)^{-n\lambda},$$

which implies that Y is Pareto with parameters $(n\alpha, n\lambda, \phi)$.

REFERENCES

1. Scott MA, Kennedy BB. Pitfalls in pathways: Some perspectives on competing risks event history analysis in education research. *J Educ Behav Stat* 2005;30:413–442.
2. Sah RK. Fallibility in human organizations and political systems. *J Econ Perspect* 1991;5:67–88.
3. Agustin MA, Pena EA. A dynamic competing risks model. *Probab Eng Inf Sci* 1999;13:333–358.
4. Agustin MA. Asymptotic and finite-sample properties of a reliability estimator for a competing risks system. *J Stat Comput Simul* 1999;64:221–234.
5. Prentice RL, Kalbfleisch JD, Peterson AV, *et al.* The analysis of failure times in the presence of competing risks. *Biometrics* 1978;34:541–554.
6. Banasik J, Crook JN, Thomas LC. Not if but when will borrowers default. *J Oper Res Soc* 1999;50:1185–1190.
7. Bier VM, Nagaraj A, Abhichandani V. Protection of simple series and parallel systems with components of different values. *Reliab Eng Syst Saf* 2005;87:315–323.
8. Hausken K. Strategic defense and attack for series and parallel reliability systems. *Eur J Oper Res* 2009;186:856–881.
9. Barlow RE, Proschan F. *Mathematical theory of reliability*. New York: Wiley; 1965.
10. Gnedenko BV, Belyayev YK, Soloviev AD. *Mathematical methods of reliability theory*. 2nd ed. London: Academic Press; 1969.
11. Leemis L. *Reliability: probabilistic models and statistical methods*. New Jersey: Prentice Hall; 1995.
12. Jelinski Z, Moranda P. Software reliability research. In: Freiberger W, editor. *Statistical computer performance evaluation*. New York: Academic Press; 1972. pp. 465–484.
13. Basu AK, Basu AB, Batra R. Modeling the response pattern to direct marketing campaigns. *J Market Res* 1995;32:204–212.
14. Osborne J, Severini T. The Lorenz curve for model assessment in exponential order

- statistic models. *J Stat Comput Simul* 2002;72:87–97.
15. Özekici S, Soyer R. Reliability of software with an operational profile. *Eur J Oper Res* 2003;149:459–474.
 16. Agustin M, Agustin Z. Modeling a system of softwares under imperfect debugging. *Inter-Stat* 2009;15(4). Available at <http://interstat.statjournals.net/YEAR/2009/articles/0907004.pdf>.
 17. Davis HT, Feldstein ML. The generalized Pareto law as a model for progressively censored survival data. *Biometrika* 1979;66:299–306.
 18. Littlewood B. Theories of software reliability: How good are they and how can they be improved? *IEEE Trans Softw Eng* 1980;6:489–500.
 19. Agustin MA. Statistical properties of a system reliability estimator using the Littlewood software reliability model. *J Appl Probab* 2003;40:766–778.

SYSTEMS MODELING TO INFORM DRUG POLICY

JONATHAN P. CAULKINS
Heinz College, Carnegie Mellon
University, Pittsburgh,
Pennsylvania

The production, distribution, use, and control of illegal drugs generates enormous social costs. In countries that have recently quantified the annual burden, the estimates are very large: \$1.4 billion in Canada [1], \$2.3 billion in France [2], \$6 billion in Australia [3], £15.4 billion in the United Kingdom [4], and \$180 billion in the United States [5].

Not surprisingly, there is energetic debate concerning how best to control drug use and its related consequences. Operations research/management science (OR/MS) has made important contributions to this debate.

This article does not attempt a comprehensive review. Rather, it highlights sample contributions from some classic studies, with additional references for those who want to dig deeper. The goal is to offer a flavor of how OR/MS perspectives can contribute to this important policy debate, not to provide a comprehensive resource for drug policy analysts.

The focus is on analysis that supports “systems understanding” and “strategic” drug policy choices. System understanding is valuable because the drug “system” is complex, and unintended consequences of interventions are common. Drug policy at the strategic level can be viewed as a resource allocation problem with resources divided across programmatic areas (source country control, interdiction, prevention, etc.) in order to optimize some social welfare function.

The discussion is organized into sections that highlight parallels between traditional interests of OR/MS and issues that are important when modeling drug policy. Specifically, the drug-related topics with their conventional OR/MS analogs in parentheses are modeling initiation (new product adoption),

demand (Markov modeling), supply (production and distribution), supply shocks (disequilibrium modeling), and policy evolution over the course of a drug epidemic (dynamic optimization).

Each section can be seen as addressing part of an integrated dynamic optimization problem. If we let $I()$ denote initiation and $\mathbf{X}()$ denote the current number of drug users, then the first two sections describe how initiation, $I()$, and endogenous growth, $g()$, depend on the state ($\mathbf{X}$) and controls ($\mathbf{u}$). So suppressing the time argument, t , for clarity, we may describe the evolution of drug use by

$$d\mathbf{X}/dt = g(\mathbf{X}, \mathbf{u}) + I(\mathbf{X}, \mathbf{u}).$$

Intersecting at any given time, the resulting demand, $Q_D(\mathbf{X})$, with a model of supply $Q_S(\mathbf{X}, \mathbf{u})$ modulated by possible supply disruptions, $\varepsilon(\mathbf{u})$, discussed in the third and fourth sections, yields a condition that determines the market clearing price, p :

$$Q_D(\mathbf{X}; p) = Q_S(\mathbf{X}, \mathbf{u}; p) + \varepsilon(\mathbf{u}).$$

This price in turn modifies the epidemic dynamics. That is, the system dynamics depend on the price, p and so might be more properly written as

$$d\mathbf{X}/dt = g(\mathbf{X}, \mathbf{u}; p) + I(\mathbf{X}, \mathbf{u}; p).$$

The final section addresses the policy question, what time-varying control vector $\mathbf{u}$ influences this dynamic system to evolve in the “best” possible way, for some definition of best.

MODELING DRUG INITIATION AS NEW PRODUCT ADOPTION

Bass’ [6] classic model initiated a significant literature on new product adoption. (See also the section titled “Forecasting” in this encyclopedia, for forecasting and diffusion models.) The key feature of Bass’ model was

that while some people (innovators) might spontaneously decide to use a new product, others (imitators) do so by following the lead of people who were already using the product. Hence, an important subset of product initiation could be modeled as being proportional to the current number of users of the product times the number of people who had not yet started using it.

Modeling “initiation” as being proportional to the product of people who are currently in the population of interest (“infected”) times the number who could join (“susceptibles”) also has a storied history within so-called *susceptible–infected–recovered* (SIR) models of the spread of contagious diseases [7]. Indeed, people have long referred to the ebb and flow of drug use as “epidemics” because current users tend to “recruit” new users [8,9].

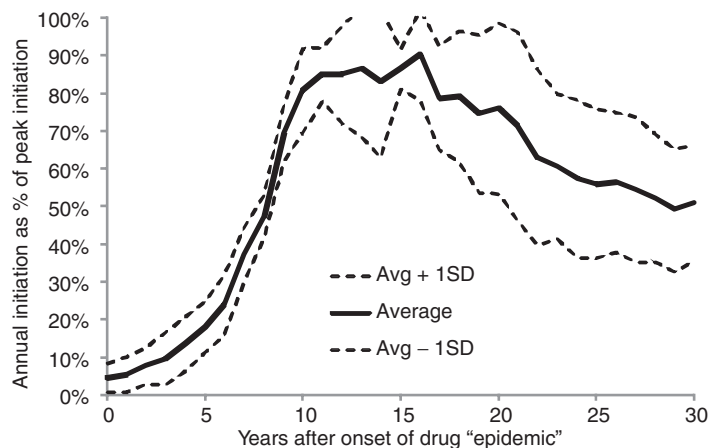
Attempts to mathematically model drug problems draw inspiration from both of these literatures. The metaphor of infectious spread is particularly potent for illegal drugs because typically the other user is not only the inspiration for initiation, but also the supplier, since prospective users cannot simply go to a store to obtain drugs. Furthermore, product illegality stifles advertising, so word-of-mouth diffusion of product information is more important than it is for most conventional goods [10].

The regularity of illegal drugs’ new product adoption trajectories is striking. Figure 1 illustrates this with retrospective self-report

data provided by 388,617 respondents to the 1999–2005 US National Household Surveys on Drug Use and Health. Respondents who had used any of these nine illegal drugs [marijuana, cocaine, crack, heroin, methamphetamine, hallucinogens, lysergic acid diethylamide (LSD), phencyclidine (PCP), and Ecstasy] were asked in which year they started using that drug. Popularity of the drugs varies considerably (many more people have tried marijuana than heroin), so the series were normalized to make their initial peak level be 100%. Similarly, some drugs became widely available earlier than others (marijuana and heroin in the 1960s, cocaine in the 1970s, and Ecstasy even later), so the normalized initiation plots were time shifted to align the convex growth part of their “S-curves.”

Plotting all nine lines creates a graph that is hard to read because the lines overlap (figure not shown), but the overlap is exactly the point. Despite enormous differences in price, dependence potential, neural pathways affected, and so on, all nine substances followed similar initiation trajectories after normalization and time shifting. Figure 1 shows this commonality by plotting the average of the nine normalized trajectories along with dotted lines showing the average ± 1 standard deviation. Where the dotted lines are close to each other, all nine substances followed very similar trajectories. Postpeak there was some variability. Initiation dropped off dramatically for some drugs (Ecstasy and PCP) and

Figure 1. Average (± 1 std dev) across nine drugs of initiation trajectories in the US. After normalizing so peak initiation = 100% and time shifting to align their stages of rapid growth.



substantially for others (heroin and cocaine), but much less so for a few (e.g., hallucinogens). However, in the early growth stages of the epidemics, there was very little variation; initiation trajectories were very similar for all drugs suggesting that common “marketing” characteristics trump “pharmacological” factors when trying to understand this spread.

Another empirical regularity observed in US cocaine data is that the “infectivity” of current light users is inversely related to the ratio of current heavy to current light users [11]. That is, word of mouth seemed to work “best” as an initiation promoter when there were lots of light users relative to the number of heavy users.

Musto [12] offers a feasible hypothesis for why this should be so. The explanation depends on the fact, elaborated below, that drug use progresses through stages. Although there is a spectrum of intensities of use, much is gained simply by differentiating between “light users” who have initiated recently and are typically happy with their drugs and “heavy users” who may have escalated to drug dependence and are a visible reminder of the dangers of drug use.

As Kleiman [13] notes, early in a drug epidemic, most users are light users because they have not had time to escalate to dependent or problematic use. Since most of the current users are experiencing no problems, the drug has a relatively benign reputation. That benign reputation makes the drug attractive to nonusers, so the nonusers are likely to initiate when given the opportunity. This positive feedback from current light users to new initiates leads to rapid growth in initiation.

Over time, some of the original users escalate to heavy use, but news of their struggles is lost amidst the burgeoning sea of relatively happy light users. Eventually, though, either enough people escalate to heavy use or the rapid growth in initiation is tempered by exhaustion of the pool of susceptibles, and the proportion of current users who are experiencing problems begins to grow. When that happens, the drug begins to be seen as dangerous, which deters some potential initiates. As the inflow of new light users

ebbs, the total number of light users falls because people who never escalate typically do not use a drug like cocaine for more than a few years [11]. This further increases the proportion of dependent users, which in turn intensifies the drug’s negative reputation, further undercutting initiation. Thus, a negative feedback loop is established, and initiation is restricted to the subset of the population who are not deterred by the drug’s negative reputation. (Eventually, the drug’s reputation and initiation may bounce back somewhat from the nadir induced by the “bulge” in dependent drug use that comes a few years after the peak in light use.)

Models that build on such dynamics [14–17] have achieved strong fits to cocaine’s historical data by adding a reputational feedback term that multiplicatively modifies initiation. A typical form of the reputational adjustment might be

$$\text{Reputational adjustment} = (f + (1 - f) \times \exp(-qH/L)),$$

where f and q are the positive constants and H , L are the respective numbers of heavy and light users. (Early models set $f = 0$, but now it appears that overstates this “Musto dynamic”.) Almeder *et al.* [18,19] extended this formula by looking at a family of age-distributed models in which this reputational feedback also depends on the relative and absolute ages of current and potential users. To caricature, a 16-year-old’s interest in trying a drug may increase if a lot of 19-year-olds use the drug, but decrease (or at least not go up as much) if people of the teen’s parents’ generation commonly use it.

State dependence of the “infectivity” of current users is an intriguing extension of the classic Bass and SIR models [20]. For new product adoption of consumer electronics or psychoactive drugs, the probability an opportunity to initiate leads to an actual initiation may depend on how popular the product is. This idea makes a certain intuitive sense. If a friend suggests trying a drug or a new gadget that few others use, you might be more skeptical than if the drug or gadget is popular. The opposite may be true for certain fashion goods: in as much as conspicuous

consumption of an exclusive good signals status, a product already in widespread use might be less, not more appealing.

Wallner [21] and Caulkins *et al.* [22] explored this possibility for drug use models parameterized for both, the US and Australia. They find that feedback from the general prevalence of use to the “infectivity” of current users creates nonlinear dynamics that can greatly enrich the range of possible system behaviors. Notably, state-dependent infectivity can create multiple equilibria separated by “tipping points.” Around those tipping points enforcement (in the case of drugs) and advertising (in the case of conventional goods) may be particularly valuable if it can tip the system from a trajectory approaching an undesirable equilibrium over onto another trajectory approaching the more desirable equilibrium (the low use equilibrium for drugs, or high use for electronics).

To summarize, drug initiation can be modeled successfully with tools from the literatures on new product adoption and infectious disease transmission. This shocks some people in the addictions field, but actually makes perfect sense. The psychopharmacological characteristics of a substance cannot possibly play a direct role in decisions to initiate for the simple reason that before someone initiates, the drug has not entered the bloodstream or the brain. So, any peculiar effects stemming from the drug’s intoxicating or dependence-inducing potential are not yet relevant. We turn next to modeling the dynamics of ongoing consumption by existing users.

“STOCKS AND FLOWS” OR MARKOV MODELS OF DRUG DEMAND

OR/MS has long used a discrete set of states to represent the state of a system; for instance, a queueing model might define the state as the number of customers in the system. Estimating the rates of transition from one state to another allows one to explore the system’s steady-state properties and model the transient behavior as well. (See also the sections titled “Stochastic Processes” and “Queueing Theory and Queueing Networks”

on Markov chains and queueing models, in this encyclopedia.) Demographic models apply the same principle at a population level, with boxes (states) representing the number of people in various life situations (e.g., being an employed 30-year-old), with arrows representing flows of people from one state to another.

These same principles have proved enormously useful in drug policy to model both individual level trajectories or “careers” of drug use and also to model the aggregated evolution of a drug “epidemic.” Without the OR/MS mindset, people often track just the total number of past-year users, without adequately differentiating between types of users, and/or view annual prevalence estimates as a series of still pictures and not a continuous process that happened to be sampled at discrete intervals.

A number of early “systems dynamics” models had been produced over time [23–26], but in the early 1990s, Everingham and Rydell [27] of RAND’s Drug Policy Research Center made a pioneering breakthrough. They observed that much of the historical dynamics of the US cocaine epidemic could be described through a simple model that differentiated between only two types of users: “light users” who have initiated recently and are typically happy with their drug use and “heavy users” who have escalated to drug dependence and are a visible reminder of the dangers of drug use. Caulkins *et al.* [11] showed that the same model with slightly updated parameters continued to have great explanatory power a decade later, even though the data gathered in the intervening decade reflected a greatly changed epidemic.

Figure 2 describes the Everingham and Rydell model [27] with typical parameter values for a “hard drug” based on Caulkins *et al.* [11] and other more recent work.

This simple model reflects a number of crucial insights. For one, most (roughly 5 in 6) initiates do not escalate to heavy or dependent use. However, those who do escalate remain in that heavy or dependent use state for many years. In Fig. 2, $1/g = 12.5$ years, but values of g as low as 0.05 are common, implying an average dwell time in dependent use as long as 20 years. When

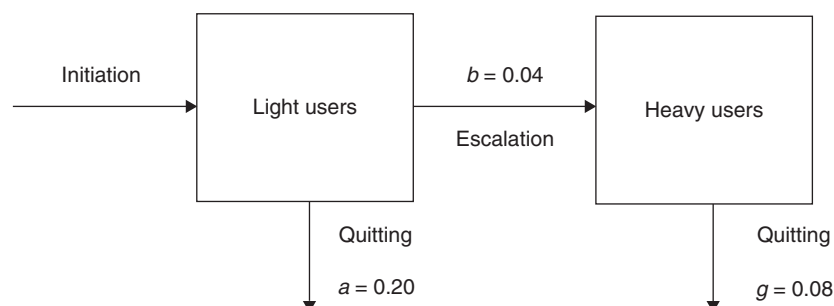


Figure 2. Everingham and Rydell's [27] light and heavy user model with flow rates updated by Caulkins *et al.* [11].

Everingham and Rydell's work [27] first appeared, audiences knew that opiate dependence could last in excess of a decade [28], but most were skeptical that cocaine dependence would last as long. (Whereas heroin addicts develop a certain routine, stimulant abuse is often characterized by bingeing, with use becoming all-engrossing until the user runs out of money. It was presumed by some that such behavior could not be sustained for many years.) Unfortunately, time has proven Everingham and Rydell's model correct; cocaine abuse is now thought of as a chronically relapsing condition despite the periods of abstinence or reduced use between binges.

The persistence of dependent use has a crucial implication for the value of what might otherwise seem to be relatively ineffective treatment interventions. In the past, the goal of treatment was to increase the out-flow rate from heavy or dependent use, and it appeared to be singularly ineffective at that goal. Relapse rates on the order of 80% were widely reported, and taxpayers became reluctant to "waste" their money on programs with such high failure rates.

Everingham and Rydell [27] looked into this issue from a systems perspective. Definitive data on treatment effectiveness were not and still are not available [29]. However, a simple calculation drawing on the model shows that, a treatment program with an 80% relapse rate can still be a terrific investment from the taxpayers' perspective.

Suppose that 80% of people enter a treatment program relapse, but 20% cease or substantially reduce their consumption in the

long run. Since from Fig. 2 we might expect something like 8% to have quit even without treatment, we might suppose that treatment led 12% of the admits to cease use. Their residual career of use could have been on the order of 10 or more years, suggesting a reduction of 1.2 person-years of dependent use. Furthermore, most (perhaps 80%) of the people cease or substantially reduce use during the 3 months (roughly) they are in treatment; so the total reduction might be closer to 1.5 person-years of use.

Social costs per person-year of dependent use vary with individual circumstances (notably, what proportion of drug purchases are financed by crime), but a figure of \$25,000 per year is not unreasonable. (For cocaine, the figures might be 120 g consumed per year with an average social cost of \$215 per gram consumed [30].) This suggests that treatment programs with 80% relapse rates could still generate social benefits of the order of $1.5 \times \$25,000 = \$37,500$ per person treated, vastly exceeding the average cost of \$2000–3000 per treatment admission. (The average cost per admission is so low because so many people drop out very quickly.) Indeed, programs that induce even 1 in 100 of those admitted to quit use permanently could be cost-effective.

It is simply not known what the exact parameters for such a calculation are, and they no doubt vary by country, drug, and type of user. Nevertheless, two conclusions can be drawn. First, if treatment is effective for even a minority of users, then it is probably cost-effective in the sense of averting more

social costs than it takes taxpayer dollars to subsidize it. Second, the performance metric used to evaluate treatment matters mightily. A treatment intervention that does very poorly when judged in terms of abstinence rates could still offer taxpayers an economical way of reducing social costs, including costs borne by nonusers, such as those associated with drug-related crime and violence. On the other hand, even treatment on demand [31] will not necessarily quickly eliminate most drug use [32,33].

Everingham and Rydell's simple two-state model has been extended in various ways. We mention just one. Kaya *et al.*'s [34] analysis of Australian heroin data shows that the number of people quitting heroin is highly correlated over time with the number initiating, presumably because the modal career of use is very short. Essentially for all illegal drugs, almost half of the people who try the drug report using it less than a total of 12 times in their lifetime. Furthermore, there was some tendency for initiation to oscillate up and down within a period of a few years. Also, Marotta [35] and Preti *et al.* [36] observe fairly high frequency oscillations in the number of heroin overdoses in Italy, and longer term oscillations have been observed for many drugs, including alcohol. Caulkins *et al.* [37] extend Everingham and Rydell's *LH* model to include two additional states. One is a very brief initial state from which users either progress to ongoing nondependent use (akin to the *L* state in Everingham and Rydell's model) or, if they do not like the drug, they quit. Those who quit more or less immediately are assumed to tell their peers that the drug is not particularly appealing, creating a negative feedback on initiation. However, inasmuch as these early quitters were never engrossed in the drug culture, they and their cautions are assumed to dissipate rather quickly. This four-state structure creates the possibility for simultaneous cycling on two different timescales (fast and slow). The resulting model displays sensible, but highly sensitive dependence on certain parameter values, with modest changes pushing the model through regions where its behavior is stationary, cyclic, quasi-periodic,

and chaotic. The model reveals that if different drugs have even modestly different attributes in terms of escalation lags or probabilities, strength of reputational feedback, and so on, that could translate into very different epidemic behaviors at the macrolevel.

To summarize, state-transition models of drug use have found wide application, not only for illegal drugs such as heroin [38,39], but also for tobacco [40,41] and it remains a promising area for further research.

MODELING THE SUPPLY CHAIN FOR ILLEGAL DRUGS

Medical and public health research focuses on drug use, but systems perspectives must also encompass drug supply. Modeling production and distribution is of course another area to which operations research has contributed. (See also the section titled "Supply Chain Management" in this encyclopedia.)

Building on Kennedy *et al.* [42], Childress [43,44] and Dombey-Moore *et al.* [45] developed descriptive systems models of the drug production and distribution chain from the source to the street. This foundation has been elaborated by many authors (e.g., Anthony and Fries [46]), but the essential observation for the typical drug is that there are several distinct stages of processing in the source country. Then, the finished product is moved across international borders where it is distributed domestically through a multilayered domestic distribution network, changing hands perhaps five or six times between import and final sale to the user [47], with prices increasing enormously as the drugs move down this distribution chain [48].

The standard paradigm for understanding why price markups are so large is Reuter and Kleiman's [49] "risks and prices" model. In it, enforcement acts like a tax, driving up the cost of distributing drugs. Since dealers distribute drugs to make money, they are assumed to pass these costs along to users in the form of higher prices. Contrary to what was once believed, there is clear evidence that drug use does respond to changes in price. Just as with other consumer goods, higher prices suppress current use (the "conditional elasticity") and also affect how many

people use (the “participation elasticity”). See Grossman [50] for a review of literature on the price responsiveness of drug use.

The risks and prices model raises the question, which types of enforcement are best in the sense of imposing the greatest costs on drug suppliers per million taxpayer dollars spent (or per some other metric, such as per person imprisoned)?

Drawing on the foundational work just described, Rydell and colleagues [32,33] developed one of the most frequently cited “systems analyses” of the relative effectiveness of different broad strategies for controlling drug use, in this case US cocaine use. Four strategies were considered: treating heavy cocaine users and three types of supply control programs (interventions in source countries, interdiction between the source countries and the United States, and law enforcement against drug suppliers operating within the United States).

The model assumed that the market was in dynamic equilibrium, so that each year suppliers’ revenues compensated them for all of the costs of business, including normal profits and monetary compensation for non-monetary risks of enforcement and violence. The model traced the evolution of the market over a 15-year planning horizon, as changes in demand and supply shifted the equilibrium from year to year. The supply model was multilayered and tracked separately the market prices and quantities in source countries, in transit, and within the United States, while factoring in both seizures and the markups needed to compensate for the risks of moving the drugs from one level to the next.

The demand at any given time had modest price responsiveness, and in addition prices influenced initiation, escalation, and cessation flows. So the price increase in a year would reduce the number of users in the following year.

The model led to a number of insights. For example, while it is cheaper for authorities to seize drugs closer to the source, it is also cheaper for the drug suppliers to replace drugs lost at those market levels. Indeed, as one moves down the distribution chain, the value of the drugs rises faster than does the difficulty of seizing them. As a result,

the “tax” imposed by enforcement seizures is actually stiffer within the United States than it is further upstream, in the sense that the costs imposed on suppliers is greater per million US taxpayer dollars spent. The overall model of enforcement included many factors besides drug seizures (including arrests, imprisonment, avoidance costs, and seizure of nondrug assets), but the same intuition holds over all. As long as the costs are passed down the distribution chain one-for-one (the so-called *additive model of price transmission*), enforcement can impose more damage on suppliers’ cost structure per million taxpayer dollars spent when it operates within the United States than it can when it operates abroad, particularly in source countries.

Within US borders, the risks and prices model has the feel of an arms race. If enforcement can impose enough cost on the suppliers per taxpayer dollar spent, it could be cost-effective. Everingham and Rydell [27] and most subsequent analyses [51,52] find that it is costly to fight this arms race once a market has matured. A mature distribution network is robust because of its many lateral linkages, independent paths, and the ability to expand quickly the capacity of individual arcs to compensate for arcs or nodes that have been removed by enforcement.

Furthermore, a concern with using enforcement to drive up drug prices is that high prices might increase crime even if they reduce use. Most drug-related crime is “economic-compulsive” (committed to obtain money to buy drugs) or “systemic” (arising from drug selling, e.g., punishment for nonpayment) and so is driven primarily by drug spending. Crime related to intoxication or use, called *psychopharmacological drug-related crime*, accounts for perhaps only one-sixth of all drug-related crime [51]. Hence, depending on the elasticity of demand, driving up prices could actually increase, not decrease, drug-related crime [53].

Hence, once the market has matured, it may be more constructive to task enforcement with controlling market-related externalities (notably crime and disorder) rather than seeking to reduce use by suppressing supply across the board [54].

To summarize, it is difficult to evaluate the effectiveness of supply control interventions. Randomized controlled trials are not possible, and even good natural experiments are rare. However, it is possible to build policy simulation models of the production and distribution networks and how they respond to drug control interventions. These models are indispensable for structuring the thinking about potential effects of supply control programs and for generating insights about the relative cost-effectiveness of at least a subset of them.

MODELING SHOCKS TO THE DISTRIBUTION SYSTEM

There has been growing interest in modeling how operations should respond to unusually stressful circumstances, ranging from humanitarian logistics responding to a war or physical disaster [55] to airlines trying to recover from severe weather [56]. (See also the section titled “Humanitarian and Public Service Applications” in this encyclopedia.)

Drug policy explores the inverse problem. Instead of asking how activities can be designed to remain robust in the face of disruptions, the goal is to understand how disruptions can be designed to interfere with

normal operations. One line of work focuses on disrupting local street markets, using differential equation models [57–62] and more recently, agent-based methods [63–65]. Another focuses on disrupting international distribution networks by drawing on models used in military logistics [66–68] and creating custom models [69,70]. A related line of research asks how to exploit a disruption once one has been created [71,72].

The potential for disruptions is clear. At one time or another, there have been major disruptions of the markets for every major illegal drug. A conspicuous recent example is the so-called Australian heroin “drought” [73]. Market disruptions appear to correlate with contemporaneous changes in measures of price, purity, drug test results (for arrestees and the workforce more generally), treatment admissions, and overdose [46,74–79].

Figure 3 illustrates such a correlation with data from the Australian state of Victoria, which experienced a severe disruption to its heroin market in early 2001. As the left panel of Fig. 3 shows, prices per “raw” gram did not change much because of the standardized pricing, but purity plummeted, so price per pure gram spiked sharply. The right panel shows both the trend in ambulance calls for

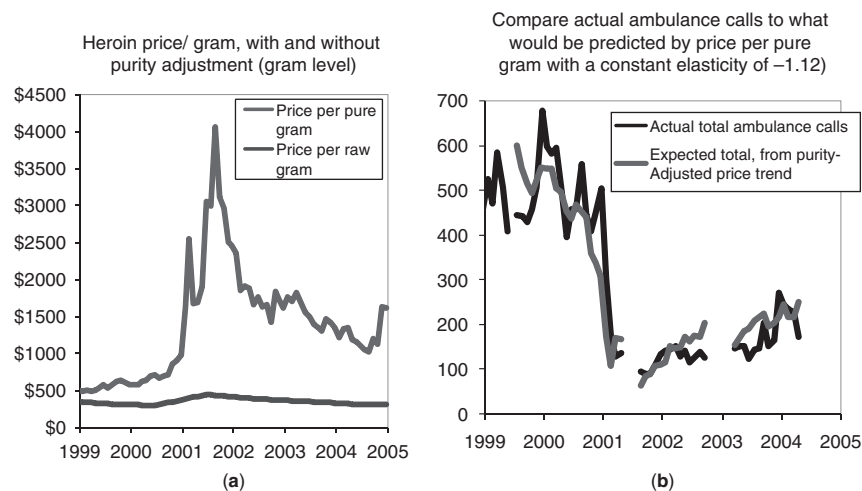


Figure 3. (a) Trends in the price per pure gram of heroin in Victoria, Australia; (b) graph of price series raised to a constant elasticity can explain most of the variation in ambulance calls for heroin overdose.

heroin overdose and the numbers that would be predicted by feeding the purity-adjusted price series into a standard constant elasticity demand model. The predicted and the actual series are very highly correlated. That is, simply raising the price per pure gram to a power (called the *elasticity*) and scaling it by a multiplicative constant creates a series that strongly parallels the actual trends in overdose.

There remains enormous debate about what exactly caused the Australian heroin drought [78]. Perhaps the most plausible explanation is that Australian Federal Police were able to dismantle the small number of organizations capable of importing shipments of more than 100 kg at a time [79]. Figuring out exactly what caused past market disruptions and how to reproduce them in the future is an important area of research to which OR/MS has much to contribute.

OPTIMAL DYNAMIC CONTROL MODELS

Drug use varies over time in a manner that is referred to as an *epidemic cycle*. From initially low levels before markets have fully developed, drug initiation and use grow rapidly, as described in Fig. 1. After a time, initiation peaks and shortly thereafter, overall prevalence and light use do too. While total prevalence falls, the number of heavy users can still grow because escalation to heavy use occurs with a lag. At this point, the epidemic switches into an endemic stage, characterized by an aging pool of dependent users that is partially but not completely replaced by an ongoing lower level of initiation and light drug use.

The initial “systems models” [27] estimated the relative cost-effectiveness of different drug control interventions under the implicit assumption that the drug problem had stabilized in this endemic stage. However, there is no reason to believe that the mix of drug control strategies that is optimal at one stage of a drug epidemic is the best for other stages. That is, it is possible that drug policy should evolve as the drug problem does.

Building on earlier research on dynamic law enforcement [80,81], researchers from

the Vienna University of Technology used optimal dynamic control theory to investigate the possibility of different control strategies being effective at different states of a drug epidemic. These models offer a number of policy recommendations. For example, law enforcement is relatively most effective early in an epidemic, when the market is still small. Hence, when policy makers become aware of a burgeoning problem, it may be advisable to make a discrete choice between an “eradication” strategy that seeks to drive the problem down to a low level equilibrium and an “accommodation” strategy that only seeks to slow its growth to a high level equilibrium, using a mix of control spending that shifts more and more over time toward treatment [82]. Zeiler [20] extends the accommodation strategy results by showing that there can also be choices between a relatively passive accommodation and a grudging accommodation that fights a holding strategy as long as possible.

Behrens *et al.* [15,16] explore timing issues for treatment and prevention. Perhaps not surprisingly, the model suggests allocating a substantially larger share of a drug control budget to prevention early in an epidemic and treatment later. (Winkler *et al.* [17] go on to distinguish a number of varieties of prevention.) Wallner [21] suggests that later in an epidemic, harm reduction interventions can be useful, even if earlier they might have risked “tipping” an epidemic to a higher level of use by promoting initiation. (Harm reduction interventions that have no adverse effect on drug prevalence could be useful at any point in an epidemic.)

Most of this optimal dynamic control work has modeled just one drug at a time, but Zeiler [20] explores multiple-interacting drug epidemics within an optimal control framework. This natural extension is akin to what Margaret Brandeau, Elisa Long, Gregory Zaric, and others have done for allocating resources across competing epidemics of HIV in different populations [20,83,84].

To summarize, people have long recognized that no one policy is best for all drugs and places, but this section observes that similarly no one policy is best for all times (epidemic stages) even for a given drug and

location. Policy makers readily accede to this idea that since drug use evolves over time, drug policy should too. However, the standard analytical tools of most of the disciplines that study drug policy are not designed to frame policy analysis in this way. In contrast, OR/MS has tools that can grapple with problems in terms of dynamic optimization, specifically optimal dynamic control theory.

RELATED TOPICS AND OPPORTUNITIES FOR FURTHER RESEARCH

Drug policy is a complicated domain. From a mathematical perspective, the system is characterized by nonlinear endogenous feedback, inertia, lags, market equilibria, disequilibrium adjustments, multiple objectives, multiple stakeholders, and great uncertainty. This complexity can overwhelm humans' intuitive reasoning, and unintended consequences are rampant even for well-intentioned interventions. Fortunately, many pieces of this complex system parallel phenomena that OR/MS has studied for decades. Hence, OR/MS tools can be enormously helpful in bringing precision and clarity to thinking about drug policy.

In particular, drug initiation, $I(\mathbf{X}, \mathbf{u})$, behaves like new product adoption. The ongoing evolution of a drug epidemic has a Markovian character, $d\mathbf{X}/dt = g(\mathbf{X}, \mathbf{u}) + I(\mathbf{X}, \mathbf{u})$. The resulting demand, $Q_D(\mathbf{X})$, intersects with supply, $Q_S(\mathbf{X}, \mathbf{u})$ and any supply shock, $\varepsilon(\mathbf{u})$, to determine a market clearing price, p , that satisfies

$$Q_D(\mathbf{X}; p) = Q_S(\mathbf{X}, \mathbf{u}; p) + \varepsilon(\mathbf{u}).$$

The equilibrium is dynamic, not static, so the overall policy problem can be thought of as finding a policy vector $\mathbf{u}$ that evolves over time in a way that minimizes the integral of the costs of drug use ($c_1(\mathbf{X}, p)$) and control ($c_2(\mathbf{u})$):

$$\begin{aligned} \text{Min } J &= \int_0^\infty e^{-rt} (c_1(\mathbf{X}, p) + c_2(\mathbf{u})) dt \\ \text{s.t. } d\mathbf{X}/dt &= g(\mathbf{X}, \mathbf{u}; p) + I(\mathbf{X}, \mathbf{u}; p) \text{ and} \\ Q_D(\mathbf{X}; p) &= Q_S(\mathbf{X}, \mathbf{u}; p) + \varepsilon(\mathbf{u}). \end{aligned}$$

Many insights stem just from analyzing the various pieces in isolation, but the dynamic optimization formulation provides a concise framework for understanding how the various pieces interact.

This review of OR/MS applied to drug policy has been selective, describing a few ideas in enough detail to appreciate their contributions rather than giving a brief paragraph about each. We have omitted much, notably a companion stream of research on the intersection of injection drug use and HIV and HCV pioneered by Kaplan's "Needles that Kill" [85,86] model and subsequent work quantifying drug treatment's cost-effectiveness when factoring direct and secondary reductions in disease transmission [87–92].

Similarly, there are exciting developments in network and agent-based models of drug markets and drug enforcement that merit their own review [93–98].

It is important to note, though, that even a comprehensive review of drug policy systems modeling would leave the majority of the drug policy questions unanswered. The corpus of research to date merely scratches the surface of what could be done.

Economists frequently talk about "diminishing returns" whereby the incremental benefit of additional units of a resource decreases the more of that resource that has been allocated. Such a principle may apply with respect to analysis of drug policy problems. While some disciplines that have studied drug policy for decades may have reached diminishing returns, there are still relatively few people with an engineering mind set who study drug control problems, so much low-hanging fruit remains to be harvested.

REFERENCES

1. Single E, Robson L, Xie X, *et al.* The economic costs of alcohol, tobacco and illicit drugs in Canada, 1992. *Addiction* 1998;93(7):991–1006.
2. Fenoglio P, Parel V, Kopp P. The social cost of alcohol, tobacco and illicit drugs in France, 1997. *Eur Addict Res* 2003;9(1):18–28.
3. Collins DJ, Lapsley HM. Counting the cost: estimates of the social costs of drug abuse in

- 1998–99. Canberra: Commonwealth of Australia; 2002.
4. Singleton N, Murray M, Tinsley L. Measuring different aspects of problem drug use: methodological developments. Home Office Online Report 16/06. London: Home Office; 2006.
 5. Office of National Drug Control Policy. The economic costs of drug abuse in the United States, 1992–2002. Washington DC: Executive Office of the President; 2004. Publication No. 207303.
 6. Bass FM. A new product growth model for consumer durables. *Manag Sci* 1969; 15(5):215–227.
 7. Anderson R, May R. Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press; 1992.
 8. DuPont RL, Greene MH. The dynamics of a heroin addiction epidemic. *Science* 1973;181:716–722.
 9. de Alarcon R. The spread of heroin abuse in a community. *Bull Narc* 1969;21:17–21.
 10. Ferrence R. Diffusion of innovation as a model for understanding population change in substance use. In: Edwards G, Lader M, editors. *Addiction: processes of change. Society for the study of addiction monograph No. 3*. New York: Oxford University Press; 1994. pp. 189–201.
 11. Caulkins JP, Behrens DA, Knoll C, *et al.* Markov Chain modeling of initiation and demand: the case of the US cocaine epidemic. *Health Care Manag Sci* 2004;7(4):319–329.
 12. Musto DF. *The American disease: origins of narcotics control*. New York: Oxford University Press; 1987.
 13. Kleiman MAR. Enforcement swamping: a positive-feedback mechanism in rates of illicit activity. *Math Comput Model* 1993;17:65–75.
 14. Behrens DA, Caulkins JP, Tragler G, *et al.* A dynamic model of drug initiation: implications for treatment and drug control. *Math Biosci* 1999;159:1–20.
 15. Behrens DA, Caulkins JP, Tragler G, *et al.* Optimal control of drug epidemics: prevent and treat—but not at the same time? *Manag Sci* 2000;46(3):333–347.
 16. Behrens DA, Caulkins JP, Tragler G, *et al.* Why present-oriented societies undergo cycles of drug epidemics. *J Econ Dyn Control* 2002;26(6):919–936.
 17. Winkler D, Caulkins JP, Behrens D, *et al.* Estimating the relative efficiency of various forms of prevention at different stages of a drug epidemic. *Soc Econ Plann Sci* 2004;38:43–56.
 18. Almeder C, Caulkins JP, Feichtinger G, *et al.* Age-specific multi-state initiation models: Insights from considering heterogeneity. *Bull Narc* 2001;53(1):105–118.
 19. Almeder C, Caulkins JP, Feichtinger G, *et al.* An age-structured single-state initiation model—cycles of drug epidemics and optimal prevention programs. *Soc Econ Plann Sci* 2004;38(1):91–109.
 20. Zeiler I. Optimal dynamic control with DNSS curves: multiple equilibria in epidemic models of HIV/AIDS and illicit drug use [Doctoral Thesis], Institute for Mathematical Methods in Economics, Vienna University of Technology. 2007.
 21. Wallner D. Dynamic models of drug users and susceptibles: optimal mix of use reduction and harm reduction in Australia and the U.S.A. [Doctoral Dissertation], Technical University of Vienna. 2009.
 22. Caulkins JP, Tragler G, Wallner D. Optimal timing and extent of use versus harm reduction in a SA model of drug epidemics. *Int J Drug Pol* 2009;20(6):480–487.
 23. Levin GE, Roberts EB, Hirsch GB. *The persistent poppy: a computer-aided search for heroin policy*. Cambridge (MA): Ballinger Publishing; 1975.
 24. Gardiner LK, Shreckengost RC. A system dynamics model for estimating heroin imports into the United States. *Syst Dyn Rev* 1987;3(1):8–27.
 25. Homer JB. A system dynamics model for cocaine prevalence estimation and trend projection. *J Drug Issues* 1993a;23(2): 251–280.
 26. Homer JB. Projecting the impact of law enforcement on cocaine prevalence: a system dynamics approach. *J Drug Issues* 1993b; 23(2):281–295.
 27. Everingham SS, Rydell CP. *Modeling the demand for cocaine*. Santa Monica (CA): RAND; 1994.
 28. Hser YI, Anglin D, Powers K. A 24-year follow-up of California narcotics addicts. *Arch Gen Psychiatr* 1993;50(7):577–584.
 29. Manski CF, Pepper JV, Petrie CV, editors. *Informing America's policy on illegal drugs: what we don't know keeps hurting us*. Washington, DC: National Academy Press; 2001.
 30. Caulkins JP, Pacula R, Paddock S, *et al.* School-based drug prevention: what kind of

- drug use does it prevent? Santa Monica (CA): RAND; 2002.
31. Kaplan EH, Johri M. Treatment on demand: an operational model. *Health Care Manag Sci* 2000;3(3):171–183.
 32. Rydell CP, Everingham SMS. Controlling cocaine: supply versus demand programs. Santa Monica (CA): RAND; 1994.
 33. Rydell CP, Caulkins JP, Everingham SMS. Enforcement or treatment: modeling the relative efficacy of alternatives for controlling cocaine. *Oper Res* 1996;44(5):687–695.
 34. Kaya YC, Tugai Y, Filar JA, *et al.* Heroin use in Australia: population trends. *Drug Alcohol Rev* 2004;23(1):107–116.
 35. Marotta E. Drug abuse and illicit trafficking in Italy: trends and countermeasures 1979–1990. *Bull Narc* 1992;44(1):15–22.
 36. Preti P, Miotto P, DeCoppi M. Deaths by unintended illicit drug overdose in Italy, 1984–2000. *Drug Alcohol Depend* 2002;66:275–282.
 37. Caulkins JP, Gagnani A, Feichtinger G, *et al.* High and low frequency oscillations in drug epidemics. *Int J Bifurcat Chaos* 2006;16(11):3275–3289.
 38. Rossi C. Estimating the prevalence of injection drug users on the basis of Markov models of the HIV/AIDS epidemic: applications to Italian data. *Health Care Manag Sci* 1999;2:173–179.
 39. Rossi C. A mover-stayer type model for epidemics of problematic drug use. *Bull Narc* 2001;53(1–2):39–64.
 40. Mendez D, Warner KE. Smoking prevalence in 2010: why the healthy people goal is unattainable. *Am J Publ Health* 2000;90(3):401–403.
 41. Levy DT, Chaloupka F, Gitchell J, *et al.* The use of simulation models for the surveillance, justification and understanding of tobacco control policies. *Health Care Manag Sci* 2002;5(2):113–120.
 42. Kennedy M, Reuter P, Riley KJ. A simple economic model of cocaine production. *Math Comput Model* 1993;17(2):19–36.
 43. Childress M. A systems description of the heroin trade. Santa Monica (CA): RAND; 1994a.
 44. Childress M. A systems description of the marijuana trade. Santa Monica (CA): RAND; 1994b.
 45. Dombey-Moore B, Resetar S, Childress M. A systems description of the cocaine trade. Santa Monica (CA): RAND; 1994.
 46. Anthony R, Fries A. Empirical modeling of narcotic trafficking from farm gate to street. *Bull Narc* 2004;LVI(1–2):1–48.
 47. Caulkins JP. Modeling the domestic distribution network for illicit drugs. *Manag Sci* 1997;43(10):1364–1371.
 48. Caulkins JP, Reuter P. What price data tell us about drug markets. *J Drug Issues* 1998;28:593–612.
 49. Reuter P, Kleiman MAR. Risks and prices: an economic analysis of drug enforcement. In: Tonry M, Morris M, editors. Volume 7. Crime and justice: an annual review of research. Chicago: University of Chicago Press; 1986. pp. 128–179.
 50. Grossman M. Individual behaviors and substance use: the role of price. In: Lindgren B, Grossman M, editors. Volume 16. Substance use: individual behavior, social interactions, markets, and politics. *Advances in health economics and health services research*. Amsterdam: Elsevier; 2005. pp. 407–439.
 51. Caulkins JP, Rydell CP, Schwabe WL, *et al.* Mandatory minimum drug sentences: throwing away the key or the taxpayers' money? Santa Monica (CA): RAND; 1997.
 52. Kuziemko I, Levitt SD. An empirical analysis of imprisoning drug offenders. *J Public Econ* 2004;88:2043–2066.
 53. Caulkins JP, Dworak M, Feichtinger G, *et al.* Price-raising drug enforcement and property crime: a dynamic model. *J Asian Econ* 2000;71(3):227–253.
 54. Caulkins JP, Reuter P. Toward a harm reduction approach to enforcement. *Safer Commun* 2009;8(1):9–23.
 55. Tomasini R, Van Wassenhove LN. Humanitarian logistics. United Kingdom: Palgrave Macmillan; 2008 pp. 1–256.
 56. Ball M, Barnhart C, Nemhauser G, *et al.* Air transportation: irregular operations and control. In: Laporte G, Barnhart C, editors. Volume 14. *Handbook in OR & MS*. North-Holland Amsterdam; 2006. pp. 1–68.
 57. Caulkins JP. Local drug markets' response to focused police enforcement. *Oper Res* 1993;41(5):848–863.
 58. Baveja A, Batta R, Caulkins JP, *et al.* Modeling the response of illicit drug markets to local enforcement. *Soc Econ Plann Sci* 1993;27(2):73–89.
 59. Baveja A, Caulkins JP, Liu W, *et al.* When haste makes sense: cracking down on street markets for illicit drugs. *Soc Econ Plann Sci* 1997;31(4):293–306.

60. Gragnani AS, Rinaldi S, Feichtinger G. Dynamics of drug consumption: a theoretical model. *Soc Econ Plann Sci* 1997;31(2): 127–137.
61. Kort PM, Feichtinger G, Hartl RF, *et al.* Optimal enforcement policies (crackdowns) on an illicit drug market. *Optim Contr Appl Meth* 1998;19(3):169–184.
62. Naik AV, Baveja A, Batta R, *et al.* Scheduling crackdowns on illicit drug markets. *Eur J Oper Res* 1996;88(2):231–250.
63. Agar MH, Wilson D. Drugmart: heroin epidemics as complex adaptive systems. *Complexity* 2002;7(5):44–52.
64. Perez P, Dray A, Ritter A, *et al.* SimDrug: a multi-agent system tackling the complexity of illicit drug markets. In: Batten D, editor. *Complex science for a complex world: exploring human ecosystems with agents*. Canberra, Australia: ANU Press; 2006. pp. 193–224.
65. Hoffer L, Bobashev G, Morris RJ. Researching a local heroin market as a complex adaptive system. *American Journal of Community Psychology* 2009;44(3-4):273–286.
66. Wood RK. Deterministic network interdiction. *Math Comput Model* 1993;17(2):1–18.
67. Washburn A, Wood K. Two-person zero-sum games for network interdiction. *Inst Oper Res Manag Sci* 1995;43(2):243–251.
68. Cormican KJ, Morton DP, Wood RK. Stochastic network interdiction. *Oper Res* 1998;46(2):184–197.
69. Caulkins JP, Padman R. Interdiction's impact on the structure and behavior of the export–import sector for illicit drugs. *Z Oper Res* 1993b;37:207–224.
70. Crawford GB, Reuter P, Isaacson K, *et al.* Simulation of adaptive response: a model of drug interdiction. Santa Monica (CA): RAND; 1988.
71. Bultmann R, Caulkins JP, Feichtinger G, *et al.* How should drug policy respond to market disruptions? *Contemporary drug problems*. 2008;35(283):371–395.
72. Bultmann R, Caulkins JP, Feichtinger G, *et al.* Modeling supply shocks in optimal control models of illicit drug consumption. In: Lirkov I, Margenov S, Wasniewski J, editors. *Large-scale scientific computing*. Heidelberg: Springer; 2008. pp. 285–292.
73. Weatherburn D, Jones C, Freeman K, *et al.* Supply control and harm reduction: lessons from the Australian heroin “drought”. *Addiction* 2002;98(1):83–91.
74. Crane BD, Rivolo AR, Comfort GC. An empirical examination of counterdrug interdiction program effectiveness. Virginia: Institute for Defense Analysis; 1997.
75. Cunningham JK, Liu LM. Impacts of federal ephedrine and pseudoephedrine regulations on methamphetamine-related hospital admissions. *Addiction* 2003;98:1229–1237.
76. Cunningham JK, Liu LM. Impacts of federal precursor chemical regulations on methamphetamine arrests. *Addiction* 2005;100(4):479–488.
77. Cunningham JK, Liu LM. Impact of methamphetamine precursor chemical regulation, a suppression policy, on the demand for drug treatment. *Soc Sci Med* 2008;66: 1463–1473.
78. Degenhardt L, Reuter P, Collins L, *et al.* Evaluating explanations of the Australian heroin drought. *Addiction* 2005;100:459–469.
79. Hawley M. Heroin shortage—the cause. *Platypus Mag* 2002;76:43–48.
80. Feichtinger G, Grienauer W, Tragler G. Optimal dynamic law enforcement. *Eur J Oper Res* 2002;141(1):58–69.
81. Fent T, Feichtinger G, Tragler G. A dynamic game of offending and law enforcement. *Int Game Theor Rev* 2002;4(1):71–89.
82. Tragler G, Caulkins JP, Feichtinger G. Optimal dynamic allocation of treatment and enforcement in illicit drug control. *Oper Res* 2001;49(3):352–362.
83. Zaric GS, Brandeau ML. Dynamic resource allocation for epidemic control in multiple populations. *Math Med Biol* 2002;19(4):235–255.
84. Long EF, Vaidya N, Brandeau ML. Controlling co-epidemics: analysis of HIV and tuberculosis infection dynamics. *Oper Res* 2008;54(6):1366–1381.
85. Kaplan EH. Needles that kill: modeling HIV transmission via shared drug injection equipment in shooting galleries. *Rev Infect Dis* 1989;11(2):289–298.
86. Kaplan EH. Probability models of needle exchange. *Oper Res* 1995;43(4):558–569.
87. Kretzschmar M, Wiessing LG. Modeling the spread of HIV in social networks of injecting drug users. *AIDS* 1998;12(7):801–811.
88. Pollack HA. Can we protect drug users from Hepatitis C? *J Pol Anal Manag* 2001;20(2):358–364.
89. Pollack HA. Methadone maintenance as HIV prevention: cost-effectiveness analysis. In: Kaplan EH, Brookmeyer R, editors. *Quantitative evaluation of HIV prevention programs*.

- New Haven (CT): Yale University Press; 2002. pp. 118–142.
90. Zaric GS, Brandeau ML, Barnett PG. Methadone maintenance and HIV prevention: a cost-effectiveness analysis. *Manag Sci* 2000;46(8):1013–1031.
 91. Richter A, Loomis B. Health and economic impacts of an HIV intervention in out-of-treatment substance abusers: evidence from a dynamic model. *Health Care Manag Sci* 2005;8(1):67–79.
 92. Caulkins JP, Feichtinger G, Gavrila C, *et al.* How much to spend on drug substitute programs? A dynamic cost–benefit analysis. *JOTA* 2006;128(2):279–294.
 93. Agar M, Wilson D. Drugmart: heroin epidemics as complex adaptive systems. *Complexity* 2002;7(5):44–52.
 94. Agar M. Agents in living color: towards emic agent-based models. *J Artif Soc Soc Simulat* 2005;8(1).
 95. Dray A, Mazerolle L, *et al.* Policing Australia’s heroin drought: using an agent-based model to simulate alternative outcomes. *J Exp Criminol* 2008;4:267–287.
 96. Perez P, Dray A, Ritter A, *et al.* SimDrug: exploring the complexity of illicit drug markets. In: Perez P, Batten D, editors. *Complex science for a complex world. Exploring human ecosystems with agents*. Canberra, Australia: ANU E Press; 2006. pp. 193–224.
 97. Hoffer L, Bobashev G, Morris RJ. Adaptive dynamics of a local heroin market and the effects of law enforcement busts. Paper Presented at the Drug Policy Modeling Workshop held in Conjunction with the 3rd Annual Meeting of the International Society for the Study of Drug Policy, Vienna, Austria, March 4th, 2009.
 98. Lee JS. Inferring adolescent social networks using partial ego-network substance use data [Ph.D. Dissertation], Pittsburgh: Department of Social and Decision Sciences, Carnegie Mellon University, 2008.

INTRODUCTION TO TABU SEARCH

J. COLE SMITH
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

JAMES J. COCHRAN
Department of Marketing and
Analysis, College of Business,
Louisiana Tech University,
Ruston, Louisiana

FOUNDATIONS AND MOTIVATION

This article regards optimization problems of the form

$$\text{minimize } f(\mathbf{x}), \text{ subject to } \mathbf{x} \in X, \quad (1)$$

where f is a (multivariate) objective function, $\mathbf{x}$ represents the solution vector, and X represents the problem's set of feasible solutions. For the sake of simplicity in this article, we assume that X is nonempty and that problem (1) has an optimal solution.

When problem (1) has certain special structures, specialized algorithms can be used to quickly obtain optimal solutions to the problem. However, especially when X contains *discreteness* restrictions on $\mathbf{x}$ (e.g., elements of $\mathbf{x}$ must be binary-valued, or represent a permutation of a set of integers), f is a nonconvex function, and/or X is not representable by (a reasonable number of) convex constraints, problem (1) may be quite difficult to solve to optimality. In fact, some of these problems are known to be *NP-hard*, and thus no algorithms that can solve these problems in polynomial time are believed to exist. (Informally, a polynomial-time algorithm is guaranteed to solve a problem in time proportional to a polynomial function of the “size” of the problem, usually given as the maximum of the number of problem elements and the number of binary bits

required to store the problem's data. See Papadimitriou [1] for more information on computational complexity, and the seminal work of Garey and Johnson [2] for the theory of NP-completeness.)

Many real-world optimization problems are indeed NP-hard, and reasonably sized instances are thus difficult or impossible to solve to optimality within reasonable computational limits. A common strategy is to relax one's insistence on obtaining an optimal solution to the problem, and instead concentrate on finding a solution that appears to be near-optimal. Hence, much research has been performed on the use of polynomial-time heuristic algorithms for solving difficult optimization problems. In particular, *metaheuristic* approaches describe general frameworks for heuristic methods that can be adapted to suit a wide array of diverse optimization problems. The topic of this article is one such metaheuristic technique called *tabu search*.

Tabu search first specifies an initial starting point for the search procedure, and then iteratively moves the current solution to a *neighboring solution* (one that is obtainable as a small perturbation from the current solution, as we will discuss in the next section). One particular search strategy, called *local search*, simply moves to a neighboring solution having the best-objective function value, until the current solution's objective function value is at least as good as all of its neighboring solutions' objective function values. This is also called *greedy search*. In this case, we say that the current solution is *locally optimal*, in the sense that it is optimal among solutions in its neighborhood. However, local optimal solutions may have objective function values that are far inferior to that of a *global optimal solution* (i.e., a solution $\mathbf{x}^*$ that optimizes problem (1)). The primary goal of tabu search is to avoid being stuck in local optima, while at the same time avoid getting stuck in a loop. We discuss the tabu search metaheuristic in the section titled “Tabu Search,” along with several issues that

arise when tailoring tabu search implementations on specific problems. We then briefly discuss a sample of applications on which tabu search has been effectively used in the section titled “Applications of Tabu Search.”

TABU SEARCH

We present the basic tabu search metaheuristic in the section titled “Basic Tabu Search Framework” and comment on its enhancements in the section titled “Extensions and Improvements.” We also refer the reader to excellent classical discussions of tabu search by Fred Glover [3–5], the manuscript by Glover and Laguna [6], and a tutorial treatment of the material by Gendreau [7]. To aid us in illustrating this technique, we consider the following two optimization problems.

Traveling Salesman Problem (TSP). Consider a set $V = \{1, \dots, n\}$ of nodes, with a distance measure of d_{ij} from node $i \in V$ to node $j \in V$. The TSP seeks a minimum-cost simple cycle (or tour) that starts and ends at node 1, and visits every node in V exactly once. We say that arc (i, j) is taken in the TSP tour if we travel directly from node i to node j in the tour. Letting P be the set of all arcs in a tour, the total cost of a tour is given by $\sum_{(i,j) \in P} d_{ij}$.

Parallel Job Scheduling Problem (PJSP). Consider a set $J = \{1, \dots, n\}$ of jobs that are processed on a set $M = \{1, \dots, m\}$ of machines, where job $j \in J$ takes p_{jk} hours to complete on machine $k \in M$ and only one job may be performed on a machine at a time. There are several possible objectives here; we will consider one of minimizing total completion time. The completion time of job j is simply the time at which it finishes processing, and the objective minimizes the sum of all job completion times.

Both of these problems are (strongly) NP-hard. The TSP is a classical domain of study for metaheuristic research; see Malek *et al.* [8] for an early study of serial and parallel tabu search algorithms for its

solution. Machine scheduling has also been a major focus of metaheuristic research in general; see Hübner and Glover [9] for an application of tabu search to multiprocessor scheduling.

Basic Tabu Search Framework

For our initial discussion, let us assume that finding feasible solutions to X is straightforward. tabu search begins by specifying an initial solution $\mathbf{x} \in X$ and characterizing a set of neighboring solutions $N(\mathbf{x}) \subseteq X$. Indeed, this step requires two of the most important decisions in designing a tabu search algorithm: (i) how to represent the decision variables $\mathbf{x}$ (i.e., the encoding of the solution), and (ii) how to construct the neighborhood $N(\mathbf{x})$. These two decisions are interrelated, and are typically the most important factors in determining how well the algorithm will perform.

Solution Encoding. In the context of the TSP and PJSP, note that either problem may be formulated as a mixed-integer programming (MIP) problem, and $\mathbf{x}$ could simply refer to the binary variables required in these formulations. For the problems at hand, however, there are more effective ways of problem representation, and we exploit the special constraint structure of the problems under investigation to devise a compact encoding for which neighboring solutions can easily be devised.

For instance, in the TSP, we use a *permutation encoding* on the integers $1, \dots, n$, that is, a sequence of integers such that every integer $1, \dots, n$ appears exactly once (in some order). We let x_j (the j th integer in a permutation solution $\mathbf{x}$) be the j th node visited in a TSP tour, where $x_1 = 1$ is fixed to indicate that the tour (arbitrarily) starts and ends at node 1. The last leg of the tour is then from the last node in this permutation back to node 1. It is easily seen that every permutation vector, starting with 1, represents a feasible solution. The computational problem is that there are $(n-1)!$ different permutations. More thought is perhaps needed for PJSP. In this case, it is possible to specify for each job the machine on which it is placed as well as the order in

which it is scheduled on the machine (e.g., x_j could refer to a pair $(k, \pi(k))$, where k represents the machine on which j is scheduled, and $\pi(k)$ represents the order in which the job is scheduled). However, this scheme would require the algorithm to ensure that a permutation sequence is enforced on each machine (e.g., two jobs cannot be processed first on the same machine). A well-known result in scheduling theory asserts that, given a minimum completion time objective and a set of jobs to be assigned to a single machine, an optimal schedule places jobs on the machine in nondecreasing order of their processing time. Hence, a more elegant (and effective) scheme is to simply let x_j represent the machine on which job j is scheduled. Given this information, a postprocessing step can then be performed to schedule the jobs on each machine (by sorting) and determine the objective function value of a candidate solution. This discussion in fact underscores the importance of exploring characteristics of a problem under study and using such knowledge within a tabu search (or indeed, any metaheuristic) algorithm, as emphasized in Gendreau [7].

Neighborhood. Given a solution encoding, the neighborhood of the solution $\mathbf{x}$ is determined by identifying an operation, or a set of operations, that slightly perturbs $\mathbf{x}$ while retaining feasibility to the overall problem. For instance, consider a solution $\hat{\mathbf{x}} = (1, 2, 5, 4, 3, 7, 6)$ to a seven-node TSP instance. Simply letting $N(\mathbf{x})$ be the set of all vectors that change only one element of $\mathbf{x}$ would result in infeasible solutions, as it would ruin the permutation property of $\mathbf{x}$. In the above example, if we change $\hat{x}_7$ to equal 7, then node 7 would be visited twice, and node 6 skipped. Instead, we could let $N(\mathbf{x})$ be a 2-opt neighborhood (which is very commonly used in permutation encodings) in which $\tilde{\mathbf{x}} \in N(\mathbf{x})$ if and only if swapping some pair of element values x_i and x_j in $\tilde{\mathbf{x}}$ yields $\mathbf{x}$. For instance, solution $\mathbf{x}^1 = (1, 2, 7, 4, 3, 5, 6)$ would be a 2-opt neighbor of $\hat{\mathbf{x}}$ in the above example.

However, once again we see that it is vital to employ knowledge of the problem under investigation. Note that $\mathbf{x}^1$ differs from $\hat{\mathbf{x}}$ by

four arcs: $\mathbf{x}^1$ uses arcs (2,7), (7,4), (3,5), and (5,6), replacing the arcs (2,5), (5,4), (3,7), and (7,6) in $\hat{\mathbf{x}}$. If the TSP is *symmetric*, where $d_{ij} = d_{ji}$ for all $i, j \in V$, then an alternative neighborhood picks two nodes to swap, but then reverses the entire subsequence in between those two nodes. Using the above example, this would yield $\mathbf{x}^2 = (1, 2, 7, 3, 4, 5, 6)$. Note that the perturbation between $\hat{\mathbf{x}}$ and $\mathbf{x}^2$ is now only two arcs: $\mathbf{x}^2$ uses arcs (2,7) and (5,6) in place of arcs (2,5) and (7,6) in $\hat{\mathbf{x}}$ (although some other arcs are traversed in the opposite direction as before). Hence, this 2-opt-reverse neighborhood, while being of the same size as a standard 2-opt neighborhood, contains solutions that are smaller perturbations of the original solution.

In general, the best results typically come from designing an encoding/neighborhood pair such that neighboring solutions $\tilde{\mathbf{x}} \in N(\mathbf{x})$ are very similar to $\mathbf{x}$, which motivates the 2-opt-reverse strategy for TSP. Note that in PJSP (unlike TSP), it is valid to let $N(\mathbf{x})$ be a set of 1-opt switches to $\mathbf{x}$, in which a single element of $\mathbf{x}$ changes its value from one machine assignment to another. However, the number of solutions in $N(\mathbf{x})$ should ideally be large enough to permit the search strategy to explore X as broadly as possible, while still limiting the overall complexity of the algorithm. Thus, it may also prove beneficial to include those vectors that are obtainable from 2-opt changes as well, in order to more broadly search X . At the same time it is of high importance to keep the size of the neighborhood limited, to ease the computational burden.

Tabu Tenure and Aspiration Criteria. Recall that local search chooses a starting point $\hat{\mathbf{x}} \in X$, moves to a best-objective neighbor in $\tilde{\mathbf{x}} \in N(\hat{\mathbf{x}})$, sets $\hat{\mathbf{x}} = \tilde{\mathbf{x}}$, and repeats until no solution in $N(\hat{\mathbf{x}})$ is strictly preferable to $\hat{\mathbf{x}}$, that is, until $\hat{\mathbf{x}}$ is a local optimum. In this sense, local search gets “stuck” at a local optimum (depending on the starting point). Tabu search avoids getting stuck in local optima by maintaining a limited memory of selected attributes of its last ℓ moves (where ℓ is a positive integer that is determined by the metaheuristic developer), and by restricting moves to neighboring solutions that contain

these attributes. Such moves are said to be tabu (or forbidden), and the limited memory used to store the selected move attributes is called the *tabu list*. The number of iterations that this information is stored for a given move is called the *tabu tenure*.

Determining how to select these attributes is an important consideration in designing tabu search algorithms. In the TSP example, it may be appropriate to remember the node indices that are swapped (e.g., nodes 5 and 7), and mark as tabu any move that swaps these nodes. An alternative would be to mark as tabu all moves containing node 5 or node 7. Alternative attributes of the move might be the set of new edges included by the move, or the set of old edges excluded (and/or both sets). For one-swap PJSP neighborhoods, we may record the job being switched and the machine from which it was switched. A separate tabu list may be stored for 2-opt changes as well, which could be implemented as multiple tabu lists, one for each type of neighborhood perturbation (see e.g. the discussion in Chapter 2 of Talbi [10] on maximum independent set problems).

An important distinction must be made between tabu search and local search at this point. Tabu search algorithms continue to move to a neighbor having the best-objective function value, provided the corresponding move attributes are not on the tabu list, *even if this best-objective neighbor has an objective that is inferior to that of the current solution*. This feature allows tabu search to move the solution away from local optima where the objective function of the current solution can deteriorate.

In a simplistic tabu search implementation, all solutions reached by tabu moves would be forbidden (although we relax this assumption in our subsequent discussion of aspiration criteria). It is easy to see that soon, if the tabu tenure were long enough, all moves would become tabu. Therefore, we retain only the selected attributes of the ℓ moves in the tabu memory. If ℓ is too small, then the metaheuristic cycles more easily: The current solution will move away from a local optimum, but fall back toward this local optimum as the tabu status of the moves it has taken to escape the local optimum ends

too quickly. If ℓ is too large, then most of the reachable neighborhood is tabu, and the search is highly restricted. One tends to tune a tabu search algorithm to find an appropriate balance between these two concerns. It is also customary to assign a tabu tenure that is randomized inside an interval (e.g., between 10 and 15) every time some move attribute is given a new tabu tenure. This helps to avoid being stuck in longer loops, and also diversifies the search.

Still, one common device used in these algorithms is that of *aspiration criteria*. The tabu criteria might sometimes be overly restrictive, in that otherwise good moves would be classified as tabu. The purpose of an aspiration criterion is to identify such moves and temporarily override the tabu status. Typically, the simplest and most effective rule is to determine whether a tabu move leads to a solution that has a new best-objective function value. (Other aspiration criteria might be devised, and the reader is encouraged to consult Glover [6] for further details.)

Overall Procedure. Having specified a solution encoding and neighborhood set, the tabu attributes and associated tabu tenure, and the aspiration criteria, we can now state the overall tabu search procedure as follows:

- Step 1.* Choose an initial solution $\hat{\mathbf{x}} \in X$. Let the incumbent solution $\mathbf{x}^* = \hat{\mathbf{x}}$ with objective function value $f^* = f(\hat{\mathbf{x}})$. Set the iteration counter $t = 0$.
- Step 2.* If $t == \text{MAXNUM}$ (where MAXNUM is a prespecified maximum number of tabu search iterations), then terminate with solution $\mathbf{x}^*$. Otherwise, continue to Step 3.
- Step 3.* Compute $f(\mathbf{x}) - f(\hat{\mathbf{x}})$ for every $\mathbf{x} \in N(\hat{\mathbf{x}})$, typically by evaluating the objective function change due to moving from $\hat{\mathbf{x}}$ to $\mathbf{x}$, and set $S = N(\hat{\mathbf{x}})$. Continue to Step 4.
- Step 4.* Let $\tilde{\mathbf{x}}$ be a solution that minimizes $f(\tilde{\mathbf{x}}) - f(\mathbf{x})$ over $\mathbf{x} \in S$. If the move from $\hat{\mathbf{x}}$ to $\tilde{\mathbf{x}}$ is tabu and does not meet the aspiration criteria, then delete $\tilde{\mathbf{x}}$ from S and repeat Step 4. Otherwise, continue to Step 5.

Step 5. Set $\hat{\mathbf{x}} = \tilde{\mathbf{x}}$. If $f(\hat{\mathbf{x}}) < f^*$, then $\hat{\mathbf{x}}$ becomes the new incumbent solution: Set $\mathbf{x}^* = \hat{\mathbf{x}}$ and $f^* = f(\hat{\mathbf{x}})$. The tabu memory is updated with the attributes and selected tabu tenure for the current move. In either case, increment t by one and return to Step 2.

Note that in Step 5, we remember the best solution found so far in the algorithm, and in Step 2, we report the incumbent solution as the final answer, rather than the last answer obtained. This is because the objectives encountered by tabu search are nonmonotonic, and hence it is quite possible that tabu search encounters its best solution early in the search.

As for termination criteria, we specify a very basic limit on the total number of iterations (solution moves) made throughout the algorithm. Several other criteria are possible, such as the maximum number of moves made without finding a new best solution in Step 5, or a time limit.

Extensions and Improvements

As this article is an introductory-level treatment of tabu search, we have only presented the foundational concepts behind the metaheuristic, leaving the more advanced notions for future reading. Here, we briefly discuss several considerations in implementing effective tabu search algorithms. More advanced treatments of this area can be found in Gendreau [11], Glover and Laguna [6], and Talbi [10].

Computational Techniques. In Step 3 of the metaheuristic description given above, we compute the change in the objective function value, $f(\mathbf{x}) - f(\hat{\mathbf{x}})$ (as opposed to recomputing $f(\mathbf{x})$ from scratch) based on the knowledge of the problem structure, in order to reduce computational effort. For instance, using a 2-opt-reverse neighborhood for a symmetric TSP, suppose that we know the objective function value $f(\hat{\mathbf{x}})$ for a current solution $\hat{\mathbf{x}}$. We can compute the objective function value of a neighboring solution by subtracting the costs of the two broken arcs in $\mathbf{x}$,

and adding the costs of the two new arcs in the neighboring solution. Each new objective function computation thus requires only four floating-point operations (or eight in an asymmetric TSP with 2-opt neighborhoods), rather than the n operations that would be required if $f(\tilde{\mathbf{x}})$ were computed from scratch for each $\tilde{\mathbf{x}} \in N(\hat{\mathbf{x}})$. (In fact, the effort here can be further reduced by exploring the neighborhood in a special order.)

Another computational consideration may arise if the number of solutions in the neighborhood is prohibitively large. In this case, it may be advantageous to explore only a subset of the neighboring solution space rather than enumerating the entire neighborhood. The proportion of neighborhood solutions evaluated can be dynamically changed throughout the algorithm, depending on the quality of solutions identified.

General Constraints. The TSP and PJSP discussed above have special constraint structures that allow solutions to be easily encoded, with their feasibility guaranteed. These structures are not necessarily present in all applications, though. Rather than explicitly enforcing the feasibility of candidate solutions to satisfy these constraints, we can instead penalize the degree to which constraints are violated. If the penalties are large, then infeasible solutions are essentially prohibited by the algorithm, which can unduly restrict the extent of the search space explored. It may be more preferable to adjust the penalties for infeasibility, starting with moderate values for the penalty and increasing them gradually during the course of the algorithm (as recommended in Gendreau [7]).

Intensification. The practice of intensification is based on the intuition that, even if tabu search fails to identify an optimal solution, it may have identified several solutions that are near an optimal solution. By keeping memory of, say, the β best unique solutions encountered, it may be appropriate to return to these solutions after the tabu search algorithm has finished executing, and explore them more intensely.

Several other intensification strategies, most depending on the problem under investigation, are also presented in the literature. Gendreau [11] prescribes fixing certain aspects of these candidate solutions (e.g., arcs that frequently recur in near-optimal TSP solutions, or machine assignments for promising PJSP solutions), and Hertz *et al.* [12] recommend partitioning and limited enumeration methods using these candidate solutions as starting points.

Diversification. The aim of diversification is to induce the search process to escape the currently explored part of the search region and explore other areas that may look promising. This feature of tabu search essentially employs a “long-term memory” by retaining information about previously visited solutions. This memory is often frequency based, as opposed to the recency-based tabu memory discussed above.

The simplest diversification strategy is to randomly restart the algorithm presented in the section titled “Basic Tabu Search Framework” by choosing different starting points $\hat{x} \in X$ in Step 1. Hence, the tabu search algorithm would essentially be run over several trials, retaining the best incumbent solution found over all of the trials. One may bias the selection of these initial solutions by encouraging them to be different in some way from prior starting solutions, and also from best solutions identified from the previous trials. In the generation of TSP tours, one might randomly generate a starting tour that attempts to avoid the inclusion of links that have appeared in these prior solutions. A PJSP diversification scheme may simply avoid assigning some jobs to machines on which they have commonly been assigned in previous solutions. Furthermore, for a case in which the search space X can be effectively partitioned into multiple sets $X_1, \dots, X_q$ (all of which are disjoint, with $X_1 \cup \dots \cup X_q = X$), Laguna *et al.* [13] prescribe a strategy in which starting solutions are generated roughly equally from among the q different sectors.

Soriano and Gendreau [14] call this method *restart diversification*, and draw a distinction between it and *continuous*

diversification, in which the objective of diversification is pursued during the evaluation of moves. Continuous diversification also attempts to avoid repeating portions of previous solutions that have been identified (e.g., links commonly appearing in a TSP solution or job-to-machine assignments occurring in a PSJP solution). However, it does so by penalizing these repeated portions within the ongoing evaluations of moves in the evaluated neighborhood. Hence, moves leading to solutions exhibiting substantial similarities to previously visited solutions are penalized, and the search process thus moves toward less frequently visited areas of the search space.

APPLICATIONS OF TABU SEARCH

Since tabu search is extremely effective in providing high-quality (and frequently optimal) solutions to complex problems, it has seen extensive application in a wide variety of problems since its creation by Glover [3–5] in the late 1980s. Our objective in this section is to foster in the reader an appreciation of the breadth of tabu search application, and not to discuss all published accounts of tabu search applications, so we focus on applications that have been published in the past few years.

Tabu search is frequently applied to complex scheduling problems. Korsvik *et al.* [15] and Korsvik and Fagerholt [16] use tabu search to increase revenues of shipping companies that transport bulk products by using space that is not contractually committed to ship spot cargoes. Hu *et al.* [17] also study shipyard operations, but focus on scheduling of a pipe-processing flowshop that comprises five stages (cutting, bending, welding preprocessing, argon-welding, and CO₂-welding), where each stage consists of identical parallel machines.

The vehicle routing problem (VRP) is a notoriously difficult combinatorial optimization problem that seeks a set of tours in a network that collectively service a set of nodes, for example, by picking up (or dropping off) items located at these nodes. Capacitated versions of the VRP include data regarding

the size of items to be picked up at each node and vehicle capacities that limit the total size of items that a vehicle can pick up. Gendreau *et al.* [18] devise one of the first tabu search approaches to solving the VRP with vehicle capacity and route length restrictions. In their implementation, neighboring solutions are created by swapping a single vertex from its current route to a new route. Oppen and Løkketangen [19] employ tabu search for solving a livestock collection problem, in which vehicle loading is subject not only to capacity limits, but also to restrictions on combinations of animals that can be loaded onto the same vehicle. The authors also consider a number of other real-world considerations in their model and demonstrate that their solutions are significantly superior to those currently in use. Zachariadis *et al.* [20] consider the case in which items are associated with two-dimensional sizes, and these items must be packed onto a vehicle's loading surface. Hence, the problem not only partitions the nodes into routes that are serviced by vehicles but also specifies the manner in which items are packed onto the vehicles. Cordeau *et al.* [21] provide a simple and effective tabu search algorithm for the VRP, in which each node is associated with a time window during which it can be visited by a vehicle. See also the two-part survey in Bräysy and Gendreau [22,23] for a more recent catalog of local search and meta-heuristic algorithms for the VRP with time windows.

Finance and capital investing are other related areas that are rife with applications of tabu search. A prime example is the work of Aldaihani and Al-Deehani [24], who focus their efforts on balancing risk and return when building a portfolio of investments available through the Kuwait stock exchange. One of their models aims to maximize expected return while maintaining a desired maximum level of risk, and their other model minimizes portfolio correlation while maintaining expected return at or above an acceptable minimum level. A mean-risk multistage capacity investment problem with irreversibility, lumpiness, and economies of scale in capacity costs is considered by Claro and Sousa [25], who use

conditional value-at-risk as their risk measure. These authors also solve this problem with a hybrid approach that combines tabu search and variable neighborhood search.

Network design is yet another area in which tabu search has been frequently employed. Pedersen *et al.* [26] develop a generic model for service network design (which includes asset positioning and utilization through constraints on asset availability at terminals) and solve both arc- and cycle-based formulations for their model with tabu search. Easwaran and Üster [27] apply two tabu search heuristics (sequential and random neighborhood search procedures) to a network design problem in a multiproduct closed-loop supply chain that incorporates remanufacturing facilities and finite-capacity manufacturing, distribution, and collection facilities that serve a set of retailers. The authors' objective is to determine the locations of the collection centers, locations of the remanufacturing facilities, and the integrated forward and reverse flows that will minimize the total cost of facility location, processing, and transportation associated with forward and reverse flows in the network. Yamada *et al.* [28] use tabu search in combination with other approaches to aid in strategic transport planning (particularly in freight terminal development and inter-regional freight transport network design) through the design of multimodal freight transport networks that facilitate regional and/or national economic development while reducing negative environmental impacts. Dengiz *et al.* apply each tabu search, genetic algorithms, and simulated annealing [29] to a different type of network—a back propagation neural network—to train the network and overcome the back propagation neural network's slow convergence rate and the tendency to converge to a local optimum.

Tabu search has also been applied to health care problems in several instances. The problem of composing medical crews in a manner that maximizes the equity and efficiency of the care provided by the crews was addressed by Aringhieri [30] with tabu search. In another health care application, Hertz and Lahrichi [31] use tabu search to assign patients to nurses for home care

services in a manner that balances the workload of the nurses while avoiding long travels for patient visits. Beaudry *et al.* [32] solve the problem of providing efficient and timely transport service to patients over several locations in a hospital campus, addressing the dynamic nature of transportation requests and varying priorities of the requests in such systems. The authors use a two-phase approach in which a simple insertion scheme is used to generate a feasible solution in the first phase and tabu search is used to refine and improve this feasible solution in the second phase. Erdoğan *et al.* [33] use tabu search to schedule ambulance crews to maximize coverage over a planning horizon.

The examples described above do not convey the breadth of application of tabu search. Gelogullari and Logendran [34] address flow-shop group-scheduling problems that are typically encountered in printed circuit board assembly. These authors specifically formulate their model to account for characteristics inherent to group-scheduling problems common in both electronics manufacturing and other production environments. Samarghandi and Eshghi [35] apply tabu search to a single-row facility layout problem (SRFLP) to linearly place a given number of rectangular facilities in a manner that minimizes the total cost associated with their known or projected interactions. Berman *et al.* [36] also apply tabu search to facility location, but consider the problem of finding the p locations for p facilities that minimize the maximum load associated with any individual facility (as a means of avoiding large discrepancies in loads associated with the facilities).

A multiorder processing procedure in a just-in-time (JIT) environment is modeled as a cost-based job shop problem, and work-in-process holding cost of half-finished orders, inventory holding cost of finished orders, and backorder cost of unfulfilled orders are minimized by Zhu *et al.* [37] through a modified tabu search method. In an effort to improve the operational efficiency of seaports and develop an analytical tool for yard operation planning, Wong and Kozan [38] use tabu search to solve a model that

incorporates relationships between quay cranes, yard machines, and container storage locations in a multiberth and multiship environment. Brusco and Steinley [39] apply neighborhood search heuristics to the statistical problem of selecting promising variable subsets for polynomial regression. Hertz *et al.* [40] report encouraging results in their application of tabu search to the auto rental company inventory management problem that was the ROADEF'99 international challenge. Mukherjee and Ray [41] search for optimal path conditions for multistage and multiresponse grinding in mass-scale manufacturing; they first consider each stage to be independent, then consider all stages as a single system, and solve both models with tabu search. Bouchard *et al.* [42] propose a specialized tabu search algorithm for finding an edge coloring with a minimum deficiency for a given graph.

As we have already seen, tabu search is often used in conjunction with other algorithmic approaches. Another recent example of this strategy includes the method of Lin and Ying [43], who develop a hybrid of tabu search and simulated annealing to minimize the makespan of the nonpermutation flow-shop scheduling problem. Rao and Shyju [44] use an algorithm that combines desirable qualities of simulated annealing and tabu search to solve multiobjective combinatorial optimization problems; they evaluate their algorithm on stacking sequence optimization problems for hybrid fiber-reinforced composite plate, cylindrical shell, and pressure vessel problems.

Finally, several authors have used tabu search to solve complex problems dealing with integrated functions or systems. Caramia and Giordani [45] propose an approach to managing the modes of task assignment and scheduling in an integrated manner. They formulate the problem as a single-mode task scheduling problem, model it as a graph interval T -coloring, and solve it using a tabu search approach. Tang *et al.* [46] examine a problem that deals with coordination of scheduling production and transportation in a steelmaking shop, and consider systems in which (i) there is a converter at the steelmaking operation and

a refining furnace at the refining operation and (ii) jobs are processed in identical parallel converters and then transported from an individual converter by dedicated trolley to the next operations. The integrated machine allocation and layout problem, that is, simultaneously assigning a set of machines to locations and product flows to machines in a manner that minimizes total material handling costs, is solved using tabu search by Jaramillo and McKendall [47]. Giallombardo *et al.* [48] integrate the berth allocation problem and quay crane assignment problem for container terminals. Their objective is to maximize the total value of chosen quay crane profiles while minimizing housekeeping costs generated by transshipment flows between ships, and they solve their problem with an algorithm that combines tabu search with classical mathematical programming techniques.

REFERENCES

1. Papadimitriou CH. Computational complexity. Reading (MA): Addison-Wesley; 1994.
2. Garey MR, Johnson DS. Computers and intractability: a guide to the theory of NP-completeness. San Francisco (CA): W. H. Freeman & Company; 1979.
3. Glover F. Future paths for integer programming and links to artificial intelligence. *Comput Oper Res* 1986;5:533–549.
4. Glover F. Tabu search—part I. *ORSA J Comput* 1989;1:190–206.
5. Glover F. Tabu search—part II. *ORSA J Comput* 1990;2:4–32.
6. Glover F, Laguna M. Tabu search. Norwell (MA): Kluwer Academic Publishers; 1997.
7. Gendreau M. An introduction to tabu search. In: Glover F, Kochenberger GA, editors. Volume 57, Handbook of metaheuristics. Norwell (MA): Kluwer Academic Publishers; 2003.pp. 37–54.
8. Malek M, Guruswamy M, Pandya M. Serial and parallel simulated annealing and tabu search algorithms for the traveling salesman problem. *Ann Oper Res* 1990;21(1–4):59–84.
9. Hübscher R, Glover F. Applying tabu search with influential diversification to multi-processor scheduling. *Comput Oper Res* 1994;8:877–884.
10. Talbi E-G. Metaheuristics: from design to implementation. Hoboken (NJ): John Wiley and Sons; 2009.
11. Gendreau M. Recent advances in tabu search. In: Ribeiro CC, Hansen P, editors. Essays and surveys in metaheuristics. Norwell (MA): Kluwer Academic Publishers; 2002.pp. 369–377.
12. Hertz A, Taillard ED, de Werra D. Tabu search. In: Aarts E, Lenstra JK, editors. Local search in combinatorial optimization. Princeton (NJ): Princeton University Press; 2003. Chapter 5.pp. 121–136.
13. Laguna M, Marti R, Campos V. Intensification and diversification with elite tabu search solutions for the linear ordering problem. *Comput Oper Res* 1999;26(12):1217–1230.
14. Soriano P, Gendreau M. Diversification strategies in tabu search algorithms for the maximum clique problems. *Ann Oper Res* 1996;63:189–207.
15. Korsvik JE, Fagerholt K, Laporte G. A tabu search heuristic for ship routing and scheduling. *J Oper Res Soc* 2010;61:594–603.
16. Korsvik JE, Fagerholt K. A tabu search heuristic for ship routing and scheduling with flexible cargo quantities. *J Heuristics* 2010;16(2):117–137.
17. Hu X, Bao J, Jin Y. A tabu search algorithm for a pipe-processing flowshop scheduling problem minimizing total tardiness in a shipyard. *Asia Pac J Oper Res* 2009;26(6):817–829.
18. Gendreau M, Hertz A, Laporte G. A tabu search heuristic for the vehicle routing problem. *Manage Sci* 1994;40(10):1276–1290.
19. Oppen J, Løkketangen A. A tabu search approach to the livestock collection problem. *Comput Oper Res* 2008;35(10):3213–3229.
20. Zachariadis EE, Tarantilis CD, Kiranoudis CT. A guided tabu search for the vehicle routing problem with two-dimensional loading constraints. *Eur J Oper Res* 2009;195(3):729–743.
21. Cordeau J-F, Laporte G, Mercier A. A unified tabu search heuristic for vehicle routing problems with time windows. *J Oper Res Soc* 2001;52(8):928–936.
22. Bräysy O, Gendreau M. Vehicle routing problem with time windows, part I: route construction and local search algorithms. *Transp Sci* 2005;39(1):104–118.
23. Bräysy O, Gendreau M. Vehicle routing problem with time windows, part II: Metaheuristics. *Transp Sci* 2005;39(1):119–139.

24. Aldaihani M, Al-Deehani TM. Mathematical models and a tabu search for the portfolio management problem in the Kuwait stock exchange. *Int J Oper Res* 2010;7(4): 445–462.
25. Claro J, Sousa JP. A multiobjective metaheuristic for a mean-risk multistage capacity investment problem. *J Heuristics* 2010;16(1):85–115.
26. Pedersen MB, Crainic TG, Madsen OBG. Models and tabu search metaheuristics for service network design with asset-balance requirements. *Transp Sci* 2009;43(2):158–177.
27. Easwaran G, Üster H. Tabu search and Benders decomposition approaches for a capacitated closed-loop supply chain network design problem. *Transp Sci* 2009;43(3): 301–320.
28. Yamada T, Russ BF, Castro J, *et al.* Designing multimodal freight transport networks: a heuristic approach and applications. *Transp Sci* 2009;43(2):129–143.
29. Dengiz B, Alabas-Uslu C, Dengiz O. A tabu search algorithm for the training of neural networks. *J Oper Res Soc* 2009;60:282–291.
30. Aringhieri R. Composing medical crews with equity and efficiency. *Cent Eur J Oper Res* 2009;17(3):343–357.
31. Hertz A, Lahrichi N. A patient assignment algorithm for home care services. *J Oper Res Soc* 2009;60(4):481–495.
32. Beaudry A, Laporte G, Melo T, *et al.* Dynamic transportation of patients in hospitals. *OR Spectr* 2010;32(1):77–107.
33. Erdoğan G, Erkut E, Ingolfsson A, *et al.* Scheduling ambulance crews for maximum coverage. *J Oper Res Soc* 2010;61:543–550.
34. Gelogullari CA, Logendran R. Group-scheduling problems in electronics manufacturing. *J Scheduling* 2010;13(2):177–202.
35. Samarghandi H, Eshghi K. An efficient tabu algorithm for the single row facility layout problem. *Eur J Oper Res* 2010;205(1): 98–105.
36. Berman O, Drezner Z, Tamir A, *et al.* Optimal location with equitable loads. *Ann Oper Res* 2009;167:307–325.
37. Zhu ZC, Ng KM, Ong HL. A modified tabu search algorithm for cost-based job shop problem. *J Oper Res Soc* 2010;61(4):611–619.
38. Wong A, Kozan E. Optimization of container process at seaport terminals. *J Oper Res Soc* 2010;61(4):658–665.
39. Brusco MJ, Steinley D. Neighborhood search heuristics for selecting hierarchically well-formulated subsets in polynomial regression. *Nav Res Log* 2010;57(1):33–44.
40. Hertz A, Schindl D, Zufferey N. A solution method for a car fleet management problem with maintenance constraints. *J Heuristics* 2009;15(5):425–450.
41. Mukherjee I, Ray PK. Quality improvement of multistage and multi-response grinding processes: an insight into two different methodologies for parameter optimisation. *Int J Prod Qual Manage* 2009;4(5–6):613–643.
42. Bouchard M, Hertz A, Desaulniers G. Lower bounds and a tabu search algorithm for the minimum deficiency problem. *J Comb Optim* 2009;17(2):168–191.
43. Lin S-W, Ying K-C. Applying a hybrid simulated annealing and tabu search approach to non-permutation flowshop scheduling problems. *Int J Prod Res* 2009;47(5):1411–1424.
44. Rao ARM, Shyju PP. A meta-heuristic algorithm for multi-objective optimal design of hybrid laminate composite structures. *Comput Aided Civ Infrastruct Eng* 2010;25(3):149–170.
45. Caramia M, Giordani S. A new approach for scheduling independent tasks with multiple modes. *J Heuristics* 2009;15(4):313–329.
46. Tang L, Guan J, Hu G. Steelmaking and refining coordinated scheduling problem with waiting time and transportation consideration. *Comput Ind Eng* 2010;58(2):239–248.
47. Jaramillo JR, McKendall AR. Metaheuristics for the integrated machine allocation and layout problem (IMALP). *Int J Oper Res* 2010;7(1):74–89.
48. Giallombardo G, Moccia L, Salani M, *et al.* Modeling and solving the tactical berth allocation problem. *Transp Res Part B* 2010;44(2):232–245.

TAKE-BACK LEGISLATION AND ITS IMPACT ON CLOSED-LOOP SUPPLY CHAINS

ATALAY ATASU
College of Management, Georgia
Institute of Technology, Atlanta,
Georgia

TAMER BOYACI
Desautels Faculty of Management,
McGill University, Montreal,
Quebec Canada

INTRODUCTION

In response to increased pressure from the public at large, as well as green organizations and the NGOs, governments around the globe have started to enact directives and pass legislation to tighten the control on waste generation and to reduce environmental damages. There are several types of environmental legislation that are currently in place or being planned, which differ in terms of the product life cycle phase the legislation is mostly concerned with. For instance, the EU's Framework Directive (2005/32/EC) on the eco-design of energy using products (EuP) regulates product design, the RoHS Directive (2002/95/EC) of the EU restricts the use of certain material and substances in production, The Basel Convention directs the export of recovered (hazardous) material and disposal, and the Waste Electrical and Electronics (WEEE) Directive of EU (Directive 2003/108/EC) regulates the take-back and recovery of end-of-life products.

Among these, take-back legislation arguably has the most comprehensive impact on closed-loop supply chains (CLSCs), and therefore constitutes the focus of this article. Our objective is to provide an overview of some of these regulations that deal with the end-of-life treatment of used products and discuss their impact on CLSCs (see **Remanufacturing** for further discussion).

There are many different types of CLSCs in practice, and these structural differences impact the nature of both forward and reverse flow of products. In this article, we utilize a generic CLSC structure, where the closing of the loop occurs after the *end-of-life* phase. Specifically, we consider five phases, representing different stages of the life cycle of the product. These are the *production*, *distribution*, *use*, *collection*, and *recovery* stages. Thus, the main stakeholders in our analysis are the *producers*, *distributor/retailers*, *end user/consumers*, *collectors*, *remanufacturers*, and *recycler/processors*. We remark that we use these terms broadly. For example, the producer refers to original equipment manufacturers (OEMs) and their entire supply chains that eventually *manufacture* or *import* new products. Logistic firms are typically involved with the transportation of new products, as well as the collection of used products from end users and their transportation for recovery operations. Likewise, municipalities, local, and national governments play crucial roles in setting up the infrastructure for collection and recovery, as well as imposing regulations that impact the operations of the CLSC.

EXISTING TAKE-BACK LEGISLATION IMPLEMENTATIONS

While all existing take-back legislations have a common goal, that is, reducing the waste generated from end-of-life products, implementations significantly differ from one country to another [1,2]. Such differences naturally result in different effects on the involved stakeholders. Below we list some important factors that could have a bearing on the impact of take-back legislation on these stakeholders.

Policy Tools

An important difference between alternative take-back legislation implementations comes from the policy tools chosen. For instance,

The European Commission has chosen collection, recycling, and recovery¹ targets as a policy tool for the end-of-life management of e-waste and end-of-life vehicles (ELVs). The WEEE Directive (2003/108/EC) of the European Commission requires Member States to ensure collection of 4 kg of such waste per capita per year starting from December 31, 2006, free of charge to consumers. The new proposal, made in December 2008, calls for 65% collection rate target (by weight) to be reached by 2016. Producers are responsible for collecting and processing end-of-life products to meet regulatory targets (percentages by weight) on product recycling or recovery. These targets differ for each category of waste. For example, in the case of large household appliances the targets are 75% for recycling and 80% for recovery, whereas for small household equipment, tools, and medical devices the rates are 50% and 70%, respectively. Similarly, the ELV Directive of the European Commission (2000/53/EC) mandates free take-back of ELVs from the last owner, with target rates of 80% for reuse and recycling and 85% for recovery (to increase to 85% and 95% by 2015, respectively).

On the other hand, The State of California [3] has chosen advance recycling fees (ARFs) for the management of e-waste. The Californian legislation requires all consumers to pay a nonrefundable fee when buying a new or refurbished product that contains a cathode ray tube (CRT) monitor or an LCD screen. The fee is used by local governments to pay authorized electronic waste collectors and recyclers.

Other policy tools, such as recycling fees (per sold or per collected unit), are also used in different parts of the world, for example, in Japan or in Taiwan (see Ref. 4 for further discussion).

¹In our discussion, *reuse* means any operation by which components of used products are used for the same purpose for which they were conceived. *Recycling* means the reprocessing in a production process of the waste materials for the original purpose or for other purposes but excluding energy recovery. *Recovery* includes both recycling and energy extraction.

Financial Responsibility

Some legislation hold the producers financially responsible for the costs of take-back (*producer pays* model), while others hold the end user or the purchaser of products for the costs of product take-back (*consumer pays* model). For instance, the producers are financially and physically responsible for the treatment of e-waste according to the European WEEE Directive, while in California consumers pay an ARF at the moment of purchase to cover for take-back related costs of certain electronics products. With a slight variation, the Specified Home Appliances Recycling Law (SHARL) of Japan [2] charges end users for the costs of take-back and recycling at the moment of disposal.

Collective versus Individual Responsibility

When producers are physically responsible for the costs of product take-back, it is important to know whether the legislation enforces collective responsibility among producers or whether producers are allowed to act individually. For instance, while the WEEE Directive states that each producer should be responsible for his own waste, the majority of Member States have their producers join *collective take-back schemes* (e.g., Walloon Region of Belgium; Denmark; France; Spain; and United Kingdom). This approach has been criticized by most producers [5] because it creates fairness concerns and lacks incentives for design improvements. Accordingly, some countries (e.g., Austria, Germany, Ireland, Italy, Sweden, and The Netherlands) allow producers to form their own *individual take-back systems* (www.iprworks.org).

Recycling Markets

Some legislation require that all producers recycle the end-of-life products through a monopolistic recycler. Other types of legislation allow competition in the recycling market. While the WEEE Directive states that each producer should be responsible for his own waste, some Member States, for example, Belgium, in EU have their producers join a *single take-back scheme*. Other countries, for example, Austria, Germany,

and The Netherlands, allows producers to join one of multiple *competing take-back systems* [4]. The most important difference between monopolistic and competitive recycling market systems appears to be on the cost side. It is argued that monopolistic take-back systems create higher recycling fees and hence higher costs to the producers [6,7].

Cost Sharing

If a collective producer responsibility system is employed, the cost allocation between producers is a very important concern. Sometimes the cost allocation can be based on market shares (e.g., the current system in WEEE) or it can be based on return shares (e.g., as allowed by the legislation in Maine (www.computertakeback.com) or Japan (<http://www.env.go.jp/en/laws/>)).

IMPACT OF LEGISLATION ON CLOSED-LOOP SUPPLY CHAIN STAKEHOLDERS

In this section, we provide a detailed account of how take-back regulations influence different stakeholders in CLSCs. Our discussion here is based on our understanding of the differences between previously discussed take-back legislation.

Consumers

Consumers represent one of the most crucial stakeholders in CLSCs. Take-back legislation imposes direct or indirect financial costs on consumers. In the context of ELVs, even in countries where free take-back is implemented, consumers are still required to bring their ELVs for dismantling. In some cases, the vehicles are not in a condition to be driven, imposing additional costs to the consumers, beyond the administrative and handling fees that may be charged.² Likewise, in the context of WEEE legislation that adopts consumer pays models, the consumers are charged fixed recycling fees

at the moment of purchase (e.g., California) or disposal (e.g., Japan). Even in implementations where take-back is free to consumers (e.g., in most of EU), the costs of collection and recovery are picked up by the producers, and it is not known as to what extent such costs are reflected on the consumers. Consequently, the most direct impact of take-back legislation on the consumers is economic in nature. Consumers are directly or indirectly paying for product take-back, meaning that the consumer surplus (CS) is likely to be lower with such legislation. An immediate impact of reduced surplus is reduced future consumption [6,8].

What is the benefit from such legislation to the consumers? The answer is probably the reduction in the environmental impact from production and disposal of hazardous goods and increased environmental sustainability due to greener designs. It is therefore no surprise that many consumer organizations and NGOs have been actively involved in the establishment of take-back legislation. Consequently, the trade-off that the consumers face is between short-term economic drawbacks and long-term environmental benefits. We remark, however, that the long-term environmental benefits are likely to result in economic benefits as well. This is because without take-back legislation, consumers will have to pay for increased taxes for toxic cleanups, water stream contamination, and increased landfill space. Take-back legislation is likely to please the consumers as long as consumers are aware of such potential future costs and they care sufficiently about the environment. Otherwise, the main impact of such legislation seems to be a reduction of the CS.

Distributors/Retailers

Distributors or retailers of new products (henceforth referred to as retailers for brevity) are ideal collection points for products to be recycled within the scope of take-back legislation. This is particularly relevant for WEEE products that are relatively less bulky, enabling collection and sorting at point of sale locations. As the closest contact point to end users, retailers can be instrumental in collecting fees under

²This extra cost and hassle is one of the main reasons for garaging or roadside abandonment of ELVs.

consumer pays models. Currently in Japan, end-of-use electronics products and the associated recycling fees are collected via retailers [4]. In California, even though retailers are not involved in the collection process, they contribute to the recycling system by collecting the ARFs [3]. In the context of ELVs, retailers' contribution is relatively insignificant, serving only as drop-off points in some cases (e.g., Belgium and Italy) [9].

From the retailer's perspective, take-back legislation may have both positive and negative impact. The negative impact stems from a possible demand reduction through an effective increase in the cost to the consumers. This follows our discussion of possible reduction in CS. In addition, additional physical and administrative costs might be incurred when retailers act as designated collection points. Some of these costs are offset by allowances made for the retailers to keep a portion of the collected recycling fees in *consumer pays* systems. Such allowances are present both in California and Japan WEEE legislation. On the less tangible but positive side, these activities can be seen as a value-adding service by the consumers. Another potential benefit to the retailers is bringing the consumers back to the retail stores. Especially, with small equipment that is easy to hand in at the retailers, consumers may end up in the retail stores which may trigger potential sales of other equipment. While there is no clear indication as to how frequently this occurs, it definitely deserves more investigation.

When the retailers do not act as designated collection points, their involvement with take-back operations is relatively limited. Hence, it is less clear what positive impact such legislation has. This is particularly relevant for *producer pays* systems with collective schemes, where collection is usually taken care of by third parties.

Collectors and Logistics Service Providers

As most take-back legislation come with target collection rates and encourage easy access of the collection facilities to end users, there is more business volume created for third party collectors. These are the independent dismantling and treatment firms

dealing with ELVs, municipalities, or public points that collect WEEE (financed by local authorities and/or producers). In the case of ELV, some countries have stricter requirements for the accessibility of the collection network, which has led to the installation of many new facilities. For example, the UK ELV regulation dictates that 75% of end users should be within 10 miles on average of the nearest facility, and no one should be more than 30 miles distance [10].

Another key player in the collection phase is third party logistics providers who carry out most of the transportation activities. Product take-back legislation is likely to create more business for them as well. Here, a distinction has to be made with respect to the nature of product that is collected. When products are collected with the purpose of recycling, there are usually certain aggregation centers from which the logistics providers access the material and transfer them to the recycling facilities. This is different from collection with the purpose of reusing products or components, where additional sorting and inspection is needed to make sure that valueless junk material is not transported.

An important factor in logistics business is economies of scale. When there is a single clearing house, a collective scheme for take-back, or when the producers form consortia to handle product take-back, greater economies of scale are likely to be observed. This may result in greater benefits for the logistics providers.

Existence of second-hand markets and restrictions put on the export of waste are likely to have a significant impact on the logistics companies that take part in CLSC operations. There is an active second-hand international market for used vehicles as well as components. Likewise, many electrical and electronic types of equipment, after being reconditioned, can be sent to developing countries where there is a market for them (see, for example, www.takebackmytv.com). With take-back legislation that restricts the use and cross-boundary transport of certain material, along with international regulations like the Basel Convention, it may not be allowed to export used products

out of the country, which limits the logistics business.³ This is likely to turn the global CLSC business into a local recycling business, in turn resulting in local logistics with lower margins instead of global logistics.

Producers

It is evident that producer responsibility measures embedded in take-back legislation ensure that producers face the most direct and significant consequences. For ELVs, producers bear most or all of the costs associated with take-back and recovery. Similarly WEEE legislation with producer pays implementation imposes the direct cost on the producers. In addition, mandatory (high) targets for recovery such as those in the EU have forced firms to go beyond the recycling of valuable material, which was for the most part economically viable, and recycle additional material and combust certain waste for heat recovery. This has further increased the costs of recovery and disposal [9]. Given the intensity of competition in global markets and low profit margins, take-back legislation imposes significant economic burden on producers, and may reduce overall profitability. Consequently, the efficiency of such legislation is extremely critical for the welfare of these sectors. For instance, in some EU countries (e.g., Belgium, see www.recupel.be) where a monopolistic, collective WEEE take-back scheme is employed, producers have been quite vocal about the high recovery fees they are charged and the fact that they cannot opt out of such systems. As a result, intensive lobbying has been observed for competitive take-back systems (see www.iprworks.org). A group of major electronics producers (Braun, Electrolux, HP, and Sony) has joined forces and established the European Recycling Platform (www.erp-recycling.org) in 2002 for this purpose, and by initiating competition between schemes, they have been able to reduce recycling costs significantly in countries such as Austria, Germany, and Ireland.

³We remark that despite these regulations, significant amount of WEEE and ELVs end up (illegally) in developing countries [2].

As the economics of recovery is the quintessential component of producer interest in take-back legislation, producers would also like to be able to influence it by designing products for recovery. A critical impact on the CLSCs is observed in this aspect. Product take-back legislation usually does not give enough incentive for reuse options [8,11]. First, since remanufactured/reconditioned products are usually targeted for developing markets and such legislation comes with export limitations, the possibility to access those markets is restricted. Furthermore, take-back legislation do not set specific targets for reuse, but rather combined targets for reuse and recycling/recovery. These rates can go up to 80–85% by weight. When such targets exist in the legislation, producers have minimal incentives to design products for reuse, because reuse is a higher-level recovery option and is considered to be quite different from recycling. Thus, product take-back legislation is likely to drive design incentives mostly for recycling and even just for the purpose of reducing the recovery costs.

Lack of standardization in take-back legislation and their implementation across different countries impose additional challenges for producers, in particular for those operating in multiple international markets. Such global producers have to deal with differing policies, financial and operational structures, and come to grips with immensely complex and diverse administrative structures. On the brighter side, the legislation induces producers operating in regions with no existing legislation to comply with these rules in order to export their products. This is likely to have positive spillover effects on their local markets as well.

Processors/Recyclers

The stakeholders that mostly benefit from the existence of take-back legislation are probably the recyclers and waste treatment and processing firms (henceforth referred to as recyclers). Clearly, take-back legislation with recycling and recovery targets brings additional business to these firms. This is evident also from the recent growth of the recycling industry. According to International Association of Electronics Recyclers,

the number of WEEE recyclers in the United States (500 recyclers and 19,000 employees in 2005) increased about 10% between 2003 and 2005 (www.iaer.org).

From the recycler's perspective, there is a growing stream of products that are being retired by end users, which have to be recycled. Before take-back legislation came to effect, recyclers would have preferred to process products or material that had high recovery value to have high profit margins. Current legislation (for both ELV and WEEE) allows recyclers to process products or new material with less recoverable value at sufficiently high margins, because there is someone paying for the cost of recycling.

Nevertheless, take-back legislation comes with certain duties and standards. Recyclers are often responsible for tracking the amount of processing they conduct and regularly report to administrating bodies. They also have to meet strict environmental standards relating to the storage, treatment, and recovery of end-of-life products to maintain authorized status. Furthermore, these standards are likely to tighten in the future. For example, while shredding is an acceptable form of recycling in the case of electronics, it may not be so in the future. Consequently, from the recyclers' point of view, investing in R&D and looking for superior recycling technologies may be essential for future profitability.

Remanufacturers

Take-back legislation has significant effect on third party remanufacturers. Given that most legislation comes with collection and recycling targets, producers have a vested interest in getting access to the used products with their own brand name. Consequently, it seems quite probable that third party access to end-of-life products will be harder in the future.

Furthermore, as we discussed earlier, exporting used products out of the country for secondary resale markets is getting more difficult due to raised awareness against exporting of waste. This may lead to reduced business opportunities for third party remanufacturers, especially in the electronics sector.

RESEARCH PROGRESS AND NEEDS

In our preceding discussion, we have highlighted the impact of take-back legislation on CLSCs based on our understanding of the existing legislation. We believe that there is a need to investigate these issues in further detail. The current academic research in this area is at its infancy. The existing operations management literature is mainly concerned about how such regulation impacts producers' operational choices and how the economics of recovery would have a bearing on their profitability.

There are two generic modeling approaches regarding product take-back and its impact on CLSCs. The first class of models is economic in nature, and includes different stakeholders and governments as decision makers. Atasu *et al.* [8] use this approach to take the first cut on how production economics would be affected by target-based take-back legislation. They discuss how recovery targets should be jointly based on recovery economics and environmental impact of products. Further, they show that intense competition leads to higher recycling and collection target requirements. Jacobs and Subramanian [12] take this one step further and discuss the possibility of sharing recovery costs within supply chains. They show that the sharing of take-back related costs between supply chain members (e.g., between producers and suppliers) can indeed increase supply chain profits, and under certain circumstances the social welfare.

The second class of models takes into account operational aspects as well. This part of the literature investigates how strategic and/or tactical decisions related to closed-loop operations, such as product design, CLSC structure, and coordination, would be affected by take-back legislation. Toyasaki *et al.* [6] model and compare collective (monopolistic) and individual (competitive) take-back schemes when collection, recycling, and treatment costs display scale economies. Assuming that target collection and recovery targets are met, they derive and compare equilibrium recycling fees, product prices, and recycler and producer profits.

They show that a win-win outcome can be achieved under the individual competitive scheme for the consumers (lower prices, higher surplus), producers (higher profits), and recyclers (higher profits). An exception to this occurs when the product substitutability is high, in which case the recyclers prefer the collective scheme since the nonprofit organization shields them from the highly competitive producer market.

Several studies investigate the impact of legislation, and in particular the minimum collection and recovery target rates, on CLSC network structures and costs. Hammond and Beullens [13] model a CLSC network of producers and consumers assuming diseconomies of scale in operational costs, and study the impact on prices and product flows. They find that minimum recovery targets can stimulate reverse supply chain activities, but are not sufficient to reduce virgin material and landfill use. Using a more detailed CLSC structure, Walther and Spengler [14] estimate the impact of WEEE Directive on reverse supply chain network and cost structures utilizing data from Germany. They show that transportation and disassembly costs increase (and hence marginal revenues decrease) significantly as the targets for recycling and recovery are elevated. In a similar vein, Quariguasi *et al.* [15] study both the economic and environmental impact in a multiple-objective framework taking European pulp and paper sector as the background, and show that mandatory recycling targets may deteriorate the efficient frontier. Krikke *et al.* [16], on the other hand, study the simultaneous CLSC network design and product design considering both economic and environmental effects. Applying their mathematical model on a data set for refrigerators, they find that the effects of minimum recovery rates are ambivalent; they reduce waste but increase energy use and costs. They emphasize the need to enhance product return rates and feasibility of recovery.

There are other papers that investigate the impact of take-back legislation on product design related issues. Zuidwijk and Krikke [11] explore two possible responses to legislation, product eco-design versus new recovery

processes, and show that investing in eco-design is preferable but has delayed effects. They conclude that legislation like the WEEE Directive in EU is appropriate in the sense that they stimulate recovery, but more incentives are needed to reward eco-design efforts. Atasu and Subramanian [17] study the environmental design incentives provided under take-back legislation and show that individual producer responsibility models create superior design incentives compared to collective producer responsibility models. Subramanian *et al.* [18] bring the supply chain considerations in the picture and investigate how take-back legislation drive product design and supply chain coordination. Plambeck and Wang [19], on the other hand, look at the impact of legislation on the rate of new product introductions in a competitive market. They find that the frequency of new product introduction depends on the specific form of producer responsibility in the legislation.

It is apparent from this brief review that there are large gaps in the current operations management literature. While some research streams consider the impact of take-back legislation on social welfare and producer profits, others deal with operational issues as well. A unifying research framework is needed and this framework should provide details on how the legislation impacts *all* of the major stakeholders. Research on the welfare implications of take-back legislation should take a broader perspective and identify the impact on these stakeholders. Similarly, the stream of research looking at the strategic and tactical decisions related to closed-loop operations should consider the specifics of the existing legislation, that is, what is relevant in practice, to identify the important research questions. Bridging the gap between the existing research streams is likely to create a better understanding of the impact of legislation and can result in improvements in the existing legislation from the perspectives of all stakeholders. We believe there are significant opportunities for applying operations research (OR) and management science (MS) methods and tools to resolve these challenges.

ILLUSTRATIVE MODELS

In this section, we illustrate how OR and MS models can be used to study the impact of legislation on closed-loop stakeholders. The first example investigates the economic and welfare implications of legislation, whereas the second example includes operational aspects as well.

The first model is based on Atasu *et al.* [8]. Economic models investigating welfare implications of take-back legislation usually include a typical utility structure for consumers, for example, consumer valuations (θ) for the product of interest are uniformly distributed between 0 and 1. Given the associated recycling fee (say ρ_c) and sales price of the product p , a consumer with valuation θ for the product gets utility $U = \theta - p - \rho_c$. In this case, the demand for the product can be developed as $d = 1 - p - \rho_c$, as a function of the sales price and take-back fees charged to consumers. As an illustrative example, consider an integrated producer (who represents distributors, manufacturers, collectors, and recyclers) who anticipates the consumers' response and tries to maximize his/her profit $\Pi = (1 - p - \rho_c)(p - \mu - \rho_m)$ with respect to the sales price p , where μ is the marginal production cost and ρ_m is the take-back fee charged to producers. The producer optimally chooses the profit maximizing price $p^* = (1 + \mu + \rho_m - \rho_c)/2$. At this optimal price, the sales quantity would be $q^* = (1 - \mu - \rho_m - \rho_c)/2$. In this case, the producer profit will be realized as $\Pi(p^*) = (1 - \mu - \rho_m - \rho_c)^2/4$ and the CS can be calculated as $CS = q^{*2}/2$. Given the producer's optimization problem, the social planner (government) tries to maximize the social welfare that consists of producer profit (Π), CS in the economy, and possible environmental externalities (E). Modeling the environmental externalities is a challenging task, because multiple factors, such as level of consumption, virgin or recycled material use, disposal amounts, recycling amounts, or recycling technologies, may be affecting the environmental impact of production and take-back. The typical trade-off the social planner faces is between the economic surplus (e.g., $\Pi + CS$) and the externalities,

where the economic surplus decreases and the environmental externalities (e.g., environmental benefits) increase with higher fees (ρ_c and ρ_m). Finally, the task of the social planner is to choose the optimal levels of ρ_m and ρ_c , such that the social welfare is maximized. For details of such a procedure under alternative cost structures, legislative implementations, or competitive settings, we refer the reader to Atasu *et al.* [8]. Note also that such models can be further developed to take into account the existence of disintegration in CLSCs, for example, how does the welfare change if producers, distributors, collectors, and recyclers are not integrated?

The second model is based on Toyasaki *et al.* [6] and compares the two prevalent implementation forms of take-back legislation—collective and individual responsibility—in a competitive setting. Suppose that the market consists of two (groups of) producers indexed by $j = 1, 2$ and two (groups of) recyclers, indexed by $i = A, B$. The demand in the product market is given by $d_j = \alpha - p_j + \beta p_{3-j}$, $j = 1, 2$ where α is the market size of each producer and $\beta < 1$ is the cross elasticity of demand. A given fraction τ of the sales are collected and sent for recycling and treatment. The unit recycling fee charged by each recycler is t_i , $i = A, B$. In the collective scheme, there is a nationwide nonprofit organization which is responsible for allocating the waste for recycling and treatment and charging the resulting costs to the producers. Suppose that the nonprofit organization employs a fixed allocation rule, where Recycler A gets a fraction of λ and Recycler B gets the remainder $1 - \lambda$ fraction. The nonprofit organization then calculates the average recycling fee $t = \lambda t_A + (1 - \lambda)t_B$ per unit collected to each producer. This ensures that the cost of recycling and treatment is borne by the producers according to market share and the nonprofit organization makes no profit. In the individual scheme, each producer makes an exclusive contract with a single recycler. Without loss of generality, Producer 1 bears the unit fee t_A set by Recycler A and Producer 2 bears the fee t_B set by Recycler B.

The strategic interactions between the CLSC stakeholders can be modeled using

a two-stage noncooperative game. At the second stage, producers simultaneously set the product prices to maximize their profits based on the recycling fees. Letting c denote the unit production cost (same for both producers), Producer j 's optimization problem in the collective scheme is as follows:

$$\begin{aligned} \max_{p_j} \pi_j = & (p_j - c)(\alpha - p_j + \beta p_{3-j}) \\ & - t\tau(\alpha - p_j + \beta p_{3-j}). \end{aligned}$$

For the individual scheme, Producer j 's problem is same as above except that the fee t is replaced with t_A for $j = 1$ and t_B for $j = 2$. For a given pair of recycling fees, π_j is concave in p_j , so there is a unique Nash equilibrium product prices $p_j(t_A, t_B)$. Based on these prices, the demand for each producer, and hence the amount of WEEE $w_i(t_A, t_B)$ destined to each recycler can be derived for both schemes. In the first stage of the game, the recyclers simultaneously select their unit fees taking the resulting producer prices and WEEE amounts into account. Let r denote the unit revenue earned from recycling a unit. The total logistic, processing, and treatment cost is taken as $(\eta w_i(t_A, t_B) - \theta w_i(t_A, t_B)^2)$, where η is the unit cost without economies of scale and θ is the economies of scale factor. Then Recycler i 's optimization problem can be formulated as

$$\begin{aligned} \max_{t_i} \pi_i = & (t_i + r)(w_i(t_A, t_B)) - [\eta w_i(t_A, t_B) \\ & - \theta w_i(t_A, t_B)^2]. \end{aligned}$$

Under mild conditions, π_i is concave in t_i , hence unique Nash equilibrium recycling fees (t_A^*, t_B^*) can be found for each scheme. The resulting equilibrium producer prices can be derived from the first stage game as $p_1^* = p_1(t_A^*, t_B^*)$ and $p_2^* = p_2(t_A^*, t_B^*)$. Comparing the equilibria, it is possible to derive conditions under which each scheme is preferable for each stakeholder. The model can be extended to the case where there is asymmetry in market sizes, operational costs, and economies of scale factors, as well as to the case where the recyclers directly compete for WEEE allocation in the recycling market. We refer the reader to Toyasaki *et al.* [6] for details.

CONCLUSION

It is evident that the take-back of end-of-life products is a vital instrument for environmental sustainability. The importance of this is borne out by the number of existing and planned take-back legislation. Interestingly, the organization and financing of these activities show disparity across different implementations, and there are diverse views regarding the preferred system. Likewise there are conflicts of interest among various stakeholders. These signify the importance of taking a holistic approach and addressing the impact of take-back legislation from a comprehensive, multiple-stakeholder- and multiple-objective perspective. To the best of our knowledge, existing literature has not yet accomplished this. The interactions between these stakeholders need to be identified and research on coordination between stakeholders is essential to determine the best case scenarios both from the industry and social points of view, considering the triple-bottom line impact—people, planet, and profits. In addition, a detailed investigation of product design implications of take-back legislation is also needed. While the existing legislation is usually local, we believe that its global impact should also be a concern for academics.

This article is aimed at providing an overview of existing take-back legislation, its impact on CLSCs, determining the pressing research issues, and illustrating how OR and MS tools can be applied to examine these research issues. We provided our perspective on the eminent effects of such legislation to the best of our knowledge, but we acknowledge that there surely exist other possible impacts that are not yet readily observable. We also realize that further research is needed to verify some of our arguments. This calls for a strong interaction with the industry and data-driven (empirical) research.

REFERENCES

1. Atasu A, Van Wassenhove LN. Environmental Legislation Regarding Product Take-back and Recovery. In: Ferguson ME,

- Souza GC, editors. Closed-loop supply chains: new developments to improve the Sustainability of business practices. Oxford: CRC Press, Taylor & Francis LLC; 2010. Forthcoming.
2. Kahhat R, Kim J, Xu M, *et al.* Exploring E-waste management systems in the United States. *Resour Conserv Recycl* 2008;52:955–964.
 3. CIWMB, California Integrated Waste Management Board. Electronic product management; 2004. Available at http://www.leginfo.ca.gov/pub/03-04/bill/sen/sb0001-0050/sb_50_bill_20040929_chaptered.pdf.
 4. Dempsey M, Van Rossem C, Lifset R, *et al.* Developing practical approaches for individual producer responsibility: industry report to the European Commission. 2008.
 5. Lifset R, Lindhqvist T. Producer responsibility at a turning point? *J Ind Ecol* 2008;12:(2): 144–147.
 6. Toyasaki F, Boyaci T, Verter V. An analysis of monopolistic and competitive take-back schemes for WEEE recycling. Working Paper, McGill University; 2009.
 7. Guilcher H. European recycling platform. INSEAD WEEE Directive Series Presentation; 2005.
 8. Atasu A, Van Wassenhove LN, Sarvary M. Efficient take-back legislation. *Prod Oper Manage* 2009;18:(3):243–258.
 9. Fergusson M. End of life vehicles (ELV) directive: an assessment of the current state of implementation by member states. Study for the European Parliament's Committee on environment, public health and food safety; 2007. Available at http://www.europarl.europa.eu/comparl/envi/pdf/externalexpertise/end_of_life_vehicles.pdf.
 10. DTI. Guidance notes on the end-of-life vehicles (producer responsibility) regulations 2005. Available at <http://www.berr.gov.uk/files/file30646.pdf>.
 11. Zuidwijk R, Krikke H. Strategic response to EEE returns: product eco-design or new recovery processes? *Eur J Oper Res* 2008;191:1206–1222.
 12. Jacobs B, Subramanian R. The Impacts of sharing responsibility for product recovery across the supply chain. Working Paper, Georgia Institute of Technology; 2008.
 13. Hammond D, Beullens P. Closed-loop supply chain network equilibrium under legislation. *Eur J Oper Res* 2006;183:895–908.
 14. Walther G, Spengler T. Impact of WEEE-directive on reverse logistics in Germany. *Int J Phys Distrib Logist Manage* 2005;35:(5): 337–361.
 15. Quariguasi Frota Neto J, Bloemhof-Ruwaard JM, Van Nunen JAEE, *et al.* Designing and evaluating sustainable logistics networks. *Int J Prod Econ* 2008;111:195–208.
 16. Krikke H, Bloemhof-Ruwaard J, Van Wassenhove LN. Concurrent product and closed-loop supply chain design with an application to refrigerators. *Int J Prod Res* 2003;41:(16): 3689–3719.
 17. Atasu A, Subramanian R. Competition under product take-back laws: individual or collective systems. Working Paper, Georgia Institute of Technology; 2009.
 18. Subramanian R, Gupta S, Talbot B. Product design and supply chain coordination under extended producer responsibility. *Prod Oper Manage* 2009;18:(3):259–277.
 19. Plambeck EL, Wang Q. Effects of E-waste regulation on new product introduction. *Management Science* 2009;55:(3): 333–347.

TEACHING ORMS/ANALYTICS WITH CASES

MEHMET A. BEGEN
SRINIVAS KRISHNAMOORTHY
JOHN WILSON
Richard Ivey School of Business,
University of Western Ontario,
Ontario, Canada

INTRODUCTION

Traditional methods for teaching Operations Research/Management Science (ORMS)/Analytics and Management Science/Operations Research have focused on the lecture model for delivery of content. This often gives rise to students being technique oriented and unable to apply the material in situations encountered outside the classroom. A major goal of the case-based approach is that students become able to identify top-level issues and apply models to situations not necessarily encountered in the classroom.

Case teaching can be implemented gradually in a traditional lecture-based *ORMS/Analytics* course; in fact, instructors can use cases in tandem with lectures. Cases can also be the core pedagogical methodology for a course.

THE CASE METHOD

The case method is an interactive style of learning. In traditional lectures, the instructor delivers a prepared talk or presentation, while the case method requires that the “talk time” be shared among the students and the instructor. The instructor’s role is to prompt students to express their opinions and conduct the discussion. Specific activities include initiating the discussion with initial questions posed to the class, selecting a student when a bunch of them raise their hands, listening to students’ comments, and responding to and amplifying comments. The

response may include a direct response to the student, probing the student for a clarification or even redirecting it to the class. A wrap up may also be required at the end of class. The case method applied to *ORMS/Analytics* poses some unique challenges. We explore those challenges in later sections but first we take a look at some articles that explore case teaching for *ORMS/Analytics*.

Many case instructors have presented their ideas on the use of cases for *OR/MS* and *Analytics* courses. Aggarwal and Khera [1] discuss the benefits of using short cases for teaching introductory management science courses. Highlighting the heterogeneity in quantitative skills among their MBA students, they argue that a case-based course works better than a traditional course. Furthermore, they believe that cases should be short and managerial in order to keep the student’s interest alive. Martin [2] supports the views of Aggarwal and Khera but argues that there is a group of business students who will benefit from pure OR in addition to managerial cases. He refers to them as model builders (as opposed to managers who may use models developed by others). Horner [3] interviews Peter Bell, Professor at the Richard Ivey School of Business, to understand the benefits of teaching OR/MS using case studies. The article is wide ranging in the perspective it offers on incorporating OR/MS cases in the business school curriculum. Bodily [4] describes how the Quantitative Analysis course at The Darden School of Business has been designed to integrate with the other MBA courses. Specifically, by using a case-based approach, the quantitative analysis teaching team at the Darden has managed to create a decision-based course rather than a “quant” course. This same principle has been adhered to while designing the quantitative analysis electives at the Darden. In similar spirit, Liberatore and Nydick [5] provide a description of their MBA Management Science course at the Villanova University. Bell and Haehling

von Lanzanauer [6] illustrate the benefits of exposing students to a real-life business problem by discussing their experience in the classroom with using a particular case study. Cochran [30] provides a detailed description of the case methodology that he adopted for teaching operational research to undergraduate business students. Hannibalsson [7] discusses how he has incorporated cases, games, and other active learning techniques in his operations course at the University of Iceland. Schmedders and Stephens [8] urge the development of new OR/MS cases that reflect current business problems. As part of his insights on teaching OR/MS to Executive MBA students, Bell [9] describes the benefits of using cases and offers specific examples of cases that he uses in the course. Long [10] shares his experience in teaching operations research using cases and presents his perspective on the requirements for successful implementation of this methodology. Cochran [11–13] provides detailed illustrations of how he has successfully used a particular case in the classroom. In his review of OR pedagogy, Cochran [12] enumerates the effectiveness of case teaching and also lists the different business schools that have adopted this approach. As part of his article on making classroom learning memorable, Cochran [14] discusses how his adoption of case-based teaching improved the student experience because of more personalized attention associated with the new teaching method. Moving from management science to other disciplines, Liang and Wang [15], Kim *et al.* [16], and Currie [17] provide perspectives on case teaching. Our goal in this article is to reinforce the ideas discussed in the above articles while adding some of our own perspective to the art of teaching ORMS/Analytics with cases.

There is empirical evidence that case teaching is more effective than lecture teaching in business education. Specifically, Bocker [18] has conducted a field study to confirm this hypothesis. Our goal in this article is to not dismiss lectures as an outdated method of teaching. We believe that cases and lectures can go hand in hand to offer students a variety of learning

experiences and instructors a variety of teaching experiences in the classroom.

Case-based teaching has spread from the traditional case-based schools of Ivey, Harvard [32], and Darden [31] to many other schools. Moreover, it is being implemented in a variety of ways, for example, a combination of case- and lecture-based teaching and mini cases. A number of universities have case publishing arms—prominent among them being the Ivey [33], the Harvard [32], and the Darden [31]. In addition, *INFORMS Transactions on Education* (Available at <http://www.informs.org/Pubs/ITE/>) publishes teaching cases online and INFORMS organizes an annual case competition (Available at <http://www.informs.org/Community/INFORM-ED/Activities/Case-Competition>) since 2000. Other sources include the ECCH (European Case Clearing House), (Available at <http://www.ecch.com/>) and *The Operations Management Education Review* (OMER) (Available at <http://www.neilsonjournals.com/OMER/>).

A CASE-BASED COURSE

It is certainly possible to use only a handful of cases in a predominantly lecture-based course. However, in what follows, we describe a case-based course. The comments and suggestions also apply if an instructor is simply using a few cases to illustrate and apply concepts from lectures. In this case, an instructor might decide to use an integrative case that requires students to use a number of different techniques to arrive at a decision or might decide to use cases that require knowledge of any a few techniques or both.

We suggest positioning a course as using the art of using data and models for decision making, with the emphasis being on decision making. For instance, such a course could be structured around “modules” such as

- Foundations of modeling;
- Models for sequential decisions;
- Data and decision making;
- Models for managing risk;
- Models for simultaneous decisions;
- Models for revenue management and pricing.

Each of these modules focuses on traditional OR/MS topics such as statistics, decision analysis, probability, simulation, and optimization. For instance, “Models for sequential decisions” refers to decision analysis. We prefer to use “business” names rather than traditional technical names for our modules. We find that students relate to our material better when there is a managerial implication, and this applies to module names as well. We have found that our course name at The Richard Ivey School of Business “Decision Making with Analytics” (DWA) can have more traction than traditional OR/MS names with both students and our other management colleagues. One model for covering the material is presented in Figure. 1. The first column refers to cases that require only one or two modeling techniques. Cases in the second column are more difficult and generally require more than one quantitative technique. The final column refers to complex integrative cases that very closely mirror the real word.

For examples of basic cases, we have ones that cover regression models for real estate price prediction, solution of newsvendor problems via simulation, product mix cases that require simulation. For instance, we can use the traditional case *Red Brand Canners* [19] to introduce linear programming. In addition, we use complex cases—especially for multiple regression—that use only one main technique. For example, the cases *The Professor Proposes* [20] and *Lanksink Appraisals* [21] are useful to tie a number of statistical

ideas together. At the final level, when the students have mastered the basic techniques, we use a number of complex cases in revenue management and the pricing of contracts and real options that not only test students’ analytical abilities but also open their minds to the applicability of the material. Some of our favorites here have been Enron Corporation’s Weather Derivatives(A) & (B) [22], Jaikumar Textiles Limited (A) & (B): The Nylon Division [23], Kimpton Hotels: Setting prices on Priceline(A), (B) & (C) [24], The Clonlara Hotel [25], and Seoul National Bank: The Chief Credit Officer’s Dilemma [26].

DWA CLASS IMPLEMENTATION

We implement interactive discussion-based case teaching. Our strategy is to let students learn by doing. The easiest way to make any class interactive is to simply ask students’ opinions on a topic rather than lecturing about one’s own ideas first. From time to time, when we feel necessary, we deliver minilectures in class to highlight important points, give an overview of the topic we are studying, or clarify concepts. For instance, we usually give a short introductory lecture on linear programming.

Teaching ORMS/Analytics interactively with case discussions may look difficult at first, that is, teaching a quantitative topic to students especially to ones who have no background in math, statistics, or any other quantitative fields. We start with a simple

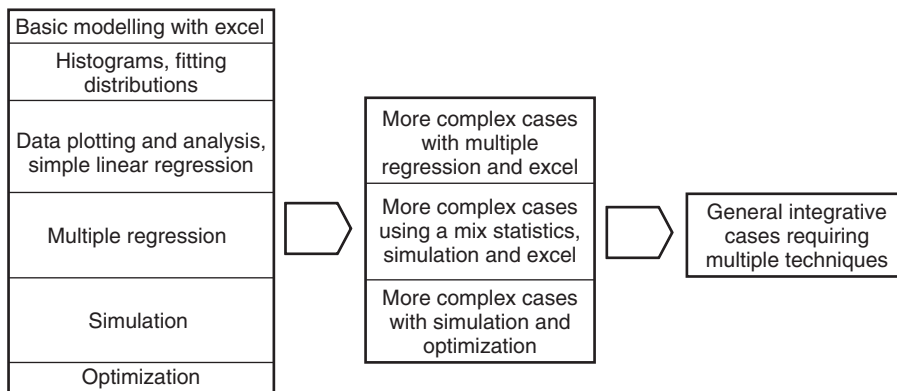


Figure 1. Case types.

case application whenever we introduce a new topic. For instance, consider regression. We give a short lecture on how to run a simple linear regression model—we take two variables, plot and fit a line. We assign introductory background readings from a textbook. We then use a case that requires the students to develop a simple linear regression model. Once this is done, we demonstrate the technique of adding extra variables and assign some more rudimentary readings. We do not comprehensively cover all the complexities of regression. (This is one of the trade-offs of the case method—depth in a technique is sometimes sacrificed so that students better understand how to apply the methods and contemplate top-level issues.) The next case is a step up in the level of challenge and could require the development of a multiple regression model. Students are required to peruse accompanying readings. We continue in this manner raising the difficulty level. This is illustrated in Table 1.

At this stage, students now have gained enough technical skills to tackle complex regression cases with regression.

When applicable and appropriate, we use related research papers or newspaper articles (e.g., a research paper relating CEOs birthdays to performance in Fortune 500 companies) to motivate the current topic (e.g., regression), show how it works in practice, and strengthen the overall understanding.

We sometimes invite guests (e.g., sometime from the organization in the case or an industry expert on the current topic) to provide additional insights and learning points.

Student Motivation

In order for students to take the process seriously, it is important that class contribution be one of the most important parts of the grade. It should be graded on an individual basis, and we have found that making it worth roughly one-third of the entire course mark is effective.

Before Class

Students are expected to attempt the case as part of their homework. There are assigned questions for each case. Student preparation involves more than simply reading the case.

Table 1. Teaching Framework with Cases and Minilectures

Order	Events	Content
1	Minilecture 1	A Y -variable (e.g., house prices) An X -variable (e.g., square footage) Plot variable in Excel Fit line on plot and obtain an equation for it Discuss the idea of an equation to predict house prices as a function of square footage Assign a reading on (simple linear) regression
2	Case	Simple case with two variables Class discussion and solution approaches Use of regression equation to solve the case and obtain insights Introduction of ideas on R -squared, residuals, and p -values
3	Minilecture 2	Demonstrate how to add new (multiple) X -variables Discuss why more than one X -variable may be needed: other factors, indicator variables Extend the example from Minilecture 1 to multiple variable regression Provide a reading on multiple regression
4	Case	Slightly more complex case with more variables Class discussion and solution approaches Discussion of R -squared, p -values, residual plots, and transformation of variables Use of regression equation for the case

A proper preparation includes perusing the case, preparing answers for the assignment questions, and finally, building and using a model to make a decision. It is very important that students attempt the case by themselves before talking to others—otherwise only limited learning will take place. After they complete their individual preparation, they discuss the case in a small group setting. This small (study) group usually consists of four to six students who come together before the class to exchange notes and ideas. The small group discussion serves as a dry run for the big group (class) discussion. The stakes and anxiety level are a lot less. Students are only within a small numbers of peers; there is no instructor present, and their performance is not graded. It is an opportunity to strengthen their understanding and preparation for the classroom discussion.

During Class

Start the class with a start-up question. In general, this should be a question that can be answered without any technical knowledge. This allows all students to feel that they can contribute irrespective of their backgrounds or trepidations about some of the later technical aspects of the case. A good example of a start-up question is “What specific decision needs to be made?” Instructors will find the answers to this question illuminating. For instance, we can use a product mix case to introduce linear programming. Here, we sometimes say “I will give you any specific numbers you want. Tell me what you want and how would you communicate this decision to those who are in charge of implementing the decision.” Many of the answers we get are along the lines of “Find the ... that optimizes ...” or “Choose values of ... that satisfy the constraints ...” None of these, of course, are satisfactory because they are related to a solution procedure and not to the final decision that needs to be communicated to workers who will actually carry out the decisions. This approach gets students to realize that the output of the case is ultimately a set of specific instructions, for example, take 10 pounds of ingredient A and 15 of ingredient B and make product and take 15 pounds of ingredient A and

3 of ingredient C and make product 2. It is remarkable how difficult this is for students of all backgrounds. This becomes one of the great advantages of the case method, being able to actually make and implement decisions rather than only knowing how to “solve” things.

It is important to carefully manage the first questions. Sometimes, there will be some students who have “solved” the case. If one of these students is chosen and care is not taken in managing the response, the instructor can find that there is not much more to discuss. So the instructor should be ready to limit students’ responses to very specific answers and may collect a number of answers before opening up the discussion around one of them.

Sometimes, the start-up question is one of the assignment questions. The student who is given the floor then provides his/her answer to the question. It is the instructor’s job to acknowledge the answer or probe for more details or even seek the opinion of other students. Sometimes, we make “cold calls,” that is, seek a response from a specific student. We find that this is a good way to ensure preparation because there is no advance warning of who could be selected. We hasten to add that the primary motivation for students to prepare or cases does not come from fear of embarrassment. It is the desire to perform under pressure under the scrutiny of the professor and fellow students. This also helps train student to be effective managers and implementers of OR/MS and Analytics. It does help that we have as much as 30% of the course grade based on class contribution.

Following the start-up question, we raise more detailed and specific questions to direct the discussion. We may use the whiteboard or overhead to summarize ideas, write down equations, or even draw pictures or graphs in order to add conceptual richness to the discussion. Each case requires building some sort of a model. This is typically a spreadsheet model; however, sometimes, it could be just a system of equations or graphs that can be done on the board. If the model can be done by hand (rarely the case) then we build it on the whiteboard or overhead in an interactive manner seeking the help and comments of students. On the other hand,

for a spreadsheet model, we pick a student who connects his/her laptop to the projector. The selected student then builds the model on his/her laptop with the help of other students. We moderate this interactive model building process. When necessary, we provide corrections to the model. Following the construction of the model, we use it to derive insights that are relevant to the problem being solved. The final step is to seek a decision from the class based on the insights provided by the model. Typically, we probe for a justification of any decision that student makes. In addition to the model output, students often bring in qualitative aspects that may affect the decision. Particularly relevant at this stage is the personal experience of a student who has worked in the industry or even the company itself. We end the class emphasizing the conclusions and learning points. In the next class, we may offer a brief review of the learning points of the previous case.

We find that the spreadsheet models are powerful for what-if analysis. Our students enjoy the insights that the output of a model provides. In the initial part of the course, students find it difficult to create a model or even conceptualize the trade-offs that a decision-making situation entails. However, this changes over the course of the semester. Students become more refined in their thinking and their ability to construct models.

We also allow room for students to share their own models. Sometimes, at the end of the class, we may ask a student to share his/her model if he/she adopted a different approach. We also allow students to post their developed models on the university's intranet so that each student in class can have access to it.

It can be important to have a clear and concise wrap-up for the case. Of necessity, the case method will involve lots of twist and turns and barking up the wrong tree. This is part of the learning process. However, many students have been conditioned to believe in a "right" solution. It can be very frustrating for them not to see a clear path to the conclusions. It is, however, important to emphasize that it is rarely possible to jump to a clean "solution." In addition, it is important to emphasize that the solution is only

as good as the inputs. How robust is the output to changes in the inputs? Maybe an input distribution is not normal or uniform or whatever. A discount rate is often used in making a decision. How sure is one of this number? What is the risk involved with a decision? Expected values for a one-shot decision can be very misleading if the variability is high. What are the consequences of not hitting the targets implied in a decision? What are the probabilities of missing a target?

A BASIC CASE

Regression analysis is very amenable to a case-based approach. We often start with a case that has time to complete assembly of a certain product as the dependant variable and the independent variables including the number of operations required per product, size of batch, priority class, and so on. An order has just been taken and the salesperson wants to give the client an idea of when the product will be ready.

"Top-Level" Questions

A reasonable starting question is "What number or numbers should be provided to the salesperson?" A number of answers can be collected from the class and written down. Some will want to provide a specific fixed number. Some will want to provide a range. This starts a discussion of risk and expected value and maybe standard deviation. Non-symmetric risk can become a discussion point: "Would you want to overestimate or underestimate the time you provide?" "What are the managerial consequences of underestimating or overestimation?"

Detailed Analysis: Input for Making the Decision

If your production manager tells you that only the number of operations is predictive, how would you go about giving a prediction? The answers from this are very illustrative of the thinking in the class. Many will try to group products in the data set with a similar number of operations and maybe just average them out. Some of the follow-up issues here

are what do we do when there are not many observations with the number of operations close to the product for which we are trying to predict? How do we deal with the issue of risk? Some will say plot the data.

Plotting the data leads to more questions. Many will now see that fitting a line might be the way to go.

At this stage, the question might be raised that other variables should be introduced, and this leads to a discussion of multiple regression.

Flow of the Case: Approach 1

Have the students read the case before but not do any formal quantitative analysis and are ready to discuss. After the discussion, give a short minilecture on fitting lines and multiple regression models with the software being used in the course. Have students do a full analysis after class and are ready to present the next day.

Flow of the Case: Approach 2

Make sure that the students know how to plot, fit lines, and multiple regression either from minilecture or by being directed to appropriate readings. Have them read the case and come up with the best approaches they can. In class, have students discuss and present their approaches. Distill the ideas at the end by having a student go through step by step the process to arrive at reasonable predictions.

A GENERIC INTEGRATIVE EXAMPLE

Consider the following generic example: a one server queue with random service and waiting times. Costs are incurred for waiting. The issue is should another server be introduced? This could incur extra fixed and variable costs. In a case setting, the problem would never be introduced in such a mathematical manner. Instead, the problem would be introduced from a business decision-making point of view. This scenario is close to a number of cases that we use.

“Top-Level” Questions: Making the Decision

The class would be asked a question such as “If you are the CEO, what crucial number or numbers—no more than two or three—would you want from an analyst/a consultant who is advising you on whether or not to invest in a second facility?” This is a question that everyone—no matter what their backgrounds—should be able to attempt. Many of the answers we get are illustrative of how students struggle in finding what is important. Some of the more popular answers we get are as follows:

- The current average waiting time;
- The current average idle time;
- The throughput;
- The work in process;
- The total current waiting time cost.

It is remarkable how difficult it is to get students to arrive at the most important issue, that is, how much more profit would be made if a new facility (server) were introduced. Then, it is even harder to obtain what needs to be calculated to obtain the answer to this question—that is, some measure of the gain in profit between the current system and the one with the new facility. Some numbers that might help if one were looking at one period would be

- expected profit for current system;
- expected profit for new system;
- difference between these.

Now suppose that the system is to be in place over a number of periods. How would one make a decision? This leads into a discussion of how decisions are made in reality. Such things as the net present value of the difference in profits between the two systems can come up.

Now, some arbitrary numbers can be assigned before the case is tackled at the OR level. The students are asked what they think about the risk involved. This leads them into asking how to refine the numbers given above and also to talk about what is meant by risk.

Detailed Analysis: Input for Making the Decision

So far no OR/MS or mathematical terminology has been introduced. Yet, in our experience, all students, both those with a good quantitative background and those who eschew quantitative things, find the above incredibly difficult. Yet, the above is often the most important part of any decision-making analysis for which OR/MS methods are appropriate and illustrates the value of the case method approach.

At this stage, one can start to get into the more detailed quantitative aspects. What are the stochastic inputs to the problem? How would one find appropriate distributions? Now, at the microlevel, for instance, one can look at the arrival distribution alone but realizing this will be an input for arriving at answers to the above questions. Distribution modeling, Excel spreadsheet modeling, and simulation will all be important here. The assumption is that cases that have made the students familiar with these techniques have been previously introduced.

Flow of the Case: Approach 1

- Have the students read the case before class but do not have them do any analysis.
- Have a discussion around the top-level decisions.
- Have the students “flowchart” the steps needed for the detailed analysis.
- Have the students present a detailed analysis for the next class that consists of a number of steps:
 - Slides that get straight to the management decision, that is, they should assume that they are consultants and give a “high-level” presentation.
 - Be ready with a detailed flow chart of how the problem was modeled.
 - Be ready to go over detailed analyses of the embedded OR/MS models, for example, precisely how they derived a service time distribution for the data provided.

Flow of the Case: Approach 2

- Have student read and try to “solve” the case before class.
- Have students present their work to the class. Have the class interact with a given solution by asking questions.
- Introduce questions such as the ones discussed above during the presentations and the ensuing discussions.

STUDENT EVALUATION AND EXPECTATIONS

It helps if the instructor assigns a participation grade to each student right after each session. For one thing, the discussion is still foremost in the instructor’s mind. This grade depends on the overall quality of everything that a particular student said in that class. We value a single, strong contribution more than several weak ones—however, quantity also counts for something. The instructor, in addition to her/his own grading, may also use peer evaluation for the class participation. Peer evaluation can be easily done by asking a few students at the beginning of each class to keep track of students who made a comment, as well as choose the ones who they think contributed the most to their learning that day. Marking class contribution is not that difficult in our classes because students sit on the same seat each class and also have nametags. (We change their seats only a few times in a year.) Therefore, there is a seating plan of the entire class that we can use in assigning grades. It also helps to have a photograph of the class because it makes it easier to assign participation marks afterwards. It is important for the instructor to be fastidious in this—often, at the end of a semester, students will try to increase their class mark because they feel they should have gotten a higher mark than others in the class based on their perception of what they said during the semester. Clear records will make the instructors job a lot easier here.

Some of the ways student can participate in class are

- starting off class/case discussion in an insightful way;

- presenting and developing models;
- advancing the comments of others;
- providing constructive criticism;
- emphasizing learning points;
- sharing personal workplace experiences from similar situations.

Grading class contribution is one of the most important aspects of interactive case-based teaching. It motivates students to come prepared to class. For example, by quick and simple questions to randomly selected students (i.e., by cold calls), the instructor can test whether students did the assignment readings and studied the case. This is an effective way of ensuring that students come to class prepared. Furthermore, we believe class contribution pressure helps students to improve their communications skills in business situations and helps them to develop public speaking skills, which is a fundamental requirement in business. Finally, grading of class contribution increases the quality of discussions and improves the overall learning and education experience of students.

For us, other components of student evaluation are quizzes, a midterm, and a final examination. We usually conduct a midterm or quiz to test the techniques and concepts taught in class. We design the examination to contain various types of questions (e.g., multiple choice, quantitative problems, true/false). On the other hand, the final examination is always a case examination. Depending on the instructor's choice, it could be either an individual hand-written examination or a take-home examination. If it is a take-home examination, we often conduct it as a 48-h report to be prepared in teams.

We expect students to be fully engaged in the entire learning process, that is, devoting time and energy to preparation before class, including small group meetings, listening to others during class discussions, and engaging in class discussions. Student participation is the key for our interactive case-based learning and teaching.

We acknowledge that the *ORMS/Analytics* subject may be difficult for some students, but we expect every student to attempt the cases on their own first. We

believe that it is okay not to be able to do the analysis correctly, but it is not acceptable to come to class unprepared. We communicate this philosophy to the students early in the semester. Lastly, we expect our students not to discuss any future cases (the ones not yet covered in class) with anyone other than their own classmates.

ADVANTAGES OF TEACHING ORMS/ANALYTICS WITH CASES

Case teaching is an interactive style of pedagogy that is in direct contrast to the traditional lecture style of teaching. Being a “science-” and “content-” based subject, *ORMS/Analytics* has conventionally been taught using lectures. We appreciate the value of lectures, and in fact, use minilectures in our classes at Ivey. Lectures allow us to transfer “fact/formula”-based knowledge to our students. On the other hand, we believe that case teaching enables “concept”-based learning. In our opinion, the best way to learn a concept is to apply it in a real-world setting, which is precisely the experience that cases provide. We structure our classes so that students attempt the cases individually before they arrive for class. In class, we orchestrate a sharing of ideas and coordinate a joint solution approach to the case problem. We believe that there are a number of advantages to this style of learning.

Students attempt to solve the case as part of their homework and often struggle to formulate a solution. This is an important part of the learning experience. We feel that students are more likely to remember a concept if they struggle a bit while learning it. Furthermore, cases usually enable students to take different sides and have dissimilar opinions on importance level of issues, effectiveness of solution methods, practicality of proposed solutions, and many others. This provides an excellent opportunity to see other views and new ways of thinking on the same case/problem. The class discussion requires students to be mentally alert and active so that they can contribute their ideas and opinions. They are more likely to be engaged in a case discussion as compared to a lecture that

requires them to be passive listeners. (Recall that we have as much as 30–35% of the grade based on class contribution that provides an incentive for students to participate in the case discussions.)

Cases are a great way for students to pull different concepts together and learn “integrated analytics.” For instance, we teach a case titled “Four Star Motorsports” [27] that requires the student to develop a pricing plan for selling tires over a season. The demand for tires is predicted using regression, and then the students apply optimization to choose the best set of prices to sell the fixed inventory of tires. Such integration of *Analytics* concepts has elicited favorable responses from our students.

Case teaching is rewarding for the instructor as well. Two class discussions of the same case are seldom alike. Unexpected surprises and unusual points of view abound in case discussions. We have personally enjoyed the mental stimulation that case teaching provides. It is exciting to stand in a theater-style classroom and direct the flow of ideas. Often there are students in a class who have practical experience that is relevant for the case discussion. Other students find this sharing of colleagues’ experiences to be a great motivator and help bring out the relevance of the material.

CHALLENGES OF TEACHING WITH CASES

The prospect of teaching cases is intimidating to many instructors—especially rookies. Case teaching requires the instructor to relinquish a degree of control over proceedings, which is a terrifying prospect for instructors who are used to delivering a prepared lecture.

Students from all backgrounds have difficulties with the case method. Those from a nonquantitative background tend to eschew anything quantitative. Surprisingly, they also find the “top-level” questions difficult because they require a certain formal discipline in how one approaches questions and makes decisions. Those from a quantitative background often want to run straight to modeling and often the wrong modeling. They tend to have too much faith in the models and the numbers and fail often

to realize that the models are only as good as the inputs. They also tend to find the “top-level” questions extraordinarily difficult.

This does raise issues for the instructor. In schools where evaluations are the most important part of the teaching experience, raising discomfort levels for students of all backgrounds can be problematic. As the course progresses, it is important to demonstrate to students that this method gives them superior analytical and decision-making skills that will be useful in very practical ways in their careers.

Substantial preparation is required for teaching a new case. The good news is that reteaching the case is much easier. Preparation includes not only knowing and solving the case and making a class plan ready but also familiarizing with the case teaching environment. Before your first case session, watching a few live case discussions would be of great help. Another way to get familiarized with the case method is to read about it [28]. The instructor has to be comfortable with students expressing their views—some of which may not be entirely correct or practical. The onus is on the instructor to listen, provide feedback to the student, and then “repackage” the student’s views for the rest of the students in class. This requires the instructor to be constantly present in the moment and exercise sound listening skills. Our experience has been that teaching cases is far more mentally draining than lecturing. Of course, this makes teaching more enjoyable as well.

Case teaching is not a good way to cover “content.” We use minilectures for that purpose.

As discussed in the previous section, often, it can help different sides in the class discuss a case. For instance, those who rush to a mathematical answer can be quizzed by students who are told to take the role of a consumer of information and ask probing questions about the practicality of the approaches. Those who want to go with a “gut feel” approach can be quizzed by other section of the class on some important managerial but very quantitative related issues. (“You say we should invest in your portfolio and that you expect to make a \$32 return. What is the probability the return will be

negative?” Of course, a good response to such questions requires that a model be run.) This banter and to-and-fro discussion (if tightly managed) can be one of the most interesting parts of the case discussion. Students start to focus on the material and how it needs to be presented and explained to their peers rather than as an academic exercise or something that will satisfy their perceptions of what an instructor wants.

The instructor will often have to deal with negative feelings of students who struggle with the cases that simulate unstructured decision-making situations experienced in real life. It is the instructor's responsibility to support the students and advise them on the benefits of learning through struggle. One effective way to calm down these students is to remind them that they will face similar unstructured difficult business situations in their career and future life and will need to make a decision then, and the sooner they can learn the better. Furthermore, we also tell our students that our classrooms and cases provide them a laboratory-like environment in which they can learn real-world decision making by making mistakes (with no cost and with no risk), learn from others, develop their skills, and build their confidence. We frequently remind our students that doing (especially management science) cases and learning from them is a nonstop process and everyone can get better at it. We tell students that learning doing/solving cases is like learning a sport, for example, swimming or playing tennis. Unless you put some effort, time, and will power, you will not learn it and you cannot learn it by just reading or going over the notes, you should practice it: attempt to solve cases, prepare for class properly, build models, discuss with your stud group, and participate classroom discussions.

Classroom discussions can get heated—especially when students challenge each other. The instructor has to maintain civility of discussion, which is an extra responsibility. Much of the ability to control a classroom comes from experience. We explore this in the section “Becoming a Good Case Teacher.”

It is the nature of the case method that once in a while (and sometimes frequently)

the class discussion will not go in the intended/correct direction, that is, it will bark up the wrong tree. The instructor should be aware that this is part of the learning and let the class go with its own way for a while. Most of the time, somebody in class will realize that they are headed in the wrong direction and change the discussion. And at other times, when no one realizes this or if the discussion is taking too long, the instructor should help the class to find the right path. One easy way to do this is to raise a question that has an answer pointing the right direction or ask a question that makes student think that they need a direction change, for example, anyone has a different/fresh idea/point of view?

Some instructors find it difficult to handle the negative feedback of students who struggle while attempting the cases on their own. Our experience has been that in the early stages of the course, students frequently approach us and express their frustration. However, they adapt over the course of a semester and realize that the struggle is a part of their learning. The case instructor has to be patient and offer gentle encouragement.

It usually helps to sit down with a struggling student and ask them how they had prepared for one of the previous cases. We usually find that insufficient preparation is to blame for unsatisfactory learning experience. From time to time, we also discuss and go over past cases with the students who had questions about them. To encourage noncontributors and frustrated students, we “warm call” them, that is, we inform them in advance of a class that we will call them for that specific case/class. This helps them to motivate and prepare for that class, and once they participate in class, they contribute to contribute in the future classes.

Classroom logistics can often be an issue. Ideally, case discussions require a theater-style classroom with stadium seating. However, this is not a disqualifying requirement. An enthralling discussion is what keeps students engaged.

Case teaching requires cases that are compatible with the teaching objectives of the course. This can be challenging sometimes. Cases can be obtained from schools such

as Ivey, Harvard, and Darden or textbooks. Another way to find cases is to write cases [29], and in fact, those cases can be the best cases to teach because as the writer of the cases, you know the case inside out and its learning/teaching objectives.

BECOMING A GOOD CASE TEACHER

Case teaching is a skill that can be learnt. It requires a period of apprenticeship in the hand of a mentor—someone who has had experience in teaching ORMS/Analytics using cases. In addition, we feel that the following will help:

- *Field Experience.* Much learning happens by actually conducting case discussions. We suggest that the apprentice be “process” oriented rather than “outcome” oriented. There will be some initial struggle, but the learning curve is steep.
- *Observing Good Case Teachers.* A lot can be learnt by watching a good case teacher facilitate a discussion. Having said that, we caution that the apprentice must not blindly copy the style of the mentor; it is important for the apprentice to develop his/her own unique style.
- *Reflection.* We recommend that the apprentice should spend some time after every class, recollecting the case discussion and evaluating the good and the bad. We also recommend that the apprentice should invite more experienced instructors to his/her class and request feedback on his/her class performance.

CONCLUSION

We have found that cases are a very effective method of enhancing ORMS/Analytics in introductory business courses. There are differences, however, between undergraduate business and MBA-level courses: frustrations can be higher for students who are less mature and have weak quantitative backgrounds. Executive MBA courses that we

have experienced are also different—there the focus needs to be much more on the top-level skills and less on the quantitative details. Course outside of business schools and higher (technical) level electives in a business school may require slightly different approaches. For instance, a course on Markov Chains requires substantial analytical and lecture content. However, cases interspersed at strategic point or toward the end of the course can substantially improve students’ abilities to focus on top-level issues related to the material.

Case preparation and discussion stimulate learning by doing for students, and case teaching can be rewarding for instructors. In order to start using cases, one does not need to give up lecturing; a hybrid teaching (a combination of cases and lectures) can be adopted. Each instructor can choose his/her comfort level and course/students needs in a wide spectrum of case teaching environments, that is, from a single case use in a course to a course where there is a case in each class. Case teaching is a skill that can be learnt and one can start using cases adding a single case to their course and go from there. Whether you consider cases as the side dish or the main entrée, they will add spice and flavor to your students’ learning and your teaching experience.

REFERENCES

1. Aggarwal R, Khara I Short cases for teaching management science: their role in closing the gap between theory and practice. *Interfaces* 1978;9(1):90–94.
2. Martin MJC. Short cases for teaching management science: an extended comment. *Interfaces* 1980;10(2):10–12.
3. Horner PR. The case for teaching MS/OR case studies in business school. *OR/MS Today* 1995; October. Available at <http://www.orms-today.org/10-95text/oct-toc.html>.
4. Bodily SE. The teachers’ forum: teaching MBA quantitative business analysis with cases. *Interfaces* 1996;26(6):132–138.
5. Liberatore MJ, Nydick RL. The teachers’ forum: breaking the mold—a new approach to teaching the first MBA course in management science. *Interfaces* 1999;29(4):99–116.
6. Bell PC, Haehling von Lanzanauer C. Teaching objectives: the value of using cases in

- teaching operational research. *J Oper Res Soc* 2000;51(12):1367–1377.
7. Hannibalsson I. Course correction. *OR/MS Today* 2001; August.
 8. Schmedders K, Stephens M. Time to update teaching cases. *OR/MS Today* 2001; February.
 9. Bell PC. The executive MBA. *OR/MS Today* 2003; August. Available at <http://www.orms-today.org/orms-8-03/frmba.html>.
 10. Long Z. On case teaching of operations research. Volume 2, Proceedings of the 2009 First International Workshop on Education Technology and Computer Science. 2009. pp. 547–550. Available at <http://toc.proceedings.com/05380webtoc.pdf>.
 11. Cochran JJ. Pedagogy in operations research: where have we been, where are we now, and where should we go?. *ORiON* 2009;25(2): 161–184.
 12. Cochran JJ. All of Britain must be stoned! An effective introductory probability case. *INFORMS Transactions on Education* 2009;10(2):62–64.
 13. Cochran JJ. Bowie Kuhn's Worst Nightmare—an integer programming & Simpson's paradox case. *INFORMS Transactions on Education* 2004;5(1):18–36.
 14. Cochran JJ. You want them to remember? Then make it memorable!. *Eur J Oper Res* 2012;219, 659–670.
 15. Liang N, Wang J. Implicit mental models in teaching cases: an empirical study of popular MBA cases in the United States and China. *Acad Manag Learn Educ* 2004;3(4):397–413.
 16. Kim S, Phillips WR, Pinsky L, Brock D, Phillips K, Keary J. A conceptual framework for developing teaching cases: a review and synthesis of the literature across disciplines. *Med Educ* 2006;40, 867–876.
 17. Currie G. Moving towards reflexive use of teaching cases. *Int J Manag Educ* 2008;7(1):41–50.
 18. Bocker F. Is case teaching more effective than lecture teaching in business administration? An exploratory analysis. *Interfaces* 1987;17(5):64–71.
 19. Wilson R. Red Brand Canners. Palo Alto (CA): Stanford University, Graduate School of Business; 1965.
 20. Leff M, Zaric G. The Professor Proposes. London, Ontario, Canada: Ivey Publishing, Ivey Management Services, c/o Richard Ivey School of Business, The University of Western Ontario; 2006.
 21. Zaric G. Lansink Appraisals. London, Ontario, Canada: Ivey Publishing, Ivey Management Services, c/o Richard Ivey School of Business, The University of Western Ontario; 2000.
 22. Capestany M, Bruner R, Bodily S. Enron Corporation's Weather Derivatives (A) & (B). Charlottesville (VA): The University of Virginia Darden School Foundation; 2000.
 23. Carraway RL. Jaikumar Textiles Ltd: The Nylon Division (A) & (B). Charlottesville (VA): The University of Virginia Darden School Foundation; 1991.
 24. Anderson C, Wilson JG, Read J. Kimpton Hotels: Setting Prices on Priceline (A), (B) & (C). London, Ontario, Canada: Ivey Publishing, Ivey Management Services, c/o Richard Ivey School of Business, The University of Western Ontario; 2010.
 25. Wilson J. The Clonlara Hotel. London, Ontario, Canada: Ivey Publishing, Ivey Management Services, c/o Richard Ivey School of Business, The University of Western Ontario; 2009.
 26. Begen MA, Jhun J. Seoul National Bank: The Chief Credit Officer's Dilemma. London, Ontario, Canada: Richard Ivey School of Business, The University of Western Ontario; 2011.
 27. Bell PC, Rosenshein JJ. Four Star Motorsports. London, Ontario, Canada: Ivey Publishing, Ivey Management Services, c/o Richard Ivey School of Business, The University of Western Ontario; 2003.
 28. Leenders MR, Mauffette-Leenders LA, Erskine JA. Teaching with Cases. Ontario, Canada: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2003.
 29. Begen MA, Krishnamoorthy S, Wilson J. Writing ORMS/Analytics cases. *EORMS Chapter* under review. 2012.
 30. Cochran JJ. Successful use of cases in introductory undergraduate business college operations research courses. *J Oper Res Soc* 2000; 51(12):1378–1385.
 31. Darden School of Business, University of Virginia. Available at <https://store.darden.virginia.edu/> (accessed 9 June 2012).
 32. Harvard University. Available at <http://hbsp.harvard.edu/product/cases> (accessed 9 June 2012).
 33. Ivey School of Business, University of Western Ontario. Available at <https://www.iveycases.com/> (accessed 9 June 2012).

TEACHING SOFT OR/MS METHODOLOGIES: WHAT, WHY, AND HOW

FRAN ACKERMANN
Curtin University, Perth,
Western Australia, Australia

INTRODUCTION

Management Science (MS)/Operational Research (OR) came of age in World War II [1,2] gaining momentum post conflict in organizations seeking to ensure that they adopted a more “scientific” basis [3]. However, while the benefits were acknowledged by businesses, application, particularly in the United States, appeared to be dominated almost exclusively by a quantitative view of the field [4]. This, it is argued [5] is partly due to the editorial policies of US journals and partly due to tenure requirements and also the virtual absence of qualitative methods from teaching [6, 7].

Over time, however, many managers while getting benefit from quantitative techniques such as optimization and so on found themselves struggling to either fully implement the outcomes from the techniques or not getting the benefits they anticipated from them. It has been argued that these issues are likely due to two factors. First, organizations are made up of people and so negotiative approaches are best, and second, problems facing managers are increasingly complex with multiple uncertainties. These observations were made by Ackoff (among others) who argued that the quantitative forms of OR (hard OR) were “mathematically sophisticated but contextually naive” [8], and furthermore that it was important that (OR) was “based on planning with people, not for them” [9]. Ackoff’s propositions stimulated a heated debate within the OR community and resulted in a more wide ranging understanding of OR’s current and potential contribution to improved decision making.

Based partly on this debate, and the work of a range of researchers, the qualitative (soft OR) field sought to incorporate the two factors mentioned above, and a growing appreciation has been gained not only for the soft qualitative approaches themselves but also for approaches that consider human factors and the management of complexity with the more mathematical techniques—recognizing that there is complementarity. This blend of soft/qualitative and hard/quantitative is gaining prominence as evidenced by the growing body of literature around mixing methods [10, 11] as well as the increasing use of combined methods when working on organizational problems [12]. Thus, understanding what constitutes Soft OR and why it is of value is important.

This article therefore explores the realm of Soft OR, considering firstly *what* is Soft OR (and what is it NOT), *why* Soft OR should be taught to students and following on from this *who* should be taught and *when*. This article concludes with some suggestions regarding *how* to teach Soft OR (and thus having answered, what, why, who, when, and how) before reflecting on how best to move forward. In essence, the article adopts Rudyard Kipling’s verse in the Just So Stories regarding the six wise men, simply missing the *where* element (which is to some extent contained within the how and who).

WHAT CONSTITUTES SOFT OR?

Soft OR methods emerged from a growing sense that the traditional “hard” methods were not able to deal with messy [13], complex [14], or wicked [15] situations. Moreover, there was a sense that OR was “losing its way and its self-confidence in the early 1970s—architecture, urban planning and policy sciences all suffered crises” ([16], p. 3) and thus alternative means of modeling were required. Researchers argued that in many cases there wasn’t a single “right” answer—based upon the fact that a number of different perspec-

tives needed to be included, different criteria taken account of, and considerations of sustainability undertaken.

It was also noted that “typically large decisions were not made by a single group of like-minded people . . . they are, rather, the result of extended negotiations, either explicit or implicit between representatives of different points of view” [17] and thus it was important that techniques that thoughtfully address contextual and socio-political considerations were developed. Groups of decision makers needed more support in terms of surfacing, structuring, and exploring knowledge—moving from a divergent explorative position to one that was more precise and detailed in its data and analysis [18]. Researchers recognized that much of knowledge is socially constructed [19] moving from a mechanistic view of an organization to one that recognizes the realities of organizational life with attendant power and politics. Organizations and decision makers needed to move from a basis of muddling through to being provided with the means to make decisions that were both procedurally rational [20] and procedurally just [21].

A range of methods, originating from different fields began to emerge to meet this need, which over time became known as problem structuring methods (PSMs). Rosenhead [22] describes PSMs as methods that seek to help managers deal with situations that “encompassed multiple actors, different perspectives, partially conflicting interests, significant intangibles, and perplexing uncertainties” ([22], p. 759). These methods focus on determining what the problem is, as much as providing assistance to its resolution.

Key contenders for PSM as noted in Ackermann [23] are soft systems methodology—SSM [24, 25], strategic options development and analysis—SODA [26–28], strategic choice [29] as well as drama theory [30], decision conferencing [31], robustness analysis [32], and viable systems model [33]. Each of these has extensive experience “in the field” as well as being supported by a considerable body of theoretical and conceptual frameworks. More recently group model building ([34, 35]) has also been identified as a contender. Good compendiums of PSM theory

and practice can be found in Rosenhead and Mingers [16] and Reynolds and Holwell [36].

PSMs allow the capture and structuring of perceptions often times through graphical representations, for example, through causal maps in the case of SODA or rich pictures when using SSM. Moreover, they explicitly encourage the surfacing of socio-political considerations weaving in stakeholder management issues, conflict, competing priorities so as to provide a strong basis from which to negotiate a way forward. Each of the methods aims to ensure that different objectives/goals/aims are revealed, different options elicited and the means for their assessment identified once a good understanding of the situation has been gained. The captured material is explored with the group, using the various modeling techniques/analyses to foster the development of an enhanced understanding by participants, to enable a shared language to be developed, and through using the representation(s) to act as transitional objects [37] helping a group negotiate toward a set of improvements and actions to resolve the situation [23]. Models can also be considered boundary objects [38]

Alongside the development of PSMs, Jackson and Keys [39] taking a systems perspective, derived SOSM—system of systems methodologies—which helped to determine which types of problem were best resolved using a PSM/soft OR type approach. Problems with a high degree of complexity in terms of problem context and in terms of a wide range of relationships with stakeholders were considered prime candidates.

However, unfortunately while there is a body of literature expounding the methods in terms of underlying theories, guidelines for use and cases illustrating their application, these are not always appreciated and considered in a rigorous manner. Probably due to their subjective, qualitative nature, there are many reported instances where researchers and practitioners have asserted that they have used a particular soft OR approach but on closer scrutiny the application is poorly designed showing little evidence of a subtle and robust understanding of the methods in

terms of theory and practice thus tarnishing the method's reputation. This is a point noted by both White [40] and Ackermann [23]. These are not applications of soft OR.

WHY TEACH SOFT OR?

Williams [4] argues for five reasons for why PSMs/Soft OR should be used and therefore by implication taught (the reasons themselves are expounded in his MS textbook). These are where problems have

- (a) Multiple actors rather than a single one,
- (b) A reality that cannot be objectively determined,
- (c) Actors who have different aims, different objectives, and often different values,
- (d) Dynamic complexity that can be ill-understood and,
- (e) Behavioral complexity—many actors interacting in complex behavioral ways with underlying power-structures organizational politics and social geographies.

William's reasons for using PSM's recognize that there are different types of problems, those that are difficult but bounded (there is an agreed scope of work) and those that are messy and complex (reflecting the above criteria). These sorts of problems are likely to be faced by OR students in their day-to-day professional life. Traditional approaches to OR are geared toward tackling the former rather than the latter and thus are "necessarily confined to one corner of the field of organizational decision making" ([16], p. 8) recognizing that extending student repertoire is necessary. Moreover, problems that demand acknowledgment of soft human factors increasingly are taking up managerial time. Consequently, it is important to widen the scope of methodological support so OR's have methods for managing the "soft" messy problems and those that are difficult but more bounded. As noted later, however, providing students with the ability and confidence to manage these types of problems is itself a challenge particularly those

without organizational experience namely, undergraduates.

In addition, as noted in Rosenhead and Mingers [16], there is an aphorism from one of the earliest classics of the OR literature [41] stating "there is an old saying that a problem well put is half solved. This much is obvious. What is not so obvious is how to put a problem well." This is echoed by Einstein's observation that paying attention to the formulation of the problem is often the most important part of solving any problem. PSMs provide powerful mechanisms for determining which problem is to be considered, what it constitutes and getting agreement for this. This is clearly reflected in SSM where one of the first stages is to begin to get an understanding of the situation under consideration through the creation of a rich picture recognizing the need to be explicit. Similarly, SODA in the generation of causal maps encourages participants to surface the issues they are concerned with along with their implications allowing for an understanding of the whole to be gained before agreeing where to focus energy. Both these are examples of Goffman's [42] notion of framing where framing is concerned with how we make sense of events through employing a scheme of interpretation.

Another reason for providing students with an understanding of soft OR methods, is that in many situations, there are "participants for whom backroom analytic modeling would have been mystifying and excluding" ([16], p. 9). Appreciating the context is, if not mandatory for many applications of more quantitative methods, certainly important. Frequently OR students experience great disquiet when their beautifully executed piece of analysis is not adopted—not understanding that either it isn't really addressing the problem being experienced by the organization, or it doesn't consider the political landscape. Finding ways to engage decision makers and encourage them to understand and agree how to move forward, and thus buy into the resultant outcomes can be achieved using soft OR methods as a means of gaining agreement for the analysis, and helping with the management of the politics.

As such, soft OR is particularly relevant for working in the strategic strata

of problems where there is no clear single answer, and where there are multiple uncertainties. It is not surprising that SODA has evolved into an approach for making strategy, Strategic Choice was adopted into regional and city planning, and that system dynamics is used in policy analysis. Strategic decisions demand attention is spent on ensuring sustainable agreements are achieved as strategies usually require a long term view to be taken. This therefore requires decisions makers to recognize the need to attend to systemicity—that issues impact other issues, that options can have potency, and that some options are not feasible in terms of achieving particular outcomes. In essence, this is about recognizing the whole can be greater than the sum of the parts.

A final consideration for teaching soft OR is that while there are very few jobs explicitly requiring soft OR skills, virtually every decision that involves people requires knowledge of methods for managing both content and process and soft OR provides this.

WHO SHOULD BE TAUGHT?

The easy answer here is to argue for all students—as messy complex problems abound. However, when faced with the varying challenges experienced in designing curricula (itself a messy problem as there is not one single, right way to teach any subject or course!) students who are comfortable with working with people rather than in the backroom are likely to be more comfortable. This is because soft OR requires facilitation skills and acknowledges Langley's [43] view that "formal analysis and social interactive processes in organizations must be viewed as being closely intertwined rather than as mutually incompatible." Soft OR methods, as noted above, help support both process management and content management.

Another tranche of students who are prime candidates are those aspiring to be, or who already are, working in consultancy as providing them with the means to determine a good understanding of the situation can be very beneficial. As such Masters in Business Analytics courses, Masters in Business Management, or Management Science (MS)

would find learning about different soft OR methods useful. This is particularly the case for MS students who would be able to not only use the soft OR methods on their own but also combine them with the hard/quantitative methods allowing for the ability to tackle a very wide range of problems and thus increasing the chance of good job opportunities.

A further cohort for consideration is engineering students. This suggestion is based on observations that many of the progenitors of the soft OR methods came from an engineering discipline (e.g., Colin Eden and the development of SODA and Peter Checkland and the development of SSM) and thus it can be postulated that it was a lack in methods for supporting engineering projects that drove them to develop the methods thus rendering these methods apposite for those with aspiration of working in engineering jobs. In addition to this, is the fact that many of the case studies reported to have benefited from the use of soft OR are based in engineering firms (or those with engineering components e.g., Shell). Organizations which routinely take on board large projects involving a plethora of different organizational contributions (both intra and inter organizational), bespoke and potentially state of the art activities, and with tight time scales clearly will benefit from use of soft OR methods as gaining an understanding of the objectives, issues, and contributions in a systemic manner will help ensure successful conclusions.

Finally, staff working, or aiming to work, in public sector organizations would benefit from being skilled in soft OR. This assertion is based on a series of characteristics of public sector working. In the public sector, a wealth of stakeholders exists and thus the need to navigate a frequently politically fraught space is mandatory. There is rarely one agreed view on how the particular public mandate should be achieved, and seldom sufficient resources. Furthermore, public sector organizations are often at the mercy of the political direction which can experience sudden and drastic shifts. Thus, it becomes necessary to be able to make adjustments to the course of action in a sensible, shared, and sustainable manner. PSMs thus are

ideally suited to these sorts of characteristics that frequently manifest in the public sector. Consequently, teaching PSMs to students attending Masters of Public Administration courses or in-house programs will clearly have benefit.

WHEN TO TEACH SOFT OR?

From personal experience, the most fruitful tranche of students to teach soft OR methods to are those students who have experience in the “real” world and thus are able to appreciate the need to pay attention to issues relating to power, politics, and personalities. As such, Master’s course students rather than undergraduate students are the ones most likely to benefit from the exposure to PSM’s. This is not to say undergraduates won’t benefit. There are enough examples of undergraduates particularly those in their honors year for whom the methods really resonate and thus the students absorb the methods and become extremely proficient. However, in an undergraduate course such students are smaller in number than is normally the case for master’s students—and particularly MBA master’s students. This is partly because MBA students are post-experience and therefore will already be sensitized to the difficulties of getting agreement and action, particularly when instigating major change. Moreover, there will be an appreciation that politics do affect decision making, that there is rarely a single right answer and that negotiation is a key aspect.

Teaching soft OR at the beginning of a course also has been found to be advantageous. This is because students are able to use the methods during the course and after. For example, students have found being able to structure the content and manage processes (particularly when doing group assignments which proliferate in many courses) helpful in completing course work. Being able to pull together the key issues and understand how they fit together to form the whole can also help with studying as it can help holistic learners get a good overview and provide an effective means of ensuring successful outcomes. This is similar to the benefits gained from using Buzan’s mind maps [44]

which have an established pedigree in teaching and learning. The other benefit students can gain is from using the structuring techniques in the design, and in some cases execution of their final project, which for many master’s students is based in organizations where issues such as multiple conflicting views and uncertainty can be encountered.

The last aspect of “when” to teach soft OR relates to those teaching the techniques. Junior academic staff with little or no experience of industry, and whose doctoral studies are significantly desk based, are likely to find teaching soft OR a challenge. This is because a key element of soft OR is familiarity of organizational behavior. Junior staff are unlikely to have a repertoire of organizational problems that they have encountered and successfully managed and thus are limited when seeking to illustrate the application of soft OR methods as well as authentically expound the rationale for considering such techniques. Senior academics, on the other hand, who have worked with industry on projects, or who have extensive experience of University management are far more likely to be able to effectively present the need to attend to the socio-political considerations. Likewise those who have completed extensively quantitative degrees (honors or doctoral) may find teaching soft OR challenging. This is because their natural preference is for the very objective, quantitative world where there is little sympathy or natural affinity for human factor considerations.

HOW TO TEACH SOFT OR?

Teaching soft OR, however, is not easy even if you have experience in organizations. To begin with there are multiple challenges ranging from (i) difficulties in helping students appreciate the implications of the socio-political world, (ii) giving students confidence in not having the “right” answer, (iii) having to teach facilitation as well as the methods, (iv) providing students with an appreciation of the whole AND the details, (v) helping students manage complexity, uncertainty and holism, (vi) covering the material necessary within the time allocated to (vii) teaching the class on line [45].

However, there are, as noted in Ackermann's [45] paper, a number of means of managing these challenges and by being aware of the challenges and their management options, many difficulties can either be alleviated or avoided.

In addition, to support the teaching of soft OR, there is a growing body of scripts emerging [34, 46, 47] where workshop design is broken into "chunks" each chunk typically lasting 20 min providing a "product" (progress step) with the entire script resulting in a clear deliverable (helping resolve the situation at hand). Augmenting this research into scripts is the development of ThinkLets [48] whereby particular workshop designs are developed between an organization and an experienced facilitator with the intention that the designs are used in house on a regular basis. This unlocking of the black box of workshops [49] goes some way to resolving the issue of transferability recognized as early as 2006 [50].

When teaching the methods, the question of whether or not a good facilitator requires subject matter/substance knowledge is also pertinent [51]. Clearly, it is important to ensure that the facilitator avoids introducing bias into the meeting or confusing their role with the roles of participants but in many instances having some knowledge of the area, particularly some of the idiom and abbreviations can prove very helpful. Thus, familiarity with soft OR doesn't only demand good modeling skills but also process management and subject matter knowledge. This is very different from hard OR which is essentially practitioner free [52].

One of the challenges is what to teach. Given the finite amount of time available and the need to teach both content management (modeling) and process management, one of the questions is whether to focus on one or two methods allowing for competences in associated techniques to be acquired or to cover a wider range but at a relatively superficial level—allowing for a good overall understanding of the range but little actual practice. This decision needs to be considered alongside the objectives of the entire the course, and the quality/experience of the enrolled students. Moreover, based on

years of teaching PSMs, starting small and providing students with exercises and activities that allow them to gain familiarity and thus comfort is best. Providing case studies (see [53]) can help, so too, where possible, encouraging groups to try out the techniques in low risk environments, for example, helping a local group think about how to grow the local soccer team also helps in gaining confidence as well as competence.

It is also worth helping students recognize the very real possibility of change/adaptation being required and being comfortable to think of their feet. This is where understanding the theory helps as it provides a foundation upon which to adapt. Reflecting after interventions can also assist with the learning process. Students benefit from being alerted also to client management matters. For example, helping students understand the dynamic nature of problems and moving objectives [54].

While an ideal way of learning PSMs is through apprenticeship, this is an extremely costly method. However, if those mentoring and those receiving the mentoring begin to note some of the more tacit insights gained from working in a partnership then these can be fed into the scripts enriching the scripts and helping to ensure wider access.

CONCLUSIONS

As evidenced from the above, focusing on teaching soft OR from the perspective of the "5 wise men" shows that they are not mutually exclusive nor discrete as it is hard to separate the who from the when, the why from the how. Thus, all have to be considered alongside one another when considering teaching soft OR. Moreover, taking cognisance of the growing benefits and traction elicited from research and practice in mixing methodologies the fusion of soft and hard OR requires attention alerting students to the pitfalls but also benefits and means of method integration.

One very particular reason for teaching soft OR is its longevity. There are real concerns for its future based on the fact that many of the progenitors are retiring and leaving the field and thus there is a need to teach the methods so that they move into the

future. Traditionally as intimated above, the field of PSMs was dominated by a UK orientation with little development of the field taking place in OR/MS groups in the United States and beyond. However, when taking a wider view, that is, one beyond the OR/MS world, evidence of problem structuring like methods appear in the United States. For example, in policy where system dynamics is used to support groups working on complex policy issues and in the group decision and negotiation field. There is also evidence in the strategy arena for the value of soft OR principles—as authors have noted that when working with complex strategic problems, actions needed to be robust, owned, and relatively quick [18]. Seeking to integrate developments from these fields into the OR/MS world will help not only widen their accessibility but provide new stimulus for their development.

There is no doubt that any field of research needs continuous development and soft OR is no exception. For example, the methods would benefit from considering new developments in technology. One particular ICT avenue that offers real potential is advances in visualization. Firstly, because using some of breakthroughs in virtual reality give rise to opportunities for training environments where students can facilitate in a virtual world experiencing many of the likely behaviors in a risk free environment. Not only will technical competence be gained but also, on review of the virtual facilitation, reflections on how best to manage the process aspects is also possible. Another avenue for using visualization is the ability to move from modeling qualitative data on two dimensions to using three and gaining even more insights into emergent structure. New forms of analysis can thus be constructed giving decision makers even more assistance in making good decisions.

As noted by Vidal in an introduction to a special issue on PSMs “from an epistemological viewpoint we can stipulate that these (PSM) methods try to combine practical experience with theoretical knowledge, and objective facts with subjective appreciation with the purpose of problem solving in very complex situations” ([55], p. 529). Thus, they meet the twin requirements of rigor

and relevance. To conclude—in returning to the title of this article—Teaching Soft OR Methodologies: What, Why and How—the answer is because there is an overwhelming need for decision makers to either have access to those who can help them navigate messy complex situations or manage them themselves. Complexity is here to stay.

REFERENCES

1. Kirby M. Operational research I war and peace: the British experience from the 1930s to 1970. London: Imperial College; 2002.
2. Waddington CH. *O.R. in World War 2: operational research against the U-boat*. London: Elek; 1973.
3. Locke R. Management and higher education since 1940: the influence of America and Japan on West Germany, Great Britain and France. Cambridge, UK: Cambridge University Press; 1981.
4. Williams TM. Management science in practice. Chichester: Wiley; 2008.
5. Mingers J, Rosenhead J. Introduction to the special issue: teaching soft O.R., problem structuring methods, and multimethodology. *INFORMS Trans Educ* 2011;12(1):1–3.
6. Mingers J. Taming hard problems with soft O.R. *OR/MS Today* 2009;36(2):48–52.
7. Mingers J. Soft OR comes of age—but not everywhere!. *Omega* 2011;39(6):729–741.
8. Ackoff R. The future of operational research is past. *J Oper Res Soc* 1979a;30:93–104.
9. Ackoff R. Resurrecting the future of operational research. *J Oper Res Soc*. 1979b;30:189–199.
10. Mingers J, Brocklesby J. Multi-methodology: towards a framework for mixing methodologies. *Omega* 1997;25(5):489–509.
11. Munro I, Mingers J. The use of multi-methodology in practice—results of a survey of practitioners. *Oper Res Soc* 2002;53:369–378.
12. Howick S, Eden C, Ackermann F, *et al*. Building confidence in models for multiple audiences the modelling cascade. *Eur J Oper Res* 2008;186:1068–1083.
13. Ackoff R. The art and science of mess management. *Interfaces* 1981;11:20–26.
14. Schon D. *Educating the reflective practitioner: toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass; 1987.

15. Rittel HWJ, Webber MM. Dilemmas in a general theory of planning. *Policy Sci* 1973;4: 155–169.
16. Rosenhead J, Mingers J. *Rational analysis for a problematic world revisited*. London: Wiley; 2001.
17. de Neufville R, Keeney RL. Systems evaluation through decision analysis: Mexico city airport. *J Syst Eng* 1972;3:34–50.
18. Eden C, Ackermann F, Bryson JM, et al. Integrating modes of policy analysis and strategic management practice: requisite elements and dilemmas'. *J Oper Soc* 2008;60:2–12.
19. Berger PL, Luckmann T. *The social construction of reality*. New York: Doubleday; 1966.
20. Simon HA. From substantive to procedural rationality. In: Latsis SJ, editor. *Method and appraisal in economics*. Cambridge: Cambridge University Press; 1976.
21. Kim WC, Maubornge RA. A procedural justice model of strategic decision making. *Org Sci* 1995;6:44–61.
22. Rosenhead J. The past, the present and the future of problem structuring methods. *J Oper Res Soc* 2006;57:759–765.
23. Ackermann F. Problem structuring methods 'in the Dock': arguing the case for soft OR. *Eur J Oper Res* 2012;219:652–658.
24. Checkland P. *Systems thinking systems practice*. Chichester: Wiley; 1981.
25. Checkland P, Holwell S. *Information, systems and information systems: making sense of the field*. Chichester: Wiley; 1998.
26. Eden C, Ackermann F. Strategic options development and analysis: the principles. In: Rosenhead J, Mingers J, editors. *Rational analysis in a problematic world*. Chichester: Wiley; 2001. pp. 21–42.
27. Ackermann F, Eden C. SODA and mapping in practice. In: Rosenhead J, Mingers J, editors. *Rational analysis in a problematic world*. Chichester: Wiley; 2001. pp. 43–60.
28. Ackermann F, Eden C. Strategic options development and analysis. In: Reynolds M, Holwell S, editors. *Systems approaches to managing change: a practical guide*. London: Springer; 2010. pp. 135–190.
29. Friend J, Hickling A. *Planning under pressure: the strategic choice approach*. Oxford: Pergamon; 1987.
30. Bryant J. The plot thickens: understanding interaction through the metaphor of Drama. *Omega* 1997;25:255–266.
31. Phillips L. People-centred group decision support. In: Doukidis G, Land F, Miller G, editors. *Knowledge based management support systems*. Chichester: Ellis-Horwood; 1987. pp. 208–224.
32. Rosenhead J. Robustness analysis: keeping your options open. In: Mingers J, Rosenhead J, editors. *Rational analysis for a problematic world*. London: Wiley; 2001. pp. 181–208.
33. Beer S. *Brain of the firm*. Chichester: Wiley; 1981.
34. Hovmand PS, Andersen DF, Rouwette EAJA, et al. Group model-building scripts as a collaborative planning tool. *Syst Res Behav Sci* 2012;29(2):179–193.
35. Andersen DF, Vennix JAM, Richardson GP, et al. Group model building: problem structuring, policy simulation and decision support. *J Oper Res Soc* 2007;58(5):691–694.
36. Reynolds M, Holwell S. *Systems approaches to managing change: a practical guide*. London: Springer; 2010.
37. de Geus A. Planning as learning. *Harv Bus Rev* March-April, 1988;66:70–74.
38. Zagonel, A. A. (2002) Model conceptualization in group model-building: a review of the literature exploring the tension between representing reality and negotiating a social order. Proceedings of the 20th International Conference of the System Dynamics Society. Palermo, Italy. Albany, NY: System Dynamics Society.
39. Jackson MC, Keys P. Towards a system of system methodologies. *J Oper Res Soc* 1984;35:473–486.
40. White L. Evaluating problem-structuring methods: developing an approach to show the value and effectiveness of PSMs. *J Oper Res Soc* 2006;57:842–855.
41. Churchman WC, Ackoff RL, Arnoff EL. *Introduction to operations research*. New York: John Wiley & Sons, Inc; 1957.
42. Goffman E. *Frame analysis*. Harmondsworth, UK: Penguin Books; 1974.
43. Langley A. In search of rationality: the purposes behind the use of formal analysis in organisations. *Adm Sci Q* 1989;34:598–631.
44. Buzan T. *The mind map book*. New York: Dutton; 1993.
45. Ackermann F. Getting messy' with problems: the challenges of teaching 'Soft' OR. *INFORMS Trans Educ* 2011;12(1):55–64.

46. Ackermann F, Andersen DF, Eden C, *et al.* ScriptsMap: a tool for designing multi-method policy-making workshops. *Omega* 2011;39(4): 427–434.
47. Andersen DF, Richardson GP. Scripts for group model building. *Syst Dynam Rev* 1997; 13(2):107–129.
48. de Vreede GJ, Briggs RO, Kolfchoten GL. ThinkLets: a pattern language for facilitated and practitioner-guided collaboration processes. *Int J Comput Appl Technol* 2006;25(2): 140–154.
49. Franco LA, Rouwette EA. Decision development in facilitated modelling workshops. *Eur J Oper Res* 2011;212(1):164–178.
50. Keys P. On becoming expert in the use of problem structuring methods. *J Oper Res Soc* 2006;57(7):822–829.
51. Huxham C, Cropper S. From many to one—and back. An exploration of some components of facilitation. *Omega* 1994;22(1):1–11.
52. Checkland P. OR and the systems movement: mappings and conflicts. *J Oper Res Soc* 1983;34:661–675.
53. Mingers J, Rosenhead J. Problem structuring methods in action. *Eur J Oper Res* 2004; 152(3):530–554.
54. Eden C, Ackermann F. Use of soft OR models by clients—what do they want from them?. In: Pidd M, editor. *Systems modelling: theory & Practice*. Chichester: Wiley; 2004. pp. 146–163.
55. Vidal RVV. Guest editor's introduction. *Eur J Oper Res* 2004;152:529.

THE ANALYTICS SOCIETY OF IRELAND

CATHAL M. BRUGHA
Centre for Business
Analytics, School of
Business, University
College Dublin,
Belfield, Dublin 4,
Ireland

ESTABLISHMENT OF THE SOCIETY

The Analytics Society of Ireland (An Cumann Eireannach d'Anailísí in the Irish language) was founded in 1964 as the Operations Research Society of Ireland (ORSI). In 1985, the phrase *management science* was included in the name of the society, making it the Operations Research and Management Science Society of Ireland (ORMSSI). Subsequently, the society was extended to include Systems Science and Decision Science, and the name was changed to the Management Science Society of Ireland (MSSI). The society was known by this name until 2011, when the society recast itself as the Analytics Society of Ireland to reflect its members' perception that their work now primarily deals with analytics.

In 1972, the ORSI hosted one of the first IFORS conferences to be held in Europe. At this conference, the important decision to establish the Association of European Operational Research Societies (or EURO) was made. Over the years, the society has published newsletters and held annual conferences and student conferences. While the demand for national conferences and written newsletters in Ireland has declined in recent years, the society was at its most active from the mid-1960s to the mid-1980s and followed the pattern of the UK Operational Research Society. It was led by a small but enthusiastic group who were employed primarily in private industry. This group

included Dr. Fred Ridgeway (banking), Dr. Roy Johnston (consulting), Dr. Frank Bannister (who later joined TCD), and others. There were operations research (OR) units in the public service, Aer Lingus (airline), and Guinness (beverage industry). Many other applications, such as rural milk collection, were also implemented in Ireland. A member of the society, Professor Harry Harrison, was an early winner of the Edelman Award for his application in pharmaceutical distribution.

EVOLUTION OF OR IN IRELAND

By the mid-1990s, most OR units in Irish companies had disappeared, with the exception of Aer Lingus, which was eventually amalgamated into business strategy and now has reappeared as revenue management. Despite this trend, the practice of OR has increased and continues to grow but is now not seen by the Irish as a separate entity but rather as a function to be executed by people in engineering, finance, software, and information and communication technologies (ICT) (growth sectors in which the Irish export-orientated economy has strong links to the US industry).

For many years, the emphasis of the society was on organizing approximately a dozen seminars annually and providing an Internet-based link to members and interested people. Society members generally prefer international conference over national conferences, but they occasionally want to give papers at seminars in Ireland to spread the word on what they are doing and to have a dry run before presenting their work at international conferences. This reflects a change in the nature of the communication and culture of the society's members. Communication is now more immediate because of the availability of e-mail and is more international because of the Internet and the wide selection of specialist and general conferences. This also reflects a change in the orientation of members toward a

“small world” perspective. The proportion of members and former members who consider themselves to be OR specialists has declined over time, and the field has become more segmented and possibly also fragmented. Members of the society can be found working in decision support systems (DSS), expert systems, geographic DSS, decision science, soft OR, systems in general, case-based reasoning, information systems, information technology management, computer science, multicriteria decision making, data mining, environmental planning, multicriteria decision making, and so on.

This phenomenon is now reflected in the Analytics Society’s approach. Several years ago, the society carried out surveys of members, and as a consequence, formulated a new strategy for the society. The survey results showed that members did not find merit or value in attending meetings because of the pressures of time, work, and travel. Many of the applications practitioners are working on are quite advanced and specialized, and these practitioners find that their efforts overlap little with the work of others.

The society has built on this role as the key (driving) node of a network, linking alumni of the MSc and practitioners through its participation in IFORS, EURO, and OR Society conferences. The society has recently extended these efforts to conferences in related areas such as systems science [International Society for the Systems Sciences (ISSS), The International Federation for Systems Research (IFSR), etc.]. When society meetings were conducted, they focused on providing a forum for members to test their ideas before presenting them at international conferences. The Analytics Society now sees

itself as a support group for early career OR practitioners and academics.

RECENT HISTORY OF THE SOCIETY

Since 1991, the office of president has been held by Professor Cathal Brugha, who is the Director of the UCD Centre for Business Analytics in the School of Business at UCD. The Secretary since 1995 has been Dr. Joseph Coughlan, who is the Head of School of Accounting and Finance in the Dublin Institute of Technology.

Having joined UCD in 1991, Cathal Brugha took over the role of president and linked it to the MSc in Operations Research/Management Science in UCD (MMS) in UCD’s Smurfit Business School. Many of the officers of the society were alumni of this MSc, including Dave Darcy, David Doody, Pidge Lynch, and later Joe Coughlan. Recently the UCD Centre for Business Analytics has played a more active role, with Sean McGarraghy dealing with international affairs, and Peter Keenan with membership.

Since 1991, the society has had a continual representation at IFORS and EURO conferences and has participated in the annual council meetings of both groups. Brugha also served as the Editor of International Transactions in Operational Research (ITOR), the Official Journal of IFORS, from 2000 to 2006.

In the future, the Analytics Society of Ireland hopes to strengthen its links with private industry, serve as an alumni network, and provide opportunities for sharing and collaboration between business, academic institutions, and the state.

THE ASSOCIATION OF ASIA PACIFIC OPERATIONAL RESEARCH SOCIETIES

YAXIANG YUAN
DEGANG LIU
Academy of Mathematics and
Systems Science, Chinese
Academy of Sciences, Beijing,
China

The Association of Asia Pacific Operational Research Societies, abbreviated as APORS, is one of the regional groupings under the International Federation of Operational Research Societies (IFORS). Its objectives are to develop operational research (OR) as a unified science and advance it in the Asia-Pacific (AP) nations through sponsoring international meetings, exchanging information on OR, encouraging AP nations to establish their national societies, and promoting OR education and applications. By 2011, APORS has 12 member societies and becomes actively involved in all IFORS activities.

ESTABLISHMENT OF APORS

In 1984 during the tenth IFORS meeting in Washington, Prof. Masao Iri, the then IFORS Vice President, called a meeting of AP OR society representatives to discuss on possible launching of a society group in AP region. Through effective communication among these societies, a formal discussion was later called in March 1985 during the Operations Research Society of Japan (ORSJ) national meeting in Tsukuba, Japan on preparing the APORS, attended by IFORS national society representatives of Japan, China, Australia, New Zealand, Korea, India, Singapore, and Hong Kong. This meeting announced the APORS with mission of encouraging OR society exchanges in the region. A council composing representatives was established. Prof. Woong Bae Rha (Korea) was elected as the first APORS president, Prof. Guang-Hui Hsu (China) as vice president, Prof. K. Wakayama (Japan) as secretary, and

Prof. R. Johnston (Australia) as treasurer. It was also decided that the *Asia-Pacific Journal of Operational Research* (APJOR) edited by Singapore OR Society be published as APORS official journal and the first APORS meeting would be convened in Seoul in 1988. This first council meeting was hosted by the OR society of Japan. Its then president J. Kondo, representative H. Takamori, and Prof. M. Fushimi and Prof. K. Yanai from ORSJ also attended the launch meeting.

In August 1988 in Seoul, the Korea OR Society hosted the first APORS conference. Before the conference opened, an APORS council meeting was called. Prof. Guang-Hui Hsu (China) was elected as the second term president, Teresa Ling (Hong Kong) as vice president, and Prof. Wakayama continued his service as secretary. It was proposed that the second APORS meeting be convened in Beijing. During this term, Malaysia and Philippines OR societies joined the APORS.

In August 1991, OR society of China hosted the second triennial APORS conference in Beijing. Regular triennial APORS meetings have been hosted by Japan, Australia, Singapore, India, and Philippine OR societies. In 2009, India became again the host society of APORS conference in Jaipur, Rajasthan. In 2010, Malaysia hosted a between-triennial meeting in the city of Penang. China is going to host the ninth APORS meeting in Xi'an. APORS2015 is now in calling for proposals.

With two new members of Iran and Nepal OR societies having joined in 2008 and 2011, APORS now has 12 member societies and become one of the active regional groupings within the IFORS. Information sharing and exchanges within APORS have been increasingly strengthened to promote OR development in its members and outreach within IFORS.

MEMBER SOCIETIES

As APORS is a regional grouping of OR societies within IFORS, member societies of

AP regions will automatically become part of APORS. By the end of 2011, APORS consisted of 12 member societies: Australia (ASOR), China (ORSC), Hong Kong (ORSHK), India (ORSI), Iran (IORS), Japan (ORSJ), Korea (KORMS), Malaysia (MSORMS), Nepal (ORSN), New Zealand (ORSNZ), Philippines (ORSP), and Singapore (ORSS).

The *Australia Society for Operational Research (ASOR)* (<http://www.asor.org.au>) was founded on January 1, 1972, and has about 400 members nationwide. ASOR serves the professional needs of OR analysts, managers, students, and educators by publishing a National Bulletin and National and Chapter Newsletters. The society also serves as a focal point for operations researchers to communicate with each other and to reach out to other professional societies. The current term ASOR President is Prof. Baikunth Nath of the University of Melbourne. ASOR is composed of six branches or chapters in the nation's states, which organize their activities. Every two years, ASOR calls a national conference. As founding member of APORS, ASOR hosted the fourth APORS meeting in Melbourne in 1997. In 2011, the society successfully hosted IFORS triennial conference at the same city with more than 1200 participants. This is the first IFORS meeting in the Oceanic. An academic journal *ASOR Bulletin* is published in four issues every year.

The *Operations Research Society of China (ORSC)* (<http://www.orsc.org.cn>) was founded in 1980 and became a member of IFORS in 1982. Now ORSC has more than 1200 registered members and around 500 are active. The current president is Prof. Yaxiang Yuan who is also the current APORS president. Within ORSC, 13 special interesting groups have been set up, for example, Mathematical Programming, Queuing Theory, Reliability, Decision Science, Uncertainty Systems, Intelligence Computing, Scheduling, Graph Theories & Combinatorics, Fuzzy Optimization, Game Theory, Financial Engineering, and recently the Computational System Biology. These technical groups meet once or every other year, apart from national society meeting once every other year, for a national conference or international conference.

ORSC has two official journals: *OR Transactions*, a quarterly journal (Prof. Yaxiang Yuan as Editor-in-Chief), that publishes original theoretical articles and *OR and MS*, a bimonthly journal (Edited-in-Chief by Prof. Xiang-Sun Zhang) that aims at both theoretical research and innovative application. Traditionally, every four years, ORSC will have a National Congress supplemented by a biannual academic conference. The most recent quadrennial meeting was held in 2008 attended by over 450 participants. The Congress featured plenary talks, paper presentations, the ORSC election, several prize competition, and committee meetings.

The *Operations Research Society of Hong Kong (ORSHK)* joined IFORS in 1980 with 120 members. The current President is Prof. Yan Chong Chen. As an active APORS member society, ORSHK work closely with universities and industries in Hong Kong to encourage operational research applications in areas such as logistics, shipping, manufacturing, and management. ORSHK has not created its web site.

The *Operational Research Society of India (ORSI)* (<http://www.orsi.in>) was founded in 1957 to provide a forum for the Operational Research Scientists as well as an avenue to widen their horizon by exchange of knowledge and application of techniques from outside the country. At present, the Society has 10 operating chapters located in Agra, Ahmedabad, Bangalore, Chennai, Delhi, Durgapur, Jamshedpur, Kolkata, Mumbai, and Tirupati.

The Society publishes a quarterly journal *OPSEARCH*, which brings out high quality and state-of-the-art papers in OR. The journal enjoys a wide spectrum of readership in both the India and abroad, covering academics, professionals, as well as industrial/service sector organizations. ORSI is conducting an examination on Graduate Diploma in Operational Research since 1973.

The *Iranian Operations Research Society (IORS)* (<http://www.iors.ir>) was established in 2005 and became the fiftieth national member of IFORS on December 16, 2009. Currently, the Society is managed by its second Executive Council to be replaced by the third Executive Council in May

2010 (the term for the council is three years). IORS currently has 134 members, mostly composed of college scholars and industry professionals. The scholars are mostly drawn from Mathematics, Industrial Engineering, and Management Sciences departments.

IORs publishes the international scientific journal, *Iranian Journal of Operations Research* (IJOR). The publication is semianual and was first published in 2009. IORS also organizes the annual International Conference of Iranian Operations Research Society. The latest International Conference of Iranian Operations Research Society on May 5–6, 2010, at the Amirkabir University of Technology, Tehran, in which ~200 were selected for oral presentations and 100 were selected as posters.

The *Operations Research Society of Japan* (ORSJ) (<http://www.orsj.or.jp>) was established on June 15, 1957. In addition to such pursuits as research on OR theory and the development of methodologies, the ORSJ explores the practical use of methods applicable to specific problems occurring in the world of business and government offices. It also actively promotes exchanges of information among its members as well as interaction on an international level. Headquartered in Tokyo, ORSJ has six chapters around the country. Each regional chapter conducts its own slate of activities, including lectures and research projects. Now ORSJ has over 2000 regular members and 170 student members, as well as 60 corporate members. *Journal of the Operations Research Society of Japan* (JORSJ) is a professional journal in Japanese language issued four times a year. From 2010, JORSJ are freely accessible at ORSJ webpage <http://www.orsj.or.jp/~archive/>. ORSJ organized two national meetings each year in spring at Tokyo and fall in elsewhere in Japan.

The *Korean Operations Research and Management Science Society* (KORMS), (<http://www.korms.or.kr>) established in June 1977 under the jurisdiction of the Ministry of Science and Technology, is a professional organization in Korea for individuals and organizations interested in the fields of OR or

Management Science (MS). KORMS has six special interest groups (SIGs). The activities of each SIG include holding international and domestic conferences; running various seminars, and publishing newsletters and proceedings. The SIGs are on Management Information Systems, Information Technology, Networks, Industrial Applications of Geographic Information Systems, Data Mining, and Material Handling System for New Millennium. The society also has several branches to promote regional cohesion and to maintain a balance of activities around the nation.

The Society publishes two journals and a review. The *Journal*, which publishes new and innovative research results, is published in two categories. *Journal* (A) is published four times a year, while *Journal* (B) is published in English twice a year and focuses on theoretical issues. The *Review* is a biannual publication that covers applied research, surveys, and articles on MS/OR trends.

KORMS holds conferences twice a year. The Fall Conference is held in the Seoul metropolitan area, while the Spring meeting convenes in other areas of the country. The conference programs include technical sessions, invited lectures, and tutorials. General assembly meetings are also held during the conferences; these comprise both regular and special sessions.

The *Management Science/Operations Research Society of Malaysia* (MSORSRM) (<http://www.msorsrm.org>) was registered as a professional society in July 1986. MSORSRM is the thirty-sixth national member of the IFORS, ninth national member of the APORS, and member of the Confederation of Scientific and Technological Association in Malaysia (COSTAM). MSORSRM offers five types of membership, that is, ordinary member, student member, corporate member, life member, and fellow. MSORSRM is currently planning for APORS triennial conference in 2015 in Kuching Sarawak, Malaysia.

The *Operational Research Society of Nepal* (ORSN) (<http://orsn.org.np/>) is founded in 2007 to capitalize its applicability in the related fields and also to develop a common platform extensively for the MS professionals. ORSN officially joined IFORS and hence

APORS in November 2011 and become the youngest member society of the association. In February 2012, ORSN held its first international conference in Kathmandu, Nepal, in which ORSN invited IFORS immediate past president, chair of the IFORS OR for Development Committee, and several internationally prominent professors from INFORMS. ORSN focuses OR methodologies and applications in the practical topics of sustained development, such as in financing, marketing, tourism, healthcare, among others.

The *Operational Research Society of New Zealand (ORSNZ)* (<http://www.orsnz.org.nz>) was founded in 1965 as a nationwide not-for-profit body comprising academics and public servants, consultants, and practitioners in industry. It currently has over 150 members within New Zealand and is affiliated to APORS, IFORS, and the Royal Society of New Zealand. The primary aim of the Society is to promote OR and MS in New Zealand to industrial and academic groups. The Society publishes a regular newsletter; organizes branch meetings in Auckland, Wellington, and Christchurch; and runs an annual two-day conference in late November/early December, with the conference location shifting between the University of Auckland, the University of Canterbury in Christchurch, and the University of Victoria in Wellington.

The *Operational Research Society of Philippines (ORSP)* (<http://www.orsp.org.ph/>) was launched in 1987 and held the first ORSP symposium in June. In December 1990, the ORSP went international and hosted its first International Conference on OR/MS (ICORMS). At about the same time, the first issue of the *Philippine Journal of Operations Research* was published. In 1997, ORSP hosted the second ICORMS, conducted jointly with the IFORS' annual International Conference on OR in Development (ICORD). Another international event, the IFORS Workshop for Teachers in Developing Countries, was successfully organized by ORSP in July 2004. In 2006, ORSP played host for the first time to the APORS) Seventh Conference. ORSP owns regular, honorary, and institutional members.

The *Operational Research Society of Singapore (ORSS)* was officially registered in November 1975. The founding president of ORSS, Professor Chew Kim Lin, is also thus far the longest serving president. The ORSS was subsequently admitted as a full member of the IFORS in 1978. ORSS is also a founding member society of APORS in 1984. It grew to a size of about more than 100 members in the late 1980s. Since then, it has maintained this number of members, though, with increasing proportion of life members. The individual members come from the academia, government agencies, and industry. A fair number of its life members are in the defense (military) and national security organizations. One of the objectives for ORSS is to promote the practical use of OR in industry, commerce, and public administration.

RUNNING THE APORS, JOURNALS, AND AFFILIATED ORGANIZATIONS

APORS, like other regional groupings within IFORS, does not support a secretariat for daily operation. Instead, its president and his/her national society provide secretariat function in communicating with other member societies. So far, APORS does not have any funding source. Therefore, the council meeting, generally during APORS and IFORS conferences, plays critical part in discussing some important matters and making decisions on election, triennial conference host, collaborative activities, and so on.

APORS officers are elected once every three years during the APORS triennial conference. The executive committee includes president, vice president, secretary, treasurer, and a representative to IFORS as IFORS Vice President for APORS. The past and current term APORS presidents are

- 1986–1988, Woong Bae Rha (Korea)
- 1989–1991, Guang-Hui Hsu (China)
- 1992–1994, Masao Iri (Japan)
- 1995–1997, Santosh Kumar (Australia)

1998–2000, Kim Lin Chew (Singapore)
 2001–2003, S.P. Mukherjee (India)
 2004–2006, Elise A. del Rosario (Philippines)
 2007–2009, Chun Kwong Han (Malaysia)
 2010–2012, Yaxiang Yuan (China)
 2013–2015, Illias Mamat (Malaysia).

APORS publishes a bimonthly journal *JAPOR (Journal of Asia-Pacific Operational Research)*, <http://www.worldscientific.com/worldscinet/apjor>. The journal places submissions in general, theoretical, OR practice, reviewer survey, OR education, and communications including short articles and letters. Emphasis is on originality, quality, and importance, with particular emphasis given to the practical significance of the results. Practical papers, illustrating the application of OR, are of special interest. Being recognized as an official publication of APORS, the journal is edited, managed, and published by ORSS in Singapore.

The ORSS started publishing the inaugural volume of *Asia-Pacific Journal of Operational Research (APJOR)* in 1984, and this was subsequently adopted as the official journal of the APORS. APJOR provides a forum for practitioners, academics, and researchers in OR and related fields, within and beyond the AP region. It is currently published by the World Scientific Publishing Co., and the ORSS continues to play an active management role in running this journal.

To promote the cooperation with OR colleagues in the AP area and the rest of the world, the Asia-Pacific Operations Research Center (APORC) was set up in 1995 in Beijing. The center is affiliated with the Chinese Academy of Sciences (CAS) and APORS. Major activities of the APORC include receiving academic visitors from APORS members and organizing regional seminars and workshops.

ACTIVITIES

Since its founding, APORS organizes a regular triennial academic conference every three

years. The past APORS meetings, locations, and hosts are listed in the following table.

Year	Venue	Host/Chair
1988	Seoul, Korea	Woong Bae Rha
1991	Beijing, China	Guang-Hui Hsu
1994	Fukuoka, Japan	Masao Iri
1997	Melbourne, Australia	Santosh Kumar
2000	Singapore	Kim Lin Chew
2003	New Delhi, India	S.P. Mukherjee
2006	Manila, Philippines	Elise A. del Rosario
2009	Jaipur, India	Bani K. Sinha
2010	Penang, Malaysia	Illias Mamat
2012	Xi'an, China	Yaxiang Yuan
2015	Kuching, Malaysia	Illias Mamat

With support from APORS member societies, China and Australia hosted successfully the IFORS1999 in Beijing and IFORS2011 in Melbourne.

By 2011, APORC has organized 10 symposiums entitled “International Symposium on Operations Research and Applications in Engineering, Technology and Management (ISORA)” that attracts regular attendees from APORS member societies. One of the main features of ISORA is that the series meetings have been located in different cultural environment in virtue of the cultural diversity of China. The first ISORA was launched in Beijing in 1995 and then followed by nine symposiums held at different towns in China. Among these places, Lhasa in Tibet, Urumqi in Xinjiang, Lijiang in Yunnan, and Dunhuang in Gansu are the most attractive ethnic minority areas. The symposiums bring OR professionals to local people and universities and meanwhile giving participants an opportunity to enjoy the highlight of the local culture. ISORA proceedings are collected into the “Lecture

Notes in Operations Research” published locally by the World Publishing Corporation.

Looking forward, we believe that APORS will play an increasingly strong role in the

future in linking OR communities in the region and demonstrating OR best practices toward a sustainable world.

THE BULGARIAN OPERATIONAL RESEARCH SOCIETY

NICOLA YANEV
Faculty of Mathematics and
Informatics, University of
Sofia "St. Kl. Ohridski",
Sofia, Bulgaria

The Bulgarian Operational Research Society (BORS) was registered as a scientific not-for-profit association on November 2, 1992 and became a member of the International Federation of Operational Research Societies (IFORS) and the European Association of Operational Research Societies (EURO) in 1994. It is a self-sufficient juridical body formed as a union of its members, in accordance with the Law for Persons and Family. The seat of BORS is in Sofia, Bulgaria, in the Institute of Informatics (now the Institute of Information Technology) within the Bulgarian Academy of Sciences where the permanent Technical Secretariat is located.

The main objectives of BORS are

- to help and encourage the scientific and creative development of its members and contribute to their efficient participation in scientific and technical activities within the field of Operational Research (OR);
- to encourage initiatives and carry out activities by
 - contributing to the development of research and applications of OR in Bulgaria, and to the improvement of their prestige and authority of OR researchers;
 - fostering, in the broadest sense, training and education in OR;
 - collaborating on scientific, methodological, and organizational cooperation with international and national scientific conferences, schools, and

other scientific OR activities held in Bulgaria and abroad;

- contributing to the activities of the IFORS and the EURO.

BORS membership categories are as follows

- *Qualified Member*. This refers to any person who has at least three years of professional or scientific experience or two publications in international journals or proceedings in the field of OR or other relevant fields may apply for qualified membership.
- *Honorable Member*. This refers to persons who have made significant contributions to the development of OR or the promotion of BORS.
- *Collective Member*. This is a juridical person from the public and the private sector, enterprises, banks, and other organizations, interested in BORS objectives, and/or giving grants to BORS.
- *Student Member*. This consists of students in universities, who have studied OR subjects.

Members of the society are entitled to the following privileges:

- reduced fees for conferences, seminars, and courses organized by the society;
- discounted subscription rates to the society's publications;
- opportunity to meet fellow OR practitioners.

The number of qualified members in 1992 was 19, and has risen to 40; at present, BORS has approximately 20 members.

BORS participated/sponsored/assisted in organizing the following conferences/seminars/workshops:

- IV International Conference on Mathematical Methods in OR, Sozopol, September 1997;
- regular seminars in the OR section of the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences;
- workshop on Well-Posedness in optimization and related topics, Sozopol, Borovets, September 2005.

The Presidents of BORS have been:

- 1992–1996: Peter Kenderov, academician in the Bulgarian Academy of Sciences;
- 1996–present: Nicola Yanev, Professor in Sofia University.

THE CANADIAN OPERATIONAL RESEARCH SOCIETY/SOCIÉTÉ CANADIENNE DE RECHERCHE OPÉRATIONNELLE

RICHARD J. CARON
Department Mathematics
and Statistics,
University of Windsor,
Windsor, Ontario,
Canada

ARMANN INGOLFSSON
School of Business,
University of Alberta,
Edmonton, Alberta,
Canada

VINH QUAN
Department of Mechanical
& Industrial
Engineering, Ryerson
University, Toronto,
Ontario, Canada

Founded in 1958 and incorporated on 25 February 1964, the Canadian Operational Research Society (CORS) takes a leadership role in the advancement of both the theory and practice of operational research (OR) in Canada, and it represents the Canadian OR community in the International Federation of Operational Research Societies (IFORS). We meet our objectives with our publications, Web site (www.cors.ca), and annual conference and by supporting a diverse array of local section activities. The society is bilingual and is also known as the Société canadienne de recherche opérationnelle (SCRO).

ORIGINS

In 1963, P. J. Sandiford [1] wrote on the origin and growth of CORS. That article gave a synopsis of the first five years of CORS history, and it appeared as the first article in the first issue of the Society's peer-reviewed journal, then called the *CORS Journal* and, since 1971, called *INFOR*. We present a brief

synopsis. In 1957, the only OR society in Canada was the Operations Research Society of Toronto. There was significant OR activity in both Montreal and Ottawa, and practitioners in those centers and elsewhere in Canada joined either ORSA (Operations Research Society of America) or TIMS (The Institute of Management Sciences), or both.

In 1957, IFORS was created, and the first secretary was a Canadian, Sir Charles Goodeve. Goodeve played a prominent role in the development of OR during wartime in Britain and was, at the time, the Director of the British Iron and Steel Research Association. In 1956, he was concerned that Canada would have no national society to join IFORS, and so he called on Dr. Omond Solandt, another significant Canadian figure of wartime OR in Britain, to try and form such a society. On 11 February 1958, representatives from Ottawa, Toronto, and Montreal met in Montreal and elected Solandt the provisional chair. The inaugural meeting of CORS was held in Toronto on 14 April 1958. Solandt was elected president. By the end of the first year, the Society had 161 members. The triangular CORS logo (Fig. 1) reflects the three founding OR centers in Canada (Ottawa, Toronto, and Montreal). Goodeve's desire for Canadian representation and membership in IFORS was realized on 1 July 1959.

CORS now has 403 members including 7 emeritus, 17 retired, and 106 student members. It is managed by a council composed of the president, vice president, secretary, treasurer, four councilors, immediate past president, designated representatives from each of the 13 local sections (including three student sections), and the chairs of the five standing committees. CORS is supported by a professional, part-time administrator.

PUBLICATIONS

The Society has two publications: a quarterly newsletter (*The CORS Bulletin*) and the peer-reviewed journal *INFOR*. *The Bulletin* is published in both English and French, and



Figure 1. Canadian Operational Research Society logo.

INFOR publishes articles in either English or French.

The peer-reviewed papers published in *INFOR* highlight recent advances in OR, and the journal contains many papers that combine theory and methodology with OR practice. The journal editors promise quick review of submissions. The first issue of the journal was published in December 1963 at which time it was called *The CORS Journal*. The first French article appeared in March 1968 [2]. In March 1971, the name of the journal was changed to *INFOR Journal* to reflect a partnership between CORS and the Canadian Information Processing Society (CIPS). In August 1985 (Vol. 23. No. 3), the name was shortened to the current *INFOR*. The relationship with CIPS ended with the publication in November 1999 of Vol. 37. No. 4. The journal is published quarterly and is distributed, either online or in print, to 121 institutions, 13 abstract and index companies, and 417 individuals in Canada and abroad.

The Bulletin provides information on conferences and awards, highlights member accomplishments, and publishes special articles highlighting OR activities across the country. It is a bilingual publication and is a membership benefit. The first issue was published in 1962, and it is now distributed electronically with separate French and English versions.

MEETINGS

The 1958 meeting that led to the establishment of CORS is considered its zeroth

meeting. The first meeting was in 1959 in Toronto and the meetings have since continued without interruption. The 53rd meeting was at St. John's Newfoundland in May 2011. In 1964, we hosted our first joint meeting, with ORSA as the partner. The success of this format led to future joint meetings with ORSA/TIMS, IFORS, Institute for Operations Research and the Management Sciences (INFORMS), and MITACS. A complete list of conferences can be found in [3].

AWARDS

The awards banquet is a feature of the annual meetings, and CORS has several prestigious prizes. Complete details, including recipients, can be found at www.cors.ca/prizes.

The Harold Larnder Prize is awarded annually to an individual who has achieved international distinction in OR. The winner delivers a keynote plenary address at the annual conference. Harold Larnder played a major role in the development of the radar-based air defense system and in the deployment of aircraft during the Battle of Britain [4]. He returned to Canada in 1951 to join the Defence Research Board and was the president of CORS in 1966–1967. The first winner of the Larnder prize was Eugene Woolsey (1986) and the most recent was John D. C. Little (2010).

The Omond Solandt Award is awarded to an organization, private or governmental, that is deemed to have made an outstanding contribution to OR in Canada. Dr. Solandt was the founder and first chairman of the Defence Research Board in Canada. He served as head of the Science Council of Canada, vice-chairman of the Canadian National Railways, and chancellor of the University of Toronto [5]. The first winner was Canada Packers (1978), and the most recent winner was Maplesoft at Waterloo, Ontario (2009).

The Award of Merit is awarded to a present or past member of CORS in recognition of significant contributions to the profession of OR. The first winner was Omond Solandt (1983), and the most recent winner was Bernard Gendron (2010).

The Service Awards recognize members of the Society who have made outstanding contributions of time and service.

COMPETITIONS

Each conference hosts a practice prize competition as well as a student paper competition. *The Practice Prize* is to recognize outstanding OR practice and to focus attention on the application of OR in Canada. The entries are judged by the impact on the client organization, contribution to the practice of OR, quality of analysis, degree of challenge, and quality of written and oral presentations. The most recent (2010) winners were Burcin Bozkaya, Erhan Erkut, Dan Haight, and Gilbert Laporte for their project entitled "Designing New Electoral Districts for the City of Edmonton."

The Student Paper Competition is to recognize the contribution of a paper either directly to the field of OR through the development of methodology or to another field through the application of OR. The competition showcases the high quality of OR education in Canada and the excellence of the next generation of operations researchers. The most recent winner (2010) in the open category was Ramon Alanis, School of Business, University of Alberta, for his paper, "A Markov Chain Model for the Performance of an EMS System with Repositioning." In the undergraduate category, the 2010 winners were Erin Hrycyszyn, Felicia Ng, Brent Ouwerkerk, Jimmie Tom, and Sean Tomilin, from the University of Alberta for the paper "Community Facility Services: Drive Time and Market Share Analysis."

THE CORS DIPLOMA

The CORS Diploma is awarded by CORS, in association with the participating Canadian universities, to recognize CORS members who have successfully completed a program of study that has included significant exposure to OR in the following areas: OR techniques, probability and statistics, computers and systems, and applications of OR. Over 900 diplomas have been given to students from over 25 Canadian universities.

THE WEB SITE

Much of the information in this article, and much more, can be found at www.cors.ca. This includes Refs 1, 3–5; complete lists of prize and award winners; past presidents; conference locations; and past issues of the Bulletin.

REFERENCES

1. Sandiford PJ. The origin and growth of the Canadian Operational Research Society. *CORS Journal* 1963;1(1):3–14.
2. Truchon M. Fonctionnement du modèle inter-industriel du Québec. *CORS Journal* 1968;6(1).
3. CORS Annual Conferences. <http://www.cors.ca/en/conferences/index.php>. Accessed 2011.
4. Larnder H. The early history of organized operational research. Conference on Operational Research; Halifax, NS, Canada; 1964. www.cors.ca/documents/EarlyHistory.pdf.
5. Ridler JS. From pioneer to president: Omond McKillop Solandt's rise in operational research. www.cors.ca/documents/RidlerSolandt.pdf. Accessed 2011.

THE CONDITION-BASED PARADIGM

WILLIAM Q. MEEKER
Department of Statistics, Iowa
State University, Ames,
Iowa

YILI HONG
Department of Statistics,
Virginia Tech, Blacksburg,
Virginia

LUIS A. ESCOBAR
Department of Experimental
Statistics, Louisiana State
University, Baton Rouge,
Louisiana

GENERAL INTRODUCTION

Today's products tend to have high reliability and can last for long periods of time. In general, doing reliability assessment and maintenance planning based only on failure-time data is difficult because these studies tend to yield few or no failures. System or component operating-condition information, however, can be obtained in a more timely manner than the corresponding failure-time information. Thus, the condition information plays an important role in reliability assessment and maintenance for many systems, especially for products or systems that require a high level of reliability or availability. The information from condition monitoring can be used for the assessment of short- and long-term reliability and scheduling of maintenance like repairs and replacements.

In practice, various system condition variables can be monitored through appropriate measurement techniques. These conditions include vibration of a system, particles and chemical elements that are released into the environment by a system, fatigue-crack sizes, temperature of a system, the resistance, conductivity of certain components, and so on. Moubray [1, Appendix 4] provides various

examples and information about existing technology used for condition monitoring. When the condition variable changes, system deterioration might be detected. Deterioration can happen in physical conditions, such as tire wear, chemical concentration, or the size of a fatigue crack. Deteriorations can also be detected in system performance such as the power output of an amplifier or the intensity of an LED. Both cases are referred to as *degradation*. Physical degradation or performance degradation can be measured as a function of time, depending on the application. These repeated measurements are generically referred to as *degradation data*. Degradation data may be available continuously or at specific time points where measurements were taken.

In many applications, there is a relationship between the system degradation and system failure. The condition information is used to obtain the reliability information. For example, when the size of a fatigue crack is increasing or the tire-tread depth is decreasing, a failure could be defined to occur when the degradation level crosses a certain threshold. Failures defined by a degradation level crossing a threshold, without complete loss of functionality, are often called *soft failures*. Failures that result in complete loss of functionality (but generally not corresponding exactly with a particular level of degradation) are called *hard failures* [2, Chapter 13]. The relationship between the amount of degradation and the failure time makes it possible to use degradation data and degradation models to make inferences and predictions about failure time. Degradation data, when properly collected and analyzed, can provide much more information than failure-time data because there are quantitative measurements on all units.

RELATED LITERATURE

During the last thirty years, there have been numerous reliability studies using

degradation data. Gertsbakh and Kordonsky [3] present and discuss some early ideas on the use of degradation models in reliability. Through the 1990s and continuing to the present, various statistical methods have been developed to get reliability information from degradation data. These methods are now beginning to be available in commercial statistical software. Murray [4] fits models to the sample paths for individual units and extrapolates the sample path until some failure level to obtain “pseudo failure” data. The “pseudo failures” are treated as failures and they are analyzed using the standard methodology available for failure-time data. Lu and Meeker [5], and Meeker *et al.* [6] use a random effects model to describe unit-to-unit variability and discuss the failure-time distribution induced by the degradation model. In some areas of application, it is useful to model the stochastic behavior in the sample paths. For this purpose, the Wiener process [7] and the Gamma process [8] are useful in stochastic modeling. Wang [9] gives a brief survey of physical degradation models.

EXAMPLES OF DEGRADATION DATA

Here, we give two examples of degradation data, which will be used to illustrate the

models in the section titled “Statistical Models and Methods for Degradation Data.”

Example 1 [Laser life data]. Figure 1 is a plot of the degradation data from 15 laser units [2, Example 13.10]. These devices, designed for telecommunication applications, have a feedback circuit to maintain constant light output. Current is increased over time to compensate for the degradation in the lasing material. The operating current was measured over time and the percentage increase in operating current was reported. For this application, a 10% increase in percentage was the specified failure level.

Example 2 [Outdoor weathering data]. Figure 2 shows damage to an epoxy compound as a function of the effective dosage for a subset of the specimens, as described in Vaca-Trigo and Meeker [10]. The damage to the compound is caused by the exposure to the outdoor ultraviolet radiation (i.e., photo-degradation). The damage was measured repeatedly after every few days of exposure and the effective dosage was also recorded. Effective dosage, which is a number that is proportional to the number of photons absorbed into the experimental specimens, is, scientifically, the appropriate timescale

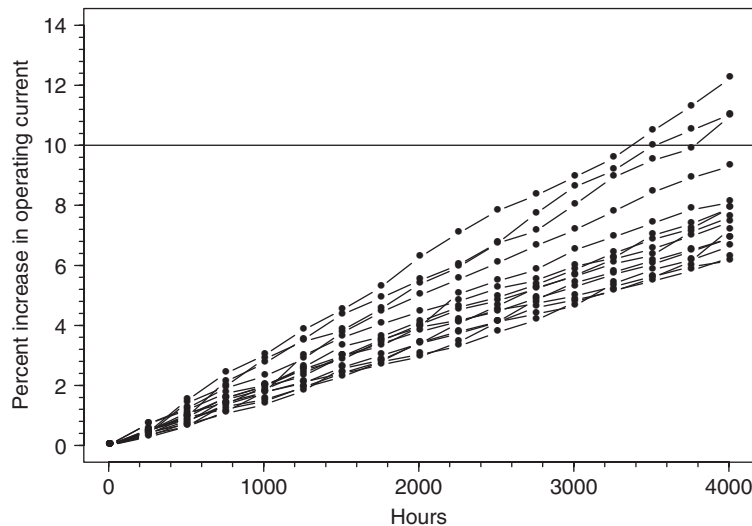


Figure 1. Laser operating current.

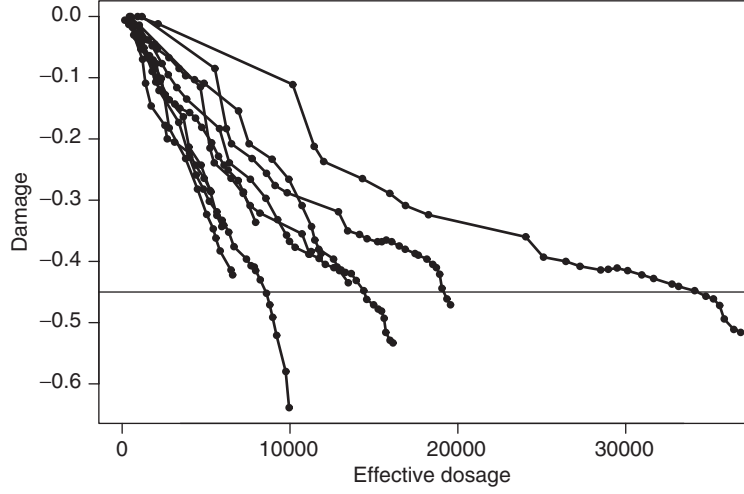


Figure 2. Plot of the damage to an epoxy compound due to outdoor weathering for a subset of specimens. A unit with damage below the critical level of -0.45 is considered a failed unit with a failure time equal to the crossing time (the timescale here is the effective dosage).

for the degradation process. There was no control of some of experimental variables such as, temperature and humidity, but these variables were also measured carefully. The degradation paths generally show more variability than laboratory test data for similar studies. For purposes of illustration, we set the definition of failure at the damage level of -0.45 , shown as a horizontal line in Fig. 2. There was a total of 13 failures that resulted for this level of damage, out of the total 55 specimens (Fig. 2 shows only a subset of specimens with 4 failures).

STATISTICAL MODELS AND METHODS FOR DEGRADATION DATA

Relationship between Degradation and Failure Time

Let $\mathcal{D}(t)$ be the degradation level at time t . Depending on the nature of the degradation process, the degradation path can be a linear, convex, or concave function of time or the amount of usage (e.g., see Meeker and Escobar [2, p. 319] for more details). A soft failure occurs when the degradation level $\mathcal{D}(t)$ reaches the failure-definition level $\mathcal{D}_f$ at time T . The failure time of the unit is T . Figure 3

gives an illustration of the possible shapes of the degradation curves, and the relationship between the degradation levels and the soft failure times.

In general, the cumulative distribution function (cdf) of T , which gives the fraction of the population failing before time t , can be obtained by

$$F(t) = \Pr(T \leq t) = \Pr[\mathcal{D}(t) \geq \mathcal{D}_f],$$

if the degradation is increasing and there is a similar definition if the degradation is decreasing. The estimate of $F(t)$ can be obtained based on the observed degradation paths of the units. The assessment of risk of failure at a future time and the corresponding maintenance scheduling can be based on the estimates of $F(t)$.

Approximate Degradation Analysis

We begin with a simple method of analyzing degradation data that is commonly used by engineers. The first step of the approximate method is to find a transformation for the degradation and/or the time such that the transformed degradation path is approximately a linear function of transformed time. No transformation is needed if the paths are

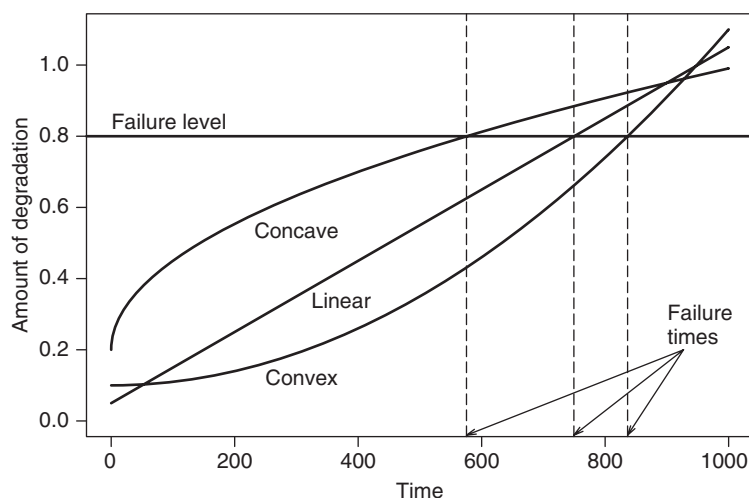


Figure 3. Illustration of the relationship between the degradation level and the failure time.

already approximately linear. Then separate simple linear regression lines are fit for each unit to predict the time at which the unit will reach the specified failure-definition level D_f . These times are called *pseudo failure times*. These pseudo failure times are then analyzed with statistical methods for failure-time data to estimate $F(t)$. The method is called the *approximate method* because the analysis is based on these pseudo failure times. See Meeker and Escobar [2, Section 13.9] for more details about the approximate method.

For simple problems, the approximate degradation analysis method is attractive because the computations are relatively simple. The approximate method is less appealing, however, when the degradation paths cannot be transformed to be approximately linear. Meeker and Escobar [2, pp. 337–338] discuss conditions for which the approximate method is adequate and potential problems with the approximate method.

Example 3 [Laser life data analysis using the approximate method]. The data for this example are shown in Fig. 1. In this case, no transformation is needed because the original paths are already approximately linear. Failure is defined as the point in time when there is a 10%

increase in current. For those lasers that did not reach this level, pseudo failure times were obtained by fitting straight lines through the sample path of the laser. These pseudo failures are analyzed using the common failure-time analysis [2, Chapter 8].

Figure 4 is a Weibull probability plot of the laser pseudo failure times showing the ML estimate for $F(t)$ and pointwise approximate 95% confidence intervals. Figure 4 also shows the Kaplan-Meier [11] estimate of the failure-time cdf, which is a step (staircase) function of time that jumps at each failure time (t_i) by an amount that is determined by the number of units at risk at t_i , the number of failures at t_i , the number of units that have been censored prior to t_i , and the original number of units in the study. The dots at each failure time are plotted halfway up each step in the Kaplan-Meier estimate, as explained in Lawless [12, Section 3.3].

Random Effects Path Model

The random effects path model is a more general and sophisticated approach to analyze degradation data. The motivation for the random effects path model is that there is unit-to-unit variability in the degradation path, which may be caused by initial conditions, material properties, and component

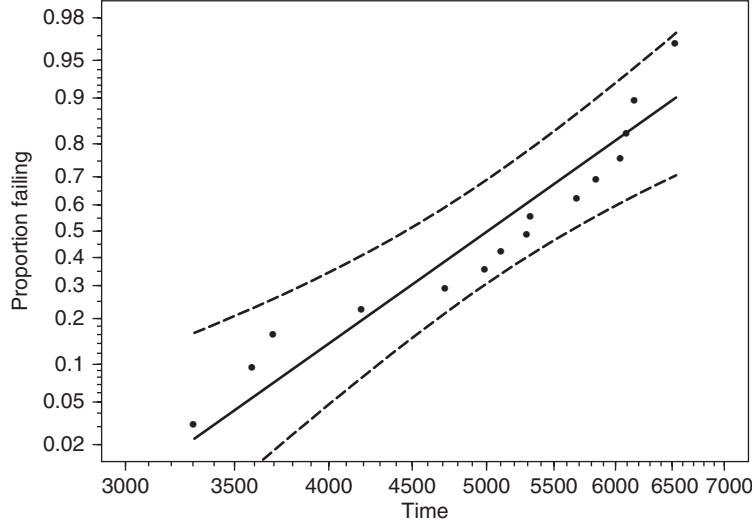


Figure 4. Weibull probability plot for the laser pseudo failure times showing the ML estimate for $F(t)$ and the pointwise approximate 95% confidence intervals. The dots track the Kaplan–Meier estimate of the cdf $F(t)$.

dimensions. That is, individual units will vary with respect to the amount of material available to wear, the initial level of degradation, amount of harmful degradation-causing material, and so on. The rate of degradation for the individual may also depend on its operating and environmental conditions. Conditional on a unit's parameters and operating/environmental conditions, however, the degradation path for the unit is assumed to be deterministic (i.e., the degradation path is deterministic function of time). This kind of degradation model is sometimes referred to as a *deterministic model*.

In practice, the values of the degradation level, $\mathcal{D}(t)$, are observed or measured at discrete points in time. The observed degradation for unit i at time t_{ij} is

$$y_{ij} = \mathcal{D}_{ij} + \epsilon_{ij},$$

where $\mathcal{D}_{ij} = \mathcal{D}(t_{ij}, \beta_{1i}, \dots, \beta_{ki})$ and ϵ_{ij} are the actual path and the residual deviation for unit i at t_{ij} , respectively. Some of the $\beta_1, \dots, \beta_k$ could be random from unit to unit to account for the unit-to-unit variability. One or more of the $\beta_1, \dots, \beta_k$, however, could be modeled as common (fixed) across all units. When the times where the degradation

is measured are not too close together, a common and plausible assumption is that the ϵ_{ij} are independent and $\text{NOR}(0, \sigma_\epsilon)$ distributed, where σ_ϵ is the standard deviation of the distribution.

The underlying model parameters, denoted by θ , consist of the parameters that determine the distribution of the random effects plus the fixed parameters. θ can be estimated using the method of maximum likelihood with the computations implemented, for example, with the `lme` or `nlme` functions in S-PLUS or R. More information can be found in Meeker and Escobar [2, Section 13.3].

For the random effects path model, the cdf of the induced failure-time distribution is

$$F(t) = F(t; \theta) = \Pr(T \leq t) = \Pr[\mathcal{D}(t) \geq \mathcal{D}_f],$$

if the degradation is increasing (and there is a similar definition for decreasing degradation). In some simple cases, it is possible to write down the explicit form of $F(t)$. For most practical path models, especially when $\mathcal{D}(t)$ is nonlinear and more than one of the $\beta_{1i}, \dots, \beta_{ki}$ parameters is random, it is necessary to compute $F(t)$ using numerical methods. For example, Meeker

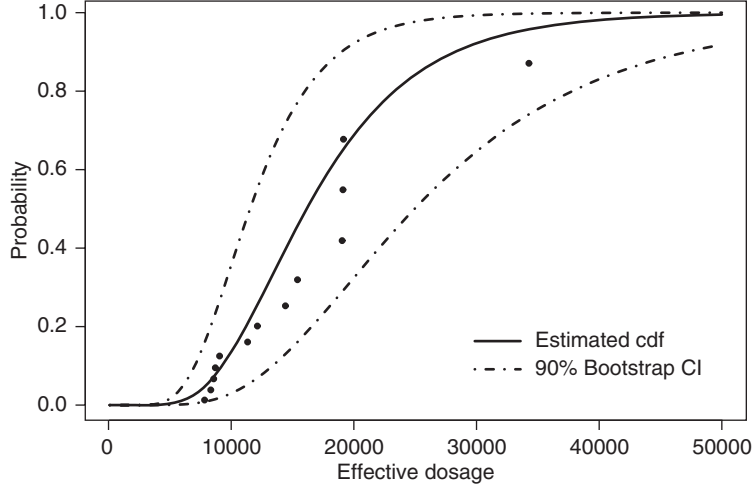


Figure 5. Degradation model estimate of $F(t)$ with pointwise approximate 90% confidence bootstrap intervals (CIs) for the outdoor weathering data, based on the random effects degradation model. The dots track the Kaplan–Meier estimate of the cdf $F(t)$.

and Escobar [2, Section 13.5] give algorithms to evaluate $F(t)$ via numerical integration and Monte Carlo simulation.

Example 4 [Linear degradation with a lognormal rate]. Here, we give a simple example where the $F(t)$ can be obtained explicitly. Suppose the actual path is

$$D(t) = \beta_1 + \beta_2 t,$$

where β_1 is fixed and β_2 has a lognormal distribution with parameters μ and σ . The cdf of T can be obtained explicitly as

$$\begin{aligned} F(t; \beta_1, \mu, \sigma) &= \Pr[D(t) > D_f] \\ &= \Phi_{\text{nor}}\left(\frac{\log(t) - [\log(D_f - \beta_1) - \mu]}{\sigma}\right), \end{aligned}$$

where $\Phi_{\text{nor}}(\cdot)$ is the standard normal cdf.

Example 5 [Analysis of weathering data using a random effects model]. This example uses the data shown in Fig. 2 which are subset of the data from Vaca-Trigo and Meeker [10]. The degradation model used here is the same as in Vaca-Trigo and Meeker [10]; that is,

$$\begin{aligned} D(t) &= \frac{A}{1 + \exp(-z)}, \quad \text{where} \\ z &= \frac{\log(t) - \beta_1}{\exp(\beta_2)}, \end{aligned}$$

t is the effective dosage, A is an unknown constant, and $\beta = (\beta_1, \beta_2)'$ is random from unit to unit. β is assumed to have a bivariate normal distribution with mean μ_β and covariance matrix Σ_β . The estimates for the unknown parameters A, μ_β , and Σ_β were obtained using the `nlme` function in R. The corresponding failure-time cdf $F(t)$ and the bootstrap confidence intervals, as shown in Fig. 5, were obtained using the algorithms in Meeker and Escobar [2, Sections 13.5.2 and 13.7].

Stochastic Models

Here, we introduce the Wiener process [7] and the Gamma process [8] which are stochastic models commonly used in the analysis of degradation data.

In the Wiener process model, the degradation path is given by

$$D(t) = \eta(t) + \sigma W[\eta(t)],$$

where $\eta(t)$ is a continuous monotone function reflecting trend and $W(t)$ is the standard Wiener process. See, for example, Cox

and Miller [13] for basic properties of the Wiener process. Under this model, the increments $\Delta\mathcal{D}(t) = \mathcal{D}(t + \Delta t) - \mathcal{D}(t)$ are independent and $\Delta\mathcal{D}(t)$ has a normal distribution with mean $\Delta\eta(t) = \eta(t + \Delta t) - \eta(t)$ and variance $\sigma^2|\Delta\eta(t)|$. For the Wiener process model case, $\mathcal{D}(t)$ has a nonmonotone and continuous path. The failure-time distribution induced by the degradation model (time to first crossing) is the inverse Gaussian distribution [14].

In some other applications [8], the degradation path $\mathcal{D}(t)$ can be modeled as a gamma process. For the gamma process model, the increments $\Delta\mathcal{D}(t) = \mathcal{D}(t + \Delta t) - \mathcal{D}(t)$ are independent and $\Delta\mathcal{D}(t)$ has a gamma distribution with a scale parameter ξ that is constant and a shape parameter equal to $\eta(t + \Delta t) - \eta(t)$, where $\eta(t)$ is a monotone increasing function of t . In this case, the degradation path $\mathcal{D}(t)$ is monotone increasing. The cdf $F(t)$ is related to the gamma function [9, Section 3].

Comparing the approximate method and the random effects path model, the stochastic process model is more advanced and requires a higher level of technical understanding. For examples where stochastic process model are used in the analysis of degradation data, see Whitmore [7], Lawless and Crowder [8], and Wang [9].

CONTINUOUS MONITORING

With the development of modern sensor and communications technology, degradation can be measured directly or indirectly in real time. Such information may also be called *system health* or *materials state* information, depending on the application. For example, aircraft engines, modern locomotives, and high-voltage power distribution transformers, system health/use rate/environmental data from a fleet of units in the field can be returned in real time to a data center. This information is used for real-time process monitoring and especially for prognostic purposes. If there is an appropriate signal in these data, rapid actions can be taken to avoid a serious system failure (e.g., by reducing the load on an unhealthy transformer). From a retrospective point of view, if some issue relating to system health arises at a

later date, it would be possible to look into the historical data that have been collected in search of causal effects. Sometimes, it is possible to find a detectable signal that could be used in the future to provide an early warning of the problem.

For example, Hahn and Doganaksoy [15, Section 9.9] describe an application involving modern locomotive engines installed with sensors that indicate the operating status such as oil pressure, oil temperature, and water temperature. Such data can be used to detect and avoid catastrophic failures by automatically shutting down a locomotive before serious damage occurs. In another application, high-voltage power transformers can be monitored by an automatic dissolved gas analyzer (DGA) system [16]. A DGA system performs periodic (typically every hour) analyses to indicate the presence of different kinds of dissolved gases in the transformer insulating oil. The chemical analyses can also measure moisture content. Certain combinations of dissolved gasses can signal a system health problem. In addition, a DGA system reports real-time dynamic loading and thermal information. This information is automatically transmitted to a control center and is useful for detecting faults that may lead to failure. Such information would enable operators to reduce the load on the transformer, or in an extreme case, to shut it down to avoid a costly failure.

FURTHER READING

This section provides a short list of useful reference in reliability. Meeker and Escobar [2, Chapters 13 and 21] introduce the concepts of degradation analysis and their relationship with the product reliability. They discuss the methods for data analysis and reliability inference for degradation data with numerous illustrations from real applications. Bagdonavičius and Nikulin [17, Chapters 3 and 13] describe statistical models and methods in degradation models and the estimation of parameters at a higher technical level.

Gertsbakh [18] has written a book for researchers and practitioners in the areas of reliability and maintainability. It

introduces the theory of reliability and how the theory can be used to develop preventive maintenance scheduling. In the area of reliability centered maintenance, Moubray [1] introduces the concepts and techniques in condition-based maintenance and condition monitoring.

Acknowledgments

Figures 1 and 4 were taken from Meeker and Escobar [2] with permission from John Wiley & Sons, Inc. We would like to thank the area editor and the reviewer, who provided comments that helped us improve this article.

REFERENCES

1. Moubray J. Reliability-centered maintenance. 2nd ed. New York: Industrial Press, Inc.; 1997.
2. Meeker WQ, Escobar LA. Statistical methods for reliability data. New York: John Wiley & Sons, Inc.; 1998.
3. Gertsbakh IB, Kordonsky KB. Models of failure. English translation from the Russian version. New York: Springer; 1969.
4. Murray WP. Archival life expectancy of 3 M magneto-optic media. *J Magnet Soc Jpn* 1993;17:(S1):309–314.
5. Lu CJ, Meeker WQ. Using degradation measures to estimate a time-to-failure distribution. *Technometrics* 1993;34:161–174.
6. Meeker WQ, Escobar LA, Lu CJ. Accelerated degradation tests: modeling and analysis. *Technometrics* 1998;40:89–99.
7. Whitmore GA. Estimation degradation by a Wiener diffusion process subject to measurement error. *Lifetime Data Anal* 1995; 1:307–319.
8. Lawless JF, Crowder M. Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Anal* 2004;10:213–227.
9. Wang X. Physical degradation models. In: Ruggeri F, Kenett R, Faltin FW, editors. *Encyclopedia of statistics in quality and reliability*. Chichester, UK: John Wiley & Sons Inc.; 2007. pp. 1356–1361.
10. Vaca-Trigo I, Meeker WQ. A statistical model for linking field and laboratory exposure results for a model coating. In: Martin J, Ryntz RA, Chin J, *et al.*, editors. *Service life prediction of polymeric materials*. New York: Springer; 2009.
11. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Amer Stat Assoc* 1958;53:457–481.
12. Lawless JF. Statistical models and methods for lifetime data. 2nd ed. New Jersey, Hoboken: John Wiley & Sons, Inc.; 2003.
13. Cox DR, Miller HD. The theory of stochastic processes. London: Chapman & Hall; 1965.
14. Chikkara RS, Folks JL. The inverse Gaussian distribution. New York: Marcell Dekker; 1989.
15. Hahn GJ, Doganaksoy N. The role of statistics in business and industry. New Jersey, Hoboken: John Wiley & Sons, Inc.; 2008.
16. Spurgeon K, Tang WH, Wu QH *et al.* Dissolved gas analysis using evidential reasoning. *IEE Proc - Sci Measure Technol* 2005;152:110–117.
17. Bagdonavčius V, Nikulin M. Accelerated life models: modeling and statistical analysis. Boca Raton (FL): Chapman & Hall/CRC; 2001.
18. Gertsbakh I. Reliability theory with applications to preventive maintenance. Berlin: Springer; 2000.

THE DECISION SCIENCES INSTITUTE

KRISHNA S. DHIR
Campbell School of Business
Berry College
Mount Berry, GA

Wickham Skinner described the origins of the discipline of Decision Sciences as “founded upon the hard-scrabble, frequently unpopular and scorned achievements of a small number of men and women who doggedly persisted in trying to find ways of making management more effective” [1]. Perhaps, Charles Babbage, the inventor of the computer, was the first to write, in 1832, about organizing the production process. Eli Whitney, Isaac Singer, Cyrus McCormick, Fredrick Taylor, and Henry Ford made major contributions to the organization of the production process, and the development of systematic planning and numerical analysis of manufacturing systems—the factories. F.W. Harris introduced modeling, Walter Shewhart demonstrated the advantages of production scheduling and controls, and LHC Tippet introduced sampling to measurement of work. Operations research demonstrated its promise in World War II, during which its principles were applied to great advantage to address the logistical challenges of moving men and materials. After great challenges of World War II, Frank and Lillian Gilbreth, Henry Gantt, and W. Edward Deming brought major advancements to time and motion studies, measurement of work, and statistical quality control [1]. In the 1950s, the discipline of decision analysis was still relatively nascent. It received a major boost in 1953, during which Irwin Bross wrote *Design for Decision* [2]. Others who made critical contributions to the discipline were Martin Starr, Jay Forrester, Herbert Simon, Oskar Morgenstern, John von Neumann, and George Dantzig. In 1960s, Howard Raiffa of Harvard University explored the process of decision making and choice when outcomes were uncertain [3]. With the advent of computers, various scholars, including coauthors

[4, 5], described how application of emerging technologies could mitigate the challenges posed by uncertainty. Linear programming software was just becoming available [6]. Before long, Lee [7] published his book on goal programming.

THE NICHE

In 1967, Dennis E. Grawoig was a chairman of the Department of Decision Sciences at Georgia State University. He was a visionary with an indomitable spirit of adventure, had an ever-expanding community of friends, and was continually creating new opportunities for students and colleagues. He quickly realized that faculty members of the various business schools, with particular interest in quantitative analysis of the decision-making process, needed a home in the form of an association that promotes their interest. Such faculty members were to be found in practically all functional areas of the broad discipline of business administration, including accounting, finance, marketing, organization behavior and theory, production and operations management, and strategic management. The new discipline of management information system was in its infancy. Grawoig understood that while addressing the managerial and administrative issues, the focus of the proposed association would cross disciplines. As such, he was keen that the association should be welcoming and open to all disciplinary orientations.

The early history of the Decision Sciences Institute has been well documented by Taylor [8]. The following narrative relies heavily on his account. On April 10, 1968, Grawoig wrote a letter to approximately a 1000 members of the faculty in various departments of collegiate schools of business at various academic institutions. In this letter, he outlined a vision for a new learned association “... for business faculty members interested in the quantitative area.” He envisaged that such an association would potentially promote scholarship through

“some publication ... to record research in our field.” He imagined that an association of this sort would meet the particular needs of the quantitatively oriented members of the business school faculty members by providing “a clearinghouse for recruitment.” He also foresaw that “meetings or seminars” could be organized under the auspices of such an association, “to provide a forum for coordination and discussion of quantitative methods programs, etc.” He inquired whether recipients of his letter would be interested in joining such an association [8]. In the 1960s, there was an Operations Management interest group active within the Academy of Management. However, some scholars in the discipline felt that at the Academy of Management the level of interest in the operational aspects of organizations was not adequate for their needs [9]. Grawoig’s query appealed to these scholars.

THE INSTITUTE IS FORMED

Grawoig’s query led to the next significant date in the history of the Institute, November 1, 1968. On that date, roughly 25 individuals responded to Grawoig’s initiative and gathered in Atlanta to hammer out a framework for a new association. Present among them were Al Heinlein, Albert Simone, George Summers, Charles Hubbard, and Roger Collons. At this session, Dennis Grawoig was elected executive director of the association, and a seven-member Executive Committee was constituted. They gave the association a preliminary name of *National Quantitative Methods Association*. However, by the time the minutes were being finalized, the formal name of the association had already emerged as the *American Institute of Decision Sciences*, with a catchy acronym of AIDS [8].

Bernard Taylor gives the credit for coining the name, American Institute of Decision Sciences, to Ann J. Hughes, a colleague of Grawoig in the Department of Decision Sciences at Georgia State University. Ann Hughes also developed the original logo for the Institute. The acronym was very apt because the very purpose of quantitative analyses of decision-making processes was to provide insights

that would *aid* decision makers in meeting their challenges. The Institute’s name was formally adopted along with the logo at the Executive Committee’s meeting in Columbus, Ohio, 2 weeks later on November 14, 1968. At this meeting, the Committee appointed Dennis Grawoig the first President of the Institute for a 1-year term. The Institute was subsequently incorporated in Atlanta, Georgia, on December 6, 1968 [8].

THE EARLY YEARS

The first conference of the Institute was held during October 30 and 31, 1969, hosted by Louisiana State University at New Orleans’ Bourbon Orleans Hotel. There were approximately 100 attendees. One month later, the Committee, now called the *Board of Directors*, met again to discuss the initiation of the institute’s first journal, *Decision Sciences*. Its first issue was published early in 1970. The Board members agreed on four areas of emphasis: (i) tutorials, (ii) original theoretical developments and extensions, (iii) application of quantitative methods in administration, and (iv) education and curriculum developments. The first editor of the journal was Albert Simone, who was later to become the President of the University of Hawaii, and subsequently, the President of the Rochester Institute of Technology [8].

The Institute rapidly gathered momentum. The second annual meeting in 1970, held in Dallas, Texas, attracted approximately 250 participants. At that point, the Institute membership had grown to 682. The same year, the Institute initiated a placement service. The Board of Directors also decided to create regional subdivisions. The first of these, the Midwest subdivision, was created in 1970. Southwest, Southeast, Western, and Northeast regional subdivisions followed, with all in place by 1972. In the following year, the Board instituted the Distinguished Service Award. In 1972, the national honorary organization for the Decisions Sciences discipline was started, called *Alpha Iota Delta*. Rodger Collons was its first President. In 1974, the newsletter, *Decision Line*, was initiated with Thad Green of Mississippi State University as its first editor [8].

A DISCIPLINE EVOLVES

The decades of the 1950s and 1960s constituted an exciting period for the discipline of Decision Sciences. Skinner describes this period as “the glory years, times of excitement and discovery ...” The leaders of the discipline “... were a kind of heroes, magicians, mysterious in dealing with forms of analysis little understood ...” [1]. As described by Skinner, dramatic changes brought about in the art and science of management through analytical techniques by a very few pioneers, who were highly successful in spite of their small population, upset the conventional arrangements of the academe. Those affected by the altered scheme of things were threatened. The counter attack described decision analytic approaches as consisting of “slick algorithms” and “merely theoretical, too impractical, limited, and simplistic to be of much use.” [1]. However, by mid-1970s, the arguments were more or less settled by dramatic benefits accrued by practitioners and industries, and Decision Sciences found acceptance among the community of scholars. Subsequently, decision scientists and members of the Decision Sciences Institute began to specialize, attending to specific problem areas such as simulation of complex technical systems through mathematical modeling and experimentation; allocation of resources through mathematical programming, including strategic planning through dynamic programming and goal programming; management of production and operations management; management of information systems; analysis of logistics; supply chain management; materials requirement planning; and service management. Application of decision analytical tools became widespread, making inroads into and interacting with other established disciplines, such as finance, marketing, medicine, biology, and even various social sciences. Specialized industries looked to Decision Sciences for benefits, as with management of health-care systems, environment, military systems, public and municipal services, and government. Today, a wide range of disciplines and methodological

approaches are routinely covered at the Decision Sciences Institute’s publications and annual meetings. These include accountability, data and information security, ethics, decision support systems, decision theory and analysis, e-commerce, expert systems, finance, hospitality management, innovative education, international business, knowledge management, logistics, management information systems, management science and operations research, organizational theory and leadership, production and operations management, project management, quality management and continuous improvement, statistical analysis and quantitative methods, strategic management, and supply chain management. Some new methods have evolved in recent years, and some old methods have been boosted with renewed interest. Among recently evolved methodologies, we find classification trees, tabu search, and neural networks. Data mining has renewed interest in some older methods, such as regression analyses and experimental design, being applied to large data sets [6].

In 1977, the Institute, still known as the *American Institute of Decision Sciences*, created the designation of Fellow to be awarded to members in recognition of distinguished achievement in the field of Decision Sciences. The award would be given to active supporters of the Institute who have made outstanding contributions to Decision Sciences in at least two of the following categories: research and scholarship, teaching and/or administration, and service to the Institute. The initial group of Fellows was the first seven Presidents of the American Institute of Decision Sciences, who held the office in turn from 1969 to 1977. These were Dennis Grawoig, George Summers, Roger Collons, Gene Groff, Albert Simone, Kenneth Uhl, and Lawrence Schkade. Since then, the list of Fellows has grown, and includes luminaries such as George Benson, Harvey Brightman, William Berry, Charles Bonini, Elwood Buffa, Richard Chase, Norman Chervany, Roscoe Davis, Ward Edwards, Fred Glover, Jack Hayya, Ira Horowitz, George Huber, Basheer Khumawala, Claude McMillan, Herbert Moskowitz, Paul Nutt, Howard

Raiffa, Aleda Roth, Thomas Schriber, Marion Sobol, Linda Sprague, Bernard Taylor, Andrew Vazsonyi, Christopher Voss, Clay Whybark, and Robert Winkler.

STEPPING BEYOND THE BORDER

By the end of its first decade, the Institute had already streamlined its constitution and bylaws. In 1980, the Executive Council was replaced by the current structure of the Board of Directors through the inclusion of the regional Vice Presidents who would be elected by the respective regional subdivisions. The Institute also began to attract members from outside the United States. Its first annual meeting outside the United States was held in November 1984 in Toronto, Canada. By mid-1980s, the American Institute of Decision Sciences was feeling the need to change its name to reflect its emerging international interests. Even as these events were underway, Dennis Grawoig announced his retirement in 1985, after a remarkable 17-year service to the Institute as its founder and only executive director. He was replaced by Carol Latta in 1986, who has continued to hold the office through the date of this writing. Two-years later, in November 1986, the Institute held its first annual meeting outside the North American continent, in Honolulu, Hawaii. The issue of the Institute's name change was finally settled by the emergence of "... widespread publicity associated with acquired immune deficiency syndrome, or as it is commonly known, AIDS." In 1986, the Institute officially changed its name to the Decision Sciences Institute and acquired a new logo based on the letter, D.

PRESIDENTS OF THE INSTITUTE

On completion of his first term as the first President of the Institute, Grawoig was persuaded to continue for a second term. Clay Whybark is the only other individual to have served two back-to-back terms as President. While the Institute has had only two executive directors, 42 individuals have served

as President of the Institute to date. These individuals are listed below:

1969–1971	Dennis E. Grawoig, Georgia State University
1971–1972	George W. Summers, University of Arizona
1972–1973	Rodger D. Collons, Drexel University
1973–1974	Gene K. Groff, Georgia State University
1974–1975	Albert J. Simone, Rochester Institute of Technology
1975–1976	Kenneth P. Uhl, University of Illinois at Urbana-Champaign
1976–1977	Lawrence L. Schkade, University of Texas at Arlington
1977–1978	Charles P. Bonini, Stanford University
1978–1979	John Neter, University of Georgia
1979–1981	D. Clay Whybark, University of North Carolina—Charlotte
1981–1982	Norman L. Chervany, University of Minnesota—Twin Cities
1982–1983	Linda G. Sprague, University of New Hampshire
1983–1984	Laurence J. Moore, Virginia Polytechnic Institute and State University
1984–1985	Sang M. Lee, University of Nebraska—Lincoln
1985–1986	Harvey J. Brightman, Georgia State University
1986–1987	William R. Darden, Louisiana State University
1987–1988	James M. Clapper, Wake Forest University
1988–1989	William L. Berry, Ohio State University
1989–1990	Bernard W. Taylor, Virginia Polytechnic Institute and State University
1990–1991	Ronard J. Ebert, University of Missouri—Columbia

1991–1992	Robert E. Markland, University of South Carolina
1992–1993	William C. Perkins, Indiana University
1993–1994	Larry P. Ritzman, Boston College
1994–1995	K. Roscoe Davis, University of Georgia
1995–1996	John C. Anderson, University of Minnesota—Twin Cities
1996–1997	Betty J. Whitten, University of Georgia
1997–1998	James R. Evans, University of Cincinnati
1998–1999	Terry R. Rakes, Virginia Polytechnic Institute and State University
1999–2000	Lee J. Krajewski, University of Notre Dame
2000–2001	Michaael J. Showalter, Florida State University
2001–2002	F. Robert Jacobs, Indiana University
2002–2003	Thomas W. Jones, University of Arkansas at Fayetteville
2003–2004	Barbara B. Flynn, Indiana University
2004–2005	Gary L. Ragatz, Michigan State University
2005–2006	Thomas E. Callarman, China Europe International Business School
2006–2007	Mark M. Davis, Bentley College
2007–2008	Kenneth E. Kendall, Rutgers University
2008–2009	Norma J. Harrison, China Europe International Business School
2009–2010	Ram Narasimhan, Michigan State University
2010–2011	G. Keong Leong, University of Nevada, Las Vegas
2011–2012	Krishna S. Dhir, Berry College
2012–2013	E. Powell Robinson, University of Houston

On April 1, 2013, Maling Ebrahimpour of University of South Florida—St. Petersburg will become the President of the Institute.

GOING GLOBAL

The Institute continued to grow and created non-US regional subdivisions. These were named Asia-Pacific, Europe, Indian subcontinent, and Mexico subdivisions, respectively. Starting in 1990, the Institute began holding biennial conferences outside the United States on a regular basis. These meetings have been held in Australia, Belgium, China, France, Greece, Mexico, Spain, Taiwan, and Thailand.

Similarly to Clay Whybark, Charles Bonini, too, attributes the success of the Decision Sciences Institute to the fact that it met a niche that was not being served by other academic or professional organizations [6]. Additionally, about the time when this Institute was formed, business education was becoming increasingly important part of academia. As they shook off their “trade school” image, business schools had started engaging in significant and rigorous academic research [10]. The Institute provided an excellent forum for faculty members to disseminate and seek collaborative efforts. Today, the Institute has become much broader in its focus. Charles Bonini points out that this broadening of scope goes beyond the fact that “the entire spectrum of quantitative techniques” has become “a significant part of management education.” [6]. The definition of Decision Sciences itself has evolved and today extends to a full gamut of emerging areas of interest, including health-care systems, supply chain systems, and even social responsibility. Applications have been most visible in production and operations management, management information systems, quality management and control, service operations, logistics, and other such areas. New perspectives continue to be developed, such as business analytics [11]. Wickham Skinner indicates that publicized areas of Decision Science, which are no less important, are “applications in finance,

marketing, biology, medicine and health systems, environmental policy, and government bureaus [1].” Nevertheless, the membership of the Institute peaked in 1991, held steady for the next about 8 years, and then began to decline at a rate of about 4% per year, until 2010. Certain innovations implemented in the annual programs in 2010 and 2011 have arrested the declining trends and increased participation by a dramatic 12.5%. However, through the years, the focus of the Institute shifted away from practical aspects of social and economic issues to topics of academic interest. Wickham Skinner examined the articles published in the Institute’s journal, *Decision Sciences*, and found that during 2007, an Institute membership of 1842 had produced merely seven articles that would be relevant and useful to practitioners [1].

THE INSTITUTE TODAY

The academic orientation of the Institute is evident in its assertions on the Institute’s official Web site, www.decisionsciences.org. The official Web site describes the Institute as “a professional organization of academicians and practitioners interested in the application of quantitative and behavioral methods to the problems of society.” However, the Web site adds that the Institute “plays a vital role in the academic community by offering professional development activities and job placement services.” In August 2008, the Board of Directors of the Institute adopted the following mission statement: “The Decision Sciences Institute advances the science and practice of decision making. We are an international learned society with an inclusive and cross-disciplinary philosophy. We are guided by the core values of high quality, responsiveness, and professional development.” The Institute’s vision is articulated as “The Decision Sciences Institute is dedicated to excellence in fostering and disseminating knowledge pertinent to decision making.” In its activities, the Institute places great emphasis on innovative education through focus on curriculum development, pedagogy, and career development. Nearly 90% of the Institute’s members are academics. A great

proportion of stakeholders among the academics are untenured faculty members who are under pressure to publish and establish themselves as scholars. Skinner describes them as the “slaves” of the academic society [1]. It is impractical for untenured faculty members to alter the priorities of the academic society. Skinner calls on those members of the academy and Institute who are tenured and those who administer the academic system to change the direction of business education and, indeed, Decision Science.

A CALL TO RELEVANCE

He reminds us that the purpose of Decision Science is to “help decision makers make better decisions.” [1]. To return the practical, “real” world from a culture of writing papers for academic peers, the Institute members need to reexamine their methodologies, reemphasize research based on empiricism, and reconnect with the practitioner. Skinner had a vision of the Decision Sciences Institute spearheading the effort to redefine the “mission, objectives, processes, and practices of Decision Sciences . . .” We exist in an environment of constant and rapid change characterized by “a high density of big, serious, varied and widely dispersed societal problems.” [1]. He cites specific issues that the discipline of Decision Science should take on to return to its glory years. These include issues pertaining to declining standard of living in the United States, rise in the negative balance of trade, climate change, weakness of the public schools, engineering and research moving offshore, federal deficits, continuing the increase in the volume of US manufacturing, apparent distaste of our workforce to work in factories, counterterrorism, and shortages of oil, food, and water [1]. This vision has implications for the way in which we practice Decision Science. Instead of engaging solitary academic research on hypothesized problems, we would have to talk to the practitioners, work in teams, be willing to cross into the boundaries of unfamiliar disciplines, collaborate with those who apply Decision Science in the “real” world—the practical decision makers, and engage in empirical research. Moreover, we have to collaborate with those who

are in disciplines different from our own, and share our intellectual property with others.

REFERENCES

1. Skinner W. Decision Sciences and the Decision Sciences Institute. *Decis Line* 2012;42(1):5–8.
2. Bross I. Design for decision. New York: The Macmillan Company; 1953.
3. Raiffa H. Decision analysis: introductory lectures on choices under uncertainty. Boston (MA): Addison-Wesley; 1968.
4. Bierman H., Hausman W.H., Bonini CP. Quantitative analysis for business decisions. Homewood (IL): R.D. Irwin; 1969.
5. McMillan C, Gonzalez R.F. Systems analysis: a computer approach to decision models. Homewood (IL): R.D. Irwin; 1968.
6. Yu W.-B. A conversation with . . . Charles Pius Bonini. *Decis Line* 2009;40(2):4, 5 and 8.
7. Lee SM. Goal programming for decision analysis. New York: Auerbach Publishers; 1972.
8. Taylor BW. The history of the Decision Sciences Institute. *Decis Line* 1989;20(4):26–28.
9. Christodoulidou, N. A conversation with . . . D. Clay Whybark. *Decis Line* 2008;39(3):4–6.
10. Benson G. The evolution of business education in the U.S.? In: Dhir K.S., editor. The Dean's perspective: issues in academic leadership in schools of business. Atlanta, Georgia: Decision Sciences Institute; 2008.
11. Evans JR. Business analytics: the next frontier for decision sciences. *Decis Line* 2012;43(2):4–6.

THE EXPONENTIALLY WEIGHTED MOVING AVERAGE

MARCUS B. PERRY
Department of Information
Systems, Statistics and
Management Science, The
University of Alabama,
Tuscaloosa, Alabama

DEFINITION

The exponentially weighted moving average (EWMA) is often applied to a time-ordered sequence of random variables. It computes a weighted average of the sequence by applying weights that decrease geometrically with the age of the observations. The EWMA is parameterized by the following. Consider the $n \times 1$ random vector $\mathbf{x}$ given by

$$\mathbf{x} = [x_1, x_2, \dots, x_n]', \quad (1)$$

with mean vector $\boldsymbol{\mu}$ and finite $n \times n$ autocovariance matrix Σ . The EWMA is defined by the linear transformation

$$\mathbf{z} = \mathbf{C}\mathbf{x} + z_0\mathbf{b}, \quad (2)$$

where $\mathbf{z}$ is $n \times 1$ and

$$\mathbf{C} = \begin{pmatrix} \lambda & 0 & 0 & \dots & 0 \\ \lambda(1-\lambda) & \lambda & 0 & \dots & 0 \\ \lambda(1-\lambda)^2 & \lambda(1-\lambda) & \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda(1-\lambda)^{n-1} & \lambda(1-\lambda)^{n-2} & \lambda(1-\lambda)^{n-3} & \dots & \lambda \end{pmatrix} \quad (3)$$

is a known $n \times n$ matrix and $\mathbf{b}$ is a known $n \times 1$ vector having the form

$$\mathbf{b} = ((1-\lambda) \quad (1-\lambda)^2 \quad \dots \quad (1-\lambda)^n)', \quad (4)$$

where z_0 is an initial (scalar) value and denotes the starting value for the EWMA. The parameter λ ($0 < \lambda \leq 1$) is called the *smoothing* coefficient and its value in practice is often selected based upon how fast the process mean changes. For more rapid mean changes one should choose λ to be “large” and for slower (or less frequent) mean changes one should choose λ to be “small.” In applications such as forecasting, λ is typically chosen between 0.05 and 0.30 [1] and [2]. It should be noted that when $\lambda = 1$ we have $\mathbf{z} = \mathbf{x}$. This is evident from Equation (2), when $\lambda = 1$ we have $\mathbf{C} = \mathbf{I}$ (where $\mathbf{I}$ is the identity matrix) and $\mathbf{b} = \mathbf{0}$ so that,

$$\mathbf{z} = \mathbf{C}\mathbf{x} + z_0\mathbf{b} = \mathbf{I}\mathbf{x} + z_0\mathbf{0} = \mathbf{x}. \quad (5)$$

To illustrate the EWMA, Fig. 1 shows a plot of $n = 100$ observations randomly generated from a normal process with a time-varying mean. The original process $\{x_i\}$ is shown as the dotted line and the EWMA process $\{z_i\}$ ($\lambda = 0.20$) is shown as the solid line. For this realization, the initial value z_0 was set to zero. The value of $\lambda = 0.20$ was chosen by minimizing the sum of the squared one-step-ahead prediction errors. This will be discussed in more detail later.

Properties

By applying the expectation operator to Equation (2), it is easily shown that

$$\mathbf{E}(\mathbf{z}) = \mathbf{C}\boldsymbol{\mu} + z_0\mathbf{b}, \quad (6)$$

and by letting $\boldsymbol{\mu} = \mu\mathbf{1}_{n \times 1}$ (i.e., stationary mean) and $z_0 = \mu$ we have

$$\begin{aligned} \mathbf{E}(\mathbf{z}) &= \mu\mathbf{C}\mathbf{1}_{n \times 1} + \mu\mathbf{b} \\ &= \mu(\mathbf{C}\mathbf{1}_{n \times 1} + \mathbf{b}) = \mu\mathbf{1}_{n \times 1}, \end{aligned} \quad (7)$$

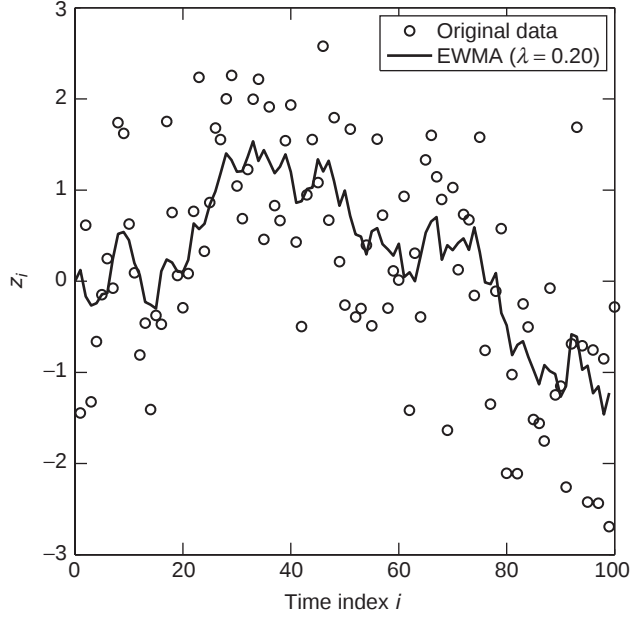


Figure 1. EWMA of a single simulated realization.

where $\mathbf{1}_{n \times 1}$ is a $n \times 1$ vector of ones. In addition, the $n \times n$ autocovariance matrix of $\mathbf{z}$ is given by

$$\text{Var}(\mathbf{z}) = \text{Var}(\mathbf{C}\mathbf{x} + z_0\mathbf{b}) = \mathbf{C}\Sigma\mathbf{C}' = \sigma_x^2\mathbf{C}\mathbf{R}\mathbf{C}', \quad (8)$$

where $\mathbf{R}$ denotes the $n \times n$ autocorrelation matrix of $\mathbf{x}$. The i th diagonal element of Equation (8) is the variance of the EWMA process at time i , and the (i, j) th off-diagonal element is the autocovariance of the EWMA process at $\text{lag}(i-j) = \text{lag}(j-i)$. It can be shown that the variance of z_i at any time $i > 0$ can be written explicitly as

$$\begin{aligned} \text{Var}(z_i) = \sigma_x^2 \lambda^2 & \left(\sum_{j=0}^{i-1} (1-\lambda)^{2j} \right. \\ & \left. + 2 \sum_{k=0}^{i-2} \sum_{\ell=k+1}^{i-1} (1-\lambda)^{k+\ell} \rho_{\ell-k} \right), \end{aligned} \quad (9)$$

where $\rho_m = \rho_{-m}$ denotes the lag m autocorrelation of the original process $\{x_i\}$. It should

be noted that $\text{Var}(z_i) \rightarrow \sigma_z^2$ as $i \rightarrow \infty$, where σ_z^2 denotes the steady-state variance of the EWMA process and can be written as

$$\sigma_z^2 = \sigma_x^2 \left(\frac{\lambda}{2-\lambda} + 2\lambda^2 \sum_{k=0}^{\infty} \sum_{\ell=k+1}^{\infty} (1-\lambda)^{k+\ell} \rho_{\ell-k} \right), \quad (10)$$

where the last term converges, as $|(1-\lambda)^{k+\ell} \rho_{\ell-k}| < 1$ for all k and ℓ . Note that for uncorrelated processes Equation (10) reduces to

$$\sigma_z^2 = \sigma_x^2 \left(\frac{\lambda}{2-\lambda} \right). \quad (11)$$

In the next section, some discussion on EWMA forecasts is provided. In subsequent sections, application of the EWMA to sales forecasting and quality control applications are presented. Finally, this tutorial will close with a discussion.

COMPUTING EWMA FORECASTS

The EWMA forecast of a future observation at time $n + \ell$ is calculated by

$$\hat{x}_{n+\ell} = z_n = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j x_{n-j} + (1-\lambda)^n z_0, \quad (12)$$

where for large n the last term is negligible and the forecast at the time $n + \ell$ can be written as

$$\hat{x}_{n+\ell} = z_n = \lambda \sum_{j=0}^{n-1} (1-\lambda)^j x_{n-j}. \quad (13)$$

Note that the forecasts generated by Equation (12) are the same for all lead times $\ell > 0$. To update forecasts with each new observation, one can use

$$z_n = \lambda x_n + (1-\lambda)z_{n-1}, \quad (14)$$

which can also be written as

$$z_n = z_{n-1} + \lambda(x_n - z_{n-1}), \quad (15)$$

where Equation (15) is written as the sum of the previous period's forecast and a fraction of the previous period's forecast error.

Suppose one uses z_n to forecast x_{n+1} , then for long running processes the variance of the one-step-ahead prediction error ($z_n - x_{n+1}$) is given by

$$\sigma_{\text{pred}}^2 = \sigma_z^2 + \sigma_x^2 \left(1 - 2\lambda \sum_{j=0}^{n-1} (1-\lambda)^j \rho_{j+1} \right), \quad (16)$$

where σ_z^2 is the steady-state EWMA variance, σ_x^2 is the steady-state variance of the original process $\{x_i\}$, and ρ_m denotes the lag m autocorrelation of $\{x_i\}$. Notice that for uncorrelated processes $\rho_m = 0$ ($m \neq 0$) and Equation (16) reduces to

$$\sigma_{\text{pred}}^2 = \sigma_z^2 + \sigma_x^2, \quad (17)$$

which is the sum of the steady-state variances of the EWMA process and the original

process. More generally, the variance of the prediction error is defined for all lead times $\ell > 0$ by

$$\sigma_{\text{pred}}^2 = \sigma_z^2 + \sigma_x^2 \left(1 - 2\lambda \sum_{j=0}^{n-1} (1-\lambda)^j \rho_{j+\ell} \right). \quad (18)$$

Initial Value for z_0

As noted previously, the influence of z_0 on z_n is negligible for large n . Thus, if n is large, one can choose any value for z_0 and its effect on z_n should be null. Some guidance for selecting z_0 in practice is given in Abraham and Ledolter [1], where these authors suggest using the arithmetic average of available historical data as the starting value.

Choice of Smoothing Coefficient

The smoothing coefficient λ determines how much the current observation affects the forecast. As mentioned previously, a small λ applies more weight to the previous observations and less weight to the current observation (and vice versa for large λ). Selection of values for λ in practice is typically accomplished via simulation. Specifically, forecasts are computed for a range of values of λ and are compared to the actual observations $x_1, x_2, \dots, x_n$. For each value of λ , the one-step-ahead forecast errors are computed by

$$e_{i-1} = x_i - z_{i-1}, \quad (19)$$

for $i = 1, 2, \dots, n$, and the sum of squared one-step-ahead forecast errors are computed by

$$\text{SSE}(\lambda) = \sum_{i=1}^n e_{i-1}^2, \quad (20)$$

where the smoothing coefficient λ that minimizes Equation (20) is then used to compute future forecasts.

To illustrate, consider the time series shown in Fig. 2. The data were simulated from an independent normal process with a time-varying mean. Suppose one considers values of $\lambda \in [0.05, 0.30]$ in increments of

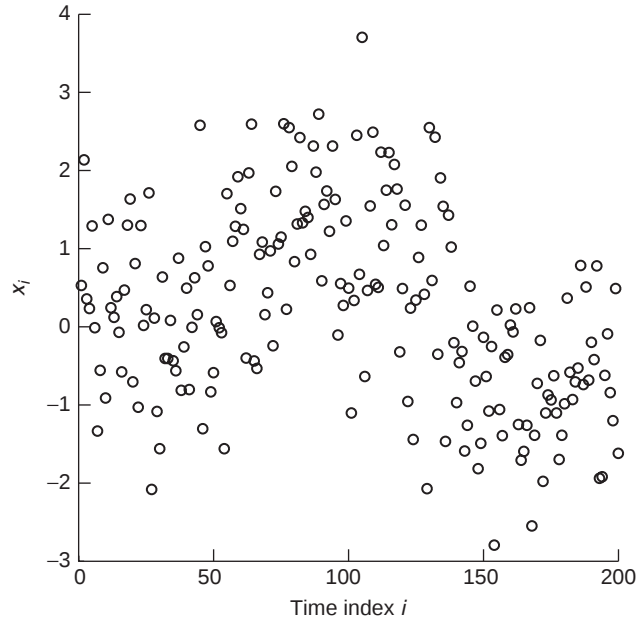


Figure 2. Single realization of simulated time series.

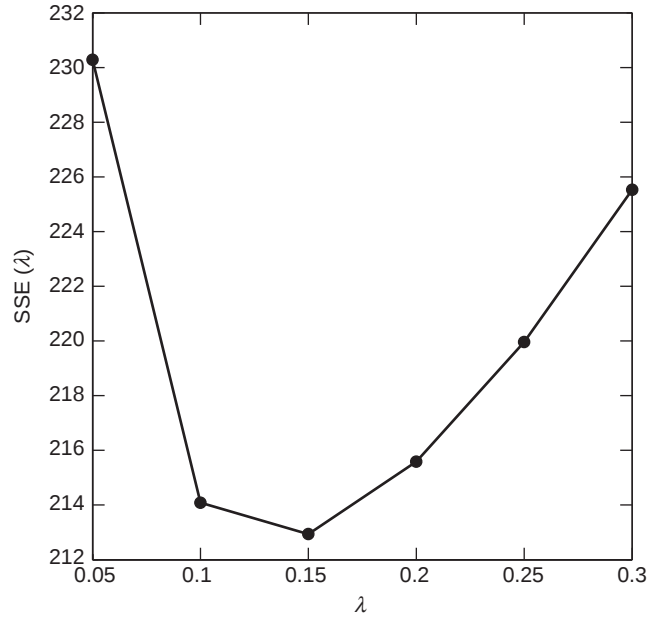


Figure 3. Plot of λ versus $SSE(\lambda)$ for simulated series.

0.05, and computes $SSE(\lambda)$ for each of these cases (i.e., $\lambda = 0.05, 0.10, 0.15, \dots, 0.30$). Figure 3 shows a plot of λ versus $SSE(\lambda)$ for the series shown in Fig. 2. Since $\lambda = 0.15$ yields the minimum sum of squared (one-step-ahead) prediction errors, this value will be used to compute future forecasts.

The EWMA process $\{z_i\}$ with $\lambda = 0.15$ is shown in Fig. 4, superimposed on the original process $\{x_i\}$. The $(n + \ell)$ th forecast ($\ell > 0$) for this process is then computed by

$$\hat{x}_{n+\ell} = z_{n-1} + 0.15(x_n - z_{n-1}). \quad (21)$$

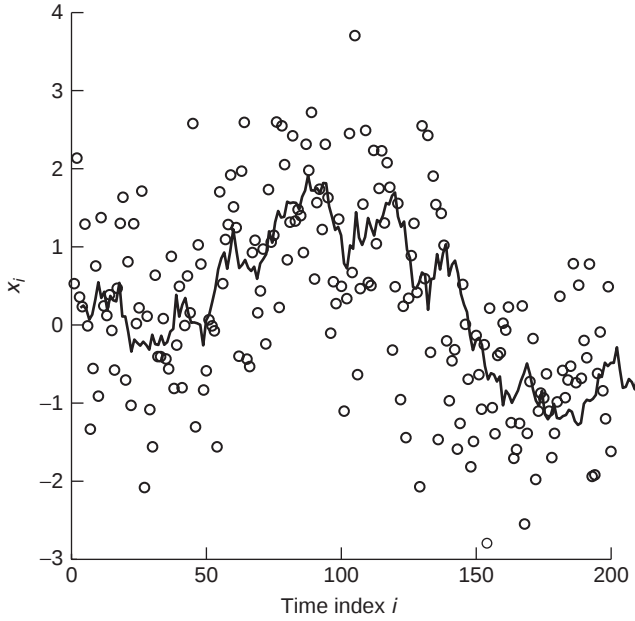


Figure 4. Original series $\{x_i\}$ and EWMA series $\{z_i\}$ ($\lambda = 0.15$) superimposed on same plot.

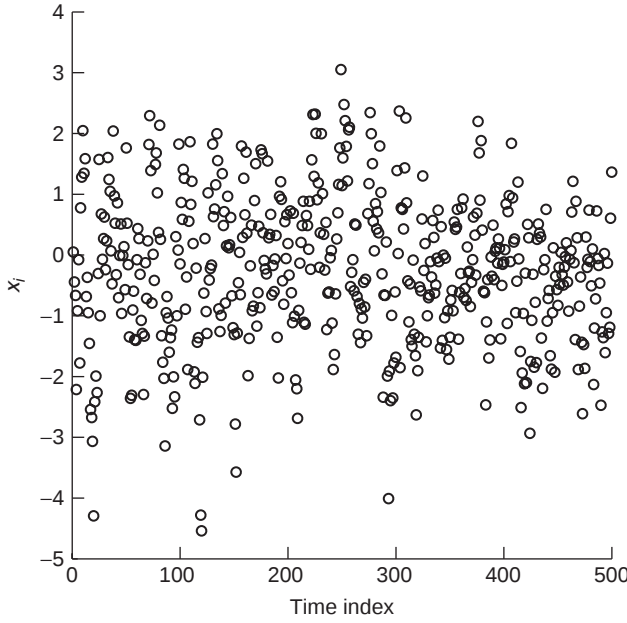


Figure 5. Single realization of simulated AR(1) process.

Consider the second example, where the process under study is well modeled by the first-order autoregressive process, or AR(1), given by

$$x_i = \mu + \phi(x_{i-1} - \mu) + \epsilon_i, \quad (22)$$

for $i = 1, 2, \dots, n$, where $|\phi| < 1$ denotes the autoregressive parameter and the ϵ_i 's are uncorrelated with constant variance σ_ϵ^2 . A single realization of the AR(1) process with $\mu = 0$, $\phi = 0.60$, and $\sigma_\epsilon^2 = 1$ is shown in Fig. 5. Considering a range of $\lambda \in [0.15, 0.90]$

in increments of 0.15, Fig. 6 shows the value of λ that yields the minimum $SSE(\lambda)$ is $\lambda = 0.60$. This is no surprise since the lag 1 autocorrelation for the AR(1) process with $\phi = 0.60$ is $\rho_1 = 0.60$. Figure 7 shows the EWMA process $\{z_i\}$ with $\lambda = 0.60$

superimposed on the original process $\{x_i\}$. For this example, $\lambda = 0.60$ should be used to compute future forecasts, where the forecast at time $n + \ell$ ($\ell > 0$) is given by

$$\hat{x}_{n+\ell} = z_{n-1} + 0.60(x_n - z_{n-1}). \quad (23)$$

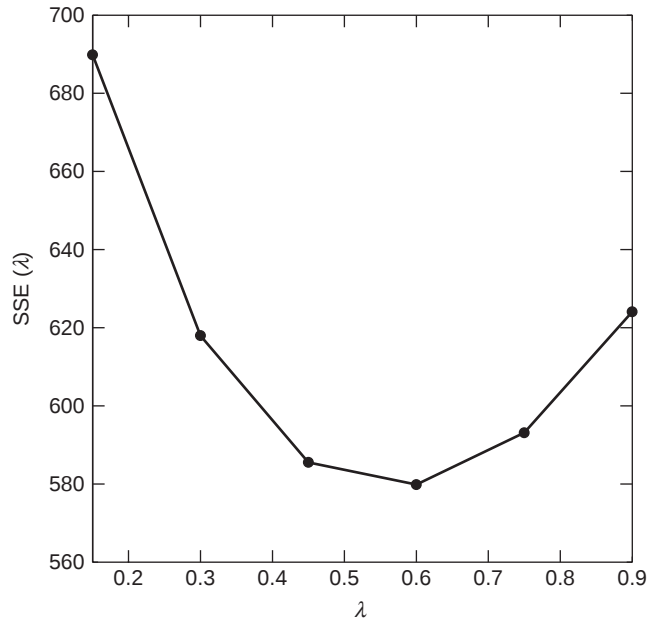


Figure 6. Plot of λ versus $SSE(\lambda)$ for realization of AR(1) process.

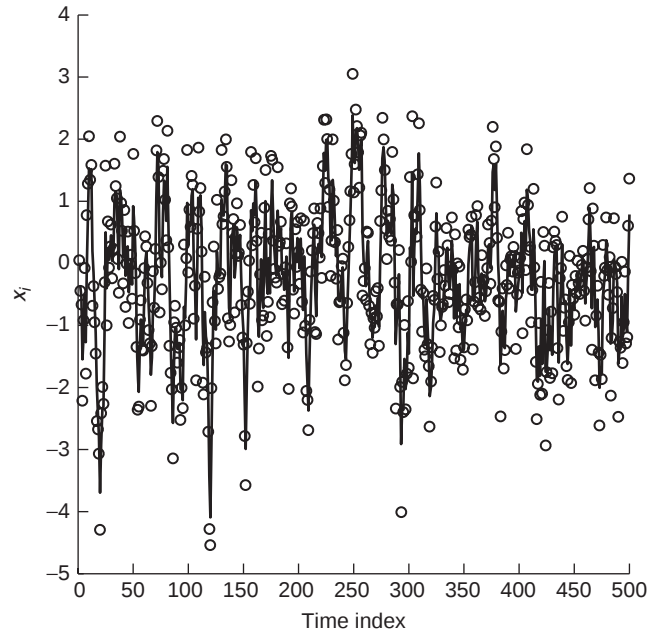


Figure 7. Original series $\{x_i\}$ and EWMA series $\{z_i\}$ ($\lambda = 0.60$) superimposed on same plot.

In the next section, some of the most common applications of the EWMA are discussed to include forecasting and application of the EWMA to quality engineering problems.

APPLICATIONS OF THE EWMA

One of the most common applications of the EWMA is generating one-step-ahead forecasts of time series. For example, consider Fig. 8, where the monthly Australian sales of rose wine is plotted over 173 months, from approximately January 1980 to April 1995. The data and its original source can be obtained from the Time Series Data Library.¹ Suppose we consider values of $\lambda \in [0.05, 0.30]$ in increments of 0.05, then the plot in Fig. 9a shows the value of λ , which minimizes $SSE(\lambda)$ is $\lambda = 0.15$. The plot in Fig. 9b shows the EWMA with $\lambda = 0.15$ superimposed on the same plot as $\{x_i\}$. The forecast for x_{174} is then given by $\hat{x}_{174} = z_{173} = 47.23$. The actual sales in the 174th month is $x_{174} = 45.00$, and the forecast error is given by $e_{173} = -2.23$.

Another common application of the EWMA is in quality control applications. Roberts [3] was the first to suggest the use of EWMA as the basis for a process monitoring and control scheme (i.e., control chart). The idea is to chart the z_i at each time i and compare to an upper and lower control limit. If z_i exceeds one of these control limits, then the process is deemed “out of statistical control.” Specifically, at each time $i > 0$, the EWMA z_i is compared to the control limits

$$\mu_0 \pm L\sqrt{\text{Var}(z_i)}, \quad (24)$$

where μ_0 is the expected value of the process, $\text{Var}(z_i)$ is defined in Equation (9) and L is a constant expressed in standard deviation units.² If z_i exceeds one of the control limits in Equation (24), then one would conclude with strong evidence that the process mean

has shifted. It is common in practice to use the steady-state control limits, which is given by

$$\mu_0 \pm L\sqrt{\sigma_z^2}, \quad (25)$$

where σ_z^2 is defined in Equation (10).

To illustrate the application of the EWMA control chart, consider the statistical monitoring of an industrial process to assess its quality over time. Suppose that the expected (or in-control) mean value of the original process $\{x_i\}$ is $\mu_0 = 0$, that samples taken over time are uncorrelated, and the variance of $\{x_i\}$ is given by $\sigma_x^2 = 1$. Suppose further that following the 50th sample collected, the process mean shifts from $\mu_i = 0$ to $\mu_i = 1$, so that the 51st sample collected is the first obtained from the changed process. Then the plot in Fig. 10a shows a single realization of this process. The plot in Fig. 10b shows the EWMA control chart; which signaled at $n = 58$, which implies eight observations were sampled from the changed process before the control chart issued a signal. Notice that the control chart limits reach steady-state after roughly 15 observations.

The above illustration is an example of how the EWMA can be used as the basis for a process change detection algorithm. Although the process in the latter example is uncorrelated, the EWMA can be used to monitor autocorrelated processes as well [4–6]. For more details on the development, design, and application of the EWMA control chart see Roberts [3], Hunter [7], Montgomery [2], and Lucas and Saccucci [8].

The above examples are not the only applications of the EWMA; however, they are common applications of the EWMA in the management science and quality engineering disciplines. In general, application of the EWMA is widespread throughout several disciplines, including health, finance, criminal justice, sports, industry, demography, and many more.

DISCUSSION

In this tutorial, the EWMA was discussed. As previously noted, EWMA is frequently

¹The Time Series Data Library is accessible at <http://www.robjhyndman.com/TSDL/>.

²The constant L is typically chosen based upon an acceptable false alarm rate.

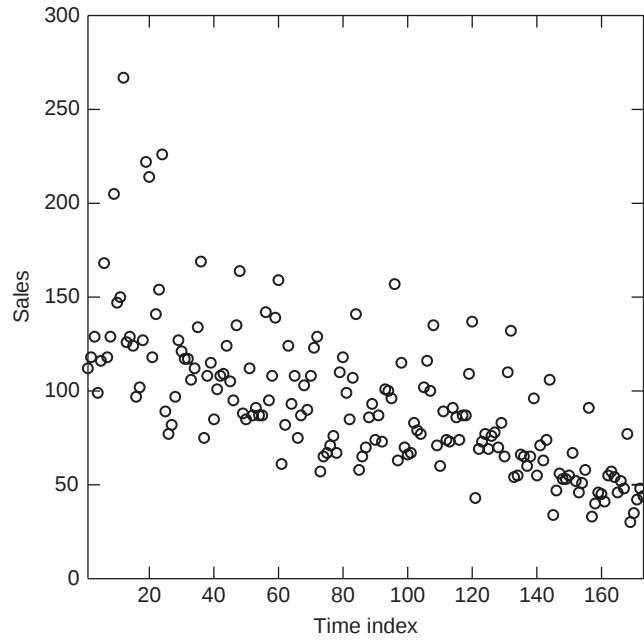


Figure 8. Plot of Australian rose wine over 173 months.

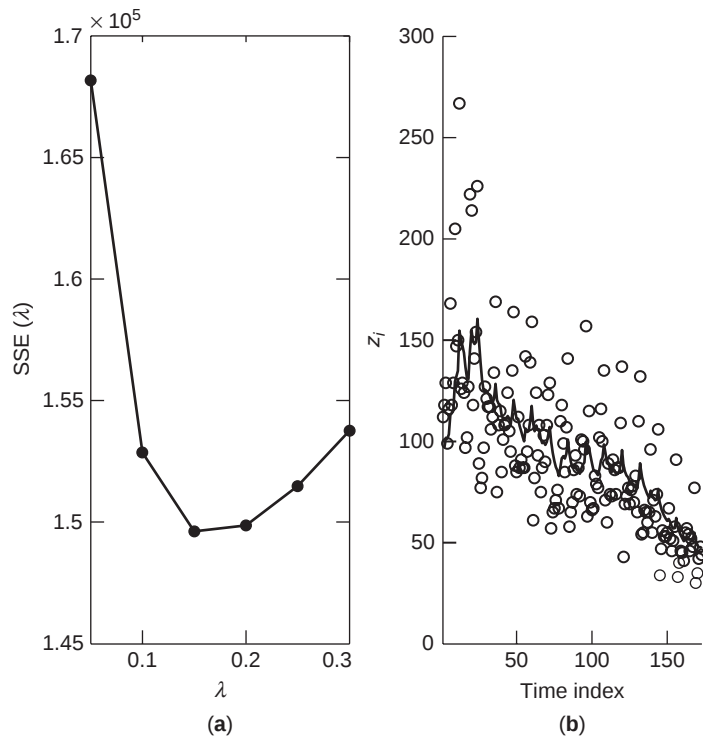


Figure 9. (a) Plot of λ versus $SSE(\lambda)$. (b) EWMA of rose wine data with $\lambda = 0.15$.

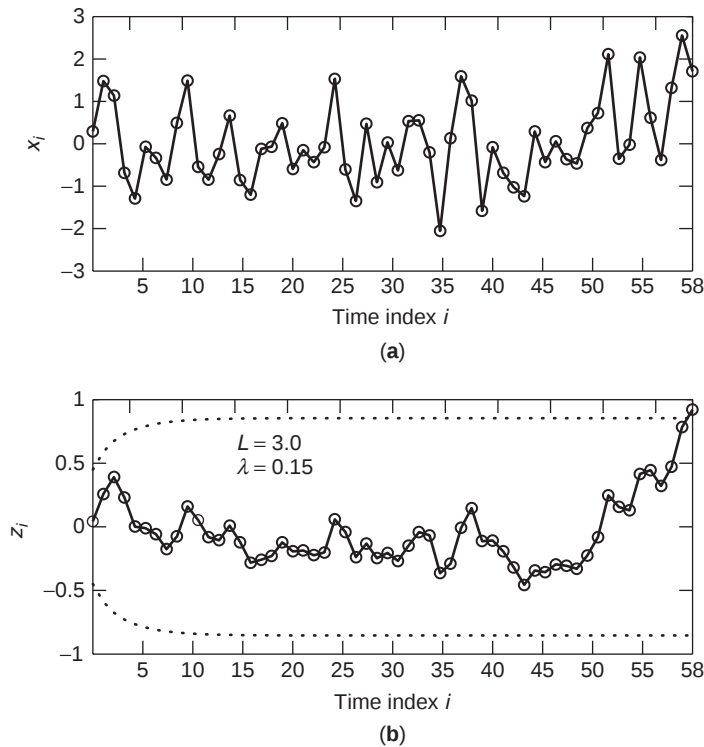


Figure 10. (a) Plot of original series $\{x_i\}$. (b) EWMA control chart with $L = 3$ and $\lambda = 0.15$. Control chart signaled at $n = 58$.

used to compute short-term forecasts of time series (e.g., sales, stocks, and so on). It is also used as a basis for developing process monitoring and control schemes. Further, some properties of the EWMA, to include its expected value and variance, as well as the ℓ -step ahead forecast error variance were provided. Some discussion on computing EWMA forecasts was also included, as well as a commonly used method for selecting a value for λ in practice. Finally, some common applications of the EWMA in the management science and quality engineering disciplines were discussed.

REFERENCES

1. Abraham B, Ledolter J. Statistical methods for forecasting. New York: John Wiley & Sons, Inc.; 1983.
2. Montgomery DC. Introduction to statistical quality control. New York: John Wiley & Sons, Inc.; 2005.
3. Roberts SW. Control chart tests based on geometric moving averages. *Technometrics* 1959;1:239–250.
4. Zhang NF. A statistical control chart for stationary process data. *Technometrics* 1998;40:24–38.
5. Montgomery DC, Mastrangelo CM. Some statistical process control methods for autocorrelated processes. *J Qual Tech* 1991;23(3):179–204.
6. Lu CW, Reynolds MR Jr. EWMA control charts for monitoring the mean of autocorrelated processes. *J Qual Tech* 1999;31(2):189–206.
7. Hunter JS. A one-point plot equivalent to the shewhart chart with western electric runs rules. *Qual Eng* 1990;2(1):13–19.
8. Lucas JM, Saccucci MS. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics* 1990;32:1–29.

THE FAILURE-BASED PARADIGM

WILLIAM Q. MEEKER
Department of Statistics, Iowa
State University, Ames,
Iowa

YILI HONG
Department of Statistics,
Virginia Tech, Blacksburg,
Virginia

LUIS A. ESCOBAR
Department of Experimental
Statistics, Louisiana State
University, Baton Rouge,
Louisiana

GENERAL INTRODUCTION

Reliability is the probability that a system, vehicle, machine, device, and so on, will perform its intended function under operating conditions, for a specified period of time [1, p. 2]. As suggested by Condra [2], reliability can also be defined as “quality over time.” It has been recognized, however, that achieving and improving high reliability requires careful focus on the time dimension, which lies beyond the standard tools used in quality control and improvement.

Reliability data are usually collected to assess product reliability. Failure-time data, which record the time-to-failure of units, are one of the most common and important types of reliability data. Statistical models and methods for such data play an important role in reliability. An important task of statistical analysis in the failure-based paradigm is to estimate product reliability based on available failure-time data. These estimates can be used to answer questions that are related to product reliability such as what will be the fraction failing for the next six months, or how much improvement in reliability can be expected if a new design is used.

Other than failure-time information, system-condition information can also be

used to assess system reliability and plan maintenance. This area is different from the failure-time analysis focus in this article. See Hong *et al.* [3] for an introduction to the condition-based paradigm.

The rest of this article describes the main sources and characteristics of failure-time data, commonly used reliability metrics, the analysis of failure-time data with explanatory variables, and the analysis of failure-time data with competing risks. Some trends in reliability data and further reading are suggested.

SOURCES OF FAILURE-TIME DATA

Failure-time data come mainly from two sources: laboratory life tests and the field. *Life test data* usually come from regular life tests or accelerated life tests. Laboratory tests are often done to study durability, wear rate, or other lifetime properties of particular materials or components. Tests are also done at a subsystem or complete system level. For example, Meeker [4] reports results from an integrated circuit life test.

An “accelerated test” can be used to obtain information in a timely manner for components that last for years or even decades. In the accelerated test, units are tested at high levels of cycling rate, temperature, voltage, stress, or some other accelerating variable to get reliability information quickly. Extrapolation based on a physically motivated model is often needed to get reliability information under the conditions of use. Nelson [5] gives examples, models, and data on accelerated test. Escobar and Meeker [6] provide a detailed review of accelerated test models.

Field data usually come from warranty databases or field tracking studies. Warranty databases, which were initially created for financial-reporting purposes, also provide a rich source of reliability information. Warranty data are generally heavily censored (i.e., not all units fail during the warranty period; see the section titled “Characteristics

of Failure-Time Data” for more details) because little or no information can be obtained after a product is out of warranty. Kalbfleisch *et al.* [7] give an example of failure-time data from an automobile warranty database.

For some products, careful field tracking can provide good reliability data. Examples include medical devices and a company’s fleet or fleets of assets. Hong *et al.* [8] give an example of a field tracking study of lifetimes of an energy company’s fleet of high-voltage power transformers.

CHARACTERISTICS OF FAILURE-TIME DATA

Failure-time data are typically censored (failure times are not known exactly). The most common reason for censoring is the need to analyze data before all units fail. Censored observations provide a bound or bounds on the actual failure times. Censoring arises in different forms such as right censoring, left censoring, and interval censoring.

For a right-censored observation, we have a lower bound on the failure time. Type I (time) censoring results when a sample of units is put on test and the units are tested simultaneously until the test is stopped at a particular time. See Meeker and Escobar [1, Examples 1.2 and 1.3] for examples. Type II (failure) censoring is less common in practice and arises when a life test is terminated after a specified number of failures have been observed. Units can be censored at different censoring times, causing multiple censoring. Multiply-censored data can result from staggered entry (units put into the field or on test at different times) or due to random censoring. Random censoring can arise from competing-risks data. For example, when a product can fail due to two or more causes of failure [1, Example 3.8] and the primary focus is on a particular cause of failure, failures from other causes can be considered as randomly censored. When the multiple causes of failure are not independent, random censoring is informative and the analysis is more complex [9,10].

Left censoring and interval censoring also arise in failure-time data. For a left-censored

observation, we have an upper bound for the failure time. For interval censoring, we know both, the lower and upper bounds for a failure time.

Truncation is similar to censoring but with an important difference. Truncation arises when only those units whose failure time lie within a certain time window are observed. The existence of units outside that window is unknown. For the censored data case, however, we have at least partial information on each unit. Like censoring, truncation can also arise as left truncation, right truncation, interval truncation, and so on. See Kalbfleisch and Lawless [11], Meeker and Escobar [1], and Hong *et al.* [8], for some examples.

Example 1 [Bearing-cage field data].

This is an example with multiple censoring. Figure 1 shows an event plot for bearing-cage fracture times for six failed units as well as running times for 1697 units that had accumulated various amounts of service time without failing. The data and analysis were given initially in Abernethy *et al.* [12]. The data are multiply censored because the failure times are intermixed with the censored times.

BASIC RELIABILITY METRICS

Here, we introduce some basic reliability metrics. An important goal of statistical analysis is to estimate one or more of these quantities and use confidence intervals to quantify statistical uncertainties.

The time-to-failure random variable is denoted by T . Here, T is a positive quantity because the failure time of a product or a component is always positive. The cumulative distribution function (cdf) of T is denoted by $F(t) = \Pr(T \leq t)$, giving the fraction of the population failing at time t . The reliability function (rf), also known as survival function, is defined by $R(t) = \Pr(T > t) = 1 - F(t)$, which gives the fraction surviving at time t . The probability density function (pdf) for a continuous random variable is defined as $f(t) = dF(t)/dt$, the first derivative of $F(t)$ with respect to t .

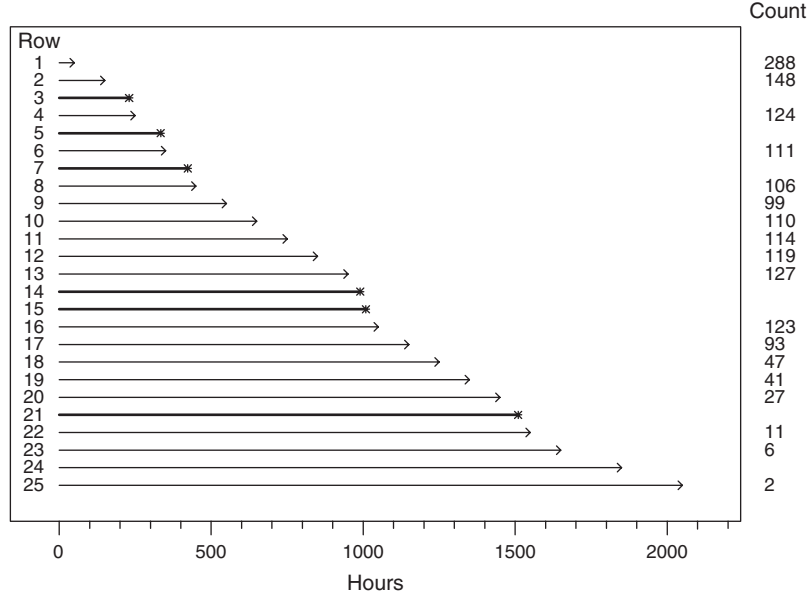


Figure 1. Event plot of bearing-cage field data (* indicates failed units, → indicates censored units).

The quantile function t_p gives the time for observing a specified proportion p of the population failing. Note that t_p is the inverse function of the cdf; that is, $t_p = F^{-1}(p)$ if $F(t)$ is strictly increasing. For example, $t_{0.10}$ is the time by which 10% of the population will fail.

The hazard function (hf), defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}$$

gives the propensity to fail in the next small interval of time for units that have survived to the beginning of that interval. There is a connection between the shape of the hf and the type of failure mode (cause of failure). For example, an increasing hf corresponds to aging/wear out and a decreasing hazard function generally suggests a mixture of defective or other weak units leading to infant mortality. The cumulative hf is defined as $H(t) = \int_0^t h(x) dx$. The relationship between $F(t)$ and $H(t)$ is $F(t) = 1 - \exp[-H(t)]$. Figure 2 gives plots

of the cdf, pdf, rf, and hf for an example distribution.

The mean time-to-failure (MTTF) also appears in the reliability literature as a measure of the central tendency of T . The MTTF is defined by $E(T) = \int_0^\infty tf(t) dt = \int_0^\infty [1 - F(t)] dt$. When the distribution of T is highly skewed, the MTTF may differ appreciably from other measures of central tendency like the median, $t_{0.5}$. MTTF, for historical reasons, is often used as a reliability metric when it should not be used. In particular, MTTF has little or no practical meaning to describe the failure-time distribution of a component or system that has only a small probability of failing during its technological life. For such components and systems, a failure probability or a lower-tail distribution quantile is much more appropriate.

NONPARAMETRIC ESTIMATION

Here, we introduce methods of nonparametric (model-free) estimation of $F(t)$. Once an estimate of $F(t)$ is available, the other functions such as $S(t)$ and $H(t)$, can be estimated

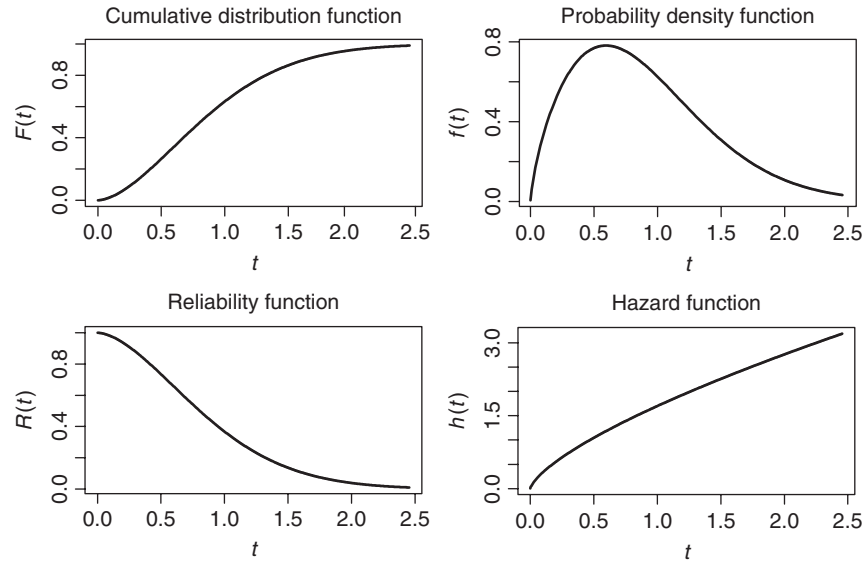


Figure 2. Plots of the cdf, pdf, rf, and hf for an example distribution.

using the relationships provided in the previous section. For complete (not censored) or right-censored data, the Kaplan–Meier (KM) estimator [13] is the most popular nonparametric estimator for $F(t)$. Details of how this estimator is computed can be found in, for example, Meeker and Escobar [1, Section 3.5]. The KM estimator is a step (staircase) up function of time that jumps at each failure time t_i by a quantity which is determined by the number of units at risk at t_i , the number of failures reported at t_i , and the number of units that have been censored prior to t_i . When there is no censoring, the KM estimate is the same as the empirical cdf estimate giving the fraction failing as a function of time. The KM estimator requires no assumptions about the form of the underlying distribution but it requires the assumption that censoring is not informative. The KM estimate is widely available in modern statistical software packages (e.g., JMP, MINITAB, R, SAS, and S-PLUS).

The KM estimator can be extended to handle data with other types of censoring and truncation. Peto [14] defined the nonparametric estimator of the cdf for arbitrary censoring, including complicated overlapping interval-censored data. Turnbull

[15] further generalized this estimator to allow for truncated data. The Peto–Turnbull estimator for censored data given in Turnbull [15] is also available in some reliability software packages.

Nelson [16] proposed a nonparametric estimator for the cumulative hf which can be converted to an estimate for $F(t)$ using the relationship $F(t) = 1 - \exp[-H(t)]$. Aalen [17] gave the corresponding asymptotic properties of this estimator.

Example 2 [KM estimate for the bearing-cage field data]. As shown in Fig. 1, the bearing-cage data come from a population of units that had been introduced into service over time. Analysts wanted to use these initial data to see if the bearing-cage design is adequate and decide if a redesign would be needed. The design-life specification was that $t_{0.10}$ be at least 8000 h. Figure 3 shows a KM estimate of the fraction failing as a function of hours. Because the largest reported time $t = 2050$ is a right-censored observation, the KM estimate is only defined up to that time. We can see that, in this application, the KM estimate, by itself, will not allow assessment of whether the design-life specification has been met.

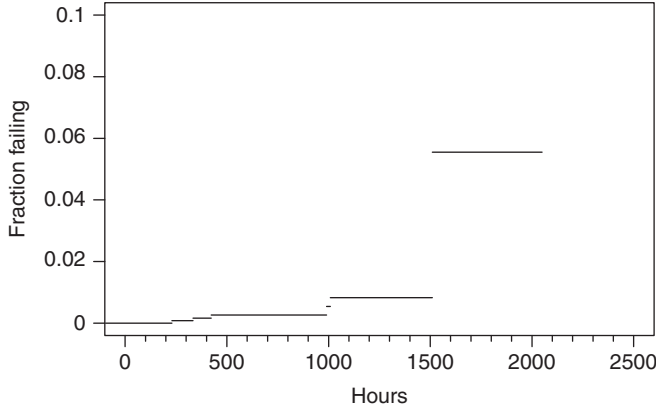


Figure 3. Plot of the KM estimate of the cdf of the bearing-cage data.

PARAMETRIC MODELS AND PARAMETER ESTIMATION

Common Parametric Distributions for Failure-Time Data

Extrapolation is often needed for reliability estimation and prediction. For example, the fraction of a product failing after three years often needs to be estimated based on one-year data. Parametric models are needed for this purpose.

The distribution of a failure time T is often modeled with a log location-scale distribution. The Weibull and lognormal are the most commonly used distributions and are members of this family. The cdf of T from a log location-scale distribution can be expressed as

$$F(t; \mu, \sigma) = \Phi \left[\frac{\log(t) - \mu}{\sigma} \right], \quad t > 0, \quad (1)$$

where $\Phi(z)$ is the standard cdf for the location-scale family of distributions (with location parameter $\mu = 0$ and scale parameter $\sigma = 1$). The corresponding pdf is

$$f(t; \mu, \sigma) = \frac{1}{\sigma t} \phi \left[\frac{\log(t) - \mu}{\sigma} \right],$$

where $\phi(z) = d\Phi(z)/dz$ is the standard pdf corresponding to $\Phi(z)$. The cdf and pdf of the lognormal distribution can be obtained by replacing Φ and ϕ above with Φ_{nor} and ϕ_{nor} , the standard normal cdf and pdf, respectively. The cdf and pdf of a Weibull random variable can be obtained by using $\Phi_{\text{sev}}(z)$

$= 1 - \exp[-\exp(z)]$ and $\phi_{\text{sev}}(z) = \exp[z - \exp(z)]$, the standard (i.e., $\mu = 0, \sigma = 1$) smallest extreme value cdf and pdf, respectively. The cdf and pdf of the Weibull random variable can equivalently, but more commonly, be written as

$$F(t; \eta, \beta) = 1 - \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right]$$

and

$$f(t; \eta, \beta) = \left(\frac{\beta}{\eta} \right) \left(\frac{t}{\eta} \right)^{\beta-1} \exp \left[- \left(\frac{t}{\eta} \right)^\beta \right],$$

where $\eta = \exp(\mu)$ is the scale parameter and $\beta = 1/\sigma$ is the shape parameter. The shape parameter β determines the shape of the Weibull hf. For example, if $\beta > 1$, the hf is increasing (corresponding to wear out). See, for example, Hahn and Shapiro [18, Chapter 8] and Meeker and Escobar [1, Chapters 4 and 5] for more details about these and other distributions that are commonly used to describe failure-time data.

Probability Plots

It is difficult for the human eye to compare data in plots like Fig. 3 with parametric cdfs like the cdf shown in Fig. 2. Probability plots display data on special probability scales such that the parametric cdf from every member of a specified family of distributions is a straight line when plotted on such a scale. For example, taking the logarithm of the Weibull quantile gives the straight line $\log(t_p) = \log(\eta) + (1/\beta) \log[-\log(1-p)]$,

which suggests plotting t on a logarithmic scale (usually on the x axis) and plotting p on a probability scale with the underlying transformation $\log[-\log(1-p)]$ (usually on the y axis), providing a Weibull probability plot. See Meeker and Escobar [1, Chapter 6] for more examples. Thus, appropriately plotting a nonparametric estimate on a probability plot, and assessing whether the plot reasonably approximates a straight line, provides a simple but effective assessment of the adequacy of an assumed distribution.

Example 3 [Probability plot of the KM estimate for the bearing-cage field data]. Figure 4 is a Weibull probability plot of the bearing-cage data showing the nonparametric estimate (the solid dots), plotted at a point half way up the jump in the KM estimate at each failure time. The dashed lines are approximate 95% nonparametric simultaneous confidence bands that help assess the sampling uncertainty in the estimate of $F(t)$. These confidence bands are obtained by using the method described in Nair [19], and are defined by the inversion of a distributional goodness-of-fit test. If a straight line can be drawn through the confidence bands, as is the case here, the distribution implied by the probability paper cannot be ruled out as a model that might have generated the data.

Maximum Likelihood (ML) Estimation of Parameters

The ML method provides a formal and objective statistical approach to fit a failure-time distribution based on an assumed model. For analysis of reliability data with censoring and using skewed distributions (e.g., the Weibull and the lognormal), the use of the ML method is strongly recommended. The idea behind the ML method is simple. For a given data set, one writes the probability of observing the data (or a function that is proportional to the probability of the data) for the proposed model as a function of unknown model parameters. This function is called the “likelihood function.” The ML estimates are those values of the parameters that maximize this likelihood function. For example, the Weibull likelihood function for the bearing-cage data is given by

$$\mathcal{L}(\mu, \sigma | \text{DATA}) = \prod_{i=1}^n \left\{ \frac{1}{\sigma t_i} \phi_{\text{sev}} \left[\frac{\log(t_i) - \mu}{\sigma} \right] \right\}^{\delta_i} \times \left\{ 1 - \Phi_{\text{sev}} \left[\frac{\log(t_i) - \mu}{\sigma} \right] \right\}^{1-\delta_i},$$

where $\delta_i = 1$ if t_i is an exact observation and $\delta_i = 0$ if t_i is a right-censored observation. Here n is the number of observations. A probability plot is also useful for displaying the ML estimate. This estimate will appear as a straight line on the Weibull probability plot (Fig. 5).

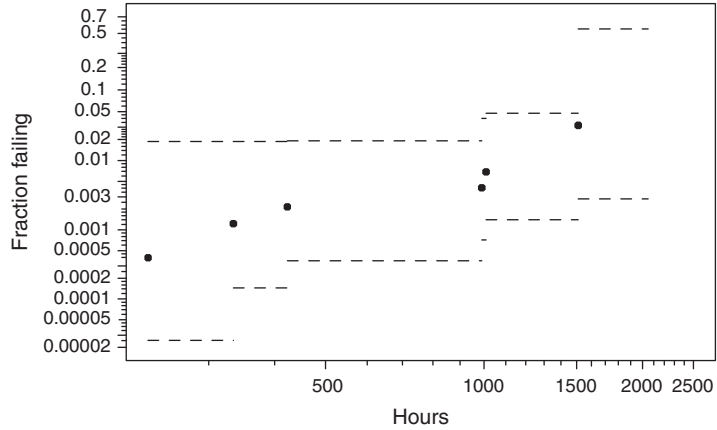


Figure 4. Weibull probability plot for the bearing-cage fracture data along with nonparametric approximate 95% simultaneous confidence bands.

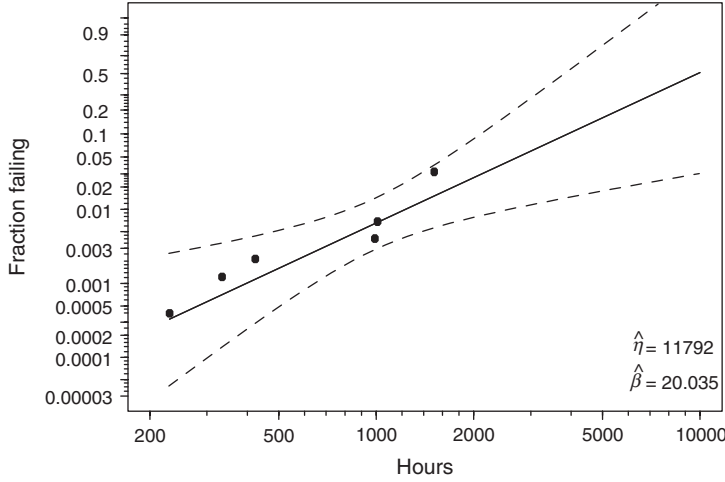


Figure 5. Weibull probability plot with the Weibull ML estimate and a set of pointwise approximate 95% confidence intervals for $F(t)$ for the bearing-cage field data.

Parametric ML estimation for data with arbitrary censoring and truncation is a natural extension of the likelihood for the right-censored data illustrated here. More details can be found in Meeker and Escobar [1].

Example 4 [Probability plot showing ML estimates for the bearing-cage field data]. Figure 5 is a Weibull probability plot for the bearing-cage data with the fitted ML line. The dotted lines depict a set of 95% parametric pointwise normal-approximation confidence intervals which are used to quantify statistical uncertainties due to limited data. See Meeker and Escobar [1, Chapter 8] for details about the computations.

REGRESSION ANALYSIS

Regression analysis of failure-time data from accelerated testing and field data with explanatory variable(s) can be useful to explain variability in failure-time data. In accelerated tests, test units of a material, component, subsystem, or entire systems are subjected to higher-than-usual levels of one or more accelerating variables such as temperature or stress. The accelerated test results are used to estimate the failure-time distribution of the units at use conditions. Regression models are needed for extrapolation to the use condition.

The single distribution models in Equation (1) can be extended to regression models. See Lawless [20] or Meeker and Escobar [1, Chapter 17] for details on parametric regression models and data analysis for failure-time data. In the most commonly used parametric failure-time regression model, the location parameter μ is modeled as a function of a vector of explanatory variables $\mathbf{x}$. For example, in a simple regression case, $\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1$ where β_0, β_1 are the regression coefficients and x_1 is the univariate explanatory variable. For categorical variables, dummy variable coding is needed.

In a general case, $\mu(\mathbf{x}) = g(\mathbf{x}, \boldsymbol{\beta})$, where $\mathbf{x} = (1, x_1, x_2, \dots, x_p)'$ is a vector of explanatory variable, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ is the regression coefficient vector, and $g(\cdot, \cdot)$ is a general relation function specified by the model for a model that is linear in the parameter, $g(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$. The ML method can be used to estimate $\boldsymbol{\beta}$ even in the presence of censoring and/or truncation.

Example 5 [Regression analysis of transformer lifetime data]. This is an example to illustrate the simple regression analysis of failure-time data with one explanatory variable. More complicated regression examples and regression analysis of data from accelerated tests can be found in Meeker and Escobar [1, Chapters 17 and 19].

We use a subset of the transformer failure-time data analyzed in Hong *et al.* [7]. The

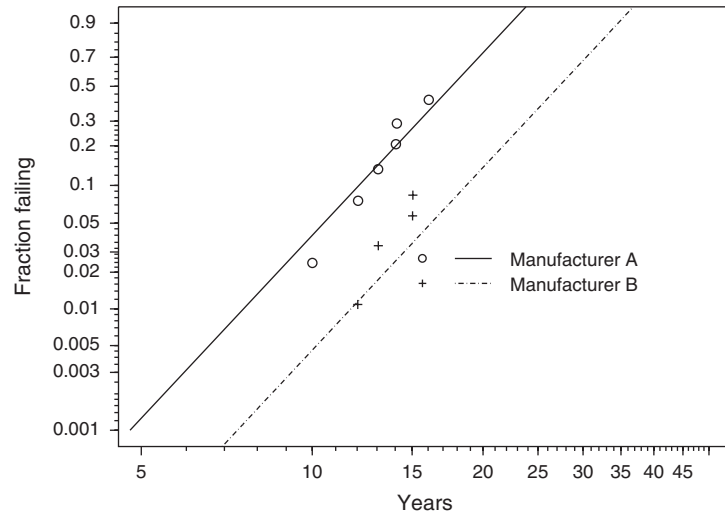


Figure 6. Weibull probability plot showing the parametric regression estimates of the transformer failure-time distribution.

data used here are the failure-time data from a fleet of high-voltage power transformers from an energy company installed after 1987. More details about the data can be found in Hong *et al.* [7]. The explanatory variable here is the manufacturer of transformers which has two levels. This categorical variable is coded as a dummy variable (e.g., $x_1 = 0$ for Manufacturer A and $x_1 = 1$ for Manufacturer B). Figure 6 is the Weibull probability plot showing the parametric regression estimates of the transformer lifetime data. This figure suggests that the transformers from Manufacturer B are more reliable.

MULTIPLE FAILURE MODES AND COMPETING RISKS

Multiple failure modes failure-time data arise when products can fail due to different causes. For some applications, the cause of failure can be identified and it is often important to distinguish among different failure modes when analyzing the failure-time data. Some possible reasons for this are as follows:

- When failure modes behave differently, it is generally easier to find well-fitting failure-time distributions for the individual failure modes [1, Example 15.6].
- Some failure modes that could cause serious harm are of critical importance while others are not.

- Some failure modes are much more expensive to fix than others when predicting warranty costs.
- Knowledge of the relative frequency of different failure modes and the effect of eliminating one or more of the individual failure modes is important for engineers from a product design point of view. In particular, such information is important for providing focus on methods for improving reliability in the next generation of the product.

The theory of competing risks provides the appropriate statistical model for failure-time data with multiple failure modes. David and Moeschberger [21] describe the theory of competing risks. For single distribution applications in reliability, these methods are described and illustrated with examples in Nelson [22, Chapter 5], Meeker and Escobar [1, Chapter 15], and Crowder [9].

When failure-mode information is available for all failed units and the different failure modes can be assumed to be statistically independent, data analysis is similar to data analysis for single failure mode. When failure-mode information is missing for some or all of the failed units (i.e., masked causes of failure), the analysis is more complicated and needs special methods. Flehinger *et al.* [23] describe statistical methods for dealing with masked failure-mode data. When the

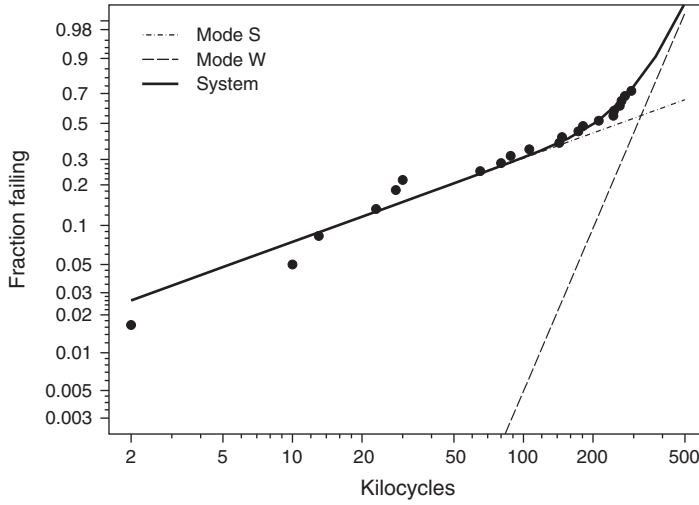


Figure 7. A Weibull probability plot showing ML estimates for the failure-time distribution of the individual failure modes and the system for Device G.

failure times for different failure modes are dependent, one approach is to use a multivariate failure-time model. In practice, however, there is often not enough information to identify such a model [10, Figure 7]. The subdistribution functions which give the fraction failing due to individual failure modes [9, p. 4], however, are always identifiable and sufficient for the data analysis in some applications.

Example 6 [Device G data]. The background for this data set is given in Meeker and Escobar [1, Example 15.6]. Device G is a component of a large system that had two failure modes: failure mode S and W, corresponding to electrical surges and wear out. The field tracking study data are available in Meeker and Escobar [1, Table 15.1]. The Weibull distribution was used to fit the failure time of individual failure modes assuming the independence of the two failure modes. Figure 7 is a Weibull probability plot showing the ML estimates for the individual failure modes. The curved line is the estimated series system failure probability for the two failure modes combined under the independence assumption, and it agrees well with the nonparametric estimate computed by ignoring the failure-mode information. Figure 8 shows the estimated hf for the failure time of Device G. The hf $h(t)$ shows an interesting “U” shape. This is because $h(t) = h_S(t) + h_W(t)$

where $h_S(t)$ and $h_W(t)$ are the hfs for Mode S and Mode W, respectively. Mode S has a decreasing hazard and dominates system failure early in life. Mode W has an increasing hazard and dominates system failure late in life.

NEW TRENDS IN FIELD RELIABILITY DATA

The next generation of reliability field failure-time data will be richer in information due to changes in data collecting technology (e.g., sensors and smart chips built into systems to measure environmental variables like stress and temperature). In some applications, this information is available dynamically. For example, medical systems (such as CAT scanners) and high-end printers and copiers can be connected to the network, providing such information on demand.

Use rate and environmental conditions are important sources of variability in product failure times. The most important differences between controlled laboratory accelerated test experiments and field reliability results are the uncontrolled operating and environmental field variation. This variation is from unit-to-unit and temporal, caused by variables such as use rate, vibration, temperature, and humidity. In the past, use-rate/environmental data has not been generally available in reliability

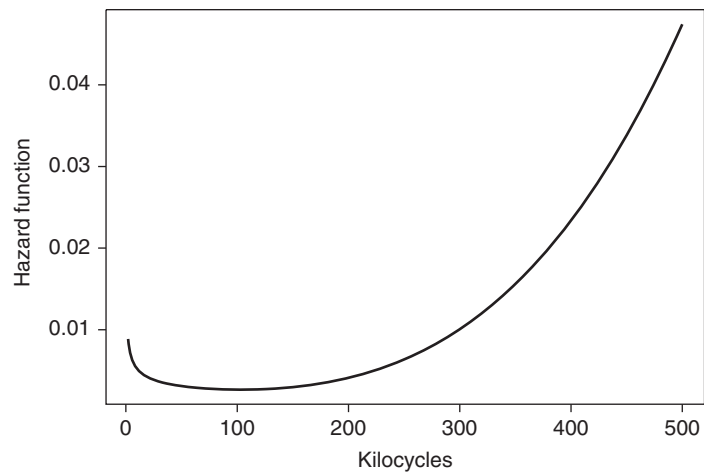


Figure 8. The estimated hazard function for Device G.

data analysis. In the future, failure-time models that use use-rate/environmental data will have the potential to explain much more variability in field data than has been possible before. The information can also be used to predict future environmental conditions and thus provide better information for predicting failure times.

FURTHER READING

This section provides a short list of useful references in reliability. Meeker and Escobar [1] describe commonly used statistical methods for reliability data analysis, providing a large number of examples and sufficient technical detail to understand how and why the methods work. Lawless [20] is an important reference for some of the technical aspects of statistical methods for failure-time data, illustrating methods with both engineering and medical applications. Tableman and Kim [24] describe methods for analyzing time-to-event data and illustrate the corresponding computations using S-PLUS. Work by Nelson [5] has been the definitive work on statistical models and methods for accelerated testing since it was published, while the work by Hahn and Shapiro [18] has been an important resource for statistical methods in engineering applications for many years. Yang [25] provides a broad overview of quantitative methods that should be used in processes

for the design of reliable systems from an engineering point of view. Crowder [9] provides a thorough treatment on the subject of competing risks models and subdistribution functions. These models are useful for modeling and prediction in reliability applications with multiple failure modes.

Acknowledgments

We would like to thank the area editor and the reviewer, who provided comments that helped us improve this article. We would also like to thank Katherine Meeker for helpful comments that improved this article.

REFERENCES

1. Meeker WQ, Escobar LA. Statistical methods for reliability data. New York: John Wiley & Sons, Inc.; 1998.
2. Condra LW. Reliability improvement with design of experiments. New York: Marcel Deller; 1993.
3. Hong Y, Meeker WQ, Escobar LA. The condition-based paradigm. Wiley Encyclopedia of Operations Research and Management Science. Hoboken (NJ): John Wiley & Sons, Inc.; 2010. In press.
4. Meeker WQ. Limited failure population life tests: application to integrated circuit reliability. *Technometrics* 1987;29:51–65.
5. Nelson W. Accelerated testing: statistical models, test plans, and data analyses. New

- York: John Wiley & Sons, Inc.; 1990. (Republished in a paperback in Wiley Series in Probability and Statistics, 2004.)
6. Escobar LA, Meeker WQ. A review of accelerated test models. *Stat Sci* 2006;21:552–577.
 7. Kalbfleisch JD, Lawless JF, Robinson JA. Methods for the analysis and prediction of warranty claims. *Technometrics* 1991;33:273–285.
 8. Hong Y, Meeker WQ, McCalley JD. Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *Ann Appl Stat* 2009;3:857–879.
 9. Crowder MJ. Classical competing risks. Boca Raton (FL): Chapman & Hall/CRC; 2001.
 10. Meeker WQ, Escobar LA, Hong Y. Using accelerated life tests results to predict product field reliability. *Technometrics* 2009;51:146–161.
 11. Kalbfleisch JD, Lawless JF. Regression models for right truncated data with applications to aids incubation times and reporting lags. *Stat Sinica* 1991;1:19–32.
 12. Abernethy RB, Breneman JE, Medlin CH, *et al.* Weibull analysis handbook. Air Force Wright Aeronautical Laboratories, Available from the National Technical Information Service, Washington D.C., Technical Report AFWAL-TR-83-2079; 1983.
 13. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–481.
 14. Peto R. Experimental survival curves for interval-censored data. *Appl Stat* 1973;22: 86–91.
 15. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Ser B* 1976;38:290–295.
 16. Nelson W. Hazard plotting for incomplete failure data. *J Q Technol* 1969;1:27–52.
 17. Aalen O. Nonparametric inference in connection with multiple decrement models. *Scand J Stat* 1976;3:15–27.
 18. Hahn GJ, Shapiro SS. Statistical models in engineering. New York: John Wiley & Sons, Inc.; 1967. (Republished in a paperback in Wiley Classics Library, 1994.)
 19. Nair VN. Confidence bands for survival functions with censored data: a comparative study. *Technometrics* 1984;26:265–275.
 20. Lawless JF. Statistical models and methods for lifetime data. 2nd ed. Hoboken (NJ): John Wiley & Sons, Inc.; 2003.
 21. David HA, Moeschberger ML. The theory of competing risks. London: Griffin; 1978.
 22. Nelson W. Applied life data analysis. New York: John Wiley & Sons, Inc.; 1982.
 23. Flehinger B, Reiser B, Yashchin E. Survival with competing risks and masked causes of failure. *Biometrika* 1998;85:151–164.
 24. Tableman M, Kim JS. Survival analysis using S: analysis of time-to-event data. New York: Chapman & Hall/CRC; 2004.
 25. Yang G. Life cycle reliability engineering. Hoboken (NJ): John Wiley & Sons, Inc.; 2007.

THE G/G/1 QUEUE

SERGEY FOSS
School of Mathematical and
Computer Sciences and
Maxwell Institute,
Heriot-Watt University,
Edinburgh, UK

The notation $G/G/1$ queue is usually referred to a single-server queue with first-in-first-out discipline and with a general distribution of the sequences of inter-arrival and service times (which are the “driving sequences” of the system). Customers are numbered $n = 0, 1, \dots$. We assume that customer 0 arrives at a system at time $t = 0$ and finds there an initial amount of work, so has to wait for $W_0 \geq 0$ units of time for the start of its service. Let t_n be the time between the arrivals of n th and $(n + 1)$ st customers, and s_n the service time of the n th customer. Let W_n be the waiting time (delay) of the n th customer, that is, the time between its arrival and beginning of its service. Then, for $n \geq 0$, the sequence $\{W_n\}$ satisfies *Lindley's equations*

$$W_{n+1} = (W_n + s_n - t_n)^+, \quad (1)$$

where we use the notation $x^+ = \max(0, x)$.

We assume the two-dimensional sequence $\{(s_n, t_n)\}$ to be stationary. We consider mostly the i.i.d. sequences ($GI/GI/1$ queue; here “ GI ” stands for “general independent”), and then the general case in brief. The list of references (Refs 1–13) contains a number of recent textbooks and surveys with the whole coverage of problems discussed in the article. The key references are Ref. 1 for the section titled “ $G/G/1$ queue,” and Refs 2 and 3 for the section titled “General $G/G/1$ queue.”

GI/GI/1 QUEUE

In this section we assume that each of the sequences $\{t_n\}$ and $\{s_n\}$ consists of i.i.d. (independent and identically distributed)

random variables (r.v.'s) and that these two sequences are independent and do not depend on W_0 . Without loss of generality, we may assume that these sequences are extended onto negative indices $n = -1, -2, \dots$. Let $T_0 = 0, T_n = \sum_{i=0}^{n-1} t_i, n \geq 1$, and $T_n = -\sum_{i=n}^{-1} t_i, n \leq -1$.

Let $a = \frac{1}{\lambda} = \mathbf{E}t_n \in (0, \infty)$ denote the Inter-arrival mean (expectation) and $b = \mathbf{E}s_n \in (0, \infty)$ the mean service time. Then $\rho = \lambda b$ is the *traffic intensity*. We consider the case of finite means only.

For the analysis of the waiting-time distributions, the key tool is the *duality* between the maximum of a random walk and the waiting time processes. Define $\xi_n = s_n - t_n$, $\mu = \mathbf{E}\xi_n = b - a, S_0 = 0, S_n = \xi_0 + \dots + \xi_{n-1}, M_n = \max_{0 \leq k \leq n} S_k$. Note that $\rho < 1$ and $\mu < 0$ are equivalent.

For the random walk $\{S_n\}$, the strong law of large numbers holds, $S_n/n \rightarrow \mu$ a.s. as $n \rightarrow \infty$. So, if $\mu < 0$, then the random variable $v = \max\{n : S_n > 0\}$ is finite a.s. and $M_n = M_v$ for all $n \geq v$. Then $M = \sup_{n \geq 0} M_n = M_v$ and, for any $n, \mathbf{P}(M_n \neq M) \leq \mathbf{P}(v > n) \rightarrow 0$, as $n \rightarrow \infty$.

Proposition 1. (1) For any $n \geq 0$, r.v.'s W_n and $\max(M_n, W_0 + S_n)$ are identically distributed,

$$W_n =_D \max(M_n, W_0 + S_n). \quad (2)$$

In particular, if $W_0 = 0$, then $W_n =_D M_n$ and probability $\mathbf{P}(W_n > x)$ increases in n , for any x .

(2) If $\rho < 1$, then a limiting steady-state waiting time W exists and coincides in distribution with that of $M, W =_D M$. Further, the distribution of W_n converges to that of W in the total variation norm, that is

$$\sup_A |\mathbf{P}(W_n \in A) - \mathbf{P}(W \in A)| \leq \mathbf{P}(v > n) \rightarrow 0, \quad n \rightarrow \infty. \quad (3)$$

For the steady-state random variable W and for an independent of it r.v. $\xi =_D \xi_1$, the random variables $(W + \xi)^+$ and W have the same distribution,

$$(W + \xi)^+ =_D W. \quad (4)$$

(3) If $\rho = 1$ and if $\sigma^2 = \mathbf{Var} \xi_n = \mathbf{Var} t_1 + \mathbf{Var} s_1$ is finite, then the distributions of $W_n/\sqrt{n}$ converge weakly, as $n \rightarrow \infty$, to the distribution of the absolute value of a normal random variable with mean 0 and variance σ^2 .

(4) If $\rho > 1$, then $W_n/n \rightarrow \mu$ a.s., as $n \rightarrow \infty$.

Regenerative Structure

Consider the case $W_0 = 0$. Let $\tau_0 = 0, \tau = \tau_1 = \inf\{n \geq 1 : W_n = 0\}$ and, for $k \geq 1$, $\tau_{k+1} = \inf\{n > \tau_k : W_n = 0\}$. Clearly, $\tau = \inf\{n \geq 1 : S_n \leq 0\}$.

Proposition 2. Assume $W_0 = 0$. If $\rho \leq 1$, then all τ_k are finite a.s. and the random elements $(\tau_k - \tau_{k-1}, (t_i, s_i, W_i), i = \tau_{k-1}, \dots, \tau_k - 1)$ are i.i.d. Moreover, if $\rho < 1$, then $\mathbf{E}\tau < \infty$ and, for any $\alpha > 1$, $\mathbf{E}\tau^\alpha < \infty$ if and only if $\mathbf{E}s_1^\alpha < \infty$. Further, if $\rho < 1$ and if $\mathbf{E}e^{c s_1}$ is finite for some $c > 0$, then $\mathbf{E}e^{c_1 \tau}$ is finite for some $c_1 > 0$.

We may interpret $\tau_k - \tau_{k-1}$ as the number of customers served in the k th busy period, the duration of the k th period is the total service time

$$B_k = \sum_{i=\tau_{k-1}}^{\tau_k-1} s_i,$$

and this busy period is followed by the *idle* period where the server is empty during

$$I_k = \sum_{i=\tau_{k-1}}^{\tau_k-1} t_i - \sum_{i=\tau_{k-1}}^{\tau_k-1} s_i = \sum_{i=\tau_{k-1}}^{\tau_k-1} (t_i - s_i)$$

units of time. The sum of a busy period and of the following idle period is a *busy cycle*, $C_k = B_k + I_k = \sum_{i=\tau_{k-1}}^{\tau_k-1} t_i$. In particular, the first idle period, I , is equal to $I = I_1 = -S_\tau$.

Proposition 3. Assume $W_0 = 0$. If $\rho < 1$, then $\mathbf{E}I$ is also finite and, for any function $f : [0, \infty) \rightarrow [0, \infty)$,

$$\mathbf{E}f((W + \xi)^-) = \frac{\mathbf{E}f(I)}{\mathbf{E}\tau} = -\mathbf{E}\xi \frac{\mathbf{E}f(I)}{\mathbf{E}I} \quad (5)$$

and, in particular,

$$\mathbf{E}(W + \xi)^- = -\mathbf{E}\xi. \quad (6)$$

Here we use notation $x^- = -\min(x, 0)$.

Moments of the Stationary Waiting Time

Proposition 4. Assume $\rho < 1$. Let W be the stationary waiting time. For any $\alpha > 0$, if $\mathbf{E}s_1^{\alpha+1} < \infty$, then $\mathbf{E}W^\alpha < \infty$. Conversely, if $\mathbf{E}W^\alpha < \infty$ and $\mathbf{E}t_1 < \infty$, then $\mathbf{E}s_1^{\alpha+1} < \infty$. Further, if, for some $k = 1, 2, \dots$, both $\mathbf{E}s_1^{k+1}$ and $\mathbf{E}t_1^{k+1}$ are finite, then

$$\begin{aligned} \sum_{l=0}^k C_{k+1}^l \mathbf{E}W^l \mathbf{E}\xi^{k+1-l} &= \mathbf{E}[-(W + \xi)^{k+1}] \\ &= (-1)^k \mathbf{E}\xi \frac{\mathbf{E}I^{k+1}}{\mathbf{E}I}. \end{aligned} \quad (7)$$

In particular,

$$\mathbf{E}W = \frac{\lambda^2(\mathbf{Var} t_1 + \mathbf{Var} s_1) + (1 - \rho)^2}{2\lambda(1 - \rho)} - \frac{\mathbf{E}I^2}{2\mathbf{E}I}, \quad (8)$$

and then

$$\begin{aligned} \frac{\rho^2 + \lambda^2 \mathbf{Var} s_1 - 2\rho}{2\lambda(1 - \rho)} &\leq \mathbf{E}W \\ &\leq \frac{\lambda(\mathbf{Var} t_1 + \mathbf{Var} s_1)}{2(1 - \rho)}. \end{aligned} \quad (9)$$

Here again, random variables W and $\xi = {}_D\xi_1$ are assumed to be independent.

Continuous Time. The Virtual Waiting Time

Recall that $C = C_1, B = B_1$, and $I = I_1$ are, correspondingly, the (duration of) the first busy cycle, the first busy period, and the first idle period.

Proposition 5. Assume $W_0 = 0$. Suppose $\rho < 1$. Then the busy cycle has mean $\mathbf{E}C = a\mathbf{E}\tau$, the busy period has mean $\mathbf{E}B = b\mathbf{E}\tau$, and the mean of the idle period is $\mathbf{E}I = \mathbf{E}C - \mathbf{E}B = -\mu\mathbf{E}\tau$.

Let $W(t)$ be the *virtual waiting time* at time t , that is, the total residual workload at

t , which is the sum of the residual service time of a customer in service and of all service times of customers in the queue. As a function of t , $W(t)$ decreases linearly between consecutive arrivals if positive and stays at zero if zero, with positive jumps at the arrival epochs. Further, $W_n = W(T_n -)$, $n = 0, 1, \dots$ and the process $W(t)$, $t \geq 0$ satisfied Lindley's equations in continuous time:

$$W(t) = (W(T_n -) + s_n - (t - T_n))^+, \quad t \in [T_n, T_{n+1}). \quad (10)$$

Non-Lattice Distribution. A random variable X has a non-lattice distribution if $\sum_{k=-\infty}^{\infty} \mathbf{P}(X = kh) < 1$, for all $h > 0$.

Proposition 6. Suppose that $\rho < 1$ and that $\{t_n\}$ have a common non-lattice distribution. Then there exists a limiting (stationary) distribution, as $t \rightarrow \infty$, of the virtual time $W(t)$ which is given by

$$\mathbf{P}(V > x) = \frac{1}{\mathbf{E}C} \mathbf{E} \left(\int_0^C \mathbf{I}(W(u) > x) du \right). \quad (11)$$

Further,

$$\mathbf{E}V = \rho \left(\frac{\mathbf{E}s^2}{2\mu_s} + \mathbf{E}W \right). \quad (12)$$

Here $\mathbf{I}(\cdot)$ is the indicator function, and $\mathbf{I}(A) = 1$ if event A occurs and $\mathbf{I}(A) = 0$ otherwise.

Proposition 7. Assume that a random variable $\hat{s}$ is independent of W and has an absolutely continuous distribution with density $\mathbf{P}(s > x)/b$ where $b = \mathbf{E}s_1$. Then

$$\mathbf{P}(V = 0) = 1 - \rho$$

and, for all $x \geq 0$,

$$\mathbf{P}(V > x) = \rho \mathbf{P}(W + \hat{s} > x).$$

Queue Length and Little's Law

We introduce three quantities: the queue length Q_n^A at the n arrival time, the queue length Q_n^D at the n departure time, and the queue length $Q(t)$ at an arbitrary time t .

Spread-Out Distribution. A probability distribution F is spread out if it can be represented as $F = pG_1 + (1-p)G_2$ where $p \in [0, 1)$, G_1 and G_2 are probability distributions, and G_2 is absolutely continuous with respect to Lebesgue measure.

Proposition 8. If $\rho < 1$, then the distribution of Q_n^A (correspondingly, of Q_n^D) converges, as $n \rightarrow \infty$, to that of a finite random variable Q^A (correspondingly, Q^D), in the total variation norm.

If $\rho < 1$ and the distribution of the inter-arrival times is non-lattice, then the distribution of $Q(t)$ converges weakly, as $t \rightarrow \infty$, to that of a finite random variable Q . If, in addition, the common distribution of inter-arrival times is spread out, then the convergence is in the total variation norm.

Proposition 9 [Little's Law]. Assume that $\rho < 1$, $\mathbf{E}s_1^2 < \infty$, and that the distribution of inter-arrival times is non-lattice. Then

$$q = \lambda(w + b), \quad (13)$$

where $q = \mathbf{E}Q$ is the mean stationary queue length, $\lambda = 1/a$ the arrival rate, and $w + b = \mathbf{E}(W + s_1)$ the mean stationary sojourn time, that is, the sum of mean waiting and mean service times.

Proposition 10 [Distributional Little's Law in Discrete Time]. Suppose $\rho < 1$. Then, for any $k = 1, 2, \dots$,

$$\begin{aligned} \mathbf{P}(Q^A \geq k) &= \mathbf{P}(Q^D \geq k) \\ &= \mathbf{P}\left(W + s_1 \geq \sum_{i=1}^k t_i\right), \end{aligned} \quad (14)$$

where the random variables W, s_1 , and $\{t_i\}_{i=1}^k$ are assumed to be mutually independent. Further, if either s_1 or t_1 has a continuous distribution, then Equations (14) are equivalent to

$$\mathbf{P}(Q^A = 0) = \mathbf{P}(Q^D = 0) = \mathbf{P}(W = 0). \quad (15)$$

and, for $k \geq 1$,

$$\mathbf{P}(Q^A \geq k) = \mathbf{P}(Q^D \geq k) = \mathbf{P}\left(W > \sum_{i=1}^{k-1} t_i\right). \quad (16)$$

Proposition 11 [Distributional Little's Law in Continuous Time]. Suppose that $\rho < 1$ and that the distribution of inter-arrival times is non-lattice. Then the stationary distribution of the queue length Q is given by

$$\mathbf{P}(Q = 0) = 1 - \rho$$

and, for $k = 1, 2, \dots$,

$$\begin{aligned} \mathbf{P}(Q \geq k) &= \mathbf{P}\left(V > \sum_{i=1}^{k-1} t_i\right) \\ &= \rho \mathbf{P}\left(W + \widehat{s} > \sum_{i=1}^{k-1} t_i\right), \end{aligned} \quad (17)$$

where we assume the mutual independence of all r.v.'s in the formula above.

Continuity Under Perturbation

Proposition 12. Consider a series of single-server queue indexed by the upper k . Assume that, as $k \rightarrow \infty$, the distributions of $t_1^{(k)}$ weakly converge to that of t_1 , the distributions of $s_1^{(k)}$ weakly converge to that of s_1 , and that $\rho < 1$ and $\rho^{(k)} < 1$, for all k . If $\mathbf{E}s_1^{(k)} \rightarrow \mathbf{E}s_1$, then the distributions of $W^{(k)}$ converge weakly to that of W . If, in addition, either the distribution of s_1 or the distribution of t_1 is continuous, then $\mathbf{P}(W^{(k)} = 0) \rightarrow \mathbf{P}(W = 0)$.

Heavy-Traffic Limits

Proposition 13. Consider a series of single-server queues indexed by the upper index k . Assume that, as $k \rightarrow \infty$, the distributions of $t_1^{(k)}$ weakly converge to that of t_1 , the distributions of $s_1^{(k)}$ weakly converge to that of s_1 , and that the distributions of s_1 and of t_1 do not both degenerate. Assume further that $\rho^{(k)} < 1$, for all k , and $\rho^{(k)} \rightarrow \rho = 1$, and that the squares $(s_1^{(k)})^2$ and $(t_1^{(k)})^2$ are uniformly integrable.

Then the distributions of random variables $Y^{(k)} = |\mu^{(k)}| W^{(k)} (\sigma^{(k)})^2$ converge weakly to an

exponential distribution with intensity 2 and also $\mathbf{E}Y^{(k)} \rightarrow \frac{1}{2}$. Furthermore, for each T and for $y \geq 0$,

$$\begin{aligned} &\mathbf{P}\left(\frac{|\mu^{(k)}|}{(\sigma^{(k)})^2} W^{(k)} \left[T(\sigma^{(k)})^2\right] > y\right) \\ &\rightarrow \mathbf{P}\left(\max_{0 \leq t \leq T} (B(t) - t) \geq y\right), \end{aligned} \quad (18)$$

where $\{B(t), t \geq 0\}$ is the standard Brownian motion, and $[x]$ is the integer part of a number x .

Rare Events

In the stable case, $\rho < 1$, the exact values of the tail probability $\mathbf{P}(W > x)$ may be found only in a number of particular cases. But the asymptotics for this probability may be written down in a great generality. There are two particular cases, of light tails and heavy tails, of the service time distributions, where the tail asymptotics differ and are caused by different phenomena.

For two positive functions f and g on the real line, we write $f(x) \sim g(x)$ if $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 1$ as $x \rightarrow \infty$.

Rare Events in the Presence of Light Tails.

Proposition 14. Assume that $\rho < 1$ and the distribution of the service times has the light tail, $\mathbf{E}e^{ds_1} < \infty$ for some $d > 0$ and, moreover, that there exists a constant $\gamma > 0$ such that $\mathbf{E}e^{\gamma(s_1 - t_1)} = 1$ and $d := \mathbf{E}(s_1 - t_1)e^{\gamma(s_1 - t_1)} \in (0, \infty)$. Assume further that the distribution of r.v. $s_1 - t_1$ is non-lattice. Then, as $x \rightarrow \infty$,

$$\mathbf{P}(W > x) \sim re^{-\gamma x}, \quad (19)$$

where $r = \mathbf{E}e^{-\gamma \chi} \in (0, 1)$ and χ is the overshoot over the infinite barrier (see Refs 1 or 2 for the definition and more detail). Further, assume that $W = W_0$ is the stationary waiting time of customer 0 in the system that runs, say, from time $-\infty$. As $x \rightarrow \infty$,

$$\mathbf{P}(A(x) | W_0 > x) \rightarrow 1, \quad (20)$$

where $A(x)$ is the event of the form:

$$A(x) = \left\{ \begin{array}{l} \sum_{i=-\lfloor x/d \rfloor}^{-\lfloor x/d \rfloor + j - 1} (s_i - t_i) \\ \in (-R(x) + j(d - \varepsilon(x)), \\ R(x) + j(d + \varepsilon(x)), \\ \text{for } j = 1, 2, \dots, \lfloor x/d \rfloor \end{array} \right\},$$

and $R(x)$ is any function tending to infinity and $\varepsilon(x)$ is any function tending to zero, as $x \rightarrow \infty$. Roughly speaking, the latter means that, given the event $\{W_0 > x\}$ occurs, all increments $(s_i - t_i)$ are approximately equal to d , for $i = -\lfloor x/d \rfloor, -\lfloor x/d \rfloor + 1, \dots, -1$.

In particular, the distribution of $s_1 - t_1$ is non-lattice if either of the distributions of s_1 or t_1 is non-lattice.

Rare Events in the Presence of Heavy Tails.

Here we assume that the distribution of service times is *heavy-tailed*, that is, does not have finite exponential moments, $\mathbf{E}e^{\alpha s_1} = \infty$, for all $\alpha > 0$.

A distribution F on the positive half-line with an unbounded support is *subexponential* if

$$\overline{F * F}(x) \sim 2\overline{F}(x), \quad \text{as } x \rightarrow \infty.$$

It is known that any subexponential distribution is *long-tailed*, that is, $\overline{F}(x + c) \sim \overline{F}(x)$ as $x \rightarrow \infty$, for any constant c . Further, any long-tailed distribution is heavy tailed.

Typical examples of subexponential distributions are log-normal distributions, distributions with power tails $(1 + x)^{-\alpha}$, $\alpha > 0$, and Weibull tails e^{-x^β} , $\beta \in (0, 1)$.

Let again r.v. $\widehat{s}$ have the absolutely continuous distribution with distribution function $F_{\widehat{s}}(x)$ and density $f_{\widehat{s}}(x) = \mathbf{P}(s > x)/b$. Clearly, the distribution of service time s is heavy-tailed if and only if the distribution of $\widehat{s}$ is.

Proposition 15. *Assume that $\rho < 1$ and that the distribution of $\widehat{s}$ is subexponential.*

Then, as $x \rightarrow \infty$,

$$\begin{aligned} \mathbf{P}(W > x) &\sim \frac{\rho}{1 - \rho} (1 - F_{\widehat{s}}(x)) \\ &= \frac{1}{a - b} \int_x^\infty \mathbf{P}(s_1 > y) dy. \end{aligned} \quad (21)$$

Further, assume that $W = W_0$ is the stationary waiting time of customer 0 in the system that runs, say, from time $-\infty$. As $x \rightarrow \infty$,

$$\mathbf{P}(B(x) \mid W_0 > x) \rightarrow 1, \quad (22)$$

where $B(x)$ is the event of the form

$$B(x) = \bigcup_{i \geq 1} \{(s_{-i} > x + i(a - b))\}.$$

This means that, for large x , the main cause for the event $\{W_0 > x\}$ to occur is a single big jump of one of previous service times.

GENERAL G/G/1 QUEUE

In this subsection we consider the single-server queue under the stationary ergodic assumptions on the driving sequences. Namely, the sequence $\{(t_n, s_n)\}$ is *stationary* if the joint distribution of random vectors $\{(t_{m+i}, s_{m+i})\}_{0 \leq i \leq k}$ does not depend on m , for any k . In addition, this sequence is *ergodic* if, for any event A generated by the driving sequence, the equality $\mathbf{P}(A \cap A^\theta) = \mathbf{P}(A)$ implies that $\mathbf{P}(A)$ is either 0 or 1. Here A^θ is defined as follows: Any event A generated by the driving sequence may be represented as $A = \{g(t_i, s_i; i = \dots, -1, 0, 1, \dots) = 1\}$ where g is a measurable function of the driving sequence. Then $A^\theta = \{g(t_{i+1}, s_{i+1}; i = \dots, -1, 0, 1, \dots) = 1\}$. In particular, the sequence is ergodic if it is i.i.d. or, more generally, the tail sigma-algebra generated by this sequence is *trivial*, that is, contains only events of probability 0 or 1.

Proposition 16. *Under the stability condition $\rho = \mathbf{E}s_1/\mathbf{E}t_1 < 1$, there exists a unique stationary workload process $\widetilde{W}(t)$, $-\infty < t < \infty$ which satisfies Equations (10) on the whole real line and is such that*

$$\tilde{W}(0) = \sup_{n \leq 0} \left(T_n + \sum_{i=0}^n s_i \right)^+. \quad (23)$$

Further, there is an infinite number of negative indices n and infinite number of positive indices n such that $\tilde{W}(T_n-) = 0$.

If $\rho > 1$, then there is no finite stationary workload process and, for any initial condition $W_0 \geq 0$, $W(t)/t \rightarrow \rho - 1$ a.s., as $t \rightarrow \infty$.

If $\rho = 1$, then there may or may not exist a finite stationary process.

REFERENCES

1. Asmussen S. Applied probability and queues. 2nd ed. New York: Springer; 2003.
2. Baccelli F, Bremaud P. Elements of Queueing Theory. 2nd ed. New York: Springer; 2003.
3. Brandt A, Franken P, Lisek B. Stationary stochastic models. New York: Wiley; 1992.
4. Borovkov AA. Stochastic processes in Queueing Theory. New York: Springer; 1976.
5. Bremaud P. Markov Chains, Gibbs Fields, Monte-Carlo Simulation and Queues. New York: Springer; 1999.
6. Cohen JW. The single server queue. Amsterdam: North-Holland Publishing Co.; 1982.
7. Cox DR, Smith WL. Queues. London: Methuen; 1961.
8. Gnedenko BV, Kovalenko IN. An introduction to Queueing Theory. 2nd ed. Boston: Burkhauser; 1989.
9. Kalashnikov VV. Mathematical methods in Queueing Theory. Dordrecht: Kluwer; 1994.
10. Kleinrock L. Volume I, II, Queueing systems. New York: Wiley; 1975, 1976.
11. Miyazawa M. Rate conservation laws: a survey. Queueing Syst Theor Appl 1994;15:1–58.
12. Robert P. Stochastic networks and queues. New York: Springer; 2003.
13. Whitt W. Stochastic-process limits. New York: Springer; 2002.

THE G/G/s QUEUE

LOTHAR BREUER
Institute of Mathematics and
Statistics, University of Kent,
Canterbury, UK

BASIC DEFINITIONS

The G/G/s queue features s servers, iid interarrival times, and iid service completion times. Arrival and service completion times are independent. Often this queue is denoted by GI/G/s in order to emphasize that the arrivals form a renewal process. We shall assume that $s > 1$ as the single-server case has been treated in the article titled *The G/G/1 Queue*. Unless stated otherwise, we further assume $s < \infty$. The most common service discipline considered is FCFS (first come first served), which we shall presume if nothing else is specified. We further assume, unless stated otherwise, that the servers are homogeneous, that is, all servers have the same distribution of service completion times. An arriving user will commence to be served immediately if at the time of arrival one of the servers is idle. Otherwise the user joins the queue at the back and will be served as soon as all users before they have commenced to be served and a server has become idle. The waiting room is infinite, unless specified otherwise.

The general G/G/s queue is of course the most difficult to analyze and a complete computationally tractable solution has not been derived yet. Functionals of interest are the number of users in the system (also called the *queue length*), the waiting times, the work load, or the length of a busy period. For the queue length, one may seek the transient distribution at a finite fixed time or the stationary distribution, which under some stability condition is attained asymptotically after the queueing system has run for a long time.

There is no mature general theory for the G/G/s queue. Results mostly concern special

aspects and employ methods quite diverse. Thus it is not possible to present a concise survey of the general multiserver queue. Indeed, most of the practically relevant questions still await a satisfying answer. We therefore present two classical results in the sections titled “The Markov Chain of Waiting Time Vectors” and “The G/M/s Queue,” while the last section contains a literature survey. This shall be roughly structured according to the functionals listed above. Although many results are included, this survey is by no means exhaustive.

THE MARKOV CHAIN OF WAITING TIME VECTORS

The construction of waiting time vectors gives rise to a (discrete-time) Markov chain analysis of the G/G/s queue, which has been initiated by the seminal paper [1]. The main results of this is shortly summarized in this section.

Let $t_0 := 0$ and denote the arrival time of the i th user by t_i , $i \in \mathbb{N}$, with $t_{i+1} \geq t_i$ for all $i \in \mathbb{N}$. The interarrival times shall be denoted by $g_i := t_i - t_{i-1}$, $i \in \mathbb{N}$. Let A denote the distribution function of the (iid) interarrival times. The service time of the i th user is denoted by R_i , $i \in \mathbb{N}$, and the distribution function of the (iid) service times R_i by H . Define $\rho := \mathbb{E}(H)/(s \cdot \mathbb{E}(A))$, where $\mathbb{E}(H)$ and $\mathbb{E}(A)$ denote the mean service and interarrival time, respectively.

We denote the waiting time of the i th user by w_i , $i \in \mathbb{N}$. This is the time duration between the arrival t_i of the i th user and the commencement of her service. Let $t_i + u_{ij}$ denote the time at which the j th server finishes serving the last of those among the $i - 1$ first users which it serves. For any real number $a \in \mathbb{R}$, define $a^+ := \max(0, a)$. Denote the order statistic by O , that is, $O(x_1, \dots, x_s) = (x_{(1)}, \dots, x_{(s)})$ is the permutation of $(x_1, \dots, x_s)$ such that $x_{(1)} \leq \dots \leq x_{(s)}$. Then the i th waiting time vector is defined as

$$w_i := (w_{i1}, \dots, w_{is}) := O(u_{i1}^+, \dots, u_{is}^+),$$

for all $i \in \mathbb{N}$. The iid nature of interarrival and service times entails that the sequence $(w_i : i \in \mathbb{N})$ is a Markov chain with $w_{i+1} = \phi(w_i, R_i, g_{i+1})$ for all $i \in \mathbb{N}$, where the function ϕ is defined by

$$\phi(w_i, R_i, g_{i+1}) := O((w_{i1} + R_i - g_{i+1})^+, (w_{i2} - g_{i+1})^+, \dots, (w_{is} - g_{i+1})^+). \quad (1)$$

Let $S := \{x = (x_1, \dots, x_s) \in \mathbb{R}^s : 0 \leq x_1 \leq x_2 \leq \dots \leq x_s\}$ denote the set of all ordered vectors with s nonnegative entries. Define the distribution functions

$$F_i(x|y) := \mathbb{P}(w_{i1} \leq x_1, \dots, w_{is} \leq x_s | w_1 = y),$$

for all $i \in \mathbb{N}$ and $x, y \in S$. Further define $F_i(x) := F_i(x|0)$, where 0 denotes the zero vector. The Markov property of $(w_i : i \in \mathbb{N})$ yields $F_{i+1}(x) = \int_{y \in S} F_i(x|y) dF_2(y)$ for all $i \in \mathbb{N}$ and $x \in S$. For $y_1, y_2 \in S$ with $y_1 \leq y_2$ (entry-wise) a sample path comparison yields $F_i(x|y_1) \geq F_i(x|y_2)$ for all $x \in S$ and $i \in \mathbb{N}$. Hence we obtain

$$F_{i+1}(x) - F_i(x) = \int_{y \in S} F_i(x|y) - F_i(x|0) dF_2(y) \leq 0,$$

which proves the inequality $F_{i+1}(x) \leq F_i(x)$ for all $i \in \mathbb{N}$ and $x \in S$. Thus the sequence $(F_i(x) : i \in \mathbb{N})$ approaches a limit $F(x)$ as $i \rightarrow \infty$. Let F denote the function on S with values $F(x)$. In Section 6 of Ref. 1, it is shown via discretisation of A and H and stochastic comparison arguments that F is a distribution function if $\rho < 1$. Under the opposite assumption $\rho \geq 1$ and $P(R_i = sg_i) < 1$, Section 7 therein shows that $\lim_{x' \rightarrow \infty} F(x', \dots, x') = 0$, that is, F is not a distribution function. The argument for this is a comparison to the G/G/1 queue with service times R_i and interarrival times $s \cdot g_i$.

Regarding ergodic properties of $(w_i : i \in \mathbb{N})$, we first observe that $F_i(x|y) \leq F_i(x) = F_i(x|0)$ for all $i \in \mathbb{N}$ and $x, y \in S$. Thus, $\lim_{i \rightarrow \infty} F_i(x|y) = 0$ for all $x, y \in S$ if $\rho \geq 1$ and $P(R_i = sg_i) < 1$. Assuming $\rho < 1$, Equation 8.1 in Ref. 1 states that $\lim_{i \rightarrow \infty} F_i(x|y) = F(x)$ for all $x, y \in S$, that is, F is the ergodic

distribution function of the Markov chain $(w_i : i \in \mathbb{N})$ which is independent of the initial value $w_1 = y$.

A fixed point equation for F can be derived as follows. Let P_i denote the probability measure on S that is determined by F_i . Recall definition (1) and define the set $\psi(x, b, c) := \{y \in S : \phi(y, b, c) \leq x\}$. Then $F_{i+1}(x) = \int P_i(\psi(x, b, c)) dH(b) dA(c)$ and $F_1(0) := 1$ provide a recursion for the distribution functions F_i . As $i \rightarrow \infty$, Lebesgue's theorem of dominated convergence allows to conclude

$$F(x) = \int P_\infty(\psi(x, b, c)) dH(b) dA(c), \quad (2)$$

where P_∞ denotes the measure determined by F . As the authors state, "it would be very desirable and interesting to solve the integral Equation (2), at least for interesting or important functions A and H . This, however, is likely to be very difficult."

THE G/M/s QUEUE

The most general special case that is solvable is the G/PH/s queue with phase-type service time distributions. Since PH distributions are dense within all distributions with positive support, the G/PH/s queue possibly might be regarded as an approximation of the G/G/s queue. Since a presentation of its solution is notationally rather extensive, we shall solve the more special case of the G/M/s queue instead. The solution has been derived in Ref. 2, while the following presentation is similar to Ref. 3 (Section 8.3). The solution of the G/PH/s queue follows along the same lines.

Assume that the interarrival times are iid with distribution function A , and every server has an exponential distribution with parameter $\mu > 0$. The service discipline is FCFS and the capacity of the waiting room is unbounded.

Since the exponential service time distribution is memoryless, the times of arrivals lead to an embedded Markov renewal chain. The system process $\mathcal{Q} = (Q_t : t \in \mathbb{R}_0^+)$ has state space $E = \mathbb{N}_0$, with Q_t indicating the number of users in the system (i.e., in service

or waiting) at time t . Define T_n as the time of the n th arrival and $X_n := Q_{T_n} - 1$ as the number of users in the system immediately before the n th arrival.

Clearly, the T_n are stopping times and X_n is a deterministic function of Q_{T_n} . We assume that A is not lattice, and $0 < \mathbb{E}(A) < \infty$. This implies, in particular, that $T_n \rightarrow \infty$ almost surely as n tends to infinity. The sequence $\mathcal{X} = (X_n : n \in \mathbb{N}_0)$ is a Markov chain since at times of arrivals the future of the system is determined only by the current number of users in the system, due to the memoryless property of the service times. The same property of the queue ensures

$$\begin{aligned} \mathbb{P}(Q_{T_n+t_1}=j_1, \dots, Q_{T_n+t_k}=j_k | Q_u \cdot u \leq T_n, X_n=i) \\ = \mathbb{P}(Q_{t_1}=j_1, \dots, Q_{t_k}=j_k | X_0=i). \end{aligned}$$

Thus the system process $\mathcal{Q}$ is Markov regenerative with embedded Markov renewal chain $(\mathcal{X}, \mathcal{T})$.

The transition probabilities p_{ij} of $\mathcal{X}$ are derived by the following considerations. Clearly, there may be only one arrival in any interval $[T_n, T_{n+1}]$. Hence $p_{ij} = 0$ for $j > i + 1$.

Now consider the case $j \leq i + 1 \leq s$. This means that during one interarrival period, no user is waiting and all are served with exponential intensity μ . Given that the interarrival period has length $t > 0$, the probability for any user to complete service is $1 - e^{-\mu t}$. A transition from state i to state j for the Markov chain $\mathcal{X}$ means that $i + 1 - j$ out of $i + 1$ users are served during one interarrival period. There are $\binom{i+1}{i+1-j} = \binom{i+1}{j}$ combinations to choose which users complete their service and which do not. Hence for $j \leq i + 1 \leq s$, we obtain

$$\begin{aligned} p_{ij} = \int_0^\infty \binom{i+1}{j} \\ \times (1 - e^{-\mu t})^{i+1-j} (e^{-\mu t})^j \, dA(t), \end{aligned}$$

by conditioning on the length t of the interarrival period.

The third case is $s \leq j \leq i + 1$. Here, all s servers are busy during the complete interarrival period. Then the number of

users served in time t is distributed like a Poisson process with intensity $s \cdot \mu$ evaluated at time t . Conditioning on the length of the interarrival period, we obtain

$$p_{ij} = \int_0^\infty \frac{(s\mu \cdot t)^{i+1-j}}{(i+1-j)!} e^{-s\mu t} \, dA(t), \quad (3)$$

for $s \leq j \leq i + 1$.

The last case to consider is $j < s < i + 1$. In this situation, there are $i + 1 - s$ users waiting in the queue at the beginning of the interarrival period, while $s - j$ servers are idle at the end of it. This is a mixture between the second and the third case. First the regime of the latter governs, until there are no waiting users any more (i.e., the queue has emptied). Then the former case applies. Thus, we condition first on the length t of the interarrival period and then on the time $u < t$ to empty the queue. The time to empty the queue has an Erlang distribution with $i + 1 - s$ stages and intensity $s\mu$. After that the number of served users has a binomial distribution with s degrees of freedom and parameter $p := e^{-\mu(t-u)}$. This leads to an expression

$$\begin{aligned} p_{ij} = \int_0^\infty \binom{s}{j} e^{-j\mu t} \int_0^t \frac{(s\mu u)^{i-s}}{(i-s)!} \\ \times (e^{-\mu u} - e^{-\mu t})^{s-j} s\mu \, du \, dA(t). \end{aligned}$$

Collecting these results, we can sketch the structure of the transition matrix as

$$P = \left(\begin{array}{ccc|ccc} p_{00} & p_{01} & & & & \\ p_{10} & p_{11} & p_{12} & & & \\ \vdots & & & \ddots & & \\ p_{s-2,0} & \dots & p_{s-2,s-1} & & & \\ p_{s-1,0} & \dots & p_{s-1,s-1} & \beta_0 & & \\ \hline p_{s,0} & \dots & p_{s,s-1} & \beta_1 & \beta_0 & \\ \vdots & & \vdots & \vdots & \ddots & \\ p_{s+n,0} & \dots & p_{s+n,s-1} & \beta_{s+1} & \beta_s & \dots & \beta_0 \\ \vdots & & \vdots & \vdots & \vdots & & \ddots \end{array} \right),$$

abbreviating $\beta_k := p_{i,i+1-k}$ in case (3). The nonspecified entries in P correspond to the case $j > i + 1$ and hence are zero. The matrix

P has an upper Hessenberg form. The most important part of it is the lower right-hand part containing the entries β_n . The other parts are boundary conditions.

In order to determine the stationary distribution of $\mathcal{X}$, we first consider the partition $E = F \cup F^c$ with $F = \{0, \dots, s-2\}$. We then obtain for the transition matrix P' of the Markov chain restricted to F^c the form

$$P' = \begin{pmatrix} r_0 & \beta_0 & & & \\ r_1 & \beta_1 & \beta_0 & & \\ r_2 & \beta_2 & \beta_1 & \beta_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

with $r_n = 1 - \sum_{k=0}^n \beta_k$. This has the same form as the respective transition matrix for the G/M/1 queue. Define $\rho = 1/(s\mu \cdot \mathbb{E}(A))$. The arguments for the G/M/1 queue all apply to the matrix P' with ρ defined as above. In the case $\rho < 1$, we obtain a stationary distribution $v' = v'P'$ given by

$$v'_n = (1 - \xi)\xi^n, \quad (4)$$

with $\xi = \sum_{n=0}^{\infty} \xi^n \beta_n$. For the case $\rho \geq 1$, it follows that P' (and hence P) is not positive recurrent.

In the following we assume $\rho < 1$. Then there is a unique extension of v' to a stationary measure for P , denoted by $v'' = v''P$. For this we have $v''_s = v'_{s-1}$ for all $n \geq s-1$ and $v''_{n+1} = \sum_{k=0}^{\infty} v''_k p_{k,n+1} = \sum_{k=n}^{\infty} v''_k p_{k,n+1}$ for all $n = 0, \dots, s-2$, which leads to

$$v''_n = \frac{1}{p_{n,n+1}} \times \left(v''_{n+1} \cdot (1 - p_{n+1,n+1}) - \sum_{k=n+2}^{\infty} v''_k p_{k,n+1} \right). \quad (5)$$

Thus v''_{s-2} and iteratively $v''_{s-3}, \dots, v''_0$ can be determined by formula (5). The stationary distribution $v = vP$ for $\mathcal{X}$ is then given as $v = c \cdot v''$ for a constant $c > 0$. This is determined by

$$c^{-1} = \sum_{n=0}^{\infty} v''_n = \sum_{n=0}^{s-2} v''_n + 1. \quad (6)$$

Altogether, formulae (4)–(6) yield the stationary distribution $v = vP$ for $\mathcal{X}$.

The asymptotic distribution of the queueing process Q is defined as the vector π on E with entries $\pi_n := \lim_{t \rightarrow \infty} \mathbb{P}(Q_t = n)$. This can now be obtained from the theory of Markov regenerative processes. For $n \geq s$, we obtain

$$\begin{aligned} \pi_n &= \frac{c}{\mathbb{E}(A)} \sum_{k=n-1}^{\infty} (1 - \xi)\xi^{k-(n-1)} \\ &\quad \times \int_0^{\infty} (1 - A(t))e^{-m\mu t} \frac{(m\mu t)^{k+1-n}}{(k+1-n)!} dt \\ &= \frac{c \cdot (1 - \xi)}{\mathbb{E}(A)} \int_0^{\infty} (1 - A(t))e^{-m\mu t} \xi^{n-m} \\ &\quad \times \sum_{k=n-1}^{\infty} \frac{(m\mu t \cdot \xi)^{k+1-n}}{(k+1-n)!} dt \\ &= \frac{c \cdot (1 - \xi)}{\mathbb{E}(A)} \xi^{n-m} \int_0^{\infty} (1 - A(t))e^{-m\mu t(1-\xi)} dt. \end{aligned}$$

Because of

$$\begin{aligned} \xi &= \sum_{n=0}^{\infty} \beta_n \xi^n = \int_0^{\infty} e^{-s\mu t} \sum_{n=0}^{\infty} \frac{(s\mu t \xi)^n}{n!} dA(t) \\ &= \int_0^{\infty} e^{-s\mu t(1-\xi)} dA(t) \end{aligned}$$

and integration by parts

$$\begin{aligned} &\int_0^{\infty} A(t)e^{-s\mu t(1-\xi)} dt \\ &= \frac{-1}{s\mu(1-\xi)} \left[A(t)e^{-s\mu t(1-\xi)} \right]_0^{\infty} - \frac{-1}{s\mu(1-\xi)} \\ &\quad \times \int_0^{\infty} e^{-s\mu t(1-\xi)} dA(t) \\ &= \frac{\xi}{s\mu(1-\xi)}. \end{aligned}$$

we obtain

$$\begin{aligned} &\int_0^{\infty} (1 - A(t))e^{-s\mu t(1-\xi)} dt \\ &= \int_0^{\infty} e^{-s\mu t(1-\xi)} dt - \int_0^{\infty} A(t)e^{-s\mu t(1-\xi)} dt \\ &= \frac{1}{s\mu(1-\xi)} - \frac{\xi}{s\mu(1-\xi)} = \frac{1}{s\mu}. \end{aligned}$$

Thus for $n \geq s$,

$$\pi_n = \rho \cdot c \cdot (1 - \xi) \xi^{n-m}.$$

Concerning the asymptotic probabilities π_n with $n = 1, \dots, s-1$, we employ the rate conservation law. We first give an intuitive explanation of this, an exact presentation can be found in Ref. 4 (Section VII.6). The arrival process of the G/M/s queue is a renewal process with renewal intervals distributed by A . Blackwell's theorem states that asymptotically an arrival occurs with constant rate $(\mathbb{E}(A))^{-1}$. Given that an arrival occurs, there is an asymptotic probability v_{n-1} of observing n users in the system. On the other hand, out of state $n \leq s-1$ (which has asymptotic probability π_n) the exponential servers provide a constant rate $n \cdot \mu$ to switch to state $n-1$. Now the rate conservation law states that asymptotically $v_{n-1}/\mathbb{E}(A) = \pi_n \cdot n\mu$, which means that the probability flow from state $n-1$ to state n equals the flow from n to $n-1$. This yields

$$\pi_n = \frac{v_{n-1}}{n\mu \cdot \mathbb{E}(A)},$$

for $n = 1, \dots, s-1$. Finally we obtain π_0 by normalization, that is,

$$\pi_0 = 1 - \sum_{n=1}^{\infty} \pi_n = 1 - \rho \cdot c - s\rho \sum_{n=1}^{s-1} \frac{v_{n-1}}{n}.$$

LITERATURE SURVEY

Waiting Times

General Results. Early results on work load and waiting time distributions can be found in Ref. 5. Here, the distribution of the actual waiting times is derived in terms of a solution of a set of integral equations. This approach has been extended in Ref. 6 and specified to (hyper)-exponential service times in Ref. 7. The stochastic chain of waiting times at the servers observed at the times that the users begin their service is analysed in Refs 1 and 8. It is shown there that an asymptotic (and hence stationary) distribution does exist if and only if $\rho < 1$. Harris

recurrence for this chain is investigated in Ref. 9. The work load vector as a process in continuous-time is analysed in Refs 10 and 11. This is generalized toward tandem queues with multiserver stations in Ref. 12.

It is shown in Ref. 13 that the working and the sojourn times of a user in a G/G/s queueing system are increasing and star-shaped with respect to the mean service time. Different service disciplines are compared in Refs 14 and 15. Some remarks on stochastic bounds for the delay distribution, commenting on Ref. 16, are given in Ref. 17.

Mean Stationary Waiting Time. Continuity of the mean stationary waiting time is proven in Ref. 18; see also Ref. 19. Bounds for this are derived in Refs 20 and 21, while conditions for it having finite moments are examined in Ref. 22. For further results on delay moments, see Refs 23–25. It is shown in Ref. 26 that the mean stationary waiting time is a convex and decreasing function of the number s of servers. This gives rise to a marginal benefit analysis for the optimal allocation of the number of servers; see also the earlier papers [27,28]. An example presented in Ref. 29 shows that even with bounded service times and an arbitrarily small traffic intensity, it may occur that the mean waiting time is infinite. Further results on the mean waiting time are given in Refs 30 and 31.

Asymptotic Results. Asymptotic growth of the maximum waiting time is analysed in Ref. 32. It is shown in Ref. 33, even for heterogeneous servers, that the waiting time is asymptotically exponential. Weaker logarithmic limits, assuming only positive Harris recurrence with no additional conditions, are derived in Ref. 34. The asymptotic behavior of the distribution tail of the stationary waiting time in the G/G/2 queue is studied in Ref. 35.

Approximations. An approximation of interarrival and service time distributions by mixed generalized Erlang distributions is analysed in Ref. 36. Some two-moment approximation formulas for the mean waiting time are proposed in Refs 37–41. It is, however, shown in Ref. 42 that any

approximation of the mean waiting time based only on the first two moments of the service time distribution must have an error with a strictly positive lower bound. Relations between the waiting time and the work load of an arbitrary server are derived in Ref. 43, where further bounds and approximations for the mean waiting time can be found.

Work Load

Existence and uniqueness of the stationary distribution of the work load at arrival times is shown in Ref. 1, for generalizations, see Refs 44–46. An alternative proof for uniqueness is given in Ref. 47. An expression for the total work load in the system is given in Ref. 43. A lower bound in terms of a related G/G/1 queue is given in Ref. 4 (Chapter 12). Sample-path inequalities regarding cyclical assignment of servers are derived in Ref. 48. Optimality of the FCFS service discipline in terms of remaining work load is shown in Refs 49–51. Asymptotics for the distribution tail of the stationary two-dimensional work load vector in the G/G/2 queue are derived in Ref. 35. Further results can be found in Refs 52 and 53.

Queue Length

General Results. The earliest papers [54–56] on the number of users in the system treat the special case M/M/s (see *The M/M/s Queue*), while the G/M/s queue is analyzed in Ref. 2. It is shown in Ref. 1 that for $\rho \geq 1$, there is no stationary distribution of the queue length, except in the trivial case of $\mathbb{P}(S = sT) = 1$. For the case of atomless interarrival and service time distributions, it is shown in Ref. 57 that the queue length converges in distribution as time passes. An implicit equation for the transient distribution is given in Ref. 58.

Stationary Distribution. It is shown in Ref. 59 that the stationary distributions at arbitrary points in time and at arrival or departure epochs are comparable with respect to the stochastic order relation. In Ref. 60, it is shown that the stationary distribution has an operator-geometric form if the interarrival

times are absolutely continuous. More results on the steady-state distribution can be found in Ref. 61. The distribution of the number of busy servers is studied in Ref. 62. Some results on the stationary probability that all servers are busy can be found in Refs 63 and 64.

Approximations and Asymptotic Results.

An approximation of interarrival and service time distributions by Coxian distributions is analysed in Ref. 65. A diffusion approximation for the stationary queue length is developed in Ref. 66; see also Ref. 67.

Asymptotic growth of the maximum queue length is analysed in Ref. 32. It is shown in Ref. 33, even for heterogeneous servers, that the queue length is asymptotically geometric. Asymptotics for the distribution tail of the queue length in the G/G/2 queue are derived in Ref. 35.

Busy Period

A conventional busy period begins when an arrival finds all servers idle and ends when, for the first time after that, a departure leaves the system empty. Finiteness of ϕ -moments for a busy period is investigated in Ref. 68. It is proved in Ref. 69 that in a GI/G/s queue with a traffic intensity $\rho < (s - m + 1)/s$, $m \geq 1$, at least m servers are simultaneously idle infinitely often with probability 1. The special case $m = s$ is treated in Ref. 70.

A slightly different concept, introduced in Ref. 71, is the full busy period: It begins when an arrival finds $s - 1$ users in the system, that is, exactly one server is idle, and ends when, for the first time after that, a departure leaves behind $s - 1$ users in the system. Conditions for the full busy period to have finite moments are derived in Ref. 72, the same for the related virtual work in Ref. 73. The distribution of the remaining time in a full busy period starting from a random epoch is given in Ref. 74.

Model Fitting

Given a sample path of the queue length process, conditions for identifiability of the interarrival and service time distributions are derived in Ref. 75. On the basis of the output epochs of a G/G/ ∞ queue, Ref. 76

provides a related result. A cross-spectral analysis of the G/G/ ∞ queue is performed in Ref. 77. A large sample test based on a normal approximation for the traffic intensity parameter ρ is proposed in Ref. 78.

Heavy Traffic

Under the assumption of heavy traffic (i.e., here $\rho > 1$), central limit theorems for the queue length process are derived in Refs 79 and 80. The analysis is based on Ref. 81. In Ref. 82, the limit as the arrival rate and the number s of servers jointly tend to infinity is investigated; see also Ref. 83. A diffusion approximation based on the first two moments of interarrival and service times is proposed in Ref. 84. Convergence theorems for the waiting time are derived in Ref. 85. An exponential limiting formula for the waiting time distribution is established in Ref. 86; see also Ref. 87 for further properties.

Special Cases

Approximations for specific PH/G/s queues are given in Ref. 88. A diffusion approximation for the G/G/s queue with group arrivals is given in Ref. 89. In the case of a LCFS service discipline, conditions for finite mean waiting times are derived in Ref. 29.

Finite Capacity Queues and Loss Systems. A diffusion approximation for the G/G/s/m finite capacity queue is proposed in Ref. 90. More results on the finite capacity case can be found in Ref. 91.

The rate of convergence to a steady-state in the G/G/s loss system is investigated in Ref. 92. The asymptotic performance of a G/G/s loss system with heterogeneous servers and retrials is analysed in Ref. 93. Some effects of the deletion of arrivals on the G/G/s loss system are investigated in Ref. 94.

Priority Classes. A conservation law for the G/G/s queue with priority classes has been proved in Ref. 95. An approximation of the mean response times in a G/G/s queue with priority classes and preemptive resume service discipline is proposed in Ref. 96. Waiting and sojourn times in a G/M/s queue with mixed priorities are determined in Ref. 97.

The G/G/ ∞ Queue. The infinite server case $s = \infty$ is clearly much simpler to analyze, as there are no waiting users. Existence of steady-state distributions for the number of users and the work load vector is proven in Ref. 98. Steady-state results for the number of users in the system are derived in Ref. 99. Results on the busy period are given in Ref. 100. The transient distribution of the number of users in the system and of the total work load is examined in Ref. 101. For approximations, see Ref. 102.

The G/PH/s Queue. It has been shown in Ref. 103 that the distribution of the stationary waiting time is phase-type. Its representation is derived in Ref. 104 in a form which is explicit up to the solution of a matrix fixed point problem. A Wiener–Hopf factorisation is developed in Ref. 105. An algorithmic solution to the stationary distribution of the number of users in the system is presented in Ref. 106 (Section 4.5). Efficient algorithms for its computation are given in Ref. 107.

REFERENCES

1. Kiefer J, Wolfowitz J. On the theory of queues with many servers. *Trans Am Math Soc* 1955;78:1–18.
2. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann Math Stat* 1953;24:338–354.
3. Breuer L, Baum D. An introduction to queueing theory. New York: Springer; 2005.
4. Asmussen S. Applied probability and queues. New York: Springer; 2003.
5. Pollaczek F. Über das warteproblem. *Math Z* 1934;38:492–537.
6. de Smit JHA. Some general results for many server queues. *Adv Appl Probab* 1973;5:153–169.
7. de Smit JHA. On the many server queue with exponential service times. *Adv Appl Probab* 1973;5:170–182.
8. Gnedenko B, Kovalenko IN. An introduction to queueing theory. 2nd ed. New York: Birkhauser; 1989.
9. Charlot F, Ghidouche M, Hamami M. Irreductibilité et récurrence au sens de Harris des “Temps d’attente” des files

- GI/G/q. *Z Wahrscheinlichkeitstheorie verw Geb* 1978;43:187–203.
10. Foss SG, Kalashnikov VV. Regeneration and renovation in queues. *Queueing Syst* 1991;8:211–224.
 11. Asmussen S, Foss SG. Renovation, regeneration, and coupling in multiple-server queues in continuous time. *Frontiers in pure and applied probability*. Utrecht: VSP; 1993. pp. 1–6.
 12. Sigman K. Regeneration in tandem queues with multi-server stations. *J Appl Probab* 1988;25:391–403.
 13. Shanthikumar J, Wu C. On the starshapeness of G/G/c queueing systems. *Adv Appl Probab* 1991;23:431–435.
 14. Gittins JC. A comparison of service disciplines for GI/G/m queues. *Math Operationsf Stat Ser Optim* 1978;9:255–260.
 15. Liu Z, Towsley D. Effects of service disciplines in G/G/s queueing systems. *Ann Oper Res* 1994;48:401–429.
 16. Wolff RW. An upper bound for multi-channel queues. *J Appl Probab* 1977;14:884–888.
 17. Whitt W. On stochastic bounds for the delay distribution in the GI/G/s queue. *Oper Res* 1981;29:604–608.
 18. Asmussen S, Johansen H. Über eine Stetigkeitsfrage betreffend das Bedienungssystem GI/G/s. *Elektron. Inf.-verarb. Kybernetik* 1986;22(10/11):565–570.
 19. Brandt A, Lisek B. On the continuity of G/GI/m queues. *Math Operationsf Stat Ser Stat* 1983;14:577–587.
 20. Brumelle SL. Some inequalities for parallel-server queues. *Oper Res* 1971;19:402–413.
 21. Scheller-Wolf A, Sigman K. New bounds for expected delay in FIFO GI/GI/c queues. *Queueing Syst* 1997;26:169–186.
 22. Scheller-Wolf A, Sigman K. Delay moments for FIFO GI/GI/s queues. *Queueing Syst* 1997;25:77–95.
 23. Scheller-Wolf A. Further delay moment results for FIFO multi-server queues. *Queueing Syst* 2000;34:387–400.
 24. Scheller-Wolf A. Necessary and sufficient conditions for delay moments in FIFO multi-server queues: why s slow servers are better than one fast server for heavy-tailed systems. *Oper Res* 2003;51:748–758.
 25. Scheller-Wolf A. Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues. *Queueing Syst* 2006;54:221–232.
 26. Weber RR. On the marginal benefit of adding servers to G/GI/m queues. *Manage Sci* 1980;26:946–951.
 27. Rolfe AJ. A note on marginal allocation in multiple-server service systems. *Manage Sci* 1971;17:656–658.
 28. Weber RR. On the optimal assignment of customers to parallel servers. *J Appl Probab* 1978;15:406–413.
 29. Ritter G, Wacker U. The mean waiting time in a G/G/m/∞ queue with the LCFS-P service discipline. *Adv Appl Probab* 1991;23:406–428.
 30. Daley DJ, Rolski T. Some comparability results for waiting times in single- and many-server queues. *J Appl Probab* 1984;21:887–900.
 31. Daley DJ. Some results for the mean waiting-time and workload in GI/GI/k queues. In: Dshalalow JH, editor. *Frontiers in queueing: models and applications in science and engineering*. Boca Raton (FL): CRC Press; 1997. pp. 35–59.
 32. Sadowsky JS, Szpankowski W. Maximum queue length and waiting time revisited: multiserver G/G/c queue. *Probab Eng Inf Sci* 1992;6:157–170.
 33. Sadowsky JS, Szpankowski W. The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. I: tight limits. *Adv Appl Probab* 1995;27:532–566.
 34. Sadowsky JS. The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. II: positive recurrence and logarithmic limits. *Adv Appl Probab* 1995;27:567–583.
 35. Foss S, Korshunov D. Heavy tails in multi-server queue. *Queueing Syst* 2006;52:31–48.
 36. Bertsimas D. An exact FCFS waiting time analysis for a general class of G/G/s queueing systems. *Queueing Syst* 1988;3:305–320.
 37. Seelen LP, Tijms HC. Approximations for the conditional waiting times in the GI/G/c queue. *Oper Res Lett* 1984;3:183–190.
 38. Kimura T. A two-moment approximation for the mean waiting time in the GI/G/s queue. *Manage Sci* 1986;32:751–763.
 39. Kimura T. Approximating the mean waiting time in the GI/G/s queue. *J Oper Res Soc* 1991;42:959–970.

40. Kimura T. Approximations for the waiting time in the GI/G/s queue. *J Oper Res Soc Jpn* 1991;34:173–186.
41. Whitt W. Approximations for the GI/G/m queue. *Prod Oper Manage* 1993;2:114–161.
42. Gupta V, Harchol-Balter M, Dai JG, *et al.* On the inapproximability of M/G/K: why two moments of job size distribution are not enough. *Queueing Syst* 2010;64:5–48.
43. Hokstad P. Relations for the workload of the GI/G/s queue. *Adv Appl Probab* 1985;17:887–904.
44. Loynes R. The stability of a queue with non-independent inter-arrival and service times. *Proc Camb Philol Soc* 1962;58:497–520.
45. Brandt A. On stationary waiting times and limiting behaviour of queues with many servers I: the general G/G/m/ ∞ case. *Elektron. Inf.-verarb. Kybernetik* 1985;21: 47–64.
46. Brandt A. On stationary waiting times and limiting behaviour of queues with many servers. II: the G/GI/m/ ∞ case. *Elektron. Inf.-verarb. Kybernetik* 1985;21:151–162.
47. Wolfson B. The uniqueness of stationary distributions for the GI/G/S queue. *Math Oper Res* 1986;11:514–520.
48. Loulou R. Two sample-path inequalities for G/G/k queues. *INFOR* 1983;21:136–144.
49. Daley DJ. Certain optimality properties of the first-come first-served discipline for G/G/s queues. *Stoch Proc Appl* 1987;25:301–308.
50. Wolff RW. Upper bounds on work in system for multichannel queues. *J Appl Probab* 1987;24:547–551.
51. Foss SG, Chernova NI. On optimality of the FCFS discipline in multiserver queueing systems and networks. *Siberian Math J* 2001;42:372–385.
52. Kopocinski B. On the virtual waiting time in the GI/G/s queue. *Bull Acad Pol Sci Ser Sci Math Astron Phys* 1976;24:779–782.
53. Kopocinski B, Rolski T. A note on the virtual waiting time in the GI/G/s queue. *Bull Acad Pol Sci Ser Sci Math Astron Phys* 1977;25:1279–1280.
54. Brockmeyer E, Halstrom H, Jensen A. The life and works of A. K. Erlang. Copenhagen; 1948.
55. Molina E. Application of the theory of probability to telephone trunking problems. *Bell Syst Tech J* 1927;6:461–494.
56. Kolmogorov A. Sur le probleme d'attente. *Rec Math* 1931;38:101–106.
57. Miller DR, Sentilles F. Translated renewal processes and the existence of a limiting distribution for the queue length of the GI/G/s queue. *Ann Probab* 1975;3:424–439.
58. Hou Z, Yuan C, Zou J, *et al.* Transient distribution of the length of GI/G/N queueing systems. *Stoch Anal Appl* 2003;21: 567–592.
59. König D, Schmidt V, Stoyan D. On some relations between stationary distributions of queue lengths and imbedded queue lengths in G/G/s queueing systems. *Math Operationsf Stat* 1976;7:577–586.
60. Breuer L. Transient and stationary distributions for the GI/G/k queue with Lebesgue-dominated inter-arrival time distribution. *Queueing Syst* 2003;45:47–57.
61. Kopocinska I, Kopocinski B. Steady-state distributions of queue length in the GI/G/s queue. *Zastosow Mat* 1977;16:39–46.
62. Borovkov AA. On limit laws for service processes in multi-channel systems. *Siberian Math J* 1967;8:746–763.
63. Sobel MJ. Simple inequalities for multiserver queues. *Manage Sci* 1980;26:951–956.
64. Heyman DP. Comments on a queueing inequality. *Manage Sci* 1980;26:956–959.
65. Bertsimas D. An analytic approach to a general class of G/G/s queueing systems. *Oper Res* 1990;38:139–155.
66. Kimura T. An M/M/s-consistent diffusion model for the GI/G/s queue. *Queueing Syst* 1995;19:377–397.
67. Wu J-S. Refining the diffusion approximation for the G/G/c queue. *Comput Math Appl* 1990;20:31–36.
68. Thorisson H. The queue GI/GI/k: finite moments of the cycle variables and uniform rates of convergence. *Stoch Models* 1985;1:221–238.
69. Kennedy DP. A note on the number of busy servers in a GI/G/s queue in light traffic. *J Appl Probab* 1972;9:868–869.
70. Whitt W. Embedded renewal processes in the GI/G/s queue. *J Appl Probab* 1972; 9:650–658.
71. Kiefer J, Wolfowitz J. On the characteristics of the general queueing process with applications to random walks. *Ann Math Stat* 1956;27:147–161.
72. Ghahramani S, Wolff RW. On finite moments of full busy periods of GI/G/c queues. *J Appl Probab* 2005;42:275–278.

73. Ghahramani S. Finiteness of moments of virtual work for GI/G/c queues. *J Appl Probab* 1989;26:423–425.
74. Ghahramani S. On remaining full busy periods of GI/G/c queues and their relation to stationary point processes. *J Appl Probab* 1990;27:232–236.
75. Ross SM. Identifiability in GI/G/k queues. *J Appl Probab* 1970;7:776–780.
76. Kendall DG, Lewis T. On the structural information contained in the output of GI/G/∞. *Z Wahrscheinlichkeitstheorie Verw Geb* 1965;4:144–148.
77. Brillinger DR. Cross-spectral analysis of processes with stationary increments including the stationary G/G/∞ queue. *Ann Probab* 1974;2:815–827.
78. Rao SS, Harishchandra K. On a large sample test for the traffic intensity in GI/G/s queue. *Nav Res Log Q* 1986;33:545–550.
79. Jin X. On Berry-Esseen rate for queue length of the GI/G/K system in heavy traffic. *J Appl Probab* 1988;25:596–611.
80. Jin X, Wang R. On the speed of convergence for the queue length process of the GI/G/K system in heavy traffic. *J Appl Probab* 1990;27:417–424.
81. Kennedy DP. Rates of convergence for queues in heavy traffic. I. *Adv Appl Probab* 1972;4:357–381.
82. Whitt W. On the heavy-traffic limit theorem for GI/G/∞ queues. *Adv Appl Probab* 1982;14:171–190.
83. Glynn PW. On the Markov property of the GI/G/∞ Gaussian limit. *Adv Appl Probab* 1982;14:191–194.
84. Halachmi B, Franta W. A diffusion approximation to the multi-server queue. *Manage Sci* 1978;24:522–529.
85. Loulou R. Multi-Channel Queues in Heavy Traffic. *J Appl Probab* 1973;10:769–777.
86. Köllerström J. Heavy traffic theory for queues with several servers. I. *J Appl Probab* 1974;11:544–552.
87. Köllerström J. Heavy traffic theory for queues with several servers. II. *J Appl Probab* 1979;16:393–401.
88. van Hoorn M, Seelen L. Approximations for the GI/G/c queue. *J Appl Probab* 1986;23:484–494.
89. Ohsone T. Diffusion approximation for a GI/G/m queue with group arrivals. *J Oper Res Soc Jpn* 1984;27:78–96.
90. Whitt W. A diffusion approximation for the G/GI/n/m queue. *Oper Res* 2004;52:922–941.
91. Schmidt V. On some relations between stationary time and customer state probabilities for queueing systems G/GI/s/r. *Math Operationsf Stat Ser Optim* 1978;9:261–272.
92. Sejdí L. Rate of convergence to a steady state in queueing systems of type G/G/m/0. *J Sov Math* 1986;32:79–89.
93. Pourbabai B. Asymptotic analysis of G/G/K queueing-loss system with retrials and heterogeneous servers. *Int J Syst Sci* 1988;19:1047–1052.
94. Ziya S, Ayhan H, Foley RD, *et al.* A monotonicity result for a G/GI/c queue with balking or reneging. *J Appl Probab* 2006;43:1201–1205.
95. Bartsch B, Bolch G. A conservation law for G/G/m queueing systems. *Acta Informatica* 1978;10:105–109.
96. Tabet-Aouel N, Kouvatsos DD. On an approximation to the mean response times of priority classes in a stable G/G/c/PR queue. *J Oper Res Soc* 1992;43:227–239.
97. Zeltyn S, Feldman Z, Wasserkrug S. Waiting and sojourn times in a multi-server queue with mixed priorities. *Queueing Syst* 2009;61:305–328.
98. Shanbhag DN. A note on the queueing system GI/G/∞. *Proc Camb Philol Soc* 1968;64:147–161.
99. Kaplan N. Limit theorems for a GI/G/∞ queue. *Ann Probab* 1975;3:780–789.
100. Kopocinska I. Busy period problems in the GI/G/∞ queue. *Zastosow Mat* 1984;18:195–201.
101. Ayhan H, Limon-Robles J, Wortman M. On the time-dependent occupancy and backlog distributions for the GI/G/∞ queue. *J Appl Probab* 1999;36:558–569.
102. Newell G. Approximate stochastic behavior of N-server service systems with large N. New York: Springer; 1973.
103. Asmussen S, O’Cinneide CA. Representations for matrix-geometric and matrix-exponential steady-state distributions with applications to many-server queues. *Stoch Models* 1998;14:369–387.
104. Asmussen S, Moller J. Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues. *Queueing Syst* 2001;37:9–29.

105. de Smit JHA. Explicit Wiener-Hopf factorizations for the analysis of multidimensional queues. In: Dshalalow JH, editor. *Advances in queueing: theory, methods, and open problems*. Boca Raton (FL): CRC Press; 1995. pp. 392–419.
106. Neuts MF. *Matrix-geometric solutions in stochastic models*. Baltimore (MD): Johns Hopkins University Press; 1981.
107. Lucantoni DM, Ramaswami V. Algorithms for the multi-server queue with phase type service. *Stoch Models* 1985;1(3):393–417.

THE $G/M/1$ QUEUE

DAVID PERRY
Department of Statistics,
University of Haifa,
Haifa, Israel

INTRODUCTION

This article constitutes a natural continuation of the article titled *The $M/G/1$ Queue* in this encyclopedia, that is focused on the $M/G/1$ queue. Accordingly, the $G/M/1$ queue is characterized as a single server queue with exponentially distributed (μ) service times which are independent of each other and of the arrival process in which the successive interarrival times are independent and identically distributed (iid) random variables. Let G denote the interarrival time distribution, with $G^*(\alpha)$ being the Laplace Stieltjes transform (LST) and $\lambda = -\frac{dG^*(\alpha)}{d\alpha}|_{\alpha=0}$ denoting the arrival rate. We proceed according to the concept of duality between the $G/M/1$ and the $M/G/1$ queues; this argument says that in order to realize the $G/M/1$ queue one can use the same set of random variables as those used for the $M/G/1$ queue, but under the modification that the interarrival times and the services are reversed. This means that in the dual $M/G/1$ queue, the arrival process is Poisson with rate μ and G is the service distribution. We designate $\rho := \mu/\lambda$. Then, the *traffic intensity* of the primal $M/G/1$ queue is ρ , but for the dual $G/M/1$ queue it is ρ^{-1} . Treatments of the $G/M/1$ queue may be found in several books on queueing theory [1–7].

For later use, we start by citing two limit theorems about irreducible aperiodic discrete-time Markov chain on a countable state space. Theorems 1 and 2 below enable us to classify the states of the chain into one out of three classes of equivalence: either all states are positive recurrent, null recurrent, or transient. Accordingly, in

Theorem 1 below we introduce a necessary and sufficient condition for the chain to be ergodic (which means that any set of states is positive recurrent). In particular, if the chain is not ergodic, then all states are either null recurrent or transient. In the nonergodic case, we give in Theorem 1 a necessary and sufficient condition for all states to be recurrent. Then, in Theorem 2 below we introduce a necessary and sufficient condition for all states to be recurrent (which implies that all states must be null recurrent).

It turns out that the embedded processes of the number of customers just after departures in the $M/G/1$ queue and the number of customers just before arrivals in the $G/M/1$ queue, respectively, are irreducible aperiodic discrete-time Markov chains with state space $E = \{0, 1, 2, \dots\}$. In the section titled “Duality,” we apply Theorems 1 and 2 to the duality between the $M/G/1$ and the $G/M/1$ queues. As will be seen, with ρ fixed, when the $M/G/1$ system is positive recurrent the $G/M/1$ system is transient and vice versa. When the $G/M/1$ system is null recurrent the $M/G/1$ system is also null recurrent.

In the sequel, we restrict the attention to the stable case of the $G/M/1$ queue (the case $\rho > 1$). We focus on the main factors of interest: limiting number of customers seen by arbitrary arrivals, waiting times and sojourn times in steady state, and busy periods and idle periods.

Remark 1. The $G/M/1$ system is a less natural model than the $M/G/1$ since, in practice, services are rarely exponential because of the memoryless property. However, as will be seen, the *attained waiting time* (AWT) process (which is the reversed version of the *work* process) of the $G/M/1$ queue can also be interpreted as the content level of an insurance risk cash fund with Poisson(μ) arrivals and generally distributed claims. Then, in the case of $\rho > 1$, the $G/M/1$ queueing system is stable, but for the insurance risk interpretation a ruin (bankruptcy), somewhere in the future, is a certain event (the latter statement

is also valid for the case of $\rho = 1$). In the case of $\rho < 1$, the $G/M/1$ queueing system is unstable, and under the insurance risk interpretation, no matter what the initial state is, a ruin is not a sure event.

TWO LIMIT THEOREMS

We introduce two well-known limit theorems that can be found in many introductory books about Markov chains (therefore, the proofs are omitted here; see Theorem 2.1, p. 152 and Theorem 5.34, p. 176 in Çinlar [8]). The relevance of the two theorems to the $M/G/1$ and the $G/M/1$ queues stems from the fact that the number of customers left behind a departing customer in the $M/G/1$ queue and the number of customers seen at the moment of arrival in the $G/M/1$ queue are irreducible aperiodic Markov chains.

Let $\mathbf{X}$ be an arbitrary irreducible aperiodic Markov chain with state space $E = \{0, 1, 2, \dots\}$ and transition matrix $\mathbf{P} := \|p(i, j)\|$.

Theorem 1. *All states are positive recurrent if and only if the system of linear equations*

$$\begin{aligned} \pi &= \pi \mathbf{P} \\ \text{and} \\ \pi \mathbf{1} &= 1 \end{aligned}$$

has a solution π . If there exists a solution π then it is strictly positive, there are no other solutions, and we have

$$\pi_j = \lim_{n \rightarrow \infty} p^n(i, j),$$

for all $i, j \in E$.

Before introducing the second theorem, we mention that if there are infinitely many states, it is possible that all of them are transient. Then, it is possible for the chain to remain in the set of transient states forever. Practically speaking, let A be a subset of the state space E , and let $\mathbf{Q} := \|Q(i, j)\|$ be the matrix obtained from $\mathbf{P}$ by deleting all the rows and columns corresponding to the states which are not in A . Then,

$$\begin{aligned} \sum_{j \in A} Q^n(i, j) &= \Pr(X_1 \in A, \dots, X_n \in A \mid X_0 = i), \\ i &\in A. \end{aligned} \quad (1)$$

Clearly, the number in Equation (1) decreases as n increases. Let

$$f(i) = \lim_{n \rightarrow \infty} \sum_{j \in A} Q^n(i, j), \quad i \in A;$$

in words, $f(i)$ is the probability that starting at $i \in A$, the chain stays in the set A forever. Then, for its computation of $f(i)$ we have the following.

Theorem 2. *The function f is the maximal solution of the system of linear equations*

$$\mathbf{h} = \mathbf{Q}\mathbf{h}, \quad 0 \leq \mathbf{h} \leq 1.$$

Either $f = 0$ or $\sup_{i \in A} f(i) = 1$.

It was already mentioned that the special duality between the $G/M/1$ queue and the $M/G/1$ queue is achieved by reversing the roles of the interarrival times and the service times. As will be seen, when $\rho > 1$ the $M/G/1$ queue size is likely to increase to infinity, but in the $G/M/1$ queue the server can keep up with arrivals and we expect the queue sizes to be positive recurrent; if $\rho < 1$ then the $G/M/1$ system is transient and the $M/G/1$ queue is positive recurrent; in the case of $\rho = 1$ both the $G/M/1$ queue and the $M/G/1$ queue are null recurrent. Obviously, when the embedded $M/G/1$ and $G/M/1$ processes are positive recurrent, null recurrent, or transient, the continuous-time processes are also positive recurrent, null recurrent, or transient, respectively.

DUALITY

The mentioned duality extends somewhat further; in the case of positive recurrence, it enables to obtain the limiting law of the queue size seen by an arriving customer in the $G/M/1$ system as a function of the probabilities of never becoming empty in the $M/G/1$ system. To see this, let X_n^* be the number of customers in the $M/G/1$

system after the instant of the n th departure. It follows from the article titled **The M/G/1 Queue** in this encyclopedia that $\mathbf{X}^* = \{X_k^* : k = 1, 2, \dots\}$ is an irreducible aperiodic Markov chain with state space $E = \{0, 1, \dots\}$ and transition matrix

$$\mathbf{P}^* = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 & \cdot & \cdot \\ q_0 & q_1 & q_2 & q_3 & \cdot & \cdot \\ 0 & q_0 & q_1 & q_2 & \cdot & \cdot \\ 0 & 0 & q_0 & q_1 & q_2 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \quad (2)$$

where

$$q_n = \int_0^\infty \frac{e^{-\mu t} (\mu t)^n}{n!} dG(t) \quad (3)$$

is the probability for exactly n arrivals within an arbitrary service.

To show that the chain $\mathbf{X}^*$ is positive recurrent and aperiodic, one has to apply Theorem 1 and solve the set of linear equations

$$\begin{aligned} \pi^* &= \pi^* \mathbf{P}^* \\ \text{and} \\ \pi^* \mathbf{1} &= \mathbf{1}. \end{aligned}$$

It follows from Theorem 1 that the solution of π exists if and only if $\rho < 1$ and is given by its generating function

$$\Pi^*(z) = \frac{(1 - \rho)(1 - z)G^*(\lambda(1 - z))}{G^*(\lambda(1 - z)) - z}, \quad (4)$$

where G^* is the LST of the service time distribution (see **The M/G/1 Queue**).

Remark 2. When restricting attention to the case of positive recurrence it should be noted that, in the M/G/1 queue, the number of customers left behind a departing customer and the number of customers in steady state are stochastically equal by ASTA [9], so that Equation (4) is also the generating function of the number of customers in steady state. However, the number of customers seen by an arriving customer and the number of customers in steady state in the G/M/1 queue are not necessarily stochastically equal.

Theorem 3 below is of crucial importance for the condition of stability for the G/M/1

queue, although it is applied to the M/G/1 queue. Based on Theorems 1 and 2 above, Theorem 3 shows that the chain $\mathbf{X}^*$ (the embedded chain just after the moments of departures in the M/G/1 queue) is recurrent non-null if and only if $\rho < 1$, null recurrent if and only if $\rho = 1$, and transient if and only if $\rho > 1$. To prove the latter statement, we interpret $f(j)$ as the probability that the M/G/1 queue, starting with j customers, never becomes empty.

Theorem 3. We have

$$f(j) = 1 - \beta^j, \quad j = 1, 2, \dots, \quad (5)$$

where β is the smallest number in $[0, 1]$ satisfying

$$\beta = q_0 + q_1\beta + q_2\beta^2 + \dots \quad (6)$$

Proof. We observe from the shape of the transition matrix $\mathbf{P}^*$ in Equation (2) that starting from $j + 1$, in order for the queue to become empty, $\mathbf{X}^*$ must first reach j . Therefore, starting from $j + 1$, the queue can avoid becoming empty either by never reaching j or, if it does reach j , by never becoming empty starting from j . Since the probability of never reaching j starting with $j + 1$ customers is the same for all j , it is equal to $f(1)$. Thus, we get

$$f(j + 1) = f(1) + [1 - f(1)]f(j), \quad j = 1, 2, \dots$$

Assigning $f(1) = 1 - \beta$ and solving this recursively, we obtain Equation (5). It remains to compute β . From Theorem 2, f is the maximal solution of $\mathbf{h} = \mathbf{Q}^*\mathbf{h}$, $0 \leq \mathbf{h} \leq \mathbf{1}$, with $\mathbf{Q}^*$ given by

$$\mathbf{Q}^* = \begin{bmatrix} q_1 & q_2 & q_3 & \cdot & \cdot \\ q_0 & q_1 & q_2 & \cdot & \cdot \\ 0 & q_0 & q_1 & q_2 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}. \quad (7)$$

From the first equation of $\mathbf{h} = \mathbf{Q}^*\mathbf{h}$ we see that f must satisfy

$$f(1) = q_1f(1) + q_2f(2) + \dots,$$

which, upon using the values in Equation (5), implies that β satisfies Equation (6). Since f

is the maximal such solution, β must be the smallest possible solution of Equation (6). We need to show that $\beta < 1$ (and therefore that $f \neq 0$, which in turn implies that $\mathbf{X}^*$ is transient) if and only if $\rho > 1$. Consider the function

$$K(\beta) = q_0 + q_1\beta + q_2\beta^2 + \dots, \quad 0 \leq \beta \leq 1.$$

Clearly, $K(\beta)$ is nonnegative, increasing, and has an increasing derivative. This implies that $K(\beta)$ can intersect the line $H : H(\beta) = \beta$ at most once in $(0, 1)$ (of course, they always intersect at $\beta = 1$). Consider the derivative $K'(\beta)$ at $\beta = 1$. If $K'(1) > 1$, the point of intersection β is strictly less than 1; otherwise, if $K'(1) \leq 1$, then K and H have no point of intersection in $(0, 1)$, and the smallest number β with the desired properties is 1. Finally, note that $K'(1) = \rho$.

Now consider the G/M/1 queue and denote the number of customers in the system by X_n , just before the time T_n of the n th arrival. We introduce the notation $r_n = q_{n+1} + q_{n+2} + \dots$. Here q_n is interpreted as the probability that the server completes exactly n services during an interarrival time provided that there are sufficiently many customers available to be served. Obviously,

$$\rho = \sum_{n=1}^{\infty} nq_n = r_0 + r_1 + \dots = \mu/\lambda$$

is the expected number of services that the server is capable of completing during an interarrival time.

Claim 1. $\mathbf{X} = \{X_n : n = 1, 2, \dots\}$ is a Markov chain with state space $E = \{0, 1, \dots\}$ and transition matrix $\mathbf{P} := \|p(i, j)\|$ given by

$$\mathbf{P} = \begin{bmatrix} r_0 & q_0 & & & 0 \\ r_1 & q_1 & q_0 & & \\ r_2 & q_2 & q_1 & q_0 & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}. \quad (8)$$

To validate Claim 1 let M_{n+1} be the number of services completed during the $(n+1)$ th interarrival time $[T_n, T_{n+1})$. Then,

$$X_{n+1} = X_n + 1 - M_{n+1}. \quad (9)$$

In other words, X_{n+1} is the number of customers that were in the system before the n th arrival plus the n th customer and less those whose services were completed during the $(n+1)$ th interarrival time. From Equation (9) it follows that $\mathbf{X}$ is a Markov chain only if M_{n+1} is conditionally independent of the past history before T_n given the present number X_n . But the latter statement is obvious because of the independence of the service times and the interarrival times, and the fact that, by the memoryless property of the exponential distribution, the remaining service time of the customer who is being served at time T_n (if any) has the same exponential distribution as any service time. The above reasoning shows that as long as there are customers to be served, the conditional distribution of the number of services completed is Poisson: with $Z = T_{n+1} - T_n$,

$$\begin{aligned} \Pr[M_{n+1} = k \mid X_n, Z] \\ = \begin{cases} \frac{e^{-\lambda Z} (\lambda Z)^k}{k!} & \text{if } X_n + 1 > k \\ \sum_{m=k}^{\infty} \frac{e^{-\lambda Z} (\lambda Z)^m}{m!} & \text{if } X_n + 1 = k \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Taking expectation with respect to Z , which is independent of X_n , we obtain

$$\begin{aligned} \Pr[M_{n+1} = k \mid X_n = i] \\ = \begin{cases} q_k & \text{if } k \leq i \\ r_{k-1} & \text{if } k = i + 1 \\ 0 & \text{otherwise.} \end{cases} \quad (10) \end{aligned}$$

Combining Equations (9) and (10), we see that the transition probabilities $p(i, j)$ are as claimed.

It is clear from an inspection of Equation (8) that the Markov chain $\mathbf{X}$ is irreducible and aperiodic. Therefore, all the states are of the same class of equivalence [8, Chapter 5; Theorem 3.16]. The proof of the next theorem is based on the duality between the G/M/1 queue and the M/G/1 queue; we show that the Markov chain $\mathbf{X}$ is positive recurrent if and only if $\rho > 1$.

Theorem 5. $\mathbf{X}$ is positive recurrent if and only if $\rho > 1$. If $\rho > 1$, the limiting distribution

$$\pi_j = \lim_{n \rightarrow \infty} \Pr(X_n = j \mid X_0 = i)$$

is given by

$$\pi_j = (1 - \beta)\beta^j, \quad j = 0, 1, 2, \dots,$$

where β is given in Equation (6). If $\rho \leq 1$, then $\pi_j = 0$ for all j .

Proof. According to Theorem 1, $\mathbf{X}$ is positive recurrent if and only if $\pi = \pi \mathbf{P}$, $\pi \mathbf{1} = \mathbf{1}$ has a solution. Writing out the equations, with $r_n = q_{n+1} + q_{n+2} + \dots$, we have

$$\begin{aligned} \pi_0 &= q_1\pi_0 + q_2\pi_0 + q_3\pi_0 + \dots \\ &\quad + q_2\pi_1 + q_3\pi_1 + \dots \\ \pi_1 &= q_0\pi_0 + q_1\pi_1 + q_2\pi_2 + \dots \\ \pi_2 &= q_0\pi_1 + q_1\pi_2 + q_2\pi_3 + \dots \\ &\vdots \end{aligned} \quad (11)$$

Let a function f be defined by

$$f(j) = \pi_0 + \pi_1 + \dots + \pi_{j-1}, \quad j = 1, 2, \dots \quad (12)$$

The equation for π_0 gives an equation for $f(1)$, summing the equations for π_0 and π_1 yields an equation for $f(2)$, and so on. We obtain

$$\begin{aligned} f(1) &= q_1f(1) + q_2f(2) + q_3f(3) + \dots, \\ f(2) &= q_0f(1) + q_1f(2) + q_2f(3) + \dots, \\ f(3) &= q_0f(2) + q_1f(3) + \dots, \end{aligned}$$

so that f satisfies

$$\mathbf{f} = \mathbf{Q}^* \mathbf{f}, \quad (13)$$

where $\mathbf{Q}^*$ is given in Equation (7). In other words, $\mathbf{Q}^*$ is the matrix obtained from $\mathbf{P}^*$ in Equation (2) by deleting the first row and the first column (corresponding to state $\{0\}$).

We are interested in the solution of Equation (13) satisfying

$$\lim_{j \rightarrow \infty} f(j) = \sum_{j=0}^{\infty} \pi_j.$$

We conclude (see Proposition 4.5, p. 167 of Çinlar [8]) that such a solution exists if and only if the chain $\mathbf{X}$ is transient. By Theorem 3, this is true if and only if $\rho > 1$. If $\rho > 1$, the solution for f is given by Equation (5), where β satisfies Equation (6). Solving for π in Equation (11) by using Equations (12) and (5) for f , we obtain

$$\begin{aligned} \pi_0 &= f(1) = 1 - \beta, \\ \pi_1 &= f(2) - f(1) = (1 - \beta)\beta, \\ &\vdots \end{aligned}$$

which is the same as the claimed limiting distribution π in Theorem 4. Finally, the statement that $\rho \leq 1$ implies that $\pi = \mathbf{0}$ follows again from the duality between the M/G/1 and the G/M/1 queues.

Next we restrict the attention to the case of $\rho > 1$ and focus on other relevant measures of the G/M/1 queue.

WORK AND ATTAINED WAITING TIME

In addition to the duality argument above, several derivations in this review are based on sample-path analysis of two dual compound processes; the so-called *work process* (or *workload process*), sometimes called the *virtual waiting time* (VWT) and the *Attained waiting time* (AWT) processes. The times between successive arrivals are iid random variables $S_1, S_2, \dots$, and the service requirements of arriving customers are iid random variables $Z_1, Z_2, \dots$.

Let $\mathbf{N} = \{N(t) : t \geq 0\}$ and $\mathbf{\Lambda} = \{\Lambda(t) : t \geq 0\}$ be counting processes such that for all $t \geq 0$, $n = 0, 1, \dots$ and $m = 0, 1, \dots$, $\{N(t) \geq n\} = \{Z_1 + \dots + Z_n \leq t\}$ and $\{\Lambda(t) \geq m\} = \{S_1 + \dots + S_m \leq t\}$. Obviously, $\mathbf{N}$ is a Poisson process with rate μ , and $\mathbf{\Lambda}$ is a renewal process whose inter-renewal distribution is $G(\cdot)$ with mean $1/\lambda$. Now define the continuous-time random walks $\mathbf{X} = \{X(t) : t \geq 0\}$ where $X(t) = t - (S_1 + \dots + S_{N(t)})$

and $\mathbf{Y} = \{Y(t) : t \geq 0\}$ where $Y(t) = (Z_1 + \dots + Z_{\Lambda(t)}) - t$. Then, construct the reflected processes $\mathbf{A} = \{A(t) : t \geq 0\}$ and $\mathbf{V} = \{V(t) : t \geq 0\}$, by $A(t) = X(t) - \inf_{0 \leq s \leq t} X(s)$ and $V(t) = Y(t) - \inf_{0 \leq s \leq t} Y(s)$. Here $\mathbf{A}$ is interpreted as the conditional AWT process of the G/M/1 queue in which the idle periods are deleted and the busy periods are glued together. The process $\mathbf{V}$ is interpreted as the VWT process (or the work process) of the same G/M/1 queue.

The processes $\mathbf{V}$ and $\mathbf{A}$ are dual processes with respect to waiting times. While $V(t)$ is interpreted as the time that a customer would have to wait in line if he arrived at t , $A(t)$ is interpreted as the conditional time (given it is strictly positive) already attained (or elapsed) since the arrival of the customer, who is being served at t . In other words, while $\mathbf{V}$ designates the waiting time of a virtual customer by looking forward in time (the customer is virtual in the sense that he did not arrive at t and thus, in practice, he contributed nothing to the work), $\mathbf{A}$ designates the conditional waiting time of a real customer (given it is strictly positive) by looking backward in time. As a result, while the VWT might sometimes be equal to 0, the conditional AWT cannot be 0 because the served customer “sees” at least himself in the system. By that interpretation, the steady-state law of $\mathbf{A}$ and that of $\mathbf{V}$ must be closely related to each other. In fact, it can be shown by construction [10] that the steady-state law of $\mathbf{A}$ is equal to the conditional steady-state law of $\mathbf{V}$, given that the idle periods (the time periods in which $V = 0$) are deleted and the busy periods are glued together. Note that $\rho > 1$ implies that both $\mathbf{X}$ and $\mathbf{Y}$ tend to $-\infty$ a.s., so that $\mathbf{A}$ and $\mathbf{V}$ are regenerative processes. Furthermore, the cycles associated with $\mathbf{A}$ are the busy periods and those associated with $\mathbf{V}$ are the busy cycles (the busy cycle is composed of busy period plus idle period). Also, it can be shown [10] that the stopping times $T = \inf\{t \geq 0 : X(t) \leq 0\}$ and $\tau = \inf\{t \geq 0 : Y(t) = 0\}$ are the same random variables that represent the busy period of the G/M/1 queue.

Note that a busy cycle generated by $\mathbf{V}$ is $C = \inf\{t \geq \tau : V(t) > 0\}$. Thus, $C - T$ and

$-A(T)$ are the same random variables representing the idle period. Also, while the sample path of $\mathbf{V}$ is continuous at τ and $V(\tau-) = V(\tau+) = 0$, $A(T)$ is a point of discontinuity since, by definition, $A(T-) > 0 > A(T)$.

WAITING TIMES AND SOJOURN TIMES

We first study the AWT process $\mathbf{A}$, showing that its steady-state distribution is exponential. We define the steady-state random variable $A = \lim_{t \rightarrow \infty} A(t)$, where the latter limit is defined in terms of weak convergence. Let $f_A(\cdot)$ be the equilibrium density of A . A level-crossings argument [11] leads to the balance equation

$$f_A(x) = \mu \int_x^\infty [1 - G(w - x)] f_A(w) dw. \quad (14)$$

To better understand the balance equation (14) note that the left-hand side (the steady-state density of $\mathbf{A}$) is interpreted as the long-run average number of up-crossings of level $x > 0$. Accordingly, the right-hand side represents the long-run average number of down-crossings of the same level. However, down-crossings of level x may occur only at moments of negative jumps. Indeed, μ is the rate of the negative jumps and $\int_x^\infty [1 - G(w - x)] f_A(w) dw$ is the probability that an arbitrary jump that starts from level $w > x$ is greater than $w - x$, which means that the latter jump is also a down-crossing of level x [11].

Rewriting Equation (14)

$$f_A(x) = \mu \int_0^\infty [1 - G(y)] f_A(y + x) dy \quad (15)$$

and differentiating, we get

$$f'_A(x) = \mu \int_0^\infty [1 - G(y)] f'_A(y + x) dy, \quad (16)$$

where $f'_A(x)$ is the derivative of $f_A(x)$ with respect to x . We know that, for $\rho > 1$, A has a unique density. Noticing that $f'_A(x)$ and $f_A(x)$ satisfy the same equation, it follows that $f'_A(x)$ equals $f_A(x)$, up to a multiplicative constant.

Solving $f'_A(x) = \eta f_A(x)$ with $\int_0^\infty f_A(x) dx = 1$ yields

$$f_A(x) = \eta e^{-\eta x}, \quad x > 0. \quad (17)$$

Here η is implicitly defined as the unique solution, in $(0, \mu)$, of

$$\eta = \mu[1 - G^*(\eta)]. \quad (18)$$

The fact that η satisfies Equation (18) follows from the substitution of Equation (17) in Equation (14). The uniqueness statement follows since $G^*(0) = 1$, $G^*(\infty) = 0$ and $G^*(\alpha)$ is a monotone decreasing convex function, which combined with $\rho > 1$ implies that the derivative of the right-hand side of Equation (18) is $1/\rho < 1$. We conclude that the steady-state law of the process $\mathbf{A}$, that is, the conditional AWT process of the $G/M/1$ queue in which the idle periods are deleted, is exponentially distributed(η).

The last argument implies that the *peak* points (i.e., points of local maximum) of the conditional $G/M/1$ queue are also exponentially distributed(η) (the latter statement is a well-known result [2]). To see this, note that the peak values of the AWT process are obtained at the arrival instants of the Poisson process $\mathbf{N}$. Hence, by PASTA [6], the limiting distribution of the peak values of the conditional AWT process equals the stationary distribution (17).

From the above, one can conclude that the sojourn times are also exponentially distributed(η). The validity of the latter statement follows from the fact that $\mathbf{A}$ is the conditional AWT process in which the idle periods are deleted. Now, consider a modified version of the AWT process $\mathbf{A}$. In this modification, the idle periods are added back at the end of the busy periods. Practically, we construct the modified version $\bar{\mathbf{A}} = \{\bar{A}(t) : t \geq 0\}$ of $\mathbf{A}$ such that at moments of negative overflows in $\mathbf{A}$, an interval whose size is equal to the negative overflow is added (during that interval the value of the modified process is 0). By this modification, the values of the peak points of $\bar{\mathbf{A}}$ and that of $\mathbf{A}$ are the same; thus, they also have the same steady-state laws. But, in the unconditional AWT process $\bar{\mathbf{A}}$, the peak points can be interpreted as the sojourn times in the $G/M/1$ queue.

Next, let $F_{\bar{A}}(x) = \lim_{t \rightarrow \infty} \Pr(\bar{A}(t) \leq x)$ be the steady-state distribution of $\bar{\mathbf{A}}$. Then,

$F_{\bar{A}}(0)$ is the probability that the $G/M/1$ system is empty. We have from Equation (17)

$$1 - e^{-\eta x} = \frac{F_{\bar{A}}(x) - F_{\bar{A}}(0)}{1 - F_{\bar{A}}(0)}, \quad (19)$$

because the right-hand side of Equation (19) represents the conditional distribution of $\bar{A}$ given $\bar{A} > 0$. To compute $F_{\bar{A}}(0)$ one can apply renewal theory to get

$$F_{\bar{A}}(0) = \frac{EI}{EI + ET},$$

where I designates the idle period and T designates the busy period (the law of I and T is analyzed in the section titled “Busy Period and Idle Period” below).

Finally, to obtain the law of the waiting time, recall that the sojourn time is the sum of the waiting time plus the service, where the latter two random variables are independent. In terms of LST this yields

$$\frac{\eta}{\eta + \alpha} = f_W^*(\alpha) \frac{\mu}{\mu + \alpha},$$

where $f_W^*(\alpha)$ is the LST of the waiting time. Thus,

$$f_W^*(\alpha) = \frac{\eta(\mu + \alpha)}{\mu(\eta + \alpha)}.$$

BUSY PERIOD AND IDLE PERIOD

For an arbitrary distribution G in the $G/M/1$ queue, the idle period I and the busy period T are not necessarily independent random variables. In the special case for which G is an arbitrary phase-type distribution, I and T are conditionally independent given the phase in which the busy period is terminated (I and T are independent random variables only when G is exponential, when the $G/M/1$ system becomes an $M/M/1$ system, and when I is also exponential).

For general G there are several methodologies for computing the joint law of I and T (different approaches can be found in text books [1–4]). We adopt an approach (based on a sequence of recent works [7,10,12–15] and use the notation as defined in the section titled “Work and Attained Waiting Time.”

In other words, the continuous-time random walk $\mathbf{X}$ is composed of a drift of rate 1 minus a compound Poisson process with rate μ and iid jumps having distribution G . Define the stopping time $T = \inf\{t : X(t) \leq 0\}$. Note that $X(T) < 0$. The random vector $(T, -X(T))$ can be interpreted as the busy period and the idle period, respectively, of the $G/M/1$ queue (note again that T and $X(T)$ are not necessarily independent, unless the jumps are exponential).

To compute the joint law of $(T, -X(T))$ in terms of LSTs, we introduce the functionals $k(x, \alpha_1, \alpha_2)$ and $K(s, \alpha_1, \alpha_2)$ defined by $Re(\alpha_1, \alpha_2) \geq 0, x \geq 0$

$$k(x, \alpha_1, \alpha_2) = E\left(e^{-\alpha_1 T - \alpha_2 I} \mid Z_1 = x\right), \quad (20)$$

and

$$K(s, \alpha_1, \alpha_2) = \int_0^\infty e^{-sx} k(x, \alpha_1, \alpha_2) dx. \quad (21)$$

Also, let $\hat{s} = \hat{s}(\alpha_1)$ be the unique zero of $1 - \frac{\mu}{\mu - s} G^*(\alpha_1 + s)$ in the right-half α_1 -plane [2, p. 226]; such that

$$1 - \frac{\mu}{\mu - \hat{s}} G^*(\alpha_1 + \hat{s}) = 0 \quad (22)$$

Then, by the law of total probability and conditioning successively on the events $\{S_1 \geq x\}$ and $\{S_1 < x\}$ we can write

$$\begin{aligned} k(x, \alpha_1, \alpha_2) &= \int_{t=x}^\infty e^{-\alpha_1 t} e^{-\alpha_2(t-x)} dG(t) \\ &\quad + \int_{z=0}^\infty \mu e^{-\mu z} \int_{t=0}^x e^{-\alpha_1 t} \\ &\quad \times k(x-t+z, \alpha_1, \alpha_2) dG(t) dz. \end{aligned} \quad (23)$$

Taking the Laplace transform with respect to x and changing the order of integration yields

$$\begin{aligned} K(s, \alpha_1, \alpha_2) &= \int_0^\infty e^{-(\alpha_1+s)x} \int_0^\infty e^{-\alpha_2 u} dG(x+u) dx \\ &= \int_0^\infty e^{-(\alpha_1+s)x} \int_0^\infty e^{-\alpha_2 u} dG(x+u) dx \end{aligned}$$

$$\begin{aligned} &+ \mu \int_0^\infty e^{-(\alpha_1+s)t} dG(t) \int_0^x e^{-su} \\ &\times \int_0^x e^{-\mu z} k(u+z, \alpha_1, \alpha_2) dz du \\ &= \frac{G^*(\alpha_1+s) - G^*(\alpha_2)}{\alpha_2 - \alpha_1 - s} + \mu G^*(\alpha_1+s) \\ &\times \frac{K(s, \alpha_1, \alpha_2) - K(\mu, \alpha_1, \alpha_2)}{\mu - s}, \end{aligned} \quad (24)$$

so that

$$\begin{aligned} K(s, \alpha_1, \alpha_2) &\left[1 - \frac{\mu}{\mu - s} G^*(\alpha_1 + s)\right] \\ &= \frac{G^*(\alpha_1 + s) - G^*(\alpha_2)}{\alpha_2 - \alpha_1 - s} \\ &\quad - \frac{\mu}{\mu - s} G^*(\alpha_1 + s) K(\mu, \alpha_1, \alpha_2). \end{aligned} \quad (25)$$

Substituting $s = \hat{s}$ in Equation (25), we get

$$K(\mu, \alpha_1, \alpha_2) = \frac{G^*(\alpha_1 + \hat{s}) - G^*(\alpha_2)}{\alpha_2 - \alpha_1 - \hat{s}} \quad (26)$$

and by definition $\mu K(\mu, \alpha_1, \alpha_2)$ is the joint LST of the busy period and idle period. Also, by using $\alpha_1 = \alpha_2 = \alpha$ in Equation (26) we get the LST of the busy cycle $C = I + T$. Note also that

$$\begin{aligned} K(s, \alpha_1, \alpha_2) &= \frac{\mu - s}{\mu - s - \mu G^*(\alpha_1 + s)} \left[\frac{G^*(\alpha_1 + s) - G^*(\alpha_2)}{\alpha_2 - \alpha_1 - s} \right. \\ &\quad \left. - \frac{\mu}{\mu - s} G^*(\alpha_1 + s) \frac{G^*(\alpha_1 + \hat{s}) - G^*(\alpha_2)}{\alpha_2 - \alpha_1 - \hat{s}} \right]. \end{aligned} \quad (27)$$

As a by-product of the functional introduced in Equation (27) one can also introduce the joint asymptotic law of the AWT and the time elapsed since the beginning of the busy period via a martingale approach. To see this, consider the process $\mathbf{M} = \{M(t) : t \geq 0\}$, where

$$\begin{aligned} M(t) &= [\varphi(\alpha_2) - \alpha_1] \int_0^t e^{-\alpha_1 s - \alpha_2 X(s)} ds \\ &\quad + 1 - e^{\alpha_1 t - \alpha_2 X(t)}, \end{aligned} \quad (28)$$

and $\varphi(\alpha_2) := \alpha_2 + \mu[1 - G^*(-\alpha_2)]$ is the exponent of $X(t)$. Then, $\mathbf{M}$ is a special case of the *Kella–Whitt* martingale [16,17]. By applying the optional sampling theorem with the stopping time T to the martingale $\mathbf{M}$ we see that $E(M(T)) = 0$, thus obtaining the fundamental identity (with the substitution $X(0) = 0$)

$$\begin{aligned} [\varphi(\alpha_2) - \alpha_1]E \int_0^T e^{-\alpha_1 s - \alpha_2 X(s)} ds \\ = -1 + Ee^{-\alpha_1 T - \alpha_2 X(T)} \\ = -1 + \mu K(\mu, \alpha_1, -\alpha_2), \end{aligned} \quad (29)$$

where the second equality in Equation (29) follows since by Equation (20)

$$E(e^{-\alpha_1 T + \alpha_2 I}) = \mu K(\mu, \alpha_1, -\alpha_2).$$

Now define the random variable L as the asymptotic age of the busy period in the $G/M/1$ queue. That is, L is the time elapsed since the beginning of the busy period in steady state; in other words, L is the asymptotic age of the busy period. Accordingly, by the theory of regenerative processes [17] it follows that

$$\frac{E \int_0^T e^{-\alpha_1 s - \alpha_2 X(s)} ds}{ET} = Ee^{-\alpha_1 L - \alpha_2 A}. \quad (30)$$

Also, by *level crossing theory* [11], level 0 is down-crossed by the AWT only once during a cycle—at the end of the busy period. Thus, by Equation (17)

$$ET = 1/f_A(0) = 1/\eta. \quad (31)$$

Substituting Equations (30) and (31) into Equation (29) we get

$$Ee^{-\alpha_1 L - \alpha_2 A} = \frac{\eta[-1 + \mu K(\mu, \alpha_1, -\alpha_2)]}{\alpha_2 + \mu[1 - G^*(-\alpha_2)] - \alpha_1}.$$

SUMMARY

The $G/G/1$ queueing system is the prototype general single server queueing model. However, for the general $G/G/1$ queue, most of the

relevant measures and functionals cannot be expressed in closed form formulas. Thus, the two dual special cases—the $M/G/1$ queue and the $G/M/1$ queue—represent the pillar models. In practice, the $M/G/1$ model is more natural than the $G/M/1$ model since the memoryless assumption of the service in the latter model is unnatural. However, one can interpret the structure of the AWT process as the content level of the cash fund of insurance risk models with a constant input of premium, and compound Poisson arrivals of claims with general claim sizes. Then, the time of ruin and the deficit in the insurance risk model are interpreted as the busy period and the idle period in the $G/M/1$ model. One could think of many ramifications of the basic $G/M/1$ model. For example, under the insurance risk interpretation it is natural to model the content level of cash funds in which the input of premiums is state dependent (a special case of the latter assumption is the finite capacity case, see Löpker and Perry [18] and Perry *et al.* [19] for some variants). Other examples are those of the models with vacations, models with Abandonments, and models with priority disciplines. In this study, only the main interest factors were reviewed, such as the number of customers in steady state and the waiting time, the sojourn time, the busy period, and the idle period. The analysis was based on the duality between the $M/G/1$ model and the $G/M/1$ model.

REFERENCES

1. Asmussen S. Applied probability and queues. 2nd ed. New York: Springer; 2003.
2. Cohen JW. The single server queue. Amsterdam: North-Holland Publishing Co.; 1982.
3. Kleinrock L. Volume 1. Queueing systems. New York: Wiley; 1975.
4. Prabhu NU. Queues and inventories. New York: Wiley; 1965.
5. Tákacs L. Introduction to the theory of queues. New York: Oxford University Press; 1962.
6. Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice-Hall; 1989.
7. Adan I, Boxma OJ, Perry D. The $G/M/1$ queue revisited. Math Method Oper Res 2005;58(2): 437–452.

8. Çinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice-Hall; 1975.
9. Melamed B, Yao D. The ASTA property. In: Dshalalow JH, editor. Advances in queueing: theory, methods and open problems. Boca Raton (FL): CRC Press; 1995. pp. 195–224.
10. Perry D, Stadje W, Zacks S. Sporadic and continuous clearing policies for a production/inventory system under $M/G/$ demand. Math Oper Res 2005;30(2):354–368.
11. Doshi BT. Level crossing analysis of queues. In: Bhat UN, Basawa IV, editors. Queueing and related models. Oxford: Clarendon Press; 1992. pp. 3–33.
12. Zacks S, Perry D, Bshouty D, Bar-Ler SK. Distributions of stopping times for compound poisson processes with positive jumps and linear boundaries. Stoch Model 1999;15:89–101.
13. Perry D, Stadje W, Zacks S. Some contribution to the theory of first-exit times of some compound processes in queueing theory. Queueing Syst 1999;33:369–379.
14. Perry D, Stadje W, Zacks S. Busy period analysis for $M/G/1$ and $G/M/1$ -type queues with restricted accessibility. Oper Res Lett 2000;27(4):163–174.
15. Perry D. Stopping problems for compound processes with applications to queues. J Stat Plan Infer 2000;91:65–75.
16. Kella O, Whitt W. Useful martingales for stochastic storage processes with Levy input. J Appl Probab 1992;29:396–403.
17. Cohen JW. On regenerative processes in queueing theory. Berlin: Springer; 1976.
18. Löpker A, Perry D. The idle period of a finite $G/M/1$ queue with an interpretation in risk theory. Queueing Syst 2010;64(4):395–407.
19. Perry D, Stadje W, Zacks S. A duality approach to queues with service restrictions and storage systems with state-dependent rates. 2010. Submitted for publication.

THE GLOBAL REPLENISHMENT PROBLEM

JAMES R. BRADLEY
HÉCTOR H. GUERRERO
Mason School of Business, College of
William and Mary, Williamsburg,
Virginia

INTRODUCTION

In recent years, manufacturers and retailers have increasingly sourced their raw materials and their retail goods overseas. This has led to the extensive use of intermodal transport to deliver globally sourced goods through the supply chain and to the eventual consumer. The value of imports into the United States grew at approximately an 8.4% compounded annual growth rate from 1990 through 2008 [1]. Even Wal-Mart, the nation's largest retailer, which in the 1980s promoted goods "Made in America," has relied on imported goods, importing \$12 billion of goods from China in 2002 [2]. While the importance of global replenishment has grown, long transportation lead times and the demand uncertainty over replenishment lead times have made managing global replenishment perhaps more difficult than domestic replenishment. Fortunately, global supply chains can be structured so that they have the flexibility to adjust replenishment quantities and destinations while goods are in transit in reaction to the revelation of demand information during long replenishment cycles. Macy's, for example, relies on such flexibility with its shipments from Asia [3], which also requires appropriate information technology and decision support systems. This flexibility to postpone the final replenishment decision is of value not only in retail distribution but also to manufacturers who might reapportion replenishment quantities among multiple factories depending on demand and production during transit. This

article discusses the replenishment decisions in flexible supply chains where the original replenishment decision can be revisited, and we place the analysis of those decisions in context with existing research, and we also suggest future research directions.

The flexibility discussed above stems from the many ways that a shipment from overseas might be routed once it arrives in the United States as shown in Figure 1:

1. Direct transit from the port to a final destination, which might be a factory for a manufacturer or a regional distribution center (RDC) or retail store for a retailer
2. To an import distribution center (IDC), which is a facility where goods not immediately needed downstream are stored [4,5]
3. To a crossdocking facility where the contents of containers can be deconsolidated and mixed in outgoing containers or trailers for immediate transfer downstream.

Thus, by providing information systems that continually monitor demand and decision models that can provide a better updated distribution solution, the original shipment decision can be revisited during transit to either (i) change the destination of whole containers, or (ii) resort the contents of containers at a crossdocking facility. A container's destination might be changed, for example, from an IDC to a crossdocking facility if a surge in retail demand causes some or all of the goods to be needed immediately downstream in the supply chain. Resorting the contents of containers at a crossdocking facility allows a response to more recent demand and provides for more accurate replenishment. IDCs might be operated by a company doing its own distribution or by a third party logistics provider (3PL). Crossdocking facilities are often operated by 3PLs such as Maersk Distribution [6] or NYK Logistics [7].

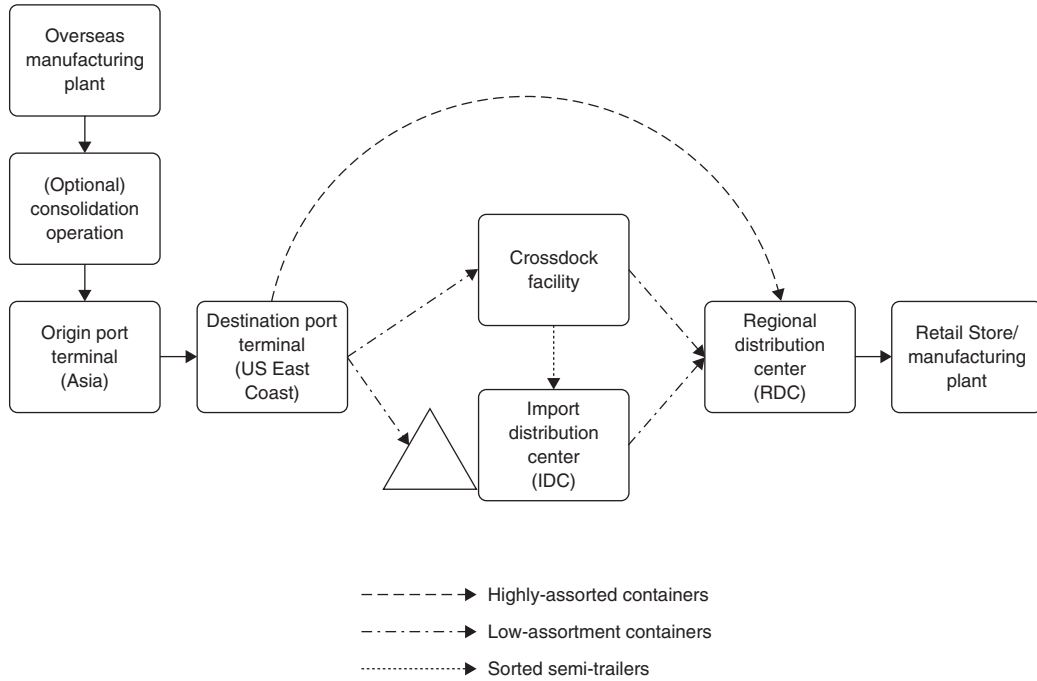


Figure 1. Alternative global supply chain routes.

One might consider the problem of determining the optimal replenishment decisions in the short term as (stochastic) demand varies in an otherwise static business environment. However, the benefits resulting from this flexibility also extend to situations where business environment parameters change. For example, when diesel fuel prices surged above \$4 per gallon in 2008, logistics managers planned changes in their supply chains [8], including (i) transitioning from just-in-time delivery to full truckloads, (ii) a renewed emphasis on increased packing density, (iii) reduced stock keeping unit (SKU) counts, and (iv) transitioning from trucking to less costly shipping modes such as rail. While we might not expect that exact same shock to the business environment to reoccur, other changes in the parameters under which replenishment decisions are made undoubtedly will occur and flexibility in supply chain operations will be rewarded in those circumstances. For example, recent increases in container rental fees due to shipping lines' efforts to increase revenue have

greatly increased the cost of using containers [9]. The cost of using containers can also change significantly for particular origins and destinations as container owners realign container charges and shipping fees in order to rebalance container flow throughout the world. The amount of time that a shipper can maintain possession of a container before additional charges accrue also changes over time. Besides the costs of shipping, fuel, and containers, the costs of labor, exchange rates, and inventory vary over time, each of which might motivate a change in supply chain flows that would reduce cost. Flexibility makes it possible to shift operation modes quickly in reaction to the changing business environment, for example, by increasing the use of crossdocking facilities to transload goods from containers into semitrailers for logistics efficiency when container fees increases.

We proceed in the next section (titled "A General Description of Decisions in the Flexible Global Supply Chain") by generally describing the replenishment decisions in

a global supply chain. In the section titled “Formulation for Replenishment Decisions in Flexible Global Supply Chains,” we present a formulation for those replenishment decisions and relate it to the existing literature. We conclude in the section titled “Conclusions.”

A GENERAL DESCRIPTION OF DECISIONS IN THE FLEXIBLE GLOBAL SUPPLY CHAIN

For the remainder of this article, we discuss the problem in the context of a large retail distribution supply chain, which is likely to be a more general structure than that for a manufacturer. The manufacturer’s network generally has fewer echelons because no link is needed between the IDC/crossdock echelon of the supply chain and the factories (which take the place of retail stores in Fig. 1). This is because a large retailer, for example, Wal-Mart, might have thousands of stores, whereas even large manufacturers might only have tens of plants. Our discussion, therefore, is without loss of generalization because the manufacturer’s problem has a simpler supply chain and, otherwise, is likely to be of smaller dimensions (e.g., number of goods and number of destinations) as well. With Wal-Mart or Target in mind, our discussion about the global supply chain in Fig. 1 references a US port throughout the article, although our discussion is relevant for import supply chains into any country.

Global supply chains due to necessity use intermodal transportation where one link is ocean transport: the only other intercontinental transportation alternative, air freight, is too expensive for most goods. Moreover, the types of goods ordered by the manufacturers and retailers that we consider in this article would be shipped in containers. In other words, this article is not concerned with bulk shipments of commodities such as coal, iron ore, or wood products. Containerized freight movements are indeed significant and constituted 70% of global trade in 2000 [10]. The percentage of freight traveling via container continues to increase at an estimated 8.7% annual growth rate [11]. Although our focus

is on globally sourced supply, the basic decisions we discuss could also be faced by domestically based supply chains that are far-flung. Thus, our concern is centered on goods that move through some intermodal network—ocean carriers via containers to truck via trailers and possibly by rail.

The presence of an IDC and/or crossdock offers great flexibility in routing containers from the receiving port (ultimately) to the RDCs, which serve many retail locations (see Fig. 1). A highly assorted container, with many SKUs and relatively small quantities of each SKU, might be shipped directly to an RDC, thus bypassing the IDC and crossdock and avoiding the associated costs. These containers might require an overseas consolidation step if a particular manufacturer does not manufacture a sufficiently wide variety of goods. Small delivery quantities can alternatively be created at a US crossdocking facility using goods from containers with larger quantities of fewer SKUs. Some containers with fewer SKUs and greater quantities of each SKU might also go directly to the IDC if they are not immediately needed downstream, and the IDC might also receive goods from the crossdocking facility if in reconsolidating a container some goods are not needed immediately downstream at an RDC. These options create flexibility because the destinations of the goods in ocean-bound containers with fewer SKUs and larger lot quantities can be postponed until a short time before landfall [3], at which time they can be directed toward the location, either a crossdocking facility or an IDC, where they would offer the greatest advantage. Taking best advantage of this flexibility requires that an appropriate assortment of two types of loads be created overseas, with some loads that are highly assorted and destined for RDCs and some containers with larger batch sizes destined for the IDC or a crossdocking facility. Creating that assortment can be broken down into three decision components as follows:

1. How many containers to ship.
2. Which SKUs and how many of each SKU to load in each container.
3. The intended destination of each container.

Making the best decision requires that the economics of each route be considered along with the uncertainty of demand. Creating an assorted load overseas avoids charges that accrue at the US IDC and crossdocking facility. Additionally, the lesser quantities of the many SKUs in an assorted container would quickly flow through an RDC, thus minimizing inventory cost, as well as helping to avoid having the problem of shipping great quantities to one RDC, which are later found to be required at another RDC. However, these savings are in some cases offset to some degree by a cost to consolidate the load overseas before shipping. We might expect that overseas consolidation would be less costly, however, than consolidation in an US crossdocking facility: labor in many countries is less than in the United States. A more subtle cost might be imposed on creating a more highly assorted load overseas if the mix of SKUs and their quantities on that load destined for an RDC no longer proved to be a wise choice once the container arrives in the United States. The difficulty of an overseas determination of the appropriate mixes of assorted shipments for RDCs due to uncertain demand and long transit times motivates a mix of (i) more highly assorted containers and (ii) other loads with greater quantities of fewer SKUs that could be used to customize loads for RDCs with a reduced lead time from the point of customization. This is called *postponed replenishment*. Of course, greater cost is incurred for handling the loads with a lesser assortments at the crossdocking facility because these goods stay in the pipeline longer (particularly those stored at the IDC), thus accruing labor and inventory holding cost.

Given that the original replenishment decision can be revisited at some point during the ocean voyage, one would ideally postpone the final decision regarding the routing and handling of the containers until they arrive at the port. Some lead time is required, however, to arrange for the transportation of the containers and for planning staffing requirements at crossdocking facilities, which is called the *logistics planning and acquisition lead time* (L_2). The total transportation lead time is $L = L_1 + L_2$, where, given L and

L_2 , L_1 is how long the final replenishment decision can be delayed, which is called the *decision postponement lead time*. See Table 1 for a breakdown of events that occur within these two lead time components. Note that the lead time to the IDC, RDCs, and retail stores on the landside of the supply chain is in addition to L and it varies depending on which facilities the goods pass through en route to the destination. The lead time L is determined by the schedule for a *trade lane* for a shipping line. Shipping lines offer scheduled service at posted origins and destinations. For example, Maersk Line offers a trade lane that takes two weeks to travel from Rotterdam to Norfolk, VA and stops in Bremerhaven, Felixstowe, and Newark before arriving in Norfolk. Multiple ships follow this sequence of ports so that a ship arrives in Norfolk every seven days. Thus, in this example, $L = 14$ days and two container ships are in the pipeline at any one point in time.

Replenishment decisions for goods sourced overseas can be either in the context of a one-time purchase or the continuous replenishment of goods stocked on retail store shelves. One-time purchases that retailers do not intend to continuously stock are made, for example, in the case of “dollar stores” for over-run or close-out goods [12], and also by traditional retailers such as Target and Walmart for seasonal goods (e.g., lawn furniture in spring and holiday decorations in Winter). Retailers, large and small, also carry items on a continuous basis for a long period: these items require ongoing replenishment orders to keep the retail shelves stocked. Manufacturers also require continuous supply of offshored raw materials for products that are produced on a sustained basis. Intermediate modes of procurement between these two end points on the continuum are also possible where, for example, a small number of replenishment orders might be placed within a selling season such as in the case with Sport Obermeyer [13]. Another such intermediate example is the case of a product for which initial demand will spike, such as a Harry Potter novel. The large initial purchase might be followed up with many replenishment orders over a finite time horizon. The formulation

Table 1. Global Transport Lead Time Components

Lead Time Components	Events
Decision postponement lead time L_1	Loading of goods into the container
	Transporting container to port terminal (truck/rail)
	Handling/waiting at origin port and loading container onto ship
	Initial portion of ocean voyage
Logistics planning and acquisition lead time L_2	Remainder of ocean voyage after recourse decision
	Unloading container from vessel
	Handling/waiting at port for pickup

that we present in the next section can be used in both contexts.

The maximum size of this problem might be fairly estimated by the parameters of Wal-Mart's supply chain and, in particular, the portion of their supply chain used for nonperishable goods. Wal-Mart supercenters stock on the order of 100,000 SKUs (<http://www.retailingworks.com/why.htm>), perhaps 30–50% of which are nonperishable and flow through their IDC–RDC supply chain echelons to retail stores. Wal-Mart manages much of its own logistics including 147 distribution centers, 7,200 tractors, and 53,000 semi-trailers (<http://walmartstores.com/download/2336.pdf>). Each RDC serves between 75 and 100 retail stores within a 250-mile radius. It is reported that \$12 billion of Wal-Mart's goods in 2002 were sourced in China, and so its imports as measured in containers are considerable given that its China-based import department sources in 18 other countries as well (<http://walmartstores.com/download/3719.pdf>).

FORMULATION FOR REPLENISHMENT DECISIONS IN FLEXIBLE GLOBAL SUPPLY CHAINS

We now present the economic parameters for global replenishment decisions, followed by a formulation for the problem. We then consider various simplifying assumptions correlating the resulting problems to the existing literature, and we also comment on the advantages and disadvantages of the

simplified problems in terms of providing managerial insight.

The Economics of Global Shipping

In this section, we describe the relevant costs for replenishment decisions in a global supply chain, as summarized in Table 2. Costs accrue due to many different bases, including per container, per unit of time, per order line item, per shipment value, and so on, at the different supply chain links. The data in Table 2 is based on the assumption that the “shipper” owns the IDC and RDC links of the distribution supply chain such as would be the case with national retail chains (e.g., Wal-Mart and Target). We also assume, as is often the case, that the crossdocking facility is operated by a 3PL provider such as Maersk Distribution or NYK Logistics. The formulation that we present in the following section takes into account the relevant costs from Table 2 that would affect both the overseas and postponed replenishment decisions.

Formulation for the Global Replenishment Problem

In this section, we construct a formulation for replenishment of the RDCs for goods that are received from one port of call on one shipping lane, which is a kernel of a potentially larger problem. We discuss the effect of expanding the scope of the problem, and also discuss relaxing the simplifying assumptions that we make below. We also address extensions of this problem that consider various important features of real-world supply chains.

Table 2. Summary of Global Supply Chain Costs and Bases of Charges

<i>Supply Chain Links</i>	
Facility/Link Type	Relevant Charges
Port marine terminal	• Per instance charges – Per container lift – Per move • Per day charges – Demurrage for storage past free-time allowance
Consolidator (third party)	• Per container charge
Crossdock/transload (third party operation)	• Per container charge for transloading
International DC	• Per value of inventory, per time – Holding cost • Fixed cost – Building and assets – Some portion of labor • Variable handling cost – Some portion of labor
Regional DC	• Per value of inventory, per time – Holding cost • Fixed cost – Building and assets – Some portion of labor • Variable, per-container/trailer handling cost – Some portion of labor
<i>Transportation Links</i>	
Ocean liner	• Per container – Transport fee • Per day charges – Container demurrage past free-time
Trucking (dray move, container)	• Per container, per trip charge
Trucking (long haul, semitrailer)	• Per load, per trip charge based on distance, weight, and other factors
Trucking (LTL)	• By weight and distance
Railroad	• By weight and distance
Barge	• Per load, per trip
<i>Logistics/Supply Chain Facilitators</i>	
Freight forwarder	• Per transaction charges • Per container charges • Per item charges – Customs tariff code classification
Insurance	• Per product value in container

Table 3. Decision Epochs

Decision	Decision Epochs
Original replenishment decision	$t = 0$
Postponed replenishment decision	$t = L_1$
Allocation of IDC inventory to RDCs	$t = 1, \dots, T$

Our simplifying assumptions begin with the assumption that the transportation lead time $L = T$ days, and thus a ship arrives every T days such that containers on only one ship are in transit at any time. Within the T -day horizon, three different types of decisions are made as shown in Table 3 where we have indexed time (days) using t . These decision epochs would be repeated for goods that are continuously replenished, whereas a single problem with a horizon of T days suffices for a one-time buy. We assume that ships arrive according to schedule (deterministic arrivals) and that lead times from ports to the IDC and RDCs, and lead times from the IDC to RDCs, are deterministic. Also, without loss of generality given the foregoing assumptions, we assume zero lead time from the IDC to each RDC, and from the port to the IDC and all RDCs. Including nonzero lead time in the formulation is straightforward.

A set of sequential problems that addresses the relevant decisions for each period are summarized in Table 4; formulations for these problems are shown below

using the notation in Table 5. Using a backward solution procedure, Θ_1 would be solved, followed by Θ_2 , and finally Θ_3 . The problems are shown as parameters that (i) indicate the current inventory state of the system, and (ii) solutions from prior problems that bear on the solution of the current problem.

In the initial formulation, we consider the stochastic demand, where $D_{p,r,t}$ is a random variable of demand for product p at RDC r in period (day) t . (See Table 5 for notation.) Later, we consider the problem simplified by assuming the deterministic demand. Our formulation considers only one IDC, although it is simple to formulate a problem with multiple IDCs. We represent shipping containers and truck trailers as having only a volume constraint, which we represent as a volume V for containers and V' for loads leaving the crossdocking facility. The specification of two volumes offers the opportunity to use trailers on loads leaving the crossdocks for RDCs. The formulation is easily extended if multiple types of loads with multiple volumes are used at the crossdock. Extending this formulation to accommodate a weight restriction is easy, although this impacts the computational requirements, which we comment on later. We also ignore determining detailed packing plans for containers, which would greatly complicate the solution.

Our formulation accommodates the varying logistics costs of different routes that goods might take. First, the formulation

Table 4. Sequential Problems

Problem	Periods	Decisions Considered
Θ_1	$t = L_1 + 1, \dots, T$	Daily shipments from IDC to RDCs
Θ_2	$t = L_1$	- Daily shipment from IDC to RDCs - Postponed replenishment decision
Θ_3	$t = 0, \dots, L_1 - 1$	- Daily shipments from IDC to RDCs in period $t = 1$ on - Original replenishment decision in period $t = 0$

Table 5. Notation*Economic Parameters*

h_p	Holding cost rate for product p at the IDC
h'_p	Holding cost rate for product p at and RDC
b'_p	Backorder cost of product p in period t
c_r	Cost of shipping a load from the IDC to RDC r
c_I	Cost of logistics route where a container is shipped from the port to the IDC
c_C	Cost of shipping a container to the crossdock and deconsolidation (handling/labor)
c'_r	Cost of shipping a container directly to RDC r
c_{CI}	Cost of shipping route where a load is shipped from the crossdock to the IDC
c_r^{PR}	Cost of shipping a load from the crossdock to RDC r

Demand

$D_{p,r,t}$	Random variable of demand for product p at RDC r in period t
-------------	--

Inventory Variables

$I_{p,t}$	Inventory of product p in period t at the IDC
$I_{p,r,t}^+$	(Positive) on-hand inventory of product p in period t at RDC r
$B_{p,r,t}$	Backorders of product p in period t at RDC r

Decision Variables

S_r^{IR}	Cost of shipping a load from the IDC to RDC r
$s_{r,u,t}$	1 if load u is used from the IDC to RDC r in period t
S^I	The number of loads shipped directly from the port to the IDC
s_w^I	1 if load w is shipped directly from the port to the IDC, and 0 otherwise
S_r^R	The number of containers shipped directly to RDC r in period L_1
$s_{r,w}^R$	1 if container w is shipped directly to RDC r , and 0 otherwise
S^C	The number of containers crossdocked in period L_1
s_w^C	1 if container w is crossdocked, and 0 otherwise
S_r^{PR}	The number of loads shipped from the crossdock to the RDC r in period L_1
S^{CI}	The number of loads shipped from the crossdock to the IDC in period L_1
$s_{r,y}$	1 if (outgoing) crossdocked load y is shipped to RDC r , and 0 otherwise
s_y	1 if (outgoing) crossdocked load y is shipped to the IDC
$x_{p,r,t,u}$	1 if product p is put on load u from the IDC to RDC r in period t
$x_{p,r,y}$	1 if product p is put on (outgoing) crossdocked y for RDC r
$x_{p,y}$	1 if product p is put on (outgoing) crossdocked y for the IDC
$q_{p,r,t}$	Quantity of product p shipped from RDC r in period t
Q'_p	Quantity of product p crossdocked to IDC in period L_1
$Q'_{p,r}$	Quantity of product p crossdocked to RDC r in period L_1

Other Parameters

v_p	Volume of each unit of product p
P	The number of products (SKUs) shipped from overseas
R	The number of RDCs
U	The available number of trucks/loads from the IDC to the RDCs
Y	The available number of trailers/loads from the crossdock to the IDC and the RDCs
W	Maximum number of containers to ship from overseas
V	Volume of a shipping container
V'	Volume of a container or a trailer for outgoing loads form the crossdocking facility
M	Sufficiently large value that exceeds the maximum units in any container or load

allows, if required, different holding costs at the IDC and RDC, which are h_p and h'_p respectively. The period backorder cost at the RDCs is b'_p per unit. Different logistics costs accrue depending on whether a container proceeds directly from the port to a particular RDC r (c'_r), or to the IDC (c_I), or to the crossdocking facility (c_C). The container cost c_C includes transportation and handling associated with deconsolidating at the crossdock. The outgoing costs at the crossdock are c_r^{PR} for consolidating a load for RDC r and shipping it, whereas c_{CI} is the cost of creating a load for the IDC and shipping it. The cost to create a load at the IDC and ship it to a particular RDC r is c_r . The handling cost of putting goods away at the IDC and then retrieving them for later shipping is consolidated in the aforementioned parameters c_I and c_{CI} : two such parameters allow accounting for differing effort levels that may be required. For example, a 40' container (from the port) would take less labor to unload than would a 53' trailer (from the crossdock). Any applicable overseas consolidation fee, as we discussed earlier, could be included in the parameters c'_r for each RDC r .

The lower case binary decision variables s_i are 1 if a particular container or trailer is used to transport a load on a particular origin–destination pair, and zero otherwise. The S_i upper case integer parameters accumulate the number of loads employed between a particular origin–destination pair. The formulation can be rewritten to avoid these latter parameters, but we include them for greater clarity.

Lastly, the decision variable $q_{p,r,t}$ is the shipping quantity of product p from the IDC

to RDC r in period t . The upper case $Q_{p,w}$ indicates quantities of each product p in container w , whereas Q'_p and $Q'_{p,r}$ indicate shipping quantities of product p to the IDC and RDC r , respectively, for goods after they have been crossdocked.

Problem Θ_1 determines the optimal shipping quantities of each product p from the IDC to each RDC r in time periods $t = L_1 + 1, \dots, T$ ($q_{p,r,t}$) in order to minimize the holding cost at both the IDC and RDCs, as well as backorder costs at the RDC. (Backorder costs are charged only at the RDC, which is the closest echelon to the final customers where inventory shortages have an immediate monetary effect in terms of lost sales.) Constraints (1) and (2) are inventory balance equations for the IDC in consideration of the quantity Q'_p of product p that arrives at the IDC in period T . Constraints (3) and (4) are inventory balance equations for each RDC r in consideration of the shipment $Q'_{p,r}$ that each RDC r receives in period T . Constraint (5) accounts for (positive) on-hand inventory, while constraint (6) accounts for backorders at the RDCs. Constraint (7) requires that the total shipping quantity of a product should not be greater than its on-hand inventory. Constraints (8) through (13) are a multiple knapsack problem that allocate goods among the U available trucks. Constraint (11) requires that the entire quantity of each product p shipped to RDC r be placed on the same load u , thus simplifying the form of constraint (8), which keeps the contents of each truck within the volume capacity V' .

$$\Theta_1(\{Q'_p\}_{p \in P}, \{Q'_{p,r}\}_{p \in P, r \in R}, \{I_{p,L_1}\}_{p \in P}, \{I_{p,r,L_1}\}_{p \in P, r \in R})$$

$$= \min \mathbf{E} \left\{ \sum_{t=L_1+1}^T \left\{ \sum_{p \in P} \left[h_p I_{p,t} + \sum_{r \in R} (h'_p I_{p,r,t}^+ + b'_p B_{p,r,t}) \right] + \sum_{r \in R} c_r S_r^{IR} \right\} \right\}$$

s.t.

$$I_{p,t} = I_{p,t-1} - \sum_{r \in R} q_{p,r,t} \quad p \in P, t = L_1 + 1, \dots, T - 1 \quad (1)$$

$$I_{p,T} = I_{p,T-1} + Q'_p - \sum_{r \in R} q_{p,r,T} \quad p \in P \quad (2)$$

$$I_{p,r,t} = I_{p,r,t-1} - D_{p,r,t} \quad p \in P; r \in R; t = L_1 + 1, \dots, T - 1 \quad (3)$$

$$I_{p,r,T} = I_{p,r,T-1} + Q'_{p,r} - D_{p,r,T} \quad p \in P; r \in R \quad (4)$$

$$I_{p,r,t}^+ \geq I_{p,r,t} \quad p \in P; r \in R; t = L_1 + 1, \dots, T \quad (5)$$

$$B_{p,r,t} \geq -I_{p,r,t} \quad p \in P; r \in R; t = L_1 + 1, \dots, T \quad (6)$$

$$\sum_{r \in R} q_{p,r,t} \leq I_{p,t-1} \quad p \in P; t = L_1 + 1, \dots, T \quad (7)$$

$$\sum_{p \in P} v_p q_{p,r,t} x_{p,r,t,u} \leq V' \quad r \in R; t = L_1 + 1, \dots, T; u = 1, \dots, U \quad (8)$$

$$s_{r,u,t} \geq \frac{1}{M} \sum_{p \in P} q_{p,r,t} x_{p,r,t,u} \quad r \in R; u = 1, \dots, U; t = L_1 + 1, \dots, T \quad (9)$$

$$S_{r,t}^{IR} = \sum_{u=1}^U s_{r,u,t} \quad r \in R; t = L_1 + 1, \dots, T \quad (10)$$

$$\sum_{u=1}^U x_{p,r,t,u} = 1 \quad p \in P; r \in R; t = L_1, \dots, T \quad (11)$$

$$x_{p,r,t,u} \in \{0, 1\} \quad p \in P; r \in R; u = 1, \dots, U; t = L_1, \dots, T - 1 \quad (12)$$

$$s_{r,u,t} \in \{0, 1\} \quad r \in R; u = 1, \dots, U; t = L_1, \dots, T - 1 \quad (13)$$

$$Q'_p \in \{0, 1, 2, \dots\} \quad p \in P \quad (14)$$

$$Q'_{p,r} \in \{0, 1, 2, \dots\} \quad p \in P; r \in R. \quad (15)$$

Problem Θ_2 determines the final replenishment decisions and, hence, the objective function accounts for logistics costs. The objective function also includes the cost of the subsequent Problem Θ_1 and the holding and backorder costs in period L_1 , analogous with Problem Θ_1 .

The binary decision variables s_w^I , s_w^C , and $s_{r,w}^R$, when set to 1, indicate that a particular container w will be sent directly to the IDC, crossdocked, or sent directly to RDC r , respectively. Constraint (3) ensures that each container is routed only once. Constraints

4–6 accumulate the number of containers that are sent to the IDC, crossdocked, and to each RDC r , respectively, in the variables S^I , S^C , and S_r^R , which are used in the objective function to account for the various logistics costs. Constraints (7) through (18) constitute a multiple knapsack problem, similar to that in Problem Θ_1 , except now the quantities of each product p that are crossdocked and reloaded into another container or trailer are limited by the quantity of incoming goods designated for crossdocking, as reflected in constraint (7).

$$\begin{aligned} & \Theta_2 \left(W, \{Q_{p,w}\}_{p \in P, w \in W}, \{I_{p,L_1-1}\}_{p \in P}, \{I_{p,r,L_1-1}\}_{p \in P, r \in R} \right) \\ &= \min E \left\{ \Theta_1(\{Q'_p\}_{p \in P}, \{Q'_{p,r}\}_{p \in P, r \in R}) + \sum_{p \in P} \left[h_p I_{p,L_1} + \sum_{r \in R} (h'_p I_{p,r,L_1}^+ + b'_p B_{p,r,L_1}) \right] \right. \\ & \quad \left. + \sum_{r \in R} c_r S_{r,L_1}^{IR} + c_I S^I + \sum_{r \in R} c'_r S_r^R + c_C S^C + \sum_{r \in R} c_r^{PR} S_r^{PR} + c_{CI} S^{CI} \right\} \end{aligned}$$

$$s_w^I, s_w^C \in \{0, 1\} \quad w \in W \quad (1)$$

$$s_{r,w}^R \in \{0, 1\} \quad w \in W; r \in R \quad (2)$$

$$s_w^I + \sum_{r \in R} s_{r,w}^R + s_w^C = 1 \quad w \in W \quad (3)$$

$$S^I = \sum_{w=1}^W s_w^I \quad (4)$$

$$S^C = \sum_{w=1}^W s_w^C \quad (5)$$

$$S_r^R = \sum_{w=1}^W s_{r,w}^R \quad r \in R \quad (6)$$

$$Q'_p + \sum_{r \in R} Q'_{p,r} \leq \sum_{r \in R} \sum_{w \in W} s_w^C Q_{p,w} \quad p \in P \quad (7)$$

$$\sum_{p \in P} v_p Q'_{p,r} x_{p,r,y} \leq V' \quad r \in R; y = 1, \dots, Y \quad (8)$$

$$\sum_{p \in P} v_p Q'_p x_{p,y} \leq V' \quad y = 1, \dots, Y \quad (9)$$

$$s_{r,y} \geq \frac{1}{M} \sum_{p \in P} Q'_{p,r} x_{p,r,y} \quad r \in R; y = 1, \dots, Y \quad (10)$$

$$s_y \geq \frac{1}{M} \sum_{p \in P} Q'_p x_{p,y} \quad y = 1, \dots, Y \quad (11)$$

$$S_r^{PR} = \sum_{y=1}^Y s_{r,y} \quad r \in R \quad (12)$$

$$S^{CI} = \sum_{y=1}^Y s_y \quad (13)$$

$$\sum_{y=1}^Y x_{p,r,y} = 1 \quad p \in P; r \in R \quad (14)$$

$$x_{p,r,y} \in \{0, 1\} \quad p \in P; r \in R; y = 1, \dots, Y \quad (15)$$

$$x_{p,y} \in \{0, 1\} \quad p \in P; y = 1, \dots, Y \quad (16)$$

$$s_{r,y} \in \{0, 1\} \quad r \in R; y = 1, \dots, Y \quad (17)$$

$$s_y \in \{0, 1\} \quad y = 1, \dots, Y \quad (18)$$

$$I_{p,L_1} = I_{p,L_1-1} - \sum_{p \in P} q_{p,r,t} \quad p \in P \quad (19)$$

$$I_{p,r,L_1} = I_{p,r,L_1-1} - D_{p,r,L_1} \quad p \in P; r \in R \quad (20)$$

$$I_{p,r,L_1}^+ \geq I_{p,r,L_1} \quad p \in P; r \in R \quad (21)$$

$$B_{p,r,L_1} \geq -I_{p,r,L_1} \quad p \in P; r \in R \quad (22)$$

$$\sum_{r \in R} q_{p,r,L_1} \leq I_{p,L_1-1} \quad p \in P \quad (23)$$

$$\sum_{p \in P} v_p q_{p,r,L_1} x_{p,r,L_1,u} \leq V' \quad r \in R; u = 1, \dots, U \quad (24)$$

$$s_{r,u,L_1} \geq \frac{1}{M} \sum_{p \in P} q_{p,r,L_1} x_{p,r,L_1,u} \quad r \in R; u = 1, \dots, U \quad (25)$$

$$S_{r,L_1}^{IR} = \sum_{u=1}^U s_{r,u,L_1} \quad r \in R \quad (26)$$

$$\sum_{u=1}^U x_{p,r,L_1,u} = 1 \quad p \in P; r \in R \quad (27)$$

$$x_{p,r,L_1,u} \in \{0, 1\} \quad p \in P; r \in R, u = 1, \dots, U \quad (28)$$

$$s_{r,u,L_1} \in \{0, 1\} \quad r \in R; u = 1, \dots, U \quad (29)$$

$$Q_{p,w}, Q'_p, Q'_{p,r} \in \{0, 1, 2, \dots\} \quad (30)$$

The objective function of Problem Θ_3 accounts for the daily logistics costs of transferring goods from the IDC to the RDCs in periods 1 though $L_1 - 1$, just as in Problem Θ_1 . A decision is made in Problem Θ_3 regarding how many containers to ship initially and the quantity of each product in each container, which is represented by the decision variables $Q_{p,w}$. Constraint (12) limits the goods packed into each container w to the volume available. Although the initial shipping decision is made in this problem, no costs for shipping containers are accrued in the objective function of this problem. Instead, all logistics costs are accounted for

in Problem Θ_2 when replenishment decisions are finalized. The cost impact of the decisions $Q_{p,w}, r \in R, p \in P$, in problem Θ_3 comes later and depends on how much the original replenishment decision must be adjusted, as reflected in Problem Θ_2 .

Problem Θ_3 includes all of the constraints pertaining to the daily transfer of goods from the IDC to the RDCs for periods 1 though $L_1 - 1$, just as in Problem Θ_1 , with the exception that no goods are received over the time frame encompassed by Problem Θ_3 so that two constraints (2 and 4) in Problem Θ_1 are not required.

$$\begin{aligned} & \Theta_3 (\{I_{p,0}\}_{p \in P}, \{I_{p,r,0}\}_{p \in P, r \in R}) \\ &= \min \mathbf{E} \left\{ \Theta_2(W, \{Q_{p,w}\}_{p \in P, w \in W}, \{I_{p,L_1-1}\}_{p \in P}, \{I_{p,r,L_1-1}\}_{p \in P, r \in R}) \right. \\ & \quad \left. + \sum_{t=L_1-1}^T \left\{ \sum_{p \in P} \left[h_p I_{p,t} + \sum_{r \in R} (h'_p I_{p,r,t}^+ + b'_p B_{p,r,t}) \right] + \sum_{r \in R} c_r S_{r,t} \right\} \right\} \\ & \text{s.t.} \\ & I_{p,t} = I_{p,t-1} - \sum_{r \in R} q_{p,r,t} \quad p \in P; t = 1, \dots, L_1 - 1 \quad (1) \\ & I_{p,r,t} = I_{p,r,t-1} - D_{p,r,t} \quad p \in P; r \in R, t = 1, \dots, L_1 - 1 \quad (2) \\ & I_{p,r,t}^+ \geq I_{p,r,t} \quad p \in P; r \in R; t = 1, \dots, L_1 - 1 \quad (3) \\ & B_{p,r,t} \geq -I_{p,r,t} \quad p \in P; r \in R; t = 1, \dots, L_1 - 1 \quad (4) \\ & \sum_{r \in R} q_{p,r,t} \leq I_{p,t-1} \quad p \in P; t = 1, \dots, L_1 - 1 \quad (5) \\ & \sum_{p \in P} v_p q_{p,r,t} x_{p,r,t,u} \leq V' \quad r \in R; t = 1, \dots, L_1 - 1; u = 1, \dots, U \quad (6) \\ & s_{r,u,t} \geq \frac{1}{M} \sum_{p \in P} q_{p,r,t} x_{p,r,t,u} \quad r \in R; u = 1, \dots, U; t = 1, \dots, L_1 - 1 \quad (7) \\ & S_{r,t} = \sum_{u=1}^U s_{r,u,t} \quad r \in R; t = 1, \dots, L_1 - 1 \quad (8) \end{aligned}$$

$$\sum_{u=1}^U x_{p,r,t,u} = 1 \quad p \in P; r \in R; t = 1, \dots, L_1 - 1 \quad (9)$$

$$x_{p,r,t,u} \in \{0, 1\} \quad p \in P; r \in R; u = 1, \dots, U; t = 1, \dots, L_1 - 1 \quad (10)$$

$$s_{r,u,t} \in \{0, 1\} \quad r \in R; u = 1, \dots, U; t = 1, \dots, L_1 - 1 \quad (11)$$

$$\sum_{p \in P} v_p Q_{p,w} \leq V \quad w \in W \quad (12)$$

$$Q_{p,w} \in \{0, 1, 2, \dots\} \quad p \in P; w \in W \quad (13)$$

The Global Replenishment Problem and Related Literature

In this section, we discuss the solution of the problem as outlined in the previous section and relate it to relevant literature. We first consider the solution of Θ_1, Θ_2 , and Θ_3 with a simplifying assumption that demand is deterministic. Subsequently, we consider possible solution methodologies with stochastic demand.

Before proceeding, we situate the framework of Problems Θ_1, Θ_2 , and Θ_3 in the contexts of ongoing replenishment and one-time purchases. The framework of Θ_1, Θ_2 , and Θ_3 was initially presented in the context of ongoing replenishment: Θ_1, Θ_2 , and Θ_3 constitute the problem for one replenishment cycle, and ongoing replenishment would be represented by a sequence of, perhaps, an infinite number of these replenishment cycles where the ending inventory of each cycle was the starting inventory of the next cycle. To represent a one-time purchase, one might consider only one ordering cycle, that is, solving Problem Θ_1, Θ_2 , and Θ_3 without repeating further cycles. In that case, we might extend the time horizon of Problem Θ_1 for some number of periods past time T , which is when selling would commence assuming that no prior inventory was in the IDC and RDCs. We would also want to include revenue in the objective function to penalize lost sales and, possibly, a salvage value at the end of the selling horizon. (In the continuous replenishment formulation, unfulfilled demand in any period is penalized by a backorder cost, which motivates some minimum service level.) This formulation structure allows investigation of the effectiveness of holding a portion of

the one-time purchase quantity back at the IDC until demand at the various RDCs is observed. Holding inventory back at the IDC in this fashion is not uncommon for seasonal products such as holiday decorations, and this practice is reminiscent of the accurate replenishment approach for replenishment of seasonal items where orders are placed two times within a selling season, and where the forecast of demand for the second order is based on sales of the goods in the first order [13]. A problem with a horizon intermediate between these two extremes might also be considered with a finite number of replenishment cycles. For example, considering two ordering cycles and enhancing the formulation by updating the demand forecast in the second cycle based on demand in the first cycle would create a generalized version of the accurate replenishment formulation.

The global replenishment problem, as we have described it, has only one stochastic element in the formulation, which is represented by $D_{p,r,t}$ for demand of product p at RDC r in period t . If we now assume that demand is deterministic, we are left with a purely deterministic, and simpler, problem. We begin by discussing that deterministic formulation. In this case, no uncertainty exists over the problem horizon so that we can determine the solution to all three problems simultaneously at the outset at time $t = 0$. Note that, although no realignment of goods at the US crossdock from container to container is needed due to demand uncertainty (the correct quantities of goods in containers can be loaded initially), realignment from container to truck trailer may be necessary for logistics efficiency (the contents of three 40' containers can normally be loaded into two 53' trailers). A single problem can thus be

written by combining all the constraints and objective function terms from all problems.

Such a problem involves (i) determining shipping quantities of each product, (ii) determining how many containers to ship and how many of each product to place in the containers, and (iii) how to deconsolidate and repack these goods that are crossdocked. The final two steps constitute two sequential multiple knapsack problems, the latter one (Problem Θ_2) being constrained by the former (Problem Θ_3) and the decision of which containers to repack. The literature on the multiple knapsack problem and other variants of the knapsack problem is extensive; two excellent references on exact and approximate solutions are Refs 14, 15. Note that the multiple knapsack problem is strongly NP-Complete [16]. Thus, two sequential multiple knapsack problems, where the first presents a constraint on the second, are even more complex than a single problem. Polynomial time approximation algorithms are available for a single multiple knapsack problem [16], which might possibly be employed for two sequential problems.

Although simplifying the problem by assuming deterministic demand would allow us some insight into how containers might be efficiently repacked on the US side of the supply chain for transportation efficiencies, it does not allow investigation of safety stock or what percentage of containers should be highly assorted for direct shipment to RDCs versus packed overseas with fewer SKUs for deconsolidation and resorting at a US crossdocking facility in reaction to demand revelation during transit. It is probable that uncertainty influences decision policies significantly and, although the global replenishment problem is still difficult with deterministic demand, this simplification will not provide some answers a manager might be seeking.

Simplifying the problem further, one might consider allowing shipping quantities to be continuous variables, which results in a binary integer program that is likely to be much simpler computationally. The binary variables that remain, s , determine how many containers or trailers are employed in each link of the supply chain. It might be

possible to prove convexity results in these variables (e.g., cost convexity in the number of containers shipped from overseas), which would make the computations easier, or perhaps an efficient branch and bound algorithm might be possible. (see the article titled **Branch-and-Bound Algorithms** in this encyclopedia.)

Solving Problems Θ_1, Θ_2 , and Θ_3 with stochastic demand adds complexity, which not only involves binary and integer decision variables but also stochastic demand in a multiple-stage multiple knapsack problem. Moreover, this sequence of problems must be solved repeatedly for continuous replenishment. Possible solution methodologies include dynamic programming, heuristic policies, and simulation.

Dynamic programming can be used to analyze this problem either analytically or numerically. A typical approach to proving structural results (e.g., convexity of an objective function) can be accomplished by applying calculus and the theory of dynamic programming (see the section titled "Dynamic Programming" in this encyclopedia). Such approaches can be used to prove the existence of a unique optimal solution, although they often fail at providing a closed-form solution. One complication of the formulation that we have presented is that each replenishment cycle requires the solution of three sequential problems. Furthermore, that sequence is repeated for continuous replenishment. A hierarchical approach to such problems has been developed whereby, in our context, the optimal solution for a single sequence of Problems Θ_1, Θ_2 , and Θ_3 is proven and, subsequently, dynamic programming theory is applied to a sequence of those embedded problems. Research that has considered embedding multiple decisions or "subperiods" within one period of a dynamic program includes Ref. 17.

When dynamic programming fails to provide for easy computation of the optimal solution, numerical methods can be used to compute an answer within an interval of the optimal solution. Two approaches are possible including value iteration and policy iteration (see **Approximate Dynamic Programming—II: Algorithms** and

Dynamic Programming, Control, and Computation.)

Discrete-event simulation models can also be used to address the global replenishment problem. Used in a “what-if” mode of investigation, this approach might help provide insight into effective decision policies, although it can neither prove the existence of an optimal solution nor can any solution be guaranteed to be optimal (see ***Introduction to Discrete-Event Simulation***). Simulation-based gradient search methods might also be applied to find optimal solutions numerically, provided the existence of the optimal solution and its structure have otherwise been proved (e.g., via dynamic programming). Infinitesimal perturbation analysis (IPA) is one such method that can be used to estimate the gradients of expected values of objective functions, although it can be applied only under certain conditions [18]. One of the conditions required by IPA is the continuity of the parameters with respect to which gradients are computed. For the decision variables that we would want to search over, our formulation contains binary variables representing decisions to load a particular container or truck and integer variables representing shipping quantities. Relaxing the problem to consider continuous shipping quantities might, perhaps, be an effective tack. However, relaxing the integrality of binary variables is very likely not a promising approach. Thus, IPA might be used to solve a portion of our overall formulation, while some algorithm to handle the binary variables would still be required.

If the complexity of the problem does not permit effective analysis by any of the foregoing methodologies, then other heuristic approaches might be applied. Specifically, one might apply policies that are known in the literature to be optimal in other circumstances, even though it is neither known nor likely that they are optimal for the global replenishment problem. For example, base-stock or (s, S) policies might be applied (using simulation to evaluate their effectiveness) and compared to determine which performs best. Other intuitive heuristics might also be applied, such as the following service level-based policy:

- specify a service level for replenishing each RDC from the IDC;
- specify a service to provide each RDC from the crossdocking facility;
- specify a service level for each RDC up to which assorted containers would be loaded overseas;
- specify a service level for aggregate RDC demand up to which nonassorted containers would be loaded overseas;
- develop an algorithm for adapting the replenishment quantities suggested by the service levels above to the loading constraints (e.g., container volume limitations).

This service level approach is analogous with the Newsvendor solution, although this problem has many more dimensions, greater complexity, and constraints not present in the basic Newsvendor model (see ***Newsvendor Models***). Thus, the extensive research on the constrained newsvendor problem or Newsvendor models with multiple periods might be applied to a portion of our formulation [19]. Of course, policies from practice might also be evaluated against competing alternatives. One such policy we have observed is replenishing RDCs based on point of sale data and, in a continuous-time mode, sending a truck directly to an RDC whenever it is filled.

Extensions

Many extensions to our formulation that reflect various facets of real-world supply chains are possible, some of which we outline in this section.

First, our formulation is for a single shipping lane where an origin port and a destination port are specified. The flows of goods are from manufacturers or sources of supply in the vicinity of the origin port. While the scope of the formulation encompasses many decisions, a higher-level formulation could be constructed with a broader scope to evaluate the value of integrating goods flow from many shipping lanes or from multiple ports on the same shipping lane, where goods from multiple origins could be crossdocked

together in the United States. That formulation could be constructed by integrating multiple replicates of our formulation.

The current formulation implicitly assumes that there are destinations for all containers before the containers arrive at the ports. This is not always the case in practice, and goods are sometimes stored in a container until a final decision regarding destination is made, in which case the containers might remain at the port and possibly incur storage charges. This extension could be incorporated by adding an additional stocking location to the formulation, which would represent storage at the port.

Our formulation assumes a zero lead time from the port to the IDC and RDCs, and from the IDC to the RDCs. We also did not incorporate a lead time to unload the goods at either the IDC or the RDC, which could be viewed as part of the lead time that we did represent. One could add a separate deterministic nonzero lead time for unloading at the IDC and RDCs, which would not greatly complicate the formulation. Further, we may want to represent the stochastic nature of transportation or unloading by representing each corresponding lead time as a stochastic random variable. Unloading lead time might vary due to fluctuations in container contents (small vs large items), daily fluctuations in staffing (e.g., absenteeism), fluctuations in the number of containers that arrive each day, or simply variations in daily efficiency. The last factor might motivate us to represent unloading lead time as an independent random variable, whereas the other circumstances might require other types of formulation enhancements. For example, fluctuations in container volume might be handled by adding constraints that limit how many containers can be unloaded on a given day. These enhancements could complicate the formulation quite significantly.

Two other possible extensions include the following:

- We have not considered downstream logistics that would incorporate truck routings that serve multiple RDCs or retail locations. A rich literature exists

on this topic of the Direct Shipping Problem [20–22]. This problem combines the problems of inventory control and vehicle routing to consider *inventory routing* models, thus combining inventory and transportation costs. In most formulations, a single depot (IDC, RDC, or crossdocking operation) services multiple retailers with a single product. We could extend the formulation developed here to incorporate this final stage of the supply chain. Of course, the solution of the extended formulation would be further complicated.

- We have offered a formulation for continuous replenishment and described how it can be adapted to one-time purchases. These approaches could be combined in a formulation that includes both one-time purchases and continuously replenished goods.

CONCLUSIONS

The global replenishment problem that we have described in this article is of growing importance in a world with increasing global trade. In relating the existing literature to the formulation that we present, we have noted where current research can address portions of the formulation or simplifications of the formulation, but no research has been carried out on the problem in total. Future research can then pursue several directions. First, efforts can be made to integrate many models, each of which addresses a portion of the problem. An approach such as that in Ref. 17 is one framework that might provide such integration, specifically among Problems Θ_1 , Θ_2 , and Θ_3 . However, additional work is needed to handle the complexity of these subproblems before they can be successfully integrated. A second direction would be to investigate simplifications of the formulation that we have provided. As is always the case, the challenge with this tack is to find a simplification that is both tractable and a meaningful representation of the problem. Finally, heuristic approaches are always possible via simulation to evaluate candidate policies against

one another. The advantage of this approach is that no simplification of the formulation may be required, although one cannot identify an optimal solution with this approach. This is indeed a complex and important problem that can be the source of much research.

REFERENCES

1. United States Census Bureau Federal Trade Statistics. Available at <http://www.census.gov/foreign-trade/statistics/historical/index.html>. Accessed 2009 Mar 20.
2. Bianco A, Zellner W. Low prices are great: But Wal-Mart's dominance creates problems—for suppliers, workers, communities, and even American culture; is Wal-Mart too powerful. *Bus Week* 2003;3852:100.
3. Gooley T. Third-party logistics: double duty. *DC Velocity* 2009;7(2):43–45.
4. James City County Office of Economic Development. Wal-Mart import distribution center hosts grand opening in James City County. Available at <http://www.jccecondev.com/news/archive/2000/walmart111700.html>. Accessed 2009 Mar 3.
5. Newswanger P. Not just for lovers any more. *J Commer* 2002;3(45):34–35.
6. Maersk Distribution. Available at <http://www.maersk-distribution.com/Documents/Maersk%20Distribution%20Brochure.pdf>. Accessed 2010 Feb 23.
7. NYK Logistics. Available at http://www.na.nyklogistics.com/wrws_crossDock.shtml?page=wrws2. Accessed 2010 Feb 23.
8. Hoffman W. Changing the fuel game. *Traffic World* 2008;272(29):8.
9. Mongelluzzo B. Trans-pacific carriers seek drastic rate hike. *J Commer Online* 2009. Available at <http://www.joc.com/node/412240>. Accessed 2010 Feb 23.
10. Rodrigue JP, Comtois C, Slack B. *The geography of transport systems*. New York: Routledge; 2009.
11. Bingham P. The demographics of freight growth. 3rd NARC Freight and Goods Movement Summit; 2006. Available at http://narc.org/uploads/File/Transportation/Freight%20Summit/Bingham_NARC.pdf. Accessed 2009 Apr 8.
12. Davis J. Dollar stores' stature grows purchasing power of suburbanites lures discount retailers to area. *Chicago Daily Herald*. 2005. p. 1.
13. Raman A, Fisher M, Hammond J, *et al.* Configuring a supply chain to reduce the cost of demand uncertainty. *Prod Oper Manage* 1997;6(3):211–225.
14. Martello S, Toth P. *Knapsack problems: algorithms and computer implementations*. New York: John Wiley & Sons, Inc.; 1990.
15. Pisinger D. An exact algorithm for large multiple knapsack problems. *Eur J Oper Res* 1999;114(3):528–541.
16. Chekuri C, Khanna S. A polynomial time approximation scheme for the multiple knapsack problem. *SIAM J Comput* 2006;35(3):713–728.
17. Huggins EL, Olsen TL. Inventory control with generalized expediting. Working paper. 2007. Available at <http://apps.olin.wustl.edu/workingpapers/pdf/2003-12-005.pdf>.
18. Glasserman P. *Gradient estimation via perturbation analysis*. Boston (MA): Kluwer Academic Publishers; 1991.
19. Nahmias S, Schmidt CP. An efficient heuristic for the multi-item newsboy problem with a single constraint. *Nav Res Logist Q* 1984;31(3):463–474.
20. Barnes-Schuster D, Bassok Y. Direct shipping and the dynamic single-depot/multi-retailer inventory systems. *Eur J Oper Res* 1997;101(3):509–518.
21. Bertazzi L. Analysis of direct shipping policies in an inventory-routing problem with discrete shipping times. *Manage Sci* 2008;54(4):748–762.
22. Gallego G, Simchi-Levi D. On the effectiveness of direct shipping strategy for the one-warehouse multi-retailer R-systems. *Manage Sci* 1990;36(2):240–243.

THE GRAPH MODEL FOR CONFLICT RESOLUTION

KEITH W. HIPEL
Department of Systems Design
Engineering, University of
Waterloo, Waterloo,
Ontario, Canada

D. MARC KILGOUR
Department of Mathematics,
Wilfrid Laurier University,
Waterloo, Ontario, Canada

LIPING FANG
Department of Mechanical and
Industrial Engineering,
Ryerson University,
Toronto, Ontario, Canada

Conflict inevitably arises whenever humans interact. For example, family members may have differences of opinion regarding the choice of restaurant for dinner together. An employee may be in dispute with her boss over how much her annual salary increase should be. In the highly competitive business world, companies aggressively promote their products to increase their share of the market within a sector. Political parties vie with one another to attract supporters by proposing different policies. Internationally, ongoing negotiations aim to find an appropriate and effective means for combating greenhouse gas emissions and thereby mitigating climate change. In all of these examples, strategy, capability, and preference play an important role in determining the resolution. As a specific illustration, the international dispute over the sharing of water in the Aral Sea Basin of central Asia is examined later in this article.

Because of the ubiquity of conflicts, which range from personal disagreements to international military campaigns, and which occur within practically every field of study including law, engineering, and health care, there is a great demand for formal methodologies to help decision makers

(DMs) facing conflicts, and to find stable resolutions. The aforementioned “social” conflicts possess certain key characteristics including the presence of

- two or more DMs, such as people, organizations, and nations, participating in the dispute;
- different options or actions under the control of each of the DMs;
- a separate value system for each DM, which has multiple dimensions or objectives, many of which are in direct conflict with the values of other participants in a conflict;
- relative preferences for a particular DM that reflect his or her value system, with respect to the conflict being investigated;
- possible moves and countermoves controlled by a given DM, which can be brought into play at any time or not at all;
- strategic moves and countermoves among the DMs that could dynamically take place at any time, in any order, as the conflict evolves from a starting point to an intermediate or final outcome;
- a range of ways people may behave under conflict situations, such as an insightful DM who can perceive many potential moves and countermoves into the future before deciding what to do, much like a clever chess player planning his or her next move;
- participants who may act completely independent of one another or sometimes in cooperation with others by forming a coalition, in which members of the coalition benefit from their cooperative actions;
- high uncertainty caused by, for instance, unknown relative preferences and many DMs taking part in a conflict;

- psychological factors like attitudes, emotions, and misunderstandings.

When designing a large building, one must take into account the local geophysical characteristics, such as the soil properties and possible seismic occurrences, as well as meteorological conditions including temperature, wind, and precipitation. Likewise, when designing a model for systematically studying social conflict, one must consider the foregoing list of key characteristics. The Graph Model for Conflict Resolution [1,2] was purposefully designed to handle all of these characteristics of real-world conflict.

Besides the physical environment, the building must be designed to provide a range of purposes or uses such as retail space for stores located at ground level and offices for different types of organizations at higher floors. Likewise, conflict models must be structured to not only take into account the key characteristics of conflict but also to furnish desired uses or purposes. Specifically, the Graph Model for Conflict Resolution has incorporated into its basic design the capability to perform the following:

- Summarize in an organized way, complex and often confusing information about a conflict into a clear model structure. This process puts the conflict into perspective by focusing on the essentials of the problem within a simple, yet revealing, model framework.
- Function when information or data are scarce or when one is drowning in an overabundance of information.
- Enhance the understanding of the dispute being investigated via this type of systems thinking.
- Facilitate meaningful communication using the structured “graph model language” among stakeholders and interested parties.
- Predict the strategic consequences of making potential choices under conflict in order to select the best possible decision, given the social and strategic constraints existing in light of what others may do to advance their own positions.

This will reduce making misinformed decisions having potentially highly negative and perhaps irreversible impacts.

- Be cognizant of the strategic implications of other characteristics of the conflict under study, such as coalition formation, preference uncertainty, and psychological factors.
- Make informed decisions based on the foregoing sensible modeling and analyses, which may lead to win/win resolutions.

The formal study of social conflict started with the launching of Metagame Analysis by Howard [3] who recognized the need for techniques that could handle the nonquantitative and interactive aspects of conflict. In an approach called *Conflict Analysis*, Fraser and Hipel [4,5] expanded metagame analysis by introducing additional ways to model human behavior under conflict, analytical methods for carrying out calculations, and computerized implementation of modeling and analytical techniques. In turn, Fang *et al.* made significant improvements to Conflict Analysis through the development of a highly flexible and comprehensive methodology named the *Graph Model for Conflict Resolution*. The word “graph” is used in the title because a graph can theoretically keep track of the potential moves and countermoves that a given decision maker controls in a conflict. Moreover, the sound theoretical design of the Graph Model provides a solid foundation upon which many novel expansions can be made, such as the ability of the Graph Model to take into account cooperation, high uncertainty, and various psychological factors discussed later. In addition, a flexible decision support system can be constructed for computerizing the Graph Model methodology. Another direction in which Metagame Analysis was extended is Drama Theory [6–9], in which a conflict is envisioned as evolving like a drama through various phases and dilemmas to a final resolution. Drama Theory is the subject of the article ***Drama Theory*** in this encyclopedia. As explained by Obeidi and Hipel [10], Drama Theory and the Graph Model provide complementary ways to investigate conflict, and the Graph Model can

be used to analyze a conflict at a particular phase in a drama.

In the next section, an overall procedure is presented for systematically studying a real-world dispute in conjunction with a decision support system. Next, a water resources conflict in the Aral Sea Basin is used to demonstrate how one can systematically model and analyze a conflict to obtain a number of strategic insights for enhancing the decision-making process. Finally, valuable extensions of the Graph Model for better reflecting the characteristics of actual conflict and expanding its domain of applicability are summarized in the last section.

SYSTEMATIC INVESTIGATION OF REAL-WORLD CONFLICT

For small conflicts, one can do much of the modeling and analysis by hand. However, for medium-size and large disputes, it is convenient and expeditious to employ a computerized system. In particular, a decision support system is a computerized system that contains all of the key components necessary to execute a thorough study of a particular topic for the purpose of supporting DMs [11].

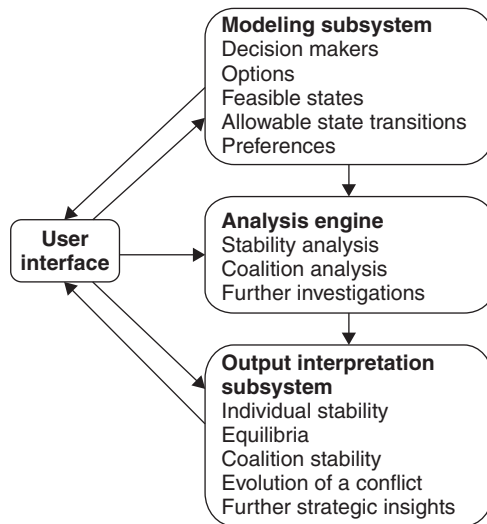


Figure 1. The decision support system, GMCR II, for conflict resolution.

Figure 1 displays the general design of a decision support system for the Graph Model for Conflict Resolution, which is referred to as GMCR II, [12,13]. As can be seen, via the user interface, an analyst can interact with the modeling subsystem, analysis engine, and output interpretation subsystem when investigating a given dispute. Within the modeling subsystem, a conflict model is developed in terms of DMs and the options or courses of action available to each DM. On the basis of the combination of ways in which DMs can separately select their options, GMCR II can automatically generate the states or scenarios that could occur. States that cannot possibly take place in reality are removed from the conflict model in order to end up with the list of feasible states. Finally, a very important step is to obtain the relative preferences of each DM with respect to the set of feasible states. In the realm of physical systems, the basic laws of nature, such as gravity, dictate how a particular system will dynamically change. In the social realm of conflict resolution, the value systems of DMs as expressed by their preferences over states are the driving forces of human behavior. Hence, ascertaining the preferences of DMs is a crucial step in the calibration of a conflict model.

After determining an initial conflict model for the dispute under study, the user can instruct the analysis engine to carry out a stability analysis and thereby ascertain the possible equilibria or compromise resolutions to the conflict, which will then appear as part of the output. A given state is deemed to be stable for a particular DM if it is not advantageous for the DM to move to another state. For example, a DM may have the ability to move on his or her own to a more preferred state by changing his or her option selection or strategy. However, if the competitors can invoke countermoves that put the original DM in a less preferred situation, he or she is better off to stay at the initial state which is said to be stable. Because DMs may behave differently in conflict situations, a range of stability definitions have been formulated to define different patterns of moves and countermoves under which these stabilities can take place.

A general approach to follow when executing a conflict study is to first determine how well a given DM can perform on his or her own via independent behavior. Subsequently, one can find out if the DM can do even better by cooperating with others by joining a coalition, in which all members of the coalition benefit when certain actions are taken [14,15].

As indicated in the output interpretation subsystem box in Fig. 1, one can also examine the dynamic evolution of a conflict over time from a given starting state to an attractive equilibrium. In some cases, it may not be possible to reach a desirable situation unless one cooperates with others. In fact, one can use a decision support system in an iterative manner. As one gains a better understanding of the conflict via, for instance, executing basic analyses and discussing these results with others, one can make meaningful changes to the conflict in the modeling subsystem, and subsequently execute another analysis. Further types of investigations that could be carried out are discussed in the section titled "Further Investigations and Insights."

The general Graph Model methodology depicted in Fig. 1 can be used in the following three main types of situations:

- *Analysis and Simulation Tool for a DM in a Conflict, or a DM's Agent.* Moves and countermoves following a DM's possible actions can be analyzed, and the potential consequences of certain actions assessed, in order to improve the DM's position. Preparations and assessments can be carried out at different times as the conflict unfolds. For example, an analyst may advise the US government about strategic initiatives to invoke in Afghanistan over time.
- *Communication and Analysis Tool in Mediation.* GMCR II can be utilized by a mediator to assess possible strategic consequences of a range of preferences for DMs, without knowing which preferences are correct. Options that are beneficial, detrimental, or irrelevant to all parties can be ascertained by using this process. In negotiations

between labor and management over designing a new contract, a mediator may be able to demonstrate how only small concessions and minimum shifts in preferences by each side may result in a mutually beneficial contract.

- *Analysis Tool for a Third Party Analyst.* On the basis of the observed outcome and evolution of a conflict, an analyst can estimate the DMs' preferences. How the structure of the conflict influenced DMs' behavior can also be studied, and better ways to structure a future conflict can be identified. For instance, the government of China may wish to know the strategic effectiveness of US policy in Afghanistan, even though China is not a combatant in Afghanistan.

STRATEGIC STUDY OF THE ARAL SEA CONFLICT

Conflict arising over the ongoing environmental damage caused by societal activities is becoming more widespread and serious. For example, the failure of the nations of the world to reach an effective agreement over the containment and significant reduction of greenhouse gases in Copenhagen in December 2009 means that the consequences of climate change will become much worse, and the ensuing conflict will intensify. Clearly, powerful negotiation procedures and tools are needed now to guide society in a responsible direction. In the past, smaller societies such as those in Easter Island, Greenland, and elsewhere have collapsed largely because of the destruction of the local environment, such as cutting down all of the trees in Easter Island. One of the most dramatic environmental nightmares of the twentieth century was the ecological destruction of the Aral Sea, which was directly brought about by human folly. The purpose of this section is to outline how the Graph Model for Conflict Resolution can be conveniently applied to real-world conflict to gain strategic insights by using it to study a current water quantity conflict existing within the Aral Sea Basin.

Background to the Aral Sea Dispute

Figure 2 displays the Aral Sea Basin, which is located in the heart of the Eurasian continent, where it extends over the territories of the five Central Asian Republics: Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, and Uzbekistan. Some smaller parts of the basin lie in Afghanistan, China, and Iran. As depicted on the map, the two main rivers flowing into the Aral Sea are the Amu Darya, which starts in the mountains of Afghanistan and Tajikistan and flows through Turkmenistan and Uzbekistan to the Aral Sea, and the Syr Darya, which originates in Kyrgyzstan and flows through Tajikistan, Uzbekistan, and Kazakhstan to the Aral Sea [16].

To obtain hard currency through exports, the Soviet Union promoted the large-scale growing of irrigated cotton in the Aral Sea Basin and became a net exporter of cotton in 1937. Beginning in 1960, large quantities of water from the rivers were diverted for irrigating millions of hectares of land, resulting in the level of the Aral Sea beginning to drop. While the sea had been receiving about 50 km^3 of water per year in 1965, by the early 1980s this amount had decreased

to zero. The shrinking of the Aral Sea caused the salinity to increase in combination with a significant decline in the previously large fish catch. Additionally, the declining sea level lowered the water table in the region, which destroyed many oases near its shores, and created dangerous levels of polluted airborne sediments in areas that became dry.

The diversion of water to irrigate agricultural land had other direct negative effects. In particular, ecologically sensitive areas in the river deltas were converted to cropland, pesticides were heavily used throughout the Aral Sea Basin resulting in heavy contaminant concentrations in the sea, and high salt concentrations built up on the irrigated fields. By the start of the 1990s, the surface area of the Aral Sea had shrunk by nearly half, and the volume was reduced by 80%. Regional climate became more continental, thereby shortening the growing season and causing some farmers to switch from cotton to rice, which required even more diverted water. The exposed area of the former seabed was more than $28,000 \text{ km}^2$ from which winds picked up millions of tons of sediments laced with salts and pesticides, with devastating health consequences for people living in the surrounding regions.

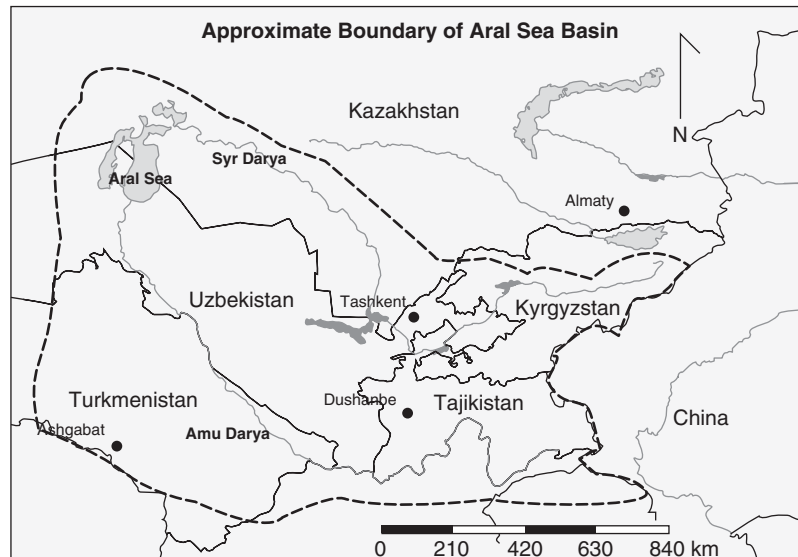


Figure 2. Aral Sea Basin.

Table 1. The Decision Makers and Their Options in the Syr Darya River Water Sharing Conflict

Decision Makers	Options
<i>Downstream</i> Two downstream countries: Uzbekistan and Kazakhstan	1. <i>Less fuel</i> : supply less fuel than specified 2. <i>Pressure</i> : exert political and military pressure 3. <i>Follow agreement</i> : follow the present agreement
<i>Upstream</i> One upstream country: Kyrgyzstan	4. <i>Ignore agreement</i> : generate more hydroenergy in winter than agreed upon 5. <i>Follow agreement</i> : do not release too much water for hydropower generation during winter and release sufficient water for irrigation during summer 6. <i>Revise agreement</i> : revise agreement to meet winter hydropower demand and allow for summer irrigation
<i>ICWC</i> Interstate Commission for Water Coordination	7. <i>Enforce agreement</i> : enforce the present agreement. 8. <i>Revise agreement</i> : revise agreement to try to meet new demands of both upstream and downstream countries.

In 1988, the Soviet Union decreed that cotton growing has to be reduced, expecting the Aral Sea to receive water in gradually increasing amounts. However, the dissolution of the Soviet Union in 1991 ended such central authority, and since then the Aral Sea crisis has been under the control of the five Central Asian independent nations consisting of Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, and Uzbekistan. Unfortunately, water use has increased rapidly and is now at an unsustainable level. Under the Soviets, water and energy resources were exchanged freely across the entire region. Since becoming independent, the five Central Asian states have failed to come up with a viable regional approach to replace the Soviet system of management because of intense competition among them. To help overcome this, in 1992, the five states set up an intergovernmental body called the *Interstate Commission for Water Coordination* (ICWC) to manage water resources in the Aral Sea Basin.

The main type of conflict existing in the Aral Sea Basin is the classic battle between downstream and upstream countries, over which nation controls and utilizes the water in the main rivers flowing across the basin to the Aral Sea. More specifically, the downstream countries require large quantities of water for their expanding agricultural sectors

and rising populations, while the economically weaker upstream countries need more water for electricity generation and farming.

Nandalal and Hipel [17] provide an overview of the recent history of the Aral Sea Basin as well as references in which detailed information and environmental studies of the area can be found [16,18–20]. These authors used the Graph Model for Conflict Resolution to carry out a strategic examination of the downstream versus upstream conflict among countries along the Syr Darya river. The conflict study presented in this section to illustrate how the Graph Model can be used in practice is based on the work of Nandalal and Hipel [17].

Modeling the Aral Sea Dispute

Under the modeling subsystem shown in Fig. 1, one can see that two key components in constructing a conflict model are the DMs and the options or courses of action controlled by each DM. The left column in Table 1 lists the three main DMs in the water sharing conflict along the Syr Darya river as Downstream (Uzbekistan and Kazakhstan), Upstream (Kyrgyzstan), and the ICWC. The right column furnishes the options under the control of each DM along with an explanation as to what each option means.

After modeling and analyzing the conflict following the steps given in the section titled

Table 2. Evolution of the Conflict from the Status Quo, Through Two Transitional States, to the Final Resolution

Decision Makers and Options	Initial State		First Intermediate State		Second Intermediate State		Final Equilibrium
<i>Downstream</i>							
1. Less fuel	N		N		N		N
2. Pressure	Y	→	N		N		N
3. Follow agreement	N	→	Y		Y		Y
<i>Upstream</i>							
4. Ignore agreement	Y		Y	→	N		N
5. Follow agreement	N		N	→	Y		Y
6. Revise agreement	N		N		N		N
<i>ICWC</i>							
7. Enforce agreement	N		N		N		N
8. Revise agreement	N		N		N	→	Y

“Systematic Investigation of Real-World Conflict” and Fig. 1, a type of dynamic insight that one can garner from a formal conflict study is depicted in Table 2. In this table, each column of Y and N stands for a possible state or scenario that could occur in the real-world where a Y means “yes” the option opposite the Y is selected by the DM who controls it, while an N indicates “no” the option is not taken. Hence, the initial state or status quo is the situation whereby Downstream does not supply less fuel, exerts pressure on Upstream not to release excess water in the winter so it is available for irrigating the summer crops downstream, and does not completely follow the agreement. Upstream ignores the agreement and uses more water in the winter, while ICWC does not enforce or revise the agreement. The arrows in Table 2 indicate how option selections change as the conflict evolves over time from the status quo on the left to the two intermediate states shown in the center to the final state or equilibrium given on the right. When moving from the initial state to the first intermediate state, which is more preferred by Downstream to the initial state, Downstream stops exerting pressure (does not take option 2) and follows the agreement (selects option 3). This in turn entices Upstream to adhere to the agreement (choose option 5 instead of 4) to move the conflict from the first intermediate state to the second more preferred intermediate state. Next, the final equilibrium is reached

from the second intermediate state when ICWC revises the agreement and both Downstream and Upstream accept this improved agreement.

Going back to the modeling stage, one must determine the feasible states in the conflict, as indicated in Fig. 1. Since the model in Table 1 contains a total of eight options, each of which can be chosen or not, mathematically there are $2^8 = 256$ possible states. Fortunately, in practical applications, the number of feasible states is much less than the number of mathematically possible states, and for the case of the Syr Darya river water sharing conflict the number of feasible states is only 21. One can easily identify which states cannot occur in this conflict and then remove them using GMCR II. For example, the three options under the control of Upstream are mutually exclusive and, hence, Upstream can only select at most one of its three options which are numbered as options 4, 5, and 6 in Tables 1 and 2. Therefore, if Upstream were to “Ignore agreement” (option 4), its other two options of “Follow agreement” (option 5) and “Revise agreement” (option 6) could not occur, as is the case for the Initial State in Table 2. Likewise, options 7 and 8 are mutually exclusive for ICWC since it only makes sense for this DM to choose at most one of those two options. Any state in which both options 7 and 8 were selected would be removed from the conflict model. For Downstream, the first and third options, and second and third options

Table 3. Solution Concepts and Human Behavior

Solution Concepts	Stability Descriptions
Nash Stability [21,22]	DM (decision maker) cannot unilaterally move to a more preferred state.
General Metarationality (GMR) [3]	All DM's unilateral improvements are sanctioned by subsequent unilateral moves by others.
Symmetric Metarationality (SMR) [3]	All DM's unilateral improvements are still sanctioned even after possible responses by the original DM.
Sequential Stability (SEQ) [4,5]	All DM's unilateral improvements are sanctioned by subsequent unilateral improvements by others.
Limited-Move Stability (L_h) [2,23,24]	All DMs are assumed to act optimally and a maximum number of state transitions (h) is specified.
Nonmyopic Stability (NM) [2,25]	Limiting case of limited-move stability as the maximum number of state transitions increases to infinity.

are mutually exclusive. Moreover, at least one of the three options for Downstream and for Upstream must be taken, since these countries are expected to take action. After removing all the states containing infeasible combinations of option selections that cannot take place, 21 feasible states remain, three of which are displayed in Table 2.

Notice in Table 2 that Downstream can cause the conflict to move from the initial state to the first intermediate state by changing its selection of options, while Upstream can unilaterally move the conflict from the first to the second intermediate state. Likewise, ICWC can make the conflict move from the second intermediate state to the final equilibrium via an option change that it controls. A movement from one state to another is called a *state transitions*; all the state transitions can easily be calculated by GMCR II, given the DMs and options. However, sometimes a transition can take place in only one direction, which is the case for the transition

from the second intermediate state in Table 2 to the final equilibrium. GMCR II can easily keep track of these irreversible moves. Since movement among states can be envisioned as a graph, in which the vertices are the states and the arcs are movements controlled by DMs, the conflict methodology described in the previous section is referred to as the *Graph Model for Conflict Resolution*.

The last, and most important, type of information needed to build a conflict model is the relative preferences of each of the DMs with respect to the states or scenarios that could take place in the conflict. Therefore, one only needs to specify each DM's ordering of the states, with ties allowed; there is no requirement to estimate the amount or intensity of any preference. Accordingly, cardinal preference information, such as utility values, is not required; the user of GMCR II enters only information that permits GMCR II to determine a ranking of states from most to least preferred for each DM, where some states may be equally preferred.

Option prioritization is a clever procedure used within GMCR II to obtain relative preference information for each of the DMs, whereby preferences are stated in terms of prioritized logical statements about which options are selected. Table 3 presents the preference statements in the order of priority for Downstream, from most important at the top of a column to least important at the bottom. The numbers in the left column of Table 3 refer to option numbers, and an English explanation for each preference statement is provided in the right column. Option number 5 located at the top means that Downstream most prefers to see Upstream follow the agreement, as explained on the right. Option number 3, written below 5, indicates that Downstream's next preference is for itself to follow the agreement. In the order of decreasing preference, Downstream would like to supply less fuel or petroleum from its oil reserves (option 1), followed by exert pressure (option 2). After that, Downstream would like Upstream not to ignore the present agreement, as indicated by the negative sign in front of option 4. Next, Downstream does not want Upstream or ICWC to revise the agreement ($-6|-8$,

Table 4. Option Prioritizing for Downstream

Preference Statements	Explanation
5	Downstream most prefers to see Upstream follow the agreement by taking Option 5.
3	Next, Downstream's preference is for itself to follow the agreement (choose Option 3).
1	Downstream would like to supply less fuel (select Option 1).
2	Downstream's next preference is to exert pressure (take Option 2).
-4	Downstream would like Upstream not to ignore the present agreement (as indicated by the negative sign before Option 4).
-6 -8	Downstream does not want to see Upstream or ICWC revise the agreement (not select Option 6 or Option 8, where stands for "or").
7	Downstream would like ICWC to enforce the agreement (take Option 7).

where | means "or". Lastly, Downstream would like ICWC to enforce the agreement (option 7).

Written horizontally in text from most to least important, the preference statements for Upstream are 5, 3, 4, 6|8, 7, -1, -2, while those for ICWC are 3 and 5, 8|6, 7 if 1|2, -4, -1| -2. A major advantage of the option prioritization approach to preference elicitation is that it closely reflects the way in which a person thinks about preferences in a given conflict. In fact, one can readily obtain prioritized preference statements in terms of options for a DM, either directly from the DM or from published descriptions of the conflict as is demonstrated here. From a theoretical viewpoint, these preference statements adhere to all of the rules of first-order logic. Under the assumption of what is called *transitivity of preferences*, GMCR II possesses an algorithm for taking the prioritized preference statements for each DM to produce a ranking of states in which ties are permitted. Frequently, a given DM's most and least preferred preference statements are initially known. When GMCR II ranks

the states based on a limited number of preference statements, there may be blocks of equally preferred states contained within the ordering of states. As more preference information becomes available, these equally preferred blocks will become less prevalent and will fully disappear when there are sufficient preference statements to produce complete ranking of states from most to least preferred with no ties. Nonetheless, even when there are blocks of equally preferred states for one or more DMs, GMCR II can still carry out stability analyses. From a practical viewpoint, this means that one can begin with a "quick and dirty" analysis and later refine preference statements as more information becomes available. In their paper, Nandalal and Hipel [17] show the ranking of states for the three DMs in the Syr Darya water sharing dispute.

Analysis and Results

As shown in Fig. 1, after a conflict model has been constructed, a stability analysis is executed using GMCR II to determine the stability of every state, for every DM, for all of the stability definitions listed in Table 4. These stability definitions describe the patterns of interaction that a DM may expect when dynamically interacting with others in terms of moves and countermoves. If it is not advantageous for a given DM to depart unilaterally from a particular state according to a given stability definition, then the state is deemed to be stable for that DM under that stability definition. For example, under the solution concept of Sequential Stability listed in Table 4, if all of a given DM's unilateral improvements from the state under consideration can be sanctioned by unilateral improvements by others, the particular state is stable according to sequential stability for that DM. If a state is stable according to a given stability definition for all of the DMs, it constitutes an equilibrium under that stability definition, and is, therefore, a compromise resolution, because no DM has an incentive to unilaterally move away from it. Fang *et al.* [2] present mathematical definitions and comparisons in graph model form for the stability definitions provided in Table 4. The original references for these

Table 5. Further Extensions of the Graph Model for Conflict Resolution

Area	Specific Topic	Explanation
Preferences	Strength of preferences	An environmentalist may greatly prefer the situation in which a company is not allowed to seriously pollute the surrounding environment. When possible unilateral improvements by a DM can be blocked by others by putting the DM in a state that is greatly less preferred, this creates a strong type of stability [34,35]. The concept of strength of preference has been generalized to multiple levels of preference [36], which can be interpreted as bringing relative preference information close to the concept of cardinal preferences.
	Unknown preferences	When the relative preference between two states is unknown, this valuable information is taken into account when calculating individual stability and potential resolutions [37]. Hence, this approach is not based upon probability or fuzzy set theory.
	Fuzzy preferences	Allowing preferences to be “fuzzy” is one way to model uncertain preference in which, for example, a DM may more or less prefer one state over another or definitely prefer one state more than the other. Fuzzy preference information can be utilized to determine fuzzy stability [38].
Psychological factors	Attitudes	In a conflict, a given DM can have a positive, neutral or negative attitude toward him or herself and others. One would expect that when all DMs in a dispute possess positive attitudes toward themselves and others a better resolution for everyone concerned can be reached. In practice, this is a common occurrence [39,40].
	Emotions	Emotions often arise in conflict situations and can propel a given dispute toward a range of final outcomes depending upon the mix of emotions that are present. The perceptual graph model is a useful approach to formally incorporating emotions into conflict investigations [41,42]. Emotions can also be expressed via strength of preference, which can be formerly modeled within the Graph Model methodology.
	Hypergames	In a hypergame, one or more DMs have a misunderstanding regarding preferences of others, available options, which DMs are participating in the conflict, or any combination of the foregoing misinterpretations. Hypergame analysis is a way for formally modeling misunderstandings and determining their strategic consequences. Because a conflict may contain misunderstandings, based on other faulty interpretations of reality, hypergames possess multiple levels of perception [43,44].

stability definitions are given in the left column in Table 4.

After carrying out an exhaustive stability analysis using GMCR II, the most likely equilibrium is the one shown on the right in Table 2, in which both Downstream and Upstream agree to follow a revised agreement proposed by ICWC. This is a particularly strong equilibrium since it is an equilibrium according to all of the stability definitions given in Table 4. The pathway for reaching this compromise resolution is indicated by the evolution of the conflict from the initial state on the left via the

intermediate states to the final outcome on the right in Table 2.

FURTHER INVESTIGATIONS AND INSIGHTS

As exemplified by the Aral Sea application in the section titled “Strategic Study of the Aral Sea Conflict,” the Graph Model provides a flexible decision technology for obtaining strategic insights with respect to this complex systems problem. At the start of this article, many of the key capabilities of this systems methodology are listed. For applications of the Graph Model to many different types of disputes in areas such as international trade

and environmental engineering, the reader can refer to Hipel *et al.* [26,27] and Hipel and Obeidi [28] as well as references cited therein.

Subsequent to carrying out a basic study similar to the one just done for the Aral Sea Basin conflict in the previous section, one may wish to execute further investigations to gain a deeper understanding about the conflict and additional strategic insights. Often this type of approach is referred to as a *sensitivity analysis*, in which parameters of a model are changed in a meaningful manner to see how the output is altered. If, for example, changes in the preferences of a given DM do not change the key equilibria, which have already been calculated, these equilibria are robust with respect to these preference changes.

As pointed out in the section titled “Systematic Investigation of Real-World Conflict,” after one ascertains the best that each DM can do on his or her own, one should also determine if DMs can fare even better through cooperation with others by forming coalitions [14,15]. In addition, one can check which equilibria can be reached from a given initial state or status quo through both independent and joint movement [29,30]. Zeng *et al.* [31] explain how the Graph Model can be extended for employment in policy analysis. Table 5 summarizes other expansions of the Graph Model that have been made, some of which are under further development. Finally, the matrix approach for equivalently defining the Graph Model for Conflict Resolution provides the systematic structure for formally incorporating all the foregoing advancements into the Graph Model as well as associated algorithms for designing the next generation of a decision support system [32,33].

REFERENCES

1. Kilgour DM, Hipel KW, Fang L. The graph model for conflicts. *Automatica* 1987;23(1): 41–55.
2. Fang L, Hipel KW, Kilgour DM. Interactive decision making: the graph model for conflict resolution. New York: John Wiley & Sons, Inc.; 1993.
3. Howard N. Paradoxes of rationality: theory of metagames and political behavior. Cambridge (MA): MIT Press; 1971.
4. Fraser NM, Hipel KW. Solving complex conflicts. *IEEE Trans Syst Man Cybern* 1979;9(12):805–816.
5. Fraser NM, Hipel KW. Conflict analysis: models and resolutions. New York: North-Holland Publishing Co.; 1984.
6. Howard N. Drama theory and its relation to game theory. Part 1: dramatic resolution vs. rational solution. *Group Decis Negot* 1994;3(2):187–206.
7. Howard N. Drama theory and its relation to game theory. Part 2: formal model of the resolution process. *Group Decis Negot* 1994;3(2):207–235.
8. Howard N. Confrontation analysis: how to win operations other than war. Pentagon, Washington (DC): CCRP Publications; 1999.
9. Bryant J. The six dilemmas of collaboration: inter-organisational relationships as drama. Chichester: John Wiley & Sons, Ltd.; 2003.
10. Obeidi A, Hipel KW. Strategic and dilemma analyses of a water export conflict. *INFOR* 2005;43(3):247–270.
11. Sage AP. Decision support systems engineering. New York: John Wiley & Sons, Inc.; 1991.
12. Fang L, Hipel KW, Kilgour DM, *et al.* A decision support system for interactive decision making, Part I: model formulation. *IEEE Trans Syst Man Cybern C Appl Rev* 2003;33(1):42–55.
13. Fang L, Hipel KW, Kilgour DM, *et al.* A decision support system for interactive decision making, Part II: analysis and output interpretation. *IEEE Trans Syst Man Cybern C Appl Rev* 2003;33(1):56–66.
14. Kilgour DM, Hipel KW, Fang L, *et al.* Coalition analysis in group decision support. *Group Decis Negot* 2001;10(2):159–175.
15. Inohara T, Hipel KW. Coalition analysis in the graph model for conflict resolution. *Syst Eng* 2008;11(4):343–359.
16. Vinogradov S, Langford VPE. Managing transboundary water resources in the Aral Sea basin: in search of a solution. *Int J Glob Environ Issues* 2001;1(3/4):345–362.
17. Nandalal KWD, Hipel KW. Strategic decision support for resolving conflict over water sharing among countries along the Syr Darya river in the Aral Sea basin. *J Water Resour Plann Manage* 2007; 133(4):289–299.

18. Antipova E, Zyrynov D, McKinney D, *et al.* Optimization of Syr Darya water and energy uses. *Water Int* 2002;27(4):504–516.
19. ICG (International Crisis Group). Central Asia: water and conflict. ICG Asia Rep 2002; No.34, Brussels.
20. Weinthal E. State making and environmental cooperation: linking domestic and international politics in Central Asia. Cambridge (MA): MIT Press; 2002.
21. Nash JF. Equilibrium points in n -person games. *Proc Natl Acad Sci U S A* 1950;36(1): 48–49.
22. Nash JF. Non-cooperative games. *Ann Math* 1951;54(2):286–295.
23. Zagare FC. Limited-move equilibria in 2×2 games. *Theory Decis* 1984;16:1–19.
24. Kilgour DM. Anticipation and stability in two-person non-cooperative games. In: Ward MD, Luterbacher U, editors. *Dynamic models of international conflict*. Boulder (CO): Lynne Rienner Press; 1985. pp. 26–51.
25. Brams SJ, Wittman D. Nonmyopic equilibria in 2×2 games. *Confl Manage Peace Sci* 1981;6:39–62.
26. Hipel KW, Kilgour DM, Fang L, *et al.* Strategic decision support for the services industry. *IEEE Trans Eng Manage* 2001;48(3): 358–369.
27. Hipel KW, Fang L, Kilgour DM. Decision support systems in water resources and environmental management. *J Hydrol Eng* 2008;13(9):761–770.
28. Hipel KW, Obeidi A. Trade versus the environment: strategic settlement from a systems engineering perspective. *Syst Eng* 2005;8(3): 211–233.
29. Li KW, Kilgour DM, Hipel KW. Status quo analysis of the Flathead river conflict. *Water Resour Res* 2004;40(5): W05S03 (9 pages).
30. Li KW, Kilgour DM, Hipel KW. Status quo analysis in the graph model for conflict resolution. *J Oper Res Soc* 2005;56:699–707.
31. Zeng DZ, Fang L, Hipel KW, *et al.* Policy equilibrium and generalized metarationalities for multiple decision-maker conflicts. *IEEE Trans Syst Man Cybern A Syst Hum* 2007;37(4):456–463.
32. Xu H, Hipel KW, Kilgour DM. Matrix representation of solution concepts in multiple decision maker graph models. *IEEE Trans Syst Man Cybern A Syst Hum* 2009;39(1):96–108.
33. Xu H, Li KW, Kilgour DM, *et al.* A matrix-based approach to searching colored paths in a weighted colored multidigraph. *Appl Math Comput* 2009;215(1):353–366.
34. Hamouda L, Kilgour DM, Hipel KW. Strength of preference in the graph model for conflict resolution. *Group Decis Negot* 2004;13(5):449–462.
35. Hamouda L, Kilgour DM, Hipel KW. Strength of preference in graph models for multiple decision-maker conflicts. *Appl Math Comput* 2006;179(1):314–327.
36. Xu H, Hipel KW, Kilgour DM. Multiple levels of preference in interactive strategic decisions. *Discrete Appl Math* 2009;157(15):3300–3313.
37. Li KW, Hipel KW, Kilgour DM, *et al.* Preference uncertainty in the graph model for conflict resolution. *IEEE Trans Syst Man Cybern A Syst Hum* 2004;34(4):507–520.
38. Al-Mutairi MS, Hipel KW, Kamel MS. Fuzzy preferences in conflicts. *J Syst Sci Syst Eng* 2008;17(3):257–276.
39. Inohara T, Hipel KW, Walker S. Conflict analysis approaches for investigating attitudes and misperceptions in the war of 1812. *J Syst Sci Syst Eng* 2007;16(2):181–201.
40. Walker S, Hipel KW, Inohara T. Strategic decision making for improved environmental security: coalitions and attitudes. *J Syst Sci Syst Eng* 2009;18(4):461–476.
41. Obeidi A, Hipel KW, Kilgour DM. The role of emotions in envisioning outcomes in conflict analysis. *Group Decis Negot* 2005;14(6):481–500.
42. Obeidi A, Kilgour DM, Hipel KW. Perceptual stability analysis of a graph model system. *IEEE Trans Syst Man Cybern A Syst Hum* 2009;39(5):993–1006.
43. Bennett PG. Toward a theory of hypergames. *OMEGA* 1977;5:749–751.
44. Wang M, Hipel KW, Fraser NM. Modelling misperceptions in games. *Behav Sci* 1988;33(3):207–223.

THE HUNGARIAN OPERATIONS RESEARCH SOCIETY

TIBOR CSENCES
Department of Computational
Optimization Szeged,
University of Szeged,
Hungary

HISTORY OF THE SOCIETY

The Hungarian Operations Research Society (HORS) was founded in Budapest in 1991, shortly after major political changes in the country. By that time, Operations Research (OR) was already represented in main research and scientific bodies, such as the Operations Research Committee of the Hungarian Academy of Sciences (HAS), the Applied Mathematics Department of the János Bolyai Mathematical Society and the Society of Economic Modeling. Still, the founders felt that it was important to have an independent association for OR. This structure of professional society life exists since that time, and obviously these scientific bodies serve the interest of the OR community quite well. Since decades, our society organizes the biannual Hungarian Operations Research Conference in rotation with the last two. Also, HORS enjoys a fruitful collaboration with all the abovementioned scientific bodies.

The honorary president of HORS is András Prékopa (EURO Gold Medal laureate, Rutgers University and Eötvös Loránd University, Budapest), member of the HAS. Béla Martos (1920–2007) also acted as the honorary president of HORS. Three honorary members were elected: Jakob Krarup, Harold W. Kuhn, and Anders Lindquist.

HORS has about 90 members and each year, young people enter our society based on the recommendation of the members. The first president was Margit Ziermann (1991–1994, Budapest University

of Economy). The subsequent presidents of the society were Tamás Rapcsák (1994–1996 and 1998–2000, Computer and Automation Institute (CAI, of the HAS), Zsolt Harnos (1996–1998, University of Horticulture and Food Sciences, Budapest), Sándor Komlósi (2000–2002, University of Pécs), Tamás Szántai (2002–2004, Budapest University of Technology and Economics), László Szeidl (2004–2006, University of Pécs), and József Temesi (2006–2008, Corvinus University of Budapest). Earlier, the executive body was elected every second year. Since 2008, this period has been extended to a term of three years.

The society is also supported by institutional members. They are mostly OR-related departments of Hungarian universities and research institutes, and they represent the most important centers where OR is investigated and taught at the highest level:

CAI of the HAS;
Department of Applications of Informatics, Juhász Gyula Teacher Training Faculty, University of Szeged;
Department of Applied Mathematics, University of Miskolc;
Department of Business Methodologies, University of Pécs;
Department of Mathematics, Corvinus University of Budapest;
Department of Mathematics and Informatics, Faculty of Horticultural Science, Corvinus University of Budapest;
Department of Operations Research, Corvinus University of Budapest;
Department of Operations Research, Eötvös Loránd University, Budapest;
Faculty of Information Technology, University of Pannonia;
Institute of Informatics, University of Szeged;
Institute of Mathematics, Budapest University of Technology and Economics;

John von Neumann Faculty of Informatics, Budapest Tech Polytechnical Institution.

HORS is a member of the international OR associations of EURO and INFORMS. We have an official collaboration agreement with the German OR Society, but we also enjoy good relations with OR societies of neighboring countries even in the absence of such official relations.

The present executive board of HORS is strengthened by the following people:

István Deák (vice-president, Corvinus University of Budapest),
Csanád Imreh (University of Szeged),
Tamás Kis (treasurer, CAI, HAS),
Tamás Szántai (Budapest University of Technology and Economics).

The web page of the HORS is http://www.opkut.hu/index_en.html, where further information is available on the members and on their publications.

ACTIVITIES AND ACHIEVEMENTS OF THE SOCIETY

Research

OR activities have a rich, over 50-year-old history in Hungary. Applied mathematics research, which provides to a certain extent the theoretical base for OR theory, dates back to the turn of the last century (Farkas theorem, 1901) and is connected later to the works of König, Egerváry, Wald, and von Neumann. Applications acquired more significance in the 1970s due to the wide introduction of computers.

In the year 2008, we had a workshop dedicated to the memory of Jenő Egerváry on the occasion of the fiftieth anniversary of his death. The nine talks highlighted his contribution to the present OR methodologies, especially to the Hungarian method for the solution of the assignment problem and its relatives. Among the presentations, one provided interesting details on how Harold W. Kuhn succeeded in translating the paper of

Egerváry (originally written in Hungarian). The translation was also discussed. Based on the outcomes of the scientific meeting, a special issue of the Central European Journal on Operations Research (CEJOR) is being edited, and will appear soon.

Education

At Eötvös Loránd University of Budapest (ELUB) students of mathematics have the opportunity to specialize in OR since 1968. This specialization was introduced and organized for many years by András Prékopa and his research group at CAI of HAS. The OR Department of ELUB was established in 1983. OR specialization is available for mathematics students at the Budapest University of Technology and Economy (BUTE), and also for computer science students at the University of Szeged. Since last year, they also have a Department of Computational Optimization.

OR and/or management science, decision theory, operational management, and industrial engineering are taught at the Faculties of Business and Economics of the Corvinus University of Budapest and the University of Pécs for students of economics and business. At the agricultural schools, OR is an integral part of the management courses, agricultural informatics, and agroeconomics. In education, the emphasis is on decision models and not on the theory of OR.

After changing the scientific qualification system in Hungary, the ELUB was the first to organize PhD courses in OR. These courses belong to a more general PhD program with the title of Applied Mathematics. The first two semesters of the program have half a dozen of students. This education form started in 1993. At the Corvinus University of Budapest, and the University of Pécs OR, management science and operational management are present in the PhD program. The University of Szeged had also an OR and combinatorial optimization doctoral subprogram. A member of HORS, Boglárka Gazdag-Tóth has won the UPS-SOLA Dissertation Award of INFORMS in 2007.

The Operations Research and the Informatics Committees of the HAS composed

recently a review on the OR doctoral education in Hungary. According to this study, several degrees are acquired in this field of science every year. Doctoral education is offered now in eight Hungarian universities. The HORS has an indirect role in reviewing, evaluating, and helping OR education on various levels.

PUBLICATIONS

The HORS is one of the professional associations responsible for the editorial tasks of the CEJOR. This scientific periodical became the main English language publication forum for our researchers. Several members of our society contribute to the success of this journal as editors, authors, or reviewers. HORS also contributes financially to the publishing of CEJOR, and as compensation, some of the major OR centers in Hungary obtain a copy of the journal regularly. Hungarian OR results related to the biannual Hungarian Operations Research Conferences are published in special issues of the CEJOR.

There are two other scientific journals, the *Alkalmazott Matematikai Lapok* and *Sigma*, that publish OR-oriented papers in Hungarian. These journals have some editorial board members from HORS. Still, they play an important role in the development of Hungarian language optimization terminology and in bringing the newest OR results to the wide Hungarian OR community.

SCIENTIFIC MEETINGS

Traditionally, we used to have two scientific meeting series in Hungary, one of which, the biannual Hungarian Operations Research Conference, still exists. We just had the

twenty-eighth conference this summer. It is organized by the János Bolyai Mathematical Society, the Society of Economic Modeling, and HORS in rotation. The official language of the meeting is Hungarian, with the exception of the talks held by foreign guests. For the past decade, the best papers emerging from the talks of these meeting have been published in special issues of the CEJOR.

The other traditional series of conferences were the international conferences on mathematical programming, held regularly in Mátrafüred. These were high-level biannual English language meetings with several respected invited speakers from all over the world. The last one, the fourteenth in a row was organized in 1999. The role of these meetings are taken over by the VOCAL conference series organized in Veszprém by the University of Pannonia with the collaboration of the HORS. The best papers related to the talks are published in special issues of the journals *Optimization and Engineering* and *Optimization Methods and Software*.

HORS is proud of the success of the EURO meeting in Budapest, 2000 chaired by András Prékopa.

HORS has a prize, the Egerváry Memorial Plaque, that is awarded for exceptional Hungarian contributions to OR. The first recipient was András Prékopa, followed by Emil Klafszky in 2004, Béla Martos and Pál Rózsa in 2005, György Mészéna in 2006, József Kolumbán in 2007, and Tamás Rapcsák (posthumous) in 2008.

The Hungarian researchers are strong in inventory control and production, linear, nonlinear, and stochastic programming, combinatorial optimization, LP software. They had successful applications in inventory control, power systems, water resources, micro and macro economics.

THE KNOWLEDGE GRADIENT FOR OPTIMAL LEARNING

WARREN B. POWELL

Department of Operations Research
and Financial Engineering,
Princeton University, Princeton,
New Jersey

INTRODUCTION

There is a wide range of problems in which we need to make a decision under some sort of uncertainty, and where we have the ability to collect information in some way to reduce this uncertainty. The problem is that collecting information can be time consuming and/or expensive, so we have to do this in an intelligent way. Some examples include the following:

- We wish to find a supplier for a component who provides the lowest cost and best level of service. We know the cost, but the only way to evaluate the level of service is to try the supplier and observe actual delivery times.
- We need to find the best path through New York City to respond to certain emergencies (e.g., getting a fire truck to city hall). To evaluate the time required to traverse each path, we can try each path, or we could collect information about specific links in the network. Which paths (or links) should we measure?
- We are trying to find the best molecular compound to produce a particular result (curing cancer, storing energy, conducting electricity, etc.). There are thousands of combinations that we can try, but each test takes a day. Which compounds should we test?
- We would like to find the best price to sell a product on the internet. We

can experiment with different prices and measure sales. How do we go about testing different prices?

- We have a simulation model of an operational problem (e.g., managing takeoffs and landings at Heathrow Airport). We have a number of control parameters that we can adjust to improve performance, but running a simulation takes half a day. How do we go about trying out different parameters?
- We are running a stochastic search algorithm to solve a specific optimization problem. Each iteration takes several minutes to perform a noisy function evaluation (this might be a simulation of the climate). How do we optimize this function as quickly as possible?

These are just a small number of examples of problems where we face the challenge of collecting information to make better decisions. Applications range from business decisions, science and engineering, and simulation and stochastic optimization.

The learning problems that we wish to address all have three fundamental components: a measurement decision, which determines what information we are going to collect; the information that we obtain; and an implementation decision which uses the information. We are primarily interested in sequential problems, where we can make multiple measurements and where each measurement may depend on the outcomes of previous measurements. Our goal, then, is to design a *measurement policy* which determines how these measurements are made.

Measurement problems come in two fundamental flavors. Off-line problems use a sequence of measurements before making a final implementation decision. These arise when we have a period of time to do research before choosing a final design. The second flavor is online problems, where we collect information each time we make an implementation decision. For example, we may

want to find the best path from our new apartment in New York City to our new job. The only way to evaluate a path is to try it. The implementation decision is the choice of path to use tomorrow, but we collect information as we try the path. Off-line problems are often grouped under labels such as ranking and selection (which generally refers to choosing among a finite set of alternatives), stochastic search, and simulation optimization. Online problems are usually referred to as *bandit problems*. Our presentation of the knowledge gradient (KG) treats both off-line and online problems in an integrated way.

Information collection problems can be approached from two perspectives: frequentist and Bayesian. In the frequentist approach, we form estimates about a truth based on information drawn from a sample (see [1–4] for excellent surveys of learning from a frequentist perspective). In this article, we focus on the Bayesian perspective, where we assume that we have a prior belief about problem parameters, along with estimates of the level of uncertainty. The Bayesian perspective assumes that there is a prior distribution of possible truths, and our goal is to design a policy that discovers this truth as quickly as possible. The frequentist perspective uses noisy observations from an unknown truth to make statistical statements about this truth; the Bayesian perspective uses a prior belief and noisy measurements to create a probability distribution to describe the truth.

In this short article, we adopt a Bayesian perspective, focusing on a concept we call the knowledge gradient, which belongs to a family of methods that guides the learning process based on the marginal value of information. The roots of this idea in the field of decision theory belong to the seminal paper by Howard [5] on the value of information. We first became aware of this idea applied as a policy for off-line ranking and selection problems from Gupta and Miescke [6]. A separate line of research that uses arose in the context of optimizing unknown functions under the general name “global optimization,” with roots in the seminal paper by Kushner [7]. When these functions

can only be measured with uncertainty, this field has been referred to as *Bayesian global optimization*, with many contributions (see Jones *et al.* [8] for a good review). Our work in this area is based on Frazier and Powell [9], which analyzes the properties of sequential measurement policies based on the marginal value of a single measurement using a Bayesian framework with variance known; Frazier and Powell [10] provide an important generalization to problems with correlated beliefs (reviewed below). In parallel research, Chick and Branke [11] derive the marginal value of information for the important case where the variance is unknown using the name $LL(1)$ (linear loss, with measurement batches of one observation).

We provide a brief summary of different policies for collecting information. Although measurement policies may have a Bayesian or frequentist foundation, it is possible (and, we would argue, most natural) to evaluate any policy by sampling a truth from a prior distribution, and then determine how well a policy learns this assumed truth. This exercise then has to be repeated over many possible truths drawn from the prior.

The essential feature that separates learning problems from traditional stochastic optimization is that we are unsure about our uncertainty. If we make an observation, we are willing to update the probability distributions we use to describe uncertain parameters. Inserting this step of updating our beliefs after a measurement is made is typically referred to as *statistical learning*. When we actively make choices of what to measure, taking into account our willingness to update our beliefs, then this is optimal learning.

This article is intended to serve as a brief introduction to some of the important problem classes in optimal learning. We provide a mathematical framework for formulating and evaluating measurement policies. We review a number of heuristic policies, along with optimal policies for special problem classes. The remainder of the article focuses on the concept of the KG algorithm, which is a measurement policy that applies to a wide range of applications.

ELEMENTARY PROBLEMS

There are two elementary learning problems that provide the foundation for our discussion. Both have a discrete set of measurements $\mathcal{X} = (1, 2, \dots, M)$, where M is an integer that is “not too large,” which means that we do not have any difficulty in enumerating the choices. The first problem is known in the literature as the *multiarmed-bandit problem*, which is an online learning problem where we learn from the rewards that we receive. The second is known as the *ranking and selection problem*, where we have a budget of N measurements to evaluate each choice, after which we have to decide which alternative appears to be best.

Multiarmed Bandit Problems

The multiarmed bandit problem is based on the story of trying to choose the best of a set of M slot machines (often known as *one-armed bandits*). We do not know how much we will win each time we play a particular slot machine, but we have a distribution of belief, which we acknowledge may be wrong. We may think one machine has the highest expected reward, but we are willing to acknowledge that we may be wrong and another machine may also be the best. The only way we will learn is to try machines that do not appear to be the best. But while trying these machines, we may be earning lower rewards than we would earn by playing the machines that we think are better. The goal is to maximize the expected discounted sum of rewards that balance what we earn against what we learn (to improve future decisions).

Let μ_x be the true mean reward if we choose x . We do not know μ_x , but assume that we believe that it is normally distributed with prior mean μ_x^0 and variance $(\sigma_x^0)^2$. For convenience, we define the precision $\beta_x^0 = 1/(\sigma_x^0)^2$. Let W^n be the reward (“winnings”) we receive in the n th iteration, and let (μ^n, β^n) be our vector of beliefs about the means and precisions for all the choices after n measurements. We can write $\mu_x^n = \mathbb{E}^n \mu_x$, where $\mathbb{E}^n$ is the expectation given the first n measurements. Let $S^n = (\mu^n, \beta^n)$ be our “state of knowledge” (often referred to as the *belief state*).

We let x^n be the choice we make after n measurements, meaning that our first choice is x^0 , which is based purely on the prior. We make these measurements using a policy π , which is allowed to depend on the history of observations $W^1, W^2, \dots, W^n$. Let $X^{\pi, n}(S^n)$ be the random variable representing the decision we make, given our state S^n , and given measurement policy π . This notation allows our policy to depend on n ; if we wish to follow a stationary policy, we would write it as $X^{\pi, n}(S^n)$ (in a finite-horizon problem, the policy can depend on the number of measurements n , as well as on the belief state S^n). Our goal is to find a measurement policy π that solves

$$\sup_{\pi} F^{\pi} = \mathbb{E}^{\pi} \sum_{n=0}^N \gamma^n \mu_{X^{\pi, n}(S^n)}, \quad (1)$$

where γ is a discount factor. We write the expectation $\mathbb{E}^{\pi}$ as dependent on the policy π , which reflects assumptions on how we construct the underlying probability space. We assume that an elementary outcome is a sequence of decisions of what to measure (which depend on the policy) and the results of a particular measurement. This is not the only way to construct the probability space, but it is the one that is most often used in the research community.

In the classical multiarmed bandit problem, $N = \infty$ and $\gamma < 1$, but the finite horizon problem (possibly with $\gamma = 1$) is also of interest.

Ranking and Selection Problems

Now imagine that we have a budget of N measurements (or B dollars to spend on measurements) after which we have to choose the best of a set of M alternatives. In the bandit problem, we learn as we go, incurring rewards (or costs) as we proceed. In the ranking and selection problem, we are not concerned with how well our choices perform during the process of collecting information. Instead, we are only concerned with how well our final choice performs.

Let μ_x^N be the posterior mean of the value of alternative x after N measurements, which we chose using measurement policy π . μ_x^N is a

random variable whose distribution depends on our measurement policy. The value of using π can be written

$$F^\pi = \mathbb{E} \max_x [\mu_x^N] = \mathbb{E}^\pi \mu_{x^\pi}^N,$$

where $x^\pi = \arg \max_x \mu_x^N$ is the optimal solution when we follow measurement policy π . The problem of finding the best policy is then given by

$$\sup_{\pi \in \Pi} F^\pi.$$

We note that it is possible to write the objective as

$$\sup_{\pi} \mathbb{E} F^\pi,$$

in which case we assume that the measurement policy is built into the objective function. When we write the objective this way, it means that the probability space consists of all potential measurements W_x^n for all alternatives x and all measurements n . Alternatively, we can write the objective as

$$\sup_{\pi} \mathbb{E}^\pi F.$$

Written this way, it means that we have imbedded the policy into the probability space (a sample outcome consists of measurement alternatives and realizations), which means that F does not explicitly depend on the policy.

Notes on Objective Functions

The formulations we have given in this section assume that we are maximizing a reward. Equivalently, we could minimize the expected opportunity cost (EOC). For the ranking and selection problem, this would be written as

$$EOC = \mathbb{E} \max_x \mu_x - \mathbb{E} \mu_{x^\pi},$$

where we are measuring the value that we could achieve if we could find the best alternative using the true means, versus the value of the alternative we did choose (but again using the true means). Keep in

mind that the expectations here are over both the distribution of truths as well as the distribution over measurements. Minimizing the EOC is the same as solving

$$\sup_{\pi} \mathbb{E} \mu_{x^\pi}.$$

It is also possible to show that

$$\sup_{\pi} \sup_{\chi} \mathbb{E}^\pi \mu_{\chi(x)} = \sup_{\pi} \mathbb{E}^\pi \max_x \mu_x^N, \quad (2)$$

where χ is the policy for choosing the best alternative x given what we measure, which for our problem is defined by

$$\chi = \arg \max_x \mu_x^N.$$

Equation (2) states that if we use a measurement policy to find the best alternative x^π and we evaluate this choice using the true values μ_x , we get the same answer if we evaluate our choice using the estimates μ_x^N .

LEARNING

At the heart of any learning problem is not only uncertainty about the value of the choices we are making but also uncertainty about our uncertainty. Learning problems are easily posed in a Bayesian framework, where we are able to capture the uncertainty in our belief about a system. In our bandit or ranking and selection problems, μ_x is the true value of x , but we do not know this value. Instead, we assign a probability distribution that describes what we think μ_x is for each x . Before we start collecting any information, we might assume that our *prior* distribution of belief about μ_x is normally distributed with mean μ_x^0 and variance $(\sigma_x^0)^2$. We adopt a common convention in Bayesian analysis and define the *precision* of our belief as $\beta_x^0 = 1/(\sigma_x^0)^2$. Now assume that when we make an observation W_x^n of choice x , the precision of this measurement is known and given by β_ϵ . To reduce notational clutter, we assume that this is constant across x , but this is easily relaxed.

Assume that our current (prior) belief about choice x is given by μ_x^n and β_x^n , and

that we choose to measure alternative $x^n = x$. We would then observe W_x^{n+1} , which is unknown when we picked x^n . Further assume that our prior is normally distributed (i.e., $\mu_x \sim N(\mu_x^0, 1/\beta_x^0)$), and that $W_x^{n+1} = \mu_x + \varepsilon_x^{n+1}$, where ε_x^{n+1} is also normally distributed with precision β_ϵ . Bayes theorem can be used to show that our updated mean and precision of the posterior belief can be computed using

$$\mu_x^{n+1} = \frac{\beta_x^n \mu_x^n + \beta_\epsilon W_x^{n+1}}{\beta_x^n + \beta_\epsilon}, \quad (3)$$

$$\beta_x^{n+1} = \beta_x^n + \beta_\epsilon. \quad (4)$$

We also have the property that, if our prior belief about μ is normally distributed, then the posterior belief is normally distributed.

MEASUREMENT POLICIES

Our central challenge is designing a policy for collecting information. For off-line problems, these policies guide information collection that determines the final decision. For on-line problems, our policy has to strike a balance between receiving rewards (or incurring costs) and collecting information that will help future decisions.

Measurement policies can be deterministic or sequential. If we are using a deterministic policy, decisions about what to measure are made before any measurements are learned. Deterministic policies can be optimal for certain types of statistical learning problems. Our interest is in sequential problems, where the next measurement decision is made only after learning the value of the previous measurement.

We begin by showing that optimal measurement policies can be characterized using a simple dynamic programming formulation. The problem is that the dynamic program cannot be solved exactly. We then describe a series of simple heuristic policies that are often used in the research community.

Optimal Sequential Policies

Dynamic programming is widely used for sequential decision problems. For our problem, we start in a (knowledge) state S^n ,

then we take an action (measurement) x^n , and observe a random outcome W_x^{n+1} , which takes us to a new state according to the transition function $S^{n+1} = S^M(S^n, x^n, W_x^{n+1})$ defined by Equations (3) and (4). Our optimal measurement x^n can, in theory, be characterized using Bellman's equation

$$V(S^n) = \max_x (C(S^n, x) + \gamma \mathbb{E} \{ V(S^{n+1}) | S^n \}). \quad (5)$$

Here, $V(S^n)$ captures the value of being in a particular knowledge state S^n . Using the principle of dynamic programming, this is given by choosing the measurement action x that maximizes the contribution earned (this might be a negative cost) from taking action x , plus the value of being in the knowledge state S^{n+1} that results from this action, given by $V(S^{n+1})$. Of course, S^{n+1} is random given S^n and x , so we have to take its expectation over the measurement W_x^{n+1} for each alternative x . For off-line problems, $C(S^n, x)$ may be zero, since we do not receive any value until we pick the final design. For online problems, $C(S^n, x)$ would be the expected reward μ_x^n .

So, if Equation (5) gives us the optimal measurement policy, why do we not just use this solution? The reason is that we just do not have algorithms to solve dynamic programs when the state variable is a vector of continuous parameters. There is, however, a special case that can be solved optimally without directly solving the dynamic program, which uses the concept of Gittins indices for multiarmed bandit problems. After presenting this idea, we review the most popular heuristic policies. Our presentation focuses purely on sequential policies where the choice of the n th measurement depends on the prior and the outcomes of $W^1, \dots, W^n$. We then close by introducing the KG policy and illustrate the wide range of information collection problems that can be addressed using this strategy.

Gittins Indices for Multiarmed Bandit Problems

In 1974, Gittins and Jones [12] found that the multiarmed bandit problem (specifically, the infinite horizon version of the multiarmed

bandit problem) could be solved optimally using an index policy computed using

$$\Gamma(\mu_x^n, \sigma_x^n, \sigma_\epsilon, \gamma) = \mu_x^n + \Gamma\left(0, \frac{\sigma_x^n}{\sigma_\epsilon}, 1, \gamma\right) \sigma_\epsilon. \quad (6)$$

In this expression, $\Gamma(\mu_x^n, \sigma_x^n, \sigma_\epsilon, \gamma)$ is the Gittins index for measurement x for which the belief is μ_x^n and the standard deviation is σ_x^n . $\sigma_\epsilon = \sqrt{1/\beta_\epsilon}$ is the standard deviation of the measurement and γ is the discount factor. The Gittins index policy, as it became known in the literature, specifies that you measure the choice x with the highest value of $\Gamma(\mu_x^n, \sigma_x^n, \sigma_\epsilon, \gamma)$. The power of this policy is that there is one index per choice, and at no time are we dealing with multidimensional state vectors.

This leaves us with the challenge of computing $\Gamma(\mu_x^n, \sigma_x^n, \sigma_\epsilon, \gamma)$. Equation (6) shows us that we can compute the index as a function of a simpler index $\Gamma\left(0, \frac{\sigma_x^n}{\sigma_\epsilon}, 1, \gamma\right)$. This has a close parallel with the standard normal distribution (indeed, this relationship depends on normally distributed beliefs). While this relationship simplifies the problem, we still have to find $\Gamma\left(0, \frac{\sigma_x^n}{\sigma_\epsilon}, 1, \gamma\right)$, which requires the fairly difficult numerical solution of an integral equation.

Fortunately, there have been recent efforts to develop numerical approximations. Brezzi and Lai [13] found that

$$\Gamma(0, s, 1, \gamma) = \sqrt{-\log \gamma} \, b\left(-\frac{s^2}{\log \gamma}\right), \quad (7)$$

where the function $b(\cdot)$ has to be approximated. Chick and Gans [14] improved on the initial approximation of this function by Brezzi and Lai [13], proposing

$$b(\zeta) \approx \begin{cases} \frac{\zeta}{\sqrt{2}} & \zeta \leq \frac{1}{7} \\ \exp\left(\frac{-0.02645(\log \zeta)^2}{+0.89106 \log \zeta - 0.4873}\right) & \frac{1}{7} < \zeta \leq 100 \\ \sqrt{\zeta} \sqrt{\zeta \log \zeta - \log \log \zeta - \log 16\pi} & \zeta > 100. \end{cases} \quad (8)$$

The Gittins index policy is specifically designed for on-line problems, and in particular infinite horizon bandit problems. However, the structure of the policy offers

some insights into the information-collection process, as we see below.

Heuristic Policies

There are a number of simple policies that can be used for collecting information. The policies reviewed below are the most popular. These can generally be adapted for on-line or off-line applications.

- *Pure Exploration.* Pure exploration involves picking choices at random. If there are M choices, we might choose $x^n = x$ with probability $1/M$ or use some other exogenously driven process. A pure exploration strategy makes little sense in an on-line application, but it can be useful (if not optimal) for off-line problems, especially for high-dimensional applications where the measurement space $\mathcal{X}$ is quite large. For off-line problems, we may use pure exploration just to collect data to fit a statistical model.
- *Pure Exploitation.* We make the decision that appears to be best, given what we know. Stated mathematically, a pure exploitation policy would be written as

$$x^n = \arg \max_{x \in \mathcal{X}} \mu_x^n.$$

Pure exploitation can be effective for online learning where the prior information is quite good.

- *Mixed Exploration and Exploitation.* Here we explore with some probability ρ , and exploit with probability $1 - \rho$. A variation is epsilon-greedy exploration, where we explore with probability $\rho^n = c/n$, where c is a tunable parameter.
- *Boltzmann Exploration.* Here we explore with probability

$$\rho_x^n = \frac{\exp(\theta \mu_x^n)}{\sum_{x' \in \mathcal{X}} \exp(\theta \mu_{x'}^n)},$$

where θ is a tunable parameter. $\theta = 0$ produces pure exploration, while as θ increases, it approaches pure

exploitation. This policy explores in a more intelligent way than the more classical exploration policies.

- *Interval Estimation.* Interval estimation is the most sophisticated of this group of heuristics. Here, we compute an index using

$$v_x^n = \mu_x^n + z_\alpha \bar{\sigma}_x^n. \quad (9)$$

In this policy, z_α is a tunable parameter, and $\bar{\sigma}_x^n$ is the standard deviation of our estimate μ_x^n . This policy strikes a balance between the estimate μ_x^n (low values are less likely to be tested), and the uncertainty with which we know μ_x^n . The policy rewards higher levels of uncertainty, but only if μ_x^n is competitive.

It is useful to compare the index policy given by interval estimation in Equation (9) with the Gittins index policy, given in Equation (6). Both policies compute an index given by the current belief plus an additional “bonus” (often referred to as the *uncertainty bonus*). How this bonus is computed is where the two policies differ. The Gittins policy uses a theoretically derived factor that declines to zero as the number of observations increase. This is multiplied by the standard deviation σ_ϵ of the measurement error. For interval estimation, the bonus is a constant factor (which has to be tuned) times the standard deviation $\bar{\sigma}_x^n$ of our estimate μ_x^n , which then declines with the number of observations. As the number of times we measure an alternative increases, the factor $\Gamma\left(0, \frac{\sigma_\epsilon^n}{\sigma_\epsilon}, 1, \gamma\right)$ in Equation (6) decreases (σ_ϵ stays the same). By contrast, with interval estimation, it is $\bar{\sigma}_x^n$ that decreases, while z_α stays the same.

Policies from Simulation Optimization

An entire body of research has developed around the problem of choosing the best set of parameters to guide a simulation. If these parameters are discrete, and if we fix the length of the simulation, this problem falls under the umbrella of ranking and selection if we are considering a finite number of alternatives. Simulation optimization

introduces the additional dimension that we can choose the length of the simulation, but we may face a budget on the total computing time. This problem was first addressed under the name of “optimal computing budget allocation” (OCBA) by Chen [15]. This idea has subsequently been studied in a number of papers [16–19]. Chick and Inoue [20] introduces the *LL(B)* strategy, which maximizes the linear loss with measurement budget B . He and Chick [21] introduce an OCBA procedure for optimizing the expected value of a chosen design, using the Bonferroni inequality to approximate the objective function for a single stage. A common strategy in simulation is to test different parameters using the same set of random numbers to reduce the variance of the comparisons. Fu *et al.* [22] apply the OCBA concept to measurements using common random numbers.

There is also a substantial body of literature that is often grouped under the heading of stochastic search, which addresses problems where we are searching for the best of a continuous set of parameters. For a thorough review of this field, see Spall [23]. We note, however, that this field covers numerous algorithms such as stochastic approximation methods that we would not classify as falling under the heading of optimal learning.

EVALUATING POLICIES

When faced with an array of different learning policies, we have to address the problem of evaluating policies. Using the setting of online problems, we start by observing that we generally cannot compute the expectation in Equation (1) exactly, but this equation hints at how we might evaluate a policy in practice. Let ω index both a prior $\mu(\omega)$ and a sequence of observations of $W^1(\omega), W^2(\omega), \dots, W^N(\omega)$, which depend on the prior, where

$$W_x^n(\omega) = \mu_x(\omega) + \epsilon_x^n(\omega).$$

Recall that $X^{\pi,n}(S^n(\omega))$ is our decision rule (policy) for choosing an alternative to test given our state of knowledge S^n . Given a sample path ω , a sample realization of a measurement policy would be computed using

$$F^\pi(\omega) = \sum_{n=0}^N \gamma^n \mu_{X^\pi, n(S^n(\omega))}(\omega). \quad (10)$$

Finally, we can repeat this simulation K times to form a statistical estimate of the value of a policy

$$\bar{F}^\pi = \frac{1}{K} \sum_{k=1}^K \sum_{n=0}^N \gamma^n \mu_{X^\pi, n(S^n(\omega^k))}(\omega^k). \quad (11)$$

Now we use standard statistical tools to compare policies.

We view this method of evaluating policies as being fundamentally Bayesian, since we depend on a prior to determine the set of truths. However, the policy we are testing could be frequentist. A frequentist testing strategy might involve simulating a policy (Bayesian or frequentist) on a single function, where we start by assuming that we have no knowledge of the function (we can treat it initially as a constant), and where measurements of the function are noisy. While this strategy is not uncommon, it introduces dangers. For example, it may be possible to choose a policy that works well on a particular function. This problem can be mitigated by running the algorithm on a family of functions, but then this is comparable to generating a truth (the set of truths would be given by the family of functions).

THE KNOWLEDGE GRADIENT POLICY

A simple idea for guiding information collection is to choose to measure the alternative that provides the greatest value from a single measurement. This idea was first introduced for ranking and selection problems by Gupta and Miescke [6] as the $(R_1, \dots, R_1)$ procedure. It has since been studied in greater depth as the knowledge gradient by Frazier and Powell [9] for problems where the measurement noise is known, and by Chick and Branke [11] under the name of $LL(1)$ (linear loss with batch size 1) for the case where the measurement noise is unknown. The idea was recently applied to on-line problems [24] for multiarmed bandit problems with both independent and correlated beliefs.

The “breakthrough,” if it can be called this, with the idea of the knowledge gradient is the growing amount of empirical evidence that it appears to work well, even when competing against optimal policies such as Gittins indices for classical multiarmed bandit problems. This seems surprising, given that optimal learning problems can be formulated as dynamic programs. As a general statement, myopic policies often work poorly for dynamic problems, and the KG policy is effectively a myopic heuristic. The insight that our numerical work appears to be showing is that, while this expectation is certainly true in the context of dynamic problems that arise in the management of physical resources, it does not appear to hold true in the context of learning problems.

The Knowledge Gradient for Off-Line Learning

As before, we assume that all beliefs are normally distributed with parameters captured by the state $S^n = (\mu_x^n, \beta_x^n)_{x \in \mathcal{X}}$. Given S^n , which means, given our beliefs about each choice μ_x^n , the value of our current state of knowledge is given by

$$V^n(S^n) = \max_{x' \in \mathcal{X}} \mu_{x'}^n.$$

Now assume that we choose to measure $x^n = x$. This means that we get to observe W_x^{n+1} and update our belief about μ_x . We write this updating process using our transition equation $S^{n+1}(x) = S^M(S^n, x^n, W^{n+1})$. This means applying the Bayesian updating Equations (3) and (4), but only for choice x . With this new information, the value of our new state of knowledge S^{n+1} would be given by

$$V^{n+1}(S^{n+1}(x)) = \max_{x' \in \mathcal{X}} \mu_{x'}^{n+1}.$$

At iteration n , however, the observation W^{n+1} is a random variable, which means that $V^{n+1}(S^{n+1}(x))$ is a random variable. We can compute the expected value of measuring x as

$$v_x^{\text{KG}, n} = \mathbb{E} [V^{n+1}(S^{n+1}(x)) - V^n(S^n) | S^n]. \quad (12)$$

We refer to $v_x^{\text{KG}, n}$ as the KG since it is the marginal value of information from

Table 1. Calculations Illustrating the Knowledge Gradient Index

Measurement	μ	$\bar{\sigma}$	β	$\tilde{\sigma}$	ζ	$f(\zeta)$	$v_x^{\text{KG},n}$
1	20.0	18.00	0.0031	17.999	-0.444	0.215	3.878
2	22.0	12.00	0.0069	11.998	-0.500	0.198	2.373
3	24.0	25.00	0.0016	24.999	-0.160	0.324	8.101
4	26.0	12.00	0.0069	11.998	-0.167	0.321	3.853
5	28.0	16.00	0.0039	15.999	-0.125	0.340	5.432

measuring x . Note that the KG policy not only captures what we learn about the mean of a choice but also the change in the precision of our belief. We write the KG policy as choosing to measure the alternative x which has the highest marginal value of information. We write this policy simply as

$$X^{\text{KG},n} = \arg \max_{x \in \mathcal{X}} v_x^{\text{KG},n}. \quad (13)$$

We start by computing the variance of the change in our estimate of μ_x given our state of knowledge S^n , given by

$$\tilde{\sigma}_x^{2,n} = \text{Var}[\mu_x^{n+1} - \mu_x^n | S^n].$$

It is fairly straightforward to show that

$$\begin{aligned} \tilde{\sigma}_x^{2,n} &= \bar{\sigma}_x^{2,n} - \bar{\sigma}_x^{2,n+1}, \\ &= \frac{(\bar{\sigma}_x^{2,n})}{1 + \sigma_\epsilon^2 / \bar{\sigma}_x^{2,n}}. \end{aligned} \quad (14)$$

We then compute the distance between our current estimate of the value of x , and the best of the rest, normalized by the number of standard deviations in the measurement of the change, given by

$$\zeta_x^n = - \left| \frac{\mu_x^n - \max_{x' \neq x} \mu_{x'}^n}{\tilde{\sigma}_x^n} \right|.$$

Next, we use a standard formula for $\mathbb{E} \max\{0, Z + \zeta\}$, where Z is the standard normal deviate. This formula is given by

$$f(\zeta) = \zeta \Phi(\zeta) + \phi(\zeta),$$

where $\Phi(\zeta)$ and $\phi(\zeta)$ are respectively the cumulative standard normal distribution and the standard normal density. Finally, the KG is given by

$$v_x^{\text{KG},n} = \tilde{\sigma}_x^n f(\zeta_x^n).$$

For a more detailed development of these equations, see [9]. Table 1 provides an illustration of the calculations for a simple problem with five alternatives.

On-Line Learning

The ranking and selection problem has a long history that has evolved completely independently of the multiarmed bandit problem. However, we can take the basic idea of the KG policy and apply it to the multiarmed bandit problem. We have to find the expected value of a single measurement. For on-line problems, we approach this by assuming that we are going to do a single step from which we can learn and update our beliefs about the problem. After that, we continue making decisions holding our state of knowledge constant.

If we have an on-line learning problem where we have only two trials, this means we can learn from the first trial and use it in the second trial. The value of measuring x would simply be $v_x^{\text{KG},n}$, which measures the improvement in our ability to solve the problem in the second trial.

Now imagine that we have N trials, where we have already made $n - 1$ measurements and are thinking about what we should do for the n th measurement. Since this is an on-line problem, the expected reward we will receive by measuring x is μ_x^n . In addition, we get the benefit of using this information $N - n$ more times in later decisions. Using this reasoning, we can find the on-line knowledge gradient (OLKG) using

$$v_x^{\text{OLKG},n} = \mu_x^n + (N - n)v_x^{\text{KG},n}.$$

If we have an infinite horizon problem (as with the standard multiarmed bandit problem) with discount factor γ , the OLKG policy would be

$$v_x^{\text{OLKG},n} = \mu_x^n + \frac{\gamma}{1-\gamma} v_x^{\text{KG},n}.$$

Again note the comparisons against the Gittins index policy and the interval estimation policy, both of which use the expected reward from a single measurement, and a term that in some way captures the value of information on future decisions (although the OLKG policy is the only one of these policies where this is explicit). Unlike these other two policies, on-line KG is not a true index policy, since $v_x^{\text{KG},n}$ depends on a calculation that requires information from all the other choices. The Gittins and IE index policies compute an index that is based purely on parameters from a particular choice.

This idea is developed with much greater rigor in Ryzhov *et al.* [24], which also reports on comparisons between the OLKG policy and other policies. It was found that the KG policy actually outperforms Gittins indices even when applied to multiarmed bandit problems (where Gittins is known to be *optimal*) when we use the approximation in Equation (8). KG also outperforms pure exploitation (i.e., for on-line problems), but it slightly underperforms interval estimation when the parameter z_α is carefully tuned. However, it was also found that IE is sensitive to this tuning, and even slight deviations from the best value of z_α can produce results that underperform KG.

The S-Curve Problem

An issue with the KG is that the value of information can be nonconcave. It is not hard to create problems where a single measurement adds very little value, since the noise in the observation is too high to change a decision. We can easily find the value of information by computing the KG where the precision of a measurement x is given by $n_x \beta_\epsilon$ rather than just β_ϵ . Think of it as making a single measurement with precision $n_x \beta_\epsilon$ instead of β_ϵ . An example of a curve where the value of information is nonconcave is given in Fig. 1.

The issue of the nonconcavity of information is well known. Howard [5] demonstrates the nonconcavity of information in the context of an auction problem. Radner and Stiglitz [25] discusses the problem in

considerably greater depth, giving general conditions under which the value of information is not concave. This research has been extended more recently by Chade and Schlee [26] and Delara and Gilotte [27] to more general settings.

The issue is discussed in Frazier and Powell [28] in the context of sequential sampling. This paper proposes a modification of the KG policy, which is dubbed KG(*). In this policy, we find the value of n_x , which maximizes the average value of information, as depicted in Fig. 1. We note that in a budget-constrained setting, we may not have the budget to do n_x measurements, in which case we have to consider the value of a smaller number of measurements. Further, we may not wish to use our entire budget on a single alternative, which may eliminate alternatives that require a large number of measurements to evaluate.

Correlated Beliefs

The KG policy is able to handle the very important case of correlated beliefs. This arises when measuring x may tell us something about x' . For example, imagine that we are testing the set of features in a product, or we want to identify the best set of drugs to put into a drug treatment, or we want to evaluate the best set of energy saving technologies in a house. These are all instances of subset (or portfolio) selection problems, where a choice x might be given by

$$x = (0, 1, 1, 0, 1, 0, 0, 0).$$

Now consider a choice x' given by

$$x' = (0, 1, 1, 0, 0, 1, 0, 0).$$

x and x' share two common members, so it might be reasonable to expect that, if x performs better than expected, that x' will perform better than expected. Let Σ^0 be our initial estimate of the covariance in our belief between different alternatives, where Σ^0 has $|\mathcal{X}|$ rows and columns. Also, let $B^n = (\Sigma^n)^{-1}$ be the inverse of the covariance matrix (think of this as the matrix version of our precision). Assume that we measure $x^n = x$, observing W_x^{n+1} . Recalling that μ^n is our vector of

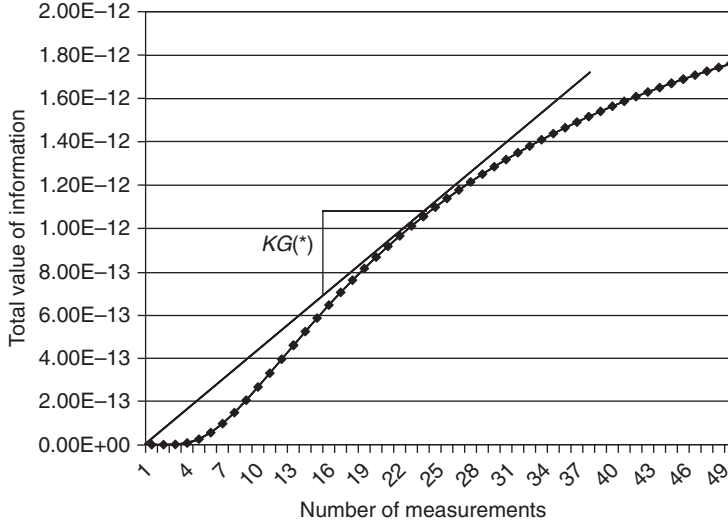


Figure 1. Illustration of the nonconcavity of information, and the maximum average value that would be used by the KG(*) algorithm.

beliefs, let $\mu^{n+1}(x)$ be the updated vector of beliefs assuming we measure $x = x^n$, and let $\Sigma^{n+1}(x)$ be the updated matrix of covariances. The Bayesian formulas for updating means and covariances, in the presence of correlated beliefs, are given by

$$\begin{aligned}\mu^{n+1}(x) &= \Sigma^{n+1}(x) \left(B^n \mu^n + \beta_\epsilon W_x^{n+1} e_x \right), \\ B^{n+1}(x) &= B^n + \beta_\epsilon e_x (e_x)^T,\end{aligned}$$

where e_x is a column vector of 0s with a 1 in the element corresponding to alternative x .

For our purposes, there is a more convenient way of expressing the updating of the vector μ^n . First, let $\tilde{\Sigma}^n(x)$ be the change in the covariance matrix for μ^n due to measuring alternative x , just as $\tilde{\sigma}_x^{2,n}$ was the change in the variance of μ_x^n due to measuring x . This is given by

$$\begin{aligned}\tilde{\Sigma}^n(x) &= \Sigma^n - \Sigma^{n+1}(x) \\ &= \frac{\Sigma^n e_x (e_x)^T}{\sqrt{\Sigma_{xx}^n + \lambda W}}.\end{aligned}$$

Next let $\tilde{\sigma}^n(x)$ be the column vector of $\tilde{\Sigma}^n(x)$ corresponding to alternative x , given by

$$\tilde{\sigma}^n(x) := \tilde{\Sigma}^n(x) e_x. \quad (15)$$

Also, let $\tilde{\sigma}_i^n(x) = (e_i)^T \tilde{\sigma}^n(x)$ be the i th component of the vector $\tilde{\sigma}^n(x)$.

Let $\text{Var}^n[\cdot] = \text{Var}[\cdot | S^n]$ be the variance given all the measurements up through iteration n . We observe that we can write

$$\begin{aligned}\text{Var}^n [W_x^{n+1} - \mu_x^n] &= \text{Var}^n [W_x^{n+1}] \\ &= \text{Var}^n [\mu_x + \varepsilon_x^{n+1}] \\ &= \Sigma_{xx}^n + \sigma_\epsilon^2,\end{aligned}$$

where the last step follows from the independence between the belief about the variance of μ_x conditioned on S^n , and the noise in the $n + 1$ st measurement ε_x^{n+1} . Now let

$$Z^{n+1} := (W^{n+1} - \mu^n) / \sqrt{\text{Var}^n [W^{n+1} - \mu^n]},$$

where $\text{Var}^n [W^{n+1} - \mu^n] = \Sigma_{xx}^n + \sigma_\epsilon^2$. It is straightforward to show that Z^{n+1} is normally distributed with mean 0 and variance 1. We can now write the updating equation using

$$\mu^{n+1} = \mu^n + \tilde{\sigma}^n(x^n) Z^{n+1}. \quad (16)$$

Also, it is easy to see that $\text{Var} Z^{n+1} = 1$ (since we constructed it that way).

The KG policy for correlated measurements is computed using

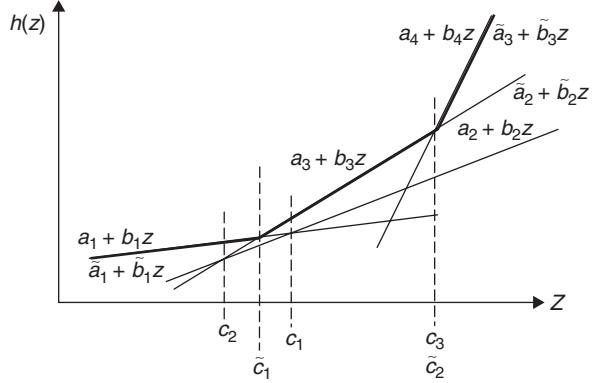


Figure 2. Regions of z over which different choices dominate. Choice 2 is always dominated.

$$\begin{aligned} \nu_x^{\text{KG}} &= \mathbb{E} \left[\max_i \mu_i^{n+1} \mid S^n = s, x^n = x \right] \\ &= \mathbb{E} \left[\max_i \mu_i^n + \tilde{\sigma}_i^n(x^n) Z^{n+1} \mid S^n, x^n = x \right], \end{aligned} \quad (17)$$

where Z is a scalar random normal variate.

Our challenge now is computing this expectation, which is much harder for correlated beliefs, but the computation is very manageable. We start by defining

$$\begin{aligned} h(\mu^n, \tilde{\sigma}^n(x)) &= \mathbb{E} \left[\max_i \mu_i^n + \tilde{\sigma}_i^n(x^n) Z^{n+1} \mid S^n, x^n = x \right]. \end{aligned} \quad (18)$$

Substituting Equation (18) into Equation (17) gives us

$$X^{\text{KG}}(s) = \arg \max_x h(\mu^n, \tilde{\sigma}^n(x)). \quad (19)$$

Let $h(a, b) = \mathbb{E}(\max_i a_i + b_i Z)$, where $a_i = \mu_i^n$, $b_i = \tilde{\sigma}_i^n(x^n)$ and Z is our standard normal variate. a and b are M (or $|\mathcal{X}|$)-dimensional vectors of means and $\tilde{\sigma}$'s. Now sort the elements of the vector b so that $b_1 \leq b_2 \leq \dots$, giving us a sequence of lines with increasing slopes. Looking at the lines $a_i + b_i z$ and $a_{i+1} + b_{i+1} z$, we find they intersect at

$$z = c_i = \frac{a_i - a_{i+1}}{b_{i+1} - b_i}.$$

Assume that $b_{i+1} > b_i$. If $c_{i-1} < c_i < c_{i+1}$, then there should be a range for z over which a

particular choice dominates, as depicted in Fig. 2. It is possible that a choice is always dominated, which will happen if $c_{i+1} < c_i$ (see choice 2 in the Fig. 2). If this happens, we simply drop the choice from the set, producing a new set of choices with coefficients $\tilde{a}_i, \tilde{b}_i$ and intersections $\tilde{c}_i$, which satisfy $c_{i-1} < c_i < c_{i+1}$.

Once we have sorted the slopes and dropped dominated alternatives, we can compute Equation (17) using

$$h(a, b) = \sum_{i=1}^M (b_{i+1} - b_i) f(-|c_i|),$$

where, as before, $f(z) = z\Phi(z) + \phi(z)$. Note that the summation over alternatives has to be adjusted to skip any choices i that were found to be dominated. For a more detailed presentation, see Frazier *et al.* [10].

The ability to handle correlated beliefs is a major generalization, and one which is not possible with the other heuristic search rules or Gittins indices. There are numerous practical problems where these correlations might be expected to arise. For example, if we are measuring a continuous function, we would expect our beliefs about x and x' to be correlated in proportion to the distance between x and x' . Correlated beliefs make it possible to tackle problems where the measurement budget is potentially much smaller than the number of potential measurements.

There is a limited but growing literature on bandit problems (online learning problems) where there is structure among the alternatives. Berry and Fristedt [29], Chapter 2 consider the case of correlated

beliefs, but do not provide a method to solve it. Pandey *et al.* [30] introduces the notion of dependent arms where alternatives can be grouped into clusters of arms with similar behavior. The choice is then reduced to one of choosing among clusters. Presman and Sonin [31] introduces the idea of bandit problems with hidden state variables, which creates dependencies among different alternatives. There is a small but growing literature on problems where correlations are introduced by assuming an underlying functional form for the belief [32].

Finally, there are a number of papers that deal with the important special case of continuous functions. Ginebra and Clayton [33] studies “response surface bandits” where the measurements are continuous; they introduce a simple heuristic (basically an uncertainty bonus similar to Gittins indices) for learning a surface. Agrawal [34] considers learning a multidimensional surface, but this is accomplished by sampling the function at equally spaced points, and then combining the results to produce an estimate of the function using kernel regression [35]. Kleinberg [35] improves on this by sampling over evenly spaced intervals, but then applying a standard bandit algorithm to the resulting finite set of points (there is no evidence that correlations are being used in the choice of what to measure).

Other Applications of the Knowledge Gradient

Now that we have shown that we can use the KG concept for on-line and off-line problems, as well as the important class of correlated beliefs, it is natural to ask the following question: what is the scope of applications of the KG for guiding information collection? We have only begun to explore the potential range of applications, but some examples include the following:

- *Learning on Graphs.* Consider the problem of making measurements so that we can better learn the costs on an uncertain graph. For example, we need to find the best path for an emergency response vehicle, and we have to send people to measure the costs on specific links of the network to produce the best

solution to our shortest path problem. Ryzhov and Powell [36] shows that we can identify the best link to evaluate using a method that is very similar to the KG for independent beliefs.

- *Drug Discovery.* Negoescu *et al.* [37] consider the problem of finding the best molecular compound, through intelligent sequencing of laboratory experiments. The research used a linear regression model to capture beliefs, reducing a problem with 87,000 possible compounds to one involving only 200 regression parameters. The KG policy was able to identify a compound very close to the best in under 100 experiments.
- *Subset Selection.* Ryzhov and Powell [38] shows how the KG can be used to guide the selection of the best set of energy-saving technologies to implement in a building, recognizing that different technologies interact in different ways for a particular building.
- *Optimizing General Surfaces.* The KG for correlated beliefs presented in this article requires that the covariance matrix Σ^0 be known. In Mes *et al.* [39], the KG concept is adapted to a problem for optimizing general functions, which may be nonconcave and which can be a function of discrete or categorical variables. The method uses statistical representation where the function is approximated using a weighted sum of estimates at different levels of aggregation. The KG is proven to find the optimum of the function in the limit, and appears to show good empirical convergence.

The KG has other useful theoretical properties. Frazier and Powell [9] shows that the KG policy is optimal (for any measurement budget) if there are only two choices, and that, for off-line problems, the policy is asymptotically optimal, which means that in the limit it will find the best alternative. Of course, many heuristic policies are asymptotically optimal (again, this applies only to off-line problems), but the KG algorithm is the only stationary policy that is both myopically optimal

(which KG is by construction) and asymptotically optimal. This may be a reason why it appears to offer good empirical performance for intermediate measurement budgets.

The KG does not always work. It might consistently work well if the value of information is concave in the measurement budget, but this is not always the case. The marginal value of information may initially be quite small, when a single measurement is not enough to change a decision. But it may happen that, after some number of measurements, the marginal value starts to grow. This might arise, for example, if we are trying to find the best hitter for a baseball team. Watching a hitter for a few at bats teaches us almost nothing. It can take several hundred at bats before we can confidently say that one hitter is better than another. The value of information can then follow an S-curve, which means that it can be highly nonconvex for some applications.

The KG offers the type of general strategy for collecting information that steepest ascent has offered for a wide range of nonlinear maximization problems. It may not work best for all applications, but it may provide a valuable starting point for a wide range of problems.

REFERENCES

1. Bechhofer RE, Santner TJ, Goldsman DM. Design and analysis of experiments for statistical selection, screening, and multiple comparisons. New York: John Wiley & Sons, Inc.; 1995.
2. Kim S-H, Nelson BL. Selecting the best system. Amsterdam: Elsevier; 2006. pp. 501–534, Chapter 17.
3. Goldsman DM, Nelson BL, Kim S-H. Statistical selection of the best system. Proceedings of the Winter Simulation Conference, 2005. Piscataway (NJ): IEEE Press; 2005. pp. 188–201. Available at <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1574251>
4. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
5. Howard RA. Information value theory. IEEE Trans Syst Sci Cybern 1966;2(1):22–26. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4082064
6. Gupta S, Miescke K. Bayesian look ahead one-stage sampling allocations for selection of the best population. J Statistical Planning and Inference 1996;54(2):229–244. Available at <http://linkinghub.elsevier.com/retrieve/pii/S0378375895001697>
7. Kushner HJ. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. J Basic Eng 1964;86:97–106.
8. Jones D, Schonlau M, Welch W. Efficient global optimization of expensive black-box functions. J Glob Optim 1998;13(4):455–492. Available at <http://www.springerlink.com/index/M5878111M101017P.pdf>
9. Frazier PI, Powell WB, Dayanik S. A knowledge gradient policy for sequential information collection. SIAM J Control and Optimization 2008;47(5):2410–2439.
10. Frazier PI, Powell WB, Dayanik S. The knowledge-gradient policy for correlated normal beliefs. INFORMS J. on Computing 2009;21(4):599–613. Available at <http://joc.journal.informs.org/cgi/doi/10.1287/ijoc.1080.0314>
11. Chick SE, Branke J, Schmidt C. Sequential sampling to myopically maximize the expected value of information. INFORMS J. on Computing 2010;22(1):71–80. Available at <http://joc.journal.informs.org/cgi/doi/10.1287/ijoc.1090.0327>
12. Gittins J, Jones D. A dynamic allocation index for the sequential design of experiments. Amsterdam: North Holland; 1974. pp. 241–266.
13. Brezzi M, Lai T. Optimal learning and experimentation in bandit problems. J Econ Dyn Control 2002;27(1):87–108. Available at <http://linkinghub.elsevier.com/retrieve/pii/S0165188901000288>
14. Chick SE, Gans N. Economic analysis of simulation selection problems. Manage Sci 2009;55(3):421–437.
15. Chen CH. An effective approach to smartly allocate computing budget for discrete event simulation. 34th IEEE Conference on Decision and Control, Volume 34, New Orleans (LA). pp. 2598–2603.
16. Yucsan E, Chen CH, Dai L, et al. A gradient approach for smartly allocating computing budget for discrete event simulation. In: Charnes J, Morrice D, Brunner D,

- et al.*, editors. Proceedings of the 1996 Winter Simulation Conference. Piscataway (NJ): IEEE Press; 1997. pp. 398–405.
17. Yucesan E, Chen HC, Chen CH, *et al.* New development of optimal computing budget allocation for discrete event simulation. Proceedings of the Winter Simulation Conference; Atlanta (GA): 1997. pp. 334–341.
 18. Chick SE, Chen CH, Lin J, *et al.* Simulation budget allocation for further enhancing the efficiency of ordinal optimization. Discrete Event Dyn Syst 2000;10:251–270.
 19. Chen HC, Chen CH, Yucesan E. Computing efforts allocation for ordinal optimization and discrete event simulation. IEEE Trans Automat Control 2000;45(5):960–964.
 20. Chick SE, Inoue K. New two-stage and sequential procedures for selecting the best simulated system. Oper Res 2001;49(5):732–743. Available at <http://www.jstor.org/stable/3088571>
 21. He D, Chick SE, Chen C-H. Opportunity cost and OCBA selection procedures in ordinal optimization for a fixed number of alternative systems. IEEE Trans Syst Man Cybern Part C Appl Rev 2007;37(5):951–961.
 22. Fu MC, Hu J-Q, Chen C-H, *et al.* Simulation allocation for determining the best design in the presence of correlated sampling. INFORMS J Comput 2007;19:101–111.
 23. Spall JC. Introduction to stochastic search and optimization: estimation, simulation and control. Hoboken (NJ): John Wiley & Sons, Inc.; 2003.
 24. Ryzhov IO, Powell WB, Frazier PI. The knowledge gradient algorithm for a general class of online learning problems; 2009. Accessed 2009. Available at <http://www.castlelab.princeton.edu/Papers/RyzhovPowellFrazier-OnlineKnowledgeGradientSept112009.pdf>
 25. Radner R, Stiglitz J. Volume 5, A nonconcavity in the value of information. Amsterdam: North-Holland Publishing Co.; 1984. chapter 3. pp. 33–52. Available at <http://pages.stern.nyu.edu/~rradner/publishedpapers/50Nonconcavity.pdf>
 26. Chade H, Schlee EE. Another look at the radner–stiglitz nonconcavity in the value of information. J Econ Theory 2002;107(2): 421–452. Available at <http://linkinghub.elsevier.com/retrieve/pii/S0022053101929606>
 27. Delara M, Gilotte L. A tight sufficient condition for Radner–Stiglitz nonconcavity in the value of information. J Econ Theory 2007; 137(1):696–708. Available at <http://linkinghub.elsevier.com/retrieve/pii/S0022053107000373>
 28. Frazier PI, Powell WB. Paradoxes in learning: the marginal value of information and the problem of too many choices; 2010. Accessed 2010. Available at <http://www.castlelab.princeton.edu/Papers/FrazierScurve-Jan192010.pdf>
 29. Berry DA, Fristedt B. Bandit problems. London: Chapman and Hall; 1985.
 30. Pandey S, Chakrabarti D, Agarwal D. Multi-armed bandit problems with dependent arms. Proceedings of the 24th International Conference on Machine Learning. Corvallis (OR): ACM; 2007. pp. 721–728. Available at <http://portal.acm.org/citation.cfm?id=1273496.1273587>
 31. Presman EL, Sonin IN. Sequential control with incomplete information: the Bayesian approach to multi-armed bandit problems. New York: Academic Press; 1990.
 32. Mersereau A, Rusmevichientong P, Tsitsiklis J. A structured multiarmed bandit problem and the greedy policy. In: Medova-Dempster EA, Dempster MAH, editors. 47th IEEE Conference on Decision and Control. Piscataway (NJ): IEEE Press; CDC 2008. pp. 4945–4950.
 33. Ginebra J, Clayton MK. Response surface bandits. J R Stat Soc Ser B (Methodological) 1995;57:771–784.
 34. Agrawal R. The continuum-armed bandit problem. SIAM J Control Optim 1995;33(6):19–26. Available at <http://link.aip.org/link/?SJCODC/33/1926/1>
 35. Kleinberg R. Nearly tight bounds for the continuum-armed bandit problem. In: Saul L, Weiss Y, Bottou L, editors. Advances in neural information processing systems. Cambridge (MA): MIT Press; 2004. pp. 697–704.
 36. Ryzhov IO, Powell WB. Information collection on a graph. Oper Res 2010;1–35. Available at <http://www.castlelab.princeton.edu/Papers/RyzhovPowell-GraphPaperFeb102010.pdf>
 37. Negoescu DM, Frazier PI, Powell WB. The knowledge-gradient algorithm for sequencing experiments in drug discovery; 2010. Available at http://www.castlelab.princeton.edu/Papers/Negoescu-OptimalLearningDrugDiscovery_August242009.pdf
 38. Ryzhov IO, Powell WB. A Monte-Carlo knowledge gradient method for learning abatement potential of emissions reduction technologies.

- In: Rossetti MD, Hill RR, Johansson B, *et al.*, editors. Proceedings of the 2009 Winter Simulation Conference. Piscataway (NJ): IEEE Press; 2009.
39. Mes MRK, Powell WB, Frazier PI. Hierarchical knowledge gradient for sequential sampling; 2009.

THE LAW AND ECONOMICS OF RISK REGULATION

CARY COGLIANESE
University of Pennsylvania Law
School, Philadelphia,
Pennsylvania

From the origins of the common law era to the present day regulatory state, governments have always used law as a means of risk management. Today law plays a central role in managing a variety of risks associated with the modern economy. The many laws governing financial institutions and investment practices seek to protect society from undesirable economic risks such as fraud or insolvency. Other laws regulate manufacturing operations in an attempt to protect workers from occupational risks and neighbors from environmental risks. Laws seek to protect consumers from injuries arising from unsafe products, travelers from transportation accidents, and all citizens from risks of criminal or terrorist violence.

Given law's important role in addressing risk in society, scholars have devoted considerable attention to risk regulation, with both normative and positive research. An initial normative question has centered on the justifications for governmental regulation of the marketplace [1]. Accepting the need for risk regulation, researchers have studied, as a positive matter, the process by which regulatory standards are established and implemented, as well as have evaluated whether alternative decision-making procedures achieve their intended objectives. A central substantive issue in risk regulation focuses on the proper level of risk protection that regulatory standards should strive to achieve. Whatever the level of appropriate risk protection, a final issue considers the form or design of risk regulatory instruments. Much research in recent decades on the regulation of health, safety, and environmental risks has focused on each of these issues in

an effort to contribute policy-relevant knowledge for those officials who seek to use law as means of risk management.

THE BASIS FOR RISK REGULATION

In theory, the common law of contracts, which allows the participants in market transactions to allocate economic risks between themselves, might initially be thought to be the only law needed for purposes of risk management. If individuals faced no transaction costs and possessed perfect information about the risks of different activities, they could presumably bargain with each other so that society's most valued economic activities would continue while taking optimal measures to protect those who would be harmed from such activities [2]. Those consumers who want to pay more for risk reduction can do so, and the market in principle should adjust to meet individuals' varied demands for safety. But "the difficulty here, of course, is that markets are far from perfect" (3, p. 82). Given the existence of transaction and information costs, contracting over risk does not occur in an optimal manner. As a result, even under the common law, judges recognized that other legal principles, namely, tort principles of nuisance and, more recently, of products liability were needed both to provide compensation to injured individuals and also to serve as a disincentive for socially suboptimal risky activity [4,5].

The threat of tort liability serves to deter risky behavior, especially given the size of some jury awards [6,7], and in this way tort law can be said to fulfill a "regulatory" function by inducing more optimal levels of care. Tort law in principle holds certain advantages as a regulatory mechanism [8]. Unlike the often tightly defined prescription of contemporary regulation, the open-ended standards of the common law (such as "unreasonable" harm) makes liability more readily adaptable to changing circumstances.

Rather than prospectively addressing sometimes speculative risks, common law liability applies only to harms that actually occur, and it also places the principal decision-making authority in the hands of judges who may be far less subject to interest group pressures than legislators or regulatory officials. Finally, when liability rules place no-fault responsibility on the least-cost avoider of the risk, this can induce firms to produce safer products and processes [9]. Insurance coverage can provide further incentives for risk reduction, because insurance companies can set premiums based on risk levels and condition coverage on compliance with basic safety principles [1].

On the other hand, some of these same purported advantages of tort law can be disadvantages [8]. For example, judges may well be more independent, but they are also much less knowledgeable about science and risk analysis than are experts within centralized regulatory agencies. More significantly, tort liability may provide less-than-optimal levels of deterrence for society. Despite their theoretical advantages for pricing risk, liability insurance markets may, like any market, operate imperfectly or may not cover all types of harms. The availability of bankruptcy may also blunt some of the deterrent effects of tort liability. Business decision makers may underestimate the probability that their activities will cause harm to others, thereby underestimating their liability for such harm [10,11]. The litigation process itself can be slow and costly, both for society as well as for plaintiffs seeking recovery, which may keep some number of meritorious cases from being pursued. The injured party bears the burden of proof, and especially when injured parties are poor, they may not pursue litigation at a level needed to create a socially optimal level of deterrence. In addition, when there are many injured parties (such as in cases of pollution or major industrial accidents), these individuals will face collective action problems in organizing to seek recovery [12]. If many individuals are harmed only slightly, the overall level of harm to society may still vastly exceed the overall benefits from the underlying economic activity, but each individual will have little reason to

pursue costly litigation. As a result of these limitations, the level of deterrence provided by tort liability will often prove to be much lower than is socially desirable [7].

For these kinds of reasons, regulation has been recognized as necessary to supplement, or even replace, common law liability [13]. Regulation—that is, imposing more specific, prospective risk reduction requirements—is preventive rather than reactive. Regulatory intervention in the marketplace has been generally viewed as justified in instances of market failure such as information asymmetries (consumer product risks or workplace health and safety risks) and negative externalities (environmental risks or major industrial or nuclear power accidents) [14,15]. As a result, although common law liability remains in place, the rate of regulation has grown dramatically over the last century and regulatory agencies, rather than courts, have today become the primary source of law affecting risk management.

THE PROCESS OF RISK REGULATION

Given the critical role of regulation, scholars have focused on how regulatory agencies make law. Although the normative case for regulation to address market failures has been generally accepted, this does not mean that all regulation is motivated by such public interested reasons nor that adequate levels of regulation will always be put in place. On the contrary, positive research on the process of regulation has suggested that risk regulation can depart from normative ideals and instead fall prey to the dictates of interest group and electoral politics [16]. Businesses that would incur costs under regulation can be expected to oppose the establishment of stringent regulation, and they are better organized in the policy process than are the beneficiaries of regulation, who may be numerous but who have only diffuse interests at stake [17]. As a result, government may undersupply public regulatory protection just as with common law protection. The same collective action problems that serve as a barrier to tort litigation in cases of diffuse and unorganized victims of pollution or other harms

may also dampen the pressure for regulatory responses to such harms.

The close, interactive relationships that arise between government and industry can potentially lead to regulatory capture, a situation where government protects an industry's interests at the expense of the broader public's interests [18,19]. In some cases, capture can be expected to lead to less regulation, or less stringent regulation, than would be optimal. In other cases, business interests actually may support stringent regulation in order to impose costs on new competitors [20,21]. Risk regulation will sometimes include "grandfather" clauses that exempt existing uses or otherwise apply more stringent control requirements on new sources of risk, which can impose cost barriers to the entry of competitors [22,23].

Despite the real pressures business organizations can exert on the regulatory process, the expansive growth of regulation over the past half-century cannot be easily squared with regulatory capture. If industry controlled the process of making regulation, developed countries presumably would not have witnessed as expansive a growth in regulation addressing consumer protection, workplace safety, and environmental protection [24]. As economies have developed and educational levels have increased, members of the public today are less tolerant of a variety of risks than were previous generations [25]. Industry, no doubt, remains a politically important force in regulatory policymaking, but the increased political demand for regulation, generally, along with the development of well-organized labor, environmental, and consumer rights movements, has led to the creation of new regulatory agencies and greater levels of regulatory control over the last fifty years [26,27]. Competition across different jurisdictions may also sometimes create a "race to the top" in regulatory protection, as politicians in states or nations with less stringent regulations try to emulate jurisdictions with more stringent regulations in order to satisfy the demands of an increasingly risk averse public [28].

Risk regulation in advanced economies like the United States has evolved through an interplay between the legislature and

the bureaucracy. Legislatures pass laws that typically provide general frameworks for risk management, leaving responsibility for implementing those general frameworks to regulatory agencies. For example, the US Clean Air Act instructs the Environmental Protection Agency (EPA) to set ambient air quality standards at a level "requisite to protect the public health" with "an adequate margin of safety," but leaves it up to the EPA to decide what that actual level should be for a given air pollutant [29]. This delegation of regulatory authority from the legislature to the agency, especially under the terms of broad statutory language like that in the Clean Air Act example, create the conditions for a principal-agent problem, that is, the likelihood that regulators will stray from what the legislature intended [30,31]. In a democracy, who or what controls the unelected bureaucrats who establish risk standards?

Two complementary theories have emerged in the study of US risk regulation, one that emphasizes control by the President and one that emphasizes ongoing control by the Congress. Presidents exert influence over regulatory agencies by appointing the individuals who run them, overseeing their budget requests, and reviewing their most significant regulatory proposals [32]. Congress can also continue to exert influence over regulatory agencies by calling their administrators to testify and, more significantly, using the appropriations process as a carrot or stick [33,34]. Researchers find evidence that both Presidents and Congress affect the behavior of regulatory agencies [35–38].

Some scholars emphasize the role of administrative procedure in affecting what bureaucrats do and in holding them accountable to the elected branches of government. For example, the US Administrative Procedure Act of 1946 requires agencies to issue public notices of proposed rules and to give the public an opportunity to comment on these proposals before enacting them. Procedural requirements like these can help ensure that the same interest groups that supported an initial legislative delegation to an agency will have continued influence over that agency's actions [39,40]. Interest groups

can serve as surrogate monitors for the legislature, pulling a “fire alarm” when an agency seems to stray too far from the direction the Congress intended [34]. At its extreme, the theory of procedural control posits that Congress uses administrative procedures to “stack the deck” in ways that ensure agencies will carry out legislative preferences [39,40].

To date, the development of the theory of procedural control seems to have outpaced empirical support for the strong, stack-the-deck version of the theory [41–43]. But scholarly interest in administrative procedure has much deeper roots than just the recent theory of procedural control. Administrative law scholars have long posited that procedures could be used to advance a variety of desirable qualities in regulatory policy [44]. For example, they have favored opportunities for judicial review in order to encourage greater adherence to law and to promote better policy reasoning [45]. They have urged the use of innovations such as negotiated rulemaking to reduce regulatory conflict and speed up regulatory decision making [46]. Both legal scholars and economists have advocated procedural requirements for economic analyses of proposed rules as a means of promoting more efficient regulatory outcomes [47]. A growing empirical literature tests the claims advocates make for various forms of administrative procedure, often finding mixed or even negative results [48].

Whether or not procedures achieve substantial benefits in terms of improved decision making, some scholars have expressed concern that the simple accretion of procedural requirements imposed on agencies has come to hamper their ability to enact needed regulatory protections. Some have posited that a “paralysis by analysis” or “ossification” now grips the US regulatory process [49–51]. As with the purported benefits of regulatory procedure, these purported costs are subject to empirical investigation. To date, however, no systematic research has confirmed fears that procedures have placed any discernible barrier in the way of new regulation [48,52–55].

Comparative research on risk regulation across different developed economies provides another way of assessing the

effects that procedural or other institutional structures have on regulatory outcomes. The management of risk varies widely across jurisdictions [56]. Do the structures of regulatory policy making, or just more conventional political factors such as political parties or voter ideologies, explain differences in risk regulation across jurisdictions? Some countries in Europe have more “corporatist” regulatory structures that rely on formal collaboration between selected industry groups, labor representatives, and government officials; others like the United States are much more “pluralist” in allowing interest groups to compete against each other in an open, adversarial manner [57]. Although early work suggested that corporatist regulatory structures might lead to greater environmental controls [58], it appears that the stringency of environmental regulation is not related to institutional structures but rather (and not surprisingly) to the political strength of green and left-libertarian political parties [59].

SETTING RISK STANDARDS

Regulatory stringency is the core substantive issue at stake in setting risk management standards. Put simply, the issue has often been characterized by the question: How safe is safe? Answering this question requires making a normative decision about how much risk members of the public should confront in their lives. Setting risk standards “necessarily requires the use of value judgments on such issues as the acceptability of risk and the reasonableness of the costs of control” [60].

In making these value judgments about risk, decision makers can adhere to one of four basic normative principles or approaches [61]:

1. eliminate all risk—or at least all man-made risk (the zero-risk approach);
2. reduce risk to an acceptable level (the acceptable risk approach);
3. reduce risk until the cost of doing so reaches an unacceptable level (the feasibility approach); and

4. balance the benefits of risk reduction with the costs (the efficiency approach).

The zero-risk approach appeals to an intuitive notion of safety that implies complete freedom from any potential harm. Some laws appear to require this approach, such as the US Clean Air Act's provisions calling for standards that "protect the public health" with "an adequate margin of safety." The European Union's precautionary principle could also be viewed as indicative of a risk elimination approach [62].

In the context of setting environmental risk standards, a zero-risk approach would require standards set at levels below some exposure threshold at which adverse health effects occur. If, as seems increasingly the case for many chemicals, a given pollutant is nonthreshold in that some adverse effects arise from even the lowest levels of exposure, a zero-risk approach may require the complete elimination of the economic activity that relies upon or generates the chemical, even if doing so would create substantial, negative economic consequences. It is also possible, of course, that eliminating one kind of risk will only exacerbate another kind of risk. A standard that requires the removal of harmful substances from automobile brake pads may protect autoworkers and mechanics but could very well increase the risks from automobile accidents. In cases of such "risk-risk trade-offs" [63], the zero-risk approach can at best become a *minimize-risk* approach under which the regulator sets a standard at the nonzero level at which the marginal adverse effects of an economic activity equals the marginal adverse effects from reducing or controlling that activity.

Rather than seeking to eliminate or minimize risk, a second principled approach would reduce risk to an *acceptable* level. For example, no matter how much transportation safety regulation exists, anything short of banning airplane flights will always leave some residual risk of injuries and fatalities from air travel. A transportation safety regulator could, nevertheless, act to reduce risks to an acceptable, even if nonzero, level. In other contexts, regulators have also followed this approach. The US Occupational

Safety and Health Administration (OSHA) uses a benchmark mortality level of 1 in 1000 as the basis for its occupational health standards and the US EPA has sometimes deemed individual mortality risks below 1 in 10,000 to be acceptable [61].

An acceptable risk approach has its limitations. For one, an acceptable risk approach needs benchmarks not just for mortality risks as EPA and OSHA have used, but also for morbidity risks. A reason that agencies have not developed as clear a set of benchmarks for morbidity risks is that it is much harder to determine what those appropriate benchmarks should be. Regulators must also grapple with the choice between individual risk and population risk, as even a tiny individual risk could lead to a large total public health effect if a vast proportion of the population is exposed to a risk [61]. Determining what level should be deemed "acceptable" will most likely call for difficult value judgments; it is far from self-evident whether 1 in 1000 or 1 in 10,000 is the right level. Moreover, an acceptable risk approach disregards costs altogether, implying that regulators must act to reduce risks to an "acceptable" level even when the costs of doing so are disproportionately (and unacceptably) high. Correspondingly, the acceptable risk approach directs regulators to avoid reducing risks below the acceptable level, even when the costs of doing so would be *de minimus*.

A third approach—the feasibility principle—emphasizes costs in much the same way that the acceptable risk approach emphasizes benefits. The feasibility principle directs regulators to tighten the stringency of risk standards to the point at which the costs of complying with the standard reaches an unacceptable level. For example, the US Congress has directed OSHA to develop toxic exposure standards that will protect workers from harm "to the extent feasible"—that is, up to the point at which costs become unacceptable [64]. Just as regulating to achieve levels of acceptable risk disregards costs, regulating to avoid levels of unacceptable costs disregards the benefits of risk control. A risk standard might be infeasible in the sense that it would force the shutdown of a major industry, but if doing so

would save tens of thousands of lives, such economic disruption is almost surely worth it. The US phase-out of lead as an additive in gasoline seems a good example where high costs, even the significant contraction of a major industry, were clearly justified in order to secure significant public health benefits [65].

Cost-benefit balancing, the fourth principle for setting risk management standards, avoids the problems of the acceptable risk and unacceptable cost approaches [66]. By taking both costs and benefits into consideration, the balancing principle directs regulators to set risk standards at the socially efficient level, that is, the level that maximizes net benefits (i.e., benefits minus costs). If tightening a risk standard will generate high costs and deliver only few benefits, it should not be adopted. On the other hand, even if only a few benefits come from making a standard incrementally more stringent, doing so will be justified if the incremental costs are even lower than the incremental benefits. In the United States, the efficiency principle has not only been favored by economists [67], but has also been embedded in an executive order applied to regulatory agencies since the Reagan administration. The current order, Executive Order 12,866, requires agencies to “propose or adopt a new regulation only upon a reasoned determination that the benefits of the intended regulation justify its costs” [68].

Those economists and legal scholars who favor cost-benefit balancing do so because they believe it can achieve the most risk reduction for the resources devoted to meeting risk standards [66,67,69]. Currently, the estimated net benefits of risk regulation vary markedly across, and within, policy areas. Some regulations have been estimated to cost only in the tens of thousands of dollars per life saved; others have been estimated to cost billions of dollars for each life saved [70].

Despite the attraction of cost-benefit balancing, many scholars and political officials have resisted its widespread application. Applying this approach requires converting the costs and benefits of risk regulation into a common metric, typically a monetary one. The monetization of benefits generates objections, in particular when these benefits

involve the protection of human life [71,72]. Of course, even if benefits are not explicitly monetized, when regulators choose between policy alternatives with greater and lesser levels of risk control they are implicitly making value judgments about health and life. Additional value judgments must be made when the costs and benefits of control are not equally distributed, as frequently is the case.

Applying cost-benefit balancing also raises a series of other choices over how to estimate benefits and costs. One choice is whether to rely on estimates of how much individuals are willing to accept to forego risk reduction versus how much they are willing to pay to secure the same reduction. Although theoretically these two choices should lead to comparable amounts, willingness-to-accept amounts have sometimes been higher than willingness-to-pay amounts for equivalent risks [73]. In practice, economists typically have relied on willingness-to-pay amounts.

Another choice centers on the methods for obtaining value estimates, primarily for benefits that do not normally have market prices associated with them. Regulatory analysts can try to extrapolate using revealed preference methods, such as those that use the differences in wages in jobs with different risk profiles to extrapolate a monetized value of a statistical life [74]. Alternatively, analysts can rely on stated preference methods, such as contingent valuation surveys that ask individuals how much they would be willing to pay (or accept) for certain levels of risk reductions (or risk exposures) [74].

A final choice in applying the efficiency principle involves the discount rate used to place all monetary estimates into present value [75]. Not all costs and benefits occur at the same time; regulations addressing health risks with long latency periods will incur immediate costs but will deliver benefits only on a longer time horizon. The choice of what discount rate to use may affect the determination of whether a new risk standard will yield positive net benefits, as higher discount rates will tend to make long-term benefits smaller in present value terms [76].

DESIGNING RISK REGULATORY INSTRUMENTS

In addition to determining risk standards' stringency, regulators also must choose the form or design of risk regulation. Many environmental, health, and safety risk standards are designed to direct regulated firms to adopt specific means of risk control, such as by mandating the use of protective devices on machinery or the installation of specific pollution control techniques. Other risk standards impose an obligation on firms to meet a stated level of performance, such as when they prohibit pollution that exceeds specified emissions limits. Such performance standards provide firms with more flexibility as each firm can choose its own means of meeting the required level of performance [77]. But both means-specific and performance standards—sometimes referred together to as *command-and-control* regulation—treat all firms equally, requiring each firm either to adopt the same type of risk control measures or achieve the same level of risk control. Since the costs of risk control can vary across firms, a one-size-fits-all approach can result in excessive overall costs from risk regulation.

A more cost-effective way of achieving the same level of aggregate risk reduction would have firms with lower costs of control reduce risk more than firms with higher costs of control. For this reason, market-based regulatory instruments have been long advocated, and occasionally implemented, because they allow firms to achieve different levels of control. Market-based instruments include taxes, ideally set equal to the costs that risky activity imposes on society [78], and tradable permit systems [79,80]. With tradable permits, the regulator determines an overall level of aggregate risk (or a level of a proxy for that risk), issues individual output permits that add up to the desired aggregate level, and then allows firms to trade their individual permits with each other. In the environmental area where this approach has been most prominently used, tradable permits have been thought to be better suited for risks that arise from mixable pollutants than for those that arise

from exposure to highly toxic chemicals. The flexibility tradable permits afford regulated entities can lead to concentrated “hot spots” with higher levels of pollution, even if overall levels decline. For risk problems where some variation in risky outputs is acceptable—because it is the overall level of such outputs that matters—market-based instruments can lower costs. With regulatory instruments like tradable permits, firms can have an incentive to make even greater reductions in risky behavior so they can sell their excess credits. Economists have analyzed the conditions for deploying either taxes or tradable permits [81,82], and a body of research on market-based risk regulation has developed as governments have experimented with its use [83,84]. The US government has successfully deployed tradable permits both to phase out the use of lead additives in gasoline [65,85,86] as well as to reduce emissions of sulfur dioxide from coal-powered electric utilities [87].

To work effectively, market-based instruments require government regulators to monitor and measure outputs. But some risks—or precursors to risks—are quite difficult or costly to monitor. This may be because (i) they arise only from relatively low-probability events, (ii) the causal pathways to the risk are complex, or (iii) a monitoring technology does not readily exist. In such circumstances, regulators have sometimes responded by adopting management-based regulations, mandating that firms themselves identify the risks posed from their own operations and develop internal policies and procedures to manage those risks [88,89]. A well-known example is the application of Hazard Analysis and Critical Control Point (HACCP) requirements on food-processing facilities by governments around the world. Under HACCP, food processors are required to identify critical points where contamination could arise in their production processes, and then to develop their own measures for preventing such contamination from arising and ensuring that their employees carry out these measures [90]. Management-based regulation has been demonstrated to achieve reductions [91]; however, especially when it is used in

situations where governmental monitoring is difficult, management-based regulation can never be said to guarantee that firms will invest the resources needed to make socially optimal risk reductions [92,93].

Much the same can be said of other alternative regulatory responses. Governments sometimes impose information disclosure requirements, either providing consumers or the government information about the risks associated with their products or processes. Such information-based strategies have long been a staple of financial regulation, but they also have been used in recent years to address health and safety risks in a variety of areas [94,95]. Information disclosure appears more likely to prove effective when it actually can affect decisions by investors or customers, giving rise to other (nonregulatory) pressures for risk reduction.

The US Toxic Release Inventory (TRI) program, adopted in the late 1980s, requires certain companies to disclose their emissions of toxic chemicals [96]. Some researchers have suggested that the pressures activated by TRI-reporting led firms to reduce their toxic emissions by substantial amounts [97,98]. Of course, these studies have not adequately accounted for the contributions of more conventional regulation to this decline, which is important to consider given that nearly contemporaneous amendments to the Clean Air Act imposed new regulatory controls on the same hazardous air pollutants that would be reported under TRI. For this reason, TRI's unique impact on toxic emissions remains, "to date, unknown" (96, p. 242).

More research will be needed to assess the effectiveness of innovative regulatory designs like market-based instruments, management-based regulation, and information disclosure. Evaluation research can be conducted on the existing application of these designs to risk problems as well as on their use to address new forms of risk. Governments are only likely to continue to experiment with these and perhaps new forms of regulatory instruments altogether in the face of pressing or emerging areas of risk ranging from global climate change to nanotechnology. Empirical evaluation of these new regulatory strategies should aim

to compare the outcomes achieved under innovative policies with a realistic appraisal of more conventional forms of regulation [99]. Conventional regulation is certainly not immune from outcomes that are ineffectual or even counterproductive. Regulation of any kind can lead to adaptations in behavior that offset the risk reductions intended to be achieved through regulation. If drivers in cars with seat belts or airbags end up driving faster [100,101], or if parents are more likely to leave medications with child-resistant packaging in areas accessible to their small children [102], protective regulations will not reduce risk as much as their designers may intend and sometimes they may even create or exacerbate other kinds of risks.

CONCLUSION

Regulation provides a central means for government to manage the risks associated with today's global, industrialized economy. Markets themselves cannot adequately protect the public from hidden hazards or spillover harms. *Ex post* liability, while useful, does not always by itself provide a socially optimal level of risk control. As such, preventative risk regulation will be needed, and the core questions will remain about how stringent should such regulation be, and what form should such regulation take. Risk regulation research will continue to be needed to provide conceptual clarity to the normative basis for risk standards as well as to generate better empirical evidence on the effectiveness, efficiency, and equity of applying particular regulatory instruments under varied risk conditions.

REFERENCES

1. Breyer S. Regulation and its reform. Cambridge (MA): Harvard University Press; 1982.
2. Coase R. The problem of social cost. *J Law Econ* 1960;3:1–44.
3. Viscusi WK. Toward a diminished role for tort liability: social insurance, government regulation, and contemporary risks to health and safety. *Yale J Regul* 1989;6:65.

4. Calabresi G. *The costs of accidents: a legal and economic analysis*. New Haven (CT): Yale University Press; 1970.
5. Morag-Levine N. *Chasing the wind: regulating air pollution in the common law state*. Princeton (NJ): Princeton University Press; 2003.
6. Rose-Ackerman S. Regulation and the law of torts. *Am Econ Rev* 1991;81(2):54–58.
7. Shavell S. *Economic analysis of accident law*. Cambridge (MA): Harvard University Press; 1987.
8. Bardach E, Kagan RA. *Going by the book: the problem of regulatory unreasonableness*. Philadelphia (PA): Temple University Press; 1982.
9. Calabresi G, Douglas Melamed A. Property rules, liability rules and inalienability: one view of the cathedral. *Harv Law Rev* 1972;85:1089.
10. Camerer C, Kunreuther H. *Making decisions about liability and insurance*. Norwell (MA): Kluwer Academic Publishers; 1993.
11. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. *Science* 1978;185:1124.
12. Olson M Jr. *The logic of collective action*. New York: Schocken Books; 1968.
13. Schroeder CH. Lost in the translation: what environmental regulation does that tort cannot duplicate. *Washburn Law J* 2002;41:583.
14. Stokey E, Zeckhauser RJ. *Primer for policy analysis*. New York: W W Norton & Co; 1980.
15. Viscusi WK, Vernon JM, Harrington JE. *Economics of regulation and antitrust*. 3rd ed. Cambridge: MIT Press; 2000.
16. Farber DA, Frickey PP. *Law and public choice: a critical introduction*. Chicago: University Chicago Press; 1991.
17. Wilson JQ. *The politics of regulation*. New York: Basic Books; 1980.
18. Huntington SP. The marasmus of ICC: the commission, the railroads, and the public interest. *Yale Law J* 1952;61:467–509.
19. Bernstein MH. *Regulating business by independent commission*. Princeton (NJ): Princeton University Press; 1955.
20. Stigler GJ. The theory of economic regulation. *Bell J Econ Manage Sci* 1971;2:3–21.
21. Peltzman S. Toward a more general theory of regulation. *J Law Econ* 1976;19:211–240.
22. Ackerman B, Hassler W. *Clean coal/dirty air*. New Haven (CT): Yale University Press; 1981.
23. Stavins RN. Vintage-differentiated environmental regulation. *Stanford Environ Law J* 2006;25(1):29–63.
24. Kamieniecki S. *Corporate America and environmental policy: how often does business get its way?* Palo Alto (CA): Stanford University Press; 2006.
25. Inglehart R. *Modernization and postmodernization: cultural, economic and political change in 43 societies*. Princeton (NJ): Princeton University Press; 1997.
26. Braithwaite J, Drahos P. *Global business regulation*. Cambridge: Cambridge University Press; 2000.
27. Coglianese Cary. Social movements, law, and society: the institutionalization of the environmental movement. *Univ PA Law Rev* 2001;150(1):85–118.
28. Vogel D. *Trading up: consumer and environmental regulation in a global economy*. Cambridge (MA): Harvard University Press; 1995.
29. Clean Air Act. 42 United States Code § 7409(b).
30. Eisner MA, Meier KJ. Presidential control versus bureaucratic power: explaining the reagan revolution in antitrust. *Am J Pol Sci* 1990;34:269–287.
31. Spence DB. Administrative law and agency policymaking: rethinking the positive theory of political control. *Yale J Regul* 1997; 14:407–450.
32. Moe T. An assessment of the positive theory of ‘congressional dominance’. *Legis Stud Q* 1987;12:475–520.
33. Weingast BR. The congressional-bureaucratic system: a principal-agent perspective. *Public Choice* 1984;44:147–191.
34. McCubbins M, Schwartz T. Congressional oversight overlooked: police patrols vs. fire alarms. *Am J Pol Sci* 1984;28(1):165–179.
35. Moe T. Regulatory performance and presidential administration. *Am J Pol Sci* 1982; 26:197–224.
36. Weingast BR, Moran M. Bureaucratic discretion or congressional control? Regulatory policymaking by the Federal Trade Commission. *J Polit Econ* 1983;91(5):765–800.
37. Wood BD, Waterman R. The dynamics of political control of the bureaucracy. *Am Polit Sci Rev* 1991;85:801.
38. Ringquist E. Political control and policy impact in EPA’s office of water quality. *Am J Pol Sci* 1995;39:336–363.

39. McCubbins M, Noll R, Weingast B. Administrative procedures as instruments of political control. *J Law Econ Organ* 1987;3:243–277.
40. McCubbins M, Noll R, Weingast B. Structure and process; politics and policy: administrative arrangements and the political control of agencies. *Va Law Rev* 1989;75:431.
41. Potoski M, Woods N. Designing state clean air agencies: administrative procedures and bureaucratic autonomy. *J Public Adm Res Theory* 2001;11(2):203–222.
42. Spence DB. Managing delegation ex ante: using law to steer administrative agencies. *J Legal Stud* 1999;28:413–459.
43. Balla SJ. Administrative procedures and political control of the bureaucracy. *Am Polit Sci Rev* 1998;92(3):663–673.
44. Landis J. *The administrative process*. New Haven (CT): Yale University Press; 1938.
45. Seidenfeld M. Demystifying deossification: rethinking recent proposals to modify judicial review of notice and comment rulemaking. *Tex Law Rev* 1997;75:483.
46. Harter PJ. Negotiating regulations: a cure for malaise. *Georgetown Law J* 1982;71:1.
47. Crandall RW, DeMuth C, Hahn RW, *et al.* *An Agenda for Federal Regulatory Reform*. Washington (DC): American Enterprise Institute and the Brookings Institution; 1997.
48. Coglianese C. Empirical analysis and administrative law. *Univ Ill Law Rev* 2002;2002(4):1111–1137.
49. Mendeloff J. *The dilemma of toxic substance regulation: how overregulation causes underregulation*. Cambridge (MA): The MIT Press; 1988.
50. McGarity TO. Some thoughts on ‘Deossifying’ the rulemaking process. *Duke Law J* 1992;41:1385.
51. Mashaw JL, Harfst DL. *The struggle for auto safety*. Cambridge (MA): Harvard University Press; 1991.
52. Balla SJ, Deets JM, Maltzman F. Outside communication and OMB review of agency regulations. Paper presented at the Annual Midwest Political Science Association Meeting; 2006; Chicago (IL).
53. O’Connell AJ. Political cycles of rulemaking: an empirical portrait of the modern administrative state. *Va Law Rev* 2008;94:889.
54. Yackee JW, Yackee SW. Administrative procedures and bureaucratic performance: is federal rulemaking “ossified”? *J Public Adm Res Theory* 2009;20(2):261–282.
55. Shapiro S. Speed bumps and roadblocks: procedural controls and regulatory change. *J Public Adm Res Theory* 2002;12(1):29–58.
56. Hammitt JK, Wiener JB, Swedlow B, *et al.* Precautionary regulation in Europe and the United States: a quantitative comparison. *Risk Anal* 2005;25(5):1215–1228.
57. Kagan RA. *Adversarial legalism: the American way of law*. Cambridge (MA): Harvard University Press; 2001.
58. Scruggs LA. Institutions and environmental performance in seventeen Western democracies. *Br J Polit Sci* 1999;29:1–31.
59. Neumayer E. Are left-wing party strength and corporatism good for the environment? Evidence from panel analysis of air pollution in OECD countries. *Ecol Econ* 2003;45:203–220.
60. National Academy of Sciences/National Research Council. *Risk Assessment in the Federal Government: Managing the Process*. Washington (DC): National Academy Press; 1983.
61. Coglianese C, Marchant GE. Shifting sands: the limits of science in setting risk standards. *Univ PA Law Rev* 2004;152:1255–1360.
62. Marchant GE, Mossman KL. *Arbitrary and capricious: the precautionary principle in the European Union Courts*. Washington (DC): American Enterprise Institute; 2004.
63. Graham JD, Wiener JB, editors. *Risk vs. risk: tradeoffs in protecting public health and the environment*. Cambridge (MA): Harvard University Press; 1997.
64. The Occupational Safety and Health Act of 1970, 29 United States Code § 6(b)(5).
65. Newell RG, Rogers K. Leaded gasoline in the United States: the breakthrough of permit trading. In: Harrington W, Morgenstern RD, editors. *Choosing environmental policy: comparing instruments and outcomes in the United States and Europe*. Washington (DC): Resources for the Future Press; 2004.
66. Sunstein CR. *The cost-benefit state: the future of regulatory protection*. Chicago: The American Bar Association; 2002.
67. Arrow KJ, Cropper ML, Eads GC, *et al.* Is there a role for benefit-cost analysis in environmental, health, and safety regulation? *Science* 1997;272:221–222.
68. Executive Order 12,866—Regulatory Planning and Review, *Federal Register* 58:51,

- 735-51,744 (Oct. 4, 1993)
69. Adler MD, Posner EA. *New foundations of cost-benefit analysis*. Cambridge (MA): Harvard University Press; 2006.
 70. Breyer SG. *Breaking the vicious circle: toward effective risk regulation*. Cambridge (MA): Harvard University Press; 1993.
 71. Ackerman F, Heinzerling L. *Priceless: on knowing the price of everything and the value of nothing*. New York: New Press; 2004.
 72. Kelman S. Cost-benefit analysis: an ethical critique. *AEI J Gov Soc Regul* 1981;5:33-40.
 73. Kahneman D, Knetsch JL, Thaler RH. Anomalies: the endowment effect, loss aversion, and status quo bias. *J Econ Perspect* 1991;3(1):192-206.
 74. U.S. Environmental Protection Agency. *Guidelines for Preparing Economic Analyses*; Sep 2000.
 75. Farber DA, Hemmersbaugh PA. The shadow of the future: discount rates, later generations, and the environment. *Vanderbilt Law Rev* 1993;46:267-304.
 76. Revesz R. Environmental regulation, cost-benefit analysis, and the discounting of human lives. *Columbia Law Rev* 1999; 99(4):941-1017.
 77. Coglianese C, Nash J, Olmstead T. Performance-based regulation: prospects and limitations in health, safety, and environmental protection. *Adm Law Rev* 2003;55 (4):705-729.
 78. Pigou AC. *The economics of welfare*. 4th ed. London: MacMillan and Co; 1932.
 79. Dales JH. *Pollution, property & prices: an essay in policy making and economics*. Toronto: University of Toronto Press; 1968.
 80. Tietenberg TH. *Emissions trading: an exercise in reforming pollution policy*. Washington (DC): Resources for the Future; 1985.
 81. Weitzman M. Prices vs. quantities. *Rev Econ Stud* 1974;41(4):477-491.
 82. Goulder LH, Parry IWH. Instrument choice in environmental policy. *Rev Environ Econ Policy* 2008;2(2):152-174.
 83. Cropper ML, Oates WE. Environmental economics: a survey. *J Econ Lit* 1992;30(2): 675-740.
 84. Stavins RN. Market-based environmental policies: what can we learn from U.S. experience (and Related Research)? In: Freeman J, Kolstad C, editors. *Moving to markets in environmental regulation: lessons from twenty years of experience*. New York: Oxford University Press; 2007. pp. 19-47.
 85. Nussbaum BD. Phasing down lead in gasoline in the U.S.: mandates, incentives, trading, and banking. In: Jones T, Corfee-Morlot J, editors. *Climate change: designing a tradeable permit system*. Paris: Organization For Economic Co-Operation and Development; 1991.
 86. Nichols AL. Lead in gasoline. In: Morgenstern RD, editor. *Economic analyses at EPA: assessing regulatory impact*. Washington (DC): Resources for the Future Press; 1997.
 87. Stavins RN. What can we learn from the grand policy experiment? Positive and normative lessons from SO₂ allowance trading. *J Econ Perspect* 1998;12:69-88.
 88. Coglianese C, Lazer D. Management-based regulation: prescribing private management to achieve public goals. *Law Soc Rev* 2003; 37:691-730.
 89. Braithwaite J. Enforced self regulation: a new strategy for corporate crime control. *Mich Law Rev* 1982;80:1466-1507.
 90. May PJ. Social regulation. In: Salamon L, editor. *Tools of government: a guide to the new governance*. New York: Oxford University Press; 2002.
 91. Benneer LS. Are management-based regulations effective? *J Policy Anal Manage* 2007;26.2:327-348.
 92. Fairman R, Yapp C. Enforced self-regulation, prescription, and conceptions of compliance within small businesses: the impact of enforcement. *Law Policy* 2005;27(4): 491-519.
 93. Gunningham N, Sinclair D. Organizational trust and the limits of management-based regulation. *Law Soc Rev* 2009;43(4): 865-899.
 94. Graham M. *Democracy by disclosure: the rise of technopopulism*. Washington (DC): Brookings Institution Press; 2002.
 95. Jin G, Leslie P. The effect of information on product quality: evidence from restaurant hygiene grade cards. *Q J Econ* 2003;118(2):409-451.
 96. Hamilton JT. *Regulation through revelation: the origin, politics, and impact of the toxics release inventory program*. Cambridge: Cambridge University Press; 2005.
 97. Konar S, Cohen M. *Information as regulation: the effect of community right-to-know*

- laws on toxic emissions. *J Environ Econ Manage* 1997;32:109.
98. Fung A, O'Rourke D. Reinventing environmental regulation from the grassroots up: explaining and expanding the success of the toxics release inventory. *Environ Manage* 2000;25(2):115–127.
 99. Coglianese C, Snyder BL. Program evaluation of environmental policies: toward evidence-based decision making. National research council, social and behavioral science research priorities for environmental decision making. Washington (DC): National Academies Press; 2005. pp. 246–273.
 100. Lave LB, Webber W. A benefit-cost analysis of auto safety features. *Appl Econ* 1970;2: 265–275.
 101. Peltzman S. The effects of automobile safety regulation. *J Polit Econ* 1975;83(4):677–726.
 102. Viscusi WK. The lulling effect: the impact of child-resistant packaging on aspirin and analgesic ingestions. *Am Econ Rev* 1984; 74(2):324–327.

THE $M/G/1$ QUEUE

ONNO J. BOXMA
Department of Mathematics,
Eindhoven University of
Technology, Eindhoven,
The Netherlands

A.K. Erlang's 1909 paper "The theory of probabilities and telephone conversations" [1] may be viewed as the birth of queueing theory. In the century following this publication, queueing theory has developed into a strong branch of both applied probability and stochastic operations research. Its development has, time and again, been fueled by new developments in communications, and it has also frequently been inspired by problems posed by production and computer networks. The sustained success of queueing theory is to a considerable extent due to the fact that there are a few very natural, basic, models which turn up all the time in new practical problems, but usually with new aspects or extensions which make it challenging to study them. These basic models are the Erlang loss model (see *The M/G/s/s Queue*), the Jackson network [see *Jackson Networks (Open and Closed)*] and the $M/G/1$ queue, which is the subject of the present contribution. They have been studied extensively, are very well understood, and are taught in courses all over the world. They have also served as testing grounds for new methods.

Accordingly, our contribution is also methodologically oriented; we shall discuss several methods to study key performance measures for the $M/G/1$ queue, like queue length, waiting and sojourn time, and busy period. There are other performance measures worthy of consideration. One example is the cycle maximum, that is, the maximum workload during a busy period. Methods for that quantity are discussed in Ref. 2.

We restrict ourselves to steady-state metrics, although the transient behavior of the $M/G/1$ queue is also an important topic.

THE $M/G/1$ MODEL

The $M/G/1$ queue is a single-server queue. Customers arrive at the service facility according to a Poisson process, with rate λ , requiring a particular amount of service. Let B_i denote the service requirement of the i th arriving customer. It is assumed that $B_1, B_2, \dots$ are independent, identically distributed (i.i.d.) random variables. In the sequel, B denotes a generic service requirement, and $\mathbb{E}[e^{-\omega B}]$ denotes its LST (Laplace–Stieltjes transform). If an arriving customer finds the server busy, then it joins the queue. The size of the waiting room is assumed to be infinite. The server serves customers in order of arrival (FCFS: first come first served), working at unit speed. Throughout the paper it is assumed that $\rho := \lambda \mathbb{E}[B] < 1$. This condition, which states that the mean offered amount of work per time unit is less than one, is necessary and sufficient for the existence of the steady-state distributions of the key performance measures.

There are numerous modifications and generalizations of the above-described basic $M/G/1$ model. We mention a few, giving one key reference for each:

- finite waiting room [3];
- customer impatience [4];
- server setup time for the first service in a busy period [5];
- server vacation at the end of a busy period [6];
- group arrivals and/or group service [7];
- workload-dependent service speed [8];
- several customer classes, but no priorities [9];
- several customer classes, with priorities (see [10]);
- polling models;
- matrix-geometric models. See, in particular, the books by Neuts [11,12], and the paper by Lucantoni [13] on the $BMAP/G/1$ queue, in which the Poisson

arrival process is generalized to a batch Markovian arrival process.

The text is organized as follows: The first section is devoted to the queue-length process, at various time epochs (arbitrary, arrival, and departure epochs). The waiting and sojourn time distributions are studied in the next section, followed by the section titled “The Busy Period.” In the last section we shall consider service disciplines other than FCFS.

QUEUE-LENGTH DISTRIBUTION

Let N^a , N^d , and N denote, respectively, the steady-state queue length (number in system) just before an arrival, just after a departure, and at an arbitrary epoch. PASTA (see **PASTA and Related Results**, [14]) implies that

$$N^a \stackrel{d}{=} N,$$

where $\stackrel{d}{=}$ denotes equality in distribution. Arguing that neither arrivals nor departures occur in groups, so that, in steady state, the number of customers increases from n to $n+1$ just as often as it decreases from $n+1$ to n , for $n = 0, 1, \dots$, it is seen that

$$N^a \stackrel{d}{=} N^d.$$

So to study N^a , N^d or N , it suffices to study just one of them. In a very influential paper, Kendall [15] has analyzed N^d . His fundamental observation was that the numbers of customers $N_1^d, N_2^d, \dots$ at successive departure epochs form a Markov chain, due to the memoryless property of the Poisson arrival process. Obviously, the number of customers at successive arrival epochs do *not* form a Markov chain, and neither is the process $\{N(t), t \geq 0\}$ of the number of customers at time t in general a continuous-time Markov process.

The following recursion holds:

$$\begin{aligned} N_{n+1}^d &= \max(0, N_n^d - 1) + A(B_{n+1}), \\ n &= 1, 2, \dots, \end{aligned} \quad (1)$$

where $A(B_{n+1})$ denotes the number of arrivals during the service time B_{n+1} . This recursion confirms that $\{N_n^d, n = 1, 2, \dots\}$ is a discrete-time Markov chain. Its transition probabilities $P_{ij} = \mathbb{P}(N_{n+1}^d = j | N_n^d = i)$ equal 0 if $j < i - 1$. Furthermore, defining the probability of having n arrivals during a service time

$$\alpha_n := \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} d\mathbb{P}(B < t),$$

we have

$$\begin{aligned} P_{0j} &= \alpha_j, \quad j = 0, 1, \dots, \\ P_{ij} &= \alpha_{j-i+1}, \quad i = 1, 2, \dots, \quad j = 0, 1, \dots. \end{aligned}$$

The steady-state probabilities $d_n := \mathbb{P}(N^d = n)$ satisfy the following equation:

$$d_k = \sum_{j=0}^k d_{k+1-j} \alpha_j + d_0 \alpha_k, \quad k = 0, 1, \dots \quad (2)$$

Introducing the generating functions

$$D(z) := \sum_{k=0}^\infty d_k z^k = \mathbb{P}(N^d = k) z^k,$$

$$A(z) := \sum_{k=0}^\infty \alpha_k z^k = \mathbb{E}[e^{-\lambda(1-z)B}],$$

it is readily verified, exploiting the convolution form of Equation (2), that

$$D(z) = \frac{A(z)}{z} (D(z) - d_0) + A(z) d_0,$$

so

$$D(z) = \frac{d_0(1-z)A(z)}{A(z) - z} = \frac{d_0(1-z)\mathbb{E}[e^{-\lambda(1-z)B}]}{\mathbb{E}[e^{-\lambda(1-z)B}] - z}.$$

The normalizing condition $D(1) = 1$, or the observation that d_0 equals the probability $\mathbb{P}(N = 0)$ of an empty system, yields $d_0 = 1 - \rho$. Finally,

$$D(z) = \mathbb{E}[z^{N^d}] = \frac{(1-\rho)(1-z)\mathbb{E}[e^{-\lambda(1-z)B}]}{\mathbb{E}[e^{-\lambda(1-z)B}] - z}. \quad (3)$$

This formula is sometimes referred to as the *Pollaczek–Khintchine* formula. Differentiation w.r.t. z yields

$$\mathbb{E}[N^d] = D'(1) = \frac{\lambda^2 \mathbb{E}[B^2]}{2(1-\rho)} + \rho. \quad (4)$$

We once more remind the reader that the same results hold for N^a and N .

WAITING AND SOJOURN TIME DISTRIBUTION

In this section, we present several methods to derive the (LST of the) waiting and sojourn time distributions in the $M/G/1$ queue with FCFS discipline. It should be noted that, nowadays, there are very efficient numerical inversion procedures known for LSTs of probability distributions.

In the sequel, W and S denote a generic waiting and sojourn time; $S = W + B$.

Method 1 [The distributional form of Little's law]. The FCFS service order implies that the customers present immediately after a customer departure are precisely the customers which have arrived during the sojourn time of the departing customer. Hence,

$$\begin{aligned} \mathbb{E}[z^{N^d}] &= \sum_{k=0}^{\infty} z^k \int_{t=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} d\mathbb{P}(S < t) \\ &= \mathbb{E}[e^{-\lambda(1-z)S}], \end{aligned} \quad (5)$$

a relation sometimes referred to as the distributional form of Little's law; Little's law itself stating that $\mathbb{E}[N] = \lambda \mathbb{E}[S]$ (see **Little's Law and Related Results**). From Equations (3) and (5),

$$\mathbb{E}[e^{-\lambda(1-z)S}] = \frac{(1-\rho)(1-z)\mathbb{E}[e^{-\lambda(1-z)B}]}{\mathbb{E}[e^{-\lambda(1-z)B}] - z},$$

so

$$\mathbb{E}[e^{-\omega S}] = \frac{(1-\rho)\omega \mathbb{E}[e^{-\omega B}]}{\omega - \lambda + \lambda \mathbb{E}[e^{-\omega B}]}. \quad (6)$$

Observing that the sojourn time of a customer is the sum of its waiting time and

its service time, and that these waiting and service times are independent, it follows that

$$\mathbb{E}[e^{-\omega S}] = \mathbb{E}[e^{-\omega W}] \mathbb{E}[e^{-\omega B}],$$

and hence,

$$\mathbb{E}[e^{-\omega W}] = \frac{(1-\rho)\omega}{\omega - \lambda + \lambda \mathbb{E}[e^{-\omega B}]}. \quad (7)$$

This formula is also known as the *Pollaczek–Khintchine* formula.

The distributional form of Little's law implies that $\mathbb{E}[N^d] = \lambda \mathbb{E}[S]$, and also yields relations between higher moments of N^d and S . From Equation (4),

$$\mathbb{E}[S] = \frac{\lambda \mathbb{E}[B^2]}{2(1-\rho)} + \mathbb{E}[B], \quad (8)$$

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[B^2]}{2(1-\rho)}. \quad (9)$$

Remark. The so-called *Mean Value Approach* provides an easy way to derive $\mathbb{E}[W]$ for the $M/G/1$ queue and for many of its variants and generalizations. It is based on the following observations: A newly arriving customer first has to wait for the *residual service time* of the customer in service (if any). It also has to wait during the service of all waiting customers. Denoting a generic residual service time by R , and the numbers of *waiting* customers at an arrival epoch and in steady state by N_w^a and N_w , respectively,

$$\mathbb{E}[W] = \rho \mathbb{E}[R] + \mathbb{E}[N_w^a] \mathbb{E}[B].$$

PASTA implies that $\mathbb{E}[N_w^a] = \mathbb{E}[N_w]$, and Little's law gives $\mathbb{E}[N_w] = \lambda \mathbb{E}[W]$.

Hence

$$\mathbb{E}[W] = \rho \mathbb{E}[R] + \rho \mathbb{E}[W],$$

so

$$\mathbb{E}[W] = \frac{\rho \mathbb{E}[R]}{1-\rho}.$$

R is the same as the asymptotic forward recurrence time of a renewal process with service times as interrenewal times. It follows from renewal theory that $\mathbb{E}[R] = \mathbb{E}[B^2]/(2\mathbb{E}[B])$, and Equation (9) follows. It is

interesting to observe that the mean waiting time is determined by the first two moments of the service-time distribution.

Method 2 [Level crossings]. In the seventies, Brill and Posner, and Cohen, independently developed a level crossing method (see the survey in Ref. 16), which is based on a balancing argument between up- and downcrossings of a particular queue length or workload level. In the case of the workload of the $M/G/1$ queue, the argument is as follows: Let $V(x)$ denote the distribution of the steady-state workload (amount of work) V in the $M/G/1$ queue (which by PASTA is in distribution equal to the waiting time), and let $v(x)$ denote its density. Equating the rates to cross any level $x > 0$ from below and from above, it is seen that

$$v(x) = \lambda \int_{0-}^x \mathbb{P}(B > x - y) dV(y), \quad x > 0. \quad (10)$$

Note that $V(x) = V(0) + \int_0^x v(y) dy$, take LSTs on both sides of Equation (10), and exploit the fact that the LST of a convolution is the product of the LSTs of the individual terms:

$$\mathbb{E}[e^{-\omega V}] - V(0) = \lambda \mathbb{E}[e^{-\omega V}] \frac{1 - \mathbb{E}[e^{-\omega B}]}{\omega}. \quad (11)$$

Remark. It should be noted that $\frac{\mathbb{P}(B > z)}{\mathbb{E}[B]}$ is the density of the residual part R of the service time B , and that its Laplace transform equals $\frac{1 - \mathbb{E}[e^{-\omega B}]}{\omega \mathbb{E}[B]}$. It follows from Equation (11) that

$$\mathbb{E}[e^{-\omega V}] = \frac{V(0)}{1 - \rho \mathbb{E}[e^{-\omega R}]}. \quad (12)$$

Taking $\omega = 0$ yields the expected result that the probability of an idle server equals $V(0) = 1 - \rho$. PASTA implies that V and W , the workload as seen by an arriving customer, have the same distribution. Hence,

$$\mathbb{E}[e^{-\omega W}] = \mathbb{E}[e^{-\omega V}] = \frac{1 - \rho}{1 - \rho \mathbb{E}[e^{-\omega R}]}. \quad (13)$$

It is easily verified that this agrees with Equation (7).

Remark. As observed by Doshi [16], the same level-crossing method readily reveals that, in the $M/G/1$ queue with server vacations, a decomposition result holds: the workload is equal in distribution to the sum of two independent quantities, one being the workload in the model without vacations. See Ref. 17 for other $M/G/1$ decomposition results, and Ref. 18 for a workload decomposition in polling models.

Remark. A powerful tool which is related to level crossings is the Rate of Conservation Law (RCL); it is surveyed by Miyazawa [19].

Remark. An attractive feature of the representation (Eq. 13) is that inversion of the LST is straightforward:

$$\mathbb{P}(W < x) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n \mathbb{P}(R_1 + \dots + R_n < x), \quad x > 0, \quad (14)$$

where $R_1, \dots, R_n$ denote n i.i.d. residual service times. See Ref. 20 for an interesting and elegant interpretation of this formula, exploiting work conservation (see **Conservation Laws and Related Applications**): the workloads under FCFS and under the LCFS-PR (last come first served preemptive resume) service discipline have the same distribution. Furthermore, the steady-state number of customers in the $M/G/1$ LCFS-PR queue is geometrically distributed with parameter ρ (see Ref. 21 for an elegant argument) and, as argued in Ref. 20, the residual service times of all customers present are i.i.d. copies of R .

Method 3 [Lindley's recursion]. We briefly outline a third method to derive Equation (7). The following relation, known as *Lindley's recursion*, holds between the waiting times W_n and W_{n+1} of the n th and

$(n + 1)$ st customers in a $GI/G/1$ FCFS queue:

$$W_{n+1} = [W_n + B_n - A_{n+1}]^+, \quad n = 1, 2, \dots, \quad (15)$$

where A_{n+1} denotes the interarrival time between the n th and $(n + 1)$ st customers and $[x]^+ = \max(0, x)$. Cohen (3, Section II.4.5) uses this recursion relation to determine the transient behavior of the waiting time in the $M/G/1$ queue. Below, we again restrict ourselves to the steady-state waiting-time behavior, rederiving Equation (7). Denoting by A a generic interarrival time, and by $(\cdot)$ an indicator function, we can write for the steady-state waiting-time LST in the $M/G/1$ case:

$$\begin{aligned} \mathbb{E}[e^{-\omega W}] &= \mathbb{E}[e^{-\omega[W+B-A]^+}] \\ &= \mathbb{E}[e^{-\omega[W+B-A]^+} (W+B \geq A)] \\ &\quad + \mathbb{E}[e^{-\omega[W+B-A]^+} (W+B < A)] \\ &= \mathbb{E}[e^{-\omega(W+B-A)} (W+B \geq A)] \\ &\quad + \mathbb{P}(W+B < A) \\ &= \mathbb{E}[e^{-\omega(W+B-A)}] - \mathbb{E}[e^{-\omega(W+B-A)} \\ &\quad \times (W+B < A)] + \mathbb{P}(W+B < A) \\ &= \mathbb{E}[e^{-\omega W}] \mathbb{E}[e^{-\omega B}] \frac{\lambda}{\lambda - \omega} \\ &\quad - \mathbb{P}(W+B < A) \left[\frac{\lambda}{\lambda - \omega} - 1 \right], \quad (16) \end{aligned}$$

where in the last step we have used that $A - W - B$, when positive, is $\exp(\lambda)$ distributed. After some rewriting, we obtain

$$\mathbb{E}[e^{-\omega W}] = \frac{\mathbb{P}(W+B < A)\omega}{\omega - \lambda + \lambda \mathbb{E}[e^{-\omega B}]}. \quad (17)$$

The normalizing condition implies that $\mathbb{P}(W+B < A) = 1 - \rho$, which makes sense as $\mathbb{P}(W+B < A) = \mathbb{P}([W+B-A]^+ = 0) = \mathbb{P}(W=0)$. We thus recover Equation (7).

Finally, we mention three other elegant derivations of Equation (7): a combinatorial method, developed by Takács [22], starting from a generalization of the classical ballot theorem; a derivation that exploits the regenerative structure of the waiting-time process

w.r.t. a busy period ([23], pp. 20–22); and one based on the Kella–Whitt martingale [24].

THE BUSY PERIOD

The following beautiful argument for deriving the joint distribution of the length of a busy period, P , and the number of customers served in it, K , is presented in Takács [25]. Consider the service time B_1 of the customer who starts a busy period by entering an empty system. Condition on the number of arrivals $A(B_1)$ during that first service time. If $A(B_1) = 0$, then $P = B_1$. Else, temporarily remove $A(B_1) - 1$ customers from the system at the end of B_1 . Serve the remaining customer and subsequent new arrivals until the busy period initiated by that customer ends. Then take the second removed customer into service, etc., until all $A(B_1)$ customers and their busy periods have been served. Thus P consists of B_1 , and of $A(B_1)$ sub-busy periods:

$$\begin{aligned} P &= B_1 + P_1 + \dots + P_{A(B_1)}, \\ K &= 1 + K_1 + \dots + K_{A(B_1)}. \end{aligned}$$

Since the order of arriving customers is irrelevant for the length of the busy period (we could, for example, serve LCFS, first taking the *last* removed customer into service), and since the joint distribution of the length of any sub-busy period and the number served in it is the same as the joint distribution of the whole busy period and the number served in it, we can write

$$\begin{aligned} \mathbb{E}[r^K e^{-\omega P}] &= r \int_{t=0}^{\infty} e^{-\omega t} \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &\quad \prod_{i=1}^k \mathbb{E}[r^{K_i} e^{-\omega P_i}] d\mathbb{P}(B_1 < t) \\ &= r \int_{t=0}^{\infty} e^{-\omega t} \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &\quad \times (\mathbb{E}[r^K e^{-\omega P}])^k d\mathbb{P}(B_1 < t) \\ &= r \mathbb{E}[e^{-(\omega + \lambda - \lambda \mathbb{E}[r^K e^{-\omega P}])B}]. \quad (18) \end{aligned}$$

It can be shown that $\mathbb{E}[r^K e^{-\omega P}]$ is the smallest zero, inside the unit circle, of the function $z - r \mathbb{E}[e^{-(\omega + \lambda(1-z))B}]$. In particular, it can

be shown that $\mathbb{E}[e^{-\omega P}]$ is the smallest zero, inside the unit circle, of the function $z - \mathbb{E}[e^{-(\omega + \lambda(1-z))B}]$.

Inversion of Equation (18) turns out to be possible [3, p. 250]:

$$\mathbb{P}(K = k, P < t) = \int_0^t e^{-\lambda u} \frac{(\lambda u)^{k-1}}{k!} d\mathbb{P}(B_1 + \dots + B_k < u). \quad (19)$$

We refer to Takács [22] for a combinatorial method of obtaining the busy-period distribution, which directly yields Equation (19).

The mean busy period is obtained via differentiation of Equation (18) (or via the simple balancing argument $\mathbb{E}[P] = \mathbb{E}[B] + \rho \mathbb{E}[P]$):

$$\mathbb{E}[P] = \frac{\mathbb{E}[B]}{1 - \rho}. \quad (20)$$

Notice that the mean busy period is only determined by the service-time distribution via its first moment.

OTHER SERVICE DISCIPLINES

So far we have restricted ourselves to the FCFS service discipline. An abundance of service disciplines has been studied for the $M/G/1$ queue. Arguably, the most important alternatives to FCFS are SRPT (shortest remaining processing time first), LCFS-PR, which was already briefly considered in the section titled “Waiting and Sojourn Time Distribution,” and PS (processor sharing).

SRPT. Under SRPT, the server always serves the customer with the shortest remaining service requirement. Intuitively, this leads to the smallest number of customers in the system, and hence in particular also to the smallest mean queue length and, by Little, to the smallest mean waiting time. Schrage [26] has indeed proven this optimality result. The queue-length distribution for $M/G/1$ SRPT is obtained by Schassberger [27], and the waiting and sojourn time distributions by Schrage and Miller [28].

LCFS-PR. Under LCFS-PR, the last arriving customer is immediately taken into service. The customer in service is temporarily set aside, and its preempted service is resumed as soon as that last arriving customer, as well as all customers which arrived during its service, during their service, and so on, have been served. LCFS-PR is particularly interesting from a theoretical point of view. Firstly, the $M/G/1$ LCFS-PR queue belongs to the class of *symmetric* queues [29, Section 3.3]. Nodes with such a discipline in a network of queues with Markovian routing allow for a joint queue-length distribution that has a product form. The steady-state queue-length distribution in an $M/G/1$ LCFS-PR queue is geometric with parameter ρ , the traffic load. Secondly, LCFS-PR has a kind of memoryless property: at an arrival epoch, the number of customers or the amount of work already present plays no role at all during the time that the arriving customer is present. This has been exploited in various studies, like those in Ref. 17 and 21. A special consequence is also that the sojourn time in $M/G/1$ LCFS-PR is in distribution equal to the busy period.

PS. Under PS, all customers present in the system are simultaneously being served, each receiving an equal fraction of the server capacity. PS was introduced by Kleinrock as the limit of the round-robin discipline in computer processors, where jobs receive a very small slice of service and subsequently return to the end of the queue, this process repeating itself until the job has been completely processed. More recently, PS has found new applications in representing bandwidth-sharing policies in modern communication networks like the Internet. The following interesting fairness property holds in $M/G/1$ PS: the mean sojourn time of a customer with service requirement x is $\frac{x}{1-\rho}$; that is, linear in x .

Like LCFS-PR, PS is a symmetric discipline which plays a key role in product form networks. Like LCFS-PR, the queue-length distribution in $M/G/1$ PS is geometric with parameter ρ . The sojourn time distribution is more complicated than for FCFS and LCFS-PR, though. It was independently derived by Yashkov, Ott and Schassberger. We refer to Ref. 30 for an interesting survey of processor-sharing models.

REFERENCES

1. Erlang AK. The theory of probabilities and telephone conversations. *Nyt Tidsskrift Mat* 1909;B 20:33–41.
2. Asmussen S. Extreme value theory for queues via cycle maxima. *Extremes* 1998;1:137–168.
3. Cohen JW. The single server queue. 2nd ed. Amsterdam: North-Holland Publ. Co.; 1982.
4. Stanford RE. On queues with impatience. *Adv Appl Probab* 1990;22:768–769.
5. Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice-Hall; 1989.
6. Takagi H. Queueing analysis. Volume 1, Vacation and priority systems. Amsterdam: North-Holland Publ. Co.; 1991.
7. Chaudhry ML, Templeton JGC. A first course in bulk queues. New York: Wiley; 1983.
8. Harrison JM, Resnick SI. The stationary distribution and first exit probabilities of a storage process with general release rule. *Math Oper Res* 1976;1:347–358.
9. Boxma OJ, Takine T. The $M/G/1$ FIFO queue with several customer classes. *Queueing Syst* 2003;45:185–189.
10. Jaiswal NK. Priority queues. New York: Academic Press; 1968.
11. Neuts MF. Matrix-geometric solutions in stochastic models. Baltimore (MD): The Johns Hopkins University Press; 1981.
12. Neuts MF. Structured stochastic matrices of $M/G/1$ type and their applications. New York: Marcel Dekker; 1989.
13. Lucantoni D. New results on the single server queue with a batch Markovian arrival process. *Stoch Models* 1991;7:1–46.
14. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
15. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann Math Stat* 1953;24:338–354.
16. Doshi BT. Level crossing analysis of queues. In: Bhat UN, Basawa IV, editors. *Queueing and related models*. Oxford: Clarendon Press; 1992. pp. 3–33.
17. Fuhrmann SW, Cooper RB. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper Res* 1985;33:1117–1129.
18. Boxma OJ. Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Syst* 1989;5:185–214.
19. Miyazawa M. Rate of conservation law: a survey. *Queueing Syst* 1994;15:1–58.
20. Cooper RB, Niu S-C. Benes's formula for $M/G/1$ -FIFO 'explained' by preemptive-resume LIFO. *J Appl Probab* 1986;23:550–554.
21. Núñez-Queija R. Note on the $GI/GI/1$ queue with LCFS-PR observed at arbitrary times. *Probab Eng Inf Sci* 2001;15:179–187.
22. Takács L. Combinatorial methods in the theory of stochastic processes. New York: Wiley; 1967.
23. Cohen JW. On regenerative processes in queueing theory. Berlin: Springer; 1976.
24. Kella O, Whitt W. Useful martingales for stochastic storage processes with Lévy input. *J Appl Probab* 1992;29:396–403.
25. Takács L. Introduction to the theory of queues. Oxford: Oxford University Press; 1962.
26. Schrage LE. A proof of the optimality of the shortest remaining processing time discipline. *Oper Res* 1968;16:678–690.
27. Schassberger R. The steady-state appearance of the $M/G/1$ queue under the discipline of shortest remaining processing time. *Adv Appl Probab* 1990;22:456–479.
28. Schrage LE, Miller LW. The queue $M/G/1$ with the shortest remaining processing time discipline. *Oper Res* 1966;14:670–684.
29. Kelly FP. Reversibility and stochastic networks. New York: Wiley; 1979.
30. Yashkov SF. Processor-sharing queues: some progress in analysis. *Queueing Syst* 1987;2:1–17.

THE $M/G/s$ QUEUE

TOSHIKAZU KIMURA
Graduate School of Economics
and Business Administration,
Hokkaido University,
Sapporo, Japan

Multiserver queues have been important models for evaluating the performance of a variety of service systems in the fields of computer, communication, transportation, and manufacturing. In particular, systems with Poisson inputs have been of general interest in these fields, due to intuitive interpretation and analytical tractability of the Poisson process. We consider the standard $M/G/s/s+r$ queue with $s(\geq 1)$ homogeneous servers in parallel, $r(\geq 0)$ extra waiting spaces, and the FCFS (first-come first-served) discipline. Customers are arriving at the system according to a Poisson process and are lost if they arrive when all s servers are busy and all r waiting spaces are full, namely they leave without receiving services and without affecting future arrivals. Their service times are iid (independent and identically distributed) random variables with a common cdf (cumulative distribution function) and independent of the arrival process. The $M/G/s/s+r$ queue is relevant to other single-station queues: It coincides with the $M/G/s/s$ queue if $r=0$ (see **The $M/G/s/s$ Queue**) and the $M/G/\infty$ queue if $s=\infty$ (see **The $M/G/\infty$ Queue** and **The $M/M/\infty$ Queue**). Also, it is a generalization of the $M/M/s$, $M/M/s/c$ and $M/G/1$ queues (see **The $M/M/s$ Queue**). In this article, we primarily focus on the case of unlimited waiting spaces ($r=\infty$), namely, the $M/G/s$ queue, with emphasis on its classical results because some recent results related to the $M/G/s$ queue are separately reviewed, for example, in the articles titled **Large Deviations in Queueing Systems** for large deviations, and **Markov Chains of the $M/G/1$ -Type** for the $M/G/1$ -type Markov chain analysis, in this encyclopedia.

EXACT RESULTS

Formulation

Let $N(t)$ denote the number of customers either waiting or being served at time t (≥ 0) in the $M/G/s$ queue. For the $M/G/1$ case, the instants of service completion are points of regeneration of the process $\{N(t); t \geq 0\}$. However, such an embedded Markov chain technique is of no use for the $M/G/s$ queue, since it is difficult to follow through successive customers due to the fact that they enter the queue in one order but leave the servers in another. Unlike the $M/M/s$ queue, the probability of service completion of a customer depends on its *age* at that instant. The dependence on the age of customers must be taken into account to analyze the $M/G/s$ queue, which can be realized by the method of inclusion of supplementary variables [1]; see Pollaczek [2] for an earlier analysis without using the supplementary variables. The supplementary variables are defined as either the elapsed or remaining service times of customers in service at arbitrary time. If we denote these supplementary variables at time t by $V_1(t), \dots, V_s(t)$, then the vector-valued process $\{(N(t), V_1(t), \dots, V_s(t)); t \geq 0\}$ becomes Markovian, and hence a set of equations for the state probabilities can be formulated by using transformations such as the generating function with respect to the discrete state $N(\cdot)$ and the multiple Laplace transform with respect to the vector $(t, V_1, \dots, V_s)$; see Syski [3, pp. 262–268] for details. See also Hokstad [4, pp. 512–515] for the formulation in equilibrium.

The equations for the transformed state probabilities is too involved to obtain its solution, even if the queue is in steady state. Using the equations, however, we can prove the PASTA property [5] that the steady-state probability of $N(\cdot)$ at an arrival epoch is equal to $P_j = P(N=j)$ ($j \geq 0$), the steady-state probability at an arbitrary epoch, where N is a generic random variable for $N(\cdot)$ at

arbitrary time in equilibrium [4, p. 514]. See also the article titled **PASTA and Related Results** in this encyclopedia for results related to PASTA.

For further analysis, we use the following notation: let λ denote the arrival rate of customers and let G denote the service-time cdf with mean μ and the squared coefficient of variation (variance divided by the square of the mean) c^2 . Let $\rho = \lambda/s\mu$ be the traffic intensity and assume that the system is stable and in steady state, that is $\rho < 1$ (for $r = \infty$) [6]. In addition, let $Q = \max(N - s, 0)$ be the number of customers in queue at an arbitrary epoch (excluding any customer in service) and let W be the waiting time of an arbitrary customer before beginning his/her service. For the $M/G/s$ queue, it has been known that the number of customers who arrived during the waiting time W has the same distribution as the queue length Q [7]. Specifically, between Q and W , we have the relation

$$E[z^Q] = E[e^{-\lambda(1-z)W}], \quad |z| \leq 1. \quad (1)$$

When necessary, we will express a quantity as a function of the service-time cdf G , for example $W \equiv W(G)$, $P_j \equiv P_j(G)$, and so on.

Asymptotic Properties

The exact formulation is analytically intractable, but we can obtain asymptotic properties of characteristic quantities of the $M/G/s$ queue in heavy or light traffic, which would be useful for generating accurate approximations. In heavy traffic when ρ is less than, but close to unity, the waiting time W approximately has a negative exponential distribution given by

$$P(W(G) \leq x) \approx 1 - \exp\left\{-\frac{2\lambda(1-\rho)}{1+\rho^2c^2}\right\}, \quad x \geq 0,$$

from which we obtain

$$\lim_{\rho \rightarrow 1} (1-\rho)EW(G) = \frac{1+c^2}{2s\mu}; \quad (2)$$

see K  llerstr  m [8]. In light traffic, on the contrary, Burman and Smith [9, Corollary

3.3] obtained

$$\lim_{\lambda \rightarrow 0} E[W(G) | W(G) > 0] = \int_0^\infty \{1 - G_e(t)\}^s dt \equiv a(s),$$

where G_e is the stationary-excess cdf associated with G , that is,

$$G_e(t) = \mu \int_0^t \{1 - G(u)\} du, \quad t \geq 0,$$

and hence

$$\lim_{\rho \rightarrow 0} \frac{EW(G)}{EW(M)} = s\mu a(s), \quad (3)$$

using the asymptotic property that $P(W(G) > 0) \approx s\mu EW(M)$ as $\lambda \rightarrow 0$; see Whitt [10] and Wang and Wolff [11] for related results.

To analyze asymptotic properties in these extreme cases in a unified way, we introduce a normalized quantity $R(G) = EW(G)/EW(M)$ as a bivariate function of s and ρ . Then, the heavy and light traffic limits in Equations (2) and (3) can be considered as the boundary values of $R(G)$ at $\rho \rightarrow 0$ and $\rho \rightarrow 1$, respectively; that is,

$$\begin{aligned} \lim_{\rho \rightarrow 1} R(G) &= \frac{1}{2}(1+c^2) \quad \text{and} \\ \lim_{\rho \rightarrow 0} R(G) &= s\mu a(s). \end{aligned} \quad (4)$$

From the exact results for the $M/G/1$ and $M/G/\infty$ queues, we have the boundary values of $R(G)$ at $s = 1$ and $s \rightarrow \infty$, which are respectively given by

$$\lim_{s \rightarrow 1} R(G) = \frac{1}{2}(1+c^2) \quad \text{and} \quad \lim_{s \rightarrow \infty} R(G) = 1. \quad (5)$$

APPROXIMATE RESULTS

To deal with the analytical difficulty, a number of approximations have been developed for characteristic quantities in the $M/G/s$ queue; see Ovuworie [12] and Kimura [13] for bibliographies. According to Miyazawa [14], the methods of generating approximations can be classified into two categories: one is to obtain approximations by presuming

a parametric formula theoretically or intuitively. The parameters included are chosen to be consistent with known exact results. This method often generates simple approximations for mean quantities, being suited for quick engineering purposes. The other is to obtain approximations by solving either exact or approximate equations for the quantities using some additional assumptions or hypotheses. This method is particularly useful when we develop approximations for the queue length distribution, for which intuitive inference is difficult. We review previous approximations from the view point of this classification.

Of course, this classification is *not* everything to approximations for the $M/G/s$ queue. For example, the method of diffusion approximation has been applied to the $M/G/s$ queue and its extensions; see Kimura [15–17], Yao [18], Whitt [19] and the references therein. The entropy maximization approach can be also applied to analyzing the $M/G/s$ queue [20,21]. In addition, Whitt [22] and Choi *et al.* [23] developed and evaluated heuristic two-moment approximations for general multiserver queues.

Mean Waiting Time

The development of approximations for the mean of queue length or waiting time has a history longer than that of their distributions. Due to Little's law, there is no essential difference between these two mean quantities. Hence, we primarily focus on the mean waiting time $EW(G)$ for the $M/G/s$ queue. An early and typical approximation for $EW(G)$ is traced back to the work of Lee and Longton [24], which is $EW(G) \approx EW^{(L)}(G)$ with

$$EW^{(L)}(G) = \frac{1}{2}(1 + c^2)EW(M), \quad (6)$$

where $EW(M)$ is the mean waiting time of the associated $M/M/s$ queue given by

$$EW(M) = \frac{(s\rho)^s}{s!s\mu(1-\rho)^2}P_0(M),$$

with the empty probability

$$P_0(M) = \left[\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!} \frac{1}{1-\rho} \right]^{-1}.$$

The formula (6) came from the consistency with both, the exact result for the $M/M/s$ queue and the cases $s = 1$ and $\rho \rightarrow 1$. Another typical approximation formula in this category is $EW^{(P)}(G)$ defined as

$$EW^{(P)}(G) = c^2EW(M) + (1 - c^2)EW(D), \quad (7)$$

where $EW(D)$ is the mean waiting time of the associated $M/D/s$ queue. The formula (7) is based on a simple idea of interpolating the exact results for the $M/M/s$ and $M/D/s$ queues, and is consistent with the cases $s = 1$, $s \rightarrow \infty$, and $\rho \rightarrow 1$; see Cosmetatos [25] and Kimura [26–28] for details of *system interpolations*. Due to the simplicity of its idea, it is difficult to identify the original citation of the formula (7). Actually, the formula (7) has often appeared in the literature [13, p. 353]. Since the formula (7) is sometimes called *Page's approximation* [29] therein, we attached his initial to the formula.

The formulas (6) and (7), being classified as the first category, can be interpreted as approximations for the normalized quantity $R(G)$, that is $R(G) \approx \frac{1}{2}(1 + c^2)$ for Equation (6) and $R(G) \approx c^2 + (1 - c^2)R(D)$ for Equation (7), where $R(D)$ is the associated quantity for the $M/D/s$ case. Most approximations for $EW(G)$ also can be regarded as those for $R(G)$, at least when they are consistent with $EW(M)$.

Table 1 summarizes those approximations from the point of view of approximating $R(G)$, together with a checklist of the consistency with the exact boundary values in Equations (4) and (5). In Table 1, two approximations specific to $R(D)$ are also added due to the difficulty of calculation. Cosmetatos' approximation [30, Equation (1)] $R^{(C)}(D)$ is sufficiently accurate for computing $EW(D)$ itself, but it has two serious defects such that $\lim_{s \rightarrow \infty} R^{(C)}(D) = \infty$ and $\lim_{\rho \rightarrow 0} R^{(C)}(D) = \infty$. (Note that $\lim_{s \rightarrow \infty} R(D) = 1$ and $\lim_{\rho \rightarrow 0} R(D) = s/(s+1) < 1$.) Hence, if one needs an approximation for $R(D)$ as a part of a certain approximation such as $R^{(B)}(G)$, it is highly recommended that Kimura's approximation is used [13, Equation (4.29)] $R^{(K)}(D)$, which is consistent with all of the exact boundary values; see Kimura [31] for a more sophisticated and accurate approximation.

Table 1. Summary of the approximations for $R(G)$ and $R(D)$

Citation	Symbol	$R(G)$	$R(D)$	$s = 1$	$s \rightarrow \infty$	$\rho \rightarrow 1$	$\rho \rightarrow 0$
Hokstad [4]	$R^{(H)}(\cdot)$	$\frac{1}{2}(1+c^2)$	$\frac{1}{2}$	✓	□	✓	□
Tijms <i>et al.</i> [32, Case A]	$R^{(T)}(\cdot)$	$\frac{\rho}{2}(1+c^2) + (1-\rho)s\mu a(s)$	$\frac{1}{2} \left\{ 1 + \frac{(1-\rho)(s-1)}{s+1} \right\}$	✓	□	✓	✓
Miyazawa [14, Case 2]	$R^{(M)}(\cdot)$	$\frac{1-\rho + \frac{\rho}{2}(1+c^2)}{1-\rho + \rho s\mu a(s)} s\mu a(s)$	$\frac{1}{2} \left\{ 1 + \frac{(1-\rho)(s-1)}{s+1-\rho} \right\}$	✓	□	✓	✓
Page [29]	$R^{(P)}(G)$	$\frac{c^2 + (1-c^2)R(D)}{(1+c^2)R(D)}$	—	✓	✓	✓	□
Boxma <i>et al.</i> [33]	$R^{(B)}(G)$	$\frac{2b(s)R(D) + 1 - b(s)}{(1+c^2)R(D)}$	—	✓	✓	✓	✓
Kimura [34]	$R^{(K)}(G)$	$\frac{(1+c^2)R(D)}{2c^2R(D) + 1 - c^2}$	—	✓	✓	✓	□
Cosmetatos [30]	$R^{(C)}(D)$	—	$\frac{1}{2} \{1 + f(s, \rho)\}$	✓	□	✓	□
Kimura [13]	$R^{(K)}(D)$	—	$\frac{1}{2} \{1 + f(s, \rho)g(s, \rho)\}$	✓	✓	✓	✓

where $a(s) = \int_0^\infty \{1 - G_e(t)\}^s dt$,

$$b(s) = \begin{cases} 1, & s = 1 \\ s+1 \left\{ \frac{1+c^2}{(s+1)\mu a(s)} - 1 \right\}, & s \geq 2, \end{cases} \quad \text{and} \quad \begin{cases} f(s, \rho) = \frac{(1-\rho)(s-1)(\sqrt{4+5s}-2)}{16s\rho} \\ g(s, \rho) = 1 - \exp \left\{ -\frac{s-1}{(s+1)f(s, \rho)} \right\}. \end{cases}$$

Queue Length Distribution

Elaborate approximations for $\{P_j(G)\}$ have been developed so far by several researchers. Among them, Hokstad [4], Tijms *et al.* [32] and Miyazawa [14] have made significant contributions to approximations in the second category.

Hokstad [4] adopted the method of supplementary variables to obtain the approximation

$$P_j^{(H)} = \begin{cases} \frac{(s\rho)^j}{j!} P_0(M), & j = 0, \dots, s-1, \\ (h_{j-s} - h_{j-s-1}) P_{s-1}^{(H)}, & j \geq s, \end{cases} \quad (8)$$

where $h_{-1} = 1$ and $\{h_j; j \geq 0\}$ is given by the generating function

$$H(z) \equiv \sum_{j=0}^{\infty} h_j z^j = \frac{1}{\gamma(z) - z} \quad \text{with} \\ \gamma(z) = \int_0^\infty e^{-\lambda(1-z)t/s} dG(t),$$

for $|z| \leq 1$. The sequence $\{h_j\}$ can be easily computed by numerical inversion of the generating function [35].

Tijms *et al.* [32] used a regenerative approach to derive three different approximations, namely, Cases A through C, of which Case B agrees with Hokstad's approximation. The numerical scheme for computing Case C requires much more work than those for Cases A and B, so that Case C is mainly of value for approximating $\{P_j\}$ for the $M/D/s$ queue. Their representative approximation of Case A, $\{P_j^{(T)}\}$, can be obtained by recursively solving the infinite system of linear equations

$$P_j^{(T)} = \begin{cases} \frac{(s\rho)^j}{j!} P_0(M), & j = 0, \dots, s-1, \\ \lambda \alpha_{j-s} P_{s-1}^{(T)} + \lambda \sum_{k=s}^j \beta_{j-k} P_k^{(T)}, & j \geq s, \end{cases} \quad (9)$$

together with the normalizing condition $\sum_{j=0}^{\infty} P_j^{(T)} = 1$, where for $j \geq 0$,

$$\alpha_j = \int_0^\infty \{1 - G_e(t)\}^{s-1} \{1 - G(t)\} e^{-\lambda t} \frac{(\lambda t)^j}{j!} dt, \\ \beta_j = \int_0^\infty \{1 - G(st)\} e^{-\lambda t} \frac{(\lambda t)^j}{j!} dt.$$

The recursion scheme for solving Equation (9) with integral-form coefficients is actually too burdensome. It is necessary to reduce the infinite system of linear equations to a relatively small system of linear equations, for example by using the geometric tail approach discussed in Tijms [36, pp. 298–299]. An alternative expression for $\{P_j^{(T)}\}$ has been obtained by Kimura [37, Proposition 3.2] as a special case of his results for the finite capacity $M/G/s$ queue, which is

$$P_j^{(T)} = \begin{cases} \frac{(s\rho)^j}{j!} P_0(M), & j = 0, \dots, s-1, \\ (t_{j-s} - t_{j-s-1}) P_{s-1}^{(T)}, & j \geq s, \end{cases} \quad (10)$$

where $t_{-1} = 0$ and $\{t_j; j \geq 0\}$ is given by the generating function

$$T(z) \equiv \sum_{j=0}^{\infty} t_j z^j = \frac{\lambda}{\gamma(z) - z} \int_0^{\infty} \{1 - G_e(t)\}^{s-1} \{1 - G(t)\} e^{-\lambda(1-z)t} dt,$$

for $|z| \leq 1$. The numerical inversion method for $T(z)$ would be a highly effective means for computing $\{P_j^{(T)}\}$ in common with $\{P_j^{(H)}\}$.

Miyazawa [14] aimed for affording theoretical insight into the approximations for $\{P_j(G)\}$ in the second category, giving a unified view of the previously established approximations. Using a rate conservation law [38], he provided reasonable interpretations of assumptions used in the approximations of Nozaki and Ross [39], Hokstad [4], and Case A of Tijms *et al.* [32]. In addition, he derived three new approximations for $\{P_j(G)\}$, Cases 1 through 3, in terms of generating functions. Among them, Case 2 has performance similar to Case A of Tijms *et al.* [32]. Case 3 is more complicated than Cases 1 and 2, and Case 1 is less accurate than Cases 2 and 3. Hence, the approximation of Case 2 would be of practical use.

In Table 1, we also give the approximations for $R(G)$ and $R(D)$ in the second category. From Table 1, we see that these approximations for $R(G)$ are *not* consistent with the asymptotic property as $s \rightarrow \infty$. However, this defect does not matter much for

practical purposes; see Ma and Mark [40] for a heuristic modification of $\{P_j^{(T)}\}$ for large s .

There are a few approximations for $\{P_j(G)\}$ in the first category. Based on the expression for $\{P_j(M)\}$, Kimura [41] suggested the geometric approximation

$$P_j^{(K)} = \begin{cases} \frac{(s\rho)^j}{j!} P_0(M), & j = 0, \dots, s-1, \\ C(s, \rho)(1-\eta)\eta^{j-s}, & j \geq s, \end{cases} \quad (11)$$

where $C(s, \rho)$ is the Erlang C formula of the associated $M/M/s$ queue; that is,

$$C(s, \rho) = \frac{(s\rho)^s}{s!(1-\rho)} P_0(M),$$

and η is a decay parameter. To obtain an appropriate approximation of the decay parameter η for the $M/G/s$ queue, Kimura [41] used a moment-matching with respect to $EW(G)$, obtaining

$$\eta = \frac{\rho R(G)}{1 - \rho + \rho R(G)}.$$

Clearly, the geometric approximation (Eq. 11) is exact for the $M/M/s$ case where $\eta = \rho$. It can be interpreted to view the $M/G/s$ queue as a system alternating between the associated $M/G/1$ and $M/G/\infty$ queues. This idea has been widely used such as in Hokstad [4], Tijms *et al.* [32], Newell [42], and Shore [43].

The delay probability is defined by $P(W(G) > 0) = \sum_{j=s}^{\infty} P_j(G)$ due to PASTA. From Equations (8), (9), and (11), these approximations give the Erlang delay probability approximation $P(W(G) > 0) \approx C(s, \rho)$, which has been known to usually be a good approximation for general service-time distributions. For large s , this approximation is certainly justified by the insensitive property of $M/G/\infty$ queue (see **The M/G/ ∞ Queue**). See also Miyazawa [14] and Kimura [44] for alternative approximations. To obtain the waiting time distribution $P(W(G) \leq t)$ ($t \geq 0$), we can apply the distributional Little's law in Equation (1) to each of the approximations for $\{P_j(G)\}$ in Hokstad [4, Equation (26)], van Hoorn and Tijms [45, Equation (5)] and Kimura [13, Equation (5.12)].

Optimal Buffer Design

In computer and communication systems, there are many important design issues that can be modeled by multiserver queues. A useful model for buffer capacity design in these systems is the $M/G/s/s+r$ queue. A buffer design problem is to find the smallest buffer capacity such that the proportion of lost customers (i.e., messages, transactions, etc.) is below an acceptable level. To obtain a concise expression for the optimal buffer size, we need an explicit formula for the loss probability, which is given by P_{s+r} due to PASTA. Approximations for $\{P_j(G)\}$ for the $M/G/s/s+r$ queue have been also proposed by Hokstad [4], Tijms and van Hoorn [46], Miyazawa [14], and Kimura [41] as extensions of their formulas for the $M/G/s$ queue; see Kimura [37] for some equivalence relations in those formulas. Among them, Miyazawa [14] and Kimura [41] obtained explicit approximation formulas for the loss probability. Based on these two explicit formulas, Kimura [47] obtained an alternative expression of the loss probability for $\rho \leq 1$, which is

$$P_{s+r}^{(K)} = \begin{cases} \frac{(1-\rho)\xi^r B(s, \rho)}{1 - \rho + \rho(1 - \xi^r)B(s, \rho)}, & \rho < 1, \\ \frac{B(s, 1)}{1 + \frac{2r}{1+c^2}B(s, 1)}, & \rho = 1, \end{cases} \quad (12)$$

where $B(s, \rho)$ is the Erlang B formula of the associated $M/G/s/s$ queue, that is

$$B(s, \rho) = \frac{(s\rho)^s}{s!} \bigg/ \sum_{j=0}^s \frac{(s\rho)^j}{j!},$$

and ξ is the exact rate of decay in the tail of $\{P_j(G)\}$ for the $M/G/s$ queue [48], which is defined by $\xi = \zeta^{-1}$ with the smallest positive root ζ of the equation $z = \gamma(z)$.

We can say that the buffer design problem is to find the smallest integer $r \geq 0$ for which $P_{s+r} \leq \epsilon$ for any acceptance level $\epsilon \in (0, 1)$. Since there exists a lower bound such that $P_{s+r} \geq 1 - \rho^{-1}$ when $\rho > 1$ [49, 50], there is no optimal value of r if $\epsilon \leq 1 - \rho^{-1}$ or, equivalently, if $\rho \geq 1/(1 - \epsilon) > 1$. Hence, for $\rho \leq 1$, there always exists the optimal

value of r for any $\epsilon \in (0, 1)$. For a given acceptance level ϵ , we define the optimal buffer size by $r_\epsilon \equiv r_\epsilon(G) = \min\{r \geq 0 \mid P_{s+r} \leq \epsilon\}$. If the acceptance level is more generous than the loss probability in the associated $M/G/s/s$ queue, it is not necessary to have extra buffer spaces. Hence, we immediately have $r_\epsilon(G) = 0$ if $\epsilon \geq B(s, \rho)$. From Equation (12), Kimura [47] proved that the optimal buffer size r_ϵ for $\epsilon \in (0, B(s, \rho))$ is given by the smallest non-negative integer r satisfying the inequality $r \geq r_\epsilon^*(G)$; that is, his *distribution-dependent* approximation $r_\epsilon^{(K)}(G)$ is given by

$$r_\epsilon^{(K)}(G) = \lceil r_\epsilon^*(G) \rceil, \quad (13)$$

where

$$r_\epsilon^*(G) = \begin{cases} \frac{1}{\ln \xi} \ln \left[\frac{\epsilon(1 - \rho + \rho B(s, \rho))}{B(s, \rho)(1 - \rho + \rho \epsilon)} \right], & \rho < 1, \\ \frac{1}{2}(1 + c^2) \left(\frac{1}{\epsilon} - \frac{1}{B(s, 1)} \right), & \rho = 1. \end{cases}$$

See Choi *et al.* [23] for a two-moment approximation similar to Equation (13). From Equation (13), the quotient $r_\epsilon(G)/r_\epsilon(M)$ can be approximated by the quantity

$$\theta(G) = \frac{r_\epsilon^*(G)}{r_\epsilon^*(M)} = \begin{cases} \frac{\ln \rho}{\ln \xi}, & \rho < 1, \\ \frac{1}{2}(1 + c^2), & \rho = 1. \end{cases}$$

Note that $\theta(G)$ is independent either of the acceptance level ϵ or of the number of servers s if ρ is fixed to a constant. From this property and numerical observations, Kimura [47] also provided a simple *two-moment* approximation $r_\epsilon^{(K)}(c^2)$, which is

$$r_\epsilon^{(K)}(c^2) = r_\epsilon(M) + \text{NINT}\left(\frac{1}{2}(c^2 - 1)\sqrt{\rho}r_\epsilon(M)\right), \quad (14)$$

where $\text{NINT}(x)$ is the function that returns the nearest integer of x . Another two-moment approximation for r_ϵ was developed by Tijms [36, Equation (4.8.20)], which is

$$r_\epsilon^{(T)}(c^2) = c^2 r_\epsilon(M) + (1 - c^2)r_\epsilon(D), \quad (15)$$

where $r_\epsilon(D)$ is the optimal buffer size for the associated $M/D/s/s+r$ queue. Clearly, $r_\epsilon^{(T)}(c^2)$ is based on the idea of system

interpolations shown in Equation (7). From numerical tests, we see that all of these approximations work well, independently on the number of servers s , even for small capacity cases; see Choi *et al.* [23] and Smith [51] for further numerical experiments of the two-moment approximations. Also, we see that the distribution-dependent approximation becomes more accurate than the two-moment approximations when service times are highly variable and a severe acceptance level is required in that system.

REFERENCES

1. Cox DR. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Math Proc Cambridge Phil Soc* 1955;51(3):433–441.
2. Pollaczek F. Über das warteproblem. *Math Z* 1934;38(1):492–537.
3. Syski R. Introduction to congestion theory in telephone systems. 2nd ed. Amsterdam: North-Holland Publishing Co.; 1986.
4. Hokstad P. Approximations for the $M/G/m$ queue. *Oper Res* 1978;26(3):510–523.
5. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30(2):223–231.
6. Kiefer J, Wolfowitz J. On the theory of queues with many servers. *Trans Am Math Soc* 1955;78(1):1–18.
7. Haji R, Newell GF. A relation between stationary queue and waiting time distributions. *J Appl Probab* 1971;8(3):617–620.
8. Köllerström J. Heavy traffic theory for queues with several servers. *J Appl Probab* 1974;11(3):544–552.
9. Burman DY, Smith DR. A light-traffic theorem for multi-server queues. *Math Oper Res* 1983;8(1):15–25.
10. Whitt W. Comparison conjectures about the $M/G/s$ queue. *Oper Res Lett* 1983;2(5):203–209.
11. Wang C-L, Wolff RW. The $M/G/c$ queue in light traffic. *Queueing Syst* 1998;29(1):17–34.
12. Ovuworie GC. Multi-channel queues: a survey and bibliography. *Int Stat Rev* 1980;48(1):49–71.
13. Kimura T. Approximations for multi-server queues: system interpolations. *Queueing Syst* 1994;17(3-4):347–382.
14. Miyazawa M. Approximations of the queue length distribution of an $M/GI/s$ queue by the basic equations. *J Appl Probab* 1986;23(2):443–458.
15. Kimura T. Diffusion approximation for an $M/G/m$ queue. *Oper Res* 1983;31(2):304–321.
16. Kimura T. An $M/M/s$ -consistent diffusion model for the $GI/G/s$ queue. *Queueing Syst* 1995;19(3-4):377–397.
17. Kimura T. A consistent diffusion approximation for finite-capacity multiserver queues. *Math Comput Model* 2003;38(11–13):1313–1324.
18. Yao DD. Refining the diffusion approximation for the $M/G/m$ queue. *Oper Res* 1985;33(6):1266–1277.
19. Whitt W. A diffusion approximation for the $G/GI/n/m$ queue. *Oper Res* 2004;52(6):922–941.
20. Kouvatso DD, Almond J. Maximum entropy two-station cyclic queues with multiple general servers. *Acta Inf* 1988;26(3):241–267.
21. Wu J-S, Chan WC. Maximum entropy analysis of multiple-server queueing systems. *J Oper Res Soc* 1989;40(9):815–825.
22. Whitt W. Approximations for the $GI/G/m$ queue. *Prod Oper Manag* 1993;2(2):114–161.
23. Choi DW, Kim NK, Chae KC. A two-moment approximation for the $GI/G/c$ queue with finite capacity. *INFORMS J Comput* 2005;17(1):75–81.
24. Lee AM, Longton PA. Queueing processes associated with airline passenger check-in. *Oper Res Q* 1959;10(1):56–71.
25. Cosmetatos GP. Some approximate equilibrium results for the multi-server queue ($M/G/r$). *Oper Res Q* 1976;27(3):615–620.
26. Kimura T. Approximating the mean waiting time in the $GI/G/s$ queue. *J Oper Res Soc* 1991;42(11):959–970.
27. Kimura T. Approximations for the waiting time in the $GI/G/s$ queue. *J Oper Res Soc Jpn* 1991;34(2):173–186.
28. Kimura T. Interpolation approximations for the mean waiting time in a multi-server queue. *J Oper Res Soc Jpn* 1992;35(1):77–92.
29. Page E. Queueing theory in OR. London: Butterworths; 1972.
30. Cosmetatos GP. Approximate explicit formulae for the average queueing time in the processes ($M/D/r$) and ($D/M/r$). *Informatics* 1975;13(3):328–332.
31. Kimura T. Refining Cosmetatos' approximation for the mean waiting time in the $M/D/s$ queue. *J Oper Res Soc* 1991;42(7):595–603.

32. Tijms HC, van Hoorn MH, Federgruen A. Approximations for the steady-state probabilities in the $M/G/c$ queue. *Adv Appl Probab* 1981;13(1):186–206.
33. Boxma OJ, Cohen JW, Huffels N. Approximations of the mean waiting time in an $M/G/s$ queueing system. *Oper Res* 1979;27(6):1115–1127.
34. Kimura T. A two-moment approximation for the mean waiting time in the $GI/G/s$ queue. *Manag Sci* 1986;32(6):751–763.
35. Abate J, Whitt W. Numerical inversion of probability generating functions. *Oper Res Lett* 1992;12(4):245–251.
36. Tijms HC. *Stochastic models: an algorithmic approach*. New York: Wiley; 1994.
37. Kimura T. Equivalence relations in the approximations for the $M/G/s/s + r$ queue. *Math Comput Model* 2000;31(10–12):215–224.
38. Miyazawa M. Rate conservation laws: a survey. *Queueing Syst* 1994;15(1–4):1–58.
39. Nozaki SA, Ross SM. Approximations in finite-capacity multi-server queues with Poisson arrivals. *J Appl Probab* 1978;15(4):826–834.
40. Ma BNW, Mark JW. Approximation of the mean queue length of an $M/G/c$ queueing system. *Oper Res* 1995;43(1):158–165.
41. Kimura T. A transform-free approximation for the finite capacity $M/G/s$ queue. *Oper Res* 1996;44(6):984–988.
42. Newell G. Approximate stochastic behaviour of n -server service systems with large n , *Lecture Notes in Economics and Mathematical Systems*, No. 87. Berlin: Springer; 1973.
43. Shore H. Simple approximations for the $GI/G/c$ queue – I: the steady-state probabilities. *J Oper Res Soc* 1988;39(3):279–284.
44. Kimura T. Approximations for the delay probability in the $M/G/s$ queue. *Math Comput Model* 1995;22(10–12):157–165.
45. van Hoorn MH, Tijms HC. Approximations for the waiting time distribution of the $M/G/c$ queue. *Perform Eval* 1982;2(1):22–28.
46. Tijms HC, van Hoorn MH. Computational methods for single-server and multi-server queues with Markovian input and general service times. In: Disney RL, Ott TJ, editors. *Volume II, Applied probability–computer science, the interface*. Boston (MA): Birkhäuser; 1982. pp. 71–102.
47. Kimura T. Optimal buffer design of an $M/G/s$ queue with finite capacity. *Stoch Models* 1996;12(1):165–180.
48. Abate J, Choudhury GL, Whitt W. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stoch Models* 1994;10(1):99–143.
49. Sobel MJ. Simple inequalities for multiserver queues. *Manag Sci* 1980;26(9):951–956.
50. Heyman DP. Comments on a queueing inequality. *Manag Sci* 1980;26(9):956–959.
51. Smith JM. $M/G/c/K$ blocking probability models and system performance. *Perform Eval* 2003;52(4):237–267.

THE $M/G/s/s$ QUEUE

JEFFREY P. KHAROUFEH

Department of Industrial Engineering,
University of Pittsburgh, Pittsburgh,
Pennsylvania

INTRODUCTION

The $M/G/s/s$ queueing system is perhaps the most important and popular model for analyzing service systems characterized by random arrival patterns, random processing times, and a limited number of resources. It has been especially useful for telephone network design, the performance evaluation of telecommunications systems, and the design and analysis of large-scale service systems (e.g., customer contact centers). Most of the main results for the model were developed within the first three decades of the twentieth century, and the model's utility and importance have grown since that time.

In the parlance of Kendall's notation [1], the abbreviation $M/G/s/s$ means that the customer arrival process is Markovian or memoryless (M), the service times follow a general distribution (G), there are s identical servers ($s \geq 1$), and the total system capacity is limited to s (i.e., there is no waiting space in the system). The absence of a waiting space implies that customers who arrive to find all s servers busy leave and do not retry their service later. These customers are said to be lost. The $M/G/s/s$ queue is often called the *Erlang loss system*, since its development can be traced primarily to the Danish mathematician, Agner Krarup Erlang (1878–1929).

The popularity of the $M/G/s/s$ model stems from the simplicity of the steady-state probability distribution of the number of busy servers, which makes it easy to estimate the proportion of customers that are lost by the system in the long run. This performance parameter is an important measure of service quality in many practical contexts. For example, the lost customers might

represent vehicles arriving to a parking garage to find no vacant parking spaces. The garage owner forfeits a fraction of the revenue that he or she would have obtained if another parking space were available. In telecommunications applications, the lost customers may represent lost data packets that contain critical information. For either case, the service provider is concerned with minimizing the proportion of lost customers to ensure a high quality of service. The steady-state distribution of the number of busy servers also provides an estimate of the utilization of the system's resources in the long run.

The next section provides a mathematical description of the $M/G/s/s$ system and presents the main results for its steady-state behavior, namely, the steady-state distribution of the number of occupied servers and the steady-state blocking probability.

MODEL DESCRIPTION AND THE ERLANG B FORMULA

In an $M/G/s/s$ queue, customers arrive to the system according to a homogeneous Poisson process with rate parameter λ ($\lambda > 0$), and the service times form an independent and identically distributed (i.i.d.) sequence of nonnegative random variables with finite mean τ . The *offered load* to the system is the dimensionless quantity

$$a \equiv \lambda\tau.$$

The offered load per server, often termed the *traffic intensity*, is given by

$$\rho \equiv \frac{\lambda\tau}{s} = \frac{a}{s}.$$

Denote by $Q(t)$ the number of occupied servers (or equivalently, the number of customers in the system) at time t . Note that $\{Q(t) : t \geq 0\}$ is a stochastic process on the state space $S = \{0, 1, 2, \dots, s\}$. Let Q denote

the steady-state number of busy servers and define the probability mass function (p.m.f.) of Q by

$$p_j = \lim_{t \rightarrow \infty} P(Q(t) = j) = P(Q = j), \\ j = 0, 1, 2, \dots, s.$$

Owing to the fact that, at most, a finite number of customers can be present in the system, the existence of the steady-state distribution p_j , $0 \leq j \leq s$, is ensured. Using a reversibility argument, as outlined by Gross and Harris [2] and detailed in Ref. 3, it can be shown that

$$p_j = \frac{a^j/j!}{\sum_{i=0}^s a^i/i!}, \quad j = 0, 1, 2, \dots, s. \quad (1)$$

Equation (1) indicates that Q has a truncated Poisson distribution with parameter a , the offered load. The attractive feature of the steady-state distribution is that it depends on the service time distribution only through its mean (since $a = \lambda\tau$) and is otherwise insensitive to the form of the distribution. This *insensitivity property* is extremely useful for real applications when service times may be distinctly nonexponential. Because Poisson arrivals see time averages [PASTA, [4,5]], the long-run proportion of arriving customers who see s servers busy (and are blocked from receiving service) is precisely p_s . The formula for this blocking probability, termed the *Erlang B formula*, is

$$B(s, a) \equiv p_s = \frac{a^s/s!}{\sum_{j=0}^s a^j/j!}. \quad (2)$$

The Erlang B formula is a fundamental result for telephone traffic engineering problems and can be used to select the appropriate number of trunks (servers) needed to ensure a small proportion of lost calls (customers).

Several important properties of the function $B(s, a)$ can easily be derived. For example, for a fixed, the blocking probability $B(s, a)$ monotonically decreases to zero as s increases, and for s fixed, $B(s, a)$ monotonically increases to unity as a increases. Other

properties, such as convexity of the loss formula, have been discussed in Refs 6–8, among others. The Erlang B formula can also provide valuable insights into the behavior of large systems. For instance, Equation (2) can be used to show that it is more efficient to use a single large system with pooled servers than to use multiple smaller systems.

The next section examines the case when service times follow an exponential distribution. We discuss loss systems as well as systems with additional waiting spaces.

EXPONENTIAL SERVICE AND THE ERLANG C FORMULA

If the service times are independent and identically distributed exponential random variables with mean $1/\mu$, the system is an $M/M/s/s$ queue and $Q(t) : t \geq 0$ is a Markov process, namely, a birth-and-death process with birth rates

$$\lambda_j = \begin{cases} \lambda, & \text{if } j = 0, 1, 2, \dots, s-1, \\ 0, & \text{if } j = s, \end{cases}$$

and death rates

$$\mu_j = \begin{cases} j\mu, & \text{if } j = 1, 2, \dots, s, \\ 0, & \text{if } j = 0. \end{cases}$$

In this case, $\lambda\tau = \lambda/\mu$, so the Erlang B formula is [2]

$$B(s, a) = \frac{(\lambda/\mu)^s/s!}{\sum_{j=0}^s (\lambda/\mu)^j/j!}. \quad (3)$$

Related to the Erlang B formula is the *Erlang C formula* (or *Erlang delay formula*) for the $M/M/s$ system (or *Erlang delay system*), which includes an infinite-capacity queue to accommodate arriving customers who find all s servers busy. For this model, p_s is interpreted as the long-run proportion of customers who experience a delay before their service begins. The model assumes that the customers are willing to wait as long as they wish to receive service. The Erlang C

formula is given as [9]

$$C(s, a) = \frac{\frac{a^s}{s!(1-\rho)}}{\sum_{i=0}^{s-1} \frac{a^i}{i!} + \frac{a^s}{s!(1-\rho)}}, \quad 0 < a < s, \quad (4)$$

where $\rho = \lambda/s\mu$. Note that Equation (4) does not hold for arbitrary service time distributions, and it requires that the traffic intensity (ρ) is strictly less than unity.

As one might expect, a simple relationship exists between the Erlang B formula of Equation (2) and the Erlang C formula of Equation (4). It can be shown that [9]

$$C(s, a) = \frac{sB(s, a)}{s - a[1 - B(s, a)]}, \quad 0 < a < s.$$

COMPUTATIONAL ISSUES

When a and s are large, it can be difficult to compute the right-hand side of Equation (2) due to the presence of factorials and potentially very large powers. However, the blocking probability can be much more easily computed by using a recursive scheme. As before, let

$$B(s, a) \equiv \frac{a^s/s!}{\sum_{i=0}^s a^i/i!}$$

denote the blocking probability when an s -server system is offered a units of Poisson traffic. By analyzing the reciprocal of $B(s, a)$, it is not difficult to derive the recursion

$$B(k, a) = \frac{aB(k-1, a)}{k + aB(k-1, a)}, \quad k = 1, 2, \dots, s, \quad (5)$$

where $B(0, a) \equiv 1$. This recursion can be easily implemented in a spreadsheet or a computer program to obtain the blocking probability for systems with very large numbers of servers that are heavily loaded. To illustrate, consider a system with $s = 10^4$ servers that is offered $a = 10^4$ units of Poisson traffic. Computing $B(s, a)$ by Equation (5), the steady-state blocking probability is

$$B(10^4, 10^4) \approx 0.00794.$$

This means that about 0.8% of customers will be blocked from service.

When designing a new system, an important consideration is whether to use multiple small systems or to pool servers in one larger system. Suppose that the first option is to operate two different loss systems, one with s_1 servers handling a_1 units of traffic and the second with s_2 servers handling a_2 units of traffic. The alternative is to operate a single larger system with $s = s_1 + s_2$ servers and a combined offered load of $a = a_1 + a_2$. Is it more efficient to operate the single larger system or the two smaller systems? This question can be answered using the Erlang B formula. Suppose that $s_1 = 100$, $s_2 = 50$, $a_1 = 100$, and $a_2 = 100$. The proportion of blocked customers in the first system is

$$B(s_1, a_1) \approx 0.0757.$$

The proportion of blocked customers in the second system is

$$B(s_2, a_2) \approx 0.5093.$$

Therefore, the combined total proportion of lost customers between the two small systems is approximately 59%. If, on the other hand, all of the servers are pooled in a single system with $s = s_1 + s_2 = 150$ servers, and the two traffic streams are combined so that the offered load is $a = a_1 + a_2 = 200$, then the blocking probability is

$$B(s, a) = B(150, 200) \approx 0.26318.$$

That is, the larger service system reduces the (overall) proportion of lost customers by more than 50%. In general, it can be shown that $B(s_1 + s_2, a_1 + a_2) < B(s_1, a_1) + B(s_2, a_2)$.

FURTHER READING

Classical textbooks that provide further details on the standard $M/G/s/s$ model include Refs 2, 3, 5, and 9.

Important properties of the loss and delay formulae have been studied by many researchers. Jagerman [8] provided convexity results for the Erlang loss function. Harel and

Zipkin [10] proved strong convexity results for the average sojourn time in an $M/M/c$ system. Harel [7] showed that the proportion of lost customers in an $M/G/s/s$ system is convex in the arrival rate, provided that the traffic intensity is below a threshold. Jagers and van Doorn [11] proved the convexity of functions that generalize the Erlang loss and delay functions. Some simple inequalities for the performance parameters of multi-server queues, including loss systems, were provided by Sobel [12]. Harel [6] provided sharp bounds and simple approximations for the loss and delay formulae. Recently, Adelman [13] contributed simple algebraic approximations to the blocking formula using Palm calculus [14], which equates long-run event averages to long-run time averages. Ross and Seshadri [15] provided methods for estimating the expected time until the first customer is lost by the system.

Other authors have extended the basic model to include more complex and realistic dynamics. For instance, Brumelle [16] considered state-dependent systems with arrival and service rates that depend on the number of busy servers. The transient analysis of the Erlang loss system was considered by Abate and Whitt [17] and Grier *et al.* [18]. Systems with time-varying arrival and/or service rates have been studied by Davis *et al.* [19], and more recently by Knessl and Yang [20]. Lin and Ross [21] considered optimal admission control policies for a single-server Erlang loss model. Ziya *et al.* [22] analyzed the optimal pricing of customers in a loss model.

REFERENCES

1. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Ann Math Stat* 1953;24:338–354.
2. Gross D, Harris C. Fundamentals of queueing theory. New York: John Wiley & Sons, Inc.; 1998.
3. Ross SM. Stochastic processes. New York: John Wiley & Sons, Inc.; 1996.
4. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
5. Wolff RW. Stochastic modeling and the theory of queues. New York: John Wiley & Sons, Inc.; 1989.
6. Harel A. Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Manage Sci* 1988;34:959–972.
7. Harel A. Convexity properties of the Erlang loss formula. *Oper Res* 1990;38:499–505.
8. Jagerman DL. Some properties of the Erlang loss function. *Bell Syst Tech J* 1974;53:525–551.
9. Cooper RB. Introduction to queueing theory. 2nd ed. New York: North-Holland; 1981.
10. Harel A, Zipkin PH. Strong convexity results for queueing systems. *Oper Res* 1987;35:405–418.
11. Jagers AA, van Doorn EA. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Rev* 1991;33:281–282.
12. Sobel MJ. Some inequalities for multiserver queues. *Manage Sci* 1980;26:951–956.
13. Adelman D. A simple algebraic approximation to the Erlang loss model. *Oper Res Lett* 2008;36:484–491.
14. Baccelli F, Bremaud P. Elements of queueing theory: Palm Martingale calculus and stochastic recurrences. Berlin: Springer; 2003.
15. Ross SM, Seshadri S. Hitting time in an Erlang loss system. *Probab Eng Inf Sci* 2002;16:167–184.
16. Brumelle SL. A generalization of Erlang's loss system to state dependent arrival and service rates. *Math Oper Res* 1978;3:10–16.
17. Abate J, Whitt W. Calculating transient characteristics of the Erlang loss model by numerical transform inversion. *Commun Stat Stoch Models* 1998;14:663–680.
18. Grier N, Massey WA, McKoy T, *et al.* The time-dependent Erlang loss model with retrials. *Telecommun Syst* 1997;7:253–265.
19. Davis JL, Massey WA, Whitt W. Sensitivity to the service-time distribution in the non-stationary Erlang loss model. *Manage Sci* 1995;41:1107–1116.
20. Knessl C, Yang Y. On the Erlang loss model with time dependent input. *Queueing Syst* 2006;52:49–104.
21. Lin KY, Ross SM. Optimal admission control for a single-server loss queue. *J Appl Probab* 2004;41:535–546.
22. Ziya S, Ayhan H, Foley RD. Optimal prices for finite capacity queueing systems. *Oper Res Lett* 2006;34:214–218.

THE $M/M/1$ QUEUE

MYRON HLYNKA
Department of Mathematics
and Statistics, University of
Windsor, Ontario, Canada

In a queueing system, customers arrive for service, may wait for some time in a queue before entering service, receive service, and then leave. But these “customers” need not be people. Customers could be computer commands, telephone messages, items on a store shelf, or cars in a parking lot. Two classic books on queueing theory are by Gross *et al.* [1] and Kleinrock [2]. Other books on queueing include Bolch *et al.* [3] and Medhi [4]. For a web site with extensive queueing references, see Hlynka [5].

The $M/M/1$ queue uses Kendall’s notation. The first M (which stands for Markov or memoryless) means that the interarrival times are independent exponentially distributed random variables with arrival rate λ customers per unit time. Equivalently, we often say that arrivals follow a Poisson process with rate λ . The second M means that the service times are independent exponentially distributed random variables with service rate μ . The number 1 in Kendall’s notation indicates that there is a single server. Other assumptions that are usually made with this model are as follows:

- (a) there is an infinite buffer (waiting area) to hold customers;
- (b) customers are served in a first-in, first-out (FIFO) manner.

We can illustrate a single server first-in first-out queue as

** ... ** Server

where each * represents a customer. Figure 1 shows the number of customers in a typical $M/M/1$ queueing system versus time.

Two books with extensive discussion of $M/M/1$ queues are Robert [6] and Asmussen [7].

WHY IS THE $M/M/1$ MODEL SO POPULAR?

There are several reasons for the popularity of the $M/M/1$ queueing model. First, the $M/M/1$ queue is the easiest queue to analyze (other than the $D/D/1$ queue). The memoryless property means that if we know the current state of the system, we do not have to keep track of the past in order to find probabilities of future events. This keeps the mathematics and notation easier than it would be in more complex models.

Second, the assumptions of the $M/M/1$ queueing model are reasonable for several common models in the real world. Interarrival times are exponentially distributed with rate λ if customers behave independently of each other in making a decision whether to seek service at a particular time. Thus, arrivals to a barber shop or telephone calls to a call center can be expected to follow a Poisson process.

Models for which service times are exponentially distributed not as common as models for which interarrival times are exponentially distributed. However, exponentially distributed service times make sense in situations where most customers have a short service time but some customers need longer time lengths. An example of exponentially distributed service times might be the times required to process vehicles crossing the Canada/US border. Most customers with a passport and nothing to declare can be processed quite quickly. Others, with passports from countries other than Canada or the United States, or with large purchases, will take a much longer time. Similarly, transactions with a bank teller might be modeled by an exponential distribution.

A third reason that $M/M/1$ queueing systems are popular in the queueing literature is that they form an excellent base case, which can be used as a comparison for generalizations of such models. So $M/G/1$ models, $GI/M/1$ models, $M/M/s$ models, or $M/E/1$ models are all generalizations of $M/M/1$ models.

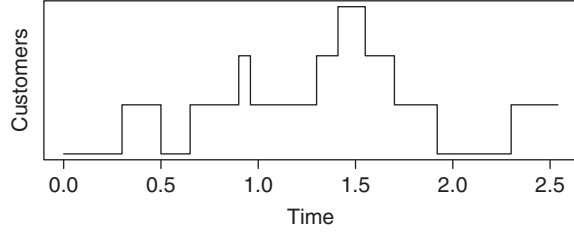


Figure 1. Number of customers versus time.

BIRTH-AND-DEATH TYPE FORMULATION

A standard $M/M/1$ queueing system is a special case of a birth-and-death process. Arrivals to the queue are considered to be births and service completions are considered to be deaths. The state of the system is taken to be the number of customers in the system (Fig. 2).

Given that there are i customers in an $M/M/1$ queueing system at a fixed point in time, let $p_{ij}(t)$ be the probability of having j customers in the system t time units later. We say that a function f is $o(t)$ if $\lim_{t \rightarrow 0} \frac{f(t)}{t} = 0$. In a small time Δt , the assumptions of a Poisson process require that the probability of two or more changes (arrivals or service completions) is $o(\Delta t)$ (“little oh”). It is also assumed that the probability of an arrival in a small time interval is approximately proportional to the length of the interval and the probability of a service completion in a small time interval is approximately proportional to the length of the interval. Thus, we have the following for $j = 1, 2, \dots$

$$p_{ij}(t + \Delta t) = p_{i,j-1}(t)\lambda\Delta t + p_{i,j+1}(t)\mu\Delta t + p_{ij}(t)(1 - \lambda\Delta t - \mu\Delta t) + o(\Delta t).$$

Subtract $p_{ij}(t)$ and divide by Δt to get

$$\frac{p_{ij}(t + \Delta t) - p_{ij}(t)}{\Delta t} = \lambda p_{i,j-1}(t) + \mu p_{i,j+1}(t) - (\lambda + \mu)p_{ij}(t) + \frac{o(\Delta t)}{\Delta t}.$$

Let $\Delta t \rightarrow 0$ and recall that $f'(t) = \lim_{\Delta t \rightarrow 0} \frac{f(t + \Delta t) - f(t)}{\Delta t}$. Thus

$$p'_{ij}(t) = p_{i,j-1}(t)\lambda + p_{i,j+1}(t)\mu - p_{ij}(t)(\lambda + \mu) \text{ for } j = 1, 2, \dots$$

When $j = 0$, we can obtain a simpler expression

$$p'_{i0}(t) = p_{i,1}(t)\mu - p_{i0}(t)\lambda.$$

These differential equations are called the *Chapman–Kolmogorov* equations and are the stepping stones to many other results about the $M/M/1$ queue.

LIMITING PROBABILITIES FOR THE M/M/1 QUEUE

We are interested in limiting probabilities $\lim_{t \rightarrow \infty} p_{ij}(t)$, which give the probability of having j customers in the system far in the future, if we begin with i customers initially. Such limiting probabilities may not exist for

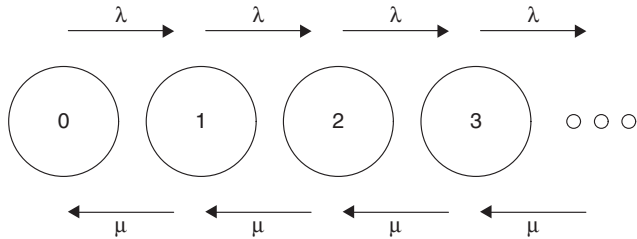


Figure 2. Transition diagram for an $M/M/1$ queue.

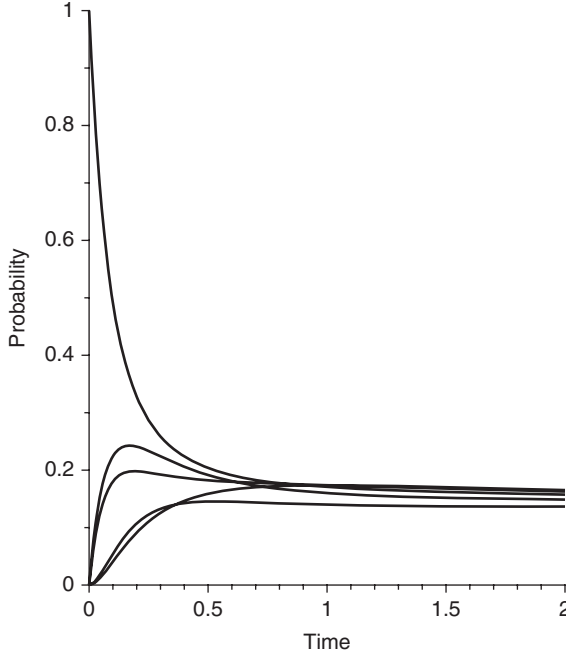


Figure 3. Transient probabilities tending to the same limit.

some queueing systems. For example, if we begin with 0 customers and exactly one customer arrives at the end of each minute and requires exactly 50 s of service then the number of customers would alternate between 0 and 1 with changes at regular intervals of time, so there would be no limit. However, this regular periodicity does not happen for an $M/M/1$ queueing system. If $\lambda > \mu$, then the number of customers in the queueing system will tend to grow larger and larger so the limiting probability of being in a particular state j would be 0 for all $j = 0, 1, \dots$. However, it can be shown that limiting probabilities for an $M/M/1$ queueing system exist and are nonzero iff the arrival rate λ is less than the service rate μ (i.e., $\lambda < \mu$).

Given that the limiting probabilities exist, it is reasonable to expect that the initial number of customers will lose importance after a long period of time and that the graph of $p_{ij}(t)$ will approach some limiting value. Thus $\lim_{t \rightarrow \infty} p'_{ij}(t) = 0$, and we can define $\pi_j = \lim_{t \rightarrow \infty} p_{ij}(t)$ which does not depend on the initial value i . In Fig. 3, we show a plot of $p_{i2}(t)$ for $\lambda = 4$, $\mu = 5$, and $i = 0, 1, 2, 3, 4$, which shows all of the probability

curves tending toward the same limiting value π_2 .

Thus far we do not know the values of $p_{ij}(t)$ and π_j . It turns out that the former are very complex and difficult to obtain, while the latter are very simple and easy to obtain. If we take the Chapman–Kolmogorov equations and let $t \rightarrow \infty$, we obtain

$$0 = \pi_{j-1}\lambda + \pi_{j+1}\mu - \pi_j(\lambda + \mu) \text{ for } j = 1, 2, \dots \quad (1)$$

$$0 = \pi_1\mu - \pi_0\lambda. \quad (2)$$

These two equations are important and can be obtained in a more intuitive manner by considering a balance equation “rate in equals rate out” approach. If $j > 0$, then state j can be entered only from states $j - 1$ or state $j + 1$. Thus the rate into state j is $\pi_{j-1}\lambda + \pi_{j+1}\mu$. The system leaves state j if there is an arrival or service completion, so the rate out of state j is $\pi_j(\lambda + \mu)$. Equating these two rates gives Equation (1). Equation (2) follows in a similar manner.

In general, we define $\rho = \frac{\lambda}{s\mu}$ where s is the number of servers. For the $M/M/1$ queue, we have $s = 1$ and in this case, ρ may be referred

to as the *traffic intensity* or the *utilization factor*.

Theorem 1. *The limiting probabilities π_j of having j customers in an M/M/1 queueing system exist if $\lambda < \mu$. In this case,*

$$\pi_j = \rho^j(1 - \rho) \text{ for } j = 0, 1, \dots$$

Proof. We can solve Equations 1 and 2 recursively to obtain

$$\begin{aligned}\pi_1 &= \frac{\lambda}{\mu}\pi_0 = \rho\pi_0, \\ \pi_2 &= \frac{\lambda}{\mu}\pi_1 = \rho\pi_1 = \rho^2\pi_0.\end{aligned}$$

In general,

$$\pi_j = \rho^j\pi_0 \text{ for } j = 0, 1, \dots$$

If we sum both sides over all j , the left hand side becomes 1 and the right hand side becomes $\frac{\pi_0}{1-\rho}$ provided $0 < \rho < 1$. Thus, we find $\pi_0 = 1 - \rho$ and the result follows.

A corollary to this result is that the probability of having at least k customers in the system is ρ^k . The limiting probabilities are also called *steady-state* probabilities and *stationary* probabilities in this setting. The probability π_i represents the probability that the system is in state i (i.e., has i customers) at any point in time far in the future and also represents that long run proportion of time that the system has i customers. We observe in the above theorem, that the limiting probabilities depend on λ and μ only through ρ .

The Camcorder Principle

Suppose we make a video recording of a queueing system with arrivals and service completions. We could replay our video either in slow motion or in high speed. Certain properties of the system would not depend on the speed of playback while other properties would change. For example, the proportion of time that the queueing system is empty

would not depend on the speed. Similarly the probability (or proportion of time) that there are i customers in the system and the expected number of customers would not depend on the speed. However, if we double the speed, the time to serve a customer W_s and the time a customer must wait in the system W would be halved.

We know that the expected number of customers in an M/M/1 queueing system depends on λ and μ as these are the only parameters of the system, but based on our heuristic argument, multiplying the speed by any c would not affect the expected number of customers in the system. Let L be the expected number of customers in the system. Then the expected number of customer is, for some function g ,

$$E(L) = g(\lambda, \mu) = g(c\lambda, c\mu) \text{ for all } c > 0.$$

Choose $c = 1/\mu$ to get $E(L) = g(c\lambda, c\mu) = g(\lambda/\mu, 1) = g(\rho, 1)$. Thus, we see that $E(L)$ is a function of ρ alone. The same argument can be made for several other characteristics of queueing systems.

Theorem 2. *For an M/M/1 queueing system with arrival rate λ , service rate μ , and $\rho = \lambda/\mu$, the expected number of customers in the system is*

$$E(L) = \frac{\rho}{1 - \rho}.$$

Proof.

$$E(L) = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i\rho^i(1 - \rho) = \frac{\rho}{1 - \rho}.$$

PASTA Principle

PASTA is an acronym for Poisson Arrivals See Time Averages [8,9]. Since an M/M/1 system has arrivals entering the system according to a Poisson process, the PASTA principle does apply. For an example of where PASTA does not apply, suppose we had a queueing system with arrivals exactly 1 min apart and service times of 50 s (i.e.,

arrivals do not form a Poisson process). If we begin with one customer in the system, then there will be 50 s periods with exactly one customer in the system and 10 s periods with exactly zero customers in the system. Thus the average number of customers (over time) is 5/6. However, every arriving customer will see a system with zero customers. In contrast, if arrivals enter the system according to a Poisson process, the expected number of customers in the system at the time of arrival will be $E(L)$, the length of the system averaged over time.

Theorem 3. *For an M/M/1 queueing system with arrival rate λ , service rate μ , and $\rho = \lambda/\mu$, the expected wait for a randomly arriving customer is*

$$E(W) = \frac{1}{\mu - \lambda}.$$

Proof. According to the PASTA principle, a randomly arriving customer will see i customers in the system with probability π_i . If there are i customers in the system when a new customer arrives, the expected time to complete service for the new customer will be the sum of the service times for the i customers in the system plus the service time for the new customer. Each customer has mean service time $1/\mu$ because service times are exponentially distributed (and hence memoryless). This even applies to the customer in service (if any). Thus the expected service time for a randomly arriving customer, given that there are i customers currently in service, is $\frac{i+1}{\mu}$. Hence

$$\begin{aligned} E(W) &= \sum_{i=0}^{\infty} \pi_i \frac{i+1}{\mu} = \frac{1}{\mu} \sum_{i=0}^{\infty} i\pi_i + \frac{1}{\mu} \sum_{i=0}^{\infty} \pi_i \\ &= \frac{E(L) + 1}{\mu} = \frac{1}{\mu - \lambda}. \end{aligned}$$

We now introduce (and repeat) some common notation:

- L the number of customers in the queueing system (includes customers waiting or in service)

- L_q the number of customers in the queue (this excludes the customer in service, if any)
- L_s the number of customers in service (this will be 0 or 1)
- W system time (includes waiting and service) for a randomly arriving customer
- W_q waiting time in the queue (excludes service time) for a randomly arriving customer
- W_s service time of a randomly arriving customers (excludes any waiting time)

The following is the most famous result in queueing theory, first proved by Little [10].

Theorem 4. [Little's Theorem (Little's Law)]. *For an M/M/1 queueing system*

$$E(L) = \lambda E(W).$$

This theorem is far more general than the special case given above. It also applies to the more general GI/G/s queues and adds the two additional results

$$\begin{aligned} E(L_q) &= \lambda E(W_q), \\ E(L_s) &= \lambda E(W_s). \end{aligned}$$

We have the following important results for an M/M/1 system:

$$\begin{aligned} E(L) &= \frac{\rho}{1 - \rho}, \\ E(L_q) &= \frac{\rho^2}{1 - \rho}, \\ E(L_s) &= \rho, \\ E(W) &= \frac{1}{\mu - \lambda}, \\ E(W_q) &= \frac{\lambda^2}{\mu(\mu - \lambda)}, \\ E(W_s) &= \frac{1}{\mu}. \end{aligned}$$

A result that relates $E(W)$ to $E(L)$ for the M/M/1 queueing system (only) using μ instead of λ is $E(W) = \frac{E(L)+1}{\mu}$.

For an M/M/1 queueing system, the system will alternate between being empty

and having at least one customer in service. The length of time between empty systems is called a *busy period*. Since there can be various numbers of customers during such times, one can reasonably expect that the busy period distribution would be quite complex. However, we can get information about the busy period in a number of clever ways.

Alternating Renewal Process Approach

An $M/M/1$ queueing system alternates between being empty and nonempty. A renewal occurs when a system becomes empty. Since interarrivals are exponentially distributed with rate λ , the system stays empty for an exponentially distributed time with expected value $1/\lambda$. Then the server stays busy for time B with expected value $E(B)$. But we know that the long run proportion of time that the system is empty is $\pi_0 = 1 - \rho$, so $1 - \rho = \frac{1/\lambda}{E(B) + 1/\lambda}$. Solving gives $E(B) = \frac{\lambda}{\mu(\mu - \lambda)}$.

Probabilistic Interpretation of Laplace Transform Approach

In queueing theory, many results are expressed in terms of Laplace Transforms. Recall that if X is a random variable with positive support and probability density function (pdf) $f(x)$, then the Laplace transform of $f(x)$ is given by

$$L_X(s) = \int_0^\infty f(x)e^{-sx} dx.$$

It can be shown that $L_X(s) = P(X < Y)$, where Y is exponentially distributed with rate s and independent of X . Here Y is sometimes called a *catastrophe*.

Theorem 5. *If B is the busy period of an $M/M/1$ queueing system, then*

$$L_B(s) = \frac{(\lambda + \mu + s) - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{\lambda}.$$

Proof. $L_B(s) = P(B < Y)$ where Y is exponentially distributed at rate s and independent

of B . A busy period begins with the arrival of a single customer to an empty system. The next event is a service completion, an arrival or a catastrophe Y . If the next event is a service completion, then $B < Y$. If the next event is arrival, then we are in the original situation except that we have two customers in the system and we must complete two busy periods before the catastrophe. Thus

$$\begin{aligned} L_B(s) &= P(B < Y) \\ &= \frac{\mu}{\lambda + \mu + s} + \frac{\lambda}{\lambda + \mu + s} L_B(s)^2. \end{aligned}$$

This is a quadratic in $L_B(s)$ and we get two possible solutions.

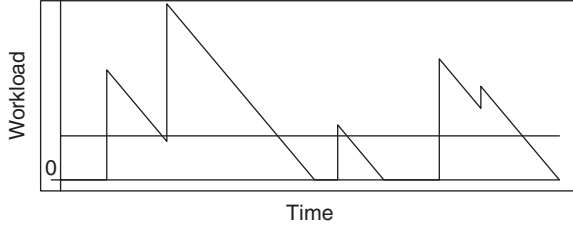
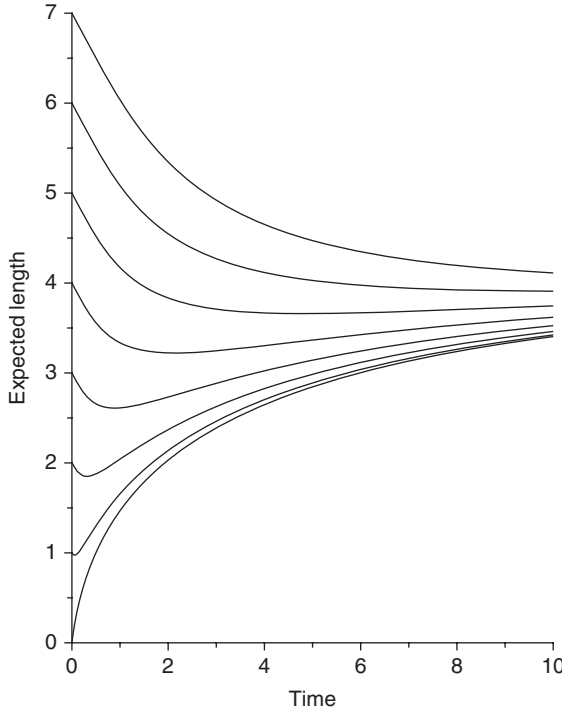
$$L_B(s) = \frac{(\lambda + \mu + s) \pm \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{\lambda}.$$

We have two possible solutions because of the $\pm$ in the numerator. However, since we now know that a Laplace Transform is actually a probability, we reject the “+” possibility as it would make the $L_B(0) = 2$.

Level Crossing Method

Another important consideration in the study of $M/M/1$ queues is the distribution of the waiting time W_q for a randomly arriving customer. One of the fastest ways of obtaining this result is to use Brill’s level crossing method [11]. In this method, the down crossing rate of any level equals the upcrossing rate. However, unlike the rate-in equals rate-out approach, the level crossing method remarkably gets enough information to determine the entire pdf. Define the server workload at time t to be the time that the server needs to complete service for all customers in the system at time t . This workload is a decreasing function (with slope -1) at every point except where an arrival occurs and the value jumps by the service time requirement of the arrival. See Fig. 4 for a typical workload graph.

By the PASTA principle, W_q has the same pdf $f(w)$ as the workload. For a fixed w , the down-crossing rate (for $w > 0$) is $f(w)$. The upcrossing rate can be partitioned into

Figure 4. Workload for an $M/M/1$ queue.Figure 5. Expected $M/M/1$ queue length at time t .

two cases—either the randomly arriving customer encounters an empty system (with probability $\pi_0 = 1 - \rho$) and the jump in workload is greater than w or the system is at level $y > 0$ and the jump in workload is greater than $w - y$. The jump size from 0 is greater than w or $w - y$ with probability $e^{-\mu w}$ or $e^{-\mu(w-y)}$ respectively. This gives rise to the integral equation

$$f(w) = (1 - \rho)\lambda e^{-\mu w} + \int_0^x \lambda f(y) e^{-(w-y)\mu} dy,$$

where $(1 - \rho) + \int_0^\infty f(w) dw = 1$. It is easy to see check that

$$f(w) = \rho(\mu - \lambda)e^{-(\mu-\lambda)w} \text{ for } w > 0,$$

with $P(W_q = 0) = 1 - \rho$, is the solution.

A related result states that the pdf of the system time of a random arriving customer to an $M/M/1$ queueing system is $g(w) = (\mu - \lambda)e^{-(\mu-\lambda)w}$ for $w > 0$.

TRANSIENT RESULTS

Steady state ($t \rightarrow \infty$) results for an $M/M/1$ queueing system are much easier to derive and much simpler to express than transient results (for finite t). Jain *et al.* [12] devote considerable attention to transient queueing results. One key result in transient analysis is given by the following theorem.

Theorem 6. Assume that there are i customers at time 0 in an M/M/1 queueing system, with arrival rate λ and service rate μ . Let $p_{in}(t)$ be the probability of having n customers at time t . Then

$$p_{in}(t) = e^{-(\lambda+\mu)t} [\rho^{(n-i)/2} I_{n-i}(2\sqrt{\lambda\mu t}) + \rho^{(n-i-1)/2} I_{n+i+1}(2\sqrt{\lambda\mu t})] + e^{-(\lambda+\mu)t} (1-\rho) \rho^n \sum_{j=n+i+2}^{\infty} I_j(2\sqrt{\lambda\mu t}) / \rho^{j/2},$$

where $I_n(t)$ is the modified Bessel function of the first kind.

Since we have an expression for the transient probabilities $p_{in}(t)$, we can find other performance measures such as the expected number of customers $E(L_i(t))$ in the system at time t if the system has i customers at time 0. Figure 5 shows $E(L_i(t))$ versus t for $\lambda = 4$, $\mu = 5$, and $i = 0, 1, \dots, 7$. All curves tend to $E(L) = \frac{\rho}{1-\rho}$ as $t \rightarrow \infty$.

Acknowledgments

The figures in this article were created using R and Maple. Figures 3 and 5 were prepared by Shan XU.

REFERENCES

1. Gross D, Shortle J, Thompson J, *et al.* Fundamentals of Queueing Theory. 4th ed. Hoboken (NJ): John Wiley and Sons; 2008.
2. Kleinrock L. Queueing systems. Volume 1, Theory. New York: John Wiley and Sons; 1975.
3. Bolch G, Greiner S, de Meer H, *et al.* Queueing networks and Markov Chains. 2nd ed. Hoboken (NJ): John Wiley and Sons; 2006.
4. Medhi J. Stochastic models in Queueing Theory. 2nd ed. Amsterdam: Academic Press; 2003.
5. Hlynka M. Myron Hlynka's Queueing Theory web site. Available at <http://web2.uwindsor.ca/mathhlynka/queue.html>. Accessed 2010 August.
6. Robert P. Stochastic networks and queues. New York: Springer; 2003.
7. Asmussen S. Applied probability and queues. 2nd ed. New York: Springer; 2003.
8. Wolff RW. Poisson arrivals see time averages. Oper Res 1982;30:223–231.
9. Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice Hall; 1989.
10. Little JDC. A proof of the queueing formula: $L = \lambda W$. Oper Res 1961;9:383–387.
11. Brill P. Level crossing methods in stochastic models. New York: Springer; 2008.
12. Jain JL, Mohanty SG, Bohm W. A course on queueing models. Boca Raton (FL): Chapman and Hall; 2007.

THE $M/M/\infty$ QUEUE

BERNARDO D'AURIA
Departamento de Estadística,
Universidad Carlos III de
Madrid, Madrid, Spain

INTRODUCTION

The $M/M/\infty$ queue is one of the simplest queueing models and as such, is often used as a first approximation for any queueing system with no waiting line. Examples of these kind of systems may be encountered in physics; for example, radioactive emissions [1,2], highway traffic [3], telecommunication systems [4], emergency appliances' management in hospitals [5], and computer systems [6], to list just a few.

Following Kendall's notation, the abbreviation $M/M/\infty$ means that the arrival process is Markovian or memoryless (M), the service times follow an exponential distribution (the only continuous memoryless distribution (M)), and ∞ signifies that the system has an infinite set of servers to satisfy customers' requests. The system has no waiting room, thus a new customer begins to be served immediately and leaves the system as soon as his/her exponential service request is satisfied. The first contribution to the analysis of the stationary behavior of the $M/M/\infty$ queue was due to Erlang [7].

The popularity of the $M/M/\infty$ queue can be attributed to a simple form of the steady-state distribution of the number of customers in the system, given by a Poisson random variable. The mean of this random variable, often referred to as the *traffic intensity*, is the ratio of the arrival and the service rates. This special form of the stationary distribution is shared by all $M/G/\infty$ systems, a property known as *insensitivity*. According to this property, the stationary distribution of the number of customers in the system depends on the

service distribution only through the service distribution mean and not its full form [8].

MODEL DESCRIPTION

We begin our considerations with the $M/M/\infty$ system starting empty at time 0. Let (τ_n, S_n) with $n > 0$ be a sequence of pairs such that for each n , τ_n is the n th interarrival time, the time interval between the arrivals of the $(n-1)$ th and the n th customers, and S_n is the service request of the n th customer.

We can easily see that $T_n = \sum_{i=1}^n \tau_i$ is the arrival time of the n th customer, $n > 0$, and taking into account that in the infinite server queue there is no waiting time, we can directly compute the departure time of the n th customer, D_n , as $D_n = T_n + S_n$.

The above definitions are valid for any infinite server queue, however in the special case of the $M/M/\infty$ queue, it is assumed that the sequences τ_n and S_n are independent of each other, both are iid exponential, the former with parameter $\lambda > 0$ and the latter with $\mu > 0$. Denoting by (τ, S) any generic pair of the underlying sequence, we see that the mean interarrival and service times are $\mathbb{E}[\tau] = 1/\lambda$ and $\mathbb{E}[S] = 1/\mu$, respectively, and hence the traffic intensity is $\rho = \lambda/\mu$.

Number of Customers in the System at Time t

The quantity of main interest in the queueing system is the number of customers present at time t denoted by $Q(t)$. This quantity can be expressed in terms of T_n and S_n as

$$Q(t) = \sum_{n=1}^{\infty} 1\{T_n \leq t < T_n + S_n\},$$

where $1\{\cdot\}$ is the indicator function of the set $\{\cdot\}$. The contribution of the n th customer to the sum takes place only if s/he is present in the queue at time t , which means that t falls in the interval $[T_n, D_n) = [T_n, T_n + S_n)$.

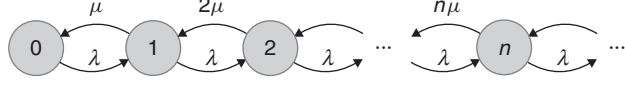


Figure 1. Transition diagram.

Keeping in mind that all variables considered are exponential, it can be easily seen that $Q(t)$ is a continuous-time Markov chain, which in particular can be viewed as a birth–death process with birth and death rates λ_n and μ_n given by

$$\begin{cases} \lambda_n = \lambda \\ \mu_n = n\mu, \end{cases} \quad n \geq 0. \quad (1)$$

In Fig. 1 the corresponding transition diagram is sketched.

If we define $p_n(t) = \Pr\{Q(t) = n\}$, it is possible to deduce the transient behavior of the system by an infinite system of differential equations given below, which are valid for the birth–death process:

$$\begin{cases} p'_0(t) = -\lambda p_0(t) + \mu p_1(t), & n = 0, \\ p'_n(t) = -(\lambda + n\mu)p_n(t) + \lambda p_{n-1}(t) \\ \quad + (n+1)\mu p_{n+1}(t), & n \geq 1, \end{cases} \quad (2)$$

which solved for $p_0(0) = 1$ leads to

$$p_n(t) = \frac{v(t)^n}{n!} e^{-v(t)}, \quad (3)$$

with $v(t) = \lambda \int_0^t e^{-\mu(t-x)} dx$. Therefore, the transient distribution of $Q(t)$ is a Poisson distribution with parameter $v(t)$

$$Q(t) \sim \text{Po}(\rho(1 - e^{-t\mu})). \quad (4)$$

STATIONARY BEHAVIOR

By allowing $t \rightarrow \infty$ in Equation (4), we get $v(t) \rightarrow \lambda/\mu = \rho$ and

$$Q(\infty) \sim \text{Po}(\lambda/\mu), \quad (5)$$

which can be interpreted as follows: under the stationary regime in the $M/M/\infty$ system with traffic intensity ρ , the number of customers in the system follows a Poisson distribution with mean ρ .

Another way of arriving at this result is to solve the stationary version of the infinite system (Eq. 2). After canceling the derivatives with respect to time, we obtain

$$\begin{cases} 0 = -\lambda p_0 + \mu p_1, & n = 0, \\ 0 = -(\lambda + n\mu)p_n + \lambda p_{n-1} \\ \quad + (n+1)\mu p_{n+1}, & n \geq 1, \end{cases} \quad (6)$$

with the ratio

$$\frac{p_n}{p_{n-1}} = \frac{\rho}{n},$$

which taking into account the normalizing condition $\sum_n p_n = 1$ yields

$$p_n = \frac{\rho^n}{n!} e^{-\rho}.$$

It is worth noticing that in this case the queue is always stable for any value of $\rho > 0$.

To construct a stationary version of the process, it is sufficient to consider an additional iid sequence of pairs (τ_n, S_n) now indexed over the negative integers $n < 0$, that corresponds to the sequence of the customers arriving before time 0. The arrival time of the n th customer, where $n < 0$, is given by $T_n = -\sum_{i=n}^{-1} \tau_i$. Subsequently, the augmented sequence $\{T_n\}_{n \neq 0}$ forms a stationary homogeneous Poisson point process on the real line.

Rate of Convergence to the Stationary Regime

Let $Q^*(t)$ be the stationary version of the process and $Q(t)$, as before, the system starting empty at time 0. Then, $Q^*(0) \geq Q(0) = 0$ a.s. and the coupling time $T^* = \inf_{t \geq 0} \{Q^*(t) = Q(t)\}$ is the moment of departure of the last of the customers present in the stationary system at time 0, with the number of those customers given by $Q^*(0)$. The tail of the

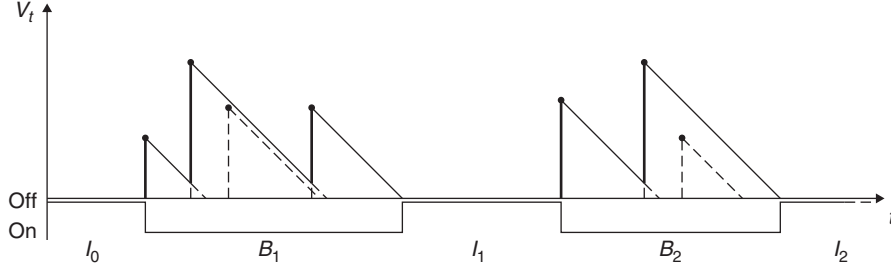


Figure 2. Busy and idle periods.

distribution of T^* can be calculated as

$$\begin{aligned}
 \Pr\{T^* > t\} &= \Pr\{\max\{S_1^*, \dots, S_{Q^*(0)}^*\} < t\} \\
 &= 1 - \mathbb{E}[\mathbb{E}[1\{\max\{S_1^*, \dots, S_{Q^*(0)}^*\} > t\} \\
 &\quad | Q^*(0)]] \\
 &= 1 - \mathbb{E}\left[(e^{-\mu t})^{Q^*(0)}\right] = 1 - \phi(e^{-\mu t}) \\
 &= 1 - e^{-\rho e^{-\mu t}},
 \end{aligned}$$

where $\phi(z) = e^{-\rho(1-z)}$ is the characteristic function of a Poisson distribution with parameter ρ , and $S_1^*, \dots, S_{Q^*(0)}^*$ are the residual service times of the customers present in the stationary system at time 0.

The upper bound on the convergence rate of the empty-started system to its stationary regime can be easily derived by employing the *coupling time inequality* (see Ref. 9)

$$d_v(Q(t), Q^*(t)) \leq \Pr\{T^* > t\} = 1 - e^{-\rho e^{-\mu t}}, \quad (7)$$

where by

$$d_v(X, Y) = \frac{1}{2} \sum_{k=0}^{\infty} |\Pr\{X = k\} - \Pr\{Y = k\}|;$$

we denote the total variation distance between the probability distributions of two arbitrary discrete random variables X and Y .

THE BUSY PERIOD DISTRIBUTION

In the M/M/∞ queue, a busy period is any time interval during which the system is continuously busy, that is with at least one customer in service. The busy period begins

when an empty system is entered by a new customer and ends as soon as the last served customer leaves the system empty again. The time interval between two consecutive busy periods is referred to as an idle period, during which the system remains empty.

The alternation of busy and idle periods can be viewed as an “On–Off” renewal process with the busy periods corresponding to the system being On and the idle to Off. In our considerations we assume that the system starts with an idle period of length I_0 followed by an interchange of iid pairs $\{(B_n, I_n)\}_n$, $n \geq 1$. In Fig. 2 a graphical representation of this renewal process is displayed, where V_t is the virtual remaining service time, that is the time the system would require to become empty if the arrival process were to be stopped at time t .

In this section we are concerned with the distribution of the generic pair (B, I) . Given the Poisson nature of the arrival stream, we immediately see that $I \sim \text{Exp}(\lambda)$. Moreover, since the sequence of the service requests is independent of that of the arrival times, we also notice that B and I are independent. In what follows, we derive the distribution of B using the original idea of Takács, see Ref. 11.

Let $N(t)$ be the process that counts the number of busy periods started before time t :

$$\{N(t) \geq n\} = \left\{ \sum_{k=0}^n C_k \leq t \right\},$$

where we set the cycle length $C_n = B_n + I_n$ with $B_0 = 0$. Notice that C_0 has a different distribution from the other C_n 's.

Let us introduce the (delayed) renewal function $m(t) = \mathbb{E}[N(t)]$, which satisfies the

following equation

$$m'(t) = \lambda p_0(t). \quad (8)$$

Hence, the rate of having a new busy period starting at time t is given by the joint occurrence of two events: the system being idle at time t^- and a new arrival at time t with rate λ . Now, $p_0(t)$, defined in Equation (3) and involved in the last equation can be written as $p_0(t) = \Pr\{Q(t) = 0\} = \Pr\{t \in \text{Idle period}\}$. Let $\pi(s) = \int_0^\infty e^{-st} p_0(t) dt$ be its Laplace transform.

We obtain the following

$$\begin{aligned} \Pr\{t \in \text{Busy period}\} &= 1 - p_0(t) \\ &= \int_0^t \Pr\{B > t - x\} dm(x), \end{aligned} \quad (9)$$

where we split the event $\{t \in \text{Busy period}\}$ into disjoint events, assuming that the busy period active at time t started at x , $x \in (0, t]$.

Let $\phi_B(s) = \mathbb{E}[e^{-sB}]$ be the Laplace–Stieltjes transform of B , and $\phi_I(s) = \lambda/(\lambda + s)$ that of I . Combining Equations (9) and (8) leads to

$$\phi_B(s) = \frac{1}{\phi_I(s)} - \frac{1}{\lambda \pi(s)} = \frac{B_1(s)}{B_0(s)}, \quad (10)$$

where

$$\begin{aligned} B_n(s) &= B_n(s; \lambda, \mu) \\ &= \int_0^1 (1-x)^n x^{s/\mu-1} e^{\lambda x/\mu} dx \\ &= \frac{\Gamma(n+1) \Gamma(s/\mu)}{\Gamma(n+1+s/\mu)} M\left(\frac{s}{\mu}, n+1+\frac{s}{\mu}, \frac{\lambda}{\mu}\right), \end{aligned} \quad (11)$$

where $M(a, b, z)$ is the hypergeometric ${}_1F_1$ generalized function; see formula (13.2.1) in Ref. 12.

The length of a busy period can also be seen as the first hitting time of level 0 by the process $Q(t)$, given that the process started at level 1, that is $B =_d T_1^*$ with $T_1^* = \inf_{t>0}\{Q(t) = 0 | Q(0) = 1\}$. This particular interpretation allows for the following generalization to be considered.

C-Congestion Periods and General Hitting Times

The C -congestion period, T_C^* , is defined as the first hitting time of level 0 for the process $Q(t)$, assuming that it started at level C . Compared with a more general notation of the hitting times,

$$T_a^*(b) = \inf_{t>0} \{Q(t) = b | Q(0) = a\},$$

we see that the argument b for the case of $b = 0$ has been suppressed. It follows that a busy period is nothing else than an 1-congestion period.

In Ref. 13 it was shown that

$$\phi_{T_C^*}(s) = \mathbb{E}[e^{-sT_C^*}] = \frac{B_C(s)}{B_0(s)},$$

whereas the Laplace transform of the hitting time $T_a^*(b)$ is

$$\phi_{T_a^*(b)}(s) = \begin{cases} B_a(s)/B_b(s), & 0 \leq a \leq b, \\ \Gamma_a(s)/\Gamma_b(s), & 0 \leq b \leq a, \end{cases} \quad (13)$$

with

$$\begin{aligned} \Gamma_n(s) &= \Gamma_n(s; \lambda, \mu) \\ &= \int_0^\infty (1 + \mu x/\lambda)^n x^{s/\mu-1} e^{-x} dx. \end{aligned} \quad (14)$$

For a proof of Equation (13) and additional information on historical references; see Proposition 6.7 in Ref. 10.

POISSON RANDOM MEASURES: A DIFFERENT PERSPECTIVE

An elegant and more powerful analysis of the model can be obtained by taking the single-sided or the double-sided sequences of pairs $\{(T_n, S_n)\}_n$ with $n > 0$ or $n \neq 0$ respectively, and regarding them as a collection of points in a two-dimensional space [14,15]. By doing so, we obtain a bidimensional random measure

$$N(A) = \sum_n \delta_{(T_n, S_n)}(A),$$

which is a Poisson Random Measure (PRM), such that for any borel set $A \subset \mathbb{R}^+ \times \mathbb{R}^+$

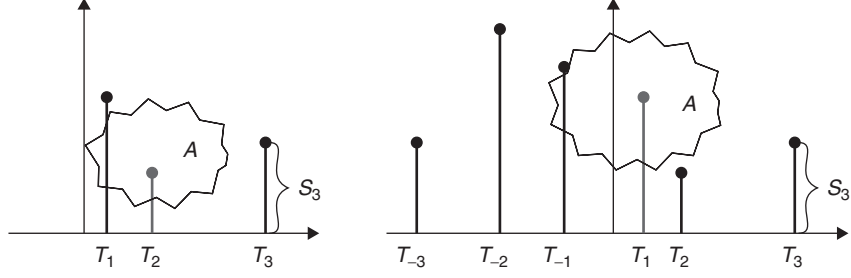


Figure 3. Poisson random measures perspective.

($A \subset \mathbb{R} \times \mathbb{R}^+$), it counts the number of points $\{(T_n, S_n)\}_n$ captured by A .

This result is a consequence of the fact that the sequence $\{T_n\}_n$ generates a Poisson point process and that the $\{S_n\}_n$ are independent. For more details see, for example, Proposition 3.8 in Ref. 16.

Figure 3 shows a measure-based representation of the number of customers in the $M/M/\infty$ system for the transient ($n > 0$) and the stationary ($n \neq 0$) scenarios.

From the PRM formulation, we get right away that for any borel set $A \subset \mathbb{R}^+ \times \mathbb{R}^+$ ($A \subset \mathbb{R} \times \mathbb{R}^+$)

$$N(A) \sim \text{Po}(|A|),$$

where $|A|$ is the measure of set A under the intensity measure $\nu(A)$ defined by

$$\nu(dt \times ds) = \lambda \mu e^{-\mu s} dt ds.$$

Additionally, any two disjoint sets A and B imply the independence of $N(A)$ and $N(B)$ written as

$$N(A) \perp N(B).$$

Putting everything together and referring to Fig. 4 we observe that

$$Q(t) = N(A_t) \quad \text{and} \quad Q(\infty) =_d Q^*(0) = N(A_0^*),$$

with

$$A_t = \{(x, y) | 0 \leq x \leq t, y > t - x\},$$

$$A_0^* = \{(x, y) | x \leq 0, y > |x|\}$$

and

$$|A_t| = \nu(A_t) = \nu(t),$$

$$|A_0^*| = \nu(A_0^*) = \rho$$

which leads to Equations (3) and (5).

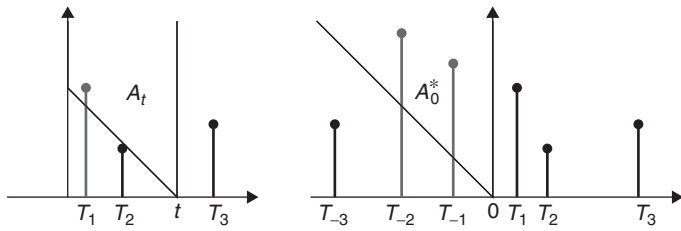


Figure 4. Number of customers in the system.

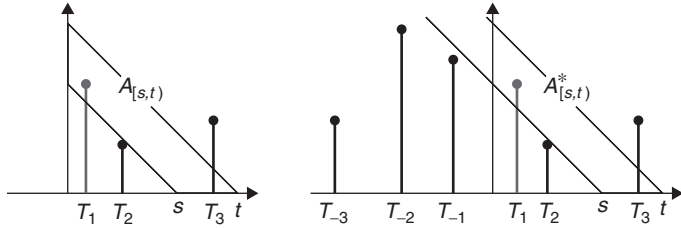


Figure 5. Number of departures in the interval $[s, t]$.

The Departure Process

As shown in Fig. 5, the number of departures in the interval $[s, t)$ corresponds to the number of points falling into

$$A_{[s,t)} = \{(x, y) | 0 \leq x \leq t, (s - x)^+ \leq y < t - x\}$$

and

$$A_{[s,t)}^* = \{(x, y) | x \leq t, (s - x)^+ \leq y < t - x\},$$

with $(a)^+ = \max\{a, 0\}$.

As such, the output process of the $M/M/\infty$ system is a Poisson process, which in the transient case has rate

$$\begin{aligned} \lambda(t) &= \frac{d}{dt} v(A_{[0,t)}) = \frac{d}{dt} \int_0^t (1 - e^{-\mu(t-x)}) dx \\ &= \lambda(1 - e^{-\mu t}) \end{aligned}$$

and λ in the stationary case.

FURTHER READING

Further details on the classical $M/M/\infty$ queue can be found in standard textbooks such as Refs 17–23, to mention just a few. Nevertheless, the standard analysis of the $M/M/\infty$ queue encountered in books is often restricted to the proof of the stationary distribution of the number of customers given by Equation (5). A deeper analysis is generally reserved for the $M/G/\infty$ system that is a generalization of it. The exception is the textbook by Robert, Ref. 10, in which an entire chapter is dedicated to the Markovian system. An in-depth analysis of that chapter is based on the original martingale technique, that employs the earlier results of Ref. 24.

The treatment of the asymptotic properties of the $M/M/\infty$ system, can be found in, for example, Refs 25–28 and the references therein, while some interesting results on the analysis of the busy and congestion periods, in Refs 13, 29–34.

The transient analysis of the $M/M/\infty$ system and its $M/G/\infty$ extension have also been studied extensively, see for example Refs 5, 35, and 36 and the references therein.

Several authors provided numerous extensions of the underlying model to incorporate more complex and more realistic dynamics. In some cases, this has been achieved by introducing some level of dependence in the customers' behavior by allowing the service and the arrival rates to depend on the external random environment. Work in this direction can be found in Refs 37–41.

REFERENCES

1. Takács L. On a probability problem arising in the theory of counters. *Proc Camb Philos Soc* 1956;52:488–498.
2. Pyke R. On renewal processes related to type I and type II counter models. *Ann Math Stat* 1958;29(3):737–754.
3. Baykal-Gursoy M, Xiao W. Stochastic decomposition in $M/M/\infty$ queues with Markov-modulated service rates. *Queueing Syst* 2004;48:75–88.
4. Kelly FP. Loss networks. *Ann Appl Probab* 1991;1(3):319–378.
5. Collings T, Stoneman C. The $M/M/\infty$ queue with varying arrival and departure rates. *Oper Res* 1976;24(4):760–773.
6. Coffman EG, Kadota TT, Shepp LA. A stochastic model of fragmentation in dynamic storage allocation. *SIAM J Comput* 1985;14:416–425.
7. Syski R. Volume 3, Introduction to congestion theory in telephone systems. 2nd ed. Amsterdam: North-Holland Publishing Co.; 1986.
8. Adan I, Resing J. Queueing theory. The Netherlands: Tue, Eindhoven; 2001. Available at <http://www.win.tue.nl/~iadan/queueing.pdf>.
9. Thorisson H. Coupling, stationarity and regeneration. New York: Springer; 2000.
10. Robert P. Stochastic networks and queues. New York: Springer; 2000.
11. Takács L. On a probability problem in the theory of counters. *Ann Math Stat* 1958;29(4):1257–1263.
12. Abramowitz M, Stegun IA. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover; 1964.
13. Guillemin F, Simonian A. Transient characteristics of an $M/M/\infty$ system. *Adv Appl Probab* 1995;27(3):862–888.
14. Franken P, König D, Arndt U, et al. Queues and point processes. New York: John Wiley & Sons, Inc.; 1983.

15. Baccelli F, Brémaud P. Elements of queueing theory: palm–martingale calculus and stochastic recurrences. Berlin: Springer; 2003.
16. Resnick SI. Extreme values, regular variation and point processes. New York: Springer; 1987.
17. Takács L. Introduction to the theory of queues. New York: Oxford University Press; 1962.
18. Feller W. Volume 2, An introduction to probability theory and its applications. 2nd ed. New York: Wiley; 1971.
19. Neuts M. Matrix–geometric solutions in stochastic models: an algorithmic approach. Baltimore: Johns Hopkins University Press; 1981.
20. Borovkov AA. Stochastic processes in queueing theory. New York: Springer; 1976.
21. Ross SM. Stochastic processes. 2nd ed. New York: John Wiley & Sons, Inc.; 1996.
22. Brémaud P. Markov chains. New York: Springer; 1999.
23. Asmussen S. Applied probability and queues. 2nd ed. New York: Springer; 2003.
24. Fricker C, Robert P, Tibi D. On the rates of convergence of Erlang’s model. *J Appl Probab* 1999;36(4):1167–1184.
25. Borovkov AA. Asymptotic methods in queueing theory. New York: John Wiley; 1984.
26. Newell GF. The $M/M/\infty$ service system with ranked servers in heavy traffic, Lecture notes in economics and mathematical systems. Berlin, New York: Springer; 1984.
27. Aldous D. Some interesting processes arising as heavy traffic limits in an $M/M/\infty$ storage process. *Stoch Processes Appl* 1986;22(2):291–313.
28. Knessl C. Some asymptotic results for the $M/M/\infty$ queue with ranked servers. *Queueing Syst Theor Appl* 2004;47(3):201–250.
29. Preater J. $M/M/\infty$ transience revisited. *J Appl Probab* 1997;34(4):1061–1067.
30. Guillemin F, Pinchon D. On a random variable associated with excursions in an $M/M/\infty$ system. *Queueing Syst Theor Appl* 1999;32(4):305–318.
31. Knessl C, Yang YP. Asymptotic expansions for the congestion period for the $M/M/\infty$ queue. *Queueing Syst* 2001;39:213–256.
32. Preater J. On the severity of $M/M/\infty$ congested episodes. *J Appl Probab* 2002;39(1):228–230.
33. Tsybakov B. Busy periods in $M/M/\infty$ systems with heterogeneous servers. *Queueing Syst Theor Appl* 2006;52(2):153–156.
34. Roijers F, Mandjes M, van den Berg H. Analysis of congestion periods of an $M/M/\infty$ -queue. *Perform Eval* 2007;64:737–754.
35. Whitt W. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Manag Sci* 1991;37:307–314.
36. Eick SG, Massey WA, Whitt W. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Manag Sci* 1993;39(2):241–252.
37. O’Cinneide C, Purdue P. The $M/M/\infty$ queue in a random environment. *J Appl Probab* 1986;23:175–184.
38. Keilson J, Servi LD. The matrix $M/M/\infty$ system: retrials models and Markov modulated sources. *Adv Appl Probab* 1993;25:453–471.
39. D’Auria B. Stochastic decomposition of the $M/G/\infty$ queue in a random environment. *Oper Res Lett* 2007;35:805–812.
40. D’Auria B. $M/M/\infty$ queues in semi-Markovian random environment. *Queueing Syst* 2008;58(3):221–237.
41. Fralix BH, Adan IJ. An infinite-server queue influenced by a semi-Markovian environment. *Queueing Syst Theor Appl* 2009;61(1):65–84.

THE $M/M/s$ QUEUE

M. YASIN ULUKUS

Department of Industrial Engineering,
University of Pittsburgh, Pittsburgh,
Pennsylvania

INTRODUCTION

The $M/M/s$ queueing system is one of the most fundamental models for designing, analyzing, and controlling stochastic service systems. Since the introduction of the model by Erlang nearly a century ago for telephone engineering applications, new applications have emerged, such as customer contact centers, hospitals, and web servers. Most of the main results for the model were developed within the first three decades of the twentieth century.

Using Kendall's notation [1], the nomenclature $M/M/s$ is described as follows: The arrival process is Markovian or memoryless (M); the service times are also memoryless (M); there are s identical servers ($1 \leq s < \infty$); and there is an infinite-capacity waiting room for customers who cannot be served upon arrival. Customers who arrive to find all s servers busy join the queue and wait as long as necessary for service, so that no customers are lost. Because there is a possibility that customers experience delay in the queue before they gain access to a server, the system is often called the *Erlang delay system*. Note that, if $s = 1$, the system is an $M/M/1$ queueing model (see **The $M/M/1$ Queue**), and when $s = \infty$, we have the $M/M/\infty$ system (see **The $M/M/\infty$ Queue**).

Although the $M/M/s$ model imposes some restrictive assumptions, such as exponential service times, it can still be utilized to design and analyze a variety of real service systems thanks, in part, to the simplicity of results describing its limiting behavior. For example, the model can be utilized to determine the optimal number of agents in a customer contact center (e.g., a call center)

to reduce the average waiting time in queue or the probability of being delayed when placing a call. In health-care service applications, the model can be used to determine the optimal number of beds, or the appropriate nurse staffing level, to achieve a minimum quality-of-service guarantee. The $M/M/s$ model can also be used to design a hospital ward that minimizes the proportion of patients who are forced to wait before being admitted. Besides these design considerations, the model provides an estimate of the utilization of the system's resources—an important indicator of efficiency.

The remainder of the article is organized as follows. The next section provides a mathematical description of the $M/M/s$ system and describes the steady-state (or limiting) behavior of the system. Subsequently, we discuss the $M/M/s/c$ system as a special case of $M/M/s$. The next two sections discuss the Erlang C formula, its structural properties, and computational issues. The final section provides suggestions for reading on the Erlang delay model.

MODEL DESCRIPTION AND PERFORMANCE PARAMETERS

The $M/M/s$ queue is a model for a stochastic service system to which customers arrive according to a homogeneous Poisson process with rate parameter λ ($\lambda > 0$), and the service requirement of each customer is distributed exponentially with mean $1/\mu$. The *offered load* to the system is the dimensionless quantity

$$a \equiv \frac{\lambda}{\mu}.$$

The offered load per server, often termed the *traffic intensity*, is given by

$$\rho \equiv \frac{\lambda}{s\mu} = \frac{a}{s}.$$

Owing to the Markovian nature of the arrival process and the service times, the state of

the system is completely characterized by the total number of customers in the system (i.e., the number of customers currently in service and those waiting in the queue to receive service). Denote by $Q(t)$ the number of customers in the system at time t , and note that $\{Q(t) : t \geq 0\}$ is a continuous-time Markov chain (CTMC) on the denumerable state space $S = \{0, 1, 2, \dots\}$. Under suitable conditions, as $t \rightarrow \infty$, the random variable $Q(t)$ converges in distribution to a random variable describing the steady-state number of customers in the system. Let Q denote this random variable and define the probability mass function (p.m.f.) of Q by

$$p_j = \lim_{t \rightarrow \infty} P(Q(t) = j) = P(Q = j), \\ j = 0, 1, 2, \dots$$

When the traffic intensity (ρ) is less than 1, the existence of the steady-state distribution, $\{p_j\}$, is ensured. The process, $\{Q(t) : t \geq 0\}$, is a birth-and-death process (see **Birth-and-Death Processes**) with birth rates $\lambda_j = \lambda$ for all j , and death rates

$$\mu_j = \begin{cases} j\mu, & \text{if } 0 \leq j < s, \\ s\mu, & \text{if } j \geq s. \end{cases}$$

The steady-state probabilities can be derived by solving the balance equations [2], and these equations, when solved simultaneously, yield

$$p_j = \begin{cases} \frac{(s\rho)^j}{j!} p_0, & \text{if } 0 \leq j < s, \\ \frac{\rho^j s^s}{s!} p_0, & \text{if } j \geq s, \end{cases} \quad (1)$$

where the normalization condition

$$\sum_{j=0}^{\infty} p_j = 1$$

shows that

$$p_0 = \left(\sum_{j=0}^{s-1} \frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!(1-\rho)} \right)^{-1}. \quad (2)$$

The mean performance parameters, namely, the steady-state mean number of

customers in the system (L), the steady-state mean number of customers in the waiting line (L_q), the steady-state mean time in the system (W), and the steady-state mean time in the waiting line (W_q), can be obtained by using Equations (1) and (2) and Little's Law [3]. The mean queue length can be directly obtained by using the steady-state probabilities

$$L_q = \sum_{j=s+1}^{\infty} (j-s)p_j = \frac{a^s \rho}{s!(1-\rho)^2} p_0. \quad (3)$$

By Little's Law, the mean time in the waiting line is $W_q = L_q/\lambda$,

$$W_q = \frac{a^s}{s!(s\mu)(1-\rho)^2} p_0. \quad (4)$$

The mean time in the system is $W = W_q + 1/\mu$,

$$W = \frac{1}{\mu} + \frac{a^s}{s!(s\mu)(1-\rho)^2} p_0. \quad (5)$$

Finally, the mean number of customers in the system is

$$L = a + \frac{a^s \rho}{s!(s\mu)(1-\rho)^2} p_0. \quad (6)$$

The cumulative distribution function (c.d.f.) of the steady-state delay, or waiting time in queue, is denoted by $F_q(t)$, and the total time in system has c.d.f. $F(t)$. By conditioning on the number of customers seen in the system by an arriving customer [3], and assuming $\rho < 1$, $F_q(t)$ and $F(t)$ are, respectively, given by

$$F_q(t) = 1 - \frac{a^s p_0}{s!(1-\rho)} e^{-(s\mu-\lambda)t}, \quad t \geq 0 \quad (7)$$

and

$$F(t) = \frac{s(1-\rho) - F_q(0)}{s(1-\rho) - 1} (1 - e^{-\mu t}) \\ - \frac{1 - F_q(0)}{s(1-\rho) - 1} (1 - e^{-(s\mu-\lambda)t}), \quad t \geq 0. \quad (8)$$

THE $M/M/s/c$ QUEUE

The $M/M/s/c$ queue is a special case of the $M/M/s$ system in which there is a limit (c) on the total number of customers in the system (i.e., the total number of customers in service and waiting in the queue is limited to c so that the capacity of the waiting area is $c - s$). An arriving customer who finds c customers in the system is lost and does not retry his/her service. The steady-state probabilities can be derived in a straightforward way (as in the $M/M/s$ case) except that the effective arrival rate of customers must account for the fact some proportion of arrivals will be turned away due to blocking. The steady-state probabilities for the $M/M/s/c$ system are given by

$$p_j = \begin{cases} \frac{(s\rho)^j}{j!} p_0, & \text{if } 0 \leq j < s, \\ \frac{\rho^j s^s}{s!} p_0, & \text{if } s \leq j \leq c. \end{cases} \quad (9)$$

To obtain p_0 , the same normalization procedure is utilized as in $M/M/s$ system; however, the series is finite, and, thus, we need not require that $\rho < 1$. Therefore,

$$p_0 = \begin{cases} \left(\sum_{j=0}^{c-1} \frac{a^j}{j!} + \frac{a^s}{s!} \frac{1 - \rho^{c-s+1}}{1 - \rho} \right)^{-1} & \text{if } \rho \neq 1, \\ \left(\sum_{j=0}^{c-1} \frac{a^j}{j!} + \frac{a^s}{s!} (c - s + 1) \right)^{-1} & \text{if } \rho = 1. \end{cases} \quad (10)$$

Note that letting $c \rightarrow \infty$ and restricting $\rho < 1$ in Equation (10) gives the steady-state probabilities of the $M/M/s$ system.

The mean performance parameters can be obtained by using Equations (9) and (10) and Little's Law [3]. The mean queue length can be obtained via the steady-state probabilities ($\rho \neq 1$)

$$L_q = \sum_{j=s+1}^c (j-s)p_j = \frac{p_0 a^s \rho}{s!} \frac{d}{d\rho} \left(\frac{1 - \rho^{c-s+1}}{1 - \rho} \right), \quad (11)$$

after applying L'Hospital's rule,

$$L_q = \frac{p_0 a^s \rho}{s!(1-\rho)^2} \left[1 - \rho^{c-s+1} - (1-\rho)(c-s+1)\rho(c-s) \right]. \quad (12)$$

For $\rho = 1$, it is necessary to apply L'Hospital's rule twice.

To obtain the other mean performance parameters, Little's formula can be used as in the $M/M/s$ system. However, customers who arrive to find c customers in the system are lost. Fortunately, for a system with Poisson arrivals, these arrivals see the time-averaged state of the system (PASTA, Poisson arrivals see time averages, property; see Wolff [4,5]). By PASTA, it follows that effective arrival rate seen by servers is $\lambda' \equiv \lambda(1 - p_c)$. Hence the mean time in the waiting line is $W_q = L_q/\lambda'$,

$$W_q = \frac{L_q}{\lambda(1 - p_c)}. \quad (13)$$

The mean number of customers in the system is

$$L = L_q + \frac{\lambda'}{\mu} = L_q + a(1 - p_c). \quad (14)$$

Finally, the mean time in the system is $W = L/\lambda'$:

$$W = \frac{L}{\lambda(1 - p_c)}. \quad (15)$$

The derivation of the c.d.f. of the delay time in the queue, $F_q(t)$, is a bit more complicated because of the finite series. The derivation can be found in Gross and Harris [3].

$$F_q(t) = 1 - \sum_{j=s}^{c-1} \frac{p_j}{1 - p_c} \sum_{i=0}^{j-s} \frac{(s\mu t)^i e^{-s\mu t}}{i!}. \quad (16)$$

In the next section, we discuss the well-known Erlang C formula which describes the probability that an arbitrary customer experiences delay in the $M/M/s$ system.

ERLANG C FORMULA

An important quantity is the probability that (in steady state) an arriving customer is delayed before receiving service. The PASTA property ensures that the steady-state probability that an arriving customer sees j customers in the system at arrival is given by p_j . Therefore, the steady-state probability that an arriving customer is delayed is $\sum_{j=s}^{\infty} p_j$. This probability, usually denoted by $C(s, a)$, is computed using the so-called *Erlang C formula* given by [2]

$$C(s, a) = \frac{\frac{a^s}{s!(1-\rho)}}{\sum_{i=0}^{s-1} \frac{a^i}{i!} + \frac{a^s}{s!(1-\rho)}}, \quad 0 < a < s. \quad (17)$$

Note that Equation (17) requires that $\rho < 1$ to ensure stability. The Erlang C formula is an important result in designing telecommunication systems, call centers, or service systems, especially to determine the number of trunks, agents, or customer service representatives for a specified desired probability of waiting.

Related to the Erlang C formula is the Erlang B formula, which gives the steady-state probability of blocking for the $M/G/s/s$ (Erlang loss) system (see *The M/G/s/s Queue*). Since there is no waiting space in that system, arriving customers who find all s servers busy are lost and do not return. The service times are assumed to be independent and identically distributed with a finite mean (not necessarily exponential). The Erlang B formula is given by [3]

$$B(s, a) \equiv p_s = \frac{a^s/s!}{\sum_{j=0}^s \frac{a^j}{j!}}. \quad (18)$$

Unlike the Erlang delay model, the Erlang loss model exhibits an *insensitivity* property, meaning that the blocking probability, $B(s, a)$, is independent of the service time distribution. By using a reciprocal analysis, one can easily show that there exists a simple relationship between the Erlang B and the

Erlang C formulae [2]

$$C(s, a) = \frac{sB(s, a)}{s - a[1 - B(s, a)]}, \quad 0 < a < s. \quad (19)$$

STRUCTURAL PROPERTIES AND COMPUTATIONAL ISSUES

Several structural properties of the function $C(s, a)$ have been extensively studied in the literature. The monotonicity of $C(s, a)$ in both a and s can be easily derived, where $C(s, a)$ is increasing in a , for a fixed s and decreasing in s , for a fixed a . Lee and Cohen [6] discussed the convexity of $C(s, a)$ as a function of the offered load, a . Additionally, a proof of the convexity of $C(s, a)$ in s is provided by Jagers and van Doorn [7].

For large a and s , it can be difficult to compute the delay probability because of the presence of factorials and potentially very large powers. However, the delay probability can be much more easily computed by using a recursive scheme. Specifically, using reciprocals, it is not difficult to show that $C(s, a)$ can be equivalently expressed by

$$C(s, a) = \frac{1}{1 + \frac{s-a}{a} \frac{s-1-aC(s-1, a)}{(s-1-a)C(s-1, a)}}, \quad s > a + 1. \quad (20)$$

Note that the Erlang C formula is defined for $a < s$, and, if we start the recursion from $C(1, a)$, we are restricted to $a < 1$, which makes the recursion less useful than the recursion for the Erlang B formula [2], which is

$$B(k, a) = \frac{aB(k-1, a)}{k + aB(k-1, a)}, \quad k = 1, 2, \dots, s, \quad (21)$$

where $B(0, a) \equiv 1$. However, we can use Equation (21) to obtain $B(s, a)$ and substitute in Equation (19) to obtain the delay probability. It is very easy to implement these recursions in a spreadsheet or a computer program to obtain the delay probability for a heavily loaded system with a very large

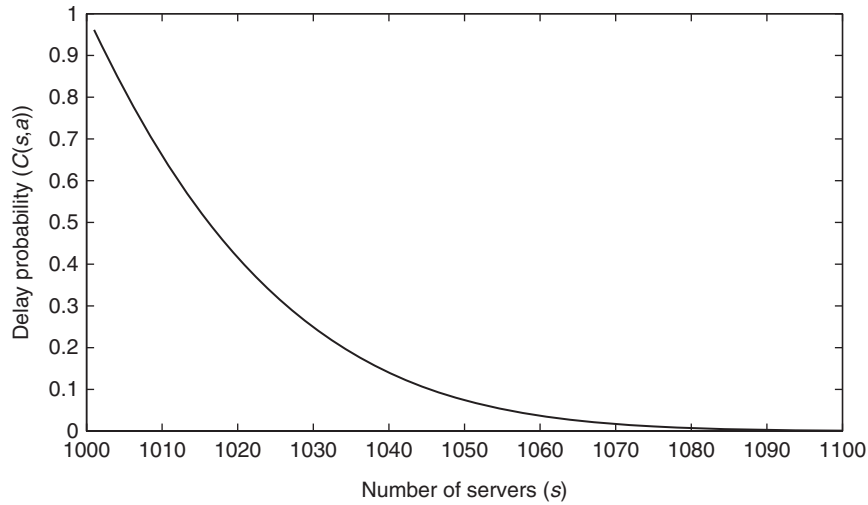


Figure 1. The probability of delay as a function of s when a is fixed.

number of servers. For example, Fig. 1 graphs the delay probability $C(s, a)$ as a function of the number of servers s when the offered load is fixed at $a = 1000$.

FURTHER READING

Some further details of the $M/M/s$ model can be found in many classical queueing textbooks including Cooper [2], Gross and Harris [3], Ross [8], and Wolff [5]. Important properties of the loss and delay formulae have been studied by many researchers. Lee and Cohen [6] discussed the convexity of $C(s, a)$ as a function of the offered load, a . Grassman [9] gives the proof of the convexity for the related performance parameters, the mean number of customers in the system L , and the mean number of customers in queue L_q . Jagers and van Doorn [7] proved the convexity of functions that generalize the Erlang loss and delay functions. Harel and Zipkin [10] proved strong convexity results for the average sojourn time in an $M/M/s$ system. Dyer and Proll [11] provided marginal analysis for allocating servers in $M/M/s$ queues. Harel [12] provided sharp bounds and simple approximations for the loss and delay formulae. Recently, Harel [13] improved the bounds [12], and obtained new bounds and simple approximations for the number of servers

that invert the Erlang delay and loss formulae. Additionally, it was shown that the steady-state idle probability is a convex and decreasing function of traffic intensity.

The extension of the model to general service times can be handled by approximation schemes that exist in the queueing literature. Kingman [14] provided heavy-traffic approximations for the $M/G/s$ queue. When s is very large, another asymptotic regime emerges, namely the quality and efficiency driven (QED) regime. Halfin and Whitt [15] were the first to analyze the $M/M/s$ queue in this regime, and they derived an asymptotic expression for the Erlang C formula. An important extension of the model, known as the *Erlang A Model*, allows for customer abandonment (i.e., customers who join the queue but may opt to leave the queue after some fixed or random time). Borst *et al.* [16], Halfin and Whitt [15] and Garnett *et al.* [17] each analyze the Erlang A model under different asymptotic regimes. Recently, Pang and Whitt [18] provided heavy-traffic extreme value limits for the Erlang delay model with abandonment. Gans *et al.* [19] provided an extensive review of call centers, which includes the Erlang delay and loss models. Additionally, Stidham [20] provided an extensive survey covering the analysis,

design, and control of a variety of queueing systems, including the $M/M/s$ queue.

REFERENCES

1. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Ann Math Stat* 1953;24:338–354.
2. Cooper RB. *Introduction to Queueing Theory*. 2nd ed. New York: North-Holland Publishing Co.; 1981.
3. Gross D, Harris C. *Fundamentals of Queueing Theory*. New York: John Wiley & Sons, Inc.; 1998.
4. Wolff RW. Poisson arrivals see time averages. *Oper Res* 1982;30:223–231.
5. Wolff RW. *Stochastic modeling and the theory of queues*. New York: John Wiley & Sons, Inc.; 1989.
6. Lee HL, Cohen MA. A note on the convexity of performance measures of $M/M/c$ queueing systems. *J Appl Probab* 1983;20:916–919.
7. Jagers AA, van Doorn EA. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Rev* 1991;33:281–282.
8. Ross SM. *Stochastic processes*. New York: John Wiley & Sons, Inc.; 1996.
9. Grassman W. The convexity of the mean queue size of the $M/M/c$ queue with respect to the arrival rate. *J Appl Probab* 1983;20:920–923.
10. Harel A, Zipkin PH. Strong convexity results for queueing systems. *Oper Res* 1987;35:405–418.
11. Dyer ML, Proll LG. On the validity of marginal analysis for allocating servers in $M/M/c$ queues. *Manag Sci* 1977;23:1019–1022.
12. Harel A. Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Manag Sci* 1988;34:959–972.
13. Harel A. Sharp and simple bounds for the Erlang delay and loss formulae. *Queueing Syst* 2010;64:119–143.
14. Kingman JFC. On queues in heavy traffic. *J R Stat Soc Ser B* 1962;24:383–392.
15. Halfin S, Whitt W. Heavy-traffic limits for queues with many exponential servers. *Oper Res* 1981;29:567–587.
16. Borst SC, Mandelbaum A, Reiman MI. Dimensioning large call centers. *Oper Res* 2000;52:17–34.
17. Garnett O, Reiman M. Designing a call center with impatient customers. *Manuf Serv Oper Manag* 2002;4:208–227.
18. Pang G, Whitt W. Heavy-traffic extreme value limits for Erlang delay models. *Queueing Syst* 2009;63:13–32.
19. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. *Manuf Serv Oper Manag* 2003;5:79–141.
20. Stidham S. Analysis, design, and control of queueing systems. *Oper Res* 2002;50:197–216.

THE MANUFACTURING AND SERVICE OPERATIONS MANAGEMENT (MSOM) SOCIETY

TAVA OLSEN
Department of Information Systems
and Operations Management,
University of Auckland Business
School, Auckland, New Zealand

OVERVIEW

The Manufacturing and Service Operations Management (MSOM) Society is one of the largest societies within the Institute for Operations Research and the Management Sciences (INFORMS) and its purpose statement reads as follows:

The *MSOM Society* promotes the enhancement and dissemination of knowledge, and the efficiency of industrial practice, related to the operations function in manufacturing and service enterprises. The methods which *MSOM* members apply in order to help the operations function add value to products and services, which are derived from a wide range of scientific fields, including operations research and management science, mathematics, economics, statistics, information systems, and artificial intelligence. The members of *MSOM* include researchers, educators, consultants, practitioners, and students, with backgrounds in these and other applied sciences.

Thus MSOM is the natural home for operations management (OM) research from a range of disciplines and approximately 950 people, particularly students and faculty from OM departments and groups, as well as from industrial engineering and operations research departments and groups, are the members. Although the society welcomes practitioners, they have traditionally been a small portion of the membership. The membership is primarily US-based, although there are a significant number of European and Asian members.

This article outlines the history and governance of the society, gives information on the society prizes and publications, and meeting information. Further details on the society may be found at <http://msom.society.informs.org/>.

HISTORY AND GOVERNANCE

MSOM began as a technical section of INFORMS following the 1995 merger of the Operations Research Society of America (ORSA) and The Institute of Management Sciences (TIMS). It was formed from a merger of the ORSA, Technical Section on Manufacturing Management (TECMAN), and the TIMS College on Production and Operations Management (COPOM). The outgoing chair of TECMAN was Candace Yano, although John Buzacott, TECMAN chair 1993–1994, was also highly involved in the formation of MSOM. COPOM appears to have been largely inactive at that time. Michael Magazine was the first official president of MSOM and served from 1995 to 1996. A complete listing of prior presidents may be found in the appendix. In 1997, under the presidency of Mark Spearman, MSOM transitioned to a society within INFORMS, with the transition becoming formal in 1998 and is now officially titled, the “Manufacturing and Service Operations Management Society: An INFORMS Organization.”

MSOM was founded with a president, president-elect, vice president (Meetings), and secretary/treasurer; all of these offices continue today. All offices involve a one-year term, voted on in the spring, traditionally with two candidates for each post although this is not required by the bylaws (http://msom.society.informs.org/pdf/MSOM_bylaws_Updated_may_9_08.pdf). The new officers take over at the end of the business meeting at the MSOM summer conference.

At the Society's founding, there was a newsletter and Frank Ciarallo served as the long-standing, and apparently only,

newsletter editor. He also served as webmaster and mailing list moderator for many years, both of which positions continue to this day, separately. Other volunteer position for MSOM include the INFORMS cluster organizer(s), the MSOM student paper competition coordinator, and the Society's representative on the Subdivisions Council.

One of the biggest changes to the Society's organization came in 2006 with the introduction of Special Interest Groups (SIGs) under the leadership of Gerard Cachon and Martin Lariviere. The initial SIGs were supply chain management and service management groups. These were quickly followed upon petition by iFORM, interface between Finance, Operations, and Risk Management. In 2009, the MSOM officers approved a fourth SIG: health care OM. The SIG chairs are officially officers in the Society. Finally, in 2009, MSOM initiated an official affiliation with the behavioral section of INFORMS, when they indicated that they were not interested in becoming SIG.

PRIZES AND PUBLICATIONS

A student paper competition was introduced at the time of the Society's founding and is now traditionally run during the summer. A distinguished service award was introduced in 1996. In 1999, the Society approved the appointment of Fellows, after the work of a committee chaired by Nicholas Hall. The first Fellows were appointed in 2000 and a complete list may be found in the appendix. From 2009, Fellows gave a presentation at the MSOM conference; previously, they had given a presentation at the INFORMS annual meeting. Details for applying for these prizes and a list of past recipients may be found at <http://www3.informs.org/index.php?c=35&kat=-+INFORMS+Subdivision+Prizes>

In 1997, Leroy Schwarz became the founding editor of the MSOM Journal. John Buzacott was awarded the 1998 MSOM distinguished service award in part because he was largely responsible for getting ORSA/TIMS/INFORMS to launch the new journal. The inaugural volume was published

in 1999 and from the beginning, has focused on high quality papers and short review cycles. In 2003, Garrett Van Ryzin took over as editor-in-chief, followed by Gerard Cachon in 2006–2008. Stephen Graves is the current editor-in-chief.

In 2007, Gerard Cachon instituted the MSOM Best Paper Award, an award for the best published work in the MSOM journal that has appeared in the past 3 years.

MEETINGS

The first unofficial MSOM meeting was actually a TECMAN meeting and was held in 1994 at Carnegie Mellon; it was coorganized by John Buzacott and Candi Yano. The first official meeting was held in 1996 at Dartmouth College and was organized by Medini Singh. These meetings had published proceedings, a practice continued until the 2000 meeting at the University of Michigan, where the proceeding were made electronic and then made optional at the 2002 meeting at Cornell. Proceedings were then brought back in 2006 by Georgia Tech and made to be refereed as a Method of Screening Articles. The conference has always been held at a university campus and a complete list of locations with dates may be found in the appendix.

The MSOM meeting started as a biannual meeting and then moved to an annual meeting when INFORMS turned its spring meeting from a general meeting to a practitioners' meeting. It is typically held in late June and the society bylaws state "The Conference shall be held no earlier than the second week of June and no later than the third week of July. Deviation from this timing requires the approval of a majority of Society officers." Until 2008, it was preceded by the one-day Multiechelon Inventory Conference, an annual conference unaffiliated with any particular society. When the MSOM conference was biannual, the Multiechelon Inventory Conference was held by itself in the off years. In 2008, the decision was made to replace the Multiechelon Inventory Conference with one-day parallel SIG conferences, to typically precede the MSOM conference.

Refereed proceedings as a part of the conference have been somewhat controversial and the subject of a number of discussions at MSOM business meetings. However, in this author's opinion, the general consensus appears to be (i) members like the meetings to be held at universities; (ii) members do not like too many parallel tracks (which are also often impractical given university facilities); and (iii) members like the meeting length to be the traditional 2 days. These constraints, along with the growing size of the conference (from 80–90 attendees in 1994 to around 340 in 2009), have meant that screening of papers for acceptance to the conference has become a somewhat accepted practice, although the precise format is still under discussion and was the subject of a 2008 sub-committee whose recommendations will be implemented in 2010.

A call for hosting the conference is typically put out by the VP (Meetings) around November/December for the conference that is 2.5 years away so that conferences are determined 2 years in advance. There is also an informal rotation that the MSOM conference is held internationally once in every 3 years, although this is not formally stated anywhere. There has also been a rotation for these international conferences between east and west of the United States, although there has actually been only one international eastern conference, which was in Beijing, China, in 2007.

The Society also sponsors tracks at the fall INFORMS meeting and at other conference (e.g., CORS-INFORMS, Euro-INFORMS, International Meeting of INFORMS, IFORS etc) on an *ad hoc* basis. At the fall INFORMS meeting, each SIG hosts a track in addition to the general track. Business meetings are held both at the fall INFORMS conference and at the summer conference. Minutes are taken and posted to the MSOM website, after the meeting (<http://msom.society.informs.org/minutes/>).

SUMMARY

As one of the largest societies within INFORMS, MSOM is a healthy and thriving society. It offers a range of prizes, has a top

quality journal, and is financially healthy. In this author's opinion, its main challenge is to remain the primary home for OM research as new OM topics grow and desire communities. The addition of SIGs has helped alleviate this problem to some extent and yet, it has also introduced some new questions. For example, what is the precise relationship between the Health Care OM SIG and the Health Care Applications section within INFORMS? This is an open question yet to be developed further. Further, MSOM currently has no formal relationship with the POM Society (<http://www.poms.org/>) and yet both are growing societies competing for similar membership bases; at some point, it may make sense for the two societies to either form a closer relationship or more clearly differentiate themselves. These will be topics that are likely to occupy the minds of future MSOM presidents.

APPENDIX

Former Presidents

John Buzacott outgoing chair of TECMAN
93–94 and Candace Yano 94–95

1995	Michael Magazine
1996	Joseph Mazzola
1997	Mark Spearman
1998	John McClain
1999	Nicholas Hall
2000	Ronald Askin
2001	Sridhar Tayur
2002	Ananth Iyer
2003	Panos Kouvelis
2004	Izak Duenyas
2005	Fangruo Chen
2006	Gerard Cachon
2007	Martin Lariviere
2008	Tava Olsen
2009	Jeanette Song
2010	Noah Gans

MSOM Fellows

2000	John Buzacott, Steve Graves, and Ed Silver
2001	Ken Baker, Hau Lee, and Paul Zipkin
2002	Marshall Fisher and Evan Porteus
2003	Lawrence Wein
2004	Morris Cohen and Awi Federgruen
2005	Wallace Hopp, Warren Hausman, and Bill Maxwell
2006	Linus Schrage and Harvey Wagner
2007	Garrett van Ryzin
2008	Sven Axsäter, J. Michael Harrison and John (Jack) Muckstadt

MSOM Conferences

1994	Carnegie Mellon (only TECMAN conference), June 27–28
1996	Dartmouth, June 24–25

1998	University of Washington, Seattle, June 29–30
2000	University of Michigan, Ann Arbor, June 26–27
2002	Cornell, June 17–18
2003	USC, Los Angeles, June 16–17
2004	Eindhoven, Netherlands, July 1–2
2005	Northwestern University, Evanston, Illinois, June 27–28
2006	Georgia Institute of Technology, June 19–20
2007	Tsinghua University, Beijing, June 18–19
2008	University of Maryland, June 5–6
2009	MIT Sloan School, June 29–30
2010	Technion—Israel Institute of Technology, Haifa, Israel, June 28–29
2011	University of Michigan, Ann Arbor, June 27–28

THE $M/G/\infty$ QUEUE

NATARAJAN GAUTAM

Department of Industrial and Systems
Engineering, Texas A&M University,
College Station, Texas

An $M/G/\infty$ queue is one where the arrivals are according to a Poisson Process, and the service time for each arriving entity is generally distributed with cumulative distribution function (CDF) $G(\cdot)$. Further, there are an infinite number of servers in the system, and therefore the sojourn time in the system equals the service time. Let λ be the average arrival rate. Also, let τ denote the mean service time. We will typically consider the mean service time to be finite, that is, $\tau < \infty$. Also, for nontriviality we assume $\tau > 0$. It is crucial to note that there is no restriction in terms of the service time random variable; it could be discrete, continuous, or a mixture of discrete and continuous.

The first time anyone hears about the $M/G/\infty$ queue, the biggest concern is the issue of infinite number of servers. It does not seem practical and if it is a purely theoretical exercise, why bother? Several reasons come to mind. There are systems where infinite servers are excellent approximations, such as conveyor belts where for all practical purposes the entities on the conveyor belt can begin service as soon as they arrive. Similarly systems with *self-service* can also be modeled as $M/G/\infty$ queues. A more common approximation is when the number of servers in an $M/G/c$ queue is large and the traffic intensity is not too high (i.e., not almost 1). Typically, the $M/G/c$ queue for $c > 1$ is difficult to analyze. Hence, the $M/G/\infty$ queue is also known as a *queue with Poisson arrivals and ample servers*.

Other motivations for studying $M/G/\infty$ queues are to obtain insights into the model and analysis that can be used in related systems. For example, to explain some of the results of $M/G/c/c$ queues, one typically turns to $M/G/\infty$ results. Moreover, among

all the systems where the arrival or service (or both) is nonexponential, $M/G/\infty$ is the only one that can be analyzed for transient behavior, departures from queues, and time-varying parameters. These reasons are usually sufficient to make a case to teach $M/G/\infty$ in courses and also include in textbooks (and encyclopedias).

Recently, there have been some renewed interest and new results for $M/G/\infty$ while being applied to modeling traffic in computer-communication systems. Parulekar and Makowski [1] considered a slotted-time system, where the arrivals in a given slot were proportional to the number in the system of an $M/G/\infty$ queue with heavy tail service time (such as a Pareto distribution). Notice that since the number in the system for an $M/G/\infty$ is correlated over time, using a Pareto distribution results in long-range dependence, which is shown to be satisfied in traffic in computer-communication systems. Chang [2] characterizes such an arrival process using large-deviations theory to obtain the tail distribution of subsequent buffers.

Further, Likhanov and Mazumdar [3] as well as Resnick and Samorodnitsky [4] extend the slotted (or discrete time) version developed by Parulekar and Makowski [1] to the continuous time case. In particular, requests arrive according to a Poisson process and every request is responded by a heavy-tailed response, where fluid is poured at a constant rate. These are excellent models for web servers where information flows out (response) to requests that arrive according to a Poisson process and file sizes are heavy tailed. These are referred to as *fluid- $M/G/\infty$ queues*. In summary, the $M/G/\infty$ queue can be used to suitably approximate a number of applications, and there is a need to understand them in order to obtain useful insights.

Having motivated the need to study $M/G/\infty$ queues, here is how the remainder of this article is organized. In the section titled "Number in System: Transient and

Steady State,” we describe transient and steady-state analysis of the number in the system. We use those results to derive the busy period distribution in the section titled “Busy Period Distribution.” Then in the section titled “Departure Process: Transient and Steady State,” we study the departures from $M/G/\infty$ queues. We consider an extension to the standard $M/G/\infty$ system by allowing time-varying arrivals in the section titled “Time-varying Arrivals: $M_t/G/\infty$.”

NUMBER IN SYSTEM: TRANSIENT AND STEADY STATE

Consider an $M/G/\infty$ queue having independent and identically distributed service times with CDF $G(\cdot)$. Let $\{N(t), t \geq 0\}$ be a Poisson process with parameter λ that counts the number of arrivals in time $(0, t]$ for all $t \geq 0$. Define $X(t)$ as the number (of entities) in the system at time t . The objective of transient and steady-state analysis is to obtain a probability distribution of $X(t)$ for finite t and as $t \rightarrow \infty$, respectively. We would first present the transient analysis and then take the limit for steady-state analysis.

Transient analysis typically depends on the initial state of the queue. To obtain simple closed-form expressions, we need to make one of the following three assumptions: (i) the queue is empty at $t = 0$, that is, $X(0) = 0$; (ii) the queue started empty in the distant past, that is, $X(-\infty) = 0$, and hence the stochastic process $\{X(t), t \geq 0\}$ is stationary; (iii) if $X(0) > 0$, then the service for all the $X(0)$ entities *begins* at $t = 0$. If one of the above three assumptions is not satisfied, we will have to know the times when service began for each of the $X(0)$ customers in order to obtain a distribution for $X(t)$.

For the transient analysis here, we make the first assumption, that is, $X(0) = 0$. Hence, we are interested in computing $p_j(t)$ defined as

$$p_j(t) = P\{X(t) = j | X(0) = 0\}.$$

Note that if we made the second assumption, then $X(t)$ would be according to the steady-state distribution for all $t \geq 0$. However, if we

made the third assumption, then

$$\begin{aligned} P\{X(t) = j | X(0) = i, B_i\} \\ = \sum_{k=0}^{\min(i,j)} \binom{i}{k} [1 - G(t)]^k [G(t)]^{i-k} p_{j-k}(t), \end{aligned}$$

with the event B_i denoting that service for all i initial customers begins at $t = 0$. Since this is straightforward, once $p_j(t)$ is known, we continue with obtaining $p_j(t)$ by making the first assumption.

Notice that $\{X(t), t \geq 0\}$ is a regenerative process with regeneration epochs corresponding to when the queueing system becomes empty. To obtain $p_j(t)$, consider an arbitrary arrival at time x such that $x \leq t$. Let q_x be the probability that this arriving entity is in the system at time t . Clearly,

$$q_x = 1 - G(t - x)$$

since it is the probability that this entity's service time is larger than $t - x$. Now consider a nonhomogeneous Bernoulli splitting of the Poisson (arrival) process $\{N(t), t \geq 0\}$ such that with probability q_x an entity arriving at time x will be included in the split process. Since the split process counts the number in the $M/G/\infty$ system at time t , we have

$$p_j(t) = \exp \left\{ -\lambda \int_0^t q_x dx \right\} \frac{\left\{ \lambda \int_0^t q_x dx \right\}^j}{j!}.$$

The proof is described in Kulkarni [5], Gross and Harris [6], and Wolff [7]. It is based on conditioning the number of arrivals in time t and using the fact that each of the given arrivals occurs uniformly in $[0, t]$. The argument is similar to the derivation of the number of departures from the $M/G/\infty$ queue in time t in the section titled “Departure Process: Transient and Steady State.”

Next, we write down $p_j(t)$ in terms of entities given in the model, namely, λ and $G(t)$. As a result of using $q_x = 1 - G(t - x)$ and change

of variables, we get

$$\begin{aligned}
 p_j(t) &= \exp \left\{ -\lambda \int_0^t q_x \, dx \right\} \frac{\left\{ \lambda \int_0^t q_x \, dx \right\}^j}{j!} \\
 &= \exp \left\{ -\lambda \int_0^t [1 - G(t-x)] \, dx \right\} \\
 &\quad \times \frac{\left\{ \lambda \int_0^t [1 - G(t-x)] \, dx \right\}^j}{j!} \\
 &= \exp \left\{ -\lambda \int_0^t [1 - G(u)] \, du \right\} \\
 &\quad \times \frac{\left\{ \lambda \int_0^t [1 - G(u)] \, du \right\}^j}{j!}.
 \end{aligned}$$

Using $p_j(t)$ we can also obtain

$$E[X(t)] = \lambda \int_0^t [1 - G(u)] \, du$$

and

$$\text{Var}[X(t)] = \lambda \int_0^t [1 - G(u)] \, du$$

by observing that for a given t , $X(t)$ is a Poisson random variable with parameter $\lambda \int_0^t [1 - G(u)] \, du$.

For the steady-state analysis, we let $t \rightarrow \infty$ in the transient analysis to derive all results. Let $X(t)$ converge in distribution to X as $t \rightarrow \infty$ such that $p_j(t) \rightarrow p_j$ where $p_j = \lim_{t \rightarrow \infty} P\{X(t) = j\}$. Taking the limit $t \rightarrow \infty$ for $p_j(t)$, we find

$$p_j = e^{-\lambda\tau} \frac{(\lambda\tau)^j}{j!},$$

where τ is the average service time. Then, with an integration-by-parts step it can be shown that

$$\tau = \int_0^\infty [1 - G(u)] \, du.$$

In addition, the mean and variance of the number in the system in steady state are both $\lambda\tau$ since X is a Poisson random variable with parameter $\lambda\tau$. Note that the mean

and variance of the sojourn times correspond, respectively, to the mean and variance of the service times since there is no waiting for service to begin. Before concluding this section, it is worthwhile observing that the steady-state probability p_j is identical to those of the $M/M/\infty$ system, and thus p_j does not depend on the CDF $G(\cdot)$ but just the mean service time.

BUSY PERIOD DISTRIBUTION

As we described earlier, $\{X(t), t \geq 0\}$ is a regenerative process with regeneration epochs corresponding to times when the number in the system goes from 1 to 0. Each regeneration time corresponds to one idle period followed by one *busy period* (time when there are one or more entities in the $M/G/\infty$ system). Let U be the regeneration time and $U = I + B$, where the idle time $I \sim \exp(\lambda)$ (i.e., time for next arrival in a Poisson process), and the busy period B has a CDF $H(\cdot)$. We need to determine $H(\cdot)$. For that we develop the following explanation based on Example 8.17 in Kulkarni [5].

Let $F(\cdot)$ be the CDF of U . Using a renewal argument by conditioning on $U = u$, we can write down a renewal-type equation for $p_0(t) = P\{X(t) = 0 | X(0) = 0\}$ as

$$\begin{aligned}
 p_0(t) &= \int_0^t p_0(t-u) \, dF(u) \\
 &\quad + \int_t^\infty P\{I > t | U = u\} \, dF(u).
 \end{aligned}$$

Since $U = I + B$, $P\{I > t | U = u\} = 0$ if $u \leq t$, we can rewrite

$$\begin{aligned}
 p_0(t) &= \int_0^t p_0(t-u) \, dF(u) \\
 &\quad + \int_0^\infty P\{I > t | U = u\} \, dF(u) \\
 &= \int_0^t p_0(t-u) \, dF(u) + P\{I > t\} \\
 &= p_0 * F(t) + e^{-\lambda t}
 \end{aligned}$$

with $p_0 * F(t)$ being the standard convolution notation.

We already have an expression for $p_0(t)$ in the section titled “Number in System: Transient and Steady State” as

$$p_0(t) = \exp \left\{ -\lambda \int_0^t [1 - G(u)] du \right\},$$

where $G(\cdot)$ is the CDF of the service times. One way to obtain the unknown $F(t)$ function is to numerically solve for $F(t)$ in $p_0(t) = p_0 * F(t) + e^{-\lambda t}$ and similarly solve another convolution equation to get $H(t)$. An alternate approach, typically standard when there are convolutions, is to use transforms. We consider the Laplace Stieltjes transform (LST) which is denoted as $\tilde{R}(s)$ for any positive-valued nondecreasing function $R(t)$ and is defined as

$$\tilde{R}(s) = \int_0^\infty e^{-st} dR(t).$$

Therefore, taking the LST on both sides of the equation $p_0(t) = p_0 * F(t) + e^{-\lambda t}$, we get

$$\tilde{p}_0(s) = \tilde{p}_0(s)\tilde{F}(s) + \frac{s}{s + \lambda}.$$

However, since we are ultimately interested in $H(t)$, we use the relation

$$\tilde{F}(s) = \left(\frac{\lambda}{\lambda + s} \right) \tilde{H}(s)$$

to obtain

$$\tilde{H}(s) = 1 + \frac{s}{\lambda} - \frac{s}{\lambda \tilde{p}_0(s)}.$$

In the most general case, the above LST is not easy to invert and obtain $H(t)$. Another challenge is to obtain $\tilde{p}_0(s)$ from $p_0(t)$, which is not trivial. However, there are several software packages available (such as Matlab and Mathematica) that can be used to numerically compute as well as invert LSTs.

Having said that, it is relatively straightforward to obtain $E[B]$, the mean busy period. Using the fact that $\tilde{p}_0(0) = p_0(\infty) = e^{-\lambda \tau}$, we can obtain

$$E[B] = -\tilde{H}'(0) = \frac{e^{\lambda \tau} - 1}{\lambda}.$$

Another way to obtain it is to use regenerative process results and solve for $E[B]$ in $1 - p_0 = \frac{E[B]}{E[B] + 1/\lambda}$.

DEPARTURE PROCESS: TRANSIENT AND STEADY STATE

Since the output from an $M/G/\infty$ may flow into some other queue, it is critical to analyze the departure process. Let $D(t)$ be the number of departures from the $M/G/\infty$ system in time $[0, t]$ given that $X(0) = 0$. For any arbitrary t , we seek to obtain the distribution of the random variable $D(t)$ and thereby characterize the stochastic process $\{D(t), t \geq 0\}$. Similar to the analysis for the number in the system, here too we first consider transient and then describe steady-state results. The results follow the analysis in Gross and Harris [6]. However, it is crucial to point out that there are many other elegant ways of analyzing departures from $M/G/\infty$ queues and extending them, some of which we will see toward the end of this section.

If we are given that n arrivals occurred in time $[0, t]$, then using the standard Poisson process results we know that the time of arrival of any of the n arrivals is uniformly distributed over $[0, t]$ and it is independent of the time of other arrival times. Therefore, consider one of the n arrivals that occurred in time $[0, t]$. The probability $\theta(t)$ that this entity would have departed before time t can be obtained by conditioning on the time of arrival and unconditioning as

$$\theta(t) = \frac{1}{t} \int_0^t G(t - x) dx = \frac{1}{t} \int_0^t G(u) du.$$

In addition, the probability that out of the n arrivals in time $[0, t]$, exactly i of those departed before time t is $\binom{n}{i} [\theta(t)]^i [1 - \theta(t)]^{n-i}$.

Now, to compute the distribution of $D(t)$, we condition on $N(t)$, which is the number of arrivals in time $[0, t]$. In order to remind us that we do make the assumption that $X(0) = 0$, we include this condition in the expressions. Therefore, we have

$$\begin{aligned} P\{D(t) = i | X(0) = 0\} &= \sum_{n=i}^{\infty} P\{D(t) = i | N(t) = n, X(0) = 0\} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\ &= \sum_{n=i}^{\infty} \binom{n}{i} [\theta(t)]^i [1 - \theta(t)]^{n-i} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \end{aligned}$$

$$\begin{aligned}
&= \frac{[\theta(t)]^i}{i!} e^{-\lambda t} (\lambda t)^i \sum_{n=i}^{\infty} [1 - \theta(t)]^{n-i} \frac{(\lambda t)^{n-i}}{(n-i)!} \\
&= \frac{[\lambda t \theta(t)]^i}{i!} e^{-\lambda t \theta(t)}.
\end{aligned}$$

Therefore, for a given t , $D(t)$ is a Poisson random variable with parameter $\lambda t \theta(t)$. In other words, the transient distribution of the number of departures in time t is Poisson with parameter $\lambda t \theta(t)$. Having described the transient analysis, we just let $t \rightarrow \infty$ for steady-state analysis.

It can be shown from the definition of $\theta(t)$ that for very large t , $t\theta(t)$ approaches $t - \tau$. However, since we assume that τ is finite, as we let $t \rightarrow \infty$, the departure becomes a Poisson process with parameter $\lambda(1 - \tau/t) \rightarrow \lambda$. Of course, we have not shown stationary and independence properties of the departure process; the reader is encouraged to refer to Serfozo [8, Section 9.6] for a derivation based on stationary marked point processes. Using the above results, as well as stationary and independence properties, we can conclude that the departure process from an $M/G/\infty$ queue in steady state is the same as the arrival process, that is, a Poisson process with parameter λ . Further, similar to the steady-state probability p_j being identical to that of the $M/M/\infty$ system and being independent of the CDF $G(\cdot)$, the departure process in steady state is also identical to that of the $M/M/\infty$ system and depends only on the mean service time (τ) and not on its distribution.

We conclude this section by emphasizing that there is substantial amount of research on departure process from an $M/G/\infty$ queue that has not been covered here. Daley [9] shows that the departure process when viewed as a random translation of the arrival process is Poisson. That argument even works in the nonhomogeneous case (i.e., the $M_t/G/\infty$ described in the next section). Also, a natural extension of output processes is to consider a network of $M/G/\infty$ queues. Analyzing a network of feedforward $M/G/\infty$ queues is particularly straightforward leveraging upon the fact that the output of an $M/G/\infty$ queue is also Poisson. For non-feedforward networks (i.e., networks

with cycles) some of the properties are more tricky to show, and the reader is referred to Serfozo [8]. Furthermore, it can be shown that (see Kobayashi and Mark [10]) the $M/G/\infty$ queue is quasi-reversible and hence useful in analyzing networks (e.g., to obtain product-form solutions).

TIME-VARYING ARRIVALS: $M_t/G/\infty$

In this section, we consider an extension to the $M/G/\infty$ queue. The arrival process is Poisson; however, the parameter of the Poisson process is time varying. The average arrival rate at time t (for all t in $(-\infty, \infty)$) is a deterministic function of t represented as $\lambda(t)$. Hence, the arrival process is defined as a nonhomogeneous Poisson process. Everything else is the same as the regular $M/G/\infty$ queue. We call such a system an $M_t/G/\infty$ queue. This summary of results for the $M_t/G/\infty$ queue is based on Eick *et al.* [11].

Recall that we need to make one of the three assumptions stated in the section titled “Number in System: Transient and Steady State,” otherwise we would need to know when service started for each of the customers at a reference time, say $t = 0$. We make the second assumption, although making one of the other two would not significantly alter the analysis but the results would not be as neat. In other words, assume that the system was empty at $t = -\infty$. Assume that for all finite t , $\lambda(t)$ is nonnegative, measurable, and integrable. With these assumptions in mind, we state the results.

The results are in terms of the equilibrium random variable of the service times. In particular, let S be a random variable corresponding to a service time. We have already defined the CDF $G(x) = P\{S \leq x\}$ and mean $\tau = E[S]$. The equilibrium random variable S_e corresponding to the service times has a CDF $G_e(x)$ defined as

$$G_e(x) = P\{S_e \leq x\} = \frac{1}{\tau} \int_0^x [1 - G(u)] du.$$

One encounters equilibrium random variables in renewal theory. If a renewal process with inter-renewal random variable Y starts

at $t = -\infty$, then at time $t = 0$, the remaining time for a renewal is according to Y_e .

Recall the notation $X(t)$. Here, it denotes the number (of entities) in the $M_t/G/\infty$ system at time t . Then for any t , $X(t)$ has a Poisson distribution with parameter $\mu(t)$ given by

$$\mu(t) = E[\lambda(t - S_e)]E[S]$$

with $E[X(t)] = \text{Var}[X(t)] = \mu(t)$. It is important to note that $\lambda(\cdot)$ is the function defined in this section and not make the mistake of thinking that at $t = 0$ the average number in the system is negative! In fact, observe that the number in the system at any time depends on the arrival rate S_e time units ago. Further, the departure process also has a similar time lag effect where the average departure rate at time t is $E[\lambda(t - S)]$ and the resulting process is nonhomogeneous Poisson.

For $u > 0$ and any t ,

$$\begin{aligned} \text{Cov}[X(t), X(t + u)] \\ = E[\lambda(t - (S - u)_e^+)]E[(S - u)^+], \end{aligned}$$

where the notation $(y)^+$ is $\max(y, 0)$. In fact, this result can be derived for the homogeneous case (which we have not done earlier, but is extremely useful especially in computer-communication traffic with long-range dependence). For an $M/G/\infty$ queue where $\lambda(t) = \lambda$,

$$\text{Cov}[X(t), X(t + u)] = \lambda P\{S_e > u\}E[S].$$

There are several results for networks of $M_t/G/\infty$ queues. Since the entities do not interact and departure processes are Poisson,

the analysis is fairly convenient. The reader is referred to Eick *et al.* [11] as well as the references therein for further results.

REFERENCES

1. Parulekar M, Makowski AM. Tail probabilities for a multiplexer driven by $M/G/\infty$ input processes (I): preliminary asymptotics. *Queueing Syst Theor Appl* 1997;27:271–296.
2. Chang CS. Performance guarantees in communication networks. London: Springer; 2000.
3. Likhanov N, Mazumdar RR. Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources. In: IEEE INFOCOM; Tel Aviv, Israel; 2000. pp. 173–180.
4. Resnick S, Samorodnitsky G. Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. *Bernoulli* 2001;7:191–210.
5. Kulkarni VG. Modeling and analysis of stochastic systems, Texts in Statistical Science Series. London: Chapman and Hall, Ltd.; 1995.
6. Gross D, Harris CM. Fundamentals of queueing theory. 3rd ed. New York: John Wiley & Sons, Inc.; 1998.
7. Wolff RW. Stochastic modeling and the theory of queues. Englewood Cliffs (NJ): Prentice Hall; 1989.
8. Serfozo R. Introduction to stochastic networks. New York: Springer; 1999.
9. Daley DJ. Queueing output processes. *Adv Appl Probab* 1976;8:395–415.
10. Kobayashi H, Mark BL. System modeling and analysis: foundations of system performance evaluation. Upper Saddle River (NJ): Pearson/Prentice Hall; 2008.
11. Eick SG, Massey WA, Whitt W. The physics of the $M_t/G/\infty$ queue. *Oper Res* 1993;41(4):731–742.

THE NASCENT INDUSTRY OF ELECTRIC VEHICLES

DIEGO KLABJAN

TIMOTHY SWEDA

Department of Industrial Engineering
and Management Sciences,
Northwestern University,
Evanston, Illinois

INTRODUCTION

Although electric vehicle (EV) technology has existed since the mid-1800s, gasoline-powered automobiles with internal combustion have largely overshadowed EVs because of their lower costs, higher speeds, and greater driving ranges. Consequently, public interest in EVs has remained low until recently, when the 2000s' energy crisis and subsequent recession sparked consumer demand for sustainable transportation. This heightened interest is not transient. As automakers race to bring mass-produced consumer EVs to market and policy makers provide incentives for both the producers and end users of these vehicles, the public's attitude toward sustainable transportation will be solidified as EV adoption rates increase.

To encourage adoption of EVs, public and private sector organizations are teaming up to provide charging infrastructure at both the local and national levels. In addition to the United States, other countries such as Canada, Australia, Denmark, Israel, China, and Japan have begun laying the groundwork for new EV fleets [1]. The teams responsible for such infrastructure seek to maximize the convenience of owning and operating an EV, and through collaboration with power utility companies, they can ensure the reliability and energy efficiency of their systems.

The time is ripe to begin planning for the rollout of EVs and their supporting

infrastructure, and research opportunities abound for those versed in analytics and operations research. In this article, the features of EVs which distinguish them from the ubiquitous gasoline-powered vehicles are outlined, and the various types of charging infrastructure are described. A series of open issues to be addressed by operations researchers is also presented to demonstrate the vast potential for new research and to map out future paths of study.

WHAT IS AN EV?

An EV is a vehicle that operates solely using electricity. It has an electric motor and stores energy in an internal battery, which can be recharged by plugging the vehicle into an outlet or charger. Because EVs do not use any gasoline, they are also known as zero-emission vehicles since they do not produce tailpipe emissions. Indirectly, though, they are responsible for a share of the greenhouse gases emitted by the coal-burning generators which produce the electricity for their operation, but the net carbon footprint of an EV is still less than that of a comparable gasoline-powered vehicle [2].

One major concern with EVs is their limited driving range as compared with regular gasoline-powered vehicles. This factor increases the likelihood of running out of charge while on the road, and because recharging can take several hours at standard voltages, the inconvenience to a stranded driver is significant. For this reason, some drivers prefer the concept of a plug in hybrid EV (PHEV). PHEVs retain most of the same traits as pure EVs, but they also utilize an internal combustion engine fueled by gasoline. Depending on the make, a PHEV may either maximize overall energy efficiency by reverting to its all-electric mode for city driving and switching to gasoline power for highway driving, or minimize gasoline usage by utilizing the combustion engine only after the battery is depleted.

Hybrid electric vehicles (HEVs), which have gained popularity over the last several years, differ from PHEVs because they have no external outlet and thus cannot draw power from the electrical grid. Instead, their electric power comes from an internal generator driven by the combustion engine, or is restored via regenerative braking, in which kinetic energy from deceleration is recaptured rather than dissipated as heat. While this characteristic does offer improved fuel efficiency over regular vehicles, the efficiency is still less than that provided by PHEVs and pure EVs [3].

INFRASTRUCTURE REQUIREMENTS

The key feature of EVs is their ability to connect to the electrical grid. In order for them to experience widespread adoption by the public, they must be able to connect to the grid efficiently and effortlessly, and operations research can aid in the development of solutions for accomplishing this goal. The three main types of EV charging infrastructure include at-home charging methods, public charging stations, and battery swapping concepts.

At-Home Charging

Currently, EV charging takes place almost exclusively in private garages, which are expected to continue to be the primary charging location as EVs gain popularity. Therefore, it is important for at-home charging solutions to be as convenient and reliable as possible.

The most common method of charging EVs at home is by conventional (Level I) charging. It involves plugging an EV into a standard 120 V wall socket. Charging the battery completely can take approximately 8–14 hours [4]. This solution suffices for the majority of EV owners since they can plug in their vehicle at the end of the day, let it charge overnight, and wake up the next morning to a fully charged vehicle. The system also works for the power companies because charging takes place during off-peak hours when demand for electric power is low. For drivers hoping to

get the most mileage out of their EVs, however, the long time required for a complete recharge is a serious drawback. Once the battery is fully depleted, the vehicle is effectively out of commission until the following day.

An alternative to Level I charging has been termed *opportunity charging*, or Level II charging, which is designed to fully recharge EV batteries in 4 hours or less [4]. To achieve the faster charging time, the outlet voltage is increased to around 240 V. Level II charging is intended to extend the daily mileage of EVs by allowing a greater number of excursions per day, since drivers can plug in their vehicles during meals or other downtime and make another trip only hours later. Given that very few drivers return home with a fully depleted battery, the actual time necessary to recharge can be much less than 4 hours if the battery still carries a partial charge. This also extends the length of outings occurring later in the day because drivers do not need to worry about limiting their range due to a partially depleted battery.

Several different pricing models for the electric power used to fuel EVs have been considered. Currently, EV owners who charge their vehicles at home pay for electric power in the same way that they do for power used by other appliances. A feasible alternative would be to meter electricity used by EVs separately, either within at-home charging units or on-board the vehicles themselves, and charge a different rate for EV electricity usage. Some studies have examined the effects of imposing a per-mile tax on gasoline-powered vehicles as opposed to a per-gallon tax [5,6], and a similar pricing mechanism could be considered for EVs.

Dynamic pricing schemes may be necessary to attenuate peak loading on the electrical grid. Some research has already been done to study the impact of EV charging on the grid [7–9], and consideration of time of use pricing, critical peak pricing, and nodal pricing will help the utility companies to optimize loading on their systems. Empirical research on managing EV charging behavior will become important as EV adoption rates rise and more data becomes available.

Public Charging

Despite the importance of at-home charging, the ability to recharge EVs while away from home will be critical to their success. PHEVs will still be able to utilize the existing network of gasoline stations in case their batteries become depleted while on the road, but pure EVs will need to be able to access the grid wherever they are. This will allow EV drivers to commute both short and long distances without needing to own a spare, gasoline-powered (or hybrid) vehicle.

So far, only a handful of charging stations in the United States have been installed and made accessible to the public, and their primary purpose is to generate publicity, spread awareness, and encourage the use of EVs. They are centrally located at sites such as community centers and major-chain supermarkets, but the eventual goal is to have one charging station per parking spot rather than one station per site. As the rate of EV ownership increases, multiple drivers will want to plug in and charge their vehicles at the same time. One way to accomplish this is by outfitting parking lots and garages with sufficient charging infrastructure so that drivers can plug into the grid wherever they park.

It is expected that most drivers who plug in away from home will be working, shopping, or participating in some other activity that will last only a few hours. Therefore, Level II charging capability will likely be the standard for public charging stations. For drivers in need of a more expedited recharge, however, Level II charging may be insufficient. Level III charging, which requires less than 10 minutes for a full recharge (approximately the same amount of time necessary to fill a tank with gasoline) [4], is instead seen as a viable solution for when time is of the essence. It remains to be seen whether Level II or Level III charging will be most popular among EV drivers, but it seems certain that a mix of the two will persist since both types have their advantages (Type III is faster, but Type II requires lower voltages and less specialized hardware).

An open area of research for operations researchers is the pricing of charging services. Previous studies on gasoline station pricing provide a starting point [10,11], but

the fact that electricity is a utility (and, thus, is regulated) poses new challenges. Currently, EV drivers must first purchase a membership in order to utilize a company's public charging stations, and then they are required to purchase a subscription plan. Subscribers can pay as they go (i.e., pay a flat fee each time they recharge), purchase a monthly or annual quota, or buy an unlimited plan that allows them to recharge as frequently as they wish. In each case, the number of recharges is measured by the number of times an EV is plugged in and unplugged. An alternative to consider would be metered charging, where rates are determined by the amount of power used rather than the total number of recharges. This option would appeal to drivers who do not deplete their batteries before recharging and require only a partial recharge. Furthermore, studies examining the effects of regulation on charging service providers would aid in determining whether or not these companies should be treated as utilities, as they sell electric power yet do not constitute a power plant.

Demand modeling is necessary to better understand pricing models of EVs and charging stations. A consumer's decision of whether or not to purchase an EV depends heavily on the cost and convenience of recharging. If recharging is expensive or if charging stations are inconveniently located, a highly priced EV will be difficult to sell. In addition, social networks may play a role in vehicle purchasing decisions, as buyers may be more likely to purchase EVs if their friends also own EVs. Knowing how all of these factors influence individuals' attitudes toward EVs and, ultimately, their decision to purchase (or not to purchase) an EV will be instrumental in determining the optimal pricing of EVs and charging stations.

Discrete choice modeling, a popular and widely accepted technique in transportation demand modeling, is a natural fit to model the demand for EVs. It has already been used in several studies examining the demand for EVs [12–14], but further research using more recent data will be required as the public becomes more informed about EVs. Predictive models that capture trends in the

demand for EVs will also be necessary to fully understand the EV market.

Battery Swapping Stations

An alternative to Level III charging stations is battery swapping stations. Rather than recharge the battery, an EV driver can simply swap a depleted battery for a fully charged one. The process is not only slightly faster than recharging, but it is also fully automated. It has already been successfully tested in Japan, and takes less than one minute from beginning to end [15].

Unlike charging stations, which have minimal spatial requirements and can be installed in home garages or parking spaces, battery swapping stations are comparable to gasoline stations and thus need dedicated plots of land. The capital costs of acquiring the land, constructing a facility, and maintaining an inventory of batteries are much greater than the cost of installing a charging station; thus, swapping stations will be less prevalent than charging stations, especially during the early phases of EV adoption. Before battery swapping stations are made available to the public, their primary users will be companies that operate large EV fleets, such as taxi, shuttle, and delivery services.

Battery swapping has a number of issues that are yet to be resolved. In order to ensure that a driver is able to swap batteries, there must be a sufficient number of charged batteries in the inventory. Because of the spatial requirements needed for storing EV batteries, it is neither practical nor cost effective to carry a large surplus. On the other hand, it would be unacceptable for a driver to be unable to swap for a fully charged battery. The number of chargers at the station for recharging depleted batteries would also need to be kept within a reasonable level to satisfy service requirements without incurring excessive cost. In addition since batteries have a limited lifetime (measured in charging cycles), a swapping station would need to manage its inventory by recycling or discarding worn out batteries and replacing them with newer ones. This task is complicated by the fact that the station has little control over the incoming stock of depleted

batteries. Some sort of safeguard would need to be established to protect the station from handling too many incoming batteries that have reached the end of their lifetime.

If a swapping station decides to recycle its used batteries, it faces the decision of when to withdraw those batteries from its inventory. A used battery has salvage value because it can still store electric charge, enough for some applications beyond EVs, although this value decreases with the age of the battery. Similarly, the swapping station may choose to purchase used batteries to replace the stock it recycles, especially since these batteries will eventually be swapped out with incoming ones. Determining the appropriate purchasing price of a used battery is yet another optimization problem to solve.

Since a proof-of-concept of battery swapping for EVs was only recently demonstrated, the literature is barren with regard to optimal swapping station management. A team from Japan has used queueing systems to analyze the operation of a single swapping station [16], but their model relies on simplistic assumptions (such as Poisson arrival and processing times) and does not take into account the presence of other battery swapping stations. Further research utilizing more sophisticated modeling techniques with more realistic assumptions is warranted.

Smart Grid Applications

The future of the EV infrastructure will be greatly enhanced by the development of the smart grid for delivering electric power to consumers. Presently, most power distribution systems provide electric power to customers without any control to curtail wasteful usage or to reallocate unnecessary peak-hour consumption to off-peak hours. (A few systems and utilities do employ real-time controls and time of use pricing [17–19], but these efforts are still in their infancy.) By incorporating such controls into their systems, power utility companies can better match supply and demand, thereby reducing costs, conserving energy, and improving reliability. The ability to optimize distribution in real-time is especially important when utilizing renewable

energy sources, such as wind and solar, which are intermittent.

EV charging would directly benefit from the presence of the smart grid. For drivers who charge their EVs overnight, the smart grid would ensure that the cost of recharging would be minimized by delaying charging until off-peak hours. Collecting data on a driver's driving activity and patterns would permit an even greater degree of charging optimization, especially for Level II charging, which is more likely to occur during peak hours. For example, if a driver's daily commuting distance is 30 miles and the EV's battery has enough charge left for 20 miles, the smart grid might decide to only recharge a portion of the battery—enough for the driver to return home, plus a buffer—and then complete the remainder of the recharge later during off-peak hours. There could always be manual overrides, of course, if the driver plans to deviate from his or her normal activities, but intelligent recharging would benefit both the power company and the customer.

As much as the smart grid could optimize energy distribution, the addition of vehicle-to-grid (V2G) technology would improve the efficiency of the grid even more. With V2G, not only can the grid provide power to EVs and other electrical appliances, but the EVs can also provide power back to the grid. Some papers have considered the potential benefits of V2G [20–22], but further analysis will be required once data becomes available. EV batteries can serve as spare energy resources during peak hours, when demand for electric power is greatest, and later draw power back from the grid during off-peak hours. Normally, during peak hours, the power companies must operate additional generators to satisfy demand, and these generators are much more expensive than the primary baseload generators. Utilizing a network of EV batteries plugged into the grid would provide a greener and more efficient alternative to satisfying customers' power needs.

To encourage EV owners to plug in their vehicles and discharge power back to the grid, monetary incentives could be offered based on the amount of power discharged. The net

difference between the incentive for discharging power during peak hours and recharging during off-peak hours would translate directly into a rebate for the EV driver, and over time, this rebate would help to offset the initial investment into the EV itself. V2G would operate in conjunction with the smart grid to ensure that the amount of power discharged during the day does not inconvenience a driver by prohibiting additional travel, and manual thresholds could be established based on the driver's personal preferences.

In order for V2G to be truly effective, the grid must know to what extent it can draw power from an EV. If it drains an EV's battery completely, and the EV is parked at its owner's workplace, then the owner cannot return home and will not likely discharge power to the grid again. However, if the owner needs to commute 5 miles to return home and the grid leaves the battery with 20 miles of charge, both parties benefit. The same considerations hold for charging EV batteries as well. Consider the previous example, except that the battery is completely discharged and the owner plugs it in to recharge. If the owner only needs enough charge for a 5-mile commute, it would not make sense to fully charge the battery, especially during peak hours when the cost of power is greatest. Instead, the grid should only recharge enough for the commute, plus a buffer (e.g., an additional 10 miles of charge).

OTHER OPEN ISSUES

Because EVs have only recently experienced a resurgence in popularity, there are still a number of open issues that have either been insufficiently addressed or completely overlooked; hence, there are plenty of opportunities for academic research to facilitate the transition to EVs and sustainable transportation. Many of these issues, including those introduced in earlier sections, involve topics which fall in the realms of operations research, management science, and analytics. Advances in EV design and battery technologies have led to several automakers announcing plans to

market EV and PHEV models, but a lack of supporting infrastructure and business models threatens to delay the EV transition.

Much of the existing literature on EVs has dealt with great amounts of uncertainty regarding the technological requirements and capabilities of EV infrastructure. The lack of data on driving behavior with respect to EVs further hampers the ability to draw conclusions and ultimately provide valid policy recommendations. As these items become better understood, though, more quantitative models can be formed and more detailed and rigorous analyses can be conducted.

In the following sections, some additional examples of potential research areas are presented that are of particular interest to industrial engineers with a focus on analytics. The goal is not to offer an exhaustive list, but rather to demonstrate the breadth of topics in need of further study and underscore some of the interrelations between the topics.

EV Pricing

One of the foremost issues for EVs is the pricing of the vehicles themselves. The price of an EV will need to be low enough to entice buyers to make the investment and switch over to sustainable transportation. Even if EVs are priced higher than comparable gasoline-powered vehicles, as long as the perceived benefits are substantial (e.g., fuel cost savings, convenience, and image), consumers beyond the current niche group of EV owners will purchase EVs. Incentives and rebates from the government could be given to boost EV sales as well.

Automakers could also offer nonmonetary incentives for early EV adopters. For example, they could provide at-home installation of charging hardware to enable Level II and Level III charging. Another possibility is to offer a program that would allow purchasers to borrow, rent, or share a gasoline-powered vehicle for a limited number of days per calendar year, since driving range will be a concern to new EV buyers. Other creative options could be considered to help tailor EVs to better fit the needs of consumers.

Infrastructure Deployment

In order to support a vast fleet of EVs, a robust charging infrastructure needs to be in place. An obvious question is where to install charging stations, but the solution is less apparent. Depending on the initial deployment priorities, charging stations may be placed in clusters near heavily trafficked areas to maximize convenience, at waypoints along highways to permit long-distance travel, or at a handful of key locations to provide a satisfactory level of service. Their placement, however, will partially be dictated by the existing infrastructure and access to the grid. If the cost of expanding the grid or installing a dedicated circuit is prohibitive, charging stations may need to be installed in less desirable spots to satisfy consumer demand.

Charging stations in densely populated areas will serve the daily commuting needs of most EV drivers, but they will not benefit drivers who must travel longer distances. One of the goals of the EV charging infrastructure will be to extend the limited driving range of EVs, which will require the strategic placement of charging stations in less heavily populated yet still highly traveled areas. Drivers who regularly commute long distances might invest in an EV with a greater battery capacity, but for those who need the extra range infrequently, a second vehicle, likely gasoline-powered, would be necessary if the existing charging infrastructure were insufficient. The additional cost and inconvenience of owning a second vehicle would not suit a majority of EV drivers.

The establishment of the charging infrastructure for EVs will not occur overnight. It will, in fact, take years, even decades, to grow and expand. To best understand how the infrastructure will evolve over time, models will need to be developed. Important considerations will include consumer demand and charging behavior, which will influence the rate of expansion as well as the prioritization of range extension and coverage, and also the coexistence of the EV charging infrastructure with the existing gasoline station infrastructure. Knowing how consumers affect and react to changes in the EV charging infrastructure will facilitate

planning and decision making for charging station locations.

In determining how to build the charging infrastructure for EVs, it will be necessary to determine the source of funding as well. Many of the existing charging stations were funded by the companies that installed them, such as Better Place and Coulomb Technologies, Inc., and those companies collect a fee when the charging stations are used. Other charging stations, paid for by city councils, are free to use. As EVs increase in popularity, the decisions to install charging stations will become more complex as the number of stakeholders increases.

For example, consider a private business that wants to install charging stations in its parking lot. Some of the stakeholders include the business itself, which benefits by providing convenient charging to its patrons; the patrons, who are able to charge their EVs; the charging station manufacturer, which collects a fee each time one of the charging stations is used; and the EV manufacturers (particularly during the earlier phases of EV adoption), whose EVs gain value from the additional infrastructure. The goal is to fairly distribute the cost of the charging stations among the various stakeholders. Such a problem is complicated even further when public funds, which are collected from taxpayers (who may not all own EVs), are used to provide publicly accessible charging stations. Weighing the benefits to individual citizens as compared with society as a whole is both technically and ethically challenging.

Vehicle Telematics

Understanding EV driving patterns will be critical to providing effective charging services. Vehicle telematics provides a means of tracking EV road movements as well as monitoring charging behavior, and communications between the vehicles and the charging infrastructure can direct drivers to the nearest available charging stations. There already exists a software application for mobile devices allowing users to locate nearby charging stations and providing them with more advanced details such as the availability of charging stations and the status of a charging session in progress [23].

As more data becomes available from vehicle telematics systems, researchers will be able to better characterize EV driving patterns and identify opportunities to improve the charging infrastructure. Techniques such as data mining and network analysis will prove useful in uncovering patterns of driving and charging behaviors, which can then be applied to simulation studies to perform experiments and gain even deeper insights into the EV charging networks.

FINAL THOUGHTS

The future of EVs and sustainable transportation is no longer a pipe dream—it is imminent. Automakers have announced plans for mass producing their EV models. Charging station manufacturers have successfully demonstrated their products and installed them for public use. The vehicles and infrastructure are ready to go, but there is still plenty of modeling and planning to do before EVs become a major method of transportation. Together with automakers, charging infrastructure companies, policy makers, and the public at large, researchers in the fields of engineering, operations research, and management science can help to make sustainable transportation a reality.

REFERENCES

1. Better Place. Global progress. Available at <http://www.betterplace.com/global-progress>. Accessed 2010 June 5.
2. Clark J. Are electric cars better for the environment? Discovery News, 2010 July 30. Available at <http://news.discovery.com/tech/are-electric-cars-better-for-the-environment.html>. Accessed 2010 August 29.
3. Markel T, Simpson A. Plug In hybrid electric vehicle energy storage system design. Proceedings of Advanced Automotive Battery Conference. Baltimore, Maryland; 2006.
4. Pacific Gas and Electric Company. Electric vehicle infrastructure installation guide. San Francisco, California; Air Transportation Department 1999.
5. Parry IWH. How should heavy-duty trucks be taxed? *J Urban Econ* 2008;63(2):651–668.

6. Parry IWH. Should fuel taxes be scrapped in favor of pay-by-the-mile charges? *World Econ J* 2005;6(3):91–102.
7. Hadley SW, Tsvetkova AA. Potential impacts of plug in hybrid electric vehicles on regional power generation. *Electricity J* 2009;22(10):56–68.
8. Kintner-Meyer M, Schneider K, Pratt R. Impacts assessment of plug in hybrid vehicles on electric utilities and regional U.S. power grids, Part 1: technical analysis. Richland (WA): Pacific Northwest National Laboratory; 2006.
9. Lemoine DM, Kammen DM, Farrell AE. An innovation and policy agenda for commercially competitive plug in hybrid electric vehicles. *Environ Res Lett* 2008;3(1): 014003.
10. Chouinard HH, Perloff JM. Gasoline price differences: taxes, pollution regulations, mergers, market power, and market conditions. *B E J Econ Anal Policy* 2007;7(1).
11. Deltas G. Retail gasoline price dynamics and local market power. *J Ind Econ* 2004;56(3):613–628.
12. Axsen J, Mountain DC, Jaccard M. Combining stated and revealed choice research to simulate the neighbor effect: the case of hybrid-electric vehicles. *Resour Energy Econ* 2009;31(3):221–238.
13. Hess S, Train KE, Polak JW. On the use of a modified latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. *Transport Res B Methodol* 2006;40(2):147–163.
14. Spissu E, Pinjari AR, Pendyala RM, *et al.* A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. *Transportation* 2009; 36(4):403–422.
15. Squatriglia C. Better place unveils an electric car battery swap station. *Wired*, 2009 May 13. Available at <http://www.wired.com/autopia/2009/05/better-place>. Accessed 2010 June 5.
16. Yudai H, Osamu K. A safety stock problem in battery switch stations for electric vehicles. *Proceedings of the 8th International Symposium on Operations Research and its Applications*. Zhangjiajie, China; 2009.
17. Ameren. Day-ahead prices in dollars/kWh. Available at <http://www2.ameren.com/RetailEnergy/realtimeprices.aspx>. Accessed 2010 June 5.
18. Consolidated Edison, Inc. Price responsive program. Available at http://www.coned.com/energyefficiency/vol_time_pricing.asp. Accessed 2010 June 5.
19. SRP. SRP time-of-use price plan. Available at <http://www.srpnet.com/prices/home/tou.aspx>. Accessed 2010 June 5.
20. Kempton W, Tomic J. Vehicle-to-grid power implementation: from stabilizing the grid to supporting large-scale renewable energy. *J Power Sources* 2005;144(1):280–294.
21. Sovacool BK, Hirsh RF. Beyond batteries: an examination of the benefits and barriers to plug in hybrid electric vehicles (PHEVs) and vehicle-to-grid (V2G) transition. *Energy Policy* 2009;37(3):1095–1103.
22. Turton H, Moura F. Vehicle-to-grid systems for sustainable development: an integrated energy analysis. *Technol Forecast Soc Change* 2008;75(8):1091–1108.
23. iTunes Store. ChargePoint. Available at <http://itunes.apple.com/us/app/chargepoint/id356866743?mt=8>. Accessed 2010 June 5.

THE NATURALISTIC DECISION MAKING PERSPECTIVE

CHRISTOPHER NEMETH

GARY KLEIN

Klein Associates Division,
Applied Research Associates,
Fairborn, Ohio

The naturalistic decision making (NDM) approach seeks to understand human cognitive performance by studying how individuals and teams actually make decisions in real-world settings. Three major criteria have appeared in the literature to describe research that counts as NDM study: The research focuses on expertise, it takes place in field settings, and it reflects the conditions (such as uncertainty) that complicate our lives. The most important reason to conduct NDM research is the desire to sympathetically address a question of how individuals or teams and organizations make certain types of decisions in natural settings.

ORIGINS

While researchers had studied behavior in real-world settings, most research through the 1970s was done in the laboratory in order to verify mathematical and statistical models. This normative research approach explored how people *ought* to make decisions. Principles and constraints for decisions were derived from formal or mathematical systems such as deductive logic, Bayesian probability theory, and decision theory. The approach contended that the need to improve decision making arises because human decision makers systematically violate normative constraints [1]. Research in this vein relied substantially on abstractions of human behavior in order to analyze it. For example, the economic theory of Von Neumann and Morgenstern [2] and the expected utility theory of Keeney and Raiffa [3] contended that humans consider available options and choose the one with the highest expected

return. Operations research [4] described normal human performance in well-bounded, concrete domains through methods such as simulation, linear and nonlinear programming, dynamic programming, queueing and other stochastic-process models, Markov decision processes, econometric methods, data envelopment analysis, neural networks, expert systems, decision analysis, and the analytic hierarchy process; see Hillier and Lieberman [5] for more on operations research methods. Nearly all of these techniques involve the construction of reductive mathematical models to describe a system. As early as the 1970s, Shackel [6, p. 436] noted that operations research techniques such as dynamic programming were available “but fail when the problem involves many criteria with changing priorities”

Unfortunately, the training methods and decision support systems that were developed according to principles from these formal systems neither improved decision quality nor were adopted in field settings. People found these tools and methods cumbersome and irrelevant to the work they needed to do [7]. Other research initiatives offered greater potential to understand behavior by studying it in real-world settings. This came to be known as the *naturalistic decision making (NDM)* approach. One of the influences on NDM was Gibson’s [8] ecological psychology demonstrations of how organisms directly perceive their environment, rather than follow abstract experimental constructs. Neisser [9] complemented the Gibsons’ view, encouraging the study of cognition to be more realistic and seek to

- understand purposeful behavior in the real world (outside of the laboratory);
- understand the real-world’s details and the fine-grain structure of available information;
- grasp the kinds of cognitive skills people acquire and develop; and
- examine the fundamental questions of human nature.

Work by deGroot and Simon also formed the foundation for NDM. From 1938 to 1943, deGroot [10], p. 320] collected verbal protocols of chess players while conducting early field studies of how people make decisions. deGroot wanted to determine what separated those who excel at chess from others who do not. He found that the chess master "...sizes up positions more easily and, especially, more accurately." This expert skill appeared to rely on the mental simulation of options and the comparison of reactions to them. Simon [11,12] introduced the notion of *satisficing*, contending that business organizations deal with real-world complexity by using procedures that find "good enough" answers to questions when best answers cannot be obtained. People satisfice by looking for the first workable option, rather than trying to find the best possible option.

By the early 1980s, Kahneman *et al.* [13] showed that individuals rely on heuristics (rules of thumb) to make decisions. Their heuristics and biases (HB) paradigm demonstrated that people did not use strategies in the form of algorithms in order to follow principles of optimal performance. This occurred even when expected utility theory, probability laws, and statistics indicated that an individual was likely to choose certain optimal courses of behavior.

PRACTICE

Researchers need several kinds of skills, knowledge, and attitudes in order to conduct NDM studies. These can be thought of as models, methods, and mind-sets.

Models

The NDM models and theories will help researchers to notice relationships and form expectancies. Without them, these relationships are likely to be invisible. Followed too closely, though, models might also get in the way of taking a fresh look at a phenomenon. For example, the recognition-primed decision (RPD) model [14] describes how people use their experience in the form of a repertoire of patterns that describe a situation's primary causal factors.

Klein *et al.* [15] deliberately chose not to use the RPD model as a framework at the start of studies in sensemaking because it would get in the way of observing the phenomenon. They instead formulated a data/frame model of sensemaking, to examine the interaction between the interpreted signals of events (data) and explanatory structures that account for the data (frames). The data/frame model is compatible with the RPD model but emphasizes different processes and relationships than RPD. Similarly, in a recent investigation of causal reasoning (which is a form of sensemaking), Klein and Hoffman [16] chose to begin with a naturalistic exploration rather than try to extend the data/frame sensemaking model.

Methods

Cognitive task analysis (CTA) [17] is a set of methods such as structured interviews and observations that are used to elicit knowledge gained from real-world experience, analyze it, and use the results to develop representations including descriptions, diagrams, and models. Competence in the use of CTA methods to pursue the question "How, and why, are they doing that?" is more important than knowledge of NDM models. These may be methods to conduct interviews, either about tough cases or about the way a domain is understood; see Crandall *et al.* [17] and Hoffman and Militello [18] for more information. They may be field observation methods or they may be methods to review the existing literature and documentation [19,20]. NDM researchers' use of CTA methods sets them apart from the ecological psychology tradition, which attempts to account for performance in terms on environmental constraints rather than mental activities. NDM researchers typically focus on mental activities such as decision-making and sensemaking strategies, while also trying to be sensitive to the context of a situation.

It is not enough to just read about methods, because methods need to be used to be understood. This is particularly true of methods used to conduct interviews and field observations. The use of NDM methods is a skill, and novices can find it difficult to notice the key indicators or events that seem

so obvious to more experienced researchers. Coaching can improve their practice. Ideally, a more experienced researcher will participate in the same interview or the same observation. Many research methods require practice, such as neuropsychological experiments that depend on inserting implants and developmental studies that depend on gaining rapport with children. Ethnography also depends on skills in observing performance [21,22], and the use of NDM methods similarly requires practice and feedback.

While NDM methods and NDM models are treated separately here, the two are actually linked. The NDM models will determine the kinds of methods researchers use. The application of NDM methods will reflect the kinds of models researchers have in mind about functions such as decision making and sensemaking.

Mind-Sets

A researcher also needs a mind-set, or sensitivity, that is borne of experience. This can make it possible to notice opportunities to learn more about tacit knowledge, or to dig deeper into an application of expertise, or learn about an exercise of decision making or sensemaking or replanning. This type of “tuning” is developed with practice, like any other kind of expertise. HB researchers are impressive in their ability to spot examples of judgment and decision biases. Their long experiences in studying these biases and thinking about them have resulted in a perceptual/conceptual skill. Similarly, the study of expertise can and should lead NDM researchers to become sensitive to subtle cues that others would likely miss.

Professionals gamble on the kinds of skills they will gain as they continue their work. A professional’s mind-sets will make certain things pop out in a situation, and will obscure others. Ericsson and Charness [23] described the importance of deliberate practice and the limited number of skills one could address with deliberate practice. Skill with noticing how people size up situations or detect problems is different than skill with the detection of decision biases. Each kind of skill pays dividends for the conduct of relevant research. The NDM mind-set helps

researchers to conduct effective observations and interviews. It enables them to capture the way people use their experience to handle uncertainty, vague goals, and high stakes, just as the HB mind-set helps HB researchers study consistent errors in judgment.

EVOLUTION OF NDM

Rather than rely on formal models of decision making, the first NDM researchers began by conducting field research to try to discover the strategies that people used. The formal standards approach we just described had looked for ways that people were suboptimal. By contrast, NDM researchers wanted to find out how people were able to make tough decisions under difficult conditions such as limited time, uncertainty, high stakes, vague goals, and unstable conditions. Researchers had already studied these kinds of issues in fields such as medicine [24] and business [25].

The basic research program at the Army Research Institute for the Behavioral and Social Sciences began to fund several NDM researchers during the mid-1980s. The US Navy became interested in naturalistic decisions following the 1988 USS *Vincennes* shoot down incident, in which the crew of a US Navy Aegis cruiser mistook an Iranian commercial airliner for a hostile attacker and destroyed it. Both the Army and the Navy wanted to help people make high-stakes decisions under extreme time pressure and under dynamic and uncertain conditions.

The first NDM conference, in 1989, assembled researchers who studied decision making in field settings. In a chapter that emerged from that meeting, Lipshitz [26] identified no less than nine NDM models that had been developed in parallel. Working separately, multiple researchers all reached similar conclusions. They found that people did not generate and compare option sets, as those who followed the formal standards approach had contended. Instead, people used their experience to rapidly categorize situations, relying on some kind of synthesis of their experience to make these judgments, and the categories suggested appropriate courses of action.

Rasmussen’s [27] model of cognitive control, which distinguished skill-based,

rule-based, and knowledge-based (SRK) behavior, was one of the nine models. SRK explained behavior in terms of a ladder-style diagram that illustrated behavior based on knowledge, rules, or skills and the direct or cut-off paths that an individual might follow depending on experience and effort involved.

Another account, the RPD model [14], describes how people use their experience in the form of a repertoire of patterns. The patterns highlight the most relevant cues, provide expectancies, identify plausible goals, and suggest typical types of reactions in that kind of situation. When people need to make a

decision, they can quickly match the situation to the patterns they have learned. If they find a clear match, they can carry out the most typical course of action. They do not evaluate an option by comparing it to others, but instead imagine—mentally simulate—how it might be carried out. In that way, people can successfully make extremely rapid decisions. The RPD model is a blend of intuition (the pattern matching) and analysis (the mental simulation). The RPD model was based on in-depth interviews with fire-ground commanders about recent and challenging incidents. Those studies found that

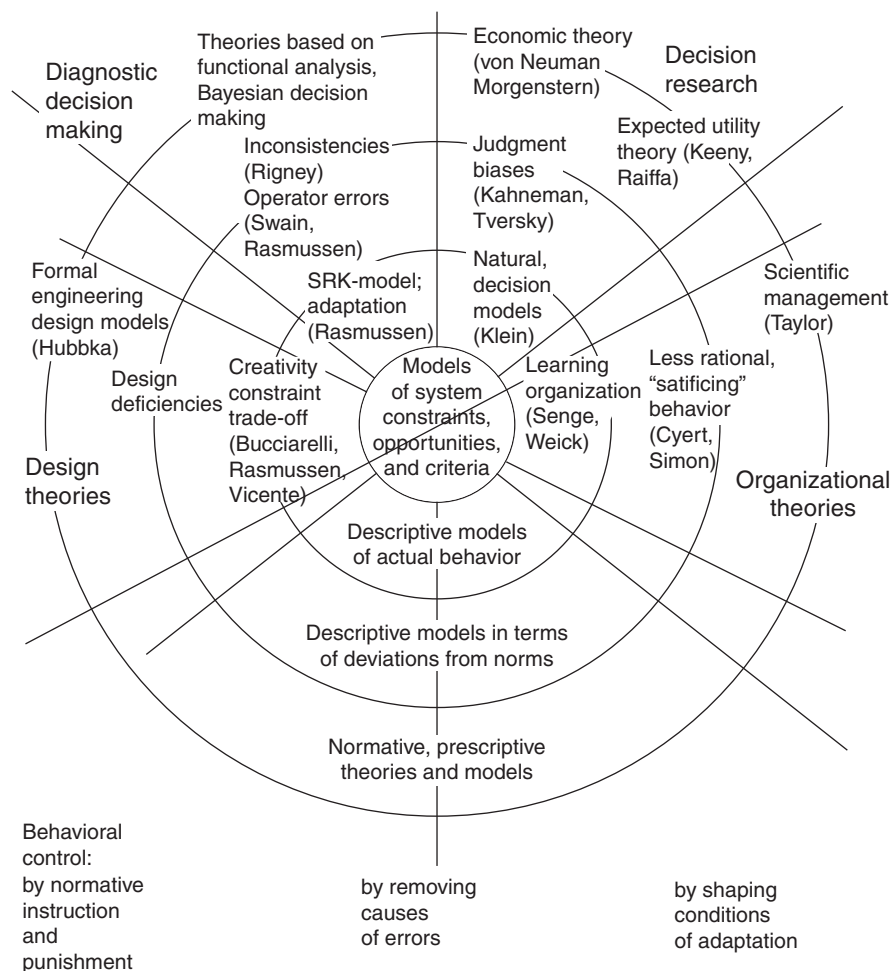


Figure 1. Behavioral modeling methods. (Source: Reprinted by permission of Ashgate Publishing [28]).

the percentage of RPD strategies generally ranged from 80% to 90%.

Figure 1 portrays Jens Rasmussen's [[28], p. 75] view of how methods to model behavior have evolved (from outer to inner circles) from normative models of rational behavior, to models in terms of deviation from rational behavior, to representations of actual behavior. The figure depicts NDM in this latter stage along with Rasmussen's SRK model, Weick's and Senge's concepts of learning organizations, and Bucciarelli, Rasmussen and Vicente's concept of creativity-constraint trade-off. Developed in the same time period, diagnostic and decision-making studies arrived at similar conclusions. For example, studies of process plant operation error reports led to the development of the SRK model as a replacement for decision-making models, and the skills and rules portion of the SRK model are similar to recognition-primed NDM model that evolved from field studies of fireground commanders.

CHALLENGES FOR NDM

Like any research approach, NDM has adherents and critics. The judgment and decision making (JDM) researchers from the formal standards tradition had largely ignored experts, or had devised paradigms to show the flaws in expert judgments. The lead article by Lipshitz *et al.* [29] on "NDM in behavioral decision making" elicited 16 commentaries by JDM researchers, many of them critical. For example, Klayman [30] questioned whether NDM's reliance on experts can produce results that can be generalized beyond the particular nature of their own experience. Tieggen [31] probed whether distinctions can be clarified between prescriptive and descriptive approaches, and between a research program and its results. Bazerman [32] queried whether the standard of expert decision making is suitable to provide a benchmark for real-world decisions by those who are not experts.

A few critical themes related to NDM have evolved since its start. The following are some of the main concerns people have with the field.

NDM Research and Science

Some people ask whether NDM research counts as science. While few NDM projects try to test hypotheses or generate falsifiable claims, they still follow a scientific approach. The scientific method begins with observation/examination of the phenomenon, followed by theory generation, and then theory testing. NDM research primarily takes place at the initial stage, with observation and examination. The basic NDM question "How/why do they do that?" is a way to gain new knowledge and formulate, or at least seed, new theories. The NDM community addresses an earlier stage of the scientific method than the JDM/HB communities. Suggestions that NDM studies institute control groups, randomize treatment effects, counterbalance conditions, and explicitly test hypotheses in order to become "more scientific" miss the added value that a front-end investigation of cognitive processes in natural settings provides.

Very few NDM studies set out to test a hypothesis, so the studies do not look like science to experimental psychologists who are used to conducting controlled experiments that manipulate variables in order to evaluate theories and describe lawful relationships. However, many NDM studies collect and analyze data. For example, Lipshitz and Strauss [33] investigated how military officers manage uncertainty. Klein *et al.* [34] studied how firefighters make decisions under time pressure and uncertainty. Both of these studies collected and analyzed data that could be used by other researchers who sought to replicate the findings. The article by Lipshitz and Strauss [33] is also widely cited in different research communities and several empirical replications of the firefighter study have been conducted.

Falsifiability is the second, related criterion for scientific research. The RPD model satisfied this criterion. Klein *et al.* [35] tested the hypothesis that the first option people considered was of reasonable quality, rather than being random. The results supported the RPD model, and Johnson and Raab [36] later replicated those results. Falsifiability encourages controversy that calls for more data collection. McLennan

et al. [37] criticized some of the limitations of the RPD model and demonstrated the importance of prepriming. Cohen *et al.* [38] similarly modified the RPD model in a recognition/metacognition model. Lipshitz and Ben Shaul [39] have criticized the data/frame model of sensemaking [15] because of vagueness regarding the concept of a frame.

Relationship between NDM and JDM/HB Traditions

The earlier discussion of mind-set noted that HB researchers have tuned themselves to notice things that NDM researchers miss, and ask tough questions about the skill level of subject-matter experts used in NDM research; see Shanteau [40], for an NDM approach to expertise assessment. Efforts are underway within JDM and HB schools to press beyond simple conclusions and to search for the cognitive basis for decision biases [41]. In the pursuit of these questions, JDM/HB researchers are making important findings about the strategies people use. If we can get past our aversion to the term “bias,” the NDM community can learn a great deal from these efforts. And if the JDM/HB community can get past its aversion to the term “expertise” we might be able to forge a productive dialog.

We can also improve NDM practice by making use of recent developments in behavioral economics, which [42] have summarized well. The behavioral engineering they describe seems entirely consistent with the NDM interest in finding natural ways to support and improve decision making.

The NDM community may also be able to use its experience in field settings and its background in human factors to suggest ways to improve JDM applications. We have all seen our share of failed systems and methods, and we have developed cognitive systems engineering and resilience engineering methods to overcome many problems. Bordley [43] suggested that these experiences and methods could be useful for JDM researchers who seek to develop applications. NDM methods of cognitive task analysis may also be useful for JDM researchers who look for insights into the strategies used by target audiences

[44]. Cooksey [45] identified several ways that NDM could partner with JDM to form an integrated decision science: emphasize the decision context, understand constraints on (as well as capabilities of) decision makers, avoid oversimplification, examine individual and organizational learning, and contrast focus of process and outcome.

The Treatment of Error

NDM has been greatly interested in error from its inception. One of the initiating events for NDM was the shoot down of an Iranian airliner by the USS Vincennes. The event contributed to a lengthy research program on decision making for Aegis cruiser command staff, “Tactical Decision Making Under Stress” (TADMUS). Misperceptions and erroneous acts that led to the incident at the Three Mile Island nuclear power facility helped to guide the cognitive systems engineering movement [46].

Reason [47] has distinguished between sharp end errors (the people in an organization who make poor judgments or make poor choices of actions) and blunt end errors (the organizational dynamics, such as inadequate training or conflicting priorities) that contributed to the sharp end errors. The JDM/HB communities have focused on the former. They have identified systematic errors that are independent of organizational pressures. These are the judgment and decision biases.

The NDM community needs to reexamine the HB literature because systematic errors clearly play a role in natural settings, and fields such as behavioral economics have been successful in generating meaningful interventions. At the same time, the JDM and HB communities can benefit from the NDM research on errors, particularly our examination of the dynamics of a situation that can increase the likelihood of errors [48]. Snook’s book on friendly fire is a good example of poor decisions that stemmed from organizational dynamics rather than from biases.

CONTINUING WORK

From the very start of the NDM community in 1989, researchers studied more than

just judgment and decision making. Theoretical constructs such as the RPD model [34] depended heavily on situation awareness, sensemaking, story construction, and mental simulation. Klein *et al.* [49] described this expansion as *macrocognition*. Continuing work in NDM seeks a better understanding of each of these areas, how they are related, and how they occur at the individual and team level.

Macro cognition has been described by others as follows: situated cognition and extended cognition [50]; the cognitive adaptation to complexity [51]; and the cognitive functions that are performed by actual practitioners in natural decision-making settings that are generally performed by a team working together in a natural situation, and usually in conjunction with computational artifacts [52]. The primary reason for describing macrocognitive functions is that many researchers have received their initial training in laboratory settings. They are not used to considering functions such as problem detection that are not easily studied under controlled conditions. The desire to control the experimental conditions is common in cognitive research. By contrast, the NDM orientation is to be inclusive and curious about different aspects of cognition that will affect the way people handle the conditions such as limited time, uncertainty, high stakes, vague goals, and instability [53]. The term *macrocognition* is intended to change the mind-set of the researcher. Instead of feeling discomfort regarding these types of complexities we want NDM researchers to feel excitement and curiosity about how people adapt to them.

The primary macrocognitive functions are decision making, sensemaking, planning/replanning, problem detection, learning/adaptation, and coordination. In most settings, these functions are related to each other. Decision making usually depends on sensemaking. Planning/replanning depends on problem detection and sensemaking and decision making. Replanning depends on the diagnosis of difficulties, which is a form of learning that relies on sensemaking. Coordination of resources and people depends on problem detection, sensemaking and

replanning, and leads to decision making. Because of this an NDM investigation will rarely be confined to only a single function. The research may highlight one or another of the functions, but the researchers will benefit from knowledge about the other functions and processes. That is the reason NDM researchers have been able to go back to the studies of a macrocognitive function such as decision making and use the incident accounts in studies of other functions such as sensemaking and problem detection.

REFERENCES

1. Lipshitz R, Cohen MS. Warrants for prescription: analytically and empirically based approaches to improving decision making. *Hum Factors* 2005;47(1):102–120.
2. Von Neumann J, Morgenstern O. *Theory of games and economic behavior*. Princeton (NJ): Princeton University Press; 1947.
3. Keeney RL, Raiffa H. *Decisions with multiple objectives: preferences and value trade-offs*. New York: Cambridge University Press; 1993.
4. Morse PM, Kimball GF. *Methods of operations research*. Cambridge (MA): The Technology Press of MIT; 1951.
5. Hillier FS, Lieberman GJ. *Introduction to operations research*. 8th ed. New York: McGraw-Hill; 2005.
6. Shackel B. Process control-simple and sophisticated display aids as decision devices. In: Sheridan T, Johannsen G, editors. *Monitoring behavior and supervisory control*. New York: Plenum Press; 1979. pp. 429–443.
7. Yates J, Veinott ES, Patalano AL. Hard decisions, bad decisions: on decision quality and decision aiding. In: Schneider SL, Shanteau J, editors. *Emerging perspectives on judgment and decision research*. New York: Cambridge University Press; 2003. pp. 13–63.
8. Gibson JJ. *The perception of the visual world*. Boston (MA): Houghton Mifflin; 1950.
9. Neisser U. *Cognition and reality: principles and implications of cognitive psychology*. New York: WH Freeman; 1976.
10. deGroot AD. *Thought and choice in chess*. 1st ed. New York: Mouton; 1946/1978.
11. Simon HA. *Models of man: social and rational*. New York: John Wiley & Sons; 1957.
12. Simon HA. *The sciences of the artificial*. Cambridge (MA): The MIT Press; 1969.

13. Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: Heuristics and biases. Cambridge (MA): Cambridge University Press; 1982.
14. Klein GA. Recognition-primed decisions. In: Rouse WB, editor. Volume 5, Advances in man-machine systems research. Greenwich (CT): JAI Press; 1989. pp. 47–92.
15. Klein G, Phillips JK, Rall E, *et al.* A data/frame theory of sensemaking. In: Hoffman RR, editor. Expertise Out of Context: Proceedings of the 6th International Conference on Naturalistic Decision Making. Mahwah (NJ): Lawrence Erlbaum & Associates; 2007.
16. Klein G, Hoffman RR. Causal reasoning: initial report of a naturalistic study of causal inferences. Proceedings of NDM9, the 9th International Conference on Naturalistic Decision Making; 2009 June. London: British Computer Society; 2009.
17. Crandall B, Klein G, Hoffman RR. Working minds: a practitioner's guide to cognitive task analysis. Cambridge (MA): The MIT Press; 2006.
18. Hoffman RR, Militello LG. Perspectives on cognitive task analysis: historical origins and modern communities of practice. New York: Taylor & Francis; 2009.
19. Allison GT. Essence of decision: explaining the Cuban missile crisis. Boston (MA): Little Brown & Co; 1971.
20. Snook SA. Friendly fire: the accidental shoot-down of U.S. Black Hawks over northern Iraq. Princeton (NJ): Princeton University Press; 2002.
21. Clancey WJ. Field science ethnography: methods for systematic observation on an expedition. *Field Methods* 2001;13:223–243.
22. Fetterman DM. Ethnography: step by step. 3rd ed. Thousand Oaks (CA): Sage Publications; 1998.
23. Ericsson KA, Charness N. Expert performance: its structure and acquisition. *Am Psychol* 1994;49(8):725–747.
24. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge (MA): Harvard University Press; 1978.
25. Isenberg DJ. How senior managers think. *Harv Bus Rev* 1984; November-December: 81–86.
26. Lipshitz R. Converging themes in the study of decision making in realistic settings. In: Klein GA, Orasanu J, Calderwood R, *et al.*, editors. Decision making in action: models and methods. Norwood (NJ): Ablex; 1993. pp. 103–137.
27. Rasmussen J. Skill, rules and knowledge: signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans Syst Man Cybern* 1983; SMC-130:257–266.
28. Rasmussen J. Merging paradigms: decision making, management, and cognitive control. In: Flin R, Salas E, Strub M, *et al.*, editors. Decision making under stress. Aldershot: Ashgate Publishing; 1997. pp. 67–81.
29. Lipshitz R, Klein G, Orasanu J, *et al.* Focus article: taking stock of naturalistic decision making. *J Behav Decis Making* 2001; 14:331–352.
30. Klayman J. Ambivalence in (not about) naturalistic decision making. *J Behav Decis Making* 2001;14(5):372–373.
31. Tieggen KH. Two fuzzy themes in a clear message. *J Behav Decis Making* 2001;14(5): 379–380.
32. Bazerman MH. The study of “real” decision making. *J Behav Decis Making* 2001;14(5): 353–354.
33. Lipshitz R, Strauss O. Coping with uncertainty: a naturalistic decision making analysis. *Organ Behav Hum Decis Process* 1997;66:149–163.
34. Klein GA, Calderwood R, Clinton-Cirocco A. Rapid decision making on the fireground. Proceedings of the Human Factors and Ergonomics Society 30th Annual Meeting; Dayton; (OH). 1986. pp. 576–580.
35. Klein G, Wolf S, Militello L, *et al.* Characteristics of skilled option generation in chess. *Organ Behav Hum Decis Process* 1995;62(1): 63–69.
36. Johnson JG, Raab M. Take the first: option generation and resulting choices. *Organ Behav Hum Decis Process* 2003;91(2): 215–229.
37. McLennan J, Holgate M, Omodei MM, *et al.* Human information processing aspects of emergency incident management decision making. In: Cook M, Noyes J, Masakowski Y, editors. Decision making in complex environments. Aldershot: Ashgate; 2007. pp. 143–151.
38. Cohen MS, Freeman JT, Wolf S. Meta-recognition in time-stressed decision making: recognizing, critiquing, and correcting. *Hum Factors* 1996;38:206–219.

39. Lipshitz R, Ben Shaul O. Schemata and mental models in recognition-primed decision making. In: Zsombok C, Klein G, editors. *Naturalistic decision making*. Mahwah (NJ): Lawrence Erlbaum & Associates; 1997. pp. 293–304.
40. Shanteau J. Psychological characteristics and strategies of expert decision makers. *Acta Psychol* 1988;68:203–215.
41. Gilovich T, Griffin D. Introduction -heuristics and biases: then and now. In: Gilovich T, Griffin D, Kahneman D, editors. *Heuristics and biases: the psychology of intuitive judgment*. Cambridge (MA): Cambridge University Press; 2002. pp. 1–18.
42. Thaler RH, Sunstein CR. *Nudge: improving decisions about health, wealth, and happiness*. New Haven: Yale University Press; 2008.
43. Bordley RF. Naturalistic decision making and prescriptive decision theory. *J Behav Decis Making* 2001;14:355–357.
44. Sieck WR, Klein G. Decision making. In: Durso FT, editor. *Handbook of applied cognition*. West Sussex: John Wiley & Sons; 2007. pp. 918.
45. Cooksey RW. Pursuing an integrated decision science: does “Naturalistic Decision Making” help or hinder? *J Behav Decis Making* 2001;14:361–362.
46. Rasmussen J, Pejtersen AM, Goodstein LP. *Cognitive systems engineering*. New York: John Wiley & Sons; 1994.
47. Reason J. *Human error*. Cambridge (MA): Cambridge University Press; 1990.
48. Dekker SWA. Illusions of explanation: essay on error classification. *Int J Aviat Psychol* 2003;13(2):95–106.
49. Klein G, Ross KG, Moon BM, *et al.* *Macro cognition*. *IEEE Intell Syst* 2003;18(3):81–85.
50. Hollnagel E. Extended cognition and the future of ergonomics. *Theor Issues Ergon Sci* 2001;2(3):309–315.
51. Schraagen JM, Klein G, Hoffman R. The macrocognitive framework of naturalistic decision making. In: Schraagen JM, Militello L, Ormerod T, *et al.*, editors. *Naturalistic decision making and macrocognition*. Hampshire: Ashgate; 2008. pp. 3–26.
52. Cacciabue PC, Hollnagel E. Simulation of cognition. In: Hoc JM, Cacciabue PC, Hollnagel E, editors. *Applications, expertise and technology: cognition and human-computer cooperation*. Mahwah (NJ): Lawrence Erlbaum Associates; 1995. pp. 55–73.
53. Orasanu J, Connolly T. The reinvention of decision making. In: Klein GA, Orasanu J, Calderwood R, *et al.*, editors. *Decision making in action: models and methods*. Norwood (NJ): Ablex; 1993. pp. 3–20.

FURTHER READING

- Flin R, Salas E, Strub M, *et al.*, editors. *Decision making under stress: emerging themes and application*. Aldershot: Ashgate Publishing; 1997.
- Hoffman RR, editor. *Expertise out of context*. New York: Taylor & Francis; 2007.
- Klein G. Recognition-primed decision making: looking back, looking forward. In: Klein G, Orasanu J, Calderwood R, *et al.*, editors. *Decision making in action: models and methods*. Norwood (NJ): Ablex Publishing; 1993.
- Klein G. *Sources of power: how people make decisions*. Cambridge (MA): The MIT Press; 1998.
- Klein G. *Streetlights and shadows: searching for the keys to adaptive decision making*. Cambridge (MA): The MIT Press; 2009.
- Klein G. *The power of intuition*. New York: Random House; 2003.
- Klein G. Naturalistic decision making. *Hum Factors* 2008;50(3):456–460.
- Montgomery J, Lipshitz R, Brehmer B, editors. *How professionals make decisions*. Mahwah (NJ): Lawrence Erlbaum & Associates; 2005.
- Mosier K, Fischer U, editors. *Expert performance in complex situations*. New York: Taylor & Francis. In preparation.
- Salas E, Klein G. *Linking expertise and naturalistic decision making*. Mahwah (NJ): Lawrence Erlbaum & Associates; 2001.
- Schraagen JM, Militello L, Ormerod T, *et al.*, editors. *Naturalistic decision making and macrocognition*. Farnham: Ashgate Publishing; 2008.
- Zsombok CE, Klein G. *Naturalistic decision making*. Mahwah (NJ): Lawrence Erlbaum & Associates; 1997.

THE NORTH AMERICAN OPERATIONS RESEARCH SOCIETIES

KARLA HOFFMAN
George Mason University,
Systems Engineering
and Operations
Research Department,
Fairfax, Virginia

Each of the operations research societies that exist in North America has similar goals. The purpose of these is to promote the advancement of knowledge, interest, and education in operations research by providing mechanisms for the exchange of information through the organization of conferences, the promotion of advances in the field, and the production of books, journals, magazines, videos, web materials, and other media that describe the developments, technologies, and successes of the field. Multiple prizes are awarded by each society honoring such achievements.

Each of the societies is also part of a larger worldwide umbrella organization known as the *International Federation of Operations Research* (IFORS; <http://www.ifors.org>). IFORS is divided into regions, and the two societies that are officially part of the North American (NORAM) group are The Canadian Operations Research Society (CORS) and the Institute for Operations Research and the Management Sciences (INFORMS). The Mexican Institute of Systems and Operations Research/Sociedad Mexicana de Investigacion de Operaciones (IMSIO/SMIO) does not belong to this regional group but rather to the Association of Latin-Iberoamerican Operations Research Societies (ALIO), which includes societies representing Bolivia, Chile, Columbia, Cuba, Ecuador, Mexico, Peru, and Venezuela as well as the Spanish Society of Statistics and Operations Research.

CORS (<http://www.cors.org>) was established in 1957 (the same year during which

IFORS was created) and held its inaugural meeting in 1958. CORS serves its members through a variety of publications and services including its quarterly *Bulletin* (started in 1962); its journal, *Information Systems and Operations Research* (INFOR), which is published jointly with the Information Processing Society of Canada, published quarterly; a traveling speaker program; an annual Graduate Student Conference; and grants to attend teaching effectiveness workshops. In addition, CORS honors its members with a variety of awards including the Harold Larnder Prize for international distinction in operations research; the CORS Award of Merit; the CORS Service Awards; and the CORS Practice Prize. CORS holds an annual conference that is often held jointly with “sister organizations” such as the Information Processing Society of Canada. CORS includes a variety of sections including Atlantic, Québec, Kingston, Toronto Student, Southwest Ontario, Waterloo Student, Winnipeg, Saskatoon, Calgary, Edmonton, Vancouver, and Vancouver Student.

INFORMS (<http://www.informs.org>) is the largest operations research society in the world. INFORMS holds two major conferences per year as well as a variety of smaller meetings that are organized by its many special-interest groups, societies, and fora. The fall meeting, known as the *National meeting*, is primarily a forum for operations research academics and has the largest attendance of any operations research meeting worldwide. The other meeting, labeled the Analytics Conference, is directed toward practitioners. The Analytics Conference includes tutorials and workshops related to the use of advanced analytics within business and government; workshops for operations research analysts on soft skills that enhance one’s ability to communicate ideas and results to management; and presentations by major industry figures on operations research successes as well as an executive forum. In addition to these meetings, INFORMS often plans a summer

International meeting which is usually organized jointly with another operations research society.

INFORMS is composed of two former organizations: The Operations Research Society of America (ORSA), which was found in 1952, and the Institute of Management Science (TIMS) found in 1953. Both these organizations had similar objectives, although ORSA was often perceived to be more engineering and academically oriented with a strong military focus, whereas TIMS was considered to be more business and applications focused with a stronger international perspective. From the time of their establishment, these societies often held joint conferences, had many officers in common, and produced joint publications. In 1974, the organizations held their first joint National meeting and continued to do so until the merging of the organizations in 1995. The discussion of merging the two organizations took place as early as 1957 [1] and in 1975, a "Blue Ribbon Committee" was formed to identify how the two societies could cooperate more effectively [2]. For a historical description of the merging of these two societies and how it was finally accomplished, refer to the article by Keller and Kirkwood [3]. For a more complete description of the activities of the professional societies, refer to the articles by Assad and Gass [4] and Salveson [5].

INFORMS publishes multiple journals (*Decision Analysis*, *Information Systems Research*, *INFORMS Journal on Computing*, *Interfaces*, *Management Science*, *Manufacturing and Service Operations Management*, *Marketing Science*, *Mathematics of Operations Research*, *Operations Research*, *Organizational Science*, *Service Science*, *Transportation Science*, and *INFORMS Transactions on Education*) as well as a tutorial and book series. It publishes three magazines: *OR/MS Today*, *Analytics*, and *OR/MS Tomorrow*. Other programs include a speakers program, a high-school teachers' program, a History and Traditions Committee that encourages and facilitates the collection and preservation of materials relating to the history and traditions of OR/MS, a doctoral colloquium, a teaching

effectiveness colloquium, and young practitioners' workshop. INFORMS is divided into communities; as of 2012, there are 10 societies (*Applied Probability*, *Computing*, *Decision Analysis*, *Health Applications Society*, *Information Systems*, *Manufacturing & Service Operations Management*, *Marketing Science*, *Military Applications Society*, *Optimization*, *Simulation*, and *Transportation Science & Logistics*), 22 subdivisions, 5 fora (interest groups that are neither discipline nor geographical), and 31 regional groups. These entities hold their own meetings as well as contribute substantially to the content of the national meeting.

The institute recognizes the achievements of its members through a variety of prizes, including the John von Neumann Theory Prize; the INFORMS President's award; the INFORMS Fellows awards; the George E. Kimball Award for distinguished service; the INFORMS Prize for the Teaching of OR/MS Practice; and the INFORMS Impact Award. Many other prizes are awarded through the subdivisions.

Almost from the beginning, CORS and INFORMS have regularly held joint meetings, and have provided each other's members with reduced prices for membership and journal subscriptions. More recently, these two societies have worked with ALIO to organize joint meetings that encourage collaboration and dissemination among the Americans.

There are a number of "sister" organizations with research interests that are closely aligned with those of INFORMS and CORS. The most closely aligned is the Institute of Industrial Engineers (IIE), which is the world's largest professional society dedicated solely to the support of the industrial engineering profession and individuals involved with improving quality and productivity. Founded in 1948, IIE is an international, nonprofit association that provides leadership for the application, education, training, research, and development of industrial engineering. For more on this organization, visit <http://www.iienet.org>. There are many other organizations that also represent individuals whose research technologies are closely related to those of operations

research analysts. We end this short note with a listing of those organizations and their web sites:

- Academy of Marketing Science (<http://www.ams-web.org/>);
- AGIFORS (Airline Group of the International Federation of Operational Research Societies; <http://www.agifors.org/>);
- Academy of Management's Operations Management Division (<http://om.aomonline.org/>);
- ASQ (American Society for Quality; <http://www.asq.org/>);
- DSI (Decision Sciences Institute; <http://www.euro-online.org/>);
- HIMSS (Healthcare Information Management and Systems Society; <http://himss.org/>);
- IFIP (International Federation for Information Processing; <http://www.ifip.or.at/>);
- LMI (Logistics Management Institute; <http://www.lmi.org/>);
- MORS (Military Operations Research Society; <http://www.mors.org/>);
- The Mathematical Programming Society (<http://www.mathopt.org/>);
- PMSA (Pharmaceutical Management Science Association; <http://www.pmsa.net/>);
- POMS (Production and Operations Management Society; <http://www.poms.org/>);
- SCS (Society for Computer Simulation International; <http://www.scs.org/>);

- SDP (The Society of Decision Professionals; <http://www.decisionprofessionals.com/>);
- SIAM (Society for Industrial and Applied Mathematics; <http://www.siam.org/>);
- SIM (Society for Information Management; <http://www.simnet.org/>);
- SJDM (Society for Judgment and Decision Making; <http://www.sjdm.org/>);
- SOLE (Society of Logistics Engineers; <http://www.sole.org/>);
- TRB (Transportation Research Board; <http://www.nas.edu/trb/>);
- WSC (Winter Simulation Conference; <http://www.wintersim.org/>).

REFERENCES

1. Lathrop JB. "A proposal for Merging ORSA and TIMS" Letter to the Editor. *Operations Research* 1957; 5(1) 123–125.
2. Blumstein A, Boodman DM, Davidson HJ, Liff J, Oliver R, Wagner H. ORSA/TIMS collaboration. *Interfaces* 1975; 3(4): 32–42.
3. Keller R, and Kirkwood CW. The founding of INFORMS: a decision analysis approach *Operations Research* 1999; 47(1): 16–28.
4. Assad AA, Gass SI. An annotated timeline of operations research: an informal history. International Series in Operations Research & Management Science. Springer; 2004. p. 211.
5. Salveson ME. The Institute of Management Sciences: a prehistory and commentary on the occasion of TIMS 40th anniversary. *Interfaces* 1997; 27(3): 74–85.

THE OPERATIONAL RESEARCH SOCIETY OF SINGAPORE

LAI KAH WAH
Operational Research Society of
Singapore, Kent Ridge,
Singapore

HISTORY OF ORSS

How was the Operational Research Society of Singapore (ORSS) founded? At a June 1974 seminar on “Application of OR” held at the (then) Nanyang University, the participants strongly felt that an association should be formed to promote OR in Singapore [1]. In December 1974, a *protem* committee was formed. By February 1975, the society name had been decided upon and the constitution was drafted. ORSS was officially registered on November 21, 1975. The inaugural Annual General Meeting (AGM) was held on March 26, 1976.

ORSS was admitted as a full member of the International Federation of Operations Research Societies (IFORS) in May 1978. ORSS is a founding member society of the Asia-Pacific Operations Research Societies (APORS) in August 1984; APORS was formally launched in March 1985. The founding president of ORSS, Prof Chew Kim-Lin, was APORS president for 1998–2000.

From its beginning with a handful of members, ORSS grew to more than a hundred members in the late 1980s, despite a deliberate decision to keep it as narrow-based as possible. The growth can be attributed to the establishment of OR [or operations analysis (OA) in military parlance] as a technical discipline in the Singapore defense community [2]. The sustained use of OR-related techniques, especially discrete-event simulation and mathematical programming, by other large home-grown organizations namely the Port of Singapore Authority (now PSA Corporation), Singapore Airlines,

and the Land Transport Authority likewise encouraged their in-house practitioners to join ORSS.

Since then, membership has been maintained at hundred-plus persons, though the proportion—and number—of life members have grown steadily each year. Along with the retirement of all the pioneer members, ORSS has seen regular, albeit small, annual enrollments of young analysts with interest in exploring OR as a career skill-set.

ACCOMPLISHMENTS

For the first decade of its existence, ORSS organized several small-scale seminars with academic focus.

Talks on actual applications in industry were gradually added, from the 1990s onwards. These talks and the academic seminars, still going on regularly, are aimed at raising current awareness of issues, techniques, and tools.

In June 1995, ORSS cohosted with the National University of Singapore (NUS), the first-ever INFORMS International Conference. This event was assessed as an unqualified success, attracting a few hundred participants from all over the world.

In July 2000, ORSS cohosted the Fifth Triennial APORS Conference. This was held over three days with more than 200 attendees from the Asia-Pacific region.

To encourage students' interest at both undergraduate and postgraduate levels, ORSS sponsors prize awards annually to the top graduating students in the NUS B Eng Industrial and Systems Engineering (ISE) and M Sc (ISE) courses who had specialized in the OR tracks.

Occasional industrial visits to organizations applying OR are also organized. Such visits serve to highlight practical aspects of ground OR to younger and student members.

ORSS has just initiated discussions with an advisory director of the US Military Operations Research Society (MORS). Given

the sizable number of ORSS members working on defense and national security, there is much interest in exploring technical dialogues and joint events with MORS.

PUBLICATIONS

The ORSS newsletter was started in 1977, averaging two to three issues per year. In the pre-Internet years, that is the 1970s–1980s, hardcopies of the newsletters were printed and distributed to both, individual members and organizations. Beginning in the early 2000s, dissemination of newsletters was switched to soft-copies. Recognizing the need for quick, timely, bite-sized information since then, ORSS now keeps its members updated via emailing lists. For members' convenience, larger files are uploaded to the ORSS' website. Some more technology-savvy members are suggesting Facebook and Twitter as alternative Internet-based channels.

ORSS first published the Asia-Pacific Journal of Operational Research (APJOR) in 1984. APJOR was subsequently adopted as APORS' main technical journal. APJOR provides a forum for practitioners, academics, and researchers in OR and related fields, within and beyond the Asia-Pacific region [3].

As for milestone publications, two special compilations of largely practice-oriented articles were published to commemorate the 10th and 20th anniversaries of ORSS.

REGULAR MEETINGS

As called for in its constitution, ORSS management committee (MC) meetings are held quarterly. Members of the MC are elected, or reelected, at each AGM, usually held in April of each year. For diversity, each MC usually comprises a balanced mix of members from different organizations and sectors. Representatives from the MC are also nominated to attend APORS and IFORS Council meetings.

OFFICE HOLDERS

So far, ORSS has had 11 presidents, from academia, government, and industry. Their

names and their terms of office are as tabulated:

S/N	Name ^a	Term of Office
1	<i>Chew</i> Kim-Lin	1976–1978, 1982, 1984–1985, 1987, 1989–1990
2	William <i>Liu</i> Wei Hai	1979–1981
3	<i>Kwok</i> Kah Chee	1983
4	<i>Yeong</i> Wee Yong	1986
5	<i>Tan</i> Thiam Soon	1988
6	<i>Liu</i> Kwai Wah	1991–1995
7	<i>Sim</i> Cheng Hwee	1996–1997
8	<i>Yu</i> Wei Shin	1998
9	<i>Poh</i> Kim Leng	1999–2001
10	<i>Tan</i> Kok Choon	2002–2005
11	<i>Lai</i> Kah Wah	2006–current

^aFamily name italicized.

The other members of the current (2009) 34th MC are: *Tan* Kok Choon (Vice President), *Ng* Kien Ming (Honorary Secretary), *Patrick Hoe* Puay Hian (Honorary Treasurer), *Ryan Chang* Khang Yang (Assistant Honorary Secretary), *Tiah* Yao Ming, *Andrew Tan* Chee Seng, *Alvin Ong* Siau Wah, and *Toh* Kaiyang (Committee Members). The two honorary auditors are *Ng Ee Chong* and *Pim Beers*.

REFERENCES

1. Chew KL. A brief history of the society. In: Hioe W, *et al.*, editors. ORSS 10th anniversary special publication. Singapore: ORSS; 1987.
2. Sim CH. Operations research in Singapore's defense establishment. In: Goh M, Tang LC, editors. ORSS 20th anniversary special publication. Singapore: ORSS; 1995.
3. Asia-Pacific Journal of Operational Research (APJOR). Available at <http://www.worldscinet.com/apjor/apjor.shtml>. Accessed 2009 Jun 30.

THE OPERATIONAL RESEARCH SOCIETY

GRAHAM SHARP
The OR Society, Birmingham, UK

HISTORY

In April 1948, the Operational Research Club was established in London as the first OR association in the world. Five years later the OR Club became The OR Society.

The term *Operational Research* was first used in 1938, as a descriptive term for the use of scientists to assess, at first hand, military situations and the deployment of devices therein. An example of this was the use of scientists in the process of evaluation of radar as an aid to air defence up to and including the War years.

During the Second World War, OR sections were set up in various branches of the British armed forces. These were followed by the establishment of similar sections in the services of the United States and Europe. After gaining distinction throughout the War years, the advocates of OR sought to secure its future in peacetime. There followed a period from 1945 until the mid-seventies when use of OR methodologies expanded rapidly into nationalised industries, the civil government, and the corporate sector.

Although OR may have been born in 1938, it was to be another 10 years before it developed a professional body. This came about when, in the autumn of 1947, a group of distinguished gentlemen met for an informal dinner at the Athenaeum Club in London. At this dinner, a momentous decision was made to form an OR Club. Those at the dinner included luminaries such as Sir Charles Goodeve, Professor P.M.S. Blackett (later Lord Blackett), Dr C. Gordon, and Sir Charles Tizard. The OR Club, the world's first body set up to cater for the OR profession, was inaugurated in April 1948 with an initial membership of 50.

In a communication received by the OR Society dated 16th August 1976, Sir Charles Goodeve explained the delay between founding the OR Club and inaugurating The OR Society: "We had to call it a club first of all to get over certain legal difficulties with regard to membership. We thought these difficulties would be solved in a couple of years, but in the event it took five or six years. However, the time was not wasted because we got the Quarterly going and our constitution finally settled." The Quarterly to which Goodeve referred was the *Operational Research Quarterly (ORQ)*, the world's first OR journal, which later became the *Journal of the Operational Research Society (JORS)*.

Volume 1, No. 1 of the *ORQ* was duly published in March 1950—the first editors were Max Davies and Roger T. Eddison. In the editorial notes of that first issue, the reasoning behind the Quarterly is clearly set out. "The main purpose of the *ORQ* is to assemble in one place as much as possible of the information that OR workers now find (or fail to find) scattered widely over the very large body of the scientific and technical literature. The method is to provide a quarterly collection of abstracts of relevant papers and articles, taken from as wide a field as possible." The first article to appear in the *ORQ* was "Operational Research" by Professor P.M.S. Blackett. The Quarterly continued as such until 1978, after which time it was renamed the *JORS* with 12 issues a year.

By 1955, interest in OR had spread to most western countries. Membership levels grew steadily, the original 50 had increased to a total of around 1250 members in 1964. Membership of the Society reached a plateau at around 3000 in the early 1970s and has remained at roughly the same level ever since.

THE OPERATIONAL RESEARCH SOCIETY TODAY

The OR Society continues to be an innovator and leader in its field. Today, The Society has

around 3000 members in 53 countries who receive a wide range of services and benefits. The Society's training programme is the most comprehensive OR training programme on offer anywhere in the world. Its annual conference is the largest national conference outside the United States, and its series of YoungOR conferences, inaugurated in 1980, is the world's leading conference for young OR workers.

Inside OR is the world's leading monthly OR magazine, carrying news, features on OR methods, personalities and events, debate, and a comprehensive notice board section providing information on what's happening in the OR world.

As a means of raising the profile of OR, The OR Society set itself an ambitious "stretch goal" that every school child should have heard of OR. The OR in Schools initiative has been the driver for the development of a website containing materials that can be used by mathematics teachers, young people, and careers advisers. The Society has also produced a DVD, downloadable from the website www.LearnAboutOR.co.uk that explains what OR is and how it helps people to make better decisions. The DVD won the British Institute of Videography Award for the Best Corporate Video 2009.

In addition to its own website www.theorsociety.com, The OR Society has two other websites—www.scienceofbetter.co.uk—developed in association with the Institute for Operations Research and the Management Sciences (INFORMS) as a mechanism to communicate the benefits of OR to the business community and www.LearnAboutOR.co.uk which was designed in response to the OR Society's OR in Schools initiative.

NOTEWORTHY ACCOMPLISHMENTS

- The world's first OR Society
- The world's first OR Journal
- Comprehensive training programme
- Largest national conference outside the United States
- Young OR conference

- Launched two specific Journals for Knowledge Management and Simulation
- Website designed in response to the OR Society's OR in Schools initiative—www.LearnAboutOR.co.uk
- Award-winning DVD about OR

PUBLICATIONS

- *Inside OR* is the magazine circulated to all members of The OR Society. The magazine is distributed both in print and electronically every month. *Inside O.R.* contains news about current events in OR, forthcoming meetings, conferences and training courses, awards and prizes that are available, and articles of interest to the OR community.
- *OR Insight* is an interesting and stimulating publication that appeals not only to OR practitioners and consultants, but also to managers and others wishing to learn more about OR in practice. The journal seeks to provide information about not just the scope and potential of OR but also about developments in related areas. Contributions are particularly sought from OR practitioners, consultants, and others with experience in OR.
- *The Journal of the Operational Research Society (JORS)* is published 12 times a year by Palgrave Macmillan on behalf of The OR Society. JORS is the world's longest established OR Journal and aims to present papers that are relevant to practitioners, researchers, teachers, students, and consumers of OR and which cover the theory, practice, history, or methodology of OR
- *Knowledge Management Research & Practice (KMRP)* is published quarterly by Palgrave Macmillan on behalf of The OR Society. KMRP provides an outlet for high quality, peer-reviewed articles on all aspects of managing knowledge, organisational learning, intellectual

capital, and knowledge economics. This includes not just those focused on the organisational level, but all levels from that of the individual to that of the nation or profession and includes both theoretical and practical aspects, and especially the relationship between the two. There is a particular emphasis on cross-disciplinary approaches and on the mixing of “hard” (e.g., technological) and “soft” (e.g., cultural or motivational) issues. Rigorous contributions from both academics and practitioners are welcomed.

- *The Journal of Simulation (JOS)* is published four times a year by Palgrave Macmillan on behalf of The OR Society. The journal publishes both articles and technical notes from researchers and practitioners active in the field of simulation. Within *JOS*, the scope of the field of simulation is taken to include the techniques, tools, methods and technologies of the application, and the use of discrete-event simulation in a wide range of domains including, for example, manufacturing, service, defence, health care and commerce.

MEETINGS

Special Interest Groups

A Special Interest Group of The OR Society is intended to serve some or all of the following objectives:

- to provide a focal point where members with a common interest in a particular OR area can meet together and exchange ideas and experience;
- to further technical development in a particular sphere;
- to promote a particular aspect of OR and to increase management understanding of the same;
- to help represent the OR Society viewpoint in joint activities with other Societies;
- to prepare written material on the “state of the art” for general publication;
- to provide education for new entrants to the profession or to the area covered by the Special Interest Group.

Groups are arranged for the following areas of special interest: Agriculture and Natural Resources, Community OR Network, Complex Systems Discussion Group, Criminal Justice, Decision Analysis, Defence, Financial Services, Forecasting, Health & Social Services, Independent Consultants’ Network, Information Systems, Local Search, Mathematical Programming, OR and Strategy, OR for Developing Countries, Problem Structuring Methods, Productivity Measurement, Simulation, and SD+.

PAST PRESIDENTS

Sir Owen Wansbrough-Jones	1954–1955
Sir William K Slater	1956–1957
Professor M G Kendall	1958–1959
The Earl of Halsbury	1960–1961
Professor B H P Rivett	1962–1963
Professor G A Barnard	1964–1965
Professor R T Eddison	1966–1967
Mr E C Williams	1968–1969
Mr S Beer	1970–1971
Professor K D Tocher	1972–1973
Mr R C Tomlinson	1974–1975
Mr A M Lee	1976–1977
Professor M G Simpson	1978–1979
Mr G H Mitchell	1980–1981
Professor K B Haley	1982–1983
Dr R S Stainton	1984–1985
Professor JV Rosenhead	1986–1987
Professor JV Rosenhead	1986–1987
Dr J C Ranyard	1988–1989
Mr P N Thornton	1990–1991
Professor C B Chapman	1992–1993
Professor L C Thomas	1994–1995
Mr I J Disley	1996–1997

4 THE OPERATIONAL RESEARCH SOCIETY

Professor R G Dyson	1998–1999
Professor M Pidd	2000–2001
Mr J Gibb	2002–2003
Professor V Belton	2004–2005
Professor J D Griffiths	2006–2007
Mrs S M Merchant	2008–2009
Professor R Eglese	2010–2011

THE OPERATIONS RESEARCH SOCIETY OF SOUTH AFRICA

HANS W. ITTMANN
CSIR Built Environment, Pretoria,
South Africa

V. S. SARMA YADAVALLI
Department of Industrial and Systems
Engineering, University of Pretoria,
Pretoria, South Africa

HISTORY OF ORSSA

The Operations Research Society of South Africa (ORSSA) was formally founded on Thursday 20 November 1969 in Johannesburg [1]. Some 150 individuals, from all over South Africa and also the then Rhodesia, were present at this founding meeting. The society is one of only a few OR societies currently active on the African continent, in fact, for many years it was the only one on the continent. It is interesting to note that the involvement of South Africans in OR goes back to World War II. South African scientists were members of the “Blackett’s Circus”, for example, Professors Solly Zuckerman and Frank Nabarro [1]. In addition Dr Basil Schonland, professor in geophysics at the University of the Witwatersrand, who was initially responsible for the South African efforts around the development and application of radar, later became the superintendent of the British Army Operational Research Group [2].

The use and practice of OR in South Africa has its roots in the mining industry [3] and in this regard, the Operational Research Bureau was founded by Herbert Sichel in 1952, mainly to act as consultants to the mining industry [4]. Various other large organizations embraced OR in the 1950s and 1960s in South Africa including the Defense Force, the main rail operator, and also the Council for Scientific and Industrial Research (CSIR), the largest public research establishment in the country. A South African,

RR Tusenius, attended the first OR conference held in Oxford in September 1957 [5]. With this growing interest in OR it was inevitable that this would lead to the establishment of a formal society, which happened in 1969. One of the biggest supporters and long-standing friends of the new society was Professor Patrick Rivett, a well-known, international OR personality from the University of Sussex in the United Kingdom. Rivett was present at the inaugural meeting in 1969 [1].

As indicated above, the mining industry was a fertile ground for OR applications in the early years of OR in South Africa. Whereas mining companies had strong OR groups, these all disappeared during the late 1970s and early 1980s when mining companies came under severe financial pressures. The development of OR was given a boost in South Africa when the CSIR established an OR group in the early 1960s. Many of the members of this initial group went to different universities where they were instrumental in establishing the first formal academic programs in South Africa [6]. The group at the CSIR continues to be one of the strongest OR groups in the country. Over the years, their work has involved more of consultations. OR also spread to industry, commerce, and defense. It is difficult to pinpoint a very strong sector where OR was used or is being used because of the wide variety of areas of application. A group in one of the banks was dominant at one stage; energy modeling was conducted during the 1970s and early 1980s by the CSIR together with the University of Pretoria; it is a well-known fact that in the mid-1970s a very strong group was established to conduct defense OR although little was known about what exactly the group was doing. Today there are still defense OR groups, possibly smaller than before democratization. Another noticeable area of application and a feature for which South Africa is famous, is wildlife management, including the conservation of the wildlife heritage and ecological and game range management. Considerable and

meaningful work was conducted in these areas [3,7,8]. Over the past 10 years a very strong modeling and simulation group has been developing at SASOL, the largest chemical company in South Africa, which is well known for the production of gasoline from coal. Today, this group is possibly the only such group in industry and by far the largest. A disappointing feature of OR in South Africa is the lack of OR usage in the public sector. The reasons are many but there is a general lack of appreciation of the value and potential contribution of quantitative methods and scientific problem solving.

In the academic environment, the University of South Africa (UNISA) was the first to establish a department that focused on traditional OR. This happened in the early 1970s and was by far the strongest OR university department for many years. The department has remained strong but the focus has shifted more toward financial modeling. One can find noticeable teaching in OR at various universities, namely, Stellenbosch University, the University of Cape Town (UCT), the University of the Witwatersrand (WITS), the North-West University, and also the University of Pretoria. One or two of the “previously disadvantaged institutions”, or Historically Black Universities, have relative large OR departments. In all cases except one, the departments that currently teach OR have different names such as decision sciences, statistical sciences or industrial engineering. At Stellenbosch University OR is located in the Department of Logistics. The teaching of OR at all these universities has always been dependent on the availability of suitable lecturing staff.

ORSSA exists primarily to further the interests of those engaged in, or interested in, OR activities. To achieve these objectives, it is involved in matters that concern OR practitioners in general, such as drawing up guidelines for OR education, presenting short courses and marketing OR, and providing information to the public on the nature of and career opportunities in OR. Ever since its inception, the society has had a variety of activities, initially mainly through three chapters and an annual conference, but these have been extended over the years.

Past and current activities of the society are the following:

- An annual conference is held in different parts of the country every year, usually in September of a year. The conference stretches typically over two or three days with around 70 to 80 delegates nowadays. In addition, the society endeavors to invite an international OR expert as the keynote speaker. In 1996, the annual conference was held in a neighboring country, namely, Swaziland. Joint conferences have also been held in the past with the Statistical Association of South Africa as well as the Institute for Industrial Engineers. In 2007 ORSSA also hosted the Operations Research Practise in Africa (ORPA) conference in Cape Town jointly with its national conference.
- The main mechanism of communicating with members is through a newsletter, which is published quarterly in printed format. The appearance of the newsletter has changed over the years and it is now a very professional looking newsletter.
- The society initially established three chapters and over the years, these have been extended to cover the entire country. Currently there is local representation in the form of five local chapters located in Cape Town, Johannesburg, Pretoria and two regions, namely, KwaZulu-Natal and the Vaal Triangle. All these chapters are active in one way or another.
- One of the founding members of the society, Tom Rozwadowski, died tragically with his family in 1971. In his honor, the society established the Tom Rozwadowski award, which is awarded annually for the best paper published by a member of the society.
- The society also runs a student competition annually where the best Honors and MSc students in the past year are recognized by the society in the form of a monetary prize and a certificate.
- In 1981 the first international meeting held in South Africa was a joint

- Israeli–South African conference on OR. It was held at the CSIR Conference Center in Pretoria and attracted nearly 200 delegates.
- The second international conference on “Operations Research in Resources and Requirements in Southern Africa” was held in 1984, also at the CSIR in Pretoria. This was an IFORS-sponsored event that included nine prominent international speakers.
 - In 1985 ORSSA established its own bi-annually published journal, ORiON. The journal has developed over the years to a quality publication with a high standard and an international advisory board. Volume 25 Number 1 has just been published.
 - A very successful Multi-Criteria Decision Making (MCDM) conference was held in January 1997 in Cape Town. Some 170 people, mainly international delegates, attended the conference.
 - The fourth International Conference on OR for Development (ICORD) was organized by ORSSA in May 2001 in the Kruger National Park. People from 14 countries attended. Although it was a very small conference it was nevertheless, a very successful event.
 - ORSSA initiated a project “OR into Africa” that endeavors to establish a community of OR workers in the development arena within South Africa.
 - In addition, ORSSA was also involved in assisting the East African OR fraternity to organize its first OR conference and to establish its own society.
 - Arguably the highlight of ORSSA’s existence was when the society’s bid to host the 18th triennial conference of the International Federation of Operations Research Societies (IFORS) was accepted by IFORS. This conference was held in July 2008 at the International Convention Center in Sandton and was, by all accounts, a very successful conference.
 - ORSSA has its own website at the following address: www.orssa.org.za.
 - The society is managed by an Executive Committee that *inter alia* consists of a President, who serves for two years, a Vice President, Secretary, Treasurer, two additional members, the editors of the newsletter and journal, respectively, as well as the business editor of the journal. They meet on a regular basis to conduct the business of ORSSA.
 - A number of honorary memberships of ORSSA have been awarded over the years and include Herbert Sichel, Pat Rivett, Gerhard Rudolph, Jos Grobelaar, Gerhard Geldenhuys, and Mike Splaine. The society also has a retired member category. At the conference in 2003 ORSSA introduced another form of recognition, ORSSA Fellows, for long-standing members and members who have over the years contributed significantly to OR in the country as well as served ORSSA in various ways.
 - ORSSA joined IFORS in 1973 while it is also a member of the Association of European Operational Research Societies (EURO) since the latter is the “closest” regional grouping of IFORS that ORSSA can belong to, given that there is no African grouping.
- All the activities of ORSSA are conducted through voluntary service of members. The membership of ORSSA is relatively small and although the number fluctuates, there are around 300 members mainly in South Africa but also from other countries. It is however, true that there are many more people practicing OR in South Africa but many of them tend to associate themselves with their functional areas of work such as industrial engineering, finance, supply-chain management, information systems, etc.
- A few additional interesting aspects of the society are firstly that the society’s constitution drafted at its inception included specific prohibitions against any discrimination on the grounds of race, gender or creed. This was done at the height of the institutionalized apartheid era [6]. Secondly, members of ORSSA have been involved in international activities, for example, one member was a Vice President of IFORS (Theo Stewart),

an ORSSA member was the editor of the IFORS newsletter for developing countries and is also the current editor of the IFORS newsletter (Hans Ittmann). The chairperson of the adjudication committee for the IFORS prize for developing countries during the past two IFORS conferences is an ORSSA member (Paul Fatti) while a South African entry to this prize has been runner-up on three occasions. One of the more recent Presidents of the International Multi-Criteria Decision Making Society is a member of ORSSA (Theo Stewart). Members from ORSSA regularly contribute to IFORS and EURO conferences although the numbers are small. There is also some involvement with working groups of EURO. The majority of ORSSA members who publish tend to publish in ORiON, the local journal, although a limited number of papers are published in international OR journals. A few books authored by ORSSA members have been published over the years. Another highlight of the South African society was when an entry from the South African National Defence Force (SANDF) won the prestigious Franz Edelman award in 1996. This was for excellent work on “determining the size and shape of the SANDF”, given the absence of a conventional military threat [9]. Support from President Nelson Mandela, then President of the country, in the form of an accompanying letter, cannot be underestimated!

From the above, it is clear that ORSSA is an active society. It is struggling with problems similar to those elsewhere in the world, where it is difficult to retain the more junior members and those finishing at universities.

Presidents of ORSSA to date are listed as follows: Dr H Sichel (1970), Mr JW Grobbelaar (1971), Mr D Masterson (1972), Dr G Rudolph (1973), Mr J Miller (1974),

Mr K Sandrock (1975), Mr HID du Plessis (1976), Mr R Eales (1977), Prof TJ Stewart (1978), Prof JS Wolvaardt (1979), Dr MJ Venter (1980), Dr AH Money (1981), Mr M Splaine (1982), Mr D Masterson (1983), Prof LP Fatti (1984), Mr DW Evans (1985), Mr MA de Vries (1986), Mr HW Ittmann (1987), Prof G Erens (1988–89), Ms A Pachyannis (1990–91), Mr LA Visagie (1992–93), Mrs E Ferreira (1994–95), Dr E Dixon (1996–97), Prof JW Hearne (1998–99), Dr PduT Fourie (2000–01), Mr HW Ittmann (2002–03), Prof WR Gevers (2004–05), Mrs MF Harmse (2006–07), Prof VSS Yadavalli (2008–09).

REFERENCES

1. Geldenhuys G, Rudolph G. A brief history of the beginnings of operations research in South Africa; 2006. www.orssa.org.za.
2. Austin B. Schonland: scientist and soldier. Johannesburg: Institute of Physics Publications and Wits University Press; 2002.
3. Fatti LP. Current practice of operational research/management science in South Africa. *Omega* 1988;16(3):181–187.
4. Sichel HS. Operational research in South Africa. *Bull Int Stats Inst* 1957;35:391–392, 394.
5. Page T. First International Conference on operations research; 1957 Sept 2–5; Oxford, England. *Oper Res* 1957;5:863–871.
6. Stewart TJ. OR practice in South Africa. *Eur J Radiol* 1995;87:464–468.
7. Hearne J. African OR adventure. *OR/MS Today* 1994;21:50–53.
8. Hearne J. Hunting for Profits. *OR/MS Today* 1999;26:40–43.
9. Botha S, Gryffenberg I, Hofmeyr FR, *et al*. Guns or butter: decision support for determining the size and shape of the South African National Defense Force. *Interfaces* 1997;27:7–28.

THE SCATTER SEARCH METHODOLOGY

MANUEL LAGUNA
Leeds School of Business, University
of Colorado at Boulder

RAFAEL MARTÍ
Departamento de Estadística e
Investigación Operativa,
Universidad de Valencia, Spain

MICAEL GALLEGO
ABRAHAM DUARTE
Departamento de Ciencias de la
Computación, Universidad Rey
Juan Carlos, Spain

INTRODUCTION

Scatter search (SS) is an evolutionary method that has been successfully applied to hard optimization problems. SS was first introduced by Prof. Fred Glover as a heuristic for integer programming and it was based on strategies presented at a management science and engineering management conference held in Austin, Texas in September 1967. Unlike genetic algorithms, it operates on a small set of solutions and employs diversification strategies of the form proposed in tabu search [1], which give precedence to strategic learning based on adaptive memory, with limited recourse to randomization.

The fundamental concepts and principles were first proposed in 1977 [2] as an extension of formulations, dating back to the 1960s, for combining decision rules and problem constraints. (The constraint combination approaches, known as *surrogate constraint* methods, now independently provide an important class of relaxation strategies for global optimization.) The SS framework is flexible, allowing the development of alternative implementations with varying degrees of sophistication. In addition, SS can be combined with other methods to enhance their

effectiveness. For example, SS has been integrated with particle swarm optimization and adaptive memory strategies of tabu search to produce a method called *cyber swarm optimization* that has provided improvements over its particle swarm component [3].

As described in Refs 4 and 5, SS consists of five methods:

1. Diversification Generation
2. Improvement
3. Reference Set Update
4. Subset Generation
5. Solution Combination

The *diversification generation method* is used to generate a set of diverse solutions that are the basis for initializing the search. The most effective diversification methods are those capable of creating a set of solutions that balances diversification and quality. It has been shown that SS produces better results when the diversification generation method is not purely random and constructs solutions by reference to both a diversification measure and the objective function.

The *improvement method* transforms solutions with the goal of improving quality (typically measured by the objective function value) or feasibility (typically measured by some degree of constraint violation). The input to the improvement method is a single solution that may or may not be feasible. The output is a solution that may or may not be better (in terms of quality or feasibility) than the original solution. The typical improvement method is a local search (LS) with the usual rule of stopping as soon as no improvement is detected in the neighborhood of the current solution. There is the possibility of basing the improvement method on procedures that use a neighborhood search but that they are able to escape local optimality. tabu search, simulated annealing, and variable neighborhood search qualify as candidates for such a design. This may seem

as an attractive option as a general approach for an improvement method, however, these procedures do not have a natural stopping criterion. The end result is that choices need to be made to control the amount of computer time that is spent improving solutions (by running a metaheuristic-based procedure) versus the time spent outside the improvement method (e.g., combining solutions). In general, LS procedures seem to work well and most SS implementations do not include mechanisms to escape local optimality within the process of improving a solution.

The *reference set update* method refers to the process of building and maintaining a reference set of solutions that are used in the main iterative loop of any SS implementation. While there are several implementation options, this element of SS is fairly independent from the context of the problem. The first goal of the reference update method is to build the initial reference set of solutions from the population of solutions generated with the diversification method. Subsequent calls to the reference update method serve the purpose of maintaining the reference set. The typical design of this method builds the first reference set by blending high quality solutions and diverse solutions. While choosing diverse solution, reference needs to be made to a distance metric that typically depends on the solution representation. That is, if the problem context is such that continuous variables are used to represent solutions, then diversification may be measured with Euclidean distances. Other solution representations (e.g., binary variables or permutations) result in different ways of calculating distances and in turn diversification. The updating of the reference set during the SS iterations is customarily done on the basis of solution quality.

The *subset generation method* produces subsets of reference solutions which become the input to the combination method. The typical implementation of this method consists of generating all possible pairs of solutions. The SS framework considers also the generation of larger subsets of reference solutions; however, most SS implementations have been limited to operate on pairs of solutions. Clearly, no context information is

needed to implement the subset generation method.

The *solution combination method* uses the output from the subset generation method to create new solutions. New trial solutions are the results of combining, typically two but possibly more, reference solutions. The combination of reference solutions is usually designed to exploit problem context information and solution representation. The strategic use of linear combinations of solutions in heuristic search has often been employed for discrete and nonlinear optimization problems, since such uses were first proposed in Ref. 2. Several proposals for combining solutions represented by permutations have also been applied [6]. The strategy known as *path relinking*, originally proposed within the tabu search methodology [1], has also played a relevant role in designing combination methods for SS implementations.

As mentioned above, we will describe both the main SS elements and the advanced search strategies, and use the maximum diversity problem (MDP), [7] to illustrate some implementation details. The MDP consists of selecting a subset of m elements from a set of n elements in such a way that the sum of the distances between the chosen elements is maximized. The definition of distance between elements is customized to specific applications. As mentioned in Refs 8 and 9, the MDP has applications in plant breeding, social problems, ecological preservation, pollution control, product design, capital investment, workforce management, curriculum design, and genetic engineering. In most applications, it is assumed that each element can be represented by a set of attributes. Let s_{ik} be the state or value of the k th attribute of element i , where $k = 1, \dots, K$. Then the distance between elements i and j may be defined as

$$d_{ij} = \sqrt{\sum_{k=1}^K (s_{ik} - s_{jk})^2}.$$

In this case, d_{ij} is simply the Euclidean distance between i and j . The distance values are then used to formulate the MDP as a quadratic binary problem, where variable x_i

```

1. Diversification generation and improvement methods
2. while (stopping criteria not satisfied) {
3.     Reference set update method
4.     while (new reference solutions) {
5.         Subset generation method
6.         Combination method
7.         Improvement method
8.         Reference set update method
9.     }
10.    Rebuild reference set
11. }

```

Figure 1. Scatter search framework.

takes the value 1 if element i is selected and 0 otherwise, $i = 1, \dots, n$

$$\begin{aligned}
 &\text{Maximize} && \sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij} x_i x_j \\
 &\text{Subject to} && \sum_{i=1}^n x_i = m \\
 &x_i = \{0, 1\} && 1 \leq i \leq n.
 \end{aligned}$$

We finally would like to highlight the SS library provided in Ref. 5. It consists of a C code that implements the search mechanisms that are discussed throughout their book. The C code can be used “as is” to replicate practical illustrations or can be modified and adapted to other optimization problems of interest. Several SS applications have been implemented from that source code.

BASIC SS DESIGN

Laguna [10] summarizes the basic SS framework as follows (Fig. 1). The search starts with the application of the diversification and improvement methods (step 1 in Fig. 1). The typical outcome consists of a set of about 100 solutions that is referred to as the *population* (denoted by P). In most implementations, the diversification generation method is applied first followed by the improvement method. If the application of the improvement method results in the shrinking of the population (due to more than one solution converging to the same local optimum) then the diversification method is applied again until the total number of improved solutions reaches the desired

target. Other implementations construct and improve solutions, one by one, until reaching the desired population size.

The main SS loop is shown in lines 2 to 11 of Fig. 1. The input to the first execution of the reference set update method (step 3) is the population of solutions generated in step 1 and the output is a set of solutions known as the *reference set* (or *RefSet*). Typically, 10 solutions are chosen from a population of 100. The first five solutions are chosen to be the best solutions (in terms of the objective function value) in the population. The other five are chosen to be the most diverse with respect to the solutions in the reference set. If the diverse solutions are chosen sequentially, then the sixth solution is the most diverse with respect to the five best solutions that were chosen first. The seventh solution is the most diverse with respect to the first six, and so on until the tenth one is added to the reference set.

The inner while-loop (lines 4 to 9) is executed as long as at least one reference solution is new in the *RefSet*. A solution is considered new if it has not been subjected to the subset generation (step 5) and combination (step 6) methods. If the reference set contains at least one new solution, the subset generation method builds a list of all the reference solution subsets that will become the input to the combination method. The subset generation method creates new subsets only. A subset is new if it contains at least one new reference solution. This avoids the application of the combination method to the same subset more than once,

which is particularly wasteful when the combination method is completely deterministic. Combination methods that contain random elements may be able to produce new trial solutions even when applied more than once to the same subset of reference solutions. However, this is generally discouraged in favor of introducing new solutions in the reference set by replacing some of the old ones in the rebuilding step (line 10).

The combination method (step 6) is applied to the subsets of reference solutions generated in the previous step. Most combination methods are designed to produce more than one trial solution from the combination of the solutions in a subset. These trial solutions are given to the improvement method (step 7) and the output forms a pool of improved trial solutions that will be considered for admission in the reference set (step 8).

If no new solutions are added to the reference set after the execution of the reference set update method, then the process exits the inner while-loop. The rebuilding step in line 10 is optional. That is, it is possible to implement a SS procedure that terminates the first time that the reference set does not change. However, most implementations extend the search beyond this point by executing a *RefSet* rebuilding step. The rebuilding of the reference set entails the elimination of some current reference solutions and the addition of diverse solutions. In most implementations, all solutions except the best are replaced in this step. The diverse solutions to be added may be either population solutions that have not been used or new solutions constructed with the generation diversification method. Note that only 10 solutions out of 100 are used from the population to build the initial reference set and therefore the remaining 90 could be used for rebuilding purposes.

The process (i.e., the main while-loop in lines 2 to 11) continues as long as the stopping criteria are not satisfied. Possible stopping criteria include number of rebuilding steps or elapsed time. When SS is applied in the context of optimizing expensive black boxes, a limit on the number of calls to the objective function evaluator (i.e., the black box) may also be used as a criterion for stopping.

We now expand our description and provide examples of each of the SS methods.

THE MAXIMUM DIVERSITY PROBLEM

Three of the five SS elements (the diversification-generation, the improvement, and the combination methods) are *problem-dependent* and should be designed specifically for the problem at hand (although it is possible to design “generic” procedures, it is more effective to base the design on the specific idiosyncrasies of the problem setting). The other two, the reference-set-update and the subset-generation methods, are *problem-independent*, and usually follow a standard implementation. In this section, we describe efficient implementations of the three problem-dependent methods for the MDP. In the next section, we cover the two problem-independent methods.

The definition of distance between solutions is a key design issue in SS implementations. Distance is used to measure how diverse one solution is with respect to a set of solutions. Specifically, for the MDP, let x_i^r be the value of the i th variable for the reference solution r (i.e., $r \in \text{RefSet}$). Also let x_i^t be the value of the i th variable for the trial solution t . Then, the distance between the trial solution t and the solutions in the *RefSet* in our SS implementation is defined as

$$\text{distance}(t, \text{RefSet}) = bm - \sum_{r=1}^b \sum_{i: x_i^t = 1} x_i^r.$$

The formula simply counts the number of times that each selected element in the trial solution t appears in the reference solutions and subtracts this value from the maximum possible distance (i.e., bm). The maximum distance occurs when no element that is selected in the trial solution t appears in any of the reference solutions. When choosing solutions to rebuild the reference set, we select the trial solution that has the maximum distance between itself and the solutions currently in the *RefSet*. Since the solutions are added one at a time, the distance calculations have to be updated before the next solution is selected.

Diversification Generation Method

Duarte and Martí [11] proposed a GRASP Hybrid [12] for generating MDP diverse solutions. It is based on randomizing *D*-2, a deterministic destructive heuristic developed by Glover *et al.* [9]. *D*-2 starts with the infeasible solution for which $x_i = 1$ for all i . That is, all n elements are originally selected. In order to reduce the set of selected elements to m , the procedure performs $n-m$ steps. At each step, the procedure deselects element i^* (i.e., x_{i^*} is set to 0), where i^* is such that

$$D(i^*) = \text{Min}_{i: x_i=1} (D(i)),$$

where $D(i) = \sum_j d_{ij}x_j$.

The randomization of *D*-2 that is employed within *RD*-2 consists of selecting i^* from a reduced candidate list formed by all those elements i such that $D(i) \leq (1 + \alpha)D(i^*)$. The value of α is initially set to 0.5 and decreased by 0.1—to a minimum of 0.1—after a prespecified CPU time is consumed without improving the incumbent. We set this value to a 20% of the total CPU time in our implementation.

Improvement Method

The LS method [13] scans the set of selected elements in search of the best exchange to replace a selected element with an unselected one. The method performs moves as long as the objective value increases and it stops when no improving exchange can be found. The improved local search method, *I_LS*, [11] selects the element i^* ($x_{i^*} = 1$) that provides the smallest contribution to the objective function value of the current solution. Then, it searches for an element j ($x_j = 0$) to be exchange with element i^* . The first element j that results in an improving move is selected and the exchange is performed without examining the remaining unselected elements. If no improving move can be found to exchange element i^* , then the selected element with the next smallest contribution is examined. This process continues until no improving exchange can be found.

Combination Method

This method consists of the application of the destructive heuristic *D*-2 [9] to the union

of the elements in the reference solutions being combined. The method starts with the selection of all elements in the union and then it deselects one element at a time until there are only m selected elements remaining. The element i that is deselected at each step is the one with the minimum $D(i)$ value.

PROBLEM-INDEPENDENT METHODS

We now discuss some generic (i.e., context-independent) implementations of the reference set update and the subset generation. While these designs are not customized for a particular problem setting, the solution representation (e.g., binary strings, permutations, real numbers, etc.) does play an important role in their development and implementation. The following two sections have been taken from Ref. 10.

Reference Set Update

The execution of the diversification generation and improvement methods in line 1 of Fig. 1 results in a population P of solutions. The reference set update method is executed in two different parts of the SS procedure. The first time that the method is called, its goal is to produce the initial *RefSet*, consisting of a mix of b (typically 10) high quality and diverse solutions drawn from P . The mix of high quality and diverse solutions could be considered a tunable parameter; however, most implementations populate the initial reference set with half of the solutions chosen by quality and half chosen by diversity.

Choosing solutions by quality is straightforward. From the population P , the best (according to the objective function value) $b/2$ solutions are chosen. These solutions are added to *RefSet* and deleted from P . To choose the remaining half, an appropriate measure of distance $d(r, p)$ is needed, where r is a reference solution and p is a solution in P . The distance measure depends on the solution representation, but must satisfy the usual conditions for a metric, that is

$d(r, p) = 0$	If and only if $r = p$
$d(r, p) = d(p, r) > 0$	If $r \neq p$
$d(r, q) + d(q, p) \geq d(r, p)$	Triangle inequality

For instance, the Euclidean distance is commonly used when solutions are represented by continuous variables

$$d(r, p) = \sqrt{\sum_{i=1}^n (r_i - p_i)^2}.$$

Likewise, the [14] distance is appropriate for two strings of equal length. The distance is given by the number of positions for which the corresponding symbols are different. In other words, the distance is the number of changes required to transform one string into the other

$$d(r, p) = \sum_{i=1}^n v_i \quad v_i = \begin{cases} 0 & r_i = p_i \\ 1 & r_i \neq p_i \end{cases}. \quad (1)$$

This distance has been typically used for problems whose solution representation is given by binary vectors (e.g., the max-cut and knapsack problems) and permutation vectors (e.g., the traveling salesman, quadratic assignment, and linear ordering problems).

The distance measure is used to calculate the minimum distance $d_{\min}(p)$ of a population solution and all the reference solutions r

$$d_{\min}(p) = \min_{r \in \text{RefSet}} d(r, p).$$

Then, the next population solution p to be added to *RefSet* and deleted from P is the one with the maximum $d_{\min}$ value. That is, we want to choose the population solution p^* that has the maximum minimum distance between itself and all the solutions currently in *RefSet*

$$p^* = \left\{ p : \max_{p \in P} d_{\min}(p) \right\}.$$

The process is repeated $b/2$ times to complete the construction of the initial *RefSet*. Note that after the first calculation of the $d_{\min}$ values, they can be updated as follows. Let

p^* be the population solution most recently added to the *RefSet*. Then, the $d_{\min}$ value for a population solution p is given by

$$d_{\min}(p) = \min(d_{\min}(p), d(p, p^*)).$$

Alternatively, the diverse solutions for the initial reference set may be chosen by solving the MDP introduced above. The special version of the MDP that must be solved includes a set of elements that have already been chosen (i.e., the high quality solutions). Mathematically, the problem may be formulated as follows:

$$\begin{aligned} & \text{Maximize} && \sum_{(p, p^*) \in P: p \neq p'} d(p, p') x_p x_{p'} \\ & \text{Subject to} && \sum_{p \in P} x_p = b \\ & && x_p = 1 \quad \forall p \in \text{RefSet} \\ & && x_p \in \{0, 1\} \quad \forall p \in P. \end{aligned}$$

The formulation assumes that a subset of high quality solutions in P have already been chosen and added to *RefSet*. The binary variables indicate whether a population solution p is chosen ($x_p = 1$) or not ($x_p = 0$). The second set of constraints in the formulation force the high-quality solutions to be included in the set of b solutions that will become the initial *RefSet*. This nonlinear programming model has been translated into an integer program for the purpose of solving it as well as for showing that the MDP is NP-hard. Martí *et al.* [15] embed the MDP in a SS procedure for the max-cut problem. Instead of solving the MDP exactly, they employ the GRASP_C2 procedure developed by Duarte and Martí [11]. The procedure was modified to account for the high-quality solutions that are chosen before the subset of diverse solutions is added to the *RefSet*. The procedure is executed for 100 iterations and the most diverse *RefSet* is chosen to initiate the SS.

The reference set update method is also called in step 8 of Fig. 1. This step is performed after a set of one or more trial solutions has been generated by the sequential calls to the subset generation, combination,

and improvement methods (see steps 5, 6, and 7 in Fig. 1). The most common update consists of the selection of the best (according to the objective function value) solutions from the union of the reference set and the set of trial solutions generated by steps 5–7 in Fig. 1. Other updates have been suggested in order to preserve certain amount of diversity in the *RefSet*. These advanced updating mechanisms are beyond the scope of this tutorial article but the interested reader is referred to Chapter 5 of Ref. 5 for a detailed description and to Ref. 16 for experimental results.

Subset Generation

This method is in charge of proving the input to the combination method. This input consists of a list of subsets of reference solutions. The most common subset generation consists of creating a list of all pairs (i.e., all 2-subsets) of reference solutions for which at least one of the solutions is new. A reference solution is new if it has not been used by the combination method. The first time that this method is called (step 5 in Fig. 1), all the reference solutions are new, given that the method is operating on the initial *RefSet*. Therefore, the execution of the subset generation method results in the list of all 2-subsets of reference solutions, consisting of a total of $(b^2 - b)/2$ pairs. Because the subset generation method is not based on a sample but rather on the universe of all possible pairs, the size of the *RefSet* in SS implementation must be moderate (e.g., less than 20). As mentioned in the introduction, a typical value for b is 10, resulting in 45 pairs the first time that the subset generation method is executed.

When the inner while-loop (steps 5–8 in Fig. 1) is executing, the number of new reference solutions at the time that the subset generation method is called depends on the strategies implemented in the reference set update method. Nonetheless, the number of new solutions decreases with the number of iterations within the inner while-loop. Suppose that b is set to 10 and that, after the first iteration of the inner while-loop, six solutions are replaced in the reference set. This means that the reference set that will serve

as the input to the subset generation method will consist of four “old” solutions and six “new” solutions. Then, the output of the subset generation method will be 39 2-subsets. In general, if the reference set contains n new solutions and m old ones, the number of 2-subsets that the subset generation method produces is given by

$$nm + \frac{n^2 - n}{2}.$$

The SS methodology also considers the generation of subsets with more than two elements for the purpose of combining reference solutions. As described in Ref. 5, the procedure uses a strategy to expand pairs into subsets of larger size while controlling the total number of subsets to be generated. In other words, the mechanism does not attempt to create all 2-subsets, then all 3-subsets, and so on, until reaching the $b-1$ -subsets and finally the entire *RefSet*. This approach would not be practical because there are 1013 subsets in a reference set of size $b = 10$. Even for a smaller reference set, combining all possible subsets would not be effective because many subsets will be very similar. For example, a subset of size four containing solutions 1, 2, 3, and 4 is almost the same as all the subsets with four solutions for which the first three solutions are solutions 1, 2, and 3. And even if the combination of subset {1, 2, 3, 5} would generate a different solution than the combination of subset {1, 2, 3, 6}, these new trial solutions would likely reside in the same basin of attraction and therefore converge to the same local optimum after the application of the improvement method. Instead, the approach selects representative subsets of different sizes by creating subset types:

- *Subset Type*. all 2-element subsets;
- *Subset Type*. 3-element subsets derived from the 2-element subsets by augmenting each 2-element subset to include the best solution not in this subset;
- *Subset Type*. 4-element subsets derived from the 3-element subsets by augmenting each 3-element subset to include the best solutions not in this subset;

- *Subset Type*. the subsets consisting of the best i elements, for $i = 5$ to b .

Campos *et al.* [17] designed an experiment with the goal of assessing the contribution of combining subset types 1 to 4 in the context of the linear ordering problem. The experiment undertook to identify how often, across a set of benchmark problems, the best solutions came from combinations of reference solution subsets of various sizes. The experimental results showed that most of the contribution (measured as the percentage of time that the best solutions came from a particular subset type) could be attributed to subset type 1. It was acknowledge, however, that the results could change if the subset types were generated in a different sequence. Nonetheless, the experiments indicate that the basic SS that employs only subsets of type 1 is quite effective and explains why most implementations do not use subset types of higher dimensions.

HYBRIDIZING SCATTER SEARCH WITH TABU SEARCH

Evolutionary approaches, such as SS, implicitly make use of memory. This is evident if one examines how the reference set update, solution combination, and subset generation methods operate. The reference set update method, in its most basic form, is designed to “remember” the best solutions encountered during the search. Some features of these solutions are used to create new trial solutions with the combination method. Hence, this method is instrumental in the transmission of information embedded in the reference solutions. But we can add extra memory structures to SS by hybridizing it with tabu search. Specifically, in this section, we briefly describe how to design tabu search based procedures to implement the diversification generation, the improvement and the combination methods for the MDP.

Diversification Generation Method

The diversification generator within the tabu search methodology, *MD-2*, is also based on

the destructive procedure *D-2*. At each step of the procedure, the element i^* to be deselected is such that

$$D(i^*) = \min_{i:x_i=1} \left(D(i) - \beta(range) \left(\frac{f(i)}{f_{\max}} \right) + \delta(range) \left(\frac{q(i)}{q_{\max}} \right) \right),$$

where

$$range = \max_{i:x_i=1} (D(i)) - \min_{i:x_i=1} (D(i)).$$

In this modified distance calculation, $f(i)$ indicates the frequency in which element i has appeared in previous solutions and $q(i)$ is the average quality (as measured by the objective function value) of past solutions that included element i . The $f_{\max}$ and $q_{\max}$ are the maximum values of f and q over all elements. The penalty factors β and δ are respectively set to 0.1 and 0.0001 according to Gallego *et al.* [7].

Improvement Method

LS_TS [11] implements a short-term tabu search method also based on exchanges. An iteration of this method begins with a random selection of an element i ($x_i = 1$). The probability of selecting element i is inversely proportional to $D(i)$. The list of unselected elements is scanned and the first improving move that exchanges elements i and j ($x_j = 0$) is selected. If no improving move is found, then the least nonimproving move is chosen. The chosen exchange is performed and both elements participating in the exchange are classified tabu-active for a number of iterations (known as the *tabu tenure*). Tabu-active elements are not allowed to participate in any exchanges. The LS_TS method stops if after a number of consecutive iterations the incumbent solution is not modified.

Combination Method

This method consists of the application of the *MD-2* procedure to the union of the elements of the reference solutions being

combined. This method uses information about solutions generated in the past as well as information associated with those solutions combined in previous iterations. Once we obtain the set of elements selected in the solutions being combined, we only need to apply *MD-2* with the modified evaluations according to the penalty terms

$$\beta(\text{range}) \left(\frac{f(i)}{f_{\max}} \right) \text{ and } \delta(\text{range}) \left(\frac{q(i)}{q_{\max}} \right).$$

The element with minimum D value is deselected. The union of elements updated, iteration are performed until the number of selected elements matches m .

- SOM** This data set consists of 20 matrices with random numbers between 0 and 9 generated from an integer uniform distribution. The problem sizes are such that for $n = 100, m = 10, 20, 30$, and 40; for $n = 200, m = 20, 40, 60$, and 80; for $n = 300, m = 30, 60, 90$, and 120; for $n = 400, m = 40, 80, 120$, and 160; and for $n = 500, m = 50, 100, 150$, and 200. These instances were generated by Silva *et al.* [18].
- GKD** This data set consists of 20 matrices for which the values were calculated as the Euclidean distances from randomly generated points with coordinates in the 0 to 10 range. The number of coordinates for each point is also randomly generated between 2 and 21. Glover *et al.* [9] developed this data generator and constructed instances with $n = 30$. We generated instances with $n = 500$ and $m = 50$.
- Type I** Diversity values are real numbers uniformly distributed between 0 and 1000. We generate 30 instances of Type I with $n = 500, 2000$ and $m = 50, 200$.
- Type II** Diversity values are real numbers uniformly distributed between 0 and 10. We generate 20 instances of Type II, as described in the section titled “Basic SS Design,” with $n = 500$ and $m = 50$.

The purpose of the first experiment is identifying the most effective method for generating subsets of reference solutions that are in turn the input to the combination method. For this experiment, we consider combinations of 2, 3, 4, and 5 solutions. Our subset generation method operates as described in the section titled “Problem-Independent Methods.” All subsets of size 2 (SG1) are considered. That is, all pairs of reference solutions are added to the list of subsets. Subsets of size 3 (SG2) are constructed by considering each subset of size 2 and adding the best reference solution that is not part of

EXPERIMENTS WITH THE MAXIMUM DIVERSITY PROBLEM

In this section, we summarize some of the experiments presented in Ref. 7 for the MDP. The SS implementation follows the basic framework outlined in Fig. 1. The diversification generation, combination, and improvement methods are those corresponding to the descriptions in the section titled “The Maximum Diversity Problem.” All the experiments were conducted on a Pentium 4 computer at 3 GHz with 3 GB of RAM. The procedures were coded in Java and executed in the Java Runtime Environment 1.5. We use the same four data sets employed in Ref. 11.

the subset. Subsets of higher dimensions are constructed following the same logic. That is, subsets of size 4 (SG3) are based on subsets of size 3. Likewise, subsets of size 5 (SG4) are constructed by adding a solution to subsets of size 4. This mechanism avoids the exponential explosion in the number of subsets generated had we considered all possible subsets of size 3, 4, and 5. The results of running this experiment on 20 instances (10 problems of each type, I and II, and with $n = 500$ and $m = 50$) are summarized in Table 1. We use the outcomes of our experiments to calculate the average percent deviation (*average*

Table 1. Alternative Subset Generation Methods

Subset Generation Method	Average Deviation (%)	Number of Best
SG1	0.0000	20
SG2	0.0017	19
SG3	0.0000	20
SG4	0.0042	18

deviation) of the solutions obtained by each procedure when compared to the best solutions during the given experiments. We also report on the number of best solutions (*number of best*) found by each method.

The results of this preliminary experiment indicate that there is no additional gain that could be realized by generating and combining subsets with more than 2 solutions. Hence, we perform step 2 (Fig. 1) of the SS implementation by limiting the subset generation to all pairs of reference solutions. These results are in line with similar experiments for other combinatorial optimization problems [17].

The objective of the second experiment is to compare the relative merit of the SS variants described in the sections titled “The Maximum Diversity Problem” and “Hybridizing Scatter Search with Tabu Search,” respectively. We set a time limit of 3 CPU min and we run each procedure with and without the improvement method. The results are summarized in Table 2.

The results in Table 2 indicate the advantage of using an improvement method within our design. In terms of average deviation from the best solutions, the improvement method has the largest impact in the case of the base design. Also, the improvement method makes a significant difference in the number of best solutions found in the tabu search hybrid. In general, we conclude that the procedures embedded in the tabu search hybrid variant results in the best SS configuration. For the next experiment, the SS that we use is the one that implements the tabu search hybrid procedures (including the improvement method), as described in the section titled “Hybridizing Scatter Search with Tabu Search.”

In the final experiment, we compare the proposed procedure to state-of-the-art methods for solving the MDP. In particular we consider

- KLD [18] with LS [13]
- KLDv2 [18] with LS [13]
- Tabu_D–2 with LS_TS [11]

In this experiment, we observed the solution quality obtained by each method after 30 s and after 3 min of search time. Table 3 reports the average percentage deviation of the solution obtained with each method with respect to the best solution known.

Table 3 shows the merit of the proposed procedure. The SS implementation consistently produces the best solutions with percent deviations that in some cases are orders of magnitude smaller than those of the competing methods. The problem instances in the GKD set do not provide a way of differentiating the performance of the methods that we are comparing. They are either easy to solve and all the methods are capable of finding the optimal solutions in a very short period of time or the problems are difficult and all the methods are attracted to the same local optima.

CONCLUSIONS

Our goal was to introduce the SS framework at a level that would make it possible for the reader to implement a basic but robust procedure. A number of extensions are possible and some of them have already been explored and reported in the literature. It is not possible within the limited scope of this encyclopedic article to detail completely many of the aspects of SS that warrant further investigation. Additional implementation considerations, including those associated with intensification and diversification processes, and the design of accompanying methods to improve solutions produced by combination strategies are found in several of the references listed below.

Table 2. Different Scatter Search Designs

Procedure	Improvement	Average Deviation (%)	Number of Best
Scatter search	No	1.16	0
(section titled “The Maximum Problem”)	Yes	0.18	2
Tabu search hybrid	No	1.07	0
(section titled “Hybridizing Scatter Search with Tabu Search”)	Yes	0.00	20

Table 3. Comparison of Average Percent Deviation at Two Times During the Search

Data Set	Time	KLD (%)	KLDv2 (%)	Tabu_D-2 (%)	SS (%)
SOM	30 s	1.056	1.463	0.138	0.002
	3 min	0.178	0.187	0.095	0.000
GKD	30 s	0.000	0.000	0.000	0.000
	3 min	0.000	0.000	0.000	0.000
Type II	30 s	0.857	1.083	0.245	0.010
	3 min	0.525	0.607	0.203	0.000
Type I	30 s	9.807	100.000	0.453	0.453
	3 min	9.807	9.828	0.331	0.331

Acknowledgments

This research has been partially supported by the *Ministerio de Ciencia e Innovación* of Spain (TIN2009-07516).

REFERENCES

1. Glover F, Laguna M. Tabu search. Boston (MA): Kluwer Academic Publisher; 1997.
2. Glover F. Heuristics for integer programming using surrogate constraints. *Decis Sci* 1977; 8:156–166.
3. Yin PY, Glover F, Laguna M, *et al.* Cyber swarm algorithms—improving particle swarm optimization using adaptive memory strategies. *Eur J Oper Res* 2010;201(2): 377–389.
4. Glover F. A template for scatter search and path relinking. In: Hao J-K, Lutton E, Ronald E, *et al.*, editors. *Artificial evolution*, Lecture Notes in Computer Science 1363. Berlin: Springer; 1998. pp. 13–54.
5. Laguna M, Martí R. Scatter search: methodology and implementations in C. Boston (MA): Kluwer Academic Publishers; 2003.
6. Martí R, Laguna M, Campos V. Scatter Search vs. Genetic algorithms: an experimental evaluation with permutation problems. In: Rego C, Alidaee B, editors. *Metaheuristic optimization via adaptive memory and evolution: tabu search and scatter search*. Norwell (MA): Kluwer Academic Publishers; 2005. pp. 263–282.
7. Gallego M, Duarte A, Lagunay M, *et al.* Hybrid heuristics for the maximum diversity problem. *Comput Optim Appl* 2009;44(3):411–426.
8. Kuo CC, Glover F, Dhir KS. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decis Sci* 1993;24(6):1171–1185.
9. Glover F, Kuo CC, Dhir KS. Heuristic algorithms for the maximum diversity problem. *J Inf Optim Sci* 1998;19(1):109–132.
10. Laguna M. Scatter search. In: Burke EK, Kendall G, editors. *Search methodologies: introductory tutorials in optimization and decision support techniques*. 2nd ed. 2009. In press.
11. Duarte A, Martí R. Tabu search and GRASP for the maximum diversity problem. *Eur J Oper Res* 2007;178:71–84.
12. Resende MGC, Ribeiro CC. Greedy randomized adaptive search procedures. In: Glover F, Kochenberger G, editors. *State-of-the-art handbook in metaheuristics*. Boston (MA): Kluwer Academic Publishers; 2001. pp. 219–250.

13. Ghosh JB. Computational aspects of the maximum diversity problem. *Oper Res Lett* 1996;19:175–181.
14. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J* 1950;26(2): 147–160.
15. Martí R, Duarte A, Laguna M. Advanced scatter search for the max-cut problem. *INFORMS J Comput* 2009;21(1):26–38.
16. Laguna M, Martí R. Experimental testing of advanced scatter search designs for global optimization of multimodal functions. *J Glob Optim* 2005;33:235–255.
17. Campos V, Glover F, Laguna M, *et al.* An experimental evaluation of a scatter search for the linear ordering problem. *J Glob Optim* 2001;21:397–414.
18. Silva GC, Ochi LS, Martins SL. Experimental comparison of greedy randomized adaptive search procedures for the maximum diversity problem, *Lecture Notes in Computer Science* 3059. Berlin Heidelberg: Springer; 2004. pp. 498–512.

THE SEARCH ALLOCATION GAME

RYUSUKE HOHZAKI
National Defense Academy,
Department of Computer
Science, Yokosuka,
Japan

INTRODUCTION

An ultimate purpose of *search theory* is to answer how efficiently we could find targets. In search theory, there are two main players: searchers and targets. As *operations research* (OR) was born in World War II as a realistic academia, search theory originated from the military research of the Anti-Submarine Warfare Operations Research Group (ASWORG) of the U.S. Navy for anti-submarine warfare (ASW). The famous books *Methods of Operations Research* by Morse and Kimball [1] and *Search and Screening* by Koopman [2] transmitted the fruitful methodology of OR to the world. In Koopman's book, he has discussed in one of his chapters on how to distribute a limited amount of time in two regions for an effective search. He made a big contribution to the theoretical progress of search theory, being conscious of the important role of optimization in OR.

Koopman submitted some articles to the early editions of *OR* and developed his theory on optimal distribution of searching effort or searching resource. In Ref. 3 published in 1953, he focused on search in two areas and generalized the theory in Ref. 4, entitled "The theory of search ‡V: The optimum distribution of searching effort." The origin of many optimization problems arising afterward in the field of search theory is Koopman's problem, which are discussed below.

Let us consider the set of all real numbers $\mathbf{R}$ as a search space. A target hides at a position $x \in \mathbf{R}$ with probability density $p(x)$, where nonnegative $p(x)$ satisfies

$\int_{\mathbf{R}} p(x)dx = 1$. A searching resource or an effort means any kind of resource to be useful for search, such as search time, sensors, and manpower. Let $\varphi \equiv \{\varphi(x), x \in \mathbf{R}\}$ be a distribution plan, where $\varphi(x)$ is the resource density to be distributed at x . If the target is at x and the resource density $\varphi(x)$ is distributed there, the searcher can detect the target with probability $1 - \exp[-\varphi(x)]$. What is an optimal distribution plan of the total amount of resource Φ to make the detection probability the largest? This problem is formulated as follows:

$$(K_c) \quad \max_{\{\varphi(x)\}} P[\varphi] \equiv \int_{\mathbf{R}} p(x) \{1 - \exp[-\varphi(x)]\} dx$$

$$\text{s.t.} \quad \int_{\mathbf{R}} \varphi(x)dx = \Phi, \quad \varphi(x) \geq 0, \quad x \in \mathbf{R}. \quad (1)$$

From the necessary condition of optimality $P(\varphi + \delta\varphi) - P(\varphi) \leq 0$ for an optimal solution φ and any little perturbed solution $\varphi + \delta\varphi$, Koopman derived an optimal solution φ^* . We can easily reach his conclusion that $p(x)\exp[-\varphi(x)] = \lambda(\text{const.})$ if $\varphi(x) > 0$, by applying calculus of variations to the variational problem (K_c) , in which we maximize the functional $F(\varphi, x) \equiv p(x)\{1 - \exp[-\varphi(x)]\}$ under Equation 1. The calculus says that an optimal function φ^* is given by the Euler–Lagrange characteristic equation $d[F(\varphi, x) - \lambda\varphi(x)]/d\varphi = 0$, where λ is a Lagrangean multiplier.

We can define Koopman's problem in a discrete cell space, as follows. For a target that is estimated to be in cells $i = 1, 2, \dots, n$ with respective probability $p_i (p_i \geq 0, \sum_i p_i = 1)$, how should we distribute Φ searching resources in an optimal manner to maximize the detection probability of the target? Our question is formulated by

$$(K_d) \quad \max_{\{\varphi_i\}} P(\varphi) = \sum_{i=1}^n p_i [1 - \exp(-\varphi_i)]$$

$$\text{s.t.} \quad \sum_{i=1}^n \varphi_i = \Phi, \quad \varphi_i \geq 0, \quad i = 1, \dots, n.$$

To deal with such optimization problems, we already have a good methodology such as convex analysis and nonlinear programming. Koopman's problem is also the first step for more general field of research named *optimal resource allocation problems* [5]. de Guenin [6] generalized Koopman's problem to a variational problem and Kadane [7] extended it to a discrete optimization problem with discrete resources.

In *Methods of Operations Research*, the authors had already considered the problem of how to intercept the passage of a submarine in straits by an ASW-airplane, modeling it as a two-player game. But such a two-sided problem is exceptional. After Koopman, there have been many studies dealing with one-sided problems for an optimal search given that the searcher knows the probabilistic law of target existence, as in Koopman's problem. From search problems with stationary targets, search theory steps forward to problems with moving targets. Stone developed a theoretical analysis of stationary target problems in a book, named *Theory of Optimal Search* [8] and won the Lanchester Prize.

As an early research on the moving target problem, there is Pollock's work [9], where he discussed an optimal search for target hiding and moving in two cells, and other research such as Dobbie [10], Hellman [11], Iida [12], and Kan [13]. Brown [14] proposed a numerical algorithm to compute an optimal distribution of searching resource, which varies on a discrete time horizon, to maximize the total detection probability. Washburn [15] extended Brown's method to a discrete resource problem. This was generalized by Stromquist and Stone [16]. From the point of view of optimization, their methods belong to convex optimization and global optimization for continuous objectives. Eagle and Yee [17] and Hohzaki and Iida [18] apply branch-and-bound methods or relaxation methods to search problems with constraints on the searcher's motion.

In the research surveyed so far, the authors do not regard the target existence or the target movement as the decision making of the target but as stochastic parameters, and discuss their problems as one-sided problems. The one-sided search problem

progresses to a two-sided problem, namely a search game, where the target is a decision maker as well as the searcher. Search games can be split into two categories, depending on the strategy of searcher: the distribution of searching resource or the movement in a search space. The former is referred to as *search allocation game* (SAG) [19]. The latter can be split into two further categories. We refer to the problem with stationary targets as the *hide-and-search game* (HSG) and the one with moving targets as the *evasion-and-search game* (ESG). For targets, almost all research takes the moving strategy, assuming that the target's mobility is much inferior to the searcher's, as seen in the ASW operation between submarines and ASW airplanes. Some researchers take exceptional models such as Baston and Garnaev [20] and the *Blotto game* [21], where the searcher and the target adopt the same type of strategy to distribute their resources or forces to overwhelm their opponents.

Norris [22] studied a two-person zero-sum search game whose payoff is the number of searches until the detection of the target. The searcher repeatedly looks in several boxes but possibly overlooks the target even though the target is there. Norris' problem is regarded as an SAG with discrete resource. Nakai [23] and Iida *et al.* [24] discussed single-stage SAGs and Kikuta [25] dealt with a single-stage HSG. Baston and Bostock [26] and Garnaev [27] analyzed an optimal strategy of an ASW helicopter to drop mines to destroy a hiding submarine as a single-stage ESG. Danskin [28] discussed the dipping strategy of sound buoys for an ASW airplane to detect a submarine moving diffusively in a two-dimensional plane. Because the submarine takes a fixed course and speed, he deals with the problem as an HSG on the plane called *speed circle*. We can take Eagle and Washburn [29] as another example of single-stage ESGs with moving searcher. Meinardi [30] discussed a competition between a searcher and a target as a multistage ESG. Both players are moving on a line. As the payoff, Washburn [31] and Nakai [32] adopt the total traveling distance of the searcher until detection of the target, which happens on the coincidence of cells selected by the searcher

and the target. Alpern [33], Gal [34,35], and Alpern and Gal [36] are big contributors to the HSG and the ESG.

We have taken an overview of the literatures on the search problem since Koopman began search theory. Koopman's formulation for an optimal distribution of resource is compatible with formulations in general optimization problems and therefore we can use the basics of nonlinear programming such as convex programming to solve the SAG. Unlike the SAG, the HSG and the ESG more likely belong to discrete optimization problem and they require special idea or customization to be solved. There are many differences between the SAG and the (H/E)SG from the methodological point of view. We focus on SAG in the following sections.

SAG FOR A STATIONARY TARGET

Let us consider the following two-person zero-sum game with a hiding target and a searcher.

- A search space consists of n cells, $\mathbf{K} = \{1, 2, \dots, n\}$, and a target hides in one of them.
- A searcher distributes Φ continuously divisible searching resources into the cells to detect the hiding target.
- If the target is in i and x searching resources are scattered there, the searcher detects the target with probability $f_i(x) \equiv 1 - \exp(-\alpha_i x)$, where α_i indicates the effectiveness of unit resource in cell i on detection.

The searcher wants to maximize the detection probability and the target desires to minimize it.

Let us denote a mixed strategy of the target by $\mathbf{p} = \{p_i, i \in \mathbf{K}\}$, where p_i is the probability with which the searcher hides in cell i . A pure strategy of the searcher is denoted by $\mathbf{x} = \{x_i, i \in \mathbf{K}\}$. x_i is the quantity of resource to be distributed in cell i . We have the expression for the expected detection probability, $P(\mathbf{x}, \mathbf{p}) = \sum_i p_i [1 - \exp(-\alpha_i x_i)]$.

We can see that there exists an equilibrium point, $\mathbf{x}^*$ and $\mathbf{p}^*$, for our search game

from known properties of the so-called convex game [37] because $P(\mathbf{x}, \mathbf{p})$ is strictly concave for $\mathbf{x}$. The following solution satisfies the optimality conditions of $\mathbf{x}^* = \arg \max P(\mathbf{x}, \mathbf{p}^*)$ and $\mathbf{p}^* = \arg \min Q(\mathbf{x}^*, \mathbf{p})$.

$$x_i^* = \frac{1/\alpha_i}{\sum_{j \in \mathbf{K}} 1/\alpha_j} \Phi, p_i^* = \frac{1/\alpha_i}{\sum_{j \in \mathbf{K}} 1/\alpha_j}.$$

As a result, the optimal strategies x_i and p_i look simple in such a manner that they are proportional to the reciprocal of α_i . An optimal solution of the one-sided problem $\max_{\mathbf{x}} P(\mathbf{x}, \mathbf{p})$ with fixed $\mathbf{p}$ is given by $x_i^* = [\log(\alpha_i p_i / \lambda)]^+ / \alpha_i$, where $[x]^+ \equiv \max\{0, x\}$ and λ is a Lagrangean multiplier. It often happens that the solution of the game is simpler than the solution of the one-sided optimization problem.

SAG FOR A MOVING TARGET

Here we discuss an SAG with a moving target. Although we consider time in our model, we define our SAG in a discrete geographic space and a discrete time space, as in the section titled "SAG for a Stationary Target."

- B1. Let us denote a search space by $\mathbf{K} \times \mathbf{T}$ consisting of a discrete cell space $\mathbf{K} = \{1, \dots, n\}$ and a discrete time space $\mathbf{T} = \{1, \dots, T\}$.
- B2. Let us denote a set of target paths by Ω , which is a set of functions from $\mathbf{T}$ to $\mathbf{K}$. A target chooses one from the set and follows it as time passes by. A path $\omega \in \Omega$ runs in cell $\omega(t) \in \mathbf{K}$ at time $t \in \mathbf{T}$.
- B3. A searcher wants to detect the target using searching resource at hand. $\Phi(t)$ is the amount of resource available at time t . The searcher can start the search from time τ .
- B4. If the target is in cell i and the searcher distributes x searching resources there, the searcher detects the target with probability $1 - \exp(-\alpha_i x)$. An event of no-detection at a time is exclusive to events at other time points.
- B5. The searcher acts as a maximizer and the target as a minimizer for the detection probability.

Let us denote a searcher's strategy by $\varphi = \{\varphi(i, t), i \in \mathbf{K}, t \in \hat{\mathbf{T}}\}$ during a time interval $\hat{\mathbf{T}} = \{\tau, \tau + 1, \dots, T\}$ starting from τ . $\varphi(i, t) \in \mathbf{R}$ is the amount of resource which the searcher distributes in cell i at time t . We denote a mixed strategy of the target by $\pi = \{\pi(\omega), \omega \in \Omega\}$. $\pi(\omega)$ is the selection probability for path $\omega \in \Omega$.

For a target path ω and a searcher's strategy φ , the detection probability is given by

$$P(\varphi, \omega) = 1 - \exp \left[- \sum_{t=\tau}^T \alpha_{\omega(t)} \varphi(\omega(t), t) \right], \quad (2)$$

which is the payoff of the game, and its expectation $P(\varphi, \pi) = \sum_{\omega} \pi(\omega) P(\varphi, \omega)$ is to be maximized for the searcher and minimized for the target. Because the expected payoff is linear for π and strictly concave for φ , its maximin value coincides with its minimax value, as for the SAG with stationary targets. The maximin optimization gives us an optimal searcher strategy φ^* and the minimax optimization gives an optimal mixed strategy for the target π^* . An equilibrium point of the game has a much simpler form than an optimal solution of the one-sided problem $\max_{\varphi} P(\varphi, \pi)$. While the solution of the one-sided problem can be calculated by Brown's algorithm [14] or Washburn's algorithm [15] using nonlinear programming, an optimal searcher strategy of the search game is given by the following linear programming problem:

$$\begin{aligned} (P_S) \quad & \max_{\varphi} \eta \\ \text{s.t.} \quad & \sum_{t=\tau}^T \alpha_{\omega(t)} \varphi(\omega(t), t) \geq \eta, \omega \in \Omega, \\ & \sum_{i \in \mathbf{K}} \varphi(i, t) \leq \Phi(t), t \in \hat{\mathbf{T}}, \\ & \varphi(i, t) \geq 0, i \in \mathbf{K}, t \in \hat{\mathbf{T}}. \end{aligned}$$

An optimal strategy of the target, π^* , is given by the following linear programming

formulation:

$$\begin{aligned} (P_T) \quad & \min_{\{\pi(\omega), \pi(\omega)\}} \sum_{t \in \hat{\mathbf{T}}} \Phi(t) v(t) \\ \text{s.t.} \quad & \alpha_i \sum_{\omega \in \Omega_{it}} \pi(\omega) \leq v(t), i \in \mathbf{K}, t \in \hat{\mathbf{T}}, \\ & \sum_{\omega \in \Omega} \pi(\omega) = 1, \pi(\omega) \geq 0, \omega \in \Omega, \end{aligned}$$

where Ω_{it} is a set of paths going through cell i at time t , namely, $\Omega_{it} \equiv \{\omega \in \Omega | \omega(t) = i\}$. Problems (P_S) and (P_T) are dual to each other and we can obtain the value of the original search game by calculating $1 - \exp(-\eta^*)$ using the optimal value η^* of the above problems (P_S) or (P_T) . As seen from the formulations (P_S) and (P_T) , we can change the exponential form of payoff $P(\varphi, \omega)$ to the linear form $R(\varphi, \omega) \equiv \sum_{t \in \hat{\mathbf{T}}} \alpha_{\omega(t)} \varphi(\omega(t), t)$ without loss of generality. The equilibrium point of the SAG is more easily computed than that of the one-sided problem.

ENERGY CONSTRAINTS ON TARGET MOTION AND LARGE-SIZE OF PROBLEMS

Real moving targets always have their constraints of motion in the real world. One of them is energy for moving. If the target is a battery-powered submarine, a captain would always be worrying about the energy of the battery to maintain the submarine's mobility. The energy constraints are implicitly involved in the formulations (P_S) and (P_T) such that all paths Ω are enumerated taking account of the constraints. But we have to think over the solution feasibility of the formulation in terms of the total number of paths. If the target can move to any of the n cells of $\mathbf{K}$ each time, we can count the total number of paths as n^T , which is too many to solve the problems (P_S) or (P_T) in feasible computational time. Let us explicitly express the energy constraints for path ω , as follows:

$$\omega(1) \in S_0, \quad (3)$$

$$\omega(t+1) \in N(\omega(t), t), t = 1, \dots, T-1, \quad (4)$$

$$\sum_{t=1}^{T-1} \mu(\omega(t), \omega(t+1)) \leq e_0. \quad (5)$$

Cells $S_0 \subseteq \mathbf{K}$ are possible areas where the target is at initial time $t = 1$. $N(i, t) \subseteq \mathbf{K}$ are cells which the target can move to from cell i at time t and $\mu(i, j)$ is the energy which the target requires to move from cell i to j . e_0 is the initial energy of the target. If the target uses it up, it must stop at the last cell it reaches.

Assuming that function $\mu(i, j)$ is integer-valued, we denote the whole energy states of the target by $\mathbf{E} = \{0, \dots, e_0\}$. Each state of the target is represented by a triplet (i, t, e) , which means that it has residual energy e in cell i at time t . Let us define new variables $q(i, t, e)$ as the probability that the target is in state (i, t, e) and $v(i, j, t, e)$ as the probability that the target is in state (i, t, e) and moves to cell j at time $t + 1$. In the formulations proposed below, we take the variables $\{q(i, t, e)\}$ and $\{v(i, j, t, e)\}$ as the target strategy while $\{\varphi(i, t)\}$ still represents the searcher's strategy. It is easy to construct the strategy $q(\cdot)$ and $v(\cdot)$ from original strategy $\{\pi(\omega)\}$ and reversely transform them. But we omit them.

As stated before, let $R(\varphi, \omega) = \sum_{t \in \mathbf{T}} \alpha_{\omega(t)} \varphi(\omega(t), t)$ be a newly defined payoff function and let variable $h(t)$ be the maximum expected payoff obtained by some optimal search after time t . As reachable cells from state (i, t, e) are denoted by $N(i, t, e) = \{j \in N(i, t) | \mu(i, j) \leq e\}$, the cells from which the target comes into state (i, t, e) at time t are given by $N^*(i, t, e) = \{j \in \mathbf{K} | i \in N(j, t - 1, e + \mu(j, i))\}$. The following dynamic programming (DP) equation is valid for $h(t)$. We can regard the formulation as an optimization problem with respect to the probability flow of target existence.

$$\begin{aligned} h(t) &= \max_{\varphi \in \Psi} \left\{ \sum_{i \in \mathbf{K}} \alpha_i \varphi(i, t) \sum_{e \in \mathbf{E}} q(i, t, e) + h(t + 1) \right\} \\ &= \Phi(t) \max_{i \in \mathbf{K}} \left(\alpha_i \sum_{e \in \mathbf{E}} q(i, t, e) \right) + h(t + 1). \end{aligned}$$

After some consideration, we obtain a linear programming formulation to give us the minimax value of the game and an optimal

target strategy q^* and v^* .

$$\begin{aligned} (E_T) \quad & \min h(\tau) \\ \text{s.t.} \quad & h(t) \geq \Phi(t) \alpha_i \sum_{e \in \mathbf{E}} q(i, t, e) + h(t + 1), \\ & i \in \mathbf{K}, t = \tau, \dots, T - 1, \\ & h(T) \geq \Phi(T) \alpha_i \sum_{e \in \mathbf{E}} q(i, T, e), i \in \mathbf{K}, \\ & q(i, t, e) = \sum_{j \in N(i, t, e)} v(i, j, t, e), \\ & i \in \mathbf{K}, t = 1, \dots, T - 1, e \in \mathbf{E}, \\ & q(i, t, e) = \sum_{j \in N^*(i, t, e)} v(j, t - 1, e + \mu(j, i)), \\ & i \in \mathbf{K}, t = 2, \dots, T, e \in \mathbf{E}, \\ & \sum_{i \in S_0} q(i, 1, e_0) = 1, \\ & \sum_{i \in \mathbf{K}} \sum_{e \in \mathbf{E}} q(i, t, e) = 1, t \in \mathbf{T}, \\ & q(i, 1, e) = 0, i \in \mathbf{K}, e \in \mathbf{E} - \{e_0\}, \\ & q(i, 1, e) = 0, i \in \mathbf{K} - S_0, e \in \mathbf{E}, \\ & v(i, j, t, e) \geq 0, \\ & i, j \in \mathbf{K}, t = 1, \dots, T - 1, e \in \mathbf{E}. \end{aligned}$$

We can derive the other linear programming problem, say (E_S) , for an optimal searcher strategy by the DP formulation for the maximin optimization in the same manner to the derivation of (E_T) (see Hohzaki *et al.* [38]). The formulations (E_T) and (E_S) require order $O(K^2 T e_0)$ of memory size. From this fact, the formulations are preferable to (P_T) and (P_S) from the point of view of computational feasibility.

EXTENSIONS OF SAG

In this section, we consider the extension of our SAG model to more practical applications. First, we discuss the continuous SAG defined in continuous space. Further we mention other extensions or modifications of our models described in the sections titled "SAG for a Moving Target" and "Energy Constraints on Target Motion and Large-Size of Problems."

Continuous SAG and Applications

We replace the discrete search space in assumption (B1) with a continuous space $\mathbf{K} = [0, K] \subseteq \mathbf{R}$ and a continuous time $\mathbf{T} = [0, T] \subseteq \mathbf{R}$. We denote a distribution density of searching resource at point $(x, t) \in \mathbf{K} \times \mathbf{T}$ by $\varphi(x, t)$ and put a limit on the usage rate of the total available resource, $\int_{\mathbf{K}} \varphi(x, t) dx \leq \phi(t)$, at $t \in \mathbf{T}$. The target adopts a moving strategy denoted by a continuous and differentiable function of time $\omega(t) \in \mathbf{K}$, as in the discrete SAG. As the constraint on target motion, we replace condition (Eq. 4 with $|d\omega(t)/dt| \leq V_c$ and energy constraint (Eq. 5 with $\int_{\mathbf{T}} \mu(d\omega(t)/dt) dt \leq e_0$.

For the continuous SAG with payoff $R(\varphi, \omega) = 1 - \exp\{-\int_{\mathbf{T}} \alpha_{\omega(t)} \varphi(\omega(t), t) dt\}$ defined by strategies $\varphi(x, t)$ and $\omega(t)$, we have the following theorem about the relationship between a continuous SAG and the corresponding discrete SAG that is derived from the continuous SAG by discretizing the continuous space $\mathbf{K} \times \mathbf{T}$ in an appropriate manner. It is proved that a sequence of equilibrium points of the corresponding discrete SAGs converges to the equilibrium of the continuous SAG [39].

Theorem1 [39]. *There exists an equilibrium point for the continuous SAG and it is given by the convergence point for the corresponding discrete SAGs when the number of pieces the search space is divided into approaches infinity.*

When we divide a multidimensional continuous space into pieces such as $\mathbf{K} = \{1, \dots, n\}$ and construct a corresponding discrete SAG, we can compute an approximation close to an optimal strategies of the continuous SAG by the formulations in the sections titled “SAG for a Moving Target” and “Energy Constraints on Target Motion and Large-Size of Problems.” For a search game with a diffusively moving target on a plane, named a *flaming datum search game*, Washburn and Hohzaki [40,41] proposed some estimations of upper bound and lower bound for the value of the game by optimal control theory. They confirm that a sequence of optimal solutions of the corresponding

discrete games converges to an optimal one of the original continuous SAG by Theorem 1 and they numerically observe that the values of the discrete SAGs always stay between the lower and upper bounds of Washburn and Hohzaki.

We can demonstrate other practical applications of the SAG, taken in Hohzaki *et al.* [38], for which we can derive optimal solutions by the minor modification of the formulations (E_T) and (E_S) in the section titled “Energy Constraints on Target Motion and Large-Size of Problems.” The first example is a search game with shelter, which gives the target a safe place. The second one is about the good timing of the target’s energy supply or charging battery by a submarine in the ASW operation.

SAG with False Contacts

False contacts by sensors are inevitable in practical search operations, especially in the ASW. The event of false contacts has been an important theme in search theory. Stone discusses in a chapter of his book about his analysis of the event [8]. Sonar could pick up signals from false targets such as whales or from noises and sonar men could fail to recognize origins of signals correctly. Therefore, they inspect the first signals several times and grade up their correctness through the inspection process. Kisi [42] and Iida *et al.* [43] consider false contacts, which are assumed to occur according to the Poisson distribution, and discuss the inspection time to maximize the detection probability of true target within a limited amount of time. Their problem is a kind of one-sided optimization problem, in which the searcher determines the inspection time for a signal each time it occurs. When a false contact occurs, much inspection time increases the probability to recognize correctly what it is but makes the searcher spend time in vain without identifying true contacts. In Hohzaki [44], the author considers the following game, assuming that the probability of false-contact occurring depends on the amount of used resource or the number of dropped sonars.

In Hohzaki’s model, the following additional assumption (F1) about false contacts is added to the basic model of (B1)–(B5) in

the section titled “SAG for a Moving Target”: (F1). Given that the search is executed at time t , a false contact occurs with probability $q_t \equiv \sum_i \beta_i \varphi(i, t) + \gamma(t)$ at the end of time t . The coefficient β_i indicates the inducing efficiency of unit searching effort in cell i and $\gamma(t)$ is the stationary occurrence rate. Both parameters are positive and small enough to satisfy $\max_i \beta_i \Phi(t) + \gamma(t) < 1$. On the occurrence of a false contact, the searcher focuses on the investigation of the contact and he can start the search again after a fixed length of time period t_f at the earliest.

In the SAG with false contacts, the searcher cannot carry out search operations during the inspection process for false contacts. If the searcher distributes many resources aiming to detect the true target, his behavior increases the probability of false-contact occurrence as well as the detection probability of the target and it results in many time periods in which the searcher cannot search. The false-contact model requires the searcher to make an adequate distribution plan of searching resource while balancing the trade-off between the detection of target and the occurrence of false contacts.

If we denote the probability that the searcher can search at time $t \in \hat{\mathbf{T}}$ by $S(t)$, the probability satisfies the following recursive equations and then we can evaluate the payoff of a searcher's strategy φ and a target's path ω by $R(\varphi, \omega) = \sum_{t=\tau}^T S(t) \alpha_{\omega(t)} \varphi(\omega(t), t)$.

$$S(t) = \begin{cases} 1, & \text{for } t = \tau \\ S(t-1)(1 - q_{t-1}), & \text{for } \tau + 1 \leq t \leq \tau + t_f - 1 \\ S(t-1)(1 - q_{t-1}) + S(t-t_f)q_{t-t_f}, & \text{for } \tau + t_f \leq t \leq T. \end{cases}$$

The equations above implicitly involve the searcher strategy φ in a complicated manner and then the payoff does so. In spite of the nonlinearity, we can propose a linear programming problem for the optimal strategies by means of transformation and substitution (see Hohzaki [44]). In the ASW, contact information is classified into several grades based on its correctness. The false-contact model is extended to the model with several types of false contacts by Hohzaki [45].

Other Applications

In this section, we show some other applications of the SAG.

SAG Taking Account Of Attributes Of Searching Resource. We use many kinds of searching resources in search operations, for example, sonar buoys for submarines, searching time or manpower for the search. Taking an example of a flare used in search-and-rescue operations, its effect lasts for some time and extends to some areas at a distance from its shooting point. Hohzaki [46,47] embedded the durability and reachability of searching resources in his SAG model. Dambreville and Le Cadre [48,49] considered a variety of constraints on the amount of resource. To the basic SAG model (B1)–(B5), Hohzaki introduced an additional assumption: (P1). The effect of the resource lasts for time duration D after its dropped time. Furthermore, it expands to an area $A(i) \subseteq \mathbf{K}$ although the intensity of the effect attenuates by $\beta(i, j)$ based on the distance from the dropped point i to a point j .

The payoff of the SAG model is

$$R(\varphi, \omega) = \sum_{t \in \hat{\mathbf{T}}} \alpha_{\omega(t)} \sum_{\xi = \max\{\tau, t-D\}}^t \sum_{j \in A^*(i)} \beta(j, \omega(t)) \varphi(j, \xi),$$

where $A^*(i)$ is the set of cells at which the resource dropped has effects on cell i and is defined by $A^*(i) \equiv \{j \in \mathbf{K} | i \in A(j)\}$.

SAG with Energy Supply Strategy. Apart from submarines or other vehicles, energy is crucial to the mobility of the target and the mobility gives the target the flexibility to avoid searchers in search games. The submarine supplies his energy at the risk of letting himself more likely to be detected by the searcher. Some articles deal with the strategy of energy supply as a target strategy. In Hohzaki and Ikeda [50], they put an extra assumption to the basic SAG model: (S1). A target can supply his energy by $\delta > 0$ during a time period. As well as the selection of his paths, he chooses one of the two options concerning the supply of energy: supply or no

supply. If he takes the strategy of supply, his energy increases by δ at the end of a relevant time point or at the beginning of the next time. But, during the supply, the effectiveness of unit searching resource in cell i on target detection changes from α_i to β_i , where $\alpha_i < \beta_i$ usually holds.

In this model, a target pure strategy is denoted by a combination of a path $\omega \in \Omega$ and a supply rule $\sigma = \{\sigma(t), t \in \mathbf{T}\}$, where $\sigma(t)$ is 1 for “Supply” and 0 for “No-supply” at time t . For a searcher strategy φ and a target strategy $\rho = (\omega, \sigma)$, the payoff of the SAG with energy supply strategy is given by

$$R(\varphi, \rho) = \sum_{t \in \mathbf{T}} \varphi(\omega(t), t) \{ \alpha_{\omega(t)} [1 - \sigma(t)] + \beta_{\omega(t)} \sigma(t) \}.$$

Miscellanea. There are other topics concerned with the SAG. We mentioned a datum search game on a two-dimensional space in the section titled “Continuous SAG and Applications.” The ASW forces sometimes repeat contact and loss of submarine’s signals in the real ASW. Therefore, we are also interested in the result that the target and the searcher fight each other in a sequence of such datum search games. The game could be modeled by a multi-stage SAG. At each stage when an element datum search game is played, the submarine loses some of his energy but gets information about searching resource of the searcher, and the searcher obtains some information about whereabouts of the target while consuming searching resource. In the multi-stage SAG, players have to make a sequence of decisions, for example, “where should I go next?” or “how much resource and where I’d better distribute?”. Hohzaki [51] analyzed the multi-stage SAG.

The SAG implicitly presumes the competition between a searcher and a target in the ASW operation. That is why almost all research on the SAG was modeled by noncooperative two-person zero-sum game. However, if there are multiple searchers involved in the SAG such as treasure hunting, each searcher would compete with others to get the value of target. In such an SAG, each searcher would behave to increase his own

interest and the payoff of the game would be nonzero-sum. Hohzaki [52] discussed a nonzero-sum search game with two competitive searchers and a target, and he showed that there is certainly an equilibrium point, in which both searchers cooperatively act as a team to get reward by detection of target if the target is stationary. In such an equilibrium, the nonzero-sum SAG with multiple searchers and a stationary target is essentially the same as a two-person zero-sum SAG with a team of searchers and a target as two competitors. Whether this property is valid for a moving target SAG is still an open problem.

For the SAG, the assumption of noncooperativeness has been kept for a long time. However we can imagine some situations in which some players might take a cooperative operation. Several searchers could help each other to search for a target. For example, some players or some companies organize a syndicate to search for valuable minerals. Coast guard or navy could take a cooperative operation to search for missing persons at sea. Hohzaki [53] discussed a cooperative SAG, where some searchers cooperate against a target. He first derives the conditions that give searchers an incentive to form a coalition and then models the SAG by the so-called coalition game of cooperative game (see Owen [37]). He generates characteristic functions customized to search operations and proposes a methodology to divide the obtained target value among members of the coalition. The cooperative SAG has just started and there is much future work to be done to clarify the properties of the cooperative SAG.

REFERENCES

1. Morse PM, Kimball GE Methods of operations research. Cambridge: MIT Press; 1951.
2. Koopman BO. Search and screening. New York: Pergamon; 1980. p 221–227.
3. Koopman BO. The optimum distribution of effort. Oper Res 1953;1:52–63.
4. Koopman BO. The theory of search III: the optimum distribution of searching effort. Oper Res 1957;5:613–626.

5. Ibaraki T, Katoh N. Resource allocation problems: algorithmic approaches. London: The MIT Press; 1988.
6. de Guenin J. Optimum distribution of effort: an extension of the Koopman basic theory. *Oper Res* 1961;9:1–9.
7. Kadane JB. Discrete search and the Neyman-Pearson lemma. *J Math Appl* 1968;22:156–171.
8. Stone LD. Theory of optimal search. New York: Academic Press; 1969.
9. Pollock SM. A simple model of search for a moving target. *Oper Res* 1970;18:883–903.
10. Dobbie JM. A two-cell model of search for a moving target. *Oper Res* 1974;22:79–92.
11. Hellman O. On the optimal search for a randomly moving target. *SIAM J Appl Math* 1972;22:545–552.
12. Iida K. An optimal distribution of searching effort for a moving target (in Japanese). *Keiei Kagaku* 1972;16:204–215.
13. Kan YC. Optimal search of a moving target. *Oper Res* 1977;25:864–870.
14. Brown SS. Optimal search for a moving target in discrete time and space. *Oper Res* 1980;28:1275–1286.
15. Washburn AR. Search for a moving target: the FAB algorithm. *Oper Res* 1983;31:739–751.
16. Stromquist WR, Stone LD. Constrained optimization of functionals with search theory applications. *Math Oper Res* 1981;16:518–529.
17. Eagle JN, Yee JR. An optimal branch-and-bound procedure for the constrained path moving target search problem. *Oper Res* 1990;38:110–114.
18. Hohzaki R, Iida K. Path constrained search problem with reward criterion. *J Oper Res Jpn* 1995;38:254–264.
19. Garnaev AY. Search games and other applications of game theory. Tokyo: Springer-Verlag; 2000.
20. Baston VJ, Garnaev AY. A search game with a protector. *Nav Res Logist* 2000;47:85–96.
21. Washburn A. TPZS applications: Blotto games, Wiley Encyclopedia of Operations Research and Management. Science 2011;8:5504–5511.
22. Norris RC. Studies in search for a conscious evader. MIT Technical Report No. 279; 1962.
23. Nakai T. Search models with continuous effort under various criteria. *J Oper Res Soc Jpn* 1988;31:335–351.
24. Iida K, Hohzaki R, Sato K. Hide-and-search game with the risk criterion. *J Oper Res Soc Jpn* 1994;37:287–296.
25. Kikuta K. A search game with traveling cost. *J Oper Res Soc Jpn* 1991;34:365–382.
26. Baston VJ, Bostock FA. A one-dimensional helicopter-submarine game. *Nav Res Logist* 1989;36:479–490.
27. Garnaev AY. A remark on a helicopter-submarine game. *Nav Res Logist* 1993;40:745–753.
28. Danskin JM. A helicopter versus submarine search game. *Oper Res* 1968;16:509–517.
29. Eagle JN, Washburn AR. Cumulative search-evasion games. *Nav Res Logist* 1991;38:495–510.
30. Meinardi JJ. A sequentially compounded search game. Theory of games: techniques and applications. London: The English Universities Press; 1964. p 285–299.
31. Washburn AR. Search-evasion game in a fixed region. *Oper Res* 1980;28:1290–1298.
32. Nakai T. A sequential evasion-search game with a goal. *J Oper Res Soc Jpn* 1986;29:113–122.
33. Alpern S. Rendezvous search games, Wiley Encyclopedia of Operations Research and Management. Science 2011;7:4499–4511.
34. Gal S. Search games. New York: Academic Press; 1980.
35. Gal S. Search games, Wiley Encyclopedia of Operations Research and Management. Science 2011;7:4743–4750.
36. Alpern S, Gal S. The theory of search games and rendezvous. Boston (MA): Kluwer Academic Publishers; 2003.
37. Owen G. Game theory. New York: Academic Press; 1995.
38. Hohzaki R, Iida K, Komiya T. Discrete search allocation game with energy constraints. *J Oper Res Soc Jpn* 2002;45:93–108.
39. Hohzaki R. Search allocation game. *Eur J Oper Res* 2006;172:101–119.
40. Washburn AR, Hohzaki R. The diesel submarine flaming datum problem. *Mil Oper Res* 2001;4:19–30.
41. Hohzaki R, Washburn A. An approximation for a continuous datum search game with energy constraint. *J Oper Res Soc Jpn* 2003;46:306–318.
42. Kisi T. Optimal stopping of the investigating search: search theory and applications (NATO Conference Series II-8). New York: Plenum Press; 1979.

43. Iida K, Hohzaki R, Kaiho K. Optimal investigating search maximizing the detection probability. *J Oper Res Soc Jpn* 1997;40:294–309.
44. Hohzaki R. Discrete search allocation game with false contacts. *Nav Res Logist* 2007;54:46–58.
45. Hohzaki R. A search game with several types of false contacts. In: Takahashi W, Tanaka T, editors. *Nonlinear analysis and convex analysis*. London: Yokohama Publishers; 2004. p 59–79.
46. Hohzaki R. A search game taking account of attributes of searching resources. *Nav Res Logist* 2008;55:76–90.
47. Hohzaki R. A search game taking account of linear effects and linear constraints of searching resource. *J Oper Res Soc Jpn* 2012;55:1–22.
48. Dambreville F, Le Cadre JP. Detection of a Markovian target with optimization of the search efforts under generalized linear constraints. *Nav Res Logist* 2002;49:117–142.
49. Dambreville F, LeCadre, J-P. Search game for a moving target with dynamically generated informations. *Proceedings of the 5th International Conference on Information Fusion (FUSION'2002)*; Annapolis 2002. p 243–250.
50. Hohzaki R, Ikeda K. A search game with a strategy of energy supply for target. In: Bernhard P, Gaitsgory V, Pourtallier O, editors. *Advances in dynamic games and their applications*. Volume 10, *Annals of the International Society of Dynamic Games*. Boston: Springer; 2009. p 313–337.
51. Hohzaki R. A multi-stage search allocation game with the payoff of detection probability. *J Oper Res Soc Jpn* 2007;50:178–200.
52. Hohzaki R. A nonzero-sum search game with two competitive searchers and a target. *Ann Dyn Games* 2013;12:351–373.
53. Hohzaki R. A cooperative game in search theory. *Nav Res Logist* 2009;56:264–278.

THE SHAPLEY VALUE AND RELATED SOLUTION CONCEPTS

MILAN VLACH
Faculty of Mathematics and
Physics, Charles University in
Prague, Czech Republic

PRELIMINARIES

Recall that basic data specifying a cooperative game with TU, that is transferable utility (or with side payments), in short a TU game or a game, are composed of a nonempty set N and a real-valued function v defined on the collection of all subsets of N and having the property $v(\emptyset) = 0$ where $\emptyset$ stands for the empty set. We shall denote the game given through N and v by (N, v) , or simply v , and the collection of all TU games with player set N by $\mathcal{G}(N)$. As usual, the elements of N are called *players*, subsets of N (inclusive of the empty set and entire set N) are called *coalitions*, the set N of all players is referred to as the *grand coalition*, and the function v is called the *characteristic function* of the game or *coalition function*. If S and T are coalitions such that $S \subset T$, then we say that S is a *sub-coalition* of T and T is a *supercoalition* of S .

It is customary to interpret the value $v(S)$ of characteristic function v at coalition S as the worth of coalition S or the total payoff that coalition S will be able to distribute among its members, provided exactly the coalition S forms. However, equally well, the number $v(S)$ may represent the total cost of reaching some common goal of coalition S that must be shared by the members of S ; or some other quantity, depending on the application field.

Note that the sum $v + w$ of games from $\mathcal{G}(N)$ defined by $(v + w)(S) = v(S) + w(S)$ is again a game from $\mathcal{G}(N)$. Moreover, if multiplication of $v \in \mathcal{G}(N)$ by a real number α is defined by $(\alpha v)(S) = \alpha v(S)$, then αv also belongs to $\mathcal{G}(N)$. An important and well-known fact is that $\mathcal{G}(N)$ endowed with these

two algebraic operations is a real linear space.

Occasionally, we also need games whose set of players is different from N . In particular, for every game v from $\mathcal{G}(N)$ and every subset S of N , the *subgame* of v is the game $(S, v|_S)$ with player set S whose characteristic function $v|_S$ is defined by $v|_S(Q) = v(Q)$ for every subset Q of S .

We shall be concerned with the class of TU games in which the number of players is finite. For simplicity of notation, if the number of players is n , then we identify (name) the players with the first natural numbers $1, 2, \dots, n$. Accordingly, the grand coalition N is the set $\{1, 2, \dots, n\}$, the coalitions are subsets of $\{1, 2, \dots, n\}$, and the characteristic function v assigns to every subset S of $\{1, 2, \dots, n\}$ a real number $v(S)$.

SOLUTION FUNCTIONS

There is a variety of solution concepts for TU games. Some, like the cores or stable sets, consist of sets of real n -vectors, while others offer as a solution of a game, a single real n -vector. We restrict attention to the latter concept; that is, we are interested in a mapping $f = (f_1, f_2, \dots, f_n)$ that assigns to every game v from some collection of games in $\mathcal{G}(N)$ a real n -vector $f(v) = (f_1(v), f_2(v), \dots, f_n(v))$ and satisfies some prescribed desirable properties. Following the basic interpretation of values of characteristic functions as payoffs, we can interpret the values of components of such a solution function as payoffs to individual players in game v .

Definition 1. Let $\mathcal{A}$ be a nonempty collection of games from $\mathcal{G}(N)$. A *solution function on $\mathcal{A}$* is a mapping f from $\mathcal{A}$ into the n -dimensional real linear space $\mathcal{R}^n$. If the domain of f is not explicitly specified, then it is assumed to be $\mathcal{G}(N)$.

The collection of such mappings is too broad to contain only the mappings that lead to sensible solution concepts. Hence, to obtain

reasonable solution concepts we have to require that the solution functions have some reasonable properties. One of the natural properties in many contexts is the following property of efficiency.

Property 1 [Efficiency]. A solution function f on a subset $\mathcal{A}$ of $\mathcal{G}(N)$ is *efficient on $\mathcal{A}$* if

$$f_1(v) + f_2(v) + \cdots + f_n(v) = v(N) \quad (1)$$

for every game v from $\mathcal{A}$.

This property can be interpreted as a combination of the requirements of the *feasibility*

$$f_1(v) + f_2(v) + \cdots + f_n(v) \leq v(N)$$

and *collective rationality*

$$f_1(v) + f_2(v) + \cdots + f_n(v) \geq v(N).$$

Note that the efficiency demands that the grand coalition forms and distributes the total payoff $v(N)$ among all its players. This is a natural requirement in many situations. Nevertheless, we should keep in mind that occasionally, forming the grand coalition may not be efficient in an economic sense. For example, in some games there may be one or more subcoalitions of the grand coalition that have strictly greater worth than the grand coalition has.

In addition to satisfying the efficiency condition, solution functions are required to satisfy a number of other desirable properties. To introduce some of them, we need further definitions.

Player i from N is a *null player* in game v if

$$v(S \cup \{i\}) = v(S) \quad (2)$$

for every coalition S that does not contain player i ; that is, participation of a null player in a coalition does not contribute anything to the coalition in question.

Player i from N is a *dummy player* in game v if

$$v(S \cup \{i\}) = v(S) + v(\{i\}) \quad (3)$$

for every coalition S that does not contain player i ; that is, a dummy player contributes to every coalition the same amount, his or her value of the characteristic function.

Players i and j from N are *interchangeable* in game v if

$$v(S \cup \{i\}) = v(S \cup \{j\}) \quad (4)$$

for every coalition S that contains neither player i nor player j . In other words, two players are interchangeable if they can replace each other in every coalition that contains one of them.

Property 2 [Null player]. A solution function f satisfies the *null player property* if $f_i(v) = 0$ whenever $v \in \mathcal{G}(N)$ and i is a null player in v .

Property 3 [Dummy player]. A solution function f satisfies the *dummy player property* if $f_i(v) = v(\{i\})$ whenever $v \in \mathcal{G}(N)$ and i is a dummy player in v .

Property 4 [Equal treatment]. A solution function f satisfies the *equal treatment property* if $f_i(v) = f_j(v)$ for every $v \in \mathcal{G}(N)$ and every pair of players i, j that are interchangeable in v .

Properties (2)–(4) are quite reasonable, especially with regard to fairness and impartiality: a player who contributes nothing should get nothing; a player who contributes the same amount to every coalition cannot expect to get anything more than what he or she contributed; and two players who contribute the same to each coalition should be treated equally by the solution function.

The next property reflects the natural requirement that the solution function should be independent of the players' names. Let v be a game from $\mathcal{G}(N)$ and $\pi: N \rightarrow N$ be a permutation of N , and let the image of coalition S under π be denoted by $\pi(S)$. It is obvious that, for every $v \in \mathcal{G}(N)$, the function πv defined on $\mathcal{G}(N)$ by

$$(\pi v)(S) = v(\pi(S))$$

is again a game from $\mathcal{G}(N)$. Apparently, the game πv differs from game v only in players' names; they are interchanged by the permutation π .

Property 5 [Anonymity]. A solution function f is said to be *anonymous* if, for every permutation π of N ,

$$f_i(\pi v) = f_{\pi(i)}(v)$$

for every game $v \in \mathcal{G}(N)$ and every player $i \in N$.

Property 6 [Additivity]. A solution function f on $\mathcal{G}(N)$ is said to be *additive* if

$$f(u + v) = f(u) + f(v),$$

for every pair of games u and v from $\mathcal{G}(N)$.

It is natural for solution functions to be additive when a game consists of two independent games played separately by the same players or if a game is split into a sum of games. However, observe that the requirement of additivity differs from the other conditions in one important aspect. It involves two different games that may or may not be mutually dependent. In contrast, the dummy player and equal treatment properties involve only one game, and the anonymity property involves only games completely determined by a single game.

On some occasions, we need the properties to be valid only on a subset $\mathcal{A}$ of $\mathcal{G}(N)$. Then, when necessary, we require that πv , or $u + v$, belong to $\mathcal{A}$ whenever v , or u and v , are in $\mathcal{A}$.

Remark 1. The terminology introduced in the literature for various properties of players and solution functions is not completely standardized. For example, some authors use the term “dummy player” and “symmetric players” (or “substitutes”), for what we call null player and interchangeable players, respectively. Moreover, the term “symmetry” is sometimes used for our equal treatment and sometimes for our anonymity.

SHAPLEY VALUE

The Shapley solution function or briefly the Shapley value, proposed by Shapley [1] in 1953, is undoubtedly one of the most studied and most influential solution concepts for TU games. In addition to its theoretical significance, the Shapley value has a remarkably wide range of applications, particularly in economic and political models. Therefore, it is not surprising that there are several ways to introduce the Shapley value.

Basic Framework

Perhaps the simplest way of introducing the Shapley value is to define it explicitly by the following well-known formula for calculation of its components.

Definition 2. The Shapley value on a subset $\mathcal{A}$ of $\mathcal{G}(N)$ is a solution function φ on $\mathcal{A}$ whose components $\varphi_1(v), \varphi_2(v), \dots, \varphi_n(v)$ at game $v \in \mathcal{A}$ are defined by

$$\varphi_i(v) = \sum_{S: i \in S} \frac{(s-1)!(n-s)!}{n!} [v(S) - v(S \setminus \{i\})], \quad (5)$$

where the sum is meant over all coalitions S containing player i and s generically stands for the number of players in coalition S .

It can easily be seen that the components of the Shapley value can be computed also by the formula

$$\varphi_i(v) = \sum_{S: i \notin S} \frac{s!(n-s-1)!}{n!} [v(S \cup \{i\}) - v(S)], \quad (6)$$

where the sum is meant over all coalitions S not containing player i .

To clarify the basic idea behind this definition we first recall the notion of players' marginal contributions to coalitions.

Definition 3. For each player i and each coalition S , a *marginal contribution* of player i to coalition S in game v from $\mathcal{G}(N)$ is the number $m_i^v(S)$ defined by

$$m_i^v(S) = \begin{cases} v(S) - v(S \setminus \{i\}), & \text{if } i \in S; \\ v(S \cup \{i\}) - v(S), & \text{if } i \notin S. \end{cases}$$

Now imagine a procedure for dividing the total payoff $v(N)$ among the members of N , in which the players enter a room in some prescribed order and each player receives as payoff his or her marginal contribution to the coalition of players already in the room. Suppose that the prescribed order is $(\pi(1), \pi(2), \dots, \pi(n))$ where $\pi: N \rightarrow N$ is a fixed permutation of N . Then the procedure under consideration determines the payoffs to individual players as follows: Before the first player $\pi(1)$ entered the room, there was the empty coalition waiting in the room. And after player $\pi(1)$ enters, the coalition in the room becomes $\{\pi(1)\}$ and the player receives $v(\{\pi(1)\}) - v(\emptyset)$. Similarly, before the second player $\pi(2)$ entered, there was coalition $\{\pi(1)\}$ waiting in the room. After player $\pi(2)$ enters, coalition $\{\pi(1), \pi(2)\}$ is formed in the room and player $\pi(2)$ receives $v(\{\pi(1), \pi(2)\}) - v(\{\pi(1)\})$. This continues till the last player $\pi(n)$ enters and receives $v(N) - v(N \setminus \{\pi(n)\})$.

Let S_i^π denote the coalition of players preceding player i in the order $(\pi(1), \pi(2), \dots, \pi(n))$; that is, $S_i^\pi = \{\pi(1), \pi(2), \dots, \pi(j-1)\}$ where j is the uniquely determined member of N such that $i = \pi(j)$. Because there are $n!$ possible orders, the arithmetic average of the marginal contributions of player i taken over all possible orderings is equal to the number $(1/n!) \sum_\pi m_i^v(S_i^\pi)$, where the sum is understood over all permutations π of N . It can easily be verified by simple computation that this number is exactly the i th component of the Shapley value. Therefore, in addition to the equality of formulas (5) and (6), we also have the formula

$$\varphi_i(v) = \frac{1}{n!} \sum_\pi m_i^v(S_i^\pi) \quad (7)$$

for computing the components of the Shapley value.

If all orders of arrival of players are equally probable, then the probability that player i arrives after the players of S_i^π have already arrived and before the remaining players come, is equal to the number of possible orders of the form: first players in S_i^π (in some order), then player i , and then players in $N \setminus (S_i^\pi \cup \{i\})$ (in some order) divided by

the total number of possible orders of players in N . And this is equal to $(s-1)!(n-s)!/n!$ where s is the number of players in coalition $S = S_i^\pi \cup \{i\}$. Consequently, we conclude that the i th component of the Shapley value can be interpreted as the expected payoff to player i under this randomization scheme.

We have seen that the marginal contributions of players with respect to the order given by a fixed permutation π are given by

$$\begin{aligned} m_1^v(S_1^\pi) &= v(\{\pi(1)\}) - v(\emptyset), \\ m_2^v(S_2^\pi) &= v(\{\pi(1), \pi(2)\}) - v(\{\pi(1)\}), \\ &\vdots \\ m_n^v(S_n^\pi) &= v(\{\pi(1), \pi(2), \dots, \pi(n)\}) \\ &\quad - v(\{\pi(1), \pi(2), \dots, \pi(n-1)\}). \end{aligned}$$

It is clear that the sum of these contributions equals $v(N)$. It follows that the sum of the arithmetic averages of players' marginal contributions taken over all possible orderings must also add up to $v(N)$. Hence, the Shapley value satisfies the requirement of efficiency.

In addition to satisfying the condition of efficiency, the Shapley value has a number of other useful properties. In particular, it satisfies all of properties 2–6. Remarkably, no other solution function on $\mathcal{G}(N)$ satisfies the properties of null player, equal treatment, and additivity at the same time.

Theorem 1 [Shapley]. *For each N , there exists a unique solution functions on $\mathcal{G}(N)$ satisfying the properties of efficiency, null player, equal treatment, and additivity; this solution function is the Shapley value introduced by Definition 2.*

The standard proof of this basic result follows from the following facts:

- The collection $\{u_T : T \neq \emptyset, T \subset N\}$ of unanimity games defined by

$$u_T(S) = \begin{cases} 1, & \text{if } T \subset S; \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

form a base of the linear space $\mathcal{G}(N)$.

- The null player and equal treatment properties guarantee that φ is determined uniquely on multiples of unanimity games.
- The property of additivity (combined with the fact that the unanimity games form a basis) makes it possible to extend φ in a unique way to the whole space $\mathcal{G}(N)$.

Alternative Characterizations

One of the most useful aspects of the Shapley value is that it can be characterized by different sets of properties. For example, an easy modification of the previous theorem can be obtained by replacing the properties of null player and equal treatment by those of dummy player and anonymity.

The formulas (5)–(7) suggest that the Shapley value of a given player at a given game depends only on the vector of his or her marginal contributions in the game under consideration. This is useful in applications because in practice only one problem at a time, represented by some specific game, is usually considered. However, unlike the efficiency, equal treatment, and null player properties, the requirement of additivity involves different games that may or may not be mutually unrelated. This, at least partly, can explain why a considerable portion of research effort in finding alternative characterizations has been focused on the possibility of eliminating the condition of additivity.

The efficiency, equal treatment, and null player property are sufficient to determine the Shapley value uniquely for some games but, in general, they are not sufficient to determine the Shapley value on $\mathcal{G}(N)$. Therefore, just omitting the condition of additivity will not help.

Monotonicity and Marginality. Young [2] introduced the following monotonicity property for solution functions that makes it possible to characterize the Shapley value without using the condition of additivity.

Property 7 [Strong monotonicity]. A solution function f is said to be *strongly monotone* if, for every player i , $f_i(u) \geq f_i(v)$

whenever $m_i^u(S) \geq m_i^v(S)$ for every coalition S .

Theorem 2 [Young]. *For each N , the Shapley value is the only solution function on $\mathcal{G}(N)$ satisfying efficiency, strong monotonicity, and equal treatment.*

Moreover, Young [3] also showed that the Shapley value is the unique anonymous efficient solution function that satisfies the following marginality principle.

Property 8 [Marginality principle]. A solution function f is said to satisfy the *marginality principle* if $m_i^u = m_i^v$ implies $f_i(u) = f_i(v)$ for every player i and every two games u and v from $\mathcal{G}(N)$.

Potential. Both of Young's characterizations of the Shapley value also involve the characteristic functions of different, possibly independent games. The following approach proposed by Hart and Mas-Colell [4] shows that it is possible to characterize the Shapley value of a game by conditions involving only games determined by the values of the characteristic functions of the game itself.

To describe these results we first recall that for each game v from $\mathcal{G}(N)$, the functions $v|_S$ defined on the collection of all subcoalitions Q of coalition S by $v|_S(Q) = v(Q)$ are called subgames of game v . Let G_v denote the set of all subgames of game v and let p be a real-valued function on G_v . We introduce the notions of players' marginal contributions according to p , and of a potential function on G_v as follows.

Definition 4. The *marginal contribution* of a player i from a coalition S according to a function p on G_v is the difference $D_i^p(S) = p(v|_S) - p(v|_{S \setminus \{i\}})$. If p is such that $p(v|_\emptyset) = 0$ and $\sum_{i \in N} D_i^p(N) = v(N)$, then p is called a *potential function*.

Theorem 3 [Hart and Mas-Colell]. *For every game v in $\mathcal{G}(N)$, there exists a unique*

potential function p on G_v , and the resulting vector

$$(D_1^p(N), D_2^p(N), \dots, D_n^p(N))$$

of players' marginal contributions according to p coincides with the Shapley value of game v .

Because the potential of every game is determined only by the game and its subgames, we have another characterization of the Shapley value without requiring additivity. Moreover, the equality $D_i^p(N) = \varphi_i(v)$ provides an alternative way for computing the components of the Shapley value. Indeed, by rewriting the efficiency condition $\sum_{i \in N} D_i^p(N) = v(N)$ into

$$p(v|_N) = \frac{1}{n} \left(v(N) + \sum_{i \in N} p(v|_{N \setminus \{i\}}) \right), \quad (9)$$

and starting with the initial condition $p(v|_\emptyset) = 0$, we obtain a recursive way for computing the values of the potential function appearing in the definition of $D_i^p(N)$, $i = 1, 2, \dots, n$.

SHAPLEY VALUE FOR SUBCLASSES

Originally, Shapley did not consider explicitly the linear space $\mathcal{G}(N)$ of all games with a fixed finite set of players but the collection of superadditive games with an arbitrary finite number of players taken from a potentially infinite set of players, where the *superadditivity* of a game v means that the inequality $v(S) + v(T) \leq v(S \cup T)$ is satisfied for every pair of disjoint coalitions S and T .

The sum of two superadditive games is again a superadditive game and a multiple of a superadditive game by a nonnegative number is also a superadditive game; in other words, the set of superadditive games from $\mathcal{G}(N)$ forms a cone in $\mathcal{G}(N)$. Because the superadditive games belong to $\mathcal{G}(N)$, it is evident that the properties that characterize the Shapley value on $\mathcal{G}(N)$ are also satisfied by superadditive games, provided that the property under consideration is applicable under the assumption of superadditivity.

Less obvious (but easily verifiable) is the fact that these properties also guarantee the uniqueness of the Shapley value on the cone of superadditive games in $\mathcal{G}(N)$. Because various special classes of games occur in practice, it is of interest to identify other subclasses of games from $\mathcal{G}(N)$ on which the Shapley value can be characterized by the same or similar properties as in the case of $\mathcal{G}(N)$.

Additive Groups of Games

Neyman [5] demonstrated that a characterization of the Shapley value by the properties of efficiency, null player, equal treatment, and additivity is possible on the set of linear combinations with integer-valued coefficients of games determined by a fixed game from $\mathcal{G}(N)$.

For a game v from $\mathcal{G}(N)$, let $\mathcal{A}_v$ be the collection of all games from $\mathcal{G}(N)$ defined by the condition that $w \in \mathcal{A}_v$ if and only if there are a positive integer m , integers (not necessarily positive) $\alpha_1, \alpha_2, \dots, \alpha_m$ and coalitions $S_1, S_2, \dots, S_m$ such that

$$w = \sum_{k=1}^m \alpha_k v_{S_k},$$

where, for each k , the game v_{S_k} is defined by $v_{S_k}(T) = v(S_k \cap T)$, $T \subset N$.

Theorem 4 [Neyman]. *If f is a solution function on $\mathcal{A}_v$ satisfying efficiency, additivity, equal treatment, and the null player property, then it is the Shapley value on $\mathcal{A}_v$.*

Theorem 5 [Neyman]. *If f is a solution function on $\mathcal{A}_v$ satisfying efficiency, strong monotonicity, and the equal treatment property, then it is the Shapley value on $\mathcal{A}_v$.*

Simple Games

The set of games from $\mathcal{G}(N)$ whose characteristic functions assume only values 0 or 1, and some of its subsets, appear naturally in everyday life. Such games can be viewed as models of yes–no voting systems; that is, systems whose rules specify which collections of

“yes” votes guarantee passage of the issue at hand. In conformity with this interpretation, a set S of voters is said to be a *winning* coalition if passage of the issue is guaranteed by “yes” votes from S . Naturally, coalitions that are not winning are called *losing* coalitions. The corresponding voting game v is then defined by setting $v(S) = 1$ if S is winning, and $v(S) = 0$ if S is losing. We refer to such cooperative games as *simple* games. The solution functions defined on the set (or on some specific subsets) of simple games are often called *indices of power*; see the article titled **Power Indices**.

Remark 2. In this generality, a simple game can be considered as a pair (N, W) where W is a given collection of coalitions called winning coalitions. Such objects are often called hypergraphs. To distinguish the simple games from hypergraphs, many authors exclude some games from what we call simple games. For example, Taylor and Zwicker [6] reserve the term “simple game” for a simple game that is monotone in the sense that every supercoalition of a winning coalition is also winning. Sometimes, only the identically zero game is excluded; sometimes it is required that the grand coalition is winning and all singleton coalitions are losing.

In contrast to the case of superadditive games, the sum of two simple games need not be a simple game. Therefore, it is immediate that the requirement of additivity needs modification. One possibility is to require additivity only for those pairs of simple games whose sum is also a simple game. Dubey [7] showed that this modification is sufficient for proving the uniqueness of Shapley’s value on the class of all simple games.

The situation is different when we consider not identically zero monotonic simple games. First we observe that, in monotonic simple games, the players’ marginal contributions also assume only the values from the set $\{0, 1\}$. Specifically, $m_i^v(S) = 1$ only in two cases: either S is losing and $S \cup \{i\}$ is winning, or S is winning and $S \setminus \{i\}$ is losing. It follows that the formulas (5)–(7) can be simplified. For example, formula (5) can be simplified to

$$\varphi_i(v) = \sum \frac{(s-1)!(n-s)!}{n!}, \quad (10)$$

where the sum is meant over all winning coalitions S containing player i and such that coalition S_i and all its subcoalitions are losing.

It follows from the monotonicity that for every permutation π , there is exactly one player such that the coalition S_i^π of players preceding player i in the order given by π is losing and the coalition $S_i^\pi \cup \{i\}$ is winning. We say that this player is *pivotal* for π . Recalling the room example with orders of equal probability, we can easily see that Equation (10) can be expressed in the form

$$\varphi_i(v) = \frac{\text{the number of permutations for which } i \text{ is pivotal}}{\text{the number of all permutations}}. \quad (11)$$

As a rule, this specification of the Shapley value to the case of monotonic simple games is called the *Shapley Shubik power index*.

In contrast to the general case, the null player and equal treatment properties, together with efficiency and the mentioned modification of additivity, do not guarantee uniqueness of the Shapley value on the set of monotonic simple games. To deal with this situation, Dubey proposed replacing additivity of the solution function by its modularity.

To present this modification, we first notice that for every pair of monotonic simple games u and v , the games $\max(u, v)$ and $\min(u, v)$ defined by

$$\max(u, v)(S) = \max\{u(S), v(S)\},$$

$$\min(u, v)(S) = \min\{u(S), v(S)\},$$

are also monotonic simple games. It is easily seen that a coalition is winning in game $\max(u, v)$ if it is winning in game u or game v , and it is winning in game $\min(u, v)$ if it is winning in both u and v . Thus, a coalition wins in games u and v as often as in games $\max(u, v)$ and $\min(u, v)$.

Property 9 [Modularity]. A solution function f on the set of simple games is *modular*

if

$$f(u) + f(v) = f(\max(u, v)) + f(\min(u, v)),$$

for every pair of simple games u and v .

Theorem 6 [Dubey]. *The Shapley Shubik index is a unique solution function on the class of monotonic simple games that is efficient and has the dummy player, anonymity, and modularity properties.*

SOLUTIONS WITH COOPERATION STRUCTURES

The class of simple games considered in the previous section arises by restricting the range of characteristic functions. On the other hand, there are situations in which it may be reasonable to relax the condition (tacitly assumed in the efficiency requirement) that only the grand coalition forms. As examples we can take the formation of cartels, trading blocs, political parties or various joint ventures. Such relaxations are usually modeled by adjoining to a game some structure on the set of coalitions. We shall call such structures *cooperation structures*. From the wide variety of such structures appearing in the literature, we consider only the structures given by the requirement that each player belongs exactly to one coalition, and the structures given by graphs on the set of players.

Aumann Drèze Value

We begin with the cooperation structures given by partitions of the grand coalition.

Definition 5. A *coalition structure* is a partition of the grand coalition; that is, a collection of pairwise disjoint nonempty coalitions whose union is the grand coalition.

Probably the first extension of the Shapley value to games with a coalition structure was proposed by Aumann and Drèze [8] who analyzed the idea that, for a game v and a coalition structure $\mathcal{C}$, the payoff to player i should be the Shapley value of player i in the game restricted to subcoalitions of the

coalition that contains player i and belongs to partition $\mathcal{C}$.

Let such a payoff to player i be denoted by $\Phi_i(v, \mathcal{C})$. Then, this suggestion requires that

$$\Phi_i(v, \mathcal{C}) = \varphi_i(v, S),$$

where coalition S is a member of partition $\mathcal{C}$, player i belongs to S , and $\varphi_i(v, S)$ is defined by

$$\begin{aligned} \varphi_i(v, S) = \sum_{T \subset S \setminus \{i\}} \frac{t!(s-t-1)!}{s!} \\ \times [v(T \cup \{i\}) - v(T)]. \end{aligned} \quad (12)$$

It is worth noticing that $\Phi_i(v, \mathcal{C})$ reduces to the ordinary Shapley value of player i when $\mathcal{C}$ is the trivial partition $\{N\}$ and that, in general, the sum $\sum_{i=1}^n \Phi_i(v, \mathcal{C})$ may be different from $v(N)$. Thus $\Phi = (\Phi_1, \dots, \Phi_n)$ is not necessarily efficient. However, it is evident that the following property of relative efficiency is satisfied.

Property 10 [Relative efficiency]. If $\mathcal{C} = \{C_1, \dots, C_m\}$ is a coalition structure for N and v is a game from $G(N)$, then

$$\sum_{j \in C_k} \Phi_j(v, \mathcal{C}) = v(C_k),$$

for every $1 \leq k \leq m$.

Definition 6. Let $\Pi(N)$ be the collection of all partitions of N and let $\mathcal{B}$ be a subset of $\Pi(N)$. A *coalitional solution function* on a subset $\mathcal{A}$ of $\mathcal{G}(N)$ with a coalition structure from $\mathcal{B}$ is a mapping $f = (f_1, f_2, \dots, f_n)$ from the Cartesian product $\mathcal{A} \times \mathcal{B}$ into n -dimensional real linear space $\mathcal{R}^n$.

Aumann and Drèze showed that $\Phi = (\Phi_1, \Phi_2, \dots, \Phi_n)$ defined by formula (12) is the unique relatively efficient coalitional solution function on $\mathcal{G}(N) \times \Pi(N)$ that satisfies the following natural extensions of the null player, anonymity, and additivity properties.

- **Null Player:** If i is a null player, then $f_i(v, \mathcal{C}) = 0$ for every pair $v \in \mathcal{G}(N)$ and $\mathcal{C} \in \Pi(N)$.

- *Anonymity*: For every permutation π on N and every coalition structure $\mathcal{C} = \{C_1, \dots, C_m\}$ from $\Pi(N)$, we have $f(\pi v, \pi \mathcal{C}) = \pi f(v, \mathcal{C})$ where $\pi \mathcal{C} = \{\pi(C_1), \pi(C_2), \dots, \pi(C_m)\}$ and $\mathcal{C}$ is invariant under π in the sense that $\pi(C_k) = C_k$ for all $1 \leq k \leq m$.
- *Additivity*: If u and v are games from $\mathcal{G}(N)$, then $f(u + v, \mathcal{C}) = f(u, \mathcal{C}) + f(v, \mathcal{C})$ for every coalition structure $\mathcal{C} \in \Pi(N)$.

Moreover, Myerson [9] proved that the Aumann–Dr ze generalization of the Shapley value is the unique relatively efficient coalitional solution function satisfying the condition:

$$\varphi_i(v, S) - \varphi_i(v, S \setminus \{j\}) = \varphi_j(v, S) - \varphi_j(v, S \setminus \{i\}), \quad (13)$$

for all i and j in S .

Owen Value

A different extension of the Shapley value to games with coalition structures was introduced by Owen [10]. As in the Shapley value, players get their expected marginal contributions with respect to random orders of players, but the orders are required to be consistent with the coalition structure.

For a given coalition structure $\mathcal{C} = \{C_1, \dots, C_m\}$, let $\Omega_{\mathcal{C}}$ be the set of permutations π on N that are consistent with $\mathcal{C}$ in the sense that if $C_k \in \mathcal{C}$, $i, j \in C_k$, and $\pi(i) \leq \pi(l) \leq \pi(j)$, then l is in C_k .

Definition 7. The *Owen value* of player i in game v with coalition structure $\mathcal{C}$ is defined by

$$\phi_i(v, \mathcal{C}) = \frac{1}{\omega} \sum_{\pi \in \Omega_{\mathcal{C}}} [v(C_i^{\pi} \cup \{i\}) - v(C_i^{\pi})], \quad (14)$$

where ω is the number of permutations in $\Omega_{\mathcal{C}}$.

Notice that the Owen value is an efficient coalitional solution function; that is, $\sum_{i=1}^n \phi_i(v, \mathcal{C}) = v(N)$ for every v and $\mathcal{C}$. Also notice that the Owen value is equal to the Shapley value when $\mathcal{C} = \{\{1\}, \{2\}, \dots, \{n\}\}$.

Like the Shapley value, the Owen value can be characterized in several ways. For example, the Owen value is the only efficient coalitional solution function on $\mathcal{G}(N) \times \Pi(N)$ that is additive, has the null player property, and satisfies the following two symmetry requirements: For a game v and a coalition structure $\mathcal{C} = \{C_1, \dots, C_m\}$, we first define a “quotient” game $v_{\mathcal{C}}$ whose players are the coalitions in the original game v that are members of the partition $\mathcal{C}$. Formally, we take as the set of players the set $M = \{1, 2, \dots, m\}$ and define $v_{\mathcal{C}}$ by

$$v_{\mathcal{C}}(T) = v(\cup_{k \in T} C_k) \text{ for every } T \subset M.$$

Then, the mentioned symmetry requirements are defined in the following way.

- *Symmetry across Coalitions*: If players k and l in game $v_{\mathcal{C}}$ are symmetric in $v_{\mathcal{C}}$, then the total payoff to coalition C_k is the same as the total payoff to coalition C_l .
- *Symmetry within Coalitions*: For any two players i and j who are symmetric in game v and belong to the same member of coalition structure $\mathcal{C}$, the values of i and j are equal.

Recently, Khmelnitskaya and Yanovskaya [11] characterized the Owen value without using additivity. Namely, they showed that the additivity and null player properties in the characterization presented above can be replaced by the marginality principle.

Myerson Value

The third extension of the Shapley value to games with a cooperation structure is that proposed by Myerson [12] for a cooperation structure given by an undirected simple (that is, without self-loops and parallel edges) graph on the set of players. The edges of such a *cooperation graph* are often interpreted as communication channels between those players that can directly negotiate with each other, but various other interpretations are possible.

Let E be a cooperation graph. We say that players i and j are connected in coalition S

by E if there is a path in E between i and j that stays in S . Let S/E denote the partition of S into the collections of players that are connected in S by E , and define a game v^E by $v^E(S) = \sum_{T \in S/E} v(T)$.

Myerson established that the function $\phi^E = (\phi_1^E, \phi_2^E, \dots, \phi_n^E)$ defined by $\phi_i^E(v) = \varphi_i(v^E)$ (where φ denotes the ordinary Shapley value) is the only solution function satisfying the following two conditions:

- If S is a connected component of a cooperation graph E , then

$$\sum_{i \in S} \phi_i^E(v) = v(S).$$

- If E is a cooperation graph in which players i and j are not connected and H is the cooperation graph obtained from E by adding the edge that connects players i and j , then

$$\phi_i^H(v) - \phi_i^E(v) = \phi_j^H(v) - \phi_j^E(v).$$

It turns out that the Myerson value coincides with the ordinary Shapley value if the cooperation graph is the complete graph on N , and that the differences in the second condition are nonnegative when v is superadditive.

PROBABILISTIC VALUES

Let us fix some player i and, for every coalition S that does not contain player i , denote by $\alpha_i(S)$ the number $s!(n-s-1)!/n!$ where, as previously, the letter s generically denotes the number of players in coalition S . It is a simple exercise to verify that $0 \leq \alpha_i(S) \leq 1$ for every coalition S not containing player i , and that $\sum_{S \subset N \setminus \{i\}} \alpha_i(S) = 1$. Put in other terms, the family $\{\alpha_i(S) : S \subset N \setminus \{i\}\}$ is a probability distribution over the set of coalitions not containing player i . Now recall, from formula (6), that the i th component of the Shapley value can be computed by

$$\varphi_i(v) = \sum_{S \subset N \setminus \{i\}} \alpha_i(S) [v(S \cup \{i\}) - v(S)].$$

This shows that the i th component of the Shapley value is the expected marginal contribution of player i with respect to the probability measure $\{\alpha_i(S) : S \subset N \setminus \{i\}\}$ and that the Shapley value belongs to the following class of solution functions.

Definition 8. A solution function f on a subset $\mathcal{A}$ of $\mathcal{G}(N)$ is called *probabilistic on $\mathcal{A}$* if, for each player i , there exists a probability distribution $\{p_i(S) : S \subset N \setminus \{i\}\}$ on the collection of coalitions not containing i such that

$$f_i(v) = \sum_{S \subset N \setminus \{i\}} p_i(S) [v(S \cup \{i\}) - v(S)], \quad (15)$$

for every $v \in \mathcal{A}$.

The family of probabilistic solution functions embraces an enormous number of functions [13]. The efficient probabilistic solution functions are often called *quasivalues*, and the anonymous probabilistic solution functions are called *semivalues*. Since the Shapley value is anonymous and efficient on $\mathcal{G}(N)$, we know that it is both a quasivalue and a semivalue on $\mathcal{G}(N)$. Moreover, the Shapley value is the only probabilistic solution function with these properties.

Theorem 7 [Weber]. *If N has at least three elements, then the Shapley value is the unique probabilistic solution function on $\mathcal{G}(N)$ that is anonymous and efficient.*

Another widely known probabilistic solution function is the function proposed originally only for voting games by Banzhaf [14].

Definition 9. The *Banzhaf value* on $\mathcal{G}(N)$ is a solution function ψ on $\mathcal{G}(N)$ whose components at game v are defined by

$$\psi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{1}{2^{n-1}} [v(S \cup \{i\}) - v(S)]. \quad (16)$$

Again, by simple computation, we can verify that the Banzhaf value is a probabilistic solution function. Consequently, the i th component of the Banzhaf value is the expected marginal contribution of player i with respect to the probability measure $\{\beta_i(S) : S \subset N \setminus \{i\}\}$.

$\{i\}$, where $\beta_i(S) = 1/2^{n-1}$ for each subset S of $N \setminus \{i\}$.

From the probabilistic point of view, the Banzhaf value is based on the assumption that each player i is equally likely to join any subcoalition of $N \setminus \{i\}$. On the other hand, the Shapley value is based on the assumption that the coalition the player i enters is equally likely to be of any size s , $0 \leq s \leq n-1$, and that all coalitions of this size are equally likely.

The following property of total power has an important role in some characterizations of the Banzhaf value.

Property 11 [Total power]. A solution function f on a subset $\mathcal{A}$ of $\mathcal{G}(N)$ satisfies the *total power property* on $\mathcal{A}$ if

$$\sum_{i=1}^n f_i(v) = \frac{1}{2^{n-1}} \sum_{i=1}^n \sum_{S \subset N \setminus \{i\}} [v(S \cup \{i\}) - v(S)],$$

for every v from $\mathcal{A}$.

For example, we have the following analogs to Theorems 1 and 2:

- The only solution function on $\mathcal{G}(N)$ that satisfies additivity, null player, equal treatment, and total power is the Banzhaf value.
- The only solution function on $\mathcal{G}(N)$ that satisfies strong monotonicity, equal treatment, and total power is the Banzhaf value.

It is clear from formula (16) that the Banzhaf value is a semivalue on $\mathcal{G}(N)$. It is known that, if a solution function f is a semivalue on $\mathcal{G}(N)$, then, for each player i and each coalition $S \subset N \setminus \{i\}$, the corresponding $p_i(S)$ in formula (15) depends only on the number of players in S . Therefore, there exists a function γ from $\{0, 1, \dots, n-1\}$ into the unit interval $[0, 1]$ such that $p_i(S) = \gamma(s)$ for every i . We have seen that $\gamma(s) = s!(n-s-1)!/n!$ for the Shapley value, and $\gamma(s) = 1/2^{n-1}$ for the Banzhaf value.

Because the Banzhaf value is not efficient when $n > 2$, it is not suitable as the tool for

sharing benefits or costs among the members of the grand coalition. However, it may compete with the Shapley–Shubik power index when considered on various classes of simple games, for which it was originally designed. Then, it is known under different names, most often as the Banzhaf index, Banzhaf absolute index, Banzhaf–Coleman index, or Penrose–Banzhaf index.

As in the case of Shapley–Shubik index, the defining equality of formula (16) of the Banzhaf value can be simplified when considered on the class of the monotonic simple games. Let v be a monotonic simple game and let a *swing* for player i in v be defined as a pair of coalitions $(S, S \cup \{i\})$ such that S is losing and $S \cup \{i\}$ is winning. Then it is evident that formula (16) simplifies to the following analog of formula (11):

$$\psi_i(v) = \frac{\text{the number of swings for } i}{2^{n-1}}. \quad (17)$$

APPLICATIONS

Recall that the concept of a solution function on $\mathcal{G}(N)$ is just a mapping from $\mathcal{G}(N)$ into n -dimensional real linear space $\mathcal{R}^n$, and observe that every point $x = (x_1, x_2, \dots, x_n)$ in $\mathcal{R}^n$ defines a game v_x on $\mathcal{G}(N)$ by setting $v_x(S) = \sum_{i \in S} x_i$ for each coalition S . Observe that, for every x , game v_x is *additive* in the sense that, for every pair of disjoint coalitions S and T , we have the equality $v_x(S) + v_x(T) = v_x(S \cup T)$. Clearly, the collection of additive games from $\mathcal{G}(N)$ is a linear subspace of $\mathcal{G}(N)$. It follows that the solution functions can also be defined as mappings from $\mathcal{G}(N)$ into the space of additive games from $\mathcal{G}(N)$, and the Shapley value can be considered as a sensible way of transforming games from various subclasses of $\mathcal{G}(N)$ into additive games. As this or similar structures appear naturally in a variety of both practical situations and theoretical disciplines, the number of applications of the Shapley value and its variations is very large, varied, and continuously growing. The richness and diversity of these applications are convincingly illustrated in a recent survey by Moretti and Patrone [15].

Here, we briefly mention only the use of the Shapley value for the resolution of cost-sharing problems. In general, a standard game-theoretic approach to a cost-sharing problem is to design for it an appropriate cooperative game, then to analyze and solve the game in some sense and finally, to explicate the obtained solution in terms of the original problem. Typically, the players in such a game are customers of a public service, or parties who want to divide fairly the cost of establishing a common facility or undertaking a joint project.

For example, suppose that the grand coalition N is a group of potential customers of some public service, and the relevant cost data are represented through a joint cost function c defined for all subsets of N , where $c(S)$ is the cost of providing the service to the members of coalition S and where the cost of serving no one is zero. Such a game is usually called a *cost-sharing* game, and its characteristic function is denoted by c instead of the usual v .

Of course, every cost-sharing game is a usual TU game but we should note that some desirable properties of the cost-sharing games may differ from those of games in which the values of the characteristic functions are interpreted as the worth or payoff. For example, the cost-sharing games arising in practice are often not superadditive but subadditive; that is, the inequality $c(S \cup T) \leq c(S) + c(T)$ holds for every pair of disjoint coalitions S and T . However, it is always possible to associate with a cost-sharing game c another game v , called a *cost-saving* game, whose values $v(S)$ are given by $v(S) = \sum_{i \in S} c(\{i\}) - c(S)$. Then, v is superadditive if and only if the original cost-sharing game is subadditive.

Every efficient solution function of a cost-sharing game c can be interpreted in an obvious way as a method of sharing the total cost $c(N)$. If the players are to be treated fairly, then the null player and equal treatment properties of the Shapley value have a natural and attractive interpretation in the cost-sharing context:

- *Null Player*: If the cost of serving any subset of players is the same whether

a player i from N belongs to the subset or not, then player i should be charged nothing.

- *Equal Treatment*: Two players for whom the marginal cost is the same, no matter which subset of customers is served, should be charged equally.

Also, the additivity has a natural interpretation: the payment of every player for two services should be the sum of the separate payments for the two services. Thus, the Shapley value can be viewed as one of the appropriate tools for solving cost-sharing problems.

FURTHER READING

A treatment of the Shapley value for TU games is presented in almost all game theory textbooks [16–20]. However, it is almost always beneficial to also examine the origins of ideas. The original papers by Shapley and Shubik, together with a number of important subsequent papers (up to 1998) of other renowned authors, are accessible in the collection of essays edited by Roth [21]. Most of the later development can be found in Chapter 53 (by Eyal Winter) and Chapter 54 (by Dov Monderer and Dov Samet) of the third volume of the Handbook of Game Theory [22] and in the book by Peleg and Sudhölter [19]. The reader interested mainly in applications will find sufficient information in Refs 15, 23, and 24.

REFERENCES

1. Shapley LS. A value for n -person games. In: Kuhn H, Tucker AW, editors. Contributions to the theory of games 2, annals of mathematics studies 28. Princeton (NJ): Princeton University Press; 1953. pp. 307–317.
2. Young HP. Monotonic solutions of cooperative games. Int J Game Theory 1985;14(2):65–72.
3. Young HP. Individual contribution and just compensation. In: Roth AE, editors. The shapley value. Cambridge: Cambridge University Press; 1988. pp. 267–278.
4. Hart S, Mas-Colell A. Potential, value, and consistency. Econometrica 1989;57(3): 589–614.

5. Neyman A. Uniqueness of the shapley value. *Games Econ Behav* 1989;1:116–118.
6. Taylor AD, Zwicker WS. Simple games: desirability relations, trading, pseudoweightings. Princeton (NJ): Princeton University Press; 1999.
7. Dubey P. On the uniqueness of the shapley value. *Int J Game Theory* 1975;4(3):131–139.
8. Aumann RJ, Drèze JH. Cooperative games with coalition structure. *Int J Game Theory* 1975;4:217–237.
9. Myerson RB. Conference structures and fair allocation rules. *Int J Game Theory* 1980;9:169–182.
10. Owen G. Values of games with a priori unions. In: Henn R, Moeschlin O, editors. *Essays in mathematical economics and game theory*. Berlin: Springer; 1977. pp. 76–88.
11. Khmelnitskaya AB, Yanovskaya EB. Owen coalitional value without additivity axiom. *Math Methods Oper Res* 2007;66:255–261. DOI: 10.1007/s00186-006-0119-8.
12. Myerson RB. Graphs and cooperation in games. *Math Methods Oper Res* 1977;2:225–229.
13. Weber R. Probabilistic values for games. In: Roth AE, editor. *The shapley value*. Cambridge: Cambridge University Press; 1988. pp. 101–120.
14. Banzhaf JF III. Weighted voting does not work: a mathematical analysis. *Rutgers Law Rev* 1965;19:317–343.
15. Moretti S, Patrone F. Transversality of the shapley value. *TOP* 2008;16(1):1–41. DOI: 10.1007/s11750-008-0044-5.
16. Owen G. *Game theory*. 3rd ed. San Diego (CA): Academic Press; 1995.
17. Osborne MJ, Rubinstein A. *A course in game theory*. Cambridge: MIT Press; 1994.
18. Myerson RB. *Game theory: analysis of conflict*. Cambridge: Harvard University Press; 1991.
19. Peleg B, Sudhölter P. *Introduction to the theory of cooperative games*. 2nd ed. Berlin Heidelberg: Springer; 2007.
20. Gura E, Maschler M. *Insight into game theory*. New York: Cambridge University Press; 2008.
21. Roth AE, editor. *The Shapley Value: Essays in honor of Lloyd S. Shapley*. Cambridge: Cambridge University Press; 1988.
22. Aumann R, Hart S, editors. Volume 3, *Handbook of game theory*. Amsterdam: Elsevier (North Holland); 2002. (reprinted 2007, 2008).
23. Young HP. Cost allocation. In: Aumann RJ, Hart S, editors. Volume 2, *Handbook of game theory*. Amsterdam: Elsevier (North Holland); 2002. pp. 1193–1235.
24. Young HP. *Equity: in theory and practice*. Princeton (NJ): Princeton University Press; 1994.

THE SIMPLEX METHOD AND ITS COMPLEXITY

NAGABHUSHANA PRABHU
School of Industrial Engineering, Purdue
University, West Lafayette, Indiana

Primal linear programming (LP)¹ algorithms can be divided into three families²—the *interior-point* algorithms, the *surface* algorithms, and the *exterior-point* algorithms—depending on whether the path traced by the algorithm lies inside, on, or outside the feasible region [2,3]. The simplex method, devised by Dantzig, is the most prominent surface LP algorithm. Although, the simplex method is more than five decades old, fundamental questions about the underlying geometry of the simplex method remain unanswered, making it a subject of continuing active investigation.

This article presents a brief overview of the simplex method and its complexity in the first and second sections. The latter part of the article, the third section, presents a summary of the ongoing research on the complexity of simplex-like edge-following LP algorithms. This is an introductory overview slanted toward a discussion of the outstanding problems related to the geometry of simplex method and, therefore, the details of the simplex method are presented only to the extent the details are required for the later discussion of the underlying geometric problems. The material on simplex method and its complexity, presented below, is standard lore and the reader is referred to textbooks on LP, particularly Chvatal [1], Dantzig [4], and Dantzig and Thapa [2,5], and the references therein for a more expansive discussion of the topics and pointers to the original papers.

¹Linear programming and linear program will both be abbreviated to LP.

²LP can be formulated either as an optimization or a solvability problem [1]. We focus on the optimization version.

THE SIMPLEX METHOD

A general LP can be formulated in the *canonical form* as the problem of computing

$$\min \left\{ \mathbf{c}^T \mathbf{x} \mid \mathbf{A} \mathbf{x} \leq \mathbf{b}; \mathbf{x} \geq 0 \right\},$$

where $\mathbf{c} \in \mathbb{R}^d$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{A}$ is an $m \times d$ real matrix. $\mathbf{c}^T \mathbf{x}$ is called the *objective function* of the LP and the inequalities the *constraints*. Each of the inequalities corresponds to a *half-space* in $\mathbb{R}^d$, and the intersection of the finite number of half-spaces is the *polyhedral* feasible region of the LP. The points in the feasible polyhedron are called the *feasible solutions* and the points at which the objective function attains its minimum value are the *optimal solutions*. In general, the underlying polyhedron can be bounded or unbounded and can be k -dimensional³ for $-1 \leq k < d$; in the following discussion we will assume that the polyhedron is bounded and d -dimensional, that is, the polyhedron is a *d-polytope*. An inequality is said to be *redundant* if dropping the inequality does not change the feasible region; in the following discussion we will make the simplifying assumption that the LP does not have any redundant constraints, in which case the boundary of each constraint determines a *facet* of the polyhedron; see Ziegler [6] for discussion on polytopes and their boundary structure.

Introducing slack variable, $\mathbf{s}$, the above *canonical form* LP can be rewritten in *standard form* as follows:

$$\min \left\{ \mathbf{c}^T \mathbf{x} \mid \mathbf{A} \mathbf{x} + \mathbf{I} \mathbf{s} = \mathbf{b}; \mathbf{x}, \mathbf{s} \geq 0 \right\}$$

If the canonical form LP does not have redundant constraints then the facets of the polyhedron correspond to the variables in the standard form LP as illustrated in the following example. Consider the LP whose

³An empty polyhedron is said to have dimension -1 .

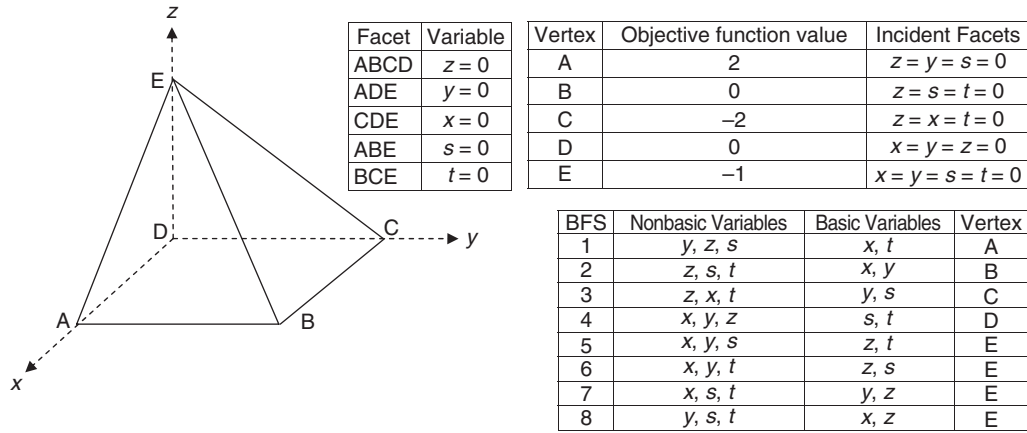


Figure 1. The polyhedron, variable-facet correspondence and the basic feasible solutions for the LP in Equation (1).

canonical and standard forms are

$$\begin{aligned} &\text{Minimize} && [2 \ -2 \ -1] \mathbf{x} \\ &\text{S.T.} && \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &&& \mathbf{x} \geq \mathbf{0} \end{aligned} \quad (1)$$

$$\begin{aligned} &\text{Minimize} && [2 \ -2 \ -1] \mathbf{x} \\ &\text{S.T.} && \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{s} \end{bmatrix} \\ &&& = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &&& \mathbf{x}, \mathbf{s} \geq \mathbf{0} \end{aligned}$$

where $\mathbf{x}^T = [xyz]$; $\mathbf{s}^T = [st]$. The feasible region of the canonical form LP is shown in Fig. 1.

Setting a variable of the standard form LP to zero forces the solution to lie on the facet corresponding to the variable as shown in the table in Fig. 1. Extending the argument, a solution of the standard form LP that represents a vertex can be obtained by insisting that the solution lie on all of the facets incident at the vertex, or in other words by setting the corresponding variables to zero; see the tables in Fig. 1. The solutions of the standard form LP that correspond to vertices are called the *basic feasible solutions* (BFSs). A minimal subset of variables that

need to be set to zero to determine the vertex constitutes the set of *nonbasic variables* of the corresponding BFS and the remaining variables are the *basic variables* of the BFS. The eight BFSs of the above LP and the basic and nonbasic variables in each BFS are tabulated in Fig. 1. The correspondence between BFSs of the standard form LP and the vertices of the polyhedron is many-to-one; for example, vertex E in Fig. 1 corresponds to four different BFSs. A vertex, such as E, that is represented by more than one BFS is called a *degenerate vertex*.

The simplex method exploits the property that *if the extreme value of a linear objective function over a polyhedron is finite then the set of optimal solutions must include at least one vertex (BFS) of the polyhedron (standard form LP)*. Therefore, it searches for an optimal solution only among the finite number of BFSs of the standard form LP. Starting with an initial BFS, the simplex method computes a sequence of BFSs, such that the objective function value improves monotonically along the sequence, until it reaches an optimal BFS. The key computation in the simplex method, called a *pivoting operation*, is the construction of the next BFS in the sequence using the current BFS.

To see the details of the pivoting operation we start at the BFS corresponding to the vertex A. Partitioning the variables into two groups $\mathbf{x}_N^T = [y \ z \ s]$ and $\mathbf{x}_B^T = [x \ t]$, corresponding to the nonbasic and basic variables of the BFS 1, the equality constraints of the LP can be rewritten as

$$\begin{aligned} \underbrace{\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}}_N \mathbf{x}_N + \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_B \mathbf{x}_B &= \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_b \\ \Rightarrow \mathbf{x}_B &= B^{-1}(\mathbf{b} - N \mathbf{x}_N) \\ \Rightarrow \begin{bmatrix} x \\ t \end{bmatrix} &= \begin{bmatrix} 1 & -z & -s \\ 1 & -y & -z \end{bmatrix} \end{aligned} \quad (2)$$

which shows that the basic variables of BFS 1 can be expressed in terms of its nonbasic variables. In general, the matrix B that premultiplies $\mathbf{x}_B$, as shown above, is invertible for every BFS and the $\mathbf{x}_B$ of a BFS can always be expressed in terms of its $\mathbf{x}_N$. Using the expression for $\mathbf{x}_B$, in terms of $\mathbf{x}_N$, the objective function can be rewritten in terms of only the nonbasic variables as follows:

$$\begin{aligned} \mathbf{c}^T \mathbf{x} &= \underbrace{\begin{bmatrix} -2 & -1 & 0 \end{bmatrix}}_{\mathbf{c}_N^T} \mathbf{x}_N + \underbrace{\begin{bmatrix} 2 & 0 \end{bmatrix}}_{\mathbf{c}_B^T} \mathbf{x}_B \\ &= \mathbf{c}_B^T B^{-1} \mathbf{b} + \underbrace{(\mathbf{c}_N^T - \mathbf{c}_B^T B^{-1} N)}_{\text{Reduced cost coefficients}} \mathbf{x}_N \\ &= 2 - 2y - 3z - 2s \end{aligned} \quad (3)$$

The above reformulation of the objective function used the constraints in the LP and, hence, is valid only in the feasible region of the LP.

Since the nonbasic variables y, z , and s have negative reduced cost coefficients, increasing any one of them from zero while keeping the other two fixed at zero will improve the objective function value. For example, we can increase z keeping s and y fixed at zero and since x and t are to remain nonnegative, Equation (2) shows that z can be increased up to one.⁴ If we set $z = 1$

then z exits the nonbasic set (since nonbasic variables have to be zero). Since $x = t = 0$ when $z = 1$ either x or t can be added to the nonbasic set to replace z . If z is replaced by x , we get BFS 5, while if t replaces z , we get BFS 8, both of which represent vertex E. Similar reasoning shows that if, instead of z , we had chosen to remove $y(s)$ from the nonbasic set we would end up at BFS 2 (BFS 4), which represents vertex B(D).

Thus, in each pivoting operation the simplex method swaps one nonbasic variable for a basic variable. If the pivoting operation is to decrease the objective function the variable exiting the nonbasic set must have a negative reduced cost coefficient.⁵ As the preceding argument shows, at BFS 1, y, z and s are all candidates to exit the nonbasic set. The criterion used to pick the variable that will exit the nonbasic set, from among the competing candidate variables, is called the *pivot rule*.⁶ For example, if simplex method uses the *maximum increase pivot rule* then, among the competing nonbasic variables, the variable that yields the greatest improvement in the objective function will be chosen to exit the nonbasic set. At BFS 1, for example, the maximum increase pivot rule chooses z over y and s , taking the algorithm to BFS 5 (or BFS 8).

From the table in Fig. 1, we see that the only BFS that has a smaller objective function value than BFS 5 (vertex E) is BFS 3 (vertex C); so, if the algorithm is to achieve monotonic improvement then it has to move from BFS 5 to BFS 3. However, the nonbasic sets of BFS 5 and BFS 3 have only one variable in common making it impossible to move from BFS 5 to BFS 3 in one pivoting operation.⁷ So, the simplex method performs

⁵Nonnegative reduced cost coefficients for all the nonbasic variables signal that the current BFS is *optimal*.

⁶This is also called the *entering variable rule* since it only specifies which variable enters the basis but does not specify which variable, among the candidate basic variables, should leave the basis.

⁷The simplex method can go from one BFS to another in one pivoting operation only if the nonbasic sets of the two BFSs have all but one of the variables in common.

⁴If z has a negative coefficient in $\mathbf{c}_N^T - \mathbf{c}_B^T B^{-1} N$ and nonpositive coefficients in $B^{-1} N$ then it follows that the objective function can be decreased indefinitely and the LP would not have a bounded optimum.

a *degenerate pivoting operation* to move from BFS 5 to another BFS representing the same vertex E, namely BFS 6 or BFS 7, from either of which the algorithm can move to BFS 3 in one pivoting operation. The degenerate pivoting operation does not improve the objective function value but is needed to represent the degenerate vertex E with an appropriate BFS from which it can move to a better vertex.

Given an initial BFS, the pivoting operation described above can be applied repeatedly to construct a sequence of BFSs that monotonically improves the objective function value. However, since the improvement of the objective function is not strictly monotonic—as illustrated by the degenerate pivoting operation above—there is the possibility that the simplex method can cycle among the multiple BFSs that represent a degenerate vertex. Such cycling has, in fact, been demonstrated and can be avoided using an *anticycling pivot rule* [1].

The construction of the first BFS is tied to the *linear solvability problem (LSP)* [1]. Given a general system of linear inequalities S , the LSP seeks a solution of S , if one exists, or an indication that S is unsolvable, if such is the case. LSP is polynomially equivalent to LP; that is, given a polynomial-time algorithm for LSP, one can use it to devise a polynomial-time algorithm for LP and conversely [1]. Therefore, the construction of the initial BFS, is almost as “hard” as the problem of LP itself. One approach used to construct the first BFS of a given LP, L , is to solve an *auxiliary LP* that is constructed using L . The solution of the auxiliary LP yields the first BFS of L , if L is solvable, or an indication that L is unsolvable, if such is the case [1].

COMPLEXITY OF THE SIMPLEX METHOD

The time requirements of simplex method have been investigated by studying both its *average-case time complexity* as well as its *worst-case time complexity*. The average-case complexity of the simplex method, for LPs drawn from certain distributions, was shown to be polynomially bounded [7–10]; simplex has also been shown to have a polynomial *smoothed* complexity [11]. A different kind of

probabilistic analysis involves randomizing the pivot rule—that is, picking an improving nonbasic variable randomly in each pivoting operation. Such randomized pivot rules are known to have subexponential upper bounds for their expected running time [12,13]. In the following discussion we focus on the worst-case complexity of the simplex method for deterministic (nonrandomized) pivot rules.

In 1972, Klee and Minty [14] constructed a family of LPs for which they showed that the simplex method, using Dantzig’s pivot rule, performs an exponential number of pivoting operations before reaching the optimal solution. The Klee–Minty polytopes are perturbed d -cubes; Figure 2 shows a three-dimensional example (not drawn to scale) that illustrates the combinatorial behavior of simplex on the Klee–Minty LPs. A combinatorial 3-cube (d -cube) is a prism whose top and bottom facets are combinatorially equivalent to squares or 2-cubes ($(d-1)$ -cubes). The Klee–Minty LP simplex visits all the 2^{d-1} vertices of the bottom $(d-1)$ -cube, moves to the top $(d-1)$ -cube and visits all the vertices of the top facet in the reverse order, thereby performing 2^{d-1} pivoting operations in all. The simplex method visits every vertex in the Klee–Minty LP even though the optimal vertex is adjacent to the initial vertex and it is possible to reach the optimal BFS from the first BFS in *one* pivoting operation, using the *maximum improvement pivot rule*⁸ instead of *Dantzig’s pivot rule* [14].

The Klee–Minty LP raised the question of *whether it is possible to make the simplex method a worst-case polynomial-time algorithm by employing a good pivot rule*; the question remains unanswered to this day. All the (nonrandomized) pivot rules whose complexities have been analyzed are known to require exponential time in the worst case; see papers [15–22] for details.⁹ Zadeh [25] made the observation

⁸Pick the nonbasic variable that yields the greatest improvement in the objective function value.

⁹In fact, even the gravitational method [23], which is not constrained to traverse the edge-graph of the polyhedron, was shown to behave like the steepest descent pivot rule and has exponential worst-case complexity [24].

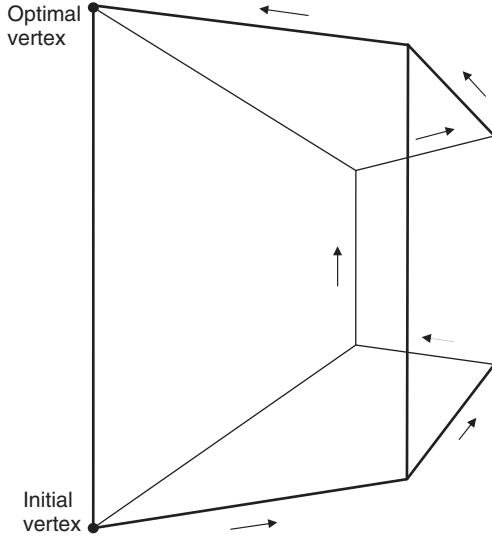


Figure 2. Simplex's trajectory on a three-dimensional Klee–Minty LP.

that the structure of the polytopes used to establish the worst-case exponential-time behavior of the analyzed pivot rules have the so-called *deformed product structure*¹⁰ and proposed a pivot rule—the *least entered pivot rule*¹¹—that was devised to avoid the pathological behavior on deformed product polytopes. The question of whether Zadeh's pivot rule makes the simplex method a worst-case polynomial-time algorithm remains open.

POLYHEDRAL DIAMETER AND THE # FUNCTION

For vertices u and v in a polyhedron P , let $\delta_P(u, v)$ denote the length¹² of the shortest

edge-path between u and v . Further, let $\delta(P)$ denote the *edge-diameter* of P , that is, the maximum value of $\delta_P(u, v)$ over all pairs of vertices (u, v) in P . Clearly, between every pair of vertices in P there is an edge-path of length of at most $\delta(P)$. The diameter functions Δ and Δ_b are defined as

$$\Delta(d, n) := \max \{ \delta(P) \mid P \text{ is a } d\text{-polyhedron} \\ \text{with } n \text{ facets} \}$$

$$\Delta_b(d, n) := \max \{ \delta(P) \mid P \text{ is a } d\text{-polytope} \\ \text{with } n \text{ facets} \}.$$

While a polynomial upper bound on $\Delta(d, n)$ does not guarantee the existence of worst-case polynomial-time pivot rule, a superpolynomial lower bound on $\Delta(d, n)$ will preclude the existence of a worst-case polynomial-time pivot rule. Hence sharp upper and lower bounds on $\Delta(d, n)$ are of interest in the context of the quest for a worst-case polynomial-time simplex algorithm. Although, neither a polynomial upper bound nor a superpolynomial lower bound has been established for $\Delta(d, n)$, the last five decades have witnessed accumulation of both theoretical and empirical results related to the diameter functions.

On the basis of the performance of the simplex method on thousands of practical LPs, Dantzig [4, p. 160] observed that if the basic set of an LP has m variables, and the initial and optimal BFSs have k basic variables in common, then the number of simplex steps is “usually close” to $m - k$; further, he added that for $k = 0$, the number of iterations rarely exceeded $3m$. Forty years later, Dantzig and Thapa [5, p. 31] reiterated the observation that simplex usually requires fewer than $3m$ iterations for problems in which the basic set has m variables. For additional computational results, see Bixby [26, 27].

The good practical performance of the simplex method and the polynomial bounds on the average-case complexity would not be surprising if polyhedra indeed have a *large number of short edge-paths* between every pair of vertices. In fact, the famous Hirsch conjecture [4] asserts that $\Delta(d, n) \leq n - d$ while the weaker *polynomial diameter conjecture* (PDC) [28] asserts that $\Delta(d, n) \leq P(d, n)$, where $P(d, n)$ is some polynomial of d and

¹⁰ The deformed product polytopes have the combinatorial structure of the form $P_d \times P_k$ where P_d (P_k) is a d -polytope (k -polytope), and k is a small number. The Klee–Minty d -polytopes are of the form $C_{d-1} \times P_1$ where C_{d-1} is the $(d - 1)$ -cube and P_1 is a line segment.

¹¹ Among the competing nonbasic variables Zadeh's *least entered pivot rule* chooses the nonbasic variable that has entered the basic set (exited the nonbasic set) least often.

¹² The length of an edge-path is the number of edges in the path.

n . The Hirsch conjecture remains open for bounded polyhedra, while the *PDC* remains open for both bounded and unbounded polyhedra.

While the diameter function provides an upper bound on the length of the shortest edge-path between two vertices, it does not contain complete information about the number of edge-paths, between the vertices, of length less than or equal to the bound provided by the diameter function. Interest in the number of paths suggests, in addition to $\Delta(d, n)$, the function $\#_f(d, n)$, defined as *the greatest lower bound on the number of edge-paths of length $f(d, n)$ or less, between a pair of vertices in a d -polyhedron with n facets, where $f(d, n)$ is a specified function of d and n* . In terms of the $\#$ function, the Hirsch conjecture can be stated as the claim that $\#_{n-d}(d, n) \geq 1$ and the PDC as $\#_{P(d, n)}(d, n) \geq 1$, for some polynomial function $P(d, n)$.

Progress in our understanding of the diameter functions has been rather modest and even less is known about the $\#$ function. We summarize the prominent results below; the reader is referred to Kalai [29], Kim and Santos [30], and Klee and Kleinschmidt [31] for a more complete survey of the results about the diameter functions.

The Hirsch conjecture is known to hold for $d = 3, n \geq 4$ [32]. The Hirsch conjecture fails for unbounded 4-polyhedra with 8 facets and further the departure from the Hirsch bound for general polyhedra tends to infinity as $\min(d, n - d)$ increases [33]. Specifically,

- $\Delta(d, n) \geq n - d + \min \left\{ \left\lfloor \frac{d}{4} \right\rfloor, \left\lfloor \frac{n-d}{4} \right\rfloor \right\}$ [33].

The Hirsch conjecture for bounded polyhedra (polytopes) is still open. The following results are known for $\Delta_b(d, n)$.

- $\Delta_b(d, n) \leq n - d$ for $n \leq d + 5$; $\Delta_b(d, 2d) = d$ for $d \leq 5$ [33].
- $\Delta_b(d, n) \leq n - d$ for $n = d + 6$; $\Delta_b(d, 2d) = d$ for $d = 6$ [34].
- $\Delta_b(4, 10) = 5$; $\Delta_b(5, 11) = 6$ [35]. $\Delta_b(4, 11) = 6$ [34]; $\Delta_b(4, 12) = 7$ [36].
- $\Delta_b(d, n) \geq n - d$ for $d \geq 7$, that is, the Hirsch bound is sharp [37–39].

- $\Delta_b(d, n) \leq n2^{d-3}$ [40].

For the general case the best known upper bound is

- $\Delta(d, n) \leq n^{\log d + 2}$ [28].

The Hirsch conjecture has been reformulated as several equivalent conjectures, the details of which, including the proofs of equivalence, can be found in Klee and Kleinschmidt [31]. For example, the seemingly less general d -step conjecture, which claims that $\Delta_b(d, 2d) = d$ for all d , is known to be equivalent to the bounded Hirsch conjecture. Of particular interest is the *Gaussian elimination sign conjecture* (GESG), which is equivalent to the bounded Hirsch conjecture and particularly well-suited for investigating the $\#_d(d, 2d)$ function for bounded polyhedra [41]. The GESG invokes the $(d - 1)$ -dimensional representation of the symmetric group $\text{Sym}(d)$ on d letters.¹³ Specifically, if $\tau \in \text{Sym}(d)$, then define the $(d - 1) \times (d - 1)$ matrix Q_τ as

$$(Q_\tau)_{i,j} = \begin{cases} 1 & \text{if } \tau(i) = j \leq d - 1 \\ 0 & \text{if } \tau(i) \neq j \text{ and } 1 \leq \tau(i) \leq d - 1 \\ -1 & \text{if } \tau(i) = d. \end{cases}$$

$S_{d-1}(d) = \{Q_\tau \mid \tau \in \text{Sym}(d)\}$ is a $(d - 1)$ -dimensional representation of $\text{Sym}(d)$. Using $S_{d-1}(d)$ the notion of a *general position* $(d - 1) \times (d - 1)$ matrix is defined as follows. A $(d - 1) \times (d - 1)$ matrix M is said to have a *nondegenerate triangular factorization* if M can be written in terms of a lower triangular matrix L and an upper triangular matrix U as,

$$M = L^{-1}U, \text{ where } L_{i,j} \neq 0 \text{ iff } j \leq i; L_{k,k} = 1, \\ 1 \leq k \leq d - 1; U_{i,j} \neq 0, \text{ iff } i \leq j.$$

The above nondegenerate triangular factorization is said to be a *positive triangular factorization* if the nonzero elements of the matrices L and U are positive. A $(d - 1) \times$

¹³ $\text{Sym}(d)$ is the group of all permutations of $(1, 2, \dots, d)$.

$(d-1)$ matrix M is said to be in *general position* if all the $(d!)^2$ matrices $Q_\sigma M Q_\tau$, $\sigma, \tau \in \text{Sym}(d)$ have nondegenerate triangular factorizations. The GESC states that *for every $(d-1) \times (d-1)$ general position matrix M there exists at least one pair $(\sigma, \tau) \in \text{Sym}(d) \times \text{Sym}(d)$ for which $Q_\sigma M Q_\tau$ has a positive triangular factorization.*

Klee and Walkup showed that $\Delta_b(d, 2d)$ can always be realized as the length of the edge-path between the antipodal vertices of a d -dimensional Dantzig figure [33, Theorem 2.8]; a d -dimensional Dantzig figure is a triple (P, u, v) where P is a simple d -polytope with $2d$ facets, and u, v are two specific *antipodal* vertices in P —that is, vertices that do not have any incident facet in common. The $(d-1) \times (d-1)$ general position matrices M mentioned in *GESC* correspond to d -dimensional *Dantzig figures*. Every pair $(\sigma, \tau) \in \text{Sym}(d) \times \text{Sym}(d)$ that yields a positive triangular factorization of M corresponds to an edge-path of length d between the antipodal vertices of the associated Dantzig figure [41, Theorem 5.1]. Therefore, the number of pairs (σ, τ) that yield positive triangular factorizations of a given $(d-1) \times (d-1)$ general position matrix M , denoted $\#^D(M; d)$, gives the number of edge-paths of length d between the antipodal vertices of the corresponding Dantzig figure.¹⁴ Further, let

$$\#^D(d) := \min \left\{ \#^D(M; d) \mid M \text{ is a } (d-1) \times (d-1) \text{ matrix in general position} \right\}.$$

Computational experiments suggested that $\#^D(d) \geq 2^{d-1}$, an inequality that was stated as the *strong d -step conjecture* [41]. The above inequality was shown to hold as an equality for the two extreme cases—Dantzig figures where P is either the dual of a cyclic polytope [which has the maximum number of vertices among $(d, 2d)$ polytopes] or the truncation polytope [which

has the minimum number of vertices among $(d, 2d)$ polytopes] [42]. The inequality also holds for $d \leq 4$ [41, 43]. Holt and Klee [39] showed that the inequality fails for $d \geq 5$ by constructing Dantzig figures for which $\#^D(M; d) = (\frac{3}{4}) 2^{d-1}$. More recently $\#^D(d)$ was shown to have an asymptotic upper bound of $O((1.8882)^d)$ [37]. The empirical experience with the simplex method [4, 5], the computational results about $\#^D(d)$ [41] and the results of [37, 41–43] suggest the question:

$$\text{Is } \#^D(d) \geq a^{d-1} \text{ for some } a > 1?$$

The answer is in the affirmative for $d \leq 4$ and the question is open for $d \geq 5$.¹⁵

Acknowledgment

I thank Mihailo Stojnic for his critical comments on the first version of this article. I also thank the referees for their many comments, which have improved the article.

REFERENCES

1. Chvatal V. Linear programming. New York: W.H. Freeman; 1983.
2. Dantzig GB, Thapa MN. Linear programming 1: introduction. New York: Springer; 1997.
3. Sherali HD, Ozdaryal B, Adams WP, *et al.* On using exterior penalty approaches for solving linear programming problems. *Comput Oper Res* 2001;28(1):1049–1074.
4. Dantzig GB. Linear programming and extensions. Princeton (NJ): Princeton University Press; 1963.
5. Dantzig GB, Thapa MN. Linear programming 2: theory and extensions. New York: Springer; 2003.
6. Ziegler G. Lectures on polytopes. New York: Springer; 1995.
7. Borgwardt KH. Some distribution-independent results about the asymptotic order of the average number of pivot steps of the Simplex method. *Mathematics of Operations Research* 1982;7:441–462.

¹⁴The superscript D indicates that we are counting the number of paths between the antipodal vertices of a Dantzig figure.

¹⁵A counterexample to the Hirsch conjecture (a 43-dimensional polytope with 86 facets) was recently announced in Santos 44.

8. Borgwardt KH. The average number of steps required by the Simplex method is polynomial. *Z Oper Res* 1982;26:157–177.
9. Borgwardt KH. The simplex method: a probabilistic analysis. New York: Springer; 1986.
10. Smale S. On the average number of steps of the simplex method of linear programming. *Math Program* 1983;27:241–262.
11. Spielman DA, Teng S. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *J Assoc Comput Mach* 2004;51(3):385–463.
12. Kalai G. A subexponential randomized simplex algorithm. Proceedings of the 24th Annual ACM Symposium on the Theory of Computing. Victoria: ACM Press; 1992. pp. 475–482.
13. Matousek J, Sharir M, Welzl E. A subexponential bound for linear programming. Proceedings of the 8th annual Symposium on Computational Geometry. 1992. pp. 1–8.
14. Klee V, Minty GJ. How good is the simplex algorithm? In: Shisha O, editor. *Inequalities III*. New York: Academic Press; 1972. pp. 159–175.
15. Avis D, Chvatal V. Notes on Bland's pivoting rule. In: Balinski ML, Hoffman AJ, editors. *Polyhedral combinatorics, Volume 8, Mathematical Programming Study*. Amsterdam, New York, Oxford: North-Holland Publishing Company; 1978. pp. 24–34.
16. Blair C. Some linear programs requiring many pivots. Faculty Working Paper No. 867. College of Commerce and Business Administration, University of Illinois at Urbana-Champaign; 1982.
17. Cunningham W. Theoretical properties of the network Simplex method. *Math Oper Res* 1979;4:196–208.
18. Goldfarb D, Sit WY. Worst-case behavior of the steepest edge simplex method. *Discrete Appl Math* 1979;1:277–285.
19. Jeroslow R. The simplex algorithm with the pivot rule of maximizing criterion improvement. *Discrete Math* 1973;4:367–377.
20. Murty KG. Computational complexity of parametric linear programming. *Math Program* 1980;19:213–219.
21. Zadeh N. Theoretical efficiency and partial equivalence of minimum cost flow algorithms: a bad network problem for the Simplex method. *Math Program* 1973;5:255–266.
22. Zadeh N. More pathological examples for network flow problems. *Math Program* 1973;5:217–224.
23. Murty KG. The gravitational method for linear programming. *Opsearch* 1986;23:206–214.
24. Morin TL, Prabhu N, Zhang Z. Complexity of the gravitational method for linear programming. *J Optim Theory Appl* 2001;108(3):633–658.
25. Zadeh N. What is the worst case behavior of the Simplex algorithm?. Technical Report No. 27. Stanford (CA): Department of Operations Research, Stanford University; 1980 May 1.
26. Bixby RE. Implementing the simplex method: the initial basis. *ORSA J Comput* 1992;4:267–284.
27. Bixby RE. Progress in linear programming. *ORSA J Comput* 1994;6(1):15–22.
28. Kalai G, Kleitman DJ. A quasi-polynomial bound for the diameter of graphs of polyhedra. *Bull Am Math Soc* 1992;26(2):315–316.
29. Kalai G. Linear programming, the simplex algorithm and simple polytopes. *Math Program* 1997;79:217–233.
30. Kim ED, Santos F. An update on the Hirsch conjecture. [arXiv:0907.1186v2](https://arxiv.org/abs/0907.1186v2); [math.CO]; 2009 Dec 22.
31. Klee V, Kleinschmidt P. The d -step conjecture and its relatives. *Math Oper Res* 1987;12(4):718–755.
32. Klee V. Diameters of polyhedral graphs. *Can J Math* 1964;16:602–614.
33. Klee V, Walkup D. The d -step conjecture for polyhedra of dimension $d < 6$. *Acta Math* 1967;133:53–78.
34. Bremner D, Schewe L. Edge-graph diameter bounds for convex polytopes with few facets. [arXiv:0809.0915v3](https://arxiv.org/abs/0809.0915v3) [math.CO]; 2009 July 15.
35. Goodey PR. Some upper bounds for the diameter of convex polytopes. *Isr J Math* 1972;11:380–385.
36. Bremner D, Deza A, Hua W, *et al.* More bounds on the diameters of convex polytopes. [arXiv:0911.4982v1](https://arxiv.org/abs/0911.4982v1) [math.CO]; 2009 Nov 25.
37. Fritzsche K, Holt FB. More polytopes meeting the conjectured Hirsch bound. *Discrete Math* 1999;205:77–84.
38. Holt FB. Blending simple polytopes at faces. *Discrete Math* 2004;285:141–150.
39. Holt FB, Klee V. Many polytopes meeting the conjectured Hirsch bound. *Discrete Comput Geom* 1998;20:1–17.
40. Larman DG. Paths on polytopes. *Proc Lond Math Soc* 1970;20(3):161–178.

- 41. Lagarias JC, Prabhu N, Reeds JA. The d -step conjecture and Gaussian elimination. *Discrete Comput Geom* 1997;18:53–82.
- 42. Lagarias JC, Prabhu N. Counting d -step paths in extremal Dantzig figures. *Discrete Comput Geom* 1998;19:19–31.
- 43. Holt FB, Klee V. Counterexamples to the strong d -step conjecture for $d \geq 5$. *Discrete Comput Geom* 1998;19:33–46.
- 44. Santos F. A counterexample to the Hirsch conjecture. 2010; DOI: arxiv.org/abs/1006.2814v1.

THE STRATEGIC CHOICE APPROACH

JOHN FRIEND
Stradspan Limited, Sheffield, UK

The *Strategic Choice Approach* (SCA) is a name that has become attached to a balanced collection of tools and concepts developed by management scientists to facilitate communication about the structuring of perplexing decision problems, and also about strategies for making progress toward timely commitments to agreed actions. The tools are primarily visual in form, yet are linked at an analytical level through a distinctive philosophy that draws on insights into how decision makers can learn to manage diverse sources of uncertainty in a strategic way. This philosophy reflects an interpretation of the practical challenges of interactive planning under real-time action pressures, as observed by management scientists working alongside their social science colleagues.

APPLICATIONS

The tools of SCA have now been successfully introduced to support people facing challenging decision problems in many contexts of management, at levels ranging from local community action to national policy development. In many instances, these tools are not employed by management science professionals but by project managers, consultants or programme coordinators with responsibility for facilitating decision processes in fields as diverse as environmental conservation, information technology strategy, urban development, product marketing, and health and welfare policy.

The approach has now been adapted for application in a wide range of cultural contexts. It is frequently used to facilitate interagency workshops, which are designed to bring together diverse organizational

and stakeholder interests; at other times, it is introduced to guide communication in more informal management settings, with optional computer support. Some users choose to apply the tools in combination with qualitative or quantitative methods drawn from other toolkits with which they are familiar. This reflects the exploratory spirit of action research that has governed the evolution of the approach since it was pioneered by a group of operational research scientists in Britain some four decades ago.

ORIGINS

The origins of the approach are to be found in the early project experiences of the *Institute for Operational Research* (IOR) in the Tavistock Institute of Human Relations in London. This was an initiative of the Council of the UK Operational Research Society in partnership with the Council of the Tavistock Institute, which is an independent not-for-profit institute in London dedicated to the application of the social sciences to important societal issues. The IOR was launched in 1963 under the direction of the late Neil Jessop, as a semiautonomous unit within the Tavistock matrix [1].

SCA grew out of two of IOR's seminal projects, both conducted by mixed teams of OR and social scientists. One of these projects addressed issues of communication in the building industry, while the other explored the processes of policy making and planning in city government. Between the early 1970s and the mid-1980s, IOR teams went on to conduct a range of further research and consulting projects, many of them addressed to public policy issues and processes under the auspices of government agencies in the United Kingdom and overseas.

Since then, the practice and theory of the approach has been further developed through projects and programmes in countries as far apart as Canada, the Netherlands, Brazil, Japan, Sweden, Venezuela, and

Italy, in many cases involving collaboration between local champions and members of the original “IOR School.” Several references are found in the third edition of the core textbook by Friend and Hickling, *Planning under Pressure* [2].

FOUNDATIONS OF THE APPROACH

SCA is one of the six main approaches to problem structuring presented in the book *Rational Analysis for a Problematic World Revisited* [3], which is recognized as the leading guide in this field. It has now established its place as one of the most widely applied of the new interactive approaches to what is sometimes called *qualitative OR*. It can be distinguished from other OR approaches, both soft and hard, with reference to a summary of its principal *emphases*.

Emphases of the Strategic Choice Approach

The distinctive emphases of SCA reflect its origins in extensive observation and facilitation of group decision processes. These emphases can be expressed in terms of the following five key dimensions.

Focus: more emphasis on *facilitating decisions* than *exploring systems*;

Knowledge: more emphasis on *managing uncertainty* than *processing information*;

Time: more emphasis on *sustaining progress* than *projecting futures*;

Skill: more emphasis on *structuring communication* than *reinforcing expertise*;

Influence: more emphasis on *guiding collaboration* than *exercising control*.

The emphasis on *facilitating decisions* serves to distinguish SCA from those approaches that aspire to take a broader, more synoptic view of systemic influences. The emphasis on *managing uncertainty* can be seen as the most distinctive feature of SCA, leading to a focus on the *strategic management of uncertainty through time*. The emphasis on *sustaining progress* reflects the observed reality that the management

of interdependent choices tends to be an incremental process, in which it can be more realistic to make progress step by step than to attempt to design a more comprehensive planning framework.

The final two emphases, on *structuring communication* and on *guiding collaboration*, reflect the observation that, the more interconnected the choices to be addressed become, the wider the range of interests and responsible decision makers that tend to be drawn in. So it becomes important to build capacities to reach out toward other decision makers, understanding the contexts and constraints within which they work and negotiating collaborative links—across organizational boundaries as well as within corporate settings.

Managing Uncertainty in a Strategic Way

One of the most important observations from the city government project was that, whenever a challenging issue arose on a group agenda, people tended to argue for different ways of overcoming the challenge, as shown in Fig. 1.

In such a situation, some people might argue for a relatively *technical* response, through some form of investigation—ranging from a quick telephone call to an expert to a more formal commissioning of some kind of survey, analysis, forecasting, or research. Other people meanwhile might argue for a more *political* response, perhaps in the form of a review of objectives, policies, goals, or priorities within some accountable group; or perhaps a consultation with representatives of any professional or community interests that will be affected by the decisions to be made. Other people again might argue for a more *structural* response, broadening the agenda to take into account other related choices, and thus moving toward placing the choice within a wider planning framework.

In Fig. 1, these three types of response are expressed in terms of calls for “deeper investigations,” “clearer policies,” and “broader perspectives,” respectively. Each of these in turn is interpreted as a response to a different kind of *source of uncertainty*.

The figure identifies three broad categories of uncertainty as follows:

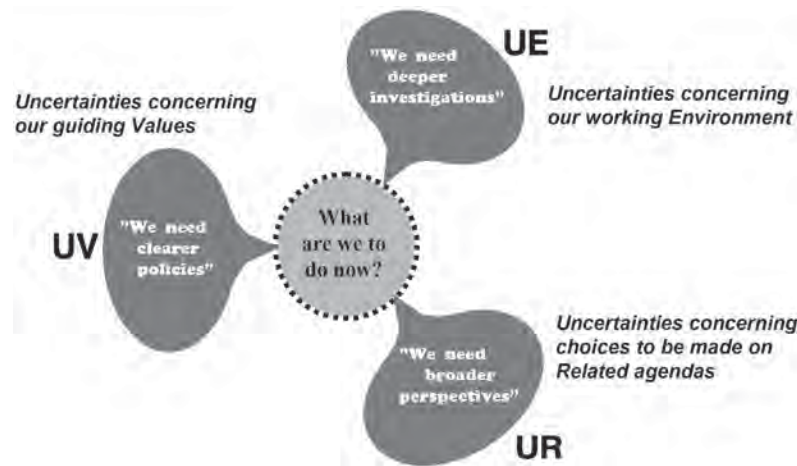


Figure 1. Sources of uncertainty in strategic choice.

Uncertainties concerning our working Environment (UE);

Uncertainties concerning our guiding Values (UV);

Uncertainties concerning choices to be made on Related agendas (UR).

It is usually not hard to agree that the reduction of significant sources of uncertainty in any of these directions is a desirable aim. Yet, wherever there are pressures on time or other resources, it can be a more controversial matter to agree on how much time and effort should be invested in actions directed at any one category of uncertainty rather than another.

Therefore, questions arise about how much *investment* of time and other resources should now be made in explorations directed toward different sources of uncertainty—broadly technical, political, and structural. These are the questions that many experienced decision makers learn to consider in an intuitive way; yet, they are rarely addressed explicitly at times when people meet to consider jointly what they should do in a specific situation, under some current combination of pressures on time and resources.

Whatever agreements may be agreed upon about actions to reduce uncertainty, these actions can have a profound influence not

only on the choices to be made on the current agenda but also on future paths of decision making. Because, depending on the choices made as to which particular package of investigations, policy consultations, or negotiations should now be set in train, different people may become involved at different times and through different channels. Hence decision making does not merely become a *process* in a sequential sense but rather an ever-shifting lattice of partially intertwined activities in which patterns of participation may shift over time in complex ways.

Therefore, while pressure to take decisions can be regarded as the primary force that drives the dynamics of choice, the *planned management of uncertainty through time* can come to play an important supporting role in shaping the agendas of the decision makers. This is why it is only realistic to view the management of uncertainty as an important secondary driver in any activity of strategic—as opposed to simple—choice.

Four Key Aspects of Decision Making

In presenting the more specific set of concepts, methods, and tools that characterize SCA, it is useful to introduce a general model of decision making that identifies four complementary aspects or *modes* of decision making. These are presented in Fig. 2 as the *shaping, designing, comparing, and choosing*

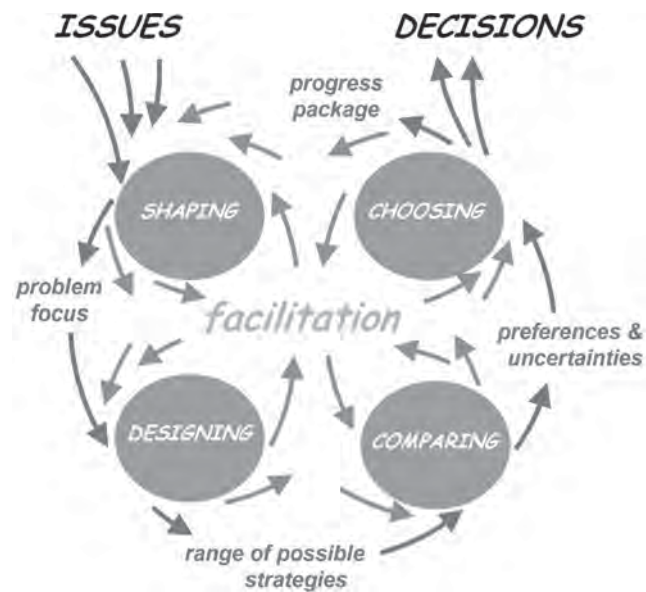


Figure 2. Four modes of decision making.

modes. They are presented here as stages in a cycle, with opportunities at any stage to revisit earlier modes.

In any simple context of decision making, where it is expected that choices in different areas of decision will be made one at a time according to prescribed guidelines, the only two significant modes are those indicated in the lower part of Fig. 2. Because, within a clear problem focus, it is sufficient to *design* some set of options to be considered, then to *compare* them in order to arrive at a preference. Sometimes, the options to be compared may be clear-cut—yes or no, this response or the other—while at other times more effort may be needed to *design* a range of feasible options. Likewise, in *comparing*, sometimes it may be clear which option is to be preferred, while at other times a careful evaluation may be required to judge which option will perform best within current guidelines. Wherever there is significant uncertainty over this, a loop back to the *designing* mode may be helpful in order to introduce further design options.

It is where there are a number of interrelated choices to be made that the two further modes of decision making shown in the upper half of Fig. 2—*shaping* and *choosing*—start to demand at least as much attention as the

more basic modes of designing and comparing. This is especially so where people bring to bear different perspectives on the nature of the decisions to be made and the configuration of links between them.

The challenge becomes all the greater where these linked areas of choice cut across different organizational or program agendas. It can now become harder to agree what areas of choice should be considered, in what words they should be expressed, and to what extent they are interlinked. Further, a collective judgment is required as to what should be the focus of concern at this stage—even accepting that it may be possible to alter such a focus by agreement later, as progress is made on the more urgent choices, and as new priorities emerge.

In such a situation, the mode of *choosing* can also become more challenging because a balance now has to be struck between areas within the present focus where decisions are to be made now, and those where choice can be deferred until some future time, while judgments are also needed regarding which areas of uncertainty are to be addressed now to increase confidence in future choices. Where there are many stakeholders, the striking of these balances can become

especially challenging, in both technical and political terms.

The concept of a *progress package* is found helpful as a means of assembling the outcomes of any cycle through the four modes. It can be presented in a simple grid format, distinguishing between decisions to be made *now* and those to be addressed at some point in the *future*, while also distinguishing between actions addressed to substantive decision areas and *exploratory* actions in response to key areas of uncertainty, as follows:

- Now: decisions agreed in some areas of choice;
explorations agreed for some areas of uncertainty.
- Future: other areas of choice in which decisions have not yet been agreed to;
other areas of uncertainty in which no explorations have been agreed.

The more challenging a process of strategic choice becomes, the more important it becomes to develop capacities to switch flexibly between one process mode and another, depending on judgments as to where the most immediate opportunities for progress are currently to be found. This is indicated in Fig. 2 by the inclusion of a loop linking each mode to each other mode, whether earlier or later in the cycle. In a group context, this kind of judgment calls for a subtle process of *facilitation*, in which the facilitator—or facilitation team—must balance a sensitivity to the dynamics of group interaction against a concern to sustain progress over the time remaining within the current session.

INTRODUCING THE TOOLS OF THE STRATEGIC CHOICE APPROACH

The sections that follow introduce, in the logical sequence of four modes introduced in Fig. 2, the set of basic concepts and visual tools that have been developed to support the application of SCA.

It would be unhelpful for us to present these concepts and tools purely in abstract terms. Therefore, they are brought to life

through an example of a situation that is presented as being small in scale and urgent, yet which illustrates the sorts of interconnected choices, uncertainties, options, and criteria that typically arise in broader planning contexts. The presenting situation is described briefly as follows:

PDQ is a management consulting group with 12 staff, which recently tendered to undertake a management review for a nearby public authority. The decision has been unexpectedly delayed for 4 months; then PDQ are unexpectedly offered the contract. *How are they now to respond, in a changed situation where some members of the proposed project team have now become committed to other assignments?*

Tools for Shaping Difficult Problems

The fundamental concept that is used to guide the work of the shaping mode is that of the *decision area*. A group of decision makers may agree to express any area of choice that they see ahead of them as a decision area if they can visualize at least two distinct *options*, and if they expect to have some level of influence in choosing between them.

The principal output of work in the shaping mode takes the form of a *decision graph*, in which significant connections between pairs of decision areas have been identified in the form of *decision links*. Where there are many interlinked decision areas, an important next step is to select a subset of these as an agreed *problem focus* before proceeding to develop and compare possible ways forward.

In logical terms, the task of shaping a decision problem in this way is very simple. Yet, as users of SCA soon discover, the art of applying these simple tools in practice can be far from straightforward. This is especially so where a workshop brings together a diverse group of participants, all with different perspectives and stakes in the current problem situation. The initial challenge is then to capture the perceptions about currently significant areas of decision, which the participants bring with them in their heads.

Figure 3 illustrates this by an example of the type of record that might be built up

about the PDQ consulting contract in the initial stages of discussion in the *shaping* mode. Figure 3a depicts an initial list of decision areas, such as might be written by a facilitator on a flip chart in response to suggestions from the group. A skilled facilitator writes legibly in freehand, using a broad-tipped colored marker pen—pausing at times to suggest shorter forms of words and to check that these capture the intended meaning. A question mark after the description of each decision area is useful as a reminder that each entry is intended to represent an *area* of choice, rather than some preferred course of action within that area.

An initial list of decision areas may cover more than one sheet of paper. Sometimes the facilitator may suggest that a proposed addition to the list may be better expressed not as a *decision area* but as an *uncertainty area*—possibly because it is more realistically treated as a matter for investigation rather than choice. Other suggestions may be better expressed as *comparison areas*, because they can more realistically be seen as high-level objectives or goals than as areas of choice between specific alternatives.

Therefore, in Fig. 3, the proposed entry *impact on our reputation* is shown struck through, to be transferred to a parallel list of comparison areas for later consideration. Similarly, the two entries “Whether to bid for HKI project?” and “Value to us of public sector work?” might be transferred to another parallel list of *uncertainty areas*, to be sorted later into the three categories indicated in Fig. 1. Some people might see the use of strikethrough lines as untidy, but it has value where it records an important moment in a group learning process.

It is worth observing here that SCA differs from more hierarchical approaches to strategic planning in not treating objectives as starting points but as criteria for comparison—and also perhaps as sources of uncertainty about values to be addressed if and when their significance to decisions becomes more evident.

Once agreement has been reached on a set of decision areas to be considered further, a shorter label for each can be chosen. If this is written on a semiadhesive sticker, as shown in Fig. 3, then it can be transferred to a blank flip chart as a candidate for inclusion in a decision graph. Figure 3

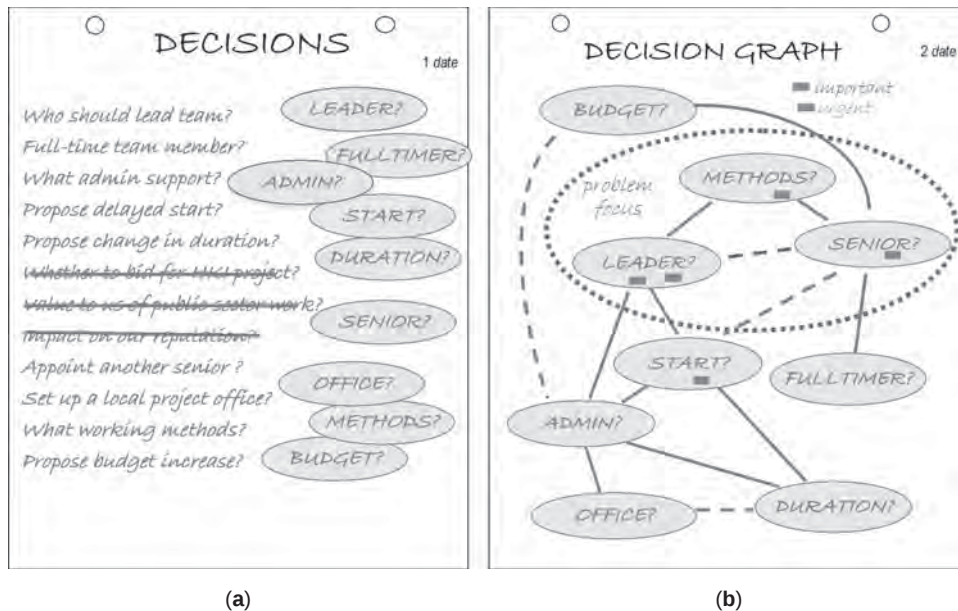


Figure 3. Tools for shaping problems.

illustrates the use of large elliptical stickers, which, while not essential, are especially convenient for this purpose; sources of supply can be located through www.ovalmap.com. The facilitator can begin moving the stickers around, with guidance from other participants, clustering together those that seem closely related until a good-enough first representation of the “shape” of the problem has been agreed. Then connecting lines—as straight as possible—can be added between pairs of decision areas that are agreed to be directly interlinked. In this visual mapping method—sometimes called *AIDA* for *Analysis of Interconnected Decision Areas*—the use of arrowheads to indicate directed links is not recommended, as questions of sequence and influence are handled in other ways.

As a further aid to selection, color coding can be introduced, as shown in Fig. 3, to identify any decision areas that are agreed to be more *important* or *urgent* than others. Then a boundary can be drawn to enclose those areas—preferably not more than three or four—which have been agreed as an initial *problem focus*. Exclusion from this initial focus does not preclude the reintroduction of decision areas later.

Where the number of decision areas is more than around 10, there may be merit in introducing more formal methods of analysis to help in agreeing on a focus, thus reducing the potential for bias from the use of visual methods alone. Computer methods may be helpful here—for examples see www.stradspan.com—so long as they do not disrupt the dynamics of the group. Users sometimes supplement these quick and intuitive methods of problem shaping by introducing mapping methods from other toolkits with which they are familiar. Several of the examples presented in *Planning under Pressure* [2] illustrate the potential for mixing methods in this way.

Tools for Designing Feasible Strategies

In the *designing* mode, the amount of effort required to develop a range of feasible strategies depends on the number of decision areas included within the current focus. Where the focus is restricted to a single decision area, the task is merely to agree to a set of two

or more feasible *options* within that decision area, which can be taken as a sufficient representation of the range of choice available. It is recommended that the number of options be limited to four or five at most, expressed where possible so as to be mutually exclusive. Occasionally, this may mean selecting a few representative points on a continuous scale.

Where the chosen focus embraces two or more linked decision areas, the process of devising a set of feasible courses of action—or strategies—becomes less simple because now it becomes necessary to consider combinations of options, one from each decision area, recognizing that some combinations may be *incompatible* for various reasons.

For example, in the PDQ case, it may be agreed upon that the three possible contenders for *Leader* of the team are *Ana*, *Bernard*, or *Colin*—with the knowledge that *Ana*, despite being originally named as leader of the team, is now less available than she was previously. It may also be thought that her senior colleague *Bernard* has less relevant skills; while their former colleague *Colin* now works for another employer, yet might be brought back to lead the team if contractual arrangements could be agreed upon. Because the inclusion of a second *Senior* team member in the team is considered optional, the options here are agreed to be *none*, *Ana*, or *Colin*. Then it may be agreed that the choice of project *Methods* be restricted to *traditional* or *group*—perhaps after some discussion of whether it might also be acceptable to suggest to this client a third option of *mixed methods*.

Once the agreed set of options in each decision area has been agreed and recorded—typically on a single flip chart—it remains to review the *mutual compatibility* of combinations of options from different decision areas. Figure 4a illustrates this in the case of PDQ by showing the incompatibilities displayed on a flip chart in terms of a set of *compatibility tables*—or matrices—one for each pair of decision areas. Alternatively, they can be displayed by expanding the decision graph into a more detailed *option graph* format [2, p.36]. In Fig. 4, an X symbol is used to indicate

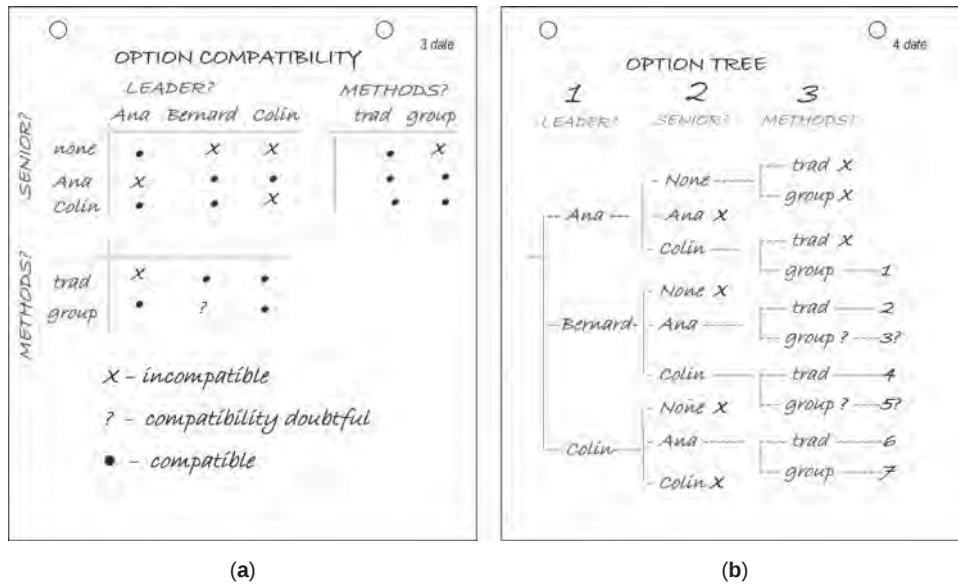


Figure 4. Tools for designing strategies.

incompatibility, while doubtful compatibility is indicated by a question mark. The use of an X is found to make for easier dialogue with nonmathematicians than the more conventional 0/1 symbols.

The reasoning behind each assumption of incompatibility may differ. For example, the inclusion of Ana as both leader and senior support person is here ruled out on logical grounds, while the choice of Bernard with no senior support is excluded here because of a judgment of his limited experience; and the choice of Colin with no senior support may be excluded because it would limit PDQ influence. The question mark over the use of group methods by Bernard reflects his limited experience in this field. It is useful to record the logic behind each assumption somewhere where it can be challenged later on.

The impact of any set of assumptions on the range of choice available across all the decision areas in focus can be explored through developing what is called an *option tree*, as shown in Fig. 4b. This involves taking the decision areas in any agreed sequence then, as each option is added, checking its compatibility with the option selected in each of the preceding decision areas. Figure 4 makes the logic explicit by

marking each closed or “dead” branch with a cross. The result in this instance is that, of the 18 theoretical combinations, only 7 are available—further reduced to 5 if questionable compatibilities are excluded. Where there are more than three decision areas in focus, this process of logical testing can be a time-consuming one, unless computer support is available. This is one reason why it is recommended that in a workshop the problem focus be restricted to only a few decision areas at a time.

The number of feasible combinations or *decision schemes* is, of course, unaffected by the sequence in which the decision areas are reviewed. However, the choice of sequence does affect the way in which structural information is displayed. For instance, the positioning of the Leader decision at the front of the tree shows that, if this is most urgent decision, then the choice of Ana will preempt all paths except one in the other decision areas, while the choice of Bernard will leave more future paths open.

Tools for Comparing Impacts under Uncertainty

As a foundation for making structured comparisons among the possible ways forward,

SCA introduces the basic concept of a *comparison area*—broadly equivalent to the familiar term *criterion*, which is indeed sometimes substituted for brevity.

Depending on the knowledge available, initial comparisons can be made either between options within a single decision area or between decision schemes spanning a set of linked decision areas. It is usual for a facilitator to start listing comparison areas on a separate flip chart from the outset of a workshop, side by side with the sheets used to record decision areas and uncertainty areas. It is normally sufficient to work with a set of between four and eight comparison areas, provided the concerns of all significant stakeholders or interest groups are recognized. Figure 5a shows a set of four comparison areas that might have been agreed upon in the PDQ case. In this instance, it is assumed that the debate has resulted in the substitution of one suggested comparison area (*contribution to profit*) by another (*contribution to cash flow*) as a preferred indicator of financial impact.

Figure 5b illustrates a visual method that has been found powerful as a means of helping a group of participants to compare *pairs* of options or schemes in terms of the perceived balance of advantage between them. This approach introduces a qualitative scale of *comparative advantage*, in which an agreed set of adjectives—marginal, significant, considerable, and extreme—is used to indicate perceived levels of advantage to either side. This forces the participants to make subjective judgments as to the relative importance of different comparison areas *in relation to their present decision situation*. So both tangible and more intangible considerations can now be brought together within a shared value framework, offering opportunities for negotiation within the group and enabling judgments to be made both within and across comparison areas.

In Fig. 5, small semiadhesive stickers are used to agree to a range of possibilities for the relative advantage between the alternatives on the comparative advantage grid, taking the comparison areas one at a time and then aggregating them. For each comparison area, discussion can then focus on

the range of possible advantage to either side, as well as the most likely position. In this instance, the judgment is that scheme 3, with Bernard as leader, is likely to have significant advantages in terms of cash flow and staff development. These are judged to outbalance the more marginal advantages of scheme 1, with Ana as leader, in terms of reputation and flexibility—remembering however that scheme 3 is subject to doubts about the capacity of Bernard to handle group methods.

The use of this kind of comparative advantage grid usually brings to the surface further areas of uncertainty relating to guiding values (UV in the strategic choice language), as well as uncertainties of types UE and UR. Therefore, this helps in building a base from which to begin considering the overall impact of uncertainty on decisions, leading into discussion of opportunities for managing the most critical areas of uncertainty in a collaborative way.

Similar comparisons might now be made between other pairs of decision schemes among the seven shown in Fig. 4. Since it can become time consuming to compare all feasible pairs of schemes in the same depth, crude methods of shortlisting are often introduced to reduce the range of pair comparisons. One approach is to look for schemes that appear to be dominated by other schemes in terms of critical comparison areas; another is to rule out those that transcend some maximum or minimum level on some of the more quantifiable criteria such as cash flow. Another approach is to aggregate separate assessments for options within decision areas as a starting point in comparing schemes. All these approaches mean temporary suppression of information about critical areas of uncertainty, accepting that these can then be explored in more depth for selected pairs.

The phrase *dynamic comparison* is sometimes used to describe a cycle in which periods of crude comparison among many schemes alternate with periods of closer pair comparison, in which some schemes may drop out while new areas of uncertainty may be revealed.

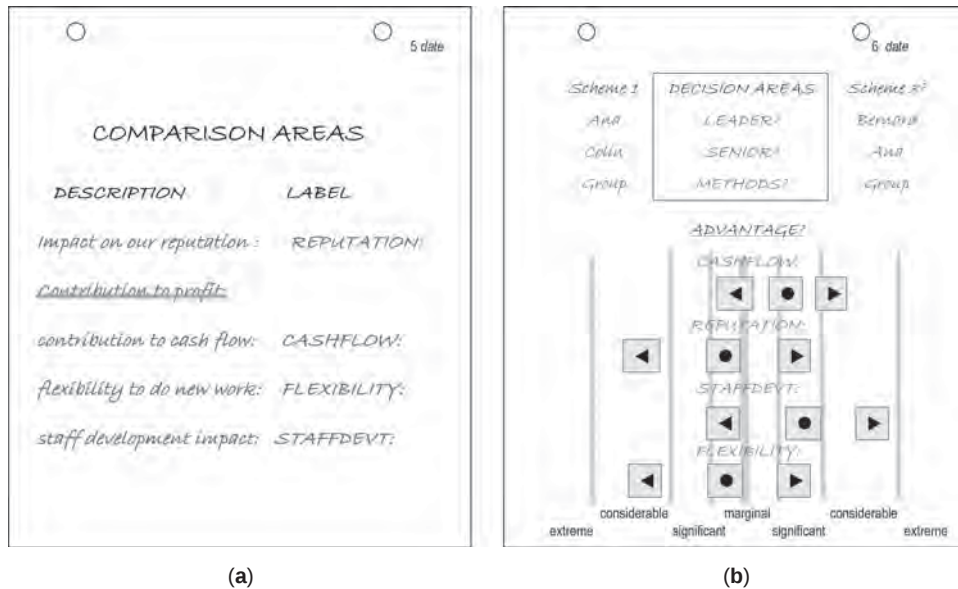


Figure 5. Tools for comparing under uncertainty.

Tools for Managing Uncertainty and Sustaining Progress

Once the most critical areas of uncertainty have been exposed, it becomes possible to discuss whether any possible *exploratory actions* can be identified to reduce their impact on the important decision areas within the present focus.

Even where a list of significant uncertainty areas has been built up from the start of a workshop, it is a good discipline to review and restructure the list when first moving into the choosing mode, to facilitate debate about any ambiguities or overlaps that may have arisen. An opportunity now arises to review the classification of each uncertainty area in terms of the three categories UE, UV, and UR, accepting that there may be some borderline cases. At this time,

symbols can also be introduced to indicate the relative *prominence* of each uncertainty area in relation to the decisions currently in focus—recognizing that this rating may have to be reviewed later if the problem focus is changed.

Among analytical participants, a bias is now likely to emerge toward sources of uncertainty of type UE that can be expressed in quantitative or probabilistic terms. Other people however may be more sensitive to political uncertainties, suggesting response in terms of policy consultations (UV), while others again may be aware of uncertainties about the intentions of other parties, prompting a more negotiative response of type UR. In the PDQ case, some key uncertainty areas that might now be identified are as shown below

- ? Will Colin be available for this project **?COLIN FREE** **UE +++**
 Exploratory options: *phone him at home; ask at next week's meeting; [no action]*
- ? Whether PDQ is to bid for HKI project **?HKI BID** **UR +**
 Exploratory options: *place on Board agenda; informal soundings; [no action]*
- ? Value to PDQ of bidding for public sector projects **?VAL PUB** **UV ++**
 Exploratory options: *policy debate; [no action]*
- ? Can Bernard handle group methods **?BERN SKILL** **UE ++**
 Exploratory options: *trial workshop; skill tests; [no action]*

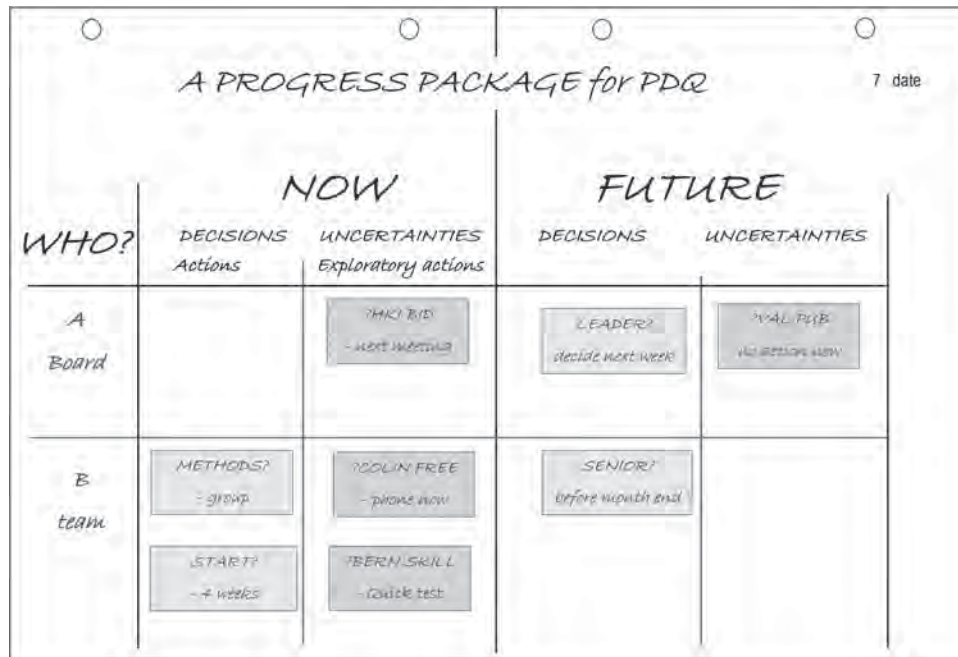


Figure 6. Tools for agreeing a way forward.

Each uncertainty area is here shown with a question mark placed *in front*—a convention used to distinguish it from a decision area—while its full description is followed by a briefer working label. Each area is also classified into one of the UE/UV/UR categories, and is given an agreed rating of *prominence* on a scale relating to the current problem focus.

Any views on possible *exploratory options* are then noted against each uncertainty area, recognizing that these can always be compared to a *null option* of taking no action. Therefore, both within and across uncertainty areas, judgments of relative value can now be made for all possible exploratory actions, each compared to the null option. Any expected benefit of an exploratory option in terms of improved *confidence* in decision making can then be balanced—however subjectively—against its expected *costs* in terms of financial or other resources, together with any penalties associated with *delay* in significant future decisions.

At a stage when a group is moving toward a point of pressure to agree on early

decisions or recommendations—as in the closing stages of a workshop session—then it is useful to introduce the framework of the *progress package grid*. A facilitator begins by posting on the wall a blank grid of the form shown in Fig.6, usually spread across two or more flipcharts side by side so as to leave space for additions and alterations. Its function is to facilitate agreement regarding the balance between those decision areas in which it is agreed to make or recommend action *now*, and those others in which commitment is to be deferred until some *future* time. This in turn raises questions about which uncertainty areas are to be addressed now through the choice of some exploratory option—explicitly recognizing, in the final column, the alternative of choosing not to take any such steps.

Figure 6 illustrates for the case of PDQ how semiadhesive stickers can be used to facilitate discussion about the allocation of both decision areas and uncertainty areas either to the *Now* or the *Future* half of the grid. It is often also useful to divide the grid

into rows to indicate different areas of responsibility. While compiling the grid, the facilitator can encourage the participants to refer flexibly to their records of earlier progress in other modes, as displayed on flipcharts elsewhere in the room. For example, when considering actions to be taken now in the more urgent decision areas, reference can be made to an option tree such as that in Fig. 4, resequencing it if necessary so as to bring the more urgent choices to the front. Then the relative *robustness* [4] of options in the more urgent decision areas can be assessed, referring back to any progress in agreeing about preferences in the comparing mode.

Where a working group has sufficient authority to act on the proposals in a progress package without referring to other people, it can be given the more definitive name *commitment package*. Once such a package has been adopted as a basis for action, or for recommendation to others, it is a good discipline to record further details, either on the grid or elsewhere, indicating who is responsible for implementing each action, over what timescale, and by what means. For purposes of presentation to policy makers, it is now often advisable to interpret the contents of the grid in a more conventional reporting format.

As most decision makers would recognize, any action plan such as that shown in Fig. 6 involves striking a judicious compromise between commitment to early action and flexibility of future choice. Its value is both as a platform for concerted implementation and as a starting point for further cycles of decision making, in which the issues to be addressed and the people involved may change in both planned and unpredictable ways. The philosophy and methods of SCA have evolved so as to help decision makers adapt to these realities; realities that

distinguish the world of strategic or *developmental* decision making in the face of multiple uncertainties from the more orderly situations to be expected in the familiar world of OR within a well-defined corporate setting.

FURTHER READING

For further insights into the many adaptations that users of SCA have now introduced, at levels of decision making varying from governmental policy to community empowerment, reference is encouraged to the diverse accounts of applications presented in Chapter 13 of *Planning under Pressure* [2] and elsewhere [5,6].

REFERENCES

1. Friend JK, Norris ME, Stringer J. The Institute for Operational Research: an Initiative to extend the scope of OR. *J Oper Res Soc* 1988;39:705–713.
2. Friend JK, Hickling DA. *Planning under pressure: the strategic choice approach*. 3rd ed. Oxford: Elsevier/Butterworth-Heinemann; 2004.
3. Rosenhead JV, Mingers J, editors. *Rational analysis for a problematic world revisited*. Chichester: Wiley; 2001.
4. Rosenhead JV. Robustness analysis: keeping your options open. In: Rosenhead JV, Mingers J, editors. *Rational analysis for a problematic world revisited*. Chichester: Wiley; 2001. pp. 181–207.
5. Hickling DA. Gambling with frozen fire? In: Rosenhead JV, Mingers J, editors. *Rational analysis for a problematic world revisited*. Chichester: Wiley; 2001. pp. 151–180.
6. Friend JK. Perspectives of engagement in community operational research. In: Midgley G, Ochoa-Arias A, editors. *Community operational research*. Dordrecht: Kluwer; 2004.

THE VEHICLE ROUTING PROBLEM WITH TIME WINDOWS: STATE-OF-THE-ART EXACT SOLUTION METHODS

GUY DESAULNIERS
École Polytechnique and GERAD,
Montreal, Canada

JACQUES DESROSIERS
HEC Montréal and GERAD,
Montréal, Canada

SIMON SPOORENDONK
DTU Management Engineering,
Technical University of Denmark,
Copenhagen, Denmark

INTRODUCTION

Vehicle routing problems are faced, in general, by companies that pickup or deliver goods to their customers by trucks. In such problems, feasible vehicle routes must be computed to fulfill given customer demands at minimum cost. Several problem variants have been studied in the literature. In this paper, we consider the vehicle routing problem with time windows (VRPTW) that can be formally stated as follows. Given an unlimited number of identical vehicles with a specified capacity, and a set of customers, each with a demand to fulfill and a delivery time interval (called a *time window*), the VRPTW consists of finding a set of routes starting and ending at a central depot such that each customer is visited exactly once within its time window, the capacity of each vehicle is respected, and the total cost (length) of the routes is minimized. The time window of a customer restricts its start of service time, but the vehicle servicing it can arrive earlier and wait until the start of the time window before beginning the service.

The VRPTW is often formulated as a set partitioning model based on the following notation. Let C be the set of customers and P the set of all feasible routes, that

is, routes starting and ending at the depot that satisfy the vehicle capacity and the visited customers' time windows. For each route $p \in P$, let c_p be its cost, a_{ip} , $i \in C$, a binary parameter equal to 1 if customer i is visited in route p , and λ_p a binary variable indicating whether or not route p is selected in the solution. The set partitioning model for the VRPTW is

$$\min \sum_{p \in P} c_p \lambda_p \quad (1)$$

$$\text{subject to: } \sum_{p \in P} a_{ip} \lambda_p = 1, \quad \forall i \in C \quad (2)$$

$$\lambda_p \text{ binary}, \quad \forall p \in P. \quad (3)$$

The objective function (1) aims at minimizing the total route cost. Set partitioning constraints (2) ensure that each customer is visited by exactly one vehicle.

Generally, this model contains a huge number of variables. To overcome this difficulty, several researchers have developed efficient branch-cut-and-price algorithms, since the early 1990s. Alternatively, Baldacci *et al.* [1] recently proposed a method that first restricts the set P to routes that have the potential to be part of an optimal solution and then solves the resulting reduced set partitioning model directly using a mixed-integer programming solver. The goal of this article is to present the main ideas behind these state-of-the-art exact methods. Additional references on these methods and alternative ones can be found in the surveys of Cordeau *et al.* [2], Kallehauge *et al.* [3], and Kallehauge *et al.* [4].

BRANCH-AND-PRICE

Branch-and-price [5–7] is the most popular methodology for solving models (1)–(3). It corresponds to a branch-and-bound algorithm where the lower bounds (LB) are computed using column generation. In this context, models (1)–(3) is called the

integer master problem (IMP) and its linear relaxation the *master problem*. Column generation is an iterative method for solving the master problem. At each iteration, it solves a so-called *restricted master problem* (RMP) and a *pricing problem*. The RMP is simply the master problem restricted to a subset of its variables λ_p . The size of this subset increases at each iteration. The RMP is thus a linear program that can be solved by the simplex algorithm. The pricing problem plays the role of determining which columns (variables) should be added to the RMP at each iteration. It aims at finding negative reduced cost columns. When no such columns exist, the current RMP solution is optimal for the master problem and the column generation method stops. Otherwise, negative reduced cost columns are added to the RMP before starting a new iteration.

For the VRPTW, each variable $\lambda_p, p \in P$, is associated with a feasible path in a network $G = (N, A)$, where N and A are the node and arc sets, respectively. The node set N contains one node for each customer $i \in C$ and two nodes, o and o' , representing the depot at the start and the end of the routes, respectively. Thus, $N = C \cup \{o, o'\}$. Each customer $i \in C$ has a demand d_i , an associated service time s_i , and a time window $[e_i, l_i]$ in which it should be visited. To simplify notation, we set $d_o = d_{o'} = s_o = s_{o'} = 0$ and $[e_o, l_o] = [e_{o'}, l_{o'}] = [0, H]$, where H is the length of the planning horizon. The arc set A is given by $A = \{(i, j) : i, j \in N, i \neq j, e_i + \tau_{ij} \leq l_j, d_i + d_j \leq D\} \setminus \{(o', o)\}$, where D is the vehicle capacity and τ_{ij} is the travel time between locations i and j plus the service time s_i at i . The last two conditions ensure that the same vehicle can visit locations i and j consecutively, according to the time and load restrictions. Denote by c_{ij} the travel cost along arc $(i, j) \in A$.

The reduced cost $\bar{c}_p$ of a variable $\lambda_p, p \in P$, is given by

$$\bar{c}_p = c_p - \sum_{i \in C} a_{ip} \pi_i = \sum_{(i, j) \in p} (c_{ij} - \pi_j) \quad (4)$$

where $\pi_j \in \mathbb{R}$ for all $j \in C$ are the dual values of Equation (2) and $\pi_{o'} = 0$. In this expression, $(i, j) \in p$ means that (i, j) is an arc of the path representing route p in G . To

determine if there exists negative reduced cost variables, one can find the shortest feasible path from o to o' in G , where the cost of each arc $(i, j) \in A$ is $\bar{c}_{ij} = c_{ij} - \pi_j$. Feasibility is enforced by resource constraints [8]. A resource is a quantity that accumulates along a path and is restricted at each node to take a value within a prespecified resource interval, called a *resource window*. For the VRPTW, the set of resources $\mathcal{R}$ contains a loading resource (*load*) to ensure that vehicle capacity is satisfied, a time resource (*time*) to respect the customer time windows, and a binary resource (*cust_i*) for each customer $i \in C$ to ensure path elementarity. The pricing problem thus corresponds to an elementary shortest path problem with resource constraints (ESPPRC). If its optimal value is negative, then the computed shortest path provides a negative reduced cost variable that can be added to the RMP (possibly with other found negative reduced cost variables). Otherwise, the column generation algorithm stops.

For solving the IMP, column generation is embedded into a branch-and-bound procedure. A classical branching rule for the VRPTW [9] is to branch on the total flow on an arc $(i, j) \in A$ (with $i \neq o$ and $j \neq o'$) which must be either 0 or 1 in any integer solution. For imposing a flow of 0 on arc (i, j) , this arc is simply removed from A in the pricing problem, while for imposing a flow of 1 on it, all arcs $(i, j') \in A$ such that $j' \neq j$ and all arcs $(i', j) \in A$ such that $i' \neq i$ are removed. For both decisions, all columns λ_p such that path p contains a removed arc are deleted from the current RMP.

FIRST-GENERATION BRANCH-AND-PRICE METHODS

2-Cycle Free Paths

Because the ESPPRC is strongly $\mathcal{NP}$ -hard [10], Desrochers *et al.* [9] proposed a branch-and-price algorithm where the pricing problem is a shortest path problem with resource constraints (SPPRC) that allows cycles (except cycles $i \rightarrow j \rightarrow i$ involving two customers $i, j \in C$, called *two-cycles*). This yields a relaxation of the master problem and possibly large integrality gaps.

The SPPRC can be solved using a labeling algorithm [8]. In such a dynamic programming method, labels represent partial paths that are extended, using so-called *extension functions*, in all feasible directions from the source node o in G . Each label L (a vector with $|\mathcal{R}| + 1$ components) stores the reduced cost of the partial path $T_{\text{rcost}}(L)$ and the current value $T_r(L)$ of each resource $r \in \mathcal{R}$. For the VRPTW, the extension along an arc (i, j) of a label L representing a partial path ending at node i proceeds as follows to create a new label L' representing a partial path ending by the arc (i, j) :

$$T_{\text{rcost}}(L') = T_{\text{rcost}}(L) + c_{ij} - \pi_j \quad (5)$$

$$T_{\text{load}}(L') = T_{\text{load}}(L) + d_j \quad (6)$$

$$T_{\text{time}}(L') = \max\{T_{\text{time}}(L) + \tau_{ij}, e_j\} \quad (7)$$

Label L' corresponds to a feasible path (that may contain cycles) if it respects the resource windows, that is, if $T_{\text{load}}(L') \in [0, D]$ and $T_{\text{time}}(L') \in [e_j, l_j]$. It is discarded if this is not the case.

To avoid enumerating all feasible paths in G , only Pareto-optimal labels (i.e., labels that are not proven to be dominated by other labels) are kept during the execution of the algorithm. When using nondecreasing extension functions as it is the case for the VRPTW, the label dominance criterion can be stated as follows:

Proposition 1 [11]. *Let L and L' be two labels representing partial paths ending at the same node. Label L dominates label L' (which can be discarded) if*

$$T_{\text{rcost}}(L) \leq T_{\text{rcost}}(L') \quad (8)$$

$$T_r(L) \leq T_r(L'), \quad \forall r \in \mathcal{R}. \quad (9)$$

When equality holds for all label components, one of the two labels must be kept.

As an extension to the idea of two-cycle free paths, Irnich and Villeneuve [12] developed a k -cycle elimination procedure for the SPPRC for arbitrary integer values of k that forbids the generation of paths containing at least one q -cycle with $q \in \{2, 3, \dots, k\}$, where a q -cycle is a subpath composed of q

arcs that start and end at the same node. This extension yields a stronger relaxation of the master problem and reduces integrality gaps.

Path Cuts

In a branch-cut-and-price method, violated valid inequalities (also called *cutting-planes* or *cuts*) are added to tighten the master problem when its computed solution is fractional. After adding cutting-planes, the master problem and the pricing problem are updated in consequence, and column generation is again performed to reoptimize the modified master problem and compute a new LB.

Kohl *et al.* [13] proposed the 2-path cuts. Such a cut forces at least two paths to visit a subset of customers when it can be proved, by solving a traveling salesman problem with time windows [14], that a single vehicle cannot serve all of them. For a subset $V \subseteq C$ of customers requiring at least two vehicles, a 2-path cut writes

$$\sum_{p \in P} \sigma_p^V \lambda_p \geq 2, \quad (10)$$

where σ_p^V is equal to the number of arcs in p entering into V . With this parameter definition, the treatment of these cuts does not pose a problem with regards to the complexity of the pricing problem because their dual values simply affect the arc costs. However, because the coefficients σ_p^V can take values greater than 1, these cuts are not as strong as they could be. Baldacci *et al.* [1] proposed a strengthened version of the 2-path cuts in which σ_p^V is equal to 1 if route p visits at least one customer in V and 0 otherwise. The downside of this improvement is that the dual value of such a cut cannot be directly transferred to the arc costs, which complicates the pricing problem. The effect is similar to that for the subset-row inequalities described in the section titled “Subset-Row Inequalities.”

RECENT BRANCH-AND-PRICE METHODS

Elementary Paths

In parallel, several researchers devised algorithms for solving the ESPPRC, allowing the generation of elementary routes and

strengthening the master problem. Feillet *et al.* [15] were the first to propose a successful labeling algorithm for the ESPPRC where one resource is required for each customer to impose elementarity. Beside the extension functions (5)–(7) for reduced cost, load, and time, the following extension functions, one per customer, are needed when extending label L along an arc (i, j) to create label L'

$$T_{\text{cust}_n}(L') = \begin{cases} T_{\text{cust}_j}(L) + 1 & \text{if } n = j \\ \max\{T_{\text{cust}_n}(L), U_n(L')\} & \text{otherwise} \end{cases} \quad \forall n \in C, \quad (11)$$

where $U_n(L')$ indicates whether or not there exist no feasible extensions of label L' reaching node $n \in C$ (n is unreachable from L). Assuming that the arc durations satisfy the triangle inequality, $U_n(L')$ is equal to 1 if $T_{\text{load}}(L') + d_n > D$ or $T_{\text{time}}(L') + \tau_{jn} > l_n$, and 0 otherwise. Label L' is discarded if $T_{\text{cust}_j}(L') > 1$.

Bidirectional labeling algorithms have been proposed by Righini and Salani [16]. These algorithms extend *forward* labels from the source node o and *backward* labels from the sink node o' before joining forward and backward labels to create feasible $o - o'$ paths. In this context, the extension of the backward labels requires backward extension functions that are different from the forward ones (Eqs 5–7 and 11). Furthermore, dominance applies only between forward labels or between backward labels.

Independently, Boland *et al.* [17] and Righini and Salani [18] proposed to initially relax the customer resources and add them iteratively until finding an elementary path. In the former paper this is referred to as *state space augmentation* and in the latter it is denoted as *decremental state space relaxation*.

Heuristic Column Generators

Solving the strongly $\mathcal{NP}$ -hard ESPPRC pricing problem to optimality at each iteration is not essential as long as negative reduced cost columns can be found. Consequently,

one can apply first a fast heuristic for the pricing problem. If it fails to produce a negative reduced cost column, then an exact labeling algorithm can be invoked. Several heuristics have been proposed in the literature. Two popular ones implement heuristic strategies in the exact labeling algorithm. The first strategy consists of eliminating arcs from network G that do not seem promising according to some criterion (for instance, their reduced cost). In the second strategy, an aggressive dominance rule is applied to eliminate a larger number of labels in the dominance procedure. For instance, one may use the standard dominance rule but considering only a (possibly empty) subset of the customer resources.

Recently, Desaulniers *et al.* [19] introduced a tabu search column generator. A tabu search method is iterative and finds at each iteration the best neighbor of the current solution (a path in the case of the ESPPRC). A neighbor is a path that is obtained from the current solution by inserting or removing a customer. To avoid cycling, the best customer insertion or removal cannot be reversed for a number of iterations: the reverse move is thus said to be tabu. To diversify the search, the tabu search method is started with different initial solutions and limited to a maximum number of iterations to perform for each initial solution. The set of initial solutions is given by the paths associated with the basic variables of the current RMP solution. All these paths are good initial candidates because they have a zero reduced cost. Using this tabu search heuristic and heuristic labeling algorithms, Desaulniers *et al.* [19] were forced to call a time-costly exact labeling algorithm only in a few column generation iterations per linear relaxation.

Subset-Row Inequalities

The subset row inequalities introduced by Jepsen *et al.* [20] are a set of valid inequalities directly defined on the master problem variables, whose dual values cannot be directly transferred to the arc costs of the pricing problem. These inequalities form a subset of the Chvátal–Gomory

(CG) rank-1 cuts for the set partitioning polytope. Given the CG multipliers $u_i^g \in [0, 1]$ for $i \in C$, a CG rank-1 cut (indexed by g) for models (1)–(3) is expressed as

$$\sum_{p \in P} \left[\sum_{i \in C} u_i^g a_{ip} \right] \lambda_p \leq \left[\sum_{i \in C} u_i^g \right]. \quad (12)$$

The subset row inequalities considered by Jepsen *et al.* [20] are obtained using a CG multiplier equal to $\frac{1}{2}$ for exactly three set partitioning constraints in Equation (2) and a zero multiplier for all the other constraints. This results in a right-hand side member $\lfloor \sum_{i \in C} u_i^g \rfloor$ of one and binary parameters $b_{gp} = \lfloor \sum_{i \in C} u_i^g a_{ip} \rfloor$ equal to 1 if and only if route p visits at least two of the three customers associated with the constraints involved in cut g . As shown by their computational results, the use of these inequalities can significantly improve the lower bound computed at the root node of the search tree. On the other hand, it increases the complexity of the ESPPRC pricing problem as each cut requires one additional resource.

Petersen *et al.* [21] extended this work and showed how any CG rank-1 cut (Eq. (12)) can be applied when using a branch-cut-and-price algorithm for solving models (1)–(3). In this case, the reduced cost of a variable λ_p also depends on the positive coefficients $\lfloor \sum_{i \in C} u_i^g a_{ip} \rfloor$ of the CG cuts and their corresponding dual values. As discussed next, the computation of these coefficients in a labeling algorithm is not straightforward because of their stepwise nature. Let $\mathcal{CG1}$ be the set of CG rank-1 cuts added to the master problem. Denote by β_g the non-positive dual variable associated with cut $g \in \mathcal{CG1}$ and by $\mathbf{u}^g$ the CG multiplier vector of size $|C|$ defining this cut. For each cut $g \in \mathcal{CG1}$, an additional resource is defined. This resource computes the current fractional part of $\sum_{i \in C} u_i^g a_{ip}$, that is, for a label L representing a partial path that visits the subset of customers $C(L)$, the value of this resource is given by $T^g(L) = \sum_{i \in C(L)} u_i^g - \lfloor \sum_{i \in C(L)} u_i^g \rfloor$. It is not restricted by resource windows. In the labeling algorithm, the forward extension of a label L along an arc $(i, j) \in A$ to create a new label L' proceeds as follows for the cut

and reduced cost components

$$T^g(L') = (T^g(L) + u_j^g) \pmod{1}, \quad \forall g \in \mathcal{CG1} \quad (13)$$

$$T_{\text{rcost}}(L') = T_{\text{rcost}}(L) + c_{ij} - \pi_j - \sum_{\substack{g \in \mathcal{CG1} \\ T^g(L) + u_j^g \geq 1}} \beta_g, \quad (14)$$

with $u_{o'}^g = 0$. The computation of the other components $T_r(L')$, $r \in \mathcal{R}$, is performed using the extension functions (6)–(7) and (11).

To discard labels in the labeling algorithm when using CG rank-1 cuts, the dominance rule of Proposition 1 can be applied considering the resource set $\mathcal{R} \cup \mathcal{CG1}$. However, this dominance criterion can be enhanced as follows:

Proposition 2 [21]. *Let L and L' be two labels representing partial paths ending at the same node. Label L dominates label L' (which can be discarded) if*

$$T_{\text{rcost}}(L) - \sum_{\substack{g \in \mathcal{CG1} \\ T^g(L) > T^g(L')}} \beta_g \leq T_{\text{rcost}}(L') \quad (15)$$

$$T_r(L) \leq T_r(L'), \quad \forall r \in \mathcal{R}. \quad (16)$$

This dominance criterion is a generalization of the one proposed by Jepsen *et al.* [20], but it is equivalent when considering only the subset row inequalities.

The separation of the CG rank-1 cuts is an $\mathcal{NP}$ -hard problem, and so is the separation of the subset row inequalities [20]. Consequently, Jepsen *et al.* [20] as well as Desaulniers *et al.* [19] and Baldacci *et al.* [1] in subsequent works considered only subset row inequalities involving exactly three constraints and used partial enumeration to find the violated ones.

SET PARTITIONING MODEL WITH ROUTE ELIMINATION

Baldacci *et al.* [1] recently proposed the most efficient exact solution method. Their algorithm combines Lagrangean relaxation and column/row generation to find a LB

on the optimal value of the IMP. The final (near optimal) dual solution is then used to eliminate all the variables whose reduced cost is larger than or equal to the gap between an upper bound (UB) on the optimal value of the IMP and LB, the LB achieved. Finally, the reduced IMP is solved with an integer programming solver.

The UB is usually computed using a fast and efficient heuristic for the VRPTW. The LB is derived from a feasible dual solution to the master problem. Baldacci *et al.* [1] find this dual solution by a column generation method that proceeds in three stages. Each stage considers as a master problem at different relaxation of the IMP. These stages are executed in sequence, the second and third ones exploiting the dual solution and the columns are generated in the previous stage. In the first two stages, the RMP is solved with a Lagrangean relaxation method based on a subgradient algorithm to avoid high degeneracy that may arise with the simplex algorithm. This dual method derives the dual variables $\pi_i, i \in C$, to price out and generate new columns.

The column generation/Lagrangean relaxation procedure $\mathcal{H}^1$ for the first stage generates *ng*-routes. An *ng*-route is obtained by adding arc (o, i) to an *ng*-path (NG, t, i) defined as a path (not necessarily elementary) starting from node $i \in C$ at time $t \in [e_i, l_i]$, visiting at least once each customer in a subset $NG \subseteq C$ within its time window, and ending at the depot at time H . In this procedure, the pricing problem generates routes with a forward labeling algorithm. The second-stage procedure $\mathcal{H}^2$ generates elementary routes satisfying both time and capacity constraints. These routes are also generated by a forward labeling algorithm that uses *ng*-paths to calculate LBs on the reduced costs of the completions of the partial elementary paths represented by the forward labels. If the calculated LB for a label is greater than 0, then this label can be discarded. The third-stage procedure $\mathcal{H}^3$ is a classical column/row generation algorithm based on the simplex algorithm to solve the master problem strengthened by the subset row inequalities involving exactly three customers. In this

procedure, the elementary routes are generated using a bidirectional labeling algorithm. As above, label fathoming using *ng*-paths (and their symmetric counterparts) is also used. Furthermore, a route can be fathomed if its reduced cost with respect to the feasible dual solution found by $\mathcal{H}^2$ exceeds the difference between *UB* and the LB achieved by $\mathcal{H}^2$. At the end of procedure $\mathcal{H}^3$, we have on hand a LB *LB* and dual variable values $\pi_i, i \in C$, associated with the set partitioning constraints, and $\beta_g, g \in SR3$, associated with the generated subset row inequalities, where *SR3* denote the subset of those cuts saturated in the solution of the final RMP.

After computing this dual solution, the proposed algorithm enumerates, using a specialized dynamic programming algorithm, all routes $p \in P$, whose reduced cost

$$\bar{c}_p = c_p - \sum_{i \in C} a_{ip} \pi_i - \sum_{g \in SR3} b_{gp} \beta_g, \quad (17)$$

satisfies $\bar{c}_p < UB - LB$, before constructing a reduced set partitioning type models (1)–(3) augmented by the subset row inequalities in *SR3*. The resulting reduced problem is then solved by a commercial branch-and-cut solver. The effectiveness of this method highly depends on the quality of the computed dual solution $\pi_i, i \in C$ and $\beta_g, g \in SR3$. Better quality solutions yield smaller reduced problems that are, in general, easier to solve.

This algorithm was tested on the Solomon's benchmark instances that are available at <http://web.cba.neu.edu/~msolomon/problems.htm>. This benchmark set comprises 168 instances that are divided into three groups of 56 instances involving 25, 50, and 100 customers, respectively. Each group contains six instance classes (C1, RC1, R1, C2, RC2, and R2) differing by the geographical distribution of the customers or the average width of the time windows. The algorithm of Baldacci *et al.* [1] was able to solve to optimality 167 of these 168 instances, outperforming all other existing exact algorithms. On average, it was five to six (respectively seven to eight) times faster than the branch-cut-and-price algorithm of Jepsen *et al.* [20] (respectively Ref. 19) on the instances solved by both algorithms. The

only instance that remains open is the R208 instance with 100 customers.

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

The first attempt to solve the VRPTW by column generation dates back two decades. A lot has been done since that time but still, one test-instance from the Solomon's benchmark set fails to be solved and its most difficult instances have only been solved in the last few years. Advances have been made by the combination of various techniques: bidirectional dynamic programming labeling algorithms to compute elementary routes, various cutting-planes to strengthen the LBs, new label dominance and fathoming rules to accelerate the labeling algorithms, Lagrangean relaxation methods to avoid degeneracy of the master problem, heuristic column generators to speed up the overall process, and so on.

Future research will certainly concern the acceleration of these techniques. In particular, three issues need to be investigated. First, efficient algorithms for solving the ESPPRC are still missing. Branch-and-cut and decomposition-based algorithms seem interesting alternatives to dynamic programming. Arc elimination to reduce the size of the pricing problem should also be considered. Second, all state-of-the-art algorithms would benefit from new types of cutting-planes that can substantially improve the LBs. Finally, good optimal dual value estimates could be exploited in a dual variable stabilization procedure to reduce the number of column generation iterations and speed up the overall solution process.

REFERENCES

1. Baldacci R, Bartolini E, Mingozzi A, *et al.* An exact solution framework for a broad class of vehicle routing problems. *Comput Manag Sci* 2010;7(3):229–268.
2. Cordeau J-F, Desaulniers G, Desrosiers J, *et al.* The VRP with time windows. In: Toth P, Vigo D, editors. *The vehicle routing problem. Monographs on discrete mathematics and applications*. Philadelphia, PA: SIAM (Society for Industrial and Applied Mathematics); 2002. pp. 157–194.
3. Kallehauge B, Larsen J, Madsen OBG, *et al.* Vehicle routing problem with time windows. In: Desaulniers G, Desrosiers J, Solomon MM, editors. *Column generation*. New York (NY): Springer; 2005. pp. 67–98.
4. Kallehauge B. Formulations and exact algorithms for the vehicle routing problem with time windows. *Comput Oper Res* 2008; 35(7):2307–2330.
5. Barnhart C, Johnson EL, Nemhauser GL, *et al.* Branch-and-price: column generation for solving huge integer programs. *Oper Res* 1998;46(3):316–329.
6. Desaulniers G, Desrosiers J, Solomon MM. *Column generation*. New York: Springer; 2005.
7. Lübbecke ME, Desrosiers J. Selected topics in column generation. *Oper Res* 2005; 53(6):1007–1023.
8. Irnich S, Desaulniers G. Shortest path problems with resource constraints. In: Desaulniers G, Desrosiers J, Solomon MM, editors. *Column generation*. New York (NY): Springer; 2005. pp. 33–66.
9. Desrochers M, Desrosiers J, Solomon MM. A new optimization algorithm for the vehicle routing problem with time windows. *Oper Res* 1992;40:342–354.
10. Dror M. Note on the complexity of the shortest path models for column generation in VRPTW. *Oper Res* 1994;42:977–979.
11. Desaulniers G, Desrosiers J, Ioachim I, *et al.* A unified framework for deterministic time constrained vehicle routing and crew scheduling problems. In: Crainic TG, Laporte G, editors. *Fleet management and logistics*. Norwell (MA): Kluwer; 1998. pp. 57–93.
12. Irnich S, Villeneuve D. The shortest path problem with resource constraints and k -cycle elimination for $k \geq 3$. *INFORMS J Comput* 2006;18:391–406.
13. Kohl N, Desrosiers J, Madsen OBG, *et al.* 2-Path cuts for the vehicle routing problem with time windows. *Transport Sci* 1999;33(1):101–116.
14. Dumas Y, Desrosiers J, Solomon MM, *et al.* An optimal algorithm for the traveling salesman problem with time windows. *Oper Res* 1993;43:367–371.
15. Feillet D, Dejax P, Gendreau M, *et al.* An exact algorithm for the elementary shortest path problem with resource constraints:

- application to some vehicle routing problems. *Networks* 2004;44(3):216–229.
16. Righini G, Salani M. Symmetry helps: bounded bi-directional dynamic programming for the elementary shortest path problem with resource constraints. *Discrete Optim* 2006;3(3):255–273.
 17. Boland N, Dethridge J, Dumitrescu I. Accelerated label setting algorithms for the elementary resource constrained shortest path problem. *Oper Res Lett* 2006;34:58–68.
 18. Righini G, Salani M. New dynamic programming algorithms for the resource constrained elementary shortest path problem. *Networks* 2008;51(3):155–209.
 19. Desaulniers G, Lessard F, Hadjar A. Tabu search, partial elementarity, and generalized k -path inequalities for the vehicle routing problem with time windows. *Transport Sci* 2008;42(3):387–404.
 20. Jepsen M, Petersen B, Spoorendonk S, *et al.* Subset-row inequalities applied to the vehicle-routing problem with time windows. *Oper Res* 2008;56(2):497–511.
 21. Petersen B, Pisinger D, Spoorendonk S. Chvátal-Gomory rank-1 cuts used in a Dantzig-Wolfe decomposition of the vehicle routing problem with time windows. In: Golden B, Raghavan R, Wasil E, editors. *The vehicle routing problem: latest advances and new challenges*. New York (NY): Springer; 2008. pp. 397–420.

THE WEIGHTED MOVING AVERAGE TECHNIQUE

MARCUS B. PERRY

Department of Information Systems,
Statistics and Management Science,
The University of Alabama,
Tuscaloosa, Alabama

INTRODUCTION

The w -term moving average (MA) is most often applied to time series data as a means to smooth out seasonal variation or irregular fluctuations in the data, permitting the analyst to more easily identify the structural patterns. It is also used as a means for computing short-term forecasts of time series. Its application in practice is widespread across many disciplines. For example, in an engineering application, one might use the MA in efforts to control the quality of a product or process. Further, in a business application, the MA might be used to forecast future sales demand or employee turnover. Depending on the application, different types of MAs can be computed (e.g., simple, weighted, and cumulative). Further, MAs can be centered by placing each w -term average in the middle of its corresponding w -length time interval. If w is odd, then by this placement, the MA is centered. On the other hand, if w is even, a 2-term moving average is required to center the original MA. Centered MAs are more typical when the primary objective of the analysis is seasonal adjustment of the data.

In this tutorial, emphasis is placed on the *weighted* moving average (WMA), where each w -term average is placed at the upper bound of its corresponding w -length time interval. In particular, consider the n -length time series $\{x_i\}$, then its w -term WMA is defined as

$$z_i = \sum_{j=i-w+1}^i \omega_j x_j, \quad (1)$$

for $i \in [w, w+1, \dots, n]$, where the ω_j 's are weights and the x_j 's are a time sequence of (possibly autocorrelated) random variables. Notice that the number of observations in the series $\{z_i\}$ is reduced to $n - w + 1$; that is, $w - 1$ terms are lost in the WMA. As noted by one of the reviewers, the right-hand side of Equation (1) defines a convex hull for the data points, and thus z_i is a point in the convex hull of data up to the present. The WMA defined in Equation (1) is useful for computing one-step-ahead (i.e., short-term) forecasts or predictions. Further, it lends nicely to the development of process monitoring algorithms useful in, say, statistical quality control applications.

The WMA in Equation (1) can be written in vector notation as

$$\begin{aligned} z_i &= (\omega_{i-w+1}, \omega_{i-w+2}, \dots, \omega_i) \begin{pmatrix} x_{i-w+1} \\ x_{i-w+2} \\ \vdots \\ x_i \end{pmatrix} \\ &= \omega_i' \mathbf{x}_i, \end{aligned} \quad (2)$$

where ω_i and $\mathbf{x}_i$ are both $w \times 1$ and $\sum_{k=1}^w \omega_{i-w+k} = 1$. If the ω_i 's are constant for all i , then z_i can be written as

$$z_i = (\omega_1, \omega_2, \dots, \omega_w) \begin{pmatrix} x_{i-w+1} \\ x_{i-w+2} \\ \vdots \\ x_i \end{pmatrix} = \omega' \mathbf{x}_i, \quad (3)$$

where $\sum_{k=1}^w \omega_k = 1$, and whereby applying larger weights to the more recent observations yields the additional constraint $\omega_1 \leq \omega_2 \leq \dots \leq \omega_w$. Consider the special case where $\omega_k = w^{-1}$ for all k , then this produces the w -term *simple* MA given by

$$z_i = w^{-1} \sum_{j=i-w+1}^i x_j, \quad (4)$$

which is just the arithmetic average of the w terms.

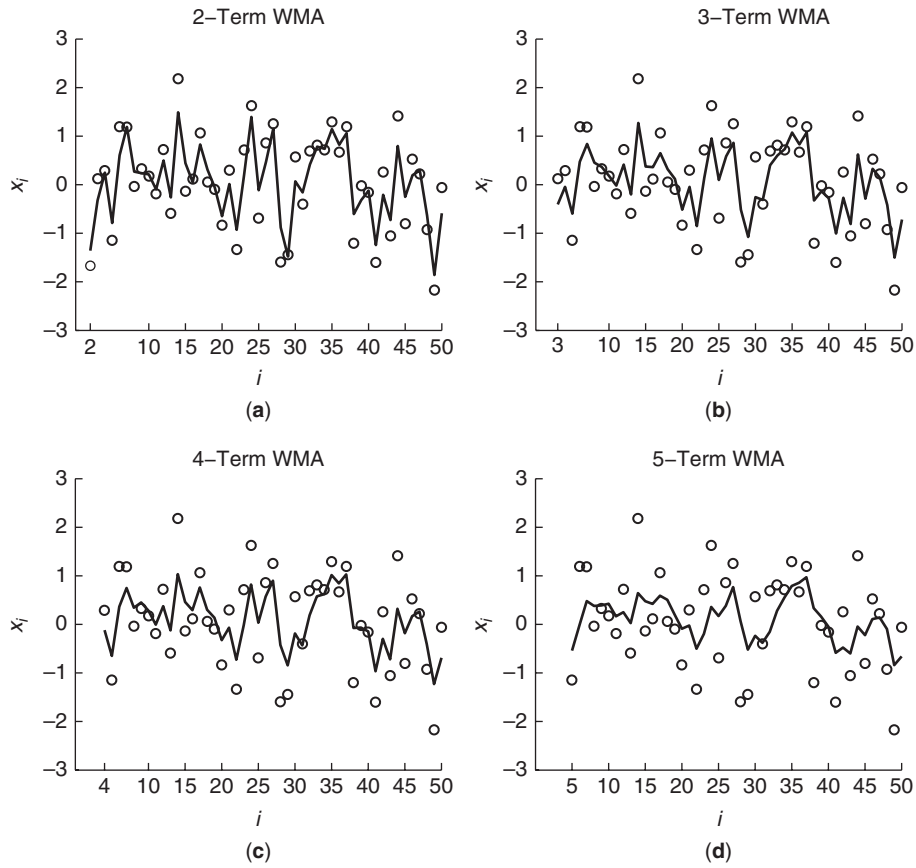


Figure 1. Examples of weighted moving averages.

To illustrate the computation of a w -term simple MA, consider $w = 3$ and the series

$$\{x_i\} = [1.3 \quad 2.5 \quad 4.1 \quad 2.9 \quad 1.6].$$

Since $n = 5$ and $w = 3$, the new series $\{z_i\}$ will contain three observations (i.e., z_3, z_4, z_5) computed, respectively, as

$$z_3 = (1.3 + 2.5 + 4.1)/3 = 2.63$$

$$z_4 = (2.5 + 4.1 + 2.9)/3 = 3.17$$

$$z_5 = (4.1 + 2.9 + 1.6)/3 = 2.87,$$

and thus the new series $\{z_i\}$ (for $i = 3, 4, 5$) is given as¹

$$\{z_i\} = [\cdot \quad \cdot \quad 2.63 \quad 3.17 \quad 2.87].$$

¹Note that the centered MA for this case is given as $\{z_i\} = [\cdot \quad 2.63 \quad 3.17 \quad 2.87 \quad \cdot]$.

To illustrate further, consider the scatter plots of $n = 50$ independently generated standard normal random deviates shown in Fig. 1. For each plot, the horizontal axis denotes time and the vertical axis denotes the observed value of X at time i . The solid line in Fig. 1a shows a 2-term WMA with $\omega_1 = 0.25$ and $\omega_2 = 0.75$. Similarly, the solid line in Fig. 1b shows a 3-term WMA with $\omega_1 = 0.15$, $\omega_2 = 0.25$, and $\omega_3 = 0.60$. A 4-term WMA with $\omega_1 = 0.10$, $\omega_2 = 0.15$, $\omega_3 = 0.25$, and $\omega_4 = 0.50$ is given in Fig. 1c, while a 5-term WMA with $\omega_1 = 0.10$, $\omega_2 = 0.15$, $\omega_3 = 0.20$, $\omega_4 = 0.25$, and $\omega_5 = 0.30$ is shown in Fig. 1d.

It seems apparent from Fig. 1 that as the number of terms in the MA increases, so does the amount by which $\{x_i\}$ is smoothed. The implication is that the variance of z_i reduces with the addition of more terms; however, at the expense of bias. This is quite

intuitive and in-line with the well-known variance-bias trade-off. This is discussed in more detail in a later section.

In the next section several properties of z_i are evaluated, including the mean, variance, and the variance of the one-step-ahead forecast errors. In subsequent sections, some common applications of the WMA in practice are discussed.

PROPERTIES

Let $\mathbf{x}_i = (x_{i-w+1}, x_{i-w+2}, \dots, x_i)'$ be a $w \times 1$ random vector with mean μ_i and finite autocovariance matrix $\Sigma = \sigma_x^2 \mathbf{R}$, where $\mathbf{R}$ denotes the $w \times w$ autocorrelation matrix of $\mathbf{x}_i$ and σ_x^2 denotes the variance of the original process $\{x_i\}$. Note the special case of the uncorrelated process when $\mathbf{R} = \mathbf{I}$. Note further that μ_i is $w \times 1$ and is given explicitly as²

$$\mu_i = (\mu_{i-w+1} \quad \mu_{i-w+2} \quad \cdots \quad \mu_i)' \quad (5)$$

and $\mathbf{R}$ has the form³

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{w-1} \\ \rho_{-1} & 1 & \rho_1 & \cdots & \rho_{w-2} \\ \rho_{-2} & \rho_{-1} & 1 & \cdots & \rho_{w-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{-w+1} & \rho_{-w+2} & \rho_{-w+3} & \cdots & 1 \end{pmatrix}, \quad (6)$$

where $\rho_m = \rho_{-m}$ denotes the lag m autocorrelation of $\mathbf{x}_i$.

Applying the expectation operator to z_i we obtain

$$E(z_i) = \omega' \mu_i \quad (7)$$

and by applying the variance operator to z_i we find

$$\sigma_z^2 = \sigma_x^2 (\omega' \mathbf{R} \omega), \quad (8)$$

where for uncorrelated processes

$$\sigma_z^2 = \sigma_x^2 \sum_{k=1}^w \omega_k^2. \quad (9)$$

Suppose that one uses z_i to forecast x_{i+1} , then applying the variance operator to $z_i - x_{i+1}$ yields the variance of the one-step-ahead forecast errors (i.e., prediction errors), or

$$\sigma_{\text{pred}}^2 = \sigma_x^2 \left(1 + \omega' \mathbf{R} \omega - 2 \sum_{k=1}^w \omega_k \rho_{w-k+1} \right), \quad (10)$$

where for the uncorrelated process $\rho_m = 0$ ($m > 0$) and Equation (10) reduces to

$$\sigma_{\text{pred}}^2 = \sigma_x^2 + \sigma_z^2, \quad (11)$$

which is the sum of the variances of the original process $\{x_i\}$ and the WMA process $\{z_i\}$.

In much of what follows, references will be made to different time series models for autocorrelated processes. These models consist of the class of stationary and invertible autoregressive moving average (ARMA) models discussed in Box *et al.* [1]. Additional references to these models include Abraham and Ledolter [2] and Pankratz [3]. The notation $\text{ARMA}(p, q)$ is often used to denote the ARMA model of order p and q ; that is, we have a model with p autoregressive terms and q moving average terms. The notation is shortened further when the model contains only autoregressive or moving average terms. That is, a model containing p autoregressive terms and no moving average terms is denoted by $\text{AR}(p)$. Similarly, a model containing q moving average terms and no autoregressive terms is denoted by $\text{MA}(q)$. This notation will be used in subsequent sections of this tutorial.

3-Term WMA Applied to AR(1) Process

Suppose that the process under consideration is well characterized by a first-order autoregressive model, or

$$x_i = \mu + \phi(x_{i-1} - \mu) + \epsilon_i. \quad (12)$$

²Note that the original process $\{x_i\}$ need not be mean stationary.

³For covariance stationary processes, $\text{Var}(\mathbf{x}_i) = \sigma_x^2 \mathbf{R}$ for all $i \Rightarrow$ covariance is only a function of lag.

Suppose further that $\sigma_\epsilon^2 = 1$ and $\phi = 0.50$, then it can be shown that $\sigma_x^2 = (1 - \phi^2)^{-1} = 1.3333$. Also, for a 3-term WMA, we have the correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.50 & 0.25 \\ 0.50 & 1.00 & 0.50 \\ 0.25 & 0.50 & 1.00 \end{pmatrix} \quad (13)$$

so that for $\omega = [0.15, 0.25, 0.60]'$ we obtain for all i

$$E(z_i) = \mu \quad (14)$$

$$\sigma_z^2 = 0.9033, \quad (15)$$

and the variance of the one-step-ahead forecast errors obtained from Equation (10) is given by⁴

$$\sigma_{\text{pred}}^2 = 1.2200. \quad (16)$$

To put this example into practical context, suppose the process in Equation (12) well models a sales demand process and one is interested in monitoring for changes in mean sales demand over time. To accomplish this, one can plot z_i for each i and compare to the limits $\mu_0 \pm L\sigma_z$, where μ_0 is some hypothesized *expected* value of sales demand, $\sigma_z = \sqrt{0.9033}$ and L is a constant expressed in standard deviation units.⁵ If any point exceeds these limits, one can conclude that the mean sales demand has changed, and thus plans might be devised at this point to increase or decrease production for the next period.

5-Term WMA Applied to ARMA(1,1) Process

Suppose now that the process under consideration is well modeled by the ARMA(1,1) model with $\phi = 0.75$ and $\theta = -0.35$, or

$$x_i = \mu + 0.75(x_{i-1} - \mu) + 0.35\epsilon_{i-1} + \epsilon_i. \quad (17)$$

Suppose again that $\sigma_\epsilon^2 = 1$, then it can be shown that

$$\sigma_x^2 = \frac{1 + \theta^2 - 2\phi\theta}{(1 - \phi^2)} = 3.7657 \quad (18)$$

and for a 5-term WMA we have the correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.8429 & 0.6322 & 0.4742 & 0.3556 \\ 0.8429 & 1.0000 & 0.8429 & 0.6322 & 0.4742 \\ 0.6322 & 0.8429 & 1.0000 & 0.8429 & 0.6322 \\ 0.4742 & 0.6322 & 0.8429 & 1.0000 & 0.8429 \\ 0.3556 & 0.4742 & 0.6322 & 0.8429 & 1.0000 \end{pmatrix}, \quad (19)$$

so that for $\omega = [0.10, 0.15, 0.20, 0.25, 0.30]'$ we obtain for all i

$$E(z_i) = \mu \quad (20)$$

$$\sigma_z^2 = 2.8163 \quad (21)$$

and the variance of the one-step-ahead forecast errors obtained from Equation (10) is given by

$$\sigma_{\text{pred}}^2 = 2.1703. \quad (22)$$

Notice that $\sigma_{\text{pred}}^2 < \sigma_z^2$ for this process,⁶ and thus for our sales demand monitoring example described above, one might consider monitoring the one-step-ahead forecasts $x_i - z_{i-1}$ in place of z_i to gain a reduction in variance of the charting statistic. That is, the quantity $x_i - z_{i-1}$ would be computed and compared to the limits $\pm L\sqrt{2.1703}$.

Figure 2 shows a single realization of the ARMA(0.75, -0.35) process with two 5-term WMAs superimposed. The WMA shown as the solid line has weights $\omega_1 = 0.10$, $\omega_2 = 0.15$, $\omega_3 = 0.20$, $\omega_4 = 0.25$, and $\omega_5 = 0.30$, while the WMA shown as the dotted line has weights $\omega_k = 0.20$ for $k = 1, 2, \dots, 5$, which is the 5-term *simple* MA. Note that for this

⁴Note also that $E(z_i - x_{i+1}) = 0$.

⁵ L is often chosen on the basis of an acceptable type-I error.

⁶There is more uncertainty associated with $\{z_i\}$ than with $\{x_i - z_{i-1}\}$ due to strong positive autocorrelation. In general, $\sigma_{\text{pred}}^2 < \sigma_z^2$ when $\sum_{k=1}^w \omega_k \rho_{w-k+1} > \frac{1}{2}$.

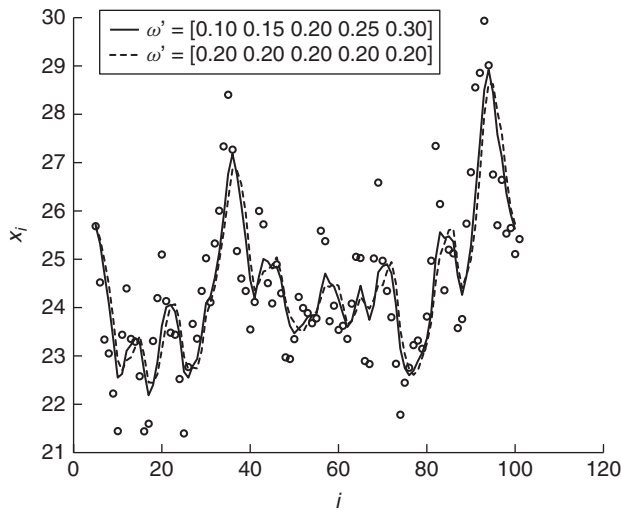


Figure 2. Single realization of ARMA(0.75, -0.35) process with two 5-term WMAs.

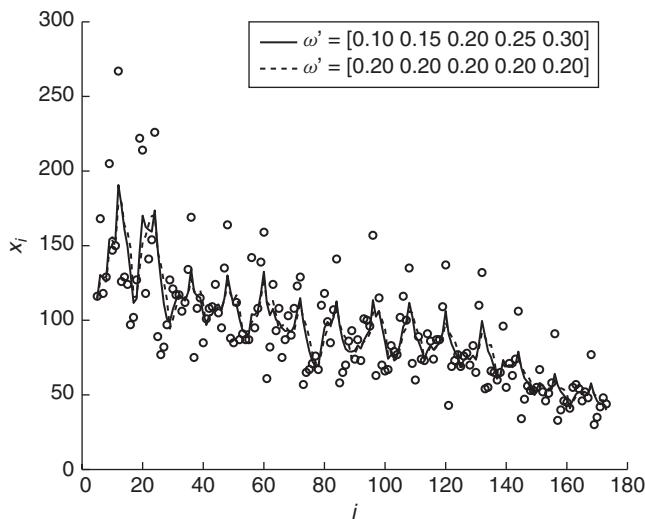


Figure 3. Monthly rose wine sales data with a 5-term simple MA and a 5-term WMA superimposed. Total number of observations is $n = 173$ and $w = 4$ terms are lost.

realization, the two MAs track closely, with only marginal differences. In the next section some of the more common applications of the WMA in practice are discussed.

APPLICATIONS OF THE WEIGHTED MOVING AVERAGE

One of the most common applications of the WMA is generating one-step-ahead forecasts of time series. For example, consider Fig. 3, where monthly Australian sales

of rose wine is plotted over 173 months from approximately January 1980 to April 1995. The data and its original source can be obtained from the Time Series Data Library.⁷ Superimposed on this data are a 5-term WMA⁸ (solid line) and a 5-term simple MA (dotted line). Notice that this

⁷The Time Series Data Library is accessible at <http://www.robjhyndman.com/TSDL/>.

⁸Note that $\omega = [0.10, 0.15, 0.20, 0.25, 0.30]$ for this application.

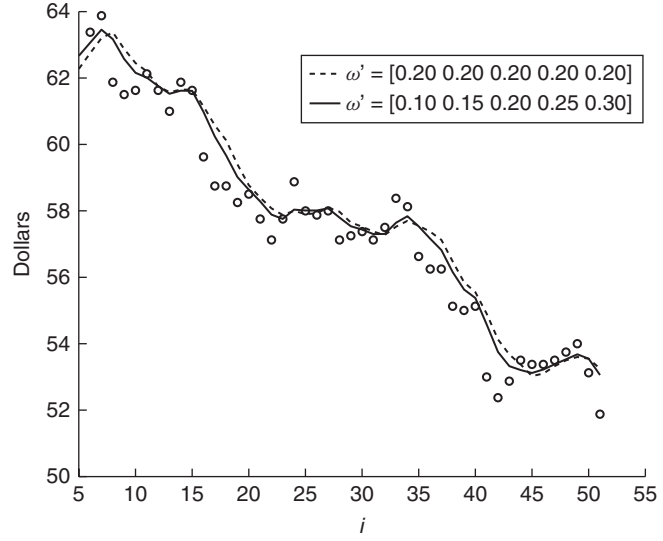


Figure 4. Weekly closing price of AT&T common shares. Total number of observations is $n = 51$ and $w = 4$ terms are lost.

process has some significant seasonal effects evidenced by the overall trend and cyclical-type fluctuations in the MAs. The forecast for x_{174} is then given by $\hat{x}_{174} = z_{173}$, which for the simple MA and WMA cases we have $\hat{x}_{174} = 39.80$ and $\hat{x}_{174} = 41.85$, respectively. The actual sales in the 174th month was $x_{174} = 45.00$, thus, for this realization, the WMA provides a more accurate forecast than that provided by the simple MA.

Consider a second example taken from Pankrats [3] involving the weekly closing price of AT&T common shares. The data set consists of $n = 52$ observations taken over each week in 1979. Figure 4 shows the data up through the 51st week along with 5-term simple and WMAs superimposed. If the WMA is used to forecast closing price in the 52nd week, then $\hat{x}_{52} = 53.05$, and the forecast using the simple MA is $\hat{x}_{52} = 53.25$. The actual closing price at week 52 was $x_{52} = 52.25$. Thus, again we see that for this realization, applying higher weights to more recent observations provides a more accurate forecast (albeit marginal in this case).

We can improve on the *accuracy* of the forecast by reducing the number of terms in the MA; however, this will result in an increase in σ_z^2 (i.e., a decrease in precision). For example, Fig. 5 shows three different WMAs with $w = 2, 3$, and 5, respectively.

Notice that the 2-term WMA forecast for x_{52} is $\hat{x}_{52} = 52.19$, which is very close to the actual value of $x_{52} = 52.25$. However, the amount by which z_i fluctuates when $w = 2$ is noticeably greater if, say, $w = 5$.

This reduction in σ_z^2 as w increases is easily shown for the uncorrelated process. If the observations are uncorrelated, then Equation (8) can be written as

$$\sigma_z^2 = \sigma_x^2(\omega' \omega) = \sigma_x^2 \sum_{k=1}^w \omega_k^2 \quad (23)$$

and for the simple MA $\omega_k = w^{-1}$ for all $k = 1, 2, \dots, w$ and

$$\sigma_z^2 = \frac{\sigma_x^2}{w} \quad (24)$$

so that as $w \rightarrow \infty$ then $\sigma_z^2 \rightarrow 0$.

The above examples demonstrate how the WMA can be applied to generate one-step-ahead forecasts of a time series. Typically, in these applications, the WMA is useful when the forecast horizon is short-term [4]. Other methods are more frequently used for computing mid-to-long-term forecasts.

Another common application of the WMA is in the area of quality control. In particular, the WMA is often used for monitoring critical process quality characteristics over time

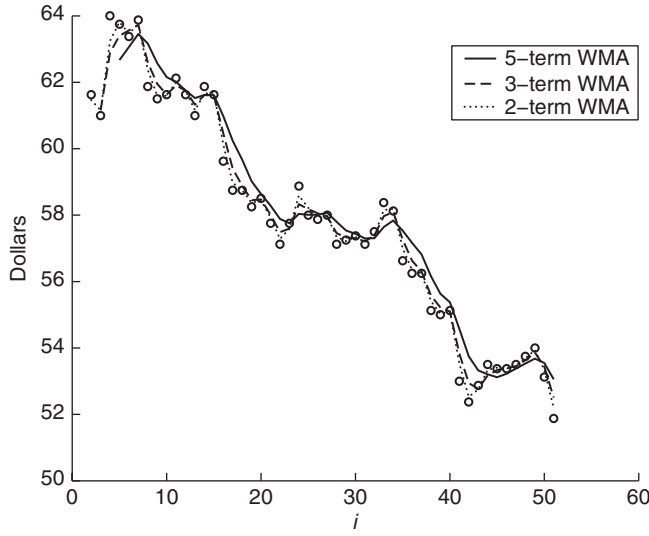


Figure 5. Three different WMAs of weekly closing price of AT&T common shares. Total number of observations is $n = 51$.

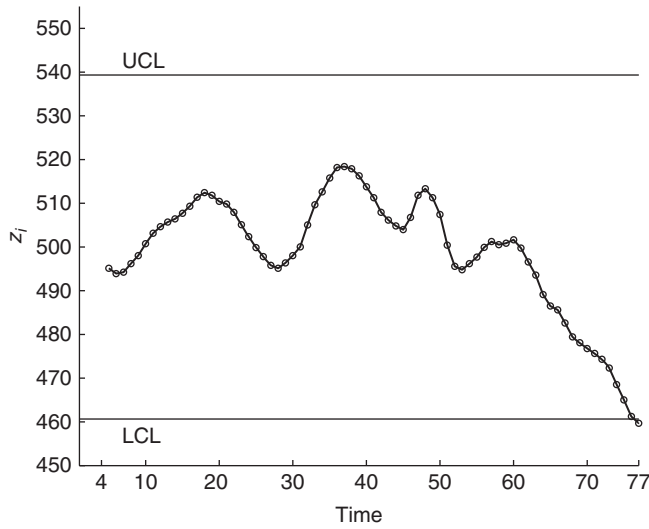


Figure 6. 5-term *two-sigma* WMA control chart applied to simulated ARMA(0.95, -0.65) process. Note the control chart signaled on the 77th observation, suggesting a decrease in the expected yield of the process.

[5,6]. For example, suppose that the process under consideration is a chemical process where the output being monitored is yield. Suppose further that the process is highly autocorrelated and its *expected* behavior is well modeled by the ARMA(0.95, -0.65) process, or

$$x_i = 500 + 0.95(x_{i-1} - 500) + 0.65\epsilon_{i-1} + \epsilon_i \quad (25)$$

for all i with $\sigma_\epsilon^2 = 15.00$. Thus, $\sigma_x^2 = 408.8462$ and for a 5-term WMA,⁹ we find

$$\sigma_z^2 = 387.0393. \quad (26)$$

Suppose it is of interest to detect changes in the expected yield of the process, say from $\mu_i = 500$ to $\mu_i = \mu_1$, where $\mu_1 \neq 500$.

⁹ $\omega' = [0.10, 0.15, 0.20, 0.25, 0, 30]$.

The WMA can be used as a basis for constructing a *control chart*.¹⁰ That is, we could compute z_i (at each time i) and compare to the upper and lower control limits $500 \pm L\sqrt{\sigma_z^2} = 500 \pm 19.6733L$, where L is a constant. If z_i exceeds either of these limits, then a signal is issued and a special-cause investigation is initiated in efforts to uncover the underlying cause of the process change. The constant L is expressed in standard deviation units and is set to achieve an acceptable false alarm rate. If we set $L = 2$, then Fig. 6 shows a *two-sigma* WMA control chart applied to a simulated realization of the ARMA(0.95, -0.65) process with $\sigma_\epsilon^2 = 15.00$. Initially, the process was simulated so that $\mu_i = 500$ for all $i = 1, 2, \dots, 50$. However, for $i > 50$ the process was simulated so that $\mu_i = 485$. Notice that $z_{77} < \text{LCL}$, which suggest that the expected yield of the process has decreased.¹¹ Since high yield is desirable, production would cease at this point and a special-cause search is initiated. When using $\{z_i\}$ as a control charting statistics, smaller shifts in the process mean are guarded against most effectively with larger w ; however, at the expense of quick response to large shifts [7]. Therefore, if the detection of smaller-sized shifts is of interest, w should be set sufficiently large. For larger-sized shifts, w should be set sufficiently small. Sparks [5] provides further discussion on strategies for choosing ω and w when using the WMA for process monitoring.

The above illustrations are not the only applications of the WMA; however, they are common applications of the WMA in the management science and quality engineering disciplines. In general, application of the WMA is widespread throughout several

disciplines, including health, finance, criminal justice, sports, industry, demography, and many more.

SUMMARY

In this tutorial, the WMA was discussed. As previously noted, the WMA is often used for smoothing irregular fluctuations (i.e., noise) in a time series to permit the data analyst to better reveal the trend-cycle patterns over time. Additionally, the WMA is frequently used to compute short-term forecasts of time series (e.g., sale and stocks). Some common applications of the WMA in the management science and quality engineering disciplines were also discussed. Further, the expected value, variance, and one-step-ahead prediction variance of the WMA were evaluated. For more information on the WMA and its application, see Bowerman *et al.* [8], Abraham and Ledolter [2], and Makridakis *et al.* [9].

REFERENCES

1. Box GEP, Jenkins GM, Reinsel RC. Time series analysis: forecasting and control. New Jersey: Prentice Hall; 1983.
2. Abraham B, Ledolter J. Statistical methods for forecasting. New York: John Wiley & Sons, Inc.; 1983.
3. Pankrats A. Forecasting with univariate Box-Jenkins models: concepts and cases. New York: John Wiley & Sons, Inc.; 1983.
4. Mentzer JT, Cox JE Jr. Familiarity, application and performance of sales forecasting techniques. J Forecast 1984;3:27–36.
5. Sparks R. Weighted moving averages: an efficient plan for monitoring specific location shifts. Int J Prod Res 2004;42(12):2521–2528.
6. Nugent T, Baykal-Gursoy M, Gursoy K. Monitoring k-step-ahead controlled processes. Qual Reliab Eng Int 2005;21:63–80.
7. Montgomery DC. Introduction to statistical quality control. New York: John Wiley & Sons, Inc.; 2005.
8. Bowerman BL, O'Connell RT, Koehler AB. Forecasting, time series, and regression. Belmont, CA: Duxbury; 2005.
9. Makridakis S, Wheelwright SC, Hyndman RJ. Forecasting: methods and applications. New York: John Wiley & Sons, Inc.; 1998.

¹⁰Control charts are often used to distinguish between common-cause and special-cause process variability. Common-cause variation is deemed “unavoidable” and special-cause variation “avoidable.”

¹¹Note that the control chart required $77 - 50 = 27$ additional samples after the change in order to detect the simulated decrease.

THEORY OF MARTINGALES

YINGDONG LU
Business Analytics and
Mathematical Sciences, IBM
T.J. Watson Research Center,
Yorktown Heights,
New York

INTRODUCTION

Suppose that we have a probability space $(\Omega, \mathcal{F}, P)$, a filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$ is defined as a family of σ -algebras $\mathcal{F}_n \subset \mathcal{F}$ satisfying $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for any $n = 1, 2, \dots$. For many practical problems, the filtration models the amount of information available to the decision makers.

Definition 1 A sequence of random variables X_n is a martingale with respect to the filtration $\{\mathcal{F}_n\}$, if it satisfies,

1. $X_n \in \mathcal{F}_n$;
2. $E|X_n| < \infty$;
3. $E[X_{n+1}|\mathcal{F}_n] = X_n$.

In the following, the filtration in each martingale is usually quite obvious; therefore, we omit mentioning the filtration unless it is necessary. While the focus of the authors' discussion is discrete-time processes in this article, most of the results have their continuous counterparts, and we have a brief discussion in one of the following sections. More detailed discussion on the basic concepts of both discrete-time and **Continuous-Time Martingales** can be found in the two articles of **Discrete-Time Martingales** and **Continuous-Time Martingales**.

Martingales, as an important class of stochastic processes, play an important role in modern probability theory. This important role is clearly manifested in such developments as the analysis of random walk, the martingale problem for Markov

processes, and the martingale methods in the construction of stochastic integration and understanding of diffusion processes. Meanwhile, being a powerful tool characterizing correlation, martingale methods and techniques are widely applied in mathematics, science, and engineering. For many problems in operations research and management science, martingales serve as an essential modeling and analytic tool for quantitatively characterizing statistical dependence that often rises when uncertainty and risk are involved. A typical example can be found in the martingale model of forecast evolution, a popular forecast update scheme, proposed in Refs 1 and 2. Finally, for such fields as mathematical finance, actuarial science, and stochastic optimal control, martingale is a necessary part of their foundations.

Like several other important probabilistic concepts, see, for example, Ref. 3, martingale finds its root in gambling. Intuitively, it characterizes the following fact, in a fair game, the knowledge of the past events will not enable the players to predict the future outcomes. This makes martingale a perfect building block in modeling decision making, which is the focus of applications in operations research and management science. Many applications in these areas often require deep understanding of the martingale model and analysis. In this article, following the two articles **Discrete-Time Martingales** and **Continuous-Time Martingales**, we discuss some advanced concepts and analysis in the theory of martingales.

The following concrete examples demonstrate that martingales are tied to some basic objects in probability and statistics, which are instrumental in modeling and analysis in operations research and management science.

Example 1 Summation of independent random variables can be studied as a martingale. More specifically, if $Y_1, Y_2, \dots$

are independent random variables satisfying $E[Y_i] = 0$, and $\text{Var}[Y_i] = \sigma_i^2$, it is easy to verify that $X_n = Y_1 + Y_2 + \dots + Y_n$ is a martingale. Meanwhile, we have, $E[(\sum_{i=1}^{n+1} Y_i)^2 | Y_1, Y_2, \dots, Y_n] = (\sum_{i=1}^n Y_i)^2 + \sigma_{n+1}^2$, therefore, $Z_n = (\sum_{i=1}^n Y_i)^2 - \sum_{i=1}^n \sigma_i^2$ is also a martingale.

Example 2 Let $Y_1, Y_2, \dots$ be independent and identically distributed random variables with finite moment generating function, define $\phi(\lambda) = E[\exp(\lambda Y_1)]$, for some $\lambda > 0$, then,

$$X_n = \phi(\lambda)^{-n} \exp\left(\lambda \sum_{i=1}^n Y_i\right)$$

is a martingale, and it is also known as the *Wald's martingale*. To see this, we have,

$$\begin{aligned} E[X_{n+1} | Y_1, Y_2, \dots, Y_n] &= \phi(\lambda)^{-(n+1)} E\left[\exp\left(\lambda \sum_{i=1}^{n+1} Y_i\right) \middle| Y_1, Y_2, \dots, Y_n\right] \\ &= \phi(\lambda)^{-(n+1)} \exp\left(\lambda \sum_{i=1}^n Y_i\right) E[\exp(\lambda Y_{n+1})] \\ &= \phi(\lambda)^{-n} \exp\left(\lambda \sum_{i=1}^n Y_i\right). \end{aligned}$$

Example 3 Suppose that $f(\cdot)$ is the density function of Y_i and $g(\cdot)$ is another probability density function, then,

$$X_n = \frac{g(Y_1)g(Y_2) \dots g(Y_n)}{f(Y_1)f(Y_2) \dots f(Y_n)}$$

is a martingale, which has its root in the likelihood ratio test. To verify this, we see,

$$\begin{aligned} E[X_{n+1} | Y_1, Y_2, \dots, Y_n] &= E\left[\frac{g(Y_1)g(Y_2) \dots g(Y_n)g(Y_{n+1})}{f(Y_1)f(Y_2) \dots f(Y_n)f(Y_{n+1})} \middle| Y_1, Y_2, \dots, Y_n\right] \\ &= \frac{g(Y_1)g(Y_2) \dots g(Y_n)}{f(Y_1)f(Y_2) \dots f(Y_n)} E\left[\frac{g(Y_{n+1})}{f(Y_{n+1})}\right] \\ &= X_n \int \frac{g(z)}{f(z)} f(z) dz = X_n. \end{aligned}$$

Example 4 Martingale is also a powerful tool to study Markov chains. For example, suppose that Y_n is a Markov chain governed by transition probability matrix $P = (P_{ij})$, then for any function $f: \mathbb{Z} \rightarrow \mathbb{R}$ that satisfies $af(i) = \sum_j P_{ij}f(j)$ for some $a > 0$, it is well known that $X_n = a^{-n}f(Y_n)$ is a martingale, as

$$\begin{aligned} E[a^{-(n+1)}f(Y_{n+1}) | Y_1, \dots, Y_n] &= E[a^{-(n+1)}f(Y_{n+1}) | Y_n] = a^{-(n+1)} \sum_j P_{Y_n j} f(j) \\ &= a^{-n}f(Y_n) = X_n. \end{aligned}$$

The condition, $af(i) = \sum_j P_{ij}f(j)$, is certainly satisfied by the eigenvector of the operator P .

Example 5 For any random variable X with $E[|X|] < \infty$, and $\mathcal{F}_n$ a filtration defined on the same probability space, define $X_n = E[X | \mathcal{F}_n]$, all the three conditions in the definition are satisfied, X_n is a martingale. It is known as the *Doob's martingale*, or the *Levy's martingale*, and a central model in stochastic approximation methods.

In many applications, condition 3 in the definition can only be satisfied in inequality form. This leads to the following pair of concepts that are closely related to martingale, submartingale, and supermartingale. When the equality in condition 3 is relaxed to $\geq$ ($\leq$) the sequence is called *submartingale* (*supermartingale*). There is a symmetry between the submartingale and supermartingale, so from now on, we just present the results for submartingale. Unless we specifically point out, the reader should know that there is a counterpart for supermartingale. The fundamental relationship between the martingale and the submartingale is best characterized in the following decomposition theorem.

Theorem 1 (*Doob's decomposition theorem*) For any submartingale Y_n , there is always a martingale X_n , and a predictable and increasing process A_n , such that $Y_n = X_n + A_n$, for any $n = 1, 2, \dots$

In the following sections, we review some basic concepts and results for martingales. In the section titled “Martingale Inequalities,” we discuss several martingale inequalities. In the section titled “Martingale Limit Theorem,” we review basic martingale convergence theorem, including law of large number and central limit theorem in general form. In the section titled “Optional Stopping Theorem,” the important topic of optional stopping is reviewed. In the section titled “Martingale with Continuous Parameters and Brownian Motion,” we discuss martingales with continuous parameter, especially Brownian motion. In the section titled “Historical Notes and Further Readings,” we review some historical development and provide some further readings.

MARTINGALE INEQUALITIES

Four types of inequalities are discussed in this section: (i) (Doob’s) maximal inequality, (ii) Burkholder’s inequality, (iii) decoupling inequalities, and (iv) concentration inequalities.

Maximal Inequalities

Doob’s maximal inequality indicates that the expected maximum of a submartingale can be controlled by the expected submartingale itself. We like to point out that maximal inequalities are extensions of the Kolmogorov’s inequality for independent and identically distributed random variables, see, for example, Ref. 4. They can be proved in the same manner. The following lemma is a key in establishing the maximal inequality, and we include the proof here for two reasons: first, we want to point out its root in the Kolmogorov’s inequality. Second, this is a typical argument in establishing many results in martingale theory, namely, identifying where the event happens and analyzing the event locally.

Lemma 1 *For a submartingale X_n , for any constant $C > 0$, we have,*

$$\mathbb{CP} \left[\sup_{1 \leq k \leq n} X_k \geq C \right] \leq \mathbb{E} \left[X_n \mathbf{1} \left\{ \sup_{1 \leq k \leq n} X_k > C \right\} \right].$$

Proof. The event under study is $\{\sup_{1 \leq k \leq n} X_k > C\}$. The key argument is to identify when the supremum is achieved. For this reason, define $A_i = \{X_i > C, \max_{j < i} X_j \leq C\}$. It is easy to see that A_i and A_j have no overlap, for any $i \neq j$, and their union recovers the event $\{\sup_{1 \leq k \leq n} X_k > C\}$. In other words, the event $\{\sup_{1 \leq k \leq n} X_k > C\}$ can be decomposed to the union of disjoint sets A_i . Hence, we have the following identity for its probability,

$$\mathbb{CP} \left[\sup_{1 \leq k \leq n} X_k \geq C \right] = C \sum_{i=1}^n \mathbb{P}[A_i].$$

Now, our problem has been reduced to estimate the probability of the smaller sets A_i , $i = 1, 2, \dots, n$.

From the way A_i is defined, we immediately have,

$$\mathbb{CP}[A_i] \leq \mathbb{E}[X_i \mathbf{1}(A_i)].$$

The definition of submartingale then implies,

$$\mathbb{E}[X_i \mathbf{1}(A_i)] \leq \mathbb{E}[\mathbb{E}[X_n | \mathcal{F}_i] \mathbf{1}(A_i)].$$

Meanwhile, as A_i is measurable with respect to $\mathcal{F}_i$, we have,

$$\mathbb{E}[\mathbb{E}[X_n | \mathcal{F}_i] \mathbf{1}(A_i)] = \mathbb{E}[\mathbb{E}[X_n \mathbf{1}(A_i) | \mathcal{F}_i]].$$

The right-hand side is, of course, equal to $\mathbb{E}[X_n \mathbf{1}(A_i)]$. The desired inequality thus follows when all the terms are summed up. $\square$

The following theorem is obtained from Lemma 1 through integration.

Theorem 2 *(Doob’s inequality) For a martingale, X_n , and any $p > 1$, we have,*

$$\|X_n\|_p \leq \left\| \sup_{1 \leq k \leq n} X_k \right\|_p \leq \frac{p}{p-1} \|X_n\|_p,$$

where $\mathbb{E}\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$.

The maximum inequality and its variations provide a vehicle to estimate the magnitude of the maximum of the martingale by its terminal value, hence, play important

roles in the analysis of random combinatorial algorithms and network scheduling, for some recent applications in these areas, see, for example, Refs 5 and 6.

Burkholder's Inequality

For a martingale X_n , the process $D_n = X_n - X_{n-1}$ is known as the *martingale difference sequence*, and $\sum_{i=1}^n D_i^2$ is known as the *quadratic variation process*, which is an important indicator of the volatility of the martingale X_n . Burkholder's inequality, sometimes is also called the *square function inequality*, establishes a fundamental relationship between the growth of martingale and that of the quadratic variation.

Theorem 3 *For any $p > 1$, there exist positive constants c and C such that the following inequality holds for martingale $X_n = \sum_{i=1}^n D_i$,*

$$\begin{aligned} c \left(\mathbb{E} \left[\sum_{i=1}^n D_i^2 \right] \right)^{p/2} \\ \leq \mathbb{E} [|X_n|^p] \leq C \left(\mathbb{E} \left[\sum_{i=1}^n D_i^2 \right] \right)^{p/2}. \end{aligned}$$

This fundamental relationship between the martingale and the square function of the martingale difference sequence is also crucial for the development of many important theories in mathematical analysis, which, in turn, has important implications on the study of random walk, a versatile model for many operations research applications, for detailed discussion on Burkholder's inequality and its applications, see, for example, Hall and Heyde [7].

Decoupling Inequalities

Understandably, independence among random variables provides convenience in probabilistic calculations. The purpose of decoupling inequalities is to use the functionals of the independent random variables to estimate the functionals of dependent random variables, such as martingales. Here

is a decoupling inequalities for martingales in a very general form.

Theorem 4 *Suppose that $X_n = \sum_{i=1}^n D_n$ and $(X_n, \mathcal{F}_n)$ is a martingale for some filtration $\mathcal{F}_n$. $\{E_i\}$ is an independent sequence of random variables such that, for each $i = 1, 2, \dots, E_i$ has the same probability law as that of D_i conditioning on $\mathcal{F}_{i-1}$. Then, for any convex function $\Phi(\cdot)$, there exist two constants c_α and C_α , such that*

$$\begin{aligned} c_\alpha \mathbb{E} \left[\Phi \left(\sup_n \left| \sum_{i=1}^n E_i \right| \right) \right] \\ \leq \mathbb{E} \left[\Phi \left(\sup_n \left| \sum_{i=1}^n D_i \right| \right) \right] \\ \leq C_\alpha \mathbb{E} \left[\Phi \left(\sup_n \left| \sum_{i=1}^n E_i \right| \right) \right], \end{aligned}$$

where $\|X\|$ denotes a general norm.

This type of decoupling inequalities are applied in the analysis of multi-item inventory systems to provide justification for reducing its analysis to that of the independent single-item inventory systems, see, for example, Ref. 8. A more comprehensive treatment of general decoupling inequalities and their application can be found in the book by de la Peña and Giné [9].

Concentration Inequalities

The concentration inequalities, closely related to the convergence theorems that are presented in the next section, describe quantitatively the phenomenon that a summation of random variables will not deviate from its expectation with a large probability. Here is a version that is accredited to Azuma and Hoeffding.

Theorem 5 (Azuma–Hoeffding). *Let $\{D_n\}$ be a martingale difference sequence, and $c_1, c_2, \dots, c_n$ are constants such that $|D_i| \leq c_i$ almost surely. Then for any $t > 0$, we have,*

$$\mathbb{P} \left[\max_{1 \leq k \leq n} \sum_{i=1}^k D_i > t \right] \leq \exp \left(-\frac{t^2}{2 \sum_{i=1}^k c_i} \right).$$

Concentration inequalities, essentially the Azuma's inequality, are instrumental in the analysis of random algorithms, see, for example, Ref. 10.

MARTINGALE LIMIT THEOREM

In this section, we first introduce Doob's martingale convergence theorems, then the law of large number and central limit theorems for martingale difference. A very important concept that is a key to some convergence theorems and optimal stopping is introduced and discussed in the next section.

Theorem 6 *If X_n is a submartingale, and there exist a $C > 0$, such that $E|X_n| \leq C$ for any $n \geq 1$, then there exists a random variable X , satisfies $E|X| < \infty$, such that $X_n \rightarrow X$ almost surely as $n \rightarrow \infty$.*

Definition 2 A family of random variables $X_n, n \geq 1$, is called *uniformly integrable* if

$$\lim_{x \rightarrow \infty} \int_{|X_n| > x} |X_n| dP = 0,$$

uniformly in $n \geq 1$.

In essence, the property of uniform integrability enables "truncation" operation, which is a very important technique in probability and analysis. In fact, we have,

Theorem 7 *The following three propositions are equivalent for a submartingale X_i :*

1. *It is uniformly integrable.*
2. *It converges in L^1 .*
3. *It converges almost surely to an integrable random variable X_∞ , with $E[X_n] \rightarrow E[X_\infty]$.*

A martingale can be treated as the summation of the martingale difference sequence; hence, it is natural to ask about its first- and second-order asymptotic behavior. In other words, do we have the law of large number and the central limit theorem?

The best result on the law of large number is obtained by Elton [11]. It indicates that, in comparison to the Marcinkiewicz–Zygmund theorem for independent and identically distributed random variables, where integrability of the random variables guarantees the almost sure convergence, a more restrictive integrability condition is required for establishing the law of large number for a martingale difference sequence. Define,

$$L \log L = \{X \in L^1 | E[|X| \log(|X| \vee 1)] < \infty\}.$$

We have,

Theorem 8 *If $D_n \in L \log L$, then $\frac{1}{n}X_n \rightarrow 0$ almost surely.*

In Elton's article, a counterexample is given for the case such that the law of large number fails to hold for a sequence that is only L^1 but not $L \log L$. The proof techniques displayed in Ref. 11 are also marvelous.

Meanwhile, requirements on second moment are needed for central limit theorem. The following version might not be in the most general form, but it provides a convenient tool for most of the problems encountered in the application.

Theorem 9 *Let S_{ni} be a zero-mean square integrable martingale for each n , and $X_{ni} = S_{ni} - S_{n,i-1}$, and η^2 is an almost surely finite random variable. Suppose that*

$$\max_i |X_{ni}| \rightarrow 0, \quad \text{in probability,}$$

$$\sum_i X_{ni}^2 \rightarrow \eta^2, \quad \text{in probability,}$$

$$E\left(\max_i X_{ni}^2\right) \text{ is bounded in } n,$$

then, $S_{nk_n} = \sum_i X_{ni} \rightarrow Z$ in distribution, where the random variable Z has the characteristic function $E \exp(-\frac{1}{2}\eta^2 t^2)$.

For more general results and deep discussions, please see Ref. 7, a wonderful resource of various limit theorems for martingales.

OPTIONAL STOPPING THEOREM

Recall that the expectation of the martingale equals to its initial value at any time. The optional stopping theorem (also known as *optional sampling theorem*) extends this property to allow the time to be random. More precisely, it allows the time to be a stopping time. This technique, as we can see from examples in this section, greatly enhances our ability to compute many probabilistic quantities.

Definition 3 A random variable τ taking value in nonnegative integers is called a stopping time (with respect to the filtration $\{\mathcal{F}_t\}$) if $\{\tau \leq n\} \in \mathcal{F}_n$ for any $n \geq 1$.

Intuitively, stopping time can be understood as a random time that depends on the history of the stochastic process. A concrete example, which may also partially explain the name, can be found in context of gambling or game, where the time to stop can depend on the outcome of the game or gambling. For example, a gambling stops when a player runs out of money, meanwhile, a tennis game stops when a player earns enough points.

An important stopping time widely studied in operations research and management science is the first passage time. For any stochastic process X_n , and a real number a , then T_a , the first passage time for X_n at level a is defined as,

$$T_a \triangleq \{n \geq 1 : X_n \geq a\}.$$

In other word, the first time that the process X_n takes a value above the level a .

One of the most important results on martingale is the following optional stopping theorem.

Theorem 10 For a martingale X_n and a stopping time τ , if the following conditions are satisfied,

- $P[\tau < \infty] = 1$,
- $E[|X_\tau|] < \infty$,
- $\lim_n \rightarrow \infty E[X_n \mathbf{1}\{\tau > n\}] = 0$.

Then, $E[X_\tau] = E[X_0]$.

There are also various optional stopping theorems, for details, see, for example, Ref. 12.

Example 6 One example will be the simple random walk, suppose that Y_i are independent and identically distributed random variables with $Y_i = 1$ or $Y_i = -1$ with probability $1/2$ respectively. Then $X_n = Y_1 + Y_2 + \dots + Y_n$ is a martingale, define, for any pair of positive integers a and b ,

$$T_{ab} = \min\{n \geq 1 : X_n = -a, \text{ or } X_n = b\}.$$

Define p_a the probability of X_n reach a first, then, optimal stopping yields,

$$E[X_{T_{ab}}] = (-a)p_a + b(1 - p_a) = 0.$$

Hence, $p_a = \frac{b}{a+b}$. Next, define $Z_n = X_n^2 - n$. It is easy to see that Z_n is a martingale, too. Apply optional stopping to Z_n , we can get that $E[T_{ab}] = ab$.

Example 7 (Walds' identities). X_i are independent and identically distributed random variables, with $E[X_1] = \mu < \infty$, $E[X_1^2] < \infty$, and T is a stopping time, then we have,

$$E[S_T] = \mu E[T], E[S_T^2] = E[T]E[X_1^2].$$

MARTINGALE WITH CONTINUOUS PARAMETERS AND BROWNIAN MOTION

The concept of martingale can be extended naturally to continuous-time processes.

Definition 4 A continuous-time stochastic process X_t is a martingale with respect to the filtration $\{\mathcal{F}_t\}$, if it satisfies,

1. $X_t \in \mathcal{F}_t$,
2. $E[|X_t|] < \infty$,
3. $E[X_t | \mathcal{F}_s] = X_s$, for any $s \leq t$.

Inequalities and convergence theorems can be adapted to the continuous version in a straightforward manner.

One of the most important **Continuous-Time Martingales** is Brownian motion. Here, we reflect on some basic properties of the Brownian motion, as well as some important applications of the optional stopping rule.

Definition 5 Brownian motion, $B(t)$, is defined as a Gaussian process with mean zero, and correlation $E[B_s B_t] = \min(s, t)$ for any s and t .

Almost surely, the sample path of Brownian motion is continuous, but nowhere differentiable. It is also an important example of diffusion processes, that is, Markov processes with continuous sample path almost surely. Diffusion processes play a vital role in the development of financial mathematics and stochastic control.

At any time $t \geq 0$, $B(t)$ follows normal distribution with mean zero and variance t , hence with density function,

$$\phi_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}, \quad x \in \mathbb{R}.$$

One of the most important results for Brownian motion is the distribution of the running maximum of the Brownian motion. Denote $M(t) = \max_{0 \leq s \leq t} B(s)$, we have,

$$\begin{aligned} P[B(t) \in dx, M(t) \in dy] &= \frac{2(2y-x)}{\sqrt{2\pi t^3}} \exp\left\{-\frac{(2y-x)^2}{2t}\right\} dx dy, \\ P[M(t) \in dx] &= \frac{2}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right) dx. \end{aligned}$$

HISTORICAL NOTES AND FURTHER READINGS

The term *martingale* was first introduced to mathematics by Jean Ville [13] in 1939. However, mathematicians Sergey Berntein and Paul Levy had investigated dependent

random variables that have martingale structures before that. There are various explanations for the origin of word “martingale” and its meanings in the game of gambling, a good resource for these stories is a recent article by Roger Mansuy [14].

Joseph Leo Doob [15] derived some fundamental results on martingales in the 1940s and 1950s. Many of these results are presented elegantly in his 1953 classic *Stochastic Processes*, and a more comprehensive treatment of martingales is presented in his later less popular book *Classical Potential Theory and Its Probabilistic Counterpart* [22]. Doob is considered by many as the “founding father of the theory of martingales” [14].

Donald Burkholder developed some fundamental results on martingales in the 1960s and 1970s. Especially, his work on martingale inequalities has deep impact in the development of harmonic analysis, some of his ideas are still actively explored by researchers today. His work on martingale transform laid foundation for the development of stochastic calculus.

The mature and fast growing of financial mathematics has brought a large new influx of ideas and problems to martingale theory and application. The key object in the development of this area is Brownian motion, a martingale with continuous parameter. While observed and studied by physicists from the nineteenth century, the first mathematical construction of Brownian motion was only obtained by Norbert Wiener [16] in 1923. During the 1940s and 1950s, Kiyoshi Ito developed the stochastic calculus method to study general Markov processes. It becomes the basis of financial mathematics.

There are some excellent textbooks available for martingales. For readers with strong mathematical background, Chapter 9 in Kai-Lai Chung’s [17] classic textbook is always considered as a wonderful introduction to the topic. Less technical savvy books, such as [12] and [18] are good ones for general audiences. The book by Chow and Teicher [19] provides deep discussions on many aspects of martingale theory. Hall and Heyde [7] discusses in details on various limit theorems. Karatzas and Shreve [20] and Reeuz and Yor

[21] are two good sources for fundamentals on continuous-time martingale.

Certainly, there are so many very interesting and important topics that we are not able to cover. These include, but not limited to, martingale problem in Markov processes, martingale's role in functional central limit theorems, martingales in Levy processes, and martingale transform.

REFERENCES

1. Graves SC, Meal S, Dasu Y, *et al.* Two-stage production planning in a dynamic environment. In: Axsater S, Schneeweiss C, Silver E, editors. Multi-stage production planning and control. Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag; 1986. p 9–43.
2. Heath D, Jackson P. Modeling the evolution of demand forecasts with applications to safety stock analysis in production/distribution systems. IIE Trans 1994;26(3):17–30.
3. David FN. Games, gods & gambling: a history of probability and statistical ideas. London: Courier Dover Publications; 1962.
4. Feller W. An introduction to probability theory and its applications. Volume 1, New York: John Wiley and Sons; 1968.
5. Naor A, Tao T. Random martingales and localization of maximal inequalities. J Funct Anal 2010;259:731–779.
6. Shah D, Tsitsiklis J, Zhong Y. Qualitative properties of α -fair policies in bandwidth-sharing networks; 2011. *arXiv:1104.2340*.
7. Hall P, Heyde CC. Martingale limit theorems and its applications. New York: Academic Press; 1980.
8. Lu Y, Song J. Order-based cost structure and its optimization in an assemble-to-order system. Oper Res 2005;53(1):151–169.
9. de la Peña V, Giné E. Decoupling: from dependence to independence. New York: Springer-Verlag; 1999.
10. Alon N, Spencer J. The probabilistic method. New York: Wiley; 1992.
11. Elton J. A law of large numbers for identically distributed martingale differences. Ann Probab 1981;9(3):405–412.
12. Karlin S, Taylor HM. A first course in stochastic processes. 2nd ed. New York: Academic Press; 1975.
13. Ville J. Etude critique de la notion de collectif. Paris: Gauthier-Villars; 1939.
14. Mansuy R. The origins of the word “martingale”. Electron J History Probab Stat 2009;5(1):1–10.
15. Doob JL. Stochastic processes. New York: John Wiley & Sons; 1953.
16. Wiener N. Differential space. J Math Phys 1923;2:131–174.
17. Chung KL. A course in probability theory. 2nd ed. New York: Academic Press; 1974.
18. Ross S. Stochastic processes. 2nd ed. New York: John Wiley & Sons; 1995.
19. Chow YS, Teicher H. Probability theory, independence, interchangeability, martingale. New York: Springer; 1978.
20. Karatzas I, Shreve S. Brownian motion and stochastic calculus, graduate texts in mathematics. 2nd ed. New York: Springer-Verlag; 1991.
21. Reez D, Yor M. Continuous martingales and Brownian motion. Volume 293, Grundlehren der mathematischen Wissenschaften. New York: Springer-Verlag; 1994.
22. Doob JL. Classical potential theory and its probabilistic counterpart. Berlin, Heidelberg, New York: Springer-Verlag; 1984.

TOTAL EXPECTED DISCOUNTED REWARD MDPs: EXISTENCE OF OPTIMAL POLICIES

EUGENE A. FEINBERG
Department of Applied
Mathematics and Statistics,
State University of New
York at Stony Brook, Stony
Brook, New York

INTRODUCTION

Deterministic optimal policies always exist for discounted dynamic programming problems with finite action sets. Such policies also exist when action sets satisfy certain compactness conditions, and transition probabilities and reward functions satisfy certain continuity conditions. If either compactness or continuity conditions do not hold, deterministic ϵ -optimal policies exist for problems with countable state spaces. For problems with uncountable Borel-state spaces, the results similar to the existence of deterministic ϵ -optimal policies hold but in this more general case, either the notion of ϵ -optimality should be replaced with the weaker notion of (p, ϵ) -optimality or a broader definition of a policy is required. Since the theory is simpler when the state space is countable, problems with countable and uncountable state spaces are considered separately in this article.

COUNTABLE STATE SPACE

Definitions

Consider a Markov decision process (MDP) with the state space X , action space A , sets of actions $A(x)$ available at states $x \in X$, transition probabilities p , and rewards r . We assume throughout this section that the state space X is countable. In particular, it can be either finite or countably infinite.

We assume that the action set A is a complete, separable metric space. For simplicity, A can be imagined as a finite set, countably infinite set, R^n , or its natural subset. All sets $A(x)$ are nonempty Borel subsets of A . In particular, if A is countable then $A(x)$ are arbitrary nonempty subsets of A .

If an action $a \in A(x)$ is selected at the state x , then the one-step reward is $r(x, a)$ and the probability that the system will be at the state y at the next step is $p(y|x, a)$. The standard assumption is that the functions $r(x, a)$ and $p(y|x, a)$ are measurable in a for all $x, y \in X$.

Unless specified, we shall assume that the sum of transition probabilities is 1 and the reward function is bounded above. The former means that $\sum_{y \in X} p(y|x, a) = 1$ for all $x \in X$ and for all $a \in A(x)$. The latter means that there exists a finite constant K such that $r(x, a) \leq K$ for all $x \in X$ and for all $a \in A(x)$.

For the classical dynamic programming problems introduced by Blackwell [1], the reward functions $r(x, a)$ are assumed to be bounded, that is, $|r(x, a)| \leq K$, $x \in X$ and $a \in A(x)$, for some finite constant K . However in many operations research applications, the reward functions are bounded above, that is, $r(x, a) \leq K$ when $x \in X$ and $a \in A(x)$. For example, in mathematical models of inventory and queueing systems, the one-step holding costs can tend to ∞ as the inventory levels or number of waiting customers increases to ∞ . Therefore, the corresponding reward functions can be unlimited from below. So, we consider the more general case when the reward function is bounded from above.

A general policy may be randomized and history-dependent. Let Π be the set of all policies. A deterministic policy is defined as a function ϕ that always selects action $\phi(x)$ at a state $x \in X$. In other words, a deterministic policy is history-independent and nonrandomized. Let $\mathcal{D}$ denote the set of deterministic policies.

Let the initial state be x and a policy π be chosen. For a positive constant $\alpha < 1$ called a *discount factor*, the expected total discounted

reward is

$$v^\pi(x) = E_x^\pi \sum_{n=0}^{\infty} \alpha^n r(x_n, a_n),$$

where E_x^π is the expectation when the initial state is x and a policy π is chosen, and x_t and a_t are states and selected actions at epochs $t = 0, 1, \dots$. The value at each state x is defined as $V(x) = \sup_{\pi \in \Pi} v^\pi(x)$.

A policy π is called optimal if $v^\pi(x) = V(x)$ for all $x \in X$. Thus, the optimality is defined with respect to all possible initial states. If an optimal policy is deterministic, it is called a *deterministic optimal policy*.

A policy π is called ϵ -optimal for a nonnegative constant ϵ if $v^\pi(x) \geq V(x) - \epsilon$ for all initial states x from X . In particular, the notions of 0-optimal and optimal policies coincide.

For $x \in X$, $a \in A(x)$, and for a bounded above real-valued function f on X , define

$$T^a f(x) = r(x, a) + \alpha \sum_{y \in X} p(y|x, a) f(y).$$

This value can be interpreted as the expected reward over two steps starting from the state x , when the action a from $A(x)$ is selected and the reward at the second step is defined by the function f . For a deterministic policy ϕ , we can consider the operator T^ϕ defined for functions f bounded above,

$$T^\phi f(x) = T^{\phi(x)} f(x), \quad x \in X.$$

Then

$$v^\phi(x) = T^\phi v^\phi(x), \quad x \in X. \quad (1)$$

The optimality operator T is defined as

$$Tf(x) = \sup_{a \in A(x)} T^a f(x), \quad x \in X.$$

The value function V satisfies the *optimality equation*

$$U(x) = TU(x), \quad x \in X. \quad (2)$$

In particular, if the reward function r is bounded, then T^ϕ and T are contraction mappings defined on the set of bounded functions

on X endowed with the metric d induced by supremum norms. This is true for an arbitrary, possibly uncountable, set X . The supremum norm $\|v\|$ of a bounded function v on X is $\|v\| = \sup_{x \in X} |v(x)|$. The metric d is defined as $d(v, u) = \|u - v\| = \sup_{x \in X} |v(x) - u(x)|$. The contraction mapping property in our case means that $d(T^\phi v, T^\phi u) \leq \alpha d(v, u)$ and $d(Tv, Tu) \leq \alpha d(v, u)$.

Therefore, when the reward function r is bounded, the Banach fixed point theorem (also known as the *contraction mapping theorem* or *contraction mapping principle*) implies that the functions v^ϕ and V are the unique bounded solutions of the equations $u = T^\phi u$ and $U = TU$, respectively. The Banach fixed point theorem also provides the convergence of the value iteration algorithm and it provides estimates for its convergence; see Bertsekas [2, Chapter 1] on such estimates and Feinberg [3] on convergence of value iterations for total-reward criteria. We notice that in addition to a unique bounded solution, each of these equations may have unbounded solutions (Example 6.4 in Ref. 3).

The analysis of discounted problems with reward functions bounded above can be reduced to the analysis of a negative dynamic programming problem by replacing the reward function r with the nonpositive reward function $\tilde{r} = r - K$. Then the expected total-reward $\tilde{v}^\pi$ for the new problem is $\tilde{v}^\pi(x) = v^\pi(x) - K/(1 - \alpha)$ for all $x \in X$ and for all $\pi \in \Pi$. Therefore, the existence of optimal and ϵ -optimal policies for discounted problems can be obtained from the corresponding results for negative programming. For example, for negative programming, the value function is the largest solution of Equation (2) with $\alpha \in [0, 1]$. Therefore, if the reward function r is bounded above for a discounted dynamic programming problem, then the value function V is the largest solution of Equation (2) satisfying the inequality $U(x) \leq K/(1 - \alpha)$ for all $x \in X$. However, stronger results sometimes hold for discounted problems rather than for negative problems. The optimality equation and its properties are summarized in the following theorem.

Theorem 1. *The value function V is the largest solution of the optimality equation (Eq. 2), satisfying the inequality $U(x) \leq K/(1 - \alpha)$ for all $x \in X$. If the reward function r is bounded, then V is the unique bounded solution of the optimality equation (Eq. 2).*

In a similar way, discounted MDPs with reward functions bounded below can be reduced to positive dynamic programming. Such reward functions are used in some economics applications.

Existence of Optimal Policies

A deterministic policy ϕ is called *conserving* if $T^\phi V(x) = V(x)$ for all $x \in X$.

Theorem 2. *A deterministic policy ϕ is optimal if and only if it is conserving.*

Therefore, the existence of a deterministic optimal policy and the existence of a conserving policy are equivalent statements. Finding a deterministic optimal policy is equivalent to finding a conserving policy. Consider the sets of conserving actions

$$A_c(x) = \{a \in A(x) \mid T^a V(x) = V(x)\}, \quad x \in X.$$

Theorem 2 implies that a deterministic optimal policy exists, if and only if $A_c(x) \neq \emptyset$. In addition, the optimality equation implies that $v^\pi(x) < V(x)$ for any policy π if $A_c(x) = \emptyset$. Therefore, Theorem 2 implies the following corollary.

Corollary 1. *An optimal policy exists if and only if for each $x \in X$ the set of conserving actions $A_c(x)$ is not empty. In addition if an optimal policy exists, then there exists a deterministic optimal policy.*

Thus, it is sufficient to verify $A_c(x) \neq \emptyset$ for all $x \in X$ to prove the existence of optimal deterministic policies. In particular, since $V = TV$, optimal policies exist if all the sets $A(x)$ are finite. However, they also exist under more general assumptions.

Definition 1. A real-valued function f defined on a topological space Z is called

sup-compact if the set $Z^\lambda = \{z \in Z \mid f(z) \geq \lambda\}$ is compact for any real number λ .

The following assumption is sufficient for $A_c(x) \neq \emptyset$ for all $x \in X$.

Assumption 1.

- (i) For each $x, y \in X$, the function $p(y \mid x, a)$ is continuous in $a \in A(x)$.
- (ii) For each $x \in X$, the function $r(x, a)$ is sup-compact in $a \in A(x)$. This means that for any $x \in X$ and for any number λ , the set $A^\lambda(x) = \{a \in A(x) \mid r(x, a) \geq \lambda\}$ is compact.

Assumption 1 implies that all the sets of conserving actions $A_c(x)$ are nonempty and therefore the following statement holds.

Corollary 2. *If an MDP satisfies Assumption 1, then there exists a deterministic optimal policy.*

The following assumption is stronger than Assumption 1.

Assumption 2.

- (i) Assumption 1(i) holds.
- (ii) The set $A(x)$ is compact for each $x \in X$.
- (iii) For each $x \in X$ the reward function $r(x, a)$ is upper semi-continuous in $a \in A(x)$.

Corollary 3. *If an MDP satisfies Assumption 2, then there exists a deterministic optimal policy.*

In particular, Assumption 2 holds when all the sets $A(x)$ are finite. Therefore, the above mentioned fact on the existence of optimal policies for finite action sets can be formulated as a particular case of Corollary 3.

Corollary 4. *If for each state $x \in X$ the set of available actions $A(x)$ is finite, then there exists a deterministic optimal policy.*

According to the above statements, the existence of optimal policies requires certain continuity and compactness conditions to ensure the existence of conserving actions for all states. The existence of deterministic ϵ -optimal policies, where ϵ is an arbitrary fixed positive number, does not require such conditions.

Let ϵ be a positive constant. Consider a deterministic policy ϕ such that $T^\phi(x)V(x) \geq TV(x) - (1 - \alpha)\epsilon$ for all $x \in X$. Then the optimality equation $V = TV$ and straightforward calculations imply that $v^\pi(x) \geq V(x) - \epsilon$ for all $x \in X$. Therefore, the following result holds.

Theorem 3. *For any $\epsilon > 0$ there exists a deterministic ϵ -optimal policy.*

BOREL-STATE PROBLEMS

A topological space $(E, \mathcal{T})$ is called *Polish* if it is separable and completely metrizable. We recall that a topological space E is called *completely metrizable*, if it admits a compatible metric d such that (X, d) is a complete metric space. A Borel σ -field on a topological space $(E, \mathcal{T})$ is defined as the minimal σ -field on E containing all the open subsets from E . A subset of B is called *Borel* if it belongs to the Borel σ -field on B .

A measurable space $(B, \mathcal{B})$ is called *Borel* (or *standard Borel*), if there is a one-to-one measurable mapping $f : B \rightarrow [0, 1]$ such that $f(B)$ is a Borel subset of the interval $[0, 1]$. Any Polish space B is a Borel space with $\mathcal{B}$ being the Borel σ -field on B . In addition, any Borel space and, therefore, any Polish space is either countable or continuum (Appendix 1 of Ref. 4 or Corollary 7.16.1 of Ref. 5). Countable sets include finite sets. If a Borel space B is continuum, there is a measurable one-to-one mapping of B on the interval $[0, 1]$. If two Borel spaces B_1 and B_2 have the same cardinality, there is a measurable one-to-one mapping of B_1 on B_2 such that the mapping f^{-1} is measurable too.

Any countable or continuum set is Polish in the appropriate topology. If a set is countable, it is Polish if the discrete topology on it is chosen. If a set B is continuum, we can choose any one-to-one correspondence

$f : B \rightarrow [0, 1]$ of B on the interval $[0, 1]$, and define the topology on B , whose basis consists of the proimages $f^{-1}(C)$ of all open intervals $C = (y, z) \subset [0, 1]$. In particular, if B is a Borel space, then f can be selected as measurable. Therefore, any Borel space is a Polish space in the described topology.

Let $(B, \mathcal{B})$ be a Borel space and $\mathcal{P}(B)$ be the set of all probability measures on $(B, \mathcal{B})$. For any probability measure p on $(B, \mathcal{B})$, consider the p -completion $\mathcal{B}_p$ of $\mathcal{B}$, that is, $\mathcal{B}_p$ is the minimal σ -field containing $\mathcal{B}$ and any subset Y of B such that $Y \subset C$ for some $C \in \mathcal{B}$ with $p(C) = 0$. The σ -field $\mathcal{U} = \bigcap_{p \in \mathcal{P}(B)} \mathcal{B}_p$ is called the *universal σ -field*. A subset C of B is called *universally measurable* if $C \in \mathcal{U}$. A function $f : B \rightarrow Y$, where B and Y are Borel spaces, is said to be *universally measurable*, if $f^{-1}(C)$ is universally measurable for any Borel set $C \subseteq Y$. If f is a universally measurable function and p is a probability measure on $(B, \mathcal{B})$, then there exists a Borel function f_p such that $P(\{z : f(z) \neq f_p(z)\}) = 0$. Thus, universally measurable functions $f : B \rightarrow R$ can be integrated with respect any probability measure on $(B, \mathcal{B})$ in the same way as Borel-measurable functions.

A Borel-state MDP is defined by the same objects as the MDP with a countable state space. The differences are that (i) the state space X is a Borel space, (ii) the transition probability $p(E|x, a)$ is a probability measure on the Borel σ -field of X and it satisfies the condition that for any Borel subset Y of X the function $p(Y|x, a)$ is measurable in $(x, a) \in X \times A$.

A general policy is defined by transition probabilities π_n , $n = 0, 1, \dots$ such that for any $h_n = x_0, a_0, x_1, a_1, \dots, a_n$, $\pi_n(da_n|h_n)$ is a probability measure on A satisfying the following two conditions: (i) $\pi_n(A(x_n)|h_n) = 1$ and (ii) for each Borel subset B of A , the function $\pi_n(B|h_n)$ is Borel-measurable on the set of all histories H_n up to time n , $H_n = X \times (A \times X)^n$. Thus, a general policy can be randomized and history-dependent.

The following assumption is standard for Borel-state MDPs and is always assumed in this article: the graph

$$\text{Gr}(A) = \{(x, a) | x \in X, a \in A(x)\}$$

is a Borel subset of $X \times A$. A deterministic policy is a measurable mapping of X to A such that $\phi(x) \in A(x)$ for all $x \in X$. In other words, $(x, \phi(x)) \in \text{Gr}(A)$ for all $x \in X$. In general, for some $D \subset X \times A$, a mapping $f: X \rightarrow A$ is called a *selector* if $(x, f(x)) \in D$ for all $x \in X$. So, a deterministic policy is a measurable selector from X to A with respect to $D = \text{Gr}(A)$. So, a deterministic policy is sometimes called a *measurable selector*.

In order to define at least one policy and avoid the possibility that $\Pi = \emptyset$, we need to assume that there exists at least one deterministic policy. We shall avoid this assumption by using the convention that $\sup\{\emptyset\} = -\infty$. Then, $V(x) = -\infty$ for all $x \in X$, if $\Pi = \emptyset$. In many cases, for example, under Assumptions S, Su, W, and Wu and for universally measurable policies described below, $\Pi \neq \emptyset$, because the existence of a deterministic policy follows from so-called *measurable selection theorems*.

A special feature of the above model is that the value function V may not be Borel-measurable. However, it belongs to a broader class of universally measurable functions, for which integration is possible. In the same way, we had for the countable state set, the value function V satisfies the optimality equation. In the case of a bounded reward function r , the value function is a unique bounded universally measurable function. This is true if either the existence of at least one deterministic policy is assumed or policies are allowed to be selected from a broader class of universally measurable policies.

For Borel-state models, the statements similar to Theorems 1 and 2 hold.

Theorem 4.

- (i) *The value function V is universally measurable and it is a solution of the optimality equation (Eq. 2).*
- (ii) *The value function V is the largest universally measurable solution U of the optimal equation (Eq. 2), such that $U(x) \leq K/(1 - \alpha)$ for all $x \in X$.*
- (iii) *If the reward function r is bounded, then V is the unique bounded solution of the optimality equation (Eq. 2).*

Theorem 5. *A deterministic policy is optimal if and only if it is conserving.*

In addition, the following statement holds.

Theorem 6. *If there exists an optimal policy, then there exists a deterministic optimal policy.*

To ensure the existence of conserving policies, some continuity and compactness properties are required. Before we describe particular cases, we formulate a general statement.

For a policy π , let v_N^π denote the N -horizon expected discounted total rewards. We have $v_0^\pi(x) = 0$, $x \in X$, and for $N = 1, 2, \dots$

$$v_N^\pi(x) = E_x^\pi \sum_{n=0}^{N-1} \alpha^n r(x_n, a_n), \quad x \in X.$$

Let $V_N(x) = \sup_{\pi \in \Pi} v_N^\pi(x)$. According to these definitions, $V_0(x) = 0$ for all $x \in X$. In the trivial case, when $\Pi = \emptyset$ we have $V_N(x) = -\infty$, for all $x \in X$ and for all $N = 1, 2, \dots$. The values $V_N(x)$ is the supremum of the expected total-reward over the finite-horizon N with the 0 terminal value V_0 . As was shown by Blackwell [1], the functions V_N , $N = 1, 2, \dots$ and N belong to the class of universally measurable functions and therefore they can be integrated in the same way as Borel functions. In addition, the function V_N satisfies the optimality equation $V_{N+1} = TV_N$, $N = 0, 1, \dots$. This is the well-known optimality equation for finite-horizon dynamic programming models. It can be proved by inductions or as a corollary from Theorem 4(i), because a finite-horizon model can be reduced to an infinite-horizon model [3, Section 4.6]. The following theorem provides a sufficient condition for the existence of optimal deterministic policies.

Theorem 7. *Suppose that there exists a nonnegative integer k such that for any $x \in X$, for any finite number λ , and for any $N \geq k$ the set*

$$A_N^\lambda(x) = \left\{ a \in A(x) \mid r(x, a) + \alpha \int V_N(y) p(dy \mid x, a) \geq \lambda \right\}$$

is a compact subset of A . Then $V(x) = \lim_{N \rightarrow \infty} V_N(x)$ for all $x \in X$, and there exists a deterministic optimal policy.

Theorem 7 is a useful tool to establish the existence of deterministic optimal policies under an assumption similar to Assumption 5. When X is continuum, there are two groups of assumptions in the literature corresponding to Assumption 5. These assumptions correspond to two different types of convergence of probability measures: setwise convergence and weak convergence. Such assumptions, the so-called S and W were introduced by Schäl [6,7] for the case of compact action spaces. Since we start with possibly noncompact action sets, we shall use abbreviations Su and Wu, where the symbol “u” indicate that the sets $A(x)$ can be unbounded and therefore noncompact. In many applications, noncompact action sets are unbounded.

Assumption Su (i) For each $x \in X$, the transition probability $p(dy|x, a)$ is setwise continuous in $a \in A(x)$. This means that $p(Y|x, a_n) \rightarrow p(Y|x, a)$ for every $x \in X$ and for any sequence $a_n, n = 1, 2, \dots$, of elements of $A(x)$ converging to a , where Y is an arbitrary measurable subset of X .

(ii) For each $x \in X$, the reward function $r(x, a)$ is sup-compact in a , that is, the set $A^\lambda(x) = \{a \in A(x) | r(x, a) \geq \lambda\}$ is compact for any real number λ .

Assumption Wu.

- (i) The transition probability $p(dy|x, a)$ is weakly continuous in $(x, a) \in X \times A$, that is, for any bounded continuous function f on X

$$\int f(y) p(dy|x_i, a_i) \rightarrow \int f(y) p(dy|x, a),$$

for any $x \in X$ and any $a \in A(x)$, if $(x_i, a_i) \rightarrow (x, a)$ and $a_i \in A(x_i)$.

- (ii) The reward function $r(x, a)$ is sup-compact on $\text{Gr}(A)$, that is, the set $\{(x, a) \in \text{Gr}(A) | r(x, a) \geq \lambda\}$ is compact for any finite number λ .

Theorem 8. *Either Assumption Su or Assumption Wu implies the existence of a deterministic optimal policy. In addition, the value function V is Borel measurable under Assumption Su and sup-compact under Assumption Wu.*

The following assumptions are sufficient for the existence of deterministic optimal policies when all the sets of available actions $A(x)$ are compact.

Assumption S.

- (i) The sets $A(x)$ are compact for all $x \in X$.
- (ii) The transition probability $p(\cdot|x, a)$ is setwise continuous in $a \in A(x)$, that is, Assumption Su(i) holds.
- (iii) The reward function $r(x, a)$ is upper semi-continuous in $a \in A(x)$ for all $x \in X$.

Assumption W.

- (i) The sets $A(x)$ are compact for all $x \in X$.
- (ii) The set-valued mapping $A(x)$ is upper semicontinuous; that is, for any open subset G of A , the set $\{x \in X | A(x) \subseteq G\}$ is open in X .
- (iii) The transition probability $p(\cdot|x, a)$ is weakly continuous on $(X \times A)$, that is, Assumption Wu(i) holds.
- (iv) The reward function $r(x, a)$ is upper semicontinuous on $\text{Gr}(A)$.

Theorem 9. *Under each Assumption S or W, there exists a deterministic optimal policy. In addition, the value function V is Borel-measurable under Assumption S and upper semicontinuous under Assumption W.*

The following corollary follows from Theorem 9 under Assumption S.

Corollary 5. *If each set $A(x), x \in X$, is finite, then there exists a deterministic optimal policy.*

Now we discuss the existence of ϵ -optimal policies. For expected discounted reward

MDPs with countable state spaces, deterministic ϵ -optimal policies always exist (Theorem 3). For a Borel-state space, this is true if all the sets of available actions $A(x)$ are countable. In this case, there always exists a deterministic policy, $V(x)$ is a Borel-measurable function, and the following theorem holds.

Theorem 10. *If all the action sets $A(x)$, $x \in X$ are countable, then for any $\epsilon > 0$ there exists a deterministic ϵ -optimal policy.*

In a general situation, the value function $V(x)$ may not be Borel measurable, but it is universally measurable, see Refs 1, 4, and 5 for detail. Since for any fixed policy π the function $v^\pi(x)$ is Borel, ϵ -optimal policies may not exist [1, Example 2]. However, such policies exist if either the definition of ϵ -optimal is changed to the definition of so-called (p, ϵ) -optimal policies or the set of policies is expanded by allowing the policies to be universally measurable.

Let p be a probability measure on X . A policy π is called (p, ϵ) -optimal if there exists a measurable subset Y of X such that $p(Y) = 1$ and $v^\pi(x) \geq V(x)$ for all $x \in Y$.

Theorem 11. *For any probability measure p on X and for any $\epsilon > 0$, there exists a deterministic (p, ϵ) -optimal policy.*

However, for any $\epsilon > 0$ there exists a deterministic ϵ -optimal policy, if the definition of a policy is expanded by allowing transition probabilities $\pi_n(B|h_n)$ to be universally measurable functions on X for all Borel-measurable subsets B of A . A deterministic universally measurable policy ϕ is a universally measurable mapping from X to A such that $\phi(x) \in A(x)$ for all $x \in X$. The Jankov–von Neumann selection theorem implies that there exists at least one universally measurable deterministic policy, details in Bertsekas and Shreve [5], Dynkin and Yushkevich [4], and Kechris [8]. We also remark that the value function V remains unchanged after the set of Π is extended by allowing universally measurable policies, except the trivial case when there is no Borel-measurable selector from X to A .

Theorem 12. *If universally measurable policies are allowed then for any $\epsilon > 0$ there exists a deterministic universally measurable ϵ -optimal policy.*

BIBLIOGRAPHIC NOTES

The literature on discounted MDPs is vast and we mention only a few references. Shapley [9] studied stochastic games with discounted payoffs. Dubins and Savage [10] and then Hordijk [11] studied the relations between conserving and optimal policies. In particular, a deterministic policy for an MDP with the expected total criterion is optimal, if and only if it is conserving and equalizing. However, any policy is equalizing for a discounted MDP with a reward function $r(x, a)$ bounded above. So, Theorem 2 can be traced to Dubins and Savage [10]. Theorem 5 is a straightforward extension of Theorem 2 to problems with Borel-state spaces.

Blackwell [1] studied a discounted Borel-state MDP with bounded rewards. Theorems 6, 10, and 11 are from Ref. 1. Strauch [12] proved the universal optimality of the value function and optimality equation (Theorem 4). Blackwell [1] and Strauch [12] considered a model when $A(x) = A$ for all $x \in X$. Dynkin and Yushkevich [10] extended the theory in several directions including the state-dependent action sets $A(x)$. Denardo [13] studied contraction properties of dynamic programming operators. In particular, these properties lead to Theorem 3. This theorem can be viewed as a particular case of Theorem 6.21 from Feinberg [3], which describes the structure of nearly optimal policies in countable MDPs with expected total rewards [14].

Theorem 7 is Proposition 9.17 from Bertsekas and Shreve [5]. Schäl [6,7] introduced Assumptions S and W, where S stands for setwise continuity of transition probabilities and W stands for weak continuity. Theorem 9 is from Schäl [7]. The monograph by Hernández-Lerma and Lasserre [15] primarily covers MDPs with setwise continuous transition probabilities. Assumption Su is from Hernández-Lerma [16] and Assumption Wu is from Feinberg and Lewis [17].

Theorem 8 presents the results from Refs 16 and 17. The answer to the question on which type of continuity is more appropriate depends on a particular application. For example, in Ref. 17, it was observed that Assumption Wu is applicable to inventory control problems with continuous demand distributions, while Assumption Su is applicable to inventory control problems with general demand distributions.

Blackwell *et al.* [18] and Freedman [19] investigated discounted MDPs with analytically measurable policies. Bertsekas and Shreve [5] developed the theory of dynamic programming on Borel-state spaces and universally measurable policies. Theorem 12 is from Ref. 5.

The theory of Borel spaces and measurable selection theorems play an important role in studying dynamic programming problems with Borel-state spaces. In addition to excellent chapters and appendices in Bertsekas and Shreve [5], Dynkin and Yushkevich [4], and Hernández-Lerma and Lasserre [15], the monograph by Kechris [8] contains important facts on these topics.

Acknowledgment

This author acknowledges support by NSF grant DMI-0600538. The author thanks Peng Dai, Jun Fei, Mark E. Lewis, and Xiaoxuan Zhang for reading a preliminary version of this article and providing their comments.

REFERENCES

1. Blackwell D. Discounted dynamic programming. *Ann Math Stat* 1965;36:226–235.
2. Bertsekas DP. Volume 2, Dynamic programming and optimal control. 2nd ed. Belmont (CA): Athena Scientific; 2001.
3. Feinberg EA. Total reward criteria. In: Feinberg EA, Shwartz A, editors. *Handbook of Markov decision processes*. Boston (MA): Kluwer; 2002. pp. 173–207.
4. Dynkin EB, Yushkevich AA. *Controlled Markov processes*. New York: Springer; 1979. (translated from the Russian 1975 edition).
5. Bertsekas DP, Shreve SE. *Stochastic optimal control: the discrete-time case*. Belmont (MA): Athena Scientific; 1996. (reprinted from the 1978 Academic Press edition).
6. Schal M. On dynamic programming: compactness of the space of policies. *Stoch Proces Appl* 1975;3:345–364.
7. Schal M. Conditions for optimality in dynamic programming and for the limit of n -stage optimal policies to be optimal. *Z Wahrsch verve Gebiete* 1975;32:179–196.
8. Kechris AS. *Classical descriptive set theory*. New York: Springer; 1994.
9. Shapley LS. Stochastic games. *Proc Natl Acad Sci U S A* 1953;39:1095–1100.
10. Dubins L, Savage L. *Inequalities for stochastic processes (how to gamble if you must)*. New York: Dover; 1976. (reprinted from the 1965 McGraw-Hill edition).
11. Hordijk A. Volume 51, Dynamic programming and Markov potential theory: Mathematical centre tracts. 2nd ed. Amsterdam: Mathematisch Centrum; 1977.
12. Strauch R. Negative dynamic programming. *Ann Math Stat* 1966;37:871–880.
13. Denardo EV. Contraction mappings in the theory of underlying dynamic programming. *SIAM Rev* 1967;9:165–177.
14. Feinberg EA. Sufficient classes of strategies in discrete dynamic programming II: locally stationary strategies. *SIAM Theory Prob Appl* 1987;32:478–493.
15. Hernandez-Lerma O, Lasserre JB. *Discrete-time Markov control processes. basic optimality criteria*. New York: Springer; 1996.
16. Hernandez-Lerma O. Average optimality in dynamic programming on Borel spaces. *Syst Control Lett* 1991;17:237–242.
17. Feinberg EA, Lewis ME. Optimality inequalities for average cost Markov decision processes and the stochastic cash balance problem. *Math Oper Res* 2007;32:769–783.
18. Blackwell D, Freedman D, Orkin M. The optimal reward operator in dynamic programming. *Ann Probab* 1974;2:926–941.
19. Freedman D. The optimal reward operator in special classes of in dynamic programming problems. *Ann Probab* 1974;2:942–949.

TOTAL EXPECTED DISCOUNTED REWARD MDPs: POLICY ITERATION ALGORITHM

Z. GOZDE ICTEN
SERDAR KARADEMIR
LISA M. MAILLART
Department of Industrial
Engineering, University of
Pittsburgh, Pittsburgh,
Pennsylvania

Policy iteration (PI), or search in policy space, was originally developed by Bellman [1] and independently by Howard [2] to solve Markov decision processes (MDPs). For an introduction to MDPs, please see the articles titled **MDP Basic Components** and **Total Expected Discounted Reward MDPs: Existence of Optimal Policies** in this encyclopedia. In this article, we restrict our attention to stationary infinite horizon MDPs under the total expected discounted reward criterion.

The PI algorithm requires one of the following sets of conditions [3]:

1. finite-state and action sets, and we assume that this assumption holds throughout the article;
2. finite-state space, compact action sets, and continuous reward and transition probabilities;
3. countable state space, compact action sets, and bounded continuous reward and transition functions.

The main idea of PI is to begin with an initial arbitrary policy and iteratively obtain improved policies until an optimal one is found, which is guaranteed to occur in a finite number of iterations under a specific set of conditions. The rest of the article is

organized as follows. We first give details of the PI algorithm and discuss its convergence properties. After demonstrating PI on a numerical example, we present the modified policy iteration (MPI) algorithm. Finally we illustrate MPI algorithm with an example and conclude the article.

POLICY ITERATION ALGORITHM

Define S as the state space of the process, A as the set of all possible actions, and D as the set of policies on $S \times A$. Let R be the set of bounded rewards defined on $S \times A$, and V be the space of bounded value vectors. Then, $A_s \subseteq A$ is the set of possible actions in state $s \in S$ and $r(s, a) \in R$ is the immediate expected reward received if action $a \in A$ is taken in state $s \in S$. The transition probability matrix defined under action $a \in A$ is denoted as $P(a)$ and the entry $p(s'|s, a)$ of $P(a)$ represents the transition probability to state $s' \in S$ from state $s \in S$ if action $a \in A_s$ is taken. Furthermore, we define λ to be the one-period discount factor, $0 < \lambda < 1$.

Let $v(s)$ be the total expected discounted reward starting in state $s \in S$ and following the optimal policy thereafter. The Bellman optimality equations are given by

$$v(s) = \max_{a \in A_s} \left\{ r(s, a) + \lambda \sum_{j \in S} p(j|s, a) v(j) \right\}, \quad \forall s \in S. \quad (1)$$

In vector notation, Equation (1) can be written as

$$v = \max_{d \in D} \{ r_d + \lambda P_d v \}, \quad (2)$$

where P_d is the transition probability matrix under policy d . We now give the details of PI algorithm.

Policy Iteration Algorithm [3]

1. Set $n = 0$ and select an arbitrary policy $d_0 \in D$.
2. (Policy Evaluation) Obtain v^n by solving the system of equations given by

$$(I - \lambda P_{d_n}) v^n = r_{d_n}, \quad (3)$$

where I denotes identity matrix of dimension $|S| \times |S|$.

3. (Policy Improvement) Choose d_{n+1} to satisfy

$$d_{n+1} \in \operatorname{argmax}_{d \in D} \{r_d + \lambda P_d v^n\} \quad (4)$$

setting $d_{n+1} := d_n$ if possible.

4. If $d_{n+1} = d_n$, return $d^* := d_n$. Otherwise set $n := n + 1$ and go back to step 2.

Each pass through the PI algorithm starts by calculating the “value vector,” v^n , for the current policy, d_n (Step 2). Remember that a policy is a mapping that assigns a single action to each state in the state space. Each element of the value vector is the corresponding total expected discounted reward for an initial state the process occupies under policy d_n . Under finite state and action sets, obtaining the value vector corresponds to solving a linear system of equations of size $|S|$. This step usually requires implementation of Gaussian elimination to solve the system. However, note that as the cardinality of the decision space, $S \times A$, increases, the computational burden of this step gets prohibitive, which is the main drawback of the policy iteration algorithm. MPI addresses this shortcoming.

After evaluating the current policy, the algorithm tries to improve the current policy (Step 3). If an improvement is found, the algorithm performs another iteration. Otherwise the algorithm stops and reports the current policy as the optimal one.

CONVERGENCE OF THE PI ALGORITHM

The policy iteration algorithm yields a sequence of policies and value functions. For processes with finite-state and action sets, this sequence is finite and the algorithm terminates in a finite number of iterations.

Finite convergence occurs with certainty, since the set of all possible policies is finite and therefore an improving policy is generated at each iteration. This convergence property of the policy iteration algorithm is its main advantage over the value iteration algorithm, which, in general, converges in an infinite number of iterations.

On the other hand, for a process with a compact action set and/or a countable state space, for which the number of policies is infinite, the stopping criterion in Step 4 may never be satisfied. In this case, a different approach must be used to demonstrate convergence of the algorithm. For further details, please refer to Puterman [3], Puterman and Brumelle [4], Puterman and Hordijk [5], and Golubin [6].

The following proposition from Puterman [3] establishes the convergence of the PI algorithm.

Proposition 1. *Let v^n and v^{n+1} be two successive value vectors generated by the PI algorithm and the operator $\succeq$ stand for componentwise comparison of two vectors. Then,*

$$v^{n+1} \succeq v^n.$$

Proof. If d_{n+1} satisfies Equation (4), then, by Equation (3)

$$r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n \succeq r_{d_n} + \lambda P_{d_n} v^n = v^n.$$

Rearranging the terms yields

$$r_{d_{n+1}} \succeq (I - \lambda P_{d_{n+1}})v^n.$$

Hence,

$$\begin{aligned} (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} &\succeq v^n \\ v^{n+1} &\succeq v^n, \end{aligned} \quad (5)$$

where inequality (5) follows from Lemma 6.1.2(b) in Puterman [3].

Proposition 1 establishes the monotonicity of the sequence $\{v^n\}$ which is key to the convergence of the PI algorithm.

Step 1. Set $n = 0$ and select the initial policy as $d_0(1) = 0$ and $d_0(2) = 1$.

Step 2. Solving of the following linear system of equations

$$\begin{aligned} v^0(1) &= -2 + 0.9 \cdot (0.75 \cdot v^0(1) + 0.25 \cdot v^0(2)) \\ v^0(2) &= -3 + 0.9 \cdot (0.25 \cdot v^0(1) + 0.75 \cdot v^0(2)) \end{aligned}$$

yields $v^0(1) = -24.12$ and $v^0(2) = -25.96$.

Step 3. We now seek an improving policy, $d_1(1)$ and $d_1(2)$, as follows

$$\max \left\{ \begin{array}{l} -2 + 0.9 \cdot (0.75 \cdot (-24.12) + 0.25 \cdot (-25.96)), \\ -0.5 + 0.9 \cdot (0.25 \cdot (-24.12) + 0.75 \cdot (-25.96)) \end{array} \right\} = \max\{-24.12, -23.45\}.$$

Hence $d_1(1) = 1$. Similarly,

$$\max \left\{ \begin{array}{l} -1 + 0.9 \cdot (0.75 \cdot (-24.12) + 0.25 \cdot (-25.96)), \\ -3 + 0.9 \cdot (0.25 \cdot (-24.12) + 0.75 \cdot (-25.96)) \end{array} \right\} = \max\{-23.12, -25.95\}$$

and $d_1(2) = 0$.

Step 4. Since $d_0(1) \neq d_1(1)$ and $d_0(2) \neq d_1(2)$, set $n = 1$ and return to the policy evaluation step.

Step 2. The solution of the linear system of equations for $d_1(1) = 1$ and $d_1(2) = 0$ yields $v^1(1) = -7.33$ and $v^1(2) = -7.67$.

Step 3. The new policy d_2 is found as follows

$$\max \left\{ \begin{array}{l} -2 + 0.9 \cdot (0.75 \cdot (-7.33) + 0.25 \cdot (-7.67)), \\ -0.5 + 0.9 \cdot (0.25 \cdot (-7.33) + 0.75 \cdot (-7.67)) \end{array} \right\} = \max\{-8.67, -7.33\}$$

and $d_2(1) = 1$. Similarly,

$$\max \left\{ \begin{array}{l} -1 + 0.9 \cdot (0.75 \cdot (-7.33) + 0.25 \cdot (-7.67)), \\ -3 + 0.9 \cdot (0.25 \cdot (-7.33) + 0.75 \cdot (-7.67)) \end{array} \right\} = \max\{-7.67, -9.83\}$$

PI: A NUMERICAL EXAMPLE

We illustrate the implementation of the PI algorithm with a small example. Consider a problem with 2 states and 2 actions, $S = \{1, 2\}$ and $A = \{0, 1\}$. Let the transition probabilities and the rewards be $p(1|1, 0) = 0.75$, $p(1|1, 1) = 0.25$, $p(2|1, 0) = 0.75$, $p(2|1, 1) = 0.25$, $r(1, 0) = -2$, $r(1, 1) = -0.5$, $r(2, 0) = -1$, and $r(2, 1) = -3$. We set $\lambda = 0.9$ and apply the policy iteration algorithm to this problem instance to find the optimal policy.

and $d_2(2) = 0$.

Step 4. Since $d_1(1) = d_2(1)$ and $d_1(2) = d_2(2)$, the algorithm terminates. The resulting optimal policy is $d^* = [1, 0]$ with value vector $v^* = [-7.33, -7.67]$.

MODIFIED PI ALGORITHM

As mentioned in the first section, for large state spaces the exact solution of the system of equations given by Equation (3) can be computationally prohibitive. The key idea behind the MPI algorithm is that it is not necessary to determine the value vector for the current policy precisely in order to identify an improving policy. Therefore, during policy evaluation, the MPI algorithm obtains an approximate solution to Equation (3) by using successive approximations with a fixed policy. That is, MPI combines

features of both policy iteration and value iteration. The reader can refer to the article titled ***Total Expected Discounted Reward MDPs: Value Iteration Algorithm*** in this encyclopedia for details. Like policy iteration, it contains an improvement and an evaluation step, but the evaluation step is iterative like the value iteration algorithm. It converges faster than value iteration and requires at least as many iterations as policy iteration. However, the computational effort per iteration exceeds that for value iteration and is less than that for policy iteration.

Modified PI Algorithm [3]

Let $\{m_n\}$ be a sequence of nonnegative integers.

1. Select $v_0 \in V$, $\epsilon > 0$, and set $n = 0$.
2. (Policy Improvement) Choose d_{n+1} to satisfy

$$d_{n+1} \in \operatorname{argmax}_{d \in D} \{r_d + \lambda P_d v^n\}$$

setting $d_{n+1} := d_n$ if possible (if $n > 0$).

3. (Partial Policy Evaluation) Obtain v^{n+1}
 - (a) Set $k = 0$ and

$$u_0^n = \max_{d \in D} \{r_d + \lambda P_d v^n\}.$$

- (b) If $\|u_0^n - v^n\| < \epsilon(1 - \lambda)/2\lambda$, go to Step 4. Otherwise go to (c).
 - (c) If $k = m_n$, go to (e). Otherwise compute u_{k+1}^n as follows

$$u_{k+1}^n = r_{d_{n+1}} + \lambda P_{d_{n+1}} u_k^n.$$

- (d) Set $k := k + 1$ and return to (c).
 - (e) Set $v^{n+1} := u_{m_n}^n$, $n := n + 1$, and go back to Step 2.
4. Return $d_\epsilon := d_{n+1}$.

The partial policy evaluation (Step 3) performs the substeps (c) and (d) for m_n iterations, where m_n may be

1. fixed for all iterations, that is, $m_n = m$ for all n ;
2. chosen according to some prespecified pattern, for example, m_n can be increasing in n ; or
3. selected adaptively, for example, $\|u^{m_n+1} - u^{m_n}\| < \epsilon_n$ where ϵ_n is fixed or variable [3].

If $m_n = 0$, then the MPI algorithm coincides with value iteration algorithm, and if $m_n = \infty$, then it coincides with the PI algorithm.

Since we partially evaluate policies in the MPI algorithm, the resulting policy is called ϵ -optimal. For a proof of convergence of the MPI algorithm and other variants of policy iteration (e.g., Gauss–Seidel, Jacobi, and/or overrelaxation) refer to Puterman [3].

MPI: A NUMERICAL EXAMPLE

For the example we solved using PI algorithm, now we consider the MPI algorithm using $m_n = 5 \forall n$, and $\epsilon = 0.01$. Setting $n = 0$, we select $v_0(1) = -25$ and $v_0(2) = -30$. Using v_0 , we implement policy improvement, obtain $d_1(1) = 0$ and $d_1(2) = 0$. Then we set $k = 0$, and implement partial policy evaluation (Step 3) for $m_n = 5 \forall n$. The result of each iteration is given in Table 1.

Then we improve the current policy (Step 2) and obtain $d_2(1) = 1$ and $d_2(2) = 0$. We continue in this manner and after 10 more iterations meet the stopping criteria resulting with no change in the last policy we had obtained. Table 2 tabulates u values obtained in the partial policy evaluation step for the last iteration. The last policy improvement step returns $d^*(1) = 1$ and $d^*(2) = 0$ with resulting $v^*(1) = -7.3290$ and $v^*(2) = -7.6739$.

SUMMARY AND FURTHER READING

Policy iteration is an algorithm used to find optimal policies for stationary, infinite horizon MDPs. Under certain conditions, it

Table 1. $u_k^0(s)$ Values for $s = 1, 2$.

$s \mid k$	0	1	2	3	4	5
1	-25.63	-24.84	-24.01	-23.26	-22.59	-22.01
2	-24.63	-23.31	-22.45	-21.75	-21.14	-20.61

Table 2. $u_k^{11}(s)$ Values for $s = 1, 2$.

$s \mid k$	0	1	2	3	4	5
1	-7.33	-7.33	-7.33	-7.33	-7.33	-7.33
2	-7.68	-7.68	-7.68	-7.67	-7.67	-7.67

can also be used under the undiscounted, expected total reward criterion, and the average reward criterion [3]. It proves to be effective in terms of its convergence properties, though it can entail some computational drawbacks with increasing problem size. However, these drawbacks can be alleviated by considering the MPI algorithm or another variant of the PI algorithm.

For a deeper understanding of the PI algorithm, refer to Bertsekas [7], Heyman and Sobel [8], or Ross [9]. Other applications of PI algorithm can be found in Meyn [10], which discusses the application of PI algorithm to MDPs with general state space, and in Hansen [11], which illustrates an improved PI algorithm for Partially Observable MDPs. Madani [12] and Tsitsiklis [13] discuss the complexity of PI algorithm and its convergence.

REFERENCES

1. Bellman RE. Dynamic programming. Princeton (NJ): Princeton University Press; 1957.
2. Howard RA. Dynamic programming and Markov processes. Cambridge (MA): The MIT Press; 1960.
3. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. Hoboken (NJ): Wiley; 1994.
4. Puterman ML, Brumelle SL. On the convergence of policy iteration in stationary dynamic programming. Math Oper Res 1979;4(1):60–69.
5. Puterman ML, Hordijk A. On the convergence of policy iteration in finite state undiscounted Markov decision processes: the unichain case. Math Oper Res 1987;12(1):163–176.

6. Golubin AY. A note on the convergence of policy iteration in Markov decision processes with compact action spaces. *Math Oper Res* 2003;28(1):194–200.
7. Bertsekas DP. *Dynamic programming and stochastic control*. Orlando (FL): Academic Press; 1976.
8. Heyman D, Sobel M, editors. *Handbook of operations research and management science, Volume 2: Stochastic Models*. North Holland Publishers; Amsterdam 1990. pp. 331–434.
9. Ross S. *Introduction to stochastic dynamic programming*. Orlando (FL): Academic Press; 1983.
10. Meyn SP. The policy iteration algorithm for average reward Markov decision processes with general state space. *IEEE Trans Automat Contr* 1997;42(12):1663–1680.
11. Hansen EA. An Improved Policy Iteration Algorithm for Partially Observable MDPs. Tenth Neural Information Processing Systems Conference; Denver (CO), 1997.
12. Madani O. On Policy Iteration as a Newton's Method and Polynomial Policy Iteration Algorithms. Eighteenth National Conference on Artificial Intelligence; American Association for Artificial Intelligence, Menlo Park (CA), 2002.
13. Tsitsiklis JN. On the convergence of optimistic policy iteration. *IEEE Trans Automat Contr* 2002;3:59–72.

TOTAL EXPECTED DISCOUNTED REWARD MDPs: VALUE ITERATION ALGORITHM

THEOLOGOS BOUNTOURELIS
Department of Industrial
Engineering, University of
Pittsburgh, Pittsburgh,
Pennsylvania

INTRODUCTION

The field of Markov decision processes (MDP) concerns the modeling and solving of dynamic decision making problems under uncertainty. The field has received considerable attention and researchers have developed the theoretical and computational tools for the application of MDPs in many scientific areas. These include the scheduling and controlling of manufacturing systems [1], finance [2], and management problems [3].

An MDP can be a *discrete-time*, a *continuous-time*, or a *semi-Markov* problem. In this article we will focus on discrete time MDPs. The basic structure of a typical discrete MDP implementation involves a *decision maker* evolving in a *discrete state space* through the execution of a series of *actions*. Starting at an initial state, at time period $t = 0$ and at every consecutive period $t = 1, 2, 3, \dots$, the decision maker observes the current *state* and chooses to execute some action that will probabilistically lead to a next state. Furthermore, the decision maker experiences a transition cost expressed by some *cost function* that assigns a numerical payoff, often random, to each executed action. The task of the decision maker is to select a sequence of actions in order to minimize some objective function of the costs collected over a *time* N .

A typical discrete-time MDP problem usually falls into one of the three main categories depending on the choice of the objective function and the time horizon. It can be a (i) finite horizon, *total cost problem*, where the objective is to minimize the cumulative

cost over some finite time horizon; (ii) an infinite horizon, *discounted cost problem*, where the objective is to minimize the sum of discounted cost over an infinite time horizon; or (iii) an *average reward problem*, where the objective is to minimize the average reward collected over an infinite time horizon.

This article is concerned with infinite horizon, discounted problems, where the cost collected at each period t is discounted by a factor γ^t , where $\gamma \in (0, 1)$ is a scalar referred to as the *discount factor*. The discounted problem is perhaps the most analytically tractable and widely studied MDP problem with its foundations set in the 1960s [4–6]. By considering the sum of discounted costs, and under some assumptions on the cost function, the discounted problem is usually well defined. By applying the discount factor at every consecutive decision epoch, the decisions made earlier have a greater effect on the cumulative cost than the decisions made later in time. This modeling framework has been proven useful in many different contexts, for example, the planning of an investment where the present value of the future payoffs is a discounted cumulative reward, or, the evaluation and control of a system that, at every time step, may fail with probability γ .

The article is structured as follows: we first provide the notation and we formally state the total cost, discounted problem. Subsequently, we characterize the problem optimal policies, and state the value iteration algorithm. In the sequel, we provide with a numerical example that illustrates the application of the value iteration algorithm on a problem instance. Finally, we conclude with some discussion on the underlying assumptions, the enhanced versions of the presented algorithm, and provide the relevant references.

A FORMAL DEFINITION OF THE TOTAL COST, DISCOUNTED, INFINITE HORIZON PROBLEM

A total cost discounted MDP $\mathcal{M}$, is defined by a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}, g, \gamma)$, where

- $\mathcal{X}$ is a finite set of states.
- $\mathcal{A}$ is a set function, defined on $\mathcal{X}$ that maps each $x \in \mathcal{X}$ to the finite, nonempty *action* set $\mathcal{A}(x)$, comprising all the actions that can be executed by the decision maker at state x .
- $\mathcal{P}$ is a *transition function*, defined on $\mathcal{X} \times \bigcup_{x \in \mathcal{X}} \mathcal{A}(x)$, that associates with every state action pair (x, a) in this set as discrete probability distribution $p(x, \cdot; a)$. The support sets of the distributions $p(x, \cdot; a)$ are subsets of the state set $\mathcal{X}$.
- $g : \mathcal{S} \times \bigcup_{x \in \mathcal{X}} \mathcal{A}(x) \rightarrow \mathbb{R}$ is a cost function, that maps each pair (x, a) , $a \in \mathcal{A}(x)$, to a cost $g(x, a)$.
- Finally, γ is a positive scalar referred to as the *discount factor*.

The decision maker starts from an initial state x_0 at period $t = 0$, and at every consecutive period $t = 1, 2, 3, \dots$ it (i) observes the current state x_t , (ii) selects and executes an action $a \in \mathcal{A}(x_t)$, and (iii) accumulates a cost of $\gamma^t \cdot g(x_t, a)$. We define a *policy* to be a sequence of functions $\mu_t : \mathcal{X} \rightarrow \mathcal{A}$, $t = 0, 1, 2, \dots$ with $\mu_t(x) \in \mathcal{A}(x)$. In words, a policy is a sequence of action selection rules that for every time step and state will return an action to be executed by the decision maker. We will call a policy *stationary* if $\mu_t = \mu$, $t = 0, 1, 2, \dots$, that is, if the action selection rule is independent of time. As is seen later, there is always an optimal stationary policy, and for the rest of this article we restrict ourselves to the class of stationary policies.

We denote by Π the set of all stationary policies. *The decision maker's objective is to determine a policy π that minimizes the cumulative cost*

$$V_\pi(x_0) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t \cdot g(x_t, \pi(x_t)) \right], \quad (1)$$

where the expectation $E_\pi[\cdot]$ is taken over all the possible realizations of the policy π . The vector V_π will be referred to as the *value function*, or the *cost-to-go* vector of the policy π . For every state $x \in \mathcal{X}$ we define

$$V^*(x) = \inf_{\pi \in \Pi} V_\pi(x). \quad (2)$$

The vector V^* will be referred to as the *optimal value function* or the *optimal cost-to-go* vector and expresses the total expected cost of initiating the process at some state $x \in \mathcal{X}$, and following the optimal policy thereafter. Furthermore, we define the *optimal policy* as

$$\pi^*(x) = \arg \inf_{\pi \in \Pi} V_\pi(x). \quad (3)$$

In the next section we state the conditions that should be satisfied by an optimal policy.

OPTIMALITY EQUATIONS

First, we introduce an operator that will play a pivotal role in the characterization of the optimal policies and the development of the value iteration algorithm. For any function $V : \mathcal{X} \rightarrow \mathbb{R}$, we define the operator $TV : \mathcal{X} \rightarrow \mathbb{R}$, as

$$(TV)(x) = \min_{a \in \mathcal{A}(x)} \left\{ g(x, a) + \gamma \sum_{y \in \mathcal{X}} p(x, y; a) \cdot V(y) \right\}, \quad x \in \mathcal{X}. \quad (4)$$

Consider the set of equations defined by $V(x) = TV(x)$, $x \in \mathcal{X}$, where the quantities $V(x)$, $x \in \mathcal{X}$ are the unknown variables. This set of equations will be described as

$$TV = V, \quad (5)$$

and referred to as the *optimality equations* or the *Bellman equations*. Then, under the assumption of (i) a countable state space $\mathcal{X}$, and (ii) a bounded cost function $g(x, a)$, it can be shown [7] that

- the optimality equations have a unique solution;
- the optimal value function, V^* satisfies the optimality equations, that is,

$$V^*(x) = \min_{a \in \mathcal{A}(x)} \left\{ g(x, a) + \gamma \sum_{y \in \mathcal{X}} p(x, y; a) \cdot V^*(y) \right\}, \quad x \in \mathcal{X}, \quad (6)$$

and

- from the optimality equations we can derive stationary optimal policies. That is, a stationary policy π is optimal if and only if it attains the minimum in Bellman's equation, that is,

$$V^*(x) = g(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(x, y; \pi(x)) \cdot V^*(y), \quad x \in \mathcal{X}. \quad (7)$$

After we established the existence and the characterization of optimal policies, we proceed to the development of an algorithmic tool to numerically obtain an approximation of the optimal value function and policy. In the next section we present the most known computational method for discounted problems, the value iteration algorithm.

THE VALUE ITERATION ALGORITHM

The value iteration algorithm is the most celebrated and easily implementable algorithm for solving discounted MDPs. Before we state the algorithm, we should first provide some intuition behind it. Let the compositions of the operator T be T^k , that is,

$$T^k V(x) = T^{k-1}(TV), \quad k = 1, 2, 3, \dots \quad (8)$$

The composition operator has an interesting interpretation. Consider the problem where the decision maker must select its actions over a horizon of k steps, and then experience a terminal cost given by the vector $\gamma^k V$. It should be intuitively clear that as $k \rightarrow \infty$, the effects of the terminal costs are diminished at a geometric rate of γ^k , and the calculated costs $T^k V_0$ should converge to the optimal value function, V^* . Indeed, it can be proved [8] that for any initial value vector V_0 , we have that

$$\lim_{k \rightarrow \infty} T^k V_0 = V^*. \quad (9)$$

The convergence of $T^k V_0$ to the optimal value function provides the basis of the value iteration algorithm that will be described, in its most basic form, in the sequel.

For any cost-to-go vector V , we define the norm $\|V\| = \sup_{x \in \mathcal{X}} |V(x)|$. Then, for any $\epsilon > 0$, we define an ϵ -optimal policy to be a policy with an expected performance that is ϵ close to the optimal, that is,

$$\|V_{\pi_\epsilon} - V^*\| < \epsilon. \quad (10)$$

Then, the following algorithm is called the *value iteration algorithm* [7] and returns an ϵ -optimal policy.

Value Iteration Algorithm

- Step 0: Select a vector V_0 , and choose an $\epsilon > 0$.
- Step 1: Set $n = 0$.
- Step 2: $V_{n+1} = TV_n$.
- Step 3: If $\|V_{n+1} - V_n\| \leq \epsilon(1 - \gamma)/2\gamma$ go to Step 4, else go Step 2.
- Step 4: Set $\pi_\epsilon(x) = \arg \min_{a \in \mathcal{A}(x)} \{g(x, a) + \gamma \sum_{y \in \mathcal{X}} p(x, y; a) \cdot V_{n+1}(y)\}$ and terminate.

Assuming the existence of an optimal solution, the value iteration algorithm can find an ϵ -optimal policy π_ϵ within a finite number of iterations. Furthermore, for the cost-to-go estimates V_{n+1} , we have

$$\|V_{n+1} - V^*\| < \epsilon/2. \quad (11)$$

Hence, by choosing an ϵ small enough, the algorithm terminates with a policy close to the optimal.

AN EXAMPLE

In this section, we report a computational study on a two-state discounted MDP instance that demonstrates the application of the value iteration algorithm presented in this article.

We consider a discounted MDP instance defined by the state space $\mathcal{X} = \{x_1, x_2\}$, an action set $\mathcal{A}(x_1) = \{a_{11}, a_{12}\}$, $\mathcal{A}(x_2) = \{a_{21}, a_{22}\}$ and transition probabilities $p(x_1, x_1; a_{11}) = 0.4$, $p(x_1, x_2; a_{11}) = 0.6$, $p(x_1, x_1; a_{12}) = 0.5$, $p(x_1, x_2; a_{12}) = 0.5$, $p(x_2, x_1; a_{21}) = 0.7$, $p(x_2, x_2; a_{21}) = 0.3$, $p(x_2, x_1; a_{22}) = 0.2$, and $p(x_2, x_2; a_{22}) = 0.8$. Furthermore, we consider the cost function $g(x_1, a_{11}) = 1.2$, $g(x_1, a_{12}) = 2.1$,

$g(x_2, a_{21}) = 1.5$, $g(x_2, a_{22}) = 1.1$, and a discount factor $\gamma = 0.9$.

The optimality equations are given by

$$\begin{aligned} V^*(x_1) &= \min_{a \in \{a_{11}, a_{12}\}} \left\{ g(x_1, a) + 0.9 \right. \\ &\quad \cdot \left. \sum_{y \in \{x_1, x_2\}} p(x_1, y; a) V^*(y) \right\} \\ V^*(x_2) &= \min_{a \in \{a_{21}, a_{22}\}} \left\{ g(x_2, a) + 0.9 \right. \\ &\quad \cdot \left. \sum_{y \in \{x_1, x_2\}} p(x_2, y; a) V^*(y) \right\}, \end{aligned}$$

and the optimal solution is given by $V^*(x_1) = 11.3415$ and $V^*(x_2) = 11.2195$.

We choose $\epsilon = 0.01$ and initialize $V_0(x_1) = V_0(x_2) = 0$. We expect the value iteration algorithm to return an approximate value vector, which is 0.005 close to the optimal. Indeed, the value iteration algorithm terminates after 72 iterations returning the value vector $V_{72}(x_1) = 11.3368$ and $V_{72}(x_2) = 11.2148$, for which $\|V_{72} - V^*\| = 0.0047$. The implied policy is given by $\pi_\epsilon(x_1) = a_{11}$ and $\pi_\epsilon(x_2) = a_{22}$.

DISCUSSION

The two sufficient conditions for the theoretical convergence of the value iteration algorithm are (i) a countable state space, and, (ii) a bounded cost function. In practice, the algorithm requires a tabular representation of the state space in order to store the value function updates. However, there is a wealth of MDP's with an infinite state space. Furthermore, there are problems with a finite state space but their size still prohibits the tabular representation of the value function. In such cases, the value iteration algorithm may still converge if one carefully designs value function approximation methods. Such methods include truncation techniques [9], where the decision maker updates a value function estimate for a finite subset of the state space. If the truncated problem is close enough to the original, the value iteration

algorithm converges. These value iteration variants are called *approximation value iteration algorithms*. On the other hand, the boundedness of the cost function is an assumption rather restrictive for problems with an infinite state space. It has been proven that the value iteration algorithm converges under more general conditions that allow unbounded cost functions with the first results appearing in Harrison [10] and Lippman [11]. For examples with truncated state spaces and unbounded cost functions, we refer the reader to Puterman [7].

An important measure of performance of the value iteration algorithm is its *convergence rate*, that is, the number of iterations required in order to obtain an ϵ -optimal policy. There are a number of value iteration algorithm variants that employ sophisticated value update routines, or *action elimination* routines, that is, eliminating early suboptimal actions, that demonstrate improved algorithmic performance. For example, the *Gauss-Seidel value iteration*, the *Jacobi value iteration*, the *action elimination value iteration*, and combinations of them are generally characterized by improved convergence rates for a variety of MDP problems. For an extensive discussion on the different value iteration variants, we refer the reader to Puterman [7], whereas for more recent developments we refer the reader to Herzberg [12,13], Shlakhter *et al.* [14], and the references therein.

Other methods for solving discounted MDPs include linear programming (LP) [15,16] and the *policy iteration* algorithm. The LP approach formulates the discounted problem as an equivalent LP with its optimal solution equal to the optimal value function. The policy iteration algorithm starts with an initial policy and, at each iteration, a new policy is constructed from the previous one with an improved cost-to-go vector. For an extensive discussion on these subjects we refer the reader to Puterman [7].

REFERENCES

1. Koole G. Stochastic scheduling and dynamic programming. Amsterdam: CWI; 1995.

2. Duffie D. Dynamic asset pricing theory. Princeton, NJ: Princeton University Press; 1996.
3. Berovic DP, Vinter RB. The application of dynamic programming to optimal inventory control. *IEEE Trans Autom Control* 2003;49(5):714–727.
4. Howard D. Dynamic programming and Markov decision processes. Cambridge, (MA): MIT Press; 1960.
5. Blackwell D. Discrete dynamic programming. *Ann Math Stat* 1962;33:719–726.
6. Blackwell D. Discounted dynamic programming. *Ann Math Stat* 1965;36:226–235.
7. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. Hoboken, New Jersey: John Wiley and Sons, Inc; 1994.
8. Bertsekas DP. Dynamic programming and optimal control. Volumes I,II, Belmont, (MA): Athena Scientific; 1995.
9. Fox BL. Finite state approximations to denumerable state dynamic programs. *J Math Anal Appl* 1971;34:665–670.
10. Harrison JM. Discrete dynamic programming with unbounded rewards. *Ann Math Stat* 1972;43:636–644.
11. Lippman JM. On dynamic programming with unbounded rewards. *Manag Sci* 1975;21:1225–1233.
12. Herzberg M, Yechiali U. Accelerating procedures of the value iteration algorithm for discounted Markov decision processes, based on a one-step lookahead analysis. *Oper Res* 1994;42(5):940–946.
13. Herzberg M, Yechiali U. A k-step look-ahead analysis of value iteration algorithm for Markov decision processes. *Eur J Oper Res* 1996;88:622–636.
14. Shlakhter O, Lee C, Khmelev D, *et al.* Acceleration operators in the value iteration algorithms for Markov decision processes. *Oper Res* 2010;58(1):193–202.
15. Trick M, Zin S. Spline approximations to value functions: a linear programming approach. *Macroecon Dyn* 1995;1:255–277.
16. Farias DPD, Roy BV. The linear programming approach to approximate dynamic programming. *Oper Res* 2003;51(6):850–865.

TOUR SCHEDULING AND ROSTERING

TOM ROBBINS
Marketing and Supply Chain,
College of Business, East
Carolina University,
Greenville, North Carolina

TOUR SCHEDULING AND ROSTERING IN SERVICE OPERATIONS

This article examines the process of scheduling individual workers to meet some form of time varying demand. This most often occurs in a service operation where the rate at which work presents itself varies significantly. Operations of this type include call centers, hospital emergency rooms, retail outlets, restaurants, and mail handling facilities.

In the *tour scheduling* process, the forecasted workload is converted into a schedule that specifies what shifts are to be staffed on each day by each employee over the course of a planning period, typically a week. The schedule is developed to meet the demand workload requirements at a minimum cost. The *rostering* process assigns the identified shifts to individual workers so as to best match individual preferences.

In this article, we first examine the problem of scheduling an inbound call center. This example illustrates the scheduling process and identifies some of the inherent issues involved in scheduling and rostering. We then discuss other environments where the same type of scheduling problems exist, and highlight some key differences in those environments.

A CALL CENTER EXAMPLE

A *call center* is a facility used to provide interactive communications with customers. In a call center, customer service representatives (CSRs), or agents, interact with customers over the telephone. A *contact center* extends

the notion of a call center to allow for other means of communication, including e-mails, faxes, and instant messages.

While call centers often deploy sophisticated telephony and computer equipment, and may require a large facility, the cost of the customer service agent is typically the most significant and controllable cost associated with call center operation [1]. Agent salaries typically account for 60–70% of the call center's total operating cost [2]. As such, effective capacity management is a critical success factor for efficient operations.

The workload presented to a call center, in terms of inbound calls, varies over the course of a week and over the course of a day. In order to efficiently staff the call center, the number of workers on duty must vary over the course of the week and the course of the day to match demand.

Call center managers are typically charged with providing a target level of overall service quality, most often specified based on some measure of caller wait time. The task of the call center manager is to develop a work schedule that assigns individual agents to work at specific time so as to achieve this service level goal at the lowest possible cost. In order to do this, call center managers must

- forecast the workload in terms of call volume;
- determine the number of agents required in each time period to meet the workload and achieve the service level target;
- package the agent level requirements into a set of shifts to which agents can be assigned;
- assign individual agents to tours taking into account the agent's availability and preferences.

Managers of small call centers may perform these tasks manually, but in most moderate to large call centers they receive automated support from a workforce management system (WMS).

SCHEDULING MODELS

The Two-Stage Scheduling Approach

Overview. The two-stage approach is a standard model for tour scheduling. In this approach, the staffing and scheduling decisions are made separately. Staffing requirements are developed based on the anticipated work in each time period. The staffing requirements are then used as an input to the scheduling model.

Staffing Requirements. The first step in the two-stage approach is to determine the staffing requirement, the number of workers required in each time period. Time periods are often 30 min in length, but in some cases 15 min intervals are used. In many environments, the number of workers required varies significantly over the course of the day.

In a call center, call volume tends to follow predictable patterns, with random variations superimposed. Consider daily call volumes shown in Fig. 1. In this call center, Mondays are in general the busiest day, Fridays are the slowest days, and the call center is closed on the weekends and holidays (such as Independence Day). The graph shows a 5-period moving average, which illustrates

that over the course of a week call volume is roughly constant with no clear trend up or down.

Seasonality also applies within the day. Consider the fairly common pattern illustrated in Fig. 2. Call volume grows throughout the morning, peaking in midmorning then dropping off in the late morning and slowing further over lunch. A second smaller peak occurs in the afternoon. Overnight volume tends to be low. Figure 2 represents a Monday; other days have similar but slightly different patterns.

Staffing requirements are determined using a queueing model, an analytical model that estimates key queue parameters such as waiting time, based on the rate at which calls arrive, the average length of calls, and the number of agents available to service those calls. Because queueing models typically require a steady rate of call arrivals, the *rate* at which calls arrive is usually assumed to be constant for each 30 min period. In the SIPP (Stationary Independent Period by Period) method, each 30 time period is analyzed independently of all other periods using the average call volume for that period as illustrated in Fig. 3.

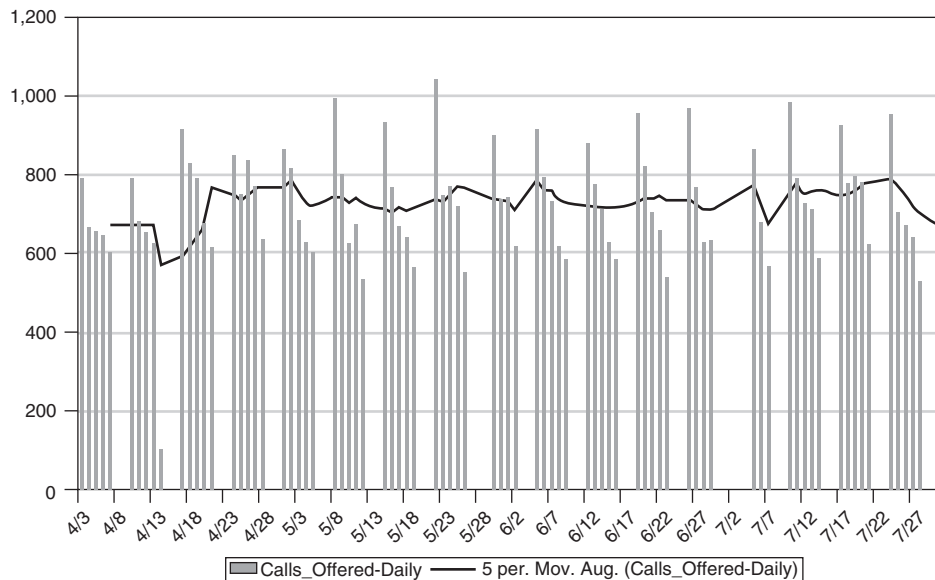


Figure 1. Calls offered daily.

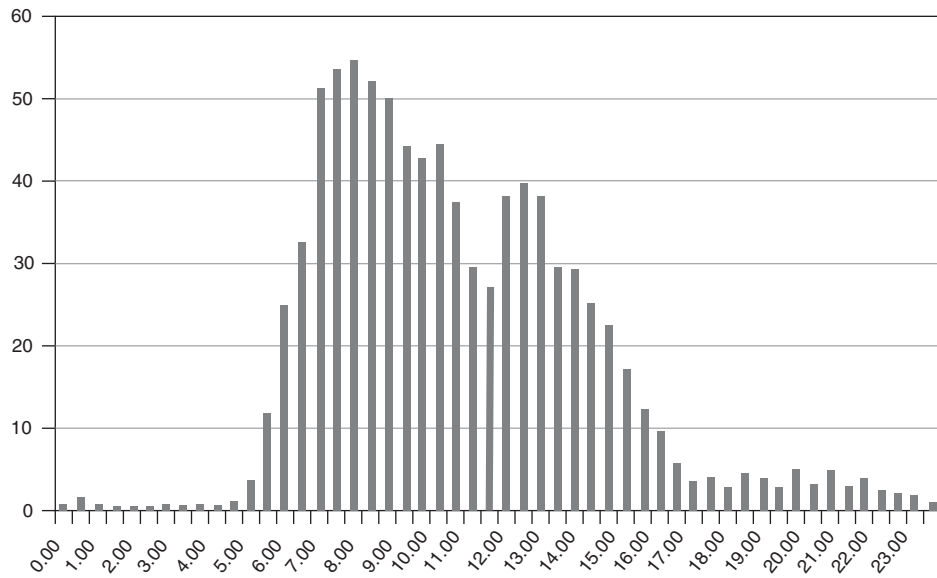


Figure 2. Average call volume.

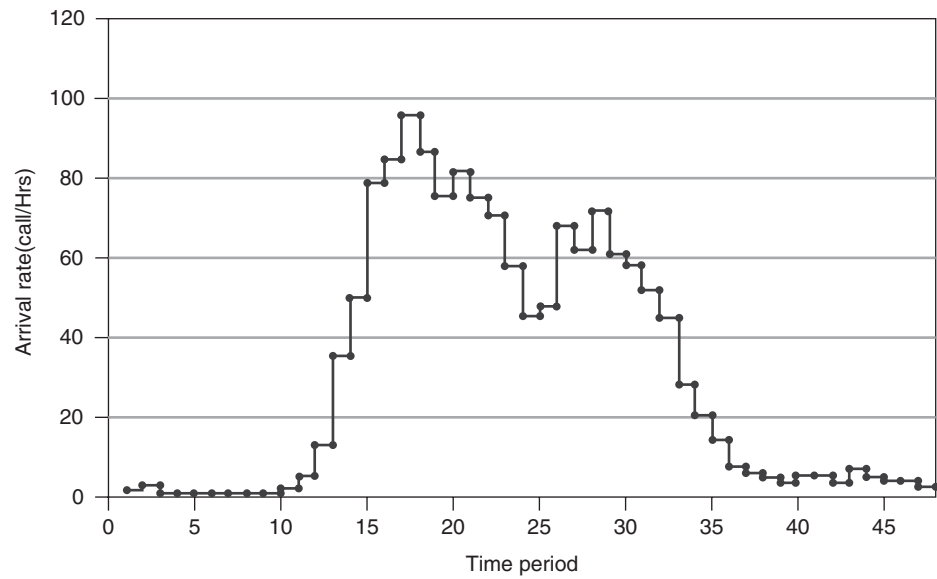


Figure 3. Arrival rates under SIPP.

The call center typically has a service level target, or a more formal service level agreement (SLA), that specifies the level of service to be provided in each period. For example, the SLA may specify that the

average time on hold should be 30 s or less, or that 80% of calls should be answered within 15 s. The expected call volume and service level requirements are inputs to the queueing model and the number of agents

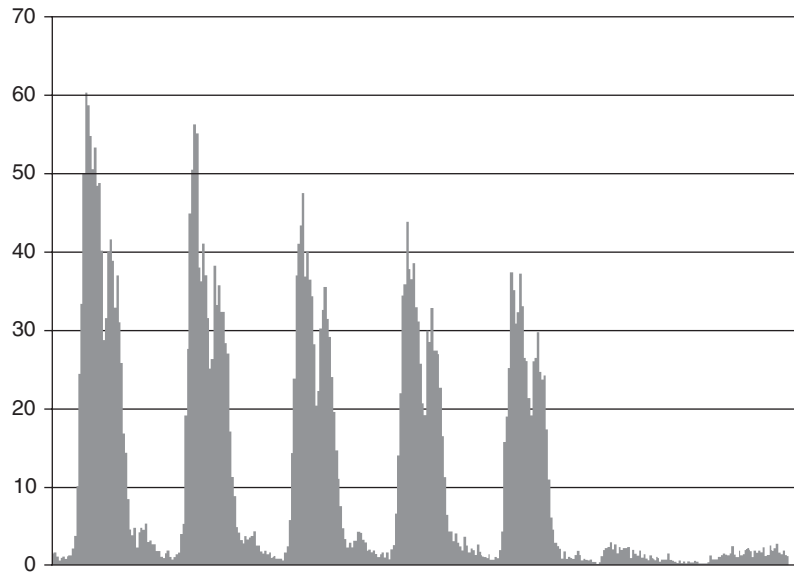


Figure 4. Minimum Agents required by 30 min period.

required per period is the output. Figure 4 graphically illustrates the output of the staffing requirements process.

Tour Scheduling. Once per period staffing requirements are defined, a tour schedule can be developed. The objective of the tour scheduling process is to determine the number of agents to assign to each possible schedule on each day so as to meet the minimum staffing requirements at the lowest possible cost. This implies minimizing the number of extra agents scheduled; that is, minimizing staffing above and beyond the minimum requirements.

A *shift* designates the specific time intervals, typically 30-min periods, during which an agent is on duty on a particular day. Consider the following simple example; a call center is open from 8:00 a.m. to 10:00 p.m., for a total of 28 half hour periods per day. Agents are scheduled to 9 h shifts, and breaks are not explicitly scheduled. A shift may start at any half hour period and lasts for 18 contiguous half hour periods. With these assumptions a total of 11 different shifts are possible, starting on each half from 8:00 a.m. to 1:00 p.m.

A schedule must further define the days of the week to be worked. Assume the call center is open seven days a week, but agents are only scheduled to work five days. Further assume that the scheduling policy is to give workers their two days off consecutively. This policy implies a total of seven different *day patterns*.

Given this relatively set of options, the scheduler has a total of 77 shift/pattern combinations ($11 \text{ shifts} \times 7 \text{ day patterns}$), which we will refer to as tours, to which agents can be assigned. The schedule covers 196 time periods ($28 \text{ half hour periods per day} \times 7 \text{ days}$). Here, we assume that each agent works the same shift every day. If this restriction is relaxed then the number of tours may increase significantly.

The Minimum Cost Schedule. Given a set of time phased staffing requirements (Fig. 4), a set of available shifts and a set of day patterns, the scheduling problem can now be formulated as an optimization problem—assign agents to tours so as to meet the minimum staffing requirements at the lowest cost. This problem can be formulated as an integer program, a linear optimization problem where

the decision variables can take on only integer values (See the section titled “Combinatorial Optimization/Integer Programming” in this encyclopedia). In this case, the decision variables are the number of agents to assign to each shift and day pattern, and since only whole agents can be assigned these numbers must be integers.

This model is an example of a *weighted set covering* problem. The basic formulation for solving this problem was first presented by George Dantzig in 1954 [3]. Dantzig looked at a similar problem, scheduling toll booth operators when the number of operators required was defined for each discrete time period.

In this model, a matrix is developed that maps tours to time periods. The matrix has I rows, one for every time period in the planning horizon (often a week), and J columns, one for each schedule (i.e., a shift and day pattern combination) to which an agent can be assigned. The cells in the matrix (denoted a_{ij}) are equal to 1 if the shift j is “on” in time period i and 0 otherwise. The cost associated with each schedule (denoted c_j) is a measure of the number of hours in the schedule. The decision variable (denoted x_j) specifies the number of agents assigned to each schedule j . The number of agents required in each period (denoted b_i) is the list of staffing levels defined in the staffing requirements step.

Mathematically, the model can be expressed as

$$\text{Minimize } \sum_{j \in J} c_j x_j$$

Subject to

$$\sum_{j \in J} a_{ij} x_j \geq b_i, x_j \geq 0, x_j \text{ integer } \forall i$$

This model seeks to minimize the number of agent/hours scheduled, subject to the constraint that the number of agents assigned in each period (found by summing up all the agents assigned to each schedule) must be greater than or equal to the number required in that period.

This model requires one decision variable for each tour a corresponding to the number of workers assigned, and one constraint for

each time period. For the example outlined above, the model would contain 77 integer decision variables and 196 constraints.

Flexible Scheduling and the Feasible Schedule Problem. While integer programs are in general difficult to solve, the model outlined above is reasonably sized and can be solved to optimality without much difficulty. However, changing a few assumptions in this model formulation can quickly result in a much larger model.

Changes that can increase the size of the model include

- *Hours of Operation.* Expanding the hours of operation of the call center; in particular, moving to a 24-h operation.
- *Shift Options.* Adding more shifts; in particular, creating a more flexible scheduling environment that includes different shift lengths and/or part-time options.
- *Breaks.* Factoring break time into the schedule; in particular, explicitly including meal and/or rest breaks in the schedule.
- *Variable Shifts.* Creating tours that include different shifts on different days.

The call center, we looked at previously, operated from 8:00 a.m. to 10:00 a.m. leading to 11 different shifts, one starting on each half hour from 8:00 a.m. to 1:00 p.m. If we allow for continuous operation (24 h a day), we have 48 possible shift patterns; one for each half hour of the day. The result is a more than fourfold increase in the number of shifts.

The number of feasible tours will also increase if more flexible scheduling options are introduced. In the example stated earlier, we scheduled each agent five days per week with consecutive days off. That led to five feasible day patterns. If we relax the consecutive days off requirement, the result is a total of 21 day patterns (calculated as a combination; 7 choose 5) and the result is a total of 1008 schedules.

Now assume that in addition to five 8 h shifts, the call center offers employees the

option of four 10 h shifts. This leads to 48 new shifts (one per half hour) and 35 day patterns (calculated as a combination; 7 choose 4.) The result is a total of 1680 schedules.

Now let us further assume that the call center allows for part-time workers. (Part-time staffing can be quite beneficial when trying to balance supply and demand.) Let us assume they allow for 8 h shifts 4 days a week. The result is another 1680 schedules.

So, if we assume we have a call center operating 24 h a day, 7 days a week (which has three basic scheduling options—five eights, four tens, and four eights), then agents can be assigned to any of 4368 schedules. The introduction of 24 h operation and moderately flexible scheduling has increased the number of integer decision variables from 77 to 4368.

The schedules described above do not explicitly account for breaks. Assume that the call center assigns agents to 9 h shifts, with a half hour lunch break and two 15 min coffee breaks. Further, assume that the call center wishes to explicitly schedule the lunch break, but not the rest breaks.

The call center may wish to have some flexibility in how breaks are scheduled. The shifts shown in Fig. 5 illustrate how adjusting when the lunch break occurs can lead to different schedules. If this flexibility is incorporated into the scheduling model, the

model has a total of 21,840 integer decision variables.

The simple model with 77 decision variables developed in the previous section has now expanded to a model with more than 21,000 decision variables. Introduction of additional flexibility, such as more part-time options, would increase this number further. The calculations shown above account only for lunch breaks not rest breaks. If the call center wanted to explicitly schedule the rest break, they would need to plan based on 15-min periods, and incorporate the rest break options. This would have the effect of doubling the number of time periods, and therefore the number of constraints. The number of decision variables would increase significantly, perhaps by an order of magnitude or more.

While adding staffing flexibility greatly increases the size and complexity of the scheduling problem, there are significant benefits. In call centers with large peaks and valleys in the demand profile, it is difficult to fit supply to demand without creating significant overcapacity in other periods. Scheduling flexibility, in particular the use of part-time workers, allows call centers to better shape supply to meet the demand curve and reduce the cost of service delivery [4].

The Complexity of Flexibility. The example discussed above shows that the number of

	S1	S2	S3	S4	S5	S6
8:00 a.m.	1	1	1	1	1	1
8:30 a.m.	1	1	1	1	1	1
9:00 a.m.	1	1	1	1	1	1
9:30 a.m.	1	1	1	1	1	1
10:00 a.m.	1	1	1	1	1	1
10:30 a.m.	1	1	1	1	1	1
11:00 a.m.		1	1	1	1	1
11:30 a.m.			1	1	1	1
12:00 a.m.	1			1	1	1
12:30 a.m.	1	1			1	1
1:00 a.m.	1	1	1			1
1:30 a.m.	1	1	1	1		
2:00 a.m.	1	1	1	1	1	
2:30 a.m.	1	1	1	1	1	1
3:00 a.m.	1	1	1	1	1	1
3:30 a.m.	1	1	1	1	1	1
4:00 a.m.	1	1	1	1	1	1

Figure 5. Example Shift Schedules with Breaks.

possible shifts, and therefore the number of integer variables in the optimization program can easily become very large. With explicit break scheduling, the number of decision variables can grow into the tens of thousands or more; test problems with millions of integer decision variables have been analyzed in the literature [5].

Integer programs are difficult (i.e., time consuming) to solve to optimality, and the amount of time and other computer resources required grows in a nonlinear fashion with problem size. Call center scheduling problems of practical size can easily become intractable; that is, solving the problem to proven optimality cannot be accomplished in a reasonable time frame. (See the sections titled “NP-hard Network Flow Problems” and/or “Combinatorial Optimization/Integer Programming” in this encyclopedia for a more detailed discussion of integer programming and complexity theory.)

Given the benefits of increased staffing flexibility, more staffing options are desirable from a cost control perspective. But, given the increased complexity of the optimization problem, staffing flexibility is undesirable from a computational perspective. Finding a balance between these competing goals has been an active area of research. Researchers have developed many approaches, most of which can be put into two basic categories.

- *Implicit Break Scheduling.* In this approach, the scheduling problem is conceptually split into two components. Schedules are first developed without breaks. Breaks are then scheduled and assigned to schedules.
- *Heuristic Solutions.* Heuristic algorithms are designed to find *good* solutions quickly, but usually cannot solve a problem to optimality.

Implicit Scheduling Models

A number of researchers have developed a modified formulation that includes an implicit definition of breaks. In the basic *implicit shift scheduling* approach, shifts are set up with no breaks, for a single day

(i.e., no breaks within a day or between days). A separate set of decision variables is established that specifies the number of employees who are on break during any given time period. Shifts and breaks are essentially established separately, although constraints are established to ensure that the breaks will fit into schedule break windows. After the optimization is complete, a relatively straightforward procedure can be used to assign breaks to individual shifts. In models where this formulation can be applied, the implicit formulation will solve much faster than the explicit set covering approach; requiring between 25% and 50% of the computer time required to solve the basic set covering model [6]. Other models have been developed that increase flexibility and may solve even faster [7,8].

Implicit scheduling models were later extended to *implicit tour scheduling* models which can be applied to 24×7 operations where employees are scheduled across multiple days [5]. This model solves the scheduling problem for a week when the days worked are continuous, and all shifts are the same length. Under the conditions, the model dramatically reduces the number of decision variables required and creates integer programs that are quite tractable.

Heuristic Solutions

A *heuristic* is a solution method that is designed to quickly find a solution that is “good” but not necessarily optimal (see the section titled “Heuristics and Meta-heuristics” in this encyclopedia). Heuristics have the advantage that they can often solve a problem much faster than a method that is guaranteed to solve the problem to optimality. A key disadvantage is that not only do they typically not solve the problem to optimality but they usually cannot provide a meaningful estimate of the optimality gap.

Given the potential size of staff scheduling problem and the issue of intractability, heuristics are often employed to solve the problem. Many types of heuristics can be applied to this problem. With one type of heuristic, the scheduling problem is still formulated as an integer program, but heuristics are used to either reduce the size of the

problem or to speed up the solution process. A second type of heuristic formulates the problem in a fundamentally different way. For the staff scheduling problem, this is most often done by formulating the problem as a discrete event simulation model and solving it using simulation-based optimization.

Integer Programming Heuristics. Given the inherent complexity of integer programs, the use of heuristics is fairly common (see the section titled “Heuristics and Meta-heuristics” in this encyclopedia). One viable approach is to formulate the integer programming (IP) model as defined above and then to apply a heuristic algorithm, such as a genetic algorithm or simulated annealing, to search for a good solution to the IP. The Lagrangian heuristic has often been used to solve set covering problems [9] (see *Relationship Among Benders, Dantzig–Wolfe, and Lagrangian Optimization*).

A second type of heuristic for solving the scheduling IP seeks to reduce the size of the problem by eliminating potential solutions. In this problem, a reasonable approach is to reduce the number of schedules that can be selected. A relatively simple method is to avoid explicit break scheduling, allowing supervisors to manage breaks off-line based on how conditions unfold over the course of the shift.

Another method for dealing with the large number of feasible schedules that can be assigned is to use some heuristic to reduce the set of available schedules to something much smaller. An example is presented in [10]. An initial set of 7,120 possible schedules is reduced to a *working set* of 100 or less using two different procedures; one that picks schedules that are the “most different,” and one that picks schedules at random. They find that scheduling against this small subset can yield solutions within a few agents of the optimal solution in much less time than required to solve the complete problem. They also find that for larger working subsets, a random selection performs nearly as well as the most different selection, and performs much faster.

Simulation-Based Heuristics. An alternate approach to solving the shift scheduling

problem is to use a discrete event simulation model (see *Introduction to Discrete-Event Simulation*). A simulation model generates and processes individual customers, calls in the case of a call center, according to the predefined probability distributions describing arrival rates, service durations, and other random factors. The simulation approach has the advantage of specifying a more realistic model of the operation; for example, by allowing for more realistic distributions of service times, varying and uncertain arrival rates, abandonment, and absenteeism.

Simulation models may be used to allow managers to evaluate policy or scheduling changes interactively. Saltzman and Mehrorta describe an application of simulation modeling to evaluate the impact that launching a priority support service would have on standard and priority customers in an IT support call center [11].

Another way to use simulation is to perform *simulation-based optimization*; a process whereby simulation is used to evaluate the schedule, and a search algorithm is used to look for better schedules (see the section titled “Simulation Optimization” in this encyclopedia). A drawback of simulation-based optimization is that it is in general impossible to determine if a solution is optimal. Instead of finding a known optimal solution, or a solution known to be within a specified percentage of optimal, the simulation optimization algorithm typically searches until it is unable to find a better solution for some designated length of time.

A common approach to simulation-based scheduling is to use an analytical method to create a rough cut schedule. The rough cut schedule is then evaluated using the simulation model. Changing the original schedule, often using randomized changes, generates new schedules, which can then be simulated. One approach uses a simple *greedy* heuristic to generate a first cut schedule for a call center with variable arrival rates [12]. A simulation model then searches for better schedules, evaluating similar schedules using a meta-heuristic search technique known as *variable neighborhood search* (VNS). Other metaheuristics include genetic algorithms,

simulated annealing or Tabu search (see the section titled “Heuristics and Metaheuristics” in this encyclopedia). Other models use simulation and cutting planes to develop a schedule for a call center where agents have different skill sets [13,14].

Joint Staffing Models

The basic set covering model splits staffing requirements and shift scheduling into two separate and distinct steps. Joint staffing models combine these two steps into a one optimization problem.

The basic set covering model implicitly makes several important assumptions that are relaxed in joint scheduling models. First, by taking as an input the number of agents required in every time period, the model implicitly assumes that service level requirements are strict in every interval. In call centers with short interval peaks (such as the call centers represented in Figs 2 and 4), staffing to satisfy the peak arrival rate may result in excess capacity in other periods. In response to this issue, some call centers seek to achieve their service level targets over an extended period; perhaps a day, week, or even a month. This is often referred to as a *global service constraint*.

A second issue is the implicit assumption that arrival *rates* are known prior to the scheduling process. While the standard queueing model used when setting staffing requirements assume that the time between call arrivals is random, the models assume the average rate at which those calls arrive is known. In many situations this is not the case, and arrivals are considered to be *doubly stochastic*; customers arrive randomly with an average rate that itself is random [15,16].

These limitations can be addressed in *joint staffing and scheduling models*; models that integrates the staffing requirement and the shift scheduling steps into one optimization problem. The simulation models discussed above are inherently joint staffing and scheduling models, but the approach can also be applied to optimization models [17,18]. Some models use a piecewise linear approximation to the service level curve; allowing for the calculation of a global service level

based on the per period staffing decision [19,20].

ROSTERING

The final step in the scheduling process is rostering. *Rostering* is the process of assigning individual agents to schedules defined in the shift scheduling process. The rostering process must consider employee preferences (e.g., I prefer to start after 10:00 a.m.), employee restrictions (e.g., I cannot work on Thursdays), management policies (e.g., all employees must work one weekend day a month), and potentially union or contract restrictions (e.g., agents must have two consecutive days off.)

While extremely important in practice, the rostering process has received limited attention in academic research, perhaps because the process is not one that has not traditionally been highly automated. The process used at New Brunswick Telephone (NBTEL) is described in Ref. 21, a case where an automated heuristic was used to improve the process. The article highlights the rostering process (referred to as the shift assignment problem in this article) and identifies many of the issues.

The input to the rostering process is the set of schedules defined in the shift scheduling process. The output is the assignment of specific individuals to all of the schedules. The objective of the process is to match employee preferences as closely as possible. Agent preferences are not all considered equal, and preferences are often evaluated relative to worker seniority with higher levels of seniority given first choice.

Agent preferences cover a range of attributes; for example, days worked per week, starting time, length of lunch break, specific days off, or length of shift. Agent preferences may be soft (e.g., I would prefer not to start before 8:00 a.m.), or hard (e.g., I cannot work on Sundays). An agent's skill level forms another hard constraint; agents cannot be assigned to shifts that require skills they do not possess.

In smaller call centers, agent preferences can be tracked by supervisors informally and

rostering can be done in a manual fashion. In larger call centers with hundreds or even thousands of agents, a manual process is not practical and call centers typically utilize an *agent self-service* module within the WMS.

The self-service module typically allows agents to indicate their hard constraints and express their preferences. An agent may, for example, indicate hours or days of the week for which they are unavailable. The system may be configured to forward this entry to a supervisor for confirmation. They may also enter a request to take a week of vacation into the WMS. The WMS may use a set of configurable rules to evaluate the request. The request may be denied, approved, or forwarded to a supervisor for evaluation and dispensation.

When it comes to shift assignment, assignments are often handled via a bidding process. Agents may use the self-service module to bid on schedules they find desirable. The WMS will typically utilize some rule set to attempt to meet agent preferences to the extent possible. Agent bids are often weighted based on agent seniority. The rules by which bids are evaluated are often complex, and a heuristic is used to balance multiple criterion (see the section titled “Heuristics for Multiple Objectives” in this encyclopedia). The heuristic used at NBTel is described in Ref. 21. The heuristic utilized at NBTel first tries to ensure that all schedules are assigned to a qualified agent; it then seeks to satisfy agent requests in order of seniority. If at the end of the run, shifts remain unassigned, the application performs a series of two and three way swaps to allocate these unassigned shifts seeking to ensure that a qualified agent is assigned to every scheduled shift.

Given the difficulty of expressing, much less satisfying agent preferences, the resulting assignments may not satisfy all agents’ desires. A common way to address this issue is for the WFM to provide automated support for a *shift swapping* process. Shift swapping is a process whereby two or more agents may agree to trade shifts. The WFM system may support this process in various ways. It may allow agents to view the detailed schedule to identify potential swaps. It may also provide automated workflow support for the

swapping process. For example, the system may allow one agent to request a swap with another agent, prompting a message to the effected agent. If that agent accepts the swap, a message may be sent to the supervisor seeking approval.

Real-Time Control

Service environments, in general, and call centers, in particular, are subject to significant uncertainty in both the demand for and supply of labor. Call center volume often varies significantly from forecast, and staff absenteeism is a common problem. The implication is there even a well developed schedule may prove inadequate when implemented.

Real-time control refers to the actions managers make during the course of the day to adjust capacity to better match realized demand. Real-time control actions can be short-lived changes that affect capacity for only a short period of time, such as rescheduling breaks, or longer-term changes that affect capacity for an hour or more, such as sending employees home, or calling in additional workers. An overview of real-time control, focused on the hospitality industry, is provided in Ref. 22. An overview of the literature associated with real-time schedule adjustments is provided in Ref. 23. In this article, the authors study the applicability of real-time control heuristics in the quick service restaurant industry.

When applied to call centers, much of the research on real-time control addresses the issue of when to take action, and focuses on determining when a statistically significant deviation from forecast has occurred [24,25]. Other papers focus on updating agent schedules [26].

Real-World Challenges

While a substantial body of academic research addresses the workforce scheduling problem, the literature often deals with models that are highly simplified versions of reality. Stated another way, workforce management application vendors, and call center managers face a reality far more complex than that addressed in the academic

literature. Assumptions made about the statistical distributions of call arrivals, talk times, and caller's willingness to wait are often inaccurate. Agents are heterogeneous with differing skills, productivities, and scheduling preferences. Agent absenteeism is a chronic problem in many call centers. Performing scheduling and rostering sequentially may create schedules that cannot be staffed given agent availability constraints. In many real-world applications, call center managers make substantial manual changes to system generated schedules.

There is a subset of the literature that attempts to quantify, and in some cases address, these complexities. Issues related to how statistical distributions in real-world call centers differ from those assumed in standard models are addressed in Ref. 16. The design of a commercial workforce management application is described in Ref. 27. A series of papers examine issues related to implementing call center solutions at L.L. Bean [28–31]. Heterogeneous agent skills and sophisticated skills based routing greatly complicate the agent scheduling problem, and while assumed away in many scheduling models, this is an area of active research (See Ref. 1 for a more detailed discussion.)

SCHEDULING AND ROSTERING IN OTHER ENVIRONMENTS

The bulk of this document looks at scheduling and rostering in call center environments. The techniques and models discussed here are also applicable to a range of other environments such as airline ticket counters, airline crews, hospitals, emergency services, restaurants, retail outlets, or mail facilities. The methods discussed here apply to any operation characterized by a time varying requirement for personnel. Ernst *et al.* give an overview of scheduling and rostering applications across a range of applications, highlighting some of the key issues that arise in various environments [9]. Call Centers are a common application for scheduling and rostering models, and as discussed in this article, call centers are characterized by

significant uncertainty in workload. Other notable application areas are listed below.

Transportation Systems

Scheduling and rostering is an important aspect of transportation systems, such as airlines. Many models have been developed to look at airline crew scheduling [9]. One important difference with airline models, in particular, is the spatial aspect of the models. Airline crew schedules are driven by point to point transportation links which move the crew from one location to another. Crews have home locations that they must eventually return to, and leaving a crew at a remote location creates an additional cost. Other transportation applications, such as bus systems, subways, and other public transport systems, are less focused on the starting and ending location of the crew as they tend to be confined to a narrow geographical area. Transportation scheduling staff requirements are also less uncertain and more controllable than those in a call center. The transportation schedule tends to drive the staff schedule.

Health Care

Many models have been developed to analyze staffing issues in health-care delivery, most notably nurse scheduling. Much of this work has focused on allocating weekday and weekend shift in accordance with work regulations. Another interesting aspect of health-care models is the ability to substitute higher skilled workers for lower skilled requirements; that is, assigning an Registered Nurse (RN) when only an Licensed practical nurse (LPN) is required.

Emergency Services

Staffing is an important issue in emergency services such as police, fire, or ambulance services. Spatial issues are also important in emergency services, in particular, ambulance services where the response time are critical and driven in large part by the distance between the incident and the responding crew. Ambulance services also create a demand profile that is highly uncertain.

Retail

Scheduling and rostering is an important characteristic in retail operations of all types. Retail scheduling applications tend to be similar to call center models, although staffing levels are often defined exogenously rather than via a queueing model approach.

REFERENCES

1. Aksin Z, Armony M, Mehrotra V. The modern call-center: a multi-disciplinary perspective on operations management research. *Prod Oper Manage* 2007;16(6):665–668.
2. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. *Manuf Serv Oper Manage* 2003;5(2):79–141.
3. Dantzig GB. A comment on Edie's "Traffic Delays at Toll Booths". *J Oper Res Soc Am* 1954;2(3):339–341.
4. Robbins TR. Managing service capacity under uncertainty [Unpublished PhD Dissertation]. Pennsylvania State University; 2007. 235p. Available at: <http://personal.ecu.edu/robbinst/>.
5. Brusco MJ, Jacobs LW. Optimal models for meal-break and start-time flexibility in continuous tour scheduling. *Manage Sci* 2000;46(12):1630–1641.
6. Bechtold SE, Jacobs LW. Implicit modeling of flexible break assignments in optimal shift scheduling. *Manage Sci* 1990;36(11):1339–1351.
7. Thompson GM. Improved implicit optimal modeling of the labor shift scheduling problem. *Manage Sci* 1995;41(4):595–607.
8. Aykin T. Optimal shift scheduling with multiple break windows. *Manage Sci* 1996;42(4):591–602.
9. Ernst AT, Jiang H, Krishnamoorthy M, *et al.* Staff scheduling and rostering: a review of applications, methods and models. *Eur J Oper Res* 2004;153(1):3.
10. Henderson WB, Berry WL. Heuristic methods for telephone operator shift scheduling: an experimental analysis. *Manage Sci* 1976;22(12):1372–1380.
11. Saltzman RM, Mehrotra V. A call center uses simulation to drive strategic change. *Interfaces* 2001;31(3):87.
12. Robbins TR. A simulation based scheduling algorithm for call centers with uncertain arrival rates. *Proceedings of the 2008 Winter Simulation Conference*; Miami (FL). 2008. pp. 2884–2890.
13. Avramidis AN, Gendreau M, L'Ecuyer P, *et al.* Simulation-based optimization of agent scheduling in multiskill call centers. 5th Annual International Industrial Simulation Conference (ISC-2007); Delft, The Netherlands. 2007.
14. Atlason J, Epelman MA, Henderson SG. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Manage Sci* 2008;54(2):295–309.
15. Jongbloed G, Koole G. Managing uncertainty in call centers using Poisson mixtures. *Appl Stoch Models Bus Ind* 2001;17:307–318.
16. Brown L, Gans N, Mandelbaum A, *et al.* Statistical analysis of a telephone call center: a queueing-science perspective. *J Am Stat Assoc* 2005;100(469):36–50.
17. Koole G, van der Sluis E. Optimal shift scheduling with a global service level constraint. *IIE Trans* 2003;35:1049–1055.
18. Thompon GM. Labor staffing and scheduling models for controlling service levels. *Nav Res Logist* 1997;44(8):719–740.
19. Robbins TR, Harrison TP. Call center scheduling with uncertain arrivals and global service level agreements. *Eur J Oper Res* 2010. In press.
20. Robbins TR. Addressing arrival rate uncertainty in call center workforce management. 2007 IEEE/INFORMS International Conference on Service Operations and Logistics, and Informatics; Penn State University, Philadelphia (PA). 2007. 6p.
21. Thompson GM. Assigning telephone operators to shifts at new Brunswick telephone company. *Interfaces*. 1997;27(4):1–11.
22. Thompson GM. Labor scheduling, part 4. *Cornell Hotel Restaur Adm Q* 1999;40(3):85–96.
23. Hur D, Mabert VA, Bretthauer KM. Real-time work schedule adjustment decisions: an investigation and evaluation. *Prod Oper Manage* 2004;13(4):322–339.
24. Shen H, Huang JZ. Interday forecasting and intraday updating of call center arrivals. *Manuf Serv Oper Manage* 2008;10(3):391–410.
25. Weinberg J, Brown L, Stroud JR. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J Am Stat Assoc* 2007;102(480):1185–1198.
26. Mehrotra V, Ozluk O, Saltzman RM. Intelligent procedures for intra-day updating of call

- center agent schedules. *Prod Oper Manage* 2009;19(3):353–367.
27. Fukunaga A, Hamilton E, Fama J, *et al.* Staff scheduling for inbound call centers and customer contact centers. Eighteenth National Conference on Artificial Intelligence; Edmonton, Alberta, Canada. 2002. Edmonton.
28. Andrews B, Parsons H. Establishing telephone-agent staffing levels through economic optimization. *Interfaces* 1993;23(2):14.
29. Andrews BH, Cunningham SM L.L. Bean improves call-center forecasting. *Interfaces* 1995;25(6):1.
30. Andrews BH, Parsons HL L.L. Bean chooses a telephone agent scheduling system. *Interfaces* 1989;19(6):1.
31. Quinn P, Andrews B, Parsons H. Allocating telecommunications resources at L. L. Bean, Inc. *Interfaces* 1991;21(1):75.

TPZS APPLICATIONS: BLOTTO GAMES

ALAN R. WASHBURN
Operations Research Department,
Naval Postgraduate School,
Monterey, California

Blotto games are a type of TPZS game that is solvable without first making a list of all pure strategies. This property is important because it permits large games to be solved, much more so than if all pure strategies had to be listed. The class is named after an example given in Ref. 1 that will be referred to and solved below. In that example, a fictional Colonel Blotto has 4 units with which to attack an enemy who has only 3 units, the object being to capture as many of four forts as possible. Each side can secretly divide its units over the four forts, and Blotto will capture each fort if and only if he assigns strictly more units to it than his opponent. If we let $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ be Blotto's and his opponent's allocations to forts, then the two sides might choose $\mathbf{x} = (2, 1, 0, 1)$ and $\mathbf{y} = (0, 1, 2, 0)$, in which case Blotto would capture the first and fourth forts, but not the other two.

Blackett [2] defined Blotto games in general. In a general Blotto game, player 1 has f units that he secretly allocates to n areas, so $\sum_{k=1}^n x_k \leq f$, and similarly player 2 has g units, with $\sum_{k=1}^n y_k \leq g$. The utility to player 1 (hereafter the "payoff") is $\sum_{k=1}^n A_k(x_k, y_k)$, where $A_k()$ is an arbitrary function representing the payoff to player 1 in the i th area. Blotto games are zero-sum, so the payoff to player 2 is implicit. Usually $A_k(x_k, y_k)$ increases with x_k and decreases with y_k , in which case each player will optimally allocate all of his units. All n of the payoff functions are identical in the introductory example, but that need not be the case in general. The characterizing features of a Blotto game are therefore four as follows:

- there are given numbers of identical units available to each side, and each

side can allocate his units as he wishes among several areas;

- neither side knows the allocations of the other, except for the totals;
- the payoff in each area depends only on the two allocations to that area;
- the payoff in the game is a sum of the payoffs in each area.

There has also been some interest in games with these four features, but which are not necessarily TPZS. Such games are covered in the article titled **Allocation Games**, with potential applications to economics. Blotto games are thus a subset of allocation games.

APPLICATIONS

There are some political situations that closely resemble Blotto games. Consider the United States Electoral College, where the collected electors from the various states meet to elect a president. The states are the areas, the resource for each party (assume only two significant political parties) is money, and the electors in each state are committed to the party that spends the most in that state. Even putting aside its cynicism about voter behavior, there are still two significant difficulties in making such a political application. One is the sequential nature of actual political campaigns, where each party closely watches the activities of the other and adjusts spending accordingly. This is not in accordance with the second characterizing feature of a Blotto game. The other difficulty is with the payoff. Laslier and Picard [3] distinguish between the "plurality game" where the payoff is the number of electors won, and the "tournament game" where the payoff is 1 if more than half of the electors are won, or otherwise 0. The plurality game is a Blotto game, but the tournament game is not because its payoff function is not a sum. The tournament game therefore does not benefit from special solution techniques that apply only to Blotto games, and it is

not true that optimal plurality strategies can be transferred to tournaments. Unfortunately, it is the tournament game that most closely resembles reality. On account of such difficulties, we cannot as yet claim any persuasive application of Blotto games to democratic politics.

Most of the applications of Blotto games are in the field of defence. One important application has been to the defense of a group of targets (ICBM silos, military bases, cities, etc.) against ICBM attack (so player 1's resource is an ICBM inventory, or perhaps an inventory of reentry vehicles) using Anti Ballistic Missiles (ABMs) (so player 2's resource is a number of ABMs). The resulting mixed strategies for the opponents are usually called *preallocated* to emphasize that the allocations require no observations about the details of the attack or defense chosen by the other side. In the simplest case the ABMs allocated to a target counter any ICBMs allocated to that target on a one-for-one basis, and the target is killed if and only if x_k strictly exceeds y_k , as in the introductory example. More generally, $A_k(x_k, y_k)$ represents a kill probability, and the sum $A(\mathbf{x}, \mathbf{y})$ is "expected number of targets killed." The Sections 4.3 and 5.6 of Ref. 4 give a good summary of specific formulations and solutions as of 1972.

In formulating a game that is potentially a Blotto game, careful attention must be paid to questions of who knows what and when. It would be significant if Blotto were able to discover the allocations of his opponent before making his own, or if the opposite situation should favor his enemy. In the introductory example, Blotto would capture at least two forts in the former situation, or only one in the latter. The value of the Blotto game ($5/3$ in the introductory example, as we will show below) will be somewhere in between those extremes, and is the value we should seek when both sides can protect the secrecy of their allocations.

Imagine that you are the defensive player in a game where there are 100 targets that you have to defend with 100 ABMs. The offense has 200 ICBMs with which he plans to attack the targets, so you are outnumbered. Each ABM will surely kill the

ICBM against which it is committed, and each surviving ICBM will kill the target against which it is committed. How should you employ the ABMs? One strategy would be the "myopic" strategy of intercepting the first 100 ICBMs that you see. Against that strategy, the offense could sacrifice 100 ICBMs and then use the remaining 100 to attack each target once. All 100 targets would be lost, so there must be a better strategy for employing the ABMs. A much better strategy would be to first observe the distribution of attackers, and then completely defend the most lightly attacked targets until you run out of ABMs. Against that strategy, the offense can do no better than to attack every target with two ICBMs, so 50 targets will survive. However, you might not be able to make the required observations. You might not be able to determine each ICBM's chosen target early enough, or the offense might make a "dribble attack" where the ICBMs show up at different times. Either possibility would foil the plan and lead to fewer than 50 targets surviving. A different defensive possibility, much easier to implement, would be to intercept each ICBM with probability 0.5. That strategy is particularly attractive because you do not need to know an ICBM's target when deciding whether to intercept it, and the strategy works just as well against dribble attacks. If the offense knows that you are using it, he should attack each target with two ICBMs, and $(0.5)^2$ of the targets will survive (about 25). The number of surviving targets varies between 0 and 50, depending on the tactics employed and the command-and-control situation.

The Blotto defense in the above situation would preallocate a certain number of ABMs to each target, keeping the exact number secret so that the offense cannot pick on the lightly defended targets. To be precise, the Blotto solution in this case has the number of ABMs assigned to each target being equally likely to be 0, 1, or 2 (note that the expected number of ABMs at each target is 1, as the ABM supply dictates). Even if the offense knows that distribution, he cannot do better than to assign 2 ICBMs to every target, thereby killing $2/3$ of the targets, on

the average, so about 33.3 will survive. This is an improvement over simply intercepting each ICBM with probability 0.5. In fact this is the best you can do as long as you know nothing about the attack other than the fact that it totals 200, and can determine each ICBM's target before committing an ABM. It would be better, of course, if you knew exactly how all the ICBMs were distributed over targets before committing any ABM. In fact, the value of that information can be quantified since it raises the average number of targets savable from 33.3 to 50.

In the above example, one-third of the targets in the optimal defense has no ABMs allocated. The logic is solid, but the result can still be bitter if you are one of the abandoned targets. Indeed, a case can be made that the only morally defensible way of defending cities is the provably ineffective myopic strategy—how can you deliberately let an ICBM attack Boston when you still have ABMs left? This is but one example of some of the ugly tactical quandaries produced by the Cold War. No ICBM attack ever occurred, but it was (and still is) necessary to plan for the possibility in the face of such issues.

There are also some military applications in areas other than missile defense. For example, strategic nuclear submarines basically have the job of remaining undetected in peacetime, awaiting the fateful day when they receive orders to launch their weapons. Those submarines must receive orders to patrol in particular parts of the ocean. If one imagines that the ocean is partitioned into n cells, one can arrive at a Blotto game where f submarines are allocated to those cells with the object of having as many submarines as possible remain undetected, while a certain amount of search effort is allocated to the same cells by an opponent with the opposite motivation. Games where one side must choose a route and the other must setup an ambush can also be viewed as Blotto games, with the areas being road segments. All of these applications share the idea that each side secretly distributes a single resource over multiple areas.

SOLUTION TECHNIQUES FOR DISCRETE, SYMMETRIC BLOTTO GAMES

In this section, we assume that the allocations of both players must be nonnegative integers; that is, that the Blotto game is discrete. We also make the symmetry assumption that all of the functions $A_k()$, are identical, and refer to each of them as $A()$. The methodology described below is essentially that of Ref. 5.

The total number of nonnegative, integer vectors $\mathbf{x}$ that sum to f is $\binom{n+f-1}{f}$, the number of combinations of $n+f-1$ things taken f at a time. This number is 35 for Blotto in the introductory game, or 20 for his opponent, so that game could be solved as a 35×20 matrix game. However, it should be clear that the brute force technique that begins by first constructing a complete payoff matrix will not work for large games. If $n = f = 10$, for example, Blotto would have 92,378 distinct strategies, even if we count only those that consume all of his resources. We seek a solution method that does not require explicitly listing all possible strategies.

We begin by seeking a bound on the game value that Blotto can enforce. A mixed strategy for Blotto is a joint distribution over the allocations to all n of the areas. Because of symmetry, each area should have the same marginal distribution for the number of units allocated—call that distribution $\mathbf{u} = (u_0, \dots, u_f)$, and let $E(\mathbf{u}, j)$ be the expected payoff at the j th area when player 2 allocates j units to that area. Thus

$$E(\mathbf{u}, j) = \sum_{i=0}^f A(i, j) u_i.$$

If Blotto can select $\mathbf{u}$ so that $E(\mathbf{u}, j) \geq a - bj$ for all j , then, since there are n identical areas and a total of g/n units per area available to player 2, the total payoff per area will be at least $a - bg/n$. Blotto wants to maximize this quantity, as in linear program P1:

Program P1: maximize $a - bg/n$
subject to the constraints

$$\begin{aligned}
\sum_{i=0}^f A(i,j)u_i &\geq a - bj; j = 0, \dots, g \\
\sum_{i=0}^f iu_i &\leq f/n \\
\sum_{i=0}^f u_i &= 1 \\
b &\geq 0, \text{ and } u_i \geq 0; i = 0, \dots, f.
\end{aligned}$$

In program P1, the first set of constraints forces the objective function to have the desired meaning, as explained above, and the rest require that $\mathbf{u}$ be a probability distribution over the integers 0 through f with a mean of at most f/n . If that distribution applies to every area, Blotto's mean total use of units will be at most f , as required. Counting a and b , which together with $\mathbf{u}$ constitute the variables, P1 has $f + 3$ variables and $g + 3$ constraints, not counting the nonnegativity constraints on $\mathbf{u}$. Neither the number of variables nor the number of constraints depends on n except through ratios, an important property.

With one caveat, the solution of P1 constitutes a lower bound on the value of the game that Blotto can guarantee. The caveat is that $\mathbf{u}$ must be *playable*, by which we mean that there must be some joint distribution of allocations over all of the targets such that

- the marginal distribution of the number of units used is $\mathbf{u}$ at every target, and
- the total number of units used never exceeds f .

The potential difficulty here is that, while P1 requires that the total number of units used should never exceed f on the average, Blotto's allocation must exceed f with probability 0, which is a stricter requirement. We will return to this potential difficulty later.

Consider the application of this idea to the introductory example. In that example, $A(i,j)$ simply indicates whether i is greater than j , so $E(\mathbf{u},j) = u_{j+1} + \dots + u_f$, and $f/n = 1$. Upon solving P1, we find that $\mathbf{u} = (1/3, 1/3, 1/3, 0, 0)$ and that the lower

bound is $5/12$. If $\mathbf{u}$ is playable, then Blotto can guarantee to capture at least $4(5/12) = 5/3$ forts, on the average. A moment's reflection reveals that $\mathbf{u}$ is in fact playable in this case. Blotto can consider the four forts in two pairs, to each of which he assigns exactly 2 units. Within each pair, Blotto attacks the first fort with 0, 1, or 2 units, equally likely, assigning the rest of the 2 units to the other fort. In this way, Blotto can make the distribution of units at every fort be $\mathbf{u}$, while simultaneously using exactly 4 units in total. Since $\mathbf{u}$ is playable, the value of the game is at least $5/3$.

We could, of course, also consider the game from the viewpoint of Blotto's opponent, who wishes to minimize the total payoff by making a clever choice of $\mathbf{v}$, the marginal distribution of the number of units assigned to the typical area. Similar logic to Blotto's leads to linear program P2

Program P2: minimize $c + df/n$

subject to the constraints

$$\begin{aligned}
\sum_{j=0}^g A(i,j)v_j &\leq c + di; i = 0, \dots, f \\
\sum_{j=0}^g jv_j &\leq g/n \\
\sum_{j=0}^g v_j &= 1 \\
d &\geq 0, \text{ and } v_j \geq 0; j = 0, \dots, g.
\end{aligned}$$

The variables in P2 are $\mathbf{v}$, c , and d , totaling $g + 3$ in number because the vector $\mathbf{v}$ has $g + 1$ components. Any feasible solution of P2 guarantees that the payoff per area is at most $c + df/n$, and the objective of P2 is to minimize that quantity. As long as the resulting $\mathbf{v}$ is playable, player 2 can guarantee that the total payoff per area will not exceed $c + df/n$, on the average.

In the case of the introductory example, the solution of P2 is that the objective function is $5/12$, and $\mathbf{v} = (5/12, 5/12, 2/12, 0)$. One way of playing $\mathbf{v}$ is as follows:

- Consider the four forts in pairs. Flip a coin to decide whether the first pair

gets 1 or 2 units, and give the rest to the second pair.

- If a pair gets 1 unit, flip another coin to decide which fort gets it.
- If a pair gets 2 units, roll a die to decide whether the allocation should be (0,2), (1,1), or (2,0), equally likely.

This method will obviously always result in using exactly 3 units, as is required. It also results in the marginal distribution $\mathbf{v}$ for the number of units assigned to any fort. The probability that a fort gets 0 units, for example, is $(1/2)(1/2) + (1/2)(1/3) = 5/12 = v_0$. Since $\mathbf{v}$ is playable, Blotto's opponent can guarantee that the payoff per fort does not exceed $5/12$, which is equivalent to guaranteeing that Blotto cannot capture more than $5/3$ forts, on the average. Since we earlier showed that Blotto can guarantee to capture at least that many forts, we have discovered that the value of the game is exactly $5/3$, and we have done so without ever making a list of all the strategies available.

It is no coincidence that linear programs P1 and P2 have the same value in the introductory game. The two linear programs are duals of each other, so both must have the same optimized objective function, and in fact either of them can be solved in order to recover all of the variables of both programs. We can therefore identify an algorithm for solving any discrete, symmetric Blotto game.

1. Solve P1 and/or P2 to discover $\mathbf{u}$ and $\mathbf{v}$, as well as a possible game value.
2. Test whether $\mathbf{u}$ and $\mathbf{v}$ are playable.
3. If both $\mathbf{u}$ and $\mathbf{v}$ are playable, then playing them is optimal and the game has been (easily) solved.
4. If either $\mathbf{u}$ or $\mathbf{v}$ is not playable, find some other way of solving the game, perhaps brute force.

Unfortunately, step 2 is, in principle, nearly as hard as solving the game by brute force in the first place, since there is no known algorithm for testing playability that does not begin by listing all possible allocations. This is disappointing, but should not be overly so. Methods for playing a given

marginal distribution can sometimes be guessed, as in the introductory example, and there are also some more systematic methods for playing arbitrary marginal distributions [6, Section 6.2]. It is encouraging that play is easiest in large games, the very ones that are most difficult to solve by brute force. Consider a Blotto game that is similar to the introductory example, except that f, g , and n are each 3 times as large. Blotto must now assign 12 units to 12 forts. The number of possible allocations of 12 units to 12 forts is over a million, but P1 and P2 do not change at all when the problem is scaled up, since the high-numbered variables are 0 and the high-numbered constraints are inactive even when there are only four forts. Furthermore, it occurs that $12\mathbf{u}$ and $12\mathbf{v}$ are both vectors of integers, so either $\mathbf{u}$ or $\mathbf{v}$ can be played by labeling 12 balls with a number of units and then randomly distributing the balls to forts. The same marginal distributions could also be played by triply replicating the strategy used earlier with four forts, but the point is that things simplify in Blotto games when the number of areas n is large. Any marginal distribution whose components are integer multiples of $(1/n)$ can be played in this manner.

As a result of considerations such as those above, large Blotto games are frequently dealt with by simply assuming playability. The entirety of the analytic effort is then in solving either P1 or P2, whichever is more convenient. This achieves the goal of "solving" Blotto games using a technique that does not require listing all possible allocations. Beale and Heselden [5] opine that the Blotto game without the playability restriction may actually be an improved model of reality, since various realistic phenomena make it impossible to forecast constraint levels exactly.

The results of this section depend on the assumption that the n areas all have the same payoff function. A similar approach will work if the payoff functions differ, although the analytic effort required increases substantially. A linear program similar to P1 must be solved for each of the different payoff functions, and it may be necessary to solve this collection of programs several times to

closely approximate the game value. Details can be found in Refs. 5 and 6, Section 6.3.

CONTINUOUS BLOTTO GAMES

If the number of units is very large, or continuous in nature (e.g., water being allocated to fires), we can consider a Blotto game where the allocations are real numbers, rather than integers. The bad news is that we can no longer make use of discrete linear programs, but the good news is that the potential for analytic solutions emerges.

Consider again the symmetric case where all n areas have the same payoff function. We can again search for Blotto's optimal marginal distribution, but the notation must change to deal with the continuous nature of the allocations. Let X be a random variable that stands for Blotto's allocation to a typical area, and let $P(X \leq x) = F(x)$; that is, let $F(x)$ be the cumulative distribution function (CDF) of X . This CDF replaces the marginal $\mathbf{u}$ that we employed in the discrete case. If player 2 allocates y units to an area, then the expected payoff is $E(F, y) \equiv \int_0^f A(x, y) dF(x)$, and if $E(F, y) \geq a - by$ for $y \geq 0$, then Blotto can guarantee that the payoff per area is at most $a - bg/n$, as before. The problem of maximizing this bound is

Problem P3: maximize $a - bg/n$

subject to the constraints

$$E(F, y) \geq a - by; y \geq 0$$

$$\int_0^f x dF(x) \leq f/n$$

$$b \geq 0.$$

Here the integral constraint on $F()$ is that the average allocation must not exceed f/n —other constraints requiring that $F()$ actually be a CDF on the interval $[0, f]$ should also be understood. The Stieltjes notation is required for the integrals because optimal CDFs in Blotto games are frequently discontinuous. Although the optimization problem is potentially difficult because it involves the unknown function $F()$, it is still solvable in some cases.

The simplest case where P3 is solvable is the majority-rules case where $A(x, y)$ is 1 if x exceeds y , or otherwise 0. In that case, $E(F, y) = 1 - F(y)$, so the corresponding constraint in P1 requires $F(y) \leq 1 - a + by$. We therefore consider CDFs of the form

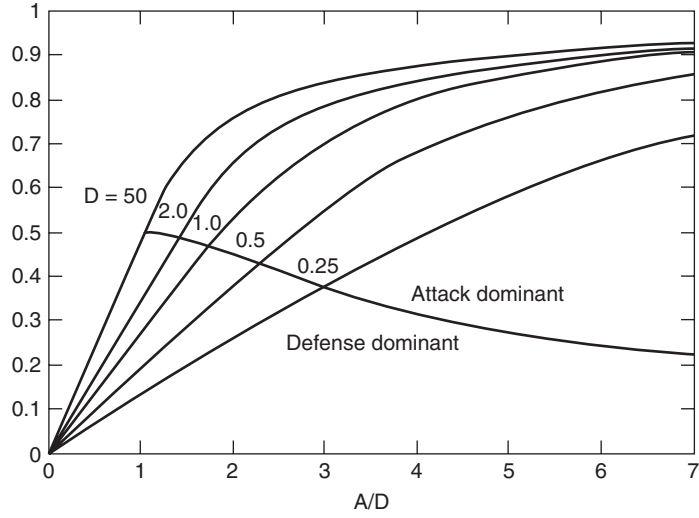
$$F(x) = \begin{cases} 0; & x < 0 \\ 1 - a + bx; & 0 \leq x \leq a/b \\ 1; & x > a/b. \end{cases}$$

Such CDFs describe an allocation that is 0 with probability $1 - a$, or otherwise uniformly distributed between 0 and a/b . The mean allocation for such distributions is $a^2/(2b)$, which we expect to equal f/n when $F()$ is optimal; that is, $b = a^2/(2(f/n))$. Substituting this into the objective, we find that a should be chosen to maximize the quantity $a - a^2g/(2f)$. Equating the derivative to 0, we find that $a = f/g$, which is optimal as long as it is smaller than 1 (parameter a cannot exceed 1 because that would make $F(0)$ negative). If $f/g \leq 1$, substituting our solution for a into the objective reveals that Blotto can guarantee a payoff of at least $f/(2g)$. If $f/g > 1$, Blotto can still guarantee an average payoff of $1 - g/(2f)$ by making $a = 1$. In summary, Blotto can guarantee a payoff per area of

$$\begin{cases} f/(2g) & \text{if } f \leq g \text{ by choosing } a = f/g, \\ 1 - g/(2f) & \text{if } f > g \text{ by choosing } a = 1. \end{cases}$$

By considering a similar problem for Blotto's opponent, we can show that the opponent can also guarantee the same payoff; that is, the game has been solved as long as playability questions are ignored. The side that has the most units is said to be "dominant." The dominant side uses a uniform marginal distribution in each area. The dominated side does the same thing, except that he first abandons enough areas to make a uniform distribution over the same interval feasible. For Blotto, the abandoned fraction is $1 - a$. If we apply this result to the ABM versus ICBM example considered above, but ignoring the integer constraints, we find that the fraction of the targets killed by the dominant attacker is $1 - 100/(400) = 0.75$. The fraction found earlier with integer constraints is $2/3$,

Figure 1. The fraction of targets killed versus the ratio of ICBMs to ABMs, when the ICBMs are imperfect.



smaller than 0.75 because the fact that the defense wins ties is significant. Analytic solutions are also available [7] for the case where the defenders are perfect, but the attackers are not. Figure 1 shows the fraction of targets killed by the attacker when an unintercepted attacker's kill probability is only $p < 1$, instead of 1 as assumed above. In that figure, $A \equiv -\ln(1-p)(f/n)$ and $D \equiv -\ln(1-p)(g/n)$, where $\ln()$ is the natural logarithm and n is as usual the number of areas. The upper curve in Fig. 1 is very close to the majority-rules game value derived above.

Majority-rules Blotto games with continuous allocations were considered by Borel [8] as early as 1921, including a demonstration of playability. See also Refs 9 and 10 for complete solutions of various continuous games of this type.

Another solvable class of Blotto games is those where $A(\mathbf{x}, \mathbf{y})$ is a concave function of $\mathbf{x}$ for all $\mathbf{y}$. For such games, it is known [6, Section 5.2] that player 1 can safely confine himself to pure strategies, so that the value of the game is $\min_{\mathbf{y}} \max_{\mathbf{x}} A(\mathbf{x}, \mathbf{y})$. While the resulting computational problem is still significant, it is at least simpler than solving a continuous game with mixed strategies. If $A(\mathbf{x}, \mathbf{y})$ is actually a *linear* function of $\mathbf{x}$ for each $\mathbf{y}$ (in which case it is automatically convex), then the problem of computing

$\max_{\mathbf{x}} A(\mathbf{x}, \mathbf{y})$ is a linear program. The dual of that program is a minimization problem, which can be combined with the minimization on $\mathbf{y}$ to produce one master minimization problem that can be solved with standard optimization packages. For an example of this technique, see Ref. 11. Similar conclusions hold if $A(\mathbf{x}, \mathbf{y})$ is a convex function of $\mathbf{y}$ for all $\mathbf{x}$.

If the payoff function is convex in $\mathbf{y}$ for all $\mathbf{x}$ and at the same time concave in $\mathbf{x}$ for all $\mathbf{y}$, then the Blotto game has a saddlepoint. Perhaps the simplest class of games where this is true are logistics games, which have a payoff function of the form $A(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^n x_k f_k(y_k)$, where $f_k()$ is convex. This function is linear in $\mathbf{x}$, so the total resource of player 1 might as well be taken to be $f = 1$, in which case x_k can be interpreted as "the probability that player 1 chooses area k ." The payoff in area k is determined entirely by player 2, who is as usual subject to a constraint g in allocating $\mathbf{y}$. The areas might correspond to different routes for a convoy, with $\mathbf{y}$ being an allocation of interdiction resources of some kind to the underlying network. This application is responsible for the name of the class, but there are other possibilities. The areas might be opportunities for terrorists, with $f_k(y_k)$ being the payoff to the terrorists if player 2 spends y_k in defending against the k th attack opportunity. The areas could also

be places for player 1 to hide, with g being a total amount of search effort for player 2. Regardless of the interpretation, logistics games can be easily solved because player 2 is effectively designing against the worst case choice of player 1. The value of the game (v) and Player 2's optimal strategy can be recovered from

$$\begin{aligned} &\text{Problem P4: minimize } v \\ &\text{subject to } f_k(y_k) \leq v, \\ &\quad \sum_{k=1}^n y_k \leq g, \text{ and} \\ &\quad y_k \geq 0; 1 \leq k \leq n. \end{aligned}$$

Such problems are usually simple enough to solve in spreadsheets.

REFERENCES

1. McDonald J. A theory of strategy. *Fortune* 1949;102.
2. Blackett DW. Pure strategy solutions of blotto games. *Nav Res Log* 1958;5:107–109.
3. Laslier JF, Picard N. Distributive politics and electoral competition. *J Econ Theory* 2002;103:106–130.
4. Eckler AR, Burr SA. Mathematical models of target coverage and missile allocation. Alexandria (VA): Military Operations Research Society; 1972.
5. Beale EML, Heselden GPW. An approximate method of solving blotto games. *Naval Res Logistics*; 1962; pp. 65–79.
6. Washburn A. Two-person zero-sum games, Topics in OR Series. 3rd ed. Linthicum (MD): INFORMS; 2003.
7. Galiano R. A simple analytic model for preallocated defense, lambda corporation paper 30. Lake Orion (MI): Lambda Corporation.
8. Borel E. La théorie du jeu et les équations intégrales à noyau symétrique. *C R Acad Sci* 1921;173:1304–1308. Translated by Savage LJ. The theory of play and integral equations with skew symmetric kernels. *Econometrica* 1953;21:97–100.
9. Gross O, Wagner R. A continuous Colonel Blotto game, RM-408. Santa Monica (CA): RAND Corp.; 1950.
10. Roberson B. The Colonel Blotto game. *Econ Theory* 2006;29:1–24.
11. Wood RK. Deterministic Network Interdiction. *Math Comput Model* 1993;17:1–18.

TRACKING TECHNOLOGIES IN SUPPLY CHAINS

MARIO M. MONSREAL
Zaragoza Logistics Center and the
Center for Latinamerican Logistics
Innovation, Avda, Zaragoza, Spain

DAI HONGYAN
TSENG MITCHELL
Department of Industrial Engineering
and Logistics Management, Hong
Kong University of Science and
Technology, Kowloon, Hong Kong

BROCK DAVID
Massachusetts Institute of
Technology, Cambridge,
Massachusetts

SUPPLY CHAIN TRACKING SYSTEM

Introduction to Tracking System

There is still a great deal of confusion on the terms involved in tracking systems. Very little attention has been paid to point out the differences among the various concepts of these systems. These concepts, however, are rather important, since they determine the objectives and functions of such systems. "Traceability" is a term often used as synonymous of tracking, and is defined as "the ability to trace the history, application, or location of that which is under consideration" [1], which is clearly related to the International Standard Organization definition "ability to trace the history, application or location of an entity by means of recorded identifications" (ISO 8402). In addition, according to McFarlane *et al.* [2], "Automated identification involves the automated extraction of the identity of an object" [2]. These definitions can let envisage the system's main function to provide visibility and traceability. The first term basically refers to the level of communication between the different nodes and agents in the supply

chain [3]. From the logistics perspective, it is considered as the transparent view of time, place, status, and content [4]. Or from the IT perspective, the ability to access or view relevant data or information as it is related to logistics and the supply chain as a whole [5]. From the latter we can infer one definition that fits the purpose of this document: A tracking system is an information system that captures current and/or past information on location, status, physical integrity, environment conditions, identity, or life cycle of moving objects. When this identification takes place in an automated way in order to provide tracking or tracing information, then it is understood as an Auto-ID system. This information could be used for collaborative decision making processes, or some other types of managerial or administrative processes. A tracking system is, in simple words, a supplier of such information.

Architecture

The tracking system architecture comprises three main layers: identification, capturing, and information exchange.

The identification layer focuses on registering information of the object for identification purposes. This information registration has activities such as writing, encoding, and decoding of the object information.

The capturing layer, in its turn focuses on retrieving the information stocked in the identification registry. Activities such as reading, decoding, and encoding object information form part of this layer.

The exchange layer takes care of storing, sharing, and exchanging specific information among the different agents and it also draws queries and documents such as certificate profiles and pedigrees.

As an example, we can clearly distinguish all the above layers in the architecture framework of an EPCglobal Standards Development Process shown in Fig. 1 [6].

There are some other architectural formats for tracking systems, as the one proposed by Cheung *et al.* [7]. They suggest

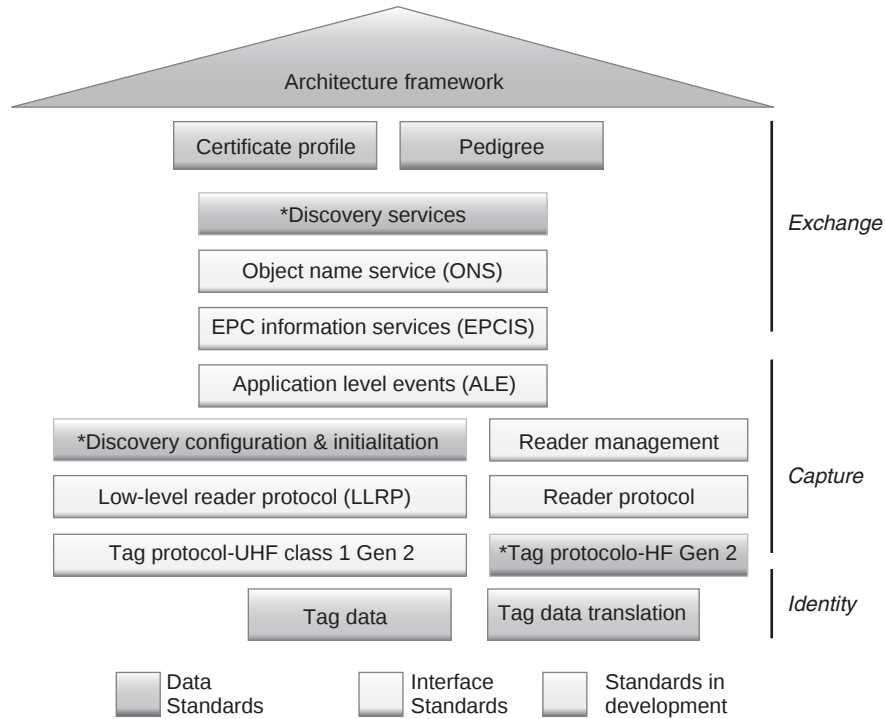


Figure 1. EPCglobal standards architectural framework [6].

a three-tier architecture with device, infrastructure, and application functions. Device tier is simply the first two layers of our previous architecture, which accounts for identification and capturing roles. Infrastructure, in turn, provides information, middleware, and discovery services. Finally, application tier provides value-added services to their users through software components (i.e., ERP and WMS).

TRACKING TECHNOLOGIES IN THE SUPPLY CHAIN

All tracking systems are made up of some basic components. These tools interact with each other to identify, register, modify, store, communicate, and emit specific and relevant information to the user, based on different technologies. As these technologies have been experiencing a wider and more global usage worldwide, even though they are more concentrated in developed countries due to their

labor-technology cost ratio in these regions, their use is also growing in current developing countries because of new international regulations and tendencies in global trade.

For operations research, tracking technologies could help leverage areas such as retailing, for example, tracking systems can help find customers' shopping behaviors at the store, or increase inventory management efficiency through an increase in inventory accuracy [8]. In product life cycle management, for example, by identifying products at the item level, there is an increased visibility that could be used to better control their entire life cycle [9]. Process optimization, for example, efficiencies in processes such as replenishment could be gained due to more precise and on time information gathered by tracking technologies [10].

In this section we have chosen a set of tracking technologies whose performance covers main functions of tracking systems such as identification, capturing, and

exchange information; and we summarize the most representative of these technologies.

Identification and Capturing Technology

These technologies are the link between the real-world and the information systems. Their main function falls in the identification and capturing layers of the tracking system architecture. They identify real objects and allow specific information to be electronically introduced in the tracking system.

Optical. Optical technologies capture data through the visual interaction of readers and labels and its main characteristic is the “line of sight” needed for reading. These technologies can read linear coded information as well as multidimensional coded information. Later we describe the most relevant.

Barcode. It is a linear binary code formed by lines and spaces (white lines) of varying thicknesses printed in different combinations. Its most common reading technologies are lasers and cameras [11]. Barcodes have been implemented in a wide variety of applications and environments in the supply chain. From all types of inventory control to check out processes in point of sales, currently barcode scanners rely on wireless data transferring to provide remote data capture through mobile devices such as cell phones, all based on barcodes [12].

GS1 128. The GS1 128 is a lineal barcode with a high density of information. It is an extension from numeric conventional barcodes such as EAN 13. The GS1 128 can include all the ASCII coding characters [13].

This code is modular, which means that we can add more information continuously. The barcode size is in direct relation to the quantity of information that it holds. For instance, for mass consumption, a code grouping ID, expiration date, and a six numerical characters describing lot number, will take up approximately 121-mm length and 164-mm width. The height will be always constant, which in this case is 33 mm.



Figure 2. Bidimensional barcode: PDF 417 Data-Matrix.

Bidimensional Coding. The 2D Technologies are usually squared or rectangular patterns of data codification. Given their increased code spatial density and code reliability capabilities, they are suitable for more challenging environments or conditions, such as smaller products or surfaces that are exposed to more deterioration due to product manipulation. Pharmaceutical industry is a good example of an implementation of this type of code [14].

There are two general categories based on coding type:

Coding by levels. It can be read by special bidimensional readers or CDD and laser readers with the help of special software to decode the information. The standard most often used in this technology is the PDF417.

Coding by matrix. It consists of a bidimensional matrix, is usually more compacted than a PDF417 code and can be read only by a bidimensional reader. The code mostly used is Datamatrix (Fig. 2).

The main advantage of bidimensional code is the capacity to codify vast amounts of information in a more reduced area. Bidimensional codes can codify approximately 2000 characters in a single label. Next we explain the two main types of bidimensional codes.

PDF-417. This code can store a maximum of 1850 alphanumeric characters (ASCII) or 1100 binary codes in a label with an area of 3 by 4.5 cm.

Mainly, PDF-417 is used in transport labels when the information has to be read although the system is not accessible. It is also used in inventory management or identification of cards. This technology incorporates “built-in redundancy,” property that enables the card been read although it is 60% damaged.

In addition PDF-417 has the capacity of storing graphics [15].

To read the PDF-417 code, apart from using bidimensional barcode readers, multipurpose scanners can be used, with specific software [16]. Conventional lineal barcodes readers are not useful. As all optical labels, PDF-417 has to be read with a direct "line sight" [17].

Datamatrix. Datamatrix is code in matrix form; it can store up to 2335 alphanumeric characters. The label is square shaped and it can extend from 0.0254 mm to 335.6 mm sideways. The size-information rate is approximately the following: the coding of 43 alphanumeric characters will be equivalent to a label of 4.2 mm sideways [18].

The Datamatrix codes are composed of squares modules placed within a pattern. This pattern consists of a frame which does not change and delimits the reading borderland; inside, the small modules are fully filled or completely blank to form a binary coding.

ECC200 is the last version of Datamatrix and has advanced coding algorithms error correction. This error correction algorithm enables that the codes damaged up to 60%, as in PDF-417, will be recognized. This version is widely being used in automobile, aerospace sector, electronics, semiconductor, medical devices, or in any case where product traceability is required. Because of the ability of Datamatrix to codify a large amount of data in a smaller code, it is being used in the cases where the reduced size of the product is a relevant factor [19].

The Datamatrix codes can be printed using most types of common thermal printers. Again, this optical symbol must also be read with a direct "line of sight."

The normative that defines this technology is ISO/IEC 16 022.

Radio Frequency. The radiofrequency technology is based on the use of radio frequencies for several purposes; these frequencies can be defined as wave oscillations within an entire radio spectrum range of 3 Hz to 300 GHz. In the supply chain environment, it is the base of contactless identification

systems without the need of line of sight. The range of frequencies where this technology is mostly used in supply chain starts from LF (low frequency) to a microwave, through HF (high frequency) and UHF (ultra high frequency). The minimum frequency used is 100 kHz while the maximum frequencies reach 2.4 GHz. The transmission frequency determinates the reading range and the ability of the signal to get over certain obstacles.

In the supply chain applications environment, there are different systems based on radio emissions, all of them have similar way of operation, but with a quite large number of possible applications such as asset and people tracking, product-user linkage, biometric measurements, product components tracking, retail operations, real-time inventories, and so on [20].

We next describe the main systems of radiofrequency that are mostly employed in supply chains.

RFID. The radiofrequency identification systems are the most known amongst the radiofrequency based systems. The first highlight is that these systems are oriented to information storage functions instead of exclusively transmitting, processing, or positioning of information/objects. RFID (radio frequency identification) is a formidable building block for an "internet of things," a network of electronically interrelated items that would allow companies to track goods through several simultaneous applications [21]. Their main function is as reflective transponders (mirror function). RFID are mainly based on two types of tags and their characteristic of a high area/reading rate. Tags can be classified into actives or passives depending on the type of the energy feed, which can be extern, (energy that tags get from the reader), or internal (energy provided by a battery incorporated in the tag). Passive tags do not have an internal power supply; thus they stay asleep until the moment they enter within a reader's range, using its energy to activate. Because of the high cost, size, and weight of the active tags, they are not considered as a good choice to be implemented in supply chains with

- 1.- A tag enters the RF field
- 2.- RF signal delivers energy to the tag
- 3.- The tag transmits data
- 4.- Antenna receives the data
- 5.- The reader receives and sends the data to the CPU
- 6.- CPU executes an action

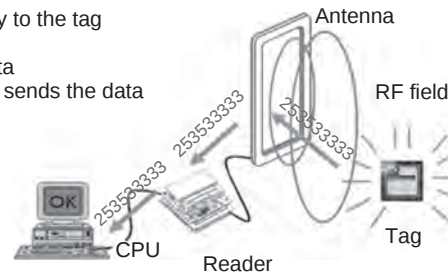


Figure 3. RFID operation.

low cost products. Passive tags are usually incorporated in a paper label jointly with a barcode and other legible information.

RFID operations have three main stages: emission, information recovery, and reception, which are carried out by three elements such as reader, antenna, and tag (label that includes the relevant information from the identification) (Fig. 3). The basic operation is the following: the reader emits a signal through the antenna, in a specific frequency captured by the tag, then the tag modifies the signal appending the required information (data) to give it back to the reader. This system enables the communication between the composing elements (tag and reader), without the need of being in line of sight, but only in range. Thus, it is possible to read items that are inside a container (box, packages, etc.) without the need of opening it up. Hundreds of tags can be read *practically* at the same time in *proper* situations.

The characteristics of an RFID system are linked directly to the type of tag, and frequencies in which the system is performing. Frequency is a major determinant of the intensity of the radio waves, and thus, a key factor in performance levels and applications. RFID mainly performs operations in four frequency bands: LF, HF, UHF and MF (microwave frequency) [22].

In Table 1 we can see how this specification relates to the devices.

According to Information Access and Modification Capabilities, Tags can be classified as follows:

1. Read Only
2. Write Once Read Only (WORM)

3. Read Write
4. Read Write (with on board sensors)
5. Read Write (with integrated transmitters)

We can see tag classes in Table 2 and their advantages, disadvantages, and remarks in Table 3.

One final issue regarding RFID description is its need for standardization, in order to achieve ubiquitous deployment. Standardization must start with the information encoding within the tag, this is closely related to the aim of the information and thus, to the characteristics of the tags. An RFID class structure is proposed by the MIT Auto-ID Labs [23], which comprises five classes, starting with identity, higher functionality, semipassive, active ad hoc, and reader tags. Each successive class includes functionality of the previous classes.

RuBee. RuBee is a specific protocol of wireless transmission to send and receive (transceiver) little information packages (about 128 bytes). This protocol (IEEE P1902.1) is magnetic (inductive) and works in low frequencies (large wavelength in 131 kHz, about 2.5 km). The latter makes it a low consumer and, probably its most important characteristic, friendly to read through and around liquids and metals.

The RuBee protocol is similar to IEEE 802's, so it is more similar to Wi-Fi (IEEE 802.11), Bluetooth (IEEE 802.15), and Zigbee (802.15.4) [24] than to RFID. Another feature of the RuBee protocol is the signal decay. The range mostly varies from 8 to 10m volumetrically. This is because signal decay

Table 2. Tag Classes

Class	Known as		Memory		Power Source	Application	
0	EAS	EPC ^a	None	EPC	Passive	Antitheft	ID
1	EPC		Read only		Any	Identification	
2	EPC		Read write		Any	Data logging	
3	Sensor tags		Read write		Semipassive/active	Sensors	
4	Smart dust		Read write		Active	<i>Ad hoc</i> networking	

^aThe reaction on EPC standards evolution shows that the EPC class 0 is likely to evolve to a read write

Table 3. Comparison of Passive and Active tags [33]

Tag	Advantages	Disadvantages	Remarks
Passive	Longer life time Wider range of form factors Tags are more mechanically flexible Lowest cost	Distance limited to 4–5 m (UHF) Strictly controlled by local regulations	Most widely used in RFID applications Tags are LF, HF or UHF
Semi passive	Greater communication distance Can be used to manage other devices like sensors (temp, pressure etc.) Do not fall under the same strict power regulations imposed on passive devices	Expensive—due to battery and tag packaging Reliability—impossible to determinate whether a battery is good or bad, particularly in multiple transponder environments	Used mainly in real time system to track high value materials or equipment throughout a factory Tags are UHF
Active		Widespread proliferation of active transponders presents an environmental hazard from potentially toxic chemicals in battery	Used in logistics for tracking of containers on trains, trucks, etc. Tags are UHF or microwave

follows an $1/R^3$ behavior instead of a most common $1/R$, (Fig. 4) [25]. Tags designed for this technology offer a 64 bits storage capacity. They can also accept temperature, humidity, acceleration, and pressure sensors. There is, however, a cost for these advantages, because low frequencies have smaller bandwidths for data transfer, the reading rate per second is lower than in HF or UHF tags. Owing to this fact, applications for RuBee should not be focused on the passage of fast-moving tagged goods through control points [26].

Communication and Information Exchange

Communication in a tracking system can be performed by a variety of technologies; this

is the task of exchanging information in the form of data. This section presents a relevant subset of these technologies.

Wireless Sensor Networks. A Wireless Sensor Network (WSN) is defined as a network of transceivers, sensors, machine controllers, microcontrollers, and user interfaces with at least two nodes communicating by wireless ways [27].

The main WSN function is to provide the required information to smart environments; their responsibility is to “sense,” as well as to execute the first stages of data processing. The WSN is generally composed of two sub networks: data gathering and data distribution (Fig. 5) [28].

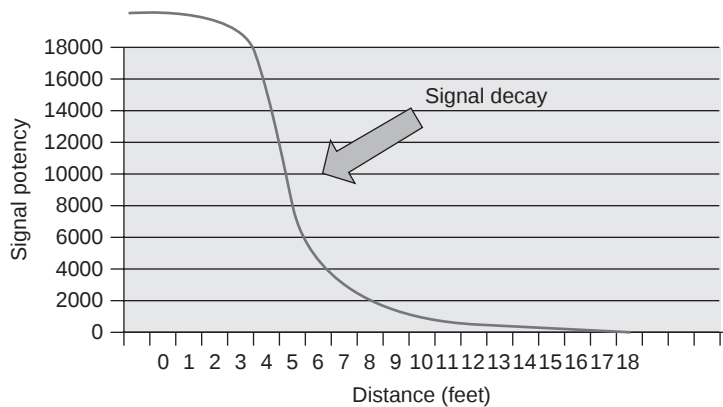


Figure 4. RuBee signal capture (low frequency) [25].

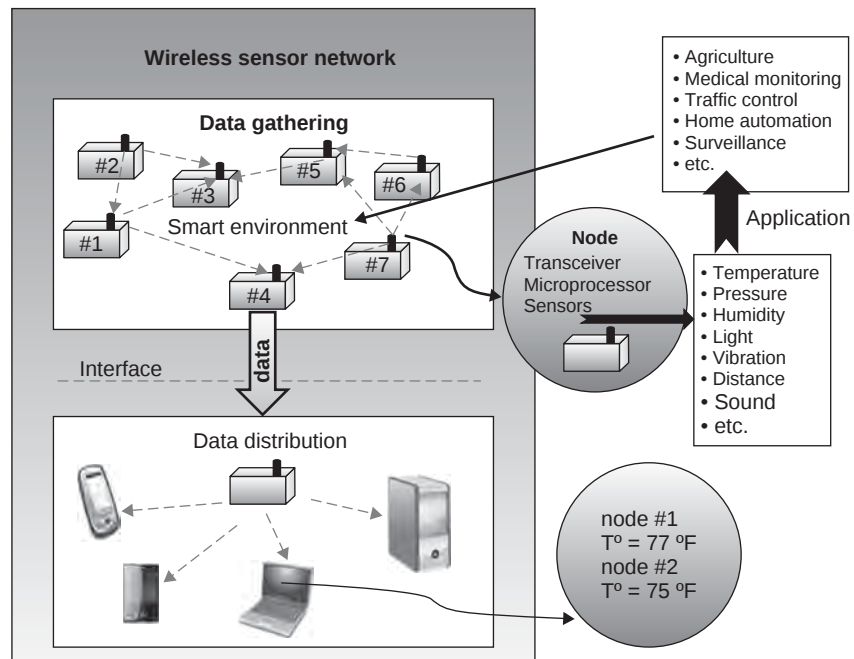


Figure 5. WSN [27].

The WSNs are composed of different spread nodes and each node is composed of sensors, microcontrollers, and radio emitters/receivers at the same time. This last point is where WSN and RFID converge to form a new technology. From this conjunction, the RFID tag is read, and the WSN uses the RFID tags as mobiles agents to enroute data. At the same time, the RFID tag uses the sensorial network to acquire data, it allows to identify and describe any specific

node in the network. A characteristic that enables the latter is that both WSN and, specifically, UHF RFID tags can work in the same ISM (*Industrial Scientific and Medical*) unlicensed frequency band [29], although WSN operates more commonly in 2.4 Ghz.

NFC (Near Field Communication). NFC is a standardized connectivity technology, wireless and short ranged, that enables

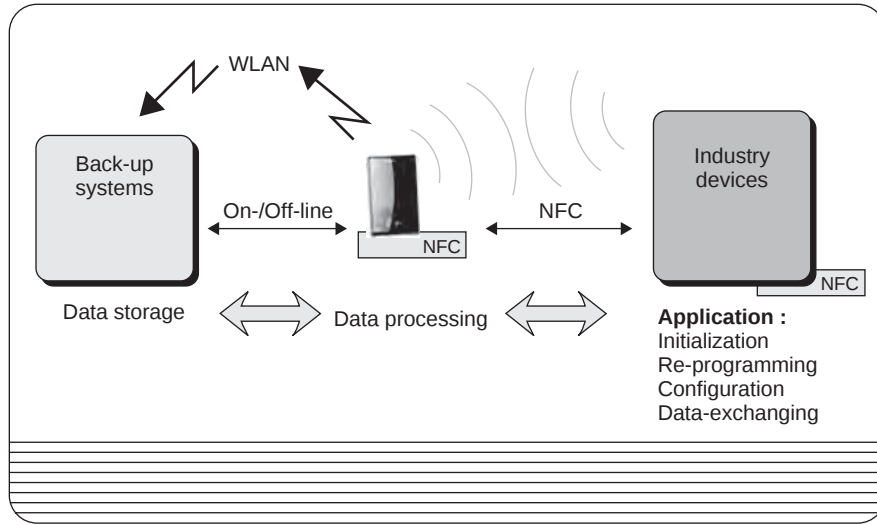


Figure 6. NFC system.

bidirectional interaction with electronic devices [30] in a simple and safety manner.

This technology has been developed from the combination of identification technologies (RFID) and contactless cards technology. NFC usually works in a frequency range of about 13.56 MHz, over an average distance of 1 dm. This technology is based on 18 092, 21 481, and ISO (*International Organization for Standardization*) standards; ECMA (European Association for Standardizing Information and Communication System) 340, 352 y 356 y ETSI TS 102, 190 [31].

NFC integrates both reading and writing components and a transponder in a single integrated circuit (chip). The different functions of this chip are controlled and triggered by means of software and combined with devices of individual use such as PDAs or mobile phones. The previous characteristic enables NFC devices perform as reading/writing or as an RFID tag. The main characteristic of the NFC is that it enables inter-connection between two available devices.

The benefits of this technology are found in the Mobile Systems:

- better use and a better user experience;
- easy access to services and physical objects content;
- access control;

- convenient exchange of digital items between devices;
- local payment capabilities and ticket issuing.

An NFC system involves different components divided mainly in three phases: data storage, processing, and application. (Fig. 6)

FFC (Far Field Communication). FFC is defined as the communication between the tag and the reader when takes place outside one full wavelength of range between them. This far field signal decays with the square of the distance from the antenna [32]. So, RFID systems that work on FFC (mainly UHF), have larger read range than those used in NFC (mainly LF and HF). Another characteristic of the FFC is that it is an electromagnetic wave radiation, rather than inductive as the NFC. The theoretical boundary between the two fields depends on the frequency used, and it is in fact directly proportional to $\lambda/2\pi$ where λ = wavelength [33]. This gives, for example, around 3.5 m for an HF system and 5 cm for UHF. NFC employs inductive coupling of the tag to the magnetic field circulating around the reader antenna (like a transformer), in its turn FFC uses similar techniques to radar (backscatter

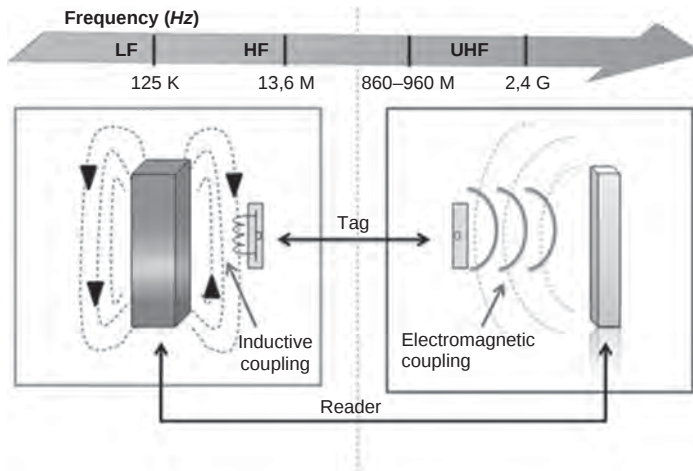


Figure 7. NFC-FFC energy and information transferring system [33].

Table 4. Near and Far Field Performance Distances

Band	Near Field Region	Far Field Region
LF	<120 m	>12 km
HF	<1 m	>110 m
UHF	<1.65 cm	>1.65 m
Microwave	<0.25 cm	>0.25 cm

reflection) by coupling with the electric field [33] (Fig. 7).

From Table 4 we can see the differences on performing distances between NFC and FFC depending on bandwidth.

Positioning

Positioning systems are a set of technologies whose main function is the location of objects in a given space. These systems use the identification and capturing layers already mentioned in the architecture framework, and provide data to the third layer of information exchange. We next detail two specific positioning systems for indoor and outdoor usage.

Wi-Fi RTLS (Real Time Location System). The RTLS is a derivation of WSNs. This technology is essentially a tracking system based on Wi-Fi 802.11(x) whose main objective is to locate objects within its network coverage, which is usually for indoor location. The main advantages are that Wi-Fi

network keeps data and voice transference in addition to location information; in fact, this technology can track any wireless device such as PDA's, laptops, mobiles phone, or scanners, without the need of a tag.

The applications of this system are vast: assets tracking, process visibility, security, and so on [34].

There are several methods used in location tracking [35], the most used by RTLS are: cell, triangulation, and RF trace.

GPS. The Global Positioning System (GPS) is a constellation of satellites designed originally for military navigation. However, it has grown into a wide gamma of commercial applications comprising people's movement, goods and information, community building, environmental administration, weather and disaster forecasting, and emergency responses.

GPS works through the signals that satellites send to earth, where they are detected by mobile or stationary receiving devices, and therefore functions in outdoor environments. These signals are used to determinate the receiver's position in the planet surface with an average accuracy of decimeters, through a triangulation system called *oversimplification*.

Combined with geometric technologies and integrated with the reference space system, GPS data can be used to locate and trace vehicles and other objects, manage infrastructures, show time information and

Table 5. Characteristics and Applications of Reviewed Technologies

Standard	Technology	Characteristics	Applications
GS1 128	Optical (barcoding)	“Line of sight” needed for reading ASCII coding characters High density of information Coding by levels	Goods ID Supply chain traceability
PDF-417	Optical (bidimensional coding)	1850 ASCII characters coding “Built-inredundancy” Capacity of storing graphics Coding by matrix	Transport labels Inventory management Identification cards
Data matrix (ECC 200)	Optical (bidimensional coding)	2335 alphanumeric characters Reduced size Squares modules placed within a pattern Algorithms error correction	Automobile, aerospace sectors Electronics Medical devices
RFID (LF, HF, UHF and MF)	Radiofrequency	No need of line of sight Updating data Frequency determinates range Passive and active tags (power supply)	Goods ID Persons ID Supply chain traceability
RuBee	Radiofrequency	Implementation of sensors Little information packages transmission (about 128 bytes) Low frequencies (inductive coupling) Low consumer No problems with metals or liquids	Security Medical devices Agriculture
Wireless Sensor Network (WSN)	Communication	Nodes spread (microcontroller, transceiver and sensors) Sensors variety	Smart environments (agriculture, security hospital)
NFC	Communication	Short range (~1 decimeter) 13.56 MHz frequency Enables interconnection between two available devices	Contactless cards PDAs or mobile phones
FFC	Communication	Signal decays with the square of the distance from the antenna	Goods ID
Wi-Fi RTLS (real time location system)	Positioning	Derivation of WSN Data and voice transference in addition of location information Dell, triangulation, and RF trace localization methods	Electronic device tracking (laptops, mobile phones, etc.) Process visibility Security

Table 5. (Continued)

Standard	Technology	Characteristics	Applications
GPs	Positioning	Signals sent from satellites Signals have no cost to commercial or civilian use worldwide	Vehicles tracking Environmental administration Supply chain
SCIS (supply chain information systems)	Information	Management information systems (MIS) Decision support systems (DSS) and expert systems Advanced planning systems (APS) Enterprise resource planning (ERP) Forecasting/demands planning systems (DPS)	Supply chain management

images, and navigate between different points of the earth globe [36].

Normally we can find a minimum of 24 satellites any time in any part of the globe. Signals have no cost to commercial or civilian use worldwide.

In the supply chain, the more GPS-relevant applications are as follows:

- fleet activity tracing and movements;
- transaction validation by location and activity and delivery checking;
- verification and tracing of assets location;
- vehicles navigation to improve routing time.

Information

According to Hoffer *et al.* [37], data is: *'The stored representation of objects and events which have a meaning and importance in the user environment.'* Information, at its turn, may also be defined as *'The data which has been processed such that the knowledge of the person that uses these types of data is increased'* [38].

Using the above, we can define an Information System as an arrangement of people, data, process, and information technologies that interact together to gather, process, store, and provide information required to support an organization [39]. Nowadays,

virtually all state-of-the-art concepts of logistic and supply chain operations use information system and hence information technologies, such as management information systems (MIS), decision support systems (DSS), and expert systems.

According to ITAA (*Information Technology Association of America*), information technology is the technology that allows the capacity of capturing, processing, storing, emitting, transmitting, and receiving data and information in an electronic way, including texts, graphics, sound, and video just as well as the ability of controlling all kinds of machines electronically.

These technologies applied to the supply chain comprise a wide range of solutions. Among these solutions, we can highlight the APS (advanced planning systems), the ERP (enterprise resource planning), and the DPS (forecasting/demands planning systems).

All these solutions manage data, and the way it obtains data is the relation they have with tracking systems; after all, the first layer of a tracking system is data capturing, or bringing real item information into an informatics system in order to complete management processes.

CONCLUSIONS

As described previously, numerous Auto-ID technologies are available to enable

traceability of objects in the supply chain. Each one of these technological options provides traceability only for certain objects and is typically suitable only in a specific setting. RFID, for example, exhibits a large potential within warehouses and stores, but is less suitable for tracking goods in transit. A compendium of previously described technologies can be seen in Table 5. To achieve complete or at least more extensive traceability in a supply chain, different Auto-ID technologies need to be integrated across multiple locations and parties. Establishing traceability involves a significant investment not only into hardware and software, but also into building up the knowledge, expertise, and capability to effectively utilize these systems. Very often, implementation of (individual or integrated) Auto-ID solutions exceeds the financial and technological resources of most companies. This is one of the reasons it is highly needed to understand these technologies' true capabilities and shortcomings, and in addition to have clear guidelines for the design of a feasible and efficient tracking system.

REFERENCES

- Golan E, Krissoff B, Kuchler F, *et al.* Traceability in the U.S. food supply: economic theory and industry studies. Agricultural Economic Report, no. 830. 2004.
- McFarlane D, Sarma S, Chirn JL, *et al.* Auto ID systems and intelligent manufacturing control. *Eng Appl Artif Intell* 2003;16(4):365–376.
- Zhang NS, He W, Tan PS. Understanding local pharmaceutical supply chain visibility. *SIMTech Tech Rep* 2008;9(4):234–239.
- Fontanella J. The supply chain management market sizing report, 2006–2011. AMR Research, Inc.; 2007. Available at www.amrresearch.com
- Tohamy N, Orlov LM, Herbert L. Supply chain visibility defined. Cambridge (MA): Forrester Research, Inc.; 2003. 24 April.
- EPCglobal. GS1 EPC global website. 2009. Available at www.epcglobalinc.org. Accessed 2010.
- Cheung SC, Chan WK, Lee PMK, *et al.* A combinatorial methodology for RFID benchmarking. 3rd RFID Academic Convocation; 2006 Oct 26–28; Shanghai: 2006.
- Rekik Y, Jemai Z, Sahin E, *et al.* Improving the performance of retail stores subject to execution errors: coordination versus RFID technology. *OR Spectr* 2007;29:597–626.
- Mitsugi J, Tokumasu O, Hada H. RF tag with RF and baseband communication interfaces for product lifecycle management. Auto-ID Labs White Paper; 2009. Available at www.autoidlabs.org.
- Fosso WS, Bendavid Y. Understanding the impact of emerging technologies on process optimization: the case of RFID technology. NSW/Montreal: Wollongong; 2009.
- Brown SA. Revolution at the checkout counter: the explosion of the bar code. Cambridge: Harvard University Press; 1997.
- Symbol Technologies. PSM20i Bar Code Scanning Module for Motorola iDEN Phones. 2004. (Online).
- Burke HE. Automating management information systems: principles of barcode applications. New York: Thomson Learning; 1989.
- Pibernik R, Blanco A, Monsreal M. Informe sobre la trazabilidad en el sector farmacéutico. Zaragoza, Spain: Zaragoza Logistics Center; 2007.
- Swartz WYPJ, McGlynn D. Postal applications of a high density bar code. US Patent 5243655. United States: Symbol Technologies, Inc.; 1993. Sep 1993.
- Wang YP. PDF417 specification. Boemia (NY): Symbol Technologies, Inc.; 1991.
- Association for Automatic Identification and Mobility. Available at www.aimglobal.org. Accessed 2010.
- GS1 Data Matrix. An introduction and technical overview. France: GS1; 2009.
- Joshi G. Information technology for retail: automatic identification & data capture systems. Oxford, England: Oxford University Press; 2009.
- Kamran A, Shah H, Kingston P. RFID applications: an introductory and exploratory study. *IJCSI Int J Comput Sci Issues* 2010;7(1):No. 3. 1–7.
- Ngai EWT, Moon KKL, Riggins FJ, *et al.* RFID research: an academic literature review and future research directions. *Int J Prod Econ* 2008;112(2):510–520.

22. Tajima M. Strategic value of RFID in supply chain management. *J Purch Supply Manage* 2007;13:261–273.
23. Engels DW, Sarma SE. Standardization requirements within the RFID class structure framework. MIT Auto-ID Labs Technical Report. 2005.
24. Wyld DC. RuBee: applying low-frequency technology for retail and medical uses. *Manage Res News* 2008;31(7):549–554.
25. Anonymous. RuBeeTM & Dot-TagTM Visibility Networks Presentation, ID World International Congress. 2007.
26. O'Connor MC. Visible assets promotes RuBee tags for tough-to-track goods. *RFID J* 2006. Available at www.rfidjournal.com/article/articleprint/2436/-1/1/.
27. Maxim Integrated Products. Available at www.maxim-ic.com. Accessed 2010.
28. Lewis FL. Wireless sensor networks, advanced controls, sensors and MEMS group automation and robotics research institute. Fort Worth (TX): The University of Texas. 2004.
29. Enache GPSGA. Wireless sensor networks: state of the art and future trends. Funchal, Madeira, Portugal: METROLOGIA E INOVAÇÃO; 2007.
30. Smartnfc. Available at www.smartnfc.com. Accessed 2010.
31. Anonymous. Available at www.nfc-global.com. Accessed 2009.
32. RFID Journal. Available at www.rfidjournal.com/glossary/73. Accessed 2009.
33. Lewis S. A basic introduction to RFID technology and its use in the supply chain. LARAN RFID; White Paper. 2004. Available at www.primtronix.com
34. Antti K. Wi-Fi based RFID and RTLS, enterprise applications and ROI Presentation. ID World Conference; Milan: 2007.
35. Cisco systems. Wi-Fi based real-time location tracking: solutions and technology. United States: Cisco systems, Inc.; White paper. 2006.
36. Forbes M. GPS Applications in Motion: Moving beyond automatic vehicle location to full enterprise integration Presentation. ID World Conference; 2007.
37. Hoffer JA, Prescott MB, McFadden FR. Modern database management. 7th ed. New Jersey: Prentice Hall; 2000
38. Pibernik R. Introduction to supply chain Information Systems. Spain: Zaragoza Logistics Center; 2004.
39. Whitten D. Ellis, and Casey, the impact of information technology outsourcing on firm profitability measures. *J Int Technol Inf Manage* 2002;11(2):9–17.

TRAFFIC NETWORK ANALYSIS AND DESIGN

STEPHEN D. BOYLES

Department of Civil and Architectural
Engineering, University of Wyoming,
Laramie, Wyoming

S. TRAVIS WALLER

Department of Civil, Architectural, and
Environmental Engineering, The
University of Texas, Austin, Texas

Networks are fundamental to the study of large-scale transportation models representing an entire metropolitan area, a state, or multistate regions. They can be applied in many contexts, including alternatives analysis, developing congestion pricing plans, identifying bottlenecks and critical infrastructure, shipping and freight logistics, multimodal planning, and disaster evacuation planning, to name only a few. The reason network models are so useful, and so broadly applicable, is because a mathematical network is a simple, compact, and flexible way to represent a large, complicated system.

This article provides an introduction to transportation network analysis, beginning with an overview of some basic network analysis problems, then describing different ways driver behavior can be represented on networks. Usually, the most important outputs of a network model are “link flows,” and the most important methods for finding these are provided next. Discussion then shifts to applications of network models, with emphases on congestion pricing and network design. The article concludes with a brief discussion of more advanced network models which may be suitable for particular applications, and a survey of important research works leading to the development of the ideas presented earlier.

Mathematically, a network consists of *links* and *nodes*. In transportation applications, a link usually represents a means of travel from one point to another: a road

segment between two intersections, a bus route between two stops, and so on, as seen in Fig. 1. The nodes, in turn, are the end points of the links. Quite often, nodes are adjacent to multiple links, so a node representing an intersection may adjoin multiple links representing road segments. Nodes and links may also be more abstract; for instance, nodes may represent “centroids” where trips begin and end, and links in a multimodal network might represent a transfer from one transport mode to another. The level of detail in a network varies from application to application. For multistate freight models, major highways may be the only links, and major cities the only nodes. For a city’s planning model, all major and minor arterials may be included as well. For a more detailed model, individual intersections may be “exploded” so that different links represent each turning movement (Fig. 2). A network is often compactly written as $G = (N, A)$ where N and A respectively represent the sets of nodes and links. A link is often written in terms of its end points, so (i, j) represents a link starting at node i and ending at node j .

TRANSPORTATION NETWORK PROBLEMS

The quintessential transportation network problem is *traffic assignment*, or translating demand for travel from all origins and destinations into *link flows* representing the number of travelers who use each link. This, in turn, can be used to create estimates of travel delay, air pollution, revenue on toll roads, and other factors which transportation planners consider when evaluating an alternative or a new policy. To represent travel demand, let $D \subseteq N \times N$ be a subset of ordered pairs (r, s) representing every combination of an origin r and destination s where there is demand for travel. We write the travel demand as d^{rs} for every $(r, s) \in D$.

Someone traveling from origin r to destination s will follow a sequence of adjacent

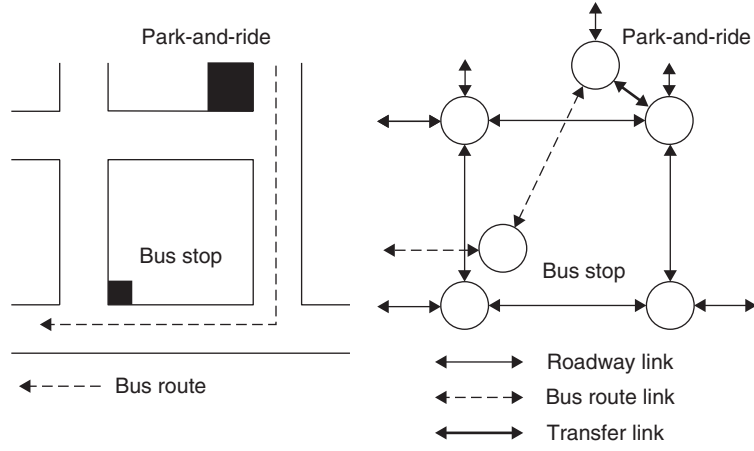


Figure 1. Physical transportation infrastructure (left) and network representation (right).

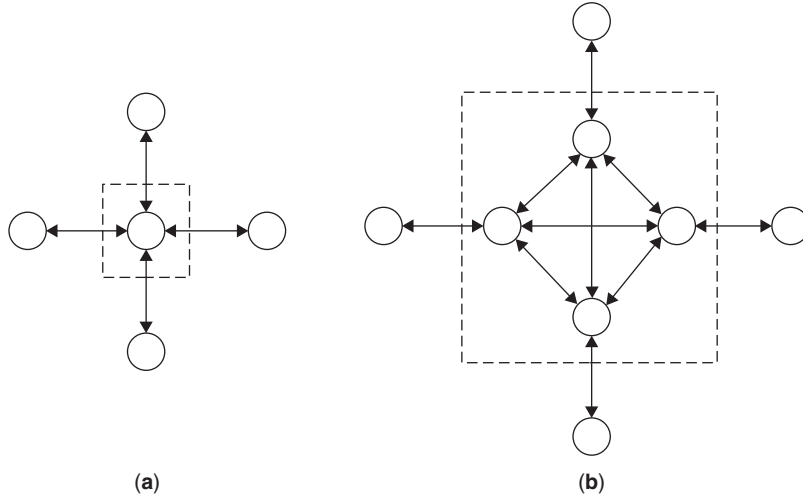


Figure 2. Two representations of the same intersection, with turning movements consolidated (a) and separated (b).

links $\{(r, i_1), (i_1, i_2), \dots, (i_k, s)\}$, called a *path*, and we write $(i, j) \in \pi$ to indicate that link (i, j) is part of path π . We write Π^{rs} to represent the set of all paths connecting origin r to destination s . The number of paths connecting an origin and destination can be very large, and grow exponentially with the size of the network. The good news is that we almost never need to list all of the elements of Π^{rs} in order to perform traffic assignment: the methods described later in this article

either generate only a small subset of all of the possible paths or work entirely with link flows, obviating the need to find *all* of the paths.

The goal of traffic assignment is to determine which links travelers will use, that is, to map the travel demands d^{rs} to the link flows x_{ij} . To do this, we need to specify two things: first, how travelers choose a path from an origin to a destination; and second, how travelers interact with each other. A common

assumption, described in more detail below, is that travelers want to choose a path which minimizes their travel time—that is, travel is not desirable in and of itself, but simply a means to move to a destination in order to conduct some more worthwhile activity. Because of congestion effects, however, the fastest path depends on the choices made by all of the other travelers. (A freeway may be very fast if few others choose to use it, but very slow if many others do.) This is commonly represented by a *link performance function* $t_{ij}(x_{ij})$ expressing the delay on link (i, j) as a function of the link flow x_{ij} . One common link performance function was developed by the Bureau of Public Roads, and is of the form

$$t_{ij}(x_{ij}) = t_{ij}^0 \left(1 + \alpha \left(\frac{x_{ij}}{c_{ij}} \right)^\beta \right),$$

where t_{ij}^0 is the free-flow travel time, c_{ij} the practical capacity,¹ and α and β shape parameters which can be calibrated. In the absence of field data, $\alpha = 0.15$ and $\beta = 4$ are common choices.

Traffic assignment is often used by itself, as planners seek to predict the impacts of various alternatives, such as construction of a new freeway or expanding capacity on existing roadways. However, it can also be used as a component of more sophisticated network models. For instance, tolls influence travelers route choices, and traffic assignment can be used to study the impact of different toll prices and identify the strategy providing maximum benefits to the region. Traffic assignment models can also be used in a prescriptive sense, rather than a descriptive one: instead of simply being used to evaluate a given set of alternatives, *network design models* aim to generate one or more alternatives which, if implemented, would provide the maximum benefit.

USER OPTIMAL AND SYSTEM OPTIMAL

The two most common rules for assigning travel demand to roadway links are to ensure that all trips follow shortest paths from the origin to the destination (called *user optimality*), or to assign demand so as to minimize the total travel time experienced by all travelers (called *system optimality*). User optimality can also be described as assigning demand in such a way that no single traveler can reduce his or her travel time by switching routes. User optimality is thought to be a better description of real traveler behavior, because travelers do not generally choose their route to minimize total congestion (nor would they know how to, even if so inclined). System optimality, while less realistic, is useful as a lower bound on congestion, for certain logistics problems, and, as described later, is central to certain congestion pricing problems. Although these rules are similar, in that assigning demand to shortest paths usually creates a low total congestion level, exceptions exist.

The most notable is the Braess paradox [2]. Consider the network in Fig. 3, where two routes connect a single origin and destination with $d = 6$, and the link performance functions are indicated on the figure.² By symmetry, the user optimal and system optimal flows are $x_{12} = x_{24} = x_{13} = x_{34} = 3$. These are indeed user optimal, because both routes have an equal and minimal travel time of 83 minutes, so no traveler has an incentive to change their decision. One can also verify that these flows are system optimal and that the total travel time is 498 vehicle-minutes, using the techniques mentioned later in the article.

Now, a new link (2, 3) is added to the network, with link performance function $t_{23}(x_{23}) = 10 + x_{23}$. Clearly, the new system optimal flows can have a minimal travel time no greater than 498 min, because the old flows still represent a valid traffic

¹The practical capacity roughly corresponds to the flow at LOS C, and is about 80% of the “true” capacity. See [1] for additional details.

²One may presume that the demand and flows are measured in some realistic units, such as hundreds or thousands of vehicles, rather than representing only six vehicles in total.

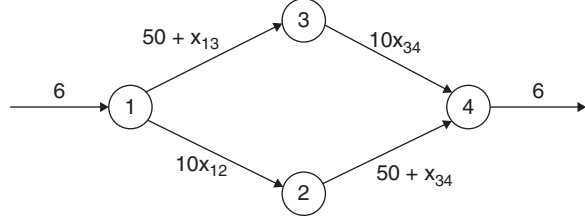


Figure 3. Network illustrating the Braess paradox.

assignment. However, the old flows are no longer user optimal: the new path has a travel time of 70 min, which is less than the other, used routes. Thus, travelers will shift onto the new path, and the new user optimum will result in equal usage of the three possible paths, resulting in link flows $x_{12} = x_{34} = 4$ and $x_{13} = x_{23} = x_{24} = 2$ and a travel time of 92 min on each path.

This should be surprising: by adding the link (2,3) the user optimal travel times *increased* for all travelers, with the total travel time increasing to 552 vehicle-min. Thus, adding capacity to the network actually degraded its performance! Whether this paradox is realistic, or just an artifact of modeling assumptions, has been debated for decades, and a full treatment is beyond the scope of this article. Still, the fact that it occurs at all shows that network models can behave counterintuitively, and justifies the existence of mathematically rigorous models and algorithms—engineering judgment alone can lead to highly faulty conclusions.

User optimal and system optimal link flows can also be represented as the solution to nonlinear optimization problems. For notational convenience, these optimization problems are expressed both in terms of the vector of path flows $\mathbf{h}$ and the vector of link flows $\mathbf{x}$, which are related by a constraint forcing consistency between these. User optimal link flows solve the program

$$\min_{\mathbf{x}, \mathbf{h}} \sum_{(i,j) \in A} \int_0^{x_{ij}} t_{ij}(x) dx \quad (1)$$

$$\text{s.t. } x_{ij} = \sum_{(r,s) \in D} \sum_{\pi \in \Pi^{rs}: (i,j) \in \pi} h^\pi \quad \forall (i,j) \in A \quad (2)$$

$$d^{rs} = \sum_{\pi \in \Pi^{rs}} h^\pi \quad \forall (r,s) \in D \quad (3)$$

$$h^\pi \geq 0 \quad \forall \pi \in \bigcup_{(r,s) \in D} \Pi^{rs} \quad (4)$$

while system optimal link flows solve

$$\min_{\mathbf{x}, \mathbf{h}} \sum_{(i,j) \in A} t_{ij}(x_{ij}) x_{ij} \quad (5)$$

$$\text{s.t. } x_{ij} = \sum_{(r,s) \in D} \sum_{\pi \in \Pi^{rs}: (i,j) \in \pi} h^\pi \quad \forall (i,j) \in A \quad (6)$$

$$d^{rs} = \sum_{\pi \in \Pi^{rs}} h^\pi \quad \forall (r,s) \in D \quad (7)$$

$$h^\pi \geq 0 \quad \forall \pi \in \bigcup_{(r,s) \in D} \Pi^{rs} \quad (8)$$

Notice that the constraints are the same for user optimal and system optimal assignment, reflecting the fact that a feasible solution to either problem is simply a way of assigning travel demand to the available paths and links: ensuring consistency between link flows and path flows with Equation (2) or (6); ensuring all demand is assigned with Equation (3) or (7); and ensuring nonnegativity of path flows with Equation (4) or Equation (8). The only difference between these two problems is their objective functions; that is, the decision rule for choosing one feasible traffic assignment over another.

It should be clear that the objective function (5) corresponds to system optimal flows; that Equation (1) corresponds to user optimal flows is less obvious. This can be seen by Lagrangianizing the constraints (3), introducing nonnegative dual variables κ , and writing the Karush–Kuhn–Tucker conditions for the resulting optimization

problem. After some simplification, these conditions include

$$\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) \geq \kappa^{rs} \quad \forall (r,s) \in D, \pi \in \Pi^{rs} \quad (9)$$

$$h^\pi \left(\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) - \kappa^{rs} \right) = 0 \quad \forall (r,s) \in D, \pi \in \Pi^{rs}. \quad (10)$$

Condition (9) implies $\kappa^{rs} = \min_{\pi \in \Pi^{rs}} \sum_{(i,j) \in \pi} t_{ij}(x_{ij})$, showing that κ^{rs} represents the travel time on the shortest path connecting origin r to destination s . Then, to satisfy the complementarity constraint (10), we either must have $\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) = \kappa^{rs}$ (path π is a shortest path between r and s) or $h^\pi = 0$ (the path is unused). This is exactly the definition of user optimal flows.

Variational inequalities provide an alternate way of characterizing user optimal and system optimal flows. A vector of feasible path flows $\mathbf{h}^*$ is user optimal if and only if

$$\mathbf{F}(\mathbf{h}^*) \cdot (\mathbf{h} - \mathbf{h}^*) \geq 0 \quad (11)$$

for all feasible path flows $\mathbf{h}$, where $\mathbf{F}(\mathbf{h}^*)$ is the vector of path travel times.³ To understand this, notice that the condition can be rewritten as $\mathbf{F}(\mathbf{h}^*) \cdot \mathbf{h}^* \leq \mathbf{F}(\mathbf{h}^*) \cdot \mathbf{h}$ for all path flows $\mathbf{h}$. That is, no other path flow could reduce total system travel time *if path costs are fixed at their current levels*. If this were not true, then certainly some traveler could switch paths and reduce his or her travel time, so $\mathbf{h}^*$ cannot be user optimal. A similar variational inequality can be written for system optimal flows, replacing the path travel times with the marginal path travel times defined in the section titled “Congestion Pricing.”

At times, it is useful to express this variational inequality in terms of link flows, rather than path flows, as

$$\mathbf{t}(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \geq 0 \quad (12)$$

for all feasible link flows $\mathbf{x}$, where $\mathbf{t}$ is the vector of link travel times. This can be derived

either as a transformation from Equation (11), or from the optimality conditions for the nonlinear program given by Equations (1)–(4). Variational inequality formulations are more flexible than nonlinear programming formulations. For instance, they can represent cases where the travel time on a link depends on the flow on other links as well. On the other hand, specialized algorithms for solving nonlinear programming formulations are usually faster than variational inequality-based formulations.

SOLVING FOR LINK FLOWS

Several approaches exist for finding user-optimal or system-optimal link flows, but all of them broadly follow the same iterative pattern, starting with any feasible assignment:

Path Loading. Given a set of feasible flows, calculate the travel times on each link.

Path Finding. Identify the optimal paths with respect to these travel times (shortest paths for user-optimal, and least marginal-cost paths, for system-optimal).

Path Adjusting. Shift some travelers from suboptimal paths onto optimal paths.

In this framework, the only difference between the user-optimal and system-optimal problems lies in the second step. Because system-optimal flows use least marginal cost paths, while user-optimal flows use least cost paths, the two assignment rules are fundamentally similar, and without loss of generality we only present algorithms for the user-optimal case. The third step requires the most care, because shifting too many travelers results in “over-correction” which does not converge toward equilibrium.

It is in this third step that most traffic assignment methods differ, in addition to the data structures used to represent the links or paths used by travelers at intermediate steps. Broadly speaking, solution methods can be classified as *link-based*, *path-based*, and *origin-based*.

Link-based methods do not actually store which paths travelers are using, and only

³That is, $F_\pi(\mathbf{h}^*) = \sum_{(i,j) \in \pi} t_{ij} \left(\sum_{\pi \ni (i,j)} h^\pi \right)$.

work with the aggregate link flows $\mathbf{x}$. Because of this, these algorithms are highly efficient in terms of computer memory, and were among the first practical traffic assignments developed. However, they are highly *inefficient* in terms of computation time, and, after making good progress in early iterations, usually bog down shortly thereafter and converge very slowly toward the final equilibrium.

The quintessential link-based method is the Frank–Wolfe algorithm, for which many variations and improvements have been proposed. The basic algorithm accomplishes path adjustment by first finding the optimal *all-or-nothing* link flow assignment with respect to the current travel times (that is, the link flows which would result from *all* travelers shifting onto the shortest path), and proportionately shifting link flows to make the current feasible assignment more closely resemble the all-or-nothing assignment. Mathematically, given a current feasible assignment $\mathbf{x}^0$ and an all-or-nothing assignment $\mathbf{x}^*$, the Frank–Wolfe algorithm seeks a flow on the line connecting these points, that is, a flow $(1 - \lambda)\mathbf{x}^0 + \lambda\mathbf{x}^*$ for $\lambda \in [0, 1]$, with λ chosen to minimize Equation (1) by a numerical line search method [3]. Thus, the Frank–Wolfe method can be summarized as

1. Start with feasible link flows $\mathbf{x}^0$, and set the iteration counter $k \leftarrow 0$.
2. Find the shortest paths between each origin and destination, with travel times $\mathbf{t}(\mathbf{x}^k)$.
3. Assign all trips to these shortest paths, obtaining an all-or-nothing flow vector $\mathbf{x}^*$.
4. Set $\mathbf{x}^{k+1} \leftarrow (1 - \lambda)\mathbf{x}^k + \lambda\mathbf{x}^*$, with $\lambda \in \arg \min_{\lambda \in [0, 1]} \sum_{(i, j) \in A} \int_0^{(1 - \lambda)x_{ij}^k + \lambda x_{ij}^*} t_{ij}(x) dx$.
5. Check convergence criterion. If satisfied, terminate. Otherwise, increment the iteration counter $k \leftarrow k + 1$ and return to Step 2.

Several convergence criteria are possible. The simplest option is to terminate the algorithm when the link flows change by a sufficiently small amount, that is, when $\|\mathbf{x}^{k+1} - \mathbf{x}^k\| < \epsilon_1$ for some predetermined ϵ_1 . This criterion is not recommended, as

it is unable to distinguish between true convergence and a scenario where a poorly-designed algorithm simply cannot find a good improvement over a nonequilibrium solution. A more theoretically sound option is to check consistency with the equilibrium principle itself. At equilibrium, all travelers are on shortest paths, so $\mathbf{t}(\mathbf{x}^*) \cdot \mathbf{x}^* = \mathbf{d} \cdot \boldsymbol{\kappa}$, with $\mathbf{d}$ and $\boldsymbol{\kappa}$ the vectors of travel demand and shortest path costs, respectively. Thus, one measure of equilibrium is the *relative gap*, defined by

$$\gamma = \frac{\mathbf{t}(\mathbf{x}) \cdot \mathbf{x}}{\mathbf{d} \cdot \boldsymbol{\kappa}} - 1$$

Note that this is zero if and only if all travelers are on shortest paths, that is, if and only if $\mathbf{x}$ is user optimal. Other definitions of relative gap exist in the literature and in commercial software, often involving the value of the objective function (1), but all share the common feature reaching zero only if equilibrium is reached.

As mentioned earlier, link-based methods are typically slow to converge. Better convergence can be seen with a *path-based* method, which tracks the paths used by individual origins and destinations, not just the aggregate link flows. Two notable path-based methods are *disaggregate simplicial decomposition* and *gradient projection*. The latter is simpler to explain, and functions as follows: the shortest path π_{rs}^* connecting each origin and destination $(r, s) \in D$ is identified, and declared the “basic” path, whose flow is defined as $h^{\pi_{rs}^*} = d_{rs} - \sum_{\pi \in \Pi^{rs}: \pi \neq \pi_{rs}^*} h^\pi$. Thus, the demand satisfaction constraint can simply be rewritten as $h^{\pi_{rs}^*} \geq 0$, so all of the constraints are captured through nonnegativity ($\mathbf{h} \geq 0$) and the only decision variables are the nonbasic path flows. With these decision variables, a descent direction for the path flows for origin-destination pair (r, s) can be identified as

$$\left[\kappa_{rs} - \sum_{(i, j) \in \pi_{rs}^*} t_{ij}(\mathbf{h}), \dots, \kappa_{rs} - \sum_{(i, j) \in \pi_{rs}^*} t_{ij}(\mathbf{h}) \right]$$

scaled by an appropriate factor. If this descent direction results in a step which

leaves the feasible region, a projection is performed to obtain the closest feasible point. Because the only constraints are nonnegativity, a projection is trivial, and simply consists of identifying a path with negative flow, and setting it to zero.

The steps of this algorithm can be summarized as

1. Start with feasible path flows $\mathbf{h}$ and a set of working paths $\hat{\Pi}^{rs} \leftarrow \{\pi \in \Pi^{rs} : h^\pi > 0\}$.
2. Find the shortest path π_{rs}^* between each origin and destination, with travel time κ_{rs} , and update the set of working paths $\hat{\Pi}^{rs} \leftarrow \hat{\Pi}^{rs} \cup \pi_{rs}^*$.
3. For each origin and destination (r, s) , and for each nonbasic path $\pi \in \hat{\Pi}^{rs} \setminus \pi_{rs}^*$ calculate the projected path flows $h^\pi \leftarrow \max \left\{ 0, h^\pi - \lambda_\pi \left(\sum_{(i,j) \in \pi} t_{ij}(\mathbf{h}) - \kappa_{rs} \right) \right\}$.
4. Update the basic path flows $h_{rs}^{\pi^*} = d_{rs} - \sum_{\pi \in \Pi^{rs} : \pi \neq \pi_{rs}^*} h^\pi$.
5. Check convergence criterion. If satisfied, terminate. Otherwise, return to Step 2.

Jayakrishnan *et al.* [4] report that a good step length is

$$\lambda_\pi = \left(\sum_{(i,j) \in (\pi \cup \pi^*) \cap (\pi \cap \pi^*)^C} \frac{dt_{ij}}{dx_{ij}}(\mathbf{h}) \right)^{-1}$$

Most recently, a class of *origin-based* (or *bush-based*) algorithms has arisen which seems to offer superior performance to both link-based and path-based methods in terms of convergence speed, with memory requirements in between the two. A *bush* (or *restricting subnetwork*) B_r , defined for each origin r , consists of all the arcs which travelers departing from that origin might plausibly use. Different algorithms use different definitions of “plausible,” but the general concept is that if an arc (i, j) is part of B_r , node j should be “farther away” from r than node i in some sense. This implies that B_r is acyclic, which makes it very easy to find shortest paths (and longest paths)

by scanning nodes in topological order. For each bush arc $(i, j) \in B_r$, an origin-based link flow x_{ij}^r is maintained, showing the number of travelers on (i, j) who departed from r . Clearly, $x_{ij} = \sum_r x_{ij}^r$.

The two steps which are iteratively performed are (i) bush equilibrating and (ii) bush updating. In the first step, the origin-based link flows x_{ij}^r are chosen so that they form an equilibrium on the bush B_r ; in the second, the bushes are updated, removing links which are unused, and adding new links which are now “plausible” for the origin r . Bar-Gera [5] uses a gradient projection approach, while Dial [6] shifts flow from longer paths to shorter ones; preliminary results seem to indicate that the latter approach is faster.

CONGESTION PRICING

That user-optimal and system-optimal traffic assignment are distinct suggests that effective transportation policy can reduce congestion and delay. Economically speaking, this distinction arises from congestion externalities—when a traveler drives on a congested link, he or she experiences delay equal to the travel time. However, this traveler’s presence adds to the congestion, causing a slight increase in delay to all of the other travelers behind him or her. The driver does *not* experience these costs and therefore, has no incentive to account for them. For system optimal link flows to result, these costs must be paid by the driver in some way.

In theory, congestion pricing can accomplish exactly this. Recall that optimality conditions for user-optimal assignment were

$$\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) \geq \kappa^{rs} \quad \forall (r, s) \in D, \pi \in \Pi^{rs} \quad (13)$$

$$h^\pi \left(\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) - \kappa^{rs} \right) = 0 \quad \forall (r, s) \in D, \pi \in \Pi^{rs} \quad (14)$$

Similar conditions can be written for the system-optimal problem:

$$\sum_{(i,j) \in \pi} \left[t_{ij}(x_{ij}) + x_{ij} \frac{dt_{ij}}{dx_{ij}}(x_{ij}) \right] \geq \kappa^{rs} \quad \forall (r,s) \in D, \pi \in \Pi^{rs}, \quad (15)$$

$$h^\pi \left(\sum_{(i,j) \in \pi} \left[t_{ij}(x_{ij}) + x_{ij} \frac{dt_{ij}}{dx_{ij}}(x_{ij}) \right] - \kappa^{rs} \right) = 0 \quad \forall (r,s) \in D, \pi \in \Pi^{rs}. \quad (16)$$

This implies that if a suitable toll, equal to $x_{ij} \frac{dt_{ij}}{dx_{ij}}(x_{ij})$, were added to each link, user-optimal flows would be identical to system-optimal flows. Note that this toll is in units of time; to convert to monetary units, we should multiply by a value of travel time (\$10/h is typical).

This toll has an intuitive explanation in terms of the congestion externalities mentioned above. The toll $x_{ij} \frac{dt_{ij}}{dx_{ij}}(x_{ij})$ is the product of the number of travelers on the link, and the marginal increase in travel time caused by another vehicle entering. Thus, charging this toll effectively eliminates this externality, bringing user optimal flows into alignment with system optimal ones.

This is called *first-best pricing* because it represents an idealized situation where a toll can be charged on every link (whether a freeway or a local street), without any regard to a maximum toll value. More realistic *second-best pricing* strategies take these practical limitations into account, but lack the simplicity and elegance of first-best pricing. An overview of second-best pricing strategies can be found in Small and Verhoef [7].

NETWORK DESIGN

Network design builds naturally on traffic assignment modeling. The network models described thus far treat the transport infrastructure as fixed, and the goal is to find path or link flows representing a certain behavior. Network design, by contrast, views the behavioral rule as fixed, and seeks a way of constructing or improving the infrastructure so as to maximize societal welfare. For

instance, one simple network design model is to make capacity improvements on certain links in order to decrease total system travel time, subject to a budget constraint. This can be formulated as

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{h}, \kappa, \mathbf{y}} \quad & \sum_{(i,j) \in A} x_{ij} t_{ij}(x_{ij}, y_{ij}) \\ \text{s.t.} \quad & \sum_{(i,j) \in A} y_{ij} \leq B \\ & x_{ij} = \sum_{\pi \ni (i,j)} h^\pi \quad \forall (i,j) \in A \\ & \sum_{(i,j) \in \pi} t_{ij}(x_{ij}, y_{ij}) - \kappa^{rs} \geq 0 \quad \forall (r,s) \in D, \pi \in \Pi^{rs} \\ & h^\pi \left(\sum_{(i,j) \in \pi} t_{ij}(x_{ij}, y_{ij}) \kappa^{rs} \right) = 0 \quad \forall (r,s) \in D, \pi \in \Pi^{rs} \\ & \sum_{\pi \in \Pi^{rs}} h^\pi = d_{rs} \quad \forall (r,s) \in D \\ & h^\pi \geq 0 \quad \forall (r,s) \in D, \pi \in \Pi^{rs}. \end{aligned}$$

Here $\mathbf{y}$ represents the amount of money spent improving each link, and B the total budget available. Notice that the link travel times t_{ij} now depend both on the flow x_{ij} as well as the amount spent y_{ij} (perhaps representing increased capacity); and also notice that the user optimal first-order conditions are included as constraints.

As formulated, this is a mathematical program with equilibrium constraints, a class of optimization problems known to be difficult to solve due to nonconvexity of the feasible region. This can also be reformulated as a bilevel optimization problem, which again is nontrivial to solve. Thus, unlike the traffic assignment models described above, one cannot generally hope to solve network design problems exactly, let alone with an efficient algorithm. Instead, solution techniques for network design problems are heuristic in nature.

ADVANCED NETWORK MODELS

Although the models presented earlier in this article demonstrate the fundamental principles behind network models, they often lack realism in several important ways. Over the past decades, researchers have developed a number of more sophisticated extensions to these models, building on similar principles but relaxing simplifying assumptions. The remainder of this article provides a brief overview of these more advanced models. Stochastic user equilibrium relaxes the assumption that users accurately perceive travel times; elastic demand relaxes the assumption that travel demand is independent of travel cost; uncertain travel times relax the assumption that link performance functions are the same at all times; and dynamic traffic assignment relaxes the assumption of steady-state congestion and travel demand.

Stochastic User Equilibrium

One assumption made above is that users correctly perceive travel times, with unlimited precision. That is, given two routes, one with a travel time of exactly 11 min, and another with a travel time of 11 min and one sec, every traveler would choose the former. In reality, most travelers would be unable to distinguish between these trip lengths, and even if they could, would likely be indifferent between the two. The stochastic equilibrium framework accounts for this by introducing a random error term $\tilde{\epsilon}$ to users' perception of travel time. That is, the travel time $\tilde{u}_i^\pi$ user i perceives on path π is related to the true travel time $t_\pi(\mathbf{x})$ by

$$\tilde{u}_i^\pi(\mathbf{x}) = t_\pi(\mathbf{x}) + \tilde{\epsilon}_i^\pi.$$

Traveler i then picks the path available to him or her with minimal perceived travel time; this may or may not be the path with the lowest actual travel time. Typically, one assumes that these perceptions are unbiased, that is, that the expected value of $\tilde{\epsilon}$ is zero. If one further assumes that the $\tilde{\epsilon}$ are independent, identically distributed Gumbel random variables, this results in a logit route choice model, where the probability that traveler

i (traveling from origin r to destination s) chooses path π is given by

$$\Pr(\tilde{u}_i^\pi = \min_{\rho \in \Pi^{rs}} \{\tilde{u}_i^\rho\}) = \frac{e^{-\theta t_\pi}}{\sum_{\rho \in \Pi^{rs}} e^{-\theta t_\rho}}.$$

The parameter θ can be seen as the degree of accuracy in users' perceptions. If $\theta = 0$, users' perceptions are completely inaccurate, so every path is equally likely to be chosen regardless of its travel time. On the other hand, as $\theta \rightarrow \infty$, the user will almost certainly choose the path whose actual travel time is minimal. Thus, if there are d^{rs} travelers departing origin r for destination s , the path flows are given by

$$h^\pi = d^{rs} \frac{e^{-\theta t_\pi}}{\sum_{\rho \in \Pi^{rs}} e^{-\theta t_\rho}}$$

One can show that, under stochastic user equilibrium, the optimal link flows satisfy

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{h}} \quad & \sum_{(i,j) \in A} \left(x_{ij} t_{ij}(x_{ij}) - \int_0^{x_{ij}} t_{ij}(x) dx \right) \\ & - \sum_{(r,s) \in D} d_{rs} E \left[\min_{\pi \in \Pi^{rs}} \{\tilde{u}_i^\pi(\mathbf{x})\} \right] \\ \text{s.t.} \quad & x_{ij} = \sum_{(r,s) \in D} \sum_{\pi \in \Pi^{rs}: (i,j) \in \pi} h^\pi \quad \forall (i,j) \in A \\ & d^{rs} = \sum_{\pi \in \Pi^{rs}} h^\pi \quad \forall (r,s) \in D \\ & h^\pi \geq 0 \quad \forall \pi \in \bigcup_{(r,s) \in D} \Pi^{rs}. \end{aligned}$$

Returning to the original example at hand, if two paths have a very similar travel time, to the extent that the perception error $\tilde{\epsilon}$ is substantially larger than the difference in true travel times, the stochastic user equilibrium model predicts nearly equal usages for the two paths, which is in accordance with intuition. Further details on stochastic user equilibrium, including proof that the above nonlinear program correctly assigns link flows and discussion on some limitations, can be found in Sheffi [8].

Elastic Demand

The demand for travel is not independent of travel costs—if congestion worsens, or if

an onerous toll is levied, travel demand will drop as users switch modes or activities, or simply reduce the number of optional trips made. A general way to represent this elasticity is to make the travel demand vector $\mathbf{d}$ a function of the trip travel times and tolls (simultaneously represented with the generalized cost κ), that is, $d_{rs} = S(\kappa_{rs})$ for some strictly decreasing (and thus invertible) function S . In this context, we seek a traffic assignment which forms a “fixed point” with respect to the demand functions and cost functions. That is, user path choices determine the travel costs between origins and destinations, which determine the demand for travel, which in turn determine the path choices. A fixed point of user path choices is consistent in this manner.

One way to formulate the user optimal problem with elastic demand is with the non-linear program

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{h}} \quad & \sum_{(i,j) \in A} \int_0^{x_{ij}} t_{ij}(x) dx \\ & - \sum_{(r,s) \in D} \int_0^{d_{rs}} S_{rs}^{-1}(s) ds \\ \text{s.t.} \quad & x_{ij} = \sum_{(r,s) \in D} \sum_{\pi \in \Pi^{rs}: (i,j) \in \pi} h^\pi \quad \forall (i,j) \in A \\ & d_{rs} = \sum_{\pi \in \Pi^{rs}} h^\pi \quad \forall (r,s) \in D \\ & h^\pi \geq 0 \quad \forall \pi \in \bigcup_{(r,s) \in D} \Pi^{rs} \\ & d_{rs} \geq 0 \quad \forall (r,s) \in D \end{aligned} \quad (17)$$

$$d_{rs} = \sum_{\pi \in \Pi^{rs}} h^\pi \quad \forall (r,s) \in D \quad (18)$$

$$h^\pi \geq 0 \quad \forall \pi \in \bigcup_{(r,s) \in D} \Pi^{rs} \quad (19)$$

$$d_{rs} \geq 0 \quad \forall (r,s) \in D \quad (20)$$

To verify this, the first-order conditions include

$$\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) \geq \kappa^{rs} \quad \forall (r,s) \in D, \pi \in \Pi^{rs} \quad (21)$$

$$\begin{aligned} h^\pi \left(\sum_{(i,j) \in \pi} t_{ij}(x_{ij}) - \kappa^{rs} \right) &= 0 \\ \forall (r,s) \in D, \pi \in \Pi^{rs} \end{aligned} \quad (22)$$

$$\kappa_{rs} \geq S_{rs}^{-1}(d_{rs}) \quad \forall (r,s) \in D \quad (23)$$

$$d_{rs}(\kappa_{rs} - S_{rs}^{-1}(d_{rs})) = 0 \quad \forall (r,s) \in D. \quad (24)$$

As before, conditions (22) and (23) enforce user optimality. The remaining conditions enforce that the travel demand is correct, given the shortest path costs between each origin and destination. To see this, note that Equation (25) requires $\kappa_{rs} = S_{rs}^{-1}(d_{rs})$ if $d_{rs} > 0$, that is, if there is positive demand from an origin to a destination, we must have $d_{rs} = S(\kappa_{rs})$. On the other hand, if the demand is zero, condition (24) allows the possibility of $d_{rs} > S(\kappa_{rs})$, as would happen if $S(\kappa_{rs}) < 0$.

Often, the elastic demand problem can be solved by introducing artificial “dummy arcs” to the network representing unsatisfied demand. If an upper bound $\bar{d}_{rs}$ on the demand between each origin and destination is defined, one can create new arcs connecting each origin to each destination, with a cost function $t_{rs} = S_{rs}^{-1}(\bar{d}_{rs} - x_{rs})$. Solving a fixed demand problem on this network (with demand for travel from each origin to each destination set at the upper bound) now produces an answer to the elastic demand problem: at equilibrium, the direct arc connecting origin r to destination s has the cost κ_{rs} (as does every other used path) and thus its flow is $x_{rs} = \bar{d}_{rs} - S(\kappa_{rs})$, which is exactly the unsatisfied demand.

Uncertain Costs

The assumption that travel times can be known with certainty limits network modelers’ ability to consider events such as incidents or poor weather, as well as their ability to evaluate innovative technologies such as traveler information systems (which are only valuable because conditions cannot be accurately predicted in advance). Furthermore, when users receive information *en route*, they may choose to change their path. This leads to the concept of *on-line routing*, where the full travel path is unknown at the start of the trip, and depends on the information received along the way. The decision rule which associates a particular path with information is known as a *policy* or *hyperpath*, and can be written as $\pi_i(\theta)$ to represent the choice made when information θ is provided at node i . This description is somewhat vague in recognition of the diversity of approaches researchers have brought to modeling information and

uncertainty; more rigorous derivation for specific problem assumptions can be found in the references that follow.

An on-line shortest path algorithm is needed to find the best policy, just as a shortest path algorithm can be used to find the best path when travel costs are fixed. A typical approach involves dynamic programming, solving a recursion of the form

$$\begin{aligned} v_s &= 0 \\ v_i &= E_\theta[\min_{(i,j) \in A} \{t_{ij} + v_j\} | \theta] \\ \pi_i(\theta) &= \arg \min_{(i,j) \in A} \{E[t_{ij} + v_j | \theta]\}, \end{aligned}$$

for an origin r and destination s where v_i represents the expected cost of the best routing policy from node i to the destination. Different assumptions on the type of information received, and on the dependence between link costs, lead to different problem scenarios and algorithms. Recently, on-line routing has been extended to traffic assignment in congested networks; the interested reader is referred to the section titled “Background” for more information.

Dynamic Traffic Assignment

Two implicit assumptions in the traffic assignment models presented above are that changes in travel demand and travel time can be ignored. On a link (i, j) , every traveler experiences the same cost $t_{ij}(x_{ij})$, based on the total demand for travel x_{ij} during the analysis period. Clearly, this is a simplification. In reality, travelers depart at different times, often in a highly nonuniform manner (for instance, shift changes at a factory). Furthermore, congestion is not static, but instead grows, shrinks, and evolves constantly based on changing demand and capacity patterns. Dynamic traffic assignment models aim to represent these changes explicitly, and thereby obtain more accurate predictions.

A considerable amount of research has been done in this field in recent years, and Peeta and Ziliaskopoulos [9] provide a comprehensive review. Broadly speaking, dynamic traffic assignment approaches can

be characterized as either analytical or simulation-based. Analytical approaches typically use exit functions or other representations of traffic flow that, while not especially sophisticated, are tractable and amenable to analysis on a large-scale. Simulation-based approaches often use a mesoscopic simulator to represent traffic flow, and seem to offer better performance on large networks, but offer less in the way of theoretical insights and properties. As the formulation of dynamic traffic assignment models diverges sharply from the network models presented in this article, the interested reader is referred to the above reference for specific details.

BACKGROUND

Traffic assignment research can be traced to Wardrop’s [10] seminal work which introduced the notion that travelers choose routes to minimize their travel time. Beckmann *et al.* [11] presented the convex programming formulation (1)–(4) and showed that the vector of link flows minimizing the objective function satisfies Wardrop’s equilibrium condition. Variational inequality formulations of the traffic assignment problem were first explored (individually) by Smith [12], Dafermos [13,14], and Marcotte [15]. Alternate representations of the traffic assignment problem not discussed in this article include nonlinear complementarity [16] and fixed-point [17] formulations.

The starting point for most commonly-used link-based solution methods is found in Frank and Wolfe’s [18] algorithm for general convex programs. Many improvements and variations of the Frank–Wolfe algorithm have been proposed for traffic assignment, as discussed in detail in by Larsson and Patricksson [19]. Route-based methods were actually proposed prior to link-based methods [20,21], but limitations in computer memory prevented their adoption for practical networks until the early 1990s. During this time, both disaggregate simplicial decomposition [19] and gradient projection [4] were found to produce solutions much more accurate than those found by link-based methods. Bush-based algorithms for user equilibrium traffic

assignment are somewhat newer. Although the concept appeared earlier in stochastic network loading [22] and roadway pricing [23] models, its usefulness for traffic assignment was independently discovered by Bar-Gera [5] and Dial [6]. More recent bush-based algorithms can be found in Nie [24] and Bar-Gera [25]. An alternate method not discussed in the article is to directly solve the variational inequality representing the equilibrium condition, as done by Marcotte [15] and Smith [26]. More details on recent variational inequality solution techniques can be found in Facchinei and Pang [27].

The concept of marginal-cost pricing leading to social optimum policy dates to Pigou [28], and early applications to adjusting user-optimal flows toward system optimality can be found in Beckmann *et al.* [11], Walters [29], Dafermos [30], and Smith [31]. Regarding network design in the traffic assignment problem, early works include LeBlanc [32], LeBlanc and Abdulaal [33,36], Dantzing *et al.* [34], and Hoang [35]. Detailed reviews of problem formulations and solution methods can be found in Magnanti and Wong [37], Boyce [38], Friesz [39], and Yang and Bell [40].

Research in stochastic user equilibrium dates to the 1970s [22,41]. The mathematical programming formulation provided in the article was developed by Sheffi and Powell [42], who built on a formulation for general discrete choice problems [43]. Representations of elastic demand can be found in Sheffi [8], Dafermos [14], and LeBlanc and Farhangian [44]. Adaptive routing problems in stochastic networks take a number of forms, depending on assumptions on link cost dependence and information structure [45–50]. Extension of adaptive routing to traffic assignment has recently been accomplished in Unnikrishnan and Waller [51]. Dynamic traffic assignment was first initiated by Merchant and Nemhauser [52,53]. Analytical models were the first to be developed, and seminal works include Carey [54,55] and Friesz *et al.* [56]. Simulation-based models grew in popularity in the late 1990s and early 2000s as computer power increased, examples of which include Jayakrishnan and Mahmassani [57], Ben-Akiva

et al. [58], Waller and Ziliaskopoulos [59], Florian *et al.* [60].

CONCLUSION

This article provided an overview of network modeling for transportation applications. The first step in developing such a model is to construct a network out of nodes and links, representing all of the features of the transportation infrastructure you wish to capture, possibly including physical infrastructure, vehicle routes, and transfer points. Next, an appropriate behavior rule must be decided for assigning travel demand onto these links. The user equilibrium rule is most commonly used, although other options are available—stochastic user equilibrium accounts for perception errors, and system optimal behavior represents the best possible network performance. Several approaches are available for solving for the link flows corresponding to the chosen behavioral rule; at present, the origin-based approaches seem to offer the best computational performance.

However, identifying link flows is only one part of the larger planning process. These link flows can then be further analyzed to obtain measures of effectiveness as part of an alternatives analysis, for instance, by using the link flows as input to an air quality model, or by using them to determine travel times between different points in the network. Equilibrium models can also be applied prescriptively, to determine congestion pricing policies or identify bottlenecks warranting improvement in a network design model.

Research in the field of transportation network analysis is extremely active, and models continue to grow more realistic in representing transportation systems, more efficient in solving large-scale problems, and more applicable to a wider variety of planning, operational, and design problems.

REFERENCES

1. Kockelman KM. Handbook of transportation engineering. New York: McGraw Hill; 2003. Chapter 12.

2. Braess D. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 1969;12:258–268.
3. Chapra SC, Canale RP. Numerical methods for engineers. 4th ed. New York: McGraw Hill; 2002.
4. Jayakrishnan R, Tsai WT, Prashker JN, *et al.* A faster path-based algorithm for traffic assignment. *Transp Res Rec* 1994;1443:75–83.
5. Bar-Gera H. Origin-based algorithm for the traffic assignment problem. *Transp Sci* 2002;36(4):398–417.
6. Dial RB. A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transp Res Part B* 2006;40:917–936.
7. Small KA, Verhoef ET. The economics of urban transportation. 2nd ed. Oxford: Routledge; 2007.
8. Sheffi Y. Urban transportation networks. Englewood Cliffs (NJ): Prentice-Hall; 1985.
9. Peeta S, Ziliaskopoulos AK. Foundations of dynamic traffic assignment: the past, the present, and the future. *Netw Spat Econ* 2001;1:233–265.
10. Wardrop JG. Some theoretical aspects of road traffic research. *Proc Inst Civ Eng Part II* 1952;1:325–378.
11. Beckmann MJ, McGuire CB, Winston CB. Studies in the economics of transportation. New Haven (CT): Yale University Press; 1956.
12. Smith MJ. The existence, uniqueness, and stability of traffic equilibria. *Transp Res Part B* 1979;13B:295–304.
13. Dafermos S. Traffic equilibrium and variational inequalities. *Transp Sci* 1980;14:42–54.
14. Dafermos S. General multimodal network equilibrium problem with elastic demand. *Networks* 1982;12:57–72.
15. Marcotte P. A new algorithm for solving variational inequalities with application to the traffic assignment problem. *Math Program* 1985;33(3):339–351.
16. Aashtiani HZ, Magnanti TL. Equilibria on a congested transportation network. *SIAM J Algebr Discrete Models* 1980;2:213–226.
17. Cantarella GE. A general fixed-point approach to multimode multi-user equilibrium assignment with elastic demand. *Transp Sci* 1987;31:107–127.
18. Frank M, Wolfe P. An algorithm for quadratic programming. *Nav Res Log Q* 1956;3:95–110.
19. Larsson T, Patriksson M. Simplicial decomposition with disaggregated representation for the traffic assignment problem. *Transp Sci* 1992;26:4–17.
20. Dafermos S. Traffic assignment and resource allocation in transportation networks [PhD thesis]. Johns Hopkins University; 1968.
21. Dafermos S, Sparrow FT. The traffic assignment problem for a general network. *J Res Natl Bur Stand* 1969;73B:91–118.
22. Dial RB. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transp Res* 1971;5:83–111.
23. Dial RB. Minimal-revenue congestion pricing part I: a fast algorithm for the single-origin case. *Transp Res Part B* 1999;33:189–202.
24. Nie Y. A class of bush-based algorithms for the traffic assignment problem. *Transp Res Part B* 2010;44:73–89.
25. Bar-Gera H. Traffic assignment by paired alternative segments. *Transp Res Part B* 2010;44(8-9):1022–1046.
26. Smith MJ. The existence and calculation of traffic equilibria. *Transp Res Part B* 1983;17B(4):291–303.
27. Facchinei F, Pang JS. Finite-dimensional variational inequalities and complementarity problems. New York: Springer; 2003.
28. Pigou AC. The economics of welfare. London: Macmillan and Co.; 1920.
29. Walters A. The theory and measurement of private and social cost of highway congestion. *Econometrica* 1961;29:676–699.
30. Dafermos S. Toll patterns for multiclass-user transportation networks. *Transp Sci* 1972;6:211–223.
31. Smith MJ. The marginal cost taxation of a transportation network. *Transp Res Part B* 1979;13B:237–242.
32. LeBlanc L. An algorithm for the discrete network design problem. *Transp Sci* 1975;9:183–199.
33. LeBlanc L, Abdulaal M. An efficient dual approach to the urban road network design problem. *Comput Math Appl* 1979;5:11–19.
34. Dantzig GB, Harvey RP, Landsowne ZF, *et al.* Formulating and solving the network design problem by decomposition. *Transp Res Part B* 1979;13:5–17.
35. Hoang JH. Topological optimization of networks: a non-linear mixed integer model employing generalized benders

- decomposition. *IEEE Trans Automat Control* 1982;27:164–169.
36. LeBlanc L, Abdulaal M. A comparison of the user-optimum versus system-optimum traffic assignment in transportation network design. *Transp Res Part B* 1984;18:115–121.
 37. Magnanti TL, Wong RT. Network design and transportation planning: models and algorithms. *Transp Sci* 1984;18:1–55.
 38. Boyce DE. Urban transportation network equilibrium and design models: recent achievements and future perspectives. *Environ Plann A* 1984;16:1445–1474.
 39. Friesz T. Transportation network equilibrium, design and aggregation: key developments and research opportunities. *Transp Res A* 1985;19:413–427.
 40. Yang H, Bell MGH. Models and algorithms for road network design: a review and some new developments. *Transp Rev* 1998;18(3):257–278.
 41. Daganzo CF, Sheffi Y. On stochastic models of traffic assignment. *Transp Sci* 1977;11(3):253–274.
 42. Sheffi Y, Powell WB. An algorithm for the equilibrium assignment problem with random link times. *Networks* 1982;12(2):191–207.
 43. Daganzo CF. Multinomial probit: the theory and its application to demand forecasting. New York: Academic Press; 1979.
 44. LeBlanc L, Farhangian K. Efficient algorithms for solving elastic demand traffic assignment problems and mode split-assignment problems. *Transp Sci* 1981;15(4):306–317.
 45. Polychronopoulos GH, Tsitsiklis JN. Stochastic shortest path problems with recourse. *Networks* 1996;27(2):133–143.
 46. Cheung RK. Iterative methods for dynamic stochastic shortest path problems. *Nav Res Log* 1998;45(8):769–789.
 47. Miller-Hooks E. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks* 2001;37(1):35–52.
 48. Waller ST, Ziliaskopoulos AK. On the online shortest path problem with limited arc cost dependencies. *Networks* 2002;40(4):216–227.
 49. Provan JS. A polynomial-time algorithm to find shortest paths with recourse. *Networks* 2003;41(2):115–125.
 50. Gao S, Chabini I. Optimal routing policy problems in stochastic time-dependent networks. *Transp Res Part B* 2006;40(2):93–122.
 51. Unnikrishnan A, Waller ST. User equilibrium with recourse. *Netw Spat Econ* 2009;9(4):575–593.
 52. Merchant D, Nemhauser G. A model and an algorithm for the dynamic traffic assignment problems. *Transp Sci* 1978;12(3):183–199.
 53. Merchant D, Nemhauser G. Optimality conditions for a dynamic traffic assignment model. *Transp Sci* 1978;12(3):200–207.
 54. Carey M. A constraint qualification for a dynamic traffic assignment model. *Transp Sci* 1986;20:55–58.
 55. Carey M. Optimal time-varying flows on congested networks. *Oper Res* 1987;35:58–69.
 56. Friesz T, Luque J, Tobin R, *et al.* Dynamic network traffic assignment considered as a continuous time optimal control problem. *Oper Res* 1989;37:893–901.
 57. Jayakrishnan R, Mahmassani H. An evaluation tool for advanced traffic information and management systems in urban networks. *Transp Res Part C Emerg Technol* 1994;2(3):129–147.
 58. Ben-Akiva M, Bierlaire M, Koutsopoulos H, *et al.* DynaMIT: a simulation-based system for traffic prediction. DACCORS Short Term Forecasting Workshop. The Netherlands; 1998.
 59. Waller ST, Ziliaskopoulos AK. A visual interactive system for transportation algorithms. 78th Annual Meeting of the Transportation Research Board. Washington (DC); 1998.
 60. Florian M, Mahut M, Tremblay N. A simulation-based dynamic traffic assignment model: DYNAMIQ. DTA 2006: 1st International Symposium on Dynamic Traffic Assignment. Singapore; 2006.

TRANSIENT BEHAVIOR OF CTMCs

SALVATORE DISTEFANO
Department of Mathematics,
University of Messina,
Messina Italy

KISHOR S. TRIVEDI
Department of Electrical and
Computer Engineering,
Duke University, Durham,
North Carolina

INTRODUCTION

Models are usually classified into two broad categories: simulation and analytic. In a *simulation model*, random variate generators or trace tapes provide time stamps of event streams (e.g., failures and end-of-repairs in a dependability model) that are processed by a computer program that mimics the system behavior. System dynamics is then observed and analyzed using statistical techniques similar to those applied to measurements. In an *analytic model*, the representation of the system is symbolic, using mathematical terms including variables, parameters, relationships such as equations and inequalities, logical statements, and data. System behavior is then studied through the solution of the derived mathematical expressions (ordinarily assisted by software tools).

Different types of analytic models can be distinguished depending on the nature of the underlying phenomena to represent, the characteristics of the system to investigate, and the solution techniques adopted in the evaluation. The models we consider in this article are based on *state space* methods due to their flexibility and power in capturing dependence conditions in the system [1–3]. The state space approach is a very general approach and can handle more cases in dependability and performance modeling than any other analytic method [4]. It can be

used when the component behaviors are *statistically independent (s-independent)* as well as for systems involving dependencies. The method proceeds by the enumeration of *system states*, that is, a collection of variables, the values of which define the state of the system at a given point in time.

State-space models may be of *deterministic* or *stochastic* nature. Stochastic models are usually the method of choice when modeling dependability since phenomena involving significant uncertainties and unpredictable variability (inherent in the system or in its inputs) frequently need to be represented. Through the probabilistic approach, the repercussions of such uncertainties in the model solution can be clearly evidenced. Stochastic models can be further classified as *Markovian* or *non-Markovian*.

The most commonly used stochastic model for performance and reliability analysis are the Markov models. A common belief, in the reliability community, is that Markov models only capture exponential distributions. This is true for the *homogeneous continuous-time Markov chain* (HCTMC), which assumes rates associated with events such as arrivals, service completions, failures, repairs, and so on are all constant in time, and the resulting sojourn times are all exponentially distributed. HCTMCs cover a wide range of real-world modeling problems. However, some behaviors of many practical systems cannot be captured by HCTMC, for example, time-dependent rates or nonexponential distributions and aging effects.

There are several interesting attempts in the literature of exploiting Markovian models such as *homogeneous* and *nonhomogeneous continuous-time Markov chains* to study systems with nonexponentially distributed behaviors. In the former case, the solution is to resort to the well-known and powerful *phase type expansion* technique that allows one to overcome the potential loss of the Markov property in dealing with nonexponential distributions. The latter alternative is to use nonhomogeneous

continuous time markov chain (NHCTMC) with time varying rates, by considering a unique and global time (no *regenerations*).

In this way, Markov models have become one of the most used paradigm in the stochastic analysis of dynamic systems in both transient and steady state, due to their modeling power and available solution techniques [3,5–7]. In this article, we deal with CTMC, and, in particular, we mainly focus on the more general NHCTMC solution techniques for their transient behavior. Thus, in the section titled “Markov Models,” we introduce the Markov modeling basics, and in the section titled “Markov Chain Analysis,” we discuss some techniques for the analysis of NHCTMC.

MARKOV MODELS

Markov models have often been used for software and hardware performance and dependability assessment. Reasons for the popularity of Markov models include the ability to capture various dependencies; the equal ease with which steady-state, transient, and cumulative transient measures can be computed; and the extension to Markov reward models, which are useful in performability analysis [8]. For example, Markov models are quite useful for systems with dependent failure and repair modes, as well as when components behave in an s-independent manner [2,5]. Furthermore, they can handle multistate devices and common-cause failures without any conceptual difficulty.

A stochastic process $\{X(t); t \in T\}$, taking values in a discrete state space Ω and indexed by a parameter set T (commonly representing the time), is called a *Markov chain* if for any $t_0 < t_1 < \dots < t_n < t$, the conditional probability mass function (pmf) of $X(t)$ for given values of $X(t_0), \dots, X(t_n)$, depends only on $X(t_n)$; that is

$$\begin{aligned} \Pr\{X(t) = x \mid X(t_n) = x_n, \dots, X(t_0) = x_0\} \\ = \Pr\{X(t) = x \mid X(t_n) = x_n\}. \end{aligned} \quad (1)$$

A Markov chain is thus a discrete state stochastic process, whose dynamic behavior is such that the probability distributions for

its future development depends only on the present state and not on how the process arrived at that state. If the conditional pmf of Equation (1), which expresses the *Markov property*, also has the property of invariance with respect to the time origin t_n ; that is

$$\begin{aligned} \Pr\{X(t) = x \mid X(t_n) = x_n\} \\ = \Pr\{X(t - t_n) = x \mid X(0) = x_n\}, \end{aligned} \quad (2)$$

the Markov chain is said to be *(time-)homogeneous*. For a homogeneous Markov chain, the past history of the process is completely summarized in the current state. Otherwise, the exact characterization of the present state needs the associated time information, and the process is said to be *nonhomogeneous*.

In the case of a *discrete-time Markov chain* (DTMC), the instants when state changes may occur are preordered to be at the integers $0, 1, 2, \dots, n, \dots \in \mathbb{N}$. In the case of a *continuous-time Markov chain* (CTMC), however, the transitions between states may take place at any instant in time.

In a homogeneous Markov chain, the Markov property imposes a heavy constraint on the distribution of time that the process may remain in a given state. In fact, the time Y_x the process spends in a given state $x \in \Omega$ is a r.v. and its distribution must be s-independent of how long it has already spent in the state. This important property, known as the *memoryless property*, is expressed as

$$\Pr\{Y_x < t + r \mid Y_x \geq t\} = \Pr\{Y_x < r\}, \quad r \geq 0. \quad (3)$$

where Y_x is a continuous random variable, and its distribution is *negative exponential*. Otherwise, the only discrete distribution with this property is the *geometric* distribution. This implies that the time that a homogeneous continuous-parameter Markov chain spends in a given state is *exponentially distributed*. Similarly, the time spent in a given state of a homogeneous discrete-parameter Markov chain is *geometrically distributed*.

Markov chains are graphically represented by state-transition diagrams, which

are directed graphs with nodes representing states and arcs representing transitions. The nodes are labeled with state names while the transitions are labeled with either transition probabilities (DTMC) or transition rates (CTMC).

As stated above, the Markov property and time-homogeneity property make the sojourn times in CTMC states exponentially distributed. This helps generating many efficient and elegant algorithms for homogeneous CTMC analysis, but it also restricts its flexibility. To make it more general, we consider CTMC without the time-homogeneity property or NHCTMC.

For NHCTMC, the following differential equation holds:

$$\frac{d\pi(t)}{dt} = \pi(t)\mathbf{Q}(t) \quad (4)$$

where $\pi(t)$ is the state probability vector at time t , $\mathbf{Q}(t) = [q_{ij}(t)]$ is the infinitesimal generator matrix, $q_{ij}(t) = \lim_{\delta \rightarrow 0} \frac{\pi_{ij}(t, t+\delta)}{\delta}$, $i \neq j$ is the transition rate from state i to state j at time t , and $q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t)$, $(\pi_{ij}(t, t+h) = \Pr\{X(t+h) = x_j \mid X(t) = x_i\})$. In the case of homogeneous CTMC, $\mathbf{Q}(t)$ is time invariant and thus, $\mathbf{Q}(t) = \mathbf{Q}$.

MARKOV CHAIN ANALYSIS

There are several methods for CTMC transient analysis (both homogeneous and nonhomogeneous): the Ordinary differential equation (*ODE method*) [9–12], *piecewise approximation* [13], and *uniformization* [13–15]. But other techniques can be exploited in the solution of CTMC, applying alternative approaches (Krylov subspace, Markov renewal processes, Petri nets, fluid models, simulation, PH (phase type) expansion, metaheuristic, etc.) [16–23].

In the following, we first describe the classical CTMC solution techniques (in the sections titled “ODE Method,” “Piecewise Approximation,” and “Uniformization Method”), in particular investigating, in the section titled “Piecewise Approximation,” a specific optimized piecewise approximation algorithm. Then, in the section titled “PH

Distributions,” we discuss an NHCTMC solution technique based on PH. Other techniques are briefly reviewed in the section titled “Other Solution Techniques.”

ODE Method

For transient analysis of nonhomogeneous CTMC, we are interested in solving Equation (4), which can be seen as the following *linear initial-value problem* [24]:

$$\dot{\mathbf{y}}(t) = \mathbf{y}(t)\mathbf{Q}(t) = f(t, \mathbf{y}(t)) \quad (5)$$

In Equation (5), if $\mathbf{Q}(t)$ is piecewise continuous (or more weakly, integrable), there will exist a unique solution of the form

$$\mathbf{y}(t) = \mathbf{y}(t_0)\Phi(t, t_0) \quad (6)$$

where, in the specific case of CTMC expressed by Equation (4), letting $\pi(t) = \mathbf{y}(t)$, $\Phi(t, t_0) \in R^{m \times m}$ is the transition probability matrix mapping the initial condition $\pi(t_0)$ into $\pi(t)$, and m is the number of states of the CTMC.

If $\mathbf{Q}(t)$ and $\int_{t_0}^t \mathbf{Q}(\tau) d\tau$ commute $\forall t$, the general solution for $\Phi(t, t_0)$ can be expressed as

$$\Phi(t, t_0) = e^{\int_{t_0}^t \mathbf{Q}(\tau) d\tau} \quad (7)$$

where the matrix exponential is defined by its Taylor series expansion $e^{\int_{t_0}^t \mathbf{Q}(\tau) d\tau} \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} \frac{1}{i!} \left(\int_{t_0}^t \mathbf{Q}(\tau) d\tau \right)^i$.

However, it is well known that Equation (7) is not a general solution, it is only valid under the assumption that $\mathbf{Q}(t)$ and $\int_{t_0}^t \mathbf{Q}(\tau) d\tau$ commute $\forall t$, always true in case $\mathbf{Q}(t) = \mathbf{Q}$ is time independent. A particular case in which $\mathbf{Q}(t)$ and its time integral commute is $\mathbf{Q}(t) = f(t)\mathbf{Q}$, where $f(t)$ is an arbitrary scalar function of time and $\mathbf{Q}$ is any valid time-invariant infinitesimal generator matrix. Methods for computing $\Phi(t, t_0)$ are well known for time-invariant systems, but computing it for time varying systems is more difficult.

A successful solution strategy is the discretization: we seek a solution on a discrete point set $\{t_i \mid i = 0, 1, \dots, n\}$ where $h = t_{i+1} - t_i = (b - a)/n$ is the step length (we assume h is fixed for analysis purpose but, in general, the step length may vary). Let $\mathbf{y}_i$ be

the approximation of the theoretical solution $\mathbf{y}(t_i)$ at time t_i . In computing $\mathbf{y}_{i+1}$, a method may incorporate the values from previously computed approximations $\mathbf{y}_j$ for $j = 0, 1, \dots, i$, or even $\mathbf{y}_{i+1}$ itself. A method that only uses $(t_i, \mathbf{y}_i)$ to compute $\mathbf{y}_{i+1}$ is said to be an *explicit single step method*. It is said to be a *multistep method* when it uses approximations at several previous steps to compute its new approximation. A method is said to be an *implicit method* when $\mathbf{y}_{i+1}$ requires a value for $\mathbf{y}_{i+1}$.

For a solution method, the *local truncation error* of a given step is $|\mathbf{y}_i - \mathbf{y}(t_i)|$. The *method's order* is the largest integer p such that the local truncation error is $O(h^{p+1})$. A method is *A-stable* if the numerical approximations $\mathbf{y}_n$ of $\mathbf{y}(t_n)$ approaches zero as $n \rightarrow \infty$, when applied to the differential equation $\dot{\mathbf{y}} = \lambda \mathbf{y}$ with a fixed positive h and a complex constant λ , with $\text{Re}(\lambda) < 0$. The method is *L-stable* when it is A-stable and, in addition, when applied to the equation $\dot{\mathbf{y}} = \lambda \mathbf{y}$, where λ is a complex constant with $\text{Re}(\lambda) < 0$, it yields $|\mathbf{y}_{i+1} - \mathbf{y}_i| \rightarrow 0$ as $\text{Re}(h\lambda) \rightarrow -\infty$ [25].

Most of the conventional ODE methods perform acceptably for transient analysis of nonstiff Markov models [26], that is, in case $\max\{|q_{ii}|t\}$ is not large. One of the commonly used methods is the fourth order Runge–Kutta [9,10], which is an explicit single step method, and is shown in Equation (8)

$$\mathbf{y}_{i+1} = \mathbf{y}_i + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4) \quad (8)$$

where

$$\begin{cases} F_1 = f(t, \mathbf{y}_i) \\ F_2 = f(t + \frac{h}{2}, \mathbf{y}_i + \frac{h}{2}F_1) \\ F_3 = f(t + \frac{h}{2}, \mathbf{y}_i + \frac{h}{2}F_2) \\ F_4 = f(t + h, \mathbf{y}_i + hF_3) \end{cases}$$

However, for stiff Markov chains explicit methods have enormous difficulty to solve the corresponding ODE. The problems stem from the fact that the solutions of stiff systems contain rapidly decaying transient terms [27]. Since explicit methods do not satisfy the A-stability property, small step sizes must be used over the entire period of integration to avoid generating large numerical errors.

Because transient solvers are often used to model a system from start-up until steady-state, explicit methods are time-consuming and often yield incorrect steady-state solution. Moreover, for extremely stiff problems, even A-stability may be inadequate. To overcome this problem, Reibman and Trivedi [11] proposed the use of TR-BDF2 method, which is an implicit two-step method with L-stability.

To compute $\mathbf{y}_{i+1}$, the TR-BDF2 method first takes a *trapezoid rule* (TR) step from t_i to $t_i + \gamma h$, where γ is a constant between 0 and 1: $0 < \gamma < 1$. The trapezoid rule applied to interval $[t_i, t_i + \gamma h]$ is

$$\mathbf{y}_{i+\gamma} - \mathbf{y}_i = \gamma h_i \cdot \frac{f(t_i, \mathbf{y}_i) + f(t_i + \gamma h_i, \mathbf{y}_{i+\gamma})}{2} \quad (9)$$

After getting $\mathbf{y}_{i+\gamma}$, we use the *second order backward differentiation formulae* (BDF2) method to step from $t_i + \gamma h$ to t_{i+1} , which is given by

$$\begin{aligned} (2 - \gamma) \mathbf{y}_{i+1} - (1 - \gamma)h_i f_{i+1} \\ = \gamma^{-1} \mathbf{y}_{i+\gamma} - \gamma^{-1}(1 - \gamma)^2 \mathbf{y}_i \end{aligned} \quad (10)$$

When computing the state probabilities with Equation (5), the TR step in Equation (9) is computationally the same as solving the first order linear algebraic system

$$\begin{aligned} \pi_{i+\gamma} [2\mathbf{I} - \gamma h_i \cdot \mathbf{Q}(t_i + \gamma h_i)] \\ = \pi_i [2\mathbf{I} + \gamma h_i \cdot \mathbf{Q}(t_i)] \end{aligned}$$

and the BDF2 step in Equation (10) becomes solving the linear system

$$\begin{aligned} \pi_{i+1} [(2 - \gamma)\mathbf{I} - (1 - \gamma)h_i \mathbf{Q}(t_i + h_i)] \\ = \gamma^{-1} \pi_{i+\gamma} - \gamma^{-1}(1 - \gamma)^2 \pi_i \end{aligned}$$

The local truncation error of TR-BDF2 method is $O(h^3)$. To better control error and provide a more efficient solution, h_i can be varied from step to step. If too much error is introduced, the step is repeated with smaller h . After a successful step, h may be increased.

Another interesting technique for dealing with stiffness in Markov chain analysis is proposed in Ref. 12 as an extension of implicit Runge–Kutta method. By exploiting the method specified in Ref. 28 and applying the Pade approximation to the resulting polynomials, the authors obtain a $2p - 1$ order method that, at each time step of constant size h , can be synthesized by the following algebraic system:

$$\pi_{i+1} \sum_{j=0}^r \alpha_j (h\mathbf{Q})^j = \pi_i \sum_{j=0}^{r-1} \beta_j (h\mathbf{Q})^j \quad (11)$$

where α_j and β_j are constants whose values are determined based upon the order $p = 2r - 1$ of the method desired. It has been shown in Ref. 29 that these methods are A-stable, and, under some specific assumptions based on the Pade approximation of polynomials, such methods are L-stable. In particular, by using $r = 2$ in Equation (11), a third order L-stable method is obtained

$$\pi_{i+1}(\mathbf{I} - 2/3h\mathbf{Q} + 1/6h^2\mathbf{Q}^2) = \pi_i(\mathbf{I} + 1/3h\mathbf{Q}). \quad (12)$$

This system can be solved by either computing the matrix polynomial directly, or by factorizing the matrix polynomial and then solving the two successive linear algebraic systems

$$\mathbf{X}(\mathbf{I} - r_2h\mathbf{Q}) = \pi_i(\mathbf{I} + 1/3h\mathbf{Q}) \quad (13)$$

$$\mathbf{X} = \pi_{i+1}(\mathbf{I} - r_1h\mathbf{Q}) \quad (14)$$

where r_1 and r_2 are the roots of the polynomial $1 - 2/3x + 1/6x^2$ corresponding to the factorization of Equation (12) into $\pi_{i+1}(\mathbf{I} - r_1h\mathbf{Q})(\mathbf{I} - r_2h\mathbf{Q}) = \pi_i(\mathbf{I} + 1/3h\mathbf{Q})$. Note that, since r_1 and r_2 are a complex conjugate pair, this requires the use of complex arithmetic.

Reference [12] provides a comprehensive overview of stiffness-tolerant methods for transient analysis of Markov chains.

Piecewise Approximation

One simple solution for the transient analysis of nonhomogeneous CTMC is the *piecewise approximation*: here a set of discrete

points $\{t_0 = 0, t_1, t_2, \dots, t_k = t\}$ is found on the continuous-time interval $[0, t]$, and the nonhomogeneous CTMC is considered to be a homogeneous CTMC in each interval $[t_i, t_{i+1}]$, $i = 0, \dots, k - 1$.

A function $f : (a, b) \subseteq \mathbb{R} \rightarrow \mathbb{R}$ is said to be *piecewise constant*, if it is locally constant in connected regions separated by a possibly infinite number of lower-dimensional boundaries. Let $t_0 = a < t_1 < \dots < t_k = b$ be such boundaries or *breakpoints*, f is constant on each interval $[t_i, t_{i+1}]$, $i = 0, 1, \dots, k - 1$. Every piecewise linear function can be written as

$$f(x) = c + \sum_{i=0}^k a_i g_i(x) = \sum_{i=0}^k f_i(x),$$

where

$$g_i(x) = \begin{cases} 1, & \text{if } t_i \leq x < t_{i+1}; \\ 0, & \text{otherwise.} \end{cases}$$

$f_i(x) = c + a_i g_i(x)$ and $a_i, c \in \mathbb{R}$ are constant values. Piecewise constant functions are often used to approximate more general functions. The piecewise constant approximation has the following three parameters that can be set in order to better fit the original function $y(x)$:

- the number of intervals k ;
- the length of each interval $[t_i, t_{i+1}]$, that can be constant or variable, depending on $i = 0, \dots, k - 1$ (adaptive according to the trend of the function to approximate);
- the value of c and a_i .

These parameters should be fixed in order to minimize the errors $e_i(x)$ and $e(x)$

$$\begin{cases} e_i(x) = y(x) - f_i(x) = y(x) - s_i, & t_i \leq x < t_{i+1}, \\ e(x) = y(x) - f(x), & a \leq x \leq b. \end{cases}$$

Each $s_i = f_i(x) \forall x \in [t_{i-1}, t_i]$, given t_i and t_{i+1} are chosen from the class of approximating functions by minimizing the norm

$$\sqrt[o]{\int_{t_{i-1}}^{t_i} |e_i(x)|^o dx},$$

where the cases $o = 1, 2, \infty$ correspond respectively to *minimum mean absolute*

deviation, least squares, and minimum maximum absolute deviation. The criterion of least squares, that is the most common and widely used in piecewise approximation problems, can be written equivalently as

$$\int_{t_i}^{t_{i+1}} e_i(x)^2 dx = \int_{t_i}^{t_{i+1}} (y(x) - s_i)^2 dx, \quad (15)$$

without taking into consideration the square root; we will use this latter form of the criterion in the following.

In the specific case, we wish to piecewise approximate the infinitesimal generator matrix $\mathbf{Q}(t) = [q_{pj}(t)]$, with $p, j = 0, \dots, m$ and $m = |\Omega|$ the number of states. In order to optimize the approximation, one way may be to separately approximate each $q_{pj}(t)$, but since each of these has its own trend, it is better to fix the number of intervals k and their common length $h = t_{i+1} - t_i$, with $i = 0, \dots, k - 1$. The only free parameter to fix is the value of $q_{pj}(t) = q_{pj}^{(i)}$ in the interval $[t_i, t_{i+1}]$, and therefore to approximate the matrix

$$\mathbf{Q}(t) = \mathbf{Q}^{(i)} = [q_{pj}^{(i)}]$$

for that interval. Given the initial probability vector $\pi(t_i)$, $\pi(t_{i+1})$ can be computed by solving the differential equation of the homogeneous CTMC:

$$\frac{d\pi(t)}{dt} = \pi(t)\mathbf{Q}^{(i)}. \quad (16)$$

For certain simple homogeneous CTMC such as the M/M/1/K queue, closed-form solution can be obtained from Equation (16) [13].

For the general case, for $t_i \leq t \leq t_{i+1}$, the solution of Equation (16) can be written as

$$\begin{aligned} \pi(t) &= \pi(t_i) \cdot \sum_{k=0}^{\infty} \frac{(\mathbf{Q}^{(i)})^k (t - t_i)^k}{k!} \\ &= \pi(t_i) \cdot e^{\mathbf{Q}^{(i)}(t - t_i)}. \end{aligned} \quad (17)$$

Given that $h = t_{i+1} - t_i$, this yields

$$\begin{aligned} \pi(t_n) &= \pi(t_{n-1}) \cdot e^{\mathbf{Q}^{(n-1)}h} = \pi(t_{n-2}) \cdot e^{\mathbf{Q}^{(n-2)}h} \\ &\cdot e^{\mathbf{Q}^{(n-1)}h} = \dots = \pi(t_0) \cdot e^{\mathbf{Q}^{(0)}h} \cdot e^{\mathbf{Q}^{(1)}h} \dots e^{\mathbf{Q}^{(n-1)}h}. \end{aligned} \quad (18)$$

Note that Equation (18) can not be simplified by summing the exponents, since $e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{\mathbf{A}+\mathbf{B}}$ except when $\mathbf{A}$ and $\mathbf{B}$ commute.

Our problem is to select the best value of s_i minimizing the approximation error of Equation (15). This problem can be reformulated by considering the error as a function of two variables $e_i(x, s_i)$; taking derivative w.r.t. s_i :

$$\begin{aligned} &\frac{d}{ds_i} e_i(x, s_i) \\ &= \frac{d}{ds_i} \left(\int_{t_i}^{t_{i+1}} (y(x) - s_i)^2 dx \right) \\ &= \frac{d}{ds_i} \left(\int_{t_i}^{t_{i+1}} y^2(x) - 2s_i y(x) + s_i^2 dx \right) \\ &= 2(t_{i+1} - t_i)s_i - 2 \int_{t_i}^{t_{i+1}} y(x) dx. \end{aligned} \quad (19)$$

By setting the first derivative equal to zero

$$\begin{aligned} 2(t_{i+1} - t_i)s_i - 2 \int_{t_i}^{t_{i+1}} y(x) dx &= 0 \Rightarrow \\ s_i &= \frac{1}{t_{i+1} - t_i} \int_{t_i}^{t_{i+1}} y(x) dx, \end{aligned}$$

that is the mean value of function $y(x)$ in the interval $[t_i, t_{i+1}]$. This value is the minimum of $e_i(x, s_i)$ while varying s_i , since the second derivative of this function is positive $\frac{d^2}{ds_i^2} e_i(x, s_i) = 2(t_{i+1} - t_i) > 0$.

In order to manage pdf that have $\lambda(t_*) = \infty$ at $t_* \in [0, +\infty[\subset \mathbb{R}$ (such as the Weibull pdf with shape parameter $\alpha < 1$ or DFR in $t_* = 0$), it is necessary to consider an $0 < \epsilon \leq h$, and to evaluate the NHCTMC in the intervals $[t_* - h, t_* - \epsilon]$ and $[t_* + \epsilon, t_* + h]$.

The piecewise approximation method is simple and easy to implement using homogeneous CTMC solvers. This motivates its wider usage. However, it does not provide a rigorous error control mechanism, which makes it less preferable in real applications. In order to achieve higher accuracy, it is necessary to reduce the approximation step width h . As a consequence, the solution's execution times increase, lowering the performance of solvers and sometimes limiting the applicability of the technique.

In addition, this method is not particularly suitable for stiff Markov chains [11,12,14,27]. The ODE method and the uniformization method introduced in the following section can be used to bypass these problems.

Uniformization Method

Another way to solve the differential equation (4) is to use *uniformization* (also known as *randomization* or *Jensen's method*). This technique is based on the reduction of a CTMC with a finite state space Ω into a DTMC subordinated to a Poisson Process. Before presenting the uniformization method for nonhomogeneous CTMC, we first review it for the homogeneous case.

From a homogeneous CTMC with m states ($|\Omega| = m$) and generator matrix $\mathbf{Q} = [q_{ij}]_{m \times m}$, we obtain a DTMC $Z = \{Z_i, i = 0, 1, \dots\}$ and a Poisson process $\{N(t), t \geq 0\}$ with rate q in the following way: let $q > \max_{1 \leq i \leq m} | -q_{ii} |$, and the transition probability matrix of Z be $\mathbf{P} = \mathbf{Q}/q + \mathbf{I}$. We then form an equivalent process $\{X(t), t \geq 0\}$ as $X(t) = Z_{N(t)}$. This process is defined by the transition probability matrix $\mathbf{P}$, and the transition rate (*fictitious rate*) from each state j is equal to q , with $1 - |q_{jj}|/q$ fraction of these transitions return immediately to state j . Therefore, the probability that the process is in state j at time t is the same as that of the original CTMC.

This process is called the *uniformization* (or *randomization*) of the original CTMC, and the transient solution can be computed by conditioning on $N(t)$, the number of transitions in $[0, t]$

$$\begin{aligned} \pi_j(t) &= \Pr\{X(t) = j\} \\ &= \sum_{i=0}^{\infty} \Pr\{X(t) = j | N(t) = i\} \Pr\{N(t) = i\}. \end{aligned}$$

Thus

$$\pi(t) = \sum_{i=0}^{\infty} \mathbf{v}(i) \cdot e^{-qt} \frac{(qt)^i}{i!}, \quad (20)$$

where $\mathbf{v}(i) = \{v_1(i), v_2(i), \dots, v_m(i)\}$ is the state probability vector of the subordinated DTMC Z after i transitions, with $\mathbf{v}(0) = \pi(0)$.

Note that $\mathbf{v}(i)$ can be computed by successive matrix-vector multiplications

$$\mathbf{v}(i) = \mathbf{v}(i-1)\mathbf{P},$$

therefore Equation (20) can also be written as

$$\pi(t) = \pi(0) \sum_{i=0}^{\infty} \mathbf{P}^i \cdot e^{-qt} \frac{(qt)^i}{i!}. \quad (21)$$

To compute the solution, the uniformization series in Equation (20) needs to be truncated so as to meet the desired error tolerance. One simple error bound is based on a right-side truncation of the Poisson distribution. If the right truncation point is k , the series is truncated after the k th term. If the truncation error is bounded by ϵ , the truncated series must satisfy the following condition:

$$\epsilon \leq 1 - e^{-qt} \sum_{i=0}^k (qt)^i / i!$$

As qt grows, the Poisson pmf thins, and the terms on the left side grows slower than the former. Therefore, left truncation can also be done in order to accelerate the computation, and the left truncation error can be bounded by using a normal approximation of Poisson distribution. For dealing with stiff CTMC, Muppala and Trivedi [14] proposed the use of steady-state detection.

For nonhomogeneous CTMC with m states and generator matrix $\mathbf{Q}(t) = [q_{ij}(t)]$, assume that there exists a $q < \infty$ such that, for $\forall t < \infty, q > \max_{1 \leq i \leq m} | -q_{ii}(t) |$. We define the Poisson process $\{N(t), t \geq 0\}$ and the DTMC $Z = \{Z_i, i = 0, 1, \dots\}$ for the nonhomogeneous CTMC in the same way as for the homogeneous CTMC. The only difference is that now the transition probability matrix $\mathbf{P}$ becomes $\mathbf{P}(t)$, which is time-dependent.

Given the definitions above, the following theorem from Ref. 15 gives the uniformization series for nonhomogeneous CTMC.

Theorem 1. For all $0 \leq t \leq \infty$,

$$\begin{aligned} \pi(t) &= \pi(0) \sum_{k=0}^{\infty} \frac{(qt)^k}{k!} e^{-qt} \cdot \int_0^t \int_0^t \cdots \int_0^t \\ &\quad (t_1 \leq t_2 \leq \cdots \leq t_k) \\ &\quad \mathbf{P}(t_1) \mathbf{P}(t_2) \cdots \mathbf{P}(t_k) d\bar{H}(t_1, t_2, \dots, t_k) \\ &= \pi(0) \cdot \hat{\mathbf{U}}(t), \end{aligned} \quad (22)$$

where $d\bar{H}(t_1, t_2, \dots, t_k)$ is the density of the order statistics $t_1 \leq t_2 \leq \cdots \leq t_k$ of a k -dimensional uniform distribution at $[0, t] \times \cdots \times [0, t] \subset \mathbb{R}^k$, and $\mathbf{P}(t_1) \mathbf{P}(t_2) \cdots \mathbf{P}(t_k)$ is the standard matrix product of the transition probability matrices at times $t_1 \leq t_2 \leq \cdots \leq t_k$.

The main complication in computing the nonhomogeneous expression in Theorem 1 is the fact that a continuum of transition matrices is required. To overcome this complication, in Ref. 15 a finite-grid approximation is used. Let $0 < h < q^{-1}$ be the step-size and denote $n_i = \lfloor t_i h^{-1} \rfloor$, $i = 1, 2, \dots$, the theorem for discrete-time approximation of the uniformization series is as follows:

Theorem 2. For arbitrary n and $t = nh$, define

$$\begin{aligned} \mathbf{U}(t) = \mathbf{U}(nh) &= \sum_{k=0}^{\infty} \frac{(qt)^k}{k!} e^{-qt} \cdot \left[\sum_{\substack{n-1 \\ 0}}^{n-1} \sum_{\substack{n-1 \\ 0}}^{n-1} \cdots \sum_{\substack{n-1 \\ 0}}^{n-1} \right. \\ &\quad \left. \mathbf{P}(n_1 h) \mathbf{P}(n_2 h) \cdots \mathbf{P}(n_k h) \bar{H}(n_1, n_2, \dots, n_k), \right] \end{aligned} \quad (23)$$

where $\bar{H}$ is the pmf of the order statistics $n_1 \leq n_2 \leq \cdots \leq n_k$ of a k -dimensional uniform distribution over $\{0, 1, \dots, n-1\}^k$. With the assumption that the transition rates are Lipschitz continuous, that is,

$$\|\mathbf{Q}(t + \Delta t) - \mathbf{Q}(t)\| \leq \Delta t \cdot k,$$

it can be shown that

$$\|\mathbf{U}(t) - \hat{\mathbf{U}}(t)\| \leq ht \cdot k,$$

where $\hat{\mathbf{U}}(t)$ is the true value given in Equation (22).

Given a truncated series of $\mathbf{U}(t)$ in Theorem 2, $\pi(t)$ can be computed using $\pi(0)\mathbf{U}(t)$.

PH Distributions

Owing to its great power and flexibility, the PH approach [16] is commonly applied in the analysis of a large class of stochastic processes such as *semi-Markov process* (SMP) and *Markov regenerative process*. A possible idea is to use this technique for the solution of NHCTMC. In a generic NHCTMC, there are some transitions ($d \geq 1$) characterized by time dependent transition rates $\lambda_i(t)$, with $i = 1, \dots, d$. The concept of time in NHCTMC is *global*, in the sense that the value assumed by each timing variant failure rate at time $t = t^*$ is uniquely determined by substituting t^* in it, and so $\lambda_i(t^*)$. In other words, all $\lambda_i(t)$ evolve according to a common time t , measured from a global clock starting at the beginning of CTMC operation. By contrast, rates in an SMP are dependent on local time, measured from the time of entry into the current state.

From the PH viewpoint, it is necessary to take into account the history/memory reached by all the $\lambda_i(t)$ by means of the phases. Therefore, in general, it is necessary to substitute each state of the NHCTMC by the vector-cross product among the PH distributions representing such $\lambda_i(t)$. In this way, the transition between two states preserves the stage reached by each of the time dependent transition rates.

The main drawback of this technique is the combinatorial growth in the state space of the PH-CTMC approximation obtained. For example, consider a simple NHCTMC composed of three states (0, 1, 2) and three time dependent transitions ($0 \rightarrow 1$, $1 \rightarrow 2$, $2 \rightarrow 0$). Suppose each transition is approximated by a three phase PH. In order to take into account the phase reached by each PH, it is necessary to consider, for each of the three states, all the possible combinations among the three PH. Thus, we represent each state of the NHCTMC by a CTMC with $3 \times 3 \times 3 = 27$ nodes, and so the overall approximation of the original NHCTMC has $27 \times 3 = 81$ states.

Table 1. Summary of the Transient Behavior Analysis Techniques Discussed

Type	Name	MC	References	Equation	Algorithm	Stiff	Pros	Cons
ODE	Fourth order Runge–Kutta	*	[9,10]	(8)	Explicit single step	No	Simple, well known and widely spread.	Not adequate for stiffness.
	TR-BDF2	*	[11]	(9), (10)	Implicit two step	Yes	Easy to implement, also in case of stiffness.	It is only second-order accurate.
	($2r - 1$)th order Runge–Kutta	*	[12]	(11)–(13)	Implicit	Yes	Greater accuracy.	Complexity.
Piecewise approximation		NH	[13]	(16)–(18)	Iterative	Partial	Easy to implement.	Not adequate for stiff models.
Uniformization		*	[13,15]	(20)–(22), (1), (2)	DTMC + PP	Partial	Recommended in case of nonstiff CTMC.	Not adequate for stiff CTMC.
Stiff uniformization		*	[14]		Steady-state detection	Yes	Recommended in case of stiff CTMC.	Adequate for stiff CTMC.

PH		NH	[16]	Phase type expansion	Partial	Simple, easy to implement, general.	
Others	Krylov subspace method	H	[17]		Yes	Deal with largeness and stiffness.	Complex algorithm and parameters tuning.
	Fluid flow models	H	[18]		Yes	Powerful and general.	Some implementation could have problems with stiffnes.
	NMSPN	H	[19]		Partial	Higher level formalism large CTMC.	Problems with stiff models may arise.
	Simulation	*	[20]		Partial	Easy to implement.	Simulation has problems in dealing with stiffness Accuracy.
	Markov renewal theory	H	21		Yes	Powerful and general.	The problem is transformed into a PDE solution.
	Metaheuristic	*	[22,23]	Bayesian net	Yes	Can start from real data.	Time, due to inference procedures and logistic regression setup.

MC Type of Markov Chain.

H Homogeneous.

NH Non-Homogeneous.

*Any type both homogeneous and non-homogeneous.

Hence the PH technique is useful only for relatively simple NHCTMC, with few states and few time varying transition rates. Note that, in some cases the complexity of the overall PH-CTMC can be significantly reduced based on their specific structure.

Other Solution Techniques

Many other solution techniques exist for Markov chains, such as the Krylov subspace method [17], fluid flow [18], non-Markovian stochastic Petri nets [19], simulation [20], Markov renewal theory [21], and meta-heuristic [22]. Specific techniques have been developed in particular cases. An interesting technique for obtaining the analytical solution of periodic and aperiodic NHCTMCs based on Floquet's method is provided in Ref. 30.

Several alternatives have been developed for solving discrete-time NHMC. For example, in Ref. 22, the authors propose a hierarchical Bayesian framework for modeling and analyzing NHMC. Also, Paroli and Spezia [23] make use of Bayesian network to solve NHCTMC. Interesting results have been obtained in Ref. 31, where the problem of evaluating the steady-state of a NHMC is transformed into a more simple problem by decomposing the original NHMC into homogeneous MCs, applying the *decomposition-separation* theorem there demonstrated.

Table 1 provides a compact overview and a comparison of the techniques for evaluating the transient behavior of NHCTMC discussed in this article.

CONCLUSIONS

Markov models have become a modeling paradigm of choice in industrial and academic environments for both practical applications and theoretical research. The availability of stable and user-friendly tools based on Markov models has largely contributed to the success of this paradigm as a general purpose, flexible and effective modeling and analysis method [7,32]. The research exploiting specific properties and structures of Markov models has rapidly increased the

size of the problems that can be effectively handled. The topic has attracted much effort in the research community as evidenced by the number of papers dealing with it, and continues to offer new avenues for research.

New challenging and promising results have been recently obtained in an attempt to overcome the memoryless/exponential assumption. The possibility of dealing with nonexponential distributions represents a realistic goal with the present state of the art. In such a context, this article makes a general overview of the existing approaches, providing a detailed survey on CTMC solution techniques, with particular regard to their transient analysis.

Acknowledgment

K.S. Trivedi's research was supported in part by the US National Science Foundation under Grant NSF-CNS-08-31325.

REFERENCES

1. Heimann DI, Mittal N, Trivedi K. Availability and reliability modeling for computer systems. In: Yovits MC, editor. Volume 31, *Advances in computers*. San Diego (CA): Academic Press; 1990; pp. 175–233.
2. Muppala JK, Malhotra M, Trivedi KS. Markov dependability models of complex systems: Analysis techniques. In: Özekici S, editor. *Reliability and maintenance of complex systems*. Berlin: Springer; 1996; pp. 442–486.
3. Sahner R, Trivedi KS, Puliafito A. *Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1995.
4. Dhillon BS. *Reliability engineering in systems design and operation*. New York: Van Nostrand Reinhold; 1983.
5. Trivedi KS. *Probability and statistics with reliability, queueing, and computer science applications*. 2nd ed. New York: John Wiley & Sons, Inc.; 2001.
6. Bolch G, Greiner S, de Meer H, *et al.* *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. 2nd ed. New York: Wiley-Interscience; 2006.

7. Trivedi KS, Sahner R. Sharpe at the age of twenty two. SIGMETRICS Perform Eval Rev 2009;36(4):52–57.
8. Trivedi KS, Muppala JK, Woollet SP, *et al.* Composite performance and dependability analysis. Perform Eval 1992;14(3 & 4):197–216.
9. Shampine LF. Stiffness and nonstiff differential equation solvers, ii: detecting stiffness with Runge-Kutta methods. ACM Trans Math Softw 1977;3(1):44–53.
10. Hairer E, Nørsett SP, Wanner G. Solving ordinary differential equations i - nonstiff problems, Springer series in computational mathematics. 2nd ed. Berlin Heidelberg, Germany: Springer; 1993.
11. Reibman A, Trivedi K. Numerical transient analysis of Markov models. Comput Oper Res 1988;15:19–36.
12. Malhotra M, Muppala JK, Trivedi KS. Stiffness-tolerant methods for transient analysis of stiff markov chains. J Microelectron Reliab 1994;34:1825–1841.
13. Takacs L. Introduction to the theory of Queues. New York: Oxford University Press; 1962.
14. Muppala JK, Trivedi KS. Numerical transient solution of finite Markovian queueing systems. Queueing and Related Models (U.Bhat, I. Basawa, E h) Oxford, UK: Oxford University Press; 1992; pp. 262–284.
15. van Dijk NM. Uniformization for nonhomogeneous Markov chains. Oper Res Lett 1992;12:283–291.
16. Neuts MF. Matrix-Geometric solutions in stochastic models: an algorithmic approach. Baltimore (MD): Johns Hopkins University Press; 1981.
17. Stewart WJ. Introduction to the numerical solution of Markov chains. Princeton (NJ): Princeton University Press; 1994.
18. Gribaudo M. Hybrid formalism for performance evaluation: theory and applications [PhD dissertation]. Italy: Dipartimento di Informatica, Università di Torino; 2001.
19. Fricks RM, Puliafito A, Telek M, *et al.* Applications of Non-Markovian stochastic petri nets. SIGMETRICS Perform Eval Rev 1998;26(2):15–27.
20. Brémaud P. Markov chains - Gibbs fields, Monte Carlo simulation, and queues. Volume 31, Texts in applied mathematics. New York: Springer; 1999.
21. Kulkarni VG. Modeling and analysis of stochastic systems. London: Chapman & Hall; 1995.
22. Soyer R, Sung M. Bayesian analysis of nonhomogeneous Markov chains. Bayesian methods with applications to science, policy, and official statistics. London, UK: Eurostat; 2001. pp. 517–526.
23. Paroli R, Spezia L. Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. Comput Stat Data Anal 2008;52(5):2311–2330.
24. MacNerney JS. A linear initial-value problem. Bull Am Math Soc. Providence (RI): 1963;69(3):314–329.
25. Lambert JD. Computational methods in ordinary differential equations. London: Wiley; 1973.
26. Grassmann W. Transient solution in Markovian queueing systems. Comput Oper Res 1977;4:47–56.
27. Hairer E, Wanner G. Solving ordinary differential equations II: stiff and differential-algebraic problems, Series in computational mathematics. Berlin, Heidelberg, New York: Springer; 1991.
28. Butcher JC. Implicit Runge-Kutta processes. Math Comput 1964;18(85):50–64.
29. Birkhoff G, Varga RS. Discretization errors for well set cauchy problems. J Math Phys 1965;44:1–23.
30. Rindos A, Woollet S, Viniotis I, *et al.* Exact methods for the transient analysis of nonhomogeneous continuous-time Markov chains. Introduction to the numerical solution of Markov chains. Princeton, New Jersey: Kluwer Academic Publishers; 1995. pp. 121–133.
31. Sonin I. The asymptotic behaviour of a general finite nonhomogeneous markov chain the decomposition-separation theorem. Volume 30, Statistics, probability and game theory: papers in honor of David Blackwell (1996), Lecture notes-monograph series. Beachwood (OH): Institute of Mathematics and Statistics; 1996. pp. 337–346. <http://www.jstor.org/stable/4355954>.
32. Ciardo G, Muppala JK, Trivedi KS, Spnp: stochastic petri net package. In: PNPM'89: The Proceedings of the 3rd International Workshop on Petri Nets and Performance Models. Washington (DC): IEEE Computer Society; 1989. pp. 142–151.

TRANSIENT BEHAVIOR OF DTMCs

NATARAJAN GAUTAM
Department of Industrial and
Systems Engineering, Texas
A&M University, College
Station, Texas

Consider a discrete-time Markov chain (DTMC) which is a special type of stochastic process where state space is discrete and the state of the process is observed at discrete-time epochs. The observations possess the Markov property, that is, if the current state is given, to predict the future states, we do not need any information about the past. In other words, if X_n is the state of a process at the n th observation, then Markov property states that

$$\begin{aligned} P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \\ X_{n-2} = i_{n-2}, \dots, X_0 = i_0\} \\ = P\{X_{n+1} = j | X_n = i\}. \end{aligned}$$

For the purpose of transient analysis, we also assume that the DTMC is time-homogeneous; that is, the probability of transitioning from a state to another state does not vary with time. In other words,

$$P\{X_{n+1} = j | X_n = i\} = P\{X_1 = j | X_0 = i\}.$$

This article assumes that the reader is familiar with modeling a process as a DTMC. If that is not the case, the reader is encouraged to review the previous article from this handbook or look through one of several texts on stochastic processes [1,2]. With that in mind, the objective of this article is to obtain expressions for the distribution of the states of a time-homogeneous DTMC at some particular time in the near future. Specifically, we will compute, say, $P\{X_m = j\}$, $P\{X_r = k | X_n = i\}$, and so on, for $0 \leq m < \infty$ and $n \leq r < \infty$. For that we will first describe some terminologies and notations used throughout

the article in the section titled “Terminology and Notation.” Then in the section titled “Key Results,” we present some of the key results of transient analysis. Finally, in the section titled “Computing Issues,” we provide some techniques to efficiently perform the transient analysis from a computational standpoint. For a more thorough treatment of computational aspects of DTMCs, please see the article titled ***Computational Methods for DTMCs*** in this encyclopedia.

TERMINOLOGY AND NOTATION

We consider a time-homogeneous DTMC $X_0, X_1, X_2, \dots$ which we denote as $\{X_n, n \geq 0\}$, where X_n is the state of a process at the n th observation. The state X_n could be multidimensional and when we say $X_n = i$, then i is also assumed to be multidimensional. S is the set of all possible values that X_n can take for all n . The state space S is assumed to be a countable set, although there could be infinite elements.

Next, we define p_{ij} which is the one-step transition probability of going from state i to j . Therefore, for a time-homogeneous DTMC, for $i, j \in S$ and $n \geq 0$,

$$p_{ij} = P\{X_{n+1} = j | X_n = i\}.$$

Notice that p_{ij} is not a function of n , because the DTMC is time homogeneous. Using the transition probabilities p_{ij} for every $i \in S$ and $j \in S$, we can build a square matrix $P = [p_{ij}]$. The matrix P is called *one-step transition probability matrix*. The rows of the matrix correspond to the given current state, while the columns correspond to the next state. The elements of each row sum to 1 for every $i \in S$, that is,

$$\sum_{j \in S} p_{ij} = 1.$$

An important notation for transient analysis is $p_{ij}^{(m)}$ which is the m -step transition probability of going from state i to j . In other

words, if the current state of a process is i , then the probability that exactly after m observations, the process will be in state j is $p_{ij}^{(m)}$. By definition,

$$p_{ij}^{(m)} = P\{X_{m+n} = j | X_n = i\}.$$

Note that $p_{ij}^{(1)} = p_{ij}$, the one-step transition probability. In matrix form, we denote $P(m) = [p_{ij}^{(m)}]$, the m -step transition probability matrix with $P(1) = P$. It is crucial to point out that unlike p_{ij} , which is part of the modeling, $p_{ij}^{(m)}$ can actually be computed and it forms the core of transient analysis. But before deriving an expression for $p_{ij}^{(m)}$, we present one last notation.

Let $a_j^{(m)}$ be the probability that the DTMC is in state j at the m th observation, that is,

$$a_j^{(m)} = P\{X_m = j\}.$$

The row vector of $a_j^{(m)}$ values is denoted as $\mathbf{a}^{(m)}$, that is, $\mathbf{a}^{(m)} = [a_j^{(m)}]$. To obtain $a_j^{(m)}$, we require not only P (which is described as part of the modeling) but also the initial distribution row-vector $\mathbf{a}$ which is made up of $a_j = P\{X_0 = j\}$ such that $\mathbf{a} = [a_j]$. Although $\mathbf{a}$ is not described in DTMC modeling, for the purpose of analysis (especially $\mathbf{a}^{(m)}$), we are required to know $\mathbf{a}$. Note that $\mathbf{a} = \mathbf{a}^{(0)}$. Having described the notation, the next section develops a technique to obtain $a_j^{(m)}$ and $p_{ij}^{(m)}$ for all $m < \infty$, $i \in S$, and $j \in S$.

KEY RESULTS

The results in this section for $a_j^{(m)}$ and $p_{ij}^{(m)}$ are due to the Chapman–Kolmogorov equation which we derive first. We start with the definition of $p_{ij}^{(m)}$ for $m > 0$ and some $0 < r \leq m$, and work our way:

$$\begin{aligned} p_{ij}^{(m)} &= P\{X_{m+n} = j | X_n = i\}, \\ &= \sum_{k \in S} P\{X_{m+n} = j, X_{r+n} = k | X_n = i\}, \end{aligned}$$

$$= \sum_{k \in S} P\{X_{m+n} = j | X_{r+n} = k, X_n = i\} P\{X_{r+n} = k | X_n = i\}, \quad (1)$$

$$= \sum_{k \in S} P\{X_{m+n} = j | X_{r+n} = k\} P\{X_{r+n} = k | X_n = i\}, \quad (2)$$

$$= \sum_{k \in S} P\{X_{m-r+n} = j | X_n = k\} P\{X_{r+n} = k | X_n = i\}, \quad (3)$$

$$\begin{aligned} &= \sum_{k \in S} p_{kj}^{(m-r)} p_{ik}^{(r)}, \\ &= \sum_{k \in S} p_{ik}^{(r)} p_{kj}^{(m-r)}. \end{aligned} \quad (4)$$

It may be worthwhile to explain the above set of equations. Equation (1) is just a standard conditional probability argument that follows the law of total probability statement. Then, the first term in Equation (2) is due to the Markov property and Equation (3) follows from the fact that the DTMC is time homogeneous. Finally, Equation (4) is by definition has implications in terms of the m -step transition probability matrix. Note that Equation (4) results in

$$P(m) = P(m-r)P(r)$$

since the ij th element of $P(m)$ is the left-hand side of Equation (4) and is equal to the i th row of $P(m-r)$ multiplied by the j th column of $P(r)$.

Using the result, $P(m) = P(m-r)P(r)$ for any $m > 0$ and $r \leq m$, as well as noticing that $P(0) = I$ and $P(1) = P$, we have $P(2) = P^2$ (by writing $m = 2$ and $r = 1$). Thereby, we can recursively show that

$$P(m) = P^m$$

and therefore

$$p_{ij}^{(m)} = [P^m]_{ij}.$$

Thus, starting in state i , the probability that the DTMC is in state j after m steps can be computed by raising P to the m th power and

taking the ij th term. Next, to compute $\mathbf{a}^{(m)}$ by conditioning on the initial state, we get

$$\mathbf{a}^{(m)} = aP(m) = aP^m.$$

To get $P\{X_m = j\}$, just take the term corresponding to j , that is, $[aP^m]_j$.

We illustrate the transient analysis by means of an example. Consider three long-distance telephone companies A , B , and C . Everytime a sale is announced, users switch from one company to another. Let X_n denote the long-distance company with which a particular user named Jill is just before the n th sale announcement. We have $S = \{A, B, C\}$. On the basis of the customers' switching in the past, it is estimated that Jill's switching patterns follow the following transition probability matrix:

$$P = \begin{array}{c} \begin{array}{ccc} & A & B & C \\ \begin{array}{l} A \\ B \\ C \end{array} & \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array} \end{array}.$$

For example, if Jill is with B before a sale is announced, she would switch to A with probability 0.5 and C with probability 0.2 or stay with B with probability 0.3 as evident from the second row in P .

If at time 0, Jill is with long-distance company B , then after the third sale (i.e., $m = 3$) we wish to compute the probability that Jill is with company C . We have

$$P^3 = \begin{bmatrix} 0.3860 & 0.2770 & 0.3370 \\ 0.3930 & 0.2760 & 0.3310 \\ 0.3870 & 0.2590 & 0.3540 \end{bmatrix}.$$

What we need is

$$P\{X_3 = C | X_0 = B\} = p_{BC}^{(3)} = [P^3]_{BC} = 0.3310$$

based on the above results.

Now we know that $a = [0.25 \ 0.5 \ 0.25]$, that is, initially Jill picks A , B , or C with probability 0.25, 0.5, or 0.25, respectively. Say we are interested in computing the probability that Jill is with A after two sale announcements. On the basis of the above results, we need

$$P\{X_2 = A\} = a_A^{(2)} = [aP^2]_A = 0.385.$$

Further, using the initial distribution $a = [0.25 \ 0.5 \ 0.25]$, if we are interested in the probability that Jill is with A after the first sale, with B after the second sale, and with C after the third, then we can obtain that as

$$\begin{aligned} P\{X_1 = A, X_2 = B, X_3 = C\} \\ &= P\{X_3 = C | X_2 = B, X_1 = A\} \\ &\quad \times P\{X_2 = B | X_1 = A\}P\{X_1 = A\} \\ &= p_{BC}p_{AB}[aP]_A = 0.2 \times 0.4 \times 0.425 = 0.034 \end{aligned}$$

with the first expression being due to Markov property and the rest from the above results.

COMPUTING ISSUES

On the basis of previous section, clearly the computationally intensive operation is in obtaining P^m especially when P is large. For most problems, mathematical software such as MATLAB, mathematica, and MAPLE have efficient routines to compute P^m ; however, if one were to use a more fundamental tool such as C-programming or VBA, it would be a good idea to incorporate a routine that would efficiently obtain P^m . Having said that, in this section, we demonstrate two other techniques to obtain P^m .

Diagonalization Method

Consider a DTMC $\{X_n, n \geq 0\}$ with transition probability matrix P that is $b \times b$ square matrix, that is, the number of states in the state space S of the DTMC is b . Here, we require that b is finite. Assume that all the eigenvalues of P are distinct. Then, we can write down P as $XD X^{-1}$ where D is a diagonal matrix of all b eigenvalues of P and X is a collection of b row vectors $b \times 1$ in size that correspond to the appropriate right eigenvectors of P . We first describe a procedure to compute D and X . Let x be a right eigenvector ($b \times 1$ column vector) and d be the corresponding eigenvalue (a scalar). Then by definition of eigenvalue and eigenvector, we have $Px = dx$, that is, $(P - dI)x = 0$. Now, if we solve for the determinant $|P - dI| = 0$, we get a b -degree polynomial and upon solving we get b solutions $d_1, d_2, \dots, d_b$ which form the b eigenvalues of P . We can obtain the corresponding

eigenvectors $x_1, x_2, \dots, x_b$ by carefully solving for x_i in $Px_i = d_i x_i$ for $i = 1, 2, \dots, b$. Note that x_i is a $b \times 1$ column vector but it is not unique, so just select one arbitrarily. Once d_i and x_i are known for all $i = 1, 2, \dots, b$, we can write down

$$X = [x_1 \ x_2 \ \dots \ x_b]$$

and

$$D = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_b \end{bmatrix}.$$

Therefore, we have

$$P = XDX^{-1}$$

and hence

$$P^m = XD^mX^{-1}.$$

Notice that

$$D^m = \begin{bmatrix} d_1^m & 0 & 0 & \dots & 0 \\ 0 & d_2^m & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_b^m \end{bmatrix}.$$

The only intensive computation is in obtaining X , X^{-1} , and D . It may be worthwhile commenting on X^{-1} as there are two ways of obtaining it. One, of course, is to invert the matrix X . The other is to realize that X^{-1} is the collection of left-eigenvalues of P , stacked top to bottom corresponding to the appropriate eigenvalue.

Generating Function Method

Consider a DTMC $\{X_n, n \geq 0\}$ with transition probability matrix P that is $b \times b$ square matrix, that is, the number of states in the state space S of the DTMC is b . We can denote the generating function of P as $\Psi(z)$ for some complex scalar z . By definition,

$$\Psi(z) = I + Pz + P^2z^2 + P^3z^3 + \dots$$

and hence can be written as

$$\Psi(z) = I + Pz\Psi(z).$$

Solving for $\Psi(z)$, we get

$$\Psi(z) = (I - Pz)^{-1}$$

provided $|z| < 1$ and b is finite (although the analysis can be extended to the infinite case). From the right-hand side of the above expression, one can typically write down $[\Psi(z)]_{ij}$ of the form $\frac{f_{ij}(z)}{g_{ij}(z)}$, where both $f_{ij}(z)$ and $g_{ij}(z)$ are polynomial functions. Now, if we write down $\frac{f_{ij}(z)}{g_{ij}(z)}$ in polynomial form, it would be equivalent to $\sum_{m=0}^{\infty} p_{ij}^{(m)} z^m$ and therefore by taking the exponent multiplying z^m in $\frac{f_{ij}(z)}{g_{ij}(z)}$, we get $p_{ij}^{(m)}$.

REFERENCES

1. Kulkarni VG. Modeling and analysis of stochastic systems. Texts in Statistical Science Series. London: Chapman and Hall, Ltd; 1995.
2. Ross SM. Introduction to probability models. San Diego (CA): Academic Press; 2003.

TRANSPORTATION ALGORITHMS

DOUGLAS S. ALTNER
Department of Mathematics,
United States Naval Academy,
Annapolis, Maryland

J. PAUL BROOKS
Department of Statistical Sciences
and Operations Research,
Virginia Commonwealth
University, Richmond, Virginia

PROBLEM OVERVIEW

The transportation problem is a classic problem in operations research. The problem has numerous applications to a wide variety of real-world situations. For example, Glover and Klingman [1] list several instances where transportation problems arise as *subproblems* in mathematical models for the production and distribution of automobiles, consumer products, textiles, lumber, agriculture as well as in applications in water management and military logistics. Although the transportation problem is not as intensively studied today as it was several decades ago, the transportation industry is still the lifeblood of the global economy and there are numerous logistics problems of modern importance where the transportation problem is an essential building block of the mathematical models used to solve these problems. These include problems such as shipping coal to power plants, crude oil to refineries, waste to recycling centers, medical supplies to hospitals, locomotives to trains, or empty shipping containers to ports. The transportation problem also arises in models other than those modeling the physical transportation of goods, including applications in designing distributed data systems [2] and large-scale circuit design [3].

This article gives an overview of the various algorithms for the transportation problem. The remainder of this section defines

the transportation problem and discusses its properties, its relationship to other problems, and its early history. The second section presents the transportation simplex method, which is the most well-known method for solving transportation problems in practice. The third section surveys other algorithms for the transportation problem. The fourth section offers concluding remarks.

Definitions and Basic Properties

Transportation Problem. The *transportation problem* asks that given a finite set of suppliers each with a finite number of units in supply, a finite set of customers each with a finite number of units in demand, and a per-unit cost of transporting between every supplier–customer pair, find a cost-minimizing transportation plan to meet all customer demands without exceeding the supply of any supplier. This problem can be expressed as a linear program. Let S be the set of suppliers and C be the set of customers. The decision variables are

$x_{ij} :=$ the number of units sent from
supplier i to customer j ,

for each $i \in S$ and $j \in C$. Let c_{ij} be the cost of sending one unit from supplier i to customer j , s_i be the number of units in supply at supplier i , and d_j be the number of units in demand at customer j . This lays the groundwork for the following linear programming formulation:

Maximize $\sum_{i \in S} \sum_{j \in C} c_{ij} x_{ij}$,

Subject to

$$\sum_{j \in C} x_{ij} \leq s_i, \quad \forall i \in S,$$

$$\sum_{i \in S} x_{ij} \geq d_j, \quad \forall j \in C,$$

$$x_{ij} \geq 0, \quad \forall i \in S, \forall j \in C.$$

Every transportation problem has an associated *transportation graph*, which is a directed graph where every supplier and

every customer has a corresponding node and there is an arc (i, j) between each supplier node i and each customer node j . In other words, the transportation graph is a directed complete bipartite graph where every arc originates at a supplier node and terminates at a customer node.

A *balanced transportation problem* is a transportation problem where $\sum_{i \in S} s_i = \sum_{j \in C} d_j$. Otherwise a transportation problem is *unbalanced*. Since any supplier can transport goods to any customer, a balanced transportation problem always has a feasible solution. A transportation problem where $\sum_{i \in S} s_i > \sum_{j \in C} d_j$ can be converted into a balanced transportation problem by adding a *dummy customer* $|C| + 1$ with demand $d_{|C|+1} = \sum_{i \in S} s_i - \sum_{j \in C} d_j$ that can absorb all of the excess demand. The cost of shipping one good from a supplier to a dummy customer is set to be equal to the cost of not shipping from that supplier. A transportation problem where $\sum_{i \in S} s_i < \sum_{j \in C} d_j$ is *infeasible* since there is not enough supply to satisfy the demands of all of the customers. Any infeasible transportation can be converted into a balanced transportation problem by adding a *dummy supplier*. The cost of shipping one good from the dummy supplier to a customer is set to be equal to the per-unit cost of not fulfilling that customer's demand.

Transshipment Problem. The *transshipment problem* is a generalization of the transportation problem that includes the possibility of *transshipment points*. Transshipment points T can receive flow from either suppliers or other transshipment points and they can send flow to either customers or other transshipment points. Transshipment points do not carry a supply nor do they have a demand, so flow balance must be maintained at every transshipment point.

The transshipment problem can be modeled with the following linear program:

$$\begin{aligned} & \text{Maximize} && \sum_{i \in S \cup T} \sum_{j \in T \cup C} c_{i,j} x_{i,j}, \\ & \text{Subject to} && \sum_{j \in T \cup C} x_{i,j} \leq s_i, \quad \forall i \in S, \end{aligned}$$

$$\begin{aligned} & \sum_{i \in S \cup T} x_{i,j} \geq d_j, \quad \forall j \in C, \\ & \sum_{i \in S \cup (T \setminus \{k\})} x_{i,k} \\ & - \sum_{j \in (T \setminus \{k\}) \cup C} x_{k,j} = 0, \quad \forall k \in T, \\ & x_{i,j} \geq 0, \quad \forall i \in S \cup T, \quad \forall j \in T \cup C. \end{aligned}$$

Input data S, C, s_i , and d_j are defined as they are in the transportation problem. The variable $x_{i,j}$ is redefined to be the number of units sent from location i to location j for each $i \in S \cup T$ and each $j \in T \cup C$ and $c_{i,j}$ is redefined similarly.

Each instance of the transshipment problem has a corresponding *transshipment graph*. There is a node for each supplier, transshipment point, and customer. In addition, there are arcs between each supplier–transshipment point pair, transshipment point–transshipment point pair, and transshipment point–customer pair.

Relationship to Other Problems. The assignment problem is a special case of the transportation problem, where $s_i = 1$ for each $i \in S$ and $d_j = 1$ for each $j \in C$. An instance of the transportation problem with integer supplies and demands can be converted into an instance of an assignment problem as follows. Replace each supplier i with s_i separate suppliers each with unit supply and replacing each customer j with d_j separate customers each with unit demand.

The minimum cost flow problem is a generalization of the transshipment problem where each arc can have both a lower bound and an upper bound on the amount of flow it can carry. In addition, the minimum cost flow problem allows arcs between any pair of nodes, whereas the transshipment problem cannot have arcs between two supply nodes or two demand nodes.

Transformations between Transportation Problem and Transshipment Problem. The transportation problem is a special case of the transshipment problem where there are no transshipment points. Moreover, a transshipment problem with nonnegative costs can be formulated as a transportation

problem. Several transformations are known [4,5]; we describe one of the transformations due to Orden [5]. Replace each transshipment point t with a new supplier u_t and a new customer v_t . Both the supply at u_t and the demand at v_t are set to $\sum_{i \in S} s_i$. The cost of sending one unit from a supplier or transshipment node i to v_t is set to $c_{i,t}$ and the cost of sending one unit from u_t to transshipment node or customer j is set to $c_{t,j}$. The cost of sending one unit from u_t to v_t is set to zero.

The intuition for how this transformation works is that each unit of flow entering a newly created customer v_t must have a corresponding unit of flow that leaves u_t . Thus, if there is a unit of flow that is sent from a supplier or transshipment node i to v_t and from u_t to a transshipment node or customer j , this is equivalent to a unit of flow being sent from i to t to j in the original transshipment problem. For any t , a unit of flow sent from u_t to v_t does not correspond to any action in the original transshipment problem and does not impact the cost of the transportation plan. Hence, setting the supply of u_t and the demand of v_t equal to $\sum_{i \in S} s_i$ is an upper bound on the amount of flow that can go through t .

Alternatively, the transshipment problem can also be reduced to the transportation problem by deleting all of the transshipment nodes and setting $c_{i,j}$ equal to the length of the shortest path from supplier i to customer j in the original transshipment graph. This reduction is polynomial if all costs are nonnegative. (This reduction is not polynomial if the costs can be negative, since the shortest path between two nodes cannot be computed in polynomial time if arc costs are negative [6].)

Properties. The constraint matrices of the transportation problem and the transshipment problem are totally unimodular. This implies, by a theorem by Hoffman and Kruskal [7], that every optimal extreme point solution to these problems is integral when the supply and demand quantities are integral. Thus, any optimal basic feasible solution to the linear programming formulation of the transportation problem or the

transshipment problem is guaranteed to have all integer values when supplies and demands are integer valued.

It is also true that a feasible solution to a balanced transportation problem is optimal if and only if there is not a *negative cost cycle* in the underlying undirected graph. In this context, a negative cost cycle is an alternating sequence of suppliers and customers:

$$s_{p(1)}, c_{q(1)}, s_{p(2)}, c_{q(2)}, \dots, s_{p(k)}, c_{q(k)}, s_{p(1)},$$

where customer $c_{q(i)}$ currently receives at least one unit from supplier $s_{p(i+1)}$, for each $i \in \{1, 2, \dots, k\}$ and $\sum_{i=1}^k c_{p(i),q(i)} < \sum_{i=1}^k c_{p(i+1),q(i)}$, where $p(k+1) = p(1)$. Such a sequence reveals how to modify the current solution to reduce its total cost. Sending one unit of supply from supplier $s_{p(i)}$ to customer $c_{q(i)}$ and sending one fewer unit of supply from supplier $s_{p(i+1)}$ to customer $c_{q(i)}$, for each $i \in \{1, 2, \dots, k\}$.

The solid arcs in Fig. 1 represent a negative cost cycle. Specifically, this cycle corresponds to the sequence: C, Y, A, X, C . To send flow around this cycle, one can send one additional unit of flow from supplier C to customer Y , send one fewer unit of flow from supplier A to customer Y , send one more unit of flow from supplier A to customer X , and send one fewer unit of flow from supplier C to customer X . Each unit of flow sent around this cycle decreases the total transportation cost by $10 - 6 + 8 - 7 = 5$ units, making it a negative cost cycle.

Both the transportation problem and the transshipment problem are strongly NP-hard if the transportation costs can be negative. This is because the strongly NP-hard shortest path problem in general networks [6] (i.e., networks where arc weights can be positive or negative) can be polynomially reduced to the transshipment problem in general networks, which in turn can be reduced to the transportation problem in general networks (see the section titled “Transformations between Transportation Problem and Transshipment Problem”). Intuitively speaking, the basic reason why the transshipment problem in general networks is NP-hard is because it is nontrivial to eliminate negative cost subtours from a solution.

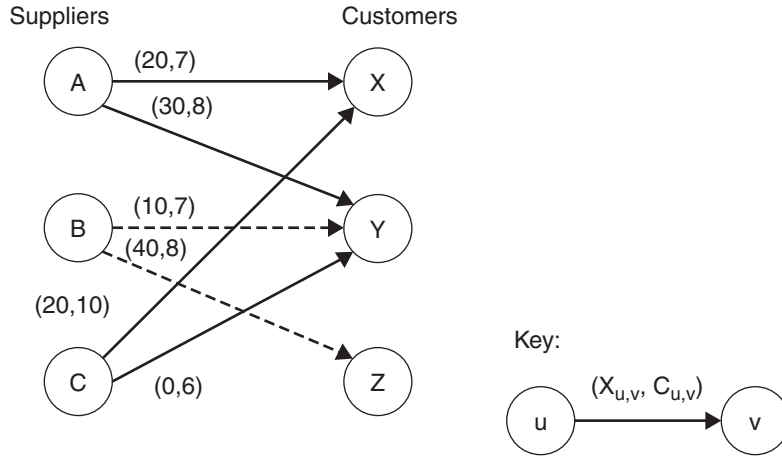


Figure 1. Example of a negative cost cycle.

Early History

Schrijver [8] contains extensive notes on the early history of the transportation problems, which are summarized here. Tolstói [9] is the earliest known example of a formal study of some form of the transportation problem, which he uses to model the problem of transporting cargo over the Soviet railroad network. Kantorovich [10] and Kantorovich and Gavurin [11] also present work on precursors to the transportation problem that were motivated by logistic problems in the Soviet railroad network. However, none of these technical reports were known to Western researchers until the 1950s.

The earliest publication on the transportation problem in Western literature is by Hitchcock [12], which is why the transportation problem is often called the *Hitchcock problem*. This paper includes a linear programming formulation of the transportation problem, a simplex-like method for solving it, complementary slackness optimality conditions, and a rule for computing an initial feasible solution. The transportation problem

is also independently introduced and studied by Koopmans [13] as part of his work for a British-American merchant shipping agency during the Second World War.

TRANSPORTATION SIMPLEX METHOD

In this section, we describe a generic transportation simplex method for balanced transportation problems. Recall that any transportation problem that is not balanced can be transformed into a balanced transportation problem by the addition of a dummy supplier or customer node, as appropriate within the problem context.

Overview of Method

Algorithm 1 gives an overview of the transportation simplex method. It is essentially the same as the regular simplex method for linear programming but it exploits the special structure of the constraint matrix of the balanced transportation problem.

Algorithm 1. Transportation Simplex Method

Start with an initial basic feasible solution

while Reduced cost optimality conditions are not satisfied **do**
 Find an entering basic variable

Determine a cycle of transfers
 Use that cycle to find a leaving basic variable
 Update current basic feasible solution
end while

return Optimal basic feasible solution

Table 1. Generic Transportation Simplex Tableau Cell

	Customer j	Supply
Supplier i	x_{ij} c_{ij}	s_i
Demand	d_j	

Example

In this section, we present an example of a single iteration of the transportation simplex method from Winston and Venkataramanan [14]. Table 1 introduces the standard layout used in a single cell of a *transportation simplex tableau*. The value of the decision variable x_{ij} is contained in the center of the cell in the i th row and the j th column and the value of the cost coefficient c_{ij} is contained in the upper right-hand corner of that cell.

Current Basic Feasible Solution. We solve the transportation problem with data given in Table 2 starting with the current basic feasible solution listed. By convention, the values of the x_{ij} variables listed in the tableau are the basic variables and the values not listed are the nonbasic variables. Nonbasic variables are set to zero.

Checking the Reduced Costs. As with the standard simplex method, the first step of an iteration is to check the reduced costs of all of the nonbasic variables. With the standard simplex method, the reduced costs are given by the formula $c_B B^{-1} N - c_N$, where B is the basis matrix corresponding to the current basic feasible solution, N is the matrix of constraint coefficients corresponding to the nonbasic variables, c_B is the vector of objective coefficients of the basic variables, and c_N is the vector of objective coefficients of the nonbasic variables. (Our convention for

the sign of the reduced cost follows that of Winston and Venkataramanan [14]; the reduced cost is frequently defined as having the opposite sign of the quantity used here.) The reduced costs are given by the same formula for the transportation simplex method. However, due to the special structure of the constraint matrix, the reduced cost for a nonbasic variable x_{ij} simplifies to $u_i + v_j - c_{ij}$, where u_i is the dual variable associated with supplier i 's constraint and v_j is the dual variable associated with customer j 's constraint.

The reduced cost of each basic variable is zero. This fact yields the following system of equations that can be used to solve for the dual variables:

$$\begin{aligned} \bar{c}_{1,1} = u_1 + v_1 - 8 &= 0 & \bar{c}_{2,3} = u_2 + v_3 - 13 &= 0 \\ \bar{c}_{2,1} = u_2 + v_1 - 9 &= 0 & \bar{c}_{3,3} = u_3 + v_3 - 16 &= 0 \\ \bar{c}_{2,2} = u_2 + v_2 - 12 &= 0 & \bar{c}_{3,4} = u_3 + v_4 - 5 &= 0, \end{aligned}$$

where $\bar{c}_{ij}$ denotes the reduced cost of variable x_{ij} . For any transportation problem with n basic variables, there are $n + 1$ dual variables. This means the above system always consists of n equations with $n + 1$ unknown variables. Mathematically, the transportation simplex algorithm is still correct if one of the dual variables is arbitrarily set to zero. This property follows from the fact that any one of the constraints is linearly dependent upon the others, and may be deleted without affecting the set of feasible solutions. Deleting a constraint is equivalent to setting its dual variable to zero. In our example, let $u_1 = 0$. The following values for the dual variables are obtained after solving the six equations above: $u_2 = 1, u_3 = 4, v_1 = 8, v_2 = 11, v_3 = 12$, and $v_4 = 1$.

Using these dual variable values, the reduced cost of the nonbasic variables can

Table 2. Starting Basic Feasible Solution for Transportation Simplex Example

	Customer 1	Customer 2	Customer 3	Customer 4	Supply
Supplier 1	35	8	10	9	35
Supplier 2	10	20	13	7	50
Supplier 3		9	16	5	40
Demand	45	20	30	30	

now be computed using the following system of equations:

$$\begin{aligned}\bar{c}_{1,2} &= u_1 + v_2 - 6 = 5 & \bar{c}_{2,4} &= u_2 + v_4 - 7 = -5 \\ \bar{c}_{1,3} &= u_1 + v_3 - 10 = 2 & \bar{c}_{3,1} &= u_3 + v_1 - 14 = -2 \\ \bar{c}_{1,4} &= u_1 + v_4 - 9 = -8 & \bar{c}_{3,2} &= u_3 + v_2 - 9 = 6.\end{aligned}$$

The transportation problem is optimal if the reduced costs of all of the decision variables are nonpositive. Thus, in our example, the current basic feasible solution is not optimal.

Performing a Pivot Operation. Since the current basic feasible solution is not optimal, an *entering basic variable* must be selected. The variable with the largest (positive) reduced cost is often chosen to be the entering basic variable, which for our example is $x_{3,2}$. However, the transportation simplex method is still correct if any variable with a positive reduced cost is selected as the entering basic variable. The section titled “Choosing an Initial Basic Feasible Solution” discusses different rules for selecting an entering basic variable.

We now describe how to select a leaving basic variable. As the value of the entering variable is increased, the objective function value improves. The transportation simplex method increases the value of the entering basic variable as much as possible until a constraint is violated. During this process, the basic variables may be adjusted as required, but the other nonbasic variables must remain fixed at zero.

For instance, suppose $x_{3,2}$ is increased by Δ . This means that supplier 3 is directly

sending Δ units to customer 2, which means that supplier 3 has Δ fewer units to send to other customers. Suppose supplier 3 reduces the number of units it sends to customer 3 by Δ . (Supplier 3 cannot decrease the amount it supplies to customer 4 because there are no other suppliers that can send an additional Δ units to customer 4. Nonbasic variables $x_{1,4}$ and $x_{2,4}$ must remain fixed at 0.) Customer 3, requiring 30 units, needs to receive an extra Δ units from another supplier. Since $x_{1,3}$ must remain fixed at 0, these Δ additional units must come from supplier 2. Moreover, since supplier 2 must supply an additional Δ units to customer 3, it now has Δ fewer units to supply to other customers. However, if supplier 2 sends Δ fewer units to customer 2, then customer 2 still receives the 20 units of supply that he/she requires since he/she is now scheduled to receive $20 - \Delta$ units from supplier 2 and Δ units from supplier 3.

In summary, the solution has been modified by redefining four variables to the following values: $x_{3,2} = \Delta$, $x_{3,3} = 10 - \Delta$, $x_{2,3} = 20 + \Delta$ and $x_{2,2} = 20 - \Delta$. All other variables remain fixed to their current values. Table 3 illustrates this *transfer cycle* and the subsequent values of the decision variables as a function of Δ .

Note the supply and demand constraints remain satisfied for any value of Δ . However, the nonnegativity constraints could be violated for a sufficiently large value of Δ . In this case, Δ is set to 10 since this is the largest value Δ can assume before a basic variable assumes a negative value.

Once the basic variables are updated, $x_{3,3}$ is the basic variable that uniquely goes to zero, which makes it the only possible choice

Table 3. Finding a Transfer Cycle

	Customer 1	Customer 2	Customer 3	Customer 4	Supply
Supplier 1	8 35	6	10	9	35
Supplier 2	9 10	12 $20 - \Delta$	13 $20 + \Delta$	7	50
Supplier 3	14	9 Δ	16 $10 - \Delta$	5 30	40
Demand	45	20	30	30	

Table 4. New Basic Feasible Solution for Transportation Simplex Example

	Customer 1	Customer 2	Customer 3	Customer 4	Supply
Supplier 1	8 35	6	10	9	35
Supplier 2	9 10	12 10	13 30	7	50
Supplier 3	14	9 10	16	5 30	40
Demand	45	20	30	30	

for the *leaving basic variable*. If multiple basic variables go to zero for a given value of Δ , then one of them can be arbitrarily selected to be the leaving basic variable while the others remain in the basis with a value of 0 (these variables are said to be *degenerate*). An iteration of the transportation simplex method is concluded once a leaving basic variable is removed from the basis. Table 4 contains the new basic feasible solution that is obtained at the end of this iteration.

The transportation simplex method requires additional iterations before termination for this example because at least one of the new reduced costs is positive. The remaining iterations are omitted from this article.

When searching for a leaving basic variable, the transfer cycle created is always unique. This is because the subgraph of the transportation graph that only contains the arcs corresponding to the basic variables is a spanning tree [15]. Thus, since adding the arc to a spanning tree creates a unique cycle, the corresponding transfer cycle is also unique.

Notes on the Transportation Simplex Method

In this section, we present several additional comments on the transportation simplex method. First, we present the advantages of using the transportation simplex method over using methods designed for general linear programs. We then discuss different methods for constructing an initial basic feasible solution. We then discuss different rules for choosing an entering basic variable. Finally, we address the issue of degeneracy.

Advantages of the Transportation Simplex Method over Linear Programming Methods.

One immediately apparent advantage of the transportation simplex method over the standard simplex method for linear programming is that inverse of a basis matrix does not need to be computed and stored. Instead, the transportation simplex method requires the less costly operation of computing reduced costs by solving a simple system of equations (see the section titled “Checking the Reduced Costs”).

Computational studies have also demonstrated that the transportation simplex method is much more effective for transportation problems than the standard simplex method. For example, the landmark computational study of Glover *et al.* [16] concluded that their implementation of the transportation simplex method was over 140 times faster than the standard simplex method.

Choosing an Initial Basic Feasible Solution.

There are several methods for constructing an initial basic feasible solution to a transportation problem. One such method is the northwest corner method, which was first introduced by Dantzig [17]. This method starts in the northern-most and western-most corner of the transportation tableau (i.e., with variable $x_{1,1}$). While the method is considering variable $x_{i,j}$ it sets $x_{i,j}$ equal to the minimum of the supply remaining at supplier i and the unfulfilled demand of customer j . If $x_{i,j}$ is set equal to supplier i 's remaining supply, then the method next moves to variable $x_{i+1,j}$. Likewise, if $x_{i,j}$ is set equal to customer j 's remaining demand, then the method next moves to variable $x_{i,j+1}$. Furthermore, if supplier i 's remaining supply equals customer j 's remaining demand, then $x_{i,j}$ is set to this value. Either $x_{i+1,j}$ or $x_{i,j+1}$ is selected to be a basic variable that is set to zero. The method then proceeds to $x_{i+1,j+1}$. The method continues until all supply is exhausted and all customer demand is satisfied.

Table 5 contains an example of an initial basic feasible solution that is constructed using the northwest corner method.

This example is taken from Winston and Venkataramanan [14]. Variable $x_{1,1}$ takes the value 2 since 2 is the minimum of supplier 1's supply, which is 5, and customer 1's demand, which is 2. Since customer 1's demand is satisfied, the method next moves to variable $x_{1,2}$. Variable $x_{1,2}$ is assigned 3 since this is the minimum of the remaining supply at supplier 1, which is 3 units, and the demand of customer 2, which is 4 units. This method continues as described above until a complete basic feasible solution is constructed.

There are other methods for constructing an initial basic feasible solution. Vogel's method [18] greedily constructs a solution by iteratively selecting a supplier (or customer) with remaining supply (or demand) that has the largest difference between its smallest associated cost coefficient and its second smallest associated cost coefficient. Dennis [19] describes the row-column minimum method, which is also known as the minimum cost method Winston and Venkataramanan [14]. This method iteratively increases the decision variable with the smallest cost coefficient whose associated supplier still has unused supply and the associated customer still has unsatisfied demand. Hadley [20] describes the row minimum method that iteratively selects the decision variable with the smallest cost coefficient from a single supplier (i.e., tableau row) until the supply from that supplier is exhausted. Glover *et al.* [16] introduce a modified row minimum method that is very similar to the row minimum method but only selects one decision variable whenever a supplier is selected.

Table 5. Illustration of the Northwest Corner Method

	Customer 1	Customer 2	Customer 3	Customer 4	Supply
Supplier 1	$c_{1,1}$ 2	$c_{1,2}$ 3	$c_{1,3}$	$c_{1,4}$	5
Supplier 2	$c_{2,1}$	$c_{2,2}$ 1	$c_{2,3}$	$c_{2,4}$	1
Supplier 3	$c_{3,1}$	$c_{3,2}$ 0	$c_{3,3}$ 2	$c_{3,4}$ 1	3
Demand	2	4	2	1	

Glover *et al.* [16] test the five aforementioned methods. Their study concludes that the transportation simplex algorithm terminates with an optimal solution in the least amount of time when either the modified row minimum method or the row minimum method are employed. They also conclude that the northwest corner method is the quickest of the five rules to execute but the transportation simplex method requires roughly three times as many iterations when it starts with the northwest corner method than when it starts with the popular Vogel's method. Finally, they conclude that the transportation simplex method tends to require the fewest number of iterations when the starting solution is constructed with Vogel's method; however, the time required to construct the initial solution with Vogel's method is much larger than what is required with other methods. They also demonstrate that even though there are fewer pivots needed when Vogel's method is used, the overall time for the transportation simplex method to terminate with an optimal solution is greater when it is started with Vogel's method than when it is started with the modified row minimum method. Similar conclusions are drawn in a concurrent study by Srinivasan and Thompson [21].

In a subsequent study, Gavish *et al.* [22] suggest a modification to the modified row minimum method. Their method uses Glover's method to select the initial basic variables that assume a nonzero value. The remaining basic variables (whose values are zero) are iteratively selected by choosing the variables with the smallest cost coefficients that preserve the tree structure of the basis.

Choosing an Entering Basic Variable.

Glover *et al.* [16] also study different ways to select an entering basic variable. They conclude that the best method is to compute the reduced costs for the nonbasic variables associated with a single supplier and to only compute the reduced costs for the nonbasic variables associated with a different supplier if those associated with the previously checked supplier(s) are all nonpositive. However, when a supplier with an associated nonbasic variable with a

positive reduced cost is encountered, then the associated nonbasic variable with the most positive reduced cost should be chosen as the entering basic variable. This pivoting rule is in contrast to the standard Dantzig pivoting rule, which selects the nonbasic variable with the most positive reduced cost across *all* suppliers as the entering basic variable. Srinivasan and Thompson [21] also reach a similar conclusion about the best rule for selecting an entering basic variable in a concurrent study.

Gavish *et al.* [22] describe a complex rule for selecting an entering basic variable that involves only considering a subset of the decision variables as potential entering basic variables. The remaining variables are considered only if the current solution cannot be improved by entering any of the variables from the subset into the basis. The authors show that their pivoting rule, when combined with a few other heuristics, leads to an implementation of the transportation simplex method that outperforms the implementation of Srinivasan and Thompson [21] by an order of magnitude.

Degeneracy. Degeneracy is always a problem with simplex methods since a degenerate pivot requires the same computational costs as a regular pivot but it does not result in a solution with a better objective value. Gavish *et al.* [22] study the frequency of degenerate pivots during the transportation simplex method. They conclude that the percentage of pivots that are degenerate can be quite high, as their study found it to range from 8% for 5×5 transportation problems to 89% for 250×250 transportation problems. In addition, Gavish *et al.* show that the percentage of degenerate pivots is not sensitive to the cost coefficients but is sensitive to the range of possible levels of supply and demand. Specifically, they conclude that there is a high percentage of degenerate pivots when this range is very small.

With regards to remedies for degeneracy, Dantzig [17] describes a simple perturbation procedure that forces all basic feasible solutions to be nondegenerate. Gavish *et al.* [22] describe a simple rule for breaking ties when selecting a leaving basic variable: of all

possible leaving basic variables, choose the one with the largest cost coefficient. As previously stated, they show that their leaving basic variable tie-breaking rule, when combined with their other heuristics, yields an implementation of the transportation simplex method that outperforms the results obtained by Srinivasan and Thompson [21] by an order of magnitude.

SURVEY OF OTHER TRANSPORTATION ALGORITHMS

In this section, we offer a survey of other algorithms for the transportation problem. Research on transportation algorithms has passed through roughly three phases. The first phase consisted of researchers focusing on developing algorithms to solve transportation problems by hand. The second phase primarily consisted of researchers developing good implementations of existing algorithms to solve transportation problems on a computer. The third phase primarily consisted of researchers striving to improve the best-known worst-case algorithmic bound for a transportation problem. As the transportation problem is a special case of the minimum cost flow problem, algorithms for minimum cost flow problem can be used to solve instances of the transportation problem. Rather than present algorithms for the more general problem, we focus on specialized algorithms for the transportation problem and refer the reader to the article *Minimum Cost Flows* for a discussion of the algorithms we omit.

We partition this survey into three sections that mirror the three main phases of research on transportation problems: initial algorithmic development, computational studies, and worst-case analysis. Previous surveys on this topic include Gass [23], which surveys several computational studies of transportation algorithms. In addition, Schrijver [8] contains a survey of the best-known worst-case analysis bounds for algorithms for the transportation problem.

Initial Algorithm Development

The first transportation simplex method was engineered by Dantzig [17] and this

method is presented in the previous section. Dantzig's transportation simplex method is alternatively called the modified distribution (MODI) method as well as the row-column sum method.

Charnes and Cooper [24] subsequently developed the stepping-stone method, which is similar in essence and simpler in explanation. The key difference is that the stepping-stone method forgoes computing the values of the dual variables and the reduced costs explicitly. Instead, it considers any nonbasic variable as a potential entering basic variable, then determines the transfer cycle that results from entering that variable into the basis and then checks if sending flow along this cycle improves the objective function. Thus, the stepping-stone method could potentially compute many transfer cycles within a single iteration of the method, whereas the transportation simplex method always computes exactly one transfer cycle during each iteration.

Charnes and Cooper note that Dantzig's transportation simplex method has certain advantages during large-scale calculations. Although Charnes and Cooper do not specify these advantages, one such advantage is surely that Dantzig's method does not iteratively compute potentially large transfer cycles for many nonbasic variables until an improving entering basic variable can be identified.

Glicksman *et al.* [25] present a modified primal transportation simplex algorithm for a transportation problem with few suppliers and many customers. Their algorithm incorporates a tree structure to represent basic feasible solutions that allows them to efficiently determine a leaving basic variable. Harris [26] also develops a primal transportation simplex algorithm for problems with many customers and few suppliers. Langley *et al.* [27] create and test a primal simplex code for the related capacitated transportation problem.

There are several other algorithms that were developed throughout the 1950s and the 1960s for the transportation problem. Ford and Fulkerson [28] and Munkres [29] create new transportation algorithms by generalizing Kuhn's Hungarian method [30] for the

assignment problem. These transportation algorithms maintain a dual feasible solution and solve a restricted primal problem using only the primal variables that are nonzero in the complementary primal solution (that comes from the complementary slackness optimality conditions). Ford and Fulkerson [28] show that the restricted primal problem can be solved as a maximum flow problem.

Ford and Fulkerson [31] propose the *out-of-kilter* method for the capacitated transportation problem, where primal and dual feasible solutions are maintained and changed until conditions equivalent to complementary slackness conditions are satisfied. Graves and Thrall [32] also propose an out-of-kilter algorithm. Balinski and Gomory [33] introduce a primal version of a generalized Hungarian method for transportation problems. At each step, primal feasibility and complementary slackness are preserved, and information from the corresponding dual variables is used to avoid degenerate pivots. Grigoriadis and Walker [34] and Gass [35] propose decomposition procedures that maintain dual feasibility and strive to achieve primal feasibility. Williams [36] applies the reformulation technique of Dantzig and Wolfe [37] to a class of transportation problems and then solves the problems using a modified simplex method.

Computational Studies

After the wave of linear programming-based algorithms, the general focus of the literature shifts to determining the best implementations of transportation algorithms to solve problems on a computer in the least amount of time. These publications mostly appeared in the 1970s and they facilitated the establishment of general standards for programming languages, computational hardware, and algorithm evaluation.

During the 1960s, the out-of-kilter algorithm was considered the fastest algorithm for solving transportation problems [38, p. 109; 16, p. 796]. This perception changed after the publication of several studies in the 1970s, most notably those of Srinivasan and Thompson [21] and Glover *et al.* [16], which present primal transportation simplex

codes that found provably optimal solutions faster than the most highly regarded implementation of the out-of-kilter algorithm (i.e., Clasen's implementation [39]). Other findings of these studies are discussed in the section titled "Notes on the Transportation Simplex Method". Hatch [40] performs a subsequent study that concludes that primal-dual transportation algorithms are faster than primal transportation simplex algorithms, but the validity of the conclusions are disputed in a letter to the editor by Glover and Klingman [41] who argue that primal transportation simplex algorithms are still faster.

Bradley *et al.* [42] develop and test an implementation of the primal transportation simplex method for the capacitated transshipment problem. Their code incorporates specially designed data structures for basis representation, pivoting and computing reduced costs. Miller *et al.* [43], Barr and Hickman [44], and Li and Zenios [45] produce primal transportation algorithms designed to operate in a parallel computing environment and discuss sources of time savings from parallelization.

Worst-Case Analysis

Edmonds and Karp [46] and Dinits [47] describe the first (weakly) polynomial-time algorithm for the transportation problem. Their approach is based on solving a series of shortest path problems, which is similar in spirit to the previous approaches of Hoffman and Markowitz [48] and Lageman [49]. Edmonds and Karp employ a technique called *scaling* where only recirculations of flow that are "sufficiently large" as determined by a parameter are considered. The parameter is iteratively reduced throughout the algorithm until an optimal solution is obtained. The work of Edmonds and Karp inspired a number of other weakly polynomial scaling algorithms, including those of Gabow [50], Gabow and Tarjan [51], and Goldberg and Tarjan [52]. The first strongly polynomial algorithm for the transportation problem is presented by Tardos [53]. (Informally, an algorithm is *strongly* polynomial if the running time does not depend on the size of the numbers in the

input). Other strongly polynomial algorithms are given by Galil and Tardos [54], Goldberg and Tarjan [52], and Orlin [55].

The scaling algorithms involve either scaling of capacities, in which case a shortest path subroutine is required, or scaling of costs, in which case a maximum flow subroutine is required. The method presented by Orlin [55] is a cost-scaling method and does not require multiplication or division of numbers. Although Orlin's algorithm is designed for the more general uncapacitated minimum cost flow problem, it currently has the best worst-case algorithmic bound among known transshipment algorithms (though it does not have the best bound among transportation algorithms). Specifically, the bound is $O(n \log n \text{SP}_+(n, m, K))$, where n is the total number of nodes, m is the number of arcs, K is the largest shipping cost, and $\text{SP}_+(n, m, K)$ is the running time of an algorithm for finding a shortest path in a network with n nodes, m arcs, and with largest arc cost K .

A number of investigators propose algorithms for the transportation problem, where the number of suppliers is different from the number of customers. Bertsekas and Castanon [56] adapt an auction algorithm for the assignment problem to a transportation problem where the objective is to maximize the utility of the transportation plan. They demonstrate that their algorithm outperforms minimum cost flow codes on instances with significantly more customers than suppliers. The auction algorithm maintains dual feasibility and approximate complementary slackness conditions, while updating one or more dual variable values at each iteration. This algorithm is called an *auction algorithm* because the dual variables can be interpreted as the prices that suppliers bid to supply the customers with goods. Matsui [57] presents an algorithm for the transportation problem where the number of suppliers is small and fixed. A worst-case analysis of Matsui's algorithm shows that it runs in $O(n)$ time, where n is the number of customers but it runs in $O((m!)^2 n)$ time with m suppliers. Ahuja *et al.* [58], Tokuyama and Nakano [2], Kleinschmidt and Schannath [59], and Brenner [60] develop variants

of the scaling transportation algorithms for instances when the difference between the number of suppliers and customers is sufficiently large, and which have better worst-case algorithmic bounds than those of nonspecialized algorithms. Tokuyama and Nakano also equate the transportation problem to the geometric problem of finding a minimum-weight partitioning of points in $\mathbb{R}^k$, where k is the number of customers, and then present a randomized algorithm that improves the worst-case bound when the ratio between the minimum and maximum supply is sufficiently small.

CONCLUSIONS

Although transportation algorithms were once a popular research topic, research on problem-specific algorithms for the transportation and transshipment problems, the pace has declined in recent years. Part of this decline can probably be attributed to the similarity between the transportation, the transshipment, and the minimum cost flow problem. Also, implementations of transportation algorithms were able to solve large instances very quickly. For example, Glover *et al.* [16]'s implementation solved 1000×1000 transportation problems in a matter of seconds in the 1970s.

This is not to say that there is no opportunity for advancement in transportation algorithms. For example, Brenner *et al.* [3] details work on problems arising in chip design that can be modeled as transportation problems with millions of suppliers and several dozen customers. Such real-world instances provide motivation for examining transportation problems of much greater scale. There has been a dramatic advancement in computing power since the last major computational studies of transportation algorithms. Because of the interest in solving larger instances and because of newly available computing power, a new computational study of transportation algorithms could be of value to the research community.

In addition, Goldberg and Tarjan [52] note that the early implementations of scaling algorithms did not fare well against

traditional primal and primal–dual implementations, but they did not rule out the possibility of better implementations of scaling algorithms. Continuing to study the best implementation of scaling algorithms for transportation problems is another possible area of research.

REFERENCES

1. Glover F, Klingman D. Real world applications of network related problems and breakthroughs in solving them efficiently. *ACM Trans Math Soft* 1975;1:47–55.
2. Tokuyama T, Nakano J. Efficient algorithms for the Hitchcock transportation problem. *SIAM J Comput* 1995;24:563–578.
3. Brenner U, Struzyna M, Vygen J. BonPlace: placement of leading-edge chips by advanced combinatorial algorithms. *IEEE Trans on Comp-Aid Des Int Cir & Syst* 2008;27:1607–1620.
4. Fulkerson DR. On the equivalence of the capacity-constrained transshipment problem and the Hitchcock problem. Research Memorandum RM-2480. Santa Monica (CA): The RAND Corporation; 1960.
5. Orden A. The transshipment problem. *Manage Sci* 1955;2:276–285.
6. Ahuja RK, Magnanti TL, Orlin JB. Network flows: theory, algorithms, and applications. Upper Saddle River (NJ): Prentice Hall; 1993.
7. Hoffman AJ, Kruskal JB. Integral boundary points of convex polyhedra. In: Kuhn HW, Tucker AW, editors. *Linear inequalities and related systems*. Princeton (NJ): Princeton University Press; 1956. pp. 223–246.
8. Schrijver A. Combinatorial optimization: polyhedra and efficiency. Berlin: Springer; 2003.
9. Tolstoĭ A. Methods of finding the minimal total kilometrage in cargo-transportation planning in space. *Planirovanie Perevozok, Sbornik pervyi* 1930;1:23–55.
10. Kantorovich LV. On the translocation of masses. *Comput Rend Akad Sci (in Russian)* 1942;37:199–201.
11. Kantorovich LV, Gavurin MK. The application of mathematical methods to problems of freight flow analysis. In: Zvonkov VV, editor. *Collection of problems concerned with increasing the effectiveness of transports*. Moscow, Leningrad: Akademii Nauk SSSR; 1949. pp. 110–138. (In Russian.)
12. Hitchcock FL. The distribution of a product from several sources to numerous localities. *J Math Phys* 1941;20:224–230.
13. Koopmans TC. Exchange ratios between cargoes on various routes (non-refrigerating dry cargoes). Memo for the Combined Shipping Adjustment Board. Washington (DC); 1942. pp. 1–12 [first published in: *Scientific Papers of Tjalling C. Koopmans*. Berlin: Springer; 1970. pp. 77–86].
14. Winston WL, Venkataramanan V. Introduction to mathematical programming. Pacific Grove (CA): Duxbury Press; 2003.
15. Johnson EL. Networks and basic solutions. *Oper Res* 1966;14:619–623.
16. Glover F, Karney D, Klingman D, *et al*. A computational study on start procedures, basis change criteria and solution algorithms for transportation problems. *Manage Sci* 1974;20:793–813.
17. Dantzig GB, editor. Application of simplex method to a transportation problem. In: Koopmans TC, editor. *Activity analysis of production and allocation*. New York: Wiley; 1951.
18. Reinfield NV, Vogel W. Mathematical programming. Englewood Cliffs (NJ): Prentice-Hall; 1958.
19. Dennis JB. A high-speed computer technique for the transportation problem. *J ACM* 1958;5:132–153.
20. Hadley G. Linear programming. Reading (MA): Addison-Wesley; 1962.
21. Srinivasan V, Thompson GL. Benefit-cost analysis of coding techniques for the primal transportation algorithm. *J ACM* 1973;20:194–213.
22. Gavish B, Schweitzer P, Shlifer E. The zero pivot phenomenon in transportation and assignment problems and its computational implications. *Math Program* 1977;12:226–240.
23. Gass SI. On solving the transportation problem. *J Oper Res Soc* 1990;4:291–297.
24. Charnes A, Cooper WW. The stepping-stone method for explaining linear programming calculations in transportation problems. *Manage Sci* 1954;1:46–69.
25. Glicksman S, Johnson L, Eselson L. Coding the transportation problem. *Nav Res Log Q* 1960;7:169–183.
26. Harris B. A code for the transportation problem of linear programming. *J ACM* 1976;23:155–157.

27. Langley RW, Kennington J, Shetty CM. Efficient computational devices for the capacitated transportation problem. *Nav Res Log Q* 1974;21:637–647.
28. Ford LR, Fulkerson DR. Solving the transportation problem. *Manage Sci* 1956;3:24–32.
29. Munkres J. Algorithms for the assignment and transportation problem. *SIAM J* 1957;5:32–38.
30. Kuhn HW. The Hungarian method for the assignment problem. *Nav Res Log Q* 1955;2:83–97.
31. Ford LR, Fulkerson DR. A primal-dual algorithm for the capacitated Hitchcock problem. *Nav Res Log Q* 1957;4:47–54.
32. Spivey WA, Thrall RM. *Linear optimization*. New York: Holt, Rinehart and Winston; 1970.
33. Balinski ML, Gomory RE. A primal method for the assignment and transportation problems. *Manage Sci* 1964;10:578–593.
34. Grigoriadis MD, Walker WP. A treatment of transportation problems by primal partitioning programming. *Manage Sci* 1968;14:565–599.
35. Gass SI. The dualplex method applied to special linear programming. *Proceedings of the International Federation of Information Processing Societies*. Amsterdam: North-Holland Publishing Co; 1972. pp. 1317–1323.
36. Williams AC. A treatment of transportation problems by decomposition. *SIAM J* 1962;10:35–48.
37. Dantzig GB, Wolfe P. A decomposition principle for linear programs. *Oper Res*. 1960;8: 101–111.
38. Ford LR, Fulkerson DR. *Flows in networks*. Princeton (NJ): Princeton University Press; 1962.
39. Clasen RJ. The numerical solution of network problems using the out-of-kilter algorithm. RAND Memo RM-5456-PR. Santa Monica (CA); 1968.
40. Hatch RS. Bench marks comparing transportation codes based on primal simplex and primal-dual algorithms. *Oper Res* 1975;23:1167–1172.
41. Glover F, Klingman D. Comments on a note by Hatch on network algorithms. *Oper Res* 1978;26:370–374.
42. Bradley GH, Brown GG, Graves GW. Design and implementation of large scale primal transshipment algorithms. *Manage Sci* 1977;24:1–34.
43. Miller DL, Pekny JF, Thompson GL. Solution of large dense transportation problems using a parallel primal algorithm. *Oper Res Lett* 1990;9:319–324.
44. Barr RS, Hickman BL. Parallel simplex for large pure network problems: Computational testing and sources of speedup. *Oper Res* 1994;1:65–80.
45. Li X, Zenios SA. Data-level parallel solution of min-cost network flow problems using ϵ -relaxations. *Eur J Oper Res* 1994;79: 474–488.
46. Edmonds J, Karp RM. Theoretical improvements in algorithmic efficiency for network flow problems. *J ACM* 1972;19: 248–264.
47. Dinits EA. The method of scaling and transportation problems. In: Fridman AA, editor. *Studies in discrete mathematics*. Moscow: Izdatel'stvo Nauka; 1973. pp. 159–160.
48. Hoffman AJ, Markowitz HM. A note on shortest path, assignment and transportation problems. *Nav Res Log Q* 1963;10: 375–379.
49. Lageman JJ. A method for solving the transportation problem. *Nav Res Log Q* 1967;14:89–99.
50. Gabow HN. Scaling algorithms for network problems. *J Comput Syst Sci* 1985;31: 148–168.
51. Gabow HN, Tarjan RE. Faster scaling algorithms for network problems. *SIAM J Comput* 1989;18:1013–1036.
52. Goldberg AV, Tarjan RE. Finding minimum-cost circulations by successive approximation. *Math Oper Res* 1990;15:430–466.
53. Tardos E. A strongly polynomial minimum cost circulation algorithm. *Combinatorica* 1985;5:247–255.
54. Galil Z, Tardos E. An $O(n^2(m + n \log n) \log n)$ min-cost flow algorithm. *J ACM*; 35: 374–386.
55. Orlin JB. A faster strongly polynomial minimum cost flow algorithm. *Oper Res* 1993;41:338–350.
56. Bertsekas DP, Castanon DA. The auction algorithm for the transportation problem. *Ann Oper Res* 1989;20:67–96.
57. Matsui T. A linear time algorithm for the Hitchcock transportation problem with fixed number of supply points. Cooperative Research Report 35 1992. Minami-Azabu, Minato-ku: The Institute of Statistical Mathematics; 128–138.

- 58. Ahuja RK, Orlin JB, Stein C, *et al.* Improved algorithms for bipartite network flow. SIAM J Comput 1994;23:906–933.
- 59. Kleinschmidt P, Schannath H. A strongly polynomial algorithm for the transportation problem. Math Program 1995;68:1–13.
- 60. Brenner U. A faster polynomial algorithm for the unbalanced Hitchcock transportation problem. Oper Res Lett 2008;36:408–413.

TRANSPORTATION RESOURCE MANAGEMENT

SUMIT KUNNUMKAL
Indian School of Business,
Hyderabad, India

HUSEYIN TOPALOGLU
School of Operations Research
and Information Engineering,
Cornell University, Ithaca,
New York

Many problems in transportation logistics involve managing a set of resources to satisfy the service requests that arrive randomly over time. For example, truckload carriers decide which load requests they should accept and to which locations they should reposition the empty vehicles. Railroad companies deploy the empty railcars to different stations in the network to serve the random customer demand in the most effective manner. Irrespective of the application setting, transportation resource management problems pose significant challenges. To begin with, these problems tend to be large in scale, spanning transportation networks with hundreds of locations and planning horizons with hundreds of time periods. Furthermore, the service requests arrive randomly, and we may not have full information about the future service requests. When making the decisions for the current time period, we need to keep a balance between maximizing the immediate profits and getting the vehicles into favorable positions to serve the potential future service requests. Finally, the decisions that we make for different locations and for different time periods interact in a nontrivial fashion. For example, the decision to serve a particular load from location i to j at time period t affects what other loads we can serve at the future time periods, since the vehicle that is used to serve the load at time period t becomes available at location j at a future time period.

In this article, we describe a variety of modeling and algorithmic approaches

for transportation resource management problems under uncertainty. We begin with a problem that takes place over two time stages. At the first stage, we decide to which locations we should reposition the resources to serve the demands that arrive later. At the second stage, we observe the demand arrivals and decide which of these demands we should serve. The first algorithmic strategy that we discuss for the two-stage problem is based on a deterministic linear programming formulation. This formulation is simple to implement, but it does not explicitly address the randomness in the demand arrivals. Motivated by this observation, we move on to other algorithmic strategies that address the randomness in the demand arrivals by using a dynamic programming formulation. The dynamic programming formulation of the two-stage problem turns out to be computationally difficult as it involves a high dimensional state variable, and we resort to approximation strategies. In particular, the first two approximation strategies that we describe decompose the dynamic programming formulation by the locations and obtain good solutions by focusing on one location at a time. We also discuss a third approximation strategy that constructs separable value function approximations by using sampled trajectories of the system. After the discussion of two-stage problems, we move on to resource management problems that take place over multiple time periods. We give a dynamic programming formulation for a multistage resource management problem and demonstrate that the algorithmic concepts that we develop for two-stage problems can be extended to multistage problems without too much difficulty.

In addition to providing algorithmic tools for two-stage and multistage problems, we also dwell on modeling issues. Most transportation resource management problems involve complex resources. For example, a driver resource in a driver scheduling application may require keeping track of its

inbound location, time to reach the inbound location, duty time within the shift, days away from home, vehicle type, and home domicile. It would be cumbersome to model such complex entities through traditional mathematical programming tools, and we provide a flexible paradigm that is useful for modeling and communicating complex resource management problems. We conclude the article with a series of numerical experiments that demonstrate the benefits from using models that explicitly capture the randomness in the system.

Transportation resource management models have a long history dating back to the earliest days of linear programming [1–4]. The early models formulate the problem over a state time network. The nodes of the state time network represent the supply of resources at different locations and at different time periods, whereas the arcs represent the movements of the resources between different locations. These models tend to ignore the randomness in the problem and simply work with the expected values of the demand arrivals. A subclass of the deterministic models is the so-called myopic models, which ignore the unknown future altogether and only work with what is known with certainty [5–7]. This approach is justified by the fact that many customers call in advance, and a large portion of the future demands is actually known with certainty. Whether they ignore the uncertainty in the future altogether or incorporate it through expectations of the demand arrivals, these kinds of deterministic models can yield tractable optimization problems and can be relatively simple to implement. As a result, they form the backbone of many commercial systems [8–10].

The majority of the models that we discuss in this article fall under the category of stochastic models, which try to capture the random demand arrivals by formulating the problem as a dynamic program and using value functions to assess the impact of the current decisions on the future evolution of the system. The dynamic programming formulations of practical resource management problems generally involve high dimensional

state variables, and this high dimensionality makes it impossible to compute the value functions by using standard Markov decision process tools. Therefore, most of the effort is concentrated around approximating the value functions in a tractable fashion [11–20]. The overview of transportation resource management models that we give in this article is rather compact, and it mostly evolves around methods that approximate the value functions. We refer the reader to Powell *et al.* [21], Powell and Topaloglu [22], and Barnhart and Laporte [23] for detailed reviews of deterministic and stochastic models. We also emphasize that our goal in this article is to give an overview of transportation resource management models in the literature, rather than to make original contributions. At the end of each section in this article, we give pointers to the relevant literature and acknowledge the original contributors.

The article is organized as follows: In the section titled “Two-Stage Problems,” we formulate a two-stage resource management problem as a dynamic program. This formulation has a high dimensional state variable and we discuss a variety of algorithmic approaches to obtain tractable approximations. In the section titled “Multistage Problems,” we extend the ideas that we discussed in the previous section to multistage problems. In the section titled “Flexible Modelling Framework,” we give a flexible paradigm for modeling complex resource management problems. In the section titled “Numerical Illustrations,” we present numerical experiments.

TWO-STAGE PROBLEMS

In this section, we consider a resource management problem that takes place over a planning horizon of two time periods. Our goal is to illustrate the fundamental algorithmic concepts by using a relatively simple problem setting. It turns out that all of the development in this section can be extended to multistage problems without too much difficulty.

We have a set of resources that can be used over a planning horizon of two time periods

to serve the demands that arrive randomly into the system. At the beginning of the first time period, there are a certain number of resources at different locations. The decisions that we make at the first time period involve repositioning the resources between different locations so that the resources end up at favorable locations to serve the demands that arrive at the second time period. At the beginning of the second time period, we observe the demand arrivals. The decisions that we make at the second time period involve serving the demands by using the resources. The objective is to maximize the total expected profit over the two time periods, which is the difference between the total cost of repositioning the resources and the total expected revenue from serving the demands.

The set of locations in the transportation network is $\mathcal{L}$. At the beginning of the first time period, we have s_i resources available at location i , and $\{s_i : i \in \mathcal{L}\}$ is a part of the problem data. We use x_{ij} to denote the number of resources that we reposition from location i to j at the first time period. The cost of repositioning a resource from location i to j is c_{ij} . We let r_i be the number of resources available at location i at the beginning of the second time period. Naturally, $\{r_i : i \in \mathcal{L}\}$ is determined by the repositioning decisions that we make at the first time period. The set of possible demand types is $\mathcal{K}$. We use y_i^k to denote the number of demands of type k that we serve by using a resource at location i at the second time period. The revenue from serving a demand of type k by using a resource at location i is p_i^k . If it is not feasible to serve a demand of type k by using a resource at location i , then we capture this infeasibility by letting $p_i^k = -\infty$. We use the random variable D^k to represent the number of demands of type k that arrive at the second time period.

Letting $r = \{r_i : i \in \mathcal{L}\}$ and $D = \{D^k : k \in \mathcal{K}\}$ to respectively capture the number of resources at different locations and the demand arrivals at the second time period, we can determine the total revenue at the second time period by computing the value

function $V(r, D)$ through the problem

$$V(r, D) = \max \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_i^k \quad (1a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad \forall i \in \mathcal{L}; \quad (1b)$$

$$\sum_{i \in \mathcal{L}} y_i^k \leq D^k, \quad \forall k \in \mathcal{K}; \quad (1c)$$

$$y_i^k \in \mathbb{Z}_+, \quad \forall i \in \mathcal{L}, k \in \mathcal{K}. \quad (1d)$$

In the problem above, constraint (1b) ensures that the number of resources that we use to serve the demands does not violate the resource availabilities at different locations, whereas constraint (1c) ensure that the number of demands that we serve does not exceed the demand arrivals. If the number of resources at different locations is given by the vector r , then the total expected revenue that we generate at the second time period is captured by $\bar{V}(r) = \mathbb{E}\{V(r, D)\}$. Therefore, we can maximize the total expected profit over the two time periods by solving the problem

$$Z = \max - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ij} + \bar{V}(r) \quad (2a)$$

$$\text{subject to } \sum_{j \in \mathcal{L}} x_{ij} = s_i, \quad \forall i \in \mathcal{L}; \quad (2b)$$

$$\sum_{i \in \mathcal{L}} x_{ij} - r_j = 0, \quad \forall j \in \mathcal{L}; \quad (2c)$$

$$x_{ij}, r_j \in \mathbb{Z}_+, \quad \forall i, j \in \mathcal{L}, \quad (2d)$$

where constraint (2b) ensures that the number of resources that we reposition does not violate the resource availabilities at different locations, and constraint (2c) computes the number of resources at different locations at the beginning of the second time period.

Conceptually, we can compute $V(r, D)$ for all possible values of the vector r and demand arrivals D by solving problem (1). In this case, we can take expectations to compute $\bar{V}(r) = \mathbb{E}\{V(r, D)\}$ for all possible values of the vector r and use this value function in problem (2) to find the optimal repositioning decisions. Practically, however, the number of possible values of the vector r and demand arrivals D can

be so large that this approach becomes computationally intractable. Therefore, we primarily focus our attention to developing methods to approximate the value function $\bar{V}(\cdot)$.

Deterministic Linear Program

A common approach for problems that take place under uncertainty is to assume that all random variables take on their expected values and to solve a deterministic approximation. In our problem setting, if we assume that all demand arrivals take on their expected values and we can utilize a fractional number of resources, then such a deterministic approximation can be written as

$$Z^{\text{DLP}} = \max - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ij} + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_i^k \quad (3a)$$

$$\text{subject to } \sum_{j \in \mathcal{L}} x_{ij} = s_i, \quad \forall i \in \mathcal{L}; \quad (3b)$$

$$- \sum_{i \in \mathcal{L}} x_{ij} + \sum_{k \in \mathcal{K}} y_j^k \leq 0, \quad \forall j \in \mathcal{L}; \quad (3c)$$

$$\sum_{i \in \mathcal{L}} y_i^k \leq \mathbb{E}\{D^k\}, \quad \forall k \in \mathcal{K}; \quad (3d)$$

$$x_{ij}, y_j^k \in \mathbb{R}_+, \quad \forall i, j \in \mathcal{L}, k \in \mathcal{K}. \quad (3e)$$

Constraint (3b) in the problem above is analogous to constraint (2b) in the problem 2, whereas constraint (3d) is analogous to constraint (1c) in problem 1. In constraint (3d), a portion of the demands $\{D^k : k \in \mathcal{K}\}$ may be known at the beginning of the first time period, in which case, these demands simply become deterministic quantities. Constraint (3c) ensures that the number of resources that we use to serve the demands at the second time period does not violate the number of resources that we reposition at the first time period. There are two uses of the problem (3). First, this problem can be used to make resource repositioning decisions. In particular, letting $\{\hat{x}_{ij} : i, j \in \mathcal{L}\}$,

$\{\hat{y}_i^k : i \in \mathcal{L}, k \in \mathcal{K}\}$ be an optimal solution to problem 3, we can use $\{\hat{x}_{ij} : i, j \in \mathcal{L}\}$ as an approximation to the optimal solution to problem 2. The solution $\{\hat{x}_{ij} : i, j \in \mathcal{L}\}$ may utilize fractional number of resources. If this is a concern, then we can impose integrality requirements in problem (3).

The second use of problem 3 is that the optimal objective value of this problem provides an upper bound on the optimal total expected profit Z . To see the reason, we let $\{\hat{x}_{ij}(D) : i, j \in \mathcal{L}\}$, $\{\hat{y}_i^k(D) : i \in \mathcal{L}, k \in \mathcal{K}\}$ be an optimal solution and $Z^{\text{DLP}}(D)$ be the optimal objective value of problem(3) when we replace the right side of constraint(3d) with the demand arrivals $D = \{D^k : k \in \mathcal{K}\}$. In this case, $Z^{\text{DLP}}(D)$ corresponds to the perfect hindsight total profit that is computed after the demand arrivals are observed to be D . Therefore, $Z^{\text{DLP}}(D)$ is an upper bound on the total profit that is obtained by any approach when the demand arrivals are observed to be D . Taking expectations, we observe that $\mathbb{E}\{Z^{\text{DLP}}(D)\}$ is an upper bound on the total expected profit that is obtained by any approach. Since the total expected profit obtained by the optimal policy is Z , we obtain

$$Z \leq \mathbb{E}\{Z^{\text{DLP}}(D)\} \leq Z^{\text{DLP}}(\mathbb{E}\{D\}) = Z^{\text{DLP}},$$

where the second inequality follows from Jensen's inequality and the equality follows from the definition of $Z^{\text{DLP}}(\mathbb{E}\{D\})$ [24]. This kind of an upper bound on the optimal total expected profit can be useful when assessing the optimality gap of suboptimal approximation strategies, such as the ones that we develop in this article. An interesting observation from the chain of inequalities above is that $\mathbb{E}\{Z^{\text{DLP}}(D)\}$ also provides an upper bound on Z and this upper bound is tighter than the one provided by Z^{DLP} . However, to compute $\mathbb{E}\{Z^{\text{DLP}}(D)\}$, we need to compute $Z^{\text{DLP}}(D)$ for all possible values of the demand arrivals D and take an expectation. Taking this expectation can be computationally intractable when the number of possible values of the demand arrivals is large, but we can resort to Monte Carlo samples to approximate the expectation $\mathbb{E}\{Z^{\text{DLP}}(D)\}$.

Deterministic approximations similar to the one in problem 3 form the core of many practical transportation models [10]. However, such approximations may not always yield satisfactory results as they completely ignore the uncertainty in the demand arrivals. Our goal is to present models that address the uncertainty in the demand arrivals more carefully.

Separable Value Functions

In this section, we describe a special case where $\bar{V}(r) = \mathbb{E}\{V(r, D)\}$ can be computed in a tractable fashion. It turns out that we can build on the results that we obtain for this special case to develop approximation strategies for more general cases.

We consider the special case where for each demand type k , there is a single location i_k such that only the resources at location i_k can serve a demand of type k . In other words, we have $p_i^k = -\infty$ for all $i \in \mathcal{L} \setminus \{i_k\}$, which implies that $y_i^k = 0$ for all $i \in \mathcal{L} \setminus \{i_k\}$ in an optimal solution to problem 1. In this case, it is easy to see that we can replace the constraint $\sum_{i \in \mathcal{L}} y_i^k \leq D^k$ for all $k \in \mathcal{K}$ in problem 1 with the constraint $y_{i_k}^k \leq D^k$ for all $i \in \mathcal{L}$, $k \in \mathcal{K}$. If we do replace the constraint $\sum_{i \in \mathcal{L}} y_i^k \leq D^k$ for all $k \in \mathcal{K}$ in problem 1 with the constraint $y_{i_k}^k \leq D^k$ for all $i \in \mathcal{L}$, $k \in \mathcal{K}$, then all of the constraints in problem 1 decompose by the locations, and we can obtain an optimal solution to problem 1 by concentrating on one location at a time. In particular, the subproblem for location i is of the form

$$V_i(r_i, D) = \max \sum_{k \in \mathcal{K}} p_i^k y_i^k \quad (4a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad (4b)$$

$$y_i^k \leq D^k, \quad \forall k \in \mathcal{K}; \quad (4c)$$

$$y_i^k \in \mathbb{Z}_+, \quad \forall k \in \mathcal{K}, \quad (4d)$$

and we have $V(r, D) = \sum_{i \in \mathcal{L}} V_i(r_i, D)$. An important observation is that problem 4 is a knapsack problem. The capacity of the knapsack is r_i . There are $|\mathcal{K}|$ items, each item occupies one unit of space, and we can put at most D^k units of item k into the knapsack.

Therefore, it is possible to solve problem 4 by sorting the objective function coefficients of the items and putting the items into the knapsack starting from the item with the largest objective function coefficient.

We need the value function $\bar{V}(r) = \sum_{i \in \mathcal{L}} \mathbb{E}\{V_i(r_i, D)\}$ to compute the optimal repositioning decisions through problem 2. It turns out that taking this expectation is tractable due to the knapsack structure of problem 4. We assume, without loss of generality, that the set of demand types is $\mathcal{K} = \{1, 2, \dots, K\}$ and the demand types are ordered such that $p_i^1 \geq p_i^2 \geq \dots \geq p_i^K$. Noting the knapsack interpretation of problem 4, we let $\psi_i^k(q)$ be the probability that the k th item uses the q th unit of available space in the knapsack for location i . In this case, the expected revenue that we generate from the q th unit of available space becomes $\sum_{k \in \mathcal{K}} \psi_i^k(q) p_i^k$ and the expectation of the optimal objective value of problem 4 can be computed as

$$\mathbb{E}\{V_i(r_i, D)\} = \sum_{q=1}^{r_i} \sum_{k \in \mathcal{K}} \psi_i^k(q) p_i^k.$$

To compute $\psi_i^k(q)$, we observe that for the k th item to use the q th unit of available space in the knapsack, the total amount of space used by the first $k-1$ items should be strictly smaller than q , the total amount of space used by the first k items should be greater than or equal to q , and item k should satisfy $i_k = i$, since otherwise we have $p_i^k = -\infty$ and it is never desirable to put the k th item into the knapsack. Therefore, we have $\psi_i^k(q) = \mathbb{P}\{D^1 + \dots + D^{k-1} < q \leq D^1 + \dots + D^k\}$ if $i_k = i$, and $\psi_i^k(q) = 0$ otherwise. This implies that we can compute $\psi_i^k(q)$ in a tractable fashion as long as we can compute the convolutions of the distributions of $\{D^k : k \in \mathcal{K}\}$, which is the case when the demand random variables have, for example, independent Poisson distributions.

The approach outlined above is proposed by Cheung [25]. Powell and Cheung [26], Powell and Cheung [27], Cheung and Powell [28], and Cheung and Powell [29] apply this approach to a variety of settings, including dynamic fleet management and inventory

distribution. An interesting aspect of this approach is that it not only allows us to compute $\bar{V}(r) = \mathbb{E}\{V(r, D)\}$ in a special case, but also allows us to construct approximations when the special case does not necessarily hold. This is the topic of the next section.

Decomposition by Demand Relaxation

In this section, we demonstrate how we can build on the special case that we discuss in the section titled “Separable Value Functions” to construct tractable approximation strategies for more general cases. The fundamental idea is to relax the constraint (1c) in the problem 1 by associating the nonnegative Lagrange multipliers $\lambda = \{\lambda^k : k \in \mathcal{K}\}$ with it. In this case, the relaxed version of problem 1 becomes

$$V^{\text{DDR}}(r, D, \lambda) = \max \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} [p_i^k - \lambda^k] y_i^k + \sum_{k \in \mathcal{K}} \lambda^k D^k \quad (5a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad \forall i \in \mathcal{L}; \quad (5b)$$

$$y_i^k \leq D^k, \quad \forall i \in \mathcal{L}, k \in \mathcal{K}; \quad (5c)$$

$$y_i^k \in \mathbb{Z}_+, \quad \forall i \in \mathcal{L}, k \in \mathcal{K}. \quad (5d)$$

The constraint (5c) would be redundant in problem 1, but we add it to the problem above to tighten the relaxation. We use the argument λ in the value function $V^{\text{DDR}}(r, D, \lambda)$ to emphasize that the optimal objective value of problem 5 depends on the Lagrange multipliers that we associate with the constraint $\sum_{i \in \mathcal{L}} y_i^k \leq D^k$ for all $k \in \mathcal{K}$. The superscript *DDR* in the value function stands for decomposition by demand relaxation.

Noting that all of the constraints in problem 5 decompose by the locations, we can obtain an optimal solution to this problem by concentrating on one location at a time. In this case, the subproblem for location i has the same form as problem 4 with the exception that the objective

function of the subproblem for location i now reads $\sum_{k \in \mathcal{K}} [p_i^k - \lambda^k] y_i^k$ instead of $\sum_{k \in \mathcal{K}} p_i^k y_i^k$. This implies that we can compute $\mathbb{E}\{V^{\text{DDR}}(r, D, \lambda)\}$ in a tractable fashion by using the approach that we describe in the previous section. Furthermore, it is possible to show that the demand relaxation strategy provides an upper bound on the optimal total expected profit. To see the reason, we let $\{\hat{y}_i^k : i \in \mathcal{L}, k \in \mathcal{K}\}$ be an optimal solution to problem 1 and note that

$$V(r, D) = \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k \hat{y}_i^k \leq \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} [p_i^k - \lambda^k] \hat{y}_i^k + \sum_{k \in \mathcal{K}} \lambda^k D^k \leq V^{\text{DDR}}(r, D, \lambda). \quad (6)$$

The first inequality follows from the fact the Lagrange multipliers are nonnegative and we have $\sum_{i \in \mathcal{L}} \hat{y}_i^k \leq D^k$ for all $k \in \mathcal{K}$ by constraint (1c) in the problem 1. The second inequality follows from the fact that $\{\hat{y}_i^k : i \in \mathcal{L}, k \in \mathcal{K}\}$ is a feasible, but not necessarily an optimal, solution to problem 5. Taking expectations, we obtain $\bar{V}(r) = \mathbb{E}\{V(r, D)\} \leq \mathbb{E}\{V^{\text{DDR}}(r, D, \lambda)\}$. Therefore, if we replace $\bar{V}(r)$ in the objective function of problem 2 with $\mathbb{E}\{V^{\text{DDR}}(r, D, \lambda)\}$ and solve this problem, then the resulting objective function value $Z^{\text{DDR}}(\lambda)$ provides an upper bound on the optimal total expected profit Z , as long as the Lagrange multipliers are nonnegative.

There are several approaches that we can use to choose a good set of Lagrange multipliers. To begin with, we note that $Z^{\text{DDR}}(\lambda) \geq Z$ whenever we have $\lambda \geq 0$. Therefore, we can solve the problem $\min_{\lambda \geq 0} Z^{\text{DDR}}(\lambda)$ to obtain the tightest possible upper bound on the optimal total expected profit and use the optimal solution to this problem as the Lagrange multipliers. Cheung and Powell [29] show that $Z^{\text{DDR}}(\lambda)$ is a convex function of λ so that the problem $\min_{\lambda \geq 0} Z^{\text{DDR}}(\lambda)$ can be solved by using standard subgradient optimization. A possible shortcoming of this approach is that solving the problem $\min_{\lambda \geq 0} Z^{\text{DDR}}(\lambda)$ can be time consuming. Kunnumkal and Topaloglu [30] propose another approach for choosing the Lagrange multipliers that uses the dual solution to problem 3. In particular, they note that constraint (3d) in problem 3 is analogous to constraint (1c) in problem 1. In this case, letting $\hat{\lambda} = \{\hat{\lambda}^k : k \in \mathcal{K}\}$ be the optimal

values of the dual variables associated with constraint(3d) in problem 3, they choose the Lagrange multipliers as $\hat{\lambda}$. Since $\hat{\lambda} \geq 0$ by dual feasibility to problem 3, we have $Z^{\text{DDR}}(\hat{\lambda}) \geq Z$. Furthermore, Kunnumkal and Topaloglu [30] show that $Z^{\text{DDR}}(\hat{\lambda}) \leq Z^{\text{DLP}}$, and we obtain

$$Z \leq \min_{\lambda \geq 0} Z^{\text{DDR}}(\lambda) \leq Z^{\text{DDR}}(\hat{\lambda}) \leq Z^{\text{DLP}}.$$

Therefore, the demand relaxation strategy provides an upper bound on the optimal total expected profit, and this upper bound is tighter than the one provided by the deterministic approximation.

The demand relaxation strategy is proposed by Cheung [25], but the idea of relaxing complicating constraints in stochastic optimization problems frequently appears in the literature. Ruszczyński [31] gives an overview of this idea within the framework of general stochastic programs. Adelman and Mersereau [32] introduce the term “weakly coupled dynamic program” to refer to a stochastic optimization problem which would decompose if a few linking constraints did not exist. They study a dual framework for such stochastic optimization problems. Cheung and Powell [28,29] provide applications of the demand relaxation strategy to dynamic fleet management and inventory distribution problems. Similar relaxation ideas are used by Castanon [33], Yost and Washburn [34], and Bertsimas and Mersereau [35] in partially observable Markov decision processes, by Federgruen and Zipkin [36], Topaloglu and Kunnumkal [37], and Kunnumkal and Topaloglu [38] in inventory distribution, and by Erdelyi and Topaloglu [39], Kunnumkal and Topaloglu [40], and Topaloglu [41] in airline network revenue management.

Decomposition by Supply Relaxation

In this section, we construct a tractable approximation strategy by using a relaxation idea that is conceptually similar to the one in the previous section, but the specific details of this relaxation idea are quite different. We begin by choosing an arbitrary location i and relax constraint (1b) in problem 1 for all other

locations by associating the nonnegative Lagrange multipliers $\mu = \{\mu_j : j \in \mathcal{L} \setminus \{i\}\}$ with them. In this case, the relaxed version of problem 1 becomes

$$V_i^{\text{DSR}}(r, D, \mu) = \max \sum_{k \in \mathcal{K}} p_i^k y_i^k + \sum_{j \in \mathcal{L} \setminus \{i\}} \sum_{k \in \mathcal{K}} [p_j^k - \mu_j] y_j^k + \sum_{j \in \mathcal{L} \setminus \{i\}} \mu_j r_j \quad (7a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad (7b)$$

$$y_i^k + \sum_{j \in \mathcal{L} \setminus \{i\}} y_j^k \leq D^k, \forall k \in \mathcal{K}; \quad (7c)$$

$$y_i^k, y_j^k \in \mathbb{Z}_+, \forall j \in \mathcal{L} \setminus \{i\}, k \in \mathcal{K}. \quad (7d)$$

The subscript i in the value function $V_i^{\text{DSR}}(r, D, \mu)$ emphasizes that the optimal objective value of the problem above depends on the choice of location i , and the superscript *DSR* stands for decomposition by supply relaxation.

It is possible to show that problem 7 has the same structure as problem 4, in which case, we can compute $\mathbb{E}\{V_i^{\text{DSR}}(r, D, \mu)\}$ by using the approach that we describe in the section titled “Separable Value Functions.” To see the equivalence between problems 4 and 7, we first observe that the decision variables $\{y_j^k : j \in \mathcal{L} \setminus \{i\}, k \in \mathcal{K}\}$ appear only in constraint (7c) in problem 7. This implies that we can replace the decision variables $\{y_j^k : j \in \mathcal{L} \setminus \{i\}\}$ with a single decision variable, and the objective function coefficient of this decision variable would be $\max_{j \in \mathcal{L} \setminus \{i\}} [p_j^k - \mu_j]$. Therefore, letting $\rho^k = \max_{j \in \mathcal{L} \setminus \{i\}} [p_j^k - \mu_j]$ for notational brevity and replacing the decision variables $\{y_j^k : j \in \mathcal{L} \setminus \{i\}\}$ with the single decision variable z^k , problem 7 can be written as

$$\max \sum_{k \in \mathcal{K}} p_i^k y_i^k + \sum_{k \in \mathcal{K}} \rho^k z^k + \sum_{j \in \mathcal{L} \setminus \{i\}} \mu_j r_j \quad (8a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad (8b)$$

$$y_i^k + z^k \leq D^k, \forall k \in \mathcal{K}; \quad (8c)$$

$$y_i^k, z^k \in \mathbb{Z}_+, \forall k \in \mathcal{K}. \quad (8d)$$

We assume without loss of generality that $\rho^k \geq 0$ for all $k \in \mathcal{K}$. If there exists some $k \in \mathcal{K}$ such that $\rho^k < 0$, then the corresponding decision variable z^k takes value zero in an optimal solution, and this decision variable can be dropped from problem 8 without changing the optimal objective value. Since we have $\rho^k \geq 0$ for all $k \in \mathcal{K}$, the decision variables $\{z^k : k \in \mathcal{K}\}$ take their largest possible values in an optimal solution and the largest possible values of these decision variables are $\{D^k - y_i^k : k \in \mathcal{K}\}$. Therefore, we can replace the decision variables $\{z^k : k \in \mathcal{K}\}$ in the problem above with $\{D^k - y_i^k : k \in \mathcal{K}\}$ and obtain the equivalent problem

$$\begin{aligned} \max \quad & \sum_{k \in \mathcal{K}} [p_i^k - \rho^k] y_i^k + \sum_{k \in \mathcal{K}} \rho^k D^k \\ & + \sum_{j \in \mathcal{L} \setminus \{i\}} \mu_j r_j \end{aligned} \quad (9a)$$

$$\text{subject to } \sum_{k \in \mathcal{K}} y_i^k \leq r_i, \quad (9b)$$

$$y_i^k \leq D^k, \forall k \in \mathcal{K}; \quad (9c)$$

$$y_i^k \in \mathbb{Z}_+, \forall k \in \mathcal{K}. \quad (9d)$$

Ignoring the constant terms $\sum_{k \in \mathcal{K}} \rho^k D^k$ and $\sum_{j \in \mathcal{L} \setminus \{i\}} \mu_j r_j$, problem 9 has the same structure as problem 4. By using an argument similar to the one in 6, we can show that $V(r, D) \leq V_i^{\text{DSR}}(r, D, \mu)$ as long as the Lagrange multipliers are nonnegative [42].

Since the choice of location i is arbitrary, we have $V(r, D) \leq V_i^{\text{DSR}}(r, D, \mu)$ for all $i \in \mathcal{L}$, and it is not clear which one of $\{V_i^{\text{DSR}}(r, D, \mu) : i \in \mathcal{L}\}$ we should use as an approximation to $V(r, D)$. We resolve this ambiguity by averaging over all $i \in \mathcal{L}$ and using $\frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} V_i^{\text{DSR}}(r, D, \mu)$ as an approximation to $V(r, D)$. Noting that $V_i^{\text{DSR}}(r, D, \mu)$ is an upper bound on $V(r, D)$ for all $i \in \mathcal{L}$, $\frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} V_i^{\text{DSR}}(r, D, \mu)$ is also an upper bound on $V(r, D)$. In this case, if we replace $\bar{V}(r)$ in the objective function of problem 2 with $\frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \mathbb{E}\{V_i^{\text{DSR}}(r, D, \mu)\}$ and solve this problem, then the resulting objective function value $Z^{\text{DSR}}(\mu)$ provides an upper bound on the optimal total expected profit Z , as long as we have $\mu \geq 0$.

The methods that we use to choose a good set of Lagrange multipliers are similar to those in the section titled “Decomposition by Demand Relaxation.” In particular, we can solve the problem $\min_{\mu \geq 0} Z^{\text{DSR}}(\mu)$ to obtain the tightest possible upper bound on Z and use the optimal solution to this problem as the Lagrange multipliers. It is possible to show that $Z^{\text{DSR}}(\mu)$ is a convex function of μ so that we can solve the problem $\min_{\mu \geq 0} Z^{\text{DSR}}(\mu)$ by using standard subgradient optimization. Another approach for choosing a good set of Lagrange multipliers is based on the dual solution of problem 3. We note that constraint (1b) in problem 1 captures the resource availabilities at different locations at the beginning of the second time period. Similarly, constraint (3c) in problem 3 captures the resource availabilities at different locations. Therefore, letting $\{\hat{\mu}_i : i \in \mathcal{L}\}$ be the optimal values of the dual variables associated with constraint (3c) in problem 3, we can choose the Lagrange multipliers as $\{\hat{\mu}_j : j \in \mathcal{L} \setminus \{i\}\}$.

The supply relaxation strategy has its roots in the network revenue management literature. In this literature, it is customary to relax the capacity constraints for all of the flight legs except for one and to obtain approximations to the value function by concentrating on one flight leg at a time [43]. Zhang and Adelman [42] establish that the supply relaxation strategy provides an upper bound on the value function. Kunnumkal and Topaloglu [44, 45] provide alternative proofs for the same result. Erdelyi and Topaloglu [46] use the supply relaxation idea to solve an overbooking problem. Topaloglu [47] extends the supply relaxation strategy to general stochastic programs and provides applications in dynamic fleet management.

Separable Projective Approximations

In this section, we describe an iterative approach that uses sampled realizations of the demand arrivals to construct a separable approximation to the value function. The fundamental idea is to approximate the value function $\bar{V}(r)$ with a separable function of the form

$$V^{\text{SPA}}(r) = \sum_{i \in \mathcal{L}} V_i^{\text{SPA}}(r_i),$$

where the approximations $\{V_i^{SPA}(\cdot) : i \in \mathcal{L}\}$ are single dimensional, piecewise linear and concave functions with points of nondifferentiability being a subset of integers. The superscript SPA stands for separable projective approximations. The concavity of the approximation $V_i^{SPA}(\cdot)$ is intended to capture the intuitive expectation that the marginal benefit from an additional resource at a location should be decreasing as we have more resources available at the location. Noting that $V_i^{SPA}(\cdot)$ is a piecewise linear function with points of nondifferentiability being a subset of integers, if we have a total of Q available resources, then we can characterize the approximation $V_i^{SPA}(\cdot)$ with a sequence of Q slopes $\{v_i(q) : q = 0, \dots, Q-1\}$, where $v_i(q)$ is the slope of $V_i^{SPA}(\cdot)$ over the interval $[q, q+1]$. In other words, we have $v_i(q) = V_i^{SPA}(q+1) - V_i^{SPA}(q)$.

We use an iterative approach to construct the approximations $\{V_i^{SPA}(\cdot) : i \in \mathcal{L}\}$. In particular, letting $\{V_i^{SPA,n}(\cdot) : i \in \mathcal{L}\}$ be the approximations at iteration n , we begin by solving problem 2 after replacing $\bar{V}(r)$ in the objective function with $\sum_{i \in \mathcal{L}} V_i^{SPA,n}(r_i)$. This provides an optimal solution $\{\hat{x}_{ij}^n : i, j \in \mathcal{L}\}$, $\{\hat{r}_i^n : i \in \mathcal{L}\}$. We then sample a realization of the demand arrivals, which we denote by $\{\hat{D}^{k,n} : k \in \mathcal{K}\}$. We solve problem 1 after replacing the right side of constraint (1b)

with $\hat{r}^n = \{\hat{r}_i^n : i \in \mathcal{L}\}$ and the right side of constraint (1c) with $\hat{D}^n = \{\hat{D}^{k,n} : k \in \mathcal{K}\}$. The challenge is to use the solution to problem 1 to improve the approximations $\{V_i^{SPA,n}(\cdot) : i \in \mathcal{L}\}$. We give a description of the general approach in Fig. 1. In Step 1, we initialize the approximations $\{V_i^{SPA,n}(\cdot) : i \in \mathcal{L}\}$ to arbitrary piecewise linear and concave functions with points of nondifferentiability being a subset of integers. In Step 2, we solve problem 2 after replacing $\bar{V}(r)$ in the objective function with the value function approximation $\sum_{i \in \mathcal{L}} V_i^{SPA,n}(r_i)$. In Step 3, we sample a realization of the demand arrivals and solve problem 1 by using the sampled demand arrivals and the resource availabilities obtained from problem 2. In Step 4, we update the approximations $\{V_i^{SPA,n}(\cdot) : i \in \mathcal{L}\}$, and this provides the approximations $\{V_i^{SPA,n+1}(\cdot) : i \in \mathcal{L}\}$ that we use at the next iteration.

The approach that we use to update the approximations is based on a stochastic approximation idea, but we need to be careful to preserve the concavity of the value function approximation. Using e_i to denote the $|\mathcal{L}|$ dimensional unit vector with a one in the element corresponding to $i \in \mathcal{L}$, we let $\hat{\vartheta}_i^n = V(\hat{r}^n + e_i, \hat{D}^n) - V(\hat{r}^n, \hat{D}^n)$. Therefore, $\hat{\vartheta}_i^n$ captures the marginal benefit from an additional resource at location i in problem 1 when the number of resources at different

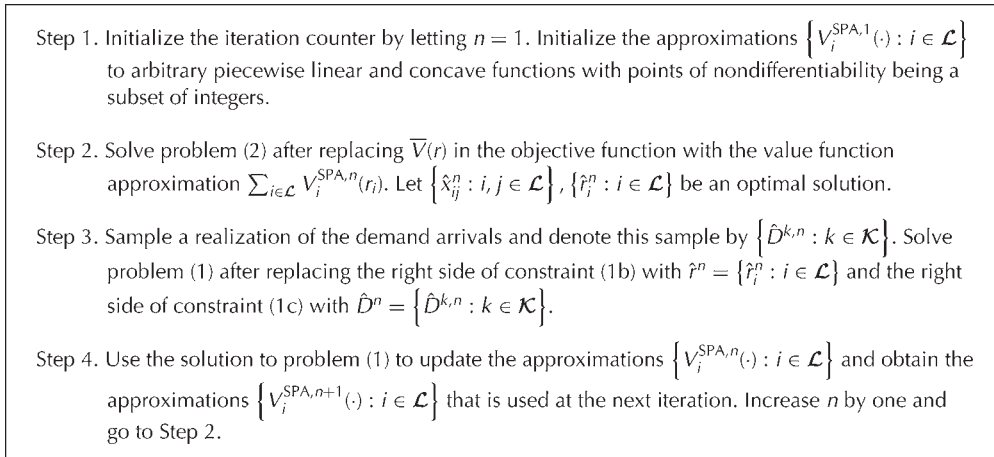


Figure 1. An iterative approach to construct a separable value function approximation.

locations is given by $\hat{r}^n = \{\hat{r}_i^n : i \in \mathcal{L}\}$ and the realization of the demand arrivals is given by $\hat{D}^n = \{\hat{D}^{k,n} : k \in \mathcal{K}\}$. In this case, letting $V_i^{SPA,n}(\cdot)$ be the piecewise linear and concave function characterized by the sequence of slopes $\{v_i^n(q) : q = 0, \dots, Q-1\}$, we update these slopes as

$$\gamma_i^n(q) = \begin{cases} [1 - \alpha^n] v_i^n(q) + \alpha^n \hat{v}_i^n & \text{if } q = \hat{r}_i^n \\ v_i^n(q) & \text{otherwise,} \end{cases}$$

where $\alpha^n \in [0, 1]$ is the step size parameter at iteration n . Since we may not have $\gamma_i^n(0) \geq \gamma_i^n(1) \geq \dots \geq \gamma_i^n(Q-1)$, the piecewise linear function characterized by the sequence of slopes $\{\gamma_i^n(q) : q = 0, \dots, Q-1\}$ is not necessarily concave. To obtain a concave approximation, we solve the problem

$$\min \sum_{q=0}^{Q-1} [w(q) - \gamma_i^n(q)]^2$$

subject to $w(q) \geq w(q+1)$, $\forall q = 0, \dots, Q-2$.

Letting $\{\hat{w}(q) : q = 0, \dots, Q-1\}$ be an optimal solution to the problem above, we use the sequence of slopes $\{\hat{w}(q) : q = 0, \dots, Q-1\}$ to characterize the approximation $V_i^{SPA,n+1}(\cdot)$. Since we have $\hat{w}(0) \geq \hat{w}(1) \geq \dots \geq \hat{w}(Q-1)$, the approximation $V_i^{SPA,n+1}(\cdot)$ is indeed concave. Intuitively speaking, the quadratic program above projects the piecewise linear function characterized by the sequence of slopes $\{\gamma_i^n(q) : q = 0, \dots, Q-1\}$ onto the set of piecewise linear and concave functions with points of nondifferentiability being a subset of integers. Throughout this article, we use $V_i^{SPA,n+1}(\cdot) = \mathcal{U}(V_i^{SPA,n}(\cdot), \hat{r}_i^n, \hat{v}_i^n, \alpha^n)$ to succinctly capture the updating procedure.

The separable projective approximations are proposed by Powell *et al.* [48]. The authors develop several convergence results for the updating procedure that we describe above, and these convergence results have close connections with the stochastic approximation theory. Topaloglu and Powell [18] and Topaloglu [49] use the separable projective approximations for dynamic fleet management problems. Godfrey and Powell [50] use

separable approximations to the value functions in a product distribution problem, but the method they use to update the approximations is different from the one that we describe above. Godfrey and Powell [14,15] apply the approach in Ref. 50 to dynamic fleet management problems. Cheung and Powell [51] and Topaloglu and Powell [52] propose alternative methods to construct separable and concave value function approximations. Both of these methods are rooted in the stochastic approximation theory as well. Cheung and Chen [53] apply the method that is developed in Ref. 51 to an empty container allocation problem. Stochastic approximation methods date back to Robbins and Monro [54]. Kushner and Clark [55], Benveniste *et al.* [56], Bertsekas and Tsitsiklis [57], and Kushner and Yin [58] provide detailed coverage of the stochastic approximation theory.

MULTISTAGE PROBLEMS

In this section, we demonstrate how the ideas that we discuss in the section titled “Two-Stage Problems” can be extended to more complex resource management problems that take place over multiple time periods. We consider a setting where we have a set of resources that can be used to serve the demands that arrive randomly over time. At each time period, we observe the realization of the demand arrivals, and we need to decide which demands we should serve and to which locations we should reposition the unused resources. The objective is to maximize the total expected profit over a finite planning horizon.

The problem takes place over the planning horizon $\mathcal{T} = \{1, \dots, \tau\}$. For notational brevity, we assume that the travel times between all pairs of locations are one time period. We continue using $\mathcal{L}$ to denote the set of locations in the transportation network. To capture the decisions, we let x_{ijt} be the number of resources that we reposition from location i to j at time period t and y_{it}^k be the number of demands of type k that we serve by using a resource at location i at time period t . The cost of repositioning one resource from location i to j is c_{ij} . If we serve a demand

of type k by using a resource at location i , then we generate a revenue of p_i^k , and the resource ends up at location δ^k at the next time period. Similar to the convention that we use in the section titled “Two-Stage Problems,” if it is infeasible to serve a demand of type k by using a resource at location i , then we assume that $p_i^k = -\infty$. We use the random variable D_t^k to represent the number of demands of type k that arrive at time period t . We assume that the demand arrivals at different time periods are independent. We let r_{it} be the number of resources available at location i at the beginning of time period t . We note that $\{r_{i1} : i \in \mathcal{L}\}$ is a part of the problem data, whereas $\{r_{it} : i \in \mathcal{L}\}$ for $t \in \mathcal{T} \setminus \{1\}$ is determined by the decisions at time period $t - 1$.

We can formulate the problem as a dynamic program by using $r_t = \{r_{it} : i \in \mathcal{L}\}$ as the state variable at time period t . Letting $\mathbf{1}(\cdot)$ be the indicator function, and using $x_t = \{x_{ijt} : i, j \in \mathcal{L}\}$ and $y_t = \{y_{it}^k : i \in \mathcal{L}, k \in \mathcal{K}\}$ to capture the decisions and $D_t = \{D_t^k : k \in \mathcal{K}\}$ to capture the demand arrivals at time period t , the set of feasible decisions and state transitions are given by

$$\begin{aligned} \mathcal{X}(r_t, D_t) = & \left\{ (x_t, y_t, r_{t+1}) : \sum_{j \in \mathcal{L}} x_{ijt} \right. \\ & \left. + \sum_{k \in \mathcal{K}} y_{it}^k = r_{it}, \quad \forall i \in \mathcal{L} \right\} \quad (10a) \end{aligned}$$

$$\begin{aligned} & \sum_{i \in \mathcal{L}} x_{ijt} + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} \mathbf{1}(\delta^k = j) \\ & \times y_{it}^k - r_{j,t+1} = 0, \quad \forall j \in \mathcal{L}; \quad (10b) \end{aligned}$$

$$\sum_{i \in \mathcal{L}} y_{it}^k \leq D_t^k, \quad \forall k \in \mathcal{K}; \quad (10c)$$

$$x_{ijt}, y_{it}^k, r_{j,t+1} \in \mathbb{Z}_+, \quad \forall i, j \in \mathcal{L}, k \in \mathcal{K}. \quad (10d)$$

Constraint (10a) ensures that the number of resources that we reposition between different locations and use to serve the demands does not violate the resource availabilities. Constraint (10b) computes the number of resources available at different locations at the beginning of the next time period. Constraint 10c ensures that the number of demands that we serve does not exceed the demand arrivals. We can

find the optimal policy by computing the value functions $\{\bar{V}_t(\cdot) : t \in \mathcal{T}\}$ through the optimality equation

$$\begin{aligned} \bar{V}_t(r_t) = & \mathbb{E} \left\{ \max_{(x_t, y_t, r_{t+1}) \in \mathcal{X}(r_t, D_t)} - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ijt} \right. \\ & \left. + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_{it}^k + \bar{V}_{t+1}(r_{t+1}) \right\}, \end{aligned}$$

with the boundary condition that $\bar{V}_{\tau+1}(\cdot) = 0$ [59]. The optimality equation above involves a high dimensional state variable, which makes it difficult to compute the value functions for all possible values of the vector r_t . This difficulty is further compounded by the fact that the decisions (x_t, y_t) in the problem above are high dimensional vectors, and we take an expectation that involves the high dimensional demand arrivals D_t . Powell [60] uses the term “three curses of dimensionality” to refer to the difficulty due to the sizes of the state, decision, and demand vectors.

It turns out that all of the ideas in the section titled “Two-Stage Problems” can be extended to multistage problems. For economy of space, we only show extensions of the deterministic linear program and the separable projective approximations. Cheung and Powell [29] extend the demand relaxation strategy and Topaloglu [47] extends the supply relaxation strategy to multistage problems.

Deterministic Linear Program

Assuming that all demand arrivals take on their expected values, we can approximate the total expected profit over the planning horizon by using the optimal objective value of the problem

$$\begin{aligned} \max - & \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ijt} \\ & + \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_{it}^k \quad (11a) \end{aligned}$$

$$\text{subject to } \sum_{j \in \mathcal{L}} x_{ijt} + \sum_{k \in \mathcal{K}} y_{it}^k - r_{it} = 0,$$

$$\forall i \in \mathcal{L}, t \in \mathcal{T}; \quad (11b)$$

$$\begin{aligned} \sum_{i \in \mathcal{L}} x_{ijt} + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} \mathbf{1}(\delta^k = j) y_{it}^k \\ - r_{j,t+1} = 0, \quad \forall j \in \mathcal{L}, t \in \mathcal{T} \setminus \{\tau\}; \end{aligned} \quad (11c)$$

$$\sum_{i \in \mathcal{L}} y_{it}^k \leq \mathbb{E}\{D_t^k\}, \forall k \in \mathcal{K}, t \in \mathcal{T}; \quad (11d)$$

$$\begin{aligned} x_{ijt}, y_{it}^k, r_{it} \in \mathbb{R}_+, \forall i, j \in \mathcal{L}, \\ k \in \mathcal{K}, t \in \mathcal{T}. \end{aligned} \quad (11e)$$

Constraints (11b), (11c) and (11d) in problem 11 are respectively analogous to constraints (3b), (3c) and (3d) in problem 3. In the problem above, our understanding is that the values of the decision variables $\{r_{i1} : i \in \mathcal{L}\}$ are fixed by the number of resources that we have at different locations at the beginning of the planning horizon. By following the same hindsight approach that we use in the section titled “Deterministic Linear Program,” it is possible to show that the optimal objective value of problem 11 provides an upper bound on the optimal total expected profit. Such an upper bound becomes useful when we want to assess the optimality gaps of suboptimal approximation strategies.

We need minor modifications in problem 11 when we use this problem to make the resource repositioning and demand service decisions dynamically over time. In particular, if we are at time period t and the demand arrivals at the current time period are given by $\{D_t^k : k \in \mathcal{K}\}$, then we replace the right side of constraint (11d) for time period t with the current realization of the demand arrivals $\{D_t^k : k \in \mathcal{K}\}$. For the other time periods, we keep using the expected values of the demand arrivals. In addition, we fix the values of the decision variables $\{r_{it} : i \in \mathcal{L}\}$ to reflect the number of resources that we have at different locations at the current time period. Finally, we replace the set of time periods $\mathcal{T}$ with $\{t, \dots, \tau\}$ so that we only concentrate on the time periods in the future. Solving problem 11 after these modifications ensures that the decisions that we obtain for time period t are

feasible for the current state of the resources and the current realization of the demand arrivals.

Separable Projective Approximations

The idea behind the separable projective approximations is to construct approximations to the value functions by using sampled trajectories of the system. In particular, assuming that we have a set of approximations $\{\hat{V}_t(\cdot) : t \in \mathcal{T}\}$ to the value functions $\{\bar{V}_t(\cdot) : t \in \mathcal{T}\}$, if the number of resources at different locations is given by r_t and the realization of the demand arrivals is given by D_t , then we can make the decisions at time period t by solving the problem

$$\begin{aligned} \max_{(x_t, y_t, r_{t+1}) \in \mathcal{X}(r_t, D_t)} & - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ijt} \\ & + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_{it}^k + \hat{V}_{t+1}(r_{t+1}). \end{aligned}$$

Therefore, repeated solutions of the problem above for the successive time periods in the planning horizon would allow us to simulate the trajectory of the system under a particular realization of the demand arrivals. The challenge is to use the information that we obtain from the simulated trajectories of the system to update and improve the value function approximations.

We use the separable functions $\{V_t^{SPA}(\cdot) : t \in \mathcal{T}\}$ as our value function approximations. Similar to the development in the section titled “Separable Projective Approximations,” each one of these value function approximations is of the form $V_t^{SPA}(r_t) = \sum_{i \in \mathcal{L}} V_{it}^{SPA}(r_{it})$, where $V_{it}^{SPA}(\cdot)$ is a single dimensional, piecewise linear and concave function with points of nondifferentiability being a subset of integers. Letting Q be the total number of available resources, we can characterize $V_{it}^{SPA}(\cdot)$ by the sequence of slopes $\{v_{it}(q) : q = 0, \dots, Q-1\}$, where $v_{it}(q)$ is the slope of $V_{it}^{SPA}(\cdot)$ over the interval $[q, q+1]$.

The updating procedure in the section titled “Separable Projective Approximations,” together with the idea of simulating the trajectories of the system, immediately provides an iterative method that can be

- Step 1. Initialize the iteration counter by letting $n = 1$. Initialize the approximations $\{V_{it}^{SPA,1}(\cdot) : i \in \mathcal{L}\}$ to arbitrary piecewise linear and concave functions with points of nondifferentiability being a subset of integers.
- Step 2. Sample a realization of the demand arrivals over the whole planning horizon and denote this sample by $\{\hat{D}_t^{k,n} : k \in \mathcal{K}, t \in \mathcal{T}\}$.
- Step 3. Simulate the evolution of the system under the current demand arrivals and the current value function approximations.
- Step 3.a. Initialize the time by letting $t = 1$. Initialize the state of the resources $\hat{r}_1^n = \{\hat{r}_{i1}^n : i \in \mathcal{L}\}$ to reflect the initial state.
- Step 3.b. Solve the problem
- $$\tilde{V}_t(\hat{r}_t^n, \hat{D}_t^n) = \max_{(x_t, y_t, r_{t+1}) \in \mathcal{X}(\hat{r}_t^n, \hat{D}_t^n)} - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{L}} c_{ij} x_{ijt} + \sum_{i \in \mathcal{L}} \sum_{k \in \mathcal{K}} p_i^k y_{it}^k + \hat{V}_{t+1}^{SPA,n}(r_{t+1}),$$
- to obtain the optimal solution $(\hat{x}_t^n, \hat{y}_t^n, \hat{r}_{t+1}^n)$. Compute and store $\hat{v}_t^n = \tilde{V}_t(\hat{r}_t^n + e_i, \hat{D}_t^n) - \tilde{V}_t(\hat{r}_t^n, \hat{D}_t^n)$ for all $i \in \mathcal{L}$.
- Step 3.c. Increase t by one. If $t < \tau$, then go to Step 3.b.
- Step 4. Using α^n to denote the step size parameter at iteration n , update the value function approximations by letting $V_{it}^{SPA,n+1}(\cdot) = \mathcal{U}\left(V_{it}^{SPA,n}(\cdot), \hat{r}_{it}^n, \hat{v}_{it}^n, \alpha^n\right)$ for all $i \in \mathcal{L}, t \in \mathcal{T}$. Increase n by one and go to Step 2.

Figure 2. An iterative approach to construct separable value function approximations when there are multiple time periods in the planning horizon.

used to construct separable value function approximations. We describe this method in Fig. 2. In Step 1, we initialize the value function approximations. In Step 2, we sample a realization of the demand arrivals over the whole planning horizon. In Step 3, we simulate the evolution of the system under the current demand arrivals and the current value function approximations. In Step 4, we update the value function approximations to obtain the value function approximations that we use at the next iteration. The updating procedure $\mathcal{U}(V_{it}^{SPA,n}(\cdot), \hat{r}_{it}^n, \hat{v}_{it}^n, \alpha^n)$ in this step is identical to the one that we describe in the section titled “Separable Projective Approximations.”

A FLEXIBLE MODELING FRAMEWORK

The resource management problems that we have considered so far are relatively simplistic. In particular, the only distinguishing characteristic of a resource is its location, whereas the only distinguishing characteristic of a demand is its type. Our motivation for working with somewhat simplistic resource management problems is that such problems allow us to demonstrate the crucial algorithmic issues more clearly, but practical resource management problems can easily take on much more complex characteristics. For example, a driver resource in a driver scheduling application can be described by its inbound location, time to reach the inbound location, duty time within

the shift, days away from home, vehicle type, and home domicile. Conceptually, we can model such a complex resource by using multiple indices in our decision and state variables to capture the different attributes of the resource, but this approach would clearly be too cumbersome. In this section, we describe a flexible paradigm that is useful for modeling complex resource management problems.

Resources

To represent the resources, we let $\mathcal{A}$ be the attribute space of the resources. Roughly speaking, the attribute space represents the set of all possible states of a particular resource. We refer to each element of the attribute space as an attribute vector. We let r_{at} be the number of resources with attribute vector a at time period t . For example, if we assume that the set of locations is $\mathcal{L}$, the set of vehicle types is $\mathcal{V}$ and the travel time between any pair of locations is bounded by $\bar{V}$, then the attribute space for the vehicles in a simple dynamic fleet management application is $\mathcal{A} = \mathcal{L} \times \mathcal{V} \times \{0, 1, \dots, \bar{T}\}$. A vehicle with the attribute vector $a = (a_1, a_2, a_3) = (\text{New York}, \text{Large}, 4) \in \mathcal{A}$ is a vehicle of type large that is inbound to New York and will reach New York in four time periods. We can access the number of resources with attribute vector a at time period t by referring to r_{at} . This implies that we can lump the resources with the same attribute vector together and treat them as indistinguishable.

We use a similar approach to represent the demands. We let $\mathcal{B}$ be the attribute space of the demands and D_{bt} be the number of demands with attribute vector b at time period t . For example, if the set of expedition options is $\mathcal{E}$, then the attribute space for the demands may be $\mathcal{B} = \mathcal{L} \times \mathcal{L} \times \mathcal{E}$. A demand with the attribute vector $b = (b_1, b_2, b_3) = (\text{New York}, \text{Delaware}, \text{Express}) \in \mathcal{B}$ is an express demand with origin location New York and destination location Delaware. It may be not possible to serve an express demand from New York to Delaware by using a vehicle in New Jersey, but if the demand is not express, then it may be the case that a vehicle in New Jersey can serve

this demand. Therefore, the characteristics of the demand may dictate the feasible resource assignments.

Decisions

We use two classes of decisions. The class of decisions $\mathcal{D}^R$ captures the repositioning decisions, whereas the class of decisions $\mathcal{D}^S$ captures the demand service decisions. We let $\mathcal{D} = \mathcal{D}^R \cup \mathcal{D}^S$ to capture the set of all decision classes and refer to each element of $\mathcal{D}$ as a decision. For example, we may have $\mathcal{D}^R = \mathcal{L} \times \mathcal{L}$, and the decision $d = (d_1, d_2) = (\text{New York}, \text{Delaware}) \in \mathcal{D}^R$ may represent the decision to reposition a resource from New York to Delaware. Similarly, we may have $\mathcal{D}^S = \mathcal{B}$, and the decision $d = (b_1, b_2, b_3) \in \mathcal{D}^S = \mathcal{B}$ may represent the decision to serve a demand with the attribute vector b . It is reasonable to assume that for each decision $d \in \mathcal{D}^S$, there exists an attribute vector $b_d \in \mathcal{B}$ such that the decision d corresponds to serving a demand with attribute vector b_d .

We use x_{adt} to denote the number of resources with attribute vector a that we modify by using decision d at time period t . If we modify a resource with attribute vector a by using decision d at time period t , then we generate a profit contribution of c_{adt} . If it is not feasible to apply decision d on a resource with attribute vector a , then we capture this situation by letting $c_{adt} = -\infty$. In this case, we can capture the decisions at time period t by $x_t = \{x_{adt} : a \in \mathcal{A}, d \in \mathcal{D}\}$, the state of the resources by $r_t = \{r_{at} : a \in \mathcal{A}\}$, and the state of the demands by $D_t = \{D_{bt} : b \in \mathcal{B}\}$.

Optimization Problem

The set of feasible decisions at time period t is given by

$$\mathcal{X}(r_t, D_t) = \left\{ x_t : \sum_{d \in \mathcal{D}} x_{adt} = r_{at} \right. \\ \left. \forall a \in \mathcal{A} \right\} \quad (12a)$$

$$\sum_{a \in \mathcal{A}} x_{adt} \leq D_{b_d, t}, \forall d \in \mathcal{D}^S; \quad (12b)$$

$$x_{adt} \in \mathbb{Z}_+ \forall a \in \mathcal{A}, d \in \mathcal{D}, \quad (12c)$$

where we use the fact that for each decision $d \in \mathcal{D}^S$, there exists an attribute vector $b_d \in \mathcal{B}$ such that the decision d corresponds to serving a demand with attribute vector b_d . We note that constraints (12c) and (12b) are respectively analogous to constraints (10a) and (10c). We capture the outcome of applying a decision d on a resource with attribute vector a by defining

$$\delta_{a'}(a, d) = \begin{cases} 1 & \text{if applying the decision } d \text{ on} \\ & \text{a resource with attribute} \\ & \text{vector } a \text{ transforms the} \\ & \text{resource into a resource} \\ & \text{with attribute vector } a'; \\ 0 & \text{otherwise.} \end{cases}$$

The precise definition of $\delta_{a'}(a, d)$ is naturally dependent on the physics of the problem and how the system reacts to decisions. Using the definition above, we can write the dynamics of the resources as

$$r_{a,t+1} = \sum_{a' \in \mathcal{A}} \sum_{d \in \mathcal{D}} \delta_{a'}(a, d) x_{a'dt}, \quad \forall a \in \mathcal{A},$$

which is analogous to constraint (10b). The implicit assumption in the definition of $\delta_{a'}(a, d)$ is that the current attribute vector of the resource and the decision that we apply on the resource deterministically give the future attribute vector of the resource. This is indeed the case in many resource management applications, but there may be settings with random travel times and equipment failures where this assumption is not necessarily satisfied.

A Markovian policy π is a collection of decision functions $\{X_t^\pi(\cdot) : t \in \mathcal{T}\}$ such that each $X_t^\pi(\cdot)$ maps the state of the system (r_t, D_t) to a decision vector $x_t \in \mathcal{X}(r_t, D_t)$. Our goal is to find a policy that maximizes the total expected profit over the planning horizon. In other words, using $C_t(x_t) = \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} c_{adt} x_{adt}$ to denote the total profit associated with the decisions x_t at time period t and Π to denote the set of all Markovian policies, we want to solve the problem

$$\sup_{\pi \in \Pi} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}} C_t(X_t^\pi(r_t, D_t)) \right\}.$$

The formulation above provides a compact representation of the resource management problem that we are interested in, but it does not help from an algorithmic perspective. In particular, it is impossible to make a direct search over a reasonably broad class of policies Π . Therefore, the modeling framework that we describe in this section is useful for communicating the problem, but the actual solution requires a concrete algorithmic approach such as the ones described in the sections titled “Two-Stage Problems” and “Multistage Problems.”

The modeling framework that we describe in this section is due to Shapiro [61] and Powell *et al.* [62]. These works are intended to capture a broad class of resource management problems arising from a variety of application settings, and their presentation is significantly larger in scope than ours. Powell [60] uses this modeling framework to come up with a sound formulation for general dynamic programs, with a particular emphasis on the evolution of information over time. Shapiro and Powell [63] demonstrate how to use the modeling framework within the context of several algorithmic approaches to decompose large-scale resource management problems. Powell *et al.* [64], Powell and Topaloglu [65], Powell *et al.* [66], Powell and Topaloglu [22], Schenk and Klabjan [20], and Simao *et al.* [67] apply the modeling framework to dynamic fleet management, military airlift operations, express package routing, and driver scheduling problems. We note that the presentation in this article uses demands as the only source of uncertainty. The driver scheduling application in Ref. 67 is particularly interesting as it models a variety of sources of uncertainty, such as random travel times, equipment failures, and driver availability.

NUMERICAL ILLUSTRATIONS

In this section, we numerically compare the performances of the different approximation strategies that we describe in the sections titled “Two-Stage Problems” and “Multistage Problems.” Our goal here is to give a feel

for the relative performances of the different approximation strategies, rather than to present a full-scale experimental study.

Numerical Results for Two-Stage Problems

In this section, we provide numerical results for two-stage problems. The test problems that we use in our experimental setup have the same structure as the problem that we describe in the section titled “Two-Stage Problems.” We generate ten locations uniformly over a 100×100 region. Using $d(i, j)$ to denote the Euclidean distance between locations i and j , the cost of repositioning a resource from location i to j is given by $d(i, j)$. We have ten demand types and each demand type is associated with a particular location. The revenue from using a resource at location i to serve a demand type that is associated with location j is given by $200 - \rho \times d(i, j)$, where ρ is a parameter that we vary. As ρ gets larger, it becomes less profitable to serve a demand type that is associated with a particular location by using a resource at another location. In our computational experiments, we vary the coefficient of variation of the demand random variables and the parameter ρ . In particular, letting v be the coefficient of variation of the demand random variables, we label our test problems by the pair $(v, \rho) \in \{0.25, 0.5, 0.75, 1.0\} \times \{0.25, 0.5, 0.75, 1.0\}$. We note that the numerical results that we present in this section do not appear in the previous literature.

We test the performances of the approximation strategies described in the sections titled “Deterministic Linear Program,” “Decomposition by Demand Relaxation,” “Decomposition by Supply Relaxation,” and “Separable Projective Approximation,” which we respectively refer to as DLP, DDR, DSR, and SPA. DLP immediately provides a set of first stage decisions $\{\hat{x}_{ij} : i, j \in \mathcal{L}\}$. On the other hand, DDR, DSR, and SPA provide value function approximations that are all separable functions of the form $\sum_{i \in \mathcal{L}} \hat{V}_i(\cdot)$. To obtain a set of first stage decisions $\{\hat{x}_{ij} : i, j \in \mathcal{L}\}$, DDR, DSR, and SPA replace $\bar{V}(r)$ in the objective function of problem 2 with $\sum_{i \in \mathcal{L}} \hat{V}_i(r_i)$ and solve this problem. For DDR, DSR, and SPA, it is possible

to show that each one of $\{\hat{V}_i(\cdot) : i \in \mathcal{L}\}$ is a piecewise linear and concave function with points of nondifferentiability being a subset of integers. In this case, we can formulate problem 2 with the value function approximation $\sum_{i \in \mathcal{L}} \hat{V}_i(\cdot)$ as a minimum cost network flow problem [68]. A set of first stage decisions allow us to compute the cost that we incur at the first stage. In addition, we can compute the number of resources at different locations at the beginning of the second stage by $\hat{r}_j = \sum_{i \in \mathcal{L}} \hat{x}_{ij}$ for all $j \in \mathcal{L}$. To check the quality of the first stage decisions, we replace the right side of constraint (1b) in problem 1 with $\{\hat{r}_i : i \in \mathcal{L}\}$ and solve this problem for many realizations of demand arrivals. In this way, we approximate the total expected revenue in the second stage. The total expected profit is the difference between the total cost and the total expected revenue in the two stages.

Table 1 summarizes our numerical results. The first column in this table shows the problem characteristics by using the pair (v, ρ) . In the sections titled “Deterministic Linear Program,” “Decomposition by Demand Relaxation,” and “Decomposition by Supply Relaxation,” we show that DLP, DDR, and DSR provide upper bounds on the optimal total expected profit. The second, third, and fourth columns in Table 1 respectively show the upper bounds obtained by these three strategies. The fifth and sixth columns show the percent gap between the upper bounds obtained by DDR and the remaining two strategies. DDR provides the tightest upper bounds and we use it as a reference point. The seventh, eighth, ninth, and tenth columns respectively show the total expected profits obtained by DLP, DDR, DSR, and SPA. Finally, the eleventh, twelfth, and thirteenth columns show the percent gaps between the total expected profits obtained by SPA and the remaining three strategies. SPA generally obtains the highest total expected profits and we use it as a reference point.

Our results indicate that the gaps between the upper bounds obtained by DDR and the other two strategies can be quite significant. We observe that the gaps between the upper bounds get larger as the coefficient of

Table 1. Numerical Results for Two-Stage Problems.

Problem (v, ρ)	Upper Bound			DDR versus		Total Expected Profit				SPA versus		
	DLP	DDR	DSR	DLP	DSR	DLP	DDR	DSR	SPA	DLP	DDR	DSR
(0.25, 0.25)	18,292	17,628	18,187	3.76	3.17	16,782	16,881	16,897	16,908	0.74	0.16	0.06
(0.25, 0.5)	18,292	17,444	18,179	4.86	4.21	16,621	16,846	16,820	16,794	1.03	-0.31	-0.16
(0.25, 0.75)	18,292	17,287	18,167	5.81	5.09	16,478	16,689	16,750	16,765	1.71	0.45	0.09
(0.25, 1.0)	18,292	17,164	18,193	6.57	6.00	16,339	16,581	16,717	16,724	2.30	0.85	0.04
(0.5, 0.25)	18,275	16,744	18,035	9.15	7.71	15,213	15,551	15,519	15,553	2.18	0.01	0.21
(0.5, 0.5)	18,275	16,307	18,004	12.07	10.40	14,890	15,345	15,311	15,321	2.81	-0.16	0.07
(0.5, 0.75)	18,275	15,932	17,972	14.70	12.81	14,605	15,020	15,127	15,147	3.58	0.84	0.13
(0.5, 1.0)	18,275	15,627	17,949	16.94	14.86	14,328	14,789	14,949	14,971	4.29	1.21	0.14
(0.75, 0.25)	18,285	15,672	17,856	16.67	13.94	13,713	14,335	14,269	14,329	4.30	-0.04	0.42
(0.75, 0.5)	18,285	15,070	17,812	21.33	18.19	13,158	13,848	13,924	13,924	5.50	0.55	0.00
(0.75, 0.75)	18,285	14,527	17,772	25.87	22.34	12,627	13,501	13,574	13,576	6.99	0.55	0.01
(0.75, 1.0)	18,285	14,048	17,747	30.16	26.33	12,117	13,207	13,302	13,306	8.94	0.74	0.03
(1.0, 0.25)	18,260	13,868	17,484	31.67	26.07	11,990	12,695	12,578	12,714	5.69	0.14	1.07
(1.0, 0.5)	18,260	13,084	17,484	39.56	33.63	11,228	12,093	12,153	12,159	7.65	0.54	0.05
(1.0, 0.75)	18,260	12,349	17,484	47.86	41.58	10,481	11,583	11,632	11,637	9.94	0.47	0.04
(1.0, 1.0)	18,260	11,736	17,484	55.59	48.98	9,764	11,077	11,109	11,143	12.37	0.59	0.30

variation of the demand random variables increases and as it becomes more costly to serve a demand type from a location that is not associated with the demand type. We use the coefficient of variation as a proxy for the level of uncertainty in the problem, and the problem clearly becomes more difficult as the level of uncertainty increases. Furthermore, the substitution possibilities become more scarce as it becomes more costly to serve a demand type from a location that is not associated with the demand type. Therefore, we also expect the problem to be more difficult as ρ gets larger. These observations imply that the upper bounds provided by DDR are especially attractive for the test problems that we expect to be more difficult. The total expected profits obtained by DDR, DSR, and SPA are quite close to each other but SPA provides a small, consistent improvement over DDR and DSR. It is interesting that although the upper bounds obtained by DDR are significantly tighter than those obtained by DSR, the performances of DDR and DSR are quite similar. As a matter of fact, DSR appears to provide slightly better performance than DDR. For both DDR and DSR, we choose the Lagrange multipliers by using the dual solution to problem 3. Similar to our observations for the upper bounds, the gaps between the total

expected profits obtained by DLP and the remaining three strategies get larger as the coefficient of variation of the demand random variables increases or as it becomes more costly to serve a demand type from a location that is not associated with the demand type. Therefore, using a model that addresses the uncertainty in the system becomes especially beneficial for the test problems that we expect to be more difficult.

Numerical Results for Multistage Problems

In this section, we provide numerical results for multistage test problems. The results that we provide are taken from Ref. 18. In our experimental setup, we generate one base test problem and modify its different attributes to come up with test problems with different characteristics. In the base test problem, we have 20 locations. The problem takes place over a planning horizon of 60 time periods. The cost of repositioning a resource from location i to j is given by $c \times d(i, j)$, and we use $c = 4$. For each demand type k , there is an origin location o^k and destination location δ^k , and a demand of type k can only be served by a resource at location o^k . The revenue from serving a demand of type k is $p \times d(o^k, \delta^k)$, and we use $p = 5$. We have a total of 200 resources at our disposal.

Table 2. Numerical Results for Multistage Problems.

Problem	Total Expected Profit		SPA versus DLP
	DLP	SPA	
Base	155,378	166,913	6.91
$ \mathcal{L} = 10$	190,681	198,391	3.89
$ \mathcal{L} = 40$	162,132	172,021	5.75
$\sum_{i \in \mathcal{L}} r_{i1} = 100$	88,766	98,083	9.50
$\sum_{i \in \mathcal{L}} r_{i1} = 400$	250,448	263,775	5.05
$c = 1.6$	220,115	232,262	5.23
$c = 8.0$	113,671	121,617	6.53

Table 2 summarizes our numerical results. The first column in this table shows the problem characteristics that we modify to obtain different test problems. The first test problem corresponds to the base case. The second and third test problems have different number of locations in the transportation network. The fourth and fifth test problems have different number of resources. Finally, the sixth and seventh test problems have different repositioning costs. The second and third columns in Table 2 show the total expected profits obtained by DLP and SPA. The fourth column shows the percent gap between the total expected profits obtained by DLP and SPA. Our numerical experiments in the previous section indicate that SPA performs at least as well as DDP and DSR. Therefore, we only compare the performance of SPA with that of DLP.

The computational results indicate that the performance gap between DLP and SPA can be on the order of 4–10%. One trend worth mentioning is that the performance gaps increase as we have fewer resources at our disposal. In particular, the performance gap between SPA and DLP increases from 5.05% to 6.91% as the number of resources decreases from 400 to 200. Similarly, the performance gap between SPA and DLP increases from 6.91% to 9.50% as the number of resources decreases from 200 to 100. We consistently observed this trend in all of our numerical experiments. As the resources become more scarce, we need to allocate the resources more carefully, and SPA appears to do a better job of allocating scarce resources.

CONCLUSIONS

In this article, we describe a variety of modeling and solution approaches for transportation resource management problems. The solution approaches that we propose either build on deterministic linear programming formulations, or formulate the problem as a dynamic program and use tractable approximations to the value functions. We use two classes of methods to construct value function approximations. The first class of methods relax certain constraints in the dynamic programming formulation of the problem by associating Lagrange multipliers with them. In this case, we can solve the relaxed dynamic program by concentrating on one location at a time. The second class of methods use a stochastic approximation idea along with sampled trajectories of the system to iteratively update and improve the value function approximations. Our numerical experiments indicate that the models that explicitly address the randomness in the demand arrivals can provide significant benefits.

REFERENCES

1. Dantzig G, Fulkerson D. Minimizing the number of tankers to meet a fixed schedule. *Nav Res Log Q* 1954;1:217–222.
2. Ferguson A, Dantzig GB. The problem of routing aircraft - A mathematical solution. *Aeronaut Eng Rev* 1955;14:51–55.
3. White W, Bomberault A. A network algorithm for empty freight car allocation. *IBM Syst J* 1969;8(2):147–171.
4. White W. Dynamic transshipment networks: an algorithm and its application to the distribution of empty containers. *Networks* 1972;2(3):211–236.
5. Powell WB. A comparative review of alternative algorithms for the dynamic vehicle allocation problem. In: Golden B, Assad A, editors. *Vehicle routing: methods and studies*. Amsterdam: North Holland Publishing Co.; 1988. pp. 249–292.
6. Powell WB. A stochastic formulation of the dynamic assignment problem, with an application to truckload motor carriers. *Transp Sci* 1996;30(3):195–219.

7. Erera A, Karacik B, Savelsbergh M. A dynamic driver management scheme for less-than-truckload carriers. *Comput Oper Res* 2008;35:3397–3411.
8. Hane C, Barnhart C, Johnson E, *et al.* The fleet assignment problem: solving a large-scale integer program. *Math Program* 1995;70:211–232.
9. Holmberg K, Joborn M, Lundgren JT. Improved empty freight car distribution. *Transp Sci* 1998;32:163–173.
10. Gorman M. Intermodal pricing model creates a network pricing perspective at BNSF. *Interfaces* 2001;31(4):37–49.
11. Frantzeskakis L, Powell WB. A successive linear approximation procedure for stochastic dynamic vehicle allocation problems. *Transp Sci* 1990;24(1):40–57.
12. Crainic T, Gendreau M, Dejax P. Dynamic and stochastic models for the allocation of empty containers. *Oper Res* 1993;41:102–126.
13. Carvalho TA, Powell WB. A multiplier adjustment method for dynamic resource allocation problems. *Transp Sci* 2000;34:150–164.
14. Godfrey GA, Powell WB. An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: single period travel times. *Transp Sci* 2002a;36(1):21–39.
15. Godfrey GA, Powell WB. An adaptive, dynamic programming algorithm for stochastic resource allocation problems II: multi-period travel times. *Transp Sci* 2002b;36(1):40–54.
16. Kleywegt AJ, Nori VS, Savelsbergh MWP. The stochastic inventory routing problem with direct deliveries. *Transp Sci* 2002;36(1):94–118.
17. Adelman D. A price-directed approach to stochastic inventory routing. *Oper Res* 2004;52(4):499–514.
18. Topaloglu H, Powell WB. Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. *INFORMS J Comput* 2006;18(1):31–42.
19. Adelman D. Price-directed control of a closed logistics queueing network. *Oper Res* 2007;55(6):1022–1038.
20. Schenk L, Klabjan D. Intra market optimization for express package carriers. *Transp Sci* 2008;42:530–545.
21. Powell WB, Jaillet P, Odoni A. Stochastic and dynamic networks and routing. In: Monma C, Magnanti T, Ball M, editors. *Handbook in operations research and management science*, volume on Networks. Amsterdam: North Holland Publishing Co., 1995. pp. 141–295.
22. Powell WB, Topaloglu H. Fleet management. In: Wallace S, Ziemba W, editors. *Applications of stochastic programming*. MPS-SIAM series in optimization. Philadelphia (PA): MPS-SIAM; 2005.
23. Barnhart C, Laporte G, editors. *Handbooks in operations research and management science*, volume on Transportation. Amsterdam: North Holland Publishing Co.; 2006.
24. Bertsimas D, Popescu I. Revenue management in a dynamic network environment. *Transp Sci* 2003;37:257–277.
25. Cheung R. Dynamic networks with random arc capacities, with application to the stochastic dynamic vehicle allocation problem [PhD thesis]. Princeton (NJ): Princeton University; 1994.
26. Powell WB, Cheung R. A network recourse decomposition method for dynamic networks with random arc capacities. *Networks* 1994a;24:369–384.
27. Powell WB, Cheung RK. Stochastic programs over trees with random arc capacities. *Networks* 1994b;24:161–175.
28. Cheung RKM, Powell WB. Models and algorithms for distribution problems with uncertain demands. *Transp Sci* 1996a;30(1):43–59.
29. Cheung RK, Powell WB. An algorithm for multistage dynamic networks with random arc capacities, with an application to dynamic fleet management. *Oper Res* 1996b;44(6):951–963.
30. Kunnumkal S, Topaloglu H. Linear programming based decomposition methods for inventory distribution systems. Technical report. Ithaca (NY): Cornell University, School of Operations Research and Information Engineering; 2008. Available at <http://legacy.orie.cornell.edu/~huseyin/publications/publications.html>.
31. Ruszczyński A. Decomposition methods. In: Ruszczyński A, Shapiro A, editors. *Handbook in operations research and management science*, stochastic programming. Amsterdam: North Holland Publishing Co.; 2003. pp. 141–211.
32. Adelman D, Mersereau AJ. Relaxations of weakly coupled stochastic dynamic programs. *Oper Res* 2008;56(3):712–727.
33. Castanon DA. Approximate dynamic programming for sensor management. *Proceedings of the 36th Conference on Decision &*

- Control. San Diego (CA); 1997.
34. Yost KA, Washburn AR. The LP/POMDP marriage: Optimization with imperfect information. *Nav Res Log Q* 2000;47:607–619.
 35. Bertsimas D, Mersereau AJ. A learning approach for interactive marketing to a customer segment. *Oper Res* 2007;55(6):1120–1135.
 36. Federgruen A, Zipkin P. A combined vehicle routing and inventory allocation problem. *Oper Res* 1984;32:1019–1037.
 37. Topaloglu H, Kunnumkal S. Approximate dynamic programming methods for an inventory allocation problem under uncertainty. *Nav Res Log Q* 2006;53:822–841.
 38. Kunnumkal S, Topaloglu H. A duality-based relaxation and decomposition approach for inventory distribution systems. *Nav Res Log Q* 2008a;55(7):612–631.
 39. Erdelyi A, Topaloglu H. Using decomposition methods to solve pricing problems in network revenue management. Technical report. Ithaca (NY): Cornell University, School of Operations Research and Information Engineering; 2009. Available at <http://legacy.orie.cornell.edu/~huseyin/publications/publications.html>.
 40. Kunnumkal S, Topaloglu H. A new dynamic programming decomposition method for the network revenue management problem with customer choice behavior. *Prod Oper Manage* 2009. In press.
 41. Topaloglu H. Using Lagrangian relaxation to compute capacity-dependent bid-prices in network revenue management. *Oper Res* 2009b;57(3):637–649.
 42. Zhang D, Adelman D. An approximate dynamic programming approach to network revenue management with customer choice. *Transp Sci* 2009;42(3):381–394.
 43. Talluri KT, van Ryzin GJ. The theory and practice of revenue management. New York: Springer; 2005.
 44. Kunnumkal S, Topaloglu H. A refined deterministic linear program for the network revenue management problem with customer choice behavior. *Nav Res Log Q* 2008c;55(6):563–580.
 45. Kunnumkal S, Topaloglu H. Computing time-dependent bid prices in network revenue management problems. *Transp Sci* 2010;44(1):38–62.
 46. Erdelyi A, Topaloglu H. A dynamic programming decomposition method for making overbooking decisions over an airline network. *INFORMS J Comput* 2010;22:443–456.
 47. Topaloglu H. A tighter variant of Jensen's lower bound for stochastic programs and separable approximations to recourse functions. *Eur J Oper Res* 2009a;199:315–322.
 48. Powell WB, Ruszczyński A, Topaloglu H. Learning algorithms for separable approximations of stochastic optimization problems. *Math Oper Res* 2004;29(4):814–836.
 49. Topaloglu H. A parallelizable dynamic fleet management model with random travel times. *Eur J Oper Res* 2006;175:782–805.
 50. Godfrey GA, Powell WB. An adaptive, distribution-free approximation for the newsvendor problem with censored demands, with applications to inventory and distribution problems. *Manage Sci* 2001;47(8):1101–1112.
 51. Cheung RKM, Powell WB. SHAPE: a stochastic hybrid approximation procedure for two-stage stochastic programs. *Oper Res* 2000;48(1):73–79.
 52. Topaloglu H, Powell WB. An algorithm for approximating piecewise linear functions from sample gradients. *Oper Res Lett* 2003;31:66–76.
 53. Cheung R, Chen C. A two-stage stochastic network model and solution methods for the dynamic empty container allocation problem. *Transp Sci* 1998;32:142–162.
 54. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat* 1951;22:400–407.
 55. Kushner HJ, Clark DS. Stochastic approximation methods for constrained and unconstrained systems. Berlin: Springer; 1978.
 56. Benveniste A, Metivier M, Priouret P. Adaptive algorithms and stochastic approximations. New York: Springer; 1991.
 57. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming. Belmont (MA): Athena Scientific; 1996.
 58. Kushner HJ, Yin GG. Stochastic approximation algorithms and applications. New York: Springer; 1997.
 59. Puterman ML. Markov decision processes. New York: John Wiley and Sons, Inc.; 1994.
 60. Powell WB. Approximate dynamic programming: Solving the curses of dimensionality. Hoboken (NJ): John Wiley & Sons, Inc.; 2007.
 61. Shapiro JA. A framework for representing and solving dynamic resource transformation problems [PhD thesis]. Princeton (NJ): Department of Operations Research and

- Financial Engineering, Princeton University; 1999.
62. Powell WB, Shapiro JA, Simão HP. A representational paradigm for dynamic resource transformation problems. In: Coullard RFC, Owens JH, editors. *Annals of operations research*. New York: J.C. Baltzer AG; 2001. pp. 231–279.
 63. Shapiro JA, Powell WB. A metastrategy for large-scale resource management based on informational decomposition. *INFORMS J Comput* 2006;18(1):43–60.
 64. Powell WB, Shapiro JA, Simão HP. An adaptive dynamic programming algorithm for the heterogeneous resource allocation problem. *Transp Sci* 2002;36(2):231–249.
 65. Powell WB, Topaloglu H. Stochastic programming in transportation and logistics. In: Ruszczyński A, Shapiro A, editors. *Handbook in operations research and management science*, volume on stochastic programming. Amsterdam: North Holland Publishing Co.; 2003. pp. 555–635.
 66. Powell WB, Wu T, Simão HP, *et al.* Using low dimensional patterns in optimizing simulators: an illustration for the military airlift problem. *Math Comput Model* 2004;29:657–675.
 67. Simao HP, Day J, George AP, *et al.* An approximate dynamic programming algorithm for large-scale fleet management: a case application. *Transp Sci* 2009;43(2):178–197.
 68. Nemhauser G, Wolsey L. *Integer and combinatorial optimization*. Chichester: John Wiley & Sons, Inc.; 1988.

TRAVEL DEMAND MODELING

FRANK SOUTHWORTH
Oak Ridge National Laboratory, Oak
Ridge, Tennessee

School of Civil and Environmental
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

LAURIE A. GARROW
School of Civil and Environmental
Engineering, Georgia Institute of
Technology, Atlanta, Georgia

INTRODUCTION

Methods for estimating and forecasting travel demands draw on a number of different disciplines, making use of optimization methods embedded in econometric as well as mathematical programming frameworks, and using microsimulation and agent-based modeling to generate travel activity patterns from the ground up for entire regional populations of travelers and freight movers. These methods are now well represented in software programs which in today's computers generate demand estimates in anything from a few seconds to a few hours, depending on the size of the task at hand. In this article, we describe the most common travel demand modeling approaches, the nature of the solution methods used, and the types of data used to fit or "calibrate" them to real-world conditions. A feature of the travel demand literature over the past half century has been the variety of methods that have been applied, with method selection often determined by the nature and quality of available data sources as well as by specific solution needs. Current models can be categorized in a number of ways. One obvious distinction is between passenger demand and freight demand models, although both use similar solution methods. From a practitioners' standpoint the following classification also proves useful:

- location-specific, time series-based demand models;
- cross-sectional modeling of origin-to-destination travel, usually broken down by travel mode and/or travel route, as well as by trip purpose or class of commodity shipped.

A few examples of combined time series/cross-sectional travel demand models have also been developed, but to date have seen limited use due principally to limited data with which to calibrate them.

A second important distinction is between what has been called the *trip based approach*, which has been the dominant approach in practice since the 1950s, and the more recent and increasingly popular "activity-based approach" to travel demand modeling. The differences between these two approaches are brought out below.

TIME SERIES-BASED DEMAND MODELS

Time series data is often used to develop forecasts of the demand for specific transportation facilities such as airports [1–3] and seaports [4,5], as well as for high volume links within a transportation network [6,7]. The empirical data used in these models may span a few weeks, a few months, or a few years of prior passenger, freight, or vehicular traffic volume data, and the forecast may be short term, covering the next few weeks or months, or long term, extending projections a number of years into the future. The historical data set may include data for a single facility or location or may combine data from a number of different locations, such as vessel arrival or departure data collected for all of the ports in a given region or country. In some cases, a simple univariate time series model is used, based solely on projecting a past trend in traffic growth or decline forward in time; or a multivariate, causal model may be applied that relates the temporal trend in facility or regional demand to

temporal trends in a number of explanatory variables: such as trends in travel prices, disposable incomes, industrial output, and growth in the size of selected traveling populations. In some instances [8], time series data from one modal activity (e.g., number of offloaded seaport containers) has been used to forecasts travel demand for another mode (e.g., the number of trucks serving a seaport).

Past studies offer a variety of methods to choose from, including simple growth factor methods, multiple regressions, exponential smoothing, Kalman filters, multinomial tobits and probits, and artificial neural network models, as well as a variety of autoregressive integrated moving average (ARIMA) models, in the tradition of Box and Jenkins [9], including variable (VARIMA) and seasonal (SARIMA) models. Commercial software is available to estimate model parameters, with most solutions based on well-established techniques for either minimizing a residual sum-of-squares function or maximizing a likelihood function associated with the observed time series. Similar techniques have been used in both passenger and freight demand forecasting.

CROSS-SECTIONAL, TRIP-BASED DEMAND MODELING

Once the issue becomes one of forecasting the demand for multiple origin-to-destination (O-D) travel movements, and the geographic context moves away from a single facility or travel corridor, then data limitations play a major role in model selection, as they have done in model development over the past 50 years. In particular, when we enter the realm of regional transportation planning, we are dealing with the estimation and forecasting of hundreds, even thousands, of spatially explicit O-D movements within, as well as into, out of, and through a planning region (see the popular textbooks by Meyer and Miller [10] and Ortuzar and Willumsen [11]).

The high costs of collecting this sort of O-D data on personal or business travel activities means that most demand models are based on the responses provided by a sample of households or companies to a travel

questionnaire, from which are developed a hopefully representative set of daily trip rates (trip frequencies) and trip lengths, broken down by trip purpose or commodity class, as well as data on the modes and sometimes also the routes selected. While this approach can work quite well for the purpose of understanding current activity rates and even the geographic range of individual travel activity patterns, it is necessarily only a surrogate for a proper behavioral analysis of how such travel decisions as mode and destination choices evolve over time, and as households and businesses themselves evolve. A preferred method for capturing such behavioral connections is to carry out panel surveys, in which the same establishment (i.e., a household or a business) is resurveyed one or more times after a suitable time interval: the challenge being to avoid attrition in the sample, especially as the time interval increases [12]. Such sampling typically occurs in two or more “waves” and as such introduces an element of time-series modeling into the estimation process. To date, however, only a limited number of such panel data sets have been collected and applied. Most planning models make use of travel surveys that draw a different sample of respondents from one survey to the next, often at an interval of 5 or more years. As a result, most applications of regional travel demand models have placed a good deal of reliance on the variability caused by a combination of differences in a respondent’s household or company structure along with differences in responses created by the different geographic locations and site-specific conditions they find themselves dealing with.

Disaggregate Behavioral Models

Discrete choice (or “logit”) models based on utility maximization theory are commonly used to represent how individuals or companies make travel and shipment decisions. Discrete choice models are used to predict the probability a decision maker will choose one alternative among a finite set of mutually exclusive and collectively exhaustive alternatives. A decision maker can represent an individual, a group of individuals, a government, a corporation, and so on. Discrete

choice models are used to estimate and forecast the total demand for a specific travel alternative as the sum of a set of individual traveler or (where freight is involved) shipper choices, usually based on traveler or shipper perceptions of the utility of one alternative over another.

Utility is a scalar index of value that is a function of attributes and/or individual characteristics. Utility represents the “value” an individual places on different attributes and captures how individuals make trade-offs among these attributes. Individuals are assumed to select the alternative that has the maximum utility. Alternative i is chosen if the utility of individual n obtains from alternative i , U_{ni} , is greater than the utility for all other alternatives. This utility U_{ni} , has an observed component, V_{ni} , and an unobserved component, commonly referred to as an *error term*, ε_{ni} . Formally, $U_{ni} = V_{ni} + \varepsilon_{ni}$, where

$$V_{ni} = \beta' x_{ni}.$$

Specific choice probabilities for different discrete choice models are obtained by imposing different assumptions on the distribution of these error terms and/or by assuming that β varies across decision makers. The assumption that unobserved error components are independently and identically distributed (iid) and follow a Gumbel distribution with mode zero and scale one, $\varepsilon \sim \text{iid } G(0, 1)$, results in the binary logit (in the case of two alternatives) or the multinomial logit model, or MNL, (in the case of more than two alternatives) [13]. Under these assumptions, MNL choice probabilities are given as

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j \in C_n} \exp(V_{nj})}, \quad (1)$$

where C_n is the set of alternatives (or choice set) faced by individual n . The assumption that the error terms are iid $G(0, 1)$ is advantageous in the sense that the choice probability takes on a closed-form expression that is computationally simple. However, the same assumption imposes several restrictions on the binary logit and MNL models. The first

key restriction is the independence of irrelevant alternatives (IIA), a property which states that the ratio of choice probabilities P_{ni}/P_{nj} for $i, j \in C_n$ is independent of the attributes of any other alternative. In terms of substitution patterns, this means a change or improvement in the utility of one alternative will draw share proportionately from all other alternatives. In many applications, this may not be a realistic assumption. For example, in mode choice model applications, one would expect that bus would compete more with other public transit modes than auto. Consequently, much of the research in the travel demand modeling community over the past 30–40 years has been focused on incorporating more flexible substitution patterns. The most commonly used model in practice that helps relax the IIA property while maintaining a closed-form expression for choice probabilities is the nested logit (NL) model [14,15]. Dozens of other models have been developed for travel demand modeling applications, including those belonging to the generalized “hierarchical” or “nested” logit (GNL, generalized nested logit) class [16] and/or the network generalized extreme value (NetGEV) class [17]. Garrow [18] provides an extensive review of these models in her text and focuses on applications of these models to the airline industry. Ben-Akiva and Lerman [19], Train [20], and Koppelman and Bhat [21] are other key introductory texts on applying discrete choice models for travel demand applications.

The second key limitation of MNL models is that the β parameters represent “average” population values. The multinomial probit (MNP) model [22] was one of the first discrete choice models that was used to incorporate more flexible substitution patterns and/or random taste variation. However, due to computational limitations, the ability to use the probit model was limited. It was not until the past decade that an alternative to the probit model, namely the mixed logit model, was introduced.

Conceptually, the mixed logit model is identical to the MNL model except that the parameters of the utility functions for mixed models can vary across individuals, alternatives, and/or observations. However,

like the probit model, this added flexibility comes at a cost—choice probabilities can no longer be expressed in closed form. Due to improvements in computational power, however, it is now feasible to use mixed logit models for travel demand modeling applications. Indeed, the number of publications using mixed logit models has expanded exponentially since 2003 and mixed logit models have been applied in numerous other transportation contexts spanning activity-based planning and rescheduling behavior models [23–25], mode choice models [26,27], residential location/relocation decisions [28], and consideration of physical activity in choice of mode [29].

Similar models have recently become popular with freight modelers. Some authors have also combined traditionally separate choice dimensions (see the section titled “Aggregate, Four-Step Planning Models”). For example, Train and Wilson [30] used an MNL logit to analyze the choice of mode by agricultural product shippers, making use of both revealed and respondent state preference data to calibrate their model. Jiang *et al.* [31] use an NL model to estimate the demand for alternative freight modes in which all for-hire modes are compared within the same nest, and their averaged, inclusive value utility was then compared with the private carriage (in Europe own-account) modal service option. A heteroscedastic extreme value (HEV) choice model has also been used in freight mode choice applications [32,33]. See Jeffs and Hills [34] for a discussion of the explanatory variables often found to be significant in freight mode choice.

Aggregate, Four-Step Planning Models

In the United States, each metropolitan area as well as each state develops databases and models to estimate current, or baseline, traffic volumes using its major transportation facilities. It then links these estimates to regional economic activity and demographic forecasts in order to predict future traffic flows. These forecasts are used in turn to inform the investment of large sums of public money in expanding or improving transportation services within the region.

For the past 50 years or so, this process has been supported, notably in large urban or metropolitan regions, by the four-step transportation planning model shown in Fig. 1. Most commonly applied to passenger, or personal travel forecasting, this approach begins by modeling the frequency with which households engage in travel, estimating the number of trips generated daily for a number of different trip purposes. This usually involves separate model calibrations, for example, for commuting to work, shopping, traveling to school, and making personal business, social and recreational trips; with a distinction also common between home-based and non-home-based trips. The four sequential modeling steps can be represented in very general terms as

Trip (or Freight) Generation:

$$O_{ip} = F(X_{ip})$$

and Attraction:

$$D_{jp} = F(X_{jp}) \quad (2a)$$

Trip (or Freight) Distribution (Flow):

$$T_{ijp} = F(O_{ip}, D_{jp}, c_{ijp}) \quad (2b)$$

Modal Split:

$$T_{ijpk} = F(T_{ijp}, g_{ijpk}, c_{ijpk}) \quad (2c)$$

Route Assignment:

$$\Sigma_p T_{ijpkr} = T_{ij^*kr} = F(T_{ij^*k}, g_{ij^*kr}, c_{ij^*kr}) \quad (2d)$$

where $F()$ means “a function of,” and where i = a trip origin ($i = 1, 2, \dots, I$), j = a trip destination ($j = 1, 2, \dots, J$), O_i = the number of trip origins out of i , D_j = the number of trip destinations into j , T = a daily (or annual) measure of travel demand, and T_{ijpkr} , for example, refers to the number of trips taken daily (or annually) from i to j for purpose p by mode k and route r ; c = transportation costs, and c_{ijkr} for example, refers to the cost of transporting a passenger or unit of freight from i to j by mode k on route r . Finally, X_i , X_j , and g refer to matrices of explanatory variables associated with specific steps in the flow estimation process.

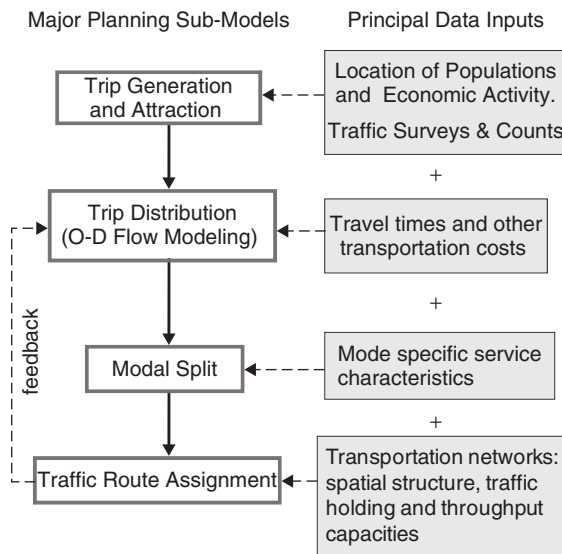


Figure 1. Four-step transportation planning model.

The principal reason for the sequential nature of this approach was the failure of early modeling efforts to capture each of these four policy relevant aspects of travel demand within a single model calibration step: with the variables best explaining mode choice, for example, differing from those yielding good estimates of the other three choices. Efforts to do so, usually referred to as *direct demand models* in the literature, are still used, and have been found to be valuable and cost effective in specific situations, such as for sketch planning purposes and/or for use in small community demand forecasting [35,36]. For the most part, these models use standard multiple regression software. However, most applications in the United States today employ the above four-step process when developing regional travel forecasts.

For the purpose of producing planning and policy relevant outputs, each step in the process is executed at an aggregate, or traffic analysis zone (TAZ) level, with the largest metropolitan regions broken down into hundreds of TAZs for the purpose of estimating O-D flows between all TAZ pairs. In practice, such TAZs are often equated with Census blocks, block-groups, or tracts, in order to make use of federally supplied estimates of the number of households and businesses

within each TAZ. At the statewide level in the United States, and at the national level in many smaller countries, such TAZs may be equated with counties. Within metropolitan areas, the focus is on automobile and truck flows, as well as public transit bus and rail movements. At the statewide level, other nonhighway modes enter the picture, notably air, rail, and water transport.

Variants on the same four-step modeling process have also been used to forecast daily and annual freight flow volumes, with the focus on truck freight movements over urban and intercity highways, and using a three-step process where modal split is not required, but adding an additional submodel where flows are estimated initially not in number of trips but in number of tons of different commodities shipped, which then need to be translated into the number of vehicle (i.e., truck) trips using suitable tons/truck loading factors. A popular functional form for all four modeling steps, for both passenger and freight demand modeling, is the logit model, with linear and log-linear regressions also popular for the trip generation stage. However, most regional transportation planning models then merge the truck freight and passenger automobile flows to produce passenger car equivalent (pce) O-D

flow matrices, in which the larger trucks are allocated pce values much greater than 1.0 before assigning them to the same highway network. Where this loading leads to high levels of traffic volume to capacity (v/c) ratios, the use of congestion-sensitive route choice algorithms have become the norm, notably traffic route assignments based on Wardrop's [37] user optimal equilibrium principle. In this approach travelers select their routes in such a way that their aggregate flow patterns lead to identical travel times between all routes used to travel between any given O-D pair of places.

Finally, a look at Figure 1 also shows an important feedback loop between the trip destination, mode and route choice submodels. By iterating in this manner, and maintaining convexity in the submodel solutions, it is possible to bring the flows and costs at each modeling step to a stable solution: such that no further significant changes occur in either O-D costs or O-D trip volumes, that is, bringing destination, mode, and route choices to a cost-flow balance. Other feedback loops can be tried, and this feedback loop has not always been applied in the past. However, extension of this procedure to trip generation is rarely attempted using traditional model formulations, largely because daily trip frequencies are relatively insensitive to trip cost differences at the level of spatial aggregation to which these planning models are typically applied. Indeed, the entire four-step sequential modeling process has received growing criticism from a number of quarters over the past four decades, not only for its limited behavioral basis but also for its lack of a cohesive conceptual framework. One response has been the rapid evolution of the disaggregate demand models reviewed above. A second response has been to develop more closely integrated, "combined" choice models that work through economically based equilibrium formulations of the entire flow modeling process [38]. The text by Oppenheim [39] describes and links both approaches together. These combined models use technical advances in nonlinear mathematical programming solution techniques, model formulations based on variational inequalities, and computationally

efficient solution methods tied to link-node, path-based network formulations of not only the physical traffic flows themselves but also of the entire process of identifying and tracking travel choices through numerous decision making steps.

Mathematical Programming Formulations

The logit discrete choice model can also be derived as an optimization problem, which in its aggregate, planning level application has well-known convex programming solution. Texts by Wilson [40], Erlander and Stewart [41], and Lee [42], among others, explore these formulations. For example, for a set of annual (or daily) trip volumes, T_{ij} , between origin locations $i = 1, 2, \dots, I$ and destination locations $j = 1, 2, \dots, J$, Wilson [40] shows that the following *entropy maximizing* trip distribution (destination) model

$$\text{Max } L(T) = -\sum_i \sum_j (T_{ij}/T) \ln(T_{ij}/T) \quad (3)$$

subject to:

$$\begin{aligned} \sum_j T_{ij} &= O_i \quad \text{for all } i \quad \text{and} \quad \sum_i T_{ij} \\ &= D_j \quad \text{for all } j \end{aligned} \quad (4)$$

$$\begin{aligned} \text{cbar} &= (1/T)^* \sum_i \sum_j T_{ij}^* c_{ij} \\ \text{and all } T_{ij} &\geq 0 \end{aligned} \quad (5)$$

was solved by the following *doubly constrained* spatial interaction model:

$$T_{ij} = A_i^* B_j^* O_i^* D_j^* \exp(-\beta^* c_{ij}), \quad (6)$$

where O_i = the volume of travel (e.g., number of trips) generated by TAZ i , D_j = the volume of travel attracted to TAZ j , and A_i and B_j are two sets of matrix "balancing factors" given as

$$A_i = 1 / \left[\sum_j B_j^* D_j^* \exp(-\beta^* c_{ij}) \right] \quad \text{for all } i \quad (7)$$

and

$$B_j = 1 / \left[\sum_i A_i^* O_i^* \exp(-\beta c_{ij}) \right] \quad \text{for all } j, \quad (8)$$

which, via iterative application, satisfy the trip end constraints given by Equation (4). Here β represents the sensitivity of travelers (or in the case of freight movements, shippers) to additional travel cost. Applying a suitable nonlinear search method, a value of β is sought that satisfies the system's average cost constraint above, while also meeting the two sets of trip end constraints. A little thought here also shows that setting all B_j 's (A_i 's) to a value of 1.0 yields a trip origination (trip attraction) constrained model in aggregate logit form.

Subsequently, Erlander [43] showed how this model can also be restated as an equivalent cost-minimization model of the form

Minimize $C(T)$

subject to:

$$\sum_j T_{ij} = O_i \quad \text{for all } i \quad \text{and}$$

$$\sum_i T_{ij} = D_j \quad \text{for all } j$$

$$L(T) = -\sum_i \sum_j (T_{ij}/T) \ln (T_{ij}/T) \geq L_0,$$

$$\text{and all } T_{ij} \geq 0,$$

where L_0 = a minimum acceptable level of trip dispersion.

Numerous extensions of this sort of approach have been used since the 1970s to formulate and solve increasingly more elaborate travel demand (O–D flow inclusive) models, including the “joint” solution of optimal O–D flow and traffic route assignment models [43] as well as jointly optimized O–D, mode, and route choice solutions in freight [44] as well as passenger travel. These approaches also set the table for the solution of much broader optimization problems that include optimal land use as well as optimal travel pricing mechanisms within their spatially explicit price-versus-volume based equilibrium solutions [42,45].

Similar programming models have also been applied to freight demand estimation and forecasting, but with the T_{ij} 's increasingly solved for in terms of tons shipped, rather than number of vehicle trips, notably by combining Leontief's input–output approach to capturing interindustry interactions with the spatial interaction models of the type represented by Equations (6)–(8)

above to create interregional or multiregional input–output models of commodity flows. When combined with logit or other discrete choice model of mode selection (e.g., of truck vs rail vs waterway vs intermodal truck–rail or truck–water transport), these modeling frameworks are now beginning to be applied in statewide freight demand forecasting in the United States as well as in nationwide freight demand forecasting in Europe and elsewhere (see the review of freight demand models by Southworth [46]).

ACTIVITY-BASED APPROACHES, MICROSIMULATION, AND AGENT-BASED MODELING TOOLS

All of the modeling approaches described in the section titled “Cross-Sectional, Trip Based Demand Modeling” have been applied to forecasting the number of passenger or freight trips associated with a region's population distribution and its travel generating land use patterns. In doing so, these trips are effectively abstracted from the often complex set of family influenced decisions being made by individual travelers and the company logistics-based decisions being made by shippers of goods. In effect, the approach has proved to be a very cost-effective way of simplifying the travel demand process: by treating nontravel decisions as largely exogenous to the production of trips. Also, beginning in the 1950s [47], but only gaining any noticeable level attention by the 1970s, was a more behaviorally attractive but at the same time more conceptually as well as computationally demanding *activity-based approach* (ABA) to travel demand modeling. In this approach, travel activities are more closely linked to the complex scheduling decisions associated with a good deal of household travel, or as McNally and Rindt [47, Section 3.1] put it: “The ABA takes as the basic unit of analysis the travel-activity pattern, defined as the revealed pattern of behavior represented by travel and activities (both in-home and non-home) over a specific time period (often a single day).” Hence an individual's single day travel activity schedule might look something like that shown in Fig. 2, in this instance for a two-adult-one-child family.

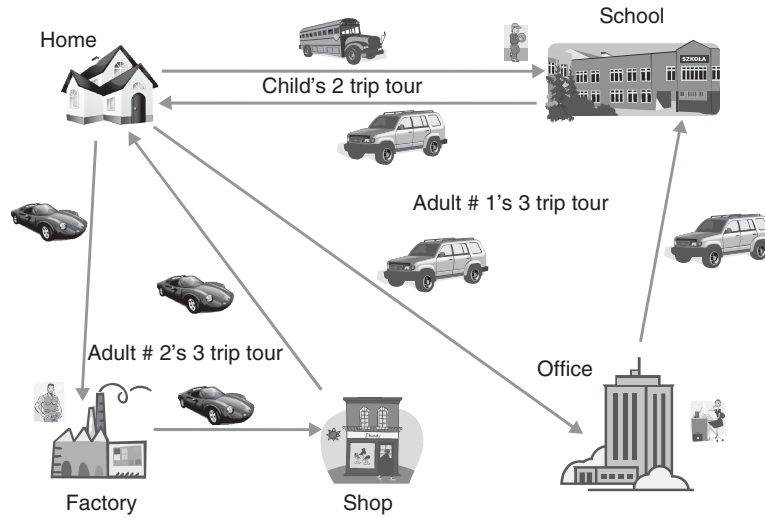


Figure 2. A daily multitrip chaining example for a three-person household.

A little thought about the many variables that might affect such a daily activity pattern indicates that this sort of modeling can become quite complex. Key improvements associated with an ABA approach over the traditional four-step process are the ability to model explicitly the multistop, tour-based structure of much of today's out-of-home travel, and the ability to link the purpose behind a trip more closely to both in- and out-of-home daily household activities. Microsimulation methods then allow these concepts to be implemented computationally. In particular, Monte Carlo-based methods are being used to associate specific attributes (of the traveler, the vehicle, the cargo, the geography) with each trip simulated, using repeated sampling from predefined statistical distributions to create a synthetic population of travelers and/or their trips. This is done by first of all building individual traveler profiles, sometimes termed *synthetic travelers*, by drawing sample traveler characteristics from statistical distributions of, for example, different age, income, and employment groups. These travelers are then assigned a specific travel activity pattern, typically a daily pattern. When the simulator is asked to select the traveler's trip frequency, mode, route, or destination, this

means drawing from a set of probabilities that have themselves already been modeled using either logit or similar demand models of the type discussed above, or using a rule-based activity scheduling algorithm [48]. In addition, it is common to use continuous duration models with activity-based models, for example, for modeling the temporal dimension of trip departure time choice. Once generated, these individual traveler activity patterns can then be aggregated to produce an estimate (and subsequently a forecast) of the number of trips taking place between any pair of locations, by vehicle type (and by vehicle occupancy if the method used is sophisticated enough to match up trips involving more than one family member). Besides this natural aggregation property, microsimulation is also an especially efficient method for representing sequential events, allowing a natural incorporation of system dynamics. This includes the generation of multitrip, multipurpose daily tours or "trip chains" of the sort shown in Fig. 2.

Miller [49] provides a review and a brief history of the best known microsimulation models in use in both North America and Europe. See also Kitamura *et al.* [50]. In the United States this includes the Transportation Analysis and Simulation System,

or TRANSIMS¹: a model, or more correctly modeling system, developed over the past decade by Los Alamos National Laboratory and its contractors for the US Department of Transportation. The modeling approach and its many software modules are downloadable and is extensively documented [51]. Reinforcing the microsimulation theme, the approach includes the use of cellular automata methods to track the detailed movements of, and interactions between, individual vehicles as they move down the highway. Among other benefits, this allows greater attention to computing such impacts as vehicle emissions that are heavily influenced by the nature of stop-go traffic movements. In Europe, the Multiagent Transport Simulation Toolkit (MATSim)² is a similarly complex, and agent-based (see below) microsimulation modeling system [52] which was also developed around an activity-based approach to household travel forecasting. The ALBATROSS model [53] uses a sequential decision process to generate the daily activity schedules of individual travelers, simulating for a given day which activities are conducted, when (start time), for how long (duration), where (location), with whom, and, if traveling is involved, the transport mode used and chaining of trips engaged in. This makes a household's daily activity schedule the unit of prediction, with some activity timing and duration decisions modeled as continuous, rather than discrete choices, subject to sensible daily space-time constraints. Unlike the traditional four-step aggregate demand models described in section titled "Aggregate, Four-Step Planning Models" earlier, these different ABA modeling systems each use their own unique sequential and iterative modeling structures, and the mix of both conceptual and computational methods used to construct a household's daily travel activity pattern continues to evolve.

Microsimulation also lends itself well to the construction of freight movements. In Canada, Hunt [54] described the use of

microsimulation to generate daily multistop pickup and delivery truck tours based on the output from logit regression models which are used to select such attributes as the next stop's trip purpose, its location, and departure time interval. The approach can also be used to build vehicle shipments from the ground up, beginning with the selection of the origination point, the size, and nature of the cargo to be moved, through the detailed scheduling as well as routing of the vehicles used to move it, while also recognizing any physical constraints on its transportation [55]. Microsimulation is also proving to be useful in the modeling of freight through complete multistage product supply chains (Boerkamps *et al.* [56] in the Netherlands; Wisetjindawat *et al.* [57] in Japan, in which the movement dynamics may play themselves out over many days or even weeks at a time). In Massachusetts, Xu *et al.* [55] use microsimulation to model the routing of trucks over transportation networks as part of a much broader approach to freight modeling that incorporates a pseudo-real-time information simulator linked to a complete supply chain decision making simulator based on a very formal mathematical treatment of the freight movement problem as a variational inequality by Nagurney *et al.* [58]. See also Kuzmyak [59] for a recent review of forecasting metropolitan commercial and freight travel.

In taking on this added detail, which also relaxes the limitations on the number of household or industry classes and trip purposes used in traditional four-step planning models, an ABA approach can become computationally very time consuming. This is especially true if thousands, or even millions, of such patterns need to be simulated in support of project assessments and other planning applications. Fortunately, the tremendous advances in computational speeds over the past five decades have now provided a practical means of addressing this problem. Raney *et al.* [60] discuss this issue, noting that rule-based computer codes of the type used to simulate individual traveler decisions do not vectorize easily, and therefore do not run efficiently on traditional vector supercomputers—but that they do

¹http://tmip.fhwa.dot.gov/community/user_groups/~transims

²<http://www.matsim.org>

however run well on modern workstation architectures, which makes traffic simulations ideally suited for clusters of PCs such as Beowulf clusters.

Some of the most promising opportunities associated with microsimulation in the near future are likely to be tied to recent advances in travel data collection, and specifically to vehicle and cargo tracking technologies such as global positioning satellites (GPS) and radio frequency identification devices (RFIDs). These "IT"-based data collection methods are bringing a much needed visibility to the step-by-step movements of both people and goods. This includes the digital recording and storage of detailed personal travel diaries as well as electronic freight data manifests, each creating a new opportunity to establish both the realism and statistical reliability of microsimulation-based travel forecasting methods.

A natural adjunct to both ABA and microsimulation modeling entering the travel demand literature with increasing frequency is an approach known as *agent-based modeling* (ABM). Given a set of allowable actions and possible strategies for action, ABM allows a population of autonomous agents, such as a group of households or freight shippers or receivers (= agents) to interact among themselves to determine how much and what types of travel to engage in, with individual agent behaviors based on an agent's current status, its objective, and its history of past actions. Like microsimulation, ABA supports a bottom-up approach to estimating travel activity patterns, and as such seems well suited to travel activity systems in which individual tripmaking behaviors can be aggregated, sometimes yielding unexpected system-level *emergent behaviors* [61]. This point gets directly to a key weakness in current trip-based aggregate passenger and freight demand models: the difficulty we have in linking our forecasts of travel activity to causal mechanisms and hence, by implication, to policy instruments based on the expected behavior of passengers, shippers, carriers, receivers, warehousemen, and third party logistics agents. Rindt *et al.* [62] and Zhang and Levinson [63] describe early efforts to develop agent-based passenger

demand models. Venkat and Wakeland [64] and Hunt and Stefan [65] describe the use of ABM modeling to simulate commercial freight movements. Takama and Preston [66] combine agent-based modeling with discrete choice binary logit, mixed nested logit (MNL), and Markov queueing models to simulate the effects of road user charges on mode choice and parking behavior. In doing so, they also demonstrate the growing tendency of modelers to solve demand estimation and forecasting problems using a mix of solution methods.

SUMMARY

The past half-century has seen considerable technical as well as conceptual innovation in travel model design, calibration, and application. One notable line of advance has been the recognition that modeling approaches with very different origins, in economic/utility theory, in system optimization, in categorically based statistical analysis, in spatial interaction theory, and in agent-based and rule-based microsimulation, can all contribute to a better understanding as well as better means of estimating the demand for travel. Just how well these and other methods contribute to better forecasts remains to be seen and needs to be supported by a retention of historical data sources that will allow modelers to look back, as well as forward, and learn from their past efforts.

Acknowledgements

The submitted manuscript has been co-authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

REFERENCES

1. Hensher DA. Determining passenger potential for a regional airline hub at Canberra International Airport. *J Air Transp Manage* 2002;8(5):301–311.

2. Andreoni A, Postorino MN. Time series models to forecast air transport demand: a study about a regional airport. Proceedings of the Eleventh IFAC Symposium on Control in Transportation Systems.; The Netherlands. 2008. Available at <http://www.ifac-papersonline.net/Detailed/30968.html>.
3. Karlaftis MG. Demand forecasting in regional airports: dynamic tobit modeling with GARCH errors. *Sitraer* 2008;7:100–111. Available at <http://www.tgl.ufrj.br/viisitraer/pdf/312.pdf>.
4. Veenstra AW, Haralambides HE. Multivariate autoregressive models for forecasting seaborne trade flows. *Transp Res E* 2001;37(4):311–319.
5. Seabrooke W, Hui ECM, Lam WHK, *et al.* Forecasting cargo growth and regional role of the Port of Hong Kong. *Cities* 2003;20(1):51–64.
6. Nihan NL, Holmesland KO. Use of the Box and Jenkins time series technique in traffic forecasting. *Transportation* 1980;9(2):125–143.
7. Ghosh B, Basu B, O'Mahoney M. Bayesian time-series model for short-term traffic flow forecasting. *J Transp Eng* 2007;133(3):180–189.
8. Klodzinski J, Al-Deek HM. Methodology for modeling a road network with high truck volumes generated by vessel freight activity from an intermodal facility. *Transp Res Rec* 2004;1873:35–44.
9. Box GEP, Jenkins GM. Time-series analysis: forecasting and control. San Francisco (CA): Holden-Day; 1970.
10. Meyer MD, Miller EJ. Urban transportation planning: a decision-oriented approach. 2nd ed. New York: McGraw-Hill; 2001.
11. Ortuzar JD, Willumsen LG. Modelling transport. 3rd ed. New York: Wiley; 2001.
12. Golob TF, Kitamura R, Long L, editors. Panels for transportation planning: methods and applications. Boston (MA): Kluwer Academic Publishers; 1997.
13. McFadden DC. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. *Frontiers in econometrics*. New York: Academic Press; 1974.
14. McFadden DC. Modeling the choice of residential location. In: Karlqvist A, *et al.*, editors. *Spatial interaction theory and residential location*. Amsterdam: North Holland Publishing Co.; 1978. pp. 75–96.
15. Williams HCWL. On the formation of travel demand models and economic evaluation measures of user benefit. *Environ Plann A* 1977;9(3):285–344.
16. Wen C-H, Koppelman F. The generalized nested logit model. *Transp Res B* 2001;35(7):627–641.
17. Daly A, Bierlaire M. A general and operational representation of generalised extreme value models. *Transp Res B* 2006;40(4):285–305.
18. Garrow LA. Discrete choice modelling and air travel demand: theory and applications. Aldershot, UK: Ashgate Publishing; 2010.
19. Ben-Akiva M, Lerman SR. Discrete choice analysis: theory and application to travel demand. Cambridge: The MIT Press; 1985.
20. Train KE. Discrete choice methods with simulation. Cambridge, UK: University Press; 2003.
21. Koppelman, F, Bhat, C. A self instructing course in mode choice modeling: multinomial and nested logit models. Report prepared for the U.S. Department of Transportation and Federal Transit Authority. 2008. Available at: http://www.civil.northwestern.edu/people/~koppelman/PDFs/LM_Draft_060131Final-060630.pdf.
22. Daganzo C. Multinomial probit: the theory and its applications to demand forecasting. New York: Academic Press; 1979.
23. Akar G, Clifton KJ, Doherty ST. How do travel attributes affect the planning time horizon of activities? CD Proceedings of the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
24. van Bladel K, Bellemans T, Janssens D, *et al.* Activity-travel planning and rescheduling behavior: an empirical analysis of influencing factors. Paper to the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
25. Bellemans T, van Bladel K, Janssens D, *et al.* Measuring and estimating suppressed travel using enhanced activity-travel diaries. CD Proceedings of the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
26. Duarte A, Garcia C, Limao S, *et al.* Experienced and expected happiness in transport mode decision making process. CD Proceedings of the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
27. Meloni I, Bez M, Spissu E. An activity-based model of women's activity-travel patterns. CD Proceedings of the 88th Annual Meeting of the

- Transportation Research Board; Washington, DC. 2009.
28. Eluru N, Senar IN, Bhat CR, *et al.* Understanding residential mobility: a joint model of the reason for residential relocation and stay duration. CD Proceedings of the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
 29. Meloni, I, Portoghese, A, Bez, M, *et al.* An activity-based model of the impact of physical activities on the propensity to choose a sustainable model of travel to work or school. Paper to the 88th Annual Meeting of the Transportation Research Board; Washington, DC. 2009.
 30. Train K, Wilson W. Spatial demand decisions in the Pacific Northwest: mode choices and market areas. *Transp Res Rec* 2006;1963:9–14.
 31. Jiang F, Johnson P, Calzada C. Freight demand characteristics and mode choice: an analysis of the results of modeling with disaggregate revealed preference data. *J Transp Stat* 1999;2(2):149–158.
 32. Holguin-Veras J. Revealed preference analysis of commercial vehicle choice process. *J Transp Eng* 2002;128(4):336–346.
 33. Norojono O, Young W. A stated preference freight mode choice model. *Transp Plann Technol* 2003;26(2):195–212.
 34. Jeffs VP, Hills PJ. Determinants of modal choice in freight transport. *Transportation* 1990;17(1):29–47.
 35. Talvitie A. A direct demand model for downtown work trips. *Transportation* 1973;2(2):121–152.
 36. Anderson MD, Sharfi K, Gholston SE. Direct demand forecasting model for small communities, using multiple linear regression. *Transp Res Rec* 2006;1981:114–117.
 37. Wardrop JG. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II*(1). 1952. pp. 325–378.
 38. Boyce DE. A practitioner's guide to urban travel forecasting models. *Proceedings of the Metropolitan Conference on Public Transportation Research*; Chicago. 2008. Available at http://tsap.civil.northwestern.edu/boyce_pubs/practitioners_guide.pdf.
 39. Oppenheim N. *Urban travel demand modeling*. New York: John Wiley and Sons Inc.; 1995.
 40. Wilson AG. *Entropy in urban and regional modelling*. London: Pion; 1970.
 41. Erlander S, Stewart NF. *The gravity model in transportation analysis—theory and extensions*. The Netherlands: VSP Books; 1990.
 42. Lee D-H, editor. *Urban and regional transportation modeling*. Cornwall: MPG Books Ltd.; 2004.
 43. Erlander S. Accessibility, entropy and the distribution and assignment of traffic. *Transp Res* 1977;11(3):149–153.
 44. Ham H, Kim TJ, Boyce DE. Implementation and estimation of a combined model of interregional, multimodal commodity shipments and transportation network flows. *Transp Res B* 2005;39(1):65–79.
 45. Wilson AG, Coelho JD, Macgill SM, *et al.* *Optimization in locational and transport analysis*. Chichester: Wiley; 1981.
 46. Southworth F. *Freight flow models. Intermodal transportation: moving freight in a global economy*. ENO Foundation; 2009. Chapter 5. In press.
 47. McNally, MG, Rindt, CR. The activity-based approach. In: Hensher D, Button K, editors. *Volume 1, Handbook of transportation modelling*. 2nd ed. Amsterdam: Pergamon; 2008.
 48. Pendyala RM, Kitamura R, Prasuna-Reddy DVG. Application of an activity-based travel-demand model incorporating a rule-based algorithm. *Environ Des B* 1998;25(5):753–772.
 49. Miller EJ. Microsimulation. In: Goulias KG, editor. *Transportation systems planning: methods and applications*. Boca Raton (FL): CRC Press; 2003. pp.12–1–12–22.
 50. Kitamura R, Chen C, Pendyala R, *et al.* Microsimulation of daily activity-travel patterns for travel demand forecasting. *Transportation* 2000;27(1):25–51.
 51. Hobeika A. *TRANSIMS fundamentals*. 2008. Available at http://tmip.fhwa.dot.gov/resources/clearinghouse/docs/transims_fundamentals/.
 52. Balmer M, Nagel K, Raney B. Agent-based demand modeling framework for large scale micro-simulations. *Transp Res Rec* 2006;1985:125–134.
 53. Arzenze T, Timmermans H. ALBATROSS: overview of the model, application and experiences. *Innovations in Travel Modeling 2008 Conference*; 2008 June 22–24; Portland, OR. 2008. Available at <http://www.trb-forecasting.org/ALBATROSS.pdf>.
 54. Hunt JD. *Calgary tour-based microsimulation of urban commercial vehicle movements. Freight Demand Modeling: Tools for*

- Public-Sector Decision Making. Issue 40, Transportation Research Board Conference Proceedings. Washington, DC: Transportation Research Board; 2008. pp. 61–71. Available at <http://www.trb.org/Conferences/2006/FDM/HuntCalgaryPaper.pdf>.
55. Xu J, Hancock KL, Southworth F. Simulation of regional freight movement with trade and transportation multi-networks. *Transp Res Rec* 2003;1854:152–161.
 56. Boerkamps JHK, van Binsbergen AJ, Bovy PHL. Modeling behavioral aspects of urban freight movements in supply chains. *Transp Res Rec* 2000;1725:17–25.
 57. Wisetjindawat W, Sano K, Matsumoto S. Supply chain simulation for modeling the interactions in freight movement. *J East Asia Soc Transp Studies* 2008;6:2991–3004. Available at http://www.easts.info/on-line/journal_06_/2991.pdf.
 58. Nagurney A, Ke K, Cruz J, *et al.* Dynamics of supply chains: a multilevel (logistical/information/financial) network perspective. *Environ Plann B Plann Des* 2002;29(6): 795–818.
 59. Kuzmyak JR. Forecasting metropolitan commercial and freight travel. NCHRP Synthesis 384. Washington, DC: Transportation Research Board; 2008.
 60. Raney B, Setin N, Vollmy A, *et al.* An agent-based microsimulation model of Swiss travel: first results. *Netw Spat Econ* 2003;3:23–41.
 61. Sanford Bernhardt KL. Agent-based modeling in transportation. Artificial intelligence in transportation: information for application. Transportation Research Circular E-C113. Washington, DC: Transportation Research Board; 2007. pp. 72–80.
 62. Rindt CR, Marca JE, McNally MG. Toward dynamical, longitudinal, agent-based microsimulation models of human activity in urban settings. Working paper UCI-ITS-AS-WP-02-5. Irvine: The Institute of Transportation Studies, University of California.; 2008. Available at <http://www.its.uci.edu/~casa/casa-pubs.html>.
 63. Zhang L, Levinson D. Agent-based approach to travel demand modeling: exploratory analysis. *Transp Res Rec* 2004;1898:28–36.
 64. Venkat K, Wakeland WW. An agent-based model of trade with distance-based transaction costs. Proceedings from the Summer Computer Simulation Conference. San Diego: The Society for Modeling and Simulation International; 2008. pp. 3–10. Available at http://www.math.ucf.edu/~kaup/kaup_files/Pub228.pdf.
 65. Hunt JD, Stefan KJ. Tour based microsimulation of urban commercial movements. *Transp Res B* 2007;41(9):981–1013.
 66. Takama T, Preston J. Forecasting the effects of road user charge by stochastic agent-based modelling. *Transp Res A* 2008;42:738–749.

TREewidth, TREE DECOMPOSITIONS, AND BRAMBLES

ARIE M.C.A. KOSTER
Lehrstuhl II für Mathematik,
RWTH Aachen University,
Aachen, Germany

INTRODUCTION

The notions of tree decomposition and treewidth were introduced by Robertson and Seymour [1] in a long series of papers to prove Wagner's conjecture. Let $G = (V, E)$ be an undirected graph consisting of a set of vertices V and a set of edges E . A tree decomposition of a graph G is a pair $(\{X_i, i \in I\}, T = (I, F))$ with $X_i \subseteq V$, $i \in I$ and $T = (I, F)$ a tree (i.e., an acyclic connected graph), such that

- (TD1) $\bigcup_{i \in I} X_i = V$;
- (TD2) for all $vw \in E$, there is an $i \in I$ with $v, w \in X_i$; and
- (TD3) for all $v \in V$, $\{i \in I : v \in X_i\}$ forms a connected subtree of T .

The *width* of a tree decomposition $(\{X_i, i \in I\}, T = (I, F))$ is $\max_{i \in I} |X_i| - 1$. The *treewidth* $\tau(G)$ of G is the minimum width over all tree decompositions of G .

Figure 1 shows a graph and a tree decomposition of the graph. The property TD3 is sometimes called the *interpolation property*. Furthermore, note that (TD1) is obsolete if G has no isolated vertices.

Several equivalent notions for treewidth have been studied over time, such as partial k trees, dimension, and k -decomposability. A graph has treewidth at most k , if and only if it is a partial k -tree, if and only if the dimension of G is at most k , if and only if G is k -decomposable. See Ref. 2 for further details. The related notions of *pathwidth* and *branchwidth* are briefly described at the end of this article.

One further equivalent notion is *bramble*, which first appeared in Ref. 3 under the name *screen*. We follow the terminology of Reed [4]. Two subsets $W_1, W_2 \subseteq V$ are said to *touch* if they have a vertex in common or E

contains an edge between them ($W_1 \cap W_2 \neq \emptyset$ or there is an edge $\{w_1, w_2\} \in E$ with $w_1 \in W_1, w_2 \in W_2$). A set $\mathcal{B}$ of mutually touching connected vertex sets is called a *bramble*. A subset of V is said to *cover* $\mathcal{B}$ if it is a hitting set for $\mathcal{B}$ (i.e., a set which intersects every element of $\mathcal{B}$). The *order* of a bramble $\mathcal{B}$ is the minimum size of a hitting set for $\mathcal{B}$. The *bramble number* $b(G)$ of G is the maximum order of all brambles of G .

Theorem 1 [3]. $\tau(G) = b(G) - 1$.

For a short proof of this theorem, we refer to Ref. 5. A class of graphs for which the bramble of maximum order can be constructed easily is the class of grid graphs. It is folklore that an x by x grid has treewidth x ; see, for example, Fig. 2 for a bramble of order $x + 1$: We have one set that contains all vertices on the bottom row, one set that contains all vertices in the last column except the last one, and $(x - 1)^2$ crosses, each consisting of the first $x - 1$ vertices of one of the first $x - 1$ rows and the first $x - 1$ vertices of one of the first $x - 1$ columns. A set that covers the bramble must contain at least $x - 1$ vertices to cover the crosses, and one vertex in each of the other two sets.

Treewidth has gained its importance by many *parameterized complexity* results [6,7]. A combinatorial optimization problem is *fixed-parameter tractable* (FPT) if there exists an exact algorithm to solve the problem in which running time is polynomial in the size of the input, except for one **fixed** parameter. Treewidth is such a parameter; for example, the weighted independent set problem can be solved in $O(2^k \cdot n)$ time, assuming $n = |V|$ and a tree decomposition of width k is part of the input [8]. Hence, maximum weighted independent sets can be solved in polynomial (linear) time if the treewidth of the graph is bounded by a constant (whereas the problem is $\mathcal{NP}$ -complete in general).

In subsequent sections, we discuss tree decomposition-based algorithms, the computation of treewidth in theory and practice, lower bounds on the treewidth including

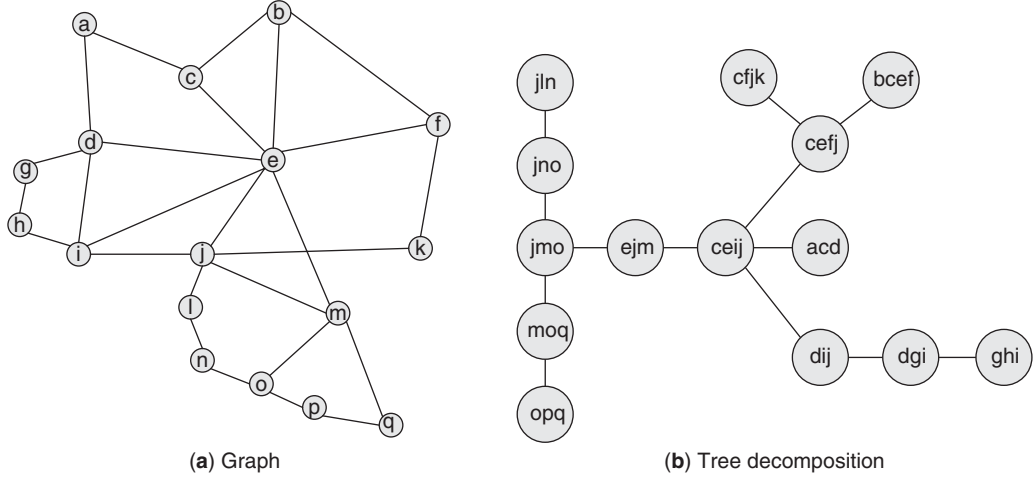


Figure 1. Graph and a tree decomposition with width 3.

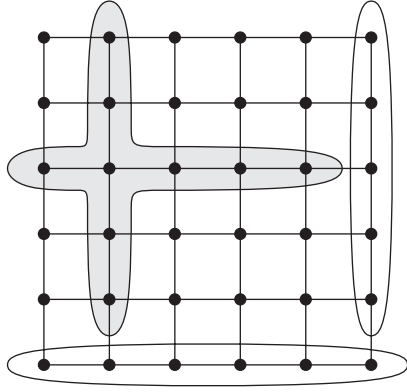


Figure 2. Illustrating a bramble of a grid.

brambles, and related notions.

TREE DECOMPOSITION-BASED ALGORITHMS

In general, an FPT algorithm for a problem that takes treewidth as a parameter has the following structure:

1. A tree decomposition (X, T) of G is found; the width of this tree decomposition is small (bounded by a constant, not necessarily optimal), see the next section for further details.

2. A dynamic programming algorithm is executed on the tree decomposition; for each node of the tree decomposition, a table is computed. For a decision problem, the table for the root of the tree T shows the answer. (For construction variants of problems, additional book-keeping and a construction step follow the run of the dynamic programming algorithm.)

For many $\mathcal{NP}$ -complete problems like independent set [8], Hamiltonian circuit, chromatic index [9], or Steiner tree [10] the above procedure results in a polynomial-time algorithm if defined on a graph of bounded treewidth. The dynamic programming algorithm works in a bottom-up fashion, that is, the tree $T = (I, F)$ is rooted at some node and the algorithm first computes results for the leaves of this rooted tree, next for the parents of the leaves and so on, until the root is reached.

To simplify explanation and implementation of the dynamic program, the tree decomposition is usually converted to a *nice tree decomposition*. In a nice tree decomposition, each node $i \in I$ is of one of the four following types:

- **Leaf:** Node i is a leaf of T , and $|X_i| = 1$.

- **Join:** Node i has exactly two children, say j_1 and j_2 and $X_i = X_{j_1} = X_{j_2}$.
- **Introduce:** Node i has exactly one child, say j , and there is a vertex $v \in V$ with $X_i = X_j \cup \{v\}$.
- **Forget:** Node i has exactly one child, say j , and there is a vertex $v \in V$ with $X_j = X_i \cup \{v\}$.

It is not hard to see that if G has treewidth at most k , then G also has a nice tree decomposition of width at most k which has $O(n)$ tree nodes. Such a nice tree decomposition can be built in linear time, given a (not nice) tree decomposition [11].

A graph $G_i = (V_i, E_i)$ is associated with each node $i \in I$. V_i is the union of all bags X_j , with $j = i$ or j a descendant of i in T , and $E_i = E \cap (V_i \times V_i)$ is the edge set induced by V_i .

For each node $i \in I$, a table, with all (in)feasible solutions for G_i and which are different with respect to X_i , is computed. The table for a leaf node is computed by just enumeration of the possible solutions. For a join node, the tables of the two children are joined by identifying the solutions. For an introduce node, the solutions are extended with the new vertex. Finally, for a forget node, all solutions that are identical with respect to X_i are identified by a single one (with best objective). A more detailed description of such a dynamic programming algorithm for a practical problem can be found in Refs 12 and 13 for frequency assignment in wireless networks.

Tree decomposition-based algorithms have been shown to be of practical interest; for example, for the inference problem for probabilistic networks in decision support systems, the traveling salesman problem TSP [14], and the partial constraint satisfaction problems with binary relations [12,13] (with frequency assignment and MAX SAT as special cases).

In real-life applications, memory consumption is a major concern as the tables contain an exponential number of entries. Techniques to reduce the memory requirements have been proposed in Ref. 15.

COMPUTING TREewidth: THEORY

The formal decision problem of treewidth asks whether there exists a tree decomposition with width at most k for some integer $k > 0$. If k is part of the input, $\mathcal{NP}$ -completeness was proved by Arnborg *et al.* [16]. If k may be considered a constant, not part of the input, the best algorithm has been given by Bodlaender [17] and checks in linear time whether or not a tree decomposition with width at most k exists, and if so, finds such a tree decomposition. The $O(n)$ notation for this algorithm however hides a huge constant coefficient that obstructs its practical computational value. An experimental evaluation by Röhrig [18] revealed that the algorithm is computationally intractable, even for k as small as four.

Graphs with treewidth of at most four can be characterized directly or indirectly: $\tau(G) = 1$ if and only if G is a forest, $\tau(G) \leq 2$ if and only if its biconnected components are series-parallel graphs [19]. Arnborg and Proskurowski [20] gave six reduction rules that reduce G to the empty graph if and only if $\tau(G) \leq 3$ (subsets suffice for $\tau(G) = 1$ and 2). Sanders [21] provided a linear-time algorithm for testing $\tau(G) \leq 4$.

Besides forests and series-parallel graphs, the complexity of treewidth is known for some special classes of graphs (by presenting either a polynomial-time algorithm or an $\mathcal{NP}$ -completeness proof). In particular, the treewidth of bipartite graphs is $\mathcal{NP}$ -complete, since every graph can be converted to a bipartite graph with the same treewidth by subdividing its edges (replacing the edge by two edges incident to a new vertex). We refer the interested reader to two surveys on the topic by Bodlaender [22,23]. Most remarkable in this context is that, so far, the complexity of treewidth for planar graphs is unknown, whereas for branchwidth a polynomial-time algorithm exists. Lapoire [24] and Bouchitté *et al.* [25] proved that the treewidth of a planar graph and of its geometric dual differ by, at most, one. The result is generalized in Ref. 26 to a difference of $\ell + 1$ for (hyper)graphs on a surface of genus ℓ .

The treewidth of a graph can also be found

by dynamic programming [27], resulting in an exact (but exponential) algorithm. The best-known exact algorithm computes the treewidth of a general graph with n vertices in $O(1.7347^n)$ if exponential space can be used [28], and $O(2.6151^n)$ with polynomial space [29].

Given a graph G with $\tau(G) = k$, the best approximation algorithm by Feige *et al.* [30] provides a tree decomposition of width at most $O(k\sqrt{\log k})$ (i.e., a $O(\sqrt{\log k})$ approximation). So far, a constant approximation algorithm is neither known nor is it proven that no such algorithm exists.

COMPUTING TREewidth: PRACTICE

Most results presented in the previous section are of theoretical interest only: the computational complexity hides huge constant coefficients that make the algorithms impractical for actual computation of treewidth. So far, only the reduction rules for treewidth, at most three, have been proved to be of practical use in preprocessing the input graph. However, in all those cases where the treewidth is larger than three, we typically have to turn to heuristics without any performance guarantee. The survey articles of Refs 31 and 32 have been dedicated to this topic. An accompanying website [33] allows experiments with some of the algorithms. Here, we briefly review preprocessing, exact algorithms, and upper-bound heuristics.

Preprocessing

The reduction rules of Ref. 20 not only reduce graphs of treewidth at most three to the empty graph, but can also be used as a preprocessing technique to reduce the size of general graphs. In Ref. 34, the rules have been adapted and extended such as to preprocess general graphs. Computational experiments have shown that significant reductions in the graph size can be achieved by these rules.

The above-mentioned preprocessing rules emphasize the removal of vertices from the graph. Another way to reduce the complexity of finding a good tree decomposition is the splitting of the input graph into smaller

graphs for which we can construct a tree decomposition independently. In Ref. 35, so-called *safe separators* are introduced for this purpose. A separator S is a set of vertices whose removal disconnects a graph G . Let V^i , $i = 1, \dots, p$ ($p \geq 2$), induce the connected components of $G - S$. On each of the connected components $G[V^i]$, a graph G^i is defined as $G[V^i \cup S] \cup K(S)$, where $K(S)$ denotes a complete graph, or *clique*, on S . If $\tau(G) = \max_{i=1, \dots, p} \tau(G^i)$, then S is called *safe* for treewidth. In particular, clique separators (i.e., S induces a clique) and almost-clique separators (i.e., S contains a $|S| - 1$ clique) are safe. Experiments have shown the effectiveness of safe separators [35].

Exact Algorithms

Although treewidth is $\mathcal{NP}$ -hard in general, there have been a couple of attempts to tackle the problem by exact approaches. Shoikhet and Geiger [36] implemented a modified version of the $O(n^{k+2})$ algorithm by Arnborg *et al.* [16]. A branch-and-bound algorithm based on vertex-ordering has been proposed by Gogate and Dechter [37]. In Ref. 27 a dynamic programming algorithm has been developed, based on the Held–Karp algorithm for TSP. Due to its exponential space requirement, only relatively small graphs can be solved exactly this way.

Upper-Bound Heuristics

Most heuristics without a performance guarantee are of a constructive nature: according to some principle, we construct a tree decomposition from scratch. Improvement heuristics as well as metaheuristics are less frequently exploited.

The construction heuristics are typically based on the equivalence of the treewidth problem with the problem of finding a triangulation of G with minimum–maximum clique size. A *triangulated* or *chordal* graph is a graph in which every cycle of length of at least four has a chord. A *triangulation* of a graph $G = (V, E)$ is a chordal graph $H = (V, F)$ with $E \subseteq F$.

Lemma 1. *Let G be a graph, and let $\mathcal{H}$ be the set of all triangulations of G . Then,*

$\tau(G) = \min_{H \in \mathcal{H}} \omega(H) - 1$, where $\omega(H)$ is the size of the maximum clique in H .

Thus, if G is triangulated, then $\tau(G) = \omega(G) - 1$, otherwise we have to find a triangulation of H with small maximum clique size. Several algorithms exist to check whether G is triangulated or to construct a triangulation of G . Examples of such algorithms are the *lexicographic breadth first search* (LEX) recognition algorithm by Rose *et al.* [38], the *maximum cardinality search* (MCS) by Tarjan and Yannakakis [39] and variants of those algorithms that guarantee to find a *minimal* triangulation (i.e., no $e \in F$ can be removed without losing the triangulation property) known as *LEX-M*, [38] and *MCS-M* [40], respectively. See Refs 31 and 41 for some experimental results for LEX-P, MCS, and LEX-M.

The *minimal fill-in* problem is another problem that is studied in relation to triangulation of graphs. The minimum fill-in of a graph is the minimum number of edges to be added to a graph such that the resulting graph is chordal/triangulated. This problem is known to be $\mathcal{NP}$ -hard [42], but it is not difficult to think of two heuristics. The first one is a greedy algorithm: select repeatedly the vertex for which the fill-in among its neighbors is minimized, turn its neighbors into a clique, and remove that vertex. This algorithm is called *Greedy Fill-In* (GFI) or simply the minimum fill-in algorithm in some articles. The *Greedy Degree* (GD) or simply minimum degree algorithm, repeatedly selects the vertex of minimum degree and turns its neighbors into a clique before removing the vertex. These algorithms typically provide very good upper bounds for both fill-in and treewidth in reasonable time. For computations and further heuristics, we refer to the survey article by Bodlaender and Koster [31].

TREewidth LOWER BOUNDS AND BRAMBLES

Another way to approach the treewidth problem is to bound the treewidth from below. In

this way, no tree decomposition is obtained but a good estimate of the span of the true treewidth of a graph can be obtained. Moreover, high-quality lower bounds are helpful within branch-and-bound approaches to bound the size of the search tree. An extensive survey on lower bounds for treewidth will be published in Ref. 43. Here we review the most important developments, with a focus on brambles.

First of all, it is easy to see that every lower-bound algorithm for treewidth can be extended by taking the maximum of the lower bound over all subgraphs or minors: given an optimal tree decomposition for G and H a subgraph (minor) of G , then we can construct a tree decomposition with equal or better width for H by removing vertices from the bags that are not part of the subgraph (minor) and replacing contracted vertices by their new vertex.

By Theorem 1, a bramble $\mathcal{B}$ of order k implies $\tau(G) \geq k - 1$. So, finding a high order bramble immediately implies getting a good lower bound for the treewidth. Unfortunately, determining the order of a bramble, is also $\mathcal{NP}$ -hard; it follows directly from the $\mathcal{NP}$ -hardness of the minimum hitting set problem [44] by taking a complete graph G and the collection of subsets as the bramble.

In Ref. 45, two algorithms for constructing a bramble of high order are presented, one for general graphs and one for planar graphs. Both algorithms aim at finding large grid minors in G , exploring the so-called *Grid Minor Theorem*.

Theorem 2 [46,47]. *There exists a function $f(r)$ such that any graph of treewidth at least $f(r)$ has a $r \times r$ grid graph as a minor. The best-known function $f(r)$ for general graphs is superexponential.*

In addition, a hitting set of a $r \times r$ grid graph can be computed easily (Fig. 2). The algorithms partition the vertex set in connected subsets representing the *vertical* connections. These subsets are connected by *horizontal* vertex-disjoint paths. By contracting a part of the subsets, grid minors in minors of G can be found. The lower bound provided by the size of the bramble in such a

grid minor is a lower bound on the treewidth of G by the above-mentioned result.

The grid minor algorithm for planar graphs is shown to give a lower bound for treewidth that is at most a constant factor away from the optimum. For both algorithms, extensive computational experiments are reported showing that the algorithms often give excellent lower bounds, in particular when applied to (close to) planar graphs.

Recently, Chapelle *et al.* [48] present an algorithm for graphs with treewidth larger than k to construct a bramble of size $k + 2$ in $O(n^{k+4})$. Brambles are also used to find minimal forbidden minors for treewidth [49].

Besides brambles, lower bounds can be computed in a variety of ways. The probably widest known lower bound is given by the maximum clique size. This can be seen by Lemma 1: the maximum clique of G will be part of a clique in any triangulation of G .

Scheffler [50] proved that every graph of treewidth at most k contains a vertex of degree at most k . Stated differently, the minimum degree $\delta(G)$ is a lower bound on the treewidth of a graph. Typically, this lower bound is of no real interest as the minimum degree can be arbitrarily small. Even if the preprocessing rules of the previous section have been applied before, only $\delta(G) \geq 3$ can be guaranteed.

As mentioned before, improvements on this bound can be obtained by taking subgraphs or minors of G . The *degeneracy* $\delta D(G)$ of G is the maximum minimum degree over all subgraphs of G , and hence is a lower bound on $\tau(G)$ (and can be computed by repeatedly removing the minimum degree vertex). Independently, Bodlaender *et al.* [51] and Gogate and Dechter [37] combined the minimum degree lower bound with taking minors. The so-called *contraction degeneracy* $\delta C(G)$ is defined as the maximum minimum degree over all minors of G . In Ref. 51 it is proven that computing $\delta C(G)$ is $\mathcal{NP}$ -hard and computational experiments are presented by applying tiebreakers to the following algorithm: repeatedly contract a vertex of minimum degree to one of its neighbors and record the maximum encountered. Significantly better lower bounds than the degeneracy are obtained this way. In Ref. 52, further

results for contraction degeneracy are discussed, showing for example that $\delta C(G) \leq 5 + \gamma(G)$, where $\gamma(G)$ is the *genus* of G .

Ramachandramurthi [53,54] introduced the parameter

$$\gamma_R(G) = \min \left(n - 1, \min_{\substack{v,w \in V, v \neq w, \\ \{v,w\} \notin E}} \max(d(v), d(w)) \right),$$

and proved that this is a lower bound on the treewidth of G . Note that $\gamma_R(G) = n - 1$ if and only if G is a complete graph on n vertices. If G is not complete, then $\gamma_R(G)$ is determined by a pair $\{v, w\} \notin E$ with $\max(d(v), d(w))$ as small as possible. From its definition it is clear that $\gamma_R(G) \geq \delta_2(G) \geq \delta(G)$, where $\delta_2(G)$ is the second smallest degree appearing in G (note $\delta(G) = \delta_2(G)$ if the minimum degree vertex is not unique). All these three lower bounds can be computed in polynomial time.

Lucena [55] proved that with the MCS upper-bound algorithm, a lower-bound $MCSLB(G)$ on the treewidth can be obtained. Bodlaender and Koster [56] proved that determining whether $MCSLB(G) \leq k$ for some $k \geq 7$ is $\mathcal{NP}$ -complete and presented computational results by constructing MCS-orderings using tiebreakers for the decisions within the MCS algorithm.

Also the lower bounds $\delta_2(G)$, $\gamma_R(G)$, and $MCSLB(G)$ can be computed over all minors. Bodlaender and Koster [51] studied the combination of $MCSLB(G)$ and taking minors, whereas Koster *et al.* [57] studied the combination of $\delta_2(G)$ and $\gamma_R(G)$ and taking minors. In practice, $\delta_2 C(G)$ and $\gamma_R C(G)$ are only marginally better (if at all) than the lower bounds computed for the contraction degeneracy.

Finally, improved lower bounds for treewidth can be obtained by the following result.

Theorem 3 [58]. *Let $G = (V, E)$ be a graph with $\tau(G) \leq k$ and $\{v, w\} \notin E$. If there exist at least $k + 2$ vertex-disjoint paths between v and w , then $\{v, w\} \in F$ for every triangulation H of G with $\omega(H) \leq k$.*

Hence, if we know that $\tau(G) \leq k$ and there exist $k + 2$ vertex disjoint paths between v

and w , adding $\{v, w\}$ to G should not hamper the construction of a tree decomposition with small width. Clautiaux *et al.* [59] explored this result by first computing a lower bound ℓ on the treewidth of G by any of the above methods (e.g., $\ell = \delta C(G)$) and next assuming $\tau(G) \leq \ell$ and adding edges $\{v, w\}$ to G for which there exist $\ell + 2$ vertex-disjoint paths in G . Let G' be the resulting graph. Now, if it can be shown that $\tau(G') > \ell$ by a lower-bound computation on G' , our assumption that $\tau(G) \leq \ell$ is false. Hence, $\tau(G) > \ell$ or stated equally $\tau(G) \geq \ell + 1$: an improved lower bound for G is determined. This procedure can be repeated until it is not possible anymore to prove that $\tau(G') > \ell$ (which of course, does not imply that $\tau(G') = \ell$). In Ref. 51 the above-described approach is nested within a minor-taking algorithm, resulting in the best known lower bounds for most tested graphs [60]. In many cases, optimality could be proved by combining lower and upper bounds.

RELATED NOTIONS

If T is restricted to be a path, we refer to $(\{X_i, i \in I\}, T = (I, F))$ as a *path decomposition* and the best width as the *pathwidth* of G . The pathwidth of a tree can be arbitrarily large. One is referred to survey papers [2, 61] for a more thorough exposition of pathwidth.

A notion closely related to treewidth is the branchwidth of a graph, also introduced by Robertson and Seymour [62, 63].

Definition 1. A *branch decomposition* of a graph $G = (V, E)$ is a pair $(T = (I, F), \sigma)$, with T a ternary tree (a tree where every non-leaf node has degree 3) with $|E|$ leaves, and σ a bijection from E to the set of leaves of T .

The *order* of an edge $f \in F$ is the number of vertices $v \in V$, for which there exist adjacent edges $\{v, w\}, \{v, x\} \in E$, such that the path in T from $\sigma(v, w)$ to $\sigma(v, x)$ uses f .

The *width of branch decomposition* $(T = (I, F), \sigma)$, is the maximum order over all edges $f \in F$. The *branchwidth* of a graph G , $bw(G)$, is the minimum width over all branch decompositions of G .

The notions of branchwidth and treewidth are closely related to each other, as expressed in the following theorem.

Theorem 4 [62]. Let $G = (V, E)$ be a graph with $E \neq \emptyset$. Then $\max(\beta(G), 2) \leq \tau(G) + 1 \leq \max(\lfloor \frac{3}{2}\beta(G) \rfloor, 2)$.

One can easily convert a branch decomposition into a tree decomposition such that the respective widths fulfill these inequalities. As a result, algorithms computing a branchwidth upper bound (or exactly) also provide an upper bound on the treewidth of the graph.

Since branchwidth can be computed in polynomial time for planar graphs [64], this algorithm provides a $\frac{3}{2}$ -approximation for treewidth of planar graphs. Improved implementations of this algorithm can be found in Ref. 65. For general graphs, heuristics have been developed by Cook and Seymour [14, 63], Hicks [66], and Robertson and Seymour [67]. Further information on branchwidth can be found in Ref. 68.

REFERENCES

1. Robertson N, Seymour PD. Graph minors. II. Algorithmic aspects of tree-width. J Algorithms 1986;7:309–322.
2. Bodlaender HL. A partial k -arboretum of graphs with bounded treewidth. Theor Comput Sci 1998;209:1–45.
3. Seymour PD, Thomas R. Graph searching and a minimax theorem for tree-width. J Comb Theor Ser B 1993;58:239–257.
4. Reed BA. Tree width and tangles, a new measure of connectivity and some applications. Volume 241, Surveys in combinatorics, LMS Lecture Note Series. Cambridge: Cambridge University Press; 1997. pp. 87–162.
5. Bellenbaum P, Diestel R. Two short proofs concerning tree-decompositions. Comb Probab Comput 2002;11:541–547.
6. Flum J, Grohe M. Parameterized complexity theory. Heidelberg, Germany: Springer; 2006.
7. Niedermeier R. Invitation to fixed-parameter algorithms, Oxford Lecture Series in Mathematics and Its Applications. New York: Oxford University Press; 2006.

8. Bodlaender HL, Koster AMCA. Combinatorial optimization on graphs of bounded treewidth. *Comput J* 2008;51(3):255–269.
9. Bodlaender HL. Polynomial algorithms for graph isomorphism and chromatic index on partial k -trees. *J Algorithms* 1990;11: 631–643.
10. Korach E, Solel N. Linear time algorithm for minimum weight Steiner tree in graphs with bounded treewidth. Manuscript; 1990.
11. Kloks T. Volume 842, Treewidth: computations and approximations, Lecture Notes in Computer Science. Berlin: Springer; 1994.
12. Koster AMCA. Frequency assignment-models and algorithms [PhD thesis]. Maastricht, The Netherlands: University Maastricht; 1999.
13. Koster AMCA, van Hoesel SPM, Kolen AWJ. Solving partial constraint satisfaction problems with tree decomposition. *Networks* 2002;40(3):170–180.
14. Cook W, Seymour PD. Tour merging via branch-decomposition. *INFORMS J Comput* 2003;15(3):233–248.
15. Betzler N, Niedermeier R, Uhlmann J. Tree decompositions of graphs: saving memory in dynamic programming. *Disc Optim* 2006;3: 220–229.
16. Arnborg S, Corneil DG, Proskurowski A. Complexity of finding embeddings in a k -tree. *SIAM J Algebr Disc Methods* 1987;8:277–284.
17. Bodlaender HL. A linear time algorithm for finding tree-decompositions of small treewidth. *SIAM J Comput* 1996;25: 1305–1317.
18. Röhrig H. Tree decomposition: a feasibility study [Master's thesis]. Germany: Max-Planck-Institut für Informatik, Saarbrücken; 1998.
19. Bodlaender HL, van Antwerpen-de Fluiter B. Parallel algorithms for series parallel graphs and graphs with treewidth two. *Algorithmica* 2001;29:543–559.
20. Arnborg S, Proskurowski A. Characterization and recognition of partial 3-trees. *SIAM J Algebr Disc Methods* 1986;7:305–314.
21. Sanders DP. On linear recognition of tree-width at most four. *SIAM J Disc Math* 1996;9(1):101–117.
22. Bodlaender HL. A tourist guide through treewidth. *Acta Cybern* 1993;11:1–23.
23. Bodlaender HL. Discovering treewidth. In: Vojtáš P, Bielíková M, Charron-Bost B, editors. Volume 3381, Proceedings 31st Conference on Current Trends in Theory and Practice of Computer Science, Liptovský Ján Slovakia: SOFSEM 2005, Lecture Notes in Computer Science. Springer; 2005. pp. 1–16.
24. Lapoire D. Treewidth and duality for planar hypergraphs; 2002. Unpublished manuscript. Available at http://www.labri.fr/perso/lapoire/papers/dual_planar_treewidth.ps.
25. Bouchitté V, Mazoit F, Todinca I. Chordal embeddings of planar graphs. *Disc Math* 2003;273:85–102.
26. Mazoit F. Tree-width of graphs and surface duality. *Electron Notes Disc Math* 2009;32:93–97.
27. Bodlaender HL, Fomin FV, Koster AMCA, *et al.* On exact algorithms for treewidth. In: Azar Y, Erlebach T, editors. Volume 4168, Proceedings 14th Annual European Symposium on Algorithms, ESA 2006, Lecture Notes in Computer Science. Zurich, Switzerland: Springer; 2006a. pp. 672–683.
28. Fomin FV, Villanger Y. Finding induced subgraphs via minimal triangulations. In: Proceedings 27th International Symposium on Theoretical Aspects of Computer Science, STACS 2010; Nancy. 2010. pp. 383–394. Available at <http://www.stacs-conf.org>.
29. Fomin FV, Villanger Y. Treewidth computation and extremal combinatorics. In: Aceto L, Damgård I, Goldberg LA, editors. Volume 5125, Proceedings of the 35th International Colloquium on Automata, Languages and Programming, ICALP 2008, Part I, Lecture Notes in Computer Science. Reykjavik, Iceland: Springer; 2008. pp. 210–221.
30. Feige U, Hajiaghayi M, Lee JR. Improved approximation algorithms for minimum-weight vertex separators. *SIAM J Comput* 2008;38(2):629–657.
31. Bodlaender HL, Koster AMCA. Treewidth computations I: upper bounds. *Inf Comput* 2010;208:259–275. Available at <http://dx.doi.org/10.1016/j.ic.2009.03.008>.
32. Bodlaender HL, Koster AMCA. Treewidth computations III. Exact Algorithms 2010. In preparation. Available at <http://www.math2.rwth-aachen.de/~koster/computeTW/>.
33. Koster AMCA. ComputeTW –An interactive platform for computing treewidth of graphs. 2009. Available at <http://www.math2.rwth-aachen.de/~koster/ComputeTW>.
34. Bodlaender HL, Koster AMCA, Eijkhof Fvd, *et al.* Pre-processing for triangulation of probabilistic networks. In: Breese J, Koller D, editors. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence. San

- Francisco (CA): Morgan Kaufmann; 2001. pp. 32–39.
35. Bodlaender HL, Koster AMCA. Safe separators for treewidth. *Disc Math* 2006;306: 337–350.
36. Shoikhet K, Geiger D. A practical algorithm for finding optimal triangulations. *Proceedings National Conference on Artificial Intelligence (AAAI '97)*. Providence (RI): Morgan Kaufmann; 1997. pp. 185–190.
37. Gogate V, Dechter R. A complete anytime algorithm for treewidth. *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence UAI-04*. Arlington (VA): AUAI Press; 2004. pp. 201–208.
38. Rose DJ, Tarjan RE, Lueker GS. Algorithmic aspects of vertex elimination on graphs. *SIAM J Comput* 1976;5:266–283.
39. Tarjan RE, Yannakakis M. Simple linear time algorithms to test chordality of graphs, test acyclicity of graphs, and selectively reduce acyclic hypergraphs. *SIAM J Comput* 1984;13:566–579.
40. Berry A, Blair J, Heggernes P, *et al.* Maximum cardinality search for computing minimal triangulations of graphs. *Algorithmica* 2004;39:287–298.
41. Koster AMCA, Bodlaender HL, van Hoesel SPM. Treewidth: computational experiments. In: Broersma H, Faigle U, *et al.*, editors. Volume 8, *Electronic Notes in Discrete Mathematics*. Enschede, The Netherlands: Elsevier Science Publishers; 2001. pp. 54–57.
42. Yannakakis M. Computing the minimum fill-in is NP-complete. *SIAM J Algebr Disc Methods* 1981;2:77–79.
43. Bodlaender HL, Koster AMCA. Treewidth computations II: Lower bounds. 2010. In preparation. Available at <http://www.math2.rwth-aachen.de/~koster/computeTW/>.
44. Garey MR, Johnson DS. *Computers and intractability, a guide to the theory of NP-completeness*. New York: W.H. Freeman and Company; 1979.
45. Bodlaender HL, Grigoriev A, Koster AMCA. Treewidth lower bounds with brambles. *Algorithmica* 2008;51:81–98.
46. Robertson N, Seymour PD. Graph minors. V. Excluding a planar graph. *J Comb Theor Ser B* 1986;41:92–114.
47. Robertson N, Seymour PD, Thomas R. Quickly excluding a planar graph. *J Comb Theor Ser B* 1994;62:323–348.
48. Chapelle M, Mazoit F, Todinca I. Constructing brambles. Volume 5734, *Proceedings MFCS* 2009, *Lecture Notes in Computer Science*. Novy Smokovec, Slovakia: Springer; 2009. pp. 223–234.
49. Lucena B. Achievable sets, brambles, and sparse treewidth obstructions. *Disc Appl Math* 2007;155:1055–1065.
50. Scheffler P. *Die Baumweite von Graphen als ein Maß für die Kompliziertheit algorithmischer Probleme* [PhD thesis]. Berlin: Akademie der Wissenschaften der DDR; 1989.
51. Bodlaender HL, Wolle T, Koster AMCA. Contraction and treewidth lower bounds. *J Graph Algorithms Appl* 2006;10:5–49.
52. Wolle T, Koster AMCA, Bodlaender HL. A note on contraction degeneracy. Technical Report UU-CS-2004-042. Utrecht, The Netherlands: Institute of Information and Computing Sciences, Utrecht University; 2004.
53. Ramachandramurthi S. *Algorithms for VLSI Layout Based on Graph Width Metrics* [PhD thesis]. Knoxville (TN): Computer Science Department, University of Tennessee; 1994.
54. Ramachandramurthi S. The structure and number of obstructions to treewidth. *SIAM J Disc Math* 1997;10:146–157.
55. Lucena B. A new lower bound for treewidth using maximum cardinality search. *SIAM J Disc Math* 2003;16:345–353.
56. Bodlaender HL, Koster AMCA. On the maximum cardinality search lower bound for treewidth. *Disc Appl Math* 2007;155(11): 1348–1372.
57. Koster AMCA, Wolle T, Bodlaender HL. Degree-based treewidth lower bounds. In: Nikolettseas SE, editor. Volume 3503, *Proceedings of the 4th International Workshop on Experimental and Efficient Algorithms WEA 2005*, *Lecture Notes in Computer Science*. Santorini Island, Greece: Springer; 2005. pp. 101–112.
58. Bodlaender HL. Necessary edges in k -chordalizations of graphs. *J Comb Optim* 2003;7:283–290.
59. Clautiaux F, Carlier J, Moukrim A, *et al.* New lower and upper bounds for graph treewidth. In: Rolim JDP, editor. Volume 2647, *Proceedings International Workshop on Experimental and Efficient Algorithms, WEA 2003*, *Lecture Notes in Computer Science*. Monte Verità, Switzerland: Springer; 2003. pp. 70–80.
60. TreewidthLIB. Treewidthlib. Available at <http://www.cs.uu.nl/people/hansb/treewidthlib>, 2004.
61. Bienstock D. Graph searching, path-width, tree-width and related problems (a survey).

- DIMACS Ser Disc Math Theor Comput Sci 1991;5:33–49.
62. Robertson N, Seymour PD. Graph minors. X. Obstructions to tree-decomposition. *J Comb Theor Ser B* 1991;52:153–190.
 63. Robertson N, Seymour PD. Graph minors. XIII. The disjoint paths problem. *J Comb Theor Ser B* 1995;63:65–110.
 64. Seymour PD, Thomas R. Call routing and the ratcatcher. *Combinatorica* 1994;14(2): 217–241.
 65. Hicks IV. Planar branch decompositions II: The cycle method. *INFORMS J Comput* 2005; 17:413–421.
 66. Cook W, Seymour PD. An algorithm for the ring-routing problem. Bellcore technical memorandum, Bellcore; 1993.
 67. Hicks IV. Branchwidth heuristics. *Congressus Numerantium* 2002;159:31–50.
 68. Hicks IV. Branchwidth and branch decompositions. In: Floudas CA, Pardalos PM, editors. *Encyclopedia of optimization*. 2nd ed. Heidelberg, Germany: Springer; 2009. pp. 332–339.

TRIAGE IN THE AFTERMATH OF MASS-CASUALTY INCIDENTS

NILAY TANIK ARGON

SERHAN ZIYA

Department of Statistics and
Operations Research,
University of North Carolina,
Chapel Hill, North Carolina

JAMES E. WINSLOW

School of Medicine, Wake Forest
University, Winston-Salem,
North Carolina

This article provides a review of the existing and ongoing research that uses mathematical modeling and analysis to better understand and help improve triage decisions in the immediate aftermath of mass-casualty incidents (MCIs). An MCI is an event that results in a massive number of casualties with severe injuries that require medical attention immediately. Typical examples include plane crashes, earthquakes, and terrorist bombings. In the wake of an MCI, emergency response resources are overwhelmed by a sudden jump in demand, and medical resources need to be rationed in some way so as to minimize the number of deaths and permanent injuries. One part of this rationing process is the determination of priorities according to which patients will have access to these limited resources. We call this *triage*. In practice, the word *triage* is typically used to refer to the process of assessing the injury severity of patients, which automatically determines their priority levels. Here, we are using the word *triage* more broadly to refer to the decision of identifying the injury severity and determining priority levels of patients.

The word *triage* comes from the French word “trier,” which simply means “to sort.” The earliest accounts of triage date back to the Napoleonic Wars that took place in the first two decades of the nineteenth century. Although, the practice of triage started in the

military, it was eventually adopted in civilian life as well, especially after World War II. The whole idea behind triage is to efficiently use scarce resources, such as medical personnel, operating rooms, and ambulances by determining the order according to which patients will have access to these resources. For more on the history and practice of patient triage, see Iserson and Moskop [1].

In daily emergencies, resources are typically sufficient to keep up with the demand that comes from patients with severe injuries or life-threatening conditions and therefore, determining priorities is relatively simpler. Even though it is crucial that patients are correctly classified into triage classes, given the triage level of a patient, there is generally no question of which class should get a higher priority (see Argon and Ziya [2] for an analysis of errors in patient classification decisions). The most severe patients are treated first and the less severe patients who can wait are treated afterwards. This ensures that all patients have the maximum chance of survival that is possible given their injuries since typically, no critical patient experiences a significant treatment delay because of others who are given higher priorities. (It is important to note that average waiting times in emergency rooms can be very high but generally not for patients in life-threatening conditions, assuming they are correctly identified to be so.)

On the other hand, MCIs require a whole new approach to triage. MCIs are very different from daily emergencies because in MCIs, unlike in daily emergencies, an overwhelming number of casualties arrive almost simultaneously and their injuries tend to be much more severe. Because of severe resource restrictions, it is typically not possible to treat every patient in a timely manner. The objective of triage in such a mass event is to allocate scarce resources to do “the greatest good for the greatest number.”

According to Lerner *et al.* [3], which is a recent review paper on mass-casualty triage, there exist nine mass-casualty triage

systems, two of which are specific to pediatric patients. Most of these triage systems classify patients into 4–5 categories based on some physiological criteria such as breathing rate and pulse. As an example, consider one of the most widely adopted and used triage procedures in the United States, which is called *START* (Simple Triage and Rapid Treatment; e.g., Nocera and Garner [4]). According to *START*, patients are classified into one of four priority groups: immediate, delayed, minor, and deceased. Time-critical patients are generally either in the immediate or delayed category, immediate patients having priority over delayed patients. Patients in the minor category can wait longer, whereas patients in the deceased category have either died or have almost no chance of survival. Each patient is given a very quick evaluation as soon as possible and is tagged according to his/her *START* category with a colored marker. Then, each patient is given treatment in the order determined by his/her priority class (i.e., *START* category).

The main problem with *START* and most other mass-casualty triage systems is that patient categorization is made independently of the size of the disaster; that is, the number of patients waiting for care (or the available resources). Obviously, a patient's health condition does not change with the number of other casualties; however, from a system's optimization point of view, this information can be very useful to achieve the greatest good for the greatest number. For example, if the situation is so severe that it is only possible to save a fraction of the casualties at best with the given resources, then it may make more sense to redistribute the resources from the more time-critical with the most severe injuries toward the less time-critical patients who are more salvageable [5].

We are aware of two lines of research that use operations research (OR) tools for the analysis of the mass-casualty triage problem. The first approach [6,7] is based on a linear program that determines the best order for transporting patients from the scene to nearby hospitals. The second approach [8–10] models the problem as a stochastic scheduling problem and uses several

techniques (such as dynamic programming and heuristic methods) to provide insights and solutions to the mass-casualty triage problem. As we discuss later in the sections titled “A Linear Programming Approach to Mass-Casualty Triage” and “A Stochastic Scheduling Approach to Mass-Casualty Triage” in detail, the two approaches share some similarities but in essence the former is more of a medical research that employs an OR methodology (namely linear programming) and proposes a real-time solution method, whereas the latter is essentially an OR work that solves a general model from which insights into mass-casualty triage can be developed. Despite the differences between these two lines of research, main insights are similar. First, in order to maximize the number of survivors, mass-casualty triage systems should take into account the severity of the event; that is, the number of casualties and the capacity of the resources. Second, a “good” categorization of patients is essential for the success of the triage effort and such a categorization should be based on how the patients' health deteriorates, their service requirements, and survival probabilities.

We review the two OR approaches discussed above in the subsequent sections. Note that this review is exclusively for patient triage and prioritization decisions at pre-hospital and hospital levels in the aftermath of mass-casualty incidents. However, the insights that come out of the reviewed work may also carry over to other prioritization decisions that are commonly faced in such emergencies. For comprehensive reviews of the OR literature on emergency response systems, the interested reader is referred to Green and Kolesar [11] and Larson [12].

A LINEAR PROGRAMMING APPROACH TO MASS-CASUALTY TRIAGE

To our knowledge, within the emergency medicine literature, the only work that uses a mathematical approach for mass-casualty triage is by Sacco *et al.* [6,7]. In these two papers, Sacco and his coauthors formulate

the mass-casualty triage problem as a linear program and call their approach the Sacco Triage Method (STM). STM is mainly developed for primary triage in which the objective is to determine which patients will be transported first from the field to nearby hospitals in case of a mass-casualty incident. Under this method, the triage procedure begins with a brief physiological assessment of patients at the field. Based on this assessment, each patient is assigned a criticality score between 0 and S , where a smaller score indicates a more severe health condition. More specifically, STM uses the RPM (respiratory rate, pulse rate, and best motor response) scoring system [13]. For RPM, $S = 12$.

Once each patient is assigned a score, a central controller enters the available data (i.e., the number of patients in each criticality state denoted by the RPM score) into a computer that solves a linear program and determines the order according to which patients should be transported. The planning horizon is divided into T time periods enumerated as $1, 2, \dots, T$. (Although the formulation is general, the authors set the length of each period to be 30 min in their numerical analysis.) The capacity of service is defined in terms of the maximum number of casualties that can be served per period, which is denoted by R_n for period n . It is assumed that transportation times are deterministic and capacity in each period is known. The patients' health conditions are assumed to deteriorate deterministically while they wait for their turn for transportation. More specifically, knowing the initial RPM score of the patient right after triage is sufficient to know exactly what the condition of the patient would be at any point in time in the future. The health condition of the patient (measured in terms of the RPM score) at the time he/she was transported to the hospital affects his/her survival probability. The objective is to determine which patients should be transported in each period to maximize the expected number of survivors.

The problem is formulated as follows. Let $V_{s,n}$ denote the decision variable that denotes the number of casualties served in time period n whose RPM score is s at the time of triage, where $n = 1, 2, \dots, T$

and $s = 0, 1, \dots, S$. Let $P_{s,n}$ be the survival probability of a patient with an initial RPM value of s when transported in period n and Q_s be the number of casualties whose initial RPM score is s .

Below is the linear programming formulation that the STM is based on:

$$\text{Maximize } \sum_{s=0}^S \sum_{n=0}^T P_{s,n} V_{s,n} \quad (1)$$

subject to

$$\begin{aligned} \sum_{s=0}^S V_{s,n} &\leq R_n, \quad \text{for } n = 1, 2, \dots, T \\ \sum_{n=0}^T V_{s,n} &= Q_s, \quad \text{for } s = 0, 1, \dots, S \\ V_{s,n} &\geq 0, \quad \text{for } s = 0, 1, \dots, S \text{ and } \\ &n = 1, 2, \dots, T. \end{aligned}$$

The performance of the solution of the above linear program has been numerically tested for two different types of injuries. In Ref. 6, the focus is on blunt trauma victims whereas Ref. 7 studies victims with penetrating injuries. In both studies, the survival probabilities are estimated using a logistic regression. More specifically, the survival probability of a patient with health score s is modeled by

$$P_s = \frac{1}{1 + e^{-(w_0 + w_1 s)}}, \quad (2)$$

where the logistic regression coefficients w_0 and w_1 are estimated from the data collected from hospitals that participated in the Pennsylvania trauma outcome study. The pattern of deterioration of patients' health conditions (indicated by their RPM scores) is estimated by the Delphi technique used by Linstone and Turoff [14]. Following the Delphi method, a panel of experts on trauma care provided estimates of scene RPM value changes in increments of 30 min during a 6-h period assuming minimal medical interventions during that period. After two iterations of taking such estimates, the median changes in RPM values for successive periods for each initial RPM value were obtained. As

Table 1. RPM Score Estimates as Function of 30-min Time Intervals Presented in Table 3 of Ref. 7

Initial RPM	Time Interval (n)											
	1	2	3	4	5	6	7	8	9	10	11	12
12	12	12	11	11	10	10	9	9	9	8	8	8
11	11	11	10	10	9	8	7	7	7	6	5	5
10	10	9	9	8	8	7	6	5	5	5	4	4
9	9	8	8	7	5	4	3	1	0	0	0	0
8	7	6	4	3	3	2	2	0	0	0	0	0
7	6	5	3	1	0	0	0	0	0	0	0	0
6	4	3	2	1	0	0	0	0	0	0	0	0
5	3	2	0	0	0	0	0	0	0	0	0	0
4	2	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

an example, see the Delphi estimates of RPM values for penetrating injuries from Ref. 7 in Table 1. These RPM estimates, as a function of time, are then used in Equation (2) to obtain $P_{s,n,s}$, which are later plugged into the linear program given in Equation (1).

In both, Refs 6 and 7, the performances of the STM and START triage method are compared by means of a numerical study. This numerical study is performed under the same assumptions that the model was built upon, that is the service times are deterministic, and success probabilities are known functions of time. Hence, not surprisingly, the STM method, which is the optimal solution under the given assumptions, yields the best performance. However, the numerical study is still useful because it provides insights on the magnitude of improvement provided by STM over START and also on types of events where using STM in place of START has a potential to make the biggest impact. In particular, the authors observe that START performs close to STM when the patients' conditions are not very critical. However, for events where the resources are overwhelmed with a large number of victims having severe injuries, STM performs significantly better than START. It is especially important to note that when resources are severely overwhelmed, worst-first triage policy (i.e., the policy that serves the patients with the smallest RPM value first) performs very badly.

As the resources become more scarce, STM tends to give priority to patients in less critical conditions but who have higher chances of survival, thereby achieving an improvement in the expected number of lives saved.

The main observation in Refs 6 and 7 is that most of the current mass-casualty triage methods such as START do not take into account the resource limitations. Furthermore, the triage categories for these methods are very crude (only 4–5 categories) meaning that among patients that are classified into the same triage class, there could be a huge variation in terms of survival probabilities. One could significantly improve the expected number of survivors by implementing a triage system that divides patients into categories based on better predictions of their survival probabilities and then applying a mathematical approach that considers resource limitations to determine the order of service for patients in different classes.

A STOCHASTIC SCHEDULING APPROACH TO MASS-CASUALTY TRIAGE

One way of modeling patient triage in MCIs is to view it as a standard job scheduling problem in a clearing system where there is a finite number of jobs all available for service at time zero but can possibly renege before their turn for service. Here, jobs are

the patients with various criticality levels and renegeing corresponds to the death of a patient (or reaching a stage where survival is almost impossible even with the appropriate medical intervention). To our knowledge, there are only three papers in the OR literature that studied this model: Refs 8, 9, and 10. Argon *et al.* [9], in particular, report several insights drawn from the analysis of this clearing model for mass-casualty triage. In this section, we will discuss the main results from these three papers and also provide insights on the mass-casualty triage problem.

Patient Triage as a Job Scheduling Problem

Suppose that in the immediate aftermath of an MCI, there is a finite number of patients at various criticality levels that are in need of access to a scarce resource, which is crucial for the patients' survival. We call this scarce resource the "server." Depending on the circumstances, the server may represent an emergency vehicle that will transport patients to nearby hospitals, a doctor or a nurse in the field, or an operating room at the hospital. Suppose that patients are instantaneously put into one of M different categories (e.g., based on their injury characteristics) and there are m_i patients of type $i \in \{1, 2, \dots, M\}$ at time zero. Each patient type is characterized by its lifetime and service time distributions, denoted by $F_i(\cdot)$ and $G_i(\cdot)$ for type i patients, respectively. Here, the lifetime of an injured patient represents the time he/she can tolerate waiting in the absence of any medical intervention. In general, the lifetime distribution would heavily depend on the injury characteristics. For example, patients with intracerebral hemorrhage are expected to have a shorter lifetime than patients with crush injuries to the leg. Note that since the model allows the service time distribution to depend on patient types, the insights that come out are relevant not only to pre-hospital triage but also to hospital-level triage where the service time is likely depend on the injury types.

The objective for this scheduling problem is to dynamically determine the order that patients are taken into service so as to maximize the number of survivors under

the assumption that patients who are admitted before their death survive with certainty. This problem can be formulated as a semi-Markov decision process where the state space at the time the server is ready to choose the next patient can be described by an $(M + 1)$ -dimensional vector $(t, x_1, x_2, \dots, x_M)$, where t denotes the time passed since the incident took place, and x_i is the number of type i patients who are still alive and in need of attention.

Analytical Results

For ease of presentation, we adapt the following terminology:

Definition 1. If the lifetimes of type i patients are smaller than those of type j patients in the hazard rate sense (i.e., $F_i \leq_{hr} F_j$), then patients of type i are said to be more *time-critical* than patients of type j , for $i, j \in \{1, 2, \dots, M\}$ and $i \neq j$. (For definitions of stochastic orders used in this article, see Ref. 15.)

With this definition, higher time-criticality indicates higher urgency to serve the patient at any point in time as hazard rate ordering implies that patients who are more time-critical have shorter remaining lifetimes in the usual stochastic sense at any point in time. Nevertheless, time-criticality by itself does not determine priorities. If it takes a long time to serve a highly time-critical patient, it might be preferable to serve those patients who are not as critical but can be served in a shorter time. However, if it does take a shorter time to serve time-critical patients in some stochastic sense, it might be reasonable to give priority to them. The following theorem formalizes this result. Note that, $G_i \leq_{lr} G_j$ means that G_i is smaller than or equal to G_j in likelihood ratio ordering, which implies $G_i \leq_{hr} G_j$.

Theorem 1. Suppose that $F_1 \leq_{hr} F_2 \leq_{hr} \dots \leq_{hr} F_M$ and $G_1 \leq_{lr} G_2 \leq_{lr} \dots \leq_{lr} G_M$. Then, the policy that prioritizes one of the patients of the type with the smallest index maximizes the total number of survivors in the usual stochastic sense.

Theorem 1, which is proved in Ref. 9, gives a complete characterization of the optimal policy when lifetimes and service times are *agreeably ordered* with respect to the hazard rate and likelihood ratio orders, respectively. Under such a condition, the optimal policy turns out to be a simple, state-independent policy. To be more specific, the optimal priority ordering of the patients is the same regardless of time and the number of patients at various criticality levels. However, this simplicity comes with the expense of a strong condition that essentially requires that it takes less time to serve more critical patients, but this is not very likely to hold in practice. In the absence of such a condition, however, the analysis becomes significantly more challenging even under certain distributional assumptions for lifetimes and service times. We next describe some results that characterize the optimal policy under such distributional assumptions.

Suppose that there are only two types of patients. For patient type $i \in \{1, 2\}$, lifetimes and service times are both exponentially distributed with respective rates r_i and μ_i . Without loss of generality, assume that $r_1 > r_2$ so that type 1 patients are more time-critical than type 2 patients. If $\mu_1 \geq \mu_2$, then the agreeability condition of Theorem 1 holds and thus, the optimal policy is to give priority to type 1 patients at all times regardless of the number of patients at various criticality levels that are still in need of service. Instead, we assume that $\mu_1 < \mu_2$ under which the

optimal policy is not known. In this case, the optimal policy is possibly state-dependent and numerical analysis suggests that in most cases it can be characterized by a threshold curve over the state space, which separates the region where type 1 patients should be served from that where type 2 patients should be served. This structure is shown in Fig. 1. Note that because of the memoryless property, the optimal policy does not depend on time and is purely described as a function of the number of patients at the two criticality levels (x_1 and x_2).

Figure 1 suggests that if the number of patients is large, then giving priority to less time-critical patients who require shorter service (i.e., type 2 patients) becomes a better strategy. This policy is quite different from the common practice of triage for daily emergencies where priority is always given to time-critical patients. In daily emergencies, resources are usually not overwhelmed by very critical patients who have life-threatening conditions and therefore the basic principle is to give them the highest priority. However, in the aftermath of mass-casualty events, available resources may not be sufficient to provide the best possible care to each patient, which may require a switch from the regular daily operating mode. As we observe from Fig. 1, when the number of patients overwhelms the resources significantly, it is optimal to give priority to less critical patients if they keep the resource occupied for a shorter period of time. However, as

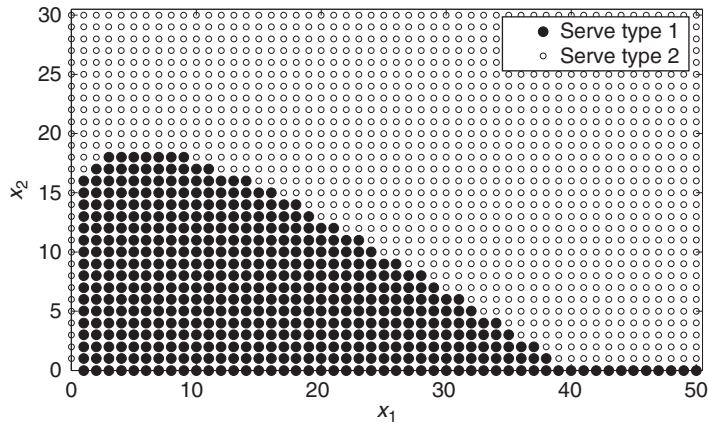


Figure 1. A typical structure for the optimal policy when $\mu_1 < \mu_2$ and $r_1 > r_2$, that is, when time-critical patients have longer service times [9].

soon as the number of patients drops, the priority policy switches back to the daily mode and priority is given to the most time-critical patients even if it takes longer to treat them.

For the Markovian model with two types of patients, Argon *et al.* [9] formulated a dynamic program and obtained a number of analytical results that provide further insights into the optimal policy in the absence of the agreeability condition of Theorem 1. We present two of these results:

Proposition 1.

- (i) If $\mu_1 < \mu_2 \leq r_2$ and $\mu_1 \leq r_1$, then for every x_1 , the optimal policy has a threshold, $t(x_1)$ (which may be infinite), such that for all $x_2 \geq t(x_1)$, it is optimal to serve a type 2 patient.
- (ii) If $r_1 \leq \mu_1 < \mu_2$ and $r_2 \leq \mu_2$, then for every $x_1 \geq 1$, the optimal policy has a threshold, $\tilde{t}(x_1)$ such that for all $x_2 \leq \tilde{t}(x_1)$, it is optimal to serve a type 1 patient, where

$$\tilde{t}(x_1) = \frac{r_1\mu_2 - r_2\mu_1}{r_2(\mu_2 - \mu_1)} - \frac{x_1r_1}{r_2}.$$

Proposition 1 formalizes the threshold structure that is observed in Fig. 1 under certain conditions on lifetimes and service times. More specifically, Proposition 1 implies that when the number of patients is large, then serving patients with shorter service times is optimal if the deaths occur faster than service (i.e., $r_i \geq \mu_i$ for $i = 1, 2$).

Proposition 2. If $\mu_1 < \mu_2 \leq r_2$, $\mu_1 \leq r_1$, and $r_1\mu_1 \leq r_2\mu_2$, then it is optimal to serve a type 2 patient in any state (x_1, x_2) such that $x_2 \geq 1$.

Proposition 2 provides a set of conditions under which the optimal policy is a state-independent policy that always calls for serving patients that are faster to serve. These conditions imply that the patients with faster service should be prioritized if the deaths occur at a faster rate than service and the conditions of the faster-to-serve patients are

sufficiently critical (if not as critical as the slower-to-serve patients).

Argon *et al.* [9] conjecture that Proposition 2 holds under less restrictive conditions based on extensive numerical experimentation. More specifically, it is conjectured that the $r\mu$ rule, which gives the priority to the type with the highest $r_i\mu_i$ score, is optimal if it also agrees with the shortest expected processing time (SEPT) rule.

Besides the above conjecture, the $r\mu$ rule has also been supported by two analytical results. A simple argument shows that if the optimal policy for the Markovian system with two types of patients is a state-independent policy, then it must agree with the $r\mu$ rule [9]. Furthermore, Glazebrook *et al.* [8] prove the “asymptotic” optimality of the $r\mu$ rule. To be more specific, let K , v_1 , and v_2 be defined so that $r_i = Kv_i$.

Theorem 2. Suppose that $r_1\mu_1 \leq r_2\mu_2$. Then the expected number of survivors under the policy that always gives priority to type 2 patients converges to the expected number of survivors under an optimal policy as K approaches to zero. In other words, the $r\mu$ policy is asymptotically optimal as the expected lifetimes for all patients go to infinity.

Theorem 2 implies that if the patients do not have serious injuries so that their lifetimes are sufficiently large, then regardless of the number of patients in the system, the priority should be given to the patient type with the largest $r\mu$ value.

Heuristic Methods and their Performance

Since it is difficult to fully characterize the optimal policy in the absence of the agreeability condition of Theorem 1 or the conditions of Proposition 2, it is of interest to develop heuristic methods. In the literature, four methods have been proposed: the $r\mu$ policy, two myopic policies (triangular and rectangular heuristics), and the PI- $r\mu$ policy.

The $r\mu$ Rule (by Glazebrook et al. [8]). As we discussed earlier, this policy is a state-independent policy according to

which the priority score of type i is $r_i\mu_i$ and the type with the higher score has a higher priority.

Myopic Policies (by Argon et al. [9]).

These policies are based on the idea of maximizing short-term (myopic) gains. More specifically, the *triangular heuristic* (one of the two myopic policies) maximizes the expected immediate “reward.” The reason that the policy is called *triangular heuristic* is that under this heuristic in the two dimensional representation of the state space (as in Fig. 1), states in which more time-critical patients are given priority forms a triangle. More specifically, the dark region in Fig. 1 is approximated by a triangle under this policy. The *rectangular heuristic* is obtained from the *triangular heuristic* by simply expanding the triangular region to a rectangle so as to obtain a simpler policy that can be described by only two threshold values that determine the lengths of the sides of the rectangle.

The PI- $r\mu$ Policy (Li and Glazebrook [10]).

This policy is obtained by applying a single step of the standard policy improvement algorithm for Markov decision processes on the $r\mu$ policy and using a fluid approximation for the value function of the $r\mu$ policy.

Among these policies, the $r\mu$ policy is the only one that is state-independent. The numerical study by Argon et al. [9] shows that this policy performs very well when the lifetimes are long, which is not surprising given Theorem 2. However, it has a poor performance when lifetimes are short. In such a scenario, the myopic policies (triangular and rectangular heuristics) perform much better than the $r\mu$ policy especially when there is a large number of casualties. On the other hand, the most recently proposed PI- $r\mu$ policy is not as simple as the other policies, but appears to perform exceptionally well across all scenarios.

Other Relevant Literature

In a working paper, Uzun et al. [16] analyze a more general version of the model described in the section titled “Patient Triage as a Job Scheduling Problem” using a model where customers possibly belong to more than two types, there are multiple servers, and each service brings a reward that depends on the patient’s type. In the context of the patient triage problem, the reward of service could be thought of as the probability of success from service, in which case the objective of the problem would be the maximization of the expected number of survivors. One could also set the reward of service to be the quality-adjusted life years (QALY) for the patient who goes through that service process. This allows taking into consideration factors such as age and pre-existing conditions as well as the injuries caused by the emergency event when making prioritization decisions. In Uzun et al. [16], the authors generalize Theorem 1, and give partial characterization of the optimal policy under certain conditions. They also numerically test a number of easy-to-implement near-optimal policies and show that they perform well.

There are also several articles in the traditional job scheduling literature on clearing systems that are closely related. However, none of the models studied in this literature suit the patient triage problem well. Among them, Refs 17, 18 and 19 appear to be the most relevant. The objective in these papers is to minimize the (weighted) number of tardy jobs (i.e., jobs not completed by their due date). Pinedo [19] and Boxma and Forst [17] consider only *list scheduling policies*, according to which the schedule is fixed at time zero and cannot be changed later on. This does not quite fit with the patient triage problem since under list scheduling policies, tardy jobs also get served. Emmons and Pinedo [18] study both, list scheduling and dynamic policies. They provide a number of cases for which optimal dynamic policies or list scheduling policies can be identified. They find that if the processing times are independent and identically distributed exponential random variables, and the due dates (e.g., lifetimes) are independent and can be ordered according to their failure rates, the optimal *preemptive*

dynamic policy is to process the jobs in the increasing order of their failure rates.

Finally, a number of articles from the stochastic scheduling literature consider systems with stationary arrival processes under the objective of minimizing weighted average number of tardy jobs with the assumption that jobs come from a single type. The main application domain for this scheduling problem is communication systems. Among these articles, Refs 20 and 21 show that when the due dates are announced at the time of arrivals, under certain conditions, one form of the “shortest-time-to-extinction” policy is optimal. Bhattacharya and Ephremides [22] and Zhao *et al.* [23] consider models where the due date is not announced but the service provider knows its probability distribution. In particular, Bhattacharya and Ephremides [22] find that if jobs have independent and identically distributed lifetimes, service times, and interarrival times, and lifetime distribution has a non-decreasing failure rate, the “earliest-arrival” policy is optimal.

CONCLUSIONS

In this article, we reviewed the growing literature on mathematical modeling and analysis of patient triage and prioritization decisions in the aftermath of mass-casualty events. In particular, we discussed two lines of research where one formulates the decision problem as a linear program and proposes that this problem be solved in real-time, whereas the other uses a stochastic job scheduling formulation in order to generate insights that can be later used in developing practical and simple policies. Despite their differences on main objectives and modeling approaches, both lines of work demonstrated the potential benefits of dynamic adaptive triage policies that take into account the number of casualties and the availabilities of resources to maximize the number of people saved.

Mass-casualty triage and patient prioritization is a complex but a very important problem that has not drawn much attention from the OR community. Considering the

typically chaotic environment that follows a mass-casualty event and the fact that each mass-casualty event typically tends to have its own unique unexpected challenges, it is difficult to imagine a sophisticated decision support tool that can be used in real-time. However, insights that come out of the mathematical models can help in the development of effective and practical policies.

More research is needed across various dimensions. In particular, injuries that could result from specific emergency events should be researched in depth to provide reliable input for the mathematical models. There is also a need for new innovative formulations that lead to fresh insights on the problem. Furthermore, extensive realistic simulation models should be developed in order to test the general validity of the insights that come out of stylized models. Operations researchers can help in developing an understanding of the potential outcomes of different kinds of priority policies, which can be very helpful to the medical community who ultimately have the responsibility of adopting policies that are practical, effective, and ethical. A successful outcome will ultimately require a significant effort from both, medical and operations research communities.

REFERENCES

1. Iserson KV, Moskop JC. Triage in medicine, part I: concept, history, and types. *Ann Emerg Med* 2007;49(3):275–281.
2. Argon NT, Ziya S. Priority assignment under imperfect information on customer type identities. *Manuf Serv Oper Manag* 2009;11(4):674–693.
3. Lerner EB, Schwartz RB, Coule PL. Mass casualty triage: an evaluation of the data and development of a proposed national guideline. *Disaster Med Public Health Prep* 2008;2, Suppl 1:S25–S34.
4. Nocera A, Garner A. An Australian mass casualty incident triage system for the future based upon triage mistakes of the past: the homebush triage standard. *Aust N Z J Surg* 1999;69(8):603–608.
5. Frykberg ER. Medical management of disasters and mass casualties from terrorist

- bombings: how can we cope? *J Trauma* 2002; 53(2):201–212.
6. Sacco WJ, Navin DM, Fiedler KE, *et al.* Precise formulation and evidence-based application of resource-constrained triage. *Acad Emerg Med* 2005;12(8):759–770.
 7. Sacco WJ, Navin DM, Waddell RK, *et al.* A new resource-constrained triage method applied to victims of penetrating injury. *J Trauma Inj Infect Crit Care* 2007;63(2): 316–325.
 8. Glazebrook KD, Ansell PS, Dunn RT, *et al.* On the optimal allocation of service to impatient tasks. *J Appl Probab* 2004;41(1):51–72.
 9. Argon NT, Ziya S, Richter R. Scheduling impatient jobs in a clearing system with insights on patient triage in mass-casualty incidents. *Probab Eng Inform Sci* 2008;22(3):301–332.
 10. Li D, Glazebrook KD. An approximate dynamic programming approach to the development of heuristics for the scheduling of impatient jobs in a clearing system. Lancaster: Lancaster University Management School; 2008. Unpublished manuscript.
 11. Green L, Kolesar P. Improving emergency responsiveness with management science. *Manag Sci* 2004;50(8):1001–1014.
 12. Larson RC. Public sector operations research: a personal journey. *Oper Res* 2002;50(1): 135–145.
 13. Sacco WJ, Champion H, Stega M. Computer implementation of severity scores for civilian trauma patient and naval combat casualty care. In: *Proceedings of the 19th Hawaii International Conference on System Sciences*. Hawaii; 1985.
 14. Linstone HA, Turoff M, editors. *The Delphi method—techniques and applications*. Reading (MA): Addison-Wesley; 1975.
 15. Shaked M, Shanthikumar JG. *Stochastic orders*. New York: Springer; 2007.
 16. Uzun E, Argon NT, Ziya S. Priority assignment in emergency response. Chapel Hill (NC): University of North Carolina; 2010. Unpublished manuscript.
 17. Boxma OJ, Forst FG. Minimizing the expected weighted number of tardy jobs in stochastic flow shops. *Oper Res Lett* 1986;5(3):119–126.
 18. Emmons H, Pinedo M. Scheduling stochastic jobs with due dates on parallel machines. *Eur J Oper Res* 1990;47(1):49–55.
 19. Pinedo M. Stochastic scheduling with release dates and due dates. *Oper Res* 1983;31(3):559–572.
 20. Bhattacharya PP, Ephremides A. Optimal allocation of a server between two queues with due times. *IEEE Trans Automat Control* 1991;36(12):1417–1423.
 21. Panwar DTSS, Wolf JK. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *J Assoc Comput Mach* 1988;35(4):832–844.
 22. Bhattacharya PP, Ephremides A. Optimal scheduling with strict deadlines. *IEEE Trans Automat Control* 1989;34(7):721–728.
 23. Zhao Z-X, Panwar SS, Towsley D. Queueing performance with impatient customers. In: *Proceedings of IEEE INFOCOM'91*. Bal Harbour (FL); 1991. pp. 400–409.

TRUST

ROBERT GODFREY
Madison, Wisconsin

The nature of trust and its role across a wide variety of social relationships—from personal friendships [1], to relations between professionals, clients, and business dealings [2], interactions within and between organizations [3], and even relations between citizens and their government [4]—has received both a wide and extensive review in the literature, across many branches of knowledge. Varied disciplines, such as psychology, social psychology, sociology, political science, economics, and mass communication, have each approached this subject in sometimes markedly different ways (Rousseau *et al.* [5] offer an extensive review). However, in general, they remain interested in learning how trust affects the workings of a relationship, how it enables cooperation, and how it is achieved, lost, and regained. Many types of research questions are posed, including: What are the sources of trust? What role does trust play in supporting institutions such as the market economy or representative government? Is trust between friends analogous to the “trust” a government might be said to inspire in its citizens? Is more trust necessarily a good thing, and when is it unwarranted, naive, and harmful? How can we best measure trust, and what kinds of evidence do we need to discriminate between rival theories?

For the purposes of this article, however, our focus on trust will largely be confined to those elements of research on the subject that can have applicability within the domains of risk, risk management, risk perception, and risk communication. We will begin with a review of the efforts to define the concept of trust, one with a long and complex history. Next, we will examine the important work that has been done exploring trust in the managers of risk. We will turn to source believability and the effects of involvement with an issue. The reader’s attention will

be then be directed to a thorough review of the importance of cognitive simplicity seeking and the role that values play in this process. We complete our survey with a short overview of the importance of assessing benefits in trust analysis.

THE IMPORTANCE OF TRUST

Trust is obviously a valued currency in use throughout society. It is a complex phenomenon to conceptualize, operationalize, and study. It also varies in meaning and context. The role of trust has been shown to be of critical importance in the literature on risk perception, risk communication, and risk analysis in general [6]. This observation is no better illustrated than by the oft-noted public concern with environmental effects stemming from the application of technology, which in turn has been paralleled by a decline in trust of those responsible for developing safe technologies and the governmental regulatory agencies meant to monitor their development [7]. At the same time, some research has shown that those who are willing to place their trust in an institution or group appear to find the risk estimates of those bodies more credible and their hazard policies more acceptable [8–10]. Indeed, as Slovic [11] and others have asserted, trust may be more fundamental to one’s ability to both understand and weigh certain potential risks than any strategy of risk communication. And as research by Mutz [12] has suggested, “levels of social trust are malleable... [and] can be altered by what [individuals] see, hear, and read” (p. 409). Consequently, discussions related to the most effective techniques for conveying information to the public about hazards and risks have quite often focused on questions of trust.

DEFINING TRUST

Perhaps one of the more notable features about the study of trust is the many seem-

ingly disparate definitions, descriptions, and characterizations that have been applied to the concept. There are a number of similar, but nonetheless multifaceted, themes and definitions.

From the field of psychology, one of the oft-noted studies of Rotter [13] defined trust “as an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon” (p. 651), while noting that one can learn to distrust large groups of people without personal experience validating such distrust. A factor analysis of the trust scale developed by Rotter [13] was found by Rotenberg [14] to yield three reliable factors: (i) political cynicism, which is skepticism about politicians and political bodies; (ii) interpersonal exploitation, which is a cautious orientation that corresponds to the perception of others as exploitive; (iii) belief in the dependability of people, which is the belief in the consistency between what others say and do. Chun and Campbell [15] identified a fourth factor that they labeled “societal hypocrisy,” the failure of social members to live up to commonly held expectations of behavior. Parks and Hulbert [16] defined trust as “the belief that others will not exploit one’s goodwill” (p. 719). They saw a firm connection between trust and cooperation. Yamagishi [17] did as well. The two most important factors observed in a factor analysis were the beliefs that people are basically honest and that trusting others is risky; both predicted subject’s cooperation level. Coleman [18] defined it in arguably more general terms as a voluntary expectation on the part of the trustor that he or she will receive a benefit from participating in an interaction with another trustee and not be hurt by it. Renn and Levine [19] characterized trust as “the generalized expectancy that a message received is true and reliable and the communicator demonstrates competence and honesty by conveying accurate, objective, and complete information” (p. 181). Rempel and Holmes [20] expressed it as, “the degree of confidence you feel when you think about a relationship” (p. 28) and Zucker [21] “as a set of expectations shared by all those involved in an exchange” (p. 54). Lewis and

Weigert [22] contended that, “members of that system act according to and are secure in the expected futures constituted by the presence of each other or their symbolic representations” (p. 965). While Koller [23] reasoned that trust is “a person’s expectations that an interaction partner is able and willing to behave promotively toward the person even when the interaction partner is free to choose among alternative behaviors that could lead to negative consequences for the person” (p. 266). Luhmann [24] defined it as, “an attitude which allows for risk taking decisions” (p. 95), and Fukuyama [25] has examined it within norms such as professional competency: for example, the audience’s belief in the professionalism of journalistic practice as the foundation for trust in journalism. Trust, is “the bedrock perception upon which society is possible” [26, p. 229].

The disciplines of political science and political communication reveal other definitions of trust. Lambert *et al.* [27], like many researchers, have produced an index of questionnaire items meant to measure trust as part of a larger study. Their questions concerning political trust dealt with dishonesty, fiduciary incompetence, and the ability to trust people in government to do what is right. Parker [28] attempted to measure a constituent trust index. Distrusting responses referred to the incumbent’s pursuit of self-interest, lack of integrity (principles), untrustworthiness, dishonesty, or lack of independence (controlled by political bosses or parties); constituent trust is reflected in responses that are the opposite of these (e.g., honesty, trustworthiness, integrity). It is interesting to note that in this study, Parker [28] found that 78% of the respondents gave more trusting evaluations of their congressperson than distrusting ones, a conclusion similar to other studies. This directly contrasts with other findings that show a strong distrust for the government in general. Howell and Fagan [29] composed an index related to questions of local trust. Again, fiduciary competence, the ability to trust people in government to do what is right, honesty, and equal representation of interests were

the items used for this purpose. Wald and Lupfer [30] used items related to ability to trust people in government to do what is right, equal representation of interests, general competence, fiduciary competence, and honesty.

Barber [31] concluded that trust has its roots in certain expectations people have of one another. There are three kinds of expectations:

“1) the expectation of the persistence and fulfillment of the natural and moral orders (e.g., the persistence of everyday routines we assume to be inevitable); 2) the expectation of technically competent role performance by the trusted individuals; 3) the expectation that these individuals will carry out their fiduciary obligations” (p. 9).

Barber [31] contended that the American political system depends on these three types of expectations, ones that are signified by the term “trust.” He maintained that since trust is an important component of power, when it is eroded, or was never present in the first place, we have distrust. Distrust is much more than the opposite of trust; it is also its equivalent. In other words, those on the lower end of the power distribution have a choice of whether to trust or not. Effective distrust leads to an equalization of power, and, consequently, the actions of the (formerly) trusted person or group are effectively monitored for competence and fiduciary responsibility. Barber [31] concluded, “trust is never enough for effective social control because the differential power of various groups in society has consequences that cannot be satisfactorily managed by trust alone” (p. 65).

Slovic [11] has noted that, generally, trust is fragile. It is slow to develop and, “can be destroyed in an instant—by a single mishap or mistake. Thus, once trust is lost, it may take a long time to rebuild it to its former state” (p. 677). For example, Rothbart and Park [32] found that, when people were asked to rate various descriptive traits of people, unfavorable traits were judged to be easier to acquire and harder to lose, taking more effort to disconfirm than a positive trait. Trustworthiness required a relatively

large number of confirming acts to establish the trait and a relatively small number of instances to disconfirm it. Slovic [11] observed that the ease with which trust is destroyed and the difficulty in creating it are due to a couple of factors. Negative (trust-destroying) events are more visible or noticeable than positive (trust-building) events. Negative (trust-destroying) events also carry much greater weight. Sources of bad (trust-destroying) news tend to be seen as more credible than sources of good news. Distrust, once initiated, tends to reinforce and perpetuate itself. This, according to Slovic, is because there is a tendency to inhibit the kinds of personal contacts and experiences that are necessary to overcome distrust. See also the notion of “recreancy” explored by Freudenburg [33,34]. At the same time, initial trust or distrust tends to color our interpretation of events—reinforcing our prior beliefs. Fischhoff [35] cited the example of the accident at Three Mile Island. For the antinuclear activists, Fischhoff suggested, the incident would have strengthened their resolve to ask only “how likely is such an accident, given a fundamentally unsafe technology?” while at the same time leading the pronuclear activists to ask only “how likely is the containment of such an incident, given a fundamentally safe technology?” (p. 300).

Standard definitions of trust from different literature share some general themes. Kasperperson *et al.* [36] summarized the most important attributes of trust to include the following:

“1) *Expectations about others and orientations toward the future.* Trust provides people with the ability to interact and cooperate without the full knowledge about others and future uncertainties. 2) *A notion of chance or risk taking.* The concept of trust instills confidence that others will act voluntarily in a manner that is beneficial. 3) *Subjective perceptions about others and situations.* These include perceptions of the intentions and attributes of others (e.g., commitment, competence, consistency, integrity, honesty), their motivations, qualities of the situation (e.g., the availability and accuracy of information), risks, and uncertainties” (p. 162).

The most basic of assumptions is that people are unlikely to attend to information, much less change their behavior or attitudes, if they do not trust the sources of risk information. Until recently, however, surprisingly little research has been initiated in this particular area of risk communication. Trust has traditionally been regarded as critical for social interaction and even for the survival of society [37,38]. It has also been found to be important during all aspects of childhood development [14]. Given that it holds such an important place in our day-to-day lives, it is surprising, as Stebbins [39] has noted, that “trust is an idea that is overused and understudied, clumsily defined and hotly contested, ideologically valued and concretely abused. While there are commentaries about and studies of trust in such varied and particular areas of life as research, family, and politics, generalized, comparative analyses of it is rare” (p. 475). Kasperson *et al.* [10] further contended that “trust is a concept widely identified as important to social interactions, but rarely well defined or characterized” (p. 164). Kramer and Tyler [40] have summarized in their literature review as many as 16 different definitions of the word trust. As Johnson [41] has observed, despite the findings from several studies pointing to the importance of trust as a factor in people’s responses to potentially risky situations, “we know very little about what drives trust judgments. This ignorance is due both to the recency of such research and use of vague (and sometimes the absence of appropriate) measures” [41, p. 348].

TRUST IN THE MANAGERS OF RISKS

Two research traditions have dominated this area of research into the overseers of risks. One includes the literature of persuasion/mass communication studies and will be discussed shortly. The second, the study of risk and science communication, has placed an especially important focus on the degree to which people trust the abilities of science, business, and government to manage danger (sometimes referred to as either fiduciary responsibility or care), and has been shown

to be one of the best predictors of the level of concern about various risks tied to the managers of those risks [7]. Frewer *et al.* [42] have noted the importance of distinguishing between the concepts of “social trust,” with its origins in sociopolitical analysis, and the concept of “source credibility” which is derived from empirical research in applied social psychology. The increasing social and technical complexity of modern living makes for more elevated probabilities that one’s expected length and quality of life are more threatened by incompetency, or even negligence, on the part of entrusted, responsible individuals and organizations, in the performance of their duties. Consequently, this division of labor in a complex society makes one more vulnerable to other’s failures. Institutional legitimacy, therefore, becomes an important variable.

Some have reasoned that social trust is the process by which individuals assign to other persons, groups, agencies, and institutions the responsibility to work on certain tasks [43]. Freudenburg [33,34] has contended that the degree to which people trust the abilities of science, business, and government to manage danger is the most important predictor of perceived risk. Some have argued that conflicts related to risk management are not due to easily dismissed factors such as public ignorance or irrationality, but instead are linked to the ongoing technological advances and social changes that have systematically destroyed trust in the sources of both control and communication of risk [44]. Viewing trust through a lens of a largely impersonal and compartmentalized technocracy-driven society would suggest less worry about technical issues related to differences in risk probabilities, and, instead, apprehension would lie with how risk decisions are made. Several similarly themed questions could arise: Are risk assessment decisions and the resultant risk management policies derived from those conclusions made fairly? Whose interests are ultimately served by the decision process? Why is one conclusion more valid than another? And, which authority is the ultimate arbitrator in the process?

As both Luhmann [24] and Earle and Cvetkovich [45] have noted, some people

make a judgment about trust and they determine that the outcome of actually being wrong is worth the risk. Others may be helped to assume this risk if they were to believe that they have the means to hold the trusted individual accountable, even if they were to learn later that trust was wrongly placed. This accountability for some may substitute for trust or, in the reverse, serve as a signal of distrust.

The concerns of the arbiters in these risk management systems (e.g., scientists, governmental regulators, etc.) are obviously quite different. For the many “experts” in positions of assessing, managing, or developing policy related to risks, there is a hope that the public will, with sufficient time and increased knowledge, view the risk in the same light as themselves. The concept of a knowledge deficit to be overcome with the help of the expert-as-information-provider, is a common paradigm. Concurrently, the very same expert can also be seen as a decision influencer, perhaps even a decision maker, by some. The noted breakdown in the process of placing trust in the knowledge of the expert has been tied to a crumbling in trust of what is assumed to be scientific rationality. An acknowledgement of that breakdown has even been sometimes used to enhance the credibility of decisions [43]. Not only is the expertise of scientists a factor, but as Fessenden-Raden *et al.* [46] found, for example, the level of trust people have in local political institutions and officials determines “the level of trust they place in the information local officials provide about a risk” (p. 96). They also reported that when state and federal agencies were suspected or even disliked, the risk information those officials provided was generally seen as less “true” than that from other sources (see also White and Eiser [47]).

As Petts [48] has observed, there are many frustrations in store when the scientific view is to be adapted to societal conditions. Petts suggested some likely outcomes: the “public are not interested; people have their own agenda and will not listen to ‘objective science’; involving activists and special interest groups will be detrimental to the education

process [of the experts]; that people inaccurately perceive risks; the scientific complexities underlying current techno-scientific debates make them difficult to discuss; and that the public will always go for the zero-risk option” (p. 359). The portrayal of scientific debates themselves are also fraught with problems in risk reporting. Frequently, stories are characterized by ambiguity and scientific debate, due in large part to the conventions developed for political reporting, utilizing conflicting sources in the debate to express the opposing spectrum of differences, while offering little interpretation of either the validity or the accuracy of the information. Consequently, this leaves consumers of risk reporting often in an uncertain situation; they may be unable to arrive at a decision regarding the veracity of the claims and counterclaims due to this lack of necessary information [49]. Along with these conditions, the intervening factor of trust can only further complicate the science communication process.

BELIEVABILITY OF SOURCES OF A RISK MESSAGE

Apart from the many hindrances that the scientific view faces in conveying information about risks to the public, the credibility of expert sources has also remained a critical area of concern for communication scholars in the literature of persuasion/mass communication studies. Understanding an individual’s perception of the credibility of the source of a risk message has been found to be an essential factor in identifying how individuals receive, process, and react to the plethora of risk messages that are received each day from various information sources [50].

Seminal work by Hovland and Weiss [51] and Hovland *et al.* [52] provided some of the early nuanced studies of source credibility and suggested it was composed of two major components—expertise and trustworthiness. Similarly, Eagly *et al.* [53] framed trustworthiness and expertise in terms of two types of expected bias on the part of the expert. Trustworthiness was transposed to a truthfulness characteristic

and termed a *reporting bias*, which corresponded to how willing a source would be to be truthful about a relevant issue. The degree of knowledge a source has about the issue was equated roughly with expertise. This was termed a *knowledge bias*—the knowledge a target audience attributes to a source. Consequently, trust will be reduced if the source is seen as having a high degree of reporting bias and/or an insufficiently informed knowledge about a particular hazard [54–57]. Hunt *et al.* [58] have noted that, while reporting bias seems to be more important than expertise, having these elements of trust present in “appropriate proportions” is more important for risk communication than simply a high trust rating alone; see also Frewer and Miles [59].

However, a rigid definition of credibility has not remained a research constant. Hovland and Weiss [51] have suggested, for example, that expertise is a less effective persuasive technique than is trustworthiness overall; see also McGinnies and Ward [60]. Issues concerning technological risk also suggest that trustworthiness is a more valued trait [61]. Gunther [62] observed that the relationship the receiver has to the content of a message will have an effect on the perception of the source’s credibility. Frewer *et al.* [63] observed that the relationship between trust and message content is a complex one. The variables in their study included both the level of perceived risk and the level of persuasiveness. If the information presented in a message was low in persuasiveness, greatest trust was observed when it was high in both personal relevance and attributed to the government, or low in personal relevance and attributed to a consumer organization. Some work by Johnson *et al.* [64] and Slater and Rouner [65] indicated that both the plausibility and quality of a message may matter more to assessments of source credibility, and to subsequent attitude change, than a source’s expertise and trustworthiness. Hamilton [66] has also described, for example, how ratings of source expertise and trustworthiness increase as argument quality increases. These different research results over the past few decades suggest that source credibility could be seen

as a relational concept and not as a fixed entity; see also Walls *et al.* [67]. However, it is important to acknowledge that contextual messages have had quite limited usage in studies of trust. A typical example can be found in [68]. An extensive 43-item trust index was employed to measure what was termed *general trust*, without reference to a context. The study listed specific, potential risks or hazards (e.g., “nuclear power”) without specifying what the actual risk was for the person completing the survey. This accounted for only a small fraction of the variance of perceived risk. An improved result was obtained when “specific trust ratings” were employed, a questionnaire asking for ratings of trust in the pertinent authorities for each of 22 hazards. These ratings of trust in specific actors tied to specific risks/benefits (e.g., university scientists, pharmaceutical companies) are a common noncontextual measurement of trust [69, p. 198].

Several other issues endemic to this line of research, typified by Sjöberg [68], have been noted by McComas and Trumbo [70]—the lack of attention to explication in past source credibility research, or indeed the need for defining a difference between the concepts of trust and credibility; quite often they are used interchangeably or combined as one concept. Further, they have detailed the “laundry lists of factors” that have been used at various times to determine what assists or detracts from source credibility, with little effort at standardization [71–73]. As Tretin and Musham [74] have noted, trust is a response that is to some extent tied to the perception of credibility, but not necessarily the only or sometimes the most appropriate response. Research has shown the trustworthiness of people’s personal doctor to be generally high, but their credibility when it came to such issues such as environmental health risks (in terms of how well-informed they were) was generally lower [50]. Metlay [75] also suggested that, depending on situational concerns, two more or less unrelated factors, which he labeled as either affective or competence components, can influence the level of public trust. Indeed, in public health and risk communication research, knowledge is

not always a preeminent factor in trustworthiness; sources knowledgeable about technological risks may sometimes be considered to be the least trustworthy (e.g., industry spokespeople), while unofficial opinion leaders could be accorded greater credibility [46]. Indeed, the findings of Jungermann *et al.* [76] suggested the pitfalls of simply looking at different sources without providing their context and individual comparisons. Some interesting research by Steel *et al.* [77] concluded that values influenced the amount of trust placed in a source of information. Gunther [78] found that opposing social groups (e.g., different political parties) could both find a given message biased, but in opposite directions, supporting “the view that a significant portion of perceived bias stems from causes other than special knowledge” (p. 161).

INVOLVEMENT WITH AN ISSUE AND SOURCE CREDIBILITY

Gunther [78] concluded that credibility is more a function of involvement than of other variables examined in the past. Instead of credibility being the property of the source, it may be more importantly related to the individual and the situation.

Additionally, Sternthal *et al.* [79] found that involvement in an issue mediates the effects of source credibility. Slater [80] noted that the degree of involvement has “emerged as the crucial contingency variable” in much of the persuasion effects research because of its predictive variation in message processing (p. 125). It has been suggested that involvement’s effects have been variable, sometimes even conflicting, due to varied ways of explication and measurement methods [81,82]. Both Johnson and Scileppi [83] and Rhine and Severance [84] noted that, when involvement with an issue was low, a highly credible source was more persuasive than a low-credibility source. Concurrently, when involvement in the issue was high, credibility did not significantly affect persuasiveness. Consequently, those who are most highly involved with a message are more likely to evaluate the quality of

the arguments contained in the message itself [85]. The less involved will, in turn, form opinions about the message based on surrounding cues of the message, such as the credibility of the source; see also Gunther [62] and Walsh-Childers *et al.* [86]. However, Sprecker [87] found that the more highly involved one is with a subject, the more credible the perception of the source (in this context, a scientist).

McComas and Trumbo [70] have observed that, when it comes to the study of trust and credibility, “depending on the tools of measurement as well as the context of the study, researchers may find very different results” (p. 471).

COGNITIVE SIMPLICITY SEEKING AND ASSESSMENT OF TRUST

Kasperson *et al.* [10] have observed that “many psychological, political, and economic perspectives conceptualize trust very narrowly in terms of individual cognition and rational, calculative behavior” (p. 167). An alternate approach of Barber [31] viewed trust as part of the need we all have to “reduce complexity” in social life. Fukuyama [25] notes that most people are simply habituated to a certain minimal degree of honesty; “gathering the necessary information and considering possible alternatives is itself a costly and time-consuming process, one that can be short-circuited by custom and habit” (p. 36). Whether to sneak out of a restaurant without paying the bill is one example; whether to be polite to a stranger is another.

The search for cognitive simplicity as an explanation for the process of acquiring and maintaining trust, led Nisbett and Ross [88] to maintain that when people have to contend with a multitude of facts and values when making a decision about risk, they simplify. Rather than attempt to think their way through to comprehensive, analytic solutions to decision-making problems, they instead rely on strategies such as habit, tradition, neighbors, media, or general rules of thumb. Several studies have also suggested that, if subjects believe that the source’s position on a topic (e.g., risks and benefits from a new technology) was influenced by their background

(i.e., personal history/affiliation and/or attitude toward the topic), then this could lead to a perception of the source as both biased and lacking in credibility and, therefore, untrustworthy [89]. If, however, they believe that the source was influenced by factual evidence alone, they would view him/her as unbiased, credible, and trustworthy [60,90].

Earle and Cvetkovich [45] have further extended the conceptualization of the search for cognitive simplicity as part of a process of acquiring and maintaining trust. They viewed much of the research outlined earlier as a somewhat simplistic rendering of how the process of trust is formulated and operates within society. Earle and Cvetkovich [45] concluded that much research on trust has employed static and unidimensional analyses, with the dualism between competence (expertise) and (fiduciary) responsibility at its core. Under such a traditional interpretation, as was explored, for example, by Ruscio [91], trust forms part of a bridge to representative democracy. If the citizens' elected representative is proven to be incompetent or irresponsible, then the citizens' trust could be "lost."

Like Barber [31] and Earle and Cvetkovich [45] believed that trust serves as a tool to satisfy an integral need we have to reduce complexity in our environment. They reasoned that when making a social judgment, "an individual in effect moves from State A (in which a person is faced with an environment that threatens to overwhelm him with complexity) to State B (in which the cognitive demands of the individual's environment are reduced to a manageable size). For the individual who makes the social trust judgment, trust acts as a bridge from State A to State B, from disequilibrium (an unbalanced, "nonnormal" state) to equilibrium (a balanced, "normal" state)" (p. 106). They maintained that the conventional monitoring process model for competence and responsibility would necessitate the individual remaining at least minimally engaged in some form of continuous supervision of information, and then be able to mindfully integrate it. At the same time, the cognitive, emotional, and behavioral demands this would place on the individual could be viewed

as substantial. Therefore, the function of trust is to reduce complexity through the acceptance of risk in exchange for "forms of cooperation which do not pay off immediately and which are not directly visible as beneficial" [92, p. 24]. Some corollary evidence supports this; individuals are often seekers of cognitive simplicity, only accepting complexity when highly motivated under certain circumstances [93,94].

In moving away from a concept of a rational decision-making process for trusting to one that reflects the need for reducing cognitive complexity, several strategies can be considered. One stratagem involves possible heuristic mental models. Herbert Simon's [95] notion of "satisficing" is but one example. The decision process for determining who or whether to trust, while concurrently deciding upon the veracity of a risk to oneself or others, may be explained by these "cognitive shortcuts." More recent work of Shah and Oppenheimer [96] on choice heuristics used in trust (since trust essentially involves a preference among options that lack a correct answer) has focused on the difficulty of retrieving cues and integrating the information associated with those cues, while concurrently examining all other alternatives. The various ways in which heuristics reduce the effort involved in judgments of trust and levels of trustworthiness is especially realized in the similarity heuristic, the notion that a perceived similarity between oneself and others leads to more positive judgements, including affinities to groups with identities similar to one's own [97], and homogeneous goals and objectives tied to value similarity (to be explored next).

THE LESSENING OF COMPLEXITY THROUGH VALUES

Aaron Wildavsky [98] has suggested that "individuals use a cognitive screen in which the individual making the risk assessment is first deemed trustworthy or not whereupon the finding of riskiness or safeness follows. But how do individuals determine who to trust?" (p. 182). Wildavsky [99] characterized social trust as a resilient strategy

that seeks to reduce cognitive complexity (while at the same time increasing social benefits) by taking risks. To trust, according to Wildavsky, is to risk. Wynne [100] has argued that an accounting of an individual's personal constructs tied to how socially dependent they perceive themselves to be upon other individuals, institutions, or organizations within a given situation, especially under circumstances tied to mediated information, is an important place to start and must be given priority in any research examining trust. Wildavsky [98,99,101] and Earle and Cvetkovich [45,102,103] have extended this reasoning further and have concluded that a better and more cogent theoretical explanation for how we develop trust in our sources of information, including the institution(s) that the sources may represent, must include values as the essential intervening variable. For, when a close symmetry of values exists between the source of information about a risk and the message receiver, much less thought is required in order to place one's trust in the source itself. In other words, we trust people we take to be similar to ourselves. This, according to Wildavsky [101] explains how "the social filter enable[s] people who possess only inches of facts to generate miles of preferences?" (p. 8). According to Wildavsky, it is not a conscious, rational thought process. Value similarity can also directly affect cooperation [104]. Jenkins-Smith [105] has even suggested that an individual's values can act as a kind of cognitive filter to screen information from which verbal images are constructed. Wildavsky [101] maintained that people can know what they believe or whom they trust without knowing how the belief is derived. By mediating their perceptions through their values, "people can grab on to any social handle to choose their preferences. All they need are aids to calculation" (p. 9). Consequently, who (or whether) people trust could be based largely on the shared values of the person, community, or institution that will potentially be the beneficiary of that trust. This is especially urgent when sizing up a risk about which we possess little knowledge or with which we have minimal personal involvement (which is true of many risks),

and could be both a predictor of what news stories we pay attention to at the outset and whether we then choose to apply further cognitive processes to an analysis of the stories themselves. McLeod *et al.* [106] suggested some conceptual justification for this avenue of research when they reported that values influenced media usage and interpersonal communication patterns.

Another approach employing values as a mediating variable, the Salient Value Similarity model [107–109] hopes to predict trust if salient values are similar and distrust if salient values are dissimilar by construing trust as a social emotion elicited by a situation that implies the question, "Should I rely on this person?"

RISK AND THE IMPORTANCE OF ASSESSING BENEFITS

Lastly, while the modern term for risk has evolved to refer to negative outcomes, perceived *benefits* still matter. Negative outcomes are, after all, usually only half of the equation. Benefits need not be material goods; convenience would be a typical example of an "intangible" benefit. The importance of benefits for our understanding of how the multifaceted processes of social interaction take place within trust-building and trust-destroying contexts would seem apparent. Research by Siegrist [69] and Siegrist *et al.* [109], for example, has shown that trust has an impact on both perceived risk and perceived benefit. Social trust was positively related to perceived benefit but negatively related to perceived risk. Despite the often-noted reduced tolerance for risk by women, in general, when trust and benefits were controlled, no difference between genders was found. However, when the level of trust was controlled for, women were more likely than men to question the perceived utility of science. As in Siegrist [69,110] also demonstrated that trust in companies and scientists conducting research in gene technologies had a strong effect on the overall levels of risk and benefit perceived to be associated with those technologies.

SUMMARY

We have explored the many varied approaches to studying trust. As with any review, the scope of this review is necessarily limited. It attempted to present some of the most influential research conducted within a number of varied fields of study and applied to it risk analysis, risk management, risk perception, and risk communication.

REFERENCES

- Ostrom E, Walker J. Trust and reciprocity: interdisciplinary lessons for experimental research. New York: Russell Sage Foundation; 2005.
- Brehm J, Gates S, editors. Teaching, tasks, and trust: functions of the public executive. New York: Russell Sage Foundation; 2008.
- Kramer RM, Cook KS, editors. Trust and distrust in organizations. New York: Russell Sage Foundation; 2004.
- Braithwaite V, Levi M. Volume 1, Trust and governance. New York: Russell Sage Foundation; 1998.
- Rousseau DM, Sitkin SB, Burt RS, *et al.* Not so different after all: a cross-discipline view of trust. *Acad Manage Rev* 1998;23:393–404.
- Slovic P. Perception of risk: reflections on the psychometric paradigm. In: Krinsky S, Golding D, editors. Social theories of risk. Westport (CT): Praeger; 1992. pp. 117–152.
- Slovic P. Risk perception and trust. In: Molak V, editor. Fundamentals of risk analysis and risk management. Boca Raton (FL): CRC-Lewis Publishers; 1997. pp. 233–245.
- Bord RJ, O'Connor RE. Determinants of risk perceptions of a hazardous waste site. *Risk Anal* 1992;12:411–416.
- Johnson BB, Slovic P. Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. *Risk Anal* 1995;15:485–494.
- Kasperson RE, Golding D, Tuler S. Social distrust as a factor in siting hazardous facilities and communicating risks. *J Soc Issues* 1992;48:161–187.
- Slovic P. Perceived risk, trust, and democracy. *Risk Anal* 1993;13(6):675–682.
- Mutz DC. Social trust and e-commerce: experimental evidence for the effects of social trust on individuals' economic behavior. *Public Opin Q* 2005;69(3):393–416.
- Rotter JB. A new scale for the measurement of interpersonal trust. *J Pers* 1967;35(4):651–665.
- Rotenberg KJ. A measure of the trust beliefs of elderly individuals. *Int J Aging Hum Dev* 1990;30(2):141–152.
- Chun K, Campbell JB. Dimensionality of the rotter interpersonal trust scale. *Psychol Rep* 1974;35(3):1059–1070.
- Parks CD, Hulbert L. High and low trusters' responses to fear in a payoff matrix. *J Conflict Resolut* 1995;39:718–730.
- Yamagishi T. The provision of a sanctioning system as a public good. *J Pers Soc Psychol* 1986;51:110–116.
- Coleman JS. Foundations of social theory. Cambridge (MA): Belknap Press; 1990.
- Renn O, Levine D. Credibility and trust in risk communication. In: Kasperson RE, Stallen PM, editors. Communicating risks to the public: international perspectives. Dordrecht, Holland: Kluwer Academic Publishers; 1991. pp. 175–218.
- Rempel J, Holmes J. How do I trust thee? *Psychol Today* 1986;19(1):28–34.
- Zucker L. Production of trust: institutional sources of economic structure 1840–1920. *Res Org Behav* 1986;8:53–111.
- Lewis JD, Weigart A. Trust as a social reality. *Soc Forces* 1985;63:967–985.
- Koller M. Risk as determinant of trust. *Basic Appl Soc Psychol* 1988;9:265–267.
- Luhmann N. Familiarity, confidence, trust: problems and alternatives. In: Gambetta D, editor. Trust: making and breaking cooperative relations. London: Basil Blackwell; 1988. pp. 94–107.
- Fukuyama F. Trust: the social virtues and the creation of prosperity. New York: Free Press; 1995.
- Cappella JN. Cynicism and social trust in the new media environment. *J Commun* 2002;52(2):229–241.
- Lambert RD, Curtis JE, Brown SD. Effects of identification with governing parties on feelings of political efficacy and trust. *Can J Polit Sci* 1986;19:705–727.
- Parker GR. The role of constituent trust in congressional elections. *Public Opin Q* 1989;53(2):175–196.
- Howell SE, Fagan D. Race and trust in government: testing the political reality model. *Public Opin Q* 1988;52(3):343–350.

30. Wald KD, Lupfer MB. The presidential debate as a civics lesson. *Public Opin Q* 1978;42(3):342–353.
31. Barber B. *The logic and limits of trust*. New Brunswick (NJ): Rutgers University Press; 1983.
32. Rothbart M, Park B. On the confirmability and disconfirmability of trait concepts. *J Pers Soc Psychol* 1986;50:131–142.
33. Freudenburg WR. Risk and recreancy: Weber, the division of labor, and the rationality of risk perceptions. *Soc Forces* 1993;71(4):909–932.
34. Freudenburg WR. The ‘risk society’ reconsidered: recreancy, the division of labor, and risks to the social fabric. In: Cohen J, editor. *Risk in the modern age: social theory, science and environmental decision-making*. New York: St. Martin’s Press; 2000. pp. 107–120.
35. Fischhoff B. Risk, a guide to controversy. In: Council NR editor. *Improving risk communication*. Washington (DC): National Academy Press; 1989. pp. 211–308.
36. Kasperson R, Kasperson JX, Renn O. The social amplification of risk: progress in developing an integrative framework. In: Krinsky S, Golding D, editors. *Social theories of risk*. London: Praeger; 1992.
37. Deutsch M. *The resolution of conflict: Constructive and destructive processes*. New Haven, Connecticut: University Press; 1973.
38. Rotter JB. Interpersonal trust, trustworthiness and gullibility. *Am Psychol* 1980;35(1):1–7.
39. Stebbins R. Book review of “The logic and limits of trust” (Rutgers Univ. Press 1983) by Barber, Bernard. *Can Rev Sociol Anthropol* 1988;25:475–477.
40. Kramer RM, Tyler TR. *Trust in organizations: frontiers of theory and research*. Thousand Oaks: Sage; 1996.
41. Johnson B. Exploring dimensionality in the origins of hazard-related trust. *J Risk Res* 1999;2(4):325–354.
42. Frewer LJ, Scholderer J, Bredahl L. Communicating about the risks and benefits of genetically modified foods: the mediating role of trust. *Risk Anal* 2003;23(6):1117–1133.
43. Wynne B. Misunderstood misunderstandings: social identities and public uptake of science. *Public Underst Sci* 1992;1(3):281–304.
44. Friedman S. Risk management: the public versus the technical experts. In: Wilkins L, Patterson P, editors. *Volume 27, Risky business: communicating issues of science, risk, and public policy*. New York: Greenwood Press; 1991. pp. 31–41.
45. Earle TC, Cvetkovich GT. *Social trust: toward a cosmopolitan society*. Westport (CT): Praeger; 1995.
46. Fessenden-Raden F, Fitchen JM, Heath JS. Providing risk information in communities: factors influencing what is heard and accepted. *Sci Technol Human Values* 1987;12:94–101.
47. White MP, Eiser JR. Marginal trust in risk managers: building and losing trust following decisions under uncertainty. *Risk Anal* 2006;26(5):1187–1203.
48. Petts J. The public-expert interface in local waste management decisions: expertise, credibility and process. *Public Underst Sci* 1997;6(4):359–381.
49. Major AM, Atwood LE. Environmental risks in the news: issues, sources, problems, and values. *Public Underst Sci* 2004;13(3):295–308.
50. McCallum DB, Hammond SL, Covello VT. Communicating about environmental risks: how the public uses and perceives information sources. *Health Educ Q* 1991;18(3):349–361.
51. Hovland CI, Weiss W. The influence of source credibility on communication effectiveness. *Public Opin Q* 1951;15:633–650.
52. Hovland CI, Janis I, Kelly HH. *Communication and persuasion*. New Haven (CT): Yale University Press; 1953.
53. Eagly AH, Wood W, Chaiken S. Causal inferences about communicators and their effect on opinion change. *J Pers Soc Psychol* 1978;36(4):424–435.
54. Gutteling JM, Wiegman O. *Volume 8, Exploring risk communication*. Dordrecht: Kluwer Academic Publishers; 1996.
55. McGinnies E, Ward CD. Persuasibility as a function of source credibility and locus of control: five cross-cultural experiments. *J Pers* 1974;8:360–371.
56. McGuire WJ. Selective exposure: a summing up. In: Abelson RP, Aronson E, McGuire WJ, *et al.* editors. *Theories of cognitive consistency: a sourcebook*. Chicago: Rand McNally; 1968.
57. McGuire WJ. Attitudes and attitude change. In: Lindzey G, Aronson A, editors. *Volume 2, Handbook of social psychology*. 3rd ed. New York: Random House; 1985. pp. 233–346.

58. Hunt S, Frewer L, Shepherd R. Public trust in sources of information about radiation risk in the UK. *J Risk Res* 1999;2(2):167–180.
59. Frewer LJ, Miles S. Temporal stability of the psychological determinants of trust: implications for communication about food risks. *Health Risk Soc* 2003;5(3):259–271.
60. McGinnies E, Ward CD. Better liked than right: trustworthiness and expertise as factors in credibility. *Pers Soc Psychol Bull* 1980;6(3):467–472.
61. Peters HP. The credibility of information sources in West Germany after the Chernobyl disaster. *Public Underst Sci* 1992;1(3):325–343.
62. Gunther AC. Attitude extremity and trust in media. *Journal Q* 1988;65(2):279–287.
63. Frewer LJ, Howard C, Hedderley D, *et al.* Reactions to information about genetic engineering: Impact source characteristics, perceived personal relevance, and persuasiveness. *Public Underst Sci* 1999;8(1):35–50.
64. Johnson B, Sandman P, Miller P. Testing the role of technical information in public risk perception. *Risk: Health Saf Environ* 1992;3(4):265–279.
65. Slater M, Rouner D. How message evaluation and source attributes may influence credibility assessment and belief change. *J Mass Commun Q* 1996;73:974–991.
66. Hamilton M. Message variables that mediate and moderate the effect of equivocal language on source credibility. *J Lang Soc Psychol* 1998;17:109–143.
67. Walls J, Pidgeon N, Weyman A, *et al.* Critical trust: understanding lay perceptions of health and safety risk regulation. *Health Risk Soc* 2004;6(2):133–150.
68. Sjöberg L. Limits of knowledge and the limited importance of trust. *Risk Anal* 2001;21(1):189–198.
69. Siegrist M. The influence of trust and perceptions of risks and benefits on the acceptance of gene technology. *Risk Anal* 2000;20(2):195–203.
70. McComas KA, Trumbo GW. Source credibility in environmental health - risk controversies: application of Meyer's credibility index. *Risk Anal* 2001;21(3):467–480.
71. Cronkhite G, Liska J. A critique of factor analytic approaches to credibility. *Commun Monogr* 1976;43:91–107.
72. Infante DA, Parker KE, Clarke CH, *et al.* A comparison of factor and functional approaches to source credibility. *Comm Q* 1983;31(1):43–48.
73. Sundar SS. Exploring receivers' criteria for perception of print and online news. *J Mass Commun Q* 1999;76(2):373–386.
74. Trettin L, Musham C. Is trust a realistic goal of environmental risk communication? *Environ Behav* 2000;32(3):410–426.
75. Metlay D. Institutional trust and confidence: a journey into a conceptual quagmire. In: Cvetkovich G, Löfstedt R, editors. *Social trust and the management of risk*. 1st ed. London: Earthscan Publications Ltd; 1999. pp. 196–207.
76. Jungermann H, Pfister HR, Fischer K. Credibility, information preferences, and information interests. *Risk Anal* 1996;16(2):251–261.
77. Steel BN, Lovrich N, Pierce J. Trust in natural resource information sources and post-materialist values: a comparative study of U.S. and Canadian citizens in the Great Lakes area. *J Environ Syst* 1993;22(2):123–126.
78. Gunther AC. Biased press or biased public? Attitudes toward media coverage of social groups. *Public Opin Q* 1992;56:147–167.
79. Sternthal B, Phillips LW, Dholakia R. The persuasive effect of source credibility: a situational analysis. *Public Opin Q* 1978;42(3):285–314.
80. Slater M. Persuasion processes across receiver goals and message genres. *Commun Theory* 1997;7(2):125–148.
81. Johnson BT, Eagly AH. Effects of involvement on persuasion: a meta-analysis. *Psychol Bull* 1989;106:290–314.
82. Salmon CT. Perspectives on involvement in consumer and communication research. In: Dervin B, Voigt M, editors. *Volume 7, Progress in communication sciences*. Beverly Hills (CA): Sage; 1987. pp. 244–268.
83. Johnson H, Scileppi J. Effects of ego-involvement conditions on attitude change to high and low credibility communicators. *J Pers Soc Psychol* 1969;13(1):31–36.
84. Rhine R, Severance L. Ego-involvement, discrepancy, source credibility, and attitude change. *J Pers Soc Psychol* 1970;16:175–190.
85. Salmon C. Perspectives on involvement in consumer and communication research. In: Dervin B, Voigt M, editors. *Volume 7, Progress in communication sciences*. Norwood (NJ): Ablex; 1986. pp. 243–268.

86. Treise D, Walsh-Childers K, Weigold MF, *et al.* Cultivating the science internet audience: impact of brand and domain on source credibility for science information. *Sci Commun* 2003;24(3):309–332.
87. Sprecker K. How involvement, citation style, and funding source affect the credibility of university scientists. *Sci Commun* 2002;24(1):72–97.
88. Nisbett R, Ross L. *Human inference: strategies and shortcomings of social judgement.* New Jersey: Prentice-Hall; 1980.
89. Wiener JL, Mowen JC. Source credibility: on the independent effects of trust and expertise. *Adv Consum Res* 1983;13:306–310.
90. Morton T, Duck J. Communication and health beliefs: mass and interpersonal influences on perceptions of risk to self and others. *Communic Res* 2001;28(5):602–626.
91. Ruscio KP. Trust, democracy, and public management: a theoretical argument. *J Public Admin Theory* 1996;6(3):461–477.
92. Luhmann N. *Trust and power.* New York: John Wiley & Sons; 1980.
93. Petty RE, Cacioppo JT. Communication and persuasion: central and peripheral routes to attitude change. New York: Springer; 1986.
94. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. In: Berkowitz L, editor. *Volume 19, Advances in experimental social psychology.* San Diego (CA): Academic Press; 1986. pp. 123–205.
95. Simon H. *Reason in human affairs.* Stanford: Stanford University Press; 1983.
96. Shah AK, Oppenheimer DM. Heuristics made easy: an effort-reduction framework. *Psychol Bull* 2008;134:207–222.
97. Foddy M, Platow MJ, Yamagishi T. Group-based trust in strangers: the role of stereotypes and expectations. *Psychol Sci* 2009;20:419–422.
98. Wildavsky A. The comparative study of risk perception: a beginning. In: Ruck B, editor. *Risk is a construct: perceptions of risk perception.* Munchen: Knesebeck; 1993. pp. 179–195.
99. Wildavsky A. *Searching for safety.* New Brunswick (NJ): Transaction Publishers; 1988.
100. Wynne B. Risk and social learning: reification to engagement. In: Krinsky S, Golding D, editors. *Social theories of risk.* Westport (CT): Praeger; 1992. pp. 275–297.
101. Wildavsky A. Choosing preferences by constructing institutions: a cultural theory of preference formation. *Am Polit Sci Rev* 1987;81(1):2–15.
102. Earle TC, Cvetkovich GT. Culture, cosmopolitanism, and risk management. *Risk Anal* 1997;17(1):55–66.
103. Earle TC, Cvetkovich GT. Volume G4, Risk judgement and the communication of hazard information: toward a new look in the study of risk perception. In: Covello VT, editor. *Environmental impact assessment, technology assessment, and risk analysis.* Berlin: Springer; 1985.
104. Chou LF, Wang AC, Wang TY, *et al.* Shared work values and team member effectiveness: the mediation of trustfulness and trustworthiness. *Hum Relat* 2009;61:1713–1742.
105. Jenkins-Smith HC. Nuclear imagery and regional stigma: testing hypotheses of image acquisition and valuation regarding Nevada. Albuquerque (NM): University of New Mexico, Institute for Public Policy; 1993.
106. McLeod JM, Sotirovic M, Holbert RL. Values as sociotropic judgements influencing communication patterns. *Communic Res* 1998;25(5):453–485.
107. Cvetkovich G, Winter PL. Trust and social representations of the management of threatened and endangered species. *Environ Behav* 2003;35(2):286–307.
108. Jackson J, Allum N, Gaskell G. Bridging levels of analysis in risk perception research: the case of the fear of crime. *Forum: Qual Soc Res [On-line Journal]* 2006;7(1). Article No. 20. Available at <http://www.qualitative-research.net/fqs-texte/1-06/06-1-20-e.htm>.
109. Siegrist M, Cvetkovich G, Roth C. Salient value similarity, social trust, and risk/benefit perception. *Risk Anal* 2000;20(3):353–362.
110. Siegrist M. A causal model explaining the perception and acceptance of gene technology. *J Appl Soc Psychol* 1999;29(10):2093–2106.

TWO-STAGE STOCHASTIC INTEGER PROGRAMMING: A BRIEF INTRODUCTION

SHABBIR AHMED

H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia

INTRODUCTION

A standard form for a two-stage stochastic integer program (SIP) is as follows:

$$\begin{aligned} \min_x \quad & c^\top x + \mathbb{E}_P[Q(x, \omega)], \\ \text{s.t.} \quad & Ax = b, \\ & x \in \mathbb{R}_+^{n_1-p_1} \times \mathbb{Z}_+^{p_1}, \end{aligned} \quad (1)$$

where

$$\begin{aligned} Q(x, \omega) := \min_y \quad & q^\top y, \\ \text{s.t.} \quad & Wy = h - Tx, \\ & y \in \mathbb{R}_+^{n_2-p_2} \times \mathbb{Z}_+^{p_2}. \end{aligned} \quad (2)$$

Above, n_1, n_2, p_1, p_2 are nonnegative integers with $p_1 \leq n_1$ and $p_2 \leq n_2$, x represents the first-stage decisions and y represents the second-stage decisions, and ω represents the uncertain data for the second stage (the parameters (q, h, T) are actual realization of the random data) with known distribution P . Throughout, we assume that the matrix W is deterministic. Much of the structural results in SIP are for this case. Problem (1) seeks a first-stage decision that minimizes first-stage cost and the *expected* cost of second-stage (recourse) decisions. Note that both first- and second-stage variables are restricted to be mixed integers ($\mathbb{R}$ and $\mathbb{Z}$ denote reals and integers, respectively). If the second-stage variables are continuous (i.e., $p_2 = 0$) then problem (1) involves minimizing a convex objective subjected to mixed-integer constraints. Much of the theory and algorithms for two-stage stochastic

linear programs (LP) (that do not rely on convexity of the first-stage constraints) are then applicable. By two-stage SIPs, we refer to problems where $p_2 > 0$.

A variety of applications in energy planning [1], manufacturing [2], logistics [3], and so on, can be formulated as two-stage SIPs of the form given in Equation (1). In this article, we briefly review some important progress in theory and algorithms for solving Equation (1). There has been significant development in extensions of the two-stage SIP framework; for example, to multistage setting [4] and involving risk averse objectives [5]. However, this article is limited to models of the form (Eq. 1).

STRUCTURE

In this section, we discuss the structure of the expected value function

$$\mathbb{E}_P[Q(x, \omega)],$$

where Q is given by Equation (2). Consider first the value function of a deterministic mixed-integer program (MIP) as a function of the objective and right-hand side vectors

$$\begin{aligned} \Phi(q, t) := \min_y \quad & q^\top y, \\ \text{s.t.} \quad & Wy = t, \\ & y \in \mathbb{R}_+^{n-p} \times \mathbb{Z}_+^p, \end{aligned} \quad (3)$$

where W is a $m \times n$ matrix, and t is an m -vector. We make the following assumptions:

1. For every $t \in \mathbb{R}^m$, there exists $y \in \mathbb{R}_+^{n-p} \times \mathbb{Z}_+^p$ such that $Wy = t$;
2. The entries of W are integers.

Assumption 1 guarantees that MIP Equation (3) is feasible for all $t \in \mathbb{R}^m$, and assumption 2 guarantees that Equation (3) has an optimal solution, if it is feasible and bounded. It is sufficient to assume that W is rational; however, we assume integrality for the ease of exposition. Let us

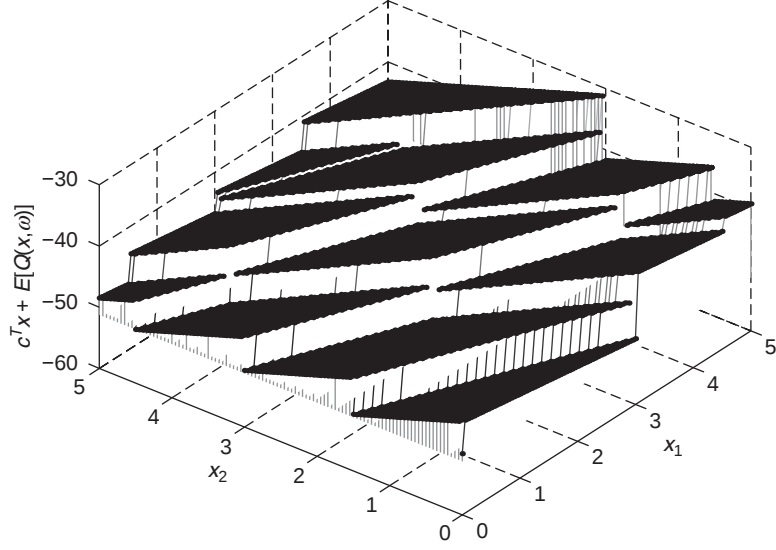


Figure 1. Objective function of a small SIP.

denote $\mathcal{Q} := \{q : \Pi(q) \neq \emptyset\}$, where $\Pi(q) := \{\pi \in \mathbb{R}^m : W^\top \pi \leq q\}$. Note that $\mathcal{Q}$ is a closed convex set in $\mathbb{R}^n$.

Theorem 1.

- (a) The value function $\Phi(\cdot, \cdot)$ is lower semi-continuous over $\mathcal{Q} \times \mathbb{R}^m$.
- (b) Given $t \in \mathbb{R}^m$, the function $\Phi(\cdot, t)$ is continuous over $\mathcal{Q}$; and given $q \in \mathcal{Q}$ the function $\Phi(q, \cdot)$ is lower semicontinuous over $\mathbb{R}^m$,
- (c) Given $q \in \mathcal{Q}$ the function $\Phi(q, \cdot)$ is continuous over $\mathbb{R}^m$ except over (at most) a countable union of hyper-planes.
- (d) Given $q \in \mathcal{Q}$ the function $\Phi(q, \cdot)$ is sub-additive over $\mathbb{R}^m$, that is, $\Phi(q, t^1 + t^2) \leq \Phi(q, t^1) + \Phi(q, t^2)$.
- (e) Given $t \in \mathbb{R}^m$, the function $\Phi(\cdot, t)$ is concave over $\mathcal{Q}$.

Next, we consider the structure of $\mathbb{E}_P[Q(x, \omega)]$. This function is given by the expectation of the MIP value function $Q(x, \omega) = \Phi(q, h - Tx)$ with respect to the distribution P of the data $\omega = (q, h, T)$.

Theorem 2. In addition to assumptions 1 and 2, assume that the random set $\Pi(q)$

is nonempty with probability one, $\mathbb{E}[||q|| ||h||] < +\infty$ and $\mathbb{E}[||q|| ||T||] < +\infty$. Then, the expected value function $\mathbb{E}_P[Q(x, \omega)]$ is well defined, real-valued, and lower semi-continuous on $\mathbb{R}^{n1}$.

The above result establishes sufficient conditions for the expected value function to be well defined and lower semicontinuous. Apart from this continuity property, very little can be said of the structure of expected value functions of MIPs, in general. Figure 1 illustrates the objective function $c^\top x + \mathbb{E}_P[Q(x, \omega)]$ of a small instance of Equation (1) with $n_1 = 2$. We see that the objective function has none of the desirable properties for optimization. It is nonconvex and even discontinuous.

When the distribution of the ω is absolutely continuous, that is, it has a density function, then it can be further shown that $\mathbb{E}_P[Q(x, \omega)]$ is continuous on $\mathbb{R}^{n1}$, however, it is in general nonconvex.

CHALLENGES AND PROGRESS

There are three levels of difficulty in solving SIPs of the form (1).

1. Evaluating the second-stage cost for a fixed first-stage decision and a

particular realization of the uncertain parameters. Note that this involves solving an instance of the second-stage problem (Eq. 2), which may be an NP-hard integer program (IP) and involve significant computational difficulties.

2. *Evaluating the expected second-stage cost for a fixed first-stage decision.* If the uncertain parameters have continuous distribution, this involves integrating the value function $Q(x, \cdot)$ of an IP, and is, in general, impossible. If the uncertain parameters have a discrete distribution, this involves solving a (typically huge) number of similar IPs.
3. *Optimizing the expected second-stage cost.* As seen in the section titled “Structure,” the value function of an IP is nonconvex and often discontinuous. Consequently, the expected second-stage cost function $\mathbb{E}[Q(\cdot, \omega)]$ is nonconvex in x . The optimization of such a complex objective function poses severe difficulties.

In the following sections, we briefly mention some of the theoretical and algorithmic progress toward addressing the aforementioned difficulties in two-stage stochastic IP.

Difficulty 1

It is typically assumed that a *single* evaluation of the second-stage problem is somehow tractable. Without this assumption, very little progress in optimizing the expected value of this IP is possible. There has been some work in obtaining approximate solutions to SIPs through approximate solutions to the integer second-stage problem [2].

Difficulty 2

Let us now consider the difficulty of evaluating the expected value of the second-stage IP $\mathbb{E}[Q(x, \omega)]$ for a given first-stage decision x . As mentioned earlier, if the distribution of the uncertain parameters is continuous or if, in case of discrete distributions, the number of possible realizations is extremely large, then it is practically impossible to evaluate $\mathbb{E}[Q(x, \omega)]$ exactly. In this case, one has to resort to approximating the underlying

probability distribution by a manageable distribution.

For example, if the underlying distribution is continuous, one may approximate it via discretization. Theoretical stability results for SIPs [6,7] suggest that if the approximate distribution is not too “far” from the true distribution, then the optimal solution to the SIP involving the approximate distribution is close to the true optimal solution.

Alternatively, one may use statistical estimates of the expected value function via Monte Carlo sampling. This can be done in one of two ways. In *interior sampling* approaches, the estimation of $\mathbb{E}[Q(x, \omega)]$ is carried out *within* the algorithm used to optimize this function. For example, in the *stochastic branch and bound algorithm* [8], the feasible domain of the first-stage variables x is recursively partitioned into subsets, and statistical upper and lower bounds on the objective function $c^\top x + \mathbb{E}[Q(x, \omega)]$ over these subsets are obtained via sampling. These bounds are used to discard inferior subsets of the feasible domain, and to further partition the promising subsets to eventually isolate a subset containing an approximate optimal solution. In *exterior sampling* approaches, the sampling and optimization are decoupled. A Monte Carlo sample of the uncertain parameters is generated, and the expectation objective in the problem is replaced by a sample average. The resulting approximation of the problem is then solved, and its solution serves as a candidate solution to the true problem. By repeating the sampling-optimization procedure several times, it is possible to obtain statistical confidence intervals on the obtained candidate solutions. It can be shown that the number of samples needed to get a fairly accurate solution with high probability is not too large. Discussion of theoretical and algorithmic issues pertaining to the above approach in the context of SIP can be found in Refs 9 and 10.

Regardless of how the underlying distribution is approximated, an evaluation of the expected second-stage objective value (under the approximate distribution) requires solving many similar IPs. Owing to the absence

of a computationally useful duality theory for IP, it is very difficult to take advantage of the similarities in the different second-stage IPs. When the second-stage variables are pure integers, several proposals for using Gröbner basis and other test set-based methods from computational algebra for exploiting IP problem similarity have been put forth in Refs 11–13. For the case of mixed-integer subproblems, if a cutting plane method is used, then under some conditions, it is possible to transform a cut (or a valid inequality) derived for one of the second-stage subproblems into a cut for another subproblem by exploiting similarity [14,15].

Difficulty 3

Much of the development in SIP has been toward the difficulty of optimizing $f(x) := c^\top x + \mathbb{E}[Q(x, \omega)]$, that is, the sum of the first-stage and the expected second-stage costs. We classify these developments as follows:

Convex Approximations of the Value Function. Consider an SIP with *simple integer recourse* with only right-hand side uncertainty, that is, q , T are deterministic; the second-stage problem has a special structure such that the value function Q is separable and is given by

$$Q(x, \omega) = \sum_{i=1}^m \phi_i(h_i - T_i x),$$

and each separable piece ϕ_i is of the form

$$\phi(t) = q^+ [t]_+ + q^- [t]_-, \quad t \in \mathbb{R},$$

where $[t]_+ = \max\{0, [t]\}$ and $[t]_- = \max\{0, -[t]\}$. In this case, a single evaluation of $f(x)$ is easy, however, owing to the nonconvex nature of ϕ , the function $f(x)$ is difficult to optimize. Fortunately, it has been shown [16] that an expectation of the continuous counterpart of the simple integer recourse function $Q(x, \omega)$ (i.e., where the recourse variables are continuous instead of being discrete valued) under a particular class of distributions of ω provides the tightest convex underestimator for $\mathbb{E}[Q(x, \omega)]$ over its entire domain. Recently, similar

results for constructing convex approximations of general integer recourse functions (SIPs involving pure integer second-stage variables) by perturbing the underlying distribution have been obtained [17]. These convex approximating functions are amenable to optimization and can be used to provide strong lower bounds within some of the algorithms for optimizing $f(x)$, discussed next. An open problem is the refinement of these approximations using additional constraints on the first-stage variables, for example, bounds on x .

Stagewise Decomposition Algorithms. This class of algorithms adopt the natural viewpoint of optimizing the objective function $f(x) := c^\top x + \mathbb{E}[Q(x, \omega)]$ over the set of feasible first-stage decisions (say denoted by X). Note that $\mathbb{E}[Q(x, \omega)]$ is not available in closed-form, nor is it suited for direct optimization. Typical algorithms in this class progress in the following manner. In an iteration i , the algorithm builds and/or refines a computationally tractable approximation (typically an underestimator) $\hat{Q}_i(x)$ of $\mathbb{E}[Q(x, \omega)]$. The underestimating function $c^\top x + \hat{Q}_i(x)$ is optimized with respect to the first-stage variables (this optimization problem is often referred to as the *master problem*) to obtain a lower bound on the true optimal objective value as well as to provide a candidate first-stage solution x^i . Corresponding to the candidate solution, the second-stage expected value function $\mathbb{E}[Q(x^i, \omega)]$ is evaluated. Assuming that the distribution of ω is discrete, this step involves independent solution of the second-stage problems for each realization of ω , allowing for a computationally convenient decomposition. The value $c^\top x^i + \mathbb{E}[Q(x^i, \omega)]$ provides an upper bound on the optimal objective value. The evaluation of $\mathbb{E}[Q(x^i, \omega)]$ also provides information on how the approximation $\hat{Q}_i$ is to be updated/refined to $\hat{Q}_{i+1}$ for the master problem of iteration $i+1$. The process continues until the bounds have converged. The details of the various stagewise decomposition algorithms differ mainly in how the approximation $\hat{Q}_i$ is constructed and updated.

For SIPs with binary first-stage variables and mixed-integer second-stage variables,

the *integer L-shaped method* [18] approximates the second-stage value function by linear “cuts” that are exact at the binary solution where the cut is generated and are underestimates at other binary solutions. The optimization of the master problem, that is, optimizing $c^\top x + \hat{Q}_i(x)$ with respect to the first-stage binary variables, is carried out via a branch-and-bound strategy. As soon as a new first-stage binary solution is encountered in the branch-and-bound search, the second-stage subproblems are solved to generate a new cut and to refine the approximation $\hat{Q}_i$. The integer L-shaped method requires that the second-stage integer problems (corresponding to the current candidate first-stage solution x^i) are all solved to optimality before a valid cut can be generated. Recall that typical IP algorithms progress by solving a sequence of intermediate LP problems. Using disjunctive programming techniques, it is possible to derive cuts from the solutions to these intermediate LPs that are valid underestimators of $\mathbb{E}[Q(x, \omega)]$ at all binary first-stage solutions [15, 19]. This avoids solving difficult integer second-stage problems to optimality in all iterations of the algorithm, providing significant computational advantage.

For SIPs where the first-stage variables are not necessarily all binary, dual functions from the second-stage IP can, in principle, be used to construct cuts to build the approximation $\hat{Q}_i$ [20]. Owing to the nonconvex nature of IP dual functions, the cuts are no longer linear, resulting in a nonconvex master problem. If the second-stage variables are pure integers (and the first-stage variables are mixed-integers), then it can be shown that $\mathbb{E}[Q(x, \omega)]$ is a piecewise constant over subsets that form a partitioning of the feasible region of x [12]. Optimization of $c^\top x + \mathbb{E}[Q(x, \omega)]$ over such a subset is easy. This leads immediately to a scheme where the subsets are enumerated, and the one over which the objective function value is least is chosen. By exploiting certain monotonicity properties, the subsets can be enumerated efficiently within a branch-and-bound strategy [21]. Additional properties of the MIP value function $Q(x, \omega)$, such as subadditivity, can be used to improve the method [22].

Scenariowise Decomposition. Assuming the distribution of ω is discrete, that is the random parameter takes one of a finite set of values (scenarios) $\{\omega_1, \dots, \omega_S\}$ having probabilities $\{p_1, \dots, p_S\}$, the two-stage SIP can be reformulated as follows:

$$\begin{aligned} \min \quad & \sum_{s=1}^S p_s (c^\top x_s + q_s^\top y_s) \\ \text{s.t.} \quad & Ax_s = b \quad s = 1, \dots, S \\ & T_s x_s + W_s y_s = h_s \quad s = 1, \dots, S \\ & x_s \in \mathbb{R}_+^{n_1-p_1} \times \mathbb{Z}_+^{p_1} \quad s = 1, \dots, S \\ & y_s \in \mathbb{R}_+^{n_2-p_2} \times \mathbb{Z}_+^{p_2} \quad s = 1, \dots, S \\ & x_1 = x_2 = \dots = x_S. \end{aligned}$$

Note that copies of the first-stage variable have been introduced for each scenario. The last constraint, known as the *nonanticipativity* constraint, guarantees that the first-stage variables are identical across the different scenarios. Consider the Lagrangian dual problem obtained by relaxing the nonanticipativity constraints through the introduction of Lagrange multipliers. Note that for a given set of multipliers, the problem is separable by scenarios. Thus, the dual function can be evaluated in a decomposed manner. Optimization of the dual function can be performed using standard nonsmooth optimization techniques. However, owing to the nonconvexities, there exists a duality gap, and one needs to resort to a branch-and-bound strategy to prove optimality [23].

Cuts for Deterministic Equivalent MIP. If the number of scenarios (assuming a finite distribution setting) is not astronomical then, a possible approach is to directly solve the deterministic equivalent MIP,

$$\begin{aligned} \min \quad & c^\top x + \sum_{s=1}^S p_s q_s^\top y_s \\ \text{s.t.} \quad & Ax = b \\ & T_s x + W_s y_s = h_s \quad s = 1, \dots, S \\ & x \in \mathbb{R}_+^{n_1-p_1} \times \mathbb{Z}_+^{p_1} \quad s = 1, \dots, S \\ & y_s \in \mathbb{R}_+^{n_2-p_2} \times \mathbb{Z}_+^{p_2} \quad s = 1, \dots, S \end{aligned}$$

using an off-the-shelf solver such as CPLEX or EXPRESS.

One of the most important features of these solvers is the generation of cutting

planes by analyzing the polyhedral structure of the problem. Such cuts significantly improve the LP relaxation of the problem, and expedite LP-based branch-and-bound search. Typically, such cuts are generated by analyzing a single row of the constraint system and do not combine information from multiple rows. In an SIP, the constraint system is repeated with small changes for each scenario. Thus, it is possible to effectively combine cuts from multiple rows corresponding to different scenarios [24]. Such cuts have been shown to significantly improve performance over single row cuts. An open issue is that such multirow cuts link second-stage variables across multiple scenarios, and hence destroy decomposability. Algorithmic techniques that can bypass this issue pose an important challenge.

CONCLUDING REMARKS

This article offers a very limited view of some of the theoretical and algorithmic concepts in SIP. The concepts alluded to here have been significantly enriched and extended in recent years. We have not discussed the large number of important developments in application-specific areas of SIP (see Ref. 25 for a bibliography of applications of SIP), and have also omitted the important progress-made in approximation algorithms for SIP. We hope that this simple introduction will pique the readers' interest toward further exploration of SIP. More detailed surveys on stochastic integer programming can be found in [26,27,28].

REFERENCES

1. Klein Haneveld WK, van der Vlerk MH. Optimizing electricity distribution using two-stage integer recourse models. In: Uryasev S, Pardalos PM, editors. *Stochastic optimization: algorithms and applications*. The Netherlands: Kluwer Academic Publishers; 2001. pp. 137–154.
2. Dempster MAH, Fisher ML, Jansen L, *et al.* Analytical evaluation of hierarchical planning systems. *Oper Res* 1981;29:707–716.
3. Laporte G, Louveaux FV, Mercure H. The vehicle routing problem with stochastic travel times. *Transport Sci* 1992;26:161–170.
4. Romisch W, Schultz R. Multistage stochastic integer programs: an introduction. In: Krumke SO, Rambau J, editors. *Online optimization of large scale systems*. Springer; 2001. pp. 581–600.
5. Schultz R, Tiedemann S. Conditional Value-at-Risk in stochastic programs with mixed-integer recourse. *Math Program* 2006;105:365–386.
6. Schultz R. Continuity properties of expectation functions in stochastic integer programming. *Math Oper Res* 1993;18(3):578–589.
7. Schultz R. On structure and stability in stochastic programs with random technology matrix and complete integer recourse. *Math Program* 1995;70(1):73–89.
8. Norkin VI, Pflug GC, Ruszczyński A. A branch and bound method for stochastic global optimization. *Math Program* 1998;83:425–450.
9. Kleywegt AJ, Shapiro A, Homem-De-Mello T. The sample average approximation method for stochastic discrete optimization. *SIAM J Optim* 2001;12:479–502.
10. Ahmed S, Shapiro A. The sample average approximation method for stochastic programs with integer recourse. Technical Report. ISyE, Georgia Tech; 2002.
11. Hemmecke R, Schultz R. Decomposition of test sets in stochastic integer programming. *Math Program* 2003;94:323–341.
12. Schultz R, Stougie L, van der Vlerk MH. Solving stochastic programs with integer recourse by enumeration: a framework using Gröbner basis reductions. *Math Program* 1998;83:229–252.
13. Tayur SR, Thomas RR, Nataraj NR. An algebraic geometry algorithm for scheduling in the presence of setups and correlated demands. *Math Program* 1995;69(3):369–401.
14. Caroe CC. Decomposition in stochastic integer programming [PhD thesis]. University of Copenhagen; 1998.
15. Higle JL, Sen S. The C^3 theorem and a D^2 algorithm for large scale stochastic integer programming. *Math Program* 2005;104:1–20.
16. Haneveld WKK, Stougie L, van der Vlerk MH. On the convex hull of the simple integer recourse objective function. *Ann Oper Res* 1995;56:209–224.
17. van der Vlerk MH. Convex approximations for complete integer recourse models. *Math Program* 2004;99:297–310.
18. Laporte G, Louveaux FV. The integer L-shaped method for stochastic integer

- programs with complete recourse. *Oper Res Lett* 1993;13:133–142.
19. Sherali HD, Fraticelli BMP. A modification of Benders' decomposition algorithm for discrete subproblems: an approach for stochastic programs with integer recourse. *J Glob Optim* 2002; 22: 319–342.
 20. Caroe CC, Tind J. L-shaped decomposition of two-stage stochastic programs with integer recourse. *Math Program* 1998;83:451–464.
 21. Ahmed S, Tawarmalani M, Sahinidis NV. A finite branch and bound algorithm for two-stage stochastic integer programs. *Math Program* 2004;100:355–377.
 22. Kong N, Schaefer AJ, Hunsaker BK. Two-stage integer programs with stochastic right-hand sides—a superadditive dual approach. *Math Program* 2006;108:275–296.
 23. Caroe CC, Schultz R. Dual decomposition in stochastic integer programming. *Oper Res Lett* 1999;24:37–45.
 24. Guan Y, Ahmed S, Nemhauser GL. Cutting planes for multi-stage stochastic integer programs. *Oper Res* 2009;57:287–298.
 25. Stougie L, van der Vlerk MH. Stochastic integer programming. In: Dell'Amico M, editor. *Annotated bibliographies in combinatorial optimization*. New York: John Wiley & Sons, Inc.; 1997. pp. 127–141.
 26. Klein Haneveld WK, van der Vlerk MH. Stochastic integer programming: general models and algorithms. *Ann Oper Res* 1999;85:39–57.
 27. Louveaux FV, Schultz R. Stochastic integer programming. In: Ruszczyński A, Shapiro A, editors. *Handbooks in operations research and management science 10: stochastic programming*. The Netherlands: North-Holland; 2003.
 28. Schultz R, Stougie L, van der Vlerk MH. Two-stage stochastic integer programming: a survey. *Stat Neerl J Neth Soc Stat Oper Res* 1996;50(3):404–416.

TWO-STAGE STOCHASTIC PROGRAMS: INTRODUCTION AND BASIC PROPERTIES

JOHN R. BIRGE

University of Chicago Booth School of
Business, Chicago, Illinois

Practical decision making often requires a current action before the resolution of some uncertainty and possibly some future action. The mathematical formulation of this basic decision sequence is known as a *two-stage stochastic program*. Examples range across all applications of operations research methodology, including production, prices, and inventory control required before demand can be observed, investment and borrowing that must be made before returns can be observed, and products and services that are designed and sold before their performance is known. A prototypical example is the *news vendor* problem in which a news vendor must purchase a quantity of papers before knowing the demand for those papers. The vendor's future action after observing the demand is then to sell or salvage the papers that were bought earlier.

To describe this model, let x represent the number of papers initially purchased at a unit price c . Let the selling price of the papers be q , the salvage value of any left-over papers be r , the demand with realization $\xi \in \Xi$ have a distribution function F , and that the news vendor is risk-neutral and indifferent between payoffs before and after the resolution of the demand uncertainty (i.e., so that future cash flows are not discounted relative to current cash flows). A two-stage formulation of this model is

$$\min_{x \geq 0} (c - q)x + \mathbb{E}[(q - r)y(\xi)] \quad (1)$$

$$\text{subject to } x - y(\xi) \leq \xi,$$

$$y(\xi) \geq 0, \text{ a.s.}, \quad (2)$$

where $\mathbb{E}$ denotes the mathematical expectation with respect to the random demand variable ξ , a.s. indicates that the constraints hold almost surely, and $y(\xi)$, the number of unsold units, is known as the *recourse* action. In this formulation, the profits on x are counted in the first period (or stage) before the realization of demand, then removed in the second period, and replaced with the salvage value in the event of insufficient demand.

The general two-stage stochastic program maintains the form of Equation (1) with a first-stage decision $x \in \mathcal{N}^1$, followed by the realization of a random vector $\xi \in \mathcal{N}^k$ and a recourse action $y(\xi) \in \mathcal{N}^2$. The formulation also includes an explicit set of first-stage constraints given by X such that $x \in X$ and second-stage constraints given by $Y(x, \xi)$ such that $y(\xi) \in Y(x, \xi)$. The general formulation then appears as

$$\begin{aligned} \min_{x \in X} \quad & c(x) + \mathbb{E}[q(y(\xi), \xi)] \\ \text{subject to} \quad & y(\xi) \in Y(x, \xi) \text{ a.s.}, \end{aligned} \quad (3)$$

where c and q now represent general functions of the decisions. The constraints in this case are often given by an explicit system of linear and nonlinear equations and inequalities but may also include integrality restrictions and conditions that may hold with a certain probability as *probabilistic* or *chance constraints* (see **Models and Basic Properties**).

The formulation in Equation (3) is often further simplified by defining a *conditional recourse* or *second-stage value function* as

$$Q(x, \xi) = \min_{y \in Y(x, \xi)} q(y(\xi), \xi), \quad (4)$$

and an *unconditional recourse* or *expected second-stage value function* as

$$Q(x) = \mathbb{E}_\xi [Q(x, \xi)] = \int Q(x, \xi) F(d\xi). \quad (5)$$

With this definition, Equation (3) becomes

the *deterministic equivalent program*:

$$\min_{x \in X} c(x) + Q(x), \quad (6)$$

which now has the form of a deterministic optimization problem.

In some cases, such as the news vendor example, Equation (6) can be solved directly (even analytically) without using the structure inherent in Equation (3), but exploiting the structure of Equation (3) is often necessary to obtain an efficient computational procedure. The properties of solution existence, optimality conditions, and alternative equivalent formulations (such as dual programs) give these fundamental structural results that then enable various computational schemes. The most basic of these properties appear below. (See Bazaraa and Shetty [1] for more details.)

SOLUTION EXISTENCE

The main question in this case is whether there exists some $x^* \in X$ such that $c(x^*) + Q(x^*) = \min_{x \in X} c(x) + Q(x)$. This may not be the case in the familiar situation that the value in Equation (6) is unbounded below over $x \in X$ as, for example, in the news vendor when the possible demand realizations, Ξ , are unbounded and the salvage value $r = q$. In other cases, $c(x) + Q(x)$ may also only approach an infimal value asymptotically as in the case where the news vendor also earns no profit on sales ($c = q$) but pays a penalty $p > 0$ for each unit of demand that is not satisfied, that is Equation (3) is

$$\min_{x \geq 0} \mathbb{E}[py_2(\xi)] \quad (7)$$

$$\text{subject to } x - y_1(\xi) + y_2(\xi) = \xi,$$

$$y_1(\xi), y_2(\xi) \geq 0, \text{ a.s.}, \quad (8)$$

where $1 - F(\xi) > 0$ for any $\xi < \infty$. In this case, the infimal value of (7) is zero, but this is only achieved as $x \rightarrow \infty$. In this case, the infimum of $f(x) = c(x) + Q(x)$ is not obtained over X because the direction $z = +1$ is such that $x + z \in X$ for any $x \in X$ and $f(x + z) - f(x) < 0$ for any $x \in X$. In terms of convex analysis (see Rockafellar [2] and *Nonlinear Conjugate*

Gradient Methods), z is a *recession direction* of both X and f . If f is closed and proper (i.e., nowhere $-\infty$ and not identically $+\infty$ convex function), X is a nonempty, closed, convex set, and f and X do not share any common recession directions, convex analysis [[2], Theorem 27.3] gives that the minimum in Equation (3) is attained. We can state these conditions for Equation (3) as follows.

Theorem 1. *If X is a nonempty, closed convex set, c and Q are both closed proper convex functions, and (i) if X is bounded or, (ii) for every z such that $x + z \in X$ for every $x \in X$, $\exists \epsilon > 0$ and $x \in X$ such $Q(x + z) - Q(x) > \epsilon$, then the minimum in Equation (3) is attained at some $x^* \in X$.*

Since Q may be hard to find explicitly, it is often useful to be able to verify that a minimum can be attained using only the properties of c and q . In this case, it is most convenient that $Q(x) = +\infty$ only when $\text{prob}\{\xi | Q(x, \xi) < +\infty\} < 1$, since then, if $Q(\cdot, \xi)$ satisfies the condition in place of Q in Theorem 1 for each ξ , attaining a minimum can be assured. This equivalence of the feasibility (finite values) in the conditional and unconditional recourse functions is the case whenever the second moments of all the random parameters exist if q is linear ($q(y(\xi), \xi) = q(\xi)^T y(\xi)$) and Y can be written as $Y(x, \xi) = \{y \geq 0 | Wy + T(\xi)x = h(\xi)\}$ (which is known as *linear, fixed recourse* when W does not depend on ξ) [1, Proposition 3.1.2].

OPTIMALITY CONDITIONS

When an optimal solution to Equation (3) exists and is attained, it is useful to have conditions that are necessary and sufficient for optimality. For these properties, assume X , c , and Q are all convex and a regularity condition is satisfied (such as a common point in the relative interior of the effective domains of c and Q , where the *effective domain* of a function f is given as $\text{dom } f = \{x | f(x) < \infty\}$), the basic optimality conditions from convex analysis [see Bazaraa and Shetty [3] (Theorem 3.4.3)] can then apply so that $x^* \in X$

is an optimal solution if and only if

$$0 \in \partial c(x^*) + \partial Q(x^*) + N_X(x^*), \quad (9)$$

where $\partial f(x^*) = \{v | f(x) + v^T(z - x) \leq f(z), \forall z\}$, the subdifferential set of all subgradients of f at x , and $N_X(x^*) = \{v | v^T(z - x^*) \leq 0, \forall z \in X\}$, the normal cone to X at x^* . When c and Q are differentiable and $X = \{x | g(x) \leq 0\}$ for some $g : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{m_1}$, where $g = [g_1, \dots, g_{m_1}]$ such that each $g_i, i = 1, \dots, m_1$ is convex and differentiable (and whose effective domain shares a common relative interior point with all other functions), then Equation (9) reduces to the standard Karush–Kuhn–Tucker conditions and there exists multipliers $0 \leq \lambda^* \in \mathbb{R}^{m_1}$ such that

$$\nabla c(x^*) + \nabla Q(x^*) + \sum_{i=1}^{m_1} \lambda_i^* \nabla g_i(x^*) = 0. \quad (10)$$

Although the recourse function $Q(x)$ is generally differentiable (see the result below) when ξ is continuously distributed, implementation often requires a finite representation of ξ ; so, the general representation in Equation (9) is more relevant to two-stage stochastic programs than the standard result in Equation (10). Most often, $Q(x)$ is not differentiable even when $q(y, \xi)$ is differentiable in y for every ξ . The potential for difficulty is that $Q(x, \xi)$ is usually not differentiable in x , although, in many cases, $Q(x, \xi)$ has useful properties for computation. The following [2, Theorems 3.5, 3.6, and 3.11] summarizes these properties for the unconditional and expected recourse functions, Q and $\bar{Q}$, for the case of linear fixed recourse, where ξ is interpreted as the vector of all random components in q , T , and h .

Theorem 2. *If $q(y(\xi), \xi) = q(\xi)^T y(\xi)$, $Y(x, \xi) = \{y \geq 0 | Wy + T(\xi)x = h(\xi)\}$, $Q(x, \xi) > -\infty$ for any x and ξ , and ξ has finite second moments, then*

1. $Q(x, \cdot)$ is piecewise-linear convex as a function of (h, T) for any $x \in X$;
2. $Q(x, \cdot)$ is piecewise-linear concave as a function of q for any $x \in X$;

3. $Q(x, \xi)$ is piecewise-linear convex as a function of x for any $\xi \in \Xi$;
4. if ξ has a finite representation, $Q(x)$ is piecewise-linear and convex;
5. $Q(x) < \infty$ whenever $Q(x, \xi) < \infty$ almost surely;
6. $Q(x)$ is convex in x and Lipschitzian; that is, there exists $K < \infty$ such that $|Q(x) - Q(z)| \leq K\|x - z\|$ for any x and z in $\text{dom } Q$;
7. if $F(\xi)$ is absolutely continuous, $Q(x)$ is differentiable on the interior of $\text{dom } Q$;
8. if $x \in X$, $\partial Q(x) = \mathbb{E}[\partial Q(x, \xi)] + N_{\text{dom } Q}(x)$.

These properties motivate the use of methods based on piecewise-linear convex approximations of Q (see the discussion in Birge and Louveaux [1]) that can also be approximated by sampling ξ [4] (see **Sampling Methods**).

DUAL AND NONANTICIPATIVE FORMULATIONS

These properties also lead to representations of dual programs to Equation (3). For these problems, we can write the Lagrangian dual function (see **Lagrangian Optimization Methods for Nonlinear Programming**) when $X = \{x | g(x) \leq 0\}$ as

$$l(\lambda) = \inf_x c(x) + Q(x) + \lambda^T g(x), \quad (11)$$

to form a dual program:

$$\max_{\lambda \geq 0} l(\lambda). \quad (12)$$

In the fully linear case where $c(x) = c^T x$, $g_i(x) = -a_i^T x + b_i$, where $A = [a_1 \dots a_{m_1}]^T$, $q(y(\xi), \xi) = q(\xi)^T y(\xi)$, and $Y(x, \xi) = \{y \geq 0 | Wy + T(\xi)x = h(\xi)\}$, Result 8 in Theorem 2 can be used by noting that $\partial Q(x, \xi) = \{-\pi(\xi)^T T(\xi)\}$ such that $\pi(\xi)$ maximizes $\pi(\xi)^T (h(\xi) - T(\xi)x)$ over $\pi(\xi)^T W \leq q(\xi)^T$ and by avoiding complications from the normal cone term in the expression for $\partial Q(x)$. This additional term occurs because some elements in $\partial Q(x)$ can be singular multipliers that cannot be expressed as expectations of

multipliers from $\partial Q(x, \xi)$. This potentially unstable situation can be avoided if the problem has a property called *relatively complete recourse* that $Q(x) < \infty$ for all $x \in X$ (or the more restrictive property of *complete recourse* that $Q(x) < \infty$ for any $x \in \mathbb{R}^n$). With the relatively complete recourse property or a finite number of realizations, the Lagrangian dual function in the fixed linear recourse case can be written as

$$l(\lambda) = \begin{cases} E[\pi(\xi)^T h(\xi)] + \lambda^T b & \text{if } c^T - E[\pi(\xi)^T T(\xi)] \\ & -\lambda^T A = 0, \\ & \pi(\xi)^T W \leq q(\xi)^T \text{ a.s.}, \\ -\infty & \text{otherwise,} \end{cases} \quad (13)$$

yielding the dual program:

$$\max b^T \lambda + E[h(\xi)^T \pi(\xi)] \quad (14)$$

$$A^T \lambda + E[T(\xi)^T \pi(\xi)] = c, \quad (15)$$

$$W^T \pi(\xi) \leq q(\xi)^T \text{ a.s.}, \quad (16)$$

$$\lambda \geq 0, \quad (17)$$

which can provide a more efficient structure for computation (particularly for interior point methods (see **Central Path and Barrier Algorithms for Linear Optimization**) as discussed in Birge and Louveaux [[1], Chapter 5]). Nonlinear versions of the dual in Equation (14) can also be useful computationally and constructed using standard convex duality arguments with attention again paid to the possible presence of singular multipliers that can be avoided when relatively complete recourse can be verified.

The major computational difficulty for stochastic programs stems from the evaluation of $Q(x)$ and discovery of potential directions of improvement. Solution methods focus on simplification and approximation of $Q(x)$, using convexity and the subdifferential structure, and on repeated structure and possible similarity among the $Q(x, \xi)$ subproblems. The computational advantages of this aspect

of the two-stage stochastic program are that the Jacobian matrix of the constraints and the Hessian matrix of the objective are sparse (since they do not involve terms directly linking $y(\xi_i)$ and $y(\xi_j)$ when $\xi_i \neq \xi_j$) and often have similar components. For direct optimization, the linear algebra can then be more efficient particularly if solutions of the linear systems recognize and exploit the structure. This observation can also lead to decomposition methods that iterate between solutions with x and $y(\xi)$, with much of the potential effort on $y(\xi)$ solutions performed in parallel. This observation also raises the possibility of decomposing the problem further by allowing x to depend on ξ as $x(\xi)$ but by then adding an explicit constraint, called *nonanticipativity* that can take the form:

$$x(\xi) - E[x(\xi)] = 0, \text{ a.s.} \quad (18)$$

The constraint (18) then becomes the only link between distinct realizations of ξ and can allow for further parallelism by relaxing Equation (18) and placing it into the objective as in Lagrangian relaxation.

The basic properties above also hold for problems with multiple stages where the action–observation–action sequence is repeated many times. These problems present additional computational challenges but the emphasis on identifying convexity, repeated structure, natural sparsity, and finding subgradients can again lead to computational efficiencies.

REFERENCES

1. Birge JR, Louveaux F. Introduction to stochastic programming. New York: Springer; 1997.
2. Rockafellar RT. Convex analysis. Princeton, NJ: Princeton University Press; 1970.
3. Bazaraa MS, Shetty CM. Nonlinear programming: Theory and algorithms. New York: John Wiley & Sons; 1979.
4. Higle J, Sen S. Stochastic decomposition: A statistical method for large-scale stochastic linear programming. Boston: Kluwer; 1996.

UNCERTAINTY IN FOREST PRODUCTION PLANNING

ANTONIO ALONSO-AYUSO

LAUREANO F. ESCUDERO

Departamento de Estadística e Investigación
Operativa, Universidad Rey Juan Carlos,
Madrid, Spain

MONIQUE GUIGNARD

Department of Operations and Information
Management, The Wharton School,
University of Pennsylvania, Philadelphia,
Pennsylvania

MARTIN QUINTEROS

Independent Consultant, Santiago, Chile

ANDRES WEINTRAUB

Department of Industrial Engineering,
University of Chile, Santiago, Chile

INTRODUCTION

Using OR models in forest planning has a long tradition. In the 1970s, the US Forest Service started using linear programming extensively for long-range planning of its forests. In that model, Timber RAM [1], decisions were made on when to harvest areas and then to replant them. Also partial harvest and other silvicultural treatments to improve tree growth were considered. Horizons considered were of almost 200 years in duration, so as to include over two tree rotations, using 10-year periods which were aggregated into longer periods as the horizon approached. In these decisions, aggregate characteristics of both, spatial areas and timber species and types were considered. This made sense, given the long range of the plans as well as the general criteria underlying the decisions. Since then, the use of OR tools has been expanded very successfully to different planning and scheduling problems; for instance Epstein *et al.* [2], Martell *et al.* [3], Weintraub *et al.* [4], and in the article titled *Operations Research in Forestry and*

Forest Products Industry. In the 1980s, tactical, medium-range planning started being implemented. Decisions involved defining specific spatial characteristics. In tactical planning, models define specific cutting units to harvest in each period. Decisions are considered only until existing trees are harvested. In this form, for native forests with growth periods of about 80 years, typical horizons are of about 40–50 years. In plantations with growth periods of about 20 years, typical horizons are of about 5 years. At this level, road building, which in native forests leads to about 40% of operating costs, is often included explicitly. Also in the 1980s, concern for the environment grew. Planning models explicitly incorporated these concerns. The most familiar forms expressing habitat protection are the well-known adjacency constraints, which indicate that when one cutting area is harvested (about 40 ha) no neighboring areas can be harvested until the recently harvested area grows for long enough, usually one or two periods. These constraints have been implemented by legislation in North America and Europe, so models to support these decisions incorporate spatial constraints. Since including adjacency or other spatial constraints leads to problems that are combinatorially complex, mostly heuristic algorithms have been implemented (Monte Carlo simulation and, lately, Tabu search or simulated annealing; [5]). In addition, exact algorithms have been developed which have solved moderate size problems [6]. In the late 1980s and 1990s, operational decisions started being implemented in forest firms. These decisions involve short-term practices related to harvesting and transporting logs to destinations such as saw mills, pulp plants or ports for export. One example is the forest firms in Chile, which in the early 1990s successfully implemented several systems for operational purposes [7]. One system supported harvesting decisions on a 45-days horizon, given known demands for logs of different sizes and trees available in different cutting units.

Another model supported decisions on location of harvesting machinery (skidders for flat terrain and cable logging for steep areas) and secondary roads to access the machinery locations. Finally, a transportation model was implemented for daily decisions on scheduling trucks to transport logs from different origins in the forests to the destinations (sawmills, pulp plant, ports, stocking areas). In the last two decades, multiple proposals for operational decisions in other countries such as Sweden, New Zealand, and Canada have appeared in the literature [8,9]. The issue of uncertainty has appeared in the literature mainly since the 1980s. There are varied causes for uncertainty:

- **Market Conditions.** Future prices of different timber products are difficult to predict. Also, the level of demands for a given firm can cause uncertainty. Given the high incidence of transportation costs within the final total cost, the location of a given firm and its main competitors in relation to specific markets is important. So, the behavior of competitors is another source of uncertainty. Note that market uncertainty is important in the long term, in relation to investments; in the medium term, in relation to tactical planning and delivery contracts. In the short term, uncertainty does not play such an important role in market conditions.
- **Growth and Silvicultural Conditions.** There exist sophisticated statistical models to predict future long-term growth of trees under different management conditions. But naturally, there are errors and uncertainties related to this growth. Climate, for example, is one cause of growth uncertainty.
- **Fires and Pests.** In addition to the silvicultural uncertainties for growth, added uncertainties are losses due to fires and pests. In North American native forest for example fires are a significant source of loss of timber, see Martell [10].
- **Short-Term Operational Uncertainties.** These uncertainties are of lesser importance, but need to be analyzed. Simulation models, based on sample lots,

try to predict how much timber will be extracted from a given cutting unit if harvested now. However, there are errors in these estimations. There are also uncertainties for example, in travel times when scheduling truck trips.

The rest of the article is organized as follows: In the next section we discuss three different aspects of uncertainty, and how they have been dealt with. In the section titled "Planning Forest Harvest and Access Road Construction under Uncertainty," we present a specific problem involving tactical planning with market uncertainty. The section titled "Stochastic Programming Approach" discusses these different aspects of uncertainty, and how they have been dealt with. In the section titled "Mixed 0–1 Model" we present one specific case of uncertainty in tactical planning, when the cause of uncertainty is in the market conditions. The next section deals with a case study and we present the conclusions in the final section.

FOREST UNCERTAINTY MODELS

In this section, we discuss the specific aspects of uncertainty and the models and algorithms that have been proposed in the literature. When uncertainty is present, planners and schedulers look for good robust and flexible solutions, which will allow for adequate recourse decisions and lead to good solutions as reality becomes known as time goes by. We will discuss the problems that have been presented in the introductory section, though naturally all aspects of uncertainty are present simultaneously in a given planning time horizon.

Market Uncertainty

Uncertainty related to markets is a major issue in planning. We can view uncertainty as long range, related to major investments, or medium range for tactical planning and operational uncertainty. Long-range uncertainty is vital when considering investments. Long-range market uncertainties reflect demand for saw timber, paper, panels, and other wood products. The investments can be in

large plantations, with horizons of typically 20 years for pine plantations for example, and investment in plants, again with long-range horizons (a pulp plant takes 5 years from planning stage to operation and costs over USD 500 million). OR models have not been really been proposed to handle long-range uncertainty. Typically, scenario analysis has proved to be a tool most used to handle uncertainty, based on long-range harvesting models. Long-range market uncertainty has been approached through real options theory. Insley [11] and Insley and Rollins [12] modeled single and multirotation stochastic optimal harvest problems as call options using a finite difference scheme. Chladna [13] includes carbon emissions permits as revenues in addition to timber. Ibañez and Zapatero [14] solve the stochastic optimal harvest problem via Monte Carlo methodology.

In the case of medium-range investments, explicit uncertainty models have been developed. As was defined, tactical models deal with shorter time periods, where spatial decisions are specified exactly. In the next section we present one case study of a medium-range harvesting and road building problem, where future market conditions present uncertainty. Adaptive optimization of forests is presented [15], see also a description of case studies. In this approach, the decision process can be viewed as a controllable Markov process. Main uncertainties are in market scenarios as well as in silvicultural and managerial issues. Stochastic dynamic programming is the basic tool for this approach.

Short-term uncertainty is present in operational decisions. In relation to markets, there is a need to determine the volumes of different products to be harvested, decisions that will depend on market prices. Uncertainties in relation to markets relate to unexpected demands and sudden variations in prices; in particular, price rises that make it very advantageous to produce additional volumes quickly. The approach in most cases has been to allow for flexibility in the decisions to be made. No specific stochastic approaches have been used in the short-term decisions.

Silvicultural and Growth Uncertainties

Again, we have long-, medium-, and short-term uncertainties since there is natural variability in tree growth. Chance-constrained programs were proposed [16,17], to handle this variability. Typical constraints related to timber production are given as a minimum and maximum production for a given period (usually derived from market demand expectations), or a maximum variation of total volume produced between periods (in general, there are important advantages to having homogeneous production along time). These constraints become probabilistic ones; for example, the probability of satisfying each constraint must be at least 95%. Assuming growth can be represented by a normal probability, these probabilistic constraints yield nonlinear deterministic models. While these models can be solved well for moderate size problems, we are not aware of any actual decision making application. Adaptive management plays an important role in real planning [18]. Forests allow for larger flexibility in adapting to future variability than other fields, such as manufacturing. This is due to the nature of tree growth. Planners may have one plan, but then when real growth becomes known, it is usually not difficult to suitably modify harvest plans. In [19], an example was shown considering two planners: one considers explicitly uncertainty (chance constrained programming) and the other one, a deterministic planner, uses expected values. It was shown in case studies that there was a significant difference in favor of the “stochastic” planner when demand was close to supply possibilities. Short-term main components of uncertainty include data errors in the availability of timber in the areas to be harvested. Simulators based on sample plots have been successful in determining the total production of timber, but not so effective in predicting the amount of timber of different lengths and diameters to be obtained once trees are bucked or cut into pieces. No models are known to explicitly include uncertainty into harvesting models. Errors in predicted amounts are difficult to measure, and actual decision making is basically adaptive and takes advantage of the

flexibility of forests (harvesting a different unit, use different bucking schemes, etc).

Fires and Pests

Martell [10] describes different decision processes related to forest fires. The main problems discussed include the initial attack resource deployment of airtankers and crews of firefighters once a fire has been detected. Proposals have included chance-constrained programming and queueing models to allocate helicopters. Large fire management, which occurs when fires cannot be controlled in the initial attack and escape, are another case of good use of OR tools. In this case, there is important uncertainty in relation to how the fire will spread; this has a stochastic nature. A few models, including stochastic programming-based models, have been developed to support decisions to contain fires but have not been widely used. Long-range decisions involve investments in firefighters and crews and also harvesting patterns to help contain fires. Notice that harvesting areas that are likely to be in the path of large fires converging on valuable land can act as useful firebreaks [20].

Operational Uncertainties

Short-term uncertainty is present in operational decisions. These decisions involve which areas to harvest in the near future (one week, one month, 4 months), how to handle the harvesting machinery, fleet planning, and daily truck scheduling. There are obvious uncertainties here, such as possible machinery or truck failures, travel times for trucks, and so on. In general, these uncertainties have not been incorporated explicitly into models, given the complexities of these models even in the deterministic versions. Robust solutions are usually determined by allowing some expected slack, such as adding a small percentage to expected travel time. Or, in case of daily transportation planning, allowing for a higher number of trucks than the expected need, which are less expensive than the loader machineries (to load logs into trucks), so that the less costly trucks seldom constitute a bottleneck.

PLANNING FOREST HARVEST AND ACCESS ROAD CONSTRUCTION UNDER UNCERTAINTY

The advantages of integrating the processes of planning forest harvest and access road construction in a single mixed 0–1 model were demonstrated [21] by obtaining solutions from 15% to 45% better than with models optimizing the processes separately; see also work by Jones *et al.* [22]. Various studies exist on the different phases of forestry planning [23,24]; for example, the problem of access road construction and timber harvest policies [25]. In essence, the problem can be formulated on the basis of a division of the forest into harvest units that we shall call “cells.” For a chosen time horizon, one must determine for each time period, which cells will be cut, which roads need to be constructed for accessing those cells, and the quantity of wood to be transported from one point to another, to maximize the expected net return under uncertainty in wood selling prices.

Approaches to solving the deterministic (one simple scenario) problem can be found [26], where model strengthening schemes and decomposition techniques such as Lagrangean relaxation are employed to obtain very good solutions in reasonable computation times with low residual gaps. See how deterministic machinery location and road design problems are solved in forestry management [27]. However, the stochastic version of the problem can be formulated as a large-scale structured mixed 0–1 linear program. As such, it is difficult to solve due to its size and the presence of several thousands of 0–1 variables.

In the following, the model for planning forest harvest and access road construction under uncertainty as proposed by Alonso-Ayuso *et al.* [28] is presented. It is a simplified version of the one actually used for planning [26]. The simplification keeps the main elements of the problem and makes it more convenient for highlighting the stochastic nature introduced in the model.

Let us consider a forest divided into a set $\mathcal{H}$ of harvest units or cells. There is a set $\mathcal{O}$ of origins where wood is processed and a set $\mathcal{S}$

of wood exits. Each origin has an associated set of cells $\mathcal{H}_o$. These origins and exits are connected through a set of existing roads, say $\mathcal{R}^E$, and there is the possibility of building several potential roads, say $\mathcal{R}^P$. Other points to be considered are the intersection between roads, $\mathcal{J}$. Note that these elements define a graph where the nodes are $\mathcal{O} \cup \mathcal{S} \cup \mathcal{J}$ and the arcs are $\mathcal{R}^E \cup \mathcal{R}^P$.

For each cell h in $\mathcal{H}$, the following parameters are given:

- a_h^t , productivity of cell h , if it is harvested in period t [m^3/Ha];
- A_h , area of the cell h to be harvested [ha];
- P_h^t , harvesting cost of one hectare of cell h in time period t [dollars/ha].

For each origin o , Q_o^t denotes the unit production cost in time period t [dollars/ m^3]. For each arc (k, l) and time period t , we know the flow capacity, $U_{k,l}^t$ [m^3], and the unit transport cost, $D_{k,l}^t$ [dollars/ m^3]. The construction cost of a potential road in arc (k, l) at time period t [dollars], say $C_{k,l}^t$, is also given.

The uncertainty in the model is due to the wood prices and demand, since these parameters have historically experienced significant fluctuations, as exemplified for pulpwood in Chile between 1994 and 2003, shown here in Fig. 1. This kind of uncertainty can be modeled by a set Ω of scenarios represented in a tree. Figure 2 shows a scenario tree of timber sale price and demand. A *scenario* is a particular realization of uncertainty through the whole time horizon and it is represented by a path from the root to a leaf in the tree. Each scenario has a weight w^ω , that represents the probability that the modeler assigns for this scenario to happen ($\sum_{\omega \in \Omega} w^\omega = 1$).

Each node in the scenario tree defines a group Ω^g of scenarios that are identical up to the stage associated to this node, and, $\mathcal{G}^t$ is the set of scenario groups (nodes) at stage t . (See **Modeling Uncertainty in Optimization Problems** for an introduction to the subject.)

The following parameters for each scenario ω are considered in the model:

- $R_s^{t,\omega}$, sale price at exit s in time period t under scenario ω [dollars/ m^3].
- $\underline{Z}^{t,\omega}, \bar{Z}^{t,\omega}$, lower and upper demand bounds in period t under scenario ω [m^3], respectively.

Stochastic Programming Approach

A key idea in stochastic programming is the nonanticipativity principle [29,30], according to which, if two different scenarios are identical up to a given period in the time horizon, the values of the decision variables must be identical up to that period. That is, the decisions for all scenarios in the same group g are identical up to the stage in which this scenario group is defined.

Consider the following deterministic model:

$$\begin{aligned} \max \quad & cx + ay \\ \text{s.t.} \quad & Ax + By = b \\ & x \in \{0, 1\}^n, y \geq 0, \end{aligned} \quad (1)$$

where c and a are the coefficient vectors of the objective function, b is the right-hand side vector, A and B are the constraint matrices, x and y are the 0–1 and continuous variables, respectively, and n is the number of 0–1 variables. The model will be extended

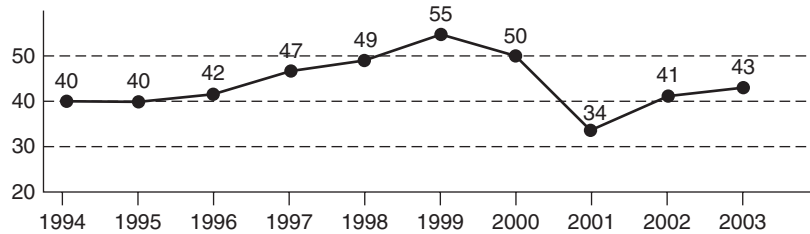


Figure 1. Annual price of pulpwood [US\$/ m^3], 1994–2003.

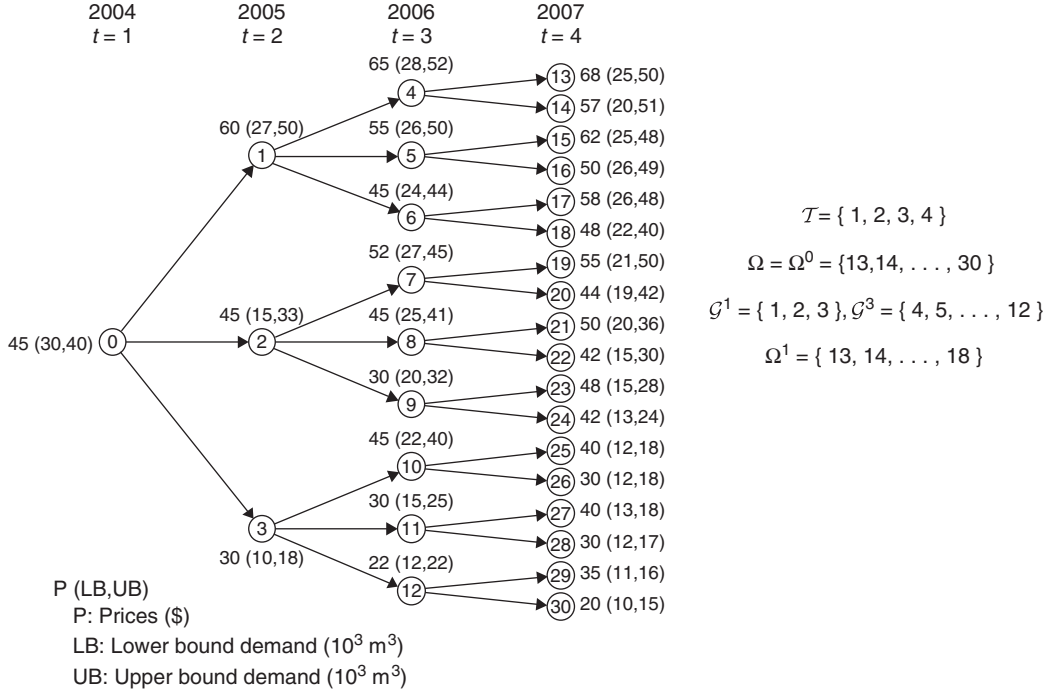


Figure 2. Scenario tree with 18 scenarios for Los Copihues case.

to incorporate the uncertainty of some of its parameters. We can now formulate the structured mixed 0–1 deterministic equivalent model (DEM) of the stochastic version of problem (1) to maximize the expected value of the objective function as follows:

$$\max Q_E = \sum_{\omega \in \Omega} w^\omega (c^\omega x^\omega + a^\omega y^\omega) \quad (2)$$

$$\text{s.t. } Ax^\omega + By^\omega = b^\omega, \quad \forall \omega \in \Omega; \quad (3)$$

$$x_t^\omega = x_t^{\omega'}, \quad \forall t \in \mathcal{T}, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g; \quad (4)$$

$$y_t^\omega = y_t^{\omega'}, \quad \forall t \in \mathcal{T}, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g; \quad (5)$$

$$x^\omega \in \{0, 1\}^n, y^\omega \geq 0, \quad \forall \omega \in \Omega. \quad (6)$$

We can observe that in the above model, the only constraints linking the various scenarios are the nonanticipativity constraints (4) and (5). It is unlikely (and we

verified this in practice) that a commercial optimization software package will be able to solve a large-scale real-life problem in a satisfactory manner. An efficient method for solving the $|\Omega|$ subproblems that arise when the constraints (4) and (5) are relaxed from the model but taken into account by the algorithm, is the branch-and-fix coordination (BFC) approach [31,32]. (See *Two-Stage Stochastic Integer Programming: A Brief Introduction* and *Stochastic Mixed-Integer Programming Algorithms: Beyond Benders' Decomposition*). The performance of the stochastic approach will be compared to the solution obtained from the average scenario.

Mixed 0–1 Model

The objective is to determine the optimal harvest and access road construction policy that will maximize expected net profit that is defined as the difference between the incomes (wood sale revenue) and the costs (wood harvest cost, production costs,

road construction cost, and wood transport cost) and satisfy the constraints for all scenarios. The decision to be taken for each scenario in the scenario tree and each time period, are the cells to be harvested, the roads to be built, the flow of good to be transported through the roads, and the demand of timber to be satisfied at each exit.

Formulation

$$\max Z = \sum_{\omega \in \Omega} \left(\sum_{s \in S} \sum_{t \in T} w^\omega R_s^{\omega,t} z_s^{t,\omega} - \sum_{h \in \mathcal{H}} \sum_{t \in T} w^\omega P_h^t A_h \delta_h^{t,\omega} - \sum_{o \in \mathcal{O}} \sum_{t \in T} w^\omega Q_o^t \right. \\ \left. \times \left(\sum_{h \in \mathcal{H}_o} a_h^t A_h \delta_h^{\omega,t} \right) - \sum_{(k,l) \in \mathcal{R}^P} \sum_{t \in T} w^\omega C_{k,l}^t \gamma_{k,l}^{t,\omega} - \sum_{(k,l) \in \mathcal{K}} \sum_{t \in T} w^\omega D_{k,l}^t f_{k,l}^{t,\omega} \right) \quad (7)$$

$$\sum_{h \in \mathcal{H}_o} a_h^t A_h \delta_h^{t,\omega} + \sum_{(k,o) \in \mathcal{K}} f_{k,o}^{t,\omega} - \sum_{(o,k) \in \mathcal{K}} f_{o,k}^{t,\omega} = 0, \quad \forall o \in \mathcal{O}, t \in T, \omega \in \Omega \quad (8)$$

$$\sum_{(k,j) \in \mathcal{K}} f_{k,j}^{t,\omega} - \sum_{(j,k) \in \mathcal{K}} f_{j,k}^{t,\omega} = 0, \quad \forall j \in \mathcal{J}, t \in T, \omega \in \Omega \quad (9)$$

$$z_s^{t,\omega} = \sum_{(k,s) \in \mathcal{K}} f_{k,s}^{t,\omega}, \quad \forall s \in S, t \in T, \omega \in \Omega \quad (10)$$

$$\underline{Z}^{t,\omega} \leq \sum_{s \in S} z_s^{t,\omega} \leq \bar{Z}^{t,\omega}, \quad \forall t \in T, \omega \in \Omega \quad (11)$$

$$f_{k,l}^{t,\omega} \leq U_{k,l}^t \sum_{1 \leq \tau \leq t} \gamma_{k,l}^{\tau,\omega}, \quad \forall (k,l) \in \mathcal{R}^P, t \in T, \omega \in \Omega \quad (12)$$

$$f_{k,l}^{t,\omega} \leq U_{k,l}^t, \quad \forall (k,l) \in \mathcal{R}^E, t \in T, \omega \in \Omega \quad (13)$$

$$\sum_{t \in T} \gamma_{k,l}^{t,\omega} \leq 1, \quad \forall (k,l) \in \mathcal{R}^P, \omega \in \Omega \quad (14)$$

$$\sum_{t \in T} \delta_h^{t,\omega} \leq 1, \quad \forall h \in \mathcal{H}, \omega \in \Omega \quad (15)$$

$$\delta_h^{t,\omega} = \delta_h^{t,\omega'} \quad \forall h \in \mathcal{H}, t \in T, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g \quad (16)$$

$$\gamma_{k,l}^{t,\omega} = \gamma_{k,l}^{t,\omega'} \quad \forall (k,l) \in \mathcal{R}^P, t \in T, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g \quad (17)$$

$$f_{k,l}^{t,\omega} = f_{k,l}^{t,\omega'} \quad \forall (k,l) \in \mathcal{R}^P, t \in T, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g \quad (18)$$

$$z_s^{t,\omega} = z_s^{t,\omega'} \quad \forall s \in S, t \in T, g \in \mathcal{G}^t, \omega, \omega' \in \Omega^g \quad (19)$$

$$f_{k,l}^{t,\omega}, z_s^{t,\omega} \geq 0, \quad \forall s \in S, (k,l) \in \mathcal{K}, t \in T, \omega \in \Omega \quad (20)$$

$$\delta_h^{t,\omega}, \gamma_{k,l}^{t,\omega} \in \{0, 1\}, \quad \forall h \in \mathcal{H}, (k,l) \in \mathcal{R}^P, t \in T, \omega \in \Omega \quad (21)$$

Variables

$\delta_h^{t,\omega} = 1$, if cell h is harvested in period t under scenario ω , and 0, otherwise.

$\gamma_{k,l}^{t,\omega} = 1$, if road in arc (k,l) is built in period t under scenario ω , and 0 otherwise.

$f_{k,l}^{t,\omega}$, flow of wood transported through arc (k,l) in period t under scenario ω [m^3].

$z_s^{t,\omega}$, demand of timber at exit s in period t under scenario ω [m^3].

The objective function of Equation (7) includes five terms representing the wood sale revenue, wood harvest cost, production costs at origin nodes, potential road construction cost, and wood transport cost. Constraints (8)–(10) are the flow balance equations at the origin nodes, intersection nodes and destination nodes, respectively. Constraint (11) is the wood demand bounds. Constraints (12) and (13) impose the flow capacity for the potential and existing roads, respectively. Constraint (14) ensures that each potential road is built at most once in the time horizon. Constraint (15) guarantees that a cell is harvested at most once in the time horizon. Constraints (16)–(17) and (18)–(19) impose the nonanticipativity principle for the 0–1 and continuous variables, respectively. Constraints (20) and (21) are the nonnegativity and integrality conditions for the variables.

Case Study

The forest property called Los Copihues, owned by the Chilean company Sociedad Forestal Millalemu has an area of 300 ha, divided into $|\mathcal{H}| = 25$ harvest cells based on company maps, with $|\mathcal{O}| = 9$ origins, $|\mathcal{J}| = 3$ junction nodes, $|\mathcal{S}| = 1$ exit node, $|\mathcal{R}^E| = 6$ existing roads, and $|\mathcal{R}^P| = 14$ potential roads. The planning horizon is divided into four time periods associated with years 2004, 2005, 2006, and 2007, respectively. The uncertainty of the problem is due to the unknown demand and timber sales prices. This uncertainty is represented in a tree with 18 scenarios as depicted in Fig. 2. For each node in the tree, the timber sales prices (\$) and the lower and upper bounds of the demand (10^3 m^3) are given.

The numerical experiments involve making comparisons between the performance of a solution adopted by a stochastic planner and a solution applied by a planner guided only by the average values of the uncertain parameters. In this experiment, it is assumed that the uncertain parameters evolve equiprobably. Given the large scale of the problem, we use the BFC approach in the search for the optimal solution. The results of the comparison are summarized in Table 1 for all scenarios. The average scenario-based

Table 1. Comparison of the Results

	Z_{AVSC}	Z_{BFC}	GAP %
SC 1	7860376.2	8141684.5	3.6
SC 2	74986706.4	7832291.9	4.5
SC 3	7272765.9	7681257.3	5.6
SC 4	6876035.1	7248863.7	5.4
SC 5	6751277.3	7288986.5	8.0
SC 6	6420668.3	6913420.5	7.7
SC 7	6440966.1	6739744.0	4.6
SC 8	6077296.2	6359584.0	4.6
SC 9	6003190.1	6111671.1	1.8
SC 10	Infeasible	5604078.0	—
SC 11	Infeasible	4945591.0	—
SC 12	Infeasible	4541990.9	—
SC 13	Infeasible	4324647.4	—
SC 14	Infeasible	4149814.2	—
SC 15	Infeasible	3335188.5	—
SC 16	Infeasible	3067968.2	—
SC 17	Infeasible	2866035.9	—
SC 18	Infeasible	2593300.9	—
tt	—	11942	—
Z_{IP}	—	5541451.0	—

solutions were obtained by simulating what happens in a given scenario, say, $\omega \in \Omega$, when applying the average scenario solution (AVSC). The headings are as follows: Z_{AVSC} , solution value of the average scenario approach for each scenario; Z_{BFC} , solution value of the BFC approach for each scenario; optimality GAP = $(Z_{\text{BFC}} - Z_{\text{AVSC}})/Z_{\text{AVSC}}$ (%); tt , total elapsed time (sec) for the solution value in all scenarios; Z_{IP} , expected solution value of the integer programming (IP) for the whole set of scenarios.

Notice that for the scenarios 10 through 18, the Z_{AVSC} column shows “infeasible.” This means that when the AVSC solution vector is substituted into these scenarios, a demand constraint is violated, thus rendering the solution mathematically infeasible.

It is of interest to analyze the possible significance of the previously discussed analysis to a forest manager. As was seen, the BFC approach led to better solutions than the deterministic approach under most scenarios, often by significant margins. But what was more significant, the deterministic approach could not find feasible solutions in multiple scenarios for all cases. The reason for these results is that in the deterministic case, where average expectations of market

conditions are taken, the possible fluctuations of future market conditions are not considered. This leads in multiple cases to decisions in the first periods that lead to conditions in future periods in relation to timber availability; this makes it difficult to react well under specific scenarios. In contrast, the BFC approach looks ahead at possible scenarios and protects itself against possible fluctuations. A forest manager might well want to consider these future possible fluctuations and protect his/her firm against them.

It is important to point out that a real planner can use the stochastic solution in a rolling time horizon, by using the decisions at stage 0 and re-running the model pushing it forward by one period after entering the realization of uncertain parameters each time. The decision to be made at the stage in a rolling horizon are the cells to be harvested, roads to be built, flow of goods to be transported through the roads, and demand of timber to be satisfied at each exit.

CONCLUSIONS

We have discussed the different forms in which uncertainty has been introduced into forest planning. We note that in general, forests have significant flexibility for harvesting, given the multiple options a planner has to harvest in a given moment. This is derived through selecting units to harvest as well as cutting patterns chosen to deliver different products. However, there are many situations where not considering uncertainty in an explicit form, working just with expected values may lead to significant losses in future periods. This happens when there is not much slack between supply and demand and decisions made early lock the planning process into patterns not easily reversible. We present one specific approach, BFC, to show how uncertainty can be expressed as a given set of scenarios of future prices. The nonanticipativity constraint is implicitly satisfied in the branching process. Computational results with a real forest data from the Chilean forest property Los Copihues show the advantages of expressing uncertainty explicitly through this approach.

Acknowledgments

This work is partially supported by the Spanish Government through the Ministerio de Ciencia e Innovación (MICINN) and the Madrid Regional Government (CAM), Fondecyt project N.1085188, and Project Millennium Complex Engineering Systems.

REFERENCES

1. Navon D. Timber RAM a long range planning method for commercial timber lands under multiple—use management. USDA Forest Service. Paper PSW-70; 1971.
2. Epstein R, Morales R, Serón J, *et al.* Use of OR systems in the Chilean forest industries. *Interfaces* 1999;29:7–29.
3. Martell DL, Gunn EA, Weintraub A. Forest management challenges for operational researchers. *Eur J Oper Res* 1998;104:1–17.
4. Weintraub A, Romero C, Bjørndal T, *et al.*, editors. Handbook of operations research in natural resources, International Series in Operations Research and Management Science. New York: Springer; 2007.
5. Murray AT. Spatial environmental concerns. In: Weintraub A, Romero C, Bjørndal T, *et al.*, editors. Handbook of operations research in natural resources. New York: Springer; 2007. pp. 419–430.
6. Goycoolea M, Murray A, Vielma JP, *et al.* Evaluating approaches for solving the area restriction model in harvest scheduling. *Forest Sci* 2007;55:149–165.
7. Epstein R, Chevalier P, Gabarró J, *et al.* A system for short term harvesting. *Eur J Oper Res* 1999;119:427–439.
8. Epstein R, Karlsson J, Rönnqvist M, *et al.* Harvest operational models in forestry. In: Weintraub A, Romero C, Bjørndal T, *et al.*, editors. Handbook of operations research in natural resources. New York: Springer; 2007. pp. 265–377.
9. Epstein R, Rönnqvist M, Weintraub A, *et al.* Forest transportation. In: Weintraub A, Romero C, Bjørndal T, *et al.*, editors. Handbook of operations research in natural resources. New York: Springer; 2007. pp. 391–414.
10. Martell DL. Forest fire management. In: Weintraub A, Romero C, Bjørndal T, *et al.*, editors. Handbook of operations research in natural resources. New York: Springer; 2007. pp. 489–510.

11. Insley M. A real options approach of the valuation of a forestry investment. *J Environ Econ Manage* 2002;44:471–492.
12. Insley M, Rollins K. On solving the multi-rotational timber harvesting problem with stochastic prices: a linear complementarity formulation. *J Agric Econ* 2005;87:735–755.
13. Chladna Z. Determination of optimal rotation period under stochastic wood and carbon prices. *Forest Policy Econ* 2007;9:1031–1045.
14. Ibañez A, Zapatero F. MonteCarlo valuation of American options through computation of the optimal exercise frontier. *J Financ Quant Anal* 2004;39:239–273.
15. Lohmander P. Adaptive optimization of forest management in a stochastic world. In: Weintraub A, Romero C, Bjørndal T, editors. *Handbook of operations research in natural resources*. New York: Springer; 2007. pp. 525–544.
16. Weintraub A, Vera J. A cutting plane approach for chance constrained linear programs. *Oper Res* 1991;39:776–785.
17. Hof JG, Joyce LA. Spatial optimization for wildlife and timber in managed forest ecosystems. *Forest Sci* 1993;38:489–508.
18. McLain RJ, Lee RG. Adaptive management: pitfalls and promises. *Environ Manage* 1996; 20:437–448.
19. Abramovich A, Weintraub A. Analysis of uncertainty of future timber yields in forest management. *Forest Sci* 1995;41:217–234.
20. Palma C, Cui W, Martell G, *et al.* Assessing the impact of stand-level harvests on the flammability of forest landscapes. *Int J Wildland Fire* 2007;16:584–592.
21. Weintraub A, Navon D. A forest management planning model integrating silvicultural and transportation activities. *Manage Sci* 1976;22:1299–1309.
22. Jones J, Hyde J, Meachan ML. Four analytical approaches for integrating land and transportation planning on forest lands. *Research vadjust* Paper INT-361. U.S. Department of Agriculture and Forest Service; 1986.
23. Constantino M, Martins I, Borges JG. A new mixed-integer programming model for harvest scheduling subject to maximum area restrictions. *Oper Res* 2008;56:543–551.
24. Diaz Legues A, Ferland JA, Ribeiro CC, *et al.* A tabu search approach for solving a difficult forest harvesting machine location problem. *Eur J Oper Res* 2007;179:788–805.
25. Henningsson M, Karlsson J, Ronnqvist M. Optimization models for forest road upgrade planning. *J Math Models Algorithm* 2007; 6:3–23.
26. Andalaft N, Andalaft P, Guignard M, *et al.* A problem of forest harvesting and road building solved through model strengthening and Lagrangean relaxation. *Oper Res* 2003;51:613–628.
27. Epstein R, Weintraub A, Sapunar P, *et al.* A combinatorial heuristic approach for solving real-size machinery location and road design problems in forestry planning. *Oper Res* 2006;54:1017–1027.
28. Alonso-Ayuso A, Escudero LF, Guignard M, *et al.* Forestry Management under price uncertainty. *Ann Oper Res* 2009. In press. DOI: 10.1007/s10479-009-0561-0.
29. Birge JR, Louveaux FV. *Introduction to stochastic programming*. New York: Springer; 2004.
30. Rockafellar RT, Wets RJ-B. Scenario and policy aggregation in optimisation under uncertainty. *Math Oper Res* 1991;16:119–147.
31. Alonso-Ayuso A, Escudero LF, Garín A, *et al.* An approach for strategic supply chain planning based on stochastic 0–1 programming. *J Glob Optim* 2003;26:97–124.
32. Alonso-Ayuso A, Escudero LF, Ortuño MT. BFC, a Branch-and-Fix Coordination algorithmic framework for solving some types of stochastic pure and mixed 0-1 programs. *Eur J Oper Res* 2003;151:503–519.

UNDERSTANDING AND MANAGING VARIABILITY

HYUN-SOO AHN
Ross School of Business,
University of Michigan,
Ann Arbor, Michigan

Variability exists in almost all production and manufacturing systems. Demand and product mix may vary over time, as may the processing capacity of the system and the availability of raw materials and components. Even in an assembly line where all tasks are standardized, unexpected machine breakdowns or shortages of people and/or materials constrains output and disrupts the flow. Variability in demand or operation processes makes seamless alignment of demand and supply impossible and degrades system performance. In order to manage an efficient production/service system, it is essential for a manager to understand the causes and effects of variability. This article discusses the sources and types of variability, its effect on a production (or service) system, and some widely used managerial tools to mitigate the adverse effects of variability.

WHAT IS VARIABILITY?

Variability refers to the quality of nonuniformity of a class or entities in an attribute of interest [1,2]. For example, if all jobs take exactly 5 min to complete, then there is no variability in processing time. On the other hand, consider a case where 50% of jobs take exactly 4 min and the other 50% take 6 min. Although the *average* processing time is 5 min, the processing time for each individual job varies as some jobs may take longer (shorter) than the average. In manufacturing processes, many attributes are prone to vary even without specific root causes (e.g., processing times, physical (or chemical) specifications of an output, quality measures, job difficulty, and order size).

In some cases, it is possible to predict how something would vary as a function of environmental variables (such as time or temperature) or management decisions. For instance, it is possible to predict a daily pattern of electricity demand as a function of the time of the day and temperature. When there is a prescheduled setup or maintenance, any job in the queue may have to wait longer to be finished. In these cases, variation and its immediate effects can be predicted or controlled.

On the other hand, some variations can occur beyond our immediate control or prediction. This “random” (stochastic) variation does not have any recognizable pattern. For example, a disruption due to power outage or contamination cannot be predicted. Even in a manufacturing system that is seemingly under control, variations arise because of differences in operators, raw materials, and job difficulty. In most made-to-order systems, the times between customer orders are random, and this random fluctuation affects the workload of a manufacturing system. In a production line, no two outputs are entirely identical. These random variations arise by chance as a result of inherent randomness of the operation process, and are prevalent in manufacturing processes.

In many systems, both types of variability coexist. For example, the number of customers arriving to a restaurant varies in a way that is partially predictable (peaks during lunch or dinner) and partially unpredictable (the exact number of customers cannot be predicted with certainty). Likewise, a plant manager expects that there will be a number of machine outages, but cannot predict which machine will fail and when. Both types of variability degrade the performance of a manufacturing system: variability increases delay, increases work-in-process inventory, and reduces effective capacity and utilization. However, because of its unpredictability, the effects of random variation are more subtle and require an understanding of the relationship between the variability and

the flow of the good. This article primarily focuses on the effect of random variation.

SOURCES OF VARIABILITY IN MANUFACTURING PROCESSES

Some of the most common causes of random variability in manufacturing processes include machine breakdown, setups, rework, availability of raw material and operators, preventive maintenance, and environmental parameters such as temperature. In addition, heterogeneity in production resources (e.g., machine and operator) or job requirements cause variability as well. Virtually every operation has a certain degree of inherent pure randomness, which does not have any assignable causes.

In a manufacturing system where several operations are linked (e.g., flowshop or job shop), variability in an earlier operation is propagated to later operations since the departure (output) from an earlier operation becomes the arrival (input) to workstations performing later operations. On a bigger scale, the variability originated in an upstream member of the supply chain (e.g., disruption, quality uncertainties, and logistic network failure) affects downstream members of the supply chain. Consequently, the variability in one link of a supply chain propagates toward the entire chain. The propagation of variability is even more pronounced in an environment where the flow of all jobs is not identical (e.g., a job shop, a repair center). Even in the case that arrival times and processing times are predictable, the heterogeneity in job routing causes variability.

In both manufacturing and service systems, demand itself is a major source of variability. While some patterns of demand can be estimated with statistical models, no one can exactly predict the size and mix of the demand for end item(s). The number of orders received in a given period (say, day) fluctuates and so does the time between orders. Although it is possible to influence demand through operational and marketing levers, in general, there is always uncertainty about how many orders will arrive and when.

MEASURING VARIABILITY

When variation occurs and has no assignable root causes, it is difficult to predict the outcome of the variation in every single unit. Instead, we describe the underlying randomness of the system using a random variable and its statistical distribution. The value of a random variable is determined by the outcome of a random experiment. The range and likelihood of values that a random variable can attain are described through its probability distribution function. For any real-valued random variable, the cumulative distribution function $F(x)$ describes the probability that the random variable X does not exceed the value of x : $F(x) = P(X \leq x)$. If X is a discrete random variable, one can describe its probability distribution with the probability mass function, $P(X = x)$. If X is a continuous random variable, one can use the probability density function, defined as $f(x) = (dF(x)/dx)$, to characterize its probability distribution. For a given random variable, the statistical parameters (e.g., mean, variance, and skewness) are numerical measures that describe properties of a random variable and its distribution. The most widely used statistical parameter is the mean, which is used as a measure for the (probability-weighted) center of the distribution. For a discrete random variable, the mean of X (the expected value of X) is given by $E[X] = \sum_x xP(X = x)$. Analogously, the mean of a continuous random variable with density $f(\cdot)$ is given by $E[X] = \int xf(x)dx$. In the absence of theoretical mean, one can estimate it with the sample mean, $\bar{X}$: $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ where n represents the number of observations in a sample.

Variability is closely related to statistical dispersion (or variation) from its mean. Statistical parameters that are commonly used to measure the statistical dispersions are the range, the variance (standard deviation), and the coefficient of variation. All assume the value of zero if there is no dispersion and become larger as a distribution becomes more spread out.

Range

The simplest measure of variation is the range, which is the difference between the

maximum value and minimum value. Although it is easy to compute, the range is very sensitive to the extreme values in the data (or the distribution). Furthermore, the range does not take into account the dispersion of the majority of the values and ignores distributional information.

Variance and Standard Deviation

The variance measures the average of the squared distances of the values from the mean of the distribution (denoted by μ) and is often denoted by σ^2 :

$$\sigma^2 = \text{Variance}[X] = E[(X - \mu)^2].$$

Since computing variance requires squaring the distances, it changes the unit of measurement to squared units. The standard deviation, which can be obtained by taking the square root of the variance, rescales the degree of dispersion to its original measurement unit. The standard deviation of a random variable X is simply the positive square root of the variance. When the population distribution is absent, one can use the sample variance, which is the best unbiased estimator for the population variance and can be computed as follows:

$$s^2 = \frac{\sum_i^n (x_i - \bar{X})^2}{n - 1}.$$

The sample standard deviation is obtained by taking the square root of the sample variance.

Coefficient of Variation

One important measure of variability is the coefficient of variation (CV), which is defined as the ratio of the standard deviation (σ) to the mean (μ), as follows:

$$\text{Coefficient of variation:} \\ \text{CV} = \frac{\text{Standard deviation}}{\text{Mean}} = \frac{\sigma}{\mu}.$$

In the absence of the population mean and standard deviation, one could estimate CV using the sample coefficient of variation with

Table 1. Examples of Coefficient of Variations for Commonly used Distributions

Random Variable	Mean	Standard Deviation	Coefficient of Variation
Uniform $[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)}{2\sqrt{3}}$	$\frac{b-a}{\sqrt{3}(b+a)}$
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$	1
Poisson	λ	$\sqrt{\lambda}$	$1/\sqrt{\lambda}$
Normal	μ	σ	$\frac{\sigma}{\mu}$

the sample standard deviation s and the sample mean $\bar{X}$.

The coefficient of variation has a number of nice properties. When there is no variation, the CV is naturally set to 0. Keeping the mean the same, the larger the CV value, the more dispersed (measured in units of standard deviation) the random variable is. For many widely used probability distributions, the CV can be characterized by the distribution parameters. Table 1 lists the mean, standard deviation, and coefficient of variation for some commonly used distributions.

The coefficient of variation measures the variation for the distribution compared to its mean and is useful when the *relative* variation is more important than the absolute variation to a decision maker. For instance, a standard deviation of 0.05 mm could represent a very high level of variation for the diameter of a vascular stent with a nominal thickness of 0.1 mm while a standard deviation of 5 mm could represent a very low level of variation for a large steel pipe. In this case, comparing variation using the standard deviation would be misleading. The CV describes the dispersion of the variable in a way that is independent of the variable's scale or dimension. However, the CV has a few discernible disadvantages. Obviously, the problem arises if the mean of a variable is zero, in which case the CV cannot be computed. Even if the mean of a variable is not zero, but the variable contains both positive and negative values and the mean is close to zero, then the CV can be misleading. However, because most quantities related to the flow of the good in the system (e.g., processing time) are positive quantities, the CV is a good indicator for describing the effects of variability for

the flow of the system. In fact, some of the most pronounced effects of variability, such as increased inventory, can be described as a function of the CV.

EFFECT OF VARIABILITY

In all manufacturing and service systems, variability inevitably arises and affects the performance of the system. Variability affects both the flow of a manufacturing system and the quality of output. Symptoms of variability include large inventory levels, long lead times, decreased utilization, and degradation of throughput.

To illustrate how variability affects the production system, consider a single workstation. Parts that arrive at the workstation can be interpreted in several ways: (i) they could be physical parts or subassemblies; (ii) they could be work orders; (iii) they could be broken machines (in case of a service station) or patients (in a hospital). If the part arrives at the workstation and sees no other part in the system, the part enters service. Otherwise, a line (queue) forms at the workstation.

Suppose that parts arrive at a rate of one every minute. In other words, the first part arrives at time zero, and at the beginning of each minute, a new part arrives. In this case, there is no variability in arrival time. If the workstation processes a part in exactly 1 min, each arriving part sees the machine idle upon arrival and enters service immediately. A queue does not form in this system. In any point in time, no more than one part is within or at the workstation. In addition, notice that demand perfectly matches processing capacity; thus the workstation is utilized 100% of the time. This is an example of a *perfectly balance deterministic production line* [1].

On the other hand, consider another workstation facing the same arriving stream. In this system, however, processing time is variable. With a 50% chance, it takes 1.25 min (75 s) to process a part. With a 50% chance, it takes 0.75 min (45 s) to process a part. Successive processing times are assumed independent. Notice that the average processing time is 1 min per part; thus, the utilization of this system is the same as that of the previous

example. However, the fluctuation in processing times results in inventory buildup. For instance, if the first part requires 1.25 min to process, the second part (which arrives exactly 1 min after the first part) cannot enter service upon arrival. Instead, it will be queued until the service for the first part is completed. Although the system looks balanced when judging by the average arrival and the service rate, occasionally parts will not arrive before the part in service is finished. How many parts will be in queue if this particular workstation operates long enough? It can be shown that, if it operates long enough, the average queue length grows to infinity. If the part requires physical space to store, no place on earth would be big enough to hold all the inventory.

The variability builds up a queue even when the overall utilization of the workstation is below 100%. For instance, in the preceding example, suppose that two consecutive parts arrive at the workstation with 66 s (1.1 min) apart. The utilization of this system, denoted by ρ , is:

$$\begin{aligned} \text{Utilization} = \rho &= \frac{\text{Arrival rate}}{\text{Process capacity}} \\ &= \frac{1}{\frac{\text{Average interarrival time}}{1}} \\ &= \frac{1}{1.1} \approx 90.91\%. \end{aligned}$$

The utilization of 90.91% indicates that, on average, the workstation does not work 100% of the time. Nonetheless, the variability in the process time allows a queue (inventory) to build up. For instance, if the processing time of the first part is 1.25 min, the second part (which arrives after 1 min) needs to wait for 15 s until the workstation clears the first job.

How does the variability in arrival time or processing time affect the time that a part spends in the queue and the inventory of the system? For the case with a single machine, an approximation formula, first developed by Kingman [3], describes the average time that a part waits in the queue. Let λ be the average arrival rate of parts to the workstation

and μ be the average rate of service completion. Thus, the average interarrival time is $1/\lambda$, the average processing time is $1/\mu$, and the utilization of the workstation is $\rho = \lambda/\mu$. From the definition, the coefficient of variation for the interarrival time (denoted by CV_a) and the coefficient of variation for the processing time (denoted by CV_s) are defined as follows:

$$CV_a = \frac{\text{Standard deviation of interarrival time}}{\text{Average interarrival time}}$$

and

$$CV_s = \frac{\text{Standard deviation of processing time}}{\text{Average processing time}}.$$

If interarrival time and/or processing time data are not available, the following approximation formulae (based on the central limit theorem for renewal processes) estimate CV_a and CV_s [4,5]:

$$CV_a = \frac{\text{Standard deviation of number of arrivals per unit time}}{\sqrt{\text{Average number of arrivals per unit time}}}$$

and $CV_s = \frac{\text{Standard deviation of number of jobs processed per unit time}}{\sqrt{\text{Average number of arrivals per unit time}}}.$

Kingman's formula states that the time that a job spends in the queue, denoted by W_q , can be approximated as follows:

$$W_q = \left(\frac{CV_a^2 + CV_s^2}{2} \right) \left(\frac{\rho}{1 - \rho} \right) \frac{1}{\mu}. \quad (1)$$

Note that the expression is valid only when $\lambda \leq \mu$. If $\lambda > \mu$, then in the long run, more parts will arrive than will be served, and thus the queue grows infinitely and so does the waiting time in the queue. Several insightful observations can be made from this simple expression:

- Unless the system is variability-free ($CV_a = CV_s = 0$), delay is an inevitable

phenomenon regardless of the utilization.

- Keeping everything else the same, the waiting time in the queue increases with the variability in interarrival and/or service times.
- The waiting time in the queue is a multiple of average processing time and increases with the utilization of the system. In particular, the waiting time will increase faster and faster as the utilization gets closer to 1. For instance, suppose that the waiting time in the queue for a given system is 1 when $\rho = 0.5$. Keeping everything else constant, the waiting time for the same system will be 1.5 if $\rho = 0.6$, 2.333 if $\rho = 0.7$, 4 if $\rho = 0.8$, 9 if $\rho = 0.9$, 19 if $\rho = 0.95$, and 99 if $\rho = 0.99$. In a highly utilized system, even a small increase in capacity (hence a small decrease in utilization) can significantly reduce waiting time.
- The total time that a part spends in the system (known as *cycle time*) W is the sum of the waiting time in the queue and its own processing time:

$$W = W_q + \text{Average processing time} \\ = \left(\frac{CV_a^2 + CV_s^2}{2} \right) \left(\frac{\rho}{1 - \rho} \right) \frac{1}{\mu} + \frac{1}{\mu}.$$

- The effect of the variability on inventory is Little's Law, which states that the average number of parts in the system, denoted by I , is equal to arrival rate times average cycle time: $I = \lambda W$ (see **Little's Law and Related Results** for details). Substituting the corresponding expression for the cycle time, the total inventory in the system is

$$I = \lambda W = \lambda \left[W_q + \frac{1}{\mu} \right] \\ = \left(\frac{CV_a^2 + CV_s^2}{2} \right) \left(\frac{\rho^2}{1 - \rho} \right) + \rho.$$

Hence, the inventory is proportional to the total time that a part spends in the system (cycle time), which increases in variability and utilization. Notice that,

if the system is highly variable and/or highly utilized, the vast majority of the cycle time is waiting in the queue. This is, in practice, why reducing cycle time often is considered equivalent to reducing inventory.

It can be shown that the expression is exact when the interarrival time is exponentially distributed (i.e., $CV_a = 1$). The formula for this special case (known as a $M/G/1$ queue) is formally known as the *Pollaczek-Khintchine* (P-K) formula [6,7]. If there are multiple workstations working in parallel, the waiting time in the queue W_q can again be approximated in terms of variability, utilization, and average processing time [8,9]. If there are m workstations in parallel, the utilization of the system will be modified to $\rho = (\text{Arrival rate}/\text{Process capacity}) = (\lambda/m\mu)$. Then, the waiting time in the queue can be approximated using the following formula:

$$W_q = \left(\frac{CV_a^2 + CV_s^2}{2} \right) \left(\frac{\rho^{\sqrt{2(m+1)}-1}}{1-\rho} \right) \frac{1}{m\mu}. \quad (2)$$

This approximation is solid as long as the number of workstations is not too large (e.g., $m \leq 20$ or so). In case there are many servers, another approximation by Halfin and Whitt [10] provides a good bound on the waiting time in the queue.

So far, we have seen that the variability in processing or arrival times results in an increase in waiting time and inventory. The above analysis assumes that there is no limit on the amount of inventory that can be queued (which we call *buffer*) in front of the workstation. In practice, physical or financial constraints may put a limit on this buffer. In such cases, the variability reduces the throughput, which is defined as the number of units finished in a given period. As Kingman's formula suggests, no inventory buffer is necessary in the perfectly balanced deterministic system since operation exactly meets demand, as no more than one part will be at any point in time. Recall the earlier example of a perfectly balanced deterministic production line where a new part arrives precisely at the beginning of each minute and the workstation processes each part in

exactly 1 min. In this system, a finished part leaves the system at the beginning of each minute, thus the long run average throughput is equal to the arrival rate, which is one part per minute (equivalently, 60 parts per hour). In a system with variability, this ideal throughput (equals to the arrival rate) can be achieved only when there is no limit on the buffer and the rate of service completion (μ) is greater than the arrival rate (λ).

What happens if there is a limit on the buffer size in a system that has variability? To illustrate this, suppose that the processing time is variable. The average production rate (μ) is still one part per minute, but the processing time for a part can be either less than or more than 1 min. Also, suppose that there is no buffer for inventory. Thus, if an arriving part sees that another part is in the workstation, it will be blocked from entering the system and leave. In this system, the variable processing time makes some parts arrive before the completion of the previous work, and thus will be blocked. Since some parts will be denied entry, the workstation will produce strictly fewer than 60 parts per hour. In the long run, the throughput of this system is smaller than the arrival rate. In addition, occasionally the workstation will be idle (because there is no part to work) and the utilization of the workstation drops. Although the case with no buffer seems extreme, the basic insights remain the same for any finite buffer: A portion of the arriving parts will be denied as the buffer becomes full, and, as a result, the throughput will be strictly lower than the arrival rate. Increasing the buffer will increase throughput and improve utilization as fewer parts will be denied entry. As the buffer grows, more parts can arrive independently from the status of the workstation, and the workstation can operate independently from arrival time fluctuations. The role of inventory buffer becomes more pronounced when operations are linked, where the output from one workstation becomes the input to another workstation. Inventory buffer between two linked operations isolates the effect of variability arising from one operation locally, and such decoupling helps both operations run smoothly.

Although the relationship between the inventory buffer and the throughput is intuitive, calculating the exact throughput for the case with general processing and interarrival time distributions is quite complicated. For a number of special cases (for instance, an $M/M/1/b$ queue or an Erlang loss systems, in both of which service and interarrival times are exponentially distributed), exact expressions are available [5,7]. For the general case, the approximation formula is available (see Ref. 11 for more details).

In most manufacturing systems, a part goes through multiple linked operations. When multiple operations are linked, departures from one workstation will be arrivals to other workstations. Therefore, the variability of the departure process at one workstation propagates to others. An example for this is a flow shop in which all parts sequentially go through a number of operations. Since the departure from workstation i becomes the arrival to workstation $i + 1$, the variability in the departure process from workstation i is indeed the variability of the arrival process to workstation $i + 1$. The exact dynamics on how variability propagates is quite complex and may depend on the utilization of the upstream workstation (workstation i). For instance, if the workstation i is lightly utilized (i.e., ρ_i is close to zero), then arriving parts rarely wait in the queue. In such a case, the variability in the departure process is almost identical to that in the arrival process. On the other hand, if workstation i is heavily utilized (i.e., ρ_i is close to 1), then it is almost always busy. Hence, the variability in the departure process resembles the processing time variability. One simple way to approximate the variability of the arrival process to workstation $i + 1$ is interpolating the variability of the arrival process to workstation i and the variability of processing times at workstation i with the following formula [2]:

$$\begin{aligned} CV_a^2(i+1) &= CV_{\text{departure}}^2(i) \\ &= \rho_i^2 CV_s^2(i) + (1 - \rho_i^2) CV_a^2(i). \end{aligned}$$

The formula implies that the effect of variability at one workstation propagates

throughout the system. The processing time variability in workstation 1 affects the arrival time variability in workstation 2, which then affects the arrival time variability at workstation 3, and so on. A corollary is that reducing the variability in an earlier stage of the operation is likely to have a greater impact than in a later stage.

Recall that, when the buffer size between two workstations is finite, the variability can decrease the throughput and the overall utilization of the system. Two events that degrade the throughput of the system are blocking (the part cannot join the buffer for the next workstation because it is full) and starving (the workstation becomes idle because there is no part to work on). Strategic placement of inventory can reduce these adverse events and increase both the throughput and the utilization. However, raising the inventory level is also detrimental to the system's performance in several ways. First, financial and/or physical resources are tied to the inventory until it is used. When the inventory level is large, it constrains the firm's resources (e.g., physical space, and working capital). Second, large inventory implies large cycle time. A longer cycle time has adverse effects on the quality of the product. Often, a long lead time is perceived as poor customer service resulting from the firm's inability to react to consumer orders in a timely manner. In addition, a longer cycle time (large inventory) makes detecting and responding to quality problems difficult, as many parts need to be worked through ahead of a defective part.

Variability affects the flow of units in the system, along with the quality of the units. In fact, the root cause of the quality problem is variability in operating processes. The variation due to inherent randomness in the operation causes output and quality to vary. Variations in physical dimensions or electronic specifications of the components affect the quality of finished goods. Since defective units either need to be reworked or scrapped, defects are not only costly to the firm but also affect the flow of the system because more units are needed to be created to ship a given number of good units. Suppose that 5% of produced parts are defective and need to be

scrapped. Accounting for possible defects, on average about 105.3 parts need to be processed in order to produce a batch of 100 good parts. The increased demand will increase the utilization of the system and affect the inventory and cycle time. In order to increase throughput without significantly alerting the inventory and cycle time, process improvements that reduce operational variability are necessary. In fact, high quality is one of the primary reasons behind the success of the Just-In-Time production system, popularized by Toyota [12].

MANAGING VARIABILITY

Variability affects the flow and the quality of the goods produced. To insulate the manufacturing process from variability, managers deploy a variety of operational tools to mitigate the effect of variability.

Inventory Buffering

A passive approach toward managing variability is to hold inventory. As illustrated earlier, inventory decouples each operation and isolates the effect of variability. Thus, inventory can alleviate some of the symptoms of variability in system: loss of utilization and throughput, blocking, and starving. Inventory placed between operations helps each workstation operate independently from its neighboring stations. By insulating each operation from its neighbors, the system is capable of producing output at a consistently high throughput rate. Strategic placement of inventory can also buffer the manufacturing operation from external variability (originated from suppliers or demand) as well. Holding safety stock of finished goods buffers the manufacturing operation from demand variability. Likewise, safety stock of key components isolates manufacturing operations from the variability arising from a supplier (e.g., lead time variability and supplier disruption). In the case where demand variability is predictable (seasonal demand or periodic discount), inventory will buffer the difference between capacity and demand.

It should be noted, however, that a buildup of inventory allows managers to work around the problems associated with variability by simply hiding them. In fact, a large inventory hides underlying problems in manufacturing operations, which can be recognized and resolved otherwise. Furthermore, as Little's law suggests, a large inventory means a longer cycle time, which has several adverse effects as well. The benefits of using an inventory buffer must be weighed against its costs.

Time and Capacity Buffering

When an inventory buffer is not feasible, the firm can use time or capacity as a buffer to mitigate the effects of variability. Adding slack time between two tasks ensures that a sequence of tasks with variable processing times is completed. Dynamic lead time quotation that takes into account of congestion in the system lets consumer know when they will receive the good (although excessively long lead times may turn away some customers). Adding temporary capacity (e.g., a floater in a manufacturing line) or allowing overbooking provides a capacity buffer against variable demand. The use of these buffers, once again, hides problems associated with variability, but does not eliminate them. In addition, deploying these buffers can be quite costly. The firm must weigh the benefits of the buffer against its direct and indirect costs associated with the buffer.

Variability Pooling

Another way to mitigate the harmful effect of variability is to pool multiple resources (or demand streams). Pooling allows the resource to be utilized more effectively, as it balances the workload evenly among production resources. In addition, pooling independent sources of variability can lead to a reduction in variability (measured by coefficient of variations).

One form of variability pooling is *queue sharing*, in which multiple workstations share a single queue. Instead of an individual queue for each workstation (hence, each workstation is dedicated to work on the parts

arriving to its own queue only), one could form a single queue shared by a number of like workstations. Pooling enables production resources to be utilized more effectively, as it prevents one workstation being idle while queues build up at other workstations, and thus it reduces cycle time. For example, consider identical workstations (with the capacity of $\mu = 1$), each having parts that arrive at the rate of $\lambda = 0.9$. (Compare this system with the pooled system, which combines the two arrival streams.) Notice that the utilization of the system remains the same since both arrival and processing rates are doubled. Suppose that the variability in both the interarrival time and processing time remains unchanged. Then, it can be shown by comparing Equations (1) and (2) that the time that a part waits in the queue of the pooled system is only 47.5% of that of the system with two separate queues. By pooling multiple workstations, the flow of the system becomes less affected by the status of any particular workstation: Even if one of the workstations experiences a longer than expected processing time, the parts move through the system. This type of pooling has been successfully implemented in many manufacturing and service organizations including banks, call centers, and semiconductor fabs [2].

Variability pooling can also be used at the supply chain level. Instead of placing safety stock at each demand point (e.g., retailer and salesperson), one can aggregate the demand and place safety stock at the distribution center level. By pooling demand from multiple sources, the total demand variability is reduced, thus the same service level can be achieved with a smaller amount of safety stock. Some companies reverse the operational sequence (e.g., knitting before dyeing, stocking a generic inventory of printer assemblies, and putting the power supply which is specific to the region it is shipped, etc.) in order to capitalize the pooling effect [13].

Operational Flexibility

Adding *flexible workstations* or *cross-training workers* can also reduce the effect of variability. When there is no job to

work on, cross-trained workers can aid the congestions at other workstations. This will reduce the starvation and blockage due to variability, and increase the throughput of the system. Examples include bucket-brigade teams used in order-picking centers [14,15] and the use of computed numerically controlled (CNC) or five-axis machine centers used on the shop floor [16]. The concept of flexibility can also be implemented at the supply chain level. The use of a back-up supplier or dual sourcing mitigates the variability originating from a supplier (e.g., disruption). In many instances, addition of limited flexibility (i.e., each plant builds only a few products) is enough to capture much of the benefits from flexibility [17].

Process Improvement

Some modern management practices call for reducing the variability itself instead of working around it. Variability reduction through continuous process improvement is a prerequisite for the success of the *Just-in-Time* (JIT) and *lean* manufacturing systems. The core of these approaches is to improve the production process, eliminating the root causes of operational variability, and to improve efficiency (measured by utilization and throughput), responsiveness (measured by lead time), and quality. For instance, Toyota reduced processing time variability through setup reductions and process standardization. Toyota also improved the quality of its finished products through total quality management (TQM), error proofing, and collaboration with vendors. It should be noted that the reduced cycle time and work in process are the outcome of successful implementation of these approaches, not the reason behind the success. One of the most common misunderstandings about the JIT system is that lowering the inventory itself is a reason behind success. Without reducing operational variability, simply reducing the inventory does not increase the throughput of the system. Once the variability is reduced with process improvement, the amount of inventory required to maintain a given

throughput level is decreased and the cycle time is reduced.

Design for Manufacture

The root causes of variations in the manufacturing process are often attributed to poor product or operation designs. In such case, addressing potential failure points at the design stage could reduce variability. *Design for manufacture* implements this concept [18]. Eliminating non-value-added steps in the manufacturing process (or dropping nonvalue-added parts) will improve the flow time and product quality, along with reducing defects. One example of the design for manufacture is to simplify product or process design to reduce operational complexity. For higher volume parts, producing the parts by casting or extruding can reduce processing time. Standardizing or fool-proofing the manufacturing process is another example of design for manufacture. Instead of using different screws, using a uniform screw size reduces the both assembly time and the chance of assembly errors. Adding dowel pins to a part will help to orient the part easily. Using interchangeable parts or modules helps to reduce variability, as well as to standardize the operation. All of these efforts help to streamline and standardize the manufacturing operations.

Demand Shaping

The variability of consumer demand can be reduced by employing smart pricing or assortment strategies [19]. One example is the use of everyday low pricing (EDLP); popularized by Walmart. Constant pricing eliminates the variations of demand attributed to the changes in price. As a result, the retailer faces a less variable demand stream. Offering a discount to predesigned bundles (e.g., preconfigured computer) is another example of demand shaping. By inducing customers to purchase a certain type of product (as opposed to each customer configuring his/her own product), the firm can make demand process more predictable. Demand shaping can also be applied in service industries. Offering

discounts during off-peak hours (e.g., early bird special, happy hour, weekend rental discount) reduces demand fluctuations and levels out the utilization of the system [20]. In the same spirit, appointment or reservation systems improve the ability to predict future demand. Although appointment or reservation systems alone may not necessarily reduce the underlying variability of the demand process, the improved visibility on future demand enables matching supply with demand better.

REFERENCES

1. Lovejoy W. Variability, buffers, and inventory note. Ann Arbor (MI): Ross School of Business Case Study; 2005.
2. Hopp W, Spearman M. Factory physics. 3rd ed. New York: McGraw-Hill/Irwin; 2008.
3. Kingman JFC. The single server queue in heavy traffic. Proc Camb Philos Soc 1961;57:902–904.
4. Heyman D, Sobel M. Volume 1, Stochastic models in operations research. New York: McGraw-Hill; 1982.
5. Ross SM. Stochastic processes. 2nd ed. New York: Wiley; 1996.
6. Wolff R. Stochastic modeling and the theory of queues. New York: Prentice-Hall; 1989.
7. Gross D, Shortle JF, Thompson JM, et al. Fundamentals of queueing theory. 4th ed. New York: Wiley; 2008.
8. Sakasegawa W. An approximation formula $L_q \cong \alpha \frac{\rho^\beta}{1-\rho}$. Ann Inst Math Stat, Part A 1977;29:67–75.
9. Whitt W. Approximating the $GI/G/m$ Queue. Prod Oper Manag 1993;2(2):114–161.
10. Halfin S, Whitt W. Heavy traffic limits for queues with many exponential servers. Oper Res 1981;29:567–588.
11. Buzacott JA, Shanthikumar JG. Stochastic models of manufacturing systems. Englewood Cliffs (NJ): Prentice Hall; 1993.
12. Liker J. The Toyota Way. New York: McGraw-Hill; 2003.
13. Lee HL, Tang CS. Variability reduction through operations reversal. Manag Sci 1998; 44(2):162–172.
14. Bartholdi JJ, Eisenstein DD. A production line that balances itself. Oper Res 1996;44(1): 21–34.

15. Bartholdi JJ, Eisenstein DD, Foley RD. Performance of bucket brigades when work is stochastic. *Oper Res* 1999;49(5):710–719.
16. Iravani SMR, Buzacott JA, Posner MJM. Operations and shipment scheduling of a batch on a flexible machine. *Oper Res* 2003;51(4):585–601.
17. Jordan W, Graves SC. Principles on the benefit of manufacturing process flexibility. *Manag Sci* 1995;41(4):577–594.
18. Anderson DM. Design for manufacturability & concurrent engineering; how to design for low cost, design in high quality, design for lean manufacture, and design quickly for fast production. Cambrian (CA): CIM Press; 2008.
19. Simchi-Levi D, Kaminsky P, Simchi-Levi E. Designing and managing the supply chain: concepts, strategies and case studies. New York: McGraw-Hill/Irwin; 2008.
20. Cachon G, Terwiesch C. Matching supply with demand: an introduction to operations management. New York: McGraw-Hill/Irwin; 2004.

UNIFORMIZATION IN MARKOV DECISION PROCESSES

OGUZHAN ALAGOZ

MEHMET U.S. AYVACI

Department of Industrial and
Systems Engineering, University
of Wisconsin-Madison, Madison,
Wisconsin

Most Markov decision process (MDP) models consider problems with decisions occurring at discrete time points. On the other hand, there are several real-life applications, particularly in queueing systems, in which the decision maker chooses actions at random times over a continuous-time interval. Such problems can be modeled using continuous-time models. Semi-Markov decision processes (SMDPs) (see *Semi-Markov Decision Processes*), a class of continuous-time models, generalize discrete-time Markov decision processes (DTMDPs) by allowing state changes to occur randomly over continuous time and letting or requiring decisions to be taken whenever the system state changes [1,2]. In SMDPs, the stochastic process defined by the state transitions follows a discrete-time Markov chain while the time between each transition is drawn from a general distribution, independent of transitions [1,3].

Continuous-time Markov decision processes (CTMDPs) constitute a special type of SMDPs in which the transition times between decisions are exponentially distributed and actions are taken at every transition [2]. Uniformization, as a tool, is used to convert a CTMDP into an equivalent DTMDP. Although uniformization has been used to analyze continuous-time Markov processes for a long time [4–7], Serfozo [8] formalized the use of uniformization in the context of countable-state CTMDPs.

In this article, we will describe uniformization in CTMDPs. Although we consider CTMDPs with stationary transition probabilities and reward functions, bounded reward functions, and finite state and action

spaces, the results can be easily extended to CTMDPs with countable state and action spaces as well as to more general spaces with appropriate measurability conditions [8]. While we focus on uniformization only for infinite-horizon CTMDPs with total expected discounted reward criterion, uniformization can also be utilized in CTMDPs with average reward criterion [2]. Assuming a unichain transition probability matrix for every stationary policy, the transformation and modeling scheme for solving CTMDPs with average reward criterion is identical to those described in this article except for the computation of the reward function. The results apply to multichain cases as well, with slight modifications. More information on uniformization in CTMDPs with average reward criterion is available elsewhere [2].

The remainder of this article is organized as follows: we next summarize how uniformization is used to convert a continuous-time Markov chain (CTMC) into an equivalent discrete-time Markov chain. Then, we describe the use of uniformization in CTMDPs. Finally, we present two examples for uniformization.

UNIFORMIZATION IN CONTINUOUS-TIME MARKOV CHAINS

CTMCs are formally defined as follows [9] (see the section titled “Continuous-Time Markov Chains (CTMCs)” in this encyclopedia): a continuous-time stochastic process $\{X(t), t \geq 0\}$ is a CTMC if for all $s, t \geq 0$ and nonnegative integers $i, j, x(u)$ where $0 \leq u < t$,

$$\begin{aligned} P\{X(t+s) = j | X(t) = i, X(u) = x(u), 0 \leq u < t\} \\ = P\{X(t+s) = j | X(t) = i\}. \end{aligned}$$

A CTMC is a stochastic process with the Markovian property; that is, the conditional distribution $P(X(s+t)|X(u))$ of the future state $X(t+s)$ is independent of the past states $X(u), 0 \leq u < t$ and only depends on

the current state $X(t)$ (see **Definition and Examples of Continuous-Time Markov Chains**).

Consider a CTMC in which the time to make a transition from its current state to a different state is exponentially distributed with β for all states. Let $P_{ij}(t)$ denote the probability of being in state j at time t starting from state i at time 0. Note that the number of transitions by time t , $\{N(t), t \geq 0\}$ is a Poisson process with rate β [9]. Therefore, $P_{ij}(t)$ can be recasted by conditioning on the number of transitions by time t as follows:

$$\begin{aligned} P_{ij}(t) &= P\{X(t) = j | X(0) = i\} \\ &= \sum_{n=0}^{\infty} P\{X(t) = j | X(0) = i, N(t) = n\} \\ &\quad \times P(N(t) = n | X(0) = i) \\ &= \sum_{n=0}^{\infty} P\{X(t) = j | X(0) = i, N(t) = n\} \\ &\quad \times e^{-\beta t} \frac{(\beta t)^n}{n!} \\ &= \sum_{n=0}^{\infty} P_{ij}^n e^{-\beta t} \frac{(\beta t)^n}{n!}, \end{aligned} \quad (1)$$

where P_{ij}^n represents the n -step stationary transition probability of an equivalent discrete-time Markov chain with transition probabilities P_{ij} . That is,

$$P\{X(t) = j | X(0) = i, N(t) = n\} = P_{ij}^n. \quad (2)$$

Equation (1) follows from the assumption that the time spent in every state is exponentially distributed with rate (β) . More specifically, the probability of moving from i to j in n transitions is equal to the probability of moving from i to j by time t since moving in n steps does not give any information on which states were visited due to identical sojourn times. Therefore, Equation (2) can be applied only if all states have identical sojourn time distributions. In order to convert a CTMC with different transition rates into a discrete-time Markov chain we use *uniformization*.

Suppose the mean sojourn time at each state is $1/\beta_i$ and there exists a finite constant β such that $\beta_i \leq \beta$, for all i . According to the new scheme, we assign the same transition

rate β to all states i , where the transition process is divided into two: the “fictitious” transitions to the state itself and the transitions to the other states. To match the actual process, we will have the process remain in each state for an exponential amount of time with rate β and the new transition probabilities will be defined as

$$\tilde{P}_{ij} = \begin{cases} 1 - \frac{\beta_i}{\beta}, & j = i, \\ \frac{\beta_i}{\beta} P_{ij}, & j \neq i. \end{cases}$$

Applying the new transition probabilities to Equation (1), we obtain

$$P_{ij}(t) = \sum_{n=0}^{\infty} (\tilde{P}_{ij})^n e^{-\beta t} \frac{(\beta t)^n}{n!}.$$

Figure 1 shows the schematic of a simple uniformization example. In summary, uniformization enabled us to convert a CTMC with state-dependent out-of-state transition rates into an analytically equivalent CTMC with uniform transition rates. This new system could be treated as a discrete-time Markov chain for the purposes of analysis [9].

UNIFORMIZATION IN CONTINUOUS-TIME MARKOV DECISION PROCESSES

In this section, we describe the uniformization process in CTMDPs. We start with the simpler case, where the transition rates are uniform and then extend this to the more general form, where the transition rates are state-and-action-dependent.

Uniform Transition Rates

Consider an infinite-horizon discounted CTMDP with the following reward (cost) function:

$$\lim_{n \rightarrow \infty} E_s \left[\int_0^{t_n} e^{-\alpha t} g(s(t), a(t)) dt \right],$$

where t_n represents the time of n th transition, $n = \{1, \dots, \infty\}$; α is the continuous-time discounting factor, $\alpha > 0$; $g(s(t), a(t))$ is the reward obtained when action $a(t)$ is selected at state $s(t)$. If we let s_n and a_n denote

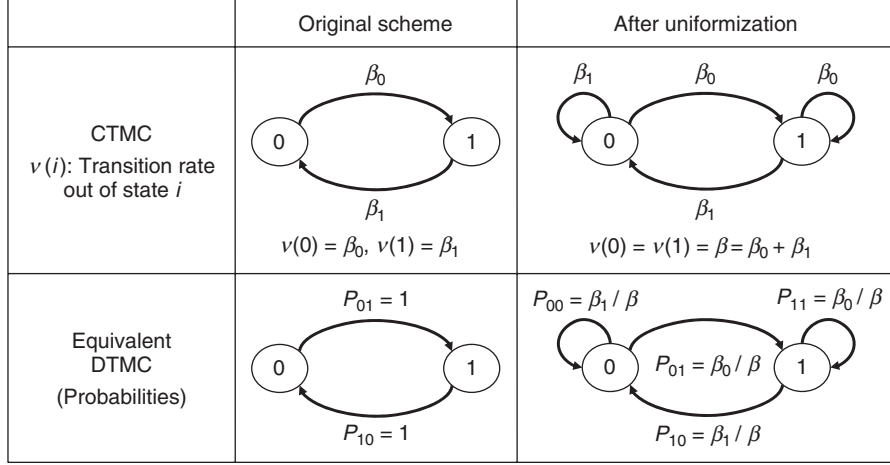


Figure 1. An illustrative example for the uniformization of a CTMC through the use of fictitious self-transitions.

the state and the action selected at time t_n , respectively, then $s(t) = s_n$ and $a(t) = a_n$ hold where $t_n \leq t < t_{n+1}$.

Suppose $g(s(t), a(t))$ consists of two parts: $K(s(t), a(t))$, the lump reward obtained when a new state and action pair is observed and $C(s(t), a(t))$, the continuous reward accumulated if the state was $s(t)$ and action $a(t)$ was taken in the last decision epoch. The state of a CTMDP does not change between decision epochs, therefore, the value of a given policy π for a CTMDP, v_α^π , that is, the total expected discounted reward over the infinite horizon for π is calculated as follows:

$$v_\alpha^\pi = E_s^\pi \left[\sum_{n=0}^{\infty} e^{-\alpha t_n} \left(K(s_n, a_n) + \int_{t_n}^{t_{n+1}} e^{-\alpha(t-t_n)} C(s_n, a_n) dt \right) \right]. \quad (3)$$

Let $\tau_{n+1} = t_{n+1} - t_n$ ($\tau_0 = t_0 = 0$) denote the time the process remains in s_n , which is distributed exponentially with parameter β for all states. Then, we can rewrite Equation (3) as follows:

$$v_\alpha^\pi = E_s^\pi \left[\sum_{n=0}^{\infty} e^{-\alpha t_n} K(s_n, a_n) \right]$$

$$\begin{aligned} & + E_s^\pi \left[\sum_{n=0}^{\infty} e^{-\alpha t_n} C(s_n, a_n) \cdot \int_0^{\tau_{n+1}} e^{-\alpha t} dt \right] \\ & = E_s^\pi \left[\sum_{n=0}^{\infty} e^{-\alpha(\tau_1 + \dots + \tau_n)} K(s_n, a_n) \right] \\ & + E_s^\pi \left[\sum_{n=0}^{\infty} e^{-\alpha(\tau_1 + \dots + \tau_n)} \frac{1}{\alpha} \right. \\ & \quad \left. \times (1 - e^{-\alpha \tau_{n+1}}) C(s_n, a_n) \right] \\ & = \sum_{n=0}^{\infty} E_s^\pi [K(s_n, a_n)] (E_s^\pi [e^{-\alpha \tau_1}])^n \\ & + \sum_{n=0}^{\infty} E_s^\pi [C(s_n, a_n)] \frac{1}{\alpha} (E_s^\pi [e^{-\alpha \tau_1}])^n \\ & \quad \times (1 - E_s^\pi [e^{-\alpha \tau_1}]), \end{aligned} \quad (4)$$

where Equation (4) follows from the assumption that $\{\tau_1, \tau_2, \dots, \tau_{n+1}\}$ are independent and identically distributed with exponential distribution and s_n is independent of $\tau_1, \tau_2, \dots, \tau_{n+1}$. Evaluating the expectation of the exponential,

$$E_s^\pi [e^{-\alpha \tau_1}] = \int_0^{\infty} e^{-\alpha t} \beta e^{-\beta t} dt = \frac{\beta}{\alpha + \beta} := \lambda,$$

and rewriting this equation with the λ substituted in Equation (4), we obtain

$$\begin{aligned}
v_\alpha^\pi &= E_s^\pi \left[\sum_{n=0}^{\infty} \lambda^n \left(K(s_n, a_n) + \frac{C(s_n, a_n)}{\alpha + \beta} \right) \right] \\
&= E_s^\pi \left[\sum_{n=0}^{\infty} \lambda^n r(s_n, a_n) \right],
\end{aligned}$$

which has the same form as an equivalent DTMDP if we redefine $[K(s_n, a_n) + C(s_n, a_n)/(\alpha + \beta)] = r(s_n, a_n)$ as the total expected discounted reward between two decision epochs for the (s_n, a_n) pair. Note that this was achieved since the sojourn times at each state are assumed to be independent and identically distributed with exponential distribution.

To summarize, a CTMDP with the reward function

$$\lim_{n \rightarrow \infty} E_s \left[\int_0^{t_n} e^{-\alpha t} g(s(t), a(t)) dt \right],$$

and a transition rate β for all states and actions is equivalent to a DTMDP with the discount factor

$$\lambda = \frac{\beta}{\alpha + \beta},$$

and the total expected discounted reward between two decision epochs is given by

$$r(s, a) = \left[K(s, a) + \frac{C(s, a)}{\alpha + \beta} \right], \quad (5)$$

where the reward functions K and C are defined as above. Let $P(j|s, a)$ represent the probability that the state at the next decision epoch will be j , given that the state is s and action a is taken at the last decision epoch. Then, the Bellman equations can be rewritten as

$$v(s) = \max_{a \in A_s} \left\{ r(s, a) + \lambda \sum_{j \in S} P(j|s, a) v(j) \right\},$$

and be solved as a DTMDP, where A_s and $v(j)$ represent the set of available actions at state s and the optimal total expected discounted reward that can be obtained when the process starts in state j , respectively.

Nonuniform Transition Rates

A major limiting assumption for the above result is the assumption of identical transition rates across all states and actions. In this section, we show that by *allowing fictitious transitions from a state to itself* as in the previous section, we can extend the results for CTMDPs with uniform transition rates to those with nonuniform transition rates.

Let $\beta(s, a)$ denote the transition rate out of state s when action a is taken and β represent a uniform transition rate satisfying $\beta(s, a) \leq \beta$, $\forall s \in S$ and $a \in A_s$, where A_s represents the action space for state s . We can then modify the transition probabilities as

$$\tilde{P}(j|s, a) = \begin{cases} 1 - \frac{\beta(s, a)}{\beta}, & j = s; \\ \frac{\beta(s, a)}{\beta} P(j|s, a), & j \neq s. \end{cases} \quad (6)$$

By creating fictitious transitions, we are creating a stochastically equivalent process in which the transitions occur more often. As an example, when the process is in state s , it will leave s at a faster rate β but will return to the same state with $1 - \beta(s, a)/\beta$ probability. Probabilistically, the new process will move to another state at the same rate as in the original one. As a result of the uniformization of the nonidentical transition rates, we can use the results for CTMDPs with uniform transition rates.

To summarize, we can analyze a CTMDP with exponential transition rates $\beta(s, a)$, transition probabilities $P(j|s, a)$, and a reward function of

$$\lim_{N \rightarrow \infty} E_s \left[\int_0^{t_N} e^{-\alpha t} g(s(t), a(t)) dt \right],$$

by converting it into an equivalent CTMDP with the discount factor $\lambda = \frac{\beta}{\alpha + \beta}$, where $\beta(s, a) \leq \beta$, $\forall s \in S$, $\forall a \in A_s$ and the transition probabilities are given by Equation (6). The total expected discounted reward between two decision epochs is given by Equation (5). The optimality equation is then written as

$$v(s) = \max_{a \in A_s} \left\{ K(s, a) + \frac{C(s, a)}{\alpha + \beta} \right\}$$

$$+ \frac{\beta}{\alpha + \beta} \sum_{j \in S} \tilde{P}(j|s, a) v(j) \Big\} \quad (7)$$

$$= \max_{a \in A_s} \left\{ r(s, a) + \lambda \sum_{j \in S} \tilde{P}(j|s, a) v(j) \right\} \quad (8)$$

and can be analyzed as a DTMDP. It can easily be shown [10] that after several simple algebraic manipulations, Equations (7) and (8) can also be presented as follows:

$$v(s) = \frac{1}{\alpha + \beta} \max_{a \in A_s} \left\{ (\alpha + \beta) K(s, a) + C(s, a) + (\beta - \beta(s, a)) v(s) + \beta(s, a) \sum_{j \in S} P(j|s, a) v(j) \right\}.$$

The optimality equations given by Equation (8) provide a compact form that is very similar to the conventional optimality equations for DTMDPs and therefore, are easier to comprehend.

EXAMPLES

In this section, we present two simple examples from queueing systems to illustrate the use of uniformization in continuous-time Markov models. More examples for the application of uniformization to CTMDPs are available elsewhere [2,10,11].

Meeting the Professor. Students come in randomly during Professor Smith's office hours and on some occasions, they find the professor busy with other students, in which case they leave and return later. The interarrival times of the students are independent and identically distributed exponentially with rate ω , and it takes exponential amount of time with rate μ for Professor Smith to finish with a student. A student arrives at the office and finds the professor busy with another student. We compute the probability that the professor will be available if the student comes back at time t .

We model the process as a birth and death process, where states 0 and 1 represent

when the professor is available and busy with another student, respectively. We can solve a set of differential equations to calculate the probability in the question. However, we will solve this problem using uniformization. Note that this problem is essentially an M/M/1/1 queue. The reader can refer to Ref. 9 for an analysis of the model to derive the probability in question.

The process has the following parameters:

$$\beta_0 = \omega, \beta_1 = \mu, \text{ and } P_{01} = P_{10} = 1.$$

By defining $\beta = \omega + \mu$, we can uniformize the CTMC to obtain

$$\begin{aligned} \tilde{P}_{00} &= \frac{\mu}{\omega + \mu} = 1 - \tilde{P}_{01}, \\ \tilde{P}_{10} &= \frac{\mu}{\omega + \mu} = 1 - \tilde{P}_{11}. \end{aligned}$$

This creates a new transition matrix with identical entries in a column:

$$\tilde{P}_{ij} = \begin{pmatrix} \frac{\mu}{\omega + \mu} & \frac{\omega}{\omega + \mu} \\ \frac{\mu}{\omega + \mu} & \frac{\omega}{\omega + \mu} \end{pmatrix} = \tilde{P}_{ij}^n$$

$$i, j = \{0, 1\} \quad n = 1, 2, \dots$$

Hence, using the uniformization for CTMCs,

$$\begin{aligned} P_{11}(t) &= \sum_{n=0}^{\infty} \tilde{P}_{11}^n e^{-(\omega + \mu)t} \frac{[(\omega + \mu)t]^n}{n!} \\ &= e^{-(\omega + \mu)t} + \sum_{n=1}^{\infty} \left(\frac{\omega}{\omega + \mu} \right) \\ &\quad \times e^{-(\omega + \mu)t} \frac{[(\omega + \mu)t]^n}{n!} \\ &= e^{-(\omega + \mu)t} + [1 - e^{-(\omega + \mu)t}] \frac{\omega}{\omega + \mu} \\ &= \frac{\omega}{\omega + \mu} + \frac{\mu}{\omega + \mu} e^{-(\omega + \mu)t}. \end{aligned}$$

The required probability is then

$$P_{10}(t) = 1 - P_{11}(t) = \frac{\mu}{\omega + \mu} [1 - e^{-(\omega + \mu)t}].$$

Professor's Dilemma. Consider a slightly modified version of the above example. Namely, we will now model the professor's

decision on how fast he should answer a student's questions.

Suppose, the professor has only three chairs in the office and students coming to the office hours get in only if there is a vacant seat in the office. Every time a student comes in, the professor sets his pace in answering questions so that he can be fair to all that are waiting. His pace in terms of time he expects to spend with a student, is distributed exponentially with a mean that is in $[1/\bar{\mu}, 1/\mu]$ interval. Each time a student comes in, the professor will accrue a reward of $U(s, \mu)$, the immediate utility he gets when the total number of students in the office is s and he chooses his pace as μ . The utility depends on the number of people waiting, as well as the pace reflecting the quality of time he expects to spend with the students. On the other hand, he accrues a utility of $u(s, \mu)$ while he is answering the student's questions and this is continuously discounted at rate α reflecting the fact that the more time he spends with a student, the less he can spend with others. We will write the optimality equations for Professor's pace decision problem.

We define the state space $S = \{0, 1, 2, 3\}$, representing the number of students that are in the office. The transition rates are

$$\beta(s, \mu) = \begin{cases} \omega, & s = 0; \\ \omega + \mu, & s = 1, 2; \\ \mu, & s = 3. \end{cases}$$

The maximum possible transition rate is $\beta = \omega + \bar{\mu}$. The new transition probability matrix for a given μ is as follows:

$$\tilde{P} = \begin{pmatrix} \frac{\bar{\mu}}{\omega + \bar{\mu}} & \frac{\omega}{\omega + \bar{\mu}} & 0 & 0 \\ \frac{\mu}{\omega + \bar{\mu}} & \frac{\bar{\mu} - \mu}{\omega + \bar{\mu}} & \frac{\omega}{\omega + \bar{\mu}} & 0 \\ 0 & \frac{\mu}{\omega + \bar{\mu}} & \frac{\bar{\mu} - \mu}{\omega + \bar{\mu}} & \frac{\omega}{\omega + \bar{\mu}} \\ 0 & 0 & \frac{\mu}{\omega + \bar{\mu}} & \frac{\omega}{\omega + \bar{\mu}} \end{pmatrix},$$

where (s, j) entry of $\tilde{P}$ represents $\tilde{P}(j|s, a = \mu)$. Uniformizing the decision process via using β and the above transition matrix followed by application of Equation (7) leads us to the optimality equations given by

$$v(s) = \max_{\mu \in [\underline{\mu}, \bar{\mu}]} \left\{ U(s, \mu) + \frac{u(s, \mu)}{\alpha + \beta} + \frac{\beta}{\alpha + \beta} \sum_{j \in S} \tilde{P}(j|s, \mu) v(j) \right\}.$$

As discussed in the previous section, the above equations could be converted to the form of Equation (8), and therefore can be solved using conventional DTMDP solution techniques such as value iteration, policy iteration, or linear programming (see **Total Expected Discounted Reward MDPs: Value Iteration Algorithm** and **Total Expected Discounted Reward MDPs: Policy Iteration Algorithm**).

Acknowledgments

This article was supported in part by National Science Foundation grant CMMI-0700094. The authors thank Jeffrey Kharoufeh and two anonymous referees for their suggestions and insights, which improved this manuscript.

REFERENCES

1. Heyman DP, Sobel MJ. Stochastic models. New York: Elsevier Science Publications; 1990.
2. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley & Sons, Inc.; 1994.
3. Cinlar E. Introduction to stochastic processes. Englewood Cliffs (NJ): Prentice Hall; 1975.
4. Howard R. Dynamic programming and Markov processes. Cambridge (MA): MIT Press; 1960.
5. Jensen A. Markov chains as an aid in the study of Markov processes. Skand Aktuarietidskr 1953;340(3):87–91.
6. Lippman SA. Applying a new device in the optimization of exponential queuing systems. Oper Res 1975;23(4):687–710.
7. Veinott AF. Discrete dynamic programming with sensitive discount optimality. Ann Math Stat 1969;40:1635–1660.
8. Serfozo R. An equivalence between discrete and continuous time Markov decision processes. Oper Res 1979;27:616–620.

9. Ross SM. Introduction to probability models. New York: Academic Press; 2007.
10. Bertsekas DP. Volumes 1 and 2, Dynamic programming and stochastic control. Belmont (MA): Athena Scientific; 2001.
11. Walrand J. An introduction to queueing networks. Englewood Cliffs (NJ): Prentice Hall; 1988.

USE OF A HIGH-FIDELITY UAS SIMULATION FOR DESIGN, TESTING, TRAINING, AND MISSION PLANNING FOR OPERATION IN COMPLEX ENVIRONMENTS

JEFFREY D. BARTON
BOHDAN CYBYK
DAVID G. DREWRY
TIMOTHY M. FREY
BRIAN K. FUNK
R. CHAD HAWTHORNE
JOHN F. KEANE
Johns Hopkins University Applied Physics
Laboratory, Laurel, Maryland

OVERVIEW AND NEED

Today's operational forces are experiencing an explosion in both the need and availability of small, agile Unmanned Air System (UAS) platforms to support small unit operations in urban and mountainous settings. However, there are a number of factors that limit the capability of these small platforms in these environments. First, small UAS are usually equipped with limited, small video payloads, are generally restricted to line-of-sight communication, and have a one-vehicle/multi-operator operational architecture, which can restrict the timely exchange of information. In addition, existing small UAS platforms do not readily lend themselves to flight within the complex, highly variable *aerodynamic* environments associated with the mission areas defined above. Because of this factor alone, it is important that a detailed understanding of the unsteady airflow environment and its impact on vehicle performance be developed to reduce the high risk of failure associated with these UAS missions. Traditionally, a combination of controlled wind tunnel experiments, field trials, and computer modeling would be used to underpin UAS technology development programs. However, experimental facilities are

not yet available that replicate the primary physics of the urban (or other complex environment) conditions on a scaled level and field trials are often prohibitively expensive or logistically impossible. As such, there is a need for accurate, high-fidelity air flow models to help reduce the risk of failure of UAS surveillance and engagement missions.

Urban aerodynamics, for example, is complicated due to the existence of complex interactions between terrain geometry, physical conditions, and varying meteorology. High-fidelity computational fluid dynamics (CFD) models exist, which capture these interactions as well as the primary first-order effects of turbulence, heat transfer, and buoyancy. The effects of turbulence, heat transfer, and buoyancy must be given special attention to accurately characterize airflow in an urban environment. Using sufficient resolution, these large eddy simulation (LES) models provide a viable means to characterize urban environments within the urban canopy layer. The authors have successfully integrated these extant CFD models and their real-time effects on embedded UAS within a synthetic modeling environment. With the understanding that aerodynamic effects are not the *only* driver of mission success, the authors have incorporated urban-terrain UAS aerodynamic data and a UAS vehicle dynamics model (which includes the systems, autopilot, algorithms, etc.) as well into a real-time, comprehensive, physics-based simulation tool for the study of vehicle–environment interactions (Fig. 1) and their impact on mission success. With further development, this tool could improve our understanding of UAS performance in complex environments, providing unique capability and flexibility to the small UAS systems engineering and operational communities.

Owing to its comprehensiveness and physics-based underpinnings, such a tool is capable of supporting the entire systems engineering life-cycle for UAS development and deployment. Because it is modular in design, individual components may be

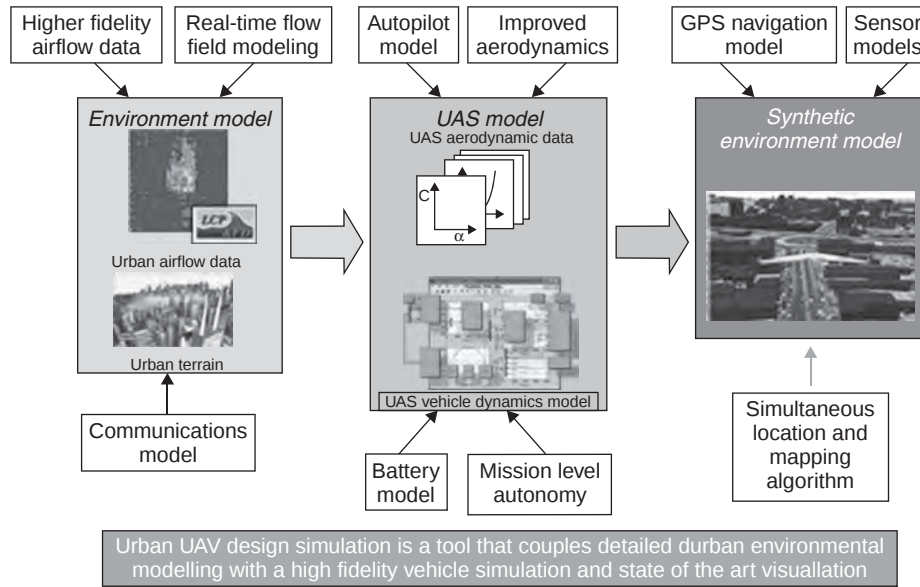


Figure 1. UAV M&S environment key components.

exchanged without fundamentally changing the overall modeling and simulation (M&S) architecture. Communications modeling and mission-level autonomy logic add to its utility as a means of exploring design and operational questions from an end-to-end perspective. With tight development and delivery schedules, not much room exists for systems engineering rigor, quantitative analysis of alternatives, or strong test and evaluation of concepts and designs for to-be fielded systems. Synthetic environments such as the one described here enable design engineers and analysts to address these needs without sacrificing the agility, which is a feature in current UAS acquisition processes.

CHARACTERIZATION OF THE ENVIRONMENT

Urban Aerodynamics

Even when considering flows around simple geometric shapes where turbulence can be a factor, the fluid dynamics is complex, as depicted in Fig. 2. Various experimental studies confirm that the structure of the flow past a surface-mounted cube is sensitive

to both the form of the upstream velocity and the plate boundary conditions. The large-scale, low-frequency vortex shedding off the back of the cube necessitates the use of a time-dependent algorithm to correctly predict time-averaged properties [1]. An understanding of this flow pattern is crucial to predicting turbulent airflows along an air vehicle's intended flight path, in addition to maintaining stable flight throughout mission execution [2].

The complex interactions between terrain geometry, varying meteorological conditions, and the overall physical conditions exhibited by man-made structures make urban aerodynamics exceptionally complicated [4–6] and the small size of the UAS that can actually be carried and deployed by individual ground troops makes them especially susceptible (e.g., stability and control) to these dynamics [7]. Complex terrain includes features such as urban canyons, mountains, and heavily wooded areas and can cover large, three-dimensional, geographic areas with widely variable temporal and spatial conditions. Figure 3 illustrates the impact of urban terrain on the boundary layer. On a macro scale, urban terrain tends to slow the airflow down

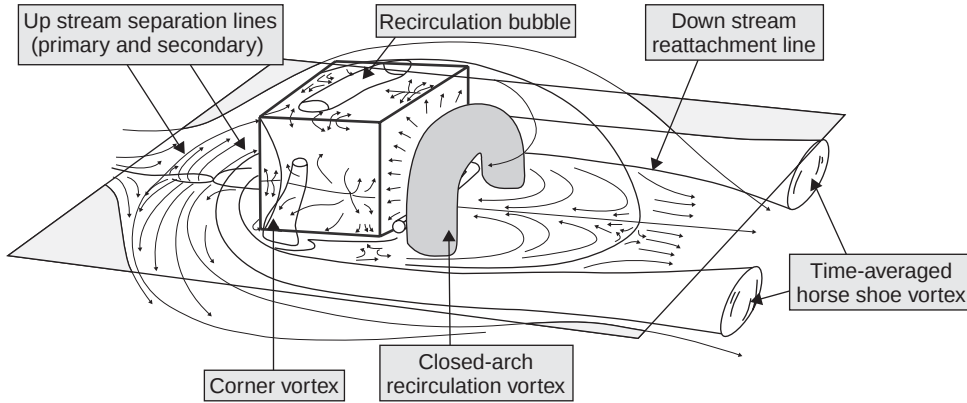


Figure 2. Idealized flow field over a surface-mounted cube [3].

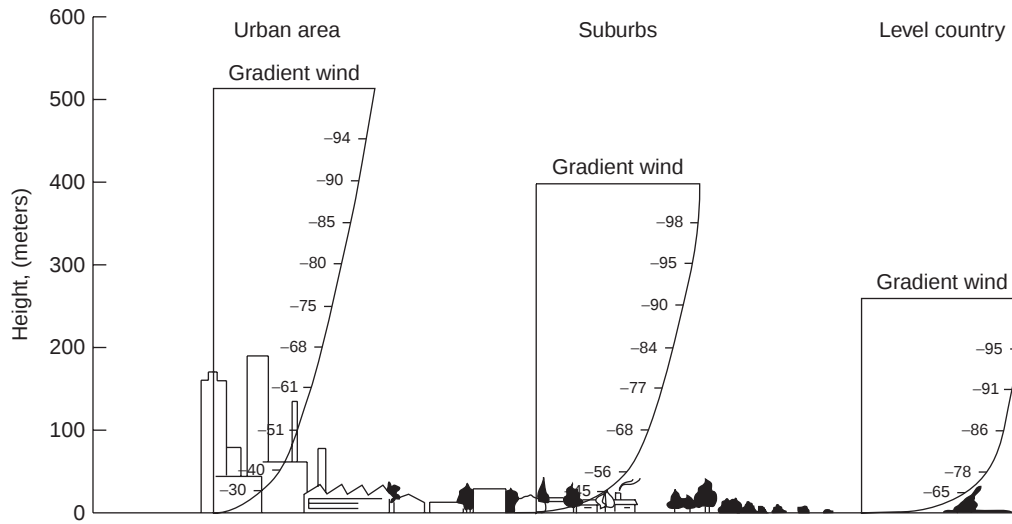


Figure 3. Schematic of urban boundary layer and effects of varying geometry on airflows [8]. Profiles are percentages of gradient wind.

and increase the height of the boundary layer; however, inside the urban canopy layer, buildings in close proximity to one another shed vortices that support large, dynamic recirculation zones which, in turn, push local winds perpendicular to, in reverse to, or even faster downstream than the prevailing wind. In the urban environment, the characteristic scale is on the order of a few meters.

The Environment Characterization Toolset

Air vehicle system designers traditionally employ a combination of controlled wind

tunnel experiments, field trials, and modeling [9–11] in order to characterize air vehicle flight performance requirements and capabilities. Under the tenets of simulation-based design, such experiments and field trials can provide useful data for model refinement, but lack the test environment control needed for validation. In many cases, such experimentation is prohibitively expensive and, under many scenarios, the return on investment is small. Additionally, when investigating the complex flow fields within environments such as the urban canopy, existing wind tunnels provide limited applicability to urban

airflow applications as they are incapable of matching flow conditions at commonly used model scales.

In order to accurately replicate flow conditions over a scaled model in a wind tunnel, the analyst/engineer must carefully match several key parameters that govern a particular class of flows. These parameters associated with external airflows which drive urban aerodynamics include geometric similarity, intensity of turbulence, turbulence length scale, surface roughness, Reynolds number (ratio of inertial to viscous forces), and Richardson number (ratio of buoyancy to vertical shear effects) [12].

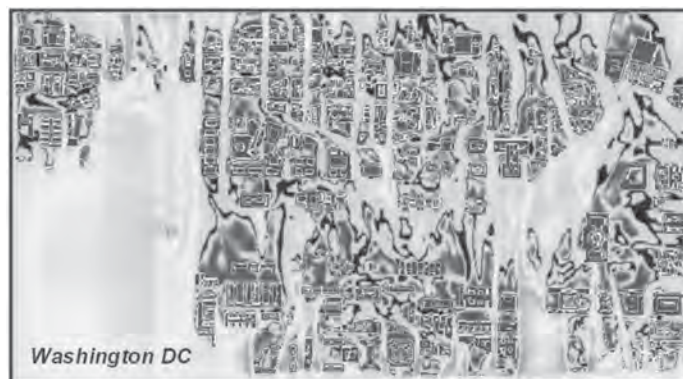
Because experimental facilities are not yet available that replicate the primary physics of the urban problem and because field trials are often prohibitively expensive or logistically impossible, there is a need for accurate, high-fidelity air flow models to help reduce risk of failure and to better understand the physics of UAS surveillance and engagement missions. Owing to the magnitude of the variables involved, the appropriate tool for studying and defining the urban environment is LES [13]. Figure 4 shows the airflow velocity contours from an LES run for Washington D.C. Solutions from these simulations can assist in defining the control authority needed for UAS on board flight control systems to maintain flight stability under varying environmental conditions. Analyzing flow characteristics such as average wind speed changes and correlation to the prevailing meteorological conditions can provide an operating envelope (i.e., area(s) in which UAS

flight is suitable) to design UAS airframes for urban employment, and provide better prediction of near-term local conditions for mission planning and rehearsal.

M&S COMPONENTS

The authors have had considerable experience over the last decade in the development of UAS technologies and the subsequent evaluation and maturation of these technologies through actual flight testing. This real-world experience and the ensuing observations became the impetus for the development of the physics-based modeling and simulation tool described herein. As operational requirements evolved to demand precision operation of UAS within more complex environments, the challenges became readily apparent, but the results were not always intuitive. To illustrate this point, the authors have focused a large portion of their research toward persistent observation and engagement of targets via placement of sensors, and lethal and nonlethal payloads. Flight experiments prior to the development of the M&S tool had shown that large inconsistencies existed in the ability to approach a moving or stationary target with reliable accuracies of less than 1 m. Observations from testing indicated that unsteady winds played a large role. However, it was not intuitively obvious as to the root cause (vehicle aerodynamics, autopilot, flight path, etc.) and the best “system” solution for improving accuracy with the necessary repeatability. The M&S tool

Figure 4. Flow over an urban environment; snapshot of velocity contours 2 m above terrain level predicted by LES [14].



was subsequently developed with a sufficient level of physics, both semi-empirically and analytically derived, to interrogate both individual and coupled mission elements.

As depicted in Fig. 1, the fundamental architecture of the UAS M&S environment comprises three major stand-alone sub-models: the environmental sub-model, the platform sub-model, and the synthetic environment sub-model. Each of these sub-models is described in more detail below.

Environmental Sub-Model

Comprised of everything necessary to model the environment in which the UAS will operate, this sub-model includes components such as the urban terrain, accurate air flow models (as described above), and representations of communication systems. The terrain must not only accurately represent elevation, structures, and other relevant features (e.g., roads and bridges), but it must also contain data about its fundamental properties to support sensor modeling. Available on-board sensors will interact differently with a glass building than a concrete bridge. This “smart” terrain also drives the underpinning of the air flow models, and for especially complex terrains, these models are quite difficult to create and validate. Not only are they greatly affected by the thousands of varying “objects” embedded in the terrain itself, but they necessarily change as a function of time—they are unsteady, requiring the model fidelity the authors have heretofore motivated. Finally, accurate communications modeling is critical to predict overall success of UAS missions in complex environments. Factors such as line of sight, propagation loss, refraction, radio type, and encryption protocols must all be accounted for (see the section titled “UAS Mission Planning Includes Communications Modeling”).

Platform Sub-Model

This sub-model integrates elements of the vehicle flight dynamics with air flow data to capture the complex air flow/vehicle interactions. These interactions are large drivers of overall mission success, are difficult to model, and are treated more thoroughly in

Cybyk *et al.* [7]. The platform sub-model must also capture the specific control algorithms for the modeled platforms, the subsystems that comprise the platforms (e.g., power and navigation), and effectively model payloads (e.g., sensors). Specific challenges in the platform model revolve around how much fidelity in each subcomponent is necessary to sufficiently answer the questions at hand. Also paramount is the feature of modularity previously mentioned. Much of the power of this M&S approach to UAS systems engineering derives from the ability to quickly exchange components (e.g., control algorithm A vs control algorithm B) to perform trade studies and analysis of alternatives in an end-to-end, system-level environment.

Synthetic Environment Sub-Model

This third sub-model integrates all the sub-models discussed above into an end-to-end analysis framework and, as such, must handle synchronization of data, interoperability of M&S components, and overall scenario control. In addition, the synthetic environment provides a visual component to the modeled terrain and platforms as well. This visualization feature not only results in a tool beneficial for training and mission planning, but also creates a useful analysis tool. Because of the complexity of UAS missions in complex environments, the potentially high numbers of entities involved (UAS, ground sensors, warfighters, etc.), and the number of variables that correlate with mission success or failure, the ability to visualize everything that is happening in a given scenario simultaneously provides the human operator the ability to synthesize data and draw conclusions that would likely not be apparent from engineering plots.

Scenario Background

Given the current US operations in the Middle East, a 5 km² area of downtown Baghdad (Fig. 5) was chosen in order to examine the operational relevance of small UAS operations to intelligence gathering in urban scenarios and to provide a use case for baseline evaluation. The flight path flown by the UAS in benign wind conditions is depicted by the



Figure 5. Urban area of 5 km² area of downtown Baghdad modeled for analysis.

dark line. Preliminary analysis using this baseline use case is presented in the following sections.

USE OF UAS M&S ENVIRONMENT

M&S can and should be employed to support the entire systems engineering life-cycle for UAS development and deployment. The development segment involves system design, analysis of alternatives, trade studies, and flight test support, and, indeed, much of our approach here leverages the work done in the systems engineering community in the area of model-based systems engineering (MBSE), as described in Estefan [15] and Weiss and Kathryn [16]. The deployment segment of the life-cycle usually leverages M&S for mission planning, training, and investigation of anomalies. This section provides examples of how the UAS M&S environment can be utilized during both segments.

Actual flight testing is performed using the Procerus Technologies Unicorn air vehicle test platform. The Unicorn is a molded-foam air vehicle that can be built to a 48 or 60 in. wingspan. Equipped with the commercial off-the-shelf Kestrel autopilot, it is

hand-launched. Its electric motor can provide test flights of approximately 1-h duration at speeds of 25–45 mph, depending on environmental conditions. Payloads of up to 1 lb can be accommodated, including a communications package operating within the 900 MHz–2.4 GHz frequency spectrum providing a line-of-sight range of 5 miles. The Procerus Technologies Unicorn air vehicle is modeled in the platform sub-model [17].

Figure 6 illustrates the Unicorn UAS operating in the synthetic Baghdad described above. In addition to the telemetry information for the vehicle, the visualization shows the direction and magnitude of the local wind (dark grey vector pointing toward the vehicle) and the ground speed of the UAS (light grey vector pointing in the direction of travel). The light grey line in the figure illustrates the intended path of the UAS, so at the moment depicted in Fig. 6, the UAS is below and slightly to the left of its desired location due to the effects of the wind.

UAS Sensor Trade Studies

In addition to providing a synthetic environment for the analyst to visualize the operation of the UAS without having to digest and

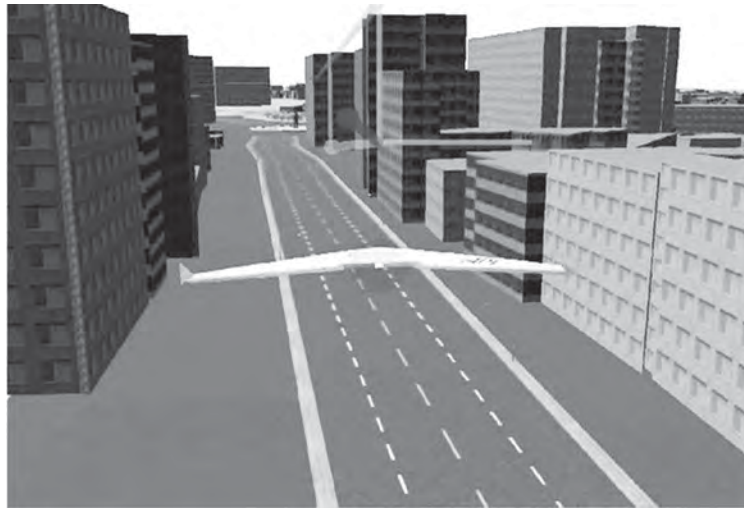


Figure 6. Procerus Technologies Unicorn UAS in synthetic environment.



Figure 7. Synthetic camera model for Procerus UAS.

fuse the information from multiple plots (e.g., latitude, longitude, altitude, pitch, yaw, and roll vs. time), the platform sub-model also enables high fidelity sensor modeling and evaluation. Figure 7 shows the camera view from the Procerus UAS. The view mimics the actual field of view and look-down angle of the on-board gimballed optical sensor.

The synthetic environment also supports the modeling of other sensor types such

as infrared (IR) (Fig. 8), provided that the appropriate material properties are modeled in the terrain, as described in the section titled “Environmental Sub-Model.”

By including the sensor models in the simulation, the M&S user is able to assess the mission impact of the sensor choice and its associated capabilities. Performing sensor trade studies (to determine which ones are optimal for which missions/platforms)

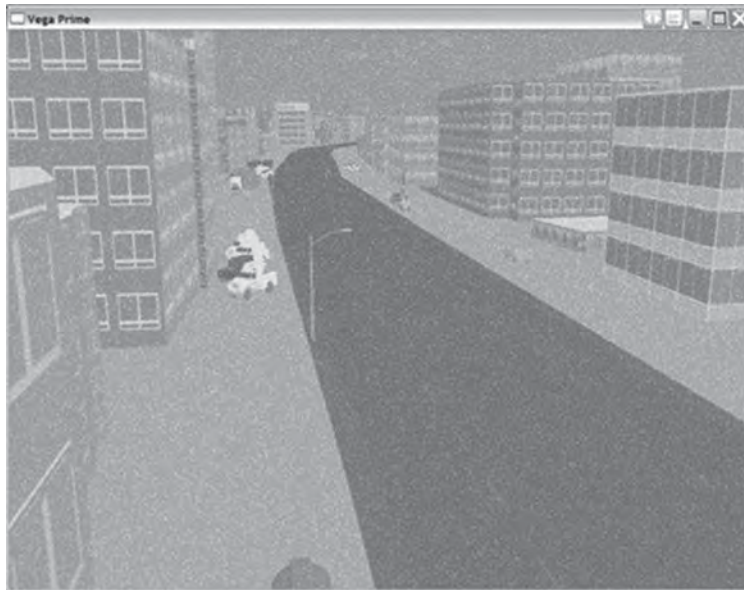


Figure 8. Synthetic infrared (IR) camera model.

by examining sensor capabilities in isolation would likely result in the selection of a sensor with the greatest capabilities. The UAS environment simulation allows the analyst to assess the performance of different sensors as they are integrated on the UAS platform in a mission-level, end-to-end context. As is sometimes the case, the most capable sensor is not always required or even able to achieve mission success. Utilization of the UAS environment simulation allows the analyst to optimize the system in terms of the mission requirements as against assuming that the optimal system is a sum of the most capable components.

Monte Carlo Analysis for UAS Mission Planning

The number of parameters associated with mission success in such dynamic environments makes it difficult to quantify this success in any kind of closed form. Therefore, optimizations and trades must be derived from intelligent use of statistical design of experiments and Monte Carlo analysis. One of the most important features of the synthetic environment model described above is the ability to remove the visualization component and analyze the rest of the M&S

components in many-iteration scenarios, varying the input parameters to derive performance sensitivities of small UAS to factors in the operational environments in which they will be deployed. This feature provides the capability to still maintain an end-to-end, mission-level perspective, yet at the same time modify input parameters “in-the-loop” and derive overall mission-level measures. It is there that synchronization and scenario control become important, as it is no trivial task to run the potentially dozens of independent models in tandem to achieve a unified mission execution.

This Monte Carlo feature certainly allows for trade studies while systems are being designed and built—with the overall goal of optimizing the airframe and mission systems to function in such complex environments. The *same* toolset, however, can be used for mission planning as well. The envisioned use of the tool in the field involves operators inputting real-time environmental and platform data, and receiving identification of no-fly zones in return. Then the mission planner could create a mission plan and use the Monte Carlo feature to compute a probability of

mission success given the environmental conditions. With the addition of autorouting functionality, the tool would be able to create mission plans that maximize or minimize an objective or a weighted combination of objectives set by the mission planner (e.g., maximize probability of mission success, maximize time on station, minimize time to target, and maximize communications connectivity). In this case, the Monte Carlo feature would be used to more accurately compute the objective measures in this dynamic environment.

Figures 9–11 illustrate example results of 30 simulation runs for a planned mission in which the prevailing wind is 27 knots from the northeast. A prevailing wind speed of

27 knots would be considered higher than normal for the area modeled, but it is not an unreasonably high value. The figures depict the position deviation from the planned route as a function of mission time (X - longitudinal, Y - lateral, and Z - vertical). These figures illustrate the uncertain and unsteady nature of the urban aerodynamic environment. At this wind speed, mission success and failure were completely determined by mission start time in this simulation run (i.e., all other input variables, such as trajectory and wind speed, were held constant).

Figures 9 and 10 illustrate several cases where a strong horizontal wind gust blew the UAS completely off course at around $T + 1000$ s (within 17 min of initialization).

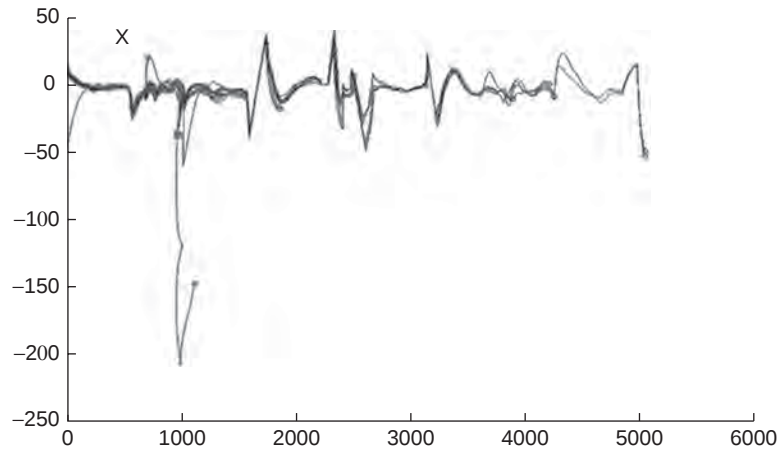


Figure 9. Example Monte Carlo results: Longitudinal path deviation.

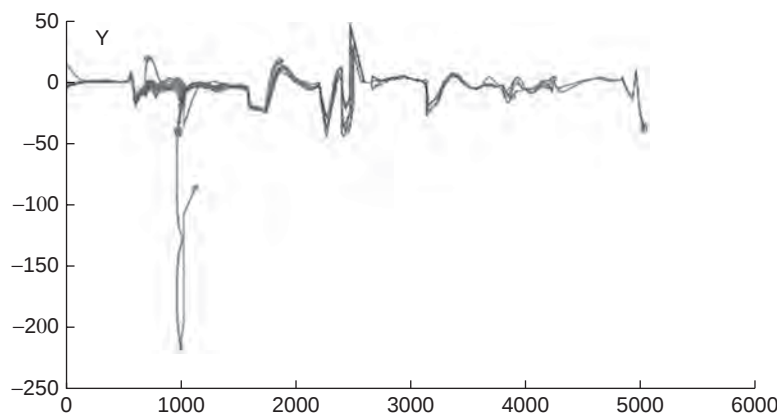


Figure 10. Example Monte Carlo results: Lateral path deviation.

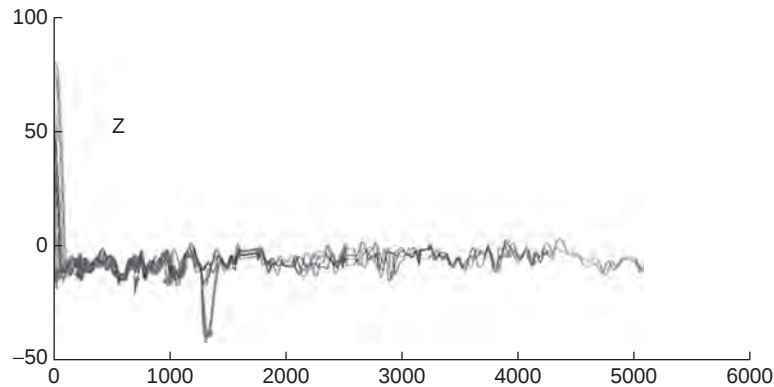


Figure 11. Example Monte Carlo results: Vertical path deviation.

Figure 11 displays several mission failures due to a large down draft at approximately $T + 1300$ s (or within 22 min of initialization).

In addition to this relatively simple example of trajectory following, the authors have used these same techniques on a broader scale to answer questions of how effectively various unmanned platforms of different size, service ceiling, endurance, and payload capabilities can respond to missions such as nuclear forensics and biochemical disaster response. While the details of these studies are beyond the scope of this article and are currently ongoing, they demonstrate the capabilities of the UAS M&S tool to address UAS performance in a mission-level operational context.

UAS Mission Planning Includes Communications Modeling

As mentioned earlier, communications between multiple UAS and/or from the UAS to a ground station is an important factor in the success of UAS operation. Without communications, the UAS is unable to send real-time sensor data to the operator. In addition, should communications be lost, current UAS systems typically terminate their mission flight path and either return to base or fly to a loiter point with the hopes of restoring communications. The ability to maintain communications throughout the mission is typically not part of the mission planning process. This problem is exacerbated by urban and mountainous

terrain and areas where communications jamming is present. As a result, there is a need to model, with high fidelity, the UAS communications environment. Such a model already exists in the Tactical Edge Network Emulation Tool (TENET) [18], which was specifically designed to handle the challenges of ad hoc, dynamic networking analysis and modeling, such as line-of-sight, propagation loss, refraction, radio type, and encryption protocols, and, as mentioned above, has been fully incorporated into the UAV M&S environment described in this article.

Figure 12 depicts a UAS and its available communication paths (light grey lines extending to the right) to the other nodes in its network at a given time in the modeled scenario. In addition to the aforementioned analysis of UAS operations in complex aerodynamic environments, the M&S operator/analyst has the capability to analyze the ability of the different communication systems and architectures to function in a complex communications environment, while the mission planner is able to plan the UAS path and associated ground station placement in order to ensure mission success from a communications standpoint.

Understanding the Effects of the Environment on Power Systems

In order to be effective, small UAS platforms will also be required to transit from their launch points to their intended patrol areas, perform their mission over a given

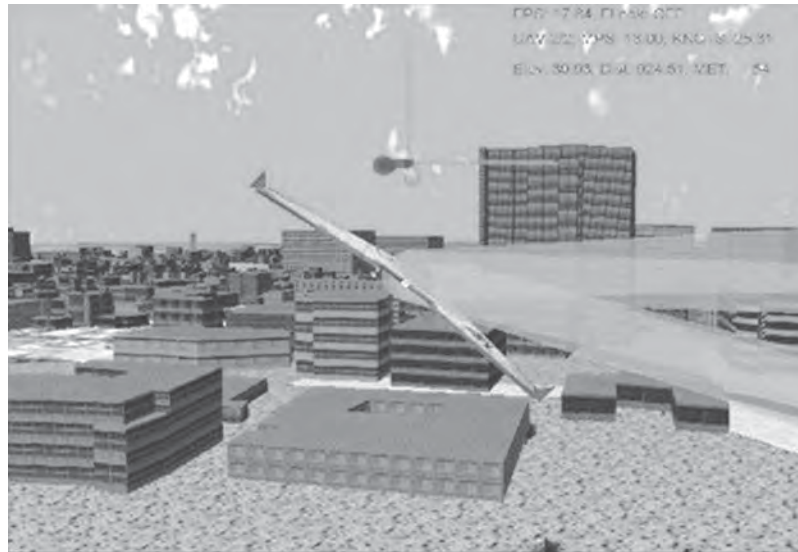


Figure 12. Lines of communication from a UAS in the UAS M&S environment.

time period, and return to friendly territory for recovery. Just as pilots of manned aircraft plan their fuel loads to accomplish similar missions (albeit on a larger scale), the UAS operator and mission planner must work together to ensure that their assets are able to meet mission goals. A part of that mission planning process is a detailed understanding of power consumption and its effect on mission completion rates. Such a tool as developed by the authors would enable mission planners to investigate alternative routes of flight to the intended operating area and return. Figure 13 shows the effects of wind on power consumption for a mission of 1-h duration.

As the air vehicle fights to overcome the effects of the wind, the engine accelerates and decelerates to maintain the required speed to transit along the intended flight path and meet its time gates (grey line with large fluctuations). The other line that drops from 11.5 VDC to 9.3 VDC depicts the effects of this “throttling” on power consumption, and therefore battery drain, with all mission systems operating. In this particular case, the air vehicle was able to meet its mission goals as planned. In other cases, should mission goals exceed the capabilities of the air vehicle systems, alternative

routes of flight can be planned and examined. Ultimately, the authors intend to incorporate an autorouting capability that will enable the simulation to determine a flight path that maximizes the probability of mission success under varying environmental conditions.

M&S VALIDATION VIA FLIGHT TESTING

M&S and flight testing have many interdependencies when it comes to assessing the capabilities and performance of UAS. M&S supplements, but does not replace, flight testing in that the operator can inexpensively run many different variations and test the system in environmental conditions or locations that are not possible to replicate in flight testing (e.g., urban). However, the M&S must be validated for its results to be useful and believable. Validation is typically achieved by comparing simulation data to flight test data. The authors have conducted several flight tests with the Procerus Unicorn UAS that is modeled in the simulation. Flight testing has resulted in the adjustment of the aerodynamic data in the simulation, validation of the autopilot model, and creation of an empirical battery model. Flight testing has also illustrated the difficulty in operating small

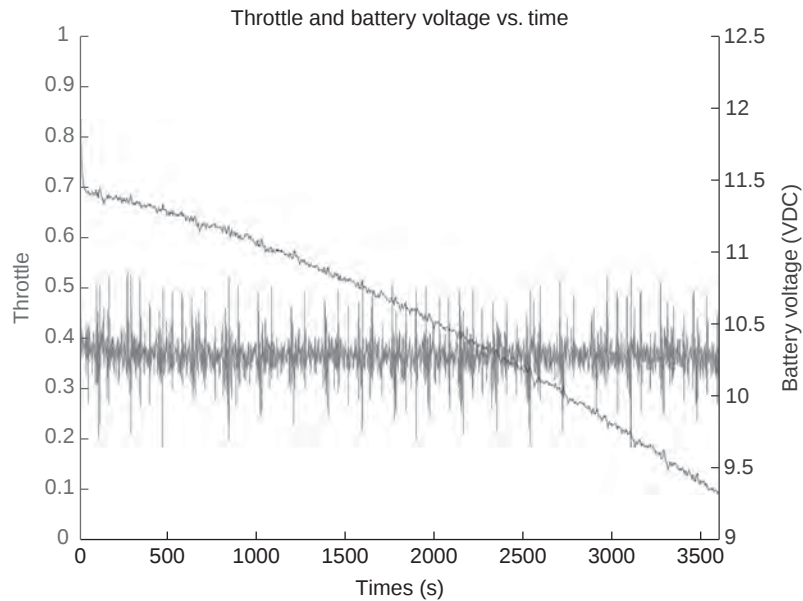


Figure 13. Effects of wind on power consumption.

UAS in unsteady aerodynamic environments. In the field tests, the unsteady aerodynamic environment was caused by gusty winds on the flight test range, but it raises the concern about operation in urban environments where there is a desire to operate but, currently, have no test facilities. For the foreseeable future, testing of UAS operation in urban environments will have to be performed in M&S similar to the UAS simulation described herein.

SUMMARY AND CIVILIAN APPLICATIONS

The maturing of UAS operations for large battlefield engagements has been ongoing for well over a decade. The utility to the warfighter of high- and medium-altitude UAS has been proven in the heat of battle, and large amounts of funding have been focused on developing and tailoring these systems to more efficiently target, track, and engage the enemy. For these systems that operate well above the battlespace and problematic terrain features, modeling of the environment/vehicle interactions is not necessary. However, for smaller, man-portable

UAS that are common at the company level and below, the need to operate within these complex environments is a fundamental element of today's warfare. Consequently, the ability to have an M&S tool that enables operators to train and conduct mission planning including determining flight paths, identifying no-fly zones, and selecting mission-specific UAS configurations is critical.

Until recently, no such tool existed and the warfighter intuitively operated significantly above terrain features that would compromise the ability of the UAS to successfully complete its mission. As the use of the small UAS within complex terrain environments becomes more commonplace, information such as assessment of enemy positions, strengths, and anticipated countermeasures can be provided at the company level in real time. As the fidelity of the M&S increases, a strong emphasis can then be placed on maturing small, low-altitude air vehicles and increasing their reliability and productivity.

Envisioned capability requirements for these small units will include persistent intelligence, surveillance, and reconnaissance (ISR) for force protection, over-the-horizon ISR, and rapid engagement of threats.

To address these requirements, advanced technologies will have to be incorporated onto smaller air vehicle platforms. These technologies include networking between vehicles for decentralized, autonomous operation of multiple UAS by one operator, non-line-of-sight (NLOS) communications, and the incorporation of targeting devices and small, lethal and nonlethal payloads. The M&S framework described in this article provides the ability to evaluate the complex interactions between numerous technological and environmental factors leading to a solution for the warfighter with measurable technological advantages.

The use of high-altitude UAS such as Predator and Global Hawk has in recent years been expanded to include numerous civilian applications. These applications use common sets of military-developed sensor payloads and have been readily adapted to safely provide persistent ISR within the continental US. Missions where these systems have been employed include fire fighting in California, humanitarian assistance after hurricane Katrina, port security, and border patrol. Similar to their military counterparts, these systems are designed for coverage of large areas and do not require the high-fidelity M&S (that capture environment/vehicle interactions) in order to achieve mission success.

The use of these small air vehicles for civilian applications in densely populated urban regions has brought to the forefront the challenge of air-space de-confliction between UAS and manned aircraft. A means to assess the issues with this challenge is an M&S tool, similar in architecture to its military counterpart that appropriately captures not only the environment/vehicle interactions, but also the integration of the UAS into an active and ever-changing airspace. When this challenge is overcome, one can envision many uses for such smaller assets such as close-in inspection from many viewing angles including persistent ISR for crowd observation (e.g., for large sporting events and protests), infrastructure surveillance (e.g., power plants, bridges, and water supplies), port security, and other law enforcement applications.

REFERENCES

1. Castro IP, Robins AG. The flow around a surface-mounted cube in uniform and turbulent streams. *J Fluid Mech* 1977;79, Part 2:307–335.
2. Lakehal D, Rodi W. Calculation of the flow past a surface-mounted cube with two-layer turbulence models. *J Wind Eng Ind Aerodyn* 1997;67&68:65–78.
3. Martinuzzi C, Tropea R. The flow around surface-mounted, prismatic obstacles placed in a fully developed channel flow. *J Fluids Eng* 1993;115:85–92.
4. Allwine KJ, Shinn JH, Streit GE, *et al.* Overview of urban 2000: a multiscale field study of dispersion through an urban environment. *Bull Am Met Soc* 2002;83:521–536.
5. Britter RE, Hanna SR. Flow and dispersion in urban areas. *Ann Rev Fluid Mech* 2003;35:469–496.
6. Snyder WH. Some observations of the influence of stratification in building wakes. In: Castro IP, Rockliff NJ, editors. *Stably stratified flows: flow and dispersion over topography*. Oxford: Clarendon Press; 1994. pp. 301–324.
7. Cybyk BZ, McGrath B, Frey TF, *et al.* Unsteady urban airflows and their impact on small unmanned air system operations. AIAA Paper 09–6049; 2009 August; Chicago, IL: 2009.
8. Turner DB. Workbook of atmospheric dispersion estimates—an introduction to dispersion modeling. CRC Press, Inc.; 1994.
9. Koomullil RP, Soni BK. Wind field simulations in urban area. AIAA Paper AIAA-2001-2621, 15th AIAA Computational Fluid Dynamics Conference; 2001 June 11–14. Anaheim (CA): 2001.
10. Lien FS, Yee E. Numerical modeling of the turbulent flow developing within and over a 3-D building array, part 1: a high-resolution Reynolds-averaged navier-stokes approach. *Bound Layer Meteorol* 2004;112:427–466.
11. Patnaik G, Boris JP, Grinstein FF, *et al.* Large scale urban simulations with the MILES approach. AIAA Paper 2003–4104, AIAA CFD Conference; 2003 June 25. Orlando (FL); 2003.
12. Cybyk BZ, Boris JP, Young TR Jr, *et al.* Simulation of fluid dynamics around complex urban geometries. AIAA Paper 01–0803, 39th Aerospace Sciences Meeting; 2001 Jan 8–11. Reno (NV); 2001.

13. Calhoun R, Gouveia F, Shinn J. Flow around a complex building: experimental and large eddy simulation comparisons. *J Appl Meteorol* 2004;43:696–710.
14. Boris JP. Dust in the wind: challenges for urban aerodynamics. AIAA-2005-5393; 2005 July; Toronto, Ontario: 2005.
15. Estefan J. Survey of model-based systems engineering (MSBE) INCOSE MBSE Focus Group, May. 2007.
16. Weiss KA, Ong EC, Leveson NG. Reusable specification components for model-driven development. Proceedings of the INCOSE 2003 International Symposium; 2003 July. Crystal City (VA); 2003.
17. Procerus Technologies Website. Available at <http://www.procerusuav.com/productsZagiTestAirframe.php>. Accessed 2010 Mar 29.
18. Tebben D, Dwivedi A, Madsen J, *et al.* TENET (Tactical edge network emulation tool): a tool for connectivity analysis for a tactical scenario. IEEE MILCOM Proceedings; 2008. 2007.

USE OF LAGRANGE INTERPOLATING POLYNOMIALS IN THE RLT

WARREN P. ADAMS
Clemson University, Clemson,
South Carolina

An effective strategy for solving a mixed 0–1 optimization program is to form tight polyhedral outer approximations of the convex hull of feasible solutions, and to then use linear programming relaxations to obtain bounds. A substantial body of work has shown how such approximations can be automatically generated in higher-dimensional spaces through a two-step *reformulation-linearization technique* (RLT). In a basic RLT implementation, polynomial representations of a mixed 0–1 program are computed in the reformulation step by multiplying the problem constraints by various “functional factors” consisting of the binary variables x and their complements $1 - x$, and by using the idempotent property that $x^2 = x$. The higher-dimensional spaces are obtained in the linearization step by substituting a continuous variable for each distinct, nonlinear product term.

It turns out that these functional factors of x and $1 - x$ for mixed 0–1 programs are part of a much larger family of functions, called *Lagrange interpolating polynomials* (LIPs), that can be used to generalize this idempotent property so as to encompass discrete variables. As a consequence, the LIPs allow the RLT to extend to mixed-discrete programs. This article reviews the role of LIPs in the RLT process to construct tight approximations and convex hull representations for mixed-discrete programs. The LIP approach was first introduced in Adams and Sherali [1,2, Chapter 4] where comprehensive descriptions are available.

The RLT for mixed 0–1 sets is briefly summarized in the next section. This summary reviews the steps of *reformulation* and *linearization*, and serves as a foundation and basis of comparison for the use of LIPs in the more general discrete framework.

RLT FOR MIXED-BINARY SETS

Consider the mixed 0–1 set

$$\begin{aligned} \mathbf{X} &\equiv \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \geq \mathbf{b}, \\ 0 &\leq x_j \leq 1 \quad \forall j \in N, \\ x_j &\in S_j \quad \forall j \in N, \mathbf{L} \leq \mathbf{y} \leq \mathbf{U}\}, \end{aligned} \quad (1)$$

where $N \equiv \{1, \dots, n\}$, and where each variable x_j can realize values in the associated discrete set $S_j \equiv \{0, 1\}$, for all $j \in N$. The matrices $\mathbf{A}$ and $\mathbf{B}$ are $m \times n$ and $m \times p$, respectively, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^p$, and $\mathbf{b} \in \mathbb{R}^m$. The $\mathbf{y}$ variables are continuous with finite lower and upper bounds $\mathbf{L}$ and $\mathbf{U}$, respectively. As noted in various contexts within Adams and Sherali [1–4], these bounds can be assumed, without loss of generality, to be scaled to lie within the unit interval.

The RLT [2–4] for sets of the form $\mathbf{X}$ constructs an $(n + 1)$ -level hierarchy of linearizations that ranges from the usual continuous relaxation at level-0 to the convex hull of feasible solutions at level- n . It operates as follows. Level-0 is that relaxation obtained by omitting the restrictions $x_j \in S_j$ for all $j \in N$. For each $d \in N$, the level- d representation is computed using the two steps of *reformulation* and *linearization* as follows.

Computation of the level- d RLT of Equation (1)

Step 1. Reformulation: For each $J \subseteq N$ with $|J| = d$, compute the product of each of the 2^d nonnegative “functional factors” $\prod_{j \in J_1} x_j \prod_{j \in J \setminus J_1} (1 - x_j)$ for all sets $J_1 \subseteq J$ with every inequality defining $\mathbf{X}$ of Equation (1), except for those bounding restrictions $0 \leq x_k \leq 1$ having $k \in J$. (For $d = n$, include $\prod_{j \in J_1} x_j \prod_{j \in N \setminus J_1} (1 - x_j) \geq 0$ for all sets $J_1 \subseteq N$ to replace the multiplication of the functional factors with the bounding restrictions $0 \leq x_k \leq 1$ since $J = N$.) Replace all the constraints found within Equation (1) with these newly defined restrictions. Use the

idempotent property of binary variables to set $x_j^2 = x_j$ for all $j \in N$.

Step 2. Linearization: For each $J \subseteq N$ with $2 \leq |J| \leq \min\{d+1, n\}$, substitute a continuous variable w_J for each resulting nonlinear term $\prod_{j \in J} x_j$. For each $J \subseteq N$ with $1 \leq |J| \leq d$, substitute a continuous variable v_{Jk} for each nonlinear term $y_k \prod_{j \in J} x_j$. Call the resulting polyhedral set in the variables $(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{v})$ as $\mathbf{X}_d$.

For clarification, given any $J \subseteq N$, the reformulation step does not multiply functional factors of the form $\prod_{j \in J_1} x_j \prod_{j \in J \setminus J_1} (1 - x_j)$ for $J_1 \subseteq J$ by any restrictions $0 \leq x_k \leq 1$ having $k \in J$ since the resulting inequalities would be implied by $\mathbf{X}_d$. Given any such $J_1 \subseteq J$, multiplication of this expression by some $x_k \geq 0$ having $k \in J \setminus J_1$ or by some $1 - x_k \geq 0$ having $k \in J_1$ trivially simplifies to $0 \geq 0$, while multiplication by some $x_k \geq 0$ having $k \in J_1$ or by some $1 - x_k \geq 0$ having $k \in J \setminus J_1$ makes this same expression nonnegative. In the latter case, for $d \leq n-1$, the inequality $\prod_{j \in J_1} x_j \prod_{j \in J \setminus J_1} (1 - x_j) \geq 0$ is implied by (summing) the two inequalities $\prod_{j \in J_1 \cup \{k\}} x_j \prod_{j \in J \setminus J_1} (1 - x_j) \geq 0$ and $\prod_{j \in J_1} x_j \prod_{j \in (J \cup \{k\}) \setminus J_1} (1 - x_j) \geq 0$ computed in level- d , where k is any index having $k \in N \setminus J$. For $d = n$, all 2^n possible such inequalities with $J = N$ are enforced. Notably, these 2^n inequalities are implied in the set $\mathbf{X}_n$ if there exist any continuous variables $\mathbf{y}$ in $\mathbf{X}$, due to their multiplication with the inequalities $\mathbf{L} \leq \mathbf{y} \leq \mathbf{U}$.

Now, for $d \in N$, the set $\mathbf{X}_d$ possesses two important properties. The first is that, for each such d , the variables $\mathbf{w}$ and $\mathbf{v}$ will equal their intended products for binary $\mathbf{x}$. This has the RLT providing a hierarchy of equivalent mixed 0–1 linear reformulations of $\mathbf{X}$, when $x_j \in S_j$ for all $j \in N$ is enforced within the sets $\mathbf{X}_d$.

The second property involves projections of the sets $\mathbf{X}_d$. Denote the projection of each set $\mathbf{X}_d$, for $d \in \{0, 1, \dots, n\}$, onto the space of the variables $(\mathbf{x}, \mathbf{y})$ as $P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_d)$, where

$$P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_d) \equiv \{(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{v}) \in \mathbf{X}_d \text{ for some } (\mathbf{w}, \mathbf{v})\}. \quad (2)$$

Then the second property is as follows:

$$\begin{aligned} \mathbf{X}_0 &= P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_0) \subseteq P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_1) \\ &\subseteq \dots \subseteq P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_n) = \text{conv}(\mathbf{X}). \end{aligned} \quad (3)$$

The first equation is due to $\mathbf{X}_0$ containing no $(\mathbf{w}, \mathbf{v})$. The family of containments follows from each inequality defining $\mathbf{X}_{d-1}$ being implied by two inequalities defining $\mathbf{X}_d$, for all $d \in N$. The convex hull is established by showing that each extreme point to the polytope $\mathbf{X}_n$ has $\mathbf{x}$ binary, which, together with the first property, ensures that all linearized variables equal their intended products at the extreme points of this highest-dimensional space.

The above construction of the sets $\mathbf{X}_d$ is the foundational RLT implementation. Extensions and generalizations exist. Special structures such as SOS type constraints [5] lead to different functional factors that promote more concise and tighter representations. Conditional logic arguments that exploit the nonlinear nature of the generated inequalities in the reformulation step also lead to tighter forms [5,6]. Equality restrictions promote more concise representations without forfeiting strength [4, pp. 104–105].

Various optimization problems are accommodated by the generated polytopes $\mathbf{X}_d$ and the linearization process. Polynomial objective functions that are linear in $\mathbf{y}$ for fixed $\mathbf{x}$ can be optimized over the sets $\mathbf{X}_d$, provided the selected level- d is large enough to handle the nonlinear terms. Polynomial terms in the constraints are also suitably treated. Extensions handle semi-infinite programs and general convex functions [7]. Moreover, tight relaxations are obtained for general polynomial programs containing products amongst the variables $\mathbf{y}$ themselves [8], such as bilinear programs [9], although the convex hull arguments are not applicable in such cases.

The sets $\mathbf{X}_d$ promote partial convex hull representations. Given any nonempty set $J \subseteq N$, the variables x_j for $j \in N \setminus J$ can be treated as continuous and grouped within $\mathbf{y}$. It then follows that the level- $|J|$ RLT representation over the full complement of remaining binary variables in this modified set, when projected onto the original $(\mathbf{x}, \mathbf{y})$ variable space, provides $\text{conv}(\mathbf{X} \cap x_j \text{ binary } \forall j \in J)$. In fact, a

special case of this argument having $|J| = 1$ is used in Balas *et al.* [10] for generating the convex hull with respect to a single variable x_j , so that $J = \{j\}$.

The increasing strengths of the hierarchical levels, coupled with the explicit convex hull representation at $\mathbf{X}_n$, also provide conditions under which optimal solutions to the continuous relaxations of mixed-binary programs afford direct information relative to the discrete optimum. Inductive arguments [11,12] that systematically progress up the hierarchical levels from $\mathbf{X}_0$ to $\mathbf{X}_n$ identify structures for which binary variables realizing values of 0 or 1 in a continuous optimum will *persist* in retaining these values at the discrete optimum.

Early theoretical and computational contributions to the RLT dealing with quadratic objective functions are recorded in the work of Adams and Sherali [13–15], and computational successes on the famous quadratic assignment problem, both in terms of bound strength and enumerative results, are collectively found in the work of Johnson [16], Adams and Johnson [17], Hahn *et al.* [18], and Adams *et al.* [19]. Other computational results are summarized in the work of Sherali and Adams [2,20].

LAGRANGE INTERPOLATING POLYNOMIALS AND KRONECKER PRODUCTS

Two critical ingredients of the RLT for mixed-binary programs are (i) the multiplication of functional factors with the constraints defining $\mathbf{X}$ and (ii) the enforcement that $x_j^2 = x_j$ for binary x_j . Indeed, this latter ingredient is the step that affords strength to the RLT representations. It turns out that both are extendable to more general regions where the variables x_j are discrete, and allowed to take values within finite sets. To make this extension, the functional factors are replaced with LIPs, and the multiplication of such LIPs with constraints of $\mathbf{X}$ is expressed as Kronecker products of matrices. These two concepts are explored below in preparation for the RLT extension to mixed-discrete programs of the section titled “RLT for Mixed-Discrete Sets.”

Lagrange Interpolating Polynomials

Reconsider the variables x_j , for $j \in N$, of the preceding section, but this time allow each x_j to realize values in some finite set $S_j \equiv \{\theta_{j1}, \dots, \theta_{jk_j}\}$ for a given nonnegative integer k_j . In this manner, each S_j found in Equation (1) is a special case having $k_j = 2$ with $\theta_{j1} = 0$ and $\theta_{j2} = 1$. Associated with each x_j are k_j LIPs (for example, see Horn and Johnson [21, pp. 29–30]). Letting $K_j \equiv \{1, \dots, k_j\}$ for each $j \in N$, these LIPs are computed in terms of the elements in S_j as follows:

$$L_{jk}(x_j) = \frac{\prod_{i \in K_j \setminus \{k\}} (x_j - \theta_{ji})}{\prod_{i \in K_j \setminus \{k\}} (\theta_{jk} - \theta_{ji})} \quad \forall(j, k), k \in K_j, j \in N. \quad (4)$$

For each $j \in N$, the k_j functions $L_{jk}(x_j)$ are polynomials of degree $k_j - 1$ for all $k \in K_j$. They possess the three important properties given below. Here, Properties 2 and 3 are derived from Property 1 but are emphasized because of their significance in an RLT implementation.

Properties of the LIPs of Equation (4), for $j \in N$

1. For each $x_j \in S_j$, $L_{jk}(x_j) = 1$ if $x_j = \theta_{jk}$ and $L_{jk}(x_j) = 0$ otherwise, for all $k \in K_j$.
2. For each $x_j \in S_j$, $L_{jk}(x_j)x_j = L_{jk}(x_j)\theta_{jk}$ for all $k \in K_j$.
3. For each $x_j \in S_j$, $L_{jk}(x_j)L_{ji}(x_j) = L_{jk}(x_j)$ if $i = k$ and $L_{jk}(x_j)L_{ji}(x_j) = 0$ otherwise, for all (i, k) with $i \in K_j$ and $k \in K_j$.

Four immediate consequences of these properties are as follows. First, for each $j \in N$, Property 1 has $L_{jk}(x_j) \geq 0$ for all $k \in K_j$ regardless of the original defining set S_j . Second, for the special case in which $k_j = 2$ with $S_j = \{0, 1\}$ so that x_j is a binary variable, the functions $L_{j1}(x_j)$ and $L_{j2}(x_j)$ reduce to the nonnegative RLT functional factors $1 - x_j$ and x_j respectively, for mixed 0–1 linear programs. Third, Properties 2 and 3, which simplify the multiplication of $L_{jk}(x_j)$ for $k \in K_j$ with the variable x_j and with the functions $L_{ji}(x_j)$ for $i \in K_j$ respectively,

can be considered as generalizations of the idempotent property for binary variables x_j . Specifically, for the special case where $k_j = 2$ and $S_j = \{0, 1\}$, Properties 2 and 3 each enforce that $x_j^2 = x_j$.

The fourth consequence follows from Property 1, and is motivated by the linearization step of the RLT for mixed 0–1 programs described in the previous section. Given a discrete variable x_j that can realize k_j values, compute the associated k_j LIPs of degree $k_j - 1$. Letting $\mathbf{x}^j$ denote the column vector in $\mathbb{R}^{k_j}$ whose i th entry is x_j^{i-1} , the (redundant) k_j inequalities restricting these LIPs to nonnegative values are represented in matrix form as $C_j \mathbf{x}^j \geq \mathbf{0}$, where C_j is that $k_j \times k_j$ matrix having entry (k, i) given by the coefficient on x_j^{i-1} in $L_{jk}(x_j)$. Here, and throughout, the notation $\mathbf{0}$ is used to represent a suitably dimensioned vector having all entries of 0. Property 1 provides that the matrix C_j is invertible with $C_j^{-1} = V_j^T$, where V_j is the $k_j \times k_j$ Vandermonde matrix whose (k, q) th entry is θ_{jk}^{q-1} for each $k \in K_j$, where $0^0 = 1$ for convenience. Now compute the polyhedral region $C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}$ in $\mathbb{R}^{k_j}$ by linearizing $\mathbf{x}^j$ to $\{\mathbf{x}^j\}_L$, obtained by replacing each nonlinear term with a continuous variable. (In the remainder of this article, the notation $\{\bullet\}_L$ is used to denote the linearized form of $\bullet$ that is obtained by substituting a continuous variable for each distinct nonlinear term.) Then, since the first entry of $\{\mathbf{x}^j\}_L$ is 1 and the first row of V_j^T consists entirely of entries of value 1, it follows that $C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}$ is a polytope in $\mathbb{R}^{k_j}$ whose extreme points are the columns of the matrix V_j^T less the first row.

This fourth consequence is stated formally as follows.

Lemma 1. *Given any variable x_j that can realize k_j values in some discrete set $S_j = \{\theta_{j1}, \theta_{j2}, \dots, \theta_{jk_j}\}$, the polyhedral set*

$$P_j = \{\{\mathbf{x}^j\}_L : C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}\},$$

where $C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}$ denotes that linear system in k_j inequalities and $k_j - 1$ variables obtained by linearizing the associated system $C_j \mathbf{x}^j \geq \mathbf{0}$ has the following property. There is a one-to-one correspondence between the

extreme points of P_j and the k_j possible realizations of x_j having $x_j \in S_j$, with $\{\mathbf{x}^j\}_L = \mathbf{x}^j$ at each such point.

Example 1. Consider the discrete variable $x_1 \in S_1 = \{-1, 0, 2\}$. Then the $k_1 = 3$ linear inequalities $C_1 \{\mathbf{x}^1\}_L \geq \mathbf{0}$ define the polytope P_1 given by

$$P_1 = \left\{ \begin{pmatrix} 1 \\ x_1 \\ \ell_1 \end{pmatrix} : \begin{bmatrix} 0 & -\frac{2}{3} & \frac{1}{3} \\ 1 & \frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{6} & \frac{1}{6} \end{bmatrix} \times \begin{pmatrix} 1 \\ x_1 \\ \ell_1 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\},$$

where ℓ_1 is used to denote the linearized version of the quadratic term x_1^2 . The inequalities defining P_1 are the nonnegative linearized LIPs, and the extreme points of P_1 are the columns of the matrix $V_1^T = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 2 \\ 1 & 0 & 4 \end{bmatrix}$ less the first row, as depicted in Fig. 1.

Kronecker Products of Matrices

As a means of applying the RLT to general discrete variables, the LIPs are now the functional factors, multiplied amongst themselves, and with other families of inequalities. These factors replace, and generalize, the expressions $\prod_{j \in J_1} x_j \prod_{j \in J \setminus J_1} (1 - x_j)$ for sets $J_1 \subseteq J \subseteq N$ used in the reformulation step of the section titled “RLT for Mixed-Binary Sets.” This multiplication of LIPs with families of inequalities is most conveniently expressed in terms of Kronecker products of matrices. A brief review of Kronecker products follows.

The Kronecker product of an $m_1 \times n_1$ matrix A_1 with an $m_2 \times n_2$ matrix A_2 , denoted by $A_1 \otimes A_2$, is the $m_1 m_2 \times n_1 n_2$ matrix

$$A_1 \otimes A_2 = \begin{bmatrix} a_{11}A_2 & \dots & a_{1n_1}A_2 \\ \vdots & \ddots & \vdots \\ a_{m_1 1}A_2 & \dots & a_{m_1 n_1}A_2 \end{bmatrix},$$

where a_{ij} represents the (i, j) th entry of A_1 . This definition gives rise to the following two properties of Kronecker products. Here, the scalar r is any integer greater than or equal to 2, and all matrices A_1 through

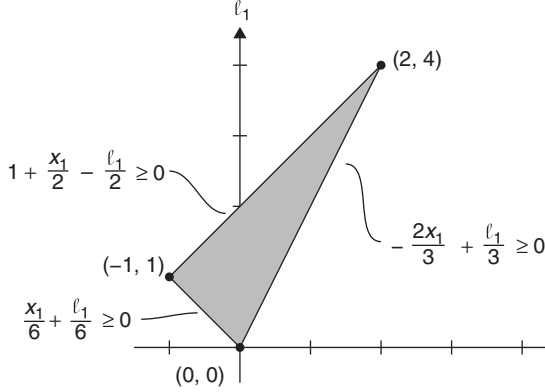


Figure 1. The polytope P_1 for x_1 realizing values in $S_1 = \{-1, 0, 2\}$.

A_{2r} are assumed compatible for the given multiplications.

Properties of Kronecker products of matrices

1. $A_1 A_2 \otimes A_3 A_4 \otimes \cdots \otimes A_{2r-1} A_{2r} = (A_1 \otimes A_3 \otimes \cdots \otimes A_{2r-1})(A_2 \otimes A_4 \otimes \cdots \otimes A_{2r})$.
2. Given that A_1 through A_r are invertible, $(A_1 \otimes A_2 \otimes \cdots \otimes A_r)^{-1} = A_1^{-1} \otimes A_2^{-1} \otimes \cdots \otimes A_r^{-1}$.

These two properties promote a generalization of Lemma 1 to any finite collection of discrete variables x_j for $j \in J \subseteq N$, as well as insight into earlier convex hull arguments for the RLT. Specifically, consider any set $J \subseteq N$ of discrete variables x_j for $j \in J$, where each variable x_j can realize k_j values. Then the $\prod_{j \in J} k_j$ inequalities obtained by computing all possible products of $|J|$ LIPs, with one LIP taken from each variable x_j , and by setting the resulting products nonnegative, can be expressed as $\otimes_{j \in J} C_j \mathbf{x}^j \geq \mathbf{0}$. Property 1 gives that these same inequalities can be expressed as $(\otimes_{j \in J} C_j)(\otimes_{j \in J} \mathbf{x}^j) \geq \mathbf{0}$, with the linearized version obtained by substituting a continuous variable for each nonlinear term given by $\otimes_{j \in J} C_j \{\otimes_{j \in J} \mathbf{x}^j\}_L \geq \mathbf{0}$. Now, since each C_j has an inverse V_j^T that is the transpose of some Vandermonde matrix V_j as explained in the previous section, Property 2 gives that $(\otimes_{j \in J} C_j)^{-1} = \otimes_{j \in J} V_j^T$. A consequence is that the polytope $\otimes_{j \in J} C_j \{\otimes_{j \in J} \mathbf{x}^j\}_L \geq \mathbf{0}$ in $\mathbb{R}^{\prod_{j \in J} k_j - 1}$ has its extreme points enumerated by the columns of the matrix $\otimes_{j \in J} V_j^T$, less the first row. A formal statement is below.

Lemma 2. *Given any $J \subseteq N$ and any collection of variables x_j , for $j \in J$, such that each x_j can realize k_j values in the discrete set $S_j = \{\theta_{j1}, \theta_{j2}, \dots, \theta_{jk_j}\}$, the polyhedral set*

$$P_J = \left\{ \{\otimes_{j \in J} \mathbf{x}^j\}_L : \otimes_{j \in J} C_j \{\otimes_{j \in J} \mathbf{x}^j\}_L \geq \mathbf{0} \right\},$$

where $\otimes_{j \in J} C_j \{\otimes_{j \in J} \mathbf{x}^j\}_L \geq \mathbf{0}$ denotes that linear system of $\prod_{j \in J} k_j$ inequalities and $\prod_{j \in J} k_j - 1$ variables obtained by linearizing the associated system $(\otimes_{j \in J} C_j)(\otimes_{j \in J} \mathbf{x}^j) \geq \mathbf{0}$, has the following property. There is a one-to-one correspondence between the extreme points of P_J and the $\prod_{j \in J} k_j$ possible realizations of $\otimes_{j \in J} \mathbf{x}^j$ having $x_j \in S_j$ for all $j \in J$, with $\{\otimes_{j \in J} \mathbf{x}^j\}_L = \otimes_{j \in J} \mathbf{x}^j$ at each such point.

Lemma 2 is illustrated in the example below.

Example 2. Consider the two discrete variables x_1 and x_2 having $S_1 = \{-1, 0, 2\}$ and $S_2 = \{0, 1\}$ with $k_1 = 3$ and $k_2 = 2$. Defining the matrices C_1 and C_2 and associated vectors $\mathbf{x}^1$ and $\mathbf{x}^2$ as in the section titled “Lagrange Interpolating Polynomials,” the products of the inequalities representing the three nonnegative LIPs for x_1 with the two nonnegative LIPs for x_2 , given by $C_1 \mathbf{x}^1 \otimes C_2 \mathbf{x}^2$, can be set to nonnegative values to obtain

$$\begin{bmatrix} 0 & -\frac{2}{3} & \frac{1}{3} \\ 1 & \frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{6} & \frac{1}{6} \end{bmatrix} \begin{pmatrix} 1 \\ x_1 \\ x_1^2 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \geq \mathbf{0}$$

$$\otimes \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} 1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Property 1 with $r = 2$ allows for the rewriting of $C_1 \mathbf{x}^1 \otimes C_2 \mathbf{x}^2$ to $(C_1 \otimes C_2)(\mathbf{x}^1 \otimes \mathbf{x}^2)$. Linearizing the vector $(\mathbf{x}^1 \otimes \mathbf{x}^2)$ to $\{\mathbf{x}^1 \otimes \mathbf{x}^2\}_L$ by substituting a continuous variable for each nonlinear term, this system yields the following polyhedral set, called P_{12} to reflect the variables x_1 and x_2 .

$$P_{12} = \left\{ \begin{pmatrix} 1 \\ x_2 \\ x_1 \\ x_1 x_2 \\ x_1^2 \\ x_1^2 x_2 \end{pmatrix}_L : \begin{bmatrix} 0 & 0 & -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & 0 & -\frac{2}{3} & 0 & \frac{1}{3} \\ 1 & -1 & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & \frac{1}{6} & -\frac{1}{6} & \frac{1}{6} & -\frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} \end{bmatrix} \times \left\{ \begin{pmatrix} 1 \\ x_2 \\ x_1 \\ x_1 x_2 \\ x_1^2 \\ x_1^2 x_2 \end{pmatrix}_L \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \right\} \right\}$$

The extreme points of P_{12} are the columns of the matrix $V_1^T \otimes V_2^T$ below, less the first row.

$$V_1^T \otimes V_2^T = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 2 \\ 1 & 0 & 4 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\ = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ -1 & -1 & 0 & 0 & 2 & 2 \\ 0 & -1 & 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 & 4 & 4 \\ 0 & 1 & 0 & 0 & 0 & 4 \end{bmatrix}.$$

Lemma 2 provides insight into a convex hull result established in an early work on the RLT. Given a collection of n binary variables indexed as in Equation (1) by the set N , the

paper by Sherali and Adams [3] constructs a special polyhedral set in the following manner. The 2^n inequalities $\prod_{j \in J} x_j \prod_{j \in N \setminus J} (1 - x_j) \geq 0$ for all $J \subseteq N$ are first computed, and then linearized as in the RLT of the section titled “RLT for Mixed-Binary Sets” by substituting a continuous variable w_J for each nonlinear term $\prod_{j \in J} x_j$, to obtain the set

$$\beta = \left\{ (\mathbf{x}, \mathbf{w}) \in \mathbb{R}^n \times \mathbb{R}^{2^n - n - 1} : \left\{ \prod_{j \in J} x_j \prod_{j \in N \setminus J} (1 - x_j) \right\}_L \geq 0 \forall J \subseteq N \right\}.$$

It directly follows from Sherali and Adams [[3] Theorem 2, p. 417] that β is a polytope with 2^n extreme points, and that each extreme point has x_j binary for all $j \in N$, with $w_J = \prod_{j \in J} x_j$ for all $J \subseteq N$ having $|J| \geq 2$. This same result is alternately established by Lemma 2.

RLT FOR MIXED-DISCRETE SETS

Reconsider $\mathbf{X}$ of Equation (1) where the sets S_j for $j \in N$ are as defined in the section titled “Lagrange Interpolating Polynomials” to be $S_j = \{\theta_{j1}, \dots, \theta_{jk_j}\}$ for given nonnegative integers k_j . Assume without loss of generality that $\theta_{j1} < \theta_{j2} < \dots < \theta_{jk_j}$ for each j . Include within $\mathbf{X}$ the nonnegative restrictions on all LIPs, given by $L_{jk}(x_j) \geq 0$ for all (j, k) , $k \in K_j$, $j \in N$, expressed in matrix form as $C_j \mathbf{x}^j \geq \mathbf{0}$ for all $j \in N$. For convenience, the set is stated below.

$$\mathbf{X} \equiv \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p : \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \geq \mathbf{b}, C_j \mathbf{x}^j \geq \mathbf{0} \forall j \in N, x_j \in S_j \forall j \in N, \mathbf{L} \leq \mathbf{y} \leq \mathbf{U}\} \quad (5)$$

Similar to the RLT for mixed-binary sets, the RLT for mixed-discrete sets constructs an $(n + 1)$ -level hierarchy of linearizations, denoted $\mathbf{X}_d$, $d = 0, \dots, n$. The level-0 relaxation is the polyhedral set of Equation (5) without the $x_j \in S_j$ restrictions, and with $\mathbf{x}^j$ replaced by $\{\mathbf{x}^j\}_L$, where the linearized variables are represented by $\mathbf{w}$. All other levels are computed as follows.

Computation of the Level- d RLT of Equation (5)

Step 1. Reformulation: For each $J \subseteq N$ having $|J| = d$, compute the Kronecker products of the $\prod_{j \in J} k_j$ nonnegative functional factors found within $\otimes_{j \in J} C_j \mathbf{x}^j$ with every inequality defining $\mathbf{X}$ of Equation (5), except for those $C_k \mathbf{x}^k \geq \mathbf{0}$ having $k \in J$. (For $d = n$, include $\otimes_{j \in N} C_j \mathbf{x}^j \geq \mathbf{0}$ to replace the multiplication of $\otimes_{j \in J} C_j \mathbf{x}^j$ with the expressions $C_j \mathbf{x}^j \geq \mathbf{0}$ since $J = N$.) Replace all the constraints found within Equation (5) with these newly defined restrictions. Use LIP Property 2 to set $L_{jk}(x_j)x_j = L_{jk}(x_j)\theta_{jk}$ for all (j, k) with $j \in N$ and $k \in K_j$.

Step 2. Linearization: Substitute a continuous variable for each distinct nonlinear term in the products $\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{x}$ and $\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{y}$ present in the reformulation. Denote the linearized vectors as $\{\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{x}\}_L$ and $\{\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{y}\}_L$ respectively, with the newly defined variables in these two sets given by $\mathbf{w}$ and $\mathbf{v}$. Call the resulting polyhedral set in the variables $(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{v})$ as $\mathbf{X}_d$.

Three comments are warranted. First, the nonnegativity of the $\prod_{j \in J} k_j$ functions found in $\otimes_{j \in J} C_j \mathbf{x}^j$, and used in the reformulation step, is due to LIP Property 1. Second, for each $J \subseteq N$ with $|J| = d$, the reformulation does not multiply the products $\otimes_{j \in J} C_j \mathbf{x}^j$ by any $C_k \mathbf{x}^k \geq \mathbf{0}$ having $k \in J$ since the resulting linearized inequalities would be implied by $\mathbf{X}_d$. This is due to LIP Property 3 reducing such (nontrivial) multiplications to $\otimes_{j \in J} C_j \mathbf{x}^j \geq \mathbf{0}$. For $d \leq n - 1$, these inequalities can be shown implied by any products of the form $\otimes_{j \in J \cup k} C_j \mathbf{x}^j \geq \mathbf{0}$ for which $k \notin J$, with the linearized version of these latter products found within $\mathbf{X}_d$. For $d = n$, all $\prod_{j \in N} k_j$ such inequalities with $J = N$ are enforced. Third, the set $\mathbf{X}_n$ does not need to include these $\prod_{j \in N} k_j$ inequalities $\otimes_{j \in N} C_j \mathbf{x}^j \geq \mathbf{0}$ if there exist any continuous variables $\mathbf{y}$ within the problem since the multiplication of such products with the inequalities $\mathbf{L} \leq \mathbf{y} \leq \mathbf{U}$ would deem them redundant. (Note the similarity of the last two comments to the redundancy of certain nonnegative

functional factors $\prod_{j \in J_1} x_j \prod_{j \in J \setminus J_1} (1 - x_j) \geq 0$ for the mixed-binary case when $J_1 \subseteq J \subseteq N$.)

The RLT hierarchy for mixed-discrete programs is a generalization of that for mixed-binary problems both in theory and in construction. Paralleling the mixed-binary case, two important theoretical properties hold true. First, for each $d \in N$, every feasible $(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{v})$ to $\mathbf{X}_d$ with $x_j \in S_j$ and $\{\mathbf{x}^j\}_L = \mathbf{x}^j$ for all $j \in N$ will have the variables $\mathbf{w}$ and $\mathbf{v}$ equal to their intended products. This gives a hierarchy of equivalent mixed-integer linear reformulations of $\mathbf{X}$ in Equation (1) for general sets S_j having $k_j \geq 2$. Second, for each $d \in \{0, 1, \dots, n\}$, letting

$$P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_d) \equiv \{(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{v}) \in \mathbf{X}_d \text{ for some } (\mathbf{w}, \mathbf{v})\}$$

denote the projection of the set $\mathbf{X}_d$ onto the space of the variables $(\mathbf{x}, \mathbf{y})$ as in Equation (2), it also follows that

$$P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_0) \subseteq P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_1) \subseteq \dots \subseteq P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_n) = \text{conv}(\mathbf{X}),$$

as in Equation (3). Hence, the highest level provides an explicit algebraic representation of the convex hull of feasible solutions in the higher-dimensional space $\mathbf{X}_n$.

A subtle difference between the mixed-binary and mixed-discrete cases is that the latter includes the restrictions $C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}$ for all $j \in N$, linearized from Equation (5), within $\mathbf{X}_0$. This has $\mathbf{X}_0$ containing variables $\mathbf{w}$, so that it lies in a different variable space than $P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_0)$. Hence, unlike Equation (3), $\mathbf{X}_0 \neq P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_0)$ for the discrete case. However, as the projection of the inequalities $C_j \{\mathbf{x}^j\}_L \geq \mathbf{0}$ for all $j \in N$ collectively reduce to $\theta_{j1} \leq x_j \leq \theta_{jk_j}$ for all $j \in N$, the set $P_{(\mathbf{x}, \mathbf{y})}(\mathbf{X}_0)$ is $\mathbf{X}$ of Equation (5) without the inequalities $C_j \mathbf{x}^j \geq \mathbf{0}$ for all $j \in N$, and with the $x_j \in S_j$ restrictions replaced by $\theta_{j1} \leq x_j \leq \theta_{jk_j}$ for all $j \in N$. This is the usual continuous relaxation of such a mixed-discrete set.

Relative to construction, the mixed-discrete hierarchy reduces to the mixed-binary hierarchy when the sets S_j , for all $j \in N$, in the former are restricted to have

$k_j = 2$ (so that $K_j = \{1, 2\}$) with $\theta_{j1} = 0$ and $\theta_{j2} = 1$. With this restriction, each $C_j \mathbf{x}^j$ reduces to $1 - x_j$ and x_j , and the substitution $L_{jk}(x_j)x_j = L_{jk}(x_j)\theta_{jk}$ for all (j, k) with $j \in N$ and $k \in K_j$ in the reformulation step becomes $x_j^2 = x_j$ for all $j \in N$. The linearization step has the replacement of $\mathbf{w}$ for $\{\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{x}\}_L$ and $\mathbf{v}$ for $\{\otimes_{j \in J} \mathbf{x}^j \otimes \mathbf{y}\}_L$ reducing to the simpler substitutions $w_J = \prod_{j \in J} x_j$ and $v_{Jk} = y_k \prod_{j \in J} x_j$ found in the mixed-binary case.

The example below illustrates the increasing strength of the hierarchical levels.

Example 3. Let $\mathbf{X}$ of (5) be given by $\mathbf{X} = \{(x_1, x_2) \in \mathbb{R}^2 : 2x_1 - 2x_2 \geq -3, -2x_1 + 2x_2 \geq -3, -2x_1 - 2x_2 \geq -11, 2x_1 + 2x_2 \geq 5, C_j \mathbf{x}^j \geq \mathbf{0} \text{ for } j \in N = \{1, 2\}, x_1 \in S_1 = \{1, 2, 3\}, x_2 \in S_2 = \{1, 2, 3\}, \text{ with } C_j \mathbf{x}^j \geq \mathbf{0} \text{ expressed as}$

$$\begin{bmatrix} 3 & -\frac{5}{2} & \frac{1}{2} \\ -3 & 4 & -1 \\ 1 & -\frac{3}{2} & \frac{1}{2} \end{bmatrix} \begin{pmatrix} 1 \\ x_j \\ x_j^2 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

for $j = 1, 2$.

Then $S_j = \{1, 2, 3\}$ with $k_j = 3$ for $j \in N = \{1, 2\}$, and there exist no continuous variables $\mathbf{y}$. The level-0 RLT projected onto the space of the variables (x_1, x_2) is the region defined in Fig. 2 by the solid lines, with the eight extreme points: $(1, \frac{3}{2}), (1, \frac{5}{2}), (\frac{3}{2}, 3), (\frac{5}{2}, 3), (3, \frac{5}{2}), (3, \frac{3}{2}), (\frac{5}{2}, 1), (\frac{3}{2}, 1)$. The projected level-1 RLT is defined by the dashed lines, with the eight extreme points: $(1, 2), (\frac{4}{3}, \frac{8}{3}), (2, 3), (\frac{8}{3}, \frac{8}{3}), (3, 2), (\frac{8}{3}, \frac{4}{3}), (2, 1), (\frac{4}{3}, \frac{4}{3})$. The projected level-2 RLT is the shaded region with the four extreme points:

$(1, 2), (2, 3), (3, 2), (2, 1)$, which is the convex hull of feasible points.

As with the mixed-binary case, partial convex hull characterizations can be constructed for the mixed-discrete case. The variable x_2 can be treated as continuous by considering it to lie within $\mathbf{y}$. Then the $C_2 \mathbf{x}^2 \geq \mathbf{0}$ restrictions simplify to $1 \leq x_2 \leq 3$. The projected level-0 RLT remains unchanged, and the projected level-1 RLT gives $\text{conv}(\mathbf{X} \cap x_1 \in S_1)$. Figure 3 gives both projected regions, with level-0 defined by the solid lines and level-1 depicted by the shaded region. The six extreme points to this shaded region are $(1, \frac{3}{2}), (1, \frac{5}{2}), (2, 3), (3, \frac{5}{2}), (3, \frac{3}{2}), (2, 1)$, with each such point having $x_1 \in S_1$.

The RLT constructs for mixed-discrete programs are sufficiently flexible to allow enhancements similar to those for the mixed-binary case. Mixed-discrete *polynomial* programs that are linear in the continuous variables $\mathbf{y}$ for fixed realizations of $\mathbf{x}$ can be readily handled by the RLT. In this case, a hierarchy of polyhedral approximations leading to the convex hull representation is again available, but each hierarchical level d must be large enough to accommodate the nonlinear terms. Problems having products of the continuous variables $\mathbf{y}$ can also be linearized; however, as with the mixed-binary case, convex hull representations are not necessarily available.

The conditional logic arguments [5,6] that tighten inequalities for mixed-binary problems are also applicable to mixed-discrete

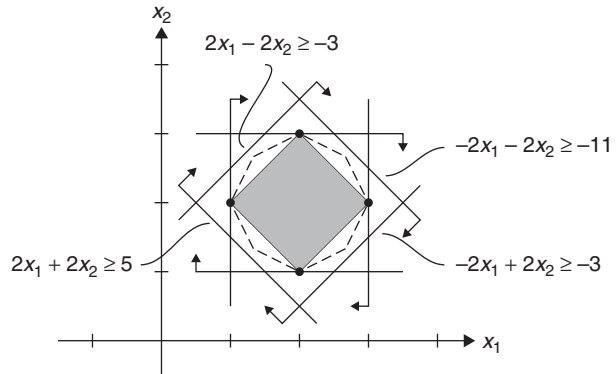


Figure 2. Projected RLT regions for x_1, x_2 discrete.

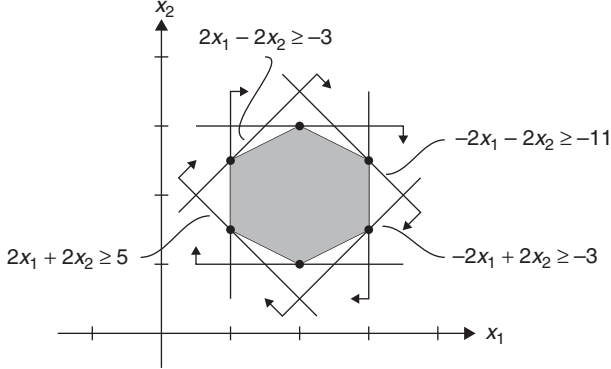


Figure 3. Projected RLT region for x_1 discrete and x_2 continuous.

programs, in a modified form. Again, the nonlinear nature of the constructed inequalities affords additional strength. To see this, consider a discrete variable x_j that can realize values in the set $S_j \equiv \{\theta_{j1}, \dots, \theta_{jk_j}\}$. Multiplication of any inequality $\alpha x + \beta y \geq b$ defining $\mathbf{X}$ in Equation (5) with the LIP $L_{jt}(x_j)$ for some $t \in K_j$ yields the quadratic inequality

$$L_{jt}(x_j)[\alpha x + \beta y - b] \geq 0.$$

Now, suppose the “condition” that $x_j = \theta_{jt}$ implies $\alpha x + \beta y \geq b'$ for $b' > b$ holds true. Then this inequality can be “logically” modified to

$$L_{jt}(x_j)[\alpha x + \beta y - b'] \geq 0,$$

which affords a strengthened form in the RLT linearization step. This strengthened inequality follows since $L_{jt}(x_j) = 0$ if $x_j \neq \theta_{jt}$ by LIP Property 1 of the section titled “Lagrange Interpolating Polynomials,” and $L_{jt}(x_j) = 1$ with $\alpha x + \beta y - b' \geq 0$ if $x_j = \theta_{jt}$, also by Property 1.

ONGOING RESEARCH

The use of LIPs, embedded within an RLT framework, provides the theoretical foundation for generating tight polyhedral outer approximations and convex hull representations of mixed-discrete sets. Such representations exhibit considerable promise for more effectively solving combinatorial

optimization problems. Accompanying this promise, however, are theoretical and computational challenges that stem from the large linear programming representations generated at the different RLT levels. Explorations are beginning into the effective handling of these forms. To this author’s knowledge, no computational experience has yet been reported on using LIPs to solve mixed-discrete programs with general integer variables.

Based on the insights and lessons learned from studies on the mixed-binary RLT, four research directions arise. The RLT representations grow in size as progression is made up through the hierarchy. However, as noted for the mixed-binary case, even the weakest level-1 form provides very tight representations that tend to dominate alternate bounding mechanisms. Moreover, this form has an inherent, block-diagonal, decomposable structure. Preliminary studies show that these same two properties of strength and structure hold true for the discrete case. A research direction is to determine means for most effectively exploiting the level-1, mixed-discrete forms.

A second promising direction is to reduce the sizes of the higher-dimensional RLT representations by recasting them onto lower-dimensional spaces through the use of projection operators. Given a mixed-discrete problem, since the highest level RLT form provides a description of the convex hull, each valid linear inequality in the original problem space must be theoretically computable via a projection.

Of course, obtaining all facets for a given problem reduces to enumerating the extreme directions of a suitably defined projection cone. The identification of all such directions is the outstanding challenge, both for general problem classes and specific instances having special structures. Efforts in this vein are underway, with initial results in Adams and Henry [22] and Henry [23]. Notably, RLT forms other than the highest level can also be used to motivate tight approximations in the original variable space.

A third direction, which closely relates to the projections mentioned above, is the generation of cutting planes. Partial convex hull representations play a critical role here, as they promote considerably smaller problems than the full RLT forms. Specifically, suppose some linear function in the discrete variables $\mathbf{x}$ and continuous variables $\mathbf{y}$ is optimized as a linear program over the set $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$ which, as described earlier, is $\mathbf{X}$ of Equation (5) without the inequalities $C_j \mathbf{x}^j \geq \mathbf{0}$ for all $j \in N$, and with the $x_j \in S_j$ restrictions replaced by $\theta_{j1} \leq x_j \leq \theta_{jk_j}$ for all $j \in N$. Further, suppose that the computed extreme point optimal $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ to $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$ does not have $\hat{x}_j \in S_j$ for all $j \in N$ (since otherwise $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ would solve the mixed-discrete program). Select any $t \in N$ such that $\hat{x}_t \notin S_t$, and consider a modified version of $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$, say $\mathbf{X}'$, that is obtained by treating all variables x_j with $j \in N \setminus \{t\}$ as continuous and *only* the variable x_t as discrete. This is accomplished by inserting the k_t linear inequalities $C_t \{\mathbf{x}^t\}_L \geq \mathbf{0}$ within $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$ to replace the $\theta_{t1} \leq x_t \leq \theta_{tk_t}$ bounding restrictions, and by treating all remaining variables $j \in N \setminus \{t\}$ to lie within $\mathbf{y}$. Now, calling the associated level-1 RLT form of $\mathbf{X}'$ as $\mathbf{X}'_1$, since $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}'_1) = \text{conv}(\mathbf{X} \cap x_t \in S_t)$, the point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ cannot be feasible to $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}'_1)$. Then a linear inequality, or cutting plane, in $(\mathbf{x}, \mathbf{y})$ which is valid for $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$ and separates the set $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$ from the point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ can be generated as a surrogate of the constraints defining $\mathbf{X}'_1$. This “projected” cut can then be included within the inequalities defining $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$, and the linear problem resolved with this additional restriction included in $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$.

There is a great deal of flexibility in computing cutting planes. The specific form of the cut generated to separate $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ from $P_{(\mathbf{x},\mathbf{y})}(\mathbf{X}_0)$

depends on the choice of surrogate multipliers. In fact, many such cuts can be derived. The selection of the index $t \in N$ having $\hat{x}_t \notin S_t$ can also be significant. If desired, more than one fractional variable can be chosen, so that the cut generation employs an explicit representation of $\text{conv}(\mathbf{X} \cap x_j \in S_j \forall j \in J)$, for some set $J \subseteq N$ having $|J| \geq 2$. The trade-off here is a tightening of the cuts at the expense of larger, level- $|J|$ RLT forms. Also, exploitation of mathematical structures to obtain simpler RLT forms appears promising. As mentioned earlier, a successful implementation of the special case dealing with mixed-binary programs is found in Balas *et al.* [10].

A fourth direction is to use the convex hull representations available at the highest level as a theoretical tool for making inferences about optimal solutions to mixed-discrete programs from optimal solutions to linear programming relaxations. From the section titled “RLT for Mixed-Binary Sets,” the works [11,12] identified conditions under which variables realizing binary values in an optimal solution to the continuous relaxation of a binary program will persist in retaining these same values in an optimal solution to the discrete problem. This was accomplished using inductive arguments that implicitly climb the RLT hierarchical ladder, without explicitly writing these formulations. Similar approaches may prove fruitful for the mixed-discrete case.

Acknowledgments

This research was supported in part by the National Science Foundation under grant number CMMI-0968909.

REFERENCES

1. Adams WP, Sherali HD. A hierarchy of relaxations leading to the convex hull representation for general discrete optimization problems. *Ann Oper Res* 2005;140(1):21–47.
2. Sherali HD, Adams WP. A reformulation-linearization technique for solving discrete and continuous nonconvex problems. Dordrecht, Boston (MA), London: Kluwer Academic Publishers; 1999.

3. Sherali HD, Adams WP. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J Discrete Math* 1990;3(3):411–430.
4. Sherali HD, Adams WP. A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems. *Discrete Appl Math* 1994;52(1):83–106. (Manuscript 1989).
5. Sherali HD, Adams WP, Driscoll PJ. Exploiting special structures in constructing a hierarchy of relaxations for 0–1 mixed integer problems. *Oper Res* 1998;46(3):396–405.
6. Lougee-Heimer R, Adams WP. A conditional logic approach for strengthening mixed 0–1 linear programs. *Ann Oper Res* 2005;139(1):289–320.
7. Sherali HD, Adams WP. A reformulation-linearization technique (RLT) for semi-infinite and convex programs under mixed 0–1 and general restrictions. *Discrete Appl Math* 2009;157(6):1319–1333.
8. Sherali HD, Tuncbilek C. A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. *J Glob Optim* 1992;2(1):101–112.
9. Sherali HD, Alameddine A. A new reformulation-linearization technique for bilinear programming problems. *J Glob Optim* 1992;2(4):379–410.
10. Balas E, Ceria C, Cornuéjols G. A lift and project cutting plane algorithm for mixed 0–1 programs. *Math Program* 1993;58(3):295–324.
11. Adams WP, Lassiter JB, Sherali HD. Persistency in 0–1 polynomial programming. *Math Oper Res* 1998;23(2):359–389.
12. Lassiter JB. Persistency in 0–1 optimization [PhD dissertation]. Clemson (SC). Department of Mathematical Sciences, Clemson University; 1993 May.
13. Adams WP, Sherali HD. A tight linearization and an algorithm for zero-one quadratic programming problems. *Manage Sci* 1986;32(10):1274–1290.
14. Adams WP, Sherali HD. Linearization strategies for a class of zero-one mixed integer programming problems. *Oper Res* 1990;38(2):217–226.
15. Adams WP, Sherali HD. Mixed-integer bilinear programming problems. *Math Program* 1993;59(3):279–305.
16. Johnson TA. New linear programming-based solution procedures for the quadratic assignment problem [PhD. dissertation]. Clemson (SC): Department of Mathematical Sciences, Clemson University; 1992 Dec.
17. Adams WP, Johnson TA. Improved linear programming-based lower bounds for the quadratic assignment problem. *DIMACS Ser Discrete Math Theor Comput Sci* 1994;16:43–76.
18. Hahn P, Hightower W, Johnson TA, *et al.* Tree elaboration strategies in branch and bound algorithms for solving the quadratic assignment problem. *Yugosl J Oper Res* 2001;11(1):41–60.
19. Adams WP, Guignard M, Hahn PM, *et al.* A level-2 reformulation-linearization technique bound for the quadratic assignment problem. *Eur J Oper Res* 2007;180(3):983–996.
20. Sherali HD, Adams WP. Computational advances on using the reformulation-linearization technique (RLT) to solve various discrete and continuous nonconvex programming problems. *Optima Math Program Soc Newsl* 1996;49:1–6.
21. Horn RA, Johnson CR. *Matrix analysis*. Cambridge, New York: Cambridge University Press; 1985.
22. Adams WP, Henry SM. On polytopes associated with products of discrete variables. Presentation at the INFORMS National Meeting. San Diego (CA): 2009.
23. Henry SM. Tight polyhedral representations of discrete sets using projections, simplices, network decompositions, and base-2 expansions [PhD. dissertation tentative title]. Clemson (SC): Department of Mathematical Sciences, Clemson University; 2011 Aug. (expected).

USING HOLISTIC MULTICRITERIA ASSESSMENTS: THE CONVEX CONES APPROACH

ÖZLEM KARSU

London School of Economics and
Political Science, London,
United Kingdom

INTRODUCTION

Consider a general multicriteria decision-making (MCDM) problem that can be formulated as follows:

$$\begin{aligned} \text{"Max"} \quad & z = f(x) = (f_1(x), \dots, f_p(x)) \quad (1) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

where x is the *decision vector*, $X \subseteq \mathbb{R}^n$ is the *feasible decision space*, $f_j(\cdot)$ is the j th *criterion (objective) function*, and z is the *criterion vector*. The above formulation uses the decision space representation of an MCDM problem. One can also formulate MCDM problems in the criterion space as follows:

$$\begin{aligned} \text{"Max"} \quad & \{z_1, \dots, z_p\} \quad (2) \\ \text{s.t.} \quad & z \in Z \end{aligned}$$

where $Z = \{z \in \mathbb{R}^p : z = f(x) : x \in X\}$ is called the (*feasible*) *criterion space*. That is, Z is the image of the feasible decision set (X) in the criterion space. Throughout the text, we refer to both x and z as the solutions (alternatives) of the MCDM problem.

A classification of the MCDM problems can be made based on whether the solutions (alternatives) are explicitly or implicitly defined. Problems where a finite set of alternatives is explicitly given are called *multiple-criteria evaluation problems* (or *multicriteria evaluation problems*) and problems where the set of alternatives is implicitly defined by constraints are called *multiple-criteria design problems* or *multiple-objective (mathematical programming) problems* (MOPs) [1]. When

the problem considered is an MOP, $X \in \mathbb{R}^n$ can be discrete or continuous.

Note that the maximization of a vector in models 1 and 2 is not a well-defined operator. Therefore, solving an MCDM model may refer to different things depending on the context. Most MCDM approaches try to identify the best alternative, that is, to find the alternative that is most preferred by the decision maker (DM). Some other cases are also possible. For example, three kinds of *problematiques* are reported to be generally used in practice in order to support DMs in multicriteria evaluation problems [2, 3]. These are as follows:

1. identify the best alternative or a small subset of good alternatives;
2. rank the alternatives from the best to the worst;
3. classify/sort the alternatives into predefined homogeneous groups.

If the MCDM problem is an MOP, that is, the set of solutions is implicitly defined by constraints, the number of solutions can be infinite (in the continuous case) or prohibitively large (in the discrete case); hence, the ranking and sorting problematiques are not typically considered. In such cases, one may want to identify the best alternative or a small subset of good alternatives.

Unless there is a single alternative that is better than the others in terms of all the criteria, which typically is not the case when there are conflicting objectives, the need to distinguish different solutions from each other makes some information on the DM's preferences necessary. This makes MCDM theory closely connected to the theory on preference relations and utility. Basic notation and definitions used in the theory of preference relations are provided in the next section.

In this study, we provide a review of the literature on interactive MCDM approaches that use convex cones as a means of representing the DM's preference structure. In the next section, we discuss the relation

between preference relations and value (utility) functions. Specifically, we provide the assumptions made on the DM's underlying preference relation by the MCDM solution methods using the convex cones approach. In the section titled "Convex Cones and Polyhedra," we cover the basic theory on the convex cones approach and show how the cones can be used to obtain the best solution in a given set. We also discuss a related concept: the use of polyhedra whose vertices are the cone generators in MCDM ranking or sorting problems. We then provide a review of the MCDM approaches that use the convex cones and we conclude the discussion in the section titled "Conclusion."

PREFERENCE RELATIONS AND VALUE FUNCTIONS

We first define and discuss the properties of a weak preference relation, denoted by $\preceq$, which completely characterizes the preference model in the criterion space. Next, we introduce the term *rational preference* along with its underlying properties and discuss the relation between preference relations and value functions. We then provide the assumptions on the DM's preference relation that allow us to use convex cones in MCDM approaches.

For two alternatives z^1 and z^2 , the statement $z^1 \preceq z^2$ is used to symbolize that z^2 is *weakly preferred* to z^1 .

Definition 1 Given a relation of weak preference, the corresponding relations of strict preference and indifference are defined as follows:

$$\begin{aligned} z^1 < z^2 & (z^2 \text{ is strictly preferred to } z^1) \\ \iff (z^1 \preceq z^2 \text{ and not } z^2 \preceq z^1) \\ z^1 \sim z^2 & (z^2 \text{ is indifferent to } z^1) \\ \iff (z^1 \preceq z^2 \text{ and } z^2 \preceq z^1). \end{aligned}$$

We now define some properties for a preference relation.

Definition 2 A preference relation $\preceq$ is complete if either $z^1 \preceq z^2$ or $z^2 \preceq z^1$, for all $z^1, z^2 \in Z$.

A preference relation $\preceq$ is transitive if $(z^1 \preceq z^2 \text{ and } z^2 \preceq z^3) \Rightarrow z^1 \preceq z^3$, for all $z^1, z^2, z^3 \in Z$.

Preference relation $<$ satisfies strict monotonicity (SM) if $z^1 < z^2$ then $z^1 < z^2$, for all $z^1, z^2 \in Z$.

The relation $\preceq$ is called a *rational preference relation* if it is complete, transitive, and its strict part ($<$) is strictly monotonic. On the basis of rational preferences, we can define (rational) dominance, which is the intersection relation of all rational preference relations [4].

Definition 3 For any two alternatives z^1 and z^2 , $z^1 <_r z^2$ (z^2 (rationally) dominates z^1) if $z^1_i \leq z^2_i$, for all $i \in \{1, 2, \dots, p\}$ where at least one strict inequality holds. $z^1 \preceq_r z^2$ (z^2 weakly dominates z^1) if $z^1_i \leq z^2_i$, for all $i \in \{1, 2, \dots, p\}$.

Having defined these preference relations, we can now discuss their relation with value functions.

Definition 4 The preference relation $\preceq$ is said to be represented by a value function $v(\cdot)$ if $\forall z^1, z^2, v(z^1) \leq v(z^2)$ if and only if $z^1 \preceq z^2$.

A preference relation is not necessarily representable by a value function. However, one can derive the conditions under which a preference relation is representable by a specified form of value function. We provide later one of the well-known results on representability by Debreu [5, 6].

Definition 5 A set A is closed if and only if it contains all of its limit points. A preference relation is continuous if for all $z \in Z$, $\{y \in Z \mid y \preceq z\}$ and $\{y \in Z \mid z \preceq y\}$ are closed.

The well-known representation theorem by Debreu [5, 6] is as follows.

Theorem 1 If a preference relation on a set $Z \subseteq \mathbb{R}^p$ is complete, transitive, and continuous, then it is representable by a continuous utility (value) function.

In an MCDM problem, if the DM has a rational preference relation representable by a value function, then this function is strictly increasing. Further structural assumptions on the value function imply further assumptions on the underlying preference relation of the DM.

MCDM solution strategies under the *value maximization approach* assume that the DM's preferences are representable by an underlying value function ($v : Z \rightarrow \mathbb{R}$) and involve maximizing this value function. Different forms of value functions have been studied such as linear [7, 8], quasiconcave [9], and monotonically increasing [10]. There are also algorithms for partially ranking [11] or sorting [12] alternatives based on an implicit quasiconcave value function assumption.

The MCDM approaches can also be categorized based on when the information is taken from the DM as follows [3], Chapter 16:

- *Methods based on the prior articulation of preferences:* In such methods, the preference information from the DM is taken at the beginning.
- *Methods based on the progressive articulation of preferences:* These approaches are called *interactive approaches*. In such methods, we iteratively reduce the solution space and approach the best solution (or a subset of good solutions), eliciting preference information from the DM at each iteration. Interactive value maximization strategies assume an implicit value function and employ an interactive search process that makes use of the structural assumptions on the value function (see Refs [13] and [14] for two reviews).
- *Methods based on the posterior articulation of preferences:* These methods try to find a good approximation of the non-dominated frontier and present it to the DM.

Preference information can be gathered in different forms. The DM can be asked to provide pairwise comparisons of alternatives, provide reference points [15], reference directions, trade-offs, information on strength of preferences [16], and so on.

A review of the literature on interactive MCDM approaches using convex cones as a means of representing the DM's preference structure is provided in this study. A convex cone is a convex polyhedral set in the objective function (criterion) space that contains solutions that are less preferred by the DM than a given set of solutions [9].

In most of the approaches using convex cones, the DM's preference model is assumed to be representable by a value function that is (strictly) quasiconcave and increasing. A function $g(\cdot)$ is strictly quasiconcave if for all $z^1, z^2 : z^1 \neq z^2$, and $\alpha \in [0, 1]$, we have $g(\alpha z^1 + (1 - \alpha)z^2) > \min\{g(z^1), g(z^2)\}$. Similarly, g is quasiconcave if $g(\alpha z^1 + (1 - \alpha)z^2) \geq \min\{g(z^1), g(z^2)\}$.

The quasiconcavity assumption of the DM's value function implies that the DM's preference relation is rational (i.e., complete, transitive, and strictly monotonic) with an additional convexity assumption.

Definition 6 A preference relation $\leq$ satisfies (weak) convexity if for all $z^1, z^2, z^3 \in Z$, such that $z^1 \leq z^2$ and $z^3 \in (z^1, z^2]$, we have $z^1 \leq z^3$. ($z^3 \in (z^1, z^2]$ means that there exists a real $\alpha, 0 < \alpha \leq 1$ such that $z^3 = \alpha z^1 + (1 - \alpha)z^2$. That is, z^3 is a convex combination of z^1 and z^2 .)

The quasiconcavity assumption for the value function corresponds to requiring the indifference curves (contours) to be convex to the origin. This assumption is quite reasonable, as it corresponds to decreasing marginal rate of substitution (DMRS), which is a commonly accepted property underlying consumer preferences in economics literature [17]. As the name implies, *marginal rate of substitution* is the maximum amount of one good a consumer would be willing to give up in order to obtain an additional unit of another. The marginal rate of substitution between two goods usually depends on the amount of goods the consumer currently has. For example, consider a case where the consumer has two types of goods: Good 1 and Good 2. When the consumer has a large amount of Good 1 and a very small amount of Good 2, he or she would probably be willing to give up quite a large amount

of Good 1 (as he or she already has plenty) to obtain one more unit of Good 2 (as Good 2 is scarce). In the opposite case where the consumer has a small amount of Good 1 and plenty of Good 2, he or she would probably be willing to give up only a very small amount of Good 1 to obtain more of Good 2. This is called the *DMRS*. As seen in the example, DMRS implies that, as the consumption of one good increases and the other decreases, the consumer would be willing to give up smaller quantities of the latter in exchange of a further unit of the former [17].

CONVEX CONES AND POLYHEDRA

We now provide a review of relevant results and studies from the literature on the use of convex cones in solving MCDM problems.

Basic Theory

We start with a discussion of two studies by Korhonen *et al.* [9] and Hazen [18], who introduce the basic theoretical results underlying the convex cones approach in independent works. The two results slightly differ in the underlying assumptions; hence, we will consider them separately. The main difference is that, the first assumes that a value function exists, whereas the second one relaxes this assumption. We start with the first results by Ref. [9] and then provide the more general case that is given by Ref. [18].

Assuming an Implicit Value Function. Given a set of k vectors, such that $z^1, \dots, z^k \in \mathbb{R}^p$, and an increasing quasiconcave function $f(\cdot)$ defined on $\mathbb{R}^p$, such that $f(z^k) < f(z^i)$ for all $i \neq k$, the following definitions are made:

Definition 7 We define the cone $C(z^1, \dots, z^{k-1}; z^k)$, where $z^i : i \neq k$ are the upper generators and z^k is the lower generator as follows: $C(z^1, \dots, z^{k-1}; z^k) = \{z \mid z = z^k + \sum_{i=1}^{k-1} \mu_i (z^i - z^k), \mu_i \geq 0\}$. The cone-dominated region of $C(z^1, \dots, z^{k-1}; z^k)$ is denoted by $CD(z^1, \dots, z^{k-1}; z^k)$ and defined as follows: $CD(z^1, \dots, z^{k-1}; z^k) = \{z' \mid z' \leq z \text{ where } z \in C(z^1, \dots, z^{k-1}; z^k)\}$.

Lemma 1 For any $z \in C(z^1, \dots, z^{k-1}; z^k)$, $f(z) \leq f(z^k)$. Moreover, for any $z' \in CD(z^1, \dots, z^{k-1}; z^k)$, $f(z') \leq f(z^k)$.

Definition 8 We define the polyhedron spanned by the vectors $z^1, \dots, z^k$ as follows: $P(z^1, \dots, z^k) = \{z \mid z = \sum \mu_i z^i, \sum \mu_i = 1, \mu_i \geq 0 \text{ for all } i\}$. The upper side of $P(z^1, \dots, z^k)$ is denoted by $UP(z^1, \dots, z^k)$ and defined as follows: $UP(z^1, \dots, z^k) = \{z' \mid z \leq z', \text{ where } z \in P(z^1, \dots, z^k)\}$.

Lemma 2 For any $z \in P(z^1, \dots, z^k)$, $f(z^k) \leq f(z)$. Moreover, for any $z' \in UP(z^1, \dots, z^k)$, $f(z^k) \leq f(z')$.

Assume that we have k points such that $z^1, \dots, z^k \in \mathbb{R}^p$ and $z^k < z^i$ for all $i \neq k$, where $<$ is the DM's preference relation. We assume that $<$ is a rational preference relation. If the DM has a quasiconcave value function $f(\cdot)$, by Lemma 1, for any $z \in C(z^1, \dots, z^{k-1}; z^k)$, we have $f(z) \leq f(z^k)$, hence $z \leq z^k$. Moreover, for each point $z' : z' \leq z$, where $z \in C(z^1, \dots, z^{k-1}; z^k)$, $z' \leq z^k$, that is, the points in the cone-dominated region will be at most as preferred as z^k . We will call each such point $z' \in CD(z^1, \dots, z^{k-1}; z^k)$ *cone dominated*. Note that if we assume strict quasiconcavity, we have $z < z^k$ for $z \neq z^k$ and $z' < z^k$ for $z' \neq z^k$.

Similarly, by Lemma 2, for any $z \in P(z^1, \dots, z^{k-1}, z^k)$, $f(z^k) \leq f(z)$, hence $z^k \leq z$. Moreover, for each point, $z' : z \leq z'$, where $z \in P(z^1, \dots, z^k)$, $z^k \leq z'$, that is, the points lying on the upper side of the polyhedron spanned by the k points will be at least as preferred as the lower generator, z^k . If we assume strict quasiconcavity, we have $z^k < z$ and $z^k < z'$.

Figure 1 shows an example of a 2-point cone in $\mathbb{R}_+^2$. The thick solid line is the cone $C(z^1; z^2)$ and the area with diagonal gray lines is the cone-dominated region, $CD(z^1; z^2)$. The line segment between z^1 and z^2 is the polyhedron $P(z^1, z^2)$ and the region above this line segment is the upper side of the polyhedron, $UP(z^1, z^2)$.

Figure 2 shows an example of a 3-point cone in $\mathbb{R}_+^2$. The region filled with the diagonal lines is $C(z^2, z^3; z^1)$ and the gray region

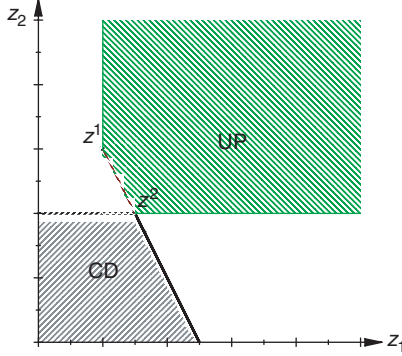


Figure 1. Illustration of a 2-point cone in $\mathbb{R}^2_+$.

including the dark gray part and the diagonal lines is the cone-dominated region. We refer the reader to Fig. 1.b in Ref. [19] for an illustration of a 3-point cone in $\mathbb{R}^3$.

We will visualize the additional information implied by the quasiconcavity using an example. Recall that a quasiconcave value function implies the convexity assumption on the DM's preference relation. See Fig. 3 for the criteria space of a bicriteria example where two of the alternatives (z^1 and z^2) are seen. Suppose that the DM has a rational preference relation that can be represented by a value function and prefers z^2 to z^1 ($z^1 < z^2$). By SM, one can say that the points in region A are less preferred to z^1 . By convexity (quasiconcavity), we gain information about region B, which is cone dominated; hence, the points in that region are less preferred to z^1 .

Similarly, using monotonicity, we can conclude that any alternative in regions C and D are preferred to z^1 . As $z^1 < z^2$, transitivity ensures that the points in region E, which are vector dominating z^2 , are also preferred to z^1 . Using convexity (quasiconcavity), we are able to say that the points in region F are preferred to z^1 . Therefore, we gain information about regions F and B by assuming a quasiconcave value function. Observe that the amount of the additional information gained depends on the two alternatives selected. One line of research in convex cones theory in MCDM focuses on finding smart ways of selecting the sets of alternatives to ask the DM.

A more General Result. Hazen [18] discusses the same results by showing that consideration of explicit responses (preference information) in the presence of increasing quasiconcave utility yields a stronger order than the strict componentwise vector order ($<$). He directly works on the DM's preference relation defined over the set of alternatives, rather than assuming that it is represented by a value function.

Hazen assumes the following axioms for the DM's preference relation.

For a relation $<$:

1. Irreflexivity (I): not $z < z$, for all $z \in \mathbb{R}^p$.
2. Transitivity (T): ($z^1 < z^2$ and $z^2 < z^3$) $\Rightarrow$ $z^1 < z^3$, for all $z^1, z^2, z^3 \in \mathbb{R}^p$.
3. Strict monotonicity (SM): $z^1 < z^2 \Rightarrow z^1 < z^2$, for all $z^1, z^2 \in \mathbb{R}^p$.
4. Weak convexity (WC): $z^1 \leq z^2$ and $z^3 \in (z^1, z^2] \Rightarrow z^1 \leq z^3$, for all $z^1, z^2 \in \mathbb{R}^p$.
5. Convexity (C): $z^1 < z^2$ and $z^3 \in (z^1, z^2] \Rightarrow z^1 < z^3$, for all $z^1, z^2 \in \mathbb{R}^p$.
6. Preference data (PD): $z^1 < z^2 < \dots < z^m$ is provided by the DM.

Note that if the binary relation $<$ is representable by a value function, WC and convexity axioms correspond to the quasiconcavity and strict quasiconcavity of the function, respectively.

The PD axiom stands for the preference information provided by the DM, which is in the form of ranking m alternatives.

Hazen uses the concept of *unanimity order* that is analogous to the dominance concept in MCDM and defines it as follows.

Definition 9 Assume that the DM's preference relation satisfies I, T, SM, WC, C, and PD; and denote the set of binary relations satisfying these properties with B . The unanimity order with respect to B is the binary relation $<^*$ over outcome vectors x and y and defined as follows: $x <^* y \iff x < y$ for all $< \in B$.

Note that rational dominance is another unanimity order where B is the set of binary relations satisfying axioms reflexivity, transitivity, and SM.

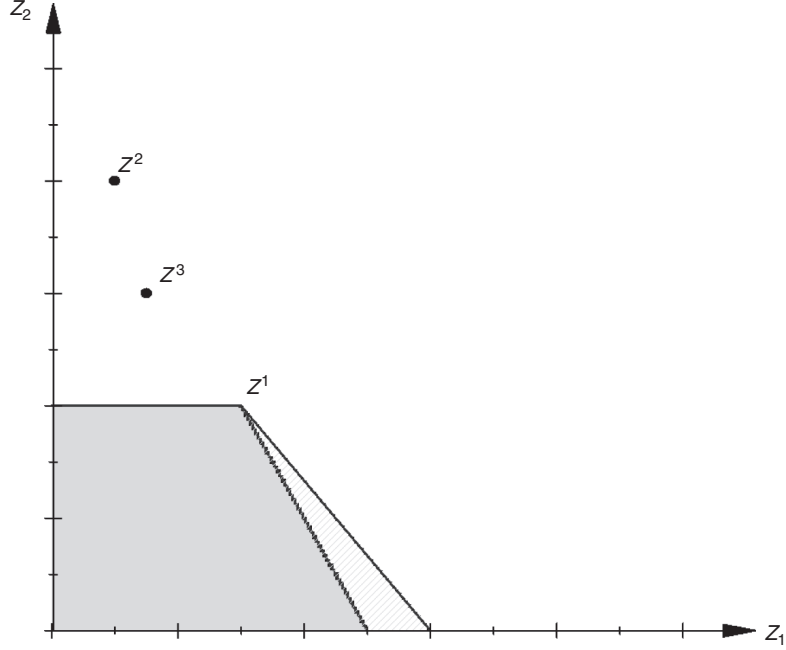


Figure 2. Illustration of a 3-point cone in $\mathbb{R}^2_+$.

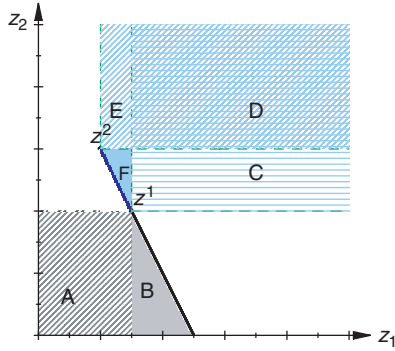


Figure 3. Using quasiconcavity.

The concept of unanimity order is important in MCDM, as it allows us to infer results without knowing the exact preference relation of the DM over the whole decision space. If we know that the preference relation of the DM satisfies a set of properties, any result that holds for the corresponding unanimity order will hold for the DM's preference relation.

Hazen's main result is the following theorem. Note that we changed the terminology to ensure consistency in presentation. (See the original paper for the original notation and the corresponding proof.)

Theorem 2 Suppose that $z^1, z^2, \dots, z^m : z^1 < z^2 < \dots < z^m$ are distinct elements of $\mathbb{R}^p$. Define the binary relation $<_m$ on $\mathbb{R}^p$ as follows: $x <_m y \iff x < y$ or $\exists j < m$ such that $x \in CD(z^{j+1}, z^{j+2}, \dots, z^m; z^j)$ and $y \in UP(z^{j+1}, z^{j+2}, \dots, z^m; x)$ and $y \neq x$. Then, $<^* = <_m$ unless $z^j \in P(z^{j+1}, z^{j+2}, \dots, z^m)$ for some j , in which case, $<^* = \mathbb{R}^p \times \mathbb{R}^p$ (in which case, the DM is inconsistent).

This theorem is valid for any case where we replace the componentwise strict vector partial order $<$ by any irreflexive conical order (see the original paper [18]). For an arbitrary vector space A , $<$ is a conical order if it is a binary relation on A represented by a convex cone K . That is, for all $x, y \in A$; $x < y \iff y - x \in K$.

Given PD $z^1 < z^2 < \dots < z^m$ and the assumptions on the DM's preference relation, Theorem 2 provides a necessary and

sufficient condition to conclude that a solution y is strictly preferred to solution x by the DM (without any further information): either $x < y$ or there is an alternative z^j in the given preference set, for which x lies in $CD(z^{j+1}, z^{j+2}, \dots, z^m; z^j)$ and y lies $UP(z^{j+1}, z^{j+2}, \dots, z^m; x)$.

When the DM is consistent, we have $x <^* y \iff x < y$ or $\exists j < m$, such that $x \in CD(z^{j+1}, z^{j+2}, \dots, z^m; z^j)$ and $y \in UP(z^{j+1}, z^{j+2}, \dots, z^m; x)$ and $y \neq x$.

It is easy to see that the sufficiency part of this proof holds as follows:

- If $x < y$, then $x <^* y$ (hence, $x < y$). This is trivial from SM.
- If $z \in CD(z^{j+1}, z^{j+2}, \dots, z^m; z^j)$ for some z^j , then $x < z^j$ by convexity, hence $x < z^{j+k}$ for $k = 1, \dots, m-j$ by transitivity. Therefore, any point $y : y \in UP(z^{j+1}, z^{j+2}, \dots, z^m; x)$ and $y \neq x$ will be strictly preferred to x because of convexity.

The necessary part of the statement is not obvious. By proving the necessary condition, Hazen shows that under given assumptions on the DM's preference relation, for any two solutions x and y , we cannot state that $x < y$; unless at least one of the two conditions is satisfied ($x < y$ or $\exists j < m$, such that $x \in CD(z^{j+1}, z^{j+2}, \dots, z^m; z^j)$ and $y \in UP(z^{j+1}, z^{j+2}, \dots, z^m; x)$ and $y \neq x$). This implies that under stated assumptions, by checking whether the conditions hold for two alternatives, a decision analyst will be making maximum use of the preference information available; if the conditions do not hold, he or she can be sure that it is not possible to conclude that one is preferred to the other without additional information.

This result is more general than the results provided in Lemmas 1 and 2, as the transitivity axiom allows indifference to be nontransitive, that is, it includes some cases where the DM's preference relation is not representable by a value function. Moreover, Theorem 2 applies to any irreflexive conical order. Therefore, Lemmas 1 and 2 are special cases of Theorem 2.

An Example. We will illustrate the results discussed so far using an example. We will show that under the same set of assumptions, the results provided in Lemmas 1 and 2 and Theorem 2 provide us the same information. In the example, we assume that the DM's preference structure is representable by a strictly quasiconcave value function and use the componentwise strict vector partial order $<$ for Theorem 2.

Figure 4 shows the criteria space of a bicriteria problem where the DM strictly prefers y to x , that is, $x < y$.

By Lemmas 1 and 2, we have the following:

1. $t \in CD(y; x)$, so $t < x$.
2. $k \in UP(y, x)$, hence $x < k$. We have $t < x < y$ and $t < x < k$. On the basis of the first ranking, one can generate $P(y, x, t)$.
3. Observing that $h \in P(y, x, t)$, we say that h is preferred to t . Therefore, we have $t < x < y, t < x < k$, and $t < h$.

According to Theorem 2, given $x < y$:

1. $t \in CD(y; x)$ and $x \in UP(y; t)$, hence $t < x$. Now we have $t < x < y$.
2. $x \in CD(y; x)$ and $k \in UP(y; x)$, hence $x < k$. We have $t < x < y$ and $t < x < k$.
3. Finally, $t \in CD(y; x)$ and $h \in UP(y; t)$, hence $t < h$. We obtain the same results: $t < x < y, t < x < k$, and $t < h$.

Operationalizing the Theory

In this section, we discuss how to operationalize the theory of convex cones. We would like to note here that the research seeking to operationalize the convex preference cone ideas originated from the team Korhonen, Wallenius, and Zionts, and their students and coworkers.

Checking Cone Dominance. The MCDM methods using convex cones gather preference information from the DM, usually in terms of pairwise comparisons. Using this information, more information is extracted about the other feasible alternatives that are not in the preference subset the DM provided.

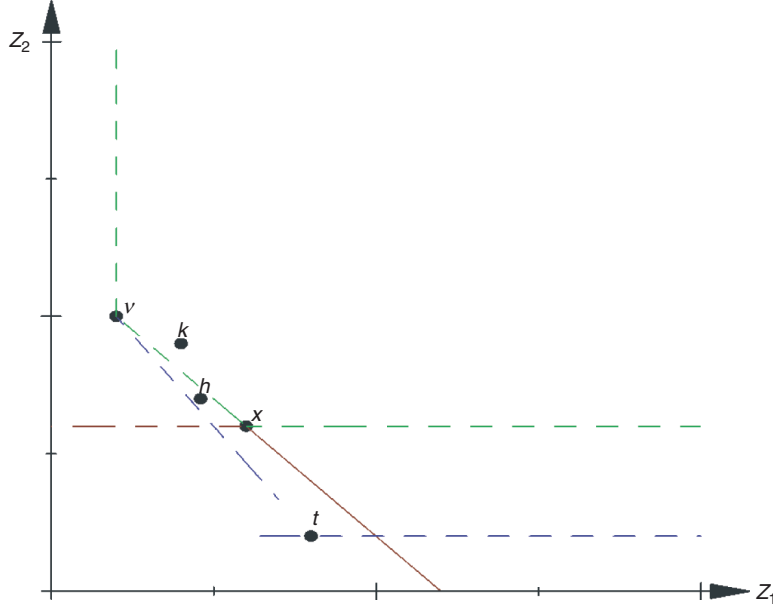


Figure 4. Example case.

Given preference information, one can generate cones and check for each candidate point whether it is in a cone-dominated region. If the objective is finding the best alternative out of a set, the cone-dominated alternatives can be eliminated from further consideration. If we have a ranking or sorting problem, the information that an alternative is inside or in the upper side of the corresponding polyhedron is also useful. Therefore, one can also check the status of an alternative with respect to the generated polyhedra based on the preference information.

We can perform these checks using linear programming (LP) problems as follows.

Suppose that we want to check whether alternative z is dominated by $C(z^1, \dots, z^{k-1}; z^k)$. Then, we solve the following LP:

$$\begin{aligned}
 & \text{Max } \epsilon & (LP_1) \\
 & \text{s.t. } \sum_{i=1}^{k-1} \mu_i (z^k - z^i) - \epsilon \geq z - z^k \\
 & \mu_i \geq 0 \text{ for } i = 1, \dots, k-1.
 \end{aligned}$$

Let us rewrite the first constraint set as follows: $z^k + \sum_{i=1}^{k-1} \mu_i (z^k - z^i) \geq z + \epsilon$. For each feasible μ_i value, the left-hand side of the constraint corresponds to a point in the cone. If z is not dominated by any of the points in the cone, the maximum value that ϵ can take should be negative. Otherwise, if $\epsilon^* \geq 0$, the first constraint set implies $z^k + \sum_{i=1}^{k-1} \mu_i (z^k - z^i) \geq z$. Then, for any nondecreasing quasiconcave function $f(\cdot)$, we have $f\left(z^k + \sum_{i=1}^{k-1} \mu_i (z^k - z^i)\right) \geq f(z)$. That is, z is at most as preferred as a point in the cone, that is, $z \in CD(z^1, \dots, z^{k-1}; z^k)$ and $f(z^k) \geq f(z)$ by Lemma 1.

To check whether $x \in UP(z^1, \dots, z^k)$, we solve the following LP:

$$\begin{aligned}
 & \text{Max } \epsilon & (LP_2) \\
 & \text{s.t. } \sum_{i=1}^k \mu_i z^i + \epsilon \leq x \\
 & \sum_{i=1}^k \mu_i = 1 \\
 & \mu_i \geq 0 \text{ for } i = 1, \dots, k.
 \end{aligned}$$

One can see that if $\epsilon^* \geq 0$, then $x \in UP(z^1, \dots, z^k)$.

In each iteration, we gain new information leading to new cones and polyhedra.

Tree Representation of Preferences and Size of Cones. Interactive approaches are based on gathering preference information from the DM throughout the process. Hence, such algorithms must keep track of the preference information gathered so far and generate the corresponding cones. A tree representation for preferences is suggested for this purpose in Ref. [20]. An example tree is given in Fig. 5, where each node represents an alternative and each arc represents the preference relation between these alternatives. Using the tree representation, it is possible to identify all distinct cones and polyhedra that can be generated, given preference information.

Note that the size of each cone as well as the number of total cones generated will affect the number of alternatives eliminated by the cones; hence, the amount of information required from the DM. The computational time will also be affected. Different strategies can be used as follows:

- One might choose to generate all the possible cones with maximum number of generators. That is, for each alternative that is a candidate to be a lower generator (less preferred to at least one solution in the tree), one can generate a cone using all the alternatives preferred to that alternative as the upper generators. By doing so, one will be making maximum use of the information provided by the DM and hence less information is required from the DM [9, 21]. In the example given in Fig. 5, this corresponds to generating $C(z^1, z^2, z^3, z^4, z^5; z^6)$, $C(z^1, z^2, z^3, z^4; z^5)$, and $C(z^1, z^2; z^3)$. This is called the *minimal set of cones*.
- As the earlier method may be cumbersome and may result in high computational effort, one might choose to generate smaller cones [9, 20]. In the literature, in most cases, only 2-point cones are used. In the example case in Fig. 5, the distinct 2-point

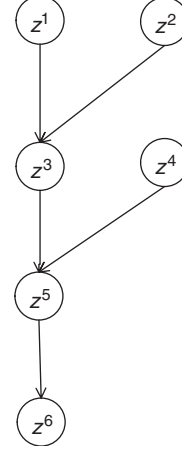


Figure 5. Tree representation.

cones are $C(z^1; z^3)$, $C(z^2; z^3)$, $C(z^3; z^5)$, $C(z^4; z^5)$, and $C(z^5; z^6)$. Note that it is also possible to generate more 2-point cones using transitivity: for example, $C(z^1; z^5)$, $C(z^1; z^6)$, $C(z^2; z^5)$, and $C(z^2; z^6)$. Korhonen *et al.* [9] report that in their experiments, the computation time saved in using 2-point cones instead of the minimal set of cones was minimal and suggest using the minimal set of cones. The use of cones with the most number of generators possible is suggested also in Ref. [19] owing to the decrease in the preference information required from the DM in that case.

Selecting the Candidate Solutions. Another issue to be considered in any interactive algorithm using convex cones is selecting the candidate solutions to be asked to the DM. The method used to select the alternatives depends on the characteristics of the problem. There are studies that use an estimated value function in an effort to select good candidates [9]. At each iteration, the parameters of the value function are updated and a solution that maximizes this value function is found and the DM is asked to compare this solution with its adjacent efficient alternatives (for an alternative z in the criterion space, adjacent efficient alternatives are the ones whose convex combinations with z are not dominated

by any other convex combination generated by the rest of the alternatives). There are also studies selecting the candidates based on their distance to an ideal point [22].

Interaction Type. The interaction type refers to the type of the questions asked to the DM.

In the solution frameworks designed to select the best alternative, the DM is usually asked for pairwise comparisons of alternatives. Other strategies are also possible such as providing the DM with m alternatives and asking him or her to select the best/worst alternative in the set or rank these alternatives.

In the solution frameworks designed for sorting alternatives, the DM can also be asked to assign some alternatives to the pre-determined groups. These alternatives are then used to generate cones and polyhedra.

Related Works in Literature

Many interactive MCDM methods using convex cones are proposed in the literature. Although not as common as convex cones, polyhedra are also used in sorting and partial ranking. The studies mainly differ in terms of the nature of the problem and the nature of the solution procedure; that is, based on the way to elicit the preference information, the way to select the alternatives that will be presented to the DM for information gathering, and the size of the generated cones.

The first problems considered are mostly multicriteria evaluation problems, where there is a finite set of alternatives. A series of papers that extend the application of convex cones to other problem types, for example, multicriteria linear problems (MCLP), multicriteria integer problems (MCIP), follow. There are also studies on increasing the efficiency of the solution procedures by finding ways to increase the region eliminated by cones and determining the number of the cone generators to be used. Recently, there has been an interest in using convex cones and polyhedra to sort and partially rank a finite set of alternatives.

We now review these applications of convex cones and polyhedra reported so far for MCDM problems.

Korhonen *et al.* [9] design an interactive algorithm for finding the best alternative in multicriteria evaluation problems. They use convex cones to represent the preference structure. They generate cones and eliminate the alternatives that are inferior to these cones. By doing so, they successively restrict the solution space and try to obtain the best alternative after a number of iterations. In order to determine the alternatives to be asked to the DM, they use a composite linear value function of the form $\sum \delta_j x_{ij}$, where $\delta_j > 0$ are multipliers and x_{ij} is the performance score of alternative i in criterion j . They find the alternative that maximizes this function and ask the DM to compare it with other selected alternatives (the adjacent efficient alternatives). They start with an arbitrary set of multipliers and use the DM's responses to update the (feasible) set of multipliers. Note that this method approximates a nonlinear value function by a composite linear value function; hence, it is possible that there are no set of multipliers consistent with the DM's preferences. This is handled by deleting the restriction on multipliers that correspond to the oldest information from the DM. In the proposed algorithm, the information from the DM is gathered in terms of pairwise comparisons. The authors conduct experiments and report statistics on the number of pairwise judgements the DM has to make for various problem sizes. They first try generating all the cones and then repeat the experiments using only 2-point cones. On the basis of the results, they conclude that the savings in the computational time when one is using only 2-point cones does not justify the increase the judgements DM has to make and hence suggest generating the minimal set of cones.

A few variations of the above algorithm is proposed by Köksalan *et al.* [23], Köksalan and Taner [24], and Köksalan and Sagala [25]. These studies address the same problem as in Ref. [9] and propose improvements in the solution algorithm. They suggest using dummy alternatives as one of the cone generators in order to increase the region eliminated by the convex cones. They discuss different ways to generate and select appropriate dummy alternatives. The authors report

improvements in results compared to the algorithm used in Ref. [9] in terms of the total number of pairwise comparisons required to find the best alternative. In Ref. [25], which generalizes the results of the previous two papers [23, 24], two dummy alternatives are used simultaneously, one of which is (hopefully) less preferred to a real alternative, so that it can be used as an upper generator of the cone. The other dummy alternative is generated as a potential lower generator. The dummy points are selected with the help of an estimated value function that is updated at each iteration.

Malakooti [26] also discusses different ways to select the cone generators as in Ref. [23], including the idea of using dummy alternatives to increase the area that is eliminated by the cone. He also discusses the use of local gradients, that is, trade-off information, at a point to eliminate worse alternatives and find better ones.

Ramesh *et al.* [20] study the underlying theory of the convex cones and the representation structure for the DM's preferences based on convex cones. They define rules to detect redundant cones, the cones that are already implied by other existing cones, in order to avoid unnecessary computations. They discuss two ways to represent the DM's preferences. The first representation scheme is an explicit representation, which is used in methods that use a single composite function of the multiple objectives. In these methods, the preferences are represented using linear inequalities on the weights of this composite function; thus, the set of feasible weights is successively reduced. The second one is based on the convex cones. They conclude that representation via cones is more accurate than representation via linear constraints on weights of the objectives. This is because by using a composite objective function, one fails to accurately represent nonlinear value functions. The authors conduct experiments on the use of convex cones in solving both multicriteria design [MCLP, MCIP (in the experiments bicriteria problems are solved)] and multicriteria evaluation problems and report the percentage of questions saved using convex cones.

Ramesh *et al.* [21] incorporate the convex cones representation method into the algorithm of Zionts and Wallenius [27] for multicriteria LP problems. The authors assume that the DM has a pseudoconcave value function. Convex cones are used to obtain an accurate and robust representation of the DM's preferences. The DM is asked for pairwise comparisons of alternatives, one of them is an alternative that maximizes a composite linear value function and the other is an adjacent efficient alternative. Only 2-point cones are generated.

Ramesh *et al.* [28] develop an algorithm for the MCIP, where the underlying value function of the DM is assumed to be pseudoconcave. They use the method previously proposed by Zionts and Wallenius [27] for MCLP in a branch and bound framework and use convex cones for the preference structure representation. The cones are used for fathoming candidate nodes in the branch and bound tree. They only generate 2-point cones and report computational results for bicriteria problems with up to 80 variables and 40 constraints.

Köksalan [22] proposes an interactive method using convex cones to identify and rank a most preferred subset of alternatives in multicriteria evaluation problems. The DM is asked for pairwise comparisons. Initially, candidate solutions are selected based on their distance to an ideal point and throughout the algorithm, new candidate solutions are generated using a weighted quadratic value function as in Ref. [23].

Taner and Köksalan [29] conduct an experimental study on how to determine the number of cone generators, select the cone generators, and determine the order of pairwise comparisons to ask the DM. They also propose an algorithm based on the results of the experiments. In the proposed algorithms, $m(m = 1, 2, \dots, 7)$ alternatives are selected by three methods, which use equal weighted linear, estimated linear, and quadratic value functions, respectively. Then, the least/most preferred alternative in the selected set is found by asking the DM a number of pairwise comparisons. On the basis of the preference information elicited

from the DM, all the possible cones are generated. They conclude that the version where $m = 3$ and the least preferred alternative is found provides the best results in terms of the average number of comparisons the DM has to make. They point out that the results of the experiments are not very conclusive; hence, a more elaborate and detailed study awaits further attention.

Prasad *et al.* [19] observe that in the computations, although most of the solutions are not cone dominated, quite a few of them are nearly cone dominated. Motivated by this observation, they introduce the concept of “near cone dominance” or “ p cone efficiency.” A p value is used to show how close an alternative is to being dominated by the generated cone. They suggest using this measure to choose the challengers of the incumbent that will be presented to the DM for pairwise comparison. That is, the alternatives are presented to the DM for comparison in an order based on how close they are to be inferior. This heuristic extension is suggested to accelerate the search by reducing preference information requirement. Another use of this idea is for early termination. That is, for an incumbent solution, the p cone efficiencies of its adjacent solutions are calculated and the algorithm is terminated if maximum of these p cone efficiencies is below a given threshold. They illustrate the idea by incorporating it within a solution framework for solving MOLP problems.

Ulu and Köksalan [12] propose interactive approaches to partition a set of discrete alternatives into acceptable and unacceptable sets. Assuming that the DM has a quasiconcave value function, they use the convex cones and polyhedra in a sorting algorithm. Note that the proposed sorting algorithm is for the special case where the number of classes is two, rather than for the general case with more than two classes.

Another study that includes a sorting approach based on convex cones is Fowler *et al.* [30]. They propose an evolutionary algorithm, a genetic algorithm, for MCDM problems assuming that the DM has a quasiconcave preference function. Genetic algorithm is a widely studied method for

finding heuristic solutions to NP-hard problems. In a genetic algorithm, the output is obtained by evolving an initial population of solutions through multiple generations by breeding and mutation. Hence, how the parents are selected becomes a critical issue to ensure offspring with good quality. The authors suggest partially sorting the population with the help of convex cones and polyhedra. The information taken from the DM is in terms of finding the best and worst in a given set of six solutions. Using this information, they generate four 2-point cones (consisting of the best as the upper generator and each of the other points except the worst) and one 6-point cone (having the worst as the lower generator and the others as upper generators). They report that using convex cones in sorting is effective and improves the output solution.

Recently, Dehnokhalaji *et al.* [11] propose an approach that uses convex cones and polyhedra to partially order a finite set of alternatives. Similar to the p cone efficiency concept discussed in Ref. [19], an alternative is classified as *surely better than*, *surely worse than*, or *possibly better/possibly worse than* the lower generator of the cone. This makes the suggested approach flexible in the sense that it can be used as an exact or approximate approach by adjusting some parameters. They generalize the idea used in Ref. [19] and employ it to obtain a strict partial order. This approach is suggested to be used to partially rank alternatives or as a supplementary method in other solution approaches such as evolutionary multiobjective optimization as discussed in Ref. [30]. Developing an interactive solution algorithm based on the approach is pointed out as a subject of future research.

Karsu *et al.* [31] use convex cones for solving MCDM problems involving equity concerns, which are encountered especially in public sector. They extend the current theoretical framework for convex cones by introducing the impartiality property to the preferences. On the basis of their theoretical findings, they propose an interactive ranking algorithm.

CONCLUSION

In this article, we consider the use of convex cones in interactive MCDM approaches, which is based on the DM's holistic assessments. We describe the assumptions on the DM's preference relation and the value function that allow us to implement the convex cones approach in solution algorithms for MCDM problems. We summarize the basic theoretical results from the literature, which show that convex cones can be used to iteratively approach the most preferred solution(s) by eliminating the ones that are less preferred. The results also show that, in multicriteria evaluation problems, polyhedra can be used alongside the convex cones for sorting or obtaining a quasi-ordering of the alternatives.

We provide a review of the studies that implement the convex cones approach over the last three decades. A large body of literature exists that use the convex cones approach in various algorithms designed to solve both multicriteria evaluation and multicriteria design problems. There are also studies on how to select good candidate solutions (cone generators) to be asked to the DM. We mention below a few challenges regarding convex cones approach awaiting further research.

- Further research can be performed on using convex cones to solve more difficult problems, such as multiobjective combinatorial optimization (MOCO) problems.
- Asking too many or relatively difficult questions increases the cognitive burden on the DM and may make the decision support system less attractive. Therefore, one needs to choose the type of interaction with the DM carefully. Further research can be conducted on finding the best way to minimize the cognitive burden while ensuring an effective decision support. Different types of questions can be asked to the DM. These include making pairwise comparisons, determining the best (worst) alternative in a given subset, or ranking a small subset of alternatives.

Experimental studies that compare these different interaction modes and discuss their advantages and disadvantages in different problem settings can be performed.

- As the size and number of cones used in an algorithm increase, more alternatives are eliminated but the computational time may increase as well. By generating merely 2-point cones, one may not be utilizing the available information fully. To the best of our knowledge, it is not studied in detail how well this approach works compared to generating the *minimal set of cones* in terms of the information required from the DM throughout the procedure. In other words, it is still an open question how to determine the cone sizes that best balance the computational time with the amount of information required from the DM for different problem settings.
- Using the information provided by the polyhedra in ranking and sorting environments is also a promising area. More research on the use of polyhedra and convex cones for sorting and ranking purposes awaits further attention.

REFERENCES

1. Köksalan M, Wallenius J. Multiple criteria decision making: Foundations and some approaches. In: Mirchandani P, editor. INFORMS TutORials in Operations Research. Hanover (MD): INFORMS; 2012. p 151–170.
2. Roy B. Problems and methods with multiple objective functions. *Math Program* 1971;1:239–266.
3. Figueira J, Greco S, Ehrgott M. Multiple criteria decision analysis: state of the art surveys. Boston (MA): Springer-Verlag; 2005.
4. Kostreva MM, Ogryczak W. Linear optimization with multiple equitable criteria. *RAIRO Oper Res* 1999;33:275–297.
5. Debreu G. Representations of a preference ordering by a numerical function. In: Thrall R, Coombs C, Davis R, editors. *Decision processes*. New York: Wiley; 1954. p 159–166.
6. Debreu G. Continuity properties of paretian utility. *Int Econ Rev* 1964;5(3):285–293.

7. Zionts S, Wallenius J. An interactive programming method for solving the multiple criteria problem. *Manage Sci* 1976;22(6):652–663.
8. Zionts S. A multiple criteria method for choosing among discrete alternatives. *Eur J Oper Res* 1981;7:143–147.
9. Korhonen P, Wallenius J, Zionts S. Solving the discrete multiple criteria problem using convex cones. *Manage Sci* 1984;30(11):1336–1345.
10. Köksalan MM, Sagala PNS. Interactive approaches for discrete alternative multiple criteria decision making with monotone utility functions. *Manage Sci* 1995;41(7):1158–1171.
11. Dehnokhalaji A, Korhonen P, Köksalan M, *et al.* Convex cone-based partial order for multiple criteria alternatives. *Decis Support Syst* 2011;51(2):256–261.
12. Ulu C, Köksalan M. An interactive procedure for selecting acceptable alternatives in the presence of multiple criteria. *Nav Res Logist* 2001;48:592–606.
13. Shin WS, Ravindran A. Interactive multiple objective optimization survey I-continuous case. *Comput Oper Res* 1991;18(1):97–114.
14. Alves MJ, Climaco J. A review of interactive methods for multiobjective integer and mixed-integer programming. *Eur J Oper Res* 2007;180:99–115.
15. Wierzbicki AP. Reference point approaches. In: Gal T, Stewart T, Hanne T, editors. *Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications*. Boston (MA): Kluwer Academic Publishers; 1999. p 9.1–9.39.
16. Argyris N, Morton A, Figueira JR. C(U){T}: a multi-criteria approach for non-additive concavifiable preferences. Technical report, LSEOR 12.134. London, UK: London School of Economics and Political Science; 2011.
17. Lipsey RG, Harbury CD. *First principles of economics*. 2nd ed. Oxford: Oxford University Press; 1992.
18. Hazen GB. Preference convex unanimity in multiple criteria decision making. *Math Oper Res* 1983;8(4):505–516.
19. Prasad SY, Karwan MH, Zionts S. Use of convex cones in interactive multiple objective decision making. *Manage Sci* 1997;43(5):723–734.
20. Ramesh R, Karwan MH, Zionts S. Theory of convex cones in multicriteria decision making. *Ann Oper Res* 1988;16:131–148.
21. Ramesh R, Karwan MH, Zionts S. Interactive multicriteria linear programming: an extension of the method of Zionts and Wallenius. *Nav Res Logist* 1989;36:321–335.
22. Köksalan MM. Identifying and ranking a most preferred subset of alternatives in the presence of multiple criteria. *Nav Res Logist* 1989;36:359–372.
23. Köksalan MM, Karwan MH, Zionts S. An improved method for solving multiple criteria problems involving discrete alternatives. *IEEE Trans Syst Man Cybern* 1984;14(1):24–34.
24. Köksalan MM, Taner O. An approach for finding the most preferred alternative in the presence of multiple criteria. *Eur J Oper Res* 1992;60:52–60.
25. Köksalan MM, Sagala P. An interactive approach for choosing the best of a set of alternatives. *J Oper Res Soc* 1992;43(3):259–263.
26. Malakooti B. A decision support system and a heuristic interactive approach for solving discrete multiple criteria problems. *IEEE Trans Syst Man Cybern* 1988;18(2):273–284.
27. Zionts S, Wallenius J. An interactive multiple objective linear programming method for a class of underlying nonlinear utility functions. *Manage Sci* 1983;29(5):519–528.
28. Ramesh R, Karwan MH, Zionts S. Preference structure representation using convex cones in multicriteria integer programming. *Manage Sci* 1989;35(9):1092–1104.
29. Taner OV, Köksalan MM. Experiments and an improved method for solving the discrete alternative multiple-criteria problem. *J Oper Res Soc* 1991;42(5):383–391.
30. Fowler JW, Gel ES, Köksalan MM, *et al.* Interactive evolutionary multi-objective optimization for quasi-concave preference functions. *Eur J Oper Res* 2010;206:417–425.
31. Karsu Ö, Morton A, Argyris N. Incorporating preference information in multicriteria problems with equity concerns. Technical report, LSEOR 12.136. London, UK: London School of Economics and Political Science; 2012.

USING OPERATIONS RESEARCH TO PLAN NATURAL GAS PRODUCTION AND TRANSPORTATION ON THE NORWEGIAN CONTINENTAL SHELF

VIBEKE STÆRKEBYE NØRSTEBØ
PETER SCHÜTZ
MARTE FODSTAD
LARS HELLEMO
KJETIL MIDTHUN
FRODE RØMO
ASGEIR TOMASGARD
Department of Applied Economics,
SINTEF Technology and Society,
Trondheim, Norway

INTRODUCTION

Since Norway started exporting natural gas in 1977, it has become the sixth largest producer and the third largest exporter of natural gas in the world. More than 30 production facilities in the North Sea send natural gas to receiving terminals in four European countries, covering approximately 15% of the total consumption in the European Union [1].

Almost all of the Norwegian natural gas is transported to the customers using a network of pipelines in the North Sea. With a length of more than 7800 km, it is the largest offshore pipeline network in the world, see Fig. 1. To ensure optimal operation of this network, the system has to be analyzed as a whole, including both production and transportation. Detailed knowledge of network integration, operational flexibility, pipeline flow and inventory, gas processing, and compressor station operation is required to analyze the behavior of the network. Variations in gas quality at the different production facilities and varying quality requirements from the customers add to the complexity of the system.

The petroleum industry has been an early adopter of operations research-based applications [3]. The Norwegian Petroleum

Directorate, for example, uses operations research for investment planning for oil and gas fields in the North Sea [4]. Other long-term planning models include designing and planning offshore field infrastructure under complex fiscal rules [5], and planning the development of offshore gas fields explicitly considering the uncertainty in reserves [6]. The upstream natural gas value chain, including production, transportation, processing, storage, and markets, is considered by Ulstein *et al.* [7], simultaneously optimizing all upstream activities on a tactical level. Pressure constraints in the transportation network, however, are not considered. Selot *et al.* [8] present an operational model for producing and routing natural gas. This model includes both a detailed description of the infrastructure and a complex contractual model for the markets.

The GassOpt model has been developed to help both natural gas producers and transportation network operators in planning their activities and maximize the value of their operations [9,10]. Given the production at the upstream facilities and the demand at the exit terminals, GassOpt determines the optimal routing, processing, and blending of natural gas ensuring that demand and quality constraints are satisfied. The development of the GassOpt decision support tool is described in more detail by Rømo *et al.* [9].

THE GASSOPT MODEL

The objective of the GassOpt model is to maximize the total flow of natural gas from the North Sea to the customers. In order to represent the real-world gas transportation problem correctly, the following elements have to be considered in the model formulation and are described hereafter: gas pressure and pipeline capacity, system effects, compressor stations, gas quality, and processing plants. Rømo *et al.* [9] and Tomasgard *et al.* [10] provide more details on how to model these elements.

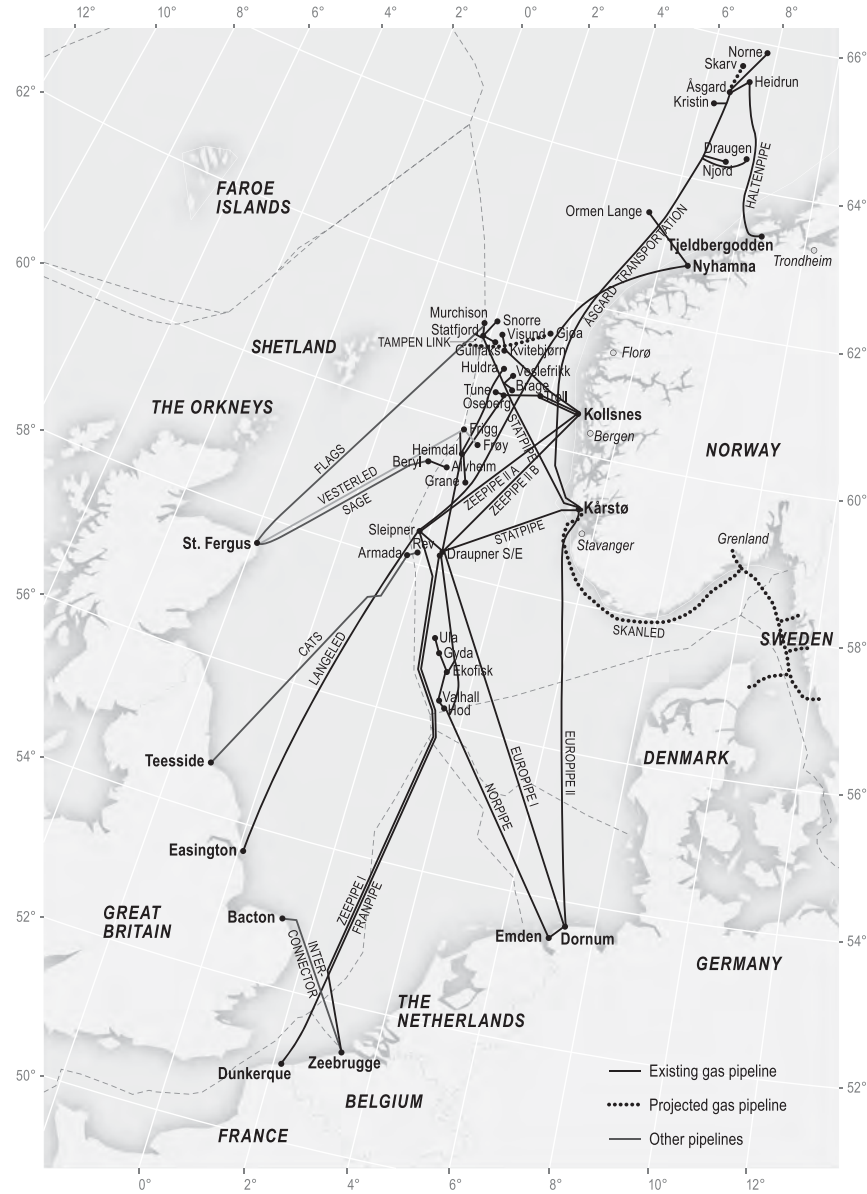


Figure 1. Natural gas production and transportation network [2].

Gas Pressure and Pipeline Capacity

The flow capacity of a pipeline depends not only on its physical properties but also on the pressure difference between inlet and outlet. Offshore gas pipelines are typically very long with high inlet pressure, large pressure drop,

and contracted pressure obligations at the pipeline outlet, making it crucial to describe the relation between gas flow q and pressure difference (between inlet pressure, p_i , and outlet pressure, p_o) correctly.

For long pipelines, this relation can be described by means of the nonlinear

Weymouth, Panhandle A, or Panhandle B equations. These equations are derived from the fundamental energy equation for compressible flow and developed for turbulent high pressure flow in long pipelines. The flow capacity of a pipeline can be calculated depending on the physical pipeline properties (such as length (L), diameter (D), friction (E), and ambient temperature (T)) and gas properties (such as the compressibility (Z)), as well as set of constants (c, n, u, x, y), which differ according to which of the equations (i.e., Weymouth, Panhandle A, or Panhandle B) is considered. For standard conditions (temperature $T_{\text{std}} = 288.15$ K and pressure $p_{\text{std}} = 1.01325$ bar), the flow rate q_{std} is given as [11]

$$q_{\text{std}} = cED^n \left[\frac{T_{\text{std}}}{p_{\text{std}}} \right]^u \left[\frac{p_i^2 - p_o^2}{G^x L T Z} \right]^y.$$

The equation above can be simplified as we can compute a constant, C , for the physical properties of a given pipeline and its surroundings. The flow rate q through the pipeline is then given as

$$q = C (p_i^2 - p_o^2)^y.$$

System Effects

For a single pipeline, the link between pressure and capacity is easy to handle. In large gas transportation systems with interconnected pipelines and processing plants, however, changes in pipeline pressure or composition of the gas flow in one part of the transportation network can affect the gas flow or the capacities in the remaining network as well. It is therefore insufficient to analyze only a part of the network. To assess the system effects, the gas production and transportation network has to be analyzed as a whole.

Compressor Stations

It is important to include pipeline pressure as a variable in the model. The ability to raise pipeline pressure results in a more realistic utilization of pipeline capacity. The total energy consumption of the compressor stations is then included in the objective

function, giving a better representation of both the real-world operations of the network and the operating costs.

The power W consumed by the compressor in order to raise pipeline pressure depends on the following variables: the flow q through the compressor, as well as inlet and outlet pressure, p_i and p_o . Thermodynamic properties of the natural gas such as its inlet temperature (T_i) and its compressibility factor (Z_i) affect the necessary compressor power as well. Compressor-specific operational characteristics for flow Q , inlet and outlet pressure, P_i and P_o , compressor efficiency (η), as well as the standard conditions (temperature $T_{\text{std}} = 288.15$ K and pressure $p_{\text{std}} = 1.01325$ bar) also have an impact on the power consumption. The nonlinear relationship between power, flow, and pressure can be linearized and approximated by the following formula [12]:

$$W = \left(\frac{P_i Q}{\left(\frac{2}{3}P_i + \frac{1}{3}P_o\right)^2 P_o} - \frac{P_o Q}{\left(\frac{2}{3}P_i + \frac{1}{3}P_o\right)^2 P_i} + \frac{P_o - P_i}{\frac{2}{3}P_i + \frac{1}{3}P_o} q \right) \cdot \frac{p_{\text{std}} Z_i T_i}{\eta T_{\text{std}} \times 24 \times 3600}.$$

Gas Quality

The quality of natural gas is described by means of its energy content. This energy content is expressed in terms of gross calorific value (GCV) or Wobbe index (WI). The WI represents the relationship between GCV and the relative density (specific gravity) of the gas. Customers make their nominations in energy content, and gas sellers must deliver gas with GCV and WI values within maximum limits.

Natural gas is a mixture of different components. Its composition and thus its energy content depend strictly on the production facility considered. Blending gas from different fields allows for a better utilization of the available resources to satisfy customer demand. This is challenging to model, as we have to keep track of the quality in every node and pipeline, especially when blending and splitting gas flows. Modeling natural gas using multiple components and considering

gas qualities introduces a so-called pooling problem [13,14] that increases the complexity of solving the mathematical model considerably.

Processing Plants

Processing plants usually remove contaminants (such as CO_2 or H_2S), heavy hydrocarbons, sour gases, and water. Some of these components have a market value and can be stored in storage tanks and are sold in a downstream market. Another important reason for modeling CO_2 removal in processing plants is to avoid corrosion problems in pipelines caused by high CO_2 content.

In addition to restrictions on the GCV and WI, customer requirements limit the allowed amount of contaminants at the exit terminals. Detailed modeling of processing plants can reduce the risk for off-spec deliveries and increase the total value of natural gas deliveries.

MODEL IMPLEMENTATION AND SOLUTION

The basic structure of the GassOpt model is a network flow model. However, modeling the different elements of the natural gas production and transportation system results in a nonlinear problem, not suited for most commercial solvers. In particular, the relationship between gas pressure and pipeline capacity as well as the pooling constraints due to gas quality requirements introduce nonlinear constraints to the model.

The GassOpt model uses the Weymouth equation to determine pipeline capacity. This equation can be linearized by means of a Taylor series expansion. In order to deal with the pooling constraints, the formulation for splitting gas flows is discretized. We introduce a set of binary variables in each split node for modeling the split fraction. See Tomasgard *et al.* [10] for a more detailed description and mathematical formulation.

In general, there will be several possible strategies for fulfilling the customer nominations. In order to have the model generate realistic flows and pressure values in the system, several penalty terms are added to

the objective function. These penalty terms apply to the total system pressure level and to deviations from the preferred flow split in certain nodes. In addition, deviations from contracted gas deliveries to customers in terms of volume are penalized in the objective function. Deviations with respect to minimum delivery pressure and gas quality limits are not allowed. Tests show that these terms do not significantly reduce the maximum flow potential but cause the model to choose more realistic flow patterns and pressure levels.

The model is implemented in Mosel and solved using XpressMP. Most of the real-world problem instances can be solved with an optimality gap of 0.5% within 20 min. There are, however, a few problem instances for which the optimality gap still exceeds 5% after reaching the runtime limit. The optimality gap seems small, but, given the market value of natural gas, this gap may represent several million USD in lost revenues.

USERS OF GASSOPT

The GassOpt model is developed in close cooperation with Statoil and Gassco. Statoil is Europe's second largest supplier of natural gas and the sixth largest in the world [1], whereas Gassco operates the transportation network on the Norwegian Continental Shelf. Both companies use the model in their planning processes, though the main focus of the analysis is different.

Statoil uses the GassOpt model in their long-term planning, evaluating the operational consequences and different investment decisions by considering all production facilities as part of the same production portfolio. In case of disruptions in production or transportation, the model is able to determine how to satisfy the customers' demand by rerouting the gas or moving production volumes from one facility to another.

Gassco administers access to the pipeline network. All producers on the Norwegian Continental Shelf have to book transportation capacity with Gassco. The GassOpt

model allows Gassco to quickly check the feasibility of the transportation bookings with respect to quality and capacity constraints. Gassco is also responsible for further developing the upstream network. Using the model enables Gassco to identify bottlenecks and propose necessary investments for increasing the network capacity and avoiding off-spec deliveries. The effect of changes in the infrastructure regarding contracted gas quality, required pressure levels, and capacities can be easily evaluated.

BENEFITS

One way of measuring the benefits of using GassOpt is, of course, calculating the monetary effect of using the model. Because of GassOpt's system perspective, Statoil can reallocate production and reroute gas to better meet customer specifications. The company estimates the accumulated savings due to an increase in fulfillments of contractual obligations at around USD 150–300 million.

GassOpt is also used to analyze the consequences of investments in infrastructure. Owing to the improved decision support that GassOpt provides to the decision makers, investments can be better targeted at removing bottlenecks in the network, ensuring that the capacity of the network meets the producers' requirements. The savings from avoiding bad investment decisions are conservatively estimated at around USD 900 million so far.

Another important factor, often observed with the use of optimization methods [15], is an improved understanding of the system's complexity. By understanding the technical aspects of the network and formulating them as a mathematical model, both researchers and practitioners increase their knowledge over the impact of decisions on the operations of the system.

FUTURE DEVELOPMENTS

Owing to the increased amount of contaminants present in recently developed gas fields, gas quality is becoming more important.

Blending and splitting of natural gas from different fields will be necessary to a greater extent, not only to meet the contractual obligations in the markets but also to fulfill the quality specifications of the processing plants. Future research will therefore focus on a system-wide quality management, extending the quality requirements upstream to ensure optimal operation of the processing plants as well. The blending and splitting processes introduce pooling constraints, such that future efforts will also be directed toward solving the resulting mathematical model.

Another possible improvement is to make the model more dynamic. As of today, the model only allows for a steady-state analysis of the production and transportation network, but cannot capture changes within the planning period. Modifying the model for a multiperiod setting will also facilitate modeling linepack and introduce additional flexibility for the operations of the network.

REFERENCES

1. StatoilHydro ASA. Energy realities—facing up to the energy challenges of tomorrow, today. Stavanger: StatoilHydro ASA; 2009.
2. Nordvik, M, Moen, T, Zenker, E, editors. Facts—The Norwegian Petroleum Sector 2009. Oslo: Ministry of Petroleum and Energy and Norwegian Petroleum Directorate; 2009.
3. Bodington CE, Baker TE. A history of mathematical programming in the petroleum industry. *Interfaces* 1990;20(4):117–127.
4. Nygreen B, Christiansen M, Bjørkvoll T, *et al.* Model Norwegian petroleum production and transportation. *Ann Oper Res* 1998;82:251–267.
5. Van den Heever SA, Grossmann IE, Vasantarajan S, Edwards K. A Lagrangean decomposition heuristic for the design and planning of offshore hydrocarbon field infrastructures with complex economic objectives. *Ind Eng Chem Res* 2001;40(13):2857–2875.
6. Goel V, Grossmann IE. A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Comput Chem Eng* 2004;28(8):1409–1429.
7. Ulstein NL, Nygreen B, Sagli JR. Tactical planning of offshore petroleum production. *Eur J Oper Res* 2007;176(1):550–564.

8. Selot A, Kuok LK, Robinson M, *et al.* A short-term operational planning model for natural gas production systems. *AIChE J* 2008;54(2):495–515.
9. Rømo F, Tomasgard A, Hellemo L, *et al.* Optimizing the Norwegian natural gas production and transport. *Interfaces* 2009;39(1):46–56.
10. Tomasgard, A, Rømo, F, Fodstad, M, *et al.* Optimization models for the natural gas value chain. In: Hasle G, Lie, K-A, Quak E, editors. *Geometric modelling, numerical simulation and optimization*. Berlin: Springer; 2007. pp. 521–558.
11. GSPA. *Engineering data book*. 11th ed. Tulsa, OK: Gas Processors Suppliers Association; 1998.
12. Van der Hoeven, T. *Math in gas and the art of linearization* [PhD thesis]. Groningen, The Netherlands: Faculty of Economics and Business, University of Groningen; 2004.
13. Haverly CA. Studies of the behaviour of recursion for the pooling problem. *ACM SIGMAP Bull* 1978;25:19–28.
14. Tawarmalani M, Sahinidis NV. *Convexification and global optimization in continuous and mixed-integer nonlinear programming*. Norwell (MA): Kluwer Academic Publishers; 2002.
15. Jeffrey P, Seaton R. The use of operational research tools: a survey of operational research practitioners in the UK. *J Oper Res Soc* 1995;46(7):797–808.

USING OR TO OVERCOME CHALLENGES IN IMPLEMENTING NEW VOTING TECHNOLOGIES

MICHAEL J. FRY
Department of Operations,
Business Analytics and
Information Systems,
University of Cincinnati,
Cincinnati, Ohio

MUER YANG
Opus College of Business,
University of St. Thomas,
Saint Paul, Minnesota

INTRODUCTION

In the United States Presidential election in 2000, candidates George W. Bush and Al Gore were locked in an extremely tight contest. Because of how the United States elects its President through the Electoral College, the contest depended on the outcome of the state of Florida; whichever candidate won the popular vote in Florida would be named the next President of the United States. Election Day on November 7, 2000 ended with no clear winner in the popular vote in Florida because the results were too close to call and many ballots had to be recounted. This set off a chain of events that culminated in the US Supreme Court case of Bush versus Gore and the eventual outcome that George W. Bush was determined to have won the Florida popular vote by a margin of 537 votes; hence Bush became the 43rd President of the United States.

Much of the controversy in Florida in 2000 centered on having to interpret votes made using punch-card ballots. Punch-card ballots allow the voter to punch out a small piece of perforated paper (typically known as a *chad*) next to the candidate's name or the ballot issue for which the voter wishes to cast her ballot. Difficulties were encountered by election officials when trying to decide the intention of voters who cast ballots that had only

partially punched out the chad or punched multiple chads for candidates for the same office; these became known as *dimpled chads*, *pregnant chads*, and so on. Matters were further complicated when some observers felt that the punch-card ballots used in Florida had been poorly designed and could have led to voters casting their vote for the wrong candidate. Other issues that may have affected the outcome of the election in Florida in 2000 included the media initially predicting that Gore would win the state before all election polls had closed [1]. It is impossible to determine the exact effect these combined events had on the election in Florida, but it was clear to election administrators that changes needed to be made in the administration of future elections.

In order to prevent a repeat of the difficulties encountered in the 2000 election, Congress passed the Help America Vote Act (HAVA) in 2002. HAVA sought "to establish a program to provide funds to states to replace punch card voting systems" [2]. Congress provided funds for these upgrades, but the decision of which exact new method to choose was left to the local election boards. Because part of the HAVA requirements for the new voting systems was that voters should be able to verify and change ballot errors, the new voting technologies were overwhelmingly of two basic types: optical-scan machines and direct-recording electronic (DRE) machines (Figs 1 and 2).

Each of these new systems (optical-scan and DRE) presented new operational challenges in voting that had to be overcome. Optical-scan machines are bulky, heavy items that are typically delivered by truck; this creates a difficult-to-solve instance of a vehicle-routing problem that administrators must solve to get the machines to the polling locations on time. DRE machines are expensive; thus, they are a scarce resource. Because DRE machines are the bottleneck in the voting process, improper allocations of DRE machines to polling locations can create queues to vote and excessive waiting



Figure 1. Example of a direct-recording electronic (DRE) voting machine (also known as a *touch-screen voting machine*).

times for voters. Interestingly, these challenges are somewhat specific to the type of voting technology being used. Unlike DRE machines, optical-scan machines are not typically the bottleneck in the voting process, so it is generally sufficient to have one optical-scan machine per precinct. Because DRE machines are typically much smaller than optical-scan machines (about the size of a laptop computer), they can usually be delivered to polls by poll workers in their personal vehicles.

Because of the widespread impact of elections and the importance of providing fair and efficient voting systems, voting presents unique opportunities for OR to improve social welfare. Previous work in OR has focused on predicting election results (e.g., Refs 3 and 4) which extends the extensive work on forecasting election results in political science and other fields ([5] for an overview of election forecasting applied to US presidential



Figure 2. Example of an optical-scan voting machine.

elections and [6] for an overview of models applied to elections in countries outside the United States). Bozkaya *et al.* [7,8] and King *et al.* [9] demonstrate the use of OR methods to problems related to political districting where geographical boundaries are defined to separate different political districts to accomplish some specific objective (balancing political representation, number of voters per district, etc.). More similar to the work presented in this article, other researchers have examined the use of OR methods to improve the efficiency and equity of voting systems. Fry and Ohlmann [10] investigate the problem of how to deliver optical-scan voting machines to precincts; Refs 11–17 specifically consider the problem of how to allocate voting machines to precincts. In this article, we examine more closely the problems and solution techniques presented in Refs 10,16,17.

CHALLENGES PRESENTED BY OPTICAL-SCAN VOTING MACHINES

Optical-scan voting machines are similar to scantron-based technology used in many

educational institutions to grade multiple-choice exams. The voter fills in a box or circle next to each candidate or issue choice for which she wishes to cast her ballot. Once the voter has completed voting, the ballot is scanned by the optical-scan voting machine to ensure that the voter has not cast any disallowed votes (e.g., voting for too many candidates for the same office). If the ballot is found to be in error, the voter can correct his/her ballot.

The operational challenge presented by these types of voting machines is that they are large and heavy, to the extent that they typically must be delivered to the polling location by a large moving truck because they do not fit in most passenger-sized cars. This creates a significant change in many localities that had previously used punch-card ballots, which require only a stack of punch cards and usually a collapsible ballot box. The punch cards and ballot box were very small, so they could be picked up by a poll worker on the morning of the election and delivered to the polling location. However, with optical-scan machines, the voting machines would have to be delivered to the polling location by a specialized delivery truck. These deliveries often had to be made days in advance of the election, in which case someone would have to be available to receive the machine and then, as mandated by law, it would have to be kept in a secured area until Election Day.

In order to accommodate the schedules of those receiving the machines, many election boards solicited delivery times when someone would be available to receive the voting machines at each polling location. Thus, the delivery requirements necessitated by the use of optical-scan voting machines created a classic version of the vehicle-routing problem with time windows (VRPTW). In even moderate-sized localities, the number of delivery locations and number of voting machines to be delivered creates a challenging problem to solve. Consider the case of Hamilton County, Ohio, which in 2006 had 966 optical-scan voting machines to deliver to 531 different polling locations. Each of these 531 different polling locations submitted time windows when they could receive the voting machines, which created a problem that was

intractable to most easily implemented solution approaches.

The VRPTW is a well-studied problem in OR; many exact and heuristic solution approaches have been examined. In order to keep solution times fast even for large problems and to prevent boards of elections from having to purchase dedicated optimization software, heuristic solution approaches are generally preferred in this setting. Fry and Ohlmann [10] present a penalty-based annealing heuristic to solve this problem. The authors' approach augments traditional simulated annealing by relaxing both time windows and capacity constraints via variable penalty terms. The objective in the authors' models is to minimize the travel distance for a fixed number of vehicles. The authors iteratively increase the number of vehicles available and resolve the problem to determine both the number of vehicles required and the total travel distance.

Fry and Ohlmann [10] report that they used this solution approach to design delivery routes for Hamilton County for several election cycles. The authors provided turn-by-turn directions to the delivery drivers who found that they were able to deliver almost all voting machines during their prescribed time window. The authors also found that they could generate much "better" solutions (in terms of requiring fewer vehicles and fewer miles traveled) by slightly modifying the time-window selection method for the delivery locations. Instead of allowing each location to choose an unconstrained time window, delivery recipients could be asked to select a set of days when delivery could be accepted as well as whether the delivery could take place in the morning, afternoon, or anytime during the day. The authors found that the responses to these solicitations gave much wider time windows than simply asking the recipient to state an available time window for delivery. Soliciting multiple possibilities for the day of delivery necessitates a two-phase solution method where first each polling location must be assigned to a delivery day and then the resulting collection of VRPTW instances must be solved using the authors' penalty-based annealing approach.

The two phases can then be solved iteratively to improve the overall solution. Fry and Ohlmann [10] claim that allowing multiple delivery days greatly reduces the required number of vehicles and required miles traveled. This new method for soliciting delivery time windows was instituted for several of the later years in which the authors performed this service for Hamilton County with great success.

CHALLENGES PRESENTED BY DIRECT-RECORDING ELECTRONIC (DRE) VOTING MACHINES

Many counties that do not use optical-scan voting machines in the United States use DRE voting machines, also known as *touch-screen machines*. Voters cast their ballots on DRE machines by touching the screen to indicate their voting preferences; DRE machines do not require any paper ballots (although many now also provide a paper record for auditing purposes). The advantage of DRE machines is that ballots can be easily modified to accommodate voters with disabilities and voters who require ballots in languages other than English. DRE machines are also much smaller and lighter than most optical-scan machines, easing the logistical burden of getting the machines to the polling locations. However, DRE machines are quite expensive which means that boards of elections must be judicious in their allocation of machines to precincts. In addition, DRE machines often require that the voter scroll through several pages of text to complete their ballot (recent ballots in Ohio have required voters to make decisions in 20 or more political offices and 9 or more ballot initiatives). Traditional voting systems often permitted voters to see all races and issues simultaneously and easily skip over items not of interest. Thus, DRE machines generally require more time for each voter to cast her ballot. An important decision for election boards is then the need to allocate the limited number of voting machines to precincts to promote efficiency and equity in the voting process.

Many voters in recent elections have suffered from inefficient and allegedly unfair

allocation policies. For instance, in the 2008 presidential election, some precincts in Franklin County, Ohio, recorded wait times greater than 5 h for many voters. Excessive wait times have also been reported in elections outside the United States; Adam [18] reports on long lines at the polls in the United Kingdom in 2010. Because it is estimated that 2–3% of voters are lost for every additional hour of wait time at the polls [13], such extended wait times can lead directly to voter disenfranchisement. However, simply minimizing some aggregate measure of voter wait times is not sufficient in this context; it is also important to ensure that voters are treated equitably across polling locations. Voters' geographic location is often strongly correlated with their political preferences [19]. Thus, if some precincts experience longer wait times than others, then this could give advantage to one political party over another or one particular outcome of a ballot initiative. This problem is exacerbated by the fact that ballot lengths often differ by precinct; hence, if voter service times are not considered in allocation decisions, then elementary queueing theory suggests that some precincts will experience longer waiting times. As an illustrative example, political science professor Walter Mebane found that African Americans waited on average 30 min longer to vote than other voters in Ohio in 2004 [20]. In a close election, such a result could clearly impact the outcome.

Voting systems should embody efficiency and equity. "Efficiency" refers to minimizing some aggregate measure of voter waiting times; and "equity" refers to having approximately equal waiting times for voters among the various precincts. The most commonly used allocation method in practice is to assign voting machines proportional to the number of registered voters in a precinct; this method is usually referred to as *proportional allocation*. However, waiting times are a function of both arrival rates and service times. Service times are affected by ballot lengths and differences in voter demographics, both of which can vary from precinct to precinct. Proportional allocation ignores differences in service times as well as any nonstationary arrival patterns of voters on Election Day.

Perhaps even more important, due to the highly charged political nature of election administration, decisions regarding voting machine allocations should be unbiased and transparent. The use of OR-based allocation methods promotes transparency and prevents bias.

On Election Day, a random number of voters arrive at each precinct throughout the day. Voters wait in line until a voting machine is available. This is an example of a parallel queueing system where the voters are the “customers” and voting machines are “servers.” The stochastic nature of both customer (voter) arrivals and service (voting) times, combined with limited server resources, results in voter queues and the potential for excessive waits.

The voting process captures many classical elements of a queueing system. Voter arrivals are random; in fact, even the total number of voters to arrive at a location is extremely unpredictable and is influenced by the nature of the election, general voter sentiment, and even the weather. The time required to cast a ballot is also uncertain, and it is highly affected by the length of the ballot, which can vary across polling locations. In most cases, voters form a single queue to use the next available voting machine. For DRE-equipped locations, the DRE machine is the bottleneck. The DRE machine is in use for the entire time that a voter is completing his or her ballot. In such cases, the DRE machines act as servers and voters as customers in a traditional $G/G/s$ -queueing system. However, there are additional complications that prevent classical queueing models from applying exactly. Voter arrivals are often nonstationary owing to voter work schedules and busy periods [12]. The short length of Election Day, usually less than 14 h, often prevents steady-state analysis and may even allow for time periods where utilizations exceed 1. Thus, traditional queueing models may be a questionable approximation of the reality of Election Day.

Determining an optimal voting-machine allocation is challenging for several reasons. (i) The arrival process is constrained in that no more voters can arrive than the number registered, a violation of the standard

queueing assumption of an infinite calling population. There are often surges in voter arrivals, contrary to the typical queueing assumption of stationary arrival processes [12]. Moreover, it is difficult to forecast turnout rates due to many uncontrollable variables (e.g., weather). (ii) Voting systems may violate traditional steady-state queueing-theory assumptions. For example, in the 2004 elections in Franklin County, Ohio, several precincts experienced average arrival rates that were higher than their average service rates, which would imply queue “explosion” in conventional queueing theory. Ohio law requires that the polls be open 13 h, plus however much time is needed to accommodate voters waiting when the polls close to new arrivals. (iii) Actual voting scenarios involve considerable computational complexity. A metropolitan area can consist of over 500 precincts and over 4500 voting machines to be allocated. The result is either a large-scale nonlinear stochastic optimization problem using traditional analytical queueing approximations that may not be valid, or a complicated simulation model. Thus, both model building and developing solution methods are challenging.

The literature directly examining voting-system operations includes Refs 11–15. Floyd [11], Edelstein [12], and Allen *et al.* [14] use simulation to model waiting times, allowing consideration of the realistic complications in their models, such as equipment failures and nonstationary arrivals. However, none of these papers explicitly considers voter equity. Allen and Bernshteyn [13] suggest simple analytical queueing models to predict average voter waiting times for given machine-allocation policies, and suggest an optimization model to allocate voting machines; they do not consider complicating issues such as nonstationary arrivals, machine failures, or specific differences in voting-time requirements due to differences in ballot lengths. Yang *et al.* [15] develop a simple heuristic to allocate voting machines based on simulation models, but they assume that voting-time distributions are identical across all precincts, and they do not provide optimality guarantees for their solutions.

In order to achieve the dual objectives of efficiency and equity, Yang *et al.* [16] build a large-scale resource-allocation model that can be solved very efficiently using heuristic methods with provable bounds for solution quality. The authors choose the maximum difference among waiting times as the inequity metric of interest, which allows for waiting times other than means (e.g., order statistics). The authors develop a two-phase algorithm utilizing simulation-optimization techniques to allocate resources equitably in a stochastic environment. To accommodate nonstationary arrivals of voters, the authors use simulation to estimate voter waiting times that allow for nonstationary arrivals, machine failures, and so on.

The authors illustrate the effectiveness of their algorithm by applying it to the allocation of voting machines in 2008 in Franklin County, Ohio, to minimize the inequity of voter waiting times among precincts. In 2008, election administrators in Franklin County were responsible for allocating 4625 DRE voting machines to 532 precincts. Table 1 [16] compares the authors' suggested allocation to the actual allocation used in 2008 based on the range of the 90th-percentile-order statistics of the waiting time distribution ($Z(X)$), the maximum waiting times experienced at a precinct ($Z_1(X)$), the system-wide expected waiting time ($Z_2(X)$), and the probability of wait times exceeding 30 min ($Z_3(X)$). The results in Table 1 show that implementing the authors' allocation during the 2008 election would have shortened voters' average wait times by 36% and resulted in 23,000 fewer voters waiting for more than 30 min—preventing voters from balking and being effectively disenfranchised. The authors' algorithms not only provide the rigor and transparency required by high-profile, and often contentious,

national elections, but these algorithms can be applied in a spreadsheet environment making them accessible to public-service officials in practice.

In addition to better allocations of available DRE voting machines, other potential operational changes in voting systems can also reduce voter waiting times. Yang *et al.* [17] discuss operational improvements related to early voting, educating voters, and providing feedback about line length at polling locations. Early voting (often called *absentee voting*) is becoming increasingly popular in many states; 16 and 22% of voters chose early voting in the 2000 and 2004 Presidential elections respectively. Early voting serves to decrease the arrival rate of voters to the polls on Election Day, thereby reducing utilizations and decreasing wait times. Early implementations suggest that informing voters about when polling stations are less busy may encourage voters to avoid peak times. Boards of elections in the District of Columbia, Utah, and some counties in Florida posted real-time waiting time estimates for early voters on their Web sites during the 2010 election [21]. Such efforts smooth the arrivals of voters over Election Day and can again reduce voter wait times. Clearly, voting queues can also be reduced if the time required for a voter to cast a ballot decreases. Boards of elections can use their websites to educate voters to become familiar with the issues on ballots and new voting machines prior to voters' casting the ballots. They can also provide sample ballots and voting machine instructions at polling stations while voters are waiting in line. Simulation results suggest that such measures can be equivalent, in terms of reducing queues, to adding a significant number of voting machines to polling locations.

Table 1. Allocation Comparisons (min)

Allocation	90th-Percentile of Waiting Time, $Z(X)$	Maximum Waiting Time, $Z_1(X)$	System-Wide Expected Waiting Time, $Z_2(X)$	Probability of Long Wait, $Z_3(X)$
Yang <i>et al.</i> (2012)	13.91 ± 0.75	36.74 ± 0.647	5.87 ± 0.0033	0.0427 ± 0.00010
Actual	72.91 ± 1.05	90.11 ± 1.321	9.17 ± 0.0048	0.1250 ± 0.00011

Note: Intervals are based on 95% confidence levels.

Yang *et al.* [17] also examine the robustness of their allocation methods since, clearly, factors such as the number of arriving voters are not known with certainty prior to Election Day. The authors show that their OR-based allocation method is fairly robust compared with competing heuristics to uncertainty of arrival rates; their allocation algorithm is shown to reduce mean voter waiting times by an average of 71% compared to the proportional allocation method even when voter turnout rates are not known with certainty.

CONCLUSIONS AND FUTURE EXTENSIONS

A common challenge facing researchers in applying OR to voting operations is that the optimality of solutions depends on the accuracy of key input parameters such as voter turnout and service times at the precinct-by-precinct level, which has proven difficult to predict with desirable confidence before Election Day. More work needs to be done in developing models to better estimate voter turnout by demographics, ballot issues, weather, and so on. Meanwhile, researchers can explore robust optimization techniques to create voting machine allocations that are designed to be insensitive to imprecise parameter estimates.

Additionally, work should be done to explore policies for dynamically allocating voting machines during Election Day. Currently, allocation decisions must be made weeks to months in advance of Election Day. Once Election Day arrives, the actual arrival of voters to different polling locations can be significantly different than expected. Complicating matters, voting machines are subject to failures, which can also create long queues to vote. Thus, methods for reallocating voting machines among precincts during Election Day are important to promote efficient and equitable voting systems. Some boards of election have experimented with keeping machines in reserve to allocate on Election Day to reduce voting queues, but these approaches have not been formally evaluated, and the implementation in practice has been frustrated owing to poor planning and execution. OR would seem to

possess many models for developing and testing potential reallocation policies.

Election operations provide an exciting opportunity for OR to have significant impact on an issue with clear social-welfare implications. The authors of Refs 16 and 17 participated in drafting House Bill #260, which was passed in the Ohio State House of Representatives and is pending in the Ohio Senate. The bill mandates that voting-machine-allocation decisions in Ohio should consider queueing effects in terms of expected voter arrivals and expected service times of voters. Passage of Ohio House Bill #260 marks a substantial achievement for applying OR methods to voting operations. It is hoped that this work and related endeavors will encourage more expanded use of OR in public-policy decisions.

REFERENCES

1. Lott Jr. JR. The impact of early media election calls on Republican voting rates in Florida's western Panhandle counties in 2000. *Public Choice* 2005;123(3):349–361.
2. 107th Congress. Help America vote act of 2002. 2002. Public Law 107 252. http://www.fec.gov/hava/law_ext.txt
3. Rigdon SE, Jacobson SH, Tam Cho WK, Sewell EC, Rigdon CJ. A Bayesian prediction model for the U.S. presidential election. *Am Polit Res* 2009;37(4):700–724.
4. Rigdon SE, Sewell EC, Jacobson SH, Cho WKT, Rigdon CJ. An analysis of daily predictions for the 2008 United States presidential election. *Case Stud Bus Ind Gov Stat* 2010;4:1–8.
5. Campbell JE, Lewis-Beck MS. US presidential election forecasting: an introduction. *Int J Forecast* 2008;24:189–192.
6. Lewis-Beck MS. Election forecasting: Principles and practice. *Br J Polit Int Relat* 2005;7(2):145–164.
7. Bozkaya B, Erkut E, Laporte G. A tabu search heuristic and adaptive memory procedure for political districting. *Eur J Oper Res* 2003;144(1):12–26.
8. Bozkaya B, Erkut E, Haight D, Laporte G. Designing new electoral districts for the city of Edmonton. *Interfaces* 2011;41(6):534–547.
9. King DM, Jacobson SH, Sewell EC, Tam Cho WK. Geo-graphs: an efficient model

- for enforcing contiguity and hole constraints in planar graph partitioning. *Oper Res* 2013;60(5):1213–1228.
10. Fry MJ, Ohlmann JW. Route design for delivery of voting machines in Hamilton county, Ohio. *Interfaces* 2009;39(5):443–459.
 11. Grant III FH. Reducing voter waiting time. *Interfaces* 1980;10(2):19–25.
 12. Edelstein WA. New voting systems for New York - long lines and high cost. Technical report, New Yorkers for Verified Voting Report; 2006. Available at <http://www.nyvv.org/doc/voterlines.pdf>. Accessed 2013.
 13. Allen TT, Bernshteyn MB. Mitigating voter waiting times. *Chance* 2006;19(4):25–36.
 14. Allen TT, Bernshteyn MB. Optimal voting machine allocation analysis. In: Hertzberg S, editor. Analysis of May 2006 Primary Cuyahoga County, Ohio. Election Science Institute; 2006. p 71–89. Available at <http://moritzlaw.osu.edu/electionlaw/news/documents/DREAnalysisforMay2006Primary.pdf>.
 15. Yang M, Fry MJ, Kelton WD. Are all voting queues created equal? Proceedings of the 2009 Winter Simulation Conference, Austin, Texas; 2009. p 3140–3149.
 16. Yang M, Allen TA, Fry MJ, Kelton WD. The call of equity: simulation-optimization models to minimize the range of waiting times. *IIE Trans* 2013;45(7):781–795.
 17. Yang M, Fry MJ, Kelton WD, Allen TA. Improving voting systems through service-operations management. *Prod Oper Manage* 2014. Forthcoming.
 18. Adam K. Hundreds of British voters were reportedly turned away. *The Washington Post*; 2010.
 19. Kohfeld CW, Sprague J. Race, space, and turnout. *Polit Geogr* 2002;21(2):175–193.
 20. Mebane WR. Timing and turnout in Ohio. Technical report, TomPaine.com; 2005. Available at http://www.tompaine.com/print/timing_and_turnout_in_ohio.php. Accessed 2013.
 21. Zambon K. Election administration 2010: innovations to watch. Technical report, electionlineWeekly; 2010. Available at http://www.pewcenteronthestates.org/uploadedFiles/wwwpewcenteronthestatesorg/Reports/Electionline_Reports/electionlineWeekly10.21.10.pdf. Accessed 2013.

USING QUEUEING THEORY TO ALLEVIATE EMERGENCY DEPARTMENT OVERCROWDING

LINDA V. GREEN
Armand G. Erpf Professor,
Columbia Business School,
New York, New York

THE GROWING PROBLEM OF ED OVERCROWDING

Timely access to care is a key component of high quality health care. Yet, patient delays are prevalent throughout the health-care system resulting in dissatisfaction and adverse clinical consequences for patients as well as potentially higher costs and wasted capacity for providers. For this reason, the Institute of Medicine has identified “timeliness” as one of the six keys in “aims for improvement” in its health-care quality initiative.

Arguably, the most critical delays for health care are the ones associated with health-care emergencies. Unfortunately, emergency department (ED) overcrowding is a continuing and growing problem. In a 2007 survey [1], nearly half of all US hospitals and 65% of urban hospitals reported being at or over capacity in their emergency rooms, resulting in long waits before being seen by a physician and delays of many hours or even days in getting a hospital bed. Each of these sources of delay can be life-threatening.

Delays for ED Physicians

Delays for emergency care generally begin with the wait to be seen by a physician. Upon arriving to the ED, patients are generally triaged into three major categories reflecting the severity and time-sensitivity of their condition. The most serious category is “emergent” (requiring “immediate” care), followed by “urgent” (requiring care within a “short” period of time) and “non-urgent.”

Though many patients are non-urgent and would not be harmed by significant delays in seeing a physician, many, if not most, are either “emergent” or “urgent.” However, recent studies have revealed that waits to see an ED physician are long and getting worse. In the United States, the overall median wait to see an ED physician increased from 22 min in 1997 to 30 min by 2004. More alarmingly, the median wait for patients diagnosed with acute myocardial infarction (AMI), who are emergent, increased from 8 min in 1997 to 14 min in 2004 [2]. These figures are considerably higher in urban areas and for teaching hospitals where most emergency care is delivered. Given the urgency of rapid treatment for victims of heart attack (as well as for sepsis, stroke, pneumonia, and trauma, among others) these figures, as well as anecdotal reports of patient deaths in EDs, indicate that these delays are a serious threat to the health and well-being of all those who need emergency care.

In addition, previous studies have established a strong link between long emergency department delays and the fraction of patients who leave without being seen (LWBS) [3,4]. The proportion of LWBS is itself an important measure of emergency department performance and quality of care. Several studies have concluded that patients who LWBS are sick and most require emergency care. One study has shown that up to 11% of patients who LWBS are hospitalized within a week and 46% of patients were judged to require immediate medical attention [5].

Delays for Inpatient Beds

Though insufficient ED physician levels are responsible for significant and serious delays in emergency care, the single biggest cause of ED overcrowding is the lack of inpatient beds [1]. About 25% of ED patients are admitted to the hospital and roughly 40% of inpatients come through the ED. For US hospitals reporting ED capacity problems, the average

time to move a patient from the ED to an inpatient bed once the decision to admit has been made is over 4.5 h and bed delays of over 24 h are not uncommon [6]. These delays lead to large numbers of patients “boarded” in the ED who strain the ability of the physicians and nurses to care for them. This ED overcrowding also compromises the ability of ED physicians to treat new arrivals to the ED.

When this kind of situation becomes untenable, a hospital will usually go on ambulance diversion—suspending new ambulance arrivals to the ED. In a 2007 survey, over half of all urban and teaching hospitals in the United States reported being on diversion in the past 12 months [1]. When hospitals are on diversion, seriously ill patients have to travel further to get to the nearest hospital increasing their risk for adverse outcomes. In addition, due to the increased travel time, waits for ambulances increase, further endangering patients. A recent study based on data from New York City showed that when there are significant levels of ambulance diversion in a borough, the number of deaths due to AMI increases by 44% [7].

Bed delays can occur even when inpatient beds are available. This is because patients and beds are not all identical. Hospitals are divided into nursing units which typically consist of between 20 and 50 beds. Each nursing unit is used for one or more clinical services (i.e., medical, surgical, pediatric, obstetrics, cardiology, neurology). So a patient with heart disease who is waiting in the ED for a bed may experience a significant delay if all the cardiology beds are occupied even though beds may be available in other clinical units. In addition, some patients need telemetry beds, which, in many hospitals, are not available in all nursing units. Finally, most hospital rooms have two or more beds which can be used only by patients of the same gender. So it is possible, for example, for a female patient to be delayed in getting an inpatient bed if the only beds that are available are in rooms occupied by male patients.

The most common delays for inpatient beds are for critical care beds [1]. Intensive care units (ICUs) are usually the most

expensive units in the hospital due to both the technology utilized and the high level of staffing needed. The full per-day cost in an ICU is about three to five times as much as in a regular inpatient unit [8] and therefore, hospital administrators believe that these beds should be highly utilized in order to be “cost-efficient.” Consequently, delays are often the longest for these beds which are used for the most seriously ill and injured patients.

Delays Due to Unavailability of Nurses

Nursing costs comprise a substantial fraction of hospital budgets and therefore cost-effective nurse staffing is very important. In most hospitals, the number of nurses assigned to a unit is determined by a specified ratio of patients to nurses. The historical norm for most types of clinical units has been 8:1, while for intensive care units it could be as little as 1:1. Though most hospitals subscribe to these standards, cost pressures and a national nursing shortage have resulted in these ratios being exceeded in many cases. Sometimes, however, this is the result of a failure to adequately plan for the daily, weekly, and sometimes seasonal variations in hospital census that are common in most clinical units of virtually every hospital.

Insufficient inpatient nursing levels are associated with patient dissatisfaction as well as increased levels of medical errors [9,10]. However, nurse unavailability is also a major factor in delays for beds in the ED. In order for a patient to be transferred to an inpatient bed from the ED, a nurse needs to be available to handle the admission to the unit. So, if the staffing level in the unit is such that the nurses are highly utilized, patients waiting for a bed in that unit will continue to wait even when a bed becomes available.

Nursing levels also play a role in delays for care within the ED. As in other parts of the hospital, nurses are the primary managers and caretakers of patients. If the number of patients in the ED becomes too high for the nurses to handle, new arrivals may not be taken into the treatment area until some of the current patients are discharged or moved

to inpatient beds. If the ED is very congested, the hospital may go on ambulance diversion.

Delays due to lack of nursing personnel may be more prevalent in the ED than in other parts of the hospital. The variability in the volume and pattern of patient arrivals, as well as the diversity of medical problems in the ED, make it a difficult environment for which to predict nursing needs. And the stress of the work environment in the ED makes it harder to attract and retain nurses and often results in high levels of absenteeism. Since the need for rapid treatment is generally greater in the ED, these delays may be particularly problematic.

USING QUEUEING THEORY TO REDUCE DELAYS FOR EMERGENCY CARE

Delays for ED Physicians

As mentioned above, visits to EDs have been steadily increasing. Constrained provider capacity relative to demand volume is exacerbated by the extreme variability in demand during each 24-h period experienced by a typical emergency department. This time-of-day pattern, as reported in the National Hospital Ambulatory Medical Care Survey for 2002, is distinguished by a relatively low level of demand during the night followed by a precipitous increase starting at about 8 or 9 a.m., a peak at about noon, and persistently high levels until late evening [6]. In addition, though the general pattern of demand is similar across the days of the week, individual days are likely to experience different overall volumes as well as slight differences in the exact timing of peaks and valleys. In particular, emergency departments are likely to have fewer visits on weekends than on weekdays.

Though physician staffing levels are generally varied over the day to meet these changing levels, this is usually done without the aid of operations research (OR) models. In a study conducted in the ED of a mid-size urban hospital in New York City [11], the overall average volume varied from a low of 63 patients per day on Saturdays to a high of 72 patients per day on Mondays. This degree of variation indicated that the then-current

policy of identical staffing levels for all days of the week was likely suboptimal. However, it was deemed impractical to have a different provider schedule every day and so it was decided to use queueing analyses (see section titled “Delays for Inpatient Beds”) to develop two schedules: weekday and weekend. This required aggregating ED arrival data into these two groups. For each, demand data were collected for each hour of the day using the hospital’s admissions database to understand the degree of variation over the day to explore the impact of different shift starting times and/or staffing levels on delays.

Estimating the average provider service time per patient was difficult. This time includes several activities including direct patient care, review of X-rays and lab tests, phone calls, charting, and speaking with other providers or consultants. In many, if not most, hospitals, these data are not routinely collected. At the time of the study, provider service times were not recorded and were estimated indirectly from direct observation and historical productivity data. Discussions with ED physicians revealed that the time spent on a patient was not necessarily linked to patient triage status and so it was decided to use a single estimate for the average time per patient.

Traditionally, in a service system with time-varying arrivals, the desired staffing levels would be determined by the *stationary independent period by period* (SIPP) approach which begins by dividing the workday into planning periods, such as shifts, hours, half-hours, or quarter-hours. Then a series of stationary queueing models, most often $M/M/s$ type models (see *The M/M/s Queue*), are constructed, one for each planning period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. In Ref. 12, the SIPP approach was shown in many cases to seriously underestimate the number of servers needed to meet a given delay performance target. This is particularly true when the mean service times are high (e.g., 30 min or more) and planning periods are long (two hours or more). In these situations, it was demonstrated that a simple variant of

SIPP, called Lag SIPP, performs far better than the simple SIPP approach. The major reason is that in cyclical demand systems, there is a time lag between the peak in the arrival rate and the peak in system congestion. This lag is significant when the mean service time is long. Lag SIPP corrects for this factor.

In our ED physician staffing study, the Lag SIPP approach was applied by first advancing the arrival rate curve by our estimate of the average physician time per patient, 30 min. A series of $M/M/s$ models were then constructed for each 2-h staffing interval, using the average arrival rate for each based on the time-advanced curve and the estimated average 30-min service time. Linear regressions for arrivals for the 2002 calendar year showed that the ratio of the mean to the variance of the number of emergency department patient arrivals each half-hour by day of week was consistently close to one yielding strong support for the assumption of time-varying Poisson arrivals. Since there was no data on provider times and estimates were only available for the mean, the assumption of exponential service times seemed most appropriate. The delay standard we choose was that not more than 20% of patients wait more than one hour before being seen by a provider. The use of one hour is consistent with the time standards associated with emergent and urgent patient groups used in the National Hospital Ambulatory Medical Care Survey [6]. The 20% criterion reflected the approximate percentage of non-urgent arrivals at the study institution.

The modeling results gave the number of ED physicians needed in each of the 2-h staffing intervals to meet the delay standard. In total, 58 physician hours were needed on weekdays to achieve the desired service standard, which represented an increase of 3 h over the existing staffing level of 55 h. Model runs for the weekend indicated that the target performance standard could be achieved with a total of 53 provider hours. In both these cases, the queueing analyses suggested that some physician hours should be switched from the middle of the night to much earlier in the day. A more subtle change

suggested by the model was that the increase in staffing level to handle the morning surge in demand needed to occur earlier than in the original schedule. Though resource limitations and physician availability prevented the staffing suggested by the queueing analyses from being implemented exactly, the insights gained from these analyses were used to develop new provider schedules. More specifically, as a result of the analyses one physician was moved from the overnight shift to an afternoon shift, 4 h were moved from the weekends and added to the Monday and Tuesday afternoon shifts (since these were the two busiest days of the week) and a shift that previously started at noon was moved to 10 a.m.

At the time of the study, the study hospital did not collect data on delays to be seen by a physician. So the impact of the new staffing levels was evaluated by focusing on the effect on the percentage of LWBS. Two 39 week periods—one before the staffing changes (August 26th, 2002–May 25th, 2003) and one after the staffing changes (September 1st 2003–May 30th 2004)—were studied. Matching weeks were chosen to better control for seasonal variation in both volume and disease states. The intervals are not aligned by exact date as to control for number of total days as well as days of the week. Despite an increase in patient arrival volume of 6.3%, an increase in provider hours of only 3.1% resulted in a decrease in the proportion of patients who left without being seen by 22.9%. Restricting attention to a 4-day subset of the week during which there was no increase in total provider hours, a reallocation of providers based on the queueing analysis resulted in a decrease in the fraction of left without being seen of 21.7% while the arrival volume increased by 5.5%.

Delays for Inpatient Beds

With the advent of price deregulation and increasing pressure from government, employers and managed care organizations, U.S. hospitals have been forced to focus on cutting costs. Due largely to decreased levels of compensation from Medicare and Medicaid, which account for over 50% of hospital revenues, about 1/3 of all hospitals

have negative operating margins and the vast majority of them struggle to break even [13]. Until the last few years, the number of hospitals had been steadily declining for decades. Many more have downsized, particularly since the mid-1990s. Many of these closures and downsizings have been directly due to the use of target occupancy levels that have been used since the mid-1970s at both the federal and state government level to evaluate and control the supply of hospital beds [14]. As recently as December 2006, a commission set up by Governor Pataki of New York State mandated the closure of almost 20 hospitals and the downsizing of dozens more using their finding of an average occupancy rate of 77% as compared with the “ideal rate of 85%” as their primary evidence of “excess capacity” [15]. This was done without any analyses of the impact on delays for emergency or inpatient care. At the institutional level, low occupancy levels are often viewed as indicative of operational inefficiency and potential financial problems. So hospital administrators generally view higher occupancy levels as desirable.

As mentioned above, the most frequently cited reason for ED overcrowding is the lack of inpatient beds and, in particular, critical care beds. Yet, ICUs and other critical care units are usually very small and intentionally operated at high levels of utilization so as to be “cost-efficient.”

In a study of the impact of target occupancy levels on delays for time-sensitive conditions [16], the basic $M/M/s$ model was used to estimate delays for ICU beds for all relevant hospitals in New York State and demonstrate the potentially adverse impact of using simple occupancy level guidelines to determine hospital bed needs. The assumption of a Poisson arrival process is reasonable in this situation since many arrivals to ICUs come through the ED and these arrivals have been shown to be well-approximated by this process, see, for example, [11].

The average size of the units in the New York hospitals was only 15 beds and the mode was 10 beds, and the data showed an average occupancy of 75%. Using the traditional criterion of hospital occupancy levels, many

of these units would be considered to have “excess” beds.

Since patients needing an ICU bed are emergent and any delay can be dangerous, we used probability of delay as our metric of performance. This measure was also chosen because it depends only on size and server utilization which were available from New York State Institutional Cost Reports. Thus, the number of beds needed to meet any delay target could be easily computed.

Adopting a standard of a maximum probability of delay of 10%, the analyses indicated that 112 of the 194 ICUs, or about 58%, were over-utilized. If that target is reduced to 5%, 143 or 74% of the units were too small to handle their experienced workloads; for a 1% target, 175 or over 90% were of inadequate size. These estimates are likely to be conservative for several reasons. First, in an analysis of intensive care units at Beth Israel Deaconess in Boston, the coefficient of variation of length-of-stay ranged from 1.1 to 1.6 [17], suggesting that the $M/M/s$ assumption of exponential service times leads to underestimates of actual congestion. Second, several of the hospitals had multiple divisions or locations, and the reported units are sometimes the sum of two or more smaller units. Third, in many larger hospitals, there are several types of ICUs (e.g., medical, surgical, neurological, and cardiac). Therefore, some of the unit sizes reported in these data likely represent the combined size of smaller, specialized units. It is also important to note that the average reported length-of-stay in these units was almost 18 days, so that when delays occur, they are likely to be long. This observation is consistent with reports by ED administrators.

It is also important to note that in performing these types of simplistic analyses, the data used is not really reflective of the true demands for the units being studied. For example, because many ICU units are often fully occupied, the hospitals in which they are housed will likely be on “critical adult” diversion on occasion and so their occupancy levels are based on some unmet demands. Another reason this may occur is that when patients are waiting in the ED for an ICU bed, less seriously ill patients in the ICU unit

may be transferred to a “step-down” unit to make room for the more needy new arrivals, hence reducing the length-of-stay.

The above analyses demonstrate that even simple models can be useful in identifying capacity problems in hospitals. But discussions with administrators and physicians who work in critical care units point to several other ways in which queueing models can be used to help alleviate emergency room delays due to unavailability of beds. First, queueing analyses can be used to show the impact of consolidating smaller, specialized nursing units into larger ones. An example of this can be seen in Green and Nguyen [17]. Since in many hospitals, the longest ED delays are for the most critically ill patients, it might be advisable to consider consolidating some of the units for which these delays are consistently long. For example, in some hospitals, stroke monitoring units have been established for patients who are victims of serious ischemic strokes. These units are characterized by more intensive nursing levels and telemetry for monitoring the stroke patients who stay in the unit until they are stabilized and can be transferred to a general neurological unit. A current study on the management of stroke patients in several hospitals has revealed anecdotal evidence that ED delays for stroke patients are often very long, jeopardizing long-term medical outcomes. Therefore, it might be sensible to consider caring for stroke patients in a neurological ICU which is also used for patients with other critical neurological problems such as head trauma, as is done in many hospitals without dedicated stroke units. Increasing the flexibility of beds can significantly decrease delays and queueing analysis can be used to estimate the magnitude of the improvement.

Other types of OR models can be used to identify admissions policies for ICUs. The decision to place a new patient into an ICU or critical care bed is a very subjective one and there are no professional or hospital guidelines. While for many patients the need for a critical care bed is clear, for many others, it is not and admission may be determined by whether or not a bed is available. Shmueli

et al. [18] developed a model for optimizing admissions to an intensive care unit based on a policy of admitting only patients whose expected incremental survival benefit from being treated in the ICU exceed some hurdle. Using data from a Jerusalem University hospital, they demonstrated that such a policy could result in saving 17.9% more lives relative to using a first come first served policy. They also explored the relative performance of policies based on both expected benefit and bed availability and showed that these had only a minimal impact on the number of lives saved.

Delays Due to Unavailability of Nurses

There has been a great deal of discussion and debate in the professional nursing world as well as in state legislatures, about how many nurses are really needed in hospitals. While nursing ratios are still the most common methodology for determining nursing levels, the old guidelines of 8:1 have been challenged with the result that in 2004, California implemented the first state law mandating minimum nurse to patient staffing levels of 1 to 6 in general medical-surgical units [19]. Other states are contemplating similar legislation. However, there have been many arguments made against the use of mandated staffing ratios [20]. Opponents of mandated ratios observe that none of the studies of staffing and quality have identified an optimal ratio and that ratios are too inflexible to account for variation in nursing skills and the severity of patients' illnesses.

From an analytical perspective, a clinical unit can be viewed as a finite source queueing system since demands are generated from the inpatients in that unit. However, due to admissions, discharges and transfers, the number of inpatients varies over a shift. Furthermore, each of these changes in inpatient census triggers a demand for nursing care. So in order to identify appropriate nursing levels, it is necessary to model a single hospital clinical unit as a queueing system with two sets of servers: nurses and beds. Though patients are usually assigned to a specific nurse for each shift, it is common practice for any available nurse to attend to a patient if the designated nurse is busy with other

patients. So it is reasonable to assume that both the nursing system and the bed system are multi-server.

Yankovic and Green [21] developed an easy-to-use two-dimensional queueing model for evaluating the impact of any given nursing level on performance measures such as the expected waiting time for a nurse, the probability of being delayed, and other performance indicators. This will enable hospital administrators and nurse managers to determine nurse staffing levels for various units and shifts that best fit their target service standards.

From a policy perspective, the model can be used to demonstrate the potential problems with using the type of mandated nurse to patient ratios legislated by California. Specifically, analyses showed that while a 1:6 nurse to patient ratio may insure a small probability of delay, that is, less than 5%, for nursing demands in a busy clinical unit of 60 beds, an otherwise similar unit with only 30 beds would require a minimum of 1:5 ratio to achieve the same delay performance. This is not surprising for anyone who understands that queueing systems exhibit economies of scale. Less obviously, the results showed that the average length-of-stay can also have an impact on nursing levels required to achieve short delays for care. In particular, since the admissions and discharge processes require a significant amount of nursing time, units with shorter lengths-of-stay will require higher staffing levels than those with longer lengths-of-stay.

Because the model tracks the number of patients in the unit and assumes, as in reality, that a new patient cannot be admitted to the unit unless a nurse is available, the model can be used to estimate the impact of nursing levels on delays for patients waiting in the ED. This is particularly important since nurse unavailability is often overlooked as a factor affecting ED delays. Using the queueing model, delays of patients in the emergency room due to both bed and nurse unavailability can be easily estimated. Figure 1 compares the expected number of patients waiting for a bed in a single clinical

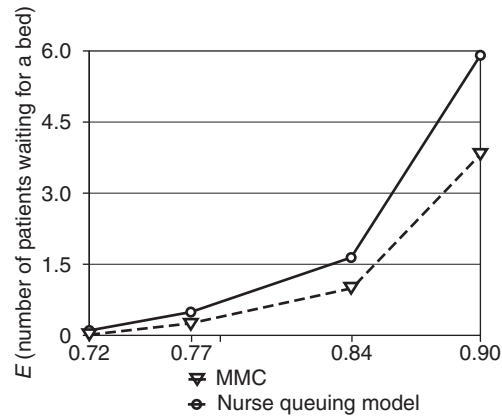


Figure 1. The impact of nurse staffing levels on ED delays.

unit by using a standard $M/M/s$ model representing just bed availability with the performance estimated from using the nurse queueing model. This demonstrates the potential impact of ignoring the effect of nurse staffing on emergency room congestion.

In the ED itself, there are several ways in which queueing and other OR analyses can be helpful in determining the impact of nurse staffing levels. Given the significant degree of variability in patient arrival times, types of problems, and time spent in the ED due to physician delays, testing delays and waits for inpatient beds, a simulation study (see section titled “Simulation Modelling and Analysis” in this encyclopedia) would be very helpful in identifying how patient levels and, therefore, nursing levels should vary by time of day and day of week. Another potentially fruitful study would be in determining how the number of scheduled nurses in each staffing period should be determined given significant absentee levels. This is the subject of a current research project being undertaken by the author.

OTHER POTENTIAL SOURCE OF DELAYS AMENABLE TO QUEUEING ANALYSIS

Though a lack of appropriate inpatient beds is usually cited as the major cause of ED overcrowding, patients often experience delays even when beds are available. In fact, as described in Yankovic *et al.* [7], one of the

two most frequent times for ED overcrowding and hence ambulance diversions in New York City is from midnight to 2 a.m. However, this is the time period when hospital occupancy levels are lowest.

One reason for this seeming anomaly was identified in one large New York hospital where a data collection effort showed that the time between bed assignment and the patient leaving the ED peaked from an average of 2.1h to between 3 and 4h during the midnight to 4 a.m., time interval. Further analysis revealed three potential reasons for this. First, the demand for transports peaked to about 8 patients per hour starting at midnight from a daytime average of about 7. This counterintuitive finding was due to the combination of peak arrival rates that started at about noon and stayed high until early evening, and an average duration of 8.2h between arrival time and bed assignment. However, because ED arrival rates drop to near their lowest during this time, hospital managers had decided that transport staff should be reduced starting at midnight from two to one. In addition, it was found that while the average transport during daytime hours was about 20 min, this increased to 27 min starting at midnight. This was attributed to the fact that during the day, ED transport personnel were used for transporting patients to diagnostic facilities located within the ED as well as to inpatient beds; while at night, when these facilities are closed, personnel were used only for transporting patients to beds. As a result of a queueing analysis that incorporated these factors, the hospital added a transporter during the midnight to 4 a.m., period with a subsequent average decrease of over an hour in transport delays.

In addition to transport personnel, most hospitals reduce other support staff at midnight. Many of these, such as nurses, who are needed to physically receive patients on the floors, housekeepers, who must make sure beds are prepared, and other personnel who are responsible for locating beds, impact ED delays. The above demonstrates the need to properly identify and analyze the impact of time-varying effects of both demands and

processing times throughout the hospital in order to alleviate ED overcrowding.

CONCLUSIONS

Health-care delivery systems are increasingly the subject of political discussions and debates as the United States and other countries struggle to reduce costs and improve quality and access. Much of these discussions center around whether there are too many or too few health-care resources such as inpatient beds, physicians, nurses, and imaging facilities. Unfortunately, politicians and policy professionals are often guided, or misguided, by simplistic frameworks and guidelines that are inappropriate and ignore the potentially dangerous impact on emergency department overcrowding. This was dramatically demonstrated in 1987–88 when New York City experienced a severe and protracted citywide shortage of inpatient hospital beds. During this period, ambulances were routinely turned away from full hospitals and urgently sick patients experienced delays of days waiting for an open bed. This hospital crisis followed a two-year period in which capacity declined by 9% because of new state regulations linking Medicaid reimbursement to achieving occupancy levels of at least 85% to reduce the number of “excess” beds in the city [22].

Queueing theory and other operations research models can and should play a role in these critical discussions and decisions regarding capacity needs and uses. The organization and use of clinical units varies significantly across hospitals and OR analyses are needed to identify best practices. Within the ED alone, there are opportunities to explore the impact of various triage policies and practices, the use of “fast-track” facilities to treat non-urgent patients, the use of “districting” to segregate patients into distinct areas either by type or to facilitate management, and many other new practices being implemented to reduce congestion. Given the growing interest in creating more cost-effective and high quality health care and the increasing prevalence

and ability of health-care information technology systems to collect and provide the needed data, this is a rich area for further exploration.

REFERENCES

1. American Hospital Association. Survey of Hospital Leaders, 2007.
2. Wilper AP, Woolhandler S, Lasser KE, *et al.* Waits to see an emergency department physician: U.S. trends and predictors, 1997–2004. *Health Aff Web Exclusive* 2008;27:w84–w95.
3. Green RA, Wyer PC, Giglio J.ED. Walkout rate correlated with ED length of stay but not with ED volume or hospital census (abstract). *Acad Emerg Med* 2002;9:514.
4. Fernandes CM, Price A, Christenson JM. Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *J Emerg Med* 1997;15:397–399.
5. Baker DW, Stevens CD, Brook RH. Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences. *JAMA* 1991;266:1085–1090.
6. McCaig LF, Burt CW. National hospital ambulatory medical care survey: 2002 emergency department summary. Advance data from vital and health statistics. Hyattsville (MD): CDC: National Center for Health Statistics; 2004. pp. 1–36.
7. Yankovic N, Glied S, Grams M, *et al.* The impact of ambulance diversion on myocardial infarction mortality. Working paper, 2009.
8. Groeger JS, Guntupalli KK, Strosberg M, *et al.* Descriptive analysis of critical care units in the United States. *Crit Care Med* 1992;21:279–291.
9. Aiken LA, Clarke SP, Sloane DM, *et al.* Hospital nurse staffing and patient mortality, nurse burnout and patient satisfaction. *J Am Med Assoc* 2002;288:1987–1993.
10. Needleman J, Buerhaus P, Mattke S, *et al.* Nurse staffing levels and the quality of care in hospitals. *N Engl J Med* 2002;346:1715–1722.
11. Green L, Soares J, Giulio J, *et al.* Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Acad Emerg Med* 2006.January
12. Green L, Kolesar PJ, Soares J. Improving the SIPP approach for staffing service systems with cyclic demand. *Oper Res* 2001;49:549–564.
13. American Hospital Association. Trendwatch chartbook, trends affecting hospitals and health systems. 2008.
14. McClure W. Reducing excess hospital capacity. Rockville (MD): Bureau of Health Planning; 1976.
15. Commission on Healthcare Facilities in the 21st Century. A plan to stabilize and strengthen New York’s healthcare system. 2006.
16. Green L. How many hospital beds? *Inquiry* 2002/2003;39:400–412.
17. Green LV, Nguyen V. Strategies for cutting hospital beds: the impact on patient service. *Health Serv Res* 2001;36:421–442.
18. Shmueli A, Sprung CL, Kaplan EH. Optimizing admissions to an intensive care unit. *Health Care Manag Sci* 2003;6:131–136.
19. California Department of Health Services. Final statement on reasons, hospital nurse staff ratios and quality of care, r37-01. 2003.
20. Lang TA, Hodge M, Olson V, *et al.* Nurse-patient ratios: a systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *J Nurs Adm* 2004;34:326–337.
21. Yankovic N, Green L. A queueing model for nurse staffing. Working paper, Columbia Business School, 2009.
22. Myers LP, Fox KS, Vladeck BC. Health services research in a quick and dirty world: the New York City hospital occupancy crisis. *Health Serv Res* 1990;25:739–755.

VALUE FUNCTIONS INCORPORATING DISAPPOINTMENT AND REGRET

LAURA MORETTI
IRENE CRISTOFORI
ANGELA SIRIGU
CNRS, Center for Cognitive
Neuroscience, Lyon, France

Economists have produced remarkable data describing how we make economic choices. Until very recently, the economic approach treated the human brain as a “black box” and suggested mathematical equations to explain behavior during complex decisions. This “rational” approach has been enormously successful at yielding a simplified theory of the brain processes underlying economic strategies.

More recently, the introduction of neuroscience tools (brain imaging techniques, lesion studies, and single-cell recording in nonhuman primates) provided evidence that other variables, such as emotions, can interact with utility-maximization in ways that challenge the standard economic theory [1].

For instance, there is evidence that the subjective impact of experiencing gains or losses is influenced by the counterfactual comparison of obtained and unobtained outcomes [2]. The intensity of an emotional response, which can transform disappointment into painful regret when the outcome of a rejected alternative turns from mildly to highly valuable, has been found to be a better predictor of subsequent choices than expected utility alone [3].

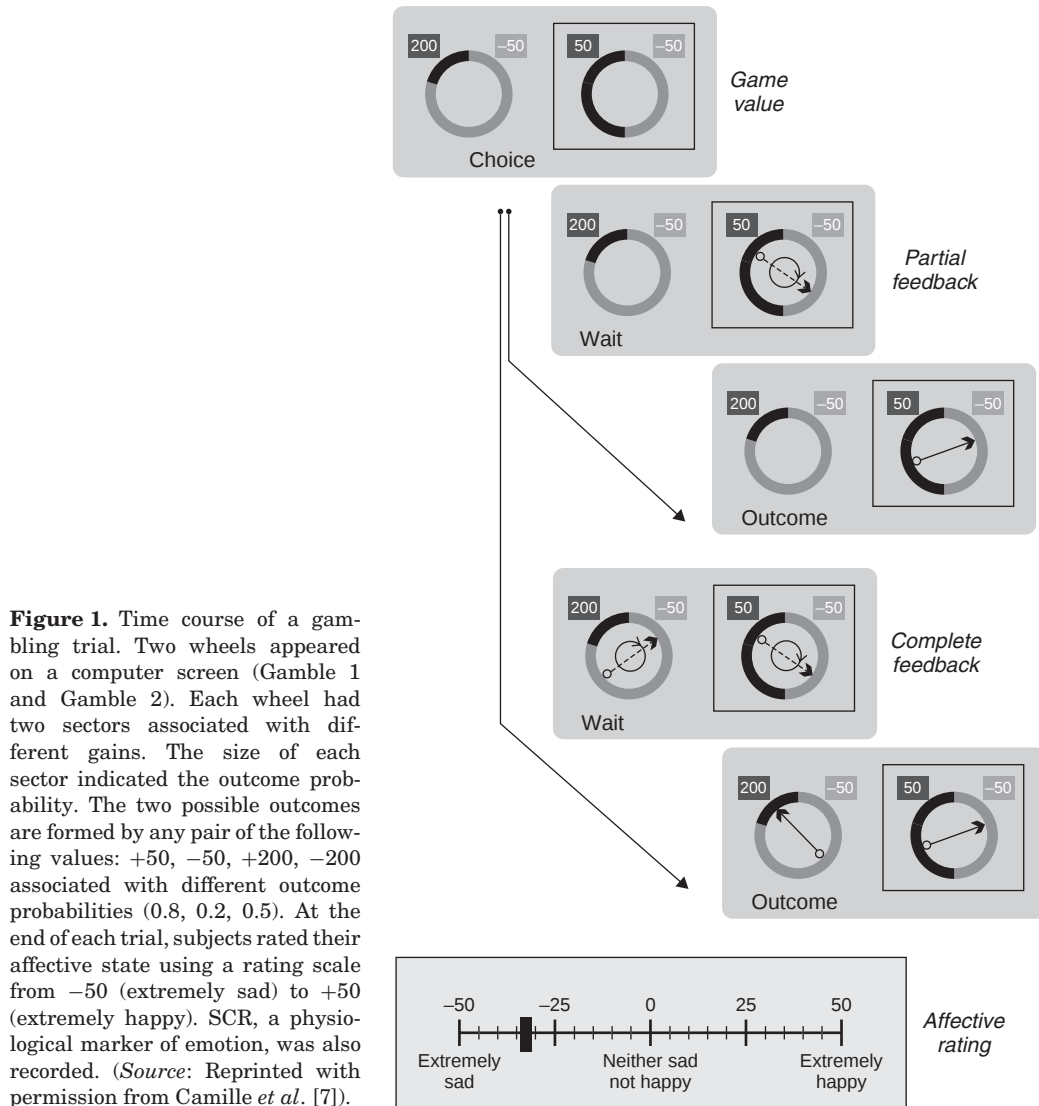
The fundamental role of regret in adaptive behavior and brain mechanisms, through which it affects the decision-making process, has been recently revealed by neuropsychological and neuroimaging studies.

The study conducted in our group by Camille and colleagues in 2004 is the first to show the neural basis of regret during decision making. The authors focused their attention on the orbitofrontal cortex (OFC),

a brain region that lies at the interface between emotion and cognition [4] and so ideally suited to control emotional experience through mechanisms such as counterfactual reasoning. The idea in this study was to test whether this cortical region was crucial to anticipate regret as an emotion that guides advantageous choices.

To this aim, Camille and colleagues adapted an experimental paradigm inspired from the work of Mellers *et al.* [5] to test normal subjects and patients with selective lesions within the OFC. The experimental task required subjects to choose between two gambles, each having different probabilities and different expected outcomes (Fig. 1). At trial start, a spinning arrow appeared; rotated for a variable duration, and then stopped indicating the outcome. Two different feedbacks were provided to subjects: a partial feedback and a complete feedback. In the partial feedback condition, the subject saw only the outcome of the selected gamble, while in the complete feedback condition, outcomes of both the chosen and unchosen gambles were provided. This latter condition was thought to induce regret because subjects could experience not only the financial consequence of their choice but also that of the alternative unselected option. At the end of each trial and after the outcome of the chosen gamble was displayed, participants were also asked to rate their emotional state. Moreover, to obtain a more implicit measure of their emotions, physiological responses (skin conductance responses, SCR) were measured both in patients and in normal subjects.

The results confirmed the hypothesis that emotions experienced for the obtained outcome are modulated by the unobtained outcome. For instance, a disappointment effect was elicited in the partial feedback condition: subjects perceived a loss as more unpleasant and a gain as less pleasant if the unobtained outcome from the same gamble won \$200 instead of losing \$200. This emotional modulation was much stronger in the



complete feedback condition where, knowing the outcome of the rejected alternative triggered the experience of regret. Participants not only reported a strong negative feeling when losing \$50 and the unchosen alternative won \$200, but they also reported being unhappy when gaining \$50 and the other option won more (\$200) (Fig. 2). The different emotional nature of disappointment and regret was revealed by the physiological index of emotional reactivity (SCR): viewing the outcome of the rejected alternative enhanced SCR as compared with viewing

only the outcome of the chosen gamble, thus, revealing that regret could engender a more intense response than disappointment.

The critical finding of this study concerned, however, the pattern of results in patients with OFC lesions (Fig. 3). Patients showed a disappointment effect comparable to control subjects: when losing, patients were indeed somewhat sadder if the unobtained outcome was a large gain than if it was a greater loss, and their SCR demonstrated clear emotional arousal when learning the outcome of the gamble.

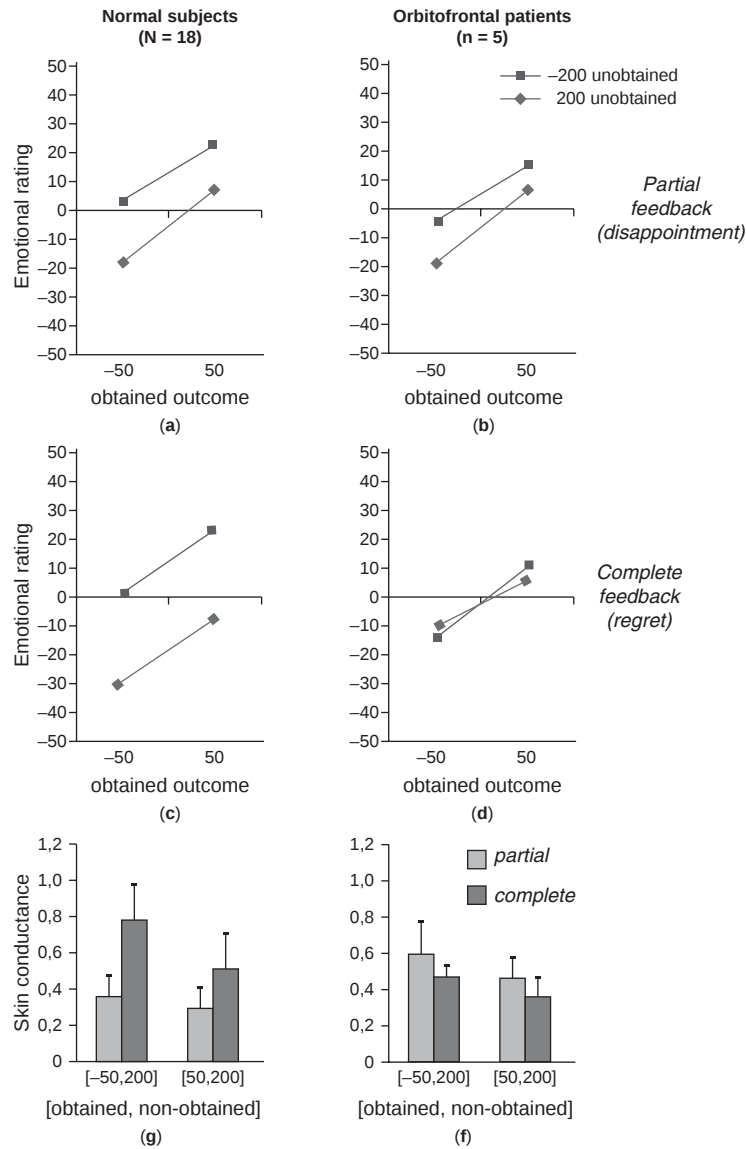


Figure 2. Effect of the unobtained outcome of the gamble in partial and complete feedback: (a and c) Mean emotional ratings made by 18 normal control subjects for two obtained outcomes (-50 or 50) as a function of the unobtained outcome (line with square points and symbols: -200 , line with triangle points and symbols: 200) in the partial and complete feedback conditions, respectively. In the partial condition, the unobtained outcome corresponds to the unobtained value of the chosen gamble. In the complete condition it corresponds to the obtained value of the nonchosen gamble. (b and d) Mean emotional ratings made by five orbitofrontal patients in the partial and complete feedback conditions, respectively. Conventions as in a and c. (e) Mean skin conductance response ($\pm$ standard error) of 15 normal subjects measured at the end of arrow rotation, for the conditions in which the unobtained outcome is more advantageous than the obtained outcome (corresponding to the lines with the triangles in the graphs above). Regret is correlated with stronger emotional arousal. (f) Same data for orbitofrontal patients, showing no regret effect. (Source: Reprinted with permission from Camille *et al.* [7]).

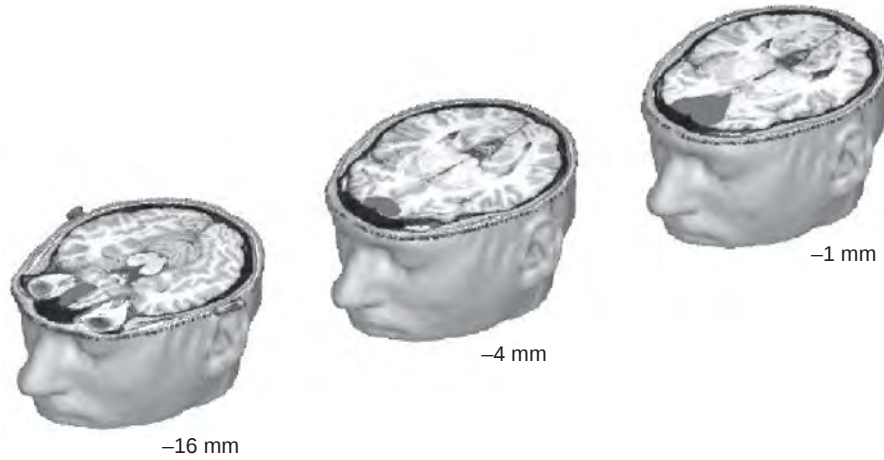


Figure 3. Lesion overlap in the OFC for the five patients. Lesion locations were reconstructed from individual MRI scans. The region of common cortical damage is located in the basal and ventromedial sector of the prefrontal cortex and involves Brodmann's areas: 10, 11, 32, 24, 47. (Source: Reprinted with permission from Camille *et al.* [7]).

This suggests that OFC patients were not emotionally unresponsive, and they were somewhat able to think counterfactually on the chosen gamble. Nevertheless, the emotions experienced by these patients were not modulated at all by the feedback of the unchosen gamble, and they seem to experience no regret. Indeed, emotional ratings showed that sadness expressed by the patients at losing \$50 was not more intense if the rejected alternative won \$200, nor the joy felt at winning \$50 was tarnished by seeing that the gain would have been larger if they had selected the alternative gamble (Fig. 2). These results confirmed the hypothesis that the ability to experience regret is mediated by the OFC.

Camille and colleagues went further by investigating how the anticipated emotion of regret influences the entire economic decision process. To this aim they tested a model of choice incorporating disappointment and regret as emotional variables and the expected values of the two gambles.

Noteworthy, the structure of the experiment was defined so that the two gambles always differed in their expected values, but the gamble with the highest expected value won less often on average than the one with the lowest expected value. This was done

to ensure that subjects would experience regret on a sufficient number of trials and so they could learn to choose advantageously by anticipating future emotional reactions and try to avoid negative emotions.

Results of this analysis showed that healthy subjects, after being exposed to several trials in which they experienced regret, began to choose the gambles with probable outcomes likely to produce minimal regret. By contrast, OFC patients persisted in choosing the more risky gamble with a high probability of producing regret.

Thus, regret experience contributes to making advantageous economic choices. The underlying process would involve integrating considerations about future emotional responses to the outcome of choices, that is, anticipating regret and maximizing expected values, allowing subjects to make advantageous decisions. This process would be impaired in patients with a lesion in the OFC. The lack of a counterfactual comparison between the value of a chosen and a rejected alternative would prevent them from anticipating the potential negative consequence of their choices. Consistent with economic models of regret [6], the regret function ψ , which is given by the comparison between the value of choice and the value

of a rejected alternative, would enter into the utility function (U). According to this theoretical model, the utility of an act (e.g., choice of x and simultaneous rejection of y) can be computed as the sum of the expected utility of the chosen act (x) and of the anticipated regret term $\psi(\cdot)$. Camille *et al.* [7] suggested that the integrity of OFC is necessary to incorporate regret function into the decision-making process, that is, into the utility function.

A lesion in the OFC would correspond to a $\psi_{\text{regret}} = 0$; and would cause subjects to make their judgments solely on the basis of what has actually occurred, and not by anticipating the negative consequences of their choices, that is, the regret. Paradoxically, the behavior of OFC-lesioned patients would be more in line with the rational theory, such that their behavior is in agreement with maximization of expected utility [3,7].

The involvement of the OFC in the experience and anticipation of regret and its role in learning processes was better defined in a second experimental study from our group by Coricelli *et al.* [8]. Here we measured the brain activity using functional magnetic resonance imaging (fMRI) while subjects participated in the regret gambling task.

The results revealed that three main regions correlated with the degree of experienced regret, namely the medial OFC region, the dorsal anterior cingulate cortex (ACC), and the anterior hippocampus. The

two latter regions are typically involved in declarative memories and cognition-induced arousal responses. These areas were activated when an outcome was compared with a more advantageous one, and were deactivated when the actual outcome was compared with a less advantageous alternative.

Interestingly, the hippocampal activity would suggest that subjects begin to learn to anticipate regret. This brain region supports indeed a modulation of declarative memory [9], such that after a bad outcome, the lesson to be learnt is: “in the future, pay more attention to the potential consequences of your choices.”

The influence of regret on subsequent choices was confirmed by behavioral data.

In line with the cumulative experience of regret, as regret increased subjects avoided risky choices over time. In fact, they first showed a bias away from choices that may potentially generate regret and, second, showed a bias away from choices that had previously led to negative outcomes. Crucially, these behavioral biases were expressed in a pattern of neural activity at the time of choice. During the time window between the trial onset and subject response, when subjects minimize regret and maximize expected values, an enhanced brain activity was found in the ACC. It is not surprising that this region, usually involved in conflict monitoring [10], is activated

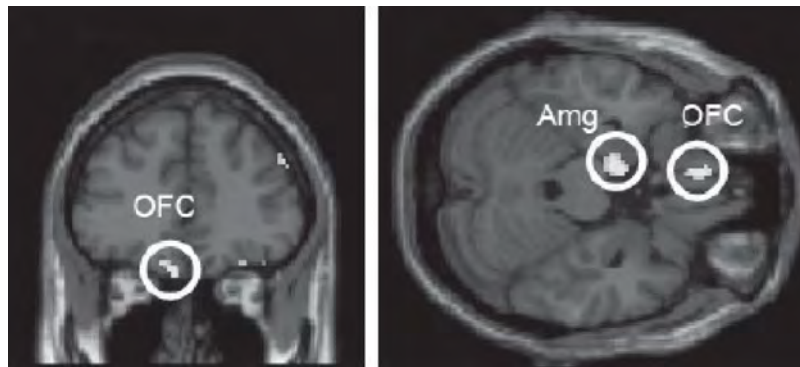


Figure 4. Activity at choice reflecting cumulative regret was found in the medial left amygdale (Amg; coordinates -8 (to -16), -4 , -25) and medial OFC (OFC; coordinates -10 , 40 , -24). (Source: Reprinted with permission from Camille *et al.* [7]).

in such kind of decisions where a conflict between maximizing the expected value and minimizing regret is at stake.

One critical finding of this study concerned the cumulative effect of regret, which makes the subject become more averse to regret over time. This effect was linked to the activity in the OFC and amygdala, the same brain regions that were found activated during the mere regret experience (Fig. 4). Thus, as a result of learning mechanisms, the same patterns of activations related to regret responses seem to be reactivated when subjects choose, thereby making it possible to anticipate the future consequences of his/her choices and therefore is anticipating the future consequences of his/her choices.

Overall, these findings show that a specific brain mechanism is responsible for attributing affective values to the alternative of our choices in terms of current or anticipated emotional experience. These results confirm that specific mechanisms of cognitive control are activated during the choice processes, involving reinforcement or avoidance of the experienced behavior. As such, the experience of regret and its anticipation would promote behavioral flexibility and minimize the negative outcomes of choices.

The interface between theory and experimentation and between mathematical models and brain function provide important insights for the understanding of the processes that underlie economic decision, preferences formation, and learning.

In line with this prospective, Marchiori and Warglien [11] have recently used neural network models that incorporate "regret" to predict the outcomes of games played by humans. They showed that also choices in economic games are better predicted by using regret-based models, than theoretical economic ones.

As the mathematical formalism of economic theories provides new ways of analyzing the neural computations involved in representing decision variables, neuroscientific approaches can contribute greatly to better understanding the cognitive

and emotional underpinnings of economic decision-making, including how people form beliefs about what other people feel or might do.

Overall, the findings reported here show that human choices and decision processes are not exclusively guided by utility considerations. Regret and the sense of responsibility that this complex emotion triggers may be an important factor for our choices and future decisions. These cognitive and emotional mechanisms are deeply rooted in specific brain regions such as the orbitofrontal, the cingulate cortex, and the amygdala.

REFERENCES

1. Lowenstein G, Rick S, Cohen J. Neuroeconomics. *Ann Rev Psychol* 2007;59:1–26.
2. Roese NJ, Olson JM. What might have been: the social psychology of counterfactual thinking. New Jersey: Erlbaum; 1995.
3. Coricelli G, Dolan RJ, Sirigu A. Brain, emotion and decision making: the paradigmatic example of regret. *Trends Cogn Sci* 2007; 11:258–265.
4. Rolls ET. The orbitofrontal cortex and reward. *Cereb Cortex* 2000;10:284–294.
5. Mellers BA, Schwartz A, Ritov I. Emotion-based choice. *J Exp Psychol Gen* 1999; 3:332–345.
6. Bell DE. Regret in decision-making under uncertainty. *Oper Res* 1982;30:961–981.
7. Camille N, Coricelli G, Sallet J, *et al.* The involvement of the orbitofrontal cortex in the experience of regret. *Science* 2004;304: 1167–1170.
8. Coricelli G, Critchley D, Joffily M, *et al.* Regret and its avoidance: a neuroimaging study of choice behavior. *Nat Neurosci* 2005;8: 1255–1262.
9. Eichenbaum H. Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron* 2004;44: 109–120.
10. Bush G, Luu P, Posner MI. Cognitive and emotional influences in anterior cingulate cortex. *Trends Cogn Sci* 2000;4:215–222.
11. Marchiori D, Warglien M. Predicting human interactive learning by regret-driven neural networks. *Science* 2008;319:1111–1112.

VARIANTS OF BROWNIAN MOTION

SHANKAR BHAMIDI

Department of Statistics and Operations
Research, The University of North
Carolina, Chapel Hill, North Carolina

PRISCILLA E. GREENWOOD

MCMSC, Arizona State University,
Tempe, Arizona

As explained in the (see ***Brownian Motion and Queueing Applications***), Brownian motion is one of the most fundamental of all stochastic processes. Brownian motion and its variants arise in a wide variety of contexts ranging from scaling limits of other processes to being used directly in models of physical systems. The main aim of this article is to introduce a number of variants of this mathematical object and pave the way for the general study of diffusions in the (see ***Introduction to Diffusion Processes***). Given the enormous range of applications of Brownian motion and its variants, it will not be possible to do justice to even a fraction of the variants of this beautiful process. Rather, we shall choose some particular examples and give the reader an idea of the diverse fields in which such models arise, ranging from queueing theory, stochastic finance, statistical physics, and models of diffusion on fractals. In each example we shall attempt to give some references for a deeper coverage of these topics. As one can imagine, for each of the models in this article, volumes have been filled developing their mathematical properties and areas of applications.

Let us first recall the definition of standard Brownian motion.

Definition 1. [Standard Brownian Motion]. A stochastic process $\{B(t) : t \geq 0\}$ is called standard Brownian motion if it has continuous sample paths with $B(0) = 0$, with stationary independent increments such that for all $0 \leq s < t < \infty$, $B(t) - B(s) \sim N(0, t - s)$.

It follows from the definition that Brownian motion is a Markov process.

For any $\mu \in \mathbb{R}$ and $\sigma > 0$, the process $\{X_{\mu,\sigma}(t) : t \geq 0\}$ defined by

$$X_{\mu,\sigma}(t) = \mu t + \sigma B(t)$$

is called *Brownian motion* with drift μ and variance parameter σ^2 . Figure 1 shows an example of the sample path of this process with positive drift.

REFLECTED BROWNIAN MOTION

A simple variant of Standard Brownian motion, is Brownian motion reflected at some level say the origin, so that it is constrained to be nonnegative. Reflected Brownian motions in one and higher dimensions arise as models of diffusing particles constrained within a region and have been used in a number of different fields, ranging from physiology to nuclear magnetic resonance. See, for example, Ref. 1 for a modern survey of such applications. These variants of Brownian motion are also very important in describing scaling limits of various discrete systems in areas such as queueing theory. While in many advanced settings one uses stochastic differential equations to construct such a process, let us give an example of one simple situation where the reflected process can be explicitly constructed using standard Brownian motion. We start with a motivating example of a simple discrete model, which converges to this process in the limit and then elucidate a bit more on its construction. Much of this section follows [2], which has a nice introduction for the derivation of such limit theorems and their applications in a wide variety of settings in queueing theory.

Consider the following simple model of a buffer of a switch in a large communication network whose role is to process inputs and transmit output. We discretize time into finite intervals called *periods*. For $k \geq 1$, let W_k represent the load on the buffer at the end

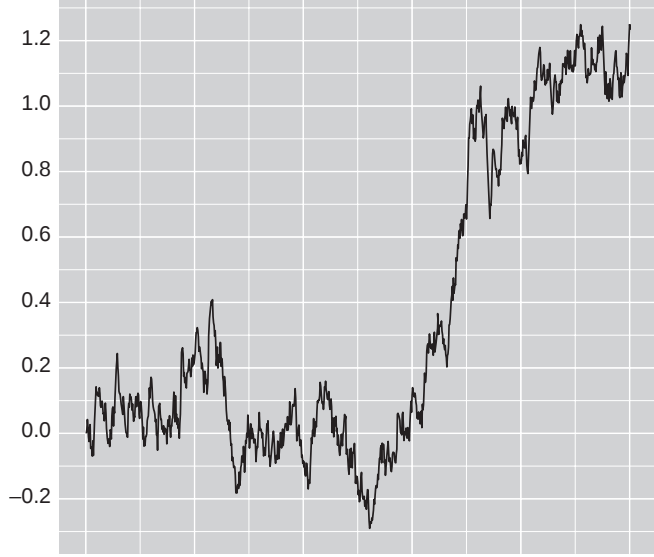


Figure 1. Brownian motion with drift $\mu = 2$ and sd $\sigma = 1$.

of period k . During this period suppose the buffer received some random input V_k and processed a deterministic output μ , assuming the buffer size is nonempty and the buffer has something to transmit. If the buffer has infinite capacity then the evolution of the workload is given by

$$W_k = \max\{0, W_{k-1} + V_k - \mu\} \quad (1)$$

while if the buffer has some finite capacity C then

$$W_k = \min\{C, \max\{0, W_{k-1} + V_k - \mu\}\} \quad (2)$$

Assume that the inputs V_k are independent and identically distributed positive random variables with finite variance σ and mean ν . Note that if the deterministic output μ is less than the expected input ν then the load on the buffer quickly increases (by the strong law of large numbers, the load on the buffer at the end of period k would be $W_k \sim k(\nu - \mu)$ as $k \rightarrow \infty$). Thus, suppose we assume that we are in the *critical* regime where the deterministic output μ is exactly equal to the expected input ν . Then since the process

$$\left\{ \frac{1}{\sqrt{n\sigma}} \sum_{k=1}^{\lfloor nt \rfloor} (V_k - \mu) : t \geq 0 \right\}$$

converges to standard Brownian motion, it is natural to postulate that the limit of the workload process in the infinite capacity regime given by Equation (1) is reflected Brownian motion, namely,

$$\left\{ \frac{1}{\sqrt{n\sigma}} W_{\lfloor nt \rfloor} : t \geq 0 \right\}$$

converges as $n \rightarrow \infty$ to Brownian Motion but reflected at 0. It turns out that the natural definition of reflected standard Brownian motion is the following.

Definition 2. [Reflected SBM]. Suppose $\{B(t) : t \geq 0\}$ is standard Brownian motion. Then the process $\{Y(t) : t \geq 0\}$ defined by

$$Y(t) = |B(t)|, t \geq 0$$

is called standard Brownian motion reflected at the origin.

Figure 2 shows a sample path of this process.

Now suppose $B(\cdot)$ is standard Brownian motion and define

$$\tilde{Y}(t) = B(t) - \inf_{0 \leq s \leq t} B(s).$$

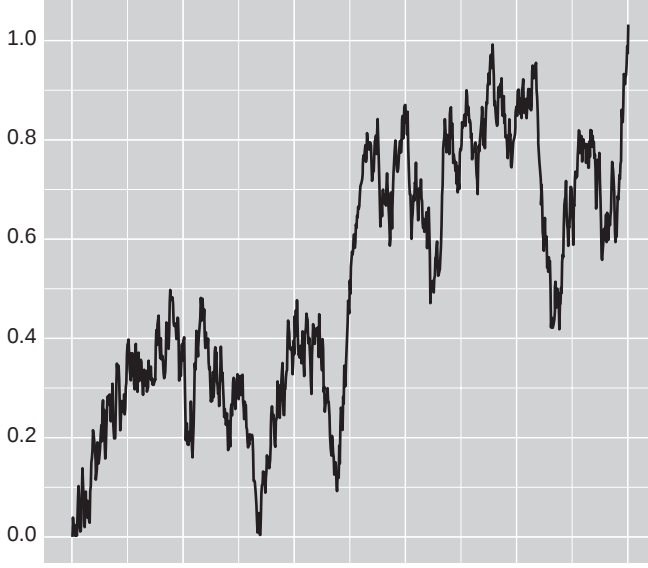


Figure 2. Reflected standard Brownian motion.

A deep and powerful result due to Paul Levy implies that this process has the same law as the process $Y(\cdot)$ in Definition 2, see Ref. 3 for a proof of this beautiful result. The advantage of this new construct is that while Definition 2 does not work for the Brownian motion $X_{\mu,\sigma}(\cdot)$ with general drift and variance, Levy's version does generalize. The reflected version at 0 of this process is defined by

$$Y_{\mu,\sigma}(t) = X_{\mu,\sigma}(t) - \inf_{0 \leq s \leq t} X_{\mu,\sigma}(s) \geq 0$$

To see an example of why this construct is so useful, consider the infinite capacity buffer as in Equation (1) but in the near critical regime, that is, with the deterministic output $\mu = \mu_n$ being near the mean $v = \mathbb{E}(V_k)$ in the sense that

$$\mu_n = v + \frac{\alpha}{\sqrt{n}}$$

for some fixed constant α . Then as before, the workload process

$$\left\{ \frac{1}{\sqrt{n}} W_{[nt]} : t \geq 0 \right\} \quad (3)$$

converges to $Y_{\alpha,\sigma}$ as $n \rightarrow \infty$. Such results give us useful information about the macroscopic (large scale) behavior of queues.

Similarly one can study Brownian motion constrained to lie in an interval $[a, b]$ and reflected at the end points. For example, consider the single buffer model but now with the capacity $C = C_n$, a function of n ,

$$C_n = b\sqrt{n}$$

for some constant b . Then the workload process in Equation (3) converges to Brownian motion reflected both at 0 and b .

GEOMETRIC BROWNIAN MOTION

Brownian motion itself is not a good model of various financial constructs. Geometric Brownian motion is a variant of Brownian motion, which has seen tremendous use in the area of stochastic modeling of processes in finance such as prices of stocks. Let us attempt to motivate the (stochastic) differential equation that a stock price process (referred to as $Z(t)$) might satisfy and then see how this process can be expressed as an explicit function of standard Brownian motion.

The starting point is the assumption that for the price of a stock, the infinitesimal change in the price is given by two components.

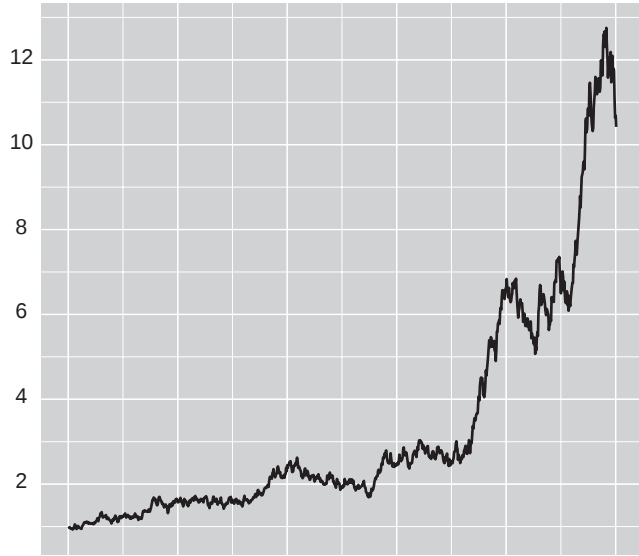


Figure 3. Geometric Brownian motion started at 1 with $\mu = 3$ and $\sigma = 1$.

1. *Deterministic Component.* For the price of the stock one expects constant expected value of return on the investment so that in a small time interval $[t, t + dt)$, the change in the price is proportional to $\mu Z(t)dt$ where μ is a constant measuring the rate of return.
2. *Stochastic Volatility.* The price of the stock is buffeted by a large number of small random factors in any time interval, and the higher the price, the more volatile is the price. This is modeled by thinking of the change in the stock price in a small interval due to random factors as being proportional to $\sigma Z(t)(B(t + dt) - B(t))$, where $B(\cdot)$ is standard Brownian motion.

This intuition yields the equation

$$dZ(t) = \mu Z(t) dt + \sigma Z(t) dB_t. \quad (4)$$

For any arbitrary starting value $Z(0)$, using Ito calculus, one can check that the process

$$Z(t) = Z(0) \exp \left[\left(\mu - \frac{\sigma^2}{2} \right) t + \sigma B_t \right]$$

satisfies Equation (4) and it is called *geometric Brownian motion*. A good reference for more intuition behind the definition of this process is Ref. 4. Note that when $\mu > \sigma^2/2$, the process tends to infinity, while for $\mu < \sigma^2/2$ this process goes to 0, as illustrated in Figs 3 and 4 respectively.

Geometric Brownian motion plays an important part in the famous Black–Scholes formula, derived in the context of option pricing in finance. We refer the interested reader to Chapter 12 of Ref. 4 and associated references for further information. For an in depth look at the mathematical aspects of general exponential functionals of Brownian motion and applications of these in various mathematical models in stochastic finance [5].

BROWNIAN BRIDGE

The Brownian bridge $\{X(t)\}_{0 \leq t \leq 1}$ sometimes referred to as *tied down Brownian motion* is a fundamental object that arises in a number of applications of Brownian process theory in statistics, in particular, the important Kolmogorov–Smirnov test. Intuitively one thinks of the Brownian bridge as standard Brownian motion *conditioned to come back to 0* at time 1, that is, conditioned on $B_1 = 0$. While this conditioning event has measure

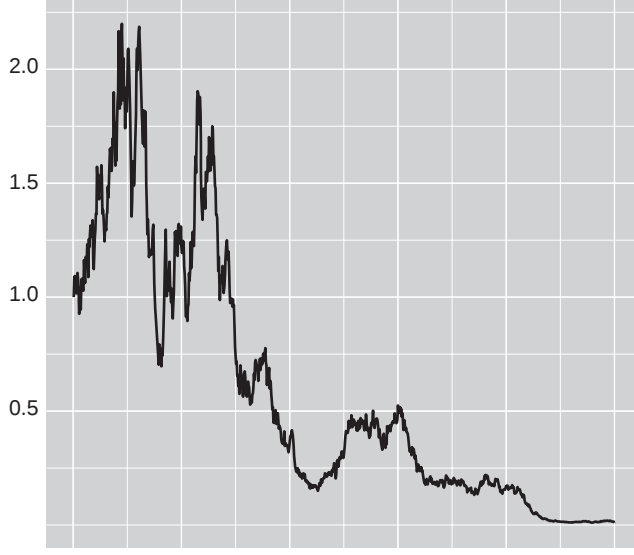


Figure 4. Geometric Brownian motion started at 1 with $\mu = 0$ and $\sigma = 2$.

0, there are a number of ways to make the notion precise. For example, for any $\epsilon > 0$ one could look at the measure of $\mathbb{P}^\epsilon(\cdot)$ of standard Brownian motion conditioned on $|B_1| \leq \epsilon$ (an event with positive probability) and then show that $\mathbb{P}^\epsilon$ are a tight sequence of probability measures on the sequence space $\mathbb{P}^\epsilon(\cdot)$ with a unique limit as $\epsilon \rightarrow 0$ [6]. Alternatively one could solve the stochastic differential equation

$$\begin{aligned} dX(t) &= \frac{-X(t)}{1-t} dt + dB(t), \\ 0 \leq t \leq 1, X_0 &= 0, \end{aligned} \quad (5)$$

where $B(\cdot)$ is standard Brownian motion. For this approach, see Chapter 5 of Ref. 3. Intuitively Equation (5) suggests that as time gets closer to 1, the process is forced to approach 0. One can check that the Brownian bridge as defined by solving Equation (5) has the same distribution as the process

$$\{B(t) - tB(1)\}_{0 \leq t \leq 1}, \quad (6)$$

where $\{B(t) : 0 \leq t \leq 1\}$ is, as usual, a standard Brownian motion. Figure 5 shows a sample path of this process.

The first important property of Brownian bridge, which is a direct consequence of Equation (6), is that it is a Gaussian process

(i.e., a process all of whose finite-dimensional distributions are Gaussian). The mean of the Brownian bridge is given by

$$\mathbb{E}(X_t) = 0 \quad \text{for } 0 \leq t \leq 1,$$

and its covariance function is

$$\rho(s, t) := \text{Cov}(X_s, X_t) = s(1-t), \quad 0 \leq s \leq t.$$

Kolomogorov–Smirnov Test

As mentioned in the introduction, the Brownian bridge is extremely important in empirical process theory in statistics. Here we give the reader an idea of how it arises as the scaling limit of fluctuations of empirical distribution functions about their true value. The setting is as follows: Suppose $\{Y_i\}_{1 \leq i \leq n}$ are independent and identically distributed random variables having a continuous distribution function $F(\cdot)$. Consider the empirical distribution function

$$F_n(t) = \frac{1}{n} \# \{j : Y_j \leq t\}, \quad t \in \mathbb{R}.$$

Define the process $Y_n(t)$

$$Y_n(t) := \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathbb{R},$$

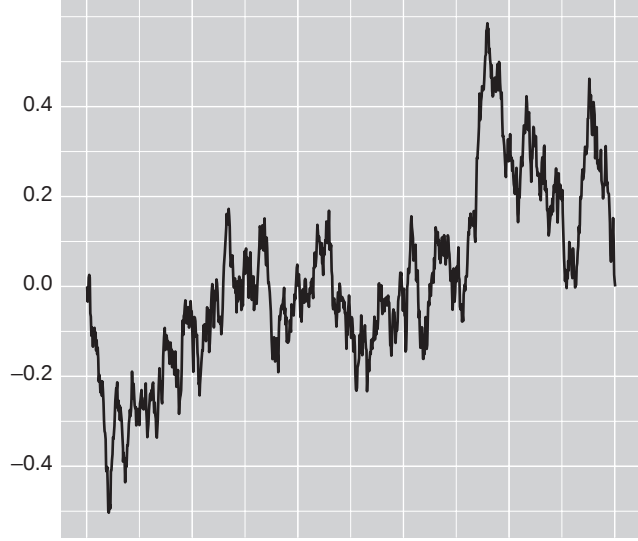


Figure 5. Brownian bridge.

and note that $\lim_{t \rightarrow \pm\infty} Y_n(t) = 0$. Using the standard central limit theorem it is not hard to see that the finite dimensional distributions of the process $Y_n(\cdot)$ converge to those of the process $Y(\cdot)$

$$Y(t) := X(F(t)),$$

where $X(\cdot)$ is the Brownian bridge. One can strengthen the above finite-dimensional convergence to weak convergence of $Y_n(\cdot)$ as a process to $Y(\cdot)$ and, in particular, show that

$$\sup_{t \in \mathbb{R}} |Y_n(t)| \xrightarrow{d} D := \sup_{0 \leq s \leq 1} |X(s)|$$

as $n \rightarrow \infty$. Another calculation implies that for any $x > 0$

$$\mathbb{P}(D > x) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}.$$

These facts are often used to test the hypothesis that a sample $\{Y_i\}_{1 \leq i \leq n}$ comes from some specified continuous distribution F .

The above is just one very small corner of the vast field of empirical process theory. We refer the interested reader to Ref. 7 or 8 for comprehensive treatments of this subject.

ORNSTEIN–UHLENBECK PROCESS

The Ornstein–Uhlenbeck process arises very naturally in various physical systems with resistance, such as the motion of a particle in a fluid facing friction or the vibration of a string. Let us first describe such a model in more detail, then state some rigorous properties of this process, and finally give references. Many of the physical motivations are paraphrased from Chapter 5 of Ref. 9. We shall be a little imprecise when describing the mental picture one has behind this process.

Let $V(t)$ denote the velocity at time t , of a one-dimensional particle with mass m , suspended in a fluid with coefficient of friction β . The frictional force is directly opposite to the direction of motion and is proportional to the magnitude of the velocity. In the absence of random collisions from other molecules in the fluid, the motion of the particle can be described by the differential equation

$$m \frac{dV(t)}{dt} = -\beta V(t).$$

However, the particle experiences a large number of random collisions from other molecules, so that the acceleration has both the above deterministic component and a random component. Assuming that the

particle experiences a large number of very small collisions in any small interval of time $[t, t + dt)$, invoking the central theorem, we conclude that the random component in the change in velocity should be a normal random variable with mean 0 and variance σdt for some constant σ , that is,

$$V(t + dt) - V(t) \sim -\beta V(t) dt + \sigma \mathcal{N}(0, dt),$$

where $\mathcal{N}(0, dt)$ is a normal random variable independent of the present velocity $V(t)$. Now note that (again at a heuristic level), for standard Brownian motion we have

$$B(t + dt) - B(t) = \mathcal{N}(0, dt),$$

so that the above equation can be written as

$$dV_t = -\beta V(t)dt + \sigma dB(t). \quad (7)$$

The solution of the *stochastic* differential equation (7) is the Ornstein–Uhlenbeck process. Each increment of the Ornstein–Uhlenbeck process can be viewed as a combination of a restorative force proportional to the present state of the system and a random perturbation. Solving the above equation with the starting value $V(0)$ having a normal distribution with mean 0 and variance σ^2/β , one can check that $V(t)$ is a stationary Gaussian process, with mean 0 and covariance function

$$\rho(s, t) := \text{Cov}(V(t), V(s)) = \frac{\sigma^2}{2\beta} e^{-\beta|t-s|}. \quad (8)$$

Note that the correlations between time points fall off exponentially in $|t - s|$.

As a model in applied probability, this process arises in a diverse number of contexts ranging from modeling neurons [10], to finance where it has been used in models of evolution of interest rates such as the famous Vasicek model (see the original paper [11] or [12]). For a beautiful introduction to the physical motivations behind the definition of the Ornstein–Uhlenbeck process as described in the beginning of this section, see Chapter 9 of Ref. 13 or Ref. 14.

MULTIDIMENSIONAL BROWNIAN MOTION

Another simple extension of standard Brownian motion is multidimensional Brownian motion. For fixed $d \geq 1$, if $\{B^{(i)}(\cdot) : i \geq 1\}$ are independent Brownian motions then the d -dimensional random process

$$\mathbf{B}(t) = (B^{(1)}(t), B^{(2)}(t), \dots, B^{(d)}(t))$$

is called the d -dimensional Brownian motion. This process arises not only as a model of a particle diffusing through d -dimensional space but has also given rise to beautiful mathematics. In addition it has proven to be very important in applied settings, including modeling of particles diffusing in various physiological and biological settings [1] and in areas such as queueing theory. For example, in two dimensions, the conformal invariance of Brownian motion can be used in a simple proof of the fundamental theorem of algebra. The following argument is based loosely on Ref. 15. Let $Z(t)$ denote complex Brownian motion,

$$Z(t) = B^{(1)}(t) + iB^{(2)}(t),$$

where $i = \sqrt{-1}$. One of the fundamental properties of this process is that for any holomorphic (i.e., differentiable) function f , the process $f(Z(\cdot))$ can be expressed as a continuous time change of another complex Brownian motion, namely, there exists a complex Brownian motion Z' such that

$$f(Z_t) = Z'_{C_t},$$

where

$$C_t = \int_0^t |f'(Z_s)|^2 ds.$$

In particular since the process Z' gets arbitrarily close to the origin, this implies that if $f(\cdot)$ is a polynomial (which automatically implies f is holomorphic) then there exists a point z such that $f(z) = 0$, and thus the polynomial has at least one root. This is the fundamental theorem of algebra. Of course this is just a very simple example of the kind of calculations that are possible. We refer the interested reader to Ref. 16 for a

beautiful mathematical exposition of various path properties of multidimensional Brownian motion (mainly focusing on the planar two-dimensional setting) as well as Chapter 9 of Ref. 17.

Recall that in the section titled “Reflected Brownian Motion”, we explored reflected versions of Brownian motion that arose in the context of a simple model of infinite capacity buffers. Reflected versions of the multidimensional process arise in more complicated models of queueing networks such as tandem queues. Such processes are usually constructed via stochastic differential equations. Even showing the existence of such processes is nontrivial in a number of settings. See Ref. 18 or the comprehensive books [2,19] for various applications in queueing theory and the mathematical issues that arise in constructing such processes in general domains.

BROWNIAN EXCURSIONS

The concept of Brownian excursions away from 0 is easy to talk about but much more nontrivial to define mathematically. The set of times

$$\mathcal{O} = \{t : B_t = 0\}$$

is a perfect set, that is, each point is a limit point of other such points. However if one looks at $\mathcal{O}^c$, then due the path continuity of Brownian motion, this set is open and can be expressed as a disjoint countable union of open intervals $\cup_j (a_j, b_j)$. Owing to the extreme irregularity of $\mathcal{O}$, we cannot speak about the “first” excursion of Brownian motion away from 0, the second, and so on. The set of excursions is the set of path segments between the 0-points, that is, we decompose the path of Brownian motion according to its excursions:

$$\{B((t + a_j) \wedge b_j) : t \geq 0\}$$

each of which is a random element of the set of continuous functions started at 0 and ending at 0 in finite time without visiting 0 in between these two times. It turns out that these paths plotted against the local time of Brownian motion at 0 form a *Poisson* point process, with an explicitly computable intensity measure called the *Ito Excursion*

measure. This fact leads to a very powerful technique for exploring the path properties of Brownian motion. Further, if one wants a notion of Brownian motion, started at 0 and conditioned to return to 0 at time $t = 1$ and always above the x axis, then a natural candidate for such an object is the following: fix any time $t > 0$. Almost surely, $B_t \neq 0$. Let α_t be the last time *before* t that the Brownian motion was at 0 and let β_t be the first time *after* t that Brownian motion visits 0 and let $\gamma_t = \beta_t - \alpha_t$ be the length of this excursion. Then a natural candidate for this process

$$e(s) = \frac{1}{\sqrt{\gamma_t}} |B(\alpha_t + s\gamma_t)|, \quad 0 \leq s \leq 1.$$

Owing to the time and scale invariance properties of Brownian motion, the choice of the fixed time t is of no consequence.

Let us give an example of how this process arises as the scaling limit of discrete random objects and thus gives an enormous amount of asymptotic information about these random structures. Trees and random models of trees play a very important role in many branches of science, ranging from algorithms in computer science to genealogies in biology.

See Fig. 6 for an example of a planar tree and a coding of this tree by a function that represents the distance traveled at any time by an individual walking at rate 1 along the boundary of the tree in lexicographical order, starting from the root (i.e., in the direction of the arrows). Now given a random tree (e.g., the uniform measure on all planar trees on n vertices), in many situations we are interested in understanding the asymptotics of such models as $n \rightarrow \infty$. Note that the contour process coding of a tree is a process that starts at 0 and eventually returns to 0 in $2(n - 1)$ steps without ever going below the x axis. For a random tree, one gets a random function satisfying these properties. What is more difficult, but is now a classical result, is that for a wide range of random tree models, the contour process properly rescaled in time and space converges to the standard Brownian Excursion. See Refs 20 and 21 or the comprehensive Ref. 22 for applications to a diverse number of models from combinatorics ranging from random trees to random graphs and random maps.

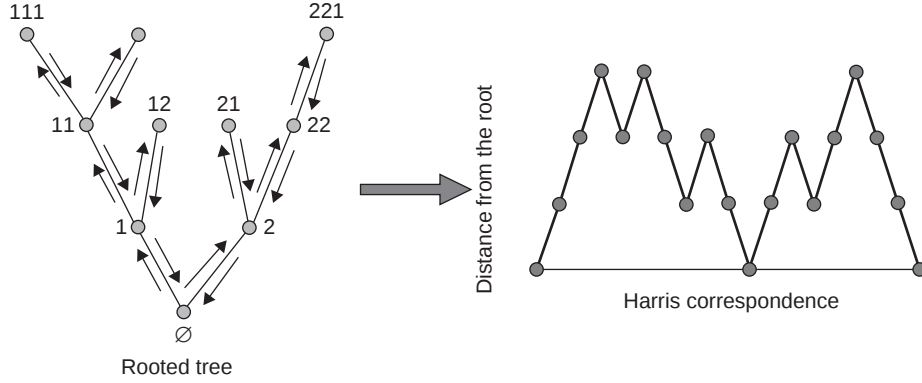


Figure 6. Planar tree on 10 vertices and the corresponding contour process.

This fundamental process and its variants such as multidimensional local time, Brownian excursions conditioned on its own local time, and nonintersecting Brownian excursions have arisen in many other diverse contexts ranging from applications in finance such as Parisian options Ref. 23, to mathematical studies of the fine scale structure of multidimensional Brownian motion [16]. For a textbook treatment of this process see Ref. 24, while Refs 25 and 26 give nice introductory treatments.

FRACTIONAL BROWNIAN MOTION

In many applications one is interested in constructing a stochastic model of how a time series evolves over time, e.g. the amount of traffic flowing through a link in a very large data transmission network such as the Internet, various economic time series such as speculative prices, and so on. Models such as the Ornstein–Uhlenbeck process defined in the section titled “Ornstein–Uhlenbeck Process” prove to be insufficient for such applications. Equation (7) implies that for the Ornstein–Uhlenbeck process, the correlation between $V(t)$ and $V(t+h)$ drops down exponentially quickly. This is not the case for many empirically observed time series, where strong temporal correlations are observed over long periods of time, staying above or below prescribed levels for much longer than one would expect under the assumptions

leading to the construction of processes such as the Ornstein–Uhlenbeck process.

Fractional Brownian motion is a variant of Brownian motion, which attempts to fill this gap. The process depends on a parameter H which controls how quickly the correlations decay. For any $0 < H \leq 1$, fractional Brownian motion $\{B_H(t) : t \in \mathbb{R}\}$ is a Gaussian process with mean 0 and covariance function given by

$$\rho_H(s, t) = \frac{1}{2} (|s|^{2H} + |t|^{2H} - |t - s|^{2H}) \quad (9)$$

Note that $H = 1/2$ corresponds to usual standard Brownian motion.

For any $H \in (0, 1)$ this process has continuous but nondifferentiable sample paths with stationary increments (i.e., for any h , the distribution of $B_H(t+h) - B_H(t)$ depends only on h and not on t). The paths of these processes get smoother as H increases from 0 to 1. One can see this by looking at the above covariance function given by Equation (8) and noticing that for any $s > 0$ and any fixed t , the consecutive increments $B_H(t+s) - B_H(t)$ and $B_H(t+2s) - B_H(t+s)$ are negatively correlated for $H < 1/2$ so that consecutive increments have opposite signs, while for $H > 1/2$ these increments are positively correlated. See Figs 7 and 8 for examples of the roughness of the sample paths for $H < 0.5$, while see Figs 9 and 10 for the relatively smoother sample paths for $H > 0.5$.

Another very important property of this process is that it is self-similar with index H , that is, for any fixed $a > 0$

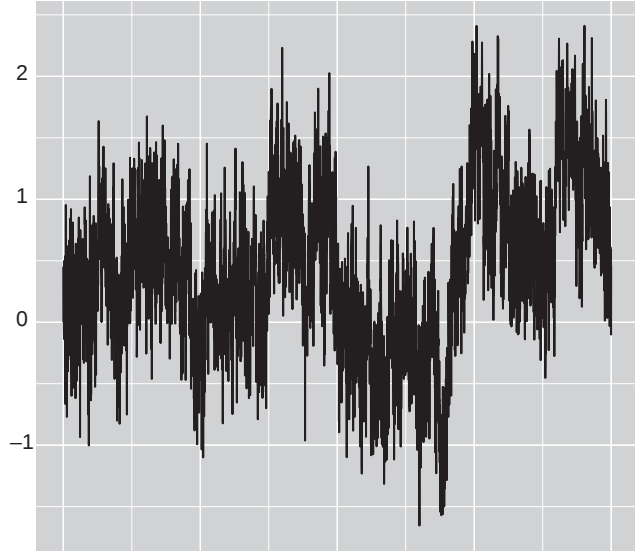


Figure 7. Fractional Brownian motion with Hurst parameter 0.1.

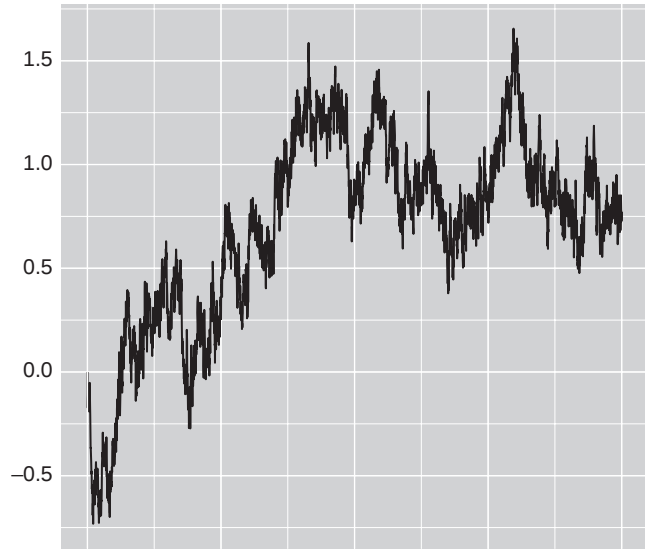


Figure 8. Fractional Brownian motion with Hurst parameter 0.35.

$$\{B_H(at)\}_{t \in \mathbb{R}} = a^H \{B_H(t)\}_{t \in \mathbb{R}}.$$

To get an intuition as to how such correlations arise and how this process arises due to long-range correlations between increments, we quote the following result from Ref. 27. Suppose $X_k = \sum_{j=-\infty}^{\infty} c_{k-j} \eta_j$, where $\{\eta_j\}_{j \in \mathbb{Z}}$ are independent and identically distributed mean 0 and finite variance and the constants satisfy $\sum_{j=-\infty}^{\infty} c_j^2 < \infty$. Suppose we define

the random walk $S_n = \sum_{k=1}^n X_k$ for $n \geq 1$. If $\text{Var}(S_n) = n^{2H} L^2(n)$ for $0 < H < 1$ and $L(\cdot)$ is a slowly varying function at infinity, then

$$\frac{1}{n^H L(n)} \sum_{k=1}^{\lfloor nt \rfloor} X_k \rightarrow B_H(t)$$

in the sense of finite-dimensional convergence.

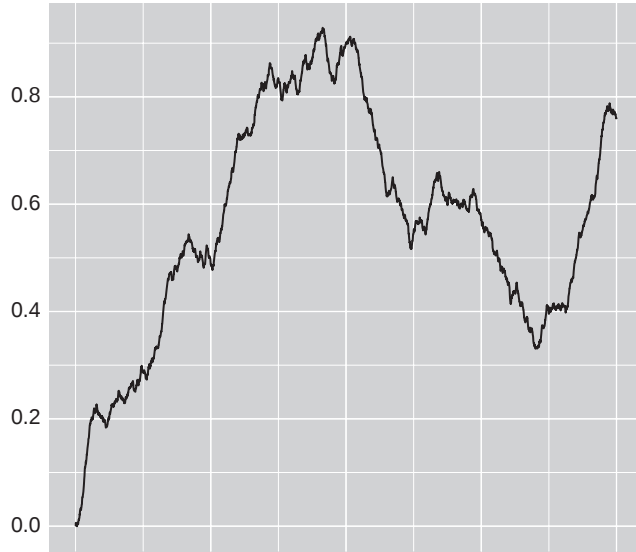


Figure 9. Fractional Brownian motion with Hurst parameter 0.75.

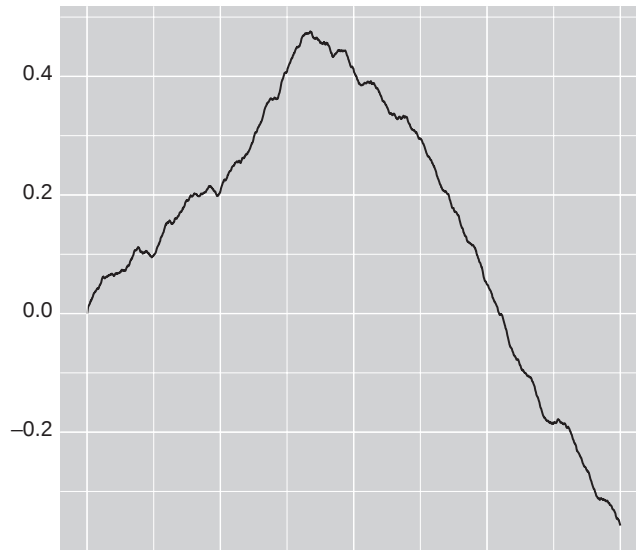


Figure 10. Fractional Brownian motion with Hurst parameter 0.95.

While this process had already arisen in the work of Kolmogorov, it was Mandelbrot and his coworkers who formulated this as a good model for many processes that arise in various applications such as finance and hydrology [28–30]. See Ref. 31 for a comprehensive account of the mathematical properties of this process, its history, and its wide range of applicability.

BROWNIAN SHEET

Almost all the processes that have been introduced in this article are indexed by time, that is, are of the form X_t for $t \in \mathbb{R}^+$. However, in the applied context, one is often interested in “Brownian processes” that are indexed by more general sets. One of the most famous of such examples is the Brownian sheet. We

shall first introduce this model, then mention how this can be viewed as the continuum limit of discrete processes (similar to viewing the Brownian motion as the continuum limit of random walks), and then give further references. We shall define the simplest version of this model, namely, the *one*-dimensional version.

The Brownian sheet with values in $\mathbb{R}$ is a two-parameter, centered (i.e. mean 0), Gaussian process $\{B(s, t) : s, t \geq 0\}$ with covariance function given by

$$\begin{aligned} \mathbb{E}(B(s, t)B(s', t')) \\ = \min(s, s') \min(t, t'), s, s', t, t' \geq 0. \end{aligned}$$

The above process can be viewed as the scaling limit of a discrete random field on the planar two-dimensional lattice. Consider $\mathbb{Z}_+^2 = \{(i, j) : i, j \geq 0\}$ and let $\{\eta_{ij} : i, j \geq 0\}$ be an independent and identically distributed array of random variables with mean μ and finite variance σ^2 . One could think of η_{ij} as the amount of insect damage at site i, j in a large forest. Thus for any $n, m \geq 0$

$$S(n, m) = \sum_{0 \leq i \leq n} \sum_{0 \leq j \leq m} \eta_{ij}$$

represents the total insect damage in the rectangle with opposite vertices $(0, 0)$ and (n, m) . By the central limit theorem, for any fixed (s, t) the fluctuations of the above random field satisfy

$$\frac{S(\lfloor Ns \rfloor, \lfloor Nt \rfloor) - N^2 st \mu}{\sigma^2 N} \xrightarrow{d} B(s, t)$$

as $N \rightarrow \infty$ and in fact we have finite-dimensional convergence of the processes $Z_N(\cdot, \cdot)$ defined as

$$Z_N(s, t) = \frac{S(\lfloor Ns \rfloor, \lfloor Nt \rfloor) - N^2 st \mu}{\sigma^2 N}, s, t \geq 0$$

to the one-dimensional Brownian bridge as $N \rightarrow \infty$. Under various assumptions one can show that the above convergence of finite-dimensional distributions can be strengthened to weak convergence of the two-dimensional processes.

It appears that this process was first constructed due to motivations from statistics,

in particular, to construct continuous models of analysis of variance [32]. For more on the early motivations and results about this process we refer the interested reader to the papers of Pyke [33, 34]. For a readable modern account see Ref. 35; for a more comprehensive textbook treatment see Ref. 36. Finally, for fascinating connections between this process and an analysis of models of the stochastic wave equation see Chapter 1 of Ref. 37.

OTHER IMPORTANT VARIATIONS

We conclude this article by mentioning a few other variants of Brownian motion, which have proven to be enormously useful in modeling various physical systems, especially at criticality but for which providing a full description is well nigh impossible in an introductory article of this nature.

Brownian Motion on Other State Spaces

Most of the processes that we have looked at above relate to Brownian motion or associated diffusions on the real line or in the Euclidean setting. However, there is a tremendous amount of interest, especially in the physics literature in formulating, rigorously constructing, and analyzing models of diffusion on more general spaces of disordered media, including various fractals such as the Sierpinski gasket illustrated in Fig. 11. The usual model of Brownian motion in Euclidean space could represent a model of transport and flow through a regular crystalline structure. However, many of the materials that physicists consider are far from regular. Physical structures such as porous fractured rock structures and superionic conductors are disordered media. The behavior of diffusing particles in such media is quite different from usual Brownian motion. For example while the usual random walk in the plane scales like $\sqrt{n}$, in the sense that the typical distance of a particle from its starting point after n steps is of the order of $\sqrt{n}$, many models of diffusion in such disordered media show “anomalous diffusion” in the sense that the distance traveled after n steps scales as some other power, n^β . See Ref. 38 for a fascinating survey of a

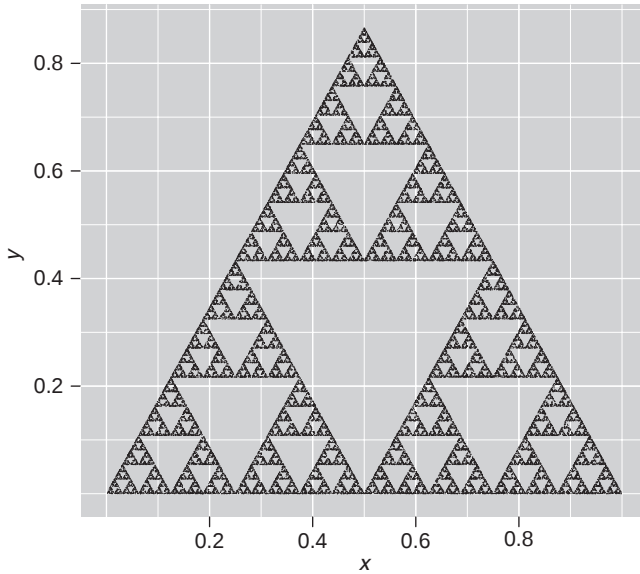


Figure 11. Sierpinski triangle.

wide array of models and settings that arise from physical motivations, while see Ref. 39 and the references therein for a survey level treatment of the rigorous mathematical results that one can derive in these contexts and connections to other techniques such as the connection between random walks and electrical networks. For diffusion models on general tree-like fractals see Ref. 40 or the comprehensive set of lectures [41].

Schramm–Loewner Evolution

One of the reasons Brownian motions and simple variants such as reflected Brownian motion are so important is that they arise as scaling limits in many important models, such as random walks and queueing systems and thus allow us to understand the macroscopic (large scale) structure of such processes. Statistical physicists are interested in the macroscopic properties of various two-dimensional systems at criticality (a parameter value at which the behavior of the system abruptly changes), including ferromagnetic systems described by the Ising and Potts model and polymer molecules.

Schramm–Loewner evolutions are a one-parameter family of growing random fractal curves in two dimensions indexed by a parameter κ grown using standard one-dimensional Brownian motion and Loewner

chains from complex analysis. The stochastic version was invented by Oded Schramm [42]. These random planar curves have proven to be enormously useful as scaling limits of many fundamental discrete planar models in two dimensions. Although it will be too difficult to describe the actual process, let us give one concrete but easily describable discrete process, whose scaling limit is such a random planar curve.

Loop Erased Random Walk

Consider the discrete integer lattice $\mathbb{Z}^2$. Since we want to understand macroscopic properties of a large random object, consider for any fixed $\delta > 0$ the graph $\delta\mathbb{Z}^2$. Let D denote the disc of unit radius in $\mathbb{R}^2$ and consider the graph $H_\delta = D \cap \delta\mathbb{Z}^2$. Now consider a simple random walk on H_δ , started from the origin and run until the first time it hits the boundary of H_δ , and let T_δ denote this time. Suppose the path of random walk is $(S_0, S_1, \dots, S_{T_\delta})$. Note that for this path, it is quite possible that there will be times when the path will intersect itself, namely, k times such that there exists $m < k$ such that $S_m = S_k$. The *loop erasure* of this path is defined as the erasure of the loops of this path in chronological order. More precisely, define $L_0 = S_0$ and for each $j \geq 0$ define $n_j = \max\{k \leq T_\delta : S_k = L_j\}$ and define $L_{j+1} = S_{n_j+1}$. We continue this

process till $j = \sigma$ such that $L_\sigma = S_{T_\delta}$. The loop erased path is given by $(L_1, L_2, \dots, L_\sigma)$.

The loop erased random walk is an extremely important process, arising in diverse areas of combinatorics and statistical physics, including the generation of spanning trees uniformly at random from graphs. In the above context, one is interested in what happens to the macroscopic (large-scale) structure of the path as the mesh size δ goes to 0. It turns out that as $\delta \rightarrow 0$, the path of loop erased walk converges to a Schramm–Loewner evolution with $\kappa = 2$. We refer the reader to Ref. 43 (by Wendelin Werner who eventually won the Fields medal for his work in this area), or Ref. 44 for an introduction to this rapidly growing field and for the diverse number of contexts where this process arises (or is conjectured to arise) as the scaling limit.

REFERENCES

1. Grebenkov DS. NMR survey of reflected Brownian motion. *Rev Mod Phys* 2007;79(3):1077–1137.
2. Whitt W. Stochastic-process limits. New York: Springer; 2002.
3. Karatzas I, Shreve SE. Brownian motion and stochastic calculus. New York: Springer; 1991.
4. Oksendal B. Stochastic differential equations. Berlin: Springer; 1998.
5. Yor M. Exponential functionals of Brownian motion and related processes. Berlin: Springer; 2001.
6. Billingsley P. convergence of probability measures. Society for Industrial Mathematics; New York: Wiley; 1968.
7. Shorack GR, Wellner JA. Empirical processes with applications to statistics, Wiley series in probability and mathematical statistics: probability and mathematical statistics. New York: John Wiley & Sons, Inc.; 1986.
8. Van der Vaart AW, Wellner JA. Weak convergence and empirical processes. New York: Springer; 1996.
9. Bhattacharya RN, Waymire EC. Stochastic processes with applications. New York: Princeton University Press; 2001.
10. Ricciardi LM, Sacerdote L. The Ornstein-Uhlenbeck process as a model for neuronal activity. *Biol Cybern* 1979;35(1):1–9.
11. Vasicek O. An equilibrium characterization of the term structure. *J Finance Econ* 1977;5(2):177–188.
12. Hull JC. Options, futures, and other derivatives. India: Pearson Education India; 2008.
13. Nelson E. Dynamical theories of Brownian motion. Citeseer; 1967.
14. Chandrasekhar S. Stochastic problems in physics and astronomy. *Rev Mod Phys* 1943;15(1):1–89.
15. Rogers LCG, Williams D. Diffusions, Markov processes, and martingales. Volume 1, Cambridge mathematical library. Cambridge: Cambridge University Press; 2000. Foundations, Reprint of the second (1994) edition.
16. Le Gall J-F. Volume 1527, Some properties of planar Brownian motion. Berlin: Springer; 1992.
17. Mörters P, Peres Y. Brownian motion. Cambridge: Cambridge University Press; 2010.
18. Harrison JM, Reiman MI. Reflected Brownian motion on an orthant. *Ann Probab* 1981;9(2):302–308.
19. Harrison JM. Brownian motion and stochastic Flow Systems. New York: John Wiley & Sons, Inc.; 1985.
20. Aldous D. The continuum random tree II: an overview, stochastic analysis (Proceedings of the Durham Symposium on Stochastic Analysis, 1990), London mathematical society lecture note series 167. Cambridge: Cambridge University Press; 1991. pp. 23–70.
21. Le Gall JF. Random trees and applications. *Probab Surv* 2005;2:245–311.
22. Pitman J. Combinatorial stochastic processes. Berlin: Springer; 2006.
23. Chesney M, Jeanblanc-Picqué M, Yor M. Brownian excursions and Parisian barrier options. *Adv Appl Probab* 1997;29(1):165–184.
24. Revuz D, Yor M. Continuous martingales and Brownian motion. Berlin: Springer; 1999.
25. Chung KL. Excursions in Brownian motion. *Ark Matematik* 1976;14(1):155–177.
26. Rogers LCG. A guided tour through excursions. *Bull Lond Math Soc* 1989;21(4):305.
27. Davydov YA. The invariance principle for stationary processes. *Teoriya Veroyatnostei i ee Primeneniya* 1970;15(3):498–509.
28. Mandelbrot BB, Van Ness JW. Fractional Brownian motions, fractional noises and applications. *SIAM Rev* 1968;10(4):422–437.
29. Mandelbrot BB. Fractals and scaling in finance: discontinuity, concentration, risk:

- selecta volume E. New York: Springer-Verlag; 1997.
30. Mandelbrot BB. Multifractals and $1/F$ noise: wild self-affinity in physics (1963–1976): selecta volume N. New York: Springer-Verlag; 1999.
 31. Doukhan P, Oppenheim G, Taqqu MS. Theory and applications of long-range dependence. Boston: Birkhauser; 2003.
 32. Kitagawa T. Analysis of variance applied to function spaces. *Mem Fac Sci Kyushu Univ Ser A Math* 1951;6(1):41–53.
 33. Pyke R. Partial sums of matrix arrays, and Brownian sheets. *Stochastic analysis (a tribute to the memory of Rollo Davidson)*. London: John Wiley; 1973. pp. 331–348.
 34. Pyke R. A uniform central limit theorem for partial-sum processes indexed by sets. Technical report no. 17. Department of Statistics, University of Washington, Seattle: Citeseer; 1982.
 35. Khoshnevisan D. Five lectures on Brownian sheet: summer internship program. Madison (WI): University of Wisconsin–Madison; 2001.
 36. Khoshnevisan D. Multiparameter processes: an introduction to random fields. New York: Springer; 2002.
 37. Walsh J. An introduction to stochastic partial differential equations. *École d'Été de Probabilités de Saint Flour XIV-1984*. Berlin: Springer; 1984. pp. 265–439.
 38. Havlin S, Ben-Avraham D. Diffusion in disordered media. *Adv Phys* 2002;51(1):187–292.
 39. Barlow M. Diffusions on fractals. *Lectures on probability theory and statistics*. Berlin: Springer; 1998. pp. 1–121.
 40. Evans SN. Snakes and spiders: Brownian motion on R-trees. *Probab Theory Relat Fields* 2000;117(3):361–386.
 41. Evans SN. Probability and real trees. Volume 1920, *Lecture notes in mathematics*. Berlin: Springer; 2008.
 42. Schramm O. Scaling limits of loop-erased random walks and uniform spanning trees. *Isr J Math* 2000;118(1):221–288.
 43. Werner W. Random planar curves and Schramm-Loewner Evolutions, in 2002 St-Flour summer school. Volume 1840, *Lecture notes in mathematics*. Berlin: Springer; 2004. pp. 107–195.
 44. Lawler GF. Conformally invariant processes in the plane. Providence, RI: American Mathematical Society; 2008.

VARIATIONAL INEQUALITIES

IGOR V. KONNOV
Department of System Analysis
and Information Technologies,
Kazan University, Kazan,
Russia

BASIC DEFINITIONS

Let X be a convex set in the n -dimensional Euclidean space $\mathbb{R}^n$ and let $F : X \rightarrow \mathbb{R}^n$ be a mapping. Then we can define the *variational inequality* (VI): Find $x^* \in X$ such that

$$\langle F(x^*), x - x^* \rangle \geq 0 \quad \forall x \in X. \quad (1)$$

Investigation of VI as a separate problem is associated with the works of Fichera [1] and Stampacchia [2]. They introduced VIs for studying certain problems in mathematical physics; see also Duvaut and Lions [3] and Kinderlehrer and Stampacchia [4] for more details. At the same time, there exist several earlier works in different fields which contain similar problem formulations. For instance, the well-known linear programming problem [5] can be treated as VI with the constant mapping F . Afterwards, a great number of applications of VIs appeared in economics, operations research, telecommunications, and many other fields [6–11]. Now VI is treated as the differential form of general equilibrium conditions [12]. The theory and methods for VIs are investigated very extensively and great advances were made recently in this field [10,11,13,14]. In this article, we intend to describe some of these results. We also discuss their possible extensions for the multivalued case.

THEORETICAL BACKGROUND

In this section, we consider VI (1) under the standing assumptions that X is a nonempty, convex, and closed set in $\mathbb{R}^n$ and that

$F : X \rightarrow \mathbb{R}^n$ is a continuous mapping. Then we can deduce the existence of a solution for Equation (1) from the Brouwer fixed point theorem if we additionally suppose X is bounded [4]. If the set X need not be bounded, we have to utilize the so-called coercivity conditions.

Theorem 1 [15,16]. *Suppose that there exists a nonempty bounded subset Y of X such that for every $x \in X \setminus Y$ there is $y \in Y$ with $\langle F(x), x - y \rangle \geq 0$. Then VI (1) has a solution.*

More general existence results are contained, for example, in papers [17–21].

We now recall some norm monotonicity concepts for mappings.

Definition 1. A mapping $Q : X \rightarrow \mathbb{R}^n$ is said to be

- (a) *strongly monotone* with modulus $\tau' \geq 0$ if, for each pair of points $x, y \in X$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle \geq \tau' \|x - y\|^2;$$

- (b) *strictly monotone*, if, for each pair of points $x, y \in X, x \neq y$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle > 0;$$

- (c) *monotone* if, for each pair of points $x, y \in X$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle \geq 0;$$

- (d) *pseudomonotone* if, for each pair of points $x, y \in X$, it holds that

$$\langle Q(y), x - y \rangle \geq 0 \quad \text{implies}$$

$$\langle Q(x), x - y \rangle \geq 0;$$

- (e) *co-coercive* (or *inverse strongly monotone*) with modulus $\tau'' > 0$, if, for each pair of points $x, y \in X$, it holds that

$$\langle Q(x) - Q(y), x - y \rangle \geq \tau'' \|Q(x) - Q(y)\|^2.$$

Clearly, we have (a) $\implies$ (b) $\implies$ (c) $\implies$ (d), and (e) $\implies$ (c). In the affine case $Q(x) = Ax - b$ is (strictly, strongly) monotone if and only if A is a positive (definite) semidefinite matrix. In the differentiable case, the monotonicity concepts are associated with the Jacobian properties [10, Proposition 2.3.2]. Also, each co-coercive mapping with modulus τ'' is Lipschitz continuous with constant $1/\tau''$, and each strongly monotone and Lipschitz continuous mapping is co-coercive, but the reverse assertions are not true in general [22]. Clearly, co-coercivity is equivalent to strong monotonicity of the inverse mapping. Concerning pseudomonotonicity [23], see also Hadjisavvas *et al.* [24] for further development of generalized monotonicity. The monotonicity properties enable us to specialize existence and uniqueness results.

Proposition 1.

- (i) If F is strictly monotone, then VI (1) has at most one solution.
- (ii) If F is strongly monotone, then VI (1) has a unique solution.

It is known that VI (1) is closely related to the following problem of finding $x^* \in X$ such that

$$\langle F(x), x - x^* \rangle \geq 0 \quad \forall x \in X. \quad (2)$$

Problem (2) seems to have appeared first in Debrunner and Flor [25] and may be termed as the *minimax dual* formulation of VI (in short, DVI); see, for example, Konnov [12] for more details. We denote by X^* (respectively, by X^d) the solution set of problem (1) [respectively, problem (2)]. Under the assumptions made, the set X^d is convex and closed. The following assertion is a slightly extended form of the Minty lemma [26], utilizing the pseudomonotonicity.

Proposition 2.

- (i) It holds that $X^d \subseteq X^*$.
- (ii) If F is pseudomonotone, then $X^* \subseteq X^d$.

Further development of this direction can be found, for example, in papers [14,18,19,24]. Besides, most results above can be extended to infinite-dimensional spaces; see papers [16,18–20,24] for more details.

RELATED PROBLEMS

It is easy to see that VI (1) with $X = \mathbb{R}^n$ reduces to the equation

$$F(x^*) = 0.$$

If X is a convex cone in $\mathbb{R}^n$, VI (1) reduces to the *complementarity problem* (CP)

$$x^* \in X, F(x^*) \in X^*, \langle F(x^*), x^* \rangle = 0,$$

where $X^* = \{z \in \mathbb{R}^n \mid \langle x, z \rangle \geq 0 \quad \forall x \in X\}$ is the conjugate cone to X [27]. Conversely, in the case when the feasible set is defined by convex functions VI (1) can be reduced to CP via the Karush–Kuhn–Tucker optimality conditions [6,11].

Next, let T be a continuous mapping from a convex set X into itself. Then the *fixed point problem*, which is to find a point $x^* \in X$ such that

$$x^* = T(x^*) \quad (3)$$

coincides with VI (1), where

$$F(x) = x - T(x); \quad (4)$$

see Kinderlehrer and Stampacchia [[4], Section 3.1].

Definition 2 [4,28]. A mapping $T : X \rightarrow X$ is said to be

- (a) *strongly contractive* with modulus $\kappa \in (0, 1)$ if for each pair of points x, y , we have

$$\|T(x) - T(y)\| \leq \kappa \|x - y\|;$$

- (b) *contractive* if for each pair of points $x \neq y$, we have

$$\|T(x) - T(y)\| < \|x - y\|;$$

- (c) *nonexpansive* if for each pair of points x, y , we have

$$\|T(x) - T(y)\| \leq \|x - y\|.$$

Clearly, we have (a) $\implies$ (b) $\implies$ (c), also, F in Equation (4) is co-coercive with modulus 0.5 if T is nonexpansive. Moreover, F in Equation (4) is strictly (strongly) monotone if T is (strongly) contractive. So, contractive properties of T in the fixed point problem imply monotonicity ones in VI (1) and (4); see also Vasin and Ageev [29] and Berinde [28] for more details. However, by using the projection mapping $\pi_X(\cdot)$ onto X , we can make some reverse transformations. Take a number $\theta > 0$ and set $T(x) = \pi_X[x - \theta F(x)]$. Then VI (1) becomes equivalent to problem (3) or to the equation $G(x^*) = 0$ with $G(x) = x - \pi_X[x - \theta F(x)]$. However, this transformation can weaken monotonicity properties. In fact, if F is co-coercive with modulus $\mu > 0$, then G is co-coercive with modulus $\mu' = 1 - \theta/(4\mu)$ only if $\theta \in (0, 4\mu)$ [[22], Chapter 5].

Let us now consider the *optimization problem* of finding a point $x^* \in X$ such that $\varphi(x^*) \leq \varphi(x)$ for all $x \in X$, or briefly,

$$\min_{x \in X} \rightarrow \varphi(x), \quad (5)$$

and suppose for simplicity that $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable function. Then Equation (5) implies VI (1), where $F(x) = \nabla \varphi(x)$, and the reverse implication holds if φ is (pseudo)convex. Note that the function φ is strongly convex (respectively, strictly convex, convex, pseudoconvex) if and only if its gradient map $\nabla \varphi$ is strongly monotone (respectively, strictly monotone, monotone, pseudomonotone) [22,24]. In this potential case, the mapping $F = \nabla \varphi$ possesses additional properties. For instance, if $\nabla \varphi$ is monotone and Lipschitz continuous, then it is co-coercive [22]. Moreover, the Jacobian $\nabla F(x) = \varphi''(x)$ must be symmetric.

Let us consider the *constrained l -person noncooperative game* with the common feasible set $X \subset \mathbb{R}^n$, where the i th player has a utility function $h_i : X \rightarrow \mathbb{R}$, which is concave in its i th variable u_i for $i = 1, \dots, m$, and $x = (u_1, \dots, u_l)^\top \in \mathbb{R}^n$. A point $x^* = (u_1^*, \dots, u_l^*)^\top \in X$ is said to be an equilibrium point for this game, if

$$h_i(u_{-i}^*, u_i) \leq h_i(x^*) \quad \forall (u_{-i}^*, u_i) \in X, i = 1, \dots, m.$$

Again suppose for simplicity that all the utility functions are differentiable. If we define

the mapping $F : X \rightarrow \mathbb{R}^n$ by

$$F_i(x) = -\frac{\partial h_i(u_{-i}, u_i)}{\partial u_i} \quad \text{for } i = 1, \dots, m; \quad (6)$$

then the equilibrium point problem becomes equivalent to VI (1) and (6) [30]. If additionally, X is the Cartesian product of the strategy sets, that is,

$$X = U_1 \times \dots \times U_l,$$

then the above concept is nothing but the known Nash equilibrium point [31]. Clearly, it reduces to a concave-convex saddle point problem if we also suppose that $l = 2$ and that the game is antagonistic, that is, $h_1(x) + h_2(x) = 0$, and involves in particular the Lagrangian saddle point problem in constrained convex optimization [6,10,11,22,32,33].

ALGORITHMS

Let us consider VI (1) where X is a nonempty, convex and closed set in $\mathbb{R}^n$ and $F : X \rightarrow \mathbb{R}^n$ is a continuous mapping. The linearization algorithms based on the sequential solution of auxiliary affine VIs constitute one of the most popular and investigated classes if F satisfies certain strengthened monotonicity properties. At the k th iteration of such an algorithm, given a point $x^k \in X$, the next iterate x^{k+1} is defined as a solution of VI: Find $\tilde{x} \in X$ such that

$$\langle F^k(\tilde{x}), x - \tilde{x} \rangle \geq 0 \quad \forall x \in X, \quad (7)$$

where $F^k(x) = F(x^k) + C_k(x - x^k)$, C_k is an $n \times n$ positive definite matrix. Different choices of C_k leads to various iterative methods. For instance, setting $C_k = \nabla F(x^k)$ gives the Newton method, setting $C_k = C$ with a symmetric and positive definite matrix C gives the (extended) projection method. Besides, one can take C_k as an approximation of $\nabla F(x^k)$, thus obtaining quasi-Newton methods. Concerning the convergence of the linearization algorithms we note that the Newton method attains the highest quadratic convergence rate, but requires rather restrictive assumptions for providing

global convergence [34, Corollary 2.6]. The simplest projection method corresponds to Equation (7) with $F^k(x) = F(x^k) + \lambda_k^{-1}(x - x^k)$, $\lambda_k > 0$ and is equivalent to the iterate

$$x^{k+1} = \pi_X[x^k - \lambda_k F(x^k)]. \quad (8)$$

It provides global convergence under co-coercivity of F and attains a linear convergence rate if F is strongly monotone and Lipschitz continuous, but requires knowledge of the corresponding constants for finding the stepsize λ_k .

In order to guarantee global convergence without *a priori* information, one can replace the initial VI by an optimization problem via introduction of some artificial *merit (gap) function*, which admits algorithms with linesearch procedures [10,11,13,35]. This approach is extended directly for more general mixed VIs and will be described in the section titled “Mixed Variational Inequalities”. However, the above methods fail to provide global convergence under only usual monotonicity conditions.

In order to essentially weaken convergence assumptions one can utilize suitable extrapolation procedures. The *extragradient method* [36] consists in replacing the usual projection iteration (8) with the double projection iteration

$$x^{k+1} = \pi_X(x^k - \beta_k F(y^k)), \quad y^k = \pi_X(x^k - \beta_k F(x^k)). \quad (9)$$

If F is a Lipschitz continuous mapping with modulus L , $X^* = X^d \neq \emptyset$, and $\beta_k = \beta \in (0, 1/L)$, then iterates in Equation (9) will converge to a solution of VI (1). In Khobotov [37], an implementable variant of the extragradient method, which finds the double stepsize β_k simultaneously was also proposed. Some other method to utilizing extrapolation is proposed in the so-called *combined relaxation (CR) methods* [38]. In this method, the auxiliary iteration (7) serves for computing parameters of a hyperplane separating, strictly, the current iterate x^k and the solution set X^* , whereas the main iteration consists in making the projections onto this hyperplane and onto the feasible set, if necessary. So, both the iterations

are different. We give now one of the CR algorithms [14,38].

Algorithm (CR). Choose a point $x^0 \in X$, numbers $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\gamma \in (0, 2)$. At the k th iteration, $k = 0, 1, \dots$, compute $\tilde{x}$ in Equation (7) and set $p^k = \tilde{x} - x^k$. If $p^k = 0$, stop. Otherwise determine m as the smallest nonnegative integer such that $\langle F(x^k + \beta^m p^k), p^k \rangle \leq \alpha \langle F(x^k), p^k \rangle$, set $y^k = x^k + \beta^m p^k$, $g^k = F(y^k)$, $\omega_k = \langle g^k, x^k - y^k \rangle$. If $g^k = 0$, stop. Otherwise set

$$x^{k+1} = \pi_X[x^k - \gamma(\omega_k / \|g^k\|^2)g^k].$$

These algorithms possess similar convergence properties, but do not require for L to be evaluated or even for F to be Lipschitz continuous. At the same time, the extrapolation methods attain a linear rate of convergence [14,39]. Similar methods can be found in Solodov and Tseng [40] and Sun [41]; see also Konnov [14,42] and Auslender [43].

Another approach consists in utilizing certain *regularization* techniques. For instance, the classical Tikhonov–Browder regularization [44,45] leads to the following auxiliary problem: Find $x^\varepsilon \in X$ such that

$$\langle F(x^\varepsilon) + \varepsilon x^\varepsilon, x - x^\varepsilon \rangle \geq 0 \quad \forall x \in X, \quad (10)$$

where $\varepsilon > 0$. In the case when F is monotone and continuous, and $X^* \neq \emptyset$, it is known [45,46] that each problem (10) has a unique solution and that the sequence of solutions $\{x^\varepsilon\}$ converges to the minimal norm solution x_n^* of VI (1) as $\varepsilon \rightarrow 0$, this property remaining valid in Banach spaces. Note also that the cost mapping $F^\varepsilon = F + \varepsilon I$ in Equation (10) is then strongly monotone. By utilizing suitable coercivity conditions one can maintain weakened convergence properties of regularization schemes in the nonmonotone case [47–51]. For instance, if the assumptions of Theorem 1 are fulfilled then each VI (10) has a solution, each sequence $\{x^{\varepsilon_k}\}$ has limit points, and if $\{\varepsilon_k\} \searrow 0$ all these limit points are solutions of VI (1) [51,52]. Some recent developments in this field can be found in Konnov [20] and Konnov and Dyabilkin [21].

The proximal point method [53,54] also consists in replacing the initial problem with

a sequence of perturbed auxiliary ones, but its regularization parameter may be in principle fixed. Given a point x^k , the next iterate is defined as a solution of VI: Find $x^{k+1} \in X$ such that

$$\langle F(x^{k+1}) + \lambda^{-1}(x^{k+1} - x^k), x - x^{k+1} \rangle \geq 0 \\ \forall x \in X,$$

with $\lambda > 0$. Again, if F is a monotone and $X^* \neq \emptyset$, the cost mapping above is strongly monotone and $\{x^k\}$ converges to a solution of VI (1). Moreover, its convergence can be established under certain generalized monotonicity conditions [55–58]. Although the auxiliary problems in regularization methods admit approximate solutions, they also may be rather expensive.

It should be noted that various constrained optimization methods are extended directly for VIs with nonlinear constraints, including exterior and interior penalty methods [59–63], dual Uzawa and multiplier type methods [22, 64–69], and hybrid schemes [10, 70, 71]. Besides, these methods are adjusted easily for GVI (12) with nonlinear constraints.

MIXED VARIATIONAL INEQUALITIES

Let X be again a nonempty, convex, and closed set in $\mathbb{R}^n$, $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ a continuous mapping, and $h: \mathbb{R}^n \rightarrow \mathbb{R}$ a convex continuous function. The mixed variational inequality problem (MVI) is the problem of finding a point $x^* \in X$ such that

$$\langle F(x^*), x - x^* \rangle + h(x) - h(x^*) \geq 0 \quad \forall x \in X. \quad (11)$$

Problem (11) was introduced in Lescarret [72] and Browder [73] and studied by many authors due to its various applications [3, 13, 14, 63]. In case $h \equiv 0$, it corresponds to the usual VI (1), whereas $F \equiv \mathbf{0}$ leads to the convex (nonsmooth) optimization problem

$$\min_{x \in X} h(x).$$

Besides, Equation (11) is equivalent to GVI (12) with the cost mapping $F + \partial h$, where

∂f denotes the subdifferential mapping of h . Nevertheless, peculiarities of MVI (11) enable one to investigate it via methods from the usual VIs. For instance, the assertions of Proposition 1 remain true for Equation (11).

By using some artificial gap (or merit) function, one can convert MVI (11) into an equivalent constrained optimization problem. The simplest regularized gap function is defined as follows:

$$\varphi_\lambda(x) = \max_{y \in X} \Phi_\lambda(x, y),$$

where $\lambda > 0$,

$$\Phi_\lambda(x, y) = \langle F(x), x - y \rangle - \|x - y\|^2 / (2\lambda) + h(x) - h(y).$$

Under the above assumptions, there exists the unique element $y_\lambda(x) \in X$ such that $\varphi_\lambda(x) = \Phi_\lambda(x, y_\lambda(x))$. It follows that $\varphi_\lambda(x) \geq 0$ for every $x \in X$, besides,

$$\varphi_\lambda(x) = 0 \text{ and } x \in X \iff x \in X^*,$$

where X^* denotes the solution set of MVI (11). Such functions for VIs were proposed in Auchmuty [74] and Fukushima [75] and extended for MVIs in Patriksson [76]. Note that the iterate $x^{k+1} = y_\lambda(x^k)$ corresponds to the forward–backward splitting method [77, 78], which clearly reduces to the projection method (8) if $h \equiv 0$ and has similar convergence properties. However, we can now utilize the descent approach with respect to φ_λ for selection of the stepsize.

Algorithm (DSE). Choose a point $x^0 \in X$ and a number $\lambda > 0$. At the k th iteration, $k = 0, 1, \dots$, we have a point $x^k \in X$, and set $p^k = y_\lambda(x^k) - x^k$. If $p^k = \mathbf{0}$, stop. Otherwise, find $\alpha_k \in [0, 1]$ such that $\varphi_\lambda(x^k + \alpha_k p^k) = \min_{\alpha \in [0, 1]} \varphi_\lambda(x^k + \alpha p^k)$, and set $x^{k+1} = x^k + \alpha_k p^k$.

Theorem 2 [13, Theorem 6.1; 76]. Suppose that F is continuously differentiable and strongly monotone on X . Then Algorithm (DSE) either terminates with a unique solution of MVI (11) or generates an infinite

sequence $\{x^k\}$, which converges to this unique solution.

A descent method with an Armijo type inexact linesearch procedure was proposed in Konnov [79]. Such methods can be based on different and more general gap functions [13,80]. They can utilize the Newton auxiliary problem instead of the projection step [13,81,82] and are adjusted for the case of nonlinear constraints [13,82,83]. One of the most interesting is the so-called D -gap function, which was proposed in Peng [84] for the usual VI when $h \equiv 0$. In Konnov [85], this gap function was applied for MVI (11). The D -gap function is defined as follows

$$\psi_{\lambda\mu}(x) = \varphi_\lambda(x) - \varphi_\mu(x),$$

where $0 < \mu < \lambda$. It follows that MVI (11) is equivalent to the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} \rightarrow \psi_{\lambda\mu}(x).$$

Besides, it was shown in Konnov [85] that, unlike the usual gap functions, $\psi_{\lambda\mu}$ remains differentiable if F is so, regardless of the function h . Next, if $\nabla F(x)$ is positive definite on $\mathbb{R}^n$, then MVI (11) becomes equivalent to the equation:

$$\nabla \psi_{\lambda\mu}(x) = \mathbf{0},$$

which can be solved via efficient descent methods without computation of derivatives.

All the above descent methods require certain strengthened monotonicity of the mapping F . In the case when the mapping F is only monotone, but not strongly monotone, one can utilize several approaches to provide convergence. For instance, if the set X is bounded, the above descent methods can be inserted in a two-level parametric iterative scheme [13,86]. Next, they can be combined with either regularization or proximal point methods, with approximate solution of their auxiliary subproblems [80,87,88]. Some other way consists in applying the *(CR) methods* [89]; see also Konnov [14,80].

Algorithm (CRM). Choose a point $x^0 \in \mathbb{R}^n$, a sequence of functions $\psi_k : \mathbb{R}^n \rightarrow \mathbb{R}$ such

that ψ_k is differentiable and strongly convex with constant $\tau' > 0$, and its gradient ψ'_k is Lipschitz continuous with constant $\tau'' < \infty$. Choose numbers $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\gamma \in (0, 2)$. At the k th iteration, $k = 0, 1, \dots$, find m as the smallest nonnegative integer such that

$$\begin{aligned} & \langle F(x^k) - F(z^{k,m}), x^k - z^{k,m} \rangle \\ & \leq (1 - \alpha)\beta^{-m} \langle \psi'_k(z^{k,m}) - \psi'_k(x^k), z^{k,m} - x^k \rangle, \end{aligned}$$

where $z^{k,m}$ is a solution of the problem

$$\begin{aligned} \min_{y \in X} \rightarrow & \left\{ \langle F(x^k) - \beta^{-m} \psi'_k(x^k), y \rangle \right. \\ & \left. + \beta^{-m} \psi_k(y) + h(y) \right\}. \end{aligned}$$

Set $\theta_k = \beta^m$, $y^k = z^{k,m}$. If $y^k = x^k$, stop. Otherwise set $g^k = F(y^k) - F(x^k) - \theta_k^{-1}[\psi'_k(y^k) - \psi'_k(x^k)]$, $\omega_k = \langle g^k, x^k - y^k \rangle$, $x^{k+1} = x^k - \gamma(\omega_k / \|g^k\|^2)g^k$.

Note that we can simply choose $\psi_k(x) = 0.5\|x\|^2$, then computation of $z^{k,m}$ is again equivalent to the forward-backward splitting iteration with $\lambda = \beta^m$.

Theorem 3 [80, Theorem 6.1]. Suppose that F is monotone and $X^* \neq \emptyset$. Then Algorithm (CRM) either terminates with a solution of MVI (11) or generates an infinite sequence, which converges to a solution of MVI (11).

This CR method attains a linear rate of convergence if F is strongly monotone. For further development, see Konnov [90–92].

GENERALIZED VARIATIONAL INEQUALITIES

Now we turn to the multivalued extension of VI (1). This requires proper generalization of continuity type concepts. In what follows, $\Pi(S)$ denotes the family of all subsets of a set S .

Definition 3. [[14], Definition 2.1.1] Let X be a convex set in $\mathbb{R}^n$. A mapping $Q : X \rightarrow \Pi(\mathbb{R}^n)$ is said to be a K (Kakutani)-mapping if it is upper semicontinuous and has nonempty, convex, and compact values.

In this section we suppose that X is a nonempty, convex, and closed set in $\mathbb{R}^n$ and that $\mathbf{F}: X \rightarrow \Pi(\mathbb{R}^n)$ is a multivalued K -mapping. Then we can define the *generalized variational inequality problem* (GVI): Find an element $x^* \in X$ such that

$$\exists f^* \in \mathbf{F}(x^*), \langle f^*, x - x^* \rangle \geq 0 \quad \forall x \in X. \quad (12)$$

The necessity to investigate and solve such problems stems from the fact that many applied equilibrium-type models involve just multivalued mappings [8,14,16,24,42,51,93]. Together with GVI (12) we can define the *minimax dual problem* (DGVI): Find $x^* \in X$ such that

$$\forall f \in \mathbf{F}(x), \langle f, x - x^* \rangle \geq 0 \quad \forall x \in X; \quad (13)$$

compare Equation (2), which also seems to have appeared first in Debrunner and Flor [25]; see, for example, Konnov [14,12] for more details. We again denote by X^* (respectively, by X^d) the solution set of problem (12) (respectively, problem (13)). The basic existence and uniqueness results from the section titled “Theoretical Background” contained in Theorem 1 and Propositions 1 and 2 are transformed directly for GVI (12), after the proper adjustment of the monotonicity type concepts [12,14,16,51]. We also notice that the results described in the section titled “Algorithms” for the regularization and proximal point methods remain true in the multivalued case [14,22,46,51,54,58,71,94]. Note that these methods only replace the initial problem with a regularized one, but do not remove the basic difficulty to solve a multivalued GVI. At the same time, the streamlined extensions of some other solution methods described in the sections titled “Algorithms” and “Mixed Variational Inequalities”, such as descent and linearization ones may fail to provide convergence. For instance, the analog of the projection method (8) which is defined by

$$x^{k+1} = \pi_X[x^k - \lambda_k f^k], f^k \in \mathbf{F}(x^k),$$

again requires the strengthened monotonicity, but its stepsize λ_k must conform to the divergent series rule, which leads to

rather slow convergence; see, for example, Gol’shtein and Tret’yakov [22] and Konnov [12] for more details. It should be noted that the *averaging* [95,96] and *iterative regularization type methods* [97,46] provide convergence under the usual monotonicity, but require the same divergent series step size rule.

Besides, one can utilize *center type methods*, whose idea consists in sequential selection of the next iterate as an approximate center of a subset containing a solution combined with a proper reduction of this subset. The most known is the ellipsoid method, which was proposed first by Yudin and Nemirovskii [98] and Shor [99] for convex programming and afterwards adjusted for saddle point problems and VIs [100,101]. The analytic center methods and proximal level (or bundle) methods also follow this direction [102–106]. These methods usually include solution of rather difficult auxiliary problems at each iteration, but also enable one to weaken convergence assumptions.

We now turn to the *CR methods* for GVI (12) [14,107–109]. Suppose in addition that the feasible set X is defined by

$$X = \{x \in \mathbb{R}^n \mid h(x) \leq 0\}, \quad (14)$$

where $h: \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex, but not necessarily differentiable function, and that the Slater constraint qualification is satisfied, that is, there exists a point $\tilde{x}$ such that $h(\tilde{x}) < 0$. Set

$$\mathbf{P}(x) = \begin{cases} \mathbf{F}(x) & \text{if } h(x) < 0, \\ \text{conv}\{\mathbf{F}(x) \cup \partial h(x)\} & \text{if } h(x) = 0, \\ \partial h(x) & \text{if } h(x) > 0. \end{cases}$$

One of suitable CR methods is then described as follows.

Algorithm (CRG). Choose a point $x^0 \in \mathbb{R}^n$, numbers $\alpha \in (0, 1)$, $\gamma \in (0, 2)$, and sequences $\{\lambda_l\} \searrow 0$ and $\{\eta_l\} \searrow 0$. Set $k = 0$, $l = 1$.

Step 1a: Choose q^0 from $\mathbf{P}(x^k)$, set $i = 0$, $p^i = q^i$.

Step 1b: If $\|p^i\| \leq \eta_l$, set $x^{k+1} = x^k$, $k = k + 1$, $l = l + 1$ and go to Step 1a.

Step 1c: Set $z^{i+1} = x^k - \lambda_i p^i / \|p^i\|$, choose $q^{i+1} \in \mathbf{P}(z^{i+1})$. If $\langle q^{i+1}, p^i \rangle > \theta \|p^i\|^2$, then set $y^k = z^{i+1}$, $g^k = q^{i+1}$, and go to Step 2. Otherwise, set $p^{i+1} = \pi_{\text{conv}\{p^i, q^{i+1}\}}(\mathbf{0})$, $i = i + 1$, and go to Step 1b.

Step 2: Set $\omega_k = \langle g^k, x^k - y^k \rangle$, $x^{k+1} = x^k - \gamma(\omega_k / \|g^k\|^2)g^k$, $k = k + 1$ and go to Step 1a.

It was shown in Konnov [107,108] (see also [[14], Sections 2.3 and 2.4]) that Algorithm (CRG) converges to a solution of GVIs (12) and (14) if $X^* = X^d \neq \emptyset$ and attains a linear rate of convergence (with respect to the index i) under certain additional assumptions.

REFERENCES

1. Fichera G. Problemi elastostatici con vincoli unilaterali; il problema di Signorini con ambigue al contorno. Atti Acad Naz Lincei Mem Cl Sci Fis Mat Natur Sez I 1964;8:91–140.
2. Stampacchia G. Formes bilinéaires coercitives sur le ensembles convexes. CR Acad Sci (Paris) 1964;258:4413–4416.
3. Duvaut G, Lions J-L. Les Inéquations en mécanique et physique. Paris: Dunod; 1972.
4. Kinderlehrer D, Stampacchia G. An introduction to variational inequalities and their applications. New York: Academic Press; 1980.
5. Dantzig GB. Linear programming and extensions. Princeton: Princeton University Press; 1963.
6. Harker PT, Pang J-S. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Math Program 1990;48:161–220.
7. Giannessi F, Maugeri A, editors. Variational inequalities and network equilibrium problems. New York: Plenum Press; 1995.
8. Ferris MC, Pang J-S. Engineering and economic applications of complementarity problems. SIAM Rev 1997;39:669–713.
9. Nagurney A. Network economics: a variational inequality approach. Dordrecht: Kluwer Academic Publishers; 1999.
10. Facchinei F, Pang J-S. Volumes I and II, Finite-dimensional variational inequalities and complementarity problems. Berlin: Springer; 2003.
11. Konnov IV. Equilibrium models and variational inequalities. Amsterdam: Elsevier; 2007.
12. Konnov IV. Generalized monotone equilibrium problems and variational inequalities. In: Hadjisavvas N, Komlósi S, Schaible S, editors. Handbook of generalized convexity and generalized monotonicity. New York: Springer; 2005. pp. 559–618.
13. Patriksson M. Nonlinear programming and variational inequality problems: a unified approach. Dordrecht: Kluwer Academic Publishers; 1999.
14. Konnov IV. Combined relaxation methods for variational inequalities. Berlin: Springer; 2001.
15. Eaves BC. On the basic theorem of complementarity. Math Program 1971;1:68–75.
16. Isac G. Complementarity problems. Berlin: Springer; 1992.
17. Isac G. A generalization of Karamardian's condition in complementarity theory. Nonlinear Anal Forum 1999;4:49–63.
18. Bianchi M, Hadjisavvas N, Schaible S. Minimal coercivity conditions and exceptional families of elements in quasimonotone variational inequalities. J Optim Theory Appl 2004;122:1–17.
19. Bianchi M, Pini R. Coercivity conditions for equilibrium problems. J Optim Theory Appl 2005;124:79–92.
20. Konnov IV. Regularization method for non-monotone equilibrium problems. J Nonlinear Convex Anal 2009;10:93–101.
21. Konnov IV, Dyabilkin DA. Nonmonotone equilibrium problems: coercivity conditions and weak regularization. J Glob Optim. DOI: 10.1007/s10898-010-9551-7. In press.
22. Gol'shtein EG, Tret'yakov NV. Augmented lagrange functions. Moscow: Nauka; 1989. (Engl. transl. in John Wiley and Sons, New York, 1996.)
23. Karamardian S. Complementarity over cones with monotone and pseudomonotone maps. J Optim Theory Appl 1976;18: 445–454.
24. Hadjisavvas N, Komlósi S, Schaible S, editors. Handbook of generalized convexity and generalized monotonicity. New York: Springer; 2005.
25. Debrunner H, Flor P. Ein erweiterungssatz für monotone mengen. Arch Math 1964;15: 445–447.

26. Minty G. On the generalization of a direct method of the calculus of variations. *Bull Am Math Soc* 1967;73:315–321.
27. Karamardian S. Generalized complementarity problems. *J Optim Theory Appl* 1971; 8:161–167.
28. Berinde V. Iterative approximation of fixed points. Berlin: Springer; 2007.
29. Vasin VV, Ageev AL. Ill-posed problems with a priori information. Ekaterinburg: Nauka; 1993. (Engl. transl. in VSP, Utrecht, 1995.)
30. Rosen JB. Existence and uniqueness of equilibrium points for concave n -person games. *Econometrica* 1965;33:520–534.
31. Nash J. Non-cooperative games. *Ann Math* 1951;54:286–295.
32. von Neumann J, Morgenstern O. Game theory and economic behaviour. Princeton: Princeton University Press; 1953.
33. Arrow KJ, Hurwicz L, Uzawa H. Studies in linear and nonlinear programming. Stanford: Stanford University Press; 1958.
34. Pang J-S, Chan D. Iterative methods for variational and complementarity problems. *Math Program* 1982;24:284–313.
35. Fukushima M. Merit functions for variational inequality and complementarity problems. In: Di Pillo G, Giannessi F, editors. *Nonlinear optimization and applications*. New York: Plenum Press; 1996. pp. 155–170.
36. Korpelevich GM. Extragradient method for finding saddle points and other problems. *Matecon* 1976;12:747–756.
37. Khobotov EN. Modification of the extragradient method for solving variational inequalities and certain optimization problems. *USSR Comput Math Math Phys* 1987;27:120–127.
38. Konnov IV. Combined relaxation methods for finding equilibrium points and solving related problems. *Russ Math (Iz VUZ)* 1993;37(2):44–51.
39. Tseng P. On linear convergence of iterative methods for the variational inequality problem. *J Comput Appl Math* 1995;60: 237–252.
40. Solodov MV, Tseng P. Modified projection-type methods for monotone variational inequalities. *SIAM J Control Optim* 1996;34:1814–1830.
41. Sun D. A class of iterative methods for solving nonlinear projection equations. *J Optim Theory Appl* 1996;91:123–140.
42. Konnov IV. Combined relaxation methods for generalized monotone variational inequalities. In: Konnov IV, Luc DT, Rubinov AM, editors. *Generalized convexity and related topics*. Berlin: Springer; 2007. pp. 3–31.
43. Auslender A, Teboulle M. Interior projection-like methods for monotone variational inequalities. *Math Program* 2005; 104:39–68.
44. Tikhonov AN. On the solution of ill-posed problems and regularization method. *Dokl Akad Nauk SSSR* 1963;151:501–504.
45. Browder FE. Existence and approximation of solutions of nonlinear variational inequalities. *Proc Natl Acad Sci U S A* 1966;56:1080–1086.
46. Bakushinskii AB, Goncharskii AV. Iterative solution methods for ill-posed problems. Moscow: Nauka; 1989 (in Russian).
47. Liskovets OA. Regularized variational inequalities with pseudo-monotone operators on approximately given sets. *Differ Equat* 1989;25:1970–1977 (in Russian).
48. Gwinner J. A note on pseudomonotone functions, regularization, and relaxed coerciveness. *Nonlinear Anal: Theory Method Appl* 1997;30:4217–4227.
49. Isac G, Bulavski VA, Kalashnikov VV. Complementarity, equilibrium, efficiency and economics. Dordrecht: Kluwer Academic Publishers; 2002.
50. Alber Y, Butnariu D, Ryazantseva I. Regularization and resolution of monotone variational inequalities with operators given by hypomonotone approximations. *J Nonlinear Convex Anal* 2005;6:23–53.
51. Konnov IV, Ali MSS, Mazurkevich EO. Regularization of nonmonotone variational inequalities. *Appl Math Optim* 2006;53: 311–330.
52. Konnov IV. On the convergence of a regularization method for variational inequalities. *Comput Math Math Phys* 2006;46: 541–547.
53. Martinet B. Regularization d'inéquations variationnelles par approximations successives. *Rev Fr Inform Rech Opér* 1970;4: 154–159.
54. Rockafellar RT. Monotone operators and the proximal point algorithm. *SIAM J Control Optim* 1976;14:877–898.
55. Billups SC, Ferris MC. QPCOMP: a quadratic programming based solver for mixed complementarity problems. *J Optim Theory Appl* 1997;76:533–562.

56. El Farouq N. Pseudomonotone variational inequalities: convergence of proximal methods. *J Optim Theory Appl* 2001; 109:311–326.
57. Konnov IV. Application of the proximal point method to nonmonotone equilibrium problems. *J Optim Theory Appl* 2003; 119:317–333.
58. Allevi E, Gnudi A, Konnov IV. The proximal point method for nonmonotone variational inequalities. *Math Method Oper Res* 2006;63:553–565.
59. Gwinner J. On the penalty method for constrained variational inequalities. In: Hiriart-Urruty J-B, Oettli W, Stoer J, editors. *Optimization: theory and algorithms*. New York: Marcel Dekker; 1981. pp. 197–211.
60. Muu LD. An augmented penalty function method for solving a class of variational inequalities. *Zhurn Vych Mat Mat Fiz* 1986;26:1788–1796.
61. Dussault J-P. Numerical stability and efficiency of penalty algorithms. *SIAM J Numer Anal* 1995;32:296–317.
62. Auslender A. Asymptotic analysis for penalty and barrier methods in variational inequalities. *SIAM J Control Optim* 1999;37:653–671.
63. Konnov IV. Mixed variational inequalities and exact penalty functions. In: Lapin A, Laitinen E, editors. *Numerical methods for continuous casting and related problems*. Proceedings of Lobachevskii's Mathematical Center. Kazan: DAS; 2001. pp. 59–63.
64. Rockafellar RT. Monotone operators and augmented Lagrangian methods in nonlinear programming. In: Mangasarian OL, et al., editors. *Nonlinear programming 3*. New York: Academic Press; 1978. pp. 1–25.
65. Rockafellar RT. Lagrange multipliers and variational inequalities. In: Cottle RW, et al., editors. *Variational inequalities and complementarity problems: theory and applications*. Chichester: John Wiley & Sons; 1980. pp. 303–322.
66. Auslender A, Teboulle M. Lagrangian duality and related multiplier methods for variational inequality problems. *SIAM J Optim* 2000;10:1097–1115.
67. Konnov IV. Approximate methods for primal-dual mixed variational inequalities. *Russ Math (Iz VUZ)* 2000;44(12):55–66.
68. Konnov IV. The Lagrange multiplier technique for variational inequalities. *Comput Math Math Phys* 2001;41:1279–1291.
69. Konnov IV. The method of multipliers for nonlinearly constrained variational inequalities. *Optimization* 2002;51:907–926.
70. Makler-Scheimberg S, Nguyen VH, Strodiot JJ. Family of perturbation methods for variational inequalities. *J Optim Theory Appl* 1996;89:423–452.
71. Kaplan A, Tichatschke R. Some results about proximal-like methods. In: Seeger A, editor. *Recent advances in optimization*. Berlin: Springer; 2006. pp. 61–86.
72. Lescarret C. Cas d'addition des applications monotones maximales dan un espace de Hilbert. *CR Acad Sci (Paris)* 1965; 261:1160–1163.
73. Browder FE. On the unification of the calculus of variations and the theory of monotone nonlinear operators in Banach spaces. *Proc Natl Acad Sci U S A* 1966;56:419–425.
74. Auchmuty G. Variational principles for variational inequalities. *Numer Funct Anal Optim* 1989;10:863–874.
75. Fukushima M. Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. *Math Program* 1992;53:99–110.
76. Patriksson M. Merit functions and descent algorithms for a class of variational inequality problems. *Optimization* 1997;41:37–55.
77. Lions PL, Mercier B. Splitting algorithms for the sum of two monotone operators. *SIAM J Numer Anal* 1979;16:964–979.
78. Gabay D. Application of the method of multipliers to variational inequalities. In: Fortin M, Glowinski R, editors. *Augmented Lagrangian methods: application to the numerical solution of boundary-value problems*. Amsterdam: North-Holland Publishing Co.; 1983. pp. 299–331.
79. Konnov IV. Descent method with inexact linesearch for mixed variational inequalities. *Russ Math (Iz VUZ)* 2009;53(8):29–35.
80. Konnov IV. Iterative solution methods for mixed equilibrium problems and variational inequalities with non-smooth functions. In: Haugen IN, Nilsen AS, editors. *Game theory: strategies, equilibria, and theorems*. Hauppauge: NOVA; 2008. pp. 117–160.
81. Taji K, Fukushima M, Ibaraki T. A globally convergent Newton method for solving strongly monotone variational inequalities. *Math Program* 1993;58:369–383.
82. Taji K, Fukushima M. A globally convergent Newton method for solving variational inequality problems with inequality

- constraints. In: Zhu DZ, Qi L, Womersley R, editors. Recent advances in nonsmooth optimization. River Ridge: World Scientific Publishers; 1995. pp. 405–417.
83. Taji K, Fukushima M. A new merit function and a successive quadratic programming algorithm for variational inequality problems. *SIAM J Optim* 1996;6:704–713.
84. Peng J-M. Equivalence of variational inequality problems to unconstrained minimization. *Math Program* 1997;78:347–355.
85. Konnov IV. On a class of D-gap functions for mixed variational inequalities. *Russ Math (Iz VUZ)* 1999;43(12):60–64.
86. Zhu DL, Marcotte P. Modified descent methods for solving the monotone variational inequality. *Oper Res Lett* 1993;14:111–120.
87. Konnov IV, Kum S. Descent methods for mixed variational inequalities in a Hilbert space. *Nonlinear Anal: Theory Method Appl* 2001;47:561–572.
88. Konnov IV, Kum S, Lee GM. On convergence of descent methods for variational inequalities in a Hilbert space. *Math Method Oper Res* 2002;55:371–382.
89. Konnov IV. Combined relaxation method for solving variational inequalities with monotone operators. *Comput Math Math Phys* 1999;39:1051–1056.
90. Konnov IV. A combined relaxation method for a class of nonlinear variational inequalities. *Optimization* 2002;51:127–143.
91. Konnov IV. A combined relaxation method for nonlinear variational inequalities. *Optim Method Softw* 2002;17:271–292.
92. Konnov IV. Splitting-type method for systems of variational inequalities. *Comput Oper Res* 2006;33:520–534.
93. Nikaido H. Convex structures and economic theory. New York: Academic Press; 1968.
94. Antipin AS. Equilibrium programming: proximal methods. *Comput Math Math Phys* 1997;37:1285–1296.
95. Bruck R. On weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *J Math Anal Appl* 1977;61:159–164.
96. Passty GB. Ergodic convergence to zero of the sum of two monotone operators in Hilbert space. *J Math Anal Appl* 1979;72:383–390.
97. Bakushinskii AB, Polyak BT. On the solution of variational inequalities. *Sov Math Dokl* 1974;15:1705–1710.
98. Yudin DB, Nemirovskii AS. Informational complexity and efficient methods for the solution of convex extremal problems, I, II. *Matecon* 1977;13:25–45,550–559.
99. Shor NZ. Cut-off method with space dilation in convex programming problems. *Cybernetics* 1977;13:94–96.
100. Shor NZ. New development trends in nonsmooth optimization methods. *Cybernetics* 1977;6:87–91.
101. Nemirovskii AS. Effective iterative methods for solving equations with monotone operators. *Ekon Matem Met (Matecon)* 1981;17:344–359.
102. Gol'shtein EG, Nemirovskii AS, Nesterov YE. Level method, its extensions and applications. *Ekon Matem Met (Matecon)* 1995;31:164–181 (in Russian).
103. Lemaréchal C, Nemirovskii A, Nesterov Y. New variants of bundle methods. *Math Program* 1995;69:111–147.
104. Kiwiel KC. Proximal level bundle methods for convex nondifferentiable optimization, saddle point problems and variational inequalities. *Math Program* 1995;69:89–109.
105. Magnanti TL, Perakis G. A unifying geometric framework and complexity analysis for variational inequalities. *Math Program* 1995;71:327–352.
106. Goffin J-L, Marcotte P, Zhu DL. An analytic center cutting plane method for pseudomonotone variational inequalities. *Oper Res Lett* 1999;20:1–6.
107. Konnov IV. On combined relaxation method's convergence rates. *Russ Math (Iz VUZ)* 1993;37(12):89–92.
108. Konnov IV. A general approach to finding stationary points and the solution of related problems. *Comput Math Math Phys* 1996;36:585–593.
109. Konnov IV. A combined relaxation method for variational inequalities with nonlinear constraints. *Math Program* 1998;80:239–252.

VENDOR-MANAGED INVENTORY

MICHAEL J. FRY

Department of Quantitative Analysis
and Operations Management,
University of Cincinnati,
Cincinnati, Ohio

Vendor-managed inventory (VMI) refers to a supply chain agreement where an upstream agent, such as a supplier or manufacturer, takes control of the inventory management decisions for one or more downstream agents, such as retailers. For simplicity, we will refer to this upstream agent that gains decision making rights as the “supplier” and the downstream agents foregoing their decision making as “retailers.” Under VMI, the supplier makes decisions concerning the timing and quantity of deliveries of material to the retailer. In some cases, the supplier may also take ownership of these items even when they are located at the retailer—an arrangement known as *consignment*—but this is not required under VMI. VMI programs utilize information sharing between the retailer and the supplier in order to facilitate the matching of inventory with customer demand. In a typical system, retailers will periodically or continuously provide data to the supplier, which describes customer demand and current inventory levels. The supplier will then decide when and how much inventory to send to each retailer. In most cases, VMI programs will be subject to an initiating contract that specifies terms of the agreement including performance measures and payment terms.

VMI programs have been in existence for quite some time. One of the earliest large implementations of VMI occurred in the 1980s between Procter & Gamble and Wal-Mart. This VMI program has been widely praised as being successful in reducing costs for both the supplier (Procter & Gamble) and the retailer (Wal-Mart). Other examples of successful VMI implementations have been documented at Barilla SpA, Guinness Beer, and Campbell Soup Company in the

grocery industry as well as agreements in the electronics industry in companies such as Intel and Lexmark and between Boeing and Alcoa in heavy manufacturing. In general, VMI programs have been applied across a wide variety of industries and settings. There is ample evidence of supply chain savings linked to VMI programs, but savings are not guaranteed—particularly when one examines the individual agent costs in the supply chain.

The most common supply chain savings associated with VMI agreements are reductions in inventory levels and increased service levels. For instance, Buzzell and Ortmeyer [1] report sales increases of up to 25% and inventory turn increases of 30% in a VMI implementation in the retail clothing industry and discount merchandising. Parks [2] reports on a VMI program instituted by Procter & Gamble and its customers (retailers). The author finds that the retailers’ inventory turns were increased over 100%, service levels and retail sales increased while storage and handling costs were reduced. Burke [3] finds that a VMI agreement at GE Supply reduced receiving and stocking costs due to increased on-time arrival of product. Note that the benefits mentioned thus far apply almost exclusively to the retailers under VMI. Retailers in VMI programs also tend to realize an immediate savings in administrative costs as suppliers assume the burden of managing inventory. However, there can also be benefits for a supplier under VMI. Because VMI requires information sharing, the supplier often gains better visibility of customer demand. There may also be opportunities for reducing supplier costs through coordinating production and shipping activities. Finally, the use of VMI programs can encourage better retailer–supplier relations that can result in longer term contracts, more favorable payment terms, and increased customer loyalty.

However, VMI has not been shown to universally result in supply chain savings. Aichlmayr [4] reports that 3 out of 10 VMI

programs studied resulted in no benefits and another 3 provided modest benefits that were less than expected. Copacino [5] discusses VMI programs that led to the supplier paying higher transportation costs than without VMI because of increased shipping frequencies. Others claim that VMI simply allows retailers to transfer costs and risk to suppliers. Yao *et al.* [6] identify supply chain cost parameters that are more likely to lead to benefits from VMI, but they also find scenarios that can lead to increased costs and simply shifting the inventory downstream in the supply chain. Cachon and Fisher [7] examine a VMI program at Campbell Soup Company through a simulation study. Here, the authors find that information sharing without VMI could have provided the same benefits. Clark and Hammond [8] come to a similar conclusion in a separate empirical study that while VMI programs have produced benefits, simply sharing information between retailers and suppliers would produce similar savings.

VMI requires both extensive information sharing and close relations between suppliers and retailers. Many researchers have examined the benefits of information sharing in the absence of VMI (see *Information Sharing in Supply Chains*). There can also be derived benefits from the collaborative relationships required under VMI between retailers and suppliers. Such concepts are often promoted as part of category management programs or collaborative planning, forecasting, and replenishment (CPFR) initiatives. However, VMI programs go further than just the sharing of information and forming partnerships. VMI agreements also transfer decision making responsibilities from retailers to the supplier and may provide incentives to the supplier to make use of shared information that are not present under these other programs. VMI programs often alter the incentive structure for suppliers by linking their costs to retailer inventory performance. Thus, there is at least the opportunity for additional gains from VMI over information sharing alone.

In the remainder of this article we provide an overview of the current state of research on vendor-managed inventory. We will first

look at several of the purported benefits of VMI including shipment consolidation, and changes to individual agent incentives. Then we will discuss several challenges to VMI implementation including resistances to sharing information and technological difficulties. We provide an example of one type of analytical model for a VMI program and discuss related findings. Finally, we present conclusions and mention several open research questions related to VMI.

BENEFITS AND CHALLENGES OF VMI PROGRAMS

Benefits

Many of the reported benefits of VMI programs can be assigned to improved forecasting and better communication offered through information sharing. Here, we attempt to differentiate between those benefits that are more directly due to increased information sharing and those that are due to the transfer of decision rights in VMI programs. Lee *et al.* [9] and Disney and Towill [10] discuss the benefits of VMI from reducing information distortion and mitigating the bullwhip effect. Because VMI promotes information transparency and effectively reduces the number of agents making inventory-related decisions in the supply chain, we see less exacerbation of the bullwhip effect. The additional benefits from VMI on which we will focus are often due to improvements in transportation and production efficiencies or to changes in the incentives offered to individual supply chain agents under VMI.

VMI programs often offer opportunities for reducing delivery-related costs by consolidating shipments and creating more efficient delivery routes. It is well known that truckload (TL) shipping rates are considerably cheaper than less-than-truckload (LTL) rates. Thus, whenever a supplier can more fully utilize cube space in its vehicles, it can reduce delivery costs. Rather than reacting to delivery requests from retailers, VMI allows the supplier to anticipate upcoming orders and make multiple deliveries on the same route (Fig. 1). This can both improve cube utilization of the vehicles and reduce

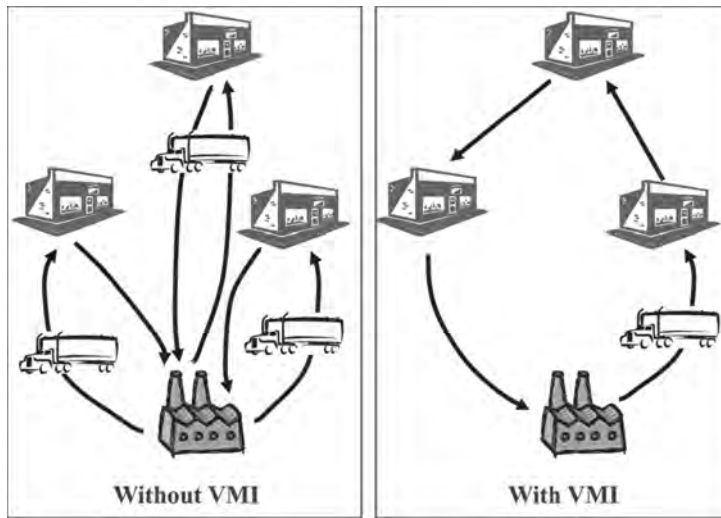


Figure 1. Example of shipment consolidation under VMI.

the number of required vehicles for delivery. It also allows for more frequent deliveries to retailers, thus reducing lead times and required safety stock. From an operational perspective, VMI often forces the supplier to solve a vehicle routing problem (see the section titled “Vehicle Routing and Scheduling” in this encyclopedia) in addition to an inventory-allocation problem. While providing many opportunities for increased operational efficiency, the combined problem is often quite complex.

Several researchers have studied the improvements offered through improved shipping efficiency from VMI. Campbell *et al.* [11], Bertazzi *et al.* [12], Campbell and Savelsbergh [13], and Archetti *et al.* [14] study the inventory-routing problem in relation to VMI. The inventory-routing problem combines inventory decisions with route design; as such, it is an exceptionally difficult problem to solve to optimality. The focus in these papers is on the solution methods that often combine optimization models and heuristic search. Owing to the inherent intractability of the models, these inventory-routing problem formulations assume that demand is deterministic. Berman and Larson [15], Adelman [16], and Kleywegt *et al.* [17] extend the inventory-routing problem to stochastic demand scenarios. In these papers as well, the focus is on finding efficient

solution methods more so than identifying the savings offered by VMI programs. Nevertheless, VMI offers a clear potential for savings in improving the efficiency of the delivery process to multiple retailers using inventory-routing models. Extending these models to solve more practical scenarios is an open area for future research.

Cheung and Lee [18] consider a stochastic demand scenario and they focus on the ability of the supplier to consolidate shipments and rebalance inventory through VMI by postponing the assignment of inventory to specific retailers. Cheung and Lee conclude that both the supplier and retailers are better off under VMI because VMI reduces the amount of variability seen by the supplier due to pooling retailer orders; meanwhile, VMI provides increased reliability for replenishment to retailers. Çetinkaya and Lee [19] and Axsäter [20] also examine shipment consolidation under VMI for stochastic demand scenarios. Through a combination of analytical and numerical results, the authors find that VMI is beneficial in some situations (generally for fast-moving items), but not in others.

Fry *et al.* [21,22] examine VMI as a means of coordinating supplier production and retailer replenishment. Because suppliers (here considered to be manufacturers following a fixed production schedule) control both

the production and replenishment decisions, they can coordinate these decisions to reduce supply chain costs. The authors also provide an explicit model for the originating VMI contract where the supplier is penalized for having too much, or too little, inventory at the retailer. This type of VMI contract is found to be quite common in interviews of VMI-participating companies by Claassen *et al.* [23]. There are many papers that investigate contracts that attempt to realign incentives within the supply chain to overcome double marginalization and other inefficiencies inherent in decentralized systems. Double marginalization refers to the successive markups present in a vertical supply chain under simple wholesale price-only arrangements. Double marginalization leads to retailers typically understocking compared to the supply chain optimal inventory levels, thus leading to suboptimal supply chain performance. VMI programs provide an opportunity for reallocating decision rights and changing incentive structures to overcome such inefficiencies in decentralized supply chains.

Other researchers have examined incentive alignment contracts for VMI programs (see the section titled “Contracts in Supply Chains” in this encyclopedia). Cachon [24] models a multiple retailer VMI program where the penalty cost for not meeting customer demand is split between the supplier and the retailer. In Cachon’s model of VMI, the supplier sets all inventory reorder points. The author shows analytically that VMI will not coordinate the supply chain without fixed transfer payments between retailers and suppliers. Furthermore, a numerical study reveals that VMI will not provide any benefit in this scenario without fixed transfer payments. Nagarajan and Rajagopalan [25], Bichescu and Fry [26] also investigate the splitting of penalty costs, holding costs, or both, in VMI contracts. These researchers show that VMI can be beneficial for both parties in the agreement, but that it is dependent on the exact contract parameters chosen. Bernstein *et al.* [27] show that a simple pricing scheme under a VMI program can guarantee supply chain coordination in many scenarios, but that perfect coordination

may require consignment-type agreements in addition to VMI or pricing decisions that could run afoul of federal anticompetitive laws.

Challenges

There are many challenges to successful implementation of VMI programs. Two of the biggest challenges are (i) communication and technology requirements, and (ii) issues related to trust. VMI programs require regular communication between suppliers and retailers to transmit inventory and demand information. While some of the earliest VMI programs relied on periodic updates sent by fax or other means, most current VMI programs are supported by common information systems that either send the information automatically or allow all parties to view information in real time. Such information systems can be very costly and can often be difficult to implement and troubleshoot. However, the advancement of new technologies may ease the technological burdens of VMI programs. Electronic data interchange (EDI) standardizes many of the information exchanges between suppliers and retailers related to order placement, inventory updates, and so on. EDI standards have been in place for quite some time, but impediments to information sharing have remained.

Many enterprise resource planning (ERP) systems are now configured to support VMI programs. This makes it easier for suppliers and retailers to share information and often allows for real-time visibility of inventory and demand information throughout the supply chain. However, ERP systems still rely on large amounts of user input and errors are not uncommon. The evolving use of radio frequency identification (RFID) technology may reduce many of these user input errors and continue to enhance the ability to share information in the supply chain (see ***Inventory Inaccuracies in Supply Chains: How Can RFID Improve the Performance?***). However, RFID still suffers from technological challenges and high costs that have slowed its widespread adoption. While the aforementioned technologies may make it easier to implement and monitor VMI programs, they still involve considerable cost. These costs are

typically ignored in the academic literature that measure the benefits of VMI but in practice, they are nevertheless a real impediment to VMI programs.

VMI programs require very close relationships between the partners. The data required for operating a VMI program includes detailed demand information and inventory levels—data that are generally considered proprietary and a possible source of competitive advantage. Thus, any successful VMI program requires a large amount of trust between the partners that these data will not be shared with outside parties or otherwise exploited. Kumar [28] highlights the importance of trust in supply chain relationships. Kumar points to the VMI program between Wal-Mart and Procter & Gamble as an example of a successful supply chain relationship built on trust. As Kumar describes, VMI requires Wal-Mart to “share sales and price data and cede control of the order process and inventory management to P&G. P&G had to trust Wal-Mart enough to dedicate a large cross-functional team to the Wal-Mart account... and invest in a customized information link” [28, p. 102].

In passing control of the inventory replenishment process to the suppliers, retailers are placing a large amount of trust in believing that suppliers will not simply pass inventory directly to the retailer—a common complaint in some VMI programs. Chopra [29] reports that Seven-Eleven Japan Co. chooses to forego VMI program with its suppliers so that Seven-Eleven Japan can maintain complete control of its supply chain to support its overall strategy of providing frequent deliveries in small time windows to its stores. Of course, retailers do not simply rely on trust to prevent the adverse actions of suppliers. Many VMI agreements contain specific metrics and contractual terms that prevent the supplier from sending large amounts of inventory to retailers that do not sell. This is also a reason why VMI is often paired with consignment where the supplier retains ownership (and financial responsibility) for the inventory at the retailers. Because the supplier must continue to pay holding costs for items, this provides a disincentive for simply shipping inventory to the retailer that

may not sell. Verification and enforcement of VMI contract parameters can also present additional costs. While these costs are not usually included in the academic models of VMI, it is important to realize that they will reduce any savings garnered from VMI.

AN EXAMPLE OF A VENDOR-MANAGED INVENTORY MODEL

Consider a finite horizon, periodic review setting with a single retailer and a single supplier following a fixed production schedule. Customer demand is stochastic and i.i.d. We denote demand at the retailer by the random variable D and we assume that demand follows the distribution function $\Phi(\cdot)$ and density function $\phi(\cdot)$. Unfulfilled demand at the retailer is assumed to be backlogged. The supplier fills retailer orders using a combination of production and outsourcing so that the retailer always receives his full order amount.

Under a traditional supply chain structure where the retailer places orders and the supplier responds by filling those orders in full, the retailer will place orders using the well-known newsvendor order amounts (see **Supply Chain Coordination**). Let y^* denote this optimal order amount without VMI, where $y^* = \Phi^{-1}\left(\frac{p}{p+h_R}\right)$, p refers to the retailer's backlogging cost for not meeting demands and h_R is the retailer's per unit, per period holding cost.

Under VMI the retailer's inventory replenishment decision now passes to the supplier. The specific VMI program we will consider here is the same as that formulated in Fry *et al.* [21] where the retailer enforces upper and lower limits on the amount of inventory delivered by the supplier in each period. These limits are referred to as z and Z , respectively, for the minimum and maximum amounts of inventory allowed at the retailer after demand has been filled. We assume that there is a penalty amount of $b^+(b^-)$ incurred by the supplier for each unit of inventory at the retailer in violation of $Z(z)$. As previously mentioned, the supplier can replenish the retailer's inventory either from the on-hand

inventory or through outsourcing. Outsourcing results in a cost premium to the supplier of b_0 per unit.

Consider first the case where the supplier has no on-hand inventory ($x_S = 0$) and, thus, employs outsourcing. Let $V_n(x_R)$ be the supplier's expected cost-to-go of the optimal policy for replenishing the retailer from period n until the end of the horizon (period N) when the retailer has x_R units of inventory on-hand.

$$V_n(x_R) = -b_0 x_R + \min_{y \geq x_R} \left\{ b^- E[(z + D - y)^+] + b^+ E[(y - Z - D)^+] + b_0 y + E[V_{n+1}(y - D)] \right\} \quad (1)$$

$$\triangleq -b_0 x_R + \min_{y \geq x_R} J_n(y). \quad (2)$$

Under the assumption that $J_N(\cdot) = V_N(\cdot) = 0$, $J_n(\cdot)$ and $V_n(\cdot)$, $n = 1, \dots, N-1$ are convex and thus, a replenish up-to policy is optimal.

Next, consider the situation where the supplier does have inventory on-hand ($x_S > 0$). We know that as soon as the supplier exhausts her on-hand inventory, she will follow a replenish up-to policy until the end of the horizon, that is the next production opportunity, because of Equations (1) and (2). Let $K_n(y, x_E)$ be the expected cost-to-go from period n until the end of the horizon when the supplier has y units of inventory and let x_E be the echelon inventory, which is defined as the combined inventory at the supplier and the retailer (i.e., $x_E = x_S + x_R$).

$$K_n(y, x_E) = b^- E[(z + D - y)^+] + b^+ E[(y - Z - D)^+] - h_S y + h_S x_E + E[U_{n+1}(y - D, (x_E - D)^+)] \quad (3)$$

$$U_n(x_1, x_2) = \min_{x_1 \leq y \leq x_2} K_n(y, x_2) \quad (4)$$

A modification of the results found in Clark and Scarf [30] prove that these cost functions are also convex; thus, a replenish up-to policy is optimal when the supplier has inventory on hand as well. Define $\bar{y}$ as the optimal replenish up-to value when the

supplier has on-hand inventory and y_n^* as the optimal replenish up-to value when the supplier does not have on-hand inventory in period n . Clearly, it is optimal to replenish the retailer from on-hand inventory before outsourcing whenever possible. It is easily shown that the derivatives of the cost-to-go functions are identical for the nonoutsourcing and the general case. These two properties allow us to state the supplier's optimal replenishment policy under VMI as follows:

- if $x_S + x_R \geq \bar{y}$, replenish the retailer up-to $\bar{y}$;
- if $\bar{y} > x_S + x_R > y_n^*$, replenish the retailer up-to $x_S + x_R$;
- if $x_S + x_R \leq y_n^*$ in period n , replenish the retailer up-to y_n^* .

The optimal production values for the supplier under VMI can also be found. Let q be the combined amount of inventory at the supplier and the retailer at the beginning of the planning horizon (production cycle). We recursively define the supplier's cost $F(q)$ as

$$F(q) = h_S \left[\sum_{n=0}^{N-1} I_n \right] + b_0 \left[\sum_{n=1}^{N-1} O_n \right] + b^+ \left[\sum_{n=0}^{N-1} (I_n - Z)^+ \right] + b^- \left[\sum_{n=0}^{N-1} (z - I_n)^+ \right] \quad (5)$$

where

- I_n = inventory at the supplier, after delivery to the retailer, in period n ,
- D_n = demand experienced at retailer in period n ,
- O_n = amount outsourced in period n .

I_n and O_n are defined as

$$I_0 = \bar{x}_S + (q - \bar{x}_S - \bar{x}_R)^+ - (\bar{y} - \bar{x}_R)^+$$

where

$\bar{x}_S$ = inventory position at supplier just before production,

$\bar{x}_R$ = inventory position at retailer just before production,

$$I_n = [I_{n-1} - (\bar{y} - (y_{n-1} - D_{n-1}))^+]^+ \\ \text{for } n = 1, \dots, N-1,$$

$$O_n = [y_n^* - I_{n-1} - (y_{n-1} - D_{n-1})]^+ \\ \text{for } n = 1, \dots, N-1$$

$$y_0 = \bar{x}_R + (\bar{y} - \bar{x}_R)^+, \text{ and for } n = 1, \dots, N-1,$$

$$y_n = \begin{cases} y_{n-1} - D_{n-1} & : y_{n-1} - D_{n-1} \geq \bar{y} \\ \bar{y} & : y_{n-1} - D_{n-1} < \bar{y} \text{ and } I_n \geq y_{n-1} - D_{n-1} \\ I_n + O_n + y_{n-1} - D_{n-1} & : y_{n-1} - D_{n-1} < \bar{y} \text{ and } I_n < y_{n-1} - D_{n-1} \end{cases}$$

and $(x)^+ = \max(x, 0)$. Fry *et al.* [21] show that $F(q)$ is convex in q ; thus, a produce up-to policy is optimal.

Now consider a traditional, non-VMI setting where the retailer places orders and the supplier responds by filling orders from manufacturing and outsourcing, which we refer to as a retailer-managed inventory (RMI) scenario. We can write the equation for the cost under RMI of having q units in inventory at the supplier and the retailer at the beginning of the production cycle as

$$G(q) = h_S \left[\sum_{n=0}^{N-1} I_n \right] + b_0 \left[\sum_{n=1}^{N-1} O_n \right] \quad (6)$$

where the I_n values are exactly the same as under VMI in Equation (5), but

$$O_n = [y^* - I_{n-1} - (y^* - D_{n-1})]^+ \\ = [D_{n-1} - I_{n-1}]^+ \text{ for } n = 1, \dots, N-1.$$

where

y_n = retailer's inventory position just after being replenished by the supplier in period n ,

Here, $G(q)$ is also convex, so a produce up-to policy remains optimal.

Once the optimal replenishment and production policies are determined for both, the VMI and RMI scenarios, we can compare the performance of the different supply chain settings and evaluate the effectiveness of VMI under different contractual parameters. Figures 2 and 3 are adapted from results in Fry *et al.* [21]. Figure 2 compares many different possible VMI contracts to a traditional RMI scenario. Each point on a line represents a particular VMI contract with a given b^- and b^+ value. All other parameters remain fixed. This figure shows that VMI reduces supply chain costs under a variety of contractual parameters; however, supply chain savings are not guaranteed (see portion of figure where RMI outperforms VMI).

There are many contracts that result in supply chain savings that are not Pareto-improving, which means that either

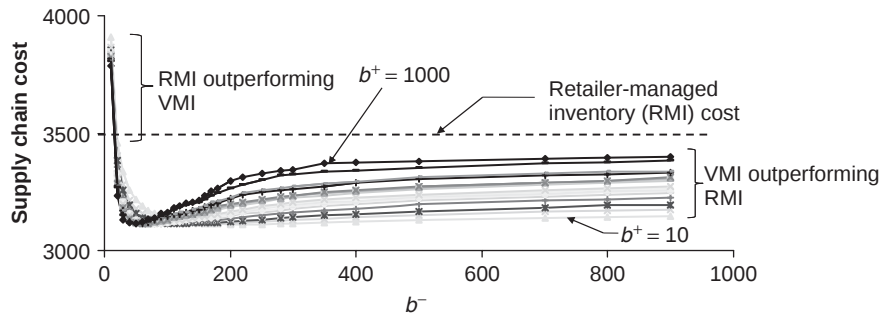


Figure 2. Supply chain costs under VMI compared to RMI.

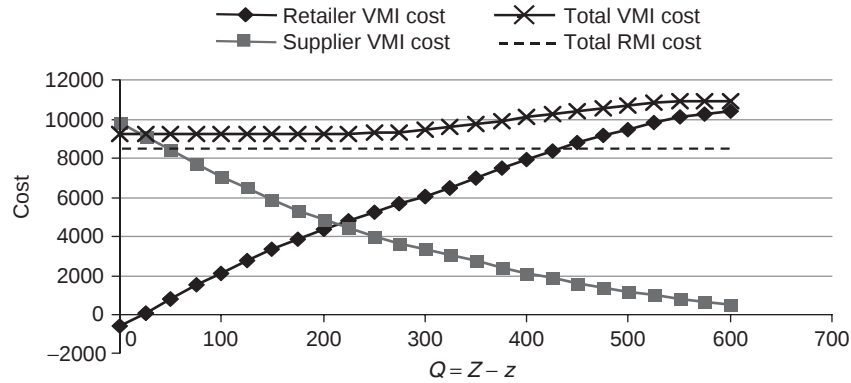


Figure 3. Individual agent costs under VMI.

the supplier or retailer would prefer the non-VMI scenario unless some sort of additional form of fixed transfer payment was instituted. Figure 3 demonstrates that individual retailer and supplier costs can be adjusted within a wide range by changing the VMI contractual parameters. Here, we adjust the range in which the supplier pays no penalties to the retailer for end-of-period inventory levels (i.e., $Q = Z - z$). All VMI contracts shown here result in supply chain savings compared to RMI, but the division of costs between supplier and retailer differ greatly under different contracts. Many of the individual agent costs shown here under VMI are greater than what the agent would pay under a traditional RMI scenario. Similar results requiring fixed transfer payments under VMI in many settings are found in Cachon [24] and Nagarajan and Rajagopalan [25]. Consistent with reports from industry, these researchers also find that when VMI does not perform well it is often the supplier that suffers.

CONCLUSIONS

The existing research demonstrates that VMI programs can be an effective means for improving supply chain efficiency and reducing system cost. VMI provides opportunities for improved supply chain visibility, shipment consolidation to reduce transportation costs, and aligning incentives between decentralized agents. Most of the existing research

papers can be divided between those that focus on the (i) transportation savings offered by VMI; and (ii) those that examine changes to the incentive structure for suppliers and retailers. Papers that examine transportation savings are often formulated in terms of the inventory-routing problem, and the focus has been on finding effective solution methods. Papers examining changes to the incentive structure under VMI generally examine the supply chain and individual agent savings offered by VMI, but typically ignore any savings offered from improved routing and shipment consolidation. Clearly there are still opportunities for more complex models that combine transportation savings with incentive contractual changes from VMI in the presence of stochastic demand.

In practice, the reports on VMI are mixed. Many implementations have led to considerable savings through increased service levels and reductions in holding costs. The existing models demonstrate that some of these failures from VMI may be due to improperly designed contracts that do not guarantee Pareto-improving solutions. Indeed, the results often show that finding Pareto-improving VMI agreements can often be difficult without the use of fixed transfer payments. Furthermore, the academic research tends to ignore many implementation and recurring costs under VMI including technology compatibility considerations as well as verification and enforcement issues. Recent work by Robinson [31] demonstrates that it

may be optimal for retailers to over-report customer demand levels under VMI. Current VMI models rely on perfect information visibility on the part of the supplier and truthful reporting by the retailer. However, such systems may not exist, and certainly are not costless; thus, further research into this area may be needed as well.

REFERENCES

- Buzzell RD, Ortmeyer G. Channel partnerships streamlining distribution. *Sloan Manage Rev* 1995;36(3):85–96.
- Parks L. CRP investment pays off in many ways. *Drug Store News* 1999;21:26.
- Burke M. It's time for vendor managed inventory. *Indus Distrib* 1996;85:90.
- Aichlmayr M. DC mart: who manages inventory in a value chain? *Transp Distrib* 2000;41(10):60–65.
- Copacino WC. How to get with the program. *Traffic Manage* 1993;32:23–24.
- Yao Y, Evers PT, Dresner ME. Supply chain integration in vendor-managed inventory. *Decis Support Syst* 2007;43(2):663–674.
- Cachon GP, Fisher M. Campbell soup's continuous replenishment program: evaluation and enhanced inventory decision rules. *Prod Oper Manage* 1997;6:266–276.
- Clark T, Hammond J. Reengineering channel reordering processes to improve total supply chain performance. *Prod Oper Manage* 1997;6:248–265.
- Lee HL, Padmanabhan V, Whang S. The bullwhip effect in supply chains. *MIT Sloan Manage Rev* 1997;38(3):93–102.
- Disney SM, Towill DR. Vendor-managed inventory and bullwhip reduction in a two-level supply chain. *Int J Oper Prod Manage* 2003;23(6):625–651.
- Campbell *et al.* The Inventory Routing Problem. In: Cranic TG, Laporte G, editors. *Fleet management and logistics*. Boston: Kluwer Academic Publishers; 1998. pp. 95–114.
- Bertazzi L, Paletta G, Speranza MG. Deterministic order-up-to level policies in an inventory routing problem. *Transp Sci* 2002;36(1):119–132.
- Campbell AM, Savelsbergh MWP. A decomposition approach for the inventory-routing problem. *Transp Sci* 2004;38(4):488–502.
- Archetti C, Bertazzi L, Laporte G *et al.* A branch-and-cut algorithm for a vendor-managed inventory-routing problem. *Transp Sci* 2007;41(3):382–391.
- Berman O, Larson RC. Deliveries in an inventory/routing problem using stochastic dynamic programming. *Transp Sci* 2001;35(2):192–213.
- Adelman D. A price-directed approach to stochastic inventory/routing. *Oper Res* 2004;52(4):499–514.
- Kleywegt AJ, Nori VS, Savelsbergh MWP. The stochastic inventory routing problem with direct deliveries. *Transp Sci* 2002;36:94–118.
- Cheung KL, Lee HL. The inventory benefit of shipment coordination and stock rebalancing in a supply chain. *Manage Sci* 2002;48(2):300–306.
- Çetinkaya S, Lee CY. Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Manage Sci* 2000;46:217–232.
- Axsäter S. A note on stock replenishment and shipment scheduling for vendor-managed inventory systems. *Manage Sci* 2001;47(9):1306–1310.
- Fry MJ, Kapuscinski R, Olsen TL. Coordinating production and delivery under a (z, Z) -type vendor managed inventory contract. *Manuf Serv Oper Manage* 2001;3(2):151–173.
- Fry MJ, Kapuscinski R, Olsen TL. Vendor-managed inventory in production constrained settings. In: Boone T, Ganeshan R, editors. *New directions in supply chain management*. New York: American Management Association; 2002. pp. 325–351.
- Claassen MJT, van Weele AJ, van Raaij EM. Performance outcomes and success factors of vendor managed inventory (VMI). *Supply Chain Manage: Int J* 2008;13(6):406–414.
- Cachon GP. Stock Wars: inventory competition in a two-echelon supply chain with multiple retailers. *Oper Res* 2001;49(5):658–674.
- Nagarajan M, Rajagopalan S. Contracting under vendor managed inventory systems using holding cost subsidies. *Prod Oper Manage* 2008;17(2):200–210.
- Bichescu BC, Fry MJ. Vendor-managed inventory and the effect of channel power. *OR Spectr* 2009;31(1):195–228.
- Bernstein F, Federgruen A. Coordinating supply chains with simple pricing schemes: the role of vendor-managed inventories. *Manage Sci* 2006;52(10):1483–1492.

28. Kumar N. The power of trust in manufacturer-retailer relationships. *Harvard Bus Rev* 1996;74:92–106.
29. Chopra S. Seven-Eleven Japan, Co., Supply Chain Management: Strategy, Planning and Operation. Evanston (IL): Kellogg School of Management case study Northwestern University; 2005.
30. Clark AJ, Scarf H. Optimal policies for a multi-echelon inventory problem. *Manage Sci* 1960;6:475–490.
31. Robinson L. The optimality of overreporting demand in vendor-managed inventory systems. Working Paper, Cornell University, Ithaca, New York, 2008.

VERY LARGE-SCALE NEIGHBORHOOD SEARCH

DOUGLAS S. ALTNER
Department of Mathematics,
United States Naval Academy,
Annapolis, Maryland

RAVINDRA K. AHUJA
Innovative Scheduling, Inc., and
Department of Industrial and
Systems Engineering,
University of Florida,
Gainesville, Florida

ÖZLEM ERGUN
H. Milton Stewart School of
Industrial and Systems
Engineering, Georgia Institute
of Technology, Atlanta, Georgia

JAMES B. ORLIN
Operations Research Center,
Massachusetts Institute of
Technology, Cambridge,
Massachusetts

INTRODUCTION

There are numerous optimization problems of interest that are too difficult to solve to optimality in a reasonable amount of time. A practical remedy is to employ heuristics that can quickly compute good quality solutions. Neighborhood search algorithms, also called *local search algorithms*, are a wide class of heuristics that iteratively search for a better quality solution in a “neighborhood” of the current solution.

A critical issue in these methods is defining the neighborhood. As a rule of thumb, larger neighborhoods contain higher quality locally optimal solutions. However, larger neighborhoods typically require more time to search. A large neighborhood will not produce an effective heuristic unless one can search the large neighborhood efficiently. This section concentrates on neighborhood search algorithms where the size of the

neighborhood is “very large” with respect to the size of the input data and where the neighborhood can be implicitly searched in an efficient manner.

Preliminaries

We introduce several basic concepts used throughout this section. A *combinatorial optimization problem (COP)* is an optimization problem of the form: Minimize $\{f(S) : S \in \mathcal{F}\}$ where $\mathcal{F} \subseteq 2^E$ is a finite set of *feasible solutions*. In this context, $E = \{e_1, e_2, \dots, e_n\}$ is the *ground set*, $f : \mathcal{F} \rightarrow \mathbb{R}$ is the *objective function*, and 2^E denotes the set of all subsets of E . For example, in the traveling salesman problem (TSP), E corresponds to the edges in the graph, $\mathcal{F}$ corresponds to the set of feasible TSP tours, and $f(S)$ for $S \in \mathcal{F}$ corresponds to the length of the tour S . COPs can also be maximization problems. However, this section characterizes COPs as minimization problems without loss of generality as a notational convenience.

A *neighborhood function* $\mathcal{N}$ is a point-to-set map $\mathcal{N} : \mathcal{F} \rightarrow 2^{\mathcal{F}}$. Given a feasible solution $S \in \mathcal{F}$, the set $\mathcal{N}(S)$ is called the *neighborhood* of the solution S . The *size* of a neighborhood is the cardinality of the set $\mathcal{N}(S)$. An operation that modifies a solution S_i to obtain a solution $S_{i+1} \in \mathcal{N}(S_i)$ is called a *move*. If $f(S_{i+1}) < f(S_i)$, then the move is *improving*. A solution $S^* \in \mathcal{F}$ is *locally optimal* with respect to a neighborhood function $\mathcal{N}$ if and only if $f(S^*) \leq f(S)$ for all $S \in \mathcal{N}(S^*)$.

A *neighborhood search algorithm*, also called a *local search algorithm*, is an iterative algorithm that begins with a solution S_{i-1} at the *i*th *iteration*, and moves to an improving solution $S_i \in \mathcal{N}(S_{i-1})$ if one exists. Otherwise, the algorithm terminates with S_{i-1} . Algorithm 1 displays pseudocode for a generic neighborhood search algorithm. This pseudocode assumes the initial solution S_0 is input. The subroutine **BestNeighbor**($S_{i-1}, \mathcal{N}, f$) returns the best solution from the neighborhood $\mathcal{N}(S_{i-1})$ according to the objective function f .

Algorithm 1. Generic Neighborhood Search

S_0 // Initial solution
 $i \leftarrow 0$

repeat

$i \leftarrow i + 1$

$S_i \leftarrow \text{BestNeighbor}(S_{i-1}, \mathcal{N}, f)$

until $f(S_i) \geq f(S_{i-1})$

return S_{i-1}

A *very large-scale neighborhood* (VLSN) is a neighborhood that is “very large” in size with respect to the problem input. Typically, these neighborhoods are exponentially large, but this term is often used informally to describe neighborhoods that are too large to explicitly search in practice. Similarly, a *VLSN search algorithm* is a neighborhood search algorithm that searches a VLSN at each iteration. The phrase *VLSN search* was originally coined by Ahuja *et al.* [1].

Although the focus is on VLSN search algorithms for NP-hard COPs, we discuss how the simplex method can be viewed as a neighborhood search algorithm to illustrate how VLSN search relates to regular neighborhood search. During the simplex method, a basic feasible solution is represented by partitioning the decision variables into a set of basic variables B and a set of nonbasic variables N . A move, which is called a *pivot* in simplex terminology, exchanges one decision variable from B with a decision variable from N . The neighborhood of a basic feasible solution is the set of all basic feasible solutions that may be constructed from it with a single move. Analogously, a simplex-based VLSN search method is column generation. In column generation, the set of decision variables is too large to be represented explicitly so the nonbasic variables are implicitly searched to determine the entering basic variable. More detailed descriptions of the simplex method and column generation can be found in Bertsimas and Tsitsiklis [2].

Previous surveys of VLSN search techniques can be found in Ahuja *et al.* [1,3] and in Altner *et al.* [4]. A survey of large

neighborhood search techniques specifically for the TSP can be found in Deineko and Woeginger [5].

Variable-Depth Neighborhood Search Algorithms

Neighborhoods based on exchanging two or three elements simultaneously can often be explicitly searched in a reasonable amount of time. Exchanging a larger number of elements may lead to neighborhoods containing better local optimal solutions, but such neighborhoods typically cannot be searched efficiently. Variable-depth search algorithms are local search algorithms that (in principle) consider neighborhoods where a variable number of elements can be exchanged; to make the search practical, heuristics are used to search the neighborhood. There are variable-depth search algorithms for a large number of applications that obtain nearly optimal solutions efficiently. Such algorithms are the focus of this section.

Let S_i and S_{i+1} be solutions of a COP with ground set E . The move from solution S_i to solution S_{i+1} is a *k-exchange* if $|S_i \setminus S_{i+1}| = |S_{i+1} \setminus S_i| = k$ for a positive integer k . A neighborhood where $|S_i \setminus S_{i+1}| = |S_{i+1} \setminus S_i| \leq k$ for all $S_{i+1} \in \mathcal{N}(S_i)$ is a *k-exchange neighborhood* of S_i . A *k-exchange neighborhood* is said to have *depth k*.

A *variable-depth neighborhood search algorithm* begins with a solution S_{i-1} at the i th iteration, and moves to an improving solution S_i in the $|E|$ -exchange neighborhood of S_{i-1} if an improving move can be identified. Otherwise, the algorithm terminates with S_{i-1} . At each iteration, a variable-depth neighborhood search algorithm typically only partially explores the *k-exchange neighborhood* for each $k \in \{1, 2, \dots, |E|\}$, thus avoiding the $\Omega(|E|^k)$ lower bound of explicitly searching the entire neighborhood.

We now illustrate an example of a variable-depth neighborhood search algorithm on the maximum cut problem (MAX CUT). MAX CUT is defined as follows: Given an undirected graph $G = (V, A)$ where each edge $e \in A$ has weight w_e , partition the vertices V into two subsets X and Y to maximize the sum of the weights of the edges in the cut

$(X, Y) = \{(i, j) : i \in X \text{ and } j \in Y\}$. Let $f(X, Y)$ denote the sum of the weights in (X, Y) .

Pseudocode for a generic variable-depth search algorithm for MAX CUT is presented in Algorithm 2. Algorithm 2 takes an initial solution S_0 as input. The i th iteration of the algorithm starts with a solution S_{i-1} and either moves to a new solution S_i where $f(S_i) > f(S_{i-1})$ or maintains the solution S_{i-1} if no better solution is discovered. Within each iteration of the variable-depth search, there are $|V|$ **rounds**. M_j denotes the set of vertices that may not be moved during round $j + 1$, and (X_j, Y_j) denotes the cut constructed during round j .

The subroutine $\text{flip}(X_{j-1}, Y_{j-1}, v)$ returns a new cut (X_j, Y_j) , where vertex v has been moved from the subset containing it to the other subset. For example, if $v \in X_{j-1}$, then $\text{flip}(X_{j-1}, Y_{j-1}, v)$ returns a cut where $X_j = X_{j-1} \setminus v$, and $Y_j = Y_{j-1} \cup \{v\}$.

Algorithm 2. Generic Variable-Depth Search for MAX CUT

```

 $S_0$  // Initial solution
 $i \leftarrow 0$ 

repeat
     $i \leftarrow i + 1$ 
     $M_0 \leftarrow \emptyset$ 
     $(X_0, Y_0) \leftarrow S_{i-1}$ 

    for  $j = 1$  to  $|V|$  do
         $v_j^* \leftarrow \text{argmax}_{v \in V \setminus M_{j-1}} \{f(\text{flip}(X_{j-1}, Y_{j-1}, v))\}$ 

         $M_j \leftarrow M_{j-1} \cup \{v_j^*\}$ 

         $(X_j, Y_j) \leftarrow \text{flip}(X_{j-1}, Y_{j-1}, v_j^*)$ 
    end for

     $S_i \leftarrow \{(X_k, Y_k) : k = \text{argmax}_{j \in \{1, \dots, |V|\}} f(X_j, Y_j)\}$ 

until  $f(S_i) \leq f(S_{i-1})$ 
return  $S_{i-1}$ 
    
```

A special type of variable-depth neighborhood search algorithm is called an *ejection chain*. Loosely defined, an *ejection chain* is a variable-depth neighborhood search algorithm that iteratively generates a structured alternating sequence of additions and deletions $S_1, S_2, \dots, S_k$ where

1. $|S_1| = |S_j|$ for all odd values of j , and
2. $|S_2| = |S_j|$ for all even values of j .

Ejections chains were first introduced by Glover [6] for solving the TSP. Glover originally developed ejection chains by extending and generalizing the ideas of Lin and Kernighan [7]. Many variable-depth neighborhood search algorithms, particularly those for vehicle routing problems (VRPs), rely on ejection chains at each iteration in order to search the neighborhood.

Problem-specific variable-depth and ejection-chain-based algorithms have been successfully used for quickly computing good quality solutions to a wide variety of difficult COPs. Gamboa *et al.* [8], Glover [6], Johnson and McGeoch [9], Lin and Kernighan [7], Mak and Morton [10], Pesch and Glover [11], Rego [12], and Zachariasen and Dam [13] consider such algorithms for the TSP. The Lin–Kernighan search was the first variable-depth neighborhood search algorithm for the TSP and is still considered the best general framework for variable-depth neighborhood search algorithms for the TSP. At the time of this publication, Helsgaun’s [14] implementation of the Lin–Kernighan search for the TSP is often considered the fastest such neighborhood search algorithm and has successfully obtained optimal solutions for all instances in TSPLIB for which optimal solutions are known and verified.

Variable-depth neighborhood search algorithms are also commonly used for VRPs. Rego [12] describes a subpath ejection chain method for a VRP with vehicle capacities and route length restrictions. Xu and Kelly [15] extended the work of Glover [6] to solve a VRP with vehicle capacities using an ejection chain method. Rego and Roucairol [16] study a parallel tabu search algorithm using ejection chains for a standard VRP. Sontrop *et al.* [17] present an ejection chain method for a VRP with time window constraints.

There are numerous other difficult COPs that have been approached with variable-depth neighborhood search algorithms. Dorndorf and Pesch [18] apply ejection chain techniques to a series of benchmark instances in the clustering literature and obtain excellent results. Yagiura *et al.*

[19,20] consider variable-depth heuristics for the generalized assignment problem, which is the problem of finding a minimum-cost assignment of jobs to agents subject to one or more additional linear constraints. In the more recent study, Yagiura *et al.* [19] obtained solutions that deviated from the best previously reported solutions for benchmark instances from the OR-Library by 0.16% at most.

Variable-depth neighborhood search algorithms have also been applied to a bottleneck assignment problem [21], nurse scheduling [22], parallel machine scheduling [23], uniform graph partitioning [24,25], and very-large-scale integration (VLSI) circuit design [26]. Kernighan and Lin's paper on uniform graph partitioning [25] is the first published work on variable-depth neighborhood search.

Cyclic Exchange Neighborhood Search Algorithms

The cyclic exchange neighborhood was developed for solving minimum-cost partitioning problems. Let $S = \{S_1, S_2, \dots, S_K\}$ be a partition of a ground set E . Roughly speaking, a move in a cyclic exchange neighborhood permits one element to be transferred from subset S_i into subset S_j , which in turn causes another element to be transferred from subset S_j into subset S_k . This chain of transfers continues until an element is transferred into S_i and thus completes a cycle. In this section, we discuss cyclic exchange neighborhoods and reference applications to a wide range of settings.

More formally, let E be the ground set for a COP. The collection of subsets $\{S_1, S_2, \dots, S_K\}$ is a K -partition of E if no subset is empty, the sets are pairwise disjoint, and the union of the sets is E . A *partitioning problem* is a COP where the objective is to find a partition $T = \{T_1, T_2, \dots, T_K\}$ of E that minimizes $\sum_{j=1}^K c(T_j)$, where $c(T_j)$ is the cost of subset $T_j \subseteq E$.

Let $S = \{S_1, S_2, \dots, S_K\}$ and $T = \{T_1, T_2, \dots, T_K\}$ be two K -partitions of E . We say that T is a *1-neighbor* of S if $|T_j \setminus S_j| \leq 1$ and $|S_j \setminus T_j| \leq 1$ for all $j \in \{1, 2, \dots, K\}$. If T is a 1-neighbor of S , then T can be obtained from S via a *1-transfer*. Being a 1-neighbor is

symmetric: if T is a 1-neighbor of S , then S is a 1-neighbor of T .

Suppose $S = \{S_1, S_2, \dots, S_K\}$ is a K -partition and $T = \{T_1, T_2, \dots, T_K\}$ is obtained from S via a 1-transfer. From S and T , we can construct the (S, T) -transfer graph as follows: For each element in the ground set E , there is a corresponding vertex in the (S, T) -transfer graph. For each $k \in \{1, 2, \dots, K\}$, if T_k is obtained from S_k by inserting element i and deleting element j [that is, $T_k = (S_k \setminus i) \cup \{j\}$], then (i, j) is an edge of the (S, T) -transfer graph. Note the (S, T) -transfer graph is a disjoint union of cycles and paths because each vertex of the graph has at most one incoming edge and at most one outgoing edge.

If the (S, T) -transfer graph consists of a single cycle, we say T can be obtained from S via a *cyclic exchange*. Specifically, if the cycle is $i_1, \dots, i_k, i_1$, then T is obtained from S as follows: for each $j = 1$ to k , one inserts i_j into the subset of S containing i_{j+1} and then deletes i_{j+1} from that subset. (Here, i_{k+1} is equivalent to i_1 .) Similarly, if the (S, T) -transfer graph consists of a single path, we say T can be obtained from S via a *path exchange*. The *cyclic exchange neighborhood* (or, *path exchange neighborhood*) of a K -partition S consists of all K -partitions that may be obtained from S by a single cyclic (or, path) exchange.

Figure 1 illustrates a cyclic exchange. Here, vertex e_1 will be transferred from subset T_1 to subset T_2 . Vertex e_4 will be transferred from subset T_2 to subset T_4 . Vertex e_{10} will be transferred from subset T_4 to T_5 . Finally, the cyclic exchange will be completed by transferring vertex e_{13} from subset T_5 to subset T_1 .

To find an improving cyclic exchange starting with a given K -partition $S = \{S_1, S_2, \dots, S_K\}$ we create a new graph called the *improvement graph*. For each element in the ground set E , there is a corresponding vertex in the improvement graph. For every pair i, j of elements of E that are in different subsets of the partition S , there is an edge (i, j) in the improvement graph. Suppose S_k is the part of the K -partition S that contains element j . Then the edge (i, j) is interpreted as a transfer of element i into S_k and a transfer of a different element j out of S_k .

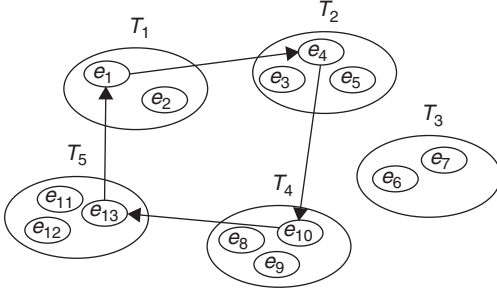


Figure 1. Illustration of a generic cyclic exchange.

Let $S'_k = (S_k \setminus \{j\}) \cup \{i\}$ be the updated subset if the cyclic exchange is executed. The cost of the edge (i, j) in the improvement graph is $c_{ij} = c(S'_k) - c(S_k)$, which is the net change in cost from transferring element i into S_k and transferring element j out.

A cycle $C = i_1, \dots, i_k, i_1$ in the improvement graph is *subset-disjoint* with respect to a K -partition S if no two elements of C belong to the same subset in S . In order for the total weight of a cycle in the improvement graph to accurately reflect the change in the objective function from a cyclic exchange, the cycle must be subset-disjoint.

Finding a minimum cost subset-disjoint cycle is NP-complete [27] but it is one of those NP-complete problems that is relatively easy to solve in practice. Thus, heuristics are used to search the improvement graph in order to construct improving cyclic exchanges. Effective heuristics for constructing negative cost subset-disjoint cycles are presented in Refs 28 and 29.

The improvement graphs and the transfer graphs are closely related. Suppose C is a subset-disjoint cycle in the improvement graph for partition S , and T is the partition obtained by performing the cyclic exchange induced by C . Then the (S, T) -transfer graph consists of the cycle C .

In addition, cyclic exchanges and subset-disjoint cycles in the improvement graph are closely related. Suppose S is a K -partition, and let G be the resulting improvement graph. There is a one-to-one correspondence between cyclic exchanges from S and subset-disjoint directed cycles in G . Moreover, the cost of a subset-disjoint cycle C is the net

change in objective value by performing the induced cyclic transfer.

There are several tricks to improving the running time of a cyclic exchange neighborhood search routine. For example, in practice, it is often much faster to update an improvement graph using the improvement graph from the previous iteration rather than constructing a new improvement graph from scratch. In addition, a path exchange may be transformed into a cyclic exchange by adding dummy vertices. This allows one to use a single search algorithm to identify both improving cyclic exchanges and improving path exchanges.

The size of the cyclic exchange neighborhood is a function of the number of partitions and the number of elements in the ground set. For K -partitions of n elements, the cyclic exchange neighborhood is of size $O(n^K)$. When K is permitted to grow with n , the size of the neighborhood is typically exponential.

Cyclic exchange neighborhoods have been successfully applied to a diverse selection of COPs. Thompson and Orlin [27] originally introduced the cyclic exchange neighborhood and the idea of VLSN search in general. Although, Thompson and Orlin limit their study to the VRP, their methods naturally generalize to other partitioning COPs. Thompson and Psaraftis [28] was the first published discussion of a cyclic exchange neighborhood search algorithm. Although they focused on the VRP, they demonstrated the potential usefulness of cyclic exchange neighborhoods for solving a wide range of partitioning problems. Agarwal *et al.* [30] extended the approach of Thompson and Psaraftis by allowing a set of consecutive cities in a vehicle tour to be transferred from one subset of a partition to another. The algorithm of Agarwal *et al.* obtained solutions within 1% of the best previously reported objective value on a series of benchmark instances that are now contained in the OR-Library. Sindhuchao *et al.* [31] applied similar ideas to an inventory routing problem.

Cyclic exchange neighborhoods have also been used in a variety of scheduling problems. Frangioni *et al.* [32] apply these ideas to

a parallel machine-scheduling problem with the objective of minimizing the makespan and showing that their algorithm outperforms other benchmark heuristics on several classes of randomly generated instances using standard instance generators. Abdullah *et al.* [33,34] designed a cyclic exchange neighborhood search algorithm for university examination timetabling and subsequently built upon this algorithm by adding tabu lists. The algorithm described in the latter paper obtains the best known results on a set of benchmark instances.

Cyclic exchange neighborhoods have also been used in a wide variety of assignment problems. Ahuja *et al.* [35] designed a VLSN search method to compute a minimum-cost fleet assignment of planes to flight legs while satisfying a variety of constraints. Ahuja *et al.* [36] generalize this approach to airline fleet assignment problems that incorporate time window constraints. Ahuja *et al.* [37] adapt a cyclic exchange neighborhood for the quadratic assignment problem (QAP). This is particularly interesting since the QAP is not a partitioning problem, which means that a cyclic exchange neighborhood is not directly applicable. Nevertheless, Ahuja *et al.* show their approach consistently produces better solutions in less time than leading two-exchange neighborhood search algorithms.

In addition to the above, Yagiura *et al.* [38] introduced a cyclic exchange neighborhood for a *multiresource generalized assignment problem* (MRGAP). Unlike most cyclic exchange neighborhood search algorithms, Yagiura *et al.* relaxed the condition that the cycles must be subset-disjoint, citing that the cost of any cycle in their improvement graph serves as a reasonably good approximation for the cost of the actual exchange. Mitrović-Minić and Punnen [39] expand VLSN research on MRGAP by combining concepts from cyclic exchange neighborhood search and the variable neighborhood search technique of Mladenović and Hansen [40]. Mitrović-Minić and Punnen also discuss how their very large-scale variable neighborhood search technique can be extended to other difficult COPs.

Other assignment problems that have been effectively solved using cyclic exchange

neighborhood search algorithms include a problem of assigning weapons to targets to minimize the survivability of all possible targets [41] and a distribution network design problem involving location, fleet assignment, and routing decisions that arises in the trucking industry [42].

Cyclic exchange neighborhoods have been used in numerous other settings. Ahuja *et al.* [29,43] introduce cyclic exchange neighborhood search algorithms for the capacitated minimum spanning tree problem (CMST). CMST is a minimum spanning tree problem subject to the constraint that if the root and its incident edges were deleted from the spanning tree, then each component has at most K vertices, where K is part of the problem input. Ahuja *et al.* [44] designed a cyclic exchange neighborhood for a multi-period production planning problem that involves constraints on production capabilities, inventory space and the perishability of products. Ahuja *et al.* [45] developed a cyclic exchange algorithm for a capacitated facility location problem with single-source constraints. Cunha and Ahuja [46] designed several cyclic exchange neighborhood search algorithms for the K -constrained multiple knapsack problem (K-MKP). K-MKP is a generalization of the knapsack problem, where each item has K types of weights and each knapsack has K types of capacity. Kulkarni and Fathi [47] develop a cyclic exchange neighborhood search algorithm for the q -Mode Problem, which is a problem that arises in cluster analysis. Magazine *et al.* [48] produced a cyclic exchange neighborhood for a problem that arises in the design of printed circuit boards. Nemani *et al.* [49] employ a cyclic exchange neighborhood search algorithm for an intermodal load planning problem, where the objective is to load containers from cargo ships and from trucks onto intermodal trains in a fashion that minimizes the transfer costs and rail car usage while ensuring all containers are arranged to limit air resistance. Finally, Scaparra *et al.* [50] created a cyclic exchange algorithm for the Capacitated Vertex p -Center Problem. The Capacitated Vertex p -Center Problem consists of locating p facilities and assigning clients to them to

minimize the maximum distance between a client and its assigned facility while not exceeding any facility's capacity for servicing clients.

The idea of finding improved solutions by determining negative cost cycles in improvement graphs has been used in several other contexts. For example, Talluri [51] identified cost-saving exchanges of equipment type between flight legs for a daily airline fleet assignment problem by finding negative cost cycles in a related network. In addition, Schneur and Orlin [52] solved a linear multicommodity flow problem by iteratively detecting and sending flow around negative cost cycles. Their technique readily extends to a cycle-based improvement heuristic for the integer multicommodity flow problem.

Other Very Large-Scale Neighborhood Search Algorithms

In designing VLSN search algorithms, researchers have drawn inspiration and techniques from more than 50 years of algorithmic development in mathematical programming and computer science. In this section, we describe a diverse collection of approaches for developing VLSN search algorithms.

One approach for developing VLSNs is to base them on combining several independent moves. compounded independent move (CIM) neighborhoods are neighborhoods based on simultaneously executing two or more moves whose effects on the objective value are independent of one another; that is, the change in the objective value due to the compound move is equivalent to the sum of the changes due to the individual moves. CIM neighborhoods were designed by Congram *et al.* [53] for a machine-scheduling problem and by Potts and van de Velde [54] for the TSP. Their techniques searched a CIM neighborhood with dynamic programming. Potts and van de Velde referred to the general idea of searching a CIM neighborhood with dynamic programming as *dynasearch*.

There are several other examples of CIM neighborhoods. Ergun *et al.* [55] and Gendreau *et al.* [56] searched CIM neighborhoods to solve VRPs. Taillard [57] used CIM neighborhoods for solving clustering problems.

Dror and Levy [58] used CIM neighborhoods for solving the inventory routing problem (IRP). The IRP combines inventory management and vehicle routing into one resource allocation problem. Brueggemann *et al.* [59] implemented a CIM neighborhood for a minimum makespan parallel machine scheduling problem where the neighborhood is searched by solving a bottleneck assignment problem. Bottleneck assignment problems are a variant of assignment problems where the objective is to choose an assignment $\mathcal{M}$ to minimize $\max\{c_e : e \in \mathcal{M}\}$ where c_e is the cost of edge e . Öncan *et al.* [60] discussed the more general problem of designing VLSN search algorithms for partitioning problems in general using a CIM neighborhood that is searched by solving a matching problem.

There are also other examples of dynasearch being used in different contexts. Ergun and Orlin [61] use dynasearch to search a VLSN for the TSP. Ergun and Orlin [62] and Grosso *et al.* [63] use dynasearch for solving single-machine total-weighted tardiness scheduling problems.

VLSNs can also be searched by solving a mathematical program. Suppose x is a solution to a mathematical program. For a given subset I of indices, we say that another feasible solution y is I -adjacent to x if $y_i = x_i$ for $i \in I$, and we say y is in the I -neighborhood of x . I -neighborhoods are conveniently searched using mathematical programming software. Davenport *et al.* [64] and Pesant and Gendreau [65] design large neighborhoods that are searched using constraint programming. In addition, Danna *et al.* [66] have proposed a method called *relaxation-induced neighborhood search (RINS)*, which uses a linear programming solution to construct a neighborhood and solves an integer program to search the neighborhood. The RINS heuristic of Danna *et al.* has been implemented in the CPLEX 9.0 MIP solver.

VLSNs can also be based on polynomially solvable special cases of NP-hard COPs. Researchers have developed a substantial literature on such cases, which often lend themselves to VLSNs that can be searched in polynomial time. For example, Carlier and Villon [67] presented a polynomial-time algorithm for finding an optimal *pyramidal tour*

for the TSP. A pyramidal tour is a sequence $i_1, \dots, i_n$ such that for some value k , the indices in the subsequence $i_1, \dots, i_k$ are in increasing order, and the indices in the subsequence $i_k, \dots, i_n$ are in decreasing order. For a given tour T , its pyramidal neighborhood is the set of pyramidal tours that would result if one first relabeled the vertices (or cities) so that $T = 1, \dots, n$. Therefore, any algorithm for finding a minimum-cost pyramidal tour can be used for finding a minimum-cost pyramidal neighbor.

As another example of this practice, for a given fixed value K and an initial ordering of cities, there is a polynomial-time dynamic program for finding an optimal tour subject to the constraint that city i precedes city j if and only if $j \geq i + K$. Simonetti and Balas [68] use this algorithm as the basis for developing a VLSN search algorithm.

VLSNs can also be constructed by exploiting the physical geography of the underlying COP. For example, Ahuja *et al.* [69] did this for a railroad blocking problem. Blocking problems, which also arise in parcel delivery, trucking, and air transportation, model how to consolidate a large number of shipments into *blocks* to reduce the total handling costs. The VLSN search algorithm devised by Ahuja *et al.* iteratively improves a blocking plan by defining a neighborhood within a single railroad yard to consider reclassifying all freight currently in the yard into different blocks.

Lastly, a unique result is that certain VLSNs can be generated using context-free grammars (CFGs). Bompadre and Orlin [70] showed how to use CFGs to generate exponentially large neighborhoods for sequencing problems. Moreover, for some sequencing problems such as the TSP and the linear ordering problem, they developed an algorithm that takes the grammar and the problem data as input and then finds the minimum-cost neighbor in polynomial time. In addition to presenting an innovative use of CFGs, this paper demonstrates how a single algorithm can be used to search a wide range of different exponentially sized neighborhoods. Mouthuy *et al.* [71] briefly overview how they used this CFG technique to generate neighborhoods to implement VLSNs for the TSP in Comet [72], an

object-oriented programming language for constraint-based local search.

REFERENCES

1. Ahuja RK, Ergun O, Orlin JB, *et al.* A survey of very large-scale neighborhood search techniques. *Discrete Appl Math* 2002;123:75–102.
2. Bertsimas D, Tsitsiklis JN. *Introduction to linear optimization*. Belmont, CA: Athena Scientific; 1997.
3. Ahuja RK, Ergun O, Orlin JB, *et al.* Very large-scale neighborhood search. In: Gonzalez TF, editor. *Approximation algorithms and metaheuristics*. Boca Raton, FL: CRC Press; 2005. p. 20–21 to 20–15.
4. Altner DS, Ahuja RK, Ergun O, *et al.* Very large-scale neighborhood search. In: Rosen KH, editor. *Handbook of discrete and combinatorial mathematics*. 2nd ed. Boca Raton, FL: CRC Press; 2009.
5. Deĭneko VG, Woeginger GJ. A study of exponential neighborhoods for the traveling salesman problem and for the quadratic assignment problem. *Math Prog* 2000; 78:519–542.
6. Glover F. Ejection chains, reference structures, and alternating path algorithms for the traveling salesman problem. *Discrete Appl Math* 1996;65:223–253.
7. Lin S, Kernighan BW. An effective heuristic algorithm for the traveling salesman problem. *Ophthalmic Res* 1973;21:498–516.
8. Gamboa D, Rego C, Glover F. Implementation analysis of efficient heuristic algorithms for the traveling salesman problem. *Comput Oper Res* 2006;33:1154–1172.
9. Johnson DS, McGeoch LA. The travelling salesman problem: a case study in local optimization. In: Aarts EHL, Lenstra JK, editors. *Local search in combinatorial optimization*. New York: John Wiley & Sons; 1997. pp. 311–336.
10. Mak K, Morton A. A modified Lin-Kernighan traveling salesman heuristic. *ORSA J Comput* 1992;13:127–132.
11. Pesch E, Glover F. TSP ejection chains. *Discrete Appl Math* 1997;76:165–181.
12. Rego C. A subpath ejection method for the vehicle routing problem. *Manag Sci* 1998;44:1447–1459.
13. Zachariasen M, Dam M. Tabu search on the geometric traveling salesman problem. In: Osman IH, Kelly JP, editors. *Metaheuristics*:

- theory and applications. Norwell, MA: Kluwer Academic Publishers; 1995. pp. 571–587.
14. Helsgaun K. An effective implementation of the Lin-Kernighan traveling salesman heuristic. *Eur J Oper Res* 2000;126:106–130.
15. Xu J, Kelly JP. A network flow-based tabu search heuristic for the vehicle routing problem. *Trans Sci* 1996;30:379–393.
16. Rego C, Roucairol C. A parallel tabu search algorithm using ejection chains for the vehicle routing problem. In: Osman IH, Kelly JP, editors. *Metaheuristics: theory and applications*. Norwell, MA: Kluwer Academic Publishers; 1996. pp. 661–675.
17. Sontrop HMJ, van der Horn SP, Teeuwen G, *et al.* Fast ejection chain algorithms for vehicle routing with time windows. In: Blesa M, Christian B, Andrea R, *et al.*, editors. *Hybrid metaheuristics, Lecture notes in computer science*. Berlin: Springer; 2005. pp. 78–89.
18. Dorndorf U, Pesch E. Fast clustering algorithms. *ORSA J Comput* 1994;6:141–153.
19. Yagiura M, Ibaraki T, Glover F. An ejection chain approach for the generalized assignment problem. *INFORMS J Comput* 2004;16:133–151.
20. Yagiura M, Yamaguchi T, Ibaraki T. A variable-depth search algorithm for the generalized assignment problem. In: Voss S, Martello S, Osman I, *et al.*, editors. *Metaheuristics: advances and trends in local search paradigms for optimization*. Norwell, MA: Kluwer Academic Publishers; 1999. pp. 459–471.
21. Aggarwal V, Tikekar VG, Hsu LF. Bottleneck assignment problem under categorization. *Comput Oper Res* 1986;13:11–26.
22. Dowsland KA. Nurse scheduling with tabu search and strategic oscillation. *Eur J Oper Res* 1998;106:393–407.
23. Agarwal R, Ergun O, Orlin JB, *et al.* Solving parallel machine scheduling problems with very large-scale neighborhood search. *J Schedul* 2009, In press.
24. Fiduccia CM, Mattheyses RM. A linear time heuristic for improving network partitions. *Proceedings of ACM IEEE Conference on Design Automation*. IEEE Computer Society; 1982. pp. 175–181.
25. Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. *Bell Syst Tech J* 1970;49:291–307.
26. Dunlop AE, Kernighan BW. A procedure for placement of standard cell VLSI circuits. *IEEE Trans Comput-Aided Des* 1985; 4:92–98.
27. Thompson PM, Orlin JB. The theory of cyclic transfers. Technical Repor OR 200-89. Cambridge (MA): ORC MIT; 1989.
28. Thompson PM, Psaraftis HN. Cyclic transfer algorithms for multivehicle routing and scheduling problems. *Oper Res* 1993;41: 935–946.
29. Ahuja RK, Orlin JB, Sharma D. Multi-exchange neighborhood structures for the capacitated minimum spanning tree problem. *Math Prog* 2001;91:71–97.
30. Agarwal R, Ahuja RK, Laporte G, *et al.* A composite very large-scale neighborhood search algorithm for the vehicle routing problem. In: Leung JY-T, editor. *Handbook of scheduling: algorithms, models and performance analysis*. Boca Raton, FL: CRC Press; 2003. pp. 49–01 to 49–23.
31. Sindhuchao S, Romeijn HE, Akcali E, *et al.* An integrated inventory-routing system for multi-item joint replenishment with limited vehicle capacity. *J Glob Opt* 2005;32:93–118.
32. Frangioni A, Necciari E, Scutellá MG. Multi-exchange algorithms for minimum makespan machine scheduling problems. *J Comb Optim* 2004;8:195–220.
33. Abdullah S, Ahmadi S, Burke EK, *et al.* Investigating Ahuja-Orlin's large neighborhood search approach for examination timetabling. *OR Spectr* 2006;29:351–372.
34. Abdullah S, Ahmadi S, Burke EK, *et al.* A tabu-based large neighborhood search methodology for the capacitated examination timetabling problem. *J Oper Res Soc* 2007;58:1494–1502.
35. Ahuja RK, Goodstein J, Mukherjee A, *et al.* A very large-scale neighborhood search algorithm for the combined through-fleet assignment model. *INFORMS J Comput* 2007;19: 416–428.
36. Ahuja RK, Liu J, Orlin JB, *et al.* A neighborhood search algorithm for the combined through fleet assignment model with time windows. *Networks* 2004;44:160–171.
37. Ahuja RK, Jha KC, Orlin JB, *et al.* A very large-scale neighborhood search algorithm for the quadratic assignment problem. *INFORMS J Comput* 2007;19:646–657.
38. Yagiura M, Iwasaki S, Ibaraki T, *et al.* A very large-scale neighborhood search algorithm for the multi-resource generalized assignment problem. *Dis. Opt.* 2004;1:87–98.

39. Mitrović-Minić S, Punnen AP. Local search intensified: very large-scale variable neighborhood search for the multi-resource generalized assignment problem. *Discrete Optim* 2009;6:370–377.
40. Mladenović N, Hansen P. Variable neighborhood search. *Comput Oper Res* 1997;24:1097–1100.
41. Ahuja RK, Kumar A, Jha KC, *et al.* Exact and heuristic algorithms for the weapon-target assignment problem. *Oper Res* 2008;55:1136–1146.
42. Ambrosino D, Sciomachen A, Scutellà MG. A heuristic based on multi-exchange techniques for a regional fleet assignment location-routing problem. *Comput Oper Res* 2009;36:442–460.
43. Ahuja RK, Orlin JB, Sharma D. A composite very large-scale neighborhood structure for the capacitated minimum spanning tree problem. *Oper Res Lett* 2003;31:185–194.
44. Ahuja RK, Huang W, Romeijn HE, *et al.* A heuristic approach to the multi-period single-sourcing problem with production and inventory capacities and perishability constraints. *INFORMS J Comp* 2007;19:14–26.
45. Ahuja RK, Orlin JB, Pallottino S, *et al.* A multi-exchange heuristic for the single source capacitated facility location. *Manag Sci* 2004;50:749–760.
46. Cunha CB, Ahuja RK. Very large-scale neighborhood search for the K-constrained multiple knapsack problem. *J Heuristics* 2005;11:465–481.
47. Kulkarni G, Fathi Y. A very large scale neighborhood search algorithm for the q-mode problem. *IIE Trans* 2007;39:971–984.
48. Magazine MJ, Polak GG, Sharma D, A multi-exchange neighborhood search heuristic for an integrated clustering and machine setup model for PCB manufacturing. *J Elec Man* 2002;11:107–119.
49. Nemani AK, Jha KC, Ahuja RK. The load planning problem at an intermodal railway terminal. Submitted. Gainesville (FL): University of Florida; 2010.
50. Scaparra MP, Pallottino S, Scutellà MG. Large-scale neighborhood heuristics for the capacitated vertex p -center problem. *Networks* 2004;43:241–255.
51. Talluri KT. Swapping applications in a daily airline fleet assignment. *Trans Sci* 1996;30:237–248.
52. Schneur RR, Orlin JB. A scaling algorithm for multicommodity flow problems. *Oper Res* 1998;46:231–246.
53. Congram RK, Potts CN, van de Velde SL. An iterated dynasearch algorithm for the single machine total weighted tardiness scheduling problem. *INFORMS J Comput* 2002;14:52–67.
54. Potts CN, van de Velde SL. Dynasearch – Iterative Local Improvement by Dynamic Programming: Part I, The Traveling Salesman Problem. Technical Report. University of Twente; 1995.
55. Ergun O, Orlin JB, Steele-Feldman A. Creating very large-scale neighborhoods out of smaller ones by compounding moves. *J Heuristics* 2006;12:115–140.
56. Gendreau M, Guertin F, Potvin JY, *et al.* Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. *Trans Res Part C* 2006;14:157–174.
57. Taillard ED. Heuristic methods for large centroid clustering problems. *J Heuristics* 2003;9:51–73.
58. Dror M, Levy L. A vehicle routing improvement algorithm comparison of a greedy and a matching implementation for inventory routing. *Comput Oper Res* 1986;13:33–45.
59. Brueggemann T, Hurink JL, Vredevelt T, *et al.* Performance of a very large-scale neighborhood for minimizing makespan on parallel machines. *Elec Not Dis Math* 2006;25:29–33.
60. Öncan T, Kabadi SN, Nair KPK, *et al.* VLSN search algorithms for partitioning problems using matching neighbourhoods. *J Op Res Soc* 2008;59:388–398.
61. Ergun O, Orlin JB. A dynamic programming methodology in very large scale neighborhood search applied to the traveling salesman problem. *Discrete Optim* 2006;3:78–85.
62. Ergun O, Orlin JB. Fast neighborhood search for the single machine total weighted tardiness problem. *Oper Res Lett* 2006;34:41–45.
63. Grosso A, Della Croce F, Tadei R. An enhanced dynasearch neighborhood for the single-machine total weighted tardiness scheduling problem. *Oper Res Lett* 2004;32:68–72.
64. Davenport A, Kalagnanam J, Reddy C, *et al.* An application of constraint programming to generating detailed operations schedules for steel manufacturing. 2007. *Proc. Con. Prog.* 64–75.
65. Pesant G, Gendreau M. A constraint programming framework for local search methods. *J Heuristics* 1999;5:255–279.

66. Danna E, Rothberg E, Le Pape C. Exploring relaxation induced neighborhoods to improve MIP solutions. *Math Prog* 2005;102:71–90.
67. Carlier J, Villon P. A new heuristic for the traveling salesman problem. *RAIRO Oper Res* 1990;24:245–253.
68. Simonetti N, Balas E. Implementation of a linear time algorithm for certain generalized traveling salesman problems. *Proceedings of IPCO*; 1996. p. 316–329.
69. Ahuja RK, Jha KC, Liu J. Solving real-life railroad blocking problems. *Interfaces* 2007;37:404–419.
70. Bompadre A, Orlin JB. Using grammars to generate very large-scale neighborhoods for the traveling salesman problem and other sequencing problems. *Proceedings of the International Conference IPCO*; Berlin, Germany; 2005. pp. 437–451.
71. Mouthuy S, Deville Y, Van Hentenryck P. Toward a generic Comet implementation of very large-scale neighborhood search. *22nd National Conference of the Belgian Operations Research Society*; Brussels, Belgium; 2008. pp. 16–18.
72. Van Hentenryck P, Michel L. *Constraint-based local search*. Cambridge, MA: MIT Press; 2005.

WARDROP EQUILIBRIA

JOSÉ R. CORREA

Industrial Engineering Department,
Universidad de Chile, Santiago, Chile

NICOLÁS E. STIER-MOSES

Graduate School of Business, Columbia
University, New York, New York

A common behavioral assumption in the study of transportation and telecommunication networks is that travelers or packets, respectively, choose routes that they perceive as being the shortest under the prevailing traffic conditions [1]. The situation resulting from these individual decisions is one in which drivers cannot reduce their journey times by unilaterally choosing another route, which prompted Knight [2] to call the resulting traffic pattern an equilibrium. Nowadays, it is indeed known as the *Wardrop (or user) equilibrium* [3], and it is effectively thought of as a steady state evolving after a transient phase in which travelers successively adjust their route choices until a situation with stable route travel costs and route flows has been reached [4]. In a seminal contribution, Wardrop [5, p. 345] stated two principles that formalize this notion of equilibrium and the alternative postulate of the minimization of the total travel costs. His first principle reads:

The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Wardrop's first principle of route choice, which is identical to the notion postulated by Kohl [1] and Knight [2], became accepted as a sound and simple behavioral principle to describe the spreading of trips over alternate routes due to congested conditions [6].

Since its introduction in the context of transportation networks in 1952 and its mathematical formalization by Beckmann

et al. [7], transportation planners have been using Wardrop equilibrium models to predict commuters decisions in real-life networks [6,8]. These models have been and are still used today to evaluate alternative future scenarios and decide a route of actions. Typical examples include allocation of investment for capacity expansion when building roads and bridges, optimizing the value of tolls, and making policy decisions. The work of Korolis *et al.* [9] introduced these concepts in telecommunication research, which have become popular ever since [10].

THE BASIC MODEL

An instance of the *traffic assignment problem* is given by the transportation offer—represented by the network topology, road geometry, road capacity, and arc link travel cost functions—and the transportation demand—represented by the list of origin–destination (OD) pairs and their demand rates.

We consider a directed network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, and a set $\mathcal{C} \subseteq \mathcal{N} \times \mathcal{N}$ of commodities represented by OD pairs. For each $k \in \mathcal{C}$, a flow of demand rate equal to d_k must be routed from the corresponding origin to its destination. The basic model assumes that demands are arbitrarily divisible; in fact, the routing decision of a single individual has only an infinitesimal impact on other users. For $k \in \mathcal{C}$, let $\mathcal{R}_k$ be the set of routes in $\mathcal{G}$ connecting the corresponding origin and destination, and let $\mathcal{R} := \bigcup_{k \in \mathcal{C}} \mathcal{R}_k$. A link flow is a nonnegative vector $f = (f_a)_{a \in \mathcal{A}}$ describing the traffic rate in each link. Furthermore, a nonnegative, nondecreasing, and continuous link travel cost function $t_a(\cdot)$, with values in $\mathbb{R}_{\geq 0} \cup \{\infty\}$, maps the flow f_a on arc a to the time needed to traverse a . A route flow is a nonnegative vector $h = (h_r)_{r \in \mathcal{R}}$ that meets the demand, that is, $\sum_{r \in \mathcal{R}_k} h_r = d_k$ for $k \in \mathcal{C}$. Given a route flow, the corresponding link flow is easily computed as $f_a = \sum_{r \ni a} h_r$, for each $a \in \mathcal{A}$. For a flow f , the travel cost along

a route r is $c_r(f) := \sum_{a \in r} t_a(f_a)$. Let X be the set of feasible flows (f, h) and X_f its projection into the space of arc flows. Note that this set is a polytope given simply by the flow-conservation constraints for each commodity on every node (see **Multicommodity Flows** and Ahuja *et al.* [11]).

Interpreting Wardrop's first principle as requiring that all flow travels along shortest paths, a flow h is called a *Wardrop equilibrium* if and only if for all $k \in \mathcal{C}$, we have that

$$c_r(h) = \min_{q \in \mathcal{R}_k} c_q(h), \quad (1)$$

for all $r \in \mathcal{R}_k$ such that $h_r > 0$. Beckmann *et al.* [7] proved that such a flow always exists by considering the following min-cost multicommodity flow problem with separable objective function:

$$\min \left\{ \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(z) dz : f \in X_f \right\}. \quad (2)$$

The previous problem is convex because the objective is the integral of a nondecreasing function and, since its domain is a compact set, attains its optimum (for some background on convex optimization. Actually, it can be proved that its first-order optimality conditions are

$$c_r(h) \leq c_q(h), \quad \text{for all } k \in \mathcal{C} \text{ and all routes } r, q \in \mathcal{R}_k \text{ such that } h_r > 0, \quad (3)$$

which is equivalent to Equation (1).

If cost functions t_a are strictly increasing, f is unique (but there can be different flow decompositions h). For the case of nondecreasing costs, the vector of costs $(t_a(f_a))_{a \in \mathcal{A}}$ is unique under the possibly nonunique equilibria. Computationally, Equation (2) implies that an equilibrium can be computed efficiently using general convex optimization techniques (see the section titled "Computation of Wardrop Equilibria")¹. A formulation similar to the one presented here can be used to perform

a sensitivity analysis of Wardrop equilibria, which can be useful when testing the robustness of the model [12].

Yet another important characterization of traffic equilibrium problems, due to Smith [13], consists of reformulating Equation (2) as a variational inequality problem, see also Dafermos [14] and **Variational Inequalities**. Accordingly, a flow f is a user equilibrium if and only if

$$\sum_{a \in \mathcal{A}} t_a(f_a) f_a \leq \sum_{a \in \mathcal{A}} t_a(f_a) x_a, \quad \text{for all flows } x \in X_f. \quad (4)$$

Note that this inequality is a direct consequence of the fact that in equilibrium, users travel on shortest paths with respect to arc costs $t_a(f_a)$. Another reformulation that is useful to characterize equilibria in very general settings is given by nonlinear complementary problems [15] (see **Complementarity Problems**).

Charnes and Cooper [16] were the first to notice that the concepts of Nash and Wardrop equilibria are related. Haurie and Marcotte [17] prove that a Nash equilibrium in a network game with a finite number of players converges to a Wardrop equilibrium when the number of players increases (for some background on game theory, see the section titled "Noncooperative Strategic-Form Games" in this encyclopedia). For this reason, although the solution concepts are different, a Wardrop equilibrium can be viewed as an instance of a Nash equilibrium in a game with a large number of players. De Palma and Nesterov [18] look at generalizations and alternative definitions of the basic model and established conditions that guarantee the existence of equilibria. For example, Wardrop equilibria still exist if cost functions are only required to be lower semicontinuous. Marcotte and Patriksson [19] also discuss alternative definitions of equilibria in network games and the relationships between them.

though, uniqueness is not guaranteed even as flow on arcs because Equation (2) is not necessarily a convex problem and hence, may admit multiple local optima.

¹In the case of nonmonotone costs, the solutions to Equation (2) are also equilibria. In this case,

To conclude, let us note that in practice a transportation planner needs to find or estimate all the elements that comprise the model. The topology of the network is usually digitized from maps, if it is not already available. Link travel cost functions are calibrated from historical information using tabulated functions that relate geometry of the road to capacity [20,21]. One may need to also add tolls or other generalized costs to the arcs, which can be converted to the same units by using the average value of time for the population. The latter can usually be estimated from socioeconomic information coming from census data. Demand can be measured directly or may come from historical OD matrices that can be calibrated using up-to-date traffic counts [8].

MORE GENERAL MODELS

Many extensions of the model presented in the previous section have been analyzed and studied since the introduction of the Wardrop equilibrium model in 1952. We now present a selection of extensions we believe are particularly interesting.

An important generalization consists in allowing link travel cost functions to depend on the full vector of flows. In that case, the function that represents congestion can be encoded by the operator $t(f) : \mathbb{R}_{\geq 0}^A \rightarrow \mathbb{R}_{\geq 0}^A$. This is of practical relevance because the cost in one arc usually depends on the load in other arcs. Typical examples in the area of transportation modeling include representing intersections more accurately when cross streets influence each other, and two-way streets where traffic going one way can impact the reverse lanes. Furthermore, to consider different vehicle types interacting in the same network, one can create one copy of the network per vehicle type and have congestion depend on the full load, defined as the weighted sum across all copies. An example in the area of wireless telecommunications is interference, which can make delays of nearby cells grow. In the nonseparable case, it is convenient to require that the operator t is monotone, meaning that $(t(f) - t(f')) \cdot (f - f') \geq 0$, for all

$f, f' \in \mathbb{R}_{\geq 0}^A$, a generalization of monotonicity in the separable case. We say that costs are nonseparable, symmetric when for any two arcs, the influence of traffic in one arc to the congestion on the other is equal to the reverse influence (with symbols, $\partial t_a(f)/\partial f_b = \partial t_b(f)/\partial f_a$, for all $a, b \in A$ and all $f \in \mathbb{R}_{\geq 0}^A$). In that case, the integral in Equation (2) can be replaced by $\int_0^f t(z) dz$, which is well defined, and an equilibrium can still be computed using convex optimization techniques. An easy example of this case is when $t(f) = \Theta f + \theta$, with Θ a symmetric matrix of dimension $|A|$. It is worth noting that this symmetry condition is equivalent to the requirement that the game is *potential* with a continuum of agents [22] (for an introduction to potential games, the reader is referred to Monderer and Shapley [23]). In the general asymmetric case, the equilibrium cannot be formulated as a convex optimization problem (but note that there are exceptions and in some cases they can be formulated as nondifferentiable optimization problems [24]) and one has to resort to abstractions such as variational inequalities and non-linear complementarity problems. Indeed, Equation (4) still holds and can be used to prove existence and uniqueness results, as well as a way to compute equilibria.

Another important extension to the basic model that goes in a different direction is referred to by elastic demands. A network model normally represents a spectrum of options that a user of the system has. In reality, a user will normally have more alternatives than those present in a model. For example, in a car transportation model, some users may decide to take a subway when roads are too congested or cancel the trip when the combination of travel time, tolls, and gas exceeds the utility of the trip. This can be captured by a model with a demand of users that will participate only when their willingness to pay is higher than the cost of trips. In this case, the link travel cost functions in the equilibrium model are complemented by a demand function $\delta_k(\cdot) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ per OD pair $k \in C$. This function specifies that when the cheapest route cost for OD pair k is π_k , its demand equals $\delta_k(\pi_k)$. Hence, an equilibrium jointly satisfies

Equation (1) and $d_k = \delta_k(\min_{q \in \mathcal{R}_k} c_q(h))$. Depending on the separability of the costs, both the optimization and variational inequality problems can be extended to solve models with elastic demand.

Finally, several authors have looked at other ways to relax some of the basic assumptions. We mention a few examples here. A classical extension to the basic model incorporates capacity and other side-constraints to the equilibrium model [25] as a way of improving the solution quality and correcting the link travel cost functions so as to bring the flow pattern into agreement with the anticipated results. This can be handled by adding the constraints to the formulations and reinterpreting their dual variables as queueing delay caused by congestion. However, this approach is controversial and several authors have looked at alternative models that work better in some situations [26]. Also, Gabriel and Bernstein [27] studied the case of nonadditive models where the cost of a path can be given by more general expressions than just the sum over all arcs in the path. Examples of this include taking into account the variability of travel times and risk-aversion [28], tolls, and valuation of time.

COMPUTATION OF WARDROP EQUILIBRIA

In this section, we discuss some computational approaches to finding Wardrop equilibria. Let us start by describing the Frank–Wolfe method [29], which is a traditional algorithm that has been used to compute equilibria. It is an iterative descent method that works with the formulation shown in Equation (2) and eventually converges to the equilibrium. The algorithm keeps a current solution, and solves a linearized version of Equation (2) at every step to determine a feasible descent direction. Referring to the objective of that problem by $T(f)$ and to the current solution by f^i , the linearized objective is $T(f^i) + \nabla T(f^i) \cdot (f - f^i)$. The linearization enables the algorithm to decompose the problem by OD pairs, allowing it to find a shortest path in $\mathcal{R}_k$ for each commodity $k \in \mathcal{C}$ independently of each other. To identify the steepest descent

direction, it computes a shortest path with respect to the prevailing traffic conditions. In the subsequent line search, the original nonlinear problem is solved restricted to the segment defined by the feasible direction of descent. The algorithm terminates when a certain precision is achieved. To determine when this is the case, the convexity of the objective function is used to derive a lower bound on the value of an optimal solution. Alternatively, one can compute the *gap*—defined as the deviation from the shortest path—in the current solution and terminate when it is smaller than a threshold. It is well known that this algorithm always converges to a global minimum because Equation (2) is a convex program.

The standard Frank–Wolfe algorithm sometimes shows poor convergence because it tends to zigzag around the equilibrium solution [8,30–32]. Because “... the (Frank–Wolfe) algorithm is considered sufficiently good for practical use” [30, p. 100], most of the commercial implementations use this procedure. Still, there are many algorithms that were developed to address the slow convergence times. Below, we summarize a few of the main approaches.

LeBlanc *et al.* [33] introduced an improved version of the previous algorithm called *Partan* (parallel tangents), which was further studied by Florian *et al.* [34] and Arezki and Van Vliet [35], among others. This improvement is based on a more intelligent line search. It determines the descent direction using the results of two consecutive iterations, thereby diminishing the zigzagging effect. These two methods belong to a class called *partial linearization algorithms* in which the objective function is simplified to be able to find a search direction.

The structure of Equation (2) leads to *decomposition algorithms*, which separate the main problem into subproblems. The Frank–Wolfe algorithm is an example of this general method since it considers OD pairs separately after fixing the prevailing flows in one iteration. But the separation can be done in other ways, and viewed as a block version of the Gauss–Seidel and Jacobi algorithms. For example, a common decomposition separates flows by node of origin, whereby

every iteration assigns all destinations for each origin at the same time. A good example of this approach is given by Bar-Gera's algorithm [36], which is one of the most efficient, in existence, to compute Wardrop equilibria.

The class of *column generation algorithms* deal with a path formulation of the model. Since it is computationally challenging to keep track of the flow along all routes as opposed to maintaining a vector of flows per arc, instead of having one variable per route initially, a column generation algorithm adds routes at the time they are needed. After discovering new routes with the search direction procedure, an algorithm of this type forms a *restricted master problem* that consists of a path formulation of Equation (2) using only the routes discovered thus far (see also **Column Generation**). These methods are especially important when costs along routes are not additive or when there are constraints based on paths because an arc formulation is not powerful enough to represent the problem in that case. The class of *simplicial decomposition* algorithms finds the next iterate using the restricted master problem. Since all the route information previously computed is badly utilized by algorithms that perform line search, this class can solve problems more efficiently, albeit doing more work per iteration.

The *method of successive averages* is a commonly used heuristic method for computing Wardrop equilibria. This method starts by computing the costs on all arcs for an arbitrary feasible flow. Iteratively, it computes a new solution using an auxiliary linear program that keeps costs fixed, and updates the current solution by averaging it with the new one using a factor that depends on the iteration. This technique is specially useful for more complicated models where exact techniques are not readily available. Some examples are the dynamic and stochastic traffic assignments, see the section titled "More Advanced Models".

The case of elastic demands can be incorporated in the previous discussion since it involves adding another term to the convex minimization problem. The case of nonseparable, symmetric cost functions can be handled similarly to what was described earlier

since it admits a convex program formulation. In contrast, the asymmetric case, requires the machinery of variational inequalities or nonlinear complementarity problems. There exist standard algorithms to solve these classes of problems and some of the variants presented earlier for the separable case can be extended to this setting. Notice that since the nonseparable case does not admit a convex programming formulation, checking convergence must rely on regret or other related measures.

The search for efficient algorithms to compute Wardrop equilibria for the various classes of models is a very active area of research. Some of the latest efficient approaches are due to Dial [37], Florian *et al.* [38], Gentile [39], and Bar-Gera [40]. To conclude the discussion on computation, we note that there are some test problems available in the *Transportation Network Test Problems* website [41] that are typically used to study new algorithms.

There exist many commercial software packages that implement some of the algorithms described in this section. These and some additional packages also implement other variants of traffic assignment problems such as dynamic models that explicitly incorporate time [42–44], and simulation models that consider finer behavioral details that analytical models cannot handle [45]. A nonexhaustive list of software implementations is AIMSUN, CUBE, CONTRAM, DYNAMIT, DYNASMART, EMME/2, PARAMICS, TRANSCAD, TRANSIMS, TSIS-CORSIM, SATURN, VISUM-VISSIM, VISTA, and UROAD-UTPS.

EFFICIENCY OF WARDROP EQUILIBRIA

Since an equilibrium model considers that users unilaterally choose their routes to minimize *their* route cost, the solution is not necessarily efficient. A natural question is thus to quantify how inefficient a Wardrop equilibrium may be, where efficiency is measured as the flow's total travel time $C(f) := \sum_{r \in \mathcal{R}} c_r(h) h_r = \sum_{a \in \mathcal{A}} t_a(f_a) f_a$. Following Wardrop's second principle [5] that states that users minimize the total travel

time in the system, a system optimum f^* is an optimal solution to the min-cost multicommodity flow problem:

$$\min \{C(f) : f \in X_f\}. \quad (5)$$

It is not hard to observe that, in general, the total travel time incurred by an equilibrium can be arbitrarily larger than that of a social optimum. Consider, for instance, a two-node two-link network with unit demand and cost functions given by $t_1(f_1) = 1$ and $t_2(f_2) = f_2^n$, for some large value of n (Fig. 1). At equilibrium, all flow will use the second link, so that the total travel time will be 1. On the other hand in the system optimum a small fraction of the flow will use the first link, so that its total travel time will be close to 0, making their ratio grow to infinity. Nevertheless, a sequence of results initiated by Roughgarden and Tardos [46] and further developed in papers [47–51] states that if we only allow link travel cost functions belonging to a certain class, then the total travel time of an equilibrium is at most a constant times that of the system optimum.

Let us illustrate this group of results by considering the case in which for all $a \in \mathcal{A}$, $t_a(\cdot)$ is an affine function with nonnegative coefficients. Consider an equilibrium flow f and a system optimal flow f^* , then we have that

$$\begin{aligned} C(f) &\leq \sum_{a \in \mathcal{A}} t_a(f_a) f_a^* = \sum_{a \in \mathcal{A}} t_a(f_a^*) f_a^* \\ &\quad + \sum_{a \in \mathcal{A}} f_a^* (t_a(f_a) - t_a(f_a^*)) \leq \sum_{a \in \mathcal{A}} t_a(f_a^*) f_a^* \end{aligned}$$

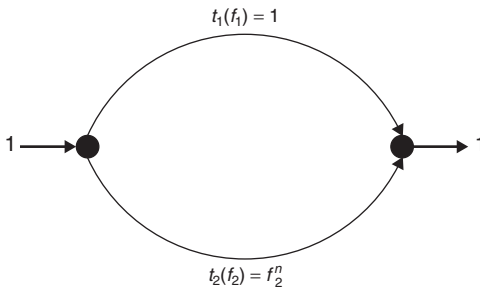


Figure 1. Instance where the Wardrop equilibrium is unboundedly worse than the system optimum.

$$\begin{aligned} &+ \sum_{a \in \mathcal{A}: f_a > f_a^*} f_a^* (t_a(f_a) - t_a(f_a^*)) \\ &\leq \sum_{a \in \mathcal{A}} t_a(f_a^*) f_a^* + \frac{1}{4} \sum_{a \in \mathcal{A}} t_a(f_a) f_a, \end{aligned}$$

implying that $C(f) \leq (4/3) \cdot C(f^*)$. The first inequality in the previous derivation holds because of Equation (4), while the last inequality follows since for affine functions the shaded area in Fig. 2 is at most 25% of the area of the big rectangle. This theorem is due to Roughgarden and Tardos [46], and the short proof presented here appeared in Correa *et al.* [49]. The result implies that, in the worst case among all possible networks, the inefficiency introduced by the self-minded behavior of an equilibrium is never worse than 1/3.

The proof above easily extends to other classes of link travel cost functions by only changing the 25% with the corresponding quantity for a given class of functions. Probably the most interesting aspect of this result is that the efficiency loss depends on the allowable functions rather than on the topology of the network. Furthermore, the idea behind this proof provides another interesting result first derived in Roughgarden and Tardos [46]. Indeed, note that if $t_a(\cdot)$ are arbitrary non-decreasing functions, then $\sum_{a \in \mathcal{A}} x_a t_a(f_a) \leq$

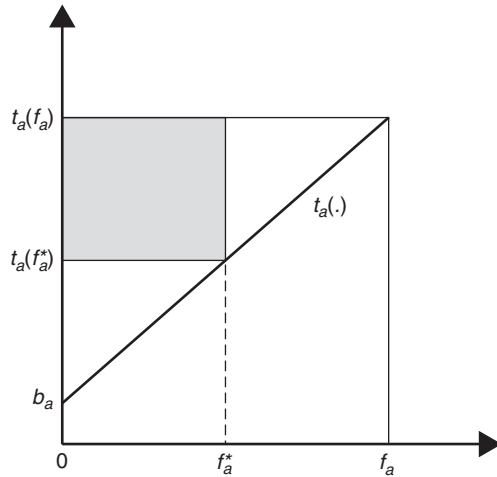


Figure 2. Illustration of the proof of the 4/3 result.

$\sum_{a \in A} \max\{f_a t_a(f_a), x_a t_a(x_a)\} \leq \sum_{a \in A} f_a t_a(f_a) + \sum_{a \in A} x_a t_a(x_a)$, where the f_a 's and x_a 's are any nonnegative numbers. Consider now an equilibrium flow f of a given instance, and a system optimal flow x of a similar instance where demands rates d are doubled. From Equation (4), and since $x/2 \in X_f$, we have

$$\begin{aligned} C(f) &= 2C(f) - C(f) \leq 2 \sum_{a \in A} t_a(f_a) \cdot (x_a/2) \\ &\quad - C(f) \leq \sum_{a \in A} t_a(f_a) \cdot x_a - C(f) \leq C(x). \end{aligned}$$

In other words, for arbitrary nondecreasing link travel cost functions, the cost of a Wardrop equilibrium is at most the cost of an optimal solution with the demand doubled. For a restricted set of cost functions, one can provide improved results [49]. For example, under affine costs one can prove that the same statement holds with 25% more demand.

As general equilibria typically do not minimize the social cost, Koutsoupias and Papadimitriou [52] proposed to analyze the inefficiency of equilibria from a worst-case perspective; this led to the notion of “price of anarchy” [53], which is the ratio of the worst social cost of a Nash equilibrium to the cost of an optimal solution. In the context of our traffic model this quantity has been analyzed in a series of papers for increasingly more general classes of cost functions and other model features. The result previously described implies that this worst-case ratio is 4/3. This was extended to more general link travel cost functions by Roughgarden [47] and by Correa *et al.* [48], who basically proved that the efficiency loss in this setting is independent of the topology of the network. Chau and Sim [50] considered the case of nonseparable, symmetric cost functions with elastic demands. They proved that the efficiency loss can be bounded in a similar way as what is described here. Perakis [51] considered general nonseparable cost functions and proved upper bounds using variational inequalities as well. Her bounds depend on two parameters that measure the asymmetry and the nonlinearity of the cost functions considered. Farzad *et al.* [54] provide results of a similar flavor in a closely related model. In their setting the flow

(players) have a priority and thus, in any given link a flow particle only experiences a travel time $t_a(x_a)$, where x_a is the amount of flow using link a , having higher priority. Interestingly in this context a system optimum is given by Equation (2).

MORE ADVANCED MODELS

This section discusses some extensions of the basic Wardrop equilibrium model that consider variations on the structure of offer and demand in the network.

In most urban transportation networks, commuters do not have to pay the cost they impose to others by a particular route choice, leading to the bad utilization of the available capacity alluded to in the section titled “Efficiency of Wardrop Equilibria”. Since congestion increases sharply with road utilization, having relatively few drivers switch to other routes may significantly improve commute times. Starting with the seminal idea of Vickrey [55,56], many transportation economists have advocated the use of congestion pricing to achieve this goal. The scheme forces drivers to pay a toll when entering congested areas. The underlying idea is to charge drivers the externality they impose to others because when commuters internalize these externalities, the corresponding choices maximize the system welfare. Singapore introduced congestion pricing in 1975, London in 2003 [57,58], and Stockholm in 2007. Increasingly, many large cities have been debating whether a congestion pricing scheme should be adopted. Nevertheless, it has been very hard to implement congestion pricing because of technical, economical, and political problems. There is a large body of research on finding the set of tolls for a given network that optimizes a given objective (e.g., social welfare, revenue, number of tollbooths) under constraints (such as a budget balance, maximum number of tollbooths, or restrictions on their location). For more details, see the books [59,60].

Stochastic user equilibrium models date back to the 1970s, when Dial [61] proposed a model where the demand on each OD is distributed among routes (with random lengths)

according to a logit distribution, in the case of uncongested traffic networks. To reduce route enumeration he considered that flow is distributed only among “efficient routes.” This model has been widely studied and extended [62,63]. Furthermore, Daganzo and Sheffi [64] looked at the case of dependent route costs, while Fisk [65] studied the model in the context of congested networks, obtaining an equivalent optimization problem. Methods avoiding route enumeration have been proposed by Bell [66], Larsson *et al.* [67], and Maher [68], also leading to equivalent optimization problems in the spirit of Fisk’s. Based on the work of Akamatsu [69], Baillon and Cominetti [70] proposed a more general concept called *Markovian traffic equilibrium*, provided an equivalent optimization problem, and established the convergence of the method of successive averages in that context. There are complementary models that consider that travel times themselves are stochastic instead of considering that the perception is stochastic. Some examples are the papers [28,71–75].

Another important extension is modeling public transportation systems. The time commuters need to wait until a bus arrives to the stop adds a difficulty that was not present in models of privately owned vehicles. Indeed, in order to balance waiting and travel time, users’ strategies may involve selecting a subset of bus lines and boarding the first available one. This idea was pioneered by Chiriqui and Robillard [76], who considered a network with a single OD pair and n bus lines serving it, each characterized by a travel time and a frequency. They provided a simple efficient algorithm to solve the problem. Spiess and Florian [77] generalized that model to arbitrary networks introducing the notion of general users’ strategies. This was further put in graph theoretic terms by Nguyen and Pallotino [78], who called these strategies *hyperpaths* and incorporated travel times dependent on congestion. However, besides increasing travel times, congestion also increases waiting times because a user may not be able to board a selected bus. This is challenging and Wu *et al.* [79] and Bouzaïene-Ayari *et al.* [80] attempted to deal with this problem. Cominetti and Correa [81]

considered the frequency of a bus line as a function of the flow in the network and called this function ‘*effective frequency*.’ They proved that an equilibrium in this context exists via an equivalent optimization problem. Cepeda *et al.* [82] obtained a new characterization of equilibria in the congested setting. This led to an effective algorithm that is currently part of EMME/2.

In some situations, such as logistics networks, it is natural to consider that some players control nonnegligible amounts of flow that can be split among several routes, as modeled by an *atomic* network game [83]. In this context, the equilibrium becomes significantly more difficult to characterize and compute because, even for the basic assumptions, the game is generally not potential. Although the existence of equilibria is still guaranteed [84], an instance may possess multiple equilibria [85], and no equivalent convex optimization problem is known even for general separable instances. Furthermore, surprisingly an equilibrium of this game can be less efficient than that of the corresponding nonatomic instance (as in the section titled “The Basic Model”), even in simple networks with two OD pairs [86,87]. Cominetti *et al.* [87] generalized the results of the section titled “Efficiency of Wardrop Equilibria” to this setting, getting a bound of $3/2$ when cost functions are affine. It is an open question whether this bound is tight. Swamy [88] proved that tolls that induce an optimal routing of the game with finite players always exist.

Generalizing a classic paper by Rosenthal [89], Milchtaich [90] considered a generalization of the basic model referred to by *congestion games*. In this model the network is abstracted away and each user selects a strategy that consists of a subset of arcs (instead of a route), selected among a set of feasible strategies defined in advance. The defining characteristic of these models is that a link travel cost function for one arc just depends on its demand. Milchtaich himself and others have looked at existence, uniqueness, and computation of equilibria in congestion games with and without atoms, and with homogeneous and heterogeneous cost functions. Actually, some of the results presented

in the section titled “Efficiency of Wardrop Equilibria” were originally presented for this kind of model.

Finally, another important area of research involves extending the basic model along the time dimension. Starting with the seminal work of Merchant and Nemhauser [42], many articles have been published about *dynamic user equilibria*. These models consider time-varying conditions on both the offer and demand side of the network. In contrast to the static model discussed here, there is less agreement on the basic characteristics of the model, and on the conditions that such a model should verify, which highlights the difficulty of the problem. Nevertheless, transportation planners commonly make use of tools that rely on analytic or simulation models of this kind. We refer the readers to Ran and Boyce [91], Peeta and Ziliaskopoulos [44], and the report by the Transportation Research Board [92] for more details about dynamic models.

Another relevant issue related to equilibria in general, and to Wardrop equilibria, in particular, is whether players “learn” the equilibria of the game and if so, how fast they do it. There is a rich literature mainly concerned with the dynamics that govern the behavior of players. The book by Sandholm [93] contains an in-depth treatment of the subject. In particular, he describes some learning dynamics under which players of a congestion game, such as those described in this article, converge to an equilibrium. Related to learning and equilibria, there has been limited empirical work to validate whether test subjects behave as the theory of Wardrop equilibria predicts they do [94–98].

FURTHER READING

For an extensive discussion on modeling and algorithmic techniques, we refer the reader to Magnanti [99], Sheffi [8], Nagurney [100], Patriksson [30], Florian and Hearn [31], and Marcotte and Patriksson [19]. Altman *et al.* [10] surveys the use of this framework in modeling telecommunication networks. Nisan *et al.* [101] and the references therein provide details on the efficiency of equilibria in network games.

Acknowledgments

We would like to thank a referee, Patrice Marcotte, and Bill Sandholm for numerous suggestions and comments regarding an earlier version of this article.

REFERENCES

1. Kohl JG. Der verkehr und die ansiedelungen der menschen in ihrer abhängigkeit von der gestaltung der erdoberfläche. Dresden/Leipzig: Arnold; 1841.
2. Knight FH. Some fallacies in the interpretation of social cost. *Q J Econ* 1924;38: 582–606.
3. Dafermos SC, Sparrow FT. The traffic assignment problem for a general network. *J Res U.S. Natl Bur Stand* 1969;73B: 91–118.
4. Larsson T, Patriksson M. Side constrained traffic equilibrium models—analysis, computation and applications. *Transp Res* 1999; 33B:233–264.
5. Wardrop JG. Some theoretical aspects of road traffic research. *Proc Inst Civ Eng Part II* 1952;1:325–378.
6. Florian M. Untangling traffic congestion: application of network equilibrium models in transportation planning. *OR/MS Today* 1999;26(2):52–57.
7. Beckmann MJ, McGuire CB, Winsten CB. *Studies in the economics of transportation*. New Haven (CT): Yale University Press; 1956.
8. Sheffi Y. *Urban transportation networks*. Englewood (NJ): Prentice-Hall; 1985.
9. Korilis Y, Lazar A, Orda A. Architecting noncompetitive networks. *IEEE J Sel Areas Commun* 1995;13(7):1241–1251.
10. Altman E, Boulogne T, El-Azouzi R, *et al.* A survey on networking games in telecommunications. *Comput Oper Res* 2006;33(2): 286–311.
11. Ahuja RK, Magnanti TL, Orlin JB. *Network flows: theory, algorithms, and applications*. Englewood Cliffs (NJ): Prentice-Hall; 1993.
12. Patriksson M. Sensitivity analysis of traffic equilibria. *Transp Sci* 2004;38(3): 258–281.
13. Smith MJ. The existence, uniqueness and stability of traffic equilibria. *Transp Res* 1979;13B(4):295–304.

14. Dafermos SC. Traffic equilibrium and variational inequalities. *Transp Sci* 1980;14(1): 42–54.
15. Aashtiani HZ, Magnanti TL. Equilibria on a congested transportation network. *SIAM J Algebr Discrete Methods* 1981;2(3):213–226.
16. Charnes A, Cooper WW. Multicopy traffic network models. In: Herman R, editor. *Theory of traffic flow*. Amsterdam: Elsevier; 1961. pp. 85–96. *Proceedings of the Symposium on the Theory of Traffic Flow held at the General Motors Research Laboratories; Warren (MI); 1959.*
17. Haurie A, Marcotte P. On the relationship between Nash-Cournot and Wardrop equilibria. *Networks* 1985;15(3):295–308.
18. de Palma A, Nesterov Y. Optimization formulations and static equilibrium in congested transportation networks. CORE Discussion Paper 9861. Belgium: Université Catholique de Louvain, Louvain-la-Neuve; 1998.
19. Marcotte P, Patriksson M. Traffic equilibrium. In: Barnhart C, Laporte G, editors. *Volume 14, Transportation, Handbooks in Operations Research and Management Science*. New York: Elsevier; 2007. Chapter 10. pp. 623–713.
20. Transportation Research Board. *Highway capacity manual 2000*; 2000.
21. Bureau of Public Roads. *Traffic assignment manual*. Washington (DC): U.S. Department of Commerce, Urban Planning Division; 1964.
22. Sandholm WH. Potential games with continuous player sets. *J Econ Theory* 2001;97: 81–108.
23. Monderer D, Shapley LS. Potential games. *Games Econ Behav* 1996;14:124–143.
24. Hearn DW, Lawphongpanich S, Nguyen S. Convex programming formulations of the asymmetric traffic assignment problem. *Transp Res* 1984;18B(4-5):357–365.
25. Hearn DW. Bounding flows in traffic assignment models. Technical Report 80-4. Gainesville (FL): Department of Industrial and Systems Engineering, University of Florida; 1980.
26. Marcotte P, Nguyen S, Schoeb A. A strategic flow model of traffic assignment in static capacitated networks. *Oper Res* 2004;52(2): 191–212.
27. Gabriel S, Bernstein D. The traffic equilibrium problem with nonadditive path costs. *Transp Sci* 1997;31:337–348.
28. Ordóñez F, Stier-Moses NE. Wardrop equilibria with risk-averse users. *Transp Sci* 2010;44(1):63–86.
29. Frank M, Wolfe P. An algorithm for quadratic programming. *Nav Res Log Q* 1956;3: 95–110.
30. Patriksson M. *The traffic assignment problem: models and methods*. Utrecht, The Netherlands: VSP; 1994.
31. Florian M, Hearn DW. Network equilibrium models and algorithms. In: Ball MO, Magnanti TL, Monma CL, editors. *Volume 8, Network routing, Handbooks in Operations Research and Management Science*. New York: Elsevier; 1995. Chapter 6. pp. 485–550.
32. Patriksson M. Algorithms for computing traffic equilibria. *Netw Spat Econ* 2004;4:23–38.
33. LeBlanc LJ, Helgason RV, Boyce DE. Improved efficiency of the Frank-Wolfe algorithm for convex network programs. *Transp Sci* 1985;19:445–462.
34. Florian M, Guélat J, Spiess H. An efficient implementation of the “Partan” variant of the linear approximation method for the network equilibrium problem. *Networks* 1987;17:319–339.
35. Arezki Y, Van Vliet D. A full analytical implementation of the Partan/Frank-Wolfe algorithm for equilibrium assignment. *Transp Sci* 1990;24(1):58–62.
36. Bar-Gera H. Origin-based algorithm for the traffic assignment problem. *Transp Sci* 2002;36(4):398–417.
37. Dial RB. A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. *Transp Res* 2006;B40(10):917–936.
38. Florian M, Constantin I, Florian D. A new look at the projected gradient method for equilibrium assignment. *Transp Res Rec* 2009;2090:10–16.
39. Gentile G. Linear user cost equilibrium: a new algorithm for traffic assignment. Working paper. 2010.
40. Bar-Gera H. Traffic assignment by paired alternative segments. *Transp Res Part B* 2010;44(8):1022–1046.
41. Bar-Gera H. *Transportation network test problems*; 2002. Available at <http://www.bgu.ac.il/~bargera/tntp/>. Accessed 2010 Aug 28.
42. Merchant DK, Nemhauser GL. A model and an algorithm for the dynamic traffic assignment problems. *Transp Sci* 1978;12: 183–199.

43. Ben-Akiva ME. Dynamic network equilibrium research. *Transp Res* 1985;19A(5/6): 429–431.
44. Peeta S, Ziliaskopoulos AK. Foundations of dynamic traffic assignment: the past, the present and the future. *Netw Spat Econ* 2001;1:233–265.
45. Kitamura R, Kuwahara M, editors. Volume 31, Simulation approaches in transportation analysis, Operations Research/Computer Science Interfaces Series. New York: Springer; 2005.
46. Roughgarden T, Tardos E. How bad is selfish routing? *J ACM* 2002;49(2):236–259.
47. Roughgarden T. The price of anarchy is independent of the network topology. *J Comput Syst Sci* 2003;67(2):341–364.
48. Correa JR, Schulz AS, Stier-Moses NE. Selfish routing in capacitated networks. *Math Oper Res* 2004;29(4):961–976.
49. Correa JR, Schulz AS, Stier-Moses NE. A geometric approach to the price of anarchy in nonatomic congestion games. *Games Econ Behav* 2008;64:457–469.
50. Chau CK, Sim KM. The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. *Oper Res Lett* 2003;31(5):327–334.
51. Perakis G. The “price of anarchy” under nonlinear and asymmetric costs. *Math Oper Res* 2007;32(3):614–628.
52. Koutsoupias E, Papadimitriou CH. Worst-case equilibria. In: Meinel C, Tison S, editors. Volume 1563, Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS), Lecture Notes in Computer Science; Trier, Germany. Heidelberg: Springer; 1999. pp. 404–413.
53. Papadimitriou CH. Algorithms, games, and the Internet. Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC), Hersonissos, Greece. New York: ACM Press; 2001. pp. 749–753.
54. Farzad B, Olver N, Vetta A. A priority-based model of routing. *Chic J Theor Comput Sci* 2008;1–29.
55. Vickrey WS. A proposal for revising New York’s subway fare structure. *J Oper Res Soc Am* 1955;3(1):38–68.
56. Vickrey WS. Congestion theory and transport investment. *Am Econ Rev* 1969;59: 251–261.
57. Santos G. Urban congestion charging: a comparison between London and Singapore. *Transp Rev* 2005;25(5):511–534.
58. Santos G, Fraser G. Road pricing: lessons from London. *Econ Policy* 2006;21(46): 263–310.
59. Yang H, Huang H-J. Mathematical and economic theory of road pricing. Oxford: Elsevier; 2005.
60. Lawphongpanich S, Smith MJ, Hearn DW, editors. Volume 101, Mathematical and computational models for congestion charging, Applied optimization. Berlin: Springer; 2006.
61. Dial RB. A probabilistic multi-path traffic assignment algorithm which obviates path enumeration. *Transp Res* 1971;5(2):83–111.
62. Trahan M. Probabilistic assignment: an algorithm. *Transp Sci* 1974;8(4):311–320.
63. Sheffi Y, Powell W. An algorithm for the traffic assignment problem with random link costs. *Networks* 1982;12:191–207.
64. Daganzo C, Sheffi Y. On stochastic models of traffic assignment. *Transp Sci* 1977;11(3): 253–274.
65. Fisk C. Some developments in equilibrium traffic assignment. *Transp Res* 1980;14B(3): 243–255.
66. Bell M. Alternatives to Dial’s logit assignment algorithm. *Transp Res* 1995;29B: 287–295.
67. Larsson T, Liu Z, Patriksson M. A dual scheme for traffic assignment problems. *Optimization* 1997;42:323–358.
68. Maher M. Algorithms for logit-based stochastic user equilibrium assignment. *Transp Res* 1998;32B(8):539–549.
69. Akamatsu T. Cyclic flows, Markov processes and stochastic traffic assignment. *Transp Res* 1996;30B(5):369–386.
70. Baillon J-B, Cominetti R. Markovian traffic equilibrium. *Math Program* 2008;111B: 33–56.
71. Liu H, Ban X, Ran B, *et al.* An analytical dynamic traffic assignment model with stochastic network and travelers’ perceptions. *Transp Res Rec* 2002;1783: 125–133.
72. Mirchandani PB, Soroush H. Generalized traffic equilibrium with probabilistic travel times and perceptions. *Transp Sci* 1987;21: 133–152.
73. Uchida T, Iida Y. Risk assignment: a new traffic assignment model considering risk of travel time variation. In: Daganzo CF, editor. Proceedings of the 12th International

- Symposium on Transportation and Traffic Theory. Amsterdam: Elsevier; 1993. pp. 89–105.
74. Bell MGH, Cassir C. Risk-averse user equilibrium traffic assignment: an application of game theory. *Transp Res* 2002;36B(8): 671–681.
 75. Lo HK, Tung Y-K. Network with degradable links: capacity analysis and design. *Transp Res* 2003;37B(4):345–363.
 76. Chiriqui C, Robillard P. Common bus lines. *Transp Sci* 1975;9:115–121.
 77. Spiess H, Florian M. Optimal strategies: a new assignment model for transit networks. *Transp Res* 1989;23B:83–102.
 78. Nguyen S, Pallotino S. Equilibrium traffic assignment in large scale transit networks. *Eur J Oper Res* 1988;37(2):176–186.
 79. Wu JH, Florian M, Marcotte P. Transit equilibrium assignment: a model and solution algorithms. *Transp Sci* 1994;28(3): 193–203.
 80. Bouzaïene-Ayari B, Gendreau M, Nguyen S. Modeling bus stops in transit networks: A survey and new formulations. *Transp Sci* 2001;35(3):304–321.
 81. Cominetti R, Correa JR. Common-lines and passenger assignment in congested transit networks. *Transp Sci* 2001;35(3):250–267.
 82. Cepeda M, Cominetti R, Florian F. A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. *Transp Res* 2006;40B: 437–459.
 83. Harker PT. Multiple equilibrium behaviors of networks. *Transp Sci* 1988;22(1):39–46.
 84. Rosen JB. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* 1965;33(3):520–534.
 85. Bhaskar U, Fleischer L, Hoy D, *et al.* Equilibria of atomic flow games are not unique. In: Mathieu C, editor. SODA. New York, Philadelphia (PA): SIAM; 2009. pp. 748–757.
 86. Catoni S, Pallotino S. Traffic equilibrium paradoxes. *Transp Sci* 1991;25:240–244.
 87. Cominetti R, Correa JR, Stier-Moses NE. Network games with atomic players. *Oper Res* 2009;57(6):1421–1437.
 88. Swamy C. The effectiveness of Stackelberg strategies and tolls for network congestion games. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). New Orleans (LA), Philadelphia (PA): SIAM; 2007. pp. 1133–1142.
 89. Rosenthal RW. A class of games possessing pure-strategy Nash equilibria. *Int J Game Theory* 1973;2(1):65–67.
 90. Milchtaich I. Congestion games with player-specific payoff functions. *Games Econ Behav* 1996;13:111–124.
 91. Ran B, Boyce DE. Modeling dynamic transportation networks: an intelligent transportation system oriented approach. 2nd ed. Berlin: Springer; 1996.
 92. Transportation Research Board. A primer for dynamic traffic assignment. 2010.
 93. Sandholm WH. Population games and evolutionary dynamics. Cambridge (MA): MIT Press; 2011.
 94. Rapoport A, Mak V, Zwick R. Navigating congested networks with variable demand: experimental evidence. *J Econ Psychol* 2006; 27(5):648–666.
 95. Selten R, Chmura T, Pitz T, *et al.* Commuters route choice behaviour. *Games Econ Behav* 2007;58(2):394–406.
 96. Ramadurai G, Ukkusuri S. Dynamic traffic equilibrium: theoretical and experimental network game results in single bottleneck model. *J Transp Res Rec* 2007;2029:1–13.
 97. Rapoport A, Kugler T, Dugar S, *et al.* Choice of routes in congested traffic networks: experimental tests of the Braess paradox. *Games Econ Behav* 2009;65(2):538–571.
 98. Morgan J, Orzen H, Sefton M. Network architecture and traffic flows: experiments on the Pigou-Knight-Downs and Braess paradoxes. *Games Econ Behav* 2009;66(1):348–372.
 99. Magnanti TL. Models and algorithms for predicting urban traffic equilibria. In: Florian M, editor. Transportation planning models. Amsterdam: North Holland Publishing Co.; 1984. pp. 153–185. Proceedings of the course given at the International Center for Transportation Studies (ICTS). Amalfi, Italy; 1982.
 100. Nagurney A. Network economics: a variational inequality approach. 2nd ed. Dordrecht, The Netherlands: Kluwer Academic; 1999.
 101. Nisan N, Roughgarden T, Tardos E, *et al.* Algorithmic game theory. Cambridge: Cambridge University Press; 2007.

WARRANTY MODELING

MARLIN U. THOMAS
Department of Electrical and
Computer Engineering, Air Force
Institute of Technology,
Wright-Patterson AFB, Ohio

Warranties are binding commitments by producers to provide acceptable quality of their products or services for some prescribed period of time following the sale or delivery of the goods. Product guarantees and warranties have been around as far back in time as is known to mankind, but it has only been during the past few decades that quantitative methods have evolved for analyzing and evaluating economic decisions relating to warranties. The early analysis methods were focused on product marketing, and accounting for the amount of warranty expenditures that should be included in manufacturing costs. However, as competition grew and became more global, manufacturers and service providers had to become more quality focused thereby making quality and warranty essential elements of corporate planning. Customers rely on warranties for protection against low quality products and manufacturers use them for marketing and building confidence in their products for purposes of market penetration and being competitive. The difficulties and challenges in providing suitable warranty programs and hence support for the need for modeling are: (i) developing appropriate product warranty policies, (ii) methods for predicting associated warranty costs and future burdens, and (iii) methods for planning quality and reliability improvement options. The relevant economic issues in dealing with these problem areas involve risk due to uncertainties in the customer preferences and expectations, and the failure characteristics of the products. Details on the history and background on warranties can be found in Refs 1 and 2.

This article provides an overview of the quantitative models for examining

economic decisions that involve warranties. Warranty models are constructed relative to the customer's costs and benefits from the warranty, or to that of the producer of the products. Throughout this article, the models will reflect the position of the producer or manufacturer. In the section titled "Product Warranties" we discuss the types of product warranties and provide an account of the basic policies that are applied to consumer products. In the section titled "Warranty Cost Models," warranty economic models developed for cost considerations under various types of warranty policies, examining optimal policies, and determining reserve funds for covering future warranty costs are discussed. We will then cover models for dealing with machine replacement and maintenance decisions. It should be noted that while our focus is still with the manufacturer's cost considerations for these models, the manufacturer is actually the customer whose goal is to maintain operational equipment, some of which is under warranties. Finally, we will provide summary remarks that will include comments on some ongoing and future modeling directions.

PRELIMINARY ASSUMPTIONS AND NOTATION

1. A warranty claim is filed each time a product failure occurs during the warranty period.
2. Failure times, T are independent and identically distributed (iid) with common probability distribution F .
3. Failure cost X_i can depend on failure time T_i but the pairs, (X_i, T_i) form an independent sequence for $i = 1, 2, \dots$

PRODUCT WARRANTIES

A warranty is a form of agreement between parties with specified conditions for: (i) the length of time during which the warranty is

to be in effect, (ii) the method of compensation for the product or service during the warranty period, and (iii) the conditions for which the warranty will apply or will not apply to the product or service. The parties are generally between customers and a producer or manufacturer, although it could also be between customers and a brokerage organization other than the producer. From a modeling perspective, the first two specifications, period of coverage and method of compensation are the salient issues for decision models with the third category being a descriptive category for the various types of warranties.

There are two distinct types of warranties relative to the period of coverage. Most consumer warranties are for a *fixed period* of time that is measured in age or product usage from the date of purchase. For a warranty of length $w > 0$, the coverage applies only for reported warranty failures during the period $(0, w]$. The second type of warranties is the class of renewable or *renewing period* warranties whereby the initial period of coverage is $(0, w]$ but when a failure occurs within this warranty period a new warranty is granted. This process is repeated until the item survives a time of length w . Sample realizations of nonrenewing and renewing period warranties are shown in Fig. 1. Renewing period warranties are generally applied to relatively inexpensive products that have a very low likelihood of failure during the warranty period. They are less common than fixed period warranties.

Free Replacement Warranty (FRW)

The most common compensation methods are based on one of the compensation schemes. The first and simplest is the class of free replacement policies, whereby the cost of repair or replacement of the warranted item is completely covered by the manufacturer during the warranty period. Let T be a non-negative random variable representing the time to failure of a product with a repair or replacement cost to the manufacturer of c dollars per unit. Under the free replacement warranty (FRW) policy, the unit cost to the manufacturer at time t , for a warranty of length $w > 0$ is

$$X(t) = \begin{cases} c, & 0 \leq t \leq w, \\ 0, & t > w, \end{cases} \quad (1)$$

where $f(t)$ and $F(t)$ are the probability density and distribution functions for T .

Prorata Rebate Warranty (PRW)

The second class of compensation policies is prorata rebate policies for which the costs of repair or replacement are shared by the customer and the manufacturer according to a prescribed prorate scheme. Prorata rebate warranty (PRW) policies, sometimes called *prorata replacement policies* are generally applied to products that are not repairable and the proportion of sharing between the manufacturer and the customer is a function of the age of the product at the time of failure. Under a PRW policy with a linear prorata function, an item that costs the manufacturer c dollars per unit that is sold with

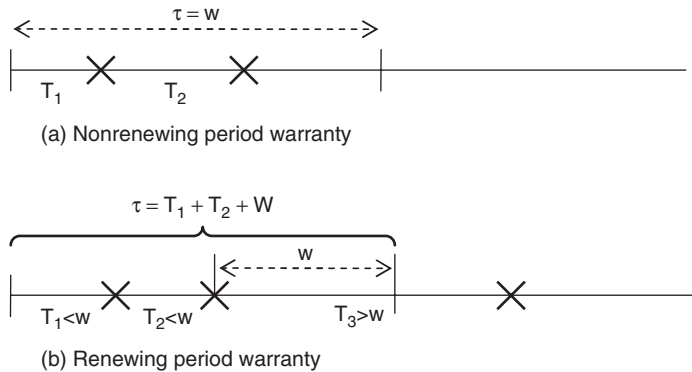


Figure 1. Cycle time τ for non-renewing and renewing warranties.

a warranty of length $w > 0$ has a unit cost to the manufacturer at time t given by

$$X(t) = \begin{cases} c \left(\frac{w-t}{w} \right), & 0 < t \leq w, \\ 0, & t > w. \end{cases} \quad (2)$$

The customer must be willing to pay the residual amount of ct/w to execute this policy. The expected unit cost to the manufacturer is

$$E[X(t)] = c \int_0^w (1 - t/w) f(t) dt \quad (3)$$

or

$$E[X(t)] = cF(w) - \frac{c}{w} \int_0^w tf(t) dt. \quad (4)$$

where $f(t)$ and $F(t)$ are the probability density and distribution functions for T .

Combined FRW PRW

Another policy that is common to many products is a combined policy where a product is covered by an FRW policy for a period $(0, w_1)$ followed by a PRW policy during (w_1, w_2) . Denoting the initial cost of the item by c , the unit cost to the manufacturer for a warranty failure is given by

$$X(t) = \begin{cases} c, & 0 \leq t \leq w_1, \\ c \left(\frac{w_2-t}{w_2-w_1} \right), & w_1 < t \leq w_2, \\ 0, & t > w_2 \end{cases} \quad (5)$$

and the expected unit cost is given by

$$\begin{aligned} E[X(t)] &= cF(w_1) + \left(\frac{cw_2}{w_2 - w_1} \right) \\ &\quad \times [F(w_2) - F(w_1)] \\ &\quad - \frac{c}{w_2 - w_1} \int_{w_1}^{w_2} tf(t) dt. \end{aligned} \quad (6)$$

An alternate form of Equation (6) that is sometimes easier to work with follows by integrating the last term by parts and simplifying to obtain

$$E[X(t)] = \frac{c}{w_2 - w_1} \int_{w_1}^{w_2} F(t) dt. \quad (7)$$

Example 1. A product sold under a PRW policy has a failure time T distributed exponential with parameter $\lambda > 0$.

$$F(t) = 1 - e^{-\lambda t}, t \geq 0. \quad (8)$$

The expected unit cost from Equation (4) is

$$E[X(t)] = c - \frac{c}{\lambda w} (1 - e^{-\lambda w}). \quad (9)$$

Example 2. A product is sold under a FRW/PRW policy with unit costs given in Equation (5). The failure time T is distributed Weibull with

$$F(t) = 1 - e^{-(t/\theta)^\beta}, t \geq 0, \quad (10)$$

where $\beta > 0, \theta > 0$.

Substituting into Equation (8),

$$E[X(t)] = c - \frac{c}{w_2 - w_1} \int_{w_1}^{w_2} e^{-(t/\theta)^\beta} dt. \quad (11)$$

Warranties with Discounting

Cost models for products under relatively long warranty periods with high repair and maintenance costs should include the provision of the time value of money. Letting $\delta, 0 < \delta < 1$ be the discount factor, the expected discounted unit cost is

$$E[X_D(t)] = E[e^{-\delta t} X(t)]. \quad (12)$$

The expected discounted unit warranty cost from Equations (1) and (3) are given by

$$E[X_D(t)] = \begin{cases} c \int_0^w e^{-\delta t} f(t) dt, & \text{FRW,} \\ \frac{c}{w} \int_0^w \int_0^t e^{-\delta \tau} f(\tau) d\tau dt, & \text{PRW.} \end{cases} \quad (13)$$

Example 3. For the case of discounting, the cost for the PRW policy with T distributed exponential in Example 1, from Equation (13) is

$$\begin{aligned} E[X_D(t)] &= \frac{c}{w} \int_0^w \int_0^t \lambda e^{-\delta \tau} e^{-\lambda \tau} d\tau dt \\ &= \frac{c\lambda}{w} \int_0^w \frac{1 - e^{-(\delta + \lambda)\tau}}{\lambda + \delta} d\tau \end{aligned}$$

or

$$\begin{aligned} E[X_D(t)] &= \frac{\lambda c}{\lambda + \delta} - \frac{\lambda c}{(\lambda + \delta)^2} [1 - e^{-(\delta + \lambda)w}], w \geq 0. \end{aligned} \quad (14)$$

Example 4. Applying discounting for the FRW/PRW policy with T distributed Weibull in Example 2, the expected discounted unit cost from Equation (7) is

$$\begin{aligned} E[X_D(t)] &= \frac{c}{w_2 - w_1} \int_{w_1}^{w_2} \int_0^t \left(\frac{\beta}{\theta}\right) \left(\frac{\tau}{\theta}\right)^{\beta-1} \\ &\quad \times e^{-[\delta\tau + (\tau/\theta)^\beta]} d\tau dt. \end{aligned} \quad (15)$$

These integrals can be solved by standard numerical integration methods.

WARRANTY COST MODELS

Warranty related decisions can be critical for manufacturers since they relate directly to the quality and reliability of their products. Still there is always some level of risk that is inherent in product failures. It is important that product failure characteristics are well understood and known in advance of products being released to the markets. Methods are therefore needed to examine these issues in managing and planning warranty programs. The following models are useful for examining policy and planning resources for warranty expenditures.

Expected Warranty Costs

Consider a product that is sold under a warranty policy ϕ , which could be FRW, PRW, or FRW/PRW. The interarrival times for warranty failures T_i are iid $F(t), t \geq 0$ with failure costs $X_i(t), i = 1, 2, \dots$. The manufacturer's expected warranty cost is then given by

$$C(w, \phi) = E \left(\sum_{i=1}^{N(w)} X_i(t) \right). \quad (16)$$

It follows that $N(w)$, the number of warranty failures occurring in $(0, w)$ is a stopping time therefore,

$$E \left(\sum_{i=1}^{N(w)} X_i(t) \right) = E[N(w)]E[X_i(t)].$$

$E[X_i(t)]$ is the expected unit cost for warranty failures given by Equations (1), (4), and (6) for $\phi = \text{FRW}$, PRW , and FRW/PRW . Thus Equation (16) more generally is

$$C(w, \phi) = M(w)E[X_i(t)], \quad (17)$$

where

$$M(t) = \sum_{n=1}^{\infty} F^{(n)}(t) \quad (18)$$

is the renewal function and $F^{(n)}(t)$ is the n -fold convolution of $F(t)$ with itself.

When failure times T_i are exponentially distributed and hence $N(t)$ is distributed Poisson with rate $\lambda > 0$, the renewal function is simply $M(t) = \lambda t$. This particular case has practical appeal for warranty and reliability related applications, since new products are designed to perform during their useful lives. This however is one of only a few special cases for which $M(t)$ can be derived in closed form. In general, it can be computationally cumbersome to evaluate. Various numerical and approximation methods have been developed to overcome these computational difficulties for analyzing warranties [3–5], as well as other applications [6–8].

Economically Equivalent Policies

It is sometimes advantageous for analysis purposes to translate a PRW policy to the more simple FRW policy.

Proposition 1. *For a nonrenewing PRW policy defined over $I_1 = \{t : t \in (0 < w_1 \leq w_2 < \infty)\}$ there is a cost equivalent policy over $I_0 = \{t : t \in (0 < w < \infty)\}$, where $w_1 < w < w_2$.*

1. *Without discounting,*

$$w = F^{-1} \left(\int_{w_1}^{w_2} \frac{F(t)}{w_2 - w_t} dt \right). \quad (19)$$

2. *With discounting, w is the value satisfying*

$$e^{-\delta w} F(w) + \delta \int_0^w e^{-\delta t} F(t) dt$$

$$= \left(\frac{1}{w_2 - w_1} \right) \int_{w_1}^{w_2} \int_0^t e^{-\delta\tau} f(\tau) d\tau dt. \quad (20)$$

The proofs of (1) and (2) follow by equating the respective expected unit costs given in Equations (1) and (3) for (1), and in Equation (13) for (2).

Example 5. [9]. A machine is produced at a cost to the manufacturer of c dollars per unit and sold with a combined warranty of 6-month FRW followed by a 30-month PRW policy. Warranty failures are distributed exponential with a mean failure time of 30 months. The interest rate over the period is assumed to be 10%.

For T_i distributed exponential with parameter $\lambda > 0$, from Equation (20), it follows that

$$w = \frac{-1}{\lambda} \ln \left[\frac{e^{-\lambda w_1} - e^{-\lambda w_2}}{\lambda(w_2 - w_1)} \right]. \quad (21)$$

Substituting parameter values $\lambda = 0.02$, $w_1 = 6$, and $w_2 = 36$,

$$w = \frac{-1}{0.02} \ln \left[\frac{e^{-(0.02)6} - e^{-(0.02)30}}{0.02(36 - 6)} \right] \\ = 28.7 \text{ months.}$$

For the case of discounting, the right side of Equation (20) simplifies to

$$\left(\frac{1}{w_2 - w_1} \right) \int_{w_1}^{w_2} \lambda e^{-(\lambda+\delta)t} dt \\ = \left(\frac{\lambda}{\lambda + \delta} \right) \left[1 - \left(\frac{e^{-(\lambda+\delta)w_1} - e^{-(\lambda+\delta)w_2}}{(\lambda + \delta)(w_2 - w_1)} \right) \right].$$

Substituting the parameter values with $\delta = 0.1$ and $\lambda + \delta = 0.12$, it follows that

$$\frac{1}{6}(1 - e^{-0.12w}) = 0.145,$$

And $w = 17$ months.

Proposition 2. For a renewing period FRW/PRW policy defined over $I_1 = \{t : t \in (0 < w_1 < w_2 < \infty)\}$ there is a cost equivalent FRW policy over $I_0 = \{t : t \in (0 < w < \infty)\}$ such that

$$w_0 = F^{-1} \left(\frac{\int_{w_1}^{w_2} F(t) dt}{(w_2 - w_1) \bar{F}(w_2) + \int_{w_1}^{w_2} F(t) dt} \right). \quad (22)$$

Proof. For a renewing warranty the period of coverage is determined by the number of successive failures

$$N = (T_1 \leq w) \cap (T_2 \leq w) \cap \dots \cap \\ (T_{n-1} \leq w)(T_n > w),$$

which is distributed geometric with

$$g_N(n) = \bar{F}(w)[F(w)]^{n-1}, n = 1, 2, \dots$$

The mean number of failures in Equation (18) is therefore

$$M(w) = 1/\bar{F}(w), w \geq 0. \quad (23)$$

Substituting the expected unit cost from Equation (7) and the mean number of warranty failures from Equation (23) into Equation (17),

$$C(w, \text{FRW/PRW}) \\ = \frac{c}{\bar{F}(w_2)(w_2 - w_1)} \int_{w_1}^{w_2} F(t) dt. \quad (24)$$

For the FRW policy,

$$C(w, \text{FRW}) = \frac{cF(w_0)}{\bar{F}(w_0)}. \quad (25)$$

Equating these two expected costs and solving for w_0 leads to the result (Eq. (22)).

Example 6. A nonrepairable product is sold with a one year warranty, providing full coverage for the first 3 months with the remaining 9 months covered under a PRW policy. A new warranty is initiated when an item is replaced during the warranty period. Warranty failures are distributed exponential with mean time to failure of 15 months. The producer is considering offering a new warranty that would provide full coverage to customers without increasing cost.

For exponentially distributed failures, in Equation (22)

$$w_0 = F^{-1} \left(\frac{\int_{w_1}^{w_2} (1 - e^{-\lambda t}) dt}{(w_2 - w_1)e^{-\lambda w_2} + \int_{w_1}^{w_2} (1 - e^{-\lambda t}) dt} \right)$$

which upon simplifying

$$\begin{aligned} w_0 &= F^{-1} \left(\frac{(w_2 - w_1) - (1/\lambda)}{(e^{-\lambda w_1} - e^{-\lambda w_2})} \right) \\ &= F^{-1} \left(\frac{(w_2 - w_1)e^{-\lambda w_2} + (w_2 - w_1)}{-(1/\lambda)(e^{-\lambda w_1} - e^{-\lambda w_2})} \right). \end{aligned} \quad (26)$$

Substituting the values $w_1 = 3$, $w_2 = 12$, and $\lambda = 1/15$,

$$\begin{aligned} w_0 &= F^{-1} \left(\frac{(12 - 3) - (15)}{(e^{-3/15} - e^{-12/15})} \right) \\ &= F^{-1}(.3195) \end{aligned}$$

and it follows that the 1-year combined FRW/PRW policy could be replaced by an FRW policy over a $w_0 = 5.77$ month period.

Optimum Policies

The manufacturer's warranty cost for a product will increase over time from warranty failures and increasing failure rates (IFR) from aging. The choice of warranty length will depend on the benefit in providing the warranty. The manufacturer benefit is generally assessed through marketing studies, consumer surveys, and other studies that can be used to quantify benefit. One method is to assume for a given product that the benefit from a warranty will vary in a decreasing manner with the cost for marketing and advertising [9]. Denoting the expected benefit as a function of warranty length by $B(w)$ and the expected warranty cost by $C(w, \phi)$, $w \geq 0$ the expected total variable cost to the manufacturer is given by

$$T(w, \phi) = B(w) + C(w, \phi). \quad (27)$$

$C(w, \phi)$ is an increasing function in w , whereas $B(w)$ will decrease with increasing warranty intervals. The rationale is that the benefit to the manufacturer is realized through lower marketing costs for promoting a product with a short warranty in comparison to one for a longer duration. We assume that $B(w)$ and $C(w, \phi)$ are monotonic and $T(w)$ is convex.

Consider the case of an FRW policy with a quadratic benefit function

$$B(w) = (b_0 - b_1 w)^2, 0 \leq w \leq b_0/b_1. \quad (28)$$

The total expected cost is then

$$T(w) = cM(w) + (b_0 - b_1 w)^2, 0 \leq w \leq b_0/b_1. \quad (29)$$

Setting $\frac{d}{dw} T(w, \phi) = 0$, it follows that the optimum choice of warranty length w^* will satisfy

$$cm(w) + 2b_1^2 - 2b_0b_1 = 0, \quad (30)$$

where $m(w)$ is the renewal density function.

For the case of the warranty failure time T distributed exponential, the number of failures in $(0, w)$ is Poisson distributed, therefore $M(w) = \lambda w$ and from Equation (30), the optimum length for the warranty is given by

$$w^* = \frac{2b_0b_1 - c\lambda}{2b_1^2}. \quad (31)$$

Example 7. A nonrepairable item has a failure time T distributed according to an Erlang distribution with probability density function

$$f(t) = \lambda t e^{-\lambda t}, \lambda > 0, t \geq 0.$$

The mean time to failure is estimated to be 1.5 years. The unit warranty cost to the manufacturer for failures occurring during an FRW period of $(0, w)$ is \$1000. The warranty benefit function is assumed to be linear with the unit cost for marketing the product given by

$$B(w) = \$500 - 100w, w \geq 0.$$

For this case, the total expected cost in Equation (29) is given by

$$T(w) = cM(w) + (b_0 - b_1w), 0 \leq w \leq b_0/b_1. \quad (32)$$

For the Erlang distribution, it can be shown (Ref. 10, p. 170) that

$$M(t) = \frac{\lambda t}{2} - \frac{1}{4} + \frac{1}{4}e^{-2\lambda t}, t \geq 0.$$

Therefore in Equation (32)

$$T(w) = c \left(\frac{\lambda w}{2} - \frac{1}{4} + \frac{1}{4}e^{-2\lambda w} \right) + (b_0 - b_1w), 0 \leq w \leq b_0/b_1. \quad (33)$$

Setting $\frac{d}{dw}T(w) = 0$, leads to the optimum

$$w^* = \frac{1}{2\lambda} \ln(1 - 2b_1/c\lambda). \quad (34)$$

Substituting the parameter values $\lambda = 1/1.5$ failures per year, $c = \$1000$, $b_1 = \$100$ per year, and noting that b_0 does not enter the solution,

$$w^* = - \left[\frac{1}{2(1/1.5)} \right] \ln[1 - 2(100)/(500(1/1.5))] \\ = 0.267 \text{ years} = 3.2 \text{ months.}$$

Other approaches have been applied in exploring optimum pricing, warranty length, and marketing strategies under various assumed demand dependence on price and warranty [10], dynamic conditions [11], and marketing behaviors [12–14]. For more details on warranty management and economic models, see the reviews in Refs 2 and 15.

Allocating Warranty Reserve Funds

Warranty reserve is the collection of funds that are set aside to cover future expenses for warranty claims. The amount and process of withholding is important to the manufacturer because if insufficient funds are available for covering the expenses, profit shares will be affected and if too much is withheld,

then there will be lost opportunity for the company to make investments. The following method initially proposed by Menke [16] and later generalized by others [17,18] consists of creating a fund through the accumulation of a fixed proportion of unit sales.

Consider a product that is produced and sold in lot sizes of K units at a price of p dollars per unit under a PRW policy over $(0, w)$. The failure time probability density function, $g(t)$, $t \geq 0$ is assumed to be known. Denoting the rate of return and expected price change per period by ρ and π , the differential for the amount of reserve is given by

$$d(R) = X(t)d[K \cdot F(t)]. \quad (35)$$

The unit price p is the cost plus a constant amount of revenue, therefore, $X(t)$ for the PRW policy is given in Equation (2) and hence

$$d(R) = p(1 - t/w)(1 + \rho)^{-t}(1 + \pi)^{-t}K \cdot f(t)dt.$$

Noting that since

$$(1 + \rho)^{-t}(1 + \pi)^{-t} \rightarrow (1 + \rho + \pi)^{-t} = b^{-t},$$

it follows that

$$R(w) = \left(\frac{Kp}{w} \right) \int_0^w (w - t)b^{-t}f(t)dt. \quad (36)$$

Example 8. An item that is sold with a 1-year PRW policy has failure times T distributed exponential with a mean failure rate of 0.05 per year. Interest is assumed to be constant at 3% per period and price changes at a rate of 4% per period.

For T distributed exponential in Equation (36), it follows that the amount of reserve funds per unit of sale is given by

$$R(w) = \frac{Kp\lambda}{\lambda + \ln b} \left[1 - \frac{1 - b^{-w}e^{-\lambda w}}{w(\lambda + \ln b)} \right]. \quad (37)$$

With $\lambda = 0.05$ failures per year, $w = 1$ year, $\pi = 0.04$ and $\rho = 0.03$, $b = 1.07$, and

$$R(w) = \frac{Kp(.05)}{05 + \ln(1.07)} \left[\frac{1 - (1.07)^{-1}e^{-.25(1)}}{1(.05 + \ln(1.07))} \right] \\ = 0.024.$$

Therefore $R(1)/Kp = 0.024$, which indicates that 2.4% of the sales price should be held in a fund to cover the warranty expenses.

The process of collecting and managing reserve resources can be quite contentious, particularly when the consequences of not having sufficient funds to fulfill commitments are devastating to the organization. The challenge is in having accurate estimates of the amount needed in the fund. This method assumes that the distribution of failure times is known. Another approach is to develop the long run distribution of required reserves when warranty claims are assumed to follow a compound Poisson process [19]. These models are based on building a fund through accruals from sales that will be available to cover forthcoming warranty costs. Still, there is the problem of managing the flow of claims over time. Even adequately planned reserve collections can have significant shortfalls at various times due to the dynamics of the claims flow. One approach for overcoming this issue is to forecast the reserve fund requirement [20,21]. Buczkowski and Kulka-rni [22] developed a control policy based on allocating a fixed level of funds initially at the beginning of a period along with a fixed amount of accrual from each sale to ensure a positive reserve at a prescribed probability level. Ja *et al.* [23] derived results for developing optimal warranty length and reserve accruals for a minimal repair policy over $(0, w)$, to ensure the adequacy of reserves for covering the costs. Another interesting approach is to model and track the total warranty exposure time and total number of items under warranty over time [24].

WARRANTY AND MAINTENANCE DECISIONS

While active warranties are liabilities for producers, they are assets to customers. The models presented in this article so far have dealt with economic decisions for the producer or manufacturer of the products. We will continue to address the position of the manufacturer, only now the manufacturer is a user of the products under warranty.

We will now discuss models for maintenance planning decisions for items under warranty. The purpose of maintenance planning is to provide policy for maintaining operationally ready equipment that will ensure reliable and efficient service. The inclusion of warranties in these analyses will alter the allocation of costs between the producer and customer or manufacturer in this case, which can have a significant impact on the replacement decisions.

In its simplest form, the machine repair and replacement problem is to determine a policy for maintaining and replacing machines that are subject to deterioration over time. There is an extensive literature on maintenance and replacement strategies that dates back to the pioneering work of Barlow and Hunter [25]. However, it is only in recent years that warranty consideration has been included in these economic decision models. Most of the replacement and maintenance planning models are renewal theoretic based cost minimization models. The decision criterion for fixed-period non-renewing warranties is to minimize the expected total cost over the warranty period. For renewing warranties, the cost objective function is structured as a renewal reward process where the renewal cycle time is the random warranty period. The criterion is to minimize the long run average cost rate, estimated by the ratio of expected costs to expected cycle time.

Repair versus Replacement

Among the earliest reported results for modeling replacement decisions for machines under warranty is the work of Nguyen and Murthy [26]. They considered the problem of finding the optimum replacement time for a machine using the maintenance strategy of replacing the machine when failure occurs in an interval $(0, w - \alpha]$, and provide minimal repair maintenance when failure occurs in $(w - \alpha, w]$, where $0 \leq \alpha \leq w$. They showed that treating the failures as renewal events, the mean number of repairs in $(w - \alpha, w]$ is $M(w - \alpha)$ and the time to first failure beyond $(w - \alpha)$ is the excess life of the machine in use at time $(w - \alpha)$ which is distributed

$$F_\gamma(x) = F(w - \alpha + x) - \int_0^{w-\alpha} \bar{F}(w - \alpha + x - y) dM(y), x \geq 0. \quad (38)$$

The expected number of repairs during the minimal repair segment, $(w - \alpha, w]$ is given by integrating the failure rate function

$$\int_0^\alpha r_\gamma(x) dx = \int_0^\alpha \left(\frac{f_\gamma(x)}{\bar{F}_\gamma(x)} \right) dx = -\ln(\bar{F}_\gamma(x)).$$

Denoting the unit replacement and repair costs by c_s and c_r , the expected warranty cost is given by

$$C(\alpha, w) = c_s M(w - \alpha) + c_r \ln[1 - \bar{F}_\gamma(\alpha)]. \quad (39)$$

Setting $\frac{d}{d\alpha} C(\alpha, w) = 0$, the value α^* , providing the minimum expected warranty cost satisfies

$$\frac{\bar{F}(\alpha^*)}{\bar{F}_\gamma(\alpha^*)} = \frac{c_s}{c_r}, 0 \leq \alpha^* \leq w. \quad (40)$$

It is easily shown that for the case of exponentially distributed failure times, $\bar{F}(\alpha^*) = \bar{F}_\gamma(\alpha^*)$ in Equation (40) and hence $c_s/c_r = 1$. Computations are generally cumbersome but can be accomplished using numerical methods. An example for the case of Erlang distributed failures can be found in (Ref. 27, p. 391).

Age-Replacement for Renewing Warranties

Age-replacement is a strategy applied to units that are considered nonrepairable. By this strategy, an item is replaced when it fails at a time T or when it reaches a fixed time τ , whichever occurs first. Suppose the item is a piece of equipment that is not repairable, but is sold with a renewing FRW policy over $(0, w]$. Denoting the replacement cost by $c_s > 0$ and the cost of downtime during a failure by $c_d > 0$, for the case of $\tau > w$, the unit cost is given by

$$X(t) = \begin{cases} c_d, & t \leq w \\ c_s + c_d, & w < t < \tau \\ c_s, & t \geq \tau \end{cases} \quad (41)$$

So the owner of this equipment would never replace it as long as it is under warranty. With the further assumption that the failure times and related cost pairs, $[T_i, X(T_i)], i \geq 1$ are iid, this cost structure forms a renewal reward process (Ref. 28, p. 77). The optimal replacement time is the minimum long-term average cost, which is the ratio of the expected costs incurred during a cycle and the expected replacement cycle time. Since failure times are renewal events, the replacement cycles are also iid and of expected length $\int_0^\tau \bar{F}(u) du$. It follows that with renewing warranties the expected cost during a cycle is given by

$$C(\tau, w) = E[X(t)] = c_d F(w) + (c_s + c_d) \int_w^\tau f(u) du + c_s [1 - F(\tau)],$$

which simplifies to

$$C(\tau, w) = c_s \bar{F}(w) + c_d F(\tau). \quad (42)$$

The long run average cost rate is then

$$CR(\tau) = \frac{c_s \bar{F}(w) + c_d F(\tau)}{\int_0^\tau \bar{F}(u) du}, 0 < w \leq \tau. \quad (43)$$

To determine the optimum replacement time τ^* from this function, we first note that as $\tau \rightarrow 0$, $CR(\tau) \rightarrow \infty$ and as $\tau \rightarrow \infty$

$$CR(\infty) \rightarrow \frac{c_s \bar{F}(w) + c_d}{\mu}, \quad (44)$$

where μ is the expected product lifetime. It further follows from $\frac{d}{d\tau} CR(\tau) = 0$ in Equation (43) that for an IFR failure rate function $r(t)$, at the optimum replacement time $\tau^* > w$, $CR(\tau^*) = c_d r(\tau^*)$.

Example 9. A particular piece of nonrepairable equipment is purchased at a cost of \$5000 with $w = 1$ -year FRW policy. Failures that occur within the warranty period cost the owner \$1000 due to downtime, while the seller replaces the item along with a renewing warranty. The failure times are Weibull distributed with

$$F(t) = 1 - e^{-t^2}, t \geq 0.$$

For the Weibull distribution with characteristic value 1 and slope of 2, the mean lifetime $\mu = \sqrt{\pi}/2$. Substituting the parameter values with $c_s = \$5000$, $c_d = \$1000$, and $\bar{F}(1) = 0.3679$ into Equation (43), the long run average cost rate is

$$CR(\tau) = \frac{1864 + 1000(1 - e^{-\tau^2})}{\int_0^\tau e^{-u^2} du}, 1 < \tau < \infty, \quad (45)$$

and

$$c_d r(\tau) = 2,000\tau, 1 < \tau < \infty. \quad (46)$$

It follows that $CR(2.2) = 4.389$ and $1000r(2.2) = 4400$, therefore $\tau^* = 2.2$.

POSTWARRANTY MAINTENANCE

Service organizations such as job shops have to ensure that they have operational-ready equipment available when they need it for orders. During warranty periods, the reliability of production equipment is at its highest and failure costs are largely covered by the seller; therefore, equipment is generally not replaced until expiration of the coverage. It is of course possible that the relative costs involved in the previously discussed optimum replacement strategies are such that the optimum replacements are at an age $\tau < w$ [29]. In either case the proprietor or job shop owner faces a postwarranty decision on the level of maintenance necessary to keep the equipment operational and cost effective. One option is to avoid or delay implementing a maintenance program by extending warranties. These decisions are much the same as “own” versus “lease” decisions and optimality conditions for given cost and failure characteristics have been developed for examining these options [30]. Maintenance program options can vary from minimal repair and replacement strategies [31] to extensive preventive maintenance (PM) programs for extending the lifetime of the equipment. Like age-replacement, maintenance planning models are renewal theory based with the cost and failure events treated as renewal rewards.

Jung and Park [32] developed conditions for optimum periodic PM following warranty coverage. Consider a machine under an FRW renewing policy of length w with failure time T distributed $F(t)$ and failure rate $r(t)$, that is, IFR. Starting from the expiration of the warranty, the cycle length is $T < w$ if failure and replacement occurs during the initial warranty period. If the item survives to w , $T > w$ then the cycle is extended by a fixed number of periodic PM actions of length τ . The expected cycle time for an item replaced at PM cycle time N is then

$$\begin{aligned} E[T(\tau, N)] &= E(T|T < w)F(w) \\ &\quad + E(w + N\tau|T > w)\bar{F}(w) \\ &= \int_0^w tf(t)dt + (w + N\tau)\bar{F}(w), \end{aligned} \quad (47)$$

During $(0, w)$, there are no costs to the user but for $T > w$ the costs for minimum repair, PM, and replacement are incurred. From earlier studies [28], the expected cost for minimum repairs between scheduled PM actions is given by

$$\begin{aligned} E(C_m) &= c_m \bar{F}(w) \left\{ \frac{N(N-1)}{2} \tau [r(w + \tau) - r(w)] \right. \\ &\quad \left. + N \int_w^{w+\tau} r(t)dt \right\}, \end{aligned}$$

where c_m is the unit cost of PM and the expected end of cycle replacement cost is $c_r \bar{F}(w)$. The long run average maintenance cost rate is then given by

$$\begin{aligned} CR(\tau, N) &= \left[\begin{aligned} &c_{PM}(N-1)\bar{F}(w) \\ &+ c_r \bar{F}(w) + c_m \bar{F}(w) \\ &\left\{ \frac{N(N-1)}{2} \tau [r(w + \tau) - r(w)] \right. \\ &\quad \left. + N \int_w^{w+\tau} r(t)dt \right\} \end{aligned} \right] \\ &\quad \times \left[\int_0^w tf(t)dt + (w + N\tau)\bar{F}(w) \right]^{-1}, \end{aligned} \quad (48)$$

where c_{PM} is the unit cost for PM. The authors show that for $r(t)$ strictly increasing and convex the optimum period length and number

of PM cycles (τ^* , N^*) can be derived. Further details on PM under various warranty conditions can be found in Ref. 33.

SUMMARY REMARKS

1. The models presented in this article deal with management and operations issues that involve product warranties. The perspective taken for the cost models is that of the producer or the manufacturer. The maintenance planning models however reflect the perspective of the customer, which is often also the manufacturer or producer. Much of the literature on warranty models includes consumer objectives as well.
2. Warranty failures and related costs are directly related to the overall effectiveness in producing products and services relative to quality and customer expectations. Consequently, important feedback information can be derived from warranty programs for planning quality improvement and serving as a management control element for the manufacturing system. Warranty information is used as a relative indicator of cumulative costs and areas for identifying reliability improvements options. There is need for more quantitative methods for developing performance tracking and improvement strategies [34].
3. This broader view of treating warranty as an integral element of a product system suggests several directions for further developments. One area of need is to better understand the classes of customer behavior effects including risk aversion to warranty options. Very little work has been done in characterizing reactions to warranty options [20]. Postwarranty options in particular are risk options that can provide mutual opportunities for both the customer and the manufacturer through shared maintenance responsibilities.
4. While the focus in this article has been on product warranties, the models

and methods presented are equally applicable to service warranties. Like products, service warranties are binding commitments by providers to ensure acceptable quality of their service for some prescribed period of time following delivery. With growing customer demands for more reliable and higher quality products and services, service warranties are likely to receive much greater attention in the future. Quantitative methods are needed for examining economic decisions for this class of warranties. The issues are essentially the same, that is, the cost, length of warranty coverage, compensation, and reserve funds. However, the analogy to failure characteristics is less likely to apply. There is need for a framework for service warranty models.

NOTATION

FRW	Free replacement warranty policy
IFR	Increasing failure rate
PM	Preventive maintenance
PRW	Prorata replacement warranty policy
$B(w)$	Expected cost-benefit function over $(0, w)$
b_0, b_1	Parameters for warranty benefit function
c, c_1	Unit warranty cost to the manufacturer
c_s, c_d, c_r, c_{PM}	Unit costs for replacement, downtime, repair, and PM
$C(w, \phi)$	Expected warranty cost to the manufacturer
$CR(\tau)$	Long-run expected cost rate
δ	Discount factor
β, θ	Weibull distribution parameters: slope and characteristic value
$F(t), f(t)$	Probability distribution and density functions for failure time T
$\bar{F}(t)$	$= 1 - F(t)$, survivor function
λ	Parameter for exponential distribution

$M(t), m(t)$	Renewal function and renewal density function
μ	Mean product lifetime
$N(t)$	Number of failures in $(0, t]$
P	Rate of return
π	Price change per period
$r(t)$	$= f(t)/\bar{F}(t)$, failure rate function
$R(w)$	Reserve resources for warranty of length $w > 0$
T	Non-negative failure time random variable
τ	Replacement time
$X_i(t)$	Unit warranty cost of failure i
$X_D(t)$	Unit warranty cost with discounting

ENDNOTE

The views in this article are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government.

REFERENCES

- Blischke WR, Murthy DNP. Product warranty handbook. New York: Marcel Dekker; 1997.
- Thomas MU, Rao SR. Warranty economic decisions: a summary and some suggested directions for future research. *Oper Res* 1999;47(6):807–820.
- Blischke WR, Scheuer EM. Application of renewal theory in analysis of the free-replacement warranty. *Nav Res Logist Q* 1981;28:193–205.
- Frees EW. Warranty analysis and renewal function estimation. *Nav Res Logist Q* 1986;33:361–372.
- Murthy DNP, Blischke WR. Product warranty management—III: A review of mathematical models. *Eur J Oper Res* 1992;62:1–34.
- Cui L, Xie M. Some normal approximations for renewal function of large Weibull shape parameter. *Commun Stat Part B Simul Comput* 2003;32(1):1–16.
- Tortorella M. Numerical solution of renewal-type integral equations. *INFORMS J Comput* 2005;17:66–74.
- Jiang R. A Gamma-normal series truncation approximation for computing the Weibull renewal function. *Reliab Eng Syst Saf* 2008;93(4):616–626.
- Thomas MU. Optimum warranty policies for irreparable items. *IEEE Trans Reliab* 1983;R-32(3):282–288.
- Glickman TS, Berger PD. Optimal price and protection period decisions for a product under warranty. *Manage Sci* 1976;22(12):1381–1390.
- Zhou Z, Yanjun L, Tang K. Dynamic pricing and warranty policies for products with fixed lifetime. *Eur J Oper Res* 2009;196(3):940–948.
- Menezes MAJ, Currim IS. An approach for determination of warranty length. *Int J Market* 1992;9:177–195.
- DeCroix GA. Optimal warranties, reliabilities and prices for durable goods in an oligopoly. *Eur J Oper Res* 1999;112:554–569.
- Wu CC, Lin PC, Chou CY. Determination of price and warranty length for a normal lifetime distributed product. *Int J Prod Econ* 2006;102:95–107.
- Thomas MU. Reliability and warranties: methods for product development and quality improvement. Boca Raton (FL): Taylor & Francis; 2006.
- Menke WW. Determination of warranty reserves. *Manage Sci* 1969;15:B542–B549.
- Amato AE, Anderson EE. Determination of warranty reserves: an extension. *Manage Sci* 1976;22:1391–1394.
- Thomas MU. A prediction model for manufacturer warranty reserves. *Manage Sci* 1989;35(12):1515–1519.
- Tapiero CS, Posner MJ. Warranty reserving. *Nav Res Logist* 1988;35(4):473–479.
- Singpurwalla ND, Wilson S. The warranty problem: Its statistical and game theoretic aspects. *SIAM Rev* 1993;35(1):17–42.
- Eliashberg J, Singpurwalla ND, Wilson SP. Calculating the reserve for a time and usage indexed warranty. *Manage Sci* 1997;43(7):966–975.
- Ja S, Kulkarni VG, Patankar JG. A renewable minimal repair warranty policy with time-dependent cost. *IEEE Trans Reliab* 2001;50:346–352.
- Buczkowski PS, Kulkarni VG. Funding a warranty reserve with contributions after each sale. *Probab Eng Info Sci* 2006;20:497–516.
- Wortman MA, Elkins DA. Stochastic modeling for computational warranty analysis. *Nav Res Logist* 2005;52(3):224–231.

25. Barlow RE, Hunter L. Optimum preventive maintenance policies. *Oper Res* 1960;8:90–100.
26. Nguyen DG, Murthy DNP. Optimum replace-repair strategy for servicing products sold with warranty. *Eur J Oper Res* 1993;39:306–312.
27. Blischke WR, Murthy DNP. Warranty cost analysis. New York: Marcel Dekker; 1993.
28. Ross SM. Stochastic processes. 2nd ed. New York: John Wiley & Sons, Inc.; 1996.
29. Yeh RH, Chen GC, Chen MY. Optimal age-replacement policy for nonrepairable products under renewing free-replacement warranty. *IEEE Trans Reliab* 2005;54:92–97.
30. Lam Y, Kwok P, Lam W. An extended warranty policy with options open to consumers. *Eur J Oper Res* 2001;131:514–529.
31. Sahin I, Polatoglu H. Maintenance strategies following the expiration of warranty. *IEEE Trans Reliab* 1996;45(2):220–228.
32. Jung GM, Park DH. Optimum maintenance policies during the post-warranty period. *Reliab Eng Syst Saf* 2003;82:173–185.
33. Boland PJ. Periodic replacement when minimal repair costs vary with time. *Nav Res Logist* 1982;29:541–546.
34. Thomas MU, Richard JP. Warranty based method for establishing reliability improvement targets. *IIE Trans* 2006;38(12):1049–1058.
- Chun YH. Optimal number of periodic preventive maintenance operations under warranty. *Reliab Eng Syst Saf* 1992;37:223–225.
- Djamaludin I, Murthy DNP. Warranty and preventive maintenance. *Int J Reliab Qual Saf Eng* 2001;8:89–107.
- Djamaludin I, Murthy DNP, Wilson RJ. Quality control through lot sizing for items sold with warranty. *Int J Prod Econ* 1994;33:99–107.
- Erke Y. Conditions for optimal opportunistic screening and warranties [PhD dissertation]. Lehigh University; 1997.
- Iskandar BP, Sandoh H. An opportunity based age-replacement policy considering warranty. *Int J Reliab* 1999;6:229–236.
- Jack N, Dagpunar JS. An optimal imperfect maintenance operations under warranty. *Microelectron Reliab* 1994;34:529–534.
- Kim CS, Djamaludin I, Murthy DNP. Warranty and discrete preventive maintenance. *Reliab Eng Syst Saf* 2004;84(3):301–309.
- Lie CH, Chun YH. Optimum single-sample inspection plans for products sold under free and rebate warranty. *IEEE Trans Reliab* 1987;R-36:634–647.
- Mitra A, Patankar JG. Market share and warranty cost for renewable warranty programs. *Int J Prod Econ* 1997;50(2):155–168.
- Sahin I, Polatoglu H. Quality, warranty and maintenance. Boston (MA): Kluwer; 1998.
- Yeh RH, Chen GC, Chen MY. Optimal age-replacement policy for nonrepairable products under renewing free-replacement warranty. *IEEE Trans Reliab* 2005;54(1):92–97.
- Yeh RH, Lo HC. Optimal preventive-maintenance warranty policy for repairable products. *Eur J Oper Res* 2001;134:59–69.
- Zuo MJ, Liu B, Murthy DNP. Replacement-repair policy for multi-state deteriorating products under warranty. *Eur J Oper Res* 2000;123:519–530.

FURTHER READING

- Blischke WR, Murthy DNP. Warranty cost analysis. New York: Marcel Dekker; 1993.
- Brennan JR. Warranties: planning, analysis, and implementation. New York: McGraw-Hill; 1994.
- Chen J, Yao DD, Zheng S. Quality control for products supplied with warranty. *Oper Res* 1999;46:107–115.

WHY RISK IS NOT VARIANCE

LOUIS ANTHONY COX, JR.
Cox Associates and Department of
Mathematical and Statistical
Sciences, University of Colorado,
Denver, Colorado

WILLIAM A. HUBER
Quantitative Decisions, Rosemont,
Pennsylvania

FRAMEWORK AND RESULTS

The following two plausible principles for managing financial investment risks are mutually inconsistent, in general

1. *Rule 1: Make Dominating Choices.* Other things being held equal, given a choice between a smaller probability of gain and a larger probability of gain, a decision maker should always choose the larger probability of gain. For example, given a choice between winning \$100 with probability 0.1 and winning \$100 with probability 0.2, rational decision makers who prefer more dollars to less should choose the option that gives a 0.2 probability of winning the \$100.
2. *Rule 2: Seek Mean–Variance Efficiency (Higher Variance Requires Higher Mean Return).* Given a choice among risky prospects, an investor should require more expected return to accept a prospect with more variance than to accept a prospect with less variance. For example, a 0.2 chance of winning \$100 (else nothing) has a higher variance than a 0.1 chance of winning \$100, but it also has a higher mean.

Rule 1 is implied by the decision-analytic principle of first-order stochastic dominance [1]: Prospects that give higher probabilities of preferred outcomes (and lower probabilities of less preferred outcomes) should be

preferred. Rule 2 provides the basis for many current efficient portfolio and mathematical optimization (e.g., quadratic programming) approaches to optimal investment (http://en.wikipedia.org/wiki/Modern_portfolio_theor). Although theorists have noted that some risk-averse decision makers may prefer some mean-preserving increases in variance (*ibid*), the idea that volatility in returns, as measured by variance or standard deviation, is generally undesirable to risk-averse investors, and that it should be avoided or compensated by higher expected returns, is still widely taught and practiced. However, *Rules 1 and 2 are incompatible*, in general, as demonstrated next. We introduce this incompatibility by first assuming that decision makers have indifference curves for mean–variance combinations, and then discuss why it holds even if no indifference curves exist.

Following the literature on mean–variance decision making, suppose that a decision maker has positively sloped continuous indifference curves in mean–variance space [2]. To any mean–variance pair (m, v) (a point in the mean–variance space), there corresponds a *certainty equivalent*, namely the point at which the indifference curve through (m, v) reaches the horizontal (mean) axis. The indifference curve through the origin $(0, 0)$ separates *acceptable risks* (those with positive certainty equivalents, lying below and to the right of the curve, if return is desirable) from *unacceptable risks* (those with negative certainty equivalents, lying above and to the left of it). To make an unacceptable risk acceptable in this framework, one must increase its mean return or reduce its variance.

The following result holds if we assume that there is at least one mean–variance combination (m, v) that the decision maker considers to be unacceptable (i.e., above the indifference curve through the origin, if such indifference curves exist).

Theorem. *If a decision maker considers all prospects corresponding to a particular mean–variance pair (m, v) to be*

unacceptable, then the decision maker considers unacceptable some prospects with positive expected values and no possibility of loss.

Proof. Huber [3] notes that one can construct explicit return distributions (e.g., with log-normally distributed returns) that have any specified positive (m, v) parameter values, but that have positive support, and hence no possibility of loss.

Such a decision maker prefers the *status quo* or “nothing ventured, nothing gained” point $(0, 0)$ to the possibility of winning a positive amount without any possibility of a loss, violating Rule 1. In this sense, Rules 1 and 2 are incompatible.

This incompatibility result does not require that a decision maker have well-defined indifference curves for mean–variance combinations, even though this is a common assumption in mean–variance risk models. Instead, the theorem only requires that there be at least one mean–variance combination, the (m, v) pair in the statement of the theorem, that the decision maker finds unacceptable.

Older, less general, results make a similar point by noting conditions under which binomial lotteries (yielding a known positive return with probability p , else nothing) can be used to construct parabolas in mean variance space, as p ranges from 0 to 1 (with $v = 0$ at both end points and rising to a maximum at $p = 0.5$) that intersect upward-sloping indifference curves twice, once for the ascending (positively sloped) portion of the parabola and once for its descending (negatively sloped) portion [4]. In any such construction, the rightmost intersection represents a stochastically dominant prospect (higher success probability p , which should be preferred, by Rule 1) compared to the leftmost intersection. That both points lie on the same indifference curve violates Rule 1.

DISCUSSION

Students of finance are often taught that “risk” is measured by the variance or

standard deviation of returns around an expected value. On the other hand, students of health, safety, and environmental (HS&E) and reliability engineering risk analysis are often taught that “risk” is determined by the probabilities of different consequences. Simple constructions show why the second approach (considering different specific consequences and their probabilities) can be preferable to considering only means and variances.

The finding that variance is problematic as a measure of risk is not new. The most common critique in the theoretical decision analysis and financial economics literature is that mean–variance analysis is compatible with von Neumann–Morgenstern expected utility theory only under very restrictive conditions (e.g., if all risky prospects have normal or location-scale distributions and utility functions are quadratic, implying that less money is preferred to more, for some amounts) [5,6]. Mean–variance dominance and stochastic dominance relations for location-scale distributions do not coincide in general [2]. Indeed, expected utility theory is inconsistent with all possible moment-based preference models (in which preferences are determined by mean, variance, skewness, kurtosis, etc.) for many utility functions [7]. Variance is also inconsistent with proposed normative axioms for coherent financial risk measures (nonnegativity, homogeneity, and subadditivity, which together imply that deterministic outcomes have zero risk; and shift-invariance, which implies that adding a constant to a random variable does not change its risk) [8]. Empirical studies since the 1960s have demonstrated that real decision makers pay attention to more than mean and variance in their choices among risky prospects [9], sparking a search for alternative measures of risk that would be consistent with stochastic dominance under usefully general conditions [10–14].

Thus, the above theorem is consistent with a long line of research. However, this recent version demonstrates a conflict between Rules 1 and 2, making only minimal assumptions (in particular, not requiring the framework of von Neumann–Morgenstern

expected utility theory or other sets of normative axioms for risk measures) and using only elementary mathematics. It may therefore be useful in illustrating why knowing the variances and expected returns from alternative investment choices (or other actions) does not necessarily suffice to identify the choice with the most desirable probability distribution of consequences.

REFERENCES

1. Sheldon A, Sproule R. Variance as a proxy for risk: the case of the binomial distribution. *Statistician* 1997;46(3):317–322.
2. Wong W-K. Stochastic dominance theory for location-scale family. *J Appl Math Decis Sci* 2006;1–19. Available at <http://downloads.hindawi.com/journals/ads/2006/082049.pdf>.
3. Huber WA. Why risk is not variance: an expository note. *Risk Anal* 2010;30(3):325–326.
4. Borch KH. A note on uncertainty and indifference curves. *Rev Econ Stud* 1969;36(1):1–14.
5. Markowitz HM. Portfolio selection: efficient diversification of investments. New York: John Wiley & Sons, Inc.; 1959.
6. Baron DP. On the utility-theoretic foundation of mean-variance analysis. *J Finance* 1977;32(5):1683–1697.
7. Brockett PL, Kahane Y. Risk, return, skewness and preference. *Manage Sci* 1992;38(6):851–866.
8. Pedersen CS, Satchell SE. Choosing the right risk measure: a survey. 1998. Available at <http://citeseer.ist.psu.edu/520923.html>. Accessed in 2010.
9. Jia J, Dyer JS, Butler JC. Measures of perceived risk *Manage Sci* 1999;45(4):519–532. Available at http://fisher.osu.edu/~butler_267/DAPapers/WP970005.pdf.
10. Ogryczak W, Ruszczyński A. On consistency of stochastic dominance and mean–semi-deviation models. *Math Program* 2001;89:217–232.
11. Föllmer H, Schied A. Convex measures of risk and trading constraints. *Finance Stochast* 2002;6(4):429–447.
12. Artzner P, Delbaen F, Eber J-M, Heath David. Coherent measures of risk. *Math Finance* 1999;9(3):203–228.
13. Rockafellar RT, Uryasev S, Zabarankin M. Generalized deviations in risk analysis. *Finance Stochast* 2006;10:51–74.
14. Ruszczyński A, Shapiro A. Conditional risk mappings. *Math Oper Res* 2006;31:544–561.

WHY TRADITIONAL KANBAN CALCULATIONS FAIL IN VOLATILE ENVIRONMENTS

COLIN KESSINGER
ALLAN GRAY
End-to-End Analytics,
Palo Alto, California

INTRODUCTION

The goal of this article is to familiarize the reader with some of the tools, techniques, and processes commonly associated with the implementation of kanbans or reorder points (ROPs), and to discuss some common pitfalls when lead times are long and demand is volatile and/or seasonal.

The use of kanbans is closely associated with the Toyota Production System and Lean Manufacturing. Without recounting the entire history of the kanban concept, it is important to recognize several aspects of the environment in which it was developed. In particular, kanbans served as the primary mechanism for pulling material between manufacturing stages in factories, where the master production schedule was fixed with minimal variation, and where the lead times between stages were measured in minutes or hours instead of weeks. Additionally, the primary purpose of kanbans was to govern a smooth flow through the factory and manage process variability.

In this environment, there was little or no need to statistically size the kanbans. The process was driven by a continuous improvement cycle in which Toyota reduced inventory levels each time it identified and eliminated a source of process variation. By design, there was no inventory within the factory to manage demand variability. Most of this variability was managed through the substantial finished goods inventory sitting on dealer lots.

In recent years, however, the kanban concept has been applied to govern replenishment not just between manufacturing cells in a single factory, but between factories, and between firms. We can see references to this both in the academic literature [1] and amongst practitioners who work in this space [2]. Indeed, in high-tech supply chains, possibly encouraged by the rise of the “Lean” movement, many managers appear to have abandoned the use of the term “reorder point” and now think exclusively in terms of kanbans.

This expansion of the scope of “kanban thinking” to cover the entire supply chain carries with it various hidden perils, because kanbans were developed in settings quite unlike many of the environments where kanbans are being implemented today. First, lead times are often much longer. Consider the typical high-tech supply chain which almost certainly includes a variety of semiconductors (processors, ASICs, FPGAs, passives, memory), power supplies, and LCDs, all of which can have lead times between 8 and 26 weeks. Second, demand volatility and variability are also considerably higher. For example, companies that manufacture consumer products face increasingly short product life cycles and are subject to seasonal demand peaks (back-to-school, Christmas, etc.) Companies that sell capital equipment or network infrastructure are also subject to these seasonal effects, but also see significant volatility within each quarter (“hockey-stick” effect) and the tremendous variability that generally accompanies “high mix, low volume” products.

To summarize, we have identified two key areas of difference between the preeminent implementation of kanbans and the environments that many of us encounter in our own supply chains. The key contrasts are shorter lead times and less variability/volatility versus longer lead times and higher demand volatility/variability. In recognition of these differences, companies have experimented with a variety of “fixes” to setting and managing kanban levels and, in our

experience, have consistently come up short. The remainder of the article will examine and address why this has been the case.

THE ORIGIN OF THE SHORTCOMINGS

There are three primary root-causes of the poor performance that we see. First and foremost, “out-of-the-box” implementations of the “standard” inventory formulas are a very poor fit for the environment in which we are interested. Second, the typical personnel responsible for analyzing the data and implementing and managing inventory programs do not have the necessary training to do so. Third, very few companies have designed a solution where they look at their entire supply chain as a system of *flows* rather than a sequence of isolated *piles* of inventory. In this paper we will focus on the first of these three shortcomings, but in no way are we suggesting that the other two are less important.

The first question that we should ask ourselves is why the standard formulas do not work. The academic community has been researching inventory theory for over 90 years and yet only a small fraction of their findings have made it into practice. We can attribute this to two factors. First, the majority of academic research tends to focus on stylized models meant to generate managerial insight. From this research, we can quantify the impact of long lead times and demand variability, and we learned about the benefits of risk-pooling and postponement, all of which are significant contributions. However, much of this research makes a series of strong assumptions which greatly facilitate our ability to extract these insights, but undermine efforts to use it in real-world situations.

The strongest of these assumptions is that the key parameters are “stationary”. In layman’s terms, this means that the parameters are constant. For demand, this implies that the run-rate (average) and the variability (standard deviation) remain constant over time. This also implies that the key cost variables remain constant over time. In such a world, forecasting is irrelevant and it is fairly straightforward to compute the

average and standard deviation of historical demand and apply the classic “Newsvendor” inventory formula. The textbook by Nahmias [3], *Production and Operations Analysis*, provides an excellent practical introduction to the Newsvendor Model and a list of useful references for readers who would like to explore more advanced models.

Unfortunately, whether working with baked goods, premium cosmetics, PC’s, routers, capital equipment, or bottled water, we simply have not found stationarity to be a very good assumption.

Aware of this deficiency, academics have responded by developing a substantial body of nonstationary models. Without providing a comprehensive review of this work, but pointing to several good examples, we can see that this work started as far back in 1962 with Iglehart and Karlin [4] looking at nonstationary demand, remained a topic of ongoing interest in the 90s with Song and Zipkin [5] looking at fluctuating demand patterns, and in the 2000s with Dong and Lee [6] examining time-correlated demand (with the added bonus of doing this in a multi-echelon setting). However, as Gavirneni and Tayur [7] note, these models “are either very complex and computationally expensive or specific to the setting they address and not easily generalized”. Despite their efforts to overcome these obstacles, the average practitioner would still find their methodology quite daunting and the research since their publication in 1993 has not overcome these obstacles. As a result, companies have to make a number of adaptations in their methodologies and processes in order to apply any kind of scientific inventory management. In the remainder of this article we will review many of the issues associated with these efforts.

CHOOSING THE RIGHT DEMAND SIGNAL TO DRIVE KANBANS

Any inventory improvement initiative should start with a characterization and quantification of demand and the uncertainty associated with it. After all, the primary role of inventory is to address this uncertainty.

This issue quickly becomes cloudy when implementing kanbans since there is considerable debate as to whether the model should be consumption-driven or forecast-driven. Purists often argue that kanbans are “pull” models and should be driven by consumption. We would argue that the execution (replenishment triggers) of the pull models should be driven by consumption but that the planning (kanban parameter setting) may need to be driven by forecasts. In either case, most companies have implemented both methodologies inadequately.

Consumption-Based ROP Calculations

The majority of textbook inventory models require the computation of a mean (or average) and standard deviation of demand over the lead time. The typical approach is to choose some standard period length (typically days or weeks), calculate the mean and standard deviation of demand of these observations and then scale by the square root of the lead time to calculate the standard deviation over the lead time. As simple as this sounds, and as much as this is encouraged by standard textbooks, there are at least five different possibilities to make significant mistakes.

Many companies start off on the wrong foot when they choose the historical horizon for analysis. In many cases, the horizon is too short, and therefore the calculations are very noisy and the fit is poor. In other cases, the horizon is too long and the standard deviation captures product life-cycle effects rather than the short-term variability. The challenge here is that the analysis horizon has to be several multiples of the lead time. However, we must also test whether the demand pattern has significantly shifted during that time period due to product life-cycle, cyclical, or other explainable factors. With long lead time items, you may often find yourself in a situation where there is no horizon that satisfies both constraints.

Companies often make a series of errors when computing the standard deviation over the lead time. Most companies started with the textbook implementation which prescribes the scaling by the square root of the lead time as described above. However,

that formula makes two strong assumptions: demand parameters do not change over time, and demand between periods is independent. For example, when looking at daily increments, you may run into trouble if you have a significant number of customers that only order on specific days. Similarly, when looking at weekly increments, you may find numerous examples where the production schedule called for 100 units for two consecutive weeks, and where you consistently produced 200 units in consecutive two-week periods, but where you did not actually produce 100 units in any given week. As a way of benchmarking your own progress, many companies have already identified this issue and pursued other methods for calculating the standard deviation of demand over the lead time.

As an alternative, some companies compute the standard deviation of lead time directly. Consider the case where your lead time is 43 days. Starting with daily data, you define each observation of lead time demand as sum of demand over the previous (or future) 43 days and then calculate the standard deviation of these observations. The good news is that this method eliminates the issues associated with the previous calculation. The bad news is that it does NOT solve the horizon (amount of data required) problem. You can see this in the following example.

Assume again that the lead time is 43 days and that you have 60 days of shipment history. How many data points do you have? One calculation would suggest that you have 17 observations: points 1 to 43, 2 to 44, 3 to 45, ..., 17 to 60. However, each of these observations has 42 points in common with each successor and predecessor, and all 17 points have at least 26 points in common. Therefore, each of the data points is highly correlated. As a result, when computing the standard deviation, you often (but not always) will dramatically underestimate the true standard deviation of the process. If you have considerable history (several multiples of the lead time) then this problem disappears. But we have already established that needing a long history is a problem in its own right.

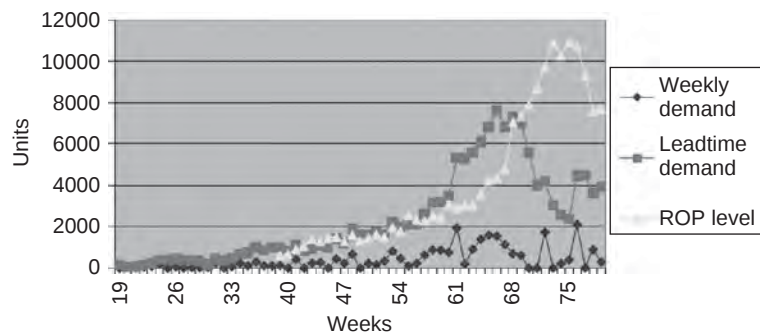


Figure 1. Example of a consumption-based ROP performing poorly in the face of ramping/declining demand.

Modified Consumption-Based ROPs

In some cases, companies acknowledged that they need a long horizon to estimate the standard deviation, but used a shorter horizon to determine the average run-rate in their kanban calculation. The motivation here stems from the notion that more recent demand is a better proxy for the upcoming demand. This approach is simply invalid. Mathematically there is no justification to use an average from one calculation and the standard deviation from another. To understand this, consider the definition of the variance (the square of the standard deviation): it is the average of the squared distances from the mean (average) and thus is a function of that mean. However, in this case, the mean of the longer horizon has been deemed invalid. Thus, if you are not satisfied with the average associated with the older data, then you cannot be satisfied with the standard deviation from that same older data.

The following example illustrates several of the issues that we see with consumption-based ROP calculations. The example shows a product with a six-week lead time that consistently ramped up for an extended period and is beginning a period of decline (Fig. 1). If shown, the company's forecast for this item would show an expected ongoing decline in volume.

As discussed above, we are interested in the demand over the lead time (labeled as lead time demand) and is the sum of weekly demand over the lead time looking forward. The ROP has been calculated using mean and

standard deviation parameters calculated on 26 weeks of history. A direct comparison of the ROP and the lead time demand gives us a visual assessment of how well the ROP would perform. If the lead time demand is higher than the ROP, then stockouts will result; if the ROP is much higher than the lead time demand, there will be excess inventory.

This example highlights three critical points. First, the consumption-based model will never catch up to ramping volumes. Second, as soon as the ramp becomes a decline, the consumption-based model will overshoot actual consumption severely. While we have shown an example with only six-week lead times, a significant portion of the typical high-tech supply chain has lead times in excess of 14 weeks. With a 14-week lead time and a 26-week horizon, it will take 2+ quarters for this problem to work itself out. Third, the company knew this was a ramping product and the consumption-based model pays the price for not using that knowledge. We see very similar behavior whenever there is explainable variability in demand: for example, within-quarter hockey-stick effects, seasonality, price changes, or product life cycles. Figure 2 shows another example without this ramp effect, but still clear cases where the reactive nature of this consumption-based ROP methodology systemically moves in the opposite direction of demand.

In summary, consumption-based ROP models face almost insurmountable challenges when applied to high-variability,

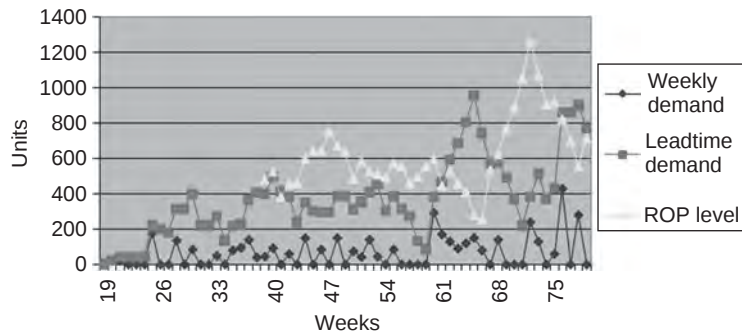


Figure 2. Example of a consumption-based ROP performing poorly even without a sustained upward or downward trend in demand.

high-volatility demand patterns, particularly as lead times grow. Too much history and your parameters no longer reflect the current reality (in the above example, history becomes irrelevant quite quickly); too little history and your critical model assumptions break down. Combine this with the fact that the methodology does not work for any new products, and you come to the conclusion that this approach has little to offer.

Before we turn our attention to using a forecast-based ROP calculation, we leave with you two last related questions. If you decide that a simple model based on historical consumption is the best predictor of future material requirements, then why have sales, marketing, and management generate forecasts? Conversely, if you have a forecasting process and it is proven to add value, then why would you choose a less predictive method (pure consumption) to drive the most critical components in your supply chain?

Partially Forecast-Based ROP Calculations

As suggested above, many companies quickly realized that a pure consumption-based ROP calculation did not respond quickly enough to changing demand situations. As such, they have experimented with a variety of efforts to incorporate forecasts into their ROP calculations.

We will start this discussion by looking at consumption-based ROPs as one class of forecast-based ROP calculation. To understand this connection, recognize that

the average and the standard deviation of historical consumption are only forecasts of the average and standard deviation of future demand. In fact, driving your ROP based on consumption is equivalent to using a moving average forecast with the moving average parameter equal to the amount of history used in the calculation (Not surprisingly, this typically is a poor forecast, but that is the subject of another paper.)

This raises the final objection to the standard practice used in estimating the standard deviation in consumption-based ROP calculations. Given that the consumption data provides only a forecast, the standard deviation should reflect the forecast error, not the demand volatility. In the example shown above (14-week lead time product) the standard deviation over the preceding 26 weeks was 2801. However, if you took the historical run-rate over the preceding 26 weeks as the forecast and measured the standard deviation of the forecast error, the resulting standard deviation is 3911, more than 30% higher. Moreover, if the business forecast (purple line) proves to be at all accurate and volume declines, the standard deviation based on historical consumption will decrease, but the forecast error associated with the consumption-based run-rate will increase. This dynamic will only worsen the performance of the consumption-based ROP.

Turning our attention to ROP calculations that explicitly use forecasts, we commonly see another set of mistakes. Similar to the above discussion, the problem originates

from efforts to set the cycle stock using the forecast, and the safety stock using historical consumption. We refer to this as the Partially Forecast-Based ROP calculation because it still uses consumption-based data to calculate the standard deviation. This fails the same test as the consumption-based ROP calculation that used one horizon to calculate the standard deviation and another to determine the average run-rate. In this case, the standard deviation is measured as the distance from an average based on a historical consumption rate which may have no functional relationship to the forecast, and which in many cases may be wildly different. Putting aside the theory, clearly, it makes little sense to take the standard deviation of the historical data if the historical average is two times or one half of the current forecast—and yet many companies do just that. It is also worth noting that this methodology fails for new product introductions because of its reliance on there being sufficient history to calculate the standard deviation.

Finally, some companies have attempted hybrid approaches in which the historical run-rate, the historical standard deviation, or both are scaled based on the difference between the historical average and the forecast. Specifically, we have seen companies take a weighted average of the historical run-rate and the forecast. Given that this results in another forecast, by our arguments above, there is no basis to use the historically generated standard deviation. By the same argument this also eliminates any effort to scale the standard deviation. Once again this should also fail the common sense test. Typically, as volume increases, forecast error (measured as a percentage) decreases in a nonlinear fashion. So again, while these approaches may feel intuitive, the culmination of all of these mistakes should leave the decision maker with no confidence that they have the “right” inventory levels, or even that they have (not?) consistently overstated or understated the inventory levels. In fact, the company is probably better served using an “unscientific” method than using any of the above methodologies.

SETTING SERVICE LEVELS

The task of determining the optimal target service level is both theoretically and practically challenging. In fact, the topic is sufficiently complex that it is more appropriate for a more advanced text. However, there are two common pitfalls that the practitioner should avoid when selecting the service levels.

One of the most common mistakes is targeting the same service level for all products, customers, or segments. The process of selecting service levels should begin with the specification of the business requirements. This entails determining the response time and service level requirements for each customer/market segment. As we would expect, not all customers, segments, or markets will have the same response time or service level requirements. Similarly, different products will have varying profit margins, lead times, and inventory carrying costs (physical storage, depreciation, risk of obsolescence, etc.). As such, it should be obvious that each of these categories should have a different service level.

Another common set of mistakes occur when companies convert target service levels into “z-scores” for the normal distribution. Putting aside that the demand variability and forecast error often are not normal, a more fundamental problem results from the confusion between two types of service level: “type 1”, which is what the academic literature refers to as service level, and “type 2”, which is also referred to as the fill rate. The type 1 service level measures the probability of satisfying 100% of the orders on time, whereas the type 2 fill-rates measure the percent of orders satisfied on time. The chart below shows how these measures differ (Fig. 3).

Some of the confusion can be traced from poor translation of the academic models into practice. It turns out that in the “Newsvendor” model (one of the best-known results in inventory theory), the optimal solution specifies the type 1 service level based purely on the economics of the situation. However, in practice, most companies cannot specify the

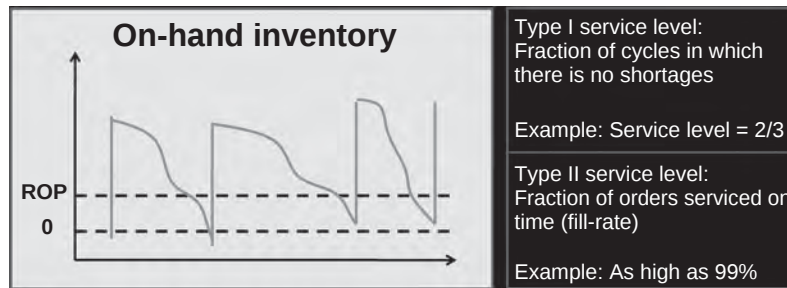


Figure 3. Distinction between Type I and Type II service levels.

necessary parameters to apply the Newsvendor formula. Instead, they (often appropriately) make an assessment of the optimal service level based on the criteria discussed, but they actually specify the type 2 fill-rate target rather than the type 1 service level. In fact, we have never met a manager who expressed any objective resembling the probability of a stock-out as a desired target. Nonetheless, each and every time the manager expressed their target service level, it was treated as a type 1 service level metric in the derivation of the z-score. As you can see from above, this almost certainly results in driving inventory levels that are much too high.

CLOSING REMARKS

In this article, we have focused on the pitfalls of trying to implement standard statistical inventory sizing techniques in environments with long lead times and elevated levels of demand volatility that result from seasonal patterns or short product life-cycles. The standard textbook formula only has three to five parameters (depending on the form of the model) and yet there are numerous opportunities to get each of them wrong. Remarkably, this is true even though we limited our discussion to the simplest of all settings: a single, stand-alone item with inventory being held in one location at one level in the supply chain.

Of course many, if not most, real-world kanban or ROP implementations face additional complexity, such as multiple items in a Bill-of-Material relationship, or

multiple inventory locations, or multiple levels ("echelons") in the supply chain. As with the single-item, single-location setting, many of the common practices and rules-of-thumb that are used to deal with this added complexity seem intuitive—and yet are fundamentally incorrect. For example, this is true of most "ABC" policies, and of the practice of multiplying service levels in assemble-to-order environments to approximate the ship-set end-product service level.

Faced with so many pitfalls, it is reasonable to ask whether there is any "right" way to statistically size inventory levels, and if there is, whether it is so complicated as to be not worth the effort required. The good news is that the right answer is usually not particularly complicated. The bad news is that the right answer is different in most situations. Nonetheless, some useful general principles are as follows:

- Make the best forecast you can and use it, unless the lead time is very short and demand variability is very low.
- Measure cumulative forecast error over lead time (rather than measuring only the one-period ahead forecast error and multiplying by the square root of the lead time).
- Where there is not enough history to measure this well, reference "similar" items with more history.
- Consider using a simulation-based approach with the real demand and forecast data to find optimal inventory settings.

- Monitor forecasts and stockouts/inventory levels to detect cases where the forecast model and/or ROP model may no longer appropriate because of some change in demand behavior.

Just as we see in effective implementations of statistical forecasting techniques, the right solution requires a combination of analytical rigor (to consistently set the appropriate inventory levels) and effective exception management (to quickly and systematically identify the inevitable misses that any analytical solution will have from time to time). If this article has equipped you with an awareness of the most common pitfalls in the former, and the consequent need for the latter, then you are well on your way toward implementing an effective kanban or ROP mechanism.

REFERENCES

1. Wang S, Sarker BR. Optimal models for a multi-stage supply chain system controlled by kanban under just-in-time philosophy. *Eur J Oper Res* 2006;172:179–200.
2. Available at http://www.flowvision.com/Lean_Supply_Chain.html
3. Nahmias S. *Production and operations analysis*. 6th ed. Chicago (IL): McGraw Hill Higher Education; 2008.
4. Iglehart D, Karlin S. Optimal policy for dynamic inventory process with non-stationary stochastic demands. In: Arrow KJ, Karlin S, Scarf H, editors. *Studies in applied probability and management science*. Stanford (CA): Stanford University Press; 1962. pp. 127–147.
5. Song J, Zipkin P. Inventory control in a fluctuating demand environment. *Oper Res* 1993;41(2):351–370.
6. Dong L, Lee HL. Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Oper Res* 2003;51(6):969–980.
7. Gavirneni S, Tayur S. An efficient procedure for non-stationary inventory control. *IIE Trans* 2001;33:83–89.

WRITING ORMS/ANALYTICS CASES

MEHMET A. BEGEN
SRINIVAS KRISHNAMOORTHY
and JOHN WILSON
University of Western Ontario,
Richard Ivey School of
Business, London, Ontario,
Canada

INTRODUCTION

Many of us have experienced the situation where students who are adept at the technical aspects of management science, statistics, or mathematics are unable to apply these methods in real-world situations. This can be the case even for students who have achieved the highest scores in their quantitative classes. We certainly have found this to be the case in our first-year MBA and HBA (honors in business administration) classes: prior success in quantitative methods is not necessarily a good indicator of adeptness in our beginning case-based classes. It is distressing to see students with good quantitative skills unable to parley this into management advantage. The case method helps with resolving this issue.

A primary goal of the case-based method is to simulate situations that a manager would experience. By learning through cases, students become more proficient in applying quantitative methods in a practical setting. In addition, there are other benefits. Students learn important presentation, writing, and discussion skills. Presentation of results is as important (or even more important) than being able to derive the results. After all, what use is a superb management science analysis if no one can be convinced of its merits or understand the managerial ramifications, strengths, or weaknesses? The case method requires that students develop the necessary verbal skills to identify and discuss managerial issues and then identify and implement appropriate quantitative

methodologies. Harvard, Ivey, and Darden are business schools that are most prominent in their use of the case method. A number of other schools have begun to use it as well. The method can be used in a wide variety of ways, for example, pure case teaching (in which there is a case in each class) or a hybrid teaching (a combination of cases and lectures).

This article provides a reference for instructors who wish to write cases. There are many reasons why an instructor may want to write a case. It is easier for an instructor to teach a self-written case rather than the one written by someone else. Writing a case compels the instructor to think hard about what and how he or she wants to teach in the course. By writing a case, an instructor can fulfill the need to find suitable cases for his or her teaching objectives. Teaching customized material and being the author of cases can greatly increase an instructor's credibility with students and enhance the learning process. In fact, students often want to be involved with case writing as they see it as a good training experience and feel that they are "leaving their mark" in the school. Sometimes, writing a case can create ideas for research. In addition, there are publication outlets for cases and innovative teaching. Often, for the purposes of examinations and assignments, an instructor looks for cases that students have not encountered before. The best way to find a case that is not widely available in the public arena is write a new one. Writing a case allows an instructor to customize the case (in terms of difficulty level, topic, industry, data, and operations research and management science (ORMS) concepts) for the examination.

There are a wide variety of cases available for the purposes of meeting different learning objectives. In addition, there are various approaches and philosophies when it comes to writing cases. Typically, cases are structured according to the following model: a background description of the company and industry; a description of the decision problem facing the manager;

related background information; and tables, graphs, and other data that could relate to the decision. It is important to be comprehensive in supplying material that might aid the decision-maker. Often, extraneous information will be included. This is in order to mimic the real world where one has to sift through the data to identify the right information. Students (and sometimes instructors) may find this frustrating but it is necessary as it reflects reality and helps them develop the skills necessary to identify information relevant to the decision at hand.

We feel that it is important to realize that there is not a “one-size-fits-all” approach to case writing. Flexibility in approaching case writing leads to a more varied and enjoyable experience. In this article, we share some of our experiences in the hope that these might be helpful to others. We certainly do not intend to be prescriptive. However, we feel that there are some common approaches and guidelines that can be of assistance. We will illustrate some points with cases that we commonly use. Sections on types of cases and publishing cases are included as these can be helpful resources for someone contemplating case writing. We feel that a teaching note is part and parcel of any case and include a section on this as well.

We believe that this article will be useful for teaching in business schools as well as industrial engineering and operations research departments. Even though engineering faculty spend more time on theoretical concepts in their courses as opposed to business faculty who focus on the managerial implications of a concept, cases can still be written and used by either faculty.

LITERATURE REVIEW

A number of authors have shared their ideas on writing cases. Linder [1] describes his ideas for writing cases using a timeline that includes preparation, interviewing, draft, quote approval, case review, and sign off. An exhaustive resource for case writing principles is available online at the National Institute for Case Teaching in Science [2]. Abel [3] provides an elegant description of the principles for a well-written case. A complete

guide to the case writing process is provided by Leenders *et al.* [4]. More recently, Lane [5] discusses the activities in organizing and presenting information in the case format. In similar vein, Roberts [6] describes the process of writing business cases. His article offers guidelines for identifying a good case, the process of working in the field with the company, and, finally, techniques for writing a good case. Drake [7] presents a compelling argument for writing and using cases. For a review of articles that deal with the case method of teaching, see Begen *et al.* [8]. These articles can certainly be useful for those contemplating writing a case.

IDEAS FOR WRITING CASES

Some instructors are intimidated by the idea of writing a case because they believe that it has to be inspired by a real-world consulting project that resulted in an Interfaces’ article or Institute for operations research and management sciences (INFORMS) Edelman Prize finalists’ presentation. Of course, real projects undertaken by the instructor are great sources for cases but ideas can also come from a variety of other sources as well. In our experience, business contacts and students are a great source of ideas. Many of our students undertake field projects for clients as part of their course requirements. We often use student internships or consulting projects to find a real-world problem and develop as cases. This was certainly the way the idea for the *Deloitte Recommends Client Selection to Regency Bank* [9] case was born. Executive MBA students are adept at basing their case ideas on their current jobs. One can also use industry contacts to identify business issues/challenges and obtain data in order to develop and write a case. *Seoul National Bank: The Chief Credit Officer’s Dilemma* [10] is an example of a case that was written using such an approach.

Cases can be fictional as well. Ideas for fictional cases are a product of the author’s imagination. The spark for the idea may come from reading an article in a newspaper or magazine or even having a conversation with a colleague. *SIGA Technologies: Profit-ing From Uncertainty* [11] was written based

on a newspaper finance article. Sometimes, the writer's research may motivate fictitious or real-world cases that capture management insights from the research. (*Kimpton Hotels* [12] and *The Clonlara Hotel* [13] are exemplars.) Whatever the source of ideas, it is important that the case follows a basic structure.

TYPES OF CASE

Cases can be used to introduce specific quantitative topics in a manner that has students immediately see the applicability of the techniques. It can help to provide a series of straightforward questions that will focus students' attention on the relevant issues and lead them toward a more complex analysis. This approach can be seen, for instance, in *Superior Grain* [14] where students are asked to fill out a worksheet to develop some important preliminary insights. The case *Buckeye* [15] is a product mix selection case. Appendices to the case contain a consultant's report and an outline of linear programming (LP) approaches. Here, students realize that LP is the most appropriate technique for the problem but they are given no help in setting up the analysis. Instead, a description of the problem and associated constraints is provided in a nontechnical matter that would closely resemble the information that managers would receive. For these cases, we sometimes would include a "mini-lecture" on some of the more technical aspects.

Cases that are written as a consultant's report can provide for lively discussion. Here, the format is similar to the classical case: company/industry description, information on the decision and possible data, and so on. However, a consultant's report is included. Here, the manager is in the situation of receiving a consultant's report and trying to understand, evaluate, and, possibly, implement the recommendations. One case along these lines is *Lansink* [16]. It is important to write the case in as neutral a manner as possible so that the students are not prejudiced to approach the consultant's report either negatively or positively. This situation of receiving a consultant's report and trying to

digest and evaluate is something that most managers will experience.

"Capstone" cases are comprehensive cases that require students to utilize many techniques. Students find these cases not only extraordinarily difficult but also most rewarding. We often have students tackle such cases in teams so that all students, no matter what their strongest skills might be, can contribute. A good example of such a case is *Enron* [17].

A case such as *Cells for Life* [18] follows a completely different tack. No data is supplied in the *A-case*. Instead, students are asked to put themselves in the positions of consultants who have been brought into help with decision-making. As such, they are to ask questions and request data that will help in making the decision. Our experience with this case has been very illuminating: for the *A-case*, no analysis is required. Indeed, no mathematical skills whatsoever are required at this stage of the analysis. Yet, most students—both those who are mathematically adept and those who claim to have no mathematical skills—seem to find this assignment remarkably difficult. Once a discussion on these issues has concluded, we then supply the students with data they requested and extra data that might be useful and ask them to prepare a consultant's report. We have used this case both at the beginning of our course to introduce the idea of the case-based method and near the end to have students tested on how their skills have evolved.

Cases are like stories: they can be real or fictional. When based on an actual company, the names and data may be frequently disguised for confidentiality purposes. When based on public events, for instance, the US presidential election or the World Series Baseball, there is rarely a requirement to disguise data although imaginary characters may be introduced to add a storytelling element to the case. Fictional cases are based on characters, situations, and data created by the author and are often sometimes to as "armchair" cases. Cases based on an actual company require the company's commitment to share information. Often requiring site visits and meetings with management, writing

a real company case is an exposure to the practicalities of real business situations. The company must “sign off” on the final version of the case, giving the author the right to publish the case and use it in a classroom or examination setting. (For instances, where the data is sensitive, a company’s consent to publish can often be obtained by an effort to disguise the data, e.g., via randomization.) Real cases are usually more difficult and challenging to write but they are believed to have more teaching and learning benefits, see Bell and von Lanzener [19].

There are also technical notes that are not really cases but are written to describe a certain topic, for example, logistic regression [20] solver in Excel and VBA [21], cost-effectiveness analysis [22], and introducing software [23]. These are meant to be assigned as customized readings and supplementary materials for students.

WRITING CASES

The case writer should have clear objective(s) in writing a case. For example, it could be to illustrate the use of a technique or techniques (e.g., LP, statistical analysis, simulation, or a combination of these concepts), to introduce a new business/opportunity (e.g., a new type of web-based auction), or to teach an ORMS concept (e.g., revenue management) or a blend of these. The difficulty level should also be properly adjusted for the desired learning purposes. For example, certain cases may provide alternative decisions and ask readers to evaluate each one of them and choose among them, whereas some other cases do not reveal alternatives or even give a hint about the decision(s) but just introduce issues. The data itself can provide difficulties: some of the facts/data provided in the case may or may not be relevant for the issue(s) and readers may need to make some sensible assumptions to see which ones matter. Perhaps, some required data is not provided and the reader evaluates alternatives for a range of inputs to see how decisions might change. Sometimes, cases are meant as introductions to certain techniques. For instance, suppose, as an instructor you are teaching optimization for the first time with cases. You may

want to start with an easy case: specifically, a case with simple LP where the body of the case provides hints about decision variables, objectives, and constraints so that readers can build their first optimization model. The next cases may require the use of a more advanced LP model (with no hints), binary variables, or non-LP. Before writing a case, the case writer has to consider the difficulty level that he or she is targeting.

A case is typically written with the following structure:

Introduction

The first paragraph of the case introduces the main character and the business problem. Some writers regard this as the most important part of the case. It entices the student to read the rest of the case. Readers are often won or lost in the introduction.

Cases should attract the interest of students by the first/opening paragraph if not by its title. Current cases usually get more attention than old ones. Real and famous companies (especially in the title) attract more interest. Cases should have a hook that sparks the curiosity of the reader after the opening paragraph.

Body

The body of the case provides further details of the business problem. Some secondary characters may be introduced as well. There are varying opinions among case writers about including extraneous material. Some believe that a more comprehensive view (even if it includes much material not immediately relevant to the problem at hand) captures the messiness of real business situations. Others feel that, while it is essential to provide the reader with the facts surrounding the business problem, care has to be taken to prune unnecessary details.

Background information is important in that it sets the problem in the broader framework of the business world and engages students on general issues beyond the quantitative methods they will eventually employ. It provides context and support for requiring an understanding of quantitative approaches. The material need not be

particularly relevant for the issue at hand as long as it enables the student to see this problem in the wider context of a general business situation. It can be helpful to personalize the issue, for example, phrases such as “Sally had a deadline for making recommendations on. . .” The goal here is to write enough so that student can envision being in the decision-maker’s shoes. The inclusion of relevant newspaper articles or company advertising material can add to a sense of comprehensiveness.

Like Aggarwal and Khera [24], we personally prefer shorter cases written in plain language. We find this to be most effective for analytics cases given the extra complications that quantitative analyses bring to cases. However, there are some case writers who have a different opinion. They prefer to write longer cases by including extra facts. For instance, Harvard cases are usually written in great detail.

Conclusion

The final paragraph reiterates the task faced by the main character. Any deadline for the completion of the task has to be emphasized here to capture the urgency of the situation.

Exhibits

The exhibits provide data that may help with the analysis. Care has to be taken to provide data in an authentic format—for instance, we have written cases in which some of the data items are missing. After all, data recording is an error-prone activity in corporations.

We feel it important that supporting data be as general as possible. It is rarely the case in the business world that one has only the data set that is relevant to the issue at hand. For instance, in a product mix problem, information on many products might be available. We often use *Red Brand Canners* [25] as an introduction to LP. This is a product mix case where the immediate decision for the decision-maker involves specific products. However, the information in the tables consists of information on many products. Also included is information on the overall and individual performance for various product lines. For the student, developing the

capability to find the right information from the plethora of available information is an important skill to learn and improve.

Cases allow students to solve real-world problems and involve them in decision-making process in a controlled laboratory-like environment. Cases should give enough background, facts, and information about the issue(s) but should not imply a solution method or provide any analysis, judgment, or prediction about the outcome. These are for readers to find and figure out themselves. Cases should not be easily “cracked”; the solution approach and answer should not be obvious but should be obtained after thoughtful analysis putting different pieces of the case together. Analysis and learning points should not be part of a case but the corresponding teaching note should have them. A case should be accompanied with a good teaching note. This increases its chances to be adopted and used by other instructors as well as forces the case writer to think thoroughly about the case, its objectives, solution approach, and its takeaways and make corrections and improvements before the case is finalized. For our cases, teaching notes are usually accompanied with (spreadsheet) models and data sets. PowerPoint slides and any video clips (e.g., of the company or process being studied) that illustrate the problem can be valuable additions.

To stimulate real-world decision-making, it is crucial to write cases that do not always lead to a single solution. Ideally, a case should lead students to identify multiple alternatives and figure out the trade-off of choosing one alternative over another. Of course, in some instances, a decision could be a yes-or-no type question but even in this situation, there can be different ways to formulate and reach an acceptable answer. A good case should allow students to identify minor and major issues and symptoms and enable them to discover their own methods/approaches to solve the case puzzle and come up with sound decisions and action plans.

It helps considerably to read and teach, see a few cases, and watch or learn about their use in classroom before attempting to write any. It is useful to read cases not only

in ORMS but also from other fields to have a broader understanding. Finally, reading cases from different publishers, authors, and books makes us more versatile case writers. For example, see Bodily [26], Liberatore and Nydick [27], Liang and Wang [28], Kim *et al.* [29], Currie [30], Long [31], and Cochran [32–36] for different use and type of cases.

Unlike many other subjects, ORMS cases usually have multiple aspects, for example, quantitative and qualitative, managerial, and data identification/extraction and analysis. (Depending on the case objectives, the case writer may give more or less emphasis to these topics, e.g., an introductory case on decision trees for business students and an advanced case on simulation for engineering students would be different in terms of technical difficulty, data, problem description, and teaching note.) Another difference is that there is usually more ambiguity, uncertainty, making sense of data, and obtaining information from it in ORMS cases. Furthermore, in order to teach a technical subject, cases difficulty levels vary more in ORMS. Writing case is a dynamic art. Often, a case will have to be revised based on the writer's experience using it in class. Students often provide unique perspectives that enrich the instructor's understanding of the business situation. By being flexible in our approach to case writing, we can enjoy this exercise of creating learning experiences for our students.

TEACHING NOTES

A teaching note is a valuable addition to the case. Indeed, we feel that a case is incomplete without an associated teaching note. A teaching note is typically written with the following structure:

Introduction

Here, the case is summarized by describing the main character and the business problem with a few details about the company.

Techniques and Concepts

A list of techniques and concepts covered helps the instructor fit the case in the course plan. An example of a list could be:

- decision analysis
- multiple regression analysis
- constructing confidence and prediction intervals
- simulation
- optimization.

Uses/Audience

Often, cases are written with a target audience in mind. By clearly defining the audience, we can ensure that the instructor chooses cases that match the expertise of his or her students. An example of a target audience would be undergraduate business students in a foundational analytics course.

Assignment Questions

Case discussions work if students spend time outside the classroom preparing for them. Assignment questions provide a framework for the preparation. It is important to keep the first couple of questions as open-ended as possible. Care has to be taken not to steer students to an “answer” or solution approach. One of the questions we like to use in our cases is what is the problem that Mr. XYZ is trying to solve or what is the decision that Ms. ABC has to make? Often, asking the students to provide a time line including when data is collected and decisions are made can help them understand the decision-making process a little better.

Suggested Teaching Plan

The teaching plan is meant to be a guideline for the instructor—not a script that the instructor should follow. A good plan will list suggestions for starting, directing, and concluding the discussion. The time to be spent in each phase should be provided as well. Once again, it has to be understood that the plan is not rigid—every instructor is going to have his or her own plan and style. Indeed, one should be wary of successful case teachers who insist that theirs is the

only way to proceed as in many cases it is their own idiosyncratic styles that successfully work for them. Others need to develop styles that work for them while, of course, observing and adapting the styles of others.

Solution Approach/Analysis

Providing a solution approach is necessary for the sake of completion. A spreadsheet file may also be provided to illustrate the modeling approach adopted. However, we recommend that the instructor solves the case on his or her own rather than mimicking the approach provided. In our experience, a case is best understood by solving it. In any case, there is rarely one correct way of solving a case.

Case Outcome

For real cases, it is useful to describe in the teaching note what actually happened in terms of the decision made and the business outcome. In our experience, memorable cases invariably arouse the interest of students with respect to the final outcome.

WRITING CASES FOR ENGINEERS

The biggest challenge that an instructor will face while writing and teaching cases for engineering students is that most engineers are used to the “teach me the tool, give me the problem, use the tool, and get the solution” approach. In reality, engineering problems in corporations are complex and ill-structured. There may not be one “solution” to the problem. We believe that the case writer should not succumb to the temptation to write a neatly packaged textbook case. Most engineering students become managers in corporations. It is imperative, therefore, that they learn to solve ill-structured problems.

We believe that engineering cases should be written in the same format and structure as business cases. Of course, the case writer may choose to include more technical details because engineering students are likely to have a stronger training in quantitative concepts as compared to business students. The case approach, however, is very useful in forcing them to realize that the communication

of their results is at least as important as the results themselves. However, the writing philosophy remains the same.

WRITING CASES IN OTHER BUSINESS DISCIPLINES

We find that cases written in other business disciplines such as marketing, strategy, and organizational behavior tend to be longer and more detailed. The data provided in the cases may be more at the strategic and structural levels, for example, industry structure, company financials, and company organizational charts. This is probably a legacy of the qualitative approaches used to tackle such cases. Students tend to spend more time reading the case and identifying key facts that may be used to make a decision. In ORMS cases, the data provided tends to be more at the tactical and operational levels, for example, demand and price schedules and processing time for a product. The students tend to spend less time reading the case and more time creating models and reaching conclusions based on the output of their model.

PUBLISHING CASES

Universities and business schools that use the case method in their teaching can sometimes be outlets for publishing cases. Often, such places will have a publishing arm devoted to cases. For example, the Harvard Business School [37], the Richard Ivey School of Business [38], Darden School of Business [39], Insead [40], and the Stanford Business School [41] accept cases from authors outside of their own faculty in addition to their own. There are also educational organizations such as the European Case Clearing House [42] that publish and distribute cases. Specialized journals such as *INFORMS Transactions on Education* [43] publish a variety of articles in the areas of ORMS teaching: cases, applications of spreadsheets, reviews, and opinion articles. The journal is available only in electronic format and allows open access of its articles to interested parties. Registered instructors can also obtain resources such as data sets, software, and spreadsheets.

The INFORMS annual case competition has resulted in the dissemination of several high quality cases. The competition was established in 2000 to promote the writing and use of cases in ORMS. The case competition web site [44] is another place to find cases. Entering this competition is a great way to get recognition for a case. (Cases submitted to the competition cannot have been published previously.)

The Journal of Business [45] is a peer-reviewed journal that provides high quality teaching and research cases in all major disciplines of management science. The Operations Management Education Review [46] publishes and sells peer-reviewed teaching cases. The Decision Science Journal of Innovative Education [47] publishes research pertaining to teaching and learning issues.

Textbooks on topics such as operations management, analytics [48], management science modeling, operations research, and quantitative decision-making are also sources for cases. Although publishing a case in a textbook is an option, unless you are authoring the textbook or know the author(s), it may not be the easiest way. Furthermore, most instructors who need cases usually find a suitable case for their needs at one of the university case-publishing web sites such as in Refs 37 and 38.

Certain academic centers, especially the ones doing projects with industry, may publish some of their projects as cases and make it available for academic use. Once such example is Centre for Operations Excellence (COE) in the Sauder School of Business at the University of British Columbia [49]. At the COE, master students are required to do a significant industry project for their degree and this is done with a team consisting of the student, technical analysts, faculty advisor, and a project manager. Projects are usually six months long and they can be thought as management consulting projects specializing in analytics and operations research. Some of these COE cases were successful in INFORMS case competitions [50, 51].

CONCLUSION

Using cases in ORMS/analytics teaching is not only effective but can also be fun and rewarding for both students and professors. One important reason for writing cases is simply the fact that case teachers need current and relevant cases. In addition, sometimes as a case teacher may need a specific case (e.g., a technique, industry, or decision type) and cannot find it elsewhere. In such a situation, he or she is motivated to write it him- or herself. (For instance, this was the case with *Clonlara* [13] and *Kimpton hotels* [12] where the authors needed cases in the relatively new area of Revenue Management.) New cases are often required for case examinations. Writing a case can be an excellent way to get started on a research project. For example, a student internship can turn into a case and continue as a research project. The cases one writes will be easier to teach. Teaching one's own case increases the instructor's credibility among students. An extra benefit of writing cases is that they help build a teaching portfolio (which can be important for career development). The cases can even be published in journals, books, or university publishing houses. There is also the possibility of participating in case competitions [24] by submitting a newly written case. Case writing can not only be enjoyable and rewarding but also expand one's teaching horizons.

REFERENCES

1. Linder J. Writing cases: tips and pointers. Boston, Massachusetts: Published by Harvard Business School Publishing; 1994.
2. National Center for Case Study Teaching in Science. Available at <http://sciencecases.lib.buffalo.edu/cs/>.
3. Abel D. What makes a good Case? European Case Clearing House: The Newsletter of European Case Clearing House 1997;17:4–7.
4. Leenders MR, Mauffette-Leenders LA, Erskine JA. Writing cases. 4th ed. Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2001.
5. Lane JL. Schreyer Institute of Teaching Excellence; 2007. Available at

- <http://www.schreyerinstitution.psu.edu/pdf/CaseWritingGuide.pdf>. Accessed 2013 July 8.
6. Roberts MJ. (Original publication in 2001). Developing a Teaching Case. Boston, Massachusetts: Published by Harvard Business School Publishing; 2012.
 7. Drake M. The case for writing cases, ORMS Today; 2012. Available at <http://viewer.zmags.com/publication/02b7ca69#/02b7ca69/10>. Accessed 2013 July 8.
 8. Begen MA, Krishnamoorthy S, Wilson J. Teaching ORMS/Analytics with Cases, to appear in EORMS; 2012. DOI: 10.1002/9780470400531.
 9. Begen MA, Yue S. Deloitte recommends client selection to regency bank. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2011.
 10. Begen MA, Jhun J. Seoul National Bank: The Chief Credit Officer's Dilemma. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2011.
 11. Begen MA, Cui C. SIGA technologies: profiting from uncertainty. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2011.
 12. Anderson CK, Wilson JG, Read J. Kimpton hotels—setting prices on priceline (A). London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario.
 13. Wilson JG. The Clonlara Hotel. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2009.
 14. Bell PC. Superior grain. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 1998.
 15. Carraway R. Buckeye power & light company. Charlottesville (VA): The University of Virginia Darden School Foundation; 1996.
 16. Zaric G. Lansink appraisals. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2000.
 17. Capestany M, Bruner R, Bodily S. Enron Corporation's Weather Derivatives (A) & (B). Charlottesville (VA): The University of Virginia Darden School Foundation; 2000.
 18. Wilson J, Young R, Oppitz G, Hoy W, Bassin D. Cells for life. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2009.
 19. Bell PC, von Lanzanauer CH. Teaching OR: a review of current practice and developments. *J Oper Res Soc* 2000;51(12):1367–1377.
 20. Odegaard F, Brennan A. Note on logistic regression—statistical significance of beta coefficients. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2010.
 21. Begen MA, Xia W. Solver in excel and VBA. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2011.
 22. Zaric GS. Difficult choices—an introduction to cost effectiveness analysis. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2010.
 23. Anderson CK, Wilson JG, Craig M. Quick tour of @RISK. London, Ontario: Ivey Publishing, Richard Ivey School of Business, The University of Western Ontario; 2009.
 24. Aggarwal R, Khara I. Short cases for teaching management science: their role in closing the gap between theory and practice. *Interfaces* 1978;9(1):90–94.
 25. Wilson R. Red brand canners. Palo Alto (CA): Stanford University, Graduate School of Business; 1965.
 26. Bodily SE. The Teachers' forum: teaching MBA quantitative business analysis with cases. *Interfaces* 1996;26(6):132–138.
 27. Liberatore MJ, Nydick RL. The Teachers' forum: breaking the mold—a new approach to teaching the first MBA course in management science. *Interfaces* 1999;29(4):99–116.
 28. Liang N, Wang J. Implicit mental models in teaching cases: an empirical study of popular MBA cases in the United States and China. *Acad Manage Learn Edu* 2004;3(4):397–413.
 29. Kim S, Phillips WR, Pinsky L, Brock D, Phillips K, Keary J. A conceptual framework for developing teaching cases: a review and synthesis of the literature across disciplines. *Med Edu* 2006;40:867–876.
 30. Currie G. Moving towards reflexive use of teaching cases. *Int J Manage Edu* 2008;7(1):41–50.
 31. Long Z. On case teaching of operations research. Proceedings of the 2009 First International Workshop on Education Technology and Computer Science; 2009 01. p 547–550.
 32. Cochran JJ. Successful use of cases in introductory undergraduate business college operations research courses. *J Oper Res Soc* 2000;51(12):1378–1385.

33. Cochran JJ. Bowie Kuhn's worst nightmare—an integer programming & simpson's paradox case. *INFORMS Trans Edu* 2004;5(1):18–36.
34. Cochran JJ. Pedagogy in operations research: where have we been, where are we now, and where should we go? *ORiON* 2009; 25(2):161–184.
35. Cochran JJ. All of Britain must be stoned! An effective introductory probability case. *INFORMS Trans Edu* 2009;10(2):62–64.
36. Cochran JJ. You want them to remember? Then make it memorable! *Euro J Oper Res* 2012;219:659–670.
37. Harvard Cases. Available at <http://hbsp.harvard.edu/product/cases>. Accessed 2013 July 8.
38. Ivey Cases. University of Western Ontario. Available at <https://www.iveycases.com/>. Accessed 2013 July 8.
39. Darden Cases. University of Virginia. Available at <https://store.darden.virginia.edu/>. Accessed 2013 July 8.
40. Insead Cases. Available at http://www.insead.edu/facultyresearch/research/search_cases.cfm. Accessed 2013 July 8.
41. Stanford Business School Cases. Available at <https://gsbapps.stanford.edu/cases/default.asp>. Accessed 2013 July 8.
42. European Case Clearing House (ecch). Available at <http://www.ecch.com/educators/>. Accessed 2013 July 8.
43. INFORMS Transactions on Education. The Decision Sciences Journal of Innovative Education, Available at <http://www.INFORMS.org/Pubs/ITE>. Accessed 2013 July 8.
44. INFORMS Case Competition. Available at <http://www.INFORMS.org/Community/INFORM-ED/Activities/Case-Competition>. Accessed 2013 July 8.
45. The Journal of Business Cases. Available at www.revuecasgestion.com/soumi2E.htm. Accessed 2013 July 8.
46. The Operations Management Education Review. Available at www.neilsonjournals.com/OMER/. Accessed 2013 July 8.
47. The Decision Sciences Journal of Innovative Education. Available at <http://dsjie.org/>
48. Bell PC, Zaric GS. *Analytics for managers*. London, Ontario: Ivey HBA printing, Richard Ivey School of Business; 2010.
49. Centre for Operations Excellence, Sauder School of Business. Vancouver, BC: University of British Columbia. Available at http://www.sauder.ubc.ca/Faculty/Research_Centres/Centre_for_Operations_Excellence/Academic_Research/Case_Studies. Accessed 2013 July 8.
50. Puterman LM, Tse S. *Whistler Blackcomb Ski and Snowboard School* by Martin L. Puterman. Vancouver, BC: Centre for Operations Excellence, University of British Columbia; 2003.
51. Smith AM, Glenn D, Begen MA, Puterman ML. *J.S. McMillan Fisheries Ltd: Optimizing Production Planning*. Vancouver, BC: Centre for Operations Excellence, University of British Columbia; 2002.